USP 410/510: Urban Informatics

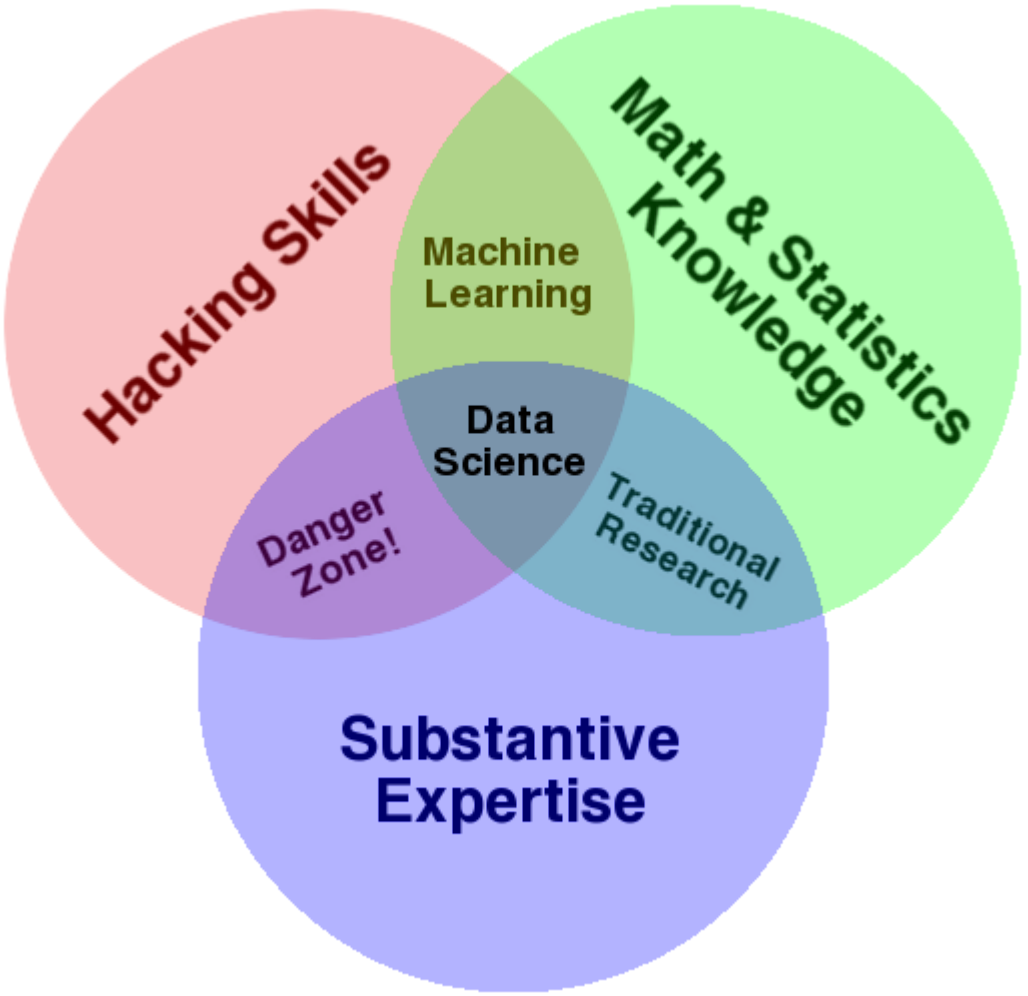
Liming Wang

January 23, 2023

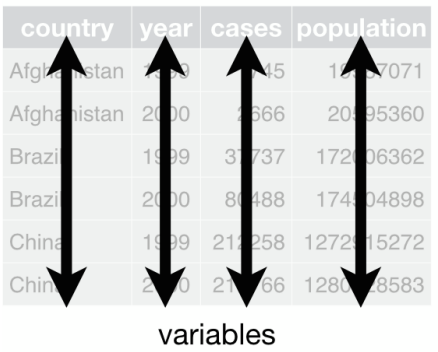
Outline

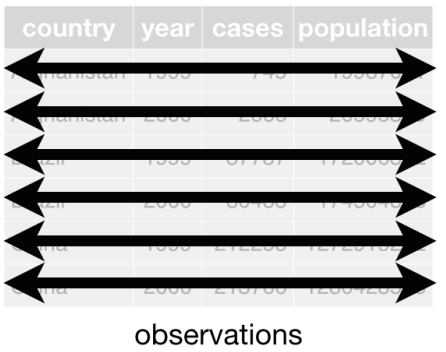
- Python vs R
- tidy data
- workflow

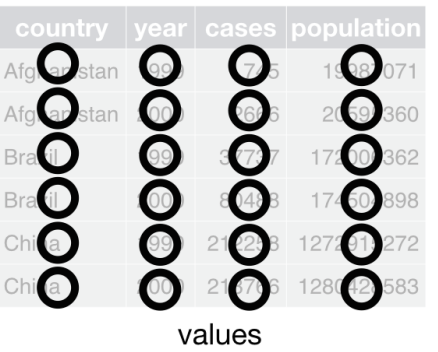
Data Scienice



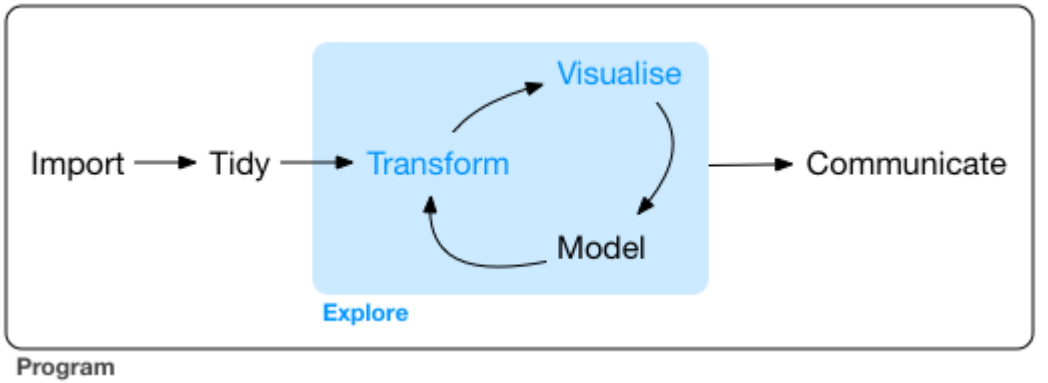
tidy data


country
year
cases
population
variables


country
year
cases
population
observations


country
year
cases
population
values

tidy workflow



import

- Packages/functions:
  - `readr`: comma-separated values (read_csv), tab delimited file (read_tsv), fixed width file (read_fwf)
  - `readxl`: Excel files (read_excel)
  - `haven`: data format from other statistics packages - SAS, SPSS, Stata, etc
  - `foreign`: read dbf (shapefile database)
- More formats/sources: https://www.datacamp.com/community/tutorials/r-data-import-tutorial
- Demo

import: data.frame vs tibble

- base R packages/functions (foreign) import data as data.frame
- tidyverse packages (readr, readxl, haven) import data as tibble
- subtle difference between the two, use `tibble` whenever possible
  - tibble print outputs are nicer to human
  - many old packages only take data.frame
- How to tell which is which
  - `class(variable_name)`
- convert between the two:
  - tibble::as_tibble
  - as.data.frame

tidy (reshape)

- rename columns of imported data (commonly a data.frame or tibble)
  - janitor::clean_names()
  - manual rename with dplyr::rename()
- tidy data with the `tidyr` package (similar packages: reshape2)
  - pivot_longer, pivot_wider
  - https://github.com/gadenbuie/tidyexplain#pivot-wider-and-longer

transform with dplyr

`dplyr` is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges:

- `mutate()` adds new variables that are functions of existing variables
- `select()` picks variables based on their names.
- `filter()` picks cases based on their values.
- `summarise()` reduces multiple values down to a single summary.
- `arrange()` changes the ordering of the rows.
- `*_join` joins two data frames