Welcome to Data Science

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

After this lesson, you should be able to:

- Describe the roles and components of a successful learning environment
- Define what is data science and who data scientists are
- Setup your development environment and practice the different workflows used in the course
- Define the data science workflow

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?
- Installfest
- An overview of the data science workflow
- Lab – Onboarding/Python Review
- Review
- Exit Tickets

Welcome to GA and DS!

Setting You Up for Success

Meet Your Team

- Ivan Corneillet, Lead Instructor

- Dan Bricarello, Associate Instructor

- Vanessa Ohta, Course Producer

Course Logistics

- Lead Instructor
  - Ivan Corneillet (ivan+GA@paspeur.com)
- Associate Instructor
  - Dan Bricarello (dabricarello@ucdavis.edu)
- Course Producer
  - Vanessa Ohta (vanessa@generalassemb.ly)
- Class
  - September 8 – November 17, Tuesdays and Thursdays, 6:30PM – 9:30PM
  - Classroom 1
- Slack
  - https://ds-sf-27.slack.com
- GitHub
  - https://github.com/ga-students/DS-SF-27
- Exit Tickets
  - http://tiny.cc/ds-sf-27

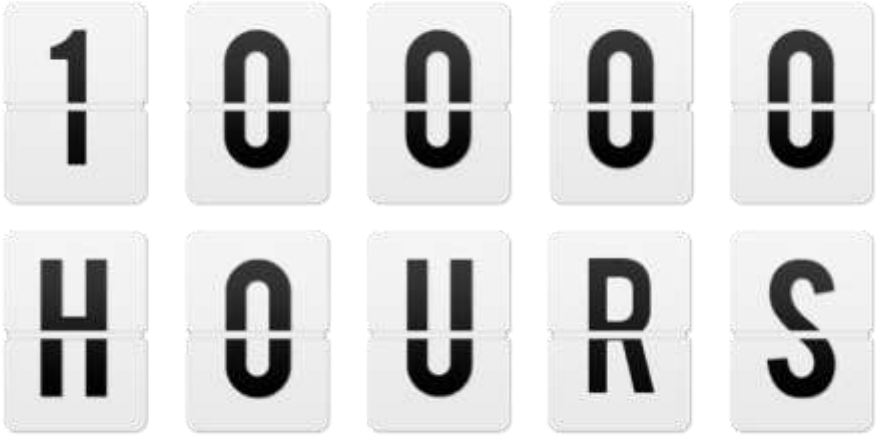
What skills will I learn in this class?

Research Design and Data Analysis	Research Design	Data Visualization in *pandas*	Statistics	Exploratory Data Analysis in *pandas*
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forests	Time Series Data	Natural Language Processing	Databases

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?

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	Logistic Modeling	Executive Summary with Findings	
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 Analysis	Notebook Draft	Final Presentation

Typical Class

- Today's objectives
- Announcements and exit tickets
- Review of the previous class

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

- Lab/Independent study
- Review of today's class
- Office hours for final projects (for the last 2-3 weeks of the course)
- Exit tickets

Setting You Up for Success

Slack (https://ds-sf-27.slack.com/)

GitHub (https://github.com/ga-students/DS-SF-27)

Exit Tickets (http://tiny.cc/ds-sf-27)

What is Data Science and Who are Data Scientists?

Activity | What is Data Science and Who are Data Scientists?

EXERCISE

DIRECTIONS (10 minutes)

1. What is data science? What are its applications? Why now? What's next?
2. Who are data scientists? How do they add value? What makes a good data scientist?
3. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

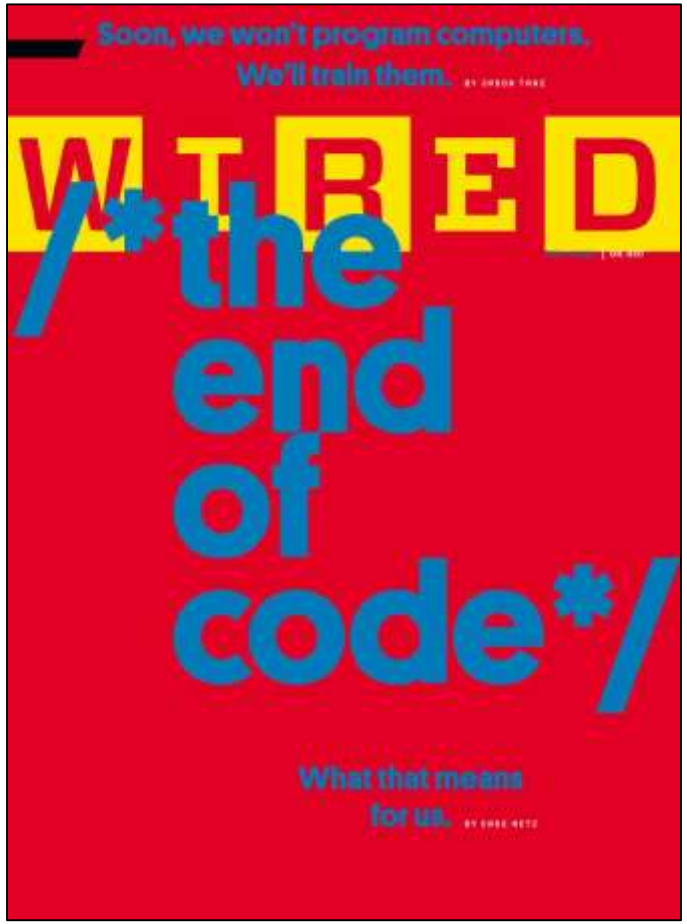
Harvard Business Review | “Data Scientists: The Sexiest Job of the 21st Century” (2012)

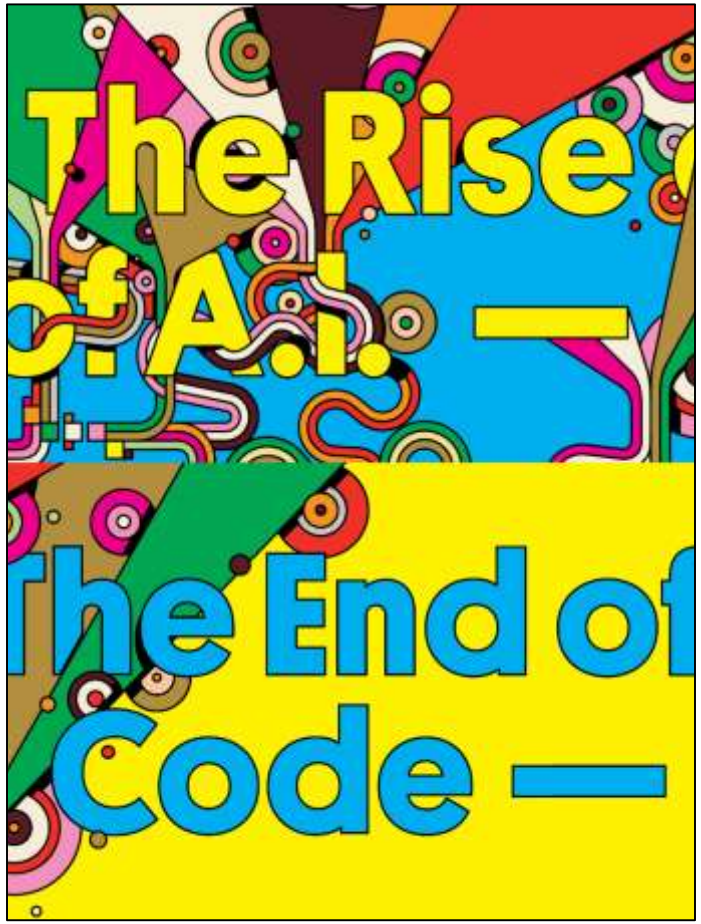
(https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/)

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/)





Source: Wired

Data science is everywhere

FiveThirtyEight

NETFLIX

Walmart

TARGET

amazon

Google

facebook

UBER

IBM Watson

LinkedIn

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
  - Linear Regression (sessions 6 and 7)
  - k-Nearest Neighbors (session 8)
  - Regression Decision Trees/Random Forests (session 12)

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 8)
  - Logistic Regression (session 9)
  - Classification Decision Trees/Random Forests (session 12)

Common questions asked in data science (cont.)

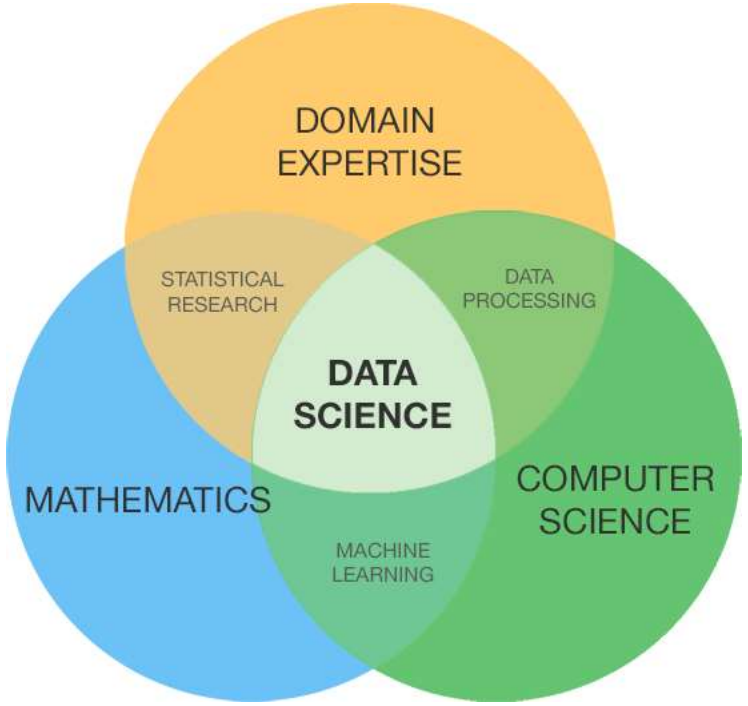
How is this Data Organized?

- 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

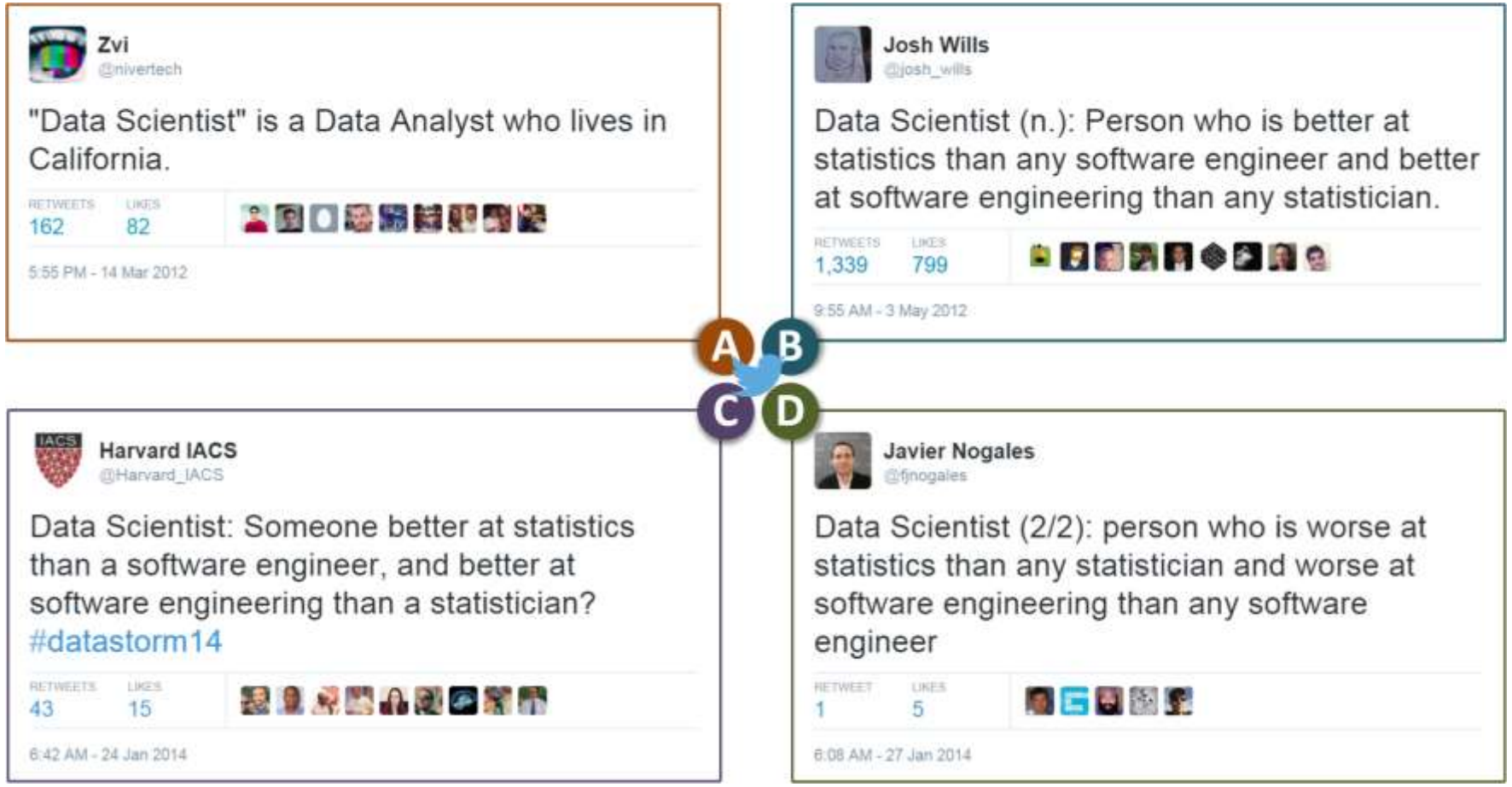
- What are the “categories” within the data?

Data science involves a variety of skillsets



Source: Data Science for the C-suite

Data scientists in ≤140 characters



Source: Twitter

Installfest

“GA” User and GitHub Desktop

Installfest

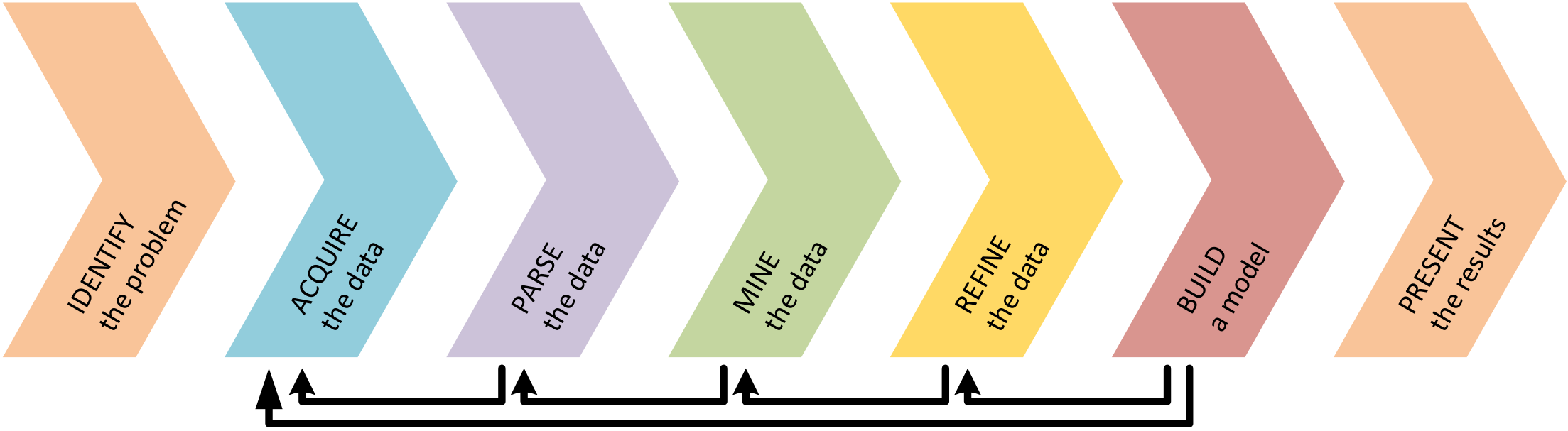
Continuum's Anaconda (Python 2.7)

Practices

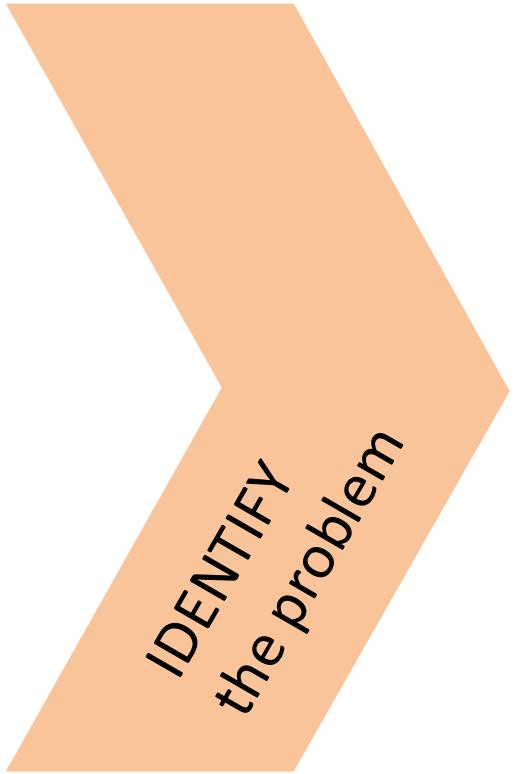
Git, GitHub, and Jupyter Notebook

Data Science Workflow

The Data Science Workflow



❶ Identify the Problem



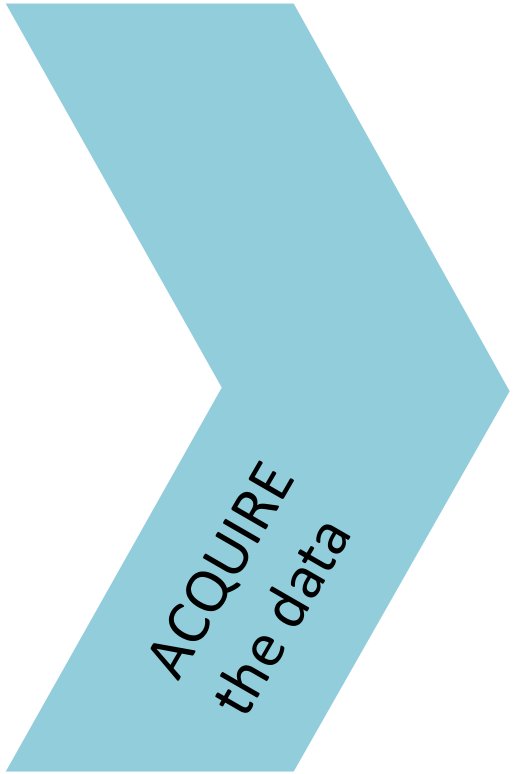
- Identify the Problem
  - Identify business/product objectives
  - Identify and hypothesize goals and criteria for success
  - Create a set of questions for identifying correct dataset

❶ Identify the Problem

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

Corina Rosu © 123RF.com

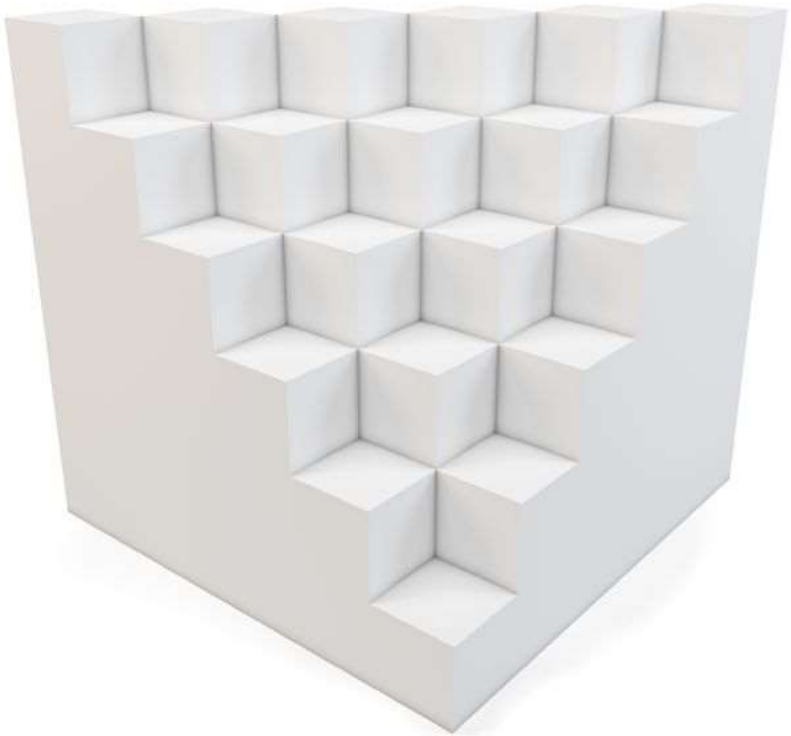
❷ Acquire the Data

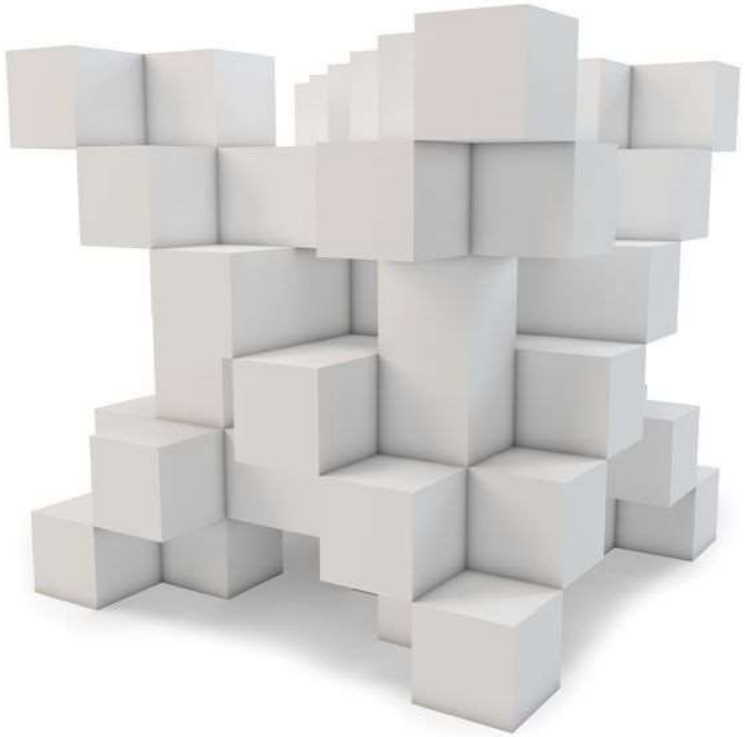


- Acquire the Data
  - Identify the “right” dataset(s)
  - Import data and set up local or remote data structure
  - Determine most appropriate tools to work with data

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?

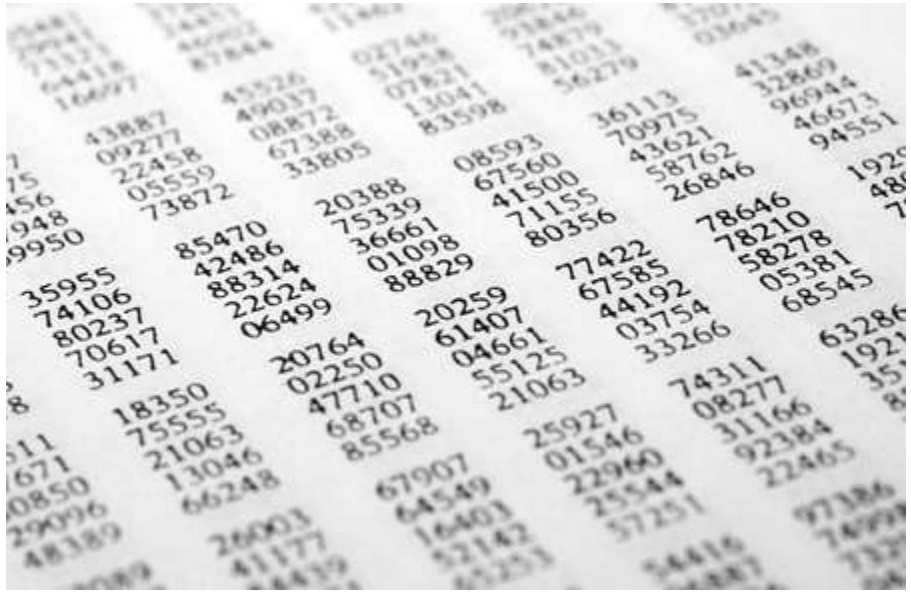
Bundit Chuangboonsri © 123RF.com

- Session 15 in Unit 3
  - Natural Language Processing

❷ Acquire the Data

Most of the course will focus on structured data

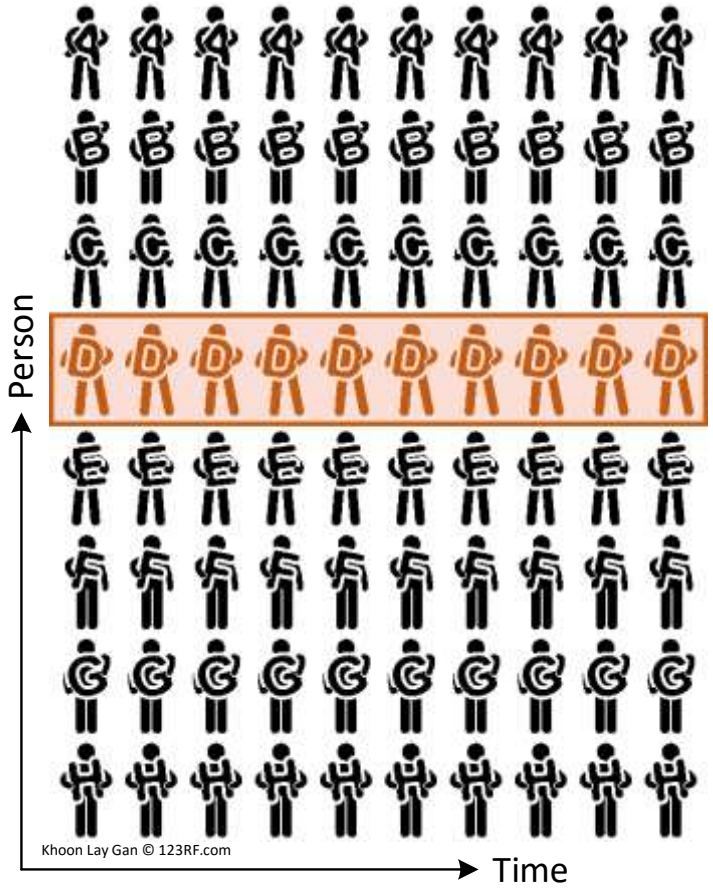
- Unit 2
  - Linear Regression (sessions 6 and 7)
  - k-Nearest Neighbors and Logistic Regression (session 8 and 9)
- Unit 3
  - Decision Trees and Random Forests (session 12)



milosb © 123RF.com

❷ Acquire the Data

Unstructured data can be longitudinal

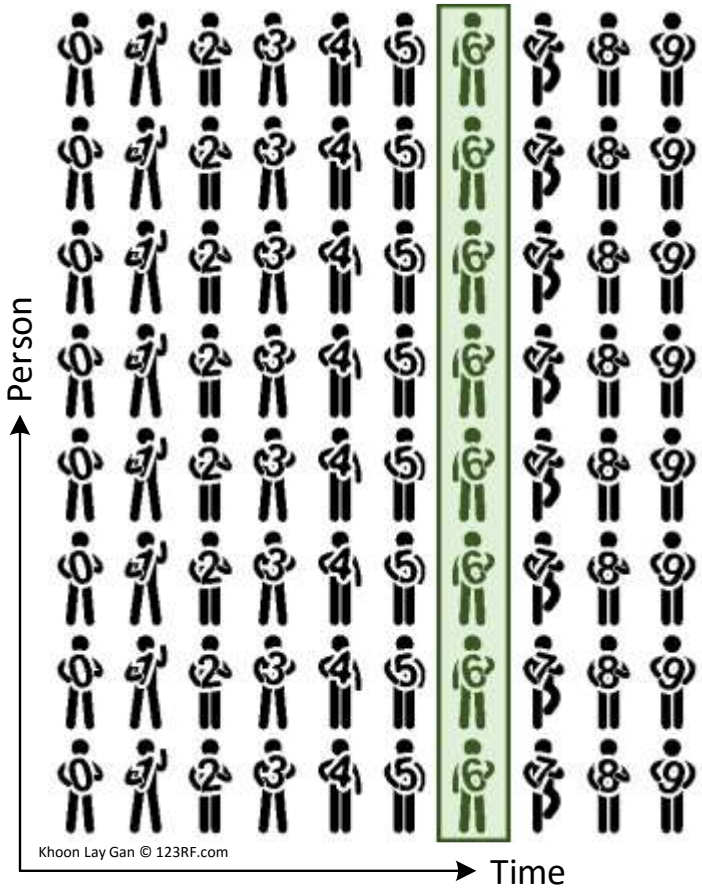


- Session 14 in Unit 3
  - Time Series

❷ Acquire the Data

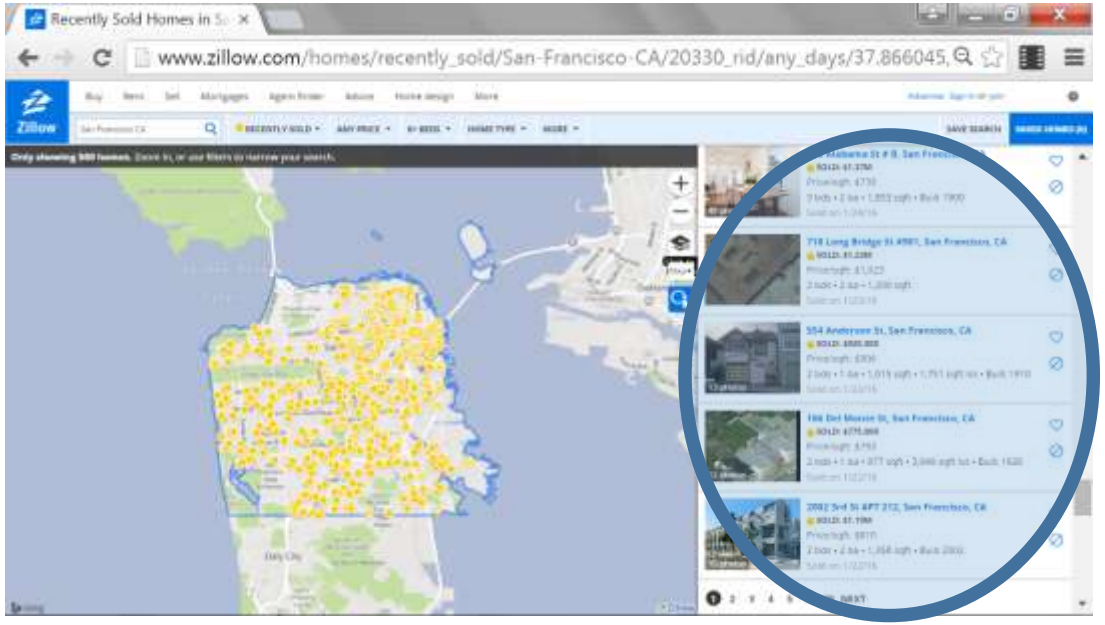
Unstructured data can be cross-sectional

- And most of the course will focus on it



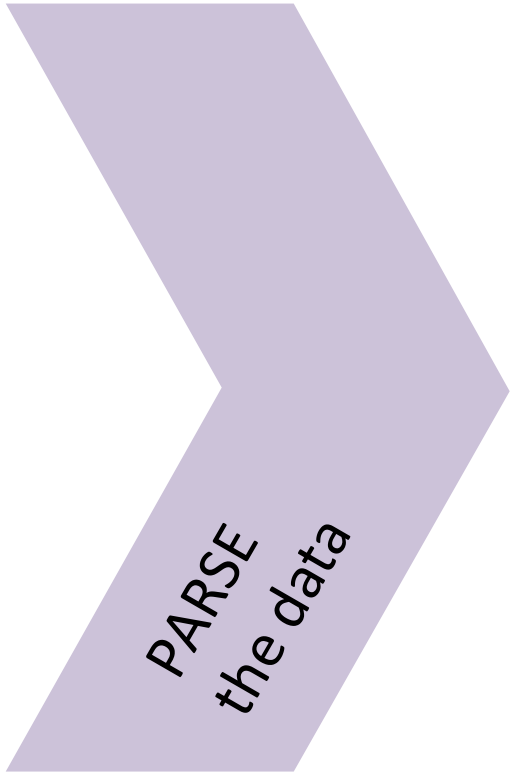
❷ Acquire the Data

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>
```

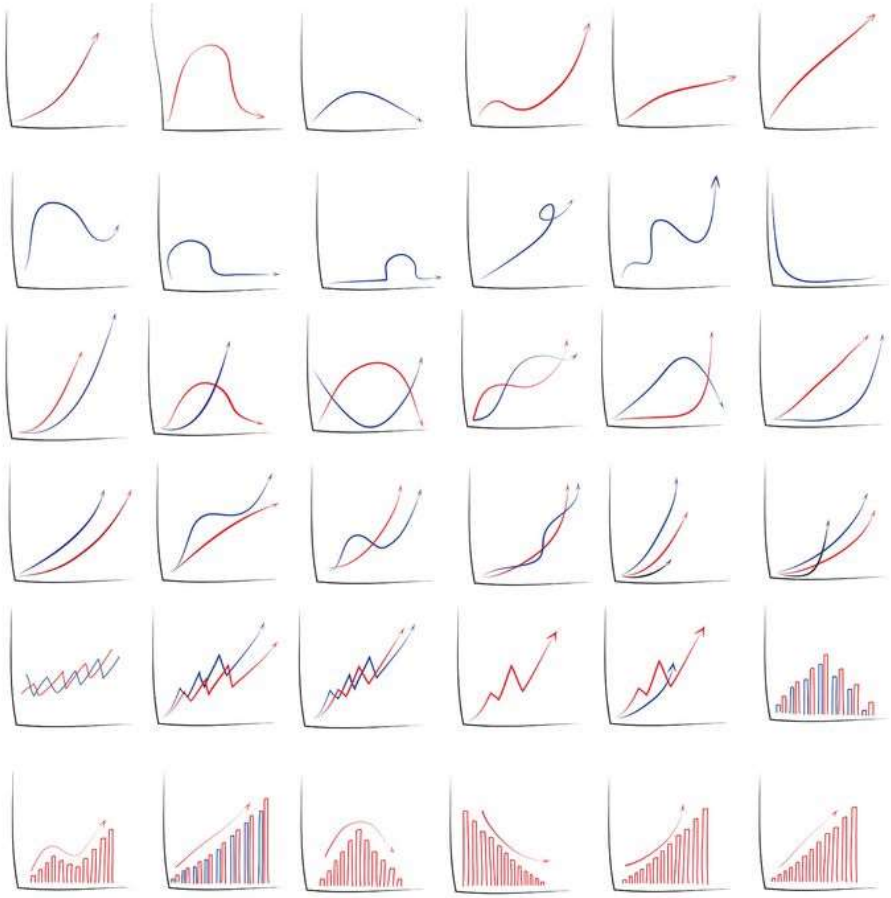
❸ Parse the Data



- Parse the Data
  - Read any documentation provided with the data
  - Perform exploratory data analysis
  - Verify the quality of the data

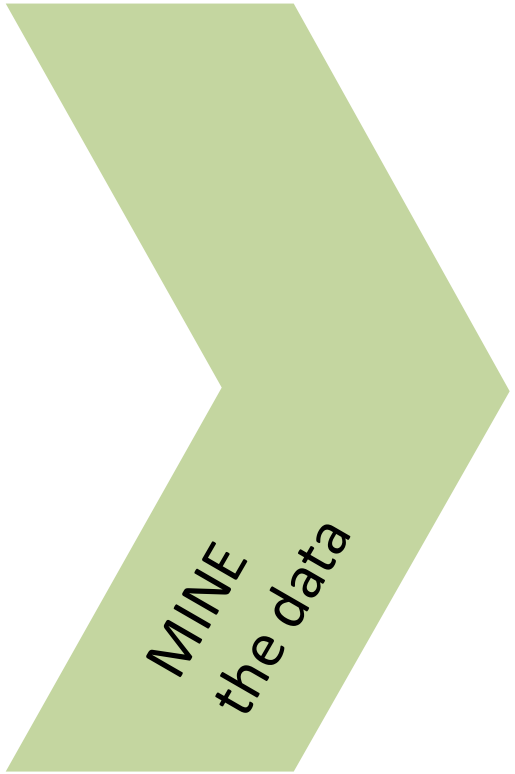
❸ Parse the Data

Exploratory Data Analysis



Napat Polchoke © 123RF.com

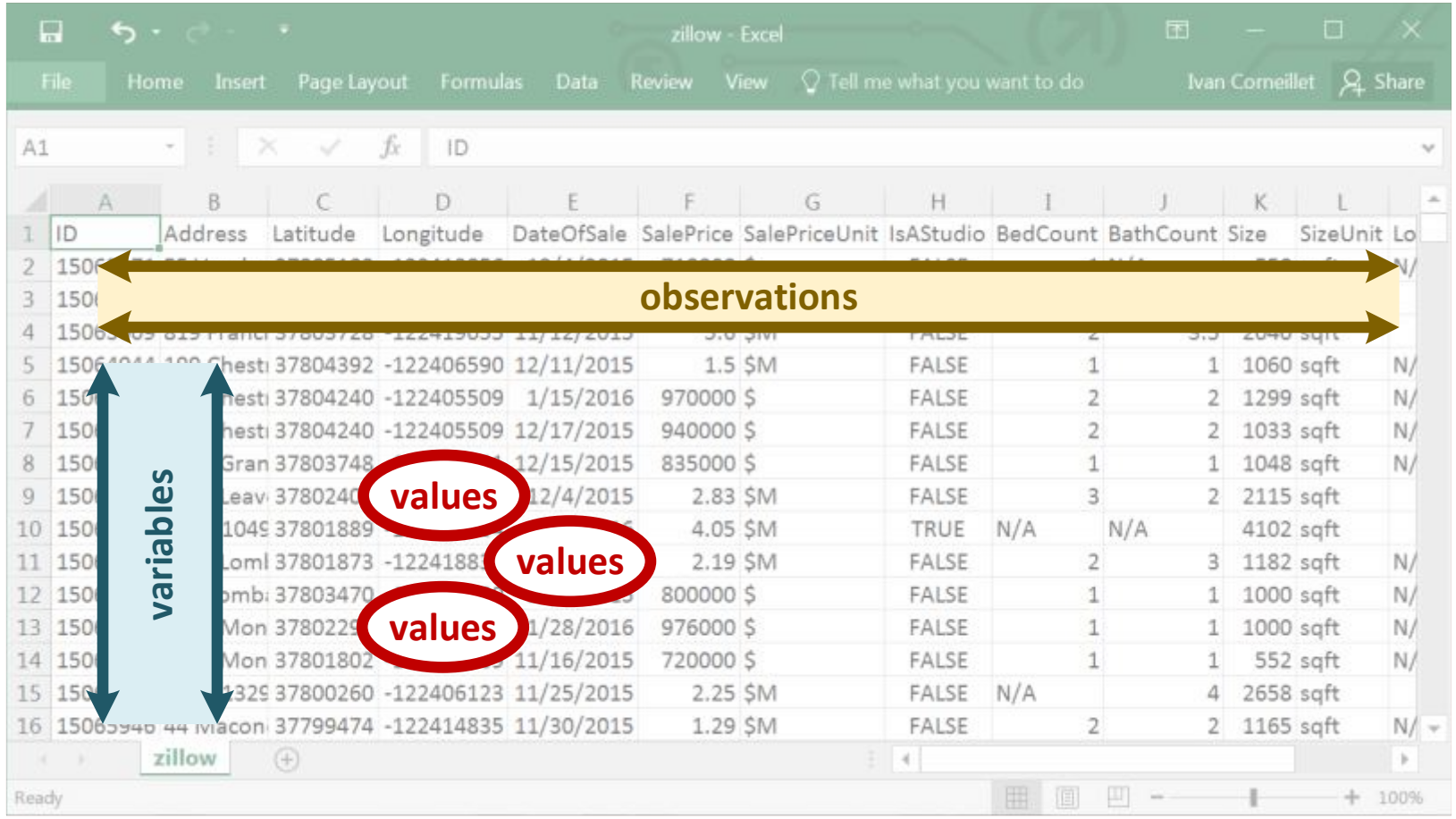
❹ Mine the Data



- Mine the Data
  - Determine sampling methodology and sample data
  - Format, clean, slice, and combine data in Python
  - Create necessary derived columns from the data (new data)

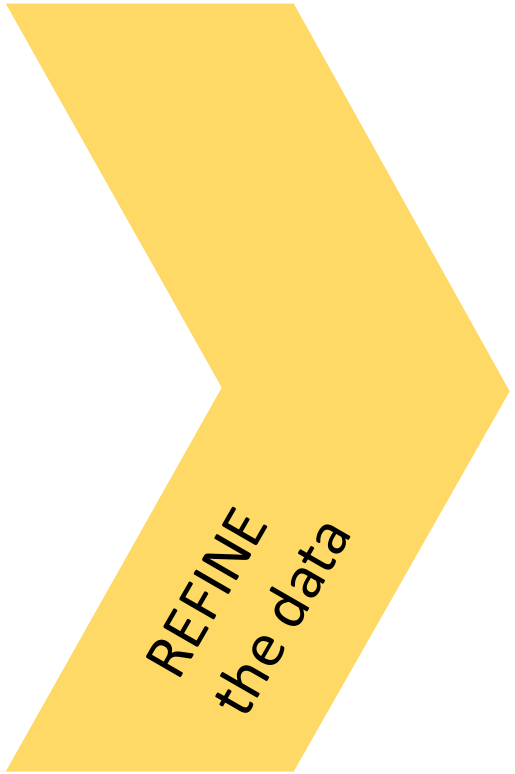
❹ Mine the Data

We will be tidying our data using the *pandas* library



ID	Address	Latitude	Longitude	DateOfSale	SalePrice	SalePriceUnit	IsAStudio	BedCount	BathCount	Size	SizeUnit	Lo
150												N/
150												
150												
150	Chesti	37804392	-122406590	12/11/2015	1.5	$M	FALSE	1	1	1060	sqft	N/
150	hesti	37804240	-122405509	1/15/2016	970000	$	FALSE	2	2	1299	sqft	N/
150	hesti	37804240	-122405509	12/17/2015	940000	$	FALSE	2	2	1033	sqft	N/
150	Gran	37803748		12/15/2015	835000	$	FALSE	1	1	1048	sqft	N/
150	Leav	3780240		12/4/2015	2.83	$M	FALSE	3	2	2115	sqft	
150	1049	37801889			4.05	$M	TRUE	N/A	N/A	4102	sqft	
150	Loml	37801873	-1224188		2.19	$M	FALSE	2	3	1182	sqft	N/
150	mbi	37803470			800000	$	FALSE	1	1	1000	sqft	N/
150	Mon	3780229		1/28/2016	976000	$	FALSE	1	1	1000	sqft	N/
150	Mon	37801802		11/16/2015	720000	$	FALSE	1	1	552	sqft	N/
150	1329	37800260	-122406123	11/25/2015	2.25	$M	FALSE	N/A	4	2658	sqft	
150	Macon	37799474	-122414835	11/30/2015	1.29	$M	FALSE	2	2	1165	sqft	N/

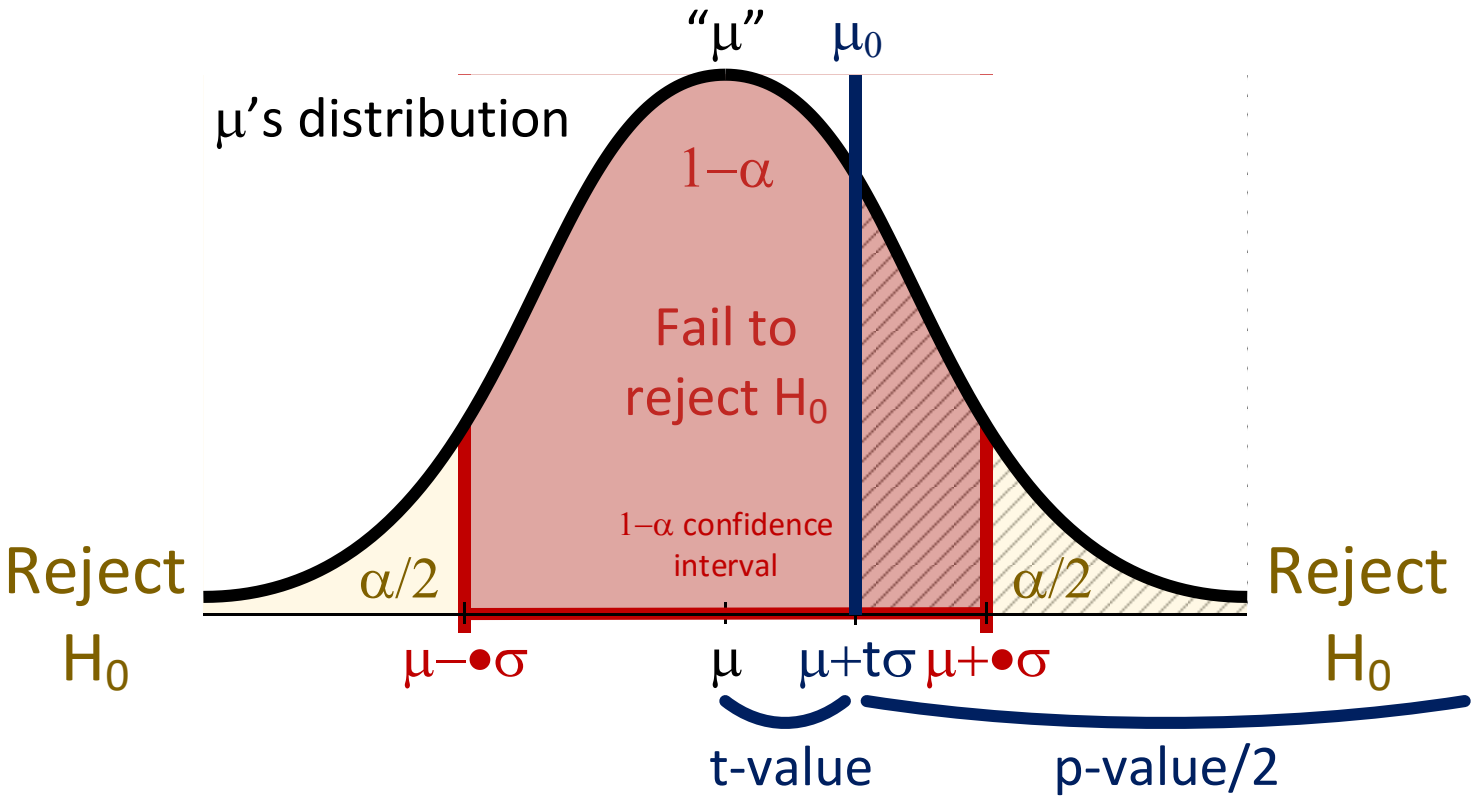
❺ Refine the Data



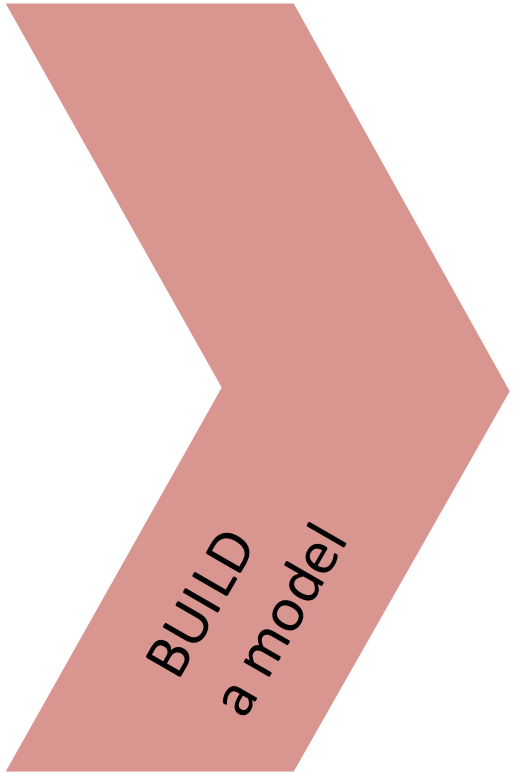
- Refine the Data
  - Identify trends and outliers
  - Apply descriptive and inferential statistics
  - Document and transform data

❺ Refine the Data

We will apply inferential statistics



❻ Build a Model



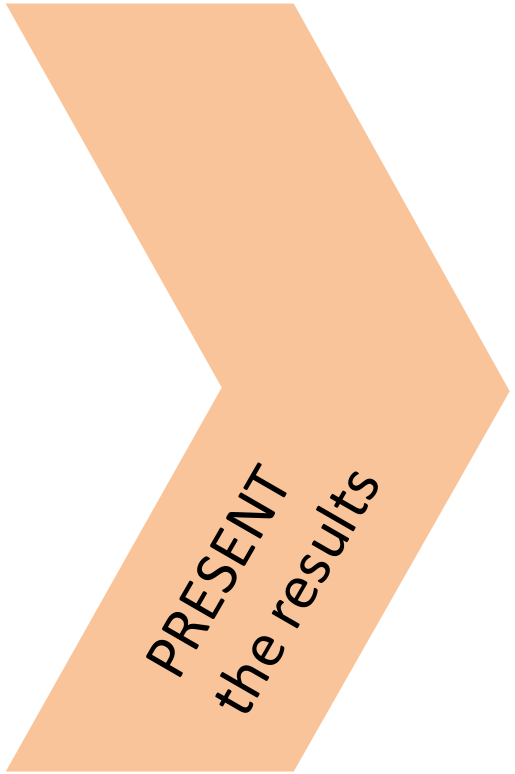
- Build a Model
  - Select appropriate model
  - Build model
  - Evaluate and refine model

❻ Build a Model

Types of machine learning algorithms we will study in this course (+ NLP)

	Continuous	Categorical
Supervised (a.k.a., predictive modeling)	Linear Regression *(sessions 6 and 7)* k-Nearest Neighbors *(session 8)* Regression Decision Trees/Random Forests *(session 12)* Time Series *(session 14)*	k-Nearest Neighbors *(session 8)* Logistic Regression *(session 9)* Classification Decision Trees/Random Forests *(session 12)*
Unsupervised	*A machine learning model that doesn't use labeled data is called unsupervised. It extract structure from the data. Goal is "representation"*	

❼ 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

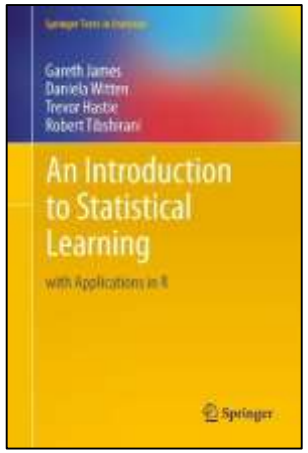
❼ Present the Results

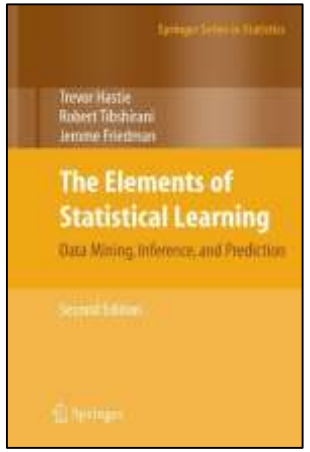
Know Your Audience

Corina Rosu © 123RF.com

Some great resources to follow along the class (or afterwards) *(optional; not required for the course)*

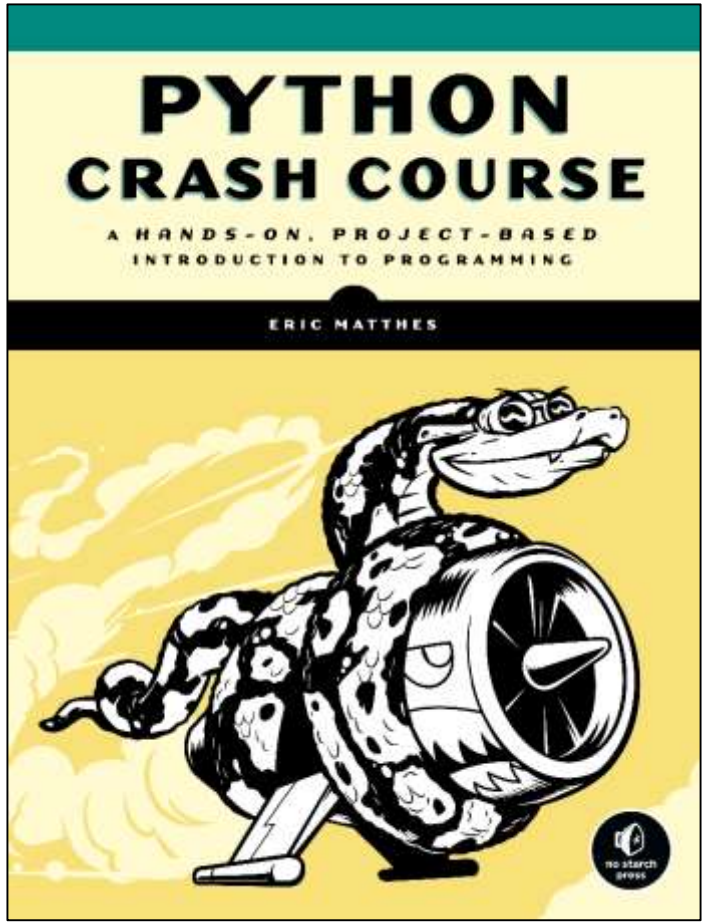
- An Introduction to Statistical Learning: with Applications in R (by James et al.). The e-book is available free-of-charge here

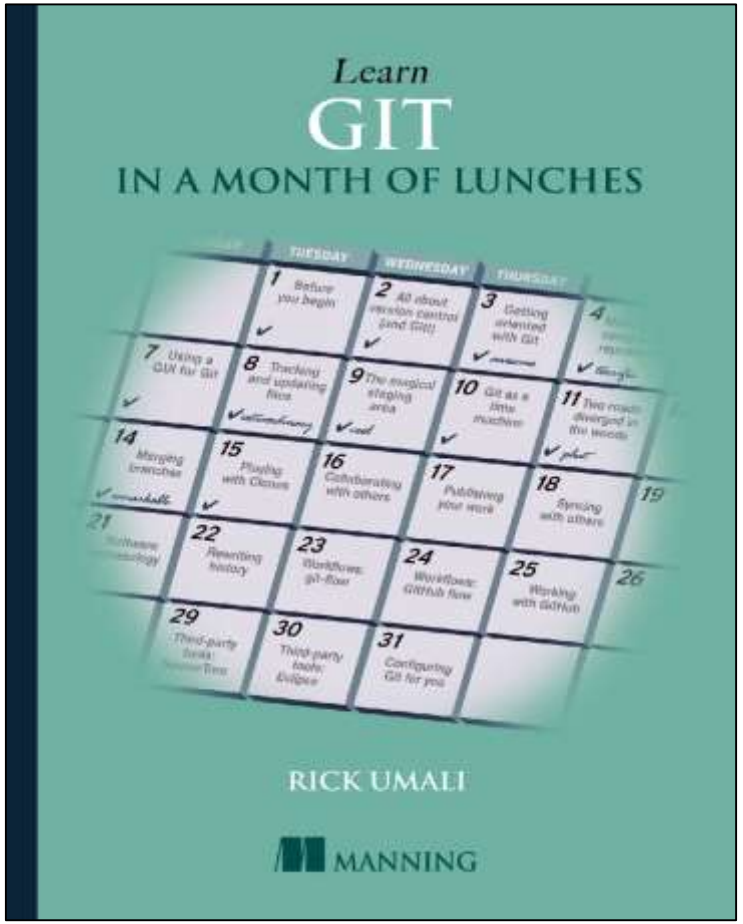




- For a more advanced treatment of these topics, check out The Elements of Statistical Learning: Data Mining, Inference, and Prediction (by Hastie et al.). The e-book is also free… (here)

A couple of resources to get started with Python and Git *(optional; not required for the course)*





Lab – Onboarding/Python Review

Review

Before Next Class

Before Next Class

- Complete your development environment setup; complete the onboarding pre-work and practice the different workflows that we will use in this course
- Look into the first unit project and start ideating about your final project's topic
- Read the two articles briefly mentioned in class, we will discuss then further in the next class:
  - 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/)
  - Wired | "The Rise of Artificial Intelligence and the End of Code" (2016) (http://www.wired.com/2016/05/the-end-of-code/)

Next Class

Research Design and pandas

Learning Objectives

After the next lesson, you should be able to:

- Define a problem and types of data
- Identify dataset types
- Apply the data science workflow in the *pandas* context
- Write an Jupyter notebook to import, format, and clean data using the *pandas* library

Exit Ticket

Don’t forget to fill out your exit ticket here