Data Science for Smart Cities

CE88

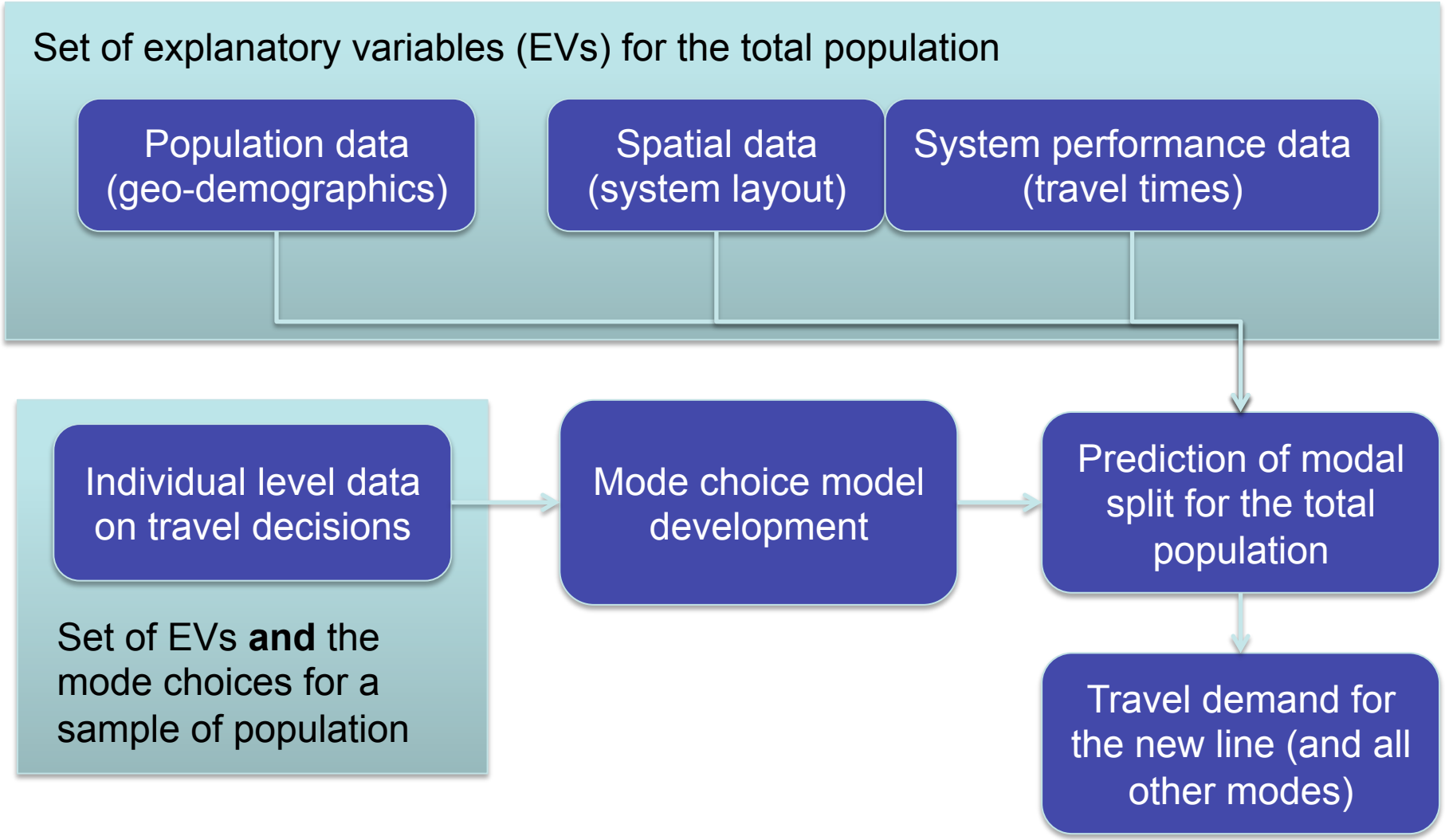
Prof: Alexei Pozdnukhov

Last time: data requirements

Describe the system in terms of **explanatory variables**

	Data source
Socio-economic characteristics of potential passengers: - total population at travel origin, car ownership - employed/occupied population, income levels, age - intended destinations of travel	Census, surveys (traditional) APIs & crowd-sensing (emerging)
Level-of-service variables - accessibility of transit, travel times - driving times (including delays due to congestion)	Routing services, online maps, APIs
System parameters and policy variables - gas prices, tolls, parking fees - transit fare - taxi/uber/lyft/your favorite TNC fare	Regional transportation agencies, APIs, local sources

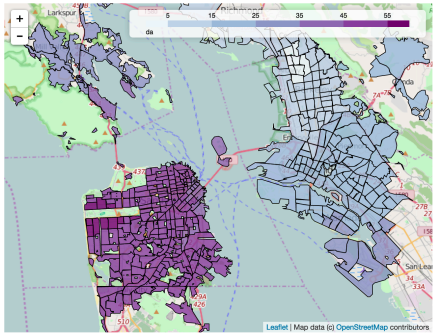
Demand forecasting: mode choice



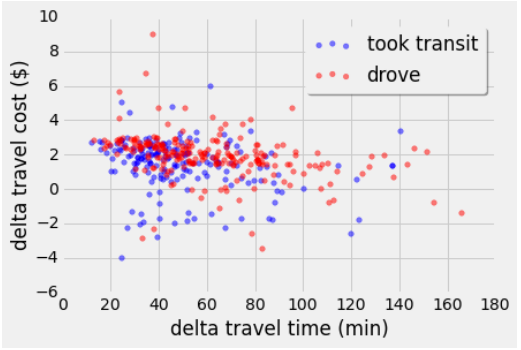
Today

Multivariate data visualization

Practice data exploration and visualization
Mini Lab



+



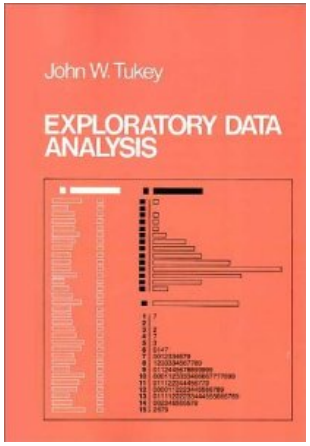
Practice knowledge discovery, hypothesis generation
Mini Lab

Exploratory Data Analysis

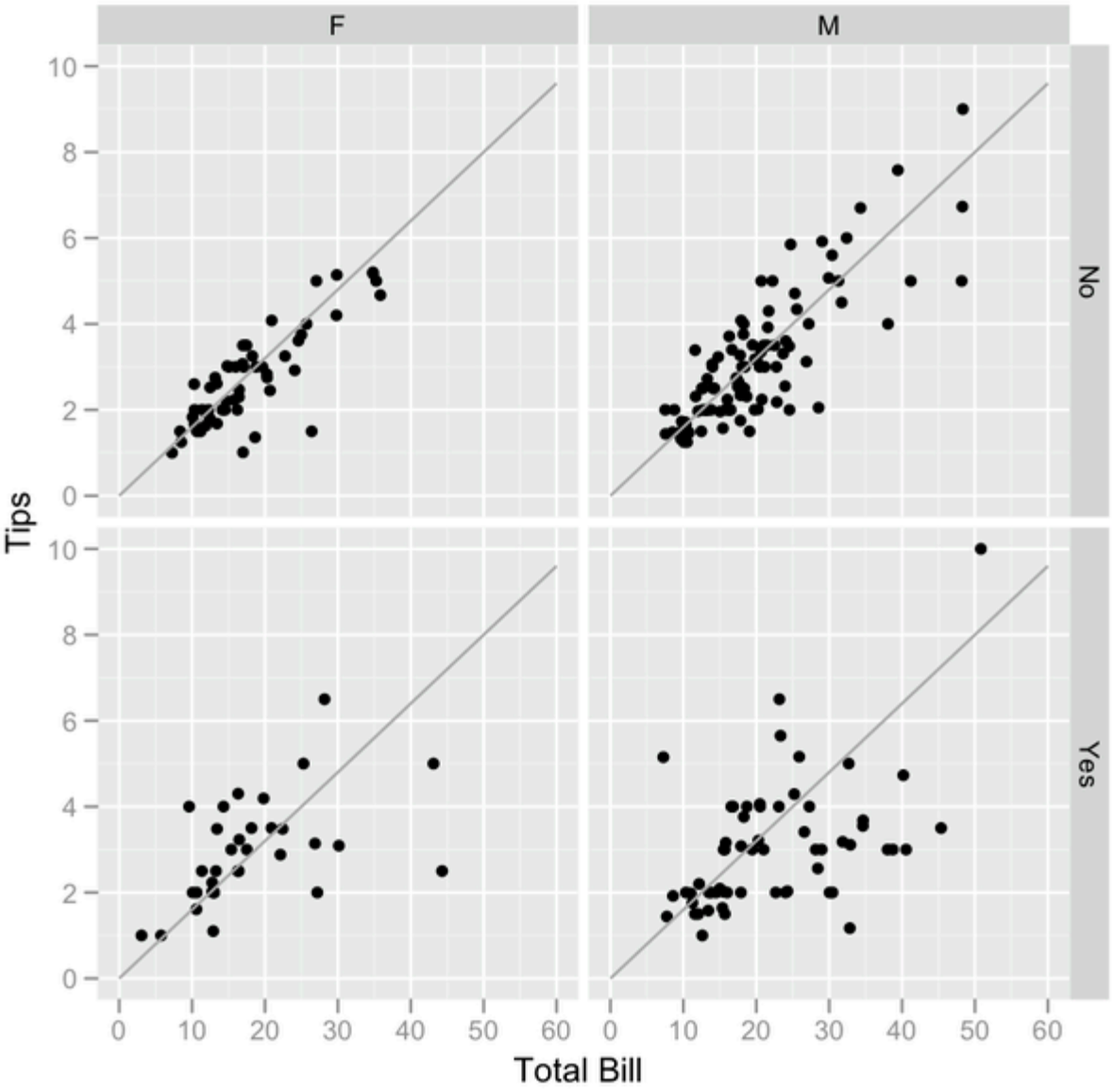
The objectives of EDA are to:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

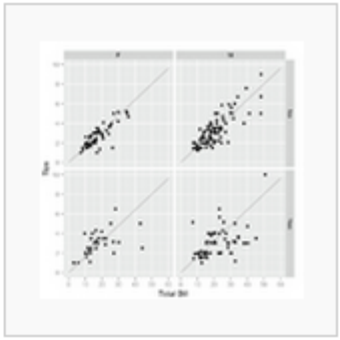
Tukey, John Wilder (**1977**). *Exploratory Data Analysis*. Addison-Wesley. ISBN 0-201-07616-0.



Exploratory Data Analysis



$$\text{tip rate} = 0.18 - 0.01 \times \text{size}$$



Scatterplot of tips vs bill separately by gender and smoking party. Smoking parties have a lot more variability in the tips that they give. Males tend to pay the (few) higher bills, and female non-smokers tend to be very consistent tippers (with the exception of three women).

Cook, D. and Swayne, D.F. (with A. Buja, D. Temple Lang, H. Hofmann, H. Wickham, M. Lawrence) (2007) "Interactive and Dynamic Graphics for Data Analysis: With R and GGobi" Springer

Visual Analytics

- Visualize: “to **make perceptible** to the mind or imagination”
  - Random House Webster’s College Dictionary
- “Visualization is the process of representing abstract business or scientific data as images that can **aid in understanding the meaning** of the data.”
  - Whatis?com computer dictionary, http://whatis.techtarget.com/whome/
- “Visualization offers a method for **seeing the unseen**.”
  - B. McCormick, T. DeFanti, and M. Brown. Definition of Visualization. ACM SIGGRAPH Computer Graphics, 21(6), November 1987, p.3
- “An estimated 50 percent of the brain's neurons are associated with vision. Visualization <…> aims to put that neurological machinery to work.”
  - Ibid.

➢ Visual analytics is called upon to **extend the perceptual and cognitive abilities** of humans, to provide them with the capability to truly understand complex information

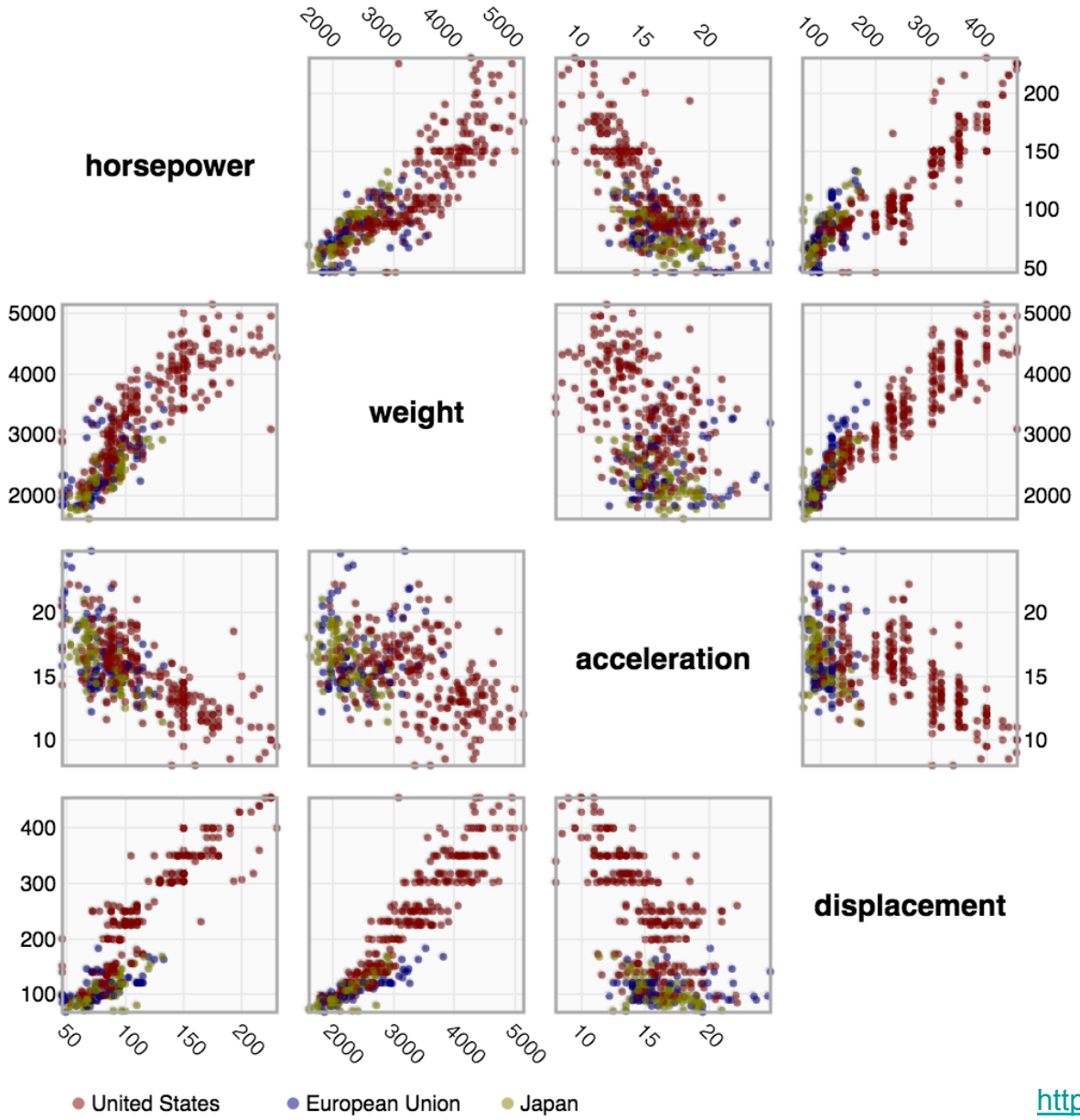
Exploratory Data Analysis

There are a number of tools that are useful for EDA, but EDA is characterized more by the attitude taken than by particular techniques.

Typical graphical techniques used in EDA are:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- **Scatter plot**
- Stem-and-leaf plot
- Parallel coordinates
- Odds ratio
- Multidimensional scaling
- Targeted projection pursuit
- Principal component analysis
- Multilinear PCA
- Projection methods such as grand tour, guided tour and manual tour

Scatter plots

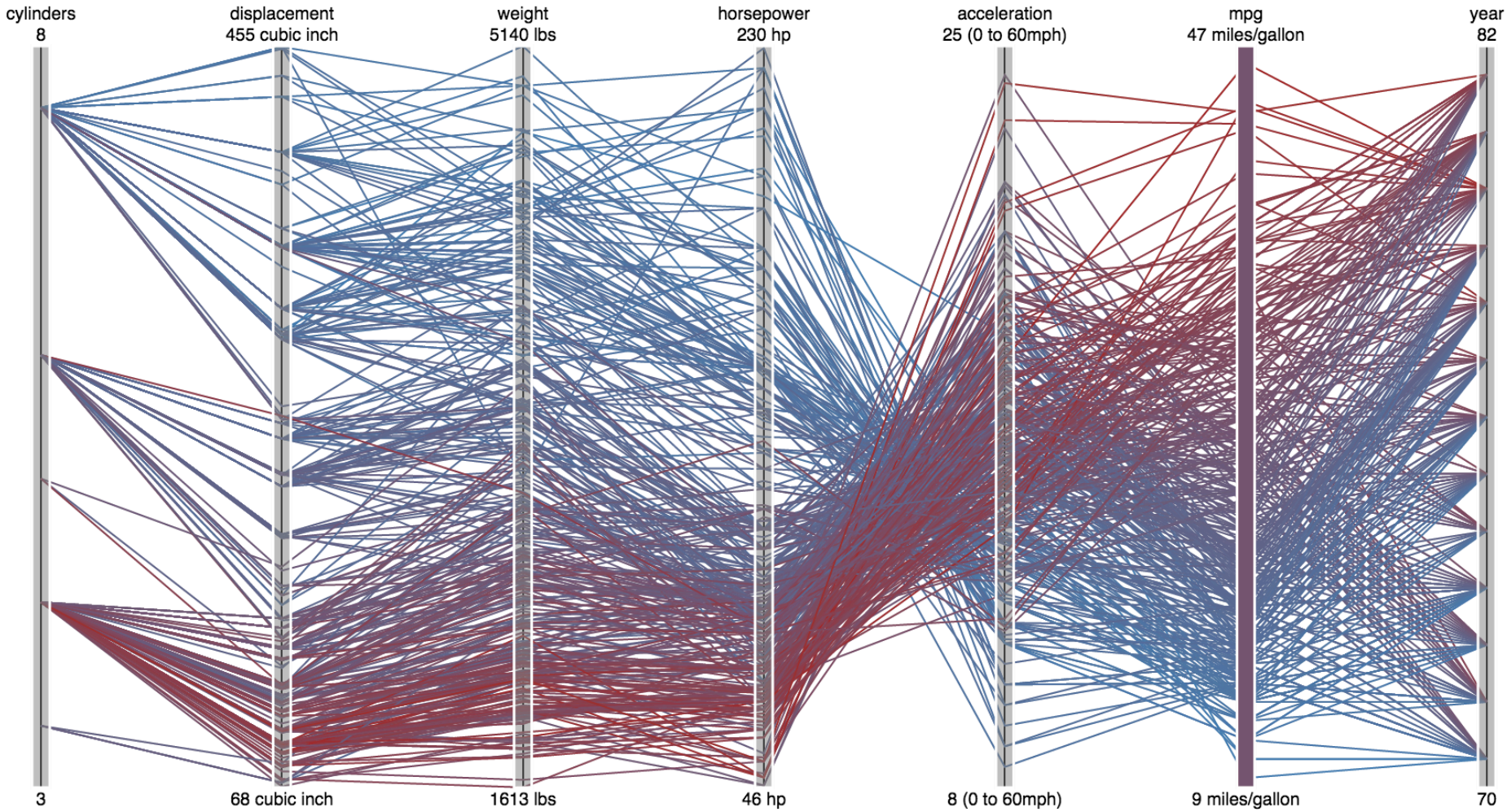


Cars performance database

http://www.ggobi.org/

http://homes.cs.washington.edu/~jheer/files/zoo/

Parallel coordinate plots

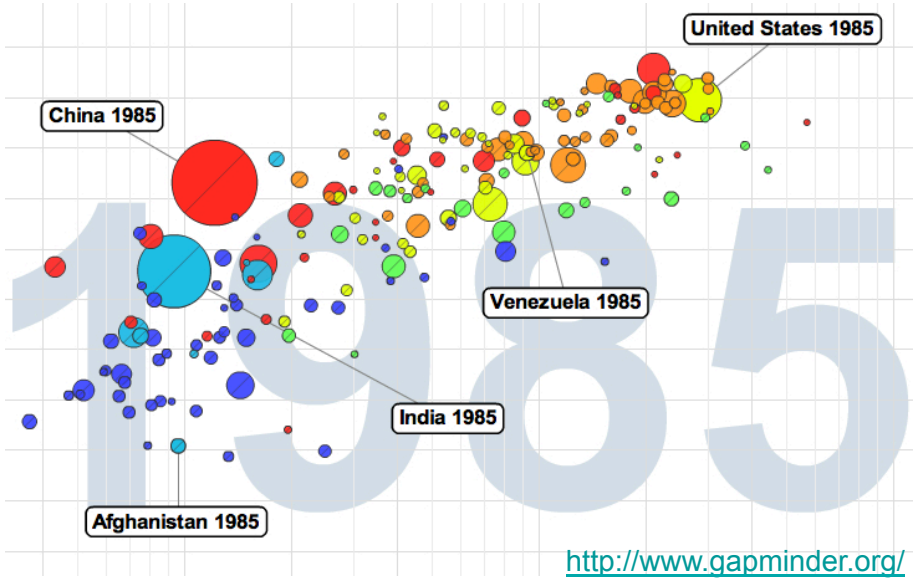


http://www.ggobi.org/

http://homes.cs.washington.edu/~jheer/files/zoo/

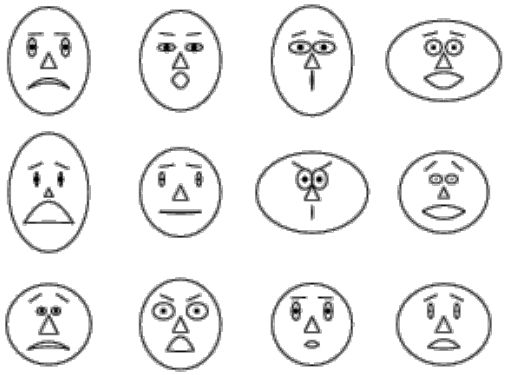
Multiple dimensions in 2D

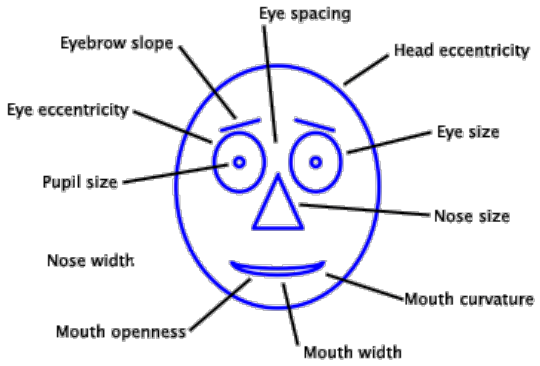
Scatter plot: x,y, color, size, transparency, shape, symbol….



http://www.gapminder.org/

such as Bar chart, Pie charts, Glyphs, Chernoff faces ☺





Exploratory Data Analysis

Homework 2: identify census tracts nearby university campuses

What is the age distribution in an ‘average’ neighborhood/tract?

What is it nearby the universities?

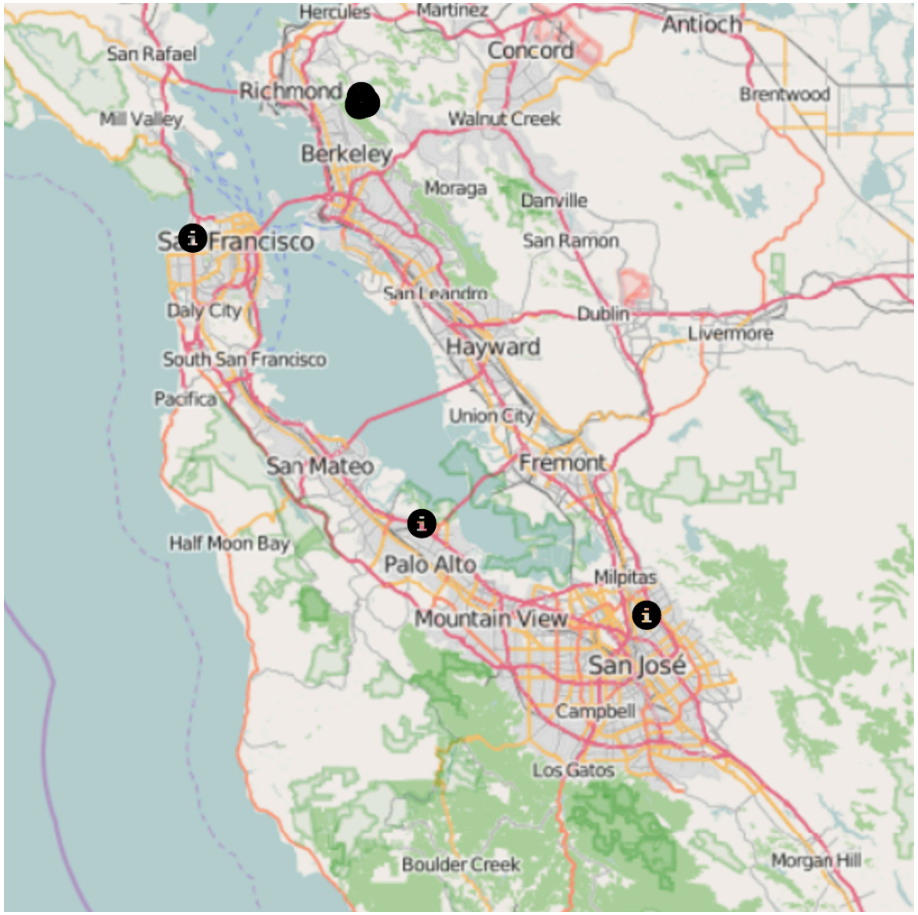
Is it different?

How is it different?

Propose a method.

Does my method makes sense?

Can I improve it? What is the threshold value? 50%? 75%?



Exploratory Data Analysis

Homework 4: forecast modal split for getting to AT&T Park stadium

What are the factors that influence mode choice?

What is the variability of factors across the region?

Can simplifying assumptions be made?

What is the most significant factor(?)

What about predictions?