

Systematic Analysis of the Factors Contributing to the Variation and Change of the
Microbiome

by

Kenneth D. Aiello

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2018 by the
Graduate Supervisory Committee:

Manfred Laubichler, Chair
Sara Walker
Michael Simeone
Kenneth Buetow

ARIZONA STATE UNIVERSITY

December 2018

ABSTRACT

Understanding changes and trends in biomedical knowledge is crucial for individuals, groups, and institutions as biomedicine improves people's lives, supports national economies, and facilitates innovation. However, as knowledge changes what evidence illustrates knowledge changes? In the case of microbiome, a multi-dimensional concept from biomedicine, there are significant increases in publications, citations, funding, collaborations, and other explanatory variables or contextual factors. What is observed in the microbiome, or any historical evolution of a scientific field or scientific knowledge, is that these changes are related to changes in knowledge, but what is not understood is how to measure and track changes in knowledge. This investigation highlights how contextual factors from the language and social context of the microbiome are related to changes in the usage, meaning, and scientific knowledge on the microbiome. Two interconnected studies integrating qualitative and quantitative evidence examine the variation and change of the microbiome evidence are presented. First, the concepts microbiome, metagenome, and metabolome are compared to determine the boundaries of the microbiome concept in relation to other concepts where the conceptual boundaries have been cited as overlapping. A collection of publications for each concept or corpus is presented, with a focus on how to create, collect, curate, and analyze large data collections. This study concludes with suggestions on how to analyze biomedical concepts using a hybrid approach that combines results from the larger language context and individual words. Second, the results of a systematic review that describes the variation and change of microbiome research, funding, and knowledge are examined. A

corpus of approximately 28,000 articles on the microbiome are characterized, and a spectrum of microbiome interpretations are suggested based on differences related to context. The collective results suggest the microbiome is a separate concept from the metagenome and metabolome, and the variation and change to the microbiome concept was influenced by contextual factors. These results provide insight into how concepts with extensive resources behave within biomedicine and suggest the microbiome is possibly representative of conceptual change or a preview of new dynamics within science that are expected in the future.

DEDICATION

To my partner and wife Ashlee Rachel Aiello.

ACKNOWLEDGMENTS

It is with great pleasure that I recognize a few special individuals who have provided guidance, direction, and support during this journey. Beginning with a number of key mentors who have inspired me throughout my education, Manfred Laubichler, Michael Simeone, Ken Buetow, Jane Maienschein, Sara Walker, Kevin McGraw, Mathieu Girardeau, Lisa Whitaker, and Jacqueline Hettel. Many thanks are due to the staff and faculty at the Center for Biology and Society and Global Biosocial Complexity Initiative including Jessica Ranney, Andrea Cottrell, Trish Yasolsky, and Karin Ellison for their assistance on the little details that were often the most important things during my educational career. I would also like to extend gratitude to Wendi Simonson and Amina Hadjarovic in the School of Life Sciences for help navigating graduate school. I am also grateful for my extraordinary colleagues Deryc Painter, Cody O'Toole, Erick Peirson, Julia Damerow, Bryan Daniels, and Nayely Velez-Cruz in the Laubichler Lab that shared their time, ideas, and opinions with me. Other motivating scientists and researchers I was fortunate enough to bounce ideas off of include: Feng "Bill" Shi, Brian Uzzi, James Evans, John Ioannadis, Robert Cook Deegan, Michael Crow, Chris Rojas, and Jose Lobo. Lastly, I am incredibly lucky to be blessed with the love and support of my family during this time. Most importantly, I would like to express gratitude to my wife Ashlee Aiello, for her patience, encouragement, and love.

TABLE OF CONTENTS

	Page
LIST OF TABLES	ix
LIST OF FIGURES	xi
CONTEXTUAL REVIEW OF THE MICROBIOME LITERATURE	13
Background	13
Historical Context	18
Language Context	23
Social Context	26
Summary	32
KNOWLEDGE: REVIEW OF PREVIOUS APPROACHES.....	33
Background	33
Biomedical Knowledge Approaches:.....	37
Systematic Reviews.....	38
Ontologies.....	40
Natural Language Processing (NLP).....	45
Topic Models.....	47
Corpus Linguistics: Corpora, Keywords, and Collocation.....	53
Publications and Citations as Knowledge Review	64
Concepts, Knowledge Maps, and Knowledge-Based Economy Review.....	69
Language as a Complex System: Studies on the Variation and Change in Language...77	
Knowledge and Context.....	85
Summary	92

	Page
Figures and Tables	94
BIG DATA AND DATA DRIVEN SCIENCE AND RESEARCH	103
Background	103
Context	104
Historical Context of Big Data.....	106
Complexity of Big Data	108
Metadata.	110
Data Driven Science and Research (DDSR).....	111
Exploratory Data Analysis (EDA).....	113
Replication.	113
Criticisms	115
Counter Argument to Criticisms: Context.....	116
Summary	118
Figures and Tables	119
CONCEPTUAL BOUNDARIES OF THE MICROBIOME, METAGENOME, AND	
METABOLOME	120
Abstract	120
Context.....	120
Objective.....	120
Results.....	120
Introduction	121

	Page
Historical Context.....	121
Objectives.....	125
Materials and Methods.....	126
Systematic Searches.....	126
Data Analysis.....	127
Results.....	132
RQ1: What are the Characteristics Differences of the Language Content of the Microbiome, Metagenome, and Metabolome?	132
Topic Models.....	132
Frequency Analyses.....	133
Keywords.....	134
RQ2: How is the Language Used with the Microbiome Similar/Dissimilar to the Language Used With Metagenome, and Metabolome?	136
Collocates & Lexical Profiles.....	136
Collocation Networks.....	138
RQ 3: What is the Microbiome Based on Usage?.....	139
Discussion.....	139
Figures and Tables	142
SYSTEMATIC REVIEW OF THE FACTORS INFLUENCING THE EVOLUTION OF THE MICROBIOME CONCEPT.....	164
Abstract.....	164
Introduction.....	164

	Page
Methodological Challenges.....	169
Contextual Review Based on History of Knowledge	174
Objectives.....	176
Results	176
How Has Microbiome Research Outputs Developed Over Time?.....	176
The Microbiome as an Innovation.....	177
Change and Variation in Knowledge.....	182
Discussion	189
Figures And Tables	194
CONCLUSION.....	214
REFERENCES	222

LIST OF TABLES

Table	Page
1. Association Measures for Words (M. Scott, 2018).....	97
2. Top 20 Collocates of Microbiome in MB Corpus by Different Association Measures.	98
3. Descriptive Statistics for the Microbiome Corpus, Metabolome Corpus, and Metagenome Corpus.	143
4. Topic Bins for Bacteria from MB Corpus Compared to Metab Corpus.....	146
5. Topic Bins for Bacteria from MB Corpus Compared to Metag Corpus.....	147
6. Comparison of Ranks and Frequencies of Words of Interest from the Corpora.	148
7. Positive Keywords in The MB Corpus Compared to the Metab Corpus.....	149
8. Comparison of Significant Keyword Categories MB Corpus Compared to Metab Corpus.....	150
9. Comparison of Significant Keyword Categories MB Corpus Compared to MetaG Corpus.....	151
10. Descriptive Statistics for Collocate Analyses of Node Word in MB Corpus, MetaG Corpus, and MetaB Corpus.....	152
11. Descriptive Statistics of 5,000 Article Random Samples (MB 1, MB 2, MB 3) Compared to MB Corpus	153
12. Comparison of Top 20 Shared Collocates	154
13. Top 20 Shared Collocates of Microbiome in MB Corpus.....	155
14. Top 20 Shared Collocates of Metagenome in Metag Corpus.....	156

Table	Page
15. Top 20 Collocates of Metabolome in Metab Corpus.....	157
16. Communal Collocates Between Microbiome and Metagenome.	158
17. Communal Collocates Between Microbiome and Metabolome.	159
18. Total Projects and Funding for Microbiome Research from Federal REPORTer....	194
19. WOS Articles Compared to Articles Collected.	195
20. Federal Reporter Microbiome Projects Compared to MB Project Corpus.	196
21. Research Outputs, NLM Mesh, and Authors Over Time.	197
22. Percent Increase Over Time.....	198
23. Comparison of Projects and Funding Between Microbiome, Metabolome, and Metagenome.....	199
24. Comparison of Microbiome Occurrences to Microbiome Collocates.	200
25. Topic Model Results on MB Corpus 2005 to 2017.	201
26. Comparison of Significant Keywords MB Corpus Pre-2007 to MB Post-2007.....	202
27. Knowledge Stability Over Time Measured by Jaccard Similarity Score.	203
28. Change in Knowledge Measured by Jaccard Similarity Score.	204
29. Spectrum of Shared Microbiome Collocates in MB Corpus.	205
30. Spectrum of Shared Microbiome Collocates in Nature Articles on the Microbiome.	206

LIST OF FIGURES

Figure	Page
1. Results For “Microbiome” in MeSH.	94
2. MeSH Descriptor Data 2019 for Microbiome.	95
3. UMLS Search Result for Microbiome.....	96
4. Frequency by Rank of One Text and Fourteen Texts.....	99
5. Examples of S-Curves Over Time.	100
6. Collocate Network for MB Corpus.....	101
7. Author to Publication Network of Microbiome Publications 1900 to 2014.....	102
8. Ansecombe’s Quartet as a Table and as a Collection of Four Graphs (Tufte, 1990).	119
9. Systematic Collection of Articles for MB Corpus.....	142
10. Word Frequency by Rank.	144
11. Topic Modeling Process and Results.....	145
12. Collocate Network of Microbiome :< Changes, Gut, Human>.....	160
13. Collocate Network of Metagenome :< Analysis, Gut, Human>.....	161
14. Collocate Network of Microbiome :< Gut, Human, Analysis, Changes>.....	162
15. Collocate Network of Metabolome: < Human, Analysis, Changes, Gut>.	163
16. Diffusion Process of an Innovation Over Time.	207
17. Adoption Rates of Microbiome in Different Systems..	208
18. Scholarly Network of MB.....	209
19. Collocates Associated With Microbiome From 2001 To 2017.	210

Figure	Page
20. Collocates Associated With Microbiome by Less Than 1.0%.	211
21. Change in Knowledge Over Time Measured by Jaccard Similarity (J).	212
22. Knowledge Convergence Over Time.....	213
23. Knowledge Convergence Toward MB 2017 Over Time.....	214

INTRODUCTION

The microbiome concept is a multidimensional object with various contextual interpretations, including but not limited to a microbial community in the human body, a microscopic biome, or the aggregate microbial genome of an organism. I argue context is necessary to better understand the different historical trajectories, variation, and changes in microbiome knowledge. In this dissertation, I use the term *context* to refer to a frame of reference. Social context includes actors, individuals, groups, institutions, settings or environments. Language context includes words, phrases, concepts, definitions, and meaning. Historical context includes events, moments, trends, people, or things.

This dissertation analyzes the contextual factors related to changes in the usage, meaning, and scientific knowledge of the microbiome. *Contextual factors* are explanatory variables which offer explanation into the microbiome phenomena, and include: historical events, words, phrases, language patterns, material artifacts, individuals, and social groups. Scientific knowledge or *knowledge*, is the encoded experience of actors with material and social dimensions that determines which actions are possible in a historical situation (Renn & Laubichler, 2017a). Understanding the transmission of knowledge and changes in knowledge requires analyzing the social and material dimensions of knowledge. Knowledge can be shared within a group or society via the use of material artifacts such as instruments or texts. Many contextual factors stemming from history, society, and language have been speculated as to influencing microbiome knowledge, but to date there is no systematic studies or empirical evidence to validate these claims. To fill this gap, this dissertation focuses on contextual factors that are directing measurable changes to the usage and meaning of the microbiome concept.

My initial unit of analysis for analyzing the knowledge of the microbiome are proxies of knowledge from the language contexts and the social contexts of the microbiome such as: words, phrases, individuals, groups, institutions, and the environment(s) of where the microbiome was used and interpreted.

This investigation answers the question, what contextual factors influenced the variation and change of the microbiome? In this project, I use an interdisciplinary approach to identify and analyze the variation and change of the microbiome within a dataset that includes over 55,000 articles from publicly available biomedical documents. Throughout this dissertation, I pay close attention to the relationships between social, language, and historical context into the development of the microbiome concept and the evolution of scientific knowledge. By using a combination of qualitative and quantitative methods, I highlight a spectrum of microbiome conceptualizations, and provide evidence of a link between the variation and changes of the microbiome concept to language and social contexts. Based on the empirical and qualitative results, this study brings new understanding to the influential historical, social, and language contexts that contributed to the development of the microbiome.

I have organized this dissertation into four self-containing chapters which detail my approach and the results of my investigation into the variation and change of the microbiome and the language and social context using the microbiome – (1) A contextual review of microbiome literature, (2) review of approaches measuring knowledge, (3) incorporating big data and data driven science and research (DDSR), (3) conceptual boundaries of microbiome, metabolome, and metagenome and (5) a systematic review of the factors influencing the language and social context of the microbiome.

In the first chapter, I provide background on the microbiome and overview the previous attempts to understanding the microbiome using context. I introduce the debates swirling the microbiome such as: the origins of the microbiome, who deserves credit for the microbiome, how the microbiome is used now, and what is the intended usage of the microbiome. I show how specific intellectual or social factors have been claimed to influence the microbiome including ecological interpretations, biomedical interpretations, government resources, and institutional norms. Ultimately, the conclusion of the review is the microbiome is still not well understood because of multiple microbiome interpretations emerging from different contexts, confusion on meaning and usage, and the extreme variation and change to the microbiome over time (Huss, 2014; Marchesi & Ravel, 2015a).

In the second chapter, I argue traditional approaches to measuring and analyzing knowledge within biomedicine, with particular emphasis on systematic reviews, ontologies, and text mining, do not accurately identify or measure changes in knowledge within biomedicine. In general, the relevant literature for analysis on context and knowledge spread across disciplines traditionally having few connections and include methods on: textual analysis, content analysis, and knowledge analysis (Stubbs, 2001; Labov, 2001; Blei & Lafferty, 2007; Cetina, 2009; P. Baker, 2010; Loet Leydesdorff, 2010; Renn & Laubichler, 2017b; Hesse, Moser, & Riley, 2015; Burkette & Jr, 2018). I connect and highlight these traditions throughout this dissertation, but due to time and space I cannot encompass all intricacies and details of each discipline and so only the most relevant information to this thesis is provided. First, I review the importance of context in assessing knowledge and point out how many approaches generally ignore context and contextual differences. I then zoom out from biomedicine and review other approaches to understanding knowledge and discuss measuring knowledge from publications and citations. I

detail the strengths and weaknesses of using publications and citations as proxies of knowledge, but argue that both are problematic and by themselves are poor indicators of knowledge changes. I move to concepts and conceptual change, and provide a summary of how concepts and conceptual change have been used historically to show how knowledge changes. I display how difficult it is to link changes in knowledge to conceptual change by emphasizing the complexity of language in regards to conceptual variation and context, specifically how one concept can have multiple interpretations, multiple concepts can have the same interpretation, and how interpretation is influenced by context. Circling back to biomedicine, I argue approaching conceptual change from a complexity science perspective provides insight into how knowledge change is dependent on context and helps motivate a study across multiple scales and dimensions. Finally, I introduce the theoretical framework which I drew upon for analyzing the history of knowledge which integrates contextual factors as explanatory variables developed by Jurgen Renn and Manfred Laubichler (Renn & Laubichler, 2017a).

In the third chapter, I argue for the necessity of incorporating big data and data driven research and science (DDSR) approaches to study the microbiome concept. To argue this claim, I engage the historical context of big data and substantiate how the current phenomena of big data is different than any previous moment in history. I stress how big data provides more contextually specific information across multiple scales and dimensions resulting in novel research methods, questions, and results such as: the decoding of the human genome and evidence of the Higgs Boson have been attributed to the use of big data. I, also, detail the complexity inherent to big data, specifically focusing on the “five V’s of big data,” these being the volume, velocity, veracity, variety, and validity of big data. I argue that the inherent complexity of big data is in fact a benefit, and provides the opportunity for researchers and scientists to design

experiments and studies across new dimensions and across larger scales than ever before. While the potential and benefits for big data are well known, there is still disagreement and confusion on how to handle and deal with big data. I argue here, that DDSR approaches through systematic and repetitive experimentation and analysis enhance the capture, curation, and analysis of big data. I, also, present the criticisms of big data and my counter arguments towards those criticisms.

In the fourth chapter, I show how the microbiome is a separate concept from the metagenome and metabolome. To prove this claim, I provide the results from a series of experiments on a corpus collected on the microbiome (MB Corpus) and compare this with corpora on concepts that are cited as being confused with the microbiome concept, specifically the concepts of metagenome and the metabolome. I apply a hybrid of qualitative and quantitative methods including topic models, frequency analyses, keywords in context, collocational analyses, lexical profiles, and collocate networks to analyze both the language context and the specific usage of the each word. My results characterize differences between the – (1) microbiome and the metabolome, and (2) the microbiome and the metagenome, specific to the larger language environment, occurrence and co-occurrence of words, and knowledge of words used. However, changes in the parameters of the analysis yielded results that suggest the difference between the concepts is not a discrete boundary.

In the fifth chapter, I explore the development of the microbiome according to different contexts by systematically comparing the variation and change of the language and social context of the microbiome. Using a combination of experiments, I analyzed the variation and change in social and language context and specific contextual factors such as: publications, citations, authors, MeSH terms, and unique MeSH terms over time in the MB Corpus. A factor

introducing variation and change in the development of the microbiome over time was attributed to the microbiome being an innovation (Rogers, 2010). Specifically, the microbiome was shown to have: a relative advantage over similar the concepts metagenome and metabolome, microbiome knowledge was highly compatible with the knowledge bases of the social systems in which it was used, the complexity of the microbiome allowed the microbiome to be separated into different conceptualizations, the range of different meanings attributed to the microbiome were evidence of the trialability of the microbiome across contexts, multiple events increased the observability of the microbiome, and the adoption rate of the microbiome across diverse systems showed the characteristic S-curve of adopted innovations, which are attributes that characterize an innovation being adopted (Kapoor, Dwivedi, & Williams, 2014). Evidence of the variation and changes in knowledge of the microbiome were described using the results from a topic model, MeSH analysis, and collocational analysis. PLOS One was unusually prolific in producing microbiome publications from 2008 to 2014 compared to all other journals based on an analysis of producers of microbiome research outputs with a focus on high producers of microbiome publications and predicted gatekeepers of knowledge (Crane, 1967; McGinty, 1999; Siler, Lee, & Bero, 2015). PLOS One was compared to other journals (Nature and Science) and the National Institutes of Health project abstracts using MeSH terms, collocates, and topic models to show similarities and differences in knowledge from cited gatekeepers of knowledge based on previous research (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011; Moore, Neylon, Eve, O'Donnell, & Pattinson, 2017). The results of the MeSH terms, topic model, and collocate comparisons of high producers to gatekeepers showed differences in the similarities of knowledge and convergence of knowledge suggesting changes to the microbiome were influenced by specific contextual factors.

This dissertation advocates that through the integration of big data and data driven science and research approaches high-dimensional data that cuts across multiple contexts are more accessible than ever before. The analysis of the dynamics of both language and social contexts and their interactions to identify contextual factors of the microbiome, provides understanding into how historical transformations in science relate to changes in knowledge, and expand the repertoire in generating hypotheses for understanding the evolution of knowledge. The strategy for understanding the evolution of scientific knowledge of the microbiome concept, with an emphasis on integrating big data for measurable changes, yields insight into future experiments and projects integrating big data and hybrid qualitative and quantitative approaches.

This dissertation supports and provides new evidence for recognizing the importance of social, language, and historical context on biomedical concepts and scientific knowledge. Big data and the novel approaches used in this dissertation makes it possible to achieve a description of how specific words, phrases, and discourses correlate to specific factors from different contexts. The results from this study can then be used to study other the trajectories of other concepts, analyze language and social contexts, and provides suggestions on what influences knowledge. Additionally, by pursuing a systematic approach that incorporates complex relational data, this study provides novel methods for research in the anticipation of the utility of big data. Big data is an important consideration for this study and future studies, as the current situation within biomedicine and science in general is that the amount of data available for analysis has far exceeded the capacity to analyze it. Therefore, it is imperative to understand how data driven science and research approaches can be utilized for real world problems and solutions within science.

Research Hypothesis and Objectives

My dissertation tests the hypothesis that knowledge of the microbiome has changed and evidence of this change is found in the changes to the usage and meaning of the microbiome or language context, and changes to the social structure and material artifacts or the social context of who uses the microbiome and where. My strategy is to identify and analyze specific contextual factors influencing specific changes to the microbiome. My project focuses on isolating, tracing, and measuring the link between contextual factors and changes to the microbiome through analysis of the microbiome concept across different scales and dimensions.

In order to test this hypothesis the following must be described:

- 1) Language context (words, phrases, sentences) characterizing the variation and changes of the microbiome in text,
- 2) Social context (social groups, categories, environment) characterizing the social variation and changes to the actors that have used the microbiome, and
- 3) Historical context or characterizing the changes in the language and social context according to time.

The primary research question for this dissertation is: *What are factors influencing the microbiome concept?* One goal of this project is detail how contextual factors, linguistic factors (frequency of use, associations with other linguistic variables, and relationships within and between these variables) and social factors (characteristics of authors, publications, journals, and the associations within and between these variables) influenced the variation and changes to the usage and meaning of the microbiome. Another goal is to create a meaning of the microbiome based on language in use, or meta-meaning, and compare the meta-meaning created with other definitions of the microbiome from other ontologies. A third goal, is to characterize the influence of these factors by assessing the similarity and convergence of microbiome knowledge between groups and across groups.

Research into contextual factors found that usage of terms in the scientific literature was a driver that impacted companies' adoption for eco-innovation policies (Bossle, 2016). *In Representing Scientific Knowledge: The Role of Uncertainty*, Chaomei Chen and Min Song found that the emergence of a specialty is determined by the two key drivers, 1) knowledge measured via terms in article titles, abstracts and 2) the research fronts associated with the knowledge (i.e. term in citations) (Chen & Song, 2017). However, this study was based on the assumption that terms or concepts across networks have the same meaning across time. Contextual factors are also dependent on the unit of analysis, such as an article, citation, author, and cannot be described as having universal properties across different networks or levels of analyses (Meyer & Goes, 1988; Loet Leydesdorff, 2013). Aaron Cicourel has pointed out, "at any given time knowledge depends on the particular state of methods in use," alluding to how results are dependent on which metrics are used when conducting multilevel analyses across different dimensions or scales (Cicourel, 1964, p. 164). Therefore, this project by identifying and evaluating contextual factors on the microbiome by integrating big data may provide vision into how to create experiments integrating multi-dimensional data across time scales.

By answering the research question insight will be gained on how scientific knowledge on the microbiome concept was influenced by social factors, language factors, or some combination of factors. Moreover, by characterizing the factors that influenced the microbiome these results can be used to assess the dynamics of other processes within biomedicine and science. Specifically, how knowledge evolves and how social and linguistic variables impact knowledge processes.

Methods

By answering the research question, *what are factors influencing the microbiome concept?* Other supplementary questions related to how knowledge on the microbiome, must also be considered. These supplementary questions range from: What is the microbiome concept? How is the microbiome dependent on context? Who has influenced the microbiome, and what is evidence of their influence? Where was microbiome research focused and how has that changed over time? Is the microbiome an innovation or rebranded concept? How did knowledge on the microbiome evolve? To answer this range of questions, an interdisciplinary approach was used to identify contextual factors including: language usage patterns, social groups, and variation and changes to the microbiome knowledge. Every publicly available scientific publication and funded project abstract with the word microbiome in the text and the metadata for each individual article or abstract was collected. The metadata collected included explanatory variables from the language context of the microbiome, or language factors including: the words used with the microbiome in text, the phrases used with the microbiome, categorizations of texts via medical subject headings (MeSH terms), and the specific discourse a phrase is a part of. The metadata, also, included explanatory variables from the social context of the microbiome or social factors: author and co-authors, journal name, institutions of the authors, country of publication, and impact of the publication. Additionally, the metadata anchored the linguistic and social factors via publication or fiscal year, and helps facilitate a constant dialogue with the historical context of the microbiome, as many of the ancillary questions posed as a part of this study are examined via specific slices of time. By identifying, analyzing, and bringing together different factors, a data synthesis from the publications, abstracts, and metadata helped

trace and map the influential words, phrases, events, individuals, groups, institutions, and networks influencing the usage and meaning of the microbiome over time.

The methods for this dissertation are organized by background chapters and experiments. However, due to the sheer amount of background information on the different approaches, only brief discussions of the experimental research questions and methods are presented here. George Lundberg warns of how brilliant scientists can make major mistakes when stepping outside of their specialties, while Pierre Bourdieu argues science advances by crossing disciplinary boundaries (Lundberg et al., 2012; Bourdieu & Wacquant, 1992). Agreeing with the Bourdieu's opinion, analyzing the historical, language, and social context of the microbiome requires an interdisciplinary collection of methods from: linguistics, sociology, digital humanities, computer science, and history of science. Each of these disciplines has rich theoretical understandings that accompany the methods. Therefore, preceding the experiments a thorough collection and review on how different approaches and an in depth literature review was conducted. The literature review investigated approaches using language, social, and historical contexts to characterize knowledge in science and biomedicine, and how knowledge on the microbiome has changed. This project is designed to understand and measure how knowledge on the microbiome evolved. To accomplish this task, five chapters with two experiments are included.

Data Collection and Curation

Prior to any experiment, the first task was to create a collection of full text documents or corpus. Publicly accessible articles were identified by searching for the microbiome as a topic or in the text from JSTOR, Web of Science, and PubMed. From these multiple databases. 32,632 articles on the microbiome and their corresponding metadata were identified. Articles were removed from the final microbiome corpus (MB Corpus) if the publication was a duplicate, was

not open access, or if the publication could not be converted to a text. This resulted in 27,977 articles on the microbiome from 1900 to 2017, a detailed description of this process is presented in the fourth chapter. After the MB Corpus was finalized the metadata was cleaned and curated for duplicates, errors, and missing values. This process took multiple iterations and included a combination of computational and hand disambiguation for the final cleaned metadata. Each full text article in the MB Corpus has a metadata line including the following information, PubMed ID (if applicable), Publication Title, Journal Title, National Library of Medicine MeSH Term(s), Author(s) Last Name and Author(s) First Name, Author Affiliation(s), and the Abstract.

Following the creation of the MB Corpus, three additional corpora were collected using a similar method based on a single concept. In total there five corpora with their accompanying metadata:

1. MB Corpus (27,977 articles): collection of publications with microbiome as the search term
2. MB Project Corpus (10,401 projects): collection of funded microbiome projects with microbiome as the search term
3. Metabolome Corpus (16,818): collection of publications with metagenome as the search term
4. Metagenome Corpus (10,741): collection of publications with metabolome as the search

These corpora were collected and used for preliminary studies prior to the final experiments.

CONTEXTUAL REVIEW OF THE MICROBIOME LITERATURE

Background

Science and scientific knowledge is influenced by the context in which it occurs (Kuhn, 1970; Popper, 2002; Cetina & Reichmann, 2015). Previous studies measuring scientific knowledge have used different methods to characterize and analyze scientific discovery. The sociology of knowledge, for example, has emphasized using knowledge itself as the unit of analysis and explores the stability/instability of knowledge dynamics through discourses, practices, communication, and details how these social processes limit or direct knowledge (Durkheim & Mauss, 1963; Bourdieu, 1977; Mannheim, 1995; Scheler, 2012). Sociology of science focuses on opposing knowledge claims and the scientific production of knowledge, examining social groups, institutional environments, laboratories, methods, and the objects of scientific inquiry as the units of analysis (Merton, 1970; Bloor, 1991; Bijker, 1993). History of science, has analyzed the “social factors” or forms of organization, social influences from politics or economics, social practices and costs of science or “intellectual factors,” and the ideas, concepts, methods, evidence of science, and scientific knowledge (Kuhn, 1970; Shapin, 1996). Linguistics, meanwhile, studies the history of words, individual and cultural knowledge, shared knowledge, and knowledge of language (Wardhaugh, 2009; Burkette & Kretzschmar, 2018). My approach meets at the intersection of these disciplines by exploring the historical, language, and social context of the microbiome.

I argue historical, language, and social context is critical in understanding history, variation, and changes to knowledge. Historical context provides information on the conditions in which something takes place and explains the development of social and intellectual factors.

Understanding historical context provides insight into who, what, where, when, and why of sources and evidence and provides understanding beyond an act or thing itself. To illustrate, an analysis of the historical context of epidemiological studies highlighted the heterogeneity and dissimilarity in epidemiologic data on mental disorders, brought to light issues with how the mentally ill were being underserved, and delivered the basis for the creation of uniform basic data on mental disorders (Regier et al., 1984). Language context is the words, concepts, phrases, definition(s), patterns, and meaning of words and has been used as both a social factor and intellectual factor influencing knowledge. William Labov is his groundbreaking work on the language context of New York City (NYC), was able to show that New York City is a single speech community, social groups within NYC were the source of language variation and change, and language context is an identifier that differentiates individuals (Labov, 2006). Social context refers to social actors, individuals, groups, institutions, setting, and the environment and has also been used as both a social and intellectual factor influencing knowledge. Studies emphasizing social context include concepts and information on the relationships among the units of analysis, focusing on individual behavior within relationships, the behavior of social groups, formal social categories or informal social connections, or social patterns that create social structures (Wasserman & Faust, 1994).

Past research on the microbiome concept has been interpreted according to specific intellectual or social factors along relatively small time scales, with many works focusing on possible origins of the microbiome and then abruptly jumping to the present, debate on who coined the microbiome, declarations of how the microbiome is being used now, or directions on how the microbiome should be used (J Lederberg & McCray, 2001; Prescott, 2017; Eisen, 2015;

Ursell, Metcalf, Parfrey, & Knight, 2012). Absent from this body of work is a comprehensive study of the context and contextual factors contributing to the development and evolution of the microbiome. I argue that in order to understand the variation and changes to the microbiome concept, a systematic study of the contextual factors influencing the microbiome is necessary. I argue context is crucial in understanding sources of variation and change. Previous research on the gene has shown that the gene was interpreted differently based on context, with some experts arguing how contextual differences of the gene concept are so different leading to a “classical gene concept” and a “contemporary molecular gene concept” (Burian, 2004; Falk, 2004). Context, also, helps to understand the influences that are directing changes, as William Labov was able to show social factors including age, sex, socioeconomic status, and community were responsible to specific variants in the English language to be adopted, diffused, and transmitted (Labov, 2001).

I use context to help understand the variation and changes to the microbiome over time. This requires a comprehensive sense of context across multiple scales and dimensions, analyzing both short and long time scales (Maasen & Weingart, 2013). In this chapter, I focus on the different interpretations of the microbiome from historical, language, and social contexts. Within the historical context, I focus on the events, moments, trends, and people or things at specific times that have been cited as the origin of the microbiome. I highlight the historical context of the microbiome links scientific luminaries such as Antonie van Leeuwenhoek, Charles Darwin, Ivan Wallin, HJ Muller, and Theodosius Dobzhansky to the origin of the microbiome (Bordenstein, n.d.; Ursell et al., 2012). Yet, the most accepted and cited origin of the microbiome belongs to the Nobel Laureate Joshua Lederberg (L. V. Hooper & Gordon, 2001). Despite the

overwhelming support to Lederberg's some contest this origin of the microbiome. Both Susan L. Prescott and Johnathon Eisen claim the current usage and interpretation of the microbiome originates from an article that predates Lederberg's claim and describe the additional evidence of other instances of the microbiome occurring in text prior to Lederberg's claim (Eisen, 2015; Prescott, 2017). I agree that Lederberg's claim to the microbiome is problematic given there are numerous other instances of the microbiome occurring in text prior to Lederberg's 2001 commentary. However, Prescott and Eisen both use a quote as proof the interpretation of the microbiome has not changed over time. I consider how Eisen and Prescott fail to provide substantial historical, linguistic, or social evidence to support their claim of how the microbiome is currently being used or what the definition of the microbiome is/was in the literature.

Adopting a language context perspective, I highlight the words, concepts, phrases, definition(s), and meaning that have been used with the microbiome concept. I argue most researchers and scientists attempting to define the microbiome create their own definition by ignoring previous interpretations. These multiple definitions for the microbiome concept, have led to confusion and misuse of the terms microbiome, metabolome, metagenome, and microbiota and an active debate has emerged on if the microbiome is a separate concept from metabolome, metagenome, and microbiota. I also advocate multiple interpretations of the microbiome have led to indecision on if there is a single core human microbiome or separate microbiomes such as the lung microbiome, skin microbiome, intestinal microbiome, and describe how some interpretations have conceptualized the microbiome to manifest beyond humans, like Julian Marchesi and Jacques Ravel, whose microbiome interpretation describes "entire habitats" and includes viruses, the genes of all microorganisms in the habitat, and environmental conditions

(Marchesi & Ravel, 2015a). I argue while most of these attempts to define the microbiome are motivated to create a single microbiome concept, the end result is more confusion. Further, I show that many studies on the microbiome have acknowledged the importance of the language context or vocabulary and lexicon of the microbiome to understand the usage and meaning of the microbiome, but studies have failed to provide evidence confirming if there is or is not a standardized microbiome vocabulary or common microbiome language.

Focusing on the social context, I maintain the social actors, or individuals, groups, institutions, setting, and environment are critical to understanding the microbiome. I argue that the microbiome has different uses across social contexts and drawing examples from biomedicine and ecology. I emphasize within biomedicine the microbiome is used to describe a single human microbiome or multiple separate microbiomes on a person, such as a lung, hand, or gut microbiome, while ecology uses the microbiome as another instance of the concept biome or microbiota. This usage within ecology removes most of the debate surrounding the microbiome as a new concept or rebranded concept. However, the comparison between ecology and biomedicine illustrates how social contexts can have different interpretations for the same concept. I also show the interpretative differences based on social context on the microbiome may not have distinct boundaries, as Gregory W. Schneider and Russell Winslow argue that both ecology and epidemiology contributed to the microbiome (Schneider & Winslow, 2014). I agree John Huss's argument several factors contributed to form the microbiome and social context can be used to inform the interpretation of the language context, but I maintain no studies have provided evidence of this connection, or on the variation in use or meaning of the microbiome across social contexts.

In summary, the purpose of this chapter is to provide the historical, language, and social contexts important in understanding the variation and range of interpretations to the microbiome and highlight changes to scientific knowledge specific to the microbiome.

Historical Context

The historical context of the microbiome describes the events, moments, trends, people, or things at specific moments related to the historical development of the microbiome. The biomedical narrative of the microbiome situates the historical context as far back as Antonie van Leeuwenhoek 1632–1723 and Leewenhoek’s discovery of different microorganisms, “Studies of the human microbiome started with Antonie van Leuwenhoek who, as early as the 1680s, had compared is oral and fecal microbiota,” (Ursell et al., 2012). Other biological and scientific luminaries ascribed to the biomedical narrative include Charles Darwin, Ivan Wallin, HJ Muller, and Theodosius Dobzhansky as part of the historical origins of the microbiome (Bordenstein, n.d.). Yet, most of these claims lack direct evidence, textual or otherwise, linking the individual to the microbiome. Thus, while Van Leeuwenhoek is considered “the father of microbiology” and was the first to observe and describe unicellular and multicellular organisms which he referred to as “animalcules”, Leeuwenhoek never explicitly used the word microbiome in text (Gest, 2004; Scher & Abramson, 2011). Similarly, though Darwin, Wallin, Muller, and Dobzhansky made major contributions to biology and microbiology, there is no direct evidence linking the microbiome to their work.

As of 2018, the most popular description of the microbiome supported by textual evidence begins in April 2001, when the Nobel laureate Joshua Lederberg claimed he coined the

microbiome in “Omics Sweet 'Omics - A Genealogical Treasury of Words,” a commentary he co-authored with Alexa McCray:

In physics, probably starting with Faraday's ion, cation, anion, the *-on* suffix has tended to signify an elementary particle, later materially focused on the photon, electron, proton, meson, etc., whereas *-ome* in biology has the opposite intellectual function, of directing attention to a holistic abstraction, an eventual goal, of which only a few parts may be initially at hand. The accompanying table illustrates a number of prevailing examples. It includes Lederberg's own recent coinage of *microbiome*, to signify the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease. (J Lederberg & McCray, 2001)

A month later in May of 2001, Lora Hooper and Jeffrey Gordon provide support for Lederberg's claim by citing a personal communication between Lederberg and the co-authors on the microbiome:

The Nobel laureate Joshua Lederberg has suggested using the term “microbiome” to describe the collective genome of our indigenous microbes (microflora), the idea being that a comprehensive genetic view of *Homo sapiens* as a life-form should include the genes in our microbiome. (L. V. Hooper & Gordon, 2001)

Lederberg would continue to advocate for his claim to the microbiome through a fixed combination of words, repeating the phrase, “*I call the microbiome*,” as seen in the first paragraph of the 2004 article *Of Men and Microbes*, Lederberg states, “Understanding this cohabitation of genomes within the human body — what *I call the microbiome* — is central to understanding the dynamics of health and disease,” and again in the same piece:

It would thus broaden our philosophical horizons if we think of a human—a body space in any human—as more than an organism. It is a superorganism with an extended genome that includes not only its own cells but also the fluctuating microbial genome set of bacteria and viruses that shares that body space. Some of these one-time invaders have become permanently established in our cells, even crossed the boundary line and entered our own genome. *I call* that extended set of companions *the microbiome*, and pray for more research on how they impact our lives, besides the flare-ups, the blunders, we call disease. (Joshua Lederberg, 2004)

Considering this evidence, Lederberg's claim to the microbiome is situated in the historical context of the year 2001 from the commentary he co-authored and his personal communication with Hooper in Gordon. Biomedicine scientists have generally supported Lederberg's claim citing Lederberg and McCray, "The concept of the human microbiome was first suggested by Joshua Lederberg, who coined the term "microbiome...", this quote comes from the paper describing the National Institutes of Health(NIH) Human Microbiome Project (HMP₁) Working Group, a collection of over 200 scientists at 80 institutions (NIH HMP Working Group et al., 2009; Schneider & Winslow, 2014). Combined, the HMP₁paper, Lederberg's personal communication, the 'Omics paper, and Hooper and Gordon's paper, have been cited thousands of times strengthening Lederberg's claim to the microbiome. However, a growing number of publications and scientists are arguing against Lederberg's claim that he was responsible for discovering, coining, and popularizing the microbiome.

One of the most vocal critics of Lederberg's claim is Susan L. Prescott, in her article the *History of Medicine: Origin of the Term Microbiome and Why It Matters*, Prescott attacks Lederberg's claim and those that have fallen victim to the "microbiome zeitgeist":

Such is the case within the microbiome zeitgeist. For example, it is continuously claimed that the term microbiome was 'coined' by Nobel laureate-microbiologist Joshua Lederberg in a 2001. This statement of coinage is presented as fact in 100's of recent research papers, including those in journals devoted to pediatrics (Tracy, Cogen, & Hoffman 2015) allergy/immunology (Bonamichi-Santos et al. 2015) and even in papers originating from the United States National Institutes of Health, Human Microbiome Project (Proctor 2016; NIH Human Microbiome Working Group et al. 2009). In an article purporting to set the record straight on microbiome terminology, it is even claimed that the term *microbiota* was defined for the first time by Lederberg in 2001 (Marchesi and Ravel 2015). Remarkably, a 'History of Medicine' article in a recent *Annals of Internal Medicine* issue makes this same claim that Lederberg *coined* the term in 2001 (Podolsky 2017)...Despite these claims, the evidence is crystal clear – Lederberg did not coin the term microbiome. (Prescott, 2017)

Prescott draws attention to historical uses of the word ‘microbiome’ in text predating Lederberg’s claim and cites a text written 13 years before Lederberg’s commentary, arguing the usage of microbiome in this text is more aligned with how the microbiome is currently used in microbiology:

A convenient ecological framework in which to examine biocontrol systems is that of the microbiome. This may be defined as a characteristic microbial community occupying a reasonably well defined habitat which has distinct physio-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatre of activity. (Whipps, Lewis, & Cooke, 1988)

Prescott’s argument also touches on falsehoods in science or what she calls ‘scientific myth’, consequences to scientific practice and science policy, and the ethical struggle within science that legitimate sources are given their due credit. She argues Lederberg’s claim to the microbiome is not simply a case of citation errors but is a case of unfounded attribution (Prescott, 2017). Though Prescott’s argument is a valid concern and highlights the importance of historical accuracy, ethics, and factual referencing, she lacks supporting evidence to substantiate her claim of how the microbiome is currently used in the literature. Further, she fails to provide multiple instances of where the microbiome has been used prior to Lederberg’s claim.

Jonathon Eisen in his *What Does the Microbiome Mean? And Where Did It Come From? A Bit of a Surprise...*, also uses the 1988 citation and usage of the microbiome as evidence against Lederberg’s claim, and makes two of the same counterclaims Prescott makes: 1) Whipps, Lewis, and Cook’s usage of the microbiome denotes how the microbiome is being used in the community right now, and 2) the interpretation of the microbiome as a community/communities of organisms aligns with the historical usage of the microbiome (Eisen, 2015). Eisen’s argument emphasizes the earliest occurrences of microbiome but fails to provide other supporting

historical, linguistic, or social evidence to support his claims of how the microbiome is being used now, or what the first definition of the microbiome is in the literature. Further, if time was the only dimension used to settle the debate, the first use of the word microbiome in print occurred in a book reviewing dentistry (odontological) techniques in 1949, “On sait, d’autre part, le rôle que joue le PH dans l’évolution du microbiome intestinal. Il est plausible d’admettre qu’à la faveur des lésions intestinales et des modifications du PH intestinal, les migrations parasitaires ou microbiennes peuvent” (*Revue odontologique*, 1949). Translating from French to English, this quote reads as, “We know, on the other hand, the role of the PH in the evolution of the gut microbiome. It is plausible to assume that the favor of the intestinal lesions and intestinal pH changes, parasitic or microbial migration” (Eisen, 2015). Yet, this text doesn’t provide any definition of the microbiome, implying that the reader either understood the word or would infer the definition from textual clues.

The first scientific article to use the microbiome was authored by John Mohr in 1952, titled *Protozoa as indicators of pollution*, Mohr writes, “The protozoan fauna (as a matter of fact, the whole microbiome) is poor in species and individuals, and those present are rather typical polysaprobies (Mohr, 1952). Similar to the text from 1949, Mohr does not provide any definition for the microbiome and the above sentence is the only occurrence of the microbiome in the text. Both of these texts only mention the microbiome once, but are rarely cited or even acknowledged as part of the microbiome’s history or development. Some researchers have erroneously pointed to Léonce Priouveau’s work in 1894, in the Archives of toxicology and of gynecology as the first instance of microbiome, but this is actually a reference to *microbisme*, a commonly used term in the late 1800’s and early 1900’s (Eisen, 2015). These occurrences of the microbiome outside

biomedicine and the misinterpretations of the microbiome for other concepts imply the microbiome is multidimensional concept with interpretations dependent on context.

Thus, Prescott and Eisen argument on the Lederberg not being the first to use the microbiome in text is accurate, as the microbiome was used prior to 2001 with the earliest instance being 1949, and the earliest use in a scientific article in 1952. Prescott and Eisen's counterclaim, Whipps, Lewis, and Cook's use of microbiome is the first and most commonly used interpretation is not substantiated by enough evidence (Eisen, 2015; Prescott, 2017). Further, this claim cannot be proven using historical context (in text occurrence) alone but requires an understanding of the language and social context of the microbiome over time. Put another way, to track how the microbiome developed, the language context (usage and meaning) of the microbiome needs to be understood at different points in time, with respect to the social context (location and social environment).

Language Context

Other researchers have emphasized language context or words, concepts, phrases, definition(s), patterns, and meaning as the most critical aspects to understanding the microbiome. Lots of attention and interest has gone into the definition and meaning of the microbiome, with the majority of scientists and experts providing their own interpretation to the microbiome. In *Defining the Human Microbiome*, Luke K Ursell and co-authors define the microbiome as “the catalog of the microbial taxa associated with humans and their genes” and argue previous misunderstandings with defining the microbiome are due to contextual factors related to language use, i.e. confusion in terminology between microbiome, microbiota, and metagenomics (Ursell et al., 2012). Yet, the authors later in the same work then proceed to contradict

themselves and explain the microbiome definition is still in flux and dependent on other concepts:

...new findings are leading us to question the concepts that are central to establishing the definition of the human microbiome, such as the stability of an individual's microbiome, the definition of the OTUs (Operational Taxonomic Units) that make up the microbiota, and whether a person has one microbiome or many. (Ursell et al., 2012)

This passage highlights another confusing aspect related to the language context and interpretation of the microbiome, specifically in some texts the microbiome and other body regions are referred to as one collective microbiome or a core microbiome, while in other texts there are multiple separate microbiomes for different parts of the body, e.g. the intestinal microbiome, vaginal microbiome, and lung microbiome (Turnbaugh et al., 2009; Preidis & Versalovic, 2009; Lamont et al., 2011; Erb-Downward et al., 2011).

Agreeing with Ursell, Julian Marchesi and Jacques Ravel in *The Vocabulary of Microbiome Research: A Proposal*, point to the misuse and misinterpretation of the microbiome with other concepts (Marchesi & Ravel, 2015a). Interestingly, Marchesi and Ravel mention and cite Lederberg and McCray's commentary as the first definition of microbiota, not the microbiome, and agree with Ursell on the point of microbiota and metagenomics being confused with microbiome, but then add the term metabolome as another term being confused with microbiome:

This rapid evolution of the field (microbiome) has been accompanied by confusion in the vocabulary used to describe different aspects of these communities and their environments. The misuse of terms such as microbiome, microbiota, metabolomic, and metagenome and metagenomics among others has contributed to misunderstanding of many study results by the scientific community and the general public alike. (Marchesi & Ravel, 2015a)

Marchesi and Ravel, also, provide another broad definition of the microbiome which extends the parameters of the microbiome past humans and their genes to now include, “the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e. genes) and the surrounding environmental conditions,” (Marchesi & Ravel, 2015a). Marchesi and Ravel address other contextual issues of language when stating their goal in providing a definition of the microbiome is to create a “common language” and “hope that a consensus use of these terms could be adopted in the near future. This editorial aims at stimulating a discussion and standardizing the vocabulary of microbiome research” (Marchesi & Ravel, 2015a). However, by adding the ambiguous terms “habitat” and “surrounding environmental conditions,” Marchesi and Ravel’s new definition is more complex and requires interpretations of habitat, surrounding, environment, and surrounding, in order to understand where a microbiome begins/ends compared to a biome, macrobiome, or an ecological web. Thus, like Ursell, Marchesi and Ravel’s microbiome definition creates more misunderstanding on the microbiome concept. Moreover, their goal in creating a common language and standardized vocabulary on the microbiome assumes there is no pre-existing common language or standardized vocabulary on the microbiome and the parameters of the microbiome are understood.

However, two years before Marchesi and Ravel, Martin Blaser argued that the *microbiome* has a “general vocabulary” and “operating lexicon”, in *The Microbiome Explored: Recent Insights and Future Challenges*, Blaser states, “For a new research field such as the study of the human microbiome, much of the conceptual and technical infrastructure had to be built from scratch. Both the HMP and MetaHit contributed greatly to these infrastructures and have helped establish a general vocabulary, operating lexicon, and tool-kit for further studies” (Blaser,

Bork, Fraser, Knight, & Wang, 2013). Marchesi and Ravel overlook Blaser's claim when pushing for a common language and standardized vocabulary, along with most of the historical context of the microbiome. Yet, no study has come forward with evidence to confirm/deny Blaser's claim of a general vocabulary. Further, Blaser while emphasizing the language context points to the importance of social context in understanding the usage of the microbiome by describing how the HMP and Metahit "helped establish" a general vocabulary and operating lexicon of the microbiome, but here again there has been no studies that have investigated a link between HMP, Metahit, or any other social groups influence on the usage and meaning of the microbiome or the microbiome lexicon.

It may be the case that the microbiome is undefinable or there is no consensus knowledge of the microbiome, as The American Academy of Microbiology, concedes in the *Human Microbiome FAQ*, "there is not yet a fully agreed upon definition" of the microbiome (American Society for Microbiology, 2013). Even Dr. Lita Proctor, the program director for the HMP₁, said after six years at the end of the HMP "...scientists are struggling to figure out how to think about the microbiome" (Mole, 2013). Still, other studies integrating social context have provided insight into the variation and changes of concepts and knowledge over time (Labov, 2001; Lawrence Edwards et al., 2016).

Social Context

The social context of the microbiome is the social actors, individuals, groups, institutions, setting, and environment, or who used the microbiome and where the microbiome was used. In general, the microbiome has been used by a broad range of individual scientists, researchers, clinicians, politicians, and the general public. Socially these individuals belong to different groups and institutions, such as biomedicine or ecology, and these groups have characteristically

different uses of the microbiome. Within the scientific community the microbiome was named as one of “the breakthroughs of the year” in 2011 and 2013, and featured in two special issues in the journal *Science* (*Science*, 2011, 2013; “*Science*,” 2012; “*Science*,” 2016). More recently, a number of journals dedicated specifically to microbiome research have popped up: *Biofilms and Microbiomes* (*Nature*), *Microbiome Journal* (biomed Central), *Microbiome Science and Medicine*, and the *Human Microbiome Journal*. Microbiome centers are now common in medical schools, and biology, engineering, ecology, and veterinary medicine departments. Additionally, the microbiome has piqued the interest of industry, as an article in the *Wall Street Journal* reported from 2011 to 2015 there was \$75.3 billion invested in microbiome based research firms and,” venture-capital interest in the microbiome- the collection of microbes that inhabit humans, animals and plants- is growing like a culture in a petri dish. ” (Gormley, 2016). Further, microbiome analysis as a service is now available via companies such as uBiome and VIOME, who offer at home clinical microbiome test or a Microbiome Counselor™ that provide “microbiome-based precision medicine” and “technology that’ll analyze your microbiome to find the best foods for your body,” and advertised on popular media outlets ranging from Forbes, Marie Claire, and cable television (“Home | uBiome,” 2018; “VIOME- ga2,” 2018). From these observations, it is no wonder the microbiome has been characterized as one of the most important biomedical innovations which touches on practically all branches of science (Prescott, 2017).

Differences in microbiome knowledge has manifested into diverse interpretations of the microbiome across scientific disciplines. Biomedicine emphasizes the discovery, diffusion, and transmission of microbiome knowledge, and generally uses the microbiome to describe either a human microbiome or microbiome specific to a region of the human body. However, there are

still disagreements within biomedicine on if there is a single human microbiome, multiple microbiomes (lung, hand, gut) per person, or a universal core microbiome (Shade & Handelsman, 2012). Within biomedicine, also, many are unsure if the microbiome is a separate concept from metagenome, metabolome, microbiota (Ursell et al., 2012; Marchesi & Ravel, 2015a).

In ecology, there is little debate on the discovery and diffusion of microbiome knowledge or concern with who deserves the credit for coining the term. Ecology uses the microbiome as a synonym to biome or microbiota and in most instances the microbiome is combined with other concepts within ecology, e.g. a root microbiome, a coral core microbiome, a soil microbiome, lake microbiome, an earth microbiome (D Ainsworth et al., 2015; Lundberg et al., 2012; Shade & Handelsman, 2012). Ed Yong, , *I Contain Multitudes: The Microbes Within Us and a Grander View of Life*, goes so far to claim the microbiome and “all zoology” is based on ecology, “All zoology is really ecology. We cannot fully understand the lives of animals without understanding our microbes and our symbioses with them. And we cannot fully appreciate our own microbiome without appreciating how those of our fellow species enrich and influence their lives” (Yong, 2016). Yong’s book displays the public interest in the microbiome and an emphasis on an ecological narrative for the microbiome, as his book has been featured across multiple mediums as an authoritative source on the microbiome: from the New York Times book review “(According to Ed Yong) In a way, the science of the microbiome had two false starts...” , National Public Radio’s book summary “ (I Contain Multitudes) Shares unique perspectives into the role of the human microbiome in human health, identity, and ability, explaining in comprehensive, lighthearted detail how microbes shape and protect life on Earth in unexpected ways, to Bill Gates blog “And he (Yong) offers realistic optimism that our growing knowledge

of the human microbiome will lead to great new opportunities for enhancing our health,” (Weiner, 2018; Gross, n.d.; Gates, 2017) . How this book was branded as an authoritative source on the microbiome is surprising, as Yong focuses more on the practice of microbiology and the study of microorganisms rather than the microbiome, his sweeping narrative includes some relatively obscure historical figures from the last 500 years of microbiology, such as: Leeuwenhoek, Hooke, Louis Pasteur, Robert Koch, Charles Darwin, Martinus Beijerinck, Thomas Huxley, Theodor Escherich, Elie Metchnikoff, Theodor Rosebury, René Dubos, Carl Woese, Ralph Woese, Norman Pace, Ed DeLong, David Relman, Jeff Gordon, and Rob Knight. Yong’s perspective on the typifies the ecological perspective, as he uses the microbiome as another instance of a different concept, in this case microbiota, “All of us have an abundant microscopic menagerie, collectively known as the microbiota or microbiome” (Yong, 2016).

Nevertheless, the ecological conceptualization of microbiome adds to the overall confusion of how to interpret the microbiome by suggesting the microbiome is synonymous with microbiota. As previously mentioned, biomedicine has an ongoing debate on if the microbiome is a separate concept from other concepts, including microbiota, metagenome, and metabolome (Ursell et al., 2012; Huss, 2014; Marchesi & Ravel, 2015a). It is critical to separate the microbiome and the microbiota as conceptual ambiguity is detrimental to concepts, according to Dr. Lesley Hoyles, “The microbiota refers to the micro-organisms and viruses associated with the human gastrointestinal tract. The microbiome refers to the genetic make-up of the whole of the microbiota: i.e. the genes from all the bacteria, eukaryotes, archaea and viruses,” (Hoyles, n.d.). Yet, The differences between the ecological interpretation and the biomedical interpretation is not so clear according to Gregory W. Schneider and Russell Winslow, as they

suggest that the microbiome is a new word and part of a larger “new lexicon” that is related to the language and social context of the Human Microbiome Project, ecology, and epidemiology:

In developing this new lexicon, the scientists who authored the June 2012 Nature article have gravitated toward the language of ecology and epidemiology, emphasizing words and phrases like “human-associated microbial communities,” ecology and populations, diversity and abundance, mutualism and competition, habitats and microhabitats. Using this language as a touchstone, the researchers approach the individual human as a collection of nested habitats, with various regions populated by various collections of microbes. One crucial word in this account, is *community*, used here in its typical ecological sense, referring to a group of plants or animals living in a specific region or habitat under relatively similar conditions, or the region in which these organisms live. (Schneider & Winslow, 2014).

Schneider and Winslow also bring to light some of the larger philosophical issues of an ecological or biological interpretation of the microbiome and how the interpretation of the microbiome changes other concepts, i.e. how does the interpretation of the microbiome as a community change the interpretation of concepts like individual, species, and human. Though Schneider and Winslow, argue that the ‘language of ecology and epidemiology’ influenced the usage and meaning of the microbiome, they fail to provide any historical or language evidence to support this claim instead offering what they call the “typical ecological sense”. Thus, while their argument connects and accentuates the importance of language and social contexts to understand the microbiome, it is difficult to support their specific claims.

John Huss, also, emphasizes the social context of the microbiome referring to an “ecological preconception” and also recognizes the importance of the language and historical context of the microbiome, acknowledging a “curious tension” between if the microbiome is a neologism or a rebranded concept from ecology:

“The status of the human microbiome—whether it should be considered an organ, an internal feature of our developmental environment, or whether it should be assimilated

into an overall ecological preconception of the human being as superorganism—is also up for discussion (Foxman 2008, Juengst 2009)”. (Huss, 2014)

Huss ultimately argues that the microbiome is a new word and that microbiome research created a new ontology based on the social context of the microbiome, a “tools-to-theories heuristic” as the main influence of the microbiome stating that, “I shall argue, the primary drivers of the ontology emerging from microbiome research are its tools, including metagenomics” and, “A look at the scientific literature reveals that there are several factors operating simultaneously to shape an emerging ontology. What I have emphasized are the role of tools- largely metagenomic methods and statistical techniques- in shaping these ontological categories” (Huss, 2014). Huss promotes an interpretation of the microbiome that integrates social, historical, and language context, “several factors operating simultaneously to shape and emerging ontology” and separates the microbiome from other concepts by providing his own definitions for the microbiome, metagenome, and enterotypes. While the most inclusive of contextual evidence, some criticize Huss for lack of empirical evidence towards his claim the social context, or tools, methods, and statistical techniques, has influenced the microbiome (Adami & Bracken, 2016).

Moving forward, I argue that the social context of where and who uses the microbiome provides insight into how the microbiome concept was used and interpreted. Looking at biomedicine compared to ecology, biomedicine emphasizes on the discovery and diffusion of the microbiome, whereas, ecology focuses on linking the microbiome to other concepts within ecology, specifically biome and microbiota. Examination of the social context of the microbiome helps to determine the general vocabulary (or lack thereof) of the microbiome concept and the microbiome lexicon. This is important as the interpretation of other concepts may change based on the interpretation of microbiome, specifically in the cases of community, organism, or human.

Social context, also, provides insight into who has influenced the usage and meaning of the microbiome. Yet, to date no studies on the microbiome that have provided evidence on the actual usage and interpretation of the microbiome across social contexts.

Summary

To understand the usage and meaning of the microbiome, the historical, language, and social context of the microbiome must be considered. Some studies have briefly discussed the influence of one context on another, i.e. how the historical context influences the language (Lederberg, Prescott, Eisen), or how the language influences the social (Ursell, Marchesi & Ravel, Blaser), but to date no study has systematically analyzed how factors from these different contexts have influenced the usage and meaning of the microbiome. Although, many of the studies do not claim to accomplish this task, understanding the microbiome along these different contexts is a prerequisite to understand the variation in the usage and meaning of the microbiome over time. By systematically evaluating the factors influencing the microbiome across different contexts, insight into who, what, where, when, and how specific to the microbiome can be answered. While research on the microbiome has been conducted with specific regards to the historical, language, or social context of the microbiome, few studies have attempted to integrate these different contexts. Additionally, all studies on the microbiome have so far, failed to provide substantial evidence towards their claims on the variation of the microbiome or influences on the microbiome.

KNOWLEDGE: REVIEW OF PREVIOUS APPROACHES

Background

In a collection of three articles from 2005 to 2016, clinician and professor of medicine John P. A. Ioannidis advocated that one of the major problems within biomedicine today is “context placement and information gain,” specifically pointing to the lack of awareness of what is already known within biomedicine so new knowledge can be placed in the appropriate biomedical context (Ioannidis, 2005, 2014, 2016b). Ioannidis’ sentiment echoes with medical professionals, scientists, and researchers within biomedicine and science at large, as increases in scientific publications, citations, and research output due to big data have led to questions about how to measure knowledge and changes in knowledge (Lazer, Kennedy, King, & Vespignani, 2014). Some of the more popular traditional approaches to measuring and understanding knowledge within biomedicine include systematic reviews, biomedical ontologies, and text mining using natural language processing (NLP) (Moher et al., 2015). Other methodologies in the humanities and social sciences have used publications, citations, or concepts to measure knowledge. In most cases, the broad argument is the study of material artifacts or research outputs are useful because these objects are proxies of knowledge (Jong & Slavova, 2014). Yet, there are criticisms on how accurately these proxies reflect knowledge and how to use these proxies to assess changes in knowledge (Bellis, 2009).

In this chapter, I review approaches used to measure knowledge, focusing on systematic reviews, ontologies, and text mining within biomedicine. First, I argue these methods by themselves are ineffectual and do not accurately reflect changes in knowledge within biomedicine. Further, I contend novel approaches are necessary to take advantage of big data and the current surge of publications, citations, and information and manage issues related to

analyzing high dimensional data. I begin by reviewing how systematic reviews, ontologies, and text mining have been used to measure and analyze knowledge within biomedicine. I highlight how systematic reviews are considered the gold standard due to the amount of time, data, and expertise required to be completed, but are ineffectual in measuring knowledge because of a dependence on rigid protocols that are unresponsive to differences in context and bias towards clinically defined outcomes. Similarly, I argue that while ontologies provide conceptual models, interpretations, and shared vocabularies, ontologies do not accurately reflect knowledge because knowledge within biomedicine is dynamic and dependent on context, specifically the cases where a single concept has separate interpretations and multiple concepts have a single interpretation. Likewise, text mining and Natural Language Processing (NLP) has been used with both systematic reviews and ontologies to improve data identification, collection, and categorization, but most text mining approaches cannot explain knowledge differences across contexts, such as when concepts change or when different contexts interpret words differently. Corpus Linguistics, on the other hand, is better suited to discovering small differences within language but require more time, effort, and human guidance than NLP methods. I argue here that a hybrid of NLP and Corpus Linguistic methods, combined with other approaches could be a powerful approach to understanding changes in knowledge.

Next, I examine other approaches used in the social sciences and humanities to measure knowledge including publication metrics, citation analysis, knowledge maps, and conceptual change. I begin by providing some of the historical context of publications as a unit of analysis in analyzing knowledge, and refer to the excellent work by Derek de Solla Price and Eugene Garfield in using publications as a means to communicate, produce, and quantify the growth of scientific knowledge (D. De Solla Price, 1986; Garfield, 1955). Publications provide the ability

to compare researchers and scientists, can predict future academic success, and have been used to map the trajectory of scientific knowledge (Garfield, 1972; D. De Solla Price, 1986). However, I point publications by themselves are inaccurate proxies of knowledge as the content and related knowledge within publications are subject to extreme variability. Citations and citation dynamics of scientific articles and patents, also, have been used to trace the spread of scientific knowledge. Here, I agree citation dynamics are useful in mapping and linking authors and scientific specialties, but I argue that citations cannot be used to analyze changes in knowledge since citations and citation dynamics are unpredictable and not well understood (Edge, 1979; van Raan, 2004; Ke, Ferrara, Radicchi, & Flammini, 2015).

Concepts and conceptual change have also been used to assess changes in scientific knowledge. I recognize both the qualitative and quantitative approaches to analyzing concepts and highlight how the History and Philosophy of Science (HPS) has focused on analyzing concepts and issues related to understanding the nature and meaning of concepts, emphasizing the work of Thomas Kuhn and Paul Feyerabend (P. Feyerabend, 1962, pp. 28–97; Kuhn, 1970). I show other approaches have analyzed concepts computationally with quantitative results including knowledge maps, maps of science, and knowledge-based economy models. Yet, while many of these approaches frame historical inquiry as an experiment with testable hypotheses and emphasis on statistical results, I argue that most of these approaches fail to take into consideration conceptual variation within and across contexts.

After teasing out the different arguments supporting and criticizing publications, citations, and concepts as proxies to understand changes in knowledge, I argue integrating insights from the science of complex systems in studying the variation and changes to concepts and knowledge.

In the second half of the paper I argue that language is a complex system, and the analysis of complex systems or complexity science can help to understand variation and changes in language and in scientific knowledge. I present an overview of complex systems or complexity science. I agree with previous studies arguing language is a complex system and consists of multiple components and variants at the same time with properties that scale. I argue complexity science has provided tools and methods to understand variation and change within a complex system, and insight from complexity science can help in understanding how knowledge changes within language, social, and historical systems and data (Kretzschmar, 2015). At this point, I review the previous studies suggesting how social and historical context influence knowledge, including William Labov's study of New York City and George Zipf's study of Ulysses by James Joyce (Labov, 2006; George Kingsley Zipf, 2012). I claim the power law distributions commonly found in language and other complex systems, or the A-curve, can be used to show the variation and range of knowledge for a specific social context by displaying variation and change of words used over time.

Lastly, I propose adopting a framework from Jurgen Renn and Manfred Laubichler for analyzing the history of knowledge which emphasizes the use of context to systematically analyze the variation and changes to scientific knowledge (Renn & Laubichler, 2017a). I show results from preliminary experiments using Renn and Laubichler's framework on the variation and change of knowledge on the microbiome. Based on these results, I argue this framework offers a method to link multiple proxies of knowledge from different contexts to analyze the contextual factors or explanatory variables which characterize the variation, changes, and evolution of the microbiome. In summary, this chapter aims to provide a better understanding of the approaches and practices used to measure and assess scientific knowledge within

biomedicine and in science and help to clarify and deflate some of the tensions in debates about complex systems and language with respect to measuring knowledge.

Biomedical Knowledge Approaches:

Understanding changes and trends in biomedicine and biomedical knowledge is crucial for individuals, groups, and institutions as biomedicine improves people's lives, supports national economies, and leads to innovation (McGuire & Albert, 2014). Recently, biomedicine has experienced a rapid growth in data as a result of the emergence of new research areas such as genomics, molecular biology, gene therapy, personalized medicine, and pharmacogenetics (Clarke, Mamo, Fosket, Fishman, & Shim, 2010). This data from new research areas, in combination with increased data from patients, populations, health care systems, have drastically increased biomedical research outputs and knowledge. The National Library of Medicine (NLM), an important resource in understanding the growth of biomedical research, provides an index of scientific articles for researchers and scientists. Using the number of articles indexed in NLM as an indicator of the growth of biomedical knowledge, 813,598 articles were added to the approximately 24 million already indexed articles in the NLM ("NLM Detailed Indexing Statistics," 2018). The tremendous increase and aggregate of biomedical research data presents a unique opportunity to better understand and further biomedical knowledge and foster biomedical discovery. However, current approaches to understanding biomedical knowledge has not met the needs of scientists, researchers, patients, clinicians, administrators, and policy makers, because the flow of knowledge from biomedicine is too slow and in most instances the scope is too narrow (Krumholz, 2014). Within biomedicine knowledge is commonly assessed via systematic reviews, ontologies, or Natural Language Processing (NLP).

systematic reviews.

Systematic reviews are an approach to synthesize evidence, summarize, and characterize the knowledge within a research area by using the data and content within papers or study results, this includes analysis of patients, interventions, comparison, outcomes, and treatments (Mulrow, 1994). The general steps in a systematic review are: define a review question, create parameters for inclusion and exclusion of studies, search for studies and collect data, analyze the data, and offer an interpretation based on the findings (Moher et al., 2015). Most systematic reviews are designed to collect data using a specific form to pull out the same numerical data from each article such as: number of participants, mean treatment outcomes, time intervals between treatments, and so on (Moher et al., 2015). Due to the amount of time, effort, and expertise required for a review, they are considered the gold standard in biomedicine for understanding the knowledge of a research domain (Green-Hennessy, 2013; Impellizzeri & Bizzini, 2012). Though, there are different review approaches to synthesizing complex biomedical data, including narrative reviews which focus on explanation and interpretation, results from a survey of different biomedical reviews found systematic reviews received more citations than narrative reviews and were thought to be superior to narrative reviews in answering research questions, reproducibility, and dealing with bias (Faggion, Bakas, & Wasiak, 2017; Greenhalgh, Thorne, & Malterud, 2018). However, most systematic reviews only provide a small reflection of knowledge directly related to social and historical context because the methods for reviews overlook qualitative data and rely on simple, discrete, and quantitative data and leads to results used for judgements rather than explanations (Greenhalgh & Peacock, 2005; Pawson, Greenhalgh, Harvey, & Walshe, 2005). Further, the enormous uptick in the use of systematic reviews has led to the mass production of unnecessary, redundant, and conflicting systematic

reviews, and in spite of the increase in available research outputs most systematic reviews have used only a relatively small fraction of the huge amounts of data available (Ioannidis, 2016a).

Systematic reviews are useful in comparing and analyzing specific forms of data and have an important purpose in biomedicine, but knowledge is context dependent and subject to variation and change. Synthesizing complex evidence requires novel approaches beyond systematic reviews. These approaches need to be able to integrate complex contextual data generated by diverse methodologies, the results from which can be used to better understand knowledge and help meet the needs of anyone who can benefit from biomedical data (Dixon-Woods, Agarwal, Shona, Young, Jones, & Sutton, 2004). Ludwik Fleck, in his book *Genesis and Development of Scientific Fact*, details how the knowledge on the concept of syphilis was influenced by social and cultural contextual factors and explains how these factors shaped the interpretation of syphilis from a broad interpretation which included other sexually transmitted diseases, a pharmacologically-based interpretation which included other diseases cured by mercury, a causal interpretation related to a pathological symptoms caused by *Treponema pallidum*, to finally a unified single concept of syphilis (Fleck & Kuhn, 1981; Löwy, 1988). Knowledge of addiction, also, has undergone multiple interpretations dependent on context influenced by social and intellectual factors related to the treatment and diagnosis of addiction. The range of interpretations characterizing addiction include: a moral emphasis and non-medical problem with diagnosis and treatment of addiction conducted by volunteers in communities and shelters, forced treatment of addiction in jails, to addiction now being diagnosed and treated by medical professionals with an emphasis on pharmacological treatments like methadone (W. L. White, 1998). A more recent study on developmental biology, using the entire catalog of the General Embryological Information Service, found that the diversification of research on

developmental biology was driven by contextual factors, including: semantic, social, cultural, intellectual, economic, and institutional, and highlighted that knowledge changes were indicators of innovation and growth within the field (Crowe et al., 2015). The cases of syphilis, addiction, and developmental biology, as with many other cases in the history of medicine and biomedicine, display how knowledge is contextual and influences interpretation, diagnosis, treatment, practice, and use (Garrison, 1921; Ackerknecht & Haushofer, 2016). Thus, context is critical to understand changes in knowledge. A contextual emphasis to the study of knowledge resonates with a growing number of researchers, scientists, and clinicians advocating scientific knowledge extends beyond systematic reviews, controlled experiments, and rigid protocols (Malterud, 2001; Maasen & Weingart, 2013).

ontologies.

Ontologies emphasize simple syntheses of the knowledge of a research domain through the creation of controlled vocabularies, conceptual relationships, and repositories of data. (Ramaprasad & Syn, 2015). In biomedicine ontologies are primarily used as repositories of knowledge that are used to catalogue individuals, concepts, publications, and citations. Repositories built from ontologies also serve as conceptual models which deliver consistency, direction, and defined knowledge for sharing and incorporation by a community or domain (Spasic, Ananiadou, McNaught, & Kumar, 2005). To integrate knowledge, ontologies link things in the world to concepts, provide meaning and definitions for concepts, and create shared understanding on concepts (Rubin et al., 2006). The four main features of ontologies in biological and biomedical research are 1) standard identifiers for classes and relations that characterize phenomena, 2) definitions and vocabulary, 3) metadata describing how the classes and relations are interpreted, and 4) machine-readable axioms and interpretations enabling

computational capacities to features of the interpretations of classes and relations. (Hoehndorf, Schofield, & Gkoutos, 2015). The Medical Subject Headings (MeSH) ontology provided by the NLM, is one of the most well-funded, curated, and used ontologies in the world (pmhdev, n.d.). MeSH are a controlled vocabulary used to index the largest biomedical literature database, MEDLINE, linking journals, documents, books, and audiovisuals through a common conceptual ontology and vocabulary for researchers (Fieschi, Coiera, & Li, 2004). As of 2018, the MeSH terminology includes 28,939 main headings [MH], which are used to index content within MeSH. The choice of main headings or terms according to the NLM website on MeSH are “set forth in standard reference works, when these can be made to agree” (“MeSH Preface,” 2014; “Whats New for 2018 MeSH,” 2017). Many studies have used the MeSH database for information retrieval, comparison of texts, text classification, citation analysis, identifying research trends, characterizing research profiles, and MeSH indexing (Kim, Fiorini, Wilbur, & Lu, 2017; Mao & Lu, 2017).

However, despite being the significant amount of resources and effort being put into MeSH, MeSH suffers from issues found with other ontologies such as lack of context, difficulty keeping up with the pace of new concepts, and differences in interpretation (Hoehndorf et al., 2015; Ramaprasad & Syn, 2015). Multiple studies found a significant amount of context missing from the MeSH representation of concepts and (Lu, Kim, & Wilbur, 2009; K. Liu et al., 2015; Peng et al., 2016). The multiple concepts listed by MeSH for the microbiome, as seen in *Figure 1*, make it difficult to understand if the microbiome is one concept or many a the result when searching for “microbiome” in the MeSH returns “microbiota” not microbiome, and on the page returned has other concepts listed with the microbiome as “Entry Term(s): Microbial Community, Microbial Community Composition, Microbial Community Structure, and

Microbiome, Human,” and the page also refers to another concept in the “See Also: Metagenome”. The MeSH Ontology results, also, creates additional confusion by listing the same interpretation or scope note for the concepts of microbiota, microbiome, and human microbiome, the interpretation listed as: “The full collection of microbes (bacteria, fungi, virus, etc.) that naturally exist within the human body as identified by their genomic sequence regardless of whether or not they can be cultured,” as seen in *Figure 2* (“MeSH Browser,” n.d.). Other problems with the MeSH ontology include: new biomedical concepts are added to the MeSH long after these new concepts have been found frequently in the literature, MeSH terms may in some cases not actually be present in the publication but added by indexers to a list of keywords for publications, changes in terms and knowledge are not re-indexed in MeSH or any of the ontologies provided by the NLM as the NLM cannot re-index or add changed or new terminology to previously published articles, and incompatible computer descriptors and manual descriptions for context within MeSH (Aronson et al., 2000; M. Huang, Névél, & Lu, 2011; Mao & Lu, 2017). Therefore, the MeSH ontology does not provide the most accurate knowledge or indicators of knowledge and as in the example of the microbiome, can create more confusion surrounding concepts and the interpretation of concepts. While, the MeSH database contains important knowledge used to understand concepts, such as words used to define a concept, author affiliations, methods, tools, and other contextual factors important to understanding a concept, the current ontology does not take advantage of the information available (M. Huang et al., 2011; Mao & Lu, 2017).

With so many fields, subfields, and divisions between biomedicine researchers, biomedical ontologies could serve an important purpose in helping to understand the knowledge landscape of biomedicine. Biomedicine is distributed into highly specialized fields and subfields with little

communication amongst disciplines, MeSH and other ontologies could be used to identify connections between individual elements of biomedical knowledge and help clarify the many overlaps and misrepresentations of the biomedical knowledge throughout biomedical literature (J. Swan, Bresnen, Newell, & Robertson, 2007; Newell, Robertson, Scarbrough, & Swan, 2009). However, without a single standardized ontology many point out that biomedicine has too many ontologies with many overlapping concepts or terms, ontologies are too specific for practical use, and ontologies within the same domain have major differences in content and interpretation (Pesquita, Faria, Falcão, Lord, & Couto, 2009; Rubin et al., 2006; Kamdar, Tudorache, & Musen, 2017; Faria et al., 2018). A promising alternative to a standardized biomedical ontology that could help ameliorate many of the problems with ontologies in biomedicine is to link concepts to textual content and emphasize contextual language, social, and historical differences.

Context is important when mapping a concept to an ontology because concepts exhibit variation in both form and function. This variation can be seen when the same concept has multiple interpretation or when multiple concepts have the same interpretation. In the case of one concept with multiple interpretations, the concept of a promoter is defined by different social contexts which results in biology defining a promoter as a “binding site in a DNA chain at which RNA polymerase binds to initiate transcription of messenger RNA by one or more nearby structural genes”, whereas a promoter in chemistry is a “substance that in very small amounts is able to increase the activity of a catalyst” (Spasic et al., 2005). Likewise, the concept Ferritin has multiple interpretations within the social context of biology, as both a protein that stores and releases iron and as a laboratory test (Jutz, van Rijn, Santos Miranda, & Böker, 2015; “Ferritin test - Mayo Clinic,” n.d.). Whereas, the concepts PTEN and MMAC1, are two different concepts for the same gene (Naguib et al., 2015). Simply stated, the problem with mapping concepts to

ontologies is that words having multiple meanings, like the word *apple* has one meaning as a fruit and another meaning as *Apple* the technology company, and different words can have the same meaning, as the past tense of the verb sneak has a standard past tense *sneaked* and a nonstandard past tense *snuck* which are both present in the English language as irregular verbs. With these issues related to concepts, meaning, and context in using ontologies, it is problematic for ontologies to measure or characterize changes knowledge without referring to context. Some have claimed computational approaches such as text mining, can help provide clarity on the knowledge about a concept by limiting the potential meanings of a concept within a text by using text and textual content to link concepts, contexts, and knowledge to ontologies, (Koehler et al., 2005; Mortensen et al., 2015).

Text is an important resource within biomedicine, as publications, citations, methods, results, and patient records are all proxies of knowledge within biomedicine based on text. Text mining approaches provide quick and computationally cheap methods to understanding knowledge and concepts within textual data and have become more popular because of the increasing volume of biomedical research and hypothesized subsequent increases in knowledge (Aronson, 2001; Fan, Wallace, Rich, & Zhang, 2006; Spasic et al., 2005; C.-C. Huang & Lu, 2016). In general, text mining approaches can be either supervised or unsupervised. Both approaches make use of algorithms and computation and focus on classification, clustering, associations, and relationships within textual data (Aggarwal & Zhai, 2012). Supervised approaches to text mining are directed methods to understanding textual data when there is a specific target in mind and unsupervised approaches focus on finding latent structures or relations among data (Aggarwal & Zhai, 2012). Both supervised and unsupervised analysis of textual data within biomedicine have been conducted using natural language processing (NLP) for named entity recognition,

relationship extraction, information retrieval, and normalization of concepts (Renganathan, 2017).

natural language processing (NLP).

NLP is a collection of methods to analyze words and the relationship of words quickly within large collections of text (K. B. Cohen & Demner-Fushman, 2014). Generally, NLP approaches to analyzing language focus on: characterizing words or texts based on the statistical analysis of the frequencies of documents, sentiments (words with pre-defined meanings), correlations between words, and clustering of words. The terms computational linguistics (CL) and Natural Language Processing (NLP) are mistakenly used interchangeably, but most of the literature acknowledges CL as the larger subfield while NLP is a part of CL (Manning & Schütze, 1999). NLP from this perspective is a methodology of tools and engineering solutions to analyze language or create results from textual data (Alexander Clark, Fox, & Lappin, 2013). In most NLP studies, meaning is directly interpreted from usage and usage is evaluated through the probabilities and patterns of linguistic variables in language. NLP offers fast computational time when analyzing large amounts of data, because in most cases model training and model output is based on a document or collection of documents reduced to a word-document matrix or bag of words (BoW). The reduction of the data increases speed but comes with deficiencies in interpretation and meaning of results as NLP ignores context.

One NLP approach to characterizing the sentiment or meaning of words and documents is sentiment analyses (SA). SA characterizes the sentiment or opinion content of a text or word via sentiments designated to specific words prior to the analysis. In most cases, the sentiments are used to create results that focus on binary oppositions, i.e. for or against, like or dislike, good or bad, positive or negative (Pang & Lee, 2008). SA can be unsupervised or supervised, with the

supervised approaches requiring a lot of work and effort to create lexicons or vocabularies to use in the analysis of text. SA has been used to improve products based on reviews, analyze attitudes towards people or events, and evaluate responses via social media sites like twitter, Facebook, and Instagram (Baccianella, Esuli, & Sebastiani, n.d.). The results of sentiment analysis have also been used to estimate the overall sentiment of texts, highlight the most used positive and negative words, and be used to understand attitudes and opinions (B. Liu, 2012). However, criticisms of SA include the validity of the meaning attributed to words and the difficulty in understanding categorizing or measuring opinion in language like cases of irony, sarcasm, and rhetoric in text (Wiegand, Balahur, Roth, Klakow, & Montoyo, 2010).

Another NLP approach for characterizing words or documents is the topic frequency – inverse document frequency (tf-idf) statistic, which classifies documents based on the frequency of their words in relation to the entire corpus. The tf-idf or text frequency- inverse document frequency statistic counts the number of occurrences of each topic or word in each document of a collection of documents or corpus and compares these frequencies with the number of occurrences of words based on the full corpus (Blei, Ng, & Jordan, 2003). Studies that use tf-idf have been used to increase the speed and efficiency of information retrieval in queries by characterizing and classifying documents in corpora (Trstenjak, Mikac, & Donko, 2014). However, criticisms of tf-idf point to how tf-idf does not assume a mathematical model (usually it is explained via Shannon's Information theory), in most documents the dimensionality or size of the topics is too large, and tf-idf does not provide information about the inter or intra-document structure (Blei et al., 2003).

Analysis of words and how they co-occur or sets of co-occurring linguistic variables within a given window or n-grams (i.e. N=2 is a bigram) is another NLP approach to characterizing text.

N-grams characterize the relationship between one item (letters, words, syllables, phonemes) in a sequence of n -items (one item= unigram, two items= bigram, three items=trigram) (P. F. Brown, deSouza, Mercer, Pietra, & Lai, 1992). N-gram analyses are based on a probabilistic model or language model (LM), with LM probabilities coming from the mathematics of Markov Chains (bigrams and trigrams) and Shannon's Entropy Theory (Seneta, 2006; Shannon, 1948). N-grams have been used to create text or subject classifiers for authors, texts, or words supporting both supervised and unsupervised learning algorithms, like decision trees, Naive Bayes, k-nearest neighbors, neural networks, and support vector machines (Miao, Kešelj, & Milios, 2005). While computational fast and easy to run, criticisms of n-grams include: high rate of data-based errors, optical character recognition of large text can lead to miscalculations (i.e. cursive and older written texts), no weight or difference in significance of one paper from another, differences in test vs. training corpus, difficulty in detecting differences in meaning or context of words due to the complexity of language, and narrow interpretative scope (Zhang, Zhao, & LeCun, 2015). Many critics of NLP argue a combination of usage, close reading, and domain expertise provides better interpretation and understanding of texts, language, and knowledge (Alexander Clark et al., 2013; Renganathan, 2017).

topic models.

Topic modeling is a flexible NLP approach used to discover the main themes of texts, provide an organizational structure to documents according to themes, and model both text and metadata of documents (Ramage, Hall, Nallapati, & Manning, 2009). Topic modeling is part of larger field of probabilistic textual modeling focused on predictive models and topic representations of a text or collection of texts and refer to a range of machine learning algorithms

including both supervised and unsupervised approaches. Generally, topic models characterize the words from a text or collection of documents into meaningful groups and helps lead to the discovery of unique patterns of words which group together (Blei, 2012). A topic model assumes a set of latent topics, a topic being a cluster of words, can be predicted via a multinomial distribution of all words, and models every document in a corpus to a set of topics (Chang, Gerrish, Wang, Boyd-graber, & Blei, 2009). Generally used in biological and medical document mining, topic modeling has also been used for biological and biomedical text mining to discover: genes and functional groups, ontological terms and their latent relationships, DNA sequences, whole genomes, shape descriptors and brain surface, and clinical events and category of patients (Lee, Liu, Kelly, & Tong, 2014; Masseroli, Chicco, & Pinoli, 2012; La Rosa, Fiannaca, Rizzo, & Urso, 2015; Dawson & Kendzierski, 2012; L. Liu, Tang, Dong, Yao, & Zhou, 2016).

Among the many different topic models, latent Dirichlet allocation (LDA) is one of the more widely used. LDA is a completely unsupervised machine learning algorithm which models a single document or a collection of documents as a mixture of topics with the total number of topics being a flexible number chosen by the researcher beforehand. Topics are hidden within text as groups of words, and LDA predicts the proportion of topics for any document using Markov chain Monte Carlo (MCMC) convergence diagnostics, Gibbs sampling, and Bayesian nonparametrics to yield fuzzy clusters of words (topics) occurring together and the probability of these clusters appearing in a document (Blei, 2012). Topic models have also been used to identify themes or patterns in a document or collection of documents, trends on the linguistic surface, and reduce the dimensionality of textual data (Blei & Lafferty, 2007; L. Liu et al., 2016). The result of a topic model are topics, words, and the probabilities of both used to discover hidden themes in large collections of documents and reduce the dimensionality of text (Blei et

al., 2003; L. Liu et al., 2016). Topic models have also been used to make specific claims about the structure or changes of knowledge. An LDA topic model from PubMed Central articles and citations was used to map the knowledge structure and the overall trends in bioinformatics field (Song & Kim, 2013). Another study modeled the interests of authors (i.e. whether authors addressed a single topic or multiple topics in their work) by creating an LDA author-topic model (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004)

Topic modeling can be used to divide a word into different senses or contexts, but a topic model cannot determine the part of speech (noun, adjective, or verb) of a word. This is a critical distinction, as language is nuanced and complicated, a word can have multiple interpretations, or multiple words can have a single interpretation, and many words carry different connotations and denotations. The denotation vs the connotation for the word *snake* (i.e. snake as an animal vs. snake as a malicious person) highlights these distinctions, as do the subtle differences in connotations in the words *unique*, *different*, and *peculiar* as all of the words can be used synonymously but have different implied meanings with *unique* being more socially acceptable than *peculiar* (English & Underwood, 2016). Other criticisms, of topic models include their sensitivity to noise, a topic model algorithm ignores the syntax of the texts it processes by looking at the groups of words occurring together, but the model does not take into account location or word order and will include words from any location within a text such as the title, headers, footnotes, page numbers, or citations (Underwood, 2012). Some critics, describe deficiencies in how LDA topic models represents texts, as there is no standard methodology for determining the number of topics to use, overfitting can occur due to the large number of free parameters, and the use of the Dirichlet in the model restricts the clusters of words and is not compatible with Zipf's law (Gerlach, Peixoto, & Altmann, 2018). Thus, topic models provide

insight into word clusters and underlying themes of text, but ignore word order and location, and should be used with other methods for interpreting the meaning of words.

Within biomedicine all of these NLP approaches have been used to speed up the text clustering and classification or text mining process leading to a substantial decrease in the amount of time required to carry out systematic reviews by (Thomas, McNaught, & Ananiadou, 2011). A systematic review on text mining approaches for study identification in systemic reviews found that text mining did save time and help in screening papers for systematic reviews (O'Mara-Eves, Thomas, McNaught, Miwa, & Ananiadou, 2015). However, as already mentioned systematic reviews are not an accurate indicator of knowledge. Systematic reviews only provide a sample of the knowledge in a given area and do not account for all the knowledge in an area or domain, systematic reviews do not measure knowledge or provide insight into how knowledge changed or what influenced change, and systematic reviews integrating text mining do not overcome these issues (Greenhalgh & Peacock, 2005; Pawson et al., 2005).

Beyond systematic reviews, text mining has been used in other ways to understand knowledge within biomedicine. A survey of biomedical text mining by Aaron Cohen and William Hersh, details how researchers have used NLP approaches combined with text mining to manage information overload within biomedicine, and found NLP was useful in Name Entity Recognition (NER), text classification, relationship extraction, and hypothesis generation (A. M. Cohen & Hersh, 2005). Yet, Cohen and Hersh stress the inadequacy of current text mining results when dealing with contextual issues such as: when a concept has multiple different interpretations based on context, many biomedical concepts have several names that refer to the same thing, gene names are often confused with the function of the gene, and biological entities have multi-word name as in “carotid artery” (A. M. Cohen & Hersh, 2005).

The National Institutes of Health has held multiple conferences, symposiums, created the NLM Indexing Initiative and other events emphasizing the use of text mining tools and NLP with biomedical data and texts. The NLM provides multiple resources detailing how to integrate NLP and text mining approaches and makes selections of biomedical data for these approaches freely available. Yet, while the NIH provides numerous tools and resources for analyzing textual data, many of these resources are outdated with broken links or links to pages not working in most cases. Reviewing these resources, they were found to be unhelpful in providing basic tasks for understanding the difference between concepts, separating unique concepts, or in assessing how knowledge within biomedicine has changed. The Unified Medical Language System (UMLS) Terminology Services, for example, is a set of files and software that integrates different biomedical and health vocabularies to link health information, medical terms, drug names, and billing codes cataloged by the NIH. Many of the tools available from the NIH use the UMLS or parts of the UMLS include: the Metathesaurus, Semantic Network, and the SPECIALIST Lexicon and Lexical Tools, the LexAccess Web Tool to search for terms, an Interactive Medical Text Indexer (MTI) which provides indexing recommendations for text based on MeSH terms, a MTI Machine Learning Package which provides machine learning algorithms to analyze texts, MEDLINE N-gram set for individual users, Phrase2MeSh that searches for matches between a phrase and MeSH indexing assigned, MeSh on Demand which identifies MeSH terms and lists similar articles to the submitted text, MetaMap which plots biomedical texts to the UMLS Metathesaurus or identifies Metathesaurus concepts in a text, and others. The UMLS can be installed locally on a computer and be used by individuals for free. According to the NIH, the UMLS and the aforementioned tools have been used to group

synonymous terms into concepts, categorize concepts, determine relationships between concepts, and for health statistics reporting (“UMLS Quick Start Guide,” n.d.).

Using these tools to discover more information on the microbiome yielded little to no useful results in understanding how the microbiome has changed or what the microbiome is, as in *Figure 3.*, the results from a search using the Metathesauras for microbiome returned “No matching results found” with suggestions for “Did you mean: microbe, microban, microbio, microbase, and microbead.” Most of the tools seem to be made for use by either computer scientists or text mining specialists and not anyone else, as many of the tools require the user to provide their own source text for analysis, a lexicon or dictionary, and an understanding of the difference between different NLP techniques, e.g. topic models, N-grams, latent semantic analysis, Vector Support Machines, and so on (Aronson, 2001; Yeh, Ke, Yang, & Meng, 2005; Z. Luo, Duffy, Johnson, & Weng, 2010; C.-C. Huang & Lu, 2016). The tools being provided by the NLM, also, were primarily built to support or use NIH data as the tools not requiring user inputs used datasets or results from NLM resources. Studies using these tools have provided interesting results, but the results are based on the MeSH and other biomedical ontologies and therefore have all the previously mentioned issues related to concepts, context, and knowledge (Holzinger & Jurisica, 2014; C.-C. Huang & Lu, 2016). Still, despite these issues with text mining approaches based in NLP, these approaches are commonly used to understand changes in knowledge within biomedicine because of the high demand for insight into biomedical knowledge and knowledge from data (Holzinger & Jurisica, 2014). While NLP has received more attention, a small handful of studies have used methods and approaches from linguistics, specifically Corpus Linguistics, to gain insight into biomedical knowledge (Seale, Ziebland, & Charteris-Black, 2006; Lawrence Edwards et al., 2016).

corpus linguistics: corpora, keywords, and collocation.

Linguistics provide models and theories for the study of language, language use, and the knowledge within language. According to linguists, language is a complex phenomenon and process. Language is a set of items, or linguistic variables such as sounds, words, grammatical structures, multi-word units (chunks, clusters, sequences, phrases), semantic categories, and so on (Hudson, 1996; Wardhaugh, 2009). Linguistic models and theories focus on linguistic variables of language as variables in models to explain language variation (Levshina, 2015). Some subfields within linguistics focus on different levels of language including: phonetics, morphology, syntax, semantics, and pragmatics, whereas other linguistic subfields study specific aspects of languages: sociolinguistics, historical linguistics, psycholinguistics, linguistics typology, and language acquisition (P. Baker, 2010). While more studies have gravitated toward computational linguistics or NLP within biomedicine, a handful of studies within biomedicine have used methods and approaches from Corpus Linguistics to study language and knowledge (Adolphs, Brown, Carter, Crawford, & Sahota, 2004; K. B. Cohen & Demner-Fushman, 2014; L. Liu et al., 2016).

Corpus linguistics studies language in use and how language occurs naturally in texts. Language in use encompasses studying language patterns of linguistic variables, language variation between a group of speakers/writers, and language of different texts or groups of texts (Stefanowitsch & Gries, 2003). In linguistics, language is a system of linguistic communication particular to a group of individuals that has a variety of modes communication and linguistic items, modes of communication include written, spoken, and signed modes of communication; and linguistic items or linguistic variables are objects in language which can include sounds, words, grammatical structures, and so on (Wardhaugh, 2009). Using these two definitions as first

principles, language variation is variation in pronunciation, spelling, or usage of linguistic variables, and is a feature of a collection of linguistic variables (P. Baker, 2006). Corpus linguistic studies emphasize the use of text-based analysis to discover and assess the extent of the complexity of patterns of linguistic variables. Empirical analyses of linguistic variables lead to the discovery of unusual cases of language, such as rare linguistic frequencies and patterns related to social or intellectual contexts. Corpus based studies, in general, analyze the occurrences of a linguistic variable, or a single text or collection of texts (Biber, 2006). To accomplish this efficiently, corpus linguistic studies utilize computers and computational techniques during the data collection, curation, and analysis of texts for faster and larger analyses than a traditional close reading or introspection of the text (Biber, Douglas, Biber, Conrad, & Reppen, 1998). Corpus linguistics have revealed complex association patterns, or how linguistic variables are used in association with other linguistic variables from language context(s) or non-linguistic variables from other contexts (Biber, 2006). When analyzed from a Corpus Linguistic perspective, association patterns have shown the variability of language specific to different social groups, across time for a specific individual, and different varieties of language based on situation or registers, e.g. newspaper, academic article, or wanted ad (Biber, 2006). The unit of analysis within Corpus Linguistics is a corpus (or corpora plural), which is a ‘body’ or large collection of objects in which language occurs naturally that acts as a representative sample of a particular language variety, e.g. computer files, sound files, video data, and scientific articles (Hettel, 2013).

Assuming a corpus is representative of a language, a corpus captures the variety and specifics of language used at a specific time and/or place and corpus linguistics studies show frequencies, patterns, variation, and changes in language (Hunston, 2002). The difference

between textual databases compared to corpora are in the size and scope of the collection of texts, a textual database will have text data arranged as part of a relational database management schema, where the text data are organized as tables, and often times in a database the textual data is a combination written language and other variables, e.g. timestamps, social demographic data, or metadata (Hettel, 2013). Corpora used for corpus linguistics are usually a collection of written or spoken examples of text sampled based on a specific question or project. Corpus linguistic analyses can be carried out on data from textual databases, and corpus linguistic analyses have been conducted on a range of textual data from a single section or passage of a text to collections of millions of documents (P. Baker, 2010). However, in most instances the results from studies using textual databases or single textual objects cannot be used to make generalizations about language use beyond the source of the text (Kretzschmar, 2009)

The majority of corpus linguistic studies create or use a general or reference corpora as a baseline for comparison between other corpora. A reference corpus is usually a large representative sample of the population under study and can consist of 1) texts from multiple sources, e.g. newspapers, articles, and books, or 2) texts from one data source, e.g. scientific articles from a specific journal. The gold standard for reference corpora is the Standard Corpus of Present-Day Edited American English for use with Digital Computers better known as the Brown corpus, created by Henry Kucera and W. Nelson Francis in 1964, which consisted of 1 million words created from multiple sources of English language text, e.g. samples from newspapers, books, articles, published in 1964 (Francis & Kucera, 1964). The Brown corpus of 1960's along with three other corpora: 1)The Lancaster Oslo-Bergen (LOB) created from British language texts from the 1970s, 2) the Freiberg-Brown (FLOB) created from British language texts from 1991, 3) and the Freiberg-Brown corpus of American English from texts published in

1992, are known as the Brown ‘Family’ of Corpora and have been used to show variation and change in multiple Corpus Linguistic studies of language across social dimensions between English and British language (P. Baker, 2010).

Yet, not all corpora or corpus-based projects require large samples of text. Some corpus linguistics projects use specialized corpora, which have specific *a priori* restrictions on data capture. The Michigan Corpus of Academic Spoken English (MICASE), consists of around 2 million words of contemporary university speech via recordings from events, e.g. lectures, classroom discussions, seminars, advising sessions, at the University of Michigan (Simpson-Vlach & Leicher, 2006). This specialized corpus has been used to study the academic language use at the University of Michigan and found differences in gender, disciplinary divisions based on frequency of language use (Reppen, Fitzmaurice, & Biber, 2002). Some linguists argue that MICASE is only contextually relevant to studies based on academic language at the University of Michigan, while others some argue that the results from studies using MICASE are representative of all academic spoken English (Aijmer & Stenström, 2004)

Most corpus linguistic studies use a combination of a specialized corpora and reference corpora to discover unusual cases of variation within language (P. Baker, 2006). In most cases, a reference corpus provides information on normal patterns of language that are then compared to a specialized corpus to identify what linguistic variables are frequent or infrequent in the specialized corpus. Some studies have compared two or more specialized corpora, but in these cases issues of size, content, composition, and generalizability between the corpora must be considered (Biber et al., 1998). Generally, Corpus Linguistics analyzes frequency of linguistic variables especially words in a corpus. Frequency is important, as frequency can highlight the focus of a text or collection, indicate markedness, and characterize the relationship between two

words (P. Baker, 2010). Differences in frequency can show power imbalances or preferences or what is considered to be the 'norm' and what is the deviant or the 'outsider' (Douglas, 1966; Cixous, Cohen, & Cohen, 1976). Results from a frequency analysis of words can indicate a preference in the use of specific phrases characteristic of social classes, identify and characterize the difference between corpora based on time or other dimensions, including geography, institutions, and other social contextual factors (Trudgill, 1979; Labov, 2001).

Using the frequency of words a frequency profile can be created to emphasize differences in the patterns in language and reflect the components we tend to use with particular people/groups or in particular situations of use, social context, historical contexts, or knowledge of a system (Burkette & Jr, 2018). Frequency profiles have been used in biomedicine to compare professional vs lay language, compute semantic similarities within ontologies, and study long-term knowledge changes in cultural, historical, and linguistic meaning (Kokkinakis & Gronostaj, 2006; Batet, Sánchez, & Valls, 2011; Guiliano, Fraistat, Brown, Muñoz, & Denbo, 2011). The frequency of all words (tokens) compared to unique words (types) or token type ratio (TTR), shows the amount of variation in a text and is used to characterize the specificity or complexity of the language context. The TTR is calculated by dividing the total number of types or different words by the total number of words or tokens (P. Baker, 2010). The TTR shows the lexical variation of a particular text, with a high TTR indicating a high lexical variation and a low TTR indicating a low lexical variation.

Concordance analyses provide context to frequency-based results and help to confirm claims about language variation and change via a table that displays all the occurrences of a linguistic variable along with the context the linguistic variable is used in, i.e. the words or sentences the linguistic variable was used with. For example, the word 'bat' without any context

can refer to a small mammal or a tool for hitting a ball, but by viewing the language context in which the word was used, a mention of someone at the plate or is ball in the sentence as opposed to mentions of nocturnal eating or winged flying animal, it can be determined which bat was referred to (Hettel, 2013). Generally, concordance analyses are used to help support interpretations of the meaning and uses of words and assist in identify linguistic patterns, and used in conjointly with other analyses such as keyword analyses and collocational analyses (Stubbs, 1995)

Keyword analysis identify the lexical focus or *lexical saliency* of texts by comparing the relative frequency of words between different corpora and reveals the words which occur more in one corpus compared to another corpus (Kronberger & Wagner, 2000a). In most cases, a keyword” or “key words” are statistically significant words based on a comparison between two corpora. Keywords are not mutually exclusive, and a keyword in one corpus can occur in the corpus used for comparison. Different statistical measures can be used to determine the keyness of a word between by comparing the frequencies of all words between two corpora, as shown in Table 1 (M. Scott, 2018). The results from a keyword analysis is a list of keywords, or a Keywords in Context list (KWIC), both will be used interchangeably from this point forward, that show both positive and negative keywords. Positive keywords are words that tend to appear in one corpus compared to the other, and negative keywords are infrequent words found in one corpus compared to the other. Keywords have been used to show the difference between word usage between countries, in groups and out groups, how identity was constructed, and provided insight into language variation and change (Gabrielatos & Baker, 2008; P. Baker, 2012). Within biomedicine, keywords have been used to analyze the of opinions of practice-based research, comments from patients, the knowledge within communication events in a clinical setting, and

the evolution of research fields (Huntley et al., 2018; Maramba et al., 2015; Adolphs et al., 2004; Pérez et al., 2016). However, criticism of keywords or KWIC stress keywords are not absolutes and reveal tendencies in language, a keyword in one corpus does not mean that the word is not a part of another corpus. These tendencies are based off statistical tests, and depending on the statistical test, size of the corpora, and the reference corpus the results may be different (Oakes & Farrow, 2007). Other criticisms of keywords are they have no standardized cut off point and larger corpora generally produce more keywords than smaller corpora. It is recommended the results of KWIC analyses are examined within the corpora with close reading and concordance lists for validation and supporting evidence.

Collocational analyses show the relationships between linguistic phenomena and provide direct evidence of the links between words within small windows of text by taking into account the position of words co-occurring with each other (Firth, 1951; Clear, 1993; Evert, 2008). Two words collocate when they frequently occur near or next to each other. Collocates are calculated using a node word of interest and a span of words to the right or left to determine the strength of relationship between words. Collocates provide contextual information about words, reveal subtle meanings about words not explicitly mentioned in text, and emphasize both word occurrence and word order (Herbst, 1996; Hettel, 2013). Though all words co-occur together to some degree, in many cases specific are statistically more likely to occur together more frequently than other words. Considering the millions of words within the English language, many of those words are not used near other words ever, any of the potential millions words could occur together in a phrase or combination as there is no rule against it, but within language certain words occur together in statistically significant ways more than other words (Burkette & Jr, 2018). In *Words and Phrases: Corpus Studies of Lexical Semantics*, Stubbs argues that

collocates detail knowledge of language, knowledge of individual words, knowledge of predictable combinations of words, and cultural knowledge of which combinations of words are used:

“The individual word *round* can mean "circular", and the individual word *table* can mean "a piece of furniture with a flat top, which people can sit at, so that they can eat, write, and so on". The phrase *round table* has one meaning which is simply due to the combination of these individual meanings: something which is both "round" and a "table". However, it is also used in longer phrases such as *round table talks*. This means that a group of people, with interests and expertise in some topic, are meeting as equals to discuss some problem. This meaning relies on additional cultural knowledge: we would not fully understand the phrase unless we also knew that it is often used of discussions between political groups who are trying to reach agreement after some conflict.” (Stubbs, 2001, p. 10).

Stubbs provides evidence of how the meaning of words is easily observable when looking at collocates, despite words having so many meanings, again *round* and *table* can be interpreted different ways, as in: *a round number, a table wine, a timetable, Knights of the Round Table, table manners, drink someone under the table, payments made under the table* (Stubbs, 2001). Another example using collocates, highlights how collocates for the word *cool* are most often used with other words indicating recipes like *dry cool place* and *allow to cool* in the 1960s, but collocates for *cool* in the 1990's are more frequently used in metaphorical expressions such as *keep cool, he looked cool* (P. Baker, 2012). Collocates are unidirectional, meaning that if a word (w_2) is a collocate of the node word (w_1), this does not provide evidence that w_1 is a collocate of w_2 . Like keywords, collocates are calculated with a range different statistical tests used for collocates including: Mutual Information (MI), MI3, Log-likelihood ratio, Z-score, T-score, Chi-square, and the Dice coefficient, see Table 1.

Each statistical test or association measure highlights different relationships of words, but some of the tests yield overlapping results. MI emphasizes exclusivity between collocates and

highlights unusual combinations that occur infrequently and in many cases MI scores will include typos or other errors in text (Evert & Krenn, 2004). MI is sometimes used with a minimum threshold for occurrence for the collocate, collocation, or both (Evert, 2008). MI3 is the cubed version of MI and also results in relatively infrequent collocations, but in contrast to the results of an MI score the MI3 score gives more weight to observed frequencies (Evert, 2008). In general, Mutual Information and Z-score both give high scores to words with low frequencies, whereas Log-likelihood, T-score, tends to favor words with high frequencies. Some studies argue the use of one score over the other depending on the question, and other studies suggest using a combination of scores to determine a range of collocates, as each test emphasizes either frequency or saliency of words (P. Baker, 2006). Using a corpus of publications created based on a search for microbiome, the top 20 collocates of microbiome with their subsequent frequencies are shown to highlight the differences and similarities in results between each statistical test or association measure, displayed in Table 2. Collocates have been used to show the variation in language use (i.e. words and their collocates) and how variation is influenced by social factors including prestige, relative location, and interactions (Kemp-Dynin, 2005). Collocates within biomedicine have been used to analyze: language used with gout, social media texts, college curriculums, relationships between diseases and geography, age stereotypes, and redundancies in electronic health records (Lawrence Edwards et al., 2016; Türker, Şehirli, & Demiral, 2016; Budgell, 2016; Porter, Atkinson, & Gregory, 2015; Ng, Allore, Trentalange, Monin, & Levy, 2015; R. Cohen, Elhadad, & Elhadad, 2013). Therefore, collocates provide understanding of the relationship between words and how those relationships determine meaning and knowledge.

However, criticism of collocates are similar to the criticisms of keywords, as the flexibility in the use of collocate approaches indicates no standardized methodology, different statistical tests will yield different results, and results require human interpretation and comparisons with other supporting evidence. Noam Chomsky and other generative linguists point out that corpus linguistics replace induction with a hypothetico-deductive mode of reasoning, and argue no corpus of data can be used as the basis for linguistic generalizations because a corpus is an incomplete and unintentional pool of words (Chomsky & Halle, 1965, p. 15). Yet, counter to this, analyses of language reveal common patterns such as Zipf's law (which will be discussed later) which have been found occurring across multiple dimensions and scales and in other complex systems. Thus, Corpus Linguistics methods identify and analyze complex association patterns in which linguistic features are used in association with other linguistic and non-linguistic features in language. Within biomedicine, Corpus Linguistic studies have been limited because of the time and expertise required to carry out experiments compared to NLP approaches (Seale et al., 2006; Lawrence Edwards et al., 2016). Corpus Linguistics provides tools to analyze and focus on the variation and change of specific linguistic variables, words, or concepts over time. These analyses provide insight into smaller changes of knowledge at the level of language and individual words or concepts and provide interpretative value beyond the model used for words and words co-occurring with each other. However, corpus linguistics is not an optimal approach when dealing with high dimensional data over large time scales and does not solve all issues of creating knowledge from data.

The NIH created the Big Data to Knowledge (BD2K) initiative to: 1) improve the ability to locate big data; 2) develop and disseminate data analysis methods and software; 3) enhance training in biomedical big data and data science; 4) establish centers of excellence in data science

to test novel ideas and approaches and as a response to the problem of creating knowledge from data (Margolis et al., 2014). With big data and the influx of textual content the NIH and other government agencies are incentivizing the use of textual data as a source of information and knowledge (Manning & Schütze, 1999; J. Swan et al., 2007). However, many complexities arise including: multiple concepts, differences in interpretation, changing concepts, and complex relationships between concepts, when using language context as a proxy of knowledge and as the unit of analysis. Previous studies have shown the processes of knowledge diffusion, dissemination, transmission, utilization, and integration for biomedicine are contextual and different than that of physics, mathematics, or other fields (M. E. J. Newman, 2001; Green, Ottoson, García, & Hiatt, 2009). NLP provides a collection of quantitative methods to provide empirical evidence for variation and change in language by classifying texts via text categorization and information retrieval, and has been used to analyze, classify, categorize, and describe biomedical data. However, NLP approaches ignore contextual nuances of language and language use and cannot explain contextual differences between things like concepts or words (Jose, 2003). Corpus Linguistic approaches are more adept at understanding more nuanced differences in language, but require more time, effort, and human direction during analyses. I argue a hybrid approach of NLP, Corpus Linguistic, and other methods should be adopted to gain insight into biomedical knowledge. While, it is recognized both NLP and Corpus Linguistics have been used to gain insight into biomedical knowledge, a review of the literature as of this dissertation points to these two approaches being used in separate studies and never in combination. Additional approaches are needed to help account for the variation and change in scientific knowledge within biomedicine. With this in mind, I turn to review common

approaches from other domains used to measure and track knowledge, including publications, citations, and concepts.

Publications and Citations as Knowledge Review

Derek de Solla Price in, *Little Science, Big Science ... and Beyond*, accurately predicted science was going to get “big” through exponential growth in manpower, publications, and citations (D. De Solla Price, 1986). Price’s novel approach to quantify the growth in science was to use publications as the unit of analysis and as a means to show how scientists’ ‘communicate knowledge’ and produce ‘new knowledge’ (D. De Solla Price, 1986, pp. 56–58). Following in Price’s example, many studies have used publication counts as a measure of the knowledge produced by individuals, institutions, and scientific fields (Tabah, 1999; Bellis, 2009).

Subsequent studies have confirmed the exponential growth of science and the growth of scientific knowledge based on scientific publications and citations, or material artifacts (Sandström & Besselaar, 2016). Publication data metrics have also been used to characterize productivity and determine funding, job promotions, and the impact of a scientists’ work (S. Carley, Porter, & Youtie, 2013; Jong & Slavova, 2014). Publication counts have also become valuable in comparing the outputs of actors (e.g. individuals, journals, institutions, research groups, countries), determining academic superstars, and analyzing trajectories of scientific knowledge (Y. Liu & Rousseau, 2010). Within biomedicine, publication analyses are often included in systematic reviews or meta analyses and include summaries of the type of study design, institutional setting, type of outcomes, number of treatment groups, and number of patients (Buchwald et al., 2004).

However, with the increased use of publication counts and publication data metrics as a way to evaluate researchers, research, and scientific knowledge, there has also been an increase

in the number of papers published and a decrease in the overall scientific contribution or content of publications (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015). Scientists and authors now under pressure to produce more papers sacrifice quality for quantity and are in a mindset of ‘publish or perish’ (Sandström & Besselaar, 2016). In an effort to produce more papers, authors have increasingly engaged in unethical research practices including: fraud, ghost authors, plagiarism, a single publication being published in multiple locations, and splitting research into multiple fragments for the sake of publication (Rawat & Meena, 2014). Scientific journals have also contributed to decreases in the quality of content within publications. Scientific papers are being retracted at alarming rates because the barriers to publication in scientific journals has lowered and predatory journals solely focused on publishing will publish anything for a fee including fake papers created by computer algorithms or papers based on pop culture references (Steen, Casadevall, & Fang, 2013; Van Noorden, 2014). Within biomedicine, the increase in papers has also seen an increase in systematic review articles with most of the reviews being characterized as “unnecessary, misleading, and conflicted” and “Instead of promoting evidence-based medicine and health care, these instruments often serve mostly as easily produced publishable units or marketing tools,” (Ioannidis, 2016a, p. 1). These combined behaviors of authors and journals to produce and publish more publications by lowering the quality of the content of papers, highlight how publications by themselves are unreliable proxies of knowledge. Even Price confirmed this point agreeing historically publications are inaccurate measures of knowledge:

Only incidentally does the paper serve as a carrier of information, an announcement of new knowledge promulgated for the good of the world, a giving of free advantage to all one’s competitors. Indeed, in past centuries it was not uncommon for a Galileo, Hooke, or Kepler to announce his discovery as a cryptogram of jumbled letters that reserved

priority without conferring the information that would help his rivals. (D. De Solla Price, 1986, p. 61).

Admittedly, some of these cryptograms were followed up by actual publications and this practice is not the norm, but given the problems with publications some have turned to using how a publication is cited or citation counts and citation metrics as a measure of knowledge.

Similar to publications, citations have been used in many studies to observe and measure knowledge (Bellis, 2009; Keramatfar & Amirkhani, 2018). One of the strongest advocates of citation analyses, Eugene Garfield, recognized the cognitive potential of citation indexes and created the *Science Citation Index* (SCI) to promote the use of citations as an approach to gain insight into knowledge and scientific disciplines (Garfield, 1955). Using citation dynamics based on citations as the unit of analysis to analyze the trajectories and collective behavior of citations, Garfield describes his vision for the citation analyses of the SCI:

...to build a model of the journal communication network that will provide more functional definitions of disciplines and specialties, that will make it possible to define in detail how different fields of knowledge interact, that will provide methods of predicting interdisciplinary impact, and that will provide more effective ways of monitoring research performance (Garfield, 1972, p. 478).

Following Garfield's lead, citation dynamics have been used to: identify key documents in a field, map scientific specialties, describe the structure and change of research fronts, intellectually link authors, quantify impact, and as a means to trace the transmission of knowledge within scientific communication (Garfield, 1955, 1972; Pinski & Narin, 1976; Lewison, Rippon, & Wooding, 2005; Bellis, 2009). Price also advocated using citation dynamics to measure knowledge by commenting how citations are a form of "scholarly bricklaying" and are used in "maintaining intellectual property" and help to establish "priority claims" by forming a "corpus of common knowledge" (D. De Solla Price, 1986, pp. 58–59, 71).

Price also noticed through analysis of citation dynamics citations followed power-law distributions and posited a model of cumulative advantage for publications (Derek De Solla Price, 1976). Price's cumulative advantage model states the probability of a publication to garner new citations is correlated to the number of citations a publication already has (Derek De Solla Price, 1976; Ke et al., 2015). Other studies have confirmed the mechanism of cumulative advantage via the study of citation dynamics or preferential attachment, by observing heterogeneous connectivity patterns in networks across domains (Barabási & Albert, 1999; M. Newman, Barabási, & Watts, 2011). Within biomedicine, the NIH offers an iCite tool to help in the analysis of citations (<https://icite.od.nih.gov/>).

However, the study of the citation dynamics has also revealed unpredictable behavior in citations which make citations an unreliable proxy of knowledge. The majority of citations have been shown to have specific lifespans which are time dependent creating radically different results for the same citation or collection of citations depending on when the analysis is conducted (Jaffe, Trajtenberg, & Henderson, 1993; Roach & Cohen, 2012). Generally, most citations have relatively small windows of time to become frequently cited, but some citations are 'sleeping beauties' and have 'second acts' or delayed impacts and see a considerable increase in citations many years after being published (van Raan, 2004; Ke et al., 2015; Wang, 2013). The time window of citations and the phenomena of sleeping beauties and second acts stress how little is understood about authors' motivations and choices on who they cite, and the difficulties in predicting publication impacts via citation dynamics. Several studies investigating author's motivation and choice have revealed: how the value of specific citations were not as important to the citing authors as it was to the citation metrics, some citations were used to bolster prestige claims for friends or themselves, highly cited papers were not related to intellectual significance

or scientific impact, and subjective criteria for citations centered on the publications age (with newer publications being used more than older publications) or specificity (with publications published from higher journal impact factors being cited more than other publications) (Cole & Cole, 1974; Bellis, 2009; MacRoberts & MacRoberts, 2018). In a study asking authors to self-report about how and why they cited specific papers, the self-reports showed the motivation for citing didn't sync with the actual use of the citation and the role of a specific citation was minimal towards knowledge claims (Amsterdamska & Leydesdorff, 1989; L. Leydesdorff, 1998). Other criticisms of citations focusing on issues of impact and social context of citations include: lack of understanding what citations say about influence, citations ignores the context of the content within the publication and only shows that a citation was used not how influential or how often a citation occurs in a text, it is not always the case that the most highly cited papers are the most important ones, it is unknown if highly cited papers are pushing science progress through evolution or revolution, and highly cited papers may consist of easier as opposed to difficult findings (Edge, 1979; Ioannidis, 2014; MacRoberts & MacRoberts, 2018). Further, the behavior and dynamics of publications and citations have major differences related to social context. Biomedical publications and citations have different characteristics and attributes than publications and citations in physics, or math. Previous results have shown biomedical publications have more collaborators per author, a larger selection of articles that could potentially be cited, on average more papers per author, and on average more authors per paper than astrophysics, condensed matter physics, high energy and physics theory (M. E. J. Newman, 2001).

In summary, both publications and citations are inaccurate measures of knowledge because of the quality and content of publications and the nescience towards the use of citations

and citation dynamics. Though, publication counts and metrics based on publications have become so ubiquitous counting publications are a standard approach to analyze, quantify, and measure individuals, institutions, and scientific knowledge, publications are unreliable as a way to understand knowledge and knowledge changes. Like publications, citations and citation dynamics are another approach to measure and empirically quantify changes in knowledge, but citations also fail to accurately depict knowledge because they are complex material artifacts with dynamics that are not well understood.

Concepts, Knowledge Maps, and Knowledge-Based Economy Review

Conceptual change is directly related to the structure and growth of scientific knowledge. History has shown the addition of new concepts such as *gravity*, *molecule*, and *virus*, were introduced as scientific knowledge changed, and the removal or replacement of concepts like *aether*, *phlogiston*, and *spontaneous generation* within the scientific discourse paralleled changes in scientific knowledge (P. R. Thagard, 1988). Similar to individuals and material artifacts, concepts can be analyzed as actors in the process of changes to knowledge and scientific advancement (Surman, Stráner, & Haslinger, 2014). Using concepts as an actor allows for a finer grained analysis of publications, citations, and other textual artifacts, and enables the use of the content within material artifacts as the unit(s) of analysis (Franzosi, 2004). This approach removes some of the ambiguity of measuring knowledge from a publication or citation, as publications or citations may contain hundreds of concepts. Analyses of concepts can highlight the variation of concepts within a language context like: how multiple concepts can have the same interpretation, a single concept has multiple interpretations, variation in form or function, and how change or differences at the word level (i.e. one word) alter in what way a concept is interpreted or used.

Traditional historical and philosophical approaches to analyzing concepts has focused on the nature of concepts and the meaning of concepts, specifically on the commensurability or incommensurability of concepts. Thomas Kuhn, argued substantively if two dissimilar theoretical methodologies, what he called paradigms, use a given word or term with fundamentally unlike and incommensurable meanings then communication is impossible across the two paradigms on a specific word and lead to no logical choice between two theories (Kuhn, 1970, p. 103). Paul Feyerabend said the information and interpretation of two theories cannot be compared if the theories contain incommensurable concepts (P. Feyerabend, 1962; P. K. Feyerabend, 1970).

The challenge of incommensurable concepts and issues related to concepts and their meaning are significant when considering if scientists can refer to the same concept across space and time, as there have been previous cases where issues related to conceptual ambiguity and the incommensurability of concepts led to diminished impact, scientific interest, and support of specific concepts (Lakoff, 1975; Zadeh, 1976; Drieschner, Lammers, & van der Staak, 2004). Gregor Mendel's groundbreaking work on genetics, for example, is one of the most well-known examples of the debates on the meaning and usage of concepts having deleterious effects to a scientist's work and career (Sandler & Sandler, 1985). Multiple studies have confirmed after Mendel's work was neglected and disregarded for 34 years, Mendel was so deeply disappointed by the lack of response to his historic papers on the concept of heredity, Mendel refused to publish his later works and became abbot of his monastery, giving up his research on heredity with his previous unpublished and potential future works permanently lost (Iltis, 1932; Henig, 2017). In psychology, the multiple uses of the concept "harm" in the psychological literature resulted in inaccurate diagnoses, scientific and public misperceptions, and splintered discourses

(Furedi, 2016; Haslam, 2016). Several studies have traced the gene concept and argued the gene has multiple meanings, anchored at different points in time, which are directly influenced by social, historical, and language context (Keller, 2002; Rheinburger & Muller-Wille, 2017). Some have gone so far as to claim biologists and molecular biologists have ascribed meanings so different to the “gene” concept there is now a “classical gene concept” and a “contemporary molecular gene concept” (Brigandt, 2010, p. 26). Like the arguments made by Kuhn and Feyerabend, researchers have argued these different characterizations of gene inhibit communication of the word gene across contexts and gene should be divided into multiple concepts (Gray, 1992; Neumann-Held, 1999; Portin, 2002; El-Hani, 2007)

Misinterpretations in conceptual use and meaning have led to exceedingly broad patent rights to industry and prohibited further research on innovative breast cancer drugs by researchers in studies from science policy (Bar-Shalom & Cook-Deegan, 2002). Further, incommensurable concepts when used for science policy can result in social and economic constraints, historical inaccuracies, underachieved outcomes, biased analyses, and misguided science policy (Grzeda, 2005; Olfati-Saber, Fax, & Murray, 2007). Contested concepts, also, have lower rates of collaboration between scientists, less evolution of scholarly activity, and received less funding than non-contested concepts (D. U. Hooper et al., 2005)

Without delving into the debates as to what a concept is, a simple representation defines a concept as a single idea represented as a word or words (e.g. a phrase), this definition of concepts allows for concepts to be analyzed computationally as parts of texts with quantifiable results (K. Carley, 1993). Paul Thagard in *Conceptual Revolutions* employs this representation in a computational analysis of new concepts to measure changes in scientific knowledge and support his argument of the slow growth of knowledge within science punctuated by major conceptual

revolutions such as: Newtonian mechanics, Lavoisier's oxygen theory, and Darwin's theory of natural selection (P. Thagard, 1992). Thagard describes the process of conceptual change by, " *conceptual combination* in which new concepts derive from parts of old ones" and uses the example of the concept of sound waves "is the result of conjoining the concept of sound with the concept of a wave" (P. Thagard, 1992, p. 8). Thagard's emphasis on past epistemology and the history and philosophy of science to better understand conceptual change is important and his computational analysis of texts yields quantifiable results on how scientific knowledge has changed, but Thagard's approach is problematic because it neglects the social context of the words being used and assumes that the meaning of words is static and unchanging.

Other approaches to measuring knowledge have used publications, citations, and concepts in varying combinations as proxies for knowledge to create knowledge maps and quantifiable results. Due to the sheer number of approaches and domains contributing to measuring knowledge including knowledge acquisition, knowledge engineering, and knowledge representation, the focus here will be to understand knowledge as it has been used to understand the social contexts or language context of concepts. For many of these studies, knowledge is defined as units of information and meaningful relationships exist between those units and sources of knowledge, which include: individuals, organizations, communities, and literature, and this definition also suggest that knowledge can be combined into knowledge collections in the form of databases, ontologies, and vocabularies (Eppler, 2008; Balaid, Abd Rozan, Hikmi, & Memon, 2016). By adopting this definition knowledge maps visualize words or keywords as a representation of a concept, and the location of these words can occur in the title, citation, or within the content of a publication. This definition also allows for the analysis of the relationships between concepts and other things like: concepts and other actors, concepts and

social variables, and concepts and events. Most knowledge maps visualize the relationships between words, articles, patents, citations, metadata, and authors allowing the unit(s) of analysis in a knowledge map to be flexible and tailored to specific question(s) and hypotheses of interest (Balaid et al., 2016; Owen-Smith & Powell, 2004; Rosvall & Bergstrom, 2008).

Some of the first knowledge maps were domain visualizations created from citation data, like “The Use of Citation Data in Writing the History of Science” by Garfield et al., which highlighted the history and discovery of the DNA code using statistics and network analysis (Garfield, Sher, & Torpie, 1964). In the article, Garfield and his co-authors argued computers could “aid the historian of science” by helping to identify and order key events and analyze the interrelationships and importance of key events (Garfield et al., 1964, p. i). Other studies have used knowledge maps or maps of science to visualize the structure of scholarly communities and networks, detail the growth and evolution of scientific fields, show how research topics diffuse, identify key actors or events within a discipline, inform policy decision, assess research performance, and provide insight into the dynamics of communicating knowledge across multiple dimensions (Noyons, 2001; Börner, Chen, & Boyack, 2003; Rafols, Porter, & Leydesdorff, 2010; Mutschke, Scharnhorst, Belkin, Skupin, & Mayr, 2017). Domain visualizations or maps of science are knowledge maps visualizing specific topics within a domain, collection of articles, or databases related to a domain. Studies using domain visualizations have depicted how methods, ideas, models, or results travel from one field to another by creating a knowledge map and tracing a pathway through the scientific literature (Almeida & Kogut, 1999; Breschi & Catalini, 2010).

With so many possible units of analysis, knowledge maps and the units on the maps can be analyzed by a range of either ‘inter-’ or ‘co-’relationships between units, including:

intercitation, interdocument, coassignment, co-classification, co-citation, co-word, or other measures from social network and network analyses such as: betweenness, indegree, outdegree, centrality, etc.(Börner et al., 2003; H. D. White & McCain, 1997). In most studies, multiple maps are used as the basis for the interpretation of the results. In *CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature*, Chaomei Chen describes a methodology to map the trends and patterns contributing to a concept's evolution in scientific literature using concepts and citation networks (Chen, 2006a). Chen's approach identifies words from scientific abstracts, titles, and cited references in combination with author data (name, discipline, who they cite, who cites them) and journal data (title of journal, publication domain, year of publication, citation years) to create networks across multiple time slices and networks of combined time slices. The software and method Chen developed connects documents that cite each other (citation networks), networks that connect documents that cite the same articles (co-citation networks), and concept networks from journal abstracts or titles, and uses a combination of co-citation, co-occurrence of words, and network analysis (Chen, 2006b, 2015). In a case study on mass-extinction research, to identify and evaluate the factors influencing concepts, Chen's effective knowledge visualization shows what group of concepts (i.e. words) has been used in mass extinction research historically and what articles, authors, and citations are connected to the concepts in mass extinction research (Chen, 2006a). Effective knowledge visualizations are designed to reduce visual search time, provide improved understanding of complex data, reveal unnoticed relations via visual inspection, and enable datasets to be observed from multiple perspectives simultaneously (Tufte, 1998). Knowledge map visualizations are useful because they are flexible and can be made from a wide range of different datatypes including: linear, planar, temporal, multidimensional, tree data; created with

different functions like: overview, enhancement, historical analysis, and extraction of data; and display the data as landscapes, circle plots, term plots, networks, bar charts, scatterplots, or correlograms to help with interpretation of the data (Shneiderman, 1996).

The “knowledge-based economy” framework from economics and organizational science has been used with knowledge maps as a framework for understanding knowledge in complex systems such as economies, government, and industries (Godin, 2006; Loet Leydesdorff, 2010). The knowledge-based economy is the “production and services based on knowledge-intensive activities that contribute to an accelerated pace of technological and scientific advance as well as equally rapid obsolescence,” (Powell & Snellman, 2004). Using this definition, most knowledge-based economy studies have measured the knowledge-base of a system via patents and the information within patents (i.e. citations or terms in citation data) as proxies for knowledge and knowledge processes. An advantage of patent and patent citation data is patent data can be analyzed across multiple dimensions including time, industries, institutions, and individuals when knowledge or a proxy of knowledge is conceptualized as a word, or citation (Roach & Cohen, 2012). Previous studies have created models from patent citation frequencies and created empirical results claiming to pinpoint where knowledge is produced, knowledge flows, changes in the knowledge related to the distribution of power in biomedicine, used to predict biomedical innovation, and used to make recommendations for policy surrounding biomedical innovation (Jaffe et al., 1993; Jaffe & Trajtenberg, 1998; Mogoutov, Cambrosio, Keating, & Mustar, 2008; Walsh, Cho, & Cohen, 2005). For example, by analyzing forward citation of patents (i.e. the references to a specific patent by later patents) the authors of, “Patents as proxies: NIH Hubs of Innovation,” advocated the volume of patents and patent citations could be used as a proxy for ‘knowledge creation’ and implemented as a metric for evaluating the return on investment for

biomedical research funding (Kalutkiewicz & Ehman, 2014). Yet, this study and others using patent data to create knowledge maps have issues with validity and interpretation, specifically empirical results from patent analyses are noisy measures of knowledge flows because patents are generally used to satisfy legal obligations and may not reflect actual intellectual influences or knowledge (Loet Leydesdorff, Kogler, & Yan, 2017).

Other criticisms of patents and patent citation studies have grown in recent years as the desire to understand knowledge and how knowledge flows have been recognized as important to understanding innovation (Jaffe & Trajtenberg, 1998; Agrawal & Henderson, 2002; Roach & Cohen, 2012). Similarly, knowledge maps and the knowledge-based economy framework provide methods to characterize and measure the structure of knowledge quantitatively and integrate multiple sources of information, but as previously mentioned publications, citations, and patents are inaccurate proxies of knowledge and their utility is limited in the interpretation of how concepts can change over time. Most studies using patents or knowledge maps, wrongly assume no change in all units of analysis from: individuals, publications, citations, and most importantly concepts and are not context specific. Rarely do knowledge maps or maps of science account for conceptual variation or differences in context, and instead create multi-modal maps across time scales creating fuzzy relationships between actors. In network analysis, it is difficult to interpret individual relationships and the dynamics of those relationships over time because node and network level relationships are subject to destabilization causing the dynamics to be unpredictable (K. Carley, n.d.; K. M. Carley, Lee, & Krackhardt, 2002; M. Newman et al., 2011). By adding other dimensions to links, the importance of links become subject to greater variation and it becomes difficult to interpret different statistical measures are actually showing and measuring. The betweenness centrality score on a multi-modal map linking concepts,

articles, people, institutions, and events in one space can highlight the most central actor but does not help to interpret the relationships among the actors or justify why these actors are related. Many of these different actors are part of complex systems which have unique properties, behaviors, and dynamics at different scales.

Previous work on complex systems has revealed language is a complex system (Beckner et al., 2009; Kretzschmar, 2015). Many studies looking at the complexity of language have provided insight and direction into interpreting language data and highlight some of the features of the complex system of language which need to be considered when analyzing concepts and knowledge related to concepts.

Language as a Complex System: Studies on the Variation and Change in Language

Language is a complex system with multiple components and multiple variants (words, multi-word phrases) co-existing and interacting synchronously (Ellis & Larsen-Freeman, 2009). Language has the same characteristics of other complex systems such as: open system, orderly, dynamic, scaling properties, and imbalance, and different factors influencing an individual's language (idiolect) and a community's language (communal language) at different strengths and at different times (Beckner et al., 2009). Studies have shown how an individual's language and communal language are interdependent; an individual's language comes from social context or interactions with other individuals in a communal language, and communal language comes from individual's languages interacting (Beckner et al., 2009; Burkette & Kretzschmar, 2018). This dynamic interaction of language and contextual factors can amplify or negate each other's effects and/or compete with each other, suggesting in order to understand the dynamic nature of language and the factors influencing language, language is best modeled as a complex adaptive system (CAS) with the following key features: (1) language as a system involves things

interacting with each other, (2) the speaker's language use (behavior) is based on their experience, (3) a speaker's language is the result of competing factors, and (4) language originates from context, which includes experience, social interaction, and cognitive processes (Beckner et al., 2009).

These ideas are common in sociolinguistics, or the study of language variation and change using language context and social context. Language variation can be studied at a specific instance of time or synchronic variation, or examined across different slices of time or events or diachronic variation (P. Baker, 2010). For the purpose of this dissertation, variation relates specifically to synchronic variation. Variation explains the possibilities or range of language context or phenomena within a specific historical and/or social context and change details what happened over time. Studies in language variation can describe how word is used differently based on geography, as in 2010 the word *fecal* in American English is spelled with the vowel /e/ and pronounced with emphasis on the vowel /e/ similar to *ear*, whereas in British English the word spelling is *faecal* with an emphasis on the vowels /æ/ similar to *bat*. Studies emphasizing synchronic language variation have studied the difference in socio-cultural variables among speakers of the same language at the same time and have been used to identify relationships between language change and social change (P. Baker, 2010). William Labov's study, *The Social Stratification of New York City*, showed how variation of macro-sociological categories of socioeconomic class, sex class, ethnicity and age were correlated with language context, specifically words and linguistic variables or parts of words (Labov, 2006). Labov demonstrated direct analysis and systematic comparison of social and linguistic factors can be accomplished with a sociolinguistic variable, or a linguistic variable which co-varies not only with other linguistic elements, but also with a number of extra-linguistic independent variables

such as social class, age, sex, ethnic group, or contextual style (Milroy, 1987, p. 10). Labov was influenced by the work of Antoine Meillet, who argued the sporadic nature of language change can only be explained by the correlations with the social structure of the speech community in which it takes place, “From the fact that language is a social institution, it follows that linguistics is a social science, and the only variable element that we can resort to in accounting for linguistic change is social change,” (Meillet, 1926). Labov extends Meillet’s claims, by emphasizing communities and individuals, stating “community is prior to the individual... language of individuals cannot be understood without knowledge of the community of which they are members” (Labov, 2006, p. 5).. Peter Trudgill building on Labov’s work highlighted how the function of language is to establish social relationships and convey clear and concise information stating “that both of these aspects of linguistic behavior are reflections of the fact that there is a close inter-relationship between language and society, “(Trudgill, 1979, p. 14).

Many studies in sociolinguistics analyze social variation in language by characterizing social groups based on their use or lack of use of linguistic variables. A study conducted by Labov showed a prestige form of specific words (*floor* and *fourth*) varied based on social class, specifically the prestige form was used most frequently by people in higher social class, used the least by those in the lowest class, and over time those in the middle-class usage of the words shifted over time toward the prestige form (Labov, 2006). Labov’s study shows two approaches to understanding language variation, 1) analysis of the linguistic forms (variables) and their distribution, and 2) analysis of the speakers and their behavior (Milroy & Gordon, 2003, p. 8). Other studies have analyzed “frequency profiles” to highlight the differences in tendencies of word usage in relation to social variables such as gender, age, and social group (Rayson & Garside, 1997; Hettel, 2013; Burkette & Kretzschmar, 2018). Variation can also refer to varieties

of language, how speakers use the same language differently, or differences in social and linguistic variables. Gender-based variation in language shows specific differences of language use in the language of men and women from a corpus of 10 million words from the British National Corpus (Schmid, 2003). Another synchronic study of language from the 1960's found American texts had more of a bias towards men than a corpus based on British texts at the same time, evident by more frequent use of masculine words such as *he*, *boy*, and *man* (Leech, Rayson, & Wilson, 2014). Another study compared seven different countries who all communicate using the same language and identified what words were unique to each of the seven different cultures (Oakes & Farrow, 2007).

Diachronic variation studies have been used to characterize and explore the changes in the distribution of words in relation to historical context, or time and events, and how changes in word usage relates to social context like a specific community or group (Bergs & Diewald, 2008; Hilpert & Gries, 2009). For this dissertation, diachronic variation or change is the phenomena where language is used differently depending on one or more contextual variables (P. Baker, 2010). Language change incorporating temporal associations for comparisons of language context can reveal how language context or linguistic variables of a language change over time and how the populations or locations related to a language change over time, or more simply how language context changes are related to social context and how social context changes are related to language context. A diachronic analysis of a corpus made from scripts from television series *Star Trek* found that societal changes in gender roles from 1966 to 1993 were reflected in the language behaviors (female language behaviors characterized as emphasizing a community among speakers and male language behaviors characterized as more concerned with the effective transmission of information) of characters in the television show (Rey, 2014). Another study

over a longer period of time found from 1961 to 2001, the use of male pronouns has decreased while the use of female pronouns has increased in the British English (Sigley & Holmes, 2002). Other diachronic studies have found authors gave more dialogue to characters of their own gender compared to characters of the opposite gender, and revealed American Spoken English is more relaxed and abstract compared to British Spoken English (S. Conrad & Biber, 2009). However, studies on language changes over time are less common and generally more difficult to conduct due to issues related to collecting data over time for communities or individuals. Generally, studies on language change will use the same actors at two different points in time. These approaches are useful and provide direction on how to conduct studies on language variation and change and incorporate links between language and social and historical contexts. However, the systematic study of changes in language and social groups is difficult because of exponential increases in variants and non-normal scaling of variants over time.

The non-linear or non-normal or power law distributions in language are also found in other complex systems and have been identified in other analyses of: cities, incomes, earthquakes, heart rate, forest fires, and the variation in words (Mitchell, 2009). George Kingsley Zipf found an A-curve or power law distribution of word frequencies by showing any word's (*word A*) frequency is approximately inversely related to the rank of *word A*'s frequency in a frequency list of all words, or $1/\text{rank}$ of all words, and this pattern of word frequencies scales as it occurs in single publications, books like *Ulysses* by James Joyce, or in collections of publications (George K. Zipf, 1935). Zipf's law helps explain a specific pattern in language which shows most words in a sample of language have a low frequency of use and only a few words have a high frequency of use. (George Kingsley Zipf, 1946). By plotting all the words from a text or collection of texts on a frequency by rank plot results in creating an A-curve,

which visualizes the variation of all words at a specific time, which words which occur more frequently and less frequently, and how the pattern of word frequency scales across multiple dimensions. Scaling within complex systems provide insight into which variables influence each other to change, specifically how variables scale in relation to other variables over time provides insight into causal mechanisms of the system under study (Barabási & Albert, 1999; Lobo, Bettencourt, Strumsky, & West, 2013). Zipf proposed his law operates because different words have nonidentical meanings and it is easier for people to attach different meanings to an old word rather than come up with a new word (George Kingsley Zipf, 2016). While this interpretation has been debated, confirmation of this phenomena is shown in Figure 4a-b., by plotting and comparing the frequency by rank graph for a single text and for a collection of 14 articles collected on the microbiome. Other studies have used the A-curve to create frequency profiles of words for different groups, and their results confirmed properties of scaling across different language contexts and the ability to characterize different social groups' use of language, how word usage differentiates groups, and the relationship between language use and social context (Kretzschmar, 2015; Burkette & Jr, 2018). Incorporating an understanding of the usage of all words at specific moment in time with insight into how most words are used infrequently and only a few words are used frequently, provides understanding into the range of knowledge of a system, and by repeating this process for multiple time slices the variation of all words can then be compared to the changes in word frequencies across different times. Studies of language incorporating Zipf's law and language as a complex system, have shown how language evolves similar to evolution of species or how the frequency of use determines selection and when new changes enter a system, the changes that are best at surviving and reproducing are used more and eventually win out (Mufwene, 2002, 2009). Thus, by comparing the frequencies of words over

time it is possible to determine which words were selected and survived and which words did not, and because this phenomena scales across dimensions it is possible to analyze both single words and aggregates of words from different social and historical contexts.

Still, complex systems are characterized by both variation and change across multiple scales and levels. To understand both variation and change across different levels (system compared to individual word) a hybrid approach incorporating analysis of all words in the system and a systematic analysis of the changes in statistical frequency of a single word is needed. Such an approach could trace the pattern of a single variant over time and indicate the specific usage and direction of change for concepts. William Kretzschmar hypothesized about an approach by modeling the frequency of single words as a “phase transition” similar to other complex systems, arguing the changes in frequency of a single word over time are indicative of the direction of change for a word (Kretzschmar, 2009). By analyzing multiple A-curves, data can be obtained on the usage of a single variant (concept) over time, which can then be used to plot the trajectory of usage over time for a single variant (Kretzschmar, 2009). Concepts which are adopted exhibit the form of an “S-curve” when plotting the frequency on the Y-axis and time on the X-axis, while other concepts which are rejected, static, or subject to problematic adoption also exhibit specific curves as shown in Figure 5a-. Similar S-curves are observed when modeling density of contact in networks, spread of variants within a population, or diffusion of within a social system (Kretzschmar, 2009; Labov, 1994; Rogers, 2010). This approach analyzes both the variation of all words at a single time and the direction of change of single concepts over time. Assuming, conceptual change is a proxy for knowledge changes, this approach can be used to compare new and old concepts, track the changes in usage of concepts, measure the range of knowledge of a concept, and the characterize the change of concepts as adoption or

rejection. Preliminary analysis of two words related to the microbiome concept show differences in usage and highlight the potential use of this approach to understand the adoption and rejection of different concepts within a collection of articles on the word microbiome, displayed in Figure 5a-d. Thus, by understanding the aggregate and individual pattern of words over time and across different scales we can provide analysis into concepts and account for variation and change in language and by proxy understand how knowledge changed.

Emphasizing the link between language change and biological change highlights how complexity science provides direction into the analysis of complex systems that consist of different components but similar properties like: multiple components interacting at the same time, multiple variants co-existing, and changes over time (Burkette & Jr, 2018). A study using A-curves and S-curves, could show both the variation of words and the direction of change of a single word in the context of language, this analysis could be done across different scales, and identify which words/concepts were adopted or rejected by a system over time.

Yet, as previously mentioned word usage is not the only thing pushing conceptual change. Through contact and interaction, language as a complex system is influenced by social and historical factors driving variation and change (Labov, 1972; Milroy & Gordon, 2003; Wardhaugh, 2009). Change in language can be caused by interactions of agents within the system including: individuals, groups, material artifacts, words, concepts, and knowledge, and these agents can change spontaneously or over time, and any agent can use concepts differently over time and can be influenced by other agents use of concepts (Bybee, 2007; Kretzschmar Jr, 2010). Bruno Latour and co-authors, articulated this point, and argued knowledge produced in laboratories that were scientific facts, are iteratively produced as the results of material, experimental practices, biased observations, contingent interpretations, and based on established

styles of thought (Knorr-Cetina & Mulkay, 1983). Therefore, determining conceptual change requires understanding variation and changes across multiple dimensions, and knowledge on concepts is dependent on language context and other contexts, specifically who is using and interpreting them, when, and for what reasons (Cetina, 2009; Fujimura, 1996). A framework is needed which allows for the analysis of actors across different systems and integrates contextual factors to measure changes in knowledge.

Knowledge and Context

Drawing on a framework based on networks and contexts developed by Jurgen Renn and Manfred Laubichler in *Extended Evolution and the History of Knowledge*, I argue the analysis of the history of knowledge requires a perspective of extended evolution. “Knowledge” is the encoded experience of actors and is a mental structure with material and social dimensions that determines which actions are possible in a historical situation, knowledge may be shared within a group or society via the use of material artifacts such as instruments or texts, and the social and material dimension of knowledge are important for understanding the transmission of knowledge from one generation to the next (Renn & Laubichler, 2017a). Institutions according to this framework, represent encoded collective experience and are a means of reproducing the social relations existing within a given group or society, which results in shared behaviors connected by cognitive, social and material links (M. Laubichler & Renn, 2015; Renn & Laubichler, 2017a). Renn and Laubichler using historical examples from biology, culture, history, technology, and language, showed the origin of variation was a result of the properties of complex systems: actors, actions, and their results; context is material means used by an actor to reach an action and the result of actions, contexts of action represent knowledge and institutions which can be used to share, transmit, and transform regulative structures; networks of human

actions include a material and social culture, and knowledge itself may be externally represented and shareable (Renn & Laubichler, 2017a). This framework provides the flexibility to model material artifacts, individuals, groups, institutions, and systems as actors. I experimentally tested this framework with a concept currently the subject of increasing scientific debate, the microbiome. The debates on the microbiome center around the knowledge on the microbiome, specifically what is the microbiome, is the microbiome a new concept or a rebranded concept and is there a distinction based on context, i.e. multiple interpretations? When did the microbiome emerge in the scientific literature? Where and how has the microbiome been used and in what contexts? What contextual factors have influenced knowledge on the microbiome?

Previous attempts to understand what has driven the use and meaning of the microbiome have focused on small changes based on qualitative interpretations of historical context, language context, or social context. A few studies have combined explanatory variables or factors from these different contexts for their arguments, but most studies emphasizing the historical context of the microbiome focus on the history of microbiology and on important dates of when the microbiome was used in text (Eisen, 2015; Prescott, 2017). Generally, the studies emphasizing the language context of the microbiome stay away from the historical context and focus on definitions, either claiming that there is no definition of the microbiome or proposing a new definition and in some cases entirely new vocabularies (Ursell et al., 2012; Blaser et al., 2013; Marchesi & Ravel, 2015a). The studies that emphasizing the social context of the microbiome focus on the specific domain of interpretation for the microbiome and the social processes related to the use of the microbiome (Shade & Handelsman, 2012; Schneider & Winslow, 2014). Admittedly, there is some overlap across the different contexts, but currently no

study has leveraged the different contexts systematically to measure and analyze the contextual factors influencing changes to scientific knowledge on the microbiome.

This framework provides an approach to use contextual factors as explanatory variables to characterize specific changes to the microbiome. By identifying the contextual variation in the microbiome over time, the specific changes to the microbiome which were adopted and were diffused can be identified. Specific changes to the usage and meaning of the microbiome are dependent on social context. By tracking the development of these changes in the historical and language context, or words, phrases, and meaning of the microbiome used over time, a determination on which social contexts were influencing the microbiome can be made. Similarly, changes in the social context of who uses the microbiome are performed by specific actors who in some cases have compatible and other cases incompatible knowledge with the microbiome. By analyzing the contextual factors of the microbiome, many of the debates surrounding the evolution and use of the microbiome including: the meaning of the microbiome, who/what influenced changes to the microbiome, and how knowledge on the microbiome transformed, can be settled.

Integrating this framework with methods and approaches to understanding knowledge variation and change, requires detailed understanding of language variation and change. Words or concepts can be used to define knowledge of language and highlight variation and change based on the contextual use of individual words and how words co-occur together (Hettel, 2013; Stubbs, 1995). As previously mentioned, methods from Corpus Linguistics and NLP can help with the understanding knowledge of concepts at different scales. NLP provides insight into the relationship between a concept and the totality of knowledge within a discourse through characterization and classification, whereas Corpus Linguistics highlights the relationship

knowledge of language, knowledge of individual words, knowledge of predictable combinations of words, and cultural knowledge of combinations of words through analyses of individual words (Stubbs, 2001). Novel methods such as lexical profiles and collocate networks, not typically used within biomedicine, have also shown potential in summarizing and displaying information in a clear and systematic manner, so as to facilitate comparison and the discovery of knowledge patterns and trajectories.

Lexical profiles, provide a coherent and interpretable summary of node words and collocates. Lexical profiles have been used to create systematic summaries of observed meanings for a word within a corpus, interpretations linking contextual factors, and differences of meaning based on knowledge (Hettel, 2013; Stubbs, 2001). In this example from Michael Stubbs, he creates a lexical profile which reveals the context, sense, and cultural knowledge of the word *chopped*, and the words in close proximity of the word *chopped* using the 20 most frequent collocates of *chopped*: *finely, fresh, parsley, onion, garlic, tbsp., tomatoes, oz, peeled, add, off, onion(s), salt, pepper, chives, herbs, tablespoons, dried, small, tsp*, and less often with collocates like: *off, up, and down* (Stubbs, 2001, p. 95). Using the node word and the collocates, these results show how the meaning of *chopped* can generally be described as a part of a recipe, and the when the words *chopped* and *add* both co-occur that the probability of the text being a recipe is near 100 percent (Stubbs, 2001, p. 95). Using Stubbs notation, a lexical profile can be shown simply by:

- *node word* (number of occurrences of node word in corpus): < collocate₁, collocate₂, ... collocate_n >

The lexical profile then for the example from above would be:

- *chopped*(n)< *finely, fresh, parsley, onion, garlic, tbsp., tomatoes, oz, peeled, add, off, onion(s), salt, pepper, chives, herbs, tablespoons, dried, small, tsp*>

The node word and the number of occurrences of a node word in the corpus are followed by a colon (:), then the commercial at (<) signifies the set of elements or collocate data, followed by the first collocate of interest, and a comma separating each subsequent collocate of interest, and then ending with a commercial at in the opposite direction signifying the end of the set (>). The lexical profile notation can also include the percentage of the occurrence of each collocate with the node word:

- *node word* (number of occurrences of node word in corpus): < collocate₁ (percentage of occurrence of collocate₁ with node word), collocate₂ (percentage of occurrence of collocate₁ with node word), ... collocate_n (percentage of occurrence of collocate_n with node word) >

The lexical profile can include collocates of collocates based on the researcher's preference:

- *node word* (number of occurrences of node word in corpus): < collocate₁ (percentage of occurrence of collocate₁ with node word) <collocate of collocate₁ (percentage of occurrence of collocate of collocate₁ with collocate 1) >>

In including collocates of collocates, the (<) is used to separate the collocate of the collocate, and again this allows the researcher to add as many collocates of collocates to the lexical profiles.

This method has also been used to compare collocates across different social and temporal contexts, being represented as:

$$node\ word_{ct}(n): <w_4, w_5, \dots, w_{n+1}>; \text{ where } c = \text{context}, t = \text{time}$$

Networks derived from combinations of the content within material artifacts, the material artifacts themselves, or the metadata associated with the material artifacts were previously described as ways to map knowledge and provide insight into social context and structures.

Collocates and collocational analyses have been reviewed as a method to understand the patterns in co-occurring words. Building on this method, collocational networks emphasize the distance, frequency, and exclusivity of collocates (Brezina, McEnery, & Wattam, 2015). Within language,

collocates are the actors in a language context which communicate the meaning and semantic structure of a text or corpus. The criteria for identifying collocates are analogous the criteria used for data collection used to create social networks, citation networks, or knowledge maps. Criterion for collocates and other networks include: distance which specifies which direct neighbors are to be included as part of the analysis, frequency of interaction or the frequency of co-occurrence between actors, and directionality or the dyadic link between actors describing asymmetrical or symmetrical relationships. By implementing these criteria the end result is a set of actors and events, and the events which detail the interaction and relationship between actors can then be visualized as a complex network of relationships (Williams, 1998). Previous studies using collocational networks have shown within discourses, or language in use, there are specific lexical patterns that can be visualized as networks representing the knowledge structure of language (G. Brown, Gillian, & Yule, 1983; Stubbs, 1983; Brezina et al., 2015). Thus, the framework by Renn and Laubichler combined with the methods mentioned in this chapter provide novel approaches to measure and understand changes in knowledge.

Preliminary analyses were completed to measure if changes in knowledge on the microbiome could be discovered based on context using a combination of methods and the framework of the history of knowledge. The first preliminary analysis completed was a concordance analyses, to discover the variation and range of words used with the word microbiome, from a corpus created from microbiome publications from 1900 to 2014. Concordance analyses show the relationship between words of interest and how the words co-occurring with words of interest change over time. The concordance analysis for each word compared collocates of the word of interest, e.g. microbiome, and words co-occurring within 4 words in front of or behind the word of interest after stop words were removed. This is

significant as the probability of two words occurring with each other is statistically extremely low (Stubbs, 2001). The criteria determining which collocates to include collocates were: (1) appeared within four words of word of interest (2) not the word combination Human Microbiome Project, (3) Mutual information Score (MI) of equal to or greater than three which indicated a strong association with the word of interest with the collocate, and (4) occurred at least 5 times within all publications. The results from the concordance analyses were used to create collocate networks. The networks were visualized as a force-directed network graph, which colored collocates based on their score according to the Newman Grouping algorithm using the ORA software package (version v3.0.9.9.33). The rapid diversification of the discourse around the microbiome concept was seen when comparing all collocates of microbiome as identified with MI Score from 2007 and 2014. The total number of collocates from the words of interest, *microbiome*, *gastrointestinal*, *ecology*, *science*, *biology*, *genomics*, and *microbiology* increased from 180 (2007) to 6352 (2014), as seen in Figure 6a-b. These graphs highlighted the possible actions or word combinations of the microbiome concept with other words of interest in 2007 and 2014 and displayed a dynamic network of microbiome knowledge across time at a finer scale than publications, citations, or static representations of concepts.

A publication to co-authorship network was created from metadata associated with the microbiome articles from 1900 to 2014. The publication network visualizes the transmission of knowledge between material artifacts and actors. The co-authorship network consisted of 7257 (out of approximately ten thousand authors) connected by publications as seen in Figure 7., highlighting the scientific and scholarly network or intellectual ties within the microbiome (J. Scott, 2017). The network of intellectual ties was created to show how people are connected by specific knowledge or what they know contained in publications specific to the microbiome.

Previous work on intellectual ties has shown=mutually relevant knowledge is the most important characteristic within scientific communities (Moody, 2004; Rodriguez & Pepe, 2008; J. Scott & Carrington, 2011). These figures demonstrated how microbiome knowledge changed specific to context and contextual factors and how it was possible to link multiple proxies of knowledge on the microbiome using Renn and Laubichler's framework.

These preliminary analyses have demonstrated the variation and change to microbiome knowledge and how differences in context must be considered when analyzing microbiome knowledge. Also, these results show how it is possible to measure and track knowledge across multiple scales and dimensions via specific variants. The first chapter of this dissertation was the result of the in depth literature review conducted to understand how language, social, and language context has been used to interpret and understand the microbiome. The results of the review on the microbiome and this chapter helped guide collection of publications and data for this study and informed how to analyze microbiome knowledge.

Summary

As something changes, what evidence do we have that something changes? Publications, citations, and concepts can be evaluated for trends and patterns, which may point to explanatory variables or contextual factors. However, what we observe in the historical evolution of a scientific field or scientific knowledge, is these changes are related to changes in knowledge, but what is not understood is how to measure and track changes in knowledge. To understand how knowledge changes and evolves, requires analysis beyond papers, citations, and concepts. It requires a framework integrating the dynamics of language, social, and historical systems and their interactions across different scales. An analysis of contextual factors analyzes both the changes in language context as a proxy to understand the changes in scientific knowledge, and

how things change in particular social context as carried out by specific individuals and institutions. Historical transformations captured by events and changes in social networks have feedbacks to scientific knowledge (Rivera, Soderstrom, & Uzzi, 2010; Shi, Foster, & Evans, 2015). Therefore, to understand how contextual factors influence scientific knowledge specific questions need to be asked in order to determine if and how knowledge has changed. The framework adopted by this dissertation engages new questions concerning the relation between biomedicine and the different contexts intersecting the microbiome, which implies new dynamics between local and global worlds of knowledge and context (Burri & Dumit, 2007). With the intention of understanding the knowledge variation and change to the microbiome, a systematic analysis of the factors contributing to the variation and change of the microbiome is needed.

Figures and Tables

Microbiota MeSH Descriptor Data 2019			
Details	Qualifiers	MeSH Tree Structures	Concepts
MeSH Heading	Microbiota		
Tree Number(s)	G06.591 G16.500.275.157.049.100.500 N06.230.124.049.100.500		
Unique ID	D064307		
Annotation	coordinate with specific organism, organ or other site with / microbial if pertinent		
Scope Note	The full collection of microbes (bacteria, fungi, virus, etc.) that naturally exist within a particular biological niche such as an organism, soil, a body of water, etc.		
Entry Term(s)	Human Microbiome Microbial Community Microbial Community Composition Microbial Community Structure Microbiome Microbiome, Human		
See Also	Metagenome		
Public MeSH Note	2014; see METAGENOME 2009-2013; see MICROBIOME 2008-2009		
History Note	2014; use METAGENOME 2008-2013		
Date Established	2014/01/01		
Date of Entry	2013/07/08		
Revision Date	2018/02/28		

page delivered in 0.102s

Copyright , Privacy , Accessibility , Site Map , Viewers and Players
U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
[National Institutes of Health, Health & Human Services, Freedom of Information Act](#)




Figure 1. Results for “microbiome” in MeSH.

Microbiota MeSH Descriptor Data 2019

Details Qualifiers MeSH Tree Structures Concepts

Microbiota Preferred		Collapse All
Concept UI	M0580963	
Scope Note	The full collection of microbes (bacteria, fungi, virus, etc.) that naturally exist within a particular biological niche such as an organism, soil, a body of water, etc.	
Terms	Microbiota Preferred Term	
Term UI	T726055	
Date	09/18/2008	
LexicalTag	NON	
ThesaurusID	NLM (2010)	
	Microbial Community	
Term UI	T000939177	
Date	01/16/2018	
LexicalTag	NON	
ThesaurusID	NLM (2019)	
	Microbial Community Composition	
Term UI	T000939883	
Date	01/30/2018	
LexicalTag	NON	
ThesaurusID	NLM (2019)	
Microbial Community Structure Related		
Concept UI	M000640513	
Terms	Microbial Community Structure Preferred Term	
Term UI	T000939178	
Date	01/16/2018	
LexicalTag	NON	
ThesaurusID	NLM (2019)	
Microbiome Related		
Concept UI	M0508691	
Scope Note	The full collection of microbes (bacteria, fungi, virus, etc.) that naturally exist within a particular biological niche as identified by the presence of their genomic sequence regardless of whether or not they can be cultured.	
Terms	Microbiome Preferred Term	
Term UI	T695479	
Date	04/13/2007	
LexicalTag	NON	
ThesaurusID	NLM (2008)	
Human Microbiome Narrower		
Concept UI	M0508728	
Scope Note	The full collection of microbes (bacteria, fungi, virus, etc.) that naturally exist within the human body as identified by their genomic sequence regardless of whether or not they can be cultured.	
Terms	Human Microbiome Preferred Term	

Figure 2. MeSH descriptor data 2019 for microbiome.

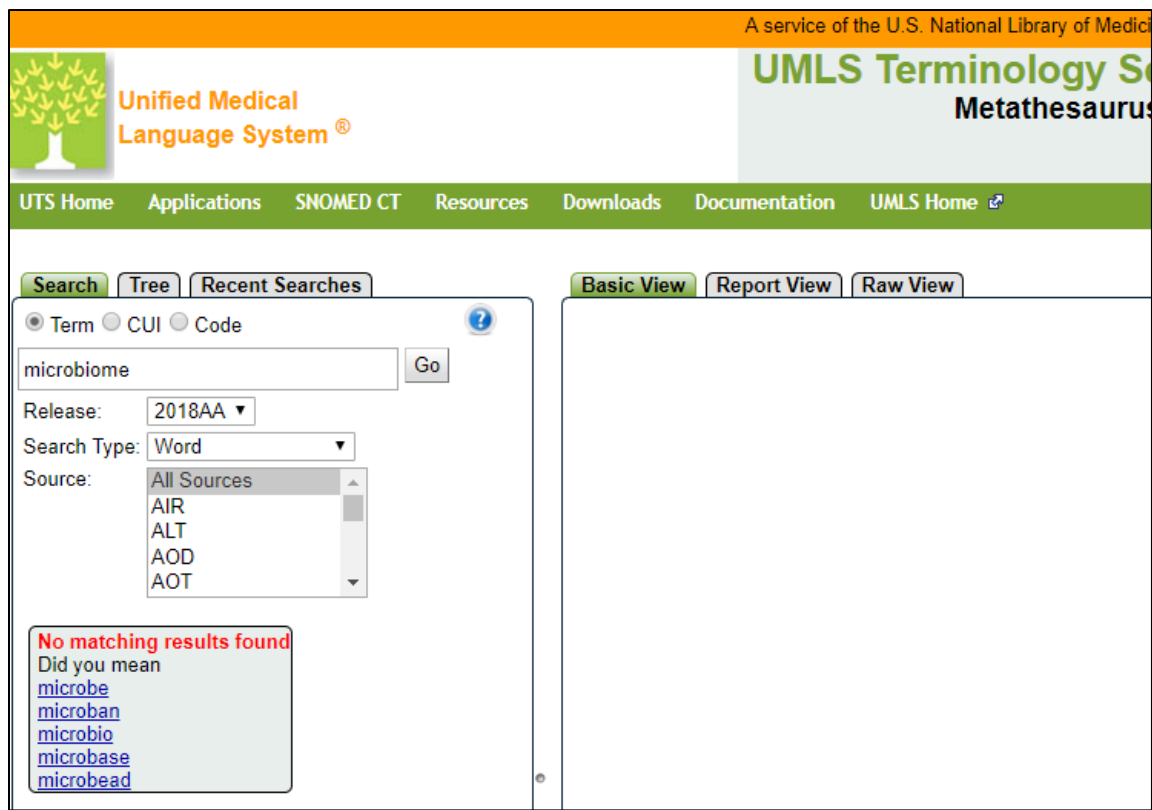


Figure 3. UMLS search result for microbiome.

Table 1. Association measures for words from (M. Scott, 2018).

Mutual Information
Log to base 2 of (A divided by (B times C))
A = joint frequency divided by total tokens
B = frequency of word 1 divided by total tokens
C = frequency of word 2 divided by total tokens
MI3
Log to base 2 of ((J cubed) times E divided by B)
J = joint frequency
F1 = frequency of word 1
F2 = frequency of word 2
E = J + (total tokens-F1) + (total tokens-F2) + (total tokens-F1-F2)
B = (J + (total tokens-F1)) times (J + (total tokens-F2))
T Score
(J - ((F1 times F2) divided by total tokens)) divided by (square root of (J))
J = joint frequency
F1 = frequency of word 1
F2 = frequency of word 2
Z Score
(J - E) divided by the square root of (E times (1-P))
J = joint frequency
S = collocational span
F1 = frequency of word 1
F2 = frequency of word 2
P = F2 divided by (total tokens - F1)
E = P times F1 times S
Log Likelihood
$2 * (a \ln a + b \ln b + c \ln c + d \ln d$
$- (a+b) \ln (a+b)$
$- (a+c) \ln (a+c)$
$- (b+d) \ln (b+d)$
$- (c+d) \ln (c+d)$
$+ (a+b+c+d) \ln (a+b+c+d)$
a = joint frequency
b = frequency of word 1 - a
c = frequency of word 2 - a
d := frequency of pairs involving neither word 1 nor word 2
and "Ln" means Natural Logarithm

Table 2. Top 20 collocates of microbiome in MB Corpus by different association measures, colored words are collocates by more than one measure.

Rank	MI3	Log-likelihood	T-score	Z-score
1	gut	gut	gut	gut
2	human	human	human	human
3	core	core	intestinal	core
4	project	oral	oral	project
5	oral	project	analysis	oral
6	intestinal	intestinal	host	viewed
7	skin	skin	core	vaginal
8	vaginal	host	changes	lung
9	lung	changes	project	skin
10	changes	healthy	healthy	lean
11	healthy	vaginal	disease	infant
12	host	lung	studies	consortium
13	analysis	analysis	skin	healthy
14	infant	studies	diversity	metabolome
15	obesity	research	research	capacity
16	research	obesity	role	changes
17	studies	role	data	alters
18	role	diversity	vaginal	distal
19	diversity	disease	lung	intestinal
20	infant	infant	obesity	jumpstart

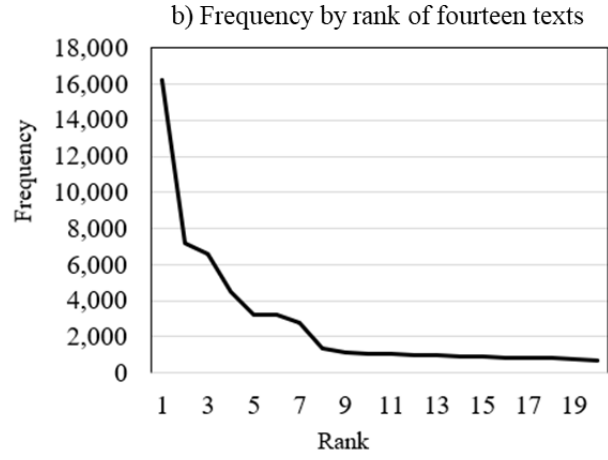
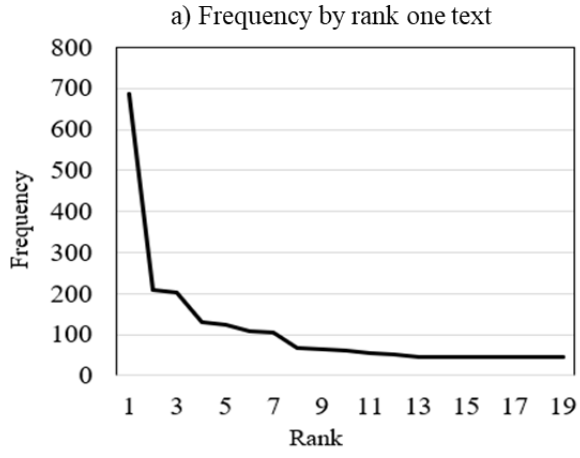


Figure 4a-b. Frequency by rank of one text and fourteen texts. 4a. Frequency by rank one text.

4b. Frequency by rank of fourteen texts

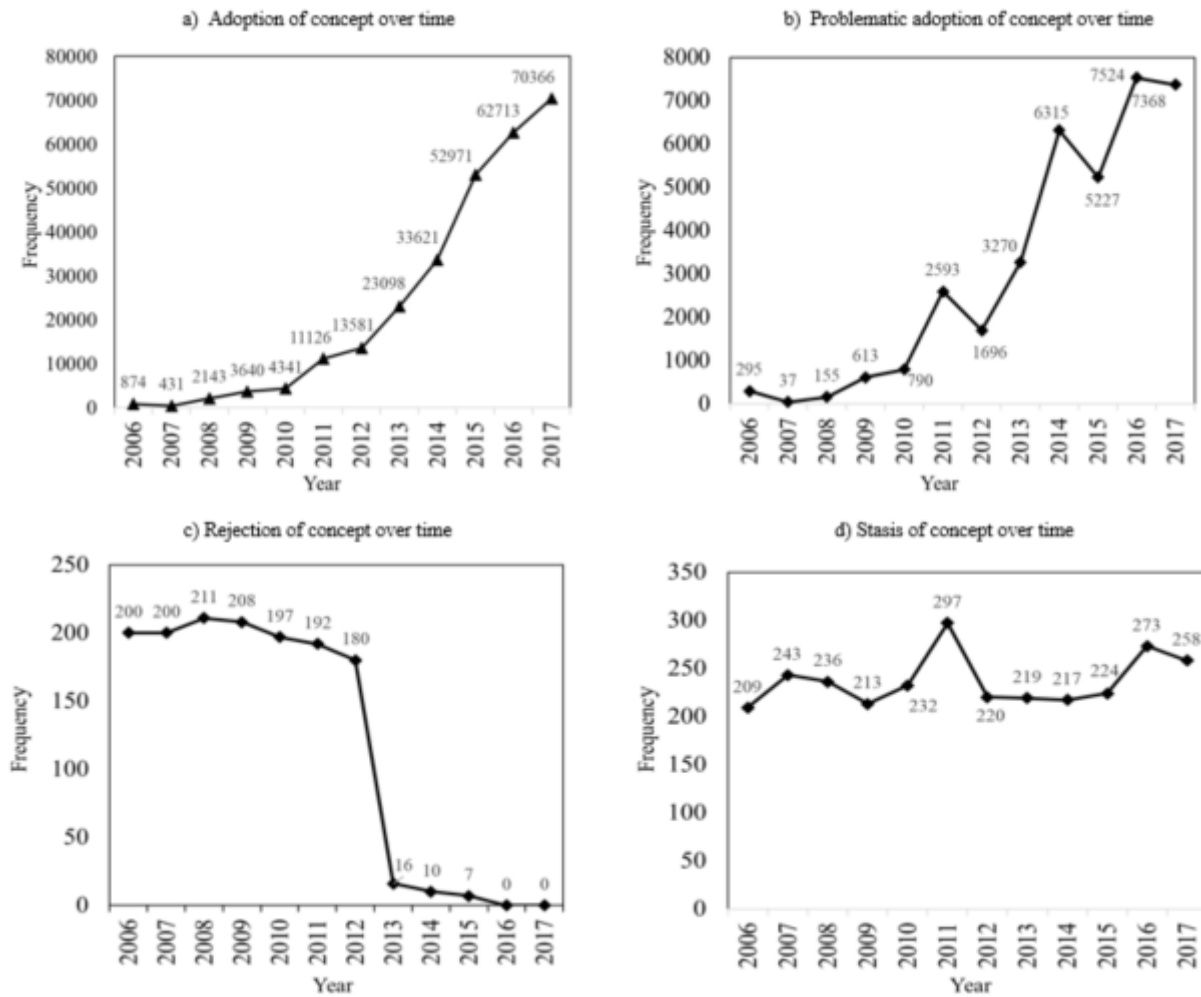
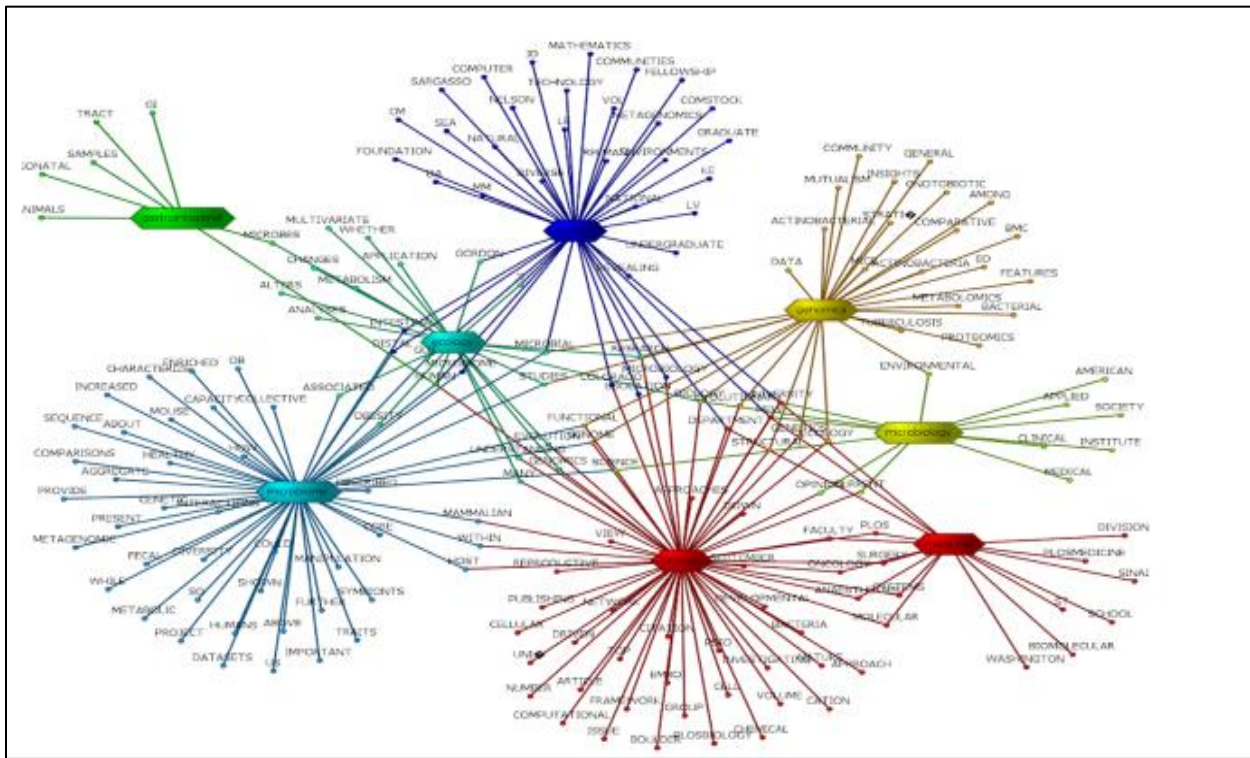


Figure 5a-d. Examples of S-Curves over time. 5a. Adoption of concept over time. 5b. Problematic adoption over time. 5c. Rejection over time. 5d. Stasis of concept over time.

a) Collocates of microbiome 2007



b) Collocates of microbiome 2014

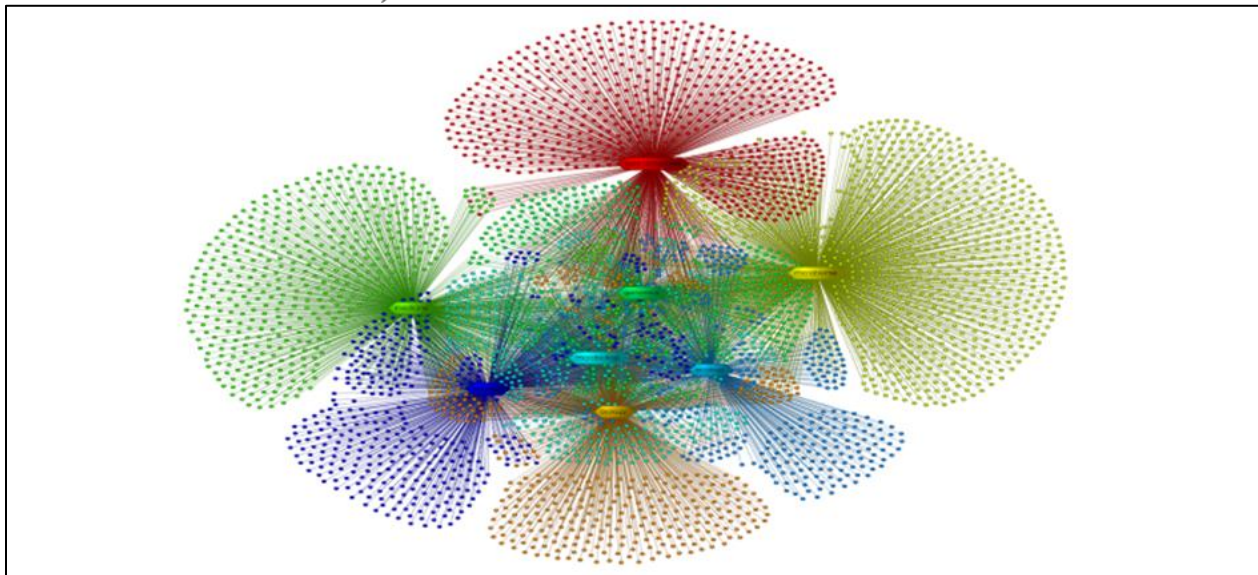


Figure 6a-b. Collocate Network for MB Corpus. 6a. Collocates of microbiome 2007. 6b.

Collocates of microbiome 2014.

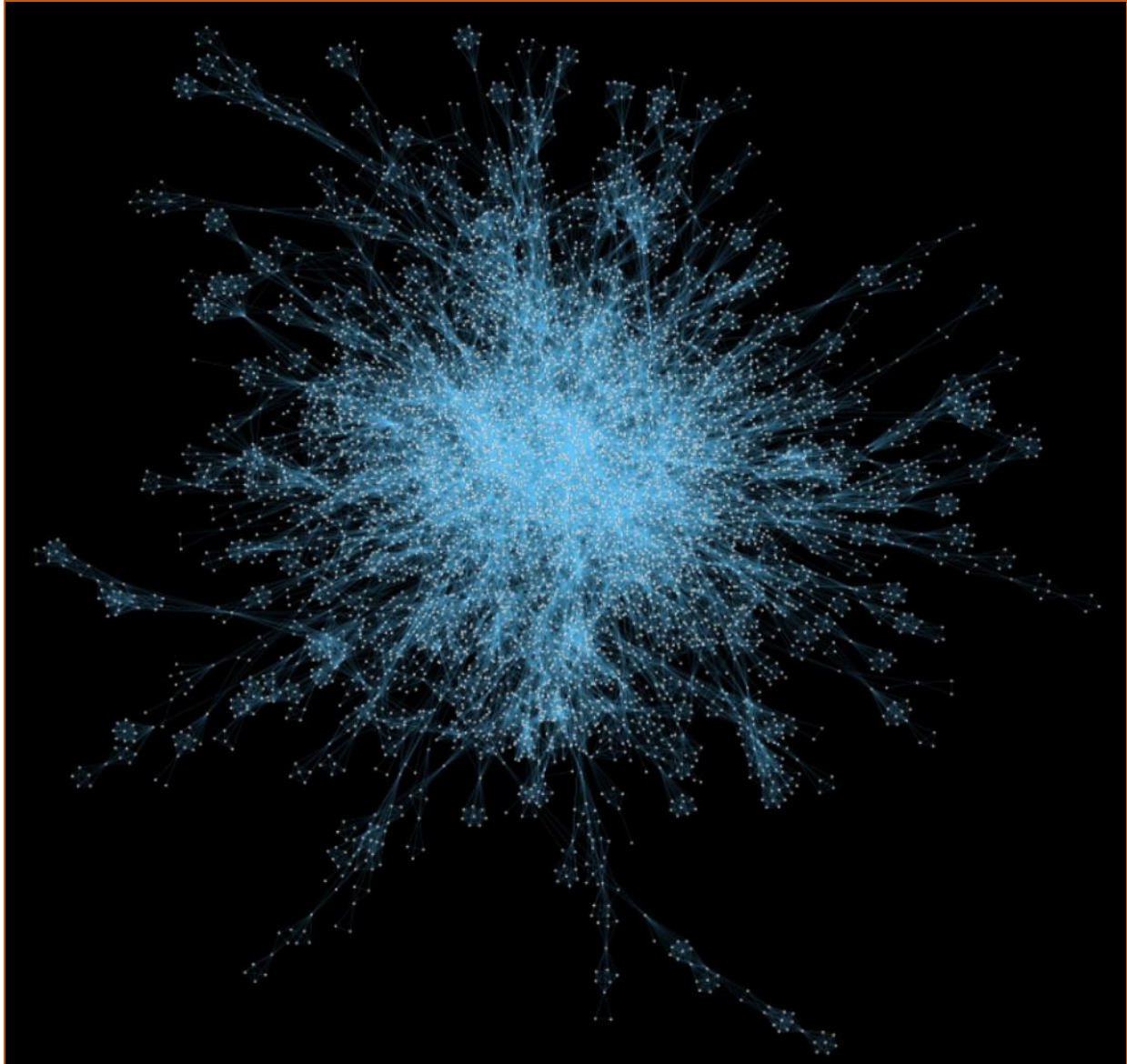


Figure 7. Author to publication network of microbiome publications 1900 to 2014.

BIG DATA AND DATA DRIVEN SCIENCE AND RESEARCH

Background

As a result of the big data, there is an amalgamation of methods from distinct cultures and methodologies with an emphasis on computation or digital, as seen in computational social science, computational and digital history and philosophy of science, and digital humanities (Guiliano et al., 2011; M. D. Laubichler, Maienschein, & Renn, 2013; English & Underwood, 2016). Researchers with from different backgrounds are trading and collaborating with each other to share tools, information, and knowledge on how to analyze and understand big data (M. E. J. Newman, 2001; Ankeny & Leonelli, 2016). Some speculate big data will produce important theoretical changes for the social sciences and for sociology in particular, aligning with Bruno Latour when he said, “change the instruments and you change the entire social theory that goes with them” (Latour, 2009, p. 156). However, some are unsure if these changes will produce any revolutionary changes in paradigms, or if these changes will be similar to other cases where a field and its traditions are subverted to other fields as in the example of social networks and network science (Fleck & Kuhn, 1981; J. Scott, 2017). Others speculate big data will push the humanities and social sciences into Pasteur’s quadrant of actionable knowledge (Underwood, 2017). To better understand big data, this chapter will explore aspects of the historical, social, and language context of big data, giving specific attention to the challenges of integrating big data and DDSR methods into experiments, how big data can be used to help interpret high dimensional data, and present a case for the necessity of big data to measure and track changes in scientific knowledge such as the microbiome concept.

Context

Big data is powerful and useful and the drastic shift in the amount of available information has made using big data and data driven science critical to all scientific disciplines. Generally, big data refers to the size or amount of data. In industry, big data is central to the operations of the largest companies in the world like Google, Amazon, and Facebook (Stephens et al., 2015). In popular media, big data has become the subject of various news articles, radio, and television programs (Labrinidis & Jagadish, 2012). In government, there are big data ‘plans’ and big data ‘working groups’, e.g. the Big Data Research and Development Strategic Plan and the Big Data Interagency Working Group, that are going to scale up big data projects in the United States (“Big Data - NITRDGROUPS,” n.d.). The challenges in translating knowledge from big data within biomedicine has been recognized by the Big Data to Knowledge or BD2K initiative by the National Institutes of Health (Lazer et al., 2009). What unites these different areas is how big data provides an unprecedented historical context which leads to novel questions, experiments, and results previous generations could not have imagined.

Big data makes it possible to do things not possible before such as identify business trends, personalize medicine at the level of the genome, fight wars using remote drones, and measure changes in scientific knowledge (Lazer et al., 2014). Big data consists of a combination of images, videos, sound recordings, and texts. Sometimes this data is messy, unorganized, and unstructured data or in other instances the data can be pristine, formatted, and accessible structured data. While data has been present throughout history and in different contexts, what makes the current situation with big data different from periods is the form and function of the data and the potential results and the ease of access to high end computational capacity. One of the most well-known and successful results from big data was sequencing of the first human

genome via the Human Genome Project (HGP) (Yang, 2013). Following the success of the HGP, a wave of large scale big data projects emerged following the success of the HGP including: CERN's large Hadron Collider (LHC), the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), and the National Oceanic and Atmospheric Administration (NOAA) Big Data Project (Choudhury, Fishman, McGowan, & Juengst, 2014; Natarajan & Philipoff, 2018). Like the HGP, all of these large international or cross-laboratory data driven projects represent the largest scale and scope of big data projects, as they generate petabytes of data per day, bring together massive teams and resources, generate multiple scientific discoveries, and form new lines of economic growth (Hey, Tansley, & Tolle, 2009). Recently, big data and big data technologies have been increasingly used for biomedical insight into biomedical collections, organisms, and knowledge (J. Luo, Wu, Gopukumar, & Zhao, 2016; Z. Luo et al., 2010; Moerchen, Fradkin, Dejori, & Wachmann, 2008).

Most big data projects take advantage of recent increases in computing power via multiple cores, multiple levels of memory, distributed data centers, and cloud computing to analyze and test data. However, recent advances in computational power show the progress made in computational speed and power. Initially the HGP took over 13 years and three billion dollars to sequence the first human genome, whereas with advances in data analytics and computation have decreased the time and cost to only a few days and around thousand dollars to conduct the same analysis (Ceri, 2018). Easily accessible and cheap cloud computing services leverages the power of multiple computers and can use thousands of computers to answer a single specific search query in less than 1 second (Marr, 2015). These increases in power and decreases in cost to integrate high end computational architecture have made the processing of bigger data sets faster and easier to implement (Guo, Ning, Hou, Hu, & Guo, 2018; Liang & Liu, 2018). In the

wake of these advances, the US government created the Big Data Research and Development Strategic Plan and the Big Data Interagency Working Group to scale up big data projects in the United States (“Big Data - NITRDGROUPS,” n.d.). Additionally, the National Institutes of Health (NIH) and the National Science Foundation (NSF) have grant openings for over \$250 million to projects using big data (Margolis et al., 2014). Still, even with increased resources to implement big data projects, more availability and access to large datasets, and diminished technological hurdles many scientists are still not taking advantage of big data (Callebaut, 2012)

Historical context of big data

Some criticize big data is not a new phenomenon but a recurring theme within science and society. These critics point to other historical examples of the ever-increasing amounts of information and the problems information overload causes. One of the staunchest advocates of this position is Ann Blair, she has pointed out society has dealt with information overload or data-deluges consistently through history, from Lucius Annaeus Seneca the 1st century philosopher “ the abundance of books is a distraction;” to Gutenberg’s printing press, suddenly, “there were far more books than any single person could master, and no end in sight,” to Erasmus the 16th century poet and priest, “ [printers] fill the world with pamphlets and books that are foolish, ignorant, malignant, libelous, mad, impious, and subversive, and such is the flood that even things that might have done some good lose all their goodness... Is there anywhere on earth exempt from these swarms of new books?” (A. Blair, 2003; A. M. Blair, 2011; “Information Overload Is Not Unique To Digital Age,” n.d.). Others have cited the foretelling of the challenges of data within science:

“The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze... The investigator is staggered

by the findings and the conclusions of thousands of other workers- conclusions which he cannot find time to grasp, much less remember.” (Bush, 1996, p. 37).

Fremont Rider, a Wesleyan University Librarian, described the problem of volume of data within libraries in 1944, as he estimated that libraries in Universities were doubling in size every sixteen years (starting in 1944), and that if this growth rate continues the Yale library collection would be around 200,000,000 volumes, that would require a library staff of over six thousand people, (“A Very Short History Of Big Data,” n.d.; Molyneux, 1994).

However, the current situation is different based on: how much new data is being created, the different types of data, the new sources of data, the capacity to use data, and the high dimensionality and scales of information the data provides. A 2016 report from IBM highlighted that 90% of the data in the world (in 2016) had been created from 2014 to 2016, this translated into 2.5 quintillion bytes of data a day, with the global total of data estimated to be around 2.7 zettabytes (“10 Key Marketing Trends for 2017,” n.d.). While, information overload has occurred over time, these numbers point to a different pace and volume in output of data. Also, drawing on extensive engagement with the literature, the current era of big data is different from other historical contexts because the current era of big data is: huge in volume, created in or near real time, diverse in structure, massive in scope (entire populations), fine grained and indexical, relational with common fields across different data sets, flexible with new fields added easily that increase the size of the data, and scalable with size expanding rapidly (Boyd & Crawford, 2012; Kitchin, 2014). To many this data represents opportunities ushering in a “data revolution” or an “era of big data.” This era of big data has created a premium on big data and has scientists, organizations, and the public looking for access and ways to use big data for data driven projects

(Breddels & Veljanoski, 2018). Yet, many of the same reasons explaining how the current historical context with big data is unique also highlight the complexity of big data.

Complexity of big data

Many emphasize the volume or amount of data and call it “big data.” Some of the most well-known projects utilizing big data or big data projects are/were large international or cross-laboratory projects using enormous volumes of data like: the Human Genome Project, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), and the National Oceanic and Atmospheric Administration (NOAA) Big Data Project (Ahmad, Testani, & Desai, 2016; Calvard, 2016). These projects constitute the largest scale and scope of big data projects, as they generate petabytes of data per day, bring together massive teams and resources, generate multiple scientific discoveries, and are the basis for new lines of economic growth (Hey et al., 2009). These projects show how big data relates to the volume or the amount of data now available, but as previously mentioned big data also includes four other characteristics.

The five characteristics of big data or ‘the five V’s’ of big data are: 1) the *volume* or size of the data, 2) the *variety* or the lack of harmonization of the data from different sources, formats, and structures, 3) the *velocity* or the rate at which data and results are being created, 4) the *veracity* of the data or uncertainty of the data quality, and 5) the *value* or insight that can be gained from the data (Gudivada, Rao, & Raghavan, 2015). Each of the five V’s has a set of inherent challenges which are directly related to the others. Curated digitized relational databases are a commonly used big data source and a brief description of an analysis using data from these databases will highlight how the five V’s are related.

Large curated relational databases have been around since the 1960’s, but only recently have these databases provided open access to their data (Boyd & Crawford, 2012). These

databases come from many different sectors and include: The Library of Congress, Amazon, YouTube, Google, the National Energy Research Scientific Computing Center (NERSC), and the National Library of Medicine (NLM) (Hitzler & Janowicz, 2013; Marr, 2015). In most cases the volume of these databases ranges from terabytes to zettabytes. The variety of data ranges from multimodal unstructured to semi-structured data which includes: text, video, audio, and metadata. Additionally, the velocity or amount of data generated by the day (or in some cases by the minute) is larger than the aggregate of the previous data stored and/or analyzed. The veracity of data from missing values or noise cannot easily be transformed and often large portions of data are unusable or alters the entire distribution of data. The value of the data depends on the usability and results from data, in some cases to update previous analysis and in others to validate new directions. Yet, while the five V's of big data highlight some of the challenges when working with big, the potential for big data and the emergence of publicly accessible databases has more scientists and looking for approaches to manage big data (Zhou et al., 2014).

To some the difficulty in incorporating big data into their projects is lack of the technical and computational skills to include a big data service or technology like: cloud-based technology, Structured Query Language (SQL), distributed storage platforms (i.e. Hadoop, Cloudera, and MongoDB), machine learning methods, or the resources to create a large project. Yet, while many of these platforms and software are used in some big data projects, they are not necessary to incorporate big data into a project or experiment. Previous studies utilizing big data have used a range of highly technical or computational tools and methods, while some projects have used simpler tools and methods such as pens, papers, completed analyses manually.

Analyzing data in multiple dimensions sometimes requires understanding how multiple systems interact and lack of this understanding turns people away from big data analytics. It is difficult to analyze differences common in big data, such as heterogeneous data types (e.g. audio, text, measurement, and metadata), possible inherent semantic associations, and the multi-dimensional relationships linking different features of the data together, especially when dealing with biomedical data (Ahmad et al., 2016; Labrinidis & Jagadish, 2012). This has resulted in some other projects utilizing big data for studies at smaller scales and with narrower scopes. Some data-driven projects are conducted in a single laboratory, on individual observations, or based on a single subject (M. Swan, 2013). Many of these projects have emphasized how big data can be used to reproduce traditional results, and create new interdisciplinary techniques, questions, and insights for researchers (M. D. Laubichler et al., 2013; Kitchin, 2014). Others complain big data results in unnecessary added complexity and some of this complexity comes from new sources of information known as “digital breadcrumbs” or “data exhaust” left from everyday transactions with phones, the internet, or attributes of the data which have never been collected before or used in an experiment (Boiten, 2016; Trusov, Ma, & Jamal, 2016). Some see data exhaust as contributing to the unknown data lakes and data sinks consisting of unusable data. But these new sources of information have changed what data-driven projects are possible and enables studies with greater breadth, depth, scale, and immediacy when compared to previous research (Lazer et al., 2014).

metadata.

New sources of information from data exhaust, generally exists in the form of an annotation or metadata about the data itself. This metadata is well suited for both quantitative and qualitative methods as it provides a link between the historical, social, and cultural context

of the data to language context. For example, longitudinal social network data was rare and costly to gather prior to the data revolution, so many studies turned to small group studies or cross-sectional analyses (Doreian & Stokman, 1997; Wasserman & Faust, 1994; Friemel, 2011). However, the data revolution has made longitudinal data and metadata more available and now more researchers are conducting data-driven longitudinal analysis across many different fields (M. Newman et al., 2011; English & Underwood, 2016). Yet, while many use these examples to highlight the insights and knowledge from big data to date there is no standardized big data methodology or protocol to guide researchers and scientists and few accessible frameworks applying big data across domains (Kelling et al., 2009).

Data driven science and research (DDSR)

DDSR approaches capture, curate, and enhance the analysis of big data through systematic and repetitive experimentation and analysis. As Bruno Strasser in “Data-Driven sciences: From wonder cabinets to electronic databases”, articulates, DDSR: 1) promotes collaboration between individuals with different disciplinary backgrounds than those who created the data, 2) facilitates statistically based analysis, and 3) creates large amounts of data from the lab (Strasser, 2012) .

Generally, a common starting point in DDSR is creating a workflow(s) of the steps required to complete the experiment or project. For example, the Large Hadron Collider (LHC) computing-grid software that led to the discovery of the Higgs Boson was based on a unique and systematic data-driven workflow which was able to capture tens of petabytes (1 petabyte=10⁶ gigabytes) of data per year, analyze data at every stage of the data collection (trigger analyses), direct the data to specific destinations within the data architecture, coordinate the analysis of the data via multiple systems on the computing grid for registered users from any location, and

repeat these processes with little user maintenance (Alekseev et al., 2016; Britton & Lloyd, 2014). While the illustrated LHC computing-grid workflow did not illustrate the analyses as part of the workflow, in many cases there can be multiple abstract steps, i.e. the results of one analysis- a univariate statistical tests- leads to a different analysis- a multivariate statistical test that are part of the process(Assunção, Calheiros, Bianchi, Netto, & Buyya, 2015). Also, any combination of the following processes: importing data, merging data, cleaning data, standardizing data, normalizing data, cleaning for duplicates, analyzing data (preprocessing, aggregation, transformation, evaluation), exporting, and repeating or modifying this process based on time and resources, could be necessary at any step in a DDSR project.

Workflows generally are flexible iterative collection of methods and processes that help guide the study or experiment. The main purposes of creating a systematic workflow for data-driven projects are: reproducibility and legitimacy of studies, expedite projects' data capture for future analyses, and highlights any possible selection/publication bias, detection bias, and reporting bias of experiments (Moher et al., 2015). For example, in a data-driven textual analysis workflow from biomedicine, a systematic workflow highlights the capture and curation of 8,394 articles from multiple databases, the rationale behind the changes to the dataset throughout the experiment, and the results of the processing of the data in a clearly articulated simplified workflow graphic (Aiello et al., 2016). Although, implementing a DDSR workflow is analogous to designing an experiment based in the scientific method, i.e. there is a question, research, hypothesis, experiment, analysis, and then results are compared to a hypothesis, most DDSR projects implement multiple exploratory analyses at many stages of the experiment rather than towards the end.

exploratory data analysis (EDA).

Exploratory data analysis (EDA) consists of summary statistics, dimensionality reduction, clustering, and visualization of the data so variables of interest, relationships between variables, and patterns of behavior are more easily identified (Tukey, 1977; Cleveland, 1985). EDA can be performed before data capture is finished, as a supplemental analysis to help with understanding primary results or can be conducted post-hoc to understand other relevant data features. Though EDA consists of many different techniques, visualization has become increasingly important as hidden patterns within data have increased the utility of visual analytics (Tufte, 1990, 1997, 1998). To illustrate, in Anscombe's Quartet, four data sets share many statistical properties, e.g. mean, variance, correlation, and regression line, shown in the tables in Figure 8. However, the bivariate scatter plots or visualization of the data shows how the visual analysis is crucial to understanding the characteristic differences in the data sets, visualized in the figures in Figure 8 (Ansecombe, 1973; Tufte, 1985). Also, visual analysis by human experts is often a critical and repeated process in data-driven projects as visual analysis of data takes advantage of inherent pattern matching abilities, edge detection, shape recognition, and can provide alternative findings to automatic algorithms (Ware, 2012). Additionally, previous studies have shown interpretation and results of data-driven analyses were improved by combining human reasoning with visualization (Keim et al., 2008; McCosker & Wilken, 2014). Therefore, exploring the data and visual analysis are important to data-driven approaches and help provide the context needed to design or redesign a model, experiment, or project.

replication.

Several projects using DDSR approaches have shown how it is possible to replicate previous findings or other studies using big data and DDSR. The inability to replicate previous

results is a major issue in science as increased reports of fraud, manipulated statistical analysis, and false research findings have led to multiple “replication projects” and intense debates on scientific practice (M. Baker, 2015; Ioannidis, 2005). The estimated replication-failure rate of scientific results is estimated at 80% and have scientists, policy makers, and the public asking for methods and approaches with repeatable (M. Baker, 2016).

Previous results from big data projects using DDSR were able to go beyond just replicating previous results and provide more sophisticated models of previously studied phenomenon. For example, in a case study by Manfred Laubichler, Jane Maienschein, and Juergen Renn, “Computational perspectives in the history of science: To the memory of Peter Damerow,” their results: 1) confirmed computational data-driven approaches could replicate known historical results of dozens of malaria scholars over multiple decades, and 2) provided evidence data-driven computational methods can be used to create more sophisticated models that are broader in scope than previous models and 3) provide insight into historical phenomena beyond traditional approaches (M. D. Laubichler et al., 2013). In this quote from the text, Laubichler, Maienschein, and Renn (2013) mention the benefits of incorporating big data within historical analysis to model dynamic processes and move past correlation into:

“...specific causal models as explanations of scientific change and innovation that are less based on singular cases and more representative of the totality of scientific activity, both within any given period or research domain as well as across a wide range of fields and historical and geographical areas.” (M. D. Laubichler et al., 2013, p. 124).

Also, other studies have provided additional evidence more sophisticated models created from DDSR outperform simpler models especially when the number of parameters is large or when there is risk of overfitting the data due to small sample size (Witten & Eibe, 2001). These results and others have provided evidence on how DDSR and big data can be used to replicate the

previous work of researchers, help to develop more sophisticated models, and provide insight into historical phenomena.

Thus, DDSR approaches have helped the science in becoming more open to scrutiny, faster, flexible, and responsive to current contextual issues (Hesse et al., 2015). Nonetheless, criticisms of DDSR have emerged which attack DDSR as different than knowledge-driven science.

Criticisms

Both critics and advocates of data-driven science have articulated the differences in methods and approaches compared to knowledge-driven science. This perspective generally appears in industry or business marketing for big-data analytics and has made erroneous claims about data-driven insight presenting a new form of knowledge production and how big data frees science from theory and tradition (Hey et al., 2009). A common theme within this view is the superior ability of machines and computers to make comparison, classifications, or make “smarter” decisions compared to human investigators (Andy Clark, 2013). For many, results from unsupervised machine learning and automated programs combined with the ability to analyze the results of hundreds of different algorithms across massive datasets are evidence of the ability of data to speak for itself (M. Anderson & Anderson, 2011). Advocates of this view see DDSR as a means to go beyond correlation into causal models and prediction using intelligent machines to reduce data and produce results (C. Anderson, 2008).

Yet, even this data first perspective can be viewed as one of the ‘old ways’ with the emphasis on big data and data-driven results simply being a reinterpretation of empiricism with renewed emphasis on results and evidence (A. M. Blair, 2011; Kitchin, 2014). The empiricist perspective articulates how data-driven science is inductive and epistemologically different from

the deductive scientific method, i.e. hypotheses and insights that emerge from data and not expertise or insight (Kelling et al., 2009; Callebaut, 2012). Others have taken the empiricist perspective of DDSR as far as to argue how the scientific method and scientists are no longer necessary for science, anyone can now construct models and test hypotheses through simulation and data-based experiments (Prensky, 2009).

counter argument to criticisms: context

Counter to these arguments is the reality of how data-driven science and research projects and results require context for interpretation. Data can never interpret itself without human interpretation providing the concepts, theories, and knowledge for the algorithm or analysis (Gould, 1981). Those who argue data is free from theory ignore the fact any captured data is always shaped or interpreted by the technology collecting the information, the platforms distributing the data, and the data ontology used to organize the data (Kitchin, 2014). Even if all the capture processes are part of an automated workflow absent of any human interaction, any algorithms used to process the data are products of science and specific scientific approaches influenced by previous theories, frameworks, and paradigms (Leonelli, 2012). Also, the idea that data can speak for itself suggests anyone can interpret the data or the results of data-driven science without contextual knowledge. However, taking an example from biomedicine a gastroenterologist can automatically identify IBD or AHI in string of text but a computer scientist analyzing data coming across would assume this was unimportant or an error in text (R. Cohen et al., 2013).

Additionally, while the possibility of analysis being conducted by anyone is democratic, it has a latent implication only statisticians or a specific kind of ‘anyone’ can interpret the results.

Even statisticians, computer, and data scientists stress the importance of context, domain expertise, and the perils of interpreting data correlations without contextual knowledge.

Context is always necessary to interpret data because without context significant results and discoveries may in fact be occurring due to chance. This problem is a major concern when using DSSR with big data as it is possible to provide supporting evidence for any argument. Essentially, the problem with seeking a needle without any parameters and devoid of meaning in massive haystacks of data, is that too many bits of straw look like needles (James, Witten, Hastie, & Tibshirani, 2013). For example, when performing a regression analysis with multiple extraneous variables the result will always produce a large and statistically significant association between two variables that may in fact have no effect on each other (Wheelan & Davis, 2014). Also, without theory and context it is difficult to ascertain correlation or causality, reverse causality or causality going in both directions, omitted variable bias where a variable takes on multiple explanatory roles, or multicollinearity (Wheelan & Davis, 2014).

Theory and context, also, provide insight into phenomenon and help guide research since any two variables with monotonic variation over time will yield a correlation. The misinterpretations of data have led to many studies citing spurious correlations and mistaking correlation for causation, i.e. how a positive or negative association between two variables does not always mean that a change in one is the source of the change in the other (Ioannidis, 2005). Some of the more outlandish published examples of spurious correlations from data-driven science include:

- 1) US spending on science, space, and technology, is correlated with the number of suicides by hanging, strangulation, and suffocation,
- 2) Number of people who drowned by falling into a pool correlates with the number of films actor Nicolas Cage appeared in, and
- 3) the Per capita consumption of chicken by the pound correlates with the total US crude oil barrel imports

(Reinhart, 2015; Vigen, 2015). Though these misinterpretations on the surface seem humorous they lead to compromised results and unreproducible research (Wheelan & Davis, 2014). Theory and context are especially critical in DDSR with big data because the size and complexity of big data increases the chances of misinterpretations occurring. Therefore, bigger data is not always better, and the interpretations of data-driven results need to be explored using theory and context from domain expertise.

Summary

Thus, the views big data and DDSR is a new form of knowledge production or empiricism reborn, falter under scrutiny and context. The path forward advocated by this dissertation is DDSR approaches create novel workflows, models, and results. Using big data with DDSR benefits from a flexible framework acknowledging knowledge within the history, context, and epistemologies of different research domains. Further, by understanding the limitations of big data and acknowledging the importance of context to interpreting results, DDSR approaches represent an opportunity to draw valuable insights from big data by exploring, extracting, and analyzing interconnected data sets, replicating previous results, fostering interdisciplinary collaborative research, and developing more sophisticated models.

Figures and Tables

Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

	1	2	3	4
Mean X	9.000	9.000	9.000	9.000
Standard Deviation X	3.317	3.317	3.317	3.317
Mean Y	7.501	7.501	7.500	7.501
Standard Deviation Y	2.032	2.032	2.030	2.031
Correlation	0.816	0.816	0.816	0.817

Summary Statistics for Anscombe's Quartet

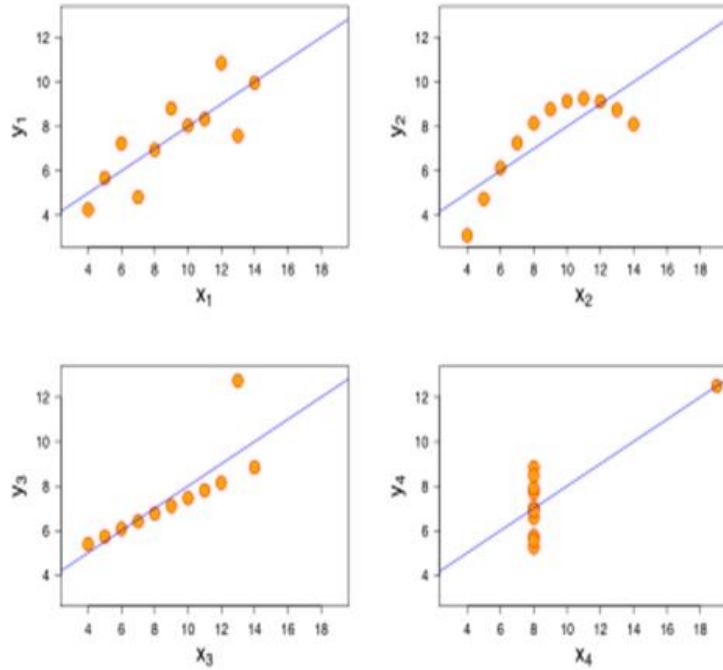


Figure 8. Anscombe's Quartet as a table and as a collection of four graphs (Tufté, 1990).

CONCEPTUAL BOUNDARIES OF THE MICROBIOME, METAGENOME, AND METABOLOME

Abstract

context.

The microbiome is a multidisciplinary concept with several interpretations describing the microbiome as a collection of microbes and microbial cells. Microbiome research outputs have experienced dramatic increases in publications, citations, and funded projects. However, amongst this proliferation there are still debates on whether the microbiome is a separate concept from metabolome and metagenome, or if these three concepts are the same concept.

objective.

This study analyzes three corpora or collections of publications created for each concept, the 224 million word Microbiome Corpus (MB Corpus) consisting of 27977 publications, the 150 million word Metabolome Corpus (MetaB Corpus) consisting of 16,818 publications, and the 78 million word Metagenome Corpus (MetaG Corpus) consisting of 10,741 publications, to describe the similarities and differences of the metabolome and metagenome in comparison to the microbiome. To accomplish this, a combination of qualitative and quantitative approaches was used to study the content of each corpora and the words used with each concept.

results.

The results indicate there are differences in the content, words, and definitions of the microbiome compared to the metagenome and metabolome. Topic models of MetaG Corpus and MetaB Corpus showed both corpora have similar topics (a cluster of words) to the MB Corpus, but no topics consisted of the same clusters of words across corpora. Categories from Keyword in Context analyses (KWIC) revealed differences in patents, approaches, diseases, and

environments between the three corpora. Concordance and collocation analyses revealed suggestively different words co-occurring with the microbiome compared to the metabolome and metagenome when considering word distance, frequency, and exclusivity to each concept. However, the dissimilarities between the co-occurring words and each concept reduce when the scope of the collocational analyses is widened to include first and second order collocates.

Introduction

This paper has two purposes. First, a novel hybrid approach is presented combining quantitative and qualitative analyses of differences between large bodies of texts. This work provides direction on how to analyze concepts and by proxy biomedical knowledge in an era of big data by integrating seemingly disconnected methods from innovative work done in the social sciences, linguistics, and the history and philosophy of science (Adolphs et al., 2004; Franzosi, 2004; Labov, 2006; M. D. Laubichler et al., 2013). The second purpose, is to report a more detailed and experimentally based description of the microbiome concept than previous studies, thereby extending the evidence base for the argument the microbiome is a separate concept from other biomedical concepts (Ursell et al., 2012; Blaser et al., 2013; Huss, 2014; Marchesi & Ravel, 2015a).

historical context.

Adults have around 40 trillion human cells but 10 times that number of microbial cells (Turnbaugh et al., 2007). Genetically we inherit 23,000 different genes from our mother and father, but our microbes provide around 3,000,000 genes (Whitman, Coleman, & Wiebe, 1998). This collection of microbes and microbial cells within our bodies, on our appendages, and in small niche ecosystems have all been called ‘the microbiome’. Recently, the microbiome has been associated as a key factor in understanding disease, health, and changes what it means to be

human, “the concept of a microbiome has implications for how we think of ourselves because it challenges the view of ourselves as atomistic individual organisms,” (Rhodes, Gligorov, & Schwab, 2013, pp. 1–2; Huss, 2014; Schneider & Winslow, 2014; Rees, Bosch, & Douglas, 2018). The microbiome is so critical to understanding humans after the publication of the human genome sequence it was argued the genome sequence would be “incomplete until the synergistic activities between humans and microbes were understood” and “Without understanding the inhabitants of the human microbiome and the mutualistic human microbial interactions that it supports, our portrait of human biology will remain incomplete (Davies, 2001; Council, Studies, Sciences, & Applications, 2007, p. 38). Some advocate the microbiome provides a new interpretation on how important microbes were to human evolution and natural selection, “the genomes from the microbiome endow us with physiologic capacities we have not had to evolve on our own and thus are both a manifestation of who we are genetically, metabolically, and influence our well-being.” (Gordon et al., 2005, p. 1).

In 2008, the National Institutes of Health (NIH) recognized the importance of microbiome research and created the Human Microbiome Project (HMP₁). The HMP₁ was a major interdisciplinary effort to study the microbiome, resulting in a comprehensive analysis of the microbial genes and genomes of the mouth, gut, vagina, and skin (NIH HMP Working Group et al., 2009). The HMP₁ was modeled after the Human Genome Project (HGP) and became another large collaborative data-driven project within biomedicine. While the scientists contributing to the HGP brought about a slow and steady conceptual evolution in biology, the architects of the HMP believed their results through would spark a Kuhnian paradigm shift in human biology (Watson & Cook-Deegan, 1990; Ankeny & Leonelli, 2016). Proponents of the microbiome believed the paradigm shift would forever change the scientific knowledge base

concerning the relationship between microbial community functions and their environment and thrust concepts such as metagenome/genomes, microbiomes, species, evolution, and ecosystem into significance (Juengst & Huss, 2009). When the HMP₁ ended in 2012, the microbiome research yielded new diagnostic biomarkers for health, a plethora of new medicinal drugs, innovative industrial applications, novel theories on nutrition requirements, and new recommendations for the cycle of food production (The Human Microbiome Project Consortium et al., 2012; Mole, 2013). However, what is undetermined is what if any changes occurred to microbiome knowledge (Juengst & Huss, 2009; Rhodes et al., 2013).

The multiple interpretations and uses of the microbiome in different context create major obstacles in assessing the knowledge on the microbiome. Some have argued the microbiome is a just a new combination of “micro” and “biome,” others have claimed there are multiple microbiome including a lung, gut, vaginal, and hand microbiomes, and some have extended microbiome past humans to include animal microbiomes, environmental microbiomes, and space microbiomes (Shade & Handelsman, 2012; Lundberg et al., 2012). Beyond the scope of the microbiome, there are arguments the microbiome is a case of conceptual innovation, while others say the microbiome is simply a rebranded concept (J Lederberg & McCray, 2001; Prescott, 2017). Contradictory studies on patent data have been released for the microbiome because of the conceptual ambiguity and misuse of the microbiome, as one study described little to no increase in the average patent activity for patents based on microbiome research and a different study reported a significant increase in the number microbiome patents (Jones, 2013; Sun, Fiala, & Lowery, 2016).

Knowledge of the microbiome is continuously changing as a result of the use, variation, and range of interpretations on the microbiome (Morton, 2013). The microbiome is at risk of

becoming a hyper object, or a concept with vast temporal and spatial dimensions with unknown contextual boundaries, as the microbiome means different things to different people.(Morton, 2013). A high-dimensional approach to analyze the contextual data is needed to understand the microbiome concept. Yet, this is problematic as it is unclear if the microbiome is a separate concept from the concepts ‘*metabolome*’ and ‘*metagenome*’ even when context is limited to the domain of biomedicine (Ursell et al., 2012; Marchesi & Ravel, 2015a). Lack of conceptual understanding of the microbiome has already led to methods, studies, and results being attributed to other concepts (Marchesi & Ravel, 2015a; Rees et al., 2018). This is problematic as other cases of conceptual ambiguity have resulted in decreased funding, weakened interest, and scientific death (Falk, 2004; Rhodes et al., 2013; Ackerknecht & Haushofer, 2016). Thus, the conceptual ambiguity of the microbiome is both a historical problem for characterizing microbiome research and a contemporary problem that needs to be dealt with urgently. In this experiment I propose a combination of both qualitative and quantitative approaches to provide clarity to the debate on if the microbiome is a distinct concept separate from other cited concepts. If the microbiome is found to be a separate concept, I hope to create a more useful interpretation of the microbiome based on language context or how the microbiome was used in scientific literature.

In this paper, I examine the microbiome using five different contextual language analysis strategies: topic models, frequency analyses, keywords or Keywords in Context (KWIC), collocates, and collocate networks. Using topic models, I analyzed the biomedical discourse surrounding the concepts microbiome, metabolome, and metagenome on topic models trained on the 224 million word Microbiome Corpus (MB Corpus) consisting of 27977 publications, the 150 million word Metabolome Corpus (MetaB Corpus) consisting of 16,818 publications, and

the 78 million word Metagenome Corpus (MetaG Corpus) consisting of 10,741 publications. Using KWIC, I characterized the differences in the tendencies of keywords that are statistically significant into inductively created categories by comparing the MB Corpus to the MetaB Corpus and the MB Corpus to the MetaG Corpus. Using collocates, I described the differences in language knowledge used in defining the microbiome, metabolome, and metagenome. Using collocate networks, I determined how word choice spans are critical in defining the concepts of microbiome, metabolome and metagenome.

Objectives

Given the multiple interpretations of the microbiome concept, the purpose of this experiment is to provide insight into the knowledge of the microbiome, with regards to the following questions:

RQ1: What are the characteristics differences of the language content of the microbiome, metagenome, and metabolome?

RQ2: How is the language used with the microbiome, metagenome, and metabolome similar/dissimilar?

RQ 3: What is the microbiome based on usage?

These questions will be answered with three datasets created from freely available publications collected from Web of Science Core Collection, JSTOR, PubMed, and PubMed Central (PMC).

The publications were identified using a search for the concept of interest (e.g. microbiome, metabolome, or metagenome) as a topic or as a word within the publication. This resulted in a collection of publications or corpus for the microbiome (MB Corpus), metagenome (MetaG Corpus), and metabolome (MetaB Corpus). These corpora provide a means to compare the language context or discourse across corpora and for examination of individual words within a specific corpus. Multiple approaches are used to triangulate the similarities/dissimilarities

between the microbiome and the metagenome, and the microbiome and the metabolome. The paper concludes with a discussion of the different results and a reflection on the limitations of this method.

Materials and Methods

systematic searches.

A qualitative review of search terms for the microbiome has shortcomings, specifically occurrences of false positives. False positives occurred within searches when the search string includes a word of interest, but the word did not occur in the actual body of the text and only occurs as a computer-generated keyword for the article, or as one of the words in the title or citations of the article. Other instances of false positive in searches for this study occurred when the only usage of a word was in reference to a database or project, as in the case of the occurrence of the word microbiome only being used in the text as part of the ‘Human Microbiome Project’. To minimize this problem, both computer and manual methods were used during the collection, curation, and review of articles within each corpus.

Three separate reviewers (Kenneth D. Aiello, Chuquan Shang, and Jasmine Ninita Riggs) reviewed and downloaded publications for each corpora as PDFs in WOS, JSTOR, PubMed, and PMC. A pilot test was conducted to remove any erroneous occurrences of the microbiome in articles only having the word “microbiome” in the title, citation, where the microbiome was only referenced as the Human Microbiome Project, or where the full text was unavailable on a corpus of microbiome articles and from 1900 to 2017. Duplicates were removed during both a computational review and manual review. After the initial cleaning was finished, the final combined list of publications was arranged by their associated metadata to be reviewed again,

and any duplicates were again removed. Articles were excluded if they were not open access during the time of collection the full collection and curation process is displayed in Figure 9. All the PDF files were converted into text files using a combination of Adobe Tools and python scripts. During this process non-English articles had a high rate of unsuccessful conversion and resulted in all non-English publications being removed, the workflow detailed in Figure 9. This process resulted in the creation of the MB Corpus and was repeated for each corpus. This resulted in a total of 27,977 publications for the MB Corpus, 16,818 publications in the MetaB Corpus, and 10,741 publications in the MetaG Corpus. To validate the corpora and the collection method, descriptive statistics combined plots of word frequency by word rank for all words were used, as seen in Table 3. and Figure 10a-c. Zipf's law was observed for each corpus, in accordance with previous studies and results (George K. Zipf, 1935; Hettel, 2013; Moreno-Sánchez, Font-Clos, & Corral, 2016).

data analysis.

topic models.

A distant view of the underlying themes, content, and language context of each corpus were created using topic models. Topic models are a collection of generative probabilistic models used to label topics in data. Latent Dirichlet allocation (LDA) topic models are an unsupervised machine learning algorithm that assumes: V words, K topics, and N document(s) in a document or collection of documents. The output of each LDA model includes the topic probability distribution over documents matrix ($N \times K$) and the word probability distribution over topics matrix ($K \times V$), resulting in a soft cluster of words or topic, as seen in Figure 11 (L. Liu et al., 2016). LDA topic models of 20, 100, and 1000 topics were created for each corpus. Due to K being specified *a priori*, model training of the word term space of documents compared with

the topic space of documents revealed a 20-topic model for each corpus was most beneficial for comparison between corpora. as topic space ($K \times V$) was smaller than word space. The two hyperparameters of alpha and beta were set at 5.0 and 0.1 respectively, and the results may include other attributes of publications including, author, title, location details, and references. The relations or distance among the words or topics are not depicted with the results of a topic. A list of words of that were to be ignored by the model, or stoplist, were created iteratively and used with each corpus to remove function words like: the, what, and, and so on. The topic models were created using the corpora and the software MALLET (McCallum, 2018).

frequency analyses.

Frequency analyses of the words in each corpus were used to discovery differences in the frequency of words and words clusters between the corpora. These frequency analyses were conducted using Wordsmith software (Version 6.0) to produce WordLists that contained the raw frequency of words in each corpus (M. Scott, 2018). The frequency analyses for each corpus yielded lists containing both words of interest and function words. Generally content words like nouns, verbs, adjectives, and adverbs were found to be words of interest, whereas function words like *the, of, a, are, etc.*, provided little insight during the analyses. Many of the function words and other words not included in the analyses were removed through the use of a stoplist. The for the frequency analyses contained words ignored during the analyses, different from the topic model stoplists. The frequency analysis stoplists were created through manual repeated examination of the frequency analyses for each corpus. The frequency analyses also showed multiple lemmas from stems of words, as the stem of *analyze* in the MB Corpus showed both American and British English lemmas such as: *analysis, analyses, analyzes, analyzed, and analysed*. Lemmas were combined to represent one word during analysis by using a

lemmatization list (one for each corpora) during the processing of the WordLists. The lemmatization lists were created through manual repeated examination of the frequency analyses for each corpus.

concordance.

Concordances provided the ability see and read the range of contexts in which the word microbiome, metagenome, and metabolome, and other words of interest were used in their original context. Concordance analyses from Wordsmith software (Version 6.0) provided a list of all the examples of a word and the surrounding language context of a word or phrase specific to each corpus. Manual reading and interpretation of the words co-occurring with the microbiome, metagenome, and metabolome were supported by usage of the words in the text. The results of the concordance analyses were often used as a means of additional evidence to help guide the interpretation of results from the keywords, collocates, lexical profiles, and collocate networks.

keywords.

The tendencies of words occurring more frequently in one corpus compared to another were discovered using keyword analyses. Using the data from the frequency analyses, keyword analyses were conducted using Wordsmith software (Version 6.0) to create Keywords in Context (KWIC) lists. The KWIC lists showed the top 500 keywords that were statistically significant (using Log-likelihood) based on differences in the frequency in the MB Corpus compared to the MetaG Corpus, and the MB Corpus compared to the MetaB Corpus were examined. The KWIC lists were useful preliminary word lists helping to isolate language patterns specific to each corpus and assisted in identifying differences between the words used in each corpus. Additionally, insight into the corpora and the keywords were found using the KWIC lists in

combination with a keyword cluster list. This facilitated comparative aspects of the range of knowledge within these corpora of text to be discovered. Categories of keywords were created as a result of applying a coding scheme to the keywords and keyword clusters, with evidence supported by the frequency lists and concordances of the keywords to evaluate the actual language use (D. H. Hymes & Gumperz, 1972; D. Hymes, 2013).

collocates.

Knowledge of individual words and the predictable combinations of words were found through collocational analyses. Collocational analyses from Wordsmith software (Version 6.0) revealed collocates, or words co-occurring with a word of interest at a greater frequency than by chance. Using the words microbiome, metagenome, and metabolome as node words, collocates were found using the frequency of occurrence of the node word compared to the frequency of occurrence of words within a span of eight total words (four words to the left or right). Results highlighting the most common interpretations and uses of words within each corpus were found by using a combination of common statistical tests used to measure the strength of association between words. These statistical tests included MI3, Log-likelihood, T-score, and Z-score. To mitigate biases related to the dispersion of words in a text, differences in total words, and difference in number of texts, only the top 20 collocates shared collocates in at least two of the four scores were used following previous studies (Rayson & Garside, 1997; Kilgarriff, 2001; Ghadessy, Henry, & Roseberry, 2001; Stubbs, 2001; Burkette & Kretzschmar, 2018).

lexical profiles.

Systematic, coherent, and comprehensive interpretations for microbiome, metabolome, and metagenome were created based on summaries of the language context derived. These summaries were derived from information on node words, collocates, and their respective

frequencies into a lexical profile. The twenty strongest shared collocates based on scores from MI3, Log-likelihood, T-score, and Z-score of microbiome, metabolome, and metagenome were used to in combination with evidence from the frequency analyses and concordance analyses to create lexical profiles for each concept. The following notation was used:

- *node word* (number of occurrences of node word in corpus): < collocate₁ (percentage of occurrence of collocate₁ with node word), collocate₂ (percentage of occurrence of collocate₁ with node word), ... collocate_n (percentage of occurrence of collocate_n with node word) >

collocate networks.

The representation of knowledge and meaning for each concept was reviewed and compared for similarities using the results from collocational networks. The collocational networks were created using each concept as the central node or ego within the collocational network. To compare networks, only communal collocates across concepts were used as direct links to the node concepts, or first order collocates. Collocational analyses were then conducted on the first order collocates to identify second order collocates or collocates of the first order collocates. Ego networks were then created for each concept with the node word (microbiome, metagenome, metabolome) being linked to shared first order collocates across the concepts, and first order collocates linked to their respective top twenty collocates (second order collocates). In total, four collocate networks were created: microbiome with shared collocates with metagenome, microbiome with shared collocates with metabolome, metagenome with shared collocates with microbiome, and metabolome with shared collocates with microbiome, represented as: microbiome *C MetaG 2001-2017* <>; metagenome *C MB 2001-2017* <>; microbiome *C MetaB 2001-2017* <>; and metabolome *C MB 2001-2017* <> respectively. Differences between networks were evaluated using the Jaccard Index or Jaccard Similarity Coefficient, which measures the

similarity between networks based on a score between 1.0 and 0.0, 1.0 representing 100% similarity (P. Jaccard, 1912; Wasserman & Faust, 1994).

Results

RQ1: What are the characteristics differences of the language content of the microbiome, metagenome, and metabolome?

topic models.

Overall, the topic models suggested there is a great deal of shared content from the microbiome to the metabolome and even more so between microbiome and the metagenome, and the word microbiome clusters with other words in each collection. The topic models for each concept also highlighted differences in latent themes represented by word clusters between each corpus. Looking at topic bins from a topic model set to 20 topics with each topic consisting of 20 words for each concept related to bacteria, 28 of the same words were found in 8 topics on bacteria in the MB Corpus and 3 topics on bacteria in the MetaB Corpus, these words were: *aureus, bacteria, bacterial, coli, community, composition, diet, disease, diversity, doi, fecal, growth, gut, host, human, infection, intestinal, mice, microbial, microbiome, microbiota, pubmed, resistance, species, strain, strains, studies, virulence*, presented in Table 4. However, differences from the topic model between MB Corpus and MetaB Corpus suggest topics on bacteria within the MB had an environmental (*environment, fungi, plant, root, soil, water*) and an oral (*dental, gingivitis, plaque, periodontitis, saliva*) focus; compared to the topics on bacteria from the MetaB Corpus had a metabolic (*acid, fermentation, glucose*) and obesity (*body, fat, food, metabolism*) focus.

Comparison of topic model bins for bacteria between 8 topic bins from the MB Corpus and 7 topic bins from the MetaG Corpus revealed 69 of the same words in topic bins which

included: *abundance, analysis, antibiotic, antibiotics, antimicrobial, bacteria, bacterial, coli, communities, community, composition, coral, data, diet, disease, diversity, doi, environmental, fecal, fermentation, fish, food, fungal, fungi, growth, gut, health, healthy, high, host, human, infection, insect, intestinal, isolates, marine, mice, microbial, microbiology, microbiome, microbiota, oral, otus, plant, plants, production, pubmed, relative, resistance, rhizosphere, root, rRNA, rumen, sample, samples, sequences, sequencing, skin, soil, soils, species strain, strains, streptococcus, studies, study, table, virulence, water*, see Table 5. Differences from the topic model between MB Corpus and MetaG Corpus suggest topics on bacteria in the MB Corpus had a gastroenterological (*acid, food, lactic, lactobacillus, milk, probiotics, yeast*), oral (*dental, gingivitis, plaque, periodontitis, saliva*), and entomological (*drosophila, insects, larvae*) focus; compared to topics on bacteria from the MetaG Corpus had a body of water (*freshwater, lake, ocean, sea*), and immunological (*immune, immunity, infection, inflammation, inflammatory*) focus. Also, the word *microbiome* was found in at least one topic in the results of all three corpora.

frequency analyses.

The frequency analyses showed differences in the words, articles, and use of words within each corpus. The type/token ratio (TTR) was used to evaluate the lexical density of each corpus. The MB Corpus had the highest TTR at 0.76, the MetaG was lower than MB Corpus at 0.69, and the MetaB had the lowest TTR at 0.54, displayed in Table 3. Differences in patterns of word usage were found when the frequencies and ranks of the words of interest were compared from the frequency analyses after the stoplists and lemmatization lists were applied, as depicted in Table 6. Biology and biomedical words like *cell, gene, analyse, and study* were found in the

three corpora. Yet, in general more supporting evidence for differences in the overall language context of each corpora were found by comparing the top-ranking words from each corpora.

keywords.

Keywords and Keyword in Context (KWIC) lists were created with the MB Corpus as the corpus of analysis and the MetaB Corpus and MetaG Corpus as the reference corpora, and vice versa using Wordsmith Tools (M. Scott, 2018). Differences in tendencies of word usage between each corpus showed positive keywords for the MB Corpus in comparison to the MetaB Corpus displayed in Table 7. The keywords were calculated using Log-likelihood with negative keywords not included as part of the analysis. The same process was used to discover the positive keywords for the MB Corpus in comparison to the MetaG Corpus

The most noticeable differences in tendencies between the MB Corpus and MetaB Corpus were articulated into selected keyword categories including: patient/population, approaches, disease/disorder/condition, symptoms, process, unit of analysis, and location/environment, presented in Table 8. Validation of keywords between corpora were supported by frequency analyses and concordance. To illustrate, the word *children* was a categorized as a patient/population keyword when the MB Corpus was compared to the both the MetaB Corpus and the MetaG Corpus. To determine which keyword category *children* belonged to, the context of the word *children* was evaluated using data from the frequency analysis of MB, which showed children occurred 73, 487 times in 7,989 texts, which represents an estimated 29% dispersion of the word *children* across the MB Corpus. These numbers indicate that the word is used multiple times in each text and not a random pattern. The actual context of use of *children*

from the concordance analysis was used to determine the category *children* belonged to the patient/population category, with typical uses of the word like:

“Data were obtained from a total number of 32,206 swabs and in twenty-two percent (N = 6,956) *S. aureus* was present. After excluding patients for whom age was unknown 6037 (87,8%) adults (aged 18+) and 840 (12,2%) *children* (aged 4 to 17) were included in our study sample, of which 56% were female” (van Bijnen et al., 2014).

“*Children* who were exclusively breastfed until at least 6 months of age (n = 71) did not show an increase in diarrheal rates associated with antibiotic exposure (Table 3). Conversely, among *children* who had discontinued exclusive breastfeeding before 6 months (n = 394), any early antibiotic exposure was associated with a 48% relative increase in diarrheal rates (IRR: 1.48, 95% CI: 1.23, 1.78) “(Rogawski et al., 2015)

“Our initial evaluation of this hypothesis used data available in the Autism Speaks Autism Treatment Network (AS-ATN) database that includes a large sample of 2-17 year old children and youth with ASD ascertained at medical centers across the U.S. and Canada ” (Marler et al., 2017).

Using this approach for all keywords, keywords distinguished the difference in the selected keywords categories between the corpora. In the patient/population category, differences in the MB Corpus and MetaB Corpus showed keywords associated with humans and animals like *adults, cattle, children, fish, infant, mice, and pigs* tended to occur in the MB Corpus; whereas the keywords of patient/population of MetaB Corpus showed tendencies towards smaller parts of organisms like *cells, genes, and tissues*, and plant related keywords *leaves, roots, seeds, and seedlings*. Examining differences in keywords based on the disease/disorder/condition category found keywords like *asthma, diarrhea, dysbiosis, enterocolitis, hiv, ibd, obesity, and periodontitis* were used more often in the MB Corpus; while keywords for the same category for the MetaB corpus exhibited tendencies to use keywords such as *alzheimer's, and cancer*.

Keywords distinguishing the patients/population of study in the MB Corpus included a greater quantity and variety of words associated with humans, including keywords such as *adult, animal, children, infant, neonatal, and women*, compared to the keywords for population/patients

of MetaG Corpus which revealed keywords like *enzyme, eukaryote, gene, prokaryote, and virus*. Significant tendencies in the location/environment category of keywords, also, showed a larger repertoire of keywords related to the body in the MB Corpus including *bile, body, brain, breast, colon, gut, intestinal, kidney, liver, lung, stomach, and vagina*; compared to keywords tending to relate to water in the MetaG Corpus like *coastal, freshwater, hypersaline, lake, marine, oceans, pacific, seawater, and waters*.

The results of the KWIC analyses confirms a difference in the words used in the MB Corpus compared to the MetaG Corpus, and the MB Corpus in comparison to the MetaB Corpus. The keywords results reinforce the findings of the topical model but provides detail and context at the word level beyond the topic model results. These results support previous results using KWIC to compare sensitive variations in the language context of large collections of biomedical text and text collections in general (Seale et al., 2006; Kronberger & Wagner, 2000b).

RQ2: How is the language used with the microbiome similar/dissimilar to the language used with metagenome, and metabolome?

collocates & lexical profiles.

The knowledge and meaning of the microbiome was dissimilar compared to the metagenome and metabolome. First, to examine if differences in corpus size, words, types, and occurrences of node word (i.e. node word representing the concept the corpus was built on and the focal point of the analysis) would have a large impact on the results (refer to Table 10.), the top 20 collocates from three 5,000 article random samples, MB 1, MB 2, and MB3 from the MB Corpus were compared with *microbiome* as the node word, see Table 11. Results from the collocational analysis showed a high similarity between the top 20 shared collocates for each random sample and the Top 20 collocates from the MB Corpus. Table 12 shows a comparison of

the top 20 collocates for each sample compared to the top 20 collocates for the MB Corpus, the collocates colored in red are shared collocates between the MB Corpus and the random sample and the collocates which were unique to the specific sample were not-highlighted. This approach was used to find the shared collocates for microbiome in the MB Corpus, metabolome in the MetaB Corpus, and metagenome in the MetaG Corpus.

Clear differences arose between the microbiome and the other concepts looking at the top shared collocates for each concept, see collocates of *microbiome* in the MB Corpus depicted in Table 13., collocates of *metagenome* in the MetaG Corpus in Table 14., and collocates of MetaB in Table 15., and when represented as lexical profiles:

microbiome₂₀₀₁₋₂₀₁₇ <gut 23%, human 17%, intestinal 4%, oral 3%, analysis 3%, host 3%, core 3%, changes 2%, analysis 3%, host 3%, core 3%, changes 2%, project 2%, healthy 2%, studies 2%, skin 2%, vaginal 2%, lung 2%, obesity 2%, disease 1%, infant 1%, role 1%, diversity 1%>

metagenome₂₀₀₁₋₂₀₁₇ <analysis 6%, sequencing 6%, data 6%, gut 4%, genome 5%, assembly 4%, soil 3%, reads 3%, study 3%, microbial 3%, genes 3%, wide 3%, human 3%, viral 2%, shotgun 2%, whole 2%, marine 2%, metatranscriptome 2%, rumen 1%, derived 1%, novo 1%, accessing 1%>

metabolome₂₀₀₁₋₂₀₁₇ <human 14%, analysis 11%, database 8%, transcriptome 8%, proteome 6%, data 5%, changes 4%, hmdb 4%, microbiome 3%, serum 3%, plasma 3%, profiling 3%, genome 2%, wide 2%, gut 2%, urine 2%, arabidopsis 2%, coverage 1%, golm 1%, fecal 1%, exo 1%, lipidome 1%, epigenome 1%>

The differences in the usage and meaning of the microbiome compared to the metagenome and metabolome are magnified when looking at the communal collocates between the concepts, as only four communal collocates (*gut*, *human*, *analysis*, and *changes*) were revealed between microbiome and metagenome, and three communal collocates (*gut*, *human*, and *analysis*) were discovered between microbiome and metabolome, displayed in Table 16 and Table 17 respectively. The similarity between the collocates was low, the Jaccard Score (J) for $J =$

$(\text{microbiome}_{2001-2017}, \text{metagenome}_{2001-2017}) = 7.5\%$, and $J = (\text{microbiome}_{2001-2017}, \text{metabolome}_{2001-2017}) = 7.7\%$. Combined, these results show the microbiome concept as a separate concept from metagenome and metabolome. These results, also, shed some light on the possible misinterpretations between the microbiome concept and the other concepts, as microbiome was a collocate of metabolome and the communal collocates *gut* and *human* are the top collocates for microbiome.

collocation networks.

Interestingly, the differences in knowledge surrounding the concepts are less pronounced when communal collocates between the concepts are used as a starting point for creating collocate networks. The node concepts and the associated communal collocates (*c*) were simplified to:

microbiome *C MetaG 2001-2017* <gut 23%, human 17%, analysis 3%>

metagenome *C MB 2001-2017* <analysis 6%, gut 4%, human 3%>

microbiome *C MetaB 2001-2017* <gut 23%, human 17%, analysis 3%, changes 2%>

metabolome *C MB 2001-2017* <human 14%, analysis 11%, changes 4%, gut 2%>

Each of the networks depicts the semantic structure and knowledge associated within close proximity of each concept, created from the node concept, first order collocates, and second order collocates, graphed in Figures 10-14. The networks were created using ORA, the central node being the concept, the first order collocates as orange nodes, and the second order nodes as grey nodes, with the links colored by source node (“CASOS Tools: Network Analysis Data | CASOS,” n.d.). The similar knowledge between the node concepts increases drastically (40% between microbiome and metagenome and 30% between microbiome and metabolome) when communal collocates with second order collocates are used: $J(\text{microbiome}_{2001-2017} :$

$\langle \text{gut}; \text{human}; \text{changes}; \rangle$, $\text{metagenome}_{2001-2017} : \langle \text{gut}; \text{human}; \text{changes}; \rangle = 47\%$,
and $J(\text{microbiome}_{2001-2017} \langle \text{gut}; \text{human}; \text{changes}; \text{analysis}; \rangle, \text{metabolome}_{2001-2017} \langle \text{gut}; \text{human}; \text{changes}; \text{analysis}; \rangle) = 37.1\%$. Therefore, these results suggest the conceptual boundaries of the microbiome concept and the concepts metagenome and metabolome are less distinct when incorporating communal collocates and second order collocates.

RQ 3: What is the microbiome based on usage?

Major differences in the knowledge of the words and combination of words within a span of 8 words to each concept were found via the shared 20 collocates for each concept. The *microbiome* concept had a total of 20 shared collocates: *analysis, changes, core, disease, diversity, gut, healthy, host, human, infant, intestinal, lung, obesity, oral, project, research, role, skin, studies, and vaginal*. Many of these collocates were found to be specific microbiomes via supporting evidence from frequency analysis and concordance. Based on this analysis, there are multiple different microbiomes and the *microbiome* can be interpreted as any 1 or all these meanings and at any given time (Lawrence Edwards et al., 2016).

Discussion

The results suggest the microbiome is a separate concept from metagenome and metabolome, based on results from the larger global language context depicted in the topic models and keywords, and the dissimilar knowledge of the concepts microbiome, metabolome, and metagenome. However, this distinction is not a discrete boundary as results from the collocate networks highlight the complex knowledge structure of biomedical concepts changes depending on the emphasis of analysis. Multiple representations of the microbiome concept were also found, ranging from the larger *human* and *core* microbiome, to smaller niches such as a *gut*,

intestinal, lung, oral, skin, and vaginal microbiome. These results suggest the microbiome concept may in fact be an amalgam of multiple microbiome concepts, separated by contextual differences.

A discussion on the status of the microbiome vocabulary is justified by these results, as the results support an evaluation and possible revision of microbiome language made by previous studies (Huss, 2014; Rhodes et al., 2013). Comprehensive reform of microbiome language is not necessary, but these results highlight possible sources of confusion between the concepts. Suggestions on how to add specificity or eliminate ambiguity on the microbiome could come from scientists, researchers, and other microbiome experts but not from these results. Refining scientific knowledge is a daunting task that requires many considerations to be weighed including preferences, states, and preferred language. The reward for which is applicability, understanding, and effectiveness of the microbiome concept.

These experiments make both a substantive and methodological contribution to the interpretation and knowledge of the microbiome by confirming and adding to hypothesized unique characteristics of the language use and knowledge of the microbiome (Bordenstein, n.d.; Marchesi & Ravel, 2015a; Prescott, 2017). Analysis of large collections of text across multiple dimensions and scales is one of the advantages of these methods over conventional analysis applied to studies of biomedical concepts and language. Possible objections to these conclusions about the microbiome based on this approach include: the methods ignore temporal fluctuations, the results emphasize majority of use and are biased towards more recent usage and knowledge of the concepts. Variation has been found at the level of individuals, and individuals have influenced the trajectory of concepts and knowledge over time (Labov, 1972; Burri & Dumit, 2007; Cetina, 2009). This is not a deficiency of the current analysis, but more of a specification

made by the objectives and the data collection. The results from the analyses, up to this point, provide evidence given a specific set of documents based on each word, there are characteristic differences at the word, combination of words, and themes which can be extended to the form, function, and interpretation of each word. However, the selection of each corpora limits the results specific to each word, i.e. results are based on a corpus built on the word microbiome compared to a corpus built on the word metabolome or metagenome there were significant differences. This does not account for the possibility the knowledge or interpretation of the microbiome is distinct based on other search criteria. Specifically, the chance the microbiome may have existed without the explicit use of the word microbiome or other possible connotations associated with the microbiome.

However, this approach provides direction for future analyses to be conducted on other groups, individuals, concepts, or knowledge if textual data is available. Specifically, understanding on how other contextual factors beyond usage, distance, and exclusivity within language influence knowledge might be compared. Social and historical factors have influenced variation and diachronic change of the microbiome may be discovered based on the MB Corpus and accompanying metadata. Lastly, this approach provides the capacity for a more interpretative and empirical construction of biomedical knowledge within corpora but can also be applied to smaller social constructions such as individuals or groups. Interpretative understanding of the differences in large collection of texts is the advantage of using this method as it removes a priori views of the researcher from the identification of features of interest, as the results are identified using statistical procedures.

Figures and Tables

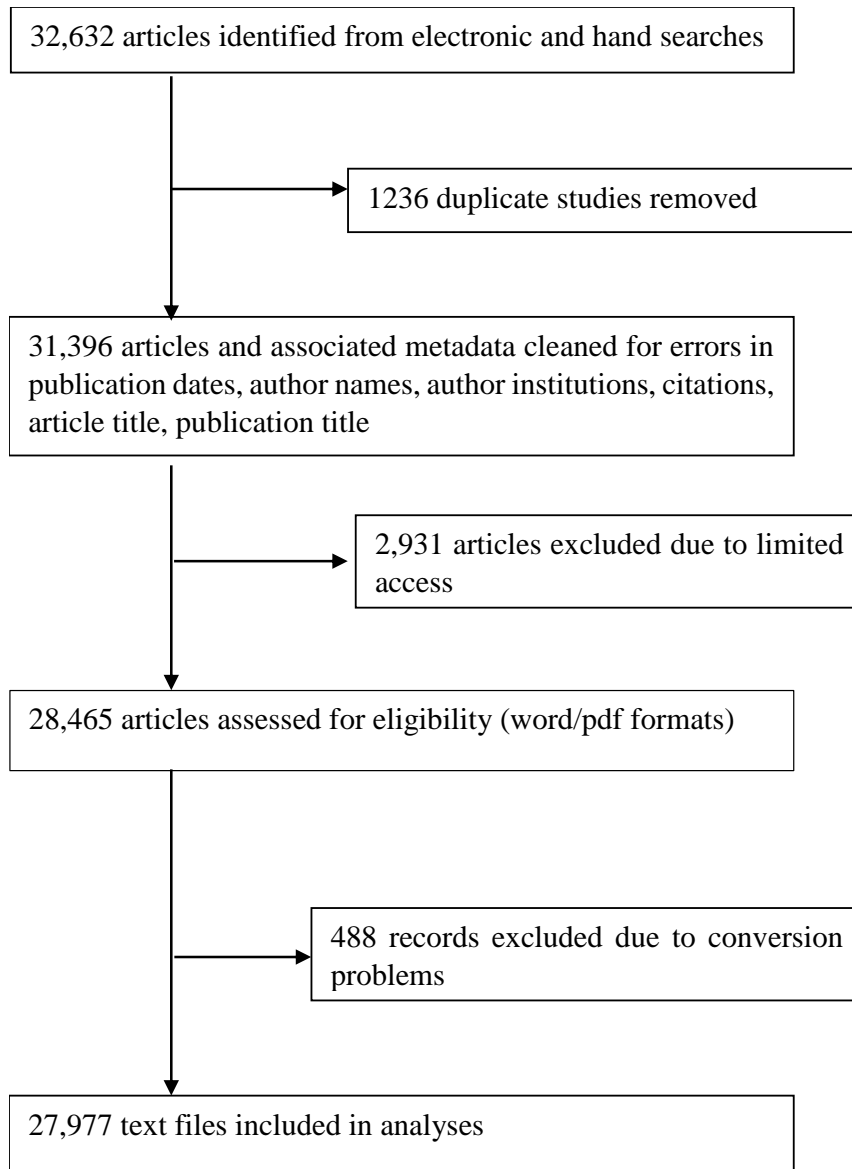


Figure 9. Systematic collection of articles for MB Corpus.

Table 3. Descriptive statistics for the Microbiome Corpus, Metabolome Corpus, and Metagenome Corpus.

	Microbiome Corpus (MB Corpus)	Metabolome Corpus (MetaB Corpus)	Metagenome Corpus (MetaG Corpus)
Total Articles	27,977	16,818	10,741
Total Characters	1,461,313,152	983,734,336	514,108,544
Tokens (all words)	224,701,200	150,675,280	78,963,584
Tokens Used for Word List	202,072,928	137,519,504	72,501,776
Types (unique words)	1,542,965	747,612	500,600
Type/Token Ratio (TTR)	0.76	0.54	0.69
Standardized TTR (STTR)	40.04	40.33	40.92
STTR Standard Deviation	60.11	59.69	59.16
STTR Basis	1,000	1,000	1,000
Mean Word Length	5.19	5.16	5.2
Word Length Standard Deviation	3.31	3.2	3.19

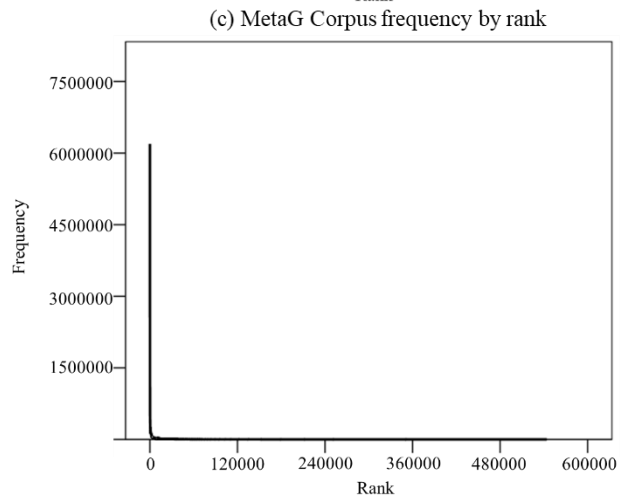
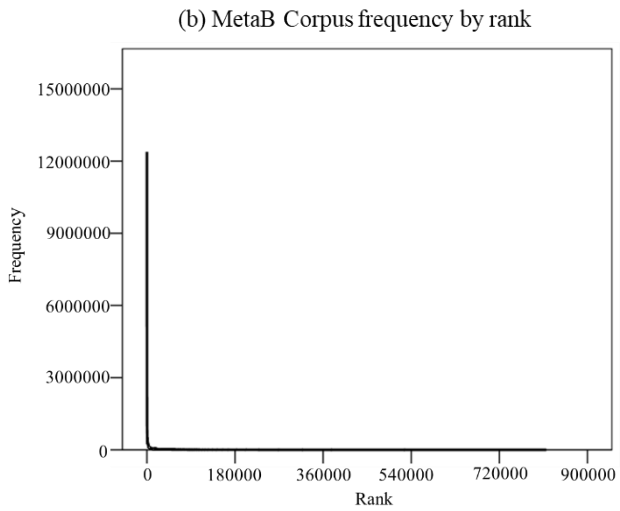
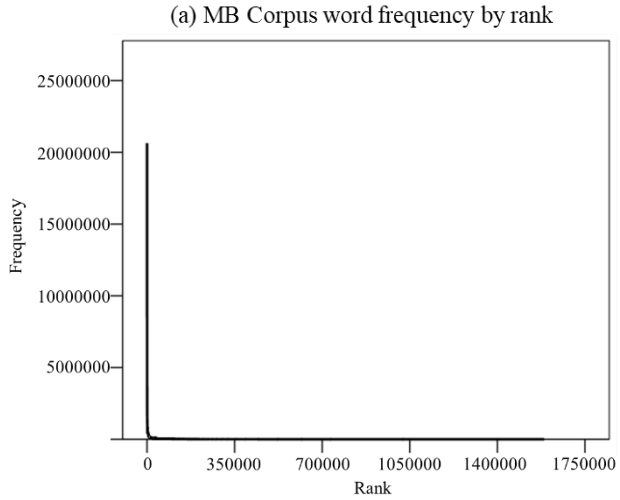


Figure 10a-c. Word frequency by rank. 10a. MB Corpus word frequency by rank. 10b. MetaB word frequency by rank Corpus. 10c. MetaG Corpus word frequency by rank.

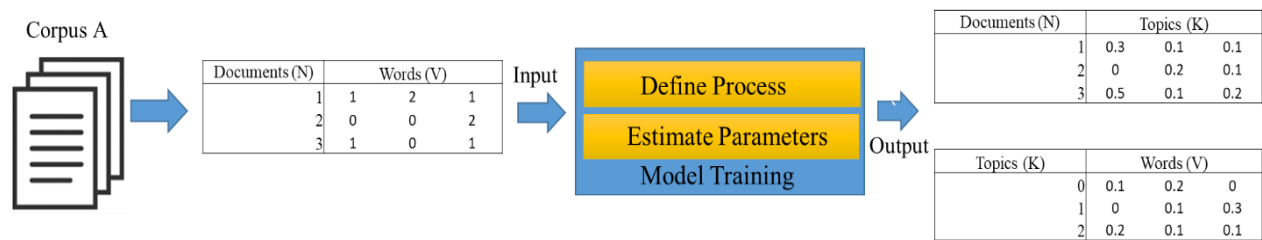


Figure 11. Topic modeling process and results.

Table 4. Topic bins for bacteria from MB Corpus compared to MetaB Corpus. The following topics (topic 0, 2, 5, 8, 11, 13, 16, 17) from the MB Corpus had the word bacteria as a result. The words in red were words also found in the topic model of the MetaB.

0	2	5	8	11	13	16	17
antibiotic	anaerobic	author	bacteria	bacteria	abundance	acid	bacteria
antibiotics	bacteria	bacteria	bacterial	bacterial	analysis	bacteria	bacterial
antimicrobial	community	bacterial	bacteroides	biol	bacterial	bacterial	communities
aureus	concentration	caries	composition	coral	communities	fermentation	community
bacteria	concentrations	dental	diet	doi	community	food	diversity
bacterial	conditions	disease	disease	drosophila	data	growth	doi
coli	degradation	gingivitis	fecal	evolution	diversity	isolated	environmental
colonization	environment	human	gut	fish	dna	isolates	fungus
human	fig	manuscript	health	gut	fig	lab	fungi
infect	high	nih-pa	healthy	host	microbial	lactic	marine
infection	methane	oral	host	host	otus	lactobacillus	microbial
infection	microbial	page	human	infection	relative	microbiol	microbiol
isolates	microbiol	periodontal	intestinal	insect	rrna	milk	plant
microbiol	process	periodontitis	mice	insects	sample	probiotic	plants
pathogens	production	plaque	microbes	larvae	samples	probiotics	rhizosphere
resistance	rumen	pmc	microbial	midgut	sequences	production	root
skin	sludge	pubmed	microbiome	species	sequencing	species	soil
staphylococcus	total	saliva	microbiota	symbionts	species	strain	soils
strain	treatment	species	species	transmission	study	strains	species
virulence	water	streptococcus	studies	wild	table	yeast	water

Table 5. Topic bins for bacteria from MB Corpus compared to MetaG Corpus. The following topics (topic 0, 2, 5, 8, 11, 13, 16, 17) from the MB Corpus had the word bacteria as a result. The words in red were words also found in the topic model of the MetaG Corpus.

0	2	5	8	11	13	16	17
antibiotic	anaerobic	author	bacteria	bacteria	abundance	acid	bacteria
antibiotics	bacteria	bacteria	bacterial	bacterial	analysis	bacteria	bacterial
antimicrobial	community	bacterial	bacteroides	biol	bacterial	bacterial	communities
aureus	concentration	caries	composition	coral	communities	fermentation	community
bacteria	concentrations	dental	diet	doi	community	food	diversity
bacterial	conditions	disease	disease	drosophila	data	growth	doi
coli	degradation	gingivalis	fecal	evolution	diversity	isolated	environmental
colonization	environment	human	gut	fish	dna	isolates	fungal
human	fig	manuscript	health	gut	fig	lab	fungi
infect	high	nih-pa	healthy	host	microbial	lactic	marine
infection	methane	oral	host	host	otus	lactobacillus	microbial
infection	microbial	page	human	infection	relative	microbiol	microbiol
isolates	microbiol	periodontal	intestinal	insect	rrna	milk	plant
microbiol	process	periodontitis	mice	insects	sample	probiotic	plants
pathogens	production	plaque	microbes	larvae	samples	probiotics	rhizosphere
resistance	rumen	pmc	microbial	midgut	sequences	production	root
skin	sludge	pubmed	microbiome	species	sequencing	species	soil
staphylococcus	total	saliva	microbiota	symbionts	species	strain	soils
strain	treatment	species	species	transmissio	study	strains	species
virulence	water	streptococcus	studies	wild	table	yeast	water

Table 6. Comparison of ranks and frequencies of words of interest from the corpora. Differences in the top 20 ranked words after stoplists and lemmatization lists were applied.

MB Corpus			MetaB Corpus			MetaG Corpus		
Rank	Word	Frequency in Corpus	Rank	Word	Frequency in Corpus	Rank	Word	Frequency in Corpus
1	cell	717,790	1	cell	652,534	1	gene	336,912
2	study	611,865	2	gene	441,569	2	sequence	264,447
3	sample	498,628	3	analyse	414,473	3	sample	234,262
4	gut	489,327	4	study	367,768	4	analyse	221,118
5	microbiota	485,372	5	protein	334,063	5	genome	208,909
6	analyse	450,097	6	acid	333,407	6	study	192,548
7	gene	444,364	7	data	317,959	7	data	181,571
8	bacterial	411,299	8	level	278,904	8	cell	168,485
9	human	402,069	9	plant	270,301	9	microbial	164,690
10	mouse	400,120	10	high	260,085	10	high	156,588
11	disease	386,346	11	show	258,553	11	protein	152,436
12	high	362,042	12	increase	243,785	12	bacterial	149,122
13	microbial	360,858	13	sample	239,825	13	show	132,543
14	bacteria	345,752	14	expression	237,704	14	gut	128,488
15	show	336,717	15	result	224,786	15	bacteria	128,377
16	increase	335,647	16	metabolic	220,067	16	human	124,336
17	group	333,972	17	patient	208,656	17	group	121,994
18	data	326,050	18	significant	205,215	18	base	120,332
19	patient	317,589	19	metabolite	204,224	19	microbiota	120,098
20	effect	306,589	20	pathway	195,084	20	dna	119,521

Table 7. Positive keywords in the MIB Corpus compared to the Metab Corpus

N	Key word	Frequency in MIB Corpus	Percent of Total Words (MIB Corpus)	Number of Texts (MIB Corpus)	Frequency in Reference Corpus (Metab Corpus)	RC. %	Keyness (Log-likelihood)	Lemmas
1	microbiota	485,372	0.0022	21380	65,287	0.0004	217,866.98	
2	gut	484,021	0.0022	19870	74,631	0.0005	195,129.17	gut[484021] guts[5262]
3	bacterial	411,299	0.0018	23942	54,923	0.0004	185,564.92	
4	microbial	360,858	0.0016	23281	48,451	0.0003	162,139.69	
5	bacteria	345,752	0.0015	23659	50,156	0.0003	146,463.34	
6	intestinal	274,831	0.0012	16660	40,510	0.0003	114,923.46	
7	community	187,682	0.0008	17647	22,922	0.0002	90,053.16	
8	diversity	187,777	0.0008	18576	26,399	0.0002	81,477.25	
9	rna	112,465	0.0005	13622	10,771	0	62,327.07	
10	dx	67,453	0.0003	4812	1,664	0	56,784.04	
11	otus	84,432	0.0004	6695	5,320	0	56,230.04	
12	immune	174,952	0.0008	15371	36,380	0.0002	52,251.88	
13	host	210,476	0.0009	19170	51,497	0.0003	50,796.82	host[210476] hosts[18500]

Table 8: Comparison of significant keyword categories MB Corpus compared to MetaB corpus

Categories of Keywords	MB Corpus	MetaB Corpus
Patient/Population	adults, animals, cattle, children, fish, host, humans(s), individuals, infants(s), mice, pigs, population, virus	cells, genes, leaves, metabolites, phenotypes, roots, seeds, seedlings, sprps, tissues,
Approaches	biodiversity, dental, ecological, gastroenterology, immunology, metagenome, microbiology, molnar, pct, phylogenetic, pyrosequencing, taxonomic	conifers, biochemistry, biology, cellular, chemistry, chromatography, electrospun, engineering, hpc (high performance liquid chromatography), ionization, lipidomics, mapping, microarray, molecular, modeling, orthogonal, pea, phytochemistry, physiological, proteomics, scan, spectrometry, spectroscopy, transcriptomics, wabburg
Disease/Disorder/Condition	allergy, asthma, atopic, autoimmune, campylobacter, candida, colitis, diarrhea, diet, difficile, dysbiosis, enterocolitis, hiv, ibd, obesity, periodontitis	alzheimer's, cancer, hypoxia, stress
Symptoms	chronic, inflammation, inflammatory, irritable,	fatty, malignant
Process	amplification, birth, colonize, commensal, contaminated, digestion, fermentation, hygiene, immunity, infect, isolated, opportunistic, pathogenic, permeability, prevent, reinfaction, succession, symbiotic, transmission, transplora, viral	abiotic, acclimation, accumulation, activity, aging, assimilation, binding, biosynthesis, catabolism, changes, circadian, collusion, death, deficiency, deprivation, derivatization, deprivation, discovery, drought, expression, flux, germination, glycolysis, heat, hypoxia, identity, imaging, intensity, kinetic, localization, magnetic, metabolism, metastasis, methylation, modifications, mutation, osmotic, oxidation, phosphorylation, photosynthesis, profiling, radiation, rate, reaction, redox, regulation, reprogramming, respiration, retention, senescence, signaling, silencing, starvation, synthesis, time, tolerance, upregulated, uptake, validation
Unit of Analysis	agar, antibiotic, antimicrobial, butyrate, dairy, food, metronidazole, microbial, milk, mucus, probiotic, svvd, toxin, vaccine, acidophilus, actinobacter, actinobacteria, akkermansia, alphaproteobacteria, amplicon, anaerobes, antigen, archaea, bacillus, bacteria, bacteroides, bacteroidetes, bifidobacterium, cation, clostridia, community, cytoine, dendritic, dna, pseudomonas, enterobacteriaceae, enterococcus, epithelium, faecalibacterium, feces, firmicutes, flora, fusobacterium, gammaproteobacteria, genomes, germ, haemophilus, helicobacter, interleukin, ketobeta, lachnospiraceae, lactobacillus, methane, microbes, microbiota, microflora, microorganism, oligosaccharides, pathogen, phylotypes, plaque, pneumonitis, porphyromonax, prevotella, primer, propionibacterium, proteobacteria, pseudomonax, pylori, rdna, rhinomonas, rna, rna/hococcus, salmonella, staphylococcus, symbionts, taxa, vancomycin, veillonella, viruses,	acetoinitrite, acetyl, acids, alanine, ascorbate, bisphosphates, citrate, compound, drug, energy, gas, glycerol, hif (hypoxia-induced factor), hydroxy, light, liquids, methanol, nicotinamide, night, noise, pathway, proline, rapamycin, serum, solvent, tobacco, in, yeast, abscisic, adenine, adp, amino, arabinoside, arginine, aspartate, atp, auxin, biomarker, carbon, carboxylate, carnitine, cdna, cells, cerevisiae, chlorophyll, chloroplast, chromatin, creatine, cysteine, cytochrome, dehydrogenase, embryo, erzwmet(s), extracts, fibroblasts, flavonoids, flower, fruit, fumarate, genes(s), glucose, glutamate, glutamine, glutathione, histone, histone, integration, ion, isocitrate, isoform, isoflavone, isolate, ketoglutarate, kinase, lactate, leaves, lipid, male, malate, markers, membrane, metabolites, methyl, mitochondria, mrra, mutant(s), myc, myo, nod, nodd, noddy, nodp, nitrogen, oxygen, pecks, penicose, pep, phenotypes, phosphate, phospholipid, phenylalanine, phosphate, phosphoglycerate, phospholipid, plant, plasma, polyamine, precursor, promoter, protein, purine, pyruvate, quadrupole, ribose, rice, roots, saccharomyces, salt, seed(s), seedlings, serine, sprps (single nucleotide polymorphism), succinate, sucrose, sugar, synthesis, synthetase, thaliana, tobacco, tomato, transcription, tumor, tyrosine, udp, valine
Location/Environment	biofilm, biome, bowel, cavity, cecal, cecum, colon, distal, ecosystem, enteric, epithelial, fecal, gastric, gastrointestinal, glioblastic, gut, habitat, ileum, intestine, lake, lamina, lumen, lymphoid, mucosal, nasal, niche, periodontal, ribosomal, sediment, skin, soil, surface, tract, ulcerative, vaginal, wastewater	brain, breast, capillary, cardiac, cytosol, heart, intracellular, kidney, leaf, muscle, network(s), ovarian, prostate, proteome, renal, ribose, skeletal, stem, subcellular, tissues, tumors, urine,

Table 9. Comparison of significant keyword categories MB Corpus compared to MetaG corpus

Categories of Keywords	MB Corpus	MetaG Corpus
Patient/Population	<i>adult, animal, children, female, fetal, human, infant, male, maternal, mouse (mice), neonatal, offspring, patient, pigs, rat, subjects, women</i>	<i>enzymes, eukaryotes, genes, prokaryotes, viruses</i>
Approaches	<i>anova, antibiotics, dose, gastroenterology, immunology, randomized, studies, test, therapeutic, therapies, transplant, trials</i>	<i>algorithm, analyses, approach, binning, biogeochemical, bioinformatics, bootstrap, classification, cluster, comparative, computational, crisispr, encode, geochemical, hydrothermal, marker, metatranscriptomic, predictions, query, sequence, shotgun, taxonomic, tools</i>
Disease/Disorder/Condition	<i>allergy, anxiety, arthritis, asthma, atopic, autoimmune, blindness, cancer, colitis, depression, diabetes, diarrhea, dysbiosis, hiv, infect, inflammation, obesity, pregnancy, ulcerative (colitis)</i>	<i>accession, annotated, biosynthesis, catalytic, conserved, construction, contained, coverage, degradation, denitification, discovery, evolution, heterologous, oxidation, photosynthetic, purification, recombination, reconstruction, uncultivated, uncultured</i>
Symptoms	<i>inflammation, irritable</i>	
Process	<i>aging, apoptosis, chronic, commensal, consumption, damage, death, diet, epigenetic, excretion, expression, factor, feeding, healthy, homeostasis, injury, mortality, nutrition, pain, pathogenesis, progress, proliferation, recurrent, reduced, regulates, signaling, smoking, stimulation, stress</i>	<i>amino, ammonia, ammonium, anoxic (waters), autotrophic, bacteriophage, bacterioplankton, base, basin, carbon, chloroflexi, clade clone, community, cyanobacteria, data, dna, domains, electron, enzymes, eukaryotes, extremophiles, families, genbank, genes, genome, heterotrophic, homologous (homologs), lineages, methane, motifs, nitrate, nucleotides, operons, organism, phages, phototrophic, phylogenetic, phytoplankton, plankton, plasmids, primers, prochlorococcus, prokaryotes, prophage, proteins, reductase, repeats, residues, rna, sites, strain, streptomycetes, substrate, sulfide, sulfur, synechococcus, vitrome, viruses</i>
Unit of Analysis	<i>alcohol, antigen, antimicrobial, antioxidant, drugs, food, insulin, metronidazole, milk, probiotics, secretion, serum, vaccine, vancomycin, vitamin, biomarkers, cation, cell, chemokine, cholesterol, clostridium, cytokine, dendritic (cells), (clostridium) difficile, (Bifidobacterium, germ, hormones, interferon, lactobacilli, lipid, lymphocytes, macrophage, metabolite, microbiota, monocytes, mucin, nucleus, neurons, neutrophils, pathogens, plasma, (Helicobacter) pylori, receptor, tissues, urine, weight</i>	
Location/Environment	<i>bile, blood, body, bone, bowel, brain, breast, colon, endothelial, enteric, faecal, fat, fecal, gastric, gut, hepatic, ileum, intestinal, kidney, liver, lung, lymphoid, marrow, meat, nasal, nervous (system), oral, organs, periodontal, pulmonary, renal, respiratory, skin, spleen, stomach, stool, tissue, tumor, urinary, vaginal, visceral (fat)</i>	<i>coastal, environmental, freshwater, habitats, hypersaline, lake, libraries, marine, natural, oceans, pacific, ribosomal, scaffolds, seawater, sediments, subsurface, subsystems, trees, waters</i>

Table 10. Descriptive statistics for collocate analyses of node word in MB Corpus, MetaG Corpus, and MetaB Corpus.

	Total Articles	Total Words	Total Types	Total Collocates (With Positive Relationship)				
				Total Occurrences of Node Word	MI3	Log-Ikelihood	T-Score	Z-Score
MB Corpus	27,977	224,064,368	1,192,009	239,819	17,133	17,100	16,542	8,003
MetaG Corpus	10,741	78,963,584	500,600	44,173	10,745	10,721	10,126	5,146
MetaB Corpus	16,818	150,675,280	747,612	31,656	12,849	12,178	12,195	6,017

Table 11. Descriptive statistics of 5,000 article random samples (MB 1, MB 2, MB 3) compared to MB Corpus

	Total Articles	Total Words	Total Types	Total Occurrences of Node Word	MI3	Log-likelihood	T-Score	Z-Score
MB1	5,000	35,326,140	403,930	40,844	10,298	9,811	9,796	5,213
MB2	5,000	55,728,168	487,718	48,389	9,817	9,397	9,287	5,017
MB3	5,000	36,194,280	413,371	42,730	10,599	10,078	10,065	5,313
Mean	5,000.0	42,416,196.0	435,006.3	43,987.7	10,238.0	9,762.0	9,716.0	5,181.0
S.D.	0.0	11,536,674.8	45,893.1	3,926.6	394.4	343.1	395.1	150.6
Median	5,000.0	36,194,280.0	413,371.0	42,730.0	10,298.0	9,811.0	9,796.0	5,213.0
Variance	0.0	133,094,865,661,488.0	2,106,172,972.3	15,418,050.3	155,581.0	117,741.0	156,121.0	22,672.0
MB Corpus	27,977	224,064,368	1,192,009	239,819	17,133	17,100	16,542	8,003

Table 12. Comparison of top 20 shared collocates

MB Corpus	MB 1	MB 2	MB 3
analysis	analysis	analysis	alters
changes	changes	capacity	analysis
core	core	changes	changes
disease	data	core	core
diversity	disease	core	disease
gut	diversity	distal	diversity
healthy	gut	gut	gut
host	healthy	healthy	healthy
human	host	host	host
infant	human	human	human
intestinal	infant	infant	infant
lung	intestinal	intestinal	intestinal
obesity	lung	lean	lung
oral	obesity	lung	obesity
project	oral	obesity	oral
research	project	oral	project
role	research	project	research
skin	role	research	role
studies	skin	research	skin
vaginal	studies	skin	studies
	vaginal	vaginal	vaginal

Table 13. Top 20 shared collocates of microbiome in MB Corpus.

Rank	MI3	Log-likelihood	T-score	Z-score
1	gut	gut	gut	gut
2	human	human	human	human
3	core	core	intestinal	core
4	project	oral	oral	project
5	oral	project	analysis	oral
6	intestinal	intestinal	host	viewed
7	skin	skin	core	vaginal
8	vaginal	host	changes	lung
9	lung	changes	project	skin
10	changes	healthy	healthy	lean
11	healthy	vaginal	disease	infant
12	host	lung	studies	consortium
13	analysis	analysis	skin	healthy
14	infant	studies	diversity	metabolome
15	obesity	research	research	capacity
16	research	obesity	role	changes
17	studies	role	data	alters
18	role	diversity	vaginal	distal
19	diversity	disease	lung	intestinal
20	infant	infant	obesity	jumpstart

Table 14. Top 20 shared collocates of metagenome in MetaG Corpus.

Rank	MI3	Log-likelihood	T-score	Z-score
1	sequencing	sequencing	sequencing	metatranscriptome
2	metatranscriptome	analysis	analysis	accessing
3	assembly	data	data	shotgun
4	analysis	assembly	gut	wide
5	wide	wide	genome	illuminates
6	shotgun	gut	assembly	assembly
7	data	shotgun	soil	anticoccidial
8	gut	soil	reads	sequencing
9	derived	metatranscriptome	study	association
10	soil	derived	microbial	chemolithoautotroph
11	whole	genome	genes	scalable
12	viral	viral	human	derived
13	genome	reads	wide	whole
14	accessing	whole	derived	soil
15	reads	study	viral	viral
16	study	human	shotgun	sargasso
17	novo	microbial	based	polytheonamides
18	rumen	genes	whole	novo
19	human	marine	samples	analysis
20	microbial	rumen	marine	urbanized

Table 15. Top 20 collocates of metabolome in MetaB Corpus.

Rank	MI3	Log-likelihood	T-score	Z-score
1	transcriptome	human	human	hmdb
2	human	transcriptome	analysis	transcriptome
3	hmdb	database	database	golm
4	database	analysis	transcriptome	database
5	proteome	proteome	proteome	proteome
6	analysis	hmdb	data	human
7	golm	changes	changes	coverage
8	microbiome	microbiome	hmdb	analysis
9	changes	data	serum	exo
10	coverage	serum	microbiome	fluxome
11	data	golm	plasma	configuring
12	serum	profiling	profiling	microbiome
13	profiling	plasma	study	lipidome
14	plasma	coverage	genome	bibliome
15	exo	wide	wide	epigenome
16	urine	urine	gut	scriptome
17	fecal	genome	arabidopsis	gmd
18	lipidome	gut	urine	profiling
19	nucleic	fecal	microbiota	serum
20	epigenome	arabidopsis	cellular	wide

Table 16. Communal collocates between microbiome and metagenome.

MB Collocate	Frequency		MetaG Collocate	Frequency with	
	with MB	Percentage		MetaG	Percentage
gut	56,587	23%	analysis	1,929	6%
human	41,851	17%	sequencing	1,907	6%
intestinal	8,844	4%	data	1,811	6%
oral	7,480	3%	gut	1,359	4%
analysis	6,745	3%	genome	1,223	4%
host	6,702	3%	assembly	1,168	4%
core	6,549	3%	soil	1,022	3%
changes	5,736	2%	reads	931	3%
project	5,579	2%	study	930	3%
healthy	5,407	2%	microbial	921	3%
studies	5,334	2%	genes	898	3%
skin	4,918	2%	wide	828	3%
research	4,574	2%	human	820	3%
vaginal	4,316	2%	viral	789	2%
lung	4,042	2%	shotgun	745	2%
obesity	3,940	2%	whole	648	2%
disease	3,154	1%	marine	598	2%
infant	2,931	1%	metatranscriptome	531	2%
role	2,773	1%	rumen	407	1%
diversity	2,567	1%	derived	404	1%
			novo	372	1%
			accessing	188	1%

Table 17. Communal collocates between microbiome and metabolome.

MB Collocate	Frequency with MB	Percentage	MetaB Collocate	Frequency with MetaB	Percentage
gut	56,587	23%	human	6,061	14%
human	41,851	17%	analysis	5,000	11%
intestinal	8,844	4%	database	3,505	8%
oral	7,480	3%	transcriptome	3,341	8%
analysis	6,745	3%	proteome	2,432	6%
host	6,702	3%	data	2,164	5%
core	6,549	3%	changes	1,788	4%
changes	5,736	2%	hmdb	1,771	4%
project	5,579	2%	microbiome	1,295	3%
healthy	5,407	2%	serum	1,269	3%
studies	5,334	2%	plasma	1,218	3%
skin	4,918	2%	profiling	1,145	3%
research	4,574	2%	genome	871	2%
vaginal	4,316	2%	wide	822	2%
lung	4,042	2%	gut	789	2%
obesity	3,940	2%	urine	755	2%
disease	3,154	1%	arabidopsis	732	2%
infant	2,931	1%	coverage	619	1%
role	2,773	1%	golm	615	1%
diversity	2,567	1%	fecal	548	1%
			exo	272	1%
			lipidome	245	1%
			epigenome	239	1%

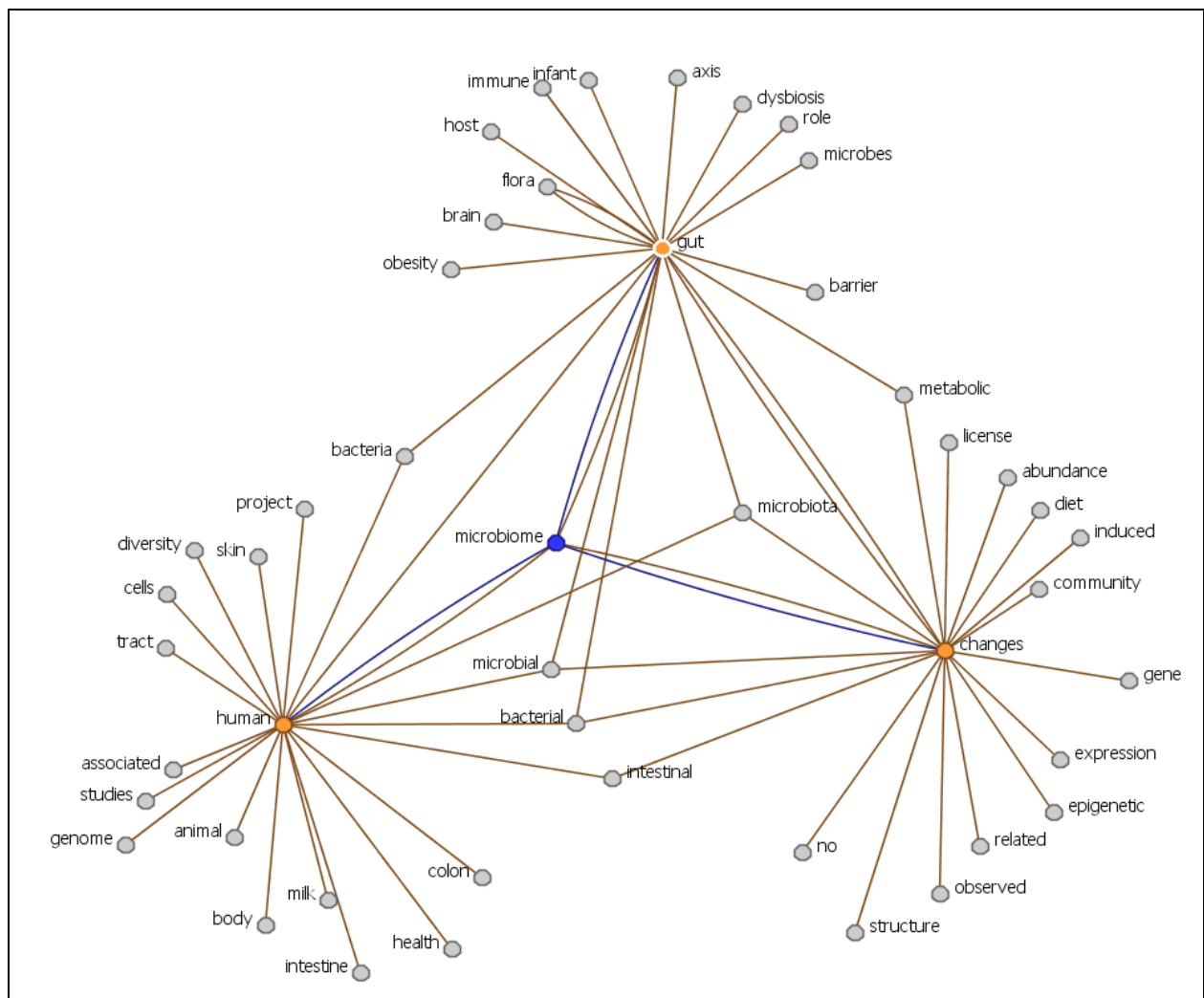


Figure 12. Collocates network of microbiome :< changes, gut, human>.

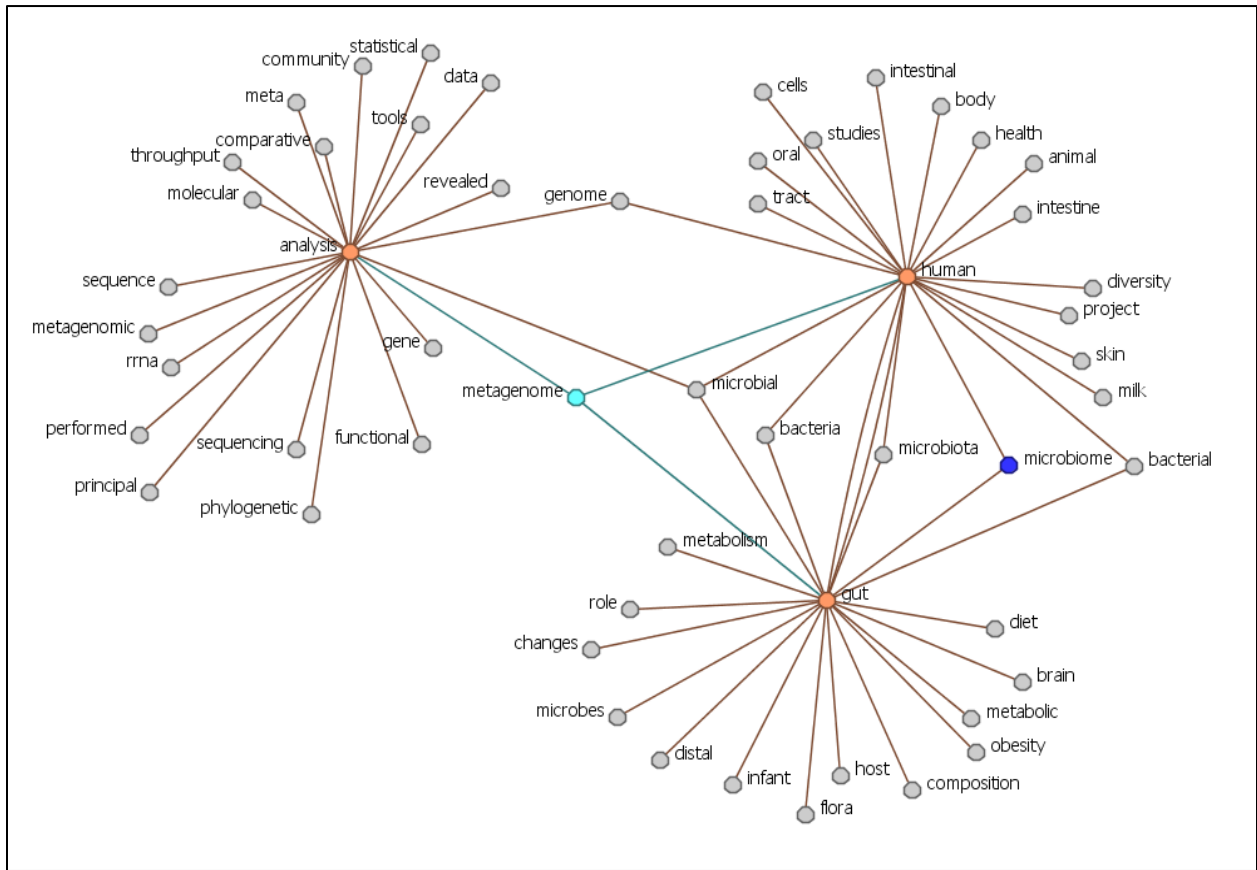


Figure 13. Collocates network of metagenome :< analysis, gut, human>.

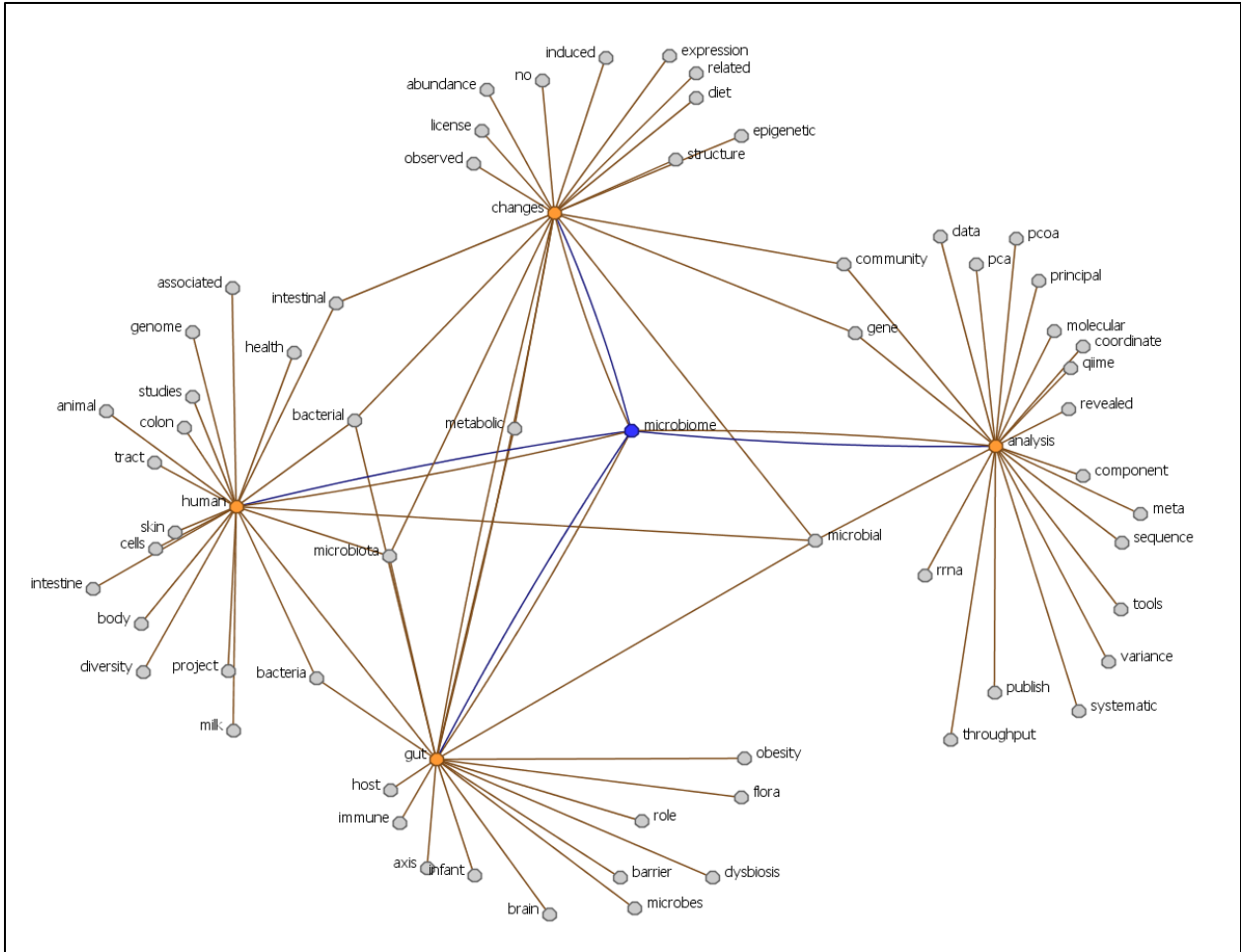


Figure 14. Collocate network of microbiome :< gut, human, analysis, changes>.

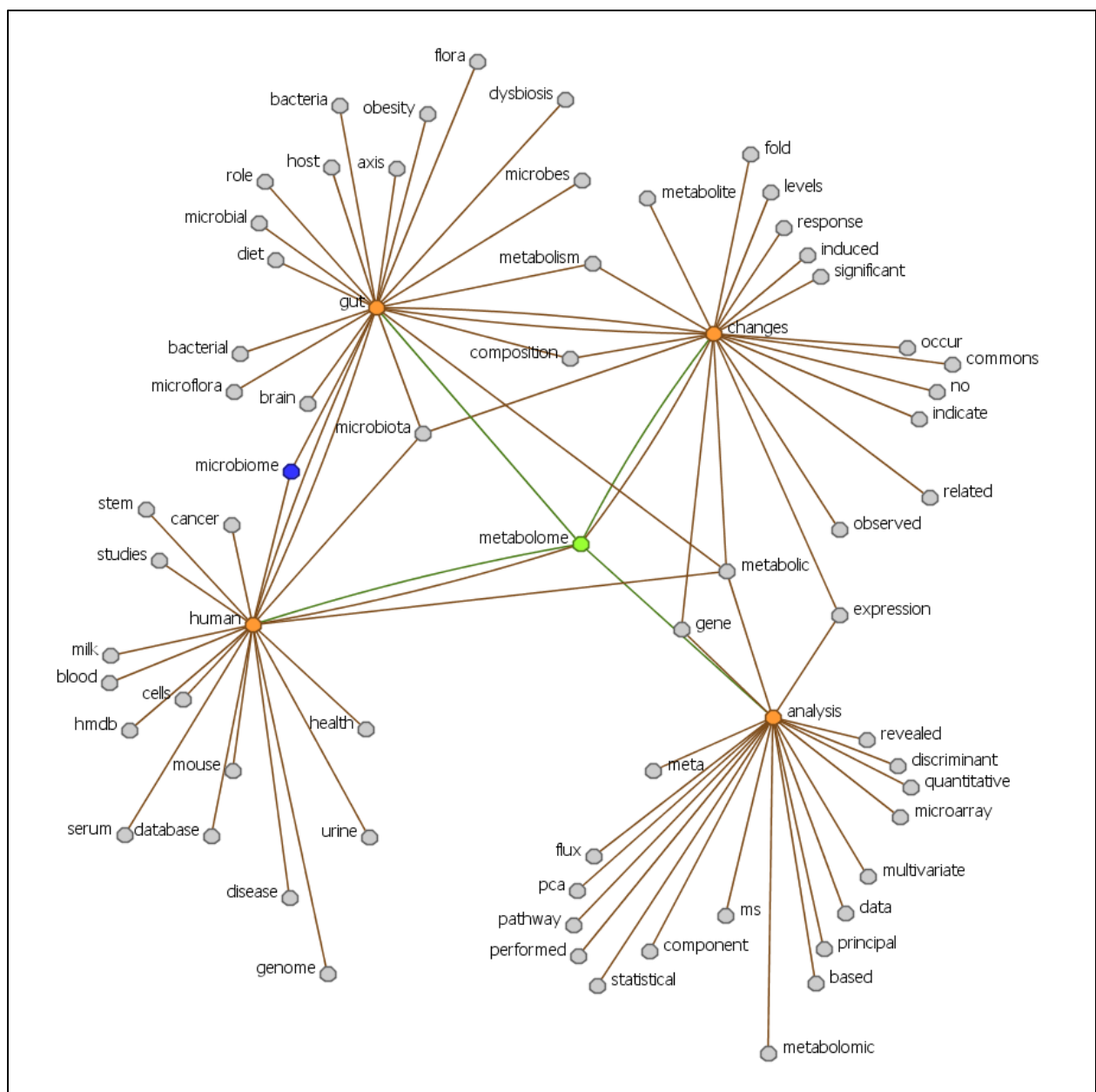


Figure 15. Collocates network of metabolome: < human, analysis, changes, gut>.

SYSTEMATIC REVIEW OF THE FACTORS INFLUENCING THE EVOLUTION OF THE MICROBIOME CONCEPT

Abstract

This article provides the results of an exhaustive review on the questions, how has microbiome research, funding, and knowledge evolved over time? This review provides both innovative methods for the process of studying literature in a systematic and ordered way, and highlights how these methods lead to novel questions and conclusions. This article studies the contextual factors influencing the microbiome via a reproducible methodology using biomedical literature and peer-reviewed articles. Drawing on a framework of extended evolution and the history of knowledge from Renn and Laubichler, variation and changes to the microbiome were characterized, clarified, and traced (Renn & Laubichler, 2017b). Characterizing the microbiome as an innovation helped to understand sources of variation and change in microbiome knowledge, social, and language context. It was discovered the microbiome is an innovation based on the relative advantage, compatibility, complexity, trialability, observability, and rate of adoption, conclusions that are in accordance with previous studies on the innovation in biomedicine, healthcare, and other domains (Greenhalgh et al., 2005; Rogers, 2010).

Introduction

Producing reviews of multi-dimensional heterogeneous evidence is a challenging methodological area. Unanswered questions remain on how to provide insight on scientific knowledge from diverse synchronic and diachronic data sets using both qualitative and quantitative methods (Greenhalgh et al., 2005; Lazer et al., 2009). Systematic reviews are used within biomedicine and other domains, as most systematic reviews are transparent and demanding syntheses of large bodies of evidence following rigid protocols to produce

quantitative data and results (Dixon-Woods & NHS Health Development Agency, 2004; Dixon-Woods, Agarwal, Jones, Young, & Sutton, 2005). Systematic reviews are a way for experts to keep up to date with their field, a starting point for practice guidelines, provide justification for granting agencies, investigate cost-effectiveness, diagnostic or prognostic questions, and inform policy (Buchwald et al., 2004; Petticrew & Roberts, 2005). A systematic review is planned around clearly formulated questions using a systematic approach to identify, organize, and analyze relevant research using quantitative data. The studies included in the review provide the data used to address a research question using any relevant type of research (Moher et al., 2015). However, valuable insight and interpretation beyond simple description have been found in studies integrating qualitative evidence in systematic reviews (Greenhalgh et al., 2018). The study described in this chapter was used as an opportunity to apply novel approaches to the synthesis of evidence across multidisciplinary fields on the microbiome concept.

Scientists have identified the microbiome as the key to understanding disease, health, and what it means to be human (“Me, myself, us,” 2012; The Human Microbiome Project Consortium et al., 2012). Researchers of the microbiome argue the human genome sequence is incomplete until the synergistic activities between humans and microbes, or the microbiome, are understood (Davies, 2001; Venter et al., 2001). In 2007, the National Institutes of Health (NIH) recognized the importance of microbiome research and announced the Human Microbiome Project (HMP). The HMP was a 5-year and \$150,000,000 interdisciplinary effort to study the microbiome and catalogue the distinct microbial communities on the human body (Relman & Falkow, 2001; The Human Microbiome Project Consortium et al., 2012). The HMP resulted in new diagnostic biomarkers for health, medicinal drugs, industrial applications, theories on

nutrition requirements, and recommendations for the cycle of food production (The Human Microbiome Project Consortium et al., 2012).

Interest in the microbiome has translated into large amounts of funding and resources. The White House Office of Science and Technology Policy (OSTP) announced a National Microbiome Initiative (NMI) with a combined Federal agency investment of more than \$121 million in Fiscal Year (FY) 2016 and 2017 to study the microbiome(s) of different ecosystems and organisms, with other stakeholders and institutions, like the Bill and Melinda Gates Foundation, Health Ministries Network, and The BioCollective LLC committing upwards of \$400 million (“Announcing the National Microbiome Initiative,” 2016). The full scope of government funded investment on microbiome research is much larger, according to data from the Federal Research Portfolio Online Reporting Tools (RePORT), the number of projects and subprojects associated with the microbiome from 2004 to 2017 increased from 1 to 1403 projects and funding for microbiome projects increased by 101% (from \$33,725,501 to \$3,411,212,659), shown in Table 18. (“Search Results - NIH RePORTER - NIH Research Portfolio Online Reporting Tools Expenditures and Results,” n.d.; “Federal RePORTER - Project Search Results,” n.d.).

Despite this wide interest, the origin of the ‘microbiome concept’ or ‘microbiome’ has not been plainly defined in research or studies on the subject. Historically, some historians and scientists speculate the microbiome concept could be traced as far back as Antonie van Leeuwenhoek 1632–1723 (Forum, 2013). Van Leeuwenhoek is considered “the father of microbiology” and was the first to observe and describe microorganisms, which he referred to as “animalcules” (Scher & Abramson, 2011). Yet, Leeuwenhoek never explicitly used the word microbiome in text and this causal link has no empirical data nor have any studies been

published investigating this matter in a systematic way. The majority of the "animalcules" Leeuwenhoek observed were unicellular and multicellular organisms he observed in pond water (Gest, 2004). In 2001, Joshua Lederberg claimed to create the concept and meaning of the microbiome. Lederberg defined the microbiome as "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space" (J Lederberg & McCray, 2001). However, the first use of the word microbiome in print occurred in a book reviewing dentistry techniques in 1949, "we know, on the other hand, the role of the pH in the evolution of gut microbiome" (*Revue odontologique*, 1949). Subsequently in 1952, John Mohr uses the microbiome concept to explain unicellular organisms (protozoans) with organisms (polysaprobies) inhabiting decomposing lakes, "the protozoan fauna (as a matter of fact, the whole microbiome) is poor in species and individuals, and those present are rather typical polysaprobies" (Mohr, 1952). These ambiguities create difficulties in understanding the evolution and history of the microbiome and the trajectory of scientific knowledge associated with the microbiome.

The conceptual debate on the microbiome occurs across multiple dimensions, as microbiome research is observed and conceptually experimented (i.e. tried out) in many different fields including but not limited to ecology, molecular biology, microbiology, and medicine. Ecology uses the microbiome to describe large geographic regions of niche bacterial environment, for example the coral core microbiome represents the coral community and niche coral habitats (D Ainsworth et al., 2015). While, microbiology uses the microbiome in to reference fungi and viruses, such as the luminal contents of ileum, cecum, and colon or specific intestinal regions in mice (Preidis & Versalovic, 2009). A search for microbiome in the PubMed Medical Subject Headings Thesaurus or MeSH , the world's largest controlled vocabulary,

resulted in unclear conceptual boundaries as the results displayed *microbiota* as the “preferred” concept, and *microbiome*, *microbial community*, and *microbiome* as “related” concepts (“MeSH Preface,” 2014; “MeSH Browser,” n.d.). Even the American Academy of Microbiology admits that, “there is not yet a fully agreed upon definition” of the microbiome (American Society for Microbiology, 2013, p. 3).

The gatekeepers of conceptual innovations and knowledge are historically high impact journals on the cutting edge of science or a field (Lyytinen, Baskerville, Iivari, & Te’eni, 2007). This implies new knowledge and the overall knowledge of the microbiome should converge with these journals. However, to ensure a study of the microbiome as an occasion that connects and produces new knowledge, there is a growing need to quickly assess and analyze the ever-growing deluge of information from multiple scientific fields and research outputs that are simultaneously defining and deploying the term (Fuller, Revere, Bugni, & Martin, 2004; Smalheiser, 2017).

The expansion and diversification associated with the term microbiome has led to large amounts of data on the microbiome including: microbiome articles, the journals in which these articles appear, the NIH Medical Subject Headings (MeSH) terms categorizing these journals, and NIH project proposals. This study analyzes the big data traces and tractable indicators of what the microbiome is and how the microbiome has evolved combined with qualitative inferences based on the contextual specifics of the usage of the microbiome in text. The results from this study provide insight into how scientific knowledge and the microbiome research field evolved and how knowledge on the microbiome was influenced.

Methodological challenges

Traditional systematic reviews collect data on a specific question with *a priori* inclusion and exclusion criteria providing parameters on what type of articles to include (e.g. randomized controlled trials, cross-sectional, double blind, other reviews, etc.) and to remove from study. In this study, all articles with *microbiome* in the text were collected, excluding articles with the only occurrence of *microbiome* in the title, abstract, or references, or instances where microbiome was used only as part of the multi-word unit “*The Human Microbiome Project.*” A wide-ranging literature search for articles in the PubMed, Web of Science (WOS), and Journal Storage (JSTOR) databases was conducted in July 2015 and again in January 2018. The specific search strategy used was dependent on the database. PubMed was searched via the [Text Word] option in the builder for ‘microbiome’ with a filter on for publication date from 1900/01/01 to 2017/12/31. In the Web of Science (WOS) the topic microbiome was searched with a timespan: 1900-2018, in all available WOS Core Collection: Citation Indexes, e.g. Science Citation Index Expanded (SCI-EXPANDED) –from 1900 to present, Social Sciences Citation Index (SSCI) from 1956 to present, Arts & Humanities Citation Index (A&HCI) from 1975 to present, and Emerging Sources Citation Index (ESCI) from 1900 to present. JSTOR was searched with microbiome full-text option and date range 1900 to 2017. All bibliographic records for relevant articles were collected. These records included: the citing article, the cited references, author(s) and co-author(s), journal of publication, publication year, and MeSH term for the journal of publication, and were downloaded from Web of Science (WOS), PubMed, National Library of Medicine (NLM), and via manual searches in October 2016 and August 2018. Additionally, manual searches were conducted in the NLM catalogue for journal MeSH terms. No language restrictions were applied.

Inclusion criteria for articles included peer-reviewed articles with no limits on: language, study design, and year of publication before 2018. Commentaries, editorial notes, and opinion articles were included. All published articles with microbiome in the text, the journal publication catalogued in NLM, and journal publication catalogued in NCBI with MeSH were included. Inclusion criteria for citations included citations with no limits on language, study design, and publication before 2018. Only cited articles from articles found from the literature search were included. Inclusion criteria for government funded microbiome projects and project data included any project funded by a US government agency prior to 2018 with microbiome in the abstract. Any projects without microbiome in the abstract were excluded.

Full text open-access articles were analyzed, as an emphasis on how the microbiome was explicitly conceptualized and how those conceptualizations changed over time was the focus of this study. This was problematic, as the copyright on articles changed over time, the total number of articles available at any given time was dependent on article copyrights, and many articles only provided abstracts. Different databases, also, had in many cases contained overlapping articles with other databases and had differences in the span of time microbiome research was published, i.e. the first article in JSTOR that matched the criteria was published in 1952 whereas the first article in PubMed that matched the criteria was published in 2002. This brought to light another issue, social contextual factors influencing the microbiome would be focused on biomedicine by collecting articles from PubMed and PubMed Central (PMC) only. Previous studies have shown social networks influence language use, patterns, and diffusion linguistic variables (Labov, 2001; Milroy, 1987; Milroy & Gordon, 2003). Other studies have shown that the diffusion of concepts and innovations is influenced by social capital and trends in use (J. Swan et al., 2007; Rogers, 2010).

To create wide-ranging datasets and overcome these issues, the full text of articles from WOS, JSTOR, PubMed, and PMC along with the associated metadata for each article, and government funded project abstracts on the microbiome and associated metadata were collected. Based on the results of the literature search, articles were collected four times, twice in 2014 and twice in 2018. The collections in 2014 were used to test the speed and efficiency of the collection methods and gathered articles and the associated metadata from articles that matched the criteria of microbiome in the text, published from 1900 to 2014 from JSTOR, Web of Science (WOS), PubMed, and PubMed Central. The collections in 2018, gathered articles and the associated metadata from articles that matched the criteria of microbiome in the text, with publication dates from 1900 to 2017 from JSTOR, WOS, PubMed, and PubMed Central.

Projects were collected twice, once in 2014 and again in 2018. The project collection in 2014, gathered projects and associated metadata from projects that matched the criteria of microbiome in the abstract, with Fiscal Year (FY) dates between 1900 to 2014 from the National Institutes of Health (NIH) Research Portfolio Online Reporting Tools (RePORTer) and the Federal Research Online Reporting Tools (RePORTer). The project collection in 2018, gathered projects and associated metadata from projects that matched the criteria of microbiome in the abstract, with Fiscal Year (FY) dates between 1900 to 2018 from the National Institutes of Health (NIH) Research Portfolio Online Reporting Tools (RePORTer) and the Federal Research Online Reporting Tools (RePORTer).

A hybrid of computational and manual approaches were used to collect, clean, and curate microbiome articles, projects, and metadata. Major inaccuracies and large portions of data were absent from the result of unsupervised data collection from well-known and highly cited biomedical ontologies, journal repositories, and databases. To illustrate, PubMed the

microbiome's representation in articles or how many articles the microbiome occurred in, from 2002 to 2014 microbiome increased by 694% (2 to 1388), but upon closer inspection of the actual use of microbiome in the articles; the word *microbiome* did not actually appear in all 1388 texts provided by PubMed as some texts labeled by curators with keyword microbiome only had *microbiome* in the title, references, or used as part of *The Human Microbiome Project*. After close reading of the texts, many of the texts identified via PubMed discussed microbiota with no mention of the microbiome. Previous results have indicated the PubMed MeSH definitions of microbiome and microbiota are the same, displayed in Figure 2. Operational definitions for microbiome and microbiota are diverse, and the boundaries between microbiome and microbiota may prove the two concepts to be related but they are not equivalent (Prescott, 2017).

Unsupervised results from the WOS had similar issues, as the WOS article count from 1900 to 2017 on the microbiome was 14986, compared to the total published articles collected from JSTOR, WOS, PubMed, and PubMed Central in the MB Corpus being 27,977, a 187% difference in articles, displayed in Table 19.

Results on total funded microbiome projects had similar issues with inaccurate data or data spread among multiple sources. The microbiome has 0 "Total Number of Research/Disease Areas" faded according to NIH categorical spending based on Research, Condition, and Disease Categories (RCDC), which provides results from NLP text mining on the annual number of grants, contracts, and other funding mechanisms for NIH ("NIH Categorical Spending -NIH Research Portfolio Online Reporting Tools (RePORT)," n.d.). Some microbiome data was found but still wildly inaccurate on the Federal RePORTer site, as there was a 187% difference in microbiome funded projects between Federal RePORTer results (5,563 projects) and the results from the MB Project Corpus, which consisted of data from Federal RePORTer combined with

NIH RePORTer (10,401), and a 183% difference in total funding between Federal RePORTer results (\$3,411,212,659) and the results from the MB Project Corpus (\$6,239,782,378), which consisted of data from Federal RePORTer combined with NIH RePORTer, depicted in Table 20. Upon closer inspection, the microbiome did not appear in all project abstracts based on automated query results. Again, a combination of computational and manual cleaning was used to provide the final accurate data for the MB Project Corpus.

Other inaccuracies based on author affiliations, multiple author representations for one author (e.g. K. Aiello, K. D. Aiello, Ken D. Aiello, Kenneth D. Aiello, Kenneth Daniel Aiello), and incomplete information required cleaning of metadata for the MB Corpus. With no standard protocol on how to collect, clean or curate metadata, a worksheet was created to organize collection of metadata. The worksheet collected: PubMed ID, PubMed Central ID, WOS Accession number, article title, publication year, journal title, National Library of Medicine Catalog Medical Subject Heading for journal, abstract and author names (last, name, middle initial) for the first 10 authors, the last author, and up to three affiliations for each author. Metadata for the MB Project Corpus had similar issues and had to be cleaned via a combination of computational and manual cleaning, the MB Project Corpus worksheet collected data on: project abstract, project terms, project title, public health relevance, administering institute or center (IC), application ID, award notice date, funding opportunity announcement, project number, type, activity (R01, R02, etc.), serial number, IC, support year, program official information (last name, first name), project start date, project end date, study section, subproject number, contact principal investigator (PI) ID, contact PI/Project leader, other PI or project leader(s), congressional district, department of PI, data universal numbering system (DUNS) Number, Federal Information Processing Standards (FIPS) Number, organization ID Institutional

Profile File (IPF), organization name, organization state, organization city, organization type, organization zip, organization country, American Recovery and Reinvestment Act (ARRA) indicator, budget start date, budget end date, Catalog of Federal Domestic Assistance (CFDA) Number, funding mechanism, fiscal year (FY), funding IC, FY direct costs, FY indirect costs, FY total cost by IC, FY total cost (sub-projects). After all information was collected the data was reviewed by the Data Mining and Informatics Team at the Laubichler Lab and a final review of the data was tested for accuracy using a combination of supervised and unsupervised methods. The final breakdown of sources contributing to the analyses consisted of 27,977 articles collected compiled into a Microbiome (MB) Corpus, displayed in Table 3. Collection of projects used the same protocol resulted in 10,401 projects, presented in Table 20 .

History of knowledge

Applying the theoretical framework developed by Jurgen Renn and Manfred Laubichler in *Extended Evolution and the History of Knowledge*, the analysis of the history of knowledge requires a perspective of extended evolution (Renn & Laubichler, 2017a). Using results from networks and contexts and the social and material dimensions of knowledge, ‘knowledge’ is the encoded experience of actors which manifests as a mental structure with material and social dimensions, determining which actions are possible in a historical situation, knowledge may be shared within a group or society via the use of material artifacts such as instruments or texts (Renn & Laubichler, 2017a). Shared behaviors connected by cognitive, social and material links are the social relations within a given group or society, and encoded collective experience can be represented by institutions according to this framework (M. Laubichler & Renn, 2015; Renn & Laubichler, 2017a). Previous results using this framework highlighted: origins of variation were the result of the properties of complex systems, context was the material means used by an actor

to reach an action and the result of actions, contexts of action represented knowledge, institutions were used to diffuse and transform regulative structures, networks of human actions included a material and social culture, and knowledge itself was externally represented and shareable (Renn & Laubichler, 2017a).

Material artifacts and the content within publications and projects, and the individuals, groups, institutions, and systems producing the material artifacts can be modeled as actors within this framework. Contextual factors suggesting explanatory variables characterizing specific changes to the microbiome can be identified from actors and determined as influential in driving changes to the microbiome based on usage. Accepted and proliferated changes to the microbiome will be investigated for historical trajectory and patterns. Previous studies have identified specific changes to the usage and meaning of the microbiome were dependent on context and influenced by contextual factors (Shade & Handelsman, 2012; Huss, 2014). At any given moment there is variation in the knowledge of the microbiome and this explains the multiple states of the microbiome concept at one time, but changes toward specific forms and usage of the microbiome in the language context, i.e. words, phrases, and meaning of the microbiome are intentional acts or processes orchestrated by actors (Knorr-Cetina & Mulkay, 1983; Labov, 2001; Owen-Smith & Powell, 2004). Similarly, variations in the social context of who uses the microbiome are performed by specific actors at a specific time but the introduction of variation and adopted changes to microbiome knowledge are the result of the accumulation of differences in the social context including impact, observability, trialability, complexity, compatibility, relative advantage, and adoption rate (Labov, 1972; Halliday & Hasan, 1991; Beckner et al., 2009; Rogers, 2010). Many of the debates surrounding the development and use of the microbiome including: the meaning of the microbiome, who/what influenced changes to

the microbiome, and how knowledge on the microbiome transformed, can be settled by systematically analyzing the contextual factors of the microbiome.

Objectives

Given the variation and changes to the microbiome, the purpose of this experiment is to provide insight into the knowledge of the microbiome, with regards to the following questions:

RQ1: How has the microbiome research outputs developed over time?

RQ2: Is the microbiome an innovation?

RQ 3: What is evidence of the variation and changes to microbiome knowledge?

To answer these questions two different datasets will be utilized, 1) a collection of articles published on the microbiome from 1900 to 2017 with associated metadata or the MB Corpus, 2) US funded projects on the microbiome with associated metadata or the MB Project Corpus. These datasets enable a multi-pronged approach for describing the scientific literature, knowledge base, and variation and change to the microbiome. The goal of this study is to determine the contextual factors influencing changes to the microbiome concept analyzing the variation and changes to the language and social context of the microbiome.

Results

How has microbiome research outputs developed over time?

increases in microbiome research, resources, and knowledge.

Microbiome research outputs, resources, and knowledge as measured by the MeSH terms, and observability and usage of the microbiome measured by the number of authors and co-authors, in general increased over time. The overall trend showed the majority of substantial increases for all research outputs, resources, and knowledge occurred prior to 2009 and smaller percent increases trended for all categories after 2010, displayed in Table 21. Prior to 2005, there was little evidence of the microbiome appearing in the text of articles, journals, government

funded project abstracts, and only a handful authors wrote microbiome articles or cited microbiome research. However, the number of articles, journals, government funded projects, MeSH terms, and authors grows substantially from 2005 to 2017. The largest percent increase for the majority of research outputs and social systems occurred between the years 2004 to 2005: the number of articles increased by 500%, the number of journals increased by 400%, the number of unique MeSH increased by 300%, the total number of MeSH increased by 333%, and number of authors and co-authors increased by 2100%, while the largest increase in citations occurred from 2003 to 2004 increasing by 333%, and government funded projects had the largest increase from 2005 to 2006 increasing by 500%, displayed in Table 22. While the smallest percent increase for all research outputs (excluding the years from 2001 to 2005) occurred between the years 2016 to 2017: the number of articles increased by 4.0 %, the number of projects increased by 18.8%, the number of journals increased by 5.8%, the number of citations increased by 16.6%, the unique number of MeSH terms had a decrease of -0.86%, the total MeSH terms had an increase by 0.45%, though the smallest percent increase for authors occurred from 2015 to 2016 only increasing by 18.6% percent, exhibited in Table 22.

The microbiome as an innovation

The four main elements in the diffusion of innovations are: 1) an innovation, or an idea, practice, or object, which undergoes diffusion; diffusion is the process by which an innovation is communicated through, 2) communication channels over, 3) time, and 4) a social system of actors (Rogers, 2010). In this study, the microbiome is the innovation, and the various channels, or means by which messages get from one actor to another, are represented as articles, citations, projects, journals, and MeSH journal categories. Five attributes known to be characteristic of

adopted innovations are described in detail with respect to the microbiome, these include: relative advantage, compatibility, complexity, trialability, and observability (Rogers, 2010).

relative advantage.

Relative advantage describes how an innovation is observed as better than other ideas and may be measured by economic benefits, and social prestige (Greenhalgh et al., 2005). The concepts metabolome and metagenome were used as comparisons to highlight relative advantage of the microbiome, as the last chapter detailed the confusion around these concepts and previous studies have shown well-funded concepts carry economic advantage and social prestige (Norredam & Album, 2007; Arcaya, Arcaya, & Subramanian, 2015). Results from combined Federal RePORTer and NIH RePORTer projects and funding from 1980 to 2017 on microbiome, metabolome, and metagenome projects, showed the microbiome had a relative advantage in more total projects (10,401) compared to metabolome (1038) and metagenome (772), and more total funding (\$6,239,782,378) compared to metabolome projects (\$516,687,021) and metagenome projects (\$663,858,34) displayed in Table 23.

Highly cited papers published in gatekeeper journals are another indicator of social prestige (Callaham, Wears, & Weber, 2002; Ioannidis, Boyack, & Klavans, 2014). The most highly cited paper on the microbiome was published in a prestigious journal and is highly cited, the article “An obesity-associated gut microbiome with increased capacity for energy harvest,” has been cited over 4,000 times averaging over 300 citations per year since being published in 2006 and was published in the high-profile journal Nature (Turnbaugh et al., 2006). In contrast, the most highly cited article on the metabolome is not as highly cited nor in a journal with a high impact factor, (“Metabolomics – the link between genotypes and phenotypes”) with 1,905 citations (Fiehn, 2002). Likewise, the most highly cited article on the metagenome (“A human

gut microbial gene catalogue established by metagenomic sequencing”) is not as highly cited as the most highly cited microbiome article with only 3,680 citations (Qin et al., 2010)

compatibility.

Compatibility describes the degree to which an innovation is perceived as being aligned with the (social, language, and historical) contextual knowledge of actors within a system (Kapoor et al., 2014). Generally, more compatible ideas fit more closely with an individual’s knowledge and skills, are less uncertain to adopters, and makes it easier for individuals to actively shape the knowledge of an idea (Denis, Hébert, Langley, Lozeau, & Trottier, 2002; Fennell & Warnecke, 2013). The word microbiome is compatible with 43 other words with the suffix -ome according to Joshua Lederberg in ”Ome Sweet ‘Omics—A Genealogical Treasury of Words” (J Lederberg & McCray, 2001). Within biomedicine, the microbiome has also been described as compatible with metagenome and metabolome based on the results from the collocate networks from last chapter, and according to the scientific literature the microbiome is related to the human genome with the microbiome being called the “Second Genome” (Marchesi & Ravel, 2015b; Grice & Segre, 2012). The microbiome vocabulary and ontology has also been found to be compatible across domains with microbiome being interpreted as another instance of a biome, niche, or habitat in ecology (Whipps et al., 1988; Shade & Handelsman, 2012). The compatibility of microbiome knowledge has been extended to individuals who never used the word microbiome in text, as some have attributed Antonie Van Leeuwenhoek’s study of animalcules as instance of microbiome knowledge (Ursell et al., 2012).

complexity.

Complexity of an innovation is the extent to which an innovation is perceived as difficult to understand and summarizes the ability of a social system to comprehend an innovation

(Rogers, 2010). Analysis of microbiome knowledge over time via the ratio of total collocates of the microbiome compared to the occurrences of microbiome in the text, showed the complexity of the microbiome reduced over time. From 2001 to 2010 the number of collocates of the microbiome exceeded the number of occurrences of microbiome, but from 2011 to 2017 the number of occurrences of microbiome exceeded the total collocates, displayed in Table 24. Complexity can also be reduced when an innovation can be fragmented into adaptable parts and adopted incrementally (Meyer & Goes, 1988; Plsek & Plsek, 2003). The microbiome has been broken down into multiple different concepts such as: gut microbiome, intestinal microbiome, lung microbiome, skin microbiome, and vaginal microbiome as argued by experts and supported by the results from the last chapter (Shade & Handelsman, 2012; Huss, 2014; Schneider & Winslow, 2014).

trialability.

Trialability explains how an innovation may be tested or experimented with (Rogers, 2010). Previous studies have shown actors giving meaning to an innovation were testing the trialability of the innovation (Yetton, Sharma, & Southon, 1999). In this regard, knowledge is unique because it provides multiple opportunities for actors to implement, on a limited-time basis, the microbiome concept even if they have had little prior exposure or experience. The creation of knowledge on the microbiome via interpretations, vocabularies, or proxies of knowledge like articles, citations, and projects provides a low-risk occasion for authors, co-authors, and government agencies to experiment with the microbiome concept. Evidence of this experimentation was exhibited through repeated attempts by experts in creating new definitions, new ontologies, or new vocabularies for the microbiome (J Lederberg & McCray, 2001; Ursell et al., 2012; Blaser et al., 2013; Marchesi & Ravel, 2015b). The broad usage of the microbiome

across domains beyond biomedicine was used as additional evidence of the trialability of the microbiome, as the microbiome has been used to describe: habitats like anaerobic environments, ecosystems like cornfields, localities like New York City, microbial communities like biofilm, nice spaces like anoxic waters, and populations like coyotes in Arizona (Shade & Handelsman, 2012).

observability.

Observability is how visible the innovation is to others as perceptibility stimulates awareness and usage (Rogers, 2010). The observability of the microbiome has peaked in relation to events in the historical and social context of the microbiome. The observability for the microbiome was high when The Human Microbiome Project Consortium announced the NIH Human Microbiome Project (HMP) in 2007 and results from the HMP 2012, as confirmed by the citations to the publications of the announcement (3158) and results (3760). Both the announcement and results were published in the high profile journal Nature (Turnbaugh et al., 2007; The Human Microbiome Project Consortium et al., 2012). Other events increasing the observability of the microbiome included: the high profile journal Science named the microbiome the breakthrough of the year in 2011 and 2013, and other research journals dedicated to the microbiome including: Biofilms and Microbiomes by Nature in 2014, Microbiome Journal (biomed Central), Microbiome Science and Medicine, and the Human Microbiome Journal (Science, 2011, 2013). Additional evidence of the observability of the microbiome includes microbiome centers in institutions, industrial investment, and home microbiome analyses advertised in magazines and televisions (Gormley, 2016; “Home | uBiome,” 2018; “VIOME- ga2,” 2018).

adoption rate.

The attributes listed here are interrelated and have contributed to increases in the adoption rate and diffusion of the microbiome within different social systems. The adoption rate of an innovation is innovation specific and system specific. In most cases, an adopted innovation will display an S-shaped curve when the number of adopters of an innovation are plotted over time, shown in Figure 16 (Valente, 1996; Rogers, 2010). The data from the MB Corpus metadata and MB Project Corpus metadata highlighted S-shaped curves or the initial characteristics of S-shaped curves (long tail to the left prior to rate of adoption taking off) for the microbiome in the following systems: publications, projects, journals, citations, project funding, authors and co-authors, total MeSH categories by year, and unique MeSH categories by year, displayed in Figure 17a-h.

Change and variation in knowledge

mesh terms.

The results from the NLM MeSH terms used with the journals for the microbiome shows a large range of knowledge. There were 4889 total unique MeSH terms from 2001 to 2017 and the 54,705 total MeSH terms from 2001 to 2017, presented in Table 21.

topic models.

Results from topic models of the microbiome for each year provided insight into the knowledge of the microbiome at the discourse level and the global language variation of the MB Corpus over time. Microbiome as a word did not appear in a topic from multiple iterations of a 20-topic model of 20 words per topic for the years 2001, 2002, 2003, and 2004, but in topic models of the years 2005 to 2017 the microbiome appeared in an average of less than 2 topics per year. The topics highlighted variation between the latent themes surrounding the

microbiome. From 2005 and 2006 topics on the microbiome consisted of bacterial, abundance, genes, and from 2007 to 2017 topics on microbiome contained words like bacterial, diseases gut, human, host, intestine, microbial, microbiota, and skin dominated, displayed in Table 25.

keywords.

Thematic differences between early years and late years of microbiome research were reinforced from the results of keyword analysis which compared: 1) MB Corpus articles from 2001 to 2006 (MB Corpus Pre-2007) and, 2) articles from the MB Corpus post 2007 to 2017 (MB Corpus Post 2007). The top statistically significant keywords (using Log-likelihood) based on differences in the frequency in the MB Corpus Pre-2007 compared to the MB Post-2007 Corpus were examined and used to create KWIC lists. Keywords and Keyword in Context (KWIC) lists were created with the MB Pre-2007 Corpus as the corpus of analysis and the MB Post 2007 as the reference corpora, and vice versa using Wordsmith Tools (M. Scott, 2018). Differences in tendencies between each corpus were calculated using Log-likelihood, which showed positive keywords for the MB Pre-2007 Corpus in comparison to the MB Post-2007 Corpus. Negative keywords were not used.

The KWIC lists were useful preliminary word lists used to help isolate language patterns specific to each corpus and assisted in identifying differences between the language context and discourse of each corpus. Differences in knowledge at the word level characterized by tendencies of word usage within these large bodies of text were identified using KWIC lists in combination with a keyword cluster list, which enabled examination of important and meaningful comparative aspects. Categories of keywords were created as a result of applying a coding scheme to the keywords and keyword clusters, with evidence supported by the frequency lists

and concordances of the keywords to evaluate contextual language use (D. H. Hymes & Gumperz, 1972; D. Hymes, 2013).

The most noticeable differences in tendencies between the MB Pre-2007 Corpus and MB Post-2007 Corpus based on selected keyword categories, exhibited in Table 26. Validation of keywords and keyword categories between corpora were supported by frequency analyses and concordance. In the patient/population category, the MB Pre-2007 Corpus resulted in keywords associated with references to larger organisms like *human*, and *humans*, *mouse*, *mice*, and *animals* as subjects; whereas the keywords of patient/population of MB Post-2007 research resulted in different age-based descriptions of humans or keywords describing relationships between humans with references like *children* and *maternal*, gender categories of humans such as *women*, and words commonly used to describe humans as a unit of analysis like *cohort*, *participants*, *patients* and *subjects*. There was an increase in the range and emphasis of diseases between corpora, as the MB Pre-2007 Corpus only disease was *enteritis*, compared to MB Post-2007 Corpus which had *disease* as a keyword and other diseases including: *clostridium difficile*, *colorectal (cancer)*, *depression*, *diabetes*, *dysbiosis*, *ibs* and *infection*. An emphasis on areas of the body were in both corpora, but the MB Pre-2007 emphasized the gastrointestinal system with keywords like *cecal*, *distal*, *gastric* and *intestine*, while the MB Post-2007 Corpus keywords gravitated towards multiple areas in the body such as: *blood*, *bowel*, *kidney*, *liver*, *skin*, *stool* and *vaginal*, also the MB Post-2007 Corpus emphasized ecological based microbiomes such as *biofilm*, *fungus*, *marine*, *root*, *soil*, and *water*.

scholarly network.

There was increased variation and drastic changes in the social context of the microbiome as visualized in a scholarly network of authors and co-authors (red nodes) linked to publications

(green nodes), Figure 18a-q. The scholarly network connected authors and co-authors to their articles and allowed for authors to be connected to multiple articles at one time. The scholarly network revealed scientific activity on the microbiome was limited and isolated to a small number of authors and their articles from 2001 to 2005, as there were no connections linking authors to multiple papers and all papers had less than three co-authors, visualized in Figure 18a-d. The interconnectivity increased starting in 2005, as among all papers there were more papers with three or more co-authors and less with two authors, displayed in Figure 18e. The interconnectivity continued to grow over time as every year witnessed increases in authors, articles, number of authors per article, and authors linked to multiple papers, demonstrated in Figure 18f-q.

collocates.

The knowledge and meaning of the microbiome emphasized two primary conceptualizations from 2001 to 2017, a *human microbiome* and a *gut microbiome*. The knowledge of *human* and *gut* microbiome accounted for a significant amount of usage, as 22% of all times *human* was used in the MB Corpus it was used with microbiome, and 19% of all times *gut* was used it was used with *microbiome*, based on strength of association within a span of 8 words (4 words to the left or right) and calculated by the overall frequency of the word with microbiome within that span divided by the how many times microbiome occurs in the corpus, presented in Figure 19 (Stubbs, 1995, 2001; M. Scott, 2018) . Other interpretations of the microbiome emerged from 2001 to 2017 including: *core microbiome*, *disease microbiome*, *healthy microbiome*, *infant microbiome*, *liver microbiome*, *lung microbiome*, *oral microbiome*, *kidney microbiome*, *role microbiome*, *root microbiome*, *skin microbiome*, *soil microbiome*, and

vaginal microbiome; but all these interpretations averaged less than 1% of occurrences used with microbiome, displayed in Figure 20.

stability.

When surveying the field of microbiome research, it is continually important to keep in mind that the concept, while gaining in popularity, does not connote the same meaning in every field or subfield. It is expected a variety in production to produce a variety in usage as the publications on the microbiome span a wide variety of journals. There is not a guaranteed mechanism to enforce any consistency across journals because papers are published from a wide variety of journals, among different actors, housed in different organizations, and with different interests and goals. The stability of the journal corpora indicated that some influence causes the collocates to change over time and a larger stability coefficient indicated that the collocates in two time periods are more similar, demonstrating the similarity of knowledge between the two corpora.

Microbiome knowledge demonstrated increased stabilization over time within the MB Corpus. The stability was measured by the Jaccard similarity score of the top 20 shared collocates between consecutive years in the MB Corpus, i.e. top shared 20 collocates of 2001 compared to top shared 20 collocates of microbiome in 2002, top 20 collocates of 2002 compared to the top 20 collocates of 2003, and so on. The Jaccard Similarity (J) has been widely used to determine the variation and changes of group and network dynamics based on the overlap between networks over time (P. Jaccard, 1912; J. Jaccard, Wan, & Turrisi, 1990; Wasserman & Faust, 1994; Li, An, Wang, Huang, & Gao, 2016). The similarity for the microbiome increased starting in 2005 to 2006, underwent one decrease between 2013 to 2014,

but then continued to increase exhibiting an extremely high average of $J= 41.2$ from 2001 to 2017 and a $J=$ of 90% from 2015 to 2017, displayed in Table 27.

The variation in knowledge of other specific social contexts cited to influence knowledge were determined by similar analyses on corpora created from journal and time based slices of the MB Corpus (Rowlands, 2002; J. Scott & Carrington, 2011; Loet Leydesdorff, 2013). Each corpus was created from all articles published by that journal from 2001 to 2017, even though some journals did not publish in every year from 2001 to 2017, the following corpora were created: PLOS One from 2008 to 2017, Nature from 2006 to 2017, Science in 2001 then 2005 and from 2007 to 2017, and a corpus based on NIH Project abstracts from 2005 to 2017. PLOS One and the NIH Project abstracts proved to have the most stable knowledge surrounding the microbiome of all the other contexts analyzed, with an average $J=38.7$ from 2008 to 2017 and $J=38.7\%$ from 2004 to 2017 respectively, while Science and Nature exhibited little knowledge stability on the microbiome with an average $J=8.2\%$ from 2005 to 2017 in the Science Corpus and $J=13\%$ from 2006 to 2017 in the Nature Corpus (the Science Corpus did not have publications in the MB Corpus in 2002 to 2004 and in 2006, so 2001 was compared to 2005, and 2005 was compared to 2007, as they were the consecutive years of published articles), displayed in Table 27 and plotted in Figure 21.

Major changes in the characteristic knowledge of the microbiome occurred in all contexts over time. The similarity for all contexts was less than 15% similar when the first year was compared to the last year for each corpus. Comparisons of the top 20 collocates from the first year for each corpus compared to the last year top 20 collocates for each year for each corpus: MB Corpus 2001 to 2017, PLOS One Corpus 2008 to 2017, Science Corpus 2008 to 2017,

Nature Corpus 2006 to 2017, NIH Abstract Corpus 2004 to 2017, showed 2.8%, 13.5%, 2.5%, and 13% similarity respectively, presented in Table 28.

convergence.

Comparison of the both the variable form or state of MB knowledge and the final form compared to the trajectory of knowledge, or convergence, provided insight into the influence of a specific journal over the microbiome knowledge over time. Convergence is the tendency of knowledge characteristics to evolve superficially similar collocates under similar environmental conditions (J. Jaccard et al., 1990; Wasserman & Faust, 1994; Li et al., 2016). Convergence for each year was determined by comparing the similarity between a one-year slice of a corpus to the collocates from the MB Corpus for each year, i.e. MB Corpus 2008 compared to Science Corpus 2008, MB Corpus 2009 compared to Science Corpus 2009, and so on, displayed in Figure 22. Convergence of the final form of microbiome knowledge compared to other social contexts was calculated using the Jaccard similarity score of the collocates from the MB Corpus 2017 compared to every year in each of the corresponding corpora, i.e. MB Corpus 2017 compared to PLOS One 2008, MB 2017 compared to PLOS One 2009, and so on, with results displayed in Figure 23.

The results showed PLOS one had unusually high convergence with the MB Corpus for each year and with the MB 2017 collocates despite the total representation of PLOS One articles decreasing over time. PLOS One articles were less than 1% of the entire MB Corpus in 2016 and 2017 but had a high collocational similarity (60%) to the MB Corpus. Science initially had a high convergence of knowledge with the microbiome but, over time the similarity between the two corpora grew and never recovered past 25%. Nature had two peaks of convergence with the microbiome in 2007 and 2016, but both never reached over 50% with the MB Corpus. NIH

Project abstracts had increased convergence with the MB Corpus, but the highest convergence with the MB 2017 was a 60% similarity to the MB Corpus collocates in the years 2015 and 2016.

Discussion

The change and variation of the microbiome was examined by using computational and manual methods to evaluate the influence of contextual factors on subsequent proxies of knowledge within a tradition. Emergent trends and transient patterns of the microbiome research field were discovered from the systematic analysis of contextual factors. We identified key factors influencing the microbiome based on differences in language, social, and/or historical context. A synthesis embracing the ambiguities and complexities of microbiome research was created by unraveling and aggregating the different uses of the microbiome specific to context. Evidence supporting and launching new narratives within microbiome research became clear through systematic analysis of the contextual factors.

It was discovered the microbiome is an innovation based on the relative advantage, compatibility, complexity, trialability, observability, and rate of adoption, conclusions that are in accordance with previous studies on the innovation in biomedicine, healthcare, and other domains (Greenhalgh et al., 2005; Rogers, 2010). This link between action and knowledge is critical, as an innovation involves both new knowledge and the decision to adopt, because someone may have known or have experienced an innovation for some time but have not acted (rejected or accepted) the innovation yet. Diffusion of innovations, simply helps characterize communication and relates the language context of new ideas and their meaning to the social context and processes evident by adoption and social change. In diffusion theory and research, other attempts have yielded understanding into the history and spread of knowledge of: new drugs, risk factors for a disease, hybrid corn, concept of preventive addiction, new administrative

practices, from a range of different social categories at different scales including: individuals, farmers, physicians, public health officers, health professionals, and other formal and informal groups, to entire states or nations (Valente, 1996; Yetton et al., 1999; Rogers, 2010; Kapoor et al., 2014). Generally, modeling the diffusion of innovation describes if an idea is accepted, directed, and managed, or disseminated, by measuring the subsequent social changes. These social changes can be seen with the S-curve and if communication is the process by which scientists and researchers create and share knowledge, then communication can be modeled using text networks, knowledge networks, co-authorship networks, co-citation networks or a combination of these networks (Valente, 1996; Fennell & Warnecke, 2013). The microbiome was characterized as an innovation according to the attributes of adopted innovations and behavior of microbiome research outputs following Roger's model (Rogers, 2010). A robust and in-depth contextual review of influential factors added value to the qualitative and quantitative analyses.

What is striking about the development of the microbiome is the volume of the research outputs, range of actors involved, and variety of usage across different contexts. Rather than emerge from antecedent science of the 19th and 20th centuries as a coherent object of scientific study, the microbiome has proliferated across medicine, microbiology, genetics, immunology, and more. Moreover, the variety of usage detected by collocate analysis indicates the functional definition of the term "microbiome" changes across both discipline and publication.

Due to the rapid growth of medical information, data, and knowledge there is a need for novel research approaches to measure conceptual change, changes in knowledge, and how contextual factors influence knowledge. Systematic reviews and ontologies like MeSH terms by design create conformity and maintain regularity within biomedicine. Systematic reviews

provide syntheses of information summarized and condensed into a single article, and MeSH serves an important role in indexing the biomedical literature through updated and controlled vocabulary, subject content, and biomedical information. However, in order to gain insight into how scientific innovation and knowledge formation occurs other avenues of engaging with the literature are becoming more necessary due to new dynamics within science, specifically open access publishing and the PLOS One model of publishing peer reviewed articles.

The results show interesting patterns for PLOS One. PLOS One stood out as unusually prolific as PLOS One led all journals publishing articles on the microbiome from 2008 to 2016, 2017 was the only year PLOS One did not publish the most articles but ended up publishing 309 articles representing the third most articles published on the microbiome. Thus, we observe PLOS taking a pivotal role in research publications about the microbiome, specifically conversations around medicine and human health. PLOS One contributes to the operational definition of microbiome as a feature of the human body that matters to the science of human health and medicine. Nevertheless, the entirety of recent microbiome research is not concentrated on applications for medicine and health. PLOS One's topical connection to the microbiome knowledge and quantity of publications make it an influential force in contemporary research on the microbiome, but publications related to medicine account for only a fraction (albeit a significant one) of the overall research corpus.

When surveying the field of microbiome research, it is continually important to keep in mind that the concept, while gaining in popularity, does not connote the same meaning in every field or subfield. The publications on the microbiome span a wide variety of journals. Because papers are published from a wide variety of journals--and there is not a guaranteed mechanism to enforce any consistency across them--we can expect a variety in production to produce a variety

in usage. Forces such as PLOS One and NIH ontologies may in the end bring more consistency, but at this point the full significance of the term is still being discovered. We are witnessing a bloom in activity that has been generative for scientific research of many types, not a culmination of longstanding biomedical discourse.

The MB Corpus and the MB Project Corpus illustrates how the availability of data has increased the amount of information that is available for analysis. This availability is a result of efforts by government agencies to become more transparent in how they fund science (Bertot, Jaeger, & Grimes, 2010; Collins & Tabak, 2014; “Enhancing Reproducibility through Rigor and Transparency,” n.d.). Also, contributing is a rise in open access journals like PLOS One as other publishers and journals have made efforts to create similar research outlets, i.e. Nature Communications, and Science Advances. Further, both government agencies and journals provide access to large repositories of data and metadata through application programming interfaces (APIs).

Discussion of this work with peer reviewers and experts on biomedical knowledge have helped articulate possible dynamics at work which have thrust PLOS One to be at the bleeding edge of research and knowledge in the case of the microbiome. PLOS One’s unique model of publishing allows authors to publish articles quickly in a broad range of disciplines based on scientific merit and not subjective judgements of impact (Savage & Vickers, 2009; Alsheikh-Ali et al., 2011). Being judged on scientific merit allows authors to contribute to compatible or interrelated research areas and experiment with different subjects for low-cost with high reward (Savage & Vickers, 2009; Rogers, 2010). Compared to other journals, every article in PLOS One is open access and freely observable to others and allows for science to be organized and communicated across contexts. Authors can create stronger applications for scientific funding via

increased publication and citation rates of their work by publishing multiple articles in PLOS One in a shorter time compared to the amount of time and knowledge required to publish an article in traditional scientific journals. The PLOS One publishing model is leading a new dynamic within science which allows for huge variation and a large variety of domains to publish original research in one place, greater influence of open-access journals on research and funding, and innovative methodological approaches and results that push research results across scientific disciplines.

The results from this study indicate scientists are still determining the definition and significance of the microbiome, and it is far from a singular concept. Further, these results combined with the results from the previous chapter suggest the microbiome concept and microbiome knowledge is a phenomena with fuzzy boundaries better interpreted as a spectrum of differences across different contexts. Understanding the variation and changes to this spectrum provides insight into the development of the microbiome according to specific influences. Microbiome knowledge from 2001 to 2017 as a spectrum is presented in Table 29, using the top shared collocates for each year from the MB Corpus. Comparing this spectrum of microbiome knowledge to a different context, displayed in Table 30, highlights the differences/similarities and convergence/divergence of microbiome knowledge between the two social contexts and provides insight into the influence and direction of microbiome knowledge.

Figures and Tables

Table 18. Total projects and funding for microbiome research from Federal RePORTer.

Fiscal Year	Total Projects Federal Reporter	Total Funding Federal Reporter All Agencies
2004	0	0
2005	0	0
2006	0	0
2007	0	0
2008	40	\$33,725,501
2009	135	\$99,717,036
2010	191	\$149,858,593
2011	238	\$144,740,020
2012	346	\$272,586,985
2013	431	\$248,898,136
2014	665	\$402,858,683
2015	894	\$550,308,247
2016	1220	\$694,093,968
2017	1403	\$814,425,490
Total	5563	\$3,411,212,659

Table 19. WOS Articles compared to articles collected.

Year	Articles Collected	Web of Science Collection
1952	1	0
2001	3	0
2002	2	2
2003	1	0
2004	1	1
2005	6	3
2006	16	6
2007	47	19
2008	185	75
2009	266	126
2010	325	202
2011	846	405
2012	1472	717
2013	2135	1180
2014	3169	1681
2015	5636	2520
2016	6596	3495
2017	6862	4554
Total	27977	14986

Table 20. Federal RePORTer microbiome projects compared to MB Project Corpus.

Fiscal Year	Total Projects Federal Reporter	Total Projects MB Project Corpus	Total Funding Federal Reporter All Agencies	Total Funding MB Project Corpus
2004	0	1	0	\$592,593
2005	0	1	0	\$688,680
2006	0	6	0	\$2,628,540
2007	0	16	0	\$5,991,477
2008	40	75	\$33,725,501	\$49,677,745
2009	135	259	\$99,717,036	\$179,547,050
2010	191	365	\$149,858,593	\$270,342,632
2011	238	445	\$144,740,020	\$251,596,328
2012	346	648	\$272,586,985	\$499,312,243
2013	431	807	\$248,898,136	\$452,924,952
2014	665	1227	\$402,858,683	\$745,570,649
2015	894	1660	\$550,308,247	\$1,020,186,492
2016	1220	2235	\$694,093,968	\$1,238,106,324
2017	1403	2656	\$814,425,490	\$1,522,616,673
Total	5563	10401	\$3,411,212,659	\$6,239,782,378

Table 21. Research Outputs, NLM MeSH, and authors over time.

Year	Publications	Projects	Journals	Citations	Unique total of NLM MesH	NLM MeSH	Authors
2001	3	0	2	0	2	3	5
2002	2	0	2	0	2	2	5
2003	1	0	1	3	2	2	3
2004	1	1	1	13	3	3	1
2005	6	1	5	40	12	13	22
2006	16	6	10	136	17	30	68
2007	47	16	31	386	28	87	209
2008	185	75	104	980	98	339	782
2009	266	259	136	2001	146	557	1218
2010	325	365	162	3834	172	642	1693
2011	846	445	326	6731	262	1722	4038
2012	1472	648	480	12090	345	3012	7208
2013	2135	807	620	19793	464	3919	10476
2014	3169	1227	863	30879	633	6210	16287
2015	5636	1660	1325	46742	857	11344	26239
2016	6596	2235	1436	65452	927	13380	31124
2017	6862	2656	1519	76306	919	13440	44681
Total	27568	10401	7023	265386	4889	54705	144059

Table 22. Percent increase over time.

Year	Articles	Projects	Journals	Citations	MeSH Unique	MeSH Total	Authors
2001			0.00	0.00	0.00	0.00	0.00
2002	-33.33	0.00	0.00	0.00	0.00	-33.33	0.00
2003	-50.00	0.00	-50.00	0.00	0.00	0.00	-40.00
2004	0.00	0.00	0.00	333.33	50.00	50.00	-66.67
2005	500.00	0.00	400.00	207.69	300.00	333.33	2100.00
2006	166.67	500.00	100.00	240.00	41.67	130.77	209.09
2007	193.75	166.67	210.00	183.82	64.71	190.00	207.35
2008	293.62	368.75	235.48	153.89	250.00	289.66	274.16
2009	43.78	245.33	30.77	104.18	48.98	64.31	55.75
2010	22.18	40.93	19.12	91.60	17.81	15.26	39.00
2011	160.31	21.92	101.23	75.56	52.33	168.22	138.51
2012	74.00	45.62	47.24	79.62	31.68	74.91	78.50
2013	45.04	24.54	29.17	63.71	34.49	30.11	45.34
2014	48.43	52.04	39.19	56.01	36.42	58.46	55.47
2015	77.85	35.29	53.53	51.37	35.39	82.67	61.10
2016	17.03	34.64	8.38	40.03	8.17	17.95	18.62
2017	4.03	4.03	4.03	4.03	4.03	4.03	43.56

Table 23. Comparison of projects and funding between microbiome, metabolome, and metagenome.

Agency	MB Projects	MetaB Projects	MetaG Projects	MB Total Funding	MetaB Funding	Meta G Funding
National Institutes of Health	4,157	934	606	\$6,011,549,139	\$451,675,395	\$581,779,891
National Science Foundation	250	48	129	\$104,721,581	\$29,662,353	\$52,297,364
National Institute of Food and Agriculture	99	15	16	\$20,476,167	\$4,346,238	\$2,369,470
Veterans Affairs	51	12	4	\$0	\$0	\$0
Congressionally Directed Medical Research Programs	40	15	5	\$32,898,326	\$13,348,022	\$6,813,010
Agricultural Research Services	24	6	10	\$42,793,196	\$8,689,157	\$19,880,622
Food and Drug Administration	8	3	0	\$272,500	\$1,200,000	\$0
Center for Disease Control and Prevention	7	3	0	\$18,026,463	\$6,215,918	\$0
Agency for Healthcare Research and Quality	6	0	0	\$2,518,326	\$0	\$0
Environmental Protection Agency	5	1	1	\$5,990,583	\$799,938	\$597,987
National Aeronautics and Space Administration	3	1	1	\$536,097	\$750,000	\$120,000
Total	4,650	1,038	772	\$6,239,782,378	\$516,687,021	\$663,858,344

Table 24. Comparison of microbiome occurrences to microbiome collocates.

Year	Total Words	Total Types	Total Occurrences of MB	Total Collocates (LL)	Total Articles
2001	16,864	2,894	16	31	3
2002	16,189	3,370	6	11	2
2003	48,086	2,690	12	19	1
2004	11,834	2,478	5	4	1
2005	78,293	6,651	43	49	6
2006	278,307	13,872	120	106	16
2007	515,514	25,110	333	499	47
2008	1,904,023	56,590	1,255	1,433	185
2009	2,793,486	72,540	2,286	2,268	266
2010	3,842,638	83,722	1,688	1,716	325
2011	9,608,905	153,843	6,573	3,958	846
2012	14,505,460	216,084	13,901	6,714	1472
2013	21,287,622	237,300	17,756	7,442	2135
2014	34,826,976	317,759	23,076	9,222	3169
2015	36,930,092	484,582	39,400	15,188	5636
2016	44,067,576	565,711	54,193	19,230	6596
2017	51,486,588	417,827	74,196	20,909	6862

Table 25. Topic model results on MB Corpus 2005 to 2017.

2005	abundance	assays	bacterial	gene	genes	manure	manures	microbiome	per	present	primer	real-time	resistance	rfp	samples	sequences	standards	swine	tet	tetracycline
2006	analysis	bacterial	clone	concentrations	diversity	dna	envion	libraries	microbial	microbiology	microbiome	primers	riboosomal	rfs	samples	sequences	standards	swine	tel	tetracycline
2006	abundance	bacteria	cell	community	diversity	genes	gut	host	human	lean	microbial	microbiome	microbiota	number	rst	rsfs	sars-v	sequence	sequencing	unclassified
2007	author	communities	community	datasets	diversity	environmental	genes	genome	gut	human	july	manuscript	metagenomic	microbial	microbiome	microbiota	nature	nhi-pa	pubmed	turkbanigh
2007	adults	aeruginosa	bacteria	cog	colonized	gpf	fig	gordon	gut	host	intestinal	intestine	microbial	microbiomes	microbiota	mutant	pk	responses	stran	zebrafish
2008	analysis	complex	function	functional	genetic	gut	high	human	including	microbial	microbiome	molecular	nature	number	potential	research	sci	species	studies	usa
2009	abundance	author	bacterial	communities	community	diversity	fig	gut	human	individuals	manuscript	microbial	microbiome	microbiota	nhi-pa	sample	samples	sequences	skin	supplementary
2010	abundance	bacterial	communities	community	composition	diversity	fecal	firmicutes	genes	gut	host	human	mice	microbial	microbiome	microbiota	phyotypes	samples	sequencing	turkbanigh
2011	antibiotic	approaches	bacteria	biology	clinical	development	disease	diseases	genetic	health	human	microbial	microbiome	molecular	research	resistance	science	studies	systems	understanding
2011	author	bacterial	cell	figure	final	function	gene	host	human	january	manuscript	microbiome	nih	nhi-pa	page	pnc	pubmed	significant	specific	studies
2011	analysis	bacterial	communities	community	cultured	diet	fecal	fig	genes	gut	human	mice	microbial	microbiome	microbiota	rna	sample	samples	species	time
2012	bacteria	bacterial	communities	community	composition	disease	diversity	fecal	gut	host	human	intestinal	microbiota	microbiota	microbiota	microbiota	pmid	samples	species	studies
2012	author	data	department	gene	hmp	human	information	manuscript	medicine	microbial	microbiome	nhi-pa	page	pnc	pubmed	samples	sequencing	supplementary	university	watermark-text
2013	bacteria	bacterial	bacteroides	colonization	composition	diet	disease	fecal	gut	healthy	host	human	intestinal	mice	microbes	microbial	microbiome	microbiota	species	tract
2014	bacteria	bacterial	composition	diet	dietary	fecal	group	gut	host	human	intestinal	lactobacillus	microbiota	microbiota	microbiota	microbiota	microbiota	probiotic	probiotics	species
2014	biology	data	department	disease	doi	environmental	genetic	health	human	information	microbial	microbiome	population	research	science	studies	systems	time	understanding	university
2015	bacteria	community	data	doi	environment	health	host	human	interactions	microbes	microbial	microbiome	model	nature	pubmed	research	species	studies	system	time
2015	bacteria	bacterial	composition	diet	asthma	birth	children	development	early	gut	human	infants	life	maternal	microbiome	microbiota	pregnancy	probiotic	risk	study
2015	age	allergic	allergy	asthma	birth	children	effects	development	early	gut	human	infants	life	maternal	microbiome	microbiota	pregnancy	probiotic	risk	study
2016	bacteria	bacterial	chronic	clinical	disease	healthy	hiv	human	infection	infectious	lung	microbiome	oral	patients	respiratory	species	streptococcus	study	vaginal	women
2016	approach	data	genetic	health	human	individual	information	interactions	metabolic	methods	microbial	microbiome	model	models	nature	network	research	science	studies	time
2016	bacteria	bacterial	composition	diet	disease	fecal	gut	health	healthy	host	human	intestinal	lactobacillus	mice	microbes	microbial	microbiome	microbiota	species	studies
2017	gut	microbiota	microbiome	human	microbial	intestinal	information	host	healthy	host	human	intestinal	lactobacillus	studies	microbes	microbial	microbiome	microbiota	healthy	studies
2017	abundance	analysis	bacterial	community	data	diversity	dia	fig	microbial	disease	composition	otus	relative	rna	samples	samples	sequencing	species	study	table
2017	age	allergy	asthma	birth	children	development	early	exposure	infant	infants	life	lung	maternal	microbiome	microbiome	pregnancy	respiratory	risk	studies	study

Table 26. Comparison of significant keywords MB Corpus Pre-2007 to MB Post-2007.

Categories of Keywords	MB Articles Before 2007	MB Articles Post-2007
Patient/Population	animals, clones, gnotobiotic, human, humans, hybrid, indigenous, mice, mouse	children, cohort, infants, maternal, otu (operational taxonomic unit), participants, patients, rats, subjects, women
Approaches	molecular, energy, biology, systems, physiology, radiation, probe, laser	analysis, antibiotic, clinical, control, correlation, counts, index, linear, pcoa, profile, qtime, qpcr, ratio, sampling, sequencing, Shannon (diversity index), statistical, surgery, test, treatment, trial, unifrac
Disease/Disorder/Condition	enteritis	(clostridium) difficile, chronic, colorectal (cancer), depression, diabetes, disease, dysbiosis, ibs (irritable bowel syndrome), infection, syndrome
Symptoms		inflammation, irritable, stress, symptoms
Process	aggregate, apoptotic, commensal, evolution, harvest, harvested, mutualism, secreted, symbiosis, tolerance, transcriptional, transport	abundance, detected, exposure, growth, identified, intervention, oxidative, pregnancy, reported, risk, therapy
Unit of Analysis	adenocarcinoma, bacteroides, capillary, carbohydrate, enzymes, flora, genechip, genes, genome, genotype, germ, glyccan, glycoside, heparin, homologs, hydrolases, lepin, longum (Bifidobacterium), (Desulfovibrio) piger, mannose, membrane, mesenchymal, molybdenum, polysaccharide, pylori (Helicobacter), symbiont	bile, biofilms, concentrations, data, fish, fluid, genus, il (interleukin), metabolites, microbiota, mitochondrial, oil, phylum, placebo, rna, sp., streptococcus, taxa
Location/Environment	cecal, distal, endothelial, epithelium, gastric, habitat, intestine, niches	biofilm, blood, bowel, fecal, fungal, kidney, liver, lung, marine, root, skin, soil, stool, vaginal, water

Table 27. Knowledge stability over time measured by Jaccard similarity score.

Years	MB <i>J</i>	PLOS One <i>J</i>	Science <i>J</i>	Nature <i>J</i>	NIH <i>J</i>
2001, 2002	0	N/A	N/A	N/A	N/A
2002, 2003	0	N/A	N/A	N/A	N/A
2003, 2004	0	N/A	N/A	N/A	N/A
2004, 2005	0	N/A	N/A	N/A	100
2005, 2006	15.6	N/A	8.6	N/A	31.6
2006, 2007	17.6	N/A	0	13.7	42.9
2007, 2008	37.9	N/A	0	14.3	14.3
2008, 2009	42.3	13.5	7.1	14.8	21.2
2009, 2010	42.3	27.6	8.1	11.1	20.8
2010, 2011	48.1	28.6	2.5	8.1	53.8
2011, 2012	51.9	37.9	5.2	11.1	11.1
2012, 2013	78.3	53.8	10.3	8.1	8.1
2013, 2014	70.8	48.1	14.3	8.1	11.1
2014, 2015	73.9	42.9	10.3	11.1	48.1
2015, 2016	90.5	48.1	14.3	25	66.7
2016, 2017	90.4	48.1	17.6	17.6	73.9

Science* Slices from the corpus compared 2001 to 2005 because no articles were published in 2002 to 2004, and 2005 compared to 2007 because no articles were published in 2006.

Table 28. Change in knowledge measured by Jaccard similarity score.

Corpus	Years	<i>J</i>
MB Corpus	2001, 2017	2.8
PLOS One	2008, 2017	13.5
Science	2008, 2017	2.5
Nature	2006, 2017	6.5
NIH	2004, 2017	13

Table 29. Spectrum of shared microbiome collocates in MB Corpus.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
anatomy	characterization	define	insect	animal	access	aggregate	analyse	analyse	analyse	analyse	analyse	analyse	across	alter	alter	alter	alter
base	disease	exceed	collective	collective	access	capability	capacity	capacity	capacity	bin	capacity	change	alter	analyse	change	alteration	analyse
billion	healthy	gene	contain	contain	capacity	capacity	cecal	change	change	capacity	change	core	analyse	change	analyse	analyse	base
collective	human	select	dna	colonic	collective	cecum	core	core	core	core	core	disease	change	core	change	change	change
contain	state	shotgun	gene	community	community	chicken	chicken	distal	distal	distal	diet	distal	core	disease	core	core	core
describe		total	genetic	contain	contain	core	core	diversity	diversity	diversity	distal	diversity	disease	diversity	disease	disease	disease
gene		type	genome	distal	distal	diet	describe	diet	energy	energy	energy	fecal	diversity	gut	diversity	diversity	diversity
genome			gut	estimate	estimate	distal	distal	distal	gut	fecal	gut	gut	gut	healthy	fecal	gut	gut
intestinal			host	feature	feature	energy	energy	energy	host	foregut	healthy	healthy	healthy	host	gut	healthy	healthy
million			human	fecal	fecal	enrich	genome	enrich	human	human	host	human	host	human	healthy	host	host
poorly			intestinal	genome	genome	gut	gut	gene	humanize	genome	human	infant	human	intestinal	host	human	human
project	physiology		physiology	give	give	healthy	healthy	gut	increase	gut	increase	intestinal	intestinal	lung	intestinal	infant	infant
remain	present		present	gut	gut	human	human	host	intestinal	human	infant	lean	lung	obese	human	infant	infant
rna	provide		provide	human	human	increase	increase	human	lean	human	lean	obese	obese	oral	intestinal	intestinal	influence
suggest	sample		sample	infer	infer	library	library	increase	metabolic	lean	obese	oral	obese	project	obese	obese	obese
term				initiative	initiative	metabolic	metabolic	metabolic	mouse	metabolic	oral	sample	oral	research	oral	oral	oral
	microbial		microbial	microbial	microbial	microbial	microbial	mouse	obese	obese	skin	skin	project	rumen	project	project	project
	obese		obese	mouse	mouse	mouse	mouse	obese	rumen	primate	skin	skin	skin	skin	role	role	role
	roughly		roughly	obese	obese	obese	obese	sample	twin	reveal	twin	twin	study	study	skin	skin	skin
	time		time	trait	trait	western	western	type	type	twin	vaginal	vaginal	twin	vaginal	study	study	study

Table 30. Spectrum of shared microbiome collocates in Nature articles on the microbiome.

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	
assignment	above	capacity	aggregate	core	complex	analyse	define	active	alter	alter	alter	analyse
capacity	capacity	cell	base	crispr	consider	base	detail	alter	alter	alteration	alteration	bacterial
cecal	characterize	disease	compare	cross	deficient	consortium	disruption	analyse	analyse	alteration	alteration	consortium
compare	common	distal	component	data	develop	data	diversity	base	cancer	base	base	data
encode	concept	energy	core	extract	distal	diversity	dynamic	cell	cirrhosis	change	change	disease
group	core	gut	distal	fecal	dynamic	framework	follow	diet	function	diet	diet	dynamic
gut	data	host	find	genbank	gene	function	genome	dive	gatherer	effect	effect	earth
human	enrich	increase	gut	gut	gut	gene	geography	gene	gut	gene	gene	enrich
increase	functional	microbe	human	lean	human	gut	gut	gut	healthy	healthy	gut	fecal
indicate	gut	mouse	individual	least	include	human	healthy	human	human	healthy	healthy	gut
obese	healthy	obese	lean	obese	lean	include	human	immune	human	human	human	healthy
pathway	human		microbe	pool	microbiota	jumpstart	infection	link	hunter	hunter	immune	human
run	increase		obese	program	mouse	press	initiative	mammalian	innovative	innovation	influence	profile
	involve		obese	prophage	obese	reference	myriad	measure	mediate	profile	link	profiling
	metabolic		revel	read	predict	research	nature	mediate	project	metagenome	module	project
	microbial		sequence	remain	repletion	review	obese	obese	say	metagenome	module	project
	mouse		size	search	return	sample	prominent	phylogeny	signature	mouse	mouse	study
	present		support	spacers	transplant	set	revel	rapidly	study	obese	obese	use
	trait		total	unclear	vaccine	sturdy	skin	respond	suggest	project	project	vaginal
	view		twin	virome	wonderful	view	view	shift	taxonomic	study	study	variation

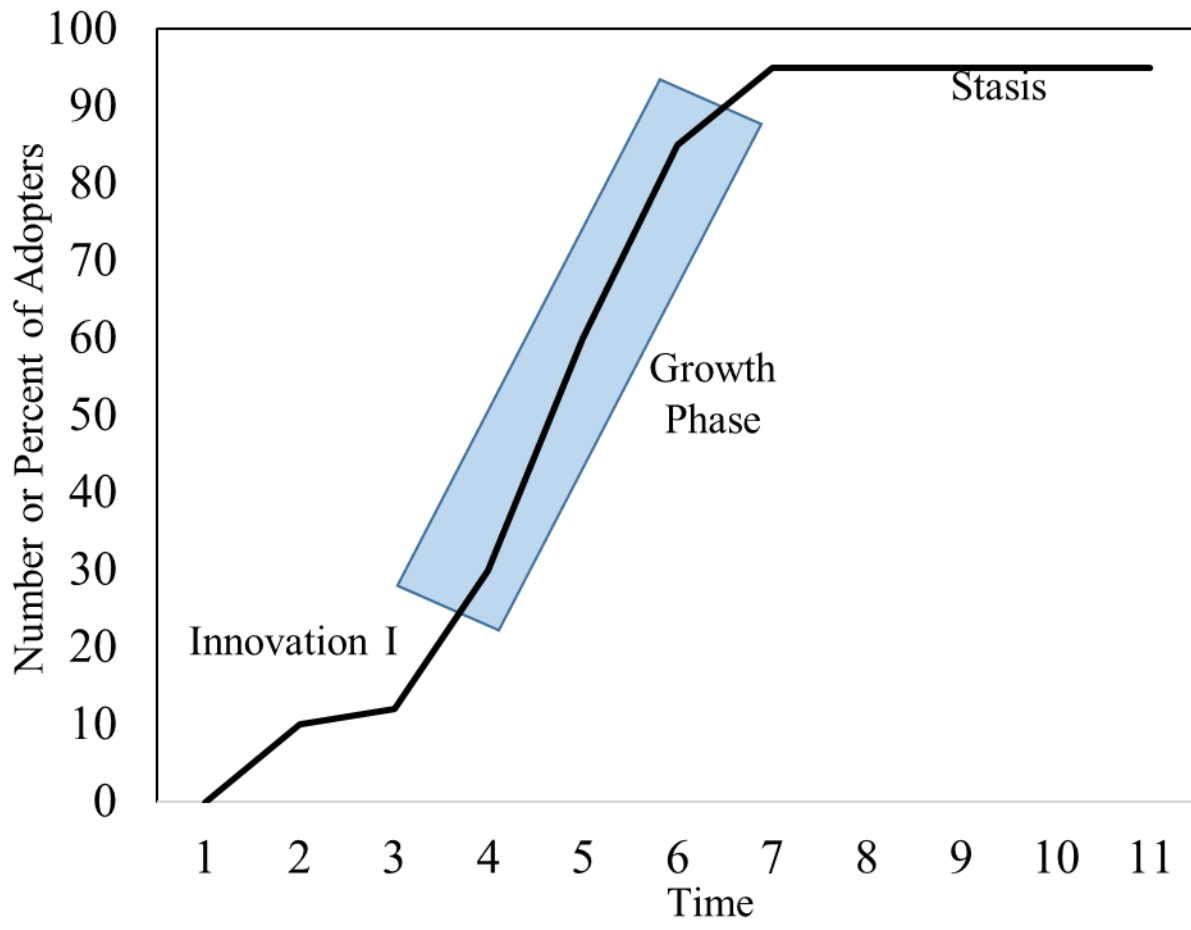


Figure 16. Diffusion process of an innovation over time.

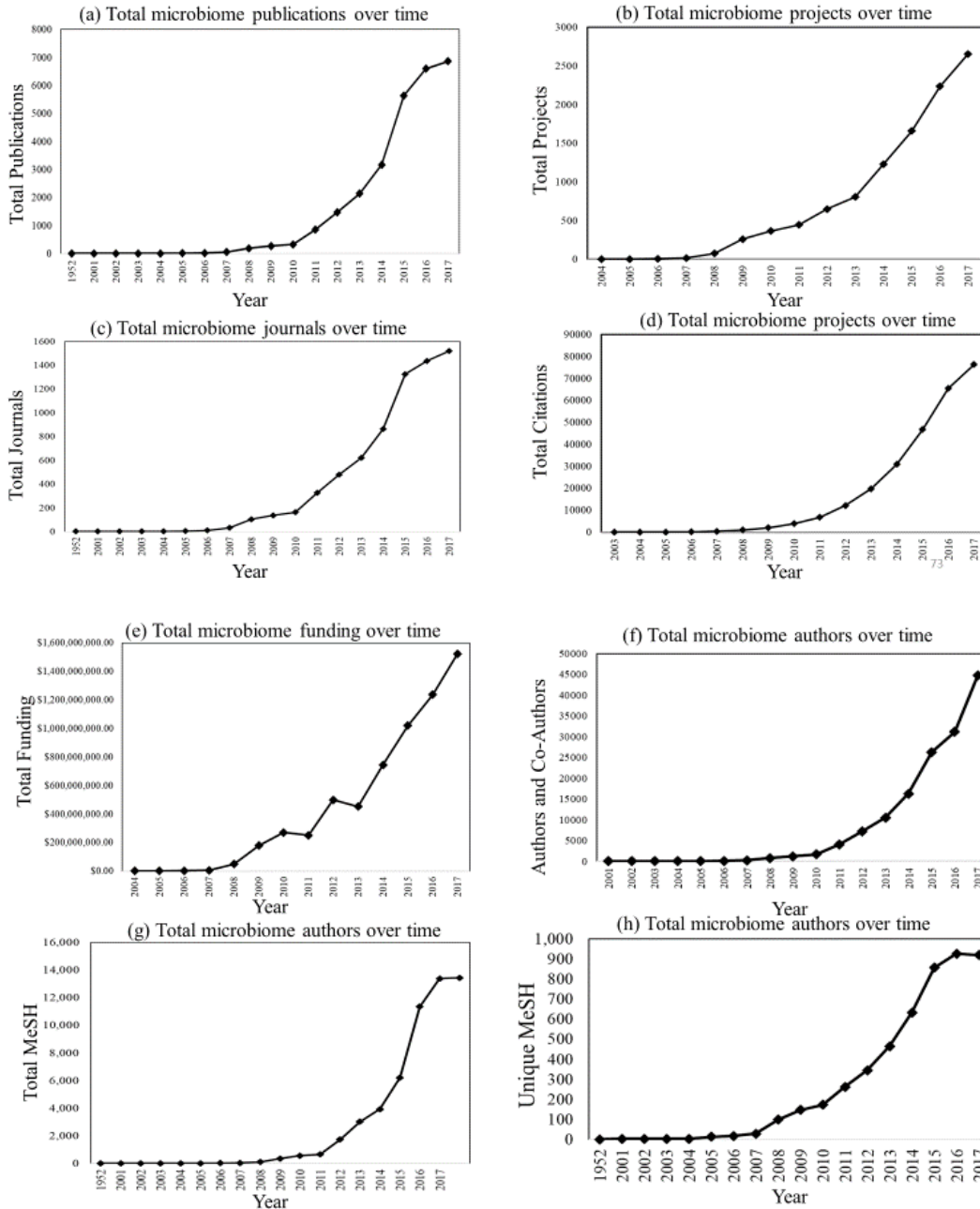


Figure 17 a-h. Adoption rates of microbiome in different systems. 17a. Total publications over time. 17b. Total projects over time. 17c. Total journals over time. 17d. Total citations over time. 17e. Total funding over time. 17f. Authors and co-authors over time. 17g. Total MeSH over time. 17h. Unique MeSH over time.

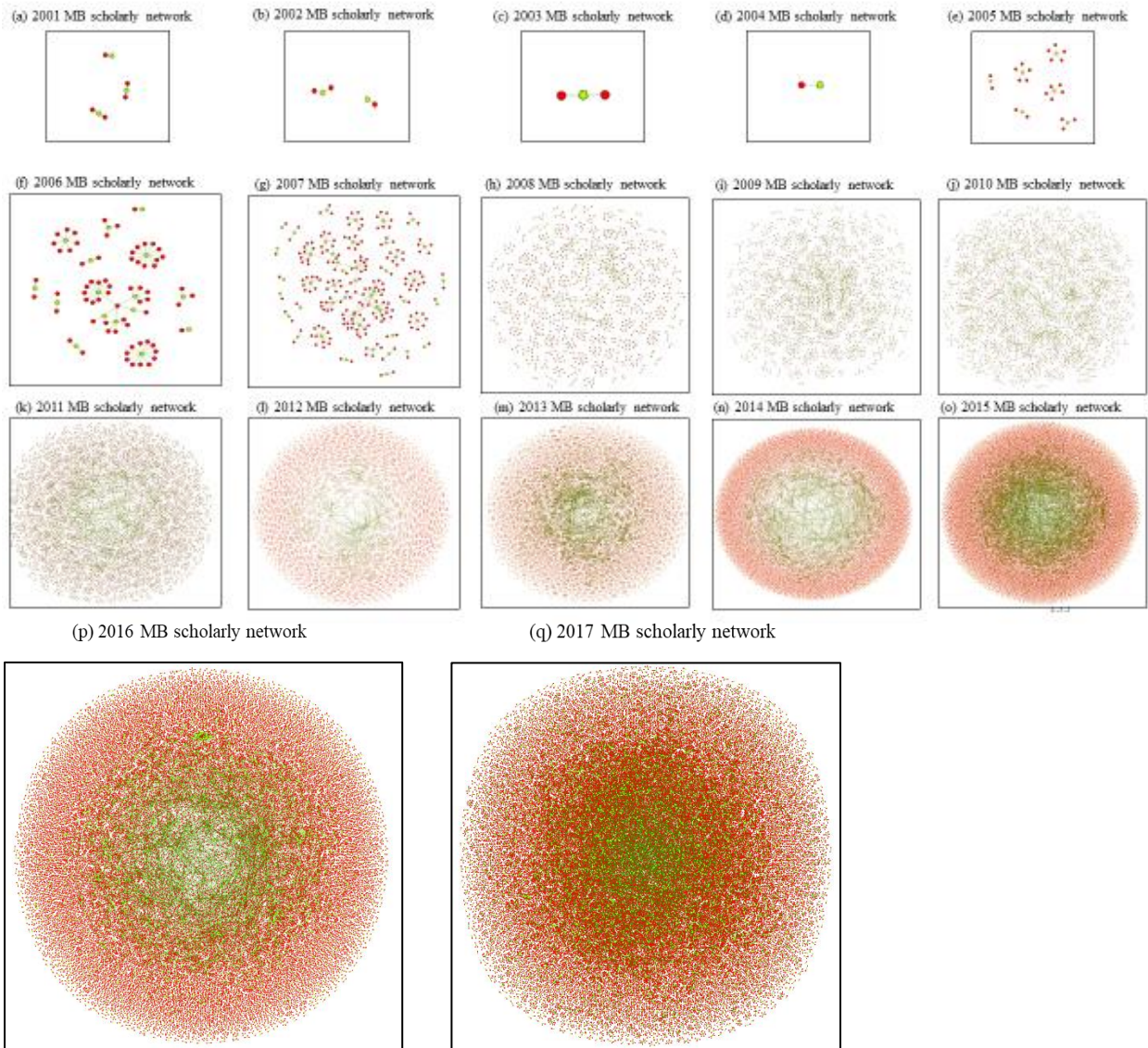


Figure 18a-q. Scholarly network of MB. 18a. 2001 MB scholarly network. 18b. 2002 MB scholarly network. 18c. 2003 MB scholarly network. 18d. 2004 MB scholarly network. 18e. 2005 MB scholarly network. 18f. 2006 MB scholarly network. 18g. 2007 MB scholarly network. 18h. 2008 MB scholarly network. 18i. 2009 MB scholarly network. 18j. 2010 MB scholarly network. 18k. 2011 MB scholarly network. 18l. 2012 MB scholarly network. 18m. 2013 MB scholarly network. 18n. 2014 MB scholarly network. 18o. 2015 MB scholarly network. 18p. 2016 MB scholarly network. 18q. 2017 MB Scholarly network.

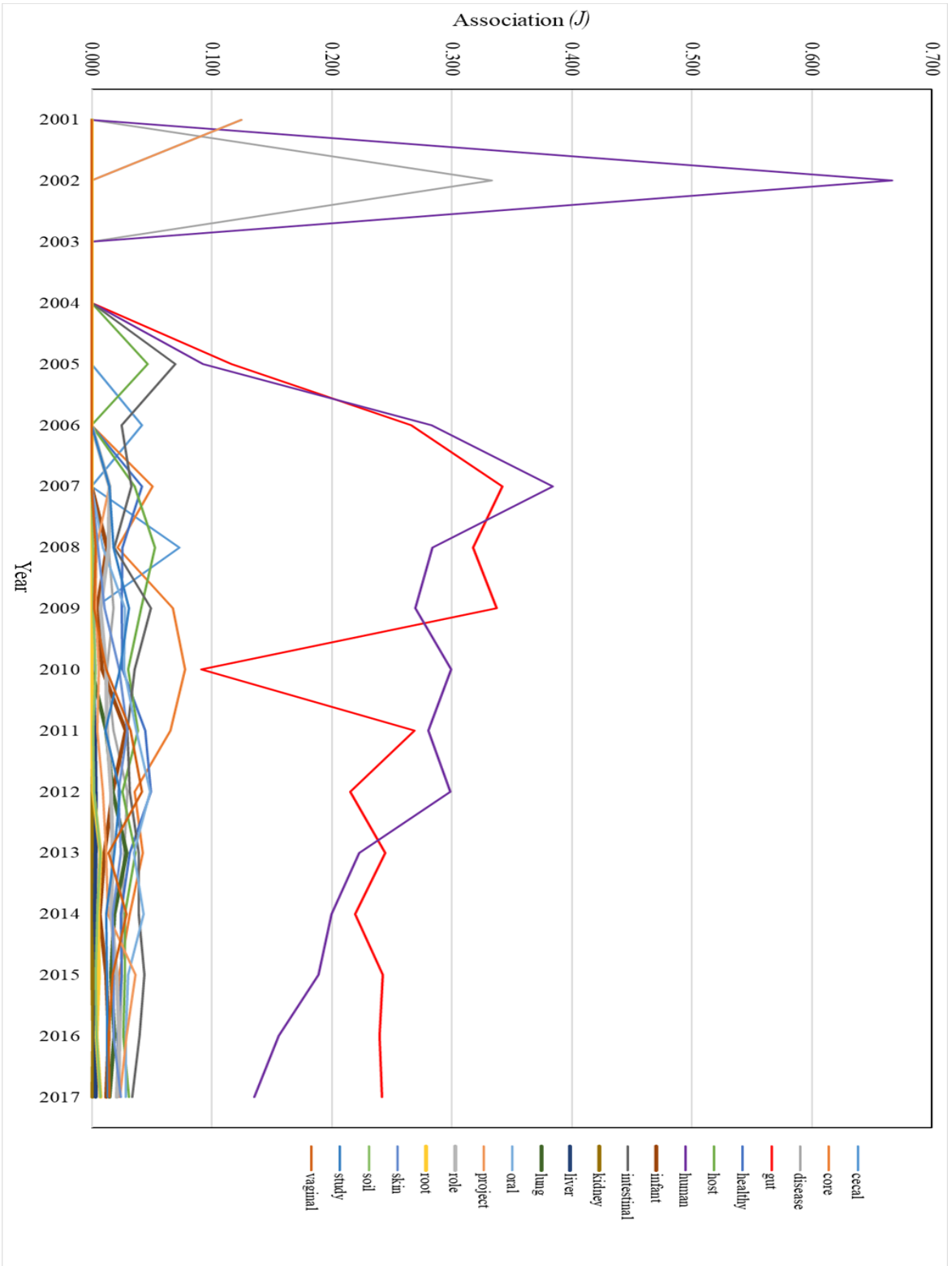


Figure 19. Collocates associated with microbiome from 2001 to 2017.

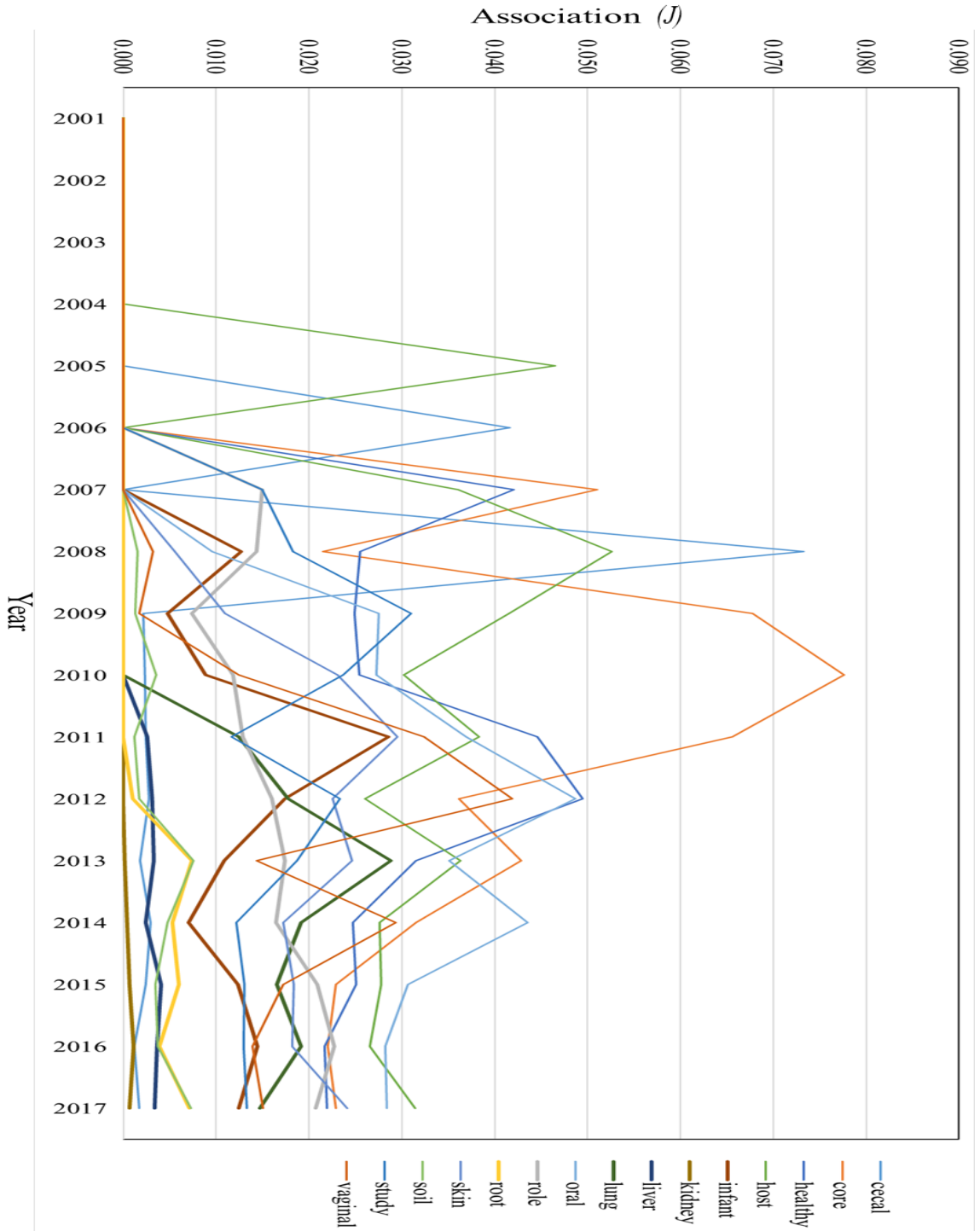


Figure 20. Collocates associated with microbiome by less than 1.0%.

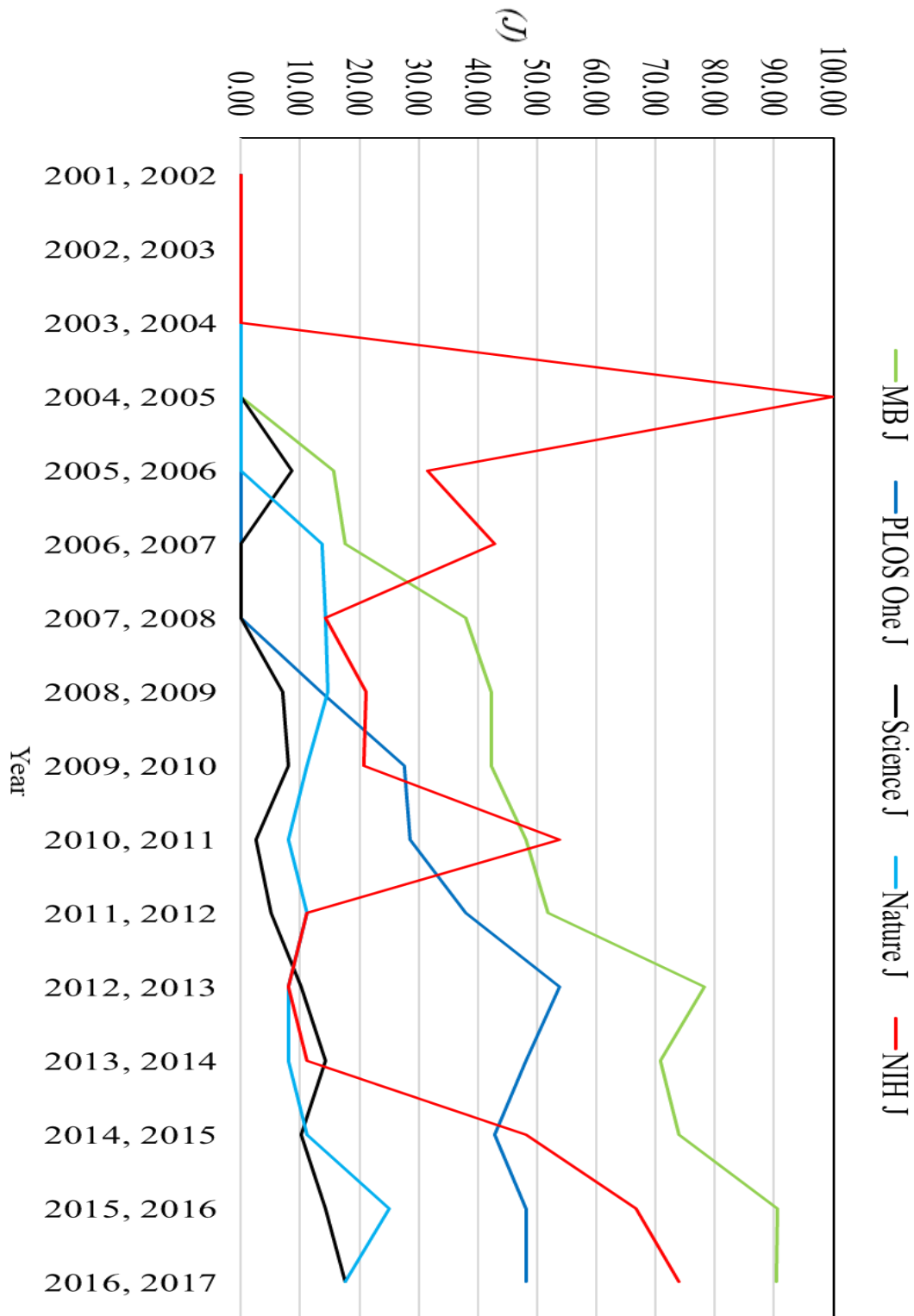


Figure 21. Change in knowledge over time measured by Jaccard similarity (J).

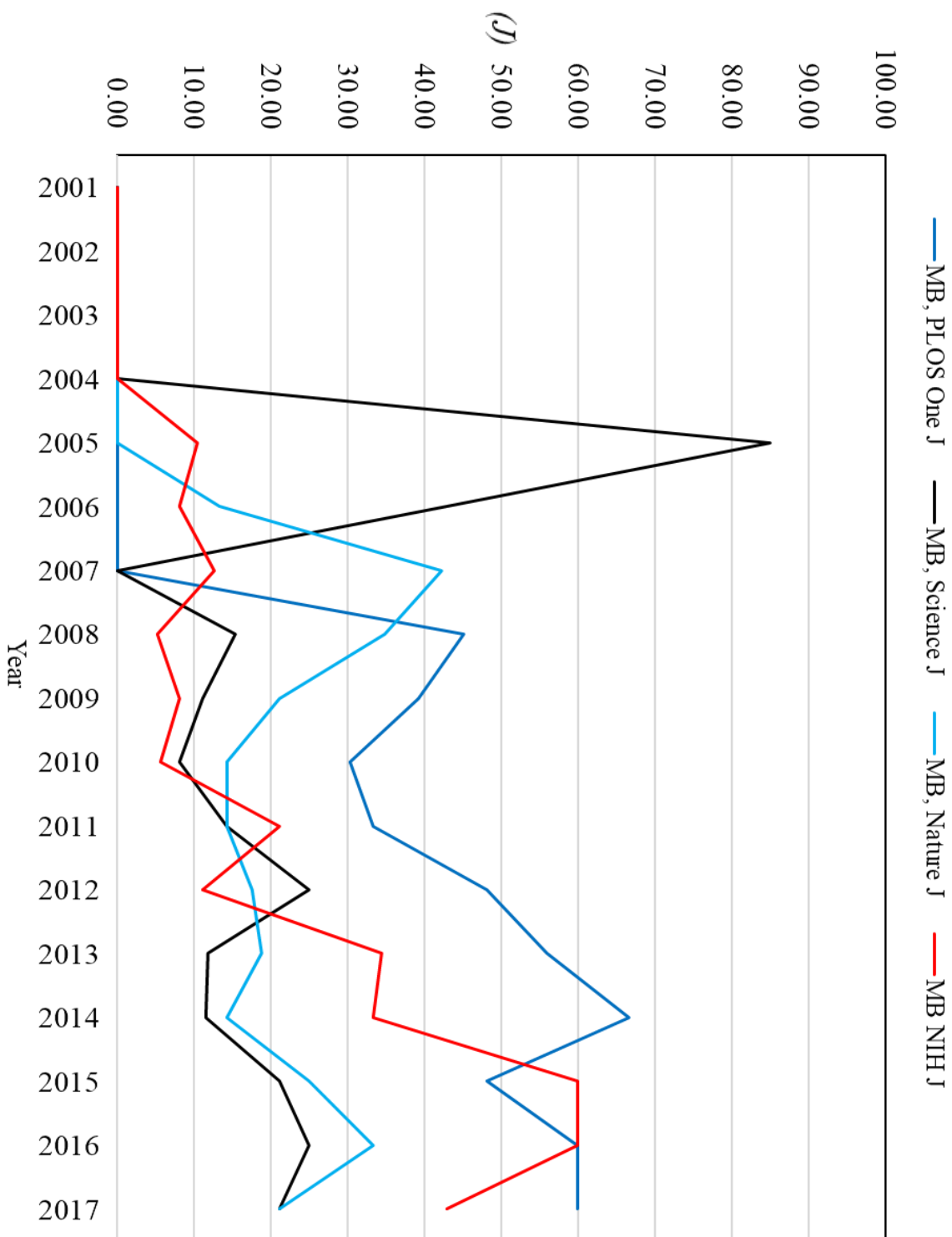


Figure 22. Knowledge convergence over time.

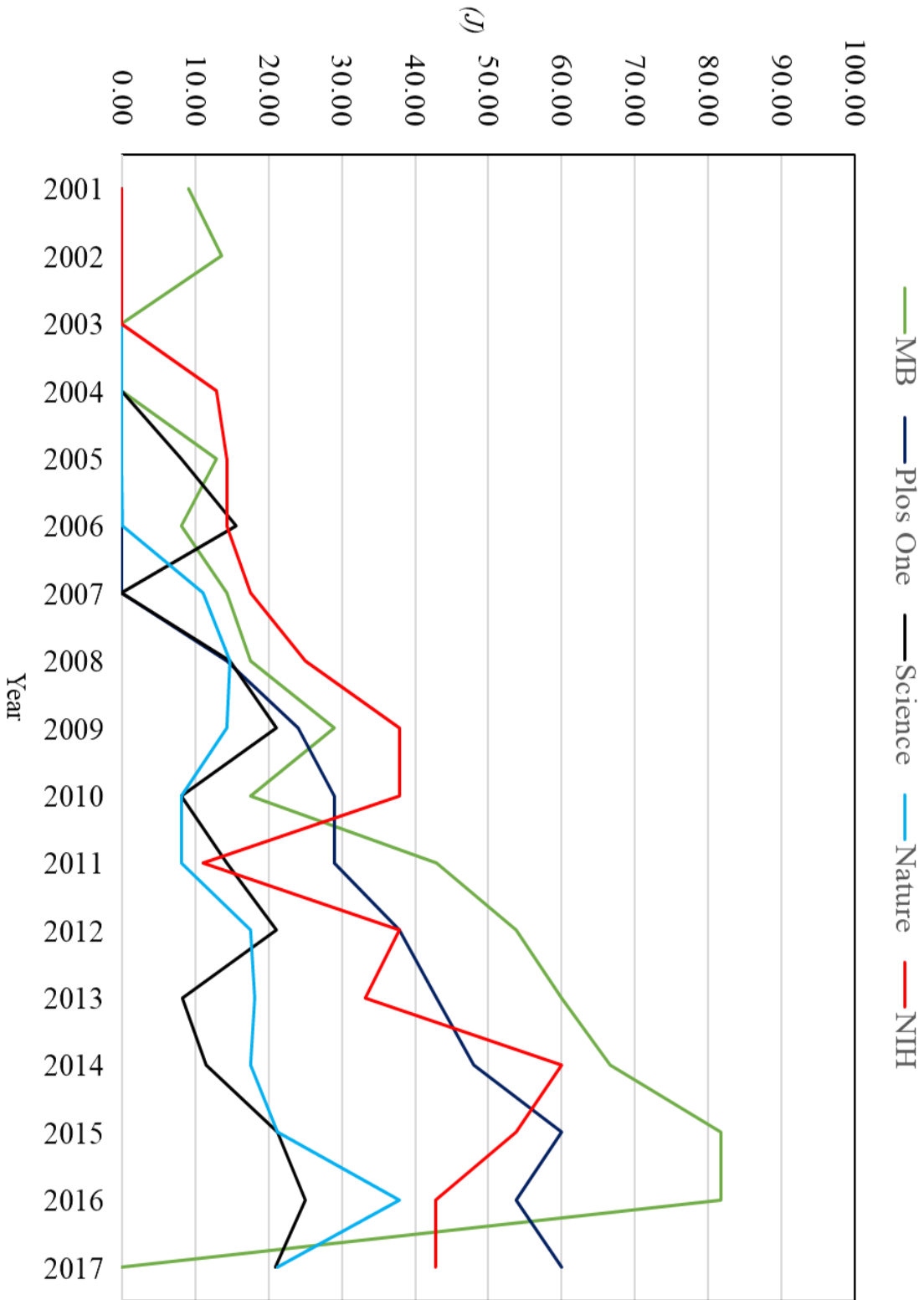


Figure 23. Knowledge convergence toward MB 2017 over time.

CONCLUSION

In this thesis, I detail a systematic strategy to identify social, language, and historical contextual factors influencing the variation and change to the microbiome. My hypothesis is specific contextual factors were related to specific changes to the microbiome concept and microbiome knowledge. The first part of the thesis, chapters one through three, are reviews and descriptions of first principles, concepts, methods, and background information. The first chapter focuses on an extensive review of the microbiome literature and aims at identifying differences in knowledge according to historical, language, and social contexts. The second chapter concentrates on a review of previous approaches to measure and understand knowledge changes within biomedicine and other domains. I adopt the Renn and Laubichler history of knowledge framework to understand changes in scientific knowledge by integrating data from language, social, and historical systems and their interactions across different scales combined with an analysis of contextual factors. The third chapter details the availability of data and describes the benefits of integrating big data and data driven science and research into novel approaches, methods, and models.

The second part of the thesis, chapters four and five, are experiments to identify specific changes to the microbiome emerging within the language and social context of the microbiome. In chapter four, I examine the similarities and differences between the microbiome and the concepts metabolome and metagenome using a hybrid approach combining qualitative and quantitative methods. Previous studies and experts have argued the microbiome, metabolome, and metagenome are often times confused as the same concept (Huss, 2014; Marchesi & Ravel, 2015b). This confusion has the potential to diminish microbiome research, funding, and impact. I find characteristic differences and patterns in the language context of the microbiome by

comparing both the metagenome and metabolome from a Microbiome Corpus of 27,977 articles, a Metabolome Corpus of 16,818 articles, and a Metagenome Corpus of 10,741 articles. Using topic models, frequency analyses, and keywords, I find differences in the discourse, themes, and words between the microbiome and the metabolome and metagenome. I also evaluate the similarity and dissimilarity in the language usage, context, and knowledge between the concepts using collocates, lexical profiles, and collocation networks. My analyses reveal differences in the usage of words and co-occurring words related to the microbiome, metagenome, and metabolome. The combined results suggest the microbiome is a separate concept from metabolome and metagenome and validate a discussion on the status of the microbiome vocabulary.

The findings of my analysis provide an opportunity to analyze concepts and scientific knowledge using tractable evidence from language usage. My analysis indicates biomedical knowledge of the microbiome is dissimilar to the metabolome and metagenome. However, I find the separation between the interpretations of the microbiome, metabolome, and metagenome are dependent on the parameters of the collocate networks and the distance between collocates and the node word(s). This implies possible connections between other collocates and representative knowledge structures. In other words, while this study highlights the differences between the concepts the results also suggest similarities of underlying knowledge components between the microbiome, metabolome, and metagenome. To better understand biomedical knowledge future research needs to explore if these are universal knowledge components to biomedicine, parts of latent knowledge structures, or socio-cultural artifacts.

In chapter five, I develop a methodology to systematically identify and analyze contextual factors influencing the microbiome concept. The methodology analyzes changes in the global language and social context of microbiome research and details the changes to individual words, combinations of words, and knowledge related to the microbiome. In order to assess the variation and change in the language and social context of the microbiome, I analyze the specific material, social, and language dimensions of the microbiome concept from a dataset of roughly 28,000 microbiome articles and 10,000 microbiome projects. My analyses show increases in microbiome articles, projects, journals publishing articles on the microbiome, MeSH terms for journals, and authors from 2001 to 2017. I implement an analysis based on Everett Roger's diffusion of innovations model to analyze the microbiome according to the attributes of innovations and the rate of adoption of innovations. The main factors motivating adoption of an innovation include relative advantage, compatibility, complexity, trialability, and observability. I find the rate of adoption of the microbiome is positively influenced by these attributes. The attributes of innovation also provides insight into sources of social and cultural variation on microbiome usage and knowledge. I also examine other contextual factors cited to influence knowledge including scholarly and scientific networks and journal specific discourses. I discover a drastic increase in the authors and papers connected by a scholarly and scientific network of the microbiome. This results offer another explanation into the variation of microbiome usage. To assess variation and change in microbiome knowledge, I analyze the stability of microbiome knowledge in the Microbiome Corpus, PLOS One Corpus, Science Corpus, Nature Corpus, and NIH Project Corpus. I analyze the knowledge stability of the microbiome corpus and evaluate the change in knowledge using collocates from 2001 comparing them to collocates in 2017. I find microbiome knowledge in the Microbiome Corpus becomes increasingly stable over time.

Repeating this analysis on journal discourses I find different patterns of variation in microbiome knowledge but only PLOS One shows indications of knowledge stability. To analyze the change in microbiome knowledge, I compare microbiome collocates from the first year of the Microbiome Corpus to the last year. I find significant amount of change as the collocates have a low similarity score. Repeating this analysis on journal discourses, my results again show low similarity scores. To identify the influence of journal discourses on microbiome knowledge, I measure the convergence of the Microbiome Corpus discourse to Microbiome Corpus, PLOS One Corpus, Science Corpus, Nature Corpus, and NIH Project Corpus. My results highlight knowledge convergence between journal discourses and the microbiome corpus.

The results from the experiments of this dissertation highlight a hybrid approach to understanding the variation and changes to scientific knowledge. I provide quantitative and qualitative evidence showing the microbiome is a separate concept from metabolome and metagenome, and variation and change to microbiome knowledge emerges from contextual factors related to language and social contexts. These results support results and evidence from previous studies and provide new directions for future research.

The future interpretations and use of the microbiome is based on knowledge. Science defines and shapes knowledge, as knowledge within science is a reflection of the social, language, and historical context. The microbiome is seen as a biomedical concept and is subject to possible constraints due to research strategies in biomedicine. According Peter Conrad, who studies medical sociology, the influx of new concepts such as: ADHD, Post traumatic stress disorder, along with the increased impact of medicine and medical concepts in the past fifty years has resulted in the jurisdiction of medicine growing to new problems that were not previously deemed to fall into the medical sphere medicine, or the medicalization of concepts (P.

Conrad, 1992) . Some analysts have suggested the growth of medical jurisdiction, medicalization increases the amount of medical social control over human behavior (Clarke et al., 2010). The key issue, remains in the context of those concepts or how they concepts are defined because the power to have a particular set of (medical) definitions dictates how a particular concept is studied and within which domain (P. Conrad, 2007). Critics have recently emphasized how medicalization has increased the profitability and markets of pharmaceutical and biotechnological firms and led to disease mongering or disease brandin .

Medicalization also suggests context is the most adequate method of understanding a concept and its causes. Previous studies have shown how alcoholism and anorexia can be better understood with a broader perspective than biomedicine, as they both can be understood as psychiatric disorders, manifestations of genetics, or as complex concepts that also intersect personal, psychological, and social factors (P. Conrad & Schneider, 2010). Understanding medicine as a culture, opens up novels perspectives on knowledge practices and epistemic features in biomedicine, insight into the construction and fashioning of knowledge objects within science, and on the arrangements and mechanisms in biomedicine that shape what is known and how it is known (J. Swan et al., 2007; Green et al., 2009; Cetina, 2009).

Considering the microbiome is shaped by contexts beyond biomedicine, provides insight that future work on the microbiome needs to go beyond the biomedical perspective into other areas. From a legal perspective, what the microbiome is determines who owns the microbiome, individuals or the environment, which then further relates to if knowledge from the microbiome is an open resource which is directly related to property rights (Rhodes et al., 2013). How the microbiome is defined and used in a language context determines who the microbiome belongs to, which is directly tied to property issues, patents, and licensing of genetic samples (Colaianni

& Cook-Deegan, 2009). As the microbiome goes beyond the practice of medicine into legal issues, the microbiome will impact social systems, legal systems, ethics, and research. Police have already started experimenting with using evidence from microbiome research as a tool for law enforcement agencies to improve trace evidence options (Hampton-Marcell, Lopez, & Gilbert, 2017). Other see a future where the microbiome as a tool for identification crosses social systems and can be used by doctors, insurers, and employers (Rees et al., 2018). Therefore, the use and meaning of the microbiome relates to larger issues within science regarding discovery, recognition, and institutional norms within science.

As science continues to grow at an ever increasing pace it is critical to understand the historical past and future trajectory of science, to understand where we have been, and where we are at in specific moments in time. This dramatic growth of information or big data has created new objects and dimensions which require new tools and perspectives. Yet, many scientists and researchers and scientists are still struggling with how to use big data. The experiments and results within this prospectus provide a methodology for data driven science and research with specific use cases and contribute to the conversations about the problems and resolutions within big data projects. Understanding how knowledge changes in the microbiome offers new tools and perspectives on how to identify and measure changes to scientific knowledge, and delivers insight into mechanisms of change within science. It is possible changes to the microbiome may be indicative of a specific pattern for the microbiome or part of a new dynamic within science. Future work needs to consider if the microbiome is a singular phenomena or how the evolution of the microbiome illustrates how concepts with resources and attention behave within biomedicine and science, or if the microbiome is a unique case and represents a change in how

concepts evolve. Understanding these dynamics are critical as the behavior of the microbiome is either representative of conceptual change or what we can expect in the near future.

REFERENCES

10 Key Marketing Trends for 2017. (n.d.). Retrieved October 27, 2018, from <https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN>

A Very Short History Of Big Data. (n.d.). Retrieved October 27, 2018, from <https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#5c0f504565a1>

Ackerknecht, E. H., & Haushofer, L. (2016). *A Short History of Medicine*. JHU Press.

Adami, A. J., & Bracken, S. J. (2016). Breathing Better Through Bugs: Asthma and the Microbiome. *The Yale Journal of Biology and Medicine*, *89*(3), 309–324.

Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, *1*(1), 9–28. <https://doi.org/10.1558/japl.1.1.9.55871>

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer Science & Business Media.

Agrawal, A., & Henderson, R. (2002). Putting Patents in Context: Exploring Knowledge Transfer from MIT. *Management Science*, *48*(1), 44–60. <https://doi.org/10.1287/mnsc.48.1.44.14279>

Ahmad, T., Testani, J. M., & Desai, N. R. (2016). Can Big Data Simplify the Complexity of Modern Medicine? Prediction of Right Ventricular Failure After Left Ventricular Assist Device Support as a Test Case. *Jacc-Heart Failure*, *4*(9), 722–725. <https://doi.org/10.1016/j.jchf.2016.06.004>

Aiello, K. D., Caughey, W. G., Nelluri, B., Sharma, A., Mookadam, F., & Mookadam, M. (2016). Effect of exercise training on sleep apnea: A systematic review and meta-analysis. *Respiratory Medicine*, *116*, 85–92. <https://doi.org/10.1016/j.rmed.2016.05.015>

Aijmer, K., & Stenström, A.-B. (2004). *Discourse Patterns in Spoken and Written Corpora*. John Benjamins Publishing.

Alekseev, A. A., Osipova, V. V., Ivanov, M. A., Klimentov, A., Grigorieva, N. V., & Nalamwar, H. S. (2016). Efficient Data Management Tools for the Heterogeneous Big Data Warehouse. *Physics of Particles and Nuclei Letters*, *13*(5), 689–692. <https://doi.org/10.1134/S1547477116050022>

Almeida, P., & Kogut, B. (1999). Localization of Knowledge and the Mobility of Engineers in Regional Networks. *Management Science*, *45*(7), 905–917. <https://doi.org/10.1287/mnsc.45.7.905>

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLOS ONE*, 6(9), e24357. <https://doi.org/10.1371/journal.pone.0024357>

American Society for Microbiology. (2013). *Human Microbiome FAQ*. American Academy of Microbiology. Retrieved from https://www.asm.org/images/stories/documents/FAQ_Human_Microbiome.pdf

Amsterdamska, O., & Leydesdorff, L. (1989). Citations: Indicators of significance? *Scientometrics*, 15(5), 449–471. <https://doi.org/10.1007/BF02017065>

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. Retrieved from <https://www.wired.com/2008/06/pb-theory/>

Anderson, M., & Anderson, S. L. (2011). *Machine Ethics*. Cambridge University Press.

Ankeny, R. A., & Leonelli, S. (2016). Repertoires: A post-Kuhnian perspective on scientific change and collaborative research. *Studies in History and Philosophy of Science Part A*, 60, 18–28. <https://doi.org/10.1016/j.shpsa.2016.08.003>

Announcing the National Microbiome Initiative. (2016, May 13). Retrieved October 30, 2018, from <https://obamawhitehouse.archives.gov/blog/2016/05/13/announcing-national-microbiome-initiative>

Arcaya, M. C., Arcaya, A. L., & Subramanian, S. V. (2015). Inequalities in health: definitions, concepts, and theories. *Global Health Action*, 8. <https://doi.org/10.3402/gha.v8.27106>

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, 17–21.

Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G., Nelson, S. J., ... Wilbur, W. J. (2000). The NLM Indexing Initiative. *Proceedings of the AMIA Symposium*, 17–21.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 79–80, 3–15. <https://doi.org/10.1016/j.jpdc.2014.08.003>

Baccianella, S., Esuli, A., & Sebastiani, F. (n.d.). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, 5.

Baker, M. (2015). Over half of psychology studies fail reproducibility test. *Nature News*. <https://doi.org/10.1038/nature.2015.18248>

- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604), 452. <https://doi.org/10.1038/533452a>
- Baker, P. (2006). *Using Corpora in Discourse Analysis*. A&C Black.
- Baker, P. (2010). *Sociolinguistics and Corpus Linguistics* (1 edition). Edinburgh: Edinburgh University Press.
- Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3), 247–256.
- Balaid, A., Abd Rozan, M. Z., Hikmi, S. N., & Memon, J. (2016). Knowledge maps: A systematic literature review and directions for future research. *International Journal of Information Management*, 36(3), 451–475. <https://doi.org/10.1016/j.ijinfomgt.2016.02.005>
- Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439), 509–512. <https://doi.org/10.1126/science.286.5439.509>
- Bar-Shalom, A., & Cook-Deegan, R. (2002). Patents and Innovation in Cancer Therapeutics: Lessons from CellPro. *The Milbank Quarterly*, 80(4), 637–676. <https://doi.org/10.1111/1468-0009.00027>
- Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, 44(1), 118–125.
- Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., ... Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, 59(s1), 1–26. <https://doi.org/10.1111/j.1467-9922.2009.00533.x>
- Bellis, N. D. (2009). *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Lanham, Md: Scarecrow Press.
- Bergs, A., & Diewald, G. (2008). *Constructions and Language Change*. Walter de Gruyter.
- Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27(3), 264–271. <https://doi.org/10.1016/j.giq.2010.03.001>
- Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2), 97–116. <https://doi.org/10.1016/j.jeap.2006.05.001>
- Biber, D., Douglas, B., Biber, P. D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Big Data - NITRDGROUPS. (n.d.). Retrieved October 27, 2018, from https://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data

Bijker, W. E. (1993). Do Not Despair: There Is Life after Constructivism. *Science, Technology, & Human Values*, 18(1), 113–138. <https://doi.org/10.1177/016224399301800107>

Blair, A. (2003). Reading Strategies for Coping with Information Overload ca. 1550-1700. *Journal of the History of Ideas*, 64(1), 11. <https://doi.org/10.2307/3654293>

Blair, A. M. (2011). *Too Much to Know: Managing Scholarly Information before the Modern Age* (First Edition edition). New Haven London: Yale University Press.

Blaser, M., Bork, P., Fraser, C., Knight, R., & Wang, J. (2013). The microbiome explored: recent insights and future challenges. *Nature Reviews. Microbiology*, 11(3), 213–217. <https://doi.org/10.1038/nrmicro2973>

Blei, D. M. (2012). Probabilistic Topic Models. *Commun. ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1), 17–35. <https://doi.org/10.1214/07-AOAS114>

Blei, D. M., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning and Research*, 3, 30.

Bloor, D. (1991). *Knowledge and Social Imagery* (1 edition). Chicago: University of Chicago Press.

Boiten, E. A. (2016). Big Data Refinement. *Electronic Proceedings in Theoretical Computer Science*, (209), 17–23. <https://doi.org/10.4204/EPTCS.209.2>

Bordenstein, S. (n.d.). *The Microbiome and Darwin's Mystery of Mysteries*.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1), 179–255. <https://doi.org/10.1002/aris.1440370106>

Bourdieu, P. (1977). *Outline of a Theory of Practice*. (R. Nice, Trans.) (1st English Ed edition). Cambridge: Cambridge University Press.

Bourdieu, P., & Wacquant, L. J. D. (1992). *An Invitation to Reflexive Sociology*. University of Chicago Press.

- Boyd, D., & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679. <https://doi.org/10.1080/1369118X.2012.678878>
- Breddels, M. A., & Veljanoski, J. (2018). Vaex: big data exploration in the era of Gaia. *Astronomy & Astrophysics*, 618, A13. <https://doi.org/10.1051/0004-6361/201732493>
- Breschi, S., & Catalini, C. (2010). Tracing the links between science and technology: An exploratory analysis of scientists' and inventors' networks. *Research Policy*, 39(1), 14–26. <https://doi.org/10.1016/j.respol.2009.11.004>
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173.
- Brigandt, I. (2010). The epistemic goal of a concept: accounting for the rationality of semantic change and variation. *Synthese*, 177(1), 19–40. <https://doi.org/10.1007/s11229-009-9623-8>
- Britton, D., & Lloyd, S. L. (2014). How to deal with petabytes of data: the LHC Grid project. *Reports on Progress in Physics*, 77(6), 065902. <https://doi.org/10.1088/0034-4885/77/6/065902>
- Brown, G., Gillian, B., & Yule, G. (1983). *Discourse Analysis*. Cambridge University Press.
- Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based N-gram Models of Natural Language. *Comput. Linguist.*, 18(4), 467–479.
- Buchwald, H., Avidor, Y., Braunwald, E., Jensen, M. D., Pories, W., Fahrback, K., & Schoelles, K. (2004). Bariatric Surgery: A Systematic Review and Meta-analysis. *JAMA*, 292(14), 1724–1737. <https://doi.org/10.1001/jama.292.14.1724>
- Budgell, B. (2016). Subluxation and semantics: a corpus linguistics study. *The Journal of the Canadian Chiropractic Association*, 60(2), 190–194.
- Burian, R. M. (2004). Molecular epigenesis, molecular pleiotropy, and molecular gene definitions. *History and Philosophy of the Life Sciences*, 26(1), 59–80.
- Burkette, A., & Jr, W. A. K. (2018). *Exploring Linguistic Science: Language Use, Complexity, and Interaction* (1 edition). New York: Cambridge University Press.
- Burkette, A., & Kretzschmar, W. A. (2018). *Exploring Linguistic Science: Language Use, Complexity, and Interaction* (1 edition). New York: Cambridge University Press.
- Burri, R. V., & Dumit, J. (Eds.). (2007). *Biomedicine as Culture: Instrumental Practices, Technoscientific Knowledge, and New Modes of Life* (1 edition). New York: Routledge.

Bush, V. (1996). As We May Think. *Interactions*, 3(2), 35–46. <https://doi.org/10.1145/227181.227186>

Bybee, J. (2007). *Frequency of Use and the Organization of Language*. Oxford University Press, USA.

Callahan, M., Wears, R. L., & Weber, E. (2002). Journal Prestige, Publication Bias, and Other Characteristics Associated With Citation of Published Studies in Peer-Reviewed Journals. *JAMA*, 287(21), 2847–2850. <https://doi.org/10.1001/jama.287.21.2847>

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 69–80. <https://doi.org/10.1016/j.shpsc.2011.10.007>

Calvard, T. S. (2016). Big data, organizational learning, and sensemaking: Theorizing interpretive challenges under conditions of dynamic complexity. *Management Learning*, 47(1), 65–82. <https://doi.org/10.1177/1350507615592113>

Carley, K. (1993). Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis. *Sociological Methodology*, 23, 75–126. <https://doi.org/10.2307/271007>

Carley, K. (n.d.). Dynamic Network Analysis. *Summary of the NRC Workshop on Social Network Modeling and Analysis*. Retrieved from http://www.casos.cs.cmu.edu/publications/protected/2000-2004/2003-2004/carley_2003_dynamicnetwork.pdf

Carley, K. M., Lee, J.-S., & Krackhardt, D. (2002). Destabilizing Networks. *Connections*, 24(3), 79–92.

Carley, S., Porter, A. L., & Youtie, J. (2013). Toward a more precise definition of self-citation. *Scientometrics*, 94(2), 777–780. <https://doi.org/10.1007/s11192-012-0745-2>

CASOS Tools: Network Analysis Data | CASOS. (n.d.). Retrieved October 27, 2018, from <http://www.casos.cs.cmu.edu/tools/data.php>

Ceri, S. (2018). On the role of statistics in the era of big data: A computer science perspective. *Statistics & Probability Letters*, 136, 68–72. <https://doi.org/10.1016/j.spl.2018.02.019>

Cetina, K. K. (2009). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press.

- Cetina, K. K., & Reichmann, W. (2015). Epistemic Cultures. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 873–880). Oxford: Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.10454-4>
- Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 288–296). Curran Associates, Inc. Retrieved from <http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf>
- Chen, C. (2006a). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359–377. <https://doi.org/10.1002/asi.20317>
- Chen, C. (2006b). *Information Visualization: Beyond the Horizon* (2nd ed.). London: Springer-Verlag. Retrieved from [//www.springer.com/us/book/9781852337896](http://www.springer.com/us/book/9781852337896)
- Chen, C. (2015). *How to Use CiteSpace*. Leanpub. Retrieved from <https://leanpub.com/howtousecitespace>
- Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2), 97–138.
- Choudhury, S., Fishman, J. R., McGowan, M. L., & Juengst, E. T. (2014). Big data, open science and the brain: lessons learned from genomics. *Frontiers in Human Neuroscience*, 8, 239. <https://doi.org/10.3389/fnhum.2014.00239>
- Cicourel, A. V. (1964). *Method and measurement in sociology*. Oxford, England: Free Press of Glencoe.
- Cixous, H., Cohen, K., & Cohen, P. (1976). The Laugh of the Medusa. *Signs*, 1(4), 875–893.
- Clark, Alexander, Fox, C., & Lappin, S. (2013). *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.
- Clark, Andy. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Clarke, A. E., Mamo, L., Fosket, J. R., Fishman, J. R., & Shim, J. K. (Eds.). (2010). *Biomedicalization: Technoscience, Health, and Illness in the U.S.* (1 edition). Durham, NC: Duke University Press Books.

Clear, J. (1993). From Firth principles: Computational tools for the study of collocation. *Text and Technology: In Honour of John Sinclair*, 271–292.

Cleveland, W. S. (1985). *The Elements of Graphing Data* (1st edition). Monterey, Cal: Wadsworth, Inc.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1), 57–71. <https://doi.org/10.1093/bib/6.1.57>

Cohen, K. B., & Demner-Fushman, D. (2014). *Biomedical Natural Language Processing*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.

Cohen, R., Elhadad, M., & Elhadad, N. (2013). Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, 14, 10. <https://doi.org/10.1186/1471-2105-14-10>

Colaianni, A., & Cook-Deegan, R. (2009). Columbia University's Axel Patents: Technology Transfer and Implications for the Bayh-Dole Act. *The Milbank Quarterly*, 87(3), 683–715. <https://doi.org/10.1111/j.1468-0009.2009.00575.x>

Cole, J. R., & Cole, S. (1974). Social Stratification in Science. *Science and Society*, 38(3), 374–378.

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature News*, 505(7485), 612. <https://doi.org/10.1038/505612a>

Conrad, P. (1992). Medicalization and Social Control. *Annual Review of Sociology*, 18(1), 209–232. <https://doi.org/10.1146/annurev.so.18.080192.001233>

Conrad, P. (2007). *The Medicalization of Society: On the Transformation of Human Conditions into Treatable Disorders*. JHU Press.

Conrad, P., & Schneider, J. W. (2010). *Deviance and Medicalization: From Badness to Sickness*. Temple University Press.

Conrad, S., & Biber, D. (2009). *Real Grammar: A Corpus-Based Approach to English* (1st edition). White Plains, NY: Pearson Education ESL.

Council, N. R., Studies, D. on E. and L., Sciences, B. on L., & Applications, C. on M. C. and F. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. National Academies Press.

Crane, D. (1967). The Gatekeepers of Science: Some Factors Affecting the Selection of Articles for Scientific Journals. *The American Sociologist*, 2(4), 195–201.

Crowe, N., Dietrich, M. R., Alomepe, B. S., Antrim, A. F., ByrneSim, B. L., & He, Y. (2015). The diversification of developmental biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 53, 1–15. <https://doi.org/10.1016/j.shpsc.2015.04.004>

D Ainsworth, T., Krause, L., Bridge, T., Torda, G., Raina, J.-B., Zakrzewski, M., ... Leggat, W. (2015). The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. *The ISME Journal*, 9(10), 2261–2274. <https://doi.org/10.1038/ismej.2015.39>

Davies, J. (2001). In a Map for Human Life, Count the Microbes, Too. *Science*, 291(5512), 2316–2316. <https://doi.org/10.1126/science.291.5512.2316b>

Dawson, J. A., & Kendzioriski, C. (2012). Survival-supervised latent Dirichlet allocation models for genomic analysis of time-to-event outcomes. *ArXiv:1202.5999 [Stat]*. Retrieved from <http://arxiv.org/abs/1202.5999>

Denis, J.-L., Hébert, Y., Langley, A., Lozeau, D., & Trottier, L.-H. (2002). Explaining Diffusion Patterns for Complex Health Care Innovations. *Health Care Management Review*, 27(3), 60.

Detailed Indexing Statistics: 1965-2017. (2018, August 22). Retrieved October 2, 2018, from http://www.nlm.nih.gov/bsd/index_stats_comp.html

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: a review of possible methods. *Journal of Health Services Research & Policy*, 10(1), 45–53.

Dixon-Woods, M., Agarwal, Shona, Young, B., Jones, D., & Sutton, A. (2004). *Integrative approaches to qualitative and quantitative evidence*. London: Health Development Agency.

Dixon-Woods, M., & NHS Health Development Agency. (2004). *Integrative approaches to qualitative and quantitative evidence*. London: Health Development Agency.

Doreian, P., & Stokman, F. N. (Eds.). (1997). *Evolution of Social Networks* (1 edition). Amsterdam: Routledge.

Douglas, M. (1966). Purity and Danger: An analysis of the concepts of pollution and taboo, 12.

Drieschner, K. H., Lammers, S. M. M., & van der Staak, C. P. F. (2004). Treatment motivation: An attempt for clarification of an ambiguous concept. *Clinical Psychology Review*, 23(8), 1115–1137. <https://doi.org/10.1016/j.cpr.2003.09.003>

Durkheim, E., & Mauss, M. (1963). *Primitive Classification*. Retrieved from <https://www.press.uchicago.edu/ucp/books/book/chicago/P/bo25149490.html>

- Edge, D. (1979). Quantitative Measures of Communication in Science: A Critical Review. *History of Science*, 17(2), 102–134. <https://doi.org/10.1177/007327537901700202>
- Eisen, J. (2015). What does the term microbiome mean? And where did it come from? A bit of a surprise .. <https://doi.org/10.15200/winn.142971.16196>
- El-Hani, C. N. (2007). Between the cross and the sword: the crisis of the gene concept. *Genetics and Molecular Biology*, 30(2), 297–307. <https://doi.org/10.1590/S1415-47572007000300001>
- Ellis, N. C., & Larsen-Freeman, D. (2009). *Language as a Complex Adaptive System* (1 edition). Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell.
- English, J. F., & Underwood, T. (2016). Shifting Scales: Between Literature and Social Science. <https://doi.org/10.1215/00267929-3570612>
- Enhancing Reproducibility through Rigor and Transparency | grants.nih.gov. (n.d.). Retrieved November 19, 2018, from <https://grants.nih.gov/reproducibility/index.htm>
- Eppler, M. J. (2008). A process-based classification of knowledge maps and application examples. *Knowledge and Process Management*, 15(1), 59–71. <https://doi.org/10.1002/kpm.299>
- Erb-Downward, J. R., Thompson, D. L., Han, M. K., Freeman, C. M., McCloskey, L., Schmidt, L. A., ... Huffnagle, G. B. (2011). Analysis of the Lung Microbiome in the “Healthy” Smoker and in COPD. *PLOS ONE*, 6(2), e16384. <https://doi.org/10.1371/journal.pone.0016384>
- Evert, S. (2008). Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2, 1212–1248.
- Evert, S., & Krenn, B. (2004). Computational approaches to collocations. *Introductory Course at the European Summer School on Logic, Language, and Information (ESSLLI 2003)*, Vienna.
- Faggion, C. M., Bakas, N. P., & Wasiak, J. (2017). A survey of prevalence of narrative and systematic reviews in five major medical journals. *BMC Medical Research Methodology*, 17. <https://doi.org/10.1186/s12874-017-0453-y>
- Falk, R. (2004). Long Live the Genome! So Should the Gene. *History and Philosophy of the Life Sciences*, 26(1), 105–121.
- Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *COMMUNICATIONS OF THE ACM*, 49(9), 7.
- Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F. M., & Cruz, I. F. (2018). Tackling the challenges of matching biomedical ontologies. *Journal of Biomedical Semantics*, 9(1), 4. <https://doi.org/10.1186/s13326-017-0170-9>

Federal RePORTER - Project Search Results. (n.d.). Retrieved October 30, 2018, from <https://federalreporter.nih.gov/projects/search/?searchId=8612652de2964f93bc587cdaa70372b3&searchMode=Smart&filters=>

Fennell, M. L., & Warnecke, R. B. (2013). *The Diffusion of Medical Innovations: An Applied Network Analysis*. Springer Science & Business Media.

Ferritin test - Mayo Clinic. (n.d.). Retrieved October 3, 2018, from <https://www.mayoclinic.org/tests-procedures/ferritin-test/about/pac-20384928>

Feyerabend, P. (1962). *Scientific Explanation, Space, and Time*. University of Minnesota Press.

Feyerabend, P. K. (1970). Against method: outline of an anarchistic theory of knowledge. Retrieved from <http://conservancy.umn.edu/handle/11299/184649>

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171. <https://doi.org/10.1023/A:1013713905833>

Fieschi, M., Coiera, E., & Li, Y.-C. J. (2004). *Medinfo*. IOS Press.

Firth, J. R. (1951). *Modes of meaning*. Bobbs-Merrill.

Fleck, L., & Kuhn, T. S. (1981). *Genesis and Development of a Scientific Fact*. (T. J. Trenn & R. K. Merton, Eds., F. Bradley, Trans.) (New edition edition). Chicago u.a: University of Chicago Press.

Forum, I. of M. (US) F. (2013). *Study of the Human Microbiome*. National Academies Press (US). Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK154091/>

Francis, N., & Kucera, H. (1964). *Brown Corpus*. Retrieved from <http://archive.org/details/BrownCorpus>

Franzosi, R. (2004). *From Words to Numbers: Narrative, Data, and Social Science* (1 edition). Cambridge, UK ; New York: Cambridge University Press.

Friemel, T. N. (2011). Dynamics of Social Networks. *Procedia - Social and Behavioral Sciences*, 22, 2–3. <https://doi.org/10.1016/j.sbspro.2011.07.050>

Fujimura, J. H. (1996). *Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer*. Harvard University Press.

Fuller, S. S., Revere, D., Bugni, P. F., & Martin, G. M. (2004). A knowledgebase system to enhance scientific discovery: Telemakus. *Biomedical Digital Libraries*, 1, 2. <https://doi.org/10.1186/1742-5581-1-2>

- Furedi, F. (2016). The Cultural Underpinning of Concept Creep. *Psychological Inquiry*, 27(1), 34–39. <https://doi.org/10.1080/1047840X.2016.1111120>
- Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. *Journal of English Linguistics*, 36(1), 5–38.
- Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, 122(3159), 108–111. <https://doi.org/10.1126/science.122.3159.108>
- Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science*, 178(4060), 471–479.
- Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The Use of Citation Data in Writing the History of Science*. Fort Belvoir, VA: Defense Technical Information Center. <https://doi.org/10.21236/AD0466578>
- Garrison, F. H. (1921). *An Introduction to the history of medicine c. 2*. W.B. Saunders Company.
- Gates, B. (2017, March 28). Of Microbes and Men: You should appreciate germs [Blog]. Retrieved September 14, 2018, from <https://www.gatesnotes.com/Books/I-Contain-Multitudes>
- Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, 4(7), eaaq1360. <https://doi.org/10.1126/sciadv.aaq1360>
- Gest, H. (2004). The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society. *Notes and Records of the Royal Society of London*, 58(2), 187–201.
- Ghadessy, M., Henry, A., & Roseberry, R. L. (2001). *Small Corpus Studies and ELT: Theory and Practice*. John Benjamins Publishing.
- Godin, B. (2006). The Knowledge-Based Economy: Conceptual Framework or Buzzword? *The Journal of Technology Transfer*, 31(1), 17–30. <https://doi.org/10.1007/s10961-005-5010-x>
- Gordon, J. I., Ley, R. E., Wilson, R., Mardis, E., Xu, J., Fraser, C. M., & Relman, D. A. (2005). *Extending Our View of Self: the Human Gut Microbiome Initiative (HGMI)*. <http://www.genome.gov/10002154>.
- Gormley, B. (2016, September 19). Microbiome Companies Attract Big Investments. *Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/microbiome-companies-attract-big-investments-1474250460>

- Gould, P. (1981). Letting the Data Speak for Themselves*. *Annals of the Association of American Geographers*, 71(2), 166–176. <https://doi.org/10.1111/j.1467-8306.1981.tb01346.x>
- Gray, R. (1992). Death of the Gene: Developmental Systems Strike Back. In P. Griffiths (Ed.), *Trees of Life: Essays in Philosophy of Biology* (pp. 165–209). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-8038-0_7
- Green, L. W., Ottoson, J. M., García, C., & Hiatt, R. A. (2009). Diffusion Theory and Knowledge Dissemination, Utilization, and Integration in Public Health. *Annual Review of Public Health*, 30(1), 151–174. <https://doi.org/10.1146/annurev.publhealth.031308.100049>
- Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ*, 331(7524), 1064–1065. <https://doi.org/10.1136/bmj.38636.593461.68>
- Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: a meta-narrative approach to systematic review. *Social Science & Medicine*, 61(2), 417–430. <https://doi.org/10.1016/j.socscimed.2004.12.001>
- Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation*, 48(6), e12931. <https://doi.org/10.1111/eci.12931>
- Green-Hennessy, S. (2013). Cochrane systematic reviews for the mental health field: is the gold standard tarnished? *Psychiatric Services (Washington, D.C.)*, 64(1), 65–70. <https://doi.org/10.1176/appi.ps.001682012>
- Grice, E. A., & Segre, J. A. (2012). The Human Microbiome: Our Second Genome. *Annual Review of Genomics and Human Genetics*, 13, 151–170. <https://doi.org/10.1146/annurev-genom-090711-163814>
- Gross, T. (n.d.). *Eating Yogurt Is Not Enough: Rebalancing The Ecosystem Of 'The Microbes Within Us'* : NPR. Retrieved from <https://www.npr.org/2016/08/18/490432969/eating-yogurt-is-not-enough-rebalancing-the-ecosystem-of-the-microbes-within-us>
- Grzeda, M. M. (2005). In competence we trust? Addressing conceptual ambiguity. *Journal of Management Development*, 24(6), 530–545. <https://doi.org/10.1108/02621710510600982>
- Gudivada, V. N., Rao, D., & Raghavan, V. V. (2015). Chapter 9 - Big Data Driven Natural Language Processing Research and Applications. In V. Govindaraju, V. V. Raghavan, & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 33, pp. 203–238). Elsevier. <https://doi.org/10.1016/B978-0-444-63492-4.00009-5>

- Guiliano, J., Fraistat, N., Brown, T., Muñoz, T., & Denbo, S. (2011). Shared Horizons: Data, Biomedicine, and the Digital Humanities. *National Endowment for the Humanities, Grant Submission, University of Maryland, College Park, MD.*
- Guo, L., Ning, Z., Hou, W., Hu, B., & Guo, P. (2018). Quick Answer for Big Data in Sharing Economy Innovative Computer Architecture Design Facilitating Optimal Service-Demand Matching. *Ieee Transactions on Automation Science and Engineering*, *15*(4), 1494–1506. <https://doi.org/10.1109/TASE.2018.2838340>
- Halliday, M. A. K., & Hasan, R. (1991). *Language, context, and text: aspects of language in a social-semiotic perspective*. Oxford University Press.
- Hampton-Marcell, J. T., Lopez, J. V., & Gilbert, J. A. (2017). The human microbiome: an emerging tool in forensics. *Microbial Biotechnology*, *10*(2), 228–230. <https://doi.org/10.1111/1751-7915.12699>
- Haslam, N. (2016). Concept Creep: Psychology’s Expanding Concepts of Harm and Pathology. *Psychological Inquiry*, *27*(1), 1–17. <https://doi.org/10.1080/1047840X.2016.1082418>
- Henig, R. M. (2017). *The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics*. Houghton Mifflin Harcourt.
- Herbst, T. (1996). What are collocations: sandy beaches or false teeth?
- Hesse, B. W., Moser, R. P., & Riley, W. T. (2015). From Big Data to Knowledge in the Social Sciences. *The Annals of the American Academy of Political and Social Science*, *659*(1), 16–32. <https://doi.org/10.1177/0002716215570007>
- Hettel, J. M. (2013). *Harnessing the Power of Context: A Corpus-based Analysis of Variation in the Language of the Regulated Nuclear Industry* (PhD Thesis). University of Georgia.
- Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Retrieved from <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature News*, *520*(7548), 429. <https://doi.org/10.1038/520429a>
- Hilpert, M., & Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, *24*(4), 385–401. <https://doi.org/10.1093/lc/fqn012>

Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semant. Web*, 4(3), 233–235.

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in Bioinformatics*, 16(6), 1069–1080. <https://doi.org/10.1093/bib/bbv011>

Holzinger, A., & Jurisica, I. (2014). Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. In A. Holzinger & I. Jurisica (Eds.), *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges* (pp. 1–18). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43968-5_1

Home | uBiome. (2018). Retrieved September 14, 2018, from <https://ubiome.com/>

Hooper, D. U., Chapin, F. S., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., ... Wardle, D. A. (2005). Effects of Biodiversity on Ecosystem Functioning: A Consensus of Current Knowledge. *Ecological Monographs*, 75(1), 3–35. <https://doi.org/10.1890/04-0922>

Hooper, L. V., & Gordon, J. I. (2001). Commensal Host-Bacterial Relationships in the Gut. *Science*, 292(5519), 1115–1118. <https://doi.org/10.1126/science.1058709>

Hoyles, L. (n.d.). What is the human gut microbiota? | Bugs In Your Guts [Blog]. Retrieved September 14, 2018, from <http://bugs-in-your-guts.com/?p=179>

Huang, C.-C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics*, 17(1), 132–144. <https://doi.org/10.1093/bib/bbv024>

Huang, M., Névéol, A., & Lu, Z. (2011). Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5), 660–667. <https://doi.org/10.1136/amiajnl-2010-000055>

Hudson, R. A. (1996). *Sociolinguistics*. Cambridge University Press.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Huntley, S. J., Mahlberg, M., Wiegand, V., van Gennip, Y., Yang, H., Dean, R. S., & Brennan, M. L. (2018). Analysing the opinions of UK veterinarians on practice-based research using corpus linguistic and mathematical methods. *Preventive Veterinary Medicine*, 150, 60–69. <https://doi.org/10.1016/j.prevetmed.2017.11.020>

Huss, J. (2014). Methodology and Ontology in Microbiome Research. *Biological Theory*, 9(4), 392–400. <https://doi.org/10.1007/s13752-014-0187-6>

- Hymes, D. (2013). *Foundations in sociolinguistics: An ethnographic approach*. Routledge.
- Hymes, D. H., & Gumperz, J. J. (1972). *Directions in sociolinguistics: The ethnography of communication*. Holt, Rinehart and Winston.
- Iltis, H. (1932). Life of Mendel. *Life of Mendel*. Retrieved from <https://www.cabdirect.org/cabdirect/abstract/19341601140>
- Impellizzeri, F. M., & Bizzini, M. (2012). SYSTEMATIC REVIEW AND META-ANALYSIS: A PRIMER. *International Journal of Sports Physical Therapy*, 7(5), 493–503.
- Information Overload Is Not Unique To Digital Age. (n.d.). Retrieved October 27, 2018, from <https://www.npr.org/2010/11/29/131671951/information-overload-is-not-unique-to-digital-age>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Medicine*, 11(10), e1001747. <https://doi.org/10.1371/journal.pmed.1001747>
- Ioannidis, J. P. A. (2016a). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *The Milbank Quarterly*, 94(3), 485–514. <https://doi.org/10.1111/1468-0009.12210>
- Ioannidis, J. P. A. (2016b). Why Most Clinical Research Is Not Useful. *PLOS Medicine*, 13(6), e1002049. <https://doi.org/10.1371/journal.pmed.1002049>
- Ioannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the Continuously Publishing Core in the Scientific Workforce. *PLOS ONE*, 9(7), e101698. <https://doi.org/10.1371/journal.pone.0101698>
- Jaccard, J., Wan, C. K., & Turrisi, R. (1990). The Detection and Interpretation of Interaction Effects Between Continuous Variables in Multiple Regression. *Multivariate Behavioral Research*, 25(4), 467–478. https://doi.org/10.1207/s15327906mbr2504_4
- Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New Phytologist*, 11(2), 37–50.
- Jaffe, A. B., & Trajtenberg, M. (1998). *International Knowledge Flows: Evidence from Patent Citations* (Working Paper No. 6507). National Bureau of Economic Research. <https://doi.org/10.3386/w6507>

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, 108(3), 577–598. <https://doi.org/10.2307/2118401>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer-Verlag. Retrieved from [//www.springer.com/us/book/9781461471370](http://www.springer.com/us/book/9781461471370)

Jones, S. (2013). Trends in microbiome research. *Nature Biotechnology*, 31, 277. <https://doi.org/10.1038/nbt.2546>

Jong, S., & Slavova, K. (2014). When publications lead to products: The open science conundrum in new product development. *Research Policy*, 43(4), 645–654. <https://doi.org/10.1016/j.respol.2013.12.009>

Juengst, E., & Huss, J. (2009). From Metagenomics to the Metagenome: Conceptual Change and the Rhetoric of Translational Genomic Research. *Genomics, Society and Policy*, 5(3), 1–19.

Jutz, G., van Rijn, P., Santos Miranda, B., & Böker, A. (2015). Ferritin: A Versatile Building Block for Bionanotechnology. *Chemical Reviews*, 115(4), 1653–1701. <https://doi.org/10.1021/cr400011b>

Kalutkiewicz, M. J., & Ehman, R. L. (2014). Patents as proxies: NIH hubs of innovation. *Nature Biotechnology*, 32, 536–537. <https://doi.org/10.1038/nbt.2917>

Kamdar, M. R., Tudorache, T., & Musen, M. A. (2017). A Systematic Analysis of Term Reuse and Term Overlap across Biomedical Ontologies. *Semantic Web*, 8(6), 853–871.

Kapoor, K. K., Dwivedi, Y. K., & Williams, M. D. (2014). Rogers' Innovation Adoption Attributes: A Systematic Review and Synthesis of Existing Research. *Information Systems Management*, 31(1), 74–91. <https://doi.org/10.1080/10580530.2014.854103>

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences*, 201424329. <https://doi.org/10.1073/pnas.1424329112>

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives* (pp. 154–175). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

Keller, E. F. (2002). *The Century of the Gene*. Cambridge, Massachusetts London: Harvard University Press.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, 59(7), 613–620. <https://doi.org/10.1525/bio.2009.59.7.12>

Kemp-Dynin, M. A. (2005). THE ‘COMPANY’ WORDS KEEP: A CORPUS-BASED ANALYSIS OF COLLOCATIONAL PATTERNING IN BUSINESS TERMINOLOGY.

Keramatfar, A., & Amirkhani, H. (2018). Bibliometrics of sentiment analysis literature. *Journal of Information Science*, 016555151876101. <https://doi.org/10.1177/0165551518761013>

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133. <https://doi.org/10.1075/ijcl.6.1.05kil>

Kim, S., Fiorini, N., Wilbur, W. J., & Lu, Z. (2017). Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. *Journal of Biomedical Informatics*, 75, 122–127. <https://doi.org/10.1016/j.jbi.2017.09.014>

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 2053951714528481. <https://doi.org/10.1177/2053951714528481>

Knorr-Cetina, K., & Mulkay, M. J. (1983). *Science Observed: Perspectives on the Social Study of Science*. Sage Publications.

Koehler, J., Rawlings, C., Verrier, P., Mitchell, R., Skusa, A., Ruegg, A., & Philippi, S. (2005). Linking Experimental Results, Biological Networks and Sequence Analysis Methods Using Ontologies and Generalised Data Structures. *In Silico Biology*, 5(1), 33–44.

Kokkinakis, D., & Gronostaj, M. T. (2006). Lay language versus professional language within the cardiovascular subdomain—a contrastive study. *Proc. of BIO'06*.

Kretzschmar Jr, W. A. (2010). Language variation and complex systems. *American Speech*, 85(3), 263–286.

Kretzschmar, W. A. (2009, March). The Linguistics of Speech by William A. Kretzschmar, Jr. <https://doi.org/10.1017/CBO9780511576782>

Kretzschmar, W. A. (2015). *Language and Complex Systems* (1 edition). New York: Cambridge University Press.

Kronberger, N., & Wagner, W. (2000a). Keywords in Context: Statistical Analysis of Text Features. Retrieved from <https://s3.amazonaws.com/academia.edu.documents/46889788/T4.5.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1540501900&Signature=iUAite6rQEqX9KGbk9MgVs0>

4Xzw%3D&response-content-disposition=inline%3B%20filename%3DKeywords_in_context_Statistical_analysis.pdf

Kronberger, N., & Wagner, W. (2000b). Keywords in context: Statistical analysis of text features. *Qualitative Researching with Text, Image and Sound. A Pratical Handbook*, 299–317.

Krumholz, H. M. (2014). Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Affairs (Project Hope)*, 33(7), 1163–1170. <https://doi.org/10.1377/hlthaff.2014.0053>

Kuhn, T. S. (1970). *The structure of scientific revolutions* ([2d ed., enl]). Chicago: University of Chicago Press.

La Rosa, M., Fiannaca, A., Rizzo, R., & Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC Bioinformatics*, 16(6), S2. <https://doi.org/10.1186/1471-2105-16-S6-S2>

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, W. (1994). Principles of linguistic change. Vol. 1: Internal features. *Language in Society*. Oxford: Blackwell.

Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Malden, Mass.: Blackwell Publishers.

Labov, W. (2006). *The Social Stratification of English in New York City*. Cambridge University Press.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and Opportunities with Big Data.

Lakoff, G. (1975). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. In D. Hockney, W. Harper, & B. Freed (Eds.), *Contemporary Research in Philosophical Logic and Linguistic Semantics: Proceedings of a Conference Held at the University of Western Ontario, London, Canada* (pp. 221–271). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-1756-5_9

Lamont, R. F., Sobel, J. D., Akins, R. A., Hassan, S. S., Chaiworapongsa, T., Kusanovic, J. P., & Romero, R. (2011). The vaginal microbiome: new information about genital tract flora using molecular based techniques. *BJOG: An International Journal of Obstetrics & Gynaecology*, 118(5), 533–549. <https://doi.org/10.1111/j.1471-0528.2010.02840.x>

Latour, B. (2009). 9 Tarde's idea of quantification, 18.

Laubichler, M. D., Maienschein, J., & Renn, J. (2013). Computational perspectives in the history of science: To the memory of Peter Damerow. *Isis*, *104*(1), 119–130.

Laubichler, M., & Renn, J. (2015). Extended Evolution.

Lawrence Edwards, N., Malouf, R., Perez-Ruiz, F., Richette, P., Southam, S., & DiChiara, M. (2016). Computational Lexical Analysis of the Language Commonly Used to Describe Gout. *Arthritis Care & Research*, *68*(6), 763–768. <https://doi.org/10.1002/acr.22746>

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. <https://doi.org/10.1126/science.1248506>

Lazer, D., Pentland, A. (Sandy), Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science (New York, N.Y.)*, *323*(5915), 721–723. <https://doi.org/10.1126/science.1167742>

Lederberg, J., & McCray, A. (2001). 'Ome Sweet 'Omics-- A Genealogical Treasury of Words. *The Scientist Magazine*®, *15*(7), 8.

Lederberg, Joshua. (2004). Of Men and Microbes. *New Perspectives Quarterly*, *21*(4), 92–96. <https://doi.org/10.1111/j.1540-5842.2004.00705.x>

Lee, M., Liu, Z., Kelly, R., & Tong, W. (2014). Of text and gene – using text mining methods to uncover hidden knowledge in toxicogenomics. *BMC Systems Biology*, *8*, 93. <https://doi.org/10.1186/s12918-014-0093-3>

Leech, G., Rayson, P., & Wilson, A. (All O. L. U. (2014). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Routledge.

Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 1–3. <https://doi.org/10.1016/j.shpsc.2011.10.001>

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.

Lewis, G., Rippon, I., & Wooding, S. (2005). Tracking knowledge diffusion through citations. *Research Evaluation*, *14*(1), 5–14. <https://doi.org/10.3152/147154405781776319>

Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, *43*(1), 5–25. <https://doi.org/10.1007/BF02458391>

- Leydesdorff, Loet. (2010). The Knowledge-based Economy and the Triple Helix Model. *Annual Rev. Info. Sci & Technol.*, 44(1), 365–417. <https://doi.org/10.1002/aris.2010.1440440116>
- Leydesdorff, Loet. (2013). Triple Helix of University-Industry-Government Relations. In *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship* (pp. 1844–1851). Springer, New York, NY. https://doi.org/10.1007/978-1-4614-3858-8_452
- Leydesdorff, Loet, Kogler, D. F., & Yan, B. (2017). Mapping patent classifications: portfolio and statistical analysis, and the comparison of strengths and weaknesses. *Scientometrics*, 112(3), 1573–1591. <https://doi.org/10.1007/s11192-017-2449-0>
- Li, H., An, H., Wang, Y., Huang, J., & Gao, X. (2016). Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network. *Physica A: Statistical Mechanics and Its Applications*, 450, 657–669. <https://doi.org/10.1016/j.physa.2016.01.017>
- Liang, T.-P., & Liu, Y.-H. (2018). Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study. *Expert Systems with Applications*, 111, 2–10. <https://doi.org/10.1016/j.eswa.2018.05.018>
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*.
- Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., & Zhu, S. (2015). MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics*, 31(12), i339–i347. <https://doi.org/10.1093/bioinformatics/btv237>
- Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1). <https://doi.org/10.1186/s40064-016-3252-8>
- Liu, Y., & Rousseau, R. (2010). Knowledge diffusion through publications and citations: A case study using ESI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology*, 61(2), 340–351. <https://doi.org/10.1002/asi.21248>
- Lobo, J., Bettencourt, L. M. A., Strumsky, D., & West, G. B. (2013). Urban Scaling and the Production Function for Cities. *PLOS ONE*, 8(3), e58407. <https://doi.org/10.1371/journal.pone.0058407>
- Löwy, I. (1988). Ludwik Fleck on the social construction of medical knowledge. *Sociology of Health & Illness*, 10(2), 133–155.
- Lu, Z., Kim, W., & Wilbur, W. J. (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*, 12(1), 69–80. <https://doi.org/10.1007/s10791-008-9074-8>

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., ... Dangl, J. L. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature*, 488(7409), 86–90. <https://doi.org/10.1038/nature11237>

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, 8, 1–10. <https://doi.org/10.4137/BII.S31559>

Luo, Z., Duffy, R., Johnson, S., & Weng, C. (2010). Corpus-based Approach to Creating a Semantic Lexicon for Clinical Research Eligibility Criteria from UMLS. *Summit on Translational Bioinformatics, 2010*, 26–30.

Lyytinen, K., Baskerville, R., Iivari, J., & Te'eni, D. (2007). Why the old world cannot publish? Overcoming challenges in publishing high-impact IS research. *European Journal of Information Systems*, 16(4), 317–326. <https://doi.org/10.1057/palgrave.ejis.3000695>

Maasen, S., & Weingart, P. (2013). *Metaphor and the Dynamics of Knowledge* (1 edition). London New York: Routledge.

MacRoberts, M. H., & MacRoberts, B. R. (2018). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology*, 69(3), 474–482. <https://doi.org/10.1002/asi.23970>

Malterud, K. (2001). The art and science of clinical knowledge: evidence beyond measures and numbers. *The Lancet*, 358(9279), 397–400. [https://doi.org/10.1016/S0140-6736\(01\)05548-9](https://doi.org/10.1016/S0140-6736(01)05548-9)

Mannheim, K. (1995). *Ideologie und Utopie*. (8. A. edition). Frankfurt am Main: Klostermann.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing* (1 edition). Cambridge, Mass: The MIT Press.

Mao, Y., & Lu, Z. (2017). MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. *Journal of Biomedical Semantics*, 8(1), 15. <https://doi.org/10.1186/s13326-017-0123-3>

Maramba, I. D., Davey, A., Elliott, M. N., Roberts, M., Roland, M., Brown, F., ... Campbell, J. (2015). Web-Based Textual Analysis of Free-Text Patient Experience Comments From a Survey in Primary Care. *JMIR Medical Informatics*, 3(2), e20. <https://doi.org/10.2196/medinform.3783>

Marchesi, J. R., & Ravel, J. (2015a). The vocabulary of microbiome research: a proposal. *Microbiome*, 3(1), 31. <https://doi.org/10.1186/s40168-015-0094-5>

Marchesi, J. R., & Ravel, J. (2015b). The vocabulary of microbiome research: a proposal. *Microbiome*, 3(1), 31. <https://doi.org/10.1186/s40168-015-0094-5>

- Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., ... Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association: JAMIA*, 21(6), 957–958. <https://doi.org/10.1136/amiajnl-2014-002974>
- Marler, S., Ferguson, B. J., Lee, E. B., Peters, B., Williams, K. C., McDonnell, E., ... Veenstra-VanderWeele, J. (2017). Association of Rigid-Compulsive Behavior with Functional Constipation in Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 47(6), 1673–1681. <https://doi.org/10.1007/s10803-017-3084-6>
- Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance* (1 edition). Chichester, West Sussex, United Kingdom ; Hoboken, New Jersey: Wiley.
- Masseroli, M., Chicco, D., & Pinoli, P. (2012). Probabilistic Latent Semantic Analysis for prediction of Gene Ontology annotations. In *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). Brisbane, Australia: IEEE. <https://doi.org/10.1109/IJCNN.2012.6252767>
- McCallum, A. (2018). MALLET. Retrieved October 26, 2018, from <https://people.cs.umass.edu/~mccallum/mallet/>
- McCosker, A., & Wilken, R. (2014). Rethinking 'big data' as visual knowledge: the sublime and the diagrammatic in data visualisation. *Visual Studies*, 29(2), 155–164. <https://doi.org/10.1080/1472586X.2014.887268>
- McGinty, S. (1999). *Gatekeepers of Knowledge*. Westport, CT: Bergen & Garvey. Retrieved from http://books.google.com/books/about/Gatekeepers_of_Knowledge.html?id=g6U4GwSgQY8C
- McGuire, W., & Albert, M. (2014). Understanding Change in Academic Knowledge Production in a Neoliberal Era. In *Fields of Knowledge: Science, Politics and Publics in the Neoliberal Age* (Vol. 27, pp. 33–57). Emerald Group Publishing Limited. <https://doi.org/10.1108/S0198-871920140000027009>
- Me, myself, us. (2012, August 18). *The Economist*. Retrieved from <https://www.economist.com/science-and-technology/2012/08/18/me-myself-us>
- Medical Subject Headings Preface. (2014, November). [Technical Documentation]. Retrieved October 3, 2018, from https://www.nlm.nih.gov/mesh/intro_preface.html#pref_hist
- Meillet, A. (1926). *Linguistique historique et linguistique generale* (2e ed). Paris: Honore Champion.

Merton, R. K. (1970). *Science, Technology and Society in Seventeenth-Century England*. Harper Torchbooks.

MeSH Browser. (n.d.). Retrieved October 3, 2018, from <https://meshb-prev.nlm.nih.gov/record/ui?ui=D064307>

Meyer, A. D., & Goes, J. B. (1988). Organizational Assimilation of Innovations: A Multilevel Contextual Analysis. *Academy of Management Journal*, 31(4), 897–923. <https://doi.org/10.5465/256344>

Miao, Y., Kešelj, V., & Milios, E. (2005). Document Clustering Using Character N-grams: A Comparative Evaluation with Term-based and Word-based Clustering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 357–358). New York, NY, USA: ACM. <https://doi.org/10.1145/1099554.1099665>

Milroy, L. (1987). *Language and Social Networks*. Wiley.

Milroy, L., & Gordon, M. (2003). *Sociolinguistics: Method and Interpretation* (2 edition). Malden, MA: Wiley-Blackwell.

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.

Moerchen, F., Fradkin, D., DeJori, M., & Wachmann, B. (2008). Emerging Trend Prediction in Biomedical Literature. *AMIA Annual Symposium Proceedings, 2008*, 485–489.

Mogoutov, A., Cambrosio, A., Keating, P., & Mustar, P. (2008). Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics*, 2(4), 341–353.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1), 1. <https://doi.org/10.1186/2046-4053-4-1>

Mohr, J. L. (1952). Protozoa as Indicators of Pollution. *The Scientific Monthly*, 74(1), 7–9.

Mole, B. (2013). Microbiome research goes without a home. *Nature News*, 500(7460), 16. <https://doi.org/10.1038/500016a>

Molyneux, R. E. (1994). What Did Rider Do? An Inquiry into the Methodology of Fremont Rider's "The Scholar and the Future of the Research Library." *Libraries & Culture*, 29(3), 297–325.

- Moody, J. (2004). The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. *American Sociological Review*, 69(2), 213–238. <https://doi.org/10.1177/000312240406900204>
- Moore, S., Neylon, C., Eve, M. P., O'Donnell, D. P., & Pattinson, D. (2017). “Excellence R Us”: university research and the fetishisation of excellence. *Palgrave Communications*, 3, 16105. <https://doi.org/10.1057/palcomms.2016.105>
- Moreno-Sánchez, I., Font-Clos, F., & Corral, Á. (2016). Large-Scale Analysis of Zipf’s Law in English Texts. *PLoS ONE*, 11(1). <https://doi.org/10.1371/journal.pone.0147073>
- Mortensen, J. M., Minty, E. P., Januszyk, M., Sweeney, T. E., Rector, A. L., Noy, N. F., & Musen, M. A. (2015). Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. *Journal of the American Medical Informatics Association*, 22(3), 640–648. <https://doi.org/10.1136/amiajnl-2014-002901>
- Morton, T. (2013). *Hyperobjects: Philosophy and Ecology after the End of the World* (1 edition). Minneapolis: Univ Of Minnesota Press.
- Mufwene, S. S. (2002). Competition and selection in language evolution. *Selection*, 3(1), 45–56.
- Mufwene, S. S. (2009). Restructuring, hybridization, and complexity in language evolution. *Complex Processes in New Languages*, 367–400.
- Mulrow, C. D. (1994). Systematic Reviews: Rationale for systematic reviews. *BMJ*, 309(6954), 597–599. <https://doi.org/10.1136/bmj.309.6954.597>
- Mutschke, P., Scharnhorst, A., Belkin, N. J., Skupin, A., & Mayr, P. (2017). Guest editors’ introduction to the special issue on knowledge maps and information retrieval (KMIR). *International Journal on Digital Libraries*, 18(1), 1–3. <https://doi.org/10.1007/s00799-016-0204-4>
- Naguib, A., Bencze, G., Cho, H., Zheng, W., Tocilj, A., Elkayam, E., ... Trotman, L. C. (2015). PTEN Functions by Recruitment to Cytoplasmic Vesicles. *Molecular Cell*, 58(2), 255–268. <https://doi.org/10.1016/j.molcel.2015.03.011>
- Natarajan, V., & Philipoff, P. (2018). Observation of surface and atmospheric parameters using “NOAA 18” satellite: a study on earthquakes of Sumatra and Nicobar Is regions for the year 2014 ($M \geq 6.0$). *Natural Hazards*, 92(2), 1097–1112. <https://doi.org/10.1007/s11069-018-3242-y>
- Neumann-Held, E. M. (1999). The Gene Is Dead — Long Live the Gene! Conceptualizing Genes the Constructionist Way. In P. Koslowski (Ed.), *Sociobiology and Bioeconomics* (pp. 105–137). Springer Berlin Heidelberg.

- Newell, S., Robertson, M., Scarbrough, H., & Swan, J. (2009). *Managing Knowledge Work and Innovation, 2nd Edition* (Second edition). Basingstoke: Palgrave Macmillan.
- Newman, M., Barabási, A.-L., & Watts, D. J. (2011). *The Structure and Dynamics of Networks*. Princeton University Press.
- Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2), 404–409. <https://doi.org/10.1073/pnas.98.2.404>
- Ng, R., Allore, H. G., Trentalange, M., Monin, J. K., & Levy, B. R. (2015). Increasing Negativity of Age Stereotypes across 200 Years: Evidence from a Database of 400 Million Words. *PLoS ONE*, 10(2). <https://doi.org/10.1371/journal.pone.0117086>
- NIH Categorical Spending -NIH Research Portfolio Online Reporting Tools (RePORT). (n.d.). Retrieved October 31, 2018, from https://report.nih.gov/categorical_spending.aspx
- NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., ... Guyer, M. (2009). The NIH Human Microbiome Project. *Genome Research*, 19(12), 2317–2323. <https://doi.org/10.1101/gr.096651.109>
- Norredam, M., & Album, D. (2007). Review Article: Prestige and its significance for medical specialties and diseases. *Scandinavian Journal of Public Health*, 35(6), 655–661. <https://doi.org/10.1080/14034940701362137>
- Noyons, E. (2001). Bibliometric mapping of science in a policy context. *Scientometrics*, 50(1), 83–98. <https://doi.org/10.1023/A:1005694202977>
- Oakes, M. P., & Farrow, M. (2007). Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries. *Literary and Linguistic Computing*, 22(1), 85–99. <https://doi.org/10.1093/lc/fql044>
- Olfati-Saber, R., Fax, J. A., & Murray, R. M. (2007). Consensus and Cooperation in Networked Multi-Agent Systems. *Proceedings of the IEEE*, 95(1), 215–233. <https://doi.org/10.1109/JPROC.2006.887293>
- O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, 4(1), 5. <https://doi.org/10.1186/2046-4053-4-5>
- Owen-Smith, J., & Powell, W. W. (2004). Knowledge Networks as Channels and Conduits: The Effects of Spillovers in the Boston Biotechnology Community. *Organization Science*, 15(1), 5–21. <https://doi.org/10.1287/orsc.1030.0054>

- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 94.
- Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review - a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, 10(1_suppl), 21–34. <https://doi.org/10.1258/1355819054308530>
- Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., & Zhu, S. (2016). DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, 32(12), i70–i79. <https://doi.org/10.1093/bioinformatics/btw294>
- Pérez, S., Laperrière, V., Borderon, M., Padilla, C., Maignant, G., & Oliveau, S. (2016). Evolution of research in health geographics through the International Journal of Health Geographics (2002–2015). *International Journal of Health Geographics*, 15(1), 3. <https://doi.org/10.1186/s12942-016-0032-1>
- Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology*, 5(7), e1000443. <https://doi.org/10.1371/journal.pcbi.1000443>
- Petticrew, M., & Roberts, H. (2005). *Systematic Reviews in the Social Sciences: A Practical Guide* (1 edition). Malden, MA ; Oxford: Wiley-Blackwell.
- Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5), 297–312. [https://doi.org/10.1016/0306-4573\(76\)90048-0](https://doi.org/10.1016/0306-4573(76)90048-0)
- Plsek, P., & Plsek, P. E. (2003). *Complexity and the Adoption of Innovation in Health Care* (p. 18). National Institute for Health Care Management Foundation.
- pmhdev. (n.d.). PubMed Health - National Library of Medicine. Retrieved October 3, 2018, from <https://www.ncbi.nlm.nih.gov/pubmedhealth/aboutnlm/>
- Popper, K. (2002). *Conjectures and Refutations: The Growth of Scientific Knowledge* (2nd edition). London ; New York: Routledge.
- Porter, C., Atkinson, P., & Gregory, I. (2015). Geographical Text Analysis: A new approach to understanding nineteenth-century mortality. *Health & Place*, 36, 25–34. <https://doi.org/10.1016/j.healthplace.2015.08.010>
- Portin, P. (2002). Historical Development of the Concept of the Gene. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 27(3), 257–286. <https://doi.org/10.1076/jmep.27.3.257.2980>

- Powell, W. W., & Snellman, K. (2004). The Knowledge Economy. *Annual Review of Sociology*, 30(1), 199–220. <https://doi.org/10.1146/annurev.soc.29.010202.100037>
- Preidis, G. A., & Versalovic, J. (2009). Targeting the Human Microbiome With Antibiotics, Probiotics, and Prebiotics: Gastroenterology Enters the Metagenomics Era. *Gastroenterology*, 136(6), 2015–2031.
- Prensky, M. (2009). H. Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom, 11.
- Prescott, S. L. (2017). History of medicine: Origin of the term microbiome and why it matters. *Human Microbiome Journal*, 4, 24–25. <https://doi.org/10.1016/j.humic.2017.05.004>
- Price, D. De Solla. (1986). *Little Science, Big Science...and Beyond*. New York: Columbia Univ Pr.
- Price, Derek De Solla. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5), 292–306. <https://doi.org/10.1002/asi.4630270505>
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., ... Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285), 59–U70. <https://doi.org/10.1038/nature08821>
- Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 1871–1887. <https://doi.org/10.1002/asi.21368>
- Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 248–256). Singapore: Association for Computational Linguistics. Retrieved from <http://www.aclweb.org/anthology/D/D09/D09-1026>
- Ramaprasad, A., & Syn, T. (2015). Ontological Meta-Analysis and Synthesis, 37, 17.
- Rawat, S., & Meena, S. (2014). Publish or perish: Where are we heading? *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, 19(2), 87–89.
- Rayson, P., & Garside, R. (1997). Comparing Corpora using Frequency Profiling, 6.
- Rees, T., Bosch, T., & Douglas, A. E. (2018). How the microbiome challenges our concept of self. *PLoS Biology*, 16(2). <https://doi.org/10.1371/journal.pbio.2005358>

Regier, D. A., Myers, J. K., Kramer, M., Robins, L. N., Blazer, D. G., Hough, R. L., ... Locke, B. Z. (1984). The NIMH Epidemiologic Catchment Area Program: Historical Context, Major Objectives, and Study Population Characteristics. *Archives of General Psychiatry*, *41*(10), 934–941. <https://doi.org/10.1001/archpsyc.1984.01790210016003>

Reinhart, A. (2015). *Statistics Done Wrong: The Woefully Complete Guide* (1 edition). San Francisco: No Starch Press.

Relman, D. A., & Falkow, S. (2001). The meaning and impact of the human genome sequence for microbiology. *Trends in Microbiology*, *9*(5), 206–208.

Renganathan, V. (2017). Text Mining in Biomedical Domain with Emphasis on Document Clustering. *Healthcare Informatics Research*, *23*(3), 141–146. <https://doi.org/10.4258/hir.2017.23.3.141>

Renn, J., & Laubichler, M. (2017a). Extended Evolution and the History of Knowledge. In *Integrated History and Philosophy of Science* (pp. 109–125). Springer, Cham. https://doi.org/10.1007/978-3-319-53258-5_9

Renn, J., & Laubichler, M. (2017b). Extended Evolution and the History of Knowledge. *Integrated History and Philosophy of Science*, 109–125. https://doi.org/10.1007/978-3-319-53258-5_9

Reppen, R., Fitzmaurice, S. M., & Biber, D. (2002). *Using Corpora to Explore Linguistic Variation*. John Benjamins Publishing.

Revue odontologique. (1949). Société syndicate odontologique de France.

Rey, J. M. (2014). *Changing Gender Roles in Popular Culture: dialogue in Star Trek episodes from 1996 to 1993*. Routledge.

Rheinburger, H.-J., & Muller-Wille, S. (2017). *The Gene: From Genetics to Postgenomics*. The University of Chicago Press. Retrieved from <https://www.press.uchicago.edu/ucp/books/book/chicago/G/bo20952390.html>

Rhodes, R., Gligorov, N., & Schwab, A. P. (2013). *The Human Microbiome: Ethical, Legal and Social Concerns*. Oxford University Press.

Rivera, M. T., Soderstrom, S. B., & Uzzi, B. (2010). Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annual Review of Sociology*, *36*(1), 91–115. <https://doi.org/10.1146/annurev.soc.34.040507.134743>

- Roach, M., & Cohen, W. M. (2012). Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research. *Management Science*, *59*(2), 504–525. <https://doi.org/10.1287/mnsc.1120.1644>
- Rodriguez, M. A., & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics*, *2*(3), 195–201. <https://doi.org/10.1016/j.joi.2008.04.002>
- Rogawski, E. T., Westreich, D. J., Adair, L. S., Becker-Dreps, S., Sandler, R. S., Sarkar, R., ... Kang, G. (2015). Early Life Antibiotic Exposure Is Not Associated with Growth in Young Children of Vellore, India. *The Journal of Pediatrics*, *167*(5), 1096-102.e3. <https://doi.org/10.1016/j.jpeds.2015.08.015>
- Rogers, E. M. (2010). *Diffusion of Innovations, 4th Edition*. Simon and Schuster.
- Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents, 8.
- Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123. <https://doi.org/10.1073/pnas.0706851105>
- Rowlands, I. (2002). Journal diffusion factors: a new approach to measuring research influence. *Aslib Proceedings*, *54*(2), 77–84. <https://doi.org/10.1108/00012530210435211>
- Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., ... Musen, M. A. (2006). National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. *OMICS: A Journal of Integrative Biology*, *10*(2), 185–198. <https://doi.org/10.1089/omi.2006.10.185>
- Sandler, I., & Sandler, L. (1985). A Conceptual Ambiguity that Contributed to the Neglect of Mendel's Paper. *History and Philosophy of the Life Sciences*, *7*(1), 3–70.
- Sandström, U., & Besselaar, P. van den. (2016). Quantity and/or Quality? The Importance of Publishing Many Papers. *PLOS ONE*, *11*(11), e0166149. <https://doi.org/10.1371/journal.pone.0166149>
- Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLOS ONE*, *4*(9), e7078. <https://doi.org/10.1371/journal.pone.0007078>
- Scheler, M. (2012). *Problems of a Sociology of Knowledge* (1 edition). London: Routledge.
- Scher, J. U., & Abramson, S. B. (2011). The microbiome and rheumatoid arthritis. *Nature Reviews Rheumatology*, *7*(10), 569–578. <https://doi.org/10.1038/nrrheum.2011.121>

- Schmid, H.-J. (2003). Do women and men really live in different cultures? Evidence from the BNC. In *Corpus linguistics by the Lune: a Festschrift for Geoffrey Leech* (p. 1). Peter Lang.
- Schneider, G. W., & Winslow, R. (2014). Parts and wholes: the human microbiome, ecological ontology, and the challenges of community. *Perspectives in Biology and Medicine*, 57(2), 208–223. <https://doi.org/10.1353/pbm.2014.0016>
- Science: 336 (6086). (2012). *Science*, 336(6086). Retrieved from <http://science.sciencemag.org/content/336/6086>
- Science: 352 (6285). (2016). *Science*, 352(6285). Retrieved from <http://science.sciencemag.org/content/352/6285>
- Science, A. A. for the A. of. (2011). The Runners-Up. *Science*, 334(6063), 1629–1635. <https://doi.org/10.1126/science.334.6063.1629>
- Science, A. A. for the A. of. (2013). Your Microbes, Your Health. *Science*, 342(6165), 1440–1441. <https://doi.org/10.1126/science.342.6165.1440-b>
- Scott, J. (2017). *Social Network Analysis*. SAGE.
- Scott, J., & Carrington, P. J. (2011). *The SAGE Handbook of Social Network Analysis*. SAGE.
- Scott, M. (2018). Wordsmith Tools. Stroud: Lexical Analysis Software. Retrieved from https://lexically.net/publications/citing_wordsmith.htm
- Seale, C., Ziebland, S., & Charteris-Black, J. (2006). Gender, cancer experience and internet use: A comparative keyword analysis of interviews and online cancer support groups. *Social Science & Medicine*, 62(10), 2577–2590. <https://doi.org/10.1016/j.socscimed.2005.11.016>
- Search Results - NIH RePORTER - NIH Research Portfolio Online Reporting Tools Expenditures and Results. (n.d.). Retrieved October 30, 2018, from https://projectreporter.nih.gov/reporter_searchresults.cfm
- Seneta, E. (2006). MARKOV AND THE CREATION OF MARKOV CHAINS., 20.
- Shade, A., & Handelsman, J. (2012). Beyond the Venn diagram: the hunt for a core microbiome. *Environmental Microbiology*, 14(1), 4–12. <https://doi.org/10.1111/j.1462-2920.2011.02585.x>
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27, 379–423, 623–656.
- Shapin, S. (1996). *The Scientific Revolution*. Retrieved from <https://www.press.uchicago.edu/ucp/books/book/chicago/S/bo3620548.html>

- Shi, F., Foster, J. G., & Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43, 73–85. <https://doi.org/10.1016/j.socnet.2015.02.006>
- Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Institute of Electrical and Electronics Engineers*, 8.
- Sigley, R., & Holmes, J. (2002). Looking at girls in Corpora of English. *Journal of English Linguistics*, 30(2), 138–157. <https://doi.org/10.1177/007242030002004>
- Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences of the United States of America*, 112(2), 360–365. <https://doi.org/10.1073/pnas.1418218112>
- Simpson-Vlach, R. C., & Leicher, S. (2006). *The MICASE Handbook*. University of Michigan Press. Retrieved from https://www.press.umich.edu/101203/micase_handbook
- Smalheiser, N. R. (2017). Rediscovering Don Swanson: the Past, Present and Future of Literature-Based Discovery. *Journal of Data and Information Science (Warsaw, Poland)*, 2(4), 43–64. <https://doi.org/10.1515/jdis-2017-0019>
- Song, M., & Kim, S. Y. (2013). Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics*, 96(1), 183–201. <https://doi.org/10.1007/s11192-012-0900-9>
- Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3), 239–251. <https://doi.org/10.1093/bib/6.3.239>
- Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why Has the Number of Scientific Retractions Increased? *PLOS ONE*, 8(7), e68397. <https://doi.org/10.1371/journal.pone.0068397>
- Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243. <https://doi.org/10.1075/ijcl.8.2.03ste>
- Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biology*, 13(7). <https://doi.org/10.1371/journal.pbio.1002195>
- Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 85–87. <https://doi.org/10.1016/j.shpsc.2011.10.009>

- Stubbs, M. (1983). *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*. University of Chicago Press.
- Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1), 23–55.
- Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics* (1 edition). Oxford England ; Malden, MA: Blackwell Publishing.
- Sun, X., Fiala, J. L. A., & Lowery, D. (2016). Patent watch: Modulating the human microbiome with live biotherapeutic products: intellectual property landscape. *Nature Reviews. Drug Discovery*, 15(4), 224–225. <https://doi.org/10.1038/nrd.2016.48>
- Surman, J., Stráner, K., & Haslinger, P. (2014). Introduction: Nomadic Concepts—Biological Concepts and Their Careers Beyond Biology. *Contributions to the History of Concepts*, 9(2), 1–17.
- Swan, J., Bresnen, M., Newell, S., & Robertson, M. (2007). The object of knowledge: The role of objects in biomedical innovation. *Human Relations*, 60(12), 1809–1837. <https://doi.org/10.1177/0018726707084915>
- Swan, M. (2013). The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data*, 1(2), 85–99. <https://doi.org/10.1089/big.2012.0002>
- Tabah, A. N. (1999). Literature Dynamics: Studies on Growth, Diffusion, and Epidemics. *Annual Review of Information Science and Technology (ARIST)*, 34, 249–286.
- Thagard, P. (1992). *Conceptual Revolutions* (Reprint edition). Princeton: Princeton University Press.
- Thagard, P. R. (1988). *Computational Philosophy of Science*. Cambridge, Mass: MIT Pr.
- The Human Microbiome Project Consortium, Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., ... White, O. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207–214. <https://doi.org/10.1038/nature11234>
- Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1), 1–14. <https://doi.org/10.1002/jrsm.27>
- Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering*, 69, 1356–1364. <https://doi.org/10.1016/j.proeng.2014.03.129>

Trudgill, P. (1979). *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting. *Marketing Science*, 35(3), 405–426. <https://doi.org/10.1287/mksc.2015.0956>

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, Connecticut: Graphics Pr.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn: Graphics Press.

Tufte, E. R. (1998). *The Visual Display of Quantitative Information* (16th Printing edition). Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis* (1 edition). Reading, Mass: Pearson.

Türker, İ., Şehirli, E., & Demiral, E. (2016). Uncovering the differences in linguistic network dynamics of book and social media texts. *SpringerPlus*, 5(1), 864. <https://doi.org/10.1186/s40064-016-2598-2>

Turnbaugh, P. J., Hamady, M., Yatsunencko, T., Cantarel, B. L., Duncan, A., Ley, R. E., ... Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228), 480–484. <https://doi.org/10.1038/nature07540>

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. *Nature*, 449, 804–810. <https://doi.org/10.1038/nature06244>

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122), 1027–1031. <https://doi.org/10.1038/nature05414>

UMLS Quick Start Guide. (n.d.). [Training Material and Manuals]. Retrieved October 22, 2018, from <https://www.nlm.nih.gov/research/umls/quickstart.html>

Underwood, T. (2017). A Genealogy of Distant Reading. *Digital Humanities Quarterly*, 011(2).

Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome. *Nutrition Reviews*, 70(suppl_1), S38–S44. <https://doi.org/10.1111/j.1753-4887.2012.00493.x>

Valente, T. W. (1996). Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory*, 2(2), 163–164. <https://doi.org/10.1007/BF00240425>

- van Bijnen, E. M. E., Paget, W. J., den Heijer, C. D. J., Stobberingh, E. E., Bruggeman, C. A., Schellevis, F. G., & APRES Study Team. (2014). Primary care treatment guidelines for skin infections in Europe: congruence with antimicrobial resistance found in commensal *Staphylococcus aureus* in the community. *BMC Family Practice*, *15*, 175. <https://doi.org/10.1186/s12875-014-0175-8>
- Van Noorden, R. (2014). Publishers withdraw more than 120 gibberish papers. *Nature*. <https://doi.org/10.1038/nature.2014.14763>
- van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, *59*(3), 467–472. <https://doi.org/10.1023/B:SCIE.0000018543.82441.f1>
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304–1351. <https://doi.org/10.1126/science.1058040>
- Vigen, T. (2015). *Spurious Correlations* (Gift edition). New York: Hachette Books.
- VIOME- ga2. (2018). Retrieved September 14, 2018, from https://www.viome.com/ga2?gclid=CjwKCAjwuO3cBRAyEiwAzOxKsliVGjkqceYEc5Oz0UjQwrl2PYm6PBUtxd3A_eFR1xLFcUt2yP-dJBoCioMQAvD_BwE
- Walsh, J. P., Cho, C., & Cohen, W. M. (2005). View from the Bench: Patents and Material Transfers. *Science*, *309*(5743), 2002–2003. <https://doi.org/10.1126/science.1115813>
- Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872. <https://doi.org/10.1007/s11192-012-0775-9>
- Wardhaugh, R. (2009). *An Introduction to Sociolinguistics* (6 edition). Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell.
- Ware, C. (2012). *Information Visualization: Perception for Design* (3 edition). Waltham, MA: Morgan Kaufmann.
- Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.
- Watson, J. D., & Cook-Deegan, R. M. (1990). The Human Genome Project and International Health. *JAMA*, *263*(24), 3322–3324. <https://doi.org/10.1001/jama.1990.03440240112027>
- Weiner, J. (2018, January 20). Human Cells Make Up Only Half Our Bodies. A New Book Explains Why. *The New York Times*. Retrieved from <https://www.nytimes.com/2016/08/21/books/review/i-contain-multitudes-ed-yong.html>

Whats New for 2018 MeSH. (2017, November). [Newsletters]. Retrieved August 13, 2018, from https://www.nlm.nih.gov/pubs/techbull/nd17/nd17_mesh.html

Wheelan, C., & Davis, J. (2014). *Naked Statistics: Stripping the Dread from the Data* (MP3 edition). Brilliance Audio.

Whipps, J. M., Lewis, K., & Cooke, R. C. (1988). *Fungi in biological control systems*.

White, H. D., & McCain, K. W. (1997). Visualization of Literatures. *Annual Review of Information Science and Technology (ARIST)*, 32, 99–168.

White, W. L. (1998). *Slaying the Dragon: The History of Addiction Treatment and Recovery in America* (1st edition). Bloomington, Ill: Chestnut Health Systems.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12), 6578–6583. <https://doi.org/10.1073/pnas.95.12.6578>

Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A Survey on the Role of Negation in Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing* (pp. 60–68). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1858959.1858970>

Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), 151–171.

Witten, I. H., & Eibe, F. (2001). *Data Mining*. München: Hanser Fachbuch.

Yang, H. (2013). HGP and -omics: Big Science and Big Data. *Febs Journal*, 280, 73–73.

Yeh, J.-Y., Ke, H.-R., Yang, W.-P., & Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management*, 41(1), 75–95. <https://doi.org/10.1016/j.ipm.2004.04.003>

Yetton, P., Sharma, R., & Southon, G. (1999). Successful IS innovation: the contingent contributions of innovation characteristics and implementation process. *Journal of Information Technology*, 14(1), 53–68. <https://doi.org/10.1080/026839699344746>

Yong, E. (2016). *I Contain Multitudes: The Microbes Within Us and a Grander View of Life* (1 edition). New York, NY: Ecco.

Zadeh, L. A. (1976). A Fuzzy-Algorithmic Approach to the Definition of Complex or Imprecise Concepts. In H. Bossel, S. Klaczko, & N. Müller (Eds.), *Systems Theory in the Social Sciences:*

Stochastic and Control Systems Pattern Recognition Fuzzy Analysis Simulation Behavioral Models (pp. 202–282). Basel: Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-5495-5_11

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *ArXiv:1509.01626 [Cs]*. Retrieved from <http://arxiv.org/abs/1509.01626>

Zipf, George K. (1935). *The psychology of language*. NY Houghton-Mifflin.

Zipf, George Kingsley. (1946). The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review*, *11*(6), 677–686. <https://doi.org/10.2307/2087063>

Zipf, George Kingsley. (2012). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Mansfield Centre, Conn: Martino Fine Books.

Zipf, George Kingsley. (2016). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Ravenio Books.