01 | What is Data Science

Ivan Corneillet

Data Scientist

Here's what's happening today:

- Welcome to GA and DS!

- Setting you up for success

- What is Data Science

  - and who are data scientists?

- Class Discussion

  - "Data Scientists: The Sexiest Job of the 21st Century" (Harvard Business Review)

  - "The Rise of Artificial Intelligence and the End of Code" (Wired)

- Installfest

- The Data Science Workflow

  - Overview

  - ❶ IDENTIFY the Problem

  - ❼ PRESENT the Results

- Python Review

Welcome to GA and DS!

Setting You Up for Success

Meet Your Team

‣ Ivan Corneillet, Lead Instructor

‣ Gus Ostow, Associate Instructor

‣ Mohit Nalavadi, Associate Instructor

‣ Matt Jones, Course Producer

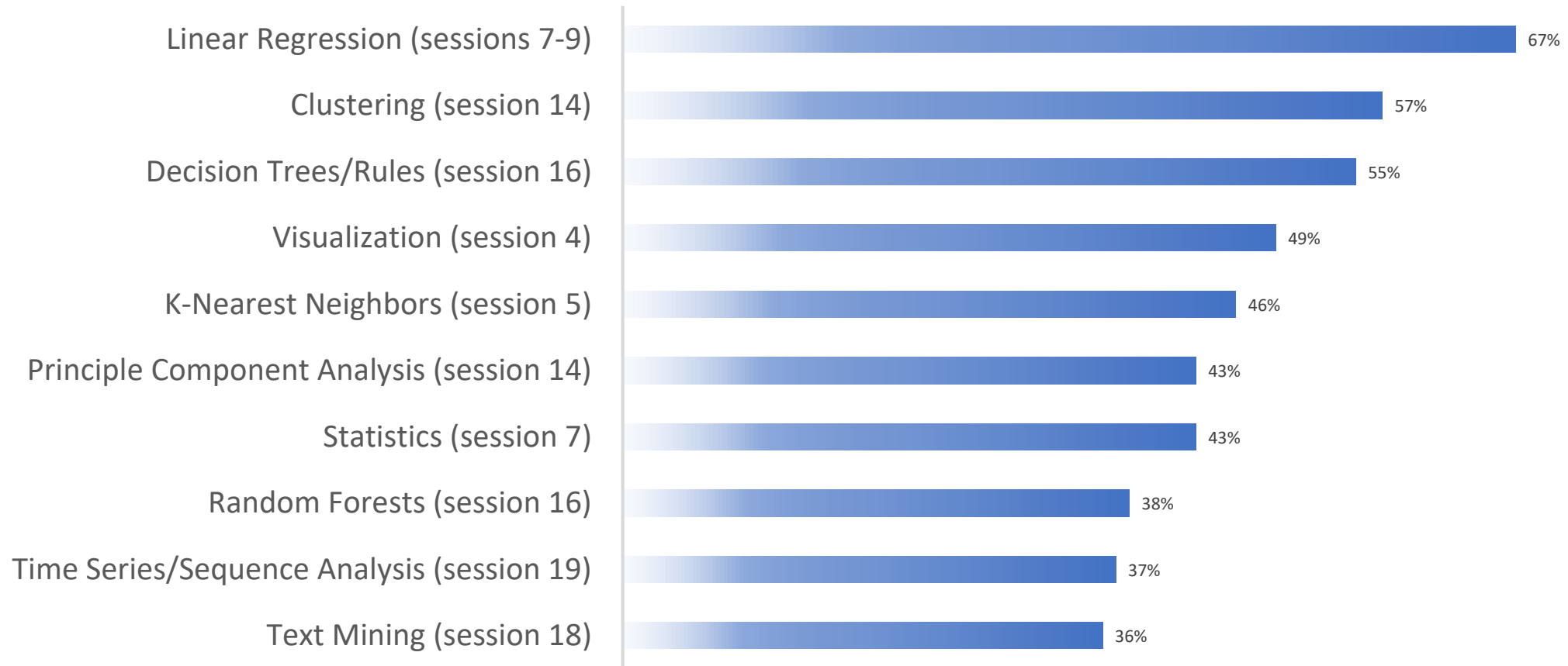# Course Logistics

- Lead Instructor

  - Ivan Corneillet

- Associate Instructors

  - Gus Ostow and Mohit Nalavadi

- Course Producer

  - Matt Jones

- Class

  - April 17 – June 26, Mondays and Wednesdays, 6:30PM – 9:30PM (no classes May 29)

  - Classroom 8

- Slack

  - https://ds-sf-34.slack.com

- GitHub

  - https://github.com/ga-students/DS-SF-34

- Exit Tickets

  - http://tiny.cc/ds-sf-34

What skills will I learn in this class?

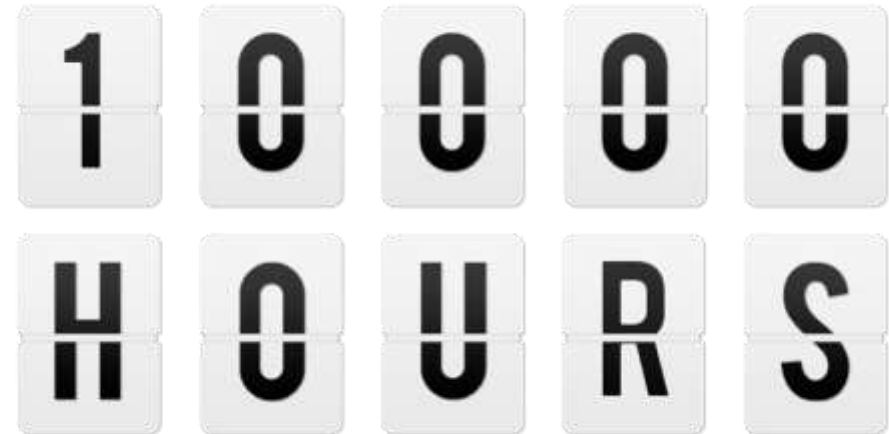|				
What is Data Science *(session 1)*	Research Design *(session 1)*	The *pandas* library *(sessions 2, 4, and 6)*	Databases and APIs *(session 3)*	Exploratory Data Analysis (and Data Visualization) *(sessions 4 and 6)*
k-Nearest Neighbors *(sessions 5, 12, and 17)*	Model Fit *(sessions 5, 7–9, 12–13, and 17)*	Linear Regression *(sessions 7–9, 12, and 17)*	Regularization *(sessions 10, 12, and 17)*	Logistic Regression *(sessions 11 and 17)*
Clustering *(sessions 14 and 17)*	Trees *(sessions 16 and 17)*	Natural Language Processing *(session 18)*	Time Series *(session 19)*	Presenting Insights from Data Models *(sessions 1, 15, and 20)*

Top 10 Algorithms & Methods used by Data Scientists

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

Algorithm	Percentage
Linear Regression (sessions 7-9)	67%
Clustering (session 14)	57%
Decision Trees/Rules (session 16)	55%
Visualization (session 4)	49%
K-Nearest Neighbors (session 5)	46%
Principle Component Analysis (session 14)	43%
Statistics (session 7)	43%
Random Forests (session 16)	38%
Time Series/Sequence Analysis (session 19)	37%
Text Mining (session 18)	36%

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?

Unit Project You will design a research project, perform exploratory data analysis and build a logistic model to determine what factors affect admission the most	Research Design	Exploratory Data Analysis	Machine Learning Modeling	Executive Summary

Final Project Using a dataset of your choosing, you will design a project, build a data science model and present their finding to the course	Lightning Presentation	Experimental Write-up	Exploratory Data Analysis	Notebook Draft	Project Presentation

Typical Class

- *Pre-readings (usually optional)*
- Class 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
- Homework *(ungraded)*
- *Post-readings (usually optional)*

What is Data Science and Who are Data Scientists?

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

Source: Harvard Business Review

Wired | "The Rise of Artificial Intelligence and the End of Code" (2016)

(http://www.wired.com/2016/05/the-end-of-code/ and https://www.wired.com/2016/05/google-alpha-go-ai/)
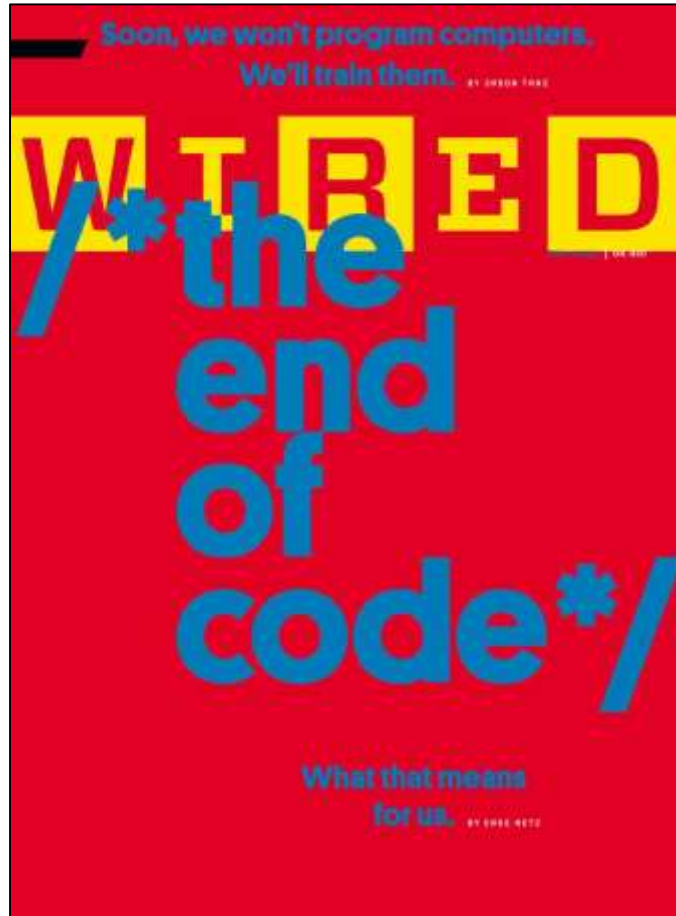
Source: Wired

Data Science is Everywhere
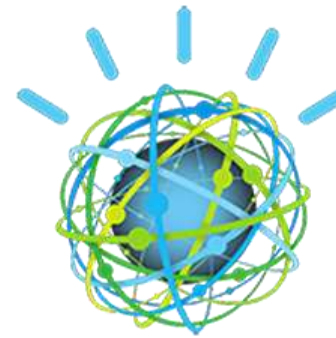
Common Questions Asked in Data Science

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 (session 5)

    ‣ Linear Regression (sessions 7–9)

    ‣ Trees (session 16)

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 (session 5)

  ‣ Logistic Regression (session 11)

  ‣ Trees (session 16)

Common Questions Asked in Data Science (cont.)
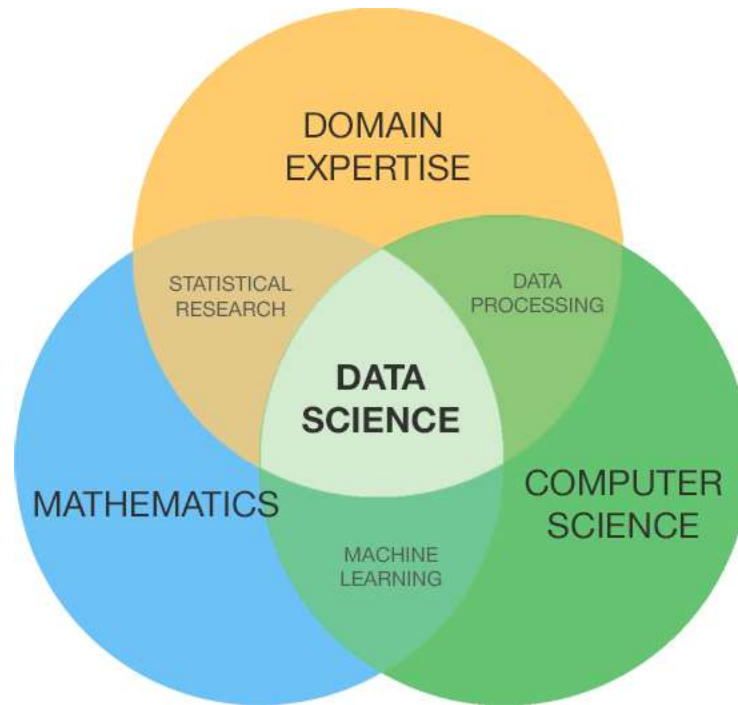
How is this Data Organized?

‣ 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?
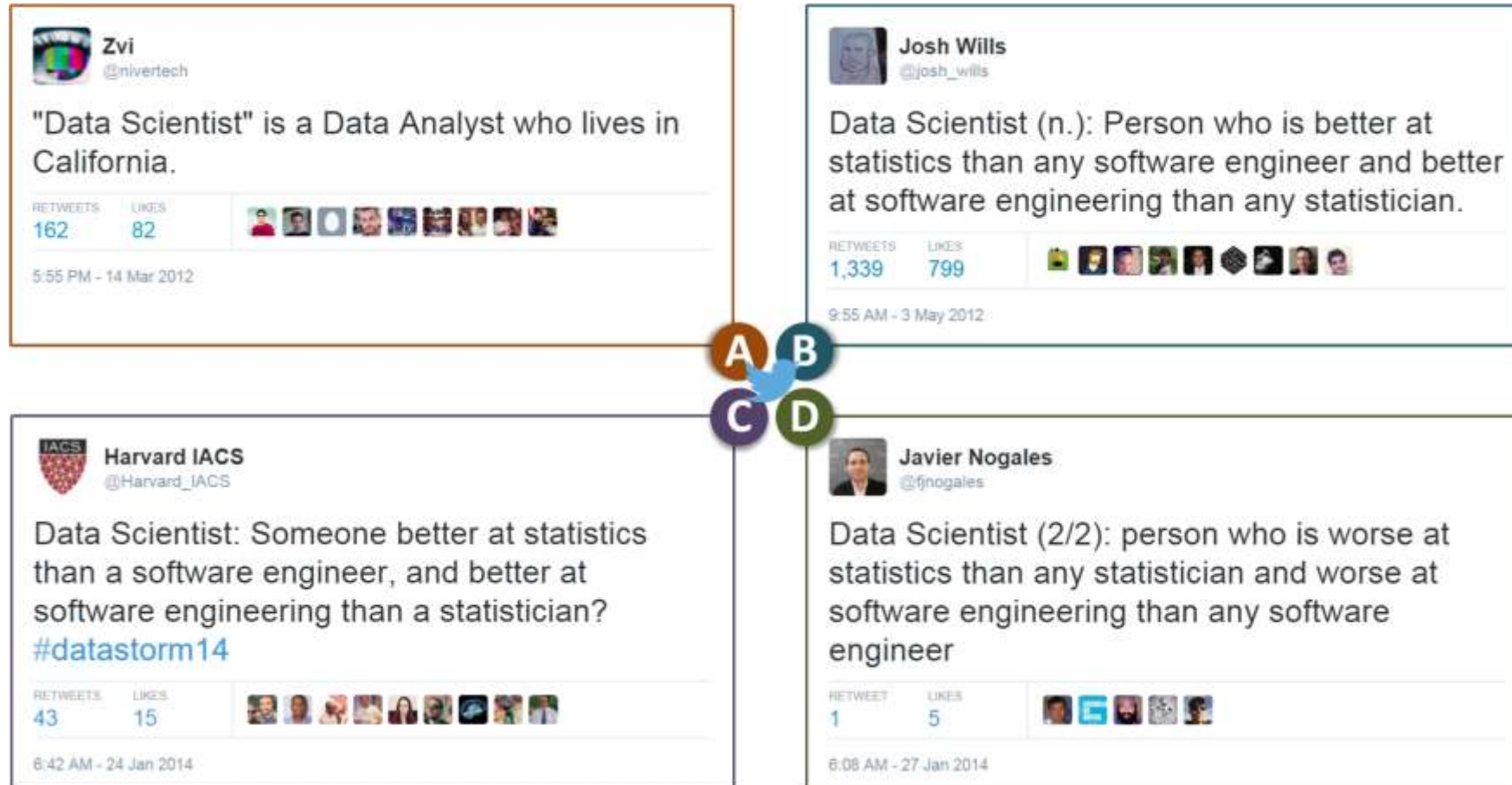
Clustering

‣ 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 "The Rise of Artificial Intelligence and the End of Code" (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 "The Rise of Artificial Intelligence and the End of Code" (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 "The Rise of Artificial Intelligence and the End of Code" (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**). These are very important data structures that will be used an 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	

Installfest

Data Science Workflow

The Data Science Workflow
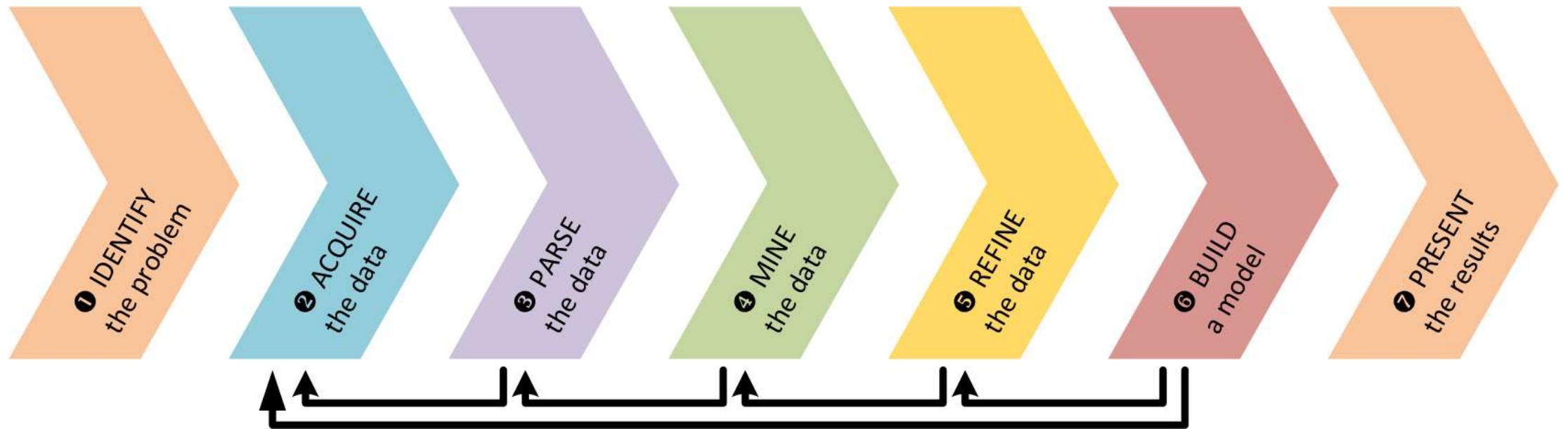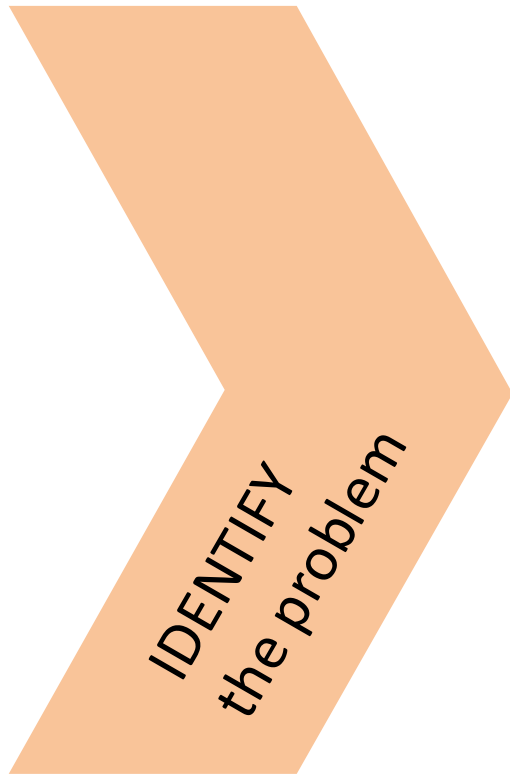
❶ IDENTIFY the problem
❷ ACQUIRE the data
❸ PARSE the data
❹ MINE the data
❺ REFINE the data
❻ BUILD a model
❼ PRESENT the results

❶ IDENTIFY the Problem

IDENTIFY
the problem

- ‣ Identify the Problem

  - ‣ Identify business/product objectives

  - ‣ Identify and hypothesize goals and criteria for success

  - ‣ Create a set of questions for identifying correct dataset

❶ IDENTIFY the Problem

The Why's and How's of a Good Question

Corina Rosu © 123RF.com

❷ ACQUIRE the Data
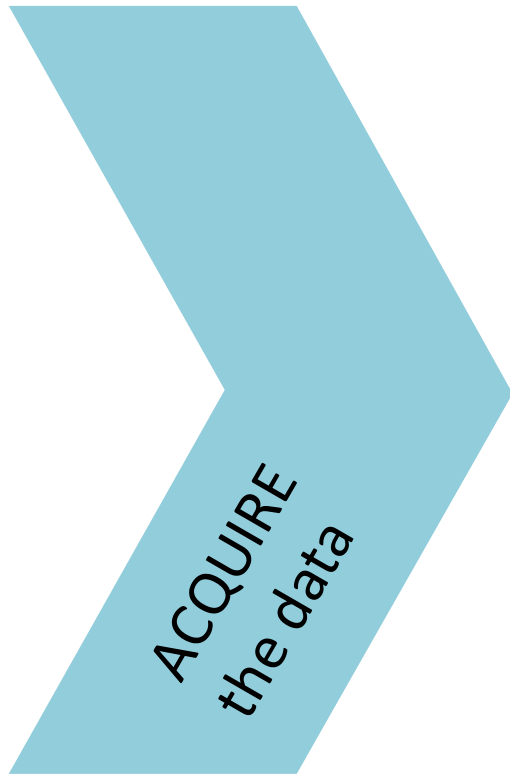
ACQUIRE
the data

‣ Acquire the Data

  ‣ Identify the "right" dataset(s)

  ‣ Import data and set up local or
  remote data structure

  ‣ Determine most appropriate
  tools to work with data

❷ ACQUIRE the Data

Data can be either unstructured or structured data
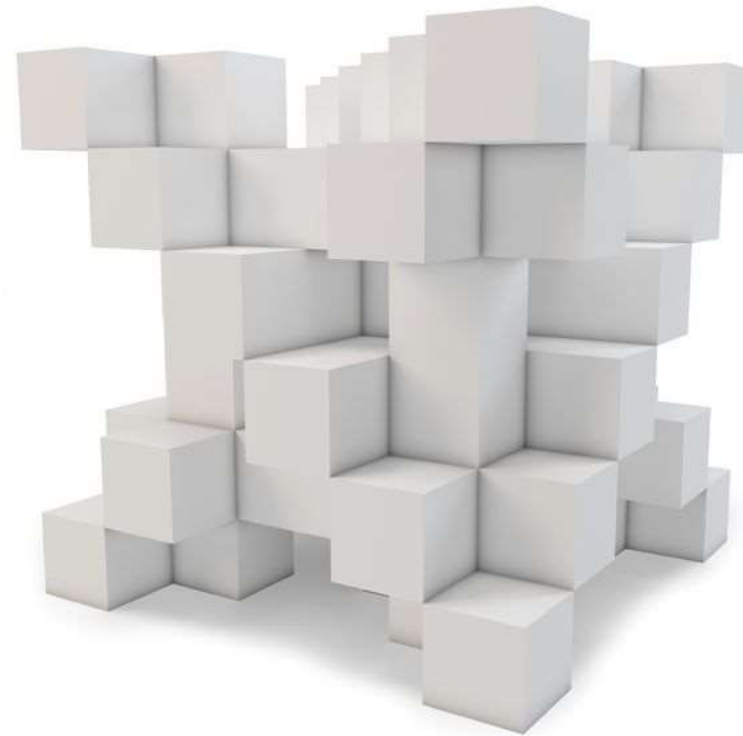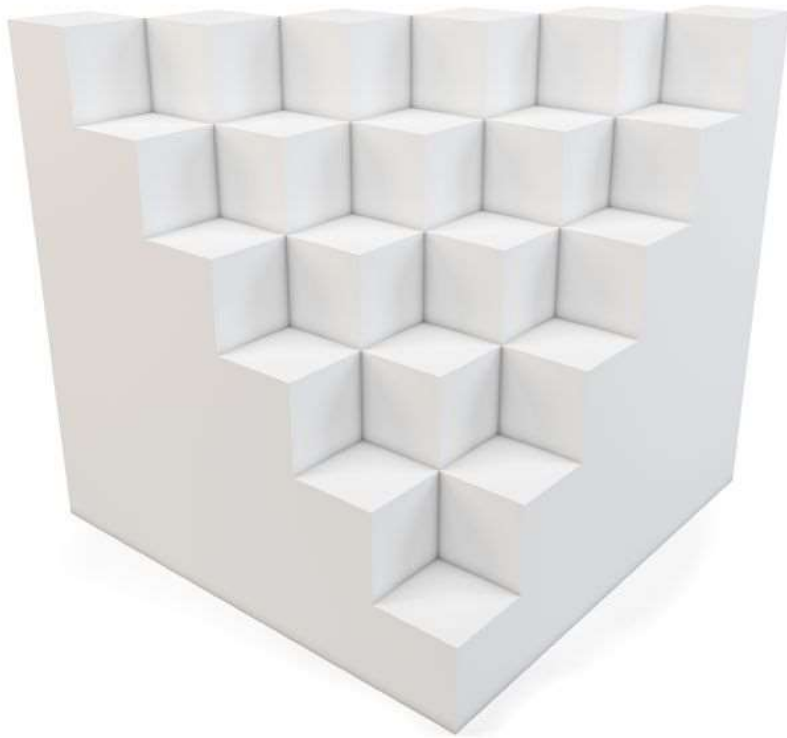
dny3d © 123RF.com

❷ ACQUIRE the Data
What's an example of unstructured data?

Bundit Chuangboonsri © 123RF.com

‣ Natural Language Processing

(session 18)

❷ ACQUIRE the Data

Most of the course will focus on structured data

‣ *k*-Nearest Neighbors *(session 5)*

‣ Linear Regression *(sessions 7–9)*

‣ Logistic Regression *(session 11)*

‣ Trees *(session 16)*

milosb © 123RF.com

❷ ACQUIRE the Data

Data can be longitudinal

Person

Khoon Lay Gan © 123RF.com
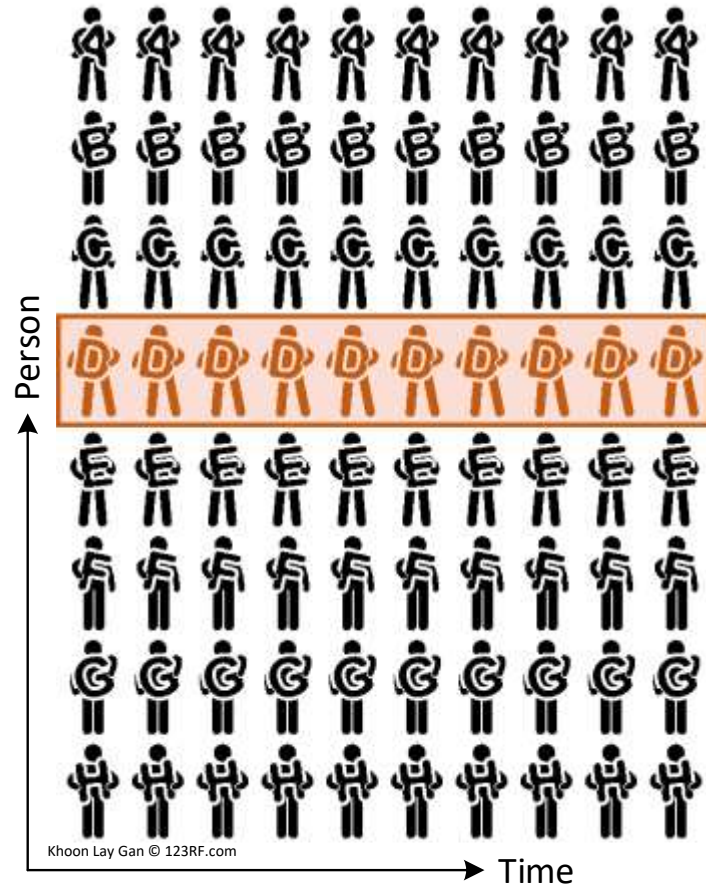
Time

‣ Time Series *(session 19)*

❷ ACQUIRE the Data

But data can also be cross-sectional

‣ And most of the course will

focus on it

Person

Time

Khoon Lay Gan © 123RF.com
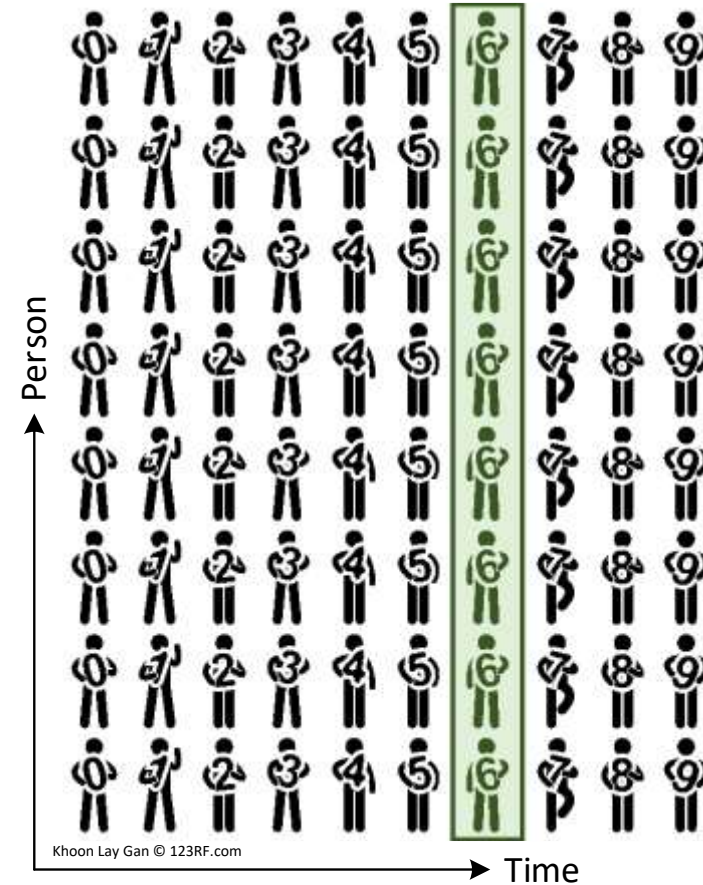
❷ ACQUIRE the Data
Raw structured data is Messy™…
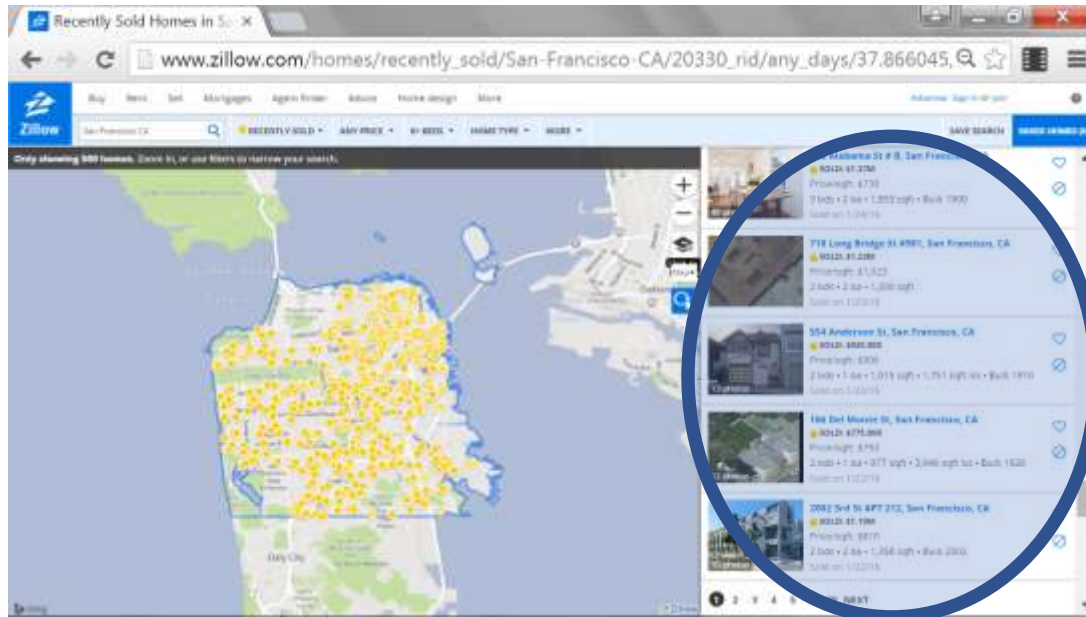
```
<div class="property-info"
id="yui_3_18_1_1_1456167242885_71870"><strong
id="yui_3_18_1_1_1456167242885_71869"><dt class="property-
address" id="yui_3_18_1_1_1456167242885_71868"><a
href="/homedetails/149-Shipley-St-San-Francisco-CA-
94107/15147894_zpid/" class="hdp-link routable" title="149
Shipley St, San Francisco, CA Real Estate"
id="yui_3_18_1_1_1456167242885_71873">149 Shipley St, San
Francisco, CA</a></dt></strong><dt class="listing-type zsg-
content_collapsed"
id="yui_3_18_1_1_1456167242885_71875"><span class="zsg-
icon-recently-sold type-icon"></span>Sold: $1.18M</dt><dt
class="zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71877">Price/sqft:
$1,116</dt><dt class="property-data"
id="yui_3_18_1_1_1456167242885_71880"><span class="beds-
baths-sqft">3 bds • 2 ba • 1,057 sqft</span><span
class="built-year" id="yui_3_18_1_1_1456167242885_71879"> •
Built 1992</span></dt><dt class="sold-date zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71975">Sold on
2/22/16</dt></div>
```
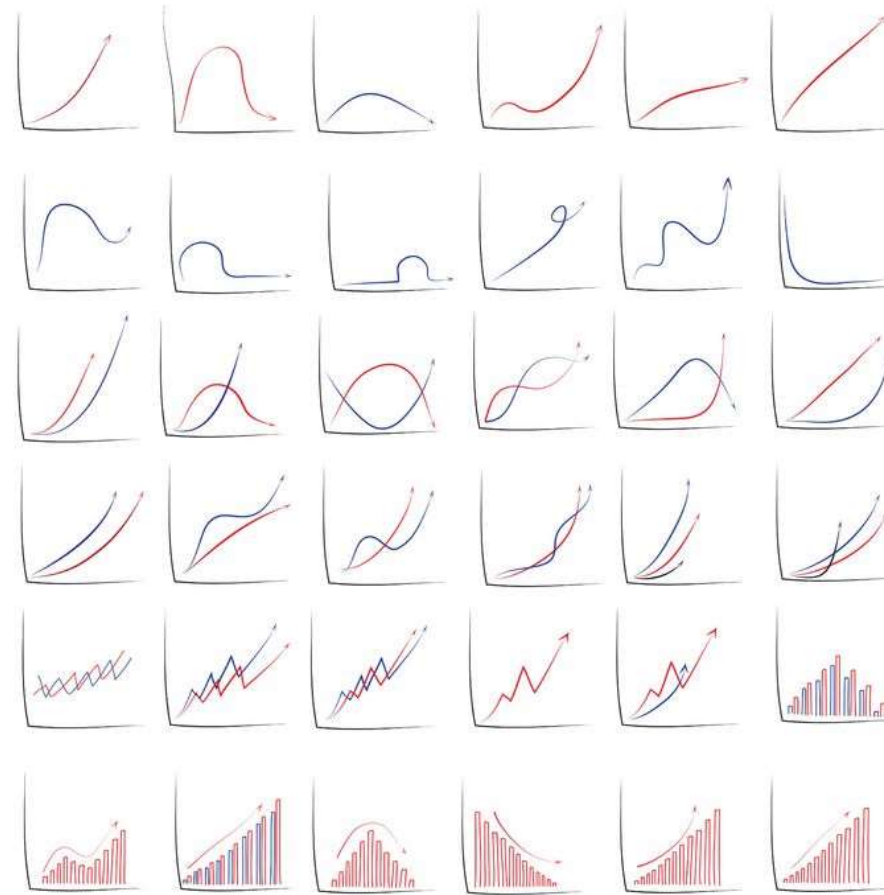
❸ PARSE the Data
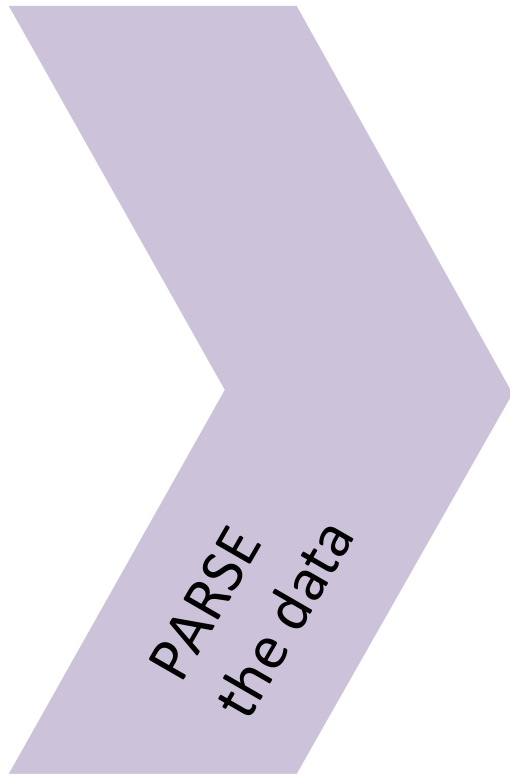
PARSE
the data

‣ Parse the Data

‣ Read any documentation provided with the data

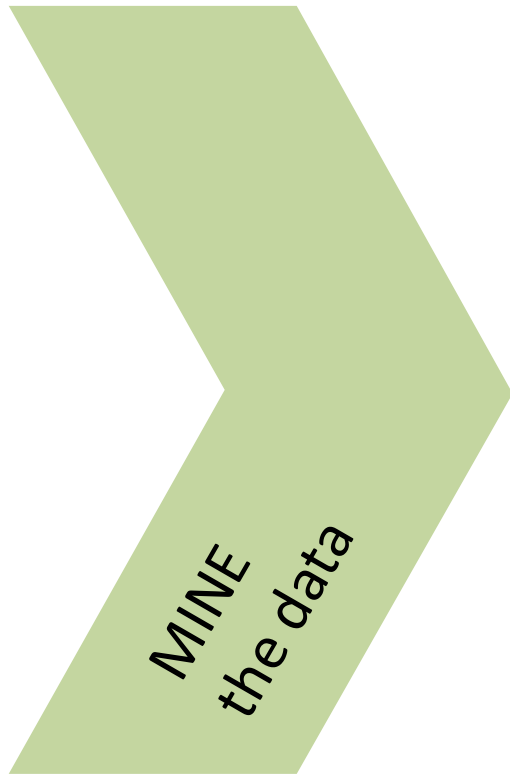‣ Perform exploratory data analysis

‣ Verify the quality of the data

❸ PARSE the Data

Exploratory Data Analysis

Napat Polchoke © 123RF.com

❹ MINE the Data

MINE the data

- ‣ Mine the Data

  - ‣ Determine sampling methodology and sample data

  - ‣ Format, clean, slice, and combine data in Python
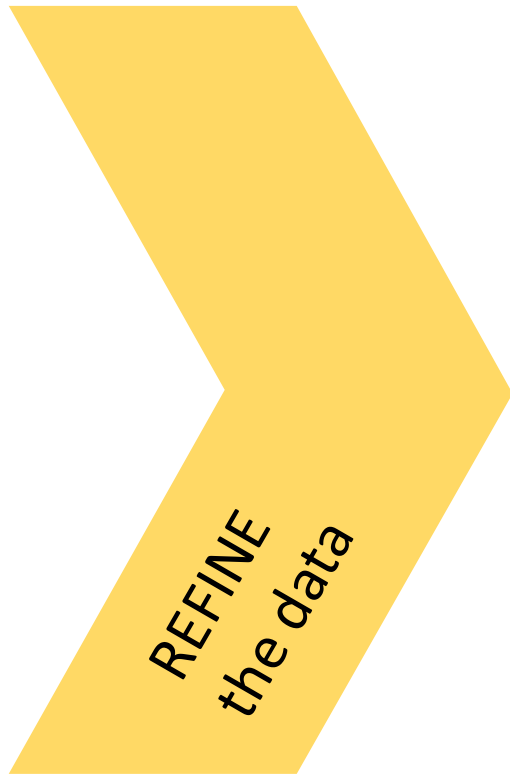
  - ‣ Create necessary derived columns from the data (new data)

❹ MINE the Data

We will be tidying our data using the *pandas* library
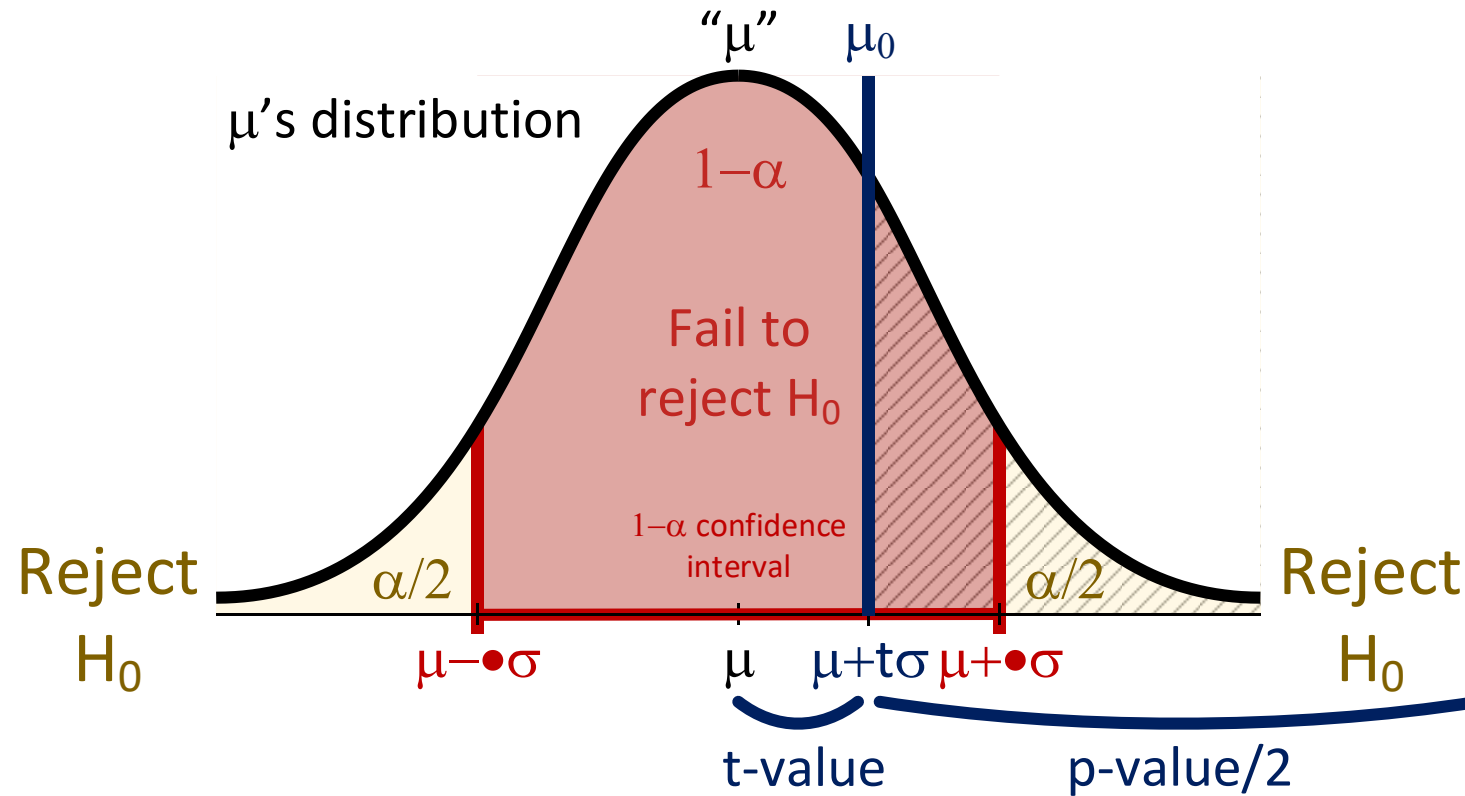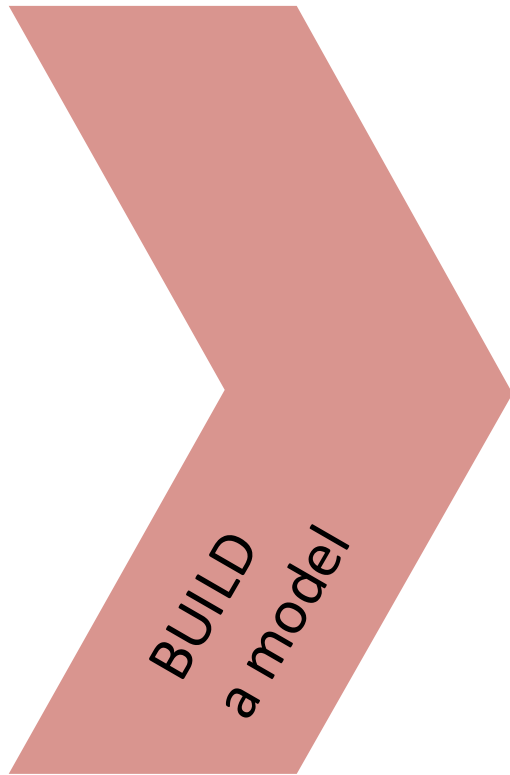
❺ REFINE the Data

REFINE the data

‣ Refine the Data

   ‣ Identify trends and outliers

   ‣ Apply descriptive and inferential statistics

   ‣ Document and transform data

❺ REFINE the Data

We will apply inferential statistics

❻ BUILD a Model

BUILD
a model

- ‣ Build a Model

  - ‣ Select appropriate model

  - ‣ Build model

  - ‣ Evaluate and refine model

❻ BUILD a Model

Types of machine learning algorithms we will study in this course

	Regression (continuous predictions)	**Classification** (categorical predictions)
Supervised a.k.a., predictive modeling (generalization; make predictions)	k-Nearest Neighbors *(session 5)* Linear Regression *(sessions 7−9)* Trees *(session 16)* Time Series *(session 19)*	k-Nearest Neighbors *(session 5)* Logistic Regression *(session 11)* Trees *(session 16)* Natural Language Processing *(session 18)*
Unsupervised *(representation; extract structure)*	Clustering *(session 14)* Natural Language Processing *(session 18)*	

❼ PRESENT the Results

PRESENT the results
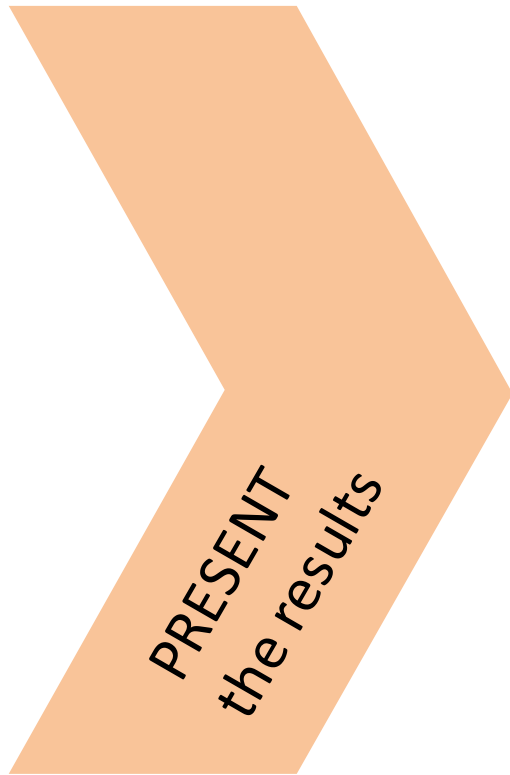
‣ Present the Results

   ‣ Summarize findings with narrative, storytelling techniques

   ‣ Present limitations and assumptions of your analysis

   ‣ Identify follow up problems and questions for future analysis
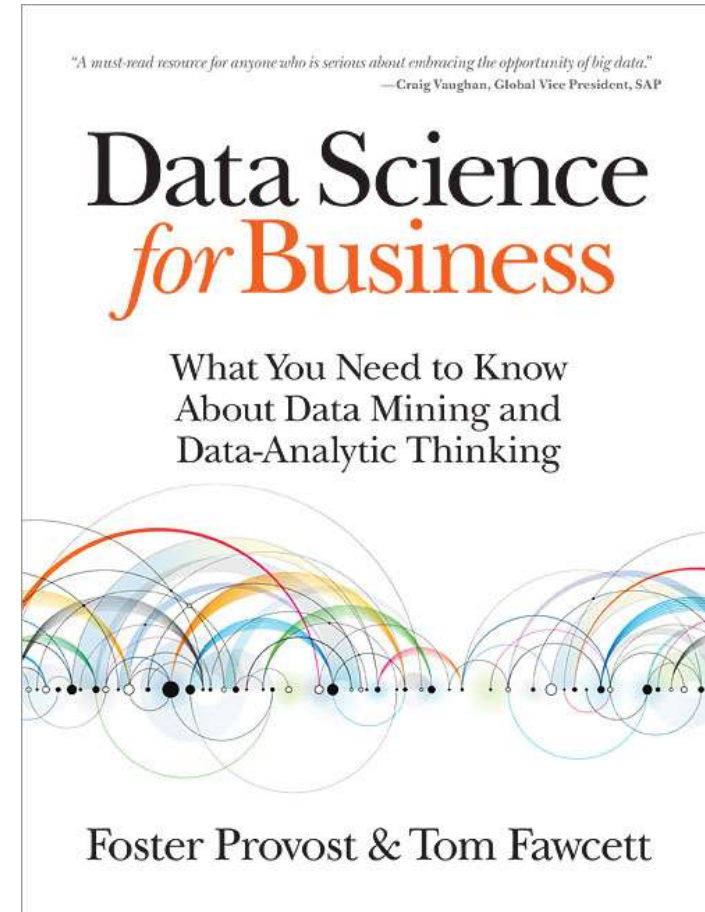
❼ PRESENT the Results

Know Your Audience

Corina Rosu © 123RF.com

A great resource to follow along the class (or afterwards...) *(will mostly reference pre-class reading materials; optional; not required but recommended for the course)*

- Data Science for Business (by Provost and Fawcett) ([link](#))

  - General Assembly holds several copies in its library

A couple of resources to follow along the class (or afterwards...) *(will mostly reference <u>post-class reading</u> materials; optional; not required for the course)*

- An Introduction to Statistical Learning: with Applications in R

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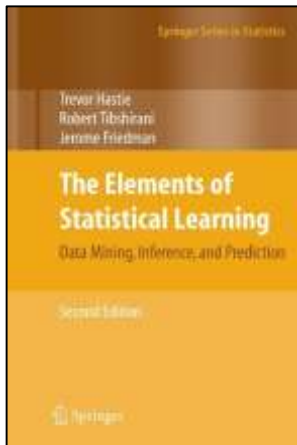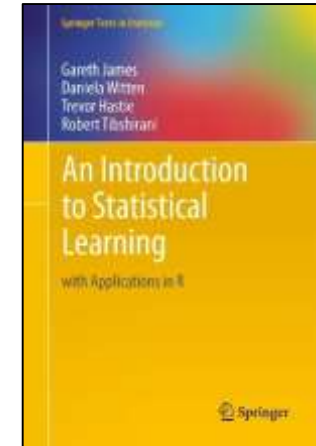
- For a more advanced treatment of these topics, check out The Elements of Statistical Learning: Data Mining, Inference, and Prediction (by Hastie et al.). The e-book is also free... ([here](#))

❶ IDENTIFY the Problem

❶ IDENTIFY the Problem

‣ **Identify the Problem**

   ‣ Identify business/product objectives

   ‣ Identify and hypothesize goals and criteria for success

   ‣ Create a set of questions for identifying correct dataset

‣ The Why's and How's of a Good Question

‣ The SMART Goals Framework

By asking a good question and setting a clear aim:

- ‣ You set yourself up for success
  - ‣ "A problem well stated is half solved" – Charles Kettering
- ‣ You help other data scientists learn from and reproduce your work
  - ‣ You establish the basis for making your analysis reproducible
- ‣ You also help them expand on your work in the future

The SMART Goals Framework for Data Science

SPECIFIC	The dataset and key variables are clearly defined
MEASURABLE	The type of analysis and major assumptions are articulated
ATTAINABLE	The question you are asking is feasible for your dataset and is not likely to be biased
REPRODUCIBLE	Another person (or you in 6 months!) can read your state and understand exactly how your analysis is performed
TIME-BOUND	You clearly state the time period and population for which this analysis will pertain

Trends often change over time and vary by the population of source of your data. It is important to clearly define who/what you included in your analysis as well as the time period for the analysis

7 PRESENT the Results

We built a model! Now what?

‣ We've built our model, but there is still a gap between our Jupyter Notebook with its plots and figures and a slideshow needed to present our results

‣ The course focus on two core concepts:

  ‣ Developing consistent practices

  ‣ Interpreting metrics to evaluate and improve model performance

‣ But what does that mean to your audience?

We built a model! Now what? (cont.)

‣ Imagine how a non-technical audience might respond to the following statements:

   ‣ "The predictive model I built has an accuracy of 80%"

   ‣ "Logistic regression was optimized with L2 regularization"

   ‣ "Gender was more important than age in the predictive model because it has a 'larger coefficient'"

   ‣ "Here's the AUC chart that shows how well the model did"

We built a model! Now what? (cont.)

‣ Who is your audience? Are they technical? What are their concerns?

  ‣ In a business setting, you may be the only person who can interpret what you've built

‣ Some people may be familiar with basic visualization, but you will likely have to do a lot of "hand holding"

‣ You need to be able to efficiently explain your results in a way that makes sense to all stakeholders (technical or not)

We built a model! Now what? (cont.)

‣ In this section, we'll focus on communicating results for "simpler" problems, but this applies to any type of model you may work with

Showing our Work

‣ We've spent a lot of time exploring our data and building a reasonable model that performs well

‣ However, if we look at our visuals, they are most likely:

‣ Statistically heavy: most people don't understand histograms

‣ Overly complicated: scatter matrices produce too much information

‣ Poorly labeled: code doesn't require adding labels, so you may not have added them

To convey important information to your audience, make sure your charts are simplified, easily interpretable, and clearly labeled

Simplified

- At most, you'll want to include figures that either explain a variable on its own or explain that variable's relationship against a target

- If your model used a data transformation (like natural log), just visualize the original data

- Remove unnecessary complexity

Easily interpretable

- Any stakeholder looking at a figure should be seeing the exact same thing you're seeing

  - A good test for this is to share the visual with others less familiar with the data and see if they come to the same conclusion

  - How long did it take them?

Clearly labeled

- Take the time to clearly label your axis, title your plot, and double check your scales – especially if the figures should be comparable

- If you're showing two graphs side by side, they should follow the same Y axis

When building visuals for another audience, ask yourself who, what, and how

Who	**What**	**How**
‣ Who is my target audience for the visual?	‣ What do they already know about this project? What do they need to know?	‣ How does my project affect this audience? How might they interpret (or misinterpret) the data?

Visualizing Models over Variables

‣ One effective way to explain your model over particular variables is to plot the predicted values against the most explanatory variables

‣ E.g., in logistic regression, plotting the probability of a class against a variable can help explain the range of effect of the model

Visualizing Performance against Baseline

‣ Another approach of visualization is the effect of your model against a baseline, or – even better – against previous models

‣ Plots like this will also be useful when talking to your peers – other data scientists or analysts who are familiar with your project and interested in the progress you've made

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