Big Data Hubris

WILL RINEHART | AUGUST 21, 2018

Executive Summary

Cutting-edge technologies such as facial recognition, self-driving cars, and AlphaGo suggest a new world of big data, automation, and artificial intelligence (AI) is upon us. Their impacts have been heavily debated, most notably in discussions of worker productivity and decreasing investment in firm capital.[1][2][3] Increasingly, however, the most contentious space with regard to big data has been in competition policy.

Rep. Zoe Lofgren has been pushing for a more expansive probe of how big data affects competition.[4] Senator Mark Warner recently released a whitepaper on how tech might be regulated, singling out big data in particular. Policymaker and antitrust scholars are trying to understand how the era of big data should change competition, if at all. The complete picture actually shows a more complex picture than one would imagine.

As this comment details, three considerations are worth keeping in mind to understand the competitive effects of big data:

- Companies can gather all the data in the world, but it only makes a difference if they use it to improve their product;
- Big data projects are rife with failure, which speaks to the limits of it as a competitive advantage; and
- Finally and most importantly, big data simply doesn't meet the four criteria of being a barrier to entry, which are inimitability, rarity, value, and non-substitutability.

What is missing is a fuller appreciation of the organizational side of the equation, i.e. an understanding of those companies or firms that actually made big data implementations possible. Embedding new technologies into the production cycle in a way that generates positive returns has long been a difficult task. The cost to start up new production methods, sometimes called switchover costs, can often delay the adoption of technology. Thus, to fully appreciate the larger discussion surrounding big data, it is first necessary to understand a bit about big data and then understand how firms integrate these processes.

The Value of Data and Big Data Hubris

Confusion abounds over fundamental statistical concepts and how they relate to economic value.

For example, Senator Mark Warner's whitepaper on regulating tech companies proclaimed, "Unlike many other assets, which tend to illustrate declining marginal utility, the value of any piece of data increases in combination with additional data."[5] A recent *Economist* article repeated this sentiment, suggesting that data "are an inexhaustible resource: the more you have, the more you get."[6] For this reason, big piles of data can become a barrier to competitors entering the market, says Maurice Stucke of the University of Tennessee.

For another example, Georgetown Professor Paul Ohm said,

Research suggests that the power of AI techniques, such as neural networks and other forms of machine learning, do not proceed linearly. An algorithm trained on a million data points beats the accuracy of one trained on only one thousand data points by much more than three orders of magnitude.[7]

Stuart Kirk, head of Deutsche Asset Management's Global Research Institute, teased out the implications. If machine learning assets live up to the hype and grow more valuable over time, then depreciation would be driven down. The result would be substantial for a company like Google. "If only a tenth of that was reversed, the company's market capitalisation would rise by about one-fifth, based on current multiples after tax."[8]

Yet, there is a decided difference in how statisticians, data scientists, and economists use specific terms and how attorneys, bankers, and boosters interpret them. Power, as Ohm is using it, relates to the probability that a statistical test will correctly reject the null hypothesis. That is, the probability of a false negative goes down with an increase in the number of observations. That isn't the same as saying that the value of data is increasing in a greater than linear fashion, but that the predictive power of the model is rising.

Big data can take on one of two broad meanings. In one regard, big data can be understood as many observations of one variable or large n. A good example of this is stock market data, where there are data second by second over years. Big data can also mean data have a large number of variables or p. This is the kind of information that Facebook and Google tend to possess.

The presence of big data doesn't mean better predictions will naturally result. As p increases, for example, the number of potential hypotheses that could explain relationships between the variables rises by 2 to the p power. The number of causes in big data sets actually do rise exponentially. In other words, as p expands, so does the possibility of a spurious correlation.

The rise and fall of Google Flu Trends serves as a parable for some of these concerns. While Google Flu Trends made headlines for its brilliant use of big data in the early 2010s, it was at first a relatively reliable estimate of flu outbreaks.[9] But over time, the prediction capability dropped. In a widely publicized paper in *Science*, a team of Harvard-affiliated researchers found that the method had over-estimated the prevalence of flu for 100 out of the 108 weeks through 2012 and 2013. Indeed, a simpler model would have forecasted flu better than GFT.

The belief that big data are a substitute for traditional data collection and analysis, rather than a supplement for these capabilities, may lead to big data hubris. As the authors of the *Science* paper warned,

> [The] quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data. The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

Expanding the pool of data for a model and the resources dedicated to modeling doesn't often yield better results, contrary to what some argue. Researchers Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta found just last year that the performance of a computer vision model increases logarithmically with the size of training data.[10] So, every additional piece of data did not grant exponential or even linear benefits, but decreasing improvements to the tune of 1/x. The hardware company NVIDIA extensively documents methods to achieve linear scaling because data processing methods can quickly be swamped by large data.[11] Even DeepMind's AlphaGo project showed declining returns about midway through the training session.[12]

For competition authorities, these examples are important. First, an enlarged pool of data needs to result in significantly better model predictions. Yet, this isn't the case for many applications. Then, those models need to provide meaningful insights that can then be used at various levels of the firms for real returns. Given that both of these steps must be present, it is hard to see how merely adding another piece of data garners an ever increasing value.

By itself, big data is simply not sufficient to create profit-enhancing opportunities. The bankruptcy proceedings for Caesars Entertainment, a subsidiary of the larger casino company, offer a unique example of this problem. As the assets were being priced in the selloff, the Total Rewards customer loyalty program got valued at nearly $1 billion, making it "the most valuable asset in the bitter bankruptcy feud at Caesars Entertainment Corp."[13] But the ombudsman's report understood that it would be a tough sell because of the difficulties in incorporating it into another company's loyalty program. Although it was Caesar's' most valuable asset, its value to an outside party was an open question.

Most valuations of big data are crude since they divide the total market capitalization or revenue of a firm by the total number of users.[14] When Microsoft bought LinkedIn, for example, reports suggested that the value of a monthly active user was $260.[15] But when that kind of data is sold in the open market, user data typically get a few cents per entry.[16] Valuing big data through a user's revenue stream exhibits omitted variable bias.[17] Human capital has to be applied to data to make it useful, and it is that applied ingenuity where value resides.

While data is helpful for companies to figure out their product niche, product differentiation makes a business successful. Differentiation is not done by having more data, but by coming up with better business forms that use that data. Startups have always succeeded and continue to despite beginning with little data at their disposal. This is why we see the likes of Snapchat, AirBnB, Waze, Spotify, Zipcar, Jet, Udacity, and Warby Parker succeed in the era of big data. Despite their rivals having copious amounts of data, startups are able to succeed by offering a better or different experience or product than their rivals.

Switchover Costs

Industries of all stripes are enamored with the possibilities of data-driven change. A recent survey by Deloitte found that 76 percent of those companies with an aggressive AI plan believe that they will "substantially transform" their operations within the next three years.[18] But the difference between the hype and the reality is stark. The transitioning to production processes that use insights derived from big data often proves costly. In a striking number of cases, the result is utter failure. That big data projects are so rife with failure speaks to the limits of IT as a competitive advantage.

For example, a study performed by PriceWaterhouseCoopers and Iron Mountain discovered that 43 percent of companies realize little actual benefit from their big data projects. Gartner analyst Nick Heudecker? put that number higher. In his estimation, 85 percent of big data projects fail.[19] A couple of years back, research from

the consulting firm IDC found that 20 to 25 percent of projects don't show any return on investment, an outright fail, and as much as 50 percent need massive reworking.

Some might be surprised given the hype, but large companies with talent and resources are still beset by failures and cost overruns. Just to name a few examples:

- Google's DeepMind lost roughly $162 million in 2016.[20]
- Facebook might have access to vast engineering resources and data about language, but still their chatbot project, M, fell short. According to one source familiar with the program, M never surpassed 30 percent automation.[21]
- Microsoft's Tay chatbot was pulled from Twitter less than 24 hours after it launched because it was repeating offensive and vile tweets.[22]
- Ocado, the UK-based online supermarket, explained in late 2017 that the company would need to spend "an extra couple of million pounds" to hire software engineers to work on automating its warehouses.[23]
- Facebook tried to largely automate their content moderation process, but had to pull back on the project and has instead upped the number of content moderators.[24]
- After years of development, T-Mobile is getting rid of its robotic customer service lines.[25]
- Throughout 2017, Tesla founder Elon Musk was touting how the company would be implementing advanced robotics in the new factories. But after missing countless production estimates, he explained that "Excessive automation at Tesla was a mistake."[26]
- Google's various end-roads in robotics, "held the industry back more than moving it forward," claimed one CEO.[27]

The examples above reveal the pecuniary costs in big data technologies, but they don't speak to the equally expensive task of overhauling management techniques to make the new systems work. New technologies can't be seamlessly adopted within firms, they need management and process innovations to make the new data-driven methods profitable. The total sum of these costs, including the potential for lost revenue, are known as switchover costs.[28]

Ryan Avent explained it this way,

> But the scarce factor isn't capital equipment. What is expensive is the intangible capital that's needed to overhaul production in ways that use cheap computing power to eliminate lots of jobs (increase productivity). It is complicated to figure out how to get these systems working and operating in a way that generates profits.[29]

Those firms most adept at integrating new technology have been steadily pulling away from their competitors.[30] While the entire economy might have sluggish productivity in recent decades, the reason isn't due to leading-edge firms slowing down their productivity, but because others in their sector aren't keeping up. These leading companies have come to be known by several monikers, like frontier firms, or superstar firms. They have similar characteristics, including high levels of productivity, high wages, and relatively large size within their sector.[31] They also commercialize more new products than their competitors, incorporate more technologies into the final product, bring their products to market in less time and compete in more product and geographic markets.[32] While Silicon Valley has come to stand for tech giants, frontier firms can be found in retail, wholesale, manufacturing, services, finance, real estate, and countless other services.[33] Throughout the economy, some firms have been pulling away from their counterparts.

Employees within these frontier firms are able to utilize capital better, compared to what they could have done on their own. Because of this, they can quickly introduce cutting-edge technologies that are both complex and have uncertain outcomes but are potentially rewarding.[34] On the other hand, firms without these workers face difficulties. Research from the OECD Economics Department finds "strong support for the hypothesis that low managerial quality, lack of ICT skills and poor matching of workers to jobs curb digital technology adoption and hence the rate of [technology] diffusion."[35]

On the face of it, older firms would seem to possess the kind of data that makes entry from new firms difficult. Yet, management surveys constantly attest to the problems that the data deluge has caused. Often, the production of this data often is by outdated technology that in turn holds an organization back from adopting newer techniques.[36] Moreover, reliability and quality issues plague managers in their ability to trust insights gleaned from these older processes. And the large amount of data means that a significant portion of the generated data is simply unused.

In brief, young companies and entrants are able to better integrate new tech because they haven't become stuck in their practices. While the overall age of companies has risen, the OECD noted that,

> Firms at the global productivity frontier…are on average younger, consistent with the idea that young firms possess a comparative advantage in commercialising radical innovations, and firms that drive one technological wave often tend to concentrate on incremental improvements in the subsequent one.[37]

The value of data comes, not in its bigness, but in its ability to help guide choices within the firm. Young firms, far from being at a disadvantage, are actually in a better place to adapt new tech since they can leapfrog their competitors.

Big Data As A Barrier To Entry

Entry barriers are understood in one of two ways. Bain was the first to dive into the subject and suggested that, "A barrier to entry is an advantage of established sellers in an industry over potential entrant sellers, which is reflected in the extent to which established sellers can persistently raise their prices above competitive levels without attracting new firms to enter the industry." A decade later, Stigler retorted that a barrier to entry should only refer to an advantage that the incumbent firm has that an entrant cannot secure. Big data needs to be understood through a Stiglerian lens. As such, it simply doesn't meet the four criteria of being a barrier to entry, which include inimitability, rarity, value, and non-substitutability.[38]

The effects of big data can be imitated. Academics, companies, and enthusiasts have built an extensive

collection of tools to understand current data sets and are working to develop techniques that don't need large data sets to be useful. Startup Geometric Intelligence has developed machine-learning software that requires far fewer examples to learn a new visual task.[39] Guides on how to deal with small data sets have popped up and can be set up with relative ease.[40][41] Undeniably, the data science community has been actively working to make tools more democratic, more useful, and easier to use.

Others have taken to creating their own data sets from scratch because their niche is more specific than what the biggest players are able to provide. Edison Software is an email management company that has taken on Google with its smart reply system. As they explained to the Wall Street Journal, the task of automating smart replies is broad and challenging. Even Google, which has access to the emails of billions of users, has only implemented smart replies in a very narrow range of cases.[42] When Edison first began, they didn't have enough data to train the algorithm, but they were able to create that data over time.

The problem of getting big data systems to work is so fundamental that many startups and large companies have pseudo-AI implementations. Some of the output comes from a big data system, but those outputs are watched over and supplemented by workers. Andrew Ingram, the widely publicized AI assistant from X.ai, is actually supported by a group of 40 Filipino workers that manage it.[43] Expensify admitted last year that it had been using humans to transcribe receipts, claiming it was instead a victory under their AI driven "smartscan technology."[44] Much like Expensify, it is common for startups to employ people via Amazon's Mechanical Turk system to code data and build a useable data set.

By its very nature, data isn't rare either. While it has often been compared to the new oil, oil cannot be duplicated like data can. My use of data doesn't mean you can't use the same piece of information. This is well summarized by Geoffrey Manne and Ben Sperry, who explain,

> To say data is like oil is a complete misnomer. If Exxon drills and extracts oil from the ground, that oil is no longer available to BP. Data is not finite in the same way. To use an earlier example, Google knowing my birthday doesn't limit the ability of Facebook to know my birthday, as well.[45]

Bots constantly scour the web looking for useful information. Google even accused Microsoft of scouring its search results to improve Bing's own search listings.[46] In an era of big data, it is difficult to see how it could be rare and thus a limitation for entering into a market.

As for its value, as detailed above, big data isn't valuable on its own. And finally, the insights produced from big data can be substituted. LiftIgniter, an AI startup, says that it has a predictive algorithm that will allow smaller players to forecast in real time what a user wants to click on next, and then provide it commercially, an advantage that Big Tech companies have.[47] But a company might not even need to consult with LiftIgniter as similar implementations already have been tested.[48]

In total, big data is of little use in itself. Startups can compete with their more established rivals if they provide customers a better experience, and in many cases, younger firms have the advantage because they are not set in their ways.

---

[1] Mark Muro and Scott Andes, *Robots Seem to Be Improving Productivity, Not Costing Jobs*, https://hbr.org/2015/06/robots-seem-to-be-improving-productivity-not-costing-jobs.

[2] David Rotman, *It Pays to Be Smart,* https://www.technologyreview.com/s/608095/it-pays-to-be-smart/.

[3] Ajay Agrawal, Joshua S. Gans and Avi Goldfarb, *Exploring the Impact of Artificial Intelligence: Prediction versus Judgment*, https://www.aeaweb.org/conference/2017/preliminary/paper/BHKkz394.

[4] David McCabe, *Lofgren wants probe of how Big Tech's data affects competition*, https://www.axios.com/lofgren-wants-probe-of-how-big-techs-data-affects-competition-1513307317-cb04e56c-a8bc-40cf-9cc0-ec0da24d203e.html.

[5] Senator Mark Warner, *Potential Policy Proposals for Regulation of Social Media and Technology Firms*, https://graphics.axios.com/pdf/PlatformPolicyPaper.pdf.

[6] The Economist, *How regulators can prevent excessive concentration online*, https://www.economist.com/special-report/2018/06/28/how-regulators-can-prevent-excessive-concentration-online.

[7] Paul Ohm, *Regulating at Scale*, https://www.georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Ohm-pp-546-56.pdf.

[8] Stuart Kirk, *Artificial intelligence could yet upend the laws of finance*,

 https://www.ft.com/content/8c263c06-fc70-11e7-9b32-d7d59aace167.

[9] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, http://static.googleusercontent.com/media/research.google.com/en/us/archive/papers/detecting-influenza-epidemics.pdf.

[10] Chen Sun, Abhinav Shrivastava, Saurabh Singh, Abhinav Gupta, *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era,* https://arxiv.org/abs/1707.02968.

[11] Adam Grzywaczewski, *Training AI for Self-Driving Vehicles: the Challenge of Scale*, https://devblogs.nvidia.com/training-self-driving-vehicles-challenge-scale/.

[12] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis, *Mastering the game of Go without human knowledge*, https://www.nature.com/articles/nature24270.

[13] Kate O'Keeffe, *Real Prize in Caesars Fight: Data on Players,*https://www.wsj.com/articles/in-caesars-fight-data-on-players-is-real-prize-1426800166.

[14] Russell Walker, *What is Your Data Worth? On LinkedIn, Microsoft, and the Value of User Data*, https://www.kdnuggets.com/2016/06/walker-linkedin-microsoft-value-user-data.html.

[15] James E. Short and Steve Todd, *What's Your Data Worth?*, https://sloanreview.mit.edu/article/whats-your-data-worth/.

[16] Frank Pasquale, *The Dark Market for Personal Data*, https://www.nytimes.com/2014/10/17/opinion/the-dark-market-for-personal-data.html.

[17] Max Chafkin, *Forget Equifax. Facebook and Google Have the Data That Should Worry You*, https://www.bloomberg.com/news/articles/2017-09-11/forget-equifax-facebook-and-google-have-the-data-that-should-worry-you.

[18] Deloitte, *The 2017 Deloitte State of Cognitive Survey*, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-da-2017-deloitte-state-of-cognitive-survey.pdf.

[19] Matt Asay, *85% of big data projects fail, but your developers can help yours succeed*, https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/.

[20] Mike Murphy, *This is how much Google is spending on cutting edge AI research*, https://qz.com/1095833/how-much-googles-deepmind-ai-research-costs-goog/.

[21] Erin Griffith and Tom Simonite, *Facebook's Virtual Assistant M Is Dead. So Are Chatbots*, https://www.wired.com/story/facebooks-virtual-assistant-m-is-dead-so-are-chatbots/.

[22] James Vincent, *Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day*, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[23] Naomi Rovnick, *Higher costs of developing robot warehouses knock Ocado shares*, https://www.ft.com/content/bafc1c00-9d0a-11e7-9a86-4d5a475ba4c5.

[24] Christopher Mims, *Without Humans, Artificial Intelligence Is Still Pretty Stupid*, https://www.wsj.com/articles/without-humans-artificial-intelligence-is-still-pretty-stupid-1510488000.

[25] Rchael Lerman, *T-Mobile gets rid of robot system for customer service calls*, https://www.seattletimes.com/business/technology/t-mobile-gets-rid-of-robot-system-for-customer-service-calls/.

[26] Timothy B. Lee, *Experts say Tesla has repeated car industry mistakes from the 1980s*, https://arstechnica.com/cars/2018/04/experts-say-tesla-has-repeated-car-industry-mistakes-from-the-1980s/.

[27] Mark Bergen and Joshua Brustein, *Google Has Made a Mess of Robotics*, https://www.bloomberg.com/news/articles/2017-10-12/google-has-made-a-mess-of-robotics.

[28] Thomas J. Holmes, David K. Levine, and James A. Schmitz, Jr., *Monopoly and the Incentive to Innovate When*

Adoption Involves Switchover Disruptions, http://users.econ.umn.edu/~holmes/papers/switchover_disrupt.pdf.

[29] Ryan Avent, *The productivity paradox*, https://medium.com/@ryanavent_93844/the-productivity-paradox-aaf05e5e4aad.

[30] Nicholas Bloom, *Corporations In The Age Of Inequality*, https://hbr.org/cover-story/2017/03/corporations-in-the-age-of-inequality.

[31] Patrizio Pagano and Fabiano Schivardi, *Firm size distribution and growth*, https://digilander.libero.it/fschivardi/images/pagaschi5.pdf.

[32] T. Michael Nevens, Gregory L. Summe, and Bro Uttal, *Commercializing Technology: What the Best Companies Do*, https://hbr.org/1990/05/commercializing-technology-what-the-best-companies-do.

[33] Müge Adalet McGowan, Dan Andrews, Chiara Criscuolo, and Giuseppe Nicoletti, *The Future of Productivity*, https://www.oecd.org/eco/OECD-2015-The-future-of-productivity-book.pdf.

[34] Edward P. Lazear, *Entrepreneurship*, http://www2.econ.iastate.edu/classes/econ521/orazem/Papers/Lazear_entrepreneurship.pdf.

[35] Dan Andrews, Giuseppe Nicoletti, and Christina Timiliotis, *Digital technology diffusion A matter of capabilities, incentives or both?*, https://www.oecd-ilibrary.org/economics/digital-technology-diffusion_7c542c16-en.

[36] Ian Barker, *80 percent of IT decision makers say outdated tech is holding them back*, https://betanews.com/2018/07/11/outdated-tech-holds-back-business/.

[37] Dan Andrews, Chiara Criscuolo and Peter N. Gal, Frontier Firms, *Technology Diffusion and Public Policy: Micro Evidence from OECD Countries*, https://www.oecd.org/eco/growth/Frontier-Firms-Technology-Diffusion-and-Public-Policy-Micro-Evidence-from-OECD-Countries.pdf.

[38] Catherine Tucker, *Is Big Data a True Source of Market Power?*, http://www.analysisgroup.com/is-big-data-a-true-source-of-market-power/.

[39] Tom Simonite, *Algorithms That Learn with Less Data Could Expand AI's Power*, https://www.technologyreview.com/s/601551/algorithms-that-learn-with-less-data-could-expand-ais-power/.

[40] Ahmed El Deeb, *What to do with "small" data?*, https://medium.com/rants-on-machine-learning/what-to-do-with-small-data-d253254d1a89.

[41] Sarthak Jain, *NanoNets : How to use Deep Learning when you have Limited Data*, https://medium.com/nanonets/nanonets-how-to-use-deep-learning-when-you-have-limited-data-f68c0b512cab.

[42] Ben Dickson, *Why Tech Companies Are Using Humans to Help AI*, https://www.pcmag.com/commentary/363079/why-tech-companies-are-using-humans-to-help-ai.

[43] John H. Richardson, *AI Chatbots Try To Schedule Meetings—Without Enraging Us*, https://www.wired.com/story/xai-meeting-ai-chatbot/.

[44] Olivia Salon, *The rise of 'pseudo-AI': how tech firms quietly use humans to do bots' work*, https://www.theguardian.com/technology/2018/jul/06/artificial-intelligence-ai-humans-bots-tech-companies.

[45] Geoffrey Manne and Ben Sperry, *Debunking the Myth of a Data Barrier to Entry for Online Services*, https://truthonthemarket.com/2015/03/26/debunking-the-myth-of-a-data-barrier-to-entry-for-online-services/.

[46] Danny Sullivan, *Google: Bing Is Cheating, Copying Our Search Results*, https://searchengineland.com/google-bing-is-cheating-copying-our-search-results-62914.

[47] Steve LeVine, *A resistance against Big Tech*, https://www.axios.com/a-resistance-against-big-tech-1513305929-4641d77e-cec4-411f-96d4-ae9554d5ca00.html.

[48] Subhojit Banerjee, *NO you don't need personal data for personalization*, https://hackernoon.com/no-you-dont-need-personal-data-for-personalization-de9222cff8e4.