Data Scientist Toolkit Notes

# Types of Questions

* In approximate order of difficulty
  + Descriptive
  + Exploratory
  + Inferential
  + Predictive
  + Causal
  + Mechanistic

## Descriptive Analysis

Goal: Describe a set of data

* First kind of analysis performed
* Commonly applied to census data
* Description and interpretation are different steps
* Cannot generalize
* Examples: US census, Google book word pairing trend

## Exploratory Analysis

Goal: Find relationships you didn’t know about

* Good for discovering new connections
* Useful for defining future studies
* Usually not the final say
* Should not be used for generalizing
* **Correlations does not imply causation**
* Examples: brain stimuli-response study, John Hopkin’s sky survey

## Inferential Analysis

Goal: Use a relatively small sample of data to say something about bigger population

* Commonly the goal of statistical models
* Involves estimating both quantity you care about and your uncertainty about your estimate
* Depends heavily on both the population and sample scheme
* Examples: effect of air pollution control on life expectancy in the US, an analysis of 545 counties

## Predictive Analysis

Goal: Use data on some objects to predict values for another object

* If X predicts Y does not mean that X causes Y
* Accurate prediction depends heavily on measuring right variables
* Although there are better and worse prediction models, more data and a simple model works really well (see link)
* “Prediction is very hard, especially about the future” -various
* Examples: Five thirty eight forecast (predict outcome of US election); Target figured out teen girl was pregnant based on her purchases

## Causal Analysis

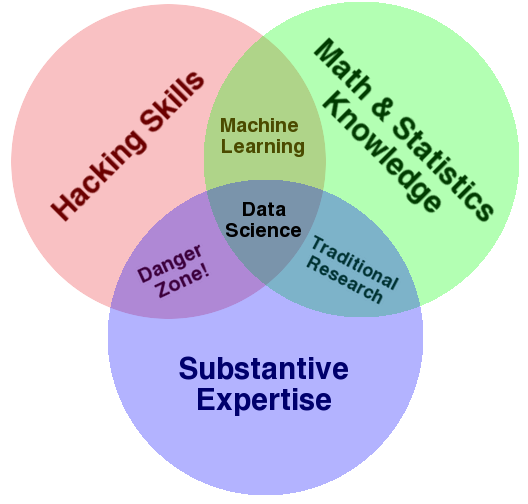
Goal: Find out what happens to one variable when you make another variable change

* Usually randomized studies are required to identify causation
* There are approaches to inferring causation in non-randomized studies but they are complicated and sensitive to assumptions
* Usually identified as average effects, but may not apply to every individual
* Usually the “gold standard” for data anlaysis
* Examples: Duodenal infusion of donor feces for recurrent C.Difficile

## Mechanistic Analysis

Goal: Understand the exact changes in variables that lead to exact changes to other variables for individual objects

* Incredibly hard to infer, except in simple situations
* Modelled by deterministic set of equations (physical/engineering science)
* Generally the random component of data is measurement error
* If equations are known but parameters are not, these may be inferred with data anlaysis
* Example: Empirical analysis of changes in pavement design on performance



Data science venn diagram

# What is Data?

Definition of data:

“Data are values of qualitative or quantitative variables, belonging to a **set of items”** -wikipedia

* Variables: A measurement or characteristic of an item.
* Qualitative: country of origin, sex, treatment
* Quantitative: height, weight, blood pressure (on a continuous scale)

## What does data look like?

* Text output of DNA sequencing machine
* Website API (e.g. twitter)
* Textual health record
* Video
* Audio
* Open government website data in any number of formats (Excel, CSV, raw)
* Rarely like nice Excel file in beginning

## Most Important Thing

* Data is second most important thing
* Most important is question
* Often data will limit or enable the question
* **But having data can’t save you if you don’t have a question**

# What about Big Data?

* Estimated 1.8 zettabytes of data in 2011
* But only tiny fraction can be used to answer questions
* Depends on perspective - now vs 50 years ago
* Why big data now? - see experimental study of small world problem using just 246 people
* Investigators took 30 billion conversations among 240 million people to determine how much separation found it is 6.6 degrees of separation
* Big data - you need the right data
* “the data may not contain the answer. The combination of some data and an aching desire for anshwer does not ensure that a reasonable answer can be extracted from a given body of data” -tukey. “No matter how big the data are” --leek

# Experimental Design

## Why You Should Care and What You Should Do

* People who developed predictive model that was wrong, then were sued
* Need to know and care about analysis plan
* Have a plan for data and code sharing - github, figshare. See leek’s for plan
* Formulate question in advance:
  + Question: Does changing the text on your website improve donations?
  + Experiment: 1. randomly show visitors one version or the other, 2. measure how much they donate, 3. determine which is better (Obama)

## Statistical inference

* Select subset of population
* Perform descriptive statistics on this sample
* Use inferential statistics to decide if small sample plays out in large sample
* Best case to perform - small variability within each type, big difference in two types

## Confounding

* Example: Shoe size is correlated with literacy, but actually age causes both
* Correlation is not causation (nobel laureates and chocolate consumption)
* Caused spurious correlation

## Randomization and blocking

* If you can (and want to) fix a variable
  + Website always says Obama 2014 n it
* If you don’t fix a variable, stratify it
  + If you are testing sign up phrases and have two website colors, use both phrases equally on both
* If you can’t fix a variable randomize it
  + use a computer program to randomly assign people to groups
  + Without randomization, cannot determine if difference in outcome is caused by difference in confounding variable or different in treatment

## Prediction

* Collect individuals from sample and put them into training set
* Build prediction function for new individual to predict if they will responded or not

## Prediction versus inference

* For prediction you need the two groups to be a lot more separated
* Pay attention to relative sizative effects

## Prediction key quantities

* Sensitivity -> Pr (positive test | disease)
* Specificity -> Pr (negative test | no disease)
* Positive predictive value -> Pr (disease | positive test)
* Negative predictive value -> Pr (no disease | negative test)
* Accuracy -> Pr (correct outcome)

## Beware of data dredging

* Keep going until you find result that you like - ignores fact that you tried a lot of different things first

## Summary

* Good experiments:
  + Have replication
  + Measure variability
  + Generalize the problem you care about
  + Are transparent
* Prediction is not inference
  + Both can be important
* Beware data dredging

# Commands

## Command Line

|  |  |
| --- | --- |
| Command (same as linux command line) | Description |
| pwd | prints working directory |
| clear | clears screen |
| ls | lists all files in directory |
| cd \c\here | change directory to \c\here |
| cd .. | changes directory to parent |
| cd | changes directory to root directory |
| rm file | removes file |
| rm -r Directory | removes directory and all contained folders and files |
| touch new\_file | create new file |
| mv file Directory | moves file to directory |
| mv file renamed\_file | renames file |
| echo | prints arguments |

## Git Config

|  |  |
| --- | --- |
| Command | Description |
| git config --list | Shows all set configurations |
| git config --global user.name “Your Name Here” | Set user name |
| git config --global user.email “you@server.com” | Set email |
| exit | exits git bash |

## Git Repository Basics

|  |  |
| --- | --- |
| Command | Description |
| git init | initialize repository |
| git remote add origin <https://github.com/kitjosh1050/test-repo.git> | link local copy with remote copy at git hub |
| git clone <https://github.com/kitjosh1050/test-repo.git> | initializes and clones remote repository within current working directory |
| git pull  git pull origin master | ??get new changes from remote  gets new files from remote |
| git push | add committed changes to remote |
| git add . | adds all new files to be tracked by Git |
| git add -u | updates tracking for files that changed names or were deleted |
| git add -A | adds all new files AND updates tracking for files that changed names or were deleted |
| git reset | undo git add |
| git log | See log of git commits |
| git commit -m “Message” | commits changes to be saved with message providing description of what was done |
| git checkout -b branchname | create a branch to work on |
| git branch | show branches with current branch highlighted green |
| git checkout master | switch back to master branch type |
| git ls-files | lists tracked files |
| git ls-files --others --exclude-standard | lists untracked files |
| git ls-files --others -i --exclude-standard | lists ignored files |
| git remote show origin | Show origin details |
| when in log - type ‘q’+Enter to quit |  |
|  |  |

## Markdown

|  |  |
| --- | --- |
| Command | Description |
| # this is title |  |
| ##this is secondary heading |  |
| ###this is tertiary heading |  |
| \* First item in list  \* second item in list  \* this item in list |  |

## Obtaining R Packages

|  |  |
| --- | --- |
| Command | Description |
| help(function) | get help about a function |
| a <- available.packages() | gets list of available packages |
| head(rownames(a),3) | shows first three |
| install.packages(“devtools”) | install package |
| install.packages(c(“devtools”, “solidify”) | install multiple packages |
| library(devtools) | load library |