DS

What is Data Science?

Data Scientists: 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
92 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

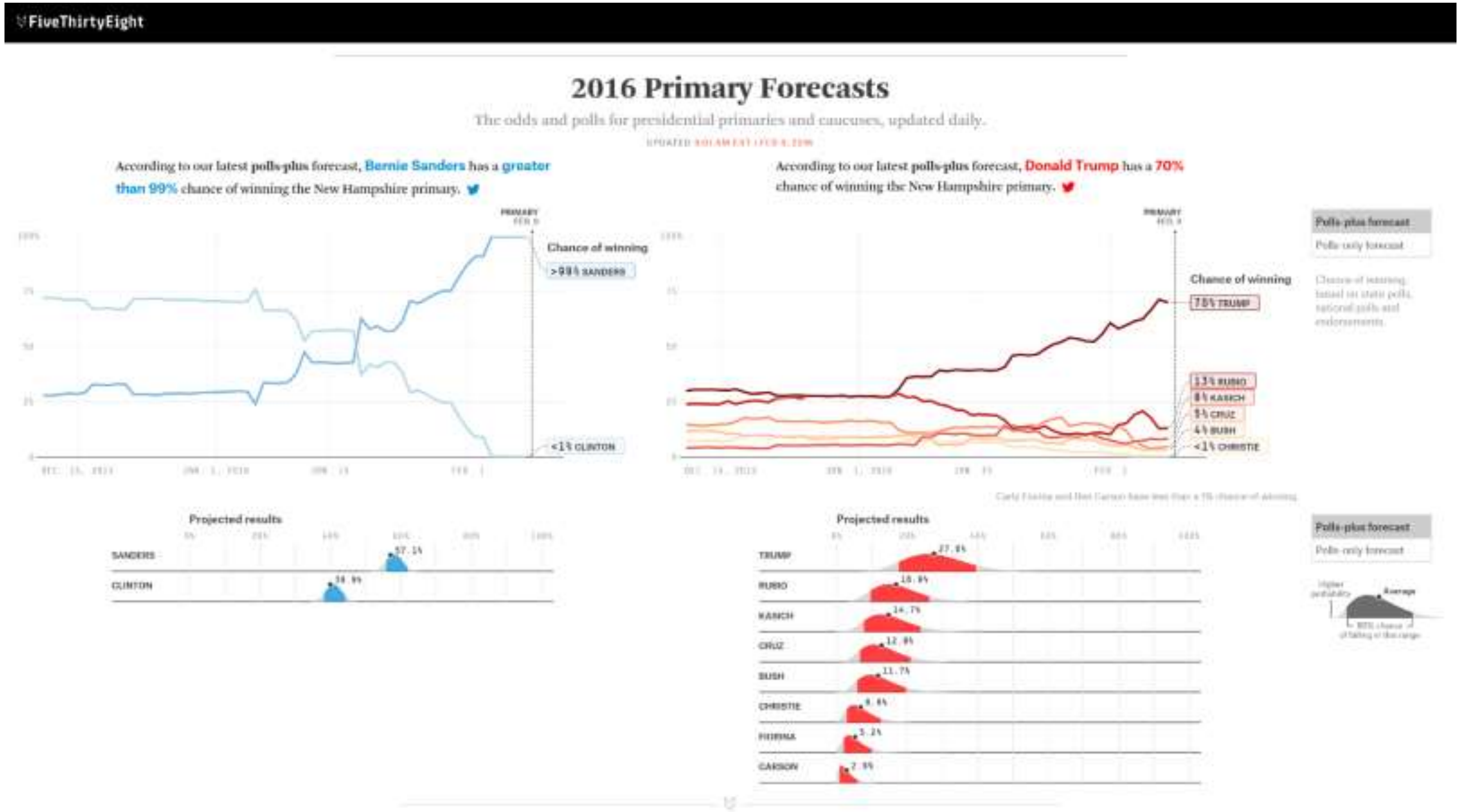
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking 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."

2 Harvard Business Review October 2012

Source: Harvard Business Review

FiveThirtyEight

Data Science by FiveThirtyEight



FiveThirtyEight

2016 Primary Forecasts

The odds and polls for presidential primaries and caucuses, updated daily.

According to our latest polls-plus forecast, Bernie Sanders has a greater than 99% chance of winning the New Hampshire primary.

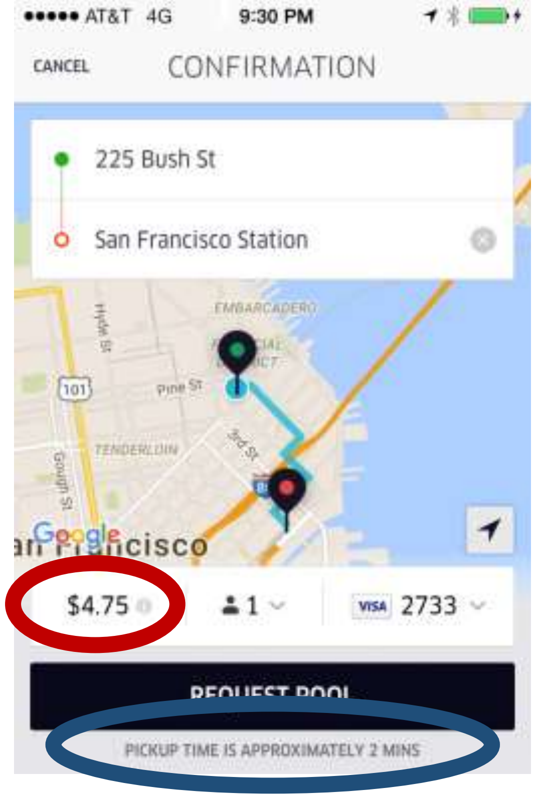
According to our latest polls-plus forecast, Donald Trump has a 70% chance of winning the New Hampshire primary.

Source: FiveThirtyEight

Uber

Data Science by Uber

For Riders…



●●●●● AT&T 4G 9:30 PM

CANCEL CONFIRMATION

225 Bush St

San Francisco Station

$4.75

1

VISA 2733

REQUEST POOL

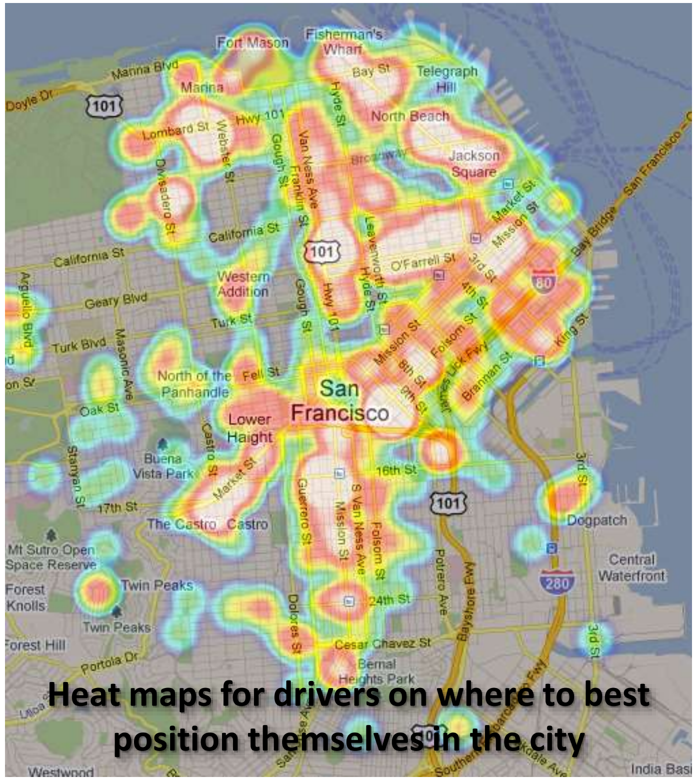
PICKUP TIME IS APPROXIMATELY 2 MINS

Fare Estimator and Dynamic Pricing (e.g., Surge Pricing) Algorithms

ETA (Estimated Time of Arrival) Algorithms

UBER

and Drivers!

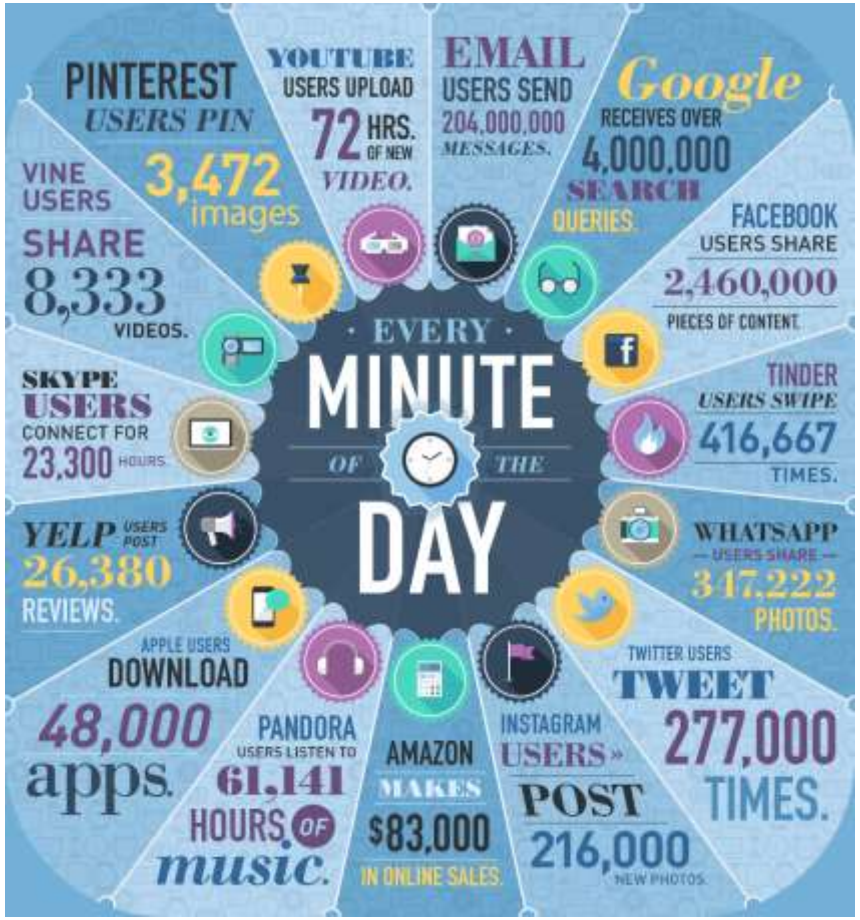


Heat maps for drivers on where to best position themselves in the city

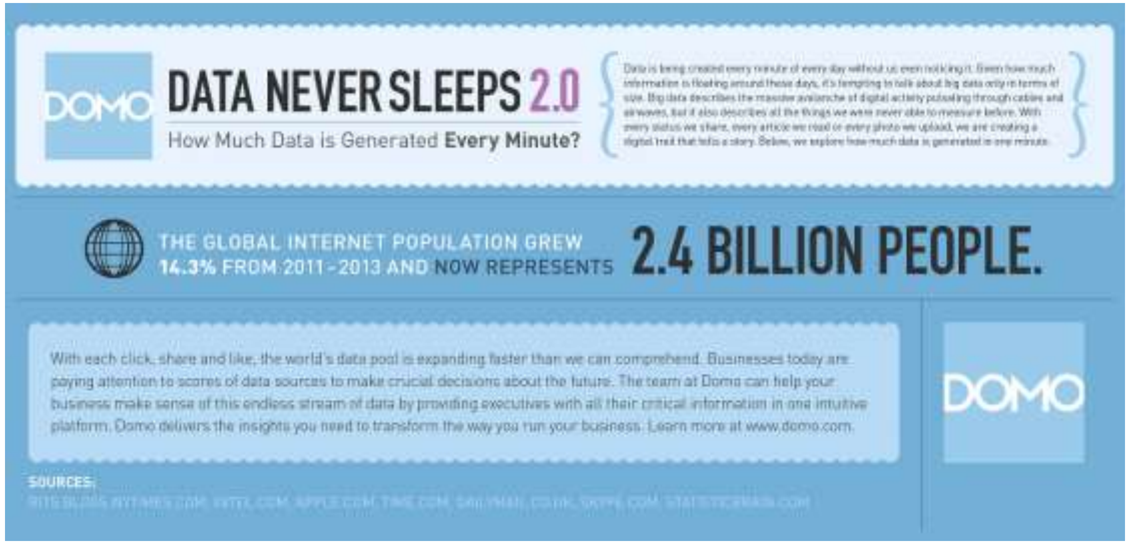
Sources: The Uber iPhone App/Wired

Data Science Based-Business Models is the New Normal

FiveThirtyEight
NETFLIX
amazon
ups
Walmart
TARGET
facebook
Google
UBER
IBM Watson
Linkedin



EVERY MINUTE OF THE DAY
PINTEREST USERS PIN 3,472 images
YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.
EMAIL USERS SEND 204,000,000 MESSAGES.
Google RECEIVES OVER 4,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.
TINDER USERS SWIPE 416,667 TIMES.
WHATSAPP USERS SHARE 347,222 PHOTOS.
TWITTER USERS TWEET 277,000 TIMES.
INSTAGRAM USERS» POST 216,000 NEW PHOTOS.
AMAZON MAKES $83,000 IN ONLINE SALES.
PANDORA USERS LISTEN TO 61,141 HOURS OF music.
APPLE USERS DOWNLOAD 48,000 apps.
YELP USERS POST 26,380 REVIEWS.
SKYPE USERS CONNECT FOR 23,300 HOURS.
VINE USERS SHARE 8,333 VIDEOS.



DOMO

DATA NEVER SLEEPS 2.0

How Much Data is Generated **Every Minute?**

Data is being created every minute of every day without us even noticing it. Given how much information is floating around these days, it's tempting to talk about big data only in terms of size. Big data describes the massive avalanche of digital activity pulsating through cables and airwaves, but it also describes all the things we were never able to measure before. With every status we share, every article we read or every photo we upload, we are creating a digital trail that tells a story. Below, we explore how much data is generated in one minute.

THE GLOBAL INTERNET POPULATION GREW 14.3% FROM 2011 - 2013 AND NOW REPRESENTS **2.4 BILLION PEOPLE.**

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. Learn more at www.domo.com.

DOMO

SOURCES:
BITS.BLOGS.NYTIMES.COM, INTEL.COM, APPLE.COM, TIME.COM, DAILYMAIL.CO.UK, SKYPE.COM, STATISTICBRAIN.COM

Source: DOMO

DS

Who are Data Scientists?

Activity: Who are Data Scientists?

EXERCISE

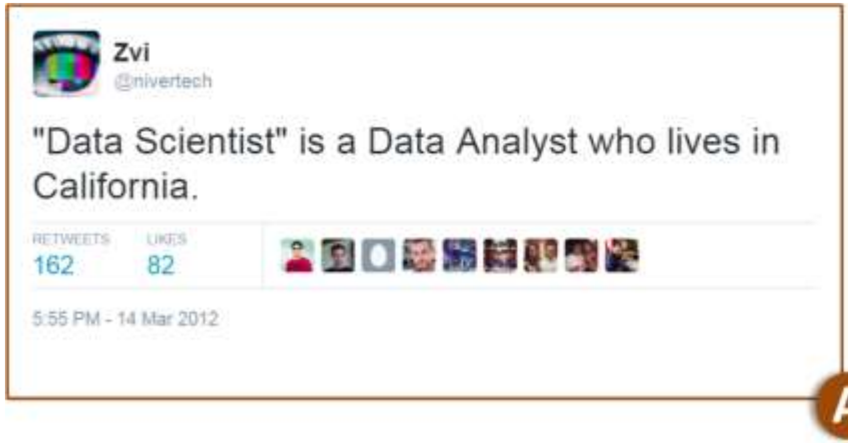
DIRECTIONS (10 minutes)

1. Who are Data Scientists?
2. How do Data Scientists add value?
3. What makes a good Data Scientist?
4. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Data Scientists in ≤140 characters



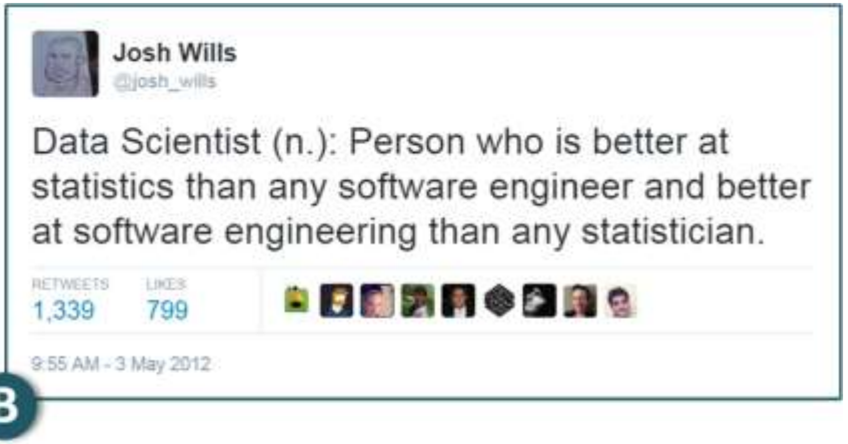
A

Zvi
@nivertech

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

RETWEETS 162 LIKES 82

5:55 PM - 14 Mar 2012



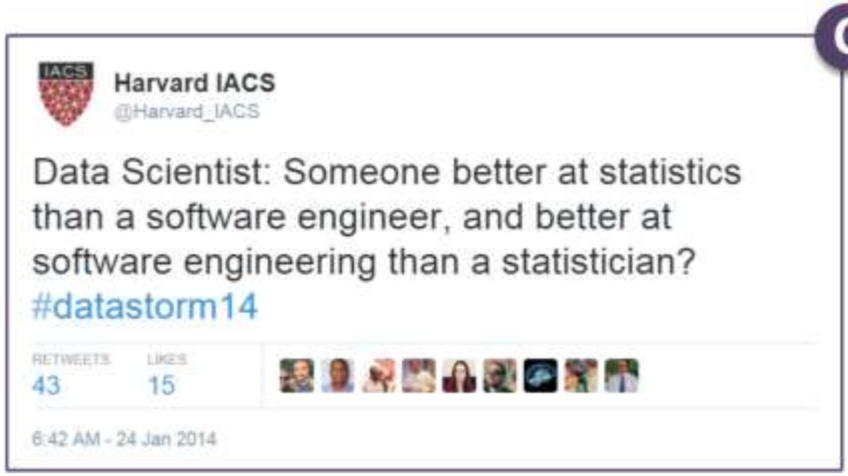
B

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



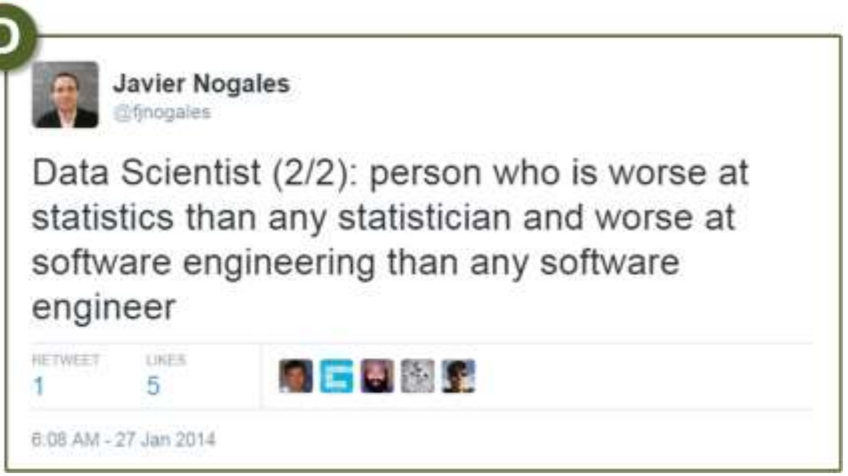
C

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

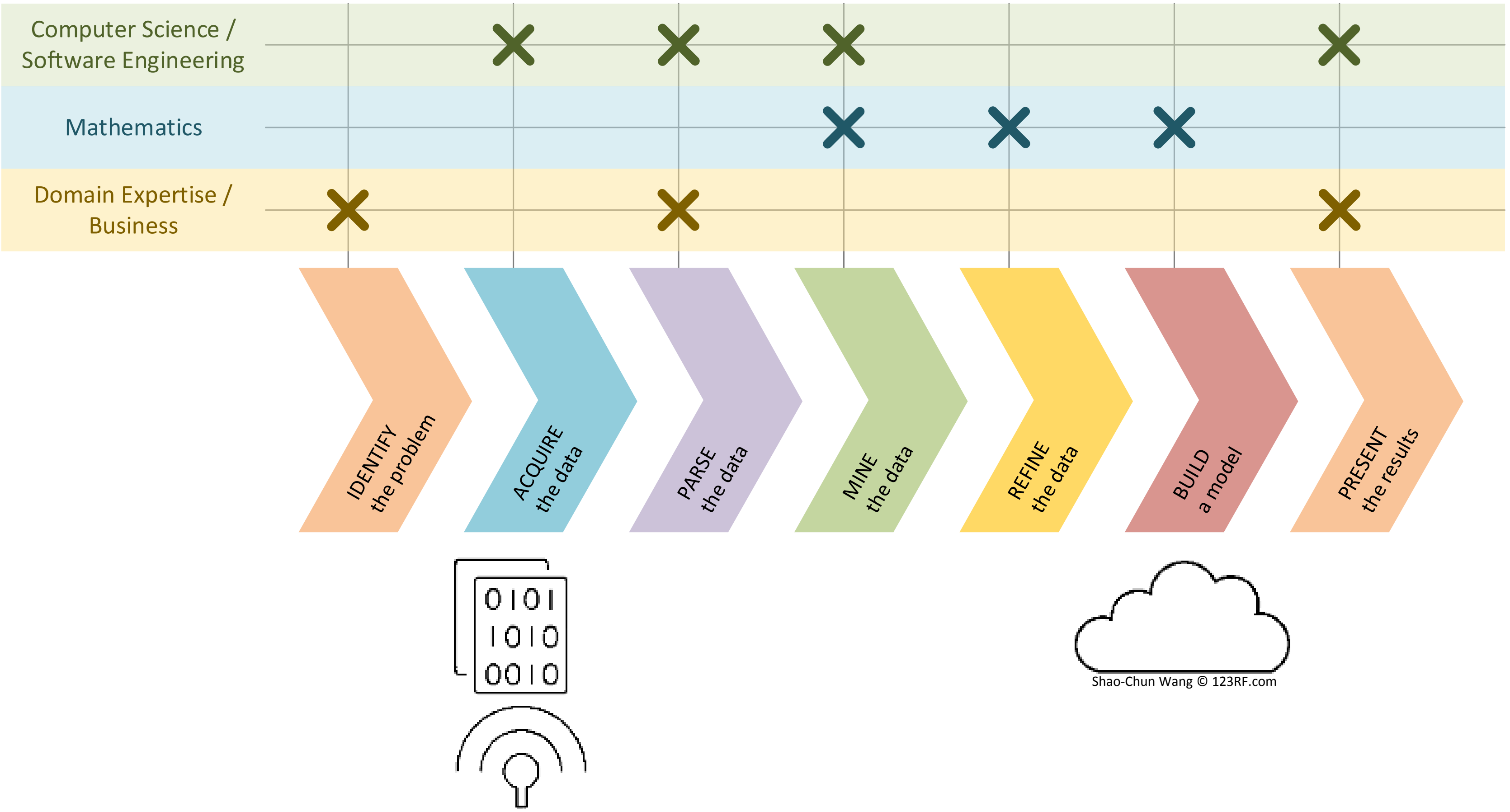
DS

What is Data Science? (cont.)

Data is the New Oil of the Digital Economy and IoT, Big Data, DS, and Cloud relate to one another

Kirill Kirsanov/Shao-Chun Wang © 123RF.com

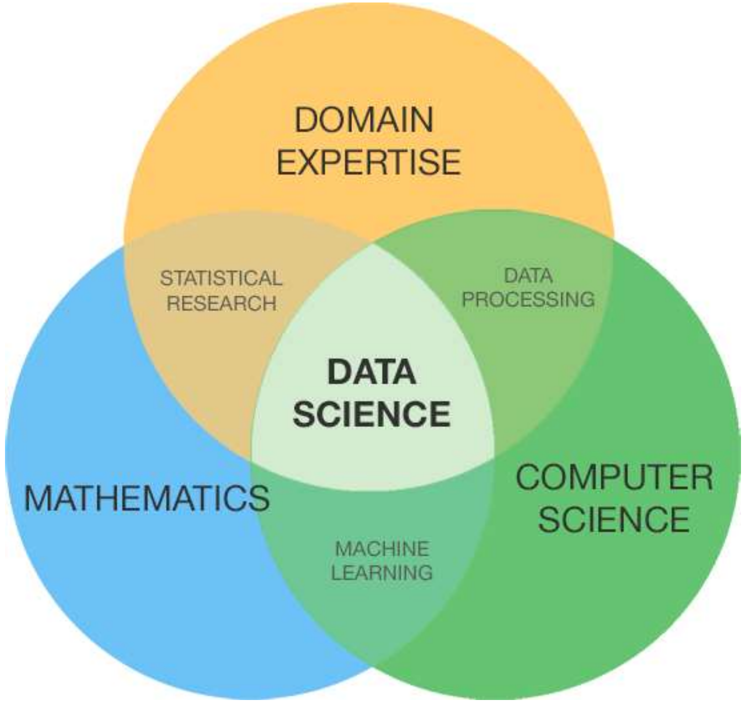
Data Science involves a variety of skillsets, not just one



	IDENTIFY the problem	ACQUIRE the data	PARSE the data	MINE the data	REFINE the data	BUILD a model	PRESENT the results
Computer Science / Software Engineering		✗	✗	✗			✗
Mathematics				✗	✗	✗	
Domain Expertise / Business	✗		✗				✗

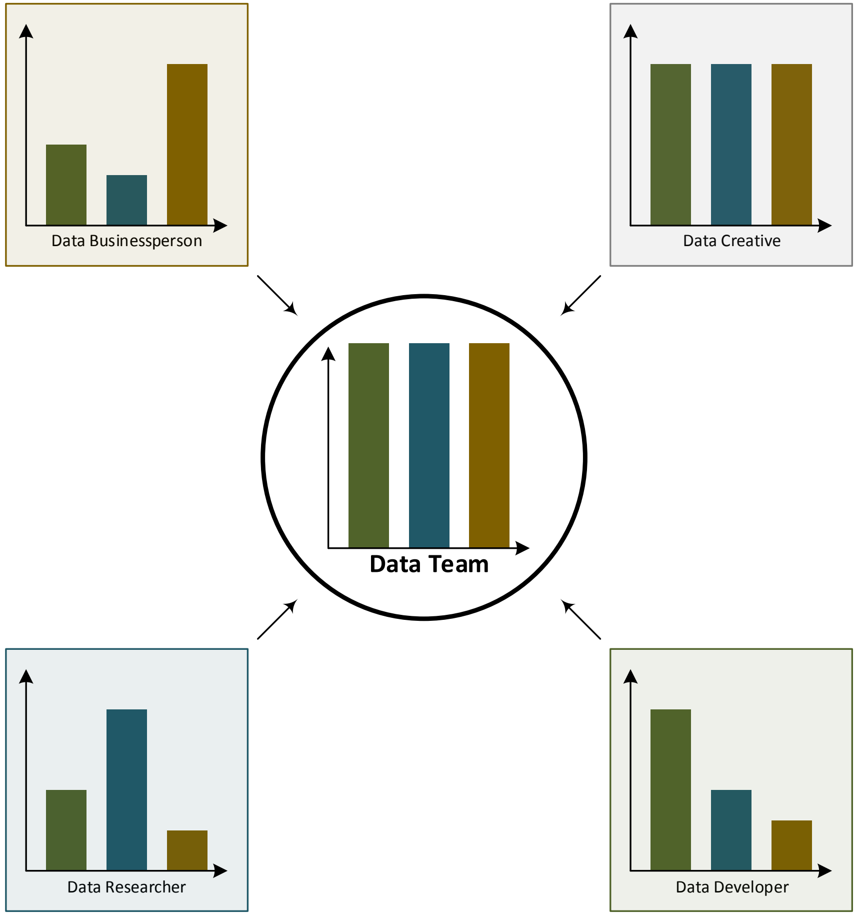
Shao-Chun Wang © 123RF.com

Data Science involves a variety of skillsets, not just one (cont.)



Source: Data Science for the C-suite

Data Scientists have different roles that prioritize different skillsets but all roles involve some part of each skillset to form strong data teams



To sum it up

- Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms

- An (ideal) data scientist is “someone who has the both the engineering skills to acquire and manage large data sets, and also has the statistician’s skills to extract value from the large data sets and present that data to a large audience” – John Rauser

Data Science Workflow
(and how it maps to the course)

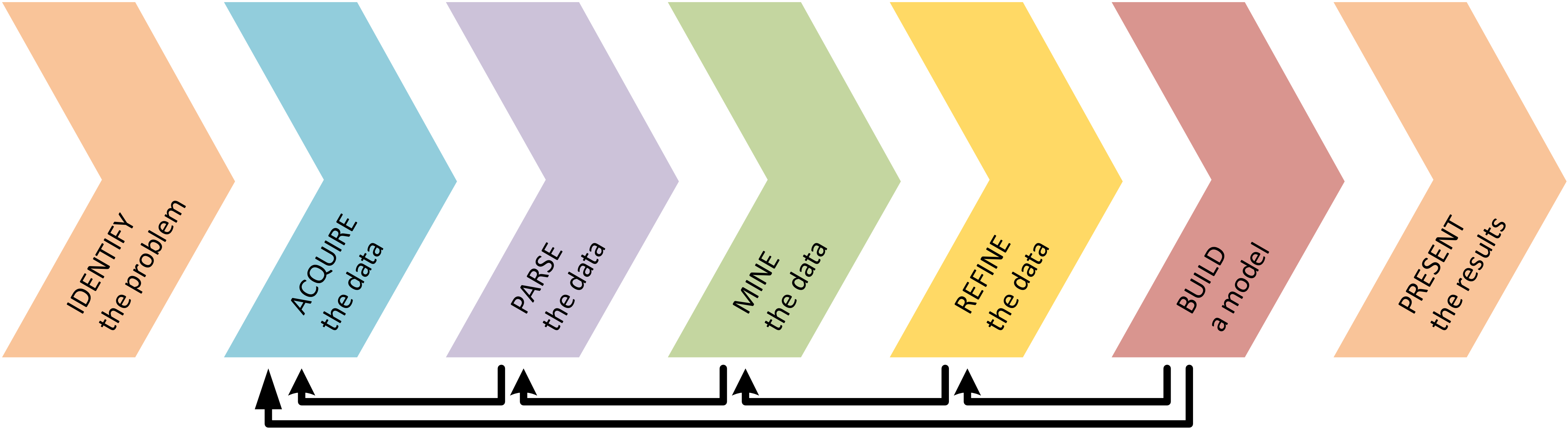
What is the Data Science Workflow for?

- A methodology for Data Science to produce *reliable* and *reproducible* results
  - **Reliable**: Accurate findings
  - **Reproducible**: Others can follow your steps and get the same results
- Similar to the scientific method

The scientific method:

- Ask a Question
- Do Background Research
- Construct a Hypothesis
- Test Your Hypothesis by Doing an Experiment
- Analyze Your Data and Draw a Conclusion
- Communicate Your Results

The Data Science Workflow
(also called the Data Science Pipeline)



IDENTIFY the problem

ACQUIRE the data

PARSE the data

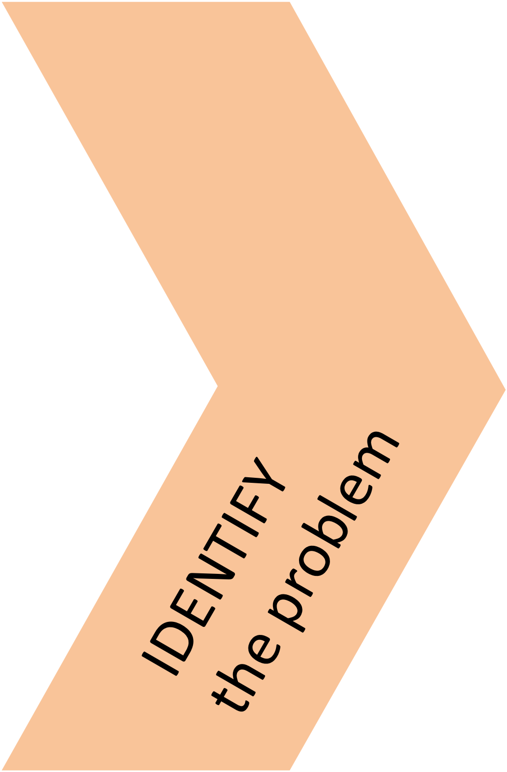
MINE the data

REFINE the data

BUILD a model

PRESENT the results

❶ Identify the Problem



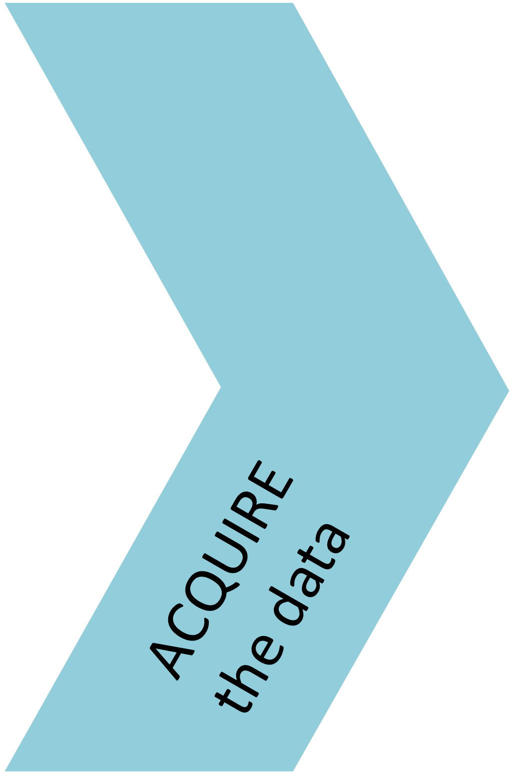
IDENTIFY
the problem

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

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

Corina Rosu © 123RF.com

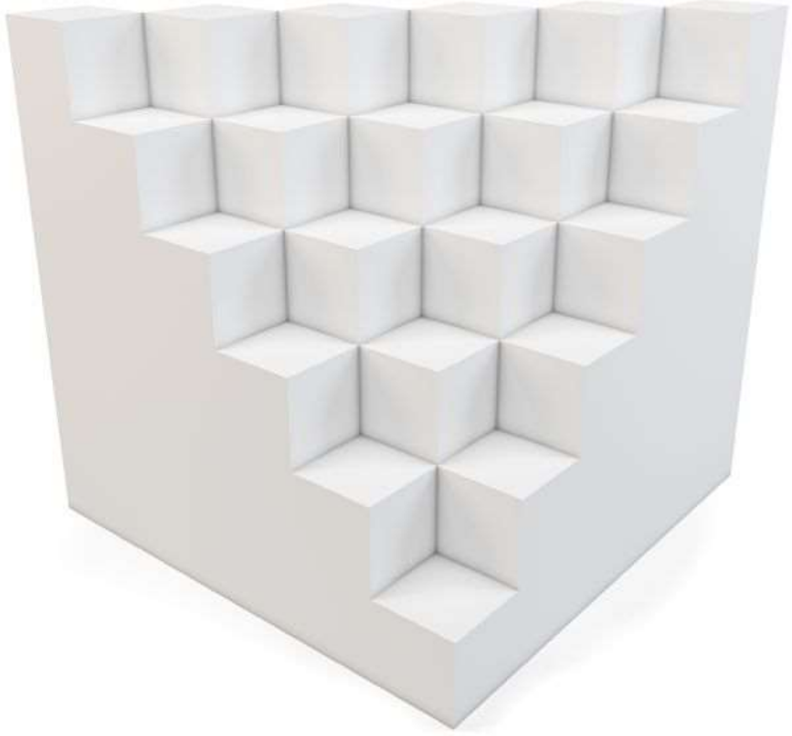
❷ Acquire the Data

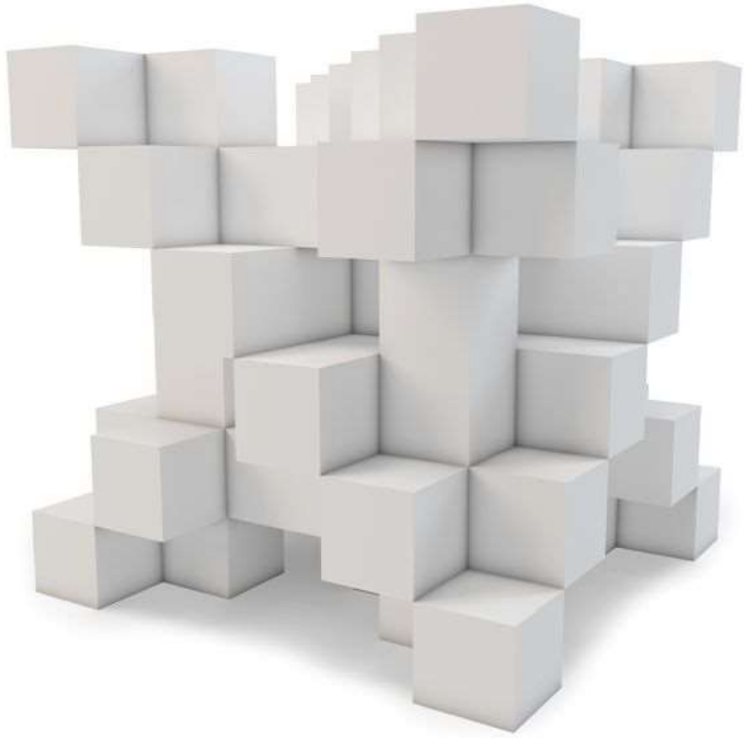


ACQUIRE
the data

- 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

The data can be either unstructured or structured data





dny3d © 123RF.com

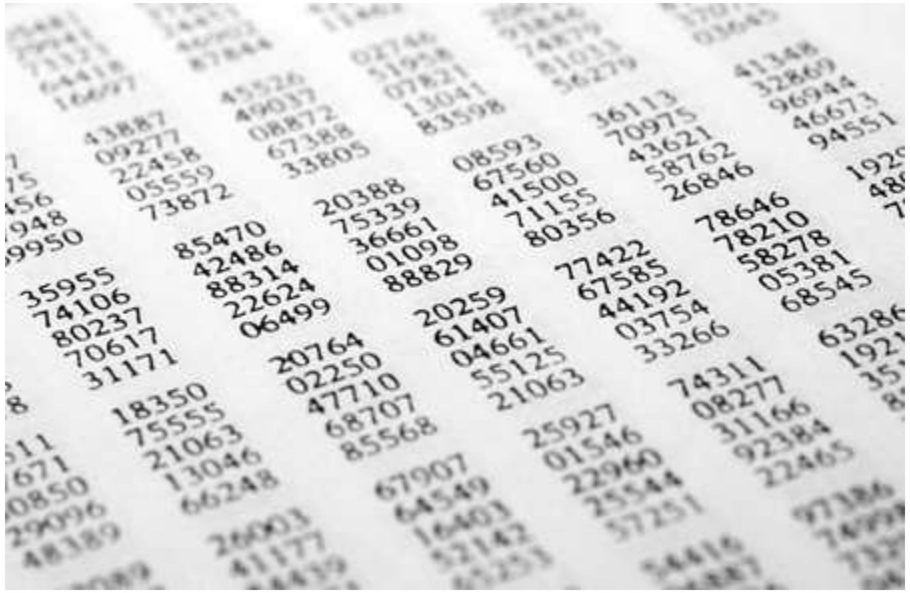
What’s an example of unstructured data?

Bundit Chuangboonsri © 123RF.com

- Sessions 13 and 14 in Unit 3
  - Natural Language Processing

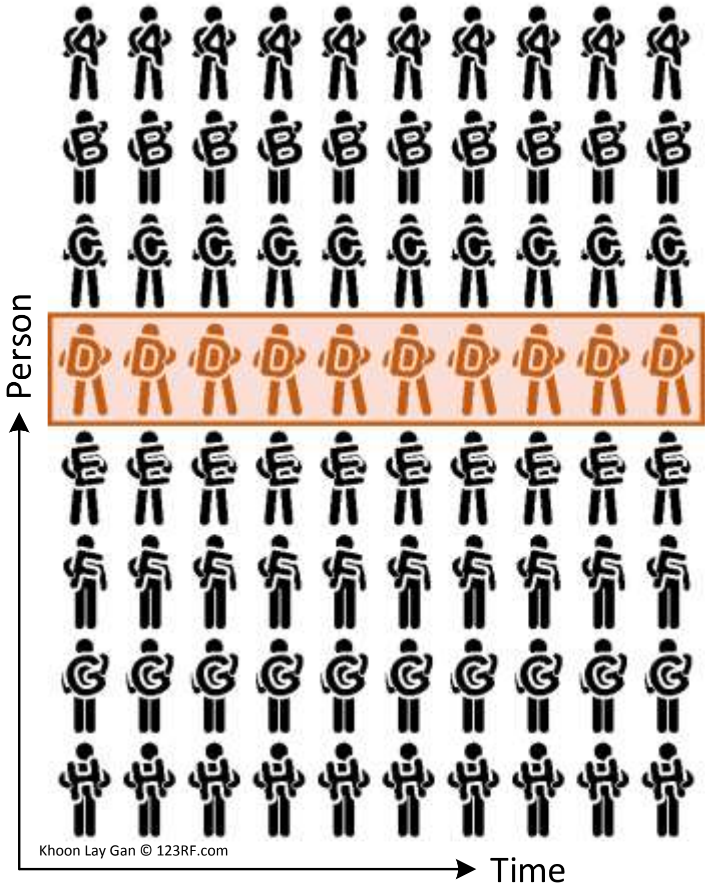
However, most of the course will focus on structured data

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



milosb © 123RF.com

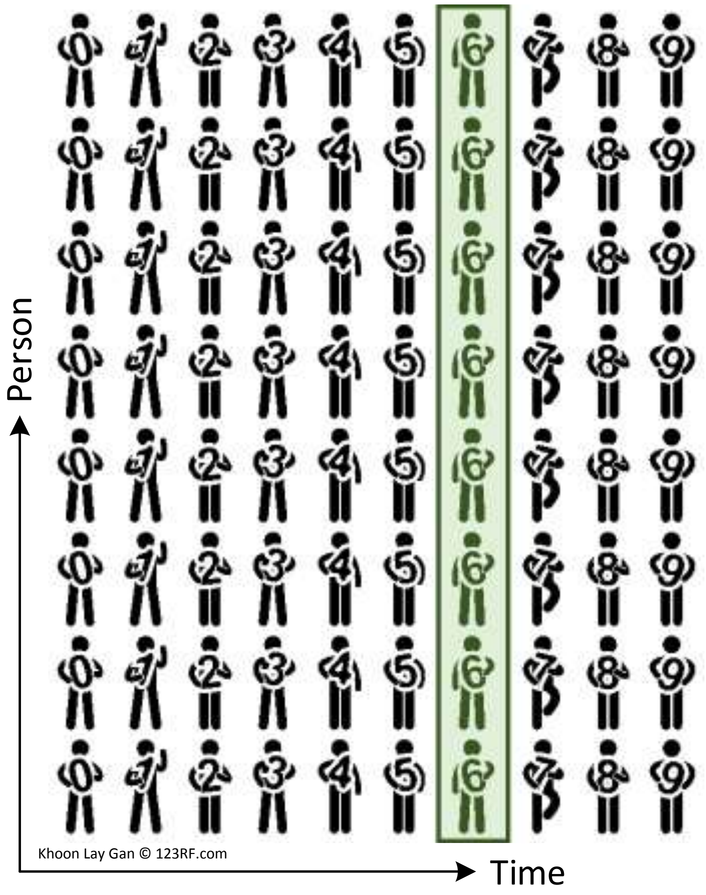
Unstructured data can be longitudinal



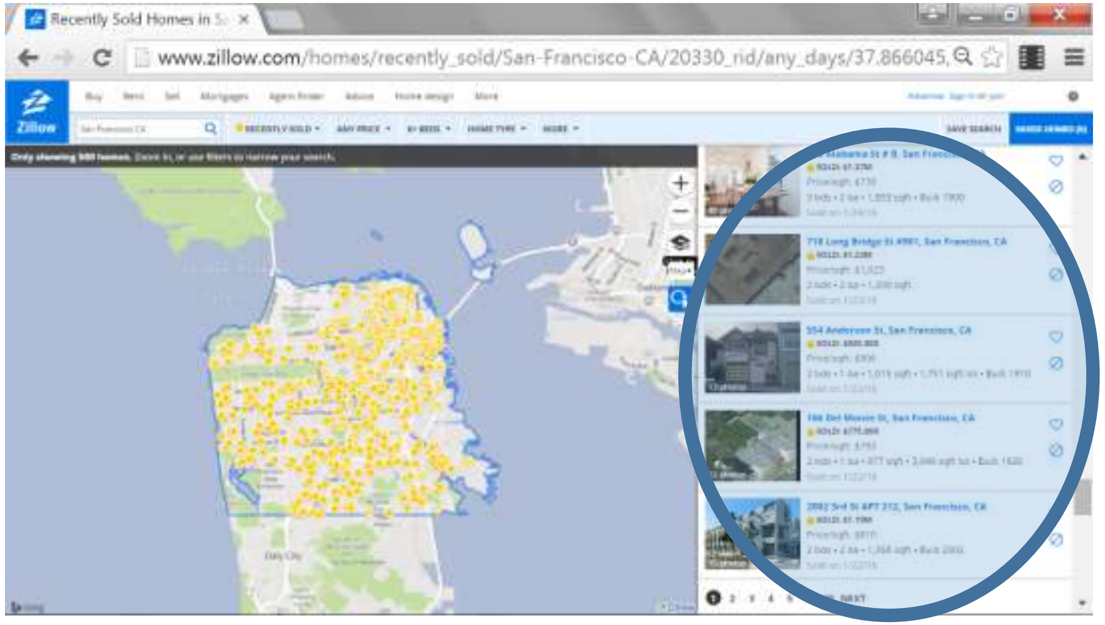
- Sessions 15 and 16 in Unit 3
  - Time Series

Unstructured data can be cross-sectional

- And most of the course will focus on it

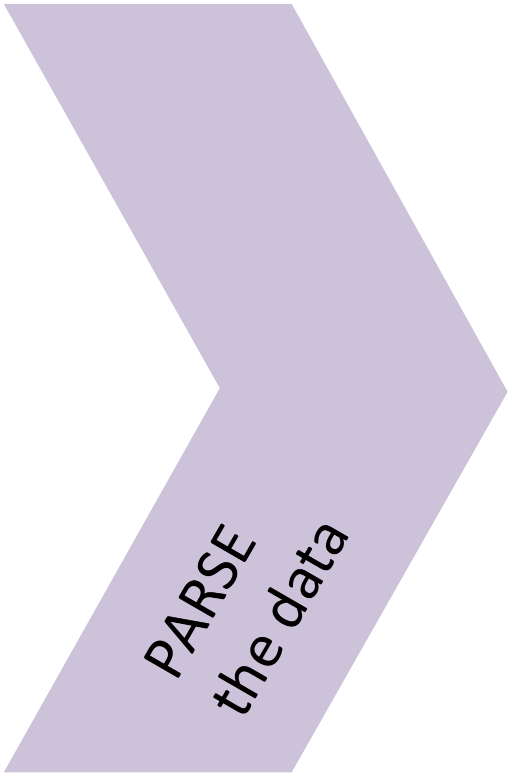


Raw structured data is Messy™…


Recently Sold Homes in S
www.zillow.com/homes/recently_sold/San-Francisco-CA/20330_rid/any_days/37.866045,
Zillow

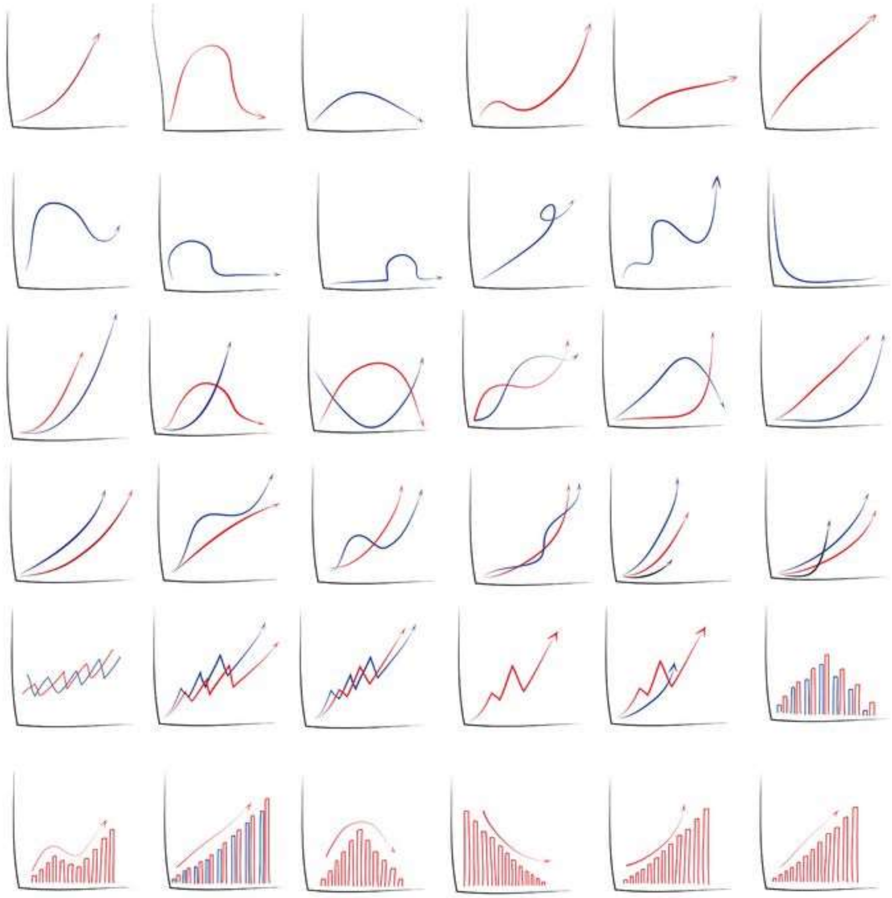
```
<div class="property-info"
id="yui_3_18_1_1_1456167242885_71870"><strong
id="yui_3_18_1_1_1456167242885_71869"><dt class="property-
address" id="yui_3_18_1_1_1456167242885_71868"><a
href="/homedetails/149-Shipley-St-San-Francisco-CA-
94107/15147894_zpid/" class="hdp-link routable" title="149
Shipley 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



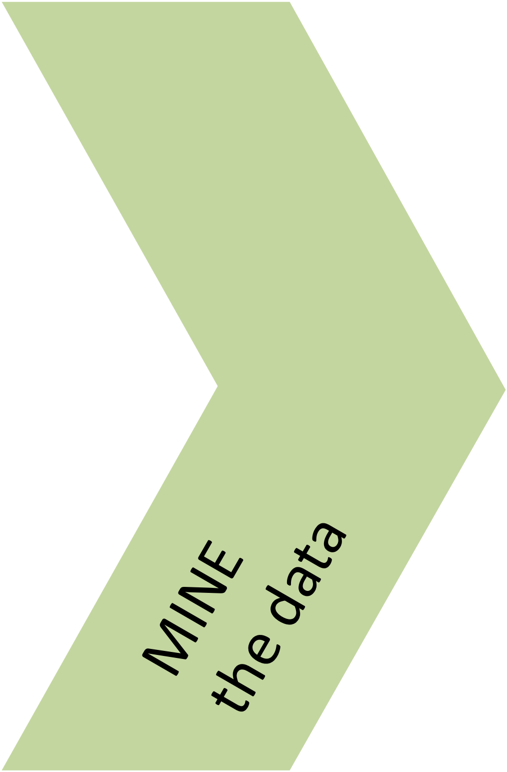
- Parse the Data
  - Read any documentation provided with the data
  - Perform exploratory data analysis
  - Verify the quality of the data

Exploratory Data Analysis



Napat Polchoke © 123RF.com

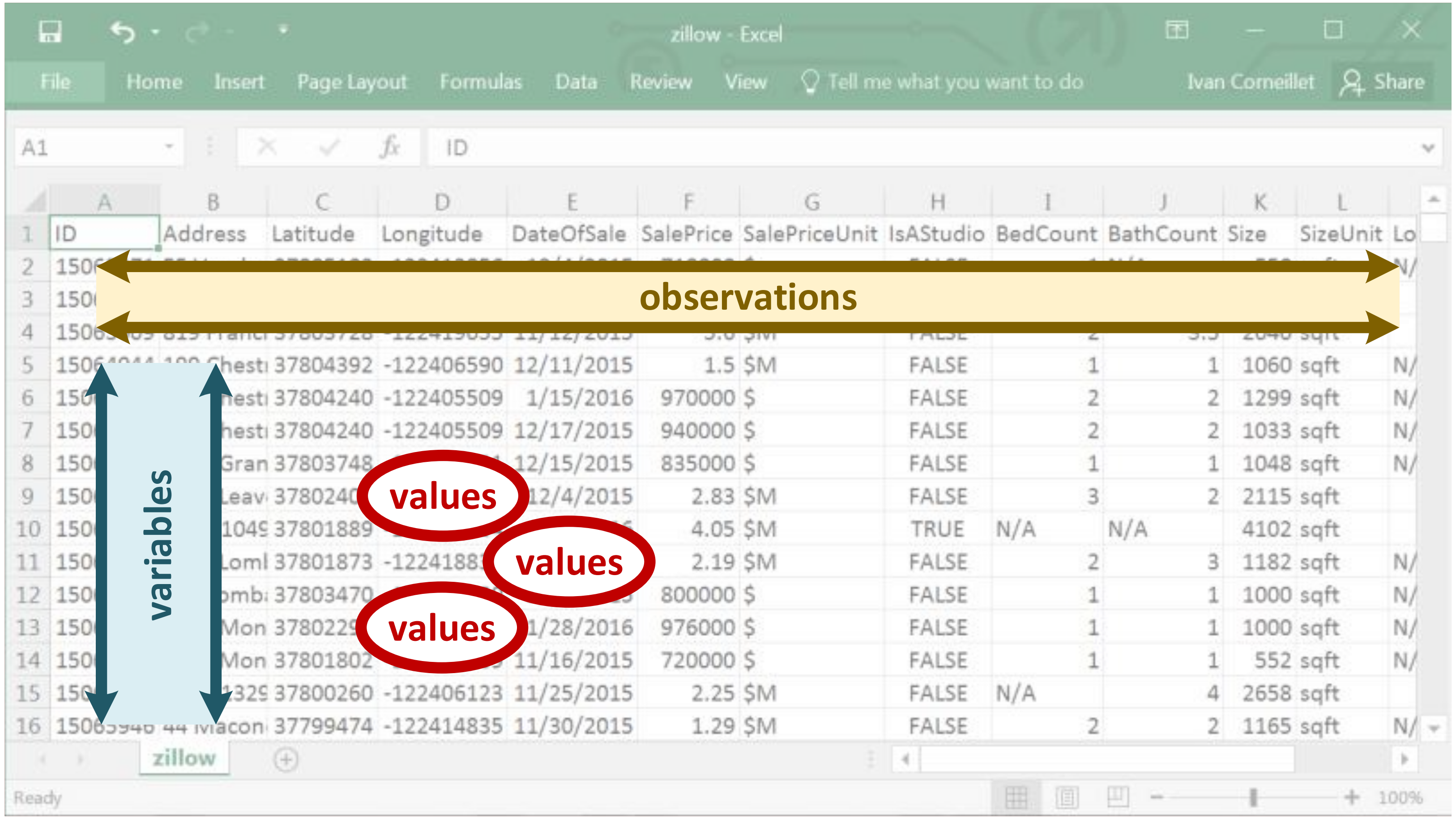
❹ Mine the Data



MINE
the data

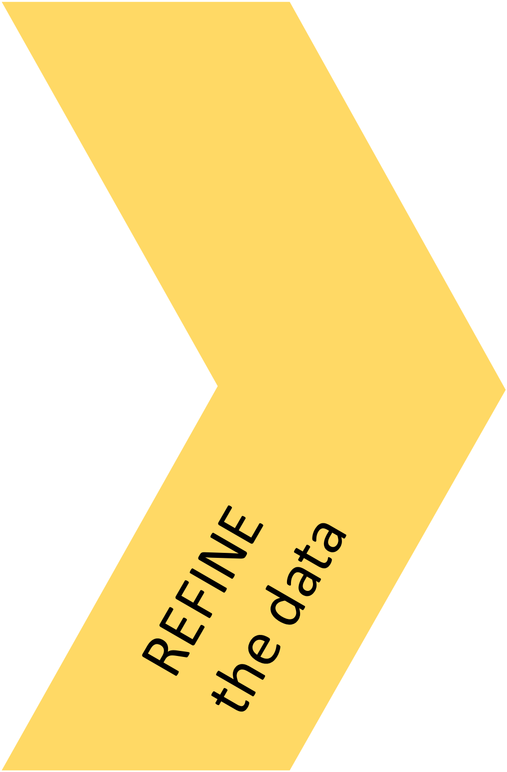
- Mine the Data
  - Determine sampling methodology and sample data
  - Format, clean, slice, and combine data in Python
  - Create necessary derived columns from the data (new data)

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



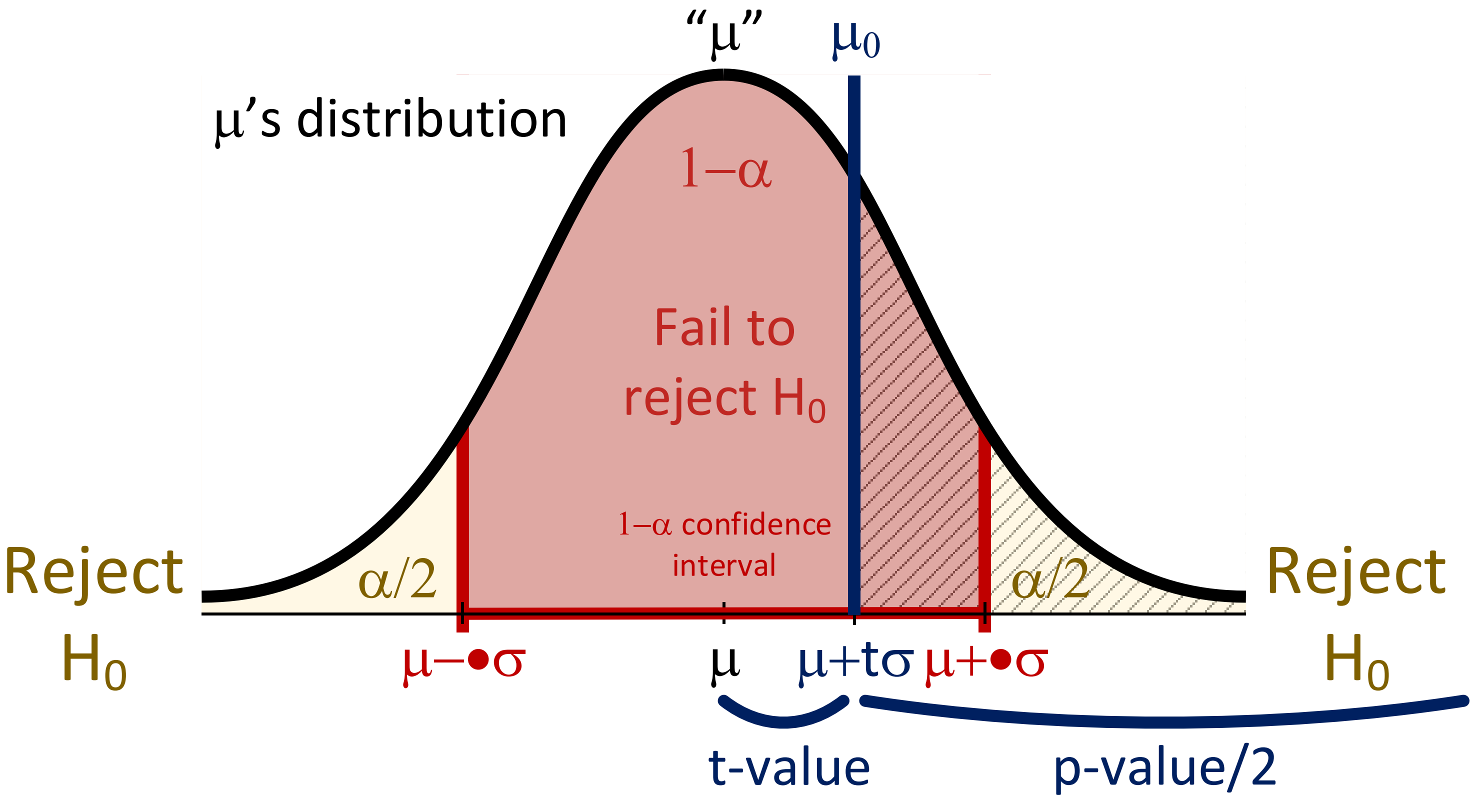
	A	B	C	D	E	F	G	H	I	J	K	L	
1	ID	Address	Latitude	Longitude	DateOfSale	SalePrice	SalePriceUnit	IsAStudio	BedCount	BathCount	Size	SizeUnit	Lo
2	150[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	N/
3	150[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
4	1506[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
5	1506[illegible]	[illegible]Chest	37804392	-122406590	12/11/2015	1.5	$M	FALSE	1	1	1060	sqft	N/
6	150[illegible]	[illegible]esti	37804240	-122405509	1/15/2016	970000	$	FALSE	2	2	1299	sqft	N/
7	150[illegible]	[illegible]hesti	37804240	-122405509	12/17/2015	940000	$	FALSE	2	2	1033	sqft	N/
8	150[illegible]	[illegible]Gran	37803748	[illegible]	12/15/2015	835000	$	FALSE	1	1	1048	sqft	N/
9	150[illegible]	[illegible]Leav	3780240[illegible]	[illegible]	12/4/2015	2.83	$M	FALSE	3	2	2115	sqft	
10	150[illegible]	[illegible]1049	37801889	[illegible]	[illegible]	4.05	$M	TRUE	N/A	N/A	4102	sqft	
11	150[illegible]	[illegible]Loml	37801873	-1224188[illegible]	[illegible]	2.19	$M	FALSE	2	3	1182	sqft	N/
12	150[illegible]	[illegible]ombi	37803470	[illegible]	[illegible]	800000	$	FALSE	1	1	1000	sqft	N/
13	150[illegible]	[illegible]Mon	378022[illegible]	[illegible]	[illegible]1/28/2016	976000	$	FALSE	1	1	1000	sqft	N/
14	150[illegible]	[illegible]Mon	37801802	[illegible]	11/16/2015	720000	$	FALSE	1	1	552	sqft	N/
15	150[illegible]	[illegible]329	37800260	-122406123	11/25/2015	2.25	$M	FALSE	N/A	4	2658	sqft	
16	1506[illegible]	[illegible]Macon	37799474	-122414835	11/30/2015	1.29	$M	FALSE	2	2	1165	sqft	N/

❺ Refine the Data

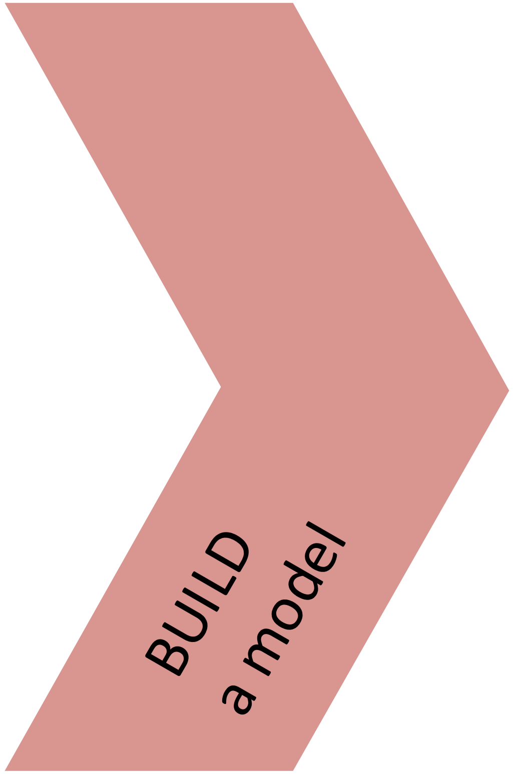

REFINE the data

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

We will apply inferential statistics



❻ Build a Model



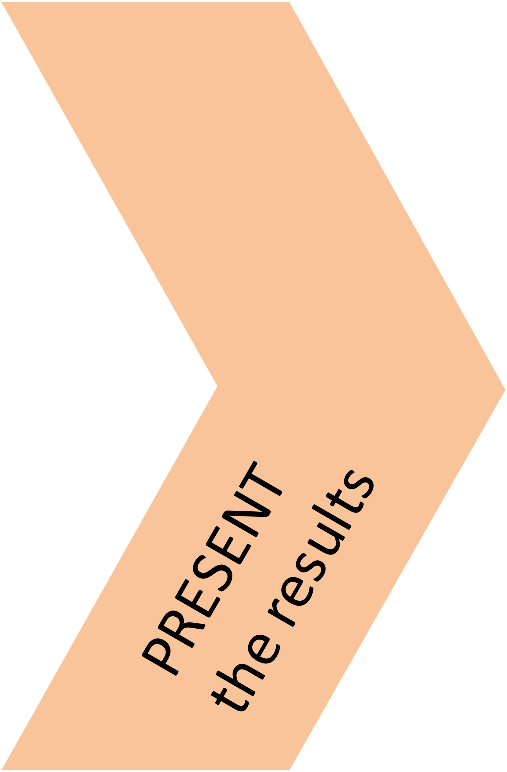
BUILD
a model

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

The types of machine learning algorithms we will study in this course

	Continuous	Categorical
Supervised (a.k.a., predictive modeling)	Linear Regression K-Nearest Neighbors Decision Trees and Random Forests	Logistic Regression K-Nearest Neighbors Decision Trees and Random Forests
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

- 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

Know Your Audience

Corina Rosu © 123RF.com

A Note About Iteration

- Iteration is an important part of *every* step in the Data Science Workflow. At any given point in the process, you may find yourself repeating or going back and re-doing elements in order to better understand your data, clarify your model, and refine your presentation

- For example, after presenting your findings, you may want to:
  - Identify follow-up problems and questions for future analysis
  - Create a visually effective summary or report
  - Consider the needs of different stakeholders and how your report might be changed for them
  - Identify the limitations of your analysis
  - Identify relationships between visualizations

Multiple variants exist but they are pretty much all doing the same thing

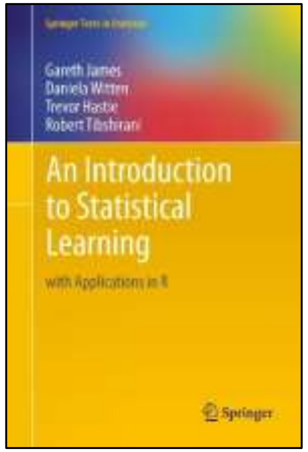
- Jeff Hammerbacher
  - Identify problem
  - Instrument data sources
  - Collect data
  - Prepare data (integrate, transform, clean, impute, filter, aggregate)
  - Build model
  - Evaluate model
- Ben Fry
  - Acquire
  - Parse
  - Filter
  - Mine
  - Represent
  - Refine
  - Interact

- Peter Huber
  - Inspection
  - Error checking
  - Modification
  - Comparison
  - Modeling and model fitting
  - Simulation
  - What-if analyses
  - Interpretation
  - Presentation of conclusions
- Dataists
  - Obtain
  - Scrub
  - Explore
  - Model
  - Interpret

- Colin Mallows
  - Identify data to collect and its relevance to your problem
  - Statistical specification of the problem
  - Method selection
  - Analysis of method
  - Interpret results for non-statisticians
- Jim Gray
  - Capture
  - Curate
  - Communicate
- Ted Johnson
  - Assemble an accurate and relevant dataset
  - Choose the appropriate algorithm

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



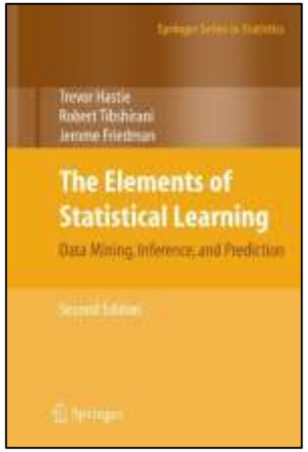
Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

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

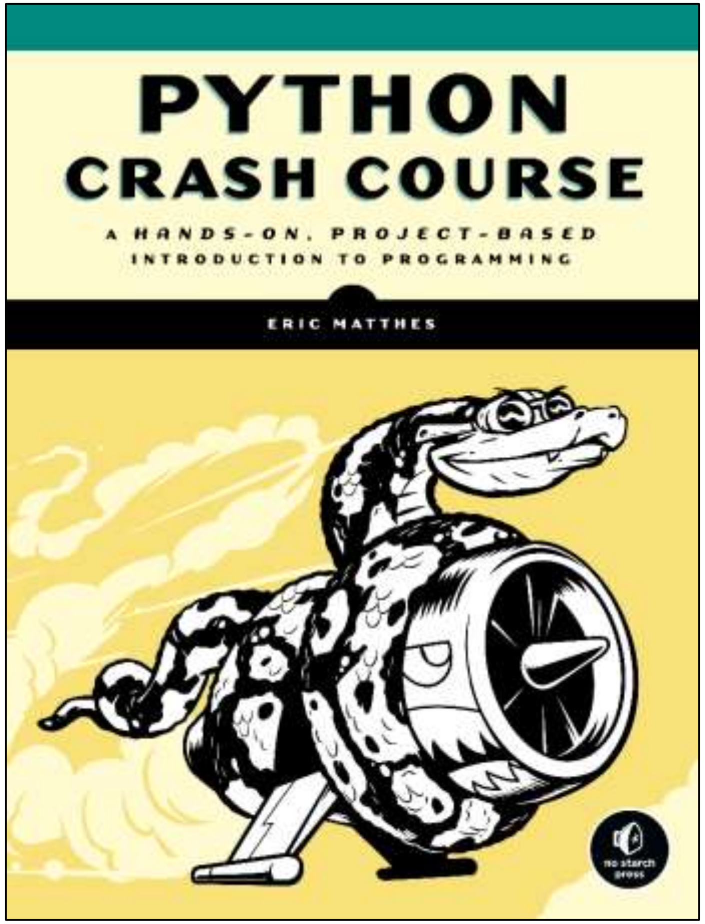
Q & A

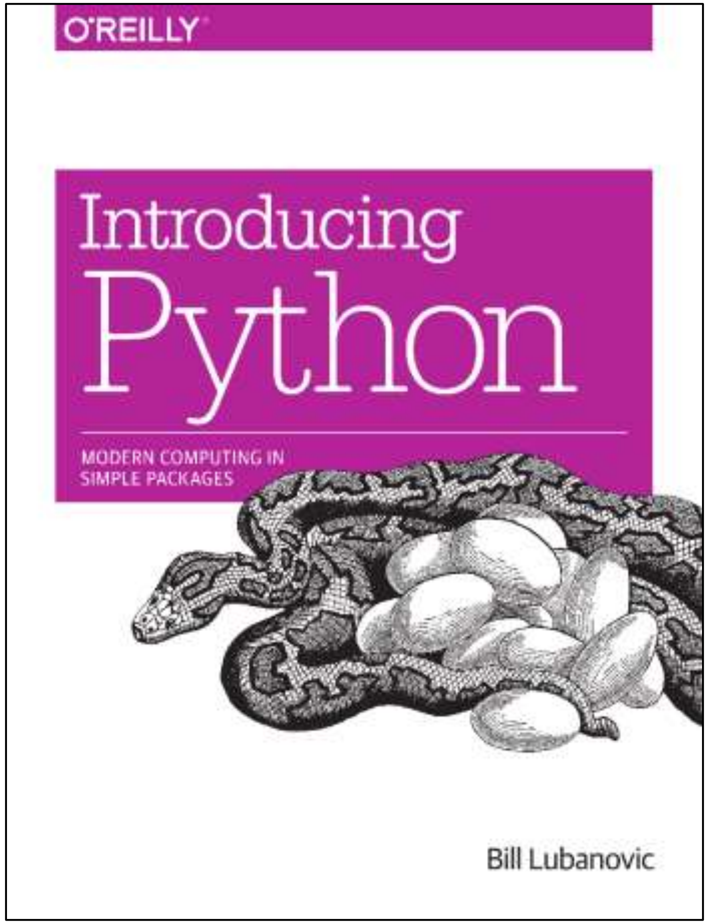
Onboarding Review

DS

Python and iPython Notebooks

A couple of resources to get started with Python *(optional; not required for the course)* (the contents overlap but the styles are different so you should only pick one if any)





Review

Review

You should now be able to answer the following questions:

- What is data science?
- What is the data science workflow?
- How can you have a successful learning experience at GA?

Q & A

Next Class

Research Design and pandas

Learning Objectives

After the next lesson, you should be able to:

- Setup and manage your personal GitHub repository for submitting assignments
- Define a problem and types of data
- Identify dataset types
- Apply the data science workflow in the *pandas* context
- Write an iPython notebook to import, format, and clean data using the *pandas* library

Exit Ticket

Don’t forget to fill out your exit ticket here