

Descriptive and Exploratory Analysis

Jason G. Fleischer, Ph.D.

Asst. Teaching Professor

Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu

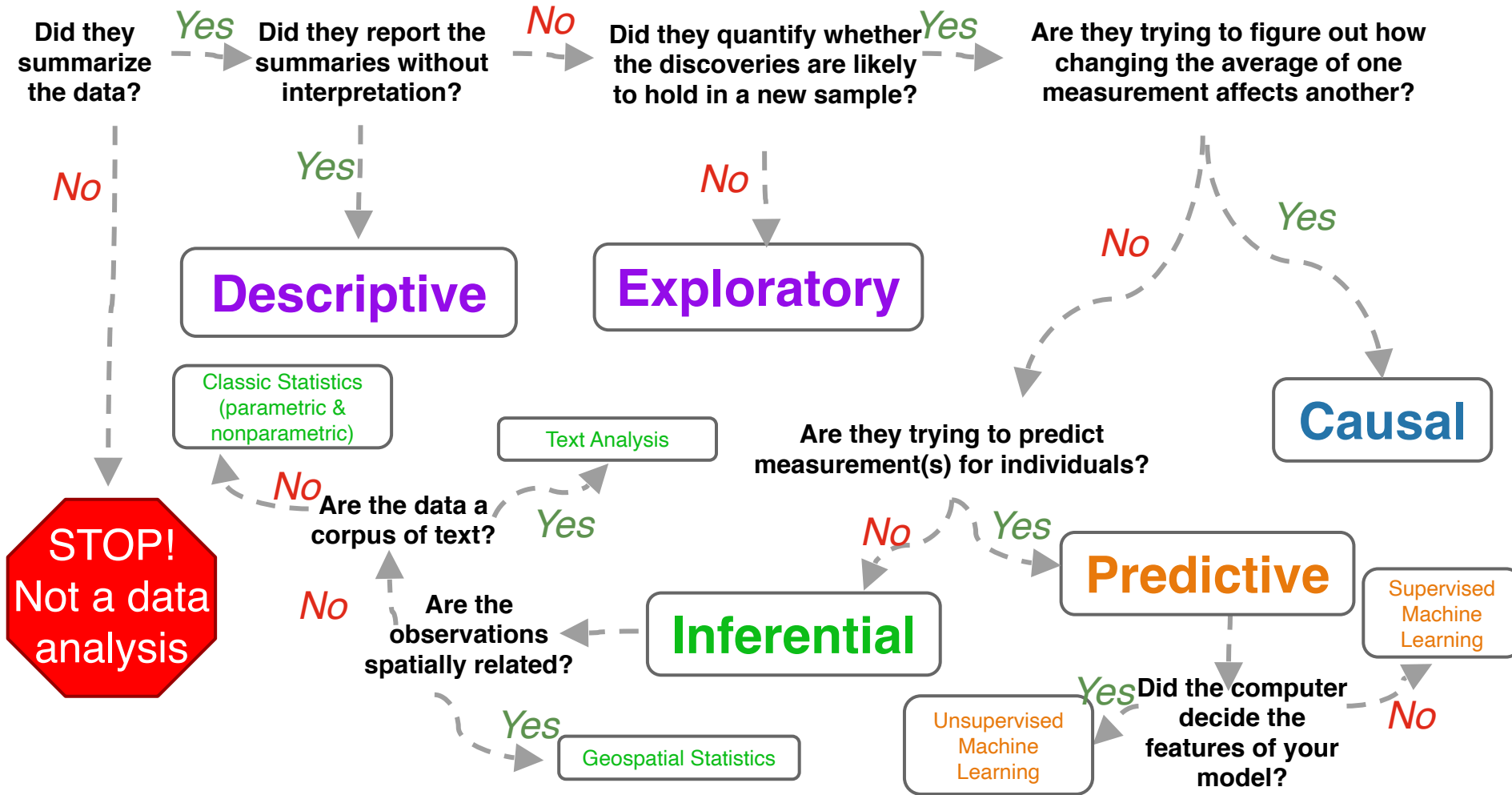


@jasongfleischer

<https://jgfleischer.com>

“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*



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?

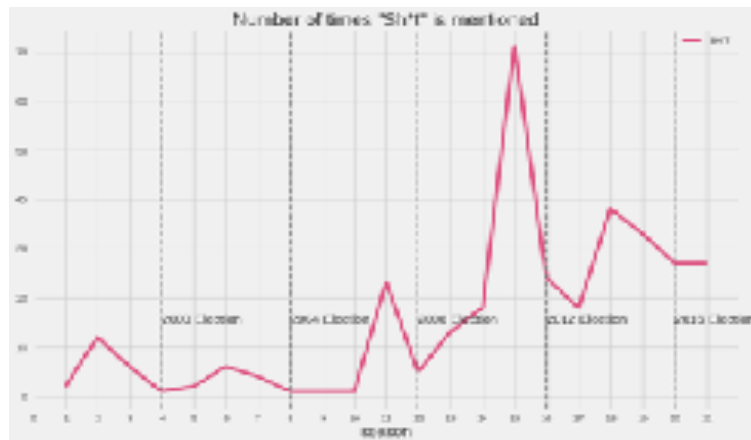
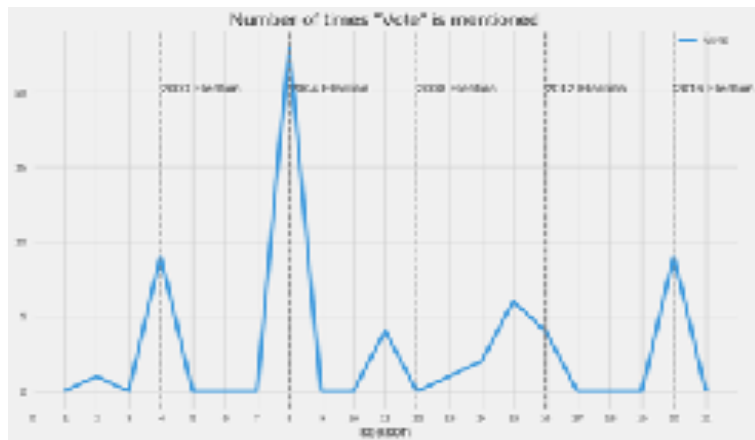
Descriptive

Exploratory

Inferential

Text Analysis

Classic Statistics
(parametric &
nonparametric)



General question: How has COVID-19 impacted students?

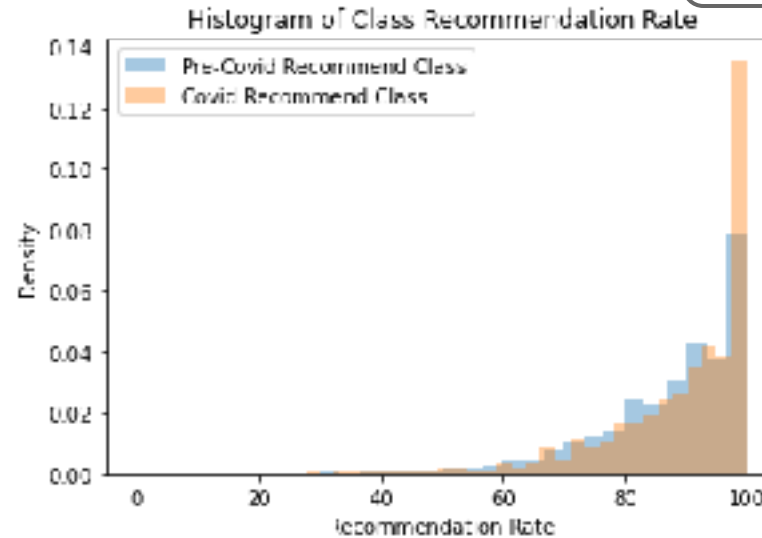
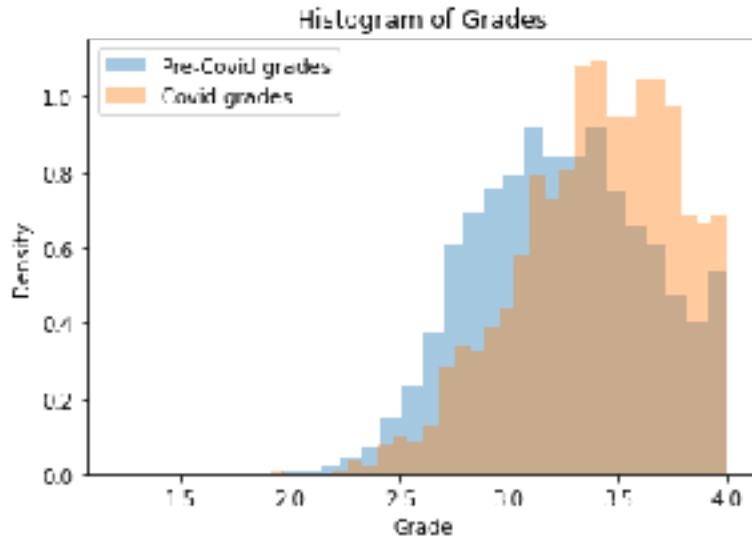
Data Science question: At UCSD, is there a difference between students' grades and how they rate their classes before COVID-19 and during remote learning, due to COVID-19?

Descriptive

Exploratory

Inferential

Classic Statistics
(parametric &
nonparametric)



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?

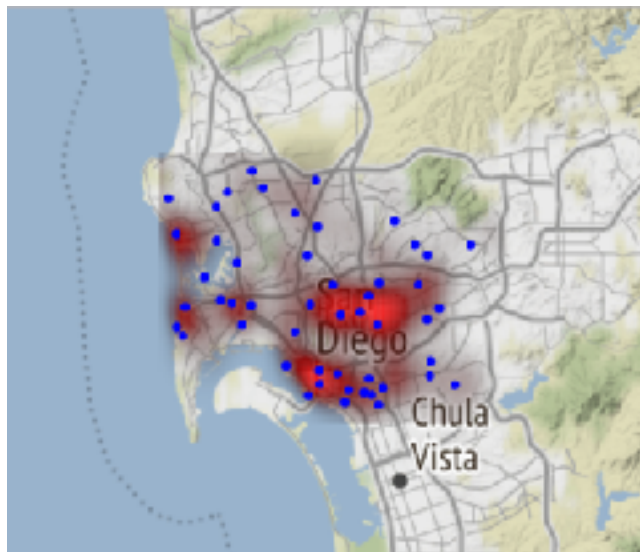
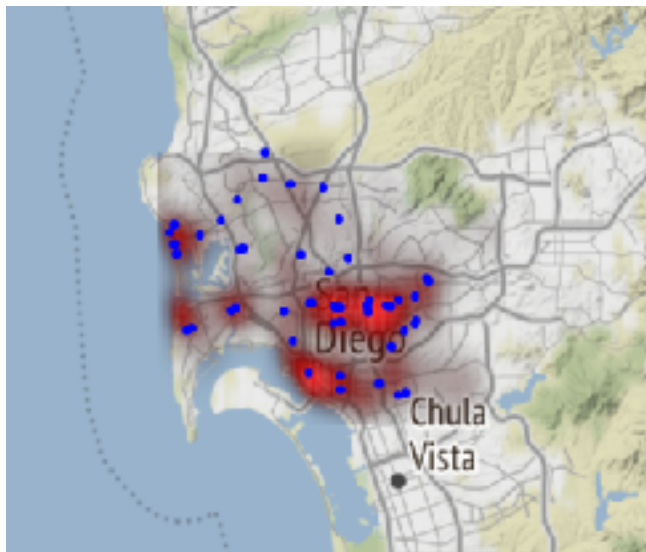
Descriptive

Exploratory

Predictive

Inferential

Geospatial Analysis



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?

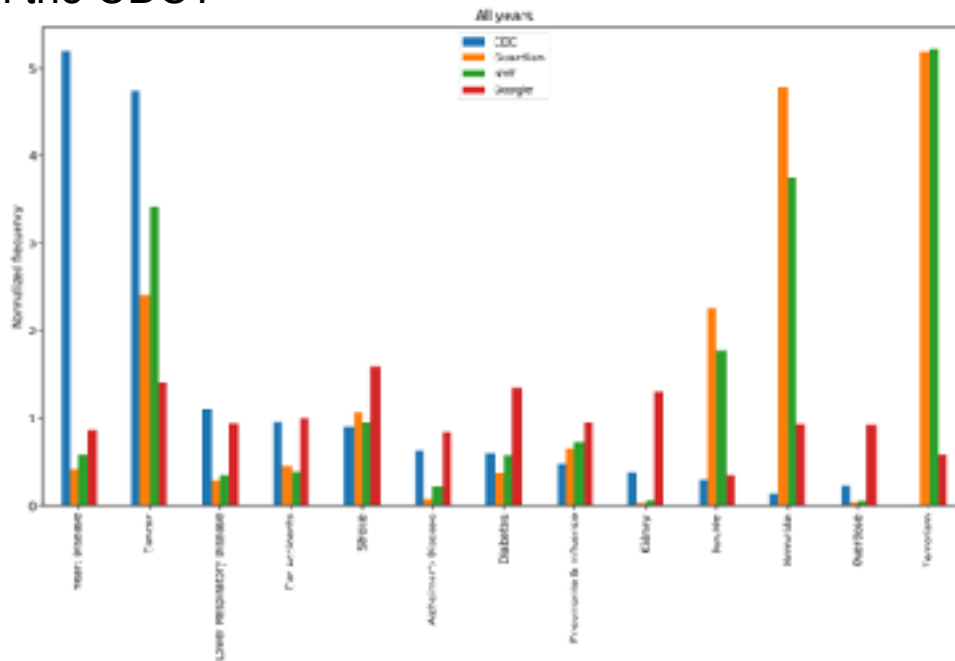
Descriptive

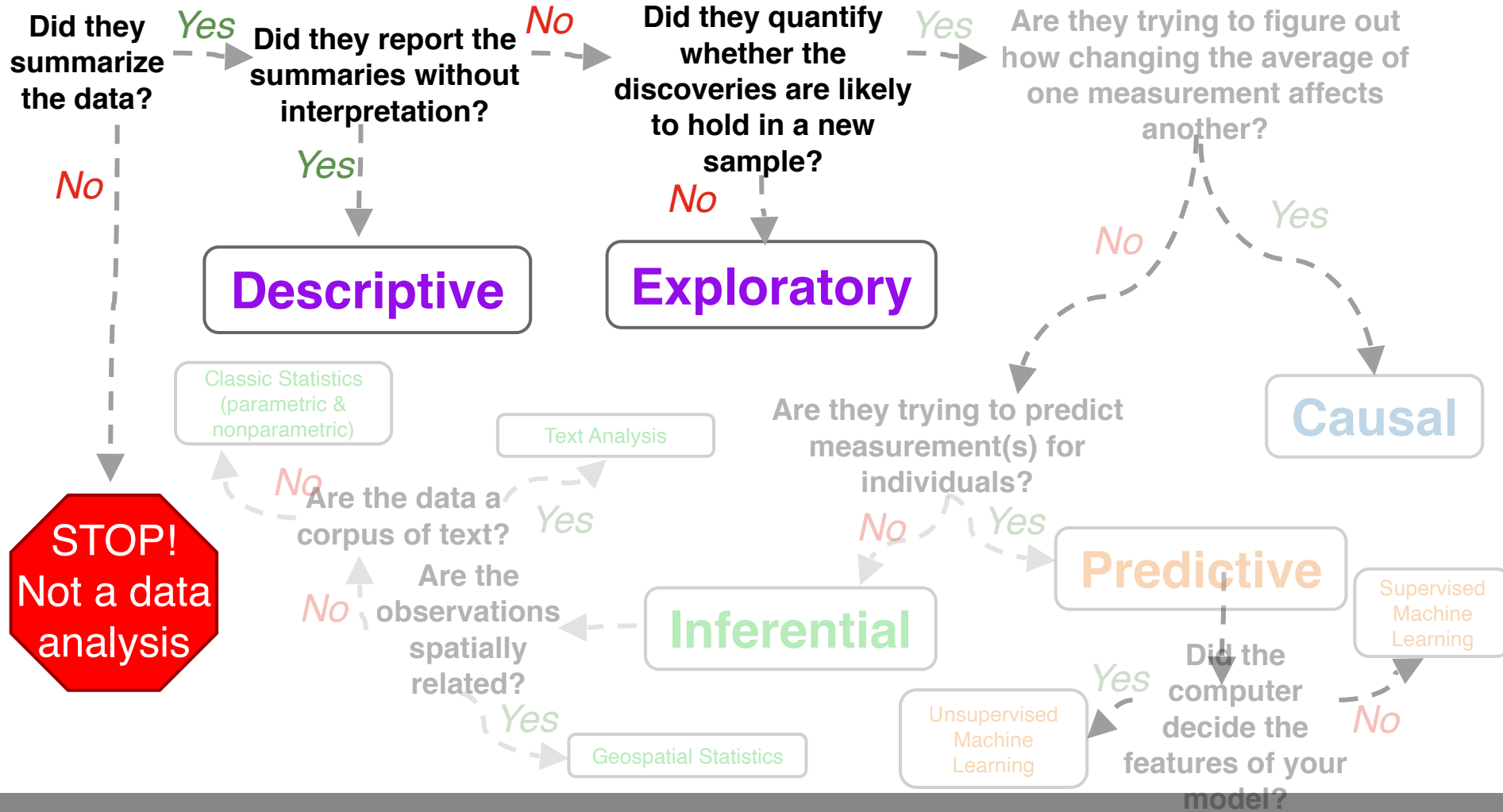
Exploratory

Inferential

Text Analysis

Classic Statistics
(parametric &
nonparametric)





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.

- Problem: Understanding whether users are nice or mean on Youtube
- Data science question: Are the words that people use in their comments more frequently positive words (great, awesome, nice, useful) or negative words (bad, stupid, lame, awful)?
- Type of analysis: Descriptive analysis

To answer this you would calculate statistics about YouTube comments



Statistics

*“the science that deals with the **collection, classification, analysis, and interpretation of numerical facts or data**”*

statistic

“A quantity computed from a sample”

statistic

“A quantity computed from a sample”



For our YouTube analysis, we could take a random sample of comments from YouTube and calculate the following statistic: *the number of positive and the number of negative words in each review.*

Population

All comments on YouTube

During the second quarter of 2020, almost 2.13 billion comments on YouTube videos were removed due to violation of the platform's community guidelines. - J Clement on

We want to learn something about this...

Sampling

Inference

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

Sample

1 million
comments from 2020

Best sampling practices:

- Always think about what your population is
- Collect data from a sample that is representative of your population
- If you have no choice but to work with a dataset that is not collected randomly and is biased, be careful not to generalize your results to the entire population



You'd want to be sure you sample randomly across *all* YouTube comments, making sure not to get more comments from one genre over another, or one location over another, etc.

Examples of bad sampling:

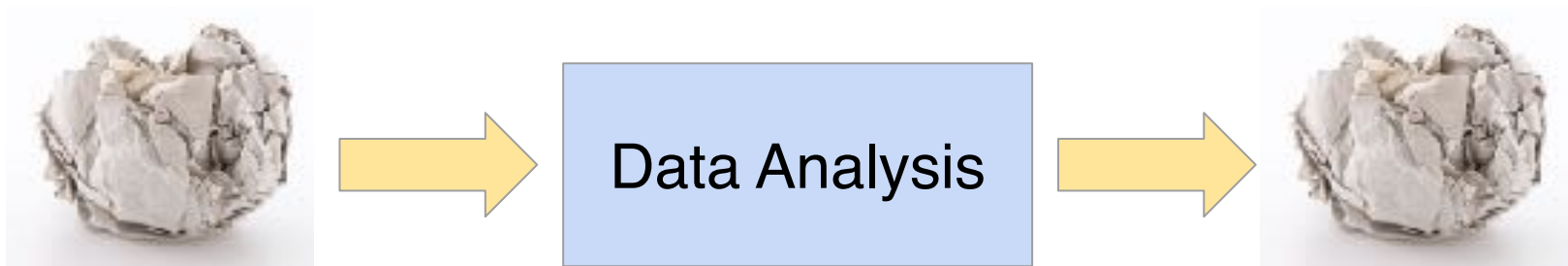
- Surveying subscribers of a gun-related magazine for research on Americans' attitudes toward owning guns
- Randomly sampling Facebook users for what TV shows people like



To understand *all* YouTube comments, you wouldn't just want to sample from one YouTube channel, or videos in a single language.


It's *always* worth spending time at the beginning 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.

GIGO : Garbage In. Garbage Out.





For the survey data I collected from you all, which of the following best describes the population I could generalize findings back to.

- 
- A** Undergraduates
 - B** Undergraduates in the US
 - C** Undergraduates at UCSD
 - D** Students aged 18-25
 - E** UCSD COGS108 students

Descriptive

Descriptive Analysis



Size



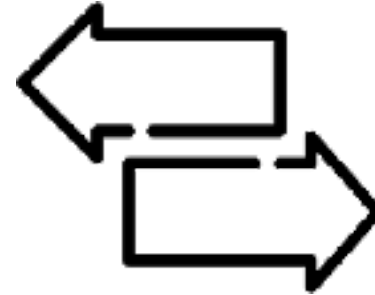
Missingness



Shape



Central
Tendency



Variability



Size

How many observations (rows) and variables (columns) you have is an important first step. You should always be aware of the **size** of your dataset



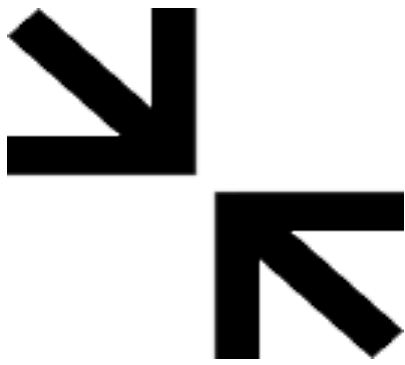
Descriptive

Missingness It's critical to know how many observations have missing data for variables of interest in your data. Knowing *why* their missing is also important.



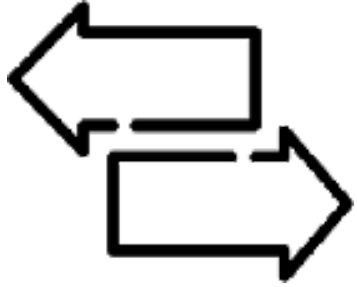
Shape

It's critical to know the distribution of the variables in your dataset. Certain statistical approaches can only be used with certain distributions.



Central Tendency

Knowing the mean, median, and/or mode can help you get an idea of what a typical value is for your variable(s) of interest



Variability

The central tendency tells you part of the story. The **variability in the values** in your observation helps fill in the rest.



Which of the following is NOT something accomplished by a descriptive analysis?

- ☐ **A** Describes typical values in your dataset
- ☐ **B** Determines the size of your dataset
- ☐ **C** Establishes causal relationships between variables
- ☐ **D** Identifies missing data
- ☐ **E** Determines how variable values in your dataset are

Descriptive Statistics & Summary

“We must suppress some of the truth to communicate the truth... In short, the techniques of descriptive statistics are designed to match the salient features of the data set to human cognitive abilities.”

-I.J. Good (1983)

Descriptive Analyses are often included as “Table 1” in academic publications

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Cardiovascular Morbidity (N = 305)	Neurovascular Morbidity (N = 358)	Ischemic Stroke (N = 150)	Nonischemic Stroke (N = 198)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (4.0)	2 (1.0)
60–69 yr	12 (4.0)	28 (8.0)	61 (40.6)	34 (17.2)
70–79 yr	302 (99.3)	329 (92.0)	115 (76.4)	133 (66.8)
80–89 yr	142 (46.2)	118 (33.4)	126 (83.0)	142 (71.0)
≥90 yr	22 (7.1)	23 (6.4)	26 (16.9)	18 (9.0)
Mean ± SD	79.2±7.4	80.1±7.3	78.4±7.3	79.3±7.4
Sex — no. (%)				
Female	163 (53.5)	138 (38.5)	135 (89.9)	184 (92.9)
Male	115 (37.5)	120 (33.5)	11 (7.1)	14 (7.1)
Race — no. (%)†				
White	297 (97.2)	331 (92.5)	136 (89.9)	154 (77.8)
Other	4 (1.3)	5 (1.4)	3 (1.9)	4 (2.0)
History of myocardial infarction — no. (%)	16 (5.3)	12 (3.4)	42 (27.9)	38 (19.2)
History of stroke — no. (%)	16 (5.3)	15 (4.2)	22 (14.7)	18 (9.1)
History of transient ischemic attack — no. (%)	12 (4.0)	15 (4.2)	12 (7.9)	15 (7.6)
Blood pressure — mm Hg				
Systolic	114±18	115±19	116±17	115±17
Diastolic	75±10	73±10	76±8	75±10
Visual acuity (normal and better) — no. (%)				
60–69 letters, 20/20–40 — no. (%)	111 (36.4)	94 (26.3)	116 (77.3)	133 (67.2)
53–59 letters, 20/50–40 — no. (%)	58 (18.7)	118 (33.0)	108 (71.3)	115 (58.1)
46–52 letters, 20/100–160 — no. (%)	67 (21.6)	13 (3.6)	58 (38.7)	58 (29.3)
23–47 letters, 20/320–120 — no. (%)	25 (8.1)	21 (5.9)	16 (10.6)	26 (13.0)
Mean score	60.1±14.7	58.3±15.1	61.5±11.2	60.4±11.4
Total cholesterol if fasting — mg/dL	418±114	483±136	458±115	461±115
Estimated creatinine plus calculated creatinine at baseline — μmol/L	211±121	254±132	247±112	252±115
Formal cardiac investigation — no. (%)				
Coronary atherosclerosis	176 (57.7)	179 (50.0)	176 (117.3)	180 (90.4)
Stroke	85 (27.9)	81 (22.6)	77 (51.3)	72 (36.4)
Heart failure	19 (6.2)	24 (6.7)	24 (15.9)	25 (12.6)
Other	15 (4.9)	23 (6.4)	15 (9.9)	18 (9.1)
No coronary atherosclerosis or not possible to give	2 (0.7)	8 (2.2)	6 (3.9)	2 (1.0)

* Percentages reflect age means ± SD.
† Race was self-reported.
‡ Total cholesterol of the serum includes the total cholesterol, high-density lipoprotein, and low-density lipoprotein cholesterol.

Descriptive

Size

Zooming in on this we see variables stratified by Age, Sex, and Race

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Ranibizumab Monthly (N = 301)	Bevacizumab Monthly (N = 286)	Ranibizumab as Needed (N = 258)	Bevacizumab as Needed (N = 300)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (2.0)	2 (0.7)
60–69 yr	33 (11.0)	28 (9.8)	31 (10.4)	34 (11.3)
70–79 yr	102 (33.9)	84 (29.4)	115 (38.6)	103 (34.3)
80–89 yr	142 (47.2)	150 (52.4)	126 (42.3)	142 (47.3)
≥90 yr	22 (7.3)	23 (8.0)	20 (6.7)	19 (6.3)
Mean — yr	79.2 ± 7.4	80.1 ± 7.3	78.4 ± 7.8	79.3 ± 7.6
Sex — no. (%)				
Female	183 (60.8)	180 (62.9)	185 (62.1)	184 (61.3)
Male	118 (39.2)	106 (37.1)	113 (37.9)	116 (38.7)
Race — no. (%)†				
White	297 (98.7)	281 (98.3)	296 (99.3)	294 (98.0)
Other	4 (1.3)	5 (1.7)	2 (0.7)	6 (2.0)

* Plus-minus values are means ± SD.

† Race was self-reported.

‡ Total thickness at the fovea includes the retina, subretinal fluid, choroidal neovascularization, and retinal pigment epithelial elevation.

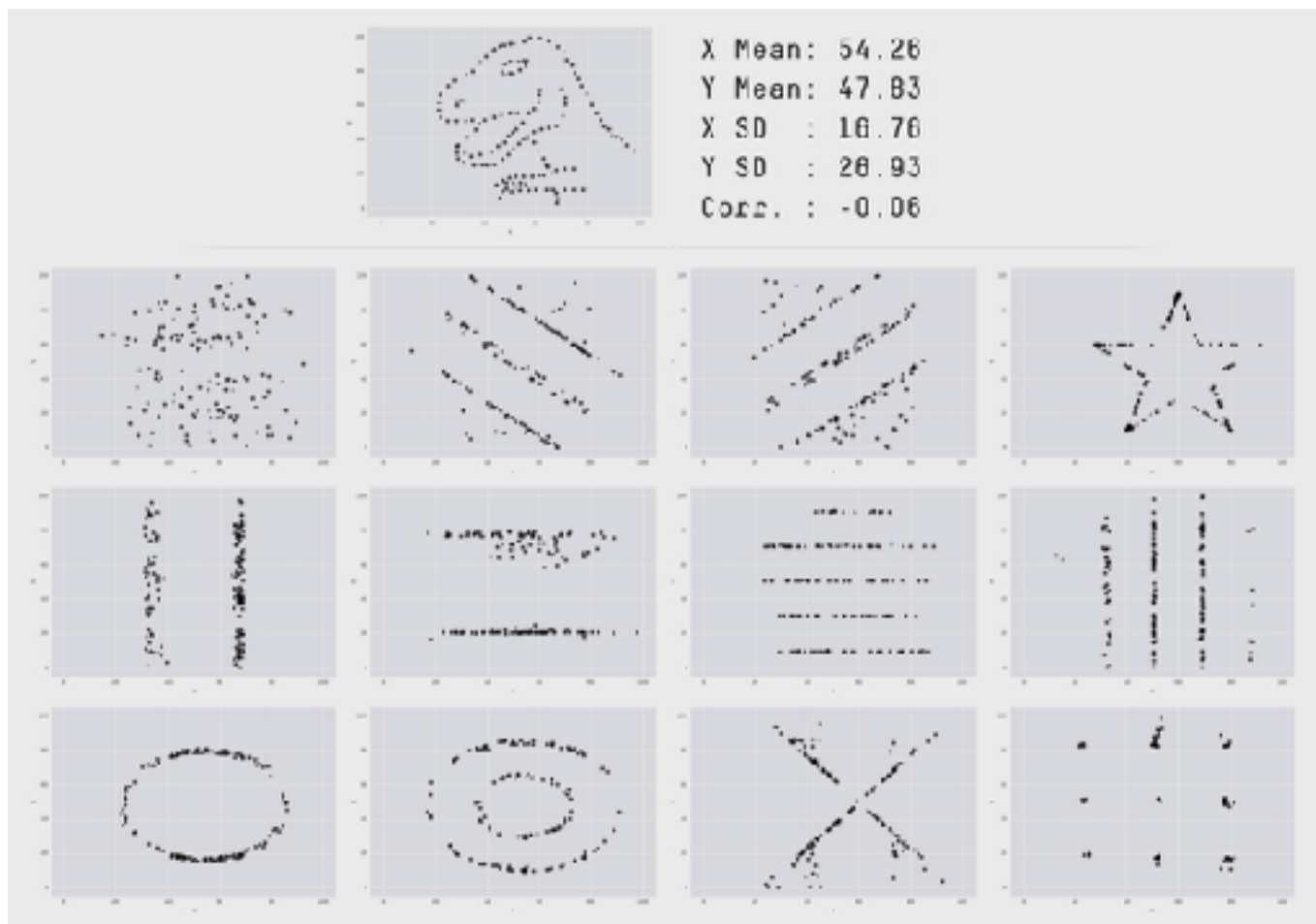
Shape

Central tendency

variability

Descriptive Statistics & Summary

Calculating descriptive statistics, understanding what they tell you about your data, and reporting them are critical steps in every analysis.



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?”

Why EDA?

- Understand data properties
- Discover Patterns
- Generate & Frame Hypothesis
- Suggest modeling strategies
- Check assumptions (sanity checks)
- Communicate results (present the data)

.....and if you don't, you'll regret it

The
dataset

You

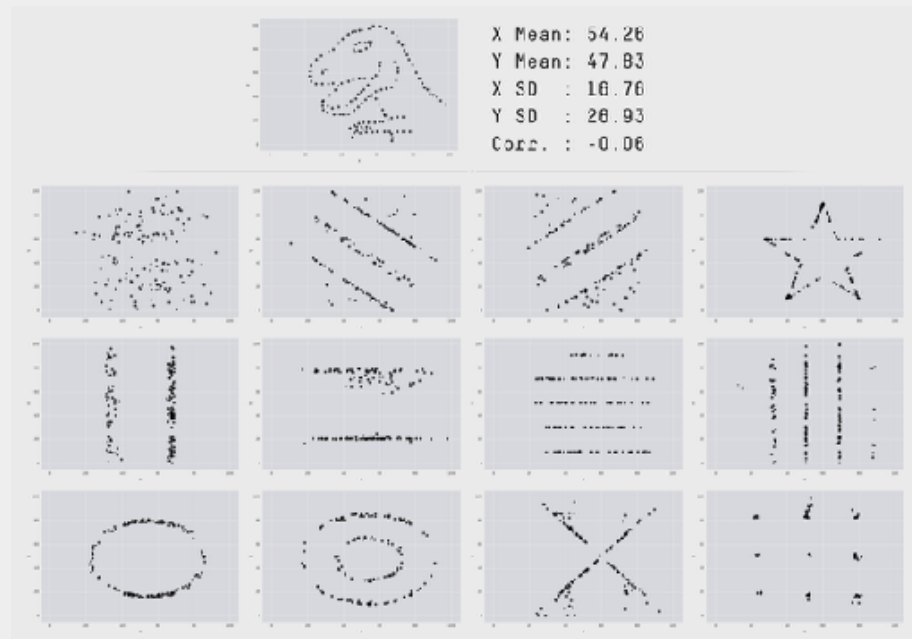


The general principles of exploratory analysis:

- 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

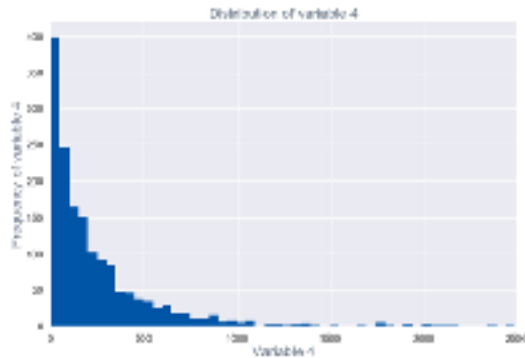
Start raw

- Examine raw data in the most direct way you can reasonably do so
- View a random sample of the data
- Plots, especially subsets of variables and dimensionality reduction
- Helpful for seeing weirdness, missingness, outliers, min/max/typical values



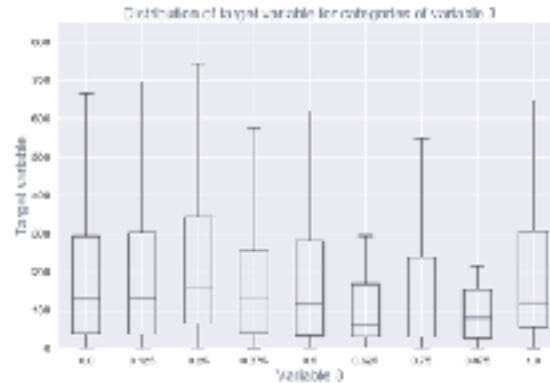
EDA Approaches to “Get a Feel for the Data”

Understanding the relationship between variables in your dataset



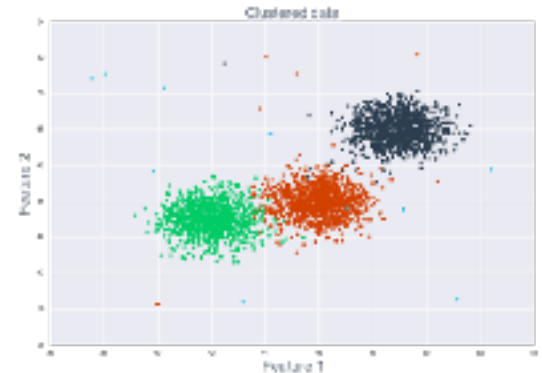
Univariate

understanding a single variable
i.e.: histogram, densityplot, barplot



Bivariate

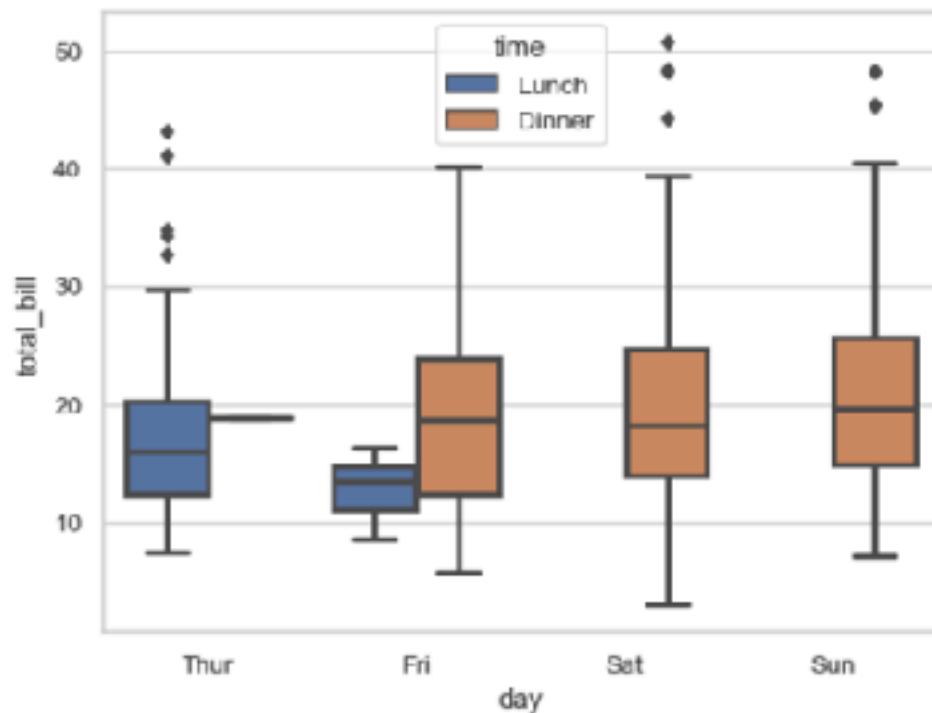
understanding relationship between 2 variables
i.e.: boxplot, scatterplot, grouped barplot, boxplot



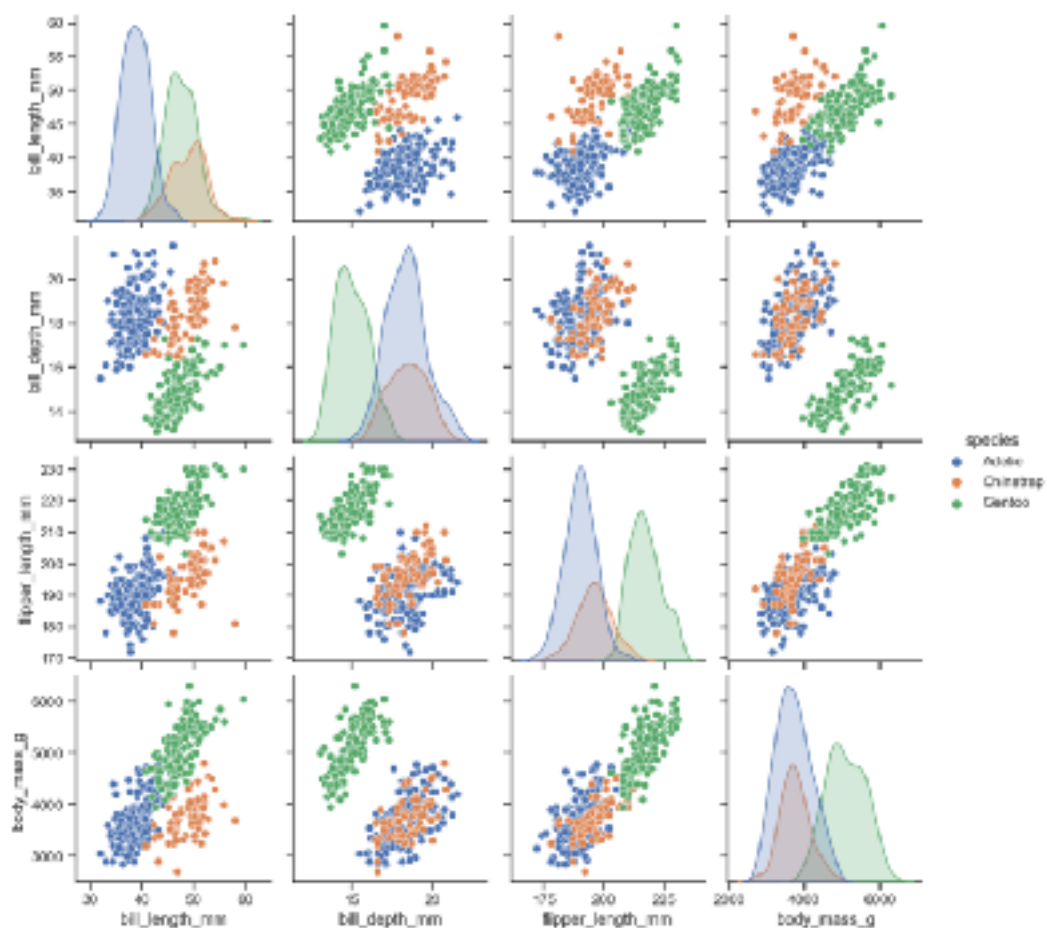
Dimensionality Reduction

projecting high-D data into a lower-D space
i.e.: PCA, ICA, Clustering

```
>>> ax = sns.boxplot(x="day", y="total_bill", hue="time",  
...                  data=tips, linewidth=2.5)
```

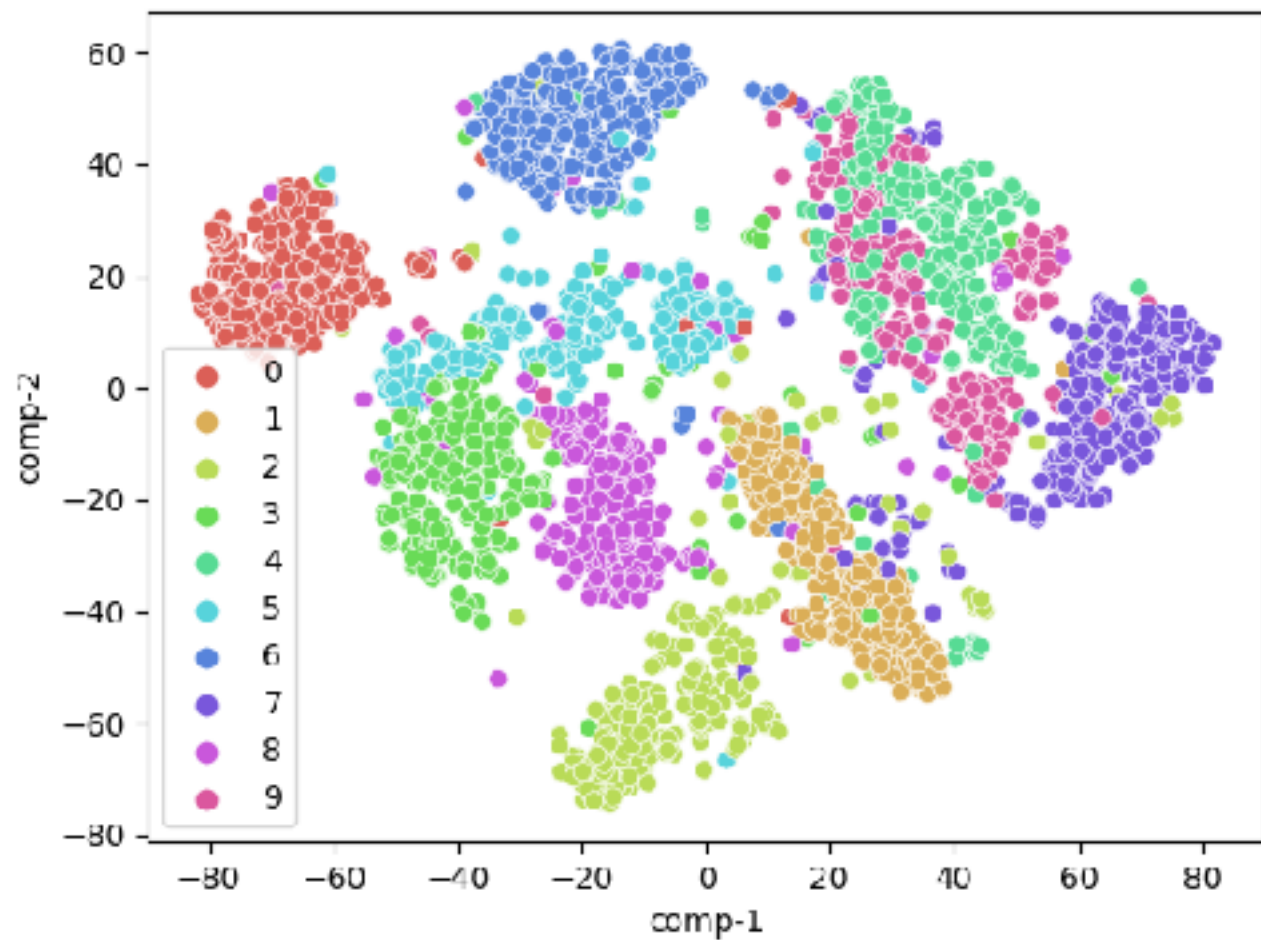



```
sns.pairplot(penguins, hue='species')
```





MNIST data T-SNE projection



1. Hyperparameters really matter
2. Cluster sizes in a UMAP plot mean nothing
3. Distances between clusters might not mean anything
4. Random noise doesn't always look random.
5. You may need more than one plot

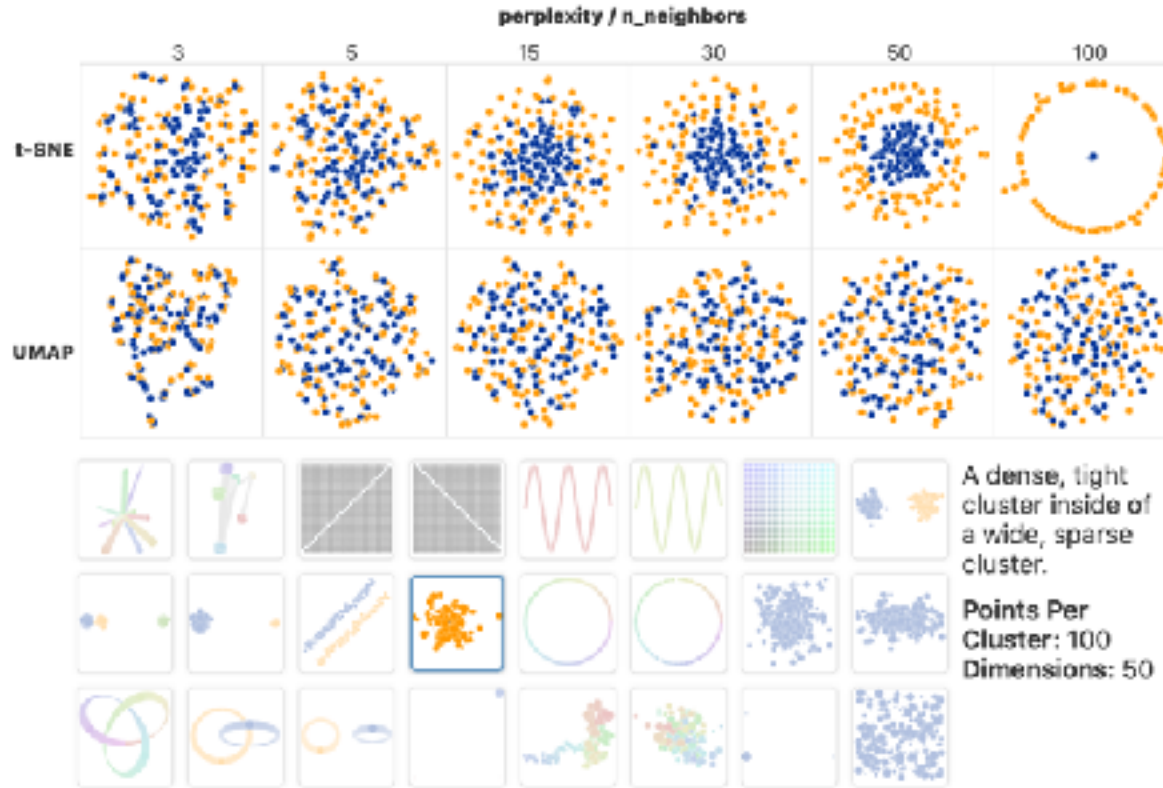


Figure 7: Comparison between UMAP and t-SNE projecting various toy datasets.