COGS138: Neural Data Science

Lecture 11

C. Alex Simpkins, PhD
UCSD Dept. of Cognitive Science, Spring 2023
RDPRobotics, LLC
http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23

rdprobotics@gmail.com | csimpkinsjr@ucsd.edu

Plan for today

- Announcements
- Previous project review
- Project overview
- Review - Last time
- Statistical data analysis, interpretations for neural data science and review

Announcements

- ITS had technical issues
  - A2 - due **Tuesday 5/9 midnight**
  - Reading 2 - Released on canvas and in web site password protected area, lecture quiz due next **Tues 5/9 R2 quiz**
  - **Quiz now due Tuesday 5/9 midnight**
- **Group formation** - check canvas for empty groups if you want to self add
  - We have assigned everyone who did not say they did not want to be assigned, please connect with the team and quickly decide if you want to stay together or move
  - Contact Siddhant to move if needed, contact me if other issues or he doesn't get back to you
- Previous project review released when we get the groups together (this week)

Last time

Course links

Website	http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23	Main face of the course and everything will be linked from here. Lectures, Readings, Handouts, Files, links
GitHub	https://github.com/drsimpkins-teaching	files/data, additional materials & final projects
datahub	https://datahub.ucsd.edu	assignment submission
Piazza	https://piazza.com/ucsd/spring2023/cogs138_sp23_a00/home (course code on canvas home page)	questions, discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/44897	grades, lecture videos
Anonymous Feedback	Will be able to submit via google form	If I ever offend you, use an example you are uncomfortable with, or to provide general feedback. Please remain constructive and polite

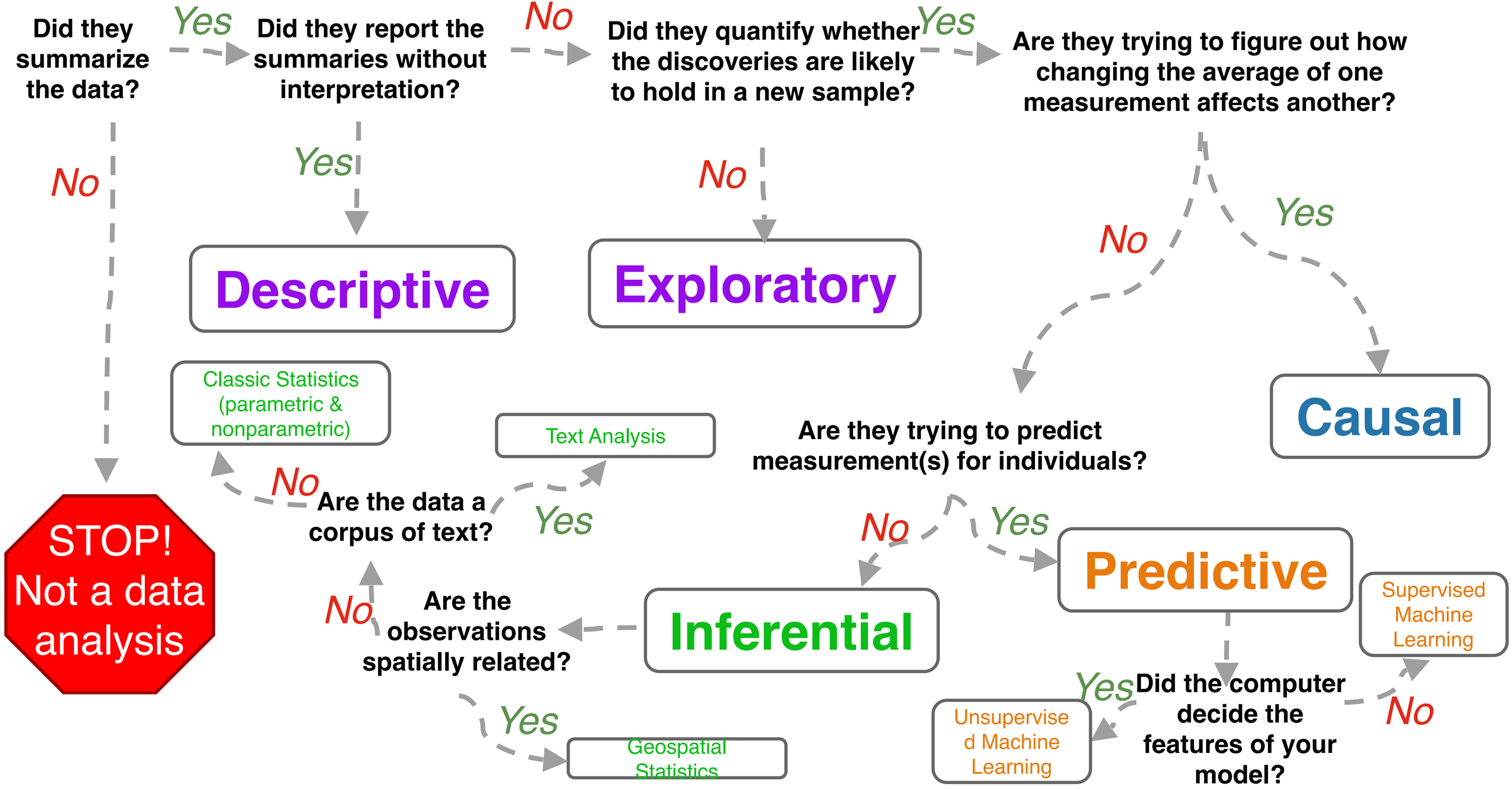
"Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience."

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/

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


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

Statistical Data Analysis

- There are various definitions
- “Statistics” - the science of gathering data and discovering patterns
- “the science that deals with the **collection, classification, analysis, and interpretation of numerical facts or data**” [dictionary.com]

About the final projects

Finding the project files

- Blank starter document for report, handouts and info to start with (draft): https://github.com/drsimpkins-teaching/cogs138/tree/main/main_project
- Links to old projects: https://github.com/NeuralDataScience/Projects

Where will you turn in, work from?

- Github - we will create a repository for each group with the files in previous slides as a starter directory
- You'll add your data, notebook file etc there
- This will be the final turning location

Final Project Overview

1. Identify **questions** that you can answer by using publicly available datasets

1. **Integrate** different datasets

1. Implement **technical skills** to answer your scientific questions

1. Work effectively with a **team**

What will you be turning in??

Proposal

- 1 page document
- Pitching your question & approach

Final Project

- Jupyter notebook
- Steps of analysis that answers your main question
- Background and discussion sections

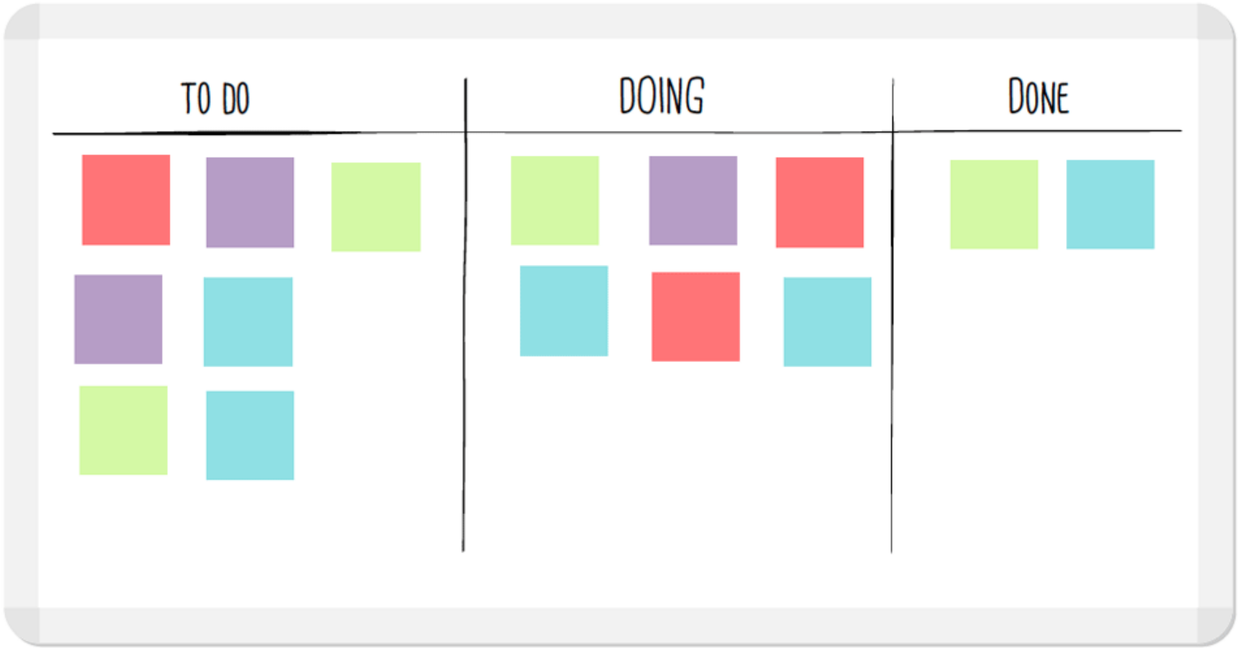
Proposal

1. Group member names

1. Experimental question

1. Background

1. Approach

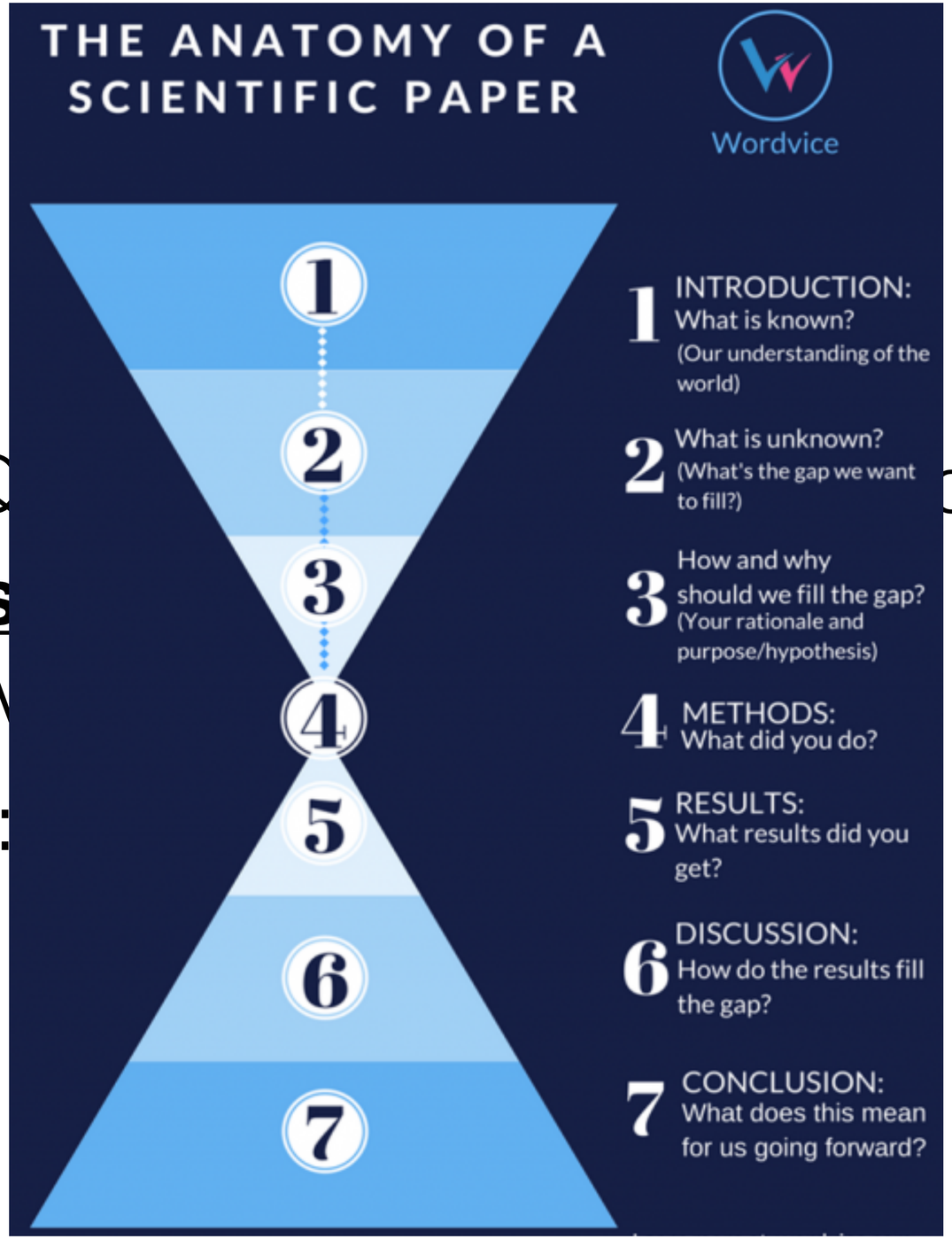


Final Project

1. **Intro**:
   a. Overview, Question, Background, Hypothesis
2. **Data Analysis**:
   a. Wrangling, Viz, Results
3. **Conclusion**:
   a. Discussion, Limitation, Future Steps

Final Project

1. **Intro**:
   a. Overview, Q… …othesis
2. **Data Analys…**
   a. Wrangling, …
3. **Conclusion**:…
   a. Discussion, …



Final Project

Overview

> In this work, we aim to develop new approaches to automated hypothesis generation by utilizing the vast neuroscience literature. Building on prior work in semi-automated hypothesis generation, we present a hypothesis-first algorithm for computing a hypothesis "attractiveness" metric which is more

Data Collection

> We based our data collection on the same 591 neuroscience terms that were analyzed in Voytek[1], to provide a reasonable benchmark to compare our models against. Using these terms, we utilized the NCBI E-Utils API to fetch intersection and union counts for the ~174k hypotheses (term-pairs) for the years 1700-2000 and 2000-2021 that were based on a term's presence in either the title and/or abstract of a given paper. For example, given a term pair (t1, t2), we would query all papers in a given time period whose titles and/or abstracts contain both t1 AND t2, and compute the sum of t1 NOT t2 and t2 NOT t1. From here, we constructed three matrices where the entries in each represented the intersection, union, and intersection over union (connectivity).
>
> To make the fetching of multiple time periods reasonable we designed a scraper, called MicroLisc (*full source: ./data_collection/microlisc.py*), that utilizes multi-threading to make requests with a throughput of ~200 requests/sec. Additionally, we were able to vastly reduce the number of requests via caching since |X NOT Y| + |Y NOT X| = |X| + |Y|, when there are no intersections between X AND Y. Hence, by checking for intersection first, we could then retrieve the cardinalities of X and Y from a cache, leaving most hypotheses only requiring only 1 HTTP request (not intersecting), and at max 3 HTTP requests (if intersecting). The combination of these two optimizations allowed us to collect the complete set of hypothesis data for an entire time period in under 1.5 hours. We believe this decrease in data collection time will facilitate further analysis of multiple time periods.

Hypothesis

Our algorithm's method of computing the "attractiveness" metric allows for the hypotheses to be compared on an ordinal scale. In doing so, our hypotheses will be more robust to spurious correlations and have greater specificity. We conjecture that these features of the resulting hypotheses make them more meaningful in terms of their utility in predicting future trends in neuroscience.

> computing percent increase in citations between the two periods.

https://github.com/NeuralDataScience/Projects/blob/main/Wi2021/FinalNotebookGroup-TextMining.ipynb

Final Project

Good commenting and following PEP guidelines (https://www.python.org/dev/peps/pep-0008/)

PEP 8 -- Style Guide for Python Code

Data Wrangling

We configure our data directories and nest terms with their synonyms, producing a l... ...tialize our scraper(data_collection/microlisc.py) with this specific term set and run a job with a specified c... ...we gather the conjunction, intersection, and connectivity counts.

Configuration and Brain Term Initialization

[2]:

```
# set up directory & file hierarchy
proj_dir = os.getcwd()
data_dir = os.path.join(proj_dir,'data')
old_data_dir = os.path.join(proj_dir,'old_data')
new_data_dir = os.path.join(proj_dir,'new_data')
neighbor_data_dir = os.path.join(new_data_dir,'neighbors_of_neighbors')
hypothesis_data_dir = os.path.join(new_data_dir,'hypothesis_first')

# load in the 591 terms from the original analysis
df = pd.read_csv(os.path.join(data_dir, 'brain_terms_new.csv'), names= None)

# extract useful information from the dataframe
term_types = df.domain.unique()
mapping = dict(df[['term','domain']].values)
mapping_lower = {k.lower(): v for k, v in mapping.items()}

# Nest all unique terms from the terms dataframe
lis = [[f'"{j}"' for j in list(set(list(df[df.term == i]['synonyms']) \
                                + list(df[df.term == i]['term'])))] for i in df.term.unique()]

assert len(lis) == 588
```

https://github.com/NeuralDataScience/Projects/blob/main/Wi2021/FinalNotebookGroup-TextMining.ipynb

Final Project

Discussion and Further Directions

Based on the approaches implemented so far and the success we have seen, there is a range of different directions we could choose to further develop our architecture for automated hypothesis generation.

A more descriptive dataset: Currently, our architecture is limited by the scope of terms we are investigating and the data we are capturing related to those terms. To be specific the current architecture leans heavily on the scraped term-pair counts of just PubMed published literature. Whilst this is sufficient as a proof of concept, it isn't comprehensive and there are many ways we can further develop this stage of the project. One idea was to explore the citations of the papers we are scraping and the h-index of the authors publishing said papers. Based on how the current architecture works, terms can co-occur frequently in a range of papers, however, these papers can be largely insignificant, rarely being cited or reviewed. By utilizing the citations of each paper and how citations are growing over time we can add an extra dimension of scrutiny to our produced hypotheses, potentially boosting the utility of the output.

Addition of NLP to add extra insight into scraped hypotheses: One proposed area for further investigation is the addition of NLP-driven insights into the data collection process. Currently, the models derive insight from data relating to term occurrence. The addition of sentiment analysis could add a level of scrutiny to the number and type of occurring matches, possibly resulting in more accurate hypotheses. This is because simply capturing term occurrence fails to account for how the terms are occurring together. For example, take into account these two passages:

- Alzheimer's disease is related to the deterioration of the prefrontal cortex.
- Alzheimer's disease has no relation to the deterioration of the prefrontal cortex.

Although these statements communicate contradictory information, our current algorithm counts them the same. By building functionality for understanding sentiment, our model would be better able to make better decisions relating to the hypothesis it produces.

Construction of a research tool: Following the testing and finalization of the hypothesis generation framework, the team is interested in developing a Python Package Index API for public use. This API would function as a streamlined way for prospective researchers to query for hypotheses based on our refined models, allowing for deployment in the wider scientific community.

https://github.com/NeuralDataScience/Projects/blob/main/Wi2021/FinalNotebookGroup-TextMining.ipynb

Groups

Distributions

- http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Distributions.ipynb
- http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Central%20Limit%20Theorem.ipynb
- http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Correlation%20resampling.ipynb

On to today…

Correlations analysis, covariance and Time series analysis

PIP package manager

- pip (package manager) – Wikipedia
- Written in python
- Used to install, remove, manage software packages
- Connects to online package repository of public software (Python Package Index)
- Most python packages come with PIP installed
- Home page:pip documentation v23.1.2 (pypa.io)

Usage 99%

- ***pip install some_package_name***
- ***pip uninstall some_package_name***

Central tendency - Mean

- Balance point
- "Expected value" (population mean)
- Computed by
  - Sum scores,
  - Divide by number of scores

$$M = \left(\sum_{i=1}^{N} x_i \right) \Big/ N$$

$$\{1.0, 1.0, 2.0, 3.0, 4.0, 4.0, 4.0, 4.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 9.0, 0.0, 0.0, 0.0\}$$

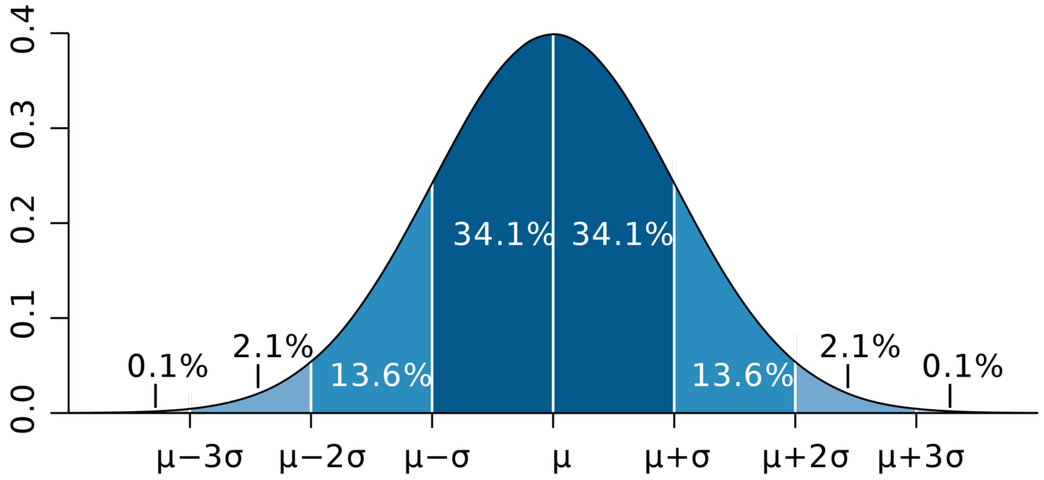
$$N = 18$$

$$\sum x_i = 80.0$$

$$M = \left(\sum x_i \right) \Big/ N = 80.0 / 18 = 4.4$$

Central Limit Theorem and Law of Large Numbers

- If X is taken independently from the same distribution, then X_i is said to be a random sample from that distribution
- X_i are said to be independent identically distributed (i.i.d.)
- **Law of large numbers** (LLN)- sample mean approaches population mean as n approaches infinity
- **Central limit theorem** (CLT) - the distribution of the sample mean approaches a normal distribution for n approaching infinity



Mean in neural data science

- Calculation in python
  - `import statistics`
  - `statistics.mean([data])`
- Application
  - DC or AC eeg?
    - How do you remove a DC bias?
  - Mean number of responses
  - Mean movement
  - Mean amplitude of oscillation in stroke, parkinson’s, etc patience
  - Where else do we see the mean in the brain or neural data science?

Mean, variance, standard deviation review

- Sample mean

- Population mean (“expected value”)

$$\mu = E(X_i)$$

Central tendency - Mode

- Most common number of a distribution
- Tells you which value has the highest frequency
- **What if there are ties?**
  - **More than one mode!**
  - **Which of the following is the mode?**

$$\{1,2,2,2,2,2,2,3,4,5,6,7,8,8,8,9,9\}$$

Mode in neural data science

- Calculation
  - `import statistics`
  - `statistics.mode(data)`
- Application
  - Some examples in NDS
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4059688/

Central tendency - Median

- The middle number of a distribution when the numbers have been ordered (sorted)
- Each score is counted separately, so if you have repeating scores such as 50 and 50, each one becomes part of the count
- Order the scores from low to high or high to low
- Count from both ends to the middle position

Central tendency - Median

- If odd number of scores, there will be one median

$$\{1,2,3,10,50\}$$

$$Median = 3$$

- If an even number of scores, count to the two closest to the middle (ie count from low towards high, high towards low) and take their average (add them up and divide by two)

$$\{1,2,2,3,3,4\}$$

$$2,3$$

$$Median = (2+3)/2 = 2.5$$

Median in neural data science

- Calculation
  - `import statistics`
  - `statistics.median(data)`
- Application
  - Median and MAD: The median and median absolute deviation (MAD)

$$x'_k = \frac{x_k - median}{MAD}$$

$$\text{where MAD} = \text{median}(|x_k - median|)$$

https://www.cs.ccu.edu.tw/~wylin/BA/Fusion_of_Biometrics_II.ppt

Mean, variance, standard deviation review

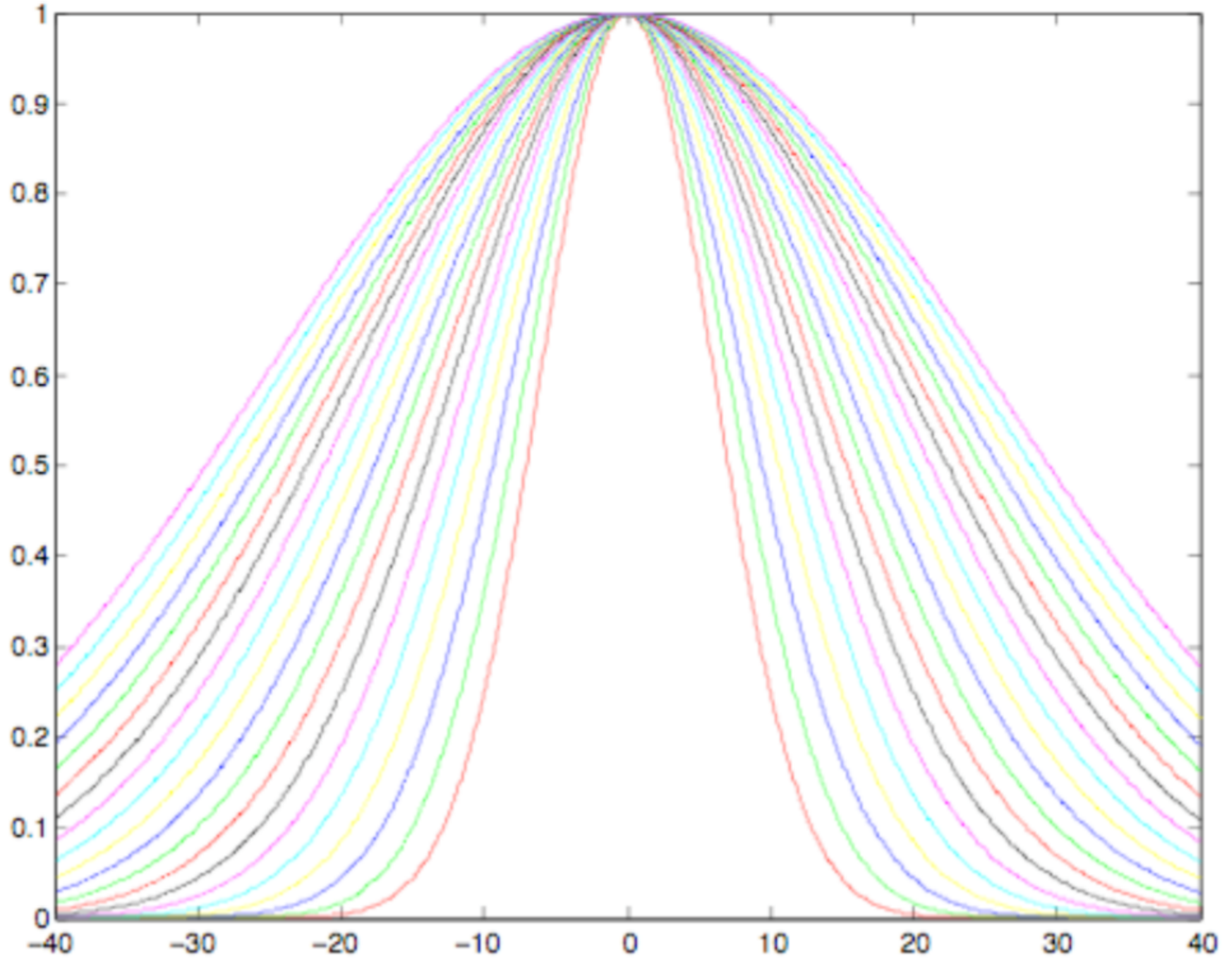
Standard deviation

How are they related?

- If you have a…
  - **Normal distribution,**
    - Mean=Median=Mode
  - **Symmetric distribution**
    - Median = Mean
  - **Skew distribution**
    - Median towards the body, mean towards the tail
      - **+skew: mean>median**
      - **-skew: mean<median**
- But this doesn't seem to be saying everything…

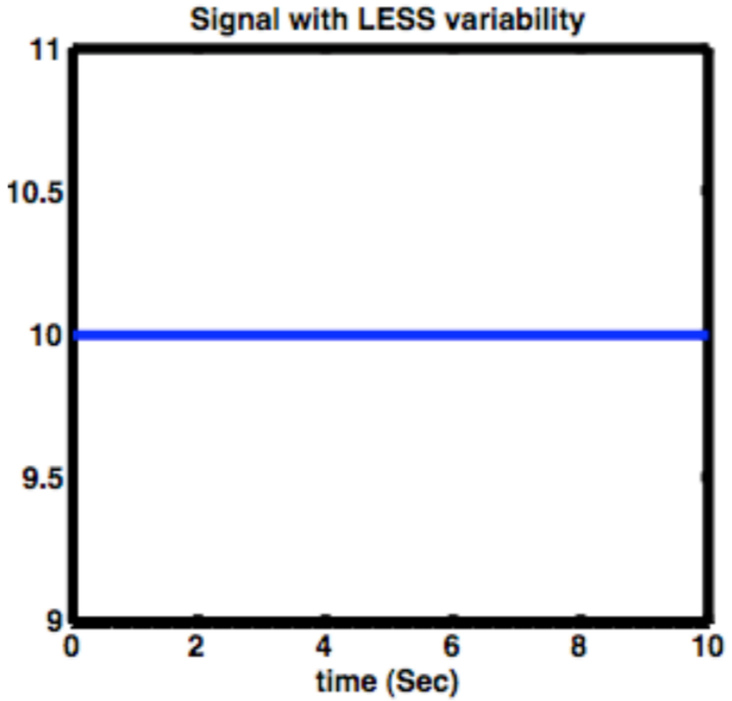
The mean isn't everything!

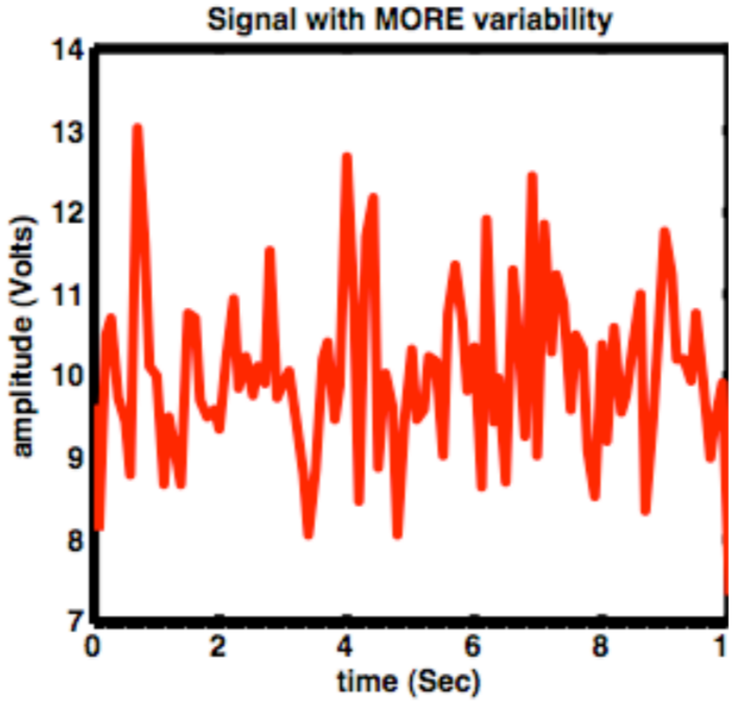
These all have the same mean



Why we need a measure of variability

Same means, different variability of the signal





We need a measure of Variability, here are a few…

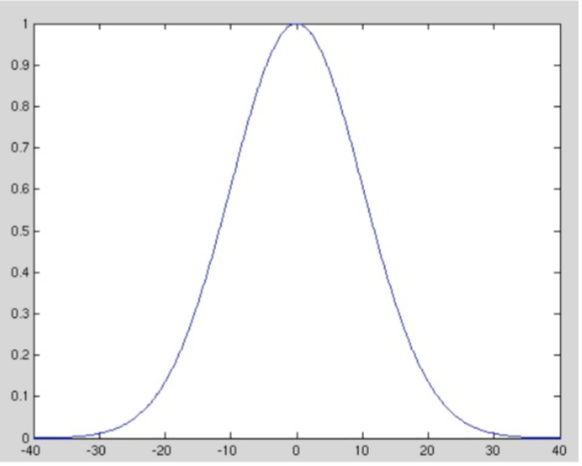
- Range
  - From math review, difference between max and min values of the data

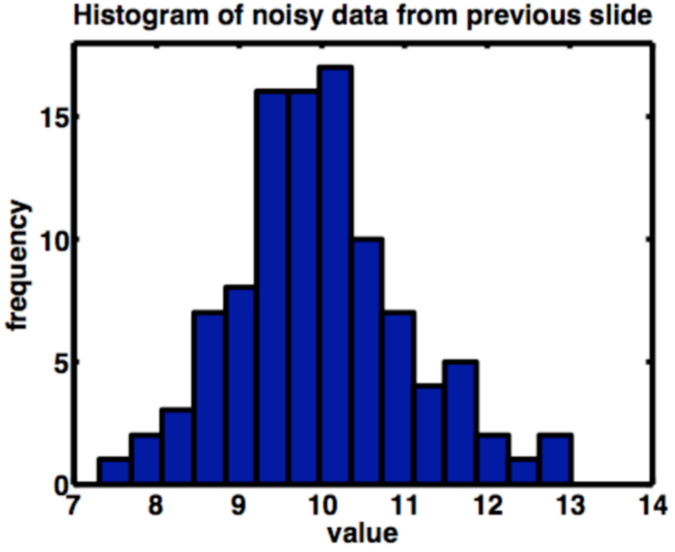
$$Range(x) = Max(x) - Min(x)$$

- Variance
  - Mean of squared deviations from the mean
  - In square units of the sample variable
- Standard deviation
  - Square root of variance
  - In units of the sample variable - sometimes easier to interpret

Returning to the normal distribution…and considering our data in terms of a histogram…

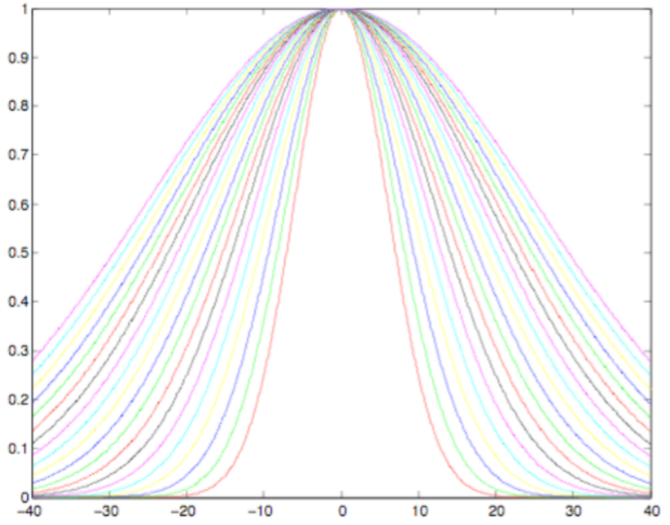
- The distribution of points about the mean can be considered in terms of probabilities
- How likely is a point to deviate from the mean?
- We call the normal distribution a *probability density function (PDF)* because it allows us to predict the likelihood that a sample will take on a particular value





Variance

- Whereas the mean defines a measure for the most likely point in state space (the center ‘location’ of a normal distribution)
- We can define the spread of the normal distribution about the mean by its *variance*



Variance (part II)

- Steps to compute the variance
  - **Compute the deviations from the mean for all the data**

$$d_i = \left(x_i - \bar{x}\right)$$

  - **Compute the square of each of the deviations**

$$sd_i = \left(d_i\right)^2$$

  - **Sum up all these squared deviations**

$$ssqd = \sum_{i=1}^{N} \left(sd_i\right)$$

  - **Divide the mean squared deviations by N, the number of observations**

$$Var = \frac{ssqd}{N}$$

Standard Deviation

- Typical ‘deviation’ from the mean
- Ie how far on average scores depart on either side from the mean
- Easy to compute after the variance - just take the square root of the variance

$$SD = \sqrt{Var} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

$$\bar{x} = \frac{\sum x_i}{N}$$

Z scores

- A Z score is simply a measure of how many standard deviations away from the mean a score is
- Units are standard deviations

$$Z_i = \frac{X_i - \mu}{SD}$$

Covariance

- Covariance is very commonly used in statistical analysis as the basis for advanced statistics
- Gives a quantitative measure of the relationship between two variables

$$Cov(X,Y) = E\left[\left(X - \mu_x\right)\left(Y - \mu_y\right)^T\right]$$

$E = \text{exp}ectation$

$\mu = mean$

More Covariance

- If the two variables are independent, the covariance is 0
  - **(BUT IF COVARIANCE IS 0 THAT DOESN'T MEAN THE VARIABLES ARE INDEPENDENT!!!)**
- If they are totally dependent the covariance of data, can be arbitrarily large
  - **(AGAIN THE CONVERSE IS NOT NECESSARILY TRUE)**
- The diagonals are the variance of each variable
- If each row is an observation, and each column a variable…

$$\text{cov}(X) = \left(\frac{1}{N-1}\right)\left(X - mean(X)\right)\left(X - mean(X)\right)^T$$

Correlation coefficient motivation

- We want to define a measure of how related our dependent and independent variables are
  - Variance, STD – variation of a single variable
  - Covariance – how two things vary in relation to each other
  - How do we compute the linear dependence of one variable to another?
- Correlation coefficient!

Intuitive arrival at the Correlation Coefficient

- Many kinds (we are going to discuss Pearson's product moment coefficient by Galton)
- A test for linear independence
- We want to measure how two things co-vary
  - We observe one thing varying (e.g. sunset)
  - We observe another thing varying (e.g. air temp. decrease)

Intuitive arrival at the correlation coefficient (II)

- **Positive Correlation** - When one thing's magnitude varies positively, and another thing's magnitude varies positively
  - **and if both vary negatively, also this is referred to as positive correlation**
- **Negative correlation** - When one thing's magnitude varies positively, and another thing's magnitude varies negatively
  - **And if one varies positively while the other varies negatively, this is also referred to as negative correlation**

Intuitive arrival at the correlation coefficient (III)

- We want our measure to be a single number
- In some way we'll need to scale the calculations so that the number is unitless
  - **The variables we're comparing may be in different units**
  - **We also don't care about bias - we're interested in variations, so we make our measures about zero, and normalize each**
  - **Remember when we presented z-scores as a normalized measure of how far from the mean a particular sample is in a dataset?**

$$Z_i = \frac{X_i - \mu}{SD}$$

Intuitive arrival at the correlation coefficient (IV)

- We arrive at the correlation coefficient by multiplying each z-score from one variable by the z-score from the other variable, then averaging all those results
  - **Thus if both tend to vary positively?**
    - Positive correlation
  - **If both tend to vary negatively?**
    - Positive correlation
  - **If one varies positively, and the other negatively?**
    - Negative correlation
  - **If sometimes they both vary positively or negatively, sometimes they vary oppositely?**
    - Small or near zero correlation

Correlation coefficient

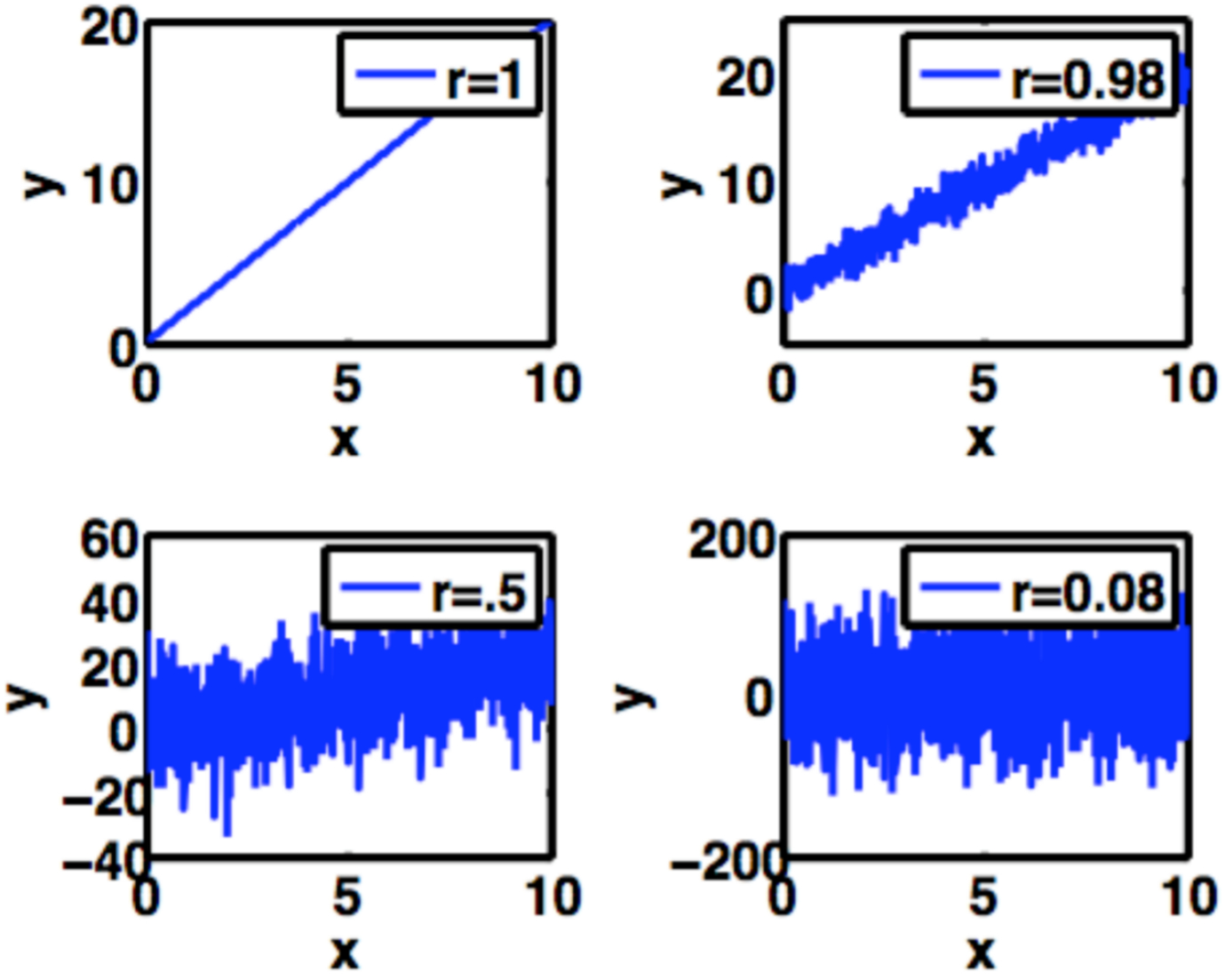
$$\rho(j,k) = \frac{\sum_{i=1}^{N} Z_{ij} Z_{ik}}{N}$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\rho(X,Y) = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

Characteristics

- Range
  - **-1<=r<=1**
- Interpretation - independence
  - **Statistical independence**
    - The more distinct and unrelated the covariation, the closer to zero the correlation coefficient
    - Statistically independent if their correlation is zero
  - **Linear independence**
    - Two things varying perfectly together are linearly dependent, variables with less than perfect correlation are linearly independent


r=1
r=0.98
r=.5
r=0.08
x
y
0
5
10
20
60
40
−20
−40
200
−200