



Is big data a big deal? A competition law approach to big data

Greg Sivinski, Alex Okuliar & Lars Kjolbye

To cite this article: Greg Sivinski, Alex Okuliar & Lars Kjolbye (2017) Is big data a big deal? A competition law approach to big data, European Competition Journal, 13:2-3, 199-227, DOI: 10.1080/17441056.2017.1362866

To link to this article: <https://doi.org/10.1080/17441056.2017.1362866>



© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 12 Sep 2017.



Submit your article to this journal [↗](#)



Article views: 13249



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 3 View citing articles [↗](#)

Is big data a big deal? A competition law approach to big data

Greg Sivinski^a, Alex Okuliar^b and Lars Kjolbye^c

^aAssistant General Counsel, Corporate, External and Legal Affairs Department of Microsoft Corporation, Competition Law Group, Redmond, WA, USA; ^bPartner, Orrick, Herrington & Sutcliffe LLP, Washington, DC, USA; ^cPartner, Latham & Watkins LLP, Brussels, Belgium

ABSTRACT

In this paper, the authors propose a framework to determine the competitive significance of data. The framework first considers whether the parties own or control the relevant data. The second consideration is whether the relevant data is commercially available as a product or as an input for products of downstream competitors. The third consideration is whether the relevant data is proprietary to the owner's or controller's products or services and a competitively critical input. The last consideration is whether reasonably available substitutes for the relevant data exist or whether the data is unique.

ARTICLE HISTORY Received 27 July 2017; Accepted 31 July 2017

KEYWORDS Big data; foreclosure; artificial intelligence; Internet of things

Several years ago, *The Economist* noted that data, particularly consumer data, had become “the new raw material of business: an economic input almost on a par with capital and labour”.¹ Modern computing power has expanded our ability to collect, store, process and analyse data on a large scale, raising complex questions about the commercial nature of the accumulated “Big Data” and the implications for competition in numerous industries across the global economy.² As artificial intelligence (AI), machine learning (ML) and the Internet of things (IoT) promise to make big data analytics a central feature of virtually every area of

CONTACT Greg Sivinski  greg.sivinski@microsoft.com  One Microsoft Way Redmond, WA 98052 USA

¹Kenneth Cukier, ‘Data, data everywhere’ *The Economist* (London, 25 February 2010) <<http://www.economist.com/node/15557443>>.

²World Economic Forum, ‘Big Data, Big Impact: New Possibilities for International Development’ (2012) (discussing value of data as a new asset class and implications for global economy) (hereinafter “Big Data Report”).

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

commerce, competition lawyers, economists and agencies race to define big data in the parlance of antitrust and analyse it under the world's competition laws.

At this point, views differ wildly regarding the relative importance of data to competition and how to analyse it. Some argue that big data is an important barrier to entry because data are difficult to collect, access, replicate and process.³ Others assert that “data-rich companies are not an economic threat, but rather an important source of innovation, which policymakers should encourage, not limit”.⁴ To them, “the notion of data as an anti-trust-relevant barrier to entry is simply a myth” because data – especially consumer data – is readily available and nonrivalrous, permitting simultaneous use by many parties and eliminating the possibility of foreclosure.⁵

Data comes in many forms. It can be as simple as a grocery list and as vast as the ocean of information generated every day by the interaction among billions of people and things across the Internet. Some observers have divided data into three types: “volunteered data” that the user shares intentionally; “observed data” that is obtained by tracking the user's activity in software or online; and “inferred data” that is derived from analysing volunteered and observed data.⁶ Volunteered data tends to be static or “persistent”, such as information and images posted online, a document, or an email. Observed data tends to be “dynamic”, such as a search query, or clicks on a search results page.

From an economics perspective, the data that a firm generates may yield value and efficiencies both along the cost and demand dimensions. On the cost side, having “more” data can have similar effects for the firm to “learning by doing” – improving the way in which the firm can provide a service by getting better information over time and becoming

³Daniel Rubinfeld and Michal Gal, ‘Access Barriers to Big Data’ [2017] *Arizona Law Review* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2830586>; Robert Mahnke, ‘Big Data as a Barrier to Entry’ *CPI Antitrust Chronicle* (May 2015) <<https://www.competitionpolicyinternational.com/assets/Uploads/Mahnke2May-152.pdf>>.

⁴Joe Kennedy ‘The Myth of Data Monopoly: Why Antitrust Concerns About Data Are Overblown’ *Information Technology & Innovation Foundation* (March 2017) <http://www2.itif.org/2017-data-competition.pdf?_ga=1.129315724.1474415869.1490229315>.

⁵Geoffrey Manne and Ben Sperry ‘Debunking the Myth of a Data Barrier to Entry for Online Services’ *Truth on the Market* (March 2015) <<http://truthonthemarket.com/2015/03/26/debunking-the-myth-of-a-data-barrier-to-entry-for-online-services/>>. See also Darren S. Tucker and Hill B. Wellford ‘Big Mistakes Regarding Big Data’ *Antitrust Source* (December 2014), <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2549044> (“Online markets are notable for their low entry barriers and typically do not require big data for entry”); Daniel O’Connor ‘Is Big Data an Entry Barrier? What Tinder Can Tell Us’ *Disruptive Competition Project* (2 April 2015) <<http://www.project-disco.org/competition/040215-big-data-entry-barrier-tinder-can-tell-us/>> (“However plausible this argument sounds, a review of the short history of the Internet economy ... seems to cast doubt on the soundness of the theory”); Andres V. Lerner ‘The Role of “Big Data” in Online Platform Competition’ (24 August 2014) <<http://ssrn.com/abstract=2482780>>.

⁶See, e.g. World Economic Forum ‘Personal Data: The Emergence of a New Asset Class’ 7 (January 2011).

more efficient at its processes. Economists rarely see “learning by doing” economies as a competition problem. On the demand side, more data can make it possible to better target offerings, and better tailor services, to consumers. This may give rise to consumer protection issues, but is unlikely to generate a need for competition intervention.

From a competition perspective, data are a class of assets that vary widely in their competitive significance.⁷ They can be a product, an input for some other product, or commercially irrelevant. Data – of all types – is best analysed using traditional tools of antitrust, albeit in new and varied contexts. Given the technology required to collect and analyse certain types of data, it is conceivable that proprietary data will exist that is necessary to compete effectively, thereby opening the possibility of anticompetitive conduct or effects.

The challenge for enforcers and courts will be to separate cases requiring closer scrutiny from the bulk of cases where data ownership and usage is economically beneficial, drives innovation and is competitively benign.⁸ Ultimately, enforcers and courts should apply traditional tools and avoid acting on models of competition analysis that do not rely on hard evidence about the nature and use of the subject data. They should be sceptical of calls to regulate access to data and take extreme care to avoid false positives. And most importantly, they should resist the temptation to engineer market outcomes that protect parochial interests and stray from protecting the competitive process.

In this paper, we propose a simple framework to determine the competitive significance of data, summarized as follows:

- (1) *Do the parties own or control the relevant data?* It is unlikely that data will be a meaningful factor in any antitrust matter involving market participants that merely *process* data but do not own or control access to the data.
- (2) *Is the relevant data commercially available as a product or as an input for products of downstream competitors?* Competition agencies have

⁷See generally, Big Data Report (n 2).

⁸A McKinsey study estimated that improved use of data could create \$3 trillion in additional value within seven industries, including \$1.3 trillion in the United States. James Manyika and others ‘Open Data: Unlocking Innovation and Performance with Liquid Information’ *McKinsey Global Institute* (October 2013) <<http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information>>. Another McKinsey study estimated that in 2010, a range of free Internet services increased the social surplus within the United States and the European Union by €120 billion, 85% of which went to consumers. ‘Consumers Driving the Digital Uptake: The Economic Value of Online Advertising-based Services for Consumers’ *IAB Europe* (September 2010), <http://youonlinechoices.com/white_paper_consumers_driving_the_digital_uptake.pdf>.

ample experience in evaluating both mergers and conduct issues involving data products and related services.

- (3) *Is the relevant data proprietary to the owner's or controller's products or services and a competitively critical input?* A category of data, like a differentiating product feature, may be a “nice to have” for competitors, but if the data is not *already* a critical input for competing downstream products or services, it is highly *unlikely* to present competition issues.
- (4) *Do reasonably available substitutes for the relevant data exist or is it unique?* The question is not whether a data set is valuable for some purpose, but whether it is unique. Many data already have or could have full or partial substitutes that can be combined, or can be collected by starting new lines of business. If obtaining unique data would create or maintain substantial market power, it is more likely warranted for an agency to assess both horizontal and vertical effects.

Before we begin, a note about privacy and data protection. This paper focuses on the role of data in competition law, particularly with respect to data aggregation. It does not address issues related to data privacy protection.⁹

We begin with a simplified overview of the ML models that make possible the business of big data and then explore our four-step analytical approach.

I. How do machines “learn” – and how does this bear upon competition?

ML and AI generally require vast amounts of data from myriad sources, significant computing power and sophisticated algorithms. Taken together, these elements allow machines to collect, scan and process data, identify behavioural patterns, and then offer predictions for future

⁹Data privacy can of course be a *quality factor* in any given case or merger, but there is not an antitrust market per se for privacy and data protection. See Maureen Ohlhausen and Alexander Okuliar, ‘Competition, Consumer Protection, and the Right [Approach] to Privacy’ (2015) 80 Antitrust Law Journal 121, 133 (“[P]rivacy protection has emerged as a small, but rapidly expanding, dimension of competition among digital platforms”). In *Microsoft/LinkedIn*, the EC also considered privacy an element of quality-based competition. Conceptually this may be true, but it remains to be seen if it is possible to make this approach operational. Consumers attach widely different values to privacy, which confounds rigorous analysis. Even if most consumers would agree that a reduction in product quality is undesirable (at least at constant prices), consumers will differ greatly in the value that they attach to privacy protection exceeding the legal standard that all operators are obliged to respect. EC in *Microsoft/LinkedIn* (Case Com./M.8124), Commission Decision C [2016] 8404 OJ L 1, 3.1.1994 paras 350, n.330 (hereinafter “EC *Microsoft/LinkedIn*”).

conduct by the group of individuals or entities contributing the data. As computers crunch data, they are said to “learn”. How a machine learns in each case will illuminate the relative value of the data, and focus the regulatory issues associated with the monetization of that data.

Data analytics often entails a search for patterns (more precisely, correlations) in data. If a pattern is found, the pattern may well hold true in the future, and so prove helpful for making predictions. In data science terms, the pattern is referred to as the “mathematical model”, which can be implemented using an “algorithm”. An algorithm is

an unambiguous, precise, list of simple operations applied mechanically and systematically to a set of tokens or objects (e.g. configurations of chess pieces, numbers, cake ingredients, etc.). The initial state of the tokens is the input; the final state is the output.¹⁰

The process of looking for patterns is often referred to as “training the algorithm”. The algorithm “learns” by analysing “training data” to find correlations.¹¹ AI employs algorithms that teach machines to learn. ML is a subfield of AI which designs intelligent machines using algorithms that iteratively learn from data and experience, which gives “computers the ability to learn without being explicitly programmed”.¹²

All data in existence are potentially useful for developing, or use in, one model or another. There is no separate market for data specifically used in the development of ML or AI applications – nor are there vendors who offer separate troves of data specifically for use in such applications. The key question rather is what model is being developed in each case and what specifically needs to be deployed to predict? In each instance, many different types of data (including captive first-party data from the

¹⁰Directorate for Financial and Enterprise Affairs Competition Committee’ *Organisation for Economic Co-operation and Development* (9 June 2017) <[https://one.oecd.org/document/DAF/COMP\(2017\)4/en/pdf](https://one.oecd.org/document/DAF/COMP(2017)4/en/pdf)> at 6.

¹¹ibid 7. ML algorithms can be classified into three broad categories, depending on their learning pattern: (1) supervised learning, where the algorithm uses a sample of labelled data to learn a general rule that maps inputs to outputs. (2) Unsupervised learning, in which the algorithm attempts to identify hidden structures and patterns from unlabeled data. (3) Reinforcement learning, case in which the algorithm performs a task in a dynamic environment, such as driving a vehicle or playing a game, and learns through trial and error.

¹²ibid. Deep learning is a subfield of ML that enables computer systems to learn using complex software that attempts to replicate the activity of human neurons by creating an artificial neural network. While traditional ML algorithms are linear, deep learning algorithms are structured in a hierarchy of increasing complexity and abstraction. As a result, deep learning enables computers to learn faster and more accurately than conventional ML, however, today there is no way to know which features or information were used by the algorithm to convert inputs into outputs. In other words, regardless of the quality of the results produced, deep learning algorithms do not provide programmers with information about the decision-making process leading to such results. *ibid*, *citing*, Ian Goodfellow, Yoshua Bengio and Aaron Courville (2016) *Deep Learning* (MIT Press 2016) <<http://www.deeplearningbook.org/>>.

customer itself or commercially available sources) are used as inputs for modelling.¹³

A. Example: linear regression, wine and weather

A simple type of mathematical model is a linear regression. Data scientists and economists use linear regressions to show a correlated relationship between two variables. We can illustrate the concept of training the algorithm by describing the analysis of a small dataset relating to the world of wine.

Princeton economist Orley Ashenfelter caused ripples in the wine world in the early 1990s by developing a model that was able to successfully predict auction prices several years out for mature Bordeaux wine at the time of its initial release, using weather data.¹⁴ This model was particularly exciting to wine enthusiasts and investors because Bordeaux does not reach its full potential until 5–10 years after release, but wine futures are traded well in advance, and wine prices reflect human affairs (subjective taste preferences, brand prestige, etc.), not just weather or other hard data (Figure 1).

Ashenfelter began by collecting the auction price of Bordeaux from dozens of chateaux and comparing that to weather data for about a 25-year period. This is the training data. The table below plots the price of wine (on a logarithmic scale to compress the wide variation in wine prices) on the y -axis against one type of weather data on the x -axis: the average temperature over the entire grape growing season. You can see that the data points are a bit scattered, but wines from warmer summers are generally priced higher. It is a straightforward mathematical exercise to calculate the best straight line through that data (starting at bottom left corner). With this line, the distance from each actual data point to the line is (on average) as small as it can be. The line can be expressed as a simple equation (in the familiar form $y = ax + b$). That equation defines a model for predicting wine prices that can be implemented in code as an algorithm. This model – based on a single data type – is exceedingly

¹³For example, the typical data sources that are used in ML applications for “CRM” lead generation include: (i) marketing automation systems (email open/click and landing page submissions); (ii) web analytics containing website visit and activity data; (iii) mobile app behaviour data; (iv) point-of-sale data from retailers; (v) order/invoice data from the CRM system itself; (vi) search engine result data; (vii) event attendance/registration data; (viii) customer survey data; (ix) e-commerce data, if the customer has an e-commerce channel; etc. And what inputs a particular customer values the most will be highly dependent on what industry they are in. Banks are highly reliant on credit score and FICO data when pitching offers to specific customers, whereas companies owning or managing hotels, bars and restaurants are much more interested in online reservation activity, social media postings, weather information, and the dates and likely popularity of sporting events, trade fairs and conferences.

¹⁴Peter Passell ‘Wine Equation Puts Some Noses Out of Joint’ *New York Times* (New York, 4 March 1990) <<http://www.nytimes.com/1990/03/04/us/wine-equation-puts-some-noses-out-of-joint.html>>.

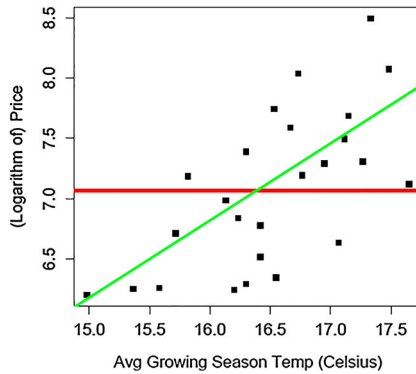


Figure 1. Ashenfelter mode¹⁵.

simple, and yet provides some predictive power. For example, if you had the historical wine prices but no weather data, and were asked to predict the likely future auction price of a new vintage, you would cite the average value of all auction wine in your dataset, here a logarithmic value of 7.07, shown by the horizontal line. The simple model presented here – the line that starts on the bottom left corner – makes considerably better predictions.

Ashenfelter improved his model by considering another type of weather data: the amount of rainfall during the grape harvest season. This could be illustrated by a 3-D chart, adding a z-axis. Again, it is straightforward to draw the best straight line through this set of data points, now called a multiple linear regression. In the same way, additional types of data could be added. Data scientists will experiment by assessing which combinations of data variables yield the greatest predictive power. Ashenfelter ultimately concluded that a good, simple model for predicting the auction price of a bottle of wine is a function of the average growing season temperature, harvest rainfall and rainfall during the preceding winter. Linear regression models work well when the subject of interest, here wine prices, is linearly related to other variables. That is not always the case. Data scientists use a range of other modelling techniques, such as logistic regression, classification, clustering and decision trees, to best uncover patterns in data.

¹⁵Reprinted with permission from Dimitris Bertsimas, 15.071 – The Analytics Edge, The Statistical Sommelier: An Introduction to Linear Regression <https://d37djuv3ytnwxt.cloudfront.net/assets/courseware/v1/7347c7fce82a329565388cd40adce2c9/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/Unit2_WineRegression_AllSlides.pdf>.

II. Do the parties own or control the relevant data?

The first step in our framework is determining who *owns or controls access* to the data? Merely having access to training and production data is not the same as being able to exploit the data for commercial purposes or to restrict access to it by others. Specifically, not all developers have the same level of permission to access and apply volunteered and observed data for the developer's software and services. Data is useful *only if* the data can be accessed and used. If not, then it is not likely to be competitively significant. As a result, competition authorities should distinguish data based on whether the company in question is a *data controller* or merely a *data processor* of that data.¹⁶

A. Data controllers

If a company is the *data controller*, the data it controls is either its own data or it is volunteered or observed data that can be used – typically with the user's consent – for specified purposes that include generating inferred data for improving its own products and services. If a party is only a *data processor*, it generally can access the data only to accomplish the agreed data processing on behalf of the customer (controller) and not for its own use. Thus, the potential competitive significance of a given data set will depend to a significant degree on who is the controller for that data.

Microsoft, for example, is the data controller for its consumer cloud suite of products and services.¹⁷ The terms of use provide that Microsoft may, within the bounds of data protection and privacy laws, aggregate each user's data with data from other users into a “graph” and analyse the data to improve its software and services.¹⁸ Even for data controllers, however, there may also be commercial reasons why it may choose to have a policy not to access or use a consumer's data.

¹⁶See, e.g. Directive 95/46/EC [1995] OJ L 281. The EU's existing data protection regime is set out in the Directive. Per Art. 2(d), “Controller” shall mean the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data ...”. Per Art.2(e), “Processor” shall mean a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller.” To address the difficulties arising under the Directive, the EU has created a new data protection regime – the GDPR. The GDPR entered into force on 24 May 2016. However, enforcement of the GDPR will not begin until 25 May 2018.

¹⁷Consumer Cloud services include Outlook.com, Office 365 Personal, Student and Home, OneNote, OneDrive, Bing, Groove, Movies & TV, etc. Microsoft runs these applications in the cloud. The user logs into the service with her or his user credentials. The user retains full ownership of the data.

¹⁸See also ‘Microsoft Privacy Statement’ Microsoft (June 2017) <<https://privacy.microsoft.com/privacystatement/>>.

B. Data processors

A data processor differs from a data controller because, while it may have *access* to data, a data processor is limited with respect to how it may *use* the data. This kind of situation can arise when one company entrusts data to another company for the limited purpose of processing that data. Enterprises insist on maintaining sole access and control over their data, whether it is stored locally – “on-premises” – or in a third-party cloud such as Microsoft Azure, which provides cloud “tenants” the processing power and applications necessary to manipulate or interact with their *own* data stored in Microsoft’s cloud. Microsoft operates as a data processor for each tenant and it cannot use tenant data generally to improve its products (unless directed to do so by its customers) or create monetization opportunities using that data.

Data for which a firm merely acts as a data processor is not competitively significant in markets where that data processor competes. In *Microsoft/LinkedIn*, for example, the European Commission (“EC” or “Commission”) inquired whether Microsoft and LinkedIn would combine their respective Office enterprise and LinkedIn user data sets. The Commission found that while millions of business users generate large amounts of data and content using Microsoft Office (on-premises or cloud-based), Microsoft was not, as described above, the data controller for enterprise customer content in, for example, Outlook. Regarding whether Microsoft might exclude competition by reserving LinkedIn data to itself, the Commission also found that:

Microsoft is subject to European data protection laws which limit its ability to undertake any treatment of LinkedIn full data. While, today LinkedIn’s privacy policy allows it to share the personal data it collects, processes, stores and uses with its controlling companies, this is only for the purposes described in the privacy policy itself.¹⁹

As a result, the Commission accepted that these constraints would prevent the merged entity from combining wholesale Microsoft business user data and LinkedIn data.

For the same reasons, there are no significant competition concerns in markets for software and services that help owners of datasets collect, store, organize and analyse their own data locally or “on-premises”. For example, Oracle, Microsoft, IBM and many others offer software to store and manage large datasets. These and dozens of other firms also

¹⁹EC *Microsoft/LinkedIn* (n 9) para 255.

offer data analytics software. Simply providing tools that customers can use to store or analyse their data does not convey any special “data asset” to the vendor of those tools. Microsoft is developing its ML and AI technology as part of Microsoft’s “Intelligent Cloud Platform”,²⁰ which will enable all developers and customers to create their own AI solutions using their own data and third-party data to which they have access.

III. Is the relevant data commercially available as a product or as an input for downstream competitors’ products?

After determining ownership or control, agencies should evaluate whether the relevant data is a factor for competition in the relevant market. Most data, big data included, can be parsed into three general categories for analysis: a *product*, such as a commercially available database; an *input* to provide and improve the functionality or utility of analytics services or products; or a *non-commercial asset* – for example, data generated as a side-product of some service that ultimately is not useful for any commercial purpose. Whatever its type, agencies should focus only on data that is relevant from a competitive perspective.

A. Data as a product

Data often possess commercial value as a commodity product. For example, data brokers collect, package and sell databases filled with personal information about consumers – name, address, age, income, job history, online site visit history, buying habits and similar data that possess commercial value for retailers and other businesses. As the Federal Trade Commission (“FTC”) noted in its study of online data brokers, these online behavioural data serve an efficiency-enhancing function in allowing for targeted advertising. The report states that businesses can “purchase information about their customers’ interests in order to market specific products to them, including using consumers’ offline activities to determine what advertisements to serve them on the Internet”.²¹ These types of commercially available data historically have been

²⁰See, e.g. Jason Zander ‘Building the Intelligent Cloud: Announcing New Azure Innovations to Transform Business’ *Microsoft Azure* (29 September 2015) <<https://azure.microsoft.com/en-us/blog/building-the-intelligent-cloud-announcing-new-azure-innovations-to-transform-business/>>.

²¹U.S. Fed. Trade Comm’n, ‘Data Brokers: A Call for Transparency and Accountability ii’ (May 2014) <<https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf>>.

treated by the antitrust agencies as any other product or service, with their competitive significance tied to the nature of demand for the product, the degree of purchaser substitution with other similar commercially available data, the extent of head-to-head price and non-price competition in the sale of the data product, and similar indicia of competitive positioning.

B. Data as an input

Data may also serve as an input to other products and services, specifically as third-party data product that is used in a vertically integrated way, such as when a manufacturer uses Nielsen point-of-sale data to inform its own product design and distribution, or as an internal input by an integrated firm, such as when a search provider uses users' interactions with its search engine to train search algorithms and improve relevance of its results. Data's role as an input is growing as an element of competition because software and online services and IoT increasingly rely on ML and AI that leverage massive data sets for their creation, operation and improvement. For example, travel metasearch site Kayak "uses data mining technology to analyze more than one billion queries run by consumers on its websites to forecast price trends on flights for specific routes".²² Kayak's service relies on volunteered and observed user search data as a key input. We can observe similar returns to scale and feedback effects with spell checkers, speech recognition software or other online shopping applications and services – the more frequently these systems are used, the more data they collect and the smarter they become. Data that serve to develop products and services in this way *and* that competitors could use to create or improve their products has competitive utility.

Key takeaway: Captive internally generated data is unlikely to give rise to competition problems, although there is the possibility that competition issues *may* arise when datasets are truly unique *and in the case of mergers* transferable under privacy and data protection laws.

C. Data that is not useful competitively

Not all data is commercially significant. In data-driven applications, such as the futures estimator in our wine example, it is usually impossible to know all the purposes for which data will be useful. First, some data has

²²Ohlhausen and Okuliar (n 9) 131 (internal citations omitted).

no predictive value in a given context and therefore is not useful competitively. Second, some data has utility but it is not known in advance. In this category, regulators assessing a merger should not speculate about some future value discovery. Indeed, ML tools are built on algorithms that learn and make predictions based on very large and often disparate sets of data. What value a given data set may have for ML may not be known or is at least uncertain until *after* a ML algorithm has been “trained” on the data as described above.

D. Data products cases

a. Horizontal mergers – direct competitors

Mergers between direct product competitors pose the most significant risk of an agency challenge, particularly where the product market is concentrated or there are limited available substitutes for the merging data products.

In *Dun & Bradstreet-Quality Education Data* (2010), for example, the FTC alleged that the parties “were the only significant U.S. suppliers of [K-12] educational marketing data”.²³ The FTC’s complaint alleged that the data included contact, demographic and other information about teachers, administrators, schools and individual school districts. The parties’ customers relied on the data to market unrelated products and services to teachers, administrators and other school personnel using both direct mail and email.²⁴ The FTC “determined that the parties’ customers did not regard other sources of marketing data as close substitutes. The data, by its unique characteristics, had greater utility and value to customers than alternative datasets”.²⁵ The agency was concerned about the unilateral effects of the deal and saw only one other competitor in the space –

²³Analysis of Agreement Containing Consent Order to Aid Public Comment, *Dun & Bradstreet Corp.*, Dkt. No. 9342’ (10 September 2010) 1 <<https://www.ftc.gov/sites/default/files/documents/cases/2010/09/100910dunbradstreetanal.pdf>>.

²⁴Edith Ramirez, *Deconstructing the Antitrust Implications of Big Data, Keynote Remarks of FTC Chairwoman Ramirez* (Fordham Competition Law Institute, 22 September 2016), 3–4 (describing matter); *Dun & Bradstreet Corp.*, Dkt. No. 9342 (F.T.C., 7 May 2010) <https://www.ftc.gov/sites/default/files/documents/cases/2010/05/100507_dunbradstreetcmpt.pdf>. The FTC pursued a similar action in 2008 against the merger of Reed Elsevier’s subsidiary, LexisNexis and Choicepoint. They were alleged to be the two largest providers – and head-to-head competitors – of “electronic public record services for law enforcement customers”. *Reed Elsevier NV et al.*, File No. 081-0133 (F.T.C., 16 September 2008) <<https://www.ftc.gov/sites/default/files/documents/cases/2008/09/080916reedelseviercpcmpt.pdf>>. In that matter, the agency noted that even though the type of records were readily available to other companies, entry would be prohibitively difficult because of the time and expense involved in creating analytical tools that would deliver comparable services and gaining acceptance with customers. *See also* Ramirez (n 24) 3.

²⁵Ramirez (n 24) 4.

a distant number three player. The matter was resolved after four months of administrative litigation with a consent agreement calling for the divestiture to a fringe player of a K-12 database and the QED brand.²⁶

In *Thomson/Reuters* (2008), the merging parties were horizontal competitors that sought to combine certain financial data sold to traders and other financial professionals. Both the Antitrust Division of the U.S. Department of Justice (“DOJ”) and the EC imposed remedies after determining that the merging parties were leading providers of specific types of financial data products, and other companies likely would be unable to offer substitutable data because of significant entry barriers.²⁷ Per the Commission: “The hurdles come, first, from the need to collect fundamental data with a global coverage and second, to collect fundamental data going back in time several years.”²⁸ The EC further explained that, per market participants, “the raw materials needed to create these databases are simply unavailable at any price”. The DOJ found that new entrants into the fundamentals data market, particularly with respect to international fundamentals data, must overcome significant barriers to entry. These include the difficulties of arranging for collection of data on tens of thousands of companies on a global basis, constructing a reliable historical database, the need to develop local expertise in each country’s accounting norms, and the ability to develop data normalization and standardization processes. Therefore, entry or expansion by any other firm will not be timely, likely, or sufficient to defeat an anticompetitive price increase.²⁹

In *Bazaarvoice/Power Reviews*, Bazaarvoice was the market-leading provider of ratings and review platforms that enable manufacturers and retailers to collect, organize and display consumer-generated product reviews and ratings. In June 2012, Bazaarvoice acquired its primary competitor, PowerReviews, and the DOJ successfully sued to unwind the deal. In opposing the deal, the DOJ cited many party statements revealing head-to-head competition, and the court found that “[t]he acquisition of PowerReviews amplifies [Bazaarvoice’s] access to ... consumer behavior data and brings significant opportunities for syndication, advertising, and

²⁶ibid.

²⁷Commission, *Mergers: Commission Clears Acquisition of Reuters Subject to Conditions* <http://europa.eu/rapid/press-release_IP-08-260_en.htm>; see also Complaint, *U.S. v. Thomson Corp. and Reuters Group*, No. 1:08-cv-262 (D.D.C., 19 February 2008) <<http://www.justice.gov/atr/cases/f230200/230281.htm>>.

²⁸EC in *Thomson/Reuters* (Case No. COMP/M.4726), Commission Decision C (2008) 654, OJ L 24, 29.1.2004, p.1, 361.

²⁹ibid, paras 365–370; see also Complaint, *U.S. v. Thomson Corp. and Reuters Group*, No. 1:08-cv-262, para 37 (D.D.C., 19 February 2008) <<http://www.justice.gov/atr/cases/f230200/230281.htm>>.

data”.³⁰ This is the type of dynamic data for which there are generally not reasonably available substitutes, and the DOJ determined that the post-merger entity would control enough such data to foreclose rivals.

The FTC found in *Nielsen Holdings/Arbitron* that the proposed merger would cause the “elimination of future competition between Nielsen and Arbitron”, which would “likely cause U.S. customers to pay higher prices for national syndicated cross-platform audience measurement services and result in less innovation for cross platform measurement services”.³¹ Per the FTC, these two companies alone were best-positioned to develop this *future product* because they were the only firms with large, representative panels capable of reporting TV programming viewership, including individual demographic data such as age and gender information. To ensure that the merger did not eliminate emerging competition for *future cross-platform products*, the FTC required Nielsen to divest and license assets, including a royalty-free license to Arbitron’s data for eight years, so that an FTC-approved buyer could successfully develop a cross-platform service to compete with Nielsen’s future offerings.³²

b. Vertical mergers – acquisitions of complementary but critical data inputs

In a vertical case, *Google/ITA* (2011), ITA was a supplier of an airline schedule database to various industry online travel intermediaries as an input for their own products. Google announced its acquisition of ITA while it announced it intended to make available certain online travel data to its search customers. The DOJ reviewed Google’s acquisition of ITA as a vertical merger involving a company purchasing a critical input supplier, and they imposed remedies designed to ensure the continued supply of those inputs to Google’s online travel competitors.³³ DOJ was preoccupied mainly with ITA’s search tools and algorithms, which could efficiently

³⁰*U.S. v. Bazaarvoice, Inc.*, No. 13-cv-00133-WHO, 2014 WL 203966, para 83 (N.D. Cal., 8 January 2014) (internal quotation and citation omitted).

³¹Analysis of Agreement Containing Consent Order to Aid Public Comment, *In the Matter of Nielsen Holdings N.V. and Arbitron Inc.*, Dkt. No. C-4439’ (20 September 2013) 3 <<https://www.ftc.gov/sites/default/files/documents/cases/2013/09/130920nielsenarbitronanalysis.pdf>>.

³²*ibid*; *Nielsen* was not without its critics. FTC Commissioner Josh Wright dissented from the FTC decision and explained the economics in a subsequent article. See Dissenting Statement of Commissioner Joshua D. Wright, *In the Matter of Nielsen Holdings N.V. and Arbitron Inc.* (20 September 2013) <https://www.ftc.gov/sites/default/files/documents/public_statements/dissenting-statement-commissioner-joshua-d-wright/130920nielsenarbitron-jdwstmt.pdf>. In any event, it should not be extended to impose the same conditions in a merger where the parties’ proprietary data was not already being offered in current products in the market. The speculative nature of the theory, in combination with the fact it would likely harm rather than promote competition in such a scenario, strongly suggests that course would be a mistake.

³³Competitive Impact Statement, *U.S. v. Google Inc. and ITA Software, Inc.*, No. 1:11-cv-688 (D.D.C., 8 April 2011) <<https://www.justice.gov/atr/case-document/file/497671/download>>.

search airline airfare databases. The airfare data that ITA accessed and processed for its customers were not owned directly by ITA. However, importantly, ITA could access, aggregate and reconfigure the data, as well as use cached results data in its most sophisticated offering, that made it a unique product able to provide accurate and very rapid – nearly instantaneous – results for its customers. The DOJ focused on the data-fed products when evaluating competitive effects and barriers to entry in this matter, although the product analysis was heavily linked with the underlying data.

IV. Is the relevant data proprietary and captive to the owner or controller's own products or services?

The more nuanced cases will involve the acquisition or use of proprietary or captive data sets used as inputs only to the products or services of the data owner or controller. These data are not available to competitors or third parties generally and are unlikely to represent unique critical inputs in any competitive sense: importantly, to the extent a vertically integrated data owner has downstream competitors, it would strongly suggest that those competitors have access to comparable internal or external data sets. Nonetheless, some commentators observe that companies undertake data-driven strategies to obtain and sustain a competitive advantage, which they argue may affect competition in three general ways:

First, the mere possession of large amounts of data gives a company a significant competitive advantage that its rivals will be unable to challenge. Second, competition policy as it is currently practiced is unable to respond to competitive threats stemming from large amounts of data. Third, the acquisition of large amounts of data about users presents a serious threat to privacy that consumer-protection authorities are unable to handle.³⁴

However, using data to seek a competitive advantage is economically efficient behaviour that drives innovation.

Thus, the threshold question is at what point can economically efficient and legal pro-competitive conduct stray into illegal exclusionary conduct? Focusing first on the *volume* of data, simply having more data than anyone else does not protect a company from competition:

The unstable history of digital business offers little evidence that the mere possession of big data is a sufficient protection for an incumbent against a superior

³⁴Kennedy (n 4) 4, responding to Maurice E. Stucke and Allen P. Grunes, 'Debunking the Myths over Big Data and Antitrust' *CPI Antitrust Chronicle* (May 2015), 2–3 and sources cited therein.

product offering. To build a sustainable competitive advantage, the focus of a digital strategy should therefore be on how to use digital technologies to provide value to customers in ways that were previously impossible.³⁵

Focusing second on the *nature* of the data, predicted anticompetitive outcomes assume to a significant degree that all data are competitively useful, and that most data are unique and without reasonable substitutes. This ignores the counterfactual reality in most cases that the data is not essential to compete or there are reasonable substitutes such that the way in which the owner or controller may choose to leverage that data should not raise a significant competition issue.

If the data is proprietary – not shared with others or only shared under certain conditions that preserve its proprietary nature – then the competition agencies are likely to follow precedent and focus their analysis on competition among products that compete with the controller’s products using the proprietary data as an input. In these situations, a category of data, like a differentiating product feature, may be a “nice to have” for competitors, but if the data is not *already* a critical input for those competitors’ downstream products or services, it is highly *unlikely* to be necessary to compete in the downstream market. Indeed, in such cases, the risk of freeriders is the greater concern.

This distinction between a “nice to have” and a “must have” is also the reason that even if a data product or a proprietary dataset is a *competitively useful* input, it does not raise competition concerns unless the company controlling that data realistically can use it to *foreclose competition in a downstream market*. As the U.S. DOJ has long observed, such an input foreclosure theory typically requires that the parties post-merger would have market power both upstream and downstream.³⁶ In addition, “before the Department would find harm to competition ... , there must be a probable downstream output or price effect”. Thus, foreclosure is less viable to the extent other market participants can compete using either alternative data sets or alternative approaches that do not depend on the parties’ data or similar data. The EC takes a similar approach. First, would the merged firm have the *ability* to foreclose its actual or potential competitors, second, would it have the *economic incentive* to do so and, third, would a foreclosure strategy have a

³⁵Anja Lambrecht and Catherine E. Tucker, ‘Can Big Data Protect a Firm from Competition?’ *Antitrust Chronicle* 1 (January 2017), no. 12, 17.

³⁶Steven Sunshine, Vertical Merger Enforcement Policy, Address of the Deputy Assistant Attorney General Before the American Bar Association (5 April 1995), <<https://www.justice.gov/atr/speech/vertical-merger-enforcement-policy>>.

“significant detrimental effect on competition, thus causing harm to consumers”.³⁷

These models for exclusion set a high bar for proof, while in today’s increasingly data-rich world, many types of data are reasonably replicable using available substitutes, meaning that one party’s control over such data is unlikely to foreclose competition. As in most cases, such substitutability should also be based on the function performed by the data, recognizing as well that the technology used to generate its substitutes can be different. This is particularly important in circumstances where substitutability is hard to measure (for example, because it is not traded on an open market) or no one knows or can show how the data might contribute to ML or AI.

V. Do reasonably available substitutes for the relevant data exist or is the data unique?

If data is necessary to compete in a relevant product market, substitutability turns mainly on the extent of overlap between the functionality of the subject data and any potential alternatives as well as the nature and extent of market demand for such data and its possible alternatives (assuming commercial availability). Only with careful study of these factors can an agency or other factfinder assess whether data that is the subject of an investigation is uniquely able to provide a necessary input – for instance, the specific training data for ML or AI – or whether reasonably available substitutes exist. To the extent that a firm can acquire or combine data from other sources or recreate relevant data or to train an algorithm by other means, even if with some effort, those data sources may be able to provide the substitute necessary for competition in that relevant market.

A. Evidence germane to the question of data substitutability

In *Microsoft/LinkedIn*, the EC found that the merged entity would not have the ability to foreclose competing providers of customer relationship management (CRM) software solutions if it, hypothetically, reduced access to LinkedIn full data because it would be unlikely to negatively affect the overall availability of substitutable data required for ML in CRM software solutions.³⁸

³⁷EC *Microsoft/LinkedIn* (n 9), para 186 (emphasis added).

³⁸*ibid*, paras 253–254.

The Commission found that LinkedIn full data, or a subset thereof, did not appear to be a unique input with respect to the provision of ML in CRM software solutions.³⁹ On the one hand, the market investigation revealed that all CRM competitors and half of the customers considered LinkedIn full data to be important for ML in CRM software solutions, either now or in the near future. However, the investigation also showed that major CRM vendors had already started offering advanced functionalities to their CRM customers based on ML, or had planned to do so in the next two to three years. None of those offerings had been developed or required for its use access to LinkedIn full data.⁴⁰

Therefore, the Commission found that even if LinkedIn full data were to be used soon in ML for CRM software solutions,

it would constitute only one of the many types of data which are needed for this purpose. Indeed, the data that are needed for ML in CRM software solutions come from essentially two data sources: in-house customer data uploaded in the CRM software and complementary third party data. In-house customer data uploaded in the CRM software relates to accounts, service tickets, interactions, leads, etc. These data are by definition available to each relevant provider of CRM software solutions and availability of such data will not be affected by the Transaction.⁴¹

The Commission also found that third-party data required for ML can be viewed and treated differently based on the use case and the relevant industry. The data collected by LinkedIn were one source of the third-party data which could be used for ML and may be relevant for certain use cases in certain industry sectors, but not for others. For example, the Commission found that LinkedIn full data may be relevant for the CRM B2B Sales and B2B Marketing sub-segments of the CRM market generally, but not for others:

In this regard, SAP stated that “LinkedIn is only one data source. Depending on the use case, other types of data might be more relevant than LinkedIn. It is difficult to predict how this will evolve in the future.” In the same vein, Oracle explained that “there is not one dataset with the highest value [as input for ML], but that it is about having numerous types of data. Therefore, not only the quality, but also the quantity and the variety are important.”⁴²

And finally, the Commission found that there are many other possible sources of data which are already available for ML, including:

³⁹ibid, para 256.

⁴⁰ibid, paras 257–258.

⁴¹ibid, paras 259–260.

⁴²ibid, paras 260–261 (emphasis original).

Dunn & Bradstreet, who explained that its data:

allows the onboarding [sic] and analysis of data from any source. To that end, data acquired from any sales intelligence provider could be used in that fashion if licensed for such a purpose either through a partnership or by the end customer. Avention Data.com Dun & Bradstreet InsideView Twitter And others.

CRM customers such as IBM explained that “LinkedIn data is very useful but is not the only source of data. There are many sources of unstructured information about commercial markets and cognitive solutions can inter-rogate and make sense of those.” Another stated that:

[t]he data needed to leverage predictive analytics to target companies using qualitative needs-based solutions is already readily available. This data includes company firmographics [sic], news, regulatory change, information about companies similar to the prospect, competitor announcements, internal intelligence on win loss reasons, publicly available leadership change information, pending legal issues and financial earnings announcements.

Where the data is *observed data*, attached to an event or user interaction after the event or interaction has taken place, the opportunity to gather that data may be lost. However, for online services, barriers to switching are low and consumers do not face either-or decisions; for example, they can choose to use more than one social network or consumer communications service. Multi-homing intensifies competition and reduces barriers to entry (including network effects). In a world of free services and multi-homing, every minute of every user’s available time is contestable every day. In *Facebook/WhatsApp*, for example, the EC found that the use of one consumer communications app does not exclude the use of competing consumer communications apps by the same user. Most users of consumer communications apps in Europe have installed and are using two or more consumer communications apps mitigating network effects. Thus, even for observed and inferred data, if there is “multi-homing” on competing services, similar data may be easily available. Moreover,

information is also generally not excludable. Platforms have a difficult time preventing their competitors from gathering data on their own. The more valuable a piece of data is, the more likely it is that more than one company will seek to acquire it.⁴³

Key takeaway: A scenario may exist where a firm (i) owns or controls a data set, (ii) uses the data set as an input, (iii) that data input is necessary

⁴³Kennedy (n 4) 7.

for competition, (iv) there are no reasonable substitutes; however, this scenario remains very rare in practice, in part given the generally vast and increasing quantities of data that are created every day. As one looks at how other antitrust theories have developed, such as essential facilities for intellectual property or physical assets, one sees those clear parallels.

B. Examples of agency analysis regarding substitutability and related issues in proprietary data input transactions

Conglomerate mergers of proprietary data inputs for which there are no reasonable substitutes, *and* where the acquisition would *also* create or maintain *significant market power* of the buyer in a related market, are a Holy Grail for enforcement agencies. There will be relatively few cases, however, where the acquired company possesses competitively significant proprietary data assets. This is because such captive data assets are most often used to compete in zero-price digital markets (for example, social networks, algorithmic search). Price is not a factor *vis-a-vis* the users, which means we can be fairly certain that quality and the rate of innovation are the primary dimensions of competition in these consumer-facing markets and that data (and the quality of data analytics) are important factors for this quality and innovation competition, such as for improving existing products, adding new features and coming up with new products. The courts and federal agencies have existing tools to evaluate this kind of non-price competition in situations where the data is captive and, therefore, not already being monetized directly.

The FTC and the DOJ have set out an analysis in the Horizontal Merger Guidelines for examining innovation where it is a prominent feature of the competitive dynamic. In mergers affecting innovation, they “consider whether a merger will diminish innovation competition by combining two of a very small number of firms with the strongest capabilities to successfully innovate in a specific direction”.⁴⁴ Where data and data analytics are closely tied to product innovation, this test may have some attraction. Innovation competition seeks to capture the dynamics in a race to market, before something has become a product. The factors that matter are know-how, talent and (perhaps)

⁴⁴U.S. Dep’t of Justice & Federal Trade Comm’n, 2010 Horizontal Merger Guidelines, §6.4 <<https://www.justice.gov/atr/horizontal-merger-guidelines-08192010#6d>>.

data and data analytics. If so, then agencies should ask whether a given merger would lead to the elimination of one “of a very small number of firms with the strongest capabilities to successfully innovate in a specific direction”.⁴⁵ Regarding data assets, almost always the answer will be “no”.

For example, in 2010 the DOJ reviewed and approved Microsoft’s search affiliation with Yahoo!, even though it would reduce the number of Internet search and search advertising competitors from three to two major players.⁴⁶ In support of its decision, the DOJ noted:

The transaction will enhance Microsoft’s competitive performance because it will have access to a larger set of queries, which should accelerate the automated learning of Microsoft’s search and paid search algorithms and enhance Microsoft’s ability to serve more relevant search results and paid search listings, particularly with respect to rare or “tail” queries. The increased queries received by the combined operation will further provide Microsoft with a much larger pool of data than it currently has or is likely to obtain without this transaction. This larger data pool may enable more effective testing and thus more rapid innovation of potential new search-related products, changes in the presentation of search results and paid search listings, other changes in the user interface, and changes in the search or paid search algorithms. This enhanced performance, if realized, should exert correspondingly greater competitive pressure in the marketplace.⁴⁷

The agency went on to observe that market participants had indicated that “combining the parties’ technology would be likely to increase competition by creating a more viable competitive alternative to Google, the firm that now dominates these markets”.⁴⁸

The EC’s decision in *Facebook/WhatsApp* also reflects a traditional approach based on aggregation of data used as an input into a related relevant market. Data was generated by an activity (consumer communications) and this data could be used as an input for another activity (online non-search advertising) and the concern was adverse price effects on advertising services. There was no link to innovation in consumer communications services or social networking services. The EC

⁴⁵ibid.

⁴⁶Press Release, Dep’t of Justice, *Statement of the Department of Justice Antitrust Division on Its Decision to Close Its Investigation of the Internet Search and Paid Search Advertising Agreement Between Microsoft Corporation and Yahoo! Inc.* (18 February 2010) 1 <<https://www.justice.gov/opa/pr/statement-department-justice-antitrust-division-its-decision-close-its-investigation-internet>>.

⁴⁷ibid.

⁴⁸ibid.

considered how data from WhatsApp user services could be used as an input in Facebook's online display advertising. It found that:

even if the merged entity were to start collecting and using data from WhatsApp users, the Transaction would only raise competition concerns *if the concentration of data within Facebook's control were to allow it to strengthen its position in advertising.*⁴⁹

The facts did not justify intervention in that case, because there continued to be “a large amount of Internet user data that are valuable for advertising purposes and *that are not within Facebook's exclusive control*”.⁵⁰ Ultimately, the Commission found the availability of substitute data to be decisive.

Microsoft controls user data from its consumer cloud services and uses that data – with users' consent – to improve its consumer offerings such as Bing search and Outlook.com. In *Microsoft/LinkedIn* the EC addressed whether a hypothetical combination of Microsoft's consumer cloud dataset with LinkedIn's user profile data might conceivably create market power in a hypothetical market for supply of data or increase barriers to entry for competitors that need the data to compete in online advertising. The Commission did not need to rely on innovation competition to deal with the case. There were no horizontal effects because the parties were small players in online advertising and there were no non-horizontal effects because many other data sources were available. Ultimately, the Commission concluded that this would not give rise to serious competition concerns for several reasons. First, pre-merger, Microsoft and LinkedIn did not make available their data to third parties for advertising purposes (with very limited exceptions). Second, large amounts of Internet user data valuable for advertising remain available and outside Microsoft's exclusive control, and third, Microsoft and LinkedIn were small players in online advertising competing only to a limited extent.

Importantly, the investigation also addressed data access and innovation, i.e. the substitutability (i.e. uniqueness) of LinkedIn data that could in theory be used with Microsoft's ML capabilities to improve lead generation capabilities of Microsoft's Dynamics CRM software. Salesforce argued that access to full LinkedIn data, including metadata, was necessary to develop ML functionality, the “key” for next-

⁴⁹EC Facebook/WhatsApp Decision (Case COMP/M.7217) Commission Decision C (2014) final para 187 (emphasis added).

⁵⁰*ibid*, para 189 (emphasis added).

generation CRM solutions. The EC dismissed the concerns for several reasons, including the following. First, there was no evidence that absent the transaction LinkedIn would license its data to third parties. The Commission found that LinkedIn did not appear to have a significant degree of market power in any potential relevant upstream market for the provision of data for the purposes of ML in CRM software solutions because

LinkedIn does not currently license any [of its] data to any third party ... Moreover, LinkedIn's internal documents show that, absent the Transaction ... no reference was made to the possible licensing of LinkedIn full data, or a subset thereof, to any third party, including for [machine learning] purposes.⁵¹

Second, Microsoft access to the LinkedIn data might enable it to offer new or improved products (e.g. a merger efficiency). Third, CRM competitors including Salesforce were already developing ML without access to LinkedIn data, i.e. LinkedIn was one of many third-party data sources and it was difficult to predict how the market would evolve. Perhaps most importantly, the EC recognized that ML is based on a vast array of data sources, including first-party data. The LinkedIn data would be one of many dozens of sources of data already being used to improve ML in CRM lead generation.

These findings reflect a critically important aspect of ML and AI, which at their core rely on dynamic experimentation. Acquiring new data that complement internally generated data enables new ways of differentiating products and services. Indeed, “big data” in ML and AI is increasingly being understood to mean the application of *a multiplicity of signals* rather than just sheer scale of data. Increasingly it is the *multidimensionality* of the data inputs that matters. Moreover, having many signals in the same set of data is not the goal; rather, the goal is to include a greater number of diverse signals, which provides more explanatory power because the results are a better fit for the intended use. This is not the same as having just *more* data – because the benefits of sheer scale can diminish rapidly – but *better* data suited to enhancing particular products or services.

Salesforce based its opposition to the *Microsoft/LinkedIn* merger on the need for a “level playing field” that would have required LinkedIn to share its heretofore proprietary data with a competitor as a condition to the merger with Microsoft. This argument of course ignored the

⁵¹EC Microsoft/LinkedIn Decision (n 9), para 254 (emphasis added).

critically important fact discussed above that a merger which allows innovative experimentation by combining complementary data sets *enables new differentiation*, and is thus *good competition*. Had this argument been accepted by the Commission, it would have relied upon an “efficiency offense” against the established presumption that deals which do not create overlaps but combine complementary assets and capabilities are pro-competitive.⁵² The core principle here is that a “level playing field” is not the goal of competition law generally. Deeper integration through mergers can allow for experimentation in a way that is not possible through just arms-length collaboration, which in turn will stimulate rivalry in innovation while not rewarding free riders.

Key takeaway: agencies and other institutions should proceed with great caution to understand in depth what data is relevant in each merger, whether those data is unique and transferable, and alternatively whether there are reasonable substitutes for the data, before taking any steps that could depart from consistent application of traditional merger review principles. This is particularly important for vertical mergers in which the parties do not have overlapping lines of business, and in cases where the relevant data is a proprietary input. Agencies worldwide should resist pressure to accept the “efficiency offense” from disgruntled competitors who must adapt and innovate to compete, or to apply novel theories to condition deals on compulsory access to proprietary data.

C. Unilateral conduct involving unique captive data sets

Outside of the merger realm, existing antitrust frameworks can apply, with adjustments to account for the volume, variety and nature of the data

⁵²See, e.g. General Court Judgment of 4 July 2006, *easyJet v. Commission*, ECLI:EU:T:2006:187, where the General Court noted that efficiencies can only lead or strengthen a dominant position in exceptional conditions that include other exclusionary conduct:

72. ... It should be noted in this regard that merger control is not premised on the prohibition of such advantages, but on the aim of avoiding the creation or strengthening of a dominant position as a result of which effective competition could be significantly impeded in the common market. The ability as a result of the merger to offer passengers services at a better price could only constitute evidence of the creation or strengthening of a dominant position in limited cases, for example where the merged entity intends or has the capacity to operate a predatory pricing policy.

assets. As it stands today, there are two antitrust theories that might be used to address data issues when it comes to unilateral conduct, both drawn from years of development around existing theories.

First, one could use the framework above to assess whether a firm that has acquired a position in data – which is controlled, necessary for competition, has no reasonably substitutable sources, etc. – *and* could engage in exclusionary conduct that prevents competitors from gaining access to that essential user-generated data. As a theory, enforcement agencies and courts are well-positioned to consider the competitive effects and foreclosure issues that would determine whether to find liability under a given scenario. The key will be whether one can credibly prove foreclosure considering growing evidence of the complementarities and multidimensionality of data as assets, multi-homing, etc. It is increasingly clear that the notion that rivals will be materially marginalized and undermined because they do not have a particular set of data at the ready is remote at best.

Second, one could use the framework above to assess whether, in rare cases, a firm should be forced to share proprietary data that some third-party claims are necessary for competition and for which there are not reasonable substitutes. Forced sharing without a prior course of dealing is virtually non-existent in the United States after *Trinko*.⁵³ Whether evaluating unilateral refusals to deal under an input foreclosure or essential facility theory, American courts and agencies generally have taken a deferential stance to companies that refuse to provide their property to competitors, grounding this position of deference in the nearly century-old Supreme Court principle that a party may “freely ... exercise his own independent discretion as to parties with whom he will deal”.⁵⁴ When confronted with these refusals to deal, they have focused their analysis mainly on whether or not the refusal is based on a legitimate business reason.

In *Morris Communications v. PGA Tour*,⁵⁵ the PGA refused to grant media companies access to its tournaments unless they had first agreed

⁵³*Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP*, 540 U.S. 398 (2004).

⁵⁴*United States v. Colgate & Co.*, 250 U.S. 300, 307 (1919); *Trinko*, 540 U.S., 407–08 (“Compelling ... firms to share the source of their advantage is in some tension with the underlying purpose of antitrust law, since it may lessen the incentive for the monopolist, the rival, or both to invest in those economically beneficial facilities. Enforced sharing also requires antitrust courts to act as central planners, identifying the proper price, quantity, and other terms of dealing – a role for which they are ill suited. Moreover, compelling negotiation between competitors may facilitate the supreme evil of antitrust: collusion. Thus, as a general matter, the Sherman Act “does not restrict the long recognized right of [a] trader or manufacturer engaged in an entirely private business, freely to exercise his own independent discretion as to parties with whom he will deal”).

⁵⁵364 F.3d 1288 (11th Cir. 2004).

not to sell the PGA's proprietary compiled real-time golf scores, known as the Real-Time Scoring System ("RTSS"), to noncredentialed third-party Internet publishers – affecting the ability of these media companies to report real-time scores in competition with the PGA's service. Morris sued the PGA for monopolization under Section 2 of the Sherman Act, claiming it should be free to resell RTSS compiled scores notwithstanding the PGA's investment in, and ownership of, the data product. Morris contended that the PGA was monopolizing the markets for the publication of real-time golf scores on the Internet, and the sale, or syndication of such scores. The PGA Tour responded that it was protecting its significant investment in RTSS – which included personnel and technology – from "free-riding" by competitors.

The district court rejected Morris's monopolization claims and the Eleventh Circuit affirmed, holding that the prevention of free-riding was a legitimate business justification. The court reasoned that the compiled real-time golf scores were not a product that Morris had a right to sell because they were a derivative product of RTSS, which PGA owned exclusively. Moreover, the court found that "Section 2 did not require PGA to give its compiled scores freely to its competitors" and observed that it "is not a function of the antitrust laws' to equip plaintiffs with defendants' competitive advantages".⁵⁶ The Eleventh Circuit also stated that "a company – even a monopolist company – that expends time and money to create a valuable product does not violate the antitrust laws when it declines to provide that product to its competitors for free".⁵⁷

Except for certain FRAND-encumbered patents, in recent years the FTC and DOJ have articulated similar positions of deference for unilateral refusals to deal.⁵⁸ In *Mylan v. Celgene*,⁵⁹ it was alleged that Celgene had declined to supply trial samples of its drugs to a potential rival, Mylan, knowing that Mylan required the samples to prepare generic versions. Mylan had no

⁵⁶*ibid*, 1296.

⁵⁷*ibid*, 1298. The courts' deference to unilateral refusals to deal is even more pronounced when the refusal involves intellectual property, with certain exceptions for FRAND-encumbered patents. See, e.g. *Image Technical Services v. Eastman Kodak Co.*, 125 F.3d 1195 (9th Cir. 1996) (providing a presumption of lawfulness that can be rebutted on a showing of pretext); *In re Independent Service Organizations Antitrust Litigation*, 203 F.3d 1322, 1327 (Fed. Cir. 2000) (providing very significant latitude for refusals to deal and noting in particular that "[i]n the absence of any indication of illegal tying, fraud on the Patent and Trademark Office, or sham litigation, the patent holder may enforce the statutory right to exclude others from making, using, or selling the claimed invention free from liability under the antitrust laws").

⁵⁸See, e.g. Decision and Order, *In re Motorola Mobility LLC and Google, Inc.*, Dkt. No. C-4410 (FTC, 24 July 2013) (determining that the pursuit of injunctive relief and related unilateral conduct involving FRAND-encumbered standard-essential patents could constitute a violation of Section 5 of the FTC Act).

⁵⁹See Fed. Trade Comm'n, Brief as Amicus Curiae, *Mylan Pharmaceuticals, Inc. v. Celgene Corp.*, No. 2:14-CV-2094-ES (D.N.J., 17 June 2014) <https://www.ftc.gov/system/files/documents/amicus_briefs/mylan-pharmaceuticals-inc.v.celgene-corporation/140617celgeneamicusbrief.pdf>.

alternative sources to gain samples. Celgene argued, among other things, that safety characteristics of the drugs required additional care in their distribution.⁶⁰

The FTC in its amicus brief examined the Supreme Court's decision in *Trinko* and highlighted three factors to guide decisions about whether refusals to deal could constitute actionable exclusionary conduct: (i) was defendant engaged in profit sacrifice without a legitimate business justification, (ii) was the defendant unwilling to sell the product at retail prices and (iii) did the refusal to deal involve "something [the alleged monopolist] was 'already in the business of providing' rather than new services or products that are 'not otherwise marketed or available to the public'". This trio of factors did not include refusal to deal after a prior course of dealing, which the FTC distinguished as evidence of willingness to sacrifice, but not a requirement to establish exclusionary conduct under *Aspen Skiing* and *Trinko*. Importantly, the FTC also acknowledged that the relief being sought in *Celgene* did not raise the "policy concerns with 'enforced sharing' the Court identified in *Trinko* ..." such as reducing the incentive to invest, setting the terms and conditions of a deal by government mandate, and inadvertently promoting collusion.⁶¹ The FTC perspective is in some respects comparable to – albeit a bit more assertive than – the DOJ's view of refusals to deal in the now withdrawn 2008 Section 2 Report,⁶² which may again align with the DOJ position under the Trump administration.

Key takeaway: In the United States, there is no competition law basis for forced sharing where a company (i) owns or controls relevant data, (ii) has never sold that data commercially, (iii) has no independent obligation to give competitors that data and (iv) has a legitimate reason to refuse to deal and prevent free-riding off its investment. Moreover, outside the United States, refusals to deal or grant access are generally addressed under the so-called *essential facility* doctrine, which can provide a competition law remedy for failure to supply or grant access to an intellectual property right, for example, proprietary data. Essential facility is perhaps most well developed in the European Union, where a compulsory license of IPRs has been imposed, but only in "exceptional circumstances".

⁶⁰ibid, 15–16.

⁶¹ibid, 11–15.

⁶²U.S. Dep't of Justice, Competition and Monopoly: Single-firm Conduct Under Section 2 of the Sherman Act, Ch. 7 (8 September 2008) <<https://www.justice.gov/atr/competition-and-monopoly-single-firm-conduct-under-section-2-sherman-act-chapter-7>> (hereinafter "Section 2 Report").

In *IMS Health*, the European Court of Justice held that exceptional circumstances include where a firm holding a dominant position unreasonably refuses to provide access to an IPR that is indispensable to the emergence of a new product for which there is a potential consumer demand, and the refusal excludes any competition on a secondary market.⁶³ In *Microsoft*, the EC was focused on promoting “follow on innovation” enabled by interoperability between different software systems, expanding the “elimination of competition” and “new product” elements of *IMS Health* in that context to find that a refusal to license IP is abusive if (i) it is imposed by a dominant company; (ii) it eliminates effective competition in a secondary market; (iii) it prevents the emergence of a new product or limits technical development; and (iv) it is not objectively justified.⁶⁴

These stringent and narrow requirements demonstrate that even within the EU and jurisdictions that follow EU principles, proprietary data that have never been shared with third parties should rarely, if ever, be expropriated as an “essential facility”. These requirements also protect legitimate property rights in derivative data against free riders who would otherwise exploit one company’s proprietary advantage rather than innovate around that advantage, and it avoids the much more significant risk of over-enforcement against conduct that is both lawful and pro-competitive. Agencies and courts should take pains to avoid what has become for some observers a creeping consensus that access is always good. Facilities-based competition is fundamentally better from most every perspective than a single monolithic network that ultimately would require regulation.

In summary, the notion of “Big Data” has caught the imagination of many, including in the competition law world. While data is “unique” in some sense in that it is not a bridge, it is not unique when one considers how to approach it under antitrust law. Data – in whatever form – may constitute an asset. To the extent it has been fully productized, it is relatively easy to identify its importance. As it moves into being a potential input, there are several preconditions to consider before determining whether an issue may exist. Before taking any action based on a company’s use or acquisition of sets of big data, agencies and other institutions should proceed with great caution to understand in depth what data is relevant in each case, how that data is used, and whether any substantial foreclosure is indeed possible.

⁶³Case C-418/01 *IMS Health v NDC Health* [2004] ECR I-5039, para 38.

⁶⁴Case T-167/08 *Microsoft v Commission* ECLI:EU:T:2012:323.

Acknowledgements

The authors thank Cristina Caffarra, Vice President, Head of European Competition, Charles River Associates for her review and invaluable input.

The views expressed in this article are Greg Sivinski's own, and do not necessarily reflect those of Microsoft. The views expressed in this article are Alex Okuliar's own, and do not necessarily reflect the views of Orrick or of any current or former clients. The views expressed in this article are Lars Kjolbye's own, and do not necessarily reflect the views of Latham & Watkins or of any current or former clients.

Disclosure statement

No potential conflict of interest was reported by the authors.