DSCI-560
Lecture 1: Introduction
Data Science Professional Practicum

Young Cho
Department of Computer Science
University of Southern California

The Goal

- Solve A Real-World Problem
  - Identify A Real-World Problem
  - Plan Your Approach with Milestones
  - Implement Your Solution
  - Demonstrate/Evaluate Your Solution
  - Present Your Solution
- Implementing Solution
  - Plan your Approach with Mitigation Plan
  - A Milestone at a Time
  - Document Your System
- No Time to Waste

Logistics

- Lectures
  - Tuesday and Thursday at 9 AM - 10:50 AM
  - Attendance is Mandatory
  - Check IN and Check OUT
- Instructor: Young Cho
- Teaching Assistant: TBD
- Web Site: Coursistant
  - https://usc.xlearnedu.com
  - Please log in for the assignments, announcements, and Q&A
  - Select Forgot Password

Grading

- Attendance
- Reading Assignment
- Laboratory Assignments
- Final Project

Project

- Laboratory Assignment
  - Individual or Team
  - YouTube Video Recordings
- Tools
  - VMWare
  - Linux Tools
  - Database
  - Amazon Web Services
  - Google Cloud
  - GitHub

Online Course Tools

- Brightspace
  - Announcements
  - Laboratory Assignments
  - Reading Assignments
- Piazza for Q&A
  - Post all of your questions and answers
  - Coursistant will answer as quickly as possible
  - Most of the time, it will give you sufficient assistant
  - However, it is still learning and sometimes needs corrections

Laboratory Assignments

- Submit Archive of Laboratory Results
  - Report with Snapshots and Results
  - Source Code and README instructions for code
  - Usually Saturday 11:59 PM
- YouTube demonstration video
  - Usually, the following Monday 11:59 PM
- Scoring
  - On-time Submissions (no exceptions)
  - Multiplied by 1.0 if on-time
  - Multiplied by 0.5 if late
  - Demonstration Videos on YouTube
  - Internet Accessible Link to the Archive (ZIP) of all

Incorporating Videos in Presentation/Demo

- Use Smartphone Camera
  - Capture your audio and video
  - Prove your work/progress
- Screen Capture Software
  - http://www.maartenbaert.be/simplescreenrecorder/
  - https://www.ispringsolutions.com/ispring-free-cam
  - https://getsharex.com/
  - Or other screen capture software
- Content
  - Capture essential steps with result narrative
  - Show the completed result of the Lab
  - Include Powerpoint Slides for diagrams and tables, if needed

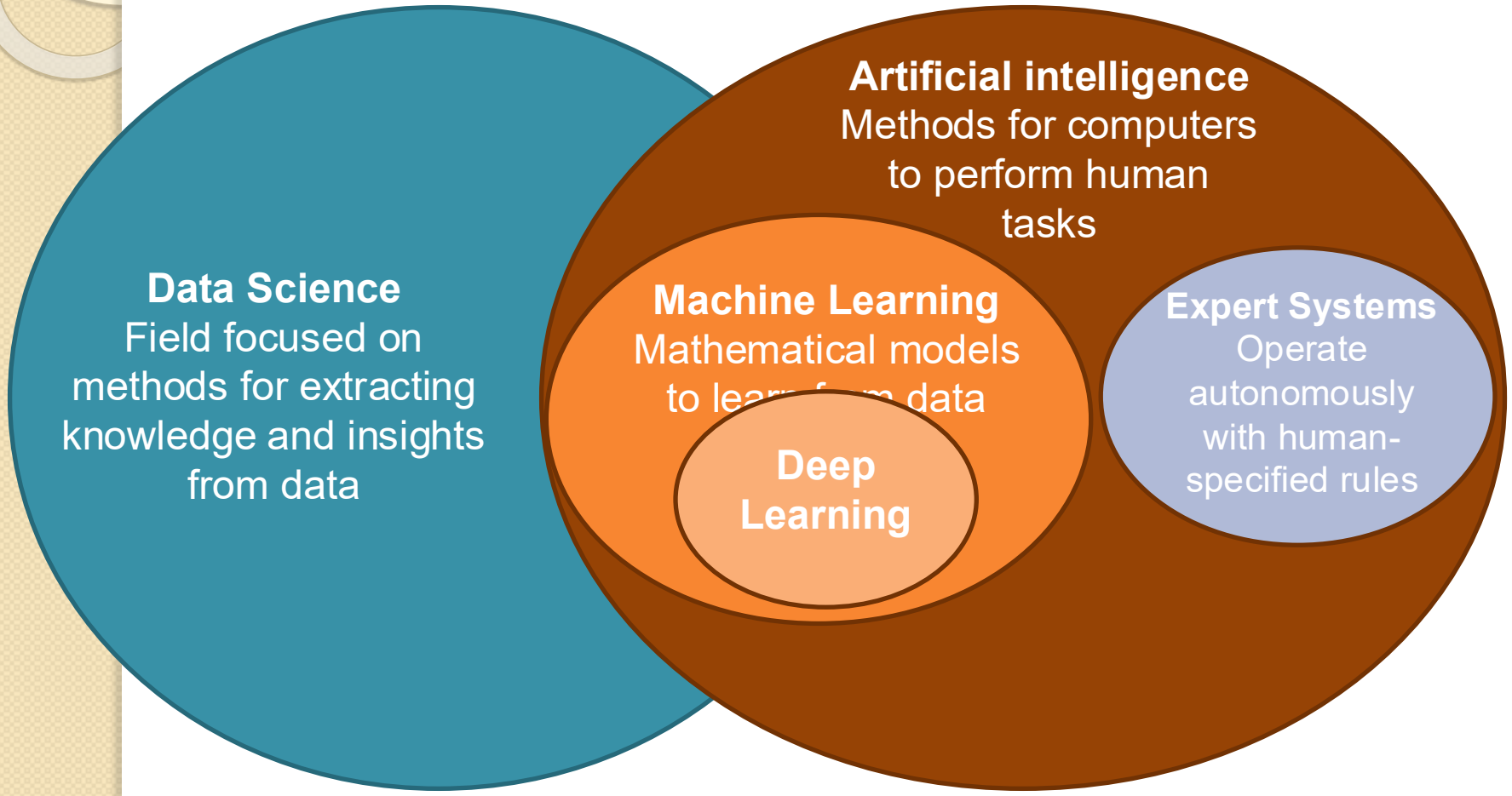
Reading Assignments

- About 10 Assignments
  - Submit Summary Slides on Blackboard
- Summary PowerPoint Slides
  - Summary slides of your understanding
  - At least one hand-drawn diagram
  - Summary tables, if possible
  - Create an intelligent Quiz Question and Answer for the paper
  - Usually, Due Tuesdays at 11:59 PM

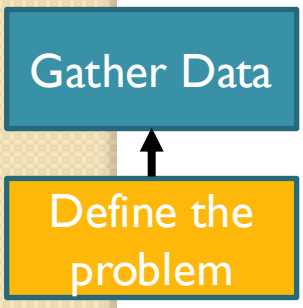
Objectives

- Apply principles studied in Data Informatics curriculum towards a real-world challenge
- Understand the requirements and objectives of clients, how these vary, and how one must tailor a solution to the expectations of a customer.
- Work in a team to accomplish a common goal
- Meet aggressive deadlines in a team effort
- Improve project-based presentation skills
- Present and sell ideas to high-level management

What is Data Science

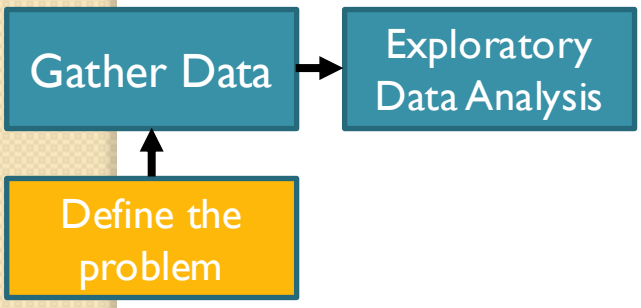


Data Science Pipeline

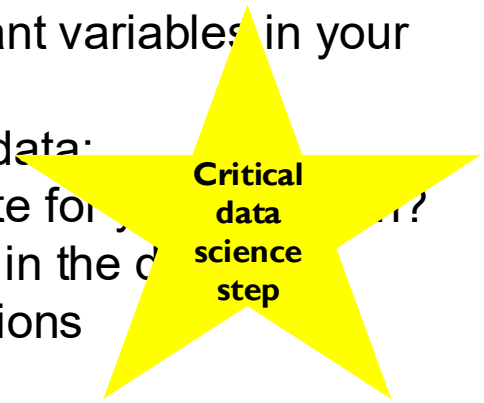


- Collect your own data
  - Surveys
  - Experiment
  - Theory-based model
- Search repositories

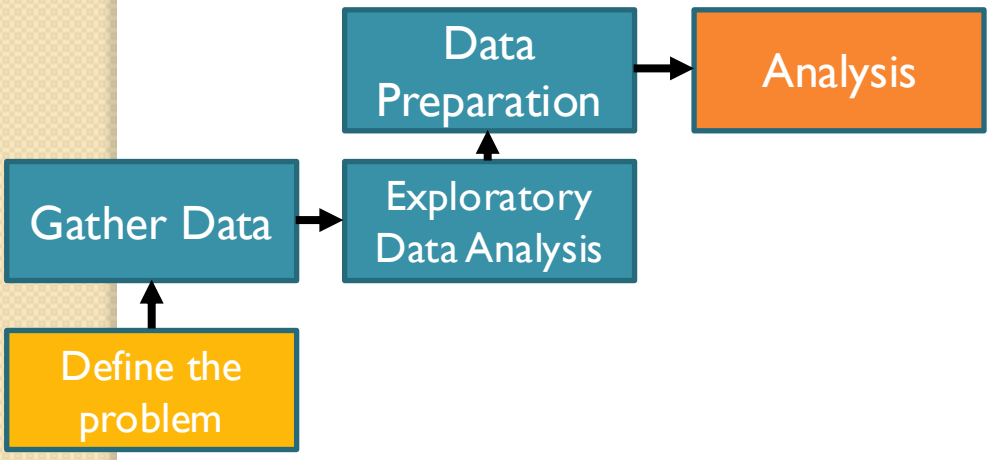
Data Science Pipeline



- Check for missing data and other mistakes
- Mapping and understanding the underlying structure of your data
- Identify the most important variables in your dataset
- Gain insight about your data:
  - Is the data appropriate for
  - Are there any biases in the d
- Often involves visualizations

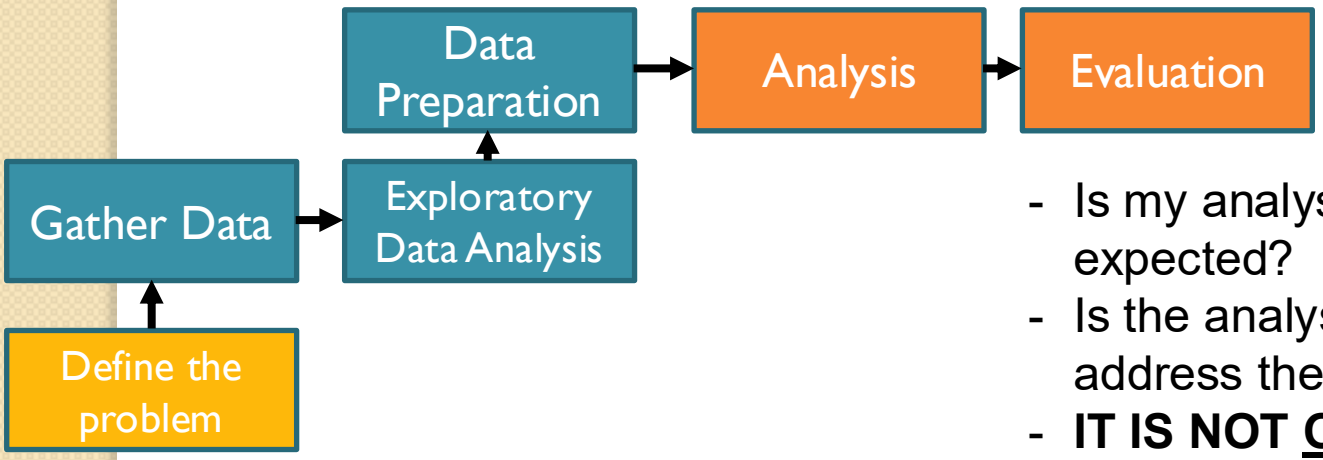


Data Science Pipeline



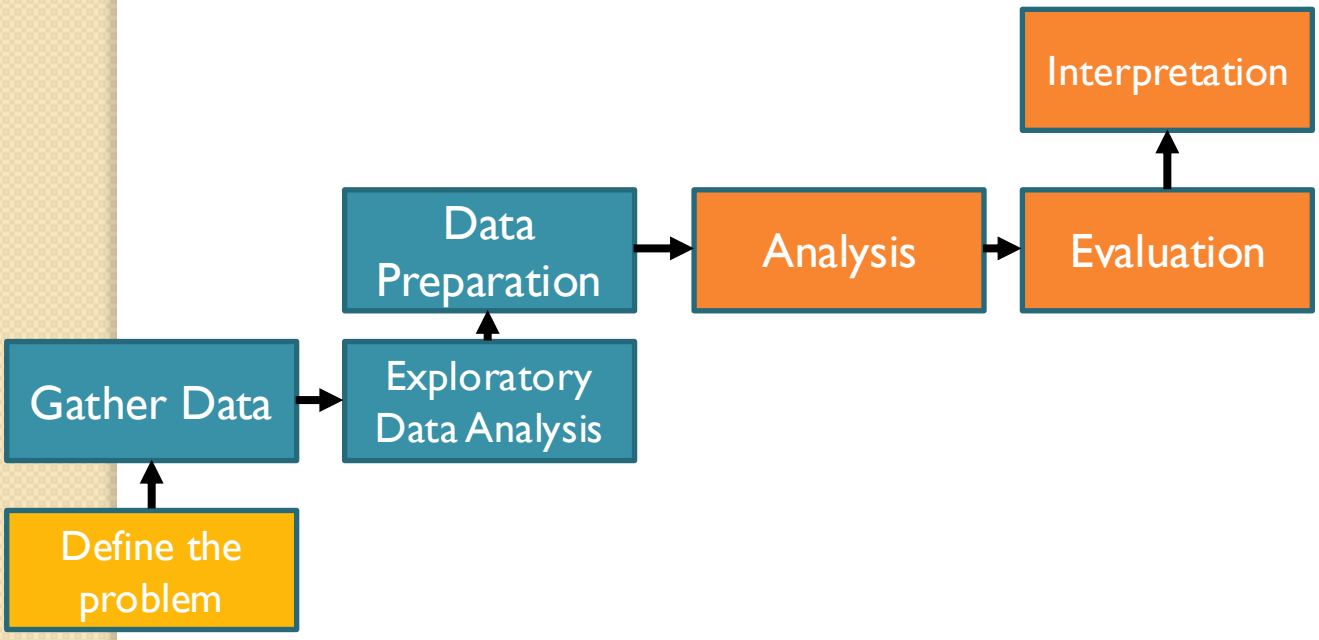
- Choose appropriate analysis for the question
- If using ML and a trained model, is the training data similar to the data to be analyzed?
- What are the pre-processing steps?

Data Science Pipeline

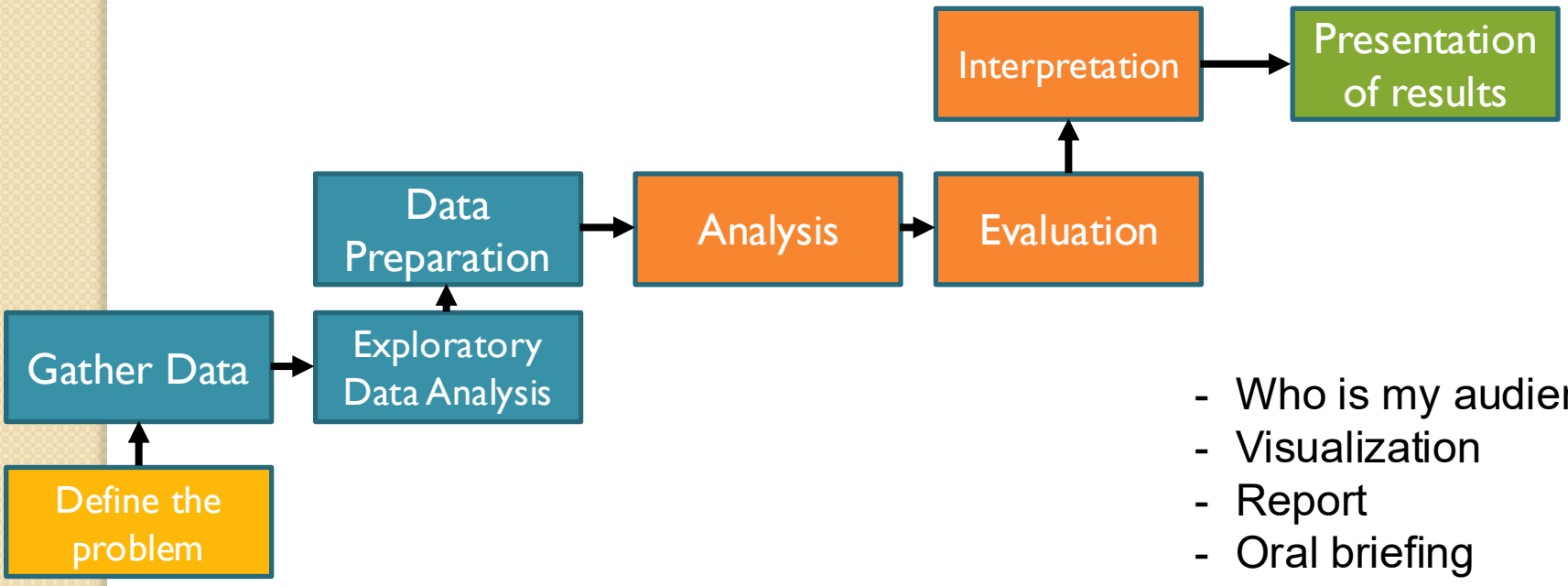


- Is my analysis performing as expected?
- Is the analysis allowing me to address the problem?
- **IT IS NOT <u>ONLY</u> ABOUT THE ACCURACY SCORE**

Data Science Pipeline

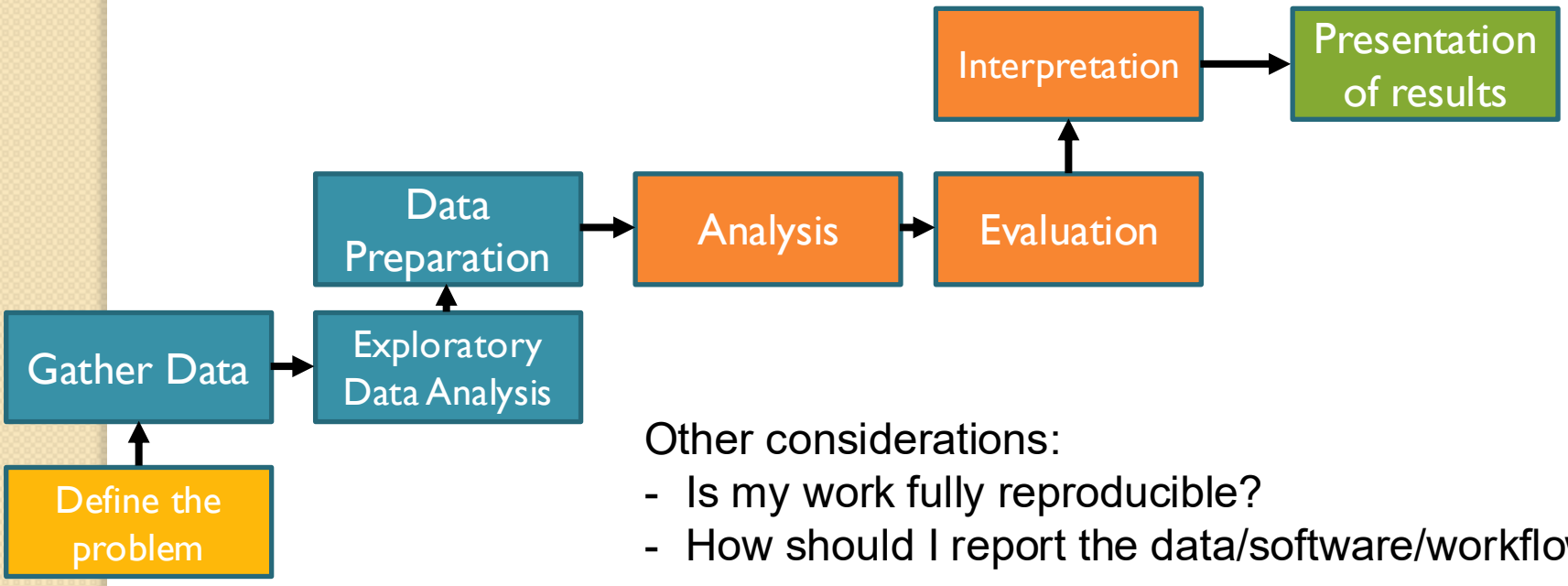


Data Science Pipeline



- Who is my audience?
- Visualization
- Report
- Oral briefing

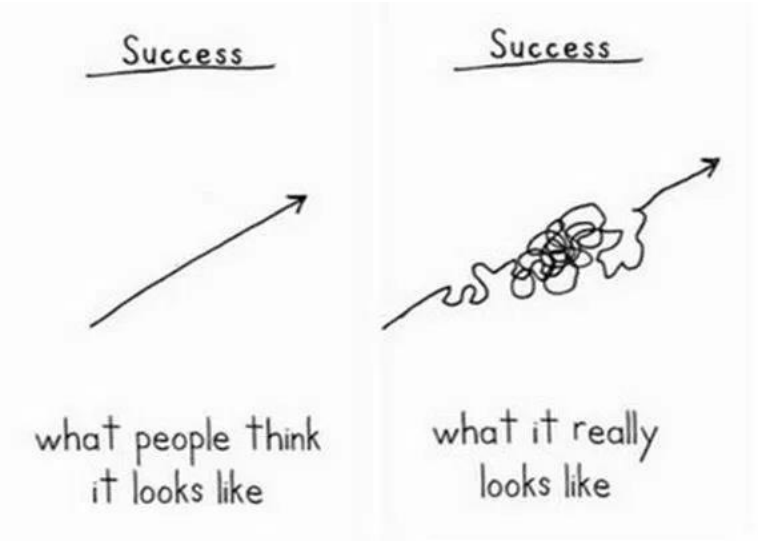
Data Science Pipeline



Other considerations:

- Is my work fully reproducible?
- How should I report the data/software/workflow?

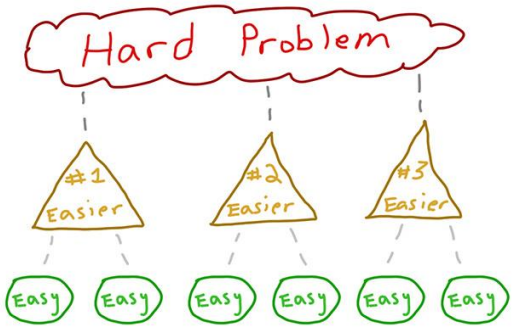
Failure is the Mother of Success



- Progress is rarely linear
- Missing/flawed data
- Problems need to be more refined
- Initial approach didn't work
- Initial hypotheses invalidated
- **Your goal is to learn how to move forward**

Debugging

- List all the components involved. Think about where the weak link might be
- Design a logical and simple troubleshooting process to find the problem
- Ask for help. Ask the internet, ask your peers…
- **<u>Research on your own</u>**



Data May Be the Problem

- Data Should be Representative of the Problem
- Data Should be Qualitatively the Same as the Training data
- Data Should be Without Bias, Sufficiently Large, and Spans All Interested Range
- **All Data Needs Cleaning before being Usable**

Product Problem

- Context
  - Company that has an on-line shopping site which would like to start to push products to customers as they browse the site
- Data
  - Data about the on-line purchases of customers for the last 5 years
  - Profile data for some customers who are repeat customers: address, credit card, shipping preferences
  - Data about customers who recommended products to their friends in order to get a discount

- Cost
  - When you push random products, 10% of customers do not like what is pushed to them and they leave the site
- Challenges
  - When you have a sale, many repeat customers buy many more items than usual
  - There is no profile data for many customers that pay through a third-party service

Bee Problem

- Context
  - Government of an island who would like to investigate how to reduce the bees so tourism can thrive again
- Data
  - Data available about the weekly water levels of all rivers and ponds for 30 years
  - Medical reports of bee bites and pollen allergies for the last 20 years
  - Climate data and population data, including rainfall and temperatures as well as pollen levels
  - A lot of data about population, pollution, pesticide use, and bird populations (bee predators)

- Challenges
  - Two bee experts in the island, but they do not know anything about data science.
  - What questions would you ask of them to help you figure out how to solve the problem?
  - May release pesticides on crops
    - Pesticides cost $1,000 per square kilometer
    - Pesticides reduce the bee population for 3 months
  - An environmental group that claims that the bee population can be reduced naturally by planting crops that have no flowers (eg corn, wheat, etc)

Fraud Problem

- Context
  - A bank, interested in detecting fraudulent activity in credit card customers
- Data
  - For each customer, there is detailed information from their card application about their address, salary and employment, and demographic
  - For each customer, there is a record of all their transactions (date, charges, and vendor) for the last 4 years
  - For 1% of customers, there is a flag that their credit card was reissued because of fraudulent use of their prior card
  - Additional data available with fee, like census data for any zip code

- Cost
  - When a fraud goes undetected, the average loss to the company is $3K
  - Reissuing a customer card costs $50
  - When card is reissued and there is no fraud, 0.5% of customers cancel their card
- Challenges
  - Volume of the data
    - There are 100M customers, with 30K transactions on average
  - Some credit cards were reissued but no fraud took place once investigated

Disaster Relief Problem

- Context
  - A non-profit organization in a remote country who would like to understand where to send relief and in what form
- Data
  - Microblog data (eg twitter), where people are posting issues with bridges, roads, and general access to remote locations
  - Many hospitals are emailing hourly reports, with number of beds occupied and available, medical inventory status, and medical personnel
- Cost
  - A number of coders have contacted your headquarters to volunteer their time to help with data analysis and any data collection needed
- Challenges
  - The remote country's government seems open to take your advice for what roads need repair, what hospitals need more personnel, etc, but will ask for detailed justifications of all your recommendations