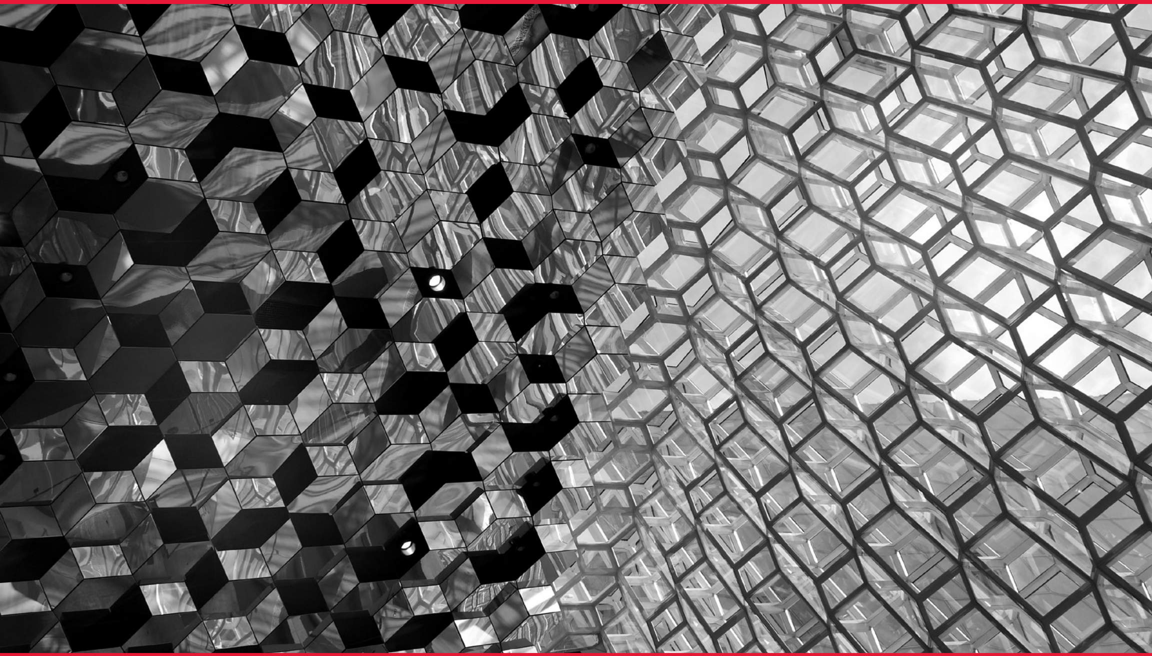
O'REILLY®

Ten Signs of Data Science Maturity



Peter Guerra
& Kirk Borne

9 781491 952528

Ten Signs of Data Science Maturity

Peter Guerra and Kirk Borne

Beijing • Boston • Farnham • Sebastopol • Tokyo

Ten Signs of Data Science Maturity
by Peter Guerra and Kirk Borne

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
Production Editor: Melanie Yarbrough
Copyeditor: Melanie Yarbrough
Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

February 2016: First Edition

Revision History for the First Edition
2016-03-07: First Release

Cover photo: *Olafur Eliasson's glass front* by tristanf.

978-1-491-95252-8

[LSI]

Table of Contents

O'REILLY®

Ten Signs of Data Science Maturity

Peter Guerra
& Kirk Borne

O'REILLY®

9 781491 952528

Ten Signs of a Mature Data Science Capability

> If you want to build a ship,
> don't drum up people to collect wood,
> and don't assign them tasks and work,
> but rather teach them to long for the endless
> immensity of the sea.
>
> —Antoine de Saint-Exupéry

Over the years in working with US government, commercial, and international organizations, we have had the privilege of helping our clients design and build a data science capability to support and drive their missions. These missions have included improving health, defending the nation, improving energy distribution, serving citizens and veterans better, improving pharmaceutical discovery, and more.

Often, our engagements have turned into exercises in transforming how the organization operates—"building a capability" means building a culture to support and make the most of data science. In many cases, this culture change has delivered significant insights into big challenges the world faces—poverty, disease outbreaks, ocean health, and so forth. We have encountered a wide variety of successful organizational structures, skill levels, technologies, and algorithmic patterns.

Based on those experiences, we share here our perspective on how to assess whether the data science capability that you are developing within your own organization is achieving maturity. In no particular order, here are our top ten characteristics of a mature data science capability.

A mature data science organization…

1. …democratizes all data and data access.

Let's make one thing clear from the start: Silos suck! Most organizations early on in the data-science learning curve spend most of their time assembling data and not analyzing it. Mature data science organizations realize that in order to be successful they must enable their members to access and use all available data—not some of the data, not a subset, not a sample, but *all* data. A lawyer wouldn't go to court with only some of the evidence to support their case—they would go with all appropriate evidence. Similarly, mature data science organizations use all of their data to understand their business domain, needs, and performance. Successful organizations take the time to understand all the data they collect, to understand its uses and content, and to allow easy access.

Some recent articles have suggested that big data and data science are mutually exclusive: Focusing on increasing data-gathering ("big data") comes at the expense of quality analysis ("data science"). We disagree. They are mutually conducive to discovery, data-driven decision-making, and big return on analytics innovation. Big data isn't about the volume of data nearly as much as it is about "all data" —stitching diverse data sources together in new and interesting ways that facilitate data science exploration and exploitation of all data sources for powerful predictive and prescriptive analysis. You can't have mature data science without democratizing access to all data. That means standardizing metadata, access protocols, and discovery mechanisms. You aren't mature until you have done that for all data.

Here is where cultural incentives are so important. We've seen too many organizations that still use data as power levers: we hear that we can't get data because a single person is the data steward and access has to be controlled. Governance is essential, but it can't be a pretext for one person or group maintaining power by controlling access to data. Let go, and let data discovery and innovation begin!

2. …uses Agile for everything and leverages DataOps (i.e., DevOps for Data Product Development).

Some traditional organizations are stuck in older ways of managing processes and development. If your IT and development departments are asking for requirements and expect to deliver a year or more out, then you may be experiencing this. These organizations are resistant to change—consequently, requests for new tools and methods go before review boards and endless architecture/design committees to justify the expenditure. Often, a large effort will be funded simply to study whether the proposed solution will work. Other times, a committee will decide which analytic problems are the most pressing. Paralysis of analysis must be broken in order to achieve data science maturity and success. Bureaucracy doesn't work well in science, and it doesn't work in data science either. Science celebrates exploratory, agile, fast-fail experimental design (see "7. …celebrates a fast-fail collaborative culture." on page 8).

Just as Agile development has championed user stories and short iterations over long drawn-out requirements and delayed delivery, Agile data science requires both close collaboration within the business and the freedom to experiment. Agile is not a software development methodology, it is a mindset. It permeates all levels of the mature organization. When was the last time your CEO or senior manager held a retrospective or SCRUM meeting? Understanding how to promote a flexible culture, organization, and technology that work together can be challenging, but immensely rewarding because of the collaboration and creativity it cultivates.

An agile DevOps methodology for data product development is critical—we call this DataOps. DataOps works on the same principles as DevOps: tight collaboration between product developers and the operational end users; clear and concise requirements gathering and analysis rounds; shorter iteration cycles on product releases (including successes and fast-fail opportunities); faster time to market; better definition of your MVP (Minimum Viable Product) for quick wins with lower product failure rates; and generally creating a dynamic, engaging team atmosphere across the organization. In addition to these general Agile characteristics, DataOps accelerates current data analytics capabilities, naturally exploits new fast data architectures (such as schema-on-read data lakes), and enables previously impossible analytics. With a sharpened focus on each MVP and the corresponding SCRUM sprints, DataOps minimizes team downtime from both lengthy review cycles and the costs of cognitive switching between different projects.

Mature data science capability reaches its full potential in an agile DataOps environment.

3. …leverages the crowd and works collaboratively with businesses (i.e., data champions, hackathons, etc.).

Data science groups that live in a bubble are missing out on the best community out there. Activities that promote data science for social good, including open or internal competitions (like Kaggle), are a great way to sharpen skills, learn new ones, or just generally collaborate with other parts of the business.

In addition, mature data science teams don't try to go at it alone, but instead work collaboratively with the rest of the organization. One successful tactic is sponsoring internal data science competitions, which are great for team building and integration. The mature data science organization has a collaborative culture in which the data science team works side by side with the business to solve critical problems using data.

Another approach is internal crowdsourcing (within your organization)—this is particularly strong for surfacing the best questions for data scientists to tackle. The mature data science capability crowdsources internally several different tasks in the data science process lifecycle, including data selection; data cleaning; data preparation and transformations; ensemble model generation; model evaluation; and hypothesis refinement (see "4. …follows rigorous scientific methodology (i.e., measured, experimental, disciplined, iterative, refining hypotheses as needed)." on page 5). Since data cleaning and preparation can easily consume 50–80% of a project's entire effort, you can accrue significant project time savings and risk reduction by parallelizing (through crowdsourcing) those cleaning and preparation efforts, especially by crowdsourcing to those parts of the organization that are most familiar with particular data products and databases.

Also, algorithms don't solve all problems. It is still incredibly difficult for an algorithm to understand all possible contexts of an outcome and pick the right one. Humans must be in the loop still, and a deep understanding of the context of the challenge is essential to solid interpretation of data and creating accurate models.

4. …follows rigorous scientific methodology (i.e., measured, experimental, disciplined, iterative, refining hypotheses as needed).

Exploratory and undisciplined are not compatible. Data science must be disciplined. That does not mean constrained, unimaginative, or bureaucratic. Some organizations hire a few data scientists and sit them in cubes and expect instant results. In other cases, the data scientists work within the IT organization that is focused on operations, not discovery and innovation.

Mature data science capability is built on the foundation of the scientific method. First, make observations (i.e., collect data on the objects, events, and processes that affect your business)—collect data in order to understand your business by embedding measurement systems or processes (or people) at appropriate places in your business workflow. Think of interesting questions to explore, and then formulate testable hypotheses with your business partners. Once you have a good set of questions and hypotheses, then test them—analyze data, develop a data science model, or design a new algorithm to validate each hypothesis, or else refine the hypothesis and iterate. This methodology will ensure that value is created when formal scientific rigor is applied. That's an undeniable sign of mature data science capability.

A key part of the scientific process is knowing the limits of your sample. Looking for and testing for selection bias is key. Similarly, it is important to understand that "big data" does not spell the end to incomplete samples (unfair sampling) or sample variance (natural diversity).

5. …attracts and retains diverse participants, and grants them freedom to explore.

The key word is diverse. What fun is a bunch of math nerds? (Three statisticians go out hunting together. After a while they spot a solitary rabbit. The first statistician takes aim and overshoots the rabbit by one meter. The second aims and undershoots it by one meter.

The third shouts out "We got it!") Some organizations are looking for data scientists who are great coders, who also understand and apply complex applied mathematics, who know a lot about the specific business domain, and who can communicate with all stakeholders. One or two such people may exist—we call them purple unicorns. Mature organizations recognize that data science is a team sport, with each member contributing valuable unique skills and points of view.

Among those skills and competencies are these: Advanced Database/Data Management & Data Structures; Smart Metadata for Indexing, Search, & Retrieval; Data Mining (Machine Learning) and Analytics (KDD = Knowledge Discovery from Data); Statistics and Statistical Programming; Data & Information Visualization; Network Analysis and Graph Mining (everything is a graph!); Semantics (Natural Language Processing, Ontologies); Data-intensive Computing (e.g., Hadoop, Spark, Cloud, etc.); Modeling & Simulation (computational data science); and Domain-Specific Data Analysis Tools.

But don't think that every person must have at least one of those technical skills at the outset—some of the best data science organizations grow those skillsets from within, by identifying the core aptitudes among their current staff that lead to data science success (even within nontechnology trained staff). Those core aptitudes include the 10 C's: curiosity (inquisitive), creativity (innovative), communicative, collaborative, courageous problem-solver, commitment to life-long learning, consultative (can-do, will-do attitude), cool under pressure (persistence, resilience, adaptability, and ambiguity tolerance), computational, and critical thinker (objective analyzer).

Diverse perspectives are beneficial on multiple fronts. They make the questions more interesting, but more importantly they make the answers even more interesting, useful, and informative. Answers are given greater context that can yield greater impact. Mature data science capability understands that you need more than just math or computer science folks on projects. The mature organization integrates business experts, SMEs, "data storytellers", and creative "data artists" seamlessly, and then grants them the freedom to explore and exploit the full power of their data assets. The output from such diverse teams will be richer than that from any purple unicorn. And remember, it is better to have both a horse and a narwhal than a unicorn!

6. …relentlessly asks the right questions, and constantly searches for the next one.

The fundamental building block of a successful and mature data science capability is the ability to ask the right types of questions of the data. This is rooted in the understanding of how the business runs or how any business challenge manifests itself. The best data science team covers all the aptitude requirements mentioned earlier (see "5. …attracts and retains diverse participants, and grants them freedom to explore." on page 5): curious, creative, communicative, collaborative, courageous problem solvers, life-long learner, doer, and resilient.

Mature data science capability is exemplified in the relentless pursuit of new questions to ask (even questions that could never be answered before) and in asking questions of the questions! Data science maturity frees the organization to ask the hard questions across the entirety of the business, is disciplined in how it asks those questions, and is not afraid of getting the "wrong answer."

In this instance, data science capability maturity tracks analytics maturity in the following sense. Advanced analytics is often described as the new stages of analytics that go beyond traditional business intelligence, which covers Descriptive Analytics (hindsight) and Diagnostic Analytics (oversight). The current view of advanced analytics includes these new stages: Predictive Analytics (foresight) and Prescriptive Analytics (insight—understanding your business sufficiently to know which decisions, actions, or interventions will lead to the best, optimal outcome). The next emerging stage of analytics maturity is Cognitive Analytics ("the right sight")—knowing the right question to ask of your data (at the right time, in the right context, for the right use case). This "cognitive" ability to come up with not just the right answers but with the right questions (especially questions that were never asked or considered before) is the highest level of both analytics maturity and data science capability maturity. As the adage says: "The only bad question is the one that you don't ask."

7. …celebrates a fast-fail collaborative culture.

Culture is a hard thing to define, but if you look at what a team celebrates, that is a good indicator. Some organizations are afraid to fail, or have a culture where that is frowned upon. They are more focused on strategy than culture. But many business experts remind us that "culture eats strategy for breakfast (or lunch)." Therefore, start working on your data science culture sooner than on your data science strategy. Admitting mistakes is one thing, but purposefully exploring the unknown with your data is not a mistake. Test your organization's maturity by asking yourself: when my hypothesis fails, then what happens? The fast-fail mindset understands and appreciates the proper meaning of this adage: "Good judgment comes from experience. And experience comes from bad judgment."

True data science (based on rigorous scientific methodology; see "4. …follows rigorous scientific methodology (i.e., measured, experimental, disciplined, iterative, refining hypotheses as needed)." on page 5) explores the limits of what can be learned quickly by iterating on multiple hypotheses with agility. This may require that you invite your business unit partners to explore with you—that's DataOps (see "2. …uses Agile for everything and leverages DataOps (i.e., DevOps for Data Product Development)." on page 3). Having the data and tools to allow you to do this is directly related to its success and maturity (see "1. …democratizes all data and data access." on page 2). Mature data science capability allows for an iterative fast-fail culture on your path to achieving the most rewarding discoveries, making the best evidence-based decisions, and delivering the most innovative choices for your organization.

The optics around a project failing is often difficult to overcome. It is hard to justify spending limited resources only to find out that the hypothesis was wrong—the value from knowing what not to do is often lost or not celebrated within the culture. A mature data science capability is familiar with traditional A/B testing—designing experiments to test and evaluate alternative hypotheses, one of which may include some sort of intervention or tuning (the treatment sample) and the other is the null hypothesis (applied to the control, untreated sample). Typically, one of those experiments will fail, and one of them will not. That's the whole point of A/B testing. If an organization cannot accept failure, then they are not doing mature data science.

One could argue that fast-fail has an analytical foundation in machine learning algorithms. Specifically, in many classification algorithms, the goal is to define as accurately as possible the boundary (however complex) that separates different classes of objects. That boundary might be linear (e.g., if your team scores more points than my team, then you win), or it might be skew (e.g., if your total score on two exams A + B is greater than 140 out of 200, then you pass the course), or it might be complex (e.g., the hyperplane separating two classes in a Support Vector Machine algorithm when you are working with complex data that has high dimensionality).

In order to circumscribe the boundaries between complex classification rules (e.g., business decisions, product choices, or class labels), the problem space can be represented as a mapping exercise in which the boundaries of the different regions are accurately defined. Determining the location along every "inch" of the border requires detailed, comprehensive probes and surveys. For example, if you are testing the hypothesis that your customers will buy your product on Black Friday only if you offer a deep discount, then you need to try multiple discounts (10%, 20%, 30%, 40%, or maybe even 0%) to see where the boundary really is. Your profit margin depends critically on identifying the boundary where your ROI is optimized, and that means finding points on both sides of the boundary (failure and success conditions) until the points along the decision boundary are finally triangulated. Fast-fail is essential in such situations—time and resource investments are being wasted otherwise.

8. …shows insights through illustrations and tells stories.

Most organizations have some form of reporting. This is often focused on producing a monthly or weekly retrospective in which a line graph, bar, or pie chart illustrates what has happened in the previous reporting period. This is a clear indication that the organization's capability is not asking the higher-order questions beyond "What happened, and when?" It is stuck in the world of descriptive analytics. It is missing out on the emerging benefits of predictive and prescriptive analytics. The mature data science organization will therefore ask: "Why did that happen, what will happen next, and what can we do to achieve a better outcome?" And the organization can mature further by asking: "What questions should I be posing to my data?" (See "6. …relentlessly asks the right questions, and constantly searches for the next one." on page 7).

When insights are generated to answer the "what if" questions ("What could happen" or "What are all the possible outcomes if we…?"), those answers can't be relegated to a line graph or a bar chart to illustrate the impact of the findings. Infographics and beautiful unique illustrations do more justice to your hard work, and are critical to having the greatest impact. Mature data science capability is focused on the harder questions and then communicates (and illustrates) in new and creative ways the answers, story, and insights that the data are revealing.

Hence, the mature data science team includes one or more people with the skills of a data artist and a data storyteller. Stories and visualizations are where we make connections between facts. They enable the listener to understand better the context (What?), the why (So what?), and "what will work" in the future (Now what?).

9. …builds proof of value, not proof of concepts.

Many organizations start down the path on which delivering a proof of concept is considered successful data science. They want to validate a particular tool that a vendor told them will fix their challenges, so they set up a Hadoop environment (or something similar), pump data into it, ask a question, and see if the system delivers the "right answer." Success! Right?

Wrong!

Mature data science capability means being methodical in how you think about your pilots. What is it that you really want your pilot to prove—a concept or real business value? Proof of value changes the value proposition of the work. Data platforms are hard enough to architect in the right way for your unique needs. So, focus more on value (answering new questions, opening new markets, deriving new insights), and not so much on answering the question to which you already know the answer. Therefore, focus on proving to the organization that the data science capabilities that you are building are on a journey that will consistently prove value (e.g., 10× in many of our experiences) and that will solve the organization's greatest "unknown unknowns."

What Is Different Now?[1]

The tangible benefits of data products include:

Opportunity Costs
: Because data science is an emerging field, opportunity costs arise when a competitor implements and generates value from data before you. Failure to learn and account for changing customer demands will inevitably drive customers away from your current offerings. When competitors are able to successfully leverage data science to gain insights, they can drive differentiated customer value propositions and lead their industries as a result.

Enhanced Processes
: As a result of the increasingly interconnected world, huge amounts of data are being generated and stored every instant. Data science can be used to transform data into insights that help improve existing processes. Operating costs can be driven down dramatically by effectively incorporating the complex interrelationships in data like never before. This results in better quality assurance, higher product yield, and more effective operations.

Build with value in mind, much as Agile forces you to do (See "2. … uses Agile for everything and leverages DataOps (i.e., DevOps for Data Product Development)." on page 3). The DataOps culture celebrates success with the MVP (Minimum Viable Product)—the product that delivers value (not proof of concept) as quickly as possible, thereby enabling the team to move on to the next success.

10. …personifies data science as a way of doing things, not a thing to do.

Data science is not just a buzzword, or a relabeling of a data analyst or business intelligence function. It is not a way to produce a better monthly report ("TPS report cover sheet, please"). It is certainly not something that someone does once and then moves on.

1 © 2015 Booz Allen Hamilton Field Guide to Data Science, page 28.

Often we find organizations that look at data science as another lever within the larger set of gears that are working together to drive an institution. The power of data science within an organization is not in being one cog, no matter how well-connected to the rest of the machine, but by being the gear shaft that turns all the other gears. It is the engine that drives all other functions in an organization. When businesses look at data science to understand their world and use that to determine the best course of action, success invariably follows.

Data science is a fundamental shift in how organizations think and operate. It is using data at the core of all functions in new and interesting ways that make the organization more innovative. The evidence of mature data science capability is an organization that believes and lives this statement: "Now is the time to begin thinking of data science as a profession not a job, as a corporate culture not a corporate agenda, as a strategy not a stratagem, as a core competency not a course, and as a way of doing things not a thing to do."

Finally, we offer some guideposts for organizations that may need some assistance in identifying indicators of their current state of maturity plus recommendations for moving forward toward greater data science maturity.

THE WILLING ORGANIZATION	THE HESITANT ADOPTER	THE DATA DISTRESSED ORGANIZATION
Characteristics	**Characteristics**	**Characteristics**
+ Beginning of the journey + Believes in the power of analytics, but overwhelmed with how to get started	+ Beginning/middle of the journey + Analytics visionaries want to invest more heavily in analytics, but lack organizational support	+ Beginning/middle of the journey + Organization wants to develop increasingly sophisticated analytics, but is stymied by an inability to get the underlying data in order
What To Do	**What To Do**	**What To Do**
Chart a Clear Path: Explore data to help inform the organization's vision for analytics, and to chart a step-by-step path to achieve that vision	**Use Prototypes to Overcome Doubt:** Think big but start small, using prototypes to prove the value of analytics to end users and help overcome doubt	**Bring the Stakeholders Together:** Collaborate and build relationships between IT and analysts to develop data that can be used by both

Figure 1-1. © 2015 Booz Allen Hamilton Tips for Building Data Science Capability Handbook

These tenets that we have outlined are key to ensuring a data science capability is successful within your organization. We believe strongly that tearing down data and organizational silos is key to transforming business and governments into agile, data-driven organizations that can only improve decision making and foster innovation. This is the way forward!

About the Authors

Peter Guerra is Chief Data Scientist and Vice President leading Booz Allen Hamilton's Data Science commercial team. He has 15 years of experience in creating big data and data science solutions for government and commercial clients. He was responsible for the architecture and implementation of one of the world's largest Hadoop clusters for the federal government. He has consulted with Fortune 500 companies and federal government organizations throughout his career. Recently, he has focused on data governance and security of large data systems, working on a book for O'Reilly titled *Data Security for Modern Enterprises*. He is a frequent speaker at large events, including Blackhat, Hadoop Summit, Strata+Hadoop World, Infosec World, Evanta CDO Council, and more. He holds an MBA from Loyola University, a B.A. degree in English and B.S. degree in Computer and Information Science from University of Maryland. Contact him on Twitter at @petrguerra.

Dr. Kirk Borne is the Principal Data Scientist at Booz Allen Hamilton (since 2015). He supports the Strategic Innovation Group in the area of NextGen Analytics and Data Science. He previously spent 12 years as Professor at George Mason University in the graduate (Ph.D.) Computational Science and Informatics program and undergraduate (B.S.) Computational Data Sciences program. Before that, he worked 18 years on various NASA contracts—as research scientist, as a manager on a large science data system contract, and as the Hubble Telescope Data Archive Project Scientist. His PhD is in Astronomy from Caltech. He has applied his expertise in science and large data systems as a consultant and advisor to numerous agencies and firms, focusing on the use of data for discovery, decision support, and innovation across many different domains and industries. He is also a blogger (rocketdatascience.org) and actively promotes data literacy for everyone by disseminating information related to data science and analytics on social media, where he has been named consistently since 2013 among the top worldwide influencers in big data and data science. Follow him on Twitter at @KirkDBorne.