An information-centric perspective on data

Jože M. Rožanec†
Artificial Intelligence Laboratory
Jožef Stefan Institute
Ljubljana, Slovenia
joze.rozanec@ijs.si

Lola Montero Santos
Department of Law PhD
European University Institute
Florence, Italy
lola.monterosantos@eui.eu

Giacomo Delinavelli
Department of Law PhD
Arthur's Legal
Amsterdam, The Netherlands
delinavelli@arthurslegal.com

ABSTRACT

While the focus of information theory, science, and technology is information, most of the current legal and regulatory frameworks focus on data and portability, disregarding the information aspect, and therefore fail to successfully achieve their goals. The paper presents an information-centric perspective on data. Furthermore, it argues that data ownership could enable additional regulatory aspects while being key to develop a data market and a data value chain. Moreover, some ideas are drafted on how the value of information could be attributed across different stages of the data value chain.

KEYWORDS

Data, Theory of value, Data value chain,

1 ECONOMIC ASPECTS OF DATA

1.1 Who or what generates data?

Data is defined by Bygrave [4] as "*signs, patterns, characters or symbols which potentially represent some thing (a process or object) from the 'real world' and, through this representation, may communicate information about that thing*". Nevertheless, Gellert [8] notes that the definition of data and the distinction between information and data remain a matter of discussion. Two kinds of data generation processes exist. First, we find sensors that observe certain phenomena (either physical or virtual) and quantify them. Second, we find processes that generate synthetic data based on previous knowledge about something they aim to emulate (e.g., heuristics or machine learning models for synthetic data generation).

1.2 What makes data valuable?

Data is not sought by the data itself, but for the information it contains. While information has been defined in many ways, it is generally understood as the knowledge communication [5]. That knowledge is sought at a particular time with a particular goal in mind, and the value of the information is related to that goal [1].

The increasing adoption and use of machine learning fosters an increasing demand for data suitable for satisfying the particular goals the machine learning models are trained for. In the machine learning realm, multiple paradigms exist and they conceive learning goals in different ways. Among these paradigms, we find unsupervised learning, supervised learning, and reinforcement learning [2]. Unsupervised learning aims to learn from unlabeled data for clustering, density estimation, or dimensionality reduction. Supervised learning aims to learn the association between input vectors and dependent variables (classification or regression settings). Finally, reinforcement learning aims to find suitable actions in a particular situation that maximize a reward and help achieve a certain goal. In reinforcement learning the algorithm interacts with the environment by trial and error, exploring actions and context to learn something new, and exploiting gained knowledge to attain the final goal. In every case, the relevant knowledge toward the specific goal is different. Furthermore, it can be conveyed using different modalities (tabular data, graph data, sequence data, or image data).

While commodities usually are subject to divisibility, appropriability, scarcity, and display decreasing returns to use, it has been observed that information is not easily divisible, and its value often increases with its use [9]. While data is abundant and can be replicated arbitrarily, the scarcity could arise from the finite amount of means to replicate, process and store the data.

From the abovementioned observations, multiple considerations arise, which we briefly introduce in the following sections.

1.3 How informative is the data?

Many approaches and metrics have been developed to measure the amount of information present in the data. Among common measures we find the Shannon entropy, mutual information, and directed information. The Shannon entropy measures the degree to which the data is unexpected: the higher the unexpectedness of the data, the higher the information value it holds. Conditional entropy measures the degree of unexpectedness of a variable given the value of another known variable. Mutual information assumes two random variables are given and measures how much information about one variable can be drawn by observing the second one. Finally, given a pair of sequences, the directed information measure the extent to which one sequence is relevant for causal inference on the other one.

In machine learning, there is an interest in understanding what is invariant and what is noise across datasets and contexts. The capacity to discriminate between information and noise is a key aspect of learning [16]. While in this context valuable data would be the one that provides information that displays little correlation to already known independent variables, such information could still be useful to a person for the sake of context (e.g., while economic growth is usually correlated with employment rates, and using both may be meaningless for a

†Corresponding author: joze.rozanec@ijs.si

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

machine learning algorithm in certain cases, they may still be valuable to a person).

1.4 Do we have substitutes?

A key aspect that defines the economic behavior of consumers with respect to a given product in the market, is whether a good substitute product exists for it. The demand for substitute products shows a negative correlation: the demand for one product reduces or replaces the need for the other. Substitutes of a particular data variable would be any kind of data that displays a high enough degree of mutual information.

1.5 Data enrichment

When considering learning goals for a specific machine learning algorithm, we may find that a single data variable will unlikely be able to describe complex relations observed in the real world. Therefore, data enrichment is required to join multiple data variables describing the different aspects of the real world, and therefore providing new information to the machine learning model or the person consuming it.

1.6 Data elasticity

The demand for a certain product is considered elastic when the demanded quantity of a product changes more than proportionally when its price increases or decreases. While product elasticity is usually considered in the realm of physical products, intangible assets could also display elastic behavior. E.g., people would be more or less likely to disclose some sensitive information based on the perceived benefit. The perceived benefit could be considered the price of that piece of data, paid either in kind (e.g., access to a product feature), money (either selling or renting the data), or both. A particular example could be access to data describing typing patterns. Such data could be used for continuous authentication of a person using a particular hardware (e.g., ensuring only the owner uses a particular device) [7, 15], or for early disease diagnosis [10]. In each case, the person could grant access to the data in exchange for (a) a digital good (e.g., a typing profile), (b) some service (e.g., authentication, (continuous) identity verification, or disease diagnostics, or (c) money obtained from data leased or sold at an aggregate level (e.g., for analytic purposes, such as its use within the scope of the research of a given disease, public health policy planning, or market research). While in (b) the person would benefit from the service and eventually pay an additional fee for it, in (c) the person could perceive a fraction of the money paid to access some of the data he owns. We devote part of Section 2.3 to weight the benefits and drawbacks of granting access to data permanently, and the benefits and drawbacks of selling or leasing data.

1.7 Data amortization

Amortization refers to the accounting method used to expense the cost of intangible assets over their expected lifetime for tax or accounting purposes. Amortization is analogous to the depreciation of physical assets. The costs are expensed to reflect the asset's loss of value over time (e.g., in physical assets this could be due to the wearing out with their use over time). Without delving into the details of data amortization, it can be observed that not all data was created equal: while certain data wears out with time (e.g., fraud patterns change over time, and, therefore, past patterns do not provide insights into current fraud strategies), some other may be lightly affected by time (e.g., prices in inflationary context), or may not be affected by time at all (e.g., landscape images). When the underlying semantics change (e.g., new types of fraud emerge and old ones disappear) there is little that can be done to avoid data depreciation. Nevertheless, when the semantics remain the same but changes in the data distribution are observed, we speak about data drift. Data drift can be mitigated to a certain extent with strategies that learn how to align past and current data distributions (e.g., through Monge mapping). While not always feasible, such alignment could extend the lifecycle for certain data if required. Anyway, the existence of different data lifecycles requires different depreciation strategies to be considered in each case.

2 DATA: ITS VALUE AND PRICING

2.1 Theories of value

A key question in economic theory regards the value of goods and their price. In his work "*An Inquiry into the Nature and Causes of the Wealth of Nations*" [13], Adam Smith presented the water-diamond paradox: water, which is required for life, is far less expensive than diamonds, which have very limited use. The subjective theory of value solved the paradox by claiming that the value of the asset is determined by the consumer, based on the marginal utility. The theory explains that while water, in total, is more valuable than the diamonds, water is plentiful, and diamonds are scarce. Therefore, an additional unit of diamonds exceeds the value of an additional unit of water. Nevertheless, does the paradox hold in the realm of data? The paradox supposes four key properties are observed in most assets: appropriability, divisibility, scarcity, and the display of decreasing returns to use. Appropriability relates to the ownership of data. While data is not divisible *per se*, divisibility could be derived from ownership: access to data could be granted by extending ownership, through a lease, or as a donation. While data is abundant and can be replicated arbitrarily, the scarcity could arise from the finite amount of means to replicate, process and store the data, and from the fact that ownership should be respected. Finally, the decreasing returns in the realm of data could be associated to the degree of information that each new piece of data provides. This is likely to diminish over time. Nevertheless, a fifth factor must be considered: the malleability of the asset under consideration, defined as how a certain asset can be used. The higher the malleability, the greater the market potential and its potential demand. While physical assets have a limited range of uses, each piece of data can be used for a virtually infinite amount of applications, and therefore directly impacting its value. Nevertheless, the subjective value assigned to data in each case may not directly correlate to its pricing. Data can be used in applications that have different value regimes, centered on different value forms (e.g., economic or aesthetic), each of them subject to different internal dynamics [3].

Bolin [3] considers that the following aspects are relevant to data valuation: (a) data is transient (the value of data diminishes over time), (b) it requires human involvement to be generated and processed, (c) data will never be exhausted as long as there is human activity, (d) and it is a non-rivalrous good. We agree with the author that data requires human involvement to be generated and processed. Furthermore, we consider both

properties as the foundation of data ownership. Nevertheless, we consider that while (a) is true for certain cases, many phenomena described by data remains invariant through time (e.g., images describing a landscape). Moreover, technological degradation could impact the ability to produce data. Finally, we agree that data is a non-rivalrous good (the use of data by a company does not infringe upon others' use of it). Jones [11] considers this has at least two consequences: (a) it cannot be priced if not legally restricted (ownership attributed to it), and (b) there may be potentially large gains by using it broadly. Furthermore, it considers that giving data property rights could generate nearly optimal allocations. While we agree that data should be given property rights, we consider that two dimensions of data value must be considered: the ownership of data and the information contained in the data. While the data ownership enables selling or renting a particular piece of data, the information contained in a piece of data may be shared by a wide range of data. We elaborate further on this concept in Section 2.3, linking this property to data pricing.

2.2 Owning data

Ownership is considered a key aspect of pricing. While some authors argue that data exhibits traits of a public good (public goods are non-excludable (it is costly or impossible to exclude someone from using the asset) and non-rivalrous) data is not non-excludable *per se*. Therefore, while some data could be legally turned into non-excludable (e.g., due to public interest or the owners' will), by default, it should be considered private property under the scheme of data markets. We ground this claim in the fact that all data is collected as a result of human intervention and certain investments, and therefore fulfilling the criteria that ownership is gained by doing some work. Nevertheless, data has the particular characteristic that its value relates to the information it holds, which (i) by the definition of information relates to a certain goal, and (ii) can be found in other pieces of data that may be owned by other people. Therefore, while data is owned by the person or entity producing it, the ownership over the information cannot be enforced and could be shared based on data ownership attribution.

2.3 Pricing data

Usually, consumers are willing to pay a higher price for products they consider to be of higher value. Therefore, how should data be priced? Spiekermann et al. [14] explored a user-centered value theory for personal data. Based on experimental research, the authors concluded that (a) most people are not aware that their data may have a market potential, (b) awareness that there is a market for data influences the perceived value of data, (c) the value of data correlates with engagement and psychological ownership (e.g., in a certain application or platform), and (d) lack of control over how data is used likely leads people to abandon the data market.

Data ownership and administration. To solve issues related to peoples' ignorance about data market potential, ensure their psychological ownership and grant them control on how the data is used, we propose regulation should mandate that browsers and devices must have a data management dashboard linked to a digital profile. Such a dashboard could display what data is being collected and provide a typified description on how this data can be used, the privacy implications, and the estimated price a piece of data has on the market. The dashboard should also display which websites /applications/legal entities are accessing the data or have accessed it in the past, the time span for which they stored the data, the purpose for which they use it, and their price offerings. Finally, it should provide data administration tools to operate with the data supporting e.g., the deletion of certain data to anyone who acquired it in the past, disable its further use, or grant it to some particular entity or anyone interested in it.

Such a dashboard could be a product created and marketed by any company interested in providing such oversight. The companies would not store the data: the dashboard would just issue API calls to any third parties and keep track of what data was given or not to particular websites/applications/legal entities. Furthermore, such implementation would provide a default and full GDPR-compliant interface e.g., ensuring the right to data deletion, which under existing implementations is hard to realize. We consider key to data privacy that such dashboards are associated with distributed identities [6]. Furthermore, such a distributed identity could be associated with multiple virtual wallets to preserve data owners anonymity and enable the trading of data.

Data intermediaries. To increase data marketing power and in the interest of privacy, persons could provide some of their data to data intermediaries who would market the data or aggregated data to interested parties under particular terms of use. This would help such parties to acquire a critical mass of data of interest while also increase price negotiation power on behalf of the data producers. The Data Governance Act has already established a legal framework and certain governance standards for data intermediation services [12].

Pricing data. When pricing data, we consider that for each piece of data two things must be considered: (i) the (ownership of the) data itself, and (ii) the information contained by the data. While the data is owned by someone, the information cannot be owned exclusively and is shared across many pieces of data. Therefore, data pricing should consider (i) the compensation paid to the owner for the right to exploit the piece of data with a particular goal, which accounts for the information value of the data in that particular case, and (ii) the compensation paid to anyone who has a piece of data that shares some amount of the information extracted from the piece of data mentioned above. The second compensation is rooted in the fact that given the data is a non-rivalrous good, a single piece of data could be arbitrarily selected and exploited without limit, inducing a certain loss to the rest of the owners of pieces of data that contains similar information. The compensation should alleviate that loss. This second component could be fixed, the amount established by a regulatory entity and paid to a third party, in a similar manner as public performance royalties are managed, collected and distributed by performance rights organizations in the music industry. The royalties would be distributed based on the fraction of information shared by a particular piece of data for which the royalty was paid, and the data owned by a particular person or legal entity. We consider that such an information-sharing-based compensation schema would help to solve attribution issues that arise from generative artificial intelligence models, where no direct attribution to a digital work exists. Furthermore, it would solve issues that arise from competing interests between open-

sourced datasets and private datasets that could contain similar information, compensating for the loss caused to owners of private datasets due to the adoption of opensource (free) ones. This is particularly relevant given the non-rivalrous nature of data.

Renting data. While data could be sold, we consider data renting to provide a more appropriate framework. By renting data, the data producers retain the rights to the data and therefore can decide at any moment to stop sharing it, relocate it, or delete it, among other choices. Data rental could provide a solution to the data portability issue: since the company would not own the data, the data generator retains the right to move the data somewhere else. Therefore, it could be considered that companies take the cost of hosting data as part of the exchange price for data. Nevertheless, they could be mandated to offer a portability service (export some or all of the data producer data on request, for a given fee), to honor the ability to relocate the data. Furthermore, such a service should guarantee that exported data can be understood (e.g., by providing a minimal amount of metadata, with a good-enough semantic description). Specialized companies could provide hosting services for exported data if a person just wants to move the data from some company to avoid losing it when denying further use of it. Furthermore, competing companies could assume the costs of porting data between platforms as a means to lure new consumers to start using their product.

When considering the data rental model, data pricing could have two components: a fixed price paid for the ability to use the data, and a variable price based on the effective value the data provides to the product. The variable component could be measured based on how many requests impact analytic outcomes leveraging certain piece of data, or if a certain piece of information is key to a machine learning model or particular request (e.g., feature significance or other explainable artificial intelligence outcomes, and how these correlate with a particular piece of data). Furthermore, in some cases, to guarantee transparency, the insights used to assess the degree up to which a piece of data is relevant for an outcome should be the ones provided to create explanations as required by regulatory normatives for the use of AI in a product (e.g., the AI Act). The price paid in the market for the (rented) data should be related with the value it provides to a particular feature or product. Furthermore, a fraction of the fixed and variable price should be assigned to the performance rights organizations established to compensate the loss suffered by other data owners whose data contains similar information as the one that was shared.

2.4 Data value chain

We envision the data value chain should have at least three parts: (a) the value of the product (e.g., some application or synthetically generated image - their value is determined by the market price based on specific value regimes), (b) the value of the information extraction process (e.g., artificial intelligence model or analytics - it considers how much of the product value can be attributed to this component (e.g., by number of requests, shared screen time, etc.)), (c) the value of data (determined through some attribution technique, e.g., which variables were most relevant to a forecast, what data contains that information, and in what degree). The data value chain also contemplates at least five distinct actors: (i) consumers (use the application), (ii) data owners renting or selling their data, (iii) data owners compensated (given data shared by third parties contains certain degree of the information contained by their data), (iv) some regulatory entity ensuring such compensations take place, and (v) a person or company that owns and develops the product.

3 CONCLUSIONS

In this paper we have briefly described some considerations regarding the value of data. We consider data ownership is key to realizing data markets, where data rental would provide means to not only pay data owners for their data, but also provide a technical solution that enables the realization of privacy rights. Furthermore, we propose the compensation of data owners based on the information contained within their data and the data shared by third parties. Finally, we propose a data value chain

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency, the European Union's Horizon Europe research and innovation program project Graph-Massivizer under grant agreement HE-101093202 and EU H2020 project STAR under grant agreement H2020-956573.

REFERENCES

[1] Bedford, Norton M., and Mohamed Onsi. "Measuring the value of information-an information theory approach." Management Services: A Magazine of Planning, Systems, and Controls 3.1 (1966): 3.

[2] Bishop, Christopher M., and Nasser M. Nasrabadi. Pattern recognition and machine learning. Vol. 4. No. 4. New York: springer, 2006.

[3] Bolin, Göran. "The Value Dynamics of Data Capitalism: Cultural Production and Consumption in a Datafied World." New Perspectives in Critical Data Studies: The Ambivalences of Data Power. Cham: Springer International Publishing, 2022. 167-186.

[4] Bygrave, Lee A. "The body as data? Biobank regulation via the 'Back Door'of data protection law." Law, Innovation and Technology 2.1 (2010): 1-25. DOI: https://doi.org/10.5235/175799610791935443

[5] Capurro, Rafael, and Birger Hjørland. "The concept of information." (2003).

[6] Dunphy, Paul, Luke Garratt, and Fabien Petitcolas. "Decentralizing digital identity: Open challenges for distributed ledgers." 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). IEEE, 2018.

[7] Ellavarason, Elakkiya, et al. "Touch-dynamics based behavioural biometrics on mobile devices–a review from a usability and performance perspective." ACM Computing Surveys (CSUR) 53.6 (2020): 1-36. DOI: https://doi.org/10.1145/3394713

[8] Gellert, Raphaël. "Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?." Regulation & governance 16.1 (2022): 156-176. DOI: https://doi.org/10.1111/rego.12349

[9] Glazer, Rashi. "Measuring the value of information: The information-intensive organization." IBM Systems Journal 32.1 (1993): 99-110. DOI: https://doi.org/10.1147/sj.321.0099

[10] Iakovakis, Dimitrios, et al. "Touchscreen typing-pattern analysis for detecting fine motor skills decline in early-stage Parkinson's disease." Scientific reports 8.1 (2018): 7663. DOI: https://doi.org/10.1038/s41598-018-25999-0

[11] Jones, Charles I., and Christopher Tonetti. "Nonrivalry and the Economics of Data." American Economic Review 110.9 (2020): 2819-2858. DOI: https://doi.org/10.1257/aer.20191330

[12] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) 2022 (OJ L) 1.

[13] Smith, Adam. "An Inquiry into the Nature and Causes of the Wealth of Nations." Readings in economic sociology (2002): 6-17.

[14] Spiekermann, Sarah, and Jana Korunovska. "Towards a value theory for personal data." Journal of Information Technology 32 (2017): 62-84.

[15] Stylios, Ioannis, et al. "Behavioral biometrics & continuous user authentication on mobile devices: A survey." Information Fusion 66 (2021): 76-99. DOI: https://doi.org/10.1016/j.inffus.2020.08.021

[16] Tian, Yonglong, et al. "What makes for good views for contrastive learning?." Advances in neural information processing systems 33 (2020): 6827-6839