CPSC 340:
Machine Learning and Data Mining

Data Exploration

Fall 2016

Admin

- Assignment 1 is coming over the weekend:
  - Start soon.
- Sign up for the course page on Piazza.
  - www.piazza.com/ubc.ca/winterterm12016/cpsc340/home
- Sign up for a CS undergrad account:
  - https://www.cs.ubc.ca/getacct
- Tutorials start next week:
  - Monday 4-5 and 5-6, Tueday 4:30-5:30, Wednesday 9-10.
  - Make sure you sign up for one.
  - No requirement to attend, but helps with assignments.
- Office hours:
  - Watch the website for details.

Outline

1) Typical steps in knowledge discovery from data.
2) Data Representations
3) Data Exploration

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

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)

Outline

1) Typical steps in knowledge discovery from data.
2) Data Representations
3) Data Exploration

What is Data?

- We'll define data as a collection of objects, and their features.

"feature"

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

"object"

- Each row is an object, each column is a feature.

Types of Data

- Discrete features come from an unordered set:
  - Binary: job?
  - Nominal/categorical: city.

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

Converting to Continuous Features

- Often want a real-valued object 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

- We can now interpret objects as points in space:
  - E.g., first object is at (23,1,0,0,22000).

Bag of Words

- Bag of words replaces document by word counts:

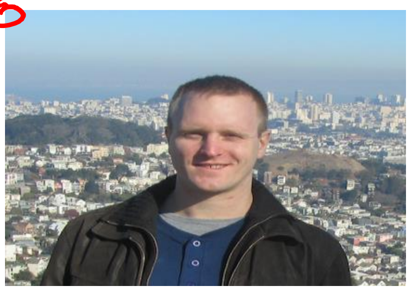
> The **International Conference on Machine Learning** (ICML) is the leading international academic conference in machine learning

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.

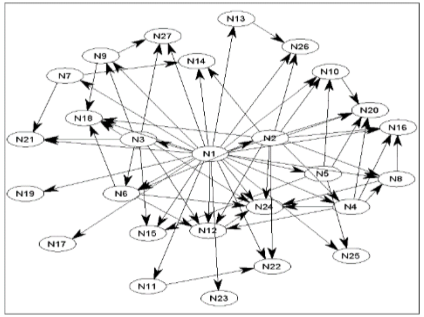
Other Data Types

- 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



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 (exact of otherwise).
- 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.

How much data do we need?

- Assume we have a categorical variable with 50 values: {Alabama, Alaska, Arizona, Arkansas,…}.
- We can turn this into 50 binary variables.
- If each category has equal probability, how many objects do we need to see before we expect to see each category once?
- Expected value is ~225.
- Coupon collector problem: O(n log n) in general.
  - Gotta Catch'em all!
- Need more data than categories:
  - Situation is worse if we don't have equal probabilities.
  - Typically want to see categories more than once.

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.

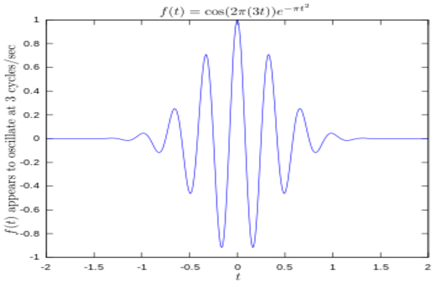
Feature Transformation

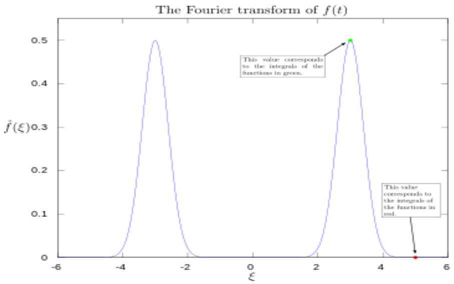
- Mathematical transformations:
  - Square, exponentiation, or take logarithm.

www.google.com/finance

Feature Transformation

- Mathematical transformations:
  - Square, exponentiation, or take logarithm.
  - Fourier or wavelet transform (signal data).

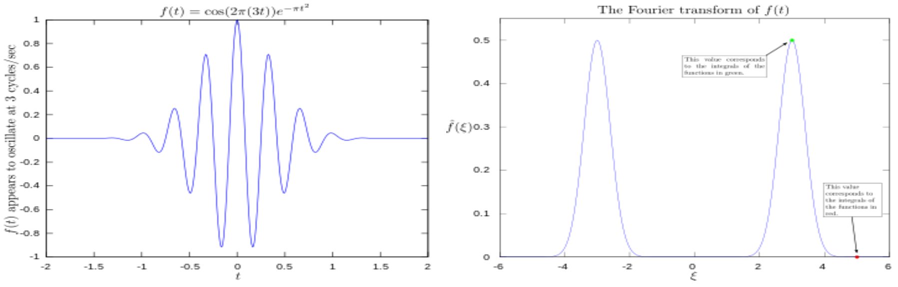


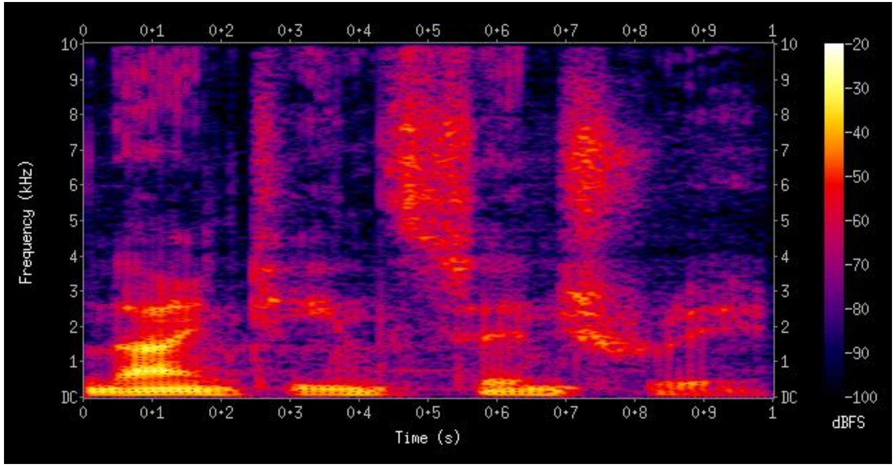


https://en.wikipedia.org/wiki/Fourier_transform
https://en.wikipedia.org/wiki/Spectrogram
https://en.wikipedia.org/wiki/Discrete_wavelet_transform

Feature Transformation

- Mathematical transformations:
  - Square, exponentiation, or take logarithm.
  - Fourier or wavelet transform (signal data).



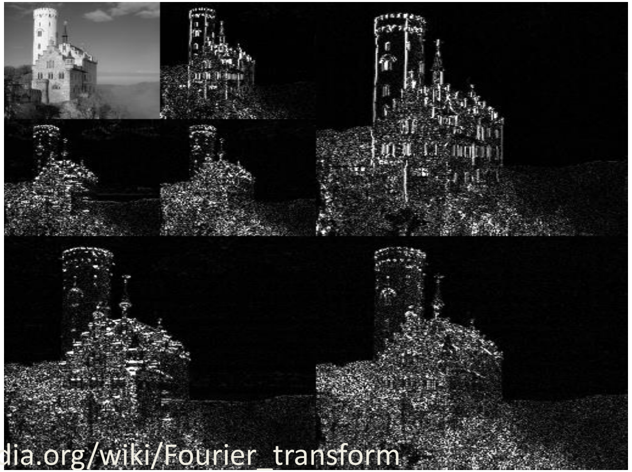


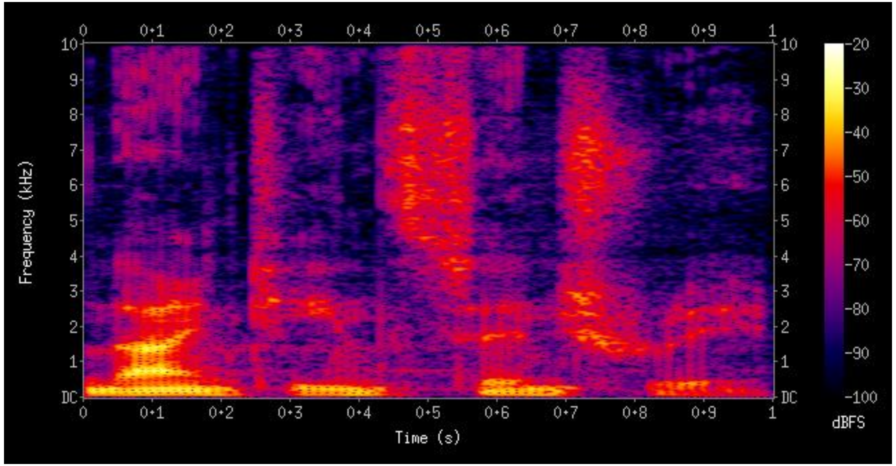
https://en.wikipedia.org/wiki/Fourier_transform
https://en.wikipedia.org/wiki/Spectrogram
https://en.wikipedia.org/wiki/Discrete_wavelet_transform

"Spectrogram"

Feature Transformation

- Mathematical transformations:
  - Square, exponentiation, or take logarithm.
  - Fourier or wavelet transform (signal data).





"Spectrogram"

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

Feature Transformation

- Mathematical transformations:
  - Square, exponentiation, or take logarithm.
  - Fourier or wavelet transform (signal data).
  - Discretization: turn continuous into discrete.

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

Feature Transformation

- Mathematical transformations:
  - Square, exponentiation, or take logarithm.
  - Fourier or wavelet transform (signal data).
  - Discretization: turn continuous into discrete.
  - Scaling: convert variables to comparable scales (E.g., convert kilograms to grams.)

Outline

1) Typical steps in knowledge discovery from data.
2) Data Representations
3) Data Exploration

Data Exploration

- You should always 'look' at the data first.

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

Discrete Summary Statistics

- Summary statistics for a discrete 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.

Discrete Summary Statistics

- Summary statistics between discrete variables:
  - Simple matching coefficient:
    - How many times two variables are the same.
    - If C_{ab} be “number of times variable 1 is a and variable 2 is b”:
      - Simple matching for binary would be $(C_{11} + C_{00})/(C_{00} + C_{01} + C_{10} + C_{11})$.
  - Jaccard coefficient for binary variables:
    - Intersection divided by union of ‘1’ values.
    - $C_{11}/(C_{01} + C_{10} + C_{11})$.

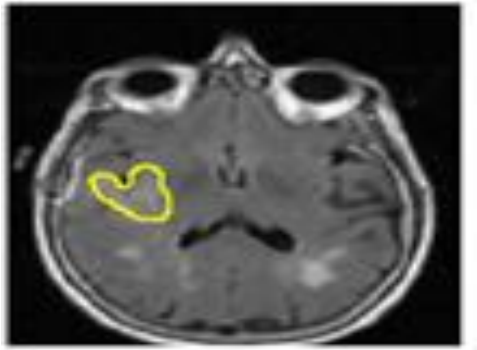
Simple Matching vs. Jaccard

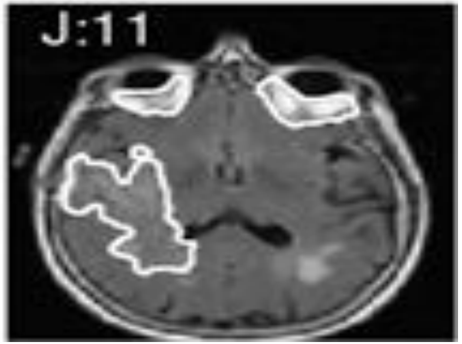
A	B
1	0
1	0
1	0
0	1
0	1
1	0
0	0
0	0
0	1

$$\begin{aligned}\mathrm{Sim}(A,B) &= (C_{11} + C_{00})/(C_{00} + C_{01} + C_{10} + C_{11}) \\ &= (0 + 2)/(2 + 3 + 4 + 0) \\ &= 2/9.\end{aligned}$$

$$\begin{aligned}\mathrm{Jac}(A,B) &= C_{11}/(C_{01} + C_{10} + C_{11}) \\ &= 0/(3 + 4 + 0) \\ &= 0.\end{aligned}$$

Simple Matching vs. Jaccard





Sim(A,B) = 0.91

Jac(A,B) = 0.11

Continuous Summary Statistics

- Measures of location:
  - Mean: average value (sensitive to outliers).
  - 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.
  - Intequantile ranges: difference between quantiles.

Continuous Summary Statistics

- Data: [0 1 2 3 3 5 7 8 9 10 14 15 17 200]
- 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.

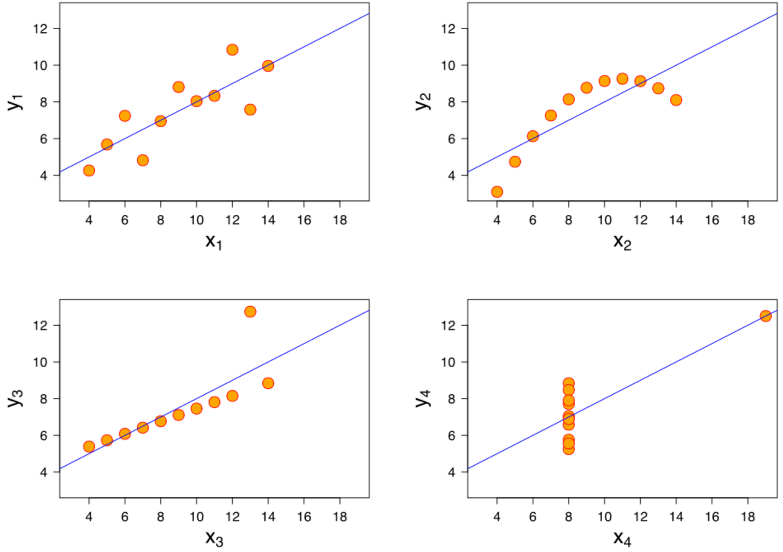
Continuous Summary Statistics

- Measures between continuous variables:
  - Correlation:
    - Does one increase/decrease proportionally as the other increases?
  - Rank correlation:
    - Does one increase/decrease as the other increases?
  - Euclidean distance:
    - How far apart are the values?
  - Cosine similarity:
    - What is the angle between them?

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.
  - Alsmot same correlations.
  - Look completely different.



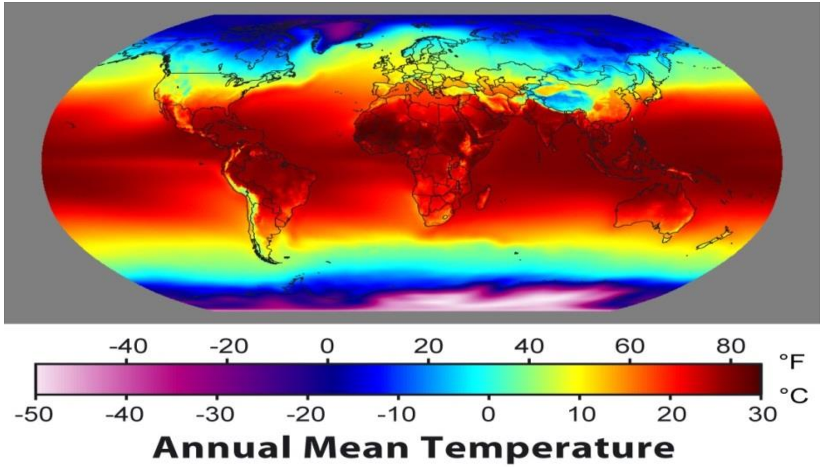
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

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
…	…	…

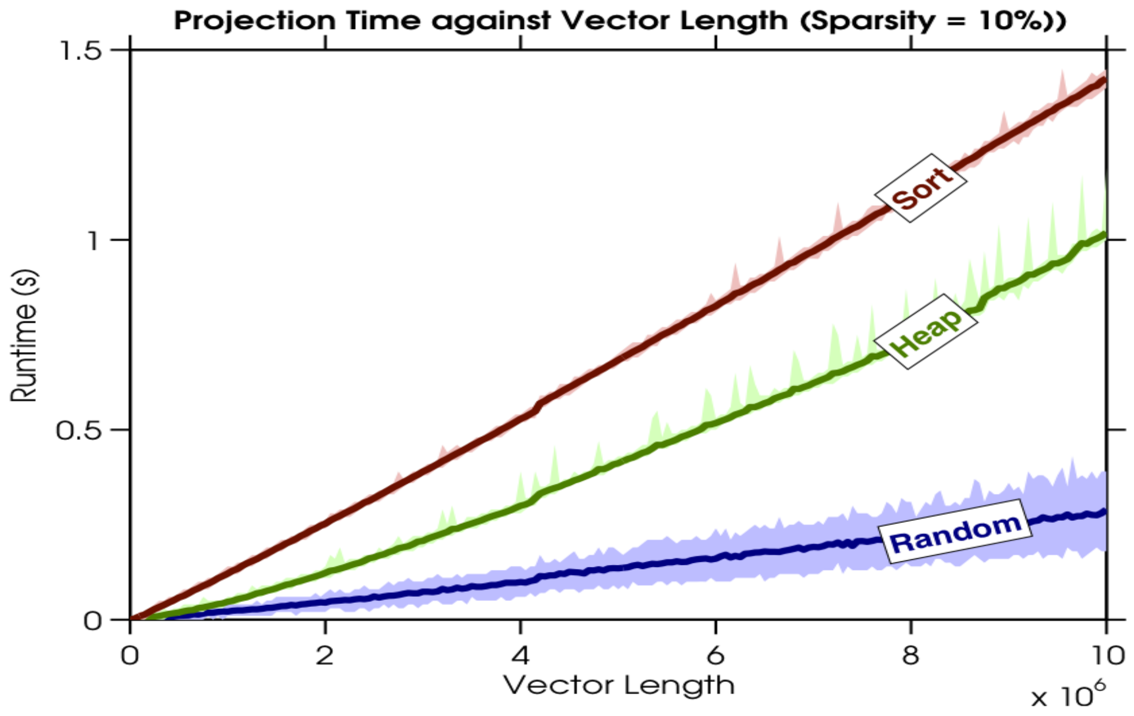
vs.



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

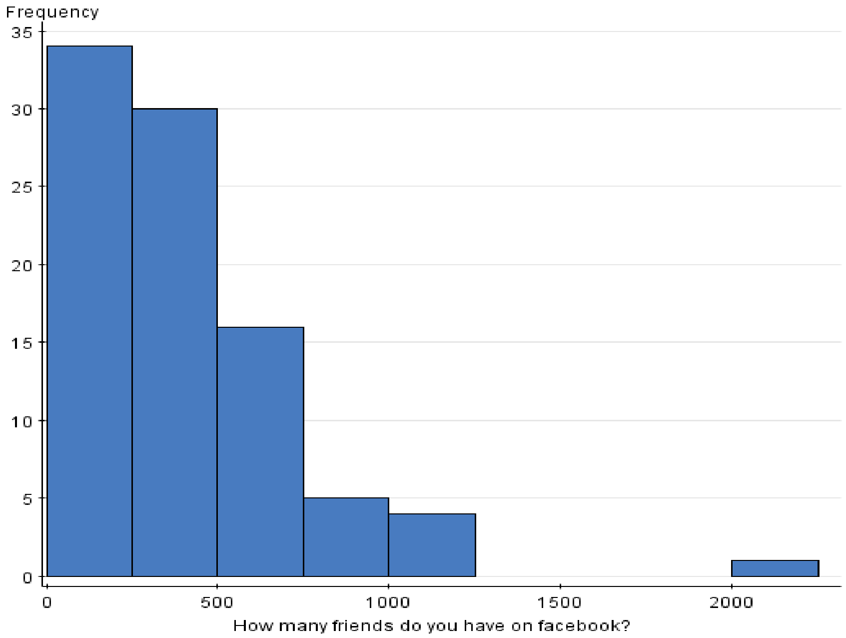
Basic Plot

- Visualize one variable as a function of another.



Histogram

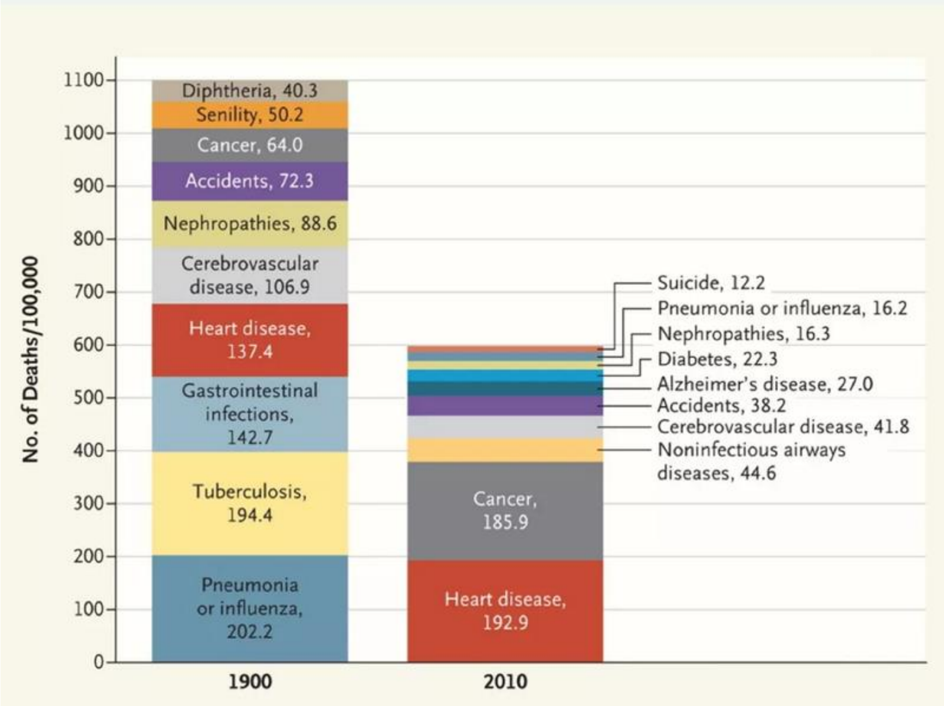
- Histograms display distribution of a variable.



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

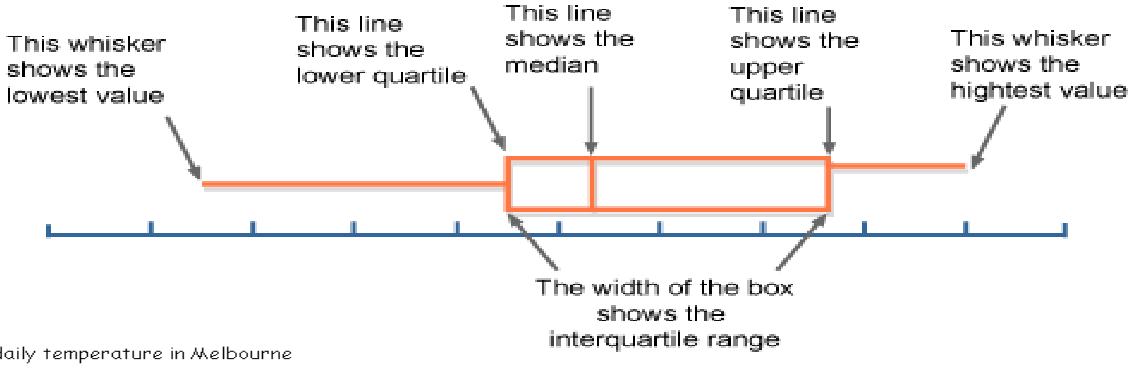
Histogram

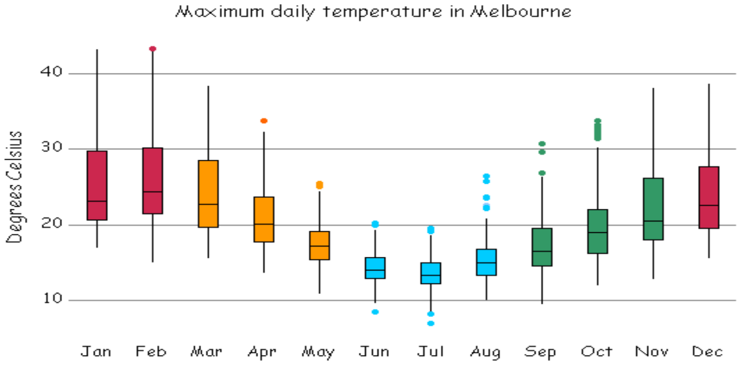
- Histogram with grouping:

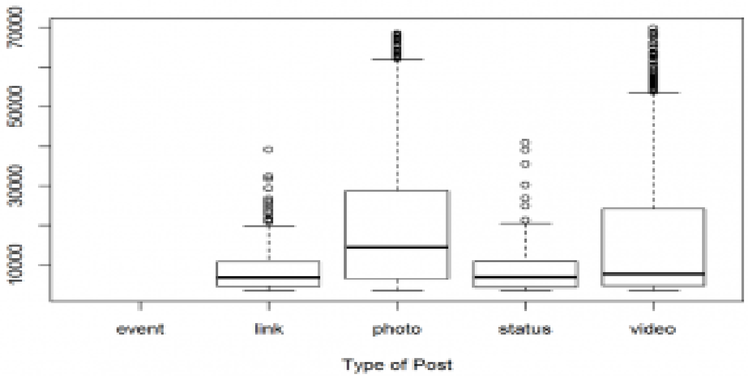


http://www.vox.com/2016/5/10/11608064/americans-cause-of-death

Box Plot







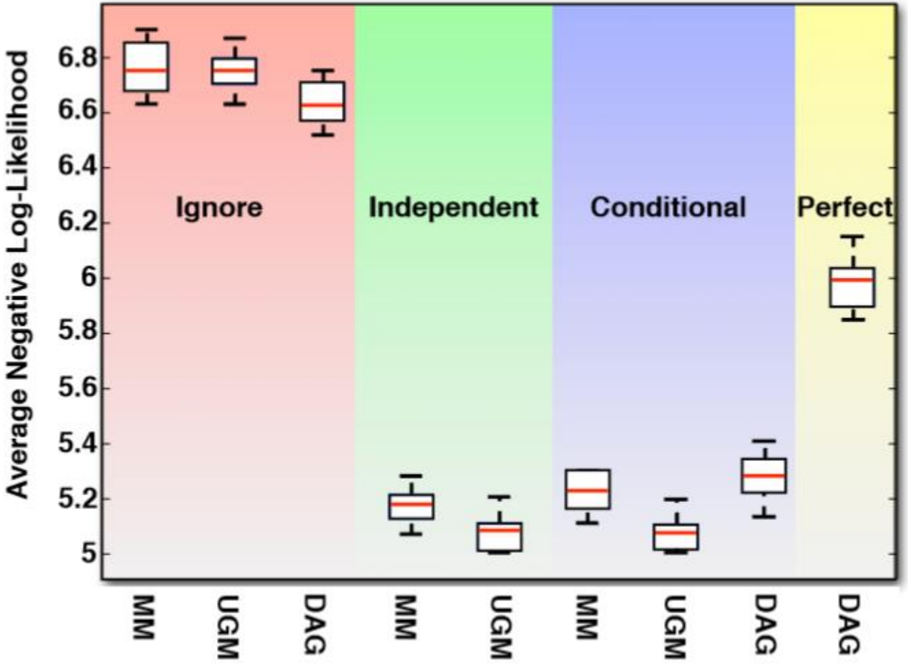
http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/representingdata3hirev6.shtml
http://www.scc.ms.unimelb.edu.au/whatisstatistics/weather.html

Box Plot

- Photo from CTV Olympic coverage in 2010:

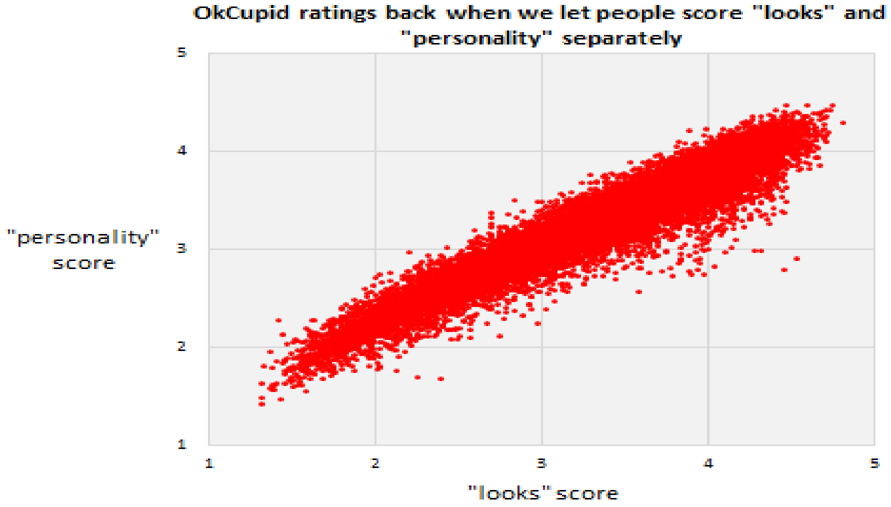
Box Plots

- Box plot with grouping:



Scatterplot

- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.

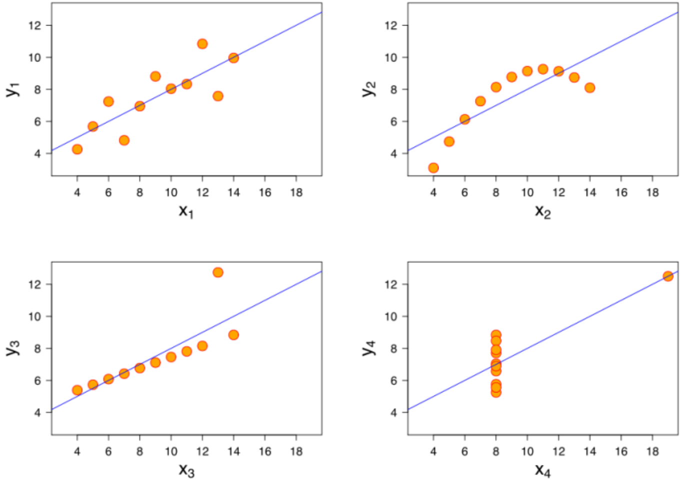


http://cdn.okccdn.com/blog/humanexperiments/looks-v-personality.png

- 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.

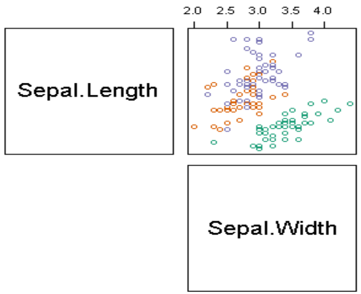


https://en.wikipedia.org/wiki/Anscombe%27s_quartet

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

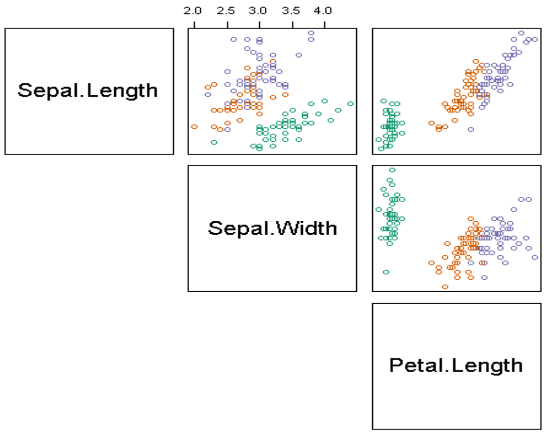
Scatterplot Arrays

- For multiple variables, can use scatterplot array.



Scatterplot Arrays

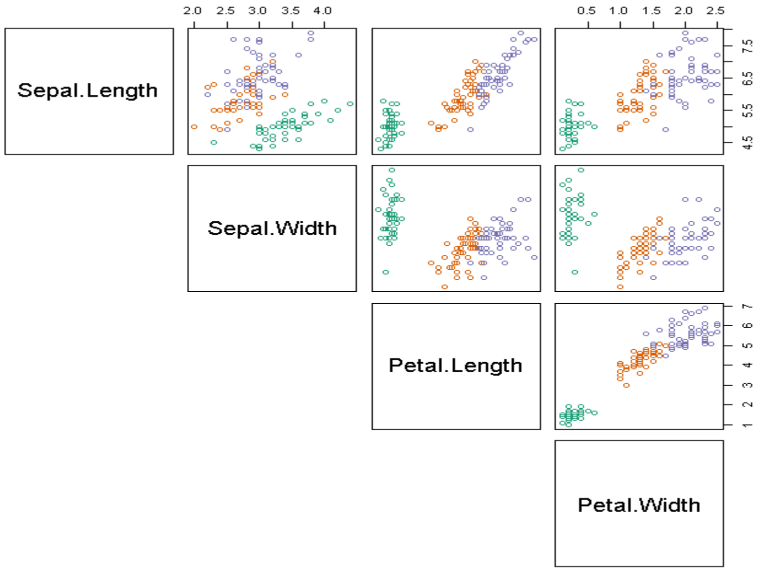
- For multiple variables, can use scatterplot array.



http://www.ats.ucla.edu/stat/r/pages/layout.htm

Scatterplot Arrays

- For multiple variables, can use scatterplot array.



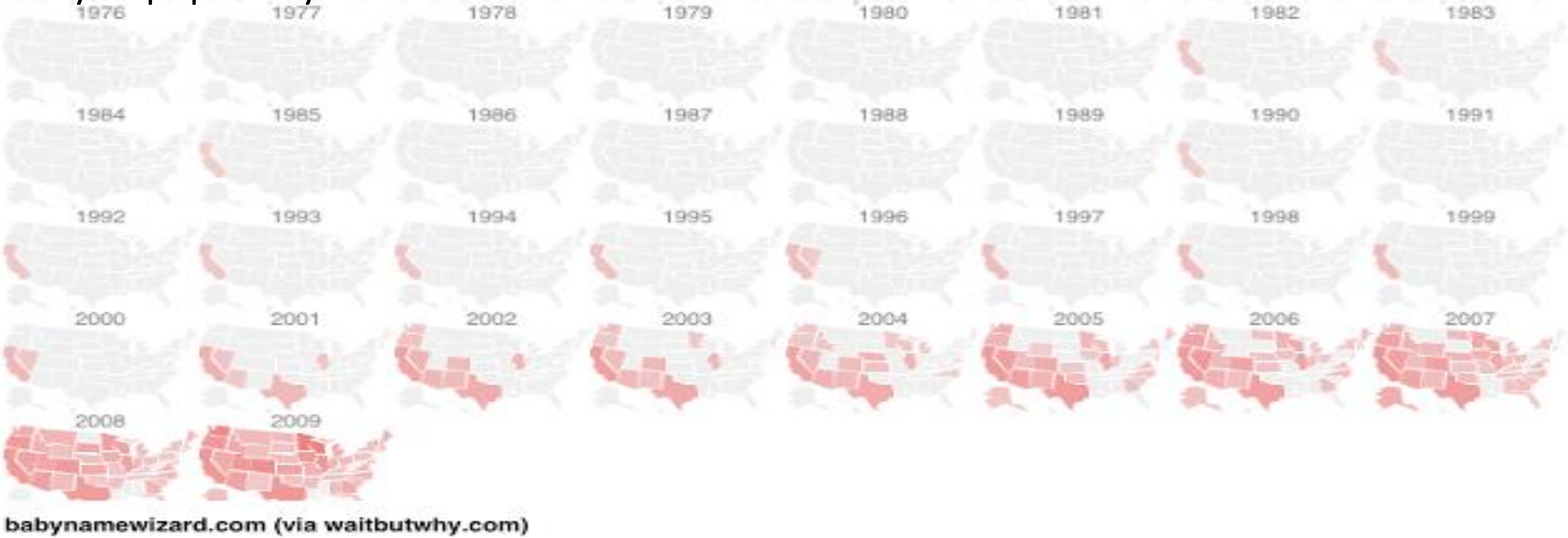
- Colors can indicate a third categorical variable.

http://www.ats.ucla.edu/stat/r/pages/layout.htm

Map Coloring

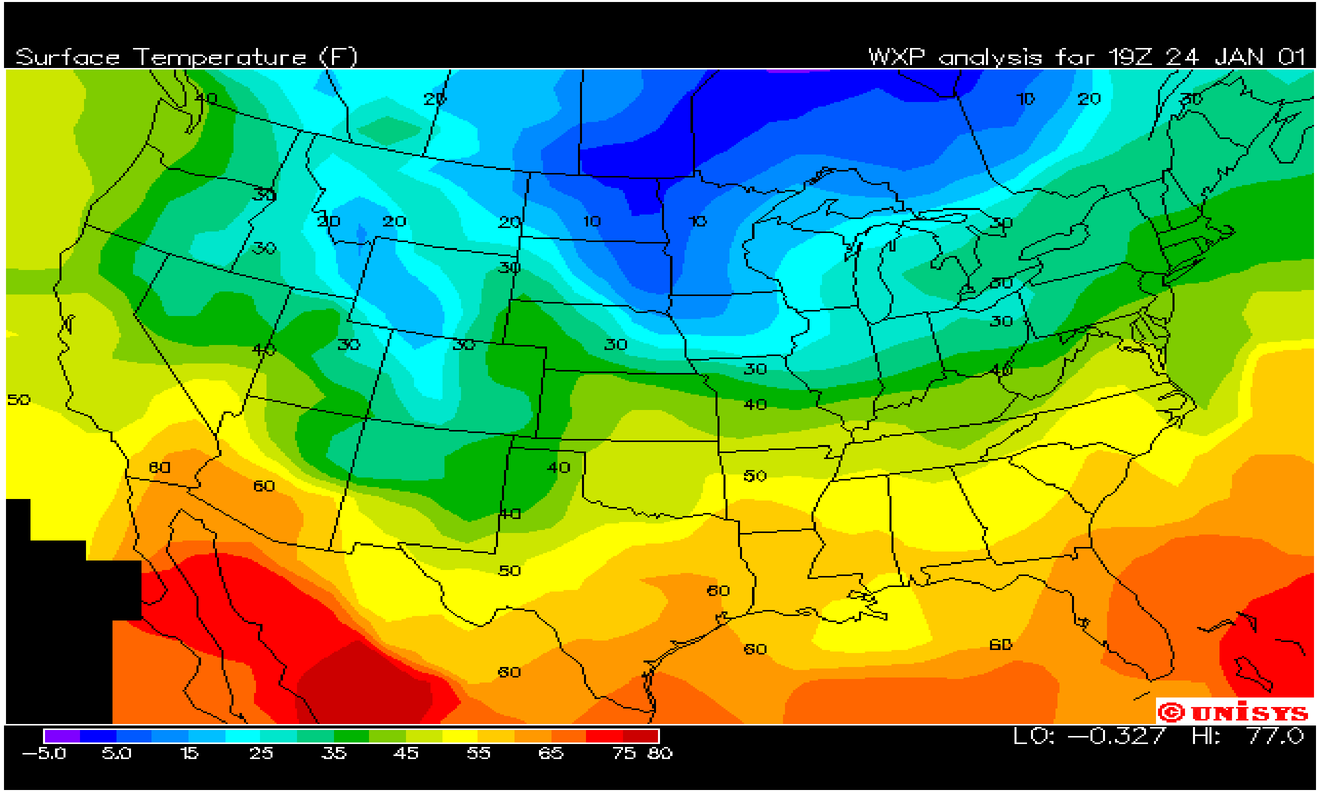
- Color/intensity can represent feature of region.

Evelyn's popularity over time:



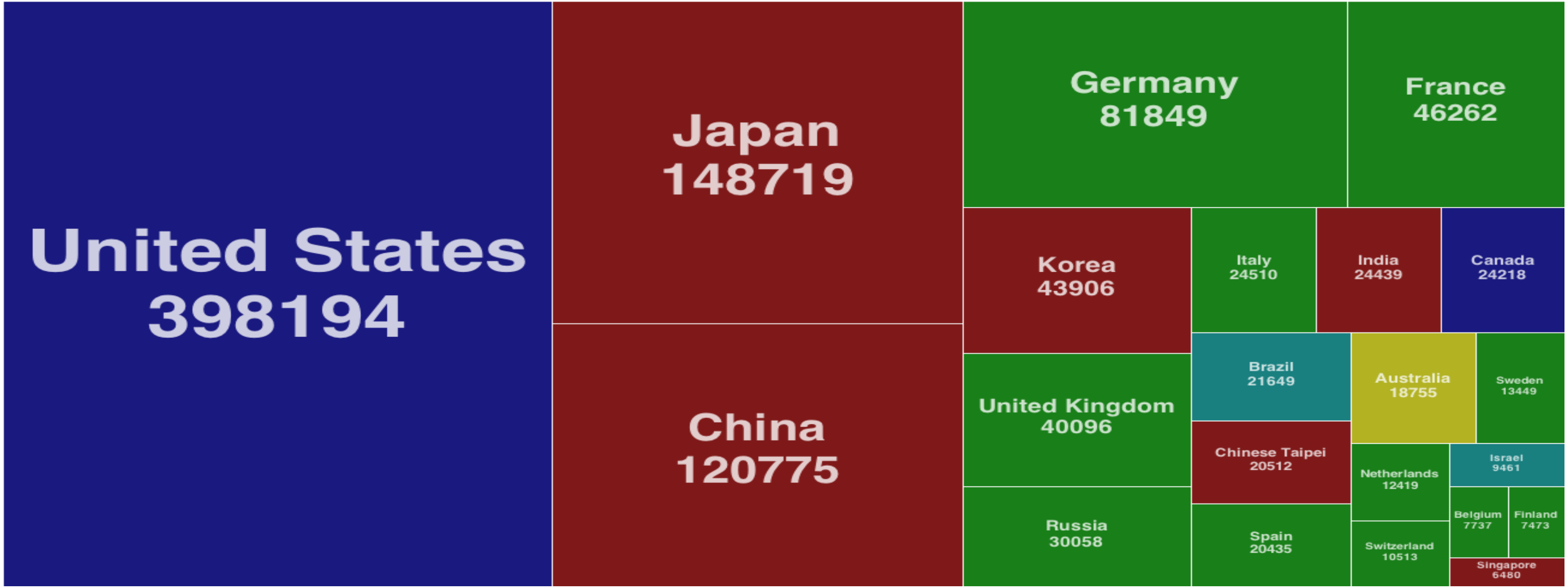
http://waitbutwhy.com/2013/12/how-to-name-baby.html

Contour Plot



Treemaps

- Area represents attribute value:



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

Cartogram

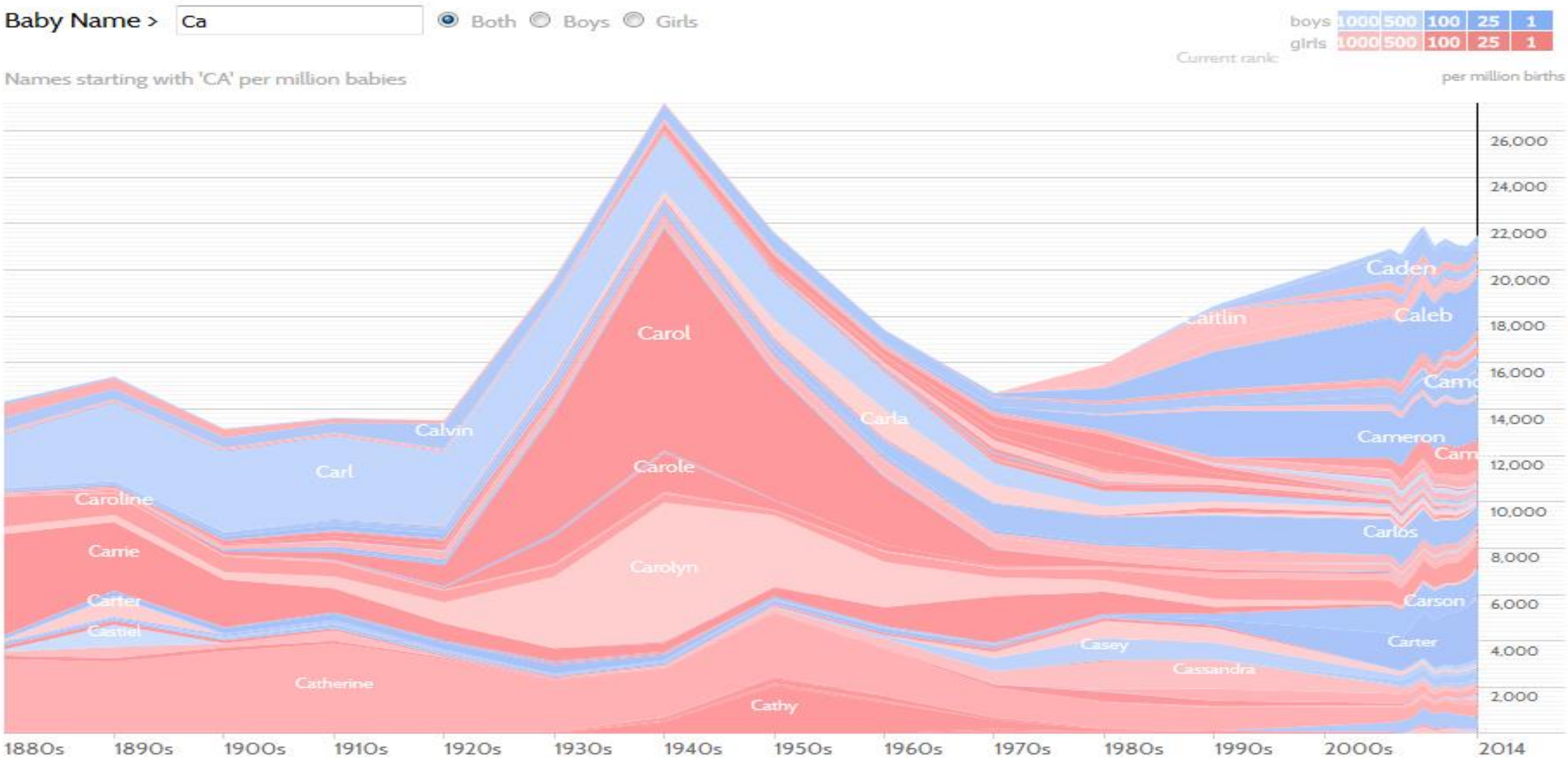
- Fancier version of treemaps:

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

Stream Graph

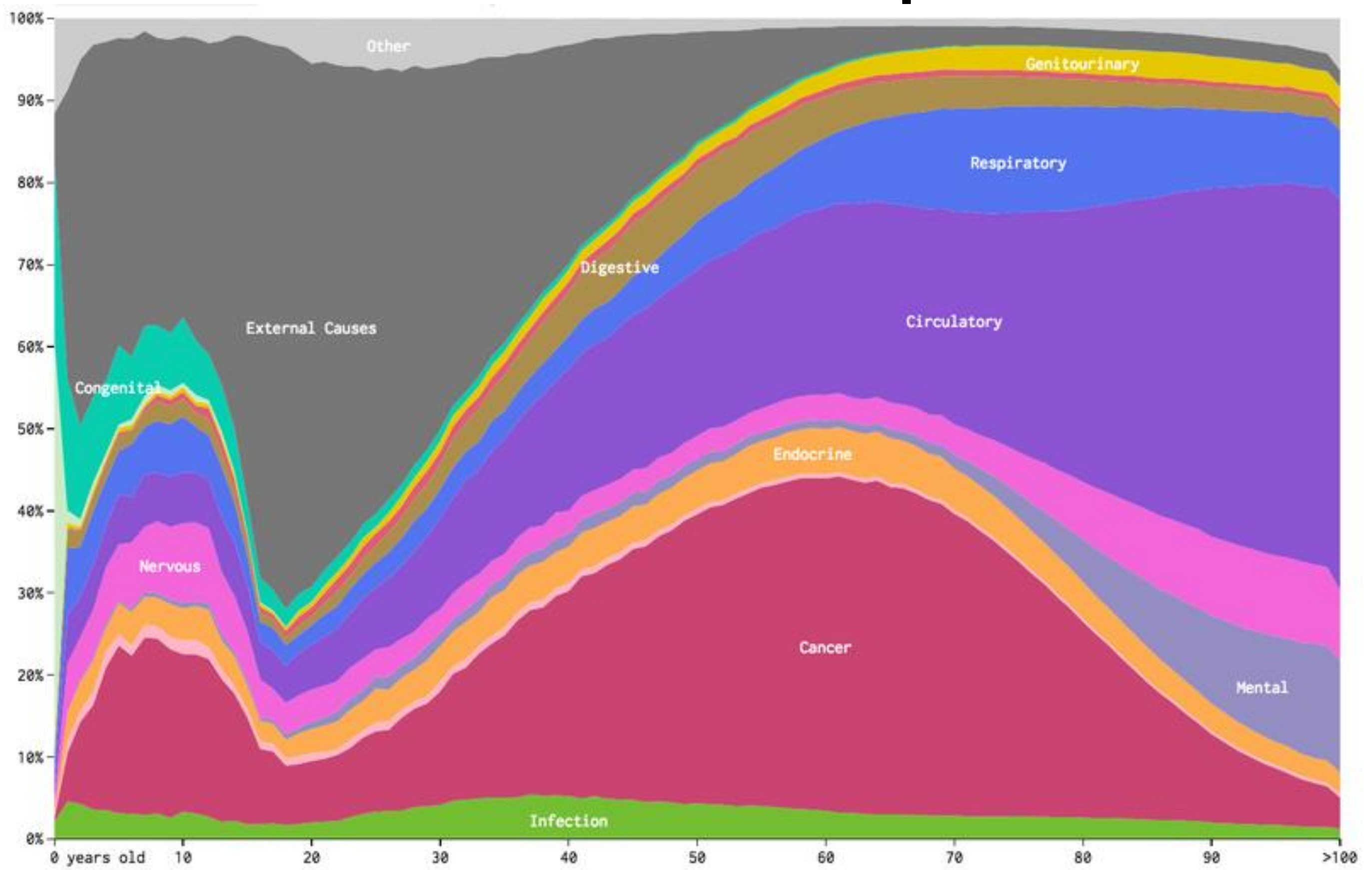
http://waitbutwhy.com/2013/12/how-to-name-baby.html

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

Summary

- Typical data mining steps:
  - Involves data collection, preprocessing, analysis, and evaluation.
- Object-feature representation and discrete/numerical features.
- Data preprocessing:
  - Data cleaning, feature transformations.
- Exploring data:
  - Summary statistics and data visualization.

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

Bonus Slide: Coupon Collecting

- Since the probability of obtaining a new state if there are 'x' states you don't have is p = x/50, the average number of states you need to pick (mean of geometric random variable with p=x/50) to get a new one is 1/p = 50/x.
- For 'n' states, summing until you have all 'n' gives:

$$\sum_{i=1}^{n} \frac{n}{i} = n \underbrace{\sum_{i=1}^{n} \frac{1}{i}}_{O(\log n)} = O(n \log n)$$