Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

# Towards Real-time Data Value Analytics in Data Service Networks

Pieter De Leenheer[a,b] and Stijn Christiaens[a]

[a]*Collibra, 61 Broadway, Suite 2401, New York, NY 10006*
[b]*Columbia University, 116th and Broadway, New York, NY 10027*

**Abstract**

In the past few decades we have seen an explosion of data. In this digital disruption, companies establish strong competitive barriers by exploiting data network effects. This involves high-scale experimentation with customer-engaging technologies, thereby relying on trusted data sharing, while leveraging human domain expertise as best as possible. Various frameworks and methodologies for data valuation are being proposed in this new research area. The challenge remains which methods works best, and how they can be compared. We propose a data governance system for data service networks in which data assets are valued in function of how much they contribute to data network effects. This system records metadata that can be used by the research community as analytical testbed for evaluating and comparing various data valuation methods and tools.



*Keywords:* data governance; data valuation; data network effect; data service network

## 1. Introduction

The world's most valuable resource is no longer oil, but data. We are amidst an explosive growth of data produced by digital technologies such as social, mobile, analytics, cloud, and the internet of things. IDC predicts that the "digital universe" (the data created and copied every year) will reach 180 zettabytes in 2025, a sign the real era of big data is still to come [24]. Data-native organizations have an unfair advantage over their incumbent counterparties to reap value from data, and research shows digital revenue growth consolidating with fewer companies at the top [20]. Indeed, only 12% percent of companies that were in the Fortune list 60 years ago are still in this list today [3].

Yet, digital disruption offers an opportunity for any company to create real value from data and provide new angles to beat competitors that were already thought invincible. Digital disruptors exploit two strategies to change the competitive setting and undermine the viability of existing product/service portfolios and go-to market strategies [2]:

(1) *Customer engagement:* the transformation of a go-to-market strategy by building an enduring relationship with your customer. For example, insurance companies that organize around life events in order to predict what you need next.

(1) *Digitized solutions:* the transformation of the business model by providing data-driven services as a byproduct from assets. For example, John Deere transformed from an equipment manufacturer into a smart farming service platform[1]; Schindler transformed from an elevator/escalator vendor (measured by assets sold) into an urban mobility platform (measured by people transported)[2].

The core challenge for any competitive digital strategy lies in leveraging *data network effects*[3]. Similar to network effects, data network effects define a virtuous circle as illustrated in Figure 1: more product/service usage leads to more data, which in turn leads to smarter products that *learn* to, e.g., monitor performance, provide personalized recommendations or predictions. More traditional businesses typically "learn" through business intelligence and analytics, with human analysts doing the bulk of the work. There is a separate process to engineer those insights back into the product or service. The more you can automate this *learning bottleneck*, the more likely you are to get the network effects going that produce the "killer app".

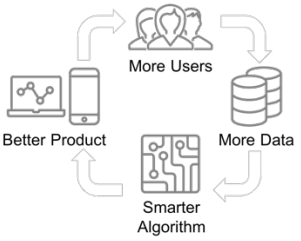


Figure 1: data network effects' virtual circle.

This requires business value-driven experimentation with customer-engaging technologies at a high scale, thereby relying on trusted data sharing, while leveraging human domain expertise as best as possible. Companies can't possibly control the entire "value chain". i.e., all the touchpoints with their customers, as well as scale the derived data network effects on their own. Instead, similar to the decades-old open innovation paradigm, they must focus on connecting to a data-partner ecosystem [6] in which they can find suitable data governance (DG) operating models that can balance out cost, risk and value associated with big data sharing and exploitation across the ecosystem participants [7].

In this paper, we propose a DG system in which data assets are valued in terms of their usage over time, i.e. in function of how much they contribute (directly or indirectly) to leveraging data network effects. This system records metadata that can be used by the research community as analytical testbed for evaluating and comparing the recent publication of various data valuation methods and tools.

## 2. Related work

The data valuation research community is young, and despite a lack of consensus as to what data value defines, we can witness an interesting variety of data value methods, some starting from traditional data quality dimensions, others from accounting principles and file systems. One of the earliest papers posing the question how to evaluate the value of a digital information was introduced in the 1999 seminal paper by Moody and Walsh [5]. Recently, Attard

[1] https://digital.hbs.edu/data-and-analysis/product-platform-john-deere-revolutionizes-farming/
[2] https://blogs.wsj.com/cio/2016/11/03/schindler-digital-chief-michael-nilles-explains-why-he-left-the-cio-title-behind/
[3] http://mattturck.com/the-power-of-data-network-effects/

and Brennan [17] provided a fair overview of methods (incl. [1,4,12,16,21,22,23]) and some open challenges. These methods value data assets by (either or a combination of) usage, cost, timeliness, applicability to business, utility, and uniqueness. Most methods, such as [21] focus on snapshots, valuing data at a certain point in time, and are merely theoretical; lacking in real-world validation. Some approaches align with our view to quantify data assets' value based on their usage over time [16]. Our proposal goes further by providing a testbed of information that relates a data asset's value explicitly to how it, directly or indirectly, contributes to data network effects, and therefore business value.

Data valuation is obviously central for data marketplaces where data providers can publish data sets and manage transactions with data consumers. E.g., Dawex.com which resembles services for a data exchange; Streamr.com for real-time data; and Dock.io for personal data using blockchain technology. Moreover, several data sharing protocols are emerging, including Iota.org to allow IoT micro-services to exchange data in using distributed ledger technologies.

After the numerous data usage fraud and breach cases, we have seen an increasing frequency of articles in popular media outlets about data rights and rewards of individuals when companies want to create value from their personal data. In 2017, The Economist made a point about data being the most valuable asset in the world, and gave some examples of recent mergers and acquisitions where company value was largely based on data they own [24]. More recently, MIT Technology Review's issue of July/August 2018 features several articles that call for a "liberalization" of the data market[4].

## 3. Data Service Networks: Metadata Requirements for Data Valuation

As we proposed in [13], we conceptually represent a data service network as a non-linear graph, i.e., a graph in which some nodes, called hubs, have more importance than other nodes in the network. Nodes denote specialized *teams*, i.e. sets of workflows between artificial and human decision-making actors working on certain data assets. Nodes can also be certified "backbone" micro-services such as policy management, glossary, reference data and helpdesk services on the Collibra DG Platform[5] or those provided by Linked Open Data APIs[6]. Nodes interact through value-exchanging "services" which through rules (such as reciprocity and regulation) bear balance in the network. E.g., one data set consumed by one team may depend on a missing values model calculated by another team.

Rich metadata for these data service networks describe *business semantics* (i.e., both networked business-model as well as DG-operational) to leverage data network effects [8,13]. In order to capture value-in-use in the network, we need metadata about a data asset for the following[7].

- *Usage*: has the data asset been used, and if so where and what business outcomes did it lead to? Has it been used multiple times? By capturing usage data in the platform, we can start to investigate the potential relation between data usage and data value (or the flip side: data risk).
- *Business Traceability*: what are the business contexts for the data asset? Which business processes *generated* (purposefully or as a byproduct) the data asset, and which ones critically depend the data asset? What policies and business rules (concerning, e.g., quality or protection) govern the data asset?
- *Technical Lineage*: what is the chain of transformations and systems this data asset has gone through? Data assets often go through complicated transportations, transformations and aggregations, each which come with value, cost or risk entropy. Transportations from one system to another could potentially decrease a data asset's value as certain information is dropped. Aggregations or transformations potentially increase an asset's value as information is added.

---

[4] https://www.technologyreview.com/magazine/2018/07/
[5] https://www.collibra.com/data-governance-solutions/
[6] e.g., https://schema.org/; http://linkeddata.org/; http://theme-e.adaptcentre.ie/odgov/
[7] see also an open data governance ontology which captures these requirements: http://theme-e.adaptcentre.ie/odgov/

- *Trust*: who is involved in the stewardship workflows (e.g., ingestion, onboarding, approval, certification, …) for this data asset? Knowing the reputation of contributors and their former collaborations gives a measure of trust, and it is reasonable to assume that there is a relation between the degree of trust and the value of the data asset.
- *Quality*: how accurate/timely/complete is this model? As many value methods (such as [21]) still take the typical data quality dimensions into account, they will give complementary dimensions to value assessment.

## 4. Example Data Service Network

Figure 2 illustrates some of the above proposed metadata requirements for a given Data Set named "Credit Risk and Customer Data" (pinned box at the bottom-center of the image) in the Collibra Data Governance Center.

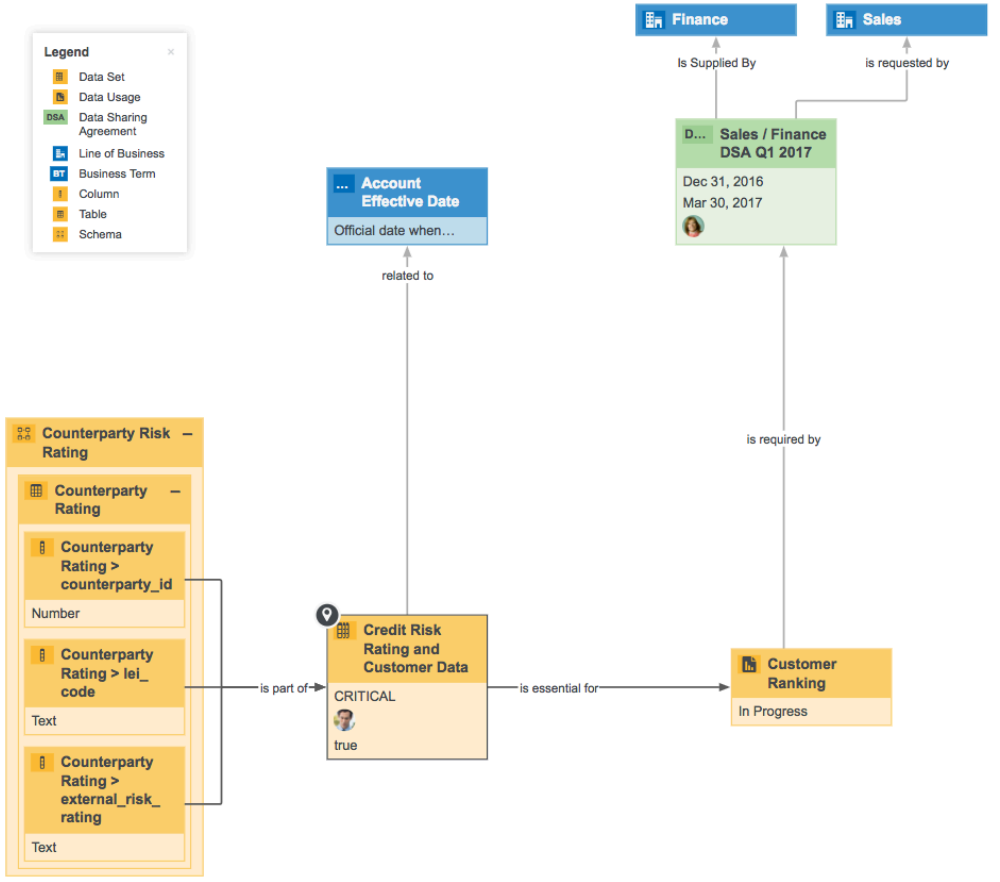


Figure 2: Example data assets (boxes) with some detailed metadata (relations and attributes) that form the basis for real-time data value analytics.

This Data Set consists of a Table named "Counterparty Risk Rating" containing three columns (counterparty_id, lei_code, external_risk_rating). The Data Set is essential for a Data Usage named "Customer Ranking", i.e., an activity in which customers are scored for processing priority. This Data Usage is required by a Data Sharing Agreement (name "Sales/Finance DSA Q1 2017") where two business lines play a role: Finance is the *data producer* and Sales the *data consumer*. Note that Finance as the data producer should be seen as a trust indicator, as Finance departments are typically measured on getting their numbers right. Stewardship roles are shown for some data assets as small avatars of the people actually responsible for executing stewardship.

The above illustrated data service network identifies a potential opportunity for a data network effect: as Sales leverages the data from Finance, it will then also extend it into its own data sets by enriching the data simply while *executing its own business process* (e.g., adding information about sales stages, opportunity estimates, deal sizes, etc.). This new data set could then circle back to Finance and serve as a new input into the Customer Ranking algorithms.

## 5. Towards real-time data-value, -cost and -risk analytics

The proposed metadata points enable a real-time capability for analyzing the value, cost and risk regarding the usage and sharing of data assets. Following are some examples of ongoing work that contributes to enriching this metadata, as well as leveraging it for analytical data value insights.

- *Human performance* In De Leenheer et al. [46], we used *social performance indicators* to gather insights of the (ever changing) social arrangement of collaboratively evolving an ontology. The analysis of user and system interactions enables clustering user types, including automatically assigning workflows of data governance operations by Debruyne and Meersman [15].
- *Data governance operating models* As with organisms the most successful data value networks will be those that can best adapt, including their governance models. The result will be a library of best-of-breed templates. To this end, we can leverage our experience with hundreds of model revisions at customers worldwide. Moreover, we should review state-of-the-art DataOps and DevOps platforms from vendors (e.g., Databricks and Streamr) and in-house (e.g., Airbnb's airflow, Facebook's FBLearner and Google's TFX).
- *Data sharing agreements* Novel decentralized technologies (e.g. the DAT PROJECT[8]) could be explored for automatically generating data sharing agreements based on agreed rule and policy frameworks (e.g. the European General Data Protection Regulation[9]). Smart agreements should automatically expire themselves, and raise unauthorized data usages. To this end, we can reuse analogous mechanisms from digital music rights clearing where fingerprinting is used ubiquitously by applications such as Shazam and passed on to clearing societies.
- *Trust* Systems of engagement (e.g., Slack) leverage systems of record to engage with trusted data. In recent work [19], we applied named entity recognition and machine learning on the social context of data to provide recommendations for the next best action. Supporting the low-level decisions business users face when interacting with data will create more time for them to focus on real impactful problems, hence reduce cost.

## 6. Conclusion

The challenge for non-digital natives is that they have to adapt their overall systems. Typically, they've built platforms designed to carry out a certain business function in a prefixed non-adaptive manner rather than an explicit understanding of the business semantics that could enable a data service ecosystem to adapt its role in changing value propositions [13]. Many of them still exhibit critical risk because of data quality and -protection issues. DG on this operational level is designed and enforced hierarchically to monitor and remediate these risks. The objective of hierarchical DG is to establish truth for decision support, but it does not enable data service networks to their full capacity.

DG in a more decentralized data service network is to be designed as a seed-model to overcome these bottlenecks by empowering workers with analytical capabilities (incl. humans, machine and data) and encourage peer-to-peer engagements within and outside the organization. The objectives are to seed innovation through speed and complexity, as well as derive reward systems to for all participants.

In this paper, we proposed a data governance system in which data assets are valued in function of how much they contribute to leveraging data network effects. This system records metadata that can be used by the research community as analytical testbed for evaluating and comparing the recent publication of various data valuation methods and tools. We also presented some research which we believe are important to leverage or enrich this metadata.

---

[8] https://datproject.org/
[9] https://www.eugdpr.org/

## Acknowledgements

We would like to thank following people for various discussions leading to this paper: Rob Brennan and Judie Attard (TCD ADAPT), Ann Nowé and Johan Loeckx (VUB AI Lab), Bart Vandekerckhove (Collibra), Henry Peyret (Forrester), Doug Laney (Gartner), James Short (San Diego Super Computer Center), Jaap Gordijn and Roel Wieringa (VU Amsterdam) and Maarten Masschelein (Shape.ai). This research has been partially funded by the Brussels Capital Region's Innoviris "Team up for AI" grant #2017DS78C.

## References

[1] Viscusi, G., Batini, C. (2014) Digital Information Asset Evaluation: Characteristics and Dimensions. pp. 77–86, Springer.

[2] Ross, J.; Weill, P. (2006) Enterprise Architecture As Strategy: Creating a Foundation for Business Execution, Harvard Business School Press.

[3] Anthony, S.; Viguerie, S.; Schwartz, E.; Van Landeghem, J. (2018) 2018 Corporate Longevity Forecast: Creative Destruction is Accelerating, Innosight

[4] Sajko, M.; Rabuzin, K.; Baca, M. (2006) How to calculate information value for effective security risk assessment. Journal of Information and Organizational Sciences, 30(2), 263–278

[5] Moody, D.; Walsh, P. (1999) Measuring The Value Of Information: An Asset Valuation Approach. Seventh European Conference on Information Systems (ECIS'99) pp. 1–17

[6] Normann, R.; Ramirez, R. (1993) Designing Interactive Strategy, Harvard Business Review

[7] Tallon, P. (2013) Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost, IEEE Computer 46(6), 32-38

[8] De Leenheer, P., Christiaens, S., and Meersman, R. (2010) Business Semantics Management: a Case Study for Competency-centric HRM. In Journal of Computers in Industry 61(8): 760-775 . Elsevier

[9] Brandenburger, A.; Nalebuff, B. (1996). Co-opetition. New York: Doubleday

[10] Yap, E. G. (2017) A Model of Trust and Collaboration in a Fresh Vegetable Supply Chain in Central Philippines. International Journal of Applied Industrial Engineering, 4(2), 47-57

[11] Gaurav Tejpal, R.; Garg, A. (2013) Trust among supply chain partners: a review, Measuring Business Excellence, 17(1), 51-71

[12] Even, A.; Shankaranarayanan, G. (2005) Value-driven Data Quality Assessment. In Proc. of MIT Conference on Information Quality

[13] De Leenheer, P.; Cardoso, J.; Pedrinaci, C. (2013) Ontological Representation and Governance of Business Semantics in Compliant Service Networks. In: Falcão e Cunha, J.; Snene, M.; Nóvoa, H. (eds.) Exploring Services Science. IESS 2013. Lecture Notes in Business Information Processing, vol 143. Springer, Berlin, Heidelberg

[14] De Leenheer, P; Debruyne, C.; Peeters, J. (2009). Towards Social Performance Indicators for Community-based Ontology Evolution. In Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK2008), collocated with the 8th International Semantic Web Conference (ISWC 2009). CEUR-WS

[15] Debruyne, C.; Meersman, R. (2012). GOSPL: A Method and Tool for Fact-Oriented Hybrid Ontology Engineering. In: Advances in Databases and Information Systems - 16th East European Conference, ADBIS 2012, Pozna, Poland, September 18-21, 2012. Proceedings II (pp.153-166)

[16] Chen, Y. (2005) Information Valuation for Information Lifecycle Management. In: Second International Conference on Autonomic Computing (ICAC'05). pp. 135–146. IEEE

[17] Attard, J.; Brennan, R. (2018) Challenges in Value-Driven Data Governance. In Proc. Of Confederated International Conferences: CoopIS, C&TC, and ODBASE, Springer

[18] Carr, N. (2004) Does IT Matter? Information Technology and the Corrosion of Competitive Advantage. HBR Press

[19] Brennan, R.; Quigley, S.; De Leenheer, P.; Maldonado, A. (2018) Automatic Extraction of Data Governance Knowledge from Slack Chat Channels. In Proc. Of Confederated International Conferences: CoopIS, C&TC, and ODBASE, Springer

[20] Bughin, J.; Catlin, T; Hirt, M.; Willmott, P. (2018) Why digital strategies fail, McKinsey Quarterly, January

[21] Al-Ruithe, M.; Benkhelifa, E.; Hameed, K. (2016) Key dimensions for cloud data governance. In Proc. of IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 379–386

[22] Ahituv, N. (1982) A Systematic Approach toward Assessing the Value of an Information System. MIS Quarterly 4(4), 61

[23] Laney, D. (2017) Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage, Gartner

[24] Klare, M. (2017) Fuel of the future: Data is giving rise to a new economy, The Economist, May 2017