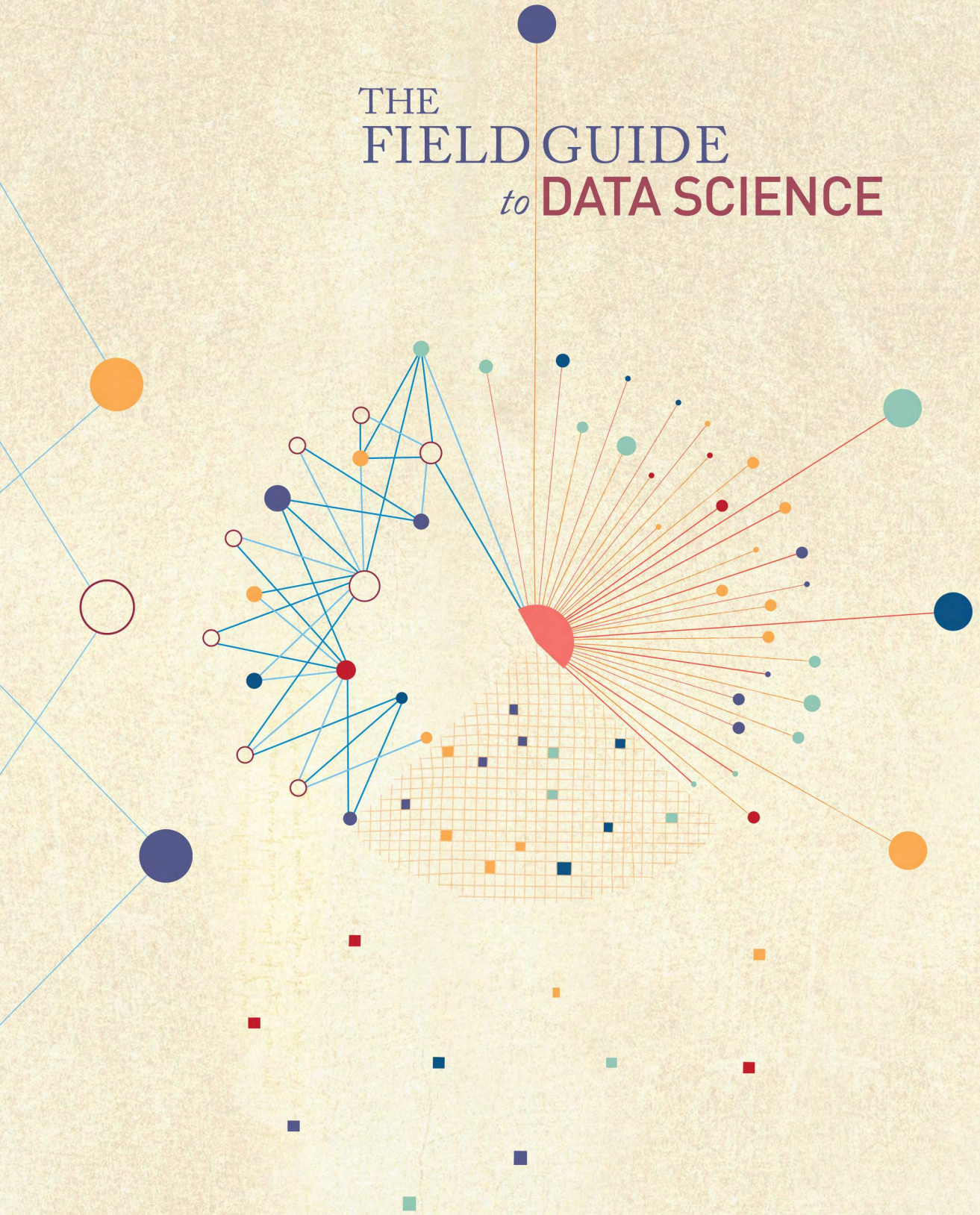



THE  
FIELD GUIDE  
*to* DATA SCIENCE



Booz | Allen | Hamilton

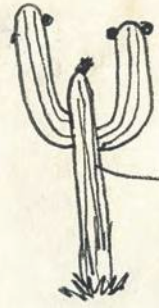
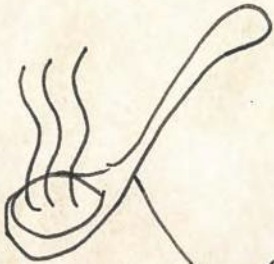
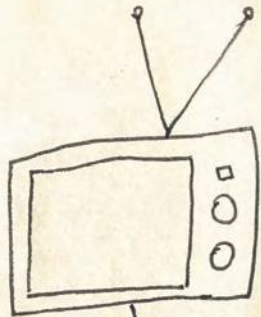


THE  
FIELD GUIDE  
*to* DATA SCIENCE

SECOND  
EDITION

Booz | Allen | Hamilton

© COPYRIGHT 2015 BOOZ ALLEN HAMILTON INC. ALL RIGHTS RESERVED.



# »» FOREWORD

*Data Science touches every aspect of our lives on a daily basis. When we visit the doctor, drive our cars, get on an airplane, or shop for services, Data Science is changing the way we interact with and explore our world.*

**Our world is now measured, mapped, and recorded in digital bits.** Entire lives, from birth to death, are now catalogued in the digital realm. These data, originating from such diverse sources as connected vehicles, underwater microscopic cameras, and photos we post to social media, have propelled us into the greatest age of discovery humanity has ever known. It is through Data Science that we are unlocking the secrets hidden within these data. We are making discoveries that will forever change how we live and interact with the world around us.

The impact of these changes is having a profound effect on humanity. We have propelled ourselves into this age of discovery through our incremental technological improvements. Data Science has become the catalyzing force behind our next evolutionary leap. Our own evolution is now inextricably linked to that of computers. The way we live our lives and the skills that are important to our very existence are directly dependent upon the functions Data Science can achieve on our behalf.

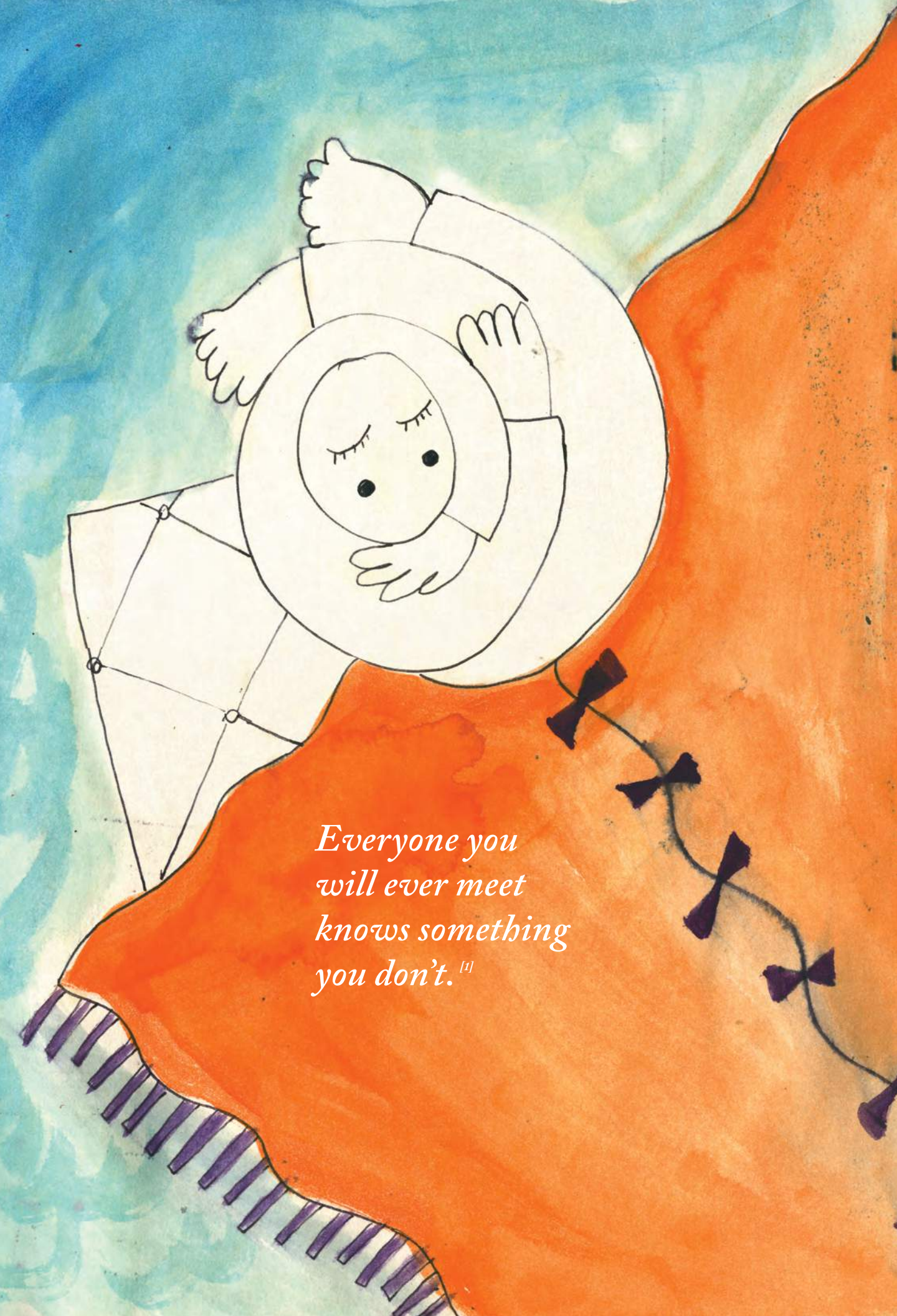
As we move into this new future, it is clearer than ever, that businesses must adjust to these changes or risk being left behind. From influencing retail markets, to setting public health and safety policies, or to addressing social unrest, organizations of all types are generating value through Data Science. Data is our new currency and Data Science is the mechanism by which we tap into it.

Data Science is an auspicious and profound way of applying our curiosity and technical tradecraft to solve humanity's toughest challenges. The growing power, importance, and responsibility of applying Data Science methodologies to these challenges is unimaginable. Our own biases and assumptions can have profound outcomes on business, national security, and our daily lives. A new class of practitioners and leaders are needed to navigate this new future. Data Scientists are our guides on this journey as they are creating radical new ways of thinking about data and the world around us.

---

We want to share our passion for *Data Science* and start a conversation with you. This is a journey worth taking.

---



*Everyone you  
will ever meet  
knows something  
you don't.<sup>[1]</sup>*

# »» THE STORY of THE FIELD GUIDE

Several years ago we created *The Field Guide to Data Science* because we wanted to help organizations of all types and sizes. There were countless industry and academic publications describing *what* Data Science is and *why* we should care, but very little information was available to explain how to make use of data as a resource. We find that situation to be just as true today as we did two years ago, when we created the first edition of the field guide.

At Booz Allen Hamilton, we built an industry-leading team of Data Scientists. Over the course of hundreds of analytic challenges for countless clients, we've unraveled the DNA of Data Science. Many people have put forth their thoughts on single aspects of Data Science. We believe we can offer a broad perspective on the conceptual models, tradecraft, processes and culture of Data Science – the *what*, the *why*, the *who* and the *how*. Companies with strong Data Science teams often focus on a single class of problems – graph algorithms for social network analysis, and recommender models for online shopping are two notable examples. Booz Allen is different. In our role as consultants, we support a diverse set of government and commercial clients across a variety of domains. This allows us to uniquely understand the DNA of Data Science.

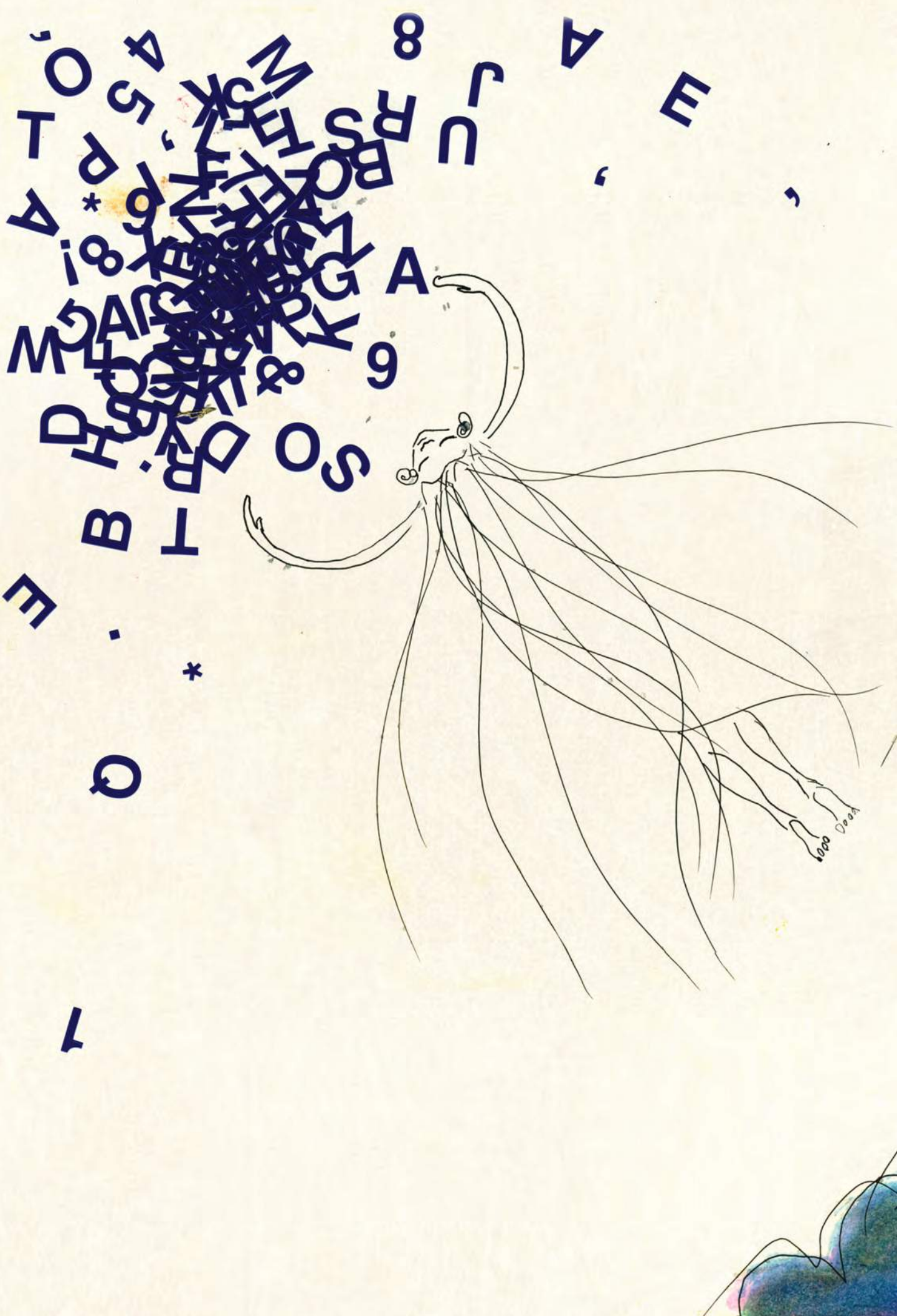
Our goal in creating *The Field Guide to Data Science* was to capture what we have learned and to share it broadly. The field of Data Science has continued to advance since we first released the field guide. As a result, we decided to release this second edition, incorporating a few new and important concepts. We also added technical depth and richness that we believe practitioners will find useful.

We want this effort to continue driving forward the science and art of Data Science.

---

This field guide came from the passion our team feels for its work. It is not a textbook nor is it a superficial treatment. Senior leaders will walk away with a deeper understanding of the concepts at the heart of Data Science. Practitioners will add to their toolbox. We hope everyone will enjoy the journey.

---



# »» WE ARE ALL AUTHORS of THIS STORY

**We recognize that Data Science is a team sport.** *The Field Guide to Data Science* provides Booz Allen Hamilton's perspective on the complex and sometimes mysterious field of Data Science. We cannot capture all that is Data Science. Nor can we keep up - the pace at which this field progresses outdates work as fast as it is produced. As a result, we opened this field guide to the world as a living document to bend and grow with technology, expertise, and evolving techniques.

Thank you to all the people that have emailed us your ideas as well as the 100+ people who have watched, starred, or forked our GitHub repository. We truly value the input of the community, as we work together to advance the science and art of Data Science. This is why we have included authors from outside Booz Allen Hamilton on this second edition of *The Field Guide to Data Science*.

If you find the guide to be useful, neat, or even lacking, then we encourage you to add your expertise, including:

- › Case studies from which you have learned
- › Citations from journal articles or papers that inspire you
- › Algorithms and techniques that you love
- › Your thoughts and comments on other people's additions

---

Email us your ideas and perspectives at [data\\_science@bah.com](mailto:data_science@bah.com) or submit them via a pull request on the [GitHub repository](#).

Join our conversation and take the journey with us. Tell us and the world what you know. Become an author of this story.

---





# »» ACKNOWLEDGEMENTS

*We would like to express our sincerest gratitude to all those who have made *The Field Guide to Data Science* such a success.*

Thank you to the nearly 15,000 people who have downloaded the digital copy from our website and the 100+ people who have connected with *The Field Guide* on our GitHub page. We have been overwhelmed by the popularity of the work within the Data Science community.

Thank you to all of the practitioners who are using *The Field Guide* as a resource. We are excited to know that the work has had such a strong influence, from shaping technical approaches to serving as the foundation for the very definition and role of Data Science within major government and commercial organizations.

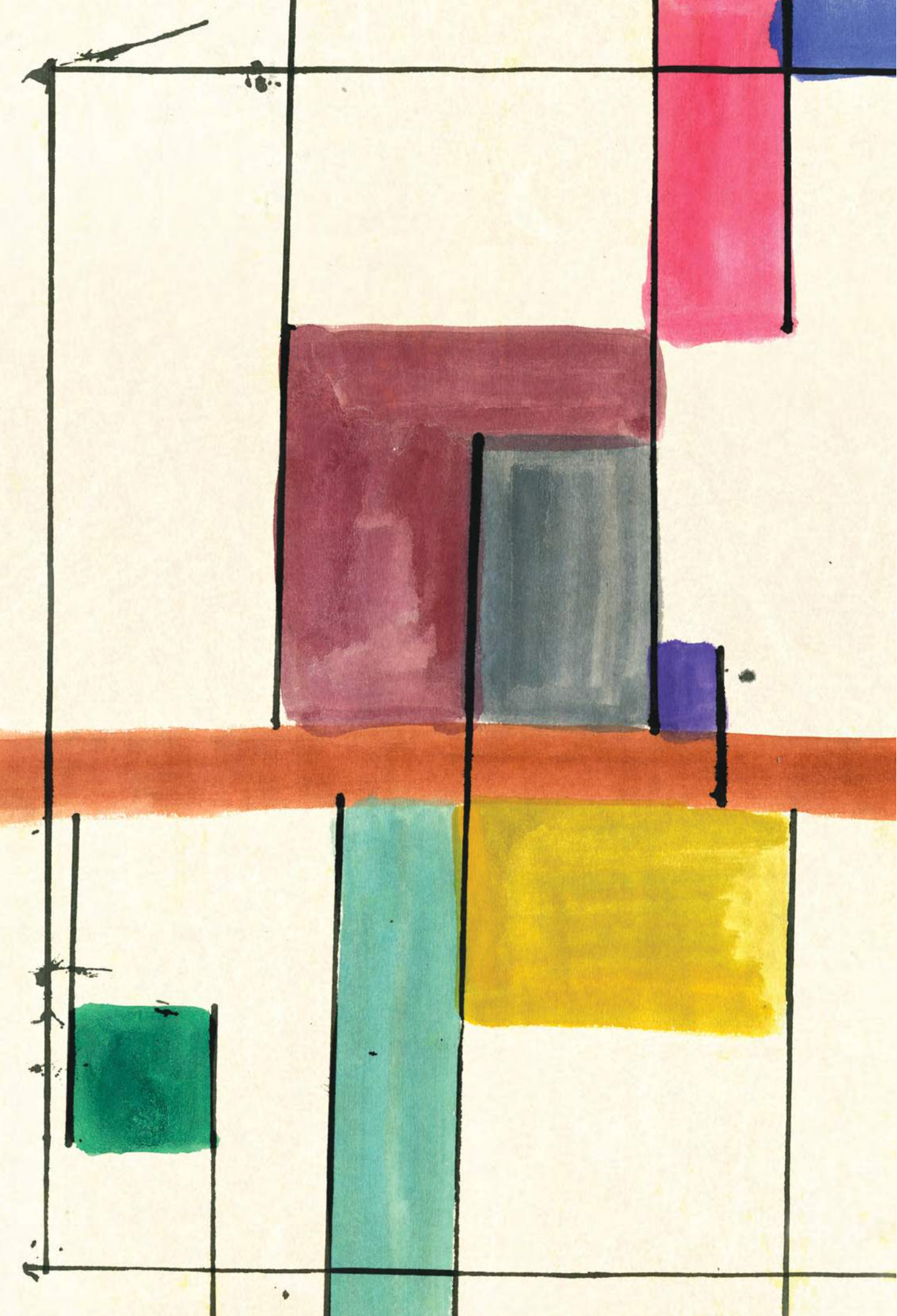
Thank you to the educators and academics who have incorporated *The Field Guide* into your course work. We appreciate your trusting this guide as a way to introduce your students to Data Science. It is an honor to know that we are shaping the next generation of Data Scientists.

Thank you to the organizational leaders who have shared your feedback, encouragement, and success stories. We are thrilled to know that *The Field Guide* has helped so many organizations, from energy, to life sciences, to retail, to begin their Data Science journeys.

---

We hope you will all continue to find value from *The Field Guide to Data Science* and to share in our excitement around the release of this second edition. Please continue to be part of the conversation and take this journey with us.

---



# »» THE OUTLINE of OUR STORY

- 12 »» **Meet Your Guides**
- 17 »» **The Short Version** – The Core Concepts of Data Science
- 18 »» **Start Here for the Basics** – An Introduction to Data Science
  - What Do We Mean by Data Science?
  - How Does Data Science Actually Work?
  - What Does It Take to Create a Data Science Capability?
- 46 »» **Take off the Training Wheels** – The Practitioner’s Guide to Data Science
  - Guiding Principles
  - The Importance of Reason
  - Component Parts of Data Science
  - Fractal Analytic Model
  - The Analytic Selection Process
  - Guide to Analytic Selection
  - Detailed Table of Analytics
- 84 »» **Life in the Trenches** – Navigating Neck Deep in Data
  - Going Deep into Machine Learning
  - Feature Engineering
  - Feature Selection
  - Ensemble Models
  - Data Veracity
  - Application of Domain Knowledge
  - The Curse of Dimensionality
  - Model Validation
- 102 »» **Putting it all Together** – Our Case Studies
  - Streamlining Medication Review
  - Reducing Flight Delays
  - Making Vaccines Safer
  - Forecasting the Relative Risk for the Onset of
    - Mass Killings to Help Prevent Future Atrocities
  - Predicting Customer Response
- 114 »» **Closing Time**
  - The Future of Data Science
  - Parting Thoughts
  - References
  - About Booz Allen Hamilton

# »» MEET *your* GUIDES



---

**Fred Blackburn**  
(@boozallen)

---

Data Science is a field that is evolving at a very rapid pace...be part of the journey.



---

**Josh Sullivan**  
(@joshdsullivan)

---

Leading our Data Science team shows me every day the incredible power of discovery and human curiosity. Don't be afraid to blend art and science to advance your own view of data analytics – it can be a powerful mixture.



---

**Peter Guerra**  
(@petrguerra)

---

Data Science is the most fascinating blend of art and math and code and sweat and tears. It can take you to the highest heights and the lowest depths in an instant, but it is the only way we will be able to understand and describe the why.



---

**Angela Zutavern**  
(@angelazutavern)

---

Data Science is about asking bigger questions, seeing future possibilities, and creating outcomes you desire.



---

**Steve Escaravage**  
(@sescarav)

---

Invest your time and energy in data that is difficult to assemble. If it doesn't exist, find a way to make it exist.



---

**Ezmeralda Khalil**  
(@ezmeraldakhalil)

---

The power of data science lies in the execution.



---

**Steven Mills**  
(@stevndmills)

---

Data Science truly can change the world.



---

**Alex Cosmas**  
(@boozallen)

---

Data scientists should be truth-seekers, not fact-seekers.



---

**Brian Keller**  
(@boozallen)

---

Grit will get you farther than talent.



---

**Stephanie Beben**  
(@boozallen)

---

Begin every new data challenge with deep curiosity along with a healthy dose of skepticism.



---

**Kirk Borne**  
(@KirkDBorne)

---

Focus on value, not volume.



---

**Drew Farris**  
(@drewfarris)

---

Don't forget to play. Play with tools, play with data, and play with algorithms. You just might discover something that will help you solve that next nagging problem.



---

**Paul Yacci**  
(@paulyacci)

---

In the jungle of data, don't miss the forest for the trees, or the trees for the forest.



---

**Charles Glover**  
(@MindAfterMath)

---

The beauty of data science lies in satisfying curiosities about important problems by playing with data and algorithms.



---

**Michael Kim**  
(@boozallen)

---

Data science is both an art and science.



---

**Stephanie Rivera**  
(@boozallen)

---

I treat Data Science like I do rock climbing: awesome dedication leads to incremental improvement. Persistence leads to the top.



---

**Aaron Sander**  
(@ajsander)

---

Data science is changing corporate culture to be more like the open source environment. More open, more collaborative, and faster paced.

*We would like to thank the following people for their contributions and edits:*

*Tim Andrews, Mike Delurey, Greg Dupier, Jason Escaravage, Christine Fantaskey, Juergen Klenk, Dan Liebermann, Mark Rockley and Katie Wilks.*

## »» COMMUNITY CONTRIBUTORS



---

**Will Cukierski**, [kaggle](#)  
([@kaggle](#))

---

Two roads diverged in a wood, and I—  
I took the one in the direction  
of the negative gradient,  
And that has made all the difference.



---

**Mark Herman**  
([@cloudEBITDA](#))

---

End every analysis with...  
'and therefore.'



---

**Ed Kohlwey**  
([@ekohlwey](#))

---

Data Science is about formally  
analyzing everything around you  
and becoming data driven.



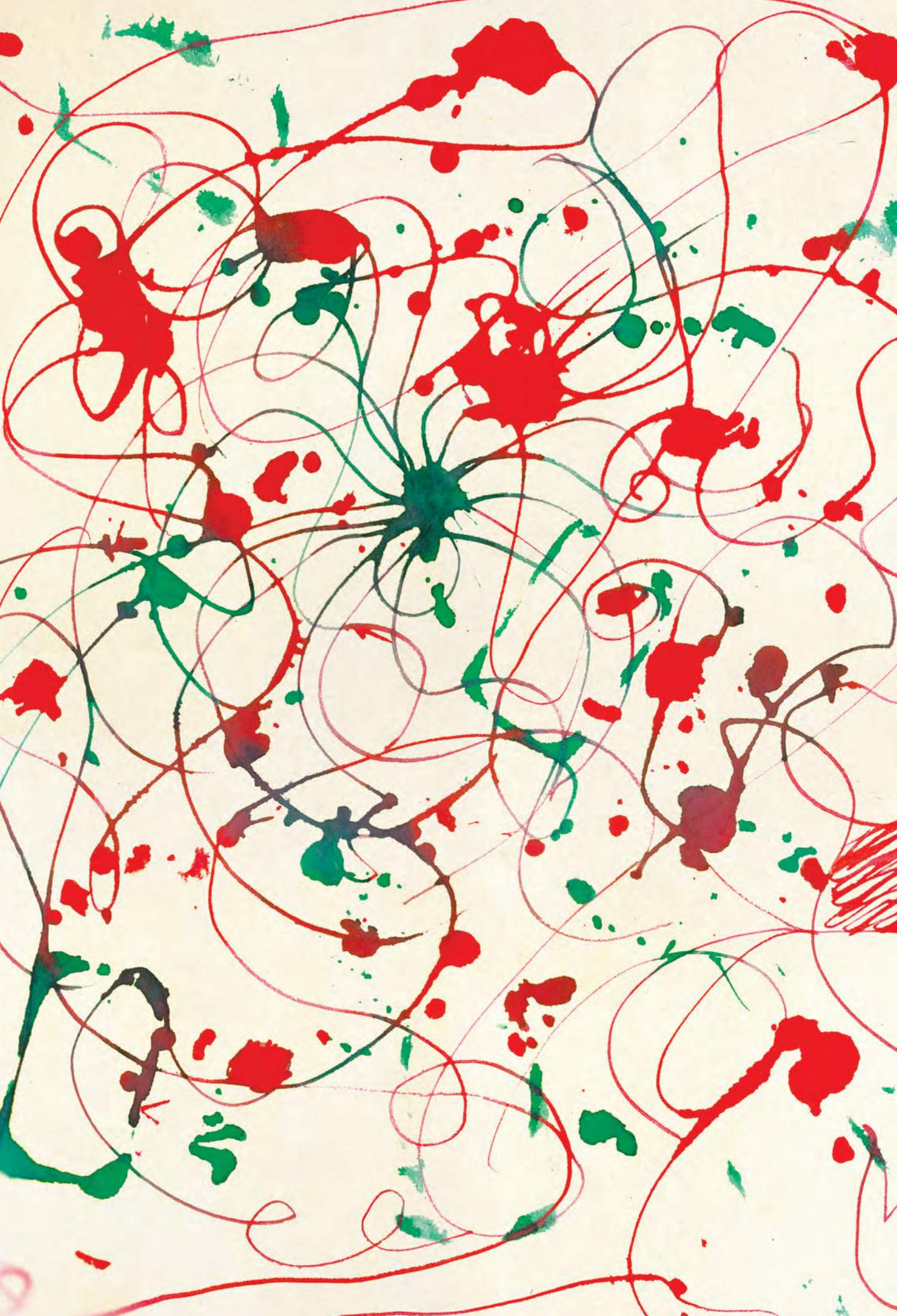
---

**Armen Kherlopian**  
([@akherlopian](#))

---

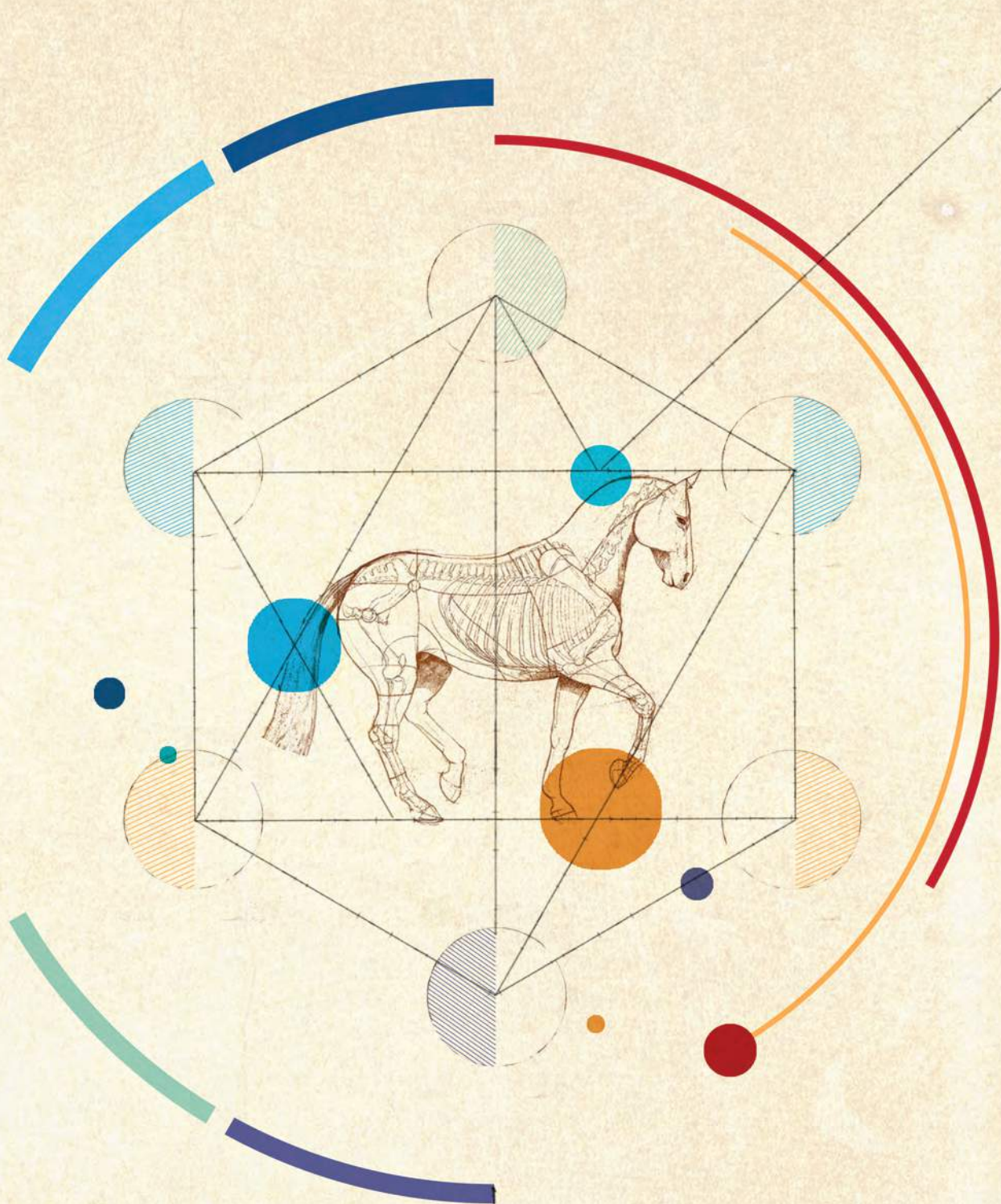
A Data Scientist must  
continuously seek truth in spite  
of ambiguity; therein rests the  
basis of rigor and insight.





# »» *The* SHORT VERSION

- › ***Data Science is the art of turning data into actions.***  
It's all about the tradecraft. Tradecraft is the process, tools and technologies for humans and computers to work together to transform data into insights.
- › ***Data Science tradecraft creates data products.***  
Data products provide actionable information without exposing decision makers to the underlying data or analytics (e.g., buy/sell strategies for financial instruments, a set of actions to improve product yield, or steps to improve product marketing).
- › ***Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning.***  
This is a fundamental change from traditional analysis approaches. Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths. Models of reality no longer need to be static. They are constantly tested, updated and improved until better models are found.
- › ***Data Science is necessary for companies to stay with the pack and compete in the future.***  
Organizations are constantly making decisions based on gut instinct, loudest voice and best argument – sometimes they are even informed by real information. The winners and the losers in the emerging data economy are going to be determined by their Data Science teams.
- › ***Data Science capabilities can be built over time.***  
Organizations mature through a series of stages – Collect, Describe, Discover, Predict, Advise – as they move from data deluge to full Data Science maturity. At each stage, they can tackle increasingly complex analytic goals with a wider breadth of analytic capabilities. However, organizations need not reach maximum Data Science maturity to achieve success. Significant gains can be found in every stage.
- › ***Data Science is a different kind of team sport.***  
Data Science teams need a broad view of the organization. Leaders must be key advocates who meet with stakeholders to ferret out the hardest challenges, locate the data, connect disparate parts of the business, and gain widespread buy-in.





# START HERE *for* **THE BASICS**

## **AN INTRODUCTION TO DATA SCIENCE**

If you haven't heard of Data Science, you're behind the times. Just renaming your Business Intelligence group the Data Science group is not the solution.



# What do We Mean by Data Science?

---

Describing Data Science is like trying to describe a sunset – it should be easy, but somehow capturing the words is impossible.

---

# Data Science Defined

Data Science is the art of turning data into actions. This is accomplished through the creation of data products, which provide actionable information without exposing decision makers to the underlying data or analytics (e.g., buy/sell strategies for financial instruments, a set of actions to improve product yield, or steps to improve product marketing).

Performing Data Science requires the extraction of timely, actionable information from diverse data sources to drive data products.

Examples of data products include answers to questions such as: “Which of my products should I advertise more heavily to increase profit? How can I improve my compliance program, while reducing costs? What manufacturing process change will allow me to build a better product?” The key to answering these questions is: understand the data you have and what the data inductively tells you.



## » Data Product

---

A data product provides actionable information without exposing decision makers to the underlying data or analytics. Examples include:

- Movie Recommendations
- Weather Forecasts
- Stock Market Predictions
- Production Process Improvements
- Health Diagnosis
- Flu Trend Predictions
- Targeted Advertising

---

### *Read this for additional background:*

The term Data Science appeared in the computer science literature throughout the 1960s-1980s. It was not until the late 1990s however, that the field as we describe it here, began to emerge from the statistics and data mining communities (e.g., <sup>[2]</sup> and <sup>[3]</sup>). Data Science was first introduced as an independent discipline in 2001.<sup>[4]</sup> Since that time, there have been countless articles advancing the discipline, culminating with Data Scientist being declared the sexiest job of the 21<sup>st</sup> century.<sup>[5]</sup>

We established our first Data Science team at Booz Allen in 2010. It began as a natural extension of our Business Intelligence and cloud

infrastructure development work. We saw the need for a new approach to distill value from our clients’ data. We approached the problem with a multidisciplinary team of computer scientists, mathematicians and domain experts. They immediately produced new insights and analysis paths, solidifying the validity of the approach. Since that time, our Data Science team has grown to 250 staff supporting dozens of clients across a variety of domains. This breadth of experience provides a unique perspective on the conceptual models, tradecraft, processes and culture of Data Science.

## What makes Data Science Different?

Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning. This is a fundamental change from traditional analytic approaches. Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths. In fact, to do the discovery of significant insights that are the hallmark of Data Science, you must have the tradecraft and the interplay between inductive and deductive reasoning. By actively combining the ability to reason deductively and inductively, Data Science creates an environment where models of reality no longer need to be static and empirically based. Instead, they are constantly tested, updated and improved until better models are found. These concepts are summarized in the figure, *The Types of Reason and Their Role in Data Science Tradecraft*.

### THE TYPES OF REASON...

#### DEDUCTIVE REASONING:

- › Commonly associated with “*formal logic*.”
- › Involves reasoning from known premises, or premises presumed to be true, to a certain conclusion.
- › The conclusions reached are certain, inevitable, inescapable.

#### INDUCTIVE REASONING

- › Commonly known as “*informal logic*,” or “everyday argument.”
- › Involves drawing uncertain inferences, based on probabilistic reasoning.
- › The conclusions reached are probable, reasonable, plausible, believable.

### ...AND THEIR ROLE IN DATA SCIENCE TRADECRAFT.

#### DEDUCTIVE REASONING:

- › Formulate hypotheses about relationships and underlying models.
- › Carry out experiments with the data to test hypotheses and models.

#### INDUCTIVE REASONING

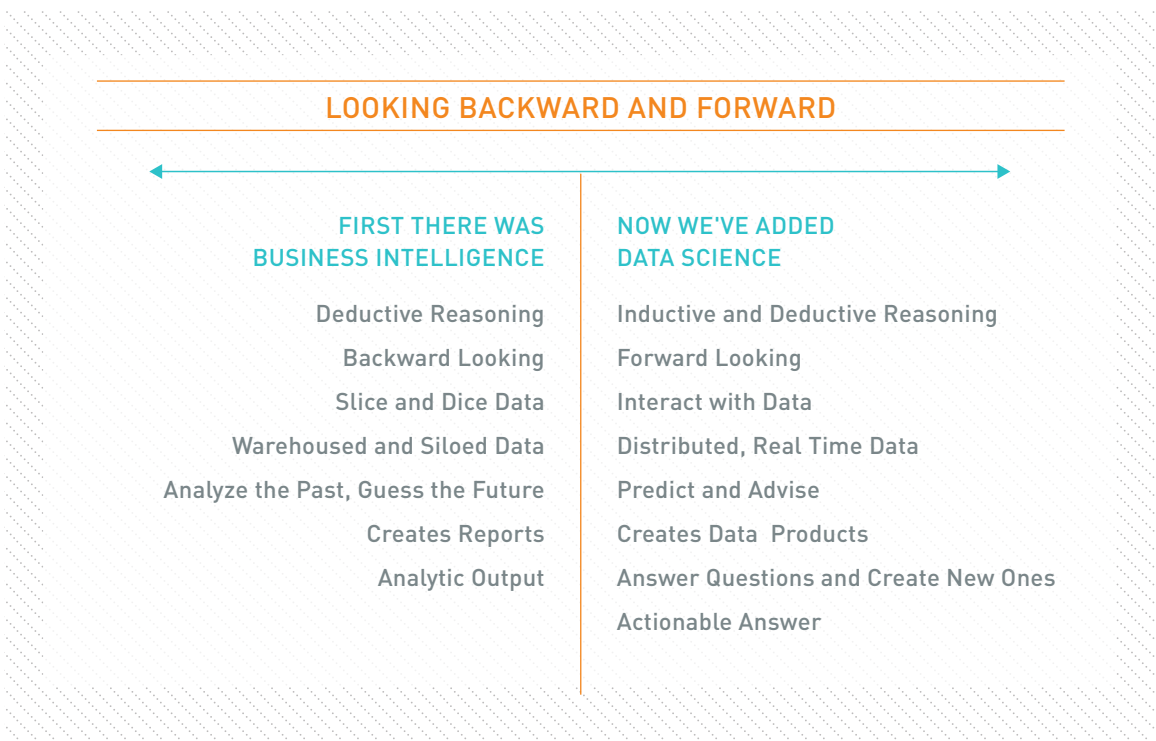
- › Exploratory data analysis to discover or refine hypotheses.
- › Discover new relationships, insights and analytic paths from the data.

Source: Booz Allen Hamilton

The Types of Reason and Their Role in Data Science Tradecraft

The differences between Data Science and traditional analytic approaches do not end at seamless shifting between deductive and inductive reasoning. Data Science offers a distinctly different perspective than capabilities such as Business Intelligence. Data Science should not replace Business Intelligence functions within an organization, however. The two capabilities are additive and complementary, each offering a necessary view of business operations and the operating environment. The figure, *Business Intelligence and Data Science – A Comparison*, highlights the differences between the two capabilities. Key contrasts include:

- › **Discovery vs. Pre-canned Questions:** Data Science actually works on discovering the question to ask as opposed to just asking it.
- › **Power of Many vs. Ability of One:** An entire team provides a common forum for pulling together computer science, mathematics and domain expertise.
- › **Prospective vs. Retrospective:** Data Science is focused on obtaining actionable information from data as opposed to reporting historical facts.



Source: Booz Allen Hamilton

Business Intelligence and Data Science - A Comparison (adapted in part from [6])



## What is the Impact of Data Science?

As we move into the data economy, Data Science is the competitive advantage for organizations interested in winning – in whatever way winning is defined. The manner in which the advantage is defined is through improved decision-making. A former colleague liked to describe data-informed decision making like this: *If you have perfect information or zero information then your task is easy – it is in between those two extremes that the trouble begins.* What he was highlighting is the stark reality that whether or not information is available, decisions must be made.

The way organizations make decisions has been evolving for half a century. Before the introduction of Business Intelligence, the only options were gut instinct, loudest voice, and best argument. Sadly, this method still exists today, and in some pockets it is the predominant means by which the organization acts. Take our advice and never, ever work for such a company!

Fortunately for our economy, most organizations began to inform their decisions with real information through the application of simple statistics. Those that did it well were rewarded; those that did not failed. We are outgrowing the ability of simple stats to keep pace with market demands, however. The rapid expansion of available data and the tools to access and make use of the data at scale are enabling fundamental changes to the way organizations make decisions.

Data Science is required to maintain competitiveness in the increasingly data-rich environment. Much like the application of simple statistics, organizations that embrace Data Science will be rewarded while those that do not will be challenged to keep pace. As more complex, disparate datasets become available, the chasm between these groups will only continue to widen. The figure, *The Business Impacts of Data Science*, highlights the value awaiting organizations that embrace Data Science.

---

## DATA SCIENCE IS NECESSARY...

---

- 17-49%** increase in productivity when organizations increase data usability by 10%
- 11-42%** return on assets (ROA) when organizations increase data access by 10%
- 241%** increase in ROI when organizations use big data to improve competitiveness
- 1000%** increase in ROI when deploying analytics across most of the organization, aligning daily operations with senior management's goals, and incorporating big data
- 5-6%** performance improvement for organizations making data-driven decisions.

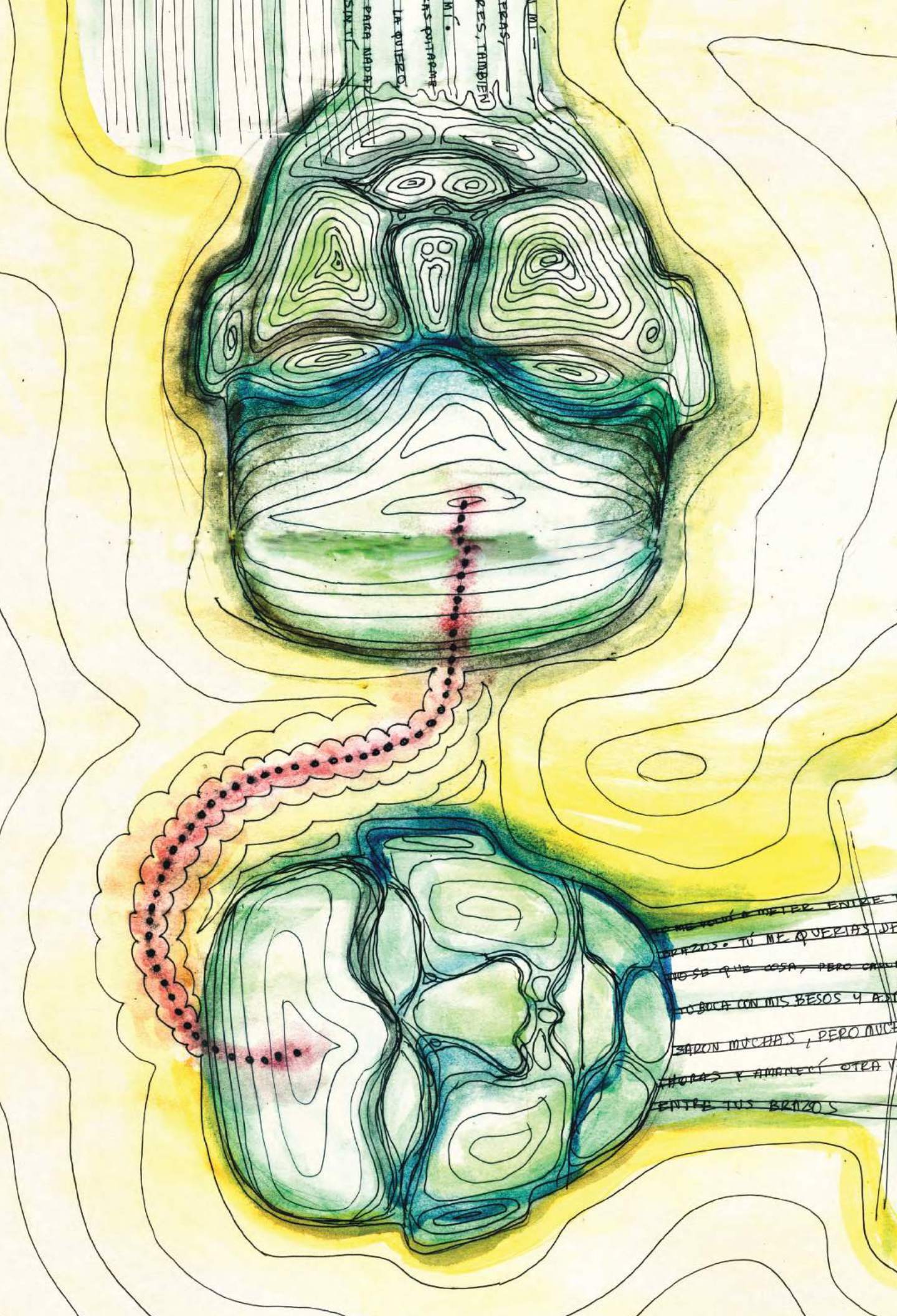
---

## ...TO COMPETE IN THE FUTURE

---

Source: Booz Allen Hamilton

The Business Impacts of Data Science (adapted from [7], [8] and [9])



MI...  
PARAS...  
RES, TAMBIEN  
MI...  
AS PUEDE  
LA QUIERO  
PARA MI  
SIN TI

...METER ENTRE  
...TU ME QUERIAS DE  
...QUE COSA, PERO CA  
...CON MIS BESOS Y A  
...MUCHAS, PERO MUCH  
...AMOR OTRO  
...TUS BRAZOS

## What is Different Now?

For 20 years IT systems were built the same way. We separated the people who ran the business from the people who managed the infrastructure (and therefore saw data as simply another thing they had to manage). With the advent of new technologies and analytic techniques, this artificial – and highly ineffective – separation of critical skills is no longer necessary. For the first time, organizations can directly connect business decision makers to the data. This simple step transforms data from being ‘something to be managed’ into ‘something to be valued.’

In the wake of the transformation, organizations face a stark choice: you can continue to build data silos and piece together disparate information or you can consolidate your data and distill answers.

From the Data Science perspective, this is a false choice: The siloed approach is untenable when you consider the (a) the opportunity cost of not making maximum use of all available data to help an organization succeed, and (b) the resource and time costs of continuing down the same path with outdated processes. 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.

# How does Data Science Actually Work?

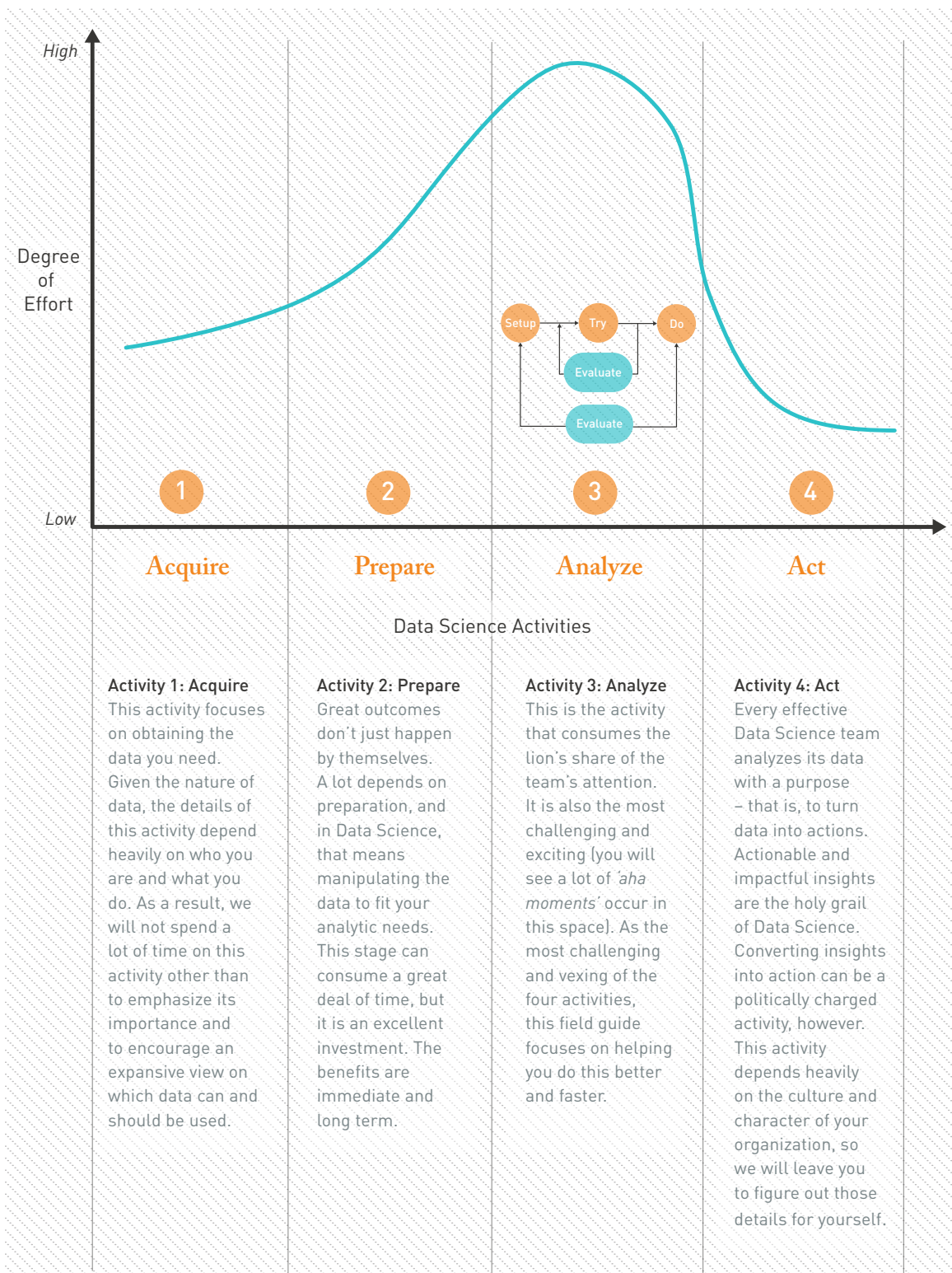
---

It's not rocket science... it's something better - Data Science

---

Let's not kid ourselves - Data Science is a complex field. It is difficult, intellectually taxing work, which requires the sophisticated integration of talent, tools and techniques. But as a field guide, we need to cut through the complexity and provide a clear, yet effective way to understand this new world.

To do this, we will transform the field of Data Science into a set of simplified activities as shown in the figure, *The Four Key Activities of a Data Science Endeavor*. Data Science purists will likely disagree with this approach, but then again, they probably don't need a field guide, sitting as they do in their ivory towers! In the real world, we need clear and simple operating models to help drive us forward.



Source: Booz Allen Hamilton

The Four Key Activities of a Data Science Endeavor

# 1 Acquire

All analysis starts with access to data, and for the Data Scientist this axiom holds true. But there are some significant differences – particularly with respect to the question of who stores, maintains and owns the data in an organization.

But before we go there, let's look at what is changing. Traditionally, rigid data silos artificially define the data to be acquired. Stated another way, the silos create a filter that lets in a very small amount of data and ignores the rest. These filtered processes give us an artificial view of the world based on the 'surviving data,' rather than one that shows full reality and meaning. Without a broad and expansive dataset, we can never immerse ourselves in the diversity of the data. We instead make decisions based on limited and constrained information.

Eliminating the need for silos gives us access to all the data at once – including data from multiple outside sources. It embraces the reality that *diversity is good and complexity is okay*. This mindset creates a completely different way of thinking about data in an organization by giving it a new and differentiated role. Data represents a significant new profit and mission-enhancement opportunity for organizations.

But as mentioned earlier, this first activity is heavily dependent upon the situation and circumstances. We can't leave you with anything more than general guidance to help ensure maximum value:



## » Not All Data Is Created Equal

As you begin to aggregate data, remember that not all data is created equally. Organizations have a tendency to collect any data that is available. Data that is nearby (readily accessible and easily obtained) may be cheap to collect, but there is no guarantee it is the right data to collect. Focus on the data with the highest ROI for your organization. Your Data Science team can help identify that data. Also remember that you need to strike a balance between the data that you need and the data that you have. Collecting huge volumes of data is useless and costly if it is not the data that you need.

- › **Look inside first:** What data do you have current access to that you are not using? This is in large part the data being left behind by the filtering process, and may be incredibly valuable.
- › **Remove the format constraints:** Stop limiting your data acquisition mindset to the realm of structured databases. Instead, think about unstructured and semi-structured data as viable sources.
- › **Figure out what's missing:** Ask yourself what data would make a big difference to your processes if you had access to it. Then go find it!
- › **Embrace diversity:** Try to engage and connect to publicly available sources of data that may have relevance to your domain area.

## 2 Prepare

Once you have the data, you need to prepare it for analysis.

Organizations often make decisions based on inexact data. Data stovepipes mean that organizations may have blind spots. They are not able to see the whole picture and fail to look at their data and challenges holistically. The end result is that valuable information is withheld from decision makers. Research has shown almost 33% of decisions are made without good data or information.<sup>[10]</sup>

When Data Scientists are able to explore and analyze all the data, new opportunities arise for analysis and data-driven decision making. The insights gained from these new opportunities will significantly change the course of action and decisions within an organization. Gaining access to an organization's complete repository of data, however, requires preparation.

Our experience shows time and time again that the best tool for Data Scientists to prepare for analysis is a lake – specifically, the Data Lake.<sup>[11]</sup> This is a new approach to collecting, storing and integrating data that helps organizations maximize the utility of their data. Instead of storing information in discrete data structures, the Data Lake consolidates an organization's complete repository of data in a single, large view. It eliminates the expensive and cumbersome data-preparation process, known as Extract/Transform/Load (ETL), necessary with data silos. The entire body of information in the Data Lake is available for every inquiry – and all at once.

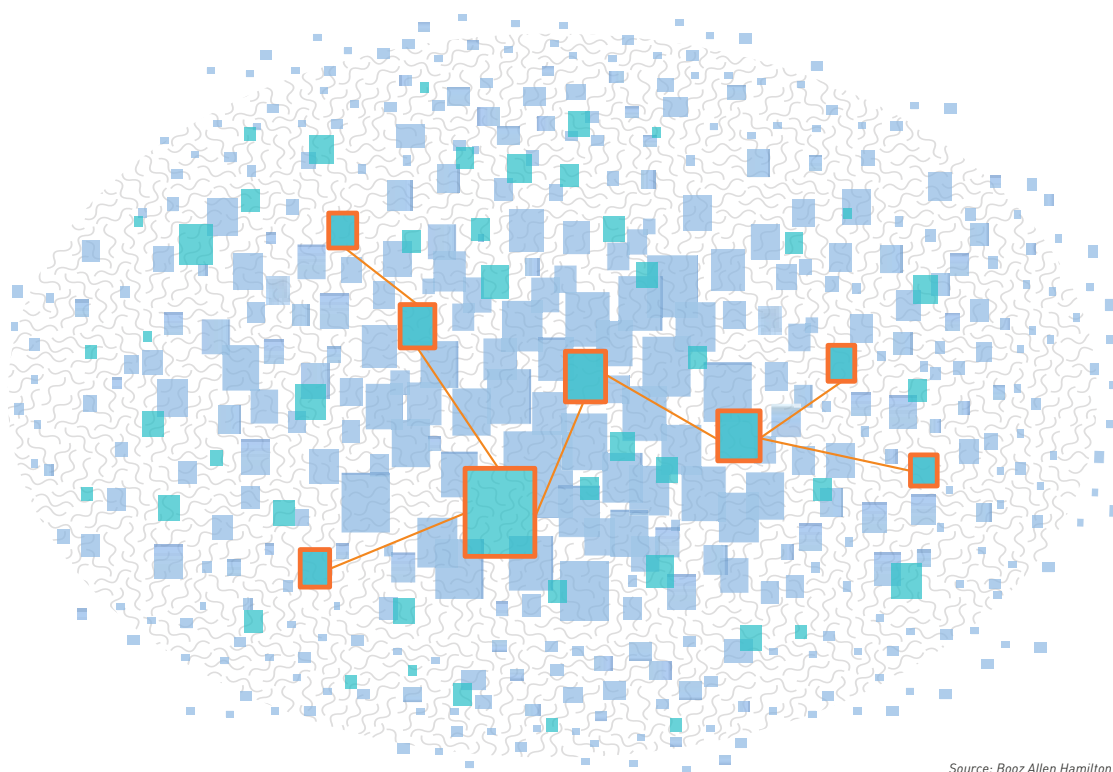


### 3 Analyze

We have acquired the data... we have prepared it... now it is time to analyze it.

The Analyze activity requires the greatest effort of all the activities in a Data Science endeavor. The Data Scientist actually builds the analytics that create value from data. Analytics in this context is an iterative application of specialized and scalable computational resources and tools to provide relevant insights from exponentially growing data. This type of analysis enables real-time understanding of risks and opportunities by evaluating situational, operational and behavioral data.

With the totality of data fully accessible in the Data Lake, organizations can use analytics to find the kinds of connections and patterns that point to promising opportunities. This high-speed analytic connection is done within the Data Lake, as opposed to older style sampling methods that could only make use of a narrow slice of the data. In order to understand what was in the lake, you had to bring the data out and study it. Now you can dive into the lake, bringing your analytics to the data. The figure, *Analytic Connection in the Data Lake*, highlights the concept of diving into the Data Lake to discover new connections and patterns.



Source: Booz Allen Hamilton

Analytic Connection in the Data Lake

Data Scientists work across the spectrum of analytic goals – Describe, Discover, Predict and Advise. The maturity of an analytic capability determines the analytic goals encompassed. Many variables play key roles in determining the difficulty and suitability of each goal for an organization. Some of these variables are the size and budget of an organization and the type of data products needed by the decision makers. A detailed discussion on analytic maturity can be found in *Data Science Maturity within an Organization*.

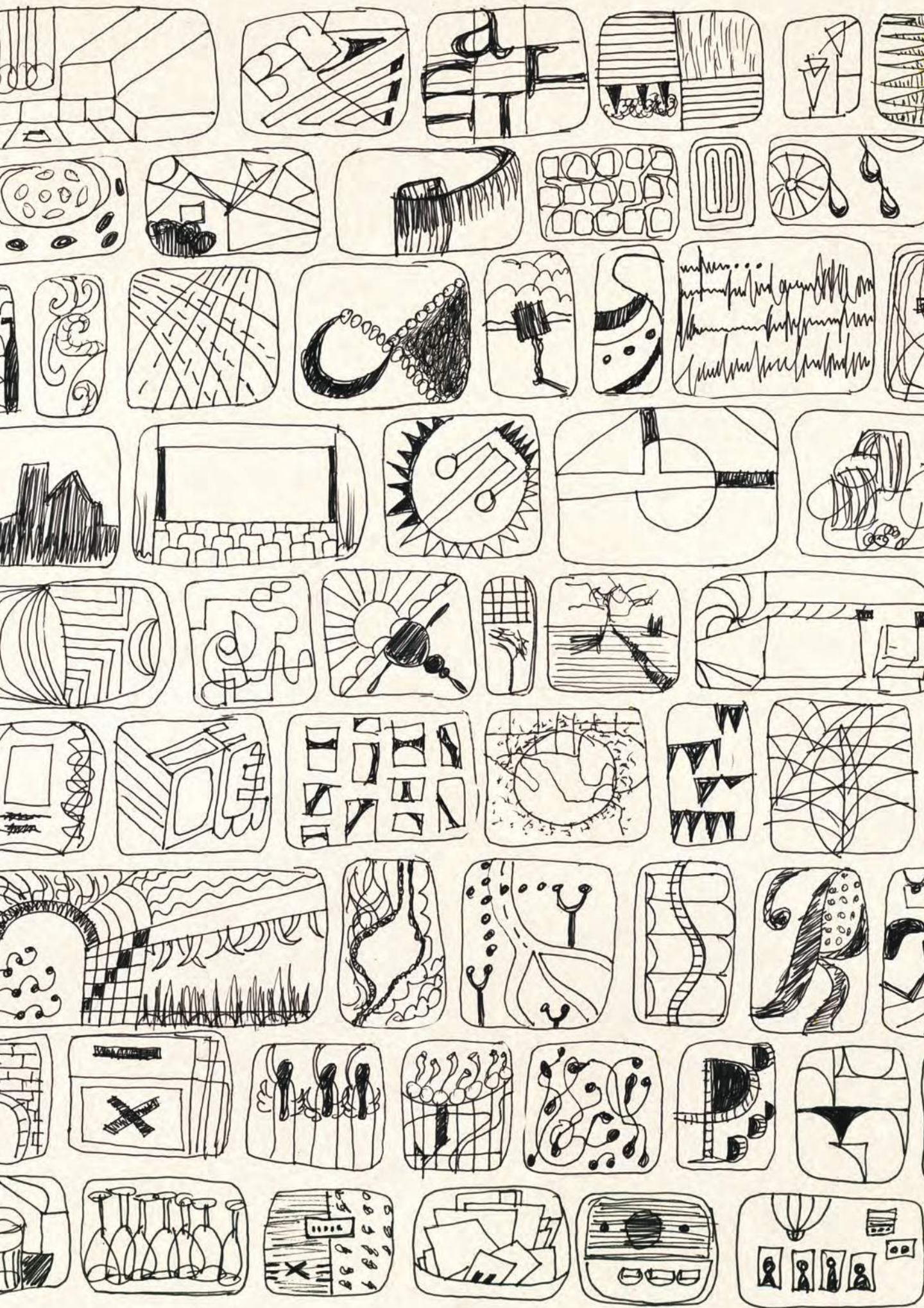
In addition to consuming the greatest effort, the Analyze activity is by far the most complex. The tradecraft of Data Science is an art. While we cannot teach you how to be an artist, we can share foundational tools and techniques that can help you be successful. The entirety of *Take Off the Training Wheels* is dedicated to sharing insights we have learned over time while serving countless clients. This includes descriptions of a Data Science product lifecycle and the *Fractal Analytic Model* (FAM). The *Analytic Selection Process* and accompanying *Guide to Analytic Selection* provide key insights into one of the most challenging tasks in all of Data Science – selecting the right technique for the job.

## 4 Act

Now that we have analyzed the data, it's time to take action.

The ability to make use of the analysis is critical. It is also very situational. Like the Acquire activity, the best we can hope for is to provide some guiding principles to help you frame the output for maximum impact. Here are some key points to keep in mind when presenting your results:

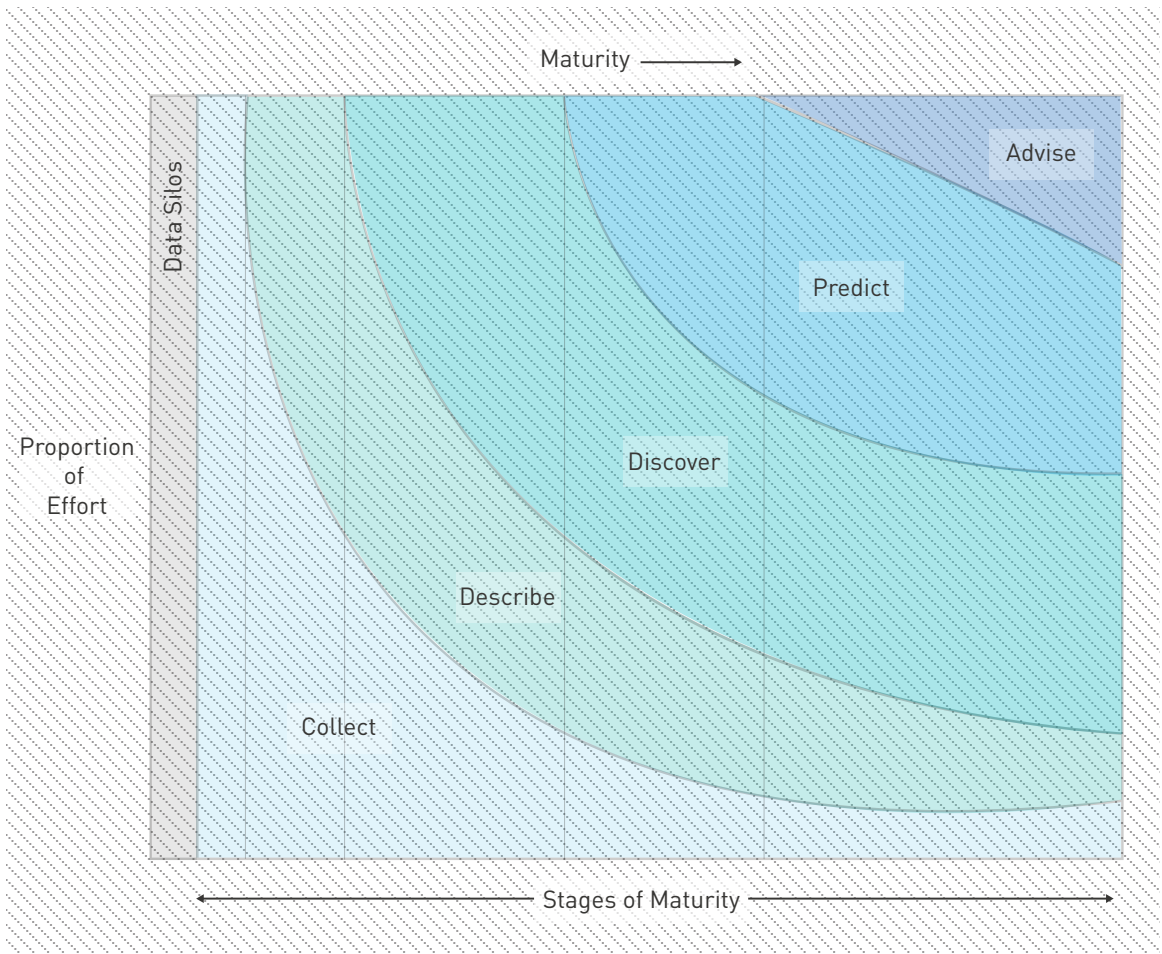
1. The finding must make sense with relatively little up-front training or preparation on the part of the decision maker.
2. The finding must make the most meaningful patterns, trends and exceptions easy to see and interpret.
3. Every effort must be made to encode quantitative data accurately so the decision maker can accurately interpret and compare the data.
4. The logic used to arrive at the finding must be clear and compelling as well as traceable back through the data.
5. The findings must answer real business questions.



# Data Science Maturity within an Organization

The four activities discussed thus far provide a simplified view of Data Science. Organizations will repeat these activities with each new Data Science endeavor. Over time, however, the level of effort necessary for each activity will change. As more data is Acquired and Prepared in the Data Lake, for example, significantly less effort will need to be expended on these activities. This is indicative of a maturing Data Science capability.

Assessing the maturity of your Data Science capability calls for a slightly different view. We use *The Data Science Maturity Model* as a common framework for describing the maturity progression and components that make up a Data Science capability. This framework can be applied to an organization's Data Science capability or even to the maturity of a specific solution, namely a data product. At each stage of maturity, powerful insight can be gained.



Source: Booz Allen Hamilton

The Data Science Maturity Model

When organizations start out, they have *Data Silos*. At this stage, they have not carried out any broad Aggregate activities. They may not have a sense of all the data they have or the data they need. The decision to create a Data Science capability signals the transition into the *Collect* stage.

All of your initial effort will be focused on identifying and aggregating data. Over time, you will have the data you need and a smaller proportion of your effort can focus on *Collect*. You can now begin to *Describe* your data. Note, however, that while the proportion of time spent on *Collect* goes down dramatically, it never goes away entirely. This is indicative of the four activities outlined earlier – you will continue to *Aggregate* and *Prepare* data as new analytic questions arise, additional data is needed and new data sources become available.

Organizations continue to advance in maturity as they move through the stages from *Describe* to *Advise*. At each stage they can tackle increasingly complex analytic goals with a wider breadth of analytic capabilities. As described for *Collect*, each stage never goes away entirely. Instead, the proportion of time spent focused on it goes down and new, more mature activities begin. A brief description of each stage of maturity is shown in the table *The Stages of Data Science Maturity*.

The Stages of Data Science Maturity

Stage	Description	Example
<b>Collect</b>	Focuses on collecting internal or external datasets.	Gathering sales records and corresponding weather data.
<b>Describe</b>	Seeks to enhance or refine raw data as well as leverage basic analytic functions such as counts.	How are my customers distributed with respect to location, namely zip code?
<b>Discover</b>	Identifies hidden relationships or patterns.	Are there groups within my regular customers that purchase similarly?
<b>Predict</b>	Utilizes past observations to predict future observations.	Can we predict which products that certain customer groups are more likely to purchase?
<b>Advise</b>	Defines your possible decisions, optimizes over those decisions, and advises to use the decision that gives the best outcome.	Your advice is to target advertise to specific groups for certain products to maximize revenue.

Source: Booz Allen Hamilton

The maturity model provides a powerful tool for understanding and appreciating the maturity of a Data Science capability. Organizations need not reach maximum maturity to achieve success. Significant gains can be found in every stage. We believe strongly that one does not engage in a Data Science effort, however, unless it is intended to produce an output – that is, you have the intent to *Advise*. This means simply that each step forward in maturity drives you to the right in the model diagram. Moving to the right requires the correct processes, people, culture and operating model – a robust Data Science capability. *What Does it Take to Create a Data Science Capability?* addresses this topic.

We have observed very few organizations actually operating at the highest levels of maturity, the *Predict* and *Advise* stages. The tradecraft of *Discover* is only now maturing to the point that organizations can focus on advanced *Predict* and *Advise* activities. This is the new frontier of Data Science. This is the space in which we will begin to understand how to close the cognitive gap between humans and computers. Organizations that reach *Advise* will be met with true insights and real competitive advantage.



## » Where does your organization fall in analytic maturity?

### Take the quiz!

#### 1. How many data sources do you collect?

- a. Why do we need a bunch of data? – 0 points, end here.
- b. I don't know the exact number. – 5 points
- c. We identified the required data and collect it. – 10 points

#### 2. Do you know what questions your Data Science team is trying to answer?

- a. Why do we need questions? – 0 points
- b. No, they figure it out for themselves. – 5 points
- c. Yes, we evaluated the questions that will have the largest impact to the business. – 10 points

#### 3. Do you know the important factors driving your business?

- a. I have no idea. – 0 points
- b. Our quants help me figure it out. – 5 points
- c. We have a data product for that. – 10 points

#### 4. Do you have an understanding of future conditions?

- a. I look at the current conditions and read the tea leaves. – 0 points
- b. We have a data product for that. – 5 points

#### 5. Do you know the best course of action to take for your key decisions?

- a. I look at the projections and plan a course. – 0 points
- b. We have a data product for that. – 5 points

### Check your score:

0 – Data Silos, 5-10 – Collect, 10-20 – Describe, 20-30 – Discover, 30-35 – Predict, 35-40 – Advise

Source: Booz Allen Hamilton

# What Does it Take to Create a Data Science Capability?

---

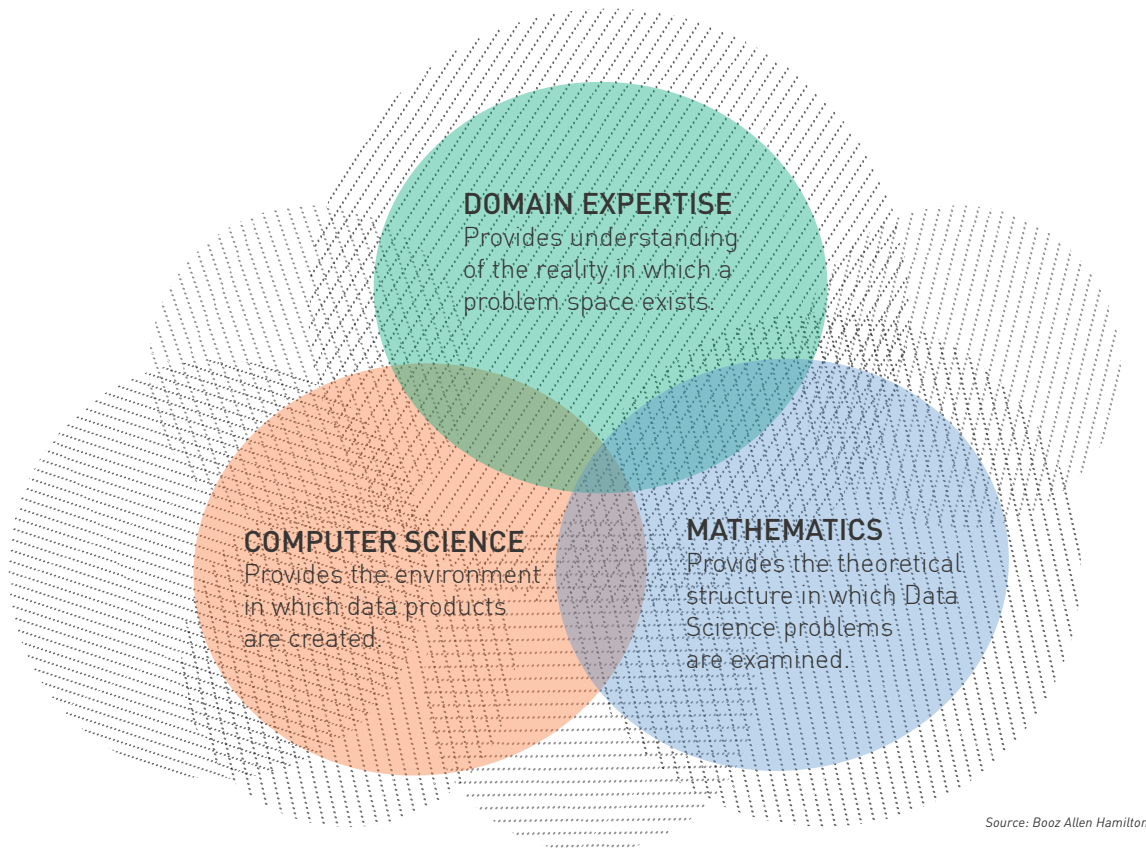
Data Science is all about building teams and culture.

---

Many organizations (both commercial and government) see the potential in capitalizing on data to unlock operational efficiencies, to create new services and experiences, and to propel innovation. Unfortunately, too many business leaders invest in one-off technical solutions— with a big price tag and mixed results— instead of investing in building a strategic Data Science capability. A Data Science capability embeds and operationalizes Data Science across an enterprise such that it can deliver the next level of organizational performance and return on investment. A Data Science capability moves an organization beyond performing pockets of analytics to an enterprise approach that uses analytical insights as part of the normal course of business. When building a capability, it is important for an organization to first identify its analytic goals (i.e., what it is trying to achieve through analytics) and then assess its readiness to achieve those goals – examining both technical readiness and organizational readiness. An organization can then make strategic choices on how to address gaps and begin to build their capability.

## Building Your Data Science Team

A critical component to any Data Science capability is having the right team. Data Science depends on a diverse set of skills as shown in The Data Science Venn Diagram. Computers provide the environment in which data-driven hypotheses are tested, and as such, computer science is necessary for data manipulation and processing. Mathematics provides the theoretical structure in which Data Science problems are examined. A rich background in statistics, geometry, linear algebra, and calculus are all important to understand the basis for many algorithms and tools. Finally, domain expertise contributes to an understanding of what problems actually need to be solved, what kind of data exists in the domain, and how the problem space may be instrumented and measured.



The Data Science Venn Diagram (inspired by <sup>[12]</sup>)

Remember that Data Science is a team sport. Most of the time, you will not be able to find the rare “unicorns” - people with expertise across all three of the skill areas. Therefore, it is important to build a blended team that covers all three elements of the Data Science Venn Diagram.



---

## BALANCING THE DATA SCIENCE TEAM EQUATION

---

Balancing the composition of a Data Science team is much like balancing the reactants and products in a chemical reaction. Each side of the equation must represent the same quantity of any particular element. In the case of Data Science, these elements are the foundational technical skills Computer Science (CS), Mathematics (M) and Domain Expertise (DE). The reactants, your Data Scientists, each have their own unique skills compositions. You must balance the staff mix to meet the skill requirements of the Data Science team, the product in the reaction. If you don't correctly balance the equation, your Data Science team will not have the desired impact on the organization.



In the example above, your project requires four parts computer science, five parts mathematics and one part domain expertise. Given the skills mix of the staff, five people are needed to balance the equation. Throughout your Data Science project, the skills requirements of the team will change. You will need to re-balance the equation to ensure the reactants balance with the products.

Source: Booz Allen Hamilton

# Understanding What Makes a Data Scientist

Data Science often requires a significant investment of time across a variety of tasks. Hypotheses must be generated and data must be acquired, prepared, analyzed, and acted upon. Multiple techniques are often applied before one yields interesting results. If that seems daunting, it is because it is. Data Science is difficult, intellectually taxing work, which requires lots of talent: both tangible technical skills as well as the intangible “x-factors.”

There are four independent yet comprehensive foundational Data Science competency clusters that, when considered together, convey the essence of what it means to be a successful Data Scientist. There are also reach back competencies that complement the foundational clusters but do not define the core tradecraft or attributes of the Data Science team.

## Data Science Competency Framework

(see <sup>[13]</sup> for complete framework)

Clusters	Competencies	Description
<b>Technical: “Knows How and What to do”</b>	Advanced Mathematics; Computer Science; Data Mining and Integration; Database Science; Research Design; Statistical Modeling; Machine Learning; Operations Research; Programming and Scripting	The technical competency cluster depicts the foundational technical and specialty knowledge and skills needed for successful performance in each job or role.
<b>Data Science Consulting: “Can Do in a Client and Customer Environment”</b>	Collaboration and Teamwork; Communications; Data Science Consulting; Ethics and Integrity	The characteristics in the consulting competency cluster can help Data Scientists easily integrate into various market or domain contexts and partner with business units to understand the environment and solve complex problems.
<b>Cognitive: “Able to Do or Learn to Do”</b>	Critical Thinking; Inductive and Deductive Reasoning; Problem Solving	The cognitive competency cluster represents the type of critical thinking and reasoning abilities (both inductive and deductive) a Data Scientist should have to perform their job.
<b>Personality: “Willing or Motivated to Do”</b>	Adaptability/Flexibility; Ambiguity Tolerance; Detail Orientation; Innovation and Creativity; Inquisitiveness; Perseverance; Resilience and Hardiness; Self-Confidence; Work Ethic	The personality competency cluster describes the personality traits that drive behaviors that are beneficial to Data Scientists, such as inquisitiveness, creativity, and perseverance.
<b>Reach Back Competencies for Data Science Teams</b>		
Business Acumen; Data Visualization; Domain Expertise; Program Management		

Source: Booz Allen Hamilton



### » ‘The Triple Threat Unicorn’

Individuals who are great at all three of the Data Science foundational technical skills are like unicorns – very rare and if you’re ever lucky enough to find one they should be treated carefully. When you manage these people:

- › Encourage them to lead your team, but not manage it. Don’t bog them down with responsibilities of management that could be done by other staff.
- › Put extra effort into managing their careers and interests within your organization. Build opportunities for promotion into your organization that allow them to focus on mentoring other Data Scientists and progressing the state of the art while also advancing their careers.
- › Make sure that they have the opportunity to present and spread their ideas in many different forums, but also be sensitive to their time.



» Don't judge a book by its cover, or a Data Scientist by his or her degree in this case. Amazing Data Scientists can be found anywhere. Just look at the diverse and surprising sampling of degrees held by Our Experts:

- › Bioinformatics
- › Biomedical Engineering
  - › Biophysics
  - › Business
- › Computer Graphics
- › Computer Science
  - › English
- › Forest Management
  - › History
- › Industrial Engineering
- › Information Technology
  - › Mathematics
- › National Security Studies
  - › Operations Research
  - › Physics
- › Wildlife & Fisheries Management

The most important qualities of Data Scientists tend to be the intangible aspects of their personalities. Data Scientists are by nature curious, creative, focused, and detail-oriented.

- › **Curiosity** is necessary to peel apart a problem and examine the interrelationships between data that may appear superficially unrelated.
- › **Creativity** is required to invent and try new approaches to solving a problem, which often times have never been applied in such a context before.
- › **Focus** is required to design and test a technique over days and weeks, find it doesn't work, learn from the failure, and try again.
- › **Attention to Detail** is needed to maintain rigor, and to detect and avoid over-reliance on intuition when examining data.

We have found the single most important attribute is flexibility in overcoming setbacks - the willingness to abandon one idea and try a new approach. Often, Data Science is a series of dead ends before, at last, the way forward is identified. It requires a unique set of personality attributes to succeed in such an environment. Technical skills can be developed over time: the ability to be flexible - and patient, and persistent - cannot.

## Finding the Athletes for Your Team

Building a Data Science team is complex. Organizations must simultaneously engage existing internal staff to create an "anchor" who can be used to recruit and grow the team, while at the same time undergo organizational change and transformation to meaningfully incorporate this new class of employee. Building a team starts with identifying existing staff within an organization who have a high aptitude for Data Science. Good candidates will have a formal background in any of the three foundational technical skills we mentioned, and will most importantly have the personality traits necessary for Data Science. They may often have advanced (masters or higher) degrees, but not always. The very first staff you identify should also have good leadership traits and a sense of purpose for the organization, as they will lead subsequent staffing and recruiting efforts. Don't discount anyone - you will find Data Scientists in the strangest places with the oddest combinations of backgrounds.

## Shaping the Culture

It is no surprise—building a culture is hard and there is just as much art to it as there is science. It is about deliberately creating the conditions for Data Science to flourish (for both Data Scientists and the average employee). You can then step back to empower collective ownership of an organic transformation.

Data Scientists are fundamentally curious and imaginative. We have a saying on our team, “We’re not nosy, we’re Data Scientists.” These qualities are fundamental to the success of the project and to gaining new dimensions on challenges and questions. Often Data Science projects are hampered by the lack of the ability to imagine something new and different. Fundamentally, organizations must foster trust and transparent communication across all levels, instead of deference to authority, in order to establish a strong Data Science team. Managers should be prepared to invite participation more frequently, and offer explanation or apology less frequently.

It is important to provide a path into the Data Science “club” and to empower the average employee to feel comfortable and conversant with Data Science. For something to be part of organizational culture, it must be part of the fabric of the employee behavior. That means employees must interact with and use data products in their daily routines. Another key ingredient to shaping the right culture is that all employees need a baseline of Data Science knowledge, starting with a common lexicon, to facilitate productive collaboration and instill confidence. While not everyone will be Data Scientists, employees need to identify with Data Science and be equipped with the knowledge, skills, and abilities to work with Data Scientists to drive smarter decisions and deliver exponential organizational performance.



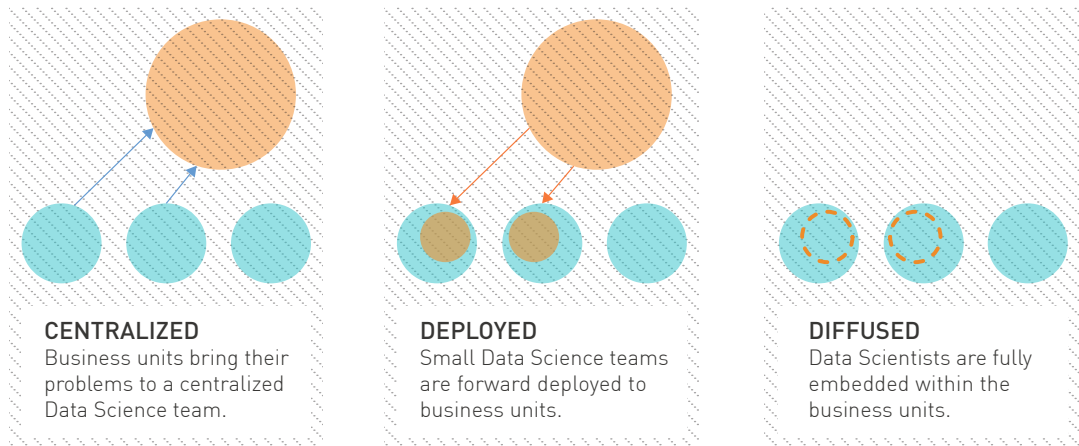
### » “I’m not nosy, I’m a Data Scientist”

---

- › Always remember that unrelenting curiosity and imagination should be the hallmarks of Data Science. They are fundamental to the success of every Data Science project.

## Selecting Your Operating Model

Depending on the size, complexity, and the business drivers, organizations should consider one of three Data Science operating models: Centralized, Deployed, or Diffused. These three models are shown in the figure, *Data Science Operating Models*.



**Data Science Operating Models**  
[see <sup>[13]</sup> for complete descriptions]

Source: Booz Allen Hamilton

**Centralized Data Science** teams serve the organization across all business units. The team is centralized under a Chief Data Scientist and they all co-locate together. The domain experts come to this organization for brief rotational stints to solve challenges around the business. This model provides greater efficiency with limited Data Science resources but can also create the perceived need to compete with other business units for Data Science talent. To address this challenge, it is important to place emphasis on portfolio management and creating transparency on how organizations will identify and select Data Science projects.

**Deployed Data Science** teams go to the business unit and reside there for short- or long-term assignments. They are their own entity and they work with the domain experts within the group to solve hard problems. In the deployed model, Data Science teams collectively develop knowledge across business units, with central leadership as a bridging mechanism for addressing organization-wide issues. However, Data Science teams are accountable to business unit leadership and their centralized leadership, which could cause confusion and conflict. In this model, it is important to emphasize conflict management to avoid competing priorities.

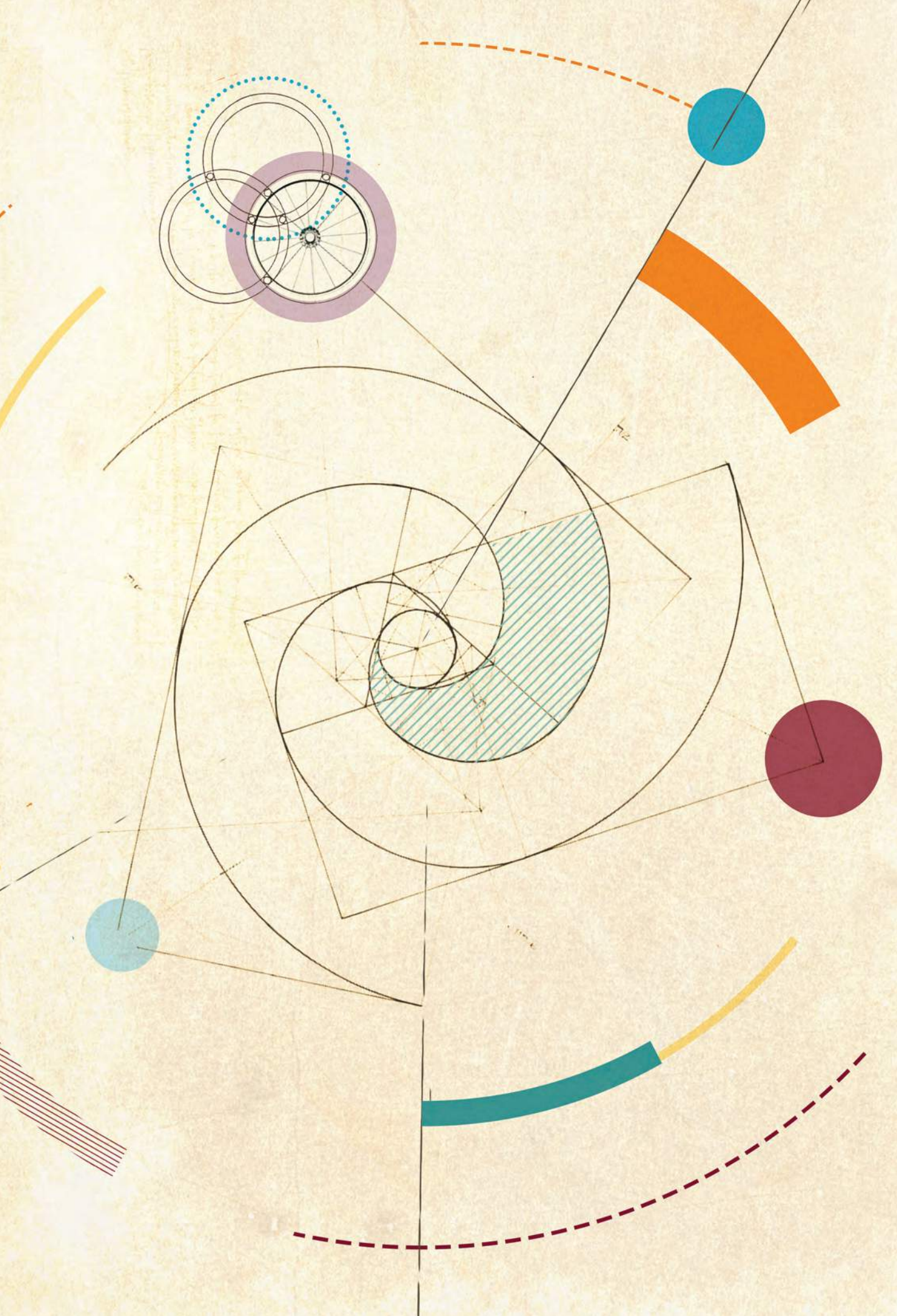
**The Diffused Data Science** team is one that is fully embedded with each group and becomes part of the long-term organization. These teams work best when the nature of the domain or business unit is already one focused on analytics. In the Diffused Model, teams can quickly react to high-priority business unit needs. However, the lack of central management can result in duplicate software and tools. Additionally, business units with the most money will often have full access to analytics while other units have none—this may not translate to the greatest organizational impact. In this model, it is important to establish cross-functional groups that promote organization-wide governance and peer collaboration.

Full descriptions of each operating model can be found in Booz Allen's *Tips for Building a Data Science Capability* <sup>[13]</sup>.

## How to Generate Momentum

A Data Science effort can start at the grass roots level by a few folks tackling hard problems, or as directed by the Chief Executive Officer, Chief Data Officer, or Chief Analytics Officer. Regardless of how an effort starts, political headwinds often present more of a challenge than solving any technical hurdles. To help battle the headwinds, it is important to generate momentum and prove the value a Data Science team can provide. The best way to achieve this is usually through a Data Science prototype or proof of concept. Proofs of concepts can generate the critical momentum needed to jump start any Data Science Capability Four qualities, in particular, are essential for every Data Science prototype:

1. **Organizational Buy-in:** A prototype will only succeed if the individuals involved believe in it and are willing to do what they can to make it successful. A good way to gauge interest is to meet with the middle managers; their views are usually indicative of the larger group.
2. **Clear ROI:** Before choosing a prototype problem, ensure that the ROI of the analytic output can be clearly and convincingly demonstrated for both the project and the organization as a whole. This outcome typically requires first reaching consensus on how the ROI will be determined and measured, so that the benefit can be quantified.
3. **Necessary Data:** Before selecting a prototype, you must first determine exactly what data is needed, whether it will actually be available, and what it will cost in terms of time and expense. It is important to note that organizations do not need all the possible data – they can still create successful analytics even with some gaps.
4. **Limited Complexity and Duration:** The problem addressed by the prototype should achieve a balance between being too complex and too easy. Organizations new to Data Science often try to show its value with highly complex projects. However, the greater the complexity, the greater the risk of failure. At the same time, if the problem is too easy to solve, senior leaders and others in the organization may not see the need for Data Science. Look for efforts that could benefit from large datasets, or bringing together disparate datasets that have never been combined before, as opposed to those that require complex analytic approaches. In these cases, there is often low-hanging fruit that can lead to significant value for the organization.



# TAKE OFF *the* **TRAINING WHEELS**

## THE PRACTITIONER'S GUIDE TO DATA SCIENCE

Read this section to get beyond the hype and  
learn the secrets of being a Data Scientist.



# Guiding Principles

---

Failing is good; failing quickly is even better.

---

The set of guiding principles that govern how we conduct the tradecraft of Data Science are based loosely on the central tenets of innovation, as the two areas are highly connected. These principles are not hard and fast rules to strictly follow, but rather key tenets that have emerged in our collective consciousness. You should use these to guide your decisions, from problem decomposition through implementation.



## » Tips From the Pros

---

It can be easier to rule out a solution than confirm its correctness. As a result, focus on exploring obvious shortcomings that can quickly disqualify an approach. This will allow you to focus your time on exploring truly viable approaches as opposed to dead ends.

- › ***Be willing to fail.*** At the core of Data Science is the idea of experimentation. Truly innovative solutions only emerge when you experiment with new ideas and applications. Failure is an acceptable byproduct of experimentation. Failures locate regions that no longer need to be considered as you search for a solution.
- › ***Fail often and learn quickly.*** In addition to a willingness to fail, be ready to fail repeatedly. There are times when a dozen approaches must be explored in order to find the one that works. While you shouldn't be concerned with failing, you should strive to learn from the attempt quickly. The only way you can explore a large number of solutions is to do so quickly.
- › ***Keep the goal in mind.*** You can often get lost in the details and challenges of an implementation. When this happens, you lose sight of your goal and begin to drift off the path from data to analytic action. Periodically step back, contemplate your goal, and evaluate whether your current approach can really lead you where you want to go.



## » Tips From the Pros

---

If the first thing you try to do is to create the ultimate solution, you will fail, but only after banging your head against a wall for several weeks.

- › ***Dedication and focus lead to success.*** You must often explore many approaches before finding the one that works. It's easy to become discouraged. You must remain dedicated to your analytic goal. Focus on the details and the insights revealed by the data. Sometimes seemingly small observations lead to big successes.
- › ***Complicated does not equal better.*** As technical practitioners, we have a tendency to explore highly complex, advanced approaches. While there are times where this is necessary, a simpler approach can often provide the same insight. Simpler means easier and faster to prototype, implement and verify.

# The Importance of Reason

---

Beware: in the world of Data Science, if it walks like a duck and quacks like a duck, it might just be a moose.

---

Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning. Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths. Models of reality no longer need to be static. They are constantly tested, updated and improved until better models are found.

The analysis of big data has brought inductive reasoning to the forefront. Massive amounts of data are analyzed to identify correlations. However, a common pitfall to this approach is confusing correlation with causation. Correlation implies but does not prove causation. Conclusions cannot be drawn from correlations until the underlying mechanisms that relate the data elements are understood. Without a suitable model relating the data, a correlation may simply be a coincidence.



## » Correlation without Causation

---

A common example of this phenomenon is the high correlation between ice cream consumption and the murder rate during the summer months. Does this mean ice cream consumption causes murder or, conversely, murder causes ice cream consumption? Most likely not, but you can see the danger in mistaking correlation for causation. Our job as Data Scientists is making sure we understand the difference.

## »» The Dangers of Rejection



Paul Yacci

In the era of big data, one piece of analysis that is frequently overlooked is the problem of finding patterns when there are actually no apparent patterns. In statistics this is referred to as Type I error. As scientists, we are always on the lookout for a new or interesting breakthrough that could explain a phenomenon. We hope to see a pattern in our data that explains something or that can give us an answer. The primary goal of hypothesis testing is to limit Type I error. This is accomplished by using small  $\alpha$  values. For example, a  $\alpha$  value of 0.05 states that there is a 1 in 20 chance that the test will show that there is something significant when in actuality there isn't. This problem compounds when testing multiple hypotheses. When running multiple hypothesis tests, we are likely to encounter Type I error. As more data becomes available for analysis, Type I error needs to be controlled.

One of my projects required testing the difference between the means of two microarray data samples. Microarray data contains thousands of measurements but is limited in the number of observations. A common analysis approach is to measure the same genes under different conditions. If there is a significant enough difference in the amount of gene expression between the two samples, we can say that the gene is correlated with a particular phenotype. One way to do this is to take the mean of each phenotype for a particular

gene and formulate a hypothesis to test whether there is a significant difference between the means. Given that we were running thousands of these tests at  $\alpha = 0.05$ , we found several differences that were significant. The problem was that some of these could be caused by random chance.

Many corrections exist to control for false indications of significance. The Bonferroni correction is one of the most conservative. This calculation lowers the level below which you will reject the null hypothesis (your  $p$  value). The formula is  $\alpha/n$ , where  $n$  equals the number of hypothesis tests that you are running. Thus, if you were to run 1,000 tests of significance at  $\alpha = 0.05$ , your  $p$  value should be less than 0.00005 ( $0.05/1,000$ ) to reject the null hypothesis. This is obviously a much more stringent value. A large number of the previously significant values were no longer significant, revealing the true relationships within the data.

The corrected significance gave us confidence that the observed expression levels were due to differences in the cellular gene expression rather than noise. We were able to use this information to begin investigating what proteins and pathways were active in the genes expressing the phenotype of interest. By solidifying our understanding of the causal relationships, we focused our research on the areas that could lead to new discoveries about gene function and, ultimately to improved medical treatments.

---

Reason and common sense are foundational to Data Science. Without these, data is simply a collection of bits. Context, inferences and models are created by humans and carry with them biases and assumptions. Blindly trusting your analyses is a dangerous thing that can lead to erroneous conclusions. When you approach an analytic challenge, you should always pause to ask yourself the following questions:

---

- › **What problem are we trying to solve?** Articulate the answer as a sentence, especially when communicating with the end-user. Make sure that it sounds like an answer. For example, “Given a fixed amount of human capital, deploying people with these priorities will generate the best return on their time.”
- › **Does the approach make sense?** Write out your analytic plan. Embrace the discipline of writing, as it brings structure to your thinking. Back of the envelope calculations are an existence proof of your approach. Without this kind of preparation, computers are power tools that can produce lots of bad answers really fast.
- › **Does the answer make sense?** Can you explain the answer? Computers, unlike children, do what they are told. Make sure you spoke to it clearly by validating that the instructions you provided are the ones you intended. Document your assumptions and make sure they have not introduced bias in your work.
- › **Is it a finding or a mistake?** Be skeptical of surprise findings. Experience says that if it seems wrong, it probably is wrong. Before you accept that conclusion, however, make sure you understand and can clearly explain why it is wrong.
- › **Does the analysis address the original intent?** Make sure that you are not aligning the answer with the expectations of the client. Always speak the truth, but remember that answers of “your baby is ugly” require more, not less, analysis.
- › **Is the story complete?** The goal of your analysis is to tell an actionable story. You cannot rely on the audience to stitch the pieces together. Identify potential holes in your story and fill them to avoid surprises. Grammar, spelling and graphics matter; your audience will lose confidence in your analysis if your results look sloppy.
- › **Where would we head next?** No analysis is ever finished, you just run out of resources. Understand and explain what additional measures could be taken if more resources are found.



---

» **Tips From the Pros**

Better a short pencil than a long memory. End every day by documenting where you are; you may learn something along the way. Document what you learned and why you changed your plan.



---

» **Tips From the Pros**

Test your answers with a friendly audience to make sure your findings hold water.

# Component Parts of Data Science

---

There is a web of components that interact to create your solution space. Understanding how they are connected is critical to your ability to engineer solutions to Data Science problems.

---

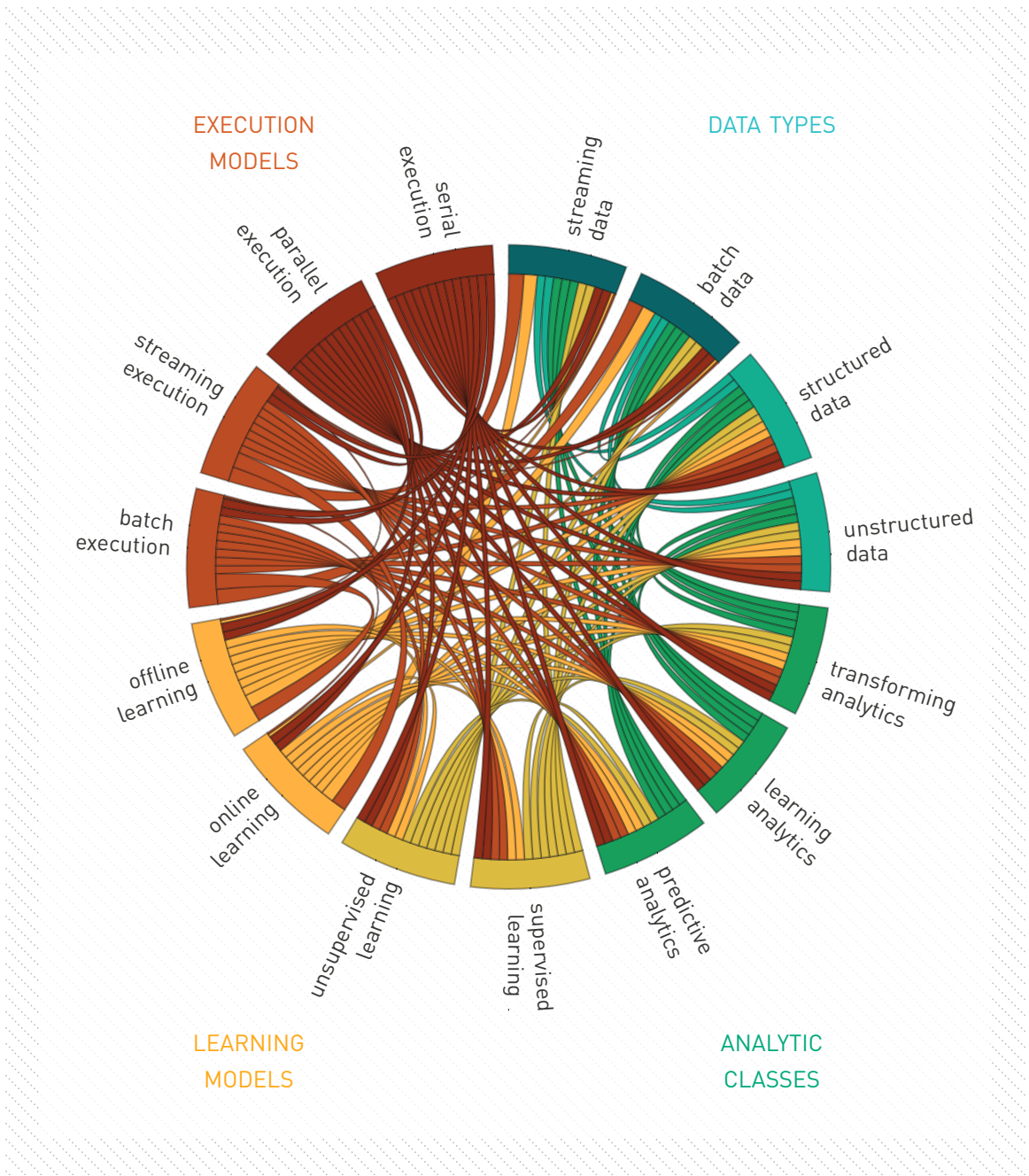
The components involved in any Data Science project fall into a number of different categories including the data types analyzed, the analytic classes used, the learning models employed and the execution models used to run the analytics. The interconnection across these components, shown in the figure, *Interconnection Among the Component Parts of Data Science*, speaks to the complexity of engineering Data Science solutions. A choice made for one component exerts influence over choices made for others categories. For example, data types lead the choices in analytic class and learning models, while latency, timeliness and algorithmic parallelization strategy inform the execution model. As we dive deeper into the technical aspects of Data Science, we will begin with an exploration of these components and touch on examples of each.

---

## *Read this to get the quick and dirty:*

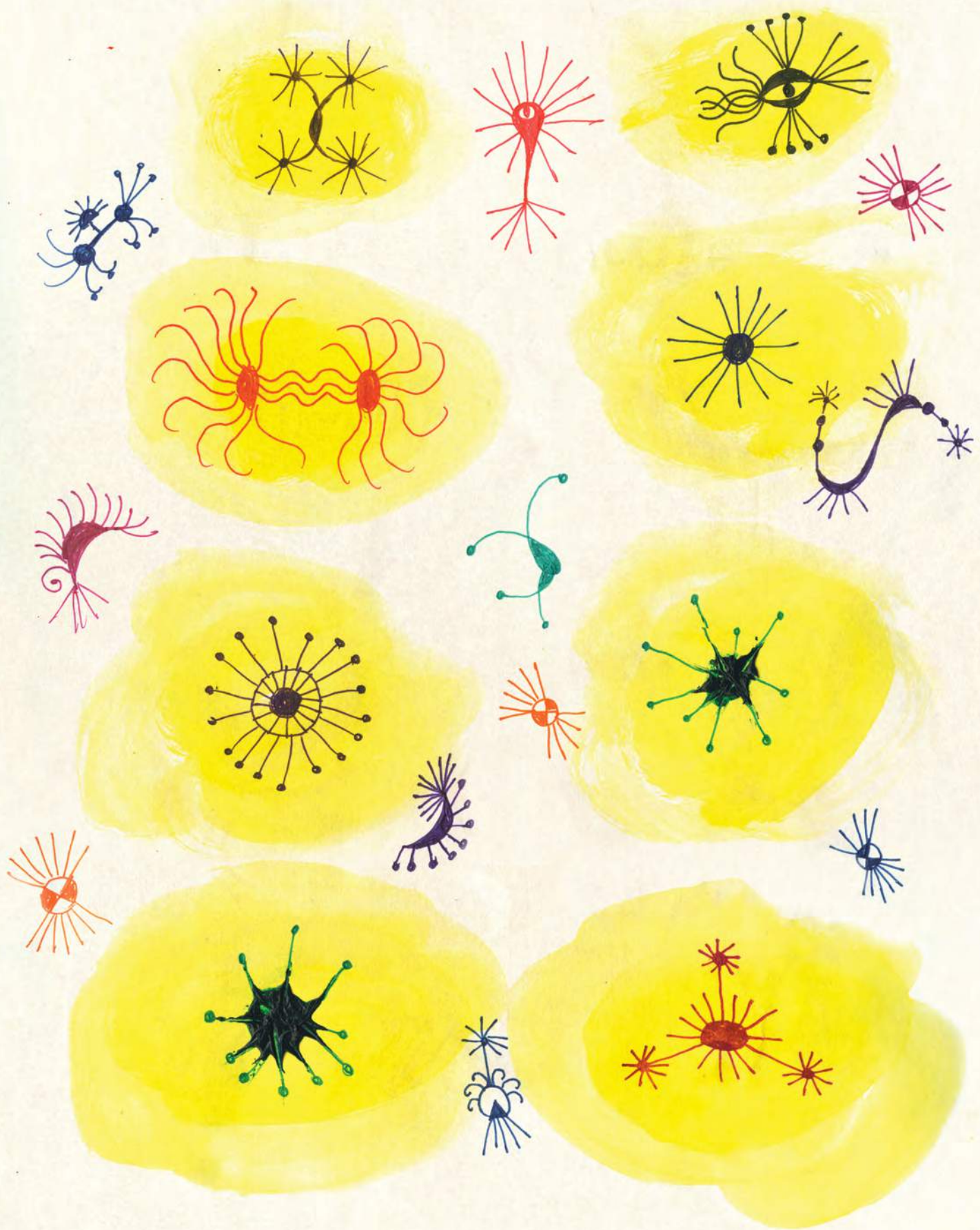
When engineering a Data Science solution, work from an understanding of the components that define the solution space. Regardless of your analytic goal, you must consider the *data types* with which you will be working, the *classes of analytics* you will use to generate your data product,

how the *learning models* embodied will operate and evolve, and the *execution models* that will govern how the analytic will be run. You will be able to articulate a complete Data Science solution only after considering each of these aspects.



Source: Booz Allen Hamilton

Interconnection Among the Component Parts of Data Science



## Data Types

Data types and analytic goals go hand-in-hand much like the chicken and the egg; it is not always clear which comes first. Analytic goals are derived from business objectives, but the data type also influences the goals. For example, the business objective of understanding consumer product perception drives the analytic goal of sentiment analysis. Similarly, the goal of sentiment analysis drives the selection of a text-like data type such as social media content. Data type also drives many other choices when engineering your solutions.

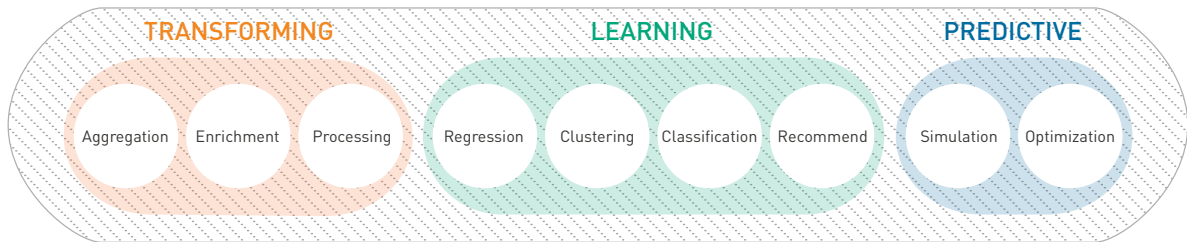
There are a number of ways to classify data. It is common to characterize data as *structured* or *unstructured*. Structured data exists when information is clearly broken out into fields that have an explicit meaning and are highly categorical, ordinal or numeric. A related category, semi-structured, is sometimes used to describe structured data that does not conform to the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers. Unstructured data, such as natural language text, has less clearly delineated meaning. Still images, video and audio often fall under the category of unstructured data. Data in this form requires preprocessing to identify and extract relevant 'features.' The features are structured information that are used for indexing and retrieval, or training classification, or clustering models.

Data may also be classified by the rate at which it is generated, collected or processed. The distinction is drawn between streaming data that arrives constantly like a torrent of water from a fire hose, and batch data, which arrives in buckets. While there is rarely a connection between data type and data rate, data rate has significant influence over the execution model chosen for analytic implementation and may also inform a decision of analytic class or learning model.



## Classes of Analytic Techniques

As a means for helping conceptualize the universe of possible analytic techniques, we grouped them into nine basic classes. Note that techniques from a given class may be applied in multiple ways to achieve various analytic goals. Membership in a class simply indicates a similar analytic function. The nine analytic classes are shown in the figure, *Classes of Analytic Techniques*.



Source: Booz Allen Hamilton

Classes of Analytic Techniques

### » Transforming Analytics

- › **Aggregation:** Techniques to summarize the data. These include basic statistics (e.g., mean, standard deviation), distribution fitting, and graphical plotting.
- › **Enrichment:** Techniques for adding additional information to the data, such as source information or other labels.
- › **Processing:** Techniques that address data cleaning, preparation, and separation. This group also includes common algorithm pre-processing activities such as transformations and feature extraction.

### » Learning Analytics

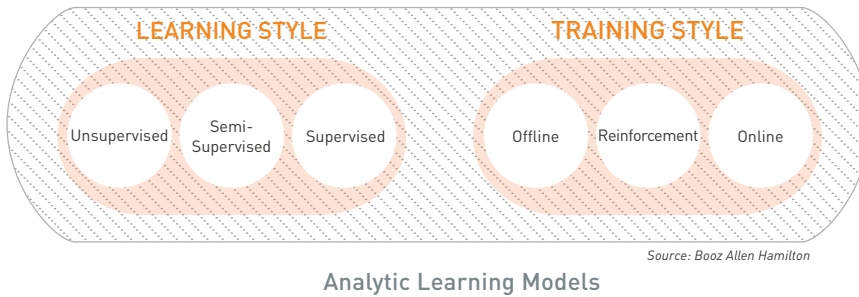
- › **Regression:** Techniques for estimating relationships among variables, including understanding which variables are important in predicting future values.
- › **Clustering:** Techniques to segment the data into naturally similar groups.
- › **Classification:** Techniques to identify data element group membership.
- › **Recommendation:** Techniques to predict the rating or preference for a new entity, based on historic preference or behavior.

### » Predictive Analytics

- › **Simulation:** Techniques to imitate the operation of a real-world process or system. These are useful for predicting behavior under new conditions.
- › **Optimization:** Operations Research techniques focused on selecting the best element from a set of available alternatives to maximize a utility function.

# Learning Models

Analytic classes that perform predictions, such as regression, clustering, classification and recommendation employ learning models. These models characterize how the analytic is trained to perform judgments on new data based on historic observation. Aspects of learning models describe both the types of judgments performed and how the models evolve over time, as shown in the figure, *Analytic Learning Models*.



Learning models are typically described as belonging to the categories of unsupervised or supervised learning. Supervised learning takes place when a model is trained using a labeled dataset that has a known class or category associated with each data element. The model relates features found in training instances with labels so that predictions can be made for unlabeled instances. Unsupervised learning involves no a-priori knowledge about the classes into which data can be placed. Unsupervised learning uses the features in the dataset to form groupings based on feature similarity. Semi-supervised learning is a hybrid between these two approaches, using a small amount of labeled data in conjunction with a large amount of unlabeled data. This is done to improve learning accuracy in cases where only a small number of labeled observations are available for learning.

There are a variety of ways to train learning models. A useful distinction is between those that are trained in a single pass, which are known as offline models, and those that are trained incrementally over time, known as online models. Many learning approaches have online or offline variants. The decision to use one or another is based on the analytic goals and execution models chosen.

Generating an offline model requires taking a pass over the entire training dataset. Improving the model requires making separate passes over the data. These models are static in that once trained, their predictions will not change until a new model is created through a subsequent training stage. Offline model performance is easier to evaluate due to this deterministic behavior. Deployment of the model into a production environment involves swapping out the old model for the new.

Online models dynamically evolve over time, meaning they only require a single deployment into a production setting. The fact that



## » Reinforcement Learning in Action

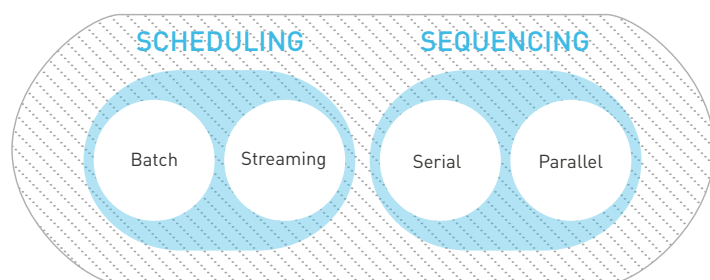
The possibilities of Reinforcement Learning captured significant attention with the publication of a study in the journal *Nature* in which a computer agent learned to play 49 different video games with accuracy rivaling a professional game tester<sup>[14]</sup>. The agent was able to achieve these results using only the raw screen pixels and game score as input. This approach represents the first artificial intelligence agent that is capable of learning complex tasks while bridging between high-dimensional sensory inputs and actions.

these models do not have the entire dataset available when being trained is a challenge. They must make assumptions about the data based on the examples observed; these assumptions may be sub-optimal. The impact of sub-optimal predictions can be mitigated in cases where feedback on the model's predictions is available. Online models can rapidly incorporate feedback to improve performance.

One such training style is known as Reinforcement Learning. Under this approach, an algorithm takes action in an environment and incrementally learns how to achieve goals based on the response to a function used to determine the quality of its results. Reinforcement learning is generally applicable to complex, real-world tasks that involve optimization, such as navigation or trading. Due to the publication of many promising results from Reinforcement Learning algorithms, the popularity of this technique has risen dramatically in recent years along with Deep Learning.

## Execution Models

Execution models describe how data is manipulated to perform an analytic function. They may be categorized across a number of dimensions. Execution Models are embodied by an execution framework, which orchestrates the sequencing of analytic computation. In this sense, a framework might be as simple as a programming language runtime, such as the Python interpreter, or a distributed computing framework that provides a specific API for one or more programming languages such as Hadoop, MapReduce or Spark. Grouping execution models based on how they handle data is common, classifying them as either batch or streaming execution models. The categories of execution model are shown in the figure, *Analytic Execution Models*.



Source: Booz Allen Hamilton

Analytic Execution Models

A batch execution model implies that data is analyzed in large segments, that the analytic has a state where it is running and a state where it is not running and that little state is maintained in memory between executions. Batch execution may also imply that the analytic produces results with a frequency on the order of several minutes or more. Batch workloads tend to be fairly easy to conceptualize because

they represent discrete units of work. As such, it is easy to identify a specific series of execution steps as well as the proper execution frequency and time bounds based on the rate at which data arrives. Depending on the algorithm choice, batch execution models are easily scalable through parallelism. There are a number of frameworks that support parallel batch analytic execution. Most famously, Hadoop provides a distributed batch execution model in its MapReduce framework.

Conversely, a streaming model analyzes data as it arrives. Streaming execution models imply that under normal operation, the analytic is always executing. The analytic can hold state in memory and constantly deliver results as new data arrives, on the order of seconds or less. Many of the concepts in streaming are inherent in the Unix-pipeline design philosophy; processes are chained together by linking the output of one process to the input of the next. As a result, many developers are already familiar with the basic concepts of streaming. A number of frameworks are available that support the parallel execution of streaming analytics such as Storm, S4 and Samza.

The choice between batch and streaming execution models often hinges on analytic latency and timeliness requirements. Latency refers to the amount of time required to analyze a piece of data once it arrives at the system, while timeliness refers to the average age of an answer or result generated by the analytic system. For many analytic goals, a latency of hours and timeliness of days is acceptable and thus lend themselves to the implementation enabled by the batch approach. Some analytic goals have up-to-the-second requirements where a result that is minutes old has little worth. The streaming execution model better supports such goals.

Batch and streaming execution models are not the only dimensions within which to categorize analytic execution methods. Another distinction is drawn when thinking about scalability. In many cases, scale can be achieved by spreading computation over a number of computers. In this context, certain algorithms require a large shared memory state, while others are easily parallelizable in a context where no shared state exists between machines. This distinction has significant impacts on both software and hardware selection when building out a parallel analytic execution environment.



### » Tips From the Pros

---

In order to understand system capacity in the context of streaming analytic execution, collect metrics including: the amount of data consumed, data emitted, and latency. This will help you understand when scale limits are reached.

# Fractal Analytic Model

---

Data Science analytics are a lot like broccoli.

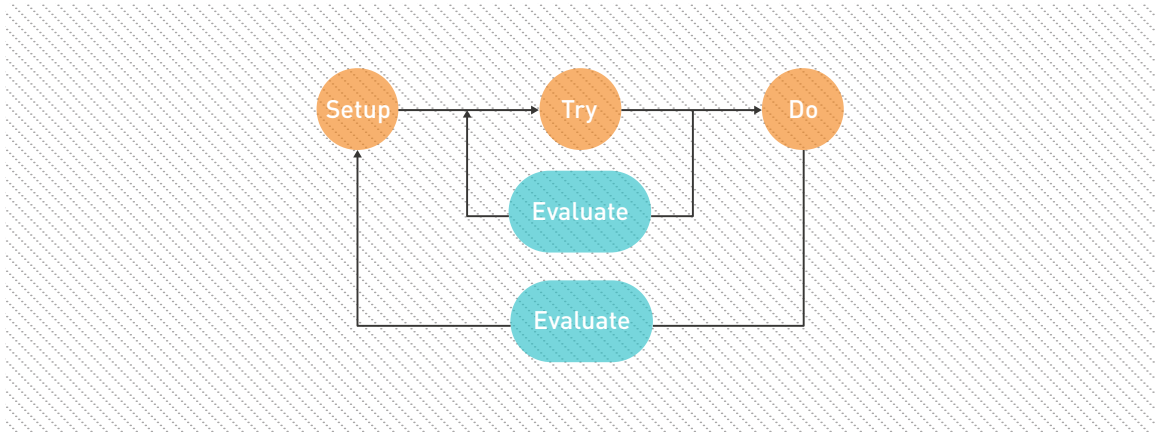
---

Fractals are mathematical sets that display self-similar patterns. As you zoom in on a fractal, the same patterns reappear. Imagine a stalk of broccoli. Rip off a piece of broccoli and the piece looks much like the original stalk. Progressively smaller pieces of broccoli still look like the original stalk.

Data Science analytics are a lot like broccoli – fractal in nature in both time and construction. Early versions of an analytic follow the same development process as later versions. At any given iteration, the analytic itself is a collection of smaller analytics that often decompose into yet smaller analytics.

## Iterative by Nature

Good Data Science is fractal in time — an iterative process. Getting an imperfect solution out the door quickly will gain more interest from stakeholders than a perfect solution that is never completed. The figure, *The Data Science Product Lifecycle*, summarizes the lifecycle of the Data Science product.



Source: Booz Allen Hamilton

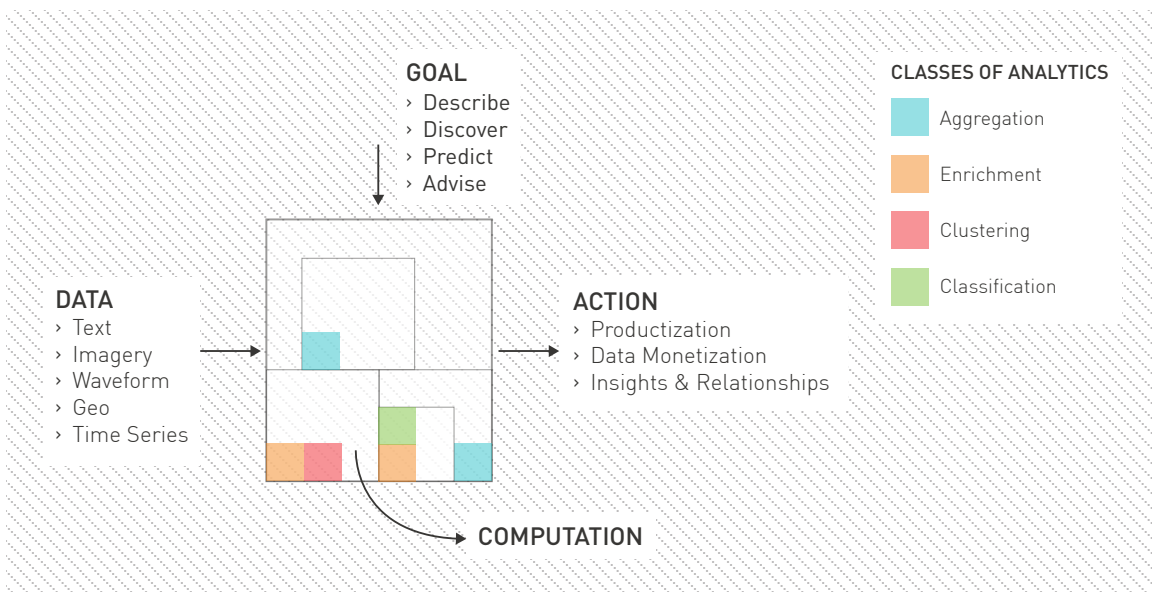
The Data Science Product Lifecycle

*Set up* the infrastructure, aggregate and prepare the data, and incorporate domain expert knowledge. *Try* different analytic techniques and models on subsets of the data. *Evaluate* the models, refine, evaluate again, and select a model. *Do* something with your models and results – deploy the models to inform, inspire action, and act. *Evaluate* the business results to improve the overall product.

## Smaller Pieces of Broccoli: A Data Science Product

Components inside and outside of the Data Science product will change with each iteration. Let's take a look under the hood of a Data Science product and examine the components during one such iteration.

In order to achieve a greater analytic goal, you need to first decompose the problem into sub-components to divide and conquer. The figure, *The Fractal Analytic Model*, shows a decomposition of the Data Science product into four component pieces.



Source: Booz Allen Hamilton

The Fractal Analytic Model

## GOAL

---

You must first have some idea of your analytic goal and the end state of the analysis. Is it to Discover, Describe, Predict, or Advise? It is probably a combination of several of those. Be sure that before you start, you define the business value of the data and how you plan to use the insights to drive decisions, or risk ending up with interesting but non-actionable trivia.

## DATA

---

Data dictates the potential insights that analytics can provide. Data Science is about finding patterns in variable data and comparing those patterns. If the data is not representative of the universe of events you wish to analyze, you will want to collect that data through carefully planned variations in events or processes through A/B testing or design of experiments. Datasets are never perfect so don't wait for perfect data to get started. A good Data Scientist is adept at handling messy data with missing or erroneous values. Just make sure to spend the time upfront to clean the data or risk generating garbage results.

## COMPUTATION

---

Computation aligns the data to goals through the process of creating insights. Through divide and conquer, computation decomposes into several smaller analytic capabilities with their own goals, data, computation and resulting actions, just like a smaller piece of broccoli maintains the structure of the original stalk. In this way, computation itself is fractal. Capability building blocks may utilize different types of execution models such as batch computation or streaming, that individually accomplish small tasks. When properly combined together, the small tasks produce complex, actionable results.

## ACTION

---

How should engineers change the manufacturing process to generate higher product yield? How should an insurance company choose which policies to offer to whom and at what price? The output of computation should enable actions that align to the goals of the data product. Results that do not support or inspire action are nothing but interesting trivia.

Given the fractal nature of Data Science analytics in time and construction, there are many opportunities to choose fantastic or shoddy analytic building blocks. *The Analytic Selection Process* offers some guidance.



# The Analytic Selection Process

---

If you focus only on the science aspect of Data Science you will never become a data artist.

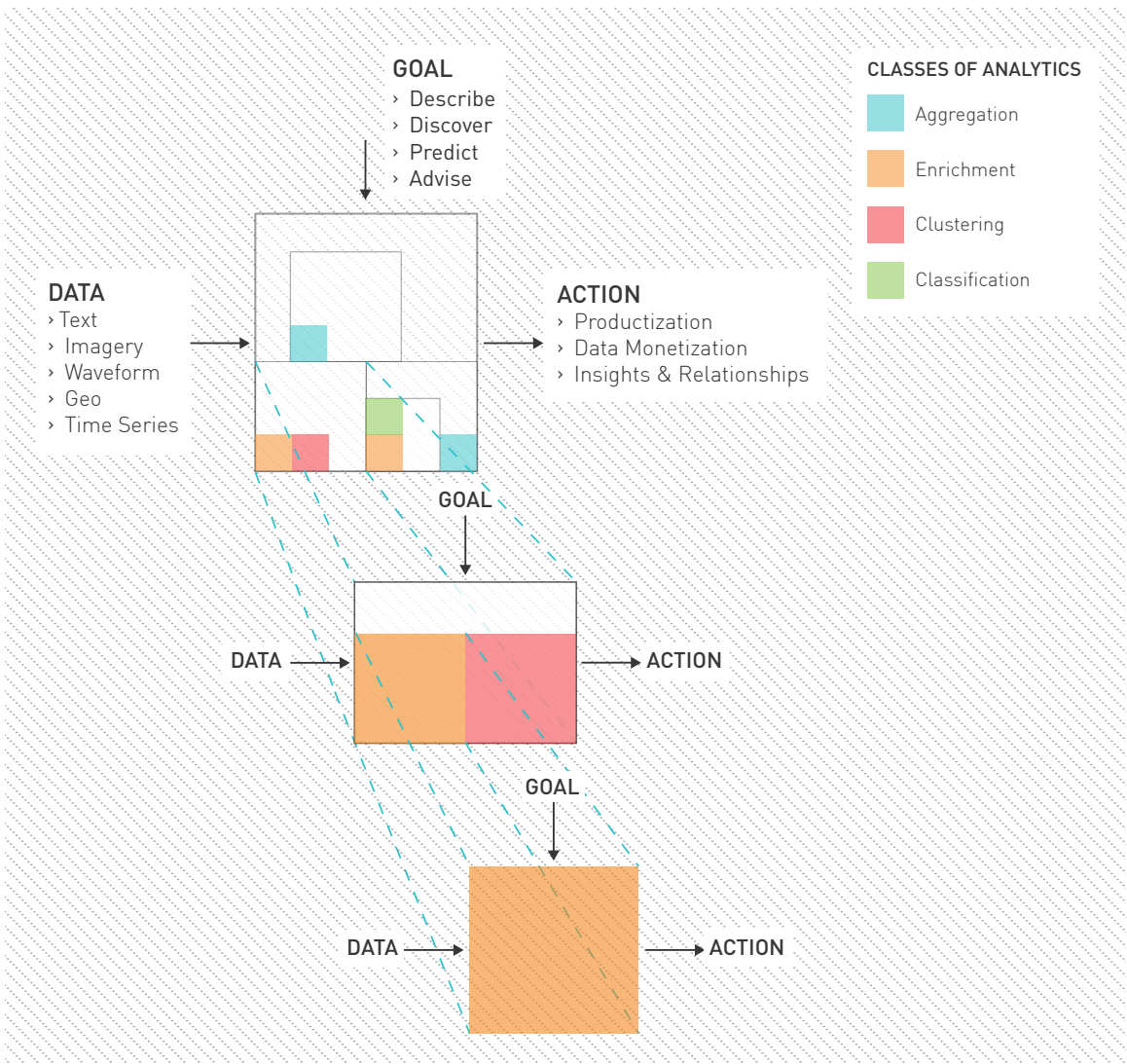
---

A critical step in Data Science is to identify an analytic technique that will produce the desired action. Sometimes it is clear; a characteristic of the problem (e.g., data type) points to the technique you should implement. Other times, however, it can be difficult to know where to begin. The universe of possible analytic techniques is large. Finding your way through this universe is an art that must be practiced. We are going to guide you on the next portion of your journey - becoming a data artist.

# Decomposing the Problem

Decomposing the problem into manageable pieces is the first step in the analytic selection process. Achieving a desired analytic action often requires combining multiple analytic techniques into a holistic, end-to-end solution. Engineering the complete solution requires that the problem be decomposed into progressively smaller sub-problems.

The *Fractal Analytic Model* embodies this approach. At any given stage, the analytic itself is a collection of smaller computations that decompose into yet smaller computations. When the problem is decomposed far enough, only a single analytic technique is needed to achieve the analytic goal. Problem decomposition creates multiple sub-problems, each with their own goals, data, computations, and actions. The concept behind problem decomposition is shown in the figure, *Problem Decomposition Using the Fractal Analytic Model*.



Source: Booz Allen Hamilton

Problem Decomposition Using the Fractal Analytic Model

On the surface, problem decomposition appears to be a mechanical, repeatable process. While this may be true conceptually, it is really the performance of an art as opposed to the solving of an engineering problem. There may be many valid ways to decompose the problem, each leading to a different solution. There may be hidden dependencies or constraints that only emerge after you begin developing a solution. This is where art meets science. Although the art behind problem decomposition cannot be taught, we have distilled some helpful hints to help guide you. When you begin to think about decomposing your problem, look for:



### » Tips From the Pros

One of your first steps should be to explore available data sources that have not been previously combined. Emerging relationships between data sources often allow you to pick low hanging fruit.

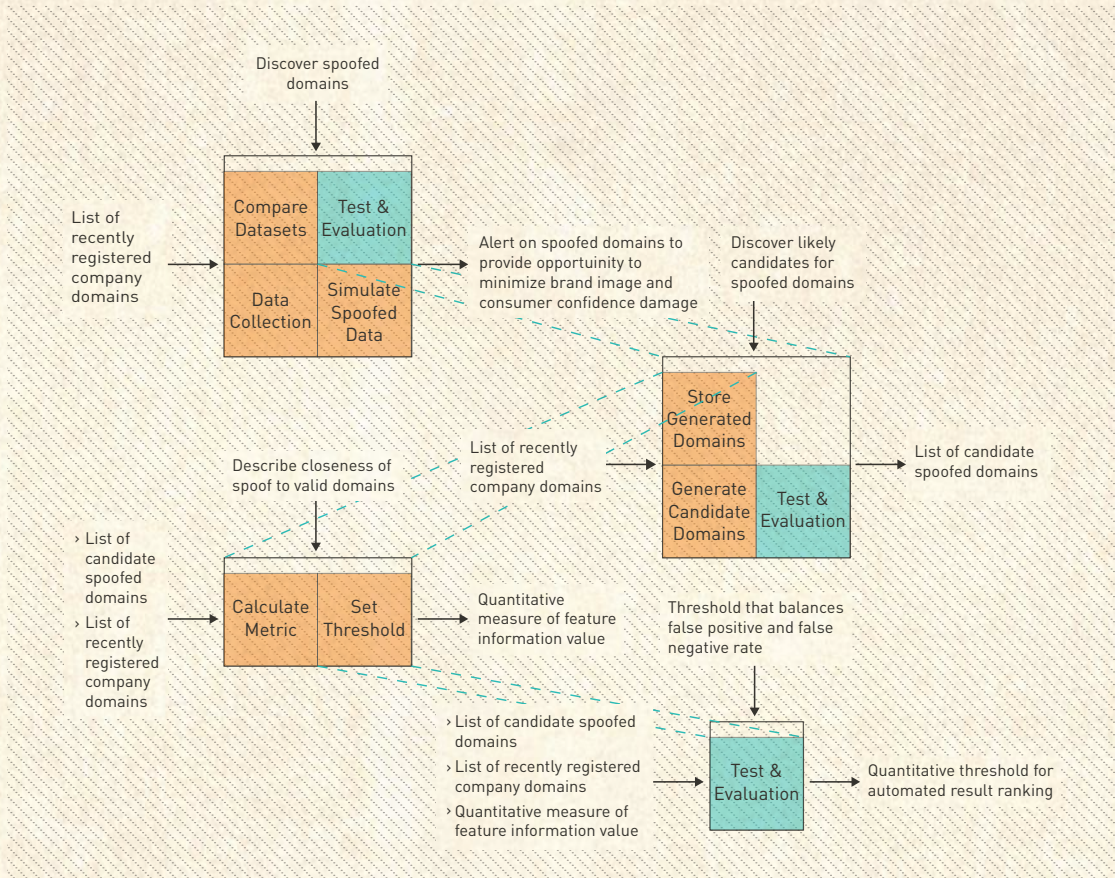
- › ***Compound analytic goals that create natural segmentation.*** For example, many problems focused on predicting future conditions include both Discover and Predict goals.
- › ***Natural orderings of analytic goals.*** For example, when extracting features you must first identify candidate features and then select the features set with the highest information value. These two activities form distinct analytic goals.
- › ***Data types that dictate processing activities.*** For example, text or imagery both require feature extraction.
- › ***Requirements for human-in-the-loop feedback.*** For example, when developing alerting thresholds, you might need to solicit analyst feedback and update the threshold based on their assessment.
- › ***The need to combine multiple data sources.*** For example, you may need to correlate two datasets to achieve your broader goal. Often this indicates the presence of a Discover goal.

In addition to problem decomposition providing a tractable approach to analytic selection, it has the added benefit of simplifying a highly complex problem. Rather than being faced with understanding the entire end-to-end solution, the computations are discrete segments that can be explored. Note, however, that while this technique helps the Data Scientist approach the problem, it is the complete end-to-end solution that must be evaluated.

# »» Identifying Spoofed Domains



Identifying spoofed domains is important for an organization to preserve their brand image and to avoid eroded customer confidence. Spoofed domains occur when a malicious actor creates a website, URL or email address that users believe is associated with a valid organization. When users click the link, visit the website or receive emails, they are subjected to some type of nefarious activity.



Source: Booz Allen Hamilton

## Spoofed Domain Problem Decomposition

Our team was faced with the problem of identifying spoofed domains for a commercial company. On the surface, the problem sounded easy; take a recently registered domain, check to see if it is similar to the company's domain and alert when the similarity is sufficiently high. Upon decomposing the problem, however, the main computation quickly became complicated.

We needed a computation that determined similarity between two domains. As we decomposed the similarity computation, complexity and speed became a concern. As with many security-related problems, fast

alert speeds are vital. Result speed created an implementation constraint that forced us to re-evaluate how we decomposed the problem.

Revisiting the decomposition process led us to a completely new approach. In the end, we derived a list of domains similar to those registered by the company. We then compared that list against a list of recently registered domains. The figure, *Spoofed Domain Problem Decomposition*, illustrates our approach. Upon testing and initial deployment, our analytic discovered a spoofed domain within 48 hours.



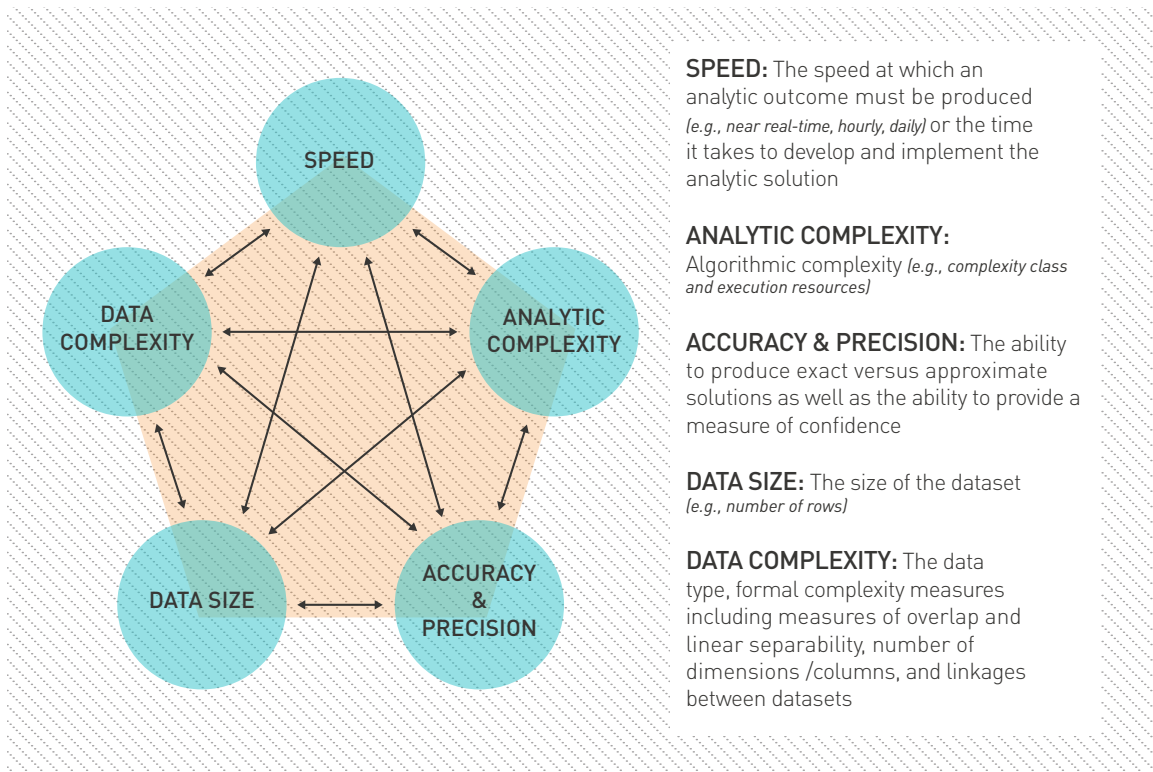
66666666 66666666



# Implementation Constraints

In the spoofed domains case study, the emergence of an implementation constraint caused the team to revisit its approach. This demonstrates that analytic selection does not simply mean choosing an analytic technique to achieve a desired outcome. It also means ensuring that the solution is feasible to implement.

The Data Scientist may encounter a wide variety of implementation constraints. They can be conceptualized, however, in the context of five dimensions that compete for your attention: analytic complexity, speed, accuracy & precision, data size, and data complexity. Balancing these dimensions is a zero sum game - an analytic solution cannot simultaneously exhibit all five dimensions, but instead must make trades between them. The figure, *Balancing the Five Analytic Dimensions*, illustrates this relationship.



Source: Booz Allen Hamilton

Balancing the Five Analytic Dimensions

Implementation constraints occur when an aspect of the problem dictates the value for one or more of these dimensions. As soon as one dimension is fixed, the Data Scientist is forced to make trades among the others. For example, if the analytic problem requires actions to be produced in near real-time, the speed dimension is fixed and trades must be made among the other four dimensions. Understanding which trades will achieve the right balance among the five dimensions is an art that must be learned over time.

As we compiled this section, we talked extensively about ways to group and classify implementation constraints. After much discussion we settled on these five dimensions. We present this model in hopes that others weigh in and offer their own perspectives.



### » Tips From the Pros

---

When possible, consider approaches that make use of previously computed results. Your algorithm will run much faster if you can avoid re-computing values across the full time horizon of data.



### » Tips From the Pros

---

Our *Data Science Product Lifecycle* has evolved to produce results quickly and then incrementally improve the solution.



### » Tips From the Pros

---

Streaming approaches may be useful for overcoming storage limitations.

## Some common examples of implementation constraints include:

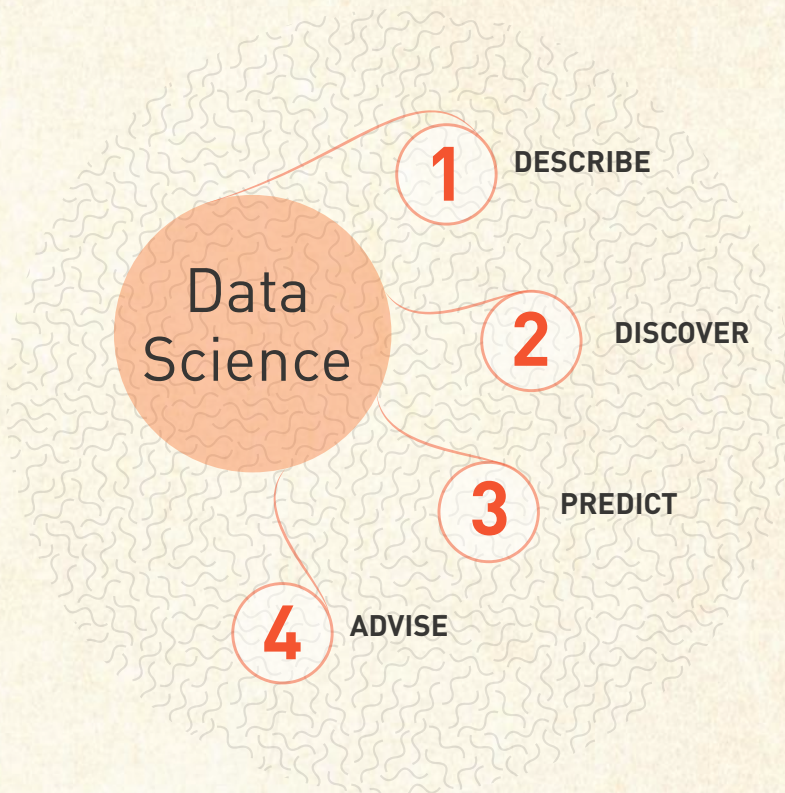
- **Computation frequency.** The solution may need to run on a regular basis (e.g., hourly), requiring that computations be completed within a specified window of time. The best analytic is useless if it cannot solve the problem within the required time.
- **Solution timeliness.** Some applications require near real-time results, pointing toward streaming approaches. While some algorithms can be implemented within streaming frameworks, many others cannot.
- **Implementation speed.** A project may require that you rapidly develop and implement a solution to quickly produce analytic insights. In these cases, you may need to focus on less complex techniques that can be quickly implemented and verified.
- **Computational resource limitations.** Although you may be able to store and analyze your data, data size may be sufficiently large that algorithms requiring multiple computations across the full dataset are too resource intensive. This may point toward needing approaches that only require a single pass on the data (e.g., canopy cluster as opposed to k-means clustering).
- **Data storage limitations.** There are times when big data becomes so big it cannot be stored or only a short time horizon can be stored. Analytic approaches that require long time horizons may not be possible.

Organizational policies and regulatory requirements are a major source of implicit constraints that merit a brief discussion. Policies are often established around specific classes of data such as Personally Identifiable Information (PII) or Personal Health Information (PHI). While the technologies available today can safely house information with a variety of security controls in a single system, these policies force special data handling considerations including limited retention periods and data access. Data restrictions impact the data size and complexity dimensions outlined earlier, creating yet another layer of constraints that must be considered.

# Guide to Analytic Selection

Your senses are incapable of perceiving the entire universe, so we drew you a map.

The universe of analytic techniques is vast and hard to comprehend. We created this diagram to aid you in finding your way from data and goal to analytic action. Begin at the center of the universe (Data Science) and answer questions about your analytic goals and problem characteristics. The answers to your questions will guide you through the diagram to the appropriate class of analytic techniques and provide recommendations for a few techniques to consider.



- ❶ **TIP:** There are several situations where dimensionality reduction may be needed:
  - › Models fail to converge feature space
  - › Models produce results equivalent to random chance › You do not know which aspects of the data are
  - › You lack the computational power to perform operations across the › the most important
- ❷ **Feature Extraction** is a broad topic and is highly dependent upon the domain area. This topic could be the subject of an entire book. As a result, a detailed exploration has been omitted from this diagram. See the *Featuring Engineering* and *Feature Selection* sections in the *Life in the Trenches* chapter for additional information.
- ❸ **TIP:** Always check data labels for correctness. This is particularly true for time stamps, which may have reverted to system default values.
- ❹ **TIP:** Smart enrichment can greatly speed-up computational time. It can also be a huge differentiator between the accuracy of different end-to-end analytic solutions.

Source: Booz Allen Hamilton





## FILTERING

How do I identify data based on its absolute or relative values?

If you want to add or remove data based on its value, start with:  
› Relational algebra projection and selection

If early results are uninformative and duplicative, start with:  
› Outlier removal › Gaussian filter  
› Exponential smoothing › Median filter

## IMPUTATION

How do I fill in missing values in my data?

If you want to generate values from other observations in your dataset, start with:  
› Random sampling  
› Markov Chain Monte Carlo (MC)

If you want to generate values without using other observations in your dataset, start with:  
› Mean › Regression models  
› Statistical distributions

## DIMENSIONALITY REDUCTION 1

How do I reduce the number of dimensions in my data?

If you need to determine whether there is multi-dimensional correlation, start with:  
› PCA and other factor analysis

If you can represent individual observations by membership in a group, start with:  
› K-means clustering  
› Canopy clustering

## NORMALIZATION & TRANSFORMATION

How do I reconcile duplication representations in the data?

If you have unstructured text data, start with:  
› Term Frequency/Inverse Document Frequency (TF IDF)

If you have a variable number of features but your algorithm requires a fixed number, start with:  
› Feature hashing

## FEATURE EXTRACTION 2

If you are not sure which features are the most important, start with:  
› Wrapper methods  
› Sensitivity analysis

If you need to facilitate understanding of the probability distribution of the space, start with:  
› Self organizing maps

## PROCESSING 3

How do I clean and separate my data?

If you suspect duplicate data elements, start with:  
› Deduplication

If you want your data to fall within a specified range, start with:  
› Normalization

If your data is stored in a binary format, start with:  
› Format Conversion

If you are operating in frequency space, start with:  
› Fast Fourier Transform (FFT),  
› Discrete wavelet transform

If you are operating in Euclidian space, start with:  
› Coordinate transform

1

## DESCRIBE

How do I develop an understanding of the content of my data?

If you are unfamiliar with the dataset, start with **basic statistics**:

› Count › Standard deviation › Box plots  
› Mean › Range › Scatter plots

2

## DISCOVER

3

## PREDICT

4

## ADVISE

Data Science

## AGGREGATION

How do I collect and summarize my data?

If your approach assumes the data follows a **distribution**, start with:  
› Distribution fitting

If you want to understand all the information available on an entity, start with:  
› "Baseball card" aggregation

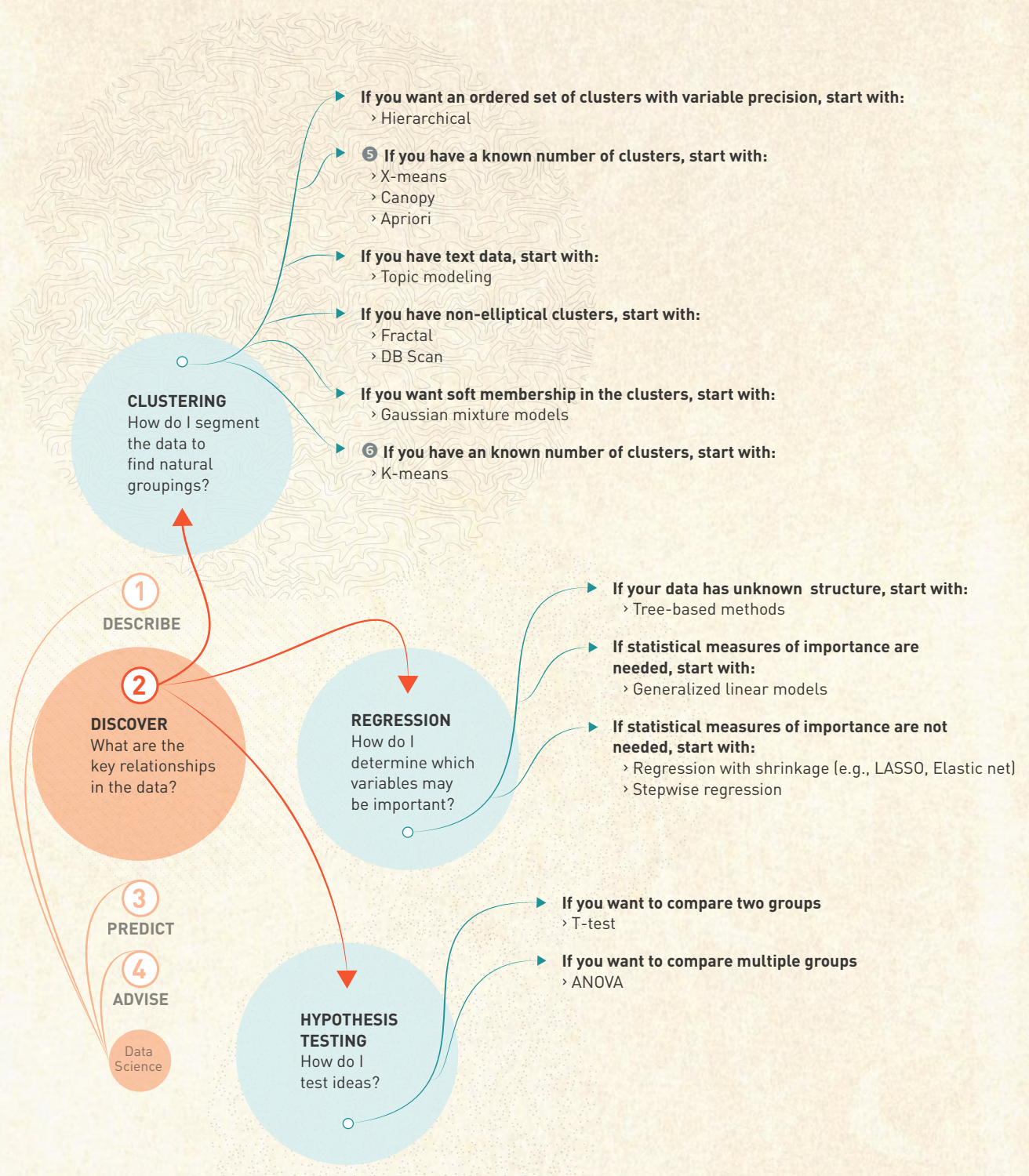
## ENRICHMENT 4

How do I add new information to my data?

If you need to keep track of source information or other **user-defined parameters**, start with:  
› Annotation

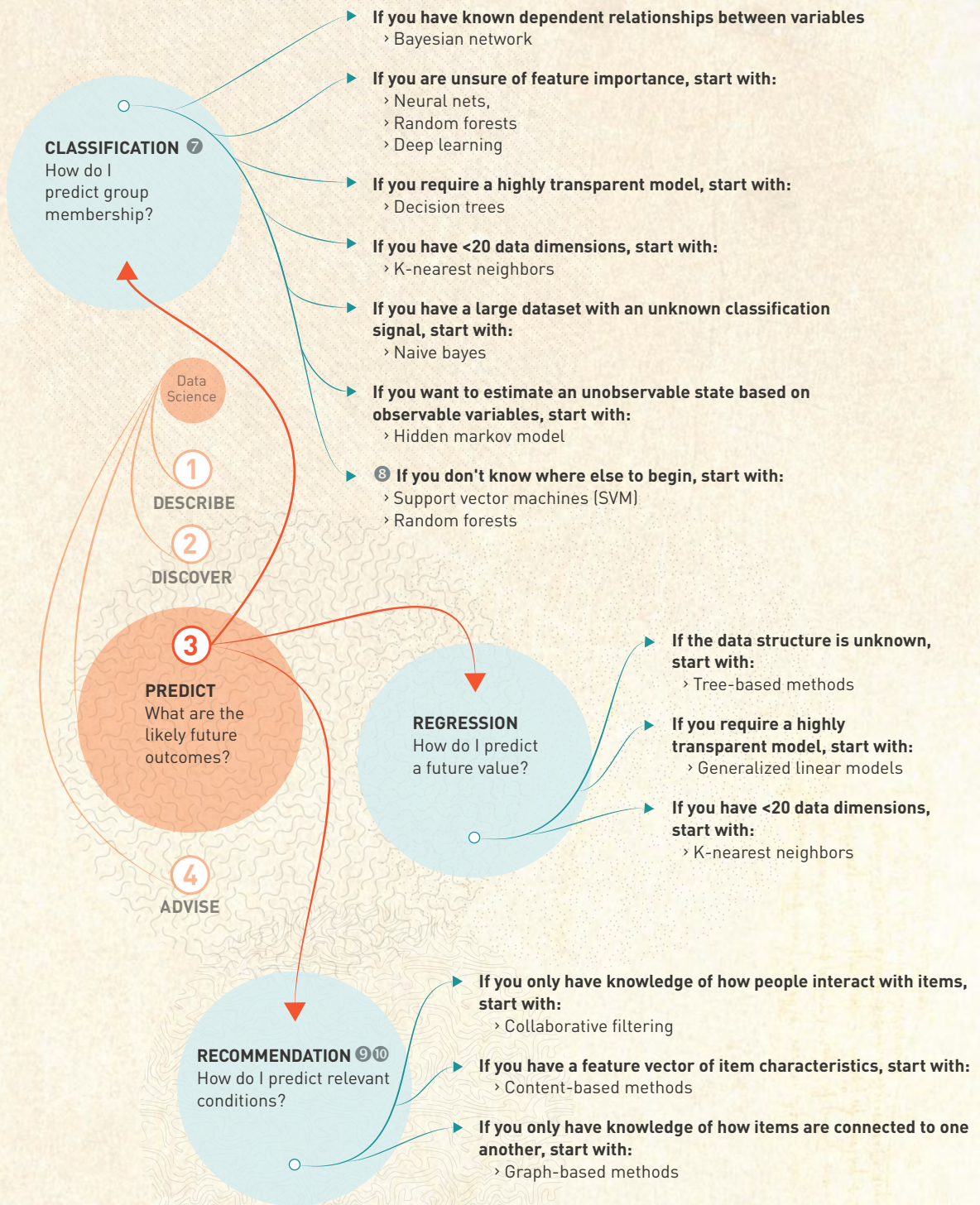
If you often process certain data fields together or use **one field to compute the value of another**, start with:  
› Relational algebra rename,  
› Feature addition (e.g., Geography, Technology, Weather)

Source: Booz Allen Hamilton



**5 TIP:** Canopy clustering is good when you only want to make a single pass over the data.

**6 TIP:** Use canopy or hierarchical clustering to estimate the number of clusters you should generate.

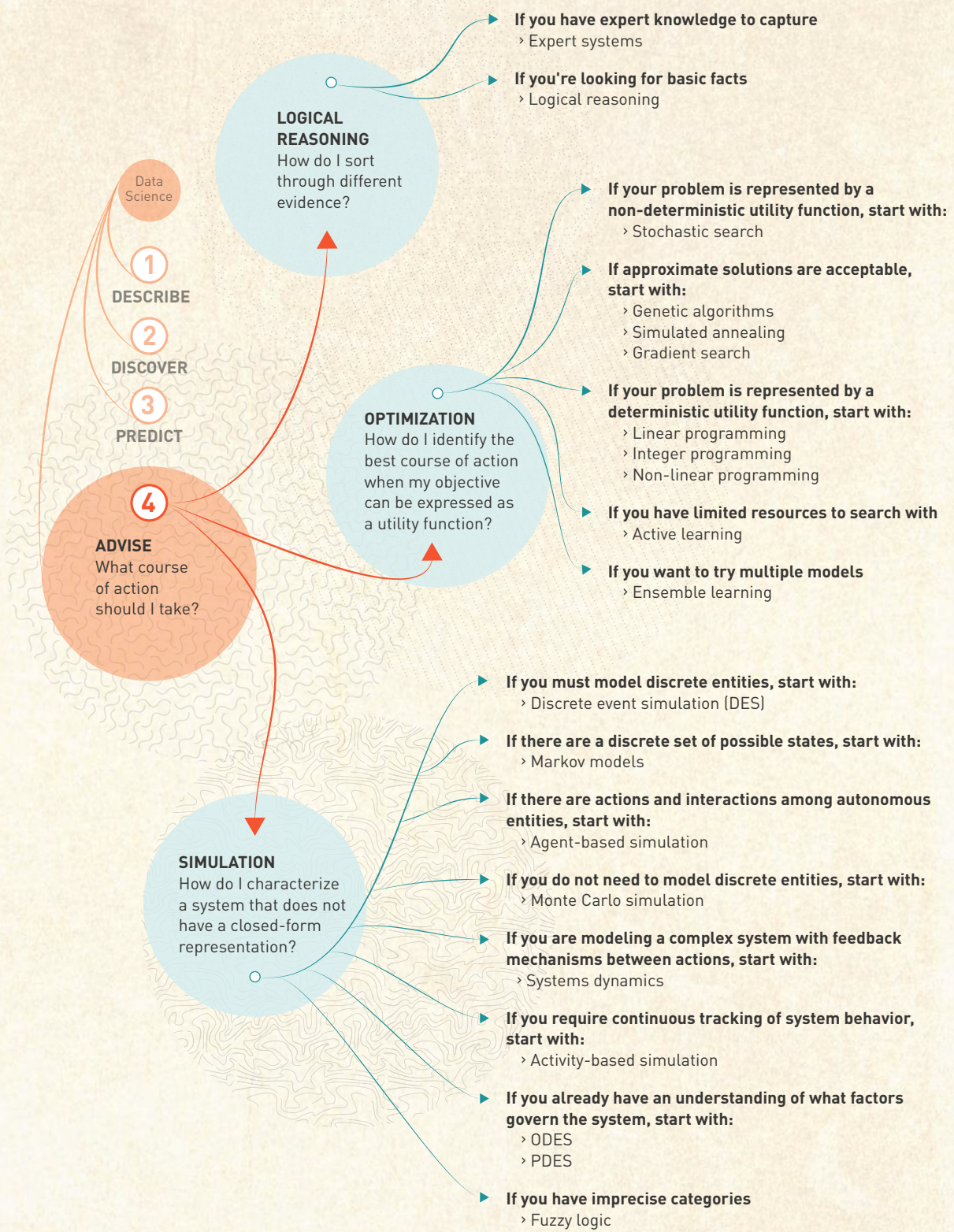


**7 TIP:** It can be difficult to predict which classifier will work best on your dataset. Always try multiple classifiers. Pick the one or two that work the best to refine and explore further.

**8 TIP:** These are our favorite, go-to classification algorithms.

**9 TIP:** Be careful of the “recommendation bubble”, the tendency of recommenders to recommend only what has been seen in the past. You must ensure you add diversity to avoid this phenomenon.

**10 TIP:** SVD and PCA are good tools for creating better features for recommenders.



Data Science

1

DESCRIBE

2

DISCOVER

3

PREDICT

4

ADVISE

What course of action should I take?

**LOGICAL REASONING**

How do I sort through different evidence?

**If you have expert knowledge to capture**

› Expert systems

**If you're looking for basic facts**

› Logical reasoning

**If your problem is represented by a non-deterministic utility function, start with:**

› Stochastic search

**If approximate solutions are acceptable, start with:**

- › Genetic algorithms
- › Simulated annealing
- › Gradient search

**If your problem is represented by a deterministic utility function, start with:**

- › Linear programming
- › Integer programming
- › Non-linear programming

**If you have limited resources to search with**

› Active learning

**If you want to try multiple models**

› Ensemble learning

**OPTIMIZATION**

How do I identify the best course of action when my objective can be expressed as a utility function?

**SIMULATION**

How do I characterize a system that does not have a closed-form representation?

**If you must model discrete entities, start with:**

› Discrete event simulation (DES)

**If there are a discrete set of possible states, start with:**

› Markov models

**If there are actions and interactions among autonomous entities, start with:**

› Agent-based simulation

**If you do not need to model discrete entities, start with:**

› Monte Carlo simulation

**If you are modeling a complex system with feedback mechanisms between actions, start with:**

› Systems dynamics

**If you require continuous tracking of system behavior, start with:**

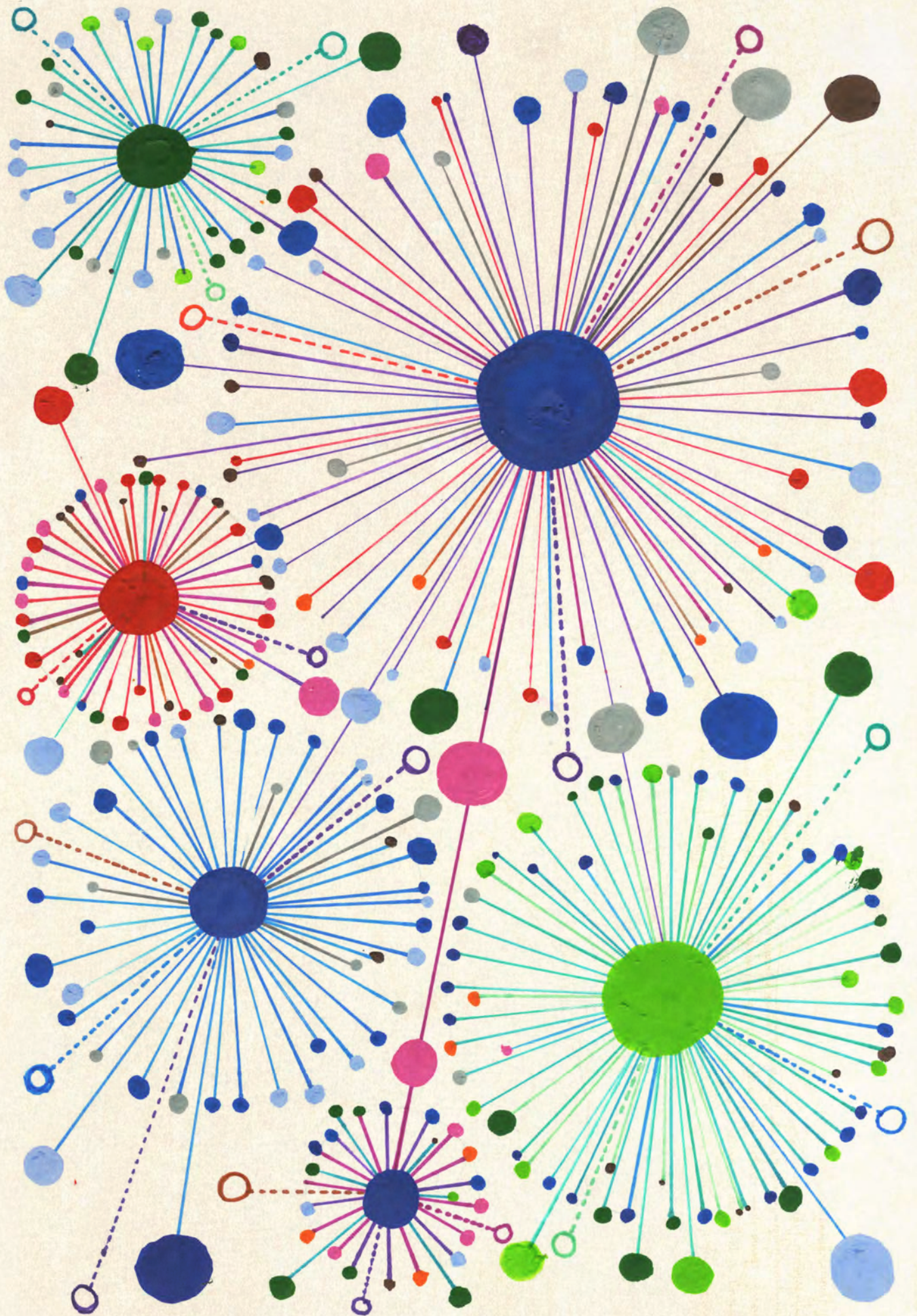
› Activity-based simulation

**If you already have an understanding of what factors govern the system, start with:**

- › ODES
- › PDES

**If you have imprecise categories**

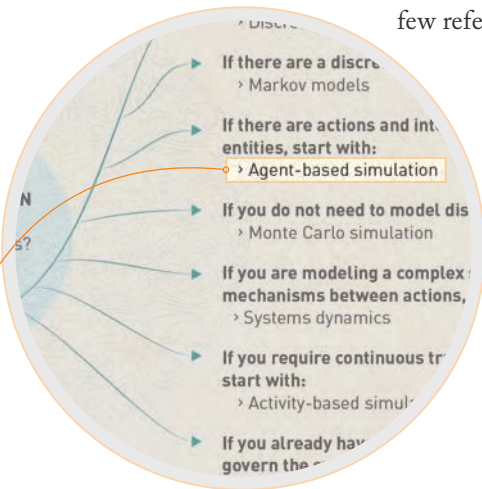
› Fuzzy logic



# Detailed Table of Analytics

Getting you to the right starting point isn't enough. We also provide a translator so you understand what you've been told.

Identifying several analytic techniques that can be applied to your problem is useful, but their name alone will not be much help. The *Detailed Table of Analytics* translates the names into something more meaningful. Once you've identified a technique in the Guide to Analytic Selection, find the corresponding row in the table. There you will find a brief description of the techniques, tips we've learned and a few references we've found helpful.



Technique	Description	Tips From the Pros	References we love to read
<b>Active Learning</b>	Intelligent sample selection to improve performance of model. Samples are selected to provide the greatest information to a learning model.	Can be paired with a human in-the-loop to help capture domain knowledge.	Burr, Settles B. "Active Learning: Synthesis Lectures on Artificial Intelligence and Machine Learning." Morgan & Claypool, 2012. Print.
<b>Agent Based Simulation</b>	Simulates the actions and interactions of autonomous agents.	In many systems, complex behavior results from surprisingly simple rules. Keep the logic of your agents simple and gradually build in sophistication.	Macal, Charles, and Michael North. "Agent-based Modeling and Simulation." Winter Simulation Conference. Austin, TX. 2009. Conference Presentation.
<b>ANOVA</b>	Hypothesis testing for differences between more than two groups.	Check model assumptions before utilizing, and watch out for Family Wise error when running multiple tests.	Bhattacharyya, Gouri K., and Richard A. Johnson. <i>Statistical Concepts and Models</i> . Wiley, 1977. Print.
<b>Association Rule Mining (Apriori)</b>	Data mining technique to identify the common co-occurrences of items.	Utilize when you have a need to understand potential relationships between elements.	Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules." <i>Proc. Of 20th Intl. Conf. on VLDB</i> . 1994. Conference Presentation.
<b>Bayesian Network</b>	Models conditional probabilities amongst elements, visualized as a Directed Acyclic Graph.	Calculate by hand before using larger models to ensure understanding.	Russel, Stuart, and Peter Norvig. "Artificial Intelligence: A Modern Approach." Prentice Hall, 2009 Print.

Compiled by: Booz Allen Hamilton

Technique	Description	Tips From the Pros	References we love to read
<b>Collaborative Filtering</b>	Also known as 'Recommendation,' suggest or eliminate items from a set by comparing a history of actions against items performed by users. Finds similar items based on who used them or similar users based on the items they use.	Use Singular Value Decomposition based Recommendation for cases where there are latent factors in your domain, e.g., genres in movies.	Owen, Sean, Robin Anil, Ted Dunning, and Ellen Friedman. <i>Mahout in Action</i> . New Jersey: Manning, 2012. Print.
<b>Coordinate Transformation</b>	Provides a different perspective on data.	Changing the coordinate system for data, for example, using polar or cylindrical coordinates, may more readily highlight key structure in the data. A key step in coordinate transformations is to appreciate multidimensionality and to systematically analyze subspaces of the data.	Abbott, Edwin A., <i>Flatland: A Romance of Many Dimensions</i> . United Kingdom: Seely & Co., 1884. Print.
<b>Deep Learning</b>	Method that learns features that leads to higher concept learning. Usually very deep neural network architectures.	Utilize a GPU to efficiently train complex models.	Bengio, Yoshua, and Yann LeCun. "Scaling Learning Algorithms towards AI." <i>Large-Scale Kernel Machines</i> . New York: MIT Press, 2007. Print.
<b>Design of Experiments</b>	Applies controlled experiments to quantify effects on system output caused by changes to inputs.	Fractional factorial designs can significantly reduce the number of different types of experiments you must conduct.	Montgomery, Douglas. <i>Design and Analysis of Experiments</i> . New Jersey: John Wiley & Sons, 2012. Print.
<b>Differential Equations</b>	Used to express relationships between functions and their derivatives, for example, change over time.	Differential equations can be used to formalize models and make predictions. The equations themselves can be solved numerically and tested with different initial conditions to study system trajectories.	Zill, Dennis, Warren Wright, and Michael Cullen. <i>Differential Equations with Boundary-Value Problems</i> . Connecticut: Cengage Learning, 2012. Print.
<b>Discrete Event Simulation</b>	Simulates a discrete sequence of events where each event occurs at a particular instant in time. The model updates its state only at points in time when events occur.	Discrete event simulation is useful when analyzing event based processes such as production lines and service centers to determine how system level behavior changes as different process parameters change. Optimization can integrate with simulation to gain efficiencies in a process.	Burrus, C. Sidney, Ramesh A. Gopinath, Haitao Guo, Jan E. Odegard and Ivan W. Selesnick. <i>Introduction to Wavelets and Wavelet Transforms: A Primer</i> . New Jersey: Prentice Hall, 1998. Print.
<b>Discrete Wavelet Transform</b>	Transforms time series data into frequency domain preserving locality information.	Offers very good time and frequency localization. The advantage over Fourier transforms is that it preserves both frequency and locality.	Burrus, C. Sidney, Ramesh A. Gopinath, Haitao Guo, Jan E. Odegard, and Ivan W. Selesnick. <i>Introduction to Wavelets and Wavelet Transforms: A Primer</i> . New Jersey: Prentice Hall, 1998. Print.
<b>Ensemble Learning</b>	Learning multiple models and combining output to achieve better performance.	Be careful not to overfit data by having too many model parameters and overtraining.	Dietterich, Thomas G. "Ensemble Methods in Machine Learning." <i>Lecture Notes in Computer Science</i> . Springer, 2000. Print.
<b>Expert Systems</b>	Systems that use symbolic logic to reason about facts. Emulates human reasoning.	Useful to have a human readable explanation of why a system came to a conclusion.	Shortliffe, Edward H., and Bruce G. Buchanan. "A Model of Inexact Reasoning in Medicine." <i>Mathematical Biosciences</i> . Elsevier B.V., 1975. Print.
<b>Exponential Smoothing</b>	Used to remove artifacts expected from collection error or outliers.	In comparison to a using moving average where past observations are weighted equally, exponential smoothing assigns exponentially decreasing weights over time.	Chatfield, Chris, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. "A New Look at Models for Exponential Smoothing." <i>Journal of the Royal Statistical Society: Series D (The Statistician)</i> . Royal Statistical Society, 2001. Print.

Compiled by: Booz Allen Hamilton



Technique	Description	Tips From the Pros	References we love to read
<b>Factor Analysis</b>	Describes variability among correlated variables with the goal of lowering the number of unobserved variables, namely, the factors.	If you suspect there are immeasurable influences on your data, then you may want to try factor analysis.	Child, Dennis. <i>The Essentials of Factor Analysis</i> . United Kingdom: Cassell Educational, 1990. Print.
<b>Fast Fourier Transform</b>	Transforms time series from time to frequency domain efficiently. Can also be used for image improvement by spatial transforms.	Filtering a time varying signal can be done more effectively in the frequency domain. Also, noise can often be identified in such signals by observing power at aberrant frequencies.	Mitra, Partha P., and Hemant Bokil. <i>Observed Brain Dynamics</i> . United Kingdom: Oxford University Press, 2008. Print.
<b>Format Conversion</b>	Creates a standard representation of data regardless of source format. For example, extracting raw UTF-8 encoded text from binary file formats such as Microsoft Word or PDFs.	There are a number of open source software packages that support format conversion and can interpret a wide variety of formats. One notable package is Apache Tika.	Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris. <i>Taming Text: How to Find, Organize, and Manipulate It</i> . New Jersey: Manning, 2013. Print.
<b>Fuzzy Logic</b>	Logical reasoning that allows for degrees of truth for a statement.	Utilize when categories are not clearly defined. Concepts such as "warm", "cold", and "hot" can mean different things at different temperatures and domains.	Zadeh L.A., "Fuzzy Sets." <i>Information and Control</i> . California: University of California, Berkeley, 1965. Print.
<b>Gaussian Filtering</b>	Acts to remove noise or blur data.	Can be used to remove speckle noise from images.	Parker, James R. <i>Algorithms for Image Processing and Computer Vision</i> . New Jersey: John Wiley & Sons, 2010. Print.
<b>Generalized Linear Models</b>	Expands ordinary linear regression to allow for error distribution that is not normal.	Use if the observed error in your system does not follow the normal distribution.	MacCullagh, P., and John A. Nelder. <i>Generalized Linear Models</i> . Florida: CRC Press, 1989. Print.
<b>Genetic Algorithms</b>	Evolves candidate models over generations by evolutionary inspired operators of mutation and crossover of parameters.	Increasing the generation size adds diversity in considering parameter combinations, but requires more objective function evaluation. Calculating individuals within a generation is strongly parallelizable. Representation of candidate solutions can impact performance.	De Jong, Kenneth A. <i>Evolutionary Computation - A Unified Approach</i> . Massachusetts: MIT Press, 2002. Print.
<b>Grid Search</b>	Systematic search across discrete parameter values for parameter exploration problems.	A grid across the parameters is used to visualize the parameter landscape and assess whether multiple minima are present.	Kolda, Tamara G., Robert M. Lewis, and Virginia Torczon. "Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods." <i>SIAM Review</i> . Society for Industrial and Applied Mathematics, 2003. Print.
<b>Hidden Markov Models</b>	Models sequential data by determining the discrete latent variables, but the observables may be continuous or discrete.	One of the most powerful properties of Hidden Markov Models is their ability to exhibit some degree of invariance to local warping (compression and stretching) of the time axis. However, a significant weakness of the Hidden Markov Model is the way in which it represents the distribution of times for which the system remains in a given state.	Bishop, Christopher M. <i>Pattern Recognition and Machine Learning</i> . New York: Springer, 2006. Print.
<b>Hierarchical Clustering</b>	Connectivity based clustering approach that sequentially builds bigger (agglomerative) or smaller (divisive) clusters in the data.	Provides views of clusters at multiple resolutions of closeness. Algorithms begin to slow for larger datasets due to most implementations exhibiting $O(N^3)$ or $O(N^2)$ complexity.	Rui Xu, and Don Wunsch. <i>Clustering</i> . New Jersey: Wiley-IEEE Press, 2008. Print.
<b>K-means and X-means Clustering</b>	Centroid based clustering algorithms, where with K means the number of clusters is set and X means the number of clusters is unknown.	When applying clustering techniques, make sure to understand the shape of your data. Clustering techniques will return poor results if your data is not circular or ellipsoidal shaped.	Rui Xu, and Don Wunsch. <i>Clustering</i> . New Jersey: Wiley-IEEE Press, 2008. Print.

Compiled by: Booz Allen Hamilton

Technique	Description	Tips From the Pros	References we love to read
<b>Linear, Non-linear, and Integer Programming</b>	Set of techniques for minimizing or maximizing a function over a constrained set of input parameters.	Start with linear programs because algorithms for integer and non-linear variables can take much longer to run.	Winston, Wayne L. <i>Operations Research: Applications and Algorithms</i> . Connecticut: Cengage Learning, 2003. Print.
<b>Markov Chain Monte Carlo (MCMC)</b>	A method of sampling typically used in Bayesian models to estimate the joint distribution of parameters given the data.	Problems that are intractable using analytic approaches can become tractable using MCMC, when even considering high-dimensional problems. The tractability is a result of using statistics on the underlying distributions of interest, namely, sampling with Monte Carlo and considering the stochastic sequential process of Markov Chains.	Andrieu, Christophe, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. "An Introduction to MCMC for Machine Learning." <i>Machine Learning</i> . Kluwer Academic Publishers, 2003. Print.
<b>Monte Carlo Methods</b>	Set of computational techniques to generate random numbers.	Particularly useful for numerical integration, solutions of differential equations, computing Bayesian posteriors, and high dimensional multivariate sampling.	Fishman, George S. <i>Monte Carlo: Concepts, Algorithms, and Applications</i> . New York: Springer, 2003. Print.
<b>Naïve Bayes</b>	Predicts classes following Bayes Theorem that states the probability of an outcome given a set of features is based on the probability of features given an outcome.	Assumes that all variables are independent, so it can have issues learning in the context of highly interdependent variables. The model can be learned on a single pass of data using simple counts and therefore is useful in determining whether exploitable patterns exist in large datasets with minimal development time.	Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris. <i>Taming Text: How to Find, Organize, and Manipulate It</i> . New Jersey: Manning, 2013. Print.
<b>Neural Networks</b>	Learns salient features in data by adjusting weights between nodes through a learning rule.	Training a neural network takes substantially longer than evaluating new data with an already trained network. Sparser network connectivity can help to segment the input space and improve performance on classification tasks.	Haykin, Simon O. <i>Neural Networks and Learning Machines</i> . New Jersey: Prentice Hall, 2008. Print.
<b>Outlier Removal</b>	Method for identifying and removing noise or artifacts from data.	Be cautious when removing outliers. Sometimes the most interesting behavior of a system is at times when there are aberrant data points.	Maimon, Oded, and Lior Rockach. <i>Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers</i> . The Netherlands: Kluwer Academic Publishers, 2005. Print.
<b>Principal Components Analysis</b>	Enables dimensionality reduction by identifying highly correlated dimensions.	Many large datasets contain correlations between dimensions; therefore part of the dataset is redundant. When analyzing the resulting principal components, rank order them by variance as this is the highest information view of your data. Use scree plots to infer the optimal number of components.	Wallisch, Pascal, Michael E. Lusignan, Marc D. Benayoun, Tanya I. Baker, Adam Seth Dickey, and Nicholas G. Hatsopoulos. <i>Matlab for Neuroscientists</i> . New Jersey: Prentice Hall, 2009. Print.
<b>Random Search</b>	Randomly adjust parameters to find a better solution than currently found.	Use as a benchmark for how well a search algorithm is performing. Be careful to use a good random number generator and new seed.	Bergstra J. and Bengio Y. Random Search for Hyper-Parameter Optimization, <i>Journal of Machine Learning Research</i> 13, 2012.
<b>Regression with Shrinkage (Lasso)</b>	A method of variable selection and prediction combined into a possibly biased linear model.	There are different methods to select the lambda parameter. A typical choice is cross validation with MSE as the metric.	Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." <i>Journal of the Royal Statistical Society. Series B (Methodological)</i> . Toronto: Royal Statistical Society, 1996. Print.
<b>Sensitivity Analysis</b>	Involves testing individual parameters in an analytic or model and observing the magnitude of the effect.	Insensitive model parameters during an optimization are candidates for being set to constants. This reduces the dimensionality of optimization problems and provides an opportunity for speed up.	Saltelli, A., Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. <i>Global Sensitivity Analysis: the Primer</i> . New Jersey: John Wiley & Sons, 2008. Print.

Compiled by: Booz Allen Hamilton

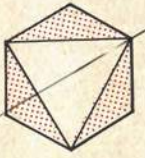
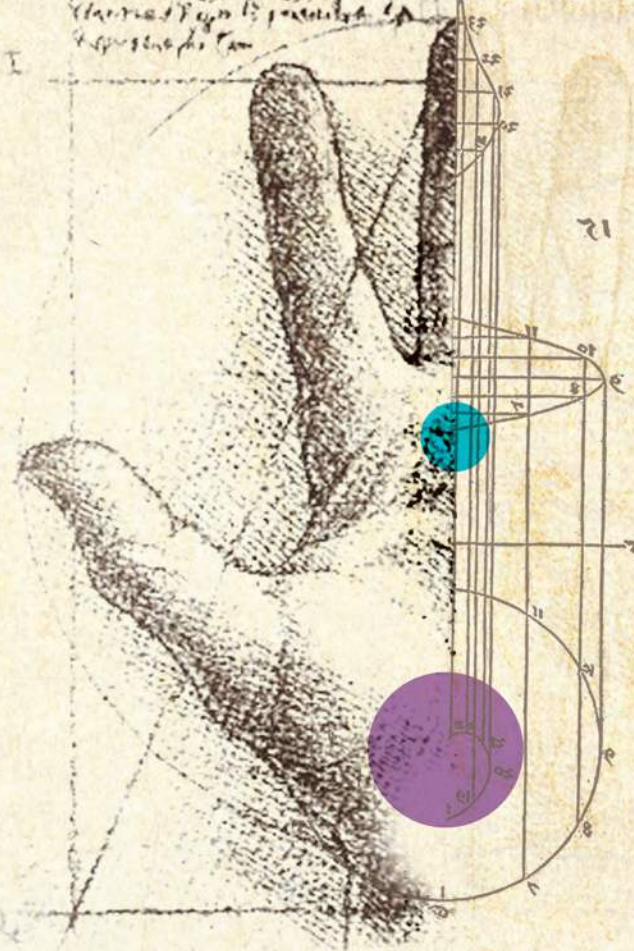
Technique	Description	Tips From the Pros	References we love to read
<b>Simulated Annealing</b>	Named after a controlled cooling process in metallurgy, and by analogy using a changing temperature or annealing schedule to vary algorithmic convergence.	The standard annealing function allows for initial wide exploration of the parameter space followed by a narrower search. Depending on the search priority the annealing function can be modified to allow for longer explorative search at a high temperature.	Bertsimas, Dimitris, and John Tsitsiklis. "Simulated Annealing." <i>Statistical Science</i> . 1993. Print.
<b>Stepwise Regression</b>	A method of variable selection and prediction. Akaike's information criterion <i>AIC</i> is used as the metric for selection. The resulting predictive model is based upon ordinary least squares, or a general linear model with parameter estimation via maximum likelihood.	Caution must be used when considering Stepwise Regression, as over fitting often occurs. To mitigate over fitting try to limit the number of free variables used.	Hocking, R.R. "The Analysis and Selection of Variables in Linear Regression." <i>Biometrics</i> . 1976. Print.
<b>Stochastic Gradient Descent</b>	General-purpose optimization for learning of neural networks, support vector machines, and logistic regression models.	Applied in cases where the objective function is not completely differentiable when using sub-gradients.	Witten, Ian H., Eibe Frank, and Mark A. Hall. <i>Data Mining: Practical Machine Learning Tools and Techniques</i> . Massachusetts: Morgan Kaufmann, 2011. Print.
<b>Support Vector Machines</b>	Projection of feature vectors using a kernel function into a space where classes are more separable.	Try multiple kernels and use k-fold cross validation to validate the choice of the best one.	Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification." National Taiwan University Press, 2003. Print.
<b>Term Frequency Inverse Document Frequency</b>	A statistic that measures the relative importance of a term from a corpus.	Typically used in text mining. Assuming a corpus of news articles, a term that is very frequent such as "the" will likely appear many times in many documents, having a low value. A term that is infrequent such as a person's last name that appears in a single article will have a higher TD IDF score.	Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris. <i>Taming Text: How to Find, Organize, and Manipulate It</i> . New Jersey: Manning, 2013. Print.
<b>Topic Modeling (Latent Dirichlet Allocation)</b>	Identifies latent topics in text by examining word co-occurrence.	Employ part-of-speech tagging to eliminate words other than nouns and verbs. Use raw term counts instead of TF/IDF weighted terms.	Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." <i>Journal of Machine Learning Research</i> . 2003. Print.
<b>Tree Based Methods</b>	Models structured as graph trees where branches indicate decisions.	Can be used to systematize a process or act as a classifier.	James, G., D. Witten, T. Hastie, and R. Tibshirani. "Tree Based Methods." <i>An Introduction to Statistical Learning</i> . New York: Springer, 2013. Print.
<b>T-Test</b>	Hypothesis test used to test for differences between two groups.	Make sure you meet the tests assumptions and watch out for Family Wise error when running multiple tests.	Bhattacharyya, Gouri K., and Richard A. Johnson. <i>Statistical Concepts and Models</i> . Wiley, 1977. Print.
<b>Wrapper Methods</b>	Feature set reduction method that utilizes performance of a set of features on a model, as a measure of the feature set's performance. Can help identify combinations of features in models that achieve high performance.	Utilize k-fold cross validation to control over fitting.	John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant Features and the Subset Selection Problem." <i>Proceedings of ICML-94, 11th International Conference on Machine Learning</i> . New Brunswick, New Jersey. 1994. <i>Conference Presentation</i> .

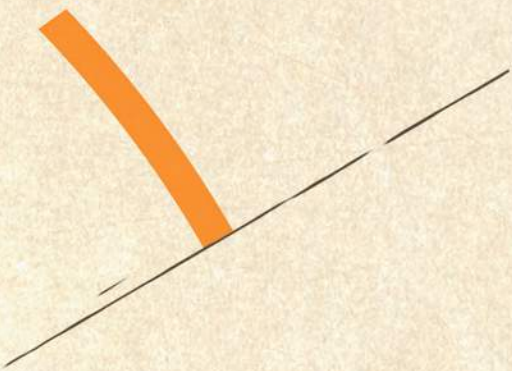
Compiled by: Booz Allen Hamilton





Handwritten text in a cursive script, likely a historical manuscript, located above the hand drawing.





# LIFE *in* THE TRENCHES

## NAVIGATING NECK DEEP IN DATA

Our Data Science experts have learned and developed new solutions over the years from properly framing or reframing analytic questions. In this section, we list a few important topics to Data Science coupled with firsthand experience from our experts.



# Going Deep into Machine Learning

---

Machines are getting better at learning by mimicking the human brain.

---

Think about where you were 10 years ago. Could computers understand and take action based upon your spoken word? Recently, speech-to-text quality has improved dramatically to nearly perfect accuracy, much to the delight of many mobile phone users. In other complex tasks, similar magic capabilities have emerged. The world-record high scores in 29 video games are now held by a machine learning algorithm with no specific knowledge of Atari or computer games in general.

These impressive feats were made possible by deep learning, a different way of approaching machine learning problems. Most approaches to machine learning require humans to encode logic and rules to create features, which are then fed into machine learning models. In some domains, such as audio, text, image, and signal processing, effective feature engineering requires considerable human expertise to achieve decent model performance. Deep learning avoids the necessity of human-encoded features and instead incorporates the feature engineering, feature selection, and model fitting into one step.

Deep learning is an extension of an idea originating in the 1950s called neural networks, which are loosely inspired by our understanding of how neurons in our brains operate. Recent hardware developments, originally designed for faster renderings of graphics, birthed a renaissance in neural networks. The latest graphical processing units, or GPUs, have more than 3,000 processing cores that are well suited for parallel processing of complex matrix manipulations required for rendering graphics – and for executing computations on neural networks.

In the late 2000s, the combination of GPUs, advances in algorithms, and collections of big data reignited interest in neural networks. GPUs enabled computers to process much larger networks in much less time, and clever advances in algorithms made the model fitting process more efficient. Large collections of image, video, and text data provided content for the deep learning algorithms to learn. The ability to train larger networks with more data drove the exploration of new neural network architectures featuring additional hidden layers and widening the breadth of the networks.

Presently, deep learning has moved beyond academic applications and is finding its way into our daily lives. Deep learning powers speech-to-text on our mobile phones and smart devices, image search provided by major tech companies, language translation services for text and spoken word, and even drug discovery within advanced pharmaceutical companies.

## »» National Data Science Bowl



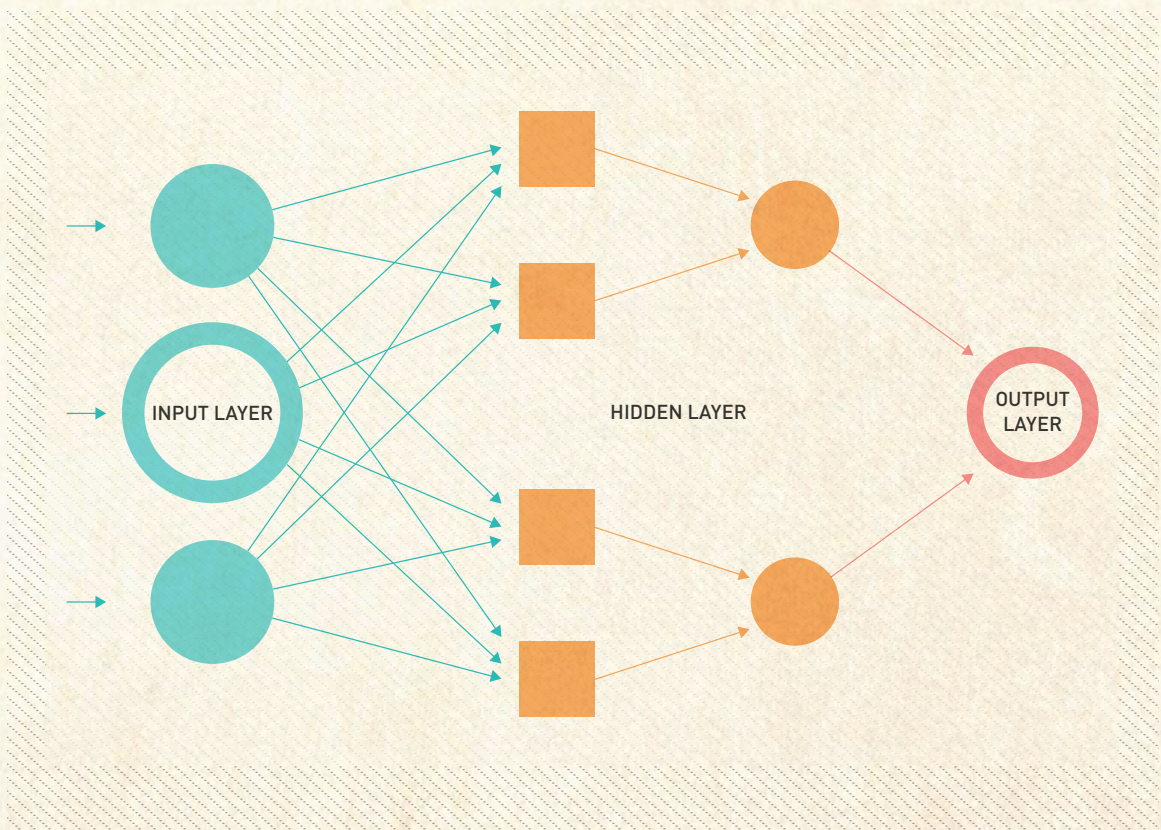
Aaron Sander

The first-ever National Data Science Bowl offered Data Scientists a platform through which individuals could harness their passion, unleash their curiosity and amplify their impact to affect change on a global scale. The competition presented participants with more than 100,000 underwater images provided by the Hatfield Marine Science Center. Participants were challenged to develop a classification algorithm that would enable researchers to monitor ocean health at a speed and scale never before possible.

More than 1,000 teams submitted a total of approximately 15,000 solutions over the 90 days of the competition. A large proportion of the participants' implemented solutions used deep learning-based approaches, specifically Convolutional Neural Nets (CNNs). The competition forum exploded with competitors collectively sharing knowledge and collaborating to advance the state-of-the-art in computer vision. Participants tested new techniques for developing CNNs and contributed to the development of open source software for creating CNN models. The top three competitors, Team

Deep Sea, Happy Lantern Festival, and Poisson Process, all used CNNs in their solutions. Their results increased algorithm accuracy by 10% over the state of the art. Without their algorithms, it would have taken marine researchers more than two lifetimes to manually complete the classification process. The work submitted by all the participants represents major advances for both the marine research and Data Science communities.

»Visit [www.DataScienceBowl.com](http://www.DataScienceBowl.com) to learn more about the first-ever National Data Science Bowl



Source: Booz Allen Hamilton

A Representation of Deep Learning



# Feature Engineering

---

Feature engineering is a lot like oxygen. You can't do without it, but you rarely give it much thought.

---



## »What differentiates a great model from a bad one?

---

In most cases, the inputs to the model matter even more than the choice of the algorithm. The traditional approach has a two-step process where heuristics and subject-matter expertise are used to find a good set of features, and then algorithms optimize model parameters. Deep learning combines these steps into one. Feature engineering, feature selection, and model parameter estimation are accomplished simultaneously. This reduces the need for highly-specialized domain knowledge and often results in better models.

For example, in the context of images from the natural world, deep networks may learn low-level features such as lines at various angles and curved lines. Middle layers may combine the lower-level features into more complex geometric shapes and patterns. Higher-level layers combine the mid-level features into more complicated features that begin to resemble faces and shapes of animals. Applied to other types of data, such as audio and text, deep neural networks learn increasingly sophisticated features at each layer in a similar manner.

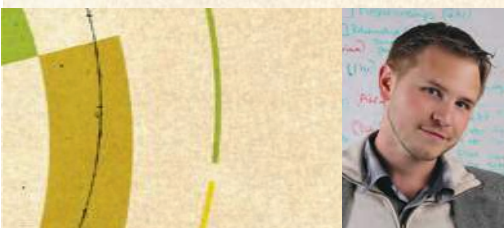
Feature engineering is the process by which one establishes the representation of data in the context of an analytic approach. It is the foundational skill in Data Science. Without feature engineering, it would not be possible to understand and represent the world through a mathematical model. Feature engineering is a challenging art. Like other arts, it is a creative process that manifests uniquely in each Data Scientist. It will be influenced substantially by the scientist's experiences, tastes and understanding of the field.

When faced with the problem of feature engineering, there are several paths that one may initially take. Generally speaking, better features can be developed with more knowledge of the domain. One approach to feature engineering is to begin by describing smaller elements of the domain and continuously constructing more intricate features as the model becomes more complex. These more complicated features can be defined by considering other attributes of the domain, aggregating features together into different groups or using advanced statistical and mathematical techniques to devise new features. The ultimate judge in this process is the performance of the machine learning algorithm which makes decisions based on this feature vector.

Consider the example of email spam classification. Because the domain is a set of emails, one possible choice of an initial feature vector is the integer array that counts the number of times a given word appears in the email. This is called the "bag of words" assumption, where the order that words appear in a text is ignored. If an algorithm with this feature vector does not adequately distinguish between spam and non-spam emails, a feature could be added that counts the number of misspelled words in the text. This new feature uses the spam recognition domain knowledge that asserts many spam emails misspell words. These misspelled words alert filters saying that the existence of certain words automatically label an email as spam. If this new feature is not enough, there are still additional features to be considered such as whether or not a first or last name is used in the email.

Feature engineering represents a complex, but crucial aspect of Data Science. The Learning Optimal Features sidebar goes into detail about feature learning – an automated approach to feature engineering that applies machine learning techniques.

## » Chemoinformatic Search



Ed Kohlwey

On one assignment, my team was confronted with the challenge of developing a search engine over chemical compounds. The goal of chemoinformatic search is to predict the properties that

a molecule will exhibit as well as to provide indices over those predicted properties to facilitate data discovery in chemistry-based research. These properties may either be discreet (e.g., “a molecule treats disease x well”) or continuous (e.g., “a molecule may be dissolved up to 100.21 g/ml”).

Molecules are complex 3D structures, which are typically represented as a list of atoms joined by chemical bonds of differing lengths with varying electron domain and molecular geometries. The structures are specified by the 3-space coordinates and the electrostatic potential surface of the atoms in the molecule. Searching this data is a daunting task when one considers that naive approaches to the problem bear significant semblance to the Graph Isomorphism Problem.<sup>[15]</sup>

The solution we developed was based on previous work in molecular fingerprinting (sometimes also called hashing or locality sensitive hashing). Fingerprinting is a dimensionality reduction technique that dramatically reduces the problem space by summarizing many features, often with relatively little regard to the importance of the feature. When an exact solution is likely to be infeasible, we often turn to heuristic approaches such as fingerprinting.

Our approach used a training set where all the measured properties

of the molecules were available. We created a model of how molecular structural similarities might affect their properties. We began by finding all the sub-graphs of each molecule with length  $n$ , resulting in a representation similar to the bag-of-words approach from natural language processing. We summarized each molecule fragment in a type of fingerprint called a “Counting Bloom Filter.”

Next, we used several exemplars from the set to create new features. We found the distance from each member of the full training set to each of the exemplars. We fed these features into a non-linear regression algorithm to yield a model that could be used on data that was not in the original training set. This approach can be conceptualized as a “hidden manifold,” whereby a hidden surface or shape defines how a molecule will exhibit a property. We approximate this shape using a non-linear regression and a set of data with known properties. Once we have the approximate shape, we can use it to predict the properties of new molecules.

Our approach was multi-staged and complex – we generated sub-graphs, created bloom filters, calculated distance metrics and fit a linear-regression model. This example provides an illustration of how many stages may be involved in producing a sophisticated feature representation. By creatively combining and building “features on features,” we were able to create new representations of data that were richer and more descriptive, yet were able to execute faster and produce better results.

# Feature Selection

---

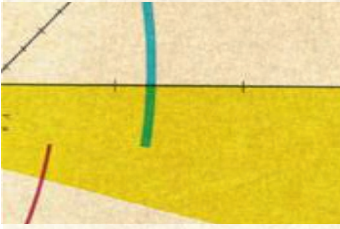
Models are like honored guests; you should only feed them the good parts.

---

Feature selection is the process of determining the set of features with the highest information value to the model. Two main approaches are filtering and wrapper methods. Filtering methods analyze features using a test statistic and eliminate redundant or non-informative features. As an example, a filtering method could eliminate features that have little correlation to the class labels. Wrapper methods utilize a classification model as part of feature selection. A model is trained on a set of features and the classification accuracy is used to measure the information value of the feature set. One example is that of training a neural network with a set of features and evaluating the accuracy of the model. If the model scores highly on the test set, then the features have high information value. All possible combinations of features are tested to find the best feature set.

There are tradeoffs between these techniques. Filtering methods are faster to compute since each feature only needs to be compared against its class label. Wrapper methods, on the other hand, evaluate feature sets by constructing models and measuring performance. This requires a large number of models to be trained and evaluated (a quantity that grows exponentially in the number of features). Why would anyone use a wrapper method? Feature sets may perform better than individual features.<sup>[16]</sup> With filter methods, a feature with weak correlation to its class labels is eliminated. Some of these eliminated features, however, may have performed well when combined with other features.

## »» Cancer Cell Classification



Paul Yacci

to classify different types of Leukemia, based on Microarray profiles from 72 samples<sup>[17]</sup> using a small set of features. We utilized a hybrid Artificial Neural Network (ANN)<sup>[18]</sup> and Genetic Algorithm<sup>[19]</sup> to identify subsets of 10 features selected from thousands.<sup>[20]</sup> We trained the ANN and tested performance using cross-fold validation. The performance measure

On one project, our team was challenged to classify cancer cell profiles. The overarching goal was

was used as feedback into the Genetic Algorithm. When a set of features contained no useful information, the model performed poorly and a different feature set would be explored. Over time, this method selected a set of features that performed with high accuracy. The down-selected feature set increased speed and performance as well as allowed for better insight into the factors that may govern the system. This allowed our team to design a diagnostic test for only a few genetic markers instead of thousands, substantially reducing diagnostic test complexity and cost.

# Ensemble Models

---

None of us is as smart as all of us, but some are smarter than others.

---

In 1906, Sir Francis Galton attended a fair at which there was a contest to guess the weight of an ox. Galton had the idea to collect the guesses of the 787 entrants and compute the mean. To his surprise, the mean was only one pound off from the ox's real weight of 1,198 pounds. Together, the estimates made by many amateurs formed a prediction that was more accurate than that of individual experts.

Galton's "wisdom of crowds" extends to Data Science in the form of ensemble learning, which is colloquially and somewhat equivalently called ensembling, blending, or stacking. An ensemble takes the predictions of many individual models and combines them to make a single prediction. Like the people guessing the ox's weight, Data Science models have unique strengths and weaknesses (i.e., determined by their design), and are influenced by varied perspectives based on past experience (i.e., the data they have observed).

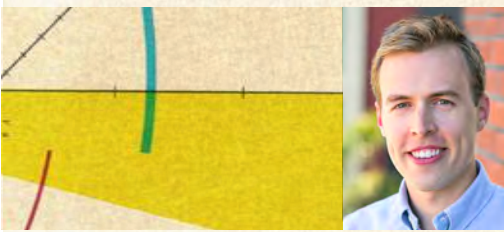
Ensembles overcome individual weaknesses to make predictions with more accuracy than their constituent models. These models need not stem from different methodologies; an ensemble might employ the same method with different parameters or weights (e.g., boosting), feature subsets (e.g., random forests), or sub-samples of data (e.g., bagging). The ensembling methodology may be as simple as averaging two outputs, or as complex as using a "meta model" to learn an optimal combination.

An ensemble's ability to reduce individual errors arises from the diversity of its members. If one model over fits the data, it is balanced by a different model that under fits the data. If one subset is skewed by outlier values, another subset is included without them. If one method is unstable to noisy inputs, it is bolstered by another method that is more robust.

In practice, ensembling typically improves a model by a few percent. The price of this accuracy is paid in complexity. The accuracy vs. complexity tradeoff can make it difficult to know when ensembling is justified. On one hand, ensembles appear to be a fit for high-stakes problems—think detecting cancer in MRI images vs. detecting unripe blueberries on a conveyer belt. On the other hand, high-stakes problems mandate higher standards for auditing model functionality.

The Data Scientist must manage a balance between ensemble interpretability and black-box complexity. If this seems easy, it isn't! Put yourself in the driver's seat of the machine learning code for a self-driving car. If a well-behaved regression model makes a right decision 99.5% of the time, but a complex, less-explainable ensemble is right 99.8% of the time, which would you pick?

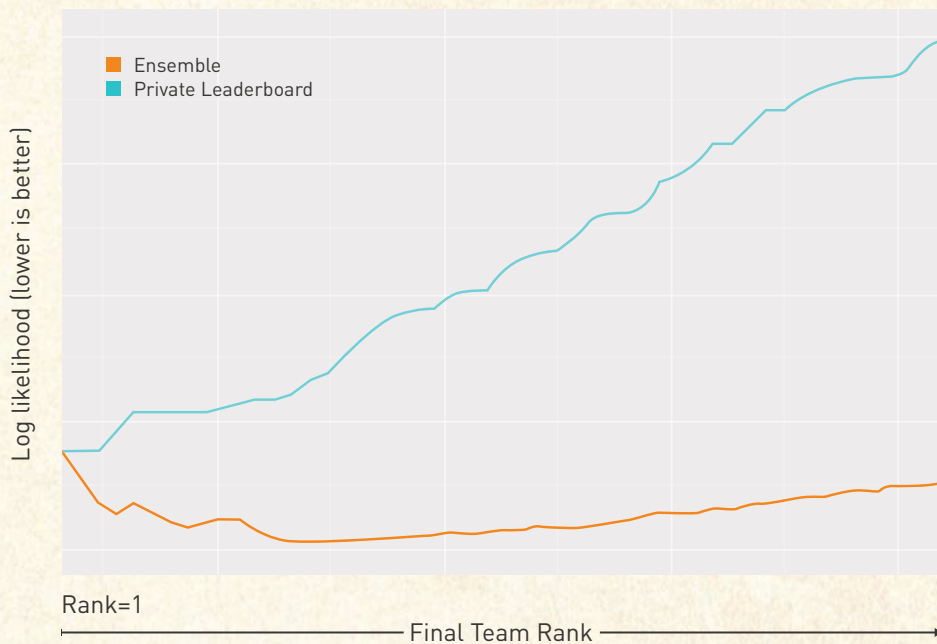
## »» The Value of Ensemble Models



Will Cukierski

Several years ago, the Kaggle Photo Quality Prediction competition posed the question “Given anonymized information on thousands of photo albums, predict whether a human evaluator would mark them as ‘good’.” Participants were supplied a large collection of user-generated photos. The goal was to create an algorithm that could automatically pick out particularly enjoyable or impressive photos from the collection.

Over the course of the competition, 207 people submitted entries. The log likelihood metric was used to evaluate the accuracy of the entries. Scores for the top 50 teams ranged from 0.18434 to 0.19884, where lower is better. Kaggle data scientist Ben Hamner used the results to illustrate the value of ensembling by means of averaging the top 50 scores. The figure below shows the results.



Value of Ensembling for the Kaggle Photo Quality Prediction Competition  
(results courtesy of Ben Hamner, Kaggle CTO)

The blue line shows the individual scores for each of the top 50 teams. The orange line shows the ensemble score for the top  $n$  teams, where  $n$  ranges from 1 to the value on the axis. For example, the ensemble point for Final Team Rank 5 is an ensemble of the entries for teams 1 through 5. As shown in the graph, the ensemble score is lower than any single individual score. The diversity of models included within the ensemble causes the respective errors to cancel out, resulting in an overall lower score. This holds true for

all points across the top 50 teams. However, after we increase the number of models in the ensemble beyond 15, we begin to see the ensemble score increase. This occurs because we are introducing less accurate (i.e., potentially overfit) models into the ensemble. The results of this simple experiment quantify the value of creating an ensemble model, while reinforcing the idea that we must be thoughtful when selecting the individual models contained within the ensemble.

# Data Veracity

---

We're Data Scientists, not data alchemists. We can't make analytic gold from the lead of data.

---

While most people associate data volume, velocity, and variety with big data, there is an equally important yet often overlooked dimension – data veracity. Data veracity refers to the overall quality and correctness of the data. You must assess the truthfulness and accuracy of the data as well as identify missing or incomplete information. As the saying goes, “Garbage in, garbage out.” If your data is inaccurate or missing information, you can't hope to make analytic gold.

Assessing data truthfulness is often subjective. You must rely on your experience and an understanding of the data origins and context. Domain expertise is particularly critical for the latter. Although data accuracy assessment may also be subjective, there are times when quantitative methods may be used. You may be able to re-sample from the population and conduct a statistical comparison against the stored values, thereby providing measures of accuracy.

The most common issues you will encounter are missing or incomplete information. There are two basic strategies for dealing with missing values – deletion and imputation. In the former, entire observations are excluded from analysis, reducing sample size and potentially introducing bias. Imputation, or replacement of missing or erroneous values, uses a variety of techniques such as random sampling (hot deck imputation) or replacement using the mean, statistical distributions or models.

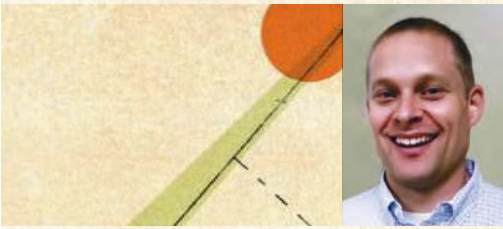


## »Tips From the Pros

---

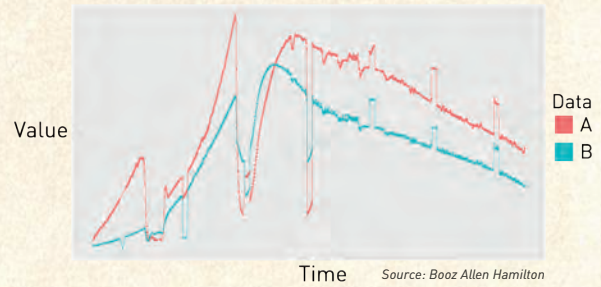
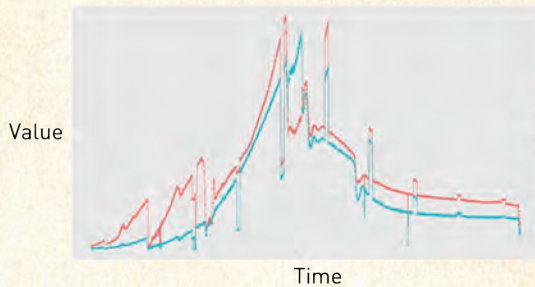
Find an approach that works, implement it, and move on. You can worry about optimization and tuning your approaches later during incremental improvement.

# » Time Series Modeling



Brian Keller

On one of our projects, the team was faced with correlating the time series for various parameters. Our initial analysis revealed that the correlations were almost non-existent. We examined the data and quickly discovered data veracity issues. There were missing and null values, as well as negative-value observations, an impossibility given the context of the measurements (see the figure, *Time Series Data Prior to Cleansing*). Garbage data meant garbage results.



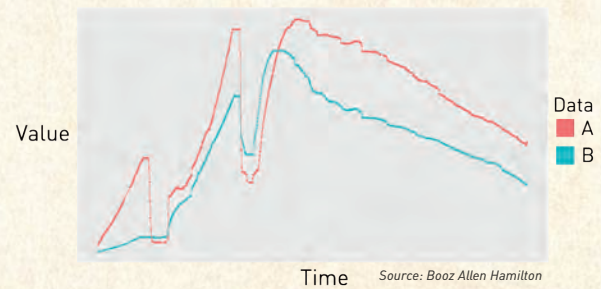
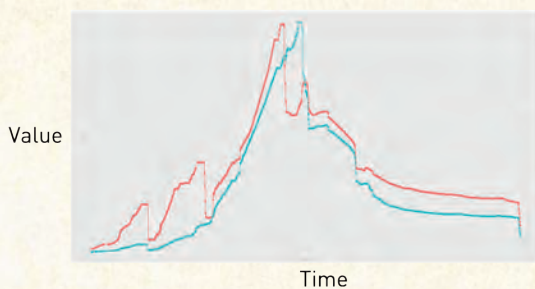
Time Series Data Prior to Cleansing

Because sample size was already small, deleting observations was undesirable. The volatile nature of the time series meant that imputation through sampling could not be trusted to produce values in which the team would be confident. As a result, we quickly realized that the best strategy was an approach that could filter and correct the noise in the data.

We initially tried a simplistic approach in which we replaced each observation with a moving average. While this corrected some noise, including the outlier values in our moving-average computation shifted the time series. This caused undesirable

distortion in the underlying signal, and we quickly abandoned the approach.

One of our team members who had experience in signal processing suggested a median filter. The median filter is a windowing technique that moves through the data point-by-point, and replaces it with the median value calculated for the current window. We experimented with various window sizes to achieve an acceptable tradeoff between smoothing noise and smoothing away signal. The figure, *Time Series Data After Cleansing*, shows the same two time series after median filter imputation.



Time Series Data After Cleansing

The application of the median filter approach was hugely successful. Visual inspection of the time series plots reveals smoothing of the outliers without dampening the naturally occurring peaks and troughs (no signal loss). Prior to smoothing, we saw no correlation in our data, but afterwards, Spearman's Rho was  $\sim 0.5$  for almost all parameters.

By addressing our data veracity issues, we were able to create analytic gold. While other approaches may also have been effective, implementation speed constraints prevented us from doing any further analysis. We achieved the success we were after and moved on to address other aspects of the problem.



# Application of Domain Knowledge

---

We are all special in our own way. Don't discount what you know.

---

Knowledge of the domain in which a problem lies is immensely valuable and irreplaceable. It provides an in-depth understanding of your data and the factors influencing your analytic goal. Many times domain knowledge is a key differentiator to a Data Science team's success. Domain knowledge influences how we engineer and select features, impute data, choose an algorithm, and determine success. One person cannot possibly be a domain expert in every field, however. We rely on our team, other analysts and domain experts as well as consult research papers and publications to build an understanding of the domain.

# » Motor Vehicle Theft

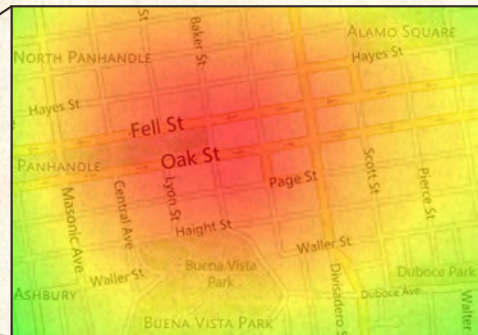
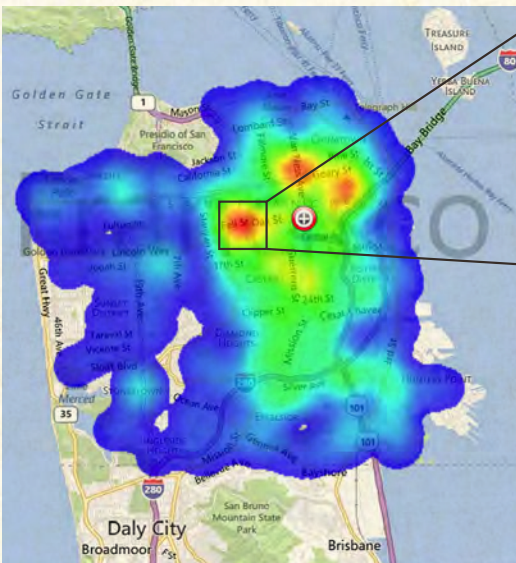


Armen Kherlopian

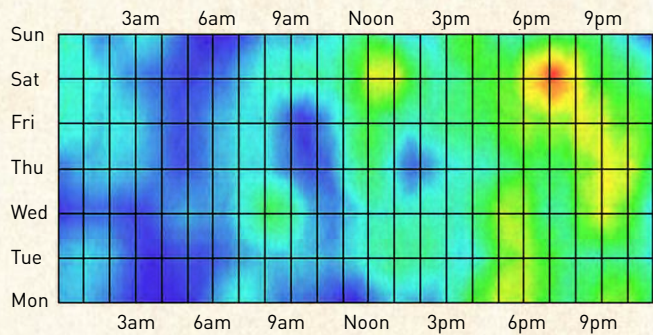
On one project, our team explored how Data Science could be applied to improve public safety. According to the FBI, approximately \$8 Billion is lost annually due to automobile theft. Recovery of the one million vehicles stolen every year in the U.S. is less than 60%. Dealing with these crimes represents a significant investment of law enforcement resources. We wanted to see if we could identify how to reduce auto theft while efficiently using law enforcement resources.

Our team began by parsing and verifying San Francisco crime data. We enriched stolen car reporting with general city data. After conducting several data experiments across both space and time, three geospatial and one temporal hotspot emerged (see figure, *Geospatial and Temporal Car Theft Hotspots*). The domain expert on the team was able to discern that the primary geospatial hotspot corresponded to an area surrounded by parks. The parks created an urban mountain with a number of over-foot access points that were conducive to car theft.

**GEOSPATIAL HOTSPOTS**



**TEMPORAL HOTSPOT**



Source: Booz Allen Hamilton

**Geospatial and Temporal Car Theft Hotspots**

Our team used the temporal hotspot information in tandem with the insights from the domain expert to develop a Monte Carlo model to predict the likelihood of a motor vehicle theft at particular city intersections. By prioritizing the intersections identified by the model, local governments would have the information necessary to efficiently deploy their patrols. Motor vehicle thefts could be reduced and law enforcement resources could be more efficiently deployed. The analysis, enabled by domain expertise, yielded actionable insights that could make the streets safer.

# The Curse of Dimensionality

---

There is no magical potion to cure the curse, but there is PCA.

---

The “curse of dimensionality” is one of the most important results in machine learning. Most texts on machine learning mention this phenomenon in the first chapter or two, but it often takes many years of practice to understand its true implications.

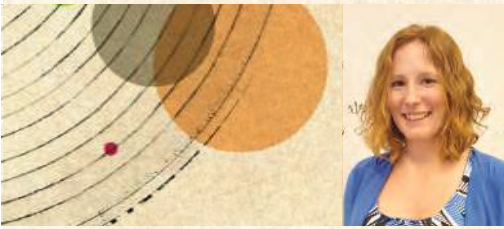
Classification methods, like most machine learning methods, are subject to the implications of the curse of dimensionality. The basic intuition in this case is that as the number of data dimensions increases, it becomes more difficult to create generalizable classification models (models that apply well over phenomena not observed in the training set). This difficulty is usually impossible to overcome in real world settings. There are some exceptions in domains where things happen to work out, but usually you must work to minimize the number of dimensions. This requires a combination of clever feature engineering and use of dimensionality reduction techniques (see *Feature Engineering* and *Feature Selection Life in the Trenches*). In our practical experience, the maximum number of dimensions seems to be ~10 for linear model-based approaches. The limit seems to be in the tens of thousands for more sophisticated methods such as support vector machines, but the limit still exists nonetheless.

A counterintuitive consequence of the curse of dimensionality is that it limits the amount of data needed to train a classification model. There are roughly two reasons for this phenomenon. In one case, the dimensionality is small enough that the model can be trained on a single machine. In the other case, the exponentially expanding complexity of a high-dimensionality problem makes it (practically) computationally impossible to train a model. In our experience, it is quite rare for a problem to fall in a “sweet spot” between these two extremes.

Rather than trying to create super-scalable algorithm implementations, focus your attention on solving your immediate problems with basic methods. Wait until you encounter a problem where an algorithm fails to converge or provides poor cross-validated results, and then seek new approaches. Only when you find that alternate approaches don't already exist, should you begin building new implementations. The expected cost of this work pattern is lower than over-engineering right out of the gate.

Put otherwise, “Keep it simple, stupid”.

## »» Baking the Cake



Stephanie  
Rivera

I was once given a time series set of roughly 1,600 predictor variables and 16 target variables and asked to implement a number of modeling techniques to predict the target variable values. The client was challenged to handle the complexity associated with the large number of variables and needed help. Not only did I have a case of the curse, but the predictor variables were also quite diverse. At first glance, it looked like trying to bake a cake with everything in the cupboard. That is not a good way to bake or to make predictions!

The data diversity could be partially explained by the fact that the time series predictors did not all have the same periodicity. The target time series were all daily values whereas the predictors were daily, weekly, quarterly, and monthly. This was tricky to sort out, given that imputing zeros isn't likely to produce good results. For this specific reason, I chose to use neural networks for evaluating the weekly variable contributions.

Using this approach, I was able to condition upon week, without greatly increasing the dimensionality. For the other predictors, I used a variety of techniques, including projection and correlation, to make heads or tails of the predictors. My approach successfully reduced the number of variables, accomplishing the client's goal of making the problem space tractable. As a result, the cake turned out just fine.

# Model Validation

---

Repeating what you just heard does not mean that you learned anything.

---

Model validation is central to construction of any model. This answers the question “How well did my hypothesis fit the observed data?” If we do not have enough data, our models cannot connect the dots. On the other hand, given too much data the model cannot think outside of the box. The model learns specific details about the training data that do not generalize to the population. This is the problem of model over fitting.

Many techniques exist to combat model over fitting. The simplest method is to split your dataset into training, testing and validation sets. The training data is used to construct the model. The model constructed with the training data is then evaluated with the testing data. The performance of the model against the testing set is used to further reduce model error. This indirectly includes the testing data within model construction, helping to reduce model over fit. Finally, the model is evaluated on the validation data to assess how well the model generalizes.

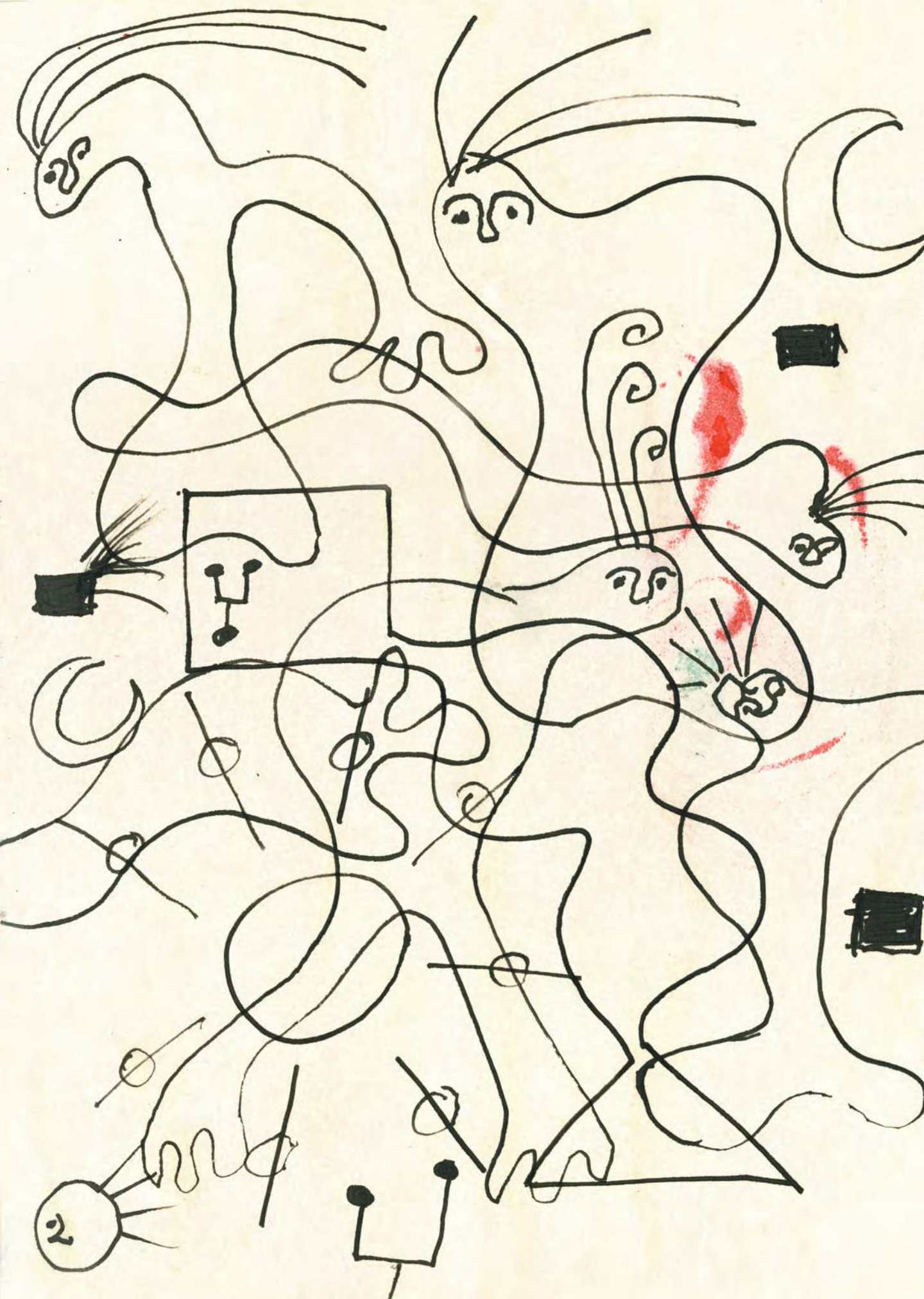
A few methods where the data is split into training and testing sets include:  $k$ -fold cross-validation, Leave-One-Out cross-validation, bootstrap methods, and resampling methods. Leave-One-Out cross-validation can be used to get a sense of ideal model performance over the training set. A sample is selected from the data to act as the testing sample and the model is trained on the rest of the data. The error on the test sample is calculated and saved, and the sample is returned to the dataset. A different sample is then selected and the process is repeated. This continues until all samples in the testing set have been used. The average error over the testing examples gives a measure of the model’s error.



»Do we really need a case study to know that you should check your work?

---

There are other approaches for testing how well your hypothesis reflects the data. Statistical methods such as calculating the coefficient of determination, commonly called the  $R$ -squared value are used to identify how much variation in the data your model explains. Note that as the dimensionality of your feature space grows, the  $R$ -squared value also grows. An adjusted  $R$ -squared value compensates for this phenomenon by including a penalty for model complexity. When testing the significance of the regression as a whole, the  $F$ -test compares the explained variance to unexplained variance. A regression result with a high  $F$ -statistic and an adjusted  $R$ -squared over 0.7 is almost surely significant.





PUTTING *it* **ALL TOGETHER**



# » Streamlining Medication Review

## Analytic Challenge

---

The U.S. Food and Drug Administration (FDA) is responsible for advancing public health by supporting the delivery of new treatments to patients; assessing the safety, efficacy and quality of regulated products; and conducting research to drive medical innovation. Although the FDA houses one of the world's largest repositories of regulatory and scientific data, reviewers are not able to easily leverage data-driven approaches and analytics methods to extract information, detect signals and uncover trends to enhance regulatory decision-making and protect public health. In addition, a rapid increase in the volume, velocity and variety of data that must be analyzed to address and respond to regulatory challenges, combined with variances in data standards, formats, and quality, severely limit the ability of FDA Center for Drug Evaluation and Research (CDER) regulatory scientists to conduct cross-study, cross-product, retrospective, and meta-analysis during product reviews.

Booz Allen Hamilton was engaged to research, develop, and evaluate emerging informatics tools, methods, and techniques to determine their ability to address regulatory challenges faced by the FDA Center for Drug Evaluation and Research (CDER). The main goal was to enable the CDER community to fully utilize the agency's expansive data resources for efficient and effective drug review through the design and development of informatics capabilities based on Natural Language Processing (NLP), data integration, and data visualization methodologies.



## » Our Case Studies

---

Hey, we have given you a lot of really good technical content. We know that this section has the look and feel of marketing material, but there is still a really good story here. Remember, storytelling comes in many forms and styles, one of which is the marketing version. You should read this chapter for what it is – great information told with a marketing voice.

## Our Approach

---

To support transformational change at CDER, we designed and developed a set of informatics prototypes for the analysis and modeling of complex structured, unstructured, and fragmented datasets. We developed multiple prototypes to enable the evaluation of emerging informatics tools, methods, and techniques and their ability to enable a critical-value driver – e.g., co-locate complex, heterogeneous data to identify patterns and foster development of strategies to protect public health. For example, we implemented NLP algorithms to compare adverse events across

datasets, and geographic visualization capabilities to support the inspection of pharmaceutical manufacturing facilities.

**Product Safety Analytics.** The review, surveillance, and analysis of adverse events throughout the product lifecycle require significant resources. The ability to identify actionable insights that lead to informed decision-making requires significant investment of effort, including the active and passive surveillance of adverse events. To address these challenges, we developed a Product Safety Dashboard that compares adverse events listed in the product label (i.e.,

package inserts) with data from the FDA Adverse Event Reporting System (FAERS). Using NLP, we extracted adverse events from the product label to create a structured table of label data out of unstructured text. This dashboard allows safety evaluators to view whether or not a reported adverse event is already known, without having to access an external data source and read through product labels.

### **Product Quality Analytics.**

To support CDER's mission of reviewing and managing product quality, novel methodologies and tools are needed to improve the efficiency and efficacy

of the product quality-review process. Integration of disparate data sources is the first step in building a comprehensive profile of manufacturers, facilities, and the products associated with individual facilities. To address these challenges, we developed a Facility Inventory Report to show the geographic location of facilities and their associated metadata. This geovisualization tool processes and transforms raw data into a user-friendly visual interface with mapping features to enhance the surveillance capabilities of CDER and provide reviewers with the ability to establish connections between facility data and product quality.

## **Our Impact**

---

Since the FDA is responsible for regulating 25 cents of every dollar that Americans spend, the agency's ability to fully use regulatory datasets and meaningfully integrate previously incompatible data to rapidly detect product quality and safety issues is critical for safeguarding public health. NLP approaches provide CDER with the ability to more efficiently search a broader range of textual data and enhances the ability to gain insight from additional data forms that may seem unrelated. Data integration and visualization directly increase the efficiency of researchers by reducing their time spent on searching for frequently-performed aggregate or granular calculations, and by proactively presenting the most frequently desired data to the reviewer through thoughtful and contextual dashboards designed to reveal patterns and trends in disparate data sources. These new capabilities position the FDA to enhance regulatory decision-making, drive advances in personalized medicine, and enable earlier detection of safety signals in the general population.

# »» Reducing Flight Delays

## Analytic Challenge

---

Domestic airline departure delays are estimated to cost the U.S. economy \$32.9 billion annually. The Federal Aviation Administration's (FAA's) Traffic Flow Management System (TFMS) is used to strategically manage flights and includes a flight departure-delay prediction engine which applies simple heuristics to predict flight delays. However, the limited predictive power of these heuristics constrains the FAA's ability to act in accordance with its existing departure-delay management plan. In response, the FAA's NextGen Advanced Concepts and Technology Development Group wanted to create a predictive probabilistic model to improve aircraft departure time predictions. This new model would help the FAA understand the causes of departure delays and develop policies and actions to improve the reliability of departure time predictions for real-time air traffic flow management.

## Our Approach

---

The commercial aviation industry is rich in flight operations data, much of which is publicly available through government websites and a few subscription vendors. Booz Allen Hamilton leveraged these sources to gather over 4 TB of data detailing tarmac and airspace congestion, weather conditions, network effects, Traffic Management Initiatives, and airline and aircraft-specific attributes for every commercial flight departing from U.S. airports between 2008 and 2012. This data included over 50 million flights and around 100 variables for each flight. The data included composite variables (e.g. incoming flight delay) that were constructed from the raw data to capture relevant dynamics of flight operations. Data acquisition, processing, quality control, and accuracy between disparate datasets were important steps during this process. The team applied supervised learning algorithms to develop Bayesian Belief Network (BBN) models to predict flight

departure deviation. The most critical steps in model development were the selection of optimal algorithms to discretize model variables, and the selection of appropriate machine learning techniques to learn the model from the data. The team followed information theory principles to discretize model variables to maximize the model's predictive power, and to represent the data as closely as possible with the least amount of network complexity. Booz Allen segmented the model variables into three different categories based on the time to flight departure: 24 hours, 11 hours, and one hour. Certain flight variables could only be known for specific pre-departure times. For example, the tarmac and airspace congestion variables for a flight are only known just before the flight, and hence those variables feature only in the one hour category. Departure delays were predicted for each of the three time horizons.

## Our Impact

---

For a typical airport, the model delivers a delay prediction improvement of between 100,000 and 500,000 minutes annually over previous FAA predictions. The model can be used by a range of aviation stakeholders, such as airlines, to better understand and predict network flight delays. This can improve the airlines' operational decisions to include more proactive schedule adjustments during times of disruption (e.g. weather or sector load). The improved reliability of departure prediction will improve FAA's predictions for airports, sectors, and other resources, and has the potential to enable improved real-time traffic flow management, which can significantly reduce airline departure, delays, and the associated economic costs. This means a more efficient and effective air transportation network.

# »» Making Vaccines Safer

## Analytic Challenge

---

The U.S. Food and Drug Administration (FDA) Center for Biologics Evaluation and Research (CBER) is responsible for protecting public health by assuring the safety and efficacy of biologics, including vaccines, blood and blood products. CBER's current surveillance process, which requires resource-intensive manual review by expert Medical Officers, does not scale well to short-term workload variation and limits long-term improvements in review cycle-time. In addition, the large volume of Adverse Event (AE) reports received by the Agency makes it difficult for reviewers to compare safety issues across products and patient populations.

CBER engaged Booz Allen Hamilton to develop advanced analytics approaches for the triage and analysis of AE reports. The main goal was to leverage (1) Natural Language Processing (NLP) to alleviate resource pressures by semi-automating some of the manual review steps through techniques, such as text classification and entity extraction, and (2) network visualizations to offer alternative interactions with datasets and support AE pattern recognition. By integrating NLP and network analysis capabilities into the Medical Officer's review process, Booz Allen successfully provided decision-makers with important information concerning product risks and possible mitigations that can reduce risk.

## Our Approach

---

We designed and developed a set of prototypes for the analysis and visualization of complex structured and unstructured AE datasets. We developed tools that leverage NLP and network analysis to extend and enhance CBER's efforts to monitor the safety of biologics and manage safety throughout the product lifecycle.

### **Adverse Event Text Mining Analytics.**

Reviewing AE reports involves sorting based on the likelihood of a relationship between a product and reported adverse events. Since much of an AE report is unstructured text, and most reports are not directly related to the use of the implicated biologic, manual review is time consuming and inefficient. To address these challenges, we enhanced and integrated tools for text mining and NLP of AE reports using open source tools, including Python and R. By

extracting diagnosis, cause of death, and time to onset, and presenting relevant information for review, the text-mining tool streamlines and enhances the post market surveillance process.

### **Adverse Event Network Analytics.**

Visualizing relationships between vaccines and AEs can reveal new patterns and trends in the data, leading reviewers to uncover safety issues. To assist CBER Medical Officers and researchers in identifying instances where certain vaccines or combinations of vaccines might have harmful effects on patients, we developed an AE network analysis tool using open source tools. This network analyzer allows users to select partitions of the FDA Vaccine Adverse Event Reporting System (VAERS) database, generate a co-occurrence matrix, view networks and network metrics (e.g., betweenness,

closeness, degree, strength), and interact with network nodes to gain insights into product safety issues.

**Other Analytics Solutions.** In addition, Booz Allen refactored, modularized, and expended the capabilities of CBER's computer simulation model of the Bovine Spongiform Encephalopathy (BSE) agent to improve estimates of variant Creutzfeldt-Jakob disease (vCJD) risk for blood products, developed code to

handle large amounts of data generated by a Monte Carlo Markov Chain analysis of the spread of influenza, developed a large database analysis strategy involving the application of classification algorithms to simulated genomic data, and implemented a Statistical Analysis Software (SAS) macro that automatically compares the relative potency of a given lot of vaccine using a matched set of dose response curves.

## Our Impact

---

New methods for post market surveillance of biologics are critical for FDA reviewers who must determine whether reported adverse events are actually a result of a biologic product. With more than 10 million vaccines administered each year to children less than one-year old, CBER reviewers are under pressure to quickly evaluate potential safety signals through manual evaluation of AE reports, review of scientific literature, and analysis of cumulative data using frequency calculations or statistical algorithms. Booz Allen's support resulted in the development of innovative and data-driven approaches for the analysis of structured and unstructured AE reports. We increased the speed of existing text mining tools by two thousand times, allowing CBER reviewers to run a text mining algorithm to extract information contained in VAERS reports in seconds, instead of hours. We also increased the productivity of Medical Officers through the implementation of text mining and network analysis tools. These new capabilities allow CBER to streamline the post market review process, extract knowledge from scientific data, and address public concerns regarding vaccine safety more quickly and efficiently.

# »» Forecasting the Relative Risk for the Onset of Mass Killings to Help Prevent Future Atrocities

## Analytic Challenge

---

Mass atrocities are rare yet devastating crimes. They are also preventable. Studies of past atrocities show that we can detect early warning signs of atrocities and that if policy makers act on those warnings and develop preventive strategies, we can save lives. Yet despite this awareness, all too often we see warning signs missed and action taken too late, if at all, in response to threats of mass atrocities.

The Early Warning Project, an initiative of the United States Holocaust Memorial Museum (Holocaust Museum), aims to assess a country's level of risk for the onset of future mass killings. Over time, the hope is to learn which models and which indicators are the best at helping anticipate future atrocities to aid in the design and implementation of more targeted and effective preventive strategies. By seeking to understand why and how each country's relative level of risk rises and falls over time, the system will deepen understanding of where new policies and resources can help make a difference in averting atrocities and what strategies are most effective. This will arm governments, advocacy groups, and at-risk societies with earlier and more reliable warning, and thus more opportunity to take action, well before mass killings occur.

The project's statistical risk assessment seeks to build statistical and machine learning algorithms to predict the onset of a mass killing in the succeeding 12 months for each country with a population larger than 500,000. The publically available system aggregates and provides access to open source datasets as well as democratizes the source code for analytic approaches developed by the Holocaust Museum staff and consultants, the research community, and the general public. The Holocaust Museum engaged Booz Allen to validate existing approaches as well as explore new and innovative approaches for the statistical risk assessment.

## Our Approach

---

Taking into account the power of crowdsourcing, Booz Allen put out a call to employees to participate in a hack-a-thon—just the start of the team's support as the Museum refined and implemented the recommendations. More than 80 Booz Allen Hamilton software engineers, data analysts, and social scientists devoted a Saturday to participate. Interdisciplinary teams spent 12 hours

identifying new datasets, building new machine learning models, and creating frameworks for ensemble modeling and interactive results visualization. Following the hack-a-thon, Booz Allen Data Scientists worked with Holocaust Museum staff to create a data management framework to automate the download, aggregation, and transformation of the open

source datasets used by the statistical assessment. This extensible framework allows integration of new datasets with minimal effort, thereby

supporting greater engagement by the Data Science community.

## Our Impact

---

Publically launched in the fall of 2015, the Early Warning Project can now leverage advanced quantitative and qualitative analyses to provide governments, advocacy groups and at-risk societies with assessments regarding the potential for mass atrocities around the world. Integration of the project's statistical risk assessment models and expert opinion pool created a publicly available source of invaluable information and positioned Data Science at the center of global diplomacy.

The machine learning models developed during the hack-a-thon achieved performance on par with state of the art approaches as well as demonstrated the efficacy of predictions 2-5 years into the future. Teams also identified approaches for constructing test/validation sets that support more robust model evaluation. These risk assessments are an important technological achievement in and of themselves, but what this initiative means for the Data Science community's position in global diplomatic dialogue marks an entirely new era for those on the frontiers of Big Data.

The data management framework developed from the lessons learned of the hack-a-thon represents a great leap forward for the Holocaust Museum. The periodicity of aggregating and transforming data was reduced from twice per year to once per week. In addition to providing the community with more up-to-date data, the reduced burden on researchers enables them to spend more time analyzing data and identifying new and emergent trends. The extensible framework will also allow the Holocaust Museum to seamlessly integrate new datasets as they become available or are identified by the community as holding analytic value for the problem at hand.

Through this project, the Holocaust Museum was able to shift the dynamic from monitoring ongoing violence to determining where it is likely to occur 12 to 24 months into the future by integrating advanced quantitative and qualitative analyses to assess the potential for mass atrocities around the world. The Early Warning Project is an invaluable predictive resource supporting the global diplomatic dialogue. While the focus of this effort was on the machine learning and data management technologies behind the initiative, it demonstrates the growing role the Data Science community is playing at the center of global diplomatic discussions.



# » Predicting Customer Response

## Analytic Challenge

---

It is very challenging to know how a customer will respond to a given promotional campaign. Together with the InterContinental Hotels Group (IHG), Booz Allen Hamilton explored approaches to predict customer-by-customer response to a state-of-the-art promotional campaign in order to better understand and increase return on investment (ROI).

In the past, conventional statistics have worked well for analyzing the impact of direct marketing promotions on purchase behavior. Today, modern multi-channel promotions often result in datasets that are highly dimensional and sometimes sparse, which strains the power of conventional statistical methods to accurately estimate the effect of a promotion on individual purchase decisions. Because of the growing frequency of multi-channel promotions, IHG was driven to investigate new approaches. In particular, IHG and Booz Allen studied one recent promotional campaign using hotel, stay, and guest data for a group of loyalty program customers.

## Our Approach

---

Working closely with IHG experts, Booz Allen investigated three key elements related to different stages of analytic maturity:

**Describe:** Using initial data mining, what insights or tendencies in guest behaviors can be identified after joining multiple, disparate datasets?

**Discover:** Can we determine which control group members would be likely to register for a promotion if offered? If so, can we also quantify their registration?

**Predict:** How would a hotel guest that received the promotion have responded if they were not offered the promotion? How would a hotel guest that did not receive the promotion have responded if they were offered the promotion?

For the promotion that was the focus of this case study, not everything about customers could be controlled as required by traditional statistics. However, because

a probabilistic Bayesian Belief Network (BBN) can learn the pairwise relationships between all individual customer attributes and their impact on promotional return, Booz Allen investigated how this technique could be used to model each treated customer without an exact controlled look-alike.

Specifically, Booz Allen developed a BBN to predict customer-by-customer impacts driven by promotional campaign offers, subsequently estimating the aggregated ROI of individual campaigns. We used six machine learning techniques (support vector machine, random forest, decision tree, neural network, linear model, and AdaBoost) in unison with the BBN to predict how each customer would be influenced by a promotional offer.

## Our Impact

---

The probabilistic model was capable of predicting customer response to the promotion, without relying on a costly control group. This finding represents millions of dollars of savings per promotional campaign. This analysis is an industry-first for the travel and hospitality sector, where demand for more data-driven approaches to optimize marketing ROI at an individual level is rapidly growing.

Because Booz Allen and IHG's approach enabled estimation of ROI for each hypothetical customer, even when no exact look-alikes exist, there are a number of valuable future applications. One such application is optimal campaign design—the ability to estimate the promotional attributes for an individual customer that are likely to drive the greatest incremental spend. Another application is efficient audience selection - which would reduce the risk of marketing "spam" that prompts costly unsubscriptions and can negatively impact a hotel's brand.

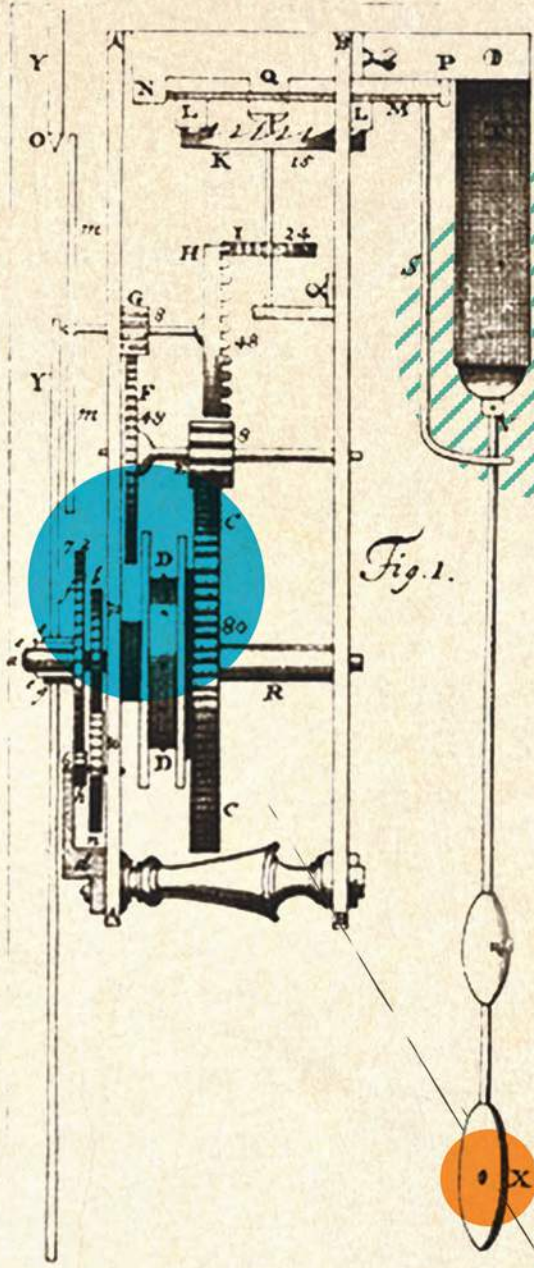
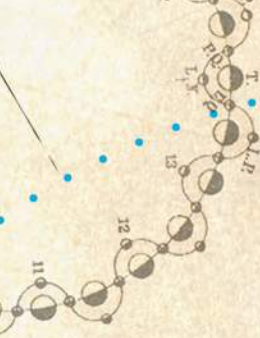


Fig. 1.

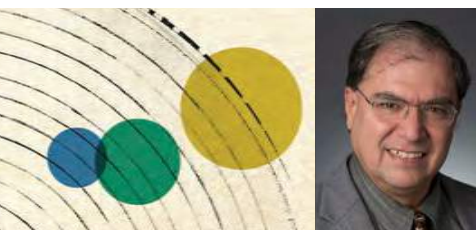




# CLOSING TIME



# »» THE FUTURE of DATA SCIENCE



Kirk Borne

**Data Science is rapidly evolving as it touches every aspect of our lives on a daily basis.** As Data Science changes the way we interact with, and explore our world, the algorithms and applications of Data Science continue to advance. We expect this trend to continue as Data Science has an increasingly profound effect on humanity. We describe here some of the trends and developments we anticipate emerging in the field of Data Science over the coming years.

The advancements in some Data Science algorithms will deliberately track the evolution of data structures and data models that Data Scientists are using to represent their domains of study. One of the clearest examples of this linkage is in the development of massive-scale graph analytics algorithms, which are deployed on graph databases (*including network data and semantically linked databases*). It is sometimes said “all the world is a graph,” and consequently the most natural data structure is not a table with rows and columns, but a network graph with nodes and edges. Graph analytics encompasses traditional methods of machine learning, but with a graph-data twist.

Another growth area in Data Science algorithms is in the domain of geospatial temporal predictive analytics, which can be applied to any dataset that involves geospatial location and time – that describes just about everything in our lives! We expect increasingly sophisticated deployments of this methodology in the areas of law enforcement, climate change, disaster management, population health, sociopolitical change, and more.

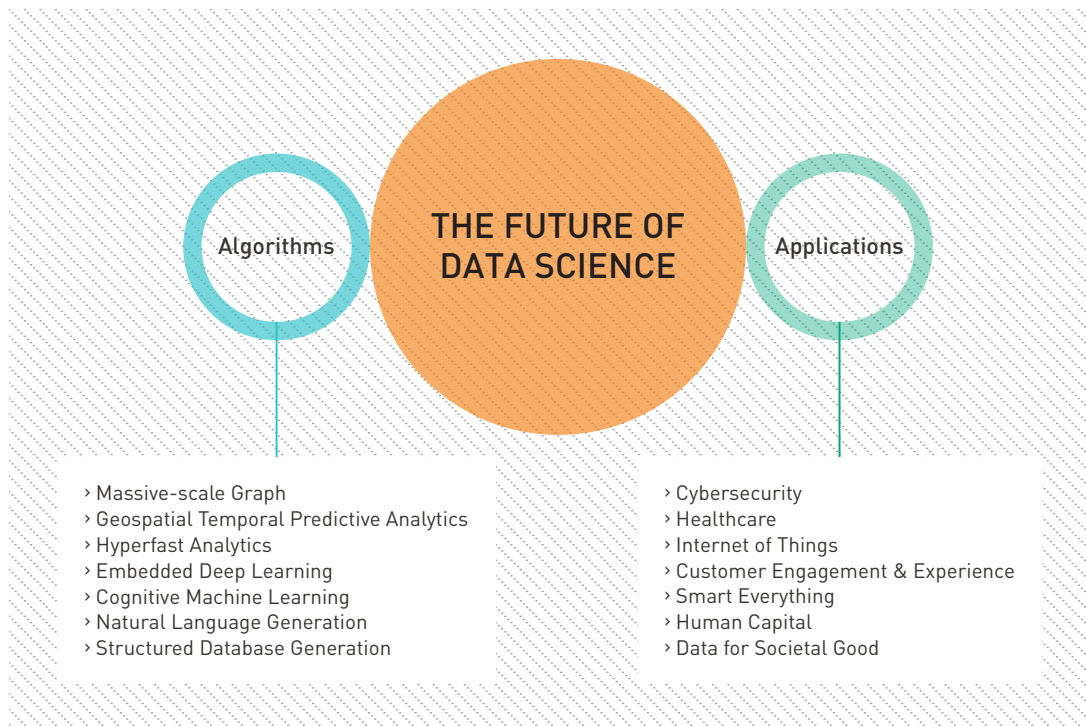
It is obvious that bigger, faster, and more complex datasets will require faster (*hyperfast!*) analytics. We anticipate advanced Data Science algorithms that take advantage of technological advancements in quantum machine-learning, in-memory data operations, and machine learning on specialized devices (*e.g., the GPU, Raspberry Pi, or the next-generation mobile handheld “supercomputer”*). In such commodity devices, we expect to see development of more embedded machine learning (*specifically, deep learning*) algorithms that perform time-critical data-to-insights transformations at the point of data collection. Such use cases will be in great abundance within the emerging Internet of Things (*IoT*), including the industrial IoT and the internet of everything.

Advances in cognitive machine learning are on the horizon, including open source and configurable algorithms that exploit streaming real-time data’s full content, context, and semantic meaning. The ability to use the 360-degree view of a situation will enable the delivery of the right action, at the right time, at the right place, in the right context – this is the essence of cognitive analytics. Another way to view cognitive analytics is that, given all of the data and the context for a given object or population, the algorithm identifies the right question that you should be asking of your data (*which might not be the question that you traditionally asked*).

Another area of Data Science evolution that tracks with the growth in a particular data type is that of unstructured data, specifically text. The growth of such unstructured data is phenomenal, and demands richer algorithms than those used on numerical data, since there are many more shades of meaning in natural language than in tables of numbers. The new Data Science algorithms for unstructured data will be applied in multiple directions. Natural Language Generation will be used to convert data points into text, which can be used to generate the data's story automatically. Structured Database Generation will transform text documents or other unstructured data into data points (*i.e., converting qualitative data into machine-computable quantitative data*).

All of these technical advancements, plus others that we cannot yet imagine, will be brought to bear on new domains. Some of the hottest, most critical domains in which Data Science will be applied in the coming years include:

- › **Cybersecurity** including advanced detection, modeling, prediction, and prescriptive analytics
- › **Healthcare** including genomics, precision medicine, population health, healthcare delivery, health data sharing and integration, health record mining, and wearable device analytics
- › **IoT** including sensor analytics, smart data, and emergent discovery alerting and response
- › **Customer Engagement and Experience** including 360-degree view, gamification, and just-in-time personalization
- › **Smart X**, where X = cities, highways, cars, delivery systems, supply chain, and more
- › **Precision Y**, where Y = medicine, farming, harvesting, manufacturing, pricing, and more
- › **Personalized Z**, where Z = marketing, advertising, healthcare, learning, and more
- › **Human capital (talent) and organizational analytics**
- › **Societal good**





# »» PARTING THOUGHTS

Data Science capabilities are creating data analytics that are touching every aspect of our lives on a daily basis. From visiting the doctor, to driving our cars, to shopping for services Data Science is quietly changing the way we interactive with and explore our world. We hope we have helped you truly understand the potential of your data and how to become extraordinary thinkers by asking the right questions. We hope we have helped continue to drive forward the science and art of Data Science. Most importantly, we hope you are leaving with a newfound passion and excitement for Data Science.

---

Thank you for taking this journey with us. Please join our conversation and let your voice be heard. Email us your ideas and perspectives at [data\\_science@bah.com](mailto:data_science@bah.com) or submit them via a pull request on the [Github repository](#).

Tell us and the world what you know. Join us. Become an author of this story.

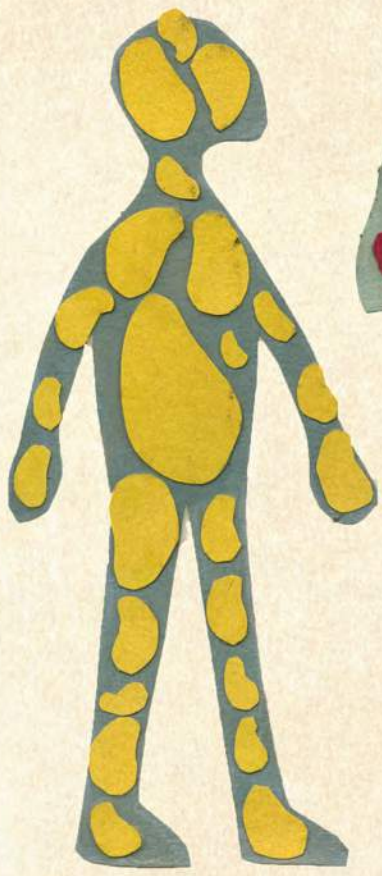
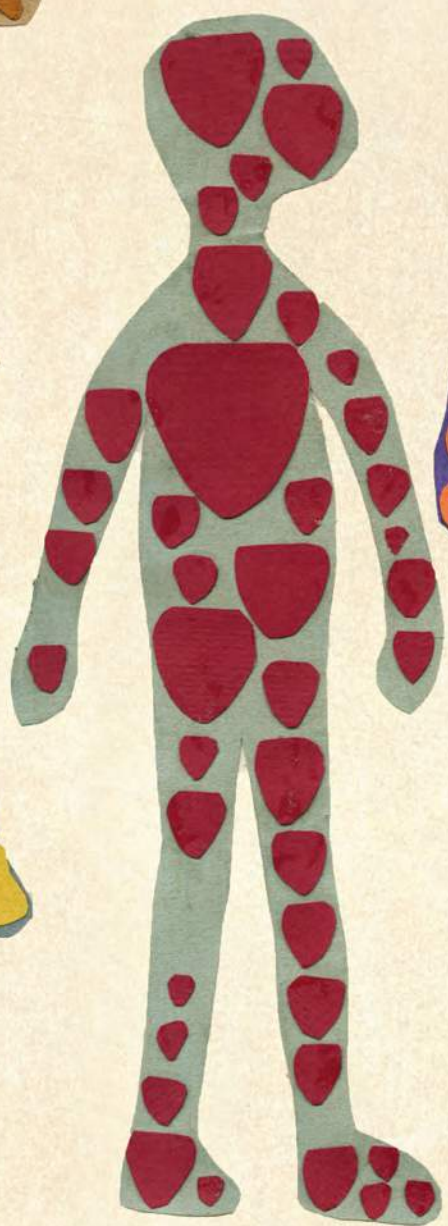
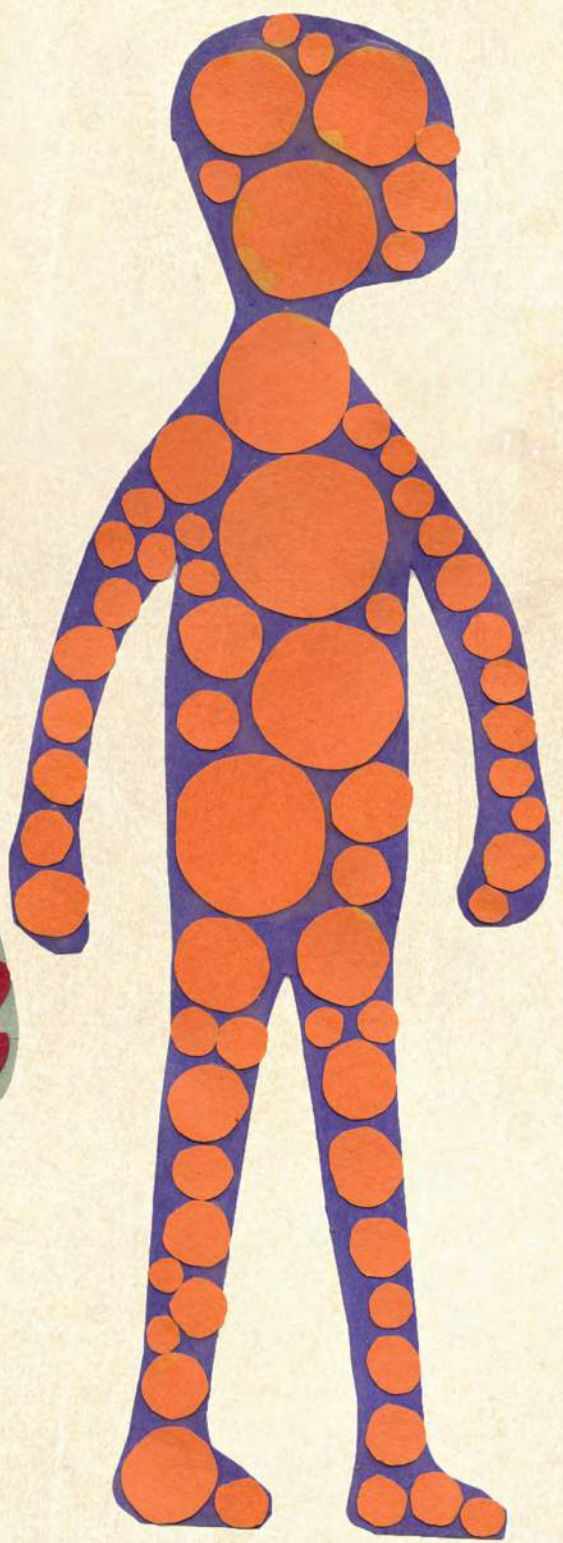
---



## »» REFERENCES

1. Commonly attributed to: Nye, Bill. *Reddit Ask Me Anything (AMA)*. July 2012. Web. Accessed 15 October 2013. SSRN: <[http://www.reddit.com/r/IAMa/comments/x9pq0/iam\\_bill\\_nye\\_the\\_science\\_guy\\_ama](http://www.reddit.com/r/IAMa/comments/x9pq0/iam_bill_nye_the_science_guy_ama)>
2. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine* 17.3 (1996): 37-54. Print.
3. "Mining Data for Nuggets of Knowledge." *Knowledge@Wharton*, 1999. Web. Accessed 16 October 2013. SSRN: <<http://knowledge.wharton.upenn.edu/article/mining-data-for-nuggets-of-knowledge>>
4. Cleveland, William S. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." *International Statistical Review* 69.1 (2001): 21-26. Print.
5. Davenport, Thomas H., and D.J. Patil. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review* 90.10 (October 2012): 70-76. Print.
6. Smith, David. "Statistics vs Data Science vs BI." *Revolutions*, 15 May 2013. Web. Accessed 15 October 2013. SSRN: <<http://blog.revolutionanalytics.com/2013/05/statistics-vs-data-science-vs-bi.html>>
7. Brynjolfsson, Erik, Lorin M. Hitt, and Heekyung H. Kim. "Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance?" *Social Science Electronic Publishing*, 22 April 2011. Web. Accessed 15 October 2013. SSRN: <<http://ssrn.com/abstract=1819486> or <http://dx.doi.org/10.2139/ssrn.1819486>>
8. "The Stages of an Analytic Enterprise." *Nucleus Research*. February 2012. Whitepaper.
9. Barua, Anitesh, Deepa Mani, and Rajiv Mukherjee. "Measuring the Business Impacts of Effective Data." *University of Texas*. Web. Accessed 15 October 2013. SSRNL: <[http://www.sybase.com/files/White\\_Papers/EffectiveDataStudyPt1-MeasuringtheBusinessImpactsofEffectiveData-WP.pdf](http://www.sybase.com/files/White_Papers/EffectiveDataStudyPt1-MeasuringtheBusinessImpactsofEffectiveData-WP.pdf)>
10. Zikopoulos, Paul, Dirk deRoos, Kirshnan Parasuraman, Thomas Deutsch, David Corrigan and James Giles. *Harness the Power of Big Data: The IBM Big Data Platform*. New York: McGraw Hill, 2013. Print. 281pp.

11. Booz Allen Hamilton. *Cloud Analytics Playbook*. 2013. Web. Accessed 15 October 2013. SSRN: <<http://www.boozallen.com/media/file/Cloud-playbook-digital.pdf>>
12. Conway, Drew. "The Data Science Venn Diagram." March 2013. Web. Accessed 15 October 2013. SSRN: <<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>>
13. Booz Allen Hamilton. *Tips for Building a Data Science Capability*. 2015. Web Accessed 2 September 2015. SSRN: <<https://www.boozallen.com/content/dam/boozallen/documents/2015/07/DS-Capability-Handbook.pdf>>
14. Mnih et al. 2015. Human-level control through deep reinforcement learning. *Nature*. 518: 529-533.
15. Torán, Jacobo. "On the Hardness of Graph Isomorphism." *SIAM Journal on Computing*. 33.5 (2004): 1093-1108. Print.
16. Guyon, Isabelle and Andre Elisseeff. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3 (March 2003):1157-1182. Print.
17. Golub T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*. 286.5439 (1999): 531-537. Print.
18. Haykin, Simon O. *Neural Networks and Learning Machines*. New Jersey: Prentice Hall, 2008. Print.
19. De Jong, Kenneth A. *Evolutionary Computation - A Unified Approach*. Massachusetts: MIT Press, 2002. Print.
20. Yacci, Paul, Anne Haake, and Roger Gaborski. "Feature Selection of Microarray Data Using Genetic Algorithms and Artificial Neural Networks." ANNIE 2009. St Louis, MO. 2-4 November 2009. Conference Presentation.

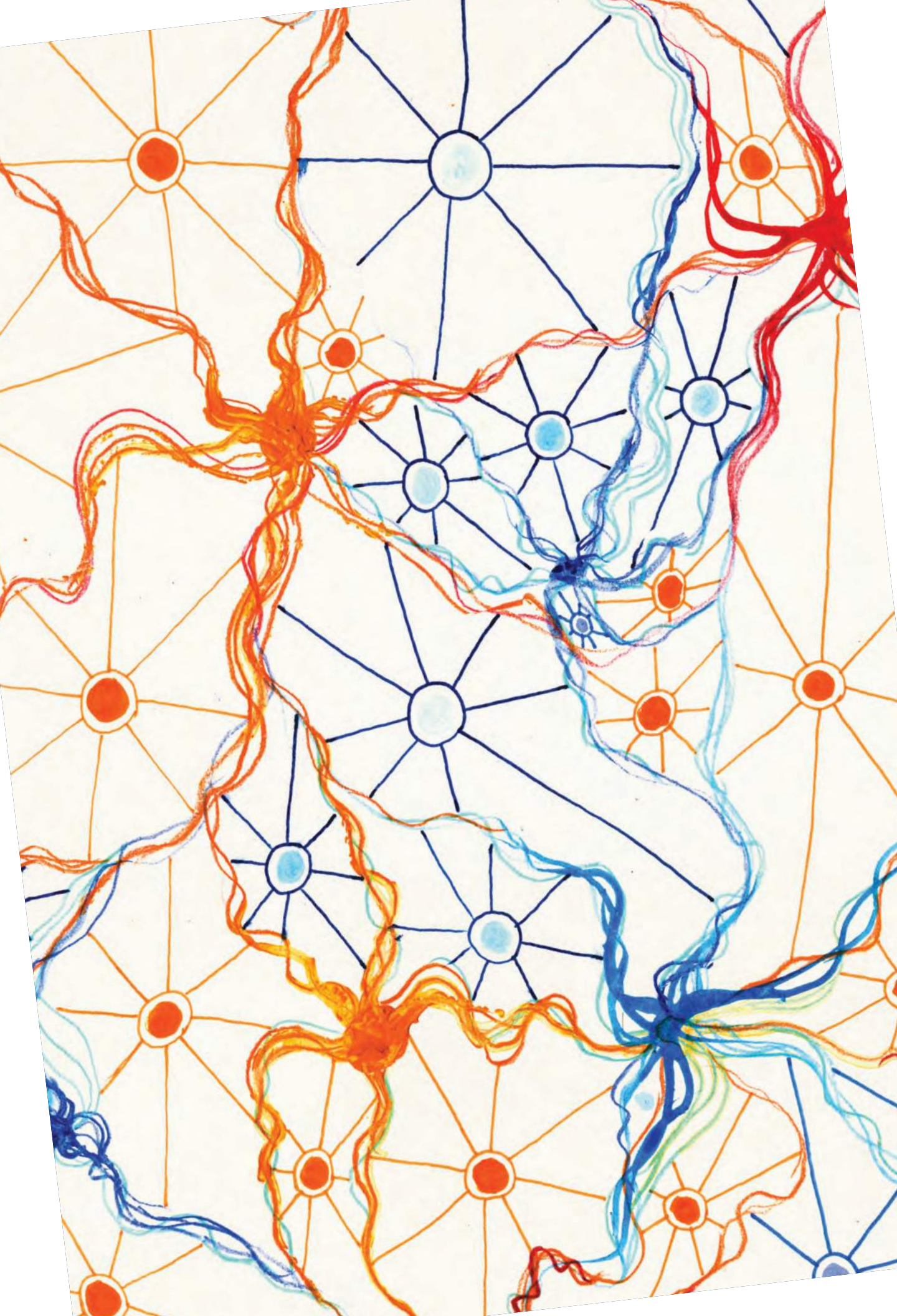


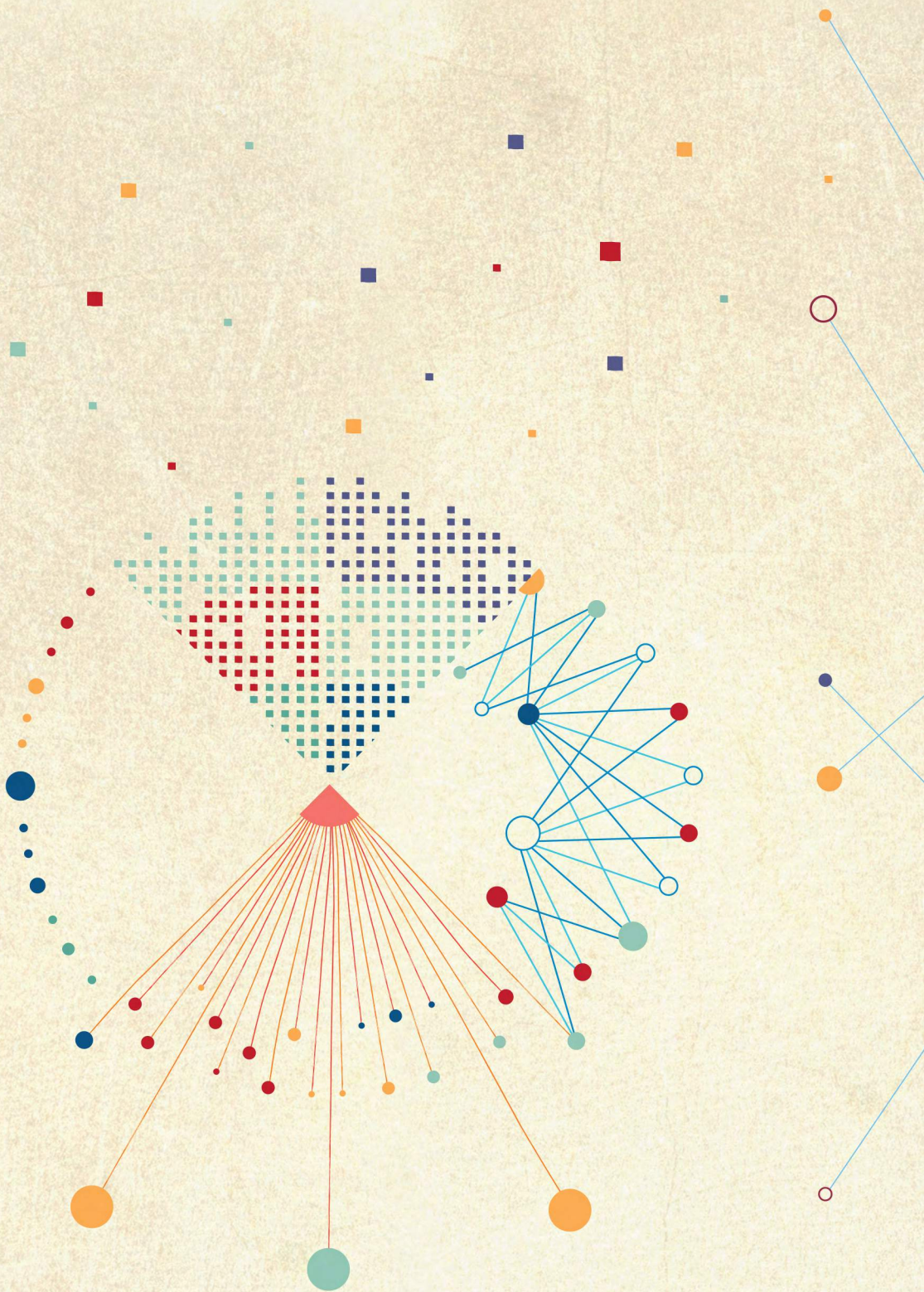
## »» *About*

# BOOZ ALLEN HAMILTON

**Booz Allen Hamilton has been at the forefront of strategy and technology for more than 100 years.** Today, the firm provides management and technology consulting and engineering services to leading Fortune 500 corporations, governments, and not-for-profits across the globe. Booz Allen partners with public and private sector clients to solve their most difficult challenges through a combination of consulting, analytics, mission operations, technology, systems delivery, cybersecurity, engineering, and innovation expertise.

With international headquarters in McLean, Virginia, the firm employs more than 22,500 people globally, and had revenue of \$5.27 billion for the 12 months ended March 31, 2015. To learn more, visit [www.boozallen.com](http://www.boozallen.com). (NYSE: BAH)





© COPYRIGHT 2015 BOOZ ALLEN HAMILTON INC. ALL RIGHTS RESERVED.

Artwork by Rafael Esquer.