$\pagebreak$

## CLI (Command Line Interface)

* / = root directory
* ~ = home directory
* pwd = print working directory (current directory)
* clear = clear screen
* ls = list stuff
  + -a = see all (hidden)
  + -l = details
* cd = change directory
* mkdir = make directory
* touch = creates an empty file
* cp = copy
  + cp <file> <directory> = copy a file to a directory
  + cp -r <directory> <newDirectory> = copy all documents from directory to new Directory \* -r = recursive
* rm = remove
  + -r = remove entire directories (no undo)
* mv = move
  + move <file> <directory> = move file to directory
  + move <fileName> <newName> = rename file
* echo = print arguments you give/variables
* date = print current date

## GitHub

* **Workflow**
  1. make edits in workspace
  2. update index/add files
  3. commit to local repo
  4. push to remote repository
* git add . = add all new files to be tracked
* git add -u = updates tracking for files that are renamed or deleted
* git add -A = both of the above
  + ***Note****: add is performed before committing*
* git commit -m "message" = commit the changes you want to be saved to the local copy
* git checkout -b branchname = create new branch
* git branch = tells you what branch you are on
* git checkout master = move back to the master branch
* git pull = merge you changes into other branch/repo (pull request, sent to owner of the repo)
* git push = commit local changes to remote (GitHub)

## Markdown

* ## = signifies secondary heading (bold big font)
* ### = signifies tertiary heading (slightly smaller font than secondary, not bold)
* \* = bullet list item

## R Packages

* Primary location for R packages CRAN
* available.packages() = all packages available
* head(rownames(a),3) = returns first three names of a
* install.packages("nameOfPackage") = install single package
* install.packages(c("nameOfPackage", "nameOfPackage", "nameOfPackage") = install multiple package
* Bioconductor Packages:
  + source("https://bioconductor.org/biocLite.R")
  + biocLite() = install bioconductor packages
* library(packagename) = load package
* search() = see all functions in package after loading

## Types of Data Science Questions

* in order of difficulty: ***Descriptive*** ***Exploratory*** ***Inferential*** ***Predictive*** ***Causal*** ***Mechanistic***
* **Descriptive analysis** = describe set of data, interpret what you see (census, Google Ngram)
* **Exploratory analysis** = discovering connections (correlation does not = causation)
* **Inferential analysis** = use data conclusions from smaller population for the broader group
* **Predictive analysis** = use data on one object to predict values for another (if X predicts Y, does not = X cause Y)
* **Causal analysis** = how does changing one variable affect another, using randomized studies, Strong assumptions, golden standard for statistical analysis
* **Mechanistic analysis** = understand exact changes in variables in other variables, modeled by empirical equations (engineering/physics

## Data

* **Data** = values of qualitative or quantitative variables, belonging to a set of items (usually population)
* **Variables** = measurement/characteristic of an item (qualitative vs quantitative)
* **Data** = not always structured, usually raw file, different formats
* Most important thing is question, then it is data
* **Big data** = now possible to collect data cheap, but not necessarily all useful (need the right data)

## Experimental Design

* Formulate you question in advance
* **Statistical inference** = select subset, run experiment, calculate descriptive statistics, use inferential statistics to determine if results can be applied broadly
* ***[Inference]*** **Variability** = lower variability + clearer differences = decision
* ***[Inference]*** **Confounding** = underlying variable might be causing the correlation (sometimes called Spurious correlation)
  + dealing with confounding: fix variables, stratify (all options), randomize
* ***[Prediction]*** collection observations for different variable values, build predictive functions
  + similar problems of probability/sampling and confounding variables
* ***[Prediction]*** Difficult to understand where observation is from from different distributions. (size of effects important)
* ***[Prediction]*** Positive/negative statuses: True positive, false positive, false negative, true negative
  + **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)
* **Data dredging** = use data to fit hypothesis
* **Good experiments** = have replication, measure variability, generalize problem, transparent
* Prediction is not inference, and be ware of data dredging