DISCUSSION PAPER SERIES

IZA DP No. 9988

Big Data Is a Big Deal But How Much Data Do We Need?

Nikos Askitas

May 2016

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Big Data Is a Big Deal
But How Much Data Do We Need?

Nikos Askitas
IZA

Discussion Paper No. 9988
June 2016

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 9988
June 2016

ABSTRACT

Big Data Is a Big Deal But How Much Data Do We Need?

The more conservative among us believe that “Big Data is a fad that will soon fade out” and they may in fact be partially right. By contrast, others – especially those who dispassionately note that digitization is only now beginning to deliver its payload – may beg to differ. We argue that all things considered, Big Data will likely cease to exist, although this will happen less because it is a fad and more because all data will eventually be Big Data. In this essay, I pose and discuss the question of “how much data do we really need” since everything in life and hence the returns from data increments ought to obey some kind of law of diminishing returns: the more the better, but at some point the gains are not worth the effort or become negative. Accordingly, I discuss small and large, specific and general examples to shed light on this question. I do not exhaustively explore the answers, rather aiming more towards provoking thought among the reader. The main conclusions, nonetheless, are that depending on the use case both a deficit and an abundance of data may be counterproductive, that individuals, data experts, firms or society have different optimization problems whereby nothing will free us from having to reach decisions concerning how much data is enough data and that the greatest challenges that data-intensive societies will face are positive reinforcement, feedback mechanisms and data endogeneity.

JEL Classification: C55

Keywords: Big Data, endogeneity, social science, causality, prediction

Corresponding author:

Nikos Askitas
Institute for the Study of Labor (IZA)
Schaumburg-Lippe-Strasse 5-9
53113 Bonn
Germany
E-mail: askitas@iza.org

1. Introduction

On September 2014, I posted a short LinkedIn[1] piece on Big Data and whether or not more data is always for the better, discussing a handful of examples that I used as metaphors for the problem. The piece was partly inspired by my participation in two multidisciplinary, two-day workshops on *Remembering and Forgetting in the Digital Age* organized by the Research Center for Information Law at the University of St. Gallen.[2] This prompted me to think about this kind of question, unexpectedly combining my mathematical heritage, information technological experience and economics research with Big Data into one single viewpoint, which was - at least for me - intellectually rewarding. I was asked to expand that short piece for a special volume of AStA on Big Data and this essay is the result[3]. It contains my thoughts on Big Data, including how and when more data may not always make us better off. In this essay, I am mostly partial to social science and in fact economics, although much of what I discuss may well be applicable to other empirical, data-intensive scientific areas. I hope that readers can benefit from this essay in a similar manner, using it in combination with their own expertise and experience to complement their understanding of Big Data and how one can optimize data utility as a function of data quantity in a balanced manner that takes into account the interests of individuals, data experts, firms and society at large.

The remainder of this essay is structured as follows. In Section 2 I use examples and metaphors to examine some of the ways in which data utility might depend to its quantity while in Section 3 I discuss digitization and Big Data to highlight some of the ways technological trends might affect data quantity and the utility we derive from measurement. In Section 4 I discuss what in my opinion will be the most severe issue we will be facing in a data-driven world, namely positive reinforcement, and in the last Section 5 I summarize some subjective conclusions from the collection of the data vignettes discussed in this essay.

2. Data Utility As a Function of Data Quantity

I think that none of the readers of AStA would find it alien if I started a piece such as this with a reference to Sir Francis Galton, given that his work still underlies much of the statistical machinery that we employ today. In 1907, he published a short note in Nature (Galton, 1907) titled *Vox Populi*. Having collected the entries rendered on cards of several hundred people in a weight judging competition, he analyzed the guesses and compared them to the true weight of a bull on display, which was the object of the guessing competition. Having proven that *"the vox populi is correct to within 1 per cent of the real value",*[4] he concluded in typical fashion for almost every empirical scientific paper today (the reader might immediately recognize having often concluded empirical work in similar fashion):

---

[1] https://www.linkedin.com/pulse/20140917093801-17164464-big-data-is-a-big-deal-but-how-much-data-do-we-need

[2] http://www.fir.unisg.ch/en/research/remembering+and+forgetting

[3] The special volume is one of many proper and necessary responses to upcoming changes as *"software and algorithms are being developed and the involvement of statisticians in these is essential to ensure that the data that become available retain their integrity and thus usability for statistical analysis"* (Shlomo & Goldstein, 2015). (Shlomo & Goldstein, 2015) (Shlomo & Goldstein, 2015).

[4] The median of the guesses and not the mean - as it is often written today - was shown to be close to the real value.

The authorities of the most important cattle shows might do service to statistics if they made a practice of preserving the sets of cards of this description, that they may obtain on future occasions, and loaned them under proper restrictions, as these have been, for statistical discussion.

Two remarks hold particular interest for the present essay and in fact for our present data reality. First, note how Galton as a data worker answers the question of *"how much data do we need"* as every data expert ought to do, namely by stating - in no uncertain terms - that *"we need more data"*. Knowing the pitfalls, caveats, nuances and shortcomings that might undermine "the story we tell" in empirical research, we all inevitably and invariably end our papers expressing our need for more data. Second, he uses the humble task of ox weight-guessing as a metaphor to prove that his result *"is more creditable to the trustworthiness of a democratic judgment than might have been expected"*.
Leaving aside the reasons why Galton found the accuracy of democratic judgment to be surprising, this second remark is perhaps the one that bestows the most merit on his appeal for more data, reflecting the best argument of every data expert today regardless of their scientific domain. Given that he did not and could not have real measurements of trustworthiness of democratic judgment on political issues, he found a situation that could proxy it in a measurable way. In other words, he substituted a problem of high scientific, political and societal interest for which he had no data with the study of a process with data readily available that had certain similarities to the original problem. In Galton's words, *"the average competitor was probably as well fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues on which he votes, and the variety among the voters to judge justly was probably much the same in either case."*

Whether or not we agree with this assessment may be a matter of debate, but we all know that such an approach is both typical and frequent in social science and economics in particular. Very often we have measurements about things that we do not necessarily care about while lacking data on vitally important questions. We subsequently often seek to argue that using the data that we have can give us results on the things that we care about in indirect ways. We use scientific intuition to argue why a certain measurement is a good proxy for a variable that is important. The practice of doing so - which comes from necessity - introduces both ingenuity in social science[5] as well as uncertainty regarding the validity of its results. The more often we must analyze variables that measure proxies rather than the actual quantities of interest, the more merit our appeal for more data gains. There are many reasons why this practice is prevalent, most of which are impossible to eradicate and related to the fact that when profound changes in socioeconomic life are in the making, the recording - by socioeconomic agents - of data measuring them understandably occupies low priority.

In economics as well as business, it is very often the case that we are forced to either work with proxies in the absence of fit-for-purpose measurements or use preliminary estimates based on proxies, in an effort to remedy our inability to produce the necessary

[5] (Schumpeter, 1961) provides a prime example of that ingenuity. In the section called the "Semeiology of Daily Life", Schumpeter proposes the participation in Sunday mass as a (countercyclical) leading indicator of the business cycle.

data in a timely manner[6]. Measuring GDP is the simplest example to demonstrate this. In order for governments to plan and submit timely annual budgets, we need to have a timely GDP measurement, which we are hence forced to estimate since we cannot measure it quickly enough. This is surprisingly the case even today, with digitization in full swing. Not only are we still unable to gather the data necessary to report GDP accurately and timely despite digitization, but we are also faced with the realization that what we are measuring might no longer correctly identify what we need to measure, if it ever did. This is the case with the so-called observed *"puzzle of productivity slow down"*,[7] to mention one of the most prominent examples. The fact that subsequent to our first estimates we need to revise them over several years means that most governments plan budgets based on incorrect or insufficient data. Of course, one could be tempted to argue that since government works the missing data nonetheless does no harm, although in the absence of counterfactuals this would be a premature and naïve conclusion.

There are many examples - and we will discuss some of them in this essay – showing that both data deficit and data overabundance may have both positive and negative effects. Such variations in data usefulness may occur along many dimensions that characterize data. It is hence along these dimensions that we need to discuss both data deficits and their overabundance as well as the positive or negative effect that these can have for the various stakeholders, including citizens, science, society, policy-makers, etc. Such parameters may refer to data itself and include such elements as data quality, quantity, accuracy, identification strategy, fit-for-purpose, frequency and reliability, to name a few. Other parameters relate with data as a commodity of sorts and include cost/benefit comparisons, price, timeliness of availability, interference with the process measured, availability for re-use, implications for individuals' privacy and introduced endogeneity[8].

We can easily imagine many situations in which we would respond differently if we vary the data on hand. Showing a hypochondriac his or her blood pressure in real time is probably more data than would be good for him. A student who is taking an exam would do well to moderate the number of times that he thinks about the number of questions answered and the time left to answer the unanswered ones or risk panicking or wasting time. Measuring the temperature of a roast frequently allows a cook to have better control of the optimal roasting point, while overdoing it will probably spoil the roast. On the other hand, measuring the temperature in the core of a nuclear power plant at high frequencies is vital for safety and in fact an absolute necessity. In all such cases, more frequent data may sometimes overwhelm the participants of the process that we are measuring or affect the process itself, albeit the absence of data can also be detrimental.

---

[6] Technological progress and falling prices of computing components together with the inherent ability of digitization to log itself in real time might give us the ability to save data pre-emptively whereby we might rewind and replay recorded reality to isolate the right measurements. My personal feeling, however, is that while this certainly will be true, we may never rid ourselves of the need to use proxies for what we really care about.

[7] http://blogs.wsj.com/cfo/2015/07/17/the-morning-ledger-does-the-classic-measure-of-productivity-fail-in-the-modern-economy/

[8] The reader may want to consult (Kenett & Shmueli, 2014) where InfoQ is defined and discussed. It is a synthetic measure of quality, which synthesizes the qualities of data, and analysis along similar dimensions as discussed here.

We could conclude that depending on the use case there is an optimum amount of data - which may be different for the data expert, the socioeconomic agent or society - that fits the purpose and hence finding it is an interesting problem to discuss and always keep in mind.

If we stand too close to an impressionistic painting, we only see random brush strokes, which amount to little more than visual noise. It is only when we gain some distance from it that the beauty and accuracy of the picture emerges and if we squint with varying intensity we can tune in and out of the picture, allowing us to focus on a spectrum of different aspects of the painting. Accordingly, in this case it is only after dropping some data that we can see meaning. On the other hand, if we look at Persian art (the kind on the beautiful rags) from too far away, it looks like a bad mix of unfinished smudges. Only from up close do we see the intricate and beautiful geometry of the rag with its symmetries, tilings and repetitions. Here, it is only after we come up close that we start to see the picture.

The trick once again is to forget and remember data at just the right proportions for the right problem. However, some further remarks illustrate how tricky the problem of data is. One may argue that a high quality photograph renders an impressionistic painting obsolete, with the latter little more than a low frequency rendering of "imperfect" brush strokes. Of course, to say this would ignore the fact that filling in the gaps that an impressionistic painter "neglected" to paint allows us a superior experience compared to viewing a photo. This is a case where less data affords us the ability to fill in the gaps in more creative ways, which even in science may lead to better overall knowledge than otherwise. In fact, this example is a metaphor for the scientific process itself.

Taking the tangent space of a differentiable geometric object or approximating a function by means of low power terms in a Taylor series are ways by which we reduce measurement not only without essentially harming the science we perform but - quite to the contrary - making it possible in the first place. Without the method of linearization and approximation, we could never use computing to aid us in mathematics, physics or economics and in fact most of the science that we now command would have remained intractable. Finally, digitization itself - by all accounts a catalytic development in matters of data - is at its core nothing more than an appropriate reduction of data compared to analog technologies of the past. The fact that we have found ways to replace the continuum of reality with a digital - i.e. discrete - representation has allowed us to make progress in strides. One might dare to state that it is only through data reduction that we can increase it.

Driving provides a further illustration of the complexities of data availability. The reason why driving at 200 km/h is possible is that we teach ourselves to feel comfortable ignoring most of the visual information coming our way, a fact that haunts every new driver and is by all accounts one of the hardest things to overcome when learning how to drive. Our visual system is so overwhelmed by the speed at which visual data reaches us that we must ignore most of it, focusing on a far-away point instead and essentially ignoring most of the rest that is passing us by. Any attempt to try to see every detail of the oncoming scenery is not only impossible but also dangerous. In other words, safe driving relates to dropping information or put differently opting for less data. However, while we are dropping data during the act of driving, we may need to mount a high-speed camera on the dashboard of our car that collects as much visual data as possible for the purpose of documenting and reviewing the drive at a later point in time. If the recording device is too large, it may interfere with the driving process, whereas if it is seamlessly embedded in the vehicle then we can preemptively record reality without affecting it, thus giving us the best of both worlds.

This example illustrates the core problem with data and measurement, namely that once reality has passed by we can no longer measure it, whereas if we are measuring too much of what we experience we might interfere with the experience itself. These are the boundaries of our data optimization problem, since we can never measure and save all the data that we need to make a choice. Doing so exposes us to the risk of having recorded the wrong variables. By the time that we adjust to account for our shortcoming or technological progress, we introduce data discontinuities, which obstruct scientific progress. Therefore, if we are asking how much data to save, the answer is that we need to save as much as we possibly can without obstructing or interfering with the process upon which we are keeping tabs. Data is unlike salt in a soup: more data can be reduced to less (i.e. micro-data can be aggregated), although the opposite is impossible. We can also clearly see the unavoidable impact of progress in digital technology on the socioeconomic progress. The less intrusive that technology becomes (smaller, powered over the air, embedded, etc.), the more data we can collect without affecting reality. We will of course see that contrary to the fact that measurement will benefit from technological progress the real time depiction and use of the data thus gathered might not be as neutral when it comes to whether or not it affects the socioeconomic process.

Comparing still photographs to film is also a pedagogically useful example to keep in mind when we think about data. Taking infrequent snapshots (measurements) of reality restricts our ability to go back and study the past (i.e. an event of scientific interest), whereas filming it allows us to rewind and replay the past, affording us revised snapshots when we need them. Therefore, clearly, if we can afford it, we should always choose filming rather than taking snapshots. On the other hand, we can trick our eyes into believing that we are watching moving pictures - i.e. a film - if we supply them with at least 25 pictures per second. Accordingly, there is our optimization problem once again. Going from low frequency of snapshots to higher frequencies improves our ability to view motion, although we get diminishing returns like with everything in life: 25 frames per second is the first local maximum.

The efficient market hypothesis (EMH) tells us that tomorrow's prices will only depend on tomorrow's news because today's news has already been taken into account. Since the news of tomorrow is by its very definition unknown and impossible to forecast, it follows that you cannot "beat the market", i.e. there is no algorithm by which you can build a portfolio that will perform better than the long term trend of the mean. The EMH has been known to be – morally - "eventually correct" even if certain short-term fads occasionally intrude. In fact, if you wish to obtain a series of pure noise, you might as well look at the differences of logs of high frequency financial prices. Two remarks on this hold interest to our matter at hand. First, the fact that these series are pure noise does not mean that they do not contain any information, but in fact quite the opposite: they contain **all information** that is known. From the perspective of complexity theory, pure noise and the aggregate average of all signals are indistinguishable from each other, providing us with yet another way to say that more data need not be better for us and that the law of diminishing returns applies. Second - and put simply - most news come from the past. This statement is both trivial and profound yet discussing it holds some value and we do so by way of an example. On September 18 2015, the US Environmental Protection Agency (EPA) found the German automaker Volkswagen to be in violation of the Clean Air Act. It was found that the car manufacturer had intentionally programmed certain diesel engines to apply artificial emissions controls during standardized emissions testing. As it emerged, the violations had been in the making since 2009. In other words, while the news of Volkswagen's wrongdoing came out in September of 2015, the news was about a fact that had been in the making for six years. This is what we mean by saying that most news comes from the past. The markets are efficient with the news, although

the news - i.e. data availability - is not. When properly adjusted by using a more pragmatic definition of news (which incorporates data lags), the EMH implies that we live most of our socioeconomic life on insufficient data. This implies that we can survive with data deficits although it leaves the question of what might have been begging for an answer.

3. Big Data and Digitization

People mean different things when they discuss Big Data. A computer expert usually means large amounts of it, such that they challenge current limits of storing and retrieving, as do the various hard- and software developers and manufacturers. A statistician might simply mean multivariate micro-data, which might come from Genetics, Astronomy, Physics, Economics, etc., whereby the more variables per object of study, the bigger the data. Some speak of Big Data in the absence of sampling, i.e. whenever we have measurements for the entire observational universe, while others speak of Big Data when it is messy rather than properly curated based on a certain methodology. Yet another way to define Big Data is by means of "an increasing number of V's" (Hitzler & Janowicz, 2013). This definition requires data to possess *"volume, velocity, variety, value and veracity"*. The reader may indulge in an exercise of assigning V's to use cases: an academic researcher may profit from any of the V's but not without *veracity*, for a business application *value* may be a sine qua non, for a central[9] bank, a policy-maker or the government in general *velocity* may be as important as *veracity*, while for a hardware manufacturer or computer scientist *volume* may be most important. Clearly all of these different definitions are covariates of each other as Big Data is the natural outgrowth of progress in ICT and the resulting digitization.

Most people site "Moore's law" (Moore, 1965) as a proxy for ICT progress, although in our context the real hero is the price of storage[10]: Since the introduction of the first hard drive by IBM in 1956, the capacity of disk drives has doubled every year and by 2011 one cent of one $US could buy 285,000 times more disk space than the first IBM disk drive offered. Highly available disk space can be purchased today for less than 10 cents per Gigabyte per year and it is often offered for free (at least in monetary terms). Random Access Memory has also improved, whereby saving large amounts of data is not only possible but also computationally useful as we have the computational speed, the storage capacity and the processing memory to manipulate that data and extract information out of it.

The progress in computing capacity goes in tandem with other socio-technological developments. We carry smartphones, which are powerful, networked computers equipped with sensors and able to transmit measurements on our location and state. We search for and retrieve digital objects from the net using Google, revealing (to Google) information about what is on our mind, we store and exchange intimate information about ourselves on Facebook and other social media of all kinds and we store our email correspondence on central, third party data silos[11]. The projected onslaught of the "Internet of Things" will intensify these trends.

---

[9] https://www.bis.org/ifc/publ/ifc-report-bigdata.pdf

[10] http://www.oecd.org/internet/ieconomy/50305352.pdf

[11] Data privacy will not directly be my subject matter in this paper, although it is an important one that cannot be neglected and warrants at least a small remark here. In the currently prevalent business model among social media companies, we are either paying customers whose privacy is covered by means of a contract or we can use their resources free-of-charge so long as we allow them to index our data and monetize it in

In a nutshell, these are some of the elements that constitute the backdrop against which we live our lives and this is a sketch of the large picture in which Big Data is being created. It can - but need not be - large in size, it can be multivariate and it no longer need be a small sample of something. It may have value for one use case but not for the other, it may have this or that degree of veracity and it may or may not have velocity. Finally, it may be too big to handle because no matter what the technological capacity is, there will always be such a thing as too much data, so much that we cannot deal with it. Whenever Big Data is big because it challenges current capacities, it drives technological innovation, which benefits applications of Big Data in the social science sense. Whether this cycle is virtuous or not is an open-ended matter.

There are three main sociological observations that allow us to more systematically summarize current trends. The first is that we have a transition from a content production model of one-to-many to a model of many-to-many. When production costs were high, a small number of experts produced audiovisual, intellectual content of all kinds, which was consumed by the large public. Now we have a more symmetric picture, one in which every content consumer is also a content producer. In summary, one might say that the democratization of content production is the main result of current socio-technological trends. One might comment negatively on the quality of such content but a dispassionate social scientist might simply see the opportunities arising. The big players in the race for data certainly see the commercial value of this fact when they build and offer storage and sharing platforms for free.

The second observation to make is about the impact that Big Data has on the way in which we go about working with data and conducting research with it. Most of our current statistical toolbox in social science is tailored to a past era of computing scarcity. Since we could neither store nor process large amounts of data, we developed techniques that could infer knowledge about the whole from a properly selected small and manageable portion of it: in other words, we developed sampling. In this way of working, we used theory to formulate our working hypothesis and subsequently crafted the questions that we thought captured the variables in which we are interested and devised a concept of random selection that later allowed us to control the error in extrapolating from the sample to the whole. This was an overall cumbersome - and often expensive - way to work even if it were an optimization problem imposed to us by real constraints. While much of this statistical machinery can and will eventually be ported to the Big Data world, we currently remain in the beginning stages of finding out how to do so.

Finally, we observe that we now have a plethora of digital, internet-based markets (Askitas & Zimmermann, 2015). Whether the marriage market or the transportation market or the market for various assets, etc., there is hardly one that does not have an online component, many are mostly digital by now and some exist only in digital form. With logging naturally built into ICT, the result is detailed, non-sampled, unsolicited data on entire markets and niches.

To recap this brief discussion, Big Data is the natural outgrowth of digitization and it is not a purely quantitative matter. By increasing the scale, volume, diversity, frequency, sampling intervals and scope of data and being able to merge disparate data, we change the quality of the game in a more profound way than simply size; rather, we change the very foundation of

---

other ways, such as targeted advertising. In the latter case, we cease to be their customer and we become their product. This trade-off implicitly says that we have placed a monetary price tag on our privacy.

socioeconomic life and of course social science as well as the research paradigm and hence consequently the current statistical toolkit.

4. Externalities, Positive Reinforcement and Endogeneity

As part of the social program of a workshop concerning the impact of social media on the small and medium firm, the event organizer offered us a wine-tasting event. A wine master would explain to us what to look for in a wine (look, smell, taste), offering us three wines to score on a scale from 1 to 10, where 10 is the best wine. We would taste the first and subsequently after writing down our score we would compare ours with that of the wine master. We were told that this was so we could all calibrate our taste buds on the same scale. After that, we would receive and evaluate our two wines. I proposed and organized a split. Of the 100 participants, 50 would score both wines in private and file their vote on a piece of paper, much like in the case of Galton, whereas the other 50 would vote sequentially in public view. We would plot the results side by side on two large charts that had the score on the x-axis, by gluing points so that that heaps of points would be formed above the score axis.

The results were interesting and can teach us a lot about data endogeneity. Both voter processes had about the same mean between 6-7, which to me was a result of the fact that - being polite – people did not want to rate the wines as too bad whereas they also did not want to rate them too high, fearing that they may reveal themselves to be naïve wine tasters. The striking difference related to the standard deviation of the two votes, given that the private voting had a much higher standard deviation than the public one. Accordingly, when people voted so that they could see the formation of the mean in real time they tended to follow each other using it as a cue for how to avoid a bad vote. To me, this example shows that living with data makes data a part of strategic behavior and changes the very data that we are collecting.

In what is known as the Fundamental Diagram of Traffic Flow, the important variables to keep track of are flux (vehicles per hour), traffic density (vehicles per kilometer) and vehicle velocity. When traffic density exceeds a certain threshold, the emergence of what are known as phantom traffic jams becomes increasingly likely (Kurtze & Hong, 1995). The reason for this is that a sudden speed reduction of any one driver will be noticed by the immediate neighbor in traffic and translated into a disproportionate speed reduction all the way along a chain of slowing vehicles until a complete stop is reached. Restoring traffic to its prior state is a time-consuming process that heavily costs society and individuals and leaves us puzzled that when motion finally resumes we never see the reason why the traffic halt occurred in the first place. This example is typical of the fact that actor density results in high likelihood of externalities, with the channel enabling it being that density increases the likelihood that a signal sent will also be received. In other words, density makes the use of more data more likely.

In social or collective contexts, everything we do has an effect on others and whenever signaling (i.e. data) is high frequency and lossless, dynamics may be "mysterious". When we offer a driver in front of us the right of way, we do so at the cost of every other driver lined up behind us, when we stand in an escalator everybody else must also stand and when a country runs current account surpluses it does so at the "expense" of other countries, which must run deficits. Indeed, there is an endless list of examples. In an urban and hyper-connected landscape, the externalities that we impose on others with everything that we do are more frequent. As soon as we enter data into the picture, the

situation becomes worse. In the digital age with social media and Big Data, in my opinion, this is the greatest problem facing us.

Market dynamics will rapidly allow us to make increasingly more data available, in real time and to more people, which means that more of the subjects whose socioeconomic behavior is reflected in the data that we use to make sense of the world will be aware of increasingly more of their own data as they are contributing to it. The aggregate effect of this fact will be both interesting and challenging for data work and "mechanism design" and may have adverse and unexpected consequences by introducing "Heisenbergian uncertainty" into the measurement process.

In 2012, Bettina Wulff - wife of the then German President Christian Wulff - filed a lawsuit against Google because the auto-completion function of Google search - a feature that Google introduced in 2010 and to which Googlers also refer as the *"psychic feature"* - was autocompleting searches for her name with terms that she found unfit and defamatory. There had been rumors that she had worked for an escort service prior to marrying the president and of course people were searching for such terms as *"Bettina Wulff escort"*. Google's auto-completion algorithm - which suggests the most popular ways in which a certain search term might end - was hence "auto-suggesting" these terms because they had become the most popular completions of how a search term might end if it started with the name of the President's wife. Bettina Wulff and Google reached an out-of-court settlement, which included Google dropping the offending terms from auto-completion, thus violating its own data recipe.
The relevance of this incident to our topic is that it provides us with a real-life incident of positive reinforcement. It exemplifies the typical way by which we might expect that what we may call "quantum effects" might arise in a data-intensive society. It may well be argued - and this could certainly explain Google's capitulation - that the act of automatically calculating and depicting in real time the most popular ways in which certain sentences end when they start in certain ways is not merely an objective depiction of a measurement of reality. It becomes an active part of that reality and shapes it due to positive reinforcement in much the same way that the real-time depiction of blood pressure might worsen the blood pressure of a hypochondriac. For example, the depiction of a top ten list can never be a neutral act as it may contribute towards establishing which item resides in the top ten list itself. We are entering the area of data endogeneity due to data overabundance and overreliance. The general fact that encapsulates this discussion is the observed winner-takes-all phenomena in socioeconomic life, long tail distributions and the prevalence of the Pareto "80-20 rule".

When Toronto traffic lights were equipped with a countdown system visible to both pedestrians and drivers (i.e. the amount of data available to the road users was increased), more accidents were observed because the pedestrians' crossings became riskier and drivers pressed for time were bumping onto prudent drivers in front of them (The murkier side of Transparency[12]).

In a world were data is being generated, reshaped and depicted in real time, endogeneity and data externalities will become part of our methodological challenges. Both the utility of collecting and using data as well as the statistical toolkit we use to evaluate the data that we collect may come into question. Put simply, in a world of positive feedback, persistent

[12] http://www.ft.com/intl/cms/s/2/785bd614-9378-11e3-b07c-00144feab7de.html

loopback effects and data endogeneity, writing *"i.i.d."* and assuming *"independent identical distribution"* of random variables might become increasingly unrealistic.

Market forces will more likely shape the race to more data. The market will leverage progress in ICT to collect and use data and by doing so it will drive ICT developments by challenging current limits. This autocatalytic process will drive developments in the immediate future, unavoidably affecting economics and social science, as we know it. We are currently already witnessing two different trends: on the one hand, business applications are driven by profit and hence care more about prediction (value) than causality (veracity) compared to academic economics, for instance; and on the other hand there are already a number of areas - even in academic research - where it is identified that prediction suffices for a large class of problems for policy purposes and causality does not have to be tackled to make useful and meaningful assertions (Kleinberg, Ludwig, & Mullainathan, 2015).

Speaking at a Caltech graduation ceremony, Richard Feynman asked students to imagine "how much harder physics would be if electrons had feelings". Regarding the data situation, we might in fact have to deal with just that. Data science might eventually need to become physics with particles that have feelings and are determined by externalities and endogeneity. Data endogeneity and adaptive behaviors are certainly not unknown to economists as Keynes' beauty contest suggests (Keynes, 2006). Shiller (2015) links the emergence of the first bubbles in speculative assets to the appearance of newspapers. Moreover, the spread of telephony lead to the appearance of the boiler rooms and the stock market instability of the 1920s, whereas the appearance of the Internet is linked to the collapse of the so-called New Economy in 2002 and social media played a role in the US housing market bubble burst in 2006. All of these instances can be viewed in our context as market endogeneity introduced by data publication, use and overreaction to.

5. Conclusions

We have used several data vignettes to show that data - like everything else in life - obeys the law of diminishing returns. While the marginal utility from an extra unit of data created is influenced by the economics of current technology, the diminishing of these returns when we make such data readily available might not necessarily be linear due to the ensuing endogeneity.

The answer to the question of how much data we really need depends on the use case, as well as who is asking the question. The individual might care about privacy and hence less of it, the data expert might care about the veracity of science and hence for more of it, the government might be hindered by regulation and firms might gain the upper hand in data ownership driven by profit, while both policy-makers and firms might care less about causality and more about prediction. Market forces in a general sense will probably shape the answers for the various stakeholders. As market competition will probably make entering the race for more data more a matter of survival than choice, working out the answers might not happen without undesired side effects. Expert vigilance will be necessary to guide us through. Such vigilance ought to be neither ignorant of this development nor overly enthusiastic because of it and ought to recognize the rare historical moment we live in but also the fact that we are leaving in some sense a very natural historical evolution: what might have worked well (normality , linearity, independence etc.) may need revisions and what did not work well or was rarely applicable might need to be reexamined (endogeneity, positive reinforcement, externalities, etc.).

Literaturverzeichnis

Askitas, N., & Zimmermann, K. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower , 36* (1), S. 2-12.
Galton, F. (1907). Vox populi (The wisdom of crowds). . *Nature , 75*, S. 450-451.
Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web , 4* (3), S. 233-235.
Keynes, J. (2006). *General theory of employment, interest and money.* Atlantic Publishers & Dist.
Kleinberg, J., Ludwig, J., & Mullainathan, S. (2015). Prediction policy problems. *The American Economic Review, , 105* (5), S. 491-495.
Kurtze, D., & Hong, D. (1995). Traffic jams, granular flow, and soliton selection. *Physical Review E , 52* (218).
Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine* , 82-84.
Schumpeter, J. (1961). *Konjunkturzyklen : eine theoretische, historische und statistische Analyse des kapitalistischen Prozesses.* Göttingen: Vandenhoeck & Ruprecht .
Shiller, R. (2015). *Irrational Exuberance.* Princeton and Oxford: Princeton University Press.
Shlomo, N., & Goldstein, H. (2015). Editorial: Big data in social research. *J. R. Statist. Soc. A , 4* (178), S. 787–790.