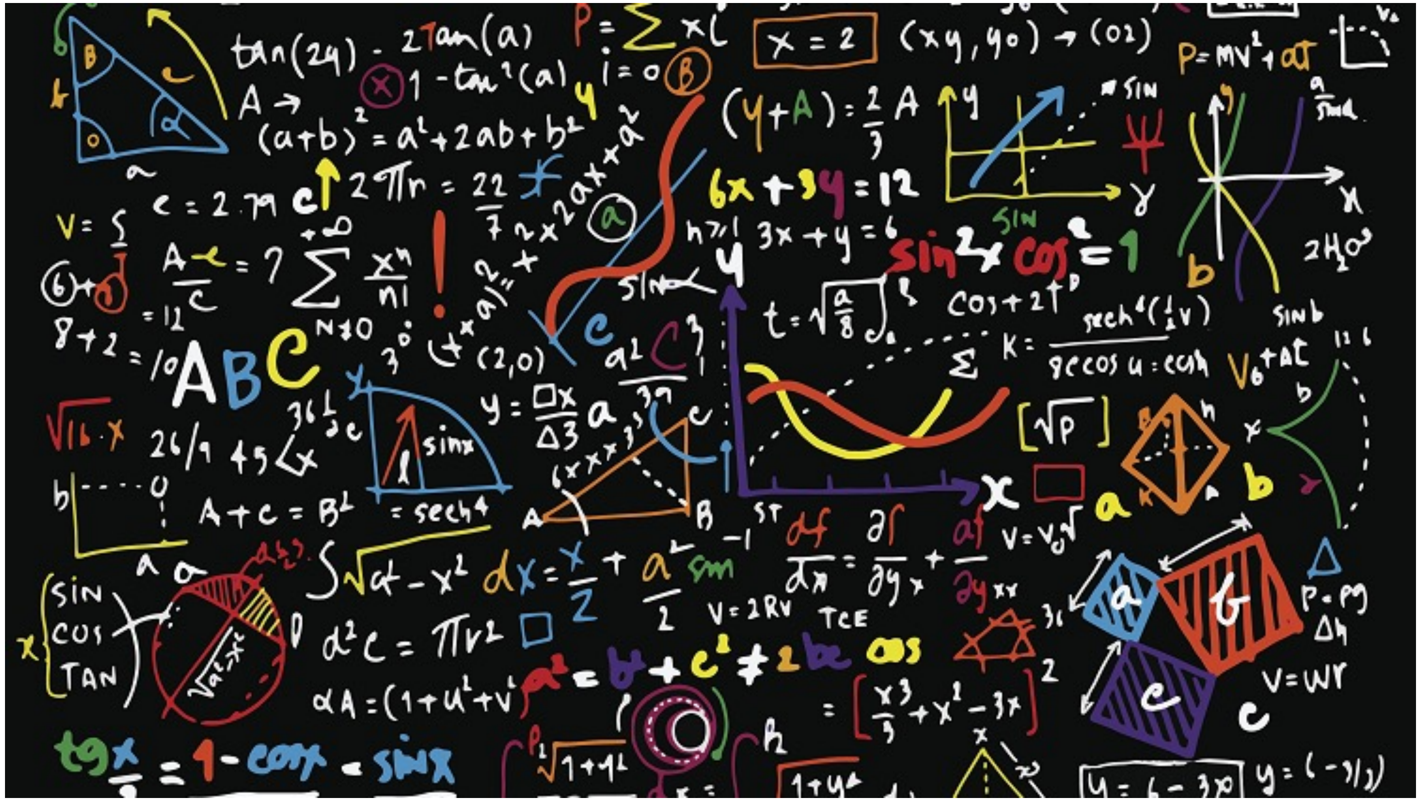


How much maths do I need to learn to be a data scientist?

Data 601 @ UMBC

Outcomes for this evening

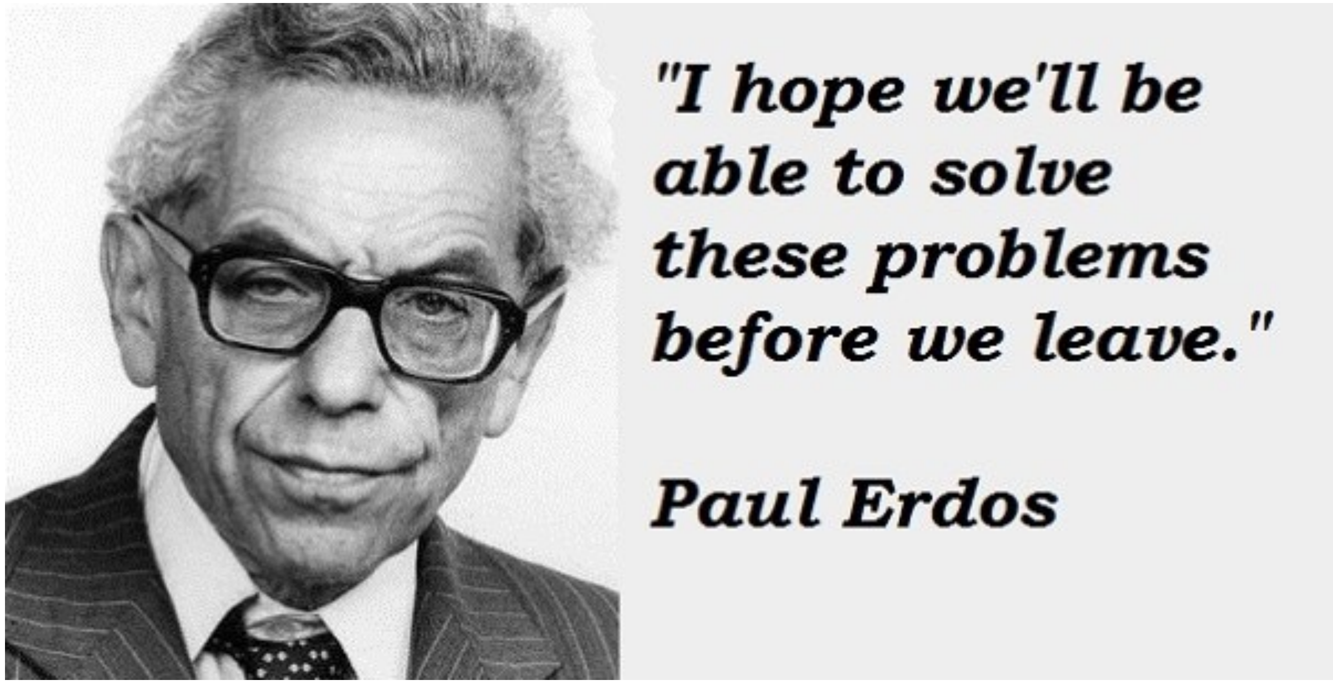
By the end of today's class, you should be able to answer the following:

- Explain the role of sampling data
- Transpose a matrix and a dataframe
- Explain uses of Gaussian and uniform distributions
- Identify misleading representations of data
- Describe three biases that can occur when gathering data

- Math and Data Science
- Probability
- Correlation
- Sets
- Linear Algebra and Numpy
- Calculus and Differential Equations
- Visualization
- Homework

Consider this a survey of an old, diverse field

- Linear algebra (vectors, matrices, cross product)
- Sets (union)
- Statistics (ie mean, median)
  - How to Lie with Statistics
  - Error bars
- Probability, combinatorics
- Calculus
- ODEs and PDEs



"I hope we'll be able to solve these problems before we leave."

Paul Erdos

Caveat: I'm not a mathematician

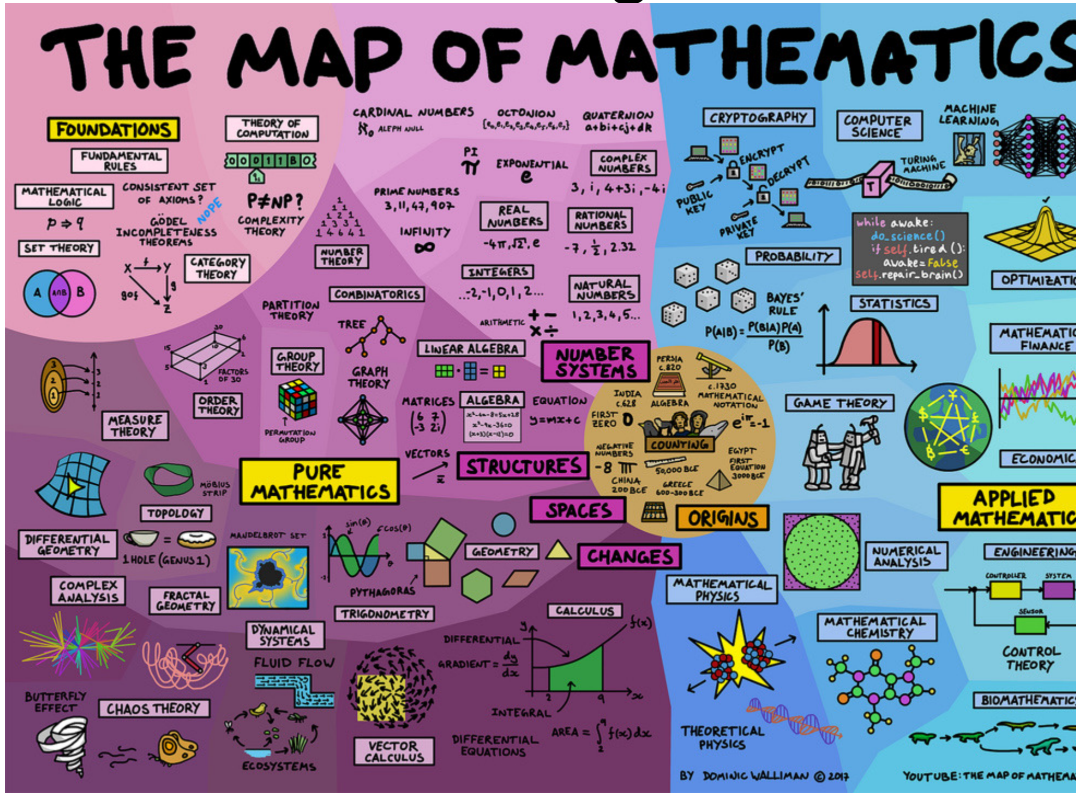
I won't be able to teach you all of Math

Where **Math** shows up in **Data Science**

- Cleaning data – *filling in gaps for missing data with interpolation*
- Modeling expectations – *sense-making, distribution of each variable*
- Generating hypotheses that are numerically testable
- Evaluating test results to validate hypotheses
- Analysis of results – *visualization, sanity checks*
- Explanation of story to audience – what do you expect customers to take away? What is their language?

Resources for learning Math

- Focus on learning the jargon; this is necessary for searching
- For a given topic, evaluate the many options before investing time
  - Teaching style
  - Level of complexity
  - Assumptions about you, the student reader

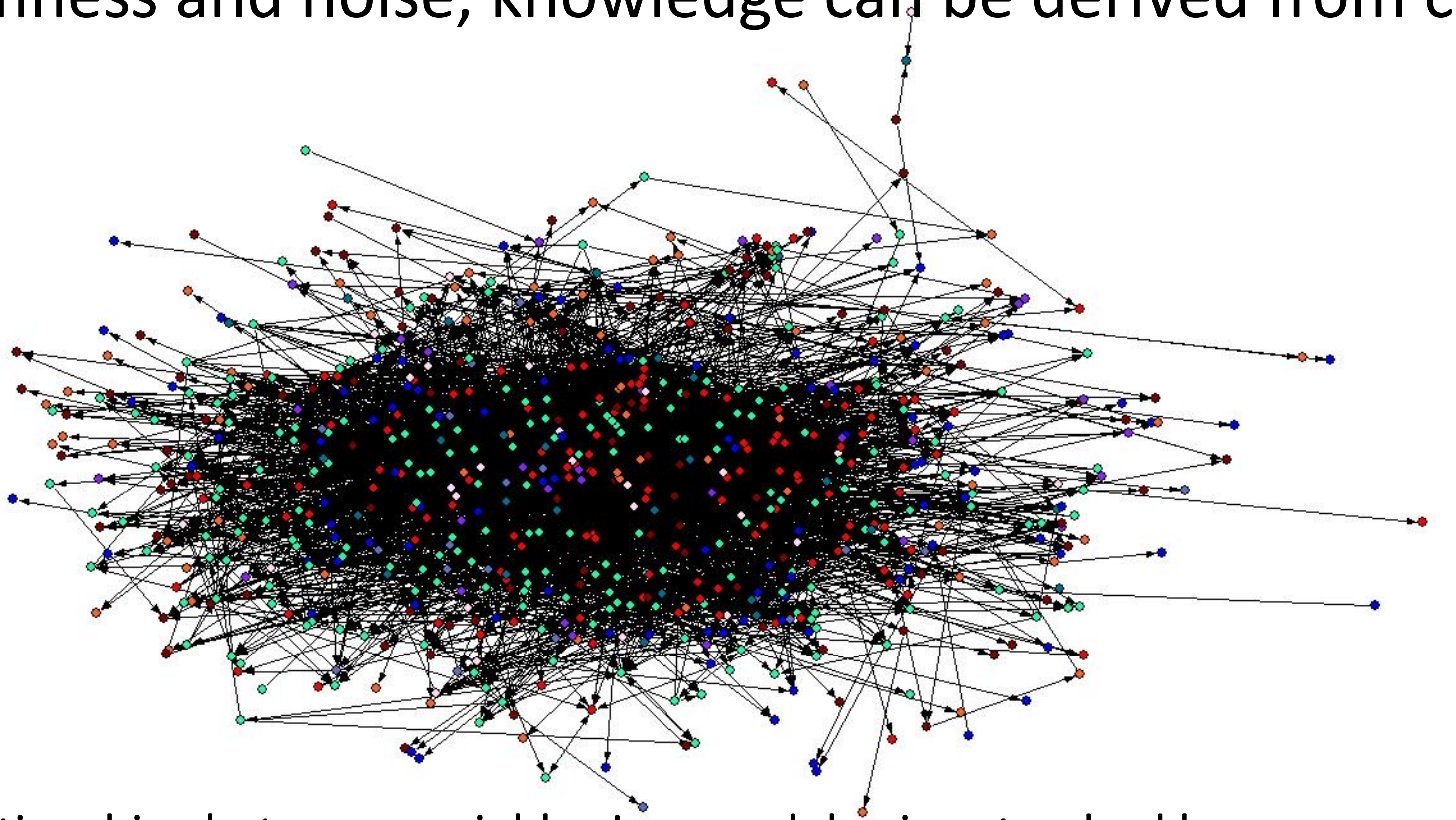


Free resources

- Online (blogs, Coursera, YouTube)
- Books

Relevance of **Statistics** in **Data Science**

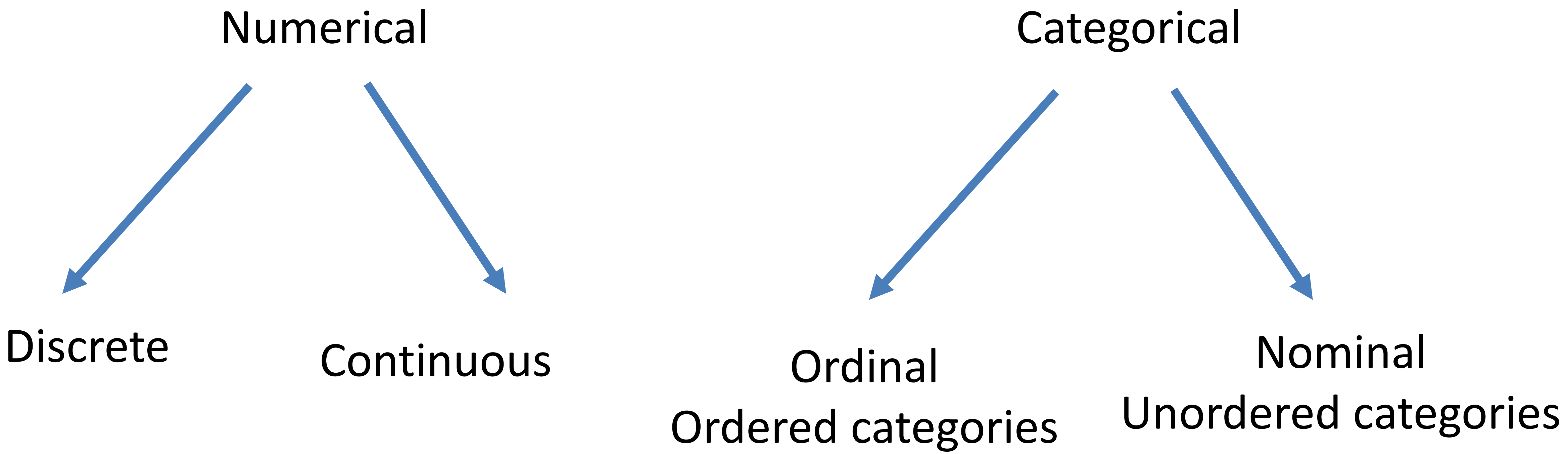
Given randomness and noise, knowledge can be derived from complex data



How:

Quantify relationships between variables in a model using standard language and techniques

Quantifying Relations



Numerical

Discrete

Continuous

Categorical

Ordinal
Ordered categories

Nominal
Unordered categories

Discrete versus Continuous variables

- Discrete variables: outcomes for coin flips, deck of cards, roll of dice
- Continuous variables: time, distance

Continuous: For any two values of a variable, it is possible to get a measurement that is between the two values.

Trick: Rounding continuous to discrete

Rounding is often applied to continuous to make the variable discrete

SPEED
LIMIT
55

- ~~Math and Data Science~~
- Probability
- Correlation
- Sets
- Linear Algebra and Numpy
- Calculus and Differential Equations
- Visualization
- Homework

Core to Statistics: Probability

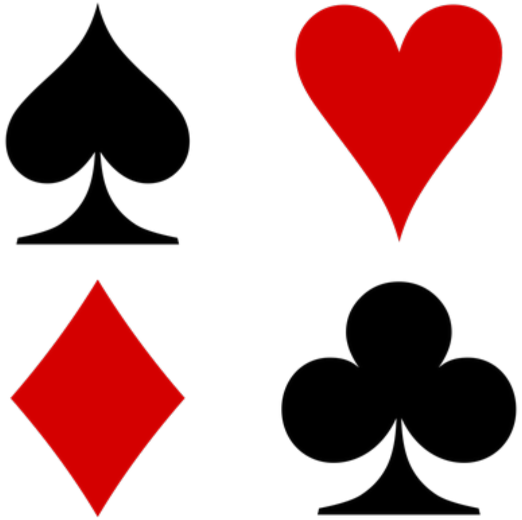
Statistics quantifies relationships
between variables in a model
using standard language and techniques

Probability is a way of figuring out an applicable model

Uniform distribution

Each outcomes is equally likely:

Chance of any of the possible options is the same as any other outcome



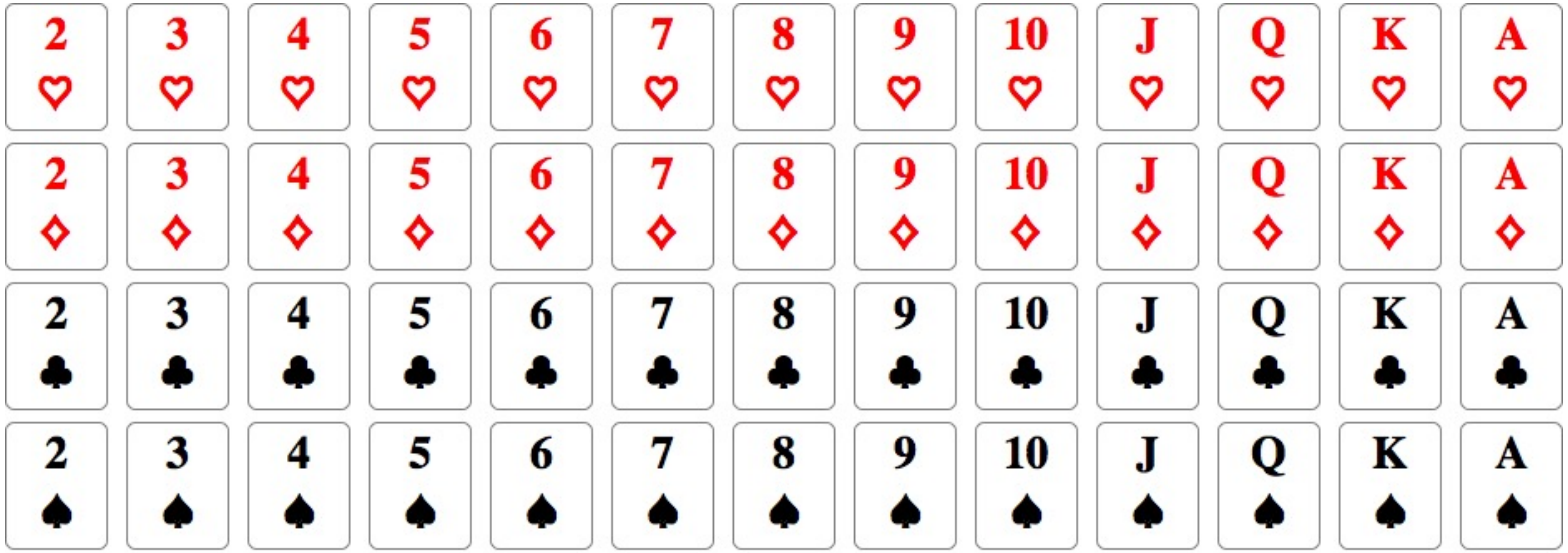
Given a standard deck of 52 cards, what is the chance of getting a card that is a diamond?

Chance of getting a heart: 13/52

Chance of getting a diamond: 13/52

Chance of getting a club: 13/52

Chance of getting a spade: 13/52



2 ♡	3 ♡	4 ♡	5 ♡	6 ♡	7 ♡	8 ♡	9 ♡	10 ♡	J ♡	Q ♡	K ♡	A ♡
2 ♢	3 ♢	4 ♢	5 ♢	6 ♢	7 ♢	8 ♢	9 ♢	10 ♢	J ♢	Q ♢	K ♢	A ♢
2 ♣	3 ♣	4 ♣	5 ♣	6 ♣	7 ♣	8 ♣	9 ♣	10 ♣	J ♣	Q ♣	K ♣	A ♣
2 ♠	3 ♠	4 ♠	5 ♠	6 ♠	7 ♠	8 ♠	9 ♠	10 ♠	J ♠	Q ♠	K ♠	A ♠

Activity: Coin toss

Using your penny, write down (in order) tails/heads for 10 flips

How many possible permutations?

- 1 flip has 2 outcomes: head (H) or tails (T)
- 2 flips --> 4 outcomes: HH or HT or TH or TT
- 3 flips --> 8 outcomes: HHH,HHT,HTH,THH,HTT,THT,TTH,TTT
- 4 flips --> 16 outcomes: HHHH,HHHT,HHTH,HTHH,THHH,HHTT,...

...

- For *N* flips there are *2^N* outcomes

- N=10 flips: 2^10= 1,024 outcomes
- N=20 flips: 2^20=1,048,576 outcomes

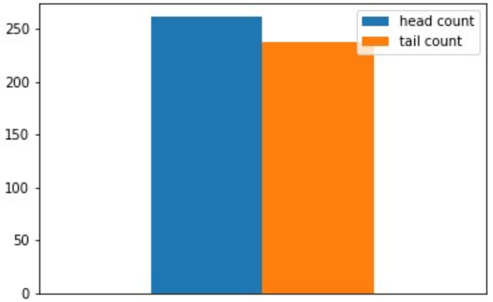
Want to watch how this executes? Check out
http://pythontutor.com/visualize.html

modeling_random_coin_flips.ipynb

Visualizing probabilistic outcomes

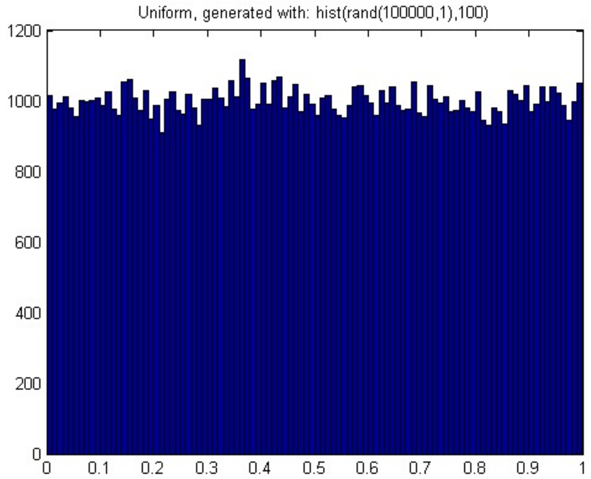
- A fair deck of cards has a uniform distribution of outcomes for a given selection
- A fair coin has a uniform distribution of outcomes

modeling_random_coin_flips_visualization.ipynb



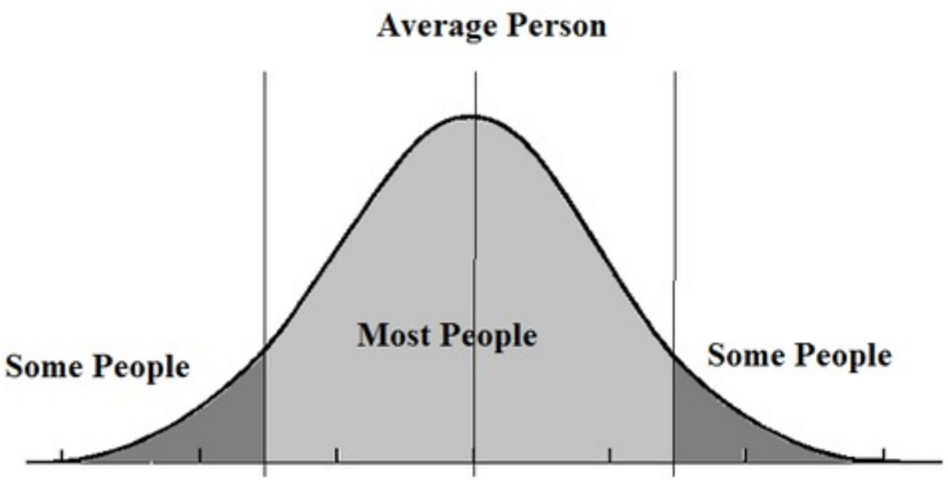
Two outcomes

Uniform distribution



many outcomes

Another distribution: the Bell curve



Gaussian and Binomial Distribution

binomial distribution is discrete; normal (Gaussian) is continuous

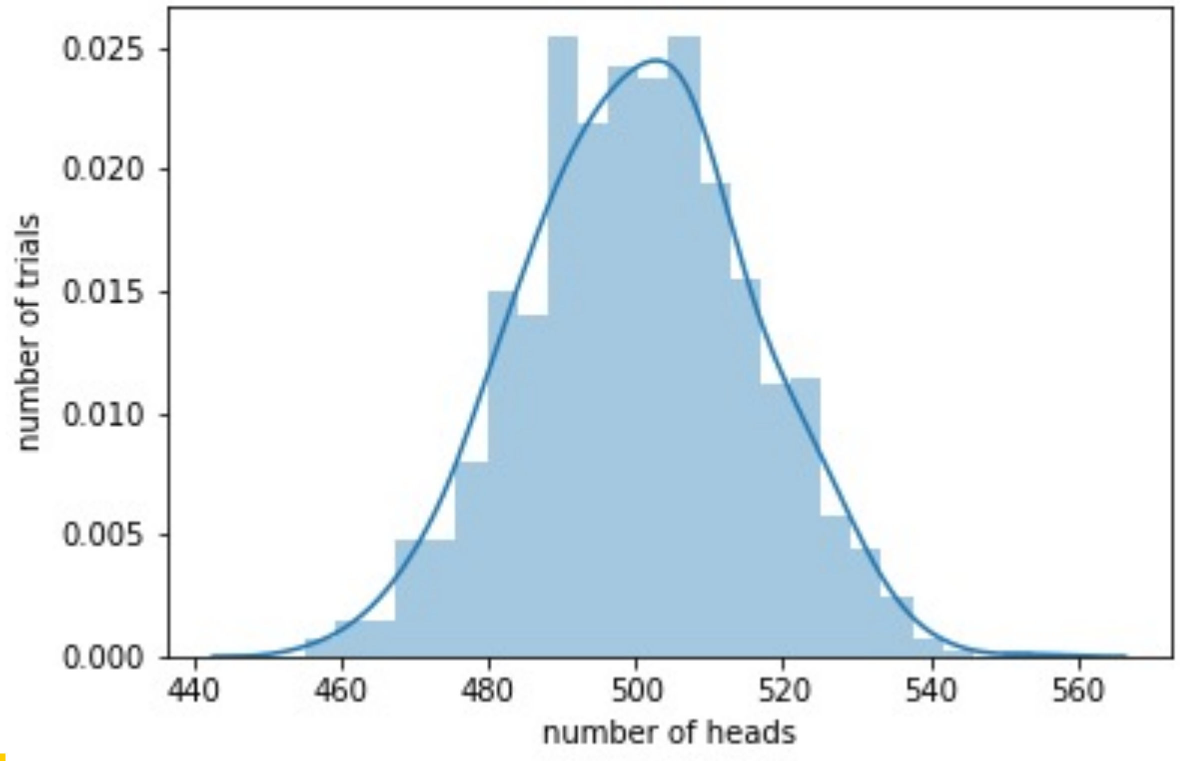
Characteristics of binomial variable:

- A fixed number of repeated, identical, independent trials. n is usually the parameter chosen to label the number of trials.
- Every trial results in either a success, with probability p, or a failure, with probability $1-p$. These must be the only two possible outcomes for a trial.
- The random variable of interest is the total number of trials that ended in a success.

Coin flips produce a bell curve!

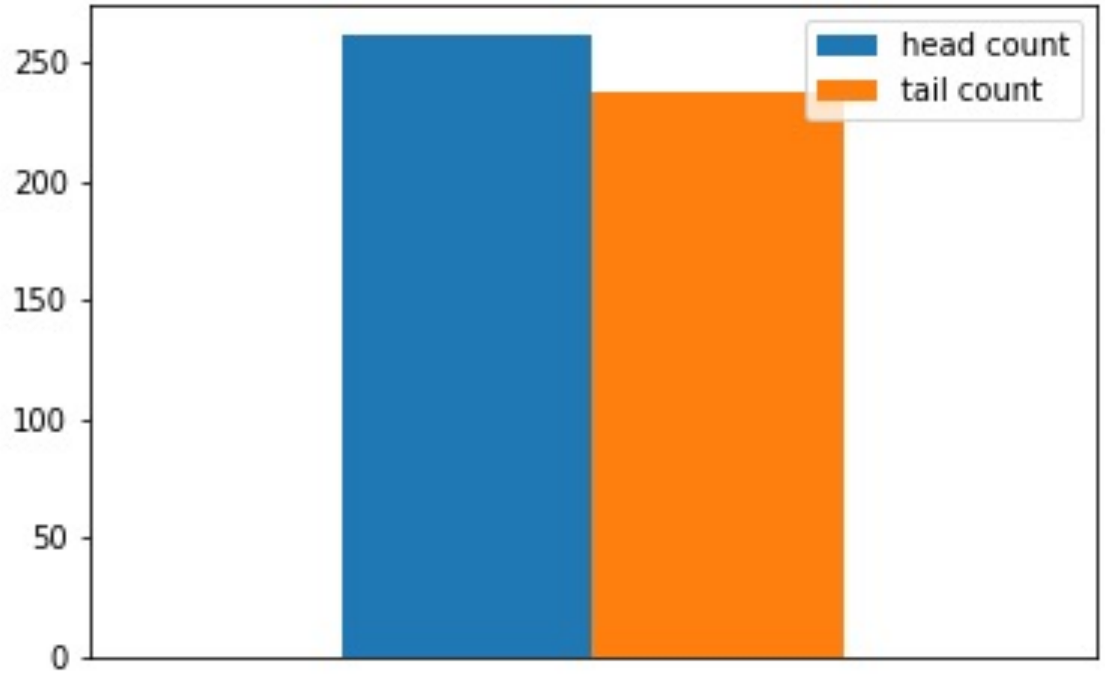
An exploration of the coin flips:

binomial_distribution_for_coin_flips.ipynb



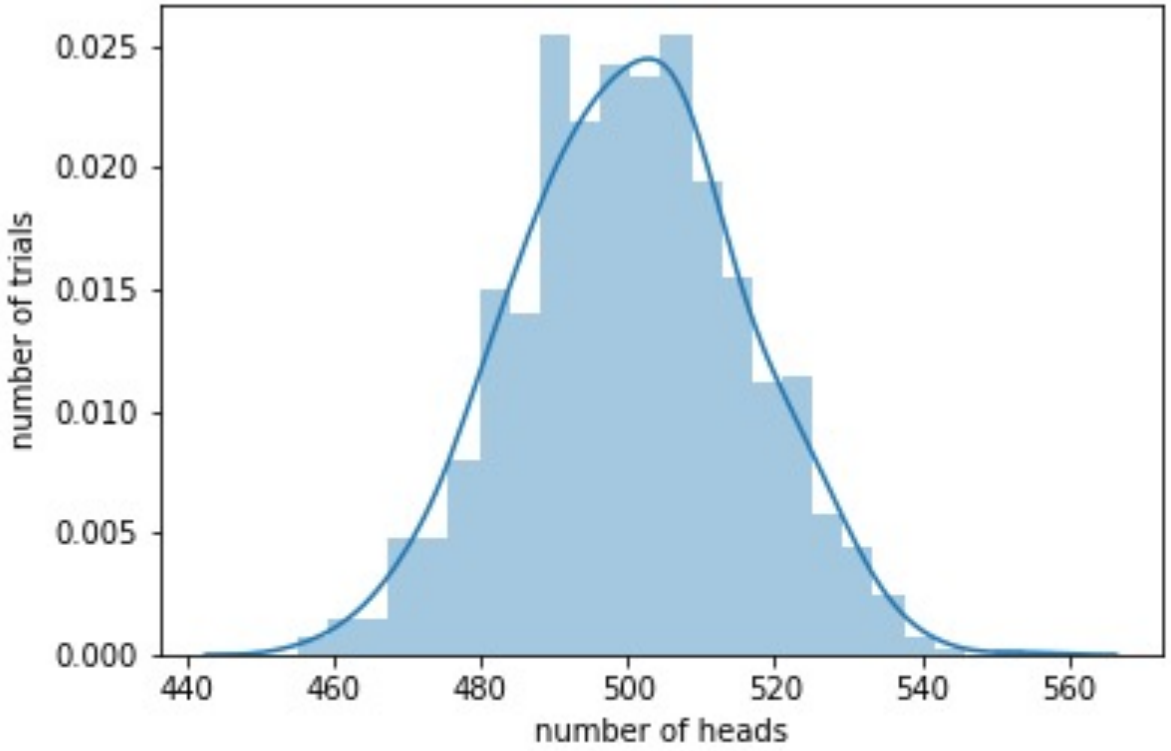
What is the relation between the curves?

Uniform distribution for single trial



Single experiment

Binomial distribution for many trials



Many experiments

Results vary: Error bars tell by how much

- Confidence Interval = certainty of what the mean value is

  Confidence interval improves when more data is added

Distinct from

- Variance measures the width of a distribution
- Standard deviation is sqrt(variance) and has same units as variable

Variance and Standard Deviation do not change as the population size increases

Visually include info about distribution of variable in Violin plots

Results vary: Error bars tell by how much

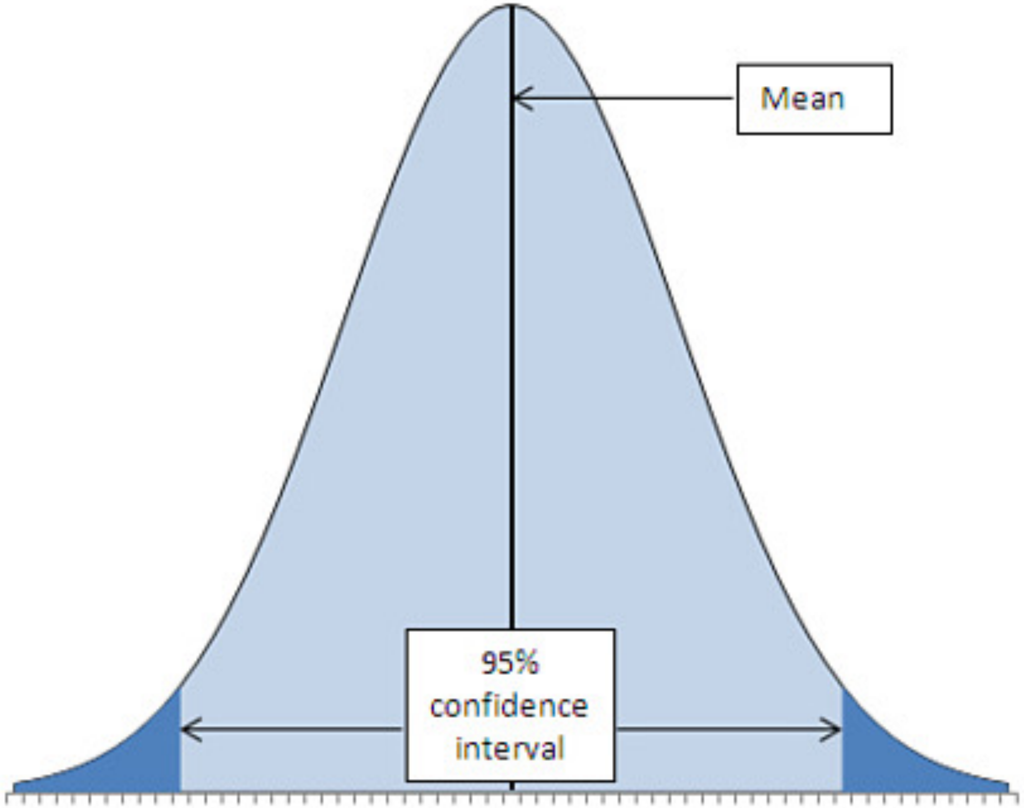
- Confidence Interval = certainty of what the mean value is

  Confidence interval narrows when more data is added

Distinct from

- Variance measures the width of a distribution
- Standard deviation is sqrt(variance) and has same units as variable

Variance and Standard Deviation do not change as the population size increases



Do we have to rely on experiments?

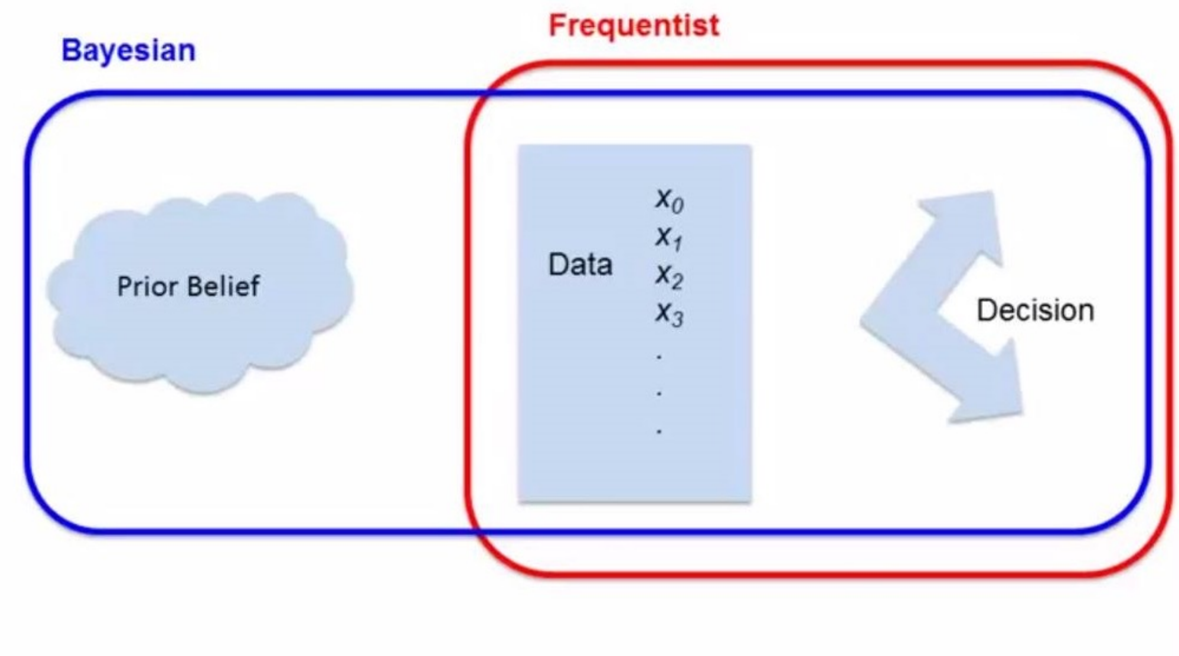
- What if the system being studied is complex?
- Expensive to replicate?

Bayesian versus Frequentist inference

- *Frequentist* approach measures repeated events and does not depend on a subjective prior that may vary from one investigator to another.

- *Bayesian inference*: "What is the probability that it recently rained given that it is wet outside?"

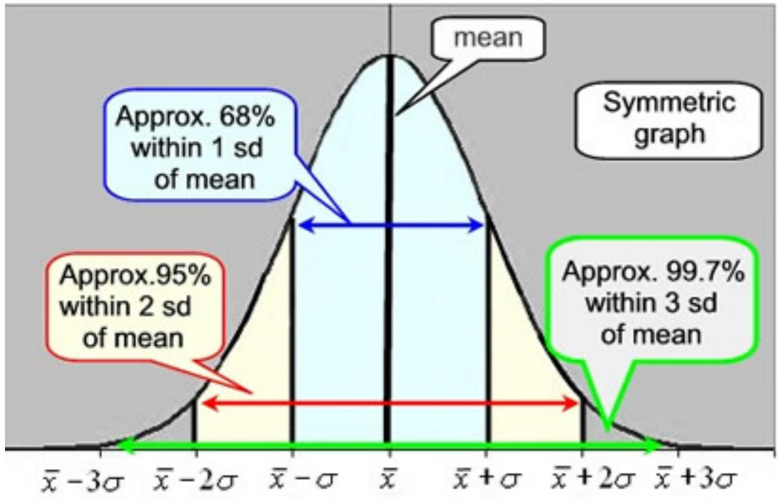
Both approaches allow evaluation of evidence about competing hypotheses.



Jake VanderPlas on difference

- ~~Math and Data Science~~
- ~~Probability~~
- Correlation
- Sets
- Linear Algebra and Numpy
- Calculus and Differential Equations
- Visualization
- Homework

Two events are statistically independent of each other when the probability that one event occurs in no way affects the probability of the other event occurring.

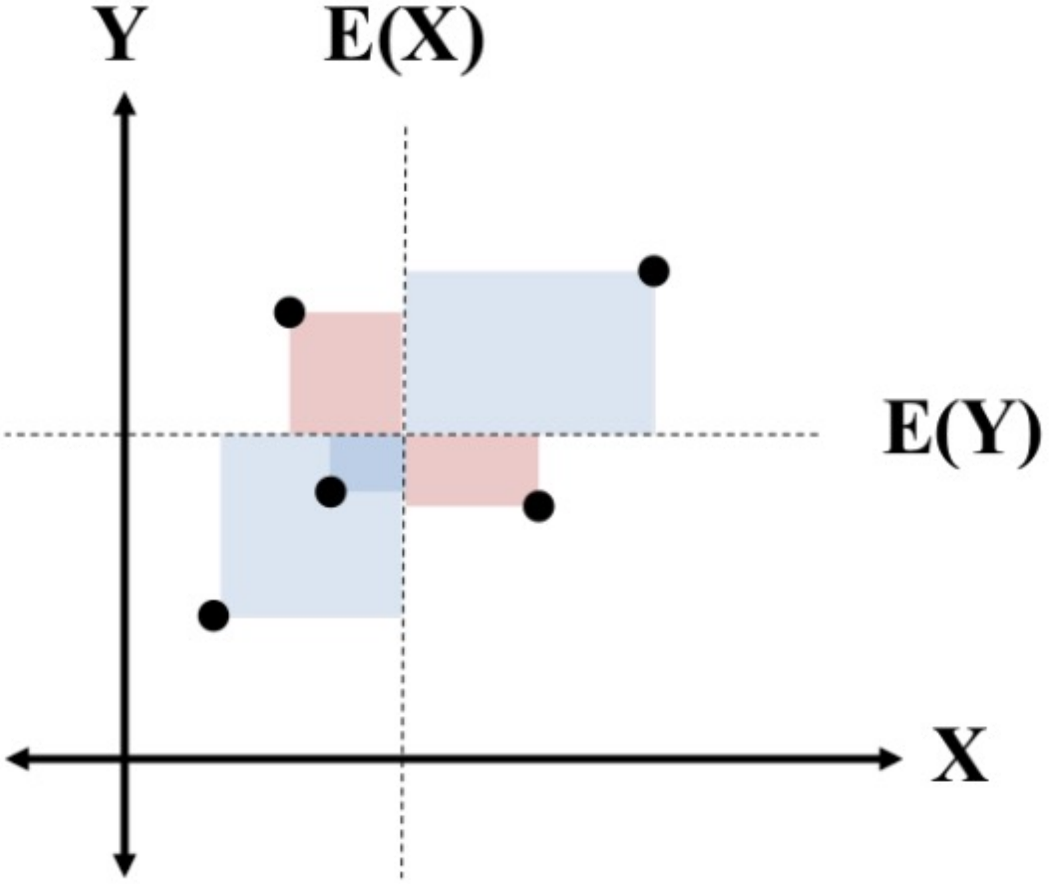


- **Variance** measures width of a distribution
- Covariance is the measure of variance for two random variables (joint variability)
- **Correlation** is the normalized covariance, from –1 to 1

More explanation here and visualizations of covariance are helpful

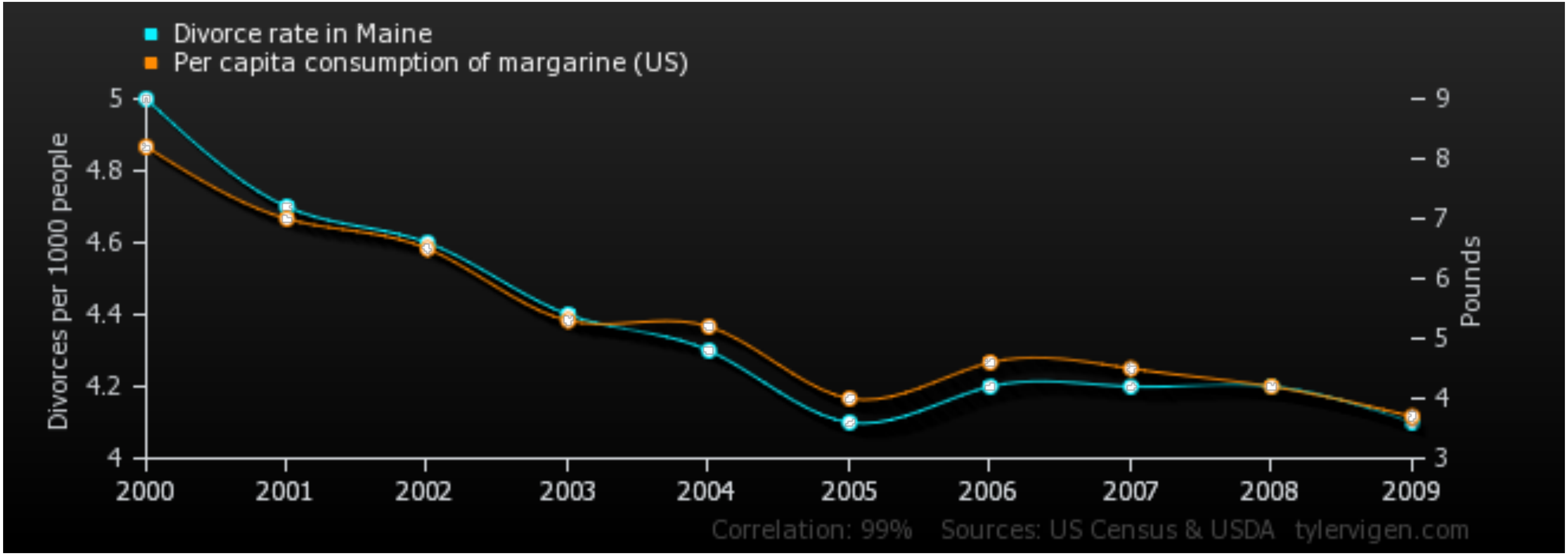
correlation.ipynb

Visualize covariance



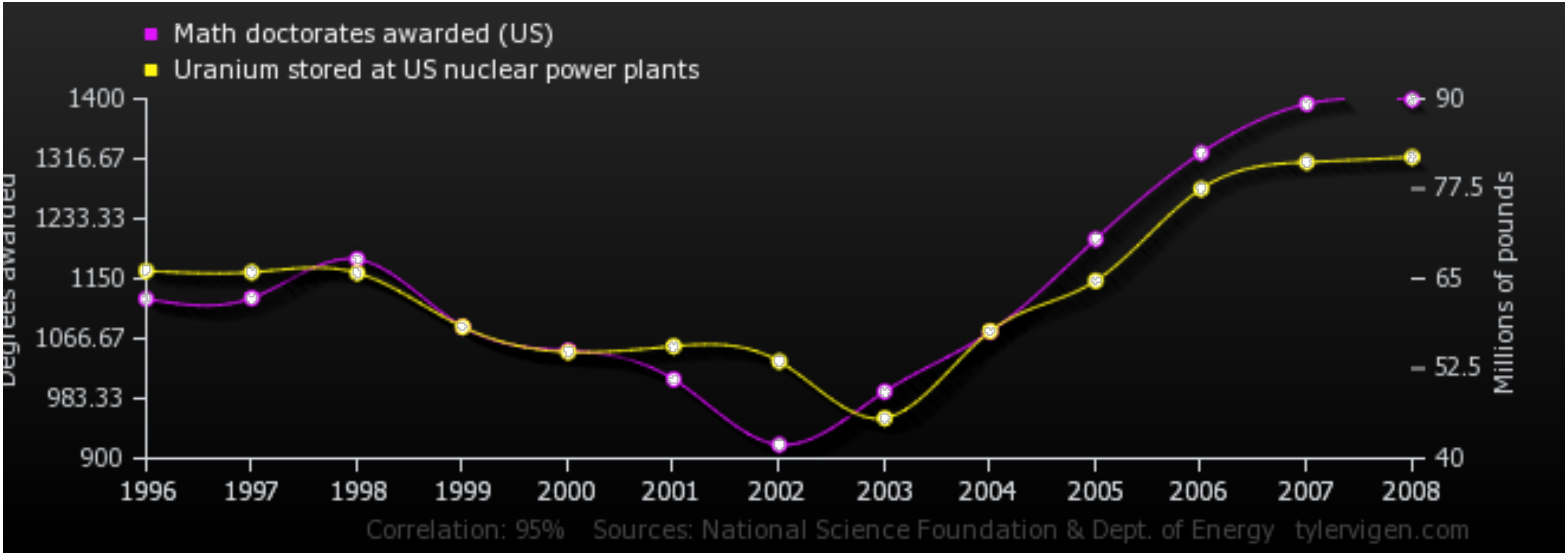
Pink areas are less than the average, so the area has a negative value when summing areas

Correlation = 0.992558 for Divorce and Margarine



- http://www.tylervigen.com/spurious-correlations

Correlation = 0.952257 for PhDs and Uranium



Correlation: 95% Sources: National Science Foundation & Dept. of Energy tylervigen.com

https://tylervigen.com/view_correlation?id=1100

How to (un)intentionally mislead

- Counting and math are objective
- Collection, analysis, and interpretation of data is implemented by humans

Consequence: data you work with may have issues you need to account for

It's a Fan!
It's a Spear!
It's a Wall!
It's a Rope!
It's a Snake!
It's a Tree!

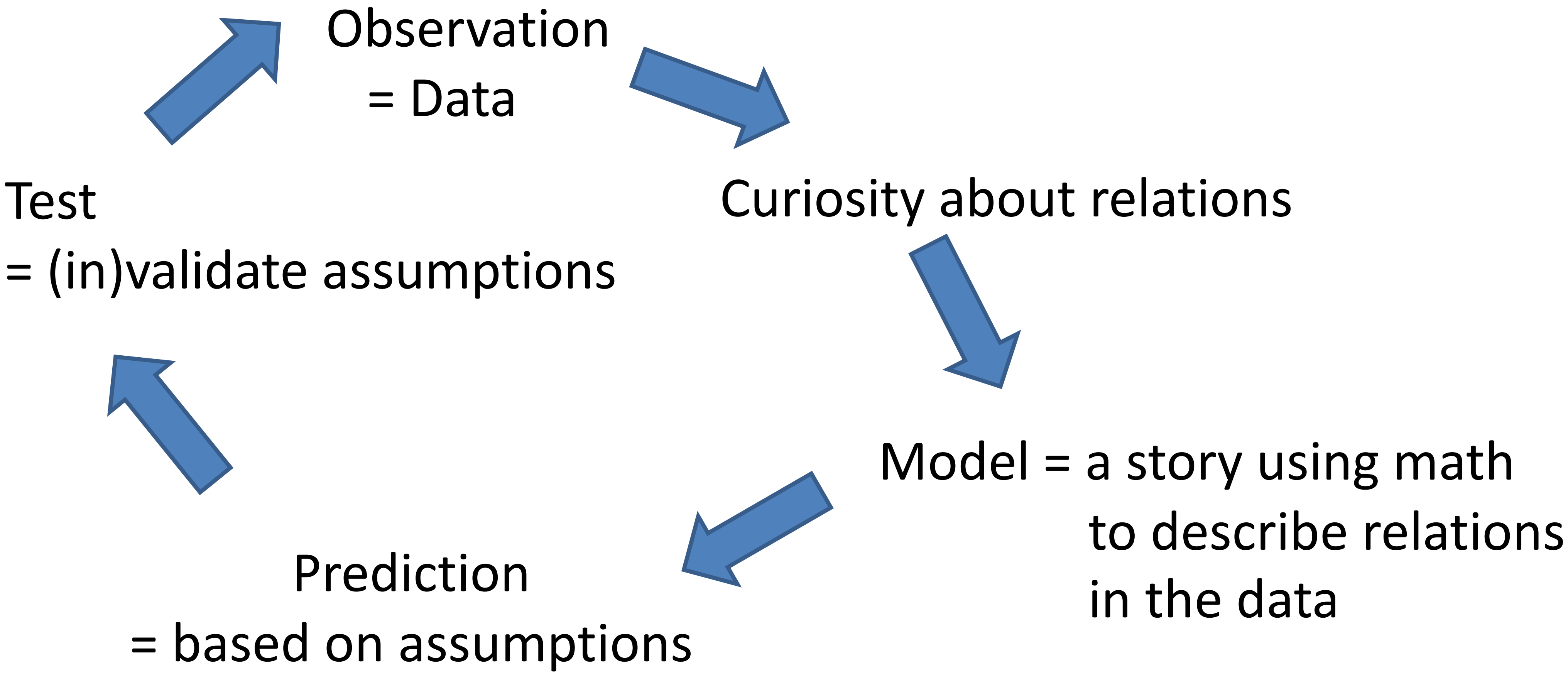
Data collection: exhaust or sample?

Sampling a population can introduce bias

- Area Bias – geographic area of sample needs to be representative of study population
- Self-Selection Bias - decision to participate may correlate with traits that affect the study
- Leading Question Bias - tone of the question suggests the answer
- Social Desirability Bias - reluctance to admit to doing something that is considered socially undesirable

- ~~Math and Data Science~~
- ~~Probability~~
- ~~Correlation~~
- Time series
- Sets
- Linear Algebra and Numpy
- Calculus and Differential Equations
- Visualization
- Homework

Reminder of the Scientific Process



Observation
= Data

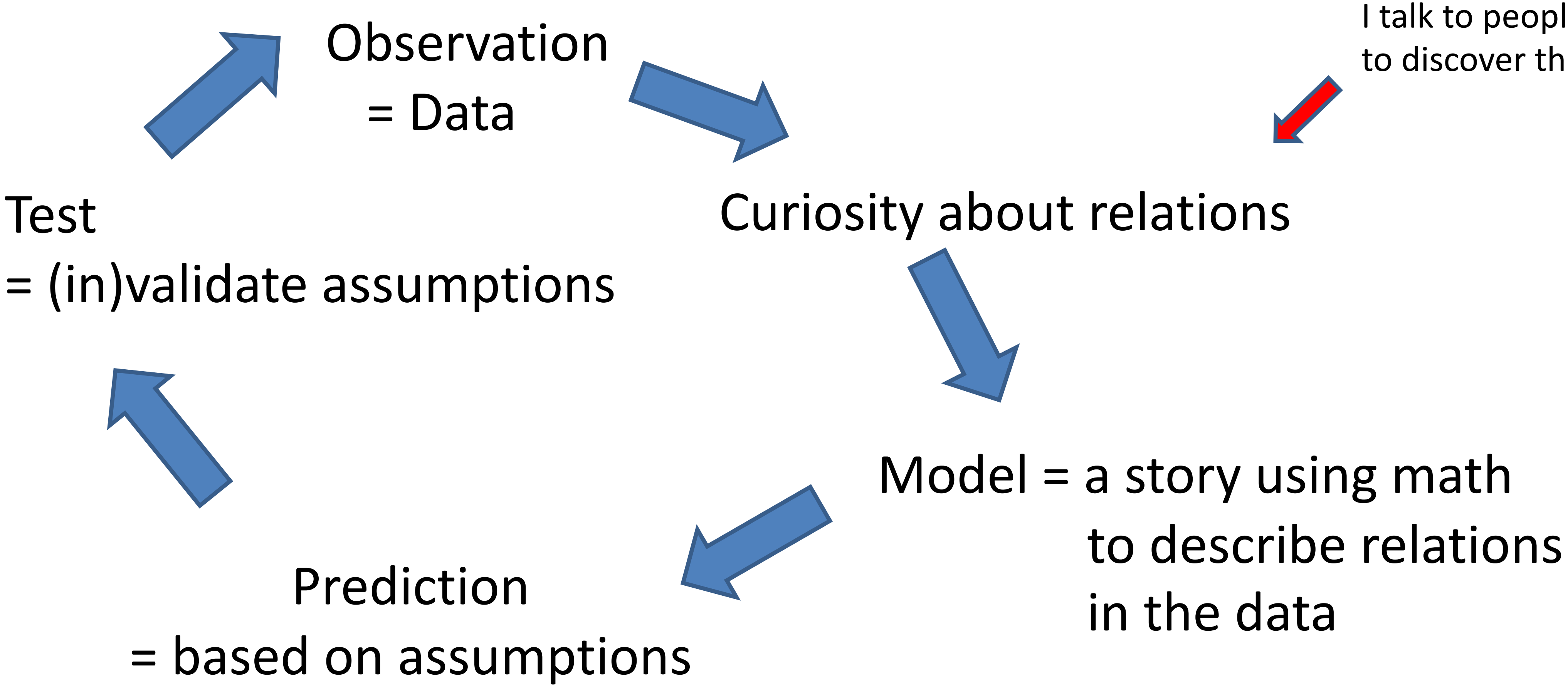
Curiosity about relations

Model = a story using math
to describe relations
in the data

Prediction
= based on assumptions

Test
= (in)validate assumptions

Relevance of folklore in Data Science



Observation
= Data

Curiosity about relations

I talk to people
to discover their folklore

Model = a story using math
to describe relations
in the data

Prediction
= based on assumptions

Test
= (in)validate assumptions

You have an internal model of the world you can leverage when starting to explore data

Activity: sketch your expectation on paper

Axes:

- Time (days)
- Power (megawatts)

Time varying data

- Source: https://www.bmreports.com/bmrs/?q=demand/rollingsystemdemand

Notebook:

visualizing_time_variation_v5_final_product_looks_easy.ipynb

Time varying data

- Source: https://www.bmreports.com/bmrs/?q=demand/rollingsystemdemand

Notebook:

visualizing_time_variation_v5_final_product_looks_easy.ipynb

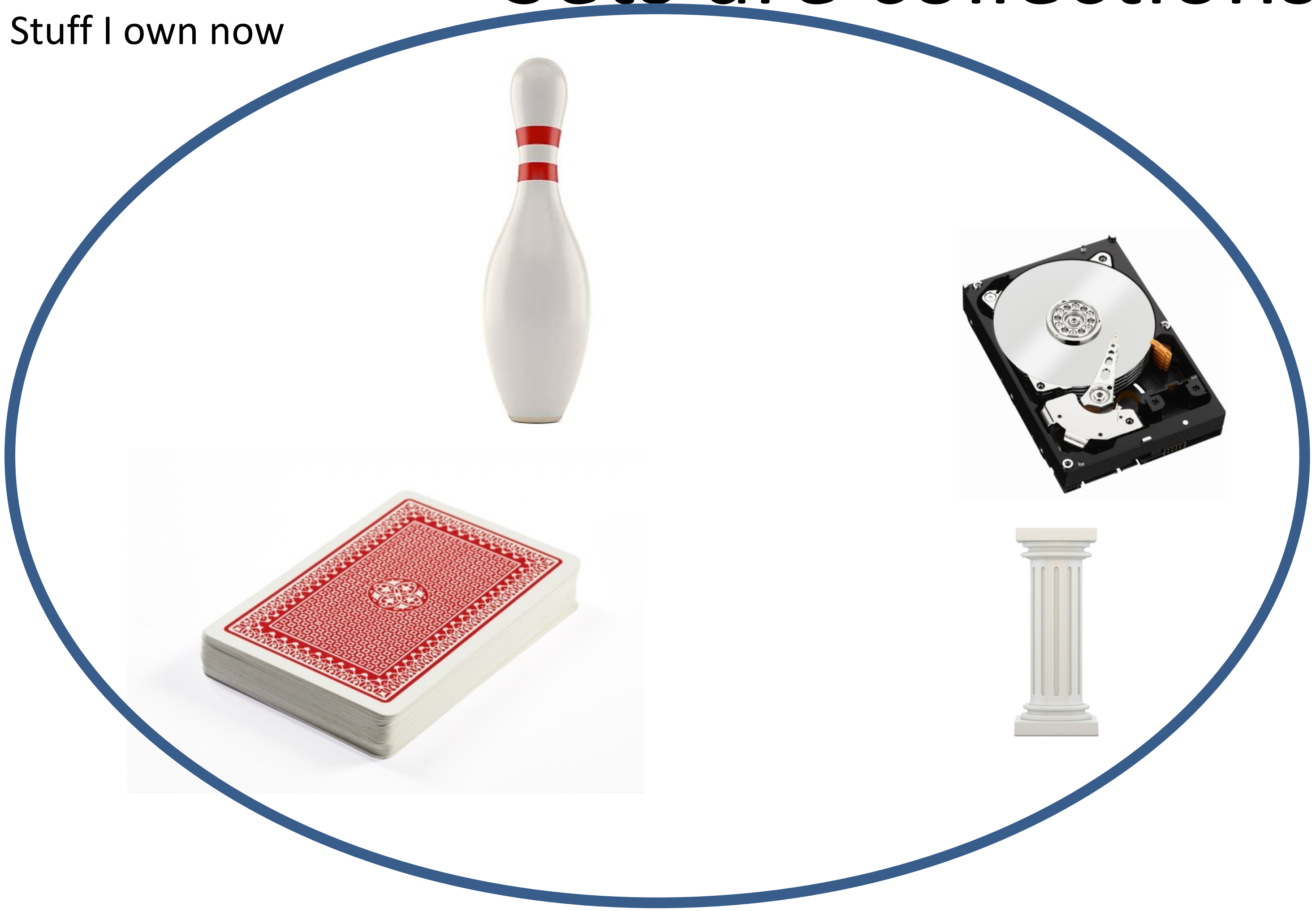
The point of this notebook is that we can tell stories about data just by starting with a scatter plot

In the future, we will discuss a mathematical approach, rather than just relying on visual analysis: Fourier transform

Browser history for two people

- How would you find which websites both have visited?

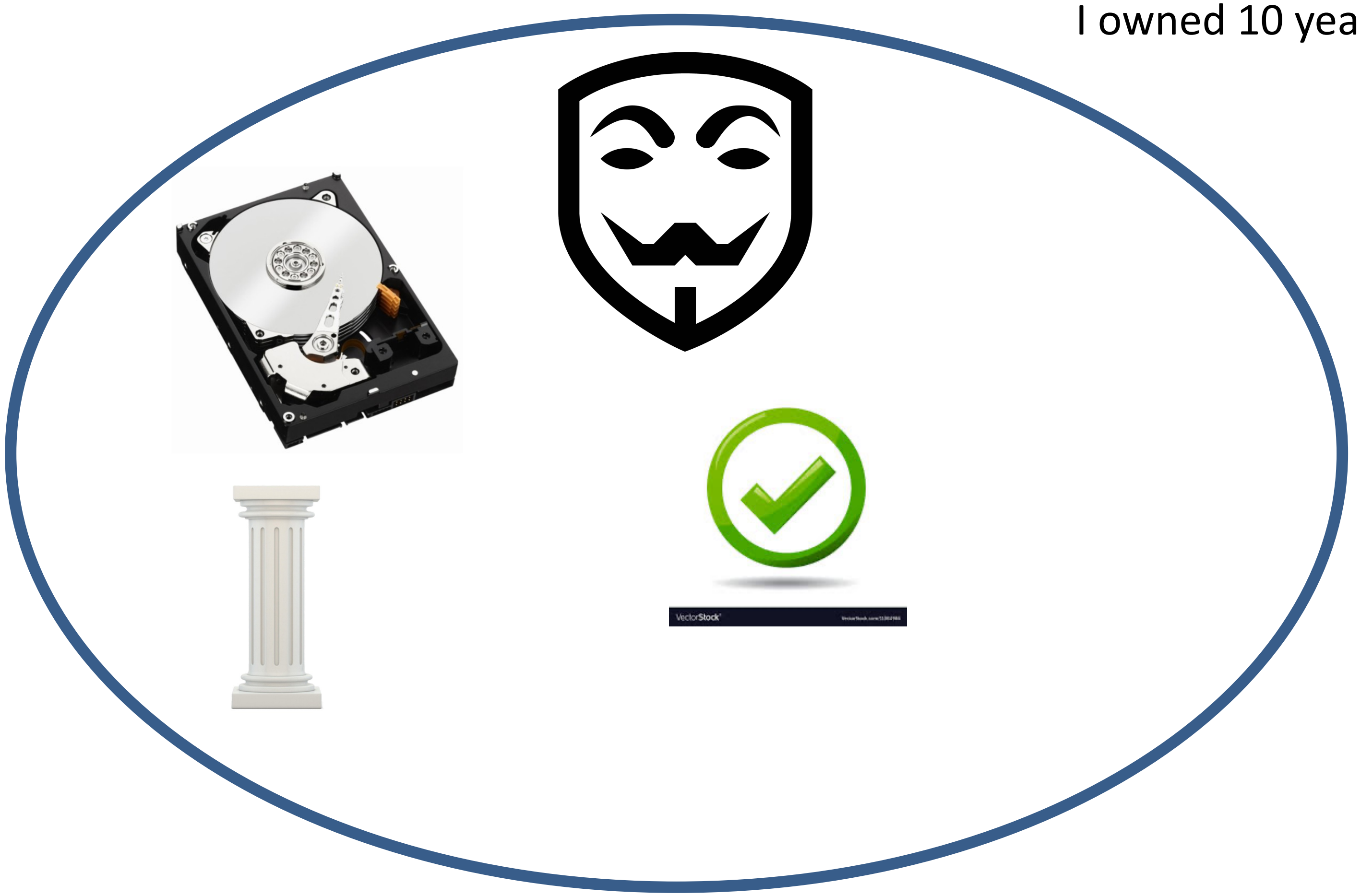
Sets are collections of items



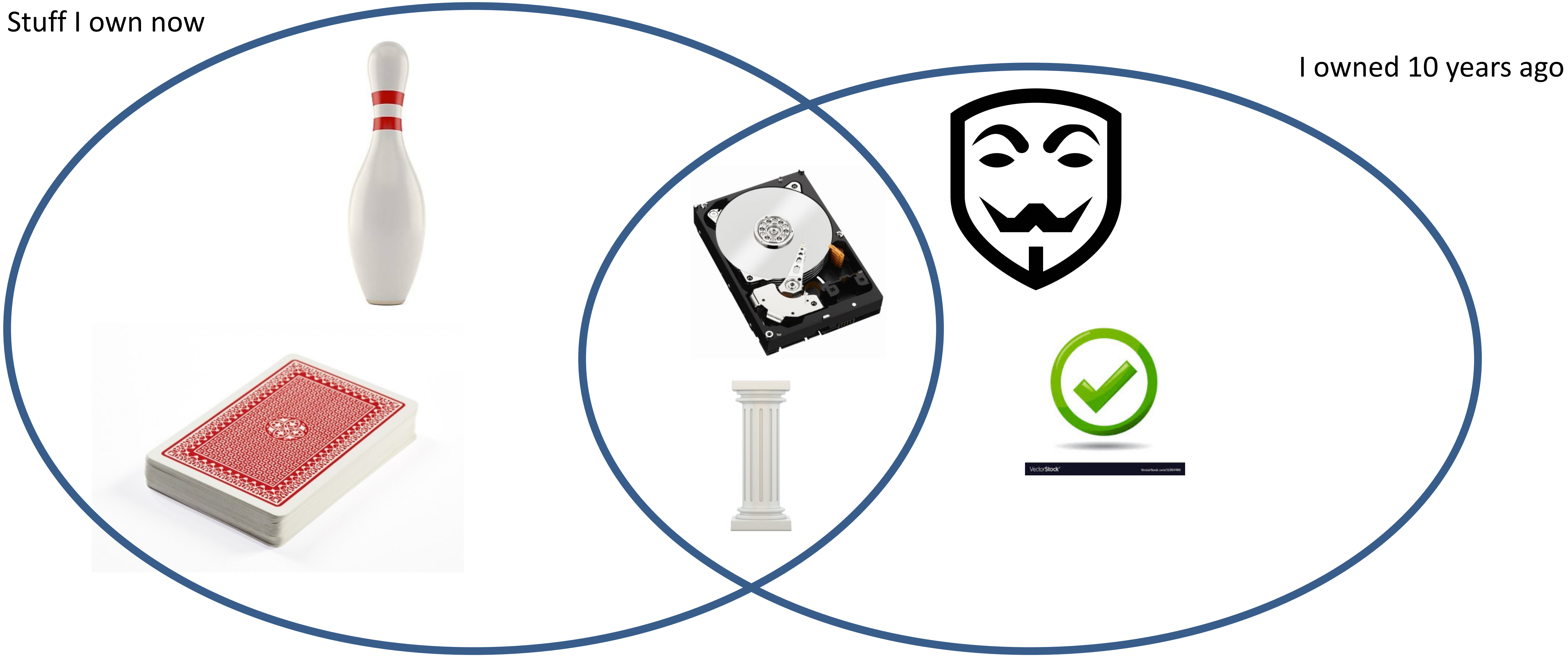
Stuff I own now

Sets are collections of items

I owned 10 years ago



Sets are collections of items



Set intersection (and); Set union (or)

- Intersection = the overlap of one or more sets
- Union = combination of one or more sets

sets.ipynb

Linear Algebra and Numpy

- ~~Math and Data Science~~
- ~~Probability~~
- ~~Correlation~~
- ~~Sets~~
- Linear Algebra and Numpy
- Calculus and Differential Equations
- Visualization
- Homework

Relevance of Linear Algebra to Data Science

Central to machine learning

- Images are arrays of numbers
- Text processing (ie word2vec)
- Neural network weights are stored as array
- Principal Component Analysis (PCA) - lossy dimensionality reduction

https://www.khanacademy.org/math/linear-algebra

Numpy and dimensions of data

NumPy is short for Numerical Python

- Scalar values, ie 5

- 1 dimension
  - Numpy: array – single type of entries
  - Python: list, set (unique elements), tuple (immutable)
  - Linear algebra: vector

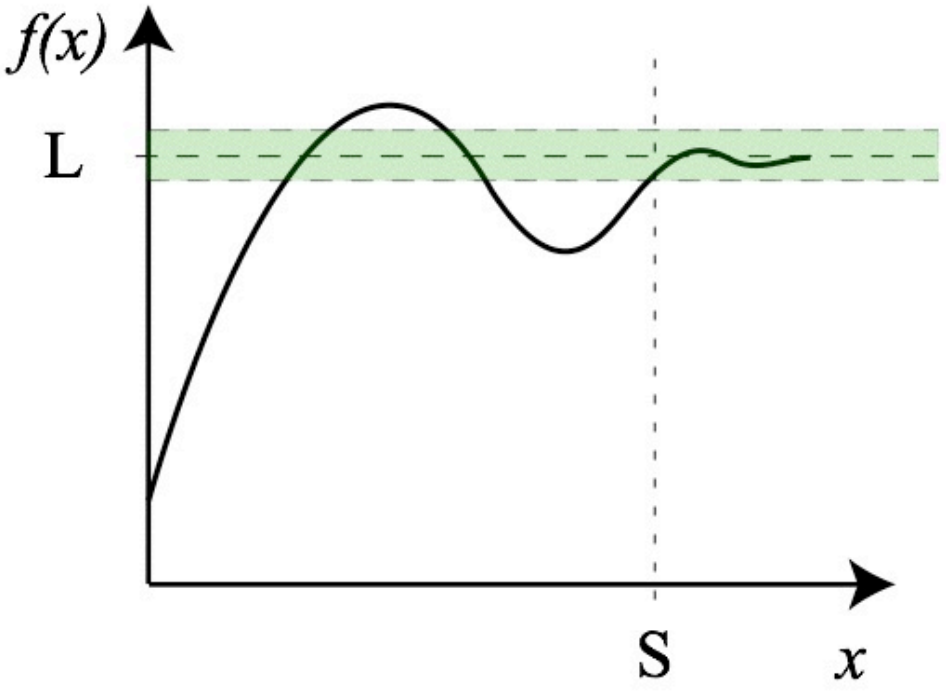
- 2 dimensions:
  - CSV: Table; Excel: worksheet
  - Numpy: array of arrays
  - Linear algebra: matrix

introduction%20to%20numpy%20and%20vectors%20and%20matrices.ipynb

- ~~Math and Data Science~~
- ~~Probability~~
- ~~Correlation~~
- ~~Sets~~
- ~~Linear Algebra and Numpy~~
- Calculus and Differential Equations
- Visualization
- Homework

Calculus as gateway to Differential Equations

- Calculus: approximation that summing a large number of small things yields a finite value.



Curious about what calculus covers?

See this review of concepts and essential concepts.

https://www.khanacademy.org/math/multivariable-calculus

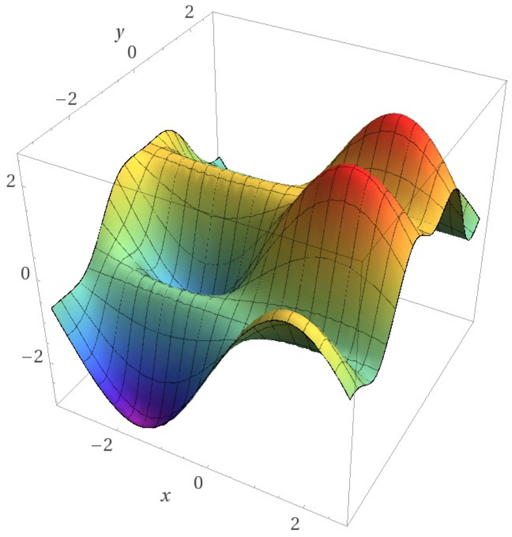
Relevance of Differential Equations in Data Science:

Differential equations enable determination of rate of change of quantities

(When the rate of change is zero, you are bored)

Relevance of Differential Equations in Data Science

- Differential equations enable determination of rate of change of quantities
- Optimization, which is central to machine learning (ie Gradient descent)
- As with most topics in this lecture, there are entire courses and textbooks dedicated to optimization.



Computed by Wolfram|Alpha

There are other applications

Free courses on differential equations are available online

What does a data scientist do all day?

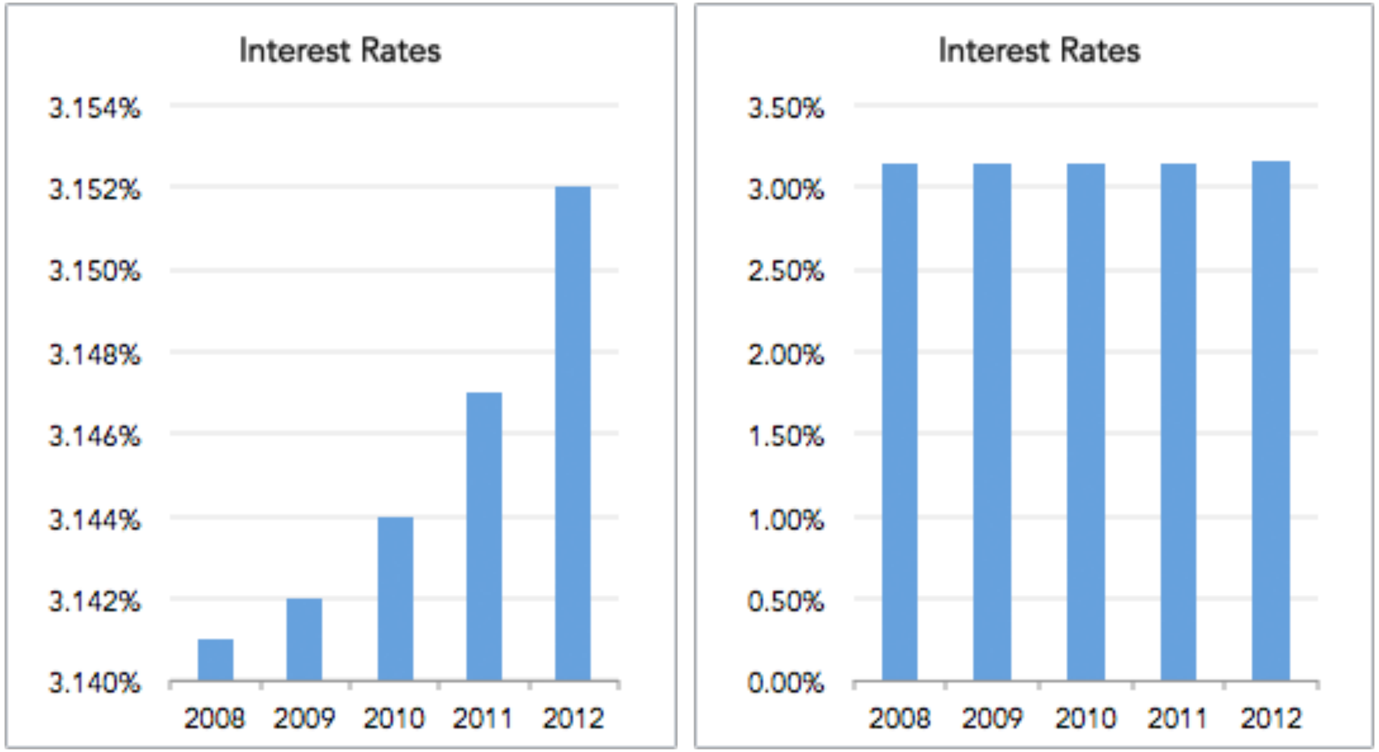
- I have autonomy because my employer expects I know what I'm doing
- Work from 9 to 5 M-F, or 10 to 4, or whatever suits my needs
- Read and write emails.
- Go to meetings; run meetings
- Have impromptu discussions with stakeholders
- Write documentation, read papers
- Write code
- Manage data

No day repeats
No day is predictable
https://www.kdnuggets.com/2017/11/day-life-data-scientist.html

- ~~Math and Data Science~~
- ~~Probability~~
- ~~Correlation~~
- ~~Sets~~
- ~~Linear Algebra and Numpy~~
- ~~Calculus and Differential Equations~~
- Visualization
- Homework

Knowing Math insufficient in story telling: Misleading Visualizations

Same Data, Different Y-Axis

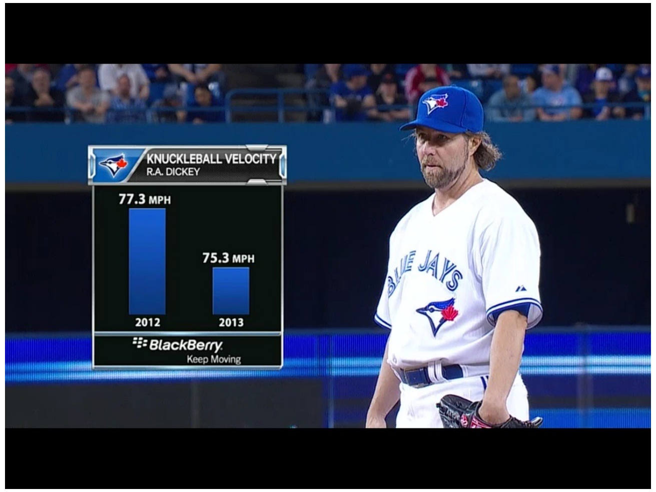


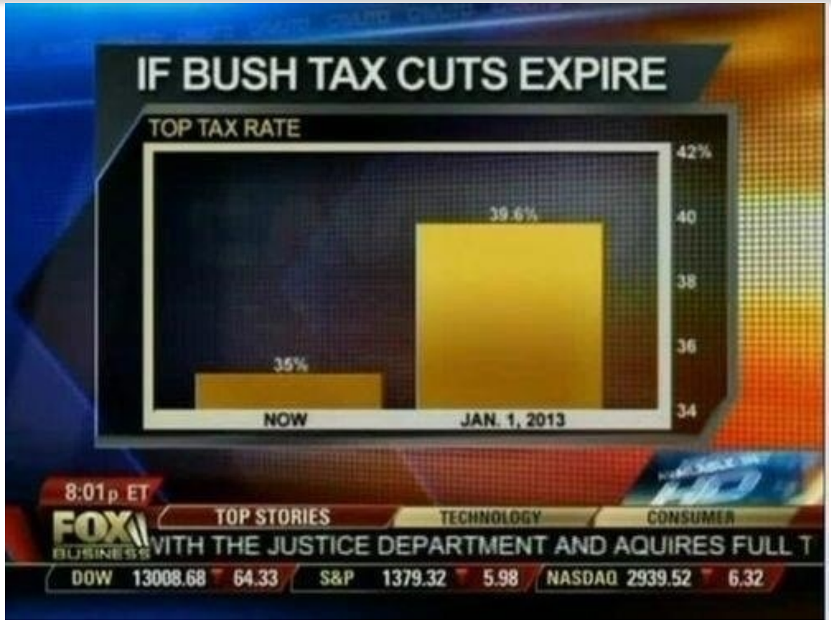
Interest Rates

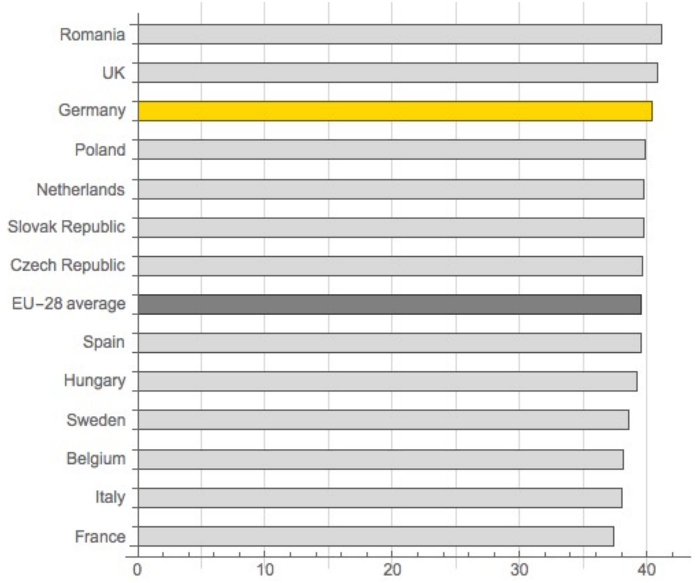
Interest Rates

No one would actually do that, right?

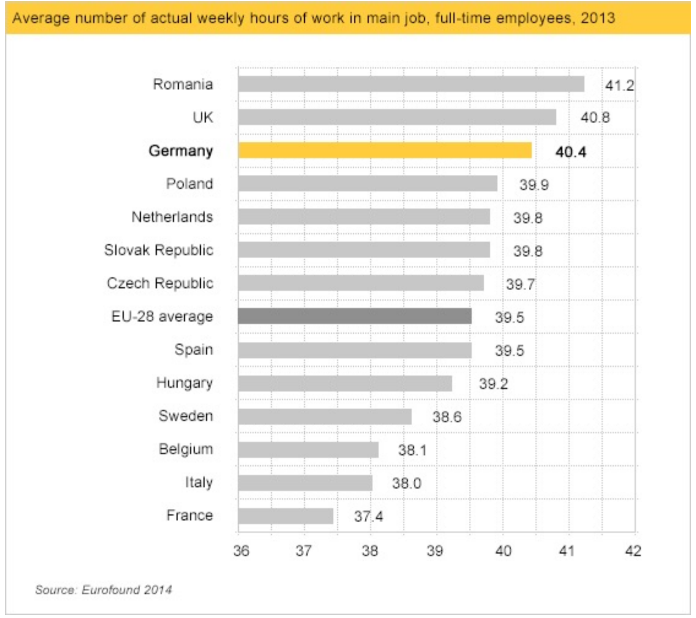
Oh yes







Complete data:



(There are conditions under which not including zero is justified.)

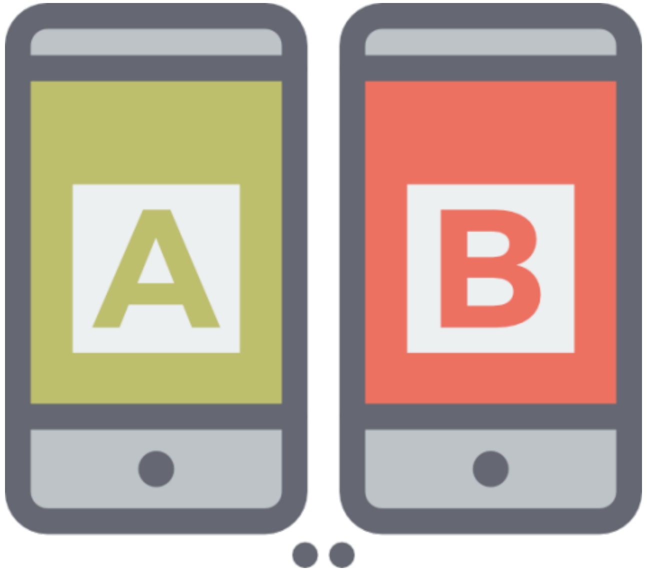
Mistakes when Characterizing data

Even "find the average" isn't as trivial as you may think

- Arithmetic mean : what you're used to
- Harmonic mean : combining multiple parameters which have different ranges so that a given percentage change in any of the properties has the same effect
- Geometric mean : appropriate when the average of rates is desired

There's more than Exploratory Data Analysis

- Comparing models using A/B testing



Where does A/B testing get used?

- Website design
- Email marketing

Useful when the audience being evaluated is sufficiently large to be statistically meaningful

Typically evaluating response with respect to an outcome