Chapter 14

DATA OVERFLOW

What's the Trend?

The combination of growing personal and corporate owned data mixed with open data creates an new challenge to go beyond algorithms to manage this data and rely instead on better artificial intelligence, smarter curation, and more startup investment.

Big data is getting bigger, small data is getting smaller – and the intersection of both is starting to cause some serious problems.

Despite the continual business conversation on big data and how to collect more information, last year I introduced the idea of "Small Data" to capture the growing practice of consumers using their *own* data collected by smart devices and social platforms and learning to leverage it for better service, pricing or outcomes. The future, I argued, would belong to the brands that could find ways to entice consumers to share this *small data* and combine it with the big data that they were already collecting.

In the coming year, this blend of big and small data will lead to a new related problem of *Data Overflow* – where any company collecting data will struggle with the fact that they will soon be buried by its sheer volume. On first glance, it may seem you've heard about this overflow before.

Sometimes termed "data overload" – much has been written about the dangers of collecting too much data or the futile journey individuals and organizations sometimes take to make sense of it once they have it. The trend of Data Overflow describes the next level of this chaos, where the intersection of self-collected consumer "small data" and corporate collected "big data" will be rendered even more confusing by a third collection of frequently useless "open data" often lacking structure or metadata and dumped online publicly by companies and governments in the name of transparency or regulatory compliance.

As one example of this rapidly growing data deluge, the GovLab Index tracks open data trends and publishes annual reports on the state of open data adoption by governments worldwide. In the most recent update, the Index report included some sobering data points:

- Over 1 million datasets have been made open by governments worldwide.
- Less than 7% of these datasets are published in both machine-readable forms and under open licenses.
- 96% of countries are sharing datasets that are not regularly updated.
- In 2015, there were nearly 400 open government data portals worldwide (up from only two just six years earlier).

While it is clear that the volume of this open data is growing exponentially year after year, the problem is that much of it may not be that inherently useful. In an interview with *The Economist*, Joel Gurin of the Centre for Open Data Enterprise estimates that perhaps four fifths of the data that has been released is *not* particularly useful due to missing the metadata or consistent standardization that would provide much needed context.

So how can we best tackle this disarray of data? And is there a cure for data overflow? To find out, let's explore the story of how one group of scientists are trying to solve a piece of challenge … from inside of one of the most dangerous and scientifically advanced locations on Earth.

Deciphering Super Collider Data

The Large Hadron Collider (LHC) is like an experiment straight from the pages of science fiction. The largest single machine in the world, the LHC was built by the European Organization for Nuclear Research (CERN) in collaboration with 10,000 scientists from 100 countries and sits in a 17 mile long tunnel underneath the France-Switzerland border. Its purpose is to allow physicists to test how particles collide and further our understanding of the physical world.

The main challenge of this technological feat is not making the collisions happen, but rather how to decode the terabytes of data the collisions generate. The volume is far too high for even the most sophisticated algorithms and too nuanced for even the smartest scientists using the fastest computers. And it turns out the team at the LHC is not the only group of scientists facing this problem.

In an article in late 2015 from the journal eLife, for example, cell biologist Robert Insall sparked an industry wide debate when he shared that the majority of senior researchers he interviewed feared a crushing volume of new biomedical research papers released each year was leading to a decline in trustworthiness for their field overall.

In the past year, both Google and IBM made big announcements to promote more development on their AI platforms (TensorFlow and Watson, respectively) by creating more open source libraries, tools and tutorials for the technical community. The aim of each is to inspire more developers to create initiatives using "deep learning," an aspect of machine learning that describes the quest to make machines more intuitive and closer to true artificial intelligence.

Over time, AI and deep learning can automatically go through a glut of research or questionably valuable open data and help make the connections that lead to real discovery. AI holds the promise to give the scientists at the LHC and the overwhelmed biomedical researchers help in their quest to make sense of all the data.

As the world of science tests new AI-driven solutions to this data challenge, their example will be watched by many other global industries – including several that are suffering from a data overflow of their own.

AgTech: Growing Data For Farmers

The agriculture industry offers the perfect example of some on the ground challenges that data overflow can cause. The industry is overrun with data. A single farm can now provide mountains of data from sensors in the soil, wearable trackers on farm animals and aerial drones for crop monitoring.

Drones in particular hold great promise and peril for the farming industry, as the Association for Unmanned Vehicle Systems International estimates that agriculture drones may make up 80 percent of the future commercial market.

All the data generated by this agricultural technology (often called "AgTech") is ushering in a new evolution of what is increasingly being described as precision farming – the ability to plant the right crop in the right spot and harvest it at the right time. Consistently achieving this precision isn't easy … and doing it takes an integration of technology into farming on a scale that has rarely been tried, or even discussed.

One of the biggest chances to have that discussion came in July of 2015 at the AgTech Summit in Salinas, California hosted by Forbes. The event was aimed at bringing agriculture industry insiders together with Silicon Valley innovators to talk about the future of the farming industry and how technology could play a role. At the event, real farmers shared the one big challenge they faced from all this new data – how to make meaning from it all.

Data was overflowing and farmers on the ground were finding it nearly impossible to make sense of it all.

Farmers are usually not scientists with PhDs in advanced particle physics trained to parse and analyze data. They are rarely doing their day's work from a comfortable chair in front of a laptop. They are mobile, savvy, time strapped and usually need data in a format that they can act on immediately.

Most agree that the solution will come not from new innovative tools to collect even more data, but by putting technologists in the same room as farmers and encouraging them to create more innovation that is actually useful at solving the data overflow. Thanks to the AgTech Summit

and events like it, there are more initiatives than ever focused on tackling this challenge.

In the agriculture driven nation of New Zealand, for example, a local business incubator has launched a 20 week accelerator program for agtech called Sprout Agritech. AgTech initiatives across the world are seeing more investment as well. By the end of 2014, industry site AgFunder tracked over $2.36 billion of investment across 264 deals – which outpaces sectors with far more hype such as financial technology ($2.1 billion) and clean technology ($2 billion).

Meanwhile many of the open data sets released by governments around the world relate to data that is highly valuable for farmers – such as regional weather data and data on regional food consumption. As this public data intersects with traditional farm data (like crop yields and soil metrics) and nontraditional new farm data (like drone based measurements) – the same data overflow challenge common in other industries is hitting agriculture as well.

Yet noted venture capitalist Randy Komisar believes this is going to inspire a golden age of investment and innovation in agtech. As a partner at noted venture capitalist firm Kleiner Perkins Caufield & Byers, Komisar is used to looking farther into the future than most. In late 2015 he shared some of his insights in an interview with *National Geographic* – predicting a more open future for agriculture where farmers could take back control from the biggest players that run the agricultural industry with a near monopoly today. This would enable them to remove themselves from being mere vendors to the big agriculture companies and instead own and share their own data more openly with one another.

In farming – innovators and new startups companies funded by growing investment are the hope for solving data overflow and creating more meaning from the intersection of data from all directions.

Curated Health Data

In the health sector, privacy is always the most important concern when it comes to sharing any data online. The first topic innovators tackle, regulators ask and patients worry about is how individual data will be used.

The health industry, though, faces the exact same data overflow issues as other industries profiled in this chapter … fueled by the same set of circumstances.

Hospitals collect data on patients and outcomes. Patients collect their own data through wearable fitness trackers and technology designed to help them manage conditions they may have such as diabetes or asthma. And, of course, there is plenty of public health data launched openly onto the Internet by governments for anyone to access.

Rather than turning primarily to AI as scientific researchers have, or inspiring more startups to tackle the challenge as in the agricultural industry – healthcare is poised to see an explosion in a different coping strategy inspired mainly by individual curators and the power of the human eye to see and solve problems that algorithms of even artificial intelligence might still struggle with.

One of the most powerful examples of this process comes from an app often described as "Instagram for doctors" called Figure1. It is a simple app that allows medical professionals to share anonymized images of patients struggling with some type of condition in order to get feedback and comments from other medical professionals. It turns out this behavior of sharing images of patient afflictions without personal details is already surprisingly common.

When Figure1 founder Josh Landy, an intensive care specialist at Scarborough Hospital in Toronto, Canada, first developed the idea his ambition was to take what he was seeing people do already in terms of sharing images individually, and transform it into "a global knowledge notebook." Today the app has more than 150,000 active users, is perfectly tailored for visual learners as many doctors happen to be, and was described by one third year medical resident in Texas as her "medical guilty pleasure."

The app offers far more than a chance for medical voyeurism though. It is also using the power of the crowd to help medical practitioners gain insight and support from those in other parts of the world to diagnose and treat conditions that may be common in one place but rare in another. In this case the data and the usefulness of it stems from a person to person interaction on a human level.

When it comes to huge data sets shared more publicly, though, the challenge gets different.

Cedric Hutchings is the co-Founder and CEO of Withings – a company that makes a suite of wearable tech and quantified self products. He is also a believer in the power of anonymous data to be useful on a regional and global level to incite change.

For the past several years, Hutchings has been driving his company to make more use of the aggregated anonymous data from all the users that they collect. Thanks to this data, he has been able to pinpoint, for example, that a tiny city on the outskirts of Paris called Argenteuil has the notorious distinction of being the most obese city – and communications around this in 2015 motivated the mayor and people of the city to issue a comprehensive plan to change school lunches and lose that dubious distinction.

Not surprisingly, Hutchings believes in the power of open data and has committed his organization to its own open data initiative, called the Withings Health Observatory. More significantly, in order to make meaning from this dataset instead of just dumping it online as many have done, Withings committed resources to helping make sense of the data.

The brand now creates content featuring key statistics, insights, and community reports. It also partners with researchers and academics to publish scientific papers through the Withings Health Institute and now offers this content as a toolkit to help researchers and other scientists to make sense of their data and to make it more valuable for everyone it is shared with.

Why It Matters

In 2016, the intersection of big data, small data and open data led to the consistently big challenge of Data Overflow. Rather than relying on algorithms alone – industry sectors from scientific research to agriculture to healthcare are all applying a slight different lens in order to solve this challenge. For some, artificial intelligence is the only viable solution to handle volumes of data and make meaning from it in any sort of valuable way. For others like agriculture, the challenge is taking pockets of

successful data analysis and merging them together into an ecosystem that helps the farmer on the ground. Investment in startups and innovation is taking off in an attempt to make this happen. And for healthcare, the solution is coming down to a rise in the art of human curation as a method to make meaning from hard to decipher visual data and data creators taking more responsibility to add context and meaning to this data before sharing it openly.

Who Should Use This Trend?

The potential and opportunity for data collection is blossoming in all types of industries, which means this trend is likely to affect almost anyone doing business in 2016. As it does, data overflow will challenge the potential value of collecting the data in the first place. The good news is that this is a highly visible problem and one that many different groups are trying to solve. Moving into 2016, industries are likely to develop their own individual solutions to this problem and the most savvy brands will be the ones keeping up to date with the latest solutions and investing time and effort to make them work for themselves.

How To Use This Trend

- ✓ **Learn data literacy** – The sad reality of business is that most of us lack basic data literacy skills. As a result, we misread statistics, misquote studies and draw incorrect or sometimes idiotic conclusions from data that is at best inconclusive. The only solution to this problem is to do something that most of us probably don't want to do … go back and improve our data literacy. This might involve taking an online class, or reading articles from those well versed in how to interpret data about a particular topic or from a particular industry.
- ✓ **Curate open data** – Central to this trend are the issues and challenges raised by a growing amount of open data being launched into the marketplace. One solution increasingly

common within healthcare, but potentially useful for any industry, is to get better at curating this data and finding the valuable pockets of information yourself and using them to sharing them with others to generate more value back to your own efforts.