Intro to Data Science

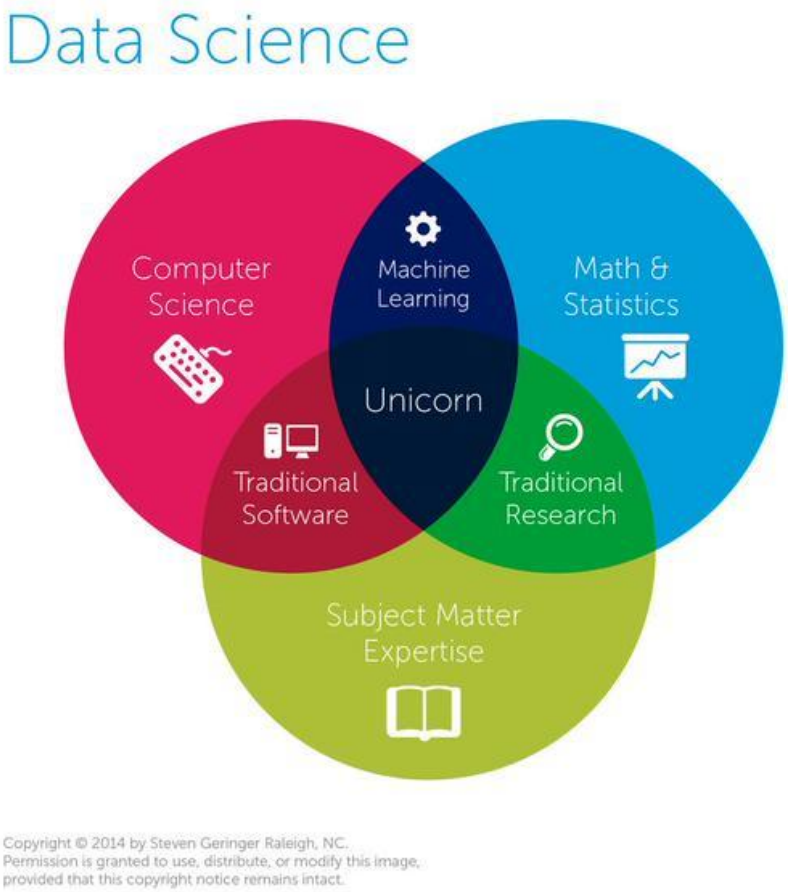
Jonathan Joa
Scott McAllister

Topics

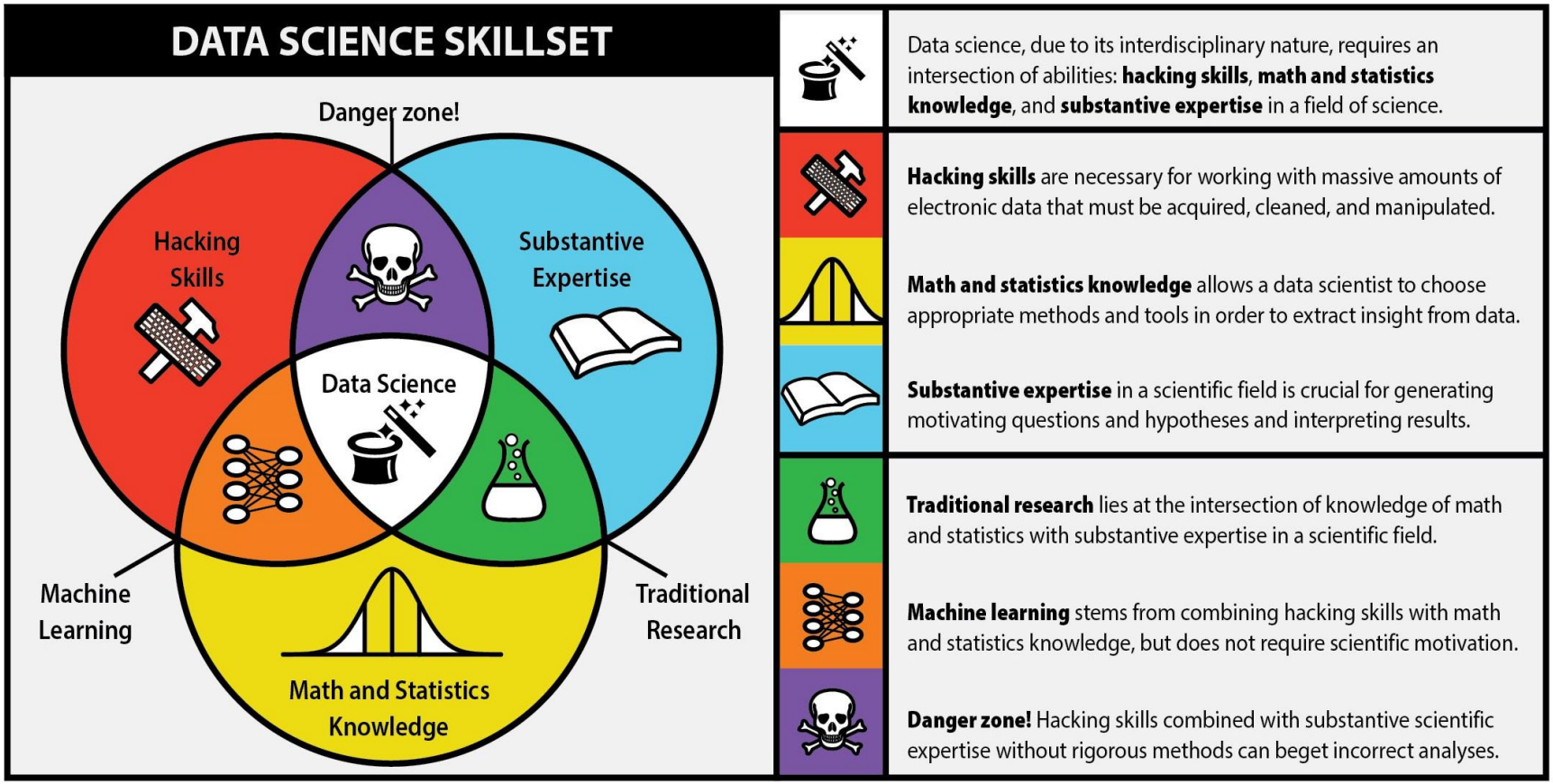
- What is Data Science?
- Why Learn Data Science?
- Netflix Example
- Data Science Roles
- Data Science Project Workflow
- Federal Government Example
- Why Python and R?
- Python vs R
- Setting expectations on what we'll learn

What is Data Science?

- Tools and techniques for data analysis
- Problem-solving
- Applying scientific techniques to practical problems



What is Data Science?



Article

Danger Zone: Spurious Correlations

Why Learn Data Science?

Data Scientist: The Sexiest Job of the 21st Century

(Harvard Business Review)

IBM Predicts Demand For Data Scientists Will Soar 28% By 2020
(Forbes)

50 Best Jobs in America, 2017 (Glassdoor)

The world's most valuable resource is no longer oil, but data

(The Economist)

Who Uses Data Science?

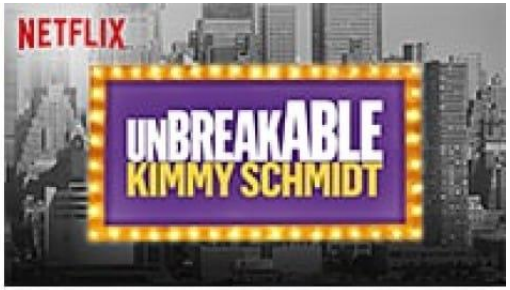
NETFLIX

amazon.com

Google

FiveThirtyEight

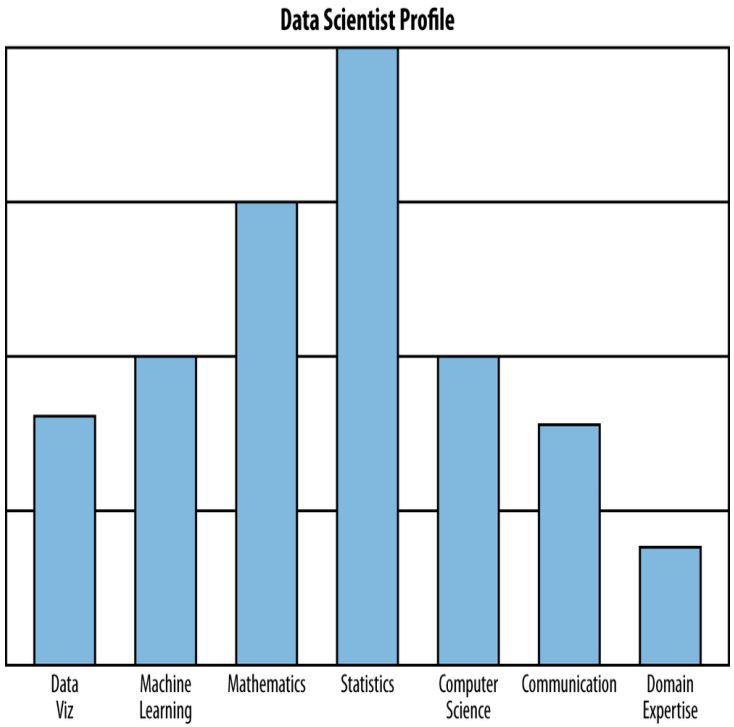
DRAFT KINGS

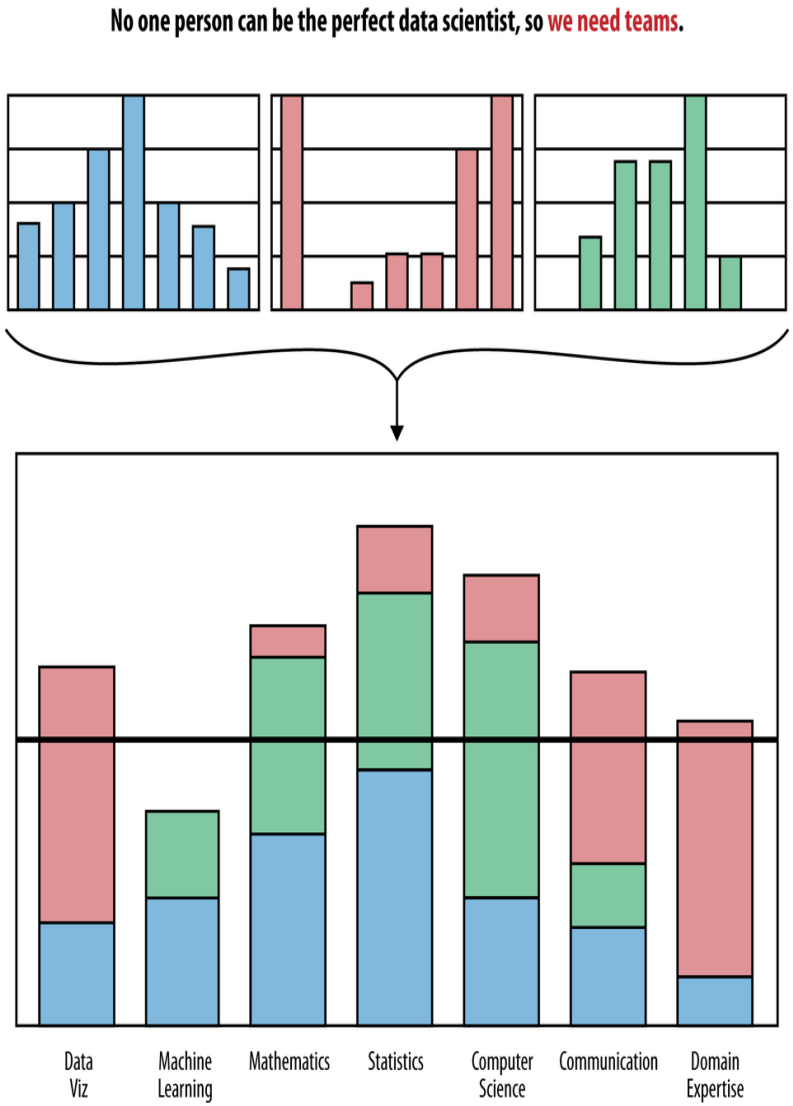


Netflix Example

Article

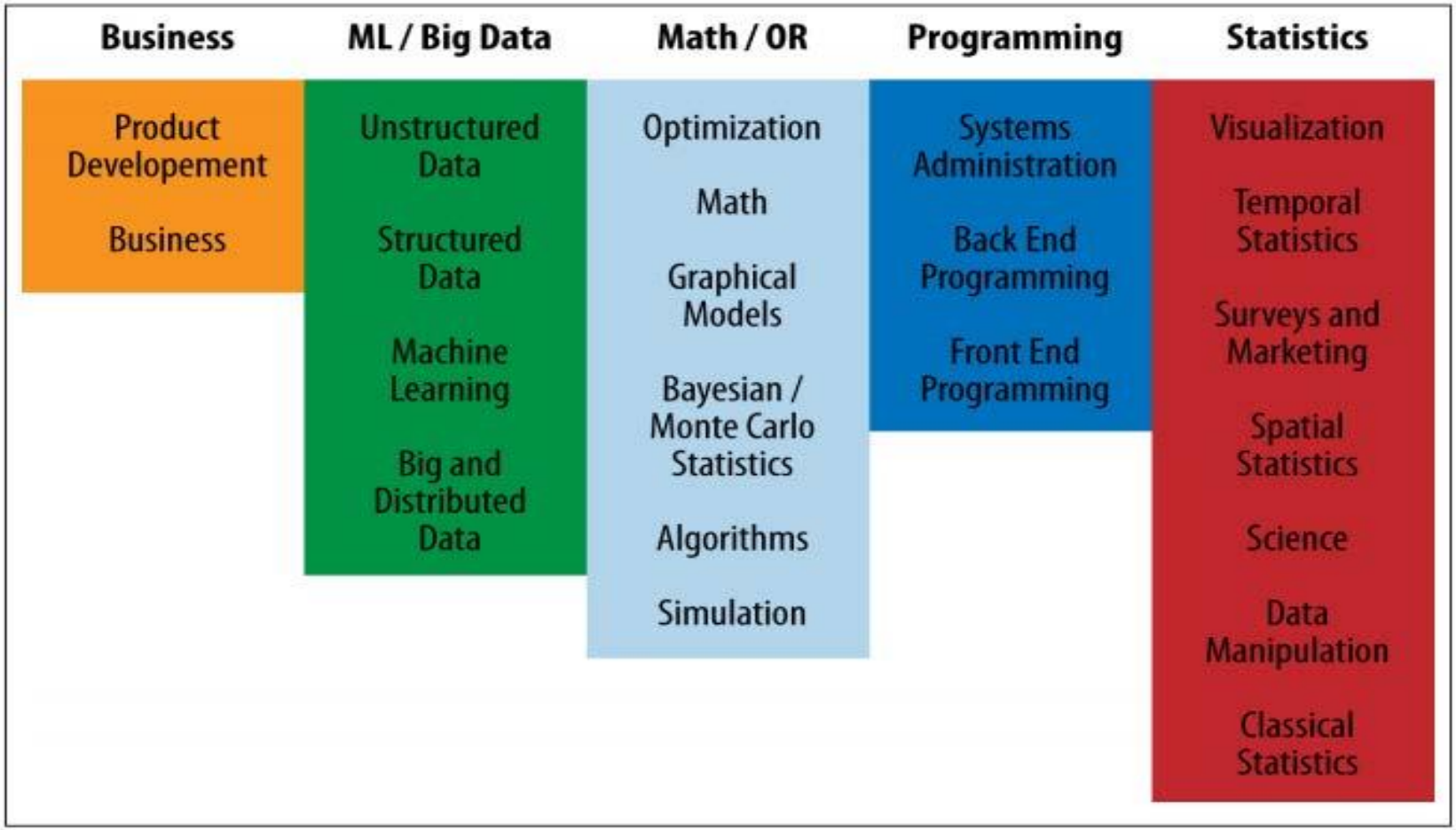
Data Science Roles





Data Science Skills

Involves a variety of skills, not just one

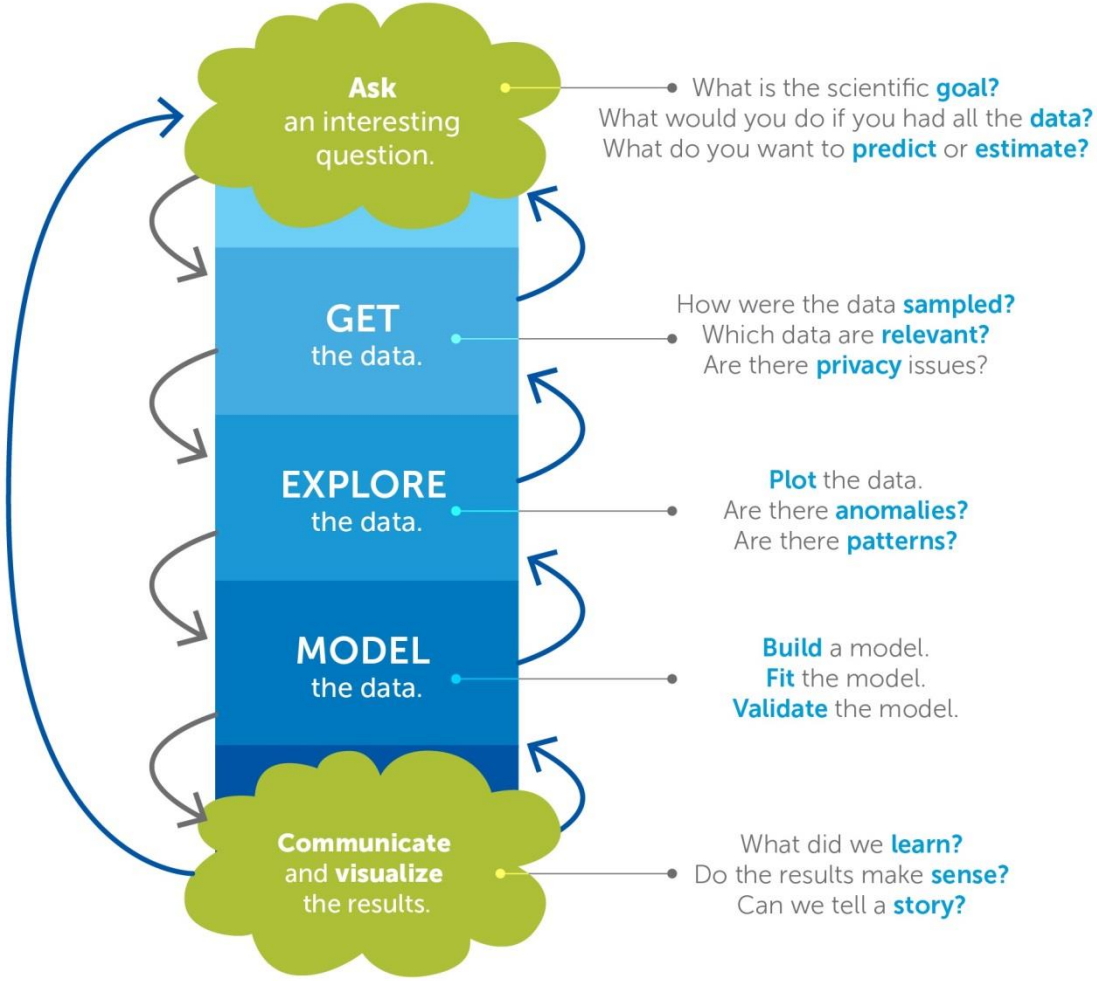


Data Science Roles

Involves a variety of roles, not just one

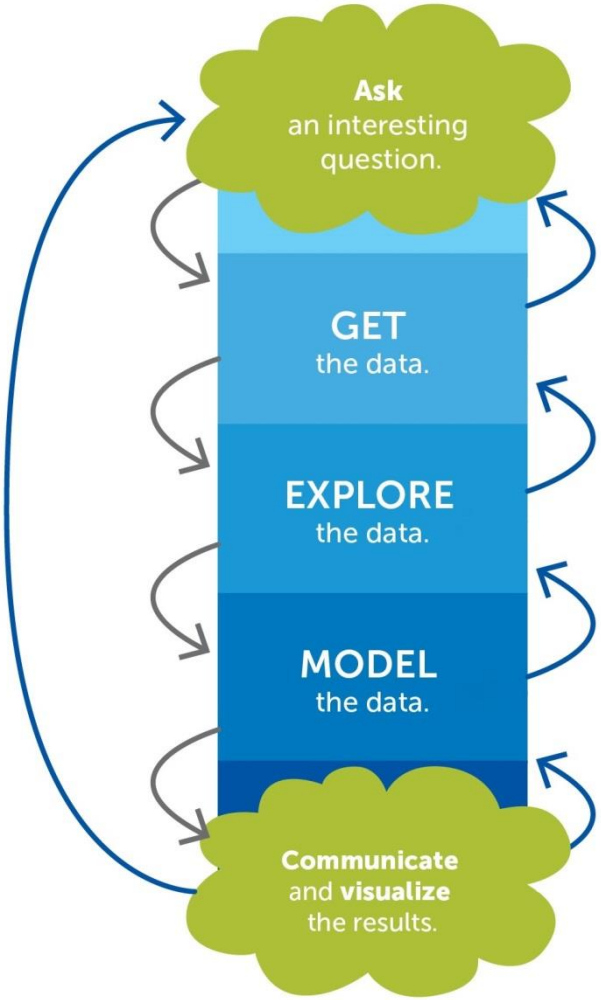
Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

Data Science Workflow



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course http://cs109.org/.

Data Science Workflow



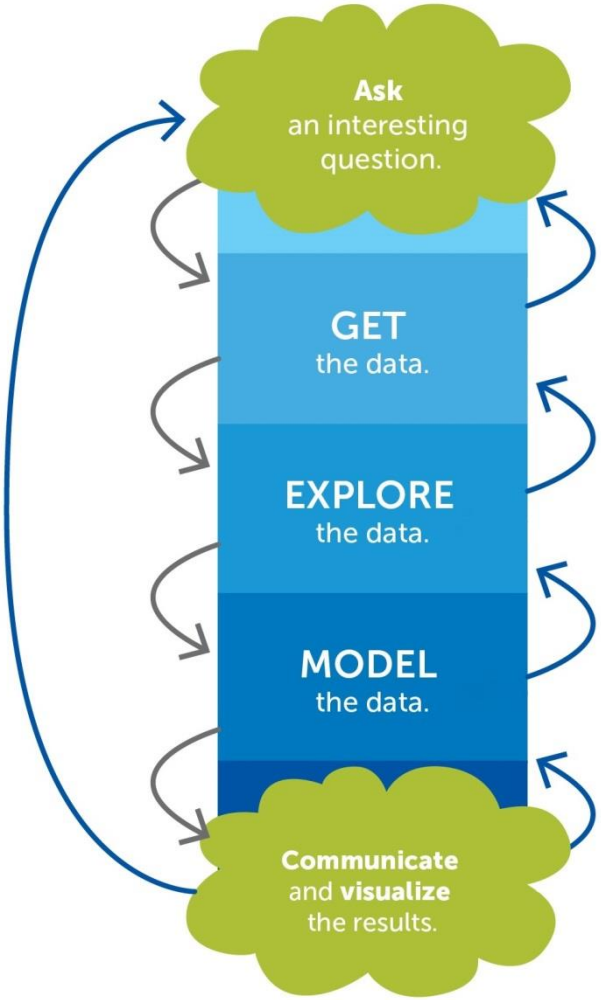
Step 1: Ask an Interesting Question

Identify business/product objectives

Identify and hypothesize goals and criteria for success

Create a set of questions for identifying correct data set

Data Science Workflow



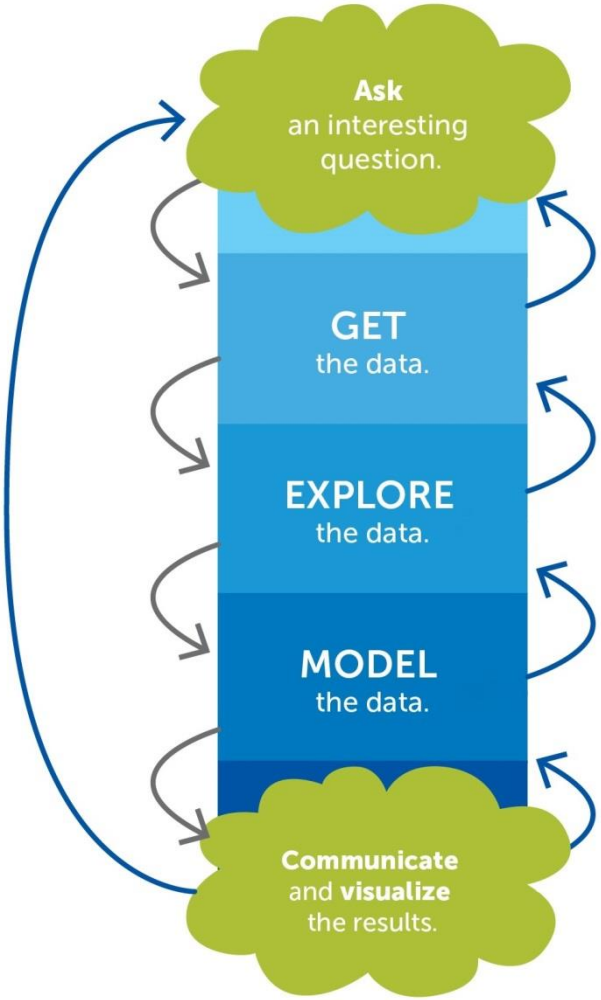
Step 2: Get the Data

Identify the “right” data set(s)

Import data and set up local or remote data structure

Determine most appropriate tools to work with data

Data Science Workflow



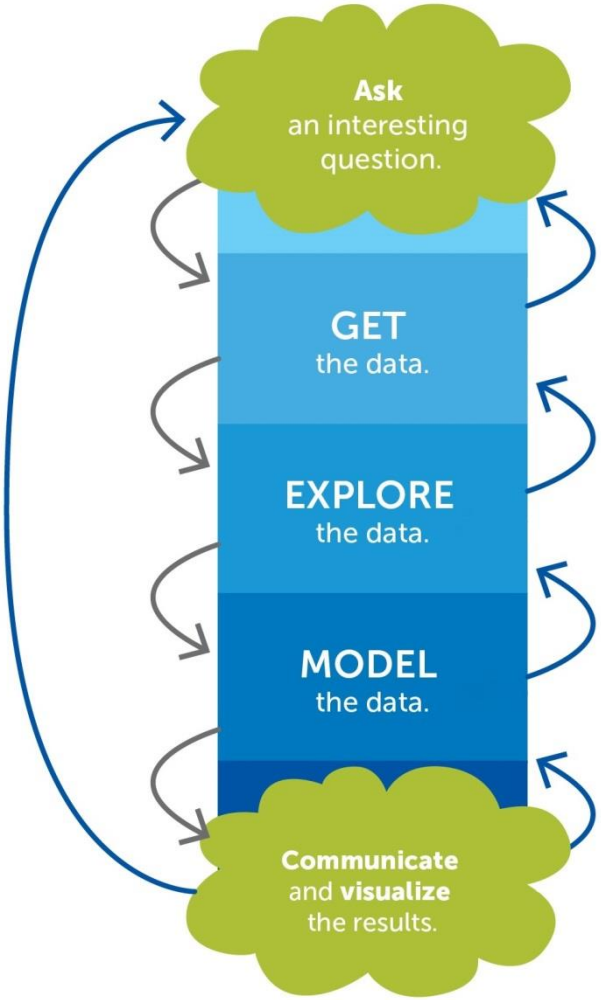
Step 3: Explore the Data

Read documentation provided with the data

Perform exploratory data analysis

Verify the quality of the data

Data Science Workflow



Step 4: Model the Data

Determine sampling methodology and sample data

Format, clean, slice, and combine data

Create necessary derived columns from the data

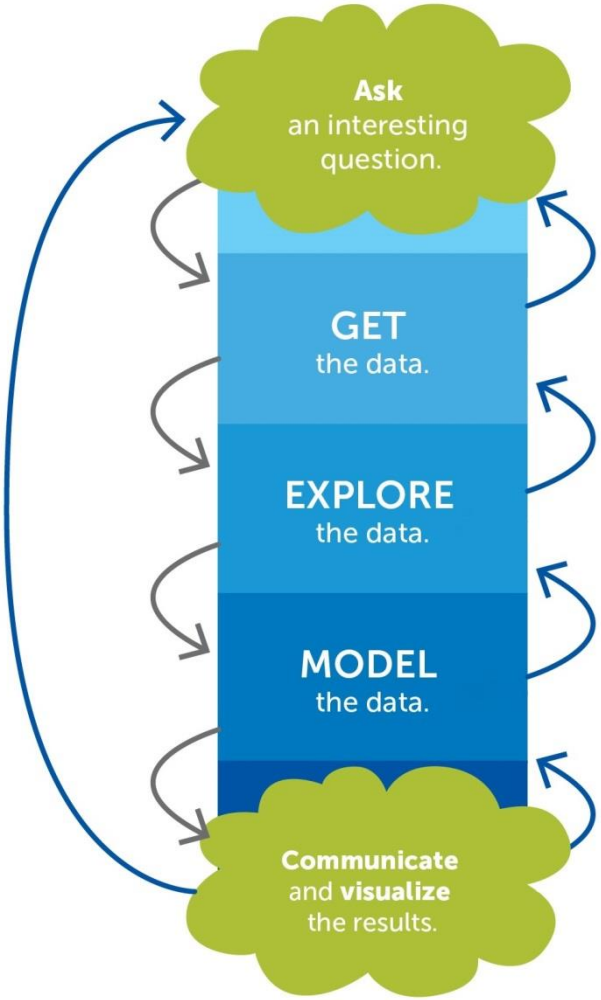
Identify trends and outliers

Document and transform data

Build model

Evaluate and refine model

Data Science Workflow



Step 5: Communicate Results

Summarize findings with narrative, storytelling techniques

Present limitations and assumptions of your analysis

Identify follow up problems and questions for future analysis

Federal Government Example

Office of Human Capital Strategy & Management (OHRM) carried out a study on Performance Management

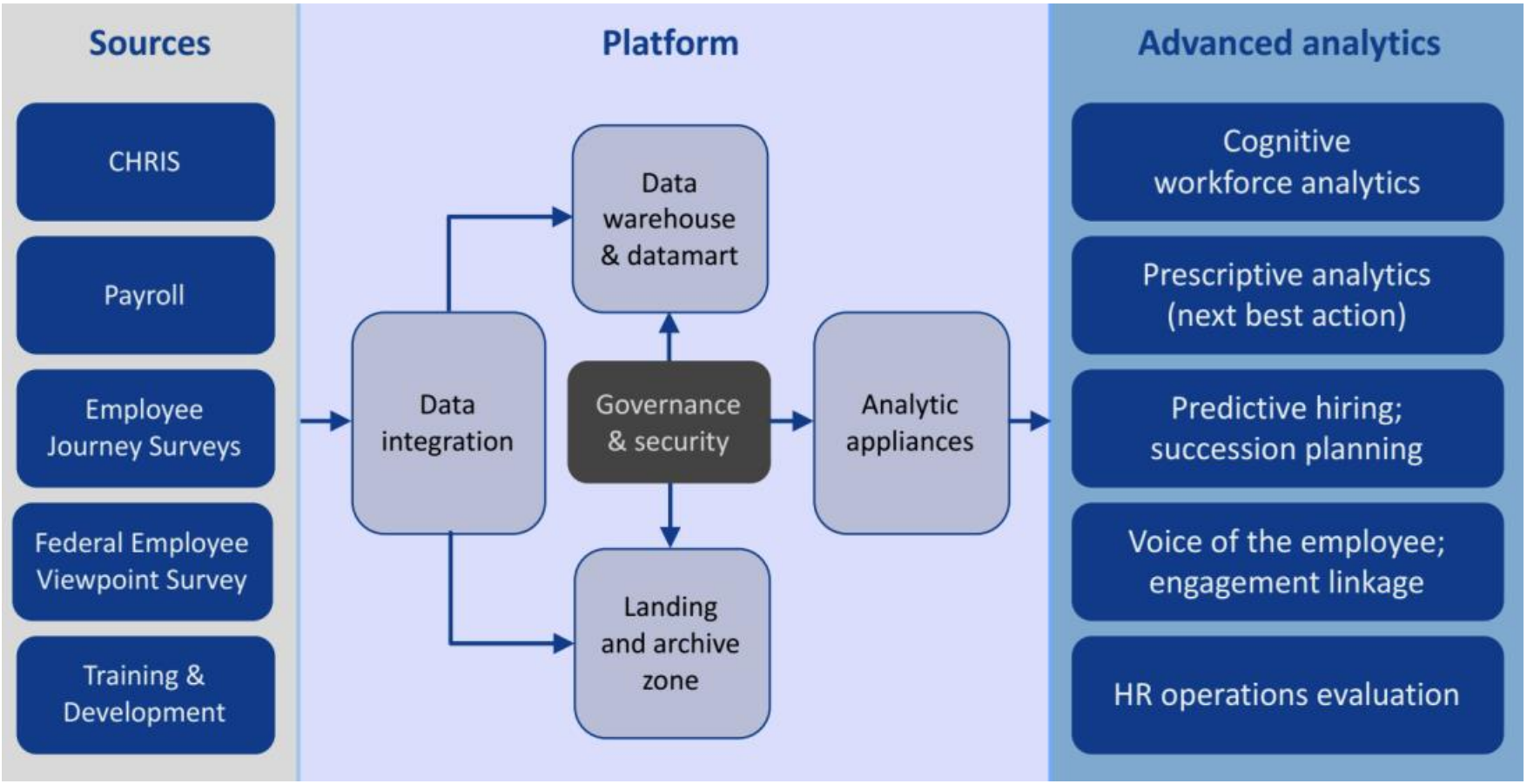


Figure 1. HRDW conceptual framework

E-mail me for further documentation if you are interested

Federal Government Example

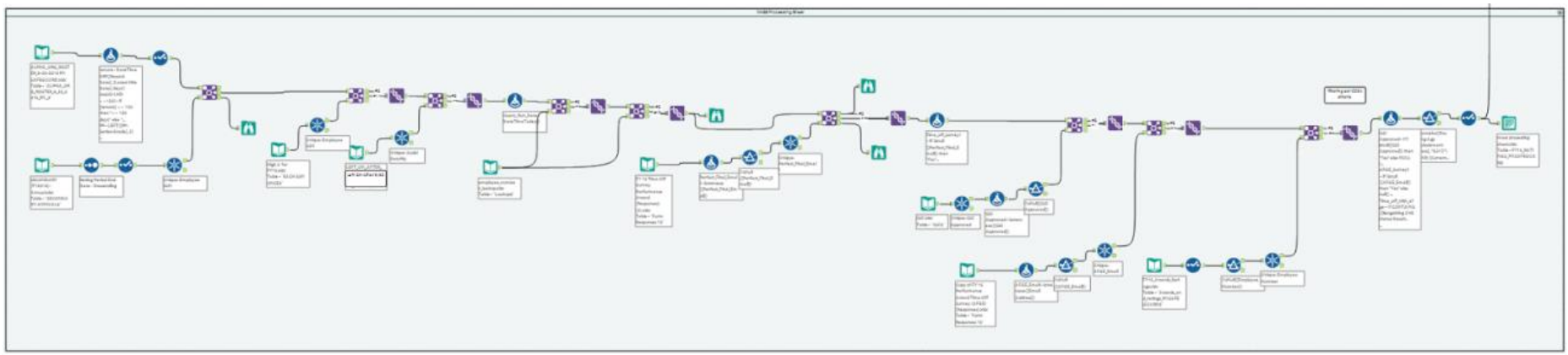


Figure 2. Data management workflow

OHRM: Real-Time Analysis

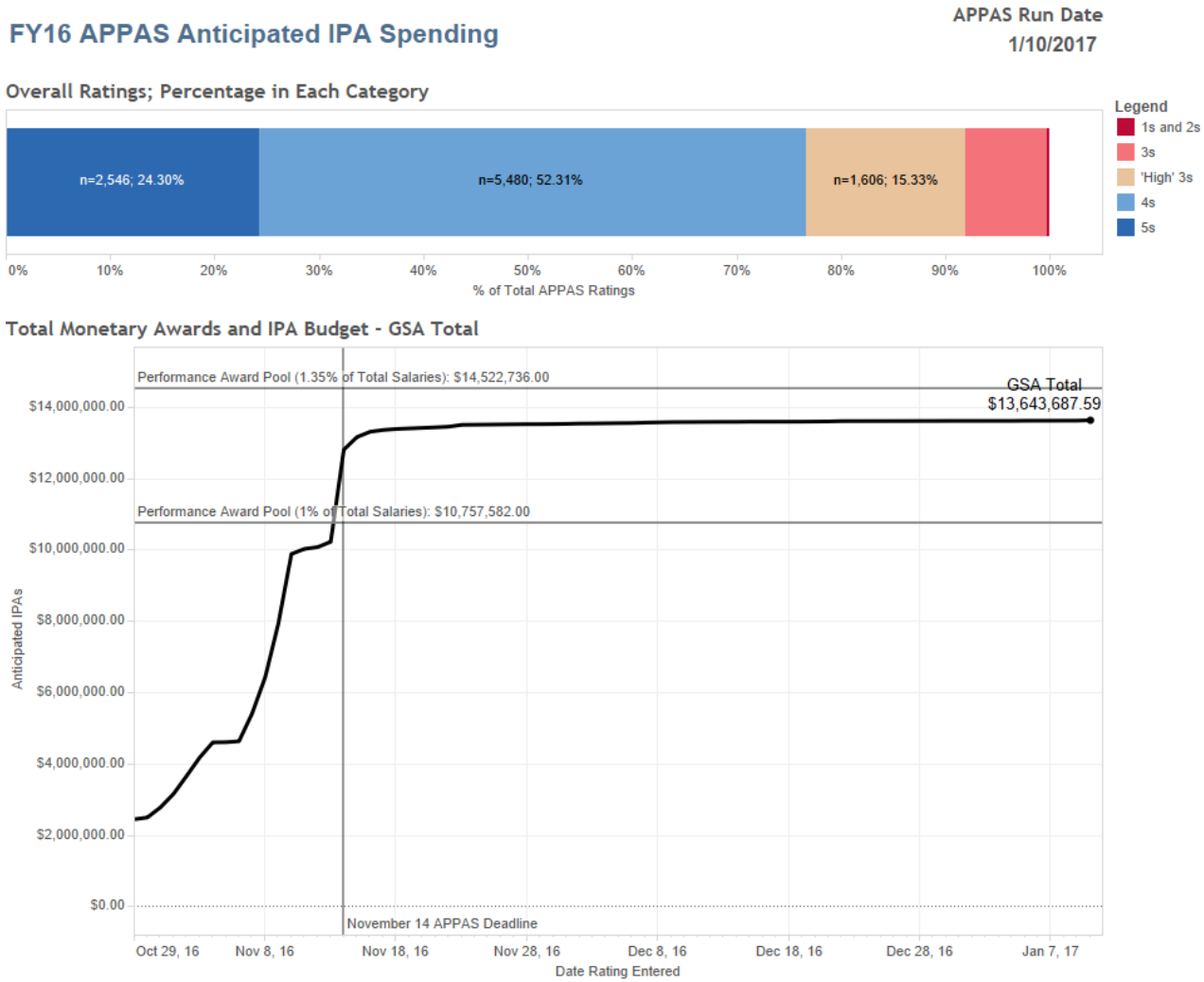


Figure 3. Real-time spend analysis: A real-time spending analysis tool displayed cost projections on over $13.6 million in award spending and over 40,000 hours of award leave. The dashboard, updated daily, provided current progress and projections towards reaching awards budget limits. Analysis at the aggregate, organizational, and individual levels helped agency leaders more proactively determine the impact of performance award percentages.

OHRM: Statistical Analysis on Bias

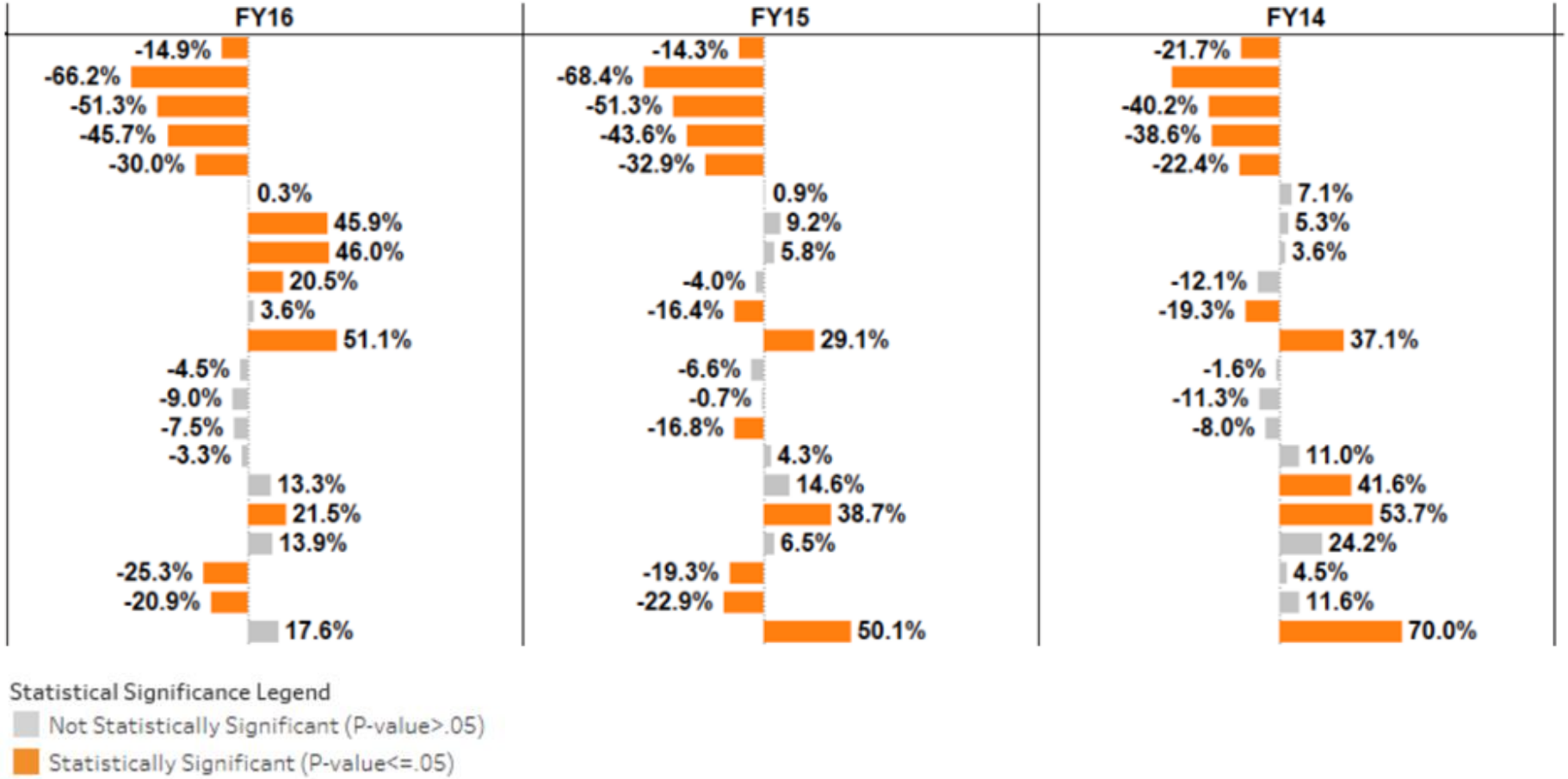
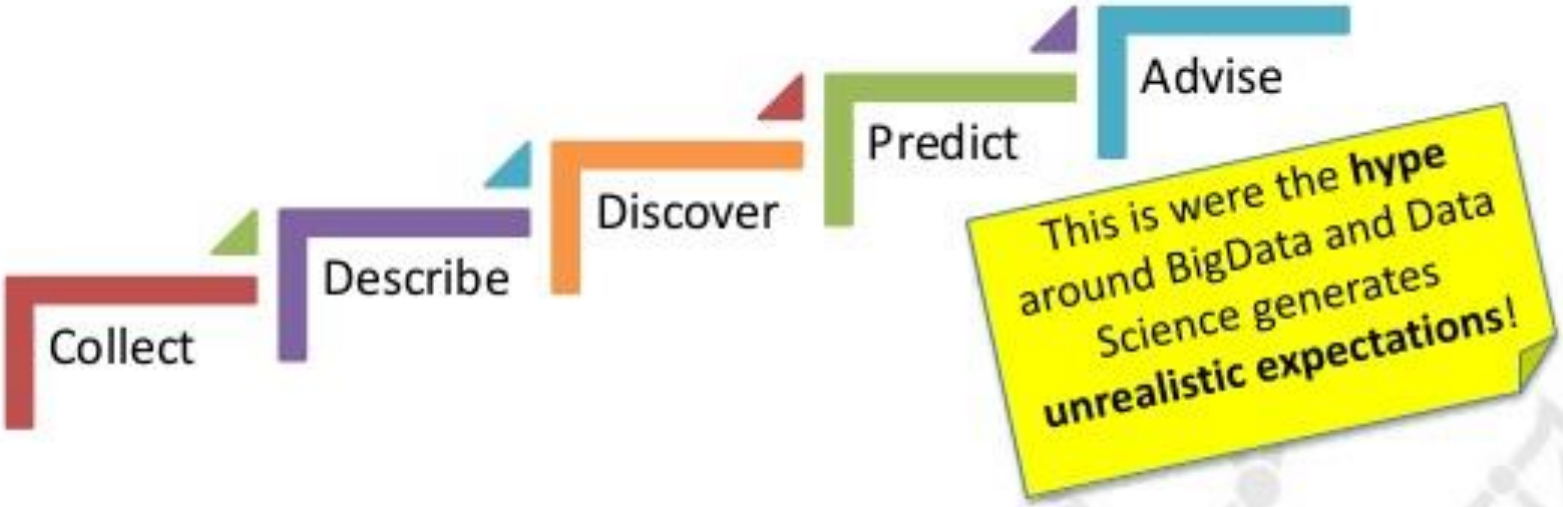


Figure 4. Descriptive statistical model: Multi-year statistical models visualized factors most associated with performance ratings outcomes, identifying statistical significance, magnitude, directionality and change over time (variable names are hidden). Percentages indicate the probability of an employee having received the next higher performance rating for each variable, holding other model variables constant.

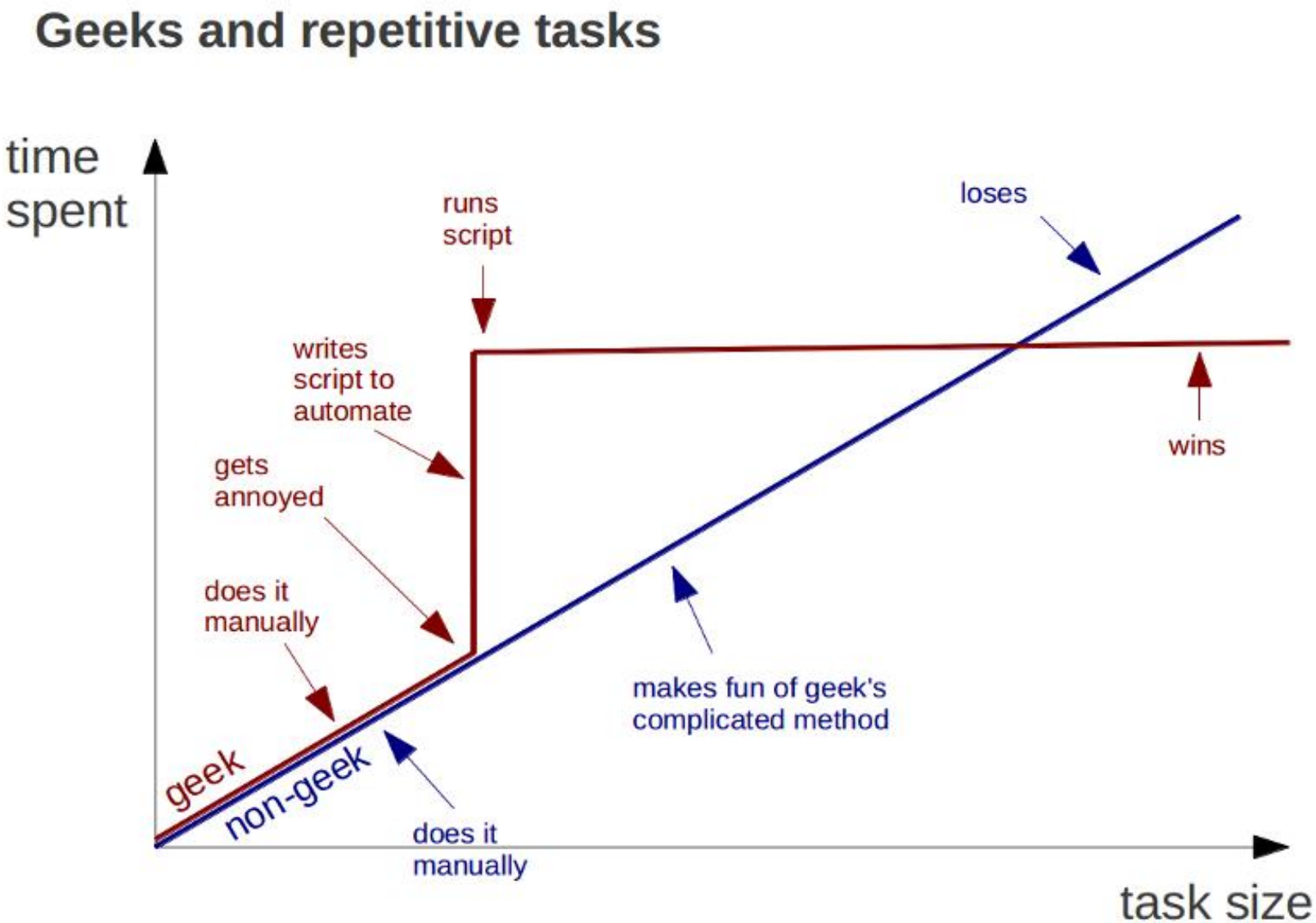
The Data Science maturity model

- Don't run before you can walk: The **Data Science Maturity model**
  Each level builds on the quality of the underlying step. It's science, not magic …

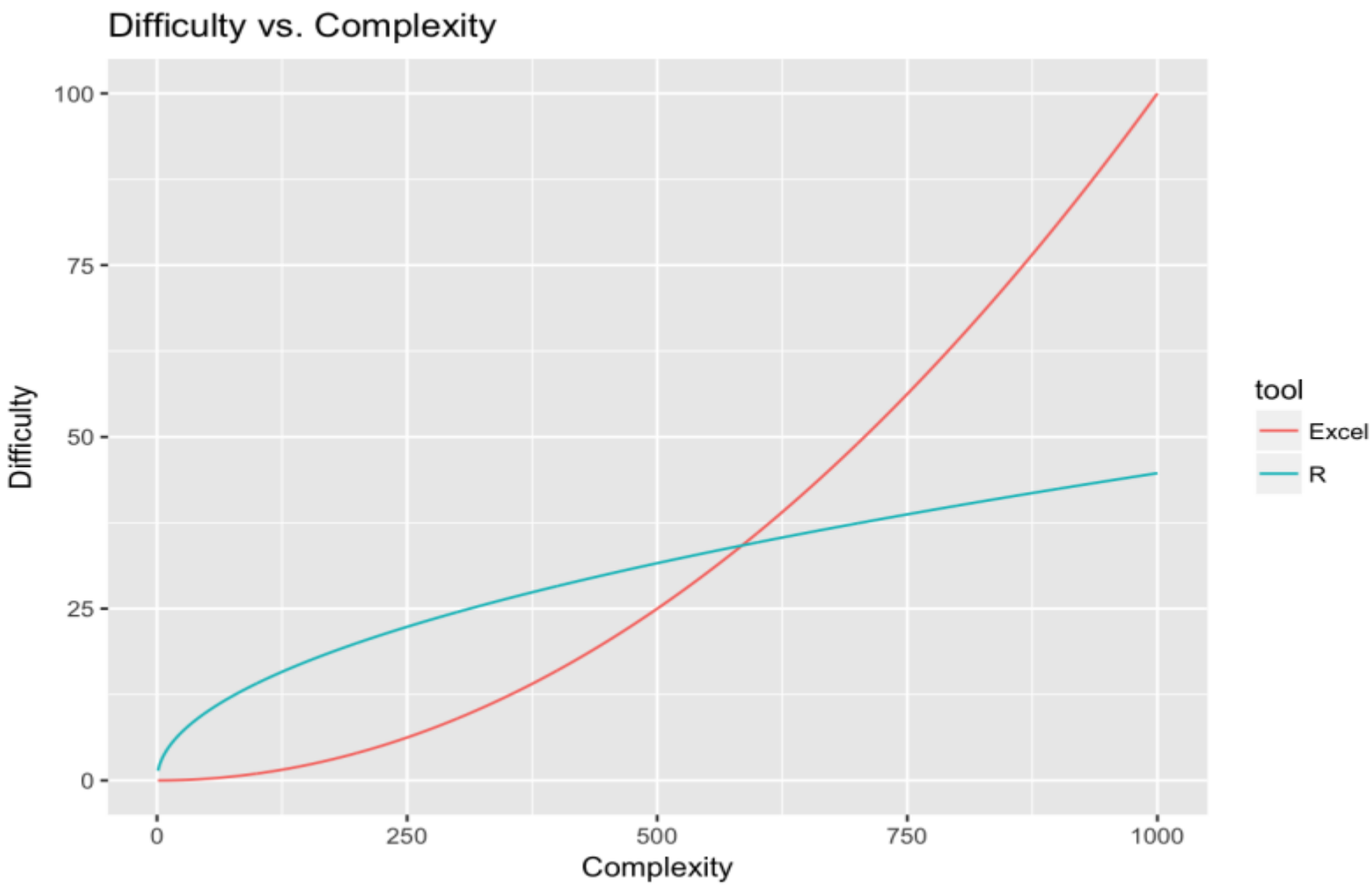


  - Start off by simply **collecting** the data you need (type, quantity, quality)
  - Then **report** on your current business (confirmative analysis)
  - **Discover** new and valuable information (exploratory analysis)
  - Build and test **prediction models** (predictive analysis)
  - Steer your business based on advise output from your predictions (data-driven)

Why not Excel?



Why not Excel?



Why Python and R?

R and Python are the two most popular programming languages used by data analysts and data scientists.

Both are free and open source, and were developed in the early 1990s—R for statistical analysis and Python as a general-purpose programming language.

Language Types (click to hide)

Web | Mobile | Enterprise | Embedded

Language Rank	Types	Spectrum Ranking
1. Python	Web, Enterprise	100.0
2. C	Mobile, Enterprise, Embedded	99.7
3. Java	Web, Mobile, Enterprise	99.4
4. C++	Mobile, Enterprise, Embedded	97.2
5. C#	Web, Mobile, Enterprise	88.6
6. R	Enterprise	88.1
7. JavaScript	Web, Mobile	85.5
8. PHP	Web	81.4
9. Go	Web, Enterprise	76.1
10. Swift	Mobile, Enterprise	75.3

Why Python and R?

- Created for simplicity and readability
- Rapid prototyping, ease of production
- Open source, importable libraries/packages
- Broad range of applications
- Fast growing community

Why Python and R?

Java

```
import javax.swing.JFrame;                                  //Importing class JFrame
import javax.swing.JLabel;                                  //Importing class JLabel
public class HelloWorld {
        public static void main(String[] args) {
                JFrame frame = new JFrame();                 //Creating frame
                frame.setTitle("Hi!");                       //Setting title frame
                frame.add(new JLabel("Hello, world!"));     //Adding text to frame
                 frame.pack();                              //Setting size to smallest
                frame.setLocationRelativeTo(null);          //Centering frame
                frame.setVisible(true);                     //Showing frame
        }
}
```

C

```
#include
int main(void)
{
                puts("Hello, world!");
}
```

R

```
cat('Hello, world! ')
```

Python

```
print(‘Hello, world')
```

Python vs R

R has an edge in statistics and visualization (these things are syntactically simpler)

Python has the edge in machine learning capabilities and connecting analyses to webapps.

Many advanced Data Scientists learn and use both, switching between the two to handle different tasks.

Choosing which language to start with depends on your situation. Here's a link for a more in-depth analysis

Expectations on What We'll Learn

You **WILL NOT** become a full-fledged Data Scientist after this course.

You **WILL** become familiar enough with Python and R to teach yourself how to read documentation and learn how to become a Data Scientist.

[The curiosity advantage: the most important skill for data science]

(O'Reilly)

Contact Information

Jonathan Joa – Jonathan.Joa@gsa.gov

Scott McAllister – Scott.McAllister@gsa.gov