Political Science Is A Data Science

SPSA Presidential Address, January 10, 2020

San Juan, Puerto Rico

Jeff Gill

Distinguished Professor

Department of Government

Department of Mathematics & Statistics

American University

jgill@american.edu

Introduction

This discussion is about changes in the world, changes that affect us as academics, changes that affect us as empirical researchers (qualitative and quantitative!), and changes that affect our students and our universities/colleges. I am an empirical social scientist, a political methodologist, and a statistician, so I will only discuss topics along these lines where I am qualified to make comments. As of this re-rewriting the country and the world is in various levels of lock-down and recovery due to Covid-19. During this difficult process it is clear that the data and privacy issues are changing rapidly. There is evidence that some governments are using the pandemic to extend their reach into data spheres that were more difficult to monitor and control before. Hopefully this is a temporary trend and people around the world can keep some control over their data as challenging world events unfold.

Broad Perspectives

We live in the *Data Century*, *whether we like it or not*. Our personal lives, our careers, our finances, our social activities, our childrens' lives, and our future prospects are all intertwined and affected by data collection, data storage, and data analysis by others (humans and machines), *whether we like it or not*. Governments have essentially lost control over this process, *whether we like it or not*. Personal education in data science, big data, statistical analysis, and data privacy is essential for people to exert some control and influence over their data future, *whether we like it or not*.

Homo sapiens are only about 200,000 years old (depending on what anthropologist you choose to talk to), whereas the earth is 4.54 billion years old. Of course we are much older than 200,000 years if count prior biological forms that we closely resemble. But the pace of change in human development has been astronomically faster in the last 100 years. Part of this is because humans now have more time to "do stuff" since 30+ years were added to our average life expectancy in the 20th century. This is remarkable if you think about it: we existed with short life spans for many tens of thousands of years and then about a few thousand years ago many humans lived until about what we call early middle-age but just recently we added an average of 30 years suddenly on top of that.

Relatedly we are now in the early-middle part of the fifth major revolution in human history: the Upper Palaeolithic revolution (about 40,000 years ago) $\longrightarrow$ the first agricultural/Neolithic revolution (about 12,000 years ago) $\longrightarrow$ the second agricultural revolution (18th century) $\longrightarrow$ the industrial revolution (1712 to the early 20th century) $\longrightarrow$ the information revolution (early 21st century onwards)

$\longrightarrow$????. This next revolution may be the genetic manipulation revolution, the space exploration revolution, or perhaps just surviving as a species in a newly dangerous era of planetary change and rampant disease propagation. It is interesting that people are typically not aware of being in a current ongoing revolution, hence this discussion. The bar is even lower than that: only a few elites were self-aware of their status during the Renaissance and the Enlightenment. Throughout history providers of low-level manual labor were rarely aware of the scope geopolitical and industrial change since their main purpose was personal survival and the survival of their immediate family.

We are now changing our environments, structures, institutions, and work-lives faster than ever before. But this change is uneven. The 20th century was the century of remarkable progress in physics, chemistry, and engineering. Achievements included: discovery of the atomic nucleus, quantum mechanics, discovery of the absolute geological timescale, general theory of relativity, discovery of antimatter, X-ray technology, evidence that Earth has a core, radio-astronomy, the development of the electrical grid, the gas turbine jet engine, rockets, radiocarbon dating, superconductivity, discovery of cosmic microwave background radiation, complete description of the human genome, discovery of exoplanets, cloning of a mammal, manned rocketry, landing men on the moon and returning them to earth, the fully digital computer, radar, sonar, nuclear power, commercial air travel, passenger automobiles, synthetic rubber, television, fiber optics, telephony, the electron microscope, atomic and nuclear weapons, laser technology, chromatography, satellites, the discovery of black holes, magnetic resonance imaging, robots, frozen food, the helicopter, air conditioning, motion pictures and digital video, microchips, mobile phones, the Hubble telescope, the Internet(!), artificial intelligence, solar cells, and more of course. In contrast, the 21st Century will be the era of monumental intellectual progress in the social and biomedical sciences. The key to research in these areas will be: digital computation, data analysis, infrastructure supporting the entire life-cycle of collecting and processing gigantic amounts of information, and the use of networked connections of information from diverse sources. It turns out that studying humans (in almost any way) is just more complicated and needs more advanced computing. We sent men to the moon with a fraction of the computing power (reportedly 2 PlayStation's worth) of the machine on my desk that (still) runs complex statistical simulations over several days for each problem. It is hard to overstate the need for advanced computational resources when attempting to understand people and how they interact with each other.

Data access and data analysis will play an indispensable part in progress to understand social, psychological, and physiological characteristics of what it means to be human. Integration of disparate

data resources will be essential to research and commercialization. Long term preservation of data involves technical challenges and new business models. The future of data analysis and provision in the social and biomedical sciences data is not going be strictly in rectangular data files, data dictionaries, and PDF codebooks. These corresponding fields have moved to new and diverse data-types: genetic/genomic, digital video, geocoding/GIS, high-resolution still imaging, high-frequency sensor data, Internet traffic, mobile phone tracing, detailed personal information, unstructured text, and more. These fields have moved to new *sources* of data: social networking and media, human physically generated, large-scale government administrative records, transactional financial information, and electronic human physical monitoring data. Note that these are both qualitative and quantitative forms. Such data require completely new documentation and archiving standards. In addition, there are important privacy/confidentiality, anonymity, government, civil law, and regulatory issues.

Universities can be positioned to be a leader in this new data revolution, or they can be left behind. This is why many universities are creating data science degrees and data science institutions (including *American University* of course). However, a data science program does not have a uniformly agreed-upon definition, so the curricular heterogeneity in graduate and undergraduate training is substantial and interesting. A key difference is in the level of computer science knowledge required on entry or covered in the course-work. Relatedly, publishers are also at a critical junction in how journals are to be delivered to university libraries. When will paper cease to be the most important distribution medium of intellectual content? When will readers demand interactive documents? When will there be a standard electronic document for academic journals? The interrelated changes in data science and electronic communication are central to how universities are going to operate in the 21st century.

Data by the Numbers

To put some context on the pace of change and the volume of data that are generated, consider the following summary statistics currently (2020) for *every single day*:

- 23 billion text messages are sent
- 5 billion searches are made (40,000 per second on google alone)
- 500 million tweets are sent
- 294 billion emails are sent

- 4 petabytes of data are created on Facebook
- 4 terabytes of data are created from each connected car
- 65 billion messages are sent on WhatsApp
- 360 terabytes are uploaded to YouTube
- 21.6 million GIFs are sent via Facebook messenger
- 149 billion spam emails are sent
- 222 million calls placed on Skype
- Venmo processes $75M peer-to-peer transactions
- The Weather Channel receives 2.6×10^{10} forecast requests
- 65M Uber bookings
- The average online person generates 10^{18} bytes of data
- The CERN Large Hadron Collider generates 864 zettabytes of data.

Of course these numbers are only estimates and have already changed between the time of writing this and someone reading. Perhaps more interesting politically, Twitter reported that on the evening of November 8th, 2016 over 75M Tweets were sent (still a record for one topic). Furthermore human civilization is moving up the data volume scale at an accelerating rate. These numbers look like this...

Abbrev.	Unit	Value	Byte Size
b	bit	0/1	1/8 of a byte
B	bytes	8 bits	1 byte
KB	kilobytes	$1,000$ bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes
BB	brontobyte	$1,000^9$ bytes	1,000,000,000,000,000,000,000,000,000 bytes
gB	geopbyte	$1,000^{10}$ bytes	1,000,000,000,000,000,000,000,000,000,000 bytes
ZB	zotzabyte	$1,000^{11}$ bytes	1,000,000,000,000,000,000,000,000,000,000,000 bytes
CB	chamsbyte	$1,000^{12}$ bytes	1,000,000,000,000,000,000,000,000,000,000,000,000 bytes

Humankind is steadily progressing up this ladder requiring more computational resources at a rapid rate. While the price of storage continues to plummet, demand by institutional repositories is outstripping this decline.

What Is Big Data?

Basically what anyone wants it to be. The classic definition is: volume, variety, velocity, value, and veracity. Meaning big data is defined by its size, its complexity of different types, how fast it comes, how much is it worth, and how reliable it may be. This mnemonic does not mean that something called big data has to meet all of these criteria but it has to meet some of them for people to care. My definition is more general: data that are large enough to challenge available computational resources. By this definition self-aware humans have always been in a "big data era". The abacus was invented thousands of years go because commercial transactions involved big data of the time. The current digital universe stored is estimated to be at least 44 zettabytes, and a large proportion of this would fit any definition of big data. Furthermore, sometime before the year 2025, we believe that 463 exabytes of stored data will be created every day. However, right now less than 1% of all generated and stored data are being

analyzed and this number is actually going *down* because the average size of datasets is going up. So an important question is what are existing tools to deal with these data-size challenges?

What is Machine Learning?

One answer is that it is a simple classifier. It is actually just statistics with an emphasis on prediction and accuracy. Machine learning, and all of its sub-types, are useful for analyzing big data because the algorithms perform work that would otherwise fall to human effort. Essentially there are four basic tools under this umbrella: Support Vector Machines, Random Forests, Neural Networks (in countless variations now, where the name comes from resembling how the neuro-cranial system works), and Logit(!). Deep learning algorithms are emblematic of these approaches as they establish initial parameters from the data and then train the computer to learn independently by recognizing data patterns using multiple layers of processing with direct human intervention. Of course machine learning is most effective when automated with *many* hopefully reliable examples to adapt to tasks independently, which is not what political scientists typically use it for because our data are often more limited than types employed by large commercial enterprises. For example credit card companies have an immense amount of spending information that can be analyzed on its own or associated with what they know about us individually. However, political scientists are now widely using these tools: we *are already* data scientists.

Privacy (or lack thereof)

The explosion of digital sensors, Internet of Things (IoT), smartphone apps, has serious and long-lasting consequences. Alexa is spying on you. Google is spying on you. The government is spying on you (online digital fingerprinting, etc.). Your phone is spying on you. If your car is recently manufactured it is spying on you. Your rental car company is spying on you. Your hotel is spying on you. Airbnb hosts are spying on you. And even more organizations are spying on you! For example, every time Amazon's Alexa AI activates on your wake command it keeps a recording of everything said in the room during operation "to improve our algorithms" (read the fine print sometime; it's scary). Substantially reduced costs for storage drives means that corporations and governments save more process traces, network logs, domain specific data, and geospatial data than ever before. This means that machine learning algorithms (generally speaking) can associate individual data across disparate data sources to search for particular behavior. A New York Times Magazine article series the week of December 16, 2019 showed

how we are all tracked by our phones and these go into commercial and government databases forever. They are also available for purchase by just about anybody, including foreign governments.

Data Science for Global Mischief

I will not comment much on this since everybody here reads the news. Except to say that it is naïve to believe that there are governments who do *not* practice it. And never mind the tens of thousands of non-governmental nefarious organizations involved. This is where it is unfortunate that most data science tools are free or easily purchased. These methods can be split into traditional hacking tools, some of which are reasonably clandestine but available to agencies of governments and some of which are publicly available, and standard data science tools and technologies that are clearly amenable for nefarious purposes. For instance, web scraping and text analysis tools can be used to gather and process topics that aid in the spread of disinformation. This is a fast moving and highly dynamic area and any comments here of a more specific nature would be quickly outdated.

Specific Trends to Pay Attention To

There are some interesting areas of data science that are becoming increasingly important to large segments of the population.

Blockchain cannot be ignored. This is a highly secured ledger that tracks and archives P2P transactions including bitcoin, but is also widely used by other electronic currencies, the US government, banks, online retailers, and others. The name is quite accurate in that it refers to blocks of transmission information in a chain of communication that have three components:

- The block contains the date, time, exchange information (money, communication, etc.).
- The block records participants in the exchange with unique digital signatures.
- Blocks are unique, chained, permanent, unalterable, and encrypted.

The last feature is important because it provides the basis of security. Malevolent actors cannot go back up the chain and observe or alter information. Each block has a unique hash code and records the hash code of the previous block. If someone attempts to change information within a block then the block's hash changes as well. This dependency within the chain means that the size of the chain

and the complexity of manipulating hash codes makes effective manipulation computationally infeasible. Naturally there are other layers of security like complex verification of systems that want to participation in a particular chain. The pervasiveness and integrity of Blockchain imply a long period of primacy in this environment.

Regulatory issues worldwide affect how data scientists at all levels do their work, including political scientists. University IRBs need to be aware of a quickly changing environment in how researchers interact with subjects. The traditional models are being superseded by online alternatives like Amazon mTurk workers. Does an Internet intermediary truly provide anonymity and protection in a way that IRBs require? In some cases this is unclear. Furthermore, government regulation of privacy, some of which affects academic research, varies widely around the world. On one extreme is the EU's aggressive attempt to preserve citizens' privacy with the European General Data Protection Regulation (GDPR) and on the other extreme is the relatively laissez-faire US federal position along with a patchwork of additional laws at the state level. Of course autocratic regimes favor a particular brand of citizen privacy where they know as much as possible and share as little as possible.

Augmented reality (AR) and virtual reality (VR) are more than just about games, although gaming still dominates in the consumer market. And there is a difference. VR technology is designed completely take over your vision (and possibly other senses) to give the mental impression that you are in a totally different environment. Conversely, augmented reality technology adds information to your current environment, usually by adding digital readouts inside something that often looks like sunglasses. This can be as simple as time, date, weather, and reminder information, or as complex as displays resembling computer screens. These two technologies are increasing being used for academic research purposes in the social sciences as a way to manipulate subject environments with an extreme and flexible form of experimental control. On the privacy side there is concern that commercial applications have easy means of collecting biometric information, with or without subject knowledge.

Returning to the privacy conversation from before, there are new technologies worth paying attention to. One interesting area is Edge Computing. This is designed to exploit the explosive growth of the Internet of Everything (IoT), which is the full set of Internet connected monitoring devices. The shear number of these devices is huge in the numerical way described above: there are over 50 billion connected devices in 2020 expecting to generate over 4.4 zettabytes of data this year according to PriceEconomics. It is well known that totalitarian regimes and other governments are tapped into this world as a way to collect immediate and extensive data on citizens. What makes edge computing important is that it

moves the computational process closer to the subjects (the "edge"), which makes it easier to distill and communicate important parts of the collected data so as not to overburden centralized resources. Such a distributed strategy makes a lot of sense given the volume of the data but it raises concern about data protection at the local distributed level.

Is Data Science a Field?

Yes! Universities have long recognized this. There has been an explosion of data science Masters programs over the last 5-10 years and rarely is a university without one (or something closely related) today. More recently there has also been a notable growth of data science bachelors and PhD programs. So if one is to ask the question in the title of this section, then empirically at the university level my affirmative assertion is supported. But data science as a field evolved from other disciplines and not as an orphan on its own. Using that analogy the parents are clearly: statistics, machine learning (i.e. computer science), mathematics, and the social sciences. The last point is the most important because the huge majority of data science work is done to understand *people*, socially, politically, biomedically, and commercially. Purely technically trained data scientists and computer scientists need social scientists to identify important, profitable, and impactful human questions to be answered, and how to interpret results. In many institutions a data science team will include statisticians, computer scientists, and social scientists with advanced degrees *as well as* members who simply identify as data scientists.

Yet there is a huge shortage of data scientists in academia, government, and industry. Data science programs, particularly at the masters level are not turning out enough qualified employees, even in an economy decimated by other factors. The recruiter Glassdoor recently ranked data science as the #1 best job in 2019 and IBM recorded that employment demand for data scientists increased 28% in 2019. There were about 30M job ads for data scientists in the US alone in 2019. The global Covid-19 pandemic does not appear to be denting this market. Into this mix there is about a 10 year long and increasing trend for PhD *political scientists* to enter this labor market, taking remunerative positions at the large technology companies as well as with government and universities. In an academic job market that may not improve for years it is reassuring to know that our PhD students have rewarding alternatives.

How the Data Century Affects Us in Political Science

Interesting and important forms of political science data are bigger and more complex than ever in the way that I have described and in additional ways. We now have more analytical tools than ever, with huge progress in *qualitative* analysis. But we need more! Regretfully, political science departments do not typically have the large and expensive infrastructure for existing and future big data challenges. This includes centralized computational resources, postdoctoral researchers, and technical support staff. Of course some departments are resource rich in this way and it shows in their output. Does the data science revolution increase the Gini Index of political science researcher resources? I think so. In addition, more presses and journals demand precise replication materials before final publication, including from large complex analyses, requiring more local resources.

The conventional model of journal publishing is becoming increasing outdated in this age of rapid knowledge transfer. Academic journals were created in the 17th century to decrease the time of dissemination of knowledge since books at the time took a very long time to be physically printed and bound. There is now a pressing need to get new knowledge out in political science and a journal review time-span that can take well over a year from submission to publication belongs in the Triassic Era. We should also consider that the traditional journal model where we give commercial entities product for free so that they can sell it back to our university libraries is increasingly obsolete, save for tenure/promotion time when we suddenly become very traditional. So the state of scholarly publishing is about to change fundamentally, and already has (`arXiv`, proceedings, etc.). We also live in a time in political science when the *achievement* of a publication often means more than the *actual content* of that publication. Almost every scholar in the discipline expects the rewards of an incremental work in a top journal to be greater than the rewards of a ground-breaking work in a third-tier journal.

The freshman you are be teaching this Fall were born *after*: the creation of the Internet, ubiquitous sophisticated mobile technology, 9/11, the end of the first Cold War, and the advent of 24 hour constant delivery of the news. And the US has been at war for *their entire lives*. Students sit in the classroom wired into their regular social environment every second of the lecture. They can immediately fact-check anything you say in class, and yet some of what they will get from that search are not actually "facts." They also increasingly want "value" out of the experience in literally the vocational sense (at least their parents do). Data information flows the other way too. Universities are increasingly tracking everything that undergraduates do through their phones: when they attend class, when they are in their

dormitories, where they go off campus, when they visit the campus health clinic, when they eat, and more (George Orwell was an unimaginative by comparison).

Universities will continue to be challenged with regard to data management and storage. There are a variety of NoSQL (not only SQL) frameworks that are necessary to manage big data applications with reasonable response times, including: hierarchical object representation (XML [extensible markup language], JSON [JavaScript object notation], BSON [binary JSON], CBOR [concise binary object representation],etc.) as well as key-value storage. But new tools are required as data get even bigger. Access and synchronization are also ongoing issues for university IT staff. The complexity of transmission, user access, and analytical processing of big data from different sources to users presents problems with synchronization where different data etiologies can be un-linked by time, have conflicting data definitions, have mis-matched metadata, and uncoordinated documentation. And there is a perennial labor shortage. There are not enough trained data scientists to support faculty ongoing and future data-intensive projects at most universities, and there is substantial competition from industry for their services. In addition, as the number of required skills increases there will be a greater need for technical specialization making this problem worse.

A Personal Big Data Challenge: Bayesian Spatial Models with Kriging

Finally I would like to show a real project where the size of data presents serious analytical challenges. This example is based on ongoing work (National Science Foundation grants SES-1630265 and SES-1630263). Start with a simple linear spatial model:

$$Y(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \omega(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{1}$$

where:

- $\mathbf{s}$ is a set of geographic locations $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$
- $Y(\mathbf{s})$ is an associated collection of observed outcomes of interest, $\mathbf{Y} = \{Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)\}$ at those sites
- $\boldsymbol{\mu}(\mathbf{s}) = \mathbf{x}(\mathbf{s})\boldsymbol{\beta}$ is the mean structure based on a linear additive component where $\mathbf{x}(s)$ are covariates at the observed $\mathbf{s}$ locations
- $\omega(\mathbf{s})$ are realizations from a mean-zero stationary (usually) Gaussian spatial process that captures spatial association

- $\epsilon(\mathbf{s})$ is a regular uncorrelated residual term (which includes the so-called nugget effect τ^2, its variance).

We can also rewrite this model more conventionally as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\Sigma} = \sigma^2 H(\phi) + \tau^2 \mathbf{I} \tag{2}$$

with the data objects:

- $H(\phi)_{ij} = \rho(\phi, \mathbf{h})$ and ρ is a spatially appropriate correlation *function* and $\mathbf{h}$ is geographic distance
- ϕ is the *decay* (its inverse is called the range), which is the rate of spatial decrease with increasing distance
- σ^2 is the *partial sill* giving the spatial variability attributable to distance (in addition to the nugget)
- τ^2 is the *nugget*, which gives the variability as the distance between points goes to zero.

The nugget can also be thought of as *microscale variability*: variability at distances smaller than the smallest inter-location distance in the data.

We are also concerned with unobserved spatial points. Full analysis requires the marginal posterior distribution of the coefficient vector $\boldsymbol{\beta}$ as well as the spatial parameters: σ^2, τ^2, and ϕ. But we also want a smooth density blanket for $Y(\mathbf{s})$, which requires predictions at unobserved sites, s_0, giving $Y(\mathbf{s}_0)$ values at chosen locations. The strategy is to use *Bayesian Kriging*, which is the process of applying mathematical decision criteria and then using constrained optimization to spatially smooth, given priors. This problem is considered in the context of Gaussian processes, but many other distributional frameworks are available. An important goal for us here is what minimizes $\mathbb{E}\left[\left(Y(\mathbf{s}_0) - f(\mathbf{Y})\right)^2 |\mathbf{y}\right]$?

Here $E[Y(\mathbf{s}_0)|\mathbf{Y}]$ is the *posterior mean* of $Y(\mathbf{s}_0)$ and therefore minimizes the posterior risk with a squared error loss function. Substituting in the regression quantities gives:

$$\mathbb{E}[Y(s_0)|\mathbf{y}] = \mathbf{X}_0\boldsymbol{\beta} + \Omega_{Y(s_0),\mathbf{y}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3}$$

$$\text{Var}[Y(s_0)|\mathbf{y}] = \sigma^2 + \tau^2 - \Omega'_{Y(s_0),\mathbf{y}}\boldsymbol{\Sigma}^{-1}\Omega_{Y(s_0),\mathbf{y}} \tag{4}$$

where $\Omega_{Y(s_0)} = \sigma^2 + \tau^2$, which is perfectly fine and normally estimable, except that...

- we have $25,000$ training observations and $250,000$ forecast points
- Σ therefore is a $25,000 \times 25,000$ matrix requiring over 300 million calculations to invert

- $\Omega_{Y(s_0),y}$ is a $25,000 \times 250,000$ matrix requiring 6.25 billion calculations for multiplication.

Furthermore, these numbers increase dramatically as the number of unobserved points of interest goes up. This big data problem taxes computation resources in many ways including RAM, storage, swapping, ability to parallelize, and the limitations of specific software packages to even perform the operations. There is no conventional way to estimate spacial models of this size. As a result we have been required to creatively re-form matrix objects, use relatively obscure linear algebra relationships, break operations up into multiple discrete tasks, and consider hardware issues in new ways. The current solution is written in `C++` code to run on AWS, which is labor-intensive and ultimately expensive.

Wrapping Up...

Human life is more complex, data-oriented, and technical than ever before. Every field, including political science, needs to understand that they are also a data science field. By "data" I mean qualitative and quantitative data (e.g. "evidence"). The ongoing Covid-19 pandemic only makes these data science challenges more difficult, more elaborate, and more important.