01 | What is Data Science

Ivan Corneillet

Data Scientist

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

▸ TBD, Associate Instructor

▸ Matt Jones, Course Producer

Course Logistics

- Lead Instructor

  - Ivan Corneillet

- Associate Instructor

  - TBD

- Course Producer

  - Matt Jones

- Class

  - August 29 – November 2, Tuesdays and Thursdays, 6:30PM – 9:30PM

  - Classroom 4

- Slack

  - https://ds-sf-38.slack.com

- GitHub

  - https://github.com/ga-students/DS-SF-38

- Exit Tickets

  - http://tiny.cc/ds-sf-38

What skills will I learn in this class?

What is Data Science *(session 1)*	Research Design *(session 1)*	Python *(session 2)*	*pandas* *(session 3)*	Databases and Scrapping *(session 4)*	Exploratory Data Analysis *(session 5)*
k-Nearest Neighbors *(session 6)*	Model Fit *(session 6)*	Data Wrangling and Exploratory Data Analysis Challenge *(session 7)*	Linear Regression *(sessions 8–10)*	Regularization *(session 11)*	Logistic Regression *(session 12)*
Machine Learning Modeling Challenge *(session 13)*	Trees *(session 14)*	Review *(session 16)*	Machine Learning Modeling Challenge, Take 2 *(session 17)*	Natural Language Processing *(session 18)*	Time Series *(session 19)*

Top algorithms and methods used by data scientists

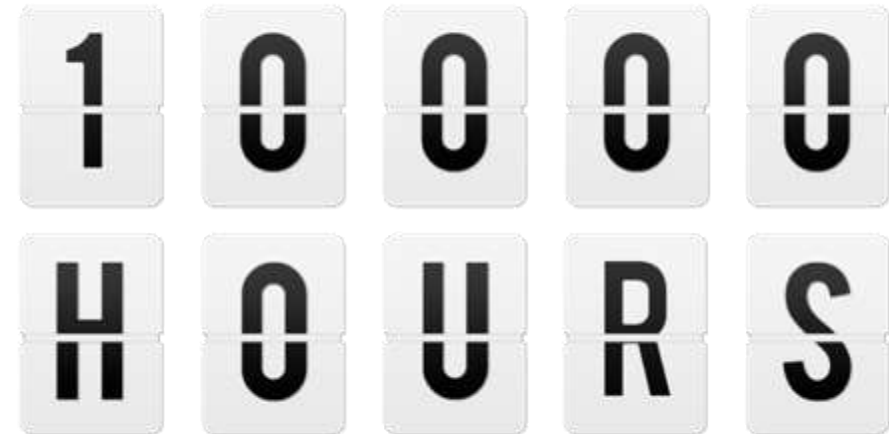(http://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html)
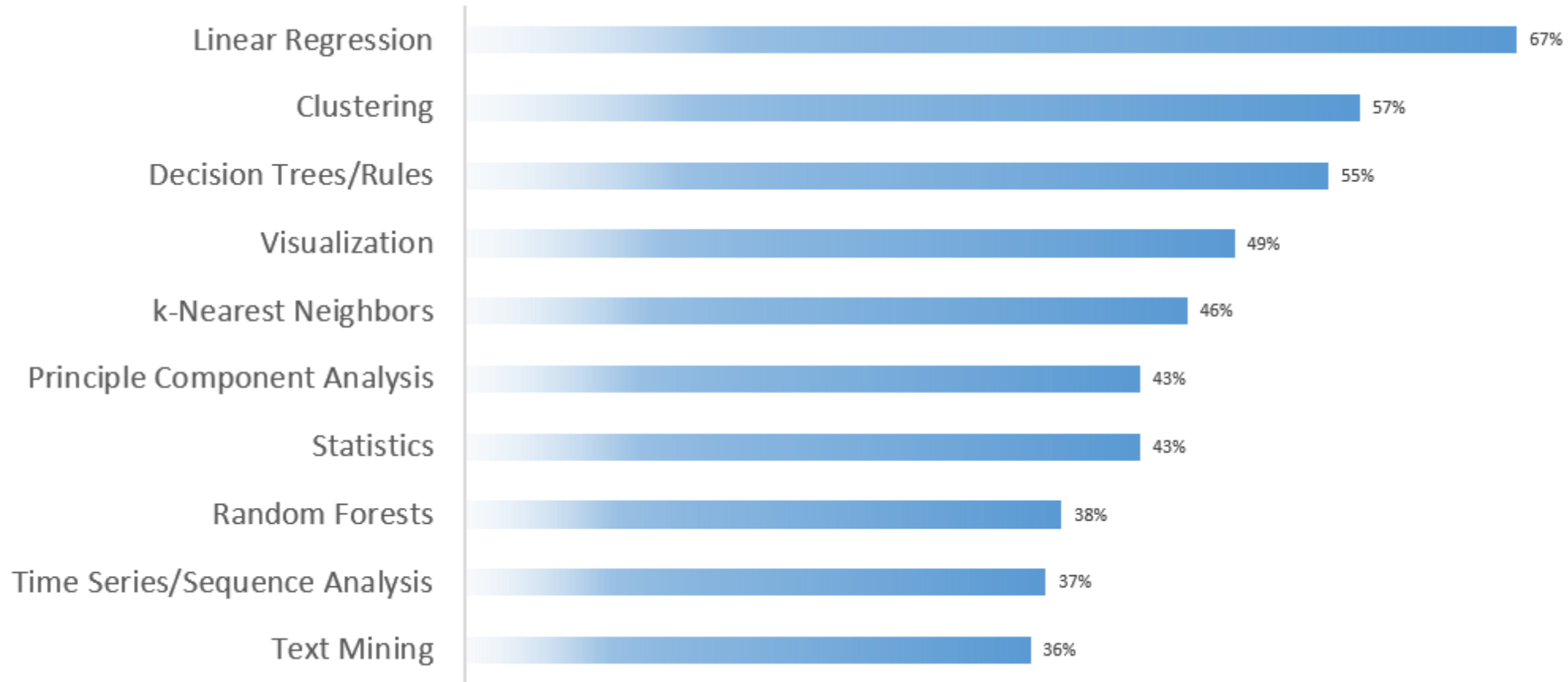
Gladwell's 10,000 Hour Rule

‣ "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 (ungraded)	Take-home assignments		
Unit Project *(graded)*	Research Design *(session 5)*	Exploratory Data Analysis *(session 9)*	Machine Learning Modeling and Executive Summary *(session 14)*
Challenges *(ungraded)*	Data Wrangling and Exploratory Data Analysis *(session 7)*	Machine Learning Modeling *(session 13)*	Machine Learning Modeling *(session 17)*
Final Project *(graded)*	Lightning Pitch *(session 10)*	Research Design, Exploratory Data Analysis, and Intermediate Presentation *(session 15)*	Machine Learning Modeling and Final Presentation *(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)
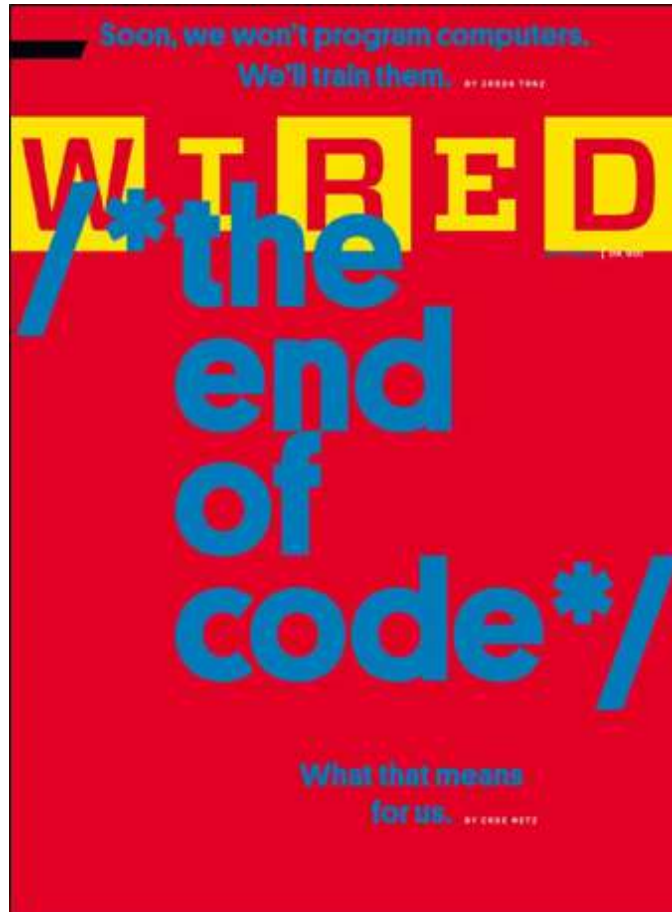
Source: Harvard Business Review

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

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
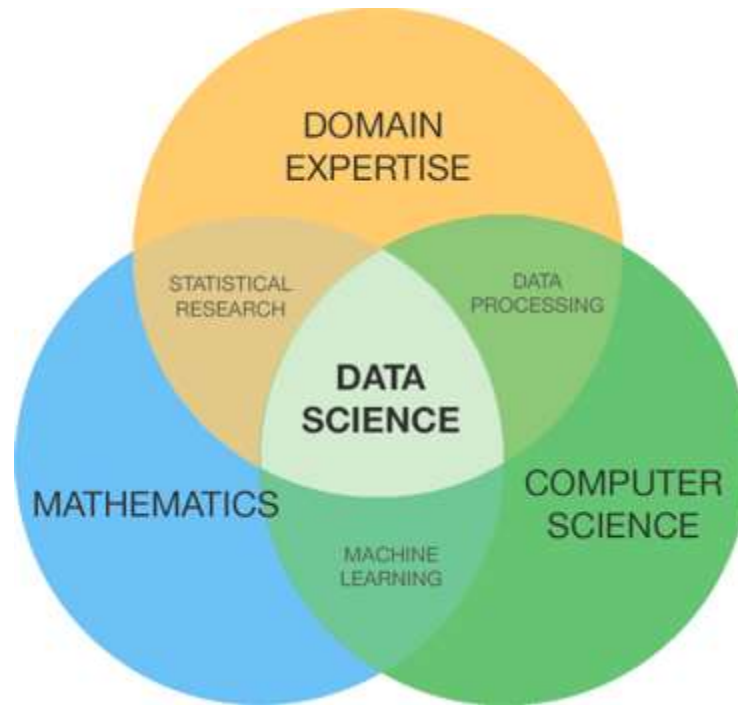
Source: Data Science for the C-suite

Data scientists in ≤140 characters

Zvi @nivertech

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

RETWEETS 162 LIKES 82

5:55 PM - 14 Mar 2012

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

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

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 X (the **feature matrix**); the response is modelled as a vector 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"

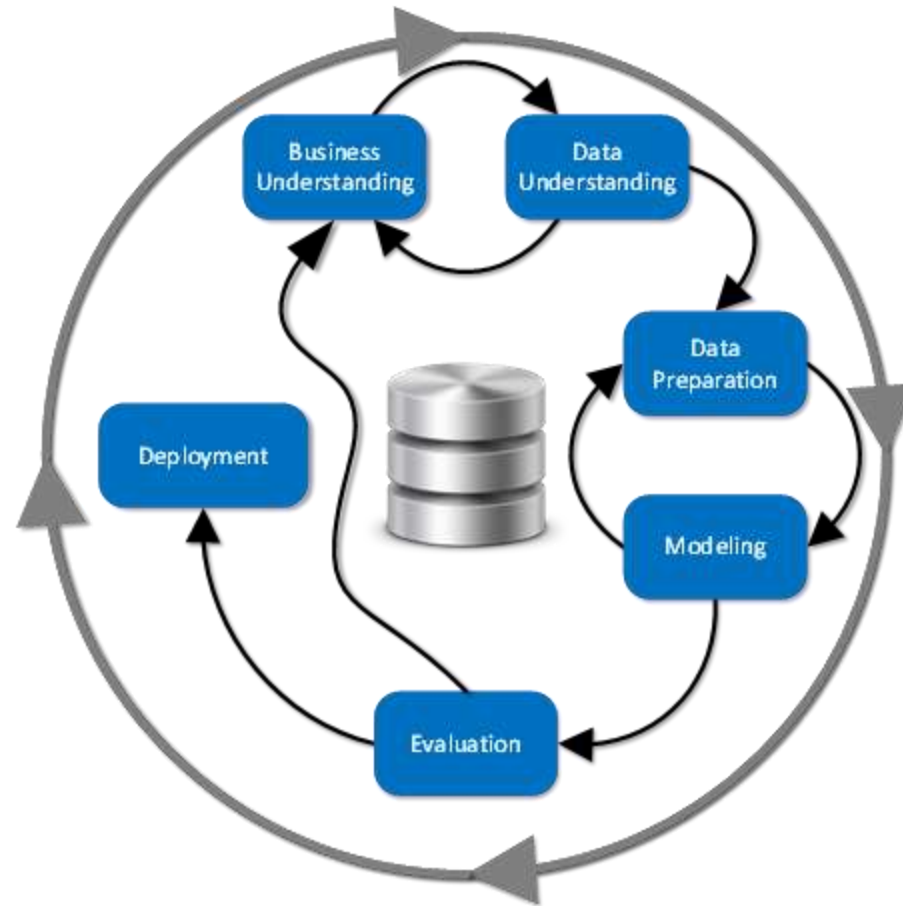|	col
row0	
row1	
row2	
row3	

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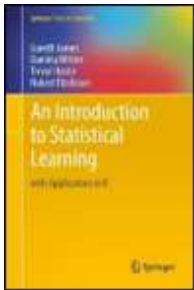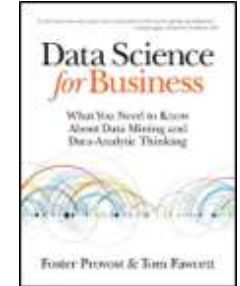
        ‣ An Introduction to Statistical Learning: with Applications in R (by James et al.) (e-book available free-of-charge here)

‣ 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)
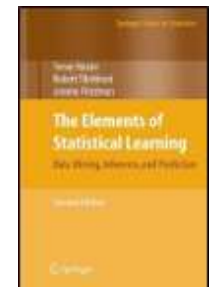
Slides © 2017 Ivan Corneillet Where Applicable
Do Not Reproduce Without Permission