Main Tools of a Data Scientist

Vladimir Garcia

January 2018

- Measurement
  - How "insights" or "policies" are constructed
- Statistical Programming Languages
  - 3 main statistical programming languages: R, Python and Julia
  - Different advantages and disadvantages
    - R: large user base, slow, no for general-purpose computing
    - Python: large user base, ubiquity, slow
    - Julia: small user base, fast, new program
- Web Scraping
  - Data is being constantly collected in publicly accessible places
  - How web scraping works?
    - APIS: employed by Twitter, Facebook, etc. impose limits on data you can download
    - Parsing: downloading html files and parsing their text to extract data. Websites monitor IP addresses of all website viewers
- Handling large data sets
  - Data sets that are too big to fit on a single hard drive
  - you may be able to split the files up into manageable chunks, though no longterm solution
  - How to solve this issue?
    - RDDs (Resilient Distributed Datasets)
      - chops your huge data set into manageable chunks and executes actions on those chunks in parallel
      - withstand any disruption in the computing cluster
    - SQL
      - transform data into a more usable form for statistical software to use

        - subset, merge, and perform other common data transformations

- Visualization
  - ggplot2 (R)
  - matplotlib (Python)
  - plots.jl (Julia)
  - Tableu
- Modeling
  - main objectives of statistical modeling are as follows:
    1. to test theories
    2. predict behavior
    3. explain behavior