Chapter 1
Statistics, Statisticians, and the Internet of Things

John M. Jordan and Dennis K. J. Lin

Abstract Within the overall rubric of big data, one emerging subset holds particular promise, peril, and attraction. Machine-generated traffic from sensors, data logs, and the like, transmitted using Internet practices and principles, is being referred to as the "Internet of Things" (IoT). Understanding, handing, and analyzing this type of data will stretch existing tools and techniques, thus providing a proving ground for other disciplines to adopt and adapt new methods and concepts. In particular, new tools will be needed to analyze data in motion rather than data at rest, and there are consequences of having constant or near-constant readings from the ground-truth phenomenon as opposed to numbers at a remove from their origin. Both machine learning and traditional statistical approaches will coevolve rapidly given the economic forces, national security implications, and wide public benefit of this new area of investigation. At the same time, data practitioners will be exposed to the possibility of privacy breaches, accidents causing bodily harm, and other concrete consequences of getting things wrong in theory and/or practice. We contend that the physical instantiation of data practice in the IoT means that statisticians and other practitioners may well be seeing the origins of a post-big data era insofar as the traditional abstractions of numbers from ground truth are attenuated and in some cases erased entirely.

Keywords Machine traffic · Internet of Things · Sensors · Machine learning · Statistical approaches to big data

J. M. Jordan
Department of Supply Chain & Information Systems, Smeal College of Business, Pennsylvania State University, University Park, PA, USA
e-mail: jmj13@psu.edu

D. K. J. Lin (✉)
Department of Statistics, Eberly College of Science, Pennsylvania State University, University Park, PA, USA
e-mail: dkl5@psu.edu

W. K. Härdle et al. (eds.), *Handbook of Big Data Analytics*, Springer Handbooks of Computational Statistics, https://doi.org/10.1007/978-3-319-18284-1_1

1.1 Introduction

Even though it lacks a precise definition, the notion of an "Internet of Things" refers generally to networks of sensors, actuators, and machines that communicate over the Internet and related networks. (Some years ago, the number of inanimate objects on the Internet surpassed the number of human beings with connections.) In this chapter, we will first elaborate on the components of the IoT and discuss its data components. The place of statistics in this new world follows, and then we raise some real-world issues such as skills shortages, privacy protection, and so on, before concluding.

1.1.1 The Internet of Things

The notion of an Internet of Things is at once both old and new. From the earliest days of the World Wide Web, devices (often cameras) were connected so people could see the view out a window, traffic or ski conditions, a coffee pot at the University of Cambridge, or a Coke machine at Carnegie Mellon University. The more recent excitement dates to 2010 or thereabouts and builds on a number of developments: many new Internet Protocol (IP) addresses have become available, the prices of sensors are dropping, new data and data-processing models are emerging to handle the scale of billions of device "chirps," and wireless bandwidth is getting more and more available.

1.1.2 What Is Big Data in an Internet of Things?

Why do sensors and connected devices matter for the study of statistics? If one considers the definition of a robot—an electromechanical device that can digitally sense and think, then act upon the physical environment—those same actions characterize large-scale Internet of Things systems: they are essentially meta-robots. The GE Industrial Internet model discussed below includes sensors on all manner of industrial infrastructure, a data analytics platform, and humans to make presumably better decisions based on the massive numbers from the first domain crunched by algorithms and computational resources in the second. Thus, the Internet of Things becomes, in some of its incarnations, an offshoot of statistical process control, six-sigma, and other established industrial methodologies.

Unlike those processes that operated inside industrial facilities, however, the Internet of Things includes sensors attached to or otherwise monitoring individual people in public. Google Glass, a head-mounted smartphone, generated significant controversy before it was pulled from distribution in 2015. This reaction was a noteworthy step in the adoption of Internet of Things systems: both technical details and cultural norms need to be worked out. Motion sensors, cameras, facial recognition, and voice recording and synthesis are very different activities on a factory floor compared to a city sidewalk.

Thus, the Internet of Things is both an extension of existing practices and the initial stage in the analysis of an ever-more instrumented public sphere. The IoT (1) generates substantially large bodies of data (2) in incompatible formats (3) sometimes attached to personal identity. Statisticians now need to think about new physical-world safety issues and privacy implications in addition to generating new kinds of quantitative tools that can scale to billions of data points per hour, across hundreds of competing platforms and conventions. The magnitude of the task cannot be overstated.

1.1.3 Building Blocks[1]

The current sensor landscape can be understood more clearly by contrasting it to the old state of affairs. Most important, sensor networks mimicked analog communications: radios couldn't display still pictures (or broadcast them), turntables couldn't record video, and newspapers could not facilitate two- or multi-way dialog in real time. For centuries, sensors in increasing precision and sophistication were invented to augment human senses: thermometers, telescopes, microscopes, ear trumpets, hearing aids, etc. With the nineteenth-century advances in electro-optics and electromechanical devices, new sensors could be developed to extend the human senses into different parts of the spectrum (including infrared, radio frequencies, measurement of vibration, underwater acoustics, etc.).

Where they were available, electromechanical sensors and later sensor networks

- Stood alone
- Measured one and only one thing
- Cost a lot to develop and implement
- Had inflexible architectures: they did not adapt well to changing circumstances

Sensors traditionally stood alone because networking them together was expensive and difficult. Given the lack of shared technical standards, in order to build a network of offshore data buoys for example, the interconnection techniques and protocols would be uniquely engineered to a particular domain, in this case, saltwater, heavy waves, known portions of the magnetic spectrum, and so on. Another agency seeking to connect sensors of a different sort (such as surveillance cameras) would have to start from scratch, as would a third agency monitoring road traffic.

[1] This section relies heavily on John M. Jordan, *Information, Technology, and Innovation* (Hoboken: John Wiley, 2012), ch. 23.

In part because of their mechanical componentry, sensors rarely measured across multiple yardsticks. Oven thermometers measured only oven temperature, and displayed the information locally, if at all (given that perhaps a majority of sensor traffic informs systems rather than persons, the oven temperature might only drive the thermostat rather than a human-readable display). Electric meters only counted watt-hours in aggregate. In contrast, today a consumer Global Positioning Satellite (GPS) unit or smartphone will tell location, altitude, compass heading, and temperature, along with providing weather radio.

Electromechanical sensors were not usually mass produced, with the exception of common items such as thermometers. Because supply and demand were both limited, particularly for specialized designs, the combination of monopoly supply and small order quantities kept prices high.

1.1.4 Ubiquity

Changes in each of these facets combine to help create today's emerging sensor networks, which are growing in scope and capability every year. The many examples of sensor capability accessible to (or surveilling) the everyday citizen illustrate the limits of the former regime: today there are more sensors recording more data to be accessed by more end points. Furthermore, the traffic increasingly originates and transits exclusively in the digital domain.

- Computers, which sense their own temperature, location, user patterns, number of printer pages generated, etc.
- Thermostats, which are networked within buildings and now remotely controlled and readable.
- Telephones, the wireless variety of which can be understood as beacons, barcode scanners, pattern matchers (the Shazam application names songs from a brief audio sample), and network nodes.
- Motor and other industrial controllers: many cars no longer have mechanical throttle linkages, so people step on a sensor every day without thinking as they drive by wire. Automated tire-pressure monitoring is also standard on many new cars. Airbags rely on a sophisticated system of accelerometers and high-speed actuators to deploy the proper reaction for collision involving a small child versus a lamp strapped into the front passenger seat.
- Vehicles: the OBD II diagnostics module, the toll pass, satellite devices on heavy trucks, and theft recovery services such as Lojack, not to mention the inevitable mobile phone, make vehicle tracking both powerful and relatively painless.
- Surveillance cameras (of which there are over 10,000 in Chicago alone, and more than 500,000 in London).[2]

[2] Brian Palmer, "Big Apple is Watching You," Slate, May 3, 2010, http://www.slate.com/id/2252729/, accessed 29 March 2018.

- Most hotel door handles and many minibars are instrumented and generate electronic records of people's and vodka bottles' comings and goings.
- Sensors, whether embedded in animals (RFID chips in both household pets and race horses) or gardens (the EasyBloom plant moisture sensor connects to a computer via USB and costs only $50), or affixed to pharmaceutical packaging.

Note the migration from heavily capital-intensive or national-security applications down-market. A company called Vitality has developed a pill-bottle monitoring system: if the cap is not removed when medicine is due, an audible alert is triggered, or a text message could be sent.[3]

A relatively innovative industrial deployment of vibration sensors illustrates the state of the traditional field. In 2006, BP instrumented an oil tanker with "motes," which integrated a processor, solid-state memory, a radio, and an input/output board on a single 2" square chip. Each mote could receive vibration data from up to ten accelerometers, which were mounted on pumps and motors in the ship's engine room. The goal was to determine if vibration data could predict mechanical failure, thus turning estimates—a motor teardown every 2000 h, to take a hypothetical example—into concrete evidence of an impending need for service.

The motes had a decided advantage over traditional sensor deployments in that they operated over wireless spectrum. While this introduced engineering challenges arising from the steel environment as well as the need for batteries and associated issues (such as lithium's being a hazardous material), the motes and their associated sensors were much more flexible and cost-effective to implement compared to hardwired solutions. The motes also communicate with each other in a mesh topology: each mote looks for nearby motes, which then serve as repeaters en route to the data's ultimate destination. Mesh networks are usually dynamic: if a mote fails, signal is routed to other nearby devices, making the system fault tolerant in a harsh environment. Finally, the motes could perform signal processing on the chip, reducing the volume of data that had to be transmitted to the computer where analysis and predictive modeling was conducted. This blurring of the lines between sensing, processing, and networking elements is occurring in many other domains as well.[4]

All told, there are dozens of billions of items that can connect and combine in new ways. The Internet has become a common ground for many of these devices, enabling multiple sensor feeds—traffic camera, temperature, weather map, social media reports, for example—to combine into more useful, and usable, applications, hence the intuitive appeal of "the Internet of Things." As we saw earlier, network effects and positive feedback loops mean that considerable momentum can develop as more and more instances converge on shared standards. While we will not

[3] Ben Coxworth, "Ordinary pill bottle has clever electronic cap," New Atlas, May 5, 2017, https://newatlas.com/pillsy-smart-pill-bottle/49393/, accessed 29 March 2018.

[4] Tom Kevan, "Shipboard Machine Monitoring for Predictive Maintenance," *Sensors Mag*, February 1, 2006. http://www.sensorsmag.com/sensors-mag/shipboard-machine-monitoring-predictive-maintenance-715?print=1

discuss them in detail here, it can be helpful to think of three categories of sensor interaction:

- Sensor to people: the thermostat at the ski house tells the occupants that the furnace is broken the day before they arrive, or a dashboard light alerting the driver that the tire pressure on their car is low.
- Sensor to sensor: the rain sensor in the automobile windshield alerts the antilock brakes of wet road conditions and the need for different traction-control algorithms.
- Sensor to computer/aggregator: dozens of cell phones on a freeway can serve as beacons for a traffic-notification site, at much lower cost than helicopters or "smart highways."

An "Internet of Things" is an attractive phrase that at once both conveys expansive possibility and glosses over substantial technical challenges. Given 20+ years of experience with the World Wide Web, people have long experience with hyperlinks, reliable inter-network connections, search engines to navigate documents, and Wi-Fi access everywhere from McDonalds to over the mid-Atlantic in flight. None of these essential pieces of scaffolding has an analog in the Internet of Things, however: garage-door openers and moisture sensors aren't able to read; naming, numbering, and navigation conventions do not yet exist; low-power networking standards are still unsettled; and radio-frequency issues remain problematic. In short, as we will see, "the Internet" may not be the best metaphor for the coming stage of device-to-device communications, whatever its potential utility.

Given that "the Internet" as most people experience it is global, searchable, and anchored by content or, increasingly, social connections, the "Internet of Things" will in many ways be precisely the opposite. Having smartphone access to my house's thermostat is a private transaction, highly localized and preferably NOT searchable by anyone else. While sensors will generate volumes of data that are impossible for most humans to comprehend, that data is not content of the sort that Google indexed as the foundation of its advertising-driven business. Thus, while an "Internet of Things" may feel like a transition from a known world to a new one, the actual benefits of networked devices separate from people will probably be more foreign than being able to say "I can connect to my appliances remotely."

1.1.5 Consumer Applications

The notion of networked sensors and actuators can usefully be subdivided into industrial, military/security, or business-to-business versus consumer categories. Let us consider the latter first. Using the smartphone or a web browser, it is already possible to remotely control and/or monitor a number of household items:

- Slow cooker
- Garage-door opener

- Blood-pressure cuff
- Exercise tracker (by mileage, heart rate, elevation gain, etc.)
- Bathroom scale
- Thermostat
- Home security system
- Smoke detector
- Television
- Refrigerator

These devices fall into some readily identifiable categories: personal health and fitness, household security and operations, and entertainment. While the data logging of body weight, blood pressure, and caloric expenditures would seem to be highly relevant to overall physical wellness, few physicians, personal trainers, or health insurance companies have built business processes to manage the collection, security, or analysis of these measurements. Privacy, liability, information overload, and, perhaps most centrally, outcome-predicting algorithms have yet to be developed or codified. If I send a signal to my physician indicating a physical abnormality, she could bear legal liability if her practice does not act on the signal and I subsequently suffer a medical event that could have been predicted or prevented.

People are gradually becoming more aware of the digital "bread crumbs" our devices leave behind. Progressive Insurance's Snapshot campaign has had good response to a sensor that tracks driving behavior as the basis for rate-setting: drivers who drive frequently, or brake especially hard, or drive a lot at night, or whatever could be judged worse risks and be charged higher rates. Daytime or infrequent drivers, those with a light pedal, or people who religiously buckle seat belts might get better rates. This example, however, illustrates some of the drawbacks of networked sensors: few sensors can account for all potentially causal factors. Snapshot doesn't know how many people are in the car (a major accident factor for teenage drivers), if the radio is playing, if the driver is texting, or when alcohol might be impairing the driver's judgment. Geographic factors are delicate: some intersections have high rates of fraudulent claims, but the history of racial redlining is also still a sensitive topic, so data that might be sufficiently predictive (postal codes traversed) might not be used out of fear it could be abused.

The "smart car" applications excepted, most of the personal Internet of Things use cases are to date essentially remote controls or intuitively useful data collection plays. One notable exception lies in pattern-recognition engines that are grouped under the heading of "augmented reality." Whether on a smartphone/tablet or through special headsets such as Google Glass, a person can see both the physical world and an information overlay. This could be a real-time translation of a road sign in a foreign country, a direction-finding aid, or a tourist application: look through the device at the Eiffel Tower and see how tall it is, when it was built, how long the queue is to go to the top, or any other information that could be attached to the structure, attraction, or venue.

While there is value to the consumer in such innovations, these connected devices will not drive the data volumes, expenditures, or changes in everyday life that will emerge from industrial, military, civic, and business implementations.

1.1.6 The Internets of [Infrastructure] Things

Because so few of us see behind the scenes to understand how public water mains, jet engines, industrial gases, or even nuclear deterrence work, there is less intuitive ground to be captured by the people working on large-scale sensor networking. Yet these are the kinds of situations where networked instrumentation will find its broadest application, so it is important to dig into these domains.

In many cases, sensors are in place to make people (or automated systems) aware of exceptions: is the ranch gate open or closed? Is there a fire, or just an overheated wok? Is the pipeline leaking? Has anyone climbed the fence and entered a secure area? In many cases, a sensor could be in place for years and never note a condition that requires action. As the prices of sensors and their deployment drop, however, more and more of them can be deployed in this manner, if the risks to be detected are high enough. Thus, one of the big questions in security—in Bruce Schneier's insight, not "Does the security measure work?" but "Are the gains in security worth the costs?"—gets difficult to answer: the costs of IP-based sensor networks are dropping rapidly, making cost-benefit-risk calculations a matter of moving targets.

In some ways, the Internet of Things business-to-business vision is a replay of the RFID wave of the mid-aughts. Late in 2003, Walmart mandated that all suppliers would use radio-frequency tags on their incoming pallets (and sometimes cases) beginning with the top 100 suppliers, heavyweight consumer packaged goods companies like Unilever, Procter & Gamble, Gillette, Nabisco, and Johnson & Johnson. The payback to Walmart was obvious: supply chain transparency. Rather than manually counting pallets in a warehouse or on a truck, radio-powered scanners could quickly determine inventory levels without workers having to get line-of-sight reads on every bar code. While the 2008 recession contributed to the scaled-back expectations, so too did two powerful forces: business logic and physics.

To take the latter first, RFID turned out to be substantially easier in labs than in warehouses. RF coverage was rarely strong and uniform, particularly in retrofitted facilities. Electromagnetic noise—in the form of everything from microwave ovens to portable phones to forklift-guidance systems—made reader accuracy an issue. Warehouses involve lots of metal surfaces, some large and flat (bay doors and ramps), others heavy and in motion (forklifts and carts): all of these reflect radio signals, often problematically. Finally, the actual product being tagged changes radio performance: aluminum cans of soda, plastic bottles of water, and cases of tissue paper each introduce different performance effects. Given the speed of assembly lines and warehouse operations, any slowdowns or errors introduced by a new tracking system could be a showstopper.

The business logic issue played out away from the shop floor. Retail and consumer packaged goods profit margins can be very thin, and the cost of the RFID tagging systems for manufacturers that had negotiated challenging pricing schedules with Walmart was protested far and wide. The business case for total supply chain transparency was stronger for the end seller than for the suppliers, manufacturers, and truckers required to implement it for Walmart's benefit. Given that the systems delivered little value to the companies implementing them, and given that the technology didn't work as advertised, the quiet recalibration of the project was inevitable.

RFID is still around. It is a great solution to fraud detection, and everything from sports memorabilia to dogs to ski lift tickets can be easily tested for authenticity. These are high-value items, some of them scanned no more than once or twice in a lifetime rather than thousands of times per hour, as on an assembly line. Database performance, industry-wide naming and sharing protocols, and multiparty security practices are much less of an issue.

While it's useful to recall the wave of hype for RFID circa 2005, the Internet of Things will be many things. The sensors, to take only one example, will be incredibly varied, as a rapidly growing online repository makes clear (see http://devices.wolfram.com/).[5] Laboratory instruments are shifting to shared networking protocols rather than proprietary ones. This means it's quicker to set up or reconfigure an experimental process, not that the lab tech can see the viscometer or Geiger counter from her smart phone or that the lab will "put the device on the Internet" like a webcam.

Every one of the billions of smartphones on the planet is regularly charged by its human operator, carries a powerful suite of sensors—accelerometer, temperature sensor, still and video cameras/bar-code readers, microphone, GPS receiver—and operates on multiple radio frequencies: Bluetooth, several cellular, and Wi-Fi. There are ample possibilities for crowdsourcing news coverage, fugitive hunting, global climate research (already, amateur birders help show differences in species' habitat choices), and more using this one platform.

Going forward, we will see more instrumentation of infrastructure, whether bridges, the power grid, water mains, dams, railroad tracks, or even sidewalks. While states and other authorities will gain visibility into security threats, potential outages, maintenance requirements, or usage patterns, it's already becoming clear that there will be multiple paths by which to come to the same insight. The state of Oregon was trying to enhance the experience of bicyclists, particularly commuters. While traffic counters for cars are well established, bicycle data is harder to gather. Rather than instrumenting bike paths and roadways, or paying a third party to do so, Oregon bought aggregated user data from Strava, a fitness-tracking smartphone app. While not every rider, particularly commuters, tracks his mileage, enough do that the bike-lane planners could see cyclist speeds and traffic volumes by time of day, identify choke points, and map previously untracked behaviors.

[5] Wolfram Connected Devices Project, (http://devices.wolfram.com/), accessed 29 March 2018.

Strava was careful to anonymize user data, and in this instance, cyclists were the beneficiaries. Furthermore, cyclists compete on Strava and have joined with the expectation that their accomplishments can show up on leader boards. In many other scenarios, however, the Internet of Things' ability to "map previously untracked behaviors" will be problematic, for reasons we will discuss later. To provide merely one example, when homes are equipped with so-called smart electrical meters, it turns out that individual appliances and devices have unique "fingerprints" such that outside analysis can reveal when the toaster, washing machine, or hair dryer was turned on and off.[6] Multiply this capability across toll passes, smartphones, facial recognition, and other tools, and the privacy threat becomes significant.

1.1.7 Industrial Scenarios

GE announced its Industrial Internet initiative in 2013. The goal is to instrument more and more of the company's capital goods—jet engines are old news, but also locomotives, turbines, undersea drilling rigs, MRI machines, and other products—with the goal of improving power consumption and reliability for existing units and to improve the design of future products. Given how big the company's footprint is in these industrial markets, 1% improvements turn out to yield multibillion-dollar opportunities. Of course, instrumenting the devices, while not trivial, is only the beginning: operational data must be analyzed, often using completely new statistical techniques, and then people must make decisions and put them into effect.

The other striking advantage of the GE approach is financial focus: 1% savings in a variety of industrial process areas yields legitimately huge cost savings opportunities. This approach has the simultaneous merits of being tangible, bounded, and motivational. Just 1% savings in aviation fuel over 15 years would generate more than $30 billion, for example. To realize this promise, however, GE needs to invent new ways of networking, storage, and data analysis. As Bill Ruh, the company's vice president of global software services, stated, "Our current jet aircraft engines produce one terabyte of data per flight. ... On average an airline is doing anywhere from five to ten flights a day, so that's 5–10 terabytes per plane, so when you're talking about 20,000 planes in the air you're talking about an enormous amount of data per day."[7] Using different yardsticks, Ruh framed the scale in terms of variables: 50 million of them, from 10 million sensors.

To get there, the GE vision is notably realistic about the many connected investments that must precede the harvesting of these benefits.

[6] Ariel Bleicher, "Privacy on the Smart Grid," IEEE Spectrum, 5 October 2010, http://spectrum.ieee.org/energy/the-smarter-grid/privacy-on-the-smart-grid, accessed 29 March 2018.

[7] Danny Palmer, "The future is here today: How GE is using the Internet of Things, big data and robotics to power its business," Computing 12 March 2015, http://www.computing.co.uk/ctg/feature/2399216/the-future-is-here-today-how-ge-is-using-the-internet-of-things-big-data-and-robotics-to-power-its-business, accessed 29 March 2018.

1. The technology doesn't exist yet. Sensors, instrumentation, and user interfaces need to be made more physically robust, usable by a global workforce, and standardized to the appropriate degree.
2. Information security has to protect assets that don't yet exist, containing value that has yet to be measured, from threats that have yet to materialize.
3. Data literacy and related capabilities need to be cultivated in a global workforce that already has many skills shortfalls, language and cultural barriers, and competing educational agendas. Traditional engineering disciplines, computer science, and statistics will merge into new configurations.[8]

1.2 What Kinds of Statistics Are Needed for Big IoT Data?

The statistical community is beginning to engage with machine learning and computer science professionals on the issue of so-called big data. Challenges abound: data validation at petabyte scale; messy, emergent, and dynamic underlying phenomena that resist conventional hypothesis testing; and the need for programming expertise for computational heavy lifting. Most importantly, techniques are needed to deal with flowing data as opposed to static data sets insofar as the phenomena instrumented in the IoT can be life-critical: ICU monitoring, the power grid, fire alarms, and so on. There is no time for waiting for summarized, normalized data because the consequences of normal lags between reading and analysis can be tragic.

1.2.1 Coping with Complexity

In the Internet of Things, we encounter what might be called "big^2 data": all the challenges of single-domain big data remain, but become more difficult given the addition of cross-boundary complexity. For example, astronomers or biostatisticians must master massive data volumes of relatively homogeneous data. In the Internet of Things, it is as if a geneticist also had to understand data on particle physics or failure modes of carbon fiber.

Consider the example of a military vehicle instrumented to determine transmission failure to facilitate predictive maintenance. The sensors cannot give away any operational information that could be used by an adversary, so radio silencing and data encryption are essential, complicating the data acquisition process. Then comes the integration of vast quantities of multiple types of data: weather (including tem-

[8] Peter C. Evans and Marco Annunziata, *Industrial Internet: Pushing the Boundaries of Minds and Machines*, 26 November 2012, p. 4, http://www.ge.com/docs/chapters/Industrial_Internet.pdf, p. 4., accessed 29 March 2018.

perature, humidity, sand/dust, mud, and so on); social network information (think of a classified Twitter feed on conditions and operational updates from the bottom of the organization up); vibration and other mechanical measurements; dashboard indicators such as speedometer, gearshift, engine temperature, and tachometer; text-heavy maintenance logs, possibly including handwriting recognition; and surveillance data (such as satellite imagery).

Moving across domains introduces multiple scales, some quantitative (temperature) and others not (maintenance records using terms such as "rough," "bumpy," and "intermittent" that could be synonymous or distinct). How is an X change in a driveshaft harmonic resonance to correlate with sandy conditions across 10,000 different vehicles driven by 50,000 different drivers? What constitutes a control or null variable? The nature of noise in such a complex body of data requires new methods of extraction, compression, smoothing, and error correction.

1.2.2 Privacy

Because the Internet of Things can follow real people in physical space (whether through drones, cameras, cell phone GPS, or other means), privacy and physical safety become more than theoretical concerns. Hacking into one's bank account is serious but rarely physically dangerous; having stop lights or engine throttles compromised is another matter entirely, as the world saw in the summer of 2015 when an unmodified Jeep was remotely controlled and run off the road.[9] Given the large number of related, cross-domain variables, what are the unintended consequences of optimization?

De-anonymization has been shown to grow easier with large, sparse data sets.[10] Given the increase in the scale and diversity of readings or measurements attached to an individual, it is theoretically logical that the more sparse data points attach to an individual, the simpler the task of personal identification becomes (something as basic as taxi fare data, which intuitively feels anonymous, can create a privacy breach at scale: http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/).[11] In addition, the nature of IoT measurements might not feel as personally risky at the time of data creation: logging into a financial institution heightens one's sense of awareness, whereas walking down the street, being logged by cameras and GPS, might feel more carefree than it

[9] Andy Greenberg, "Hackers Remotely Kill a Jeep on the Highway - With Me in It," Wired, 21 July 2015, http://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/, accessed 29 March 2018.

[10] Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," no date, https://www.cs.utexas.edu/∼shmat/shmat_oak08netflix.pdf, accessed 29 March 2018.

[11] Anthony Tockar, "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset," neustar Research, 15 September 2014, accessed 29 March 2018 [Note: "neustar" is lower-case in the corporate branding].

perhaps should. The cheapness of computer data storage (as measured by something called Kreider's law) combines with the ubiquity of daily digital life to create massive data stores recording people's preferences, medications, travels, and social contacts. The statistician who analyzes and combines such data involving flesh-and-blood people in real space bears a degree of responsibility for their privacy and security. (Researchers at Carnegie Melon University successfully connected facial recognition software to algorithms predicting the subjects' social security numbers.[12]) Might the profession need a new code of ethics akin to the Hippocratic oath? Can the statistician be value-neutral? Is there danger of data "malpractice"?

1.2.3 Traditional Statistics Versus the IoT

Traditional statistical thinking holds that large samples are better than small ones, while some machine learning advocates assert that very large samples render hypotheses unnecessary.[13] At this intersection, the so-called the death of p-value is claimed.[14] However, fundamental statistical thinking with regard to significance, for example, still applies (although the theories may not be straightforwardly applied in very large data sets). Big data on its own cannot replace scientific/statistical thinking. Thus, a wishlist for needed statistical methodologies should have the following properties:

- High-impact problems

  Refining existing methodologies is fine, but more efforts should focus on working high-impact problems, especially those problems from other disciplines. Statisticians seem to keep missing opportunities: examples range from genetics to data mining. We believe that statisticians should seek out high-impact problems, instead of waiting for other disciplines to formulate the problems into statistical frames. Collaboration across many disciplines will be necessary, if unfamiliar, behavior. This leads to the next item.
- Provide structure for poorly defined problems

  A skilled statistician is typically most comfortable and capable when dealing with well-defined problems. Instead, statisticians should develop some methodologies for poorly defined problems and help devise a strategy of attack. There are many opportunities for statistical applications, but most of them are not in the "standard" statistics frame—it will take some intelligent persons to

[12] Deborah Braconnier, "Facial recognition software could reveal your social security number," Phys.org, 2 August 2011, https://phys.org/news/2011-08-facial-recognition-software-reveal-social.html, accessed 29 March 2018.

[13] Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," Wired, 23 June 2008, https://www.wired.com/2008/06/pb-theory/, accessed 29 March 2018.

[14] Tom Siegfried, "P value ban: small step for a journal, giant leap for science," ScienceNews, 17 March 2015, https://www.sciencenews.org/blog/context/p-value-ban-small-step-journal-giant-leap-science, accessed 29 March 2018.

formulate these problems into statistics-friendly problems (then to be solved by statisticians). Statisticians can devote more efforts to be such intelligent persons.

- Develop new theories

  Most fundamental statistical theories based upon iid (independently identically distributed) for one fixed population (such ascentral limit theorem, or law of large number) may need to be modified to be appropriately applied to big data world. Many (non-statisticians) believe that big data leads to "the death of p-value." The logic behind this is that when the sample size n becomes really large, all p-values will be significant—regardless how little the practical significance is. This is indeed a good example of misunderstanding the fundamentals. One good example is about "small n and large p" where the sparsity property is assumed. First, when there are many exploratory variables, some will be classified as active variables (whether or not this is true!). Even worse, after the model is built (mainly based on the sparsity property), the residuals may highly correlate with some remaining variables—this contradicts the assumption for all fundamental theorems that "error is independent with all exploratory variables." New measurement is needed for independence in this case.

1.2.4 A View of the Future of Statistics in an IoT World

Having those wishlist items in mind, what kinds of statistics are needed for big data? For an initial approximation, here are some very initial thoughts under consideration.

- Statistics and plots for (many) descriptive statistics. If conventional statistics are to be used for big data, and it is very likely there will be too many of them because of the heterogeneity of the data, what is the best way to extract important information from these statistics? For example, how to summarize thousands of correlations? How about thousands of p-values? ANOVAs? Regression models? Histograms? etc. Advanced methods to obtain "sufficient statistics" (whatever it means in a particular context: astrophysics and biochemistry will have different needs, for example) from those many conventional statistics are needed.
- Coping with heterogeneity. Numbers related to such sensor outputs as check-engine lights, motion detectors, and flow meters can be extremely large, of unknown quality, and difficult to align with more conventional measurement systems.
- Low-dimension behavior. Whatever method is feasible for big data (the main concern being the computational costs), the reduction in resolution as it is converted to low-dimension resolution (especially 2D graphs) is always important to keep in mind.
- As we have mentioned, analyzing real-time measurements that are derived from actual ground truth demands stream-based techniques that exceed standard practice in most statistical disciplines.

- Norm or Extreme. Depending on the problem, we could be interested in either norm or extreme, or both. Basic methods for both feature extraction (mainly for extremes) and pattern recognition (mainly for norm) are needed.
- Methods for new types/structures of data. A simple example would be "How to build up a regression model, when both inputs and outputs are network variables?" Most existing statistical methodologies are limited to numbers (univariate or multivariate), but there is some recent work for functional data or text data. How to extract the basic information (descriptive statistics) or even analysis (inferential statistics) of these new types of data are highly demanding. This includes network data, symbolic data, fingerprints data, 2D or 3D image data, just to name a few. There is more that can be done, if we are willing to open our minds.
- Prediction vs estimation. One difference between computer science and statistics methods has to do with the general goal—while CS people focus more on prediction, statisticians focus more on estimation (or statistical inference). Take Artificial Neural Networks (ANN) as an example: the method can fit almost anything, but what does it mean? ANN is thus popularly used in data mining, but has received relatively low attention from statisticians. For big data, it is clear that prediction is probably more feasible in most cases. **Note:** in some very fundamental cases, we believe that statistical inference remains important, always bearing in mind the essential research question at hand.

1.3 Big Data in the Real World

Moving statistical and analytical techniques from academic and laboratory settings into the physical world sensed and measured by the IoT introduces new challenges. Not surprisingly, organizational and technical matters are emerging, and even the limits of human cognition must be appreciated and accounted for.

1.3.1 Skills

Here's a quiz: ask someone in the IT shop how many of his of her colleagues are qualified to work in Hive, Pig, Cassandra, MongoDb, or Hadoop. These are some of the tools that are emerging from the front-runners in big data, web-scale companies including Google (that needs to index the entire Internet), Facebook (manage a billion users), Amazon (construct and run the world's biggest online merchant), or Yahoo (figure out what social media is conveying at the macro scale). Outside this small industry, big data skills are rare; then consider how few people understand both data skills and the intricacies of industrial and other behind-the-scenes processes, many of them life critical (e.g., the power grid or hospital ICU sensor networks).

1.3.2 Politics

Control over information is frequently thought to bring power within an organization. Big data, however, is heterogeneous, is multifaceted, and can bring performance metrics where they had not previously operated. If a large retailer, hypothetically speaking, traced its customers' purchase behavior first to social media expressions and then to advertising channel, how will the various budget-holders respond? Uncertainty as to ad spend efficacy is as old as advertising, but tracing ad channels to purchase activity might bring light where perhaps it is not wanted. Information sharing across organizational boundaries ("how are you going to use this data?") can also be unpopular. Once it becomes widely understood how one's data "bread crumbs" can be manipulated, will consumers/citizens demand stricter regulation?

1.3.3 Technique

Given that relational databases have been around for about 35 years, a substantial body of theory and practice makes these environments predictable. Big data, by contrast, is just being invented, but already there are some important differences between the two: Most enterprise data is generated by or about humans and organizations: SKUs are bought by people, bills are paid by people, health care is provided to people, and so on. At some level, many human activities can be understood at human scale. Big data, particularly social media, can come from people too, but in more and more cases, it comes from machines: server logs, point of sale scanner data, security sensors, and GPS traces. Given that these new types of IoT data don't readily fit into relational structures and can get massively large in terms of storage, it's nontrivial to figure out what questions to ask of these data types.

When data is loaded into relational systems, it must fit predefined categories that ensure that what gets put into a system makes sense when it is pulled out. This process implies that the system is defined at the outset for what the designers expect to be queried: the questions are known, more or less, before the data is entered in a highly structured manner. In big data practice, meanwhile, data is stored in as complete a form as possible, close to its original state. As little as possible is thrown out so queries can evolve and not be constrained by the preconceptions of the system. Thus, these systems can look highly random to traditional database experts. It's important to stress that big data will not replace relational databases in most scenarios; it's a matter of now having more tools to choose from for a given task.

1.3.4 Traditional Databases

Traditional databases are designed for a concrete scenario, then populated with examples (customers, products, facilities, or whatever), usually one per row: the questions and answers one can ask are to some degree predetermined. Big data can be harvested in its original form and format, and then analyzed as the questions emerge. This open-ended flexibility can of course be both a blessing and a curse.

Traditional databases measured the world in numbers and letters that had to be predicted: zip codes were 5 or 10 digits, SKU formats were company specific, or mortgage payments were of predictable amounts. Big data can accommodate Facebook "likes," instances of the "check engine" light illuminating, cellphone location mapping, and many other types of information.

Traditional databases are limited by the computing horsepower available: to ask harder questions often means buying more hardware. Big data tools can scale up much more gracefully and cost-effectively, so decision-makers must become accustomed to asking questions they could not contemplate previously. To judge advertising effectiveness, one cable operator analyzed every channel-surfing click of every remote across every household in its territory, for example: not long ago, such an investigation would have been completely impractical.

1.3.5 Cognition

What does it mean to think at large scales? How do we learn to ask questions of the transmission of every car on the road in a metropolitan area, of the smartphone of every customer of a large retail chain, or of every overnight parcel in a massive distribution center? How can more and more people learn to think probabilistically rather than anecdotally?

The mantra that "correlation doesn't imply causation" is widely chanted yet frequently ignored; it takes logical reasoning beyond statistical relationships to test what's really going on. Unless the data team can grasp the basic relationships of how a given business works, the potential for complex numerical processing to generate false conclusions is ever present. Numbers do not speak for themselves; it takes a human to tell stories, but as Daniel Kahneman and others have shown, our stories often embed mental traps. Spreadsheets remain ubiquitous in the modern enterprise, but numbers at the scale of Google, Facebook, or Amazon must be conveyed in other ways. Sonification—turning numbers into a range of audible tones—and visualization show a lot of promise as alternative pathways to the brain, bypassing mere and non-intuitive numerals. In the meantime, the pioneers are both seeing the trail ahead and taking some arrows in the back for their troubles. But the faster people, and especially statisticians, begin to break the stereotype that "big data is what we've always done, just with more records or fields," the faster the breakthrough questions, insights, and solutions will redefine practice.

1.4 Conclusion

There's an important point to be made up front: whether it originates in a financial system, public health record-keeping, or sensors on electrical generators, *big data is not necessarily complete, or accurate, or true.* Asking the right questions is in some cases learned through experience, or made possible by better theory, or a matter of luck. But in many instances, by the time investigators figure out what they should be measuring in complex systems, it's too late to instrument the "before" state to compare to the "after." Signal and noise can be problematic categories as well: one person's noise can be a goldmine for someone else. Context is everything. Value is in the eye of the beholder, not the person crunching the numbers. However, this is rarely the case. Big data is big, often because it is automatically collected. Thus, in many cases, it may not contain much information relative to noise. This is sometimes called a DRIP—Data Rich, Information Poor—environment. The IoT is particularly prone to these issues, given both (a) notable failure and error rates of the sensors (vs the machines they sense) and (b) the rarity of certain kinds of failures: frequencies of 1 in 10,000,000 leave many readings of normal status as their own type of noise. In any event, the point here is that bigger does not necessarily mean better when it comes to data.

Accordingly, big data skills cannot be purely a matter of computer science, statistics, or other processes. Instead, the backstory behind the creation of any given data point, category, or artifact can be critically important and more complex given the nature of the environments being sensed. While the same algorithm or statistical transformation might be indicated in a bioscience, a water main, and a financial scenario, knowing the math is rarely sufficient. Having the industry background to know where variance is "normal," for instance, comes only from a holistic understanding of the process under the microscope. As we move into unprecedented data volumes (outside the Large Hadron Collider perhaps), understanding the ground truth of the data being collected and the methods of its collection, automated and remote though they may be, will pose a significant challenge.

Beyond the level of the device, data processing is being faced with new challenges—in both scope and kind—as agencies, companies, and NGOs (to name but three interested parties) try to figure out how to handle billions of cellphone chirps, remote-control clicks, or GPS traces. What information can and should be collected? By what entity? With what safeguards? For how long? At what level of aggregation, anonymization, and detail? With devices and people opting in or opting out? Who is allowed to see what data at what stage in the analysis life cycle? For a time, both Google (in its corporate lobby) and Dogpile (on the web) displayed real-time searches, which were entertaining, revealing, and on the whole discouraging: porn constituted a huge percentage of the volume. Will ski-lift webcams go the same way in the name of privacy?

Once information is collected, the statistical and computer science disciplines are challenged to find patterns that are not coincidence, predictions that can be validated, and insights available in no other way. Numbers rarely speak for themselves, and the context for Internet of Things data is often difficult to obtain or manage given the wide variety of data types in play. The more inclusive the model, however, the more noise is introduced and must be managed. And the scale of this information is nearly impossible to fathom: according to IBM Chief Scientist Jeff Jonas, mobile devices in the United States alone generated 600 billion geo-tagged transactions every day—as of 2010.[15] Finally, the discipline of statistics is being forced to analyze these vast bodies of data in near real time—and sometimes within seconds—given how many sensors have implications for human safety and well-being.

In addition to the basic design criteria, the privacy issues cannot be ignored. Here, the history of Google Glass might be instructive: whatever the benefits that accrue to the user, the rights of those being scanned, identified, recorded, or searched matter in ways that Google has yet to acknowledge. Magnify Glass to the city or nation-state level (recall that England has an estimated 6 million video cameras, but nobody knows exactly how many[16]), as the NSA revelations appear to do, and it's clear that technological capability has far outrun the formal and informal rules that govern social life in civil society.

In sum, data from the Internet of Things will challenge both the technical capabilities and the cultural codes of practice of the data community: unlike other categories of big data, people's faces, physical movements, and public infrastructures define much of their identity and well-being. The analytics of these things becomes something akin to medicine in the gravity of its consequences: perhaps the numbers attached to the IoT should be referred to a "serious data" rather than merely being another category of "big."

[15] Marshall Kirkpatrick, "Meet the Firehose Seven Thousand Times Bigger than Twitter's," Readwrite 18 November 2010, http://readwrite.com/2010/11/18/meet_the_firehose_seven_thousand_times_bigger_than#awesm =∼oIpBFuWjKFAKf9, accessed 29 March 2018.

[16] David Barrett, "One surveillance camera for every 11 people in Britain, says CCTV survey," The Telegraph, 10 July 2013, https://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html, accessed 29 March 2018.