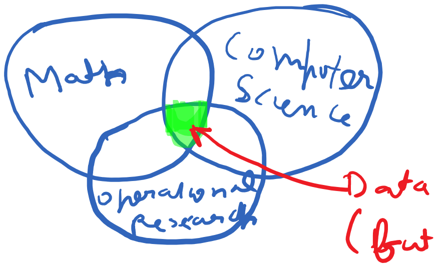
Tuesday, December 26, 2017 11:19 PM

We talked about the following things in the class today.

Data Science Venn Diagram



Data Science
(but this is just one perspective)

Topics for data sets of interest

- financial data (like stock market)
- ??? (Sangit please remind us)
- Astronomy data (lets say maps for now)
- FDA data (food borne diseases)

I will send out some useful links related to these datasets (See slack channel).

API (Application Programming Interface)

Think of this as someone creating a set of "functions" (reusable code) which you can use in your code without having to deal with the implemantation details. All you need to know is how to call the function/API i.e. what arguments/parameters to pass as inputs and what return values to expect.

R/Python and all other Programming languages are accompanied with a whole host of libraries that provide APIs to do a lot of useful work (like downloading data from the internet).

As part of this course we would be using a lot of APIs. The first example of a library that we would be using would be "httr", we will talk about it in the next class.

CSV (Comma Seperated Value) format

CSV format is one of the most commonly used formats for storing data (especially if we are not dealing with the Big Data).

It simply refers to data stored in a tabular format with coloumn values being separated by "," (comma).

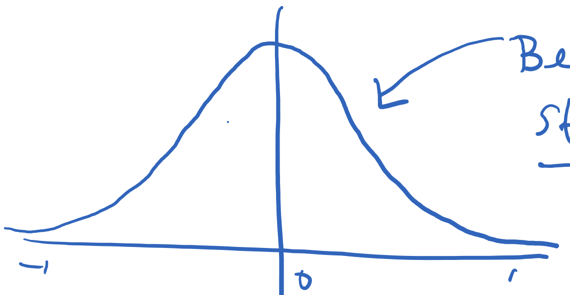
For example:

```
Name, Age, Place
John, 10, Clarksburg
```

Note that each field is seperated by a comma each row is on a different line.

We might also talk about another very commonly used format called JSON (Javascript Object Notation). More convenient for programming rather than for human beings. (If you are interested we may talk about this some more.)

Normal distribution



Bell curve

Standard Normal

→ Mean(average) = 0

→ Standard deviation = 1

One of the most commonly found distributions. We will talk more about this (just a little bit more) in next class.

Percentile / Quantile / decile

We are all familiar with "Percentage" but ther is a more interesting term called Percentile. We have heard it used in context of test scores. For example: someone scored better than the 80th percentile! So what does this mean? An 80th percentile means a value which is greater than 95% of the values in the dataset. So someone scoring greater than 80th percentile means they scored more than the score of 80% of the students in the class. This has universal interpretation: meaning that this statement holds regardless of whether the person scored 65/100 or 40/50 or 7/10 i.e. the absolute value does not matter at all the moment we express things in term of percentile.

Quantile: based on "quarters" i.e. 1st quantile is 25th percentile, 2nd quantile is 50th percentile, 3rd quantile is 75th percentile and 4th quantile is the 100th percentile decile: 10th percentile, 20th percentile and so on and so forth!

Review of dplyr "Verbs"

Select: only keep selected features

filter: only keep observations that satisfy a selection criterion (or criteria)

called verbs because these are action words.

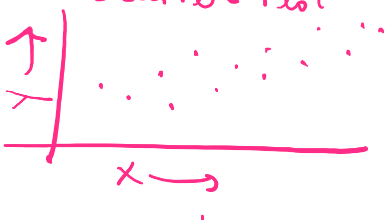
group_by: rearrange the rows in the dataset to keep rows with the same value of the group-by features together. Can have multiple features in a single group-by operation. Usually done in preparation of an "aggregation" type operation.

summarise: calculate a summary statistic like min, max, mean, quantile etc. This is the "aggregation" type operation.

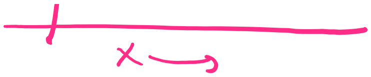
Review of ggplot2 (just the part that we used in class)

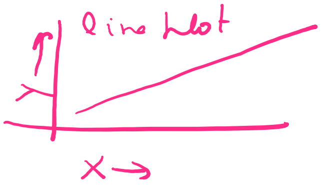
grammar of graphics → each plot is composed of multiple layers i.e. the basic geometery of the plot (line, bar, scatter, map etc.), the use of line colors

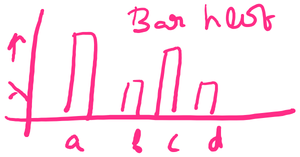
scatter plot



scatter, map etc.), the use of line colors, fill colors, labels, legends etc. everything is a layer on top of the base plot.







⇒ not same as histogram (categorical data on x axis)