CPSC 340:
Machine Learning and Data Mining

Data Exploration

Fall 2018

This lecture roughly follow:
http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap2_data.pdf

Admin

- Assignment 1 is due next Friday: start early.
- Waiting list people: you should be registered soon-ish.
  - Start on the assignment now.
- Bookmark the course webpage:
  - www.ugrad.cs.ubc.ca/~cs340
- Sign up for a CS undergrad account:
  - https://www.cs.ubc.ca/getacct
  - Needed to access lectures, notes, assignment, and to submit assignment.
- Sign up for the course Piazza group:
  - http://piazza.com/ubc.ca/winterterm12018/cpsc340
- Tutorials start next week.
- Office hours start next week (see the calendar).
- Auditing: message instructors on Piazza if you want to audit.

Data Mining: Bird's Eye View

1) Collect data.
2) Data mining!
3) Profit?

Unfortunately, it's often more complicated…

Data Mining: Some Typical Steps

1) Learn about the application.
2) Identify data mining task.
3) Collect data.
4) Clean and preprocess the data.
5) Transform data or select useful subsets.
6) Choose data mining algorithm.
7) Data mining!
8) Evaluate, visualize, and interpret results.
9) Use results for profit or other goals.

(often, you'll go through cycles of the above)

Data Mining: Some Typical Steps

1) Learn about the application.
2) Identify data mining task.
3) Collect data.
4) Clean and preprocess the data.
5) Transform data or select useful subsets.
6) Choose data mining algorithm.
7) Data mining!
8) Evaluate, visualize, and interpret results.
9) Use results for profit or other goals.

(often, you'll go through cycles of the above)

What is Data?

- We’ll define data as a collection of examples, and their features.

Age	Job?	City	Rating	Income
23	Yes	Van	A	22,000.00
23	Yes	Bur	BBB	21,000.00
22	No	Van	CC	0.00
25	Yes	Sur	AAA	57,000.00
19	No	Bur	BB	13,500.00
22	Yes	Van	A	20,000.00
21	Yes	Ric	A	18,000.00

"feature"

"example"

- Each row is an “example”, each column is a “feature”.
  - Examples are also sometimes called “samples”.

Types of Data

- Categorical features come from an unordered set:
  - Binary: job?
  - Nominal: city.

- Numerical features come from ordered sets:
  - Discrete counts: age.
  - Ordinal: rating.
  - Continuous/real-valued: height.

Converting to Numerical Features

- Often want a real-valued example representation:

Age	City	Income
23	Van	22,000.00
23	Bur	21,000.00
22	Van	0.00
25	Sur	57,000.00
19	Bur	13,500.00
22	Van	20,000.00

→

Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00

- This is called a “1 of k” encoding.
- We can now interpret examples as points in space:
  - E.g., first example is at (23,1,0,0,22000).

Approximating Text with Numerical Features

- Bag of words replaces document by word counts:

The **International Conference on Machine Learning** (ICML) is the leading international academic conference in machine learning

ICML	International	Conference	Machine	Learning	Leading	Academic
1	2	2	2	2	1	1

- Ignores order, but often captures general theme.
- You can compute ‘distance’ between documents.

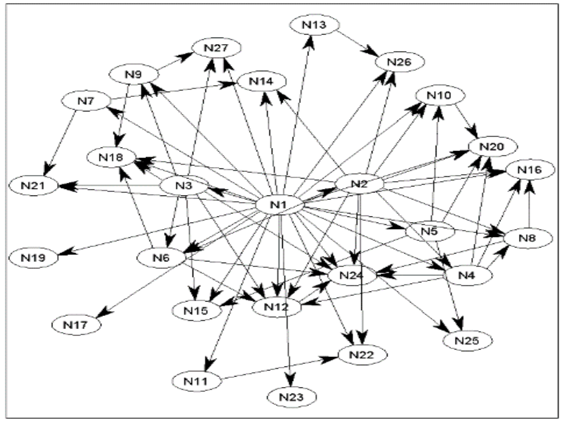
Approximating Images and Graphs

- We can think of other data types in this way:
  - Images:

→ graycale intensity

(1,1)	(2,1)	(3,1)	…	(m,1)	…	(m,n)
45	44	43	…	12	…	35

  - Graphs:



→ adjacency matrix

N1	N2	N3	N4	N5	N6	N7
0	1	1	1	1	1	1
0	0	0	1	0	1	0
0	0	0	0	0	1	0
0	0	0	0	0	0	0

Data Cleaning

- ML+DM typically assume ‘clean’ data.
- Ways that data might not be ‘clean’:
  - Noise (e.g., distortion on phone).
  - Outliers (e.g., data entry or instrument error).
  - Missing values (no value available or not applicable)
  - Duplicated data (repetitions, or different storage formats).
- Any of these can lead to problems in analyses.
  - Want to fix these issues, if possible.
  - Some ML methods are robust to these.
  - Often, ML is the best way to detect/fix these.

The Question I Hate the Most…

- How much data do we need?

- A difficult if not impossible question to answer.

- My usual answer: “more is better”.
  - With the warning: “as long as the quality doesn’t suffer”.

- Another popular answer: “ten times the number of features”.

A Simple Setting: Coupon Collecting

- Assume we have a categorical variable with 50 possible values:
  - {Alabama, Alaska, Arizona, Arkansas,…}.
- Assume each category has probability of 1/50 of being chosen:
  - How many examples do we need to see before we expect to see them all?
- Expected value is ~225.
- Coupon collector problem: O(n log n) in general.
  - Gotta Catch'em all!
- Obvious sanity check, is need more samples than categories:
  - Situation is worse if they don't have equal probabilities.
  - Typically want to see categories more than once to learn anything.

Feature Aggregation

- Feature aggregation:
  - Combine features to form new features:

Van	Bur	Sur	Edm	Cal
1	0	0	0	0
0	1	0	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	0	1
0	0	1	0	0

→

BC	AB
1	0
1	0
1	0
0	1
0	1
1	0

- More province information than city information.

Feature Selection

- Feature Selection:
  - Remove features that are not relevant to the task.

SID:	Age	Job?	City	Rating	Income
3457	23	Yes	Van	A	22,000.00
1247	23	Yes	Bur	BBB	21,000.00
6421	22	No	Van	CC	0.00
1235	25	Yes	Sur	AAA	57,000.00
8976	19	No	Bur	BB	13,500.00
2345	22	Yes	Van	A	20,000.00

  - Student ID is probably not relevant.

Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.

Age
23
23
22
25
19
22

< 20	>= 20, < 25	>= 25
0	1	0
0	1	0
0	1	0
0	0	1
1	0	0
0	1	0

- Only need consider 3 values.

Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.
  - Square, exponentiation, or take logarithm.

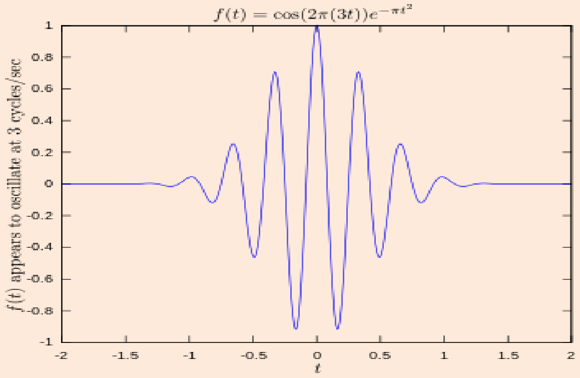
www.google.com/finance

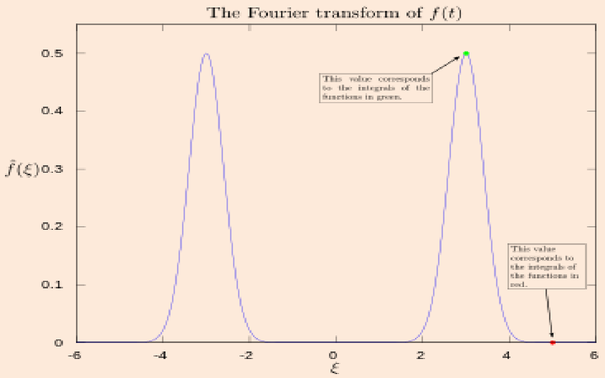
Feature Transformation

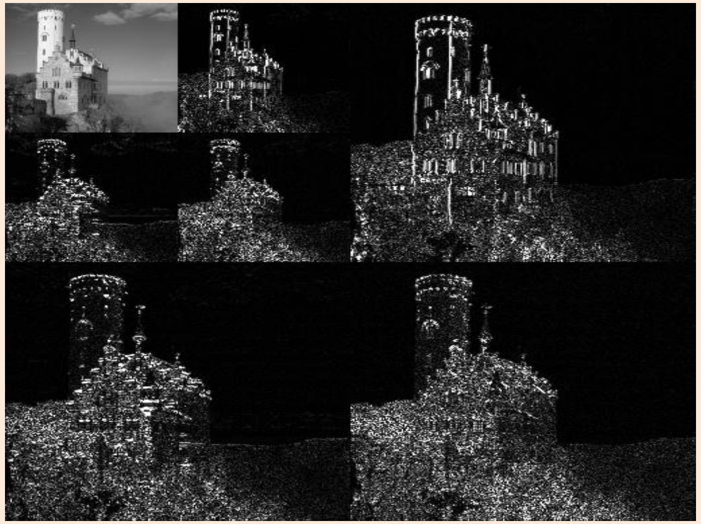
- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.
  - Square, exponentiation, or take logarithm.
  - Scaling: convert variables to comparable scales
    (E.g., convert kilograms to grams.)

Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.
  - Square, exponentiation, or take logarithm.
  - Scaling: convert variables to comparable scales.
  - Fourier coefficients, spectrograms, and wavelets (signal data).







https://en.wikipedia.org/wiki/Fourier_transform
https://en.wikipedia.org/wiki/Spectrogram
https://en.wikipedia.org/wiki/Discrete_wavelet_transform

Links to figure sources will be here.

(pause)

Exploratory Data Analysis

- You should always ‘look’ at the data first.

- But how do you ‘look’ at features and high-dimensional examples?
  - Summary statistics.
  - Visualization.
  - ML + DM (later in course).

Categorical Summary Statistics

- Summary statistics for a categorical variable:
  - Frequencies of different classes.
  - Mode: category that occurs most often.
  - Quantiles: categories that occur more than t times.

Population by year, by province and territory (Number)

	2014
Canada	**35,540.4**
Newfoundland and Labrador	527.0
Prince Edward Island	146.3
Nova Scotia	942.7
New Brunswick	753.9
Quebec	8,214.7
Ontario	13,678.7
Manitoba	1,282.0
Saskatchewan	1,125.4
Alberta	4,121.7
British Columbia	4,631.3
Yukon	36.5
Northwest Territories	43.6
Nunavut	36.6

Frequency: 13.3% of Canadian residents live in BC.
Mode: Ontario has largest number of residents (38.5%)
Quantile: 6 provinces have more than 1 million people.

Continuous Summary Statistics

- Measures of location:
  - Mean: average value.
  - Median: value such that half points are larger/smaller.
  - Quantiles: value such that ‘t’ points are larger.
- Measures of spread:
  - Range: minimum and maximum values.
  - Variance: measures how far values are from mean.
    - Square root of variance is “standard deviation”.
  - Intequantile ranges: difference between quantiles.

Continuous Summary Statistics

- Data: [0 1 2 3 3 5 7 8 9 10 14 15 17 200] "outlier"
- Measures of location:
  - Mean(Data) = 21
  - Mode(Data) = 3
  - Median(Data) = 7.5
  - Quantile(Data,0.5) = 7.5
  - Quantile(Data,0.25) = 3
  - Quantile(Data,0.75) = 14
- Measures of spread:
  - Range(Data) = [0 200].
  - Std(Data) = 51.79
  - IQR(Data,.25,.75) = 11
- Notice that mean and std are more sensitive to extreme values ("outliers").

Entropy as Measure of Randomness

- Another common summary statistic is entropy.
  - Entropy measures "randomness" of a set of variables.
    - Roughly, another measure of the "spread" of values.
    - Formally, "how many bits of information are encoded in the average example".
  - For a categorical variable that can take 'k' values, entropy is defined by:

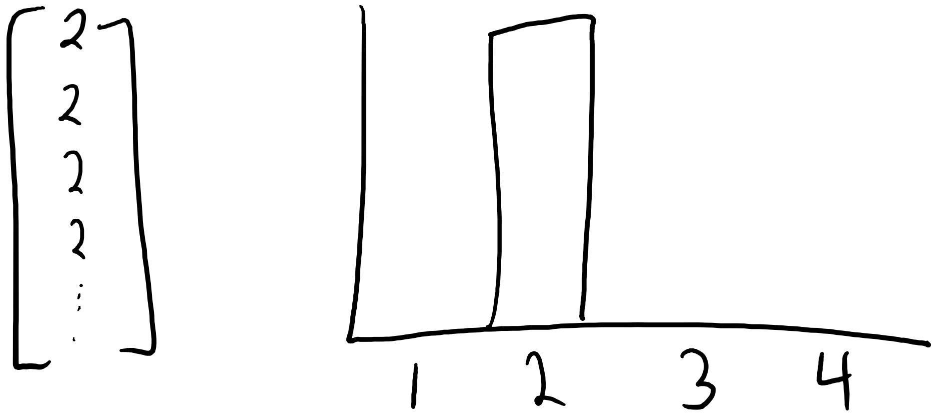
$$\text{entropy} = -\sum_{c=1}^{k} p_c \log p_c$$

where p_c is the proportion of times you have value 'c'.

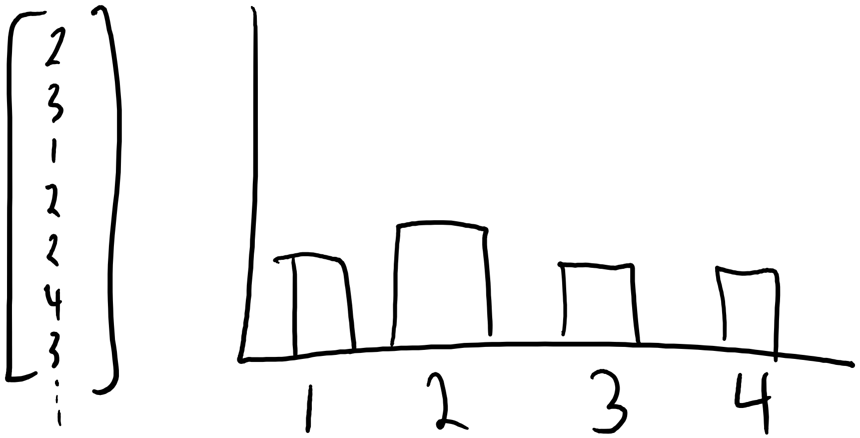
  - Low entropy means "very predictable".
  - High entropy means "very random".
  - Minimum value is 0, maximum value is log(k).
    - We use the convention that 0 log 0 = 0.

Entropy as Measure of Randomness

Low entropy means "very predictable"



High entropy means "very random"



- For categorical features: uniform distribution has highest entropy.
- For continuous densities with fixed mean and variance:
  - Normal distribution has highest entropy (not obvious).
- [Entropy and Dr. Seuss]() (words like "snunkoople" increase entropy).

Distances and Similarities

- There are also summary statistics between features ‘x’ and ‘y’.
  - Hamming distance:
    - Number of elements in the vectors that aren’t equal.
  - Euclidean distance:
    - How far apart are the vectors?
  - Correlation:
    - Does one increase/decrease linearly as the other increases?
    - Between -1 and 1.

x	y
0	0
0	0
1	0
0	1
0	1
1	1
0	0
0	1
0	1

Distances and Similarities

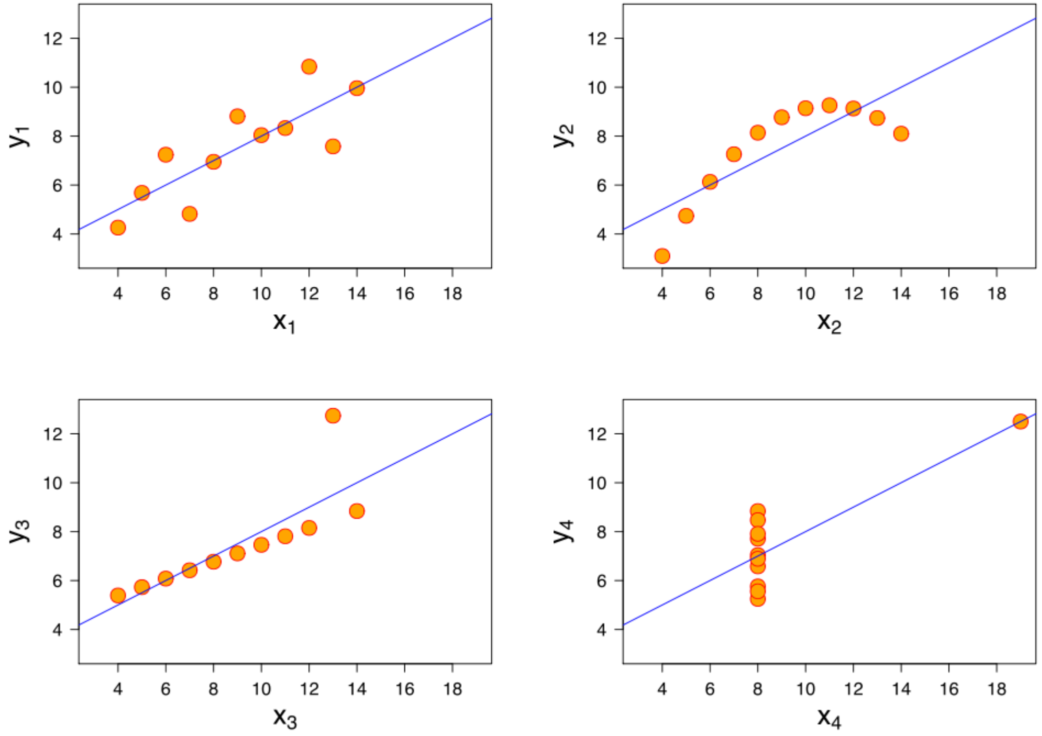
- There are also summary statistics between features 'x' and 'y'.
  - Rank correlation:
    - Does one increase/decrease non-linearly as the other increases?
- Distances/similarities between other examples:
  - Jaccard coefficient (distance between sets):
    - (size of intersection of sets) / (size of union of sets)
  - Edit distance (distance between strings):
    - How many characters do we need to change to go from x to y?
    - Computed using dynamic programming (CPSC 320).

x	y
0	0
0	0
1	0
0	1
0	1
1	1
0	0
0	1
0	1

Limitations of Summary Statistics

- On their own summary statistic can be misleading.
- [Why not to trust statistics]

- Amcomb's quartet:
  - Almost same means.
  - Almost same variances.
  - Almost same correlations.
  - Look completely different.
- [Datasaurus dozen].



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

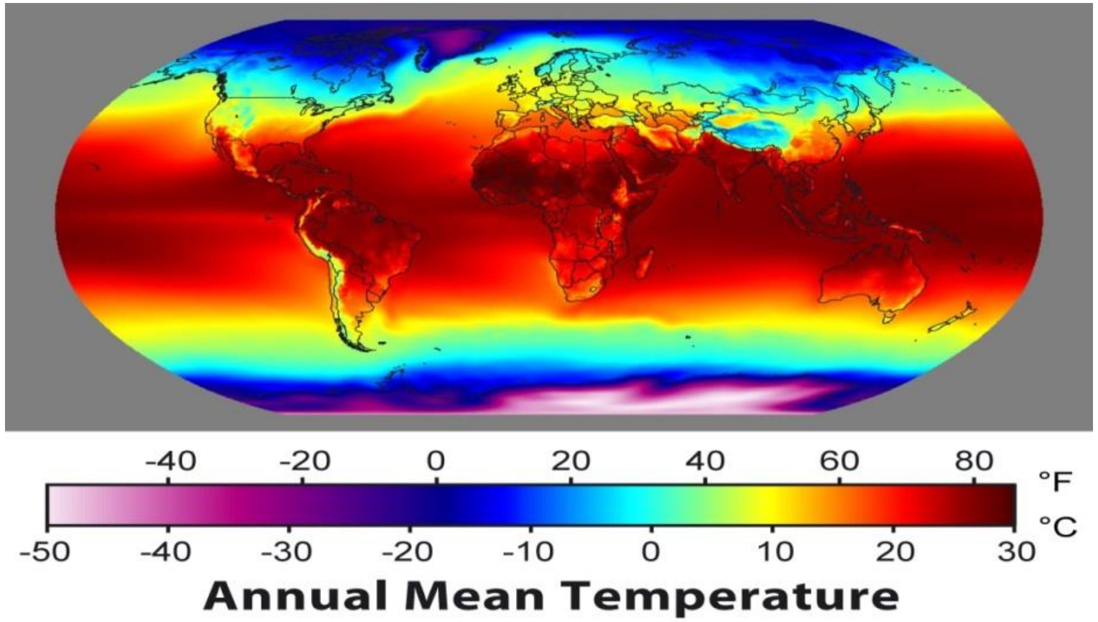
(pause)

Visualization

- You can learn a lot from 2D plots of the data:
  - Patterns, trends, outliers, unusual patterns.

Lat	Long	Temp
0	0	30.1
0	1	29.8
0	2	29.9
0	3	30.1
0	4	29.9
…	…	…

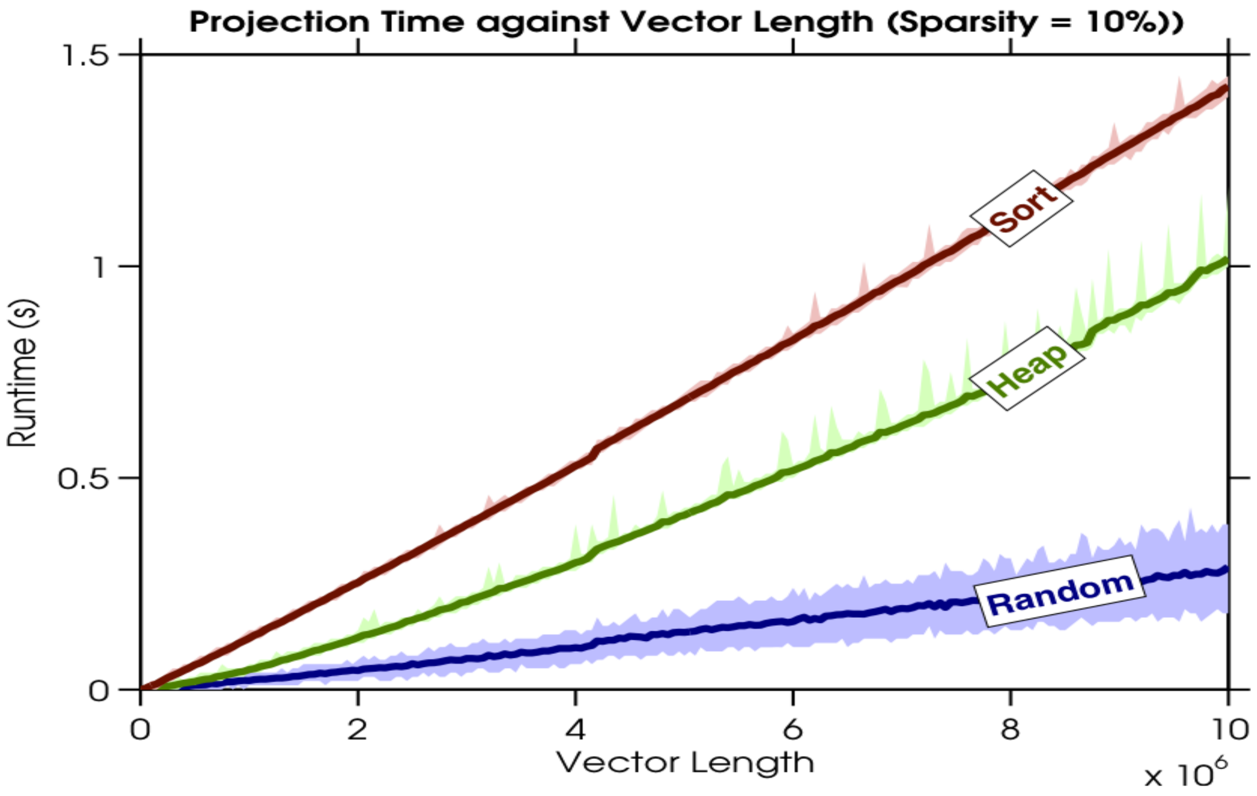
vs.

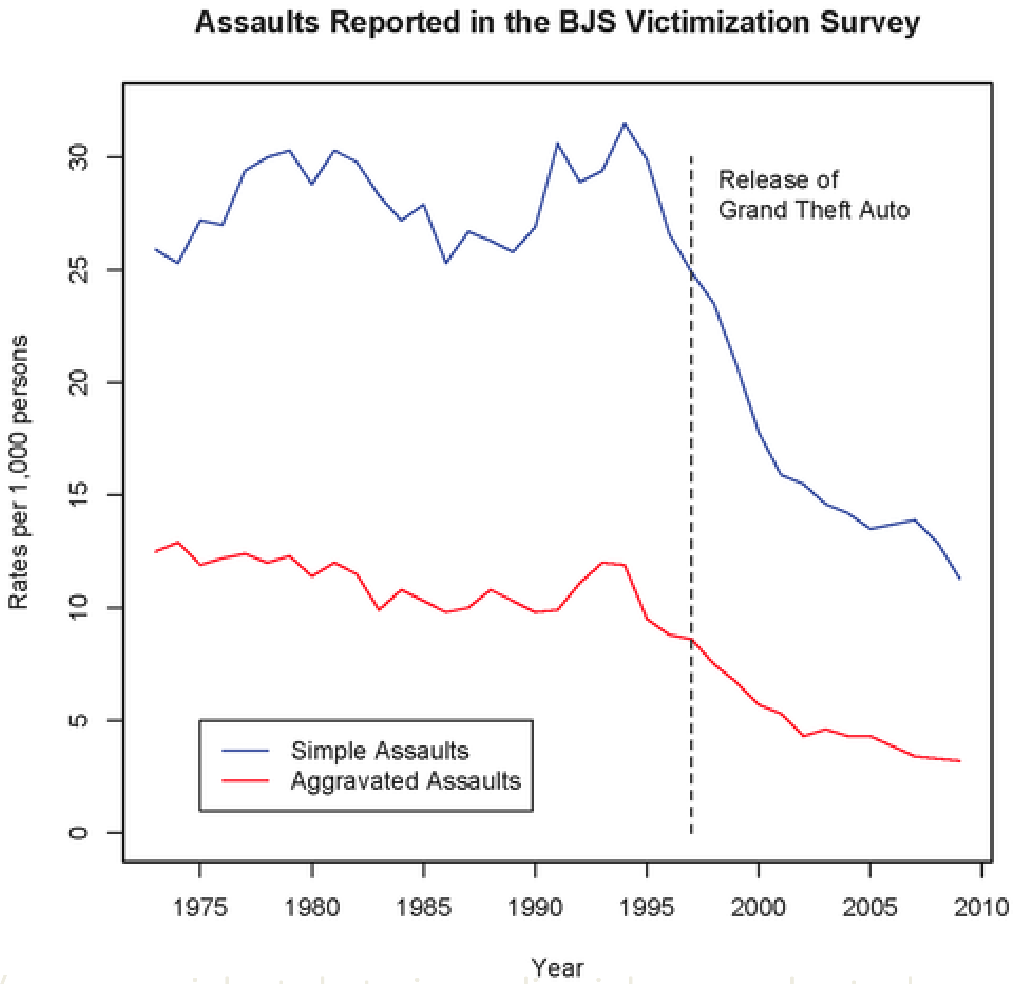


https://en.wikipedia.org/wiki/Temperature

Basic Plot

- Visualize one variable as a function of another.



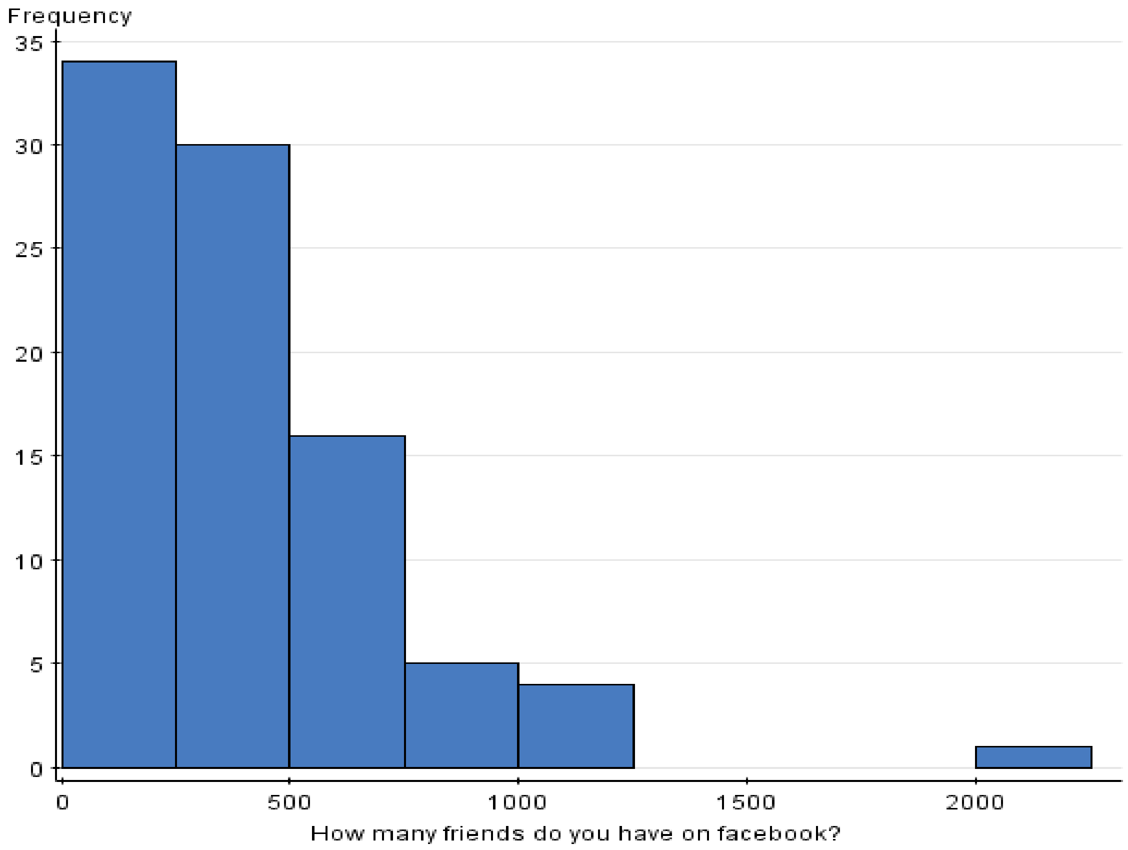


- [Fun with plots](http://notunlikeresearch.typepad.com/something-not-unlike-rese/2011/01/more-on-violent-rhetoric-media-violence-and-actual-).

http://notunlikeresearch.typepad.com/something-not-unlike-rese/2011/01/more-on-violent-rhetoric-media-violence-and-actual-

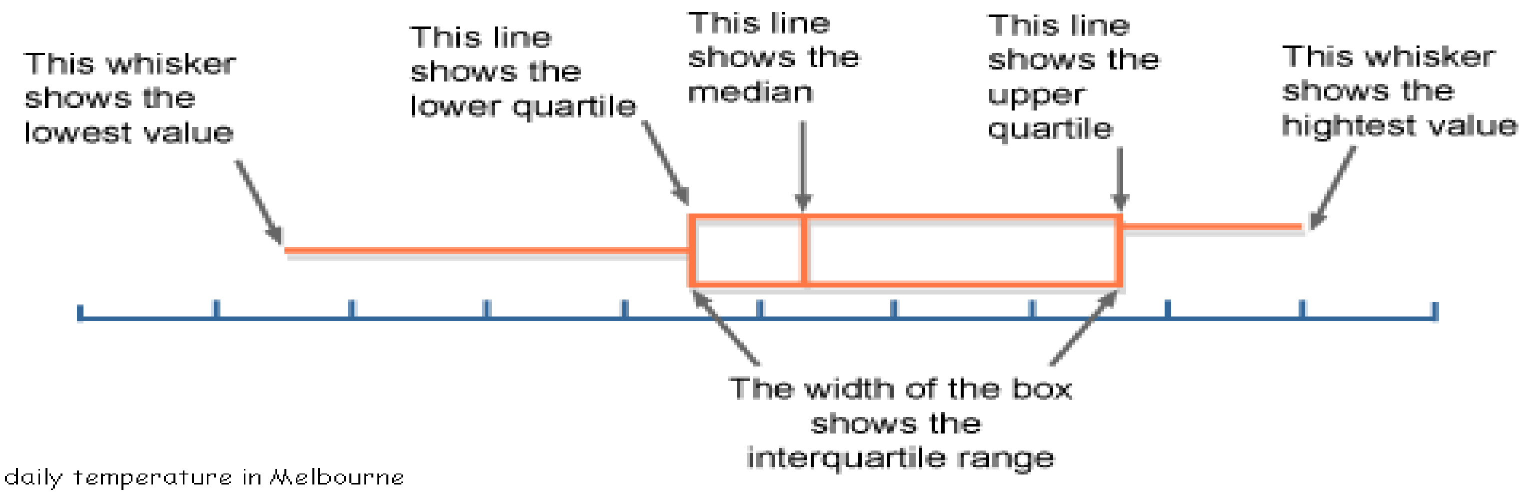
Histogram

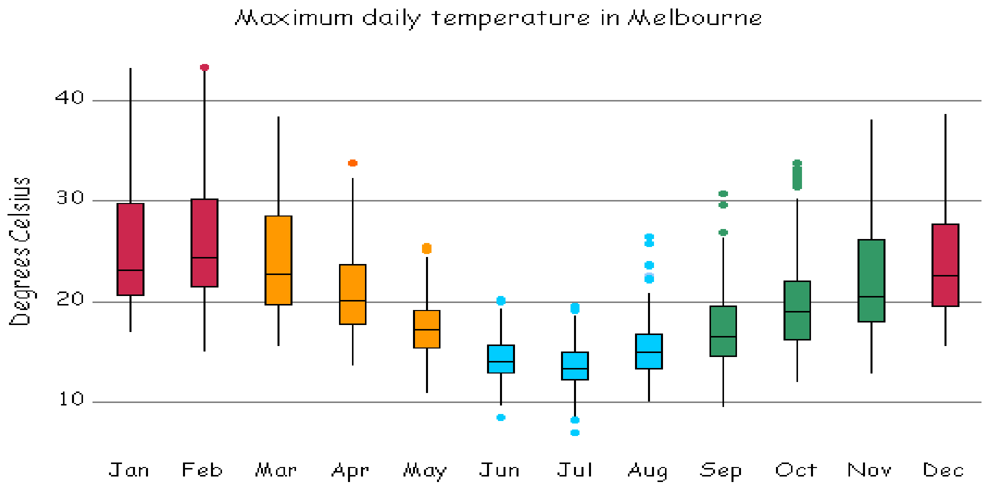
- Histograms display distribution of a variable.

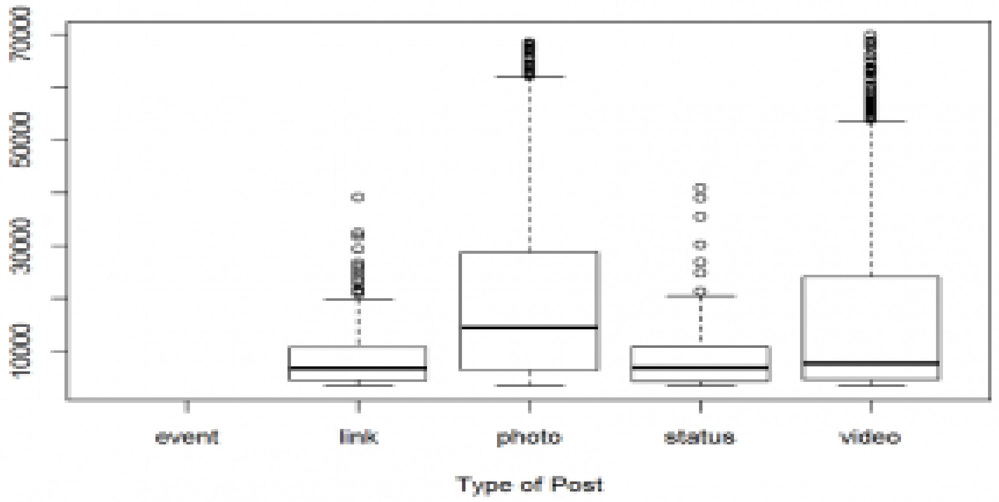


http://www.statcrunch.com/5.0/viewresult.php?resid=1024581

Box Plot







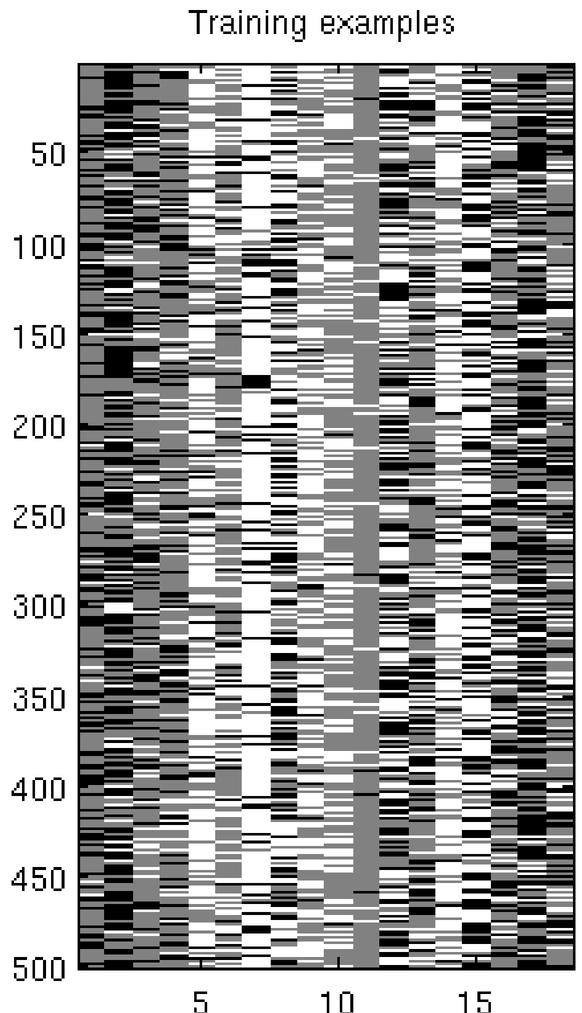
http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/representingdata3hi
http://www.scc.ms.unimelb.edu.au/whatisstatistics/weather.html
http://r.ramganalytics.com/r/facebook-likes-and-analytics/

Box Plot

- Photo from CTV Olympic coverage in 2010:

Matrix Plot

- If our features are real-valued, we can view data as a picture:
  - "Matrix plot".
  - May be able to see trends in features.

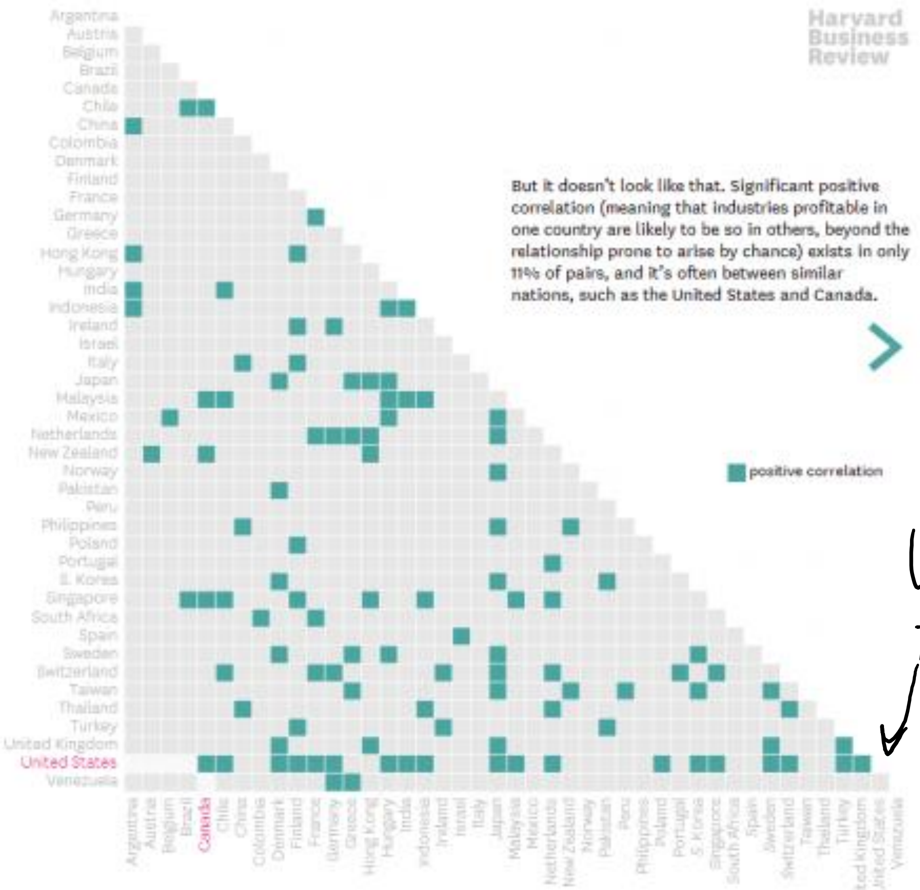


Matrix Plot

- A matrix plot of all similarities (or distances) in a data set:

"Correlation plot"

	BTC	ETH	XRP	XEM	ETC	LTC	DASH	XMR
BTC	1.00	0.61	0.36	0.51	0.60	0.56	0.55	0.66
ETH	0.61	1.00	0.28	0.49	0.68	0.43	0.70	0.64
XRP	0.36	0.28	1.00	0.48	0.08	0.35	0.40	0.44
XEM	0.51	0.49	0.48	1.00	0.40	0.43	0.47	0.52
ETC	0.60	0.68	0.08	0.40	1.00	0.47	0.56	0.53
LTC	0.56	0.43	0.35	0.43	0.47	1.00	0.59	0.67
DASH	0.55	0.70	0.40	0.47	0.56	0.59	1.00	0.74
XMR	0.66	0.64	0.44	0.52	0.53	0.67	0.74	1.00

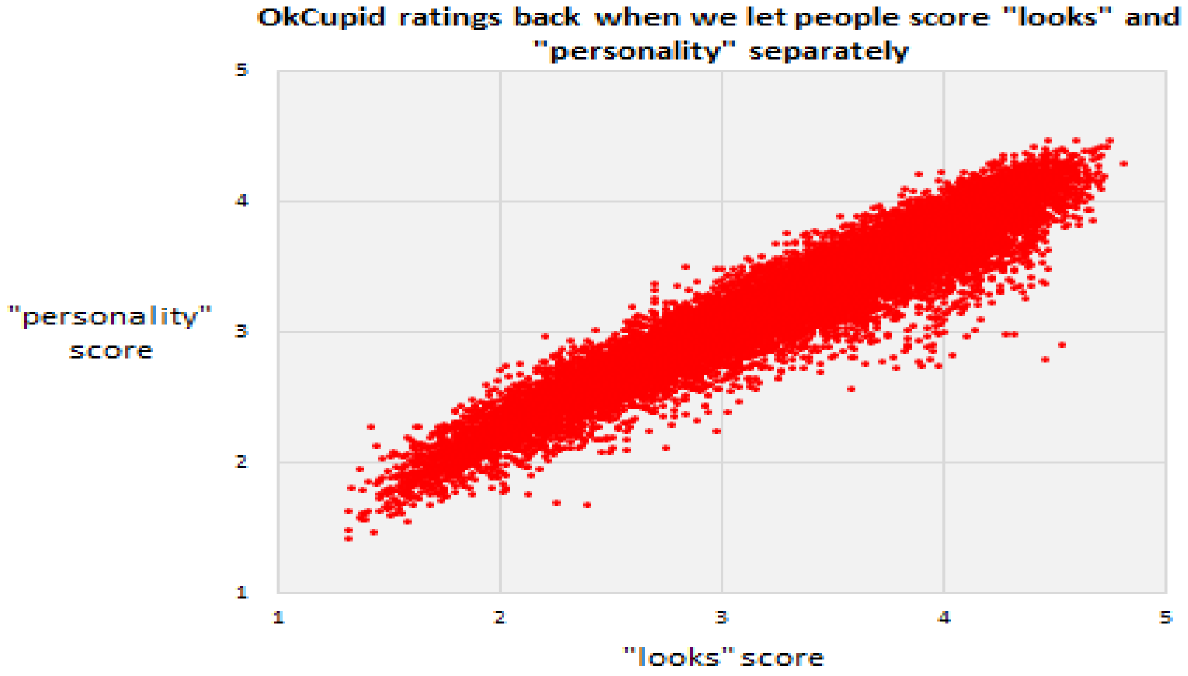


Using colour to highlight high correlations

https://www.crypto-news.net/portfolio-diversification-with-low-correlatio
https://hbr.org/web/infographic/2014/09/does-what-you-know-about-making-widgets-in-america-apply-in-china

Scatterplot

- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.
  - Basically a "plot without lines" between the points.

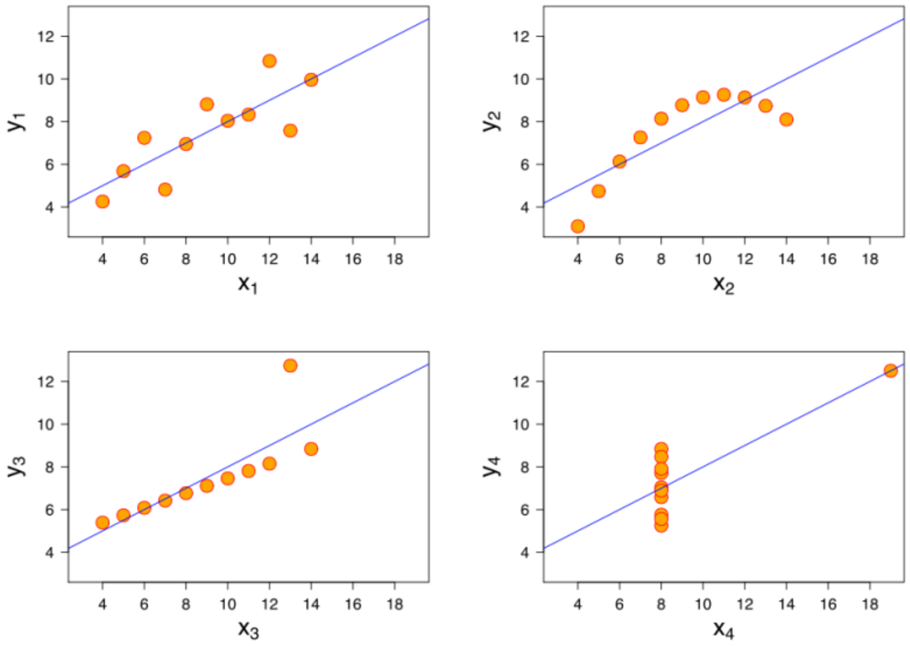


http://cdn.okccdn.com/blog/humanexperiments/looks-v-personality.png

- Shows correlation between "personality" score and "looks" score.

Scatterplot

- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.
  - Basically a “plot without lines” between the points.



- Shows correlation between “personality” score and “looks” score.
- But scatterplots let you see more complicated patterns.

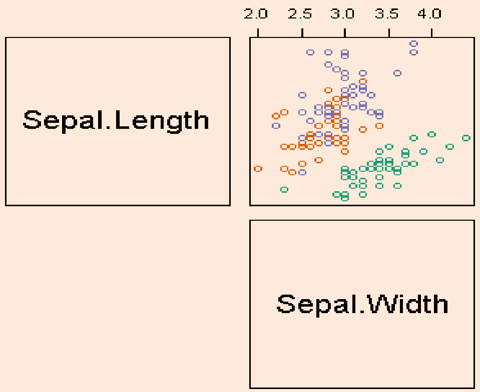
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Scatterplot Arrays

- For multiple variables, can use scatterplot array.

Fisher's *Iris* Data [hide]

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
4.5	2.3	1.3	0.3	*I. setosa*
5.0	2.3	3.3	1.0	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
5.5	2.4	3.8	1.1	*I. versicolor*
5.1	2.5	3.0	1.1	*I. versicolor*



- Colors can indicate a third categorical variable.

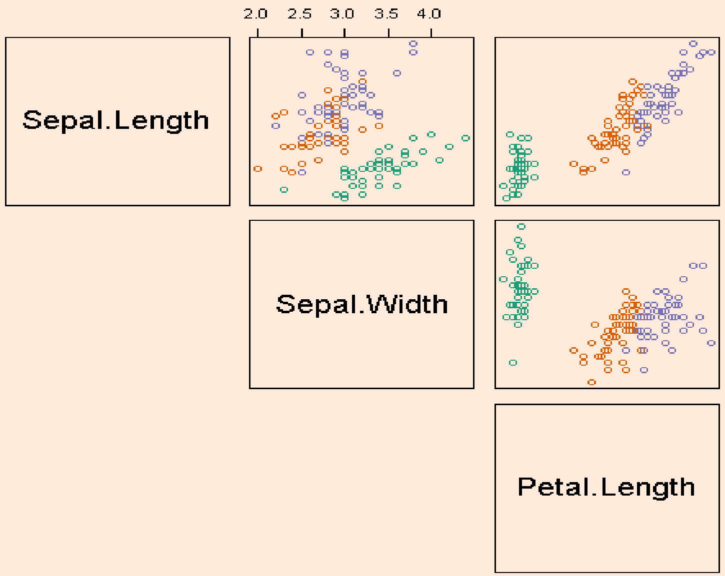
https://en.wikipedia.org/wiki/Iris_flower_data_set
http://www.ats.ucla.edu/stat/r/pages/layout.htm

Scatterplot Arrays

- For multiple variables, can use scatterplot array.

Fisher's *Iris* Data [hide]

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
4.5	2.3	1.3	0.3	*I. setosa*
5.0	2.3	3.3	1.0	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
5.5	2.4	3.8	1.1	*I. versicolor*
5.1	2.5	3.0	1.1	*I. versicolor*



- Colors can indicate a third categorical variable.

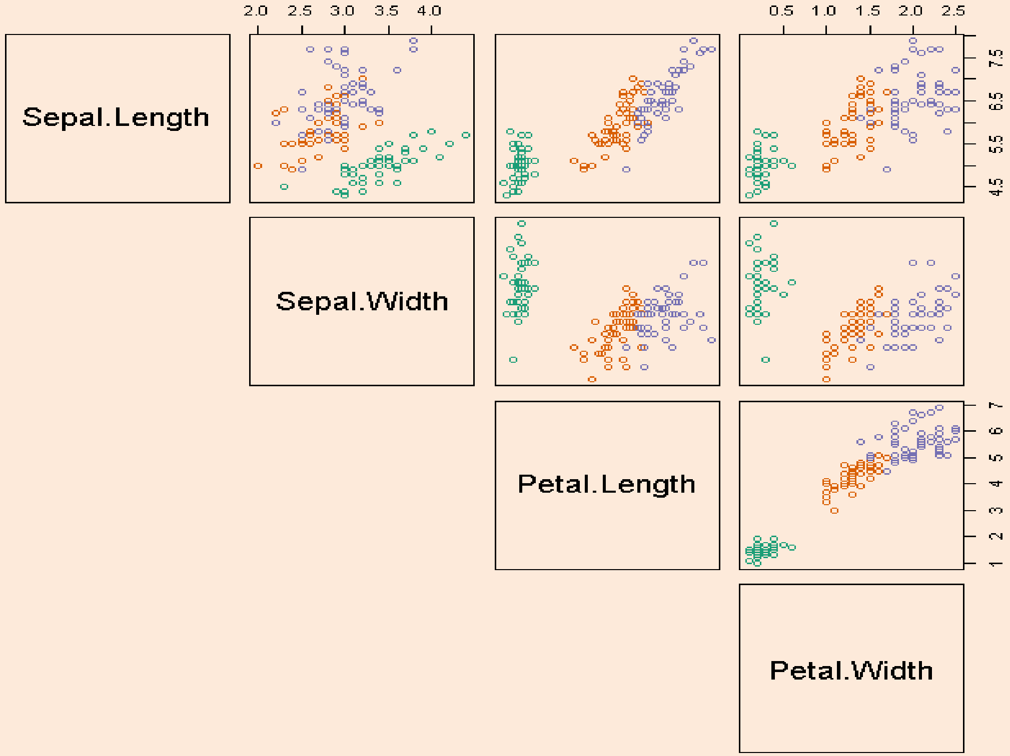
https://en.wikipedia.org/wiki/Iris_flower_data_set
http://www.ats.ucla.edu/stat/r/pages/layout.htm

Scatterplot Arrays

- For multiple variables, can use scatterplot array.

Fisher's *Iris* Data [hide]

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
4.5	2.3	1.3	0.3	*I. setosa*
5.0	2.3	3.3	1.0	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
5.5	2.4	3.8	1.1	*I. versicolor*
5.1	2.5	3.0	1.1	*I. versicolor*



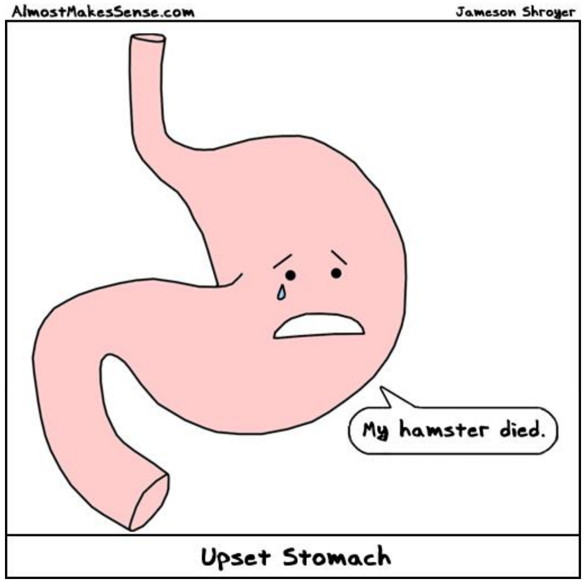
- Colors can indicate a third categorical variable.

https://en.wikipedia.org/wiki/Iris_flower_data_set
http://www.ats.ucla.edu/stat/r/pages/layout.htm

(pause)

Motivating Example: Food Allergies

- You frequently start getting an upset stomach



- You suspect an adult-onset food allergy.

http://www.cliparthut.com/upset-stomach-clipart-cn48e5.html

Motivating Example: Food Allergies

- To solve the mystery, you start a food journal:

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	…	Sick?
0	0.7	0	0.3	0	0		1
0.3	0.7	0	0.6	0	0.01		1
0	0	0	0.8	0	0		0
0.3	0.7	1.2	0	0.10	0.01		1
0.3	0	1.2	0.3	0.10	0.01		1

- But it’s hard to find the pattern:
  - You can’t isolate and only eat one food at a time.
  - You may be allergic to more than one food.
  - The quantity matters: a small amount may be ok.
  - You may be allergic to specific interactions.

Supervised Learning

- We can formulate this as supervised learning:

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	…	Sick?
0	0.7	0	0.3	0	0		1
0.3	0.7	0	0.6	0	0.01		1
0	0	0	0.8	0	0		0
0.3	0.7	1.2	0	0.10	0.01		1
0.3	0	1.2	0.3	0.10	0.01		1

- Input for an example (day of the week) is a set of features (quantities of food).
- Output is a desired class label (whether or not we got sick).
- Goal of supervised learning:
  - Use data to find a model that outputs the right label based on the features.
  - Model predicts whether foods will make you sick (even with new combinations).

Supervised Learning

- General supervised learning problem:
  - Take features of examples and corresponding labels as inputs.
  - Find a model that can accurately ***predict the labels of new examples***.

- This is the most successful machine learning technique:
  - Spam filtering, optical character recognition, Microsoft Kinect, speech recognition, classifying tumours, etc.

- We'll first focus on categorical labels, which is called "classification".
  - The model is a called a "classifier".

Naïve Supervised Learning: “Predict Mode”

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	…	Sick?
0	0.7	0	0.3	0	0		1
0.3	0.7	0	0.6	0	0.01		1
0	0	0	0.8	0	0		0
0.3	0.7	1.2	0	0.10	0.01		1
0.3	0	1.2	0.3	0.10	0.01		1

- A very naïve supervised learning method:
  - Count how many times each label occurred in the data (4 vs. 1 above).
  - Always predict the most common label, the “mode” (“sick” above).
- This ignores the features, so is only accurate if we only have 1 label.
- We want to use the features, and there are MANY ways to do this.
  - Next time we’ll consider a classic way known as decision tree learning.

Summary

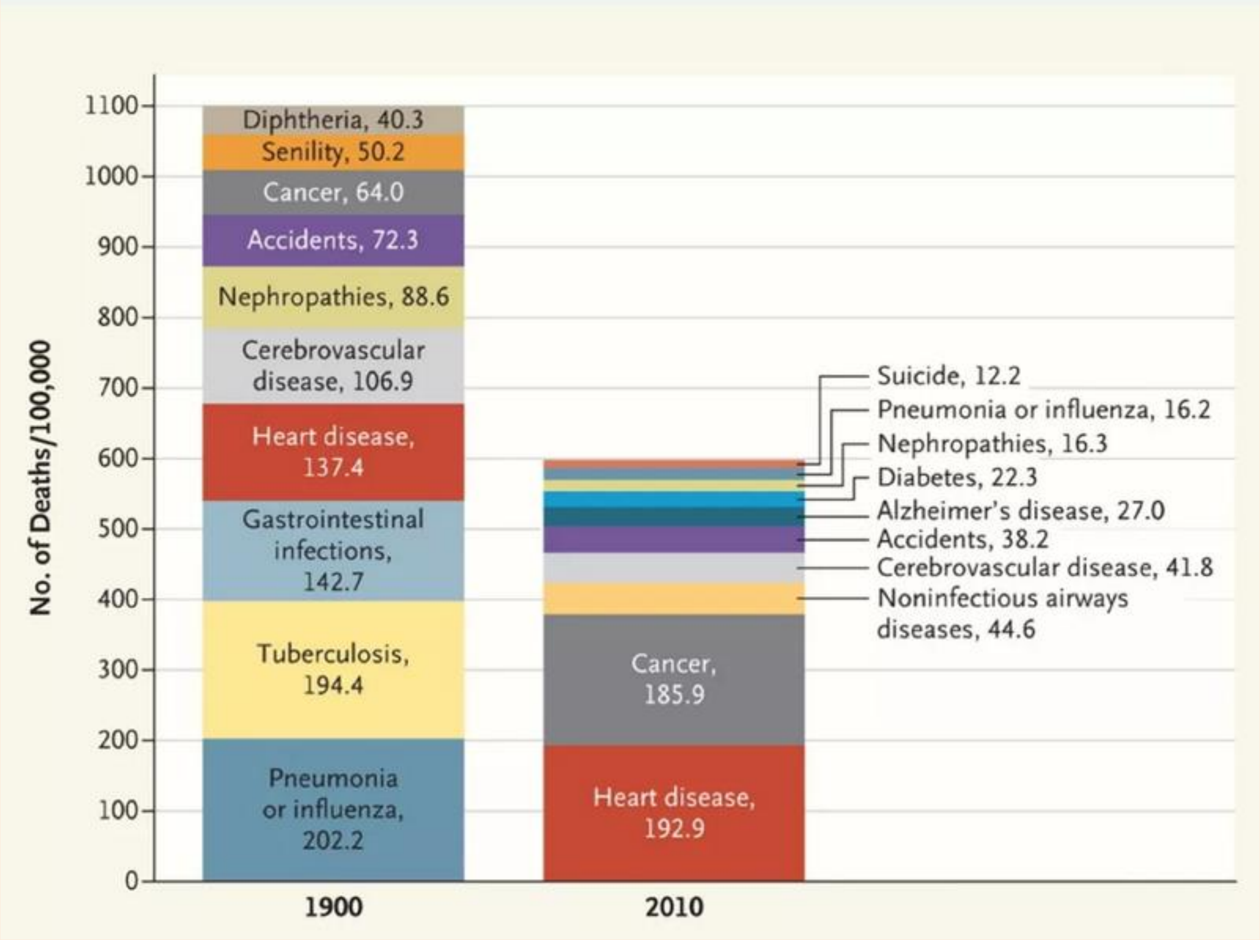
- Typical data mining steps:
  - Involves data collection, preprocessing, analysis, and evaluation.
- Example-feature representation and categorical/numerical features.
  - Transforming non-vector examples to vector representations.
- Feature transformations:
  - To address coupon collecting or simplify relationships between variables.
- Exploring data:
  - Summary statistics and data visualization.
- Supervised learning:
  - Using data to write a program based on input/output examples.

- Post-lecture bonus slides: other visualizations, parallel/distributed.

- Next week: let’s start some machine learning…

Histogram

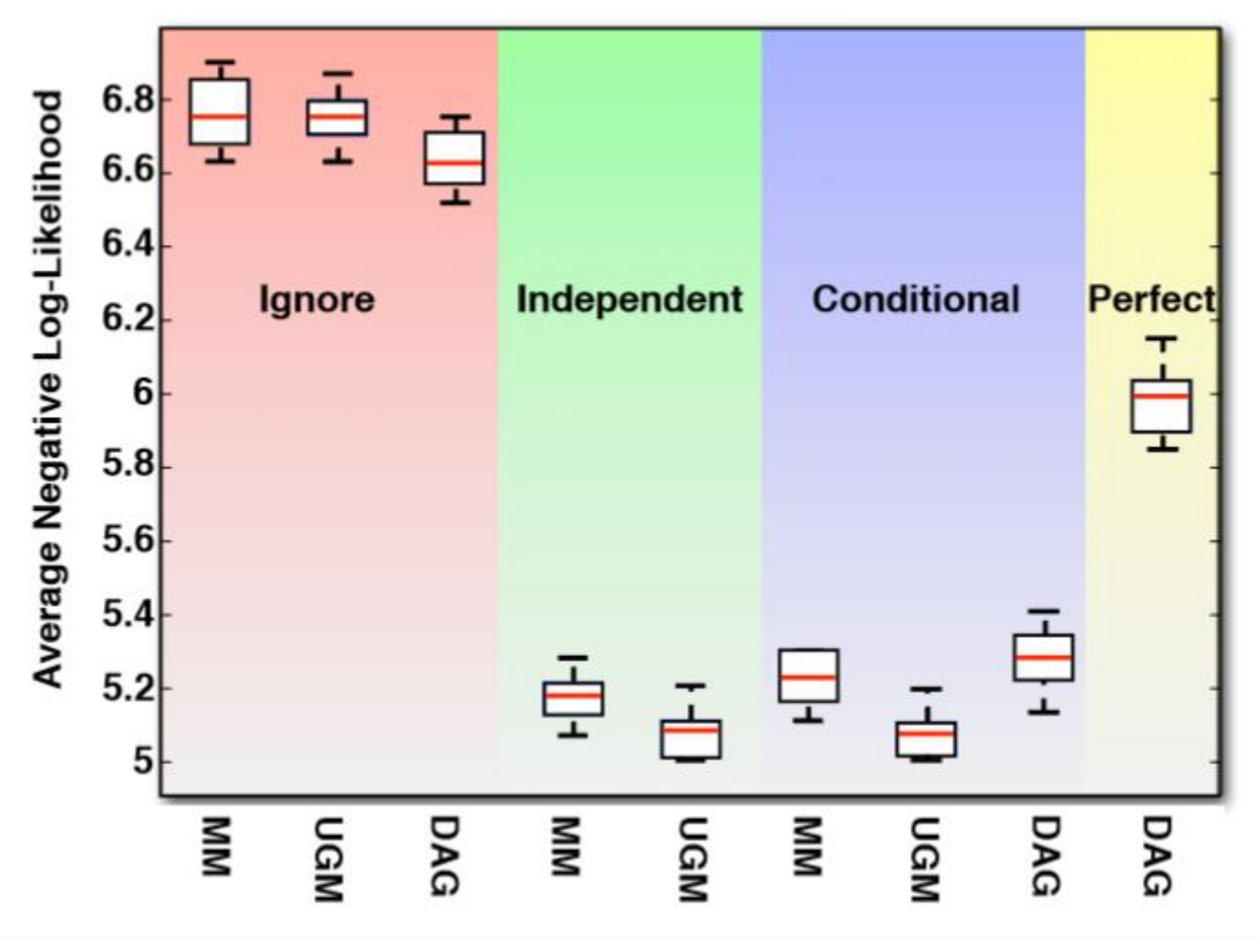
- Histogram with grouping:



http://www.vox.com/2016/5/10/11608064/americans-cause-of-death

Box Plots

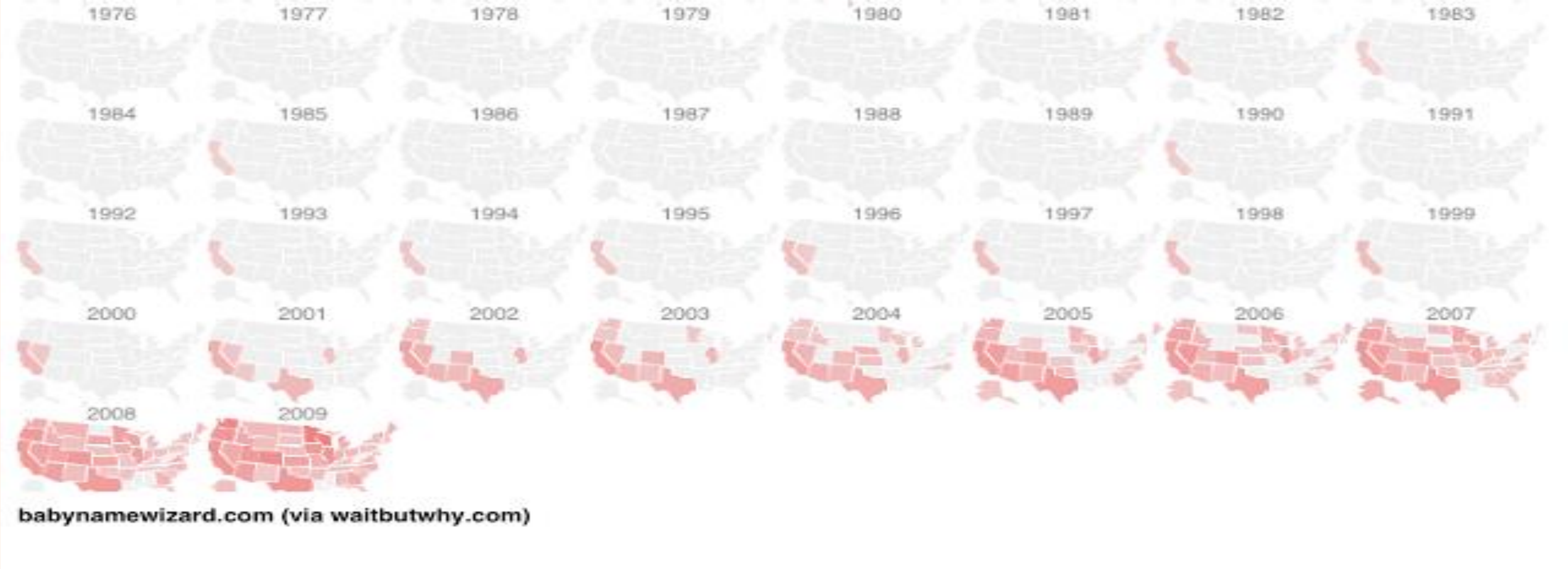
- Box plot with grouping:



Map Coloring

- Color/intensity can represent feature of region.

Popularity of naming baby “Evelyn” over time:



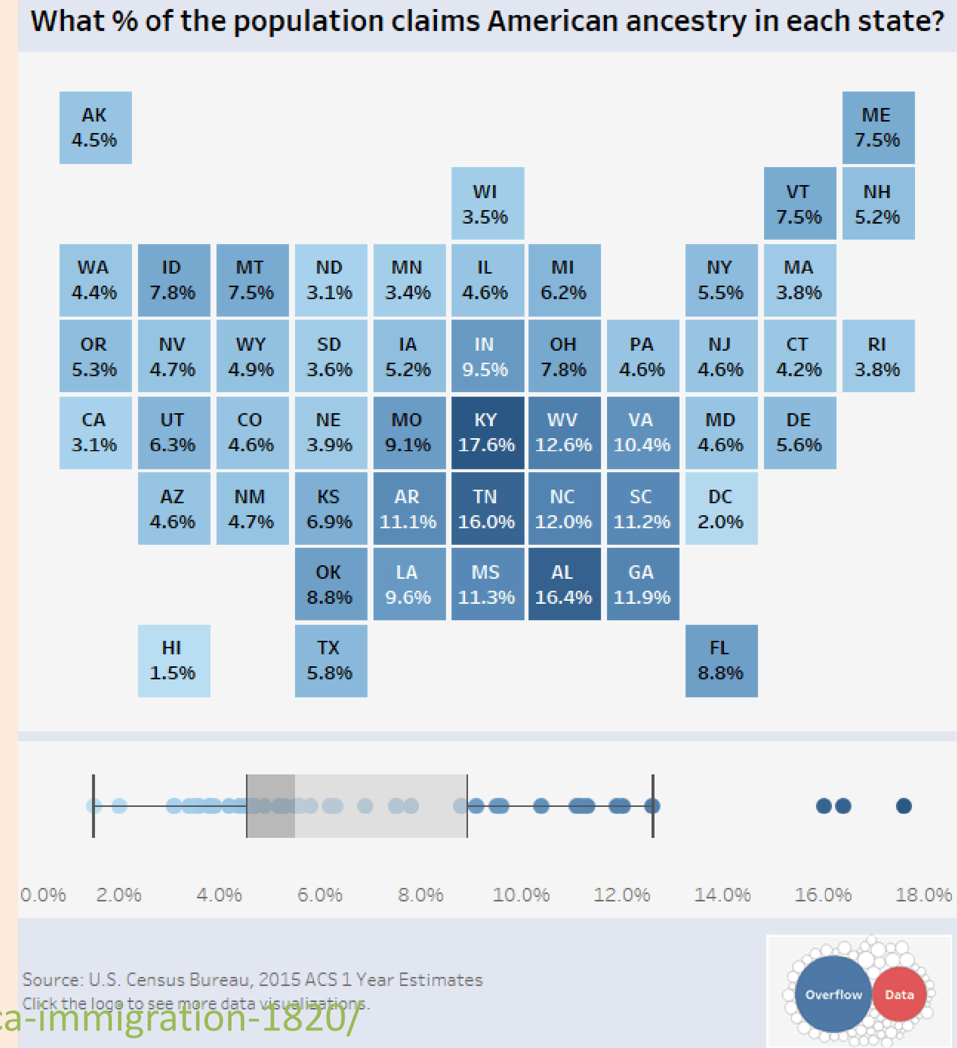
But not very good if some regions are very small.

[Canadian Income Mobility]

http://waitbutwhy.com/2013/12/how-to-name-baby.html

Map Coloring

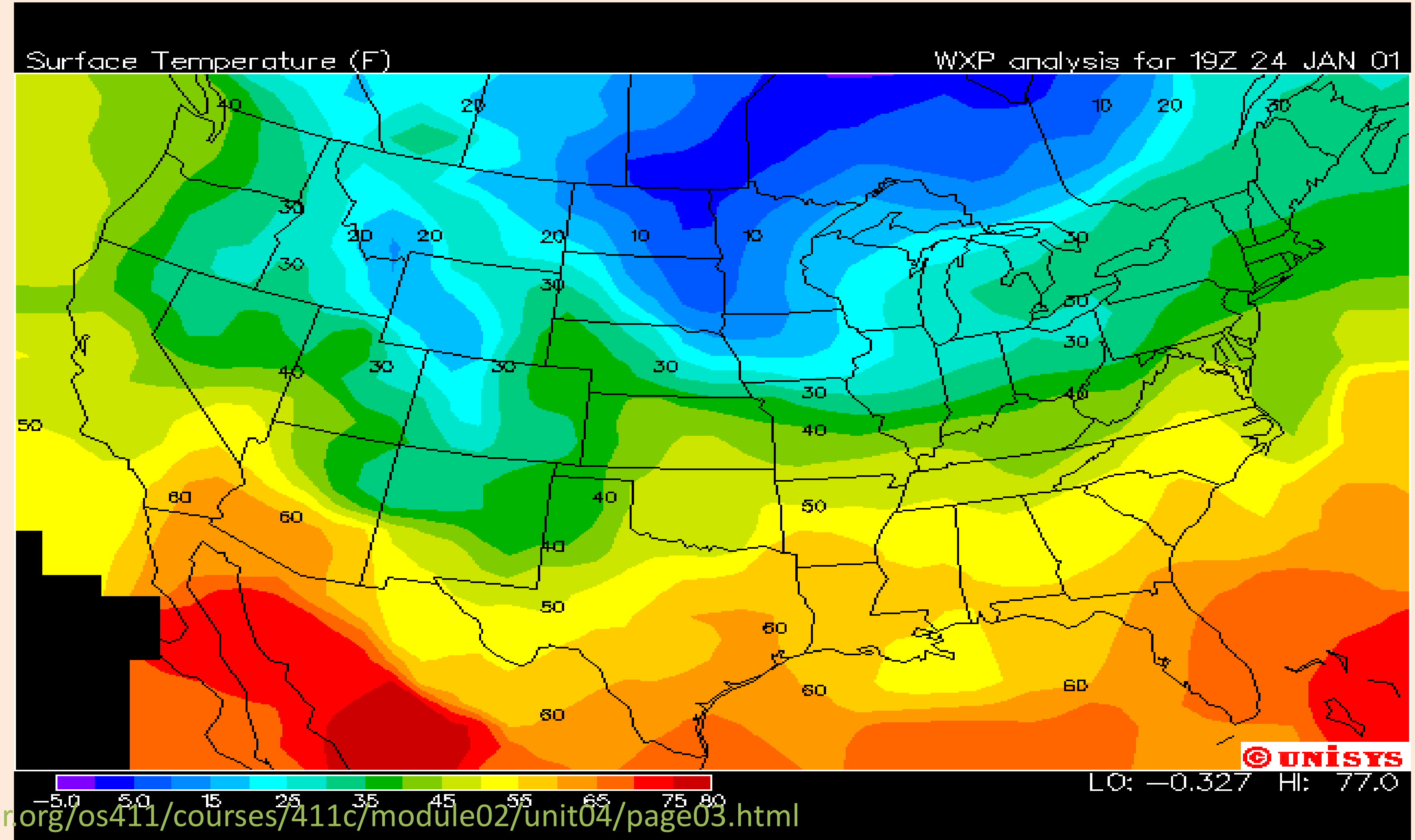
- Variation just uses fixed-size blocks and tries to arrange geographically:



http://www.visualcapitalist.com/america-immigration-1820/

Contour Plot

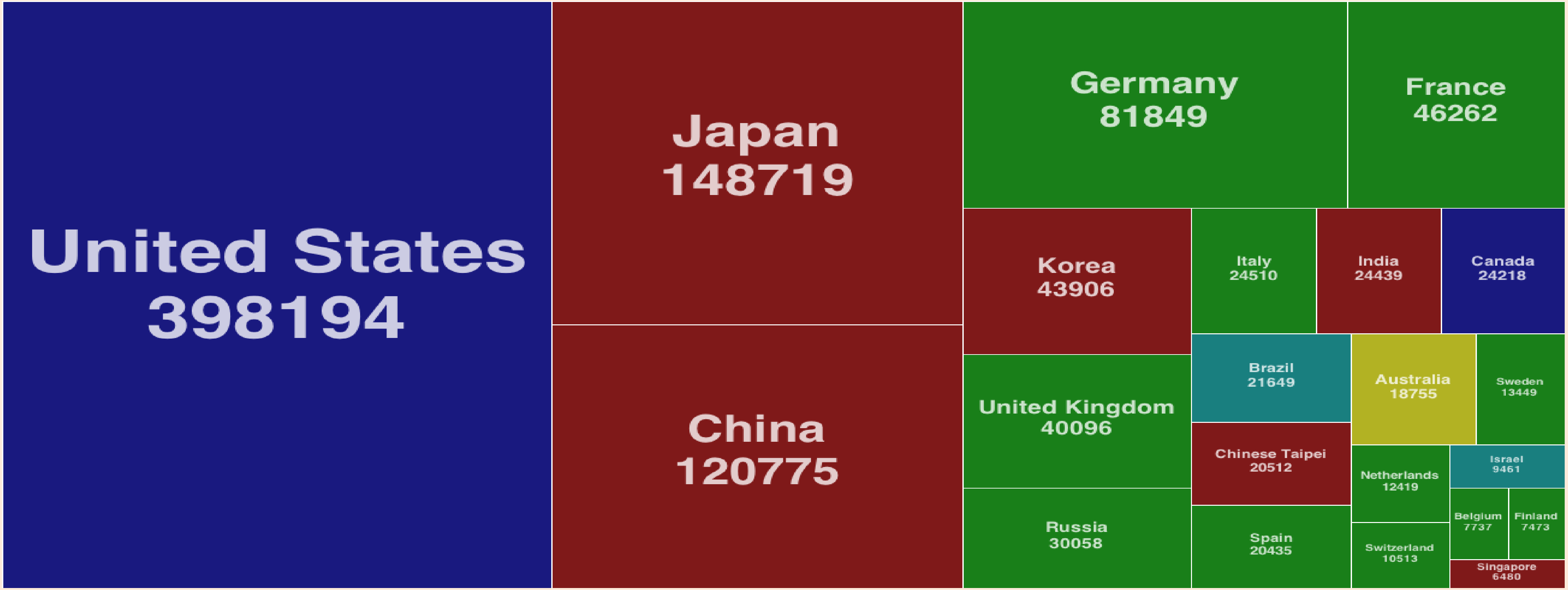
- Colour visualizes ‘z’ as we vary ‘x’ and ‘y’.



http://www.shodor.org/os411/courses/411c/module02/unit04/page03.html

Treemaps

- Area represents attribute value:



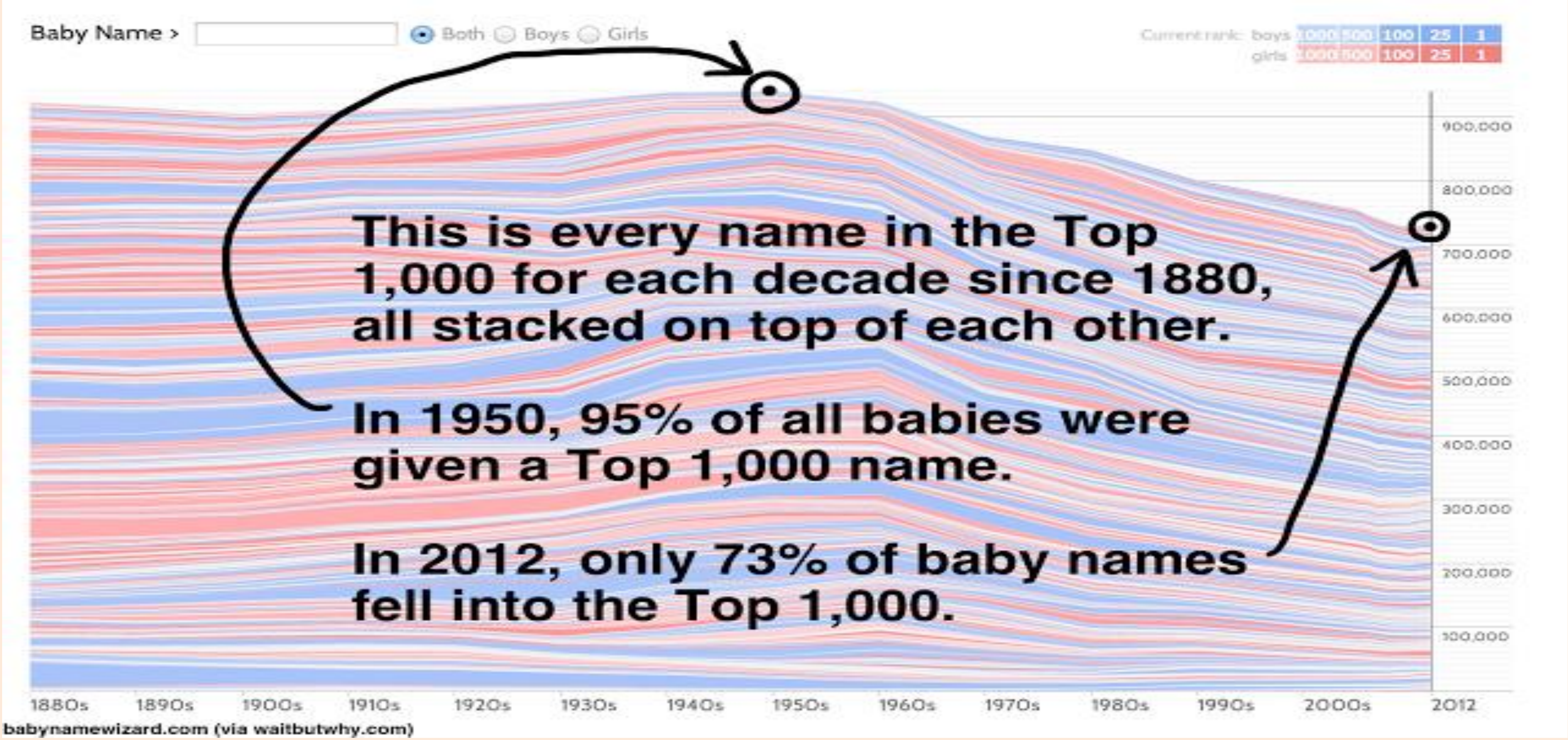
http://mjwest.org/vis/treemaps/

Cartogram

- Fancier version of treemaps:

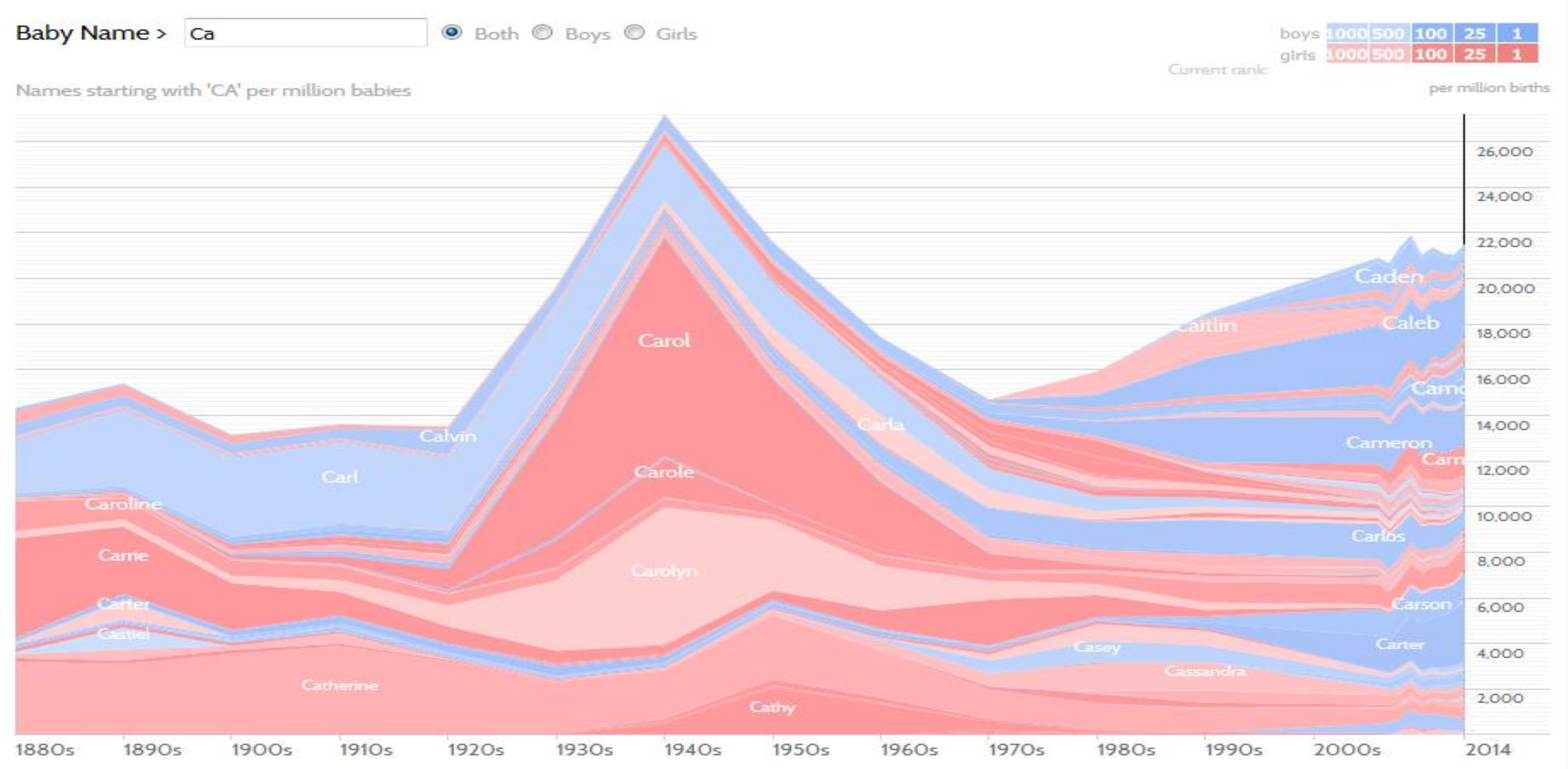
http://www-personal.umich.edu/~mejn/cartograms/

Stream Graph



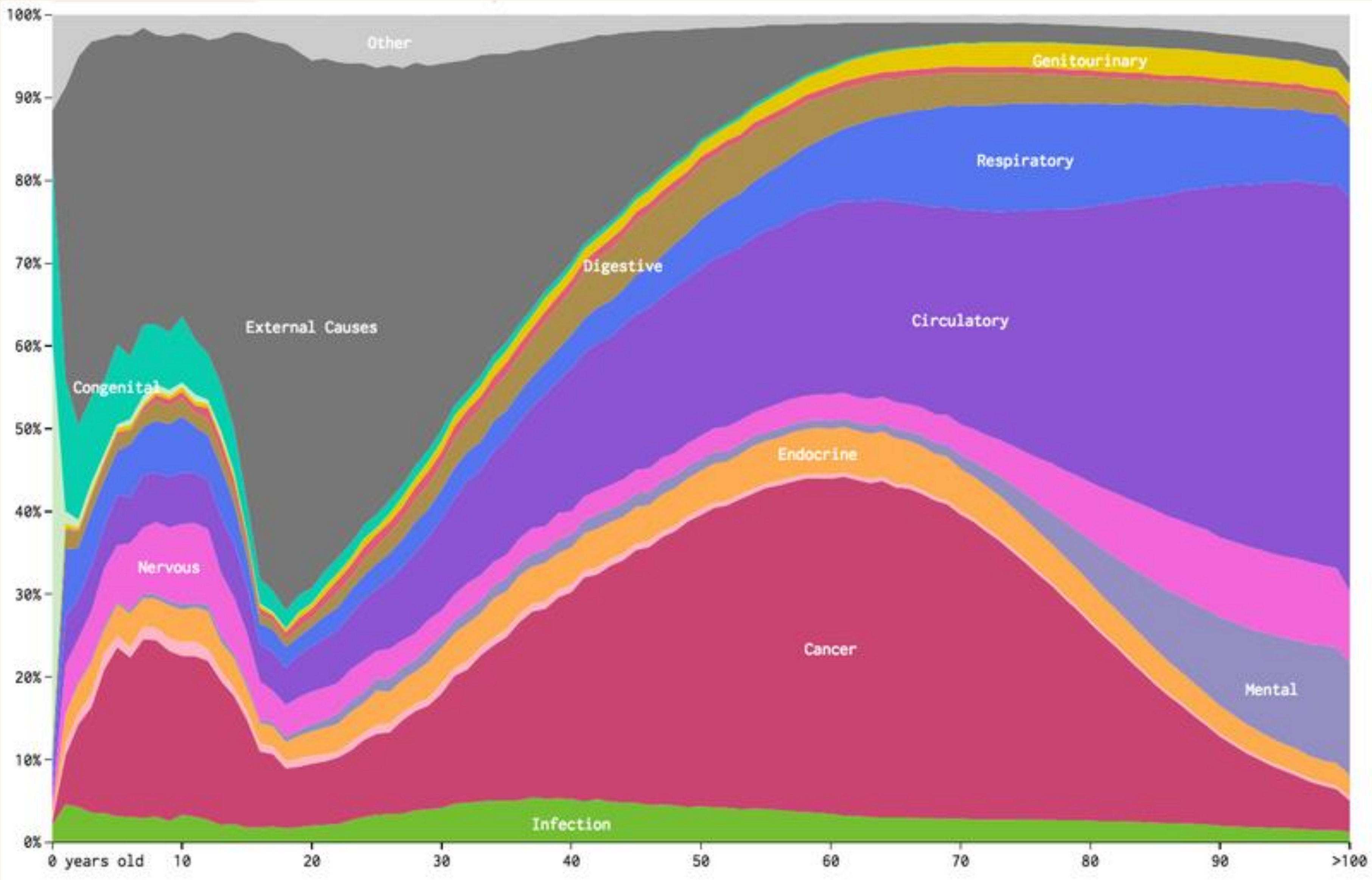
http://waitbutwhy.com/2013/12/how-to-name-baby.html

Stream Graph



http://www.babynamewizard.com/d3js-voyager/popup.html#prefix=ca&sw=both&exact=false

Stream Graph



http://www.vox.com/2016/5/10/11608064/americans-cause-of-death

Videos and Interactive Visualizations

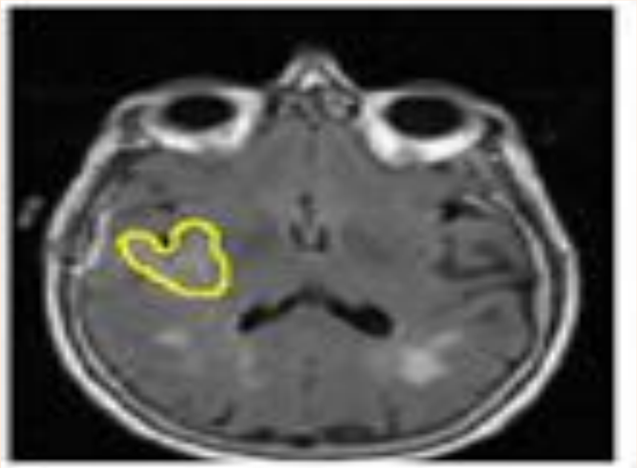
- For data recorded over time, videos can be useful:
  - Map colouring over time.
- There are also lots of neat interactive visualization methods:
  - Sale date for most expensive paintings.
  - Global map of wind, weather, and oceans.
  - Many examples here.

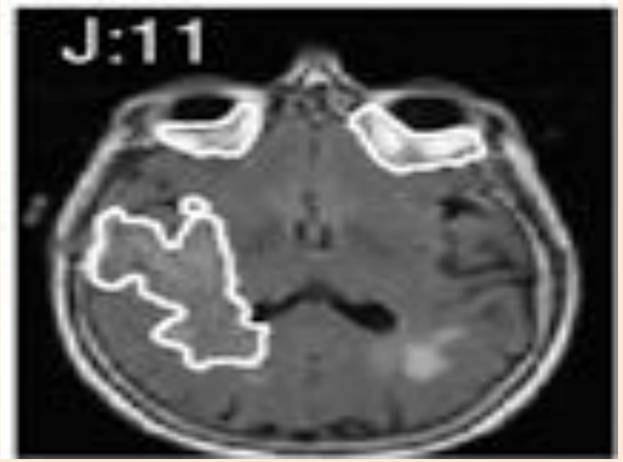
Hamming Distance vs. Jaccard Coefficient

A	B
1	0
1	0
1	0
0	1
0	1
1	0
0	0
0	0
0	1

- These vectors agree in 2 positions.
  - Normalizing Hamming distance by vector length, similarity is 2/9.
- If we're really interested in predicting 1s, we could find set of 1s in both and compute Jaccard:
  - A -> {1,2,3,6}, B -> {4,5,9}
  - No intersection so Jaccard similarity is actually 0.

Hamming Distance vs. Jaccard Coefficient





- Let’s say we want to find the tumour in an MR image.
- We have an expert label (top) and a prediction from our ML system (bottom).

- The normalized Hamming distance between the predictions at each pixel is 0.91. This sounds good, but since there are so many non-tumour pixels this is misleading.
- The ML system predicts a much bigger tumour so hasn’t done well. The Jaccard coefficient between the two sets of tumour pixels is only 0.11 so reflects this.

Coupon Collecting

- Consider trying to collect 50 uniformly-distributed states, drawing at random.
- The probability of getting a new state if there 'x' states left: p=x/50.
- So expected number of samples before next "success" (getting a new state) is 50/x.
  (mean of geometric random variable with p=x/50)
- So the expected number of draws is the sum of 50/x for x=1:50.
- For 'n' states instead of 50, summing until you have all 'n' gives:

$$\sum_{i=1}^{n} \frac{n}{i} = n \sum_{i=1}^{n} \frac{1}{i} \leq n(1 + \log(n)) = O(n \log n)$$

Huge Datasets and Parallel/Distributed Computation

- Most sufficient statistics can be computed in linear time.
- For example, the mean of 'n' numbers is computed as:

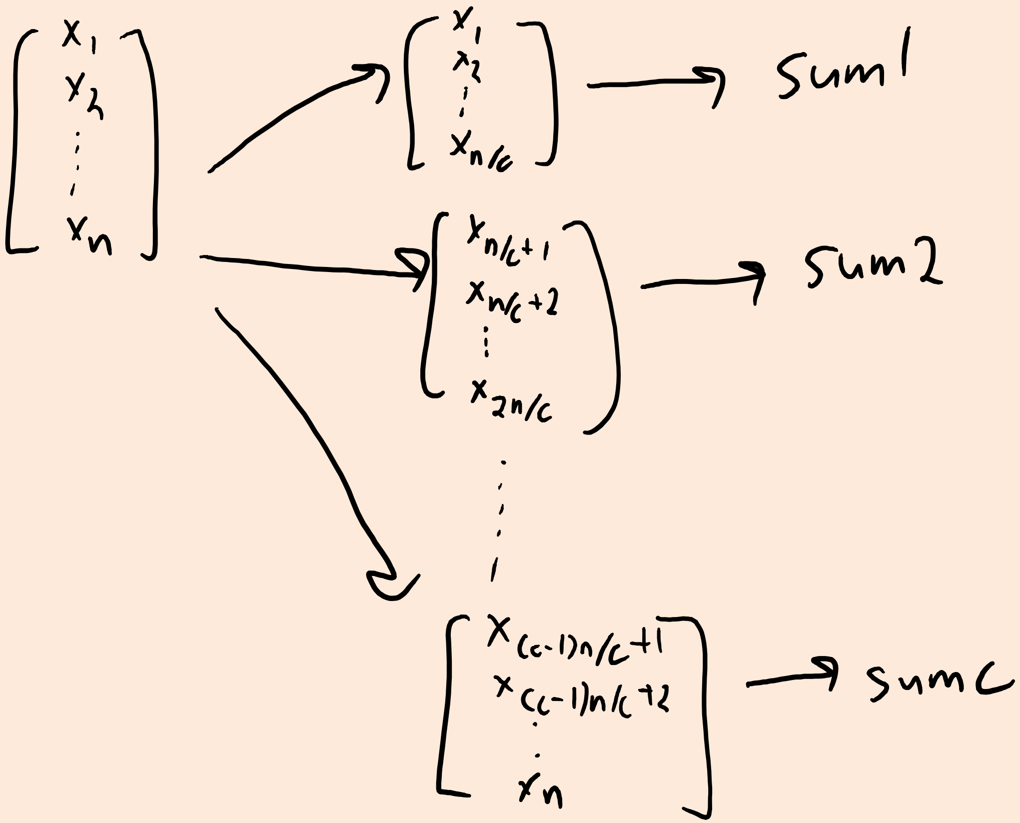
$$\text{mean}(x_1, x_2, x_3, \ldots, x_n) = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

- This costs O(n), which is great.

- But if 'n' is really big, we can go even faster with parallel computing…

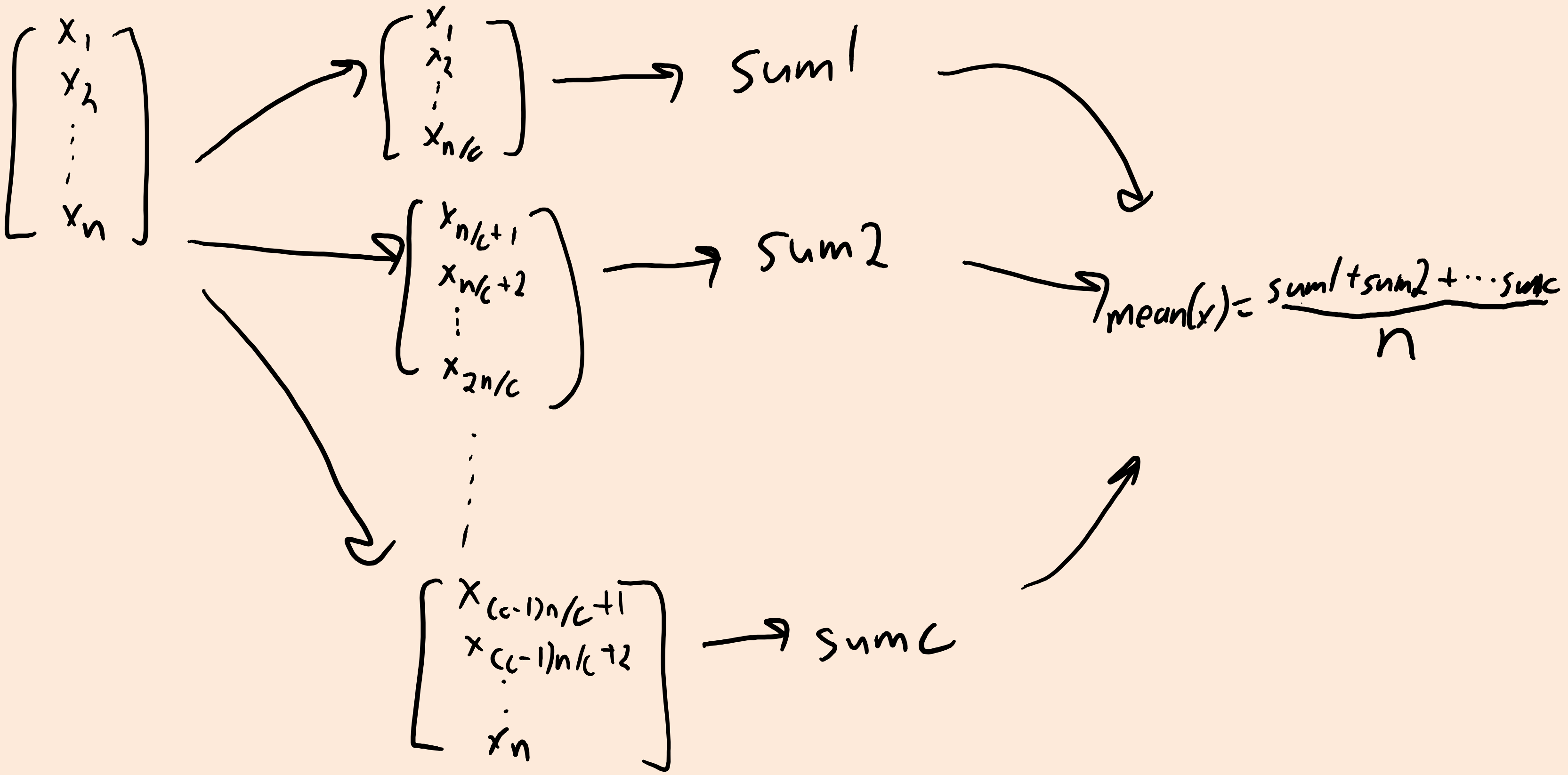
Huge Datasets and Parallel/Distributed Computation

- Computing the mean with multiple cores:
  - Each of the ‘c’ cores computes the sum of O(n/c) of the data:



Huge Datasets and Parallel/Distributed Computation

- Computing the mean with multiple cores:
  - Each of the ‘c’ cores computes the sum of O(n/c) of the data:
  - Add up the ‘c’ results from each core to get the mean.



Huge Datasets and Parallel/Distributed Computation

- Computing the mean with multiple cores:
  - Each of the ‘c’ cores computes the sum of O(n/c) of the data.
  - Add up the ‘c’ results from each core to get the mean.
  - Cost is only O(n/c + c), which can be much faster for large ‘n’.

- This assumes cores can access data in parallel (not always true).
- Can reduce cost to O(n/c) by having cores write to same register.
  - But need to “lock” the register and might effectively cost O(n).

Huge Datasets and Parallel/Distributed Computation

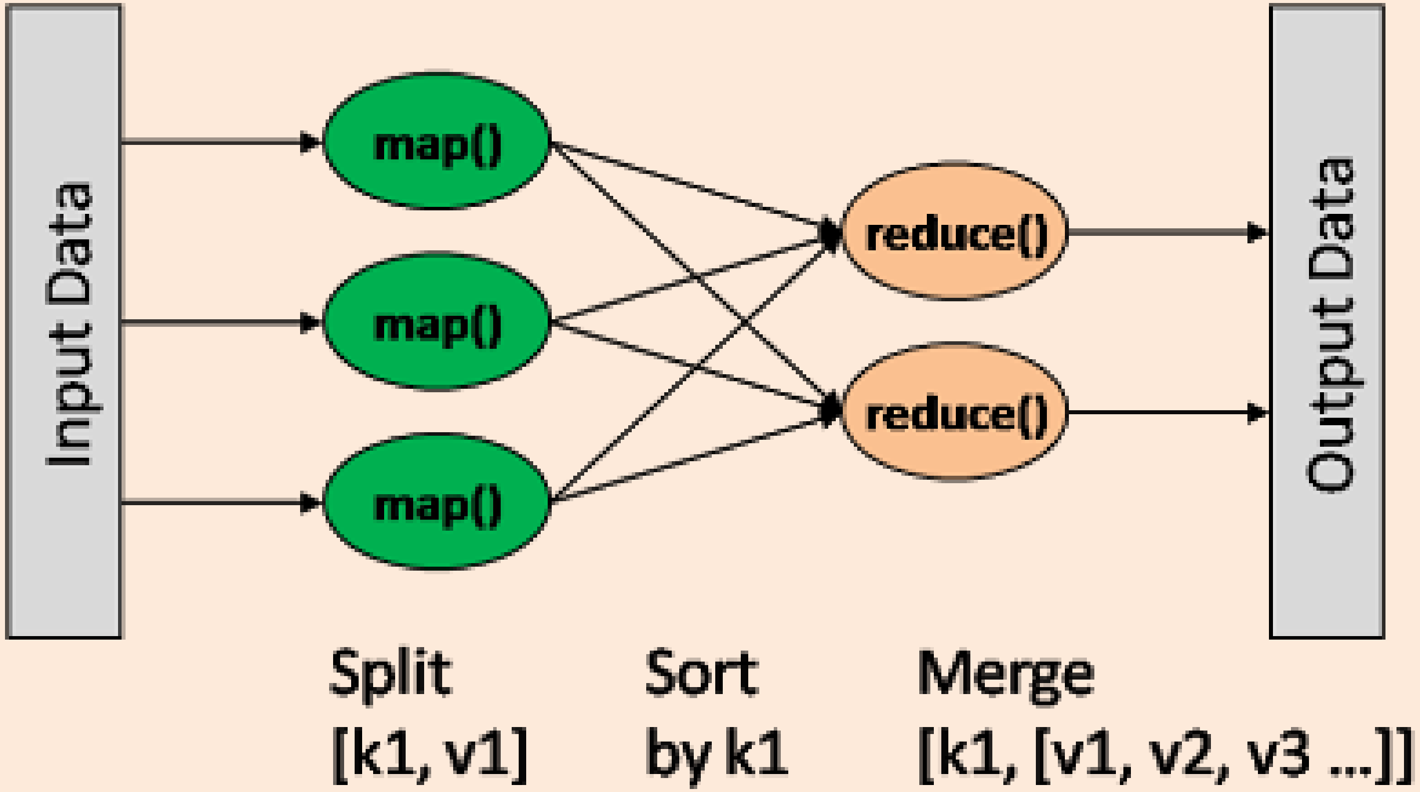
- Sometimes ‘n’ is so big that data can’t fit on one computer.

- In this case the data might be distributed across ‘c’ machines:
  - Hopefully, each machine has O(n/c) of the data.

- We can solve the problem similar to the multi-core case:
  - “Map” step: each machine computes the sum of its data.
  - “Reduce” step: each machine communicates sum to a “master” computer, which adds them together and divides by ‘n’.

Huge Datasets and Parallel/Distributed Computation

- Many problems in DM and ML have this flavour:
  - "Map" computes an operation on the data on each machine (in parallel).
  - "Reduce" combines the results across machines.



https://dzone.com/articles/how-hadoop-mapreduce-works

Huge Datasets and Parallel/Distributed Computation

- Many problems in DM and ML have this flavour:
  - "Map" computes an operation on the data on each machine (in parallel).
  - "Reduce" combines the results across machines.
  - These are standard operations in parallel libraries like MPI.

- Can solve many problems almost 'c' times faster with 'c' computers.

- To make it up for the high cost communicating across machines:
  - Assumes that most of the computation is in the "map" step.
  - Often need to assume data is already on the computers at the start.

Huge Datasets and Parallel/Distributed Computation

- Another challenge with "Google-sized" datasets:
  - You may need so many computers to store the data, that it's inevitable that some computers are going to fail.
- Solution to this is a distributed file system.

- Two popular examples are Google's MapReduce and Hadoop DFS:
  - Store data with redundancy (same data is stored in many places).
    - And assume data isn't changing too quickly.
  - Have a strategy for restarting "map" operations on computers that fail.
  - Allows fast calculation of more-fancy things than sufficient statistics:
    - Database queries and matrix multiplications.

Data Clean and the Duke Cancer Scandal

- See the Duke cancer scandal:
  - http://www.nytimes.com/2011/07/08/health/research/08genes.html?_r=2&hp

- Basic sanity checks for data cleanliness show problems in these (and many other) studies:
  - E.g., flipped labels, off-by-one mistakes, switched columns etc.
  - https://arxiv.org/pdf/1010.1092.pdf