GENERAL ASSEMBLY

"Your move, chief."
2 HOUR
The Tremont Street Subway in Boston's MBTA Subway system is the oldest subway tunnel in North America and the third oldest worldwide.
ESTABLISHED
1898
BEACON ST.
3295

Welcome to General Assembly

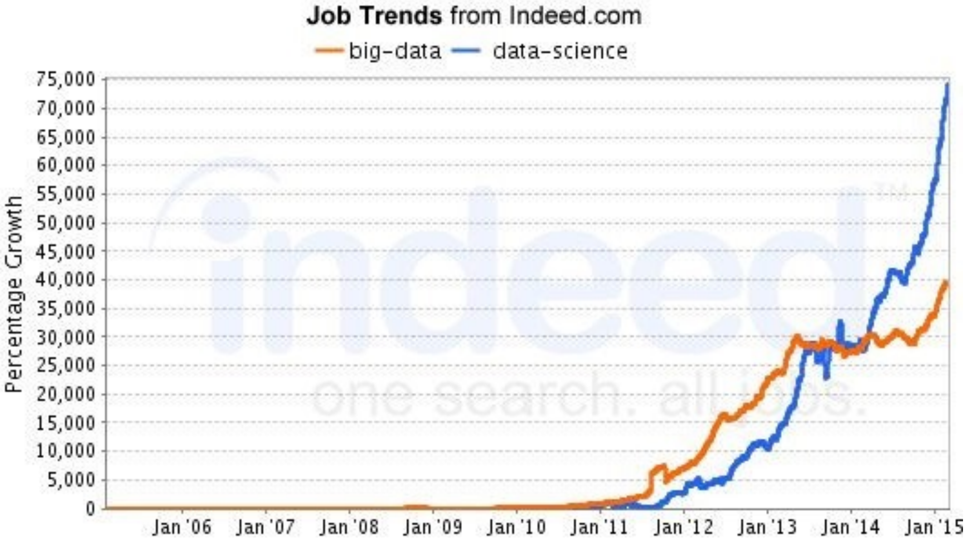
Part-Time Data Science

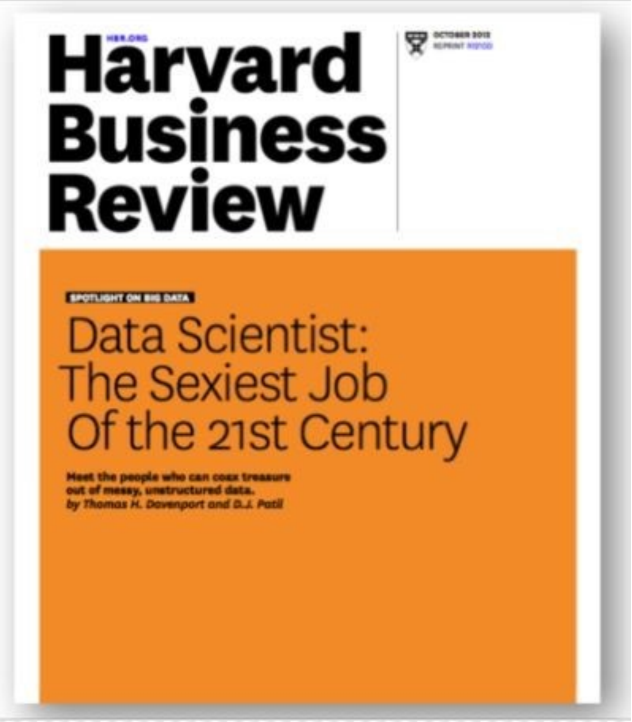
Schedule

7:00 PM: Introductory Slides
8:00 PM: Trello, Github, Anaconda
8:30 PM: Pivot to Python

WHAT IS DATA SCIENCE?

WHATIS DATA SCIENCE?





HBR.ORG

Harvard Business Review

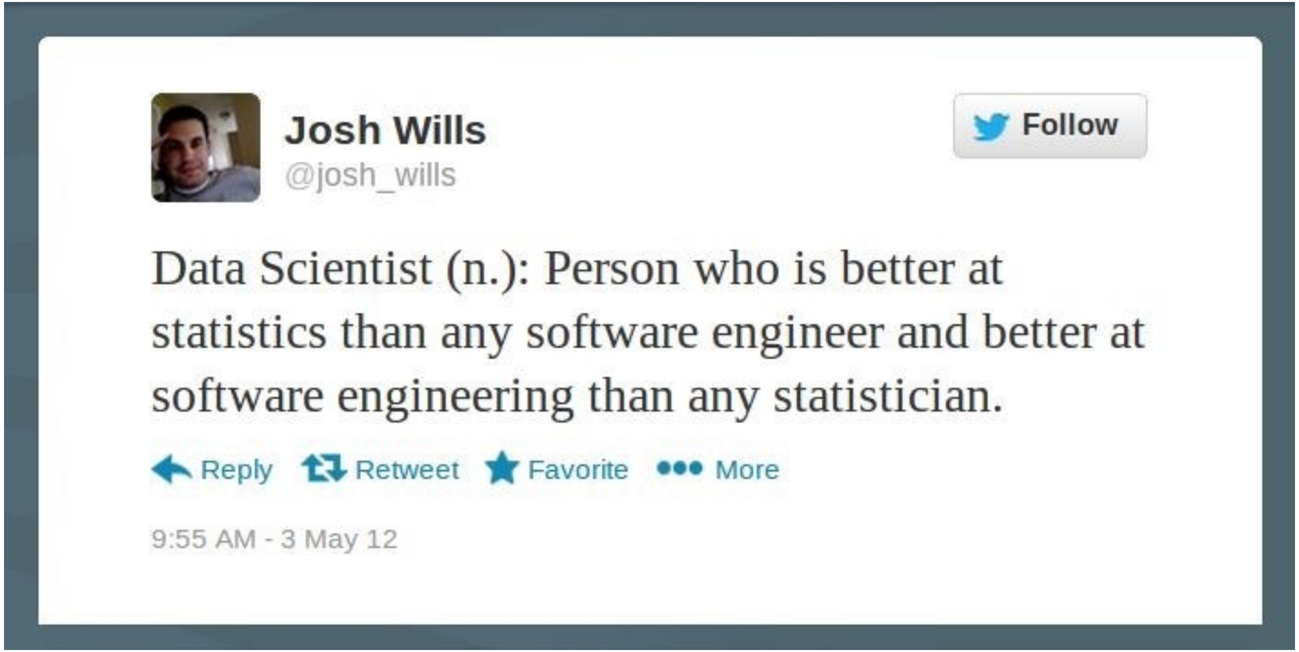
OCTOBER 2012
REPRINT R1210D

SPOTLIGHT ON BIG DATA

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

WHATIS DATA SCIENCE?



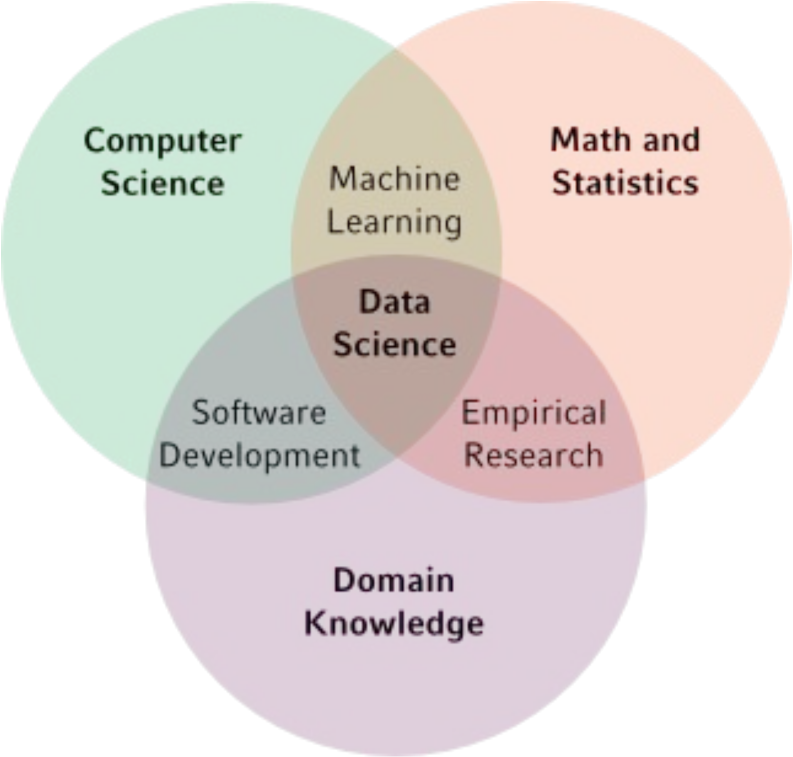
Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

9:55 AM - 3 May 12

WHAT IS DATA SCIENCE?

BY FIELD



WHAT IS DATA SCIENCE?

- **How They’re Using Data Science:**

- Airbnb prioritizes listings in popular areas, making desirable Airbnbs easier for users to find.
- KAID Health uses natural language processing to mine clinical notes, allowing providers to find patients for clinical trials
- UPS optimizes package drop-off and delivery transport using machine learning and AI to predict delivery obstacles (e.g., weather, traffic).

WHAT IS DATA SCIENCE?

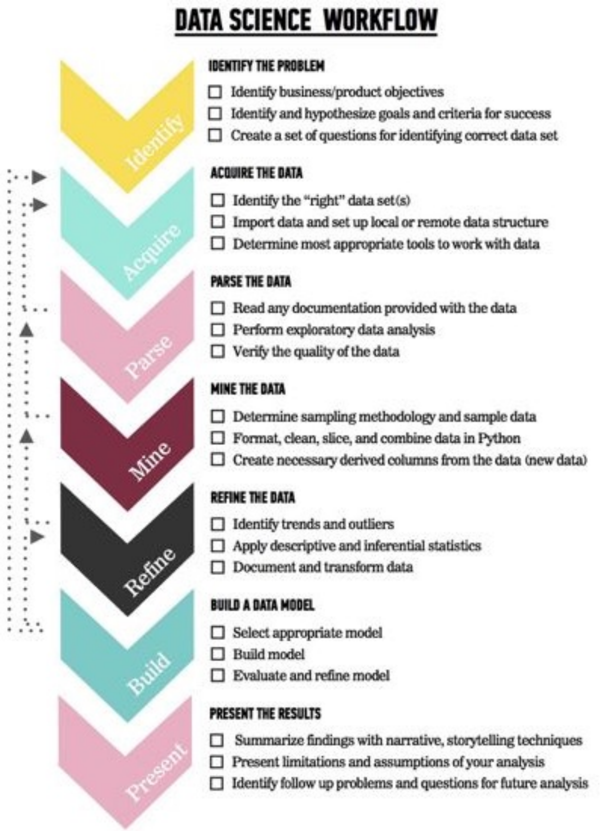
Consider these products and services:

- How do they utilize data science?
- What kinds of data do you think they use?
- How might they leverage data science in other parts of their business?

WHATIS DATA SCIENCE?

DATASCIENCEWORKFLOW

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 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 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 THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of 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)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

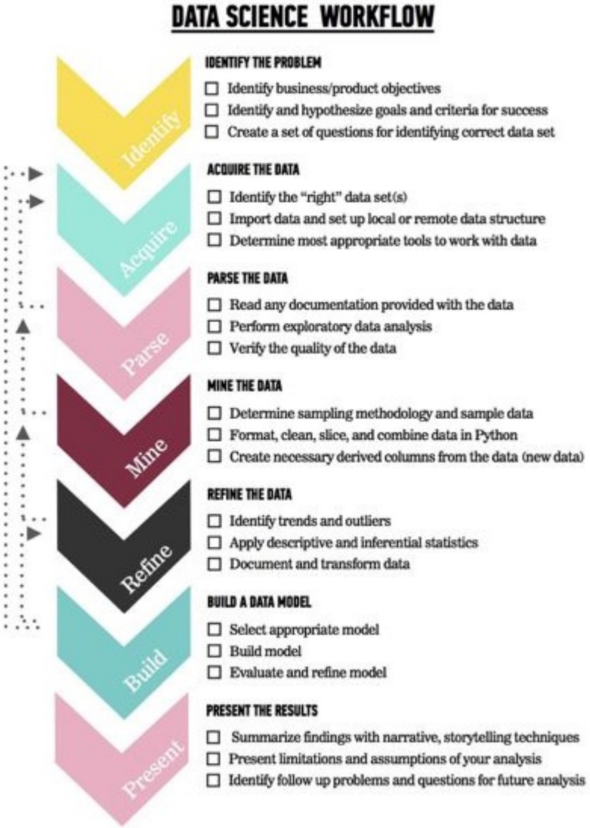
PRESENT THE RESULTS

□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

WHATIS DATA SCIENCE?

IDENTIFY

- **Why are you doing this in the first place?**
- *We believe there is a market for automating detailed medical forecasts for individual claims*
- **Who are the stakeholders?**
- **What data will you need?**
- *Is it available to us?*
- *Is is public or proprietary? Is your work easily duplicated if the former?*
- **How will you define success?**



DATA SCIENCE WORKFLOW

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 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 THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of 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)

REFINE THE DATA

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

BUILD A DATA MODEL

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

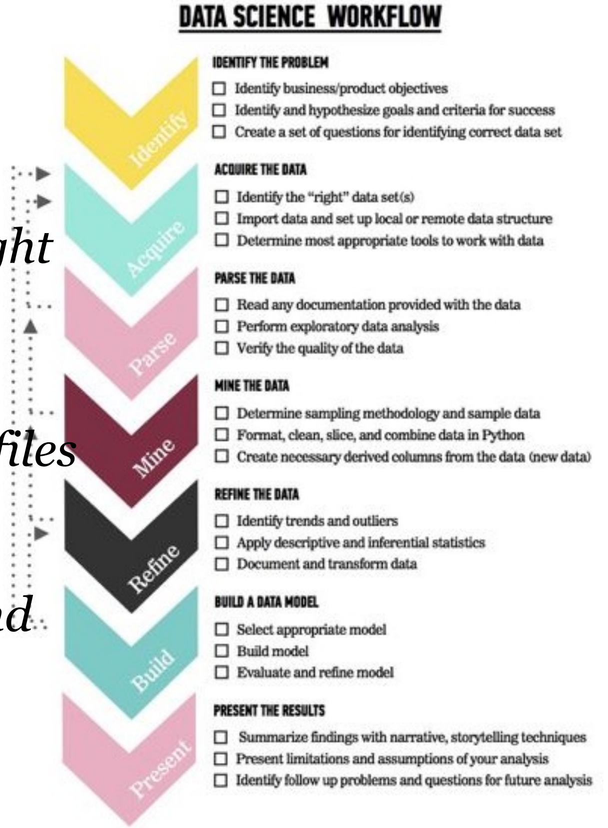
PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

WHATIS DATA SCIENCE?

ACQUIRE

- **Can you supplement your data?**
- *Is there information in the clinical notes that might not appear in bills?*
- **How is it stored?**
- *CBI stores data in relational databases and .csv files*
- **What tools will you need to work with it?**
- *Software (to manipulate data; fit algorithms) and hardware (to handle computations)*



DATA SCIENCE WORKFLOW

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 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 THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of 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)

REFINE THE DATA

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

BUILD A DATA MODEL

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

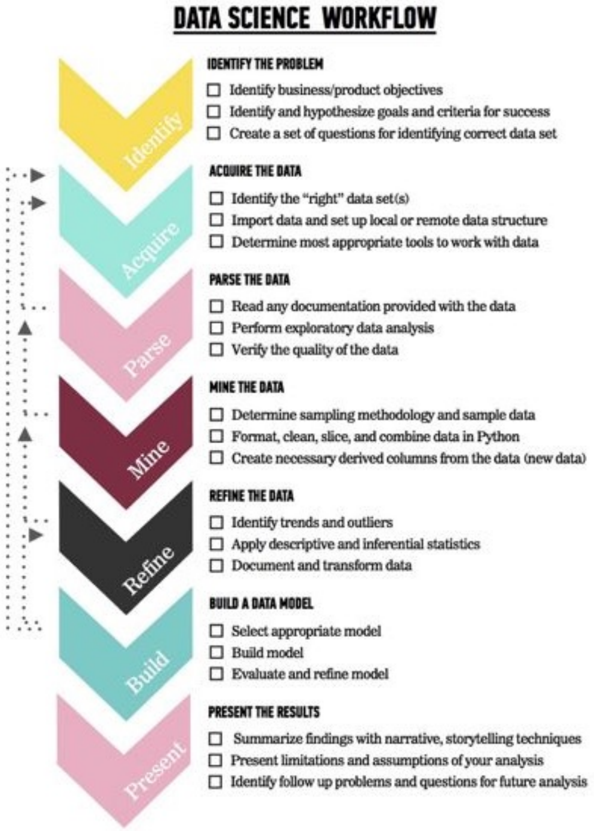
PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

WHATIS DATA SCIENCE?

PARSE

- **How do you get raw data into a format you can work with?**
- *This is the purview of "Data Engineers"*
- **What documentation is available, if any?**
- *Data dictionaries are ideal*
- **How much munging will it require?**
- *Are all your date fields valid dates?*
- *Are some fields mysteriously empty?*



DATA SCIENCE WORKFLOW

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 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 THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of 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)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

PRESENT THE RESULTS

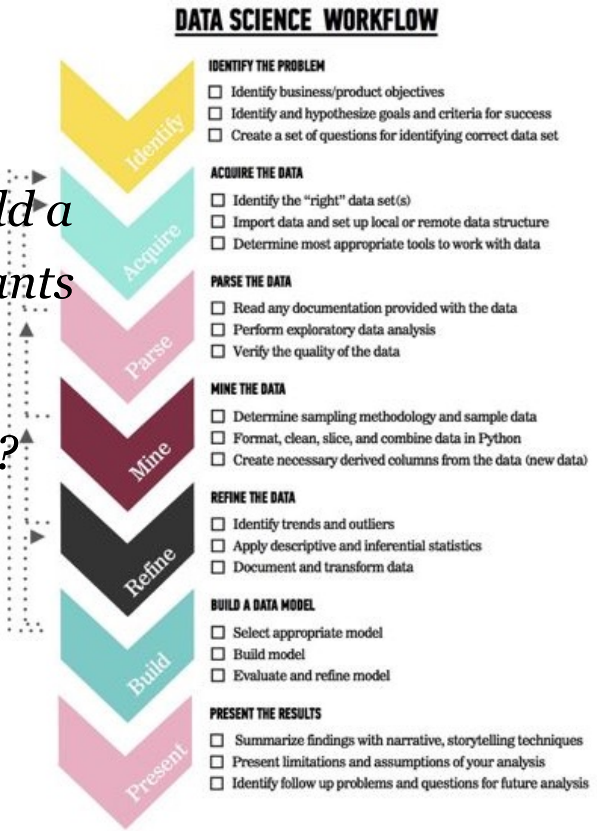
□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

WHATIS DATA SCIENCE?

MINE & REFINE

- **Combining and Transforming the data**
- *Aggregating inpatient and outpatient bills to yield a longitudinal service history for individual claimants*
- **Mining the data to find predictive insights**
- *Example: what's a predictor of shoulder surgery?*

Claimant	Age	Torn Rotator Cuff	Shoulder Sprain	Rotator Cuff Surgery
Jim	28	True	True	True
Pat	40	False	True	False



DATA SCIENCE WORKFLOW

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 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 THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of 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)

REFINE THE DATA

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

BUILD A DATA MODEL

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

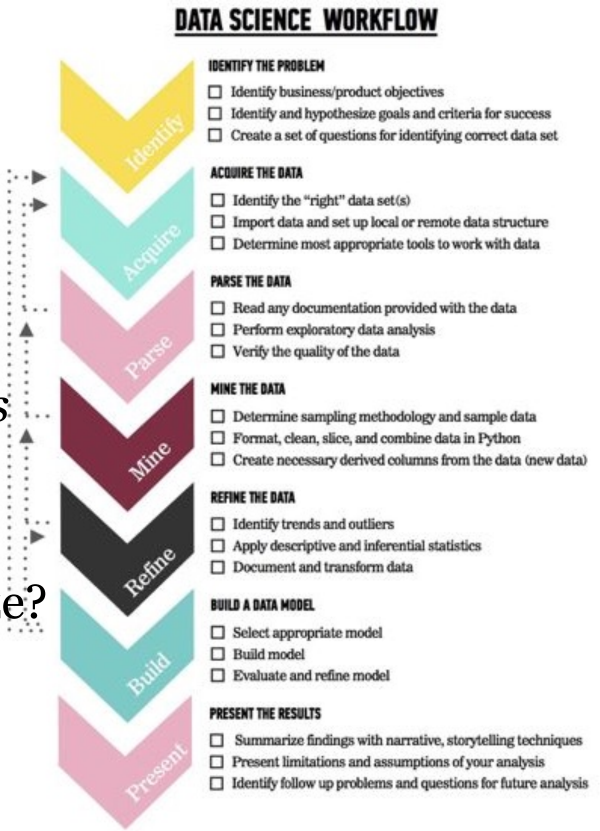
PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

WHATIS DATA SCIENCE?

MODEL

- What model or models are most appropriate for the data and the problem?
- *Is your data linear (each additional square foot yields a higher price) or non-linear (small houses in desirable neighborhoods cost more than large houses in undesirable ones)*
- How can you be sure your model results generalize?
- *We need to evaluate out-of-sample data that our model(s) haven't trained on to understand this*



DATA SCIENCE WORKFLOW

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 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 THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of 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)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

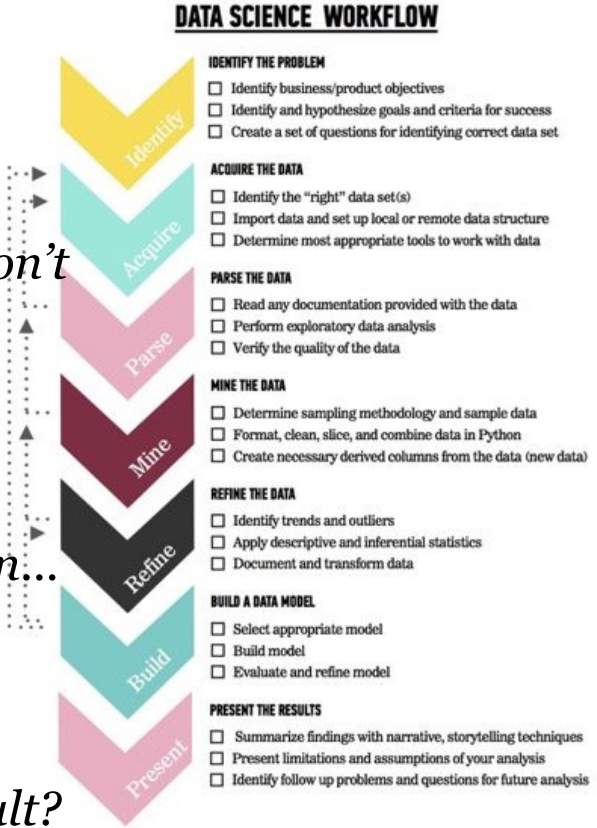
PRESENT THE RESULTS

□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

WHATIS DATA SCIENCE?

PRESENT

- **What narrative do I want to tell?**
- *Why does the model predict claimant x will or won't get surgery?*
- **What inherent limitations should be disclosed?**
- *Did you collect the data your models are based on...*
- *Will your model need to be retrained?*
- **Were the criteria for success met?**
- *With more time, how would we improve our result?*



DATA SCIENCE WORKFLOW

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 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 THE DATA

- □ Read any documentation provided with the data
- □ Perform exploratory data analysis
- □ Verify the quality of 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)

REFINE THE DATA

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

BUILD A DATA MODEL

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

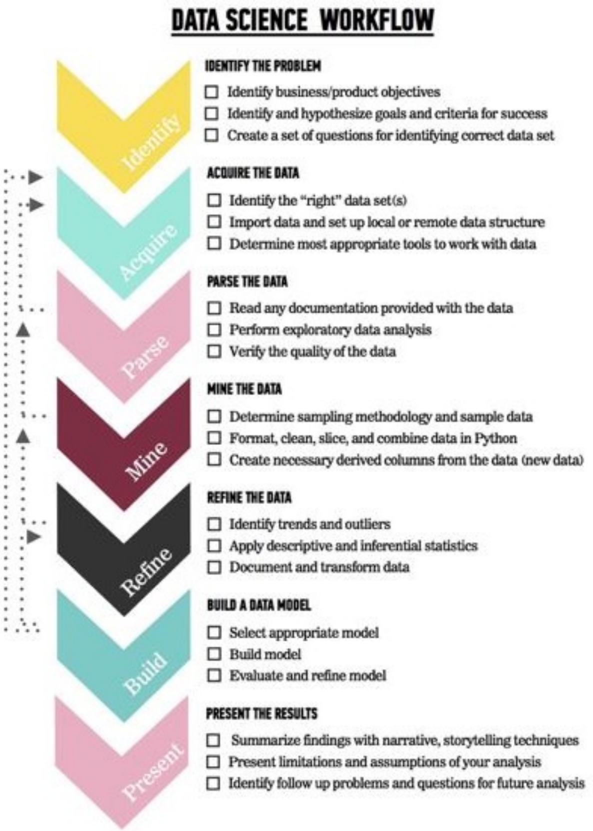
PRESENT THE RESULTS

- □ Summarize findings with narrative, storytelling techniques
- □ Present limitations and assumptions of your analysis
- □ Identify follow up problems and questions for future analysis

WHATIS DATA SCIENCE?

(DEPLOY)

- **How do we run the model in production?**
- *I pickle the model*
- *When a user calls our API, the model is loaded*
- *But we have to provide the same features to our model as I did fitting it, so the data submitted on a web form has to go through a series of transformations first. (For example, taking a diagnosis and passing a rate into the model.)*



DATA SCIENCE WORKFLOW

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 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 THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of 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)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

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

MODELING

MODELING

Video #1 - Margaritas

- https://www.youtube.com/watch?v=t_3fnVqNOUc

Video #2 - Tennis

- https://www.youtube.com/watch?v=eKD5gxPPeY0

Comparing these examples

- How did we cluster the margaritas together?
- How did we make a prediction about tennis?
- What seems different about the approaches?

MODELING

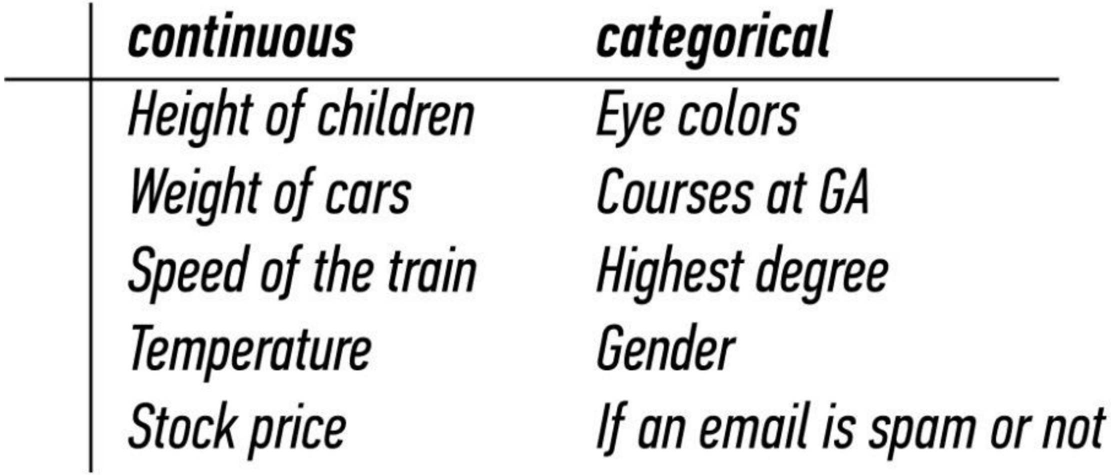
Features & Target Variables

- x: features (inputs)
- y: target variable (output; can be numeric or binary)
- Models: values in, value(s) out

$$y = f(x_1, x_2, x_3\ldots)$$

MODELING

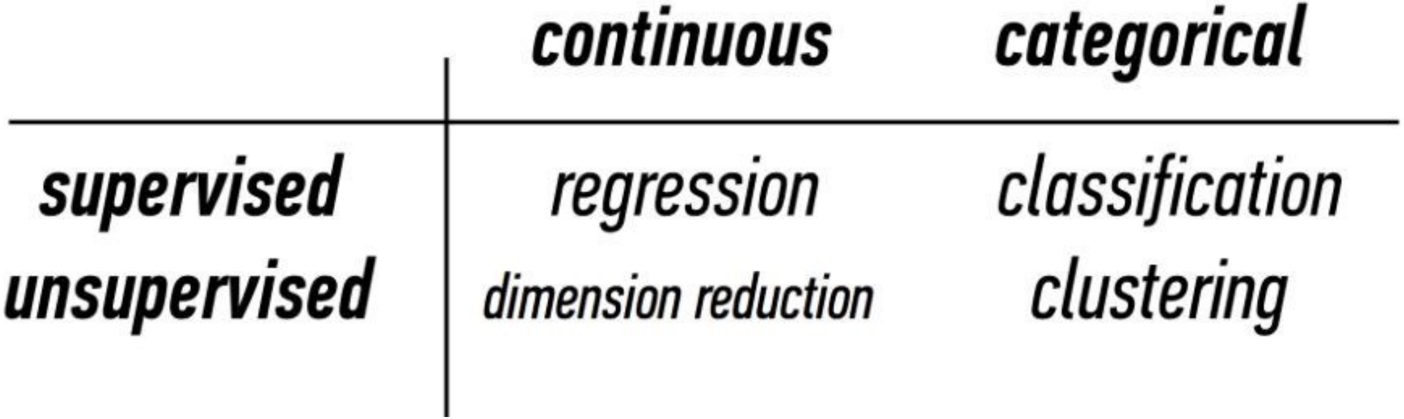
TARGET VARIABLE FLAVORS



continuous	categorical
Height of children	Eye colors
Weight of cars	Courses at GA
Speed of the train	Highest degree
Temperature	Gender
Stock price	If an email is spam or not

MODELING

MACHINE LEARNING PROBLEMS

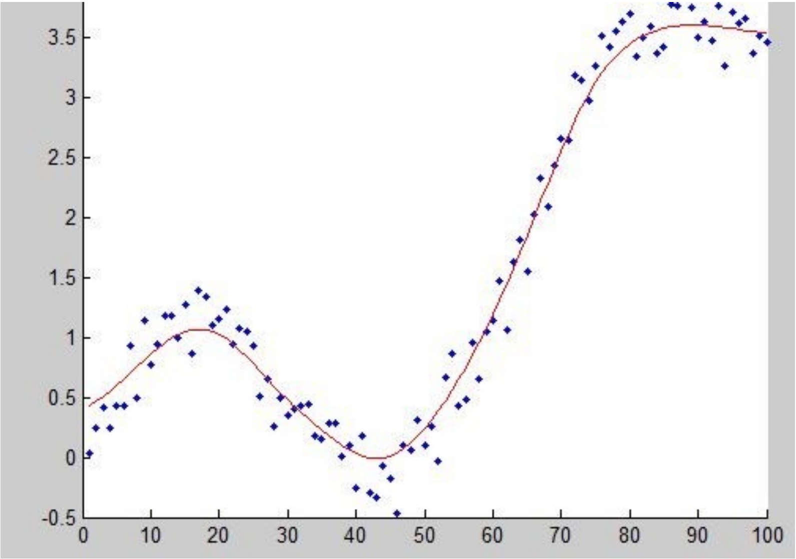


	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

MODELING

REGRESSION
(CONTINUOUS, SUPERVISED)

- Build a model to predict a continuous value that best fits data
- Minimize error without overfitting
- *Example: Linear Regression* http://setosa.io/ev/ordinary-least-squares-regression/



MODELING

CLASSIFICATION
(CATEGORICAL, SUPERVISED)

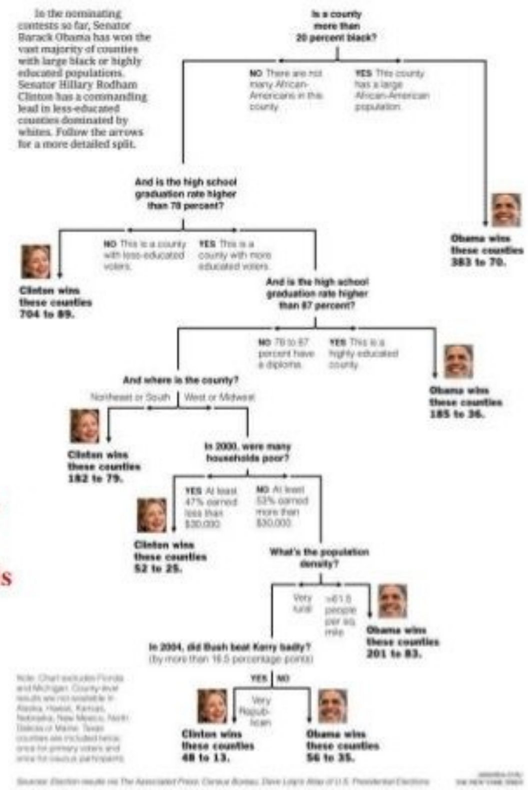
- Map features to <u>categorical</u> <u>target classes</u>.

Can we learn how counties vote?

New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

Decision Tree: The Obama-Clinton Divide


Is a county more than 20 percent black?
NO There are not many African-Americans in this county
YES This county has a large African-American population
Obama wins these counties 383 to 70.
And is the high school graduation rate higher than 78 percent?
NO This is a county with less-educated voters.
YES This is a county with more educated voters.
Clinton wins these counties 704 to 89.
And is the high school graduation rate higher than 87 percent?
NO 78 to 87 percent have a diploma
YES This is a highly educated county
Obama wins these counties 185 to 36.
And where is the county?
Northeast or South
West or Midwest
Clinton wins these counties 182 to 79.
In 2000, were many households poor?
YES At least 47% earned less than $30,000
NO At least 53% earned more than $30,000
Clinton wins these counties 52 to 25.
What's the population density?
Very rural
>61.5 people per sq. mile
Obama wins these counties 201 to 83.
In 2004, did Bush beat Kerry badly?
(by more than 16.5 percentage points)
YES
NO
Very Republican
Clinton wins these counties 48 to 13.
Obama wins these counties 56 to 35.

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.

Note: Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice: once for primary voters and once for caucus participants.

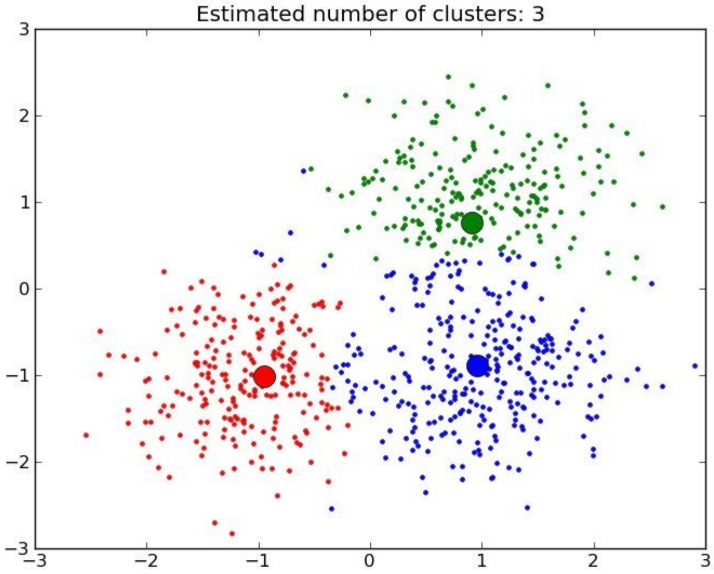
Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

THE NEW YORK TIMES

MODELING

CLUSTERING
(CATEGORIAL, UNSUPERVISED)

- Purpose is representation – anything that helps you better understand the data
- Finding common threads that a human couldn’t see (imagine 10 or 100 or 1000 inputs)
- *Example: K-Means* https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



MODELING

Regression: How Much Physical Therapy?

Claimant	Split	Severe Underlying Injury	Surgery After Injury	PT Units
Jim	Train	Yes	Yes	13
Pat	Train	No	No	4
Dimitri	Test	Yes	No	**?**

MODELING

Classification: Likelihood of Major Surgery

Claimant	Split	Rate That Diagnosis Receives Surgery	Proximity to Accident (years)	Surgical Likelihood
Deb	Training	.5	<1	100%
Jack	Training	.2	3	0%
Dimitri	Test	.15	<1	**?**

CASESTUDY

Item	Orders Since Customer Included Item	Customer Order Rate	Order Probability
Bananas	1	90%	?
Baking Soda	6	5%	?

With (a) partner(s), map out where this fits in the data science workflow – high level:
what phase is it, what comes before it, what comes after it

HOWEVER

There is a lot of variation in what data scientists do. Most of my experience is… *not* modeling.

- I write rules to e.g. deidentify documents
- I label data
- I’ve built dashboards
- I’ve handled ETL
- I’ve migrated data
- I’ve contributed to research

GO FURTHER

Understanding Exploratory Data Analysis (EDA)

https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Annotated Titanic Workflow

https://www.kaggle.com/headsortails/pytanic

That’s a wrap!