# 01 | What is Data Science

*Ivan Corneillet*

*Data Scientist*

GENERAL ASSEMBLY

# Learning Objectives

## After this lesson, you should be able to:

- Describe the components of a successful learning environment
- Define what is data science and who data scientists are; define the data science workflow
- Setup your development environment and practice the different workflows we will use in this course

# Setting You Up for Success

# Meet Your Team

- Ivan Corneillet, Lead Instructor

- George McIntire, Associate Instructor

- Matt Jones, Course Producer

# Course Logistics

- Lead Instructor
  - Ivan Corneillet
- Associate Instructor
  - George McIntire
- Course Producer
  - Matt Jones
- Class
  - June 21 – August 30, Mondays and Wednesdays, 6:30PM – 9:30PM; no class on July 3
  - “Library” on June 21 and 26; classroom 8 thereafter

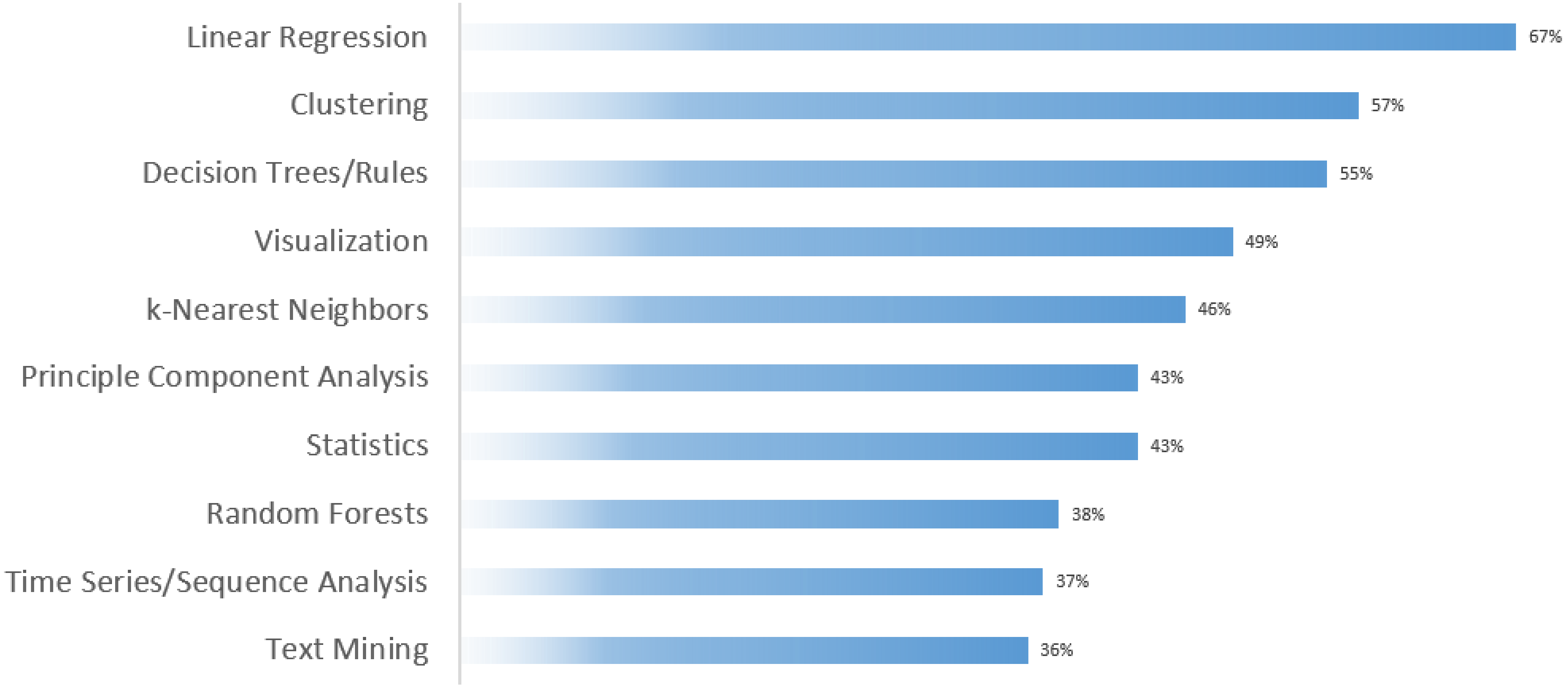
- Slack
  - https://ds-sf-36.slack.com
- GitHub
  - https://github.com/ga-students/DS-SF-36
- Exit Tickets
  - http://tiny.cc/ds-sf-36

# What skills will I learn in this class?

| | | | | |
|---|---|---|---|---|
| What is Data Science<br>*(session 1)* | Research Design<br>*(session 1)* | Python<br>*(session 2)* | *pandas*<br>*(session 3)* | Databases and Scrapping<br>*(session 4)* |
| Exploratory Data Analysis<br>*(session 5)* | $k$-Nearest Neighbors<br>*(session 6)* | Model Fit<br>*(session 6)* | Linear Regression<br>*(sessions 8–10)* | Regularization<br>*(sessions 11)* |
| Logistic Regression<br>*(sessions 12)* | Advanced Metrics<br>*(sessions 14)* | Trees<br>*(sessions 16)* | Natural Language Processing<br>*(session 18)* | Time Series<br>*(session 19)* |

# Top algorithms and methods used by data scientists

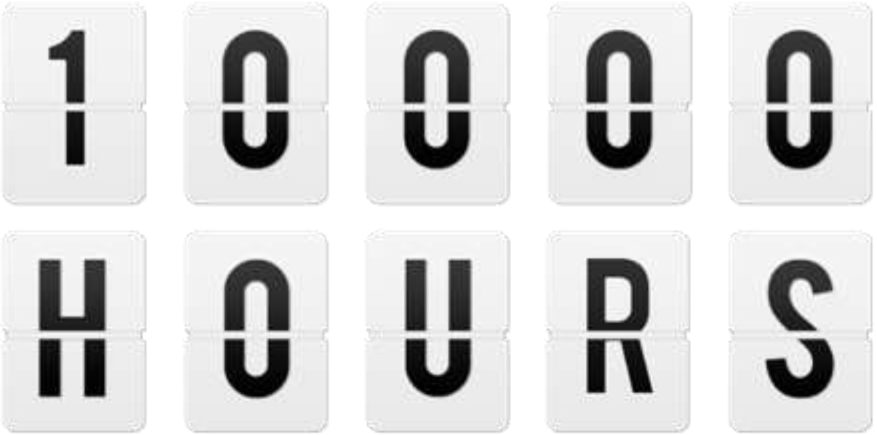
(http://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html)



# Gladwell’s 10,000 Hour Rule

(http://www.wisdomgroup.com/blog/10000-hours-of-practice)

- “Greatness requires enormous time”
  - It takes roughly ten thousand hours of practice to achieve mastery in a field



# How will I apply and reinforce these new skills?

| **Assignments**<br>*(ungraded)* | Take-home assignments | | |
|---|---|---|---|
| **Unit Project**<br>*(graded)* | Research Design<br>*(session 5)* | Exploratory Data Analysis<br>*(session 9)* | Machine Learning Modeling and Executive Summary<br>*(session 14)* |
| **Applied Sessions**<br>*(ungraded)* | Data Wrangling and Exploratory Data Analysis<br>*(session 7)* | Machine Learning Modeling<br>*(session 13)* | Machine Learning Modeling<br>*(session 17)* |
| **Final Project**<br>*(graded)* | Lightning Pitch<br>*(session 10)* | Research Design, Exploratory Data Analysis, and Intermediate Presentation<br>*(session 15)* | Machine Learning Modeling and Final Presentation<br>*(session 20)* |

# Typical Class

- *Pre-readings (usually optional)*
- Objectives
- Announcements
- Previous class review

- Series alternating between:
  - Lectures
    - (deck, whiteboard, codealongs, and demos)
  - Activities
    - (cold-calling, individual and group exercises, and codealongs)

- Class review
- Exit tickets
- *Post-readings (usually optional)*

# What is Data Science?
# Who are Data Scientists?

# Harvard Business Review | “Data Scientists: The Sexiest Job of the 21st Century” (2012)

(https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century)

HBR.ORG
Harvard Business Review
OCTOBER 2012
46 The Big Idea
The True Measures Of Success
Michael J. Mauboussin
84 International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
93 Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management
PAGE 59

SPOTLIGHT ON BIG DATA

Spotlight

ARTWORK Tamar Cohen, Andrew J Buboltz
2011, silk screen on a page from a high school yearbook, 8.5" x 11"

## Data Scientist:

*The Sexiest Job of the 21st Century*

***Meet the people who can coax treasure out of messy, unstructured data.***

*by Thomas H. Davenport and D.J. Patil*

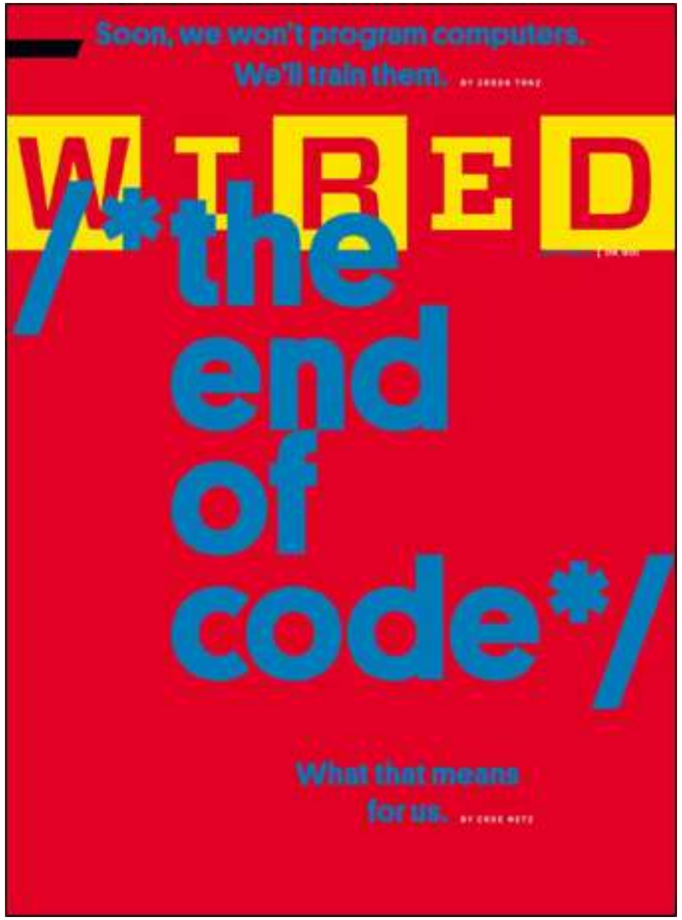
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

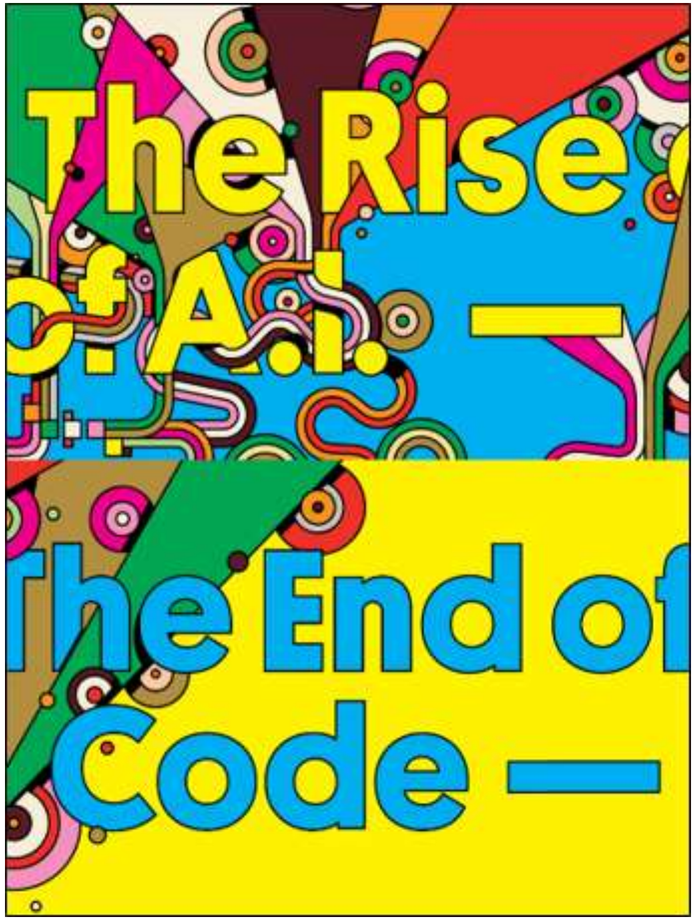
4 Harvard Business Review October 2012

Source: Harvard Business Review

# Wired | “The End of Code” | “Soon We Won't Program Computers.  We'll Train Them Like Dogs” (2016)

([http://www.wired.com/2016/05/the-end-of-code](http://www.wired.com/2016/05/the-end-of-code))





Source: Wired

# Data science is everywhere

# Common questions asked in data science

## How much?  How many?

- What will the temperature be next Tuesday?
- What will my fourth quarter sales in France be?
- How many kilowatts will be demanded from my wind farm 30 minutes from now?
- How many new followers will I get next week?

## Regression

- Predict a continuous outcome
  - *k*-Nearest Neighbors
  - Linear Regression
  - Trees

# Common questions asked in data science (cont.)

## Is this A, B or C?

- Will this customer default on their loan?
- Is this an image of a man, a cat, or a dog?
- Will this customer click on the advertisement?
- Which team will win the championship?
- Is this mole malignant or benign?

## Classification

- Predict a discrete outcome
  - $k$-Nearest Neighbors
  - Logistic Regression
  - Trees

# Common questions asked in data science (cont.)

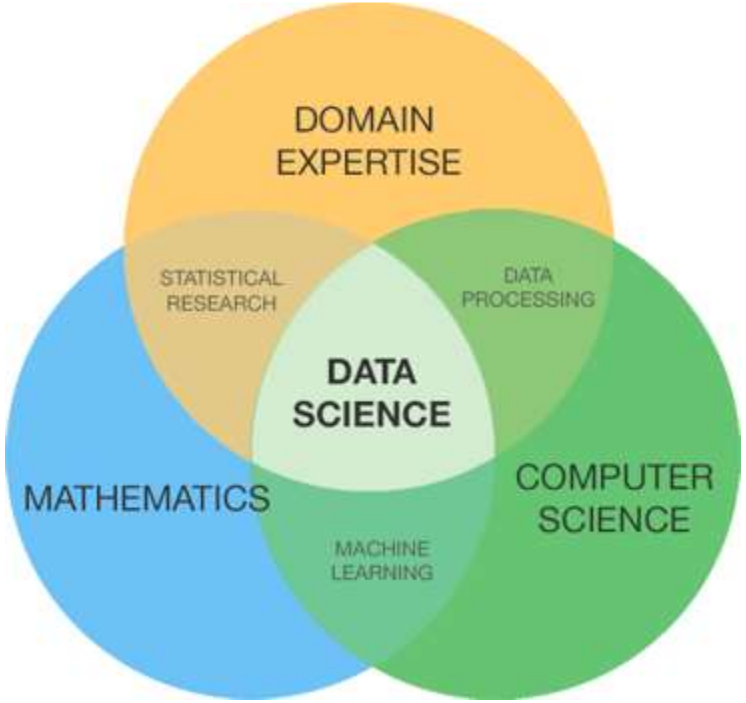
## How is this Data Organized?

- What are the different types of coffee drinkers?
- Which viewers like the same kind of movies?
- What kinds of car models does GM produce?
- Are there common clusters of cable channels that customers tend to purchase together?
- What is a natural way to break these documents into five topics?

## Clustering

- What are the “categories” within the data?

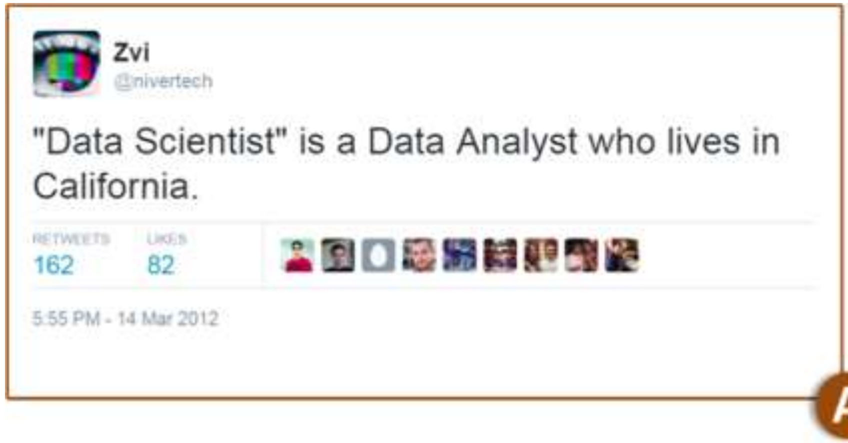
# Data science involves a variety of skillsets



Source: Data Science for the C-suite

# Data scientists in ≤140 characters

A



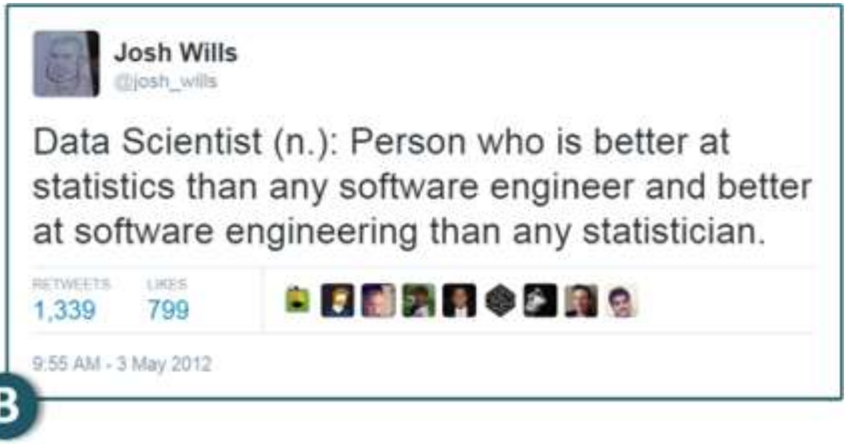
Zvi
@nivertech

"Data Scientist" is a Data Analyst who lives in California.

RETWEETS 162 LIKES 82

5:55 PM - 14 Mar 2012

B



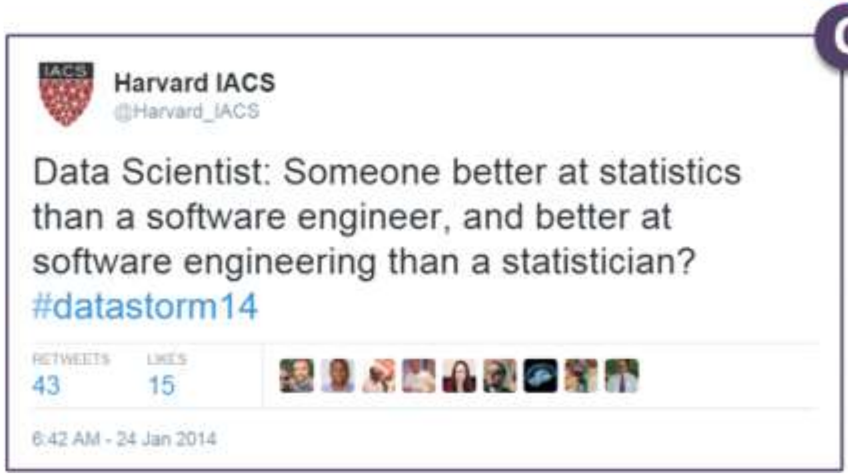
Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS 1,339 LIKES 799

9:55 AM - 3 May 2012

C



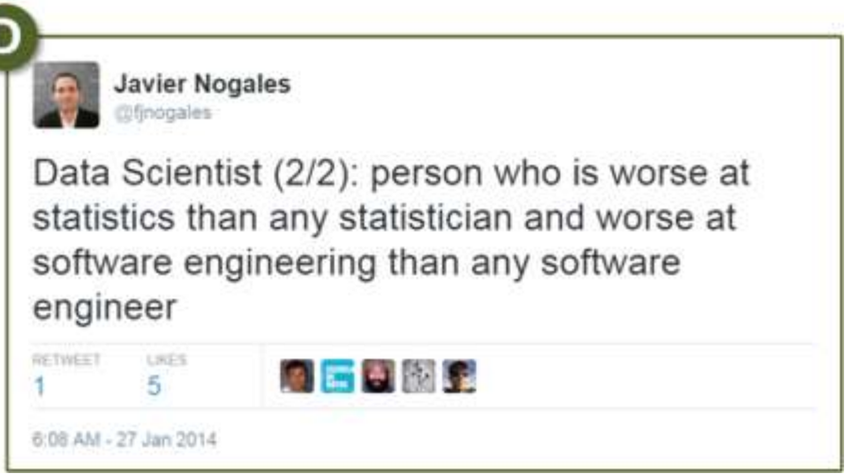
Harvard IACS
@Harvard_IACS

Data Scientist: Someone better at statistics than a software engineer, and better at software engineering than a statistician? #datastorm14

RETWEETS 43 LIKES 15

6:42 AM - 24 Jan 2014

D



Javier Nogales
@fjnogales

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET 1 LIKES 5

6:08 AM - 27 Jan 2014

Source: Twitter

# Wired’s “Soon We Won’t Program Computers. We’ll Train Them Like Dogs” (2016) (cont.)

**Behaviorism/Behavioral Psychology**

- Brain as a black box
  - Stimulus and response, feedback and reinforcements
    - “ring bell, dog salivates”

**Cognitive Psychology**

- Brain more like a computer
  - Thoughts as programs
  - Absorb, process, and act upon information

# Wired's "Soon We Won't Program Computers. We'll Train Them Like Dogs" (2016) (cont.)

## Machine Learning

- Humans *train* computers
  - Keep showing cats to a computer and eventually it will *learn* to recognize cats (https://www.wired.com/2012/06/google-x-neural-network)
  - No symbols, no rules; instead an unparsable machine learning

## Traditional Programming

- Humans *write code* (as explicit step-by-step-instructions) for computers to follow
  - Rule-based determinism
    - "Write enough rules and eventually, we'd create a system sophisticated enough to understand the world"
  - For years, Google Search relied mostly on these human-written rules (https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches)

# Wired's "Soon We Won't Program Computers. We'll Train Them Like Dogs" (2016) (cont.)

**Age of Entanglement**

- Outside-in view of how machine work
  - "Code doesn't just determine behavior, behavior also determine code"

**Age of Enlightenment**

- Inside-out view of how machine work
  - "First, we write the code, then the machine expresses it"

In this course we will model the stimuli as a matrix $\boldsymbol{X}$ (the **feature matrix**); the response is modelled as a vector $\boldsymbol{y}$ (the **response vector**). We will use these key data structures as inputs to our machine learning algorithms

## Feature Matrix $X$

*Stimulus/feedback*
*“ring bell”*

| | col0 | col1 | col2 | col3 |
|---|---|---|---|---|
| row0 | | | | |
| row1 | | | | |
| row2 | | | | |
| row3 | | | | |

## Response Vector $y$

*Response/reinforcements*
*“dog salivates”*

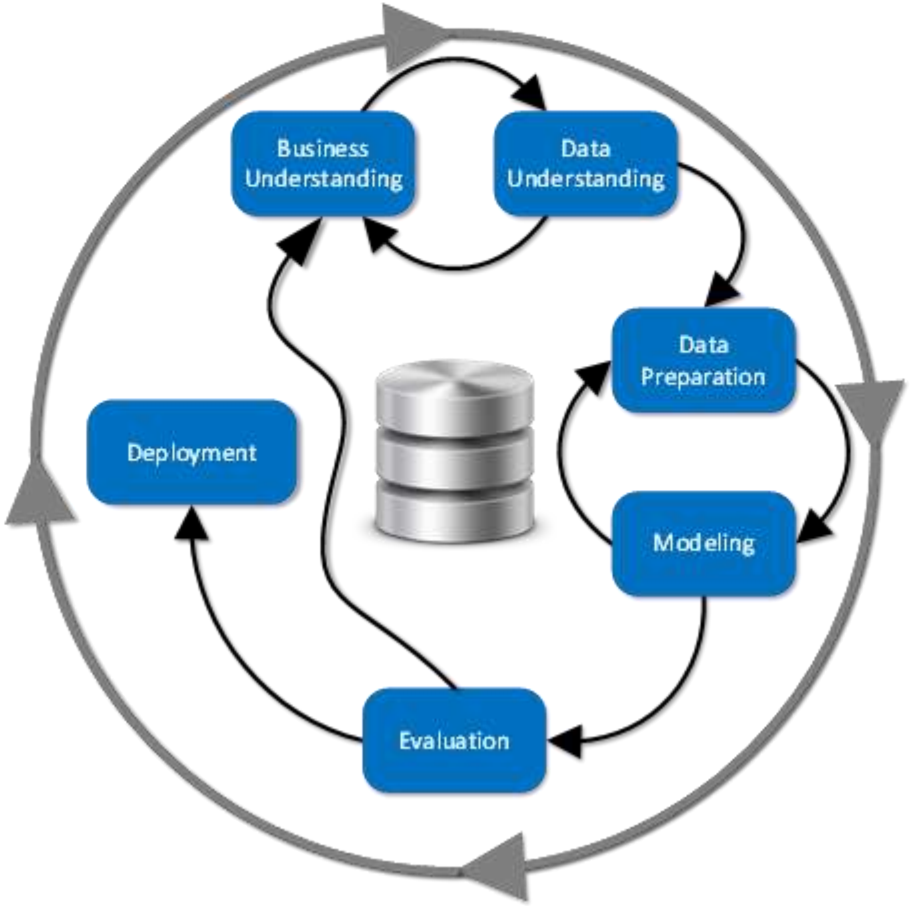
| | col |
|---|---|
| row0 | |
| row1 | |
| row2 | |
| row3 | |

DS

# Data Science Workflow

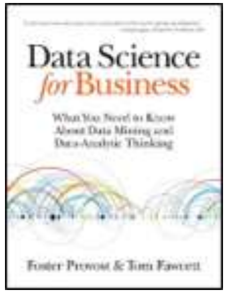
# Data Science Workflow

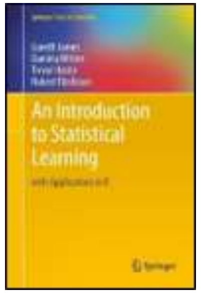
## (a.k.a., Cross-Industry Standard Process for Data Mining, or CRISP-DM)



# Some resources to follow along the class (or afterwards…) *(will reference either <u>pre-class reading</u> and/or <u>post-class reading</u> materials; optional; not required for the course)*

- Data Science for Business (by Provost and Fawcett) (link) (General Assembly holds several copies in its library)


Data Science for Business
What You Need to Know About Data Mining and Data-Analytic Thinking
Foster Provost & Tom Fawcett



- An Introduction to Statistical Learning: with Applications in R (by James et al.) (e-book available free-of-charge here)

- For a more advanced treatment of these topics, check out The Elements of Statistical Learning: Data Mining, Inference, and Prediction (by Hastie et al.). (e-book also available free-of-charge here)