INTRODUCTION TO DATA SCIENCE

Mason Gallo

GETTING STARTED

WHO AM I

‣ Data Scientist

‣ Open Source Contributor

‣ ML Researcher - Educational Technology

‣ Data Science Instructor @ GA

MY PHILOSOPHY

‣ If you're not sure, please ask!

‣ Participate!!!

‣ NO ONE knows everything

‣ Anything worth knowing is hard

‣ Examples then theory

‣ Break

4

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

DATA SCIENCE

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

‣ A set of tools and techniques for data

‣ Interdisciplinary problem-solving

‣ Application of scientific techniques to practical problems
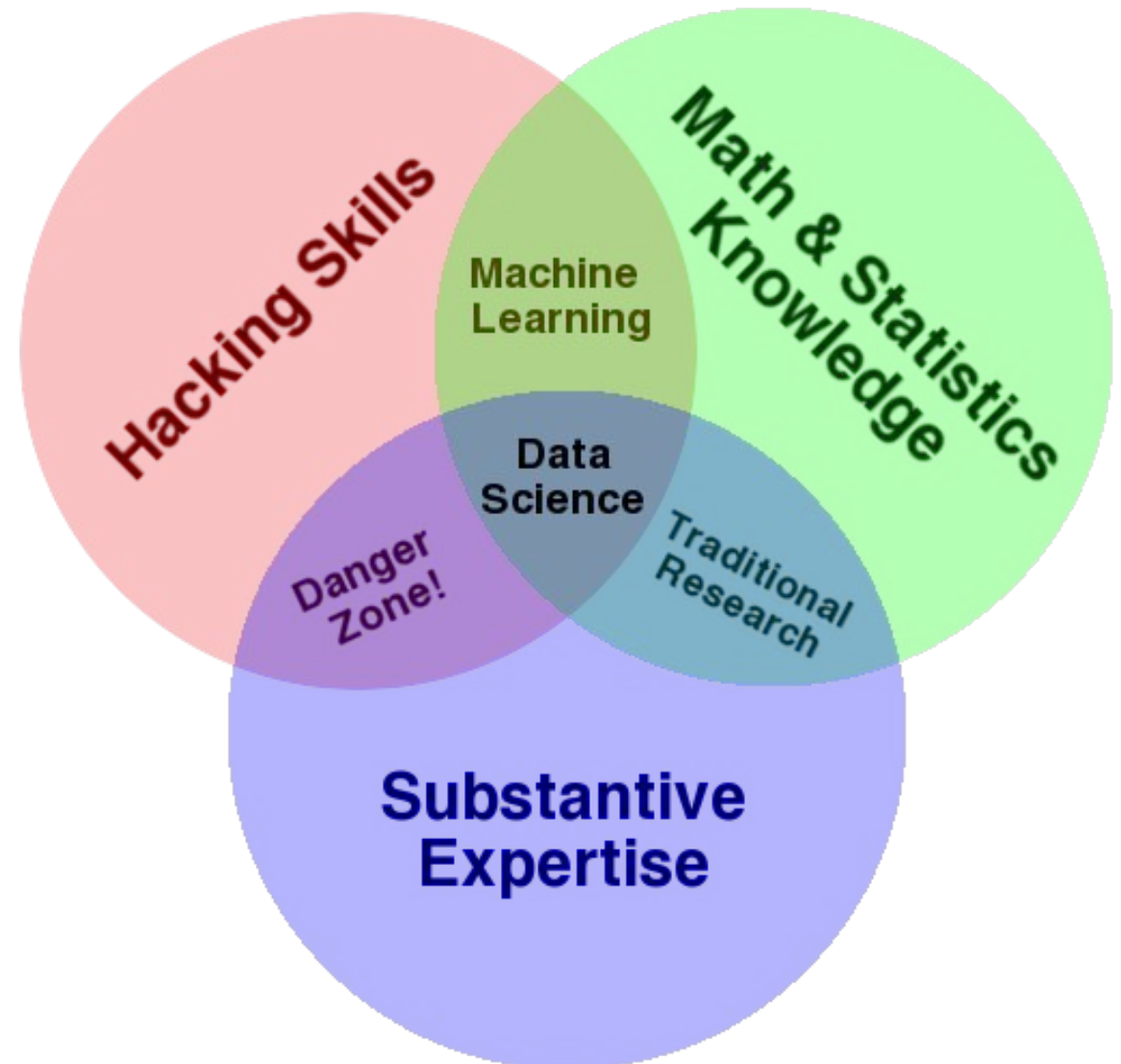
WHO USES DATA SCIENCE?

QUOTES

Michael E. Driscoll
@medriscoll

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

RETWEETS FAVORITES
35   26

4:57 PM - 17 Jul 2012
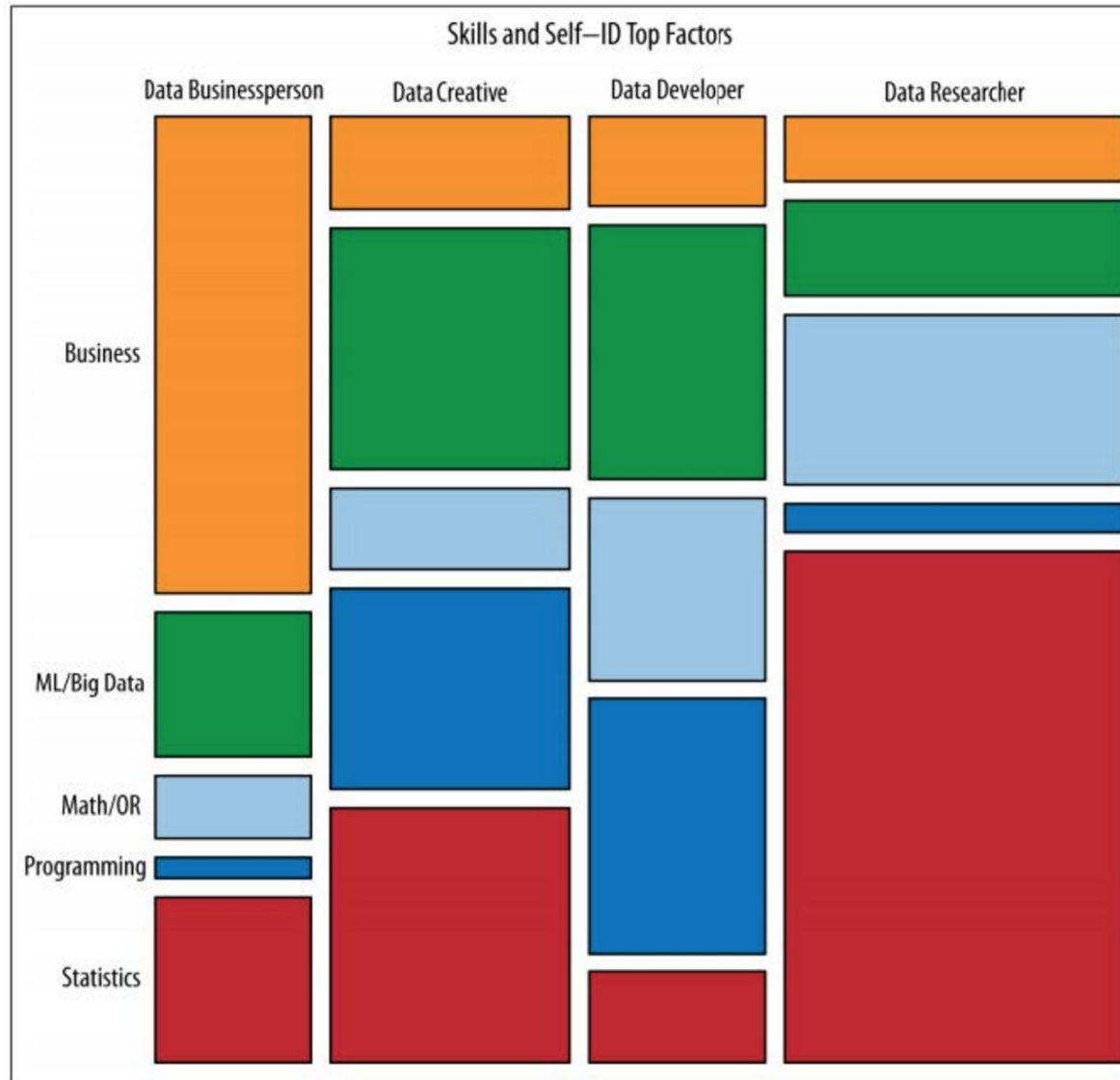
WHAT ARE THE ROLES IN DATA SCIENCE?

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

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

TYPES OF DATA SCIENTISTS

YOU MIGHT SEE THIS ON BLOGS...

HOW I THINK ABOUT THE SKILLS

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 models in production

‣ Design, build, and launch data extraction, transformation, and loading processes in production

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"

DATA VISUALIZATION AND D3

‣ Very important but this is evolving into its own field

‣ Learning beyond the basics of d3.js requires significant effort

‣ Most data scientists are not experts

‣ Designer vs analyst

https://d3js.org/

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

WHAT IS MACHINE LEARNING ANYWAY?

BETTER INGREDIENTS...BETTER DATA

Ingredients (data) matter!

Garbage in → garbage out

SO WHAT IS MACHINE LEARNING ANYWAY?
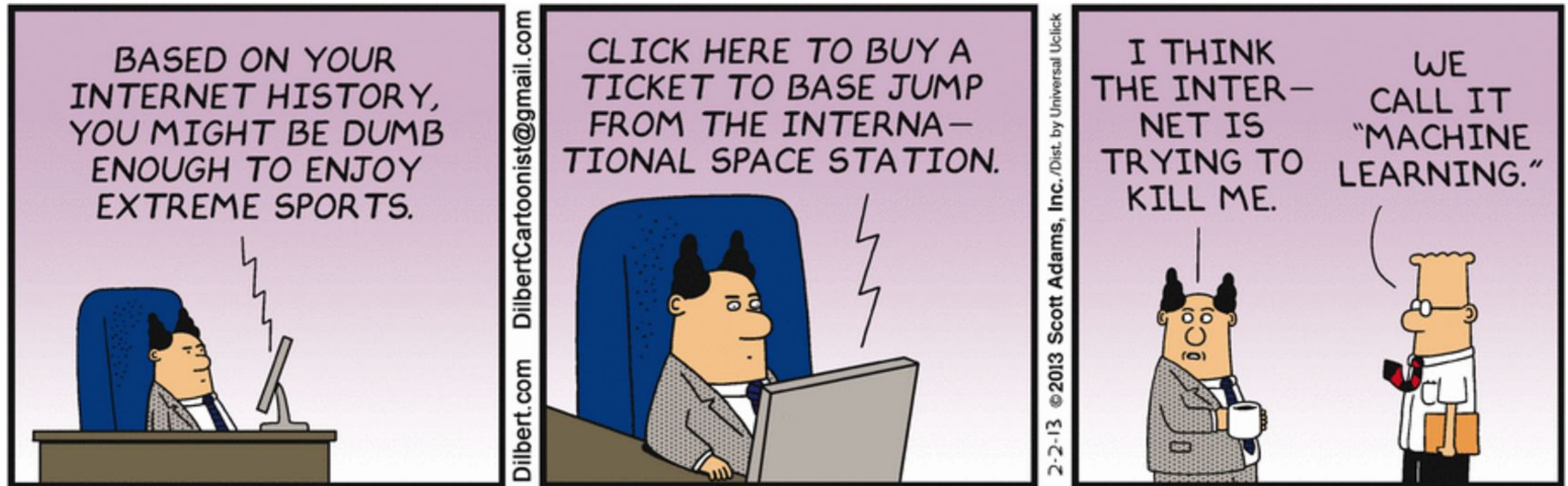
SO WHAT IS MACHINE LEARNING ANYWAY?

"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
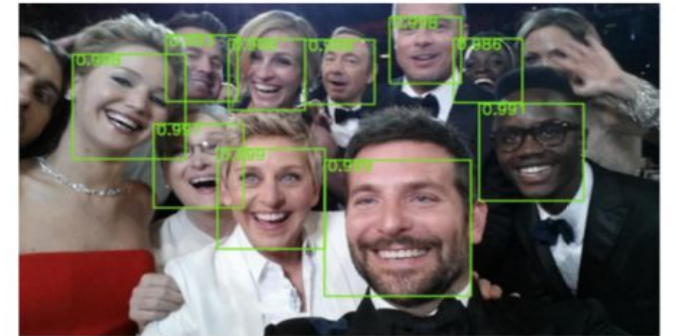
Catching fraudulent credit card transactions

MOST POPULAR BRANCHES OF ML

2 official branches of machine learning: **supervised** and **unsupervised**
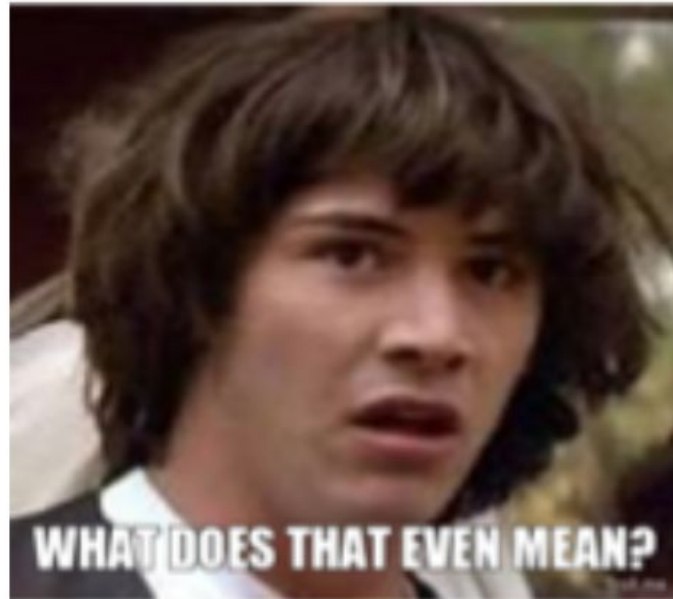
Supervised means we have examples of the correct answer

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

DATA SCIENCE

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

‣ However, the languages have different syntax and differing opinions on how to do things

‣ 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 PYTHON OR 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

DATA SCIENCE WORKFLOW

Identify

IDENTIFY THE PROBLEM
- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

Acquire

ACQUIRE THE DATA
- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

Parse

PARSE THE DATA
- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

Mine

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)

Refine

REFINE THE DATA
- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

Build

BUILD A DATA MODEL
- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

Present

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

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Identify

IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Acquire

ACQUIRE THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

PARSE THE DATA

☐ Read any documentation provided with the data

☐ Perform exploratory data analysis

☐ Verify the quality of the data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Mine

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)

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Refine

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Build

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Present

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

FUTURAMA EXAMPLE

‣ Problem Statement: "Using Planet Express customer data from January 3001-3005, determine how likely previous customers are to request a repeat delivery using demographic information (profession, company size, location) and previous delivery data (days since last delivery, number of total deliveries)."

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

FUTURAMA EXAMPLE: IDENTIFY THE PROBLEM

‣ Identify the business/product objectives.

‣ Identify and hypothesize goals and criteria for success.

‣ Create a set of questions to help you identify the correct data set.

FUTURAMA EXAMPLE: ACQUIRE THE DATA

‣ Ideal data vs. data that is available

‣ Learn about limitations of the data.

‣ What data is available for this example?

‣ What kind of questions might we want to ask about the data?

FUTURAMA EXAMPLE: ACQUIRE THE DATA

‣ Questions to ask about the data

‣ Is there enough data?

‣ Does it appropriately align with the question/problem statement?

‣ Can the dataset be trusted? How was it collected?

‣ Is this dataset aggregated? Can we use the aggregation or do we need to get it pre-aggregated?

FUTURAMA 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
Company Size	1- small, 2- medium, 3- large	Categorical
Location	Planet of the company	Categorical
Days Since Last Delivery	Integer	Continuous
Number of Deliveries	Integer	Continuous

FUTURAMA EXAMPLE: PARSE THE DATA

‣ Questions to ask while parsing

  ‣ Is there documentation for the data? Is there a data dictionary?

  ‣ What kind of filtering, sorting, or simple visualizations can help understand the data?

  ‣ What information is contained in the data?

  ‣ What data types are the variables?

  ‣ Are there outliers? Are there trends?

FUTURAMA EXAMPLE: MINE THE DATA

‣ Think about sampling

‣ Get to know the data

‣ Explore outliers

‣ Address missing values

‣ Derive new variables (i.e. columns)

FUTURAMA EXAMPLE: MINE THE DATA

‣ Common steps while mining the data

  ‣ Sample the data with appropriate methodology

  ‣ Explore outliers and null values

  ‣ Format and clean the data

  ‣ Determine how to address missing values

  ‣ Format and combine data; aggregate and derive new columns

FUTURAMA EXAMPLE: REFINE THE DATA

‣ Use statistics and visualization to identify trends

‣ Example of basic statistics

Variable	Mean (STD) or Frequency (%)
Number of Deliveries	50.0 (10)
Earth	50 (10%)
Amphibios 9	100 (20%)
Bogad	100 (20%)
Colgate 8	100 (20%)
Other	150 (30%)

FUTURAMA 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

FUTURAMA EXAMPLE: CREATE A DATA MODEL

‣ Select a model based upon the outcome

‣ Example model statement: "We completed a logistic regression using Statsmodels v. XX. We calculated the probability of a customer placing another order with Planet Express."

‣ Steps for model building

FUTURAMA 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

FUTURAMA 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.

FUTURAMA EXAMPLE: PRESENT THE RESULTS

Key factors of a good presentation include

- Summarize findings with narrative and storytelling techniques

- Refine your visualizations for broader comprehension

- Present both limitations and assumptions

- Determine the integrity of your analyses

- Consider the degree of disclosure for various stakeholders

- Test and evaluate the effectiveness of your presentation beforehand

FUTURAMA EXAMPLE: PRESENT THE RESULTS

‣ Example presentations and infographics

    ‣ [512 Paths to the White House](#)

    ‣ [Who Old Are You?](#)

    ‣ [2015 NFL Predictions](#)

ACTIVITY: YOU'RE A DATA SCIENCE CONSULTANT

DIRECTIONS

EXERCISE

1. Form into 4 groups.
2. Each group will be assigned a question 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.

‣ What's a good research question? Intuitively, what factors might lead to fraud? What's the role of outliers here?

CASE STUDY: SALES

‣ Problem:

  ‣ Hershey's needs to decide how much candy to produce this week in order to sell it next week.

‣ What kind of questions might we want to ask Hershey's before we collect data? How does timing factor into solving this problem?

CASE STUDY: SEGMENTATION

‣ Problem:

  ‣ Estee Lauder wants to figure out to whom they should market mascara.

‣ What's a good research question? Why is this a different type of problem? What type of data could we use and should we trust it?

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".

‣ What's a good research question? Why is this more difficult? What information is missing here?

WHERE TO GO NEXT - DATA SCIENCE SKILLSET

Requests: HTTP for Humans

‣ Python package

‣ Allows easy interaction with APIs

‣ Helps Data Scientists "get" data from APIs

‣ R equivalent: httr

Requests

http://docs.python-requests.org/en/master/

VISUALIZING THE DATA

‣ Important before and after the modeling process

‣ Many packages created to make visualization easier

‣ Many are good at it but few are great

numpy

‣ Numpy provides fast matrix computation

‣ Most scientific python packages depend on numpy in the backend

‣ API is a bit harder to use for beginners than pandas (IMO) but often faster for large datasets
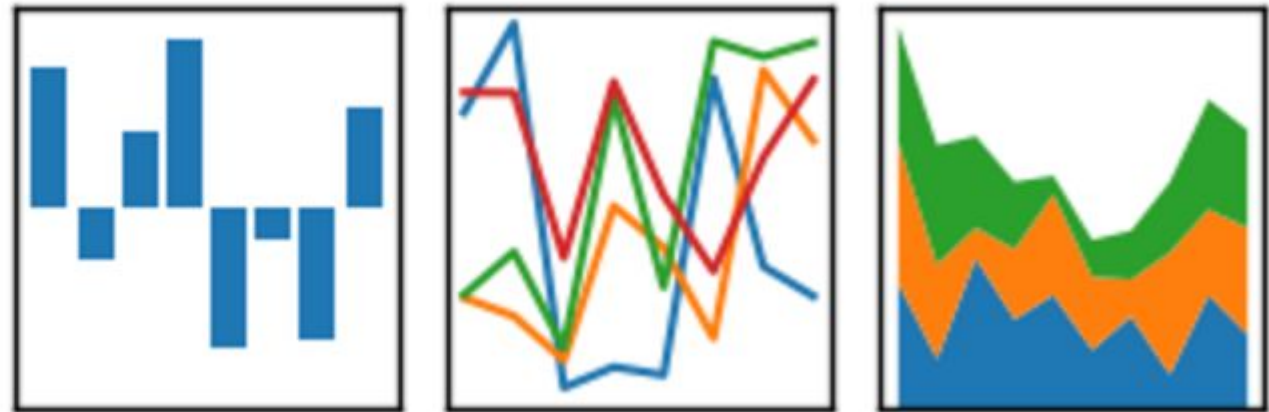
pandas

‣ Depends on numpy

‣ Makes data munging (and some visualization) much easier

‣ Think of as the next step after Excel for manipulating data

‣ "Port" of dataframe from R

matplotlib

‣ Depends on numpy

‣ Makes scripted / programmatic visualization easier

‣ Deeply engrained in the scientific/engineering community

‣ You probably wouldn't use it in non-technical client presentations

‣ R equivalent is ggplot

scikit-learn

‣ Depends on numpy

‣ The most popular destination for machine learning in Python

‣ Contains the "most popular" machine learning algorithms

‣ Standardized way to use machine learning algorithms

‣ Example R equivalent: mlr

FREE MATH AND STATISTICS RESOURCES

The entire math track on **KHAN**ACADEMY

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 **KHAN**ACADEMY

Intro to Computer Sci on U D A C I T Y

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...

Creating a community for my current/former students to discuss data science and computer science

My blog: http://masongallo.github.io/

Collaborating on open source: https://github.com/MasonGallo

Follow me on twitter for data science: https://twitter.com/ohheyitsmason

Many of my students have gone on to jobs in tech, startups, and education!

DATA SCIENCE

Q/A

DATA SCIENCE

Keep in touch!

masonagallo@gmail.com
https://www.linkedin.com/in/masongallo
http://masongallo.github.io/
[@ohheyitsmason](https://twitter.com/ohheyitsmason)