Are data a source of unassailable competitive advantage in retailing?

Benoît Durand and Iestyn Williams[*]

This version: May 2022

[*] The authors are both partners at RBB Economics. We are grateful to Amazon for funding the underlying research, which we have relied upon to write this paper. We also want to thank Else Christensen and Marta Gligorijevic for excellent research assistance. The views expressed is this paper are our own, however.

1 Introduction

The role of data as a source of competitive advantage has emerged as a key issue in recent debates on competition policy in the digital age.[1] Much of this attention has focused on a select group of companies that are believed to have access to very large amounts of data. Notably, European Commissioner Margrethe Vestager has highlighted the so-called "*gatekeeper*" positions of "*a few huge digital platforms*".[2] The concern is that these platforms benefit from access to large troves of data that are not available to smaller competitors, and that such data asymmetries will deliver unassailable competitive advantages for the leading digital players. The identified threat is that "[o]*nce a digital company gets to a certain size, with the big network of users and the huge collections of data that brings, it can be very hard for anyone else to compete – even if they develop a much better service*".[3] Indeed, this *"could mean that companies which control vital data could be in a position to drive out competition*".[4]

The President of Germany's Bundeskartellamt, Andreas Mundt, has echoed this data-centred concern, observing that "[a]*ccess to data can also create significant competitive advantages for some companies. As competition authorities we have to ensure that markets are kept open and newcomers have a chance*".[5] In a similar vein, the UK Competition and Markets Authority considers that unequal access to data is one of the specific market features, together with network effects and economies of scale, that contributes to a lack of effective competition in digital markets.[6]

A number of scholars have also highlighted the potential competitive advantages derived from data asymmetries. In their 2019 report for the European Commission, for instance, Jacques Crémer, Yves-Alexandre de Montjoye and Heike Schweitzer note that competition in an increasing number of sectors of the economy revolves around machine-learning algorithms. In particular, the authors suggest that insights that these algorithms yield can play a key role improving products and tailoring services to individual users. Critically, access to large amounts of individual-level data may be essential in order to train these machine learning algorithms effectively.[7]

[1] For a useful overview see Jan Krämer, Daniel Schnurr, and Sally Broughton Micova, The role of data for digital market contestability: case studies and data access remedies, Centre on Regulation in Europe, September 2020 (the 'September 2020 CERRE Report').

[2] Speech by EC Executive Vice-President Margrethe Vestager, 29 October 2020.

[3] *Ibid*.

[4] Speech by Commissioner Vestager, 9 March 2018.

[5] CPI Talks… With Andreas Mundt, President of the German Bundeskartellamt (Federal Cartel Office), *CPI Antitrust Chronicle*, April 2018.

[6] See, for example, paragraph 3 of Competition and Markets Authority, A new pro-competition regime for digital markets: Advice of the Digital Markets Taskforce, December 2020.

[7] See Crémer et al., 2019, , *op. cit.*, page 103, for example.

As the importance of algorithms and various data-informed business tools has grown, so has concern that firms with access to more data may hold a substantial competitive advantage vis-à-vis rivals. The 2019 report of the Digital Competition Expert Panel established by the UK Government ("the Furman Report") emphasised that exclusive possession of data can give rise to a competitive advantage in many digital markets.[8] In particular, "[t]*he extent to which data are of central importance to the offer but inaccessible to competitors, in terms of volume, velocity or variety, may confer a form of unmatchable advantage on the incumbent business, making successful rivalry less likely*".[9]

A further worry is that even relatively modest initial disparities in access to data could lead a market to tip in favour of a single supplier. The concern is that positive feedback loops will take hold as a result of so-called data-driven network effects, favouring those holding an initial data advantage. According to this feedback hypothesis, access to more data will allow more appealing/better tailored products to be developed, resulting in increased sales, generating even better access to data which, in turn, will enable the development of even better quality products. In this way, a single supplier might eventually come to dominate a market.

Some have even proposed that data sharing should be mandated to remedy these alleged effects.[10]

Many of the concerns about data and competition have originated in specific settings, notably online search and digital advertising markets. At the same time, the broader policy debate around these concerns has been rather undifferentiated, with quite different businesses being grouped under a broad "digital platform" characterisation. Against this background, in this paper we investigate the extent to which the general concerns expressed about access to and use of data are likely to apply to the retail sector specifically.

There is a long history of small and large retailers using various types of data, such as detailed sales data and volunteered customer information, to improve performance. However, the range and scale of data that can feasibly be collected and processed has grown enormously over time, as technology, computer power and data storage have evolved. In recent years, the increasing availability of user-generated online data has expanded that range still further. A variety of data analytic methods are also regularly deployed to provide insights into customer behaviour and preferences, as well as to assess the impact of business strategies. Data tools

[8] See Furman et al., 2019, *op. cit.*, page 34.

[9] *Ibid.*, pages 32-35.

[10] Krämer et al., 2020, *op. cit.*.

can help retailers make better decisions and deliver superior performance - through enhanced customer propositions and more effective supply chain management, for example. As in other areas, the advent of advanced data-based tools such as machine learning is having a significant impact, notably in enabling better tailored, data-driven promotions and recommendations for customers.

All else being equal, a retailer that has access to more relevant data might be expected to be able to generate better predictions about customer demands. At the same time, diminishing returns are also likely to apply with respect to the additional accuracy delivered by additional data. Hence, even if more data might improve certain aspects of a firm's performance, the incremental advantage is likely to become smaller and smaller as more and more data are amassed. Moreover, simple comparisons of the overall amounts of data available to different retailers may not be that informative. Differences in data needs, and in the uses of these data, must also be taken into account.

A retailer that is active across many retail segments, for instance, may well have access to much more customer data overall than more specialised retailers. However, this broad-based retailer is also likely to need access to more data overall in order to match the positions of more specialised retailers across all the product categories in which it competes. (A specialist retailer that is focused on selling certain categories of products may have much more relevant data for this purpose than the broad-based retailer, even if it has access to much less data overall.) A relevant issue then is the usefulness of information in respect of one set of consumer demands for predictions about a different set of demands. Does a broad-based retailer's access to data on sales of different brands of footwear, say, help it sell more televisions to its customers?

Furthermore, data are useful for a retailer when they feed into data analytics that improve the retailer's performance, notably in terms of the sale of products and/or services to customers. Ultimately, what matters is whether more data translate into improved sales and profits. Importantly, therefore, the value of given data to a retailer will also depend on the quality of its complementary data-analytical capabilities. In turn, that implies that an assessment of the impact of asymmetries in the underlying data themselves must also account for differences in these capabilities.

Significantly, data are rarely an end in themselves in retailing.

Crucially, retail competition is about more than data. It spans a number of dimensions, including price, range and quality of products offered, service levels and store design, branding and promotional effort. While relevant data analytics will help retailers make better decisions with respect to these variables, such analytics are just one element shaping a retailer's business proposition and performance. The other elements will tend to limit the competitive effect of any data-driven advantage. Moreover, this competition is dynamic, with innovative new entrants ready and able to target unfulfilled opportunities. The growth of online retailing has facilitated this process.

Retailing is a differentiated activity not just in terms of the products that retailers sell, but also in terms of the service they offer, the environment in which they distribute products and the commercial strategies they pursue. Different retailers will appeal to different types of customers and for different shopping purposes. Importantly, low online search costs for retail customers as well as the ease with which customers can visit and purchase from multiple retailers encourage multi-homing (i.e., individual customers using a number of different competing retailers). These features facilitate new entry, promoting a dynamic competitive environment, and militate against market tipping in retailing.

Given the different way in which retailers (can) compete, it is questionable whether symmetric access to data would be useful, let alone necessary, to protect effective competition. In that case, a requirement for costly data sharing does not seem sensible.

The rest of the paper is organised as follows:

- In Section 2, we summarise the main concerns that have been put forward with regard to data asymmetries and their impact on competition, and address the key concepts and general mechanisms that appear to underpin these data-driven competition concerns.
- In Section 3, we consider the main types and sources of data typically used by retailers, as well as the principal ways in which these data are used by them.
- In Section 4, we assess the likely effects of data asymmetries in a retailing context. We investigate the benefits that retailers obtain from having access to more data, focusing in particular on the performance of the data analytics underpinning forecasting and recommendations for customers. In doing so, we address the implications of data-driven network effects whilst also highlighting that data and data analytics are only one of several dimensions of retail competition.

- In Section 5, we discuss the case for mandated data sharing. We highlight that since access to competitor data may be not only unnecessary for a retailer to be able to compete effectively, but also of very limited value in practice, the costs of any data access regime may outweigh the competition benefits.
- We offer conclusions in Section 6.

2 Data as a source of competitive advantage

Data are considered an important driver of better decision-making and improved business performance. Examining the impact of data use on company performance, McAfee and Brynjolfsson (2012) observe:

> *[t]he more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results … companies in the top third of their industry in the use of data*-driven *decision making were on average, 5% more productive and 6% more profitable than their competitors.*[11]

Data are *"a core input factor for production processes, logistics, targeted marketing, smart products and services, as well as Artificial Intelligence (AI)*".[12] In particular, data are playing an increasingly significant role in the ways firms, including retailers, design and implement their customer propositions.

In this environment, the role of data as a source of competitive advantage has emerged as a key issue in the debate on competition policy in the digital age.[13]

Two main data-related competition issues have been highlighted, namely the impact of:

- asymmetries in access to data; and
- so-called data-driven network effects, which may transform even small initial data disparities into substantial competitive asymmetries.

Before addressing the role of data and data asymmetries in the retail sector specifically, we first consider more generally the basis of the concern that "more" data will translate into a decisive competitive advantage.

[11] Andrew McAfee and Erik Brynjolfsson, Big data: The management revolution, *Harvard Business Review*, 90(10), October 2012. Note that, arguably, firms that view themselves as data-driven might outperform other firms for other reasons. In this sense, it is not clear that using more data is necessarily the cause of better performance.

[12] Jacques Crémer, Yves-Alexandre de Montjoye, and Heike Schweitzer, Competition Policy for the digital era, *European Commission*, April 2019, page 73.

[13] See, for example, speech given by European Commissioner Margrethe Vestager on 9 March 2018. See also Jason Furman, Diane Coyle, Amelia Fletcher, Derek McAuley, and Philip Marsden, Unlocking digital competition: Report of the Digital Competition Expert Panel, *HM Treasury*, March 2019; Crémer et al., 2019, *op. cit.*; and Krämer et al., 2020, *op. cit.*

2.1 Asymmetric access to data

2.1.1 What constitutes "more" data?

"More" data can mean different things. (One can think of this in terms of a spreadsheet, where more data may translate into more rows or more columns or both.) In one dimension, it may equate to a larger number of observations. This might correspond to information in respect of a greater number of users/customers, for instance.[14] In another dimension, it might correspond to an increase in the amount of information that is held in respect of each observation. Such information might include variables capturing various attributes of the user/customer in question (e.g., age, gender, location, etc.) and measures of behaviour (e.g., purchases, online clicks, etc).[15] Firms that are active in multiple markets may be able to "*combine different data sets to create a wide variety of data about [their] users*".[16]

Significantly, the impact of increasing the number of observations will not, in general, be the same as expanding the amount of information held in respect of each observation, i.e., the number of variables in the dataset. The distinction is important to an appreciation of how more data improves the precision of analytics.

2.1.2 The benefits of "more" data

As noted above, asymmetric access to data has been identified as a feature that can undermine competition. In this section, we discuss the potential advantages that firms that hold "more" data can enjoy relative to those that have less data.[17]

2.1.2.1 The accuracy of prediction

That access to data can benefit businesses seems plainly evident. For example, data on user or customer behaviour can be processed and analysed to become valuable information that allows firms to hone their marketing efforts, as well as to develop new products and services or to improve existing ones.

[14] Note though that an increase in the number of observations might also relate to other aspects of a dataset too. For example, a dataset may contain data on customer purchases of various products over time. In this case, the number of observations can be increased in different ways, either by adding more customers, or more products, or by collecting data over a longer period, or all of these ways together.

[15] Some commentators distinguish between economies of scale and economies of scope on this basis. For example, Martens (2020) considers that economies of scale arise when the number of observations in the spreadsheet increases and gives rise to gains in precision. On the other hand, economies of scope occur when an increase in the number of variables leads to greater accuracy. (See Bertin Martens, Data access, consumer interests and social welfare: An economic perspective on data, *SSRN 3605383*, May 2020.)

[16] See Expert Group for the Observatory on the Online Platform Economy, Work stream on Data: Progress Report, 2020 at page 17. The report considers that platforms that are active in multiple markets or sectors can merge and combine data to gain insights about its users.

[17] In doing so, the relevant focus is on the benefits associated with the data themselves. Naturally, access to more customer data may imply more customers, which is an advantage of itself.

Most often, data analytics are essentially about making predictions. Machine leaning tools have been developed for a multitude of applications in a variety of sectors: for instance, to improve medical diagnostics, in speech recognition, for self-driving cars, etc. Data are used to make predictions in these cases. Similarly, in retailing, the output of the analysis is typically a set of predictions regarding user behaviour and/or customer demands.

To assess the benefit of more data, we focus therefore on the predictive accuracy of machine learning tools, which are utilised for most of the sophisticated data analytics deployed in the retail sector, as a relevant measure of performance.[18] To illustrate: over the last two decades machine learning tools have been developed to predict customer churn in order to identify and target customers that have the highest risk of switching (see Ascarza (2018) and Ascarza and Hardie (2013)).[19,20] In this case, the smaller the prediction error regarding those most likely to switch, the better able the firm is to target measures effectively at relevant 'at-risk' customers.

Thus, the key issue becomes the extent to which prediction errors are reduced when more data are available. If more data allows a material improvement in the accuracy with which firms can predict what customers want, this suggests that firms that have access to more data may have the ability to satisfy those demands more effectively (e.g., by generating better targeted products or services), which may, in turn, give them a competitive edge.

2.1.2.2 Economies of scale and scope: a clarification

Several policy papers on data and digital platforms have suggested that firms holding large amounts of data could potentially benefit from economies of scale and/or scope.[21]

Conventionally, these concepts have been defined in terms of costs. For instance, data collection and management activities might be expected to involve some fixed cost investments. Once those investments are made, the additional – or marginal – cost of recording

[18] Machine learning includes many different methods, including statistical methods such as linear regression and decision trees. Machine learning has evolved and new methodologies are developed all the time. Machine learning is now used in many fields, notably in robotics and autonomous vehicles, speech processing and language translation, neuroscience research, as well as in computer vision, and the number of applications continues to expand. See M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science*, 349(6245), July 2015. Machine learning is also used in marketing. For more, see Liye Ma and Baohong Sun, Machine learning and AI in marketing – Connecting computing power to human insights, *International Journal of Research in Marketing,* 37(3), September 2020. This study provides an overview of the use of supervised and unsupervised machine learning methods in marketing. It is important to note that machine learning can be divided between supervised learning and unsupervised learning tasks. In this paper, we only focus on supervised machine learning, the most common form of machine learning, which typically focusses on making accurate predictions. See Sendhil Mullainathan and Jann Spiess, Machine learning: an applied econometric approach, *Journal of Economic Perspectives*, 31(2), 2017. See also presentations by Stephen Hansen on Machine Learning Methods for Economists.

[19] Eva Ascarza, Retention futility: Targeting high-risk customers might be ineffective, *Journal of Marketing Research,* 55(1), February 2018.

[20] Eva Ascarza and Bruce G.S. Hardie, A joint model of usage and churn in contractual settings, *Marketing Science,* 32(4), 2013.

[21] See, for example, Krämer et al., 2020, *op. cit.,* and Expert Group for the Observatory on the Online Platform Economy, 2020, *op. cit.*

and handling additional observations is likely to be very low where relevant processes are largely automated. As such, these activities are likely to benefit from economies of scale.[22]

Economies of scope also lead to reduced average unit costs. They arise when the production of several different goods involves common fixed costs, for example. In relation to data, economies of scope may occur where a firm collects and manages data in respect of different activities and can share resources across these activities.

However, it appears that a different interpretation of economies of scale and scope is adopted by the commentators that have highlighted competition concerns stemming from data asymmetry. By economies, they refer to the benefits associated with more data more broadly. Thus, the September 2020 CERRE report notes that "*economies of scale (e.g., data about more users or 'broad data') and economies of scope (e.g., more heterogeneous data about users or 'deep data') allow firms to make better recommendations and to offer better fitting products, services and content, among other things"*. Similarly, the Expert Group for the EU Observatory on the Online Platform Economy considers that economies of scope arise when "*a platform with a large ecosystem operating in multiple markets can merge and combine different datasets to create a wide variety of data about its users*".

In line with these interpretations, we focus in this paper only on the benefits that more data can provide – notably, increased prediction accuracy – without addressing the cost advantage implications of economies of scale and scope as such.

2.1.3 "More" data and the accuracy of predictions

Statistical theory indicates that, in general, as the number of relevant observations increases (e.g., because the number of users from which data are collected increases), the accuracy of predictions based on statistical analysis will improve. However, importantly, this gain in precision is subject to diminishing returns.[23] In other words, whilst adding more observations improves precision, the incremental gains diminish as more and more data become available. In fact, once the sample reaches a certain size, there will be no appreciable gain in precision from adding further observations.[24]

[22] Economies of scale arise where average costs fall with increases in production. This occurs, in particular, when high fixed costs are present. Fixed costs are costs that are incurred irrespective of the amount of the good produced. For example, the cost of setting up an automotive plant is fixed, because it does not vary with the number of cars actually produced. As the production of cars increases, however, that fixed cost is spread over a greater number of units, such that the average cost of manufacturing each car falls.

[23] It has been noted that in statistics and econometrics, the marginal return to additional data is decreasing. See, for example, Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Leaning: Data Mining, Inference, and Prediction*, 2nd edition, Springer: New York, 2009.

[24] Specifically, as more observations are used, the variance of the prediction error will be reduced, indicating that individual predictions are grouped more closely around the average prediction. Note that the average prediction may still be biased, however.

As with statistical and econometric techniques, in general the performance of machine learning methods improves as more observations are available. That is, the prediction errors are reduced on average. Typically, however, the incremental gains in accuracy of machine learning tools become smaller as the number of observations is increased – again, as with statistical and econometric methods.[25]

The findings of the September 2020 CERRE report, which focuses on data and digital markets, are consistent with this general principle. The report provides mostly an account of studies that have looked at the benefit of greater online search query volumes for the quality of search results. It finds that "*broader data collection leads to quality improvements, albeit at a decreasing marginal rate*".[26]

As noted, more data can also mean having access to a broader array of information about users/customers and their actions. The question then is whether having more such information improves the accuracy with which user behaviour can be predicted. The answer is not clear-cut. More data will only be valuable in respect of particular analysis if they are relevant to, and therefore inform, that analysis. Otherwise, they will simply contribute noise.[27]

Ultimately, the nature of the additional data will determine the extent to which the performance of the algorithm at issue is improved. In other words, the case for more data is specific to individual situations. It will be stronger in some settings, weaker in others, and sometimes using more data will even be positively undesirable.

As the discussion above highlights, more data means different things, and depending on the situation, more data can have different incremental value for the analysis in question. The

[25] For example, an early study by Cui and Curry uses simulated data to assess how the support machine vector (SVM) performs in predicting customer choice compared to the standard logit model. Whilst this study shows that SVM is a better predictor, the hit-rate of both methods improves with sample size, albeit the marginal return decreases. (See Dapeng Cui and David Curry, Prediction in marketing using the support vector machine, *Marketing Science*, 24(4), November 2005.) Another study, Perlich et al. (2003), corroborates this finding. It compares the performance of two classical methods, tree induction and logistic regression, on classification accuracy over 36 different datasets. This analysis confirms that the prediction accuracy improves with the size of the training data, albeit with diminishing returns. (See Claudia Perlich, Foster Provost, and Jeffrey Simonoff, Tree induction vs logistic regression: A learning curve analysis, *Journal of Machine Learning Research*, 4, June 2003.)

[26] Krämer et al., 2020, *op. cit.*, section 3. The studies reviewed in the CERRE report assess the extent to which the quality of online search results is improved using more data. The studies are not comparable as they use different measures of quality for the search results (e.g., by click-through rate or the propensity of users to return to the search engine). Importantly, their findings are not always consistent: one study finds little benefit while two studies find significant benefits.

[27] For example, the performance of the linear regression model, which is one of many machine learning tools that can be used to predict user behaviour or user attributes, can be improved by adding relevant explanatory variables. The linear regression approach draws on the correlation between an array of variables (e.g., gender, income level, nationality, profession, education level, family status, etc.) associated with individual customers and the incidence of particular actions, such as purchasing a particular product or clicking on a link. The accuracy with which linear regression predicts the actions in question will depend on the strength of this correlation. In general, adding more relevant variables can increase the correlation and therefore improve the accuracy of the predictions. There is, however, some limit to this. For instance, if the additional information is redundant, then using these extra data will not increase the accuracy of the prediction. Worse, when the data are irrelevant, the performance of the machine learning algorithm can deteriorate. This is illustrated with an example based on a support vector machine (SVM) model that is used to predict hotel cancellations in Portugal. If the training dataset includes observations from domestic and international customers, the SVM model performs poorly in years when most customers are domestic. That is, in these years, the model's performance deteriorates when using additional data on international customers. (See the blog entry by Michael Grogan, *Why more data is not always better*, 8 August 2020, retrieved from https://towardsdatascience.com/why-more-data-is-not-always-better-de96723d1499.)

statement that firms that have access to more data will obtain a decisive competitive advantage as a result does not apply generally.

2.1.4 The choice of methods and the complexity of the machine learning model also determine prediction accuracy

More data is not the only factor that can impact the accuracy of predictions produced by machine learning tools. The choice of the method as well as the complexity of the model are also recognised as important aspects that influence the accuracy of prediction. Importantly, more data will affect differently the performance of the machine learning tool depending on the method used or the model complexity.

2.1.4.1 The choice of methods

Machine learning encompasses many different methods and, depending on the setting, some will perform better than others. Moreover, and importantly, several studies show that the point at which diminishing returns start to become material in respect of data can vary depending on the method employed.

For instance, an early study, Perlich et al. (2003), compares two standard, off-the-shelf machine learning methods, showing that one method – logistic regression – performs better with smaller datasets, while the other approach – tree induction – outperforms the other with larger datasets.[28] A more recent study, Crone and Finlay (2012), compares the performance of four different machine learning methods – logistic regression, linear discrimination analysis, classification and regression tree, and artificial neural networks – in a credit scoring application.[29] In all cases, not only are there diminishing returns to increasing sample size, but the marginal return from adding more observations is zero after a certain point. The study also shows that, for this particular problem, logistic regression is the best method of those studied. Significantly, the benefits of adding observations were exhausted at relatively small sample size for this technique.

2.1.4.2 Model complexity

Another important factor in the design and performance of machine learning tools that is discussed in the literature is so-called model complexity. Model complexity can mean different things depending on the method used. For a linear regression model, this corresponds to the number of parameters, which is related to the number of so-called input variables that

[28] Perlich et al., 2003, *op. cit*. The study focuses on these two standard methods for binary classification.

[29] Sven Crone and Steven Finlay, Instance sampling in credit scoring: An empirical study of sample size and balancing, *International Journal of Forecasting*, 28(1), 2012.

are included to predict the variable of interest (e.g. input variables such as gender, age, and past clicks can be used to predict user clicks).[30] For more sophisticated methods, such as artificial neural networks, which are suited to capturing non-linear relationships between the input and output variables, the degree of complexity depends on the number of so-called hidden layers or – more commonly in the literature – the "width" of the network used to fit the training data.[31]

The research literature on machine learning supports the view that increased complexity may not always improve the accuracy of predictions. Beyond a certain point, further increasing model complexity would be counter-productive for most machine learning methods.[32] Hence, the relationship between model complexity (i.e., adding more input variables in the case of a linear regression) and prediction error is then U-shaped, as represented in Figure 1 below.[33] Since accuracy is greatest when the prediction error is smallest, this means that accuracy is maximised at an intermediate level of model complexity.[34]

[30] The number of parameters in the linear regression model depends on the number of independent variables that are included. Note, however, that to fit a nonlinear relationship, the regression may use a polynomial, whereby the model includes an nth degree polynomial in the set of independent variables.

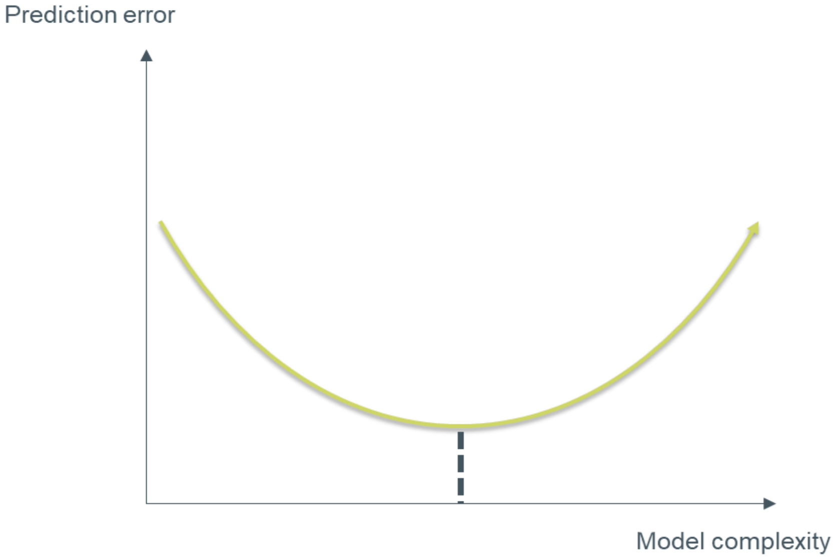
[31] In very simple terms, a neural network will consider the multiple interconnections between the input and the output variables. It is a called neural network because the process that is used mimics the way the human brain operates. In a network with only one hidden layer, the layer will comprise a series of nodes (or neurons), each of which consists of a combination of the various input variables, where the weight given to each variable varies. Increasing the number of nodes essentially increases the number of combinations that the network considers, making the model therefore more complex. To summarise: each node is called a neuron, and the number of neurons is what determines the width of the network, and also its degree of complexity. For more see Michael A. Nielsen, *Neural Networks and Deep Learning*, Determination Press, 2015, retrieved from http://neuralnetworksanddeeplearning.com/index.html.

[32] This relationship between model complexity and prediction error is related to the bias-variance trade-off. For a detailed and formal presentation of the bias-variance decomposition, see Chapter 7 of Hastie et al., 2009. For a formal introduction of the bias-variance trade-off for neural networks see Stuart Geman, Elie Bienenstock, and René Doursat, Neural networks and the bias/variance dilemma, *Neural Computation*, 4(1), January 1992. As the authors put it, "*the price to pay for achieving low bias is high variance*". The representation of the U-shaped curve is due to Scott Fortmann-Roe. (See Scott Fortmann-Roe, Understanding the bias-variance trade-off, June 2012, retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html.

[33] The reason for this relationship can be explained as follows. Machine learning algorithms learn from the pattern of the training data to form predictions. Once it is developed, the algorithm is then applied more widely to other data. A simple model will not fit the training data very well, and thus its predictions will not be accurate, resulting in large variance in the prediction error. A more complex model will perform better on the training data. However, there is a point at which the model becomes so complex that it fits extremely well the training data (it is said to overfit) but its general performance will worsen. This is because if the more complex model fits the training data perfectly, it accounts for all the idiosyncrasies of the training data, which are not present elsewhere. This means that the model will tend to perform poorly with other datasets. When the model becomes too complex, it is likely to make significant prediction errors when applied more widely.

[34] This inverted U-shape is also known as the bias-variance trade-off. Prediction error is made up of two components, the bias of the estimator and the variance of the estimator. The bias measures how far the expected value of the estimator is from the true value of the variable of interest. Econometrics always seeks to use an unbiased estimator. This is why the ordinary least square estimator (OLS) is so popular in econometrics (see Gauss Markov theorem). This is not the case in machine learning, however. This is because it is worth having a biased prediction if ultimately the prediction error is smaller. This can happen if the variance of the estimator is low, which means that predictions will be bunched together around the expected value of the estimator. This means that some bias can be tolerated if this leads to significantly less variance. Typically, the more complex the model is, the lower the bias, but the higher the variance. (Prediction error also contains a third component, which is called the irreducible error as it cannot be avoided.) See Chapter 7 of Hastie et al., 2009, *op. cit.*, for a formal discussion of the bias-variance trade-off.

Figure 1: The relationship between prediction error and model complexity (a representation)



As noted, more data can refer to information on additional input variables. Using more data may involve adding more input variables, i.e., expanding model complexity. For example, adding more input variables in a linear regression model increases the number of parameters to be estimated and in a neural network this might imply increasing the width of the network. In these cases, the conventional view is there is a point beyond which too much complexity diminishes the performance of machine learning tools, as prediction errors increase.[35]

2.2 Data-driven network effects

In addition to strong economies of scale and scope, network effects are often cited as a key feature of many digital markets that act as barriers to entry, contributing to making these markets highly concentrated. (See, for example, Crémer et al (2019) and the Furman Report.) Data-driven network effects is a related, but more recent, concept that is used to describe a

[35] The field of machine learning is evolving rapidly, notably deep learning and neural networks. Therefore, to be complete, we noted that some very recent studies have started to challenge this classic trade-off for neural networks. These new findings suggest that increased complexity can improve substantially the accuracy of predictions without any clear limit. Studying the performance of neural networks on actual datasets, Neal et al. (2018) finds that the prediction error is reduced as the width (or the number of neurons) of the network increases. The authors show that the two components of the mean squared prediction error, the bias and the variance, are reduced with the width of the neural networks (i.e., by increasing the number of hidden layers). (See Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas, A modern take on the bias-variance tradeoff in neural networks, *arXiv preprint arXiv:1810.08591*, 2018.) This finding has also been corroborated by a recent study by Belkin et al. (2019), which finds similar results for other variants of neural networks but also for decision tree and random forest methods. (See Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, *Proceedings of the National Academy of Sciences*, 116(32), August 2019.)

virtuous circle benefiting digital platforms that have access to more data, which may ultimately cause users to gravitate towards a single platform.

2.2.1 A new variant of network effects related to data?

The presence of "*strong network effects*" is a key feature of many digital markets, which can contribute to insulating large incumbents from the competitive pressure of challengers.[36] Network effects (or network externalities) arise when the value of a network to one user increases when another user joins and, therefore, enlarges the network. Social networks, such as Facebook or LinkedIn, are modern examples of direct network externalities at work. The value of joining these networks increases as their membership expands.

Network effects describe a virtuous circle whereby, as a platform's user base grows, it becomes even more attractive to new users, causing even more of them to join. This effect may snowball to the point that the market tips and only one platform remains. This winner-takes-all situation will result in an entrenched incumbent position because smaller-sized challengers will not be able to draw members from the incumbent's platform, as their own platforms have too few users to be attractive.

The competitive advantage that asymmetric access to data can confer may also generate a virtuous circle, through so-called data-driven network effects. These *"can give rise to self-reinforcing feedback effects*"[37] whereby initial data asymmetries *"can become sustained competitive advantages*"[38] for incumbents. Turk (2016) describes these positive feedback loops as follows:

> *"Data network effects occur when your product, generally powered by machine learning, becomes smarter as it gets more data from your users. In other words: the more users use your product, the more data they contribute; the more data they contribute, the smarter your product becomes (which can mean anything from core performance improvements to predictions, recommendations, personalization, etc.); the smarter your product is, the better it serves your users and the more likely they are to come back often and contribute more data – and so on and so forth. Over*

[36] See Furman et al., 2019, *op. cit.*, page 32.

[37] Krämer et al., 2020*., op. cit.*, page 64.

[38] *Ibid.*

> *time, your business becomes deeply and increasingly entrenched, as nobody can serve users as well".*[39]

In other words, even relatively modest initial differences in access to data may lead to more appealing/better tailored products and more sales, generating even better access to data.[40] As indicated by Crémer et al. (2019), the accumulation of data will improve the performance of machine learning algorithms, for instance, allowing the incumbent to enhance their customer offering. This, in turn, is liable to attract more customers, generating even more data, increasing the extent of the data asymmetry and the resulting competitive advantage.

Another prediction of this theory is that new entrants – in contrast – will face a classic "chicken and egg" problem: without data they will struggle to generate the data-driven insights needed to attract customers, but without customers they will not have the customer data to generate the insights required to compete effectively. The concern is that these positive feedback effects could transform even small initial competitive advantages into unassailable market positions.

The threat that a market may tip as a result of data-driven network effects has been stressed by several commentators and competition officials. For instance, the Furman Report highlights that data-rich incumbents can cement their positions thanks to the virtuous circle described above.[41] In the same vein, the September 2020 CERRE report observes that the collection of large amounts of data can lead markets to tip in favour of dominant online platforms.[42,43] Moreover, in launching the European Commission's June 2020 consultation on a new competition tool, Margrethe Vestager noted that "*there are certain structural risks for competition, such as tipping markets, which are not addressed by the current rules"*.[44]

2.2.2 The limits of data-driven network effects

The feedback loop arising from data-driven network effects hinges critically on the extent to which the performance of the machine learning tool in question is improved by using more data. This is different from traditional network effects, which rely primarily on the attractiveness of a platform to a user increasing with the number of users it has.

[39] See the blog entry by Matt Turck, *The power of data network effects*, 4 January 2016, retrieved from: https://mattturck.com/the-power-of-data-network-effects/.

[40] Crémer et al. also postulate other positive feedback mechanisms whereby control over some of an individual's data may enable a platform to collect more of it. (See Crémer et al., 2019, *op. cit.*, page 31.)

[41] Furman et al., 2019, *op. cit.*, page 33.

[42] Krämer et al., 2020, *op. cit.*

[43] In a speech on 29 October 2020, for example, Margrethe Vestager noted that "[o]*nce a digital company gets to a certain size, with the big network of users and the huge collections of data that brings, it can be very hard for anyone else to compete – even if they develop a much better service. So we face a constant risk that big companies will succeed in pushing markets to a tipping point, sending them on a rapid, unstoppable slide towards monopoly – and creating yet another powerful gatekeeper*".

[44] See European Commission press release of 2 June 2020: Antitrust: Commission consults stakeholders on a possible new competition tool, retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_20_977.

As indicated above, even where more observations can be beneficial, there are good reasons to think that the gain in performance will exhibit diminishing returns. The virtuous circle that is fuelled by more data can quickly attenuate. In other words, the incremental feedback effect may become smaller and smaller as more data are used, possibly quite rapidly. Indeed, in some settings there might be a point beyond which more data do not add any material gains at all. Crucially, in these circumstances, any counteracting forces may be expected to offset the concentrating effects of more data.

Furthermore, even if more data improves the performance of the machine learning tool at issue, this needs to translate into increased value of the offering to the customer. In some cases, the value of learning from more data is unlikely to deliver any competitive advantage. To illustrate: Hagiu and Wright (2020) explains that smart thermostats need only a few days' worth of data to learn users' temperature preferences over the course of the day. For this particular product, in this setting, the use of machine learning tools is unlikely to yield a competitive advantage.[45] More generally, it suggests that a case-by-case assessment is required to determine the extent to which more data deliver a significant competitive advantage, which could result in substantial feedback loops.

The countervailing forces that limit the concentration tendencies of traditional network effects would also apply to so-called data-driven network effects. Indeed, factors such as multi-homing and product differentiation are features that will serve to reduce the impact of network effects. This implies that markets may not tip, even where data asymmetries arise.

[45] Andrei Hagiu and Julian Wright, When data creates competitive advantage, *Harvard Business Review*, 98(1), January-February 2020.

3 The use of data in retailing is not a new phenomenon

As we describe below, the use of data is not a new phenomenon in the retail industry. All retailers have access to a range of data obtained as a result of their own retail activities, be they physical or online. These may include scanner data, data from store membership cards, and online data from customer purchases and browsing histories, as well as information from customer surveys. In addition, data can also be acquired from third party sources. For example, demographic information to augment consumer profiles can be purchased from data brokers. While some types of data, such as detailed product sales data, have been utilised by retailers for decades, as in other industries it has only recently become feasible to collect and process other types of data, notably user-generated data.

Retailers have employed a range of data analytics that provide insights about customer behaviour and preferences as well as assist in assessing the impact of business strategies. These data tools help retailers make better business decisions and enhance performance, e.g., by improving customer propositions and supply chain management.[46] Using data to gain insights and improve business performance is not only not new but it is also widespread in the retail industry.

Importantly, despite the ubiquitous and historical use of data in retailing, and substantial differences in the extent of the data available to different retailers, this has not generally resulted in monopolised retail markets. This fact suggests that, while data asymmetries may be a source of some competitive advantage, such advantage is unlikely to translate into a position of unassailable competitive strength.

3.1 Sources of relevant data

The EU's Margrethe Vestager has observed that "*controlling a lot of data isn't such a big issue, if others can easily get hold of the same information, from their own customers or simply by buying it in the market*".[47] Before considering the potential impact of any data asymmetries and the potential need for intervention to address this, it is necessary, therefore, to establish the extent of the data available to competing retailers in practice and, also, the nature of those asymmetries. We will consider the data that are likely to be available to retailers directly,

[46] Daniel Gutierrez, Reinventing the retail industry through machine and deep learning, *insideHPC Special Report*, 2018.

[47] Speech by European Commissioner Margrethe Vestager, 9 March 2018, retrieved from https://wayback.archive-it.org/12090/20191129214609/https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/fair-markets-digital-world_en.

noting that these data are also likely to be the most useful to retailers in any event, as well as the possibilities that retailers have to obtain relevant data from other sources.

3.1.1 Direct retailer access to data

The potential for data asymmetry starts with retailers' access to so-called first-party data, i.e., data that a firm obtains directly from its customers/audience, as large retailers will typically hold more of these types of data than small retailers. Most obviously, first-party data are likely to include the retailer's own sales data. Firms may also be provided with a range of personal data volunteered by customers (e.g., name, address, contact details, as well as demographic information) when they set up online or loyalty card accounts, for example. Where retailers are able to associate specific purchases with a unique customer identifier, they can construct a profile of the individual customer's purchase history. In some cases, they will be able to identify the user concerned, as well as connecting the purchase history with other user information.

In an online context, retailers are also able to track customer search and browsing activity on their own websites, even where this does not result in actual purchases. Moreover, customer logins, cookies, and other mechanisms may allow them to associate these patterns with individual users, and even to identify those users. At the same time, increasingly sophisticated methods are also being developed to gather comparable information offline, e.g., by monitoring customers' physical browsing activities.[48]

Simply on account of their greater size – they have more customers, record more sales, have potentially greater product ranges, etc. – large retailers are likely to be able to collect and process more data. Importantly, however, irrespective of size, a retailer will have direct access to data in respect of its existing customers and sales, which are also likely to be the most pertinent to its ongoing activities, e.g., sales forecasts, product recommendations, design and evaluation of marketing programmes.[49] In other words, the data that are most pertinent to a retailer's needs are likely to relate to its own activities, which implies that an established retailer will already have direct access to the data that are most relevant for its operations.

3.1.2 Data volunteered by customers

Retailers could use data that customers volunteer themselves. In principle, customers should be able to provide the data that they control to retailers and might be expected to have an

[48] See Section 3.2.3 below.

[49] For the same reasons, data relating to another retailer's operations may be less relevant in this regard.

incentive to do so, where this would enable those customers to receive better products, pricing and service, including by enhancing the competitive rivalry between retailers.[50] Such data could include a range of personal information, as well as purchase and online browsing histories, for instance. (At least some information of this sort may be available to multiple retailers in any event where a customer multi-homes.) A key issue then is the ease with which such data can be extracted and shared by customers, which highlights the relevance of data portability. That also raises its own set of issues.[51]

3.1.3 Third-party data sources

For some applications, a retailer may also draw on a range of third-party sources to complement its own data. These data may include:

- (Aggregated) scanner data for a large number of retail outlets from third-party providers such as Nielsen Market Research and Information Resources (IRI), allowing general sales patterns to be identified at a disaggregated product level, as well as enabling performance relative to competitors to be evaluated.
- Household panel datasets (from providers such as Nielsen) that offer data on purchases made by a sample of households over a given period of time, as well as demographic information on the participating households, allowing retailers to evaluate product performance in specific consumer segments, including consumer responses to marketing campaigns and store layout changes, for instance.
- Datasets (from brokers such as Acxiom) that combine publicly available data and data from surveys as well as from other providers to offer profiles of consumers.
- Detailed information on individual customer behaviour obtained through tracking (e.g., via cookies) of online and smartphone activity.

In addition, retailers may also be able to purchase information in respect of search and browsing histories for third-party websites, including information obtained via third-party cookies.

Hence, even where there are significant differences in the extent of retailers' own first-party data and even where these differences matter (which may not be the case, as indicated above), individual retailers may already be able to address relevant asymmetries through commercial

[50] Under GDPR, customers may be able to request personal data to port to other retailers.

[51] Portability is a central component of the obligations on so-called "gatekeepers" proposed by the European Union in its draft Data Markets Act.

arrangements to obtain data from third-party data providers, some of which (e.g., Nielsen, IRI) offer data obtained from other retailers.

The fact that retailers already contribute data to third-party aggregators such as Nielsen or IRI suggests that the perceived benefits of doing so, including receiving valuable data and insight in exchange, exceed the costs. If user-generated online data become important to retailers, third-party operators may develop similar services for this type of data. Retailers may also have incentives to contribute their own user data to such initiatives, in exchange for reciprocal access to valuable data (presumably anonymised) and insights based on the contributions of other retailers.

3.2 Uses of data in the retail industry

As shown above, retailers have access to a range of data. These data are utilised by retailers in a variety of applications to improve their customer propositions as well as their performance. The more important of these applications include marketing activities, the selection of product assortment, the design of store layouts, pricing, customer recommendations, and sales forecasts. As noted, these uses are not new, though the ways data are employed, and the types of data used have evolved.

Importantly, whilst retailers have, for some time, relied on a variety of data analytics tools based on different types of data to improve performance, the use of these tools does not appear to have delivered entrenched, monopolised retail positions.

3.2.1 Marketing activities

Marketing spans a range of activities that retailers engage in to promote the products they sell. These activities form a critical aspect of competition between retailers.[52] Equipped with data to inform relevant analytics, retailers can make more informed decisions about the set of marketing activities (i.e., the so-called marketing mix) used to increase customer satisfaction, improve performance, and boost sales.

[52] Among the marketing activities that retailers consider are, for instance, the so-called “4 Ps” - product, price, place, and promotion. Further “Ps” have been added to the list, or marketing mix, that a retailer considers – for instance, personnel, process, and physical assets. Each of the Ps involves various marketing actions with various possible strategies. For more, see Ronald E. Goldsmith, The personalised marketplace: beyond the 4Ps, *Marketing Intelligence & Planning*, 17(4), 1999, and E. Jerome McCarthy, *Basic Marketing: A Managerial Approach,* R. D. Irwin: Homewood, IL:, 1960.

Models to measure the performance of a firm's marketing mix by now date back several decades. For example, already in the 1970s, Nielsen's SCAN*PRO model utilised store scanner data to analyse the effects of promotions.[53]

Other quantitative methods and types of data can be used to measure, analyse, and predict how different marketing actions (or combinations of actions) perform and, over the years, marketing analytics has evolved as more and richer datasets have become available, but also as quantitative methods have improved, and computer processing power has increased.[54]

One consequence of having more data at hand – in particular, customer level data – and of the improvements in analytics has been a greater focus on the management of customer relationships, and on the personalisation of the marketing mix, to improve sales performance and gain new customers.[55] In the past twenty years, both offline and online retailers have developed extensive customer relationship management strategies that utilise individual-level data to make more personalised promotions and recommendations.[56] Using data to identify and target specific market segments is, however, not a new concept. It has been common practice since at least the 1980s, only the extent has changed.[57]

3.2.2 Product assortment

There are at least two good reasons why retailers must select their product assortments carefully. First, retailers typically face capacity constraints, notably limited physical shelf-space in the case of bricks-and-mortar stores, and cannot offer an unlimited product range.[58] Second, whilst consumers have been shown to value product breadth and depth, having too many options within a category can "confuse" consumer choice. Hence, sales can actually be increased by eliminating low-performing SKUs within a category.[59]

Retailers will therefore typically want to curate their product assortments. Data analysis can help retailers make informed decisions about their optimal product assortments, in particular

[53] Harald J. Van Heerde, Peter S.H. Leeflang, and Dick R. Wittink, How promotions work: SCAN*PRO-based evolutionary model building, *Schmalenbach Business Review,* 54(3), July 2002.

[54] Michel Wedel and P.K. Kannan, Marketing analytics for data-rich environments, *Journal of Marketing: AMA/MSI Special Issue,* 80(6), November 2016.

[55] *Ibid.*

[56] Peter C. Verhoef, Rajkumar Venkatesan, Leigh McAlister, Edward C. Malthouse, Manfred Krafft, and Shankar Ganesan, CRM in data-rich multichannel retailing environments: a review and future research directions, *Journal of Interactive Marketing,* 24(2), May 2010.

[57] Caroline A. Tynan and Jennifer Drayton, Market segmentation, *Journal of Marketing Management,* 2(3), 1987.

[58] Online retailers may also face some capacity constraints at their warehouses for example.

[59] Alexander Chernev, Product assortment and consumer choice: An interdisciplinary review, *Foundations and Trends in Marketing,* 6(1), 2011.

regarding the breadth (the range of product categories they offer) and depth (the number of SKUs they maintain within a product category).[60]

Retailers also use insights about demand trends and customer take-up of products offered by rivals to identify new products to add to their own assortments, including through new own-product development. A survey conducted by the Observatory on the Online Platform Economy found that e-commerce sellers place considerable value on data concerning their competitors' performance.[61] Both online marketplaces and data brokers provide analytics and actionable market insights based on consumer and rival behaviour, such as key word searches and search volumes, listing and pricing data, page views and click-through rates, and details on top-performing products and sellers.[62] Retailers use these insights to better understand shifts in demand, as well as seasonal variations, to improve the positioning of their products, and to identify new products and product niches with high revenue potential to add to their assortments.[63]

3.2.3 Placement and layout

Effective store layout ultimately boosts sales by improving the customer experience and enhancing customer loyalty. Using different types of data, retailers have sought to analyse consumer behaviour in order to optimise store layout. For instance, consumers report a greater level of satisfaction when choosing from a product assortment grouped according to benefit as opposed to other attributes, e.g., choosing from an assortment of teas grouped into wellness, weight loss, and energy boost teas, as opposed to an assortment grouped into black, green, and herbal teas.[64,65]

[60] For example, the Dutch grocery retailer Albert Heijn reviews sales data to allocate shelf space to each category based on the highest marginal profit, and to identify SKUs with the highest substitution rates which they eliminate if the depth of category has to be reduced. (See A. Gürhan Kök, Marshall L. Fisher, and Ramnath Vaidyanathan, Assortment planning: Review of literature and industry practice, in N. Agrawal, S. Smith (eds), *Retail Supply Chain Management, International Series in Operations Research & Management Science*, 122, 2008.) The Dutch online retailer CoolBlue has said that it uses data to refine its assortment and "weed out" inferior options within each product category. (See Robert P. Rooderkerk and A. Gürhan Kök, Omnichannel Assortment Planning, in S. Gallino, A. Moreno (eds), *Operations in an Omnichannel World*, *Springer Series in Supply Chain Management*, 8, 2019.)

[61] See Vaida Gineikytė, Egidijus Barcevičius, and Loreta Matulevič, Platform data access and secondary data sources: Analytical paper 1, *Observatory on the Online Platform Economy*, 2020.

[62] Amazon provides sellers with data on search terms (including the term frequency rank, the top three most-clicked products for each search terms and the products' click shares and conversion rates), details of the three rival products most frequently purchased together with each of the seller's products, details of which rival products were most frequently viewed by customers in the same day as each of the seller's products, as well as demographic characteristics of the seller's customers. eBay provides sellers with data on competitor pricing and competitor performance in comparison to the seller's pricing/performance, as well as listing data (including ranks, click-through rates, page views, conversion rates, transactions, details of where and how often sellers appear in search results). Data companies specialising in e-commerce provide data on market trends (sales and inventory, pricing, keyword trends), competitor business performance (pricing, product assortment, product ranks, sales, keyword rankings, pricing, listing details) and customer behaviour (browsing histories and search volumes, page visits and time spent viewing each product).

[63] Gineikytė at al., 2020, *op. cit*.

[64] Panagiotis Sarantopoulos, Aristeidis Theotokis, Katerina Pramatari, and Anne L. Roggeveen, The impact of a complement-based assortment organization on purchases, *Journal of Marketing Research*, 56(3), March 2019.

[65] Cait Poyner Lamberton and Kristin Diehl, Retail choice architecture: The effects of benefit- and attribute-based assortment organization on consumer perceptions and choice, *Journal of Consumer Research*, 40(3), October 2013.

Novel technology, such as video feed analysis, has made it possible for bricks-and-mortar retailers to observe customer search and choice processes more closely and to use the insights that this offers to optimise their assortment structures.[66] Similarly, online retailers can utilise clickstream data to track customer movements through their virtual stores and, based on the analysis of these data, can adapt the layout of the store to each customer's preferences while the customer is still browsing.

Further, the appearance and design of store websites can be changed to match a particular customer's cognitive style; whether that is deliberative or impulsive, analytic or holistic, visual or verbal.[67] For example, the detail with which product characteristics are displayed and whether the information is displayed primarily as numbers or primarily as text can be varied from customer to customer.[68]

3.2.4 Pricing

The combined use of sophisticated software and data have allowed managers to develop pricing strategies that are more closely based on customer preferences.[69]

Since the early 2000s, retailers have used price-setting software to ease the informational burden of optimising prices for tens of thousands of SKUs through category pricing. Early category pricing software solutions, such as those introduced by DemandTec and KhiMetrics (now part of SAP), grouped products into families using data on customer demand elasticities and item affinities, so that the optimisation problem for a given product would only take into account products with related demands, as opposed to the entire product assortment. Incorporating other inputs, namely data on past sales and prices, competitor prices, seasonality effects, planned promotional events and objectives, the pricing software would determine optimal prices at the SKU level.[70]

Retailers also rely on commercial software programs to formulate optimal dynamic pricing rules for markdowns and clearances, including which products to clear, the magnitudes of the discounts and the timing of those discounts. Markdown software programmes require data on

[66] Sam K. Hui, Yanliu Huang, Jacob Suher, and J. Jeffrey Inman, Deconstructing the "first moment of truth": Understanding unplanned consideration and purchase conversion using in-store video tracking, *Journal of Marketing Research*, 50(4), October 2018.

[67] Impulsive consumers make decisions quickly, whereas deliberative consumers prefer to explore all their options in depth before they make a decision. Analytic consumers will prefer to be presented with as many details as possible, while holistic consumers will care more about the bottom line. Visual consumers will prefer information presented graphically (e.g., in charts or graphs), while verbal consumers will have a preference for text, sound and numbers.

[68] John R. Hauser, Glen L. Urban, Guilherme Liberali, and Michael Braun, Website morphing, *Marketing Science*, 28(2), February 2009.

[69] Dhruv Grewal, Kusum L. Ailawadi, Dinesh Gauri, Kevin Hall, Praveen Kopalle, and Jane R. Robertson, Innovations in retail pricing and promotions, *Journal of Retailing*, 87, July 2011.

[70] Symphony-IRI and ACNielsen provide promotional software based on demand elasticities to set optimum prices. See Grewal et al., 2011, *op. cit*.

inventory and various costs associated with markdowns, specified "business rules" (e.g., the allowed number, frequency, and magnitude of markdowns), as well as a demand-price model.[71]

Dynamic pricing algorithms are used to implement frequent price changes with the aim of increasing revenue. To this end, these algorithms are fed with sales data and information on competitors' prices, as well as actual inventory levels. In the online world, prices can be changed quickly and relatively costlessly. These advantages allow online retailers to carry out pricing experiments and, on the basis of customer responses to these experimental price changes, to adjust prices swiftly. This makes dynamic pricing especially popular online.[72]

3.2.5 Recommendations and promotions

Access to more sophisticated analytics and the availability of fine-grained data has allowed retailers to develop more tailored product recommendations and promotions, and better targeted marketing campaigns.

Such targeting is not a new phenomenon. Retailers have long operated loyalty cards and programmes, which have enabled them to formulate and deliver tailored promotions. Retailer CVS, for example, has used data collected through its Extra Care loyalty programme to develop promotional offers aimed at different customer segments.[73] Retailers can now also use various other means – such as data generated through mobile phone apps – to identify individuals' preferences, movement patterns and co-located social connections, and thereby personalise promotions and other marketing communications.[74] The global e-marketplace Groupon, for example, sends out personalised promotions and coupons to customers on a daily basis and refines these deals as it collects more data on users' profiles.[75]

Since the 1990s, online retailers have relied increasingly on recommender systems to make personalised, data-based promotions and recommendations to customers, and these have become widespread in the 21st century.[76] Advanced data tools can be used to generate lists of

[71] Wedad Elmaghraby and Pınar Keskinocak, Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions, *Management Science*, 49(10), October 2003.

[72] An original example of dynamic pricing strategy is found with an online fashion retailer RueLaLa. Its business model consists of organising flash sales, or so-called "sales events". RueLaLa plans sales events for an item, and repeats the events until the item is sold out. Each event is limited in time, and the price level remains fixed during an individual event. However, from event to event, RueLaLa can adjust the price of the item, accounting for customers' reactions. This strategy allows RueLaLa to determine the optimal price point for each item. (See Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi, Analytics for an online retailer: Demand forecasting and price optimization, *Manufacturing & Service Operations Management*, 18(1), 2016.)

[73] Grewal et al., 2011, *op. cit*. CVS, notably, is reported to evaluate the effectiveness of its promotional campaigns by comparing outcomes with those for matched control groups.

[74] Unnati Narang and Venkatesh Shankar, Mobile marketing 2.0: State of the art and research agenda, *Marketing in a Digital World* (*Review of Marketing Research*), 16, 2019.

[75] Wedel and Kannan, 2016, *op. cit*.

[76] Dokyun Lee and Kartik Hosanagar, "People Who Liked this Study Also Like": An empirical investigation of the impact of recommender systems on sales volume and diversity, December 2015, retrieved from: https://www.krannert.purdue.edu/academics/mis/workshop/Rec%20Diversity%20Empirical%20DEC%202015.pdf.

recommended products that can be proposed to customers during a shopping session. These recommendations can appear at various stages of the session. They can be shown when the customer enters the webstore (by logging in), presented as similar or complementary products when a customer views a product, or displayed next to the shopping basket before check-out.

Importantly, data-driven recommendation tools are known to be effective in increasing sales by enabling the targeted promotion of products that are most likely to appeal to specific customers or customer segments, e.g. as alternatives or complements to other potential purchases.[77] Using a data sample from a large retailer of women's clothing, for instance, De et al. (2010) investigates the effect of a particular product recommendation system on the retailer's online sales.[78] The authors find that this particular system increased monthly sales by more than 5.5%, including both products on promotion and not on promotion.

Customers shopping online are increasingly likely to see recommended products during their visits. Indeed, many online stores use some form of recommendation system, whether they have designed them in-house or they are utilising the services of third-party providers.[79]

3.2.6 Sales forecasting

For retailers, sales forecasts are important for many business decisions. For example, accurate forecasts allow retailers to improve their inventory management, notably by avoiding overstocking products when demand is low, whilst making sure to avoid running out of stock during periods of high demand and that suitable stock is available to meet emerging and growing demands. Sales forecasts also contribute to other important strategic decisions, notably regarding the allocation of marketing resources, determining product offerings at the store level, as well as the opening or closing of stores.

There are two broad types of forecasts that retailers rely on to achieve different objectives:

- Short-term sales forecasts at the store and product level help retailers to make tactical decisions, e.g., optimising inventory management and staffing plans.

[77] Krämer et al., 2020, *op. cit*.

[78] Pabuddha De, Yu Hu, and Mohammad S. Rahman, Technology usage and online sales: An empirical study, *Management Science,* 56(11), November 2010.

[79] The list of on-line retailers using recommender system is very long but notable names include Amazon, eBay, Walmart, Etsy, Zalando. There are also many third-party providers of such systems, such as Early Birds, which is the leading personalisation provider in France and works with Fnac Darty, Cdiscount and Maisons du Monde among others. Prudsys is a major provider of recommender systems in Germany. Among its customers are OTTO, Douglas, C&A, and Thallia. Shopify is a third-party service provider that focusses on smaller retailers. Although its main service is setting up online stores and payment functionalities rather than recommendations, it has a basic recommendation system for e-commerce stores. Monetate also offers a personalised recommendation tool to retailers and has a number of British and American companies as customers, including, in particular, Waitrose and Partners, Johnston and Murphy, Travelodge, Office Depot, The Scotch Malt Whisky Society, JoJo Maman Bebe. Google is one of the companies offering services targeted to smaller retailers. Since 2014 it has offered a recommender system that, as of 2018, was used by tens of thousands of companies.

- Longer horizon market-level forecasts allow retailers to respond to shifts in demand and supply, making strategic decisions that have longer-term consequences.

Retailers have historically drawn on various data to produce sales forecasts, including past sales, anticipated rival behaviour, and data on the economic environment. However, for a long time, data-based forecasting was secondary to so-called judgemental forecasts, which typically drew on the expertise of the sales force or the views of industry experts.[80] The introduction of electronic scanners in stores in the 1970s and the technological advances in data storage in the 1980s facilitated a move away from a reliance on the experience/judgement of individuals and towards the more quantitative, data-based methods that are at the core of forecasting practice among retailers today.

Today, retailers can rely on a variety of statistical and econometric methods to forecast sales, depending on the context.[81] Indeed, a number of these methods have become standardised tools, as specialist software suppliers such as SAS and SAP offer a range of forecasting options in their demand planning suites. Specialist software providers are also increasingly including machine learning methods, which are quickly becoming an industry standard, in their offerings.[82,83]

Modern sales forecasts draw on a variety of information sources, including granular store- and SKU-level sales data, customer demographics, aggregate market data and macroeconomic indicators.[84] More recently, forecast models have started employing user-generated data, as mined from social media posts and online reviews but also from in-store traffic and path data.[85]

[80] Many retailers relied on their sales force and store audits to produce forecasts before point-of-sale (scanner) data came into wider use. See Heidi Winklhofer, Adamantios Diamantopoulos, and Stephen F. Witt, Forecasting practice: A review of the empirical literature and an agenda for future research, *International Journal of Forecasting,* 12(2), June 1996.

[81] One approach involves the use of univariate time series models. These rely on aggregate sales data and are commonly used to produce market level forecasts. Another type of forecasting model used to predict SKU-level demand is the multivariate regression model. These models typically incorporate data on own and competitor prices, promotional events, and seasonality in addition to data on past sales. Spatial interaction models can use a wide range of inputs, including data from household surveys and censuses, data from geographical information systems and data on the performance of similar stores. These are used to forecast local demand for new stores. In more recent times, various machine learning algorithms have also been added to the suite of techniques employed.

[82] See https://www.relexsolutions.com/solutions/demand-forecasting-software/. Relex's "pragmatic AI" solution combines retailers' sales data with external factors such as traffic, weather and holiday events to improve forecasts.

[83] Robert Fildes, Shaohui Ma, and Stephan Kolassa, Retail forecasting: Research and practice, *International Journal of Forecasting,* 2019.

[84] Fildes et al., 2019, *op. cit*.

[85] Tonya Boone, Ram Ganeshan, Aditya Jain, and Nada R. Sanders, Forecasting sales in the supply chain: Consumer analytics in the big data era, *International Journal of Forecasting,* 35(1), 2019.

4 The competitive benefits of 'more' data in retailing

The central issue addressed in this paper is whether the large – and larger – amounts of data that some retailers control, notably through their online presence, give them an insurmountable competitive advantage over smaller rivals. In other words, does asymmetric access to data imply ineffective retail competition?

The substantial heterogeneity (or differentiation) found in retailing means that a 'one size fits all' approach to assessing this competition concern is not appropriate. Not only will different retailers have access to different amounts, as well as different types of data, but they are also likely to have different data analytical capabilities. Furthermore, and importantly, one retailer's need for data may also differ significantly from that of another, depending on the scale and scope of their operations, as well as the business models they operate.

Data are rarely an end in themselves in retailing. They are typically an input into analytics that are themselves just one element shaping a retailer's business proposition and performance. Moreover, the quality of those analytics will depend on the analytical capabilities of the retailers concerned, as well as the data that are available.

Retailing is about delivering the products and services that customers want. Crucially, whilst data can contribute usefully to a retailer's performance, as has been described, retail competition is about much more than data. Hence, an evaluation of the competitive effects of data asymmetries requires a detailed assessment of the impact that these ultimately have on a retailer's market performance and on competition generally, recognising that the latter will depend on the overall competitive strengths – data-related and otherwise – of rivals.

4.1 The benefits of more data in retail applications

In Section 3 above, we set out a brief review of several retail business applications for which data are a valuable input. This showed that a range of different types of data are used for different kinds of analysis. Against this background, the concern is that retailers that have access to more, relevant data could gain an unchallengeable competitive advantage.

Data are useful – or relevant – for a retailer when they feed into data analytics that improve the retailer's performance, notably in terms of the supply of products and/or services to customers. Ultimately data are useful to the extent that they improve the analytics that the retailer employs to pursue its business strategy.

Here, we assess the benefits of having access to more data in a retail setting specifically, focusing in particular on the performance of the data analytics underpinning forecasting and recommendations for customers.

4.1.1 The benefits of more data in forecasting sales

As presented in Section 3.2.6, retailers use data, notably their own data, to forecast future sales. Having more accurate forecasts helps them to make better decisions.

The general statistical principle that having more observations increases the accuracy of prediction also applies to retail forecasting. All else being the same, a retailer with a larger amount of relevant sales data can be expected to be able to generate more accurate sales forecasts. However, there are likely to be diminishing returns in terms of the additional accuracy delivered. Using data from Amazon, Bajari et al. (2019 and 2018) show that having more data improves the quality of forecasts, but with diminishing returns.[86,87] At the same time, the study also shows that having sales data for a broader portfolio of products, even within the same product category, does not appear to have any material impact on the accuracy of the sales forecast.

4.1.2 Use of user-generated online data in forecasting

Nowadays, it is also possible to enrich forecasting models by combining a company's own data with user-generated online data that relate to the product at issue, such as trends in search queries or comments on social networks. Machine learning algorithms can be fed with input variables that draw on these additional data and, to the extent that these data are relevant, the algorithms might be expected to produce more accurate predictions as a result. In practice, the publicly available evidence on whether adding user-generated data really improves the accuracy of sales forecasts is mixed.

Some studies report some improvement. For example, Cui et al. (2018) shows how data from user interactions on Facebook can be used to improve the daily product sales forecasts generated by a range of machine learning methods for an online retailer that sells men's clothing.[88] The results show that augmenting the forecasting model with social media

[86] Patrick Bajari, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki, The impact of big data on firm performance: An empirical investigation, *NBER Working Paper 24334*, February 2018.

[87] Patrick Bajari, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki, The impact of big data on firm performance: An empirical investigation, *AEA Papers and Proceedings*, 109, May 2019.

[88] Ruomeng Cui, Santiago Gallino, Antonio Moreno, and Dennis Zhang, The operational value of social media information, *Production and Operations Management*, 27(10), October 2018. The name of the online retailer is not disclosed but it is listed among the top 500 largest e-retailers in America in 2014. This study compares simple linear regression, linear regression with forward selection, lasso, support vector machines with linear and radial kernels, gradient boosting method and random forests.

information improved the accuracy for all the machine learning models, except linear regression. The random forest model provided the most accurate forecast, and the use of social media information improved accuracy by 20%. Boone et al. (2018) also shows that using data from Google Trends, which provides data on search queries, yields positive but modest improvements in the accuracy of sales forecasts.[89] Specifically, weekly sales of five SKUs by a speciality retailer of food and cookware were forecast. The study showed that adding Google trend data increased forecast accuracy by between 2.2% and 7.7%, depending on the product.

Other studies, however, highlight the limited usefulness of user-generated online information. For example, Schaer et al. (2019) develops two case studies, one involving sales forecasts for video games and the other forecasts of YouTube views of corporate online videos.[90] In each case, the authors augmented the forecasting exercise with data from the internet. For the video games, the study used information from Google Trends to predict the frequency of relevant search queries. The study also measured the number of times that online corporate videos posted on YouTube were shared on social networks (Facebook, Twitter, Google+ and LinkedIn). In both cases the accuracy of the forecast was not improved materially by using the additional online information.

In summary, so far, few studies have examined how adding user-generated online information can enrich retailer forecasting sales analysis. Whilst the benefits of doing so may depend on the circumstances of each case, these studies have only considered the case of incorporating fairly aggregated data (not individual level data), such as Google Trends, which are publicly available.

4.1.3 Fine-grained data and the quality of data-driven recommendations and promotions

The possibilities provided by a broader range of so-called fine-grained data have begun to be exploited in practice as well as in the academic literature. Retailers are increasingly able to build detailed user profiles by accumulating a range of information on their customers' behaviour, notably their interactions with a webstore, such as click patterns and purchase and

These machine learning techniques can handle the high dimension data. That is, when combining social media with the retailer's own operation data (sales from new customers, sales from repeat customers, advertising, promotion schedule) to forecast sales, the number of variables become large relative to the number of observations. Traditional econometric forecast models cannot handle more variables than observations, whilst machine learning techniques can.

[89] Tonya Boone, Ram Ganeshan, Robert L. Hicks, and Nada R. Sanders, Can Google Trends improve your sales forecast?, *Production and Operations Management*, 27(10), October 2018. This study uses a traditional time series econometric model to generate sales forecasts.

[90] Oliver Schaer, Nikolaos Kourentzes, and Robert Fildes, Demand forecasting with user-generated online information, *International Journal of Forecasting*, 35(1), January-March 2019. This study relies on linear time series regression models with auto-regressive terms for sales and contemporaneous and lagged explanatory online information variables to generate sales forecasts. Because the number of variables is greater than the number of observations, this study relies on lasso regression. Essentially lasso regression forces the coefficients on uninformative variables to zero.

browsing histories – but also product ratings they submit or comments they provide, for instance.[91] Moreover, advanced analytical tools, such as machine learning algorithms, may enable these data to be used to generate tailored, data-driven promotions, advertisements and product recommendations, targeted at specific customer segments or even individual customers. (That said, as Wedel and Kannan (2016) observes, "*the availability of big data with extensive individual-level information does not necessarily make it desirable for companies to personalize at the most granular level*".[92])

A critical question is whether and, if so, the extent to which more data improve the quality of the recommendations generated, with more of them being converted into purchases as a result. In principle, the availability of richer data should help. That said, not all data are relevant. A retailer that sells millions of products can collect large amounts of data on what its customers are viewing and purchasing. However, most of these data will have little relevance for a particular customer at a particular point in time. For example, having information on, say, a customer's television purchases, such as which televisions they searched for and comments they provided on various TV brands, is unlikely to provide much insight into the preferences of customers looking for footwear or crime novels. Hence, whilst retailers that sell many items are likely to collect more data about individual customers, many of those data will not be relevant to a particular query.

We have not identified many published studies to date that investigate the extent to which the performance of product recommendation systems improves with more data. There are possibly several reasons that explain this. One is that this is a recent area, and another reason might be the fact that evaluating such performance would involve assessing the extent to which recommendations translate into actual sales. Such information is not typically publicly available.[93]

Martens et al. (2016) is notable as one such study. It evaluates the performance of machine learning tools for personalised marketing. Specifically, this study uses a large, fine-grained dataset from a bank to evaluate the performance of pseudo-social network algorithms for

[91] Personalisation platform Reflektion, for example, offers a service which "*is constantly learning, ingesting each new data point (click, view, add-to-cart) and product details to show targeted products in real-time*" (see https://reflektion.com/modules/#recommendations), whilst e-commerce revenue optimization platform Granify claims that it "*has ingested detailed behavioral shopping data on more than 10 billion shopper sessions in the last 3 years",* with over 500 data points being collected every second in each session, allowing the platform to decide "*exactly to which shoppers to display a specialized message, what sort of message to use, and the precise moment in which it will have the greatest positive effect on the session's expected revenue*" (see https://www.granify.com/our-solution/how-it-works.)

[92] Wedel and Kannan, 2016, *op. cit*.

[93] The literature suggests using customer satisfaction, which is inherently difficult to measure. See Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, Evaluating collaborative filtering recommender systems, *ACM Transactions on Information Systems,* 22(1), January 2004; Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef, Recommendation systems: algorithms, challenges, metrics, and business opportunities, *Applied Sciences,* 10(21), November 2020.

targeted marketing, namely offering financial products to customers that are similar to those who already bought these products. The study considers whether adding fine-grained data on consumer transaction behaviour (e.g., payments made by customers) delivers any benefit. The results indicate that the performance of the predictive analytics is improved by adding fine-grained data into the analysis. However, the improvement comes with diminishing returns. That is, beyond a certain point, the performance improvements achieved by adding more data become very small.[94]

4.1.4 Scalability and real-time use of data

One of the major challenges associated with the use of recommender engines is scalability. Whilst having rich user data can help improve the quality of recommendations, making them more relevant to particular customers or customer segments, the capacity of algorithms and computers to provide high-quality recommendations, notably in real time, is a major challenge.[95] In fact, many algorithms that are currently used for making personalised recommendations are based on so-called item-based collaborative filtering for precisely this reason.[96]

Real time scalability remains a critical issue, meaning that the algorithms struggle to handle large volumes of data to produce recommendations in real time. One approach deployed is to reduce the data by sampling users, but other methods to cut the data burden can be used too, depending on the setting.[97] This means that, even though having more data might still allow the analyst to improve the performance of the algorithm offline, the usefulness of more data to real time analysis is likely to be limited.

4.2 Asymmetric data requirements

An overall asymmetry of data does not necessarily imply a competitive disadvantage for the party with "less" data. For instance, a retailer that is active across many retail segments may well have many more customers overall than more specialised retailers with which it competes, and access to much more customer data than any of those specialists individually as a result.

[94] David Martens, Foster Provost, Jessica Clark, and Enric Junqué de Fortuny, Mining massive fine-grained behavior data to improve predictive analytics, *MIS Quarterly*, 40(4), December 2016.

[95] Recommendations must often work in real-time. That is, based on the current behaviour of the shopper in question, either searching, browsing, or purchasing, the engines will make immediate data-driven recommendations. The objective is to stimulate impulse purchases.

[96] User-based collaborative filtering is an alternative family of algorithms that can be used to make recommendations for customers. However, scalability issues have affected its development. On the other hand, item-based collaborative filtering algorithms recommend items that are similar to those that the customer is viewing or purchasing. Amazon pioneered this approach, and this algorithm has been used by Amazon since 1998, when it was a bookstore. See Brent Smith and Greg Linden, Two decades of recommender systems at Amazon.com, *IEEE Internet Computing,* 21(3), May-June 2017. Others, such as Netflix and YouTube and many other large and small players in the online world, are employing this type of algorithm.

[97] Julian Jarrett, A present-day perspective on recommendation and collaborative filtering, in Brent Smith and Greg Linden, Two decades of recommender systems at Amazon.com, *IEEE Internet Computing*, 21(3), May-June 2017.

However, the broad-based retailer is also likely to need access to more data overall simply to hold an equivalent amount of relevant data to its more specialised rivals in relation to each of the segments in which it competes. Hence, an overall asymmetry of data will not generally imply a competitive advantage for the general retailer in any individual segment. In other words, simply comparing the overall volumes of data available to different retailers, without also considering the differences in the product portfolios and other attributes of those retailers, will not shed light on the question of whether the data asymmetry confers a competitive advantage on one retailer over another.

Some retailers will also be more dependent on data-based analytics than others. This may be a consequence of the products they sell and/or the customers they serve. It may also arise because a retailer has adopted a business strategy that is more data reliant. For example, effective product search and/or recommendation algorithms are liable to play a more important navigational role for a general online retailer that sells an extensive and diverse range of products to a heterogeneous set of customers than for a specialised retailer, whose product range and sales activity is already likely to be more targeted. For the former, the task of matching customers with the products that would best meet their needs and preferences (from a potentially enormous array of possibilities within the range of products offered) is a key challenge. In contrast, a retailer with a narrower proposition, or with a focus on bricks-and-mortar operations, may rely less on sophisticated data analytics to steer customers through its product offering.

In other cases, a retailer that naturally attracts a more homogeneous customer base may have less need/use for data and sophisticated analytics to tailor its recommendations to the demands of different customer types.

Thus, some retailers are likely to need access to more data in order to compete effectively than others. Hence, even if they have access to more data, this may not translate into a competitive advantage. (It should not be a surprise either if these retailers also place a greater emphasis on obtaining data.) A simple comparison of the overall data available to different retailers is unlikely to be very informative in these circumstances.

It is possible, of course, that a general retailer will have access to more data that are relevant to competition in a particular segment than more specialised rivals that are focused on that segment. (For instance, despite its broader outlook, the general retailer may simply make more sales in a particular segment. Or information on sales in one product area may inform better recommendations in another, e.g., where there is an inherent complementarity or correlation in demands.) The issue then is the extent of any competitive advantage that these additional data would deliver, bearing in mind the likely diminishing returns associated with more data.

4.3 The role of data analytics

The competitive impact of asymmetric data access will also depend on the retailer's data analytical capabilities. Indeed, a key insight from the literature is that the value that can be extracted from data often depends on these complementary capabilities. For example, as the volume of data increases, the ability to process and analyse those data effectively becomes an issue. Bradlow et al. (2017), for instance, discusses the difficulties of compressing a large volume of data for effective use.[98] Unless a retailer is able to process and analyse data effectively, it may not be able to make good use of an expanded set of even highly relevant data.[99]

Furthermore, the value that a retailer extracts from given data may have more to do with the quality of its analytics than the inherent value of the data concerned. For example, a study run by Aarki, a company that uses machine learning tools for mobile advertising, compares four different methods for predicting the performance of mobile app marketing campaigns using a very large random sample.[100] This study reveals significant differences in prediction accuracy between the different approaches that were evaluated.[101] This suggests that the choice of methods is a key factor in determining the performance of the analytics. Moreover, Aarki notes that designing and fine tuning the algorithm is also important to achieve better performance.

It follows that the value of given data to a retailer that has access to the requisite data analytical assets, including human capital, may be very different to the value of the same data to a retailer that does not have the same resources. This means that evaluating the importance of data to a retailer should also account for the appropriate data analytical resources that this retailer can access. A simplistic assessment that omits to do this risks overstating the value of the underlying data themselves.

[98] Eric T. Bradlow, Manish Gangwar, Praveen. Kopalle, and Sudhir Voleti, The role of big data and predictive analytics in retailing, *Journal of Retailing,* 93(1), March 2017.

[99] As noted, that is a particular issue where such processing must occur in real time.

[100] Aarki indicates that it used 280 million transactional records for this study. For more, see https://www.aarki.com/blog/using-machine-learning-to-predict-campaign-performance.

[101] Aarki used the same training data for all four models. Once the models were trained, they were used to predict the outcome for the test data. The performance of the models was evaluated using mean squared error (MSE). Specifically, Aarki compared the performance of artificial neural network, random forest and support vector machine against the more traditional Tobit model. All the three methods outperformed the Tobit model. However, support machine vector was clearly the best approach here. Its MSE was almost 25% lower than that of artificial neural network and about 17% lower than random forest.

4.4 Retail competition is about more than data

Retailing is, ultimately, about delivering the products and services that customers want, and almost never about data as such. Whilst data can contribute usefully to a broad range of retailer activities, as has been described, retail competitiveness is typically about much more than access to and use of data. In that respect, the retail sector at large is quite different to the data-centred sectors that have been the principal motivation for the data-related concerns we reported in Section 2 above.

Factors other than data are highly relevant to retail competitiveness. Competition in retailing spans a number of dimensions, such as:

- price, range and quality of products;
- service levels and store design; and
- branding and promotional effort, etc.

Importantly, such competition is dynamic, with innovative new entrants demonstrably able to target unfulfilled opportunities in retail markets where they arise. The growth of online shopping has facilitated this process.

The trajectories of numerous online retail start-ups suggest that pre-existing data are not a pre-requisite for success. For example, recognising that digitalisation lagged in DIY, online marketplace ManoMano was able to grow its sales from €90 million in 2016 to €1,200 million in 2020.[102] Online fashion and lifestyle retailer Zalando, founded in 2008, reported revenues of almost €8 billion in 2020.[103] Cdiscount is a popular French general online retailer, which is part of the Casino Group, recording sales of more than €4.2 billion in 2020, including through a marketplace that started ten years ago. In fact, in the last ten years, Cdiscount doubled in size.[104] Moreover, new retail businesses are able to draw on e-commerce platforms such as Shopify, for example, for a range of services supporting effective online strategies, enabling them to establish standalone routes to market or to utilise marketplaces, for instance.[105,106]

[102] See https://www.ft.com/content/8208168a-697a-4dd5-876e-3ca2c20374b5

[103] See Zelando Annual Reports for 2016 and 2020. Zalando started as an on-line shoe store in Germany but is now active in at least 17 European countries, and appears to be the largest online fashion platform in Europe. (See https://blog.brandsom.nl/en/selling-on-zalando)

[104] See https://fr.wikipedia.org/wiki/Cdiscount and https://www.franceinter.fr/emissions/histoires-economiques/histoires-economiques-19-janvier-2021.

[105] Shopify claims to serve over 1,700,000 businesses in 175 countries (see https://www.shopify.com/) and to support around 9% of all e-commerce (see https://finance.yahoo.com/news/decoding-shopify-fashion-future-100009112.html). The number of consumers buying from Shopify merchants grew 52% in 2020 compared to 2019 (see https://news.shopify.com/shopify-announces-fourth-quarter-and-full-year-2020-financial-results).

[106] Shopify offers a number of illustrative case studies on its website (see https://www.shopify.com/plus/customers/use-cases/growth-and-scale).

Data analytics may help retailers make better decisions with respect to these various aspects of competition. However, such analytics are just one element shaping a retailer’s customer proposition.

A range of other factors can also affect the performance of a retail business, including efficient logistic organisation and purchasing economies. Whilst data and data science may also contribute to these aspects of performance, it would be wrong, therefore, to associate a retailer’s relatively strong or weak competitive performance with data asymmetries alone.

For example, a retailer that operates at greater scale than its rivals may be better able to achieve purchasing economies, e.g., because of superior negotiating positions vis-à-vis its suppliers. In turn, these cost advantages may translate into an ability to offer lower prices to customers, which can be expected to further boost sales, etc. Hence, there are entirely conventional reasons why a larger retailer may have gained a competitive advantage over rivals that have nothing to do with data, as such.[107]

Data asymmetries may not translate into material competitive advantages for data-rich retailers, let alone into unassailable competitive positions. The latter would rely on initial data asymmetries delivering decisive and entrenched competitive advantages either directly or via positive feedback effects. There are, however, important countervailing forces at work in retail markets that act to limit such concentrating tendencies.

Furthermore, some retailers may combine data-related strengths with relative weakness in areas that are not data-focused, such as product sourcing, branding, or promotional skills. A focus on data in isolation may then misleadingly suggest competitive advantages which are illusory.

4.5 Data-driven monopolisation

As introduced in Section 2.2 above, commentators have stressed the potential role of data-driven network effects in generating and augmenting competitive advantages for the leading digital players. Particular concern has been expressed that this may lead to tipping, whereby even an initially modest competitive advantage may be transformed into an unassailable market position.

[107] As noted, those advantages might also have been expected to give rise to positive feedback effects and potentially to market tipping if this were a significant consideration in retailing.

However, as noted, the positive feedback loop stimulated by data-driven network effects will operate only to the extent that more data deliver better data analytics, e.g., more accurate forecasts or improved recommendations for customers, and these are useful in the context of the retailer's particular business strategy. Hence, the impact of the data-driven network effect is closely connected to the competitive benefits associated with additional data. As also discussed above, even where more data can improve the performance of the analytics, this is typically subject to diminishing returns. This means that the feedback effect is likely to become smaller and smaller as more data are analysed.

Moreover, even if more data would improve the performance of some of the analytics used by retailers, there are other important factors at work shaping competition in retail markets that will tend to limit the competitive impact of any data-driven advantage. We have highlighted some of these factors already. Importantly, retailing is a differentiated activity, not just in terms of the products and services that retailers sell, but also in terms of the way that they do this. For instance, whereas some retailers specialise in particular segments, others have positioned themselves as generalists; whereas some are positioned as high quality, even luxury, retailers, others have adopted 'budget' positions. Some retailers are vertically integrated, sometimes only stocking their own products. Some now are exclusively online businesses, while others are still focused offline (though most will have at least some online presence). Different retailers will appeal to different customers. Customers will also use different retailers for different shopping purposes.

In addition, low costs of searching, visiting, and purchasing from multiple retailers in an online context will encourage multi-homing. In that respect, online search engines and comparison websites can also play an important facilitating role, allowing customers to navigate between retailers in search of the products they want on the best available terms.

Retailer differentiation and customer multi-homing are features (amongst others) that militate against monopolisation.[108] Indeed, they facilitate entry by new retailers where these are better able to address consumer demands.

Significantly, moreover, a number of more traditional mechanisms – e.g., larger retailers obtaining discounts from suppliers in exchange for greater purchase volumes, thereby improving their cost competitiveness still further – might also have been expected to result in

[108] In a very recent study, Akman (2021) evaluates the results of an extensive online survey exploring consumer attitudes to and use of online platforms. The study highlights the extent of multi-homing, amongst other observations. (See Pinar Akman, A web of paradoxes: Empirical evidence on online platform users and implications for competition and regulation in digital markets, *SSRN 3835280*, March 2021.)

positive feedback. That they have not given rise to widespread retail tipping further suggests that data-driven network effects are unlikely to do so either.

In summary, the fact that data analytics are subject to diminishing returns and that a number of features naturally limit market concentration suggests that data-driven network effects are unlikely to give rise to market tipping in the retail sector.

5 Costs and benefits to mandated data access

Calls have been made for data sharing to be mandated in order to address asymmetries in access to data. As set out in the previous section, asymmetries are a natural outcome of the differentiated strategies adopted by different retailers and need not undermine effective competition. An individual retailer's use/need for data will depend on the scope of its activities as well as its business model and the importance of data to these. This means that data that are important to one retailer's ability to compete may make little or no contribution to another retailer's competitiveness. In this context, a requirement for equal access to data may not be useful.

At the same time, mandatory data sharing is likely to be costly – both for the provider of the data and for the recipient. Where access to competitor data may anyway be not only unnecessary for a retailer to be able to compete effectively, but also of very limited value in practice, the costs of any data access regime may easily outweigh the competition benefits.

5.1 Factors affecting the benefits of symmetric access to data

5.1.1 Differences in the relevance of particular data

Data that are highly relevant for one retailer may have limited usefulness for another retailer. For example, one retailer may rely heavily on the specific purchase histories of its customers to generate product recommendations. Those data may be of much less relevance and value to another retailer. For instance, the information used by the first retailer may relate to customer and product types for which there is limited overlap or correlation with the customers served/targeted and products offered by the second retailer. Hence, even where access to certain data is critical to one retailer's competitiveness, another retailer may not require access to the same, or as much, data in order to be able to compete effectively.

As noted, a retailer's own customer/sales data are likely to be most useful to it. These are also likely to be most accessible to it.

5.1.2 Relevance of data and the scope of a retailer's activities

Some retailers, such as Alibaba, Amazon, or eBay, operate across an array of retail segments and in multiple geographies. Others will be more specialised, either in terms of the products they sell or the geographies they serve, or both (e.g., Zalando, which specialises in fashion items such as apparel and shoes in various European countries, or ManoMano, a DIY retailer which initially was mostly present in France but has expanded in other European countries).

A wider array of data are likely to be relevant to general retailers than more narrowly focused rivals. This would be the case, for instance, if the usefulness of data to a general retailer mirrors that for the specialists on a segment-by-segment basis but extends across the many more segments in which the general retailer is active. Equally, in that case, much of the general retailer's customer data, drawn from a range of retail segments, may be of limited relevance to a specialised retailer that is active in one segment only.

It is possible, of course, that data from across a broad range of segments could also be of value within an individual segment. That might be the case if there are important correlations in sales of products in different segments that transcend segment boundaries. For example, knowing that a customer browsing potential book purchases is also a keen purchaser of cycling clothing might suggest that a cycling book recommendation would be well received. It might also be the case where personal information for a customer that is active across multiple segments has been gathered in one segment but can then be deployed in other segments too.

5.1.3 The relevance of data and a retailer's business model

The importance of particular data to a retailer's performance will also depend on its business models and strategies, which may differ substantially from those of its rivals. For example, as noted, product search and/or recommendation algorithms may play a more critical navigational role for retailers that sell an extensive and diverse range of products than for more focused, specialised retailers. Equally, whilst some retailers may have developed a customer offering that relies heavily on data-centred functionality, others may have decided not to invest as much in this space.

5.1.4 The relevance of data and a retailer's data analytical capabilities

As also noted above, the ability of a retailer to make effective use of data will depend on its data analytical capabilities too. Hence, even data that would be highly relevant when combined with appropriate data analytical tools may not contribute substantially to a retailer's ability to compete where that retailer does not possess the complementary capabilities required.

5.2 Costs of data sharing

Interventions to require sharing of data will impose costs on both providers and recipients.

Firms that bear the duty to share their data will incur associated costs. These will include the direct costs of establishing and maintaining the systems and processes required to implement the data sharing. For example, a retailer may be required to make data available in particular formats, which might be very different from its existing ones. The provider may also incur indirect costs if its own uses of data are constrained by the requirement to adhere to data access protocols.

From a simplistic viewpoint, the option to access more data might be expected always to deliver positive value for the recipient. However, that supposes that the costs incurred by the recipient in this process would be limited, or could be compared with the likely benefits of receiving data before such a request was made, at least. The recipient may have costly obligations attached to even the *possibility* of receiving data – notably the requirement to protect personal data. Further, dealing with extraneous data (that are not useful to the recipient) may also impose additional filtering and processing costs on the retailer where they are commingled with useful data.

Importantly, a requirement to share data may give rise to significant dynamic costs where it disincentivises data collection and processing or the development of valuable new data uses. More generally, regulatory interventions which would limit or eliminate data-driven competitive advantages risk stymieing a healthy competitive process, limiting or eliminating the incentives that retailers have to invest in new sources of such competitive advantage, to the of detriment of customers. Therefore, whilst interventions to redress data asymmetries might yield short term competitive benefits in some cases, these must be balanced against the potentially significant costs, especially over the longer term.

Interventions that impose a duty to deal are rightly contemplated only in exceptional cases, such as in the utilities sector, where the prospects of effective competition would otherwise be very limited. These circumstances are quite different to those that prevail in the retail environment, where differentiated competition, as opposed to natural monopoly, is the norm.

6 Conclusions

Very general concerns have been voiced regarding the competition implications of some digital platforms' superior access to data. A key insight from our review is that such generalisations are not appropriate. A one size fits all approach to data, and access to data, is not justified. This is because the relevance of particular data (and particular data asymmetries) for competition depends on the type of data involved, the extent of the benefit of more data and their impact on competition, as well as market context. Thus, a concern that is grounded in a given feature of online search, say, cannot be simply translated to a retailing environment.

Data used in retailing are not homogeneous. Indeed, all sorts of data are employed by retailers in a variety of different applications. This is because firms have different uses for data, and different data needs therefore, depending on their business models and commercial strategies, as well as the type and range of customers they seek to serve. The incremental value of particular data to a retailer will depend on all of these aspects, and on how much data the firm already has, with diminishing returns to additional data typically applying.

Crucially, data analytics are not an end in themselves in retailing. The results of these analytics guide retailers' strategies, as well as shape their business operations. However, many other factors affect retail success and the outcomes of retail competition. Any assessment of the competitive effects of data in a retail setting must, therefore, involve a holistic appraisal of the different aspects of competition more generally, whether data-driven or not.