# Seminar 1

# Introduction to Data Science

Mikhail Kamrotov

Data Analysis in R

# Grades

- 50% - home assignments, 50% - group project
- 96-100% - 10, 90-95% - 9, 80-89% - 8, 75-79% - 7, 65-74% - 6, 55-64% - 5, 45-54% - 4, 35-44% - 3, 25-34% - 2, 0-24% - 1
- You can work in pairs
- Best solutions could be presented in class (5 minute talk) to get some extra points

# Definition

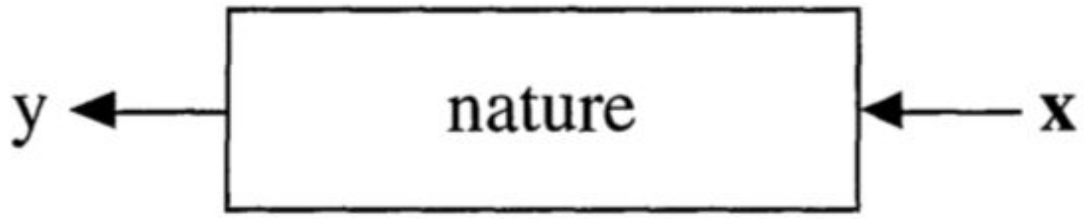
- Data analysis is **the process of transforming raw data into usable information**, often presented in the form of a published analytical article, in order to add value to the statistical output. (OECD)
- Data analysis is **a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information**, informing conclusions, and supporting decision-making (Wikipedia)
- Both miss one important step – collecting data.
- Most theories are about modeling, but 80% of the time a data scientist spends on data collection and cleansing

# Data analysis techniques

- Data mining
  - automatic discovery of useful information in large data repositories
- Descriptive statistics
  - summarizing features of data
- Exploratory data analysis
  - finding new features in data
- Confirmatory data analysis
  - hypotheses testing
- Predictive analytics
  - deriving predictions from data
- Text analytics
  - extracting information from textual (i.e. unstructured) data

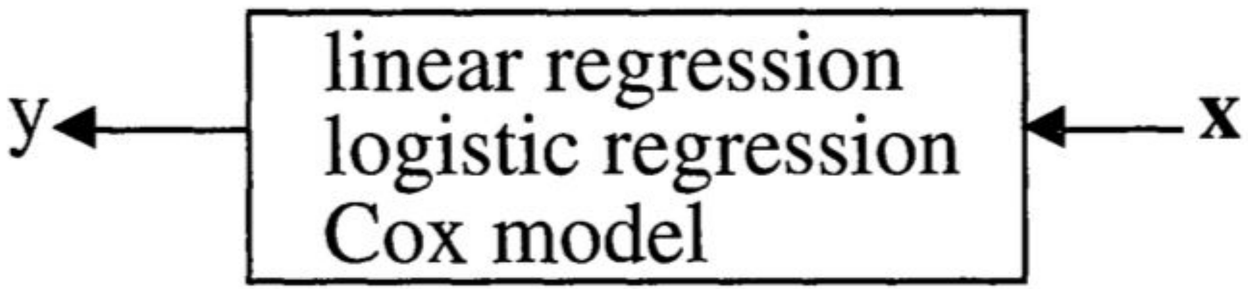
# Two cultures of data analysis

- Data is generated by a black box
- Input variables **x** (independent variables) go in one side (time you spend on your home assignments)
- On the other side the response variables **y** come out (your grades)
- Two main goals: prediction and information
- Two approaches: data modeling culture and algorithmic modeling culture



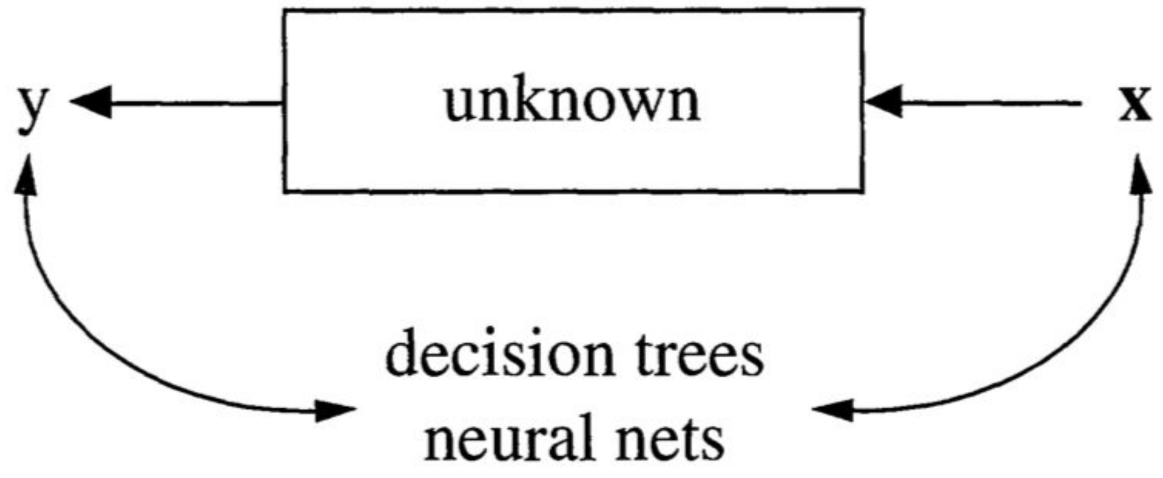
# Data modeling culture

- Starts with assuming a data model for the inside of the black box
- The values of the parameters are estimated from the data and the model then used for information and/or prediction
- Model validation: goodness-of-fit tests



# Algorithmic modeling culture

- Considers the inside of the box complex and unknown
- Tries to find a function **f(x)** - an algorithm that operates on **x** to predict the responses **y**
- Model validation: predictive accuracy



# Why do you need to learn data analysis

- Valuable skill that is highly remunerative
- Things sometimes are not as obvious as they seem at first sight
- Ability to verify results produced by your colleagues
- The only way to make scientific contribution and verify theories, especially in social sciences

# Data manipulation by Tim Cook

- https://www.statschat.org.nz/2013/09/11/cumulative-totals-tend-to-increase/

# Even academic superstars may be wrong

- http://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646

# A lot of fraud in science (especially in social sciences)

- https://www.financial-math.org/blog/2015/10/is-research-in-finance-and-economics-reproducible/

# Random chance plays a huge role in social sciences

- http://www.tylervigen.com/spurious-correlations

# Intuition might be wrong

Simpson's paradox: graduate admissions to UCB

| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| Total | 8442 | 44% | 4321 | 35% |

# Intuition might be wrong

Simpson’s paradox: graduate admissions to UCB

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| A | 825 | 62% | 108 | **82%** |
| B | 560 | 63% | 25 | **68%** |
| C | 325 | **37%** | 593 | 34% |
| D | 417 | 33% | 375 | **35%** |
| E | 191 | **28%** | 393 | 24% |
| F | 373 | 6% | 341 | **7%** |

# Intuition might be wrong, part 2

- Monty Hall problem
- https://en.wikipedia.org/wiki/Monty_Hall_problem
- Humans vs birds: birds win (Herbranson, 2010)

# R

- R is a language of statistical computing
- Modern social sciences speak mostly this language (and Python as well)
- R download link:  https://cran.r-project.org
- RStudio download: https://www.rstudio.com/products/rstudio/download/#download

# P.S.

Calling Bullshit is a highly recommended online course at the University of Washington http://callingbullshit.org/syllabus.html#Introduction