CPSC 340:
Machine Learning and Data Mining

Data Exploration

BONUS SLIDES

How much data do we need?

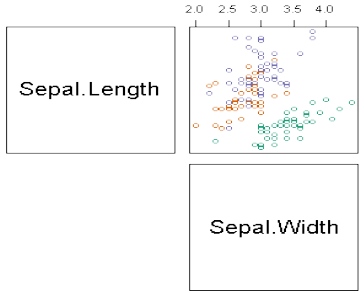
- Assume we have a categorical variable with 50 values: {Alabama, Alaska, Arizona, Arkansas,…}.
- We can turn this into 50 binary variables.
- If each category has equal probability, how many objects do we need to see before we expect to see each category once?
- Expected value is ~225.
- Coupon collector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

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?

Scatterplot Arrays

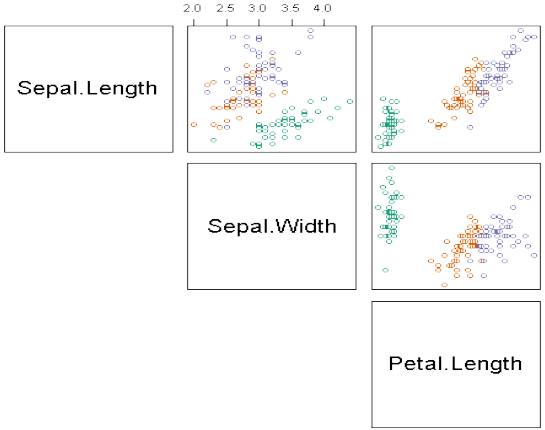
- For multiple variables, can use scatterplot array.



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

Scatterplot Arrays

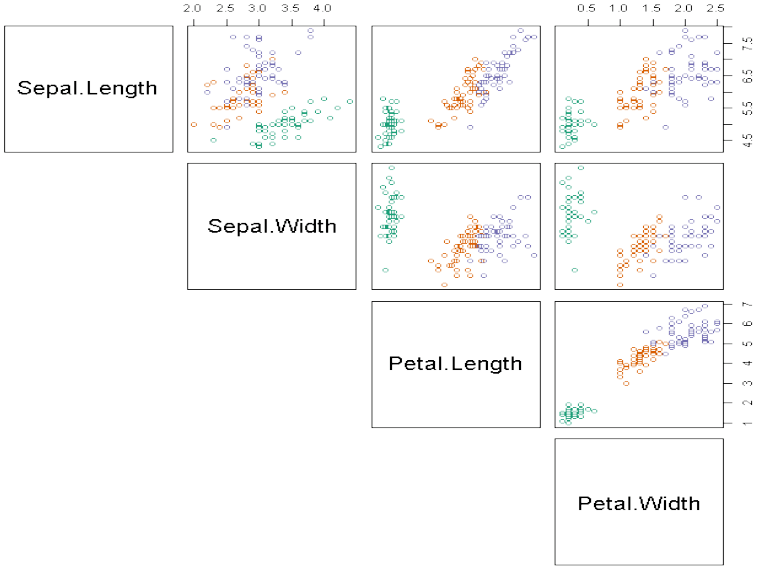
- For multiple variables, can use scatterplot array.



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

Scatterplot Arrays

- For multiple variables, can use scatterplot array.



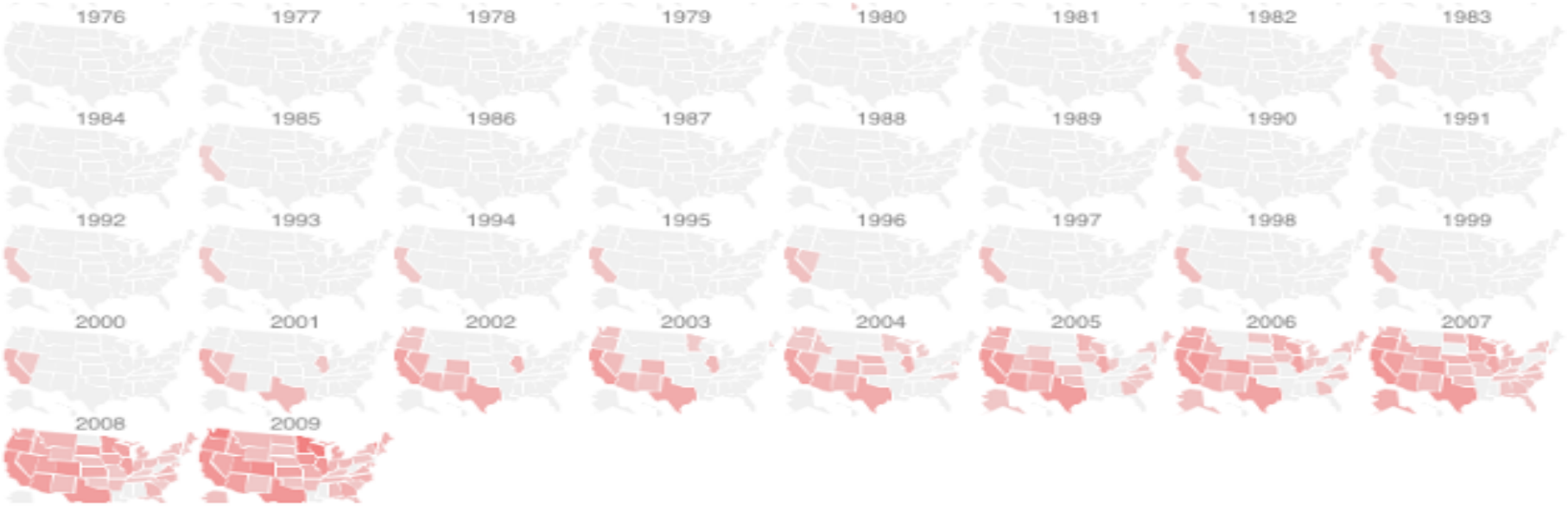
- Colors can indicate a third categorical variable.

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

Map Coloring

- Color/intensity can represent feature of region.

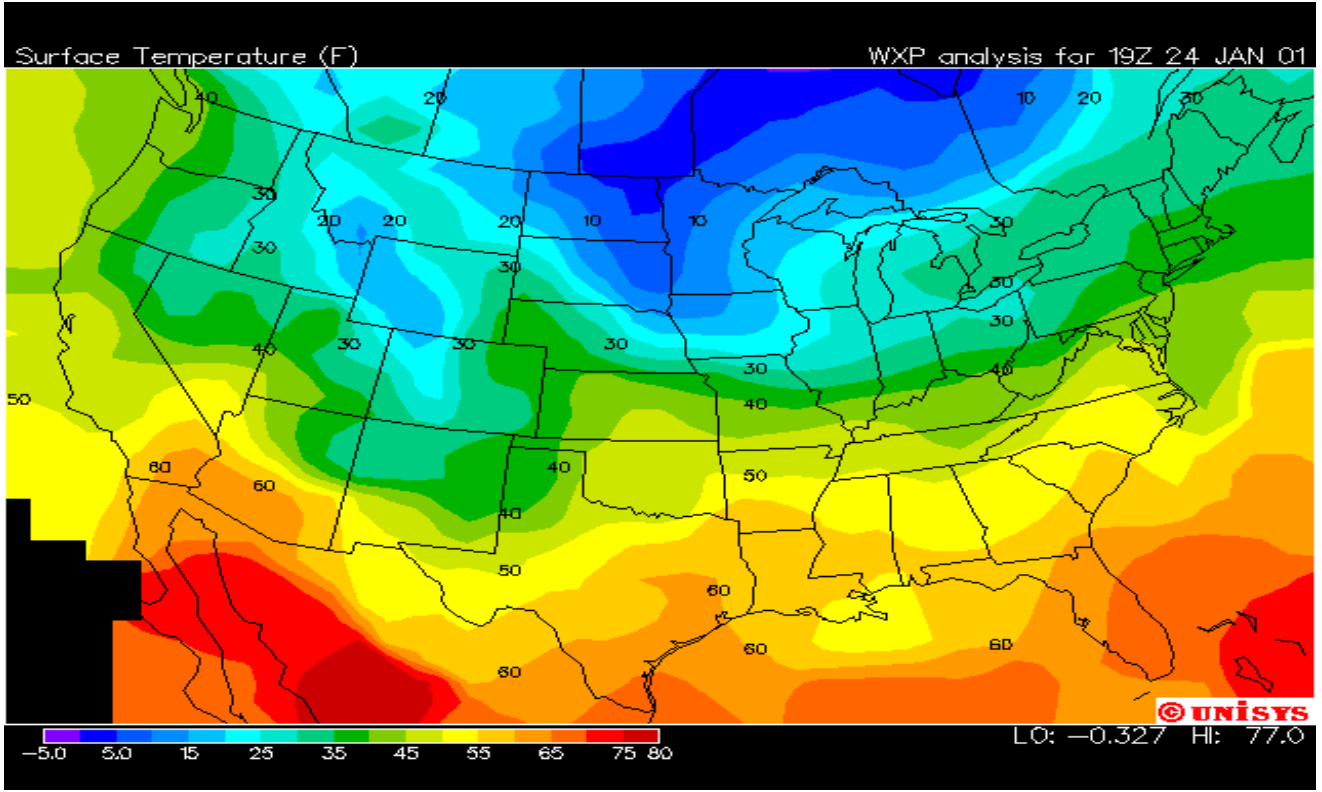
Popularity over time of the name “Evelyn”:



babynamewizard.com (via waitbutwhy.com)

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

Contour Plot



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

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:

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

- The actual sum is slightly more than log(n) since $\int_1^n \frac{1}{x} dx = \log(n)$

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.
    - $C_{11}/(C_{01} + C_{10} + C_{11})$.

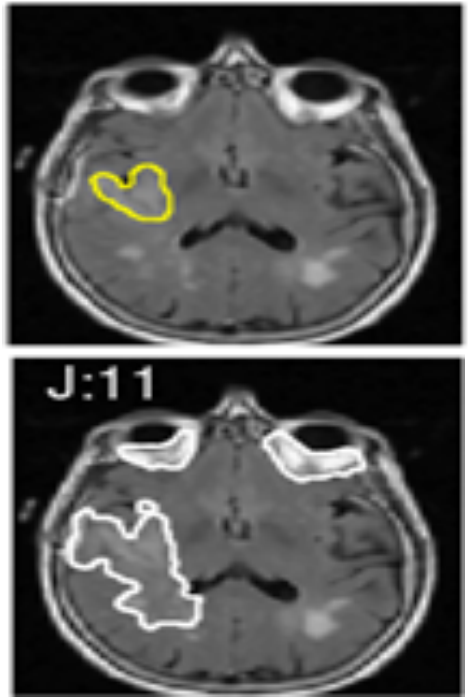
Simple Matching vs. Jaccard

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

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

$$\text{Jac}(A,B) = C_{11}/(C_{01} + C_{10} + C_{11})$$
$$= 0/(3 + 4 + 0)$$
$$= 0.$$

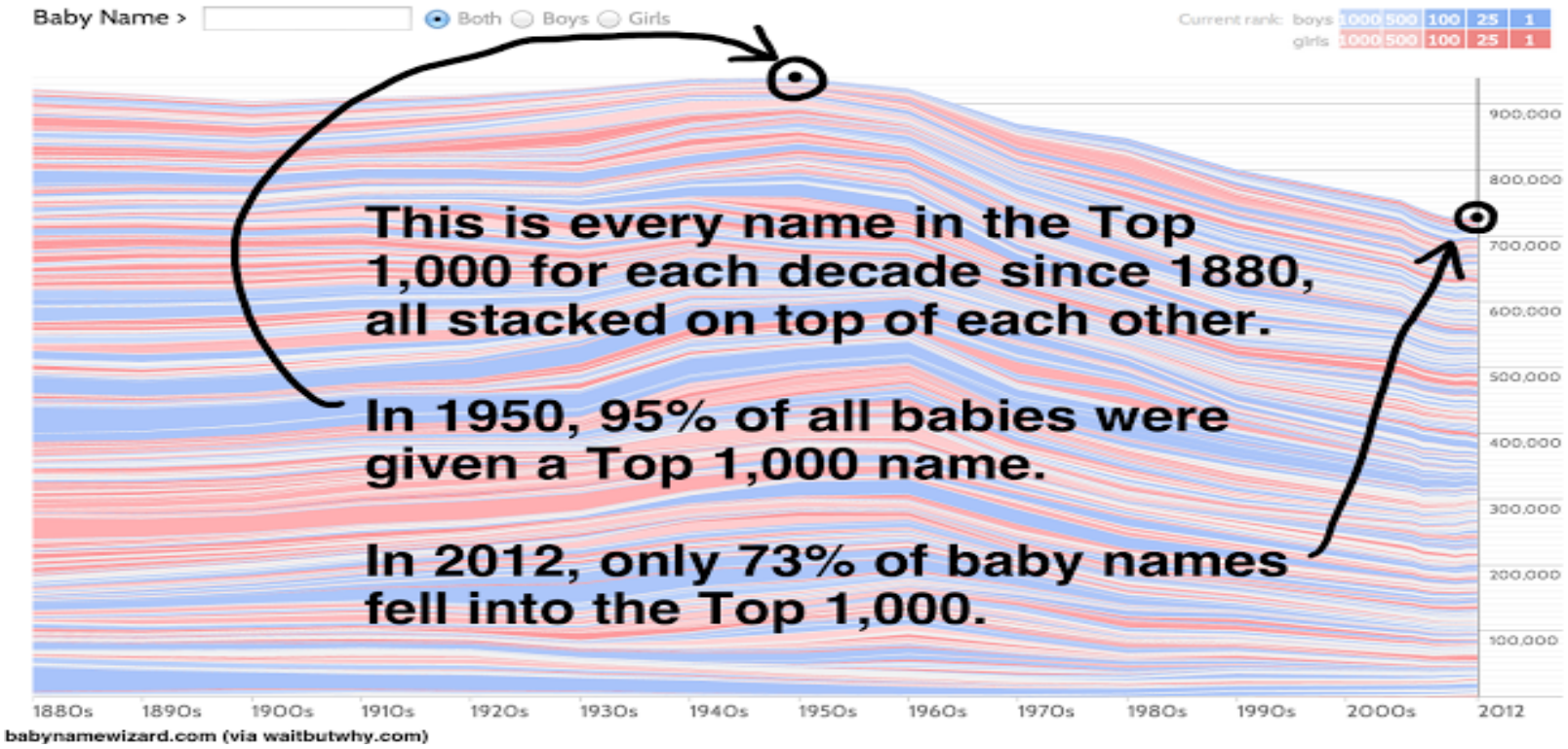
Simple Matching vs. Jaccard



Sim(A,B) = 0.91

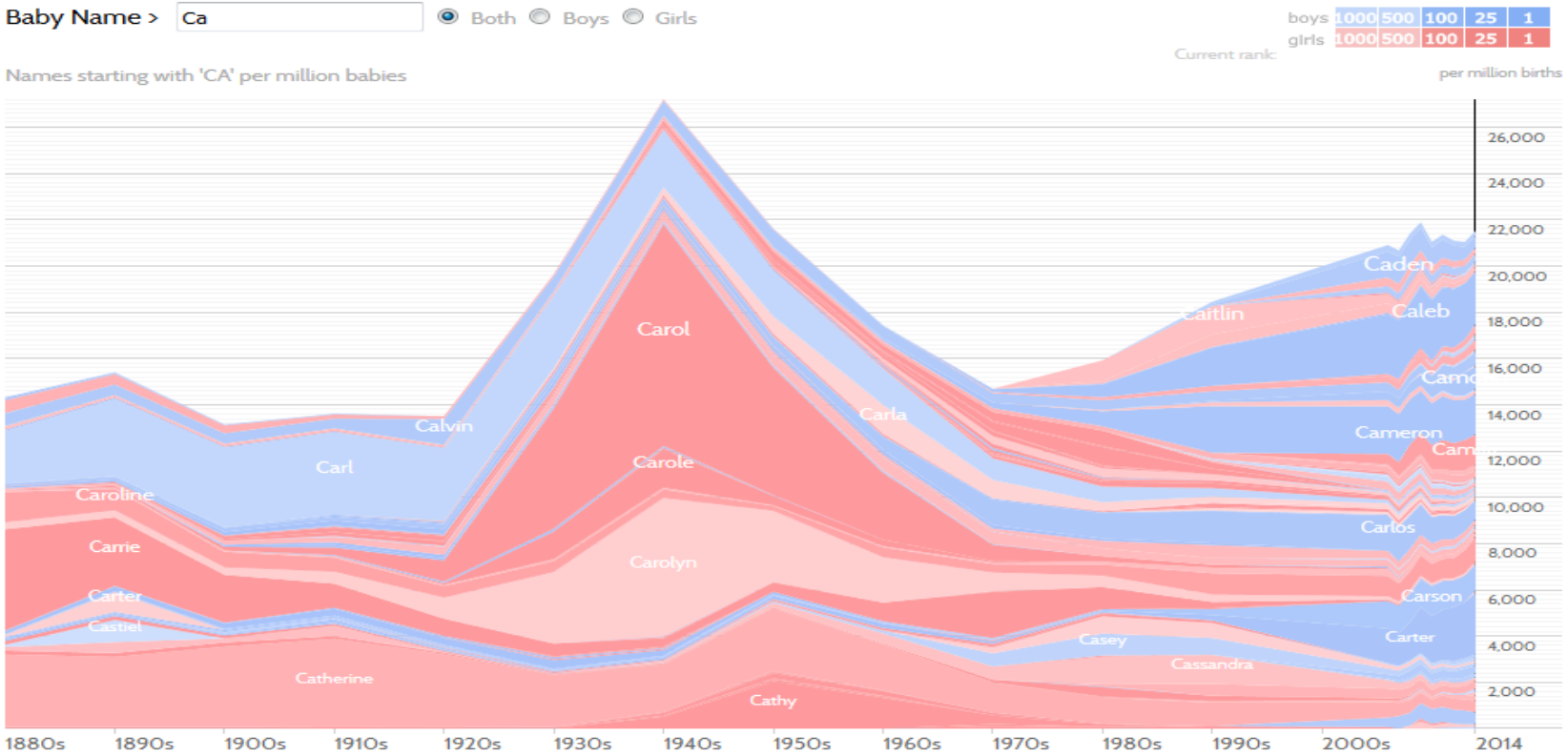
Jac(A,B) = 0.11

Stream Graph



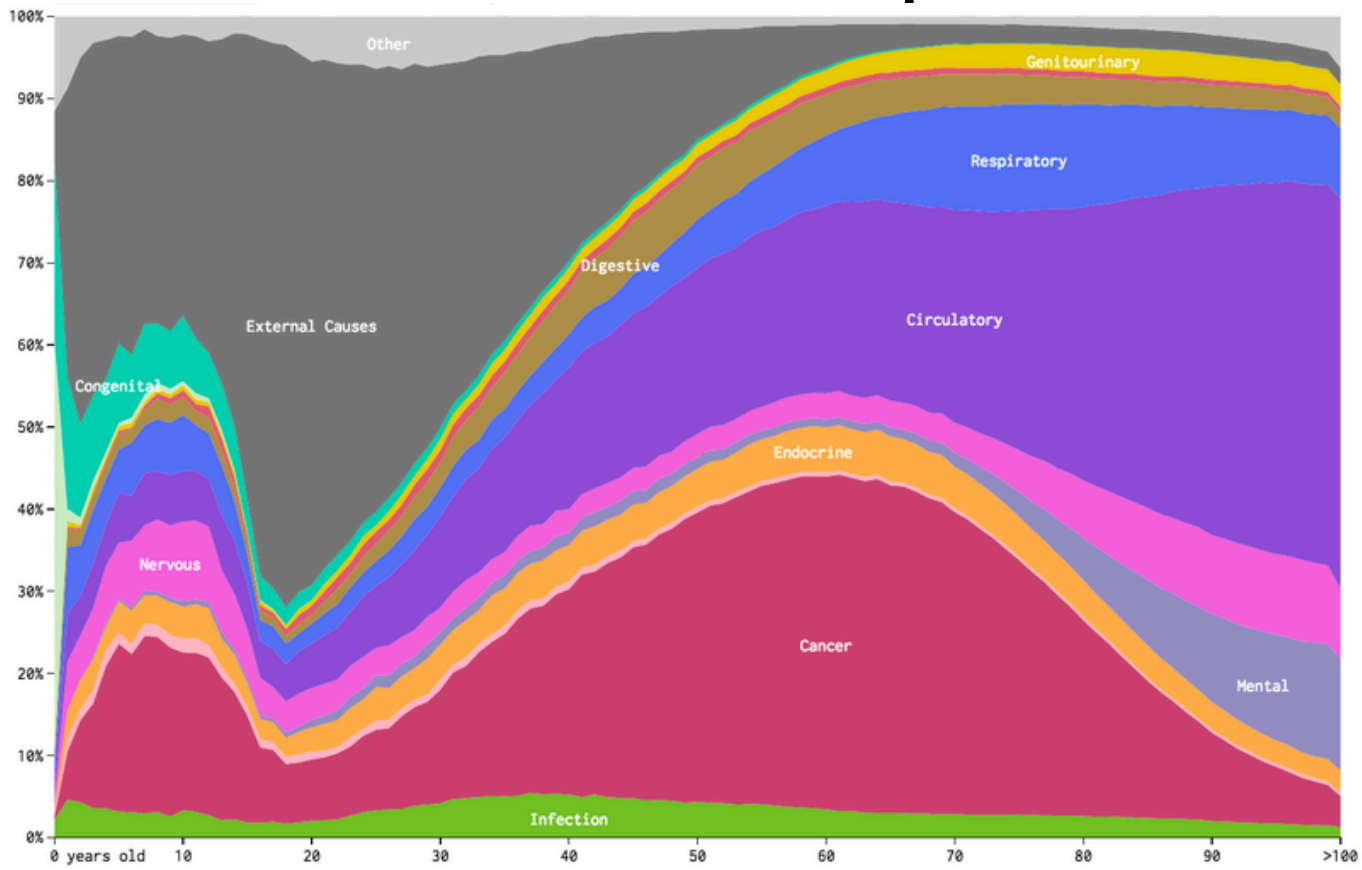
http://waitbutwhy.com/2013/12/how-to-name-baby.html

Stream Graph



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

Stream Graph



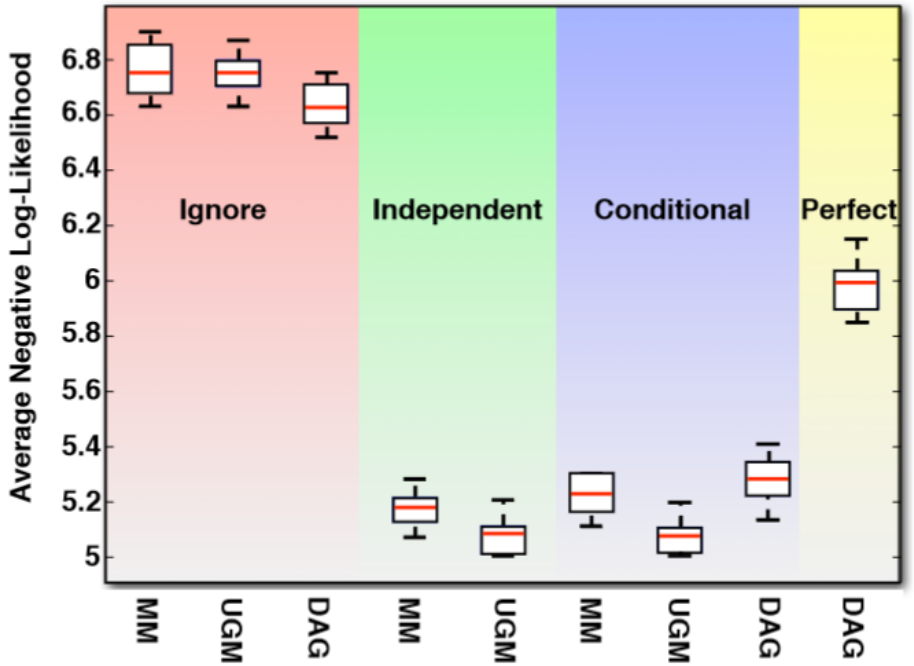
http://www.vox.com/2016/5/10/11608064/americans-cause-of-death

Box Plot

- Photo from CTV Olympic coverage in 2010:

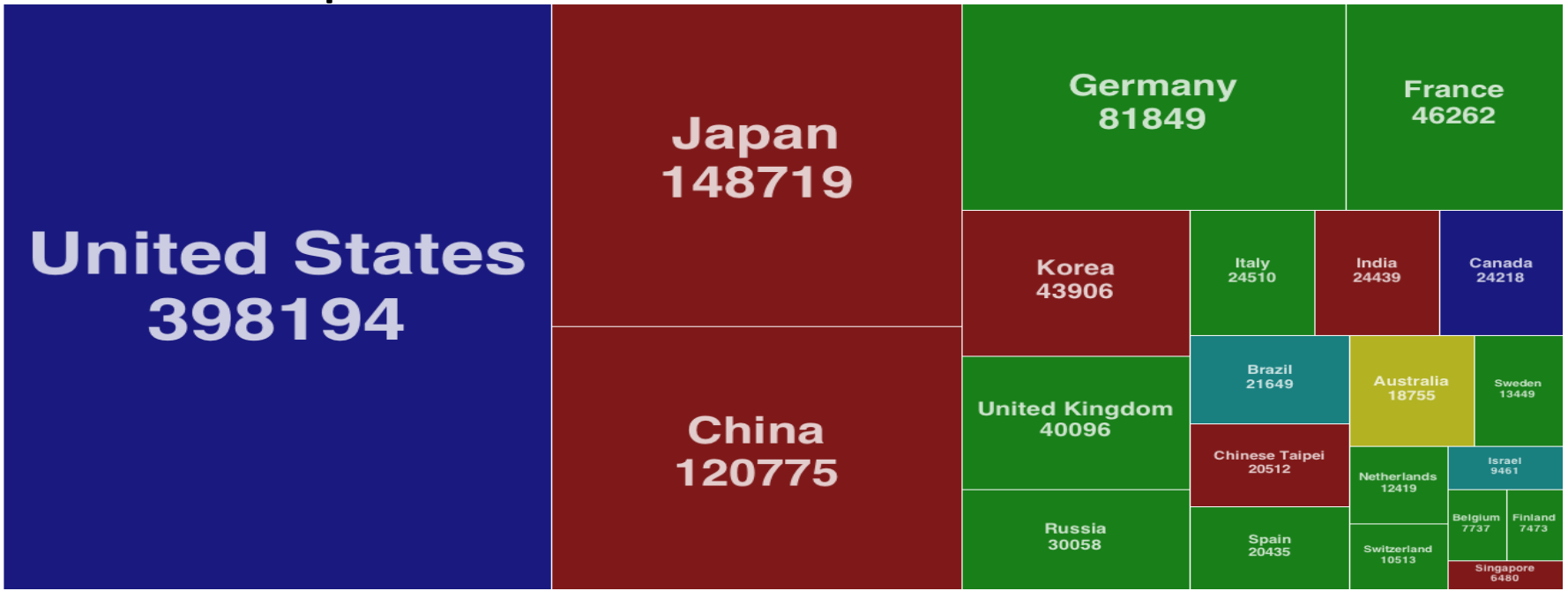
Box Plots

- Box plot with grouping:



Treemaps

- Area represents attribute value:



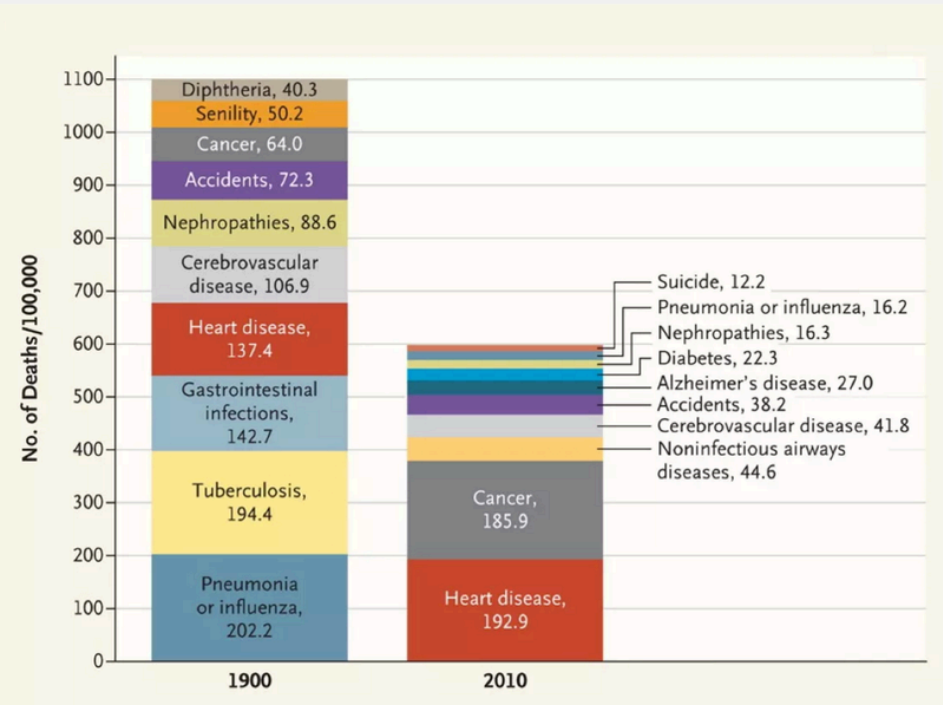
http://mjwest.org/vis/treemaps/

Cartogram

- Fancier version of treemaps:

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

- Bar chart with grouping:



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