INTRO TO DATASCIENCE

Week 1

AGENDA

- Intro to Data Science powerpoint
  - Group Tasks/Discussions
- Intro to SQL by Amulya!
  - Useful for Data Science role of project!

WHAT IS DATA?

- Data is “Information, especially facts or numbers, collected to be examined and considered and used to help decision-making.” - Cambridge Dictionary
- Data is “A set of values of qualitative or quantitative variables” - Wikipedia
- Parts of Data
  - “A set of values”: a set of items to measure from. Called the population in statistics
    - What you are trying to discover something about
  - Variables: measurements of characteristics of an item
    - Qualitative variables: information about qualities
    - Quantitative variables: information about quantities
- Data is rarely presented in a clean, formatted way!
  - E/ Sequencing Data, Population census data, Geographic information system data

WHAT IS DATA SCIENCE?

- A data scientist is broadly defined as.. “who combines the skills of software programmer, statistician and storyteller slash artist to extract the nuggets of gold hidden under mountains of data”
- Why do we need Data Science?
  - The rise in data science is directly correlated to the vast amount of data currently available and being generated as we speak!
    - The more data there is out there, the more questions we can answer with that data!
  - Simultaneously, there has been the rise of inexpensive computing…
    - Rising computer memory capabilities, better processors, more software and now, more data scientists with the skills to put this to use and answer questions using this data!

DISCUSS IN YOUR PROJECT GROUPS

- How and why data is essential to your project
- What data are you hoping to extract from your survey/database?

BIG DATA

- 3 qualities of big data
  - **Volume**
    - Big data involves large data sets that are continuously increasing!
      - Fun Fact: Youtube has 300 hours of video uploaded every minute!
  - **Velocity**
    - Data sets are continuously and exponentially increasing!
      - E/ Real time tracking
  - **Variety**
    - The million types of data out there varies enormously
      - Audio and video from youtube
      - Location and transit from real time tracking

DISCUSS!

- What types of big data do you rely on everyday?
- Ideally, what big data set would you most want to access for your project? Why?

WHAT IS A DATA SCIENTIST?

- A data scientist should embody…
  - Hacking Skills, Math & Stats knowledge, Substantive Expertise
    - Substantive Expertise: we need to have enough expertise in the area that we want to ask about in order to formulate our questions and to know what sorts of data are appropriate to answer that question.
    - Hacking Skills: Once we have our data, it often must undergo cleaning and formatting
    - Math & Stats Knowledge: Analyzing our acquired data
      - Looking back on what big data set we would idealistically access, how would you analyze it?

COOL EXAMPLES OF DATA SCIENCE!

- Google analyzed 50 million common search terms over 5 years & compared them against CDC flu outbreaks. They wanted to see what search terms coincided with outbreaks of the flu, since they have been able to predict flu outbreaks from google searches
- Get your laptops out!
  - myactivity.google.com

TYPES OF DATA SCIENCE QUESTIONS

- **Descriptive Analysis:** to describe or summarize a set of data
  - usually the first kind of analysis you will perform
- **Exploratory Analysis:** to examine and explore the data and find the relationships previously known
- **Inferential Analysis:** to use a relatively small set of data to infer or say something about the population at large
- **Predictive Analysis:** to use current data to make predictions about the future
- **Causal Analysis:** looking at the cause and effect of a relationship!
  - Other analysis can only identify correlations, this can determine the case. Very hard to prove!
- **Mechanistic Analysis:** to understand the exact change in variables that lead to the exact change in other variables

Can you fit the examples that we've gone over into a certain type of analysis?
What type of analysis will your project have to do?

EXPERIMENTAL DESIGN

- Experimental design is organizing an experiment so that you have the correct data (and enough of it!) to clearly and effectively answer your data science question.

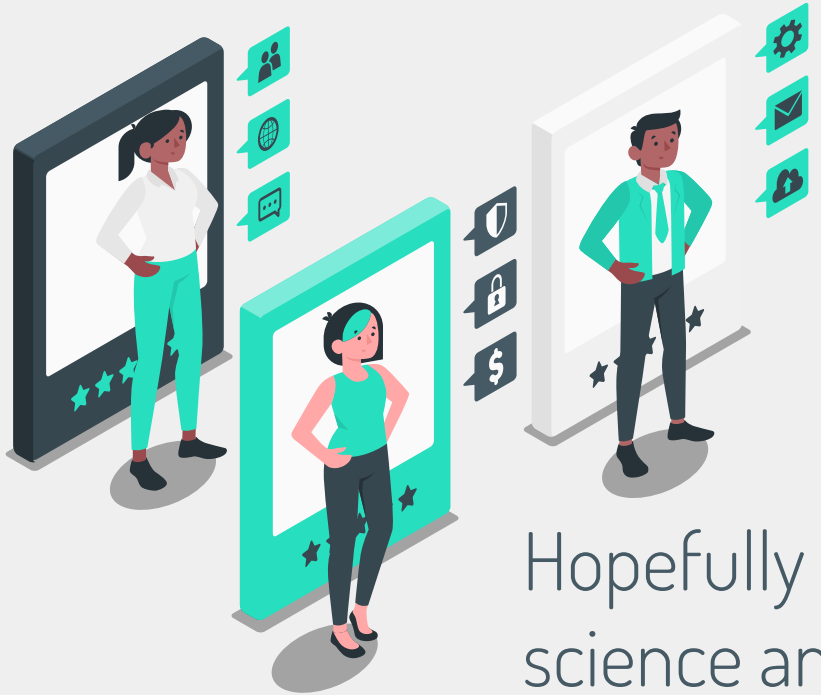
- Going into an analysis, you need to have a plan in advance of what you are going to do and how you are going to analyse the data. If you do the wrong analysis, you can come to the wrong conclusions!
    - Important this doesn't happen in high stake studies

- **Independent variable (AKA factor):** The variable that the experimenter manipulates; it does not depend on other variables being measured. Often displayed on the x-axis.

- **Dependent variable:** The variable that is expected to change as a result of changes in the independent variable. Often displayed on the y-axis, so that changes in X, the independent variable, effect changes in Y.

EXPERIMENTAL DESIGN

- **Hypothesis**: An essentially an educated guess as to the relationship between your variables and the outcome of your experiment.
- **Sample size**: The number of experimental subjects you will include in your experiment.
- **Confounder:** An extraneous variable that may affect the relationship between the dependent and independent variables.
- **Control group**: When you have a group of experimental subjects that are *not* manipulated.
- **Treatment group:** The group of experimental subjects that are manipulated
- **Blinded:** Subjects don’t know what group they are in
- **Placebo effect:** A beneficial effect produced by a placebo drug or treatment, which cannot be attributed to the properties of the placebo itself, and must therefore be due to the patient's belief in that treatment.



THANKS FOR LISTENING

Hopefully you better understand data science and feel prepared to tackle your data science role! Next: Data Science coding demo.

Any Questions?