THE FIRM AS A DATA-GENERATING MECHANISM

By Mark A. DeWeaver*

Abstract

This article shows that when big data is a strategically important resource, traditional accounts of internalization must be augmented with a view of the firm as a data-generating mechanism. I argue that big data is more likely to be a source of competitive advantage if it is generated in-house, implying that optimal firm characteristics will parallel those of valuable datasets. As data-generating mechanisms, firms will be larger and more diversified because when datasets are larger and more diverse, data analysis benefits from economies of scale. Other things being equal, firms will expand and diversify into activities generating datasets that are complimentary in the sense that their combination leads to reduced uncertainty in the results given by operationally relevant supervised learning algorithms.

Keywords Big data, Theory of the firm, Resource-based view, Internalization, Organizational knowledge

1. Introduction

As big data analysis begins to play a central role in corporate strategies for "digital transformation" (Davenport, 2006; Adner *et al.*, 2019; Vial, 2019), the subject is becoming increasingly important for business strategy researchers. During the last few years, a new research agenda has taken shape focusing on questions such as how firms derive actionable insights (Elgendy & Elragal, 2016; Erevelles *et al*., 2016; Ghasemaghaei & Calic, 2019) and dynamic capabilities (Corte Real *et al.*, 2017; Wamba *et al.*, 2017; Lin & Kunnathur, 2019) from big data, its impact on decision making (Janssen *et al.*, 2017; Merendino *et al.*, 2018), the magnitude of the potential benefits (Brynjolfsson *et al.*, 2011), the organizational requirements for realizing them (McAfee & Brynjolfsson, 2012; Davenport *et al.*, 2012; Kiron *et al.*, 2014; Peter *et al.*, 2020), and the obstacles to doing so (Sivarajah *et al.*, 2017). These contributions have demonstrated the importance of the big data opportunity while at the same time emphasizing the fact that vast quantities of data do not in themselves create new competitive advantages. The organization and its culture must be transformed in ways that promote its use. Davenport *et al.* (2012), for example, call for firms to "integrate analytics into core business and operational functions." Janssen *et al.* (2017) stress the need for "relational and contractual governance mechanisms to ensure BD quality and the ability to contextualize data." Peter *et al.* (2020) highlight the importance of "strategic leadership" in "achieving shared understandings of digital transformation processes within and across businesses." Getting the human factors right is clearly at least as important as the technology.

So far, however, relatively little attention has been paid to the implications of the big data revolution for the nature of the firm itself. The literature generally starts from the assumption that managers have access to an exogenously given database from which it is possible, subject to a variety of cognitive and managerial constraints, to derive competitive advantages. Yet the contents of many such databases are endogenously determined by the firm's own operations (Davenport, 2006)—clickstreams from corporate websites, video images from retail outlets, signals from monitoring devices, operating system logs, and so on (Splunk, 2017). The firm's data resources are not only an input into its strategic decision-making process but also a consequence of the activities determined by its prior decisions. When firms expect big data to play a central role in their strategy, we should therefore expect them to develop in ways that promote specific forms of

* Kogod School of Business, American University, 400 Massachusetts Ave NW, Washington, DC, USA 20016, deweaver@american.edu

in-house data generation. Big data can be expected to drive changes not only in the firm's organization and management but in its boundaries as well.

This paper explores the potential for such changes from the standpoint of the resource-based view. I argue that for big data to be considered a resource it must be not only "big"—in the sense of containing large amounts of potentially valuable information, but also "decipherable," which requires not only size but also sufficient diversity for data mining to be effective. For this resource to create a sustained competitive advantage, in addition to being valuable it must also be rare, imperfectly imitable, and nonsubstitutable (Barney, 1991; Braganza *et al.*, 2017). Only data generated by the firms' own businesses is likely to have these attributes (Dierickx & Cool, 1989), implying that publicly available or non-bespoke third-party data will be ruled out in most cases. As Beck and Limbert (2018) put it "Looking for connections among the same sets of material that everyone else has isn't going to help companies win."

If the value of big datasets is increasing in their size and diversity and their status as a source of sustained competitive advantage requires that they originate within the firm, there are clear implications for optimal firm size and scope. The larger and more diversified the firm, the "bigger" and more diversified will be the data it generates in-house. There may also be early mover advantages similar to those associated with other knowledge-related resources (Lockett and Thompson, 2001) for firms that are able to take advantage of higher volume, higher velocity data streams to move up the learning curve more rapidly than their competitors. When firms become more big data dependent, we should therefore expect them to expand and diversify to a greater extent than would otherwise be the case.

The characteristics of valuable big data thus have direct implications for the nature of firms in big-data driven industries. There we may view the firm as not only a "hierarchical" alternative to the market (Coase, 1937; Williamson, 1975), a "nexus of contracts" (Jensen & Meckling, 1976; Aoki *et al.*, 1990), an idiosyncratic collection of productive assets (Wernerfelt, 1984, 2016; Barney, 1991; Connor, 1991), or an "institution for integrating knowledge" (Grant, 1996) but also as a data-generating mechanism whose very operations continuously add to an information-based resource. Like knowledge derived from learning by doing, though unlike many other resource types, the firm's in-house database is a direct result of its own activity. It is dynamic and path dependent, "evolving along its own unique trajectory as a consequence of the firm's unique history" (Lockett & Thompson, 2001).

In the remainder of this paper, I consider this data-generating-mechanism view in more detail. Section 2 describes the relevant attributes of big data, focusing on the characteristics of this resource and the implications for the nature of the firm. Section 3 explores how these attributes affect optimal firm size and scope, arguing that both will tend to be greater for big-data reliant firms. Section 4 shows how enterprise expansion and diversification can make data more "decipherable." I find that a firm will benefit from expanding and diversifying in cases where entropy and forecast error variance in operationally relevant classification and regression models would be subadditive if datasets from its original and potential new businesses were combined. Section 5 presents some implications for the concept of "relatedness" among different firms while Section 6 concludes.

2. The firm as a data-generating mechanism

Articles on big data often begin with a list of 'V's'—attributes that distinguish it from traditional types of data. (Somehow these tend to begin with the letter 'V'.) The most frequently cited are "volume, velocity, and variety," referring to the massive size, real-time frequency, and variety of formats, respectively, that characterize big datasets. Kitchen and McArdle (2016) provide an extensive list including descriptors such as veracity (i.e. reliability), value (for decision making), variability (of meaning in different contexts) and "non-V's" such as exhaustivity (including the entire population rather than a random sample), extentionality

(amenable to the addition of new fields and modifications to existing ones), and scalability (capable of rapid expansions in size).

Many of these terms refer primarily to technical properties requiring nontraditional analytical techniques—for example the "velocity" with which databases are updated or the "variety" found in the "digital exhaust" produced by corporate IT systems. Others concern opportunities to leverage big data across different divisions of an organization—"portability" and "interconnectivity," for example. Gunther *et al.* (2017) define the former as "the possibility to transfer and remotely access digitized data from one context of application to be used in other contexts," the latter as "the possibility to synthesize data from various big data sources." My emphasis is instead on two properties that can potentially make "bigger" data easier to interpret—volume and diversity. As these increase, there may be increases not only in information content but also in the potential for this to be "mined," as will be demonstrated below. I focus not on the distinction between "big" and "traditional" from a computer science or managerial point of view but rather on the consequences of economies of scale in the derivation of information from data using computer algorithms.

There is an important distinction to be made between such information and the knowledge it makes possible. Sobel and Clark (2018, p. 196) characterize information as "static," "subject to a systematic search," and "accessible to all." Knowledge, on the other hand, "represents the discovery of something previously unknown," which implies the involvement of a human being—a knower. It consists of insights about Hayek's (1945) "particular circumstances of time and place" that are relevant to some specific opportunity. Consider, for example, the famous case of Walmart's use of data analytics to predict a jump in demand for strawberry Pop-Tarts in areas about to be hit by Hurricane Frances (Hays, 2004). The empirical fact, discovered by Walmart's analysts, that demand for this item tends to rise when shoppers are expecting a storm, is information. The realization that particular outlets along Florida's east coast needed to increase their stock in the week prior to landfall on September 5, 2004 would be an example of knowledge.

Information in this sense is inherent in the data in a way that knowledge is not. Any two analysts with the same dataset could apply the same algorithm and derive the same information. But the knowledge to which this gives rise would not be the same in all cases—it would vary with the identity and the situation of the information recipient. As Machlup (1980) puts it: "The contents of the information received may be the same as the contents of what is known as a result, but not necessarily so, because the merging of the new inflow with the preexisting stock of knowledge may result in a reordering, restructuring, or revised understanding of the latter...that which I know after the reception of the new information may be different from what I had known before, not in quantity but in kind or structure. As a matter of fact, the contents of the new information may have negated, disconfirmed, refuted some of what I had known before" (p. 57). While information is an objective phenomenon—a function of the data input and the analysis technique, the knowledge any particular person derives from it will depend on subjective factors such as the extent of her previous knowledge, her preferences and priorities, and the nature of her thought processes.

The big data literature sometimes explicitly draws this distinction, but the two are often subsumed under the term "insight," which may refer to either information derived by data scientists or knowledge acquired by decision makers. Yet the size of the dataset is only really an issue for the data scientist. The value of information to the decision maker will be the same whether it is the product of "big" or traditional data. As I am concerned only with the former, my argument necessarily centers exclusively on information resources.

Big data, consisting as it does of nothing more than a series of zeros and ones in computer memory, is neither information nor knowledge. It cannot, contrary to what some have imagined, "speak for itself." "Outside of analytical frameworks and architectures of interpretation, data is dumb" (Schwarz *et al.*, 2019) or, we might equally well say, tacit. There are, in fact, interesting parallels between dumb data and tacit knowledge. Both are byproducts of organizational experience that may serve as inputs to a model in which "concepts become transferable through consistent and systematic logic" (Nonaka, 1994, p. 21), which then forms the basis for useful knowledge. At the same time, of course, unanalyzed data resides in a databank

rather than a human mind. Unlike tacit knowledge, it cannot be used by decision makers in the absence of interpretation but is universally available to anyone authorized to access it.

Data and knowledge assets are thus tacit in different ways. Tacit knowledge, as Grant (1996) notes, is characterized by limited transferability and capacity for aggregation. It is necessarily local in the sense of Hayek (1945) and therefore of limited usefulness for centralized decision making. Its conversion into explicit form typically relies on social interactions (Nonaka, 1994). Data, on the other hand, can easily be transferred within the firm, is by nature relatively straightforward to aggregate, and can be interpreted with statistical methods and provided to individuals lacking direct access to its sources. It is unlike big-data related resources such as the firm's data analytics capabilities and its ability to convert the information it extracts into useful knowledge, which are likely to rely heavily on tacit understandings of analysts and managers.

Grant (1996) defines appropriability as "the ability of the owner of a resource to receive a return equal to the value created by that resource." For both data and knowledge assets this is limited by the difficulty of revealing their value to a would-be buyer without giving them away in the process (Arrow, 1971, p. 152) and by their character as public goods, which makes the idea of transferring ownership rights problematic (Arrow, 1973, p. 12). In the case of big data, it will not necessarily be obvious to outsiders, and perhaps initially not even to the firm itself, what potential the data has to generate useful information. Data may not be useful at all outside of its original context. And it may also be less valuable in the absence of other complementary datasets retained by the firm, which may be examples of what Teece (1980) refers to as co-specialized assets. A subset of a bank database that did not include any cases of fraud, for example, would be of little use in developing a fraud-detection algorithm.

A view of the firm as a data-generating mechanism follows directly from the fact that data from the firm's own operations is transferrable and aggregable (thanks to its "portability" and "interconnectivity") but generally inappropriable. When there are economies of scale in data mining, all else being equal it will clearly be optimal for the activities that give rise to the data to be carried out within the same enterprise, given the obstacles to market transactions. And the data's transferability and aggregability make it possible to achieve efficiency gains through the combination of datasets. This advantage may be even greater than those realized through the intra-firm integration of tacit knowledge (Teece & Al-Aali, 2011), which lacks these characteristics. As with other resource-based theories, the essential issue is the importance of unmarketable firm-specific assets resulting from the firm's unique historical trajectory. Data generated in-house is an obvious example of such an asset.

These same considerations imply that big databases are resources in the sense of Wernerfelt's (2016) definition: "productive assets that are economically inalienable in the sense that it is more efficient to use excess capacity inside the boundary of the firm. (More precisely, an asset is economically inalienable from a firm *f* if it is inefficient to have any part of its capacity used by a firm other than *f*.)" (p. 12). In the case of data resources, this excess capacity would take the form of data that the firm does not need (or does not realize it needs) for its operations as it currently performs them. It may also consist of data that the firm could potentially use for new economically profitable activities. When economies of scale are possible, transferability and aggregability make data "inalienable" and therefore a resource by this definition.

Capacity for data generation can then itself be considered a competency, which has the potential to strengthen if the firm expands and diversifies. Provided that the firm possesses the necessary IT and analytical resources, its activities play a dual role: They are at once part of the firm's operations and contributors to its stock of data. In the latter capacity, they enhance the dynamic capabilities Teece *et al.* (1997) refer to as abilities to "sense" and "seize" opportunities and to "transform" the organization in ways that allow it to adjust to changes in its environment (Corte Real *et al.*, 2017), making possible both improvements in internal process management and new understandings of markets, customers, and competitors.

Naturally, all of these observations are true of data in general, regardless of how "big." The big-data revolution implies a change not so much in the character of this resource as in its potential to add value. Increases in computer memory and processing power allow both for heightened transferability and aggregability and for the use of techniques capable of analyzing inputs that are orders of magnitude larger and more complex in terms of volume, velocity, and variety than traditional statistics. As a result, new assets are being created from previously unanalyzable data streams, creating an incentive to gain access to them through expansion and diversification.

3. Decipherability and competitive advantage

While the usefulness of big data will naturally depend on the extent of the firm's data analytics and knowledge derivation capabilities, it will also depend in a more fundamental way on the properties of the data itself. Abstracting from the contributions of analysts and decision makers, the usefulness of the information that the optimal data-mining procedure can derive from a dataset will be the ultimate determinant of the dataset's value. This will depend not only on the relevance of the data to the problem at hand but also on the extent of the conclusions that can be derived from it and their statistical significance.

The ideal database will not only be "big"—in the sense of comprising a large number of instances—but also diverse, covering a wide variety of features (the data scientist's term for variables) and including instances in which the values of these features are highly variable. The importance of having features in the dataset that take on a wide range of values is easy to see from the example of churn prediction for a telecoms company, where the objective is to predict which customers are likely to "defect" to another carrier. (This example is from Provost and Fawcett, 2013). In this case it will be desirable to consider many potentially relevant variables—consumer attributes such as age, zip code, gender, annual income, and so on. In order to draw any meaningful conclusions, the analyst will also need data covering multiple demographics as well as significant numbers of customers that switched plans. Data exclusively on, say, 35-year-old males living in downtown Seattle earning 200,000 dollars a year will obviously be impossible to generalize out of sample. And if almost none of them were "defectors," there will be little to learn about churning.

A problem such as this is not unlike that of breaking a code. We may think of the data each customer provides to the firm as analogous to samples of cyphertext received from an adversary. Decryption becomes easier as more of this cyphertext becomes available (Shannon, 1949), just as the potential for deriving information from a dataset is increasing in the number of instances it includes. But just as cyphertext samples become more useful when they encode a greater variety of plaintexts, so the value of having a larger number of instances depends on how varied they are. "More" will not be better if we merely obtain extra copies of the same string of characters or additional data covering the same features for only a single type of consumer. We may thus add yet another item to the lexicon of big-data attributes mentioned at the beginning of Section 2—"decipherability," a function of both volume and diversity.

Large and diverse datasets are therefore valuable for two reasons: They not only contain more information but can also be mined more effectively—they are more "decipherable." The firm with access to a greater volume and diversity of data will not only have the potential to access a greater volume of information but will also be better able to unlock this potential through big data analytics.

Volume and decipherability are a necessary but not sufficient condition for big data to be a source of sustained competitive advantage. There must also be some barrier to competitors who might seek to recreate and use it. Either the data must be difficult to acquire or there must be some obstacle to interpreting it and/or applying the information it generates. As my focus is on the data resources themselves rather than on the capabilities of the individuals and organizations involved in exploiting them, I will consider only the first possibility.

Data provided by a third party will generally be easy for a competitor to obtain from the same provider. Similarly, data produced by an in-house research team using publicly available information may also be straightforward to recreate simply by setting up a similar team. Only data that is the unique by-product of the firm's own activities is likely to be sufficiently difficult for outsiders to replicate to make it truly "rare, inimitable, and nonsubstitutable."

This is, of course, not to say that other data sources cannot be valuable to the firm. Data that is widely available is often an operational necessity. There will be many situations in which internally generated and publicly available data are complementary. It may be useful to combine a retailer's own sales history with statistics on weather conditions available from open sources, for example, as Walmart was able to do prior to Hurricane Frances. There may also be cases in which the firm can obtain a temporary advantage from a new dataset that its competitors have not yet discovered, as has become common in the world of asset management. There, "finding the new data source that generates alpha has become the next arms race" with investors analyzing "satellite images, earnings conference call recordings and transcripts, social media postings, consumer credit and debit card data, and e-commerce transactions" in the hopes of stealing a march, however fleeting, on the competition (Cao, 2019).

If big data is only likely to be "rare, inimitable, and inappropriable" when the firm itself is the data-generating mechanism, the size and diversity required for this resource to be valuable become optimal characteristics for the firm's operations as well. Larger, more diversified, firms will have an advantage over their smaller, less diversified, competitors. But the datasets must cover similar populations and have some related features. In the absence of "relationality," which characterizes data "containing common fields that enable the conjoining of different datasets" (Kitchen & McArdle, 2016), diversifying expansion will not lead to improvements in big data decipherability. To return to our cryptography analogy, obtaining additional cyphertext samples will only facilitate decryption if they were all encrypted with the same key.

4. Subadditive uncertainty and internalization

Diversification and expansion for the sake of data acquisition is likely to be a risky, though probably increasingly necessary, strategy. New activities, whether developed internally or acquired through mergers and acquisitions, will obviously generate new data but there will generally be no guarantee *ex ante* of the extent to which this can be integrated with the firm's existing datasets or of the value of the resulting information. The big data revolution may thus lead not only to larger and more diversified firms but also to many failures to achieve expected synergies with the enlarged firm's original businesses. Avoiding this fate will require correct assessments of the potential for internalization to reduce the uncertainty of analytics results by making datasets generated by the firm's existing or new activities more decipherable than they would be if each were analyzed in isolation.

Data mining techniques may be categorized as either supervised or unsupervised learning. Supervised learning refers to situations in which the analyst has specified a target (the dependent variable in the case of a regression) and seeks to find a relationship between this and a set of features (the independent variables). Unsupervised learning does not involve a particular target and seeks instead to identify similarities among the feature variables without imposing any hypotheses in advance. For example, the objective might be to use clustering to answer a question such as "Do our customers naturally fall into different groups?" (Provost and Fawcett, 2013, p. 24). As unsupervised learning is an open-ended exploration process, it would be difficult to specify the characteristics of new data that would be likely to add value. In this case, the concept of reducing uncertainty is itself problematic. Improving decipherability is only a well-defined goal in supervised learning, where the criterion for added value is increased accuracy in mapping features to a target.

Supervised learning techniques may be classified as regression and classification, depending on whether the target is numerical or categorical (e.g. having target values such as "true" or "false"), respectively (Provost and Fawcett, 2013, Chapter 2). In regression problems, uncertainly may be measured by the variance of the forecast error. (Note that for data scientists the term "regression" includes both linear and non-linear methods.) Naturally, the original and the new data must have some common features—otherwise, it will be impossible to combine them. They must also come from similar populations so that the estimated parameters will be universally valid.

For any particular dataset (D_i), we may think of the firm as choosing the regression algorithm that minimizes the variance of the forecast error (i.e. the mean squared error) given by:

$$(1)\ \underset{\{D_i\}}{\text{Var}}\ [\ \hat{Y}_j\ (\beta_i^*) - Y_j\ ] = [\ 1 / N_i\ ] \sum_{\{D_i\}} [\ \hat{Y}_j\ (\beta_i^*) - Y_j\ ]^2$$

where the Y_j are the actual values of the target, β_i^* is the vector of parameters for the optimal regression algorithm, the $\hat{Y}_j$ are the predicted values of the target, and N_i is the number of elements in D_i. Note that the argument will be the same in the presumably more likely case that the firm minimizes forecast error variance for a holdout subset of D_i rather than D_i itself.

If the firm has access to two datasets (i = 1, 2) including the same target variable, it may analyze them separately using a different algorithm for each, in which case the forecast error variance will be given by the weighted average of the separate variances:

$$(2)\ \underset{\{D_1 U\, D_2\}}{\text{Var}}\ [\ \hat{Y}_j\ - Y_j\ ] = [\ N_1 / (N_1 + N_2)] \sum_{\{D_1\}} [\ \hat{Y}_j\ (\beta_1^*) - Y_j\ ]^2 + [\ N_2 / (N_1 + N_2)] \sum_{\{D_2\}} [\ \hat{Y}_j\ (\beta_2^*) - Y_j\ ]^2$$

For the activities that generate D_1 and D_2 optimally to take place within the same firm, it must be possible to reduce the variance for the combined dataset below the value given by (2) through the use of some alternative regression algorithm. This might, for example, be the case for a convenience store chain that could use data on sales of a potential new product line (D_2) in a regression with sales of one of its existing products as the target (Y). Similarly, it might be possible for the chain's original dataset (D_1) to improve the accuracy of a model used by a rival chain to predict sales at its outlets in the event that the two rivals merged. (In this case, D_1 would initially be an example of a resource for which there was excess capacity that was "more efficient to use inside the boundary of the firm," making it a resource by Wernerfelt's (2016) definition.) Such arguments for internalization might be strongest for expansion into similar neighborhoods or the addition of similar products, where D_1 and D_2 would be more likely to cover consumers with similar shopping behavior or items that formed part of the same consumption pattern.

The situation is much the same for classification problems, where the model uses the features of each instance to assign it to a subset of similar instances. The relevant measure of uncertainty is the extent to which each subset's members actually share the same target category (k). With a quality control algorithm, for example, the goal might be to sort components into "defective" and "nondefective" groups based on characteristics detected by a sensor. Ideally, these two groups would be "pure," one containing only parts that were labelled as defective, the other only parts that were not.

Measures of subset impurity include misclassification error, the Gini index, and the Shannon entropy (Hastie *et al.*, 2009, p. 309). Here I will only consider the Shannon entropy but the conclusions apply equally to the other two measures as well. It is given by:

$$(3)\ \sum_k - p_k \log p_k$$

where p_k is the probability that a randomly chosen member of the subset in question will actually belong to category k and the log is usually taken to be base 2. For a pure subset, entropy is zero; the maximum occurs when all K categories are equally likely so that $p_k = 1/K$.

The entropy (H) of the entire classification can be found by taking a weighted sum of the entropies of the subsets, weighting each subset's entropy by the percentage of the total number of instances that subset contains. So we have:

$$(4)\quad H = \sum_{s} [N_s/T] \sum_{k} - p_{k,s} \log p_{k,s}$$

where N_s is the number of elements in subset s, T is the total number of instances, and $p_{k,s}$ is the probability that an element of subset s belongs to category k (Provost and Fawcett, 2013, p. 53).

As in the regression case, the firm will choose the optimal algorithm. Here, this will generate a classification having minimum entropy (H^*_i) for some particular dataset (D_i). With two datasets (i = 1,2) analyzed separately, this minimized total entropy will be given by:

$$(5)\quad H^*_T = [N_1 / (N_1 + N_2)] H^*_1 + [N_2 / (N_1 + N_2)] H^*_2$$

For the activities that generate D_1 and D_2 optimally to take place within the same firm, it must be possible to reduce the minimized entropy for the combined dataset below the value given by (5). (Again, we may make the same argument if this algorithm minimizes entropy for a holdout subset of D_i rather than for D_i itself).

Additional data generated through expansion or diversification might reduce the value of this aggregate entropy in two ways. First, the additional data may lead to an alternative, entropy-reducing, classification scheme. In the case of consumer banking, for example, we might find that the introduction of a new feature enabled more reliable identifications of potential customers as "defaulters" or "non-defaulters." Second, the added data might increase the purity of some of the subsets in the original scheme and/or increase/decrease the weightings of those with lower/higher-entropy. The bank's new dataset might include more "default" instances, for example, which might increase both the defaulters' share of majority-default subsets (thereby decreasing entropy) as well as these subsets' share of the total instances (the weights, N_s/T).

For both regression and classification, the central issue is the potential for the combination of data from different activities to reduce uncertainty in the results of supervised learning; in other words, for uncertainty—whether measured as variance or entropy—to be subadditive in the datasets. This account of internalization parallels that of Wernerfelt (2016), which is based on subadditivity in contracting costs. It also resembles theories involving economizing on costs related to opportunism, agency, or risk-bearing, which similarly explain the existence of firms on the basis of efficiency gains achievable when operations take place within the same entity rather than in markets. Indeed, the role of big data resources in uncertainty reduction is ultimately also a matter of economizing on costs, be they costs resulting from inefficiencies in internal processes or from strategic mistakes.

5. Redefining Relatedness

Adner *et al.* (2019) point out that "the potential for synergies appears much greater in digital-enabled contexts than in the all-physical world," suggesting that "traditional notions of relatedness may benefit from re-examination" (p. 258). The analysis in Section 4 provides a starting point for thinking about this issue. In addition to relying on conventional criteria such as shared operations or markets, we may also classify activities as related whenever uncertainty is subadditive in the datasets they generate. Traditionally, industries have been thought of as nodes in a hierarchical tree such as those of the NAICS, SIC, GICS, or ICB systems, which begin with broad sectoral aggregates and branch into increasingly more specialized subsectors based on the nature of an establishment's principle business. When firms are viewed as data-

generating mechanisms, this structure may be overlaid with a "hyperlinked" topology in which multiple connections exist among establishments that would previously have been assigned to entirely distinct industry groups. Relatedness then depends on the characteristics of the data to which their operations give rise in addition to the nature of those operations themselves.

Such a scheme will naturally replicate many familiar categorizations since activities that share operational procedures and objectives will be more likely to generate data that includes statistically related features covering instances drawn from similar populations. Consider Occidental's acquisition of Anadarko, for example, both companies that operate primarily in SIC code 1311 (Crude Oil and Natural Gas). This takeover was partly justified on the basis that it would allow for the combination of technical data from the two companies' holdings in the Delaware basin (located in West Texas and Southern New Mexico). This combined dataset will include many of the same features and cover an expanded range of locations, potentially leading to entropy and variance reductions in supervised learning models by merging observations of dissimilar instances drawn from related populations (Hollub, 2019).

Other recent data-driven acquisitions show how connections between companies may be much closer than those implied by conventional understandings of relatedness. Amazon's 2017 acquisition of Whole Foods, for example, brings together online and in-store purchase history data covering many similar individuals. Combining these datasets creates the potential for model improvements through the addition of new features. As data generators, the two have more in common than their SIC classifications would suggest—Amazon is in SIC code 5961 (Catalog and Mail-Order Houses), Whole Foods, in 5411 (Grocery Stores). Or consider Verizon's 2016 acquisition of Yahoo**,** two names even farther apart on the SIC tree, with Verizon in 4813 (Telephone Communications, except Radiotelephone); Yahoo, in 7374 (Computer Processing and Data Preparation and Processing Services). Here, the relationship results from the fact that the two have a shared user base. As Verizon's customers are often also users of Yahoo's internet offerings, Verizon has data that can potentially help Yahoo target those individuals for online advertising clients.

In all of these examples, relatedness is established because companies are monitoring the same populations. These may consist of human beings—as in the cases of Amazon/Whole Foods or Verizon/Yahoo—or of inanimate objects—as in the Occidental/Anadarko case. The key point is that shared observation creates the relationship, irrespective of the operational considerations that determine how and why the data is collected. Common business models are not a necessary condition for data synergies.

6. Conclusion

An important aspect of the nature of the firm is the firm's role as a data generator. More than a "functional information system and decision-making system" (Machlup, 1967, p. 27), the firm is also the source of data derivable from its own operations. As big data analytics makes this information ever more accessible, strategic decision making will have to take this into account, directing assets into activities that are not only attractive on the basis of traditional metrics but also have the potential to generate datasets that are complimentary in the sense that their combination makes them more decipherable.

Grant (1996, p. 109) notes that "Although economists use the term 'theory of the firm' in its singular form, there is no single, multipurpose theory of the firm." There are rather "many theories of the firm which both compete in offering rival explanations of the same phenomena, and complement one another in explaining different phenomena." A theory of the firm as a data generating mechanism is a case in point. Naturally the firm cannot be viewed solely in this way—without some operational objective there would neither be data generation nor any need for information. Furthermore, the fact that it might be optimal for firms to expand and diversify for the sake of acquiring data does not imply that other considerations will not be equally or even more important. Synergies resulting from data sharing with a rival might justify an acquisition all else being equal yet not result in sufficient additional revenue to cover costs expected to result from the

integration of the acquiree. Any real-world optimization calculation would clearly have to include not only informational factors but a variety of other dimensions as well.

Similarly, in-house data cannot be a firm's sole source of competitive advantage any more than battles can be won on the basis of superior cryptography alone. Data must be complemented with the analytical capabilities necessary to interpret it and the operational competencies required to make use of the information it contains. It is but one of the many knowledge-related resources contemplated by the resource-based view, though one with the unusual characteristic that while it is inappropriable, like most knowledge, it is aggregable and neither local nor tacit.

Where it is possible to improve decipherability by combining datasets acquired through expansion and diversification, there may be significant changes in industry structure. In sectors where larger players are able to gain informational advantages over their smaller rivals in this way, firm numbers will fall as the average firm size rises. For first movers, there will be advantages similar to those resulting from network effects and other forms of economies of scale. Their operational and data assets will coevolve, making the former more productive, the latter more comprehensive and more actionable, and the firm more efficient and more agile.

The value of any form of intelligence is determined by the extent to which it reduces uncertainty. For firms able to act as effective data generating mechanisms, such uncertainty reduction will be a byproduct of their own activities. For them, big data will imply a shift in their optimal boundaries and a transformation in the nature of their operations as traditional functions take on new significance in the production of information resources.

References

Adner R, Puranam P, Zhu F (2019) What Is different about digital strategy? From quantitative to qualitative change. *Strategy Science* 4(4): 253–261. https://pubsonline.informs.org/doi/full/10.1287/stsc.2019.0099

Aoki M, Gustafsson B, Williamson O (1990) *The Firm as a Nexus of Treaties* (Sage Publications, London).

Arrow KJ (1971) *Essays in the Theory of Risk Bearing* (Markham Publishing Company, Chicago, IL).

Arrow KJ (1973) Information and economic behavior. Office of Naval Research, Technical Report No. 14, Washington, DC. https://pdfs.semanticscholar.org/8000/cc9d8a0af5ffd73991d80c68d2360ceac96b.pdf

Barney J (1991) Firm resources and sustained competitive advantage. *Journal of Management* 17(1): 99–120.

Braganza A, Brooks L, Nepelski D, Ali M, Moro R (2017) Resource management in big data initiatives: Processes and dynamic capabilities. *Journal of Business Research* 70: 328–337.

Beck M, Libert B (2018) The machine learning race is really a data race. *MIT Sloan Management Review Blog* (December 14).

Brynjolfsson E, Hitt, L, Kim H (2011) Strength in numbers: How does data-driven decision-making affect firm performance?" http://dx.doi.org/10.2139/ssrn.1819486

Cao L (2019) AI pioneers in investment management. CFA Institute. https://www.cfainstitute.org/en/research/industry-research/ai-pioneers-in-investment-management

Chaterjee S, Wernerfelt B (1991) The link between resources and type of diversification: Theory and evidence. *Strategic Management Journal* 12: 33–48.

Coase R (1937) The nature of the firm. *Economica* 4(16): 386–405.

Conner K (1991) Historical comparison of resource-based theory and five schools of thought within industrial organization economics: Do we have a new theory of the firm? *Journal of Management* 17(1): 121–154.

Côrte-Real N, Tiago O, Ruivo, P (2017) Assessing business value of big data analytics in European firms. *Journal of Business Research* 70: 379–390.

Davenport TH (2006) Competing on analytics. *Harvard Business Review* 84: 98–107.

Davenport TH, Barth P, Bean R (2012) How 'big data' is different. *MIT Sloan Management Review* 54: 43–46.

Dierickx, I, Cool K (1989) Asset stock accumulation and sustainability of competitive advantage. *Management Science* 35(12): 1504–1511.

Elgendy N, Elragal, A (2016) Big data analytics in support of the decision making process. *Procedia Computer Science* 100: 1071–1084.

Erevelles S, Fukawa N, Swayne L (2016) Big data consumer analytics and the transformation of marketing. *Journal of Business Research* 69: 897–904.

Ghasemaghaei M, Calic G (2019) Does big data enhance firm innovation competency? The mediating role of data-driven insights. *Journal of Business Research* 104: 69–84.

Grant RM (1996) Towards a knowledge-based theory of the firm. *Strategic Management Journal*, Winter Special Issue 17*:* 109–122.

Günther WA, Mehrizi MHR, Huysman M, Feldberg F (2017) Debating big data: A literature review on realizing value from big data. *Journal of Strategic Information Systems* 26: 191–209.

Hastie T, Tibshirani R, Friedman JH (2009) *The elements of statistical learning: Data mining, inference, and prediction*. (Springer, New York).

Hayek F (1945) The use of knowledge in society. *American Economic Review* 35(4): 519–530.

Hays CL (2004) What Walmart knows about customers' habits. *The New York Times* (November 14), https://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html

Hollub V (2019) Presentation, 35th Annual Bernstein Strategic Decisions Conference, May 29, https://www.oxy.com/investors/Documents/05.29.19%20Bernstein%20Podium%20Presentation%20-%20Transcript.pdf

Janssen M, van der Voort H, Wahyudi, A (2017) Factors influencing big data decision-making quality. *Journal of Business Research* 70: 338–345.

Jensen MC, Meckling WH (1976) The theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3(4): 305–360.

Kiron D, Prentice PK, Ferguson RB (2014) The analytics mandate. *MIT Sloan Management Review* 55: 1–25.

Kitchen R, McArdle G (2016) What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data and Society* January-June: 1–10.

Lin C, Kunnathur A (2019) Strategic orientations, developmental culture, and big data capability. *Journal of Business Research* 105: 49–60.

Lockett A, Thompson, S (2001) The resource-based view and economics. *Journal of Management* 27: 723–754.

Machlup F (1967) Theories of the firm: Marginalist, behavioral, managerial. *The American Economic Review* 57(1): 1–33.

Machlup F (1980) *Knowledge: Its creation, distribution and economic significance,* Vol. I (Princeton University Press, Princeton, NJ).

Mcafee A, Brynjolfsson E (2012) Big data: The management revolution. *Harvard Business Review*, October: 1–9.

Merendino A, Dibb S, Meadowsa M, Quinna L, Wilson D, Simkin L, Canhotoc, A (2018) Big data, big decisions: The impact of big data on board level decision making. *Journal of Business Research* 93: 67–78.

Nonaka I (1994) A dynamic theory of organizational knowledge creation. *Organization Science* 5(1): 14–37.

Paolini G (2019) "Tesla, the data company." *CIO* (August 28), https://www.cio.com/article/3433931/tesla-the-data-company.html

Peter M, Kraft C, Lindeque J (2020) Strategic action fields of digital transformation: An exploration of the strategic action fields of Swiss SMEs and large enterprises. *Journal of Strategy and Management* 13(1): 160-180.

Provost F, Fawcett T (2013) *Data Science for Business* (O'Reilly Media, Sebastopol, CA).

Schwarz E, Mckeil A, Dean M, Duffield M, Chandler D (2019) Datafying the globe: Critical insights into the global politics of big data governance. *Big Data and Society*, January 26, https://westminsterresearch.westminster.ac.uk/item/q9z00/datafying-the-globe-critical-insights-into-the-global-politics-of-big-data-governance

Shannon CE (1949) Communication theory of secrecy systems. *Bell System Technical Journal* 28(4): 656–715.

Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. *Journal of Business Research* 70: 263–286.

Sobel RS, Clark JR (2018) The use of knowledge in technology entrepreneurship: A theoretical foundation. *Review of Austrian Economics* 31: 195–207.

Splunk Inc. (2017) "The Essential Guide to Machine Data." https://www.splunk.com/en_us/form/the-essential-guide-to-machine-data.html

Stevens L, Haddon H (2019) Big prize in Amazon-Whole Foods deal: Data." *Wall Street Journal* (July 18), *https://www.wsj.com/articles/big-prize-in-amazon-whole-foods-deal-data-1497951004*

Teece DJ (1980) Economies of scope and the scope of the enterprise. *Journal of Economic Behavior and Organization* 1(3): 223–247.

Teece DJ, Al-Aali A (2011) Knowledge assets, capabilities, and the theory of the firm. Easterby-Smith M, Lyles MA, eds. *Handbook of Organizational Learning and Knowledge Management*, Second Edition (John Wiley and Sons, Chichester, UK), 505–534.

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. *Strategic Management Journal* 18(7): 509–533.

Vial, G (2019) Understanding digital transformation: a review and a research agenda. *The Journal of Strategic Information Systems* 28(2): 118-144.

Wamba SF, Gunasekaran A, Akter S, Ren SJF, Dubey R, Childe, SJ (2017) Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research* 70: 356–365.

Wernerfelt B (1984) A resource-based view of the firm. *Strategic Management Journal* 5: 171–80.

Wernerfelt B (2016) *Adaptation, Specialization, and the Theory of the Firm: Foundations of the Resource-based View* (Cambridge University Press, Cambridge, UK).

Williamson O (1975) *Markets and Hierarchies* (Prentice-Hall, Englewood Cliffs, NJ).