O'REILLY®

Data Driven

Creating a Data Culture

DJ Patil & Hilary Mason

SAN JOSE

LONDON

NEW YORK

SINGAPORE

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

Data Driven

Creating a Data Culture

DJ Patil and Hilary Mason

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

Data Driven
by DJ Patil and Hilary Mason

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: Timothy McGovern
Copyeditor: Rachel Monaghan
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

January 2015: First Edition

Revision History for the First Edition
2015-01-05: First Release

978-1-491-92119-7

[LSI]

Table of Contents

Data Driven: Creating a Data Culture

The data movement is in full swing. There are conferences (Strata +Hadoop World), bestselling books (*Big Data*, *The Signal and the Noise*, *Lean Analytics*), business articles ("Data Scientist: The Sexiest Job of the 21st Century"), and training courses (An Introduction to Machine Learning with Web Data, the Insight Data Science Fellows Program) on the value of data and how to be a data scientist. Unfortunately, there is little that discusses how companies that successfully use data actually do that work. Using data effectively is not just about which database you use or how many data scientists you have on staff, but rather it's a complex interplay between the data you have, where it is stored and how people work with it, and what problems are considered worth solving.

While most people focus on the technology, the best organizations recognize that people are at the center of this complexity. In any organization, the answers to questions such as who controls the data, who they report to, and how they choose what to work on are always more important than whether to use a database like PostgreSQL or Amazon Redshift or HDFS.

We want to see more organizations succeed with data. We believe data will change the way that businesses interact with the world, and we want more people to have access. To succeed with data, businesses must develop a data culture.

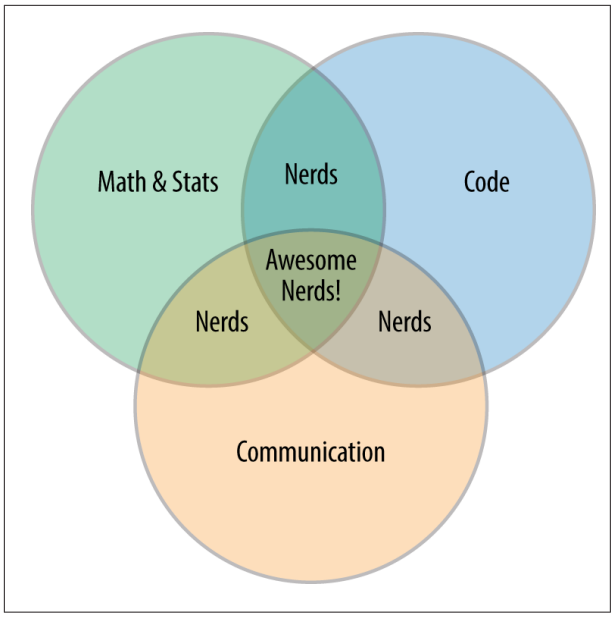
What Is a Data Scientist?

Culture starts with the people in your organization, and their roles and responsibilities. And central to a data culture is the role of the data scientist. The title *data scientist* has skyrocketed in popularity over the past five years. Demand has been driven by the impact on an organization of using data effectively. There are chief data scientists now in startups, in large companies, in nonprofits, and in government. So what exactly *is* a data scientist?

A data scientist doesn't do anything fundamentally new. We've long had statisticians, analysts, and programmers. What's new is the way data scientists combine several different skills in a single profession. The first of these skills is mathematics, primarily statistics and linear algebra. Most scientific graduate programs provide sufficient mathematical background for a data scientist.

Second, data scientists need computing skills, including programming and infrastructure design. A data scientist who lacks the tools to get data from a database into an analysis package and back out again will become a second-class citizen in the technical organization.

Finally, a data scientist must be able to communicate. Data scientists are valued for their ability to create narratives around their work. They don't live in an abstract, mathematical world; they understand how to integrate the results into a larger story, and recognize that if their results don't lead to action, those results are meaningless.

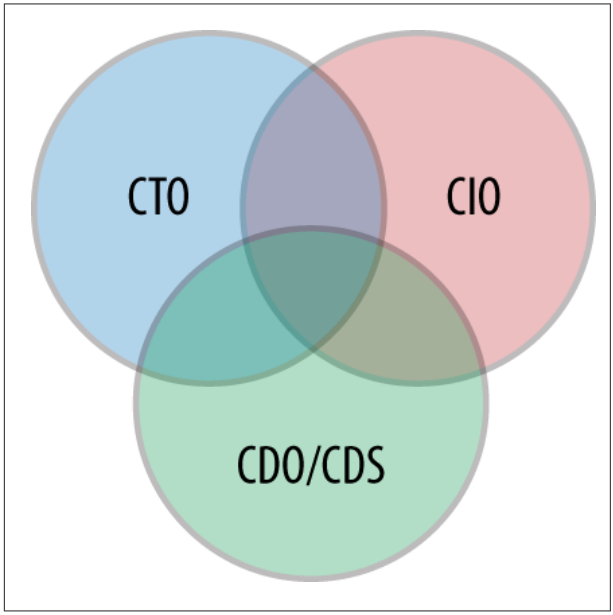


In addition to these skills, a data scientist must be able to ask the right questions. That ability is harder to evaluate than any specific skill, but it's essential. Asking the right questions involves domain knowledge and expertise, coupled with a keen ability to see the problem, see the available data, and match up the two. It also requires empathy, a concept that is neglected in most technical education programs.

The old *Star Trek* shows provide a great analogy for the role of the data scientist. Captain Kirk is the CEO. Inevitably, there is a crisis and the first person Kirk turns to is Spock, who is essentially his chief data officer. Spock's first words are always "curious" and "fascinating"—he's always adding new data. Spock not only has the data, but more importantly, he uses it to understand the situation and its context. The combination of data and context allows him to use his domain expertise to recommend solutions. This combination gives the crew a unique competitive advantage.

Does your organization have its version of a Spock in the boardroom? Or in another executive meeting? If the data scientists are isolated in a group that has no real contact with the decision makers, your organization's leadership will suffer from a lack of context and expertise. Major corporations and governments have realized that

they need a Spock on the bridge, and have created roles such as the chief data scientist (CDS) and chief data officer (CDO) to ensure that their leadership teams have data expertise. Examples include Walmart, the New York Stock Exchange, the cities of Los Angeles and New York, and even the US Department of Commerce and National Institutes of Health.



Why have a CDO/CDS if the organization already has a chief technology officer (CTO) or a chief information officer (CIO)? First, it is important to establish the chief data officer as a distinct role; that's much more important than who should report to whom. Second, all of these roles are rapidly evolving. Third, while these roles overlap, the primary measures of success for the CTO, CIO, and CDS/CDO are different. The CIO has a rapidly increasing set of IT responsibilities, from negotiating the "bring your own device" movement to supporting new cloud technologies. Similarly, the CTO is tasked with an increasing number of infrastructure-related technical responsibilities. The CDS/CDO is responsible for ensuring that the organization is data driven.

What Is a Data-Driven Organization?

The most well-known data-driven organizations are consumer Internet companies: Google, Amazon, Facebook, and LinkedIn. However, being data driven isn't limited to the Internet. Walmart has pioneered the use of data since the 1970s. It was one of the first organizations to build large data warehouses to manage inventory across its business. This enabled it to become the first company to have more than $1 billion in sales during its first 17 years. And the innovation didn't stop there. In the 1980s, Walmart realized that the quality of its data was insufficient, so to acquire better data it became the first company to use barcode scanners at the cash registers. The company wanted to know what products were selling and how the placement of those products in the store impacted sales. It also needed to understand seasonal trends and how regional differences impacted its customers. As the number of stores and the volume of goods increased, the complexity of its inventory management increased. Thanks to its historical data, combined with a fast predictive model, the company was able to manage its growth curve. To further decrease the time for its data to turn into a decision, it became the first large company to invest in RFID technologies. More recently it's put efforts behind cutting-edge data processing technologies like Hadoop and Cassandra.

FedEx and UPS are well known for using data to compete. UPS's data led to the realization that, if its drivers took only right turns (limiting left turns), it would see a large improvement in fuel savings and safety, while reducing wasted time. The results were surprising: UPS shaved an astonishing 20.4 million miles off routes in a single year.

Similarly, General Electric uses data to create improve the efficiency of its airline engines. Currently there are approximately 20,000 airplanes operating with 43,000 GE engines. Over the next 15 years, 30,000 more engines are expected to be in use. A 1% improvement in efficiency would result in $30 billion in savings over the next 15 years. Part of its effort to attack these problems has been the new GEnx engine. Each engine weighs 13,740 pounds, has 4,000 parts with 18 fan blades spinning at 1,242 ft/sec, and has a discharge temperature of 1,325°F. But one of the most radical departures from traditional engines is the amount of data that is recorded in real time. According to GE, a typical flight will generate a terabyte of data.

This data is used by the pilots to make better decisions about efficiencies, and by the airlines to find optimal flight paths as well as to anticipate potential issues and conduct preventative maintenance.

What about these data-driven organizations enables them to use data to gain a competitive advantage? In *Building Data Science Teams*, we said that a data-driven organization

> acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on and develop new products, and navigate the competitive landscape..

Let's break down the statement a little. The first steps in working with data are *acquiring* and *processing*. But it's not obvious what it takes to do these regularly. The best data-driven organizations focus relentlessly on keeping their data clean. The data must be organized, well documented, consistently formatted, and error free. Cleaning the data is often the most taxing part of data science, and is frequently 80% of the work. Setting up the process to clean data at scale adds further complexity. Successful organizations invest heavily in tooling, processes, and regular audits. They have developed a culture that understands the importance of data quality; otherwise, as the adage goes, *garbage in, garbage out*.

A surprising number of organizations invest heavily in processing the data, with the hopes that people will simply start creating value from it. This "if we build it, they will come" attitude rarely works. The result is large operational and capital expenditures to create a vault of data that rarely gets used. The best organizations put their data to use. They use the data to understand their customers and the nuances of their business. They develop experiments that allow them to test hypotheses that improve their organization and processes. And they use the data to build new products. The next section explains how they do it.

Democratizing Data

The democratization of data is one of the most powerful ideas to come out of data science. Everyone in an organization should have access to as much data as legally possible.

While broad access to data has become more common in the sciences (for example, it is possible to access raw data from the National Weather Service or the National Institutes for Health), Facebook was one of the first companies to give its employees access to data at

scale. Early on, Facebook realized that giving everyone access to data was a good thing. Employees didn't have to put in a request, wait for prioritization, and receive data that might be out of date. This idea was radical because the prevailing belief was that employees wouldn't know how to access the data, incorrect data would be used to make poor business decisions, and technical costs would become prohibitive. While there were certainly challenges, Facebook found that the benefits far outweighed the costs; it became a more agile company that could develop new products and respond to market changes quickly. Access to data became a critical part of Facebook's success, and remains something it invests in aggressively.

All of the major web companies soon followed suit. Being able to access data through SQL became a mandatory skill for those in business functions at organizations like Google and LinkedIn. And the wave hasn't stopped with consumer Internet companies. Nonprofits are seeing real benefits from encouraging access to their data—so much so that many are opening their data to the public. They have realized that experts outside of the organization can make important discoveries that might have been otherwise missed. For example, the World Bank now makes its data open so that groups of volunteers can come together to clean and interpret it. It's gotten so much value that it's gone one step further and has a special site dedicated to public data.

Governments have also begun to recognize the value of democratizing access to data, at both the local and national level. The UK government has been a leader in open data efforts, and the US government created the Open Government Initiative to take advantage of this movement. As the public and the government began to see the value of making the data more open, governments began to catalog their data, provide training on how to use the data, and publish data in ways that are compatible with modern technologies. In New York City, access to data led to new *Moneyball*-like approaches that were more efficient, including finding "a five-fold return on the time of building inspectors looking for illegal apartments" and "an increase in the rate of detection for dangerous buildings that are highly likely to result in firefighter injury or death." International governments have also followed suit to capitalize on the benefits of opening their data.

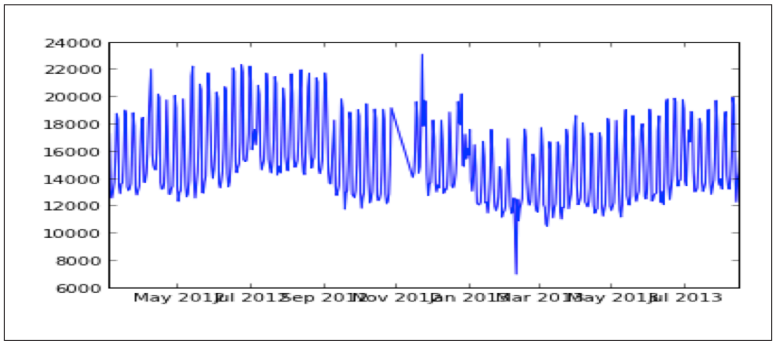
One challenge of democratization is helping people find the right data sets and ensuring that the data is clean. As we've said many

times, 80% of a data scientist's work is preparing the data, and users without a background in data analysis won't be prepared to do the cleanup themselves. To help employees make the best use of data, a new role has emerged: the data steward. The steward's mandate is to ensure consistency and quality of the data by investing in tooling and processes that make the cost of working with data scale logarithmically while the data itself scales exponentially.

What Does a Data-Driven Organization Do Well?

There's almost nothing more exciting than getting access to a new data set and imagining what it might tell you about the world! Data scientists may have a methodical and precise process for approaching a new data set, but while they are clearly looking for specific things in the data, they are also developing an intuition about the reliability of the data set and how it can be used.

For example, one of New York's public data sets includes the number of people who cross the city's bridges each day. Let's take just the data for the Verrazano-Narrows Bridge. You might imagine that this would produce a very predictable pattern. People commute during the week, and perhaps don't on the weekends. And, in fact, we see exactly that for the first few months of 2012. We can ask a few straightforward questions. What's the average number of commuters per day? How many people commuted on the least busy day? On the most busy day? But then something strange happens. There's a bunch of missing data. What's going on?



A bit of digging around those dates will show you that there's no conspiracy here: that data represents Hurricane Sandy, when the bridges and tunnels were deliberately closed. It also explains the spike that happens when the bridges reopened. You also see traffic drop sharply for the blizzard of February 2013. The data set is as simple as they come—it's just one integer per day—and yet there's a fascinating story hiding here.

When data scientists initially dive into a data set, they are not just assembling basic statistics, they're also developing an intuition for any flaws in the data and any unexpected things the data might be able to explain. It's not a matter of checking statistics off a list, but rather of building a mental model of what data says about the world.

The process is similar, though on a larger scale, for organizations with a data culture. One of the most important distinctions between organizations that are data driven and those that are not is how they approach hypothesis formulation and problem solving. Data-driven organizations all follow some variant of the scientific method, which we call the data scientific method:

1. Start with data.
2. Develop intuitions about the data and the questions it can answer.
3. Formulate your question.
4. Leverage your current data to better understand if it is the right question to ask. If not, iterate until you have a testable hypothesis.
5. Create a framework where you can run tests/experiments.
6. Analyze the results to draw insights about the question.

In 2009, Twitter was faced with a challenge. There was tremendous excitement about the service, but people were just not using it regularly: three out of four people would stop using it within two months. To solve the engagement problem, Twitter started by asking questions and looking at its current data. It found a number of surprising results. First, users who had used the service at least seven times in their first month were over 90% likely to return in subsequent months. For many organizations, identifying this magic number would be more than sufficient. But Twitter continued to study the data, and was well rewarded. Among users with high retention, it found that once a user followed 30 or more people, that person was almost certainly going to become a long-term user. The company continued to dig, and found that the nature of the people followed was also essential. Two-thirds of the people who new users followed were purely for content, but one-third had to follow the new users back.

Armed with these facts, the Twitter team was able to discover a solution that was counter to the conventional thinking about onboarding new users. Sites like Facebook and LinkedIn presented new users with an "address book importer" that would crawl the user's email addresses. Next, a new user would see a page filled with suggestions for "people you may know." Any other pages that might add "friction" to the user's experience would cause a significant (20%+) number of users to abandon the onboarding.

The analysis showed that Twitter needed to (a) teach new users what a tweet was, (b) suggest accounts that had high-quality content seg-

mented by categories (e.g., NFL, NBA, news sites), and then (c) suggest other users who were highly likely to follow someone once they knew that person was on Twitter. Implementing these ideas adds friction to the onboarding process by teaching users about the tweet; it also puts people a user is likely to interact with last. However, the result wasn't a decrease in new users, but instead a 30% increase in people completing the experience and a 20% increase in long-term engagement!

In hindsight, the process and results almost look magical. They're far from that; they represent dedicated adherence to the data scientific method. It took roughly 2.5 years to arrive at and test these results—and the process is nowhere near complete. Regular tests are ongoing to further improve what happens when new users arrive. The data scientific method never stops.

Twitter isn't the only place that employs the data scientific method. Google is famous for testing hundreds of experiments a day to improve its search functionality. LinkedIn and Facebook are constantly conducting experiments to learn how to improve the experience of new users. Netflix is well known for testing and adjusting its entire experience to reduce the probability that users will cancel their subscriptions. It is using its data to make very costly investments into the types of shows that need to be created to keep its users engaged. The Obama campaign, in its record-breaking fundraising effort, did over 500 A/B tests over 20 months, resulting in increasing donation conversions by 49% and signup conversions by a whopping 161%.

Managing Research

Once you have a sense of the problems that you'd like to tackle, you need to develop a robust process for managing research. Without a process, it's easy to spend too much time on unimportant problems, and without a research-specific process, it's easy to get drawn into the engineering world where research is not a priority.

Here's a set of questions that can be asked about every data science problem. They provide a loose framework for managing a robust portfolio of research efforts with both short- and long-term rewards. For each research problem, we ask:

What is the question we're asking?

It's important to state the question in language that everyone in the company can understand. This is harder than you think it will be! Most companies have teams with diverse backgrounds, so it's important to articulate clearly the question that you are addressing so that everyone in the organization can imagine why it might be relevant and useful.

How do we know when we've won?

Once you have defined the question, you need to define the metrics by which you will evaluate your answer. In many cases, these are quantitative metrics (e.g., cross-validation), but in some cases the metric may be qualitative, or even a "looks good to me." Everyone on the team needs to be clear about what a success will look like.

Assuming we solve this problem perfectly, what will we build first?

This question is designed to assess the solution's potential to impact your business. What capabilities will you have that you don't have now? Is this an important problem to solve right now?

While you should always have a "first thing" in mind, we recommend coming up with further questions that you'll be able to investigate once you have answered the first one. That way, you can manage both short- and long-term value.

If everyone in the world uses this, what is the impact?

What's the maximum potential impact of this work? If it's not inspiring, is it worth pursuing at all? It's vital to make sure that data science resources are invested in projects that will have a significant impact on the business. There is no greater insult than "You've created an elegant solution to an irrelevant problem."

What's the most evil thing that can be done with this?

This question is a bit different! Don't ask it if you work with people who enjoy doing evil. Instead, save this one for groups that are so lawful and good that they limit their thinking. By asking the team to imagine what their impact could be *if you abandon all constraints*, you allow for a conversation that will help you identify opportunities that you would otherwise miss, and refine good ideas into great ones. We don't want to build

"evil" products, but subversive thinking is a good way to get outside the proverbial box.

One of the challenges with data is the power that it can unleash for both good and bad, and data scientists may find themselves making decisions that have ethical consequences. It is essential to recognize that just because you can, doesn't mean you should. It's important to get outside input. When uncertain, we turn to well-regarded experts on privacy and legal matters (e.g., the Electronic Frontier Foundation).

Designing the Organization

In the last few years, a lot of attention has been focused on the celebrity data scientist. But data science isn't about celebrities; it's a team sport. While a single person who has access to data and knows how to use it can have a huge impact, relying on a single celebrity isn't scalable. A culture that is dependent on one individual is fragile and won't be sustainable. It's more important to think about the composition of the team and how it should be organized.

Should the data team be centralized or decentralized? Should it be part of Engineering, a product group, or Finance, or should it be a separate organization? These are all important questions, but don't focus on them at first. Instead, focus on whether you have the key ingredients that will allow the team to be effective. Here are some of the questions you should ask:

- What are the short-term and long-term goals for data?
- Who are the supporters and who are the opponents?
- Where are conflicts likely to arise?
- What systems are needed to make the data scientists successful?
- What are the costs and time horizons required to implement those systems?

Ask these questions constantly. As the data culture emerges and gains sophistication, periodic restructuring will be necessary.

LinkedIn's early data efforts were split between data scientists supporting the CFO (dashboards and basic reporting) and data scientists building products. When I (DJ) joined LinkedIn we had a debate about how we should structure the team. The conventional options were to build a team under the CEO, under the CTO, or in Engineering. We tried something different. We put the data team in

the Product organization. First, we wanted this team to be able to drive and implement change while ensuring ownership and accountability. Second, we had a phenomenal Engineering team, and we realized that we could bring the Product and Engineering teams into better alignment through common DNA.

Over time, we realized that this model couldn't support the speed at which we were growing (from 200 to 2,000 employees in under four years), so we decentralized the team. The unfortunate consequence was that data scientists would end up isolated, supporting a specific group. We tried many solutions, but eventually decided that a decentralized organization worked only when there were at least three data scientists supporting a given area.

All of these organizational changes took place as we implemented new technologies, built out our data warehouse, and grew our data operations. We constantly needed to rethink and reevaluate our organizational structure to provide the best career growth and impact. However, we always had one central tenet in mind: to grow a massive company, every part of the organization must be data driven. This means that the data would be fully democratized, and everyone would be sufficiently data proficient. Naturally, we would still need those with a specific skill set, but data would become an intrinsic skill and asset for every team.

Process

While most organizations focus on corporate structure, they give less attention to the processes and technology needed to build a data-driven culture. The next three sections outline some of the most essential processes and ways to evaluate technologies.

Figure 1-1. DILBERT © 2007 Scott Adams. Used By permission of UNIVERSAL UCLICK. All rights reserved.

Daily dashboard

Data-driven organizations look at their data every morning. Starting every day with a review of the data isn't just a priority, it's a habitual practice. The simplest way to review the data is by looking at dashboards that describe key metrics. These dashboards might be implemented by a spreadsheet that is emailed, or by a business intelligence application accessed through the Web.

There are two classic complaints around dashboards. The first is that they don't contain enough data; the second is that there is too much data. How do you find the right balance? Here are some hints.

Data vomit
: Don't fall for the urge to add "just one more thing" to the dashboard. Adding more information at greater density creates data vomit. Data vomit is bad and leads to frustration. The data becomes intimidating and, as a result, is just ignored.

Time dependency
: Put data on your dashboard only if you know what you will do if something changes. For example, if there is a significant change on an hourly dashboard, does someone's pager go off? Does the appropriate team know what to investigate? Similarly, display the data in a form that allows action to be taken. If the dashboard contains a pie graph, will you be able to tell if there is a change? Instead of a single dashboard, create different dashboards that reflect different time scales. For example, some dashboards might be on an hourly scale, while others could be on a quarterly scale. These simple measures prevent your dashboard from turning into data vomit.

Value
: Manage your dashboards instead of letting them manage you. Review them and ask whether they are still giving you value. If not, change them. It's surprising how often people consider their dashboard "fixed" and unchangeable. It's quite the opposite: the dashboard is a living entity that allows you to manage your organization. As the organization's data sophistication increases, it's likely that old ways of measuring the system will become too simplistic. Hence, those older measures should be replaced with newer ones.

Visual

Make your data look nice. It's surprising how ugly most dashboards are. The font is too small or in a typeface that isn't clearly readable. If there is a line graph, it looks like it came from the 1980s. Sometimes dashboards are put in 3D, or have a color palette with no real meaning. Turn the data you regularly look at into something that you'd *like* to look at.

Fatigue

Finally, watch out for "alert/alarm fatigue." We like to create alerts when something changes. But if there are too many alarms, you create alarm fatigue: the team becomes desensitized to the alerts, because they're occurring so often and they're frequently meaningless. Review alarms and ask what actions are taken once the alarm is activated. Similarly, review false positives and false negatives to see if the alerting system can be improved. Don't be afraid to remove an alert or an alarm if it isn't serving a purpose. Some well-regarded teams carry pagers, so they can share the pain of unnecessary alarms. There is nothing like successive wake-up calls at 3 a.m. to motivate change.

As a general rule of thumb, we like to ask four questions whenever data is displayed (in a dashboard, a presentation, or in a product):

- What do you want users to take away? In other words, what information do you want the user to walk away with? Is it that things are good or things are bad?
- What action should you take? When presenting a result, ask what you want your audience to do. For example, if there is a problem with sales, and the recipient is the CEO, you might want the CEO to call the head of sales right away. You may not be able to convey this in the dashboard, but you certainly can discuss it in your data meetings.
- How do you want the viewer to feel? Most effective organizations embrace putting emotion and narrative around the data. If the goal is to make people feel excited, use green. If the feeling is neutral, use black or blue. If you want to express concern or urgency, use yellow or red. Great data teams spend time and energy ensuring the narrative provides adequate context, is compelling, and is intellectually honest.

- Finally, is the data display adding value regularly? If not, don't be afraid to "prune" it. Removing items that are no longer valuable keeps the dashboard effective and actionable.

Metrics meetings

One of the biggest challenges an organization faces isn't creating the dashboard, it's getting people to spend time studying it. Many teams go to great lengths to create a dashboard, only to learn how rarely people use it. We've seen many attempts to force people to look at the dashboards, including automated delivery to mobile devices, human-crafted email summaries, and print copies of the dashboards left on the chairs of executives. Most of these techniques don't work. When we've used tracking codes to measure the open rates of these emails and see who has viewed the dashboards, the numbers were atrocious.

The model that I've found the best is inspired by Sustained Silent Reading, or SSR (a popular school reading program in the United States). Instead of assuming that people looked at the data on their own, we spent the first part of the meeting looking at the data as a group. During this time, people could ask questions that would help them understand the data. Everyone would then write down notes, circle interesting results, or otherwise annotate the findings. At the end of the reading period, time was dedicated to a discussion of that data.

I've seen dramatic results from this method. During the first few meetings, the conversation is focused on basic questions, but those are quickly replaced by deeper questions. The team begins to develop a common language for talking about the data. Even the sophistication of the data presented begins to change.

This process prevents data from being used as a weapon to push an agenda. Rather than jumping straight into decision making, we start with a *conversation about the data*. At the end of the conversation, we can then ask if we have enough information to make a decision. If the answer is yes, we can move forward. If it's no, we can ask what it will take to make an informed decision. Finally, we ask if there is any reason to think that we should go against the data.

As a result, everyone becomes smarter. A discussion makes everyone more informed about the data and its different interpretations. It also limits mistakes. This kind of forum provides a safe environ-

ment for basic questions that might otherwise seem dumb (such as "what do the labels on the axis mean?"). Simple questions often expose flaws in the way data was collected or counted. It's better to find the flaws before a decision has been made.

Second, the conversation disarms the data as a political weapon. All too often, Team 1 shows data defending its argument, and Team 2 shows similar but conflicting data for its argument. Who is right? Focusing the conversation on the data rather than the decision makes it possible to talk openly about how data was collected, counted, and presented. Both teams might be right, but their data may be addressing different issues. In this way, assessment through discussion wins, not the best graph.

One word of caution: don't follow the data blindly. Being data driven doesn't mean ignoring your gut instinct. This is what we call "letting the data drive you off a cliff." Do a web search for "GPS" and "cliff" and you'll find that a surprising number of people actually crash their car when the GPS is giving them bad information. Think about that for a second. The windshield is huge relative to the screen size of the GPS. As a result, the data coming through to the driver is massive relative to the information that is output by the voice or the screen of the GPS. By likewise hyper-optimizing to a specific set of metrics, you too can drive (your business) off a cliff. Sometimes it is necessary to ignore the data. Imagine a company that is trying to determine which market to enter next. There is stronger user adoption in Market A, but in Market B there is a new competitor. Let's suppose all the data says that the company should go after Market A. It still might be better to go after Market B. Why? The data can't capture everything. And sometimes you have to trust your gut.

How can you prevent these kinds of catastrophic failures? First, regularly ask "are we driving off a cliff?" By doing so, you create a culture that challenges the status quo. When a person uses that phrase, it signals that it's safe to challenge the data. Everyone can step back and take into account the broader landscape.

Standup and domain-specific review meetings

We've discussed a number of meetings that are needed as part of a data-driven culture. But we all know that endless, dull meetings kill creativity and independent thought. How do we get beyond that? I've found that it works to borrow processes and structures from companies that implement agile methods.

The first of these is the standup meeting. These are short meetings (often defined by the time that a person is willing to stand) that are used to make sure everyone on the team is up-to-date on issues. Questions or issues become action items that are addressed outside of the standup meeting. At the next standup meeting, the action items are reviewed to see if they have been resolved, and if not, to determine when resolution is expected. If you're literally standing up, standup meetings should be no longer than 30 minutes. They're a great way to start the day and enhance communication.

It's a misconception that standup meetings and other ideas from the agile movement work only for high-tech Silicon Valley firms. Standup meetings are used to monitor situations in the US Department of Defense. The National Weather Service has a daily meeting where the weather forecasters gather (both in person and virtually) to raise any issues. The key to success in all of these forums is to be ruthlessly efficient during the meeting and to make sure issues raised (action items) are acted upon.

It's also important for the data team to hold product review, design review, architecture review, and code review meetings. All of these meetings are forums where domain-specific expertise can provide constructive criticism, governance, and help. The key to making these meetings work is to make sure participants feel safe to talk about their work. During these meetings, definitions of metrics, methodologies, and results should be presented before being deployed to the broader organization.

Tools, Tool Decisions, and Democratizing Data Access

Data scientists are always asked about their tools: What tools do you use? How do I become an expert user of a particular tool? What's the fastest way to learn Hadoop?

The secret of great data science is that the tools are almost irrelevant. An expert practitioner can do better work in the Bash shell environment than a nonexpert can do in R. Learning how to understand the problem, formulate an experiment and ask good questions, collect

or retrieve data, calculate statistics or implement an algorithm, and verify the accuracy of the result is much more important than mastering the tools.

However, there are a few attributes of tools that both are timeless and enable stronger teamwork:

- The best tools are **powerful**. They aren't visual dashboards that offer a limited set of options, but Turing-complete programming languages. Powerful tools allow for unconventional and powerful analysis techniques.
- The best tools are **easy to use** and learn. While tools should be powerful, it should be easy to understand how to apply them. With programming languages, you want to see tutorials, books, and strong communities.
- The best tools **support teamwork**. These tools should make it easier to collaborate on analysis and to make data science work reproducible.
- The best tools are beloved by the **community**. A tool that's popular in a technical community will have many more resources supporting it than a proprietary one. It'll be easier to find people who already use the tools your team uses, and it'll give your employees the ability to demonstrate your company's expertise by participating in that community.

We've stressed the democratization of data access. Democratization doesn't come without costs, and may require rethinking your organization's data practices and tools. Not that long ago, it was common for organizations with a lot of data to build a "data warehouse." To create a data warehouse, you would build a very robust database that assumed the data types wouldn't change very much over time. Access to the warehouse was restricted to those who were "sanctioned"; all others were required to go to them. This organizational structure created a data bottleneck. The tools in the data warehouse might have been powerful, but they weren't easy to use, and certainly didn't support teamwork or a larger data community. If you needed data, you had to go through the data bureaucracy, which meant that you'd get your data a day (or days) later; if you wanted data that didn't fit into the warehouse's predefined schema, you might be out of luck. Rather than teaching people to fish, data bureaucrats opted to create dashboards that had limited functionality outside of the key questions they were designed to answer.

Requiring users to go through a data bureaucracy to get access to data is sure to halt democratization. Don't force the users who need data to go through channels; train them to get it themselves. One challenge of this approach is the ability to support a large number of users. Most data solutions are evaluated on speed—but when you're supporting large numbers of users, raw speed may not be relevant. Almost anything will be faster than submitting a request through data warehouse staff. We made this mistake early in the process of building LinkedIn's data solutions, and learned that the following criteria are more relevant than pure speed:

- How well will the solution scale with the number of concurrent users?
- How does it scale with the volume of data?
- How does the price change as the number of users or the volume of data grows?
- Does the system fail gracefully when something goes wrong?
- What happens when there is a catastrophic failure?

Price is important, but the real driver is the ability to support the broadest possible set of users.

Consider what Facebook found when it first allowed everyone to access the data. Simply providing access to data didn't help people make better decisions. People couldn't find the data they needed, and there remained a huge gap in technical proficiency. Meanwhile its data was growing at an unprecedented rate. To scale more efficiently, Facebook invested aggressively in tooling. One of its first investments was a new way to store, access, and interact with the data. And with that the company realized that it would need a new language that would be easier for a broader set of employees to use.

While Hadoop, the underlying technology, showed great promise, the primary language supported, Pig, was not friendly for analysts, product managers, or anyone accustomed to languages like SQL. Facebook decided to invest in Hive, a SQL-like language that would be more optimal for Hadoop, and a unique tooling layer called HiPal that would be the primary GUI for Hive. HiPal allowed any user to see what others in the company were accessing. This unique form of transparency allowed a new user to get up to speed quickly by studying what other people on the team were requesting and then building on it.

As powerful as HiPal was, it still was insufficient at helping users who were unfamiliar with coding or languages. As a result, the Facebook data team started a series of training classes. These classes allowed the team to educate staff about HiPal and simultaneously share best practices. Combined with training, HiPal jumpstarted the data capabilities of Facebook's teams and fostered a strong sense of data culture. It lowered the cost of data access and created the expectation that you needed data to support your business decisions. It was a major foundation for Facebook's growth strategy and international expansion.

Creating Culture Change

We want organizations to succeed with data. But succeeding with data isn't just a matter of putting some Hadoop in your machine room, or hiring some physicists with crazy math skills. Succeeding with data requires real cultural change. It requires learning how to have a discussion about the data—how to hear what the data might be saying rather than just enlisting it as a weapon in company politics. It requires spreading data through your organization, not just by adding a few data scientists (though they are critical to the process), but by enabling everyone in the organization to access the data and see what they can learn.

As Peter Drucker stated in *Management Challenges for the 21st Century*, "Everybody has accepted by now that change is unavoidable. But that still implies that change is like death and taxes—it should be postponed as long as possible and no change would be vastly preferable. But in a period of upheaval, such as the one we are living in, change is the norm." Good data science gives organizations the tools to anticipate and stay on the leading edge of change. Building a data culture isn't easy; it requires persistence and patience. You're more likely to succeed if you start with small projects and build on their success than if you create a grandiose scheme. But however you approach it, building a data culture is the key to success in the 21st century.