O'REILLY®

Compliments of ANACONDA Powered by Continuum Analytics

Breaking Data Science Open

How Open Data Science is Eating the World

Michele Chambers
& Christine Doig

Breaking Data Science Open

How Open Data Science is Eating the World

This Preview Edition of *Breaking Data Science Open* is a work in progress. The final version is expected to publish in the Fall of 2016.

Michele Chambers and Christine Doig

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Breaking Data Science Open
by Michele Chambers and Christine Doig

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com .

Editor: Tim McGovern **Cover Designer:** Karen Montgomery
Interior Designer: David Futato

September 2016: First Edition

Revision History for the First Edition
2016-09-01: First Release

978-1-491-97299-1

[LSI]

Table of Contents

Breaking Data Science Open: How Open Data Science is Eating the World 1
How Data Science Entered Everyday Business 2
Modern Data Science Teams 5
Data Science for All 7
Open Data Science Applications 13
The Journey to Open Data Science 16
Conclusion 20

Breaking Data Science Open: How Open Data Science is Eating the World

Data science has burst onto the public's attention over the past few years as perhaps the hottest and most lucrative technology field. No longer just a buzzword for advanced analytical software, data science is poised to change everything about an organization: its potential customers, its expansion plans, its engineering and manufacturing process, how it chooses and interacts with suppliers, and more. The leading edge of this tsunami is a combination of innovative business and technology trends that promise a more intelligent future based on the combination of open source software and cross-organizational collaboration called *Open Data Science*. Open Data Science is a movement that makes the open source tools of data science - data, analytics, and computation - work together as a connected ecosystem.

The Open Data Science wave is cresting now because of a happy confluence of trends. The ongoing consumerization of technology has brought open source to the forefront, creating a virtual free market where innovation quickly emerges and is vetted by millions of demanding users worldwide. These users industrialize products faster than any commercial technology company could possibly accomplish. On top of this, the Agile trend fosters rapid experimentation and prototyping that prompt modern data science teams to constantly generate and test new hypotheses, discarding many ideas and quickly arriving at the top 1% of ideas that can generate value and are worth pursuing. Agile, in turn, has led to the fusing of devel-

opment and operations into DevOps, where the top ideas are quickly pushed into production deployment to reap value. This continuous cycle of innovation requires that modern data science teams utilize a continuously evolving set of open source innovations to continually add higher levels of value without recreating the wheel.

This report discusses the evolution of data science including open data science, data science collaboration, self-service data science, and data science deployment. Continuum Analytics, the driving force behind Anaconda, the leading Open Data Science platform powered by Python, is the sponsor of this report.

How Data Science Entered Everyday Business

Business intelligence, which takes historical data and summarizes it into charts, has been evolving for decades as data has become cheaper, easier to get, and easier to share. The outputs of business intelligence are "known knowns" that are manifested in stand-alone reports examined by a single business analyst or shared among a few managers. *Predictive analytics* has been unfolding on a parallel track to business intelligence. With predictive analytics, numerous tools allow analysts to gain insight into "known unknowns," such as where their future competitors will come from. These tools track trends and make predictions, often within specialized programs designed for statisticians and mathematicians.

Data science is a multi-disciplinary field that fuses the latest innovations in advanced analytics, including machine learning and deep learning, combined with high performance computing and visualizations. The tools of data science originated in the scientific community, where researchers used them to verify hypotheses that include "unknown unknowns," and they have entered business, government, and other organizations gradually over the past decade as computing costs have shrunk and software has grown in sophistication. The finance industry was an early adopter of data science. Now it is a mainstay of retailers, city planners, and—particularly salient in this US election year—political campaigns.

Data science is a significant breakthrough from traditional business intelligence and predictive analytics. It brings in data orders of magnitude larger than what previous generations of data warehouses can store, and even works sometimes on streaming data. The analytical tools used in data science are also increasingly powerful, using artifi-

cial intelligence techniques to identify hidden patterns in data and pull new insights out of data. The visualization tools used in data science leverage modern web technologies to deliver interactive browser based applications. These are not only visually stunning, they provide rich context and relevance to the consumers of intelligent applications. Some of the changes driving the wider use of data science include:

- Open Data Science - Open source communities want to break free from the shackles of proprietary tools and embrace a more open and collaborative work style that reflects that way they work with their teams all over the world. These communities are not just creating new tools; they're calling on enterprises to use the right tools for the problem at hand. Increasingly that's a wide array of programming languages, analytic techniques, analytic libraries, visualizations, and computing infrastructure. Popular tools for Open Data Science include the R programming language, which provides one-line functions to run statistical tests, and Python as a quick-to-learn, fast prototyping language that can easily be deployed into production as it integrates with existing systems easily. Both of these languages have thousands of analytics libraries that deliver everything from basic statistics to linear algebra, machine learning, deep learning, simulation, and genetic algorithms used to address complexity and uncertainty. Additionally, powerful visualization libraries range from basic plotting to fully interactive browser based visualizations that scale to billions of points.
- Data science collaboration - The much-sought-after unicorn data scientist who understands everything about algorithms, data collection, programming, and the business you're in, exists, but more often it's modern, collaborating data science teams that get the job done for enterprises. Modern data science teams are a composite of skills represented by the unicorn data scientist and work in multiple areas of a business. These teams have varied skills and backgrounds that range from databases, statistics, development, ETL, high performance computing, Hadoop, and open source to subject matter expertise, business intelligence, and visualization. Data science collaboration tools facilitate workflows and interactions, typically based on an agile methodology, so that work seamlessly flows between various team members. This highly interactive workflow helps teams

progressively build and validate early stage proof-of-concepts and prototypes while moving towards production deployments.

- Self-service data science - While predictive analytics was relegated to the backoffice and developed by mathematical purists, data science has empowered entire data science teams including frontliners—often referred to as *citizen data scientists*--with intelligent applications and ubiquitous tools that are familiar to business people, using spreadsheet- and browser-based interfaces. With these powerful applications and tools, citizen data scientists can now perform their own analyses and predictions immediately to make evidence-based decisions regularly.
- Data science deployment - In the past, technology and cost barriers prevented many predictive analytics from moving into production. Today, with Open Data Science, both of these barriers are significantly reduced and there is a rise of intelligent applications and intelligence embedded into devices and legacy applications.

What do the new data science capabilities mean for business users? Businesses are continually seeking competitive advantage, where there are a multitude of ways to use data and intelligence to underpin strategic, operational, and execution practices. Business users today, especially with millennials entering the workforce, expect a Siri-like intelligent personalized experience that can help them create value for their organization.

Data science is, in short, driving innovation by arming everyone in an organization - from frontline employees to the board - with intelligence that connects the dots in data, bringing the power of the new analytics to existing business applications and unleash new intelligent applications. Data science can:

- Uncover totally unanticipated relationships and changes in markets or other patterns
- Help you change direction instantaneously
- Constantly adapt to changing data
- Handle streams of data—in fact, some embedded intelligent services make decisions and carry out those decisions automatically in microseconds

Data science enriches the value of data, going beyond what the data says to what it *means* for your organization—in other words, it turns raw data into intelligence that empowers everyone in your organization to discover new innovations, increase sales, and become more cost-efficient. Data science is not just about the algorithm, but about deriving value.

Modern Data Science Teams

At its core, data science rests on mathematics, computer science, and subject matter expertise. A strong statistical background has traditionally been assumed necessary to work in data science. However, data science goes far beyond statistics, and turns an expertise in statistics, data, and software development into a practical real-world discipline that solves a wide range of problems. Some of the additional skills that are part of a data science team include:

- Defining business needs and understanding what is of urgent interest to the business
- Determining what data is relevant to the organization, and balancing the value of the data against the cost and risk of collecting and storing it
- Data collection and cleansing: learning the tools to collect all kinds of data ranging from social media to sensors, and doing the necessary initial cleaning, such as removing errors and duplicates
- Exploring data to develop an understanding of the data and to discover patterns and identify anomalies in data
- Identifying the analytic techniques and models that will connect data sources to business needs
- Performing feature engineering to prepare the data for analysis including normalizing data, feature reduction, and feature generation
- Building, testing and validating data science models
- Creating powerful visualizations to support the data science model narrative and make the analysis easy to consume by end-users

- Using the data science model and visualization to build an intelligent application or embed the data science model into an existing application or device

Good statisticians are a hot commodity, and people can who do all the things just listed are even rarer. It is no surprise that an urgent shortage of data scientists plagues multiple industries in all countries. Given the complexity and technical sophistication of the requirements, we can't expect individuals to enter the field quickly enough to meet the growing need—which includes any company that doesn't want to fall behind and see its business taken by a more data-savvy competitor.

We must therefore form teams of people that embody all the necessary skills. A 2013 article in *CIO magazine* points out that asking a technologist to think like a businessman, or vice versa, is rarely successful, and that the two sides must know how to talk to each other. Modern data science team typically include:

- Business analysts: subject matter experts in the organization, good at manipulating spreadsheets and drawing conclusions; they're used to exploring data through visualizations
- Data scientists: good at statistics, math, computer science and machine learning, perhaps natural language text processing, geospatial analytics, deep learning, and various other techniques
- Developers: knowledgeable in computer science and software engineering, responsible for incorporating the data scientist's models into queries or programs that process the data and generate the final output as report, intelligent application, or intelligent service
- Data engineers: responsible for building and maintaining the data pipelines and storage mechanisms used for organizational data
- DevOps engineers: shepherd programs from test environments to production environments, ensuring that production runs finish successfully and meet requirements

These, painted with a broad brush, are the participants in an organization's data science team. We'll explore how to get them engaged in activities that uncover key information needed by your organization, and facilitate their working together.

The pace of business today demands responsive data science collaboration from empowered teams with a deep understanding of the business that can quickly deliver value. As with agile approaches, modern data science teams are being tasked to continuously deliver incremental business value. They know that they have to respond quickly to trends in the market, and they want tools that let them produce answers instantly. The new expectations match the experiences they're used to online, where they can find and order a meal, or record their activities and share videos with friends, within seconds. Increasingly, thanks to the incorporation of artificial intelligence into consumer apps, people also expect interfaces to adapt to their interests and show them what they want at the moment. With decreasing costs of computing, on-demand computation in the cloud, and new machine learning algorithms that take advantage of that computing power, data science can automate decisions that depend on complex factors, and present other decisions in a manner that is easier for people to visualize.

Data Science for All

Thanks to Big Data, data science has gone mainstream. Executives are spending billions of dollars collecting and storing data and they are demanding return on their investment. Simply getting the data faster is of limited value, so they are seeking to use the data to enrich their day-to-day operations and get better visibility into the future.

Data science is the path to monetizing the mounds of data now available. But old school tools are laden with with technical hurdles and huge costs that don't align well with the Big Data stack and aren't agile enough to keep up with the almost continuously evolving demands driven by changes in the Big Data stack and marketplace changes.

Enter Open Data Science. Open Data Science is a big tent that welcomes and connects many data science tools together into a coherent foundation that enables the modern data science team to solve today's most challenging problems. Open Data Science makes it easy for modern data science teams to use all data - big or small or anything in between. Open Data Science also maximizes the plethora of computing technologies available, including multi-core CPUs, GPUs, in-memory architectures, and clusters. Open Data Science takes advantage of a vast array of tried and true algorithms, plus the

latest and most innovative algorithms available. This is why Open Data Science is being used to propel science, business, and society forward.

Take for example, the recent discovery of gravitational waves by the Ligo project team, utilizing Python, NumPy, SciPy, matplotlib, and Jupyter Notebooks. And the Darpa Memex project that crawls the web to uncover and prosecute human trafficking rings using Anaconda, nutch, Bokeh, Tika and Elastic Search. Or the startup biotech firm Recursion Pharmaceuticals, which is on a mission to eradicate rare genetic diseases by discovering immune therapies from existing pharmaceutical shelved inventories that uses Anaconda and Bokeh.

What tools and practices enable open data science and the growing number of new opportunities to apply data science to real-world problems? Open source software, and a new emerging Open Data Science stack. In this section, we'll dig into each of these further.

Open Source Software and Benefits of Open Data Science

At the heart of Open Data Science lies open source software with huge and vibrant communities. Open source provides a foundation where new contributors can build upon the work of the pioneers who came before them. As an open source community matures and grows, its momentum increases since each new contributor can reuse the underlying work to build high level software that makes it easier for more people to use and contribute. When open source software is made available, the software is tested by far greater numbers in a wider range of situations - not just the typical cases the original designers envisioned but also the edge cases - which serves to quickly industrialize the software.

Open source is an ideal partner to the fast-paced technology shifts occurring today. The success of any open source project is based on market demand and adoption. Let's take a look at the reasons that open source has become the underpinning of this new work model:

- Availability. There are thousands of open source projects, offering any number of tools and approaches that can be used to solve problems. Each open source project initially is known to only a small community that shares a common interest. The projects quite naturally grow if they meet a market demand, or

simply wither away or linger on in obscurity if there isn't a strong market demand. But as adoption increases, the project matures and the software is used to solve more problems. This gives everyone access to the accumulated experience of thousands of data scientists, developers, and users.

- Robustness. Because every alpha and beta release goes out to hundreds or thousands of knowledgeable users who try them out on real-world data sets and applications, errors tend to get ironed out before the first official release. Even when the tools are out in the field, someone usually steps up to fix a reported bug quickly—and it the error is really holding up your own organization, your development team can fix it themselves. Open source software also guarantees continuity: you are not at the mercy of a vendor that may go out of business or discontinue support for a feature you depend on.
- Innovation. In the past, when a new algorithm was invented (typically by academics), they would present a paper about it at a conference. If the presentation was well received, they'd typically return the following year with an implementation of the algorithm and some findings to present to peers. By the time a vendor discovered the new algorithm or there was enough customer demand for it, three to five years had passed since the discovery of the algorithm. Contrast that to today. The best and brightest minds in universities invent a new algorithm or approach, often by collaborating, and immediately open source it and start to build a community that provides feedback and helps evolve the technology. The new algorithm finds its way into many more applications than initially intended and by doing so evolves faster. Users can choose from a plethora of algorithms and use them as-is or adjust them to the particular requirements of the current problem. This is exactly what you see unfolding in many open source communities: Python, R, Java, Hadoop, Scala, Julia and others. Because there are so many tools and they are so easy to exchange, new ideas fueled by the power of collective intelligence can be put into practice quickly. This experimentation and prototyping with near instantaneous feedback spreads ideas and encourages other contributors to deliver cutting edge innovations also.
- Transparency. In proprietary tools, algorithms are opaque and change requests are subject to the pace of the vendor. Open

source provides the information you need to determine whether algorithms are appropriate for your data and population.Thanks to decades of academic research, there is an abundance of open data science tools that disclose the algorithms and processing techniques to the public via open source, so that data scientists can ensure the technique is appropriate to solving the problem at hand. Additionally, data scientists can leverage open source algorithms and improve them to suit their problems and environments. This flexibility makes it easier and faster for the data science team to deliver higher value solutions. With open source, Data Scientists no longer have to blindly trust a black-box algorithm. They can read the code of the algorithms they are going to be executing in production, to make sure they are correctly implemented.

- Responsiveness to user needs. Open source software was usually developed to scratch someone's own itch (to use a metaphor popularized by Eric Raymond in his book *The Cathedral & the Bazaar*), and is extended over time by its users. Some proprietary vendors, certainly, stay very attuned to their customers and can produce new features at frequent intervals, but open source communities are uniquely endowed with the ability to shape software to meet all their requirements. At the same time, many of the new features announced with great fanfare by proprietary vendors are of little value to most customers and end up bloating the product. It's typical that the bulk of an organization's analytics uses only a dozen common features.
- Interoperability. Open source communities pay attention to each other's work and converge quickly on simple data formats, so tying tools from different projects together is easy. In contrast, proprietary vendors deliberately seek incompatibility with competing products (most readers will remember one vendor's promise to "embrace and extend" some years ago, although that vendor is now working very well with open source communities), and their own formats tend to become complex over time as they strive for backward compatibility. Open source communities are also very practical and create bridges to legacy technology that allow organizations to redeploy these systems into modern architectures and applications where necessary.
- Efficient investment. Open source projects do demand an initial investment of team time to evaluate the maturity of the software

and the community and to install software that is often packaged in a difficult manner. But over time, it is much more cost effective to run and maintain open source software than to pay the licensing fees for proprietary software. Open source software is therefore democratizing: it brings advanced software capabilities to residents of developing countries, students, and others who might not be able to afford the expensive, proprietary tools.

- Knowledgeable users. Many programmers and other technical teams learn popular open source tools in college because they can easily learn via the endless online resources and freely download the software. This means they come to their first jobs trained and can be productive with those tools the moment they first take their seats. It is harder to hire expertise in proprietary products from students straight out of college. Moreover, many open source users are adept at navigating the communities and dealing with internals of the code.

In short, modern data science teams have many reasons to turn to open source software. It makes it easy for them to choose the right tool for the right job, to switch tools and libraries when it is useful to do so, and to easily staff up their teams.

The Future of the Open Data Science Stack

Data is everywhere. There's more of it and it's messier. But it's just data. Everything around the data is changing - compute, hardware, software, analytics - while the structure and characteristics of the data itself are also changing.

For the last 30 years programmers have basically lived in a monoculture of both hardware and software. A single CPU family, made by Intel and running the x86 instruction set, has been coupled with a single succession of operating systems: DOS, then Windows and more recently, Linux. For the last 30 years, the design of software and business data systems started with this foundation. You placed your data in some kind of relational database, you hired a crew of software developers to write Java or .NET, and maybe a roomful of business analysts who used Excel.

But those software developers didn't generally have to think about potentially scaling their business applications to multiple operating system instances. They certainly almost never concerned themselves

with thinking about low-level hardware tradeoffs like network latency and cache coherence.

For business applications, almost no one was really tinkering with exotic options like distributed computing and, yes, cache coherence.

This siloed monoculture has been disrupted. At the most fundamental level, computer processors and memory—two tectonic plates under all the software we rely upon for business data processing and analytics - are being fractured, deconstructed, and revolutionized, and all kinds of new technologies are emerging.

The industry is really struggling with making computer chips run faster—or at least with keeping computer chips running at high speeds without completely melting down. So now, distributed and parallel computing are mainstream concepts that we have to get good at. This is much, much more complex than simply swapping out a CPU with one that's twice as fast.

NVIDIA's latest generation of GPUs delivers 5 TERAflops on a single chip. Depending on workload, that's roughly 100x faster than a vanilla Intel CPU. And people are buying them - Wall Street, all the major tech companies doing artificial intelligence and deep learning.

On the other end of the spectrum, Amazon and the other cloud vendors want us to stop thinking about individual computers, but rather move to a new paradigm where all computational resources are elastic and can be dynamically provisioned to suit the workload need. You want a thousand computers for the weekend? Click a button. This is a new way of thinking. Anyone who has had to deal with traditional IT departments can testify as to how long would it take them to get a new datacenter setup with 1,000 computers—about 25 racks of 42 1U servers. How many years? Now you can do it almost instantly. A "PC", a "server" - this has dissolved into a mere slider on a web page to indicate how many you want, and for how long.

While the cloud vendors are abstracting away the computer, a raft of technologies are emerging to abstract away the operating system. The technology space around containers, virtualization and orchestration is churning with activity right now, as people want to deconstruct and dissolve the concept of an "operating system" that is tied to a single computer. Instead, you orchestrate and manage an entire datacenter topology to suit your computational workload. So Windows? Linux? Who cares. It just needs an HTTP port.

And that's all just at the hardware and operating system level. If we go anywhere up the stack to applications, data storage, etc., we find similar major paradigm shifts. You're probably intimately familiar with the technology hype and adoption cycle around Hadoop. That same phenomenon is playing out in many other areas: IoT, business application architecture, you name it.

What may prove to be the largest disruption yet is about to hit next year: a new kind of storage/memory device called 3D Xpoint. A persistent class of storage like disk or SSD, it's 100x faster than SSDs and almost as fast as RAM. It's 10x more dense than RAM. So instead of a 1TB memory server, you'll have 10TB persistent memory.

To make this concrete: the new storage fundamentally changes how software is written, and even the purpose of an operating system has to be redefined. You never have to "quit" an application. All applications are running, all the time. There's no "Save" button because everything is always saved.

The rate of fundamental technology innovation - and not just churn - is accelerating. This will hit data systems first, because every aspect of how we ingest, store, manage, and compute business data will be disrupted.

This disruption will trigger the emergence of an entire new data science stack, one that eliminates components in the stack and blurs the lines of the old stack. Not only will the data science technology stack change, but costs will be driven down and the old-world proprietary vendors that didn't adapt to this new world order will finally tumble as well.

Open Data Science Applications

Open data science has brought the ingredients of data science - data, analytics, and computation - within everyone's reach. This is fueling a new generation of innovative intelligent applications that solve intractable problems and facilitate innovative discoveries. Here are a few Continuum Analytics case studies that utilize open data science.

Recursion Pharmaceuticals

This biotech startup found that the enormous size and complex interactions inherent in genomic material made it hard for biolo-

gists to find relationships that might predict disease or optimize treatment. Through a sophisticated combination of analytics and visualization, their data scientists produced "heat maps" that compared diseased samples to healthy genetic material, allowing differences to pop out. The biologists can not only identify disease markers more accurately and faster, but can also run intelligent simulations that apply up to thousands of potential drug remedies to diseased cells to identify treatments.

This greatly accelerated treatment discovery process, fueled by open data science, allows Recursion Pharmaceuticals to find treatments for rare genetic diseases —specifically, unanticipated uses for drugs already developed by their client pharmaceutical companies. The benefits to patients are incalculable, because treatments for rare diseases don't provide the revenue potential to justify costly drug development. Furthermore, small samples of patients mean that conventional randomized drug trials can't produce statistically significant results, and therefore means that drugs might otherwise not be approved for sale.

TaxBrain

The Open Source Policy Center was formed to "open source government" by creating transparency around the models used to formulate policies. Until now those models have been locked up in proprietary software. However, the Open Source Policy Center created an open source community seeded by academics and economists to create economic models using open data science tools. To make these models accessible to citizen data scientists and journalists, the Open Source Policy Center created a web interface, TaxBrain, that allows anyone to predict the economic impact of tax policy changes.

Having represented the tax code in a calculable form, this team can now ask questions: what will be the results of increasing or decreasing a rate? How about a deduction? By putting their work on the Web, the team allowed anyone with sufficient knowledge to ask such questions and get instant results. People concerned with taxes (and who is not?) can immediately show the effects of a change, instead of depending on the assurances of the Treasury Department or a handful of think tank experts. This is not only an open data science project, but an open data project (drawing from published laws) and an open source software project (the code was released on GitHub).

TaxBrain is a powerful departure from the typical data science project, where a team of data scientists create models that are surfaced to end users via reports. Instead, the underlying models in TaxBrain were created by subject matter experts who easily picked up Python and created powerful economic models that simulate the complexities of U.S. tax code to predict future outcomes based on citizen data scientists using an interactive visual interface to "what if" potential changes to policies.

Lawrence Berkeley National Laboratory/University of Hamburg

In academia, scientists often collaborate on their research as the physicists at and University of Hamburg have done. As with many scientists today, they are also data scientists that have to back up their research with data and reproducible results as their scientific research applies to medical research and treatment, national security, as well as computing.

Vying for time on one of the world's most advanced plasma accelerators is highly competitive. Your research has to be innovative and there needs to be evidence that your time on the accelerator will produce game-changing results that push the forefront of science. Particle physicists from Lawrence Berkeley National Laboratory (LBNL) and the University of Hamburg worked together to create a new algorithm and approach, using cylindrical geometry, that they embedded in a simulator to identify a set of the best experiments to run on the plasma accelerator. Even though the scientists are on separate continents, they were able to easily collaborate by using open data science tools, boosting their development productivity and allowing them to scale out complex simulations across a 128 GPU cluster that resulted in a 50% speedup in performance. This cutting-edge simulation optimized their time on the plasma accelerator, allowing them to zero in on the most innovative cutting edge research quickly.

As more businesses and researchers try to rapidly unlock the value of their data in modern architectures, open data science becomes essential to their strategy.

The Journey to Open Data Science

Organizations around the world, both small and large, are embarking on the journey to realize the benefits of Open Data Science. To succeed, they need to establish the right team, use the right technology to achieve their goals, and reduce migration risks. For most organizations, this journey removes barriers between departments as teams start to actively engage across departments and shift from incremental change to bold market moves.

Modern Data Science Team Organization and Roles

The concept of a *team* is being reimagined as organizations embark on the journey to Open Data Science. They are establishing centers of excellence, emerging technology groups, or innovation centers to bring together the right talent for their modern data science teams to jumpstart changes in their organization. There are typically two functions of these teams:

- Identify, evaluate, acquire, and absorb new technology - often open source - into their organization to empower the data science team and bridge the gap between traditional IT and the line of business
- Establish an agile or lean methodology that allows the data science team to quickly prototype new ideas and either discard them as failures or evolve them and move them into production

These teams are interdisciplinary and typically include both existing personnel and new talent already familiar with open source. Members coming from within the organization are also well equipped with subject matter expertise and the experience to inform the new world they are building. The combination of experiences and specialized knowledge allow the team to function as powerhouse data scientists. Roles often shift from focus on a single discipline to multidisciplinary knowledge. For example, a statistician may take on additional responsibilities as a programmer and become a data scientist. A database administrator may take on responsibilities for data prep - not just ETL but also analytic prep - and move into a data engineering role. A data scientist may take on responsibility for deploying data science on Hadoop and shift to a computational scientist. And so the team forms and evolves to serve the needs of the

organization as the journey to open data science progresses for the organization.

Although there is no standard data science team, at a minimum the modern team typically includes:

- Business analyst - A subject matter expert who typically uses spreadsheets and visualization tools to explore and analyze data to identify trends and anomalies. This role often identifies new opportunities for applying data science to long standing problems and challenges.
- Data scientist - An expert in statistics, mathematics, machine learning and perhaps other techniques such as deep learning, simulation, and optimization. This person is also familiar with modern compute infrastructures including Hadoop, multi-core CPUs, and GPUs. They typically know multiple data science languages, including Python and R, but could also include Scala, Lua and Julia. They're also familiar with plotting and visualization techniques and tools. They work with programming languages, integrated development environments, and online notebooks, a quintessential Open Data Science tool that facilitates the sharing of code, data, narratives and visualizations. These staff possess a strong working knowledge of the open source analytics ecosystem and familiarity with popular libraries and packages that are available for data science.
- Developer - A programmer who often creates the end application in which the data science work is encapsulated. They are typically familiar with multiple programming languages, including Python and Java, and use integrated development environments, notebooks and libraries - both analytic and visualization - to create the intelligent application.
- Data engineer - Typically has a strong database and/or ETL background and is very familiar with SQL. They move data from source systems and make it available to the data science team, often transforming the data so it is ready to be explored and analyzed.
- DevOps engineer - This role is responsible for taking proof of concept models and prototypes and readying them to be deployed into production environments. DevOps engineers typ-

ically know multiple programming languages, middleware, databases, Hadoop, encapsulation, and orchestration tools.

Given the pace of change in the data world, even the most well-qualified the data science team will need additional training to become or stay proficient in Open Data Science tools. Although instructor-led training is still the norm, there are also many online learning opportunities for Open Data Science, where the team can self-teach using the tools of Open Data Science itself. With Open Data Science, recruiting knowledgeable resources is much easier across disciplines — scientists, mathematicians, engineers, business, and more — as open source is the de facto standard in most universities worldwide. This results in a new generation of talent that can be brought onboard for data science projects.

Whether trained at university or on-the-job, the data science team needs the ability to integrate multiple tools into their workflow quickly and easily in order to be effective and highly productive. Most of the skills-ready university graduates are very familiar with collaborating with colleagues across geographies in their university experience. Many are also familiar with data science notebooks. This familiarity is critical because collaboration in data science is crucial to its success.

By forging a team that crosses organizational boundaries and skills, modern data science teams are empowered to create innovative, high-value data science applications that are helping enterprises make bold moves that are changing the world.

New Trends Critical to Success

The journey to Open Data Science is being forged with new practices that accelerate the time-to-value for organizations. In the past, much of the analysis has resulted in reports that delivered insights but required a human-in-the-loop to review and take action on the insights. Today organizations are looking to directly empower frontliners and even to embed intelligence into the devices and operational processes so that the action happens automatically and instantaneously rather than as an afterthought.

The key trends that support this transformation include the ability of the data science team to:

- Collaborate on data science projects

- Self-serve data science insights
- Deploy data science into production

Migration

A migration strategy to Open Data Science should align with the business objectives and risk tolerance of the organization. It is not necessary to commit to fully recoding old analytic methods as Open Data Science from the start. There is a range of strategies from completely risk averse (do nothing) to higher risk (recode), each with its own pros and cons.

A coexistence strategy is fairly risk-averse and allows the team to learn the new technology, typically on greenfield projects, while keeping legacy technology in place. This minimizes disruption while the Data Science team becomes familiar and comfortable with Open Data Science tools. Existing projects can then migrate to Open Data Science when limits are reached with the proprietary technology. The proprietary technologies are then phased out over time.

A migration strategy is slightly riskier and moves existing solutions into Open Data Science by reproducing the solution as-is with any and all limitations. This is often accomplished by outsourcing the migration to a knowledgeable third party who is proficient in the proprietary technology as well as Open Data Science. A migration strategy can take place over time by targeting low-risk projects with limited scope, until all existing Data Science code has been migrated to Open Data Science. Migration strategies can also migrate all the legacy code via a "big bang" cutover. The Data Science solutions are improved to remove the legacy limitations over time, usually using a continuous integration continuous delivery (CICD) methodology.

A recoding strategy is higher-risk and takes advantage of the entire modern analytics stack to reduce cost, streamline code efficiency, decrease maintenance, and create higher-impact business value more often, through faster performance or from adding new data to drive better results and value. The objective of recoding is to remove limitations and constraints of legacy code by taking full advantage of Open Data Science on modern compute infrastructure. With this strategy, a full risk assessment is often completed to determine the prioritization of projects for recoding. The full risk assessment

includes estimates for cost reduction and improved results to determine the risk.

The introduction of Big Data projects has become an ideal scenario for many companies using a coexistence strategy; they leave legacy environments as-is and use Open Data Science on Hadoop for their Big Data projects.

Conclusion

Data science has moved from the back office out into the limelight of boardrooms, operations, and the frontline. It is now a concern for the whole organization. And you will find, if you provide opportunities for your team to work together on data science, many if not most will welcome the chance. Your team wants to be useful to the organization; they want their unique insights to be exploited. Organizations that do so will become knowledge leaders and thrive in an economy that shows little mercy toward back-sliders. Industries that cannot exploit data science to make their processes more intelligent will end up commoditized, and if their products prove inferior to some newly developed competitor, will disappear into the mists of history.

Open data science combines the rich array of open source tools with modern computing architectures, hardware advances, new sources of data, and collaborative processes to transform organizations. We suggest you find a bold new project to empower your modern data science team. Armed with the right support, your team will find opportunities to reduce costs, open new revenue channels, delight your customers and change the world with innovative discoveries.

About the Authors

Michele Chambers is an entrepreneurial executive with over 25 years of industry experience. Prior to Continuum Analytics, Michele held executive leadership roles at database and analytic companies, Netezza, IBM, Revolution Analytics, MemSQL and RapidMiner. In her career, Michele has been responsible for strategy, sales, marketing, product management, channels and business development. Michele is a regular speaker at analytic conferences including Gartner and Strata and has books published by Wiley and Pearson FT Press on big data and modern analytics.

Christine Doig is a senior data scientist at Continuum Analytics, where she has worked, among other projects, on MEMEX, a DARPA-funded project helping stop human trafficking through Open Data Science. She has 5+ years of experience in analytics, operations research and machine learning in a variety of industries, including energy, manufacturing and banking. Christine loves empowering people through open source technologies. Prior to Continuum Analytics, she held technical positions at Procter and Gamble and Bluecap Management Consulting. Christine is a regular speaker and trainer at open source conferences such as PyCon, EuroPython, SciPy and PyData, among others.

Christine holds a M.S. in Industrial Engineering from the Polytechnic University of Catalonia in Barcelona.