Communications of the Association for Information Systems

Volume 47 Article 11

11-5-2020

Discovering and Transforming Exhaust Data to Realize Managerial Value

Daniel OLeary
University of Southern California, oleary@usc.edu

Veda C. Storey
Georgia State University, vstorey@gsu.edu

Recommended Citation
OLeary, D., & Storey, V. C. (2020). Discovering and Transforming Exhaust Data to Realize Managerial Value. Communications of the Association for Information Systems, 47, pp-pp. https://doi.org/10.17705/1CAIS.04715

Communications of the Association for Information Systems

Research Paper DOI: 10.17705/1CAIS.04715 ISSN: 1529-3181

Discovering and Transforming Exhaust Data to Realize Managerial Value

Daniel E. O'Leary
Marshall School of Business
University of Southern California
oleary@usc.edu

Veda C. Storey
Computer Information Systems
J. Mack Robinson College of Business
Georgia State University

Abstract:

"Exhaust data" is "extra data" or "left over" data from "core data" digital transactions collected either intentionally or unintentionally but for which there is no initial, specific purpose for its collection. In this paper, we differentiate core data from exhaust data, define and describe exhaust data, and propose how to turn it into core data to provide value for firms. We present a framework for discovering and transforming exhaust data and apply it to four case studies involving Internet search data, accounting entries and data security, social media disclosures, and EDGAR use logs. From the cases, we extract five managerial challenges and generate five recommendations to help managers identify exhaust data applications for realizing potential value.

Keywords: Innovation, Creativity, Exhaust Data, Core Data, Big Data, Analytics, Framework for Exhaust Data Value Creation, Theory of Core Data, Theory of Core Systems.

This manuscript underwent peer review. It was received 08/04/2019 and was with the authors for six months for two revisions. Jennifer Xu served as Associate Editor.

1 Introduction

Core data is data generated or captured for a specific reason, intention, or core system. However, with the advent of social media, the Internet of things, and other emerging technologies, organizations are capturing data that goes beyond core data needs and beginning to understand additional uses for data collected along with the core data. In this era of big data and analytics, data is being gathered from people or devices, users or third parties, and agents, some of whom who are cognizant of the data gathering process and others who are not. This has resulted in the capture of some data that is not used in core applications and data that has unintended uses. Organizations may not even know that they possess additional usable data. Instead, that extra data is data that is "not originally sought", data that is "left over", data that is "exhaust", or data that is "unintended" (O'Leary & Storey, 2017b)

In the classic 1966 film Blow-Up, David Hemmings, playing a photographer, captures some photos of a couple in a park. The woman confronts Hemmings and asks for the photos. The inquiry sparks Hemmings' interest and curiosity about the pictures and inspires him to give the woman a different roll of film. After developing the original film, he notices something "different" in the pictures' background and enlarges them. He then spots what appears to be a dead body and a killer hiding in the grass. The immediate ("core") purpose of Hemmings' character was to capture interesting pictures ("core data") of the couple in the park. However, the background contained valuable information about a murder. The images of the dead body and killer in the grass were, in essence, "exhaust data" that he did not intend to capture or use. Nor did he know that the data existed until he examined the picture in detail. This example illustrates both the existence of exhaust data and the potential substantial value. However, identifying and using exhaust data is not cost-free; rather, persistence, creativity and insight are required to extract value.

This Blow-Up example also suggests how exhaust data, analogously, can be particularly important to management for at least two reasons. First, managers should be interested in *leveraging exhaust data to create value for their firm*. Like Hemmings, they want to find value from both the data intended to be captured (core data) and any additional data that is captured (exhaust data). Second, managers want to *ensure that their publicly available data provides limited exhaust data about them from which others can create value.* Like the couple in Blow-Up, users may not want others to discover critical information about them in pictures or other information disclosures. In the first case, management is interested in using the additional data to make inferences; in the second, management is interested in limiting or constraining data that can be used to make inferences about their firm.

As another example, consider the set of "connections" to someone on LinkedIn. The core purpose of connecting to someone is to keep each other informed about their activities. However, such connections may provide exhaust data information to an observer. An individual who works in sales may have connections to customers on LinkedIn, and competitors could use such connections to obtain knowledge about these customers. Knowledge of these customers may provide knowledge (sales leads) to *competitors* of the salesperson. Alternatively, perhaps the connections could be used to gather information about the creditworthiness of someone based upon whom they are connected to. The connection data exists and so do the unintended consequences of disclosing it by identifying an alternative meaning or value-creating use of the data.

1.1 Core: Competencies, Capabilities, Systems, Data—and More

Consistent with management and strategy research (e.g., Prahalad & Hamel, 1990; Leonard-Barton, 1993), we assume that organizations create strategies based on their "core competencies" and "core capabilities". Organizations build and align their "core systems" to actualize these strategies, and to create a "core" data architecture that provides access to necessary core data. Unfortunately, a focus on the core competencies, core knowledge, core systems, and core data can result in core rigidities, which limit the scope and types of analysis possible. We, therefore, investigate what it means to go beyond core data as a way to facilitate and generate innovations. The search beyond the core processes, systems, and data, leads us to analyze exhaust data for innovation, evolution and value.

1.2 Purpose and Plan of this Paper

The purpose of this paper is to review the literature and emerging definitions of exhaust data and propose a *framework for exhaust data value creation* to support the identification and effective use of exhaust data. The framework, derived from an analysis of multiple case studies, is a systems and data architecture-based approach for identifying exhaust data and turning it into core data. We illustrate the application of

the framework by applying it to additional case studies that highlight managerial challenges. We also extend discussions of core competency to core information systems and their corresponding core data. Our focus is on using exhaust data to generate innovation and organization evolution to create value beyond existing core capabilities.

This paper proceeds as follows: in Section 2, we describe related research that focuses on core issues of core capabilities, competences, and values in information systems, the limitations of focusing on those core issues, and the impact of focusing on core data in business intelligence. In Section 3, we present a framework for exhaust data value creation. In Section 4, we present four case studies, which provide the basis for deriving managerial challenges and implications for value creation in Section 5. In Section 6, we conclude the paper.

2 Core vs. Exhaust: Related Research and Previous Definitions

Although the notion of core competency has permeated management research, it has received limited attention in information systems. We, therefore, summarize some of the core competency literature and apply it to information systems and data, to create notions of core information systems and core data. We also examine prior research on exhaust data and analyze the differences between business intelligence (BI) using data-driven or purpose and need-driven approaches, as it relates to organizational innovation or evolution processes.

2.1 Core Issues: Competencies, Capabilities, Values, and Information Systems

Management researchers have long emphasized the importance of organizations focusing on “core” issues. Prahalad and Hamel (1990) were among the first to note the importance of management focusing on its “core competencies” and “core products”. Leonard-Barton (1992) noted the importance of “core capabilities” in management strategies. Urde (2003) and others have examined “core values” as a basis to build brand value.

According to Leonard-Barton (1992, p. 113) core capability is the knowledge set that is embodied in knowledge and skills, technical systems, managerial systems and values and norms. “Core systems”, whether managerial or technical, capture, generate, and report on “core data”. Viscusi, Huber and Bell (2011) expand values and norms to include incentives. Firms focus by aligning “core” knowledge, competencies and capabilities with “core systems” using “core data”. Organizations can facilitate meeting core incentives by implementing their core strategy. As a result, the notion of core competencies provides organizations with the importance of “focus” and a concern with knowledge, incentives, processes, systems and data that support and align with core strategy. Prahalad and Hamel (1990, p. 1), for example, have noted that “focusing on core competencies creates unique, integrated systems that reinforce fit among your firm’s diverse production and technology skills—a systemic advantage your competitors can’t copy”.

2.2 Core Rigidities: Limitations of Focusing on the Core:

Potential problems arise by focusing only on the “core” as identified by the same researchers. Prahalad and Hamel (1990) noted that it is important that organizations “not (be) so narrowly focused that they cannot recognize the opportunities for blending their functional expertise with those of others in new and interesting ways”. Leonard-Barton (1992) suggested that core capabilities have a downside that inhibits innovation, referred to as “core rigidities”. Leonard-Barton (1992, p. 118) indicated that core rigidities are the “flip side of core capabilities”. Core rigidity is reflected in missing or inflexible knowledge, incentives, processes, systems, or data. Unfortunately, as also noted by Leonard-Barton (1992), core rigidities are “problematic for projects that are deliberately designed to create new nontraditional capabilities” because core rigidities can limit the scope and types of analysis, as well as the data used. Thus, core rigidities occur because of a focus on the core, but they can limit innovation.

In addition, Leonard-Barton (1992) and others have concluded that “core capabilities are evolving and corporate survival depends on successfully managing that evolution” (p. 112). They argue that firms must continue to evolve and innovate, but that “core capabilities simultaneously enhance and inhibit development” (p. 112). As a result, organizations are interested in both identifying core (knowledge, incentives, processes, systems, and data), and allowing some variation in that core to facilitate innovation and evolution.

2.3 Core Data: Comparison to Exhaust Data

Core systems, typically, are used to accomplish a specific purpose (provide a service, develop a product, etc.). However, as part of implementing that specific purpose, those systems may generate additional data or facilitate the identification of additional data that can be used in different, non-core manners. For example, the Ring doorbell has a core purpose of providing its owners information about what is happening near their front door. Ring owners typically put the app on their phone, which becomes part of the "fingerprint" of the phone. The fingerprint, which is exhaust data, enables data about the owners to be captured and used for various purposes, which is an unintended consequence (Hussain, 2020). As another example, a mobile phone call generates contextual core data such as the location at the time of the call. Core data determines the charge for, or feasibility of, a phone call. The location of the phone, when not in use, is not required to assess use charges and is, therefore, exhaust data. However, information gathered by aggregating the number of phones at some location at particular points in time, independent of call information, could be useful for understanding the amount of traffic that goes through a given location at different times during the day[1]. Governments, retail stores, restaurants and hotels, and other organizations, could be particularly interested in such information derived from this type of exhaust data.

The term "exhaust data" captures the general notion of data that, initially, is not core data, but may be collected as a byproduct of some event (transaction, event, search, disclosure, etc.), has unknown value, and ultimately might be used for another purpose to create value. Furthermore, it may exist as an unintended consequence of gathering the core data. Exhaust data could be transient and never used or saved; for example, because of the cost of storing it (Ojo & Heravi, 2018). Although exhaust data has been classified as having "low value density" (Banno et al., 2015), it still has the potential to create value or mitigate loss as it evolves into core data, even if this value is not immediately recognized.

There are many characteristics and definitions of exhaust data (see Table 1 on the next page) perhaps because there are different technologies and domains where exhaust data has been created and analyzed. The role of technology and the participation of users may also vary by source or type of exhaust.

Although we differentiate between core and exhaust data, the differences can occur at different times and settings. Exhaust data for one organization can be core data for another. If an organization can effectively leverage its exhaust data, that data could become core data and an integral part of the organization's core capabilities as the organization evolves.

2.4 Core Processes and Innovation: Business Intelligence Applications

Consistent with notions of core competence, core business intelligence (BI) focuses on analysis of core data identified as being core to business operations and competitors. Using core process knowledge, business intelligence uses models of core data as a basis of analysis. For example, core key performance indicators associated with core business processes could be monitored continuously to determine the status of a range of core data, such as sales, whether by the originating organization or their competition. Core-based BI is likely to be problem, or need driven, because the core knowledge can drive an understanding of the relationships between the data.

On the other hand, innovators could also apply business intelligence to exhaust data. They would first need to identify what data might be useful and how it could be processed. In the case of Blow-Up, this could be systematically examining the edges of pictures. For social media, such as LinkedIn, this would require constantly assessing the opportunity to gain insights from disclosed data. However, it can be difficult to discover useful data, which might have been generated unintentionally, and to derive value from that data. These issues are the primary concerns of this paper.

Unlike core data, initially, there is usually not sufficient knowledge, processes, or systems to use exhaust data in business intelligence. However, if exhaust data can be identified and analyzed, then organizations may be able to identify innovations that can leverage the exhaust data, make it core data, which, in turn, may generate innovations or facilitate evolution. Below, we provide a framework and examples to illustrate additional innovative uses of exhaust data.

[1] This approach helps Google Maps determine the amount of car traffic (https://electronics.howstuffworks.com/how-does-google-maps-predict-traffic.htm)

Finally, although we suggest that physically and conceptually, core and exhaust data can be different, the roles may also differ based on which organization or which parts of an organization are investigating the data. For example, location data might be exhaust data to a phone service company that does not need it after a call is made. It might, however, be core data to a company that specializes in selling phone location data. Accountants may not identify exhaust data in LinkedIn. Marketers, however, might find the exhaust data from LinkedIn as important and critical as core data when exploring potential sales opportunities.

Table 1. Characteristics of Exhaust Data

Characteristic	Description / explanation	Example
Left-over, extra, or remnant data (Davis & Davidson, 1992; Davidson, 2016)	Not originally intended for additional use beyond core transaction	Travel app with origin, destination and device data
Context / background data (O'Leary & Storey, 2017a)	Originally from identifiable data but not intended for use.	Location data from a call; name associated with a transaction
Inadvertent, fortuitous, or over-disclosed data (O'Leary & Storey, 2017a)	Captured coincidentally along with core data, including data disclosures that may go beyond requirements	Pile of money in a picture, address in a picture
Inferred data (Ginsberg et al., 2009; O'Leary, 2013)	Generated because a group of "symptoms" infer a cause.	Searches about stomach ache, vomiting, fever data indicate flu or food poisoning
Structured, unstructured, or non-standard data (O'Leary & Storey, 2017a; George et al., 2014)	Exhaust data appears in a variety of forms, depending on source, application, technology and domain.	Pictures, social media text, maps, addresses, co-occurrence of objects
Repurposed (George et al., 2014) or stolen (O'Leary & Storey, 2017a) data	Typically used for a different purpose than its original intent	Social media text or pictures
Passively collected transactional data or ambient (George, Haas, & Pentland, 2014)	Extracted from use of digital services or Internet of things; limited or zero value to original data collection purposes, but can be recombined with other data sources	Purchases, even at informal markets, or when customers interact; humidity, temperature, movement, noise levels, lack of noise
Ephemeral data by-products (George et al., 2014)	Obtained from conversations or interactions	Saved internet searches using Google, Yahoo, etc. to measure interest or activity.
Device and program data (Johnson, Gray, & Sarker, 2019) or Internet-use data (Schweidel, 2014)	Often not intended for human use, but for device and program communication	Phone location Information, cookies, temporary files
"Hidden" or "deceptive" intentions to gather data (e.g., user information)[2]	In the sharing economy, the core transaction would relate to the specific shared asset, but user information would be one type of exhaust.	Uber, Airbnb, bike rentals capture user information

2.5 "Theory of Core Data" and the "Theory of Core Data Systems"

This discussion up to this section leads to what we call the "theory of core data" and the "theory of core data systems", which assert that, in order to manage core competency, organizations focus on core data and core systems designed to gather, process and report on core data. The core data and core data systems are based on the core knowledge set (Leonard-Barton, 1992). Further, in order to generate appropriate management behavior and support key decisions consistent with core competencies, management's incentives would be based on those core competencies and implemented in the core data systems. These systems include transaction-processing systems, supply chain systems, and business intelligence systems, which are all designed to process core data that helps firms maintain their core competencies and focus on their core purposes. Together, these two theories of "core data" and "core systems" help form the basis of our exhaust data framework, presented in the next section.

[2] We thank one of the anonymous referees for this example.

3 Framework for Exhaust Data Value Creation

Figure 1 presents our proposed framework for understanding how to discover and use exhaust data, as well as how to generate potential value-creating exhaust data applications. We base the framework on both a systems theory and data approach to capturing meaning from data, focusing specifically on the bifurcation of data into core and exhaust data. (The Appendix provides additional details.) Each step is described in detail below. The framework is continuous because it reflects the need for an iterative approach to the capture and use of exhaust data. At any point in the framework, the analysis could be either data driven or purpose and need driven, depending on where the organization is in the innovation or organizational evolution process.

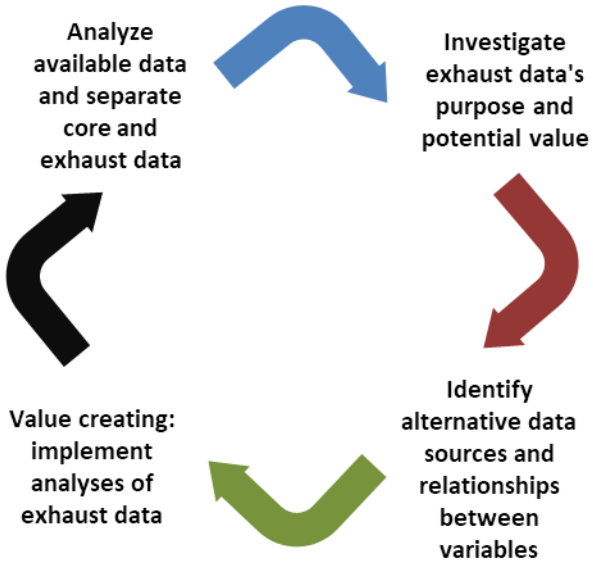

Figure 1. Framework for Exhaust Data Value Creation

3.1 Analyze Available Data and Separate Core and Exhaust Data

This step is data driven, analyzing whatever data is available for potential use, whether core or not. Identifying available data can require direct analysis of the data, not analysis of a high-level statement of what the data is intended to be. Rather, it is a "hands-on" investigation into actual examples. For example, Stolen911 was a website where users report stolen household (or other) goods. Sometimes users "over disclosed" the information requested by the web site, creating "extra information" beyond the core data required by the site that might be useful in some other settings (O'Leary & Storey, 2017a). In another example involving Alzheimer's brain research, researchers analyzed (usually discarded) exhaust data. As noted by Healy (2018, p. 1), "[t]he data they mined is usually discarded, but was archived instead by the National Institutes of Health in a bid to accelerate the discovery of new treatments by fostering 'big data' collaborations". Another source of usually discarded exhaust data is search engine queries or analysis of machines on the Internet of things (Winnig, 2016).

At this step, to the extent feasible, core data should be categorized and separated from the rest of the data. It can help to ask: "What data is part of the core process and what is supportive of the core process?". Unfortunately, distinguishing between core and exhaust data and finding potential data exhaust may not be easy, but the notion of core data can help facilitate the analysis. In Blow-Up, the photographer had some intuition, since the behavior of the people involved was anomalous, suggesting that he look beyond the core picture. In the analysis of LinkedIn data above, the difference was between the original "intended" use of staying connected and the "unintended" use of finding customers of another firm. In the case of the Ring doorbell, the difference was between the data generated on the app about the front porch, versus using the occurrence of the app with other apps to capture a device fingerprint. Thus, identifying important, and meaningful, exhaust data may not be obvious, and likely requires cleverness and insights, thus creating a clear challenge.

We add some additional questions to facilitate additional “out-of-the-box” thinking, which we summarize below:

- What available data is not on anyone’s dashboard?
- What data is not on anyone’s list of incentives?
- Is all location data being gathered also being used effectively?
- Does data gathered directly from users contain "over disclosures"?
- Is there any “temporary data” that is discarded?
- Is user "query data" (from databases, Internet, etc.) gathered and used?
- Does new technology create new data? Is that data available and used?
- Is there any log data that is not being used?
- What program to program data is not being used?

3.2 Investigate Purposes and Potential Value of Exhaust Data

This step is purpose and value oriented, so the focus aligns with the core purposes, knowledge, systems and data. In some cases, organizations may wish to start with this step. However, those seeking innovation and evolution, seeking to break away from their core, might wish to start with the previous step.

The need for innovation or the ability to facilitate the organization’s evolutionary path can determine the fit of the exhaust data with an organization’s needs or objectives. Thus, this step becomes purpose, need, or theory-oriented, in determining potential value. Because the previous step is data oriented and not focused on any specific objective, any new data from that step might not be aligned with the core business, or even be interesting to the analysts, or consistent with their overall objectives. For example, accountants might realize that the LinkedIn data is a list of clients, but they might not be interested in that list. In the Alzheimer’s data, previous researchers had not been interested in the archived data. As another example, knowing that different devices have different app fingerprints will not be useful to all firms. In summary, this step helps to guide the analysis towards deriving value for a firm.

Since the previous step is data driven, we need to ask: “What experiences or issues of value might these variables affect in our organization?”. This may require a multiple person team to represent different points of view. For example, some might be security experts, whereas others might be marketing experts or experts from other domains, such as health and medicine. At this step, the biggest challenge is likely to be finding alternative purposes or repurposes for the data that are useful within the context of the specific organization. In some cases, this can generate an opportunity to provide data to others. For example, potentially, a business could be built around analyzing LinkedIn networks and capturing information about the links, whether they are sales opportunities, consulting opportunities, etc., and selling that information to others. As another example, phone location information can be provided to those interested. For example, recently, federal agencies bought access to a commercial database and used phone location information to facilitate border enforcement (Tau & Hackman)

Thus, the analysis of exhaust data can generate uses that are beyond the focus of an organization and maybe even result in new businesses. Identifying new uses may require a team that represents multiple points of view to link the discovered exhaust data to core purposes or evolving organizational needs. Some additional questions include:

- Can we repurpose existing data to help create value?
- Is there a way to provide structure to data not on dashboards?
- What dependent value variables might these data affect?

3.3 Identify Alternative Data Sources and Relationships between Variables

After finding and matching new data to potential needs and purposes, the next concern is to identify the data sources and find relationships between the data. This raises several potential questions:

- What other data do we need?
- What relationships are in the data?
- What theories do we have to drive the analysis?

Up until this point, the potential data use has been speculative and the link to a purpose or potential use not fully established. Therefore, the purpose of this step is to find the appropriate data platforms and empirically investigate the relationships among the data.

We need to manipulate the data using data mining, in an attempt to discover useful relationships in the data. Throughout the process, managers need to use their knowledge, intuition, and creativity to identify potential relationships: different domain expertise could result in different findings. Potential relationships could then be coupled with relationships gathered using other approaches. Thus, this step could require substantial data science modeling efforts, as well as human investigation and insights, in an effort to discover unexpected relationships and patterns. There are many approaches to exploring all of the potential relationships, which depend on the structure of the data and other concerns, so complete enumeration is unrealistic.

This step could include linking other databases to the original data to ascertain or understand interesting relationships. Technologies, such as the big data lake (O'Leary, 2014), may facilitate finding and implementing such patterns and relationships between those different databases: are the variables unique and are the relationships or patterns discovered actionable? If they are not, then the firm may need to go beyond their own data and invest in other databases. The main challenges include retaining a balance between management input and data scientist analysis and finding approaches to focus the analysis on value creation in the next step. Some additional potential questions are summarized below.

- If there was a big data lake, could users identify additional useful data and relationships?
- Is there an additional data set that would allow value creation?
- Is there a way to have users experiment with data to help study these relationships?
- Can machine learning provide any unseen or undetermined relationships?

3.4 Value Creating? Implement Analyses of Exhaust Data

After management has identified and studied a range of variables, management needs to investigate implementing the recommendations. At this step, there are several different approaches that could be used, such as prototyping, but detailed technology implementation is beyond the scope of this paper. In any case, managers need to identify what types of investment are required. Managers must ensure that the analysis is performed within a context of generating value, focusing on key problems and opportunities at the business level, not just because there is a perceived "cool" (emerging, interesting, or engaging) technology. Some additional questions include:

- Are the new variables/data useful?
- Are there any relationships between value measures and exhaust data?

3.5 Multiple Iterations through the Framework

Multiple iterations through the framework will most likely be necessary, because of the nature of the process of discovery, the search for innovations, and the nature of exhaust data. This makes a cyclical, rather than linear, process. However, any single pass through the cycle would likely be linear. Furthermore, one might go backwards (e.g., "investigate purposes…" to "analyze available…") if the exhaust data being examined does not have critical or valuable purposes.

A key challenge of multiple iterations through the framework is the perception that "we already considered this data and did not discover anything useful, or we have already found all that we are going to find". One way to circumvent this argument is to focus on other non-core data sets.

4 Sample Exhaust Data Applications

We apply the framework for exhaust data value creation to four examples: Internet searches, accounting entries, social media disclosures, and Electronic Data Gathering, Analysis, and Retrieval (EDGAR) [3] use logs. The examples demonstrate: how core and exhaust data interact, how value is created from exhaust data, how exhaust evolves to core data, and how users might mitigate potential disclosures of some exhaust data. For each example, we examine the steps of the framework and identify potential approaches to mitigate information disclosure.

4.1 Case 1: Internet Searches (Google Flu Trends)

People, typically, generate an Internet search and then forget it. However, Internet searches can be saved, and aggregated so information in them can be used to make inferences. Associations and relationships between search terms and concepts of interest can be used to generate inferences about queries and, potentially, about the person making the search query. For example, search queries for information about the flu might include fever, cough, sore throat, runny or stuffy nose, muscle or body aches, headaches, fatigue (tiredness) vomiting, and diarrhea (Centers for Disease Prevention and Control, 2019).

Perhaps the best known of such efforts is Google Flu Trends (GFT), initiated in 2008 (see https://www.google.org/flutrends/about/). GFT contained information about the date, country, and the number of flu-based queries according to a geographic unit (e.g., state or city) in that country. Although Google no longer publishes information about GFT, the initial reported results of GFT were quite good (even impressive), adequately predicting doctor visits from flu searches (Ginsberg et al., 2009).

Google is not the only organization to use search data. Researchers have also used search requests, or proposed using them, in novel ways (Ginsberg et al., 2009). For example, researchers have linked Google search terms to housing sales and prices of real estate in the United States (Brynjolfsson & McAfee, 2015). As another example, researchers have proposed using search for other health concerns, including food poisoning (Brownstein, Freifeld, & Madoff, 2009). Recently, data from Wikipedia searches, Google searches and Google Trends were used to forecast coronavirus hospitalizations and deaths from COVID-19 (O'Leary & Storey, 2020).

4.1.1 Analyze Available Data and Separate Core and Exhaust Data

The minimum core data associated with an Internet search is simply the specific search terms of interest. The creativity was treating the queries as a database, thus saving something that could have been discarded. Researchers have also demonstrated this step in analyzing saved query information from over 50 million queries to build a GFT model (Ginsberg et al., 2009)

4.1.2 Investigate Purposes and Potential Value of Exhaust Data

GFT provides the potential for an early warning system for flu and other diseases, which could be used in public health settings. the approach that generated GFT was also investigated by Google to study dengue fever (Google Dengue Trends (GDT)), which researchers used to create accurate forecasts of the expected number of new dengue fever cases (Strauss, Castro, Reintjes, & Torres, 2017). Clearly, search data have value creating capabilities.

4.1.3 Identify Alternative Data Sources and Relationships in Data

GFT researchers matched 50 million queries to 1152 doctor visits from the Center for Disease Control and Prevention (CDC), requiring integration with another database. Although GFT researchers found interesting and statistically significant relationships, their use of those two databases has been criticized for a lack of granularity alignment between them (Lazer, Kennedy, King, & Vespignani, 2014). Accordingly, database choice is an important issue. The development of the GFT model has also been criticized for overfitting, a lack of transparency (limited information about query words used), and the existence of other good predicting models (Lazer et al., 2014).

[3] See https://www.sec.gov/edgar/searchedgar/legacy/companysearch.html

There is limited information about the actual query words investigated in GFT, but the search terms could include the term “flu” or refer to a set of flu-like symptoms as noted above. A process to identify additional relationships could be studying the co-occurrence of terms with the primary search term, say “flu”.

4.1.4 Value Creating? Implement Analyses of Exhaust Data

The GFT showed that the flu diffused at a different speed and to a different extent in different geographic units. Such information could be publicized and made broadly available, and, thus, be useful public health information. There are other potential uses of this approach besides understanding disease diffusion. For example, it could be used to measure the diffusion of knowledge about technology awareness, product awareness, or restaurant awareness. How one implements and uses [what] would depend on the resulting findings. For example, Internet searches have been used to find the most frequent food queries by state (Sweeney, 2014). This information could be valuable for Restaurants for investment and diversification decisions.

4.1.5 Mitigating Potential Disclosure

It is difficult for a user to mitigate inferences from search data. Users typically can limit disclosure of their location using access approaches, such as a VPN. However, queries can still be retained, and inferences made from them.

4.2 Case 2: Accounting Entries, Auditing, and System Security[4]

Accounting entries relating to the debit (e.g., cash) and credit accounts (e.g., accounts receivable) help companies keep track of core financial information. However, exhaust information gathered at the time of the transaction can be used in auditing and system security. This application is included because it has become so integrated into normal systems activity that what was originally exhaust data in this application now is likely widely regarded as core data, especially for security and audit purposes (Tenor, 1988). Regardless, our framework is still applicable.

4.2.1 Analyze Available Data and Separate Core and Exhaust Data

Core financial accounting transaction data includes the debit/credit entry, the other companies involved in the transaction, and the date. This core data enables companies to keep track of the basic financial information they need to run their business and meet regulation requirements. However, there is additional potential information in the form of exhaust data that can be gathered with each transaction and used for other audit and security purposes. For example, for each transaction, additional data could be captured about the person entering the transaction (e.g., user name), the normal role of that person (e.g., manager, accounting clerk) the location (e.g., office, elsewhere), time of day, and so forth.

4.2.2 Investigate Purposes and Potential Value of Exhaust Data

The additional data (e.g., who, what time of day, etc.) can help provide additional insights, such as “who is using the system from where and when?” Profiles of expected user behavior can be generated and used to assess whether actual data meets expectations. Analysis of tuples including user, location, date and time can be used to ascertain, for example, that a manager has entered a transaction in a computer he or she does not normally use at an anomalous time, such as a Sunday evening. Analysis of that exhaust data can be used to assess whether usage is expected or anomalous, and to facilitate security analysis and auditing of system access and generated transactions, to provide reliability to the financial data.

4.2.3 Identify Relationships in Data

Data mining can be used to analyze the set of debit and credit entries to find potentially anomalous pairs of transactions. For example, accounts receivable are typically matched to either cash or sales. If accounts receivable is matched in an entry with other accounts, then those entries are likely anomalous and deserve a deeper analysis. The original core purpose was to include the entry in the books of the firm,

[4] Some users may now see this application as using core data, however, at one point. All of the listed data except for the accounting entry was simply exhaust data.

but the exhaust data purpose was to discover unusual relationships. Thus, the same data can have both core and exhaust uses.

Further, with additional information of "who" made the entry, data mining can be used to find a distribution of the number of entries made by each employee. Such data might be used to examine the tail of that distribution to discover who only makes a few entries, leading to questions such as "why" they made those entries. The exhaust data can also be used to find a distribution of the time of day the accounting entries were made to guide the identification of potentially anomalous entries and who made them.

4.2.4 Value Creating? Implement Analyses of Exhaust Data

Additional data, not core to the financial results, can provide information about the system's security, leading to an improved ability to audit the system. Exhaust data captured as part of the transaction can be used to both help prevent and detect fraud. Usual participants to the debit/credit process can be isolated (e.g., those with a limited number of transactions or those with roles that generally do not make accounting entries). Entries generated at unusual times and locations can also be isolated. Generating these lists of anomalies can help detect audit risks. The continued gathering of such data can facilitate the system's security and quality of the information, providing substantial value. As a result, this approach to investigating accounting relationships and the analysis of that data is becoming embedded in emerging audit analytics approaches at larger accounting firms.

4.2.5 Mitigating Potential Disclosure

One approach to mitigating potential disclosure is to "masquerade" as someone else, for example, using another's login credentials. Alternatively, someone might try to access log data and change that data to cover up (hide/mask) their inappropriate accounting entries.

4.3 Case 3: Social Media Disclosures

There are many different types of social media, with new approaches continually being developed. Each type of social media is likely to have its own potential exhaust data. This section provides an overview of general issues, such as over-disclosure, in social media. The core question of Twitter has been "what are you doing?" or "what's happening?" Reportedly, Facebook "was built to accomplish a social mission—to make the world more open and connected" (Chaykowski, 2017). Social media users disclose substantial information. Some disclosures can provide deep insights that might be used by others to create value for themselves.

4.3.1 Analyze Available Data and Separate Core and Exhaust Data

Posts can take the form of text or pictures with each post providing the potential for exhaust data. Text posts contain sentiment and other information. Picture posts can provide unstructured data for insights: "What do they look like?", "Who are their friends?", "Where do they live?", "What is their standard of living?" or "Are they home right now?". There are numerous reported instances of people disclosing information that allowed inferences that were undesirable. Floyd Mayweather is a famous boxer whose house was robbed after he posted pictures of himself on the Great Wall of China (Charles, 2017). An Australian family was robbed after their teenaged daughter posted a picture of a large quantity of cash (Golijan, 2012). A California family was robbed after posting pictures of vacationing in Las Vegas (Suter, 2014).

4.3.2 Investigate Purposes and Potential Value of Exhaust

Social media exhaust data has many other potential uses, including marketing, robbery, identity theft, and election fraud. The choice of purpose may be limited only by the creativity of the investigator. For example, Cambridge Analytica apparently was interested in information about potential voters and concerned about gaining insights into voters in the 2016 United States presidential election, and possibly, even affected the voter's choices (Rosenberg & Dance, 2018). Although Cambridge Analytica employed and analyzed a mix of core and exhaust data, unfortunately, some of these purposes violated personal privacy and raised other concerns.

4.3.3 Identify Alternative Data Sources and Relationships between Variables

Social media data can be enriched with additional databases to generate insights. For example, each of the instances where someone was robbed effectively required that the robber has access to the address and additional information such as "is anyone watching the home?". Cambridge Analytica, reportedly, used Facebook to gather data about users and their friends (Rosenberg & Dance, 2018). Users were paid a few dollars to take a psychological-based quiz. However, to take the quiz, they were required to log into their Facebook accounts. Then, the researchers had access, not only to the user's Facebook information (name, birthday, etc.), but also to the same information about their friends, ultimately generating and linking multiple databases. The full extent of the data gathered by Cambridge Analytica is unclear but, most likely, includes both core and exhaust data.

Social media also can be mined for the sentiment of text and other issues. It is not unusual for Companies to mine blogs or other types of social media-based information to determine whether an expressed sentiment is positive or negative in order to determine attitudes toward different entities.

4.3.4 Value Creating? Implement Analyses of Exhaust Data

There are many instances where information gathered from social media is used to create value (e.g., Mount & Martinez, 2014). Unfortunately, these examples are not always for the betterment of society or the user. Robbery and voter manipulation are simply some of the uses of exhaust that create value for others from social media disclosures.

4.3.5 Mitigating Potential Disclosure

Mitigating the impact of exhaust data may be a matter of timing (post pictures after returning home from vacation) or discretion (limit disclosures that provide insights). However, the approach used by Cambridge Analytica illustrates that the actions of any one user can affect a broad range of users from their network of social media friends, resulting in a cascade of problems.

4.4 Case 4: EDGAR and the EDGAR Use Logs

EDGAR is a government-mandated effort to make financial information (e.g., Form 10-K and other disclosures) about publicly traded firms broadly available. EDGAR use logs (EUL) are a source of exhaust data.

4.4.1 Analyze Available Data and Separate Core and Exhaust Data

Although EDGAR provides core financial data, EUL provides exhaust data that relates to the use of financial information by different users. At their heart, the extensive and broad-based use of EDGAR data proves that the available EDGAR resources are useful and provide a service to taxpayers and stock market participants. Thus, the EUL substantiate the Government's investment in the Web-based system. The use logs include exhaust data on which IP address downloaded which financial information resources (e.g., 10-K) and the time at which those resources were downloaded.

4.4.2 Investigate Purposes and Potential Value of Exhaust Data

The EDGAR use logs provide additional value by capturing usage data. The IP address could be used to identify a user (e.g., the investment firm) and the activities of that user at a specific time. There are many financial documents with each likely to have different purposes because they contain different information. In any case, use of specific financial information and frequency of use of specific documents can be indicative of "interest" in a specific company, perhaps for acquisition or for stock purchases.

4.4.3 Identify Alternative Databases and Relationships between the Variables

If the EUL IP address is interfaced with another database that relates IP addresses to a company or user, then it is possible to discover who has been examining which company's financial statements. User information can provide insights into which investors are considering investing in a company, or which companies are potentially planning to acquire or merge with other companies. If available in real time, the use logs provide inside information about competitor investments.

4.4.4 Value Creating? Implement Analyses of Exhaust data

Users (investors, professional investing firms and others) likely are examining EDGAR data for potential investment purposes. Downloading documents may signal that a merger or acquisition is under consideration and the log indicates which firms are considering investment. Knowing who is considering investing in which companies can provide unique investment opportunities. Correctly anticipating mergers and acquisitions can provide substantial wealth to a stock market investor, by anticipating changes in stock prices that reflect the impact of the merger.

Companies that provide financial information to others might monitor and mine the use data that users download from company websites. Interestingly, based on this analysis, it appears that almost any data vendor could make a similar use of their own log of user activity in their databases.

4.4.5 Mitigating Potential Disclosure

How might a firm camouflage their EDGAR use? One approach is to use a VPN (virtual private network) service. Alternatively, analysts could use a different data source or go to the actual company to obtain the necessary financial information.

5 Discussion: Managerial Challenges and Implications

There are many opportunities for creating value from exhaust data. Generating exhaust data applications might be costly, but the potential exists to gain large benefits for both innovation and organizational and system evolution. Applications developed from exhaust data are not likely to be immediately imitable by others because of the need for appropriate systems, knowledge, and processes. The result could be sustainable value creation. However, there are associated challenges and implications.

5.1 Managerial Challenges

Based on the above cases, a set of challenges can now be described, which may need to be overcome before value can be realized from exhaust data. We assume that core data is associated with core applications, and that exhaust data can provide opportunities to innovate and facilitate organizational and system evolution. As a result, we articulate the challenges and implications for how exhaust data might be used to benefit managerial decision making. In these settings, we might observe how exhaust data evolves into core data.

5.1.1 Challenge 1: Creativity

The most striking observation from the cases throughout this paper is the important role of creativity. Table 2 summarizes notable creative ways in which exhaust data was captured in the case studies. In each of the examples, creativity most likely required substantial persistence and effort, suggesting that managers devote sufficient resources to facilitate such persistence and effort. Unfortunately, it is almost always difficult to systematically generate creative solutions.

Table 2. Exhaust Data Characteristics

Case	What we thought evidenced creativity
GFT and GDT	Uses search queries as a database, integrating additional databases with the query databases and mining the relationships.
Analysis of Accounting entries	Captures user data with transactions which provides ability to monitor characteristics of system use and provide insights by isolating usage anomalies
Twitter and Stolen911	Provides range of opportunities, based on user disclosures, under-disclosures or over-disclosures. Can gather unintentional disclosures from video, pictures, or text.
EDGAR user logs	Facilitates analysis of information concerns of different users. Analyzing that data can enable extrapolation about what users are trying to accomplish and can create value from monitoring user activities. For example, can identify stocks that firms are investigating, potentially signaling a merger.

5.1.2 Challenge 2: Technology-generated Data Varies by Domain and Application

The capture of core and exhaust data varies by technology and domain. In social media, much data is generated directly from users, whereas on the Internet of things, machines generate and capture data that users may not even know is being gathered (e.g., location). In addition, as new technologies emerge, there is likely to be different exhaust data associated with those technology uses. As examples, with phones, there is exhaust location information, whereas with LinkedIn, there exists the potential to determine company client information. Thus, each new technology will require different technology expertise.

Exhaust data in one domain may not be "directly applicable" in another. As noted above, LinkedIn data may allow access to a company's customers, but if accountants analyzed it, they might not recognize that possibility, or even be interested in the data. Different applications, domains, and industries each generate different kinds of exhaust data. Internet search applications generate data that is different from usage log applications. Exhaust data in accounting applications is likely different from exhaust data in marketing applications. Therefore, experts from different domains and industries are likely needed to help identify and fully develop exhaust data applications, recognizing that exhaust data applications in one domain may not be directly applicable in other domains. Hence, different domains are likely to require domain expertise.

5.1.3 Challenge 3: User-generated Data

Since the emergence of Web 2.0 and as we move beyond Web 2.0, users are increasingly responsible for providing data to systems. This is consistent with classic reengineering design concepts and other trends (Hammer, 1990). Further, there is hardly an application where customers do not provide at least a portion of the data. Unfortunately, it can be difficult for firms to control the quality and quantity of user-provided data (O'Leary & Storey, 2017a). Users may not fully anticipate the extent to which the data they disclose reveals information about them. For example, a vacation picture posted on Facebook announces to the world that the user is not at home, so perhaps access to their home, automobile, and other assets may now be more easily available. However, users may disclose more than just personal information. Such disclosures could relate to the organization for which they work.

5.1.4 Challenge 4: Interfacing Multiple Databases

Many of the uses of exhaust data interface multiple databases to create value. For example, discovering a company's clients from LinkedIn connections may require additional search and information about each of the connections. As another example, use logs typically require matching information in some other database to determine information about the users. As still another example, to leverage information disclosed in social media, additional databases are often required, such as an address database. Further, as seen with Google Flu Trends, some of the creativity needed to generate exhaust data applications may come from knowing about database availability and matching the granularity in the exhaust data to the granularity in other databases. Finally, the ownership or costs of obtaining these databases, and potentially even access to the exhaust data, may be an issue (Davenport, 2014).

5.1.5 Challenge 5: Privacy

Exhaust data may generate privacy issues since the data may be repurposed in ways that the originator of the data probably did not anticipate. Unfortunately, this repurposing is contrary to the notion that the purpose of any data collection should be identified in advance, and an organization should stick to that purpose. Different privacy frameworks could be applied to help clarify and manage privacy, including the European General Data Protection Regulation (GDPR).

5.2 Managerial Implications

It is, generally, difficult to identify exhaust data because it is unknown in advance, will vary depending upon the application, and exhibits the other characteristics we identified. Table 3 summarizes a number of different applications. Based on these examples, and the cases described above, several managerial recommendations emerge.

Table 3. Summary of Exhaust Examples

Information source	Core data	Exhaust data
Accounting entries	Accounting information—debits / credits, account numbers, dates and amounts	Can be used for security purposes, by capturing and using information about who, when and where, that can provide anomaly information
Automobile identification	Rental car identification number for billing purposes	Thieves can identify which cars are rentals so as to identify Los Angeles visitors who will not come back to testify about car break-ins
Email	Communications to /from, including dates, “amount” of text, “sentiment” of text, device used	Personal networks can be identified for frequency and depth of communication; devise use can capture location (away from computer) and other issues.
Google searches	Number and location of real estate searches	Suggest how the housing market in a particular area is behaving
	Number, location and type of symptoms or how to fix a specific illness.	Can be used for capturing flu trends, dengue fever, food poisoning, allergies, etc.
	Number, type and location of job searches	Suggests the condition of economy and status of specific industries
Parking spots used	Number of spaces available in some place at a particular time indicates availability	Number of spaces at an airport can identify demand for air travel
		Number of spaces at a shopping mall can identify demand for goods at stores
		Number of spaces at a car dealership or manufacturer, indicates the number of cars available for sale or sold, suggesting the quality of the current market
		Number of spaces on city streets can help identify existence of an event
Phone logs	Communications to/from, dates, number of communications, length of the calls, indicates activity	Personal networks and depth of relationships; device use can indicate the extent of use of different modes of communication or mobility
Phone location	Cost, feasibility, location and time of the call	Aggregate number at some location / time may help choose location for hotel, restaurant, etc.
Pictures	Core Image	Images at fringes or accidental images can capture other events or locations
	Date, time and location	Aggregate information for inferences about how many people are located in any one place, at one time, and what specific identity is of most interest
Social media	Events: what are you doing?	Can disclose location or vulnerability (no one else around, cash, not at home, etc.)
Ring Door Bell	Video and interaction with those at front door	Provides an identifier of user's mobile device based on portfolio of apps on phone
Use of Internet Information about companies	EDGAR use logs	Who is using what information, can suggest takeover attempts, short sales, or long sales
	Financial company information	Depending on who is the downloading, the information downloads can suggest takeover attempts, short sales, long sales

5.2.1 Recommendation 1: Be “Data Curious”

At one level, this paper suggests a data-driven approach to generating innovations: analyze your data, assess whether it is being used for all of its intended core purposes, and ask “are there any other uses of this data?”. Data curiosity appears to be at the heart of many of the examples, starting with the opening example of the photographer in Blow-Up. Exhaust data might be used to solve an existing problem (what is the progress of the flu?), provide a solution to an unknown problem (generate a device fingerprint based on apps) or lead to a new business (commercial database of phone locations).

5.2.2 Recommendation 2: Understand What Data is Actually Available

Consider the actual data available and attempt to analyze it. Sometimes the actual available data may be different than a manager's expectations. When users input the data, do not assume they have provided the expected data and do not expect the controls over the data to have worked exactly as expected. In the Stolen911.com data, users often provided more or different information than they were asked, frequently over disclosing (O'Leary & Storey, 2017a). Thus, it is important to actually examine the data, rather than descriptions of it.

5.2.3 Recommendation 3: Review and Analyze Other Examples of Exhaust Data

The mathematician George Polya (2004), in his book *How to Solve It*, suggested that one effective approach to solving a problem is to identify similar, solved problems and the approach used to solve them. Developing exhaust data solutions are likely similar, although they may vary by industry. Table 3 includes four examples of how parking space information can provide unanticipated insight through data exhaust, each from a different setting.

5.2.4 Recommendation 4: Become an Expert in Privacy Regulations

Creating value from exhaust data often requires using data differently than originally anticipated. Using data "differently" has been a privacy concern for much of the digital age. For example, privacy agencies typically suggest that information particularly, health-related or personal) be used only for the "same purpose for which they collected that information".

Although researchers in big data and analytics consistently have been concerned with privacy issues, the implementation of Europe's GDPR in May 2018 resulted in a new, recognized importance of information privacy. Since the GDPR regulations are new, there may be limited knowledge about the required privacy restrictions. Further, GDPR is not the only set of privacy concerns. Its implementation has led others to begin to consider such issues in detail. Therefore, a chief privacy officer should be consulted when assessing the feasibility of new, exhaust data applications. As an example, the privacy issues associated with the Ring Doorbell have garnered the attention of a number of observers (e.g., Schwarz 2020).

5.2.5 Recommendation 5: Keep an "Exhaust Data Mindset"

Potential exhaust data applications are everywhere; if you want to find them keep looking. On a visit to a McDonalds' restaurant one of the authors received a receipt stating "eat-in" stamped with the date and time. Although we are not deeply familiar with McDonalds' information systems, it is easy to anticipate that the core use of the "eat-in" tag was to tell employees that the order would go on a tray, not in a bag to go. However, that exhaust data could be used to analyze the demand for seating throughout the day and potentially investigate if, or when, to expand/decrease restaurant capacity.

6 Summary, Contributions and Extensions

We present a framework for exhaust data value creation that exploits the basic differences between core and exhaust data and uses those differences to help recognize characteristics of exhaust data-based applications. Managers can use this framework to seek out additional uses of exhaust data, including identifying alternative data sources and relationships among variables. We applied the framework to four case studies to illustrate the value and potential innovations that exhaust data can provide. We also identified challenges and recommendations facing managers in the pursuit of discovering how to derive value from exhaust data.

Organizations typically focus on their core capabilities, using core knowledge, systems, processes, and supporting business intelligence. These efforts to process and analyze core data are captured in core data systems and managed using those systems. However, focusing on core capabilities can result in core rigidities, limiting innovation and evolution opportunities. One approach to break away from core rigidities is to investigate opportunities associated with data exhaust. Analysis of data exhaust can ultimately generate new uses for the data, turning exhaust data into core data. As a result, this paper suggests a data-driven approach to generating new innovations.

6.1 Contributions

This paper has a number of contributions. First, we summarize some of the previous literature from previous exhaust data applications, gathering information from multiple sources. Second, we introduce the core competency literature into the exhaust data and information systems literature. The core competency literature allows us to address issues such as “core data” and “core data systems”. Third, we introduce the notion of using exhaust data as a tool for innovation, to mitigate against issues such core rigidities. Fourth, we provide a framework designed to facilitate finding exhaust data applications. Fifth, we review four case studies that provide insight into innovations associated with exhaust data. Sixth, we investigate some potential challenges associated with trying to use exhaust data and provide some recommendations designed to facilitate the capture of exhaust data innovations.

6.2 Extensions

There are a number of ways that we can extend this research. First, additional case studies illustrating exhaust data in different settings with different technologies in different domains potentially could illustrate additional concepts in exhaust data as shown in a COVID-19 analysis (O’Leary & Storey, 2020). Second, empirical research could be used to investigate the extent to which our theory of core data and theory of core data systems affect the core competence capabilities and access to data exhaust in business firms. It is easy to imagine that a lack of such a focus on core data systems could make it difficult for firms to maintain and grow their core competence. Third, case studies of the use of our framework could help the framework grow and evolve. Fourth, as firms increase their use of exhaust data, researchers can capture the additional challenges and opportunities that they faced. Finally, perhaps additional theories could be generated beyond our use of the management theory of core competency, to help guide further research. Such theory use could elicit additional issues, concerns and opportunities of exhaust data.

Acknowledgements

We thank the review team for their valuable feedback on our work and for their most timely effort in processing our paper.

References

Banno, R., Takeuchi, S., Takemoto, M., Kawano, T., Kambayashi, T., & Matsuo, M. (2015). Designing overlay networks for handling exhaust data in a distributed topic-based pub/sub architecture. *Journal of Information Processing*, *23*(2), 105-116.

von Bertalanffy, L., & Rapoport, A. (1956). General systems. In L. von Beralanffy & A. Rapoport (Eds.), *General systems: Yearbook of the society for the advancement of general system theory* (vol. 1, pp. 1-10). Ann Arbor, MI: The Society.

Boulding, K. (1956). General systems theory—the skeleton of science. *Management Science*, *2*(3), 197-208;

Brownstein, J., Freifeld C., & Madoff, L. (2009). Digital disease detection—harnessing the Web for public health surveillance. *New England Journal of Medicine*, *360*(21), 2153-2157.

Brynjolfsson, E., & McAfee, A. (2015). The digitization of just about everything. *Rotman Management.*

Centers for Disease Control and Prevention. (2019). *Flu symptoms & complications*. Retrieved from https://www.cdc.gov/flu/symptoms/symptoms.htm

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *CRISP-DM 1.0.* Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Chaykowski, K. (2017). Mark Zuckerberg gives Facebook A new mission. *Forbes*. Retrieved from https://www.forbes.com/sites/kathleenchaykowski/2017/06/22/mark-zuckerberg-gives-facebook-a-new-mission/#2ea10d3e1343

Charles, D. (2017). Floyd Mayweather got paid $3 million to visit China, shared pics, then his mansion got robbed. *Brobible*. Retrieved from https://brobible.com/sports/article/floyd-mayweather-mansion-robbed-china-visit/

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165-1188.

Davenport, T. (2014). Who owns your data exhaust? *International Institute for Analytics*. Retrieved from https://www.iianalytics.com/blog/2014/8/7/who-owns-your-data-exhaust

Davidson, A. (2016). *Big data exhaust for origin-destination surveys: Using mobile trip-planning data for simple surveying.* Paper presented at the Transportation Research Board 95th Annual Meeting.

Davis, S., & Davidson, B. (1992). 2020 vision: Transform your business to succeed in tomorrow's economy. New York, NY: Simon and Schuster.

George, G., Haas, M. R., & Pentland, B. (2014) Big data and management. *Academy of Management Journal*, *57*(2), 321-326.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L, Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014.

Golijan, R. (2012). Family robbed after teen posts photo of money on Facebook. *USA Today*. Retrieved from https://www.today.com/money/family-robbed-after-teen-posts-photo-money-facebook-800882

Hammer, M. (1990). Reengineering work: Don't automate, obliterate. *Harvard Business Review*, *68*(4), 104-112.

Healy, M. (2018). Surprising discovery about viruses and Alzheimer's disease could open new avenues for treatment. *Los Angeles Times*. Retrieved from https://www.latimes.com/science/sciencenow/la-sci-sn-alzheimers-herpes-viruses-20180621-story.html

Hussain, S. (2020). Ad industry seeks to delay new California data privacy law. *Los Angeles Times*. Retrieved from https://www.latimes.com/business/technology/story/2020-01-29/ad-trade-groups-delay-california-data-privacy-law

Johnson, S. L., Gray, P., & Sarker, S. (2019). Revisiting IS research practice in the era of big data. *Information and Organization*, *29*(1), 41-56.

Katz, D., & Kahn, R. (1966). *The social psychology of organizations*. New York, NY: John Wiley & Sons.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Leonard-Barton, D. (1992). Core capabilities and core rigidities: A paradox in managing new product development. *Strategic Management Journal*, *13*(S1), 111-125.

Mount M., & Martinez, M. G. (2014). Social media: A tool for open innovation. *California Management Review*, *56*(4), 124-143.

Ojo, A., & Heravi, B. (2018). Patterns in award winning data storytelling: Story types, enabling tools and competences. *Digital Journalism*, *6*(6), 693-718.

O'Leary, D. (2008). Gartner's hype cycle and information system research issues. *International Journal of Accounting Information Systems*, 9(4), 240-252.

O'Leary, D. E. (2013). Exploiting big data from mobile device sensor-based apps: Challenges and benefits. *MIS Quarterly Executive*, *12*(4), 179-187.

O'Leary, D. (2014). Embedding AI and crowdsourcing in the big data lake. *IEEE Intelligent Systems*, *29*(5), 70-73.

O'Leary, D. E. (2019). Technology life cycle and data quality: Action and triangulation. *Decision Support Systems*, *126*.

O'Leary, D., & Storey, V. C. (2017a). Data exhaust: Life cycle, framework and a case study of Stolen911.com. In *Proceedings of the International Conference on Information Systems.*

O'Leary D. E., Storey V. C. (2017b). Data Exhaust. In L. Schintler & C. McNeely (Eds.), *Encyclopedia of big data.* Cham: Springer.

O'Leary, D.E. and Storey, V.C., 2020. A Google–Wikipedia–Twitter Model as a Leading Indicator of the Numbers of Coronavirus Deaths. *Intelligent Systems in Accounting, Finance and Management*, *27*(3), pp.151-158.

Polya, G. (2004). *How to solve it: A new aspect of mathematical method.* Princeton, NJ: Princeton University Press.

Prahalad, C. K., & Hamel, G., (1990). The core competence of the corporation. *Harvard Business Review*, *68*(3), 79-91.

Rosenberg, M., & Dance, G. J. X. (2018). "You are the product": Targeted by Cambridge Analytica on Facebook. *The New York Times*. Retrieved from https://www.nytimes.com/2018/04/08/us/facebook-users-data-harvested-cambridge-analytica.html

Schweidel, D. (2014). *Profiting from the data economy: Understanding the roles of consumers, innovators and regulators in a data-driven world.* Upper Saddle River, NJ: Pearson.

Schwarz, J., (2020). Ring gets "dinged" for its video doorbell privacy, https://thehill.com/opinion/cybersecurity/485449-ring-gets-dinged-for-its-video-doorbell-privacy

Strauss, R., Castro, J. S., Reintjes, R., & Torres, J. (2017). Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case. *International Journal of Medical Informatics*, *104*, 26-30.

Stone, B. (2009). What's happening? *Twitter*. Retrieved from https://blog.twitter.com/official/en_us/a/2009/whats-happening.html

Storey, V. C., & Song, I. Y. (2017). Big data technologies and management: What conceptual modeling can do. *Data & Knowledge Engineering*, *108*, 50-67.

Suter, L. (2014). Family's home burglarized after posting vacation status on Facebook. *ABC7*. Retrieved from http://abc7.com/archive/9482852/

Sweeney, J. (2014). The most popular unique food-related Google searches by state. *First We Feast*. Retrieved from https://firstwefeast.com/eat/2014/05/the-most-popular-unique-food-related-google-searches-by-state

Tau, B., & Hackman, M. (2020). Federal agencies use cellphone location data for immigration enforcement. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/federal-agencies-use-cellphone-location-data-for-immigration-enforcement-11581078600

Tenor, W. (1988). Expert systems for computer security. *Expert Systems Review*, *1*(2), 3-6.

Urde, M., (2003). Core value-based corporate brand building. *European Journal of Marketing*, *37*(7/8), 1017-1040.

Viscusi, W. K., Huber, J., & Bell, J. (2011). Promoting recycling: Private values, social norms, and economic incentives. *American Economic Review*, *101*(3), 65-70.

Winnig, L. (2016). GE's big bet on data and analytics. *MIT Sloan Management Review*, *57*(3).

Appendix: Methodology for Developing Framework for Exhaust Data Value Creation

There are a number of frameworks for business intelligence and data mining designed to facilitate analysis and create value from data (Chapman et al., 1999; Chen, Chiang, & Storey, 2012; Storey & Song, 2017). However, since bifurcation of data into "core" and "exhaust" is relatively new, there are no directly applicable frameworks that could be adopted to structure our analysis. We, therefore, developed the framework for exhaust data value creation (Figure 1). We used general systems theory (Boulding, 1956; von Bertalanffy & Rappaport, 1956; Katz & Kahn, 1966) to derive our framework. However, our framework also has a data science / big data focus on distinguishing between two types of data: core data and exhaust data. The framework evolved as we used it to investigate core versus exhaust data in different settings. The framework is general enough (and its basis general enough) that it could continue to evolve as our knowledge of exhaust data evolves.

In part, the framework in Figure 1, uses a data-based approach ("analyze available data…") that typically starts with analyzing existing data, in order find additional data or alternative uses for some of the data, to find innovative values/uses in data, core or exhaust. This approach is based on the notion of core competency, which suggests that in order to assure alignment of incentives, strategy, and systems, the focus of information systems is on core data, and on the core data that drives those systems.

Potential exhaust data may sit side by side with the core data or may be resident on other systems but requires recognition. For example, there may be an over-disclosure by social media users that can be used for additional purposes beyond the core. Alternatively, the core data itself may be used for a different purpose, beyond core purposes. However, rather than an automated approach, at this point in the exhaust data life cycle we expect both efforts to be human driven.

As in general systems theory, the framework treats finding exhaust data and creating value as a "cycle of events" with "patterns of activities" associated with those events (Katz & Kahn, 1966). Our primary pattern is illustrated in Figure 1. The pattern is sequential and based on the analysis of the exhaust data. However, at this point in the exhaust data life cycle, capturing and creating value with, or about, exhaust data, may be the result of a clever insight. Since the focus of our framework is on value creation, and since the framework is based on the difference between core and exhaust data, those two characteristics (value creation and core vs. exhaust data) drive the framework's usage.

In its life cycle, exhaust data typically starts as "context", "left over", "unintended", or "remnant" data. Thus, in the beginning, developers do not, or cannot, recognize the potential value or uses of the data beyond their original core applications. However, at some point, in an analysis of available data and relationships discovered within that data, someone (likely not the developers of the original application) recognizes an alternative potential use. If that analysis finds a purpose and potential value, then the exhaust data could become core data. At that point, the "intent" for the capture of the data changes, potentially making previously considered remnant exhaust data into core data. For example, initially, "click logs" from Internet applications were not of interest, but, over time, they have become core data in security, marketing and other applications. As seen in the detailed analysis of accounting transaction data, what was, at one point, exhaust data (location, date and time) has now become core security data.

In part, finding useful exhaust data applications can depend on creativity, imagination, cleverness, and even luck. However, as demonstrated in examples, there are existing blueprints that provide various potential related applications. In addition, it appears that keeping an "exhaust data mindset" can keep managers constantly analyzing potential opportunities.

When the potential value is realized, the data shifts from "left over" to core data and becomes "found". Then, the data is re-used for a different purpose—stolen from one use for another activity. As part of this shift, typically, the data life cycle includes a creative analysis and use of the data for a purpose for which it was not necessarily designed. Rather than being useless, the data becomes the basis for value creation.

Our framework captures a potentially critical bifurcation of data into core and exhaust data. Based on the general systems theory literature and that difference, we iteratively generate and use the framework. Case studies are particularly useful to obtain an understanding of technologies and concepts early in the life cycle before there are sufficient data to allow substantial empirical analysis (O'Leary, 2008, 2019). As additional uses of exhaust data become available, such notions can be expected to evolve to obtain a greater depth of understanding.

Our research approach toward developing this framework employed a prototyping-based approach summarized in Figure A1. We developed an initial model by analyzing previous research and examples identifying exhaust data. Our initial analysis was on an example system that employed substantial social media capabilities. In that case, looking at the actual data, rather than the data architectural design led us to find that people using the system provide substantial data beyond what is required or requested by the system. Accordingly, we found substantial amounts of exhaust data that could be used for other purposes. In the next case, we investigate "use logs" to determine if they could be used for other purposes. For our third case we choose to analyze parking place data and see if it could be used to find exhaust data. Additional cases that we examined are scattered through the paper. After examining multiple aspects of the analysis, we made minor revisions and then used the framework on the four detailed case studies, leading to a final minor revision.

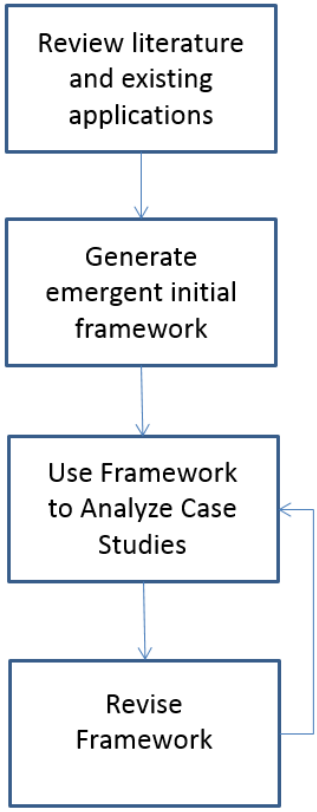

Figure A1. Research Approach

About the Authors

Daniel O'Leary is a Professor in the Marshall School of Business at the University of Southern California, focusing on artificial intelligence, big data and analytics, emerging technologies, crowdsourcing and innovations. He received his PhD from Case Western Reserve University. He is the former editor of *IEEE Intelligent Systems* and current editor of *Journal of Organizational Computing and Electronic Commerce*. His book, *Enterprise Resource Planning Systems*, published by Cambridge University Press, has been translated into both Chinese and Russian. Much of his research has studied emerging technologies and their use in business settings.

Veda C. Storey is the Tull Professor of Computer Information Systems and professor of computer science at the J. Mack Robinson College of Business, Georgia State University. Her research interests are in intelligent information systems, data management, conceptual modeling, and design science research. She is particularly interested in the assessment of the impact of new technologies on business and society from a data management perspective. She is a member of the steering committee of the International Conference of Conceptual Modeling and a member of the AIS College of Senior Scholars.