GENERAL ASSEMBLY

INTRODUCTION TO DATA SCIENCE

Mason Gallo

GETTING STARTED

WHO AM I

- Data Scientist
- Open Source Contributor - R & Python
- ML Researcher - Educational Technology
- Data Science Instructor @ GA

Washington University in St. Louis

HAVAS helia. creative with data

Georgia Tech

GA GENERAL ASSEMBLY

HAVAS WORLDWIDE

Omnicom Media Group OMD phd

Duke Clinical Research Institute
From Thought Leadership to Clinical Practice

MY PHILOSOPHY

- If you're not sure, please ask!
- Participate!!!
- NO ONE knows everything
- Anything worth knowing is hard
- Examples then theory
- Break

LEARNING OBJECTIVES

- What is data science and what types of problems does it solve?
- What is machine learning anyway?
- FAQ for newcomers to data science
- The data science workflow
- Where to go next

LEARNING ABOUT YOU

Why did you sign up for Intro to Data Science?

- Thinking about a career change
- Integrate into your current career
- Hiring / managing a data science team
- Expanding your horizons

Anyone that doesn’t fit into the above buckets?

WHAT IS DATA SCIENCE?

WHO USES DATA SCIENCE?

NETFLIX

amazon.com

Google

FiveThirtyEight

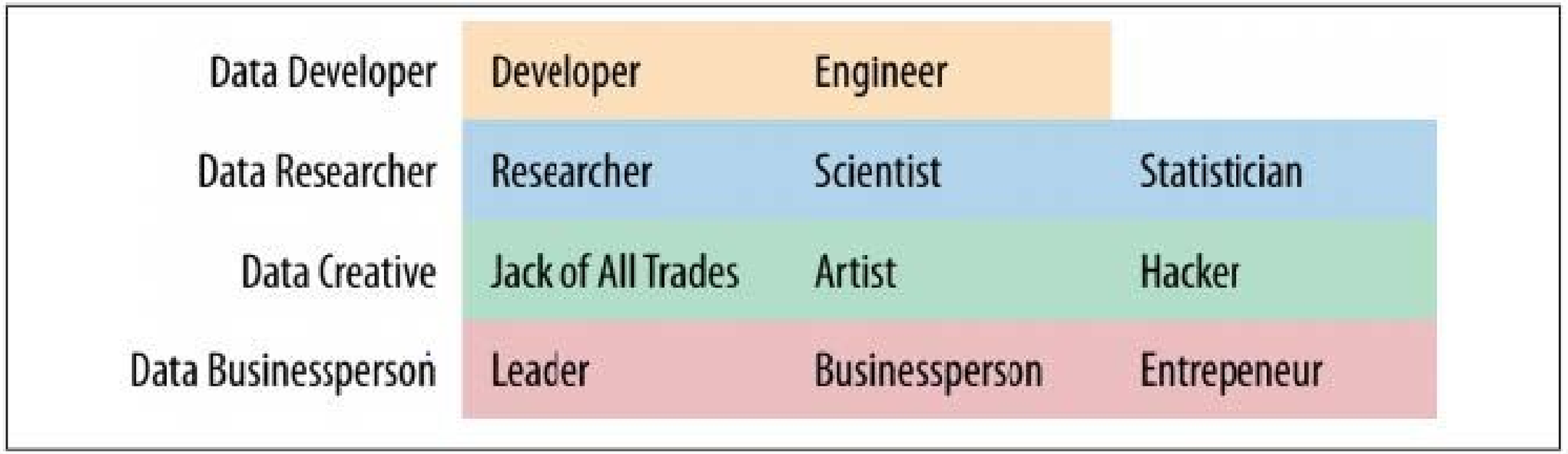
EXAMPLE TASKS

A data scientist might be asked to create a program that:

- Recommends songs users might like
- Predicts the sales of a product next week
- Guesses if a transaction is fraudulent
- Guesses who is in a picture posted to facebook
- Recommends a group of people to target with advertising materials

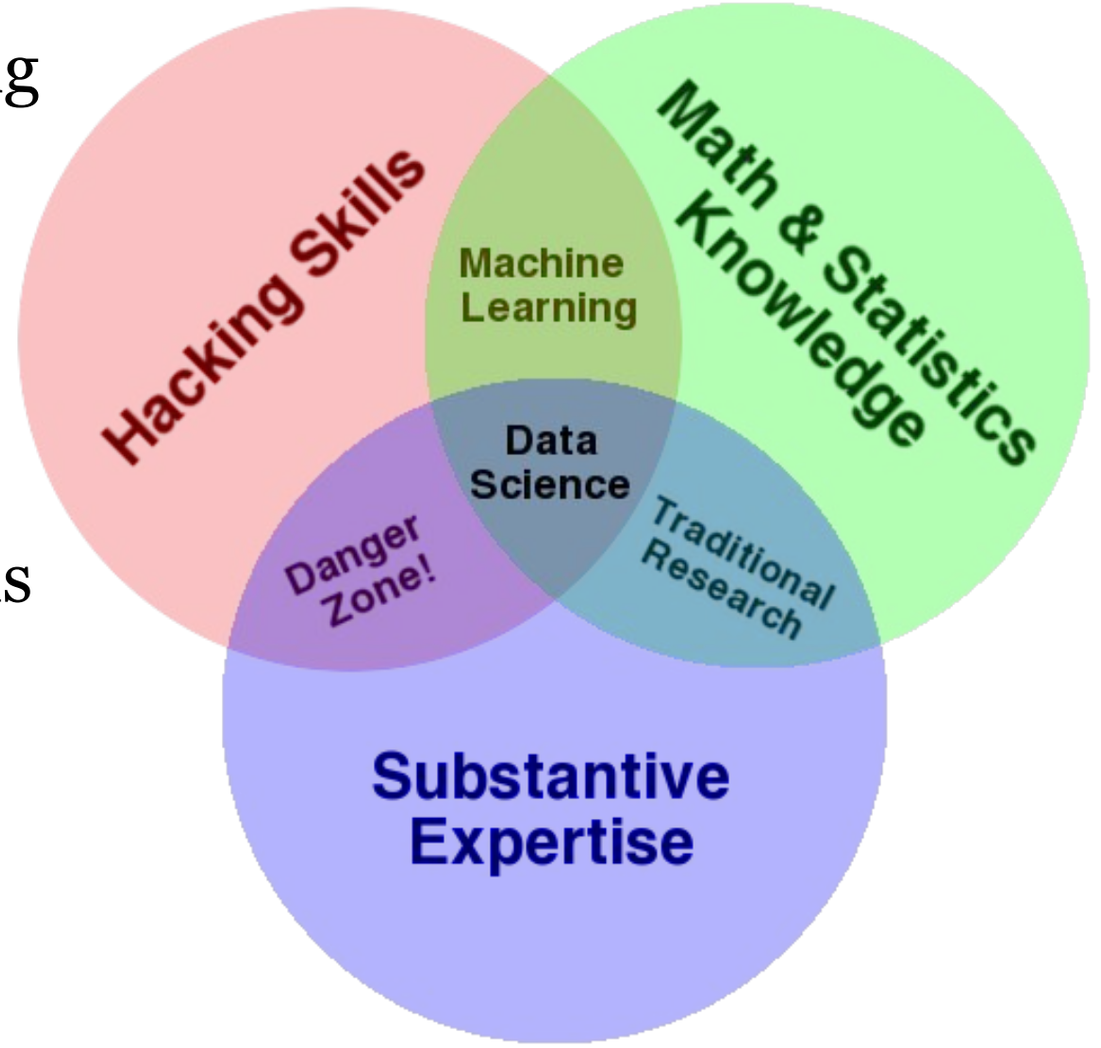
WHAT ARE THE ROLES IN DATA SCIENCE?

- Data Science involves a variety of roles, not just one.



WHAT IS DATA SCIENCE?

- A set of tools and techniques for solving problems with data
- Interdisciplinary problem-solving
- Typically this means applying machine learning to business problems



WHAT IS MACHINE LEARNING ANYWAY?

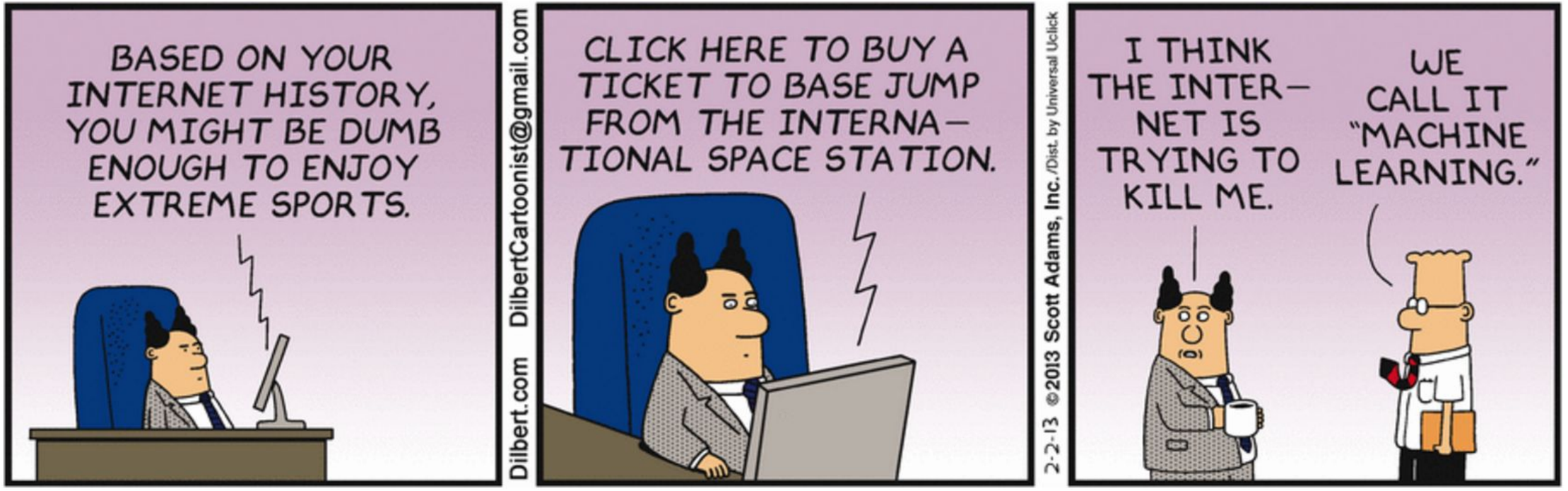
BETTER INGREDIENTS...BETTER DATA

Ingredients (data) matter!

Garbage in → garbage out

SO WHAT IS MACHINE LEARNING ANYWAY?

Saturday February 02, 2013



SO WHAT IS MACHINE LEARNING ANYWAY?

"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)

Arthur Samuel, AI pioneer
Source: Stanford

SO WHAT IS MACHINE LEARNING ANYWAY?

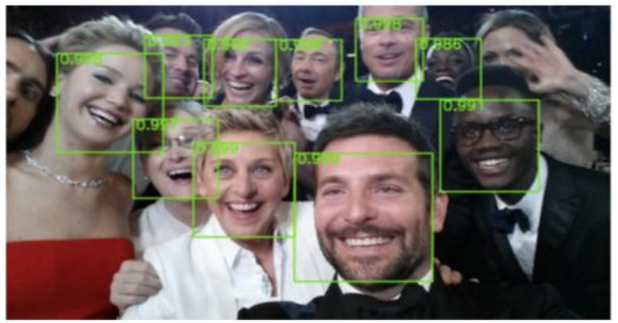
How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

Tom Mitchell, Professor, CMU
(Source: CMU)

WHAT CAN WE DO WITH IT?

Self-driving cars

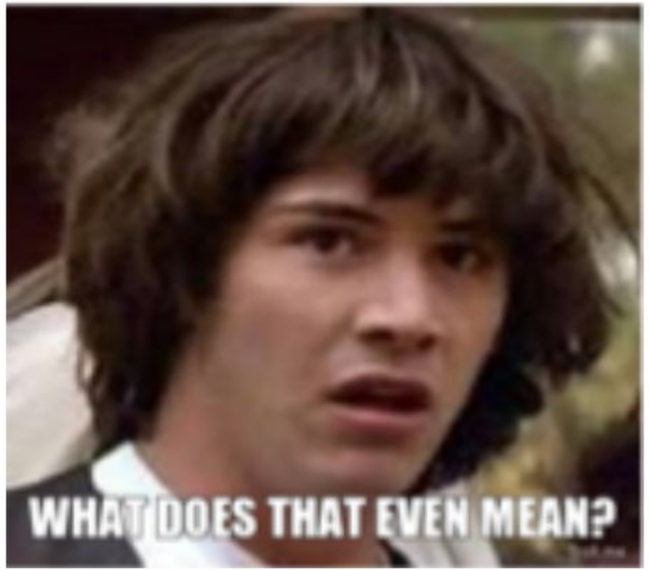
Detecting faces



Catching fraudulent credit card transactions

MOST POPULAR BRANCHES OF ML

2 official branches of machine learning: **supervised** and **unsupervised**



Supervised means we have examples of the correct answer

Unsupervised means we're just exploring to learn more

QUIZ: TYPES OF MACHINE LEARNING

1. Predicting whether a book was actually written by JK Rowling

2. Are there any interesting segments of Estee Lauder customers?

3. How does the value of a purchase affect the probability of fraud?

VOCAB!!!!

- **Training data**: example data (may or may not contain “answers”) that you want to use to learn the answer to a research question
  - Ex: Past year’s sales data, health history of all patients with cancer
- **Machine Learning Algorithm**: a recipe that takes your training data and spits out an “answer” or prediction

TYPES OF DATA SCIENTISTS

HOW I THINK ABOUT THE SKILLS

	Data Engineer	*Data Scientist*	*Data Strategist*
Communication / Domain Expertise			
Computer Science			
Statistics			

DATA STRATEGIST RESPONSIBILTIES MIGHT INCLUDE…

- Interface with interdisciplinary teams to determine business problems
- Serve as subject matter expert for data science and data engineering
- Lead presentations to clients and senior management
- Ensure modeling and data work are “actionable”

Skills needed: SQL, some Python/R

DATA SCIENTIST RESPONSIBILTIES MIGHT INCLUDE…

- Interface with interdisciplinary teams to determine business problems
- Prototype and design machine learning models to solve problems
- Query data in conjunction with data engineering
- Present outcomes of machine learning models to clients / senior management

Skills needed: advanced Python/R, SQL

DATA ENGINEER RESPONSIBILTIES MIGHT INCLUDE...

- Manage data warehouse (think databases and reporting)
- Interface with interdisciplinary teams to determine data storage needs
- Design, build, and launch machine learning models in production
- Design, build, and launch data extraction, transformation, and loading processes in production

Skills needed: advanced Python/Java, SQL

DATA VISUALIZATION AND D3

- Very important but this is evolving into its own field
- Learning beyond the basics of d3.js requires significant effort
- Great data scientists may not be great at visualization (design)
- Designer vs analyst

Skills needed: “design”, front-end skills, SQL

https://d3js.org/

LANGUAGE WARS - R vs Python

STUDENTS ASK ME

- Why did you choose to teach with Python?
- Why not R?
- Will I need to learn R?
- Which language is best?

THERE REALLY ISN'T MUCH OF A DIFFERENCE THAT MATTERS

- For someone new to data science, both languages have the tools to accomplish what you need

- Just pick one (at first)

THIS ISN'T PYTHON EVANGELISM

- The goal isn't to "teach Python"
- Certain domains traditionally prefer Python: tech, startups
- Certain domains traditionally prefer R: health, econometrics
- Even then, if you're good at one, you'll be fine

WHICH DATA SCIENCE TEAMS USE BOTH PYTHON AND R?

- Google
- Facebook
- Twitter
- Spotify
- Dropbox
- Why am I even making a list…?

THE LANGUAGE DOESN'T REALLY MATTER

- Stop worrying!
- You're learning a way of thinking and the tools needed
- What if the languages change in 5 years?

THE DATA SCIENCE WORKFLOW

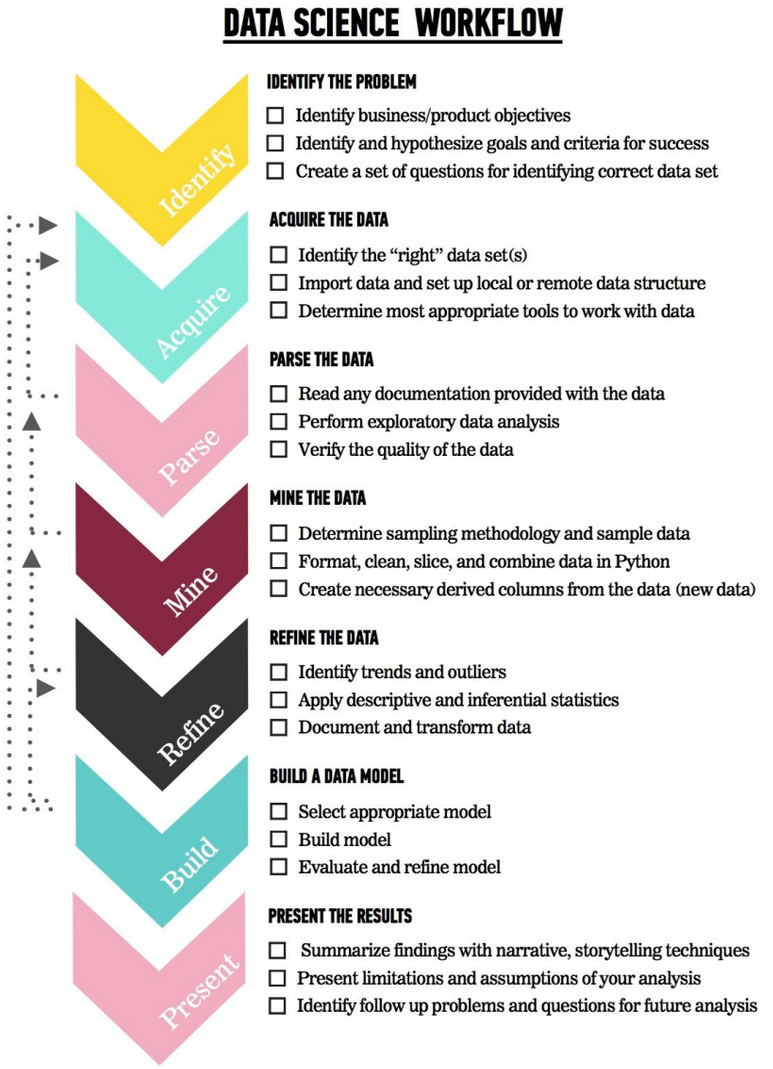
OVERVIEW OF THE DATA SCIENCE WORKFLOW

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

OVERVIEW OF THE DATA SCIENCE WORKFLOW

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



MACY'S EXAMPLE

- Problem Statement: "Using credit card transaction data from the past 2 years at Macy's, determine the factors that lead to increased customer basket size."

- We can use the Data Science workflow to work through this problem.

MACY'S EXAMPLE: IDENTIFY THE PROBLEM

- The objective is basket size $
- Do we need to describe the relationship between predictors and basket size? Do we simply need to make a prediction for each customer visit?
- Create a set of questions to help you identify the correct data set.

MACY'S EXAMPLE: ACQUIRE THE DATA

- Ideal data vs. data that is available: credit card / loyalty card purchases
- Learn about limitations of the data: what about cash purchases?
- What data is available for this example?
- What kind of questions might we want to ask about the data?
  - Representative of the general population?

MACY'S EXAMPLE: ACQUIRE THE DATA

- Questions to ask about the data
  - Is there enough data? % of total purchases / revenue?
  - Does it appropriately align with the question/problem statement?
  - Can the dataset be trusted? How was it collected?
  - Do we have customer level data? How is the data grouped?

MACY'S EXAMPLE: PARSE THE DATA

- Secondary data = we didn't directly collect it ourselves
- Example data dictionary

Variable	Description	Type of Variable
Profession	Title of the account owner	Categorical
Gender	0- male, 1- female	Categorical
Location	Zip code	Categorical
Days Since Last Purchase	Integer	Continuous
Age	Integer	Continuous

MACY'S EXAMPLE: MINE THE DATA

- Think about sampling: can we take a random sample? What about timing?
- Get to know the data
- Explore outliers: extremely large or small basket sizes?
- Address missing values: any trends in what is missing?
- Derive new variables (i.e. columns)

MACY'S EXAMPLE: REFINE THE DATA

- Use statistics and visualization to identify trends

- Example of basic statistics

Variable	Mean (STD) or Frequency (%)
Age	45.7
Gender	70% Female
Days Since Last Purchase	22.4

MACY'S EXAMPLE: REFINE THE DATA

- Descriptive stats help refine by
  - Identifying trends and outliers
  - Deciding how to deal with outliers
  - Applying descriptive and inferential statistics
  - Determining visualization techniques for different data types
  - Transforming data

MACY'S EXAMPLE: CREATE A DATA MODEL

- Select a model based upon the outcome

- Example model statement: "We performed ridge regression to predict the customer basket size at Macy's using credit card transactional data."

MACY'S EXAMPLE: CREATE A DATA MODEL

- The steps for model building are
  - Select the appropriate model
  - Build the model
  - Evaluate and refine the model
  - Predict outcomes and action items

MACY'S EXAMPLE: PRESENT THE RESULTS

- You have to effectively communicate your results for them to matter!
- Ranges from a simple email to a complex web graphic.
- Make sure to consider your audience.
- A presentation for fellow data scientists will be drastically different from a presentation for an executive.

MACY'S EXAMPLE: PRESENT THE RESULTS

- Example presentations and infographics
  - 512 Paths to the White House
  - Who Old Are You?
  - 2015 NFL Predictions
  - NYT Graphics

ACTIVITY: YOU'RE A DATA SCIENCE CONSULTANT

DIRECTIONS

1. Form into 3 groups.
2. Each group will be assigned questions from the following 4 slides
3. Pretend you are a newly-hired Data Scientist Consultant at each company
4. Think of the data science workflow we just discussed!

CASE STUDY: FRAUD

- Problem:
  - Criminals are stealing American Express credit cards and performing fraudulent transactions. Fraud costs the company money and hurts loyal customers.

- Intuitively, give 3-5 factors that might lead to fraud
- Give 2 examples of why learning the factors that contribute to fraud might be difficult or biased

CASE STUDY: SALES

- Problem:
  - Hershey's wants a projection of their sales for the next 8 weeks. They've given you access to their database with the past 5 years of sales data.

- Give 3 questions we want to ask Hershey's before we begin
- Give 2 examples of how seasonality affects this problem

CASE STUDY: ONLINE ADVERTISING

- Problem:
  - TD Ameritrade wants to determine if a green background color or yellow background color in a banner ad is better.

- Give 2 questions you would ask TD Ameritrade before getting started
- Give 2 examples why the nature of the internet might hurt our ability to conduct a valid experiment

WHERE TO GO NEXT - DATA SCIENCE SKILLSET

FREE MATH AND STATISTICS RESOURCES

The entire math track on KHANACADEMY

Andrew Ng's Machine Learning course on coursera

A/B Testing course on UDACITY

Elements of Statistical Learning book (FREE)
http://statweb.stanford.edu/~tibs/ElemStatLearn/

FREE COMPUTER SCIENCE RESOURCES

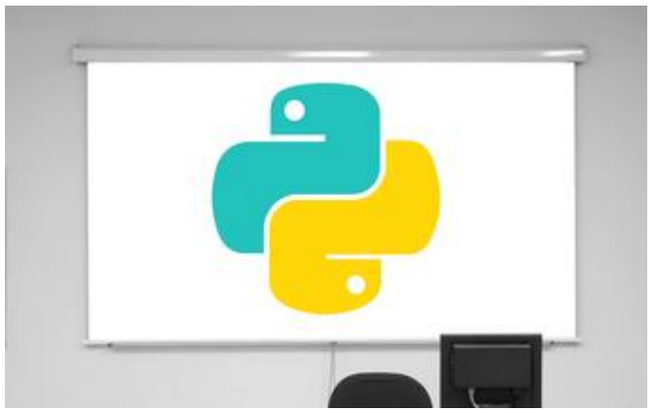
The computer science track on KHANACADEMY

Intro to Computer Sci on 

Think Python book (FREE) http://greenteapress.com/wp/think-python/

INTRODUCTION TO PYTHON PROGRAMMING @ GA

In this class, you'll learn all about Python - including how to get started, what advantages and disadvantages Python provides as a programming language, the essentials of programming in Python, and what tools are available to build applications in Python.



Bias: I teach this class

https://generalassemb.ly/education/introduction-to-python-programming/new-york-city

SQL BOOTCAMP @ GA

SQL provides powerful but reasonably simple tools for data analysis and handling. This bootcamp will take absolute beginners through the basics of SQL to an ability to write queries with confidence. We will use a combination of lecture and in-class exercises to ensure that students leave with a working grasp of SQL fundamentals.

Bias: I teach this class

https://generalassemb.ly/education/sql-bootcamp/new-york-city/

PART TIME DATA SCIENCE @ GA

Big picture of Data Science with Python toolset. Expect 6 hours of instruction per week and several more hours per week of homework and projects. Must apply and pass interview to be accepted.

Bias: I teach this class

https://generalassemb.ly/education/data-science

DATA SCIENCE IMMERSIVE @ GA

Intensive 12-week full-time career accelerator. Meant for those who want to career switch ASAP. Must apply and pass skills test to be accepted.

https://generalassemb.ly/education/data-science-immersive

IF YOU LIKE WHAT YOU'VE BEEN HEARING...

Connect with me on LinkedIn:

https://www.linkedin.com/in/masongallo

Don't forget to connect with your fellow classmates!

Q/A

Keep in touch!

masonagallo@gmail.com
https://www.linkedin.com/in/masongallo
http://masongallo.github.io/
@ohheyitsmason