Network ID: GA-Guest

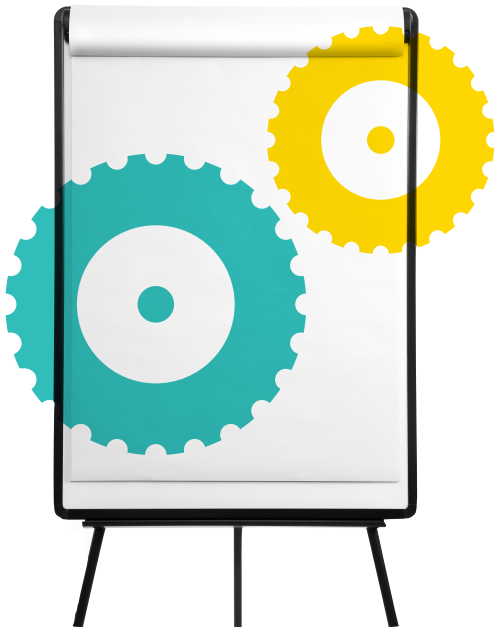
Password: yellowpencil

WELCOME TO OPENING DAY OF

DATA SCIENCE

ORIENTATION AGENDA

1. Meet Your Course Producer
2. Meet The Instructional Team
3. General Course Info
4. GA Mission
5. Course Expectations
6. Student Experience
7. Meet Your Classmates



OPENING DAY

HELLO

WHAT IS A PROGRAM PRODUCER?

FOCUS: STUDENT SUCCESS + INSTRUCTOR SUPPORT

timothy.payne@ga.co

OPENING DAY

MEET YOUR INSTRUCTIONAL TEAM

INSTRUCTOR & TA

IVAN CORNEILLET
BOB STARK

COURSE INFO

DS23

May 5th - July 12th

Tuesdays and Thursdays 6:30-9:30

Classroom 7

GENERAL ASSEMBLY IS A GLOBAL COMMUNITY OF INDIVIDUALS EMPOWERED TO PURSUE THE WORK WE LOVE.

GENERAL ASSEMBLY'S MISSION IS TO BUILD OUR COMMUNITY BY TRANSFORMING MILLIONS OF THINKERS INTO CREATORS.

COURSE EXPECTATIONS

SUCCESSFUL GRADUATE PROFILE

COMPLETE 80% OF HOMEWORK/LABS

BE PRESENT FOR AT LEAST 18 CLASSES

ROCK IT!

GET INVOLVED!

COURSE EXPECTATIONS

1. Arrive on time.
2. Turn in your assignments.
3. Ask questions.
4. Share with peers.
5. Complete exit tickets.
6. Complete mid & end of course surveys.
7. Make friends.

STUDENT SPACE

STUDENT EXPERIENCE

DISCOUNTS!

15% off all classes and workshops

code: currentstudentdiscount15

MEET YOUR CLASSMATES

WE'RE ALL IN THIS TOGETHER.

QUESTIONS?

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 data science and the data science workflow
- Setup your development environment and review Python basics

Here's what's happening today:

- Welcome to GA and DS!
- Pre-work
- Setting You Up for Success
  - Slack
  - GitHub
  - Exit Tickets
- What is Data Science and who are Data Scientists?
- The Data Science Workflow
  - And how it maps to the course
- Onboarding Review
- Python and iPython Notebooks
- Review
- Projects
  - Unit Project 1 (due in 1 week)

Welcome to GA and DS

WELCOME TO GA!

WELCOME TO DS!

DS

slack

scikit learn

GitHub

python

Pre-Work

Pre-Work

Before this lesson, you should already be able to:

- Define basic data types used in object-oriented programming
- Recall the Python syntax for lists, dictionaries, and functions
- Create files and navigate directories using the command line interface

Setting You Up for Success

Course Logistics

- Instructor
  - Ivan Corneillet (ivan+GA@paspeur.com)
- Expert-in-Residence
  - Bob Stark (bobness@gmail.com)
- Course Producer
  - Tim Payne (timothy.payne@generalassemb.ly)
- Class
  - May 5 – July 12, Tuesdays and Thursdays, 6:30PM – 9:30PM
  - Classroom 7
- Slack
  - https://ds-sf-23.slack.com
- GitHub
  - https://github.com/ga-students/DS-SF-23
- Exit Tickets
  - http://tiny.cc/ds-sf-23

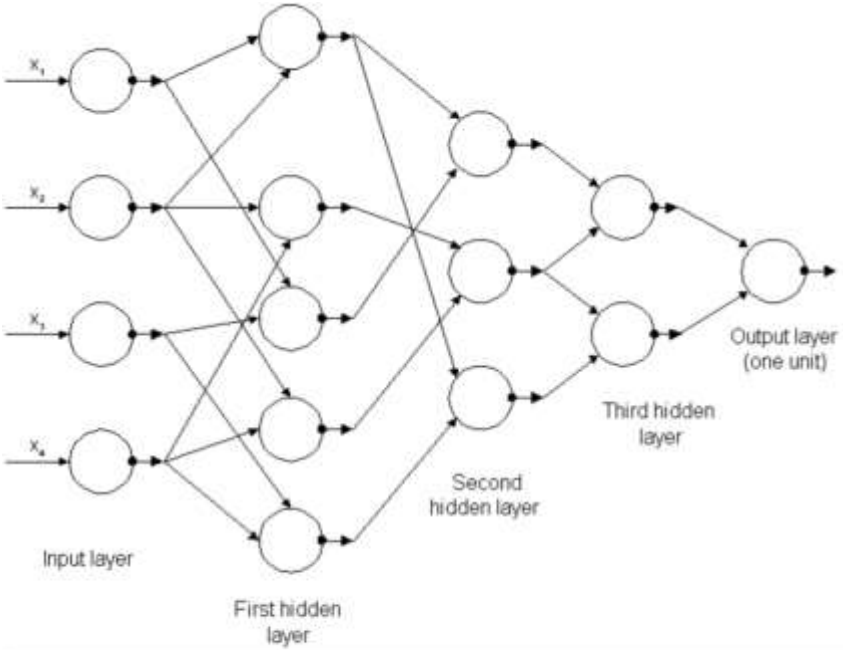
Meet Your Team

- Ivan Corneillet, Instructor

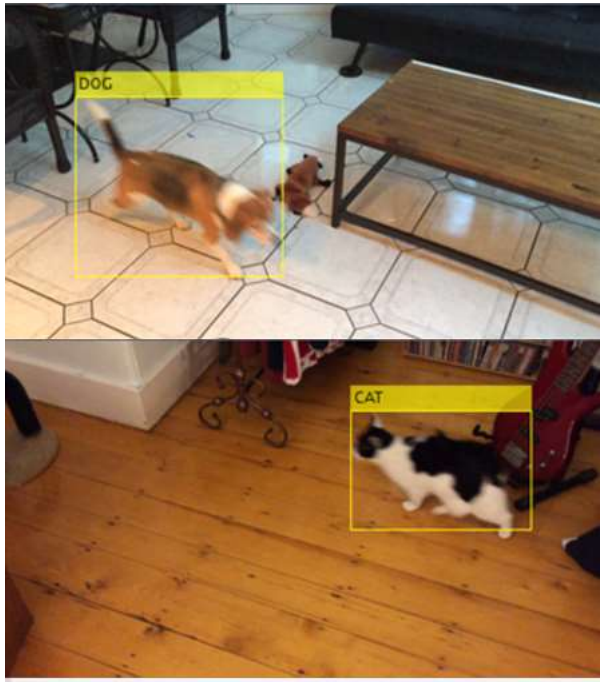
- Bob Stark, Expert-in-Residence

- Tim Payne, Course Producer

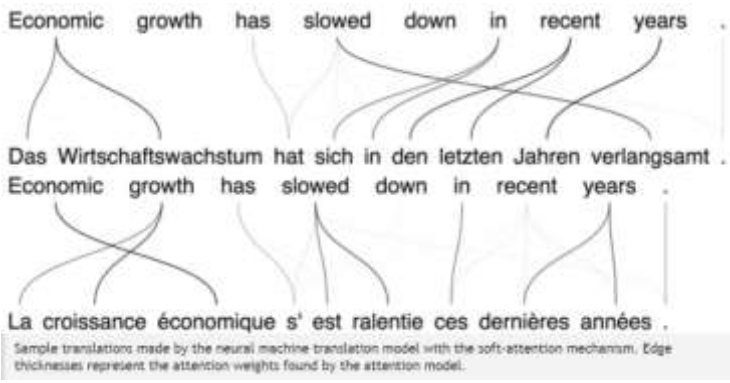
Deep Learning and High-Performance Computing



The achitecture of the first known deep network which was trained by Alexey Grigorevich Ivakhnenko in 1965. The feature selection steps after every layer lead to an ever-narrowing architecture which terminates when no further improvement can be achieved by the addition of another layer.



Pet detection and recognition system.



Sample translations made by the neural machine translation model with the soft-attention mechanism. Edge thicknesses represent the attention weights found by the attention model.

Source: NVIDIA

mettus © 123RF.com

What skills can I expect to learn in this class?

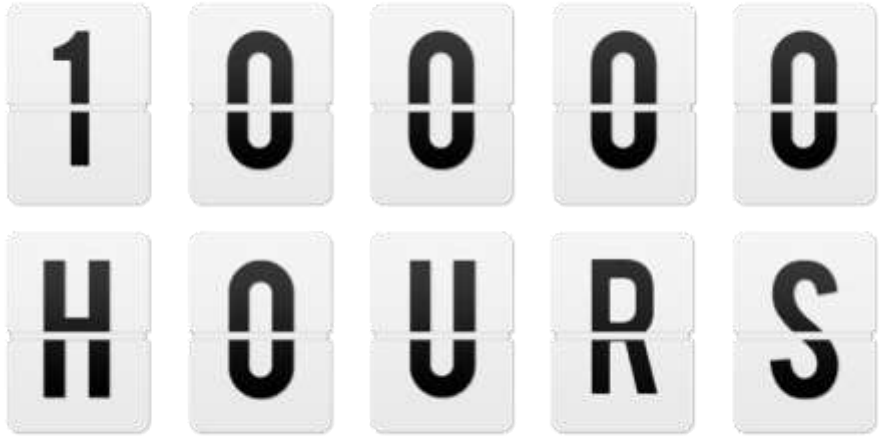
Research Design and Data Analysis	Research Design	Data Visualization in *pandas*	Statistics	Exploratory Data Analysis in *pandas*
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and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

What other skills can I expect to learn in this class?

- The onboarding Python tutorial covered the fundamentals of the Python programming language (basic syntax and data structures such as dictionaries)
- Over the course, you will get hands-on programming experience in Python and several data science tools to progressively hone both your Python and data analysis skills
- In particular, we will cover:
  - Tabular data structure with *pandas* dataframes
  - Control flow: Conditional statements and loops
  - Defining and using functions; simplify our life and avoid code duplication (e.g., transforming variables on the training set then on the testing set)
  - Primer on object-oriented programming

But you need to practice, practice, and… practice

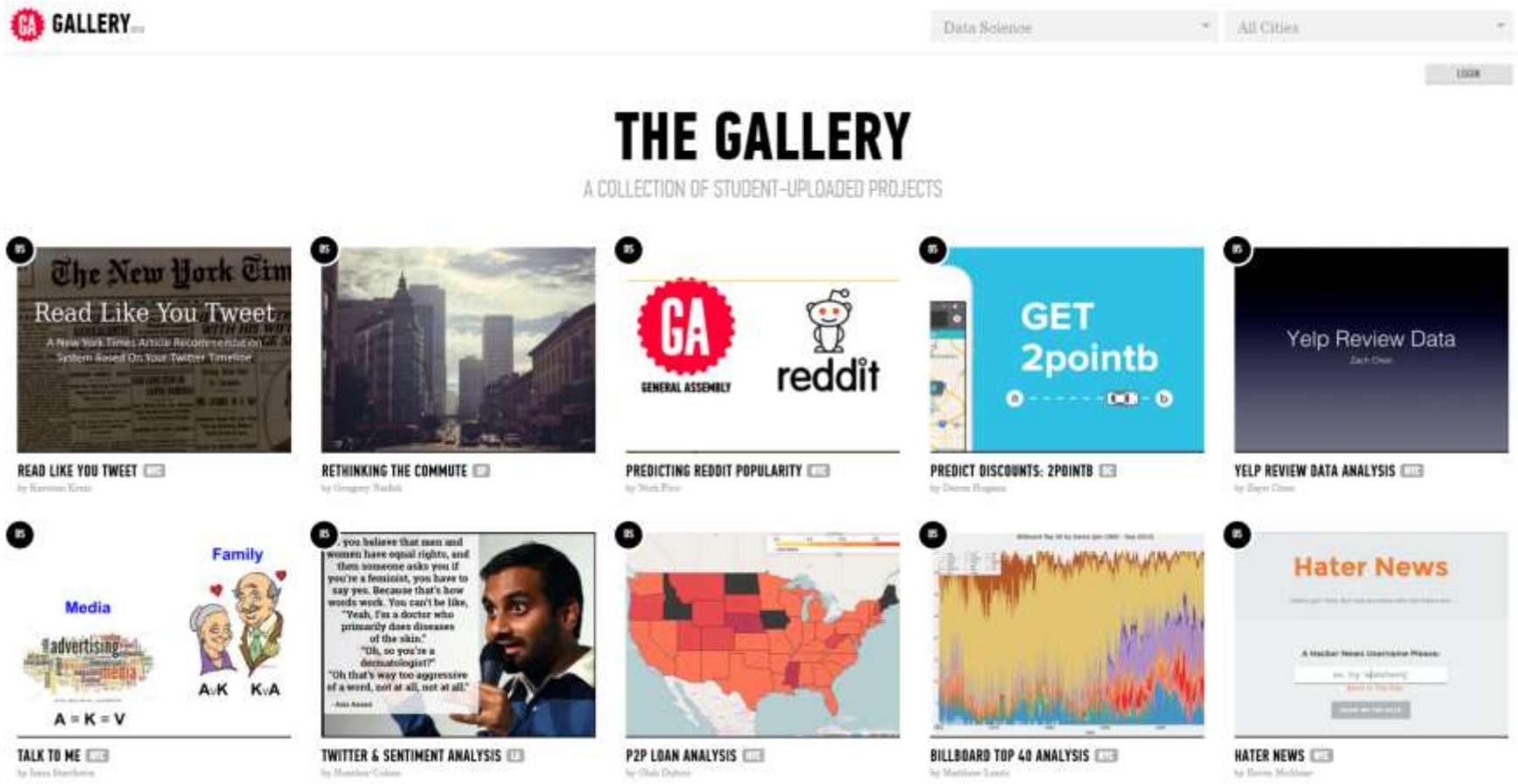
- It takes roughly ten thousand hours of practice to achieve mastery in a field (Malcolm Gladwell)



How can 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

Past Student Projects
(https://gallery.generalassemb.ly/DS)



Course Schedule (and yes, you are starting to work on your projects today…)

	Tuesdays				Thursdays			
						5	L1	What is Data Science
May	Research Design and *pandas*	L2		10	U1	12	L3	Statistics Fundamentals
	Statistics Fundamentals, Part 2	L4		17	U2	19	L5	*Flexible Class Session*
	Introduction to Regression and Model Fit	L6	F1	24		26	L7	Introduction to Regression and Model Fit, Part 2
	Introduction to Classification	L8		31	F2	2	L9	Introduction to Logistic Regression
June	Advanced Metrics and Communicating Results	L10	U3	7		9	L11	*Flexible Class Session*
	Decision Trees and Random Forests	L12	F3	14		16	L13	Natural Language Processing
	Natural Language Processing, Part 2	L14	U4	21		23	L15	Time Series Data
	Time Series Data, Part 2	L16		28	F4	30	L17	Introduction to Databases
July	Wrapping Up and Next Steps	L18		5	F5	8	L19	Final Project Presentations
	Final Project Presentations, Part 2	L20		12				

Unit 1 – Research Design and Exploratory Data Analysis	Ux – Unit Project x Due
Unit 2 – Foundations of Data Modeling	
Unit 3 – Data Science in the Real World	Fx – Final Project Part x Due

Typical Class

- Recap of last time
- Series of mini-lectures and practices (exercises and codealongs)
- Lab

Philosophy

Instructor

- Embrace diversity
- Seek an optimal pace
- Communicate early and often
- Success is not a grade

Course

- Application-based approach
- Understand key principles
- Balance depth with breadth
- Course project

What do we expect from you?

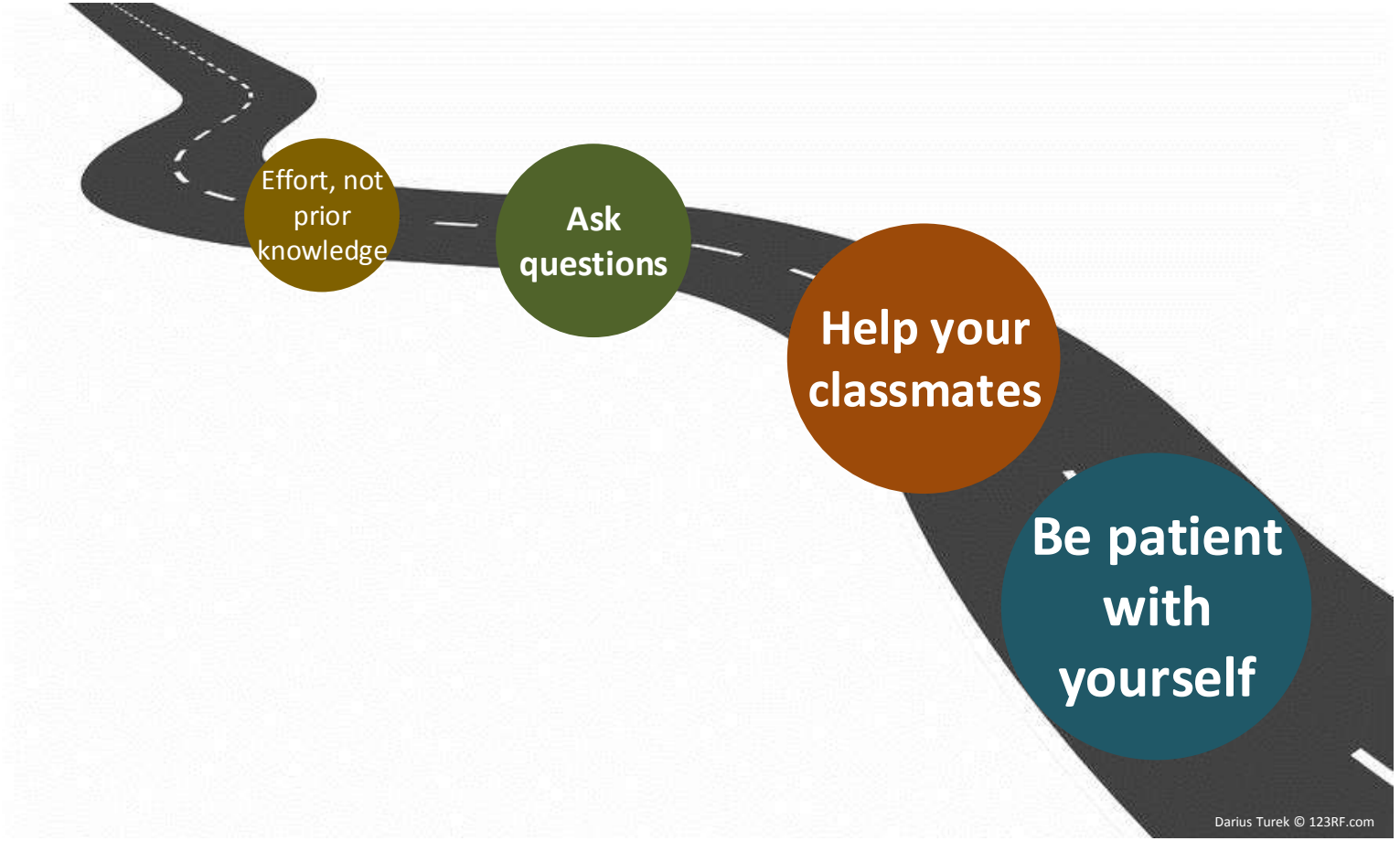
- Arrive on time!
- Turn in your assignments
- Ask questions
- Share with your fellow students
- Learn lots and make friends
- Complete exit tickets and mid-/end-course surveys

Your 4 Steps to Graduation

Life happens so if you miss a class:

- Let your instructional team know ASAP
- Review the class materials (lecture and lab)
- Grab coffee with a classmate and go over his/her notes with him/her
- Attend office hours

Your DS Road to Success



Your GA Road to Success

Build your network

It's not just about altruism; your network is your most valuable asset

13,000+ strong

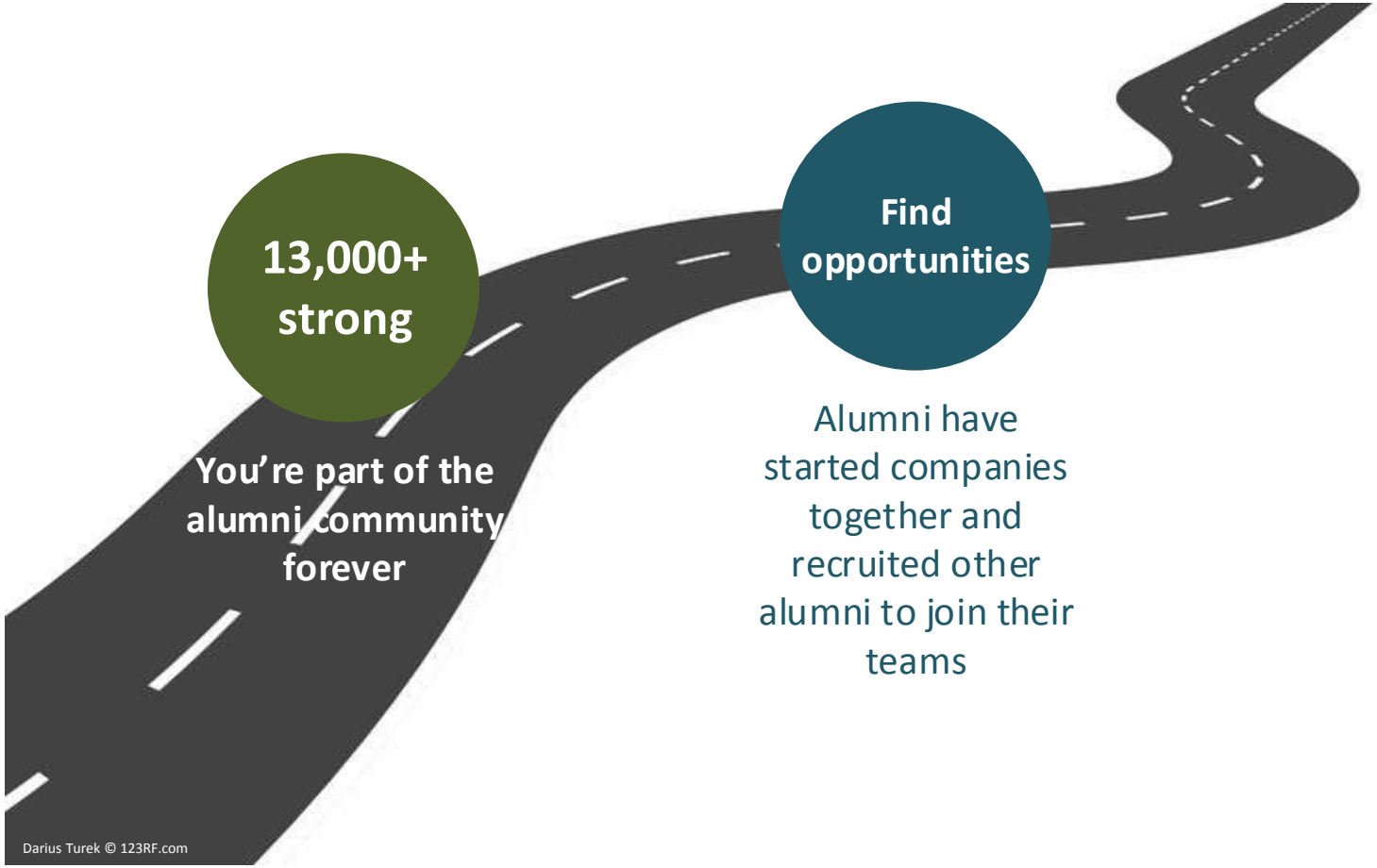
You're part of the alumni community forever

Find opportunities

Alumni have started companies together and recruited other alumni to join their teams

Perks!

We can't wait to have you back on campus



Q & A

Setting You Up for Success

Slack (https://ds-sf-23.slack.com)

GitHub (https://github.com/ga-students/DS-SF-23)

Exit Tickets (http://tiny.cc/ds-sf-23)

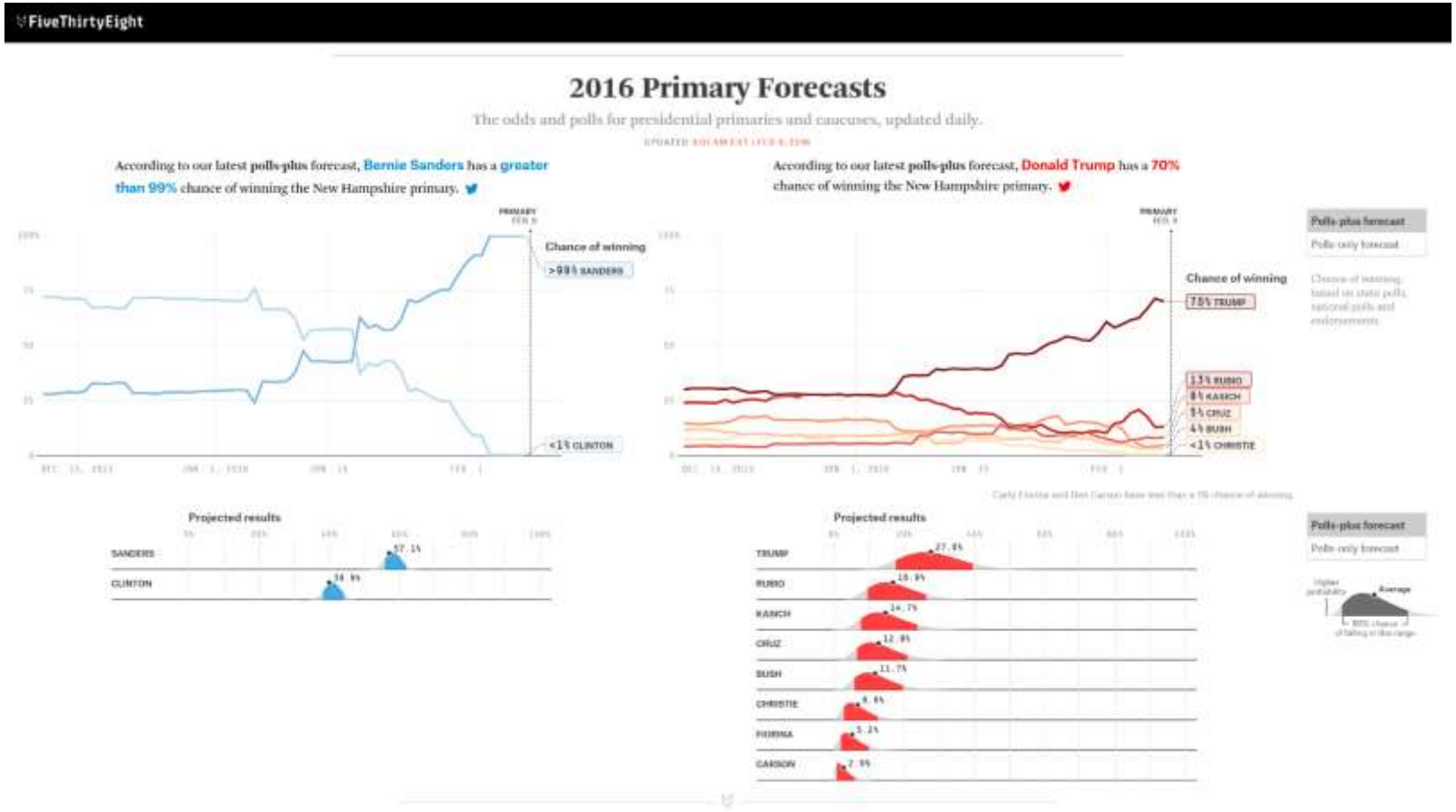
What is Data Science?

Data Scientists: The Sexiest Job of the 21st Century

Source: Harvard Business Review

FiveThirtyEight

Data Science by FiveThirtyEight

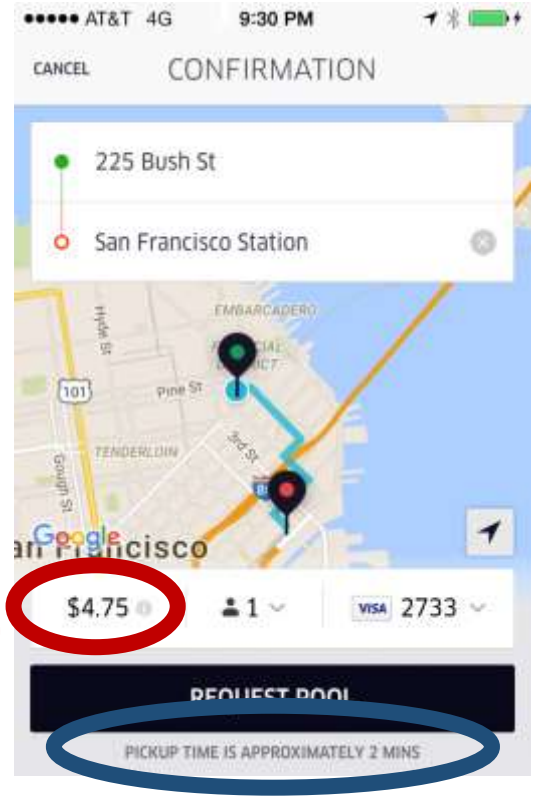


Source: FiveThirtyEight

Uber

Data Science by Uber

For Riders...

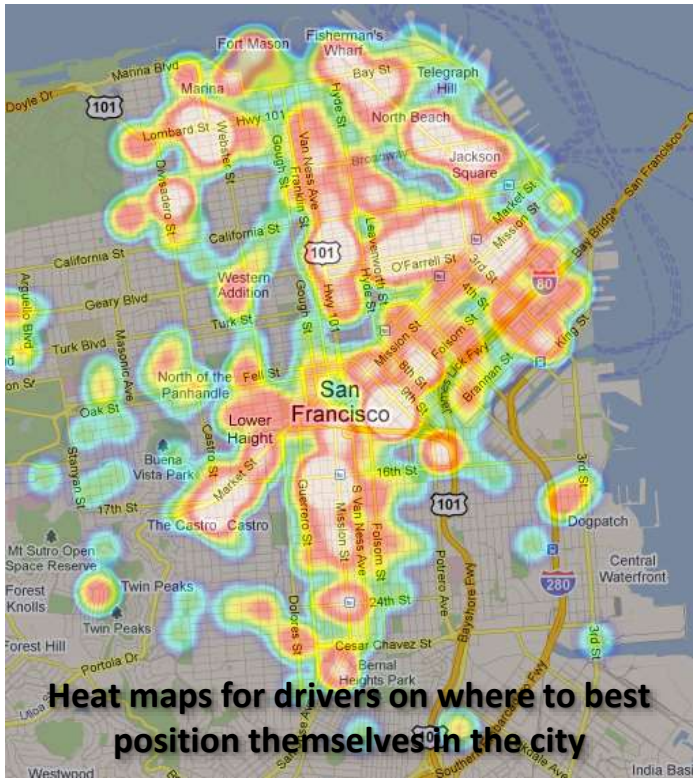


Fare Estimator and Dynamic Pricing (e.g., Surge Pricing) Algorithms

ETA (Estimated Time of Arrival) Algorithms

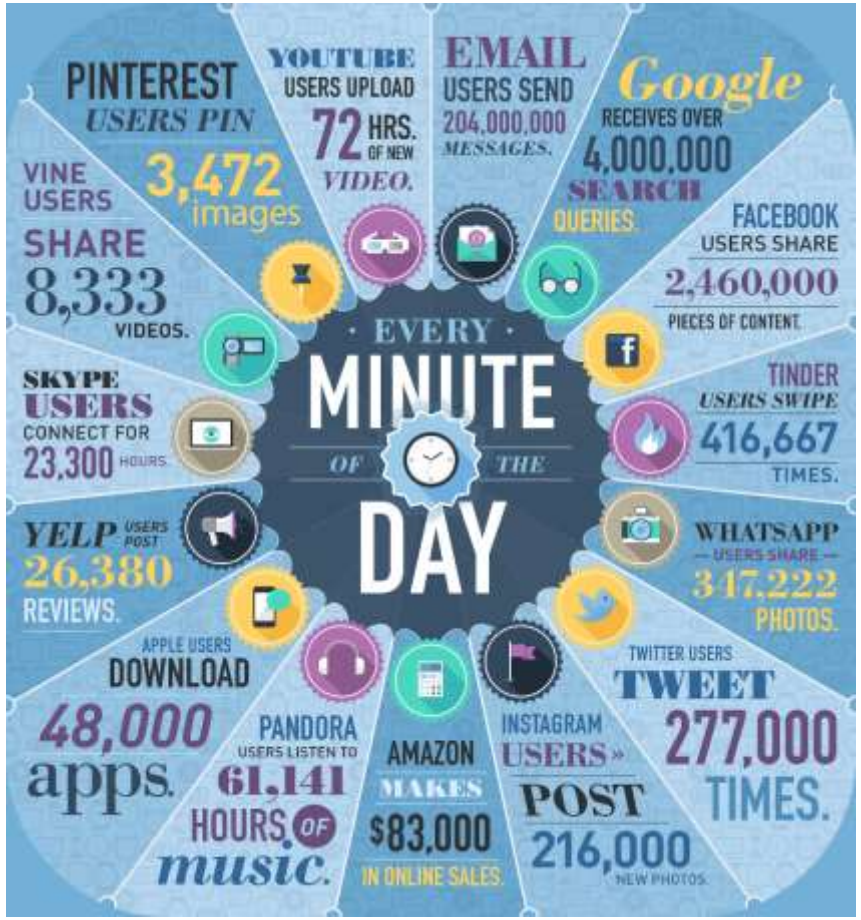
UBER

and Drivers!



Sources: The Uber iPhone App/Wired

Data Science Based-Business Models is the New Normal


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.
VINE USERS
SHARE
8,333
VIDEOS.
FACEBOOK
USERS SHARE
2,460,000
PIECES OF CONTENT.
EVERY
MINUTE
OF THE
DAY
SKYPE
USERS
CONNECT FOR
23,300 HOURS.
TINDER
USERS SWIPE
416,667
TIMES.
YELP USERS POST
26,380
REVIEWS.
WHATSAPP
USERS SHARE
347,222
PHOTOS.
APPLE USERS
DOWNLOAD
48,000
apps.
TWITTER USERS
TWEET
277,000
TIMES.
PANDORA
USERS LISTEN TO
61,141
HOURS OF
music.
AMAZON
MAKES
$83,000
IN ONLINE SALES.
INSTAGRAM
USERS
POST
216,000
NEW PHOTOS.


DOMO
DATA NEVER SLEEPS 2.0
How Much Data is Generated Every 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:

Source: DOMO

Who are Data Scientists?

Activity: Who are Data Scientists?

EXERCISE

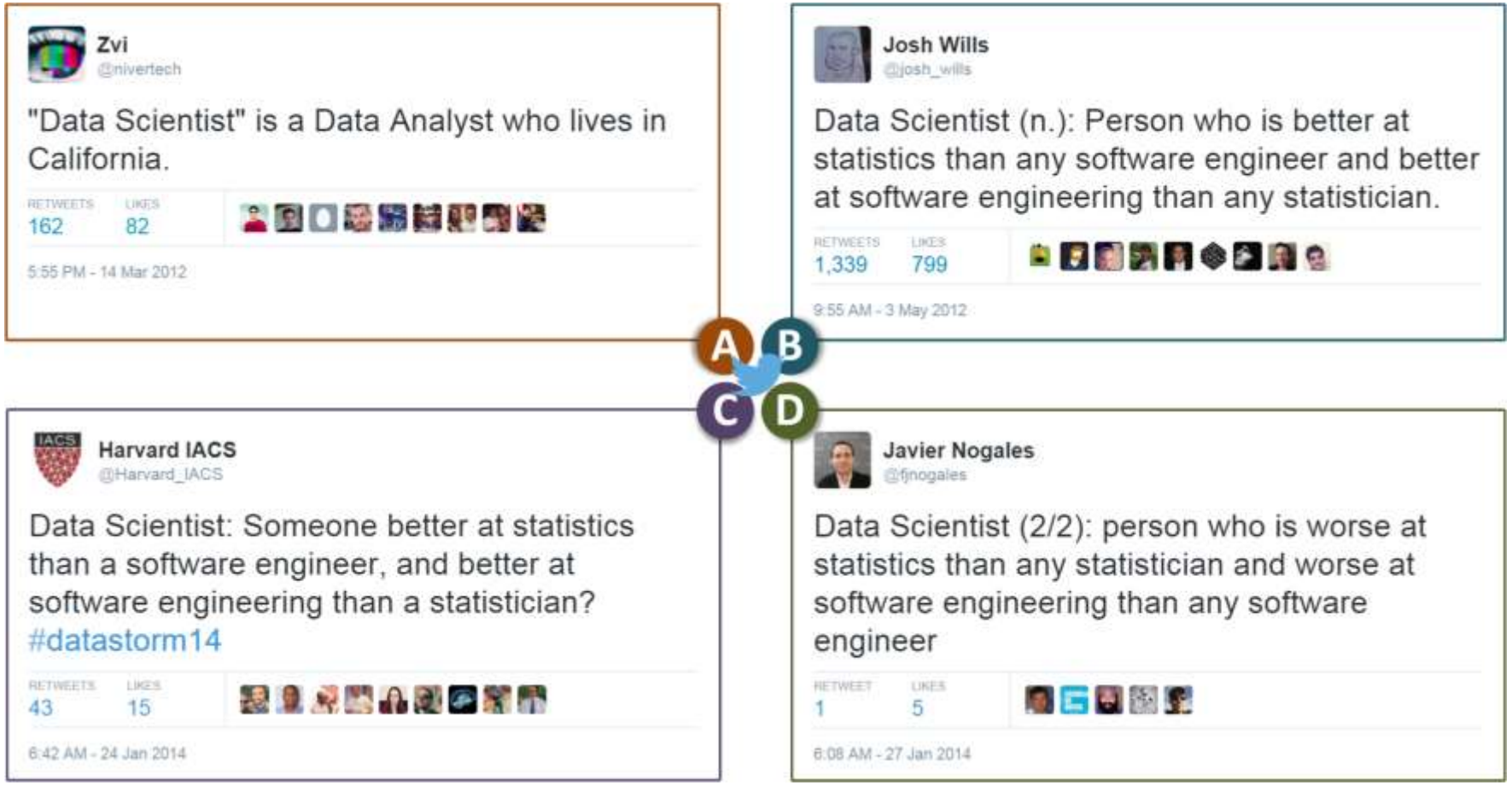
DIRECTIONS (10 minutes)

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



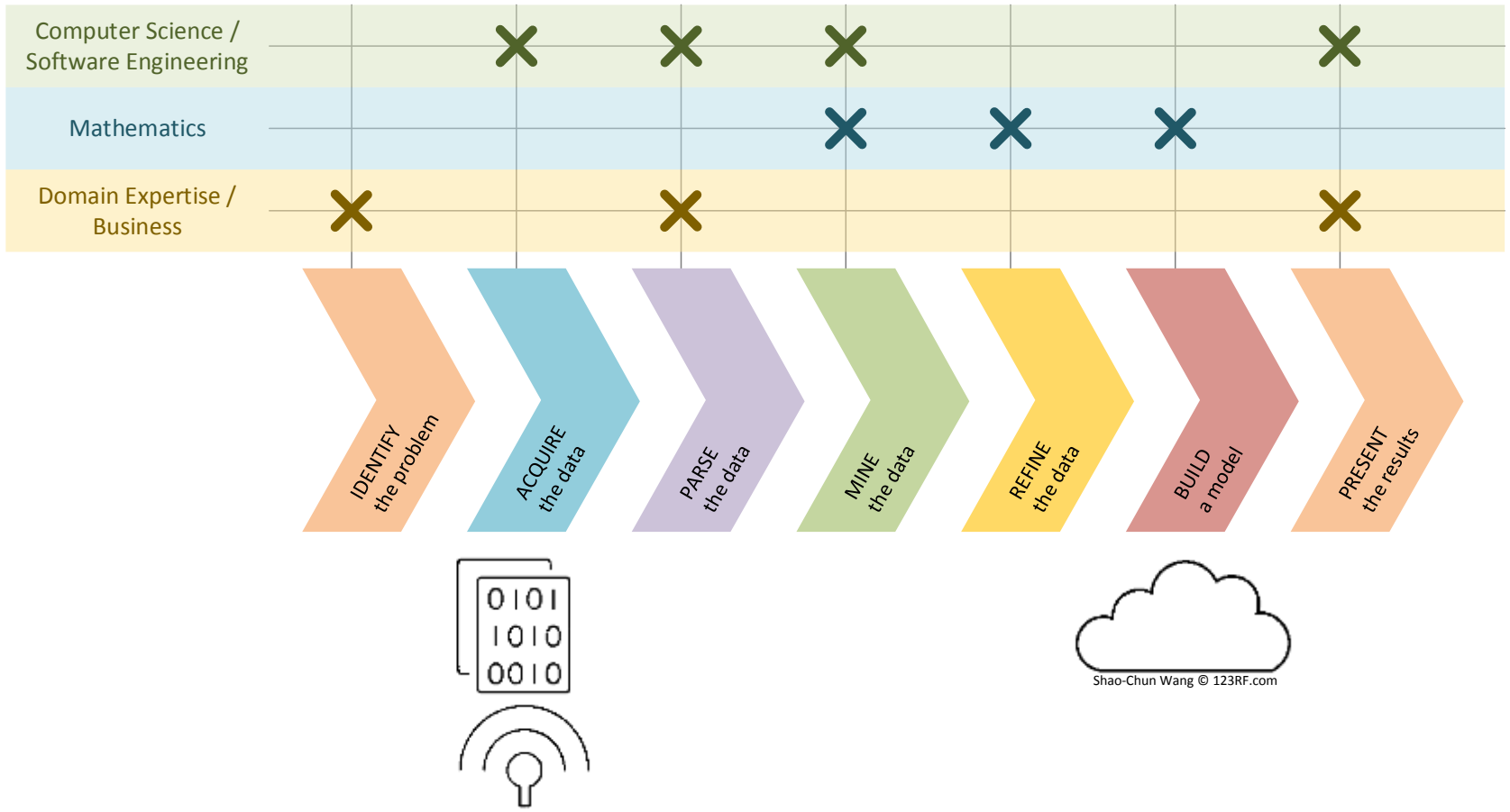
Source: Twitter

What is Data Science? (cont.)

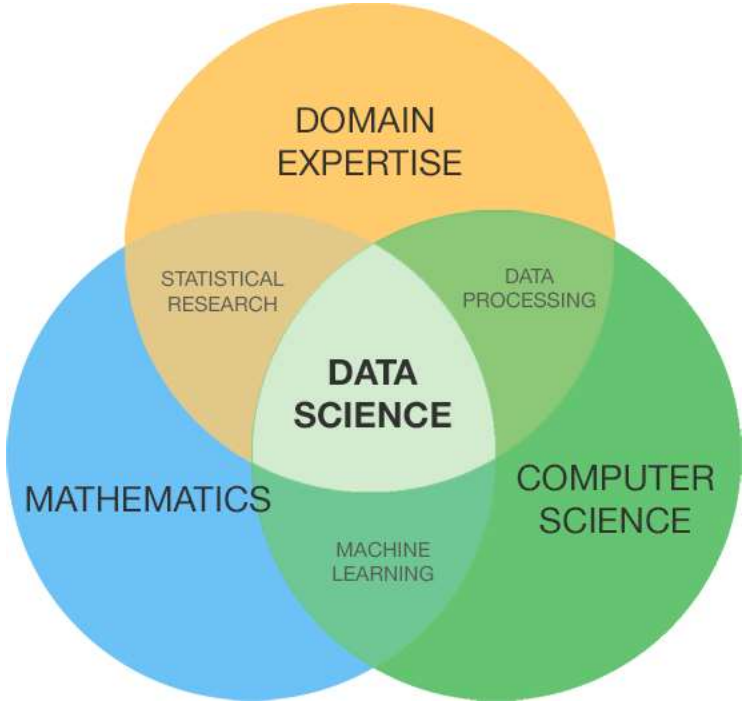
Data is the New Oil of the Digital Economy and IoT, Big Data, DS, and Cloud relate to one another

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

Data Science involves a variety of skillsets, not just one

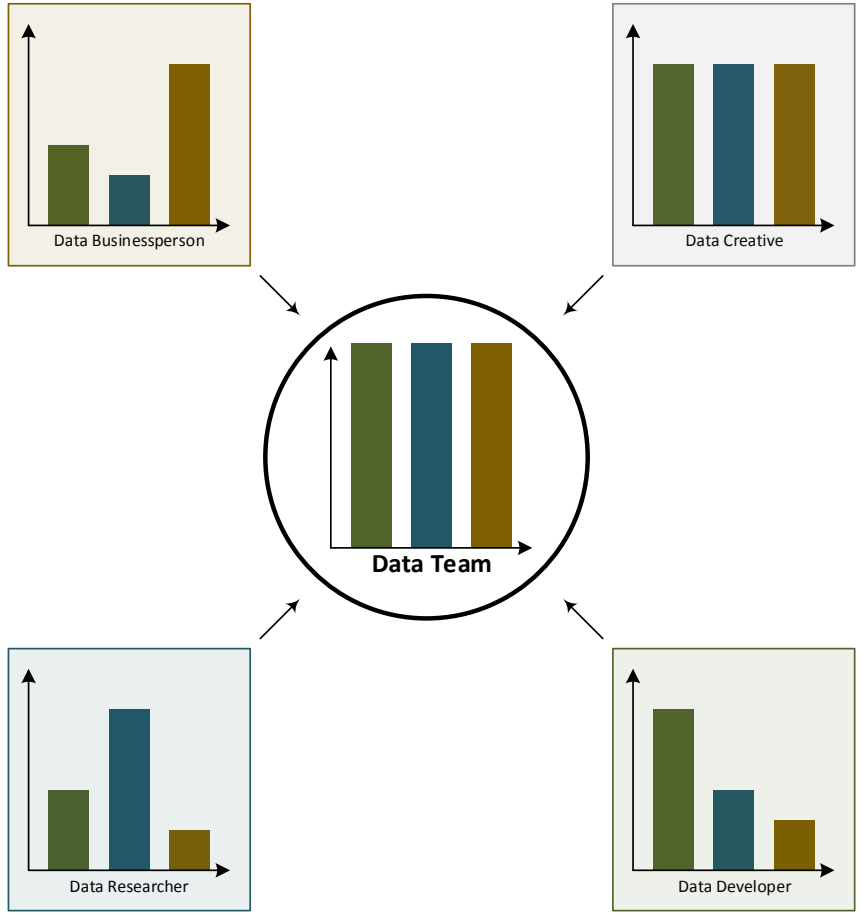


Data Science involves a variety of skillsets, not just one (cont.)



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)

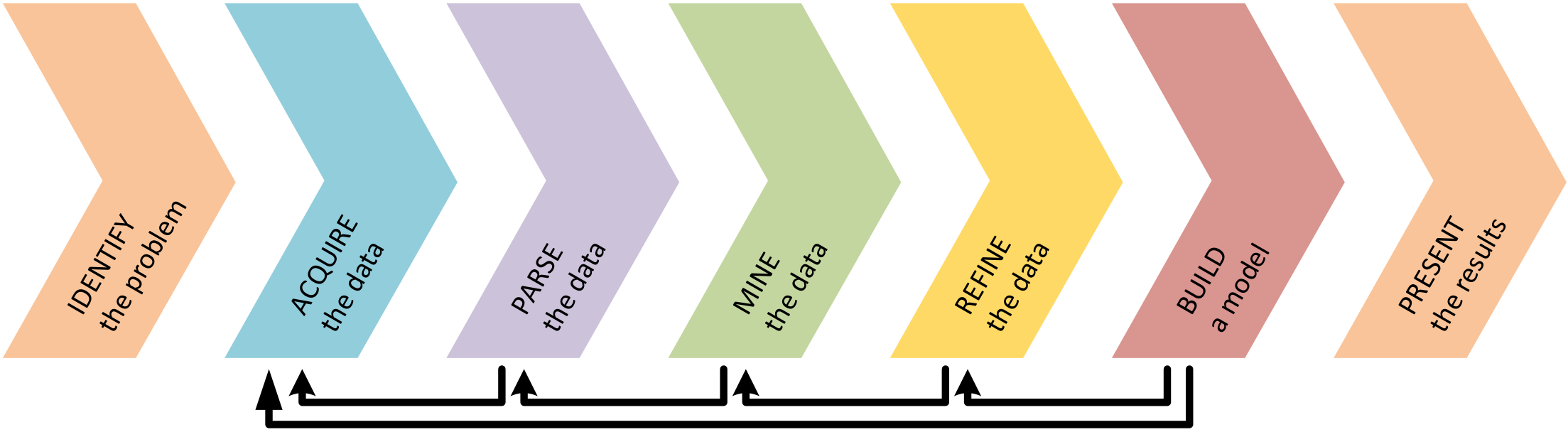
What is the Data Science Workflow for?

- 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

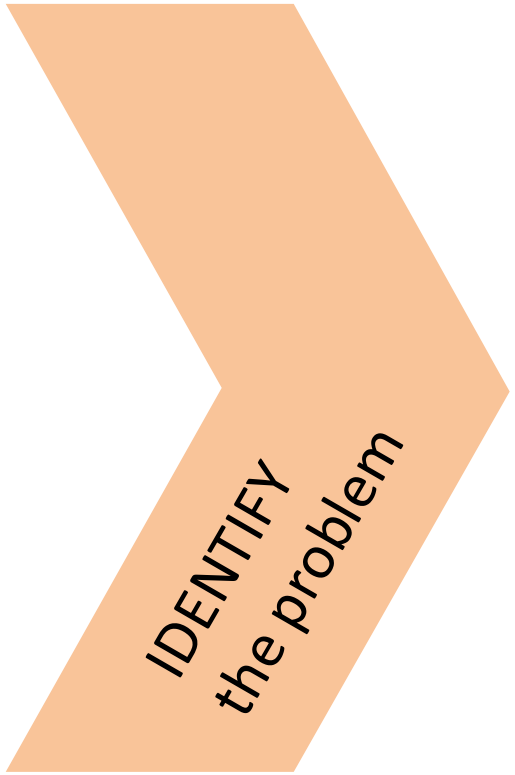
The scientific method:

- 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)



❶ Identify the Problem

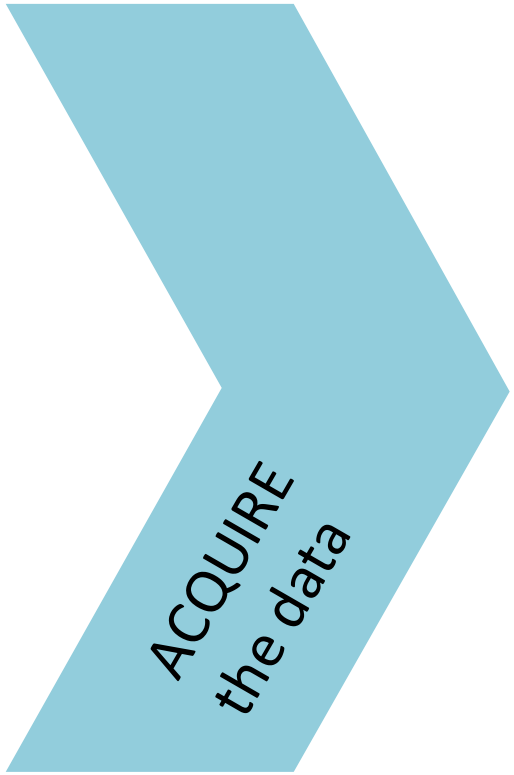


- Identify the Problem
  - Identify business/product objectives
  - Identify and hypothesize goals and criteria for success
  - Create a set of questions for identifying correct dataset

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

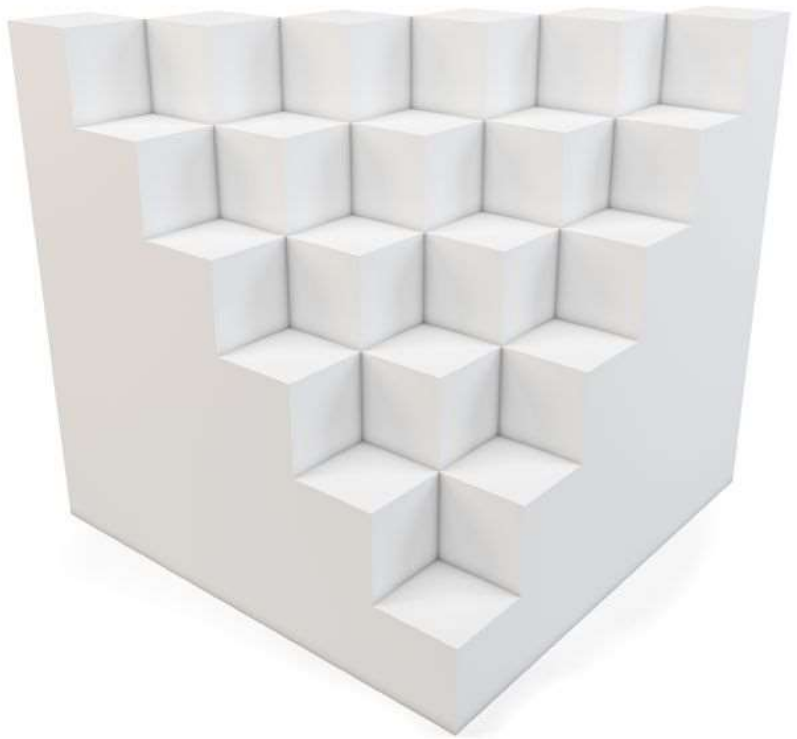
Corina Rosu © 123RF.com

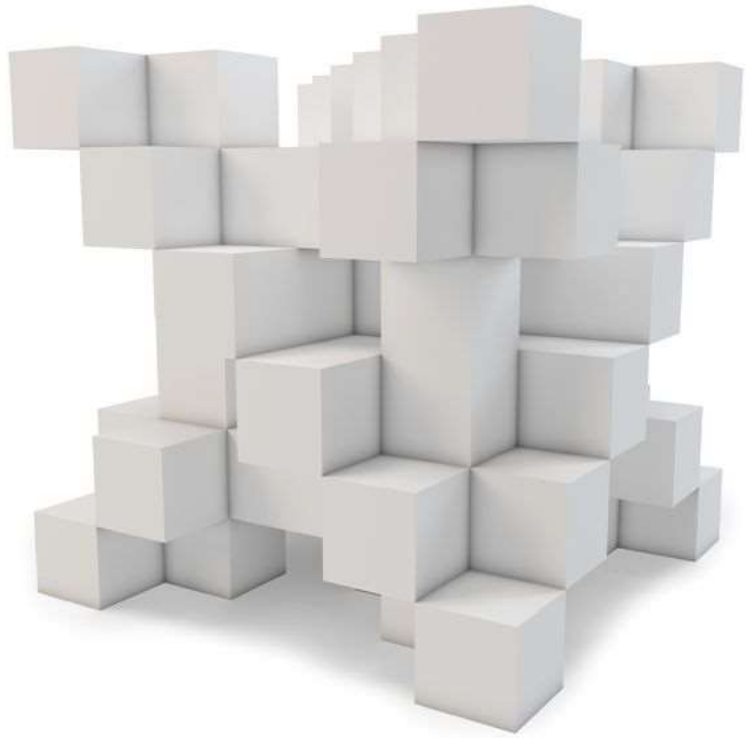
❷ Acquire the Data



- 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

The data can be either unstructured or structured data





dny3d © 123RF.com

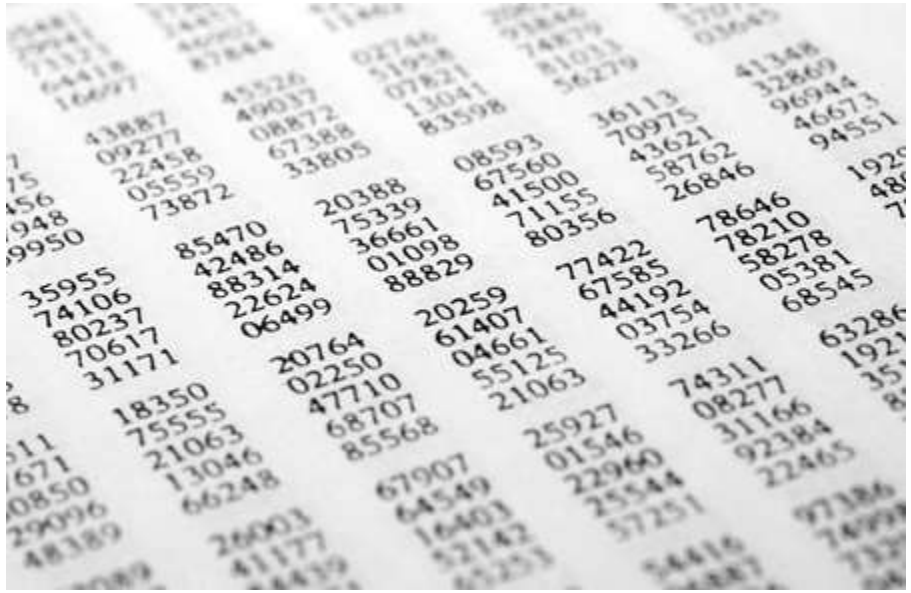
What's an example of unstructured data?

Bundit Chuangboonsri © 123RF.com

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

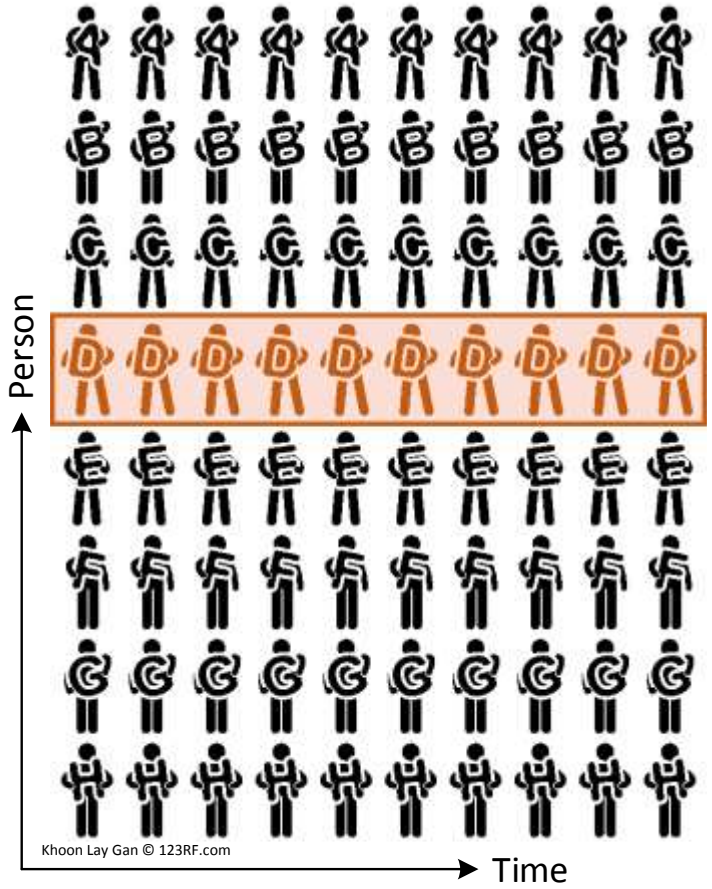
However, most of the course will focus on structured data

- 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

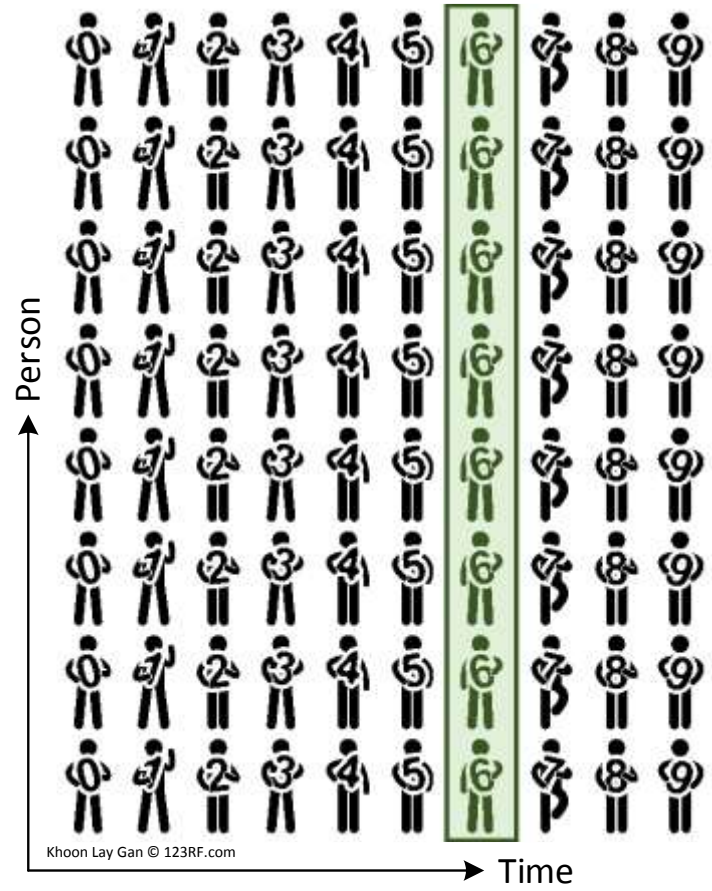
Unstructured data can be longitudinal



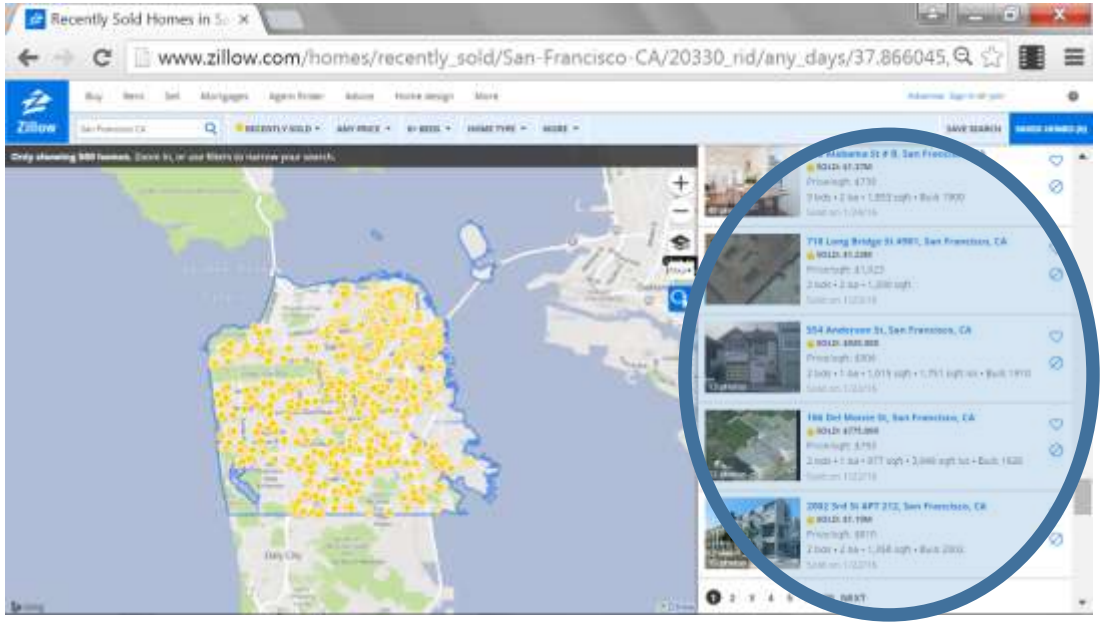
- Sessions 15 and 16 in Unit 3
  - Time Series

Unstructured data can be cross-sectional

- And most of the course will focus on it

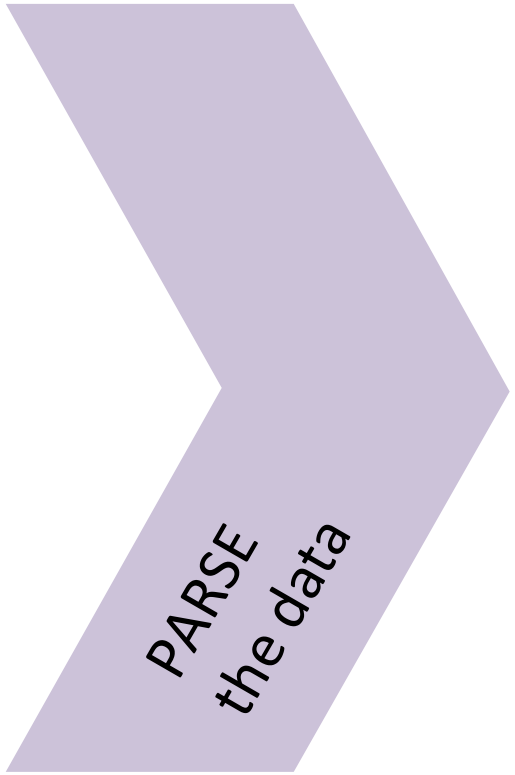


Raw structured data is Messy™…



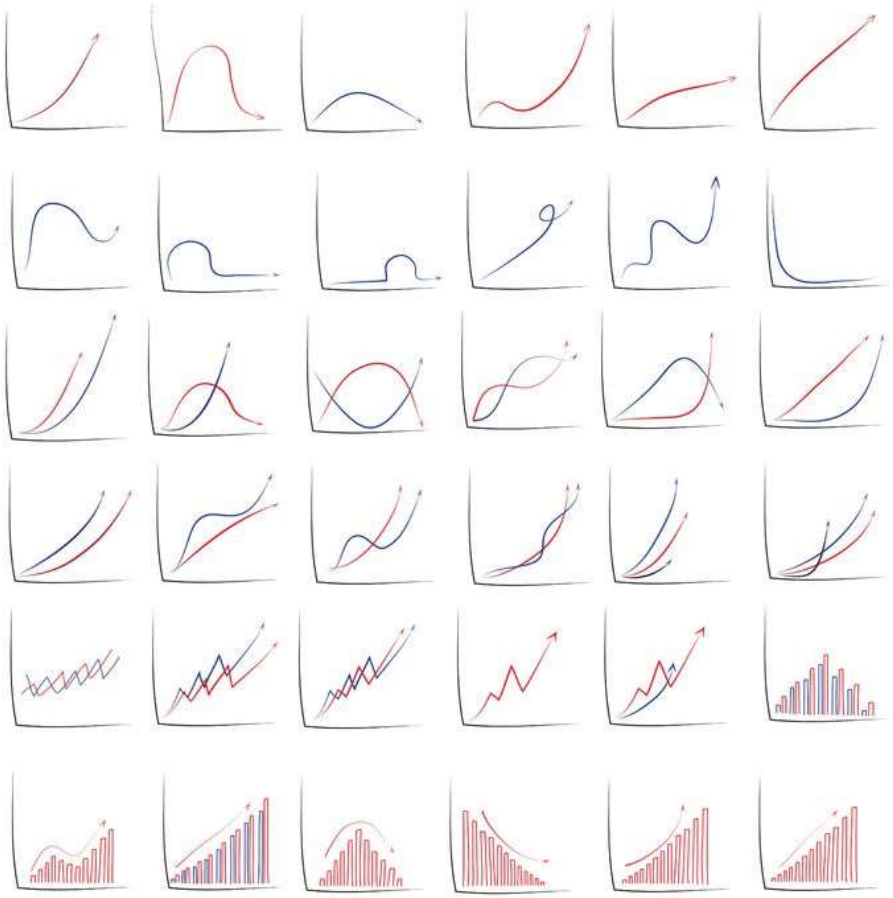

```
<div class="property-info"
id="yui_3_18_1_1_1456167242885_71870"><strong
id="yui_3_18_1_1_1456167242885_71869"><dt class="property-
address" id="yui_3_18_1_1_1456167242885_71868"><a
href="/homedetails/149-Shipley-St-San-Francisco-CA-
94107/15147894_zpid/" class="hdp-link routable" title="149
Shipley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



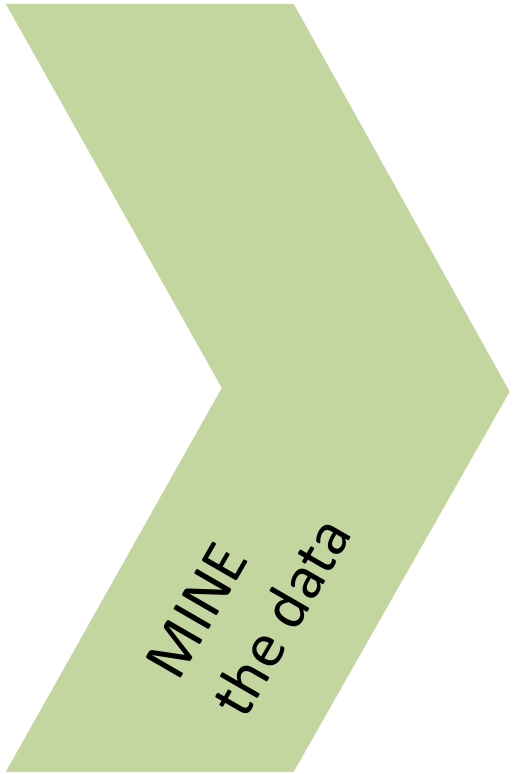
- Parse the Data
  - Read any documentation provided with the data
  - Perform exploratory data analysis
  - Verify the quality of the data

Exploratory Data Analysis



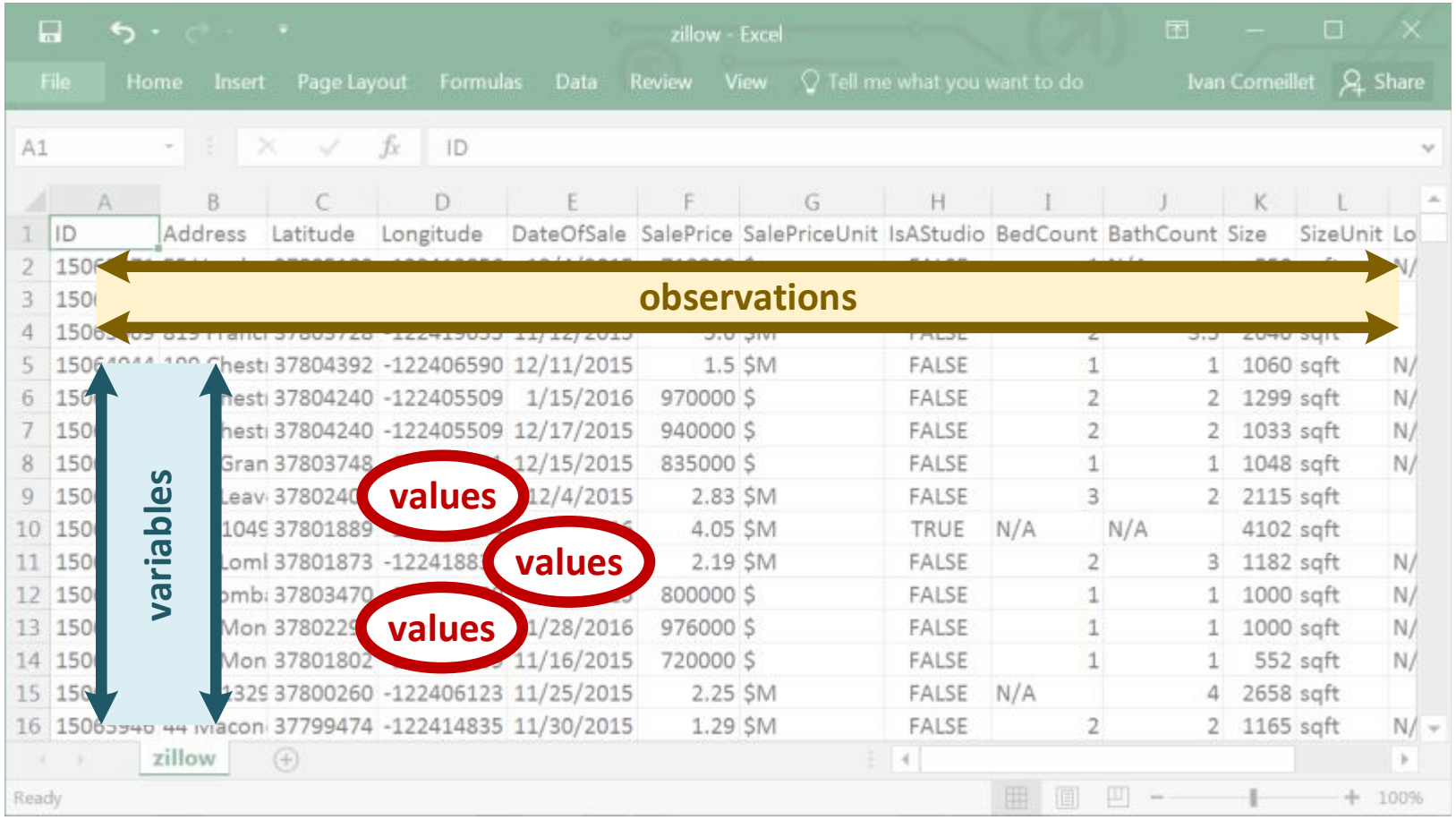
Napat Polchoke © 123RF.com

❹ Mine the Data

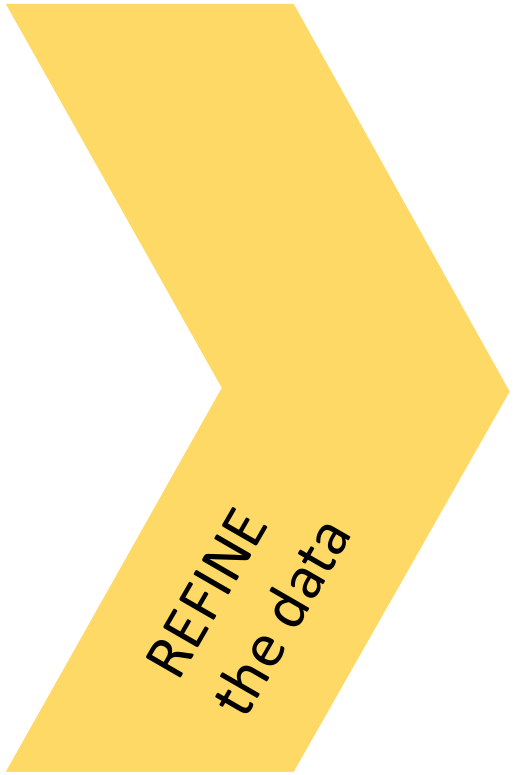


- Mine the Data
  - Determine sampling methodology and sample data
  - Format, clean, slice, and combine data in Python
  - Create necessary derived columns from the data (new data)

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

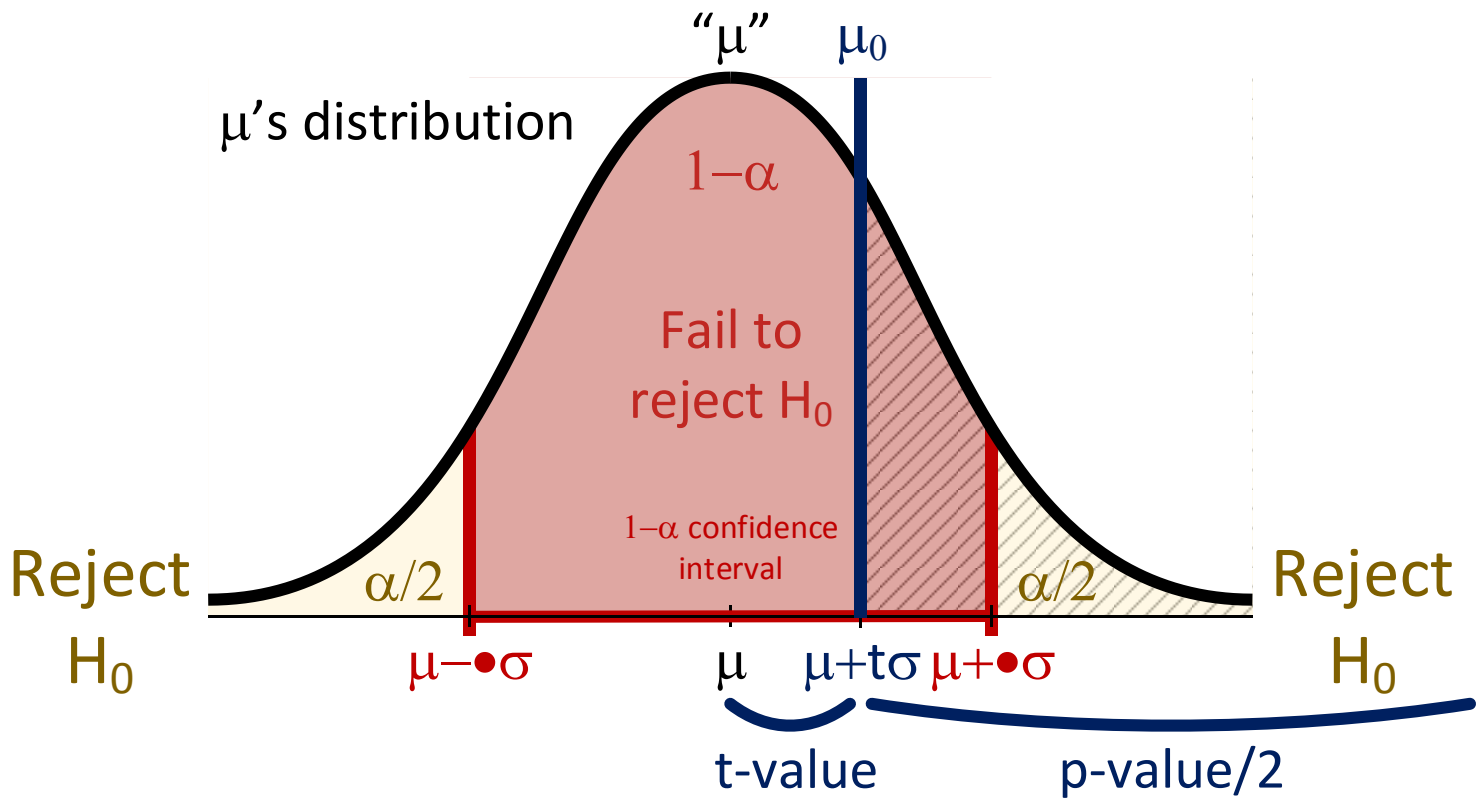


❺ Refine the Data

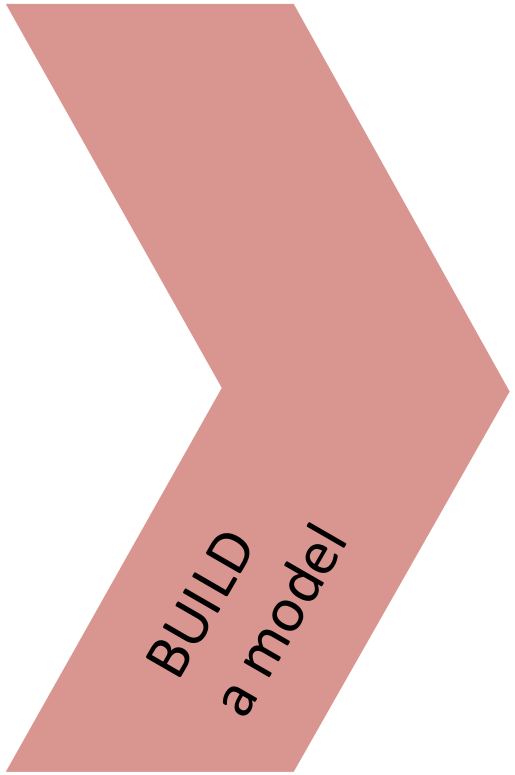


- Refine the Data
  - Identify trends and outliers
  - Apply descriptive and inferential statistics
  - Document and transform data

We will apply inferential statistics



❻ Build a Model

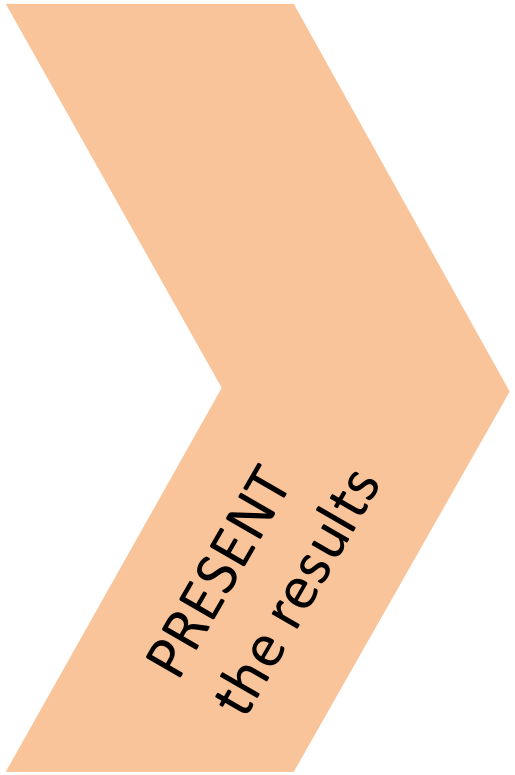


- Build a Model
  - Select appropriate model
  - Build model
  - Evaluate and refine model

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
  - 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

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

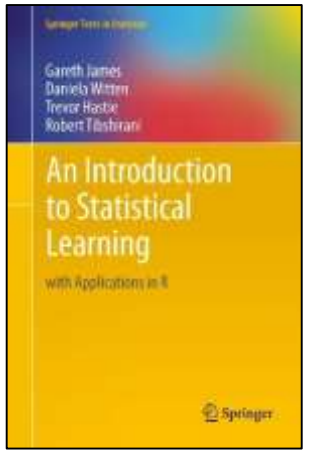
Multiple variants exist but they are pretty much all doing the same thing

- 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

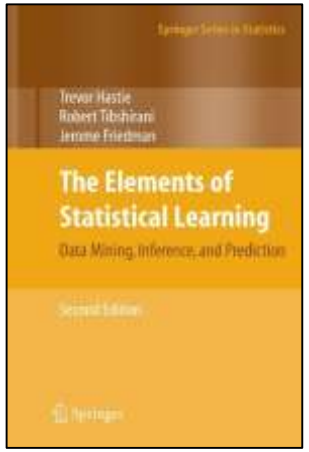
- Peter Huber
  - Inspection
  - Error checking
  - Modification
  - Comparison
  - Modeling and model fitting
  - Simulation
  - What-if analyses
  - Interpretation
  - Presentation of conclusions
- Dataists
  - Obtain
  - Scrub
  - Explore
  - Model
  - Interpret

- 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



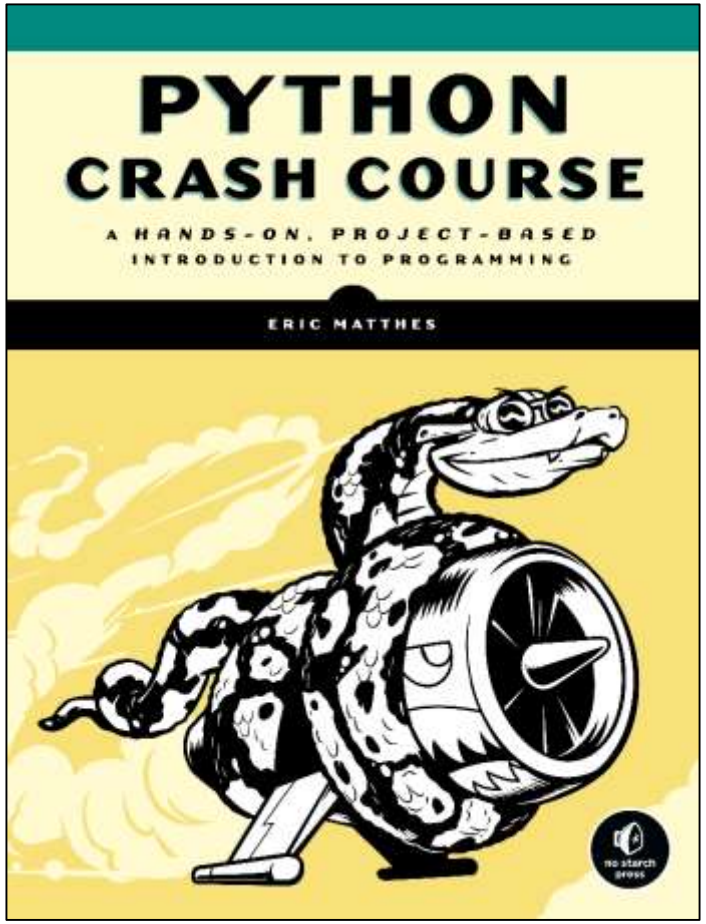
- For a more advanced treatment of these topics, check out The Elements of Statistical Learning: Data Mining, Inference, and Prediction (by Hastie et al.). And yes, the e-book is also free… (here)

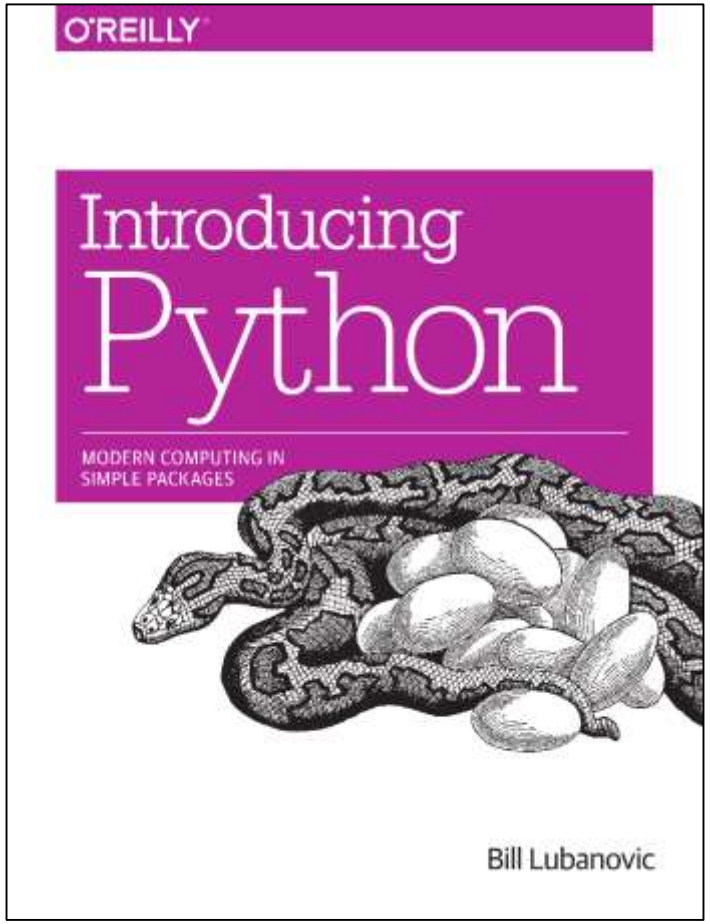
Q & A

Onboarding Review

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

Before Next Class

Before Next Class

- Pre-Work for the next session
  - Have completed the onboarding pre-work
- Install either GitHub Desktop or SourceTree (really a matter of which one you prefer) on your laptop
  - These applications provide a GUI to interact with Git repositories; they also seamlessly setup SSH keys (to authenticate you for your own repositories) should you decide to use the `git` command line tool
- Start looking first unit project
  - Unit Project 1 (due 1 week from now on 5/12)

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