# MACHINE LEARNING DATA - INTRODUCTION

# WHO AM I

- Rémy Cazabet (remy.cazabet@univ-lyon1.fr)
- Associate professor, LIRIS Laboratory, Lyon 1 University
- Team: Data Mining and Machine Learning (DM2L)
- Lyon's Institute of Complex Systems (IXXI)

# WHO AM I

- Research topics:
  - Large Network Analysis (Cryptocurrencies…)
  - Graph Clustering
  - Dynamic network
  - Graph Embedding
  - Graph Neural Networks

- Interns application welcomed

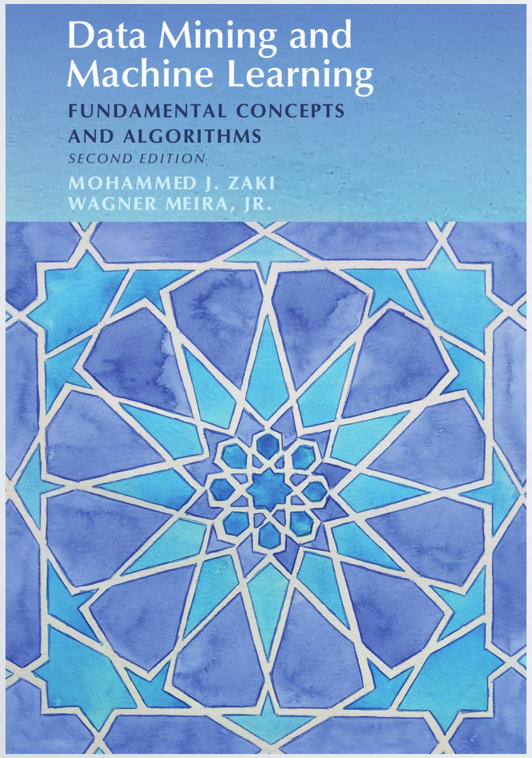
# CLASS OVERVIEW

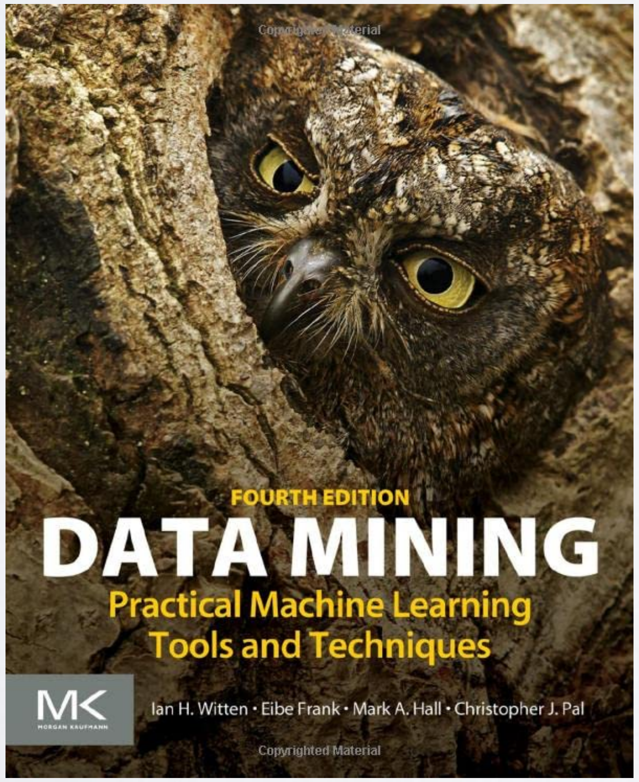
| **Topic** |
| --- |
| Mardi 12 Sep.(9:45-13h) - Introduction, Data Description |
| **Vendredi** 15 Sep.(14h-17h) - Clustering beyond k-means |
| Mardi 19 Sep.(9:45-13h) - Networks 1 - Centralities |
| **Jeudi** 21 Sep.(14:00-17h) - Networks 2 - Community Detection |
| Mardi 26 Sep.(9:45-13h) - Projet |
| Mardi 3 Oct. (9:45-13h) - 18/10: Dimensionality reduction beyond PCA |
| Mardi 10 Oct.(**8:00**-13h) - : Recommendation (TP libre de 8 à 9h45) |
| Mardi 17 Oct. (**8:00**-13:00) - : Frequent Patterns (TP libre de 8 à 9h45) |
| **Mercredi** 18 Oct. (14:00-17h) - : Frequent Patterns / Projet |
| 07/11: Final Exam |

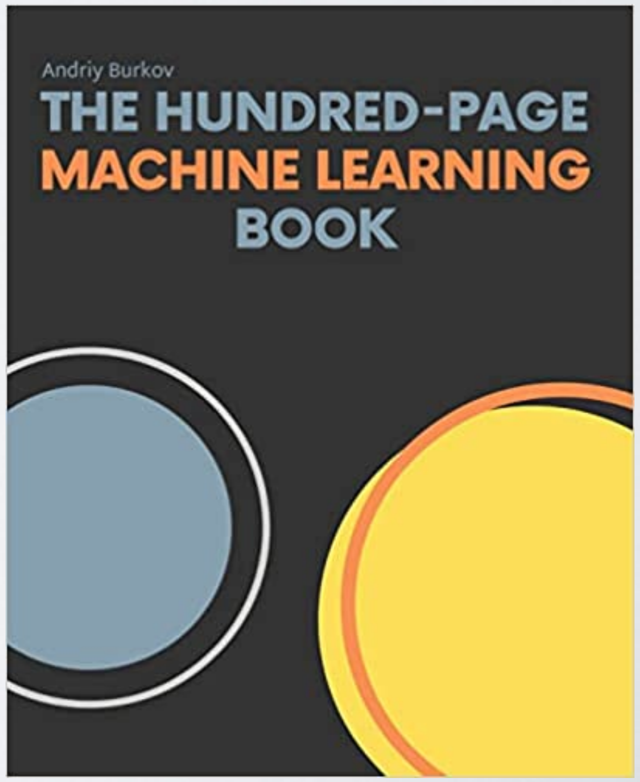
# THIS CLASS

- This class is based on:
  - Countless Wikipedia and blogs (use them too!)

- Some books
  - Borrow at my office



# CLASS OVERVIEW

- Class with me: lecture + practical
- Two other lecturers
- Details on the lecture page:
  - http://cazabetremy.fr/Teaching/DSIA/DM.html
- Exam:
  - Final project 50% (small groups)
  - Final Exam 50%

# TYPES OF DATA

# DATA TYPES

- Data types : What kind of data (feature, variables) can we encounter?
  - People
    - Name, Age, Gender, Revenue, Birth Date, Address, etc.
  - House/Apartment
    - Surface area, Floor, Address, # of rooms, # of Windows, Elevator, etc.

- Types of features?

# DATA TYPES

- Nominal
  - From “names”. No order between possible values
  - Color, Gender, Animal, Brand, etc. (Numbers:Participant ID, class…)
- Ordinal
  - Order between values, but not numeric
  - Size[small, medium, large], [Satisfied, …, Unsatisfied]
- Interval
- Ratio

# INTERVAL

- Numeric values, <u>Difference is meaningful</u>
  - T°: 30°-20° = 15°-5°, But 30° ≠ 2*15°
  - 2022-2020=1789-1787, but 1011 ≠ 2022/2
  - =>0 is not a meaningful value, is arbitrary

  - No multiplicative relation, no ratio => You should not log-transform…
    - Log10: Increasing the value by 1 means multiplying by 10. But multiplying is wrong!

# RATIO

- Numerical values, all operations are valid
  - Height, Duration, Revenue…
  - =>0 means "absence of value".

# OTHER TYPES

- Real Data can have many other forms
  - Textual
  - Relational (networks)
  - Complex objects (picture, video, software…)

# TRICKY CASES

- Real life is complex
- You will have to do modeling choices (feature engineering…)
- Possibles values: Blue, Cyan, White, Yellow, Orange, Red.
  - Nominal or Ordinal ?
- Survey: "rate X on a scale from 0 to 5"
  - What if labels are associated ? ("Bad", "average", …)

# TRAPS

- Latitude and Longitude
- Hours expressed between 0 and 12/24, day of month, etc.
  - Convert in time since beginning of dataset ?
- => Space and Time often handled with specific ML methods

# MISSING VALUES

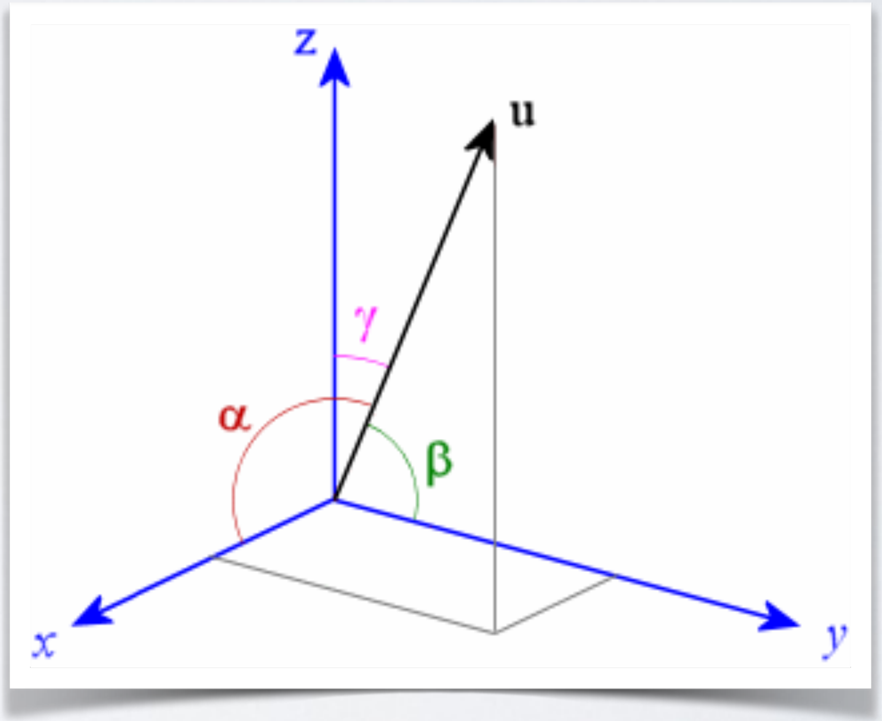
- Real life datasets are full of missing values
  - Impossible data: hair color for a bald person
  - More generally, failed to obtain them
- Few ML methods can deal with missing values
  - =>Imputation
    - Naive: fill with average value
    - Use ML to fill missing values (other problems, introduce biases…)
    - Large literature, no good solution

# DATA QUALITY

- Data coming from the real world is often incorrect
  - Malfunctioning sensors (T°, speed…)
  - Human error or falsification (e.g., entered 100 instead of 1.00)
  - Undocumented change (e.g., Bicycle sharing station was moved…)
- If the data is plausible, no simple solutions
- Two common problems can be detected
  - Out-of-range values (e.g., a person's weight is negative or above 1000kg…)
  - Zeros. (Weight of the person is 0. But in many cases, zero is possible too…)
    - Variant: 01/01/1970…

# UNIVARIATE / MULTIVARIATE

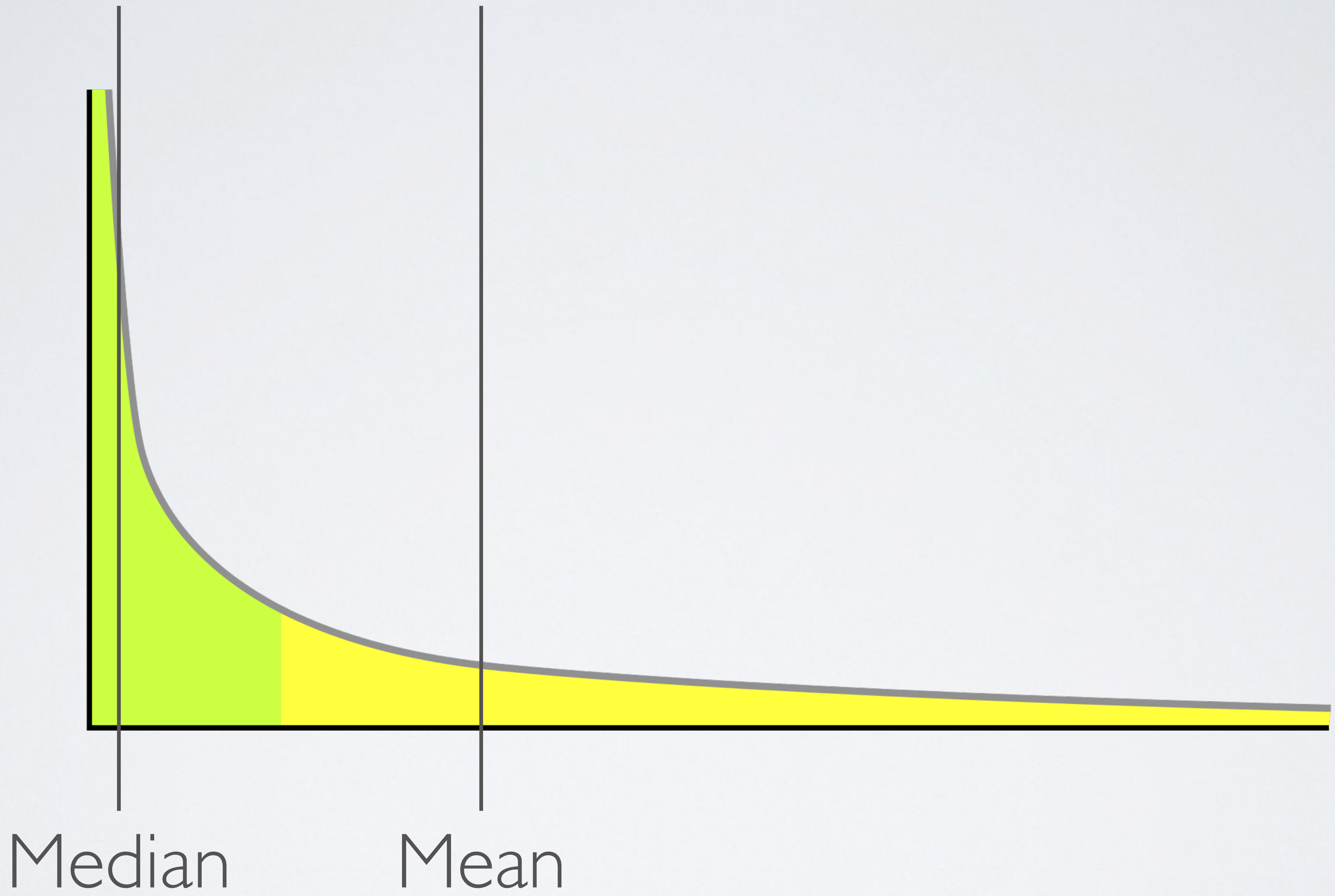
- Single *feature*: univariate
  - Age
- Real life: multivariate.
  - 2D (age, weight)
  - 3D (age, weight, height)
  - 4D (age, weight, height, genre)
  - …



# DESCRIBING A VARIABLE

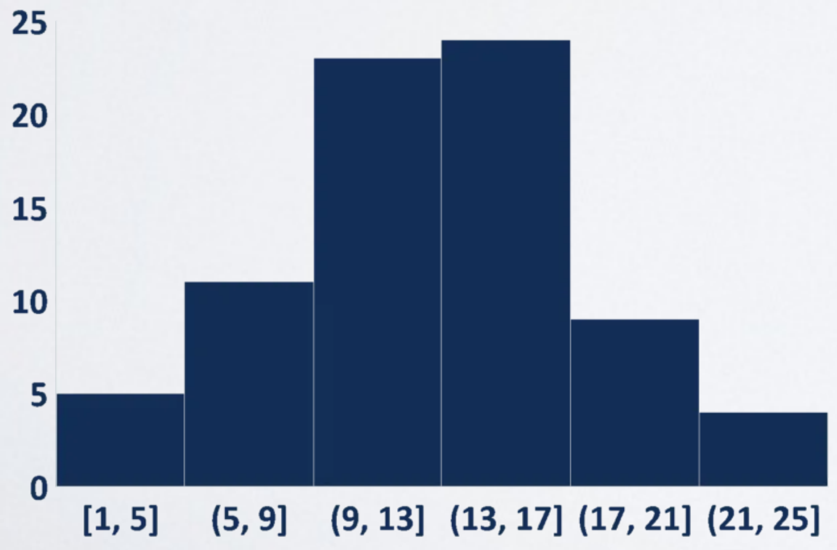
# DESCRIBING VALUES

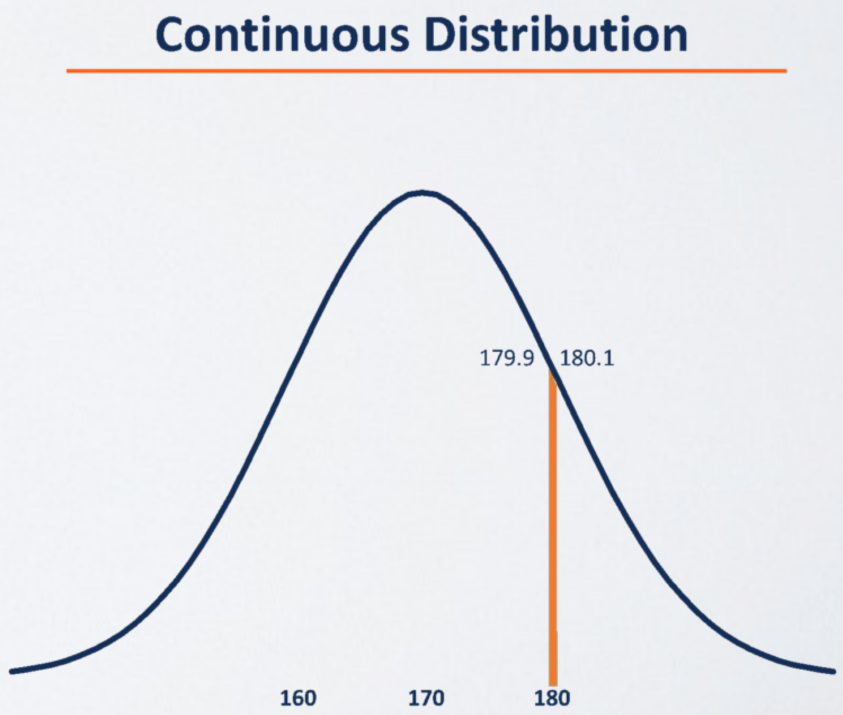
- Mean / Average
  - Be careful, not necessarily representative !
- Median
  - Be careful, not necessarily representative !
- Mode
  - Not necessarily representative
- Min/Max
  - …


Median
Mean

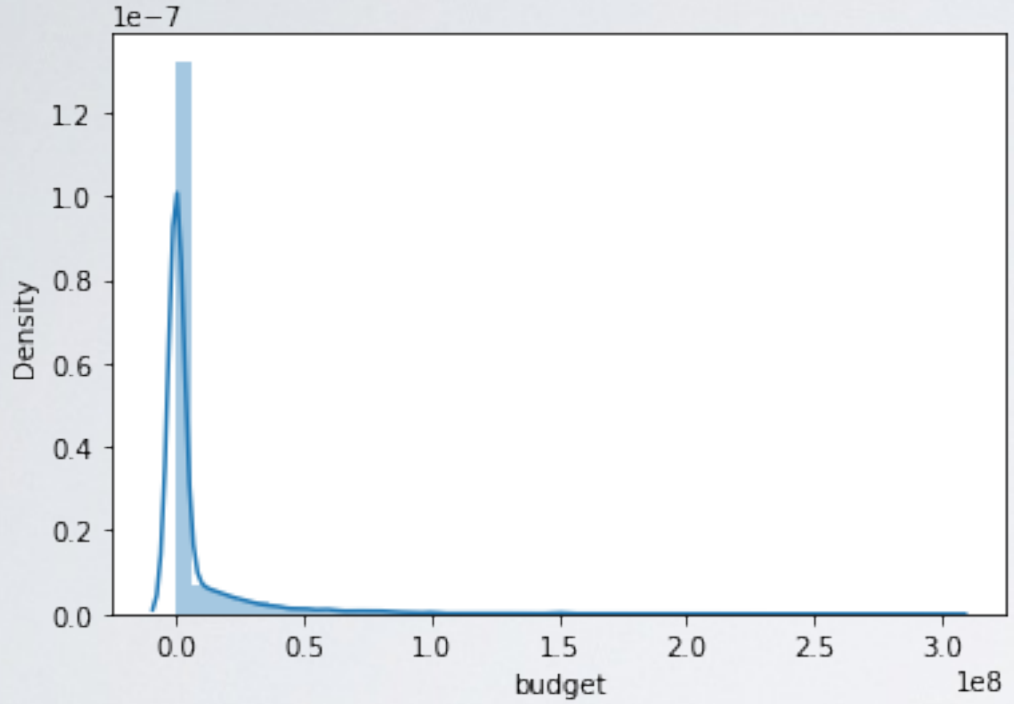
# DISTRIBUTION

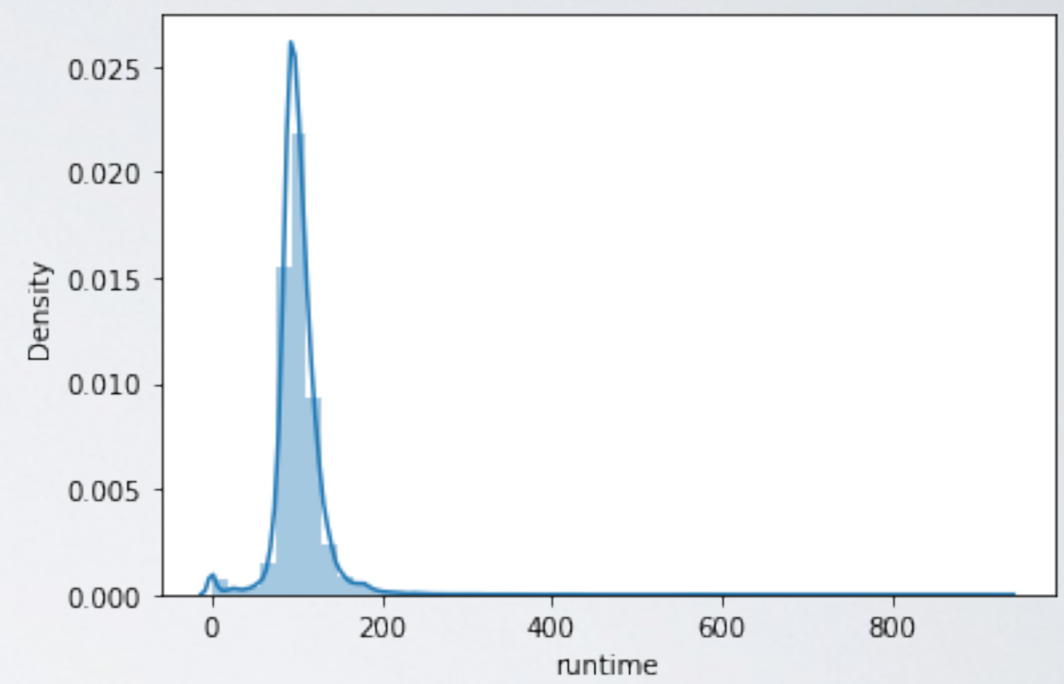
- What is a distribution?
  - A description of the frequency of occurence of items
  - A generative function describing the probability to observe any of the possible events
  - Discrete or continuous

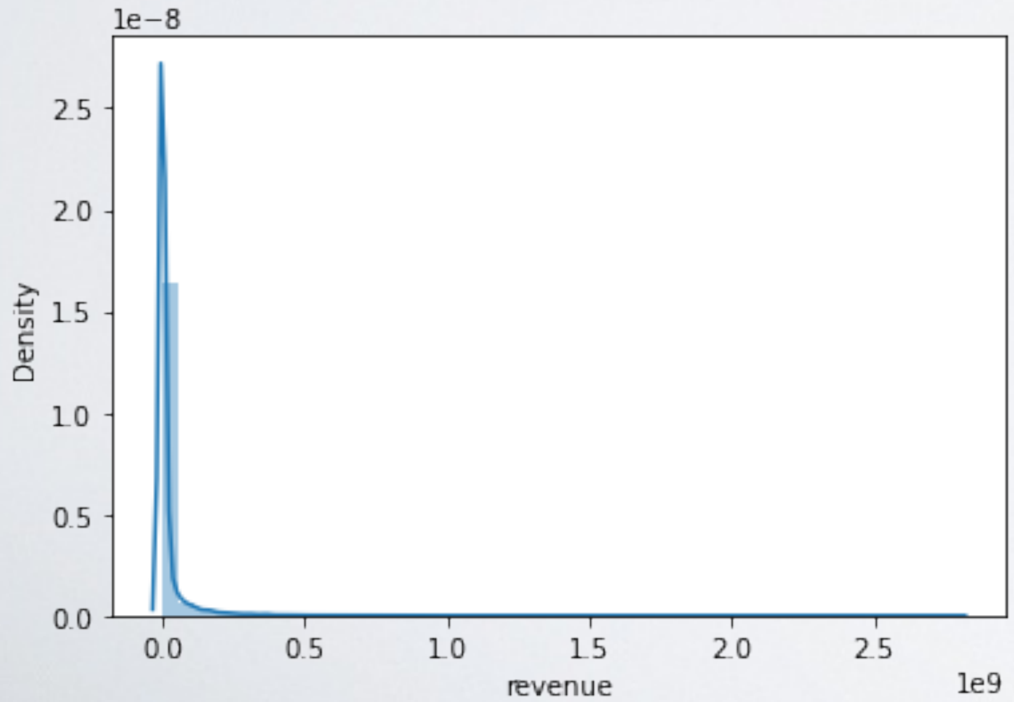


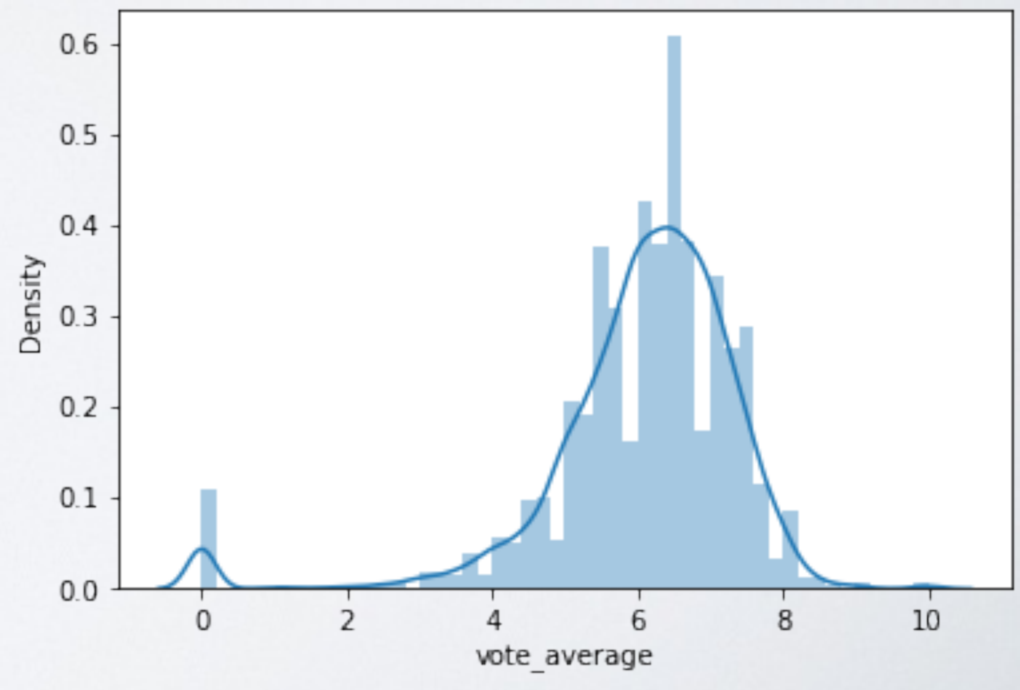



# EMPIRICAL DISTRIBUTIONS



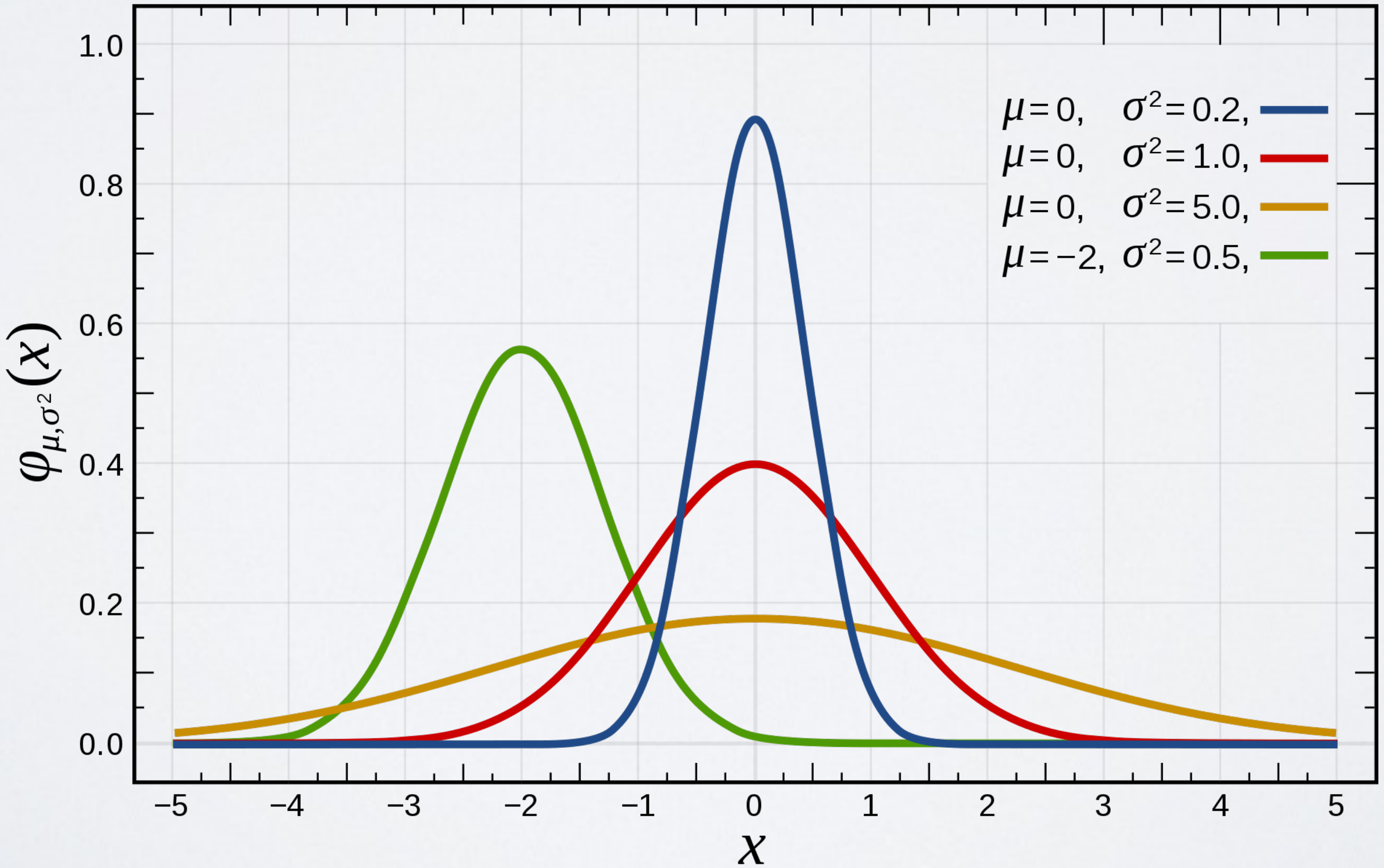



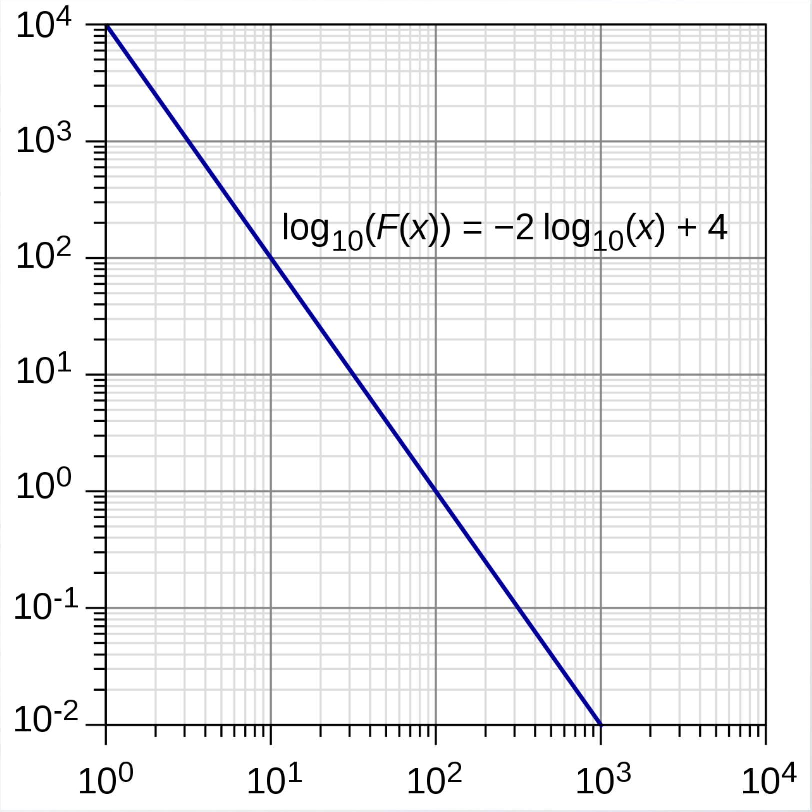
# THEORETICAL DISTRIBUTIONS

- Normal distribution
  - Many real variables follow it approximately (height, weight, price of a given product in various locations…
  - Random variations around a well-defined mean
  - Central limit theorem: <u>average</u> of many samples of a random variable converges to a normal distribution



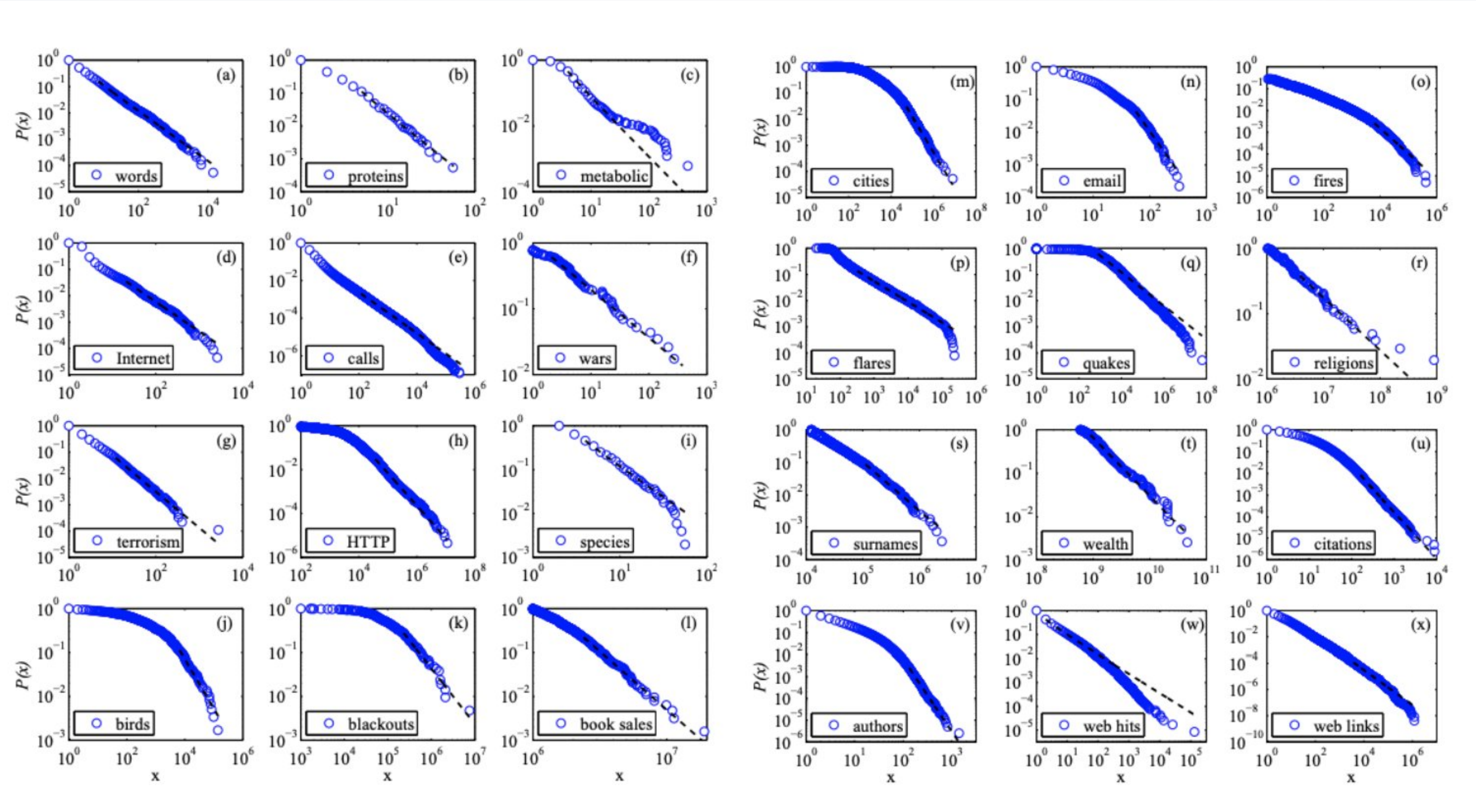
# THEORETICAL DISTRIBUTIONS

- Power Law distribution
  - A relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another.
    - e.g., earthquakes 10 times more powerful are $x$ times less frequent.
    - e.g., cities 10 times bigger are $x$ time less frequent



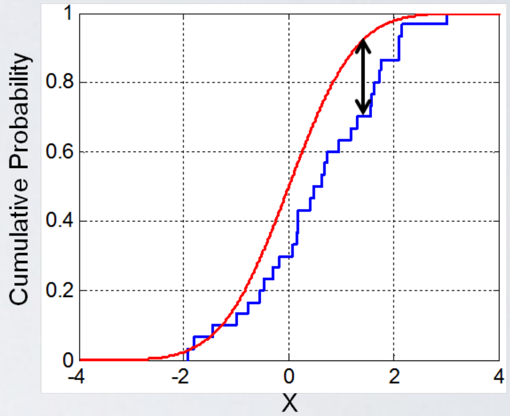
# THEORETICAL DISTRIBUTIONS

- Power Law distribution

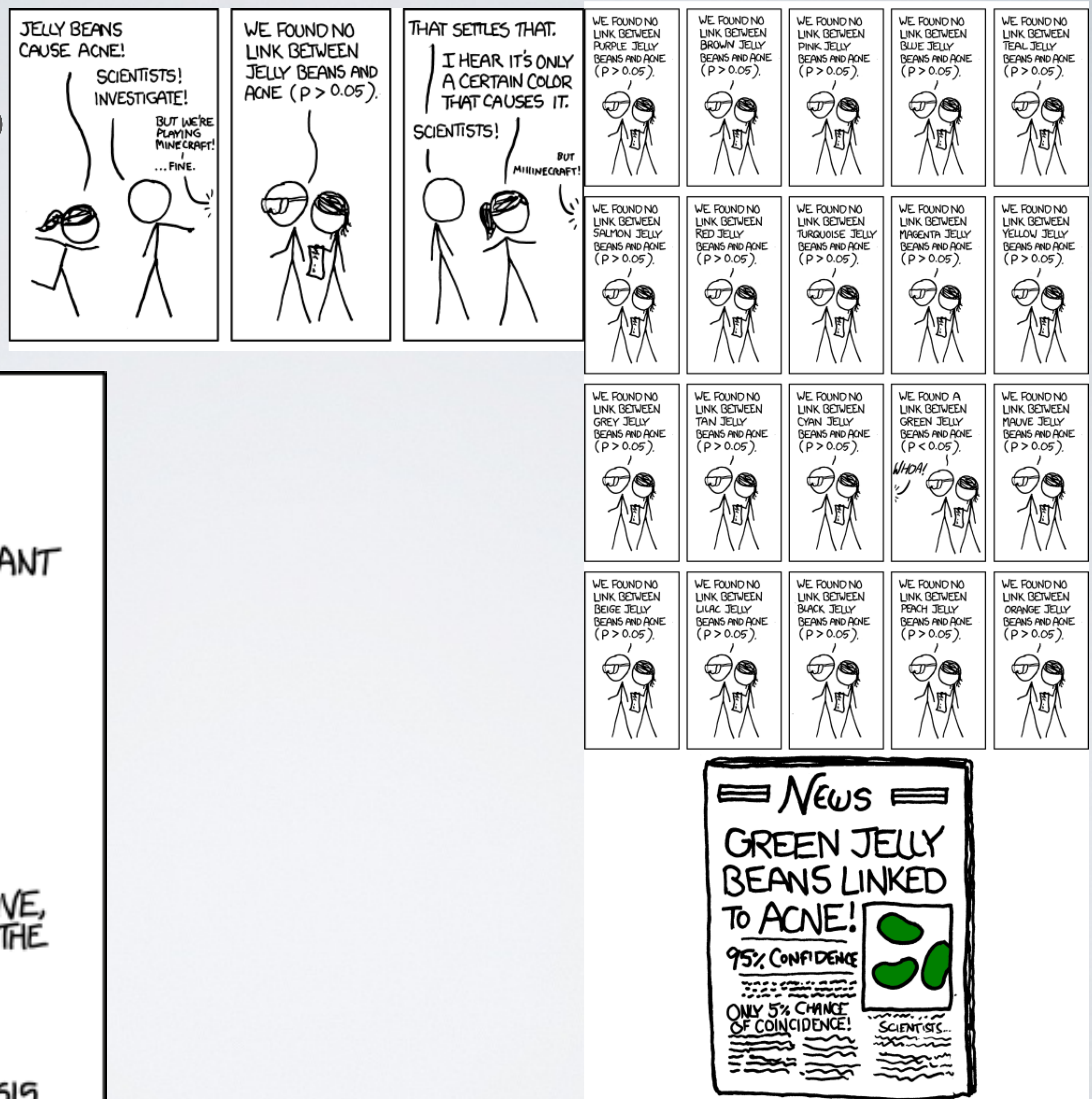


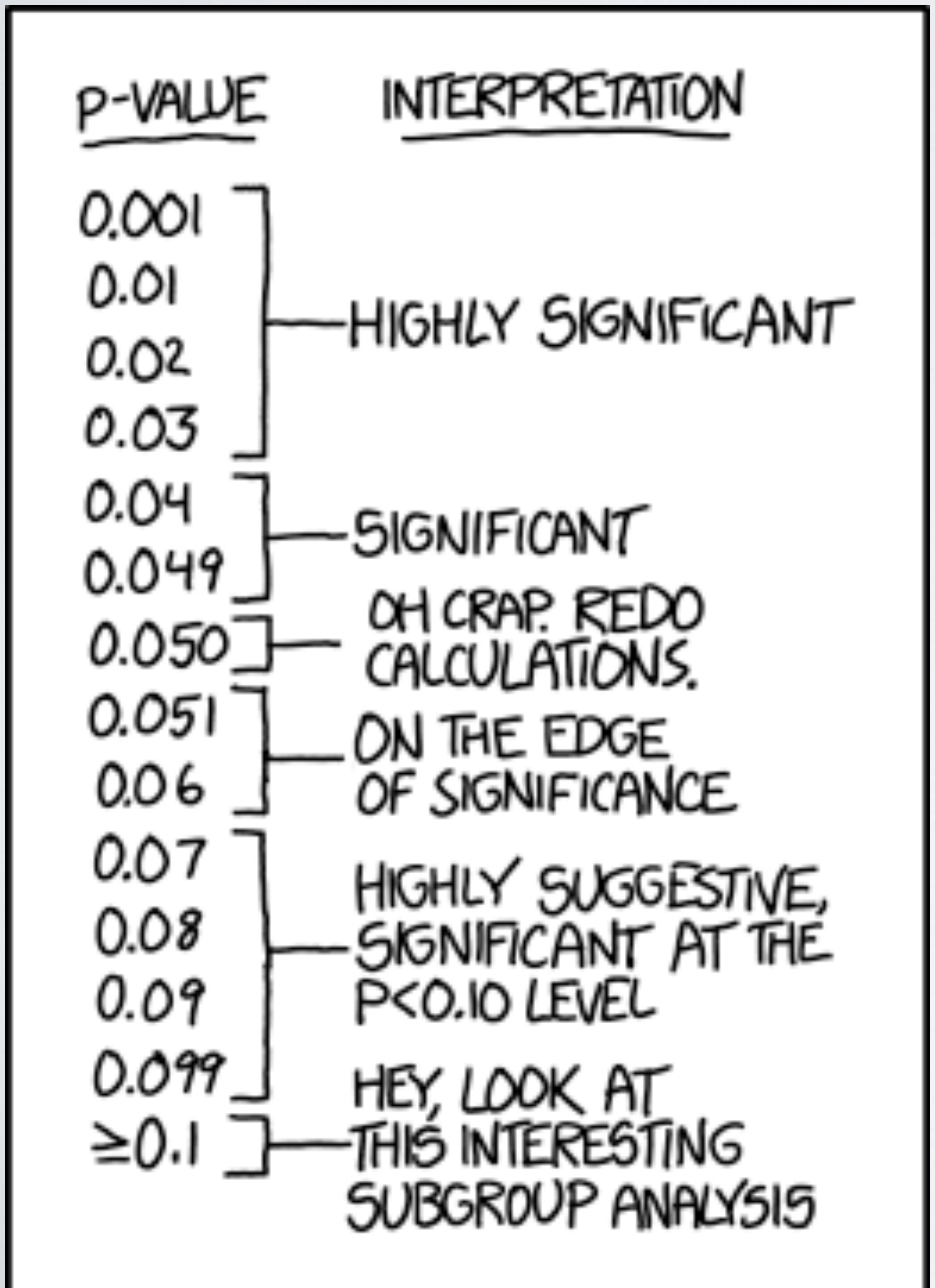
# DISTRIBUTION COMPARISON

- Statistical test
  - P-value: **The probability** that my observed data could be observed if it were generated by the theoretical distribution XXX (null hypothesis)
    - Normality: Shapiro-Wik, etc.
    - Categorical variables : Chi-squared $\chi^2$
    - Etc. (search for the right test if you need it)
  - High p-value: high probability to come from the null hypothesis
    - We usually set a p-value threshold, i.e., 0.05. (5% chance)
    - IF the p-value is below it, **I can conclude** that it is unlikely that my data has been generated by this exact null model (I can never be 100% sure)
    - IF the p-value is above, I can say that it is *possible* that it has been generated by it. However, it could also have been generated by another null hypothesis that I have not tried. **I cannot conclude.**



# P-VALUES


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

# VARIANCE

- Variance:
  - Expectation of the squared deviation of a random variable from its mean

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}\left[(X-\mu)^2\right]$$

Also expressed as average squared distance
between all elements

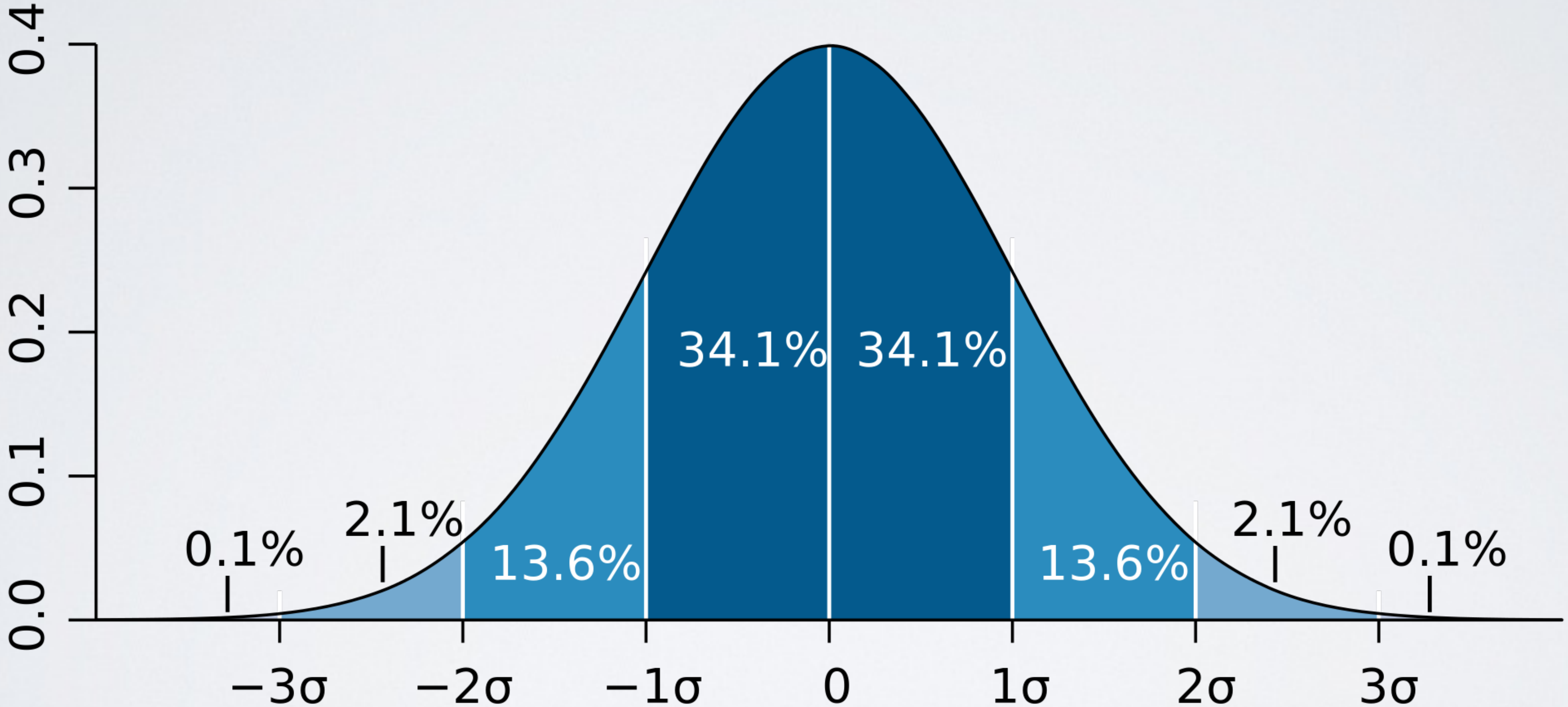
$$\sigma^2 = \frac{1}{N^2}\sum_{i<j}\left(x_i - x_j\right)^2$$

# STANDARD DEVIATION

- Squared root of the Variance

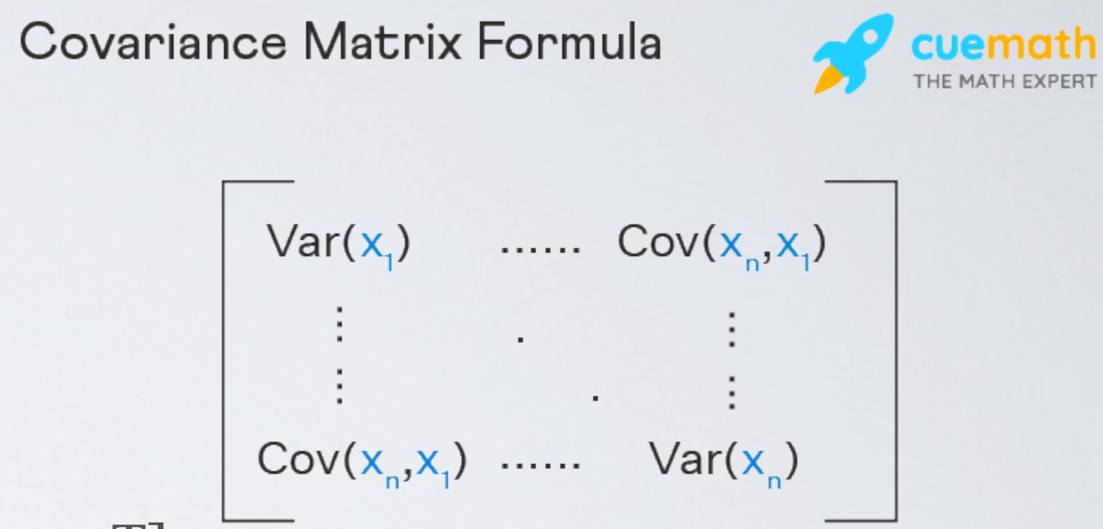
$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{E}\left[(X - \mu)^2\right]}$$

# RELATION WITH NORMAL DISTRIBUTION



# VARIABLE INTERACTIONS

# COVARIANCE MATRIX



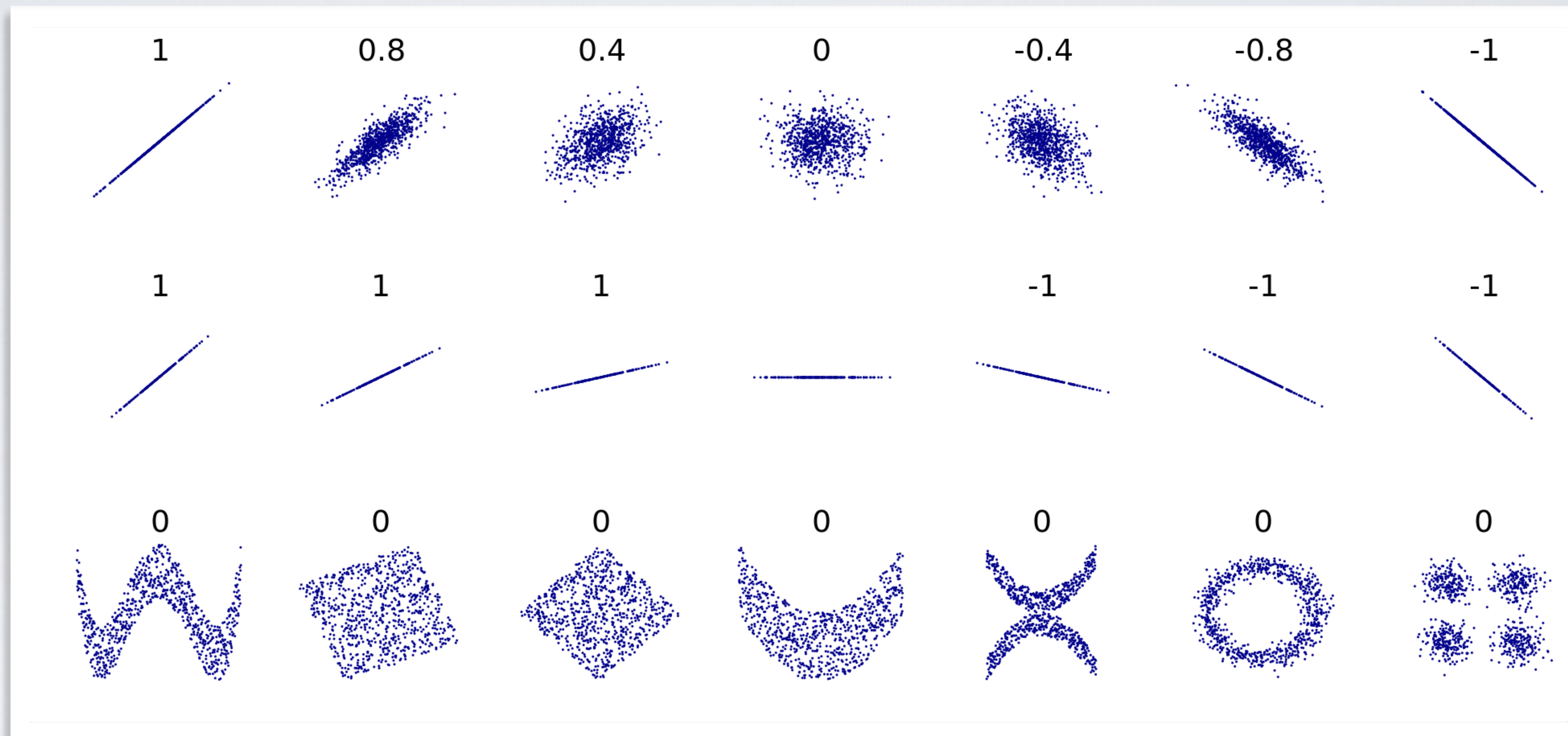

- Covariance matrix $\mathbf{K}$
  - Extension of Variance to multivariate data
  - $\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right]$
  - $\mathrm{cov}(\mathbf{X},\mathbf{Y}) = \mathrm{K}_{\mathbf{XY}} = \mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{Y}-\mathrm{E}[\mathbf{Y}])^{\mathrm{T}}\right]$
    - How much observation X differs from the mean ? And Y ?
    - Multiply the respective divergences of X and of Y for each item
    - Take the average
  - => $\mathrm{cov}(\mathbf{X},\mathbf{X}) = \mathrm{Var}(\mathbf{X})$

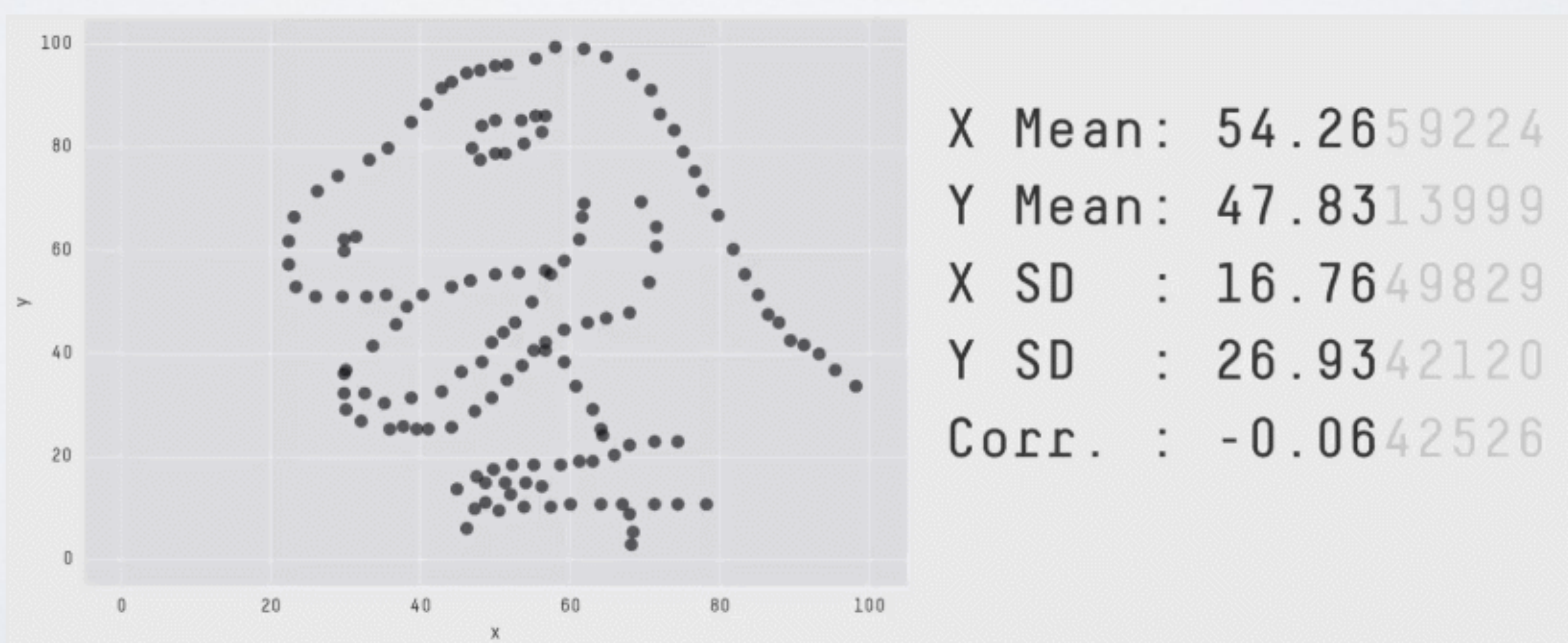
- Covariance is hardly interpretable by itself.
  - If >0, divergences tend to be in the same direction
  - Normalize it to obtain the “correlation coefficient”

# CORRELATION COEFFICIENT

- Pearson correlation coefficient : $\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$
  - Normalize the Covariance by the Standard deviation.
  - Independent from magnitude, i.e., no need to have normalized data
  - Value in -1, +1.
    - +1 means a perfect positive linear correlation, i.e., X=aY
    - -1 a negative one, i.e., X=-bY
  - 0 can mean many different things

# CORRELATION COEFFICIENT

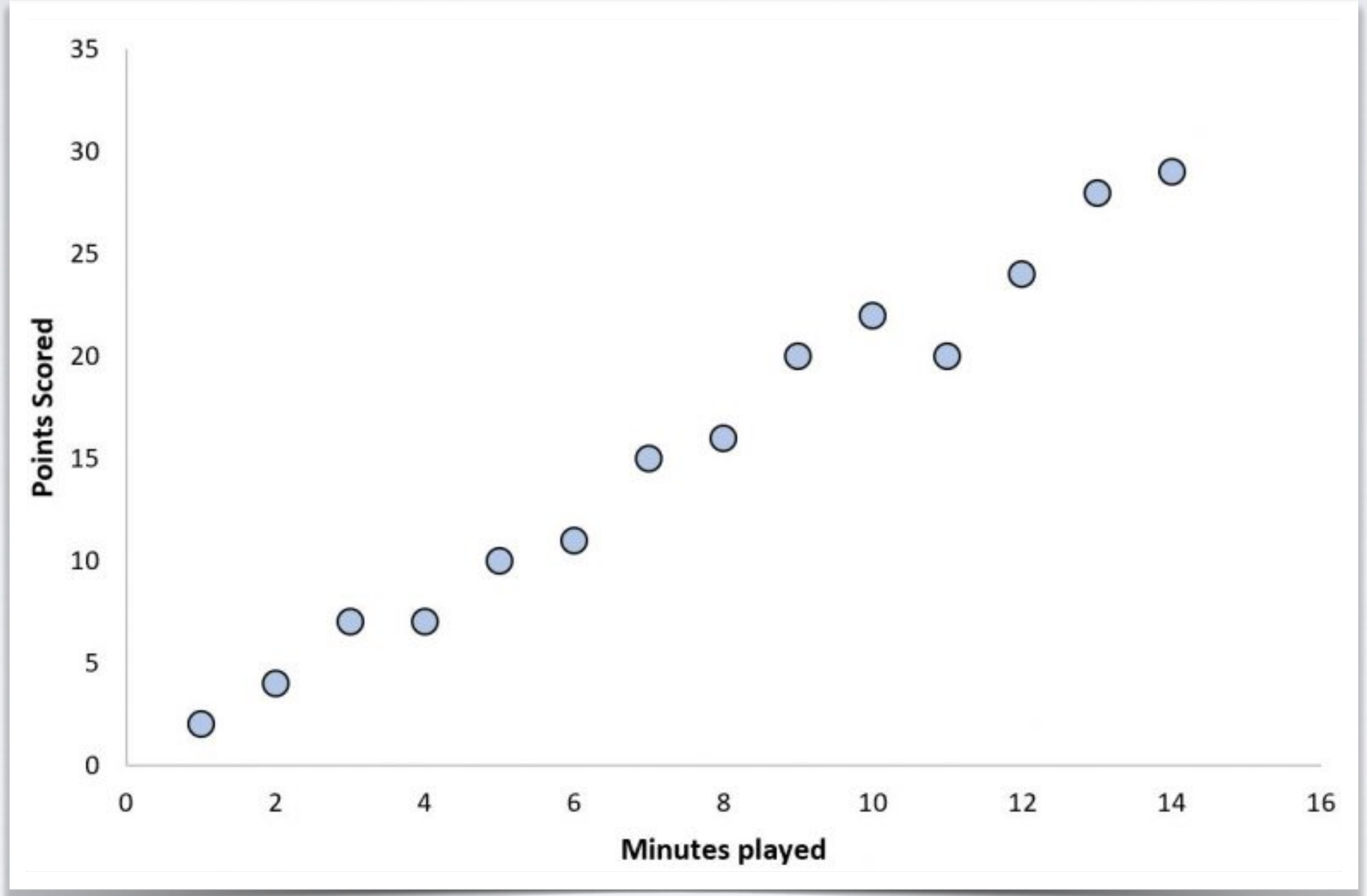


# CORRELATION COEFFICIENT

- Other possible interpretation, e.g.
  - Cosine similarity of the vectors defined by the observations…
- 0.7 ? Is it a high or low value ?
  - It depends.

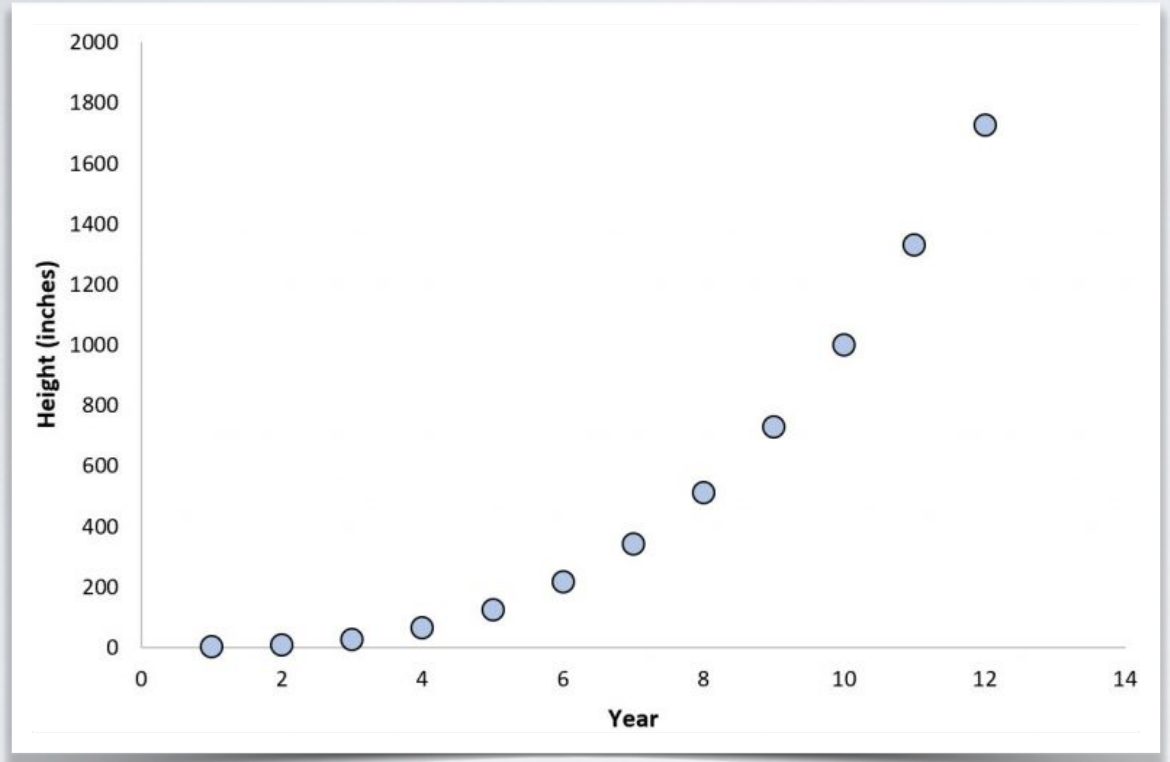
# NONLINEAR RELATIONSHIPS

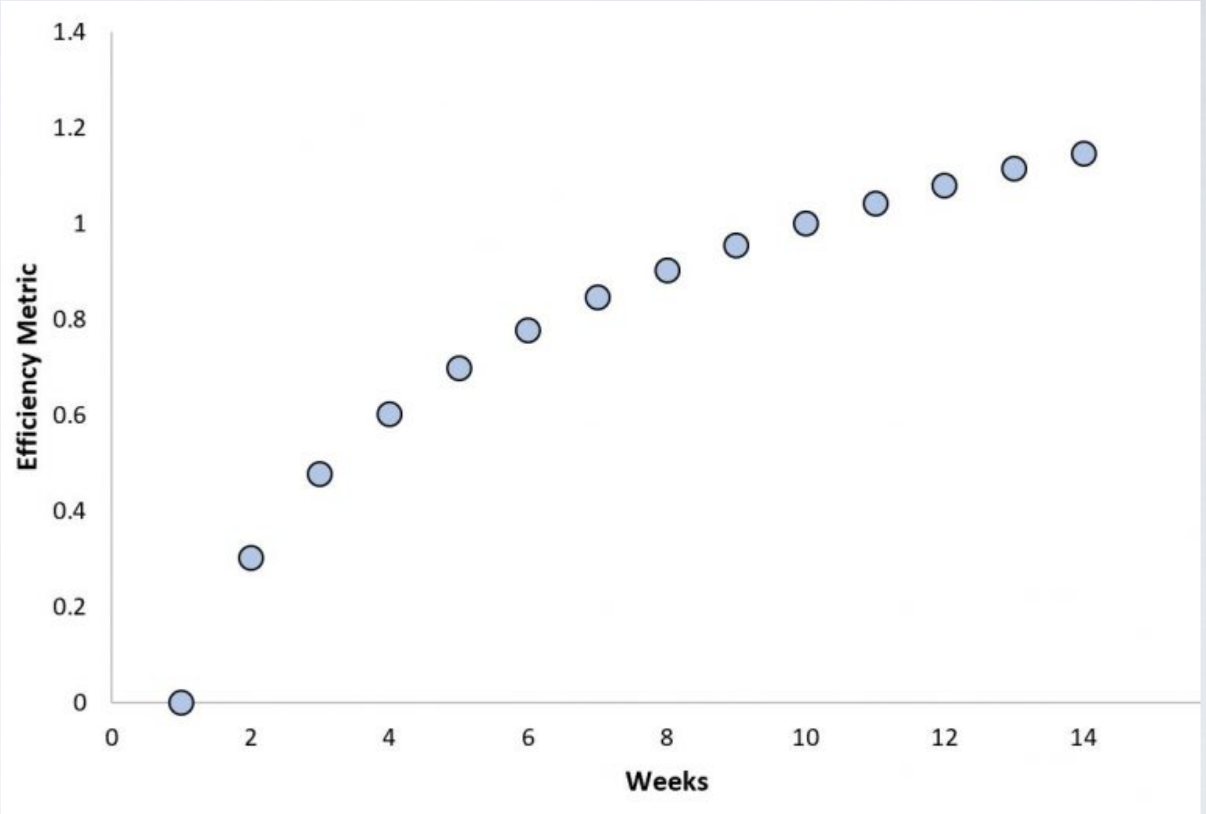



Linear relationship

$Y=a+bX+e$

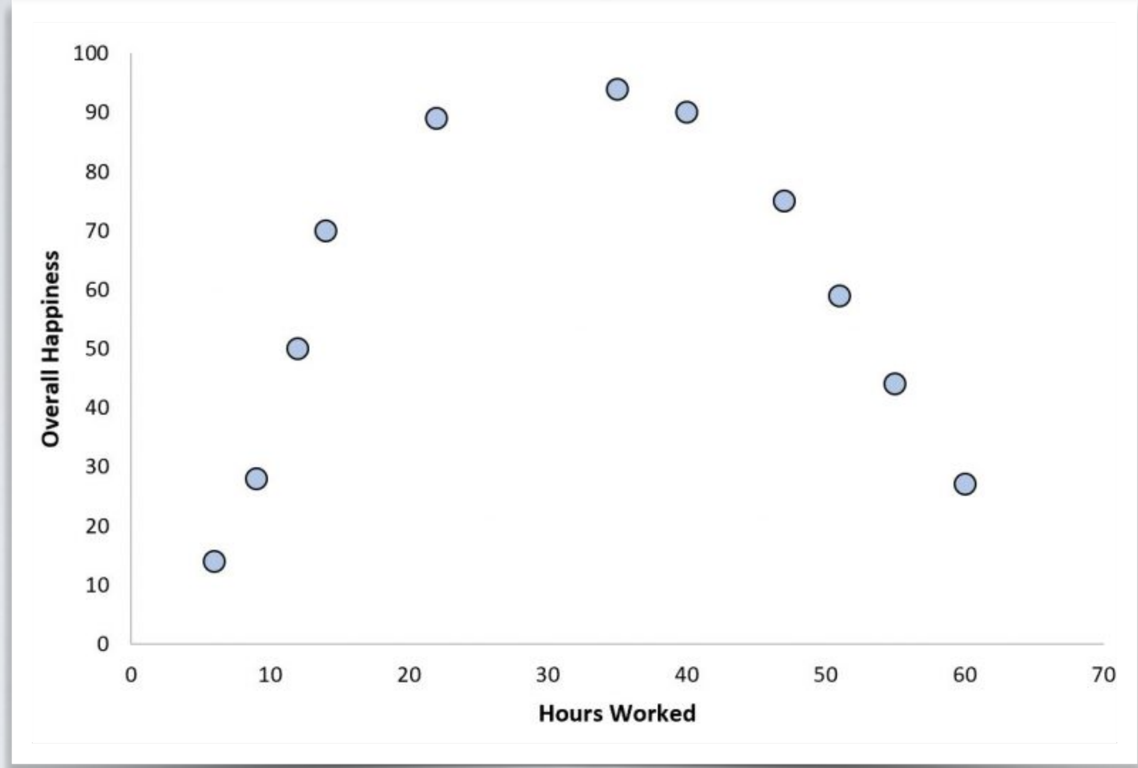
# NONLINEAR RELATIONSHIPS

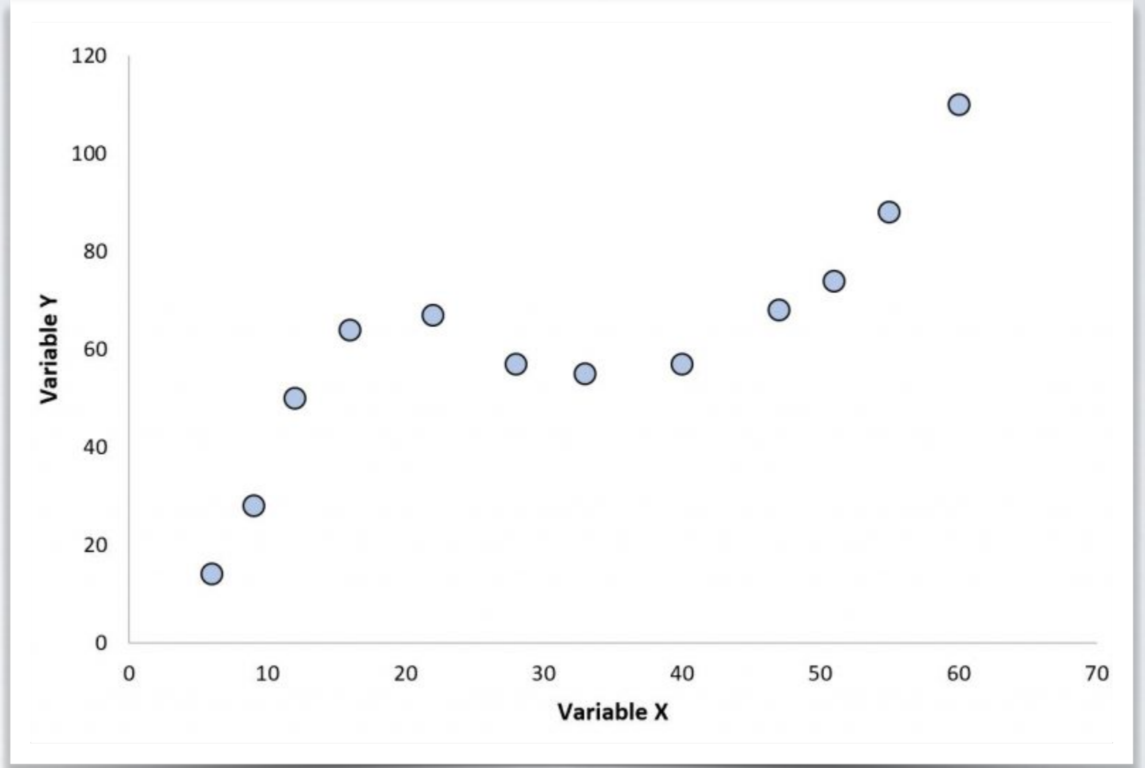


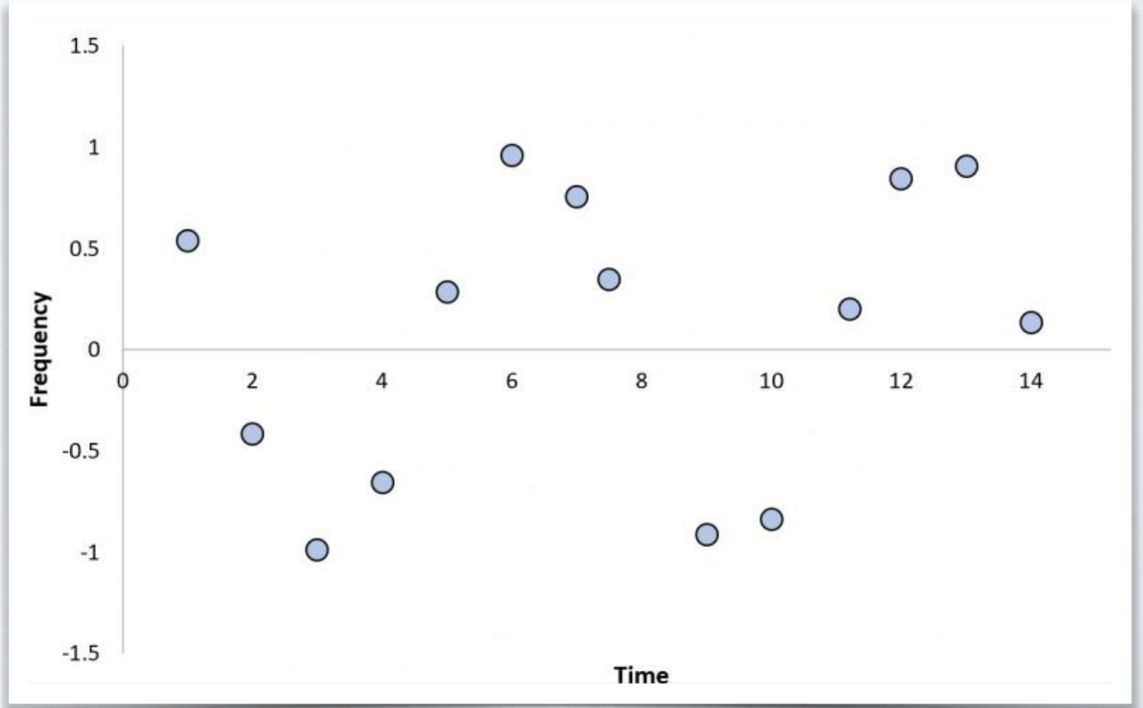

Monotonous, non-linear

# NONLINEAR RELATIONSHIPS



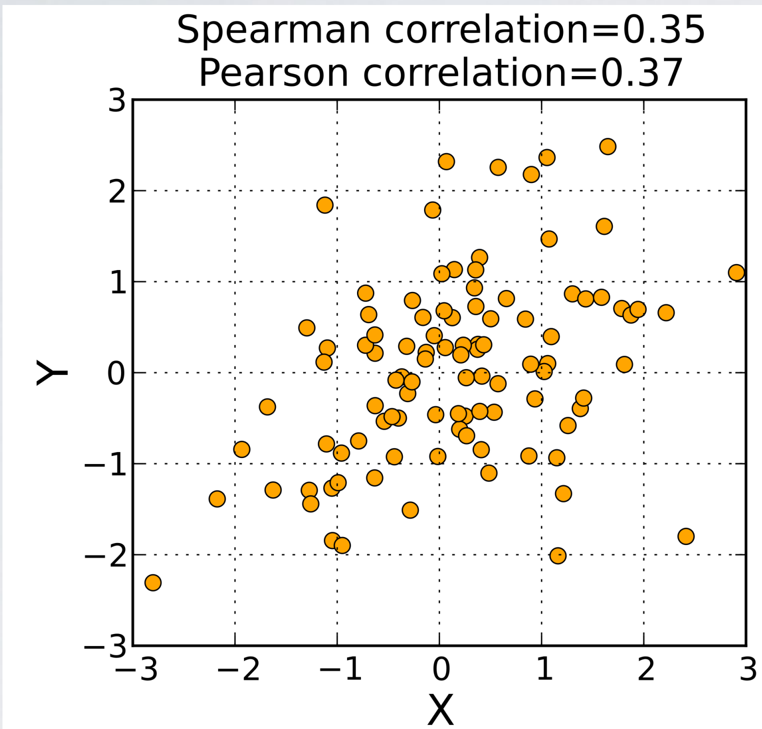

Non-monotonous,
Non-linear

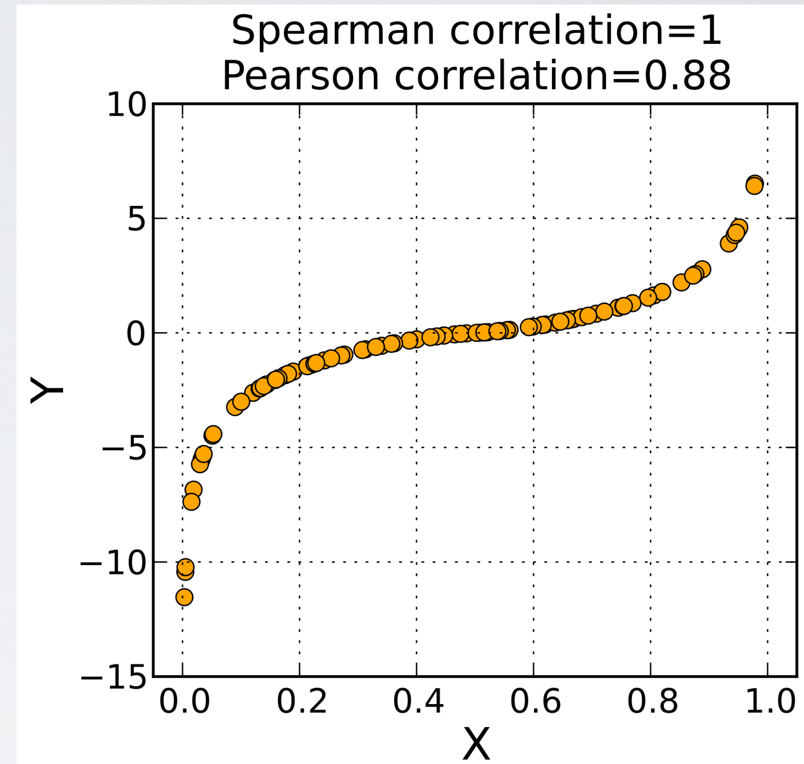


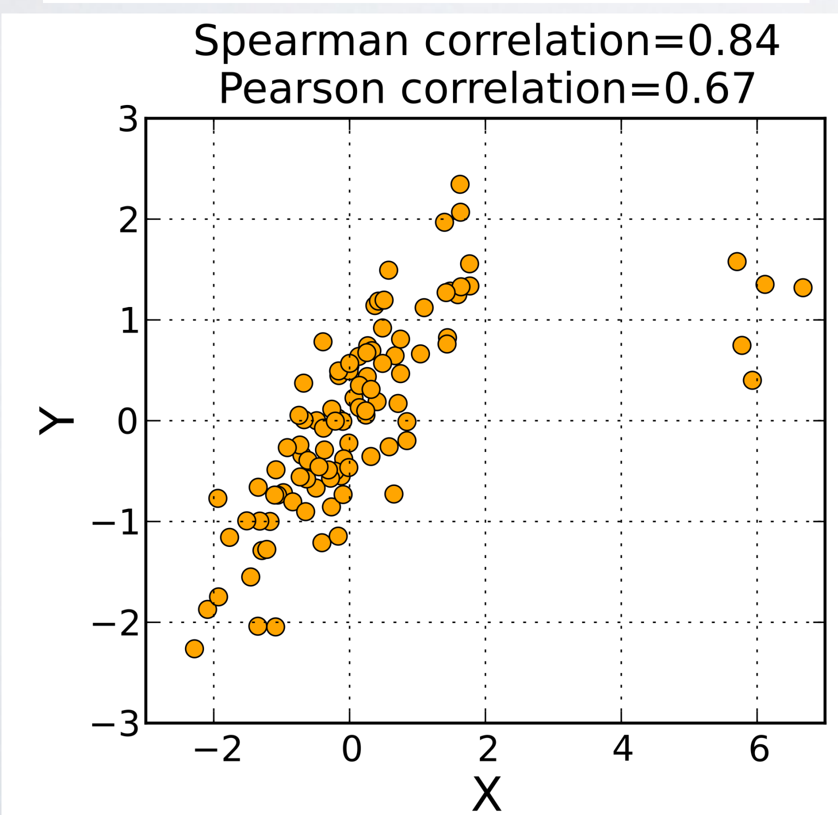
# SPEARMAN'S CORRELATION

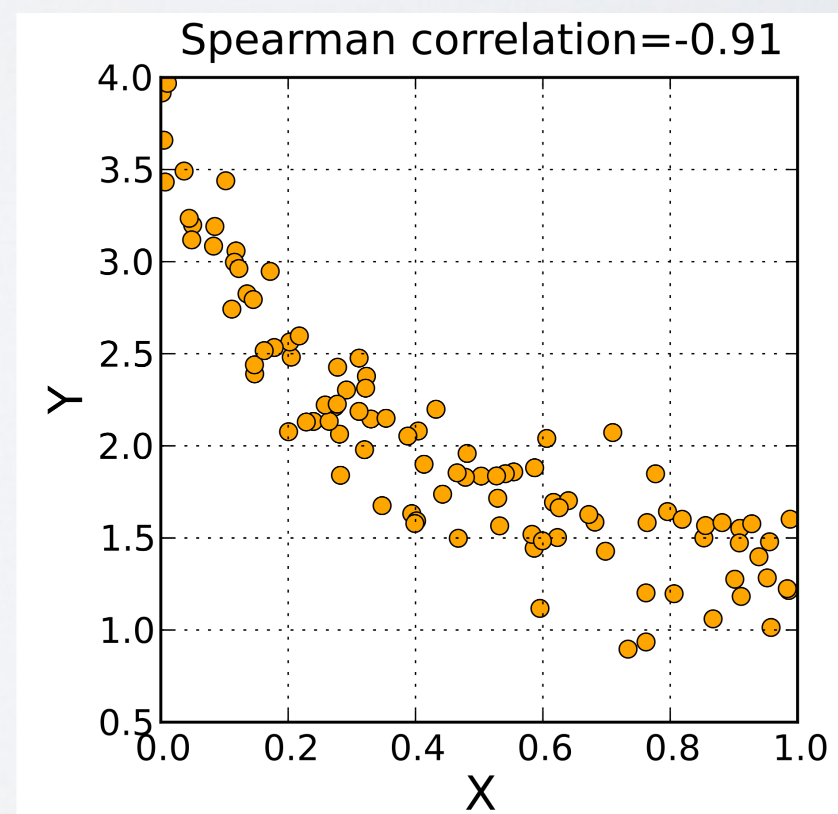
- Spearman's **rank** correlation coefficient
- Assesses how well the relationship between two variables can be described using a monotonic function
  - Not assuming a linear relation
- Pearson correlation coefficient between the rank variables
  - $r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \dfrac{\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}$

# SPEARMAN'S CORRELATION

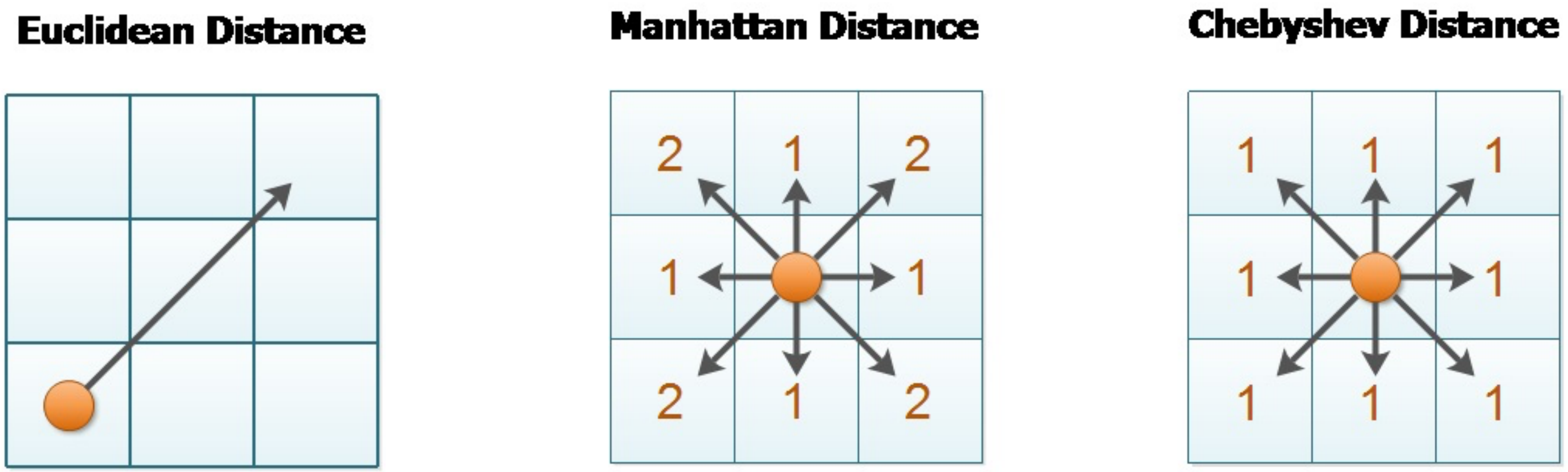


# NOTIONS OF DISTANCE




$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ $|x_1 - x_2| + |y_1 - y_2|$ $\max(|x_1 - x_2|, |y_1 - y_2|)$

# FEATURE SCALING

- We want to use euclidean distance to compute the "distance" between 2 people based on attributes age(y), height(m), weight(g).
  - a= (y:20,m:1.82,g:80 000), b=(y:20,m:1.82,g:81000), c=(y:90,m:1.50,g:80 020)
    - d(a,b)=1000.0005
    - d(a,c)=72.8
  - That is not what we expected from our expert knowledge!
    - We should normalize/standardize data

# FEATURE SCALING

- Rescaling (Normalization): $x' = \dfrac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}$ :[0,1]
- Mean normalization: $x' = \dfrac{x - \text{average}(x)}{\text{max}(x) - \text{min}(x)}$: 0=mean
- Standardization (z-score normalization): $x' = \dfrac{x - \bar{x}}{\sigma}$
  - 0: mean, -1/+1: 1 standard deviation from the mean

# SOME “GOLDEN RULES”

# SOME "GOLDEN RULES"

- In real life:
  - Your data does not follow a normal distribution. Nor a power law, nor any other theoretical distribution
  - Your features are always correlated
  - You always have non-linear relationships

# SOME "GOLDEN RULES"

- GIGO: Garbage in, Garbage out

# SOME "GOLDEN RULES"

- Real data is always garbage

# SOME "GOLDEN RULES"

- Get to know your data
  - Exploratory Analysis

# EXPERIMENTS

- Go to the webpage of the class and do today's experiments
- The "Advanced" section is not mandatory, you can do it if you have time