


7-7-2016

Relevance is in the Eye of the Beholder: Design Principles for the Extraction of Context-Aware Information

Arturo Castellanos

Florida International University, acast317@fiu.edu

Follow this and additional works at: <http://digitalcommons.fiu.edu/etd>

 Part of the [Business Administration, Management, and Operations Commons](#), [Business Intelligence Commons](#), [Health Services Administration Commons](#), [Management Information Systems Commons](#), and the [Organizational Behavior and Theory Commons](#)

Recommended Citation

Castellanos, Arturo, "Relevance is in the Eye of the Beholder: Design Principles for the Extraction of Context-Aware Information" (2016). *FIU Electronic Theses and Dissertations*. Paper 2543.
<http://digitalcommons.fiu.edu/etd/2543>

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

RELEVANCE IS IN THE EYE OF THE BEHOLDER: DESIGN PRINCIPLES FOR
THE EXTRACTION OF CONTEXT-AWARE INFORMATION

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BUSINESS

by

Arturo Castellanos

2016

To: Acting Dean Jose M. Aldrich
College of Business

This dissertation, written by Arturo Castellanos, and entitled *Relevance is in the Eye of the Beholder: Design Principles for the Extraction of Context-Aware Information*, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Benjamin Amick III

Richard Klein

Roman Lukyanenko

Monica Chiarini Tremblay, Major Professor

Date of Defense: July 6, 2016

The dissertation of Arturo Castellanos is approved.

Acting Dean Jose M. Aldrich
College of Business

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2016

© Copyright 2016 by Arturo Castellanos

All rights reserved.

ACKNOWLEDGMENTS

It is difficult to capture in a few paragraphs how grateful I am to those who have made this journey possible. First and foremost, I am grateful to God for constantly looking out for me and surrounding me with such an amazing and talented group of individuals that pushed me everyday to be a better person. I am grateful to my parents for their continuous support and being my role models. They always encouraged me throughout the program even when I was not able to explain what it is that I was doing. To my brother and sister for always being there for me and for reminding me of the important things in life. To my brother from another mother, Alfred Castillo, your friendship is one of things I treasure the most about the Ph.D. program.

I want to give special thanks to my advisor Dr. Monica Tremblay, my mentor, to whom I deeply respect and admire as a scholar and (more importantly) as a person. It is a blessing to have crossed paths with you. So many things to give you thanks for... including your family adopting me in Puerto Rico. I would not be here without your continuous support and encouragement. It all started with the opportunity you and Dr. Gloria Deckard gave to me.

I am indebted to my committee members: Roman Lukyanenko, Rich Klein, and Benjamin Amick. This would have not been possible without your continuous support and feedback. I want to give special thanks to Roman, my respect and admiration; I would not be writing this without your guidance and mentorship. Thank you for your unconditional support and always inspiring those around you. Thanks for sharing with me your passion for research on psychology, design work, conceptual modeling, and IQ —

key components of my dissertation and future career. Thank you Dr. Klein for giving me the opportunity to teach the analytics class to undergrads and for trusting me a full class of business majors the day before classes started. Thanks to Dr. Amick for your feedback and giving a broader view to my work. I look forward to continue working with you.

I would like to share this accomplishment with all the faculty, staff, and PhD students in the Department of Information Systems and Business Analytics at FIU. I will forever cherish the memories I have with you and I hope we keep in touch for the years to come. Last but not least I want to thank FIU for a world-class education on a beautiful campus in an amazing city.

ABSTRACT OF THE DISSERTATION

RELEVANCE IS IN THE EYE OF THE BEHOLDER: DESIGN PRINCIPLES
FOR THE EXTRACTION OF CONTEXT-AWARE INFORMATION

by

Arturo Castellanos

Florida International University, 2016

Miami, Florida

Professor Monica Chiarini Tremblay, Major Professor

Since the 1970s many approaches of representing domains have been suggested. Each approach maintains the assumption that the information about the objects represented in the Information System (IS) is specified and verified by domain experts and potential users. Yet, as more IS are developed to support a larger diversity of users such as customers, suppliers, and members of the general public (such as many multi-user online systems), analysts can no longer rely on a stable single group of people for complete specification of domains –to the extent that prior research has questioned the efficacy of conceptual modeling in these heterogeneous settings. We formulated principles for identifying basic classes in a domain. These classes can guide conceptual modeling, database design, and user interface development in a wide variety of traditional and emergent domains. Moreover, we used a case study of a large foster organization to study how unstructured data entry practices result in differences in how information is collected across organizational units. We used institutional theory to show how institutional elements enacted by individuals can generate new practices that can be adopted over time as best practices. We analyzed free-text notes to prioritize potential

cases of psychotropic drug use—our tactical need. We showed that too much flexibility in how data can be entered into the system, results in different styles, which tend to be homogenous across organizational units but not across organizational units. Theories in Psychology help explain the implications of the level of specificity and the inferential utility of the text encoded in the unstructured note.

TABLE OF CONTENTS

CHAPTER	PAGE
Chapter 1: Dissertation Overview	1
Overview of the Essays.....	1
Chapter 2: Basic Classes in Conceptual Modeling: Theoretical Foundations and Practical Guidelines.....	3
Emergent Challenges of Selecting Classes in Conceptual Modeling	6
Open information environments online.....	7
Other emerging domains	10
Conceptual Motivation for Guidelines	15
Guidelines for Identifying Basic Classes in Conceptual Modeling.....	25
Guideline 1. Middle level.....	25
Guideline 2. Entry Category.....	28
Guideline 3. Frequent Words	34
Guideline 4. Cohesion and Coupling	36
Guideline 5: Object Visualization	38
Guideline 6: Simplest Words	40
Guideline 7: Original Words	42
Guideline 8: CU Coefficient	44
Summary of Guidelines.....	45
Implications, Contributions, and Conclusions.....	47
Chapter 3: Identifying Organizational Style: An Institutional Theory Perspective	53
Motivation.....	58
Child Welfare.....	59
Federal and State Legislation	60
Foster Care and Case Management	61
Identifying Psychotropic Drug Use.....	63
Theory and Propositions	67
Institutional Theory.....	68

Psychological Foundation.....	76
Method.....	80
Solution Approach.....	82
Text Mining.....	82
Stylometry.....	85
Data Preparation.....	86
Analysis and Results.....	89
Discussion.....	100
REFERENCES.....	104
VITA.....	126

LIST OF TABLES

TABLE	PAGE
Chapter 2:	
Table 1. Some examples of basic-level categories from psychology studies	20
Table 2. Guidelines for Identifying Basic Classes in Conceptual Modeling	25
Table 3. Combinations and relation between the different term pairs	38
Table 4. Feature probabilities to illustrate Corter and Gluck model	45
Table 5. Category probabilities and CU measures to illustrate Corter and Gluck model ..	45
Chapter 3:	
Table 1. Unit of Analysis of Institutionalization	69
Table 2. Institutional Theory Elements	72
Table 3. Parts-of-speech	83
Table 4. Stylometric Identification	86
Table 5. Evaluation Metrics	89
Table 6. Results in difference between proportions for Precision (P) and Recall (R)	92
Table 7. Case Distribution across Agencies	94
Table 8. Prediction results with features disabled	96
Table 9. Results in difference between proportions for Precision (P) and Recall (R)	97
Table 10. Home-visit notes of children taking psychotropic medication	99

LIST OF FIGURES

FIGURE	PAGE
Chapter 2:	
Figure 1. Online quiz on iSpot that trains online volunteers to identify species of interest.....	9
Figure 2. Fragment of an ENT conceptual model for PatientsLikeMe.....	13
Figure 3: Modeling citizen science	32
Figure 4: Modeling the admission process.....	34
Chapter 3:	
Figure 1. Evaluation Metrics for each of the FCMA's (Agencies)	66
Figure 2. SQL Database Structure.....	88
Figure 3: Data mining process (process followed for Agency A).....	90
Figure 4. Intra and Inter-Agency Data mining process.....	91
Figure 5: Predicting Agency (Stylometry).....	94

Chapter 1: Dissertation Overview

In the course of normal business, organizations generate electronic documentation describing daily operations and transactions. The purpose of this documentation is generally tactical. For example, an IT help desk staff documents reported technical issues, a police officer enters the details of an incident, or a clinician documents a case in progress notes. These data are often stored and organized at the point of capture, and reflects the daily transactions of the organization's business activities –as modeled by the information system. This dissertation research explores two main topics: deriving design principles to guide conceptual modeling of open information environments and the institutionalization of data-entry practices of unstructured and semi-structured data in an organization and its implications.

Overview of the Essays

Since the 1970s many approaches to representing domains have been suggested. Each approach maintains the assumption that the information about the objects represented in the Information System (IS) is specified and verified by domain experts and potential users. Yet, as more IS are developed to support a larger diversity of users, analysts can no longer rely on a stable single group of people for complete specification of domains. This first chapter provides theoretical guidelines rooted in psychology for the existence and the importance of special classes termed in psychology basic level categories. We formulate principles for identifying basic classes in a domain. These classes can guide conceptual modeling, database design, and user interface development in a wide variety of traditional and emergent domains. Previous research has leveraged

ontologies to add a common understanding in communicating information (Gruber 1995). We illustrate these principles in a healthcare setting, particularly in the context of an Ear, Nose, and Throat (ENT) ontology. These guidelines can be generalized to other domains.

Given the shortcomings of traditional approaches to modeling structured IS and the extent to which existing IS relies on unstructured data, the third chapter proposes theory-based propositions that can provide guidance in designing and modeling information systems that rely on unstructured data-entry formats. One of the challenges of unstructured data is the inherent flexibility of how these data are entered/captured in an information system (e.g., free-form text) as opposed to a less flexible structured format (e.g., selecting from drop-down lists).

In the third chapter, we show that in the day-to-day operations individuals may deviate (to different extent) on how they input data into the system. These deviations can be based on the individual's training or based on immediate needs/pressures demanded by their units, impacting the effectiveness of their practice. We study this in the context of case management in a large foster care organization, where different caseworkers (from different agencies) report on the home visits made to the foster children. We found that unstructured data entry may result in differences in how information is collected across different organizational units in the organization. Institutional theory helps explain how institutional factors shape practices by individuals across organizational units, and how these practices can become stable over time and adopted by other individuals, making the practice persistent.

Chapter 2: Basic Classes in Conceptual Modeling: Theoretical Foundations and Practical Guidelines

It is widely held that a key role of an information system (IS) is to represent the world (Burton-Jones and Grange, 2012; Kent, 1978; Wand and Weber, 1995). This assumption suggests that one of the most important questions in IS development is “*How can we model the world* to better facilitate the development, implementation, use, and maintenance of information systems that provide value?” (Wand and Weber 2002; emphasis added). This makes conceptual modeling, a process by which representations of the world get translated into IS objects, a prominent aspect of IS development and use (Kung and Soelberg, 1986; John Mylopoulos, 1998; Rossi and Siau, 2000; Wand and Weber, 2002).

Conceptual modeling refers to the “activity of formally describing some aspects of the physical and social world around us for the purposes of understanding and communication” (J Mylopoulos, 1992). Conceptual modeling involves documenting knowledge about a domain, defining its scope, and outlining constraints. Once developed, conceptual models typically guide database and application design and often become legally binding documents that contain information specifications of the IS.

Conceptual models depict information about the kinds of objects that an IS needs to represent. Since the 1970s many approaches to representing domains have been suggested, including the Unified Modeling Language (UML) (Grossman et al., 2005; Jacobson et al., 1999), Entity-Relationship (ER) Diagrams (P. P.-S. Chen, 1976), Object-role modeling (ORM) (Halpin, 2007), and i* (Yu, 2001). Each approach maintains the

assumption that the information about the objects is specified and verified by domain experts and future users of the IS (Appan and Browne, 2010; Browne and Ramesh, 2002).

Among other things, the information elicited from users imply knowledge of structures in a domain (Cooke, 1994). Major conceptual modeling grammars, such as UML and ER Diagrams, organize domain objects into classes (e.g., similar to concepts, categories, kinds, or entity types). For example, in communicating with potential users of a university registrar system, analysts could derive and include classes such as *students*, *courses*, and *instructors* into the conceptual model. Notably, some users might prefer different structures (e.g., distinguishing between *faculty* and *instructors*), but ultimately classes reflect a consensus among all involved parties (Parsons, 2002). Once specified, classes constrain the kind of information to be managed by the IS (e.g., information about specific students, courses, and instructors), directly impacting such IS objects (e.g., database tables, data collection fields, user interface options, and reports)(Hirschheim et al., 1995; Teorey et al., 1986).

To elicit classes accurately and reach consensus on which classes to use, it is important to be in frequent communication with users. Maintaining close contact with users is a commonly prescribed guideline in systems development (Moody, 2005; Gould and Lewis, 1985; John Mylopoulos, 1998), whereas “lack of user input” is considered among the “leading reasons for project failures” (Gemino and Wand, 2004, p. 248). This issue is less problematic when an IS is developed and used within organizational boundaries (Fry and Sibley, 1976; Mason, 1978; Zuboff, 1988). Yet, as more IS are developed to support

more diverse uses by customers, suppliers, and members of general public (such as many multi-user online systems), analysts could no longer rely on a stable single group of people for complete specification of domains (P. P. Chen, 2006; Gumm, 2006). Indeed, many online projects (e.g., social media, crowdsourcing) foster open participation to any interested online user, resulting in extremely wide and diverse audiences. In such cases it is becoming nearly infeasible to elicit all possible structures that would be congruent with the domain views of every user (Lukyanenko and Parsons, 2013a).

In response to the growing challenge of modeling when user views are extremely diverse, recent research suggested to abandon conceptual modeling entirely - “no conceptual modeling” - and provide flexible database structures that will accept any user input (Lukyanenko and Parsons, 2013a, 2013b). This input can then be structured after data is collected based on ad hoc needs. This strategy, allows the collection of diverse user information, creates novel challenges such as having the resulting sparse and heterogeneous data useful for analysis. Moreover, it obviates important traditional benefits of conceptual models such as supporting communication, facilitating domain understanding among development teams, and supporting information retrieval and use of data. Although prior research has assessed the efficacy and limitations of conceptual modeling in novel settings, the proposed solutions themselves have their own limitations.

In this paper we propose an alternative approach: rather than eschewing conceptual modeling (and its benefits), we suggest to select few “basic” classes for which user consensus is likely to be high regardless of the diversity of the user-base. This approach is motivated by recent experimental findings in conceptual modeling that show that some

classes (coined “basic-level categories”) result in high accuracy and may therefore be used by most people no matter their level of domain expertise or motivation to contribute content (Lukyanenko et al., 2014). This finding raises the possibility of using such classes in conceptual modeling. Yet, to use such classes, we need to have a better understanding of their nature and have specific guidelines that can support their practical application.

This paper aims to bridge this gap by providing theoretical foundations rooted in psychology for the existence and the importance of these special classes termed in psychology basic level categories (Harper and Schoeman, 2003; Klibanoff and Waxman, 2000; Lassaline et al., 1992; Rosch et al., 1976). Investigating basic level categories led psychologists to propose (and evaluate) a number of criteria that helps in the identification and selection of basic level categories in a domain. In this paper, we formulate principles for identifying basic classes in a domain. Once identified, these classes can guide conceptual modeling, database design, and user interface development in a wide variety of traditional and emergent domains.

Emergent Challenges of Selecting Classes in Conceptual Modeling

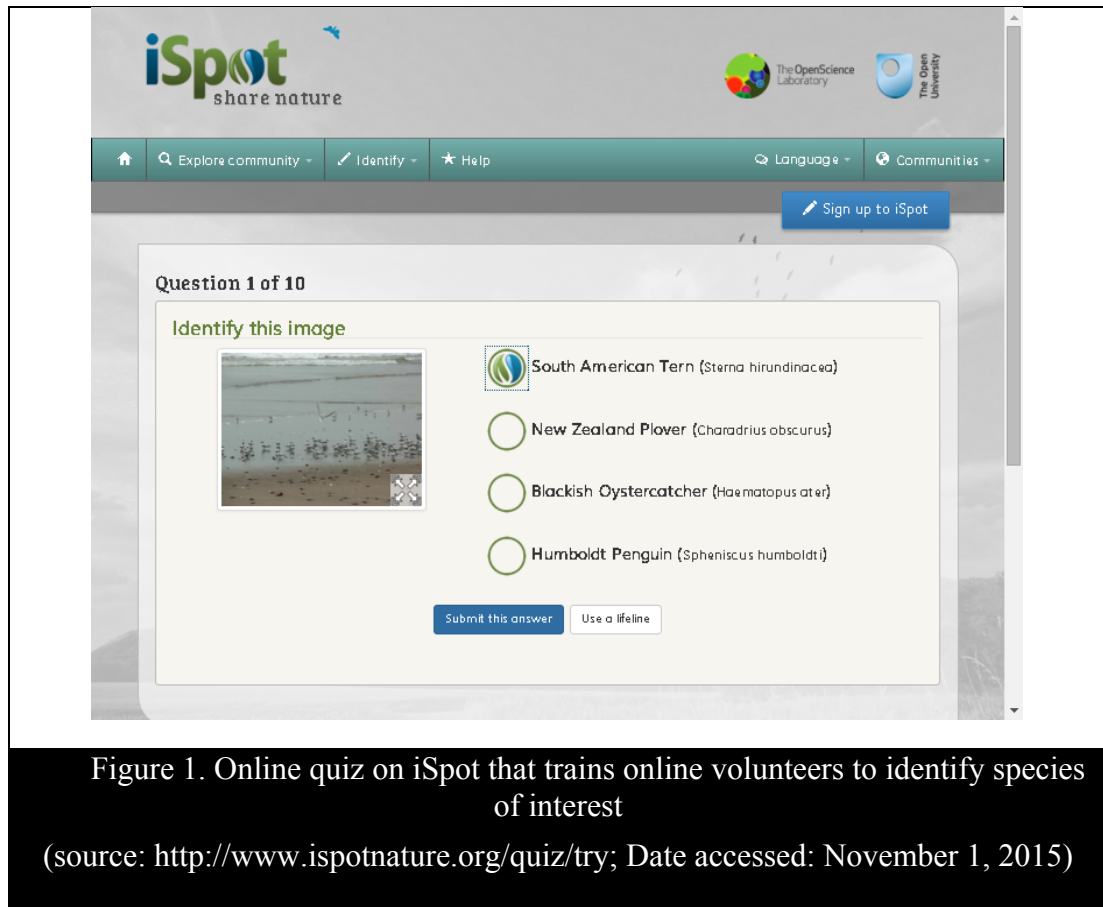
Much of traditional conceptual modeling has been conducted in corporate settings and has hence shaped the grammars and practices employed in conceptual modeling (e.g., how to determine relevant classes for the IS in advance). In this context, information systems users were typically employees or those with close ties to the organization. However, today this is not the only paradigm in which IS exist and for which there is a need for conceptual modeling. We highlight some other paradigms below.

Open information environments online

With the growth of the Internet and mobile computing, organizations increasingly allow users to contribute content. The result is the growth of open information environments (OIEs) in which organizations “have access to sources over which they may have no control; new sources of data may emerge; applications of data might change radically over time; and new uses of data might emerge” (Parsons and Wand, 2014). A prime example of OIE is *user-generated content* (UGC) created by ordinary people online that an organization can access and use in its own decision making in operations (e.g., forum posts, tweets, tags, product reviews, digital artwork, blogs)(Cha et al., 2007; Levina and Arriaga, 2014; Susarla et al., 2012). To harness the power of UGC, organizations are rapidly developing online platforms such as BeingGirl.com by Procter & Gamble, eBird.org by Cornell University, or FixMyStreet.com by the UK organization mySociety. These platforms are completely open, inviting anybody to join and participate. In such projects, the possibility of determining in advance all classes that would reflect the views of every single potential user for that project is impractical. Thus, establishing relevant classes in this context is infeasible and researchers have concluded that conceptual modeling may not be appropriate in such settings (Lukyanenko and Parsons, 2013a). We believe the idea of basic-level categories could be quite effective in these settings.

To better understand conceptual modeling challenges in OIEs and potential applications of basic level categories consider the case of citizen science – a type of UGC and OIE that harnesses contributions of ordinary people for scientific research (Bonney et al., 2014; Rossiter et al., 2015). Citizen science is built on the premise of open participation

(Hand, 2010). As a result, placing any limits on the information users can input is in many ways contrary to the spirit of citizen science. A high-profile example of a citizen science project is iSpot (www.ispotnature.org), run by The Open University in the UK (Clow and Makriyannis, 2011; Scanlon et al., 2014; Silvertown, 2010). The objective of iSpot is to expand scientific knowledge by asking people to observe plants, animals, and other taxa across the globe and report these sightings to their custom online platform. The data collection on iSpot is at the *species level* of classification (e.g., *Spotted sandpiper*, *American robin*, *Atlantic salmon*, *Black bear*). Thus, while participants can report observations at different classification levels, the focal classes for iSpot are classes of species (Crall et al., 2011; Mayden, 2002). This is consistent with the prevailing scientific interest of the project and is similar to other natural history citizen science projects, including the Cornell University's eBird (www.ebird.org), Atlas of Living Australia (<http://www.ala.org.au>), and Canada's GEIODE network (e.g., www.geog.ubc.ca/biodiversity/eflora) (Bonney et al., 2009; Mayden, 2002). Figure 1 shows a sample online quiz on iSpot that trains online volunteers to identify species of interest to the project.



Once the system is developed, the extent to which users are able to navigate its structures, search, and contribute information depends on their ability to interpret and understand the underlying conceptual model (Burton-Jones and Grange, 2012; Lukyanenko et al., 2014). The unique challenge in OIEs, however, is that while the conceptual model may faithfully capture classes that could be suggested by subject-matter experts following biological nomenclature, the model may be unable to fully support the citizen science project it was designed for. The above representation may be incongruent with views of some non-experts, which may be the actual contributors of the information – the citizen scientists. For example, it is possible that some non-expert users may prefer (or be only familiar with) certain classes other than those modeled by the system designer. For

instance, the fact that *polar bears* are *bears* and spend considerable amount of time on land may lead non-experts to conceptualize them as *land mammals*. Similarly, since “many shorebirds are long-distance migrants and can show up far from their normal ranges” (Kaufman, 1999), some users may fail to classify *Spotted sandpipers* as *shorebirds*. Non-expert users may be uncomfortable with *species* at the focal level of classification—even if they are familiar with actual instances belonging to that class. Recent empirical research in citizen science demonstrates that non-experts are generally unfamiliar with more specific scientific classes such as *genus* or *species*, leading to inaccurate classifications when data collection and storage is based on such classes (Lukyanenko et al., 2014). Each misalignment between the chosen conceptual model and the views of the people who are going to use the system, has an impact on data quality and may also preclude users from effectively navigating, searching, and contributing information.

Other emerging domains

While OIEs are an increasingly important setting, to further motivate the research on basic level classes, we suggest three additional scenarios where such classes could support more effective IS development. While the primary motivation of this work is to support conceptual modeling in the context of extreme user view diversity – as demonstrated from the additional scenarios below – the concept of basic level categories can be potentially effective in a wide range of applications.

With the explosive growth of mobile and wearable devices, an increased conceptual modeling challenge is modeling domains in a way that is congruent with affordances and limitations of mobile and miniaturized settings. While traditionally conceptual modeling research aimed at modeling application domains without being concerned with implementation issues, mobile and wearable settings may preclude realization of complete specifications because of the inherent functional and spatial limitations of the devices (e.g., screen size, hardware constraints). Successful mobile devices typically contain few menu options and provide limited (compared with desktop equivalents) data collection support. This suggests that some basic, high level, or intentionally constrained specification may be more appropriate for mobile settings. Similarly, mobile applications tend to take place online with no constraints on who can participate and engage a broader audience. Thus modeling for mobile devices may entail similar challenges to those in open information environments (OIEs).

The term ‘Big Data’ has been defined in a few different ways. One definition suggests that the volume, variety, and velocity of data created and accessible to individuals and organizations are growing at unprecedented levels – and will only increase. There are many sources such as social media outlets where ordinary people are writing the way they see the world. These descriptions of the world are often generalized as a basic notion of a “post” (e.g., Facebook status update, Twitter tweet, blog posts, etc.). The more generic notion of a post also provisions for the inclusion of content that is unpredictable to structure further. Individuals are given space to create and share their conceptualizations with other users. For example, the content of a post can include text,

symbols, numbers, URLs to other webpages, etc. While these data can be parsed through text mining algorithms, posts can also include location data or content such as documents, videos, pictures, or audio that are less structured compared to traditional fields of a database table. While it is important to have a bird's eye view of these domains and select few classes that would summarize the data sources effectively, it becomes challenging to predict how users will engage in these creative outlets in advance. Thus, the basic level class "post" provisions for such variety of data while still providing some mechanism for organization.

In addition to the emergent online contexts, organizations are increasingly opening their internal systems to customers—many of whom may not be sufficiently familiar with the conceptual structures behind such systems. For instance, consider the case of patient-facing applications in healthcare or the proliferation of online health systems (e.g., WebMD or PatientsLikeMe)(Angst et al., 2010). WebMD allows individuals to research conditions, check their symptoms, and access drug and treatment information. Another example is PatientsLikeMe (www.patientslikeme.com), which allows patients to share their own health experiences with other patients with similar conditions. In these customer-facing applications, it may be more effective to have information at a level of abstraction that is congruent to the individual's knowledge. For instance, a physician might be comfortable with the patient's record being organized based on symptoms or conditions (e.g., whether they have acute bacterial rhinosinusitis, acute rhinosinusitis, or chronic rhinosinusitis). For the patient, however, this level of detail may be unfamiliar to them or, even when known, too specific to make the information actionable (e.g., be able

to plan a course of treatment). Such applications may leverage on the notion of basic level categories by identifying information that may become understood by the individual looking at the information.

Consider the case of selecting classes for PatientsLikeMe. Analysts may elicit a list of conditions from physicians (subject matter experts) together with higher-level classes to group these conditions. Alternatively, the list can be sourced from many available medicine ontologies or scientific publications, among others.

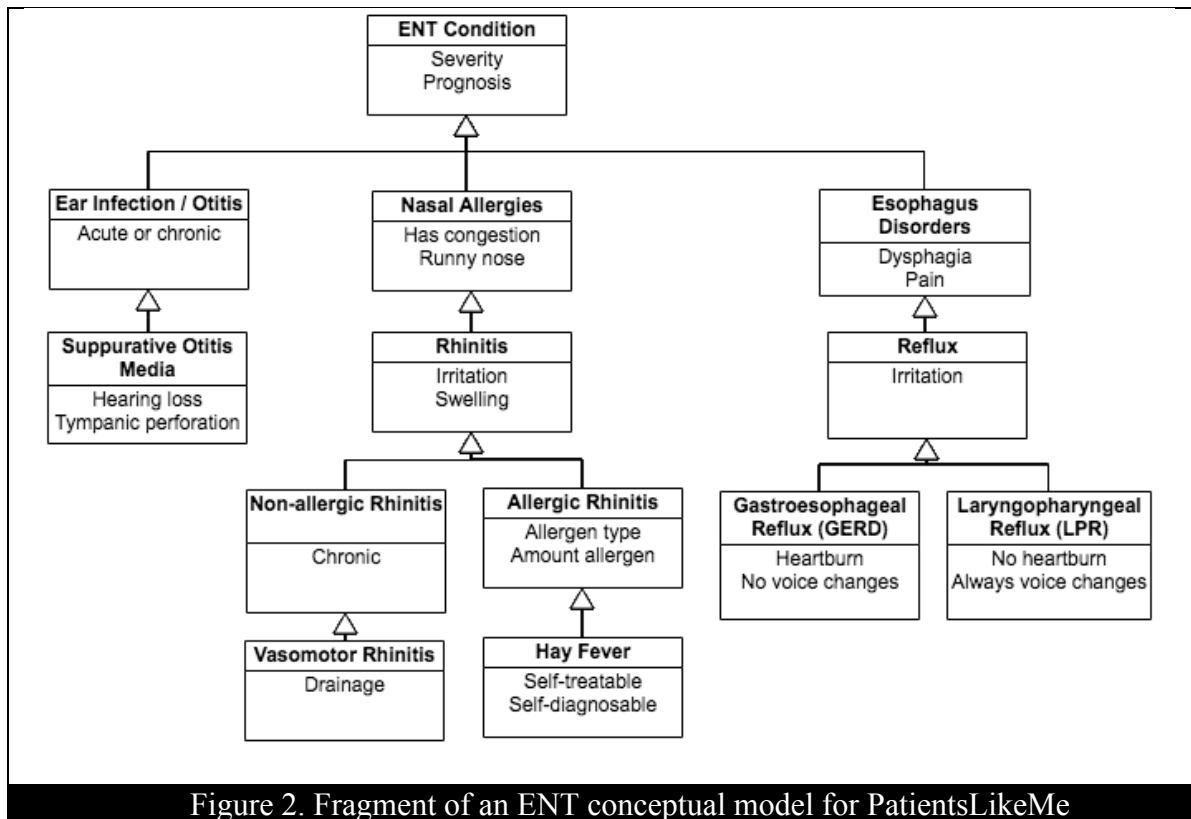


Figure 2. Fragment of an ENT conceptual model for PatientsLikeMe

Based on models similar to that of Figure 2, developers can then create database tables and user interfaces. Once the classes are established, online users can be trained in their ability to identify instances of these classes and report observations accurately.

While the three additional scenarios described above will have challenges with traditional conceptual modeling practices, research in conceptual modeling began to address the challenges of modeling in OIEs and other similar settings. There is growing evidence supporting the premise that in open and highly heterogeneous environments (e.g., citizen science, health forums, and other UGC), reaching an agreement on all valid domain conceptualizations by all potential system users is infeasible (Lukyanenko and Parsons, 2013a). One solution that has been proposed is to skip conceptual modeling entirely and not develop traditional domain representations such as those shown in Figure 2. The information systems development under this “lightweight” or “no conceptual modeling” approach then simply selects a flexible data model (e.g., a schema less no-SQL database), and presents users with an interface where users, in a free-form manner, could suggest any attributes or classes they wish to report (Lukyanenko and Parsons, 2013a, 2013b).

While the no conceptual modeling approach has advantages (e.g., ability to generate more quantity of information, ability to record novel classes and attributes), the proponents of this approach themselves concede that the resulting data is highly heterogeneous and inconsistent (Lukyanenko and Parsons, 2013b; Lukyanenko et al., 2014). For example, one user could describe instances of hay fever as *rhinitis* and another user may describe it as a *nasal allergy* or simply an *allergy*. In the absence of other information, linking these three records becomes problematic, negatively affecting retrieval, aggregation, and analysis of the data resulting from the “no conceptual modeling approach”. Motivated by these limitations, we develop an alternative approach that seeks classes for which the inter-user agreement is maximal. It is based on the

premise that while heterogeneous online audiences may have many disagreements, there could also be a significant number of classes that are universally accepted by almost all potential users. Having identified these classes, developers could then follow traditional phases of the information systems development that rely on conceptual models for the underlying structure of the domain. These classes can inform traditional database development and user interface design and drive the data collection choices. Studies in conceptual modeling that follow theories in psychology suggested the existence of such classes termed in psychology basic-level categories (Lukyanenko et al., 2014; McGinnes, 2011). In the next section we turn to psychology in search for theoretical guidance for the method for identification and application of basic-level categories.

Conceptual Motivation for Guidelines

The notion of different hierarchies of classes is not new to conceptual modeling research and practice. Conceptual grammars used in Information Systems such as the Entity-Relationship (ER) model, the Unified Modeling Language (UML) Class Diagrams, and Object-oriented programming, have a conceptual and philosophical root in theories of classification (e.g., modeling the real world). For example, upper-level classes in Object-oriented programming are defined in terms of shared properties (e.g., inheritance) that are consensus-driven. Yet, there are no widely accepted rules for creating or evaluating collections of classes (Parsons and Wand, 1997). There is no *perfect* design since it is subject to someone's perceived reality (Wand et al., 1999; Taivalsaari, 1996). The level of categorization depends on who is doing the categorizing and on what basis – the categories of objects are defined by properties shared by the objects themselves and the

abstractions of similarity to one or more individuals (Lakoff, 1987; Taivalsaari, 1996). Previous research acknowledges the need for design principles to guide conceptual modeling (Parsons and Wand, 1997) and the potential value of basic-level classes (Lukyanenko et al., 2014). The aim is to derive design principles from psychology to facilitate the identification of these basic level classes.

Information systems researchers have used two main theoretical foundations in understanding conceptual modeling: ontology and cognition. Ontology deals with models of reality. Bunge (1977) ontology has been popular in IS (and conceptual modeling) as it maps well to IS constructs (*things* – individuals or entities) (Wand and Weber, 1990) and predicts information systems phenomena (Gemino and Wand, 2004; Siau and Wang, 2007). Cognitive processes, on the other hand, moderate human understanding of the real world and provide theories of cognition, particularly, theories of classification, to identify fundamental concepts (e.g., classes) that describe an application (Rumbaugh et al., 1991; Lukyanenko et al., 2014). In the development of our guidelines, we complement ontology with cognition since classification is intended to represent human knowledge and thus the importance of cognition in deriving principles to choose classes (Parsons and Wand 1997).

According to cognitive psychology, classes support vital functions of an organism via *cognitive economy* and *inductive inference* (Lakoff, 1987; Roach et al., 1978; E.E. Smith and Medin, 1981). Both functions compete for the same *limited* cognitive resources of human *memory* and *processing power*. Cognitive economy is achieved by maximally abstracting from individual differences among objects and then grouping objects in

classes of larger scope (Fodor, 1998; G.L. Murphy, 2004; E.E. Smith and Medin, 1981). In biology such classes could be *animals* and *plants*. By storing only a few classes, humans can easily memorize identifying characteristics of the different classes. Having only a few classes in the vocabulary maximizes the likelihood that two different people would have the same classes. This promotes communication efficiency and social interaction – an important function of classification in human society (G.L. Murphy, 2004). Cognitive economy becomes further vital as the environment continuously supplies organisms with massive amounts of unique sensory data, thus having fewer classes helps people cope with the changing diversity of the world. Strictly focusing on the benefit of cognitive economy therefore suggests that the best candidates for maximal agreement classes are those classes with the broadest scope – those at the top of the taxonomic tree.

Overemphasizing cognitive economy, however, comes at the expense of ignoring certain individual characteristics of organisms that may be vital for the organism's function and survival. For example, suppose we are interested in a particular property of an object we encounter (e.g., we wish to discern if a rhinitis is *allergic* or *non-allergic*). Classifying a condition as a *rhinitis* (a high-level class) versus *Hay Fever* (a lower-level, particular kind of allergic rhinitis) gives different probabilities of this object having the property of interest. The probability that a *Hay Fever* is due to an *allergic* reaction is substantially higher than the probability that any *rhinitis* is produced as an *allergic* reaction. This example also demonstrates why a domain, such as healthcare, is interested in a finer (specific) level of classification. Knowing that a phenomenon is *Hay Fever* affords

greater inferences and action than knowing it is a rhinitis or nasal congestion. The ability to predict attributes of instances of a class, or the inferential power, increases as the scope of the class decreases. While cognitive economy mainly deals with communication, memory, and processing, inferences are the primary drivers of human behaviour and decisions (Tsui et al., 2010; E. Smith, 1989).

It follows then that to maximize predictive power, humans should prefer classes with narrower scope. Thus while classes with narrower scope are useful in many ways, memorizing, organizing, and communicating these categories require more cognitive effort. The trade-off between these competing functions is considered one of the defining mechanisms of human cognition and behavior (Corter and Gluck, 1992; Roach et al., 1978). Based on the tradeoff between cognitive economy and inferential utility, psychology hypothesized that humans favor (e.g., learn, communicate) those classes that maximally exploit both predictive power of classes and their cognitive economy. Rosch et al. (1976) argued that in the world of “infinite number of discriminately different stimuli” and facing the tradeoff between cognitive economy and inferential power, humans favor classes that are most capable of supporting these competing objectives of classification. Based on converging evidence from anthropology (Berlin et al., 1973; Raven et al., 1971), Rosch et al. (1976) proposed that there is a set of “privileged” classes coined *basic level categories*.

Basic level categories became the subject of active research in psychology and cognitive sciences, generating considerable amount of evidence and making this concept one of the most established propositions in psychology (for reviews, see (G.L. Murphy, 2004;

Lassaline et al., 1992)). Below we review conclusions on basic level categories as a result of forty years of studies in psychology.

First, as follows from the special function of basic-level categories of optimizing the tradeoff between cognitive economy and inferential utility, basic level tends to be a *taxonomic middle*. Concepts that belong to this level tend to reside between the highest and lowest level in a conceptual hierarchy (e.g., “dog” is higher than “collie” and lower than “animal”).

Second, it has been suggested that a basic level category is often an *entry category* – the first concept thought by a user when encountering a phenomenon (Jolicoeur et al., 1984). Gregory L Murphy and Brownell (1985) called it the “necessary first step” of identification (p. 72). Being the entry points, these classes tend to be retrieved extremely quickly and accurately (Lukyanenko et al., 2014; Zhou et al., 2010). While the use of basic level categories, as opposed to more accurate subcategories, may be contingent on one’s expertise (e.g., dog experts may bypass basic and immediately think of a breed), experts readily relate to the basic level (in contrast to lower levels that require familiarity and expertise for use)(Tanaka and Taylor, 1991).

Third, basic-level categories tend to be common words such as *bird, tree, fish, cup, chair, and house* (for more examples, see Table 1). Psychologists further demonstrated that children learn basic categories before superordinate ones (Carolyn B. Mervis et al., 1994) and consequently adults use more frequently in ordinary (non-specialized) day-to-day communication (Lassaline et al., 1992; Rosch et al., 1976) since these categories apply across domains. In addition to categories for nature, researchers have demonstrated basic

level categories for events (Rifkin, 1985), personality types (Cantor and Mischel, 1979), environmental scenes (Tversky and Hemenway, 1983), and for psychiatric diagnostic categories (Cantor et al., 1980).

Table 1. Some examples of basic-level categories from psychology studies	
Basic-level category	Source
Bird, dog	Tanaka and Taylor (1991)
Bear, rhino, pig, seal, bug, cat, turtle, crab, dog, fish,	Waxman and Klibanoff (2000)
Horse, rhino, lizard, pig, hippo, bug, duck, turtle,	Klibanoff and Waxman
Tree, fish, bird	Rosch et al. (1976)
Flower	Mervis et al. (1994)
Dog, duck, cat	Rhemtulla and Hall (2009)
Mouse, fish, butterfly, bird, rabbit, beetle, dolphin,	Op de Beeck and Wagemans
Apple, pear, orange, lime, coconut, pineapple, carrot,	Jolicoeur et al. (1984)
Birds, dogs, fish, other common animals	(Johnson and Mervis, 1997)
Bird, dog	(Macé et al., 2009)
Apple, melon, berry	Wales et al. (1983)
Horse, spider, chicken, fish, dog	Mandler and Bauer (1988)
Cat, dog, horse, bird, bat	Younger and Fearing (2000)
Bush, tree, flower	Murphy and Wisniewski (1989)
Cow, sheep	Zhou et al. (2010)
Bird, dog, flower, fish	Grill-Spector and Kanwisher
Cat, dog, horse, cow, apple, pear, daffodil, sunflower	Bowers and Jones (2008)
Dog, tree	Rorissa (2008)
Bird, flower, tree	Barr and Caplan (1987)

Fourth, compared to other levels, subcategories within basic category are perceived to be most similar to each other (Rhemtulla and Hall, 2009) while two neighboring basic-level categories have many psychologically relevant differences (Markman, 1991). In general, basic level maximizes “both within-category similarity and between-category dissimilarity” (Mandler and Bauer, 1988). Rosch et al. (1976) proposed that basic-level categories have the most defining attributes (e.g., more diagnostic attributes that describe *bird* than those that describe a specific bird).

Fifth, Basic level categories are at a level in the taxonomy at which category members can be visualized. In cognitive science, a category exists whenever two or more distinguishable objects are treated equivalently for some purpose (C. B. Mervis and Rosch, 1981). Categories can be derived as a result of sensory perception, cognitive, conceptual, and emotional processing of objects (Ozcan et al., 2014).

Sixth, basic level categories tend to be short (see Table 1 for examples). Word length is associated to the frequency of its use – things have many equally correct names some of which are more common than others. Objects named with infrequent words take longer to name than objects named with frequent words (Oldfield and Wingfield, 1965).

Seventh, Adults have notions about the kind of language appropriate for use with children (e.g., long names are troublesome for children). Carolyn B Mervis and Crisafi (1982) suggested that children's categorization ability is acquired in the order: basic, superordinate, and subordinate. The options are constrained by the contextual contrasts to be expressed rather than by the linguistic ability of the interlocutor (Wales et al., 1983).

Last, psychology research further explored formal models of basic-level categories. An early model by Rosch et al. (1976) advocated *cue validity*, a sum of the conditional probabilities that an object is in a target class (e.g., fish) given that it possesses a set of attributes (e.g., can swim, has scales). Rosch et al. (1976) argued that since basic-level categories hold the greatest number of attributes, cue validity of such classes would be maximal. This argument was refuted by Murphy (1982), who pointed out that cue validity model lacked constraints (e.g., limited cognitive capacity constraint) and was unbounded. To balance cue validity, another measure, *category validity* was proposed. It

reversed the conditional probability of cue validity and measured the probability of an object having features of interest (e.g., can fly, has wings) given that it is assigned a particular category (e.g., bat).

Combining cue and category validity models appeared to offer a mathematical balance to compensate for lack of binding constraints. The problem, however, is that it is unclear how to combine category and cue validity in such a way that their individual contributions genuinely reflect the importance of these functions to humans. Several heuristic approaches and algorithms, mainly in artificial intelligence, cognitive science, and economics have been proposed. Jones (1983) developed a *collocation model*. In this model, cue and category validity were *multiplied* to produce a concave function with a unique maximum. The collocation measure was argued to be the greatest for basic-level categories. While the collocation model resolved the unboundedness issue of cue and category validity, it still lacked a theoretical rationale for combining the two measures in a particular way (Corter and Gluck, 1992).

Building on the above theories, a model of classification optimality and category utility was proposed by Corter and Gluck (1992; 2012). This model is designed to directly operationalize the tradeoff between cognitive economy and inferential utility in a way that adheres to the widely held propositions about human cognition in psychology. This model has been applied in artificial intelligence and used as part of more complex algorithms (Gennari et al., 1989; Nakamura et al., 1993). The model assumes a class hierarchy (e.g., ENT condition – Rhinitis – Hay Fever, such as in Figure 2). Corter and Gluck (1992; 2012) argue that the usefulness of a class is rooted in the ability to predict

unobservable attributes (inferential utility). Moreover, a class is designed to optimize information processing and transfer (cognitive economy). An optimizing function of *category utility* (CU) was then defined. Corter and Gluck (1992) posit that classes with the highest CU will also be most universal among all humans, since knowing and storing them provides the greatest value. Classes with greater CU therefore can be considered basic. The category utility function for the domain is:

$$\max CU = f(c, F) = P(c) \sum_{k=1}^m [P(f_k|c)^2 - P(f_k)^2]$$

In this formula, some class, c is defined by a set of objects o . Each object is characterized by a finite feature (attribute) set, $F = \{f_1, f_2, \dots, f_m\}$. Consider that with no knowledge about a class membership, f_1 (or a set F) can be predicted using its base-rate probability $P(f_1)$. This probability, in turn, reflects the occurrence of that feature in reality. Such random guess, will be, on average, correct $P(f_1)$ times, leading to the final probability of correct guessing in the absence of a class being the product of the two probabilities, or $P(f_1)^2$. Extending the same rationale to the probability of guessing a feature under the assumption of a class membership the correct guess will be $P(f_1|c_1)^2$. Thus, the difference between $P(f_1)^2$ and $P(f_1|c_1)^2$ denoted the additional benefit gained from the class membership. This difference, however, needs to be weighted by the probability of a class c_1 occurring in the world, since the guess is made under the condition of c_1 identification.

Category utility ranges between 0 (when predicted frequencies are equal to base-rate) and 1 (if the base-rate frequencies are low, while conditional probabilities are high). An

interesting property of CU is its relationship to the communication theory by Shannon and Weaver (Shannon, 1948). CU can be considered as the expected reduction of uncertainty due to communication of category information through some cue. The uncertainty is maximal when no category is present and it is being reduced the more “informative” the category becomes (but balanced by the frequency of the category). The category utility offers opportunities for computational approaches to conceptual modeling and automatic discovery of basic-level categories.

To summarize, classification theory in psychology amasses considerable evidence for the existence of classes that maximize agreement among people with different backgrounds, education, and functional needs. Coined basic level categories, these classes have been shown to carry a multitude of benefits resulting in a significant cognitive bias toward these classes. Furthermore, studies in psychology proposed methods for identification and selection of these classes. In the next section we use and expand upon the conceptual motivations highlighted in this section to develop guidelines for identifying basic classes in conceptual modeling.

Guidelines for Identifying Basic Classes in Conceptual Modeling

A natural application of the theoretical propositions in psychology is to construct a set of design guidelines (see **Table 2**) an analyst (or generically, agent) could follow. In proposing the conceptual modeling guidelines, we first consider relevant evidence in psychology (reviewed in (Lassaline et al., 1992; G.L. Murphy, 2004) and highlighted above) and then derive specific design propositions based on widely-held psychological propositions. We then illustrate the application of each guideline with at least one example.

Table 2. Guidelines for Identifying Basic Classes in Conceptual Modeling	
Guideline Name	Guideline Description
G1: Middle level	Identify classes in a domain in the middle of the conceptual hierarchy.
G2: Entry Category	Elicit entry categories from a sample of potential users for the domain objects of interest.
G3: Frequent Word	Identify the most frequent domain words used in a typical discourse.
G4: Cohesion and Coupling	Find a taxonomic level, for which sibling categories have maximal difference and their respective children have maximal similarity.
G5: Object Visualization	Find the highest category in the taxonomy for which category members can be easily visualized.
G6: Simplest Words	Among the classes in a domain, identify shortest and morphologically simple words.
G7: Original Words	If applicable, identify the first words or concepts learned by children or used by
G8: Cognitive Utility	Identify classes with the greatest CU coefficient.

Guideline 1. Middle level.

Knowledge about objects in the world typically has a hierarchical organization (de Beeck and Wagemans, 2001; Roach et al., 1978). Indeed the conceptual model in Figure 2 depicts classes that are organized in a hierarchy from more abstract (e.g., ENT Condition) to more specific (e.g., Hay Fever). Organizing knowledge hierarchically is important for both cognitive economy and inference. According to psychology, inferences about

properties of abstract objects are less reliable than inferences drawn from more specific objects (e.g., knowing something is a Laryngopharyngeal Reflux suggests there is no heartburn). As discussed before, psychologists contend that humans favor those classes that maximally exploit predictive power of classes and their cognitive economy. In our conceptual model, knowing the condition is rhinitis or reflux allows us to better characterize the condition as opposed to knowing something is a nasal allergy or an esophagus disorder.

One of the most widely accepted propositions about basic level categories is that they tend to be in the middle of taxonomic hierarchies (Lassaline et al., 1992; Rosch et al., 1976). The basic level falls somewhere in the middle of taxonomic hierarchies, regardless of how many levels they contain (Neisser, 1987). Objects at the subordinate (lower than basic) level need higher perceptual processing compared to that of basic-level categorization (Jolicoeur et al., 1984) whereas middle-level categories are learned most quickly or can be named more quickly after they were learned (Corter and Gluck, 1992). Incorporating the notion of basic level being taxonomic middle leads to the following conceptual modeling guideline:

Guideline 1 (G1): Identify classes in a domain in the middle of the conceptual hierarchy.

To apply this guideline, analysts could arrange classes in a domain in a hierarchy (e.g., such as that in Figure 2) and identify classes in the middle. As much human knowledge is organized hierarchically, analysts could also leverage many existing repositories (e.g., research databases, wikis, books) to identify core concepts within a particular domain.

This process can also be automated whereby ontology is being provided as an input to an algorithm that outputs classes in the taxonomic middle. However, it is important to mention one cautionary note when applying this guideline. Psychology research does not offer precise guidance on determining which classes should be selected in the case when the hierarchy is deep (e.g., containing more than 3 levels). It is further unclear how to select the middle class when the number of levels is even. As a general rule that applies broadly to the guidelines presented in this paper, we suggest to consider all eight guidelines together when making the final determination. Indeed, the seminal paper on basic level categories by Rosch et al. (1976) introduces this concept in psychology, proceeded under the assumption of two competing levels (e.g., rose vs. flower, eagle vs. bird) and in the course of a dozen experiments, settled on the level of *bird*, *flower* rather than rose, eagle as basic. Thus, one approach to the practical application of G1 is to select more than one level from the middle of the hierarchy. These levels can then be refined by considering other guidelines.

Domain ontologies are typically represented in a hierarchy that may span both in depth (vertical axis) and breadth (horizontal axis). If the number of classes in the vertical axis of the hierarchy was odd – ENT condition, rhinitis, and hay fever ($n = 3$) – the basic level category would be that of the taxonomic middle, in this example *rhinitis*. If the number of classes in the vertical axis of our taxonomy is even – ENT condition, esophagus disorder, reflux, and gastroesophageal reflux ($n = 4$) – we could argue that both *esophagus disorder* and *reflux* are both in the taxonomic middle. When n is even and greater than four – ENT condition, nasal allergy, rhinitis, non-allergic rhinitis, and vasomotor rhinitis

($n = 5$) – the taxonomic middle, as described above, could be every class between the superordinate (e.g., ENT condition) and the subordinate (e.g., vasomotor rhinitis), which we refer as *inclusive middle* or the classes closest to the hierarchy’s middle (e.g., rhinitis –one class since it is a hierarchy with an odd number of classes), which we refer in this paper as *exclusive middle*. In the ENT example in Figure 1, the tuple of classes at the inclusive middle are: {Ear infection/Otitis; Nasal allergies, rhinitis, non-allergic rhinitis; nasal allergies, rhinitis, allergic rhinitis; esophagus disorders, reflux}. The tuple of classes at the exclusive middle are: {Ear infection/Otitis; rhinitis; esophagus disorder, reflux}. The excluded classes, in both cases, would be the superordinate class {ENT condition} and the subordinate classes {Suppurative otitis media, vasomotor rhinitis, hay fever, gastroesophageal reflux, and laryngopharyngeal reflux}.

Guideline 2. Entry Category

It has been suggested that basic categories often become an *entry category* – the first concept thought by a user when encountering a phenomenon (Jolicoeur et al., 1984). Gregory L Murphy and Brownell (1985) called it the “necessary first step” of identification (p. 72). Being the entry points, these classes tend to be retrieved extremely fast, accurately, and efficiently (Zhou et al., 2010). Naturally, the entry point process is context-sensitive (Tanaka and Taylor 1991). Basic-level is the most abstract level at which people are able to form an integrated perceptual representation of a category. These basic-level concepts are activated faster than subordinate concepts because they are perceptually distinctive. For example, an apple is matched with the name “apple” faster than with “delicious apple” or with “fruit” (Rosch et al., 1976).

Psychology research demonstrates that subjects first categorize objects at the basic level before evaluating membership at other levels via additional perceptual processing (Jolicoeur et al., 1984; Rosch et al., 1976). However, there are some exceptions in which atypical subordinates are differentiated and informative enough that are considered as basic rather than subordinate (Gregory L Murphy and Brownell, 1985). An entry category may be different in situations when a phenomenon is an atypical representative of its basic class (e.g., subordinate penguin of the basic category bird). In this case, humans tend to ignore a general basic category and reason about an object using specialized categories that seem more fitting to an atypical stimuli (e.g., duck, penguin, chicken). This raises the question that there might be multiple basic level categories (e.g., bird, duck; bird; chicken) within the same taxonomic tree. Entry level categories explain the shorter reaction times found at the subordinate level for some atypical members of basic categories (Macé et al., 2009). Incorporating the notion of entry category leads to the following conceptual modeling guideline:

Guideline 2: Elicit entry categories from a sample of potential users for the domain objects of interest.

It should be noted that an entry category may be contingent on domain expertise, general familiarity with objects in a domain, and are also affected by typicality of objects. Studies show that people may use subordinate names more often for typical exemplars (e.g., penguin vs. bird) (Jolicoeur et al., 1984). Research further suggests that experts may categorize things at the subordinate level as fast as they can categorize them at the basic level whereas non-experts use basic level names (Johnson and Mervis, 1997; Tanaka and

Taylor, 1991). Expertise does not have to span an entire domain; it could be narrow in scope (e.g., a single subordinate category). For example, a person who owns a collie and spends a lot of time with the dog could be considered a “collie expert” (Tanaka and Taylor, 1991). Likewise, people are faster at categorizing a boxing glove as a boxing glove than as a glove, even though the latter is the basic category (Gregory L Murphy and Brownell, 1985). Such a person might be aware of the distinguishing features of collies, but know little about other sub-level species of dogs. Similarly, Boster (1986) found that women from Aguarana, who typically are engaged in cultivating manioc, tended to refer to manioc plants with highly specific (species-level) names. Other members who interacted less with manioc named these plants at the basic level (Wales et al., 1983; Brown, 1958). These individual differences of classification can be a function of idiosyncratic life experiences which analysts may not be aware of. Thus, it is important to elicit entry categories from potential users regardless of their perceived basic level status. Other guidelines can be then considered for narrowing the set to those that are entry for most potential users.

Consider the case of selecting classes for an OIE like the one in Figure 2. It may be more effective to have information organized in a way that is aligned to the individual’s knowledge. Knowing a patient has a nose allergy does not give the individual enough information to select an effective course of treatment. At a general level, we know the patient has an abnormal condition (e.g., which may prevent him from performing daily activities). At a specific level, differentiating between the two is critical since non-allergic rhinitis should be treated differently from allergic rhinitis. Yet, some users may

not be familiar with the condition and course of treatment—especially for non-chronic conditions. Patients familiar with a condition will post content at a specific level (e.g., “I never get sick but I suffer from vasomotor rhinitis which means I often can't breathe through my nose”) or less informed patients may post at a more general level (e.g., “I wish I never had this kind of allergy...Go away, rhinitis. Shoo!”). Patients not aware of the condition may refer to the symptoms in an attempt to assess their medical condition (e.g., “Googling the symptoms I am experiencing for several weeks now I suspect I suffer from allergic rhinitis - I can't breathe through my nose!”). Guideline 2 provides a mechanism to elicit relevant classes from users, including non-experts users. We illustrate the need for model inclusion with two examples:

Example 1: Modeling a symptom checker

The conceptual structure of a symptom checker project such as the WebMD symptom checker (symptoms.webmd.com) allows the patient to input their symptoms to learn about plausible conditions and next steps. These models can be sourced from available medical ontologies, scientific publications, or subject experts. The structure used in WebMD symptom checker is one that requires the input of symptoms by the user. In developing such system, we argue that users with varying knowledge can inform the structure of the information systems developed. For instance, following the hierarchy on Figure 2, if a non-expert user reflects a very specific symptom such as lacrimation, the probability of accurately inferring the patient has Hay Fever is higher than if the user had input the symptom fatigue, which would yield a higher number of potential conditions. Thus, we can complement existing models by eliciting potential categories from non-

expert users. In Figure 3 a user is asked to input potential classes for specific instances they observe of a patient with Hay Fever. For instance, there could be some atypical subordinate categories (e.g., lacrimation) that may be familiar to users based on their personal experiences (e.g., chronic conditions or past conditions). Other users may state broader categories such as having a runny nose or having red eyes.

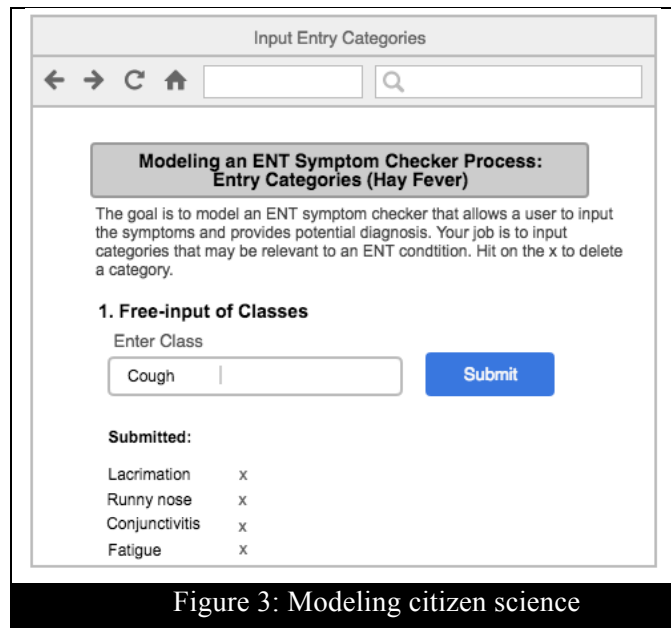


Figure 3: Modeling citizen science

A second illustration we use is the example of modeling the graduate business school admissions process adapted from (Saaty and Vargas, 2012). The goal of the admission process is to select the *best* candidates from a prospective pool of applicants by analyzing their qualifications and fit to the program.

Example 2: Modeling the graduate business school admission process

The selection process involves choice and logical decisions. Saaty and Vargas (2012) use the Analytic Hierarchy Process (Saaty, 1988) to mathematically model the relevant

characteristics of a candidate. Ideally we need “as much information about the candidate”. For instance, information about their learning style (e.g., active, visual, sequential), use of technology (e.g., computer literacy), self-efficacy, reasons for education, academic literacy, Intellectual Quotient (IQ), among many other individual traits that completely characterizes the individual. In reality, however, an admission committee focuses on a limited set of *basic* categories (e.g., scores, years of work experience) that characterize the student. These characteristics could be elicited from interested parts (e.g., admission committee members, faculty, students) through different methods (e.g., interviews, questionnaires, brainstorming sessions, use cases, or role-playing, among others). For example, in designing the system, we could gather a pool of potential users to elicit potential classes (See Figure 4). From the inputs introduced in the UI in Figure 4, someone familiar with the admission process and comprehensive adaptive exams (CATs) may refer to the potential class GMAT (an atypical subordinate) as opposed to score (a basic level) or they may refer to GPA as opposed to grade. The result is a comprehensive list of potential classes – both general and specific.

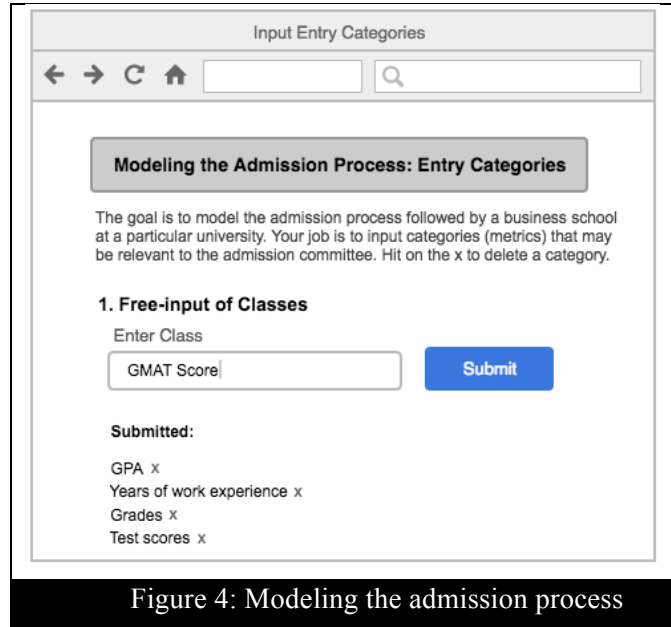


Figure 4: Modeling the admission process

Guideline 3. Frequent Words

Basic level categories are words that occur most often in ordinary daily discourse. Zipf (1935) stated that the length of a word is inversely related to its frequency (e.g., there is a small number of words that occurs frequently, while most words occur infrequently). Folk taxonomists have demonstrated an indexical relationship between the length of a name and the rank of that name in the hierarchical nomenclature system (Brown, 1958), since objects named with infrequent words take longer to name than objects named with frequent words (Oldfield and Wingfield, 1965). As categories become more differentiated, they become more basic. This idea leads to the third guideline:

Guideline 3: Identify the most frequent domain words used in a typical discourse.

The list of potential classes in guideline 2 depends on the sample of users the analyst is eliciting information from. Guideline 3 provides a mechanism to expand the candidate

classes. The goal is to extract data from different sources and to identify common words. For a citizen science scenario or a healthcare scenario, the analyst could parse information from scientific publications, biology ontologies, or user generated content, to identify common words that can suggest potential classes. For instance, the Catalogue of Life (www.catalogoflife.com), a comprehensive index of species containing information on names and relationships of over 1.6 million species. In this catalogue each instance has a taxonomic hierarchy that contains information that range from most abstract (e.g., kingdom), middle (e.g., class, order, family), to the most specific (e.g., genus, species, and subspecies). In the healthcare scenario, the analyst can leverage on existing ENT ontologies such as the Ear, Nose, and Throat Findings Ontology (<http://purl.bioontology.org/ontology/AIR/U000041>) or the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (Stearns et al., 2001) to retrieve potential classes that relate to the domain of interest.

Following the hierarchy in Figure 2, for each of the instances in the ENT domain, we could automate the process by parsing exemplars in our domain of interest and extract concepts at any of the levels in the taxonomy. Next we filter out the most common terms (e.g., based on frequency or any other established metric) and ask potential users – both experts and non-experts – to identify basic classes. A simple framework to increase the potential terms to our basic taxonomy is to collect reports and documents in the field of interest (e.g., parse from domain ontologies, UGC mediums) and plot the frequency of each of the terms, keeping the most relevant ones (e.g., top N).

Guideline 4.Cohesion and Coupling

The basic-level effect arise because exemplars of categories are quite similar to one another and quite distinct from the exemplars of other categories – by knowing that a canary is a bird, we can generalize to items with similar characteristics (e.g., other kind of birds) but not with items that are dissimilar (e.g., other kind of animals) (Rogers and Patterson, 2007). Basic level categories are the most differentiated (Gregory L Murphy and Brownell, 1985) and can be seen as a compromise between the accuracy of classification at a maximally general level and the predictive power of a maximally specific level (G.L. Murphy, 2004). People are able to list more attributes for different objects belonging to the same basic level concept than for objects belonging to more abstract concepts (Rosch et al., 1976). Individual dogs are all represented with quite similar patterns, whereas other kinds of animals (e.g., pigs, goats, birds, etc.) are represented with somewhat different patterns, and non-animals are represented with dramatically different patterns. In other words, basic-level categories correspond to relatively tight and widely separated clusters of distributed representations in the network of categories – they are both distinctive and informative (Rogers and Patterson, 2007).

Basic level categories are in general more distinctive than subordinate categories. Subordinate level categories are more specific but include only small sets of members (Schmid, 2007; Rosch et al., 1976; Lakoff and Johnson, 2008). Members of a basic category tend to resemble each other – and do not resemble members of neighboring basic categories from the same superordinate, maximizing both the within-category similarity and between-category dissimilarity (Macé et al., 2009). Rifkin (1985) found

evidence that the basic level would be the most inclusive level in event taxonomies at which clusters of features are attributed to categories.

Subordinate concepts correspond to smaller and less well-separated clusters within the basic-level cluster and have many near neighbors from different subordinate groups – they are informative but not distinctive. Superordinate concepts correspond to more inclusive but sparser clusters – they are distinctive but not as informative (Rogers and Patterson, 2007). Category membership has a degree (gradient) of membership rather than a binary membership (member/non-member) and objects that are highly typical of a category have a high degree of membership in the category as opposed to less typical objects – lower degree of membership or no membership if it is a completely unrelated object (McCloskey and Glucksberg, 1978).

Guideline 4: Find a taxonomic level, for which sibling categories have maximal difference and their respective children have maximal similarity.

The amount of terms will depend on the scope of the domain we are trying to model. Notwithstanding, the next step is to understand the relationship among concepts in our pool of potential categories. We can calculate the total possible combinations of r objects from a set of n objects $C(n,r)$. In our ENT example in Figure 2 we have 13 classes in total, thus, a total of 120 different combinations (see Formula 1).

$$C(n,r) = \frac{n!}{r!(n-r)!} \Rightarrow C(13,2) = \frac{13!}{2!11!} = \frac{13 * 12}{2} = 78 \text{ different pairs}$$

Formula 1: Combination of pairs from a set of 16 objects

For illustration purposes, we are going to use only three terms (objects): <hay fever, reflux, ENT condition>. The total number of possible combinations of 2 terms from a set of 3 terms $C(3,2)$ is 3 (See **Table 3**). We then classify whether two terms are related or not. This task can be automated and validated by multiple users (e.g., crowdsourcing task, or by the internal team). For example, the outcome for pair <term 1, term 2> could be: a) *term 1* is associated to *term 2*; b) *term 1* is not associated to *term 2*; c) *term 1* is similar to *term 2* (e.g., eagle vs. bird, boxing glove vs. glove).

Table 3. Combinations and relation between the different term pairs		
$C(n,r) = \frac{n!}{r!(n-r)!} = \frac{3!}{2!1!} = 3$	Pair	Relation
	<hay fever, reflux>	0: hay fever is not reflux
	<hay fever, ENT condition>	1: hay fever is an ENT condition
	<reflux, ENT condition>	1: reflux is an ENT condition

The binary classification allows the designer to create a diagram with the mappings between concepts. The relevance of each class can be assessed by the number of connections a class has to other classes (e.g., ENT condition is connected to reflux and hay fever). The next guideline attempts to account for the multi-level aspect of related concepts by finding the highest category in the taxonomy (e.g., reflux is an ENT condition but not all ENT conditions are reflux— thus ENT condition is above reflux in the hierarchy).

Guideline 5: Object Visualization

A concept is a mental representation of an object or a class of similar objects. Concepts can also represent abstract notions, which are implicitly experienced (e.g., adventure) or emotions (e.g., love) (Gregory L Murphy, 1996; Lakoff and Johnson, 2008). Categories can occur as a result of sensory perception, cognitive, conceptual, and emotional

processing of objects (Ozcan et al., 2014). Basic level categories are the most inclusive categories that allow for the construal of a visual gestalt (e.g., an organized whole that is perceived as greater than the sum of its parts) image of a category schema compatible with most category members. For example, the outer shapes of most members of the category *dog* are so similar that it is possible to imagine a picture of a dog “as such”. This is clearly impossible for superordinate categories because their members’ outer shapes are divergent. Basic level categories are those categories for which this informativeness and facilitation of feature prediction is maximal – compared with superordinate and subordinate categories.

For instance, a visual stimulus such as a shore birds first activates the bird node, providing rapid access to the name bird and other typical bird properties (e.g., has wings and can fly) (Rogers and Patterson, 2007).

Guideline 5: Find the highest category in the taxonomy for which category members can be easily visualized.

Following the citizen science taxonomy from Figure 1, an analyst may list the categories at the bottom of the hierarchy and ask users to identify a single visual object that represents that category. For instance, in the ENT hierarchy in Figure 2, the classes at the bottom of the hierarchy would be suppurative otitis media, vasomotor rhinitis, hay fever, gastroesophageal reflux, and laryngopharyngeal reflux. The task for the user is to identify a visual object at the most abstract level but that is still attributable to that class. For example, for a gastroesophageal reflux, the highest category the user may think of is himself or someone else having a reflux. If the user goes to a more abstract category (e.g.,

ENT condition) it is difficult to derive the class reflex because an ENT condition could also refer to an otitis or a nasal allergy– both different from the instance reflex. On the other hand, an expert user may visualize a rhinitis as the highest category for hay fever. Once we have the list of visual objects, the designer could then validate whether these visual objects can be considered basic categories.

Guideline 6: Simplest Words

Things have many equally correct names, but some of these names are more common than others. Typically, things are first named so as to categorize them in a useful way (e.g., spoon rather than silverware) but these categorizations may change over time (and context). Nonetheless, shorter names tend to be the most frequently used names for a thing. Zipf's law predicts that the basic taxonomic level, because of its frequent use, will be labeled with shorter, morphologically simpler terms than superordinate and subordinate levels (Craig, 1986). In other words, word length is primarily determined by frequency of use.

Psychologists have shown that human memory is both flexible and extendable, provided the information is structured. Lexical development is characterized with an increasing morphological complexity. Basic names tend to be shorter termed primary lexemes (Brown, 1958; Rosch et al., 1976) whereas subordinate terms tend to be secondary lexemes that are formed from the basic level term and a modifier (Berlin et al., 1973). Objects named with infrequent words take longer to name than objects named with frequent words (Oldfield and Wingfield, 1965).

Adults have notions about the kind of language appropriate for use with children. The sequence in which words are acquired is set by adults rather than by children and may be based on utility. Children have trouble pronouncing long names and so should always be given the shortest possible names. A word is preferable to a phrase and a monosyllable is better than a polysyllable – this predicts the preference for dog over boxer or animate being.

Sometimes the frequency-brevity principle makes the wrong prediction. For instance, a pineapple is called a pineapple and not a fruit, which is the shorter and more frequent term. Similarly, they will say apple, banana, orange – rather than fruit (Brown, 1958). Brown (1958) argues in favor of referent-name counts (local frequencies), which may be unique for some, while general for others. The best generalization seems to be that each thing is first given its most common name.

Guideline 6: Among the classes in a domain, identify shortest and morphologically simple words.

The probabilistic reduction technique posits that words that are more commonly used tend to be shorter. Short words are used to make communication more efficient – because of pressure for communication efficiency. Short words tend to be predictable, and, on average, convey relatively little information (Piantadosi et al., 2011). For example, we may refer to a vasomotor rhinitis as a rhinitis or refer to a gastroesophageal reflux as simply a reflux.

Guideline 7: Original Words

The formal models of classification proposed in psychology can also inform identification of basic classes. The most complete and comprehensive model is that of (Corter and Gluck, 1992).

Adults have notions about the kind of language appropriate for use with children (e.g., long names are troublesome for children). The most common name is at the level of usual utility but adults do not necessarily provide a child with the name that is at the level of usual utility in the adult world (e.g., a child would refer to a coin as a coin rather than a dime since children do not necessarily focus on the monetary value of the coin) (Brown, 1958).

Objects tend to be named first at a generic level that is perceptually primary (Berlin, 2014). The naming practices of adults determine the child's early vocabulary (Brown, 1958). Mothers use more frequent and more general terms for their children (Wales et al., 1983). The names used to refer to categories at this level tend to be brief. Considerable agreement exists across time, languages, and children in the first words children acquire (Clark, 1979). The options are constrained by the contextual contrasts to be expressed, than by the linguistic ability of the interlocutor (Wales et al., 1983). For example, when naming the same object for a child and an adult, adults will sometimes provide the child with a different name than the name they use with the adult (Anglin, 1977). Carolyn B Mervis and Crisafi (1982) suggested that children's categorization ability is acquired in the order basic, superordinate, and subordinate.

Guideline 7: If applicable, identify the first words or concepts learned by children or used by mothers to talk to children.

Following the ENT taxonomy in Figure 2, we could parse the content from medical books and store the terms in these documents. We then remove words that do not add value to the analysis (e.g., a, an, and, be, at, among others) and perform statistical analysis (e.g., term frequency-inverse document frequency, latent semantic analysis) to identify common words and or concepts and build a dictionary of common words used in medical books (e.g., ENT specialty). These common words are then added to the pool of potential basic categories. In the healthcare ENT taxonomy, we could use the Otolaryngology Head and Neck Surgery: Clinical Reference Guide (Pasha and Golub, 2013), which covers rhinology and paranasal sinuses (e.g., allergy, rhinitis, immunology), endocrinology (e.g., thyroid, parathyroids), among others. For the citizen science, for example, the Kingfisher First Encyclopedia of Animals covers mammals (e.g., lion, elephant, wolf, bear, polar bear, walrus, etc.), reptiles (e.g., lizard, rattlesnake), birds (e.g., eagle, vulture, parrot, gull, and penguin), fish (e.g., fish, goldfish, salmon, seahorse, shark), or invertebrates (e.g., worm, spider, crab, fly, bee, wasp), among others. As we see from these examples, some of these exemplars are subordinates (e.g., parathyroid is a subordinate of thyroid, rattlesnake is a subordinate of snakes, or goldfish is a subordinate of fish), some of these overarching categories are those that are identified as *basic* (e.g., allergy, bird, fish), and some are more abstract (e.g., immunology, mammals).

Guideline 8: CU Coefficient

The perceived world is not an unstructured total set of equiprobable co-occurring attributes. Rather, the material objects of the world are perceived to possess high correlational structure (e.g., wings co-occur with feathers more than with fur).

Categories group together non-identical elements, which, by virtue of their common membership, can be treated as equivalent (Gregory L Murphy and Brownell, 1985). The main benefit of categories is to aid in prediction of feature values (J. R. Anderson and Matessa, 2014). A category is useful to the extent that it can be expected to improve the ability of a person to accurately predict features of instances of that category. Category utility provides a quantitative measure of the goodness of a category for summarizing and transmitting information (Corter and Gluck, 1992). The best categories are those that maximize feature predictability and optimize information transfer (Corter and Gluck, 1992). C. B. Mervis and Rosch (1981) found that basic level categories are those that carry the most information about attributes.

The main function of semantic memory is to support inferences about the unobserved properties of objects and events from partial information (J. R. Anderson, 1991).

Guideline 8: Identify classes with the greatest CU coefficient.

To demonstrate application of the CU function, consider the iSpot example in Figure 1 and a hierarchy animal-bird-osprey. Assume the corresponding hypothetical feature probabilities given in Table 4.

Table 4. Feature probabilities to illustrate Corter and Gluck model				
	Base-rate	P(k animal)	P(k bird)	P(k osprey)
motile	0.9	1	1	1
can fly	0.4	0.5	0.95	1
eats fish	0.006	0.007	0.01	0.9

Computing these probabilities for each category gives CU measure shown in Table 5.

Table 5. Category probabilities and CU measures to illustrate Corter and Gluck model			
Class	animal	bird	osprey
Probability of category, $P(c)$	0.9	0.33	0.005
CU measure*	0.25	0.31	0.01

Based on these calculations, **bird** has the greatest CU coefficient. According to Corter and Gluck (1992), this result is explained by the relative balance between the frequency of the class *bird* and its predictive power relative to other classes.

Summary of Guidelines

The guidelines are not mutually exclusive and can be applied sequentially – the output of one guideline is the input for the next guideline. Some categories may overlap across different principles (e.g., a particular class can be in the middle of a taxonomy for a particular domain but can also be a word that was elicited from users). For example, G3 provides a list of frequent words for a particular domain (e.g., animal, dog, cat, collie, snowshoe siamese). G1 represents a subset of keywords that are in the middle of the hierarchy (e.g., dog, cat). There is a significant overlap between G1 and G3 since G1 is a subset of G3. There may be words that are not frequent yet represent atypical subordinates that can be considered basic (e.g., bulldog). G2 represents the categories that

were classified by users (experts and non-experts). G8 represents the classes with the greatest category utility. Many of the words elicited from users will fall under the basic category (hence the overlap of G2 with G1 and G8). Some guidelines depend on the existence of a prior categorization (e.g., to identify classes with the greatest CU we need a subset of potential classes on which to perform the category utility calculation). The analyst may evaluate the final pool by having a group of users (e.g., domain experts, regular users, designers) rank these classes (e.g., based on a pre-established criteria) and select the best candidates. Alternatively, the analyst may leverage the overlap and retain only the classes that are identified by most or all guidelines. Each strategy would be contingent on situational demands of the project and available resources.

Once these guidelines are followed, analysts should generate a list of candidate basic classes. It is entirely possible that some guidelines may be more applicable than others (e.g., analysts may not have the knowledge of the first words used by children relevant to the domain of interest). Rather than seeing these guidelines as necessary and sufficient, we suggest considering them as cumulative evidence in support of a hypothesis for a particular class. This is consistent with psychology, as psychologists widely recognize that no single guideline is necessary or sufficient for the definitive identification of basic-level categories (Lassaline et al., 1992). Thus, analysts are encouraged to consider the totality of evidence when making the determination.

There are many potential ways these guidelines can be applied in the context of IS development. For example, these classes can be the only classes used in the information system if the objective is to capture the objects in the domain in terms of basic classes.

Earlier we discussed several scenarios where such strategy can be effective. Specifically, when dealing with heterogeneous information sources resorting to basic classes may be a reasonable strategy. At the same time, we note that other implementation alternatives may be pursued. For example, a system can be designed following traditional approaches to conceptual modeling premised on the elicitation of all classes provided by the users. Once these classes are elicited the analyst can apply the guidelines and identify those classes that are basic. This knowledge can then guide user interface design and the functionality of the system. For example, when building a multiuser system to support healthcare applications (where both doctors and patients are expected to use the same system) the knowledge of basic classes can be instrumental in personalizing user experiences to different user groups (e.g., structures that are patient-facing can be based primarily on basic classes whereas doctor-facing interface can use a wider gamut of classes). We hope the guidelines proposed in this work can be used in these and other fruitful ways to make information systems more effective at accomplishing their objectives.

Implications, Contributions, and Conclusions

Traditionally, conceptual modeling research has relied extensively on users for the identification and selection of classes in a domain. However, in an increasingly expanding range of applications, this practice becomes problematic. For example when modeling systems to capture user-generated content, analysts may no longer rely on the ability to reach all relevant users. Even if each user is reached, these users may not be subject matter experts and their requirements may not be as accurate and reliable as in

traditional settings. In online settings, user views may be extremely diverse – further complicating the ability to achieve consensus and generate a common unified view of the domain. In each case, traditional approaches to conceptual modeling may be limited. This paper contributes to the theory and practice of conceptual modeling and development of emerging information systems by proposing a novel approach to conceptual modeling based on the notion of basic-level categories, a widely researched topic in psychology.

The paper contributes to theory of conceptual modeling by surveying theoretical foundations in psychology. The review of psychology provides strong motivation for the importance of special kinds of classes referred to as basic level categories. Following psychology research, we believe the special classes are those for which agreement among heterogeneous online users is the highest. In particular, whereas specialized classes require specialized training and familiarity, which may be absent for some user groups, basic level categories are equally familiar to subject matter experts and non-experts alike. This important property of basic level categories makes them applicable to modeling heterogeneous online contexts. Indeed, recent research in conceptual modeling has already benefited from the concept of basic level categories to operationalize a condition in an experimental study (Lukyanenko et al., 2014). This paper contributes by providing strong theoretical justification for the importance and utility of basic level categories in conceptual modeling research.

Having identified basic level categories as a potentially useful construct in conceptual modeling, this paper proposes guidelines for identifying basic level categories in a domain. These guidelines are derived from well-established propositions in psychology

research that were corroborated in numerous empirical studies. These guidelines provide concrete practical procedures analysts could follow when performing conceptual modeling. It is notable that the guidelines we proposed in this paper can be automated enabling discovery of basic level categories in big data sets. To further increase practical utility of this research, we illustrated the application of each guideline using examples, in addition, as there can be substantial procedural ambiguity when applying a theoretical design guideline in practice (Dreyfus, 1992; Gregor and Jones, 2007; Lukyanenko and Parsons, 2013c), we discussed potential pitfalls in implementation by referencing the relevant work in psychology. Taken together, we believe the proposed guidelines constitute an important novel addition to the conceptual and practical toolbox in IS development.

An important theoretical implication of the notion of basic level categories is a novel opportunity to use the properties of this classification level in explaining experimental findings in conceptual modeling research. Experimental work in conceptual modeling often involves giving analysts and users a conceptual modeling script that represents a domain and then asking questions about the domain based on the script (Bodart et al., 2001; Burton-Jones and Meso, 2008; Burton-Jones et al., 2009; Gemino and Wand, 2003; Parsons and Cole, 2005). While such script can be constructed using meaningless words (Parsons, 2011), often the scripts contain meaningful concepts that vary in their level of familiarity. Some of these concepts could be deemed basic level categories. The presence of basic level categories in such scripts can potentially confound experimental findings, as people might be attracted to those levels and leverage their familiarity with these levels

in answering these questions. We are not aware of any work so far that considers the potential confounding effects due to the presence of basic level categories in the scripts. We hope, with this work, to raise more attention to the important properties of basic level categories in knowledge representation that can be used to better explain experimental findings. We hope to have provided an increased understanding of the role of basic level categories and that this knowledge can be leveraged in future experimental research in conceptual modeling – when constructing experimental stimuli and measures.

Conceptual modeling research generally does not distinguish classes within the taxonomy (e.g., assumes all classes may be equally relevant), yet not all classification levels are equally salient for different people. We show that some classes in a domain have particularly interesting properties. An intriguing theoretical consequence of the basic class concept is the idea of an *information gradient*. The salience of basic level categories for people suggests that classes in a domain can be arranged in the order of their category utility, salience, and familiarity, rather than taxonomically. For example, using the category utility criteria used in the example in Guideline 8, the hierarchy can be arranged in the descending order of the category utility, which would result in the sequence of *bird, animal, osprey*. We call such arrangement of classes an information gradient to contrast it with the traditional generalization and specialization hierarchy that is based on property inheritance.

The gradient concept can be used as an alternative to hierarchical representations of knowledge that are based on category utility, category salience, and other functions. We hope future research will build on the intriguing possibilities implied by the special status of basic level categories and expand the notion of the information gradient.

Another intriguing possibility is whether the concept of basic level categories can become a modeling construct. For example, identifying a class in a conceptual modeling script as a basic level category can send important signals for other stages of IS development and inform database and interface design. Thus having a list of basic level categories can suggest navigational structures and high-level menu items. As conceptual models are widely used to develop other IS objects, the question becomes whether it is advantageous to identify basic level categories inside of conceptual modeling scripts. We believe this possibility should be explored in future studies.

While the discussions in this paper focused on the conceptual modeling phase of the information systems development, the concept of basic level categories carries important implications for other aspects of information systems development and use. This involves selecting navigational structures in the project, presenting choices to users, particularly in mobile settings where there could be space constraints. The concept of basic level categories can also be helpful for information retrieval and query processing. For example, if a non-expert user is trying to learn about Hay Fever or any low-level category within a taxonomy, upper level categories and sibling categories would be helpful for him to make a decision on what to query next to achieve a certain goal. In addition, by adding contextual information to the query, the system can determine the right level of

abstraction at which to present the information to the user. J. R. Anderson (1990) has referred to this as the development of cognitive schemas that individuals are using to create a hierarchy of their knowledge. We hope future research will benefit from the survey of psychology research provided in this paper and the proposed guidelines by applying the arguments and procedures proposed here to address problems in other domains.

While in this paper we painted a positive role of basic level categories in knowledge representation, it is important to also acknowledge the potential negative consequences of dealing with basic level categories. Psychologists have argued that basic level constitutes an important psychological bias. Due to the privileged status of basic level categories people may prefer to use this level at the expense of other levels. As we argued, in many applications this is a desirable outcome. However one should also be cautious and recognize that this behavior can also be detrimental under certain circumstances. For example, following the theories discussed above, we can predict that if a non-expert user is given a choice of different classification levels the user will tend to prefer working with the basic level (e.g., navigating structures based on this level, providing information, querying the information base, acting upon information). If it is more important that the user attends to other levels, inclusion and the availability of basic level categories may preclude users from considering these other levels. We hope future researchers will begin to consider negative applications of including basic level categories as well and propose strategies for mitigating them.

Chapter 3: Identifying Organizational Style: An Institutional Theory Perspective

In the course of normal business, organizations generate electronic documentation describing daily operations and transactions. The purpose of this documentation is generally tactical. For example, IT help desk staff documents reported technical issues, a police officer enters the details of an incident, or a clinician documents a case in progress notes. What is similar about all these examples is that each report is unique, but all reports within an organization are guided by the organizational objectives. Effectively, these data represent the daily transactions of the organization's daily business activities. In many cases, the data collected is used for purely tactical purposes. How does the IT staff resolve the issued ticket? How does the court system resolve a traffic violation? How does a clinician decide when to discharge a patient?

Information systems (IS) should be able to faithfully represent the world they are trying to model—by observing the behavior of an information system, we obviate the need to observe the behavior of the real-world system it represents) (Weber, 2003, 1997; Burton-Jones and Grange, 2012). IS provide the structure necessary to support the organization's business needs' and allow the organization to conduct their daily operations (S. March et al., 2000). Planning and successful decision-making requires processing and analyzing the data assets of the organization. These data reside in different forms, depending on the system design and range from unstructured data to structured data that lives in a relational database (Abiteboul, 1997; Skoutas and Simitsis, 2007). The information required to solve a task at hand can be encoded in free-text whereby a user reviews the data and takes action.

Information is generated at a faster pace than individuals and organizations can make sense of it (Lerch and Harter, 2001). The challenge is not collecting and storing more information, but utilizing the data for better decision-making. Organizations need to cope with limited resources to analyze available data –both structured and unstructured. One of the challenges in doing so, particularly with unstructured data, is the inherent flexibility on how these data are entered/captured in an information system (e.g., free-form text, selecting from drop-down lists, templates). Users may deviate from the deep structure (“the meaning”) of the system by capturing different information in a field that was not originally intended for (Boudreau and Robey, 2005; Wand and Weber, 1995). For example, in a study of an electronic patient record for hypertensive patients, Berg and Goorman (1999) found that although physicians were able to successfully enter coded complaints, diagnosis, blood pressure results, and medication, many physicians complained that the system was too “rigid” to capture the core reason of the patient’s visit. To overcome this limitation physicians started to use a text field labeled as *conclusion* to enter such information and regarded it as a central field for subsequent patient’s visits (Berg and Goorman, 1999; Berg, 2001).

Traditional IS development assumes a fixed schema that can be defined a priori to introducing any data needed to support the business needs’ (e.g., screens that allow for coded input of data that adheres to a regular schema, facilitating the extraction from a database and analysis)(Ramakrishnan and Gehrke, 2000; Shneiderman, 1996). An instantiation of this are relational databases, which have been used in banking, insurance, enterprise resource planning (ERP), finance, and healthcare among other fields. The fixed

schema format of relational databases may be less practical when designing systems that need to support changing needs', require the aggregation of data, or are structured in ways that are unknown to the existing schema (Abiteboul, 1997; Parsons and Wand, 2014). No question IS can facilitate and support work routines, but as seen in the previous example, it may also constrain the workflow of individuals using the system. Nevertheless, organizations should be able to analyze data in the aggregate to enable effective decision-making.

Structured information has the advantage of consistency (e.g., the form in which the data is stored has been modeled in advance), facilitating analysis, aggregation, and integration with other systems (Lukyanenko, 2014; John Mylopoulos, 1998; Fry and Sibley, 1976). Yet, current data/knowledge-bases do not support schema changes and rely on predefined entities of interest and static relationships between them (P. P. Chen, 2006; Parsons and Wand, 2000). Ultimately, how designers choose to model the world (as reflected by the structure imposed by the system) constrains the degree to which the system is able to reflect reality without neglecting the "dynamic" nature of the world it represents. Lukyanenko (2014)) found that relaxing rigid constraints of a system may help in capturing user input more objectively and completely –and even allow to extend the original scope of the system.

Despite much research has focused on well-defined information needs via structured data entry, IDC estimates that more than 90% of the data generated is unstructured (Gantz and Reinsel, 2011). We operationalize unstructured data as data without a predefined data model (e.g., free-form text) and we include semi-structured data that is

neither raw (e.g., images, sound) nor explicitly structured (e.g., data in a relational database) in our definition (Zhu and Azar, 2015; Abiteboul, 1997; Silberschatz et al., 1996; Buneman, 1997). The IS field defines information as “data that has been processed into a form that is meaningful to the recipient and is of real or perceived value in current or prospective actions or decisions” and defines information technology (IT) as the artifacts used to “acquire and process information in support of human purposes” (Davis and Olson, 1984; S. T. March and Smith, 1995). Yet, traditional IS research offers limited guidance in studying the effect of unstructured data-entry practices in decision-making (e.g., alignment between the information needs of data consumers and data contributors or promoting effective data-entry practices). To generate competitive edge, organizations should be able to leverage their existing stored data to solve tactical needs and be able to integrate these data with both internal and external data sources to solve emerging tasks efficiently.

Business analytics is an emerging area that organizations are leveraging on to develop competitive edge (Davenport and Harris, 2007). The increasing computational power and the availability of analytical tools allow organizations to use these tools to solve unanticipated tasks. Data mining tools can help organizations identify patterns in complex data sets (Davenport and Harris, 2007). For unstructured data, additional insight can be uncovered using knowledge discovery strategies using ontologies, natural language processing, and semantics to generate structure meaningful to solving the organization’s business needs’. For example, previous research has predicted fall in the elderly by analyzing unstructured progress notes (Tremblay et al., 2009), identified

patient smoking status from discharge records (Uzuner et al., 2008), classified breast carcinomas based on variations in gene expression patterns and then correlate tumor characteristics to clinical outcome (Sørli et al., 2001), created a cardiovascular profile score to predict presence of congestive heart failure (Hofstaetter et al., 2006), identified intellectual communities in the field of information systems (Larsen and Bong, 2016), detected discordant naming practices of constructs (e.g., same term to refer to different phenomena or using different terms to refer to the same phenomena) (Larsen and Bong, 2016), identify adverse drug interactions (Iyer et al., 2014), or extract information from textual documents in the electronic health record (Meystre et al., 2008).

Motivated by the ever increasing growth of unstructured data in organizational settings and the need of organizations to leverage on existing data for decision making, in this paper (1) we hope to understand the underpinnings of *unstructured-data-entry* formats in the data collected by an organization; (2) the impact *unstructured-data-entry* formats have in solving a tactical need and (3) what are effective practices in *unstructured-data-entry* and how effective practices can help the organization in decision making. Our research aims to increase understanding of the implications of free-text-data-entry strategies and its implications to solving tactical needs'. In the following section we further motivate the importance of adopting effective strategies. In subsequent sections we introduce our case study, which is in the context of case management in a large foster care organization.

Motivation

In the context of foster care, it is particularly important to track the status of at-risk patients, (e.g., those with conditions or history associated with adverse outcomes). Clinicians must identify the population of clients with the condition(s) and/or history of interest, and then track the status of urgency over time to try to avoid unexpected adverse events that could otherwise have been prevented. Thus, task prioritization is one important aspect in achieving timely interventions and better outcomes. This is challenging in real-world settings because each patient's case history is encoded in a set of unstructured encounter notes. In fact, even the first step, correctly classifying clients into at-risk groups, can be difficult, because the encounter notes may or may not be explicitly coded with markers indicating conditions or history of interest.

Failure to identify at-risk clients is highly problematic, because adverse outcomes can include serious health issues—including death. It is well known that decision making performance is directly tied to the quality of the information used to make decisions (O'Reilly, 1982; Zmud, 1978). In health care settings, clinicians and administrators adopt medical quality management and case review practices to ensure compliance. Since data is often encoded in free-text form (e.g., reports, case notes, progress notes), we want to study the impact of data-entry formats have in solving a tactical need.

Our goal is to comprehend whether different data entry practices lead to different outcomes. To do so we are going to address the following research questions:

Research Question 1: Does unstructured data entry result in differences in how information is collected across organizational units in an organization?

Research Question 2: Do individuals from different organizational units adopt consistent practices when entering free-text-notes into the information system?

Research Question 3: Can organizations foster effective unstructured-data-entry practices that result in more effective data collection?

Child Welfare

The child welfare system describes a continuum of services that include child protective services and foster care. Children in foster care are at increased risk of child abuse and neglect (e.g., emotional, behavioral, developmental, and physical health problems) (Halfon and Klee, 1991; Simms et al., 2000). Many of these children remain for significant periods of time in the foster care system. The child's background (e.g., age, race, health status) is what most likely determines the services needed by them. Some of the risk factors experienced by these children include low levels of parental education, residential mobility, poverty, and poor parenting (e.g., parental substance abuse, maltreatment). It is likely for these children to undergo anxiety, depression, suicidal thoughts, eating disorders, hostile behavior, and substance dependency (Barbell and Freundlich, 2001; Schneiderman, 2003). Moreover, due to the extenuating conditions these children face, they are more likely to suffer from acute and chronic health conditions (e.g., respiratory infections, dental caries, and malnutrition). The goal is to keep children safe and protect them from harm (Whitaker, 2004).

Since the service systems have not kept pace with demand, prevention and early intervention services required for these children becomes a challenge. A study found that children that have experienced multiple placements tend to have higher levels of

behavioral and emotional problems and as a result, remained longer in foster care (Barbell and Freundlich, 2001).

Federal and State Legislation

Federal legislation has played and continues to play a key role in shaping foster care through public policy and legislation. U.S. Congress enacted the Adoption and Safe Families Act (ASFA) in 1997 to ensure the child's safety and to promote adoption with the principle of "reasonable effort". ASFA was also designed to hold states more accountable for achieving positive outcomes for children and families (Whitaker, 2004). The U.S. Department of Health and Human Services (HHS) mandates periodic child and family service reviews to assess each state's performance on three critical outcomes of foster care: child safety (e.g. protection from abuse and neglect), permanence (e.g., stability of children's living arrangements), and well being (e.g., adequate education and physical and mental health needs). Thus, it is imperative to have an experienced and competent workforce. Yet, the high staff turnover, the ever-increasing workloads per caseworker, and the rising rate at which children are entering the foster care system makes it challenging. This is further aggravated by high stress from poor administrative support, bureaucracy, insufficient salaries, or budget-driven staff reductions (Rycraft, 1994; Barbell and Freundlich, 2001; D. G. Anderson, 2000). It is estimated that child welfare workers spend 50 to 80 percent of the time on paperwork (Office, 2003).

To support the ever-increasing workload, States are leveraging on technology to make the case management process more efficient for the caseworkers and safer for the children by reducing information silos, increasing accountability, and ensuring data

quality. In Florida, data is stored in the Florida Safety Families Network (FSFN), a Statewide child welfare and client management information system developed by HHS to document child protective investigations and child welfare case management (e.g., reporting abuse and neglect, adoptions, permanency planning). HHS contracts with Community-Based Care (CBC) agencies to provide services for vulnerable children and their families. CBC agencies may subcontract some services to specialized providers. The organization completes quarterly quality reviews with the Department of Children and Families (DCF) covering tasks ranging from family engagement to supervisory reviews. Ultimately, the goal is to use the resources necessary to achieve better outcomes. Information is retrieved from different systems –both internal and external. Examples of external systems include the school system, juvenile justice system, medical system, and legal system.

Foster Care and Case Management

This chapter focuses on how unstructured data shapes practice across organizational units in the context of case management in a foster care organization—where different caseworkers (from different agencies) report on the home visits made to the foster children. The focus of this essay is to understand the dynamics in how different full case management agencies (FCMAs), which we refer to as organizational units of a governing parent organization, collect data via unstructured formats (Eisenhardt, 1989). The focus becomes in studying whether the effectiveness of solving a tactical need is dependent on the organizational unit doing the reporting.

The organization is a non-profit corporation created by advocacy communities in response to the need for leadership, oversight, and coordination of a system of care for abused and/or neglected children, and children at-risk of abuse and/or neglect. The organization's purpose is to develop, operate, expand, and enhance initiatives aimed at the prevention of child abuse and neglect; to support networks of coordinated resources and activities to better strengthen and support families; and reduce the likelihood of child abuse and neglect. The organization's approach is designed to address the individual needs of children and their families and articulates specific principles of care, including the requirement that all child-serving sectors (mental health, education, child welfare, juvenile justice, and physical health care) integrate and coordinate their service provision.

The social work profession involves caring for those who are most likely poor, neglected, and vulnerable. Social workers often make personal and professional commitments to protect children. We surveyed 30 staff members at the organization that work in different capacities (e.g., supervision, management, services). At the time of the interview, the median for the time of employment in the organization was 23 months with a range from 4 to 156 months.

The Office (2003) found that the average tenure of child welfare workers is less than 24 months. The median age of the staff was 31 years old with a range from 24 to 66 years old. Almost 43% of the caseworkers held Bachelor degrees (2 respondents held a Bachelor's in Social Work [BSW] degree), almost 36% held a master's degree (4 held a master of social work [MSW] degree), and the rest have either registered nurse or business associate degrees. Seven respondents did not report their ethnicity. Out of the 23

left, close to 30% were Hispanic, 30% black, 17% African American, and 22% are White.

The Child Welfare League of America (CWLA) recommends caseloads not exceed 18 per worker. Supervisors are required to complete quarterly supervisory reviews with the staff and case managers they supervise. In our study, the full case management agencies seem homogenous on the surface—they all report to the same overseeing organization; use the same systems to carry out their functions; are geographically collocated in the same city; have employees with similar educational background and similar demographics; train caseworkers in the same facility, by the same trainer, and use the same information systems and devices.

Identifying Psychotropic Drug Use

Children in foster care are three to ten times more likely to suffer from mental health conditions (Harman et al., 2000) thus receiving behavioral health services to a greater extent compared to other children. Psychotropic medication is prescribed to help them cope with behavioral problems such as attention-deficit/hyperactivity disorder (ADHD), depression, bipolar disorder, and psychotic disorders—in many cases these children are prescribed concomitant medication with dosages that are regularly used for adults. Despite their challenging lives as foster children, those with behavioral problems are frequently the ones that do not find a stable placement, limiting the possibility of reliable and consistent treatment as a result of inaccurate medical, behavioral, and psychological history from previous care providers (Zima et al., 1999). This is threatening for the children especially when using: 1) antidepressants, in which adverse side effects include

suicidal thoughts; 2) anti-anxiety medications, which side-effects could trigger blurred vision, drowsiness and dizziness, and nightmares; or 3) mood stabilizers, which treat bipolar disorders but may have side effects such as hallucinations and suicidal thoughts (GAO, 2011).

In April of 2009, Gabriel Myers, a 7-year-old child who had been taken from his drug-abusing mother and who had been sexually abused in a previous foster home, locked himself in the bathroom and hanged from a detachable showerhead and committed suicide. At that time he had several psychiatric drugs prescribed –three of which were labeled as “black box” medication (the strongest advisory alert that the U.S. Food and Drug Administration issues, indicating that the drug can pose life-threatening adverse effects including suicidal tendencies in children) (Martinez, 2010). Another impactful case is that of Denis Maltez, a 12-year-old autistic boy who died of “serotonin syndrome” after being prescribed several psychotropic drugs in the highest doses, dosages that are typically given to adults. In this particular case, DCF received a report from a school teacher stating that Denis was “sleeping in class, shaking, and trembling” and a second medical report from the hospital which stated that “Denis was sleepy because he was over-medicated” (Miller, 4/18/2010).

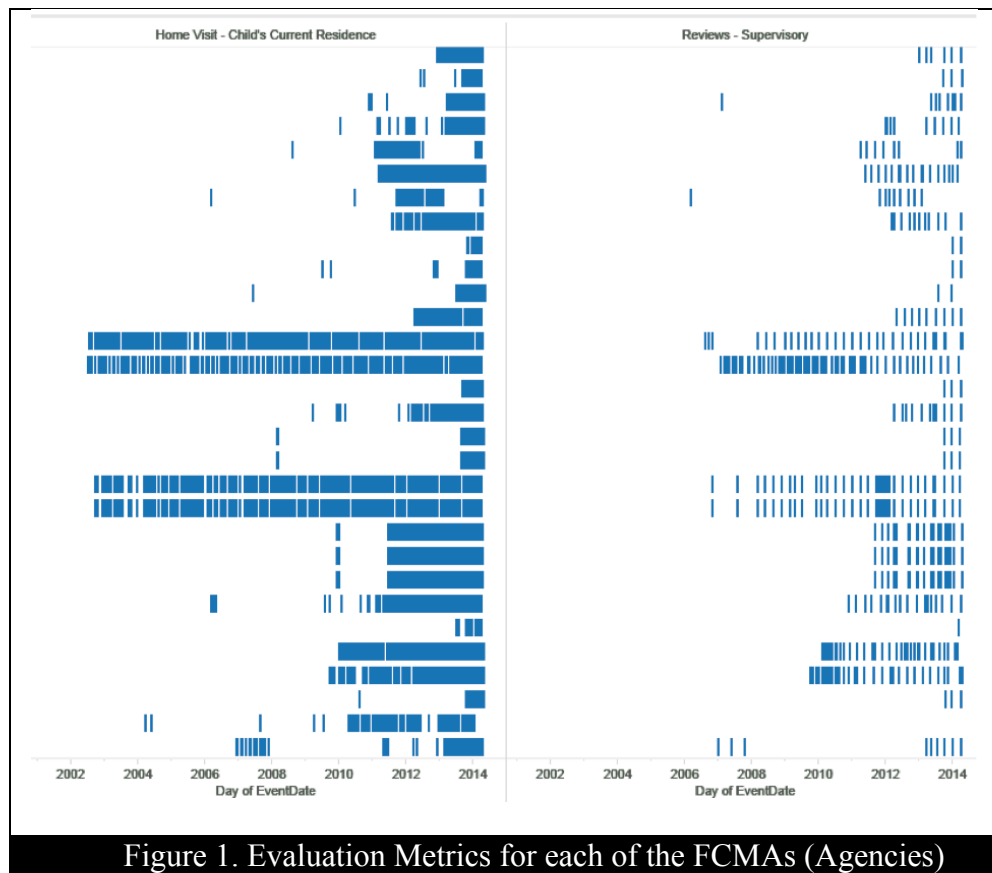
To ensure safety and well being of the children, DCF tracks, via the state run FSFN, all psychotropic drugs provided to children in foster care. Some of the fields include, but are not limited to, medication name, dosage prescribed, number of refills, prescribing physician, whether the drug is used as psychotropic medication, and the start and stop date. This system assumes that the information introduced by the case manager is reliable

and complete, but unfortunately there are few built-in mechanisms to prevent data quality issues. For example, a generic medication to which the caseworker does not know the brand equivalent is placed as “other”. Adding to data quality problems is the ability to leave blank fields (Group, 2009).

According to Section 65C-30.007 of the Florida Administrative Code (F.A.C.), children under the state’s supervision need to have a face-to-face visit at least every 30 days. The record of this visit should include developmental, physical, emotional, and mental health needs and whether those needs are being met. These visits should also be documented in the child’s FSFN within two working days. Although not explicitly stated in the case notes whether the child takes psychotropic medication or not, we attempt to use these *home visit notes* to identify children taking psychotropic medication, a tactical purpose for which home visit notes were not originally intended for.

Supervisory notes have been a good proxy to identify children on psychotropic medication as they contain explicit notes regarding the use of psychotropic medication (e.g. drug name, dosage). However, using visualization techniques, we found that the frequency in which these notes type were reported was very scarce (see Figure 1). Upon further investigation, we found out that the DCF OP under section 3-14 (n-p), which refers to supervisory reviews for children prescribed psychotropic medication, state that behavioral events should be reported by DCM and CPI supervisors on an “ongoing” basis, without stating specifically the frequency in which these reports should be written. The organization conducts monthly reviews of 100% of all children listed on FSFN that are taking psychotropic medication. Additionally, a 10% random sample of all the out-of-

home children who do not have an active medication profile in the system notes are reviewed, in an attempt to identify false negatives. We found out that, among children under psychotropic medication, the frequency of supervisory notes varied drastically—in many cases they were non-existent. Home visit notes, on the other hand, need to be filed at least once every 30 days. Consequently, we want to determine whether these home visit notes can serve as a proxy for identifying children on psychotropic medication.



Although supervisory notes might contain more explicit information about psychotropic drug use, these notes are written less frequently than home visit notes. Supervisory notes are written by case managers when they engage with foster children

through visits or telephone calls and may include interrelated behavioral indicators such as symptoms (e.g. aggression, lack of eye contact, bedwetting, extreme distraction) of a child taking psychotropic medication. The amount of case notes of all the foster children in this particular organization is large, making it difficult to oversee more than a portion of records manually. Although the problem of predicting psychotropic medication use is a big data problem, we have observed that institutional factors such as organizational norms and procedures can indicate differences in how different organizational units document –what is emphasized or de-emphasized in a document.

Theory and Propositions

In this section we seek theoretical foundation to understand what makes effective practices in settings that rely on unstructured data-entry. We review research in organizational behavior to try to explain the impact unstructured data formats have on data collection practices and ultimately assess the impact on the organization's performance in decision-making.

For research questions one and two which seek to study whether there are any differences in the way different organizational units adopt and are consistent in how to document home visit notes, we turn to organizational behavior theories to try to understand why individuals in organizations adopt established practices and create new ones that are adopted over time. To answer research question three, which relates to effective practices of data-entry, we turn to psychology theories to explain the tradeoff of generalization/specification in data collection practices.

Institutional Theory

Organizational activity (social and non-social) can become a pattern that is repeated by individuals in the organization. The concept of institution has been operationalized in diverse ways throughout the years. Early versions of institutional theory viewed institutionalization as a process of adaptively changing commitments and by which individuals come to accept a shared definition of social reality (W. R. Scott, 1987).

Rules, norms, and meanings arise in interaction, and they are preserved and modified by the behavior of individuals (Giddens, 1979; Sewell Jr, 1992). Social order is created as a shared reality by which individuals interpret actions that are then internalized and shared with others as a socially defined reality (Berger and Luckmann, 1991). These interpretations enable actors to respond in a similar way as taken-for-granted realities that reach stability and which structure's may evolve over time (Barley, 1986; Giddens, 1984; Selznick, 1984). Formally, institutional theory considers the processes by which structures, including schemes, rules, norms, and routines become established as authoritative guidelines for social behavior (W Richard Scott, 1995; W. R. Scott, 1987).

Institutions have been studied at various levels of analysis - from micro interpersonal systems to transnational or world systems. Of particular interest to our research is to understand institutionalization at the organizational field level and at the population level (see Table 1). Field refers to a community of organizations that partakes on a common meaning system and whose participants interact with one another more frequently than with actors outside the organizational field (W Richard Scott, 1995). Populations refer to groups of organizations that are "alike in some respect" (Hannan and Freeman, 1977).

Table 1. Unit of Analysis of Institutionalization Adapted from Scott (2001 p. 85)	
Level of Analysis	Example in the Context of the Case Study
Societal	Children’s well-being in Foster Care
Organizational Field	Foster Care Management
Organizational Population	CBC Agencies
Organization	Agency A, B, or C
Organizational Subsystem	Case Management at Agency A

Institutions are built on three pillars –regulative, normative, and cognitive (W Richard Scott, 1995). The *regulative* pillar regulates individual actions and behaviors to avoid the violation of institutional rules and prevent organizational sanctions (e.g., rule-setting, monitoring, and sanctioning activities). The *normative* pillar provides a structure for legitimization by which specific behaviors are believed appropriate and introduces a prescriptive, evaluative, and obligatory dimension through values (desirable outcome) and norms (how things should be done) (W Richard Scott, 1995). Some values and norms are applicable to all members in the organization and others apply to a subgroup via roles, which can be viewed as patterns, goals, attitudes, and behaviors that are characteristic of individuals under certain situations –becoming the controlling character of institutionalization (Berger and Luckmann, 1991; Searing, 1991).

The *cognitive* pillar provides a structure of signification via cognitive guides that help individuals understand how they should act. The regulatory processes, normative systems, and cultural frameworks shape the tasks of the individuals and ultimately shape the design and use of technical systems (e.g., which systems to use, what data to input into the information system).

In the absence of contextual change, actors are more likely to replicate scripted behavior, making institutions persistent (Hughes, 1936; Barley and Tolbert, 1997). Yet, this behavior can evolve over time as a result of changing regulations and norms (e.g., solving an emergent tactical purpose or when solving wicked problems). The demand for such coherence also cultivates strong expectations regarding styles, creating preferred forms of knowledge representation and production. The process of standardizing procedures among members of a population from these pillars is referred to as institutional isomorphism, which is triggered by coercive, normative, and mimetic forces—constraining the ways in which individuals perform their activities (DiMaggio and Powell, 1983). *Coercive* isomorphism stems from political influence and formal and informal pressures exerted on organizations by other organizations of which they are dependent from or by cultural expectations in society (e.g., laws, policies, social norms). *Mimetic* isomorphism results from the adoption of existing practices to reduce uncertainty and achieve legitimization (e.g., mimicking the behavior of other organizations perceived as legitimate). *Normative* isomorphism is associated with professionalization and the collective struggle of members to define conditions and methods of their work (e.g., formal education, network, skills, and knowledge of the workforce) (DiMaggio and Powell, 1983). Institutions are made up of different combinations of these institutional elements—varying among one another and over time in the elements given priority.

Although regulative features are more visible, they can also be more superficial, "thinner," and less consequential than normative and cultural elements (W Richard Scott,

2008). Governmental regulations have traditionally been depicted as forms of coercive power, imposing conformity on affected actor (whether individual or collective). Neoinstitutionalists emphasize the extent to which such "requirements" are subject to interpretation, manipulation, revision, and elaboration by those subject to them, implying a transfiguration over time of regulative into normative and cultural-cognitive elements. For example, we mentioned earlier the example of vague operating procedures in the context of foster care management, such as the DCF OP section 3-14 (n-p), which state that behavioral events should be reported on an "ongoing" basis, without stating specifically the frequency in which these reports should be written, leading to an incomplete set of supervisory notes of children that may be taking psychotropic medication or children. This regulative measure in practice can be improved by the organization (or organizational units) by specifying what "ongoing" should be. Table 2 summarizes (from the literature) the theory elements of institutionalization, indicators, and predictors of isomorphic change. For illustration purposes we provide, for each of the theory elements, an example of potential triggers of isomorphism in the context of foster care.

Table 2. Institutional Theory Elements
Adapted from (DiMaggio and Powell, 1983; W Richard Scott, 1995)

Theory element	Indicator	Predictors of Isomorphic Change	Isomorphism Triggers in Foster Care
Regulative (Coercive)	Rules, laws, and sanctions	<p>“The greater the dependence of an organization on another organization, the more similar it will become to that organization in structure, climate, and behavioral focus”.</p> <p>“The greater the centralization of an organization A’s resource supply, the greater the extent to which organization A will change isomorphically to resemble the organizations on which it depends for resources” (DiMaggio and Powell, 1983).</p>	<p><i>Compliance with the Adoption and Safe Families Act (ASFA).</i></p> <p>Funding is determined based on compliance with statutory requirements. This is done through external quality assurance to monitor and support services.</p> <p><i>Measurement:</i></p> <p>Compliance can be monitored through the use of performance scorecards, corrective action plans, customer satisfaction surveys, and complaint monitoring and investigation.</p>
Normative (Normative)	Certification, accreditation	<p>“The greater the extent of professionalization in a field (e.g., credentials, certificates, training programs), the greater the amount of institutional isomorphic change”</p> <p>“The greater the reliance on academic credentials in choosing managerial and staff personnel, the greater the extent to which an organization will become like other organizations in its field” (DiMaggio and Powell, 1983).</p>	<p>Foster care staff requires a strict set of qualifications to be able to work with foster children. The organization requires for its case managers to have at least a bachelor’s degree in social work [BSW]. Many of the case managers also hold a master’s degree in social work [MSW] with similar demographics.</p>
Cognitive (Mimetic)	Prevalence, isomorphism	<p>“The more uncertain the relationship between means and ends the greater the extent to which an organization will model itself after organizations it</p>	<p>Foster care organizations may adopt practices (imitate) from institutions they perceive to be successful as to avoid uncertainty. An example</p>

		perceives to be successful”. “The more ambiguous the goals of an organization, the greater the extent to which the organization will model itself after organizations that it perceives to be successful” (DiMaggio and Powell, 1983)	at the organization was to add GPS tracking to enhance the accountability of home visits and safety of the children. Mobility solutions like this have worked successfully in other industries such as fleet tracking, police force tracking, or tracking services for the elderly.
--	--	--	---

A more balanced rationale for understanding institutional order involves “softer” cultural, cognitive, and normative elements (W. W. R. Scott, 2013)—looking at institutions and actions as intertwined together in a process of structuration (Barley and Tolbert, 1997; Orlikowski and Barley, 2001; DeSanctis and Poole, 1994). Social structure (as defined by (Barley, 1986)) can be influenced by the interaction of institutionally-triggered and technology-triggered change processes. If organizational practices are deeply influenced by historical traditions and enduring value (and if they are supported by societal sources of legitimacy), strong resistance to transformation can be expected (Robey and Boudreau, 1999; Boudreau and Robey, 2005). Practices and behavioral patterns may not be equally institutionalized –institutions that have a relatively short history or that have not yet gained widespread acceptance by members of a collective are more vulnerable to change and less apt to influence action (Tolbert and Zucker, 1999).

Organizations are involved in both horizontal (cooperative-competitive) and vertical (power and authority) connections. They operate in systems composed of both similar and diverse forms. These organizations typically establish processes to solve tactical

problems—as reflected by the standard operating procedures, which are followed by individuals in the organization. Despite the quest for isomorphic practice, individuals in different organizational units may deviate from these procedures. For example, in the context of foster care management, when leaving too much flexibility to caseworker’s data entry (e.g., free-form text notes), the style of the notes (although adhering to the general standard) can differ from that of other individuals within and across organizational units—based on internal norms of a subgroup (e.g., an internal tactical purpose they are interested in capturing or by inherent styles in data collection practices). The question then becomes, how does unstructured data entry result in differences in how information is collected across organizational units in the organization?

Organizational decision-making is not just a byproduct of individual intellectual information processing, it also involves social information processing (M. S. Feldman and March, 1981; P. A. Anderson, 1983) that in the absence of contextual change, actors are more likely to replicate scripted behavior, making institutions persistent (Hughes, 1936; Barley and Tolbert, 1997). Institutional isomorphism constrain the ways in which individuals perform their daily activities (DiMaggio and Powell, 1983). This coherence cultivates expectations regarding the style of knowledge representations and production. The concept of institutional isomorphism in organizational behavior theory leads to our first proposition:

Proposition 1: *Data collected using unstructured-data-entry formats become isomorphic within organizational units.*

Organizations collect data in different forms (e.g. structured, unstructured) following a pre-defined process of reporting established by the organization –and its goals. The inherent flexibility on how free-text data are entered/captured in an information system allow users to deviate from the original structure and capture different information in a field that was not originally intended for (Boudreau and Robey, 2005; Wand and Weber, 1995; Berg and Goorman, 1999). Despite this, organizations should be able to analyze data in the aggregate and enable effective decision-making.

Institutionalization has been viewed as a bottom-up social process by which individuals come to accept a shared definition of social reality (W. R. Scott, 1987). Organizations are coerced to conform (imitate) to the existing status quo—that allows the organization to gain the legitimacy and resources needed to survive (Meyer and Rowan, 1977). By incorporating institutional rules within their own structures, organizations become more homogenous over time, achieving stability, which is reflective of the influences in this shared definition of social reality (Selznick, 1984; W Richard Scott, 1995). The resulting structure and processes can be a formal social order (e.g. table of organization), or an informal social order (e.g. cross-functional actors involved in a given process). The social order may vary from the expectations but it is also based on a shared reality between the social actors.

Social actors can create semi-institutional structures (that differ from the norm) that can be subject to objectification and become diffused despite having a short history. While these informal structures may acquire some degree of *normative* acceptance, adopters nonetheless are apt to remain cognizant of the effectiveness of adopting such

structure, which can be legitimized and achieve stability over time (Barley, 1986; Giddens, 1984; Selznick, 1984).

The *data collection isomorphism* principle would suggest the potential for organizations to adopt standard practices in how they collect and use the information to solve a tactical need. The effectiveness on their decision-making is tied to the information at hand to solve such tactical purpose. As the number of autonomous decision-making is minimized, the risk associated with having to make a choice is also minimized, reaching isomorphism. Institutional features of organizational environments, however, can shape the actions actors take (e.g., the level of detail –specificity or focus– at which they input the information into the IS). This notion of institutional factors of reporting leads to our second proposition:

Proposition 2: *Institutional factors can establish data entry practices that result in highly cohesive (similar within the same organizational unit) and loosely coupled (different across organizational units) data collection.*

Psychological Foundation

To address research question three, which relates to effective practices of data-entry, we turn to psychology theories to explain the tradeoff of generalization/specification in data collection practices. According to psychology, classes support vital functions of an organism via *cognitive economy* and *inductive inference* (Lakoff, 1987; Roach et al., 1978; E.E. Smith and Medin, 1981; Edward E Smith, 1988; Parsons, 1996). Both functions compete for *limited* cognitive resources of human *memory* and *processing*

power. Cognitive economy is achieved by maximally abstracting from individual differences among objects and then grouping objects in categories of larger scope (Fodor, 1998; G.L. Murphy, 2004; E.E. Smith and Medin, 1981). Overemphasizing cognitive economy, however, comes at the expense of ignoring certain individual characteristics of organisms that may be vital for the organism's function and survival.

A category groups together non-identical elements, which, by virtue of their common membership, can be treated as equivalent (Gregory L Murphy and Brownell, 1985). Categories improve the ability of a person to accurately predict features of instances of a category. The best categories are those that maximize feature predictability and optimize information transfer (Corter and Gluck, 1992). For example, suppose we wish to discern if a mushroom is *poisonous* or *edible*. Classifying it as a *fungus* (a less specific high-level object) versus *Clitocybe rivulosa* (a more specific kind of poisonous mushrooms) provide a higher likelihood of this object having the property of interest. The likelihood of a *Clitocybe rivulosa* being *poisonous* is substantially higher than the likelihood that any *fungus* is *poisonous*. This example also demonstrates why a domain, such as biology, is interested in a finer species level of classification. Knowing that a phenomenon is *Clitocybe rivulosa* affords greater inferences and action than knowing it is a Fungi. Thus, the ability to predict attributes of instances of a class, or the inferential power, increases as the scope of the class decreases.

The trade-off between these competing functions is considered one of the defining mechanisms of human cognition and behavior (Corter and Gluck, 1992; Roach et al., 1978). According to cognitive theories and theories of classification, classes provide

cognitive economy and inferential utility, enabling humans to efficiently store and retrieve information about phenomena of interest (Parsons, 1996; Roach et al., 1978). A class is a mental model of perceived reality learned or derived from prior experience (G.L. Murphy, 2004). Psychology hypothesize that humans favor (e.g., learn, communicate) those classes that maximally exploit both predictive power of classes and their cognitive economy. Rosch et al. (1976)) argued that humans favor classes that are most capable of supporting these competing objectives of classification. While cognitive economy mainly deals with communication, memory, and processing, inferences are the primary drivers of human behaviour and decisions (Tsui et al., 2010; E. Smith, 1989). Thus, specificity allows for unanticipated uses and increases the predictive accuracy—since it provides the ability to make more inferences from the data (Cruse, 1977; Brown, 1958; Tanaka and Taylor, 1991).

The basic-level advantage changes with expertise (Johnson and Mervis, 1997; Tanaka and Taylor, 1991). Experts in some domain of knowledge can make use of attributes that are ignored by the average individual. Expertise does not have to span an entire domain. Instead, it could be quite narrow in scope, perhaps limited to a single specific category. For example, a person who owns a collie and spends a lot of time with the dog could be considered a “collie expert.” Such person might be aware of the distinguishing features of collies, but know very little about distinctive properties of other breeds of dogs. Thus, individual differences in how objects are categorized can be a function of idiosyncratic life experiences and/or culture (Tanaka and Taylor, 1991; Wales et al., 1983; Brown, 1958). In the healthcare domain, a patient with a chronic condition

such as diabetes may have developed some expertise in the care of that condition but does not make him an expert in other medical conditions. Experts frequently use the subordinate name in their field of expertise whereas non-experts use basic level names (Macé et al., 2009; Tanaka and Taylor, 1991; Jolicoeur et al., 1984). Research has shown that for categories outside the domain of expertise (e.g., bird categories for dog experts), subjects are able to list more features for basic-level categories than for subordinate-level (more specific) categories. Experts, however, know as much about the features of basic-level categories as they know features of the subordinate-level categories, whereas novice individuals may only be familiar with categorization at the basic level (Tanaka and Taylor, 1991). In general, as people specialize they are more comfortable using specific language, which has higher inferential utility.

Users with different levels of expertise tend to produce information that differs in quality. Accuracy is contingent on providing users with classification structures more congruent with the level of expertise of the user. Lukyanenko et al. (2014) suggests that in a free-form data entry task, non-experts will classify more accurately at the basic level than at a more specific level. When we collect structured data the level of specificity is fixed at the time of system design. Users entering unstructured data, on the other hand, can adjust to their level of specificity—by being more or less detailed. Since specificity results from expertise, unstructured data collection can capture expertise better, which may lead to better performance (e.g., providing relevant information for decision-making).

Our research questions reflect on whether organizations can foster effective unstructured-data-entry practices that could result in richer data collection. We do so through the following propositions:

Proposition 3: *Unstructured data formats can help shape effective data-entry practices in solving well-defined needs.*

Proposition 3a: *Higher levels of specificity in the data collected leads to increased inferential utility.*

Proposition 3b: *Higher levels of specificity in the data collected facilitate unanticipated use of the data.*

We evaluate the propositions presented here via a case study of case management practices in a large foster care organization. The next section describes the characteristics of the organization, their practice, and the tactical purpose studied.

Method

A case study is a suitable observational evaluation method of an artifact in a business environment (von Alan et al., 2004). The case method allows us to understand the nature and complexity of the processes taking place by answering "how" and "why" questions by examining a phenomenon in its natural setting (Benbasat et al., 1987; Dubé and Paré, 2003; Lee, 1989). The case study method has been an essential form of research in the social sciences and management (Chetty, 1996). Yin (2013) defined case studies as a research strategy that focuses on understanding the dynamics within single or multiple settings (Eisenhardt, 1989). Case studies can employ multiple levels of analysis within a

single study and can combine different data collection methods –both qualitative and quantitative (e.g., interviews, questionnaires, physical artifacts, and observations) (Yin, 2013).

Case studies can provide description (Kidder, 2011), test theory (P. A. Anderson, 1983), or build theories (Eisenhardt, 1989; Gersick, 1988; Harris and Sutton, 1986; Dubé and Paré, 2003). In this paper, we adopt Walls et al. (1992) definition of information system design theory (ISDT) as a prescriptive theory to produce more effective information systems through design propositions (Dubin, 1970; Simon, 1996). What distinguishes design theory is the inclusion of a kernel theory to explain testable propositions or design principles in developing comprehensive bodies of knowledge (Gregor and Hevner, 2013; Gregor and Jones, 2007). Analyzing data constitutes the “heart” of building theory from case studies (Eisenhardt, 1989). Two key features of analysis are: within-case analysis (to provide familiarity with the case at hand) and cross-case analysis (look at the data using different lenses). Tying the emergent propositions to existing organizational theory enhances the internal validity and generalizability of the theoretical propositions (Eisenhardt, 1989; Chetty, 1996). The proposed design propositions are an approximation to what will work in different contexts and can be tested through an instantiation or deductive logic that lead to conclusions with some generality (Gregor and Jones, 2007; Gregor, 2006).

Solution Approach

We adopt a mixed-method approach to evaluate the propositions derived from theory. To evaluate proposition one in this case study, we use text mining techniques to discover and extract knowledge from unstructured data (Hearst, 1999). To evaluate proposition two we use a particular application of text mining named Stylometry. To evaluate proposition three we adopt both a quantitative and qualitative approach to assess any similarities or differences within notes from different organizational units. From a quantitative standpoint, we use text mining techniques to assess whether there are any significant lexical, syntactic, or semantic differences in the text authored by different organizational units.

Text Mining

Text mining is a process of knowledge discovery via a set of techniques and tools that allow for “nontrivial extraction of implicit, previously unknown, and potentially useful information from given [free-form, or textual] data” (R. Feldman and Dagan, 1995). We use an inductive classification approach to classify children taking psychotropic medication and evaluate the results of our design by benchmarking with the results given by expert case managers.

Text mining has had significant improvement over the years and has shifted from simple metrics (e.g., word frequency) to more complex use of natural language processing (NLP) techniques. Common NLP tools include document tokenizing, stemming, parts-of-speech tagging, noun group extraction, applying stop lists, entity identification, and multiword terms handling (Christopher D Manning and Schütze,

1999). The document is parsed and tagged based on the syntactical relationship between terms –based on the position in a sentence and rules of grammar (Berry and Castellanos, 2004). The aim is to convert human language into formal representations computers can manipulate, including part-of-speech tagging (POS), POS sequences, or n-gram models (see Table 3) (Abbasi and Chen, 2008; Holmes, 1998; Christopher D Manning and Schütze, 1999). Because authors do not always follow grammatical rules, the complexity of multiple meanings for words, and the domain specific use of vocabulary may require some additional considerations.

Table 3. Parts-of-speech (adapted from (Bird et al., 2009))		
Part-of-speech	Example	Example Text
Adjective	Psychotropic, happy, clean, visible	“Prescribed psychotropic medication”, “child has a visible bruise”
Adverb	Reportedly, temporarily, friendly	“Child was reportedly not home”, “ he was friendly”
Conjunction	If, but, and, or	“Appropriately dressed and groomed”
Determiner, article, quantifier	The, a, few, most, little, no, which	“No signs of abuse and neglect”
Noun	Husband, guardian, mommy	“To see his mommy and daddy”
Pronoun	I, that, he, who, them	“I went to school”
Verb	Risk, reunify, hit, fight, approve	“She doesn’t fight anymore”

Text classification is a discipline at the crossroads of machine learning and information retrieval—an inductive process of building a text classifier that is able to learn from a training set of labeled documents without being explicitly programmed. In information retrieval, a document is parsed and then transformed into a vector space model (VSM), a numerical representation of the document (G. Salton et al., 1975). To algorithmically process the text it is necessary to create a term-by-document matrix, a

matrix in which table columns represent all terms in a document and each row represents a document. The resulting cells represent either the existence (*term frequency – local weight*) or relevance (*term weighting – global weight*).

Term weighting techniques provide a greater degree of discrimination among terms by modifying the frequency weights to adjust for document size and term distribution, distinguishing individual documents from a collection of documents (Sparck Jones, 1974; Singhal et al., 1996; Gerard Salton and Buckley, 1988). A common weighting technique is term-frequency-inverse-document-frequency (tf-idf), which produces a composite weight that increases proportionally to the term frequency but is sensitized by the number of documents (Powers, 1998). Other weighting techniques include probabilistic idf, information gain, or chi-square (Lan et al., 2009). In our VSM, each document is a vector that captures the relative important terms. Representing these documents as vectors allows us to perform operations such as scoring documents on a query, document classification, and document clustering (Christopher D. Manning et al., 2008). Deerwester et al. found a way to improve document similarity based on linear algebra called latent semantic indexing (LSI) (Deerwester et al., 1990). LSI assumes a “latent” semantic structure, reducing the dimensionality by using a singular value decomposition (SVD) –a technique related to eigenvector decomposition and factor analysis (Furnas et al., 1988; Dumais et al., 1988; Deerwester et al., 1990).

Stylometry

A particular application of text mining is stylometry, which refers to the statistical analysis of writing style. Stylometry has been influenced by techniques from computer science and artificial intelligence which regards stylometry as a problem of pattern recognition that may distinguish one author from another (Holmes, 1998; Forsyth, 1999; Ramyaa and Rasheed, 2004). The premise is that authors have an inherent writing style that makes their work distinct to that of others. So far most of the efforts in Stylometry research have been in author identification, also known as author categorization, and in similarity detection, which is calculating the similarity between two or more documents (De Vel et al., 2001) (Holmes, 1998; Zipf, 1935; Mosteller and Wallace, 1964). Both of these applications fall under a more general “authorship analysis” (AA).

AA seeks to uncover unconsciously written features from the documents including, but not limited, to lexical (e.g., word or sentence length), semantic, syntactic (e.g., frequency of words), structural (e.g., paragraph length), or content-specific topic (e.g., keywords) features (Holmes, 1998) (See Table 4). A well-known application of authorship analysis is that of the Federalist Papers (Mosteller and Wallace, 1964). In the Mosteller and Wallace study they used author-specific features to establish the authorship of some of the Federalist Papers. Some of these features included unusual diction, frequency in which words appear, and habits of hyphenation and grammar style. Zheng et al. (2006) proposed a framework for authorship identification that includes (besides lexical, syntactic, and structural features) content-specific features.

Table 4. Stylometric Identification. Adapted from (Abbasi and Chen 2008)			
Category	Feature Group	Examples	Information Type
Lexical	Word Lexical	words count, words size	Opinions Style Genres
	Character Lexical	total characters, alphanumeric characters	
	Vocabulary Richness	Hapax legomana, Yules K	
	Word Length Distribution	frequency of various word sizes	
	Character N-grams	ut, utt, utte	
	Digit N-grams	150, 50, 5	
Syntactic (NLP)	POS Tag N-Grams	combinations of parts of speech	Opinions Style Genres
	Word N-grams	to be, be or not	
	Noun Phrases	child, caretaker, parent	
	Named Entities	United States, Dr. Phil	
	Bag-of-Words	all words except function words	
Structural	Document Structure	interview parent, interview child, concerns	Style

In this study we extend the concept from an individual level to the aggregate identity of an organization. If we are able to accurately identify authorship at the organizational unit level, we would demonstrate that case notes coming from different organizations have enough differences between them as to agree there is indeed an “organizational unit style”.

Data Preparation

Among the most important data preparation activities was to solicit the help of nurse case managers. We developed an accurate data sample to construct a “gold standard” dataset with correctly labeled cases of our target variable –psychotropic medication use. The organization agreed to dedicate resources to make sure each of the case notes in our

gold standard were coded and labeled correctly. All case notes were manually checked and coded and include, among other attributes, the content of the home visit note, the authoring organizational unit, and a flag indicating whether the child is taking psychotropic medication.

Many of the cases were difficult to categorize. For example, some of the cases were of children that were previously on psychotropic medication but switched mental health providers and the new physician felt they were too young to take medications and discontinued their treatment. In other cases, the children were refusing medications, or indicating that they are continuing their treatments yet they were not showing up in the system as having received a refill of a prescribed medication. Another limitation of using home visit notes is an issue of cardinality. There is a case note for a home but there can be many foster children living in the same home. This is problematic in cases where one foster child is taking psychotropic medication and the others are not. The caseworker may emphasize one foster child over another yet the home visit is for all.

Data was retrieved from the secure front-end website via Ruby on Rails scripts and were stored in a database on a secure internal server (see Figure 2 for the database structure). Since there is health-related information contained in the various case notes we had to follow strict HIPAA guidelines for personal health information (PHI) de-identification. Using sentence processing heuristics can help in situations where individuals' names can coincide with dictionary words by either removing them or labeling them as "ambiguous" for manual processing (Neamatullah et al., 2008). All labels related to location, contact information, and other PHI was removed. For the free-

text section, because the child and other individuals directly involved in the care were identified in structured sections of the data, those names formed the lookup table for parsing the free-text. Although the dates for the case notes themselves were not particularly important in this study, the order is important, as it tells a chronologically ordered story. For example, some cases resulted in multiple visits, and therefore multiple entries in the system for “Home Visit-Child’s Current Residence”. The dates were replaced with an ordinal number (1, 2, 3...n) to represent the recency of it relative to the child’s other case notes in that month (if any). Children and case identifiers were de-identified by generating a random number and using that number for reference of the child/case. The dataset contained only information relevant for our purpose, including the free-text in the case notes and the target variable.

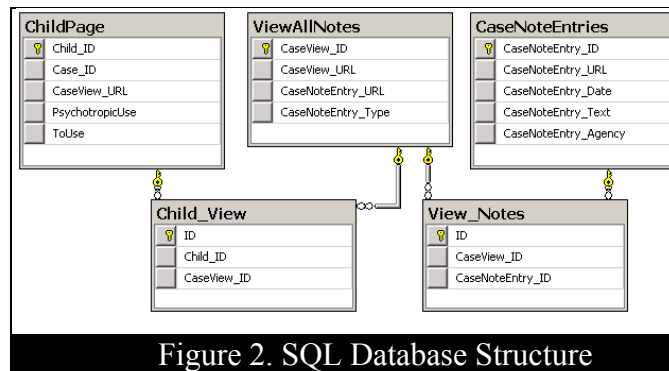


Figure 2. SQL Database Structure

We removed any guiding templates (questions intended to guide the caseworkers’ narrative of the home visit) to (1) avoid misleading results, and (2) saturating the existence of a word/phrase due to inclusion of it in a template (Luther et al., 2011). An example would be if we left a template question that asks to “...list any bruises or markings you observed...”. Due to its existence in much of the data, the signal of what

would have been an interesting stemmed word, “bruise”, would be degraded in the rest of the non-template body of text.

Analysis and Results

To evaluate the differences in performance between the models (for propositions 1 and 2) different predictive models were evaluated and compared using commonly accepted metrics: recall, precision, and F-measure. Recall (R) reflects the percentage of correct positive predictions out of all the possible positives; precision (P) reflects the percentage of correct positive predictions out of the predicted positives; and the F-measure represents a ratio of overall goodness of fit for precision and recall. The F-measure is better suited for evaluation since it provides a harmonic mean of the precision and recall (Christopher D. Manning et al., 2008). The definitions are provided in Table 5 where TP represents true positives, FP represents false positives, TN represents true negatives, and FN represents false negatives.

Table 5. Evaluation Metrics		
Precision (P)	Recall (R)	F-measure
$P = \frac{TP}{TP + FP}$	$R = \frac{TP}{TP + FN}$	$F = \frac{2(P * R)}{P + R}$

Proposition 1: Data collected using *unstructured-data-entry* formats become isomorphic *within organizational units*.

In this section we focus on an inductive (classification) text mining technique. First, an expert case manager provides a gold standard with labeled instances. Case notes are labeled “Yes” (uses psychotropic medication) or “No” (no use of psychotropic

medication), depending on whether the child is taking psychotropic medication or not. The data mining process followed is shown in Figure 3.

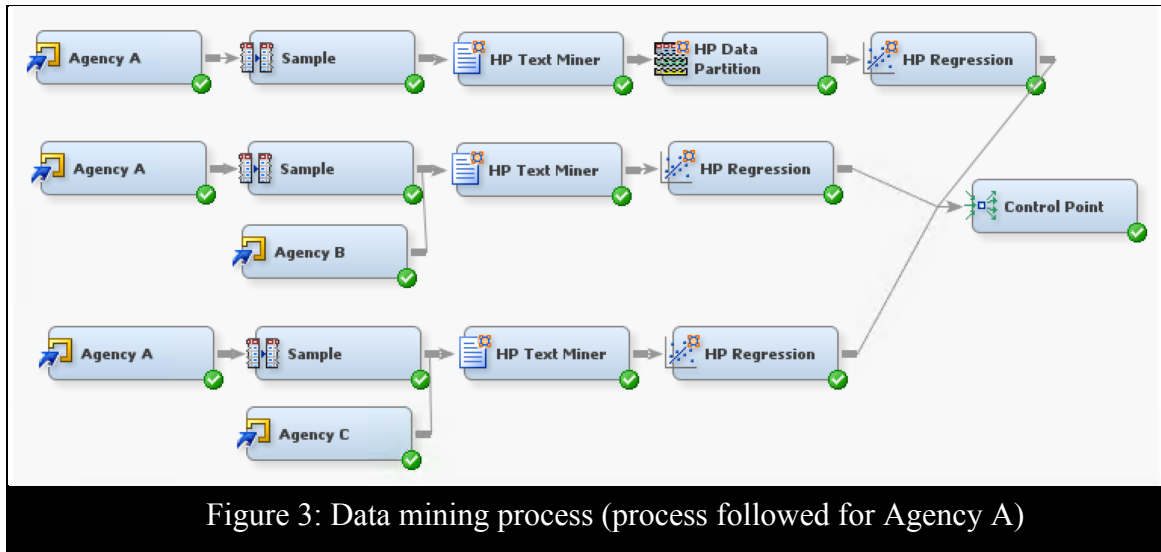


Figure 3: Data mining process (process followed for Agency A)

We create individual models for each organizational unit (Agency A, B, and C) and we evaluate each within its own organizational unit (intra) and across organizational units (inter)(see Figure 4). We first filter out case notes by organizational unit (e.g., Agency A, Agency B, and Agency C). We follow the same process shown in Figure 3 for Agency B and Agency C. We split the data using a random sample (for each organizational unit) into a training set containing 70% of the cases and a test set containing the remaining 30% of the data.

Using SAS Text Miner 9.4 (as shown in Figure 3), we evaluate the performance of these models and all the permutation comparisons across organizational units. We use a z-test for proportions for precision and recall as a mechanism for statistically comparing results from the different models (Kachigan, 1986; Adomavicius et al., 2005).

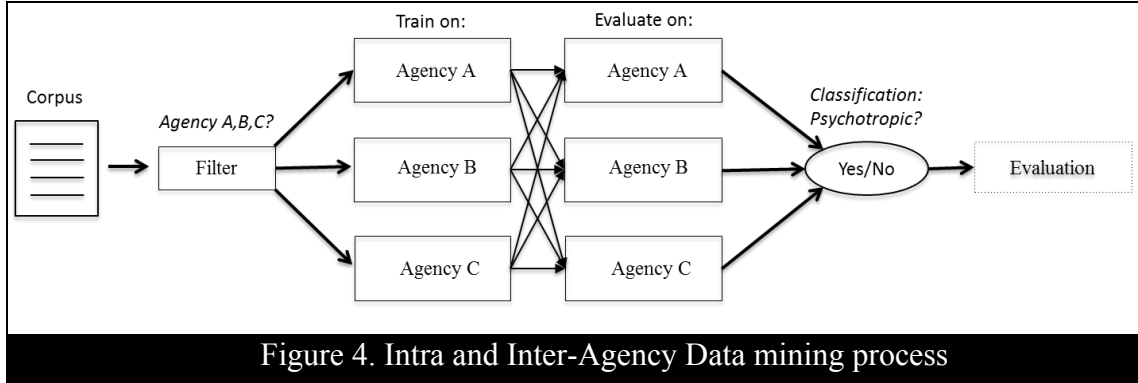


Figure 4. Intra and Inter-Agency Data mining process

The precision, recall, and F-measure metrics reflect the performance of the classifier on the binary outcome (e.g., classifying an individual as a psychotropic medication user or not). In this context, due to the negative consequences of not identifying a positive case, we consider a better predictive model to be one that has a higher recall –minimizing the number of false negatives (e.g., classifying children as not taking psychotropic medication when in fact they do take medication). The z-test for proportions evaluates the statistical difference between two population proportions p_1 and p_2 (Kachigan, 1986; Fleiss et al., 2013). To test the difference between proportions we compute the following:

$$z_{proportions} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

In Table 6, we evaluate each Agency by comparing the performance when tested with data from the same organizational unit (intra-agency) and across organizational units (inter-agency). We highlight in bold any statistically significant differences for precision and recall using a z-test for proportions (two-tailed test at the 95% confidence level). There is no standard definition of what a substantial difference in F-measure

improvement should be. In the field of information retrieval a 5% performance improvement is considered a substantial improvement (Adomavicius et al., 2005; Sparck Jones, 1974). The difference in F-measure is substantial if the difference between F-measures is more than 0.05 and the difference in precision or recall is statistically significant (determined using the z-test for proportions and highlighted in bold and with a * symbol)(Adomavicius et al., 2005).

Table 6. Results in difference between proportions for Precision (P) and Recall (R)				
Train	Evaluation	Precision	Recall	F-Measure
Agency A	Agency A	78.57	70.97	74.58
	Agency B	65	52.7	58.21
	Agency C	31.94	30.67	31.29*
	Z-Value (Agency A-Agency B)	1.2858	1.7303	
	Z-Value (Agency A-Agency C)	4.2082 (p<0.01)	3.9911 (p<0.01)	
Agency B	Agency B	46.15	54.54	50
	Agency A	45.59	30.69	36.69*
	Agency C	32	21.33	25.6*
	Z-Value (Agency B-Agency A)	0.1377	2.1261 (p<0.05)	
	Z-Value (Agency B-Agency C)	1.1864	3.0229 (p<0.01)	
Agency C	Agency C	64.71	50	56.41
	Agency A	33.33	16.83	22.37*
	Agency B	59.26	21.62	31.68*
	Z-Value (Agency C-Agency A)	2.2762 (p<0.05)	3.3621 (p<0.01)	
	Z-Value (Agency C-Agency B)	0.3613	2.5992 (p<0.01)	

The differences in F-measure are substantial in five out of the six pairs. The results of the analysis show that two of the agencies (Agency A and Agency C) consistently perform better in classifying cases of psychotropic drug use. Our initial explanation for

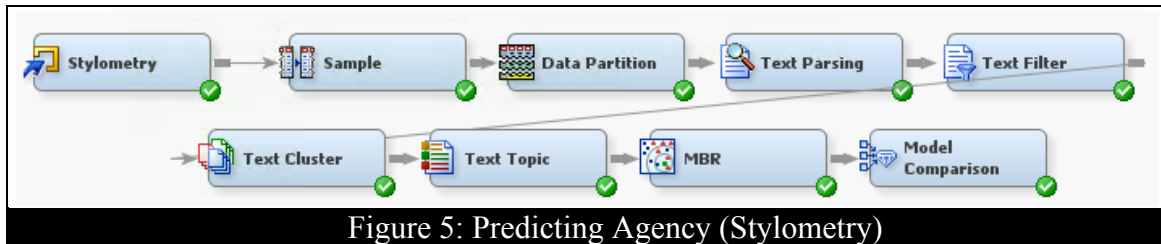
this phenomenon is that, despite the isomorphism directed by regulative, normative, and cognitive components, different agencies may have adopted institutional elements that may influence the information being recorded in the home-visit notes. In addition to what caseworkers are required to document, an organizational unit may have requested its caseworkers to include additional information that can be specific to a particular need (e.g., monitoring psychotropic drug use).

It is reasonable that the best performing model for each organizational unit is a model trained with data from that same organizational unit. In the next section we analyze whether different organizational units are consistent in the way they encode home-visit notes. In other words, based on the content of a particular case note, can we predict to which agency that particular case note belongs? By doing so, we can assess what it is in the content of these case notes that makes them more amenable to solving a specific tactical need effectively.

Proposition 2: Institutional factors establish data entry practices that result in data that is highly cohesive (similar within the same organizational unit) and loosely coupled (different across organizational units).

Stylometric analysis is simply an application of text mining that uncovers metadata from the documents and allows for statistical comparisons of these metadata as a proxy for “style”. Using statistical text mining software (SAS Text Miner 9.4), we predict, based on the text in the case note, to which agency a particular case note belongs. Our training set consists of all the case notes from the three agencies assigned to a mutually exclusive train and test set. We follow the data mining process shown in Figure 5 for the

predictive modeling. We train a classification model that has the case note text and our target variable—the agency from which that note is coming from. This target variable takes one of three levels: Agency A, Agency B, and Agency C.



The results show that we could classify case notes and attribute to which agency they belong to with a high degree of certainty (see Table 7). These results show that each organization has their own style, which is consistently used by its caseworkers. Based on the results from the psychotropic models and the results of the stylometry analysis, we could argue that the structure of these notes is similar within organizational units (highly cohesive) and different from that of other organizational units (loosely coupled)—based on the ability of the predictive model to discriminate, with high degree of certainty, the authoring agency of a particular case note.

Table 7. Case Distribution across Agencies			
Agency	Precision (%)	Recall (%)	F-measure (%)
Agency A	76.99	75.65	76.31
Agency B	79.81	76.15	77.94
Agency C	76.74	81.15	78.88

Some researchers have argued that an author’s style is comprised of a limited number of distinctive features inherent to the author, neglecting the content/context-

dependency of the writing (De Vel et al., 2001). In our study, these notes are all in the same context, reflect a specific aspect of case management, and are written by experienced staff acting as representatives of these agencies. If we combine the results of the analysis for proposition one and those of proposition two, we could argue that perhaps there is inherent features that are included in one of the agency's case notes that others may be lacking and vice versa. Could we identify best reporting practices that help solve a tactical need effectively?

Proposition 3: Unstructured data formats can help shape effective practices in solving well-defined needs.

We turn to psychology research to illustrate how despite subtle differences in human language, there are unquantifiable yet salient qualities, such as specificity, that can provide inference. Psychology research reveals a tradeoff between *cognitive economy* and *inductive inference* (Lakoff, 1987; Roach et al., 1978; E.E. Smith and Medin, 1981; Edward E Smith, 1988; Parsons, 1996). Categories improve the ability of a person to accurately predict features of instances of a category. Some members of a category are more central than others and different concepts have a degree of membership to some of these categories.

Proposition 3a: Higher levels of specificity in the data collected leads to increased inferential utility.

Computers understand very little of the meaning of human language. Information retrieval research has assumed the meaning of words is closely connected to the statistics

of word usage (Turney and Pantel, 2010). For instance, (Sparck Jones, 1972) adopted a statistical interpretation of the concept of specificity as a function of term use rather than having to do with the accuracy of the concept representation. This measure of term specificity is what later became the concept of inverse document frequency (idf) in information retrieval research. Landauer (2002) estimates that 80% of the meaning of a passage comes from word choice and the remaining 20% comes from word order.

From a quantitative approach, we assess language use (in terms of structure and meaning of the case notes) by including/excluding NLP features. We then assess whether there are any performance differences in the prediction accuracy of the models. For the text analysis we use SAS Text Miner 9.4, which has built-in text parsing and text filtering features that use natural language processing (NLP). The results are shown in Table 8. We present three different models: The first one without removing any NLP features and using a mutual information weighting scheme, a second one without part-of-speech (POS) and noun group (NG) features, and a third one only with POS and NG features but no term weighting scheme. We evaluate the performance of these models by evaluating their precision, recall, and F-measure (see Table 8). The precision, recall, and F-measure metrics reflect the performance of the classifier on the nominal outcome –agency to which a particular case note belongs to).

Table 8. Prediction results with features disabled			
Features Used in Model	Precision (%)	Recall (%)	F-measure (%)
POS, NG, TF, TW	76.15	79.81	77.94
TF + TW	70.94	79.81	75.11
POS + NG	60.53	66.35	63.31

Similarly to the analysis in proposition one, we use a z-test for proportions for precision and recall as a mechanism for statistically comparing results from the different models (see Table 9). The difference in F-measure is substantial if the difference between F-measures is more than 0.05 and the difference in precision or recall is statistically significant (determined using the z-test for proportions and highlighted in bold and with a * symbol)(Adomavicius et al., 2005).

Table 9. Results in difference between proportions for Precision (P) and Recall (R)			
Components	Precision	Recall	F-measure
POS, NG, TF, TW	76.15	79.81	77.94
No POS, No NG, TF, TW	70.94	79.81	75.11
Z-values	0.8857	0	
POS, NG, No TF, No TW	60.53	66.35	63.31*
Z-values	2.503	2.1885	
Scores are z-values for Precision (P) and Recall (R). Significance values are $p < 0.05^*$ (two-tailed)			

The results in Table 9 show a statistical significant difference between the full model (one that contains NLP features and a term weighting scheme) and the model that only has a weighting scheme (TF, TW) but no NLP features. Results also show that there is no statistical significant difference between the full model and the model that has no POS and NG features but does have a term weighting scheme (TF, TW). Consistent with previous research, the terms used are a more salient factor of prediction compared to the language structure of a case note. Institutional factors can provide two plausible explanations for this: (1) different organizational units focus on different aspects when reporting a home-visit and (2) the depth at which they encode their notes can be more general/specific.

In the next section we provide an example of an application of the generalization/specification concept in effective data-entry practices.

***Proposition 3b:** Higher levels of specificity in the data collected facilitate unanticipated use of the data.*

Human language is subtle, with many unquantifiable yet salient qualities. Users with different levels of expertise tend to produce information that differs in quality and level of abstraction. For example, within the category “taking medication”, a concept hierarchy can be the following: (a) medication (b) psychotropic medication (c) Lisdexamfetamine (d) Vyvanse, which goes from the most general (a) to the most specific (d). Knowing a child is taking Vyvanse (d) gives more information than just knowing a child is taking medication (a). Based on the results in the previous sections and in line with the literature on cognitive psychology (e.g., specifically on categorization and inference), we could argue that text with higher content specificity could be then abstracted for use in unanticipated applications. For an application such as psychotropic drug use monitoring (one for which the notes were not originally intended for) we could use the concept hierarchy introduced before as a qualitative mean to identify potential cases of psychotropic medication-use. In Table 10 we show fragments of home-visit notes from different agencies. These three case notes were cases in which the child was taking psychotropic medication. If an individual were to rank these based on the likelihood of being a case of psychotropic medication-use, which one would rank first? We argue based on psychology research that the higher the specificity, the higher the inferential utility. Knowing a child is taking 40 mgs of Vyvanse to cope with ADHD provides more

information than knowing the child is taking its medication—since the more general category medication also includes non-psychotropic medication.

Table 10. Home-visit notes of children taking psychotropic medication

Home-Visit Note	Predicted Target	Actual Target (Taking Psychotropic Medication)	Specificity
<p>regional supervisor met with the youth's mother and the youth at the paternal grandmother's house to go over paws sexual abuse psycho educational materials to document the youth's current medications and to take a photograph of the youth to promote a safe home environment... the youth expressed that she has been feeling better and that her mom has been spending more time with her the is then asked the youth how have things been going at school since she elicited the self harming behaviors the prior week the youth responded and told the regional supervisor that she has been doing better... then asked the youth if she was currently taking any medications the youth replied and said that she is currently taking two medications but that she saw her doctor yesterday to follow up about the self harming behavior she had elicited the week prior and prescribed her two new medications in addition to the two she is already taking the is then proceeded to ask if it was okay if he could please see her current medications so that he may document it for the youth's chart the youth and the mother were okay with this and the is proceeded to document the information on each medication... cm observed the children to be free of any visible marks or bruises the home was neat and clean and free of any hazardous ...</p>	1	1	More general - Medication
<p>was appropriately dressed she is in a licensed mental facility citrus no signs of abuse or neglect at the time of the visit he has a normal relationship with the staff and with this counselor the facility was clean and organized with a lot of restriction access...has no medical concerns but he is in a mental health facility due his diagnosis and behavior citrus educational evaluation individualized education plan yes not extracurricular activities child is in a mental health facility he receive visits from his adoptive father every weekend not safety plan he is in a mental health facility child request a lot of belongings from the house he wants shoes more clothes staff states he is improving but he has some behavior issues overall caregiver comments and questions staff says he is doing well he improve from the last visit worker comments...looks stable but he does not want to receive visits from his mother... house was clean organized and no hazards was visible no medical concerns he receive individual and cba services...individualized education plan yes not age appropriate not age appropriate he has therapeutic visits with his parents twice a week no safety plan child states he is doing well and no complaints caregiver states he is doing well but sometimes has problems with his sister... caregiver states the children needs to be constantly monitore</p>	1	1	Less specific - Mental health facility
<p>child was seen face to face at the hometoday he was observed with weather appropriate clothings the child was observed to be free of abuse and neglect as evidenced of no visible marks or bruises on his body when this ca arrived in the home the child's therapist from... child continues to receive behavioral services... child continues to be seen at...for medication managements the home was observed to be free of hazards the child's bedroom was observed clean as well the child continues to take the following medications vyvanse mg po qam for adhd singular mg po q day for lung diseaseenalapril mg po qdayvisit date begin time pment time pmthe caretaker child have not been notified of the date time location and type of next court hearing... copy insurance card court order parental consent for treatment shelter order copy disposition ... child's appearance well groomed in age appropriate and well fitting clothingthe child is free of visible bruises injuries and or abrasions since the last visit the child has not been a victim of a reported abuse and or neglect the child has alerts the alert codes are c d there are no persons years or older residing in the home who has not had a background screening physical condition of placement free of visible hazardsafter reviewing all of the information regarding the safety of this child in this placement the risk of harm is rated as low...child's behavior age appropriate for teenager years</p>	1	1	More specific - Taking mg of Vyvanase

Discussion

Motivated by the ever increasing growth of unstructured data in organizational settings in this paper we address (1) the implications of *unstructured-data-entry* in the data collected by an organization; (2) how it helps in solving a tactical need; and (3) effective *data-entry* strategies. Traditional IS research offers limited guidance in the effect of different data-entry practices and decision-making (e.g., alignment between the information needs of data consumers and data contributors or promoting effective data-entry practices).

Organizations collect data to solve tactical needs'. The data stored from daily transactions supports effective decision-making. These data may be encoded in free-text, which may hinder the organization from effectively using it—due to the inherent flexible structure of free-text. Nevertheless, trying to impose too much structure (e.g., guiding templates) may cause an unintentional focus on what needs to be recorded that may result in an omission of potentially interesting information. Future research should focus on finding what the optimal level on this dichotomy is. As reported in our case study, allowing some degree of freedom can prove beneficial in solving tactical needs if effective data collection strategies are put in place by the organization. By adopting such practices, organizations can leverage on their data to solve needs that may have not been anticipated at the time of the system's development. Moreover, it would allow the organization to adapt such information to a different context—a limitation of fixed schemas.

In this paper we do not argue in favor of unstructured notes over structured notes. However, we argue that for certain applications, although structured information has the advantage of consistency and ability for integration, it may hinder user input.

Future studies should focus in analyzing the importance of a good system design (e.g. structured vs. free-flow). This complements research by Lukyanenko et al. (2014) which argues that by limiting data-entry to experts in a citizen science project (e.g., data input by users at the species level) it can preclude the input of valuable information from non-experts and can lead to data accuracy problems (e.g., non-experts trying to “guess” species-level attributes). Our research encourage experts to be as specific as they can while allowing non-experts to input information at a more basic-level.

We found that unstructured data entry may result in differences in how information is collected across different organizational units in the organization. Institutional theory helps explain how institutional factors shape practices by individuals across organizational units, and how these practices can become stable over time and adopted by other individuals, making the practice persistent. The analysis of the data showed that data collected through free-text formats become isomorphic *within* organizational units and that individuals from different organizational units adopt consistent practices when entering free-text-notes into the information system. Effectively, institutional factors shape data-entry practices that result in highly cohesive and loosely coupled data collection.

We illustrate the impact data-entry formats using a case study in the context of case management in a foster care organization. Our tactical purpose is to monitor cases of psychotropic medication use. From the analysis of the data, we found that higher levels of specificity in the data collected leads to increased inferential utility, which can ultimately help the organization solve unanticipated tasks using these data. As shown in proposition 1, treating all data in the aggregate can have a detrimental effect in the performance of predictive models. Future work should focus in providing a method to evaluate when using data in the aggregate is justified as opposed to highlighting meaningful segments for separate analysis. In this study, we use organizational units as boundaries but a generalizable approach should be able to inductively select what these segments should be. We also introduce the idea of organizational stylometry. To our knowledge, the use of stylometry at the population level has yet to be explored—in which there are many contributors to a body of text. Our research objective was to show that despite the fact that organizations have established guidelines of reporting, employees adopt new guidelines that become established over time.

The results of this study can be generalized to other domains and can provide insight to effective system design—the effect of particular designs (that are more/less flexible). Moreover, we hope to increase understanding of the implications of free-text-data-entry strategies and its implications to solving tactical needs’. A practical implication is that depending on whether the individuals looking at the text is a non-experts vs. expert, the individual writing the text can choose to contribute beyond what he believes is the information required for the reader. This allows for increased inferential utility that can

prove beneficial when dealing with unanticipated use of the data. Higher specificity, however, requires higher expertise. Thus, requiring higher levels of specificity when capturing data may hinder collaboration from non-experts.

REFERENCES

- Abbasi, A., & Chen, H. (2008). CyberGate: a design framework and system for text analysis of computer-mediated communication. *Mis Quarterly*, 811-837.
- Abiteboul, S. (1997). *Querying semi-structured data*. Springer.
- Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)*, 23(1), 103-145.
- Anderson, D. G. (2000). Coping strategies and burnout among veteran child protection workers. *Child abuse & neglect*, 24(6), 839-848.
- Anderson, J. R. (1990). *Cognitive psychology and its implications*. WH Freeman/Times Books/Henry Holt & Co.
- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409.
- Anderson, J. R., & Matessa, M. A rational analysis of categorization. In *Proceedings of the Second International Conference on Machine Learning, 2014* (pp. 76-84)
- Anderson, P. A. (1983). Decision making by objection and the Cuban missile crisis. *Administrative Science Quarterly*, 201-222.
- Anglin, J. M. (1977). *Word, object, and conceptual development* (Vol. Accessed from <http://nla.gov.au/nla.cat-vn2178543>). Norton, New York.
- Angst, C. M., Agarwal, R., Sambamurthy, V., & Kelley, K. (2010). Social contagion and information technology diffusion: the adoption of electronic medical records in US hospitals. *Management Science*, 56(8), 1219-1241.
- Appan, R., & Browne, G. J. (2010). Investigating Retrieval-Induced Forgetting During Information Requirements Determination. *Journal of the Association for Information Systems*, 11(5), 250.

- Barbell, K., & Freundlich, M. (2001). *Foster care today*. Citeseer.
- Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative science quarterly*, 78-108.
- Barley, S. R., & Tolbert, P. S. (1997). Institutionalization and structuration: Studying the links between action and institution. *Organization studies*, 18(1), 93-117.
- Barr, R., & Caplan, L. (1987). Category representations and their implications for category structure. *Memory & Cognition*, 15(5), 397-418, doi:10.3758/bf03197730.
- Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS quarterly*, 369-386.
- Berg, M. (2001). Implementing information systems in health care organizations: myths and challenges. *International journal of medical informatics*, 64(2), 143-156.
- Berg, M., & Goorman, E. (1999). The contextual nature of medical information. *International journal of medical informatics*, 56(1), 51-60.
- Berger, P. L., & Luckmann, T. (1991). *The social construction of reality: A treatise in the sociology of knowledge* (Vol. 10). Penguin UK.
- Berlin, B. (2014). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton University Press.
- Berlin, B., Breedlove, D. E., & Raven, P. H. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist*, 75(1), 214-242.
- Berry, M. W., & Castellanos, M. (2004). Survey of text mining. *Computing Reviews*, 45(9), 548.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. "O'Reilly Media, Inc."

- Bodart, F., Patel, A., Sim, M., & Weber, R. (2001). Should optional properties be used in conceptual modelling? A theory and three empirical tests. *Information Systems Research*, 12(4), 384-405.
- Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., et al. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11), 977-984.
- Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., et al. (2014). Next steps for citizen science. *Science*, 343(6178), 1436-1437.
- Boster, J. S. (1986). Exchange of varieties and information between Aguaruna manioc cultivators. *American Anthropologist*, 88(2), 428-436.
- Boudreau, M.-C., & Robey, D. (2005). Enacting integrated information technology: A human agency perspective. *Organization science*, 16(1), 3-18.
- Bowers, J. S., & Jones, K. W. (2008). Detecting objects is easier than categorizing them. [Article]. *Quarterly Journal of Experimental Psychology*, 61(4), 552-557, doi:10.1080/17470210701798290.
- Brown, R. (1958). How shall a thing be called? *Psychological review*, 65(1), 14.
- Browne, G. J., & Ramesh, V. (2002). Improving information requirements determination: A cognitive perspective. *Information & Management*, 39(8), 625-645.
- Buneman, P. Semistructured data. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, 1997* (pp. 117-121): ACM
- Bunge, M. A. (1977). *Treatise on Basic Philosophy: Ontology I: The Furniture of the World*. Dordrecht, Holland: D. Reidel Publishing Company.
- Burton-Jones, A., & Grange, C. (2012). From use to effective use: A representation theory perspective. *Information Systems Research*, 24(3), 632-658.
- Burton-Jones, A., & Meso, P. N. (2008). The effects of decomposition quality and

multiple forms of information on novices' understanding of a domain from a conceptual model. *Journal of the Association for Information Systems*, 9(12), 748.

Burton-Jones, A., Wand, Y., & Weber, R. (2009). Guidelines for empirical evaluations of conceptual modeling grammars. *Journal of the Association for Information Systems*, 10(6), 495.

Cantor, N., & Mischel, W. (1979). *Prototypes in person perception* (Vol. 12, *Advances in Experimental Social Psychology*). Academic Press.

Cantor, N., Smith, E. E., French, R. D., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology*, 89(2), 181.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., & Moon, S. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007* (pp. 1-14): ACM

Chen, P. P. (2006). Suggested research directions for a new frontier—active conceptual modeling. In *Conceptual Modeling-ER 2006* (pp. 1-4). Springer.

Chen, P. P.-S. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1), 9-36.

Chetty, S. (1996). The case study method for research in small-and medium-sized firms. *International small business journal*, 15(1), 73-86.

Clark, E. V. (1979). Building a vocabulary: Words for objects, actions and relations. *Language Acquisition*, 149-160.

Clow, D., & Makriyannis, E. iSpot analysed: Participatory learning and reputation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge, 2011* (pp. 34-43): ACM

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies*, 41(6), 801-849.

- Corter, J., & Gluck, M. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*(2), 291-303, doi:citeulike-article-id:1116860.
- Craig, C. G. (1986). *Noun classes and categorization* (Vol. 7, Proceedings of a symposium on categorization and noun classification, Eugene, Oregon, October 1983). John Benjamins Publishing Company.
- Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, *4*(6), 433-442.
- Cruse, D. A. (1977). The pragmatics of lexical specificity. *Journal of linguistics*, *13*(02), 153-164.
- Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business Press.
- Davis, G. B., & Olson, M. H. (1984). *Management information systems: conceptual foundations, structure, and development*. McGraw-Hill, Inc.
- de Beeck, H. O., & Wagemans, J. (2001). Visual object categorisation at distinct levels of abstraction: A new stimulus set. *Perception*, *30*(11), 1337-1361.
- De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, *30*(4), 55-64.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.
- DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization science*, *5*(2), 121-147.
- DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review*, 147-160.

- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.
- Dubé, L., & Paré, G. (2003). Rigor in information systems positivist case research: current practices, trends, and recommendations. *MIS quarterly*, 597-636.
- Dubin, R. (1970). Theory building.
- Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems, 1988* (pp. 281-285): ACM
- Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of management review*, 14(4), 532-550.
- Feldman, M. S., & March, J. G. (1981). Information in organizations as signal and symbol. *Administrative science quarterly*, 171-186.
- Feldman, R., & Dagan, I. Knowledge Discovery in Textual Databases (KDT). In *KDD, 1995* (Vol. 95, pp. 112-117)
- Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.
- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Clarendon Press.
- Forsyth, R. S. (1999). New directions in text categorization. In *Causal models and intelligent data management* (pp. 151-185). Springer.
- Fry, J. P., & Sibley, E. H. (1976). Evolution of data-base management systems. *ACM Computing Surveys (CSUR)*, 8(1), 7-42.

- Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., et al. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, 1988* (pp. 465-480): ACM
- Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iView, 1142*, 1-12.
- GAO (2011). HHS Guidance Could Help States Improve Oversight of Psychotropic Prescriptions. *Foster Children*: U.S. Government Accountability Office.
- Gemino, A., & Wand, Y. (2003). Evaluating modeling techniques based on models of learning. *Communications of the ACM, 46*(10), 79-84.
- Gemino, A., & Wand, Y. (2004). A framework for empirical evaluation of conceptual modeling techniques. *Requirements Engineering, 9*(4), 248-260, doi:10.1007/s00766-004-0204-6.
- Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence, 40*(1-3), 11-61, doi:10.1016/0004-3702(89)90046-5.
- Gersick, C. J. (1988). Time and transition in work teams: Toward a new model of group development. *Academy of Management journal, 31*(1), 9-41.
- Giddens, A. (1979). *Central problems in social theory: Action, structure, and contradiction in social analysis* (Vol. 241). Univ of California Press.
- Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Univ of California Press.
- Gould, J. D., & Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM, 28*(3), 300-311, doi:10.1145/3166.3170.
- Gregor, S. (2006). The nature of theory in information systems. *MIS quarterly, 611-642*.

- Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS quarterly*, 37(2), 337-355.
- Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems*, 8(5), 312.
- Grill-Spector, K., & Kanwisher, N. (2005). Visual Recognition. *Psychological Science*, 16(2), 152-160.
- Grossman, M., Aronson, J. E., & McCarthy, R. V. (2005). Does UML make the grade? Insights from the software development community. *Information and Software Technology*, 47(6), 383-397.
- Group, G. M. W. (2009). PSYCHOTROPIC MEDICATION: For Children in Out of Home Care Business Rules with Data Entry Guidelines and Frequently Asked Questions (DCF, Trans.). *Special Initiatives*: Florida Department of Children and Families.
- Gumm, D. C. (2006). Distributed Participatory Design: An Inherent Paradoxon? *Proc. of IRIS29*.
- Halfon, N., & Klee, L. (1991). Health and development services for children with multiple needs: The child in foster care. *Yale Law & Policy Review*, 9(1), 71-96.
- Halpin, T. (2007). Fact-oriented modeling: Past, present and future. In *Conceptual Modelling in Information Systems Engineering* (pp. 19-38). Springer.
- Hand, E. (2010). People power. *Nature*, 466(7307), 685-687.
- Hannan, M. T., & Freeman, J. (1977). The population ecology of organizations. *American journal of sociology*, 929-964.
- Harman, J. S., Childs, G. E., & Kelleher, K. J. (2000). Mental health care utilization and expenditures by children in foster care. *Archives of Pediatrics & Adolescent Medicine*, 154(11), 1114-1117, doi:10.1001/archpedi.154.11.1114.

- Harper, M., & Schoeman, W. J. (2003). Influences of gender as a basic-level category in person perception on the gender belief system. *Sex Roles, 49*(9-10), 517-526.
- Harris, S. G., & Sutton, R. I. (1986). Functions of parting ceremonies in dying organizations. *Academy of Management journal, 29*(1), 5-30.
- Hearst, M. A. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999* (pp. 3-10): Association for Computational Linguistics
- Hirschheim, R., Klein, H. K., & Lyytinen, K. (1995). *Information systems development and data modeling: Conceptual and philosophical foundations*. Cambridge University Press.
- Hofstaetter, C., Hofstaetter, C., Hansmann, M., Eik-Nes, S. H., Huhta, J. C., & Luther, S. L. (2006). A cardiovascular profile score in the surveillance of fetal hydrops. *The Journal of Maternal-Fetal & Neonatal Medicine, 19*(7), 407-413.
- Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and linguistic computing, 13*(3), 111-117.
- Hughes, E. C. (1936). The ecological aspect of institutions. *American sociological review, 1*(2), 180-189.
- Iyer, S. V., Harpaz, R., LePendur, P., Bauer-Mehren, A., & Shah, N. H. (2014). Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association, 21*(2), 353-362.
- Jacobson, I., Booch, G., Rumbaugh, J., Rumbaugh, J., & Booch, G. (1999). *The unified software development process* (Vol. 1). Addison-Wesley Reading.
- Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General, 126*(3), 248.
- Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. [doi: DOI: 10.1016/0010-0285(84)90009-4]. *Cognitive Psychology, 16*(2), 243-275.

- Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin*, 94(3), 423-428, doi:10.1037/0033-2909.94.3.423.
- Kachigan, S. K. (1986). *Statistical analysis: An interdisciplinary introduction to univariate & multivariate methods*. Radius Press.
- Kaufman, K. (1999). *Advanced birding* (Vol. 39). Houghton Mifflin Harcourt.
- Kent, W. (1978). *Data and reality: Basic assumptions in data processing reconsidered*. Elsevier Science Inc.
- Kidder, T. (2011). *The soul of a new machine*. Back Bay Books.
- Klibanoff, R. S., & Waxman, S. R. (2000). Basic Level Object Categories Support the Acquisition of Novel Adjectives: Evidence from Preschool - Aged Children. *Child Development*, 71(3), 649-659.
- Kung, C., & Soelberg, A. Activity modeling and behavior modeling. In *Proc. of the IFIP WG 8.1 working conference on Information systems design methodologies: improving the practice, 1986* (pp. 145-171): North-Holland Publishing Co.
- Lakoff, G. (1987). *Women, fire, and dangerous things*. University of Chicago Press, Chicago.
- Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.
- Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4), 721-735.
- Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *Psychology of learning and motivation*, 41, 43-84.
- Larsen, K., & Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and metaanalyses. *MIS Quarterly*.

- Lassaline, M. E., Wisniewski, E. J., & Medin, D. L. (1992). Basic Levels in Artificial and Natural Categories: Are All Basic Levels Created Equal? *Percepts, Concepts and Categories: The Representation and Processing of Information*, 327.
- Lee, A. S. (1989). A scientific methodology for MIS case studies. *MIS quarterly*, 33-50.
- Lerch, F. J., & Harter, D. E. (2001). Cognitive support for real-time dynamic decision making. *Information Systems Research*, 12(1), 63-82.
- Levina, N., & Arriaga, M. (2014). Distinction and status production on user-generated content platforms: Using bourdieu's theory of cultural production to understand social dynamics in online fields. *Information Systems Research*, 25(3), 468-488.
- Lukyanenko, R. (2014). *An information modeling approach to improve quality of user-generated content*. Memorial University of Newfoundland,
- Lukyanenko, R., & Parsons, J. (2013a). Is traditional conceptual modeling becoming obsolete? In *Conceptual Modeling* (pp. 61-73). Springer.
- Lukyanenko, R., & Parsons, J. (2013b). Lightweight Conceptual Modeling for Crowdsourcing. In *Conceptual Modeling* (pp. 508-511). Springer.
- Lukyanenko, R., & Parsons, J. (2013c). Reconciling theories with design choices in design science research. In *Design Science at the Intersection of Physical and Virtual Design* (pp. 165-180). Springer.
- Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2014). The iq of the crowd: Understanding and improving information quality in structured user-generated content. *Information Systems Research*, 25(4), 669-689.
- Luther, S., Berndt, D., Finch, D., Richardson, M., Hickling, E., & Hickam, D. (2011). Using statistical text mining to supplement the development of an ontology. *Journal of Biomedical Informatics*, 44, S86-S93.
- Macé, M. J.-M., Joubert, O. R., Nespoulous, J.-L., & Fabre-Thorpe, M. (2009). The time-course of visual categorizations: You spot the animal faster than the bird. *PloS one*, 4(6), e5927.

- Mandler, J. M., & Bauer, P. J. (1988). The cradle of categorization: Is the basic level basic? *Cognitive development*, 3(3), 247-264.
- Manning, C. D., Raghavan, P., Schütze, H., & Elvén, A. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.
- March, S., Hevner, A., & Ram, S. (2000). Research commentary: an agenda for information technology research in heterogeneous and distributed environments. *Information Systems Research*, 11(4), 327-341.
- March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, 15(4), 251-266.
- Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. MIT Press.
- After 7-Year-Old Gabriel Myers' Suicide, Fla. Bill Looks To Tighten Access To Psychiatric Drugs. (2010, 03/17). *CBS News*.
- Mason, R. O. (1978). Measuring information output: A communication systems approach. *Information & management*, 1(4), 219-234.
- Mayden, R. L. (2002). On biological species, species concepts and individuation in the natural world. *Fish and Fisheries*, 3(3), 171-196.
- McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 6(4), 462-472.
- McGinnes, S. (2011). Conceptual modelling for web information systems: What semantics can be shared? In *Advances in Conceptual Modeling. Recent Developments and New Directions* (pp. 4-13). Springer.
- Mervis, C. B., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, 53(3), 258-266.

- Mervis, C. B., Golinkoff, R. M., & Bertrand, J. (1994). Two-Year-Olds Readily Learn Multiple Labels for the Same Basic-Level Category. [Article]. *Child Development*, 65(4), 1163-1177, doi:10.1111/1467-8624.ep7252870.
- Mervis, C. B., & Rosch, E. (1981). Categorization of Natural Objects. *Annual Review of Psychology*, 32, 89-115.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American journal of sociology*, 340-363.
- Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35, 128-144.
- Miller, C. M. (4/18/2010). Red Flags Overlooked In Prescription Drug Death Of 12-Year-Old. <http://www.psychsearch.net/red-flags-overlooked-in-prescription-drug-death-of-12-year-old/>.
- Moody, D. L. (2005). Theoretical and practical issues in evaluating the quality of conceptual models: Current state and future directions. *Data & Knowledge Engineering*, 55(3), 243-276, doi:10.1016/j.datak.2004.12.005.
- Mosteller, F., & Wallace, D. (1964). Inference and disputed authorship: The Federalist.
- Murphy, G. L. (1982). Cue validity and levels of categorization. *Psychological Bulletin*, 91(1), 174-177, doi:10.1037/0033-2909.91.1.174.
- Murphy, G. L. (1996). On metaphoric representation. *Cognition*, 60(2), 173-204.
- Murphy, G. L. (2004). *The big book of concepts*. MIT Press.
- Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(1), 70.
- Murphy, G. L., & Wisniewski, E. J. (1989). Categorizing Objects in Isolation and in Scenes - What a Superordinate Is Good For. *Journal of Experimental Psychology-*

Learning Memory and Cognition, 15(4), 572-586.

Mylopoulos, J. (1992). Conceptual Modeling and Telos. *Conceptual Modeling, Databases, and CASE. An Integrated View of Information Systems*. P. Loucopoulos and R. Zicari. John Wiley & Sons.

Mylopoulos, J. (1998). Information modeling in the time of the revolution. *Information Systems*, 23(3-4), 127-155, doi:10.1016/s0306-4379(98)00005-2.

Nakamura, G. V., Medin, D. L., & Taraban, R. (1993). *Categorización by Humans and Machines*. Academic Press.

Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., et al. (2008). Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1), 32.

Neisser, U. (1987). From direct perception to conceptual structure.

O'Reilly, C. A. (1982). Variations in decision makers' use of information sources: The impact of quality and accessibility of information. *Academy of Management journal*, 25(4), 756-771.

Office, U. S. G. A. (2003). HHS could play a greater role in helping child welfare agencies recruit and retain staff. *CHILD WELFARE*.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 17(4), 273-281.

Op de Beeck, H., & Wagemans, J. (2001). Visual object categorisation at distinct levels of abstraction: A new stimulus set. *Perception*, 30(11), 1337-1361.

Orlikowski, W. J., & Barley, S. R. (2001). Technology and institutions: what can research on information technology and research on organizations learn from each other? *MIS quarterly*, 25(2), 145-165.

Ozcan, E., van Egmond, R., & Jacobs, J. (2014). Product sounds: Basic concepts and categories. *International Journal of Design*, 8(3), 97-111.

- Parsons, J. (1996). An information model based on classification theory. *Management Science*, 42(10), 1437-1453.
- Parsons, J. (2002). Effects of local versus global schema diagrams on verification and communication in conceptual data modeling. *Journal of Management Information Systems*, 19(3), 155-183.
- Parsons, J. (2011). An Experimental Study of the Effects of Representing Property Precedence on the Comprehension of Conceptual Schemas. *Journal of the Association for Information Systems*, 12(6), 401.
- Parsons, J., & Cole, L. (2005). What do the pictures mean? Guidelines for experimental evaluation of representation fidelity in diagrammatical conceptual modeling techniques. [doi: 10.1016/j.datak.2004.12.008]. *Data & Knowledge Engineering*, 55(3), 327-342.
- Parsons, J., & Wand, Y. (1997). Choosing classes in conceptual modeling. *Communications of the ACM*, 40(6), 63-69.
- Parsons, J., & Wand, Y. (2000). Emancipating instances from the tyranny of classes in information modeling. *ACM Transactions on Database Systems (TODS)*, 25(2), 228-268.
- Parsons, J., & Wand, Y. (2014). A foundation for open information environments.
- Pasha, R., & Golub, J. S. (2013). *Otolaryngology-Head and Neck Surgery: Clinical Reference Guide*. Plural Publishing.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526-3529.
- Powers, D. M. Applications and explanations of Zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning, 1998* (pp. 151-160): Association for Computational Linguistics

- Ramakrishnan, R., & Gehrke, J. (2000). *Database management systems*. Osborne/McGraw-Hill.
- Ramyaa, C. H., & Rasheed, K. Using machine learning techniques for stylometry. In *Proceedings of International Conference on Machine Learning, 2004*
- Raven, P. H., Berlin, B., & Breedlove, D. E. (1971). The origins of taxonomy. *Science*, *174*(4015), 1210-1213.
- Rhemtulla, M., & Hall, D. (2009). Basic-level kinds and object persistence. *Memory & Cognition*, *37*(3), 292-301, doi:10.3758/mc.37.3.292.
- Rifkin, A. (1985). Evidence for a basic level in event taxonomies. *Memory & Cognition*, *13*(6), 538-556.
- Roach, E., Lloyd, B. B., Wiles, J., & Rosch, E. (1978). Principles of categorization.
- Robey, D., & Boudreau, M. C. (1999). Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications. *Information Systems Research*, *10*(2), 167-185, doi:DOI 10.1287/isre.10.2.167.
- Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, *136*(3), 451.
- Rorissa, A. (2008). User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. *Information Processing & Management*, *44*(5), 1741-1753, doi:DOI 10.1016/j.ipm.2008.03.004.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyesbraem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382-439.
- Rossi, M., & Siau, K. (2000). *Information Modeling in the new Millennium*. IGI Global, Hershey, PA.

- Rossiter, D. G., Liu, J., Carlisle, S., & Zhu, A.-X. (2015). Can citizen science assist digital soil mapping? *Geoderma*, 259, 71-80.
- Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., & Lorenzen, W. E. (1991). *Object-oriented modeling and design* (Vol. 199, Vol. 1). Prentice-hall Englewood Cliffs.
- Rycraft, J. R. (1994). The party isn't over: The agency role in the retention of public child welfare caseworkers. *Social Work*, 39(1), 75-80.
- Saaty, T. L. (1988). *What is the analytic hierarchy process?* Springer.
- Saaty, T. L., & Vargas, L. G. (2012). *Models, methods, concepts & applications of the analytic hierarchy process* (Vol. 175). Springer Science & Business Media.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11), 613-620, doi:10.1145/361219.361220.
- Scanlon, E., Woods, W., & Clow, D. (2014). Informal Participation in Science in the UK: Identification, Location and Mobility with iSpot. *Educational Technology & Society*, 17(2), 58-71.
- Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. *The Oxford Handbook of Cognitive Linguistics*, 117-138.
- Schneiderman, J. U. (2003). Health issues of children in foster care. *Contemporary Nurse*, 14(2), 123-128.
- Scott, W. R. (1987). The Adolescence of Institutional Theory. *Administrative Science Quarterly*, 32(4), 493-511, doi:Doi 10.2307/2392880.
- Scott, W. R. (1995). *Institutions and organizations*. Sage Thousand Oaks, CA.

- Scott, W. R. (2008). Approaching adulthood: the maturing of institutional theory. *Theory and society*, 37(5), 427-442.
- Scott, W. W. R. (2013). *Institutions and organizations: Ideas, interests, and identities*. Sage Publications.
- Searing, D. D. (1991). Roles, Rules, and Rationality in the New Institutionalism. *American Political Science Review*, 85(04), 1239-1260.
- Selznick, P. (1984). *Leadership in administration: A sociological interpretation*. Univ of California Press.
- Sewell Jr, W. H. (1992). A theory of structure: Duality, agency, and transformation. *American journal of sociology*, 1-29.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(4), 623-656.
- Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on, 1996* (pp. 336-343): IEEE
- Siau, K., & Wang, Y. (2007). Cognitive evaluation of information modeling methods. *Information and Software Technology*, 49(5), 455-474.
- Silberschatz, A., Korth, H. F., & Sudarshan, S. (1996). Data models. *ACM Computing Surveys (CSUR)*, 28(1), 105-108.
- Silvertown, J. (2010). Taxonomy: Include social networking. *Nature*, 467(7317), 788-788.
- Simms, M. D., Dubowitz, H., & Szilagyi, M. A. (2000). Health care needs of children in the foster care system. *Pediatrics*, 106(Supplement 3), 909-918.
- Simon, H. A. (1996). *The sciences of the artificial*. MIT press.

- Singhal, A., Buckley, C., & Mitra, M. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 1996* (pp. 21-29): ACM
- Skoutas, D., & Simitsis, A. (2007). Ontology-based conceptual design of ETL processes for both structured and semi-structured data. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 3(4), 1-24.
- Smith, E. (1989). *Concepts and Induction, Foundations of Cognitive Science*, A Bradford Book. The MIT Press.
- Smith, E. E. (1988). 2 Concepts and thought. *The psychology of human thought*, 19.
- Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Harvard University Press.
- Sørli, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19), 10869-10874.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1), 11-21.
- Sparck Jones, K. (1974). Automatic indexing. *Journal of documentation*, 30(4), 393-432.
- Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium, 2001* (pp. 662): American Medical Informatics Association
- Susarla, A., Oh, J.-H., & Tan, Y. (2012). Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube. *Info. Sys. Research*, 23(1), 23-41, doi:10.1287/isre.1100.0339.
- Taivalsaari, A. (1996). Classes vs. prototypes-some philosophical and historical observations. *Journal of Object-Oriented Programming*, 10(7), 44-50.
- Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in

the eye of the beholder? [doi: DOI: 10.1016/0010-0285(91)90016-H]. *Cognitive Psychology*, 23(3), 457-482.

Teorey, T. J., Yang, D., & Fry, J. P. (1986). A logical design methodology for relational databases using the extended entity-relationship model. *ACM Computing Surveys (CSUR)*, 18(2), 197-222.

Tolbert, P. S., & Zucker, L. G. (1999). The institutionalization of institutional theory. *Studying Organization. Theory & Method. London, Thousand Oaks, New Delhi*, 169-184.

Tremblay, M. C., Berndt, D. J., Luther, S. L., Foulis, P. R., & French, D. D. (2009). Identifying fall-related injuries: Text mining the electronic medical record. *Information Technology and Management*, 10(4), 253-265.

Tsui, E., Wang, W. M., Cheung, C. F., & Lau, A. S. (2010). A concept–relationship acquisition and inference approach for hierarchical taxonomy construction from tags. *Information Processing & Management*, 46(1), 44-57.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1), 141-188.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, 15(1), 121-149.

Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1), 14-24.

von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 28(1), 75-105.

Wales, R., Colman, M., & Pattison, P. (1983). How a thing is called—A study of mothers' and children's naming. *Journal of Experimental Child Psychology*, 36(1), 1-17.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information systems research*, 3(1), 36-59.

- Wand, Y., Storey, V. C., & Weber, R. (1999). An ontological analysis of the relationship construct in conceptual modeling. *ACM Transactions on Database Systems (TODS)*, 24(4), 494-528.
- Wand, Y., & Weber, R. (1990). An ontological model of an information system. *Software Engineering, IEEE Transactions on*, 16(11), 1282-1292.
- Wand, Y., & Weber, R. (1995). On the deep structure of information systems. *Information Systems Journal*, 5(3), 203-223.
- Wand, Y., & Weber, R. (2002). Research commentary: Information systems and conceptual modeling—a research agenda. *Information Systems Research*, 13(4), 363-376.
- Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*, 36(5), 571-581, doi:10.1037/0012-1649.36.5.571.
- Weber, R. (1997). *Ontological foundations of information systems*. Coopers & Lybrand and the Accounting Association of Australia and New Zealand Melbourne.
- Weber, R. (2003). Still desperately seeking the IT artifact. *MIS Quarterly*, 27(2), 183-183.
- Whitaker, T. (2004). *If You're Right for the Job, It's the Best Job in the World: The National Association of Social Workers' Child Welfare Specialty Practice Section Members Describe Their Experiences in Child Welfare*. National Association of Social Workers.
- Yin, R. K. (2013). *Case study research: Design and methods*. Sage publications.
- Younger, B. A., & Fearing, D. D. (2000). A Global-to-Basic Trend in Early Categorization: Evidence From a Dual-Category Habituation Task. *Infancy*, 1(1), 47-58, doi:10.1207/s15327078in0101_05.
- Yu, E. (2001). Agent-oriented modelling: Software versus the world. In *Agent-Oriented Software Engineering II* (pp. 206-225). Springer.

- Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing - style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3), 378-393.
- Zhou, H., Liu, J., Jing, W., Qin, Y., Lu, S., Yao, Y., et al. (2010). Basic Level Advantage and Its Switching during Information Retrieval: An fMRI Study. In Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, & J. Huang (Eds.), *Brain Informatics* (Vol. 6334, pp. 427-436, Lecture Notes in Computer Science). Springer Berlin / Heidelberg.
- Zhu, Q., & Azar, A. T. (2015). *Complex system modelling and control through intelligent soft computations*. Springer.
- Zima, B. T., Bussing, R., Crecelius, G. M., Kaufman, A., & Belin, T. R. (1999). Psychotropic medication use among children in foster care: relationship to severe psychiatric disorders. *American Journal of Public Health*, 89(11), 1732-1735, doi:10.2105/AJPH.89.11.1732.
- Zipf, G. K. (1935). The psycho-biology of language.
- Zmud, R. W. (1978). An empirical investigation of the dimensionality of the concept of information*. *Decision sciences*, 9(2), 187-195.
- Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. Basic Books.

VITA

ARTURO CASTELLANOS

www.arturocastellanos.com

2012-2016	Ph.D., Business Information Systems and Business Analytics College of Business – Florida International University Miami, FL
2011-2012	M.S. Management Information Systems Decision Sciences and Information Systems College of Business – Florida International University Miami, FL
2004-2009	B.S. and M.S. in Telecommunications Engineering University of Navarra San Sebastian, Spain

PUBLICATIONS AND PRESENTATIONS

Lukyanenko, R., and Castellanos, A. (2016). “Introducing Information Gradient Theory” presented at DESRIST 2016 (St. Johns, Newfoundland, Canada).

Castellanos, A. (2015) “Relevance is in the Eye of the Beholder: Design Principles for the Extraction of Context-Aware Information from Healthcare Data” presented at the Workshop on Information Technologies and Systems (WITS) 2015 (Fort Worth, Texas).

Castillo, A., Castellanos, A., & Tremblay, M. C. (2014). “Improving Case Management via Statistical Text Mining in a Foster Care Organization” In *Advancing the Impact of Design Science: Moving from Theory to Practice* (pp. 312-320). Springer International Publishing. Presented at DESRIST 2014 (Miami, FL).

Castellanos, A., Castillo, A., and VanderMeer, D. (2013) “ExUp: Inferring Multiple-Dimension Rating System” Presented in WITS, Milan, Italy, December 2013.