Data as a Strategic Resource beyond Predictive Analytics

Tom Steinberger
KAIST
tomsteinberger@kaist.ac.kr

Ju Yeon Jung
University of Notre Dame
jjung8@nd.edu

Lily Cho
KAIST
lilicho@kaist.ac.kr

Abstract

Extending IS theories of data and strategy that assume data are ultimately used for predictive analytics, this paper explores how data may be used as a strategic resource beyond the statistical predictions of analytics tools. Our point of view is that a choice exists of which relations in data — abstract statistical relations for predictive analytics, or domain-specific conceptual relations for understanding — are to be enrolled in knowledge creation. We present evidence from the choice of data variables in 162 scientific papers in a subfield of metagenomics, supplemented by analysis of 231 patents from the same subfield. We discuss how accounting for the strategic use of data beyond analytics has important implications for IS theories regarding the value of domain knowledge and the location of bottlenecks in digital ecosystems.

Keywords: Data, strategy, metagenomics, data relations, knowledge creation, data analytics.

1. Introduction

Amid the proliferation of data and data tools in the digital economy, many researchers and practitioners observe that data have become strategic resources [1, 2]. Data can be strategic across firm, industry, and ecosystem levels. Firms perform better when data-driven predictions routinely inform their decisions [3]. Data pipelines for analytics are strategic for innovation in many industries, such as the real-time processing of incoming sensor data from cars to develop capabilities in autonomous vehicles. The massive scale of data servers and processors that power predictive analytics are key bottlenecks of digital ecosystems occupied by firms such as Amazon and Google [4].

Much strategic management discourse highlights how the strategic value of data has been driven by the greater volume, variety, or velocity of data now available to firms [5]. IS researchers have argued for attention also to how the value of data is contingent on the context-specific mechanisms by which data are constructed into data objects [6]. Structuring data into data objects involves the "definition and specification of formal relations between data items and fields" [7]. Insight into how such relations in data are defined and specified is important in that data are not simply raw materials like oil, so much as they are flexible mediums for shaping knowledge creation processes underlying sense-making and decision-making [8].

As in the strategic management field, however, emerging IS theories of data and strategy have mostly taken for granted that the ultimate use of data, and thus their ultimate strategic value, is for predictive analytics — techniques such as machine learning that generate statistical predictions from data to inform decision-making [9]. In predictive analytics, the strategic value of data derives from abstract statistical relations in the data — that is, from patterns among large numbers of data points that can be interpreted based on statistical ideas (e.g., mean, variance) rather than based on domain knowledge. Using data, however, does not necessarily draw on their statistical relations. Many types of more semantically rich relations among data specific to the conceptual structure of a domain can also inform knowledge creation [10]. A physician's insights about a patient, to pick just one example, are commonly informed by scanning over data whose relations are purposely structured and displayed in an electronic health record (EHR) [11].

The potential for domain-specific conceptual relations in data to be of strategic value can be observed across the data infrastructure that firms use. The vast majority of firms still rely on relational databases (e.g., Oracle, Microsoft Access) that enable creating queries or functions based on how columns of data are related conceptually within a domain. While digital twins are often understood as technologies to

URI: https://hdl.handle.net/10125/107133
978-0-9981331-7-1
(CC BY-NC-ND 4.0)

support analytics, they are also designed with the motivation of enabling rich communication about how data across entire physical processes specific to a production setting are related [13]. Emerging generative AI tools allow domain experts to interact in natural language with semantically rich relations among data, beyond the abstract statistical relations in their data.

While we have a growing understanding of the use of abstract statistical relations in data as a strategic resource for predictive analytics, we have less insight into the question of how domain-specific conceptual relations in data may be used strategically. The common use of such conceptual relations in data to inform sensemaking and decision-making in firms suggest room to broaden theories of data as a strategic resource beyond the ultimate use of data for predictive analytics.

To develop insight into this gap, we explore evidence from the use of metagenomics data in R&D on nuruk, a Korean fermentation agent with diverse consumer and industrial biotechnology applications. Metagenomics data refer to data on the genetic material of communities of microorganisms. Analyzing the use of such data can offer broader theoretical insight into the strategic value of data. For one, scientific data has become a strategic resource across many industries, from drug discovery in the pharmaceutical industry, to boosting crop yields in agriculture. More generally, metagenomics data typically have high volume and variety, characteristics that are central to the emergence of data as a strategic resource in other industries, such as the explosion in the volume and variety of sensor data available to manufacturers. Our particular setting within metagenomics was ideal in that nuruk R&D is characterized by two distinct uses of metagenomics data — both for abstract statistical relations and domain-specific conceptual relations — that allowed close comparison and surfacing the strategic choices at play in defining and specifying relations in data.

Based on analyzing 487 data variables from 162 scientific papers published between 2011 to 2022, and supplemented by analysis of 231 patents from the same period, we surface how the uses of metagenomics data were contingent on two strategies for knowledge creation. First, a "predictive analytics strategy" involved specifying statistical relations to optimize a mapping between isolated strains and specific functions (e.g., maximizing the production of a specific enzyme), while ignoring most other process and outcome variables. The other "conceptual relations strategy" involved specifying relations to understand how whole metagenomes, and multiple process and outcome variables were interdependent, while ignoring the optimization of any specific function.

Our findings contribute to IS theories of data by showing how the strategic value of data is a choice contingent on strategies for knowledge creation. Our findings offer novel implications, firstly, regarding the relationship between data and domain knowledge in the digital economy [7]. Using data as a resource for prediction has been found to diminish the value of domain knowledge in place of domain-independent knowledge of analytics techniques [13, 14]. In contrast, domain-specific knowledge should be a core complementary resource to data stored and processed in terms of their conceptual relations. Our findings also contribute to IS theories of digital ecosystems strategies that have been premised on massive scale in storing and processing data [12]. Making available domain-specific conceptual relations in data should enable a digital ecosystem's strategic use of data to be less about scale in data storage and processing and more about the richness of interaction with data (e.g., through domain-specific user interfaces).

2. Data as Strategic Resources

Predictive analytics extract statistical relations in data to enable a form of knowledge creation based on a process of "explor[ing] through perpetual experimentation" [8:1]. Certain features of a dataset are treated as if they are a vast number of isolated "trials" with which to experiment. Analytics techniques are then applied to generate statistical correlations between each of the vast number of isolated "trials" and a specific function [16]. Optimizing the function often depends on processing a large volume and variety of data to make the statistical correlations precise. The rise of predictive analytics has led to a shift towards viewing knowledge creation as increasingly a process of algorithm-driven predictions based on statistical relations in large volumes of data.

Consider the synthetic biology company Ginkgo Bioworks, which engineers isolated strains of organisms to perform specific functions (e.g., maximizing the production of a certain enzyme) for clients from many domains. Ginkgo's[1] strategy is driven by its database of over 35 million genetic sequences that identify strains and their characteristics (proteins, enzymes, metabolic pathways, etc.). To engineer a strain, Ginkgo treats its 35 million sequences as isolated "trials" to experiment with, using analytics to statistically correlate data about each sequence to the target function and predict which "trials" (e.g., which genetic sequences) optimize the target function. Ginkgo codifies knowledge gained from the statistical correlations (e.g., X strain maximizes Y enzyme) to inform how it engineers compounds for future clients. The potential strategic value of its data rests in statistical relations among the abstract data that can be "discovered" by analytics tools. The data in such predictive analytics uses are largely "homogenized" or "liquified" into sets of features, tokens, or bitstrings that are "divested from the material forms and situations to which they refer" [17:404].

By conceptual relations in data, we refer not so much to conventional uses of data for operations or managerial decision-making that are collected in a stable way, such as the data for a traditional sales forecasting tool used in a classic multidivisional firm. We use the term conceptual relations to refer to how the principles of domains, or material realities of situations, can be defined and specified in data, in ways that can be flexibly interacted with and interpreted by domain experts. Data are not so much abstract features, as they are domain- and situation-specific "annotations" that are enrolled in expert-driven knowledge creation processes of diagnosing, troubleshooting, and otherwise qualitatively reasoning to understand the world [18]. Such conceptual relations in data may be contained informally in any data object (e.g., a summary of key variables in a dashboard), or formally in a database schema or a query of the schema. Tables that display conceptual relations can be flexibly used by experts to reason about data from many points of view to piece together knowledge of a situation specific to their domain.

Compare the use of data at Ginkgo to the use of data by the biotechnology company Biome Makers[2] to analyze soil for winemakers. As at Ginkgo, Biome Maker's core data are genetic sequences of microorganisms. Yet the value comes from drawing also on the data on 14 million taxonomic references (phylum, order, species, etc.) and over 35,000 soil samples from 40 countries to generate a "soil assessment report" for each vineyard. The report is essentially a collection of tables of data, each displaying multiple variables of data that are known to be important to managing soil in vineyards. One six-page report, for example, displayed 23 tables on 77 types of data about soil composition. The tables of the assessment report reflect knowledge about subsets of data, with tables of data given domain-specific headings such as "biocontrol", "hormone production", and "stress adaptation". The value of the report is to enable clients to view the various tables together to reason about the current state of their soil and diagnose problems or consider possible treatments. Rather than algorithmically creating knowledge based on massive numbers of experiments, the data are used to describe the single situation of a specific soil sample.

While extant theory has focused on the use of statistical relations in data for predictive analytics, we know less about how firms such as Biome Makers may use domain-specific, conceptual relations in data as a strategic resource. The uses are hardly mutually exclusive. As Rutschi, Berente, & Nwanganga [12] note, even within predictive analytics, there is growing acknowledgement that value depends on "thorough assessment of the [domain]" and of "distinctive situations". Suggestive in the two examples of Ginkgo Bioworks and Biome Makers, however, is that views of the value of data may influence firms' underlying processes for storing, structuring, and otherwise interacting with data to create knowledge.

Different views of the value of data in knowledge creation have important and novel

[1] "Grow with Ginkgo, 2021 Update and Business Review" https://s28.q4cdn.com/823357996/files/doc_financials/2021/q4/Q4-2021-Earnings-Slide-Flow-FINAL-3.30.2022.pdf; Ginkgo Bioworks, 2021 Annual Report https://s28.q4cdn.com/823357996/files/doc_financials/2021/ar/2021-Annual-Report-(1).pdf

[2] Biome Makers, "BeCrop Microbiome Analysis Report". https://biomemakers.com/becrop-test; Biome Makers, "BeCrop Technology, Setting the Standard for Soil Health", https://biomemakers.com/becrop-technology/.

implications for strategy making. For example, in an analytics view, data generation and use are more in service of abstract knowledge contained in optimized functions, rather than the knowledge of "expert cultures" in a specific domain and in line with the stable and long-run goals of an organization [7]. In this paper, our point of view is that the relations in data that underpin the mechanisms of data generation are a strategic choice — the choices are not just towards this metric or that, but between abstract statistical and domain-specific conceptual relations in data as a basis for knowledge creation. We ask: can data be used as strategic resources not just for their statistical relations as inputs to predictive analytics, but for their domain-specific, conceptual relations? What are the contingencies in determining how data are used strategically? To explore such questions, we first theoretically frame, then present empirical evidence from, the use of metagenomics data.

3. Setting: Metagenomics Data

Metagenomics data refer to data on the genetic material of microbiomes, or communities of microorganisms (yeasts, bacteria, molds) contained in biological samples such as soil, water, or the human gut. The emergence of cost-efficient technologies for DNA sequencing over the past few decades has led to a proliferation in the availability of metagenomics data [19]. These data include genetic sequences (the "genome") of microorganisms, as well as data on transcription sequences, protein sequences, metabolites, and metabolic pathways. Metagenomics datasets also may include other data describing microorganisms, such as taxonomic references, enzymatic or chemical functions, data on microbiome samples (pH, humidity, temperature, etc.), and details on collection techniques or experimental methods.

The strategic value of metagenomics data is in how they may help address a fundamental bottleneck in microbiology, which is that the vast majority of microorganisms cannot be cultured in isolation in a lab. Analysis of microbiome samples address this bottleneck by allowing the genetic sequencing of hundreds or thousands of unculturable microorganisms. In terms of downstream strategic value for businesses, the advantages of metagenomics data in enabling the study of unculturable microorganisms include: (1) dramatically expanding the overall population of microorganisms that a firm can collect data on when innovating new products (as in Ginkgo Bioworks' strategy to innovate novel compounds); (2) dramatically expanding data about specific populations of microorganisms that a firm can analyze to create and share knowledge about specific settings (as in Biome Makers' strategy to provide site-specific "assessment reports" to clients).

The value of metagenomics data is, however, contingent on their use. One use of metagenomics data is simply to expand the population of *individual organisms* that can be experimented with in biotechnology applications that use predictive analytics. As a famous example, Craig Venter's Global Ocean Sampling Expedition collected a dataset of 7.7 million genetic sequences from samples of ocean water, where one of the goals was to use predictive analytics to identify novel biocatalysts for specific functions required in healthcare applications [19]. Another use of metagenomics data focuses more on understanding the diverse ways in which a *community of organisms* in a microbiome interacts with its environments. For example, in the Human Genome Project, one of the goals was to understand the diverse ways in which variations in the flora of the gut microbiome affected health outcomes.

Stevens [20] notes that genetic sequence data are typically stored similarly to sequences of characters in text, such that the proliferation of this data has led to an understanding of genetic processes as resembling the techniques of search engines that find statistical correlations between strings of characters in web text. He points out that, given that even the genome of a single organism may be more dense and non-linear than correlations in text, such assumptions may "limit the kinds of ways in which [the genetic sequence data] can be understood and manipulated" [20:353]. Such assumptions may also have scale-related implications for the strategic use of data. Using genetic sequence data for predictive analytics requires larger teams of data scientists and engineers and the ability to store and process data at massive scale.

Though different uses of data are not mutually exclusive, the use of metagenomics data in practice can be viewed as a strategic choice to the extent that it is driven by assumptions about how data

are to be enrolled in processes of knowledge creation. To explore contingencies that may drive the strategic use of data, we next present evidence from the use of metagenomics data for both abstract statistical relations for analytics and for domain-specific conceptual relations in R&D on the Korean fermentation agent nuruk.

4. The Contingent Value of Metagenomics Data in Nuruk R&D

Nuruk is a fermentation agent traditionally used in Korean brewing. Nuruk is made by shaping finely or coarsely milled grain (wheat, rice, etc.) into discs and exposing the discs to air under controlled temperature and humidity. Each nuruk disc has a unique microbiome of hundreds or thousands of strains of microorganisms (yeasts, molds, bacteria) derived from the surface of the grains in the nuruk and ambient air in which the nuruk is fermented. Beyond brewing, nuruk has diverse consumer and industrial applications, with nuruk-related patents issued for functions ranging from food additives to therapeutics, cosmetics, textiles, and biomaterials. At the basic science level, nuruk microbiomes are a potentially valuable source of novel organic compounds for synthetic biology.

Cost-efficient DNA sequencing technologies for collecting and analyzing metagenomics data became used to study nuruk from the late 2000s. The newfound data helped enable a steady stream of scientific papers on nuruk, with an average of 15 papers published per year between 2011 and 2022 (based on our review of the scientific literature on nuruk). Some of these papers studied individual strains isolated from nuruk metagenomes to optimize specific functions, analogous at a small scale to how Ginkgo Bioworks engineered organisms for its clients from its massive library of individual genetic sequences. For example, one paper extracted 481 fungi strains from 16 nuruk samples, then selected 11 strains for high ethanol production. Other papers did not isolate strains, and instead measured aggregate properties to map entire nuruk metagenomes to a network of functions, roughly analogous to how Biome Makers provides winemakers with "soil assessment reports" that analyze how properties of the soil may affect diverse aspects of winemaking. For example, one paper analyzed the fungal and bacterial diversity of 58 samples of nuruk for brewing, then analyzed the effects of each sample on 12 outcome variables, ranging from pH to aromatic compounds.

The different uses of metagenomics data in these scientific papers provide an empirical illustration of how the use of data can be contingent and suggest a strategic choice. To explore where these contingencies come from, and thus to gain insight into the strategic use of nuruk data, we compared the full set of data variables used across scientific papers on nuruk. We take the full set of data variables as fundamental constraints on how data may be used as a resource for creating and capturing value from nuruk R&D in downstream applications developed in firms.

We extracted the data variables from the tables and figures of 162 scientific papers (link to reference list here) on nuruk published between 2011 and 2022. We compiled a dataset of 487 data variables (the columns of our dataset) along with the papers in which each of the variables appeared (the rows of our dataset). We then manually coded each paper into a type ("statistical relations" type or "conceptual relations" type) based on the outcome variables of interest, drawing on the first author's domain knowledge about the nuruk setting. Based on our coding, we identified 86 of the 162 papers as primarily studying the effects of *isolated strains*, or individual microorganisms isolated from the nuruk metagenome. We identified the other 76 papers as instead primarily studying the effects of properties of *whole metagenomes*, or properties of the communities of microorganisms that made up the nuruk metagenome (properties included: taxonomic composition, yeast diversity, bacterial diversity, fungal diversity, relative abundance).

Second, we analyzed the use of explanatory (non-outcome) data variables for the two types of paper. We ran pairwise comparisons tests across all data variables, which provided initial evidence that the two types of scientific papers had systematic differences in their composition of variables. We next aggregated the data variables into coarser-grained categories to make sense of and summarize the systematic differences. Drawing on the table and figure headings, and based on interviews and conversations with three nuruk scientists, we divided the remaining data variables into the subcategories of: (i) *Process variables*, or variables (inputs, times,

temperatures, etc.) regarding a downstream value-adding process in which nuruk is used (e.g., the fermentation process for brewing alcohol using nuruk, or the manufacturing process for making cosmetics with nuruk); (ii) *Outcome variables*, or variables on any outcomes of a process, such as assays of the aromatic or volatile compounds in brewing or cosmetics, or basic sensory properties such as flavor, taste, and color; (iii) *Specific functions*, which refer to highly targeted metabolite functions (e.g., cytokine production, ethanol tolerance), as well as enzyme activity.

Based on the three categories of data variables, we constructed simple frequency tables to explore the hypothesis that differences in the use of metagenomics data ("isolated strains" versus "whole metagenome") imply differences in the type of relations (statistical versus conceptual) in data that are used.

Metagenomic Data	**Other Data Variables**		
	Process Variables	Outcome Variables	Specific Functions
Isolated Strains	31.0%	35.5%	67.4%
Whole Metagenome	69.0%	64.5%	32.6%

Table 1. Data Used in Scientific Papers on Nuruk

We find two complementary results, depicted in Table 1 above. First, the two categories of papers differed dramatically not just in their use of metagenomic data ("isolated strains" versus "whole metagenome"), but also in the other data variables that were included. The scientific papers in the "whole metagenome" category were far more likely than the "isolated strains" category to include data on process and outcome variables, while papers in the "isolated strains" category were far more likely to include data on specific functions. Second, the analysis of these data variables differed. Papers in the "whole metagenome" category presented values of process variables, outcome variables, and the metagenomic properties of the nuruk without characterizing any optimal value for these data variables. Given the lack of any optimized functions, we infer that the ultimate use of this data depended on conceptual reasoning by a domain expert about how metagenomic properties related to multiple process and outcome variables and, equally, about how process and outcome variables were themselves related. Conversely, papers in the "isolated strains" category highlighted which isolated strains produced optimal values for specific functions, consistent with a predictive analytics approach.

We interpret these results as evidence of two strategies for using data in nuruk R&D. We characterize a "whole metagenome strategy" in nuruk R&D as using a semantically rich set of conceptual relations in data among diverse process and outcome variables (brewing, physicochemical, and sensory characteristics) to understand the effects of nuruk at the metagenome level. In contrast, we characterize an "isolated strains strategy" as generating statistical relations in a narrower set of data, between a population of isolated strains and specific functions. We depict these two strategies in Figure 1 below.

To uncover evidence of a strategic choice specifically about the value of data, we conducted a further round of coding of the downstream applications targeted by the scientific papers, which included brewing, cosmetics, therapeutics, and food additives. We found that "isolated strains" and "whole metagenome" categories had no significant differences, such that there was no clear evidence that downstream applications drove the use of data at the research-level. For example, 52 of the 162 scientific papers related to brewing applications, and these were virtually equally likely to be from either category. Overall, we interpret our results as evidence that the use of metagenomics data in our setting was contingent on the strategy of the researcher to create knowledge from either statistical relations or conceptual relations in their data.

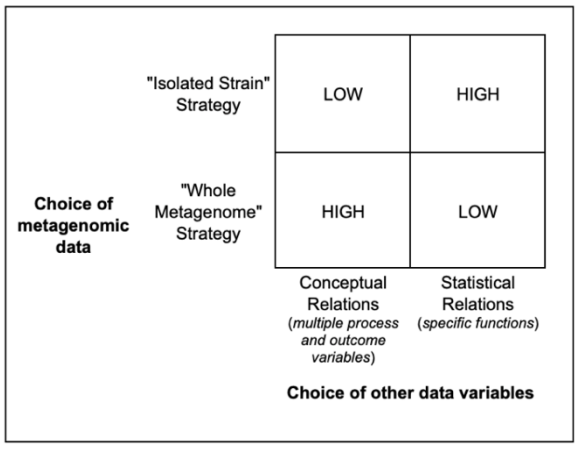

Figure 1. Strategies for Nuruk Metagenomic Data

To explore how our findings on scientific papers may connect to the strategic use of data in downstream applications of nuruk, we also analyzed 231 nuruk-related patents issued between 2011 and 2022. We manually coded the patents into the same categories of "isolated strains" (84 patents) and "whole metagenome" (147), based on whether an isolated strain was patented. The "whole metagenome" category was distributed across the full range of categories. The "isolated strains" category had little or negative correlation with most applications, but it was strongly correlated with 52 of the patents on brewing methods — that is, with the joint patenting of both a brewing method and an isolated strain. We interpret these latter results as evidence that the "isolated strains strategy" consistent with predictive analytics involved the strategic use of metagenomics data under an assumption that conceptual relations in data (e.g., across brewing processes) are held constant (e.g., a brewing method is fixed and patented). This assumption is well-known in the Korean brewing industry, which has long been oriented towards a cost (as opposed to differentiation) strategy based on tightly controlling an industrial-style brewing process using isolated strains. The strategic use of statistical, as opposed to conceptual, relations in data about strains of isolated nuruk may thus be contingent on industry-level strategy, and not inherent to the strategic value of nuruk data.

5. Discussion and Conclusions

We have challenged an assumption in most extant theories of data as a strategic resource, which is that the value of data ultimately is in their statistical relations, extracted using predictive analytics techniques and for the purposes of optimizing specific functions. We have framed the strategic value of data as instead also potentially deriving from how conceptual relations in data enable domain experts to understand a network of process and outcome variables, beyond optimizing specific functions. Our setting provided evidence that the ultimate use of the same type of metagenomics data was contingent on how relations in data — statistical or conceptual — were emphasized in processes of knowledge creation.

We open up a novel view of data as strategic resources not just in the competitive strategy sense of storing and processing proprietary data at massive scale, but in the resources and capabilities sense — that a firm faces a strategic choice of how to link its data resources to firm-specific processes of knowledge creation. Our analysis showed how the use of metagenomics data in nuruk R&D for predictive analytics (e.g., to discover isolated strains that optimized specific functions, such as ethanol production) involved ignoring semantically rich and domain-specific relations among multiple downstream process and outcome variables. Conversely, accounting for these conceptual relations (e.g., to understand how whole metagenomes behave, such as the overall characteristics of a particular type or batch of nuruk) involved ignoring the specific functions. The strategies can be framed as two views of how to use data for knowledge creation that are driven by either "perpetual experimentation" (statistical relations view) or making sense of multiple domain-specific variables (conceptual relations view).

We argue that our framing of this strategic choice can be applied to a broad range of datasets beyond the idiosyncrasies of the metagenomics setting studied here. Even in classic settings for predictive analytics (e.g., behavioral data on a social media site), we propose that firms have a choice of whether to map statistical relations to specific functions (e.g., users' "engagement" with the site) or to extract conceptual relations that reflect a deeper understanding of users' engagement with the site in particular situations.

Based on our framing and analysis of our findings, we draw two broader implications for IS theories of data and strategy, one regarding the value of domain knowledge in the strategic use of digital data and another regarding digital ecosystems strategy.

First, we heartily agree with the point of view in recent IS research on data and strategy in which the constructive processes and mechanisms underlying how data are generated, transformed and used are given a central role [17]. Yet in this literature's focus on the ultimate use of data for predictive analytics, it has been rather pessimistic about the prospects for the strategic value of domain experts' knowledge. Strategies for using data to routinely inform decisions also depend on how well conceptual relations in data are made accessible to the firm's domain experts [15]. We argue for extending theories of how data objects are constructed to account for how experts interact with conceptual relations in data. We have in mind processes of constructing data objects from conceptual relations that are far more domain-specific than constructing abstract metrics for analytics, while extending well beyond conventional domain-specific uses of data in firms (e.g., generating periodic reports from an ERP system).

Second, the rise of data as a strategic resource for predictive analytics has shifted IS research on strategy making from a traditional focus on the organization level towards digital platforms and ecosystems. Extant ecosystems theories of data as a strategic resource for predictive analytics emphasize that the key strategic bottlenecks to value creation and capture are the capacity to store and process large volumes of data [3]. In these theories, layers of the ecosystem that directly support the end-users of data (e.g., dashboards or other interfaces, development frameworks) are of strategic value mostly for how they attract users to contribute their data. These user-facing layers have no strategic value on their own if they are not coupled with massive capacities for data storage and processing. Hence, it is well-observed that the use of data as a strategic resource for analytics has been dominated by large technology firms such as Amazon, Microsoft, Google and Alibaba that have the resources to invest in massive data storage and processing capacity and mostly open-source tools to attract users.

Our framing and findings regarding the contingent value of data as a strategic resource point to one way in which current strategic bottlenecks could also concern user-facing layers for interacting with data specific to a domain. In our metagenomics setting, for instance, a strategy of focusing on conceptual relations across multiple process and outcome variables in nuruk R&D would have vastly lower requirements for data storage and data processing capacity than the variables (genetic sequence data) that were identified with the use of predictive analytics. For example, whereas the data variable "yeast diversity" has a single measure for a single sample of nuruk, the same sample may have hundreds or thousands of genetic sequences and dozens of columns of metadata about each sequence. The value of the "yeast diversity" variable would be unlocked more by a domain expert analyzing its many possible relations to other data variables (e.g., as in the dozens of data variables displayed in user-friendly tables by Biome Makers to its clients), which depends on their ability to richly interact with data more than their ability to store and process large volumes of data.

In conclusion, theories from IS strategy could be extended to explore how the current "ontological reversal", in which digital data increasingly determine the material world [21], might be a result not just of properties inherent in digital data but in how data is viewed as a strategic resource. For example, theories of data objects and digital ecosystems strategies might draw on this paper's arguments, as well as other emerging literatures that study end-users' interactions with data in their work practices [15, 22, 23], to investigate or hypothesize how the strategic use of data may be evolving, and the implications for the locus of value creation and capture.

6. Acknowledgements

The authors would like to acknowledge the support from the National Research Foundation of Korea by its grant N01230396.

7. References

[1] Hartmann, P., and J. Henkel, "The rise of corporate science in AI: Data as a strategic resource.", *Academy of Management Discoveries 6*(3), , pp. 359–381.

[2] Grover, V., R.H. Chiang, T.P. Liang, and D. Zhang, "Creating strategic business value from big data analytics: A research framework.", *Journal of management information systems 35*(2), 2018, pp. 388–423.

[3] Lo, A., and E. Brynjolfsson, *The Rise of Data Capital*, 2016.

[4] Jacobides, M.G., S. Brusoni, and F. Candelon, "The Evolutionary Dynamics of the Artificial Intelligence Ecosystem", *Strategy Science 6*(4), 2021, pp. 412–435.

[5] Adner, R., P. Puranam, and F. Zhu, "What Is

different about digital strategy? from quantitative to qualitative change", *Strategy Science 4*(4), 2019, pp. 253–261.
[6] Aaltonen, A., and E. Penttinen, "What makes data possible? A sociotechnical view on structured data innovations.", *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*, (2021), 5922–5931.
[7] Kallinikos, J., and I.D. Constantiou, "New games, new rules: big data and the changing context of strategy.", *Journal of Information Technology 30*(1), 2015, pp. 44–57.
[8] Alexander, D., and K. Lyytinen, "Organizing Successfully for Big Data to Transform Organizations.", *Proceedings of the Americas Conference on Information Systems.*, (2017).
[9] Davenport, T.H., "From analytics to artificial intelligence.", *Journal of Busines Analytics 1*(2), 2018, pp. 73–80.
[10] Koller, D., "Probabilistic relational models.", *Inductive Logic Programming: 9th International Workshop, ILP-99 Bled, Slovenia, June 24–27, 1999*, Springer Berlin Heidelberg. (1999), 3–13.
[11] Kellogg, K.C., M. Sendak, and S. Balu, "AI on the Front Lines.", *MIT Sloan Management Review 63*(4), 2022, pp. 44–50.
[12] Rutschi, C., N. Berente, and F. Nwanganga, "Data Sensitivity and Domain Specificity in Reuse of Machine Learning Applications.", *Information Systems Frontiers*, 2023, pp. 1–8.
[13] Alaimo, C., and J. Kallinikos, "Organizations Decentered: Data Objects, Technology and Knowledge", *Organization Science 33*(1), 2022, pp. 19–37.
[14] Sambasivan, N., and R. Veeraraghavan, "The Deskilling of Domain Expertise in AI Development", *CHI Conference on Human Factors in Computing Systems*, ACM (2022), 1–14.
[15] Park, S., A.Y. Wang, B. Kawas, Q.V. Liao, D. Piorkowski, and M. Danilevsky, "Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models", *26th International Conference on Intelligent User Interfaces*, Association for Computing Machinery (2021), 585–596.
[16] Balasubramanian, N., Y. Ye, and M. Xu, "Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning", *Academy of Management Review*, 2020.
[17] Aaltonen, A., C. Alaimo, and J. Kallinikos, "The Making of Data Commodities: Data Analytics as an Embedded Process", *Journal of Management Information Systems 38*(2), 2021, pp. 401–429.
[18] Forbus, K., and J. De Kleer, *Building Problem Solvers.*, MIT Press, Cambridge MA, 1993.
[19] Bietz, M.J., and C.P. Lee, "Collaboration in metagenomics: Sequence databases and the organization of scientific work.", *ECSW 2009*, Springer London (2009), 243–262.
[20] Stevens, H., "Hadooping the genome: The impact of big data tools on biology.", *Biosocieties 11*, 2016, pp. 352–371.
[21] Baskerville, R.L., M.D. Myers, and Y. Yoo, "Digital First: The Ontological Reversal and New Challenges for Information Systems Research 44 (2) (2020), pp. 509-523", *MIS Quarterly 44*(2), 2020, pp. 509–523.
[22] Steinberger, T., and M. Wiersema, "Data Models as Organizational Design: Coordinating beyond Boundaries Using Artificial Intelligence", *Strategic Management Review 2*(1), 2021, pp. 119–144.
[23] Jung, J.Y., T. Steinberger, J.L. King, and M.S. Ackerman, "How Domain Experts Work with Data: Situating Data Science in the Practices and Settings of Craftwork.", *Proceedings of the ACM on Human-Computer Interaction 6(CSCW1), pp.1-29.*, 2022.