Your Data Science in Practice Course

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC

Department of Cognitive Science
rdprobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures :http://casimpkinsjr.radiantdolphinpress.com/pages/cogs108_ss1_23/index.html

Plan for today

- Announcements
- Assignment status
- Project status
- Review of the course
- The future of data science
- Gathering the course, final parting thoughts

Remaining assignments and project parts

- A4, D7, Q4 due Sunday 8/9/23 11:59pm
  - 1 Lab/7 will be extra credit, you don't have to choose which, just do as many as possible, and total possible points will be >max by 1 lab worth
  - OR if pressed for time, you can just not do D7
- Final Project due Fri, 8/4/2023 (11:59 PM)
  - Report (GitHub)
  - Video (shared via github, link, youtube, etc such that we can view it)
  - Team Evaluation Survey:Link will be on canvas (link also on Canvas; required)
- Post COGS 108 Survey: link to be posted (link also on Canvas; *optional* for EC)
- Evaluations - reminder they are different this quarter

What you all have done

Strap in!

COGS 108: What we've learned

Week	Topic(s)
1	Data Science & Version Control, Datahub, Jupyter, python I
1	Data Intuition & Wrangling
2	Data Ethics & Questions
2	Data Visualization & Data Analysis
3	Inference
3	Text Analysis
4	Machine Learning
4	Nonparametric Analysis
5	Geospatial Analysis
5	Data Science Communication & Jobs

We defined data science

Data scientist is actually MANY jobs

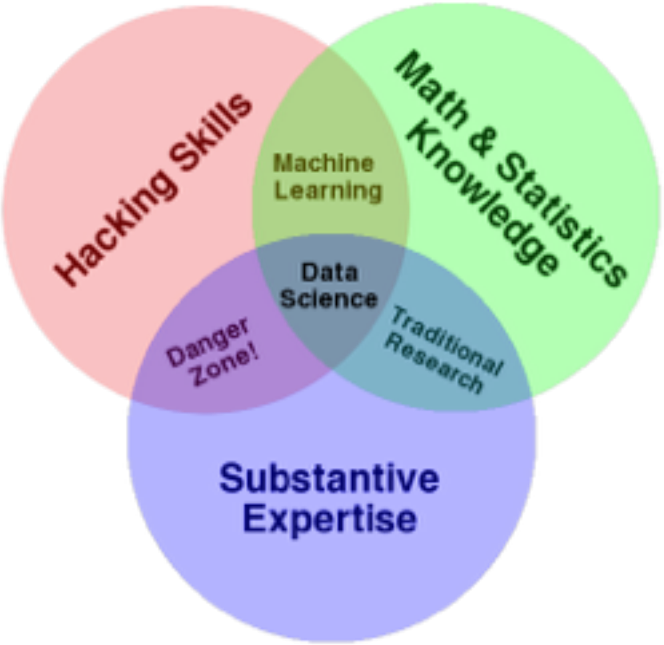
https://hbr.org/2018/11/the-kinds-of-data-scientist

A final piece of advice for those hiring data scientists: Look for people who are in love with solving problems, not with specific solutions or methods, and for people who are incredibly collaborative. No matter what kind of data scientist you are hiring, to be successful they need to be able to work alongside a vast variety of other job functions — from engineers to product managers to marketers to executive teams. Finally, look for people who have high integrity. As a society, we have a social responsibility to use data for good, and with respect. Data scientists hold the responsibility for data stewardship inside and outside the organization in which they work.

Data science for humans

Data science for computers

What is data science?



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

We discussed version control motivation and technique

This sucks

main_simple_bak9-pretty-good.c	Aug 1, 2008, 1:01 AM	33 KB	C Source
main_simple_bak9-pretty-good.o	Aug 1, 2008, 1:00 AM	303 KB	object code
main_simple_bak9-pretty-goodv2.c	Aug 2, 2008, 1:16 AM	33 KB	C Source
main_simple_bak10.c	Sep 28, 2008, 1:16 PM	33 KB	C Source
main_simple_bak11-workingUART_correctspeed.c	Aug 30, 2008, 2:49 AM	27 KB	C Source
main_simple_bak11-workingUART_correctspeed.o	Aug 2, 2008, 1:17 AM	303 KB	object code
main_simple_bak12_willspin.c	Aug 2, 2008, 1:30 AM	28 KB	C Source
main_simple_bak12_willspin.o	Aug 2, 2008, 2:35 AM	301 KB	object code
main_simple_bak13-worksA-D-nonoise-spins.c	Aug 7, 2008, 12:57 PM	26 KB	C Source
main_simple_bak14-widersinefunctionsworkingrotation.c	Aug 8, 2008, 5:02 PM	26 KB	C Source
main_simple_bak15-spins-stillneedsquadrantfixed.c	Aug 15, 2008, 7:32 PM	30 KB	C Source
main_simple_bak16-15backup-spins-needs-improvement.c	Oct 15, 2008, 8:54 PM	31 KB	C Source
main_simple_bak17-smoother-stillnostandingstart.c	Aug 16, 2008, 6:50 PM	30 KB	C Source
main_simple_bak17-smoother-stillnostandingstart.o	Aug 18, 2008, 9:41 PM	305 KB	object code
main_simple_bak18-notgood.c	Aug 18, 2008, 9:42 PM	31 KB	C Source
main_simple_bak20SIMPLE-DCnotbrushless.c	Sep 17, 2009, 11:02 PM	27 KB	C Source
main_simple_bak20WORKS_PWM_COMMAND_CONTROL.c	Aug 19, 2008, 12:54 AM	29 KB	C Source
main_simple_timer_intrpt_bak.c	Aug 12, 2008, 12:16 AM	13 KB	C Source
main_simple_timer_intrpt_bak2.c	Aug 12, 2008, 2:00 PM	13 KB	C Source
main_simple_timer_intrpt_bak3.c	Aug 18, 2008, 12:14 AM	13 KB	C Source
main_simple_timer_intrpt.c	Aug 18, 2008, 12:17 AM	13 KB	C Source
main_simple_workingHWPWM.c	Aug 18, 2008, 7:19 PM	15 KB	C Source
main_simple.c	Sep 17, 2009, 11:02 PM	29 KB	C Source

Version Control

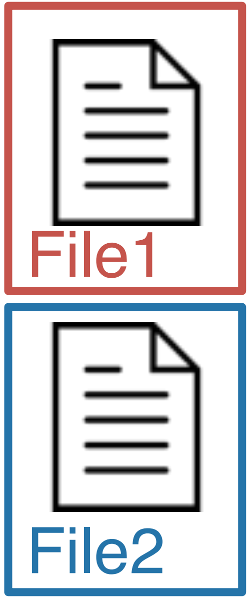
- Enables multiple people to simultaneously work on a single project.
- Each person edits their own copy of the files and chooses when to share those changes with the rest of the team.
- Thus, temporary or partial edits by one person do not interfere with another person's work

What is version control?

A way to manage the evolution of a set of files

What is version control?

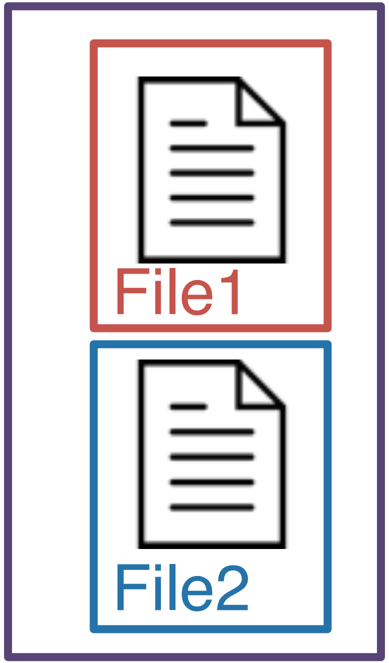
A way to manage the evolution of a set of files



When using a version control system, you have **one copy of each file** and the *version control system tracks the changes* that have occurred over time

What is version control?

A way to manage the evolution of a set of files



The set of files is referred to as a **repository (repo)**

git & GitHub

"Global Information Tracker"

the version control system

~ Track Changes from Microsoft Word….on steroids

GitHub (or Bitbucket or GitLab) is the home **where your git-based projects live** on the Internet.

~ Dropbox….but way better

What version control looks like

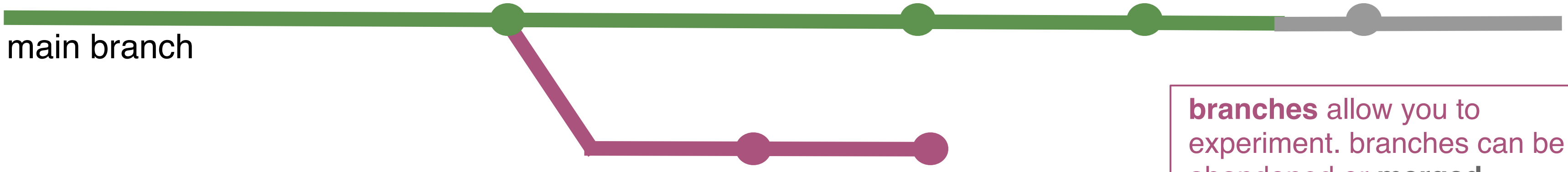
```
$ git clone https://www.github.com/username/repo.git
$ git pull
$ git add -A
$ git commit -m “informative commit message”
$ git push
```

Terminal
git

GitHub

377dfcd00dd057542b112cf13be6cf1380b292
ad

commits allow you to time travel
because each commit is assigned
a unique **hash**



main branch

try-something-cool

branches allow you to
experiment. branches can be
abandoned or **merged**

someone
else's
repo

fork

your
GitHub

You can work on others'
repos by first **forking** their
repository onto your GitHub

Pull requests allow you to make
specific edits to others' repos

Issues allow you to make general
suggestions to your/others' repos

One more git recap...

We learned about

- repos
- git, github
- clone
- merge
- branch
- push
- pull
- fork
- commits
- staging
- issues
- merge conflicts

We discussed data structures, data intuition, tidy data

Data Structures Review

Structured data

- can be stored in database SQL
- tables with rows and columns
- requires a relational key
- 5-10% of all data

Semi-structured data

- doesn’t reside in a relational database
- has organizational properties (easier to analyze)
- CSV, XML, JSON

Unstructured

- non-tabular data
- 80% of the world’s data
- images, text, audio, videos

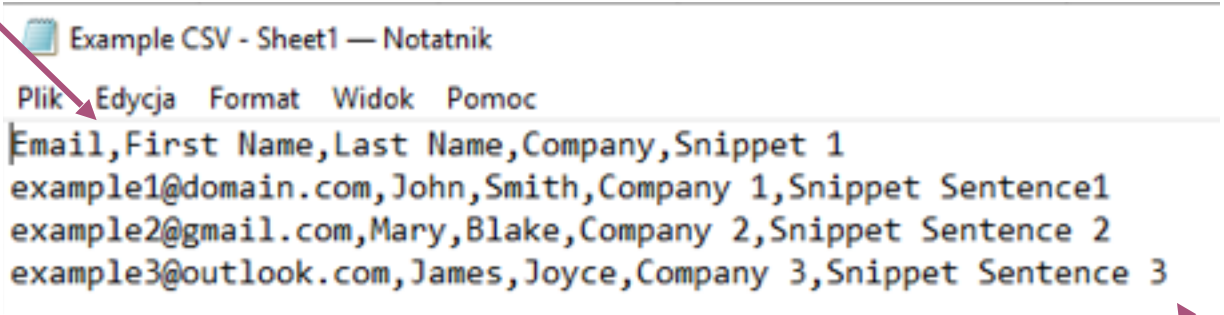
(Semi-)Structured Data

Data that is stored in such a way that it is easy to search and work with. These data are stored in a particular format that adheres to organization principles imposed by the file format. These are the data structures data scientists work with most often.

CSVs

Has the extension ".csv"

Each column separated by a comma


Example CSV - Sheet1 — Notatnik

Plik Edycja Format Widok Pomoc

```
Email,First Name,Last Name,Company,Snippet 1
example1@domain.com,John,Smith,Company 1,Snippet Sentence1
example2@gmail.com,Mary,Blake,Company 2,Snippet Sentence 2
example3@outlook.com,James,Joyce,Company 3,Snippet Sentence 3
```

Each row is separated by a new line

JSON: key-value pairs

nested/hierarchical data

{"Name": "Isabela"}

key

value

JSON

These are all nested within `attributes`

```
"attributes": {
  "Take-out": true,
  "Wi-Fi": "free",
  "Drive-Thru": true,
  "Good For": {
    "dessert": false,
    "latenight": false,
    "lunch": false,
    "dinner": false,
    "breakfast": false,
    "brunch": false
  },
```

These are all nested within `"Good For"`

https://blog.exploratory.io/working-with-json-data-in-very-simple-way-ad7ebcc0bb89

```xml
<?xml version="1.0" encoding="UTF-8"?>
<customers>
    <customer>
        <customer_id>1</customer_id>
        <first_name>John</first_name>
        <last_name>Doe</last_name>
        <email>john.doe@example.com</email>
    </customer>
    <customer>
        <customer_id>2</customer_id>
        <first_name>Sam</first_name>
        <last_name>Smith</last_name>
        <email>sam.smith@example.com</email>
    </customer>
    <customer>
        <customer_id>3</customer_id>
        <first_name>Jane</first_name>
        <last_name>Doe</last_name>
        <email>jane.doe@example.com</email>
    </customer>
</customers>
```

XML

Relational Databases: A set of interdependent tables

1. Efficient Data Storage
2. Avoid Ambiguity
3. Increase Data Privacy

Results | Messages

	EmployeeId	FirstName	LastName	DepartmentName
1	1	Ken	Sanchez	Executive
2	2	Terri	Duffy	Engineering
3	3	Roberto	Tamburello	Engineering
4	4	Rob	Walters	Engineering
5	5	Gail	Erickson	Engineering
6	6	Jossef	Goldberg	Engineering
7	7	Dylan	Miller	Support
8	8	Diane	Margheim	Support
9	9	Gigi	Matthew	Support
10	10	Michael	Raheem	Support

relational database

Relational database

restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	**JJ29JJ**	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

health inspections

id	inspection_date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

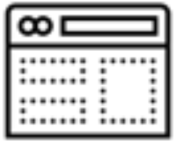
Two different restaurants with the same name will have different unique identifiers

Unstructured Data

Some datasets record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.

Unstructured Data Types

Text files and documents



Websites and applications

Sensor data

Image files

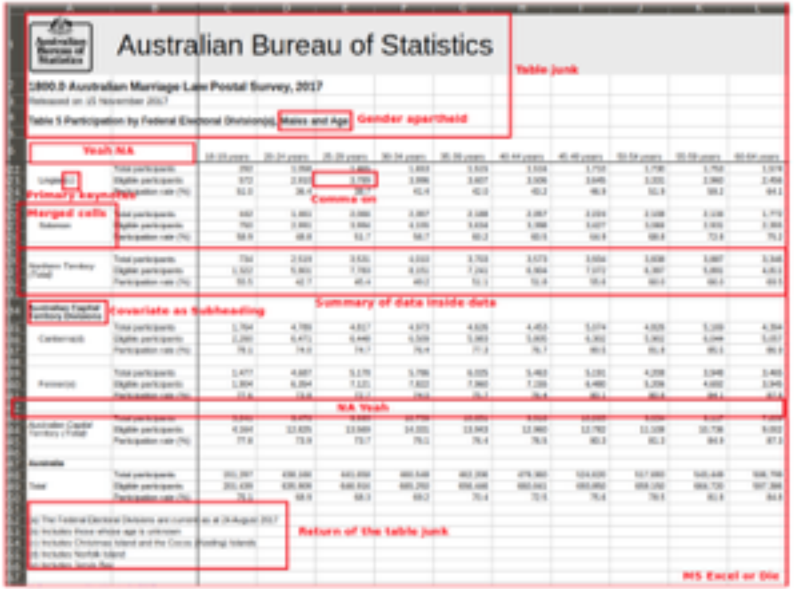
Audio files

Video files

Email data

Social media data

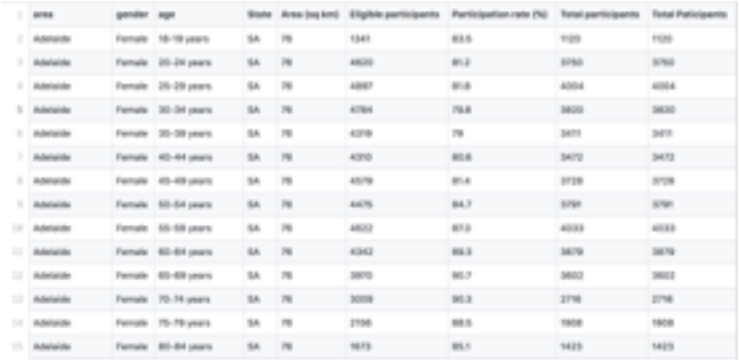
untidy data



data

wrangling

tidy data



https://medium.com/@miles.mcbain/tidying-the-australian-same-sex-marriage-postal-survey-data-with-r-5d35cea07962

Tidy data == rectangular data

A

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints* 5:e3183v1 https://doi.org/10.7287/peerj.preprints.3183v1

Data Wrangling defined

- The process of restructuring a dataset from whatever form it is initially in to a computationally usable form suitable for data science

Tidy Data Benefits

1. consistent data structure
2. foster tool development
3. require only a small set of tools to be learned
4. allow for datasets to be combined

What is data cleaning?

- Fixing/removing incorrect, corrupted, incorrectly formatted, duplicate, incomplete, data within a dataset
- Many issues combining data sources and types, researcher styles, standards, recording errors, etc

Data Wrangling vs. Data Cleaning

- Data wrangling focuses on transforming the data from a ‘raw’ format into a format suitable for computational use
- Data cleaning focuses on, as discussed, fixing/removing incorrect, corrupted, incorrectly formatted, duplicate, incomplete, data within a dataset

https://www.youtube.com/watch?v=0YzvupOX8Is

Has humanity produced enough paint to cover the entire land area of the Earth?

—Josh (Bolton, MA)

Data Intuition

1. Think about your question and your expectations
2. Do some Fermi calculations (back of the envelope calculations)
3. Write code & look at outputs <- think about those outputs
4. Use your gut instinct / background knowledge to guide you
5. Review code & fix bugs
6. Create test cases - “Sanity checks”

Always consider ethics.

<u>ETHICS</u>

“Moral principles that govern a person's behavior or the conducting of an activity.”

Source: Oxford Dictionary

On INTENT and OBJECTIVITY

- Intent is not required for harmful practices to occur
- Data, algorithms and analysis are not objective.
    - They are created and executed by people, who have biases
    - They use data, which have biases
- Data Science is powerful
- Bias & discrimination driven by data & algorithms can give new scale to pre-existing inequities, and create new inequalities that never existed

NINE THINGS TO CONSIDER TO NOT RUIN PEOPLE'S LIVES WITH DATA SCIENCE

NINE THINGS TO CONSIDER TO NOT RUIN PEOPLE'S LIVES WITH DATA SCIENCE

1. THE QUESTION
2. THE IMPLICATIONS
3. THE DATA
4. INFORMED CONSENT
5. PRIVACY
6. EVALUATION
7. ANALYSIS
8. TRANSPARENCY & APPEAL
9. CONTINUOUS MONITORING

Integrity

- The quality of being honest and having strong moral principles, moral uprightness.
- The state of being whole and undivided

Integrity

- Integrity is very important
- It can help you make decisions when life gets murky
- Maintain your integrity
- It is difficult to get back once lost (but possible)
- One particular position is less important than your integrity

Ethical Data Science

- Data Science pursued in a manner that
  - Minimizes bias, discrimination and exclusion
  - Respects privacy and consent
  - Minimizes and avoids undue harm now and in the future

What is a program?

- Generally a **program** is a **set of instructions** the programmer defines for a device or entity (usually a computer but not always) to follow
- Regarding computers-> programmer writes a set of instructions (“program”) that tells the computer to perform a set of operations
- When the program is executed, the instructions are carried out
    - How does this work (big picture)?
        - Relates to the speed discussion we are about to get into…

Programming languages

- **Low level** machine language (binary/hex) provides instructions for the processor to execute
- **Mid-level** language is called ‘assembly’ language
- **High-level** languages such as C, C++, Fortran, BASIC, etc
- **Very high-level** languages (‘scripting’ languages) such as Python, MATLAB

Terminal and command line review

- **pwd**
- **cd**
- **cd ..**
- **/, ~**
- **file structure, remote and local**
- **executing commands, options**
- **git over command line**

What is python?

- A high-level (sometimes called ‘very high level) programming language (scripting/interpreted)
- Emphasizes readability
- Highly extensible via ‘modules’
- First released in 1991, written by Guido van Rossum

Guido van Rossum

source: https://en.wikipedia.org/wiki/Python_(programming_language)#/media/File:Guido_van_Rossum_OSCON_2006_cropped.png

Python's extensibility

- The extensible core of python is where the true power lies
- Python is great, but without expansion it is not useful for scientific computing - not originally designed for numerical computing
    - Lacks matrix and linear algebra operations
    - No scientific visualization in 2d and 3d
    - Slow, memory intensive

Modules to the rescue!

- You will learn and gain experience with:
  - ***NumPy***
  - ***Pandas***
  - ***Matplotlib***
  - ***Seaborn***
  - ***SciKitLearn***
- And learn how to acquire new module skills as needed

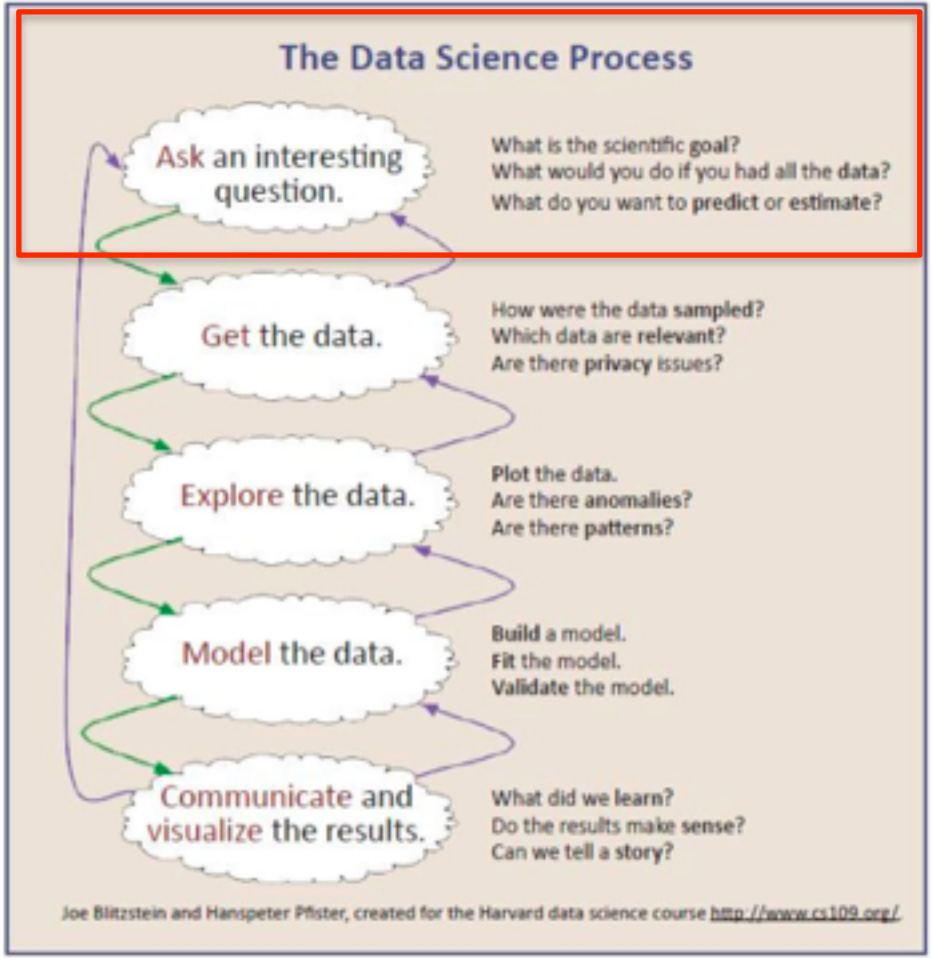
Why Jupyter Notebooks

- Mixed media is excellent for data exploration and communication
- Don’t have to write a separate program from your notes, results, etc
- Easy to experiment in nonlinear and compartmentalized ways

Formulating Data Science Questions

When you and your group sit down to figure out what you're going to do for your final project in this class, you'll have to formulate a strong question - one that is

- ***Specific,***
- ***Can be answered with data,***
- ***Makes clear what exactly is being measured*.**



adapted from Chris Keown

Hypothesis testing

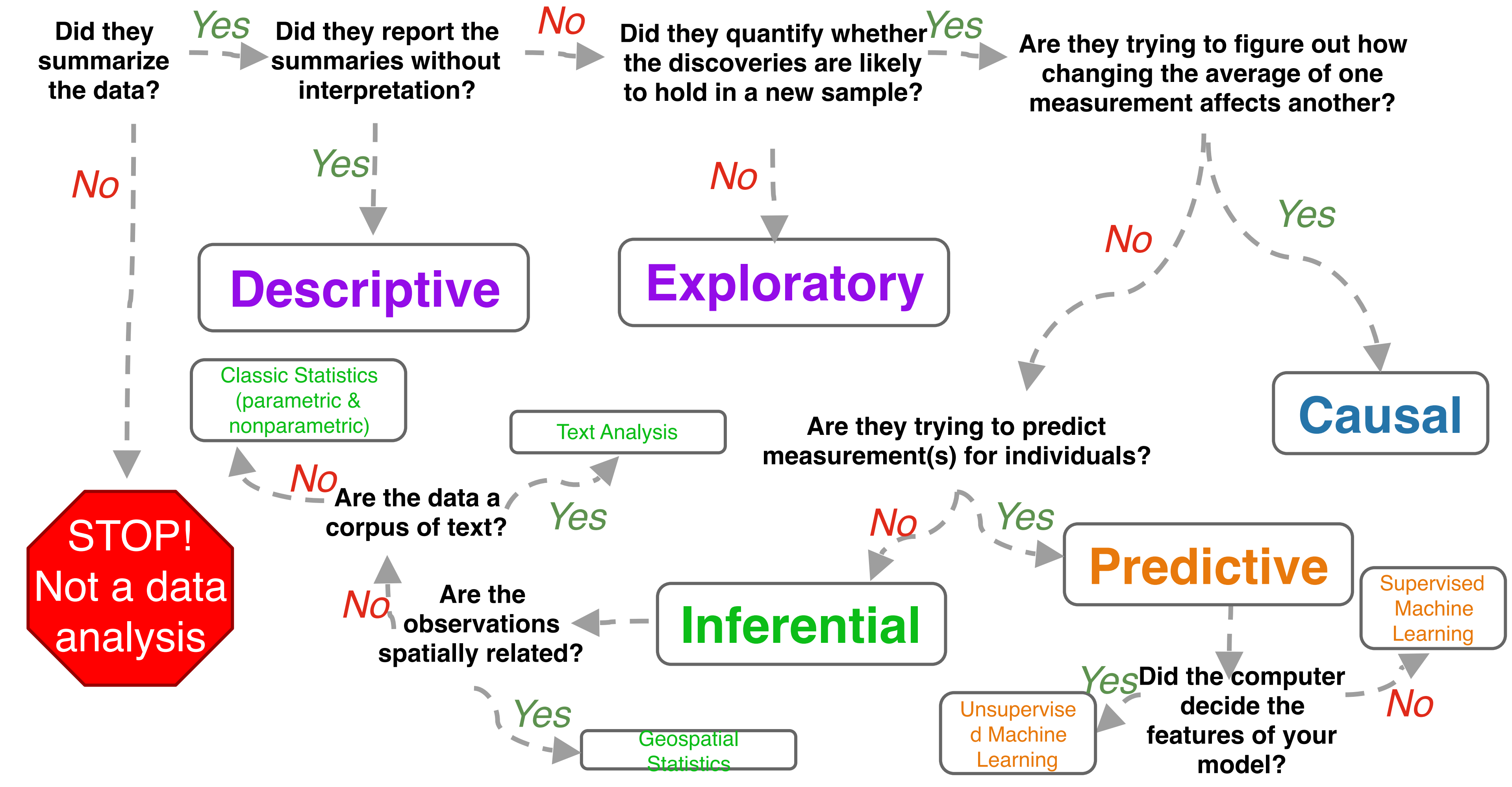
- *Cannot prove hypothesis*
- *Can only reject or fail to reject null hypothesis*
- *Why?*
  - There is always the possibility that there is an underlying variable, effect correlation, connection, direction of connection etc. that might be really affecting things causally which we are not modeling
  - Un-modeled dynamics

A hypothesis should be

- Narrow
- Very specific
- ***Not*** include a conclusion or interpretation
- Consist of a research and null hypothesis
- Remember we are trying to reject or fail to reject the null, which basically says we either
    - ***'didn't find anything' or***
    - ***'failed to not find anything'***

Hypothesis : Simplicity, narrowness

- KISS principle
- Boiled down to the essence of the relationship you are testing
  - **Null** is the thing being tested, **Alternative** is everything else
- Research/Alternative and Null are opposites
  - H_0 - Null Hypothesis
  - H_a or H_1 - Research/Alternative Hypothesis



Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
No
Yes
Yes
No
No
Yes
Descriptive
Exploratory
Causal
Classic Statistics (parametric & nonparametric)
Text Analysis
Are they trying to predict measurement(s) for individuals?
No
Are the data a corpus of text?
Yes
No
Yes
STOP! Not a data analysis
Predictive
Supervised Machine Learning
No
Are the observations spatially related?
Inferential
Yes
Did the computer decide the features of your model?
No
Unsupervised Machine Learning
Yes
Geospatial Statistics

Summary: Analytical Approaches

1. Descriptive (and Exploratory) Data Analysis are the first step(s)
2. Inference establishes relationships
   a. Classic Statistics
   b. Geospatial Analysis
   c. Text Analysis
3. Machine Learning is for prediction
   a. Supervised
   b. Unsupervised
4. Experiments best way to establish the likelihood of causality
   a. Remember you ***cannot*** establish causality with computational methods only correlations along with statistical beliefs
   b. More you are establishing if they are NOT related or 'NOT NOT' related

Descriptive: The goal of descriptive analysis is to understand the components of a data set, describe what they are, and explain that description to others who might want to understand the data.

Statistical Data Analysis

- There are various definitions
- *"The science that deals with the **collection, classification, analysis, and interpretation of numerical facts or data**"*
- The science of gathering data and discovering patterns

Source: dictionary.com

What are the 2 types of statistics?

- **Descriptive** - Summarizing the characteristics of data
- **Inferential** - Modeling, making ‘inferences’ from data

Descriptive statistics

- **Summarizing** the **characteristics** of data
    - <u>*Central tendency*</u> - (“center”) mean, median, mode
    - <u>*Variability*</u> - (“dispersion”) variance, standard deviation
    - <u>*Frequency distribution*</u> - (“occurrence within data”) counts
- Charts, plots, probability distribution shapes

Inferential statistics

- "Modeling" or making 'inferences' from the data
- Taking data from samples and making predictions about populations
- 2 types
  - *Estimating parameters*
  - *Hypothesis tests*

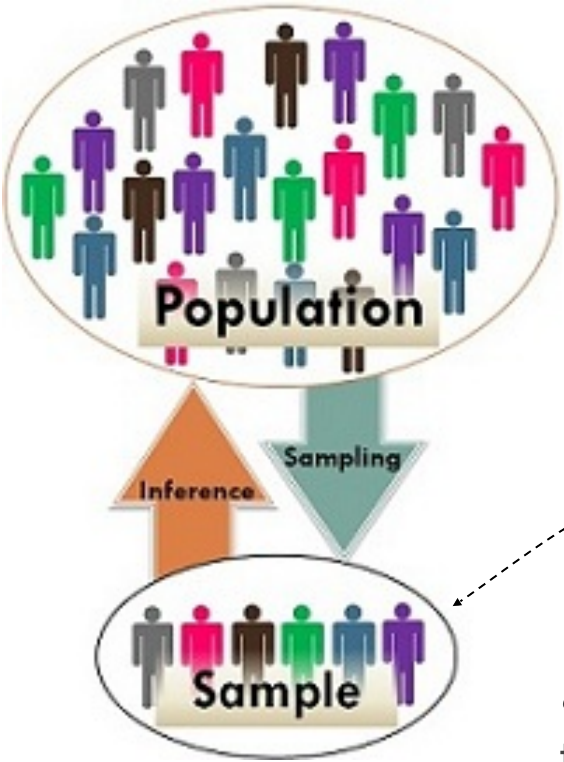
Statistic

"*A quantity computed from a sample*"

Source: dictionary.com

Populations & Samples

We want to learn something about this..



Our population: *all* YouTube comments

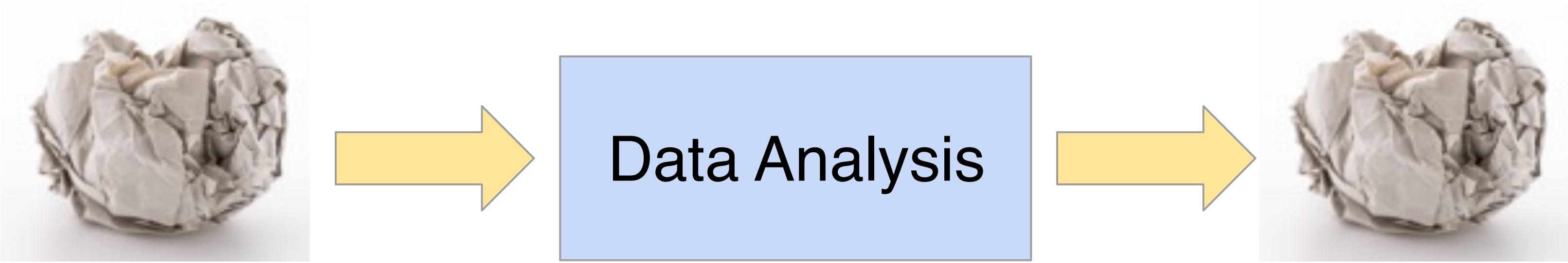
Our sample: 100,000 comments

....but we can only *actually* collect data from this

Image source: keydifferences.com

GIGO : Garbage In. Garbage Out.

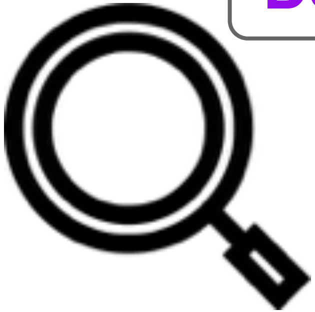
It's *always* worth spending time at the <u>beginning</u> of a project to determine whether or not the data you have are garbage. Be certain they are actually able to help you answer the question you're interested in.



Descriptive

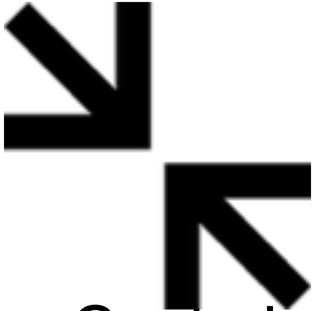
Descriptive Analysis

Size



Missingness

Shape



Central Tendency

Variability

Exploratory: The goal is to find unknown relationships between the variables you have measured in your data set. Exploratory analysis is open ended and designed to verify expected or find unexpected relationships between measurements.

Exploratory

Exploratory Data Analysis (EDA)
detective work answering the question:
"*What can the data tell us?*"

Exploratory

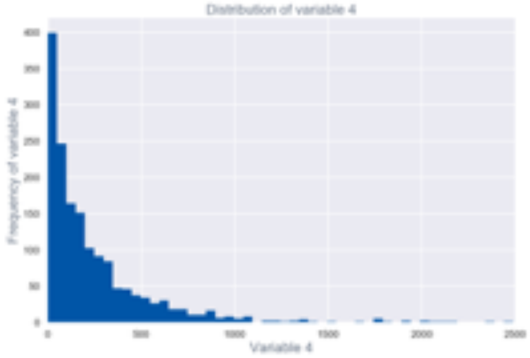
The general principles of exploratory analysis:

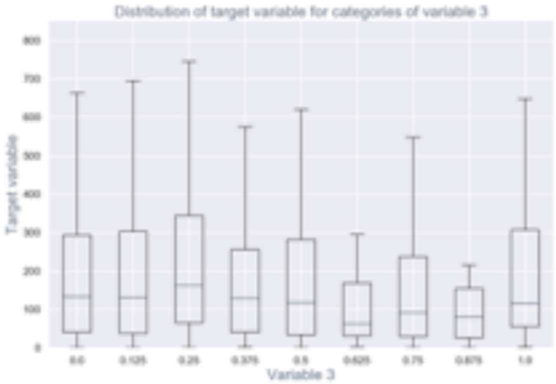
- Look for missing values
- Look for outlier values
- Calculate numerical summaries
- Generate plots to explore relationships
- Use tables to explore relationships
- If necessary, transform variables

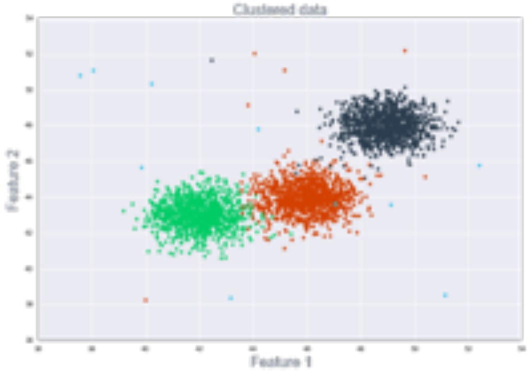
EDA Approaches to “Get a Feel for the Data”

Understanding the relationship between variables in your dataset

Exploratory







<u>Univariate</u>

understanding a single variable

i.e.: histogram, densityplot, barplot

<u>Bivariate</u>

understanding relationship between 2 variables

i.e.: boxplot, scatterplot, grouped barplot, boxplot

<u>Dimensionality Reduction</u>

projecting high-D data into a lower-D space

i.e.: PCA, ICA, Clustering

Image source: https://www.svds.com/value-exploratory-data-analysis/

Text Analysis

Sentiment analysis defined

- Process of analyzing digital or digitized text in order to determine if the emotional tone is positive, negative or neutral
- Large volumes of text data are now available in forms of
  - Emails
  - Messages
  - Support transcripts
  - Social Media interactions
  - Reviews
  - Digitized phone messages and interaction records (i.e. UCSD)

When doing sentiment analysis...

Token - a meaningful unit of text

- What you use for analysis
- *Tokenization* takes corpus of text and splits it into tokens (words, bigrams, etc.)

Stop words - words not helpful for analysis

- Extremely common words such as “the”, “of”, “to”
- Are typically removed from analysis

When doing sentiment analysis...

Stemming - lexicon normalization

- Identifying the root for each token
- Jumping, jumped, jumps, jump all have the same root ‘jump’
- Where things get tricky: jumper???

In text analysis, your choices matter:

1. How to tokenize?
2. What lexicon to use?
3. Remove stop words? Remove common words?
4. Use stemming?

TF-IDF

Term Frequency - Inverse Document Frequency

Term Frequency can only tell us so much….

2017

2018

2019

2020

TF-IDF:

Term Frequency - Inverse Document Frequency

Term Frequency (TF) : how frequently a word occurs in a document

Inverse document frequency (IDF) : intended to measure how important a word is to a document

decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

TF-IDF:

Term Frequency - Inverse Document Frequency

the frequency of a term adjusted for how rarely it is used

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

https://www.tidytextmining.com

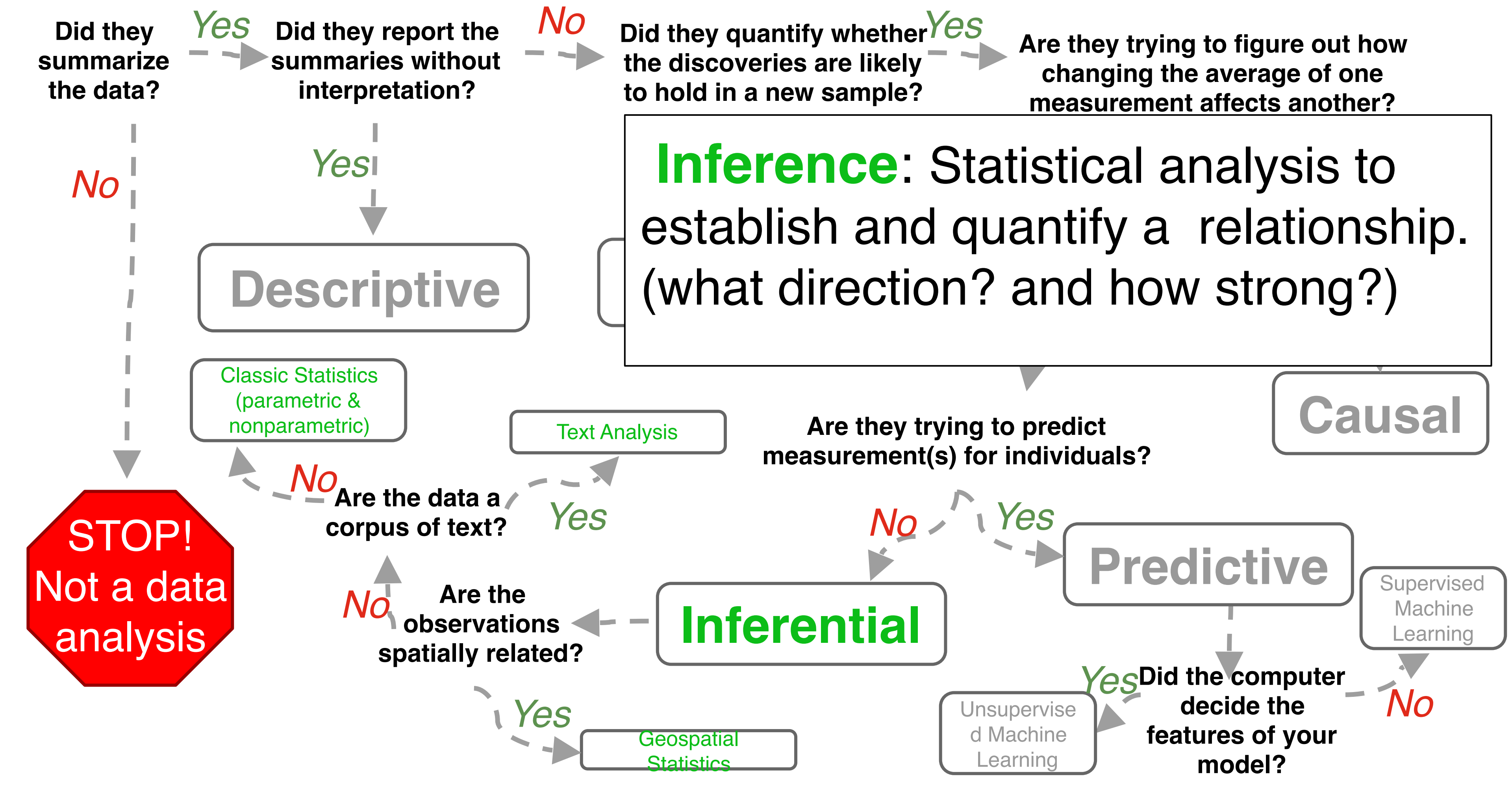
Questions we can ask...

1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are the sentiment of the #1 songs?
9. What words contribute to the sentiment of these #1 songs?
10. ...what about bigrams? N-grams?

EDA

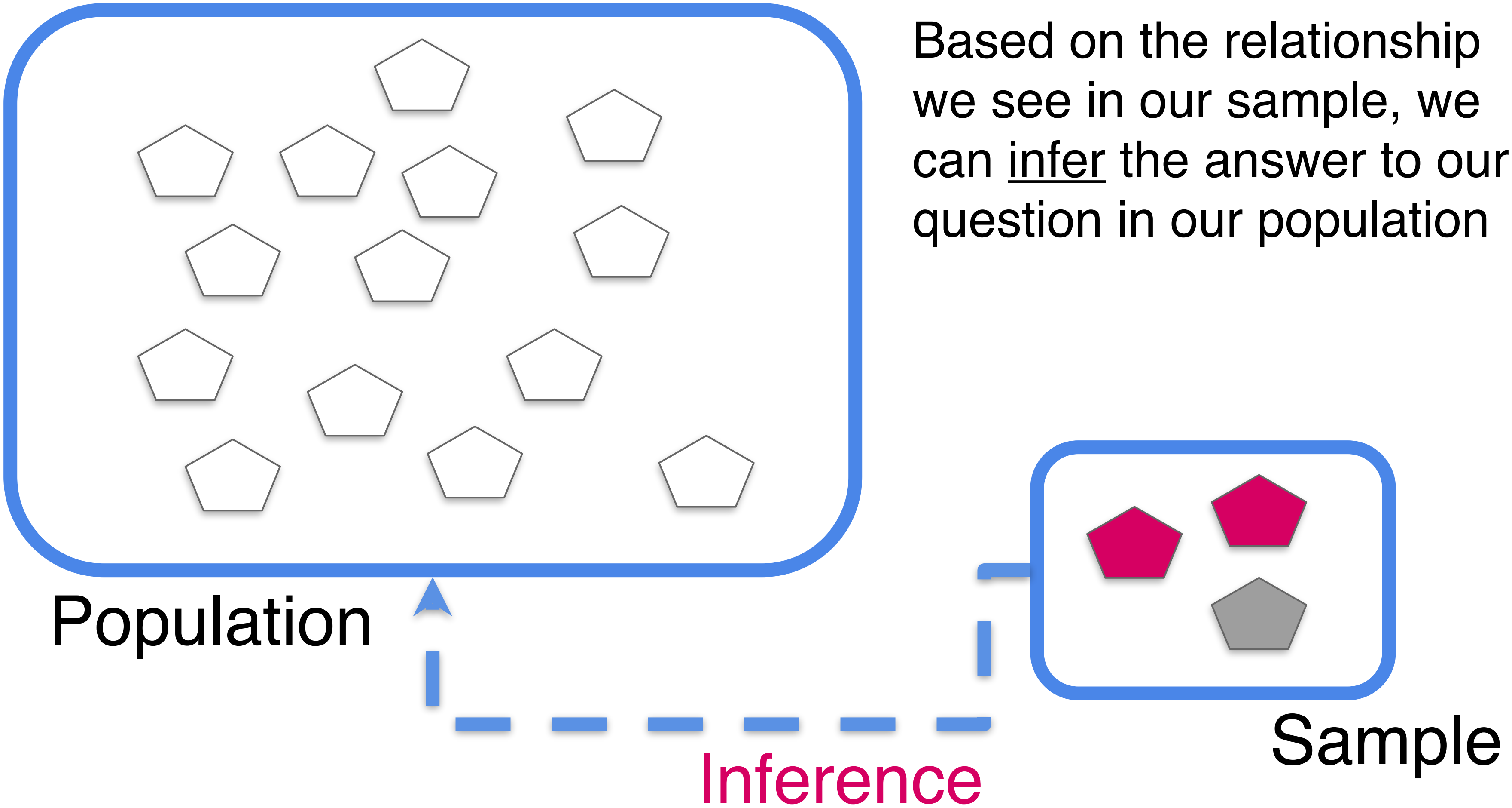
TF-IDF

Sentiment Analysis



Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
No
Yes
Descriptive
Inference: Statistical analysis to establish and quantify a relationship. (what direction? and how strong?)
Classic Statistics (parametric & nonparametric)
Text Analysis
Are they trying to predict measurement(s) for individuals?
Causal
No
Are the data a corpus of text?
Yes
STOP! Not a data analysis
No
Yes
Predictive
No
Are the observations spatially related?
Inferential
Supervised Machine Learning
Yes
Did the computer decide the features of your model?
Unsupervised Machine Learning
No
Yes
Geospatial Statistics

Based on the relationship we see in our sample, we can infer the answer to our question in our population



!

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

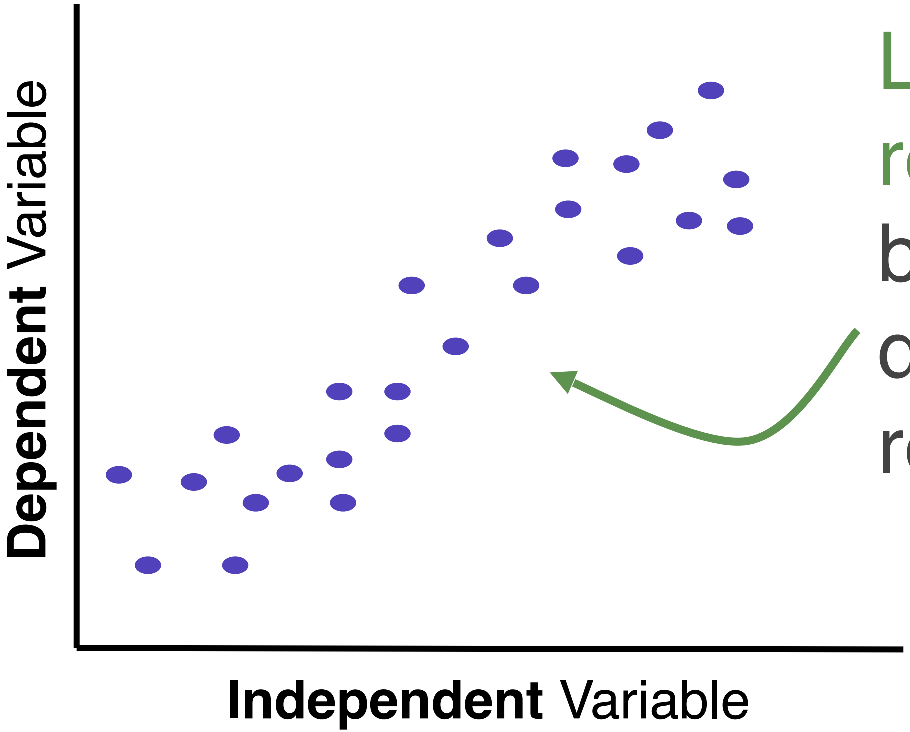
DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

NON-PARAMETRIC TESTS

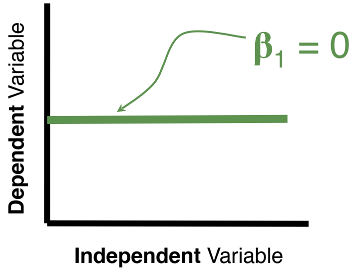
FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

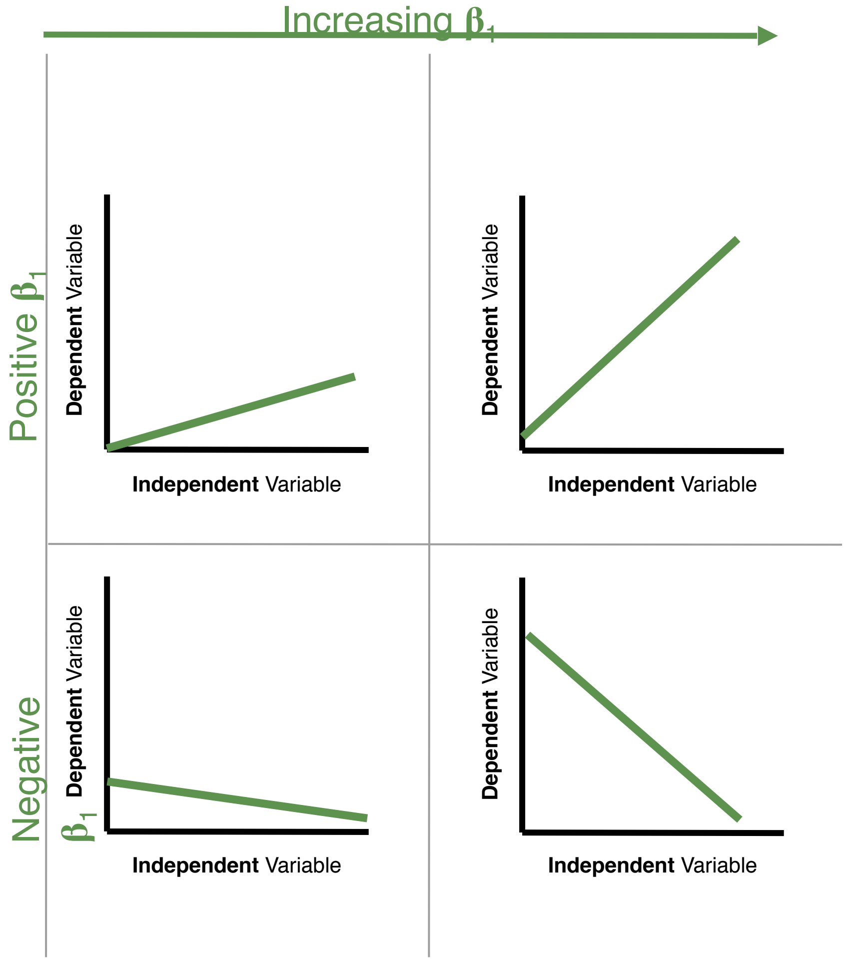
i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test



Linear regression can be used to describe this relationship

Effect size (β_1) can be estimated using the slope of the line





Assumptions of linear regression

1. Linear relationship
2. No multicollinearity
3. No auto-correlation
4. Homoscedasticity

p-value : the probability of getting the observed results (or results more extreme) by chance alone

The probability of getting 10 heads *or something more extreme* is

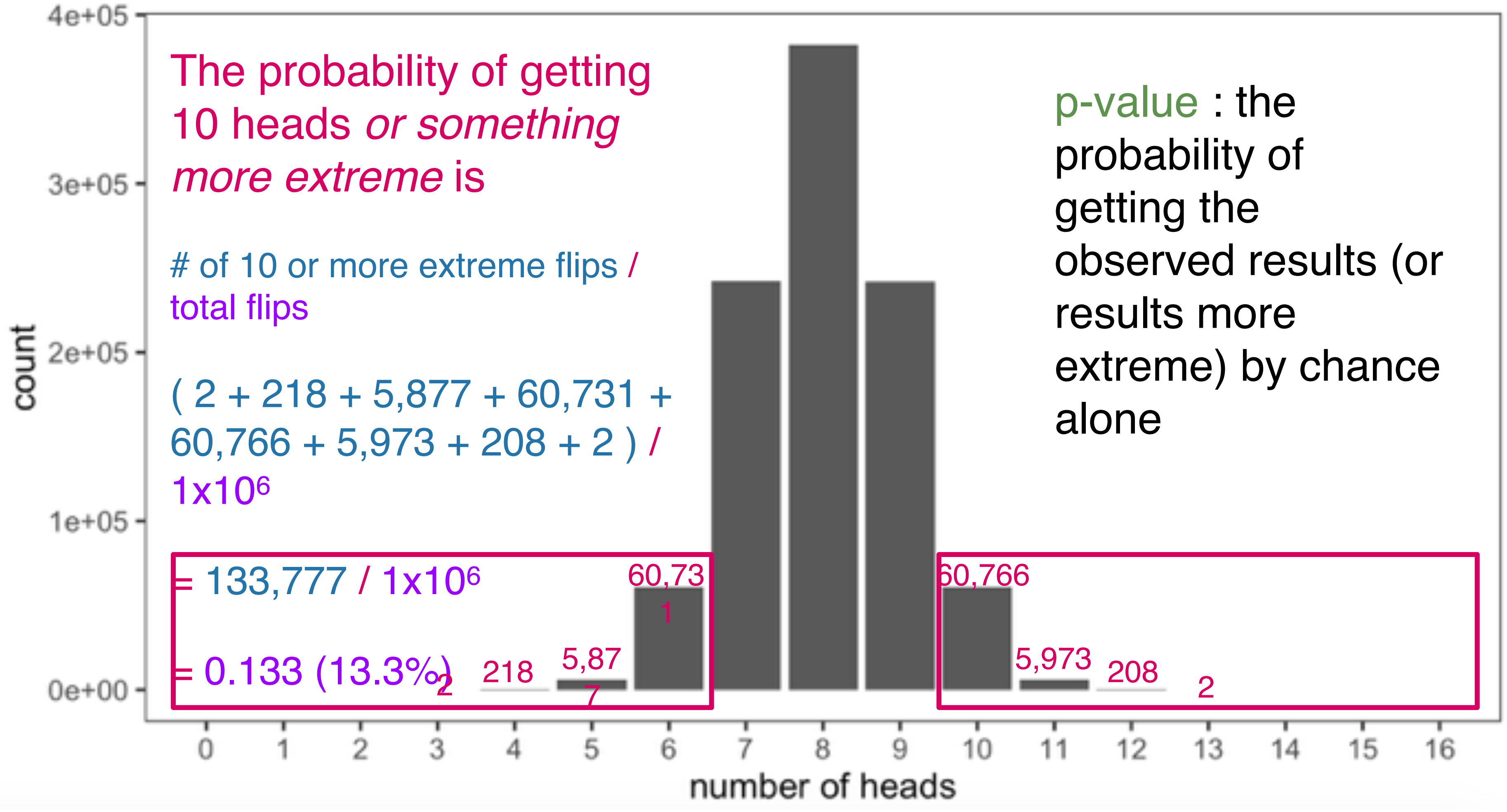
# of 10 or more extreme flips / total flips

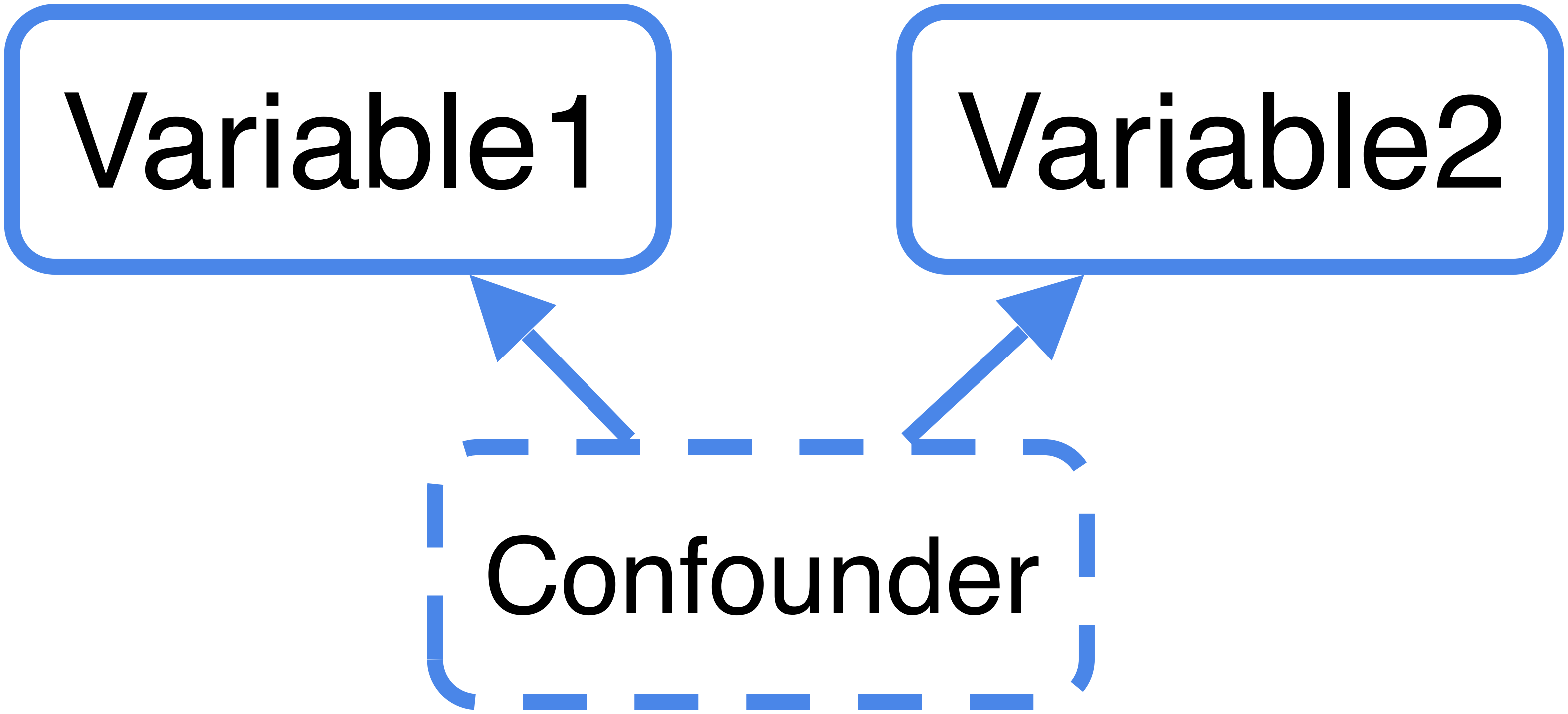
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / $1x10^6$

= 133,777 / $1x10^6$

= 0.133 (13.3%)

p-value : the probability of getting the observed results (or results more extreme) by chance alone




Variable1
Variable2
Confounder

Visualization

- Human brain has trouble making sense of large amounts of data produced by computational modeling and experimentation
- As more computational methods are applied, more and more information is being created
- Scientific visualization is one way of making important information explicit and simple to process
- http://svs.gsfc.nasa.gov/

Perceptual example - afterimages

Additive vs. Subtractive Color

- RGB
  - **red-green-blue**
  - **Additive scheme**
- CMY
  - **Cyan-magenta-yellow**
  - **Subtractive scheme**
  - **Black (CMYK) is typically added to inkjet printers**
    - Difficult to make exact black by mixing CMY, requires precision
    - Typically one uses black the most so it makes sense to have a separate ink cartridge for black
- HSV
  - **Hue-saturation-value**
  - **Many feel this is a more natural way to describe color for humans**

Example: Bad color matching

- Eeeghh!
- The red and blue are on opposite ends of the visual color spectrum, so we have trouble focusing on both colors simultaneously
- I could have made this worse by adding all equations, but last time too many people passed out!
- AVOID REDS ON BLUES OR BLUES ON REDS

Example: Good color matching

- Ahhh…
- This is much more comfortable for the eyes.
- Choose colors which are based on luminance differences
- generally avoid two fully saturated colors as foreground and background
- Increase contrast by reducing the perceived intensity of either the foreground or background

Luminance Equation

$$Y = 0.30 * Red + 0.59 * Green + 0.11 * Blue$$

- Perceived intensity due to a color
  - Different contributions of red/green/blue components
  - Empirically determined

Contrast tables

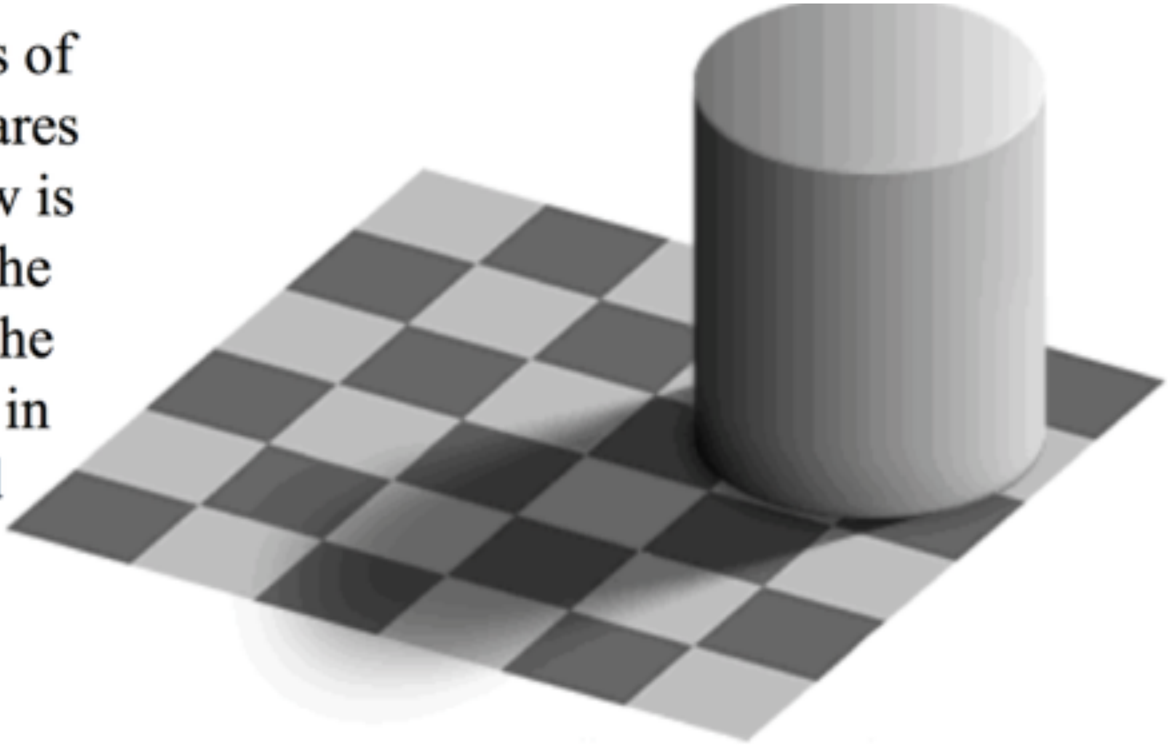
	Black	White	Red	Green	Blue	Cyan	Magenta	Orange	Yellow
Black	0.00	1.00	0.30	0.59	0.11	0.70	0.41	0.60	0.89
White	1.00	0.00	0.70	0.41	0.89	0.30	0.59	0.41	0.11
Red	0.3	0.7	0.00	0.29	0.19	0.40	0.11	0.30	0.59
Green	0.59	0.41	0.29	0.00	0.48	0.11	0.18	0.01	0.30
Blue	0.11	0.89	0.19	0.48	0.00	0.59	0.30	0.49	0.78
Cyan	0.70	0.30	0.40	0.11	0.59	0.00	0.29	0.11	0.19
Magenta	0.41	0.59	0.11	0.18	0.30	0.29	0.00	0.19	0.48
Orange	0.60	0.41	0.30	0.01	0.49	0.11	0.19	0.00	0.30
Yellow	0.89	0.11	0.59	0.30	0.78	0.19	0.48	0.30	0.00

Table 5.1: A color contrast table can be formed by subtracting the luminance equation values for two different colors, then taking the absolute value.

From C.A. Simpkins, *Design,* Manuscript

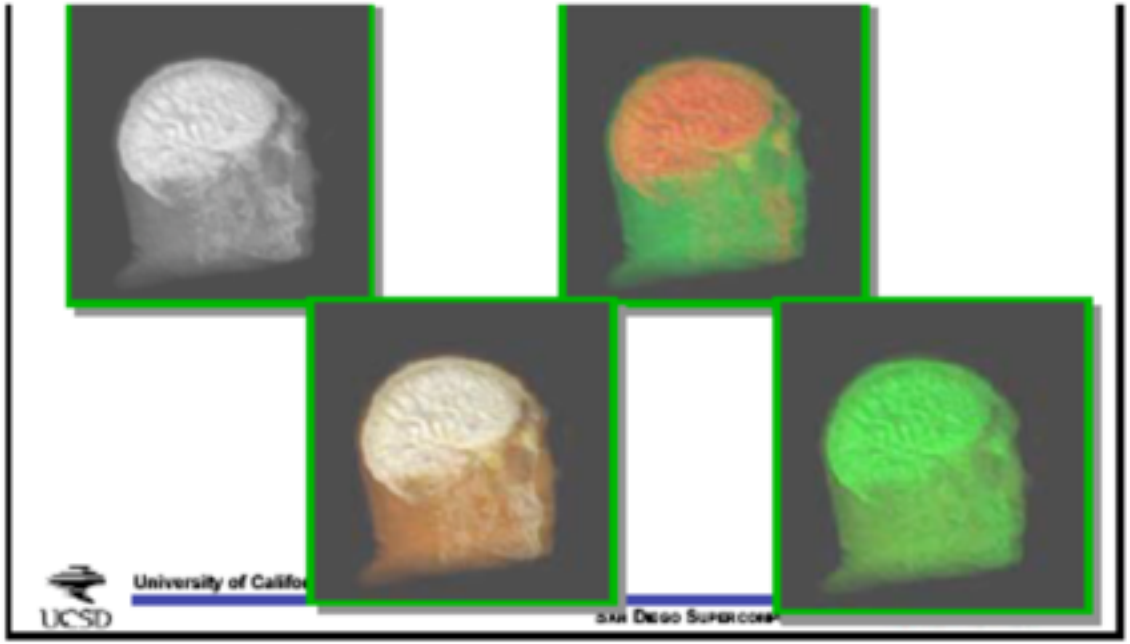
Perceived lightness is context dependent as well

- The lightness of the light squares in the shadow is the same as the lightness of the dark squares in the unshaded region



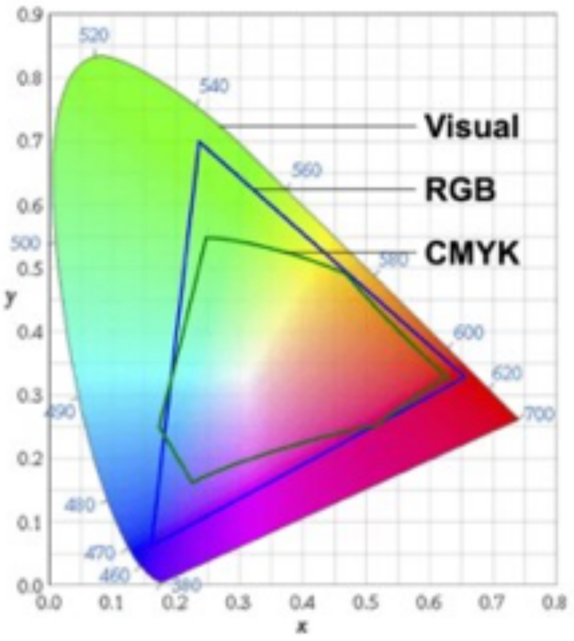
False color representation and color maps

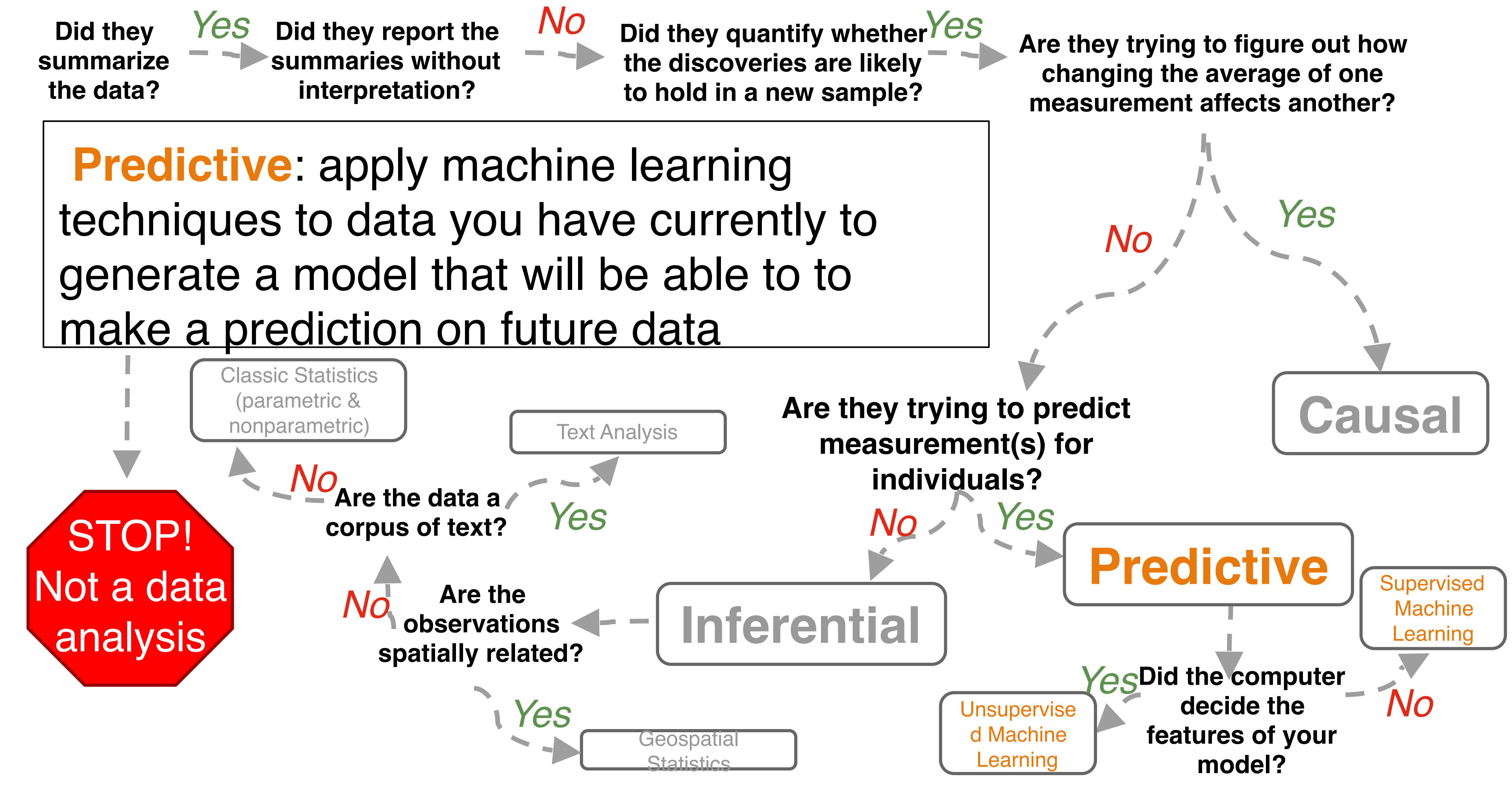
- Map values from any range to a map of colors
  - i.e. a matrix of 0-1 range-> white-black



Color **Gamut comparison**

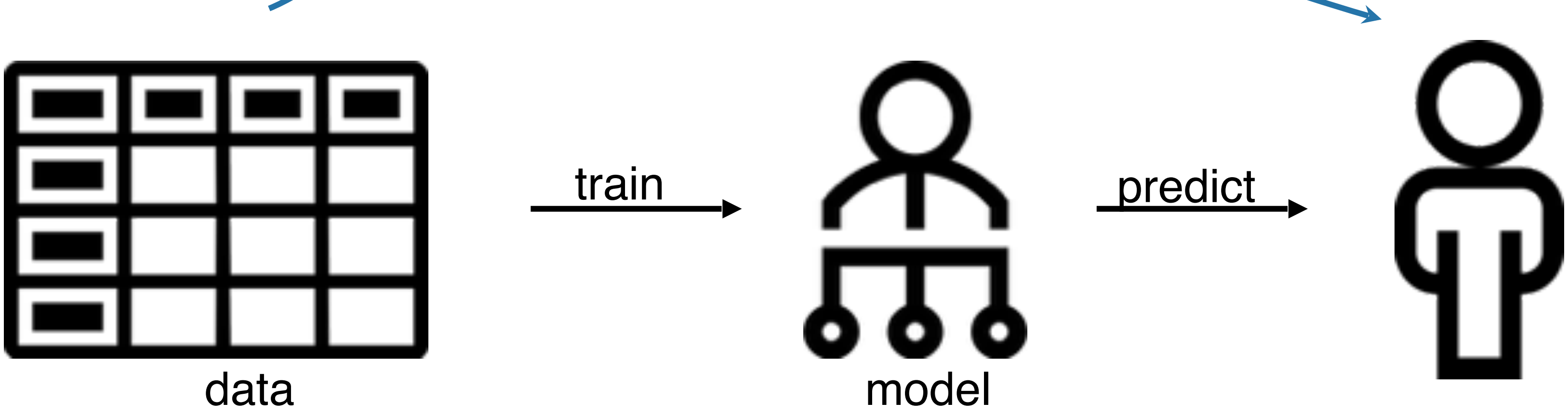
- The range of colors a device can display
- This can be a triangle or more complex shapes
- Typically a subset of human perception
    - Stay away from what cannot be printed when creating for papers




Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
Predictive: apply machine learning techniques to data you have currently to generate a model that will be able to to make a prediction on future data
No
Yes
Causal
Are they trying to predict measurement(s) for individuals?
No
Yes
Predictive
Inferential
Classic Statistics (parametric & nonparametric)
Text Analysis
No
Are the data a corpus of text?
Yes
STOP! Not a data analysis
No
Are the observations spatially related?
Yes
Geospatial Statistics
Supervised Machine Learning
Yes
Did the computer decide the features of your model?
No
Unsupervised Machine Learning

predictive analysis
uses data you have now
to make predictions in
the future

machine learning
approaches are used for
predictive analysis!



So what does that mean?

- ‘learning’ parameters from data in order to map state into action
- Learning essentially boils down to some sort of calculation
- Why implicit vs. explicit programming?
    - Camera example and parameters
    - Robotics
    - Expert systems

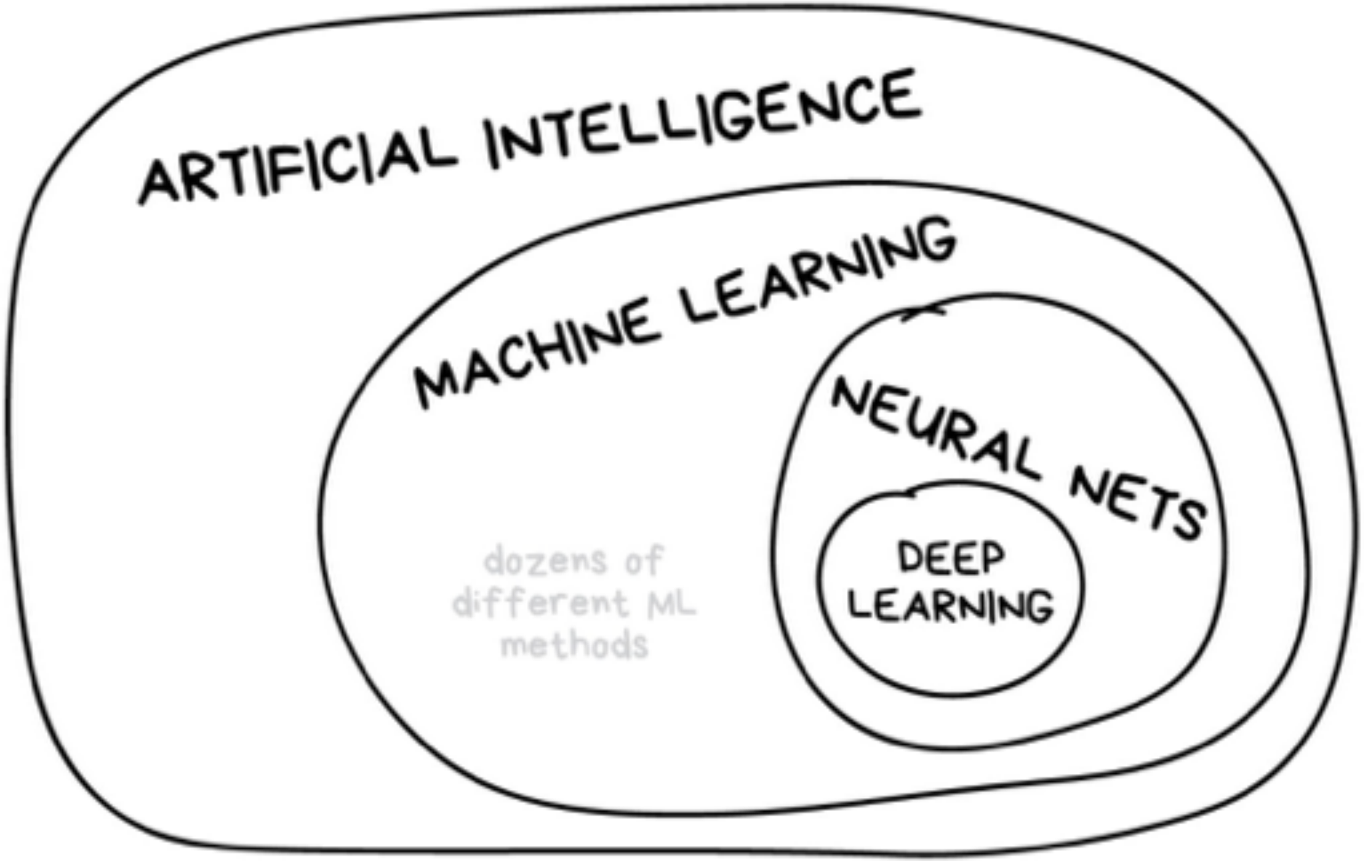
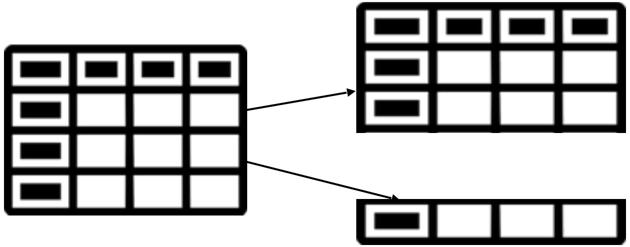
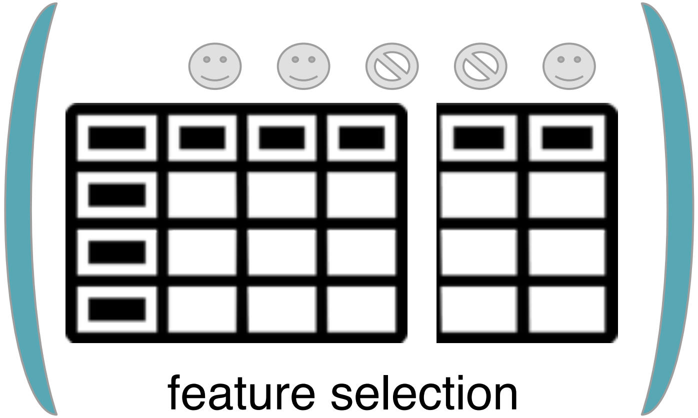


Image source: https://vas3k.com/blog/machine_learning/

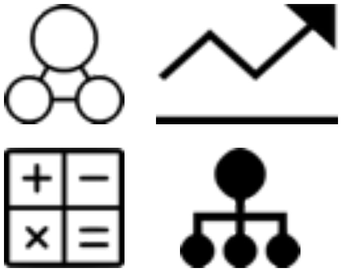
Basic Steps to Prediction



data
partitioning

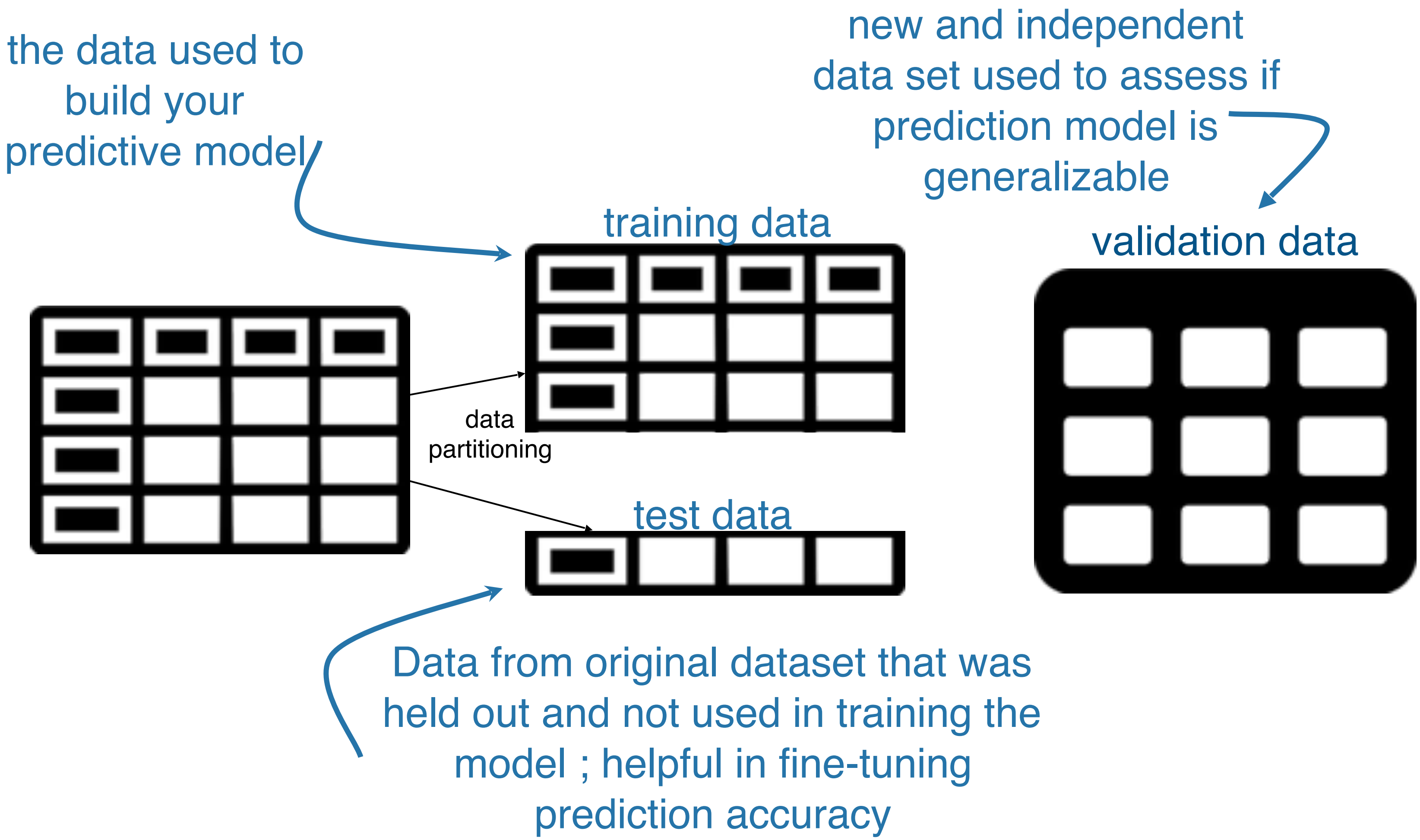


feature selection



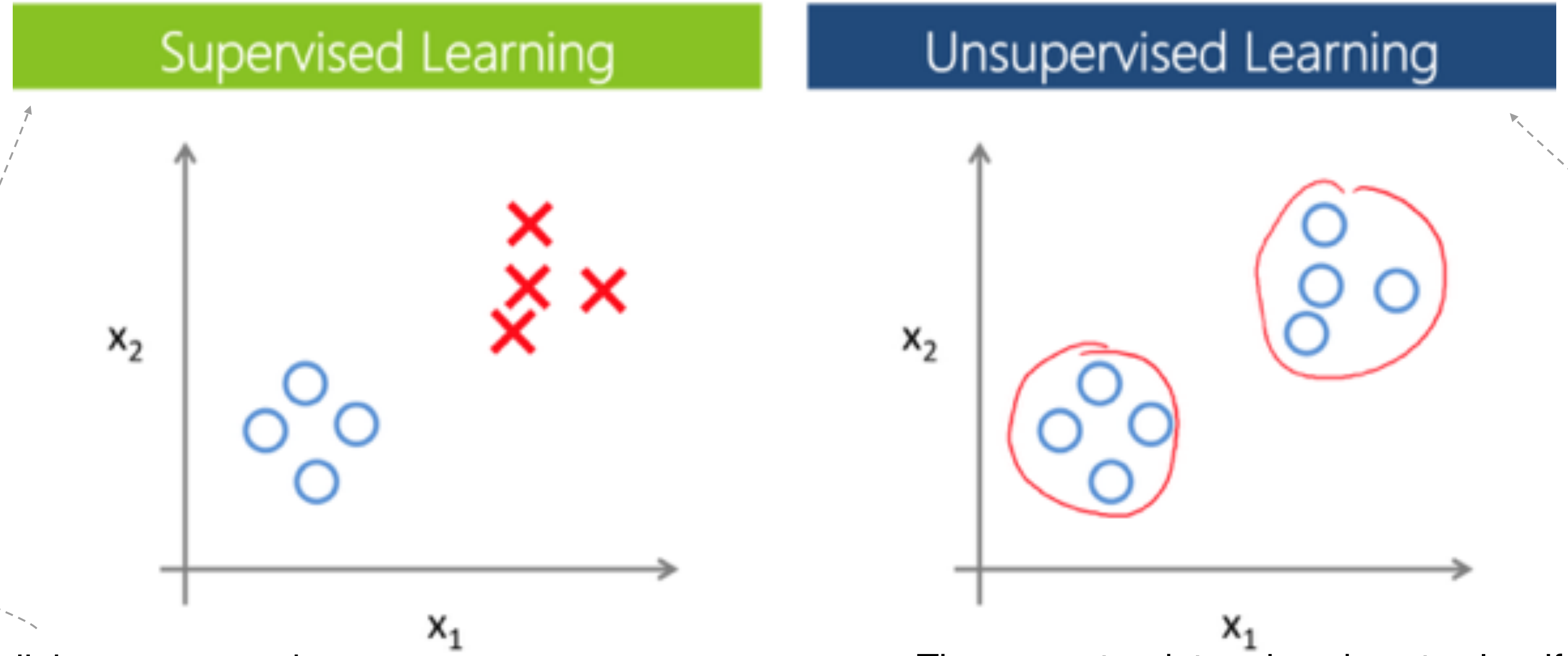
model selection

model assessment



the data used to build your predictive model
new and independent data set used to assess if prediction model is generalizable
training data
validation data
data partitioning
test data
Data from original dataset that was held out and not used in training the model ; helpful in fine-tuning prediction accuracy

Two modes of machine learning



You tell the computer what features to use to classify the observations

The computer determines how to classify based on properties within the data

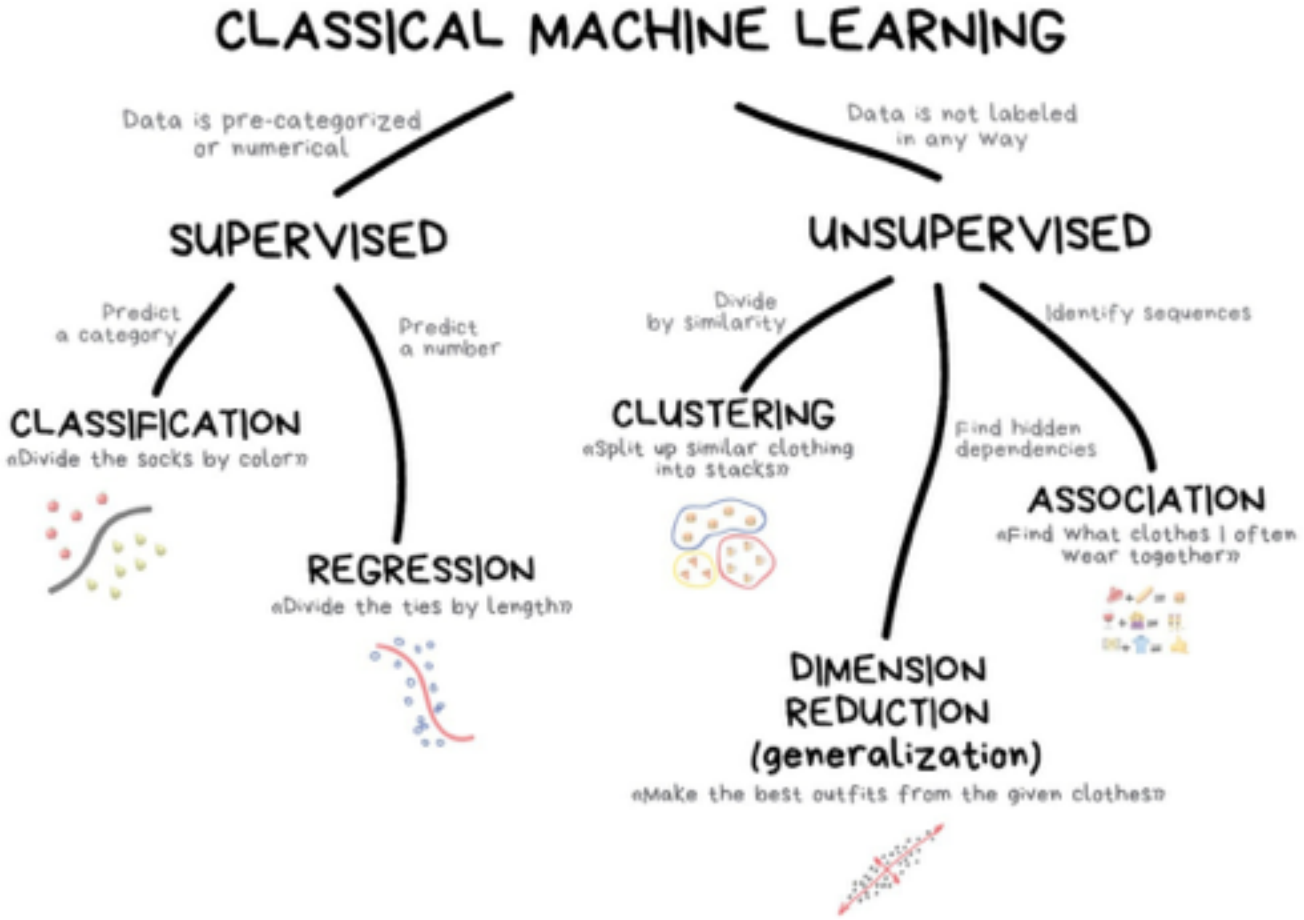


Image source: https://vas3k.com/blog/machine_learning/

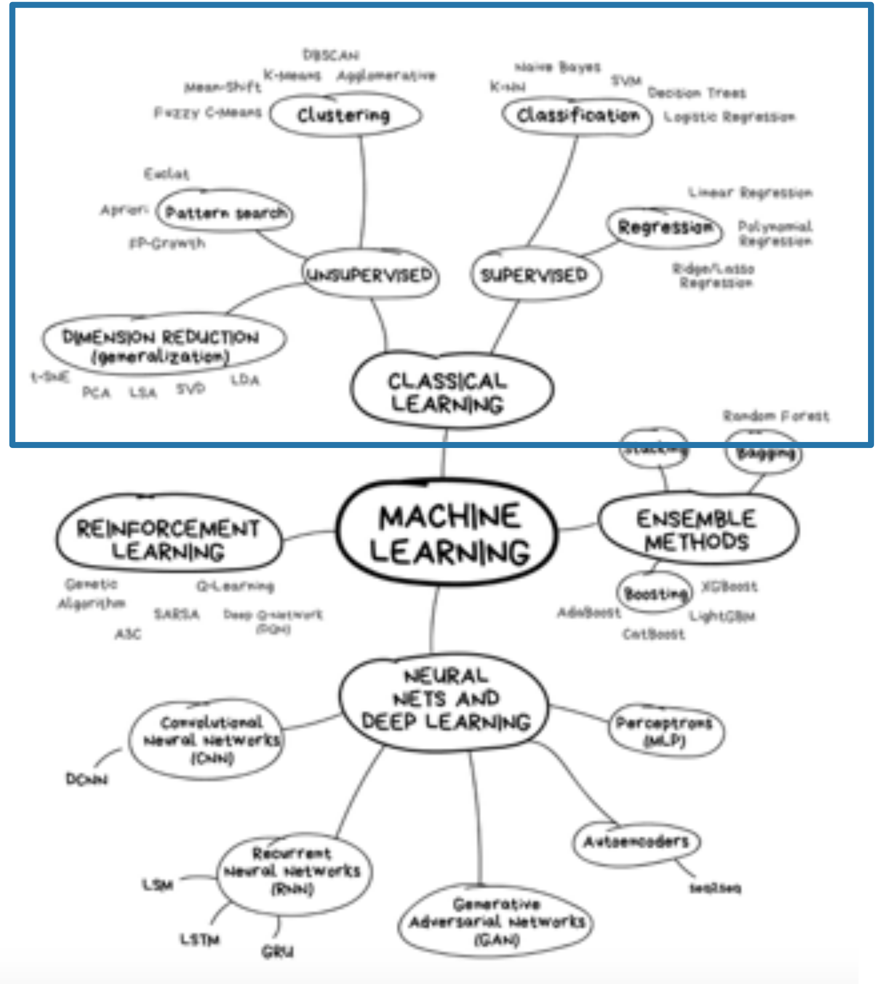


Image source: https://vas3k.com/blog/machine_learning/

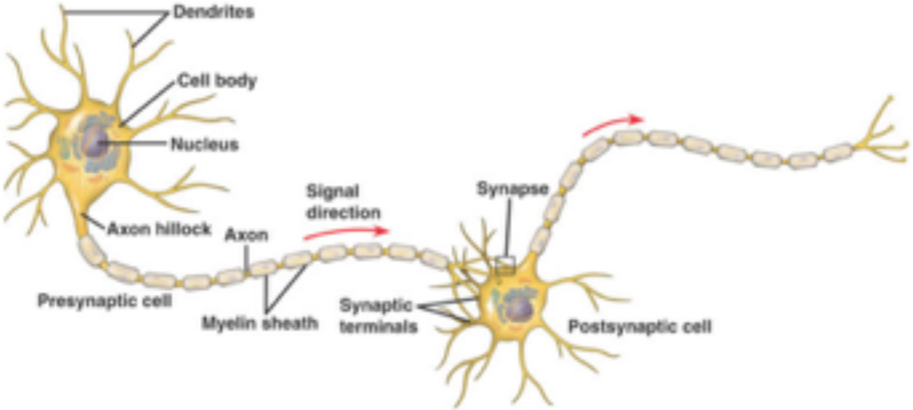
Regression:
predicting <u>continuous</u>
variables
(i.e. Age)

<u>continuous</u> variable prediction

Classification:
predicting <u>categorical</u>
variables
(i.e. education level)

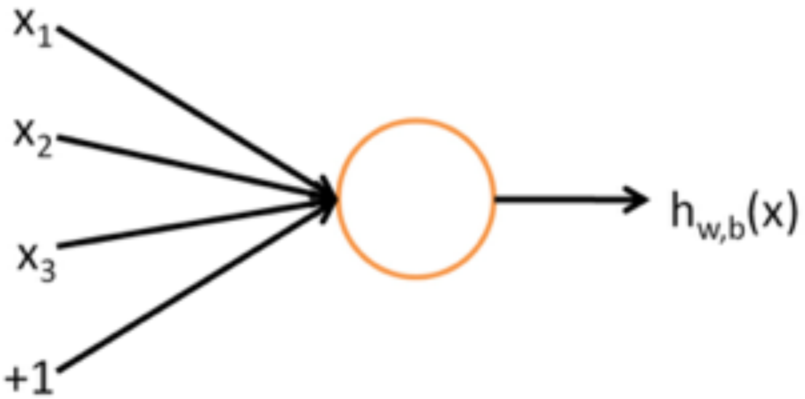
<u>categorical</u> variable prediction

WHAT IS A NEURON?



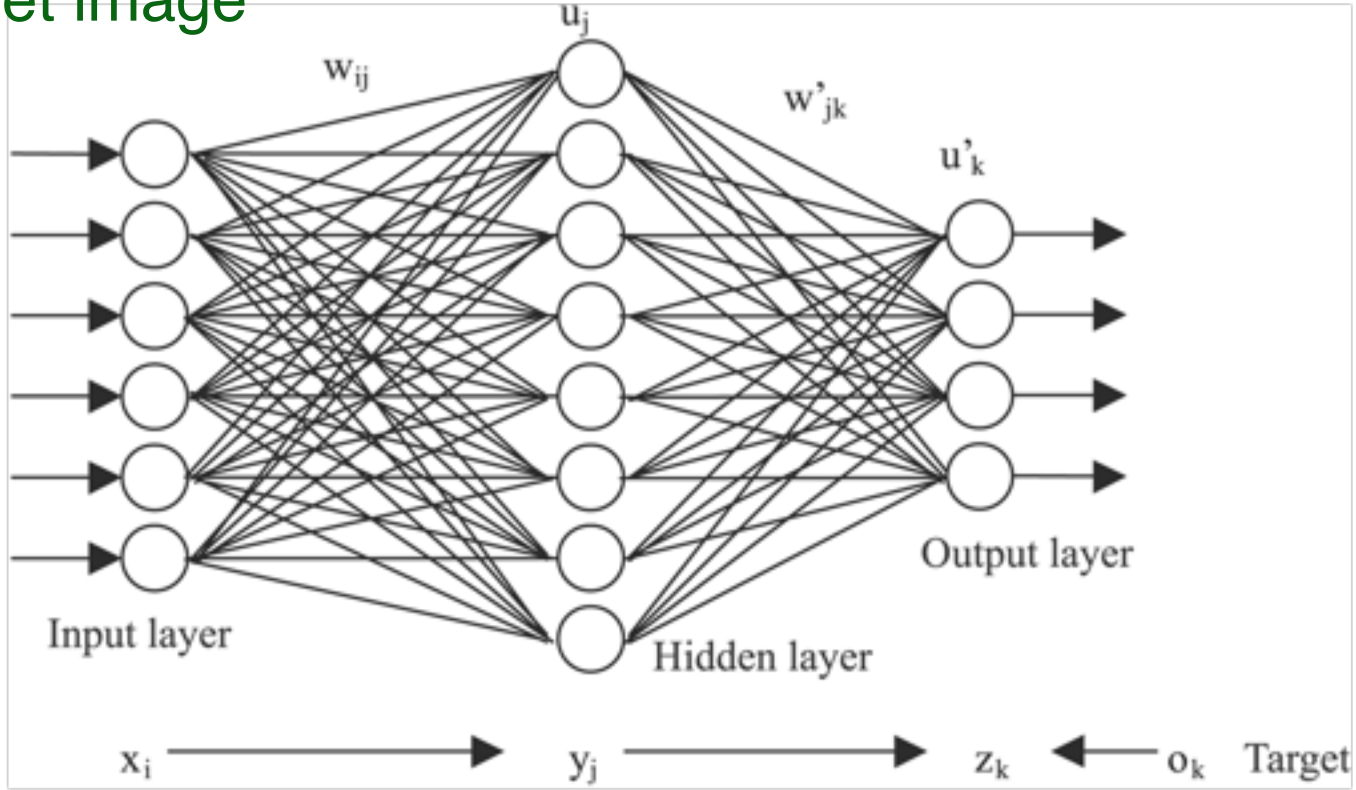
- Receives signal on synapse
- When trigger sends signal on axon

MATHEMATICAL NEURON



- Mathematical abstraction, inspired by biological neuron
- Either on or off based on sum of input

This will likely not be the last time you see this (mostly unhelpful) neural net image



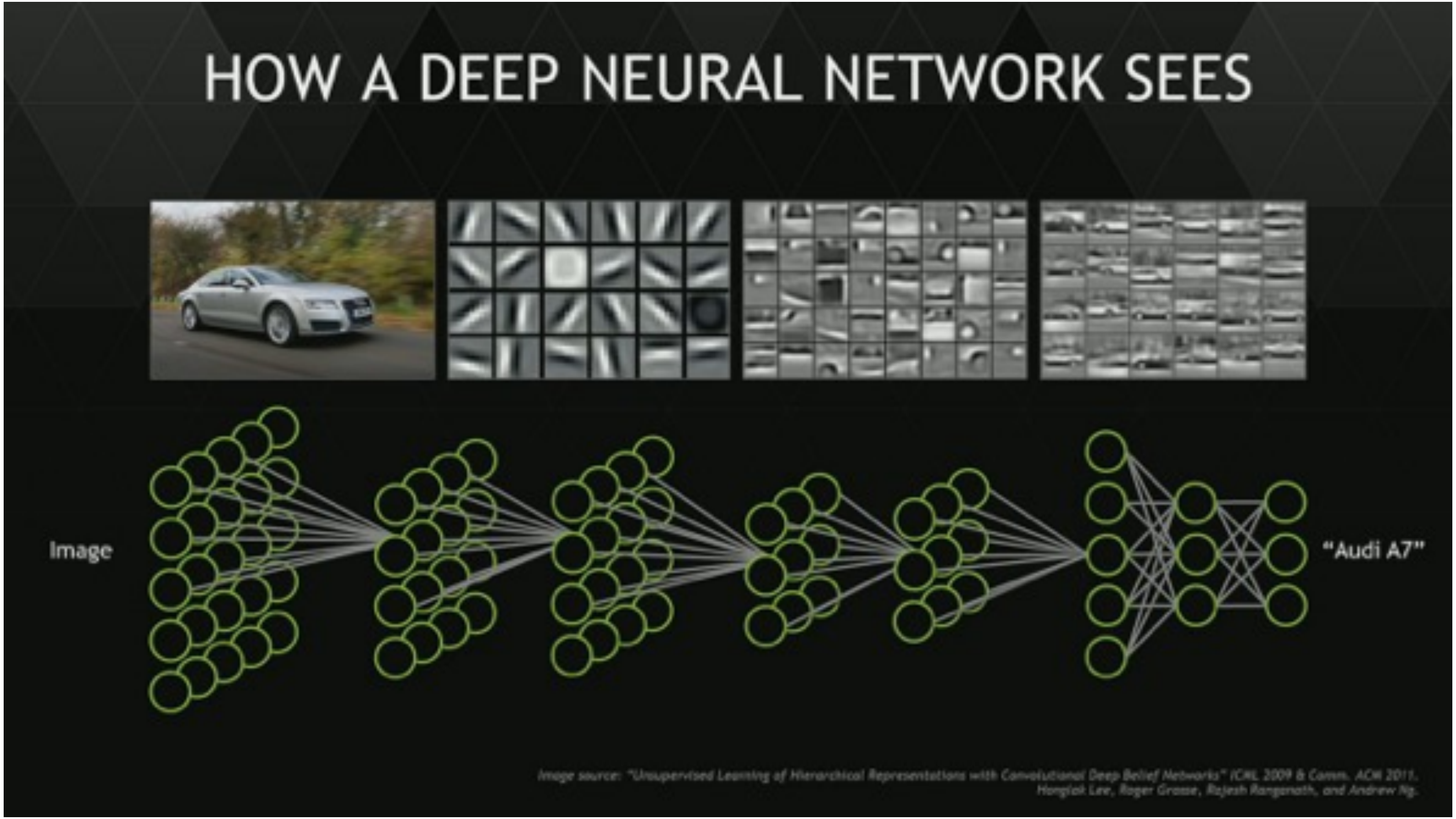


Image Source: https://towardsdatascience.com/understanding-residual-networks-9add4b664b03

These weights tell the neuron to respond more to one input and less to another. Weights are adjusted when training — that's how the network learns. Basically, that's all there is to it.

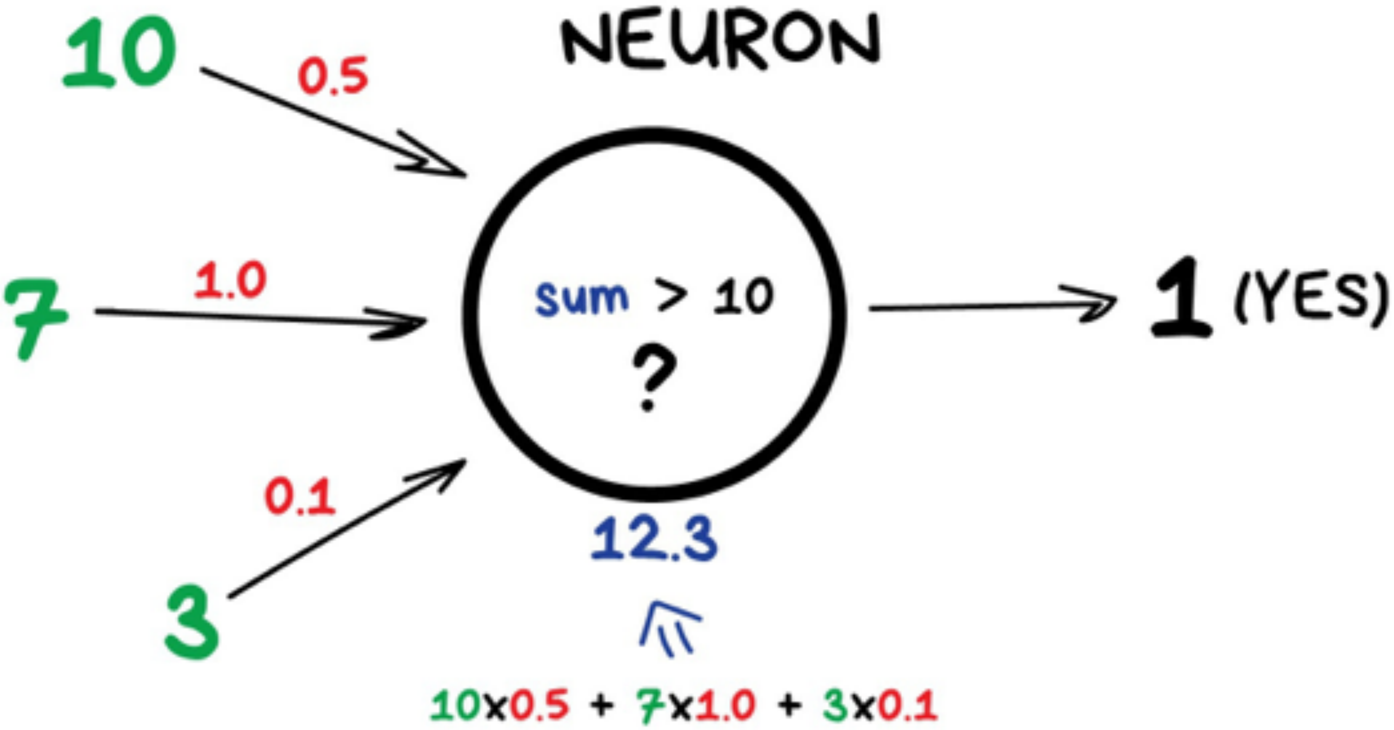


Image source: https://vas3k.com/blog/machine_learning/

model assessment

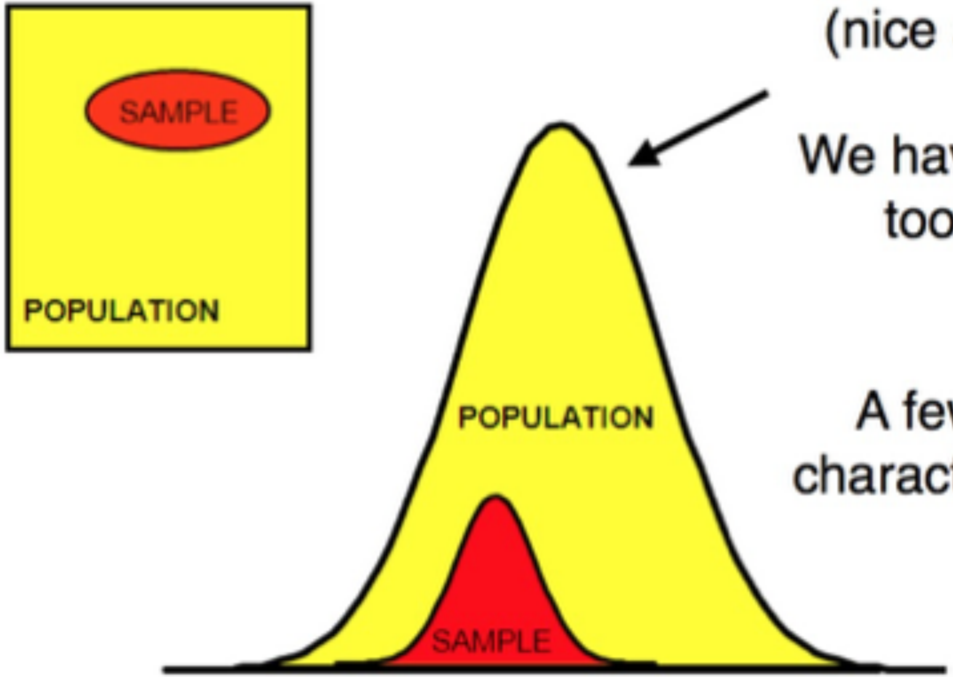
<u>categorical</u> variable prediction

<u>continuous</u> variable prediction

$$\text{Accuracy} = \frac{\#\text{ of samples predicted correctly}}{\#\text{ of samples predicted}} * 100$$

Accuracy	What % were predicted correctly?
Sensitivity	Of those that *were* positives, what % were predicted to be positive?
Specificity	Of those that were *negatives*, what % were predicted to be negative?

Non-parametric Statistics: The Why



Normal distribution
(nice and friendly)

We have good math tools for this.

A few parameters **fully** characterize the distribution.

Resampling statistics: The What

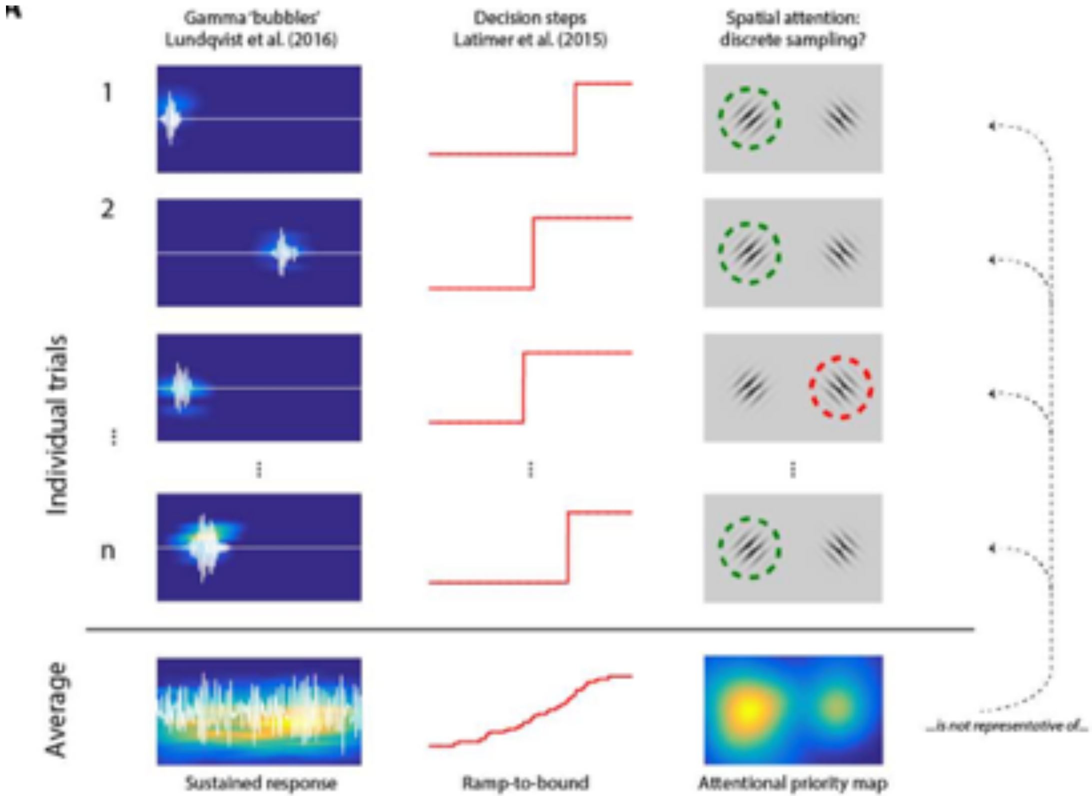
- Bootstrap (Monte Carlo)
- Rank Statistics (Mann Whitney U)
- Kolmogorov-Smirnoff Test
- Non-parametric prediction models

Why do we even teach/use parametric statistics anyway?

Parametric approaches:

- Lots of data follow expected patterns
- Require less data
- More sensitive
- Quicker to run/train/predict
- More resistant to overfitting

Combining by computing mean doesn't necessarily create a good representation

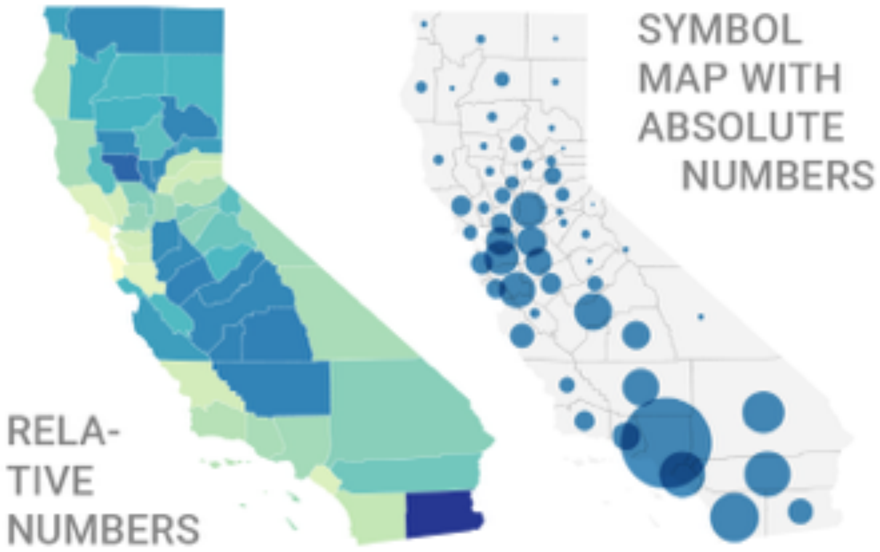


Geospatial Analysis?

Choropleth should display relative differences, *not* absolute numbers

NOT IDEAL

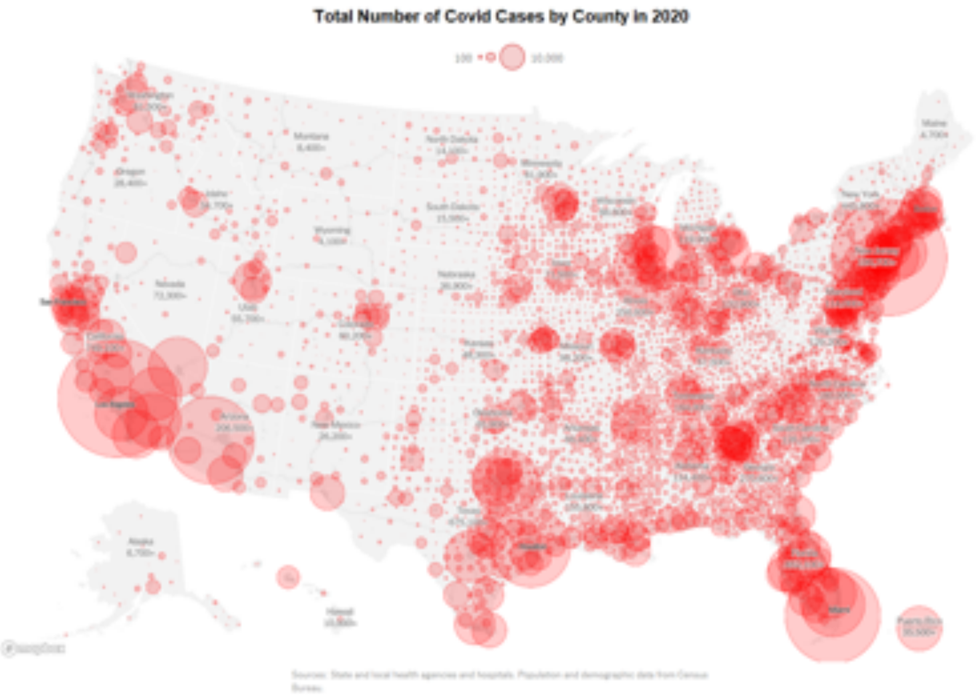
BETTER



https://blog.datawrapper.de/choroplethmaps/

Bubble maps

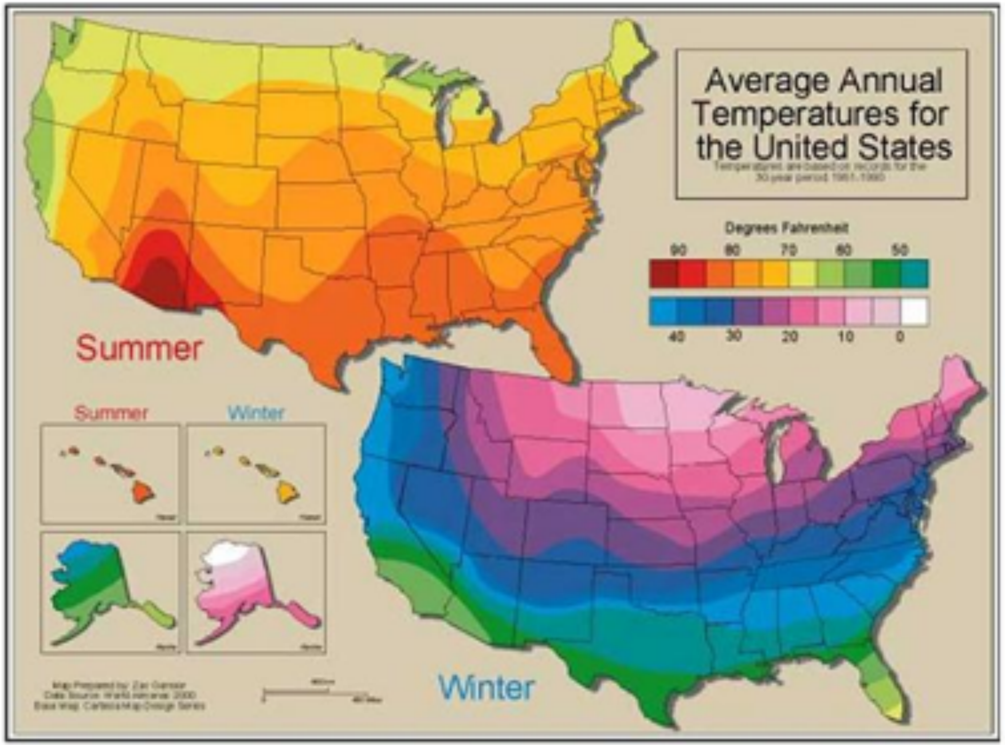
- Coordinates of latitude and longitude
- Bubble size is third axis, such as population density, COVID cases, etc
- Notes:
    - *Consider using area rather than radius to avoid exaggerating bubble sizes*
    - *Transparency for bubbles*
    - *Legend!*



https://www.nytimes.com/2020/09/10/learning/whats-going-on-in-this-graph-covid-19-cases-in-america.html

Isarithmic maps demonstrate smooth, continuous phenomena

(temperature, elevation, rainfall, etc.)



http://enb105-2012s-dem.blogspot.com/2012/03/isarithmic-map.html

Spatial Statistics

The statistical techniques we've discussed so far don't work well when considering spatial distributions…

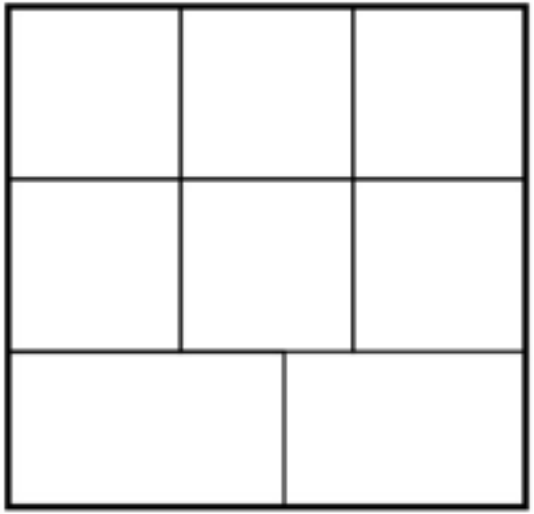
Spatial data violate conventional statistics:

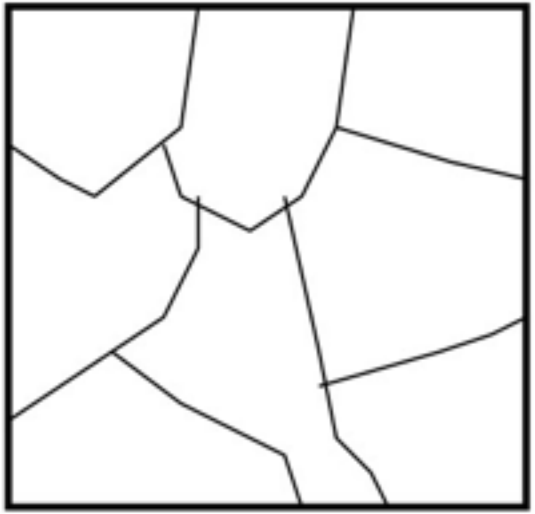
Violations of conventional statistics:

- Spatial autocorrelation
- Modifiable areal unit problem (MAUP)
- Edge effects (Boundary problem)
- Ecology fallacy
- Nonuniformity of space

Modifiable Areal Unit Problem (MAUP)

<u>modifiable area:</u> Units are arbitrarily defined and different organization of the units may create different analytical results.





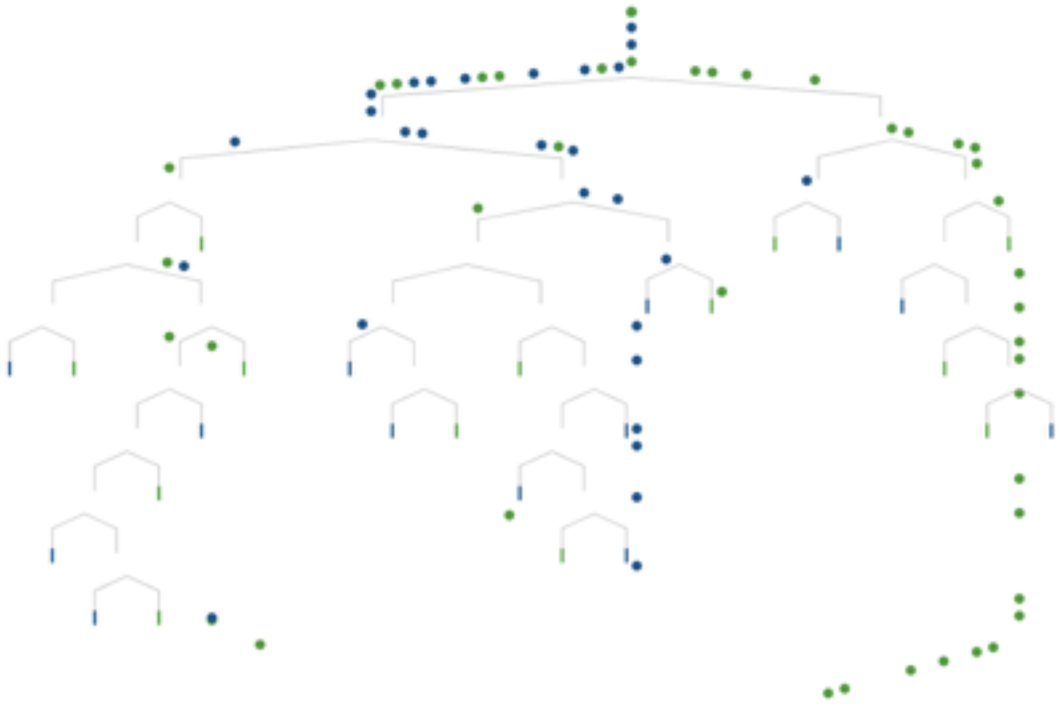
Basic Geospatial Analysis: Summary

1. Considerations when visualizing spatial data important to conclusions drawn
    a. values to plot?
    b. map type?
    c. color scale?
2. Traditional statistics fail with geospatial data:
    a. Spatial autocorrelation
    b. MAUP
    c. Edge effects
    d. Ecological fallacy
    e. Nonuniformity of space
3. Analysis still possible
    a. Global Point Density, Quadrat Density, Kernel Density
    b. Poisson Point Process
    c. K-Nearest Neighbor (KNN)
    d. Comparison to a CRP (using simulation)

Making predictions

The decision tree **model** can then predict which homes are in which city.

Here, we're using the **training data**.



http://www.r2d3.us/visual-intro-to-machine-learning-part-1/?from=@

Recap

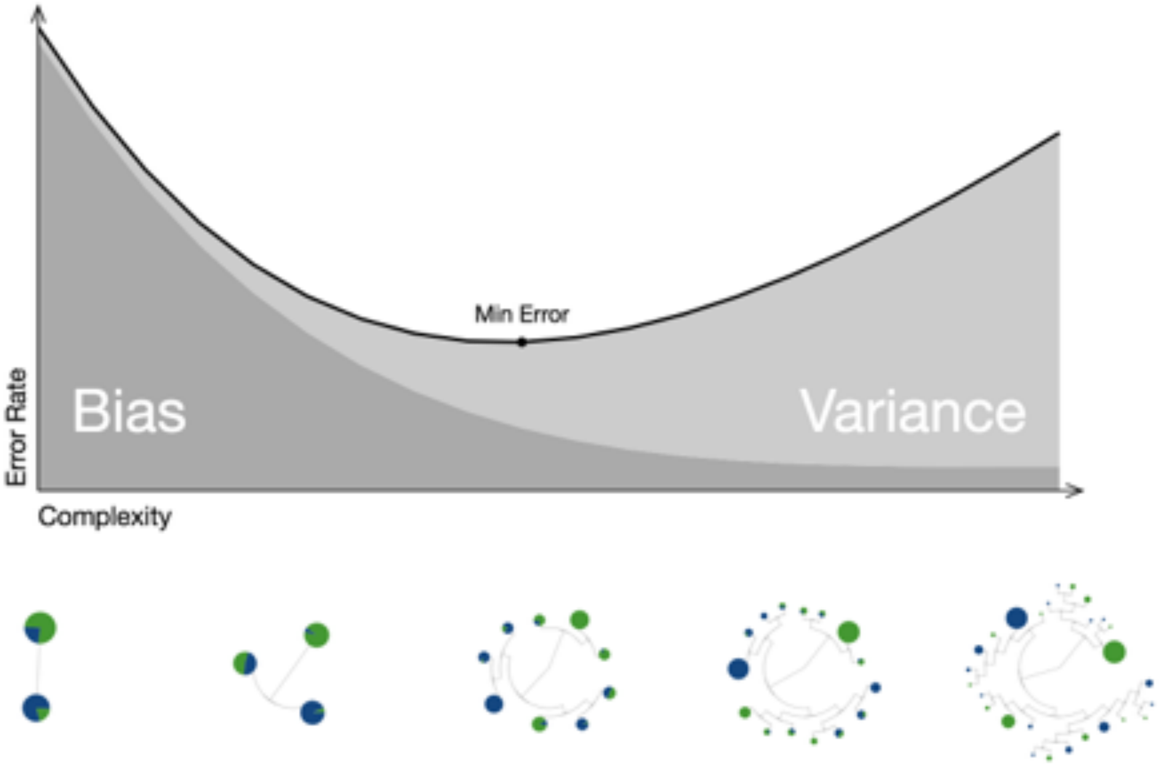
1. Machine learning identifies patterns using **statistical learning** and computers by unearthing **boundaries** in data sets. You can use it to make predictions.

2. One method for making predictions is called a decision tree , which uses a series of if-then statements to identify boundaries and define patterns in the data.

3. **Overfitting** happens when some boundaries are based on on *distinctions that don't make a difference*. You can see if a model overfits by having test data flow through the model.

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/?from=@

Bias-variance tradeoff



http://www.r2d3.us/visual-intro-to-machine-learning-part-1/?from=@

Bias-variance tradeoff

- **High variance** models make mistakes in *inconsistent* ways.
- **Biased models** tend to be overly simple and not reflect reality
- What to do:
    - Consider tuning parameters in the model
        - Can avoid overfitting by setting minimum node size threshold (fewer splits; variance decreased)
    - Changing model approach
        - Bagging, boosting, & ensemble methods
    - Re-consider data splitting approach
        - Training + test?
        - LOOCV
        - K-fold CV

What to do about bias...

1. Anticipate and plan for potential biases before model generation. Check for bias after.
2. Have diverse teams.
3. Test test test! Test all possible situations and scenarios
4. Use machine learning to improve lives rather than for punitive purposes.
5. Revisit your models. Update your algorithms. Take feedback to improve the tech
    1. e.g. IR emitter-detector issues
6. You are responsible for the models you put out into the world, unintended consequences and all.

- Checklists are helpful, but they’re not and excuse for thoughtlessness.

- Ultimately you have to keep in mind that science and engineering are about ***increasing human knowledge and improving the human condition***

- Beware of de-humanizing people with technology

- Consider the big picture, take a step back periodically

Dimensionality Reduction

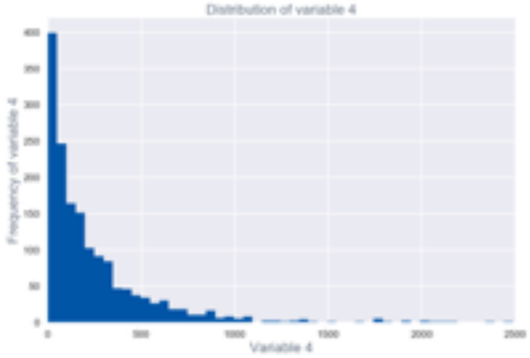
A mathematical process to reduce the number of random variables to consider

Discuss: why may we want to do this?

EDA Approaches to “Get a Feel for the Data”

Understanding the relationship between variables in your dataset

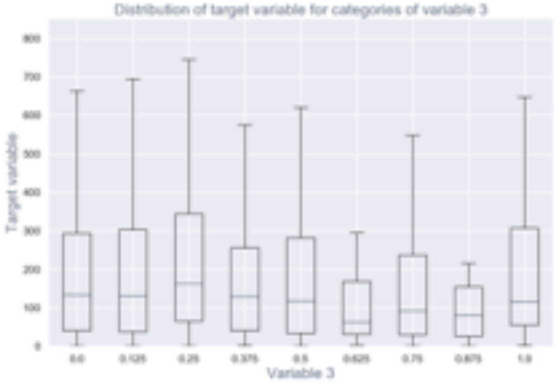
Exploratory



Univariate

understanding a single variable

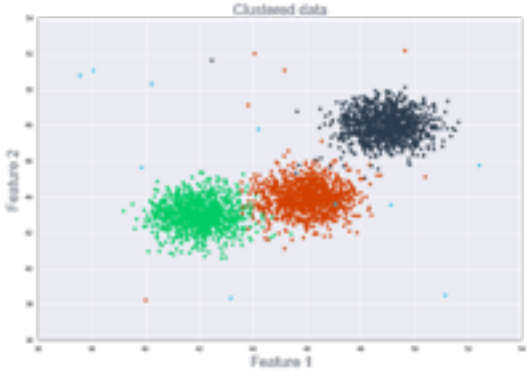
i.e.: histogram, densityplot, barplot



Bivariate

understanding relationship between 2 variables

i.e.: boxplot, scatterplot, grouped barplot, boxplot



Dimensionality Reduction

projecting high-D data into a lower-D space

i.e.: PCA, ICA, Clustering

Image source: https://www.svds.com/value-exploratory-data-analysis/

Major example methods

- **PCA** - (Linear) Find projections of the data into lower dimensional space that captures most of the variations in the data
- **ICA** - (Linear) Separate mixed additive independent signals into separate sources
- **CCA** - (Linear) Looks for relationships between two multivariate data sets
- **Clustering** - (Nonlinear) We have discussed this - uses machine learning to extract features from data

Image source: https://www.svds.com/value-exploratory-data-analysis/

Principal Component Analysis (PCA)

Key Terms:

- **Principal Component (PC)** - a linear combination of the predictor variables
- **Loadings** - the weights that transform the predictors into components (aka weights)
- **Screeplot** - variances of each component plotted

PCA : Key Ideas

1. PCs are linear combinations of the predictor variables (numeric data only)
2. Calculated to minimize correlation between components (minimizes redundancy)
3. A limited number of components will typically explain most of the variance in the outcome variable
4. Limited set of PCs can be used in place of original predictors (dimensionality reduction)

For more on PCA:

- https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/
- http://setosa.io/ev/principal-component-analysis/

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

Dimensionality Reduction with PCA: Pros & Cons

Pros:

- Helps compress data; reduced storage space.
- reduces computation time.
- helps remove redundant features (if any)
- Identifies outliers in the data

Cons:

- may lead to some amount of data loss.
- tends to find linear correlations between variables, which is sometimes undesirable.
- fails in cases where mean and covariance are not enough to define datasets.
- may not know how many principal components to keep
- highly affected by outliers in the data

https://www.geeksforgeeks.org/dimensionality-reduction/

Written Communication

Data Science Reports

1. In-depth details of analysis
2. Full Explanation (nothing extra)
3. A handful of figures (w/ interpretation)
4. Tell a Story

Final Project: Video

3% of Final Grade
3-5 minutes

All members must be involved but it's not required that all members speak or that members' faces are on video.

Can be a slideshow presentation w/ voiceover.
Can be something more creative. Has to effectively communicate your project.

Oral Communication

Overview

01:

Your Audience

02:

Storytelling

03:

The Grammar of Graphics

04:

The Glamour of Graphics

Your Audience

Consider your audience.

- General vs. technical?
- Audience background?
- Setting?

Your Audience

General

✔ background

🚫 limit technical details

🎉 emphasize take-home

Technical

⬇ limit background

💻 all-the-details

🎉 emphasize take-home

Your Audience

Presentation: General Audience

Introduction & Background

- Details on background material
- Full explanation of question and *why* doing analysis

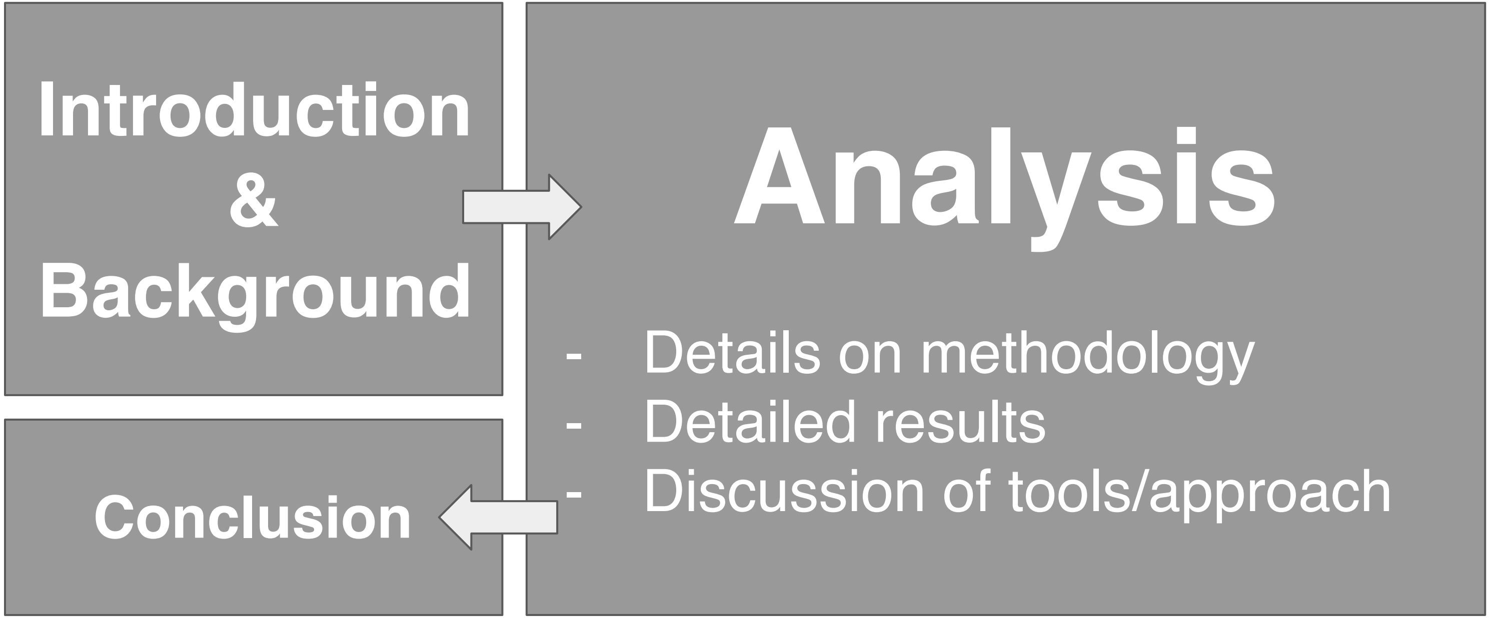
Analysis

- Limited discussion of approach
- Focused presentation of results

Conclusion

Your Audience

Presentation: Technical Audience



02:

Storytelling

Storytelling: Ground Rules

1. Enticing, short title
2. Clear presentation
3. All the necessary info
4. Nothing extra

On your slides...

- Limit number of ideas
- Limit words
- Choose good fonts
- Make text readable
- Include references

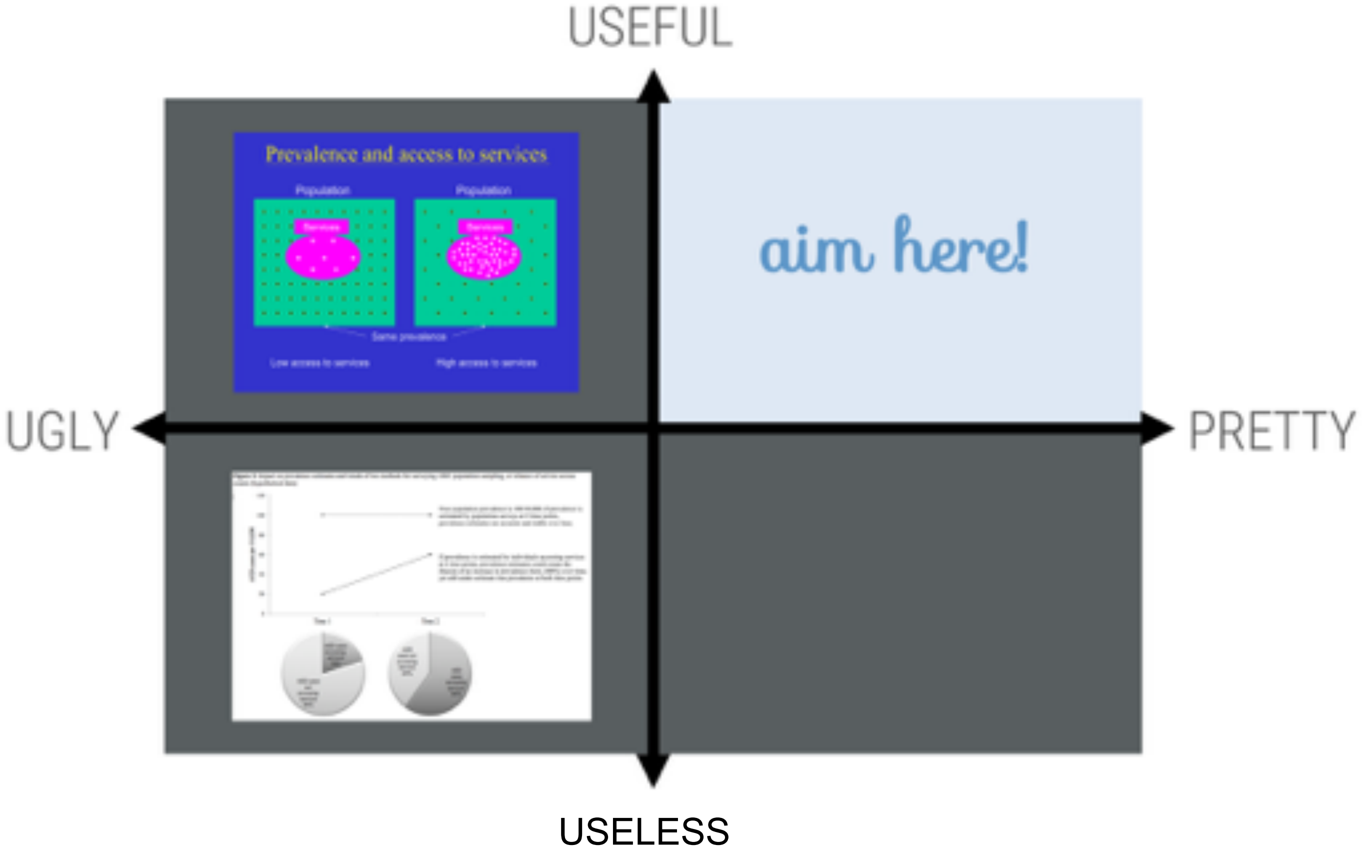
Horse

Fonts matter

MEGAFLICKS

Fast Taco

Iteration



Source: Jackie Wirz

Your Audience

Presentations: for listening

- don’t read directly off slides
- use animation to build your story (not to distract)
- introduce your axes
- benefit: words to explain out loud what you’re showing

Reports: for reading

- more on a single visualization
- explanation must be there in text
- Benefit: people have time to look at what you’ve sent

Different types of applications of data science

- **Jobs**
  - Large company
  - Small company
- **Academic**,
  - PhD - professor, scientist, staff, lecturer, research professor, teaching professor
  - MS - Staff scientist (higher rank), lecturer (some programs)
  - BS - Staff scientist (initially lower rank)
- **Entrepreneurial**
  - Start a company
  - Work with startups
  - Consult
- **Creative**
  - Art and entertainment
  - Writing, content creation

Data Science Jobs

Personas

Finding Your Job

Minimal Advice

Build A Career In Data Science (2023)

Part I: Getting Started With Data Science

- What is Data Science?
- Data Science Companies
- Getting the Skills
- Building a Portfolio

Part II: Finding Your Data Science Job

- The Search: Identifying the Right Job for You
- The Application: Resumes and Cover Letters
- The Interview: What to Expect and How to Handle It
- The Offer: Knowing What to Accept

Part III: Settling Into Data Science

- The First Months on the Job
- Making an Effective Analysis
- Deploying a model into production
- Working with Stakeholders

Part IV: Growing In Your Data Science Role

- When your Data Science Project Fails
- Joining the Data Science Community
- Leaving Your Job Gracefully
- Moving up the Ladder

A job by many names...

Data analyst
entry level
Analyze data & create reports

Product analyst
job varies
Focuses on one part of the company

ML engineer
software focused
Build ML models to power the business

Research scientist
theoretical
Research focused job, requires advanced degree

2. The Application

- Resume *and* cover letter should be compelling
    - <u>Resume</u>:
        - goal is to get you an interview, not a job
        - Better be skimmable
        - Includes: contact info, education, experience, and skills
    - <u>Cover Letter</u>
        - Should highlight both why you want *this* job and why *you* are a particularly good fit
        - Demonstrate your research/knowledge about the company and position
    - Tailor these for each job
    - Allow them to be machine-searchable
    - Referrals are a way to back-door past the algorithms
        - If contacting someone (LinkedIn, Twitter), give them a reason to read your message

SARA JONES

New York, NY · 534-241-6264

sarajones@gmail.com · linkedin.com/in/sarajones · sarajones.github.io · github.com/sarajones

GREETING

Dear Jared,

INTRODUCTORY PARAGRAPH

I am writing to express my strong interest in applying for the Data Scientist position at Awesome Company. I've enjoyed reading Awesome Company's data science blog since it started 8 months ago. The post on using topic modeling to automatically generate tags for your support articles was immensely helpful in one of my own projects to classify articles in the New York Times business section.

1-2 PARAGRAPHS OF DATA SCIENCE WORK EXAMPLES

I recently graduated from Awesome Bootcamp, a full-time, 3-month Data Science immersive. At Awesome Bootcamp, I designed, implemented, and delivered data science projects in Python involving data acquisition, data wrangling, machine learning, and data visualization. For my final project, I gathered 3,000 neighborhood reviews and ratings from Neighborhood Company. By using natural language processing on the reviews and available listings from Real Estate Company's API, I built a recommendation system that will match you to a neighborhood based on your budget, preferences, and a free-text description of your ideal neighborhood. You can try it out here: myawesomewebapp.com.

Prior to Awesome Bootcamp, I was an Investment Consultant at BigCo. When I joined, my team of six was all using Excel. While exceeding my targets, I began automating common tasks in Python, such as generating a weekly market and industry trends report, saving the team hours each week. I then developed a tailored curriculum to teach them Python. The initiative was so successful the company asked me to develop a full 2-day workshop and flew me out to three other offices to teach it, reaching over 70 consultants.

CLOSING PARAGRAPH

I am confident that my expertise in Python, academic training in Economics and Statistics, and experience delivering business results would make me a great fit for the Data Science team. Thank you for your consideration.

SIGNOFF

Sincerely,
Sara Jones

Figure 6.2 An example cover letter with highlights showing the different components

3. The Interview

Basic understanding of you/position; assessment of fit	1. Phone interview
Are you able to do the job? Are you a good fit?	2. In-person interview
Take-home assignment to determine your problem-solving and technical skills	3. Case study
Tie up loose ends, presentation,	4. Leadership interview and offer

4. The Offer

1. Offer is coming - general outline of offer coming your way
2. Company makes an offer - often in email; get it in writing; includes salary, start date,
3. You respond - thank them for offer and let them know you're excited to look it over in detail
4. You negotiate - lay out what you want/need to accept the offer; best for you
    a. What is negotiable? Salary (5%), start date, vacation, flexibility, earlier review (earlier raise), educational benefits, budget for travel/conferences, benefits (less often), options
    b. Best lever: a competing offer
5. You decide - communicate final decision

Getting Your First Job in Data Science in Summary:

- Learn one programming language extraordinarily well (Python, R).
- Learn SQL extraordinarily well.
- Learn how to set up and interact with cloud computing services.
- Know how to *think* and *communicate* about data

- Create a resume and have a few people with relevant knowledge help you revise it.
- Establish a professional web presence.
- Be prepared to apply to many dozens of jobs.

Break time…