Approaches to Analysis

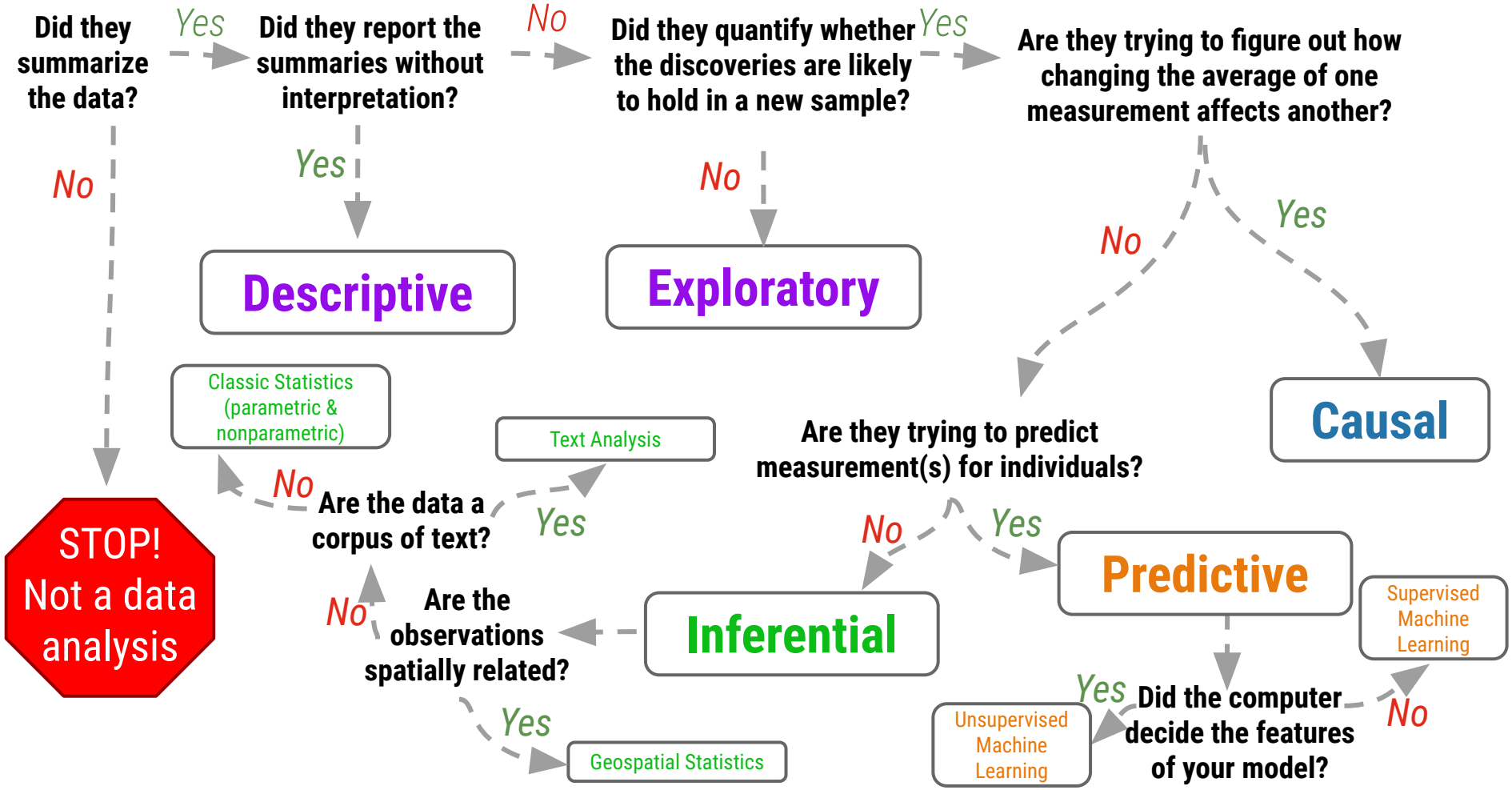
Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

*"Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, **analyzing the data**, and communicating the answer to the question to a relevant audience."*

To do this, you have to *look at, describe, and explore* the data

https://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-that-is-a-big-problem-for-data-scientists/


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
Causal
No
Descriptive
Yes
No
Exploratory
Yes
STOP! Not a data analysis
No
Classic Statistics (parametric & nonparametric)
Are the data a corpus of text?
Yes
Text Analysis
Are they trying to predict measurement(s) for individuals?
No
Yes
Predictive
Inferential
No
Are the observations spatially related?
Yes
Geospatial Statistics
Supervised Machine Learning
Yes
Unsupervised Machine Learning
Did the computer decide the features of your model?
No

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 causality

Exploring Analyses

General question: What impacts politics in America?

Data Science question: Is there a relationship between the sentiment of political words in South Park and America's presidential approval rating?

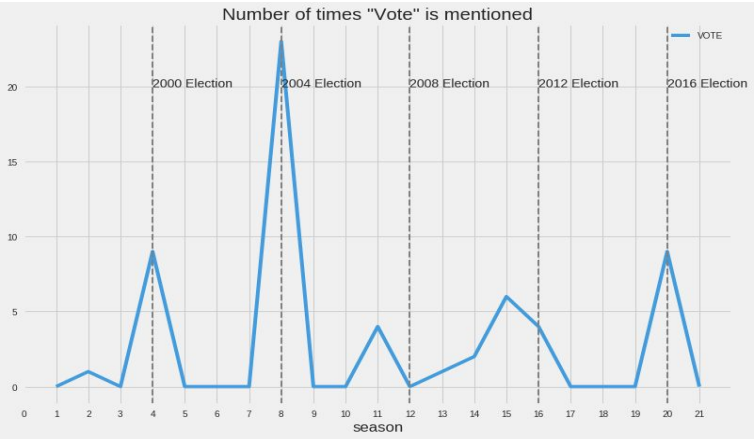
Descriptive

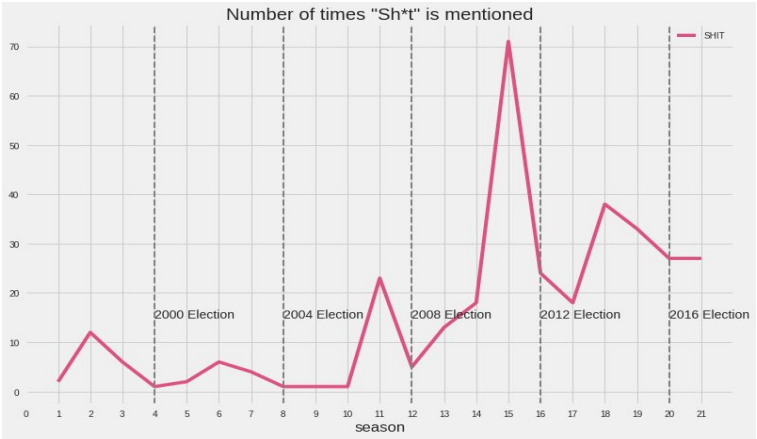
Exploratory

Inferential

Text Analysis

Classic Statistics (parametric & nonparametric)

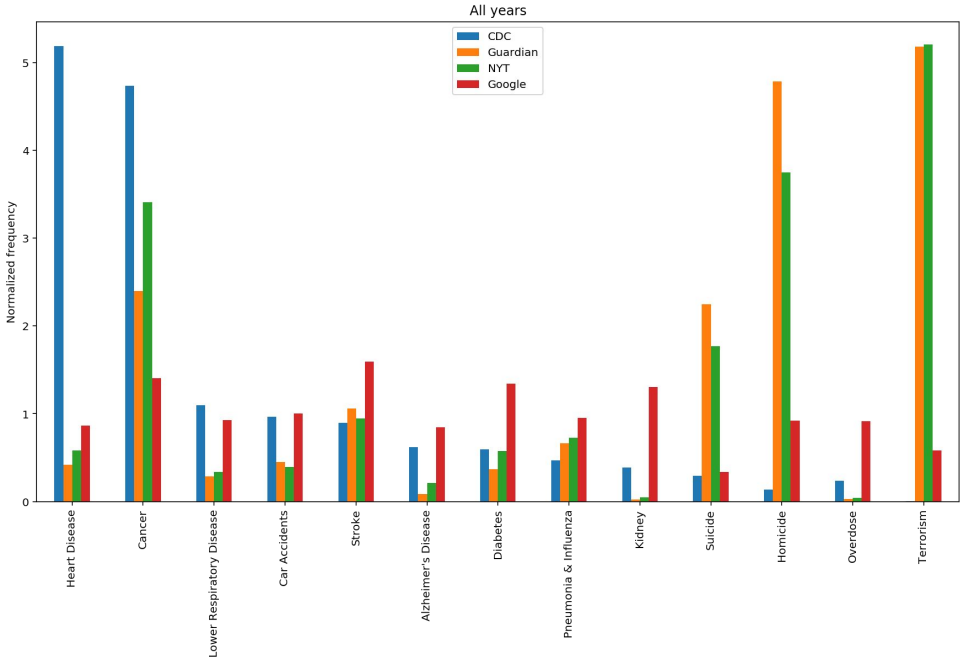




https://github.com/COGS108/FinalProjects-Wi18/blob/master/071-FinalProject.ipynb

General question: What gets too much attention in the news?

Data Science Question: Is there a relationship over time between cause of death terms in the *NYT*, The Guardian, and Google trends data relative to data from the CDC?



Descriptive

Exploratory

Inferential

Text Analysis

Classic Statistics (parametric & nonparametric)

https://github.com/COGS108/FinalProjects-Wi18/blob/master/010-FinalProject.ipynb

Descriptive

Exploratory

Predictive

Classification: Often we seek to assign a label to an item from a discrete set of possibilities.

Can we predict who will win next year's NCAA tournament? The Masters? The Super Bowl? The pennant? A game?

Can we predict the genre of a given movie (comedy, drama, or animation?) from just its script?

Descriptive

Exploratory

Predictive

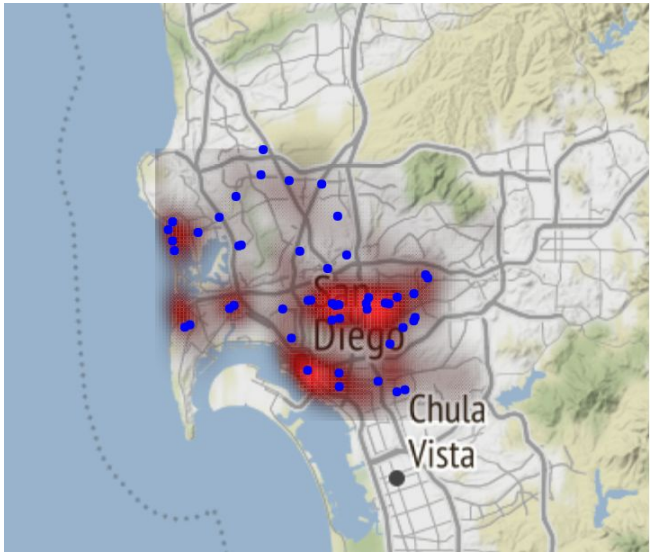
Regression: A way to forecast a given numerical quantity using other relevant features.

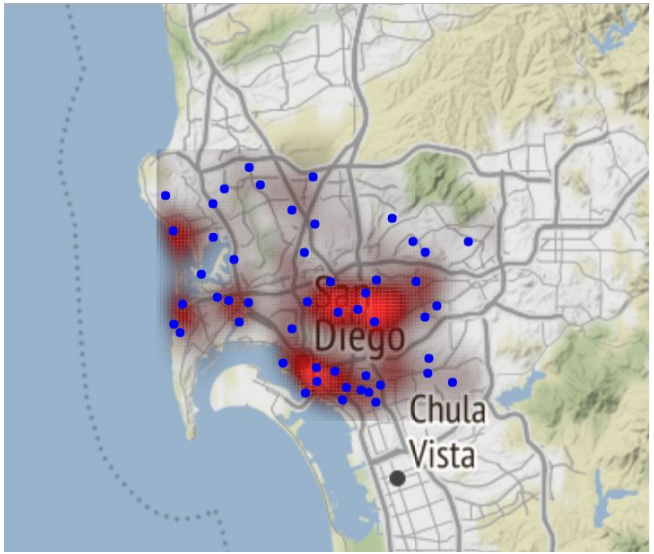
Can we predict someone's weight given other information?

How much snow will the East Coast get this year?

General question: Why isn't police response time always the same?

Data Science question: Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable throughout San Diego?





Descriptive

Exploratory

Predictive

Inferential

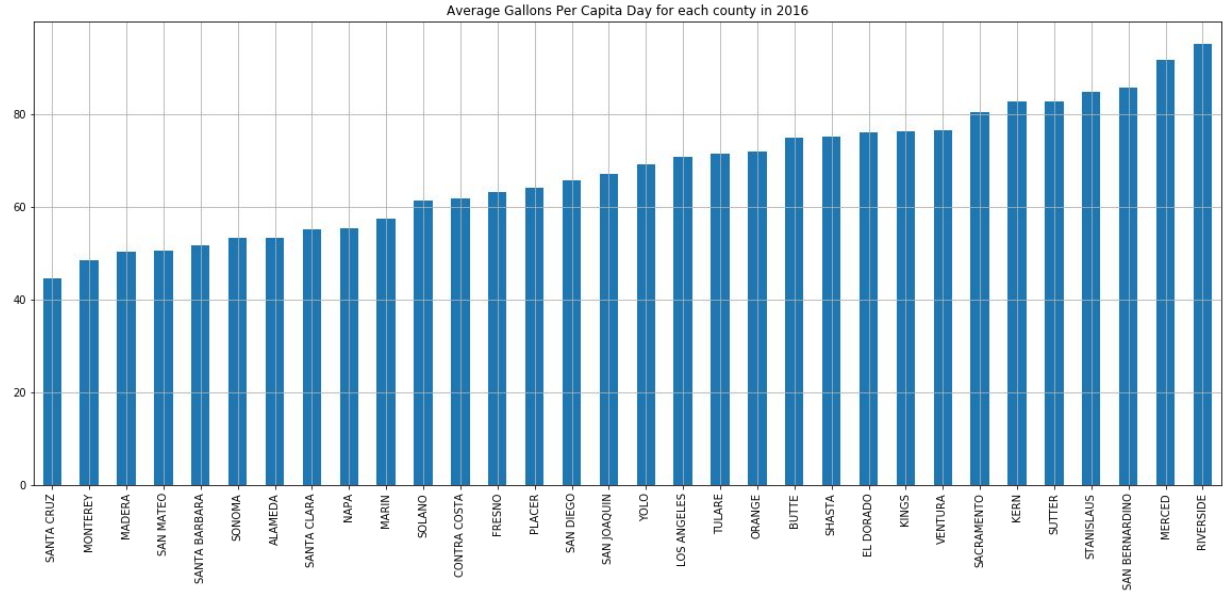
https://github.com/COGS108/FinalProjects-Sp17/blob/master/029-FinalProject.ipynb

In case of the total drought in California, how many desalination plant projects we need to supply residential use water for population who live in urban areas in California?

Descriptive

Exploratory

Predictive



https://github.com/COGS108/FinalProjects-Sp17/blob/master/018-FinalProject.ipynb