Bridging the Cyber-Analysis Gap: The Democratization of Data Science

John Healy
Leland McInnes
Colin Weir

ABSTRACT

The challenges of ever growing and ever changing Big Data are broad and far-reaching, particularly in the cyber-defense domain. The task of analyzing and making sense of this data is difficult, and only getting worse. We propose that by democratizing data science and making it accessible to everyone, we can expand the breadth and depth of analytics available to a point where we can potentially meet the challenges of Big Data.

THE ONCOMING WAVE OF BIG DATA

As computing and sensor facilities become ever more pervasive and interconnected, the amount and diversity of data available to cyber-analysts continue to grow at an exponential rate. The cyber-analyst's ability to analyze and leverage all the data is currently lacking. The tools available to analysts do not scale to the volumes of data we have now, let alone the volumes which are expected in the future. This is a significant challenge that must be addressed by current and future cyber-defense organizations.

To face the challenges of ever growing data, the data science community needs to empower cyber-analysts with more intelligent tools. We need tools that provide complex, nuanced analyses and intelligent summarizations of the data. Such tools are the domain of machine learning and data science. This growing field can provide powerful models for analyzing data. For example, analyses capable of automatically determining the state actor behind a newly discovered piece of malware, or tools capable of automatically detecting and blocking malicious web traffic never previously identified. Since analytic tasks are incredibly diverse and always changing to respond to new data, the cyber-analyst community needs machine learning tools that are general and flexible enough to cope with this evolving diversity. This might appear to require something more general than the narrow intelligence that traditional machine learning provides. We claim that this is not the case.

John Healy, Leland McInnes and Colin Weir are senior researchers at the Tutte Institute for Mathematics and Computing in Ottawa, Canada. The Tutte Institute brings together leading researchers from academia, industry and government to solve the most challenging classified and unclassified problems facing the security and intelligence community. The authors exemplify this cross-disciplinary approach, with diverse backgrounds in pure mathematics, computer science, statistics, machine learning, and cyber-security. Combined, the authors have eight advanced degrees and over 40 years of research experience spanning the gamut of government, industrial, and academic positions.

The functionalist school of philosophy of mind holds that human intelligence is merely the composition of a vast array of specialist systems [Den92, Den98, Den06]. This composite has no central point of control but is instead a swarm of specialist systems that are continuously co-operating, competing, and interacting. Such a system potentially provides the generality and flexibility of human intelligence without ever having a singular general intelligence. This provides a compelling analogy for how many interacting specialists models can come to provide a whole that is greater than the sum of their parts. We propose that the solution to the ever changing diversity of data lies in a vast army of models. Each model can be highly specialized but, with enough interacting models, a great diversity of tasks can be easily accomplished. The ecosystem of models is *smarter* than any individual model. This is "the wisdom of the crowd" writ nanoscale ([Gal07]).

Currently, building computational predictive models is the domain of machine learning experts. This is a bottleneck on model construction and on deployment to cyber-analysts. Most tellingly, it puts significant constraints on the latency of cyber-analyst feedback for improving or specializing models. In a world where every analyst can build their own machine learning models, specialized to their own needs, this feedback loop is dramatically tightened. A model can efficiently be tailored to each analyst's specific needs, offloading cognitive tasks to the machine, and a small army of analytic models can quickly be promulgated among analysts in a shared collaborative workspace. With such a system of co-operating, competing and interacting models, the system-as-a-whole begins to resemble functional machine intelligence.

In this view, our goal is not to build ever more complex and general models. Instead, the answer

lies in democratizing machine learning and making it pervasive. The Internet changed the world by democratizing data generation: data was no longer sequestered in carefully controlled databases, but generated by everyone, everywhere, all the time. To build the dynamic ecology of machine learning models that we propose, we need to democratize data analytics and put the power of machine learning directly in the hands of analysts. Thus, the question we should be asking is "how can we transform the technological landscape to make machine learning and data science ubiquitous?"

TRANSFORMATIVE TECHNOLOGIES AND THE END OF DATA SCIENCE

Simple technologies can have remarkable transformative powers, fundamentally changing the landscape of ideas. A simple example of this kind of subtle revolution is the development of spreadsheet software. The first real spreadsheet program was VisiCalc, developed by Dan Bricklin and Bob Frankston in 1979. The concept was simple, elegant and seems, on reflection, obvious: allow data to be entered in rows and columns and allow arithmetic formulas to be computed across those rows and columns. Most importantly, if a data entry is changed, then the change should propagate through the formulas and be instantly visible to the user. VisiCalc was an instant success and became a driving factor in the rise of personal computers: for many buyers, VisiCalc was the motivating reason to purchase a computer! By the mid-1980s spreadsheets were everywhere, and the market was dominated by Lotus 1-2-3, which integrated charting and plotting to spreadsheets. Spreadsheets became so central that, on the first release of Microsoft Windows, Excel was the flagship product designed to draw users to the fledgling operating system ([Pow04]).

To face the challenges of ever growing data, the data science community needs to empower cyber-analysts with more intelligent tools.

What was the change that spreadsheets spurred? They powered the first data revolution. As long as data lived in carefully curated databases on distant mainframes, it remained sparse. Once it became possible to create and work with data locally and visually via spreadsheets, the amount of data generated exploded. Spreadsheets made it possible for everyone to work with data, and so everyone did. Data was entered, plotted, linked and transformed on a scale never seen before. In short, spreadsheets changed the very way people look at and think about data—it became something that everyone has, and everyone can use. The expansion of data enabled by spreadsheets was but the first ripples of the oncoming wave of Big Data. With the democratization of data generation provided by the Internet, data has grown far beyond the analytic power of spreadsheets, and we are only seeing the early warning signs of a wave of data that threatens to wash us away.

With the arrival of internet-connected sensor networks and the Internet of Things, both the cyberattack surface and associated data will quickly grow far beyond our limited data-analytic capabilities. We need a second data revolution. Spreadsheets provided arithmetic analytics and visualization; today the cyber-analyst community needs radically new ways to summarize the immense volumes of data at their disposal—the next wave in the data revolution appears to be driven by machine learning analytics.

The second data revolution will arrive with new tools and new transformative technologies. The keys to the success of spreadsheets were their low barrier to entry, their remarkable versatility, and the powerful tools that could be built with them. With data now beyond the scope and capability of spreadsheets there is a need for new tools that ask a little more of the user, but exponentially increase versatility and analytic power. When spreadsheets were unleashed upon the world computers were a foreign concept to most, so the extremely straightforward and visual interface provided by VisiCalc was critical. Now, however, computing is pervasive, and with movements like code.org [Cod15] and President Obama's "Computer Science for All" initiative [Smi16], basic programming skills are rapidly becoming part of the mandatory curriculum ([Nor16], [Wat16]). We no longer need to assume users are unable to cope with simple programming tasks, and this opens up a vast untapped wealth of flexibility and power. Under this rubric, the transformative tools that are needed are already on the horizon.

Much of the open-source community, faced with the requirements for Big Data analytics, is consolidating on infrastructure to power the human-computer interface for data analytics. For example, the Jupyter notebook interface ([Jup16c]) provides rich tools for interactive programming. Notebooks are living interactive documents that contain explanatory text, live code, visualization, and rich visual display of interactive content. The versatility of the system is incredible ([GP15, She14, Jup16a, Jup16b]) while still providing a simple and intuitive interface. From the convergence of the pervasive programming skills of the coming generation and the powerful visual interface of tools such as Jupyter, the cyber-analyst community can expect a transformation of analysis tools from the outdated and ill-equipped to a shared collaborative ecosystem of living notebooks.

If tools similar to Jupyter provides the surface interface, what can provide the substrate? Python is the lingua franca of data science and machine learning.[1] It has spawned a growing ecosystem of data analytics and machine learning tooling built upon it (including Jupyter itself). This is an open-source ecosystem, and, in the spirit of the source language, focused on intuitive ease of use ([Pet04, Oli15]). The result is not a product, but a collaboratively built platform: data science tools by the masses for the masses. This is the democratization of data analytics underway as we speak.

In short, the second data revolution is almost upon us. Powered by machine learning, soon to be accessible to all in a vast collaborative workspace of notebooks, the cyber-data

challenges of the future will be tamed—not by specialist data scientists, but by shared efforts of ordinary analysts, newly empowered by transformative tools. Data science will become so pervasive, so ingrained in every mind that it will cease to exist as a separate concept. Much like the spreadsheet, we won't be able to imagine a world without it.

THE WHOLE IS MORE THAN THE SUM OF ITS PARTS

In 2001, a small upstart encyclopedia arrived to challenge the reign of *Encyclopedia Britannica.* At the time, *Wikipedia* seemed like a toy project, without any of the expert research and editing staff available to a giant like *Britannica.* Instead, *Wikipedia* has come to completely eclipse any other encyclopedia on the planet for both the breadth and depth of knowledge that it successfully captures and presents. It achieved this remarkable feat by democratizing the task of compiling human knowledge through the wisdom of the crowd. Anyone with access to a computer can use and edit Wikipedia. The feedback loop is swift—if you see something wrong or that can improve you can edit it immediately and see the results. Better still, everyone else also immediately sees the results of your edit and can adapt it, comment on it, or revert it. With enough people making edits, the text slowly but surely lurches its way toward a consensus description of the topic at hand. No single edit is necessarily *right*, nor final. The result is something better than any of its individual authors may have produced. In short, *Wikipedia* is more than the sum of its edits.

At the Chesapeake Large Scale Analytics Conference, a survey of attendees demonstrated that the expected time-frame for delivery of a new predictive analytic model to production was three months and could often be as long as a year or more. That represents a delay of months, or even a year before front-end analysts can evaluate the usefulness of the model on current, real-world data. For problems that remain relatively stable over time, this may be a reasonable approach. In the dynamic adversarial world of cyber-defense, such a delay is potentially devastating. Dramatically shrinking the cyber-analyst feedback loop on models and enabling a *fail-fast* approach is critical to the wider success of machine learning in cyber-analytics. To do this, we need to embrace the democratizing approach and rapid feedback that made *Wikipedia* so successful.

> In this view, our goal is not to build ever more complex and general models. Instead, the answer lies in democratizing machine learning and making it pervasive.

As we have already seen, Jupyter and Python provide a powerful infrastructure for collaborative data science for analysts. Furthermore, with robust machine learning tools the data science community can empower cyber-analysts to make use of state of the art machine learning. Bringing all of this together in a shared collaborative workspace can

enable analysts to co-operatively develop machine learning models and analytics. The result of this confluence of technologies is an open and flexible ecosystem that can evolve and grow with analysts' needs. This will require further development of the software infrastructure, however, with sufficient work, it can become the *Wikipedia* of data analytics, with a breadth and depth of models and analysis that eclipses anything that has come before. It can be an analytic platform that is far more than the sum of its models.

MACHINE LEARNING THAT JUST WORKS

Cars have existed, in various forms, since the late 18th century [Eck01]. Despite their long history, it wasn't until the 20th century that cars became the transforming societal force that they are today. The catalyst for that transition was the introduction of the Ford Model T. In the year that the Model T was introduced the world land speed record was held by a car—a steam powered car. This was the car technology of the 18th century with literally centuries of steady improvements and refinements creating a finely tuned, precision engineered racing machine. Ford's genius was realizing that car design had been solving the wrong problem. High end, high-performance cars were both expensive, and temperamental. In contrast, the Model T was not designed to be the best, nor fastest, car, but a car that was inexpensive and reliable. By making the car available to everyone Ford democratized personal transport, and in so doing disrupted the entire industry and changed the world.

Simple technologies can have remarkable transformative powers, fundamentally changing the landscape of ideas.

Machine learning has been around since the 1960s and has made many remarkable advances in that time. More recently machine learning has become a competition; from the KDD Cup ([KDD16]) and the Netflix Prize ([Net07]) to the ImageNet Challenge ([Ima15]) and Kaggle ([Kag16]). The metric for all these competitions is model accuracy. Model accuracy is the land speed record of machine learning. The models produced are near miraculous in their accuracy, but are also extremely complex and intricate, requiring considerable expertise to build and maintain. What is needed for today's analysts are machine learning tools that make model construction inexpensive and reliable—without necessarily optimizing solely for model accuracy. Simple, robust models would bring the power of machine learning to the masses. This is a different approach to designing machine learning tools and algorithms, and deserves significant research effort—since the result, the democratization of machine learning will be as revolutionary as the democratization of transport enabled by Ford's Model T.

Inexpensive, robust models are also required for machine learning in production environments. In 2014 Google published a highly influential paper titled *Machine Learning: The High Interest Credit Card of Technical Debt* [SHG + 14]. The primary thesis was that while machine learning was extremely powerful and could bring quick wins, it could also prove to be a maintenance nightmare. This was predicated on intricate traditional machine learning models which, due to their expert tuning and calibration, were hard to modify or update. On the other hand, if democratized robust models are used the problem evaporates. Models that are inexpensive to build are disposable—this accumulates very little debt, it is paid down by simply building a new model. We are even beginning to see such thinking taking hold in practice: the winning Netflix Prize entry was not implemented at Netflix due to its complexity and the vast amount of delicate hand tuning—a much simpler to maintain a model that was mere fractions of a percentage point less accurate was deemed to be the most effective solution.

The move from complex traditional models to simple practical models can be achieved by a directed research program on techniques for robust models. The foundation for such a research program is already beginning to take shape. The generalized low-rank model framework ([UHZB15]) from Stanford provides a powerful and general framework for automated feature engineering. Random Forest models ([Bre01]) provide classification models that 'just work'. Recent advances in clustering ([CMAS15, CM10]) show promise for robust unsupervised learning, including anomaly detection. Neural network motivated techniques such as word2vec and GloVe ([MCCD13, PSM14]) offer a foundation for research into text analytics for the masses. Building upon this work to fill out a complete set of machine learning tools that *just work* will bring robust models to the heart of machine learning research.

> We claim that the resulting increase in both the scope and the power of analytics can meet the challenges of the ever-growing and ever changing data landscape of cyber-analytics.

An immediate proposal for such democratization might look like a shared ecosystem of Jupyter notebooks overtop of Python and its suite of rapidly developing tools. A small cadre of more technologically literate cyber-analysts could be trained with minimal effort to be able to leverage the machine learning models of data science in their everyday work. In the longer term, research into more intuitive and powerful techniques and languages along with an increase in general programming literacy may alter this framework and help to both empower and reduce the cognitive load upon the broader analyst community.

Ultimately our proposal is to bring ordinary analysts and machine learning closer together. This involves trained cyber-analysts working with machine learning techniques designed specifically for cyber-analysts, bridging the gap by bringing each closer to the other. Closing this gap remarkably expands the user base of machine learning and data science, and shrinks the feedback loop allowing rapid evolution of models and analytics. In turn, this is a catalyst creating an ever-growing breadth and depth of analytic capabilities. We claim that the resulting increase in both the scope and the power of analytics can meet the challenges of the ever-growing and ever changing data landscape of cyber-analytics.

NOTES

1. Other languages such as R and Julia compete in this space, but currently the momentum is with Python in the machine learning (as opposed to general statistics) fields – see scikit-learn and tensorflow for examples.

NOTES

[Bre01] Leo Breiman, Random forests, *Machine Learning,* 2001.

[CM10] Gunnar Carlsson and Facundo Memoli, Multiparameter hierarchical clustering methods, *In Classification as a Tool for Research,* 63–70, 2010.

[CMAS15] R.J.G.B Campello, D. Moulavi, A.Zimek, and J. Sander, *Hierarchical density estimates for data clustering,* ACM Transactions on Knowledge Discovery, 1–51, 2015.

[Cod15] Code.org https://www.code.org, 2016.

[Den92] Daniel C. Dennett, *Consciousness Explained,* Back Bay, 1992.

[Den98] Daniel C. Dennett, *Brainchildren: Essays on Designing Minds,* Bradford, 1998.

[Den06] Daniel C. Dennett, *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness,* Bradford, 2006.

[Eck01] Erik Eckermann, *World History of the Automobile,* 2001.

[Gal07] Francis Galton, *Vox Populi,* Nature, 75, 450–451.

[GP15] Brian Granger and Fernando Perez, *Computational Narratives as the Engine of Collaborative Data Science,* https://archive.ipython.org/JupyterGrantNarrative-2015.pdf, 2015.

[Ima15] ImageNet Challenge, https://www.image-net.org/challenges/LSVRC, 2015.

[Jup16a] Jupyter Dashboards, https://github.com/jupyter-incubator/dashboards, 2016.

[Jup16b] Jupyter Kernel Gateway Bundlers, https://github.com/jupyter-incubator/kernel_gateway_bundlers, 2016.

[Jup16c] Jupyter Project, https://www.jupyter.org, 2016.

[Kag16] Kaggle, https://www.kaggle.com, 2016.

[KDD16] KDD Cup Archives, https://www.kdd.org/kdd-cup, 2016.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffery Dean, *Efficient estimation of word representations in vector space,* arXiv, 2013.

[Net07] Netflix Prize, https://www.netflixprize.com, 2007.

[Nor16] Anna North, *Should We All Learn to Code,* http://op-talk.blogs.nytimes.com/2014/06/17/should-we-all-learn-to-code/?_r=0, 2016.

[Oli15] Travis Oliphant, *Python as the Zen of Data Science,* http://youtube.com/watch?v=mNvPiV37F7Q, 2015.

[Pet04] Tim Peters, *The Zen of Python,* http://www.thezenofpython.com, 2004.

[Pow04] Power, D. J., "A Brief History of Spreadsheets", DSSResources.COM, http://dssresources.com/history/sshistory.html, version 3.6, 08/30/2004.

[PSM14] Jeffery Pennington, Richard Socher, and Christopher D. Manning, Glove: *Global vectors for word representation,* In Empirical Methods in Natural Language Processing, 1532–1543, 2014.

[She14] Helen Shen, *Interactive notebooks: Sharing the code,* Nature, 515:151–152, 2014.

[SHG + 14] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young, *Machine learning: The high interest credit card of technical debt,* In SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop), 2014.

[Smi16] Megan Smith, *Computer Science for All,* https://www.whitehouse.gov/blog/2016/01/30/computer-science-all, 2016.

[UHZB15] Madeleine Udell, Corrine Horn, Reza Zadeh, and Stephen Boyd, *Generalized low rank models,* arXiv, 2015.

[Wat16] Jackie Wattles, *GE CEO Jeff Immelt says all new hires will learn to code,* http://money.cnn.com/2016/08/04/technology/general-electric-coding-jeff-immelt/, 2016.