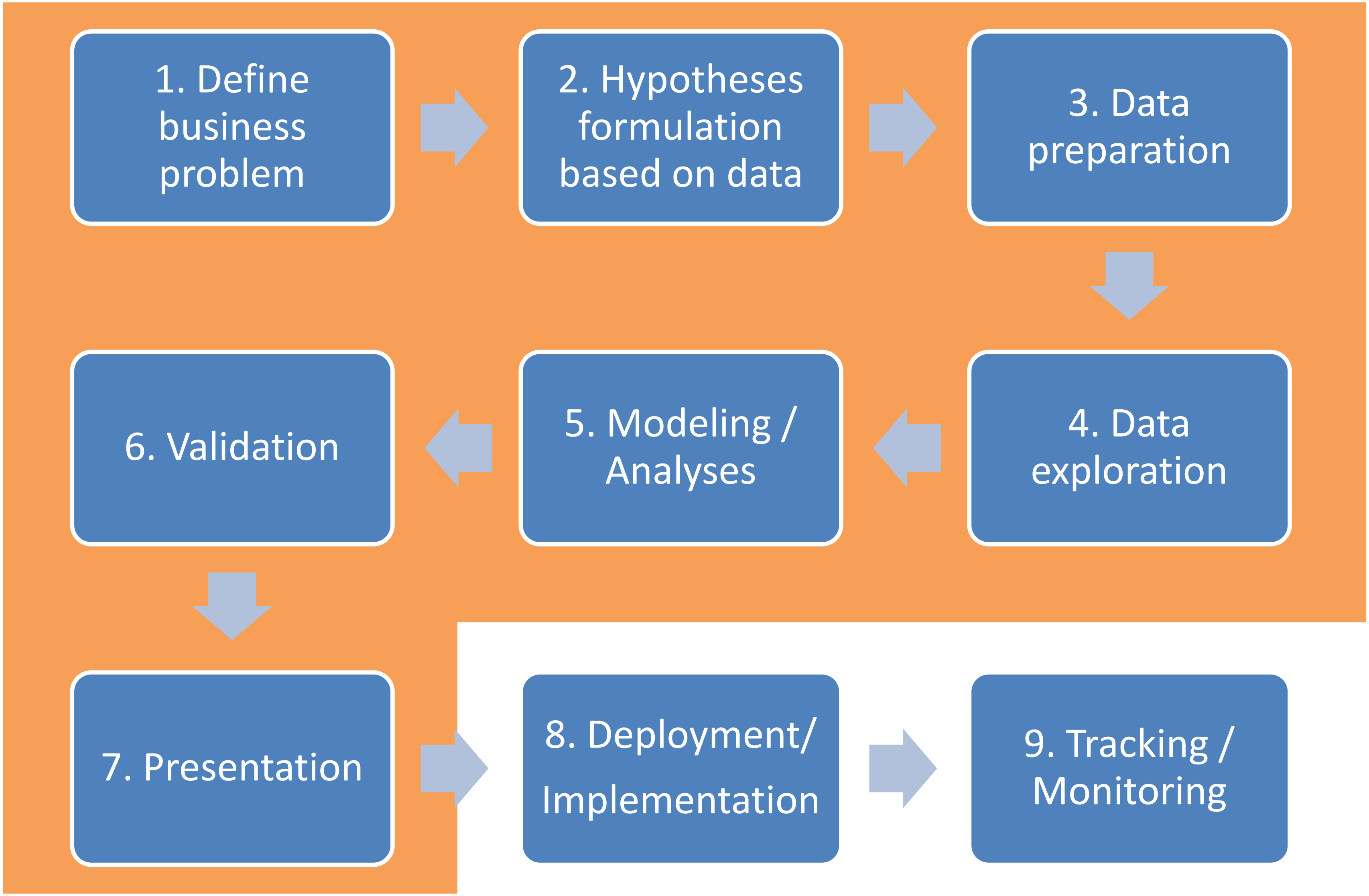
Information on Group Project

Tongil “TI” Kim

Analytics Project Cycle



1. Define business problem

2. Hypotheses formulation based on data

3. Data preparation

4. Data exploration

5. Modeling / Analyses

6. Validation

7. Presentation

8. Deployment/ Implementation

9. Tracking / Monitoring

Overview

- Go through each step of the cycle and present your results at the end of the semester.

- Data
  - BYOD (bring your own data) through public data, cold call, personal connections, scrapping websites, etc.
  - Obtain data from 3rd party platforms
    - Google dataset search (link)
    - Kaggle datasets (link), Kdnuggets (link)
    - ICPCR (all kinds of social science data) (link)
    - Amazon Data Exchange (link). Not all of them are free.
    - Stanford Large Network Datasets (link)
    - National Bureau of Economic Research data (link)
    - data.world (link)
    - Yelp datasets (link)
    - Airbnb datasets (link)
    - More Airbnb data (link)
    - National Youth Tobacco Survey by CDC (link)
  - You need to clearly disclose the source of the data. Also make sure that you have the right to use the data for class presentation.

Step 1: Define Business Problem

- All business analytics should start with a business problem
  - What is the issue?
  - What are potential causes?

- The business problem should be relevant and interesting…. not just to you, but to the right stakeholder (firms or governments)
  - How much will managers or the CEO want to know the answer to this question to better serve their customers and improve business processes?
  - Will answering this question help government agencies to formulate policies?

- Always start small, and scale if possible. A small, complete project is better than an ambitious, incomplete project

Step 2: Hypotheses Formulation Based on Data

- Formulate testable hypotheses to solve the business problem

- Data analytics = science + art. This is the ‘art’ part. Be creative

- You should also think about the quality of your data. Is your data good enough to test the hypotheses. Do your data have relevant variables (i.e., columns)? Do you have a large enough sample (i.e., number of rows)?

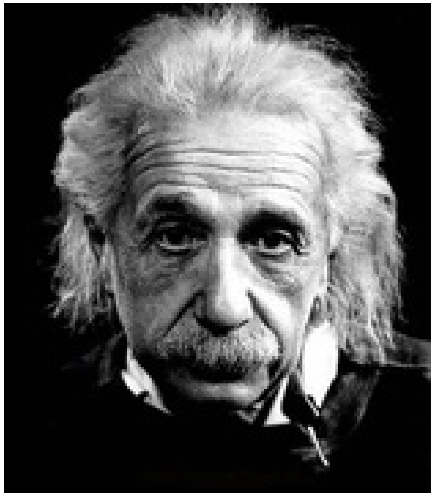
Steps 3 - 6: Data Work

- Prepare data (i.e., data cleaning)
- Explore data
  - Visualize data (histogram, scatter plots, bar graphs, etc.)
  - Check simple correlations and run simple regressions
- Modeling / Data Analyses
- Validation

Steps 7: Presentation

- This is your group project deliverable
- Scenario
  - You are presenting your project to managers in companies or government agencies with various background (some with advanced degrees in statistics and computer science, other with general business degree)
  - The goal is to convince the group that your analyses identify an interesting question and deliver clear insights and/or recommendations, even to those managers with minimal training in statistics

Steps 7: Presentation



"You do not really understand something unless you can explain it to your grandmother."

– Albert Einstein

Evaluation Criteria

Data originality, quality

Interesting question(s)?

Rigorous analyses?

Presentation for any manager

Deliverables

- Group project outline (5%)
  - Decide on the data set. If data collection is in progress, you need to explain the process and provide an expected date of data receipt
  - State business question(s) and identify parties who will find the question relevant and interesting.
  - Plans for data analyses

- Presentation (15%)

Example 1

- **College Admissions Scandal**
  - Imagine that you are a data scientist who’s been recruited to help detect fraud during the college admissions process. For this conversation, we shall narrow the focus to fraudulent information submitted in the college application forms, whether it is an inflated GPA, an invented sports achievement, or a fake community service achievement, or other types of forgeries.
  - You will be building a set of fraud detection models. Tell me what your first model will do, and why you choose that as your first model.
  - Imagine I am the director of admissions. Tell me why I should pay for your model.
  - What training data will you need to run that model? Where and how will you obtain the data?

Example 2

- **Blue Apron post-IPO**
  - In Q2 2018, Blue Apron, the meal-kit delivery business, reported that about 700,000 customers, 24 percent lower year over year while revenue per customer was $250, slightly down by $1 year over year. https://investors.blueapron.com/press-releases/2018/08-02-2018-120246005
- Senior management is desperate to stem the customer churn. You are tasked with finding the reasons for the customer churn.
- Come up with 5 hypotheses for why the number of customers dropped drastically.
- Pick one of those hypotheses, and describe how you'd validate it.
- Assume you are able to validate the hypothesis, explain what you would recommend to reverse the customer churn.