STAT 1291: Data Science

Lecture 20 - Summary

Sungkyu Jung

Semester recap

- *data visualization*
- *data wrangling*
- *professional ethics*
- *statistical foundation*
- *Statistical modeling: Regression*
- *Cause and effect: Causality and confounding*
- *More statistical modeling: Machine learning*

1. Data visualization

A powerful tool in exploring, analyzing, and conveying infomation

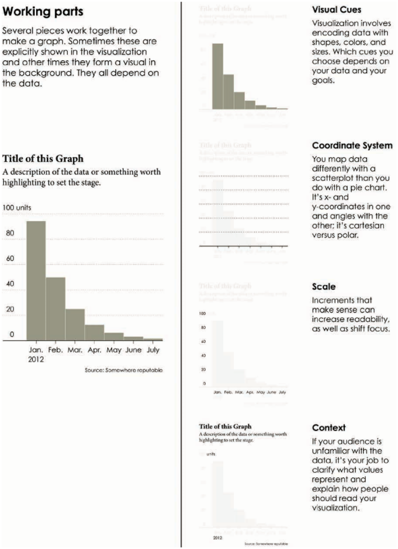
Good graphics vs bad graphics - Edward Tufte

- Maximize data-ink ratio
- Avoid chart junk
- Clear, detailed, and thorough labeling and appropriate scales

A taxonomy for data graphics

- Nathan Yau provides a systematic way of thinking about how data graphics convey specific pieces of information, and how they could be improved.
- Data graphics can be understood in terms of four basic elements:

1. Visual cues - position, length, angle, direction, shape, color
2. Coordinate system - Cartesian, polar, geographical
3. Scale - numeric (linear, logarithmic), categorical, time
4. Context - title, axis labels, references



Working parts

Several pieces work together to make a graph. Sometimes these are explicitly shown in the visualization and other times they form a visual in the background. They all depend on the data.

Visual Cues

Visualization involves encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals.

Coordinate System

You map data differently with a scatterplot than you do with a pie chart. It's x- and y-coordinates in one and angles with the other; it's cartesian versus polar.

Scale

Increments that make sense can increase readability, as well as shift focus.

Context

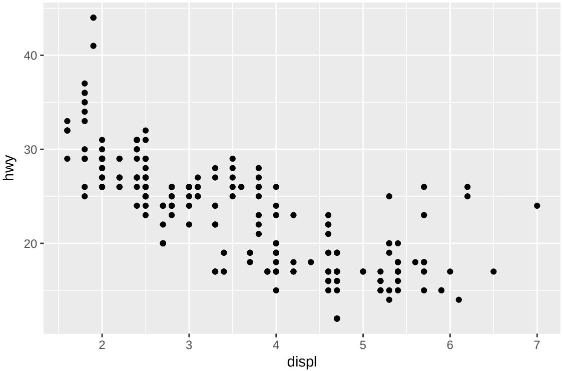
If your audience is unfamiliar with the data, it's your job to clarify what values represent and explain how people should read your visualization.

Title of this Graph
A description of the data or something worth highlighting to set the stage.
100 units
80
60
40
20
0
Jan. Feb. Mar. Apr. May June July
2012
Source: Somewhere reputable
100
80
60
40
20
0
Jan. Feb. Mar. Apr. May June July
Title of this Graph
A description of the data or something worth highlighting to set the stage.
units
2012
Source: Somewhere reputable

Figure 1:

Grammar of graphics

- Hadley Wickham, in the `ggplot2` package in R



```r
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```

In connection to the four elements of data graphics,

1. `ggplot()` (by default) sets the coordinate system as the *Cartesian coordinate system*;
2. Visual cue used is the *position*, set by `mapping = aes(x = ..., y = ...)`, paired with the use of `geom_point()`;
3. *scale* is automatically chosen as appropriate as possible;
4. *context* is (minimally) given by the axis labels.

Principles last long; technologies live short

- David Tufte, Visual display (1982)
- Package `ggplot2` (2005)
- RStudio (2011)
- Tableau (interactive data visualization product, 2003)

2. Data wrangling

> *"Tidy datasets are all alike, but every messy dataset is messy in its own way." – Hadley Wickham*

What makes a dataset tidy?

- Each variable must have its own column.
- Each observation must have its own row.
- Each value must have its own cell.

Tidying data set

We used `tidyr` package to make untidy data tidy.

- `gather()`
- `spread()`

Transforming data set

- following a grammar of data manipulation
- `dplyr` provides a small set of “verbs” that can be combined by `%>%` (pipes) to perform complex tasks
- Single table vers
- Two table verbs

Single Table Verbs

`dplyr` provides a suite of verbs for data manipulation:

- `filter()`: select rows (observations) in a data frame;
- `arrange()`: reorder rows in a data frame;
- `select()`: select columns (variables) in a data frame;
- `mutate()`: add new columns to a data frame;
- `summarise()`: collapses a data frame to a single row;

Two table operations

- **Mutating joins**, which add new variables to one table from matching rows in another.
- `inner_join()`, `left_join()`, `right_join()`
- **Filtering joins**, which filter observations from one table based on whether or not they match an observation in the other table.
- **Set operations**, which combine the observations in the data sets as if they were set elements.

What we did not discuss

- A full-on relational database management (SQL; MDSR Chapters 12,13)
- Cutting-edge database management beyond SQL
  - Distributed storage and processing of dataset of big data, e.g.
  - Hadoop using MapReduce programming (`http://hadoop.apache.org/`)
  - Spark (`http://spark.apache.org/`)
- R data intake: API, .json, etc (MDSR Section 5.5)
- R programming: `apply()` family, `for` and `while` (MDSR Section 5.4)

3. Professional ethics

Some principles to guide ethical action

1. Common sense: lying, cheating, and stealing are unethical
2. Do not take advantage of your professional skills
3. Draw on generally recognized professional standards
4. Be open and honest
5. Have a professional responsibility to particular stakeholders

Reproducible analysis

- Scriptable statistical computing (e.g. R)
- Separating data and analysis
- Repeatable analysis for different data sets
- Literate programming

  "Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do" –Donald Knuth

- The `rmarkdown` and `knitr` packages: Analysis in your report

4. Statistical foundation

- Statistical methods
  - Quantify patterns and **their strength**
  - Find patterns that are too complex to be seen visually
  - Interpreting data
  - involves modeling

Data Science workflow

A typical data science project looks something like this:

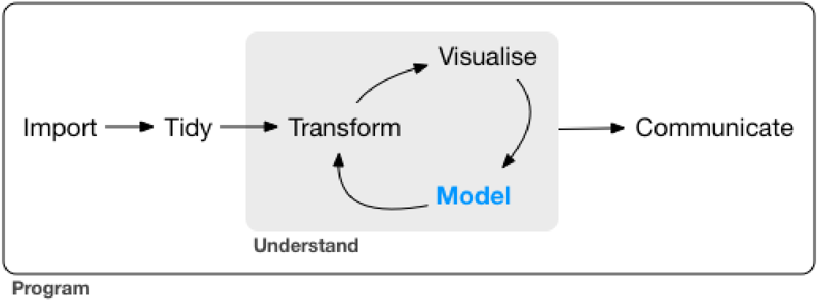


Figure 2: (from r4ds)

Uncertainty quantification

- Samples and Populations

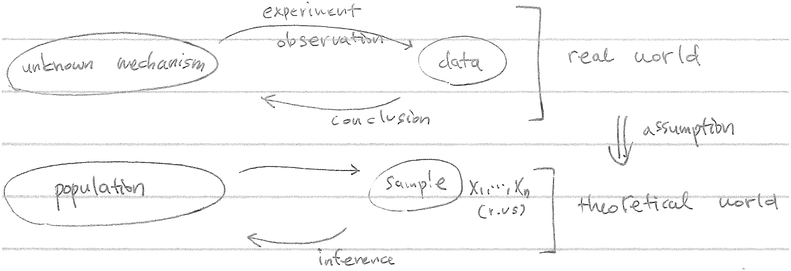


Figure 3: (from my lecture notes for mathematical statistics)

- How reliable a statistic computed from the data?
  - Confidence interval: Statistic ± 2 Standard error
  - Use Bootstrap to estimate "Standard error" from the data

5. Statistical modeling: Regression

- Information in data is identified by conditional modeling
- Regression is a fundametal tool in modeling

6. Causality and confounding

- correlation does not imply causation

- Because of potentially lurking counfounding variables
- To infer causal relations, do randomized trials (as opposed to observational studies)
- Almost all examples of Big data are observational
- Controling confounding variables: conditional modeling (stratification / multiple regression), matching.

7. Machine learning

Supervised learning vs unsupervised learning

- *Prediction* and *classification* are examples of *supervised learning*
- *clustering* and *dimension reduction* is an example of *unsupervise learning*

Model evaluation

- **Resampling** is a key in data-driven model evaluation.
- **Cross-validation** is widely used for prediction and classification.

Outro

- ____________ of data science change slowly;
- ____________ of data science change rapidly

Outro

- Theories of data science change slowly;
- Tools of data science change rapidly

Theories of data science change slowly

- Least-squares (Gauss, 1795)
- Regression (Galton, 19th century)
- Generalized linear models (Nelder, 1972)
- Lasso + Elastic net regression (Tibshirani, Zou, 1996–2005)
- Perceptron–Deep Neural Netwook (1940–2010)

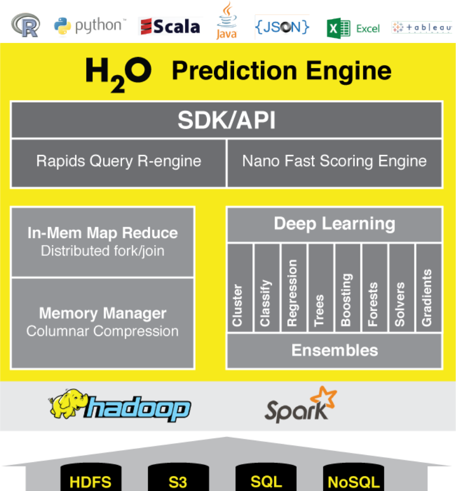
Tools of data science change rapidly

- C (1972)
- SQL (1979)
- Python (1989)
- R (2000)
- RStudio (2011)

Professional data scientists

Will you *use* tools? or *design* tools?

H2O example



H2O.ai example

- Theories of data science change slowly
  - Generalized linear models (Nelder, 1972)
  - Elastic net (Zou and Tibshirani, 2005)
- Tools of data science change rapidly
  - Fast computing by GPU (recent)

Visit `https://www.h2o.ai/gpu/`

(or `https://youtu.be/KRAMtvwlgmM`)

Advertisement

STAT 1361 (formerly STAT 1291)

“Statistical Learning and data science”

by Lucas Mentch

Thank you!