Missing Data

Lecture 13

Xiaofei Shi

Learning Objectives

- Exposure to the different sources of missing data
- Learn how to deal with missing data

Where have you seen missing data before?

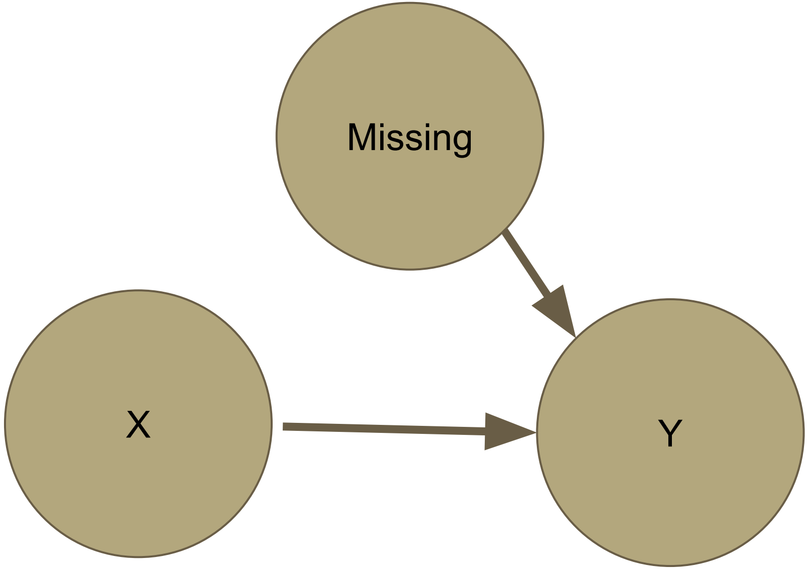
You should have seen some in the previous classes by now….

Why do we care about missing data?

- Statistical power
- Bias
    - Representativeness
- Make analysis more difficult

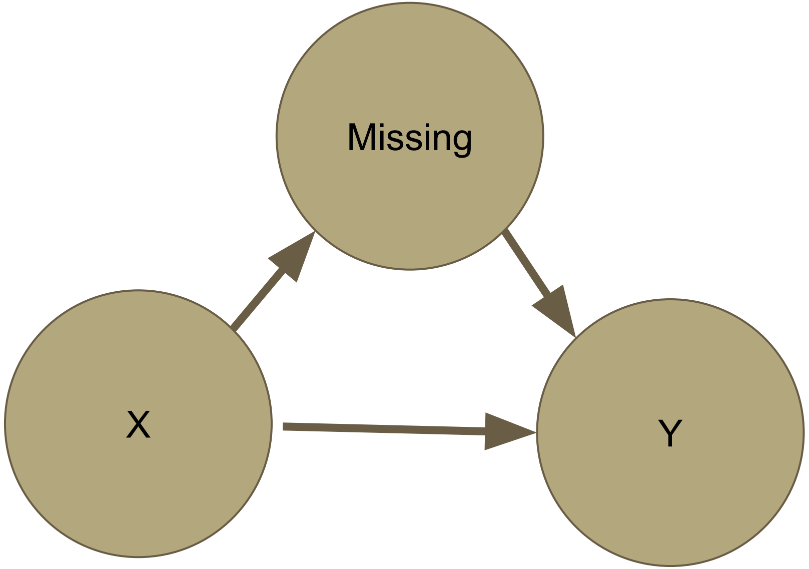
Framework on missing data

- Missing completely at random (MCAR)



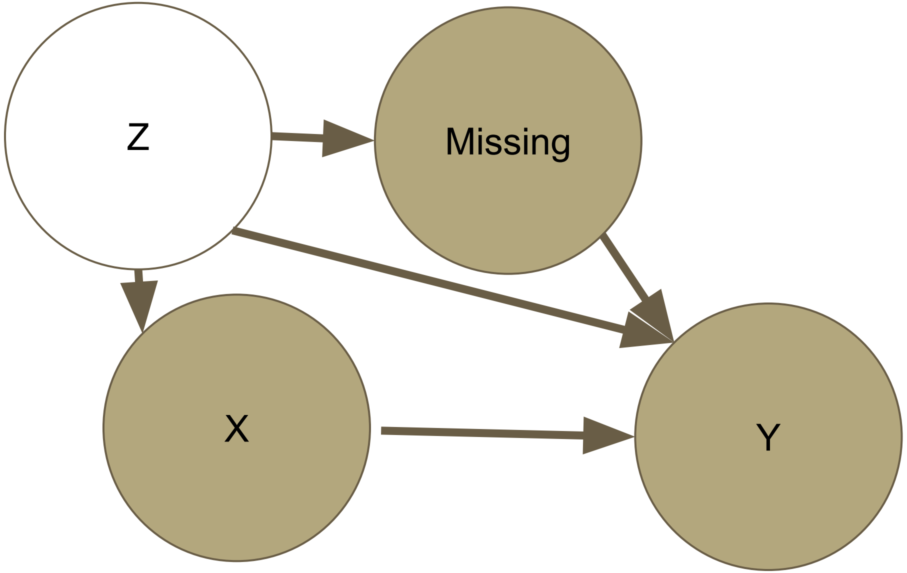
Framework on missing data

- Missing completely at random (MCAR)
- Missing at random (MAR)



Framework on missing data

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing NOT at random (MNAR)



Framework on missing data

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing NOT at random (MNAR)

Traps that invite missing values

- Too many “nice-to-have” questions in surveys
- Not understanding how data is measured/collected
- Not asking or explaining for the data you need precisely
    - Vague: Rate X on their statistical understanding
    - 2 questions in 1: Do you want higher taxes to fund public education?
    - Not gaining trust: “just collect the data”
- Poor design
- ….

Handling missing values

- Best strategy: avoid having missing values
- Ignore the missing values
- Predict the missing values
- Model the missing values

Missing Words in Job Descriptions

- What type of missing data is this?
- How should we handle it?

Non-response in census

- What type of missing data is this?
- How should we handle it?

Outcome from the opposite treatment

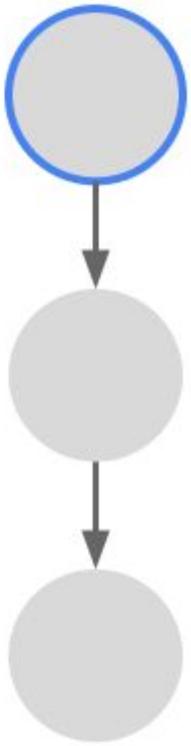
- What type of missing data is this?
- How should we handle it?

About HW 5 and the project

- Make sure you know how to submit on Kaggle!
- You can assume the comparison baseline on kaggle is indeed the ground truth!

Propose a detailed pipeline to solve the problem

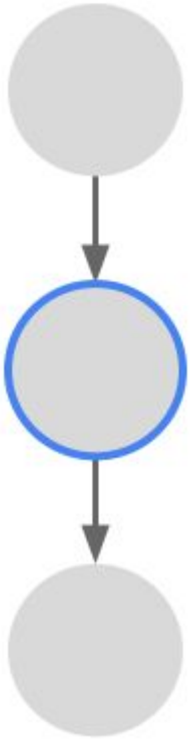
- Preprocessing
- Feature engineering



Propose a detailed pipeline to solve the problem

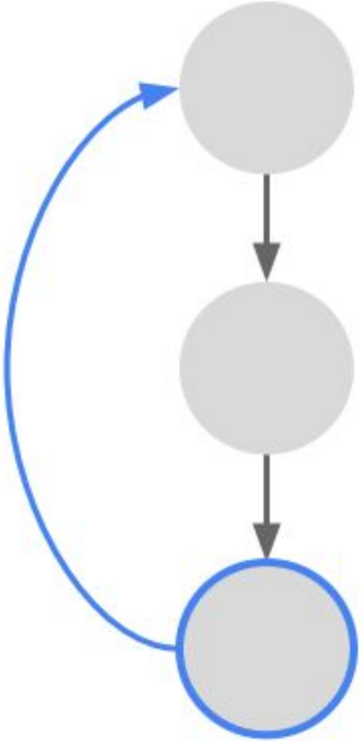
- Split into train/validation/testing sets

- Potential methods:
  - linear/ polynomial regression
  - Ridge regression
  - Lasso
  - Naive Bayes
  - Ensemble methods
  - ...

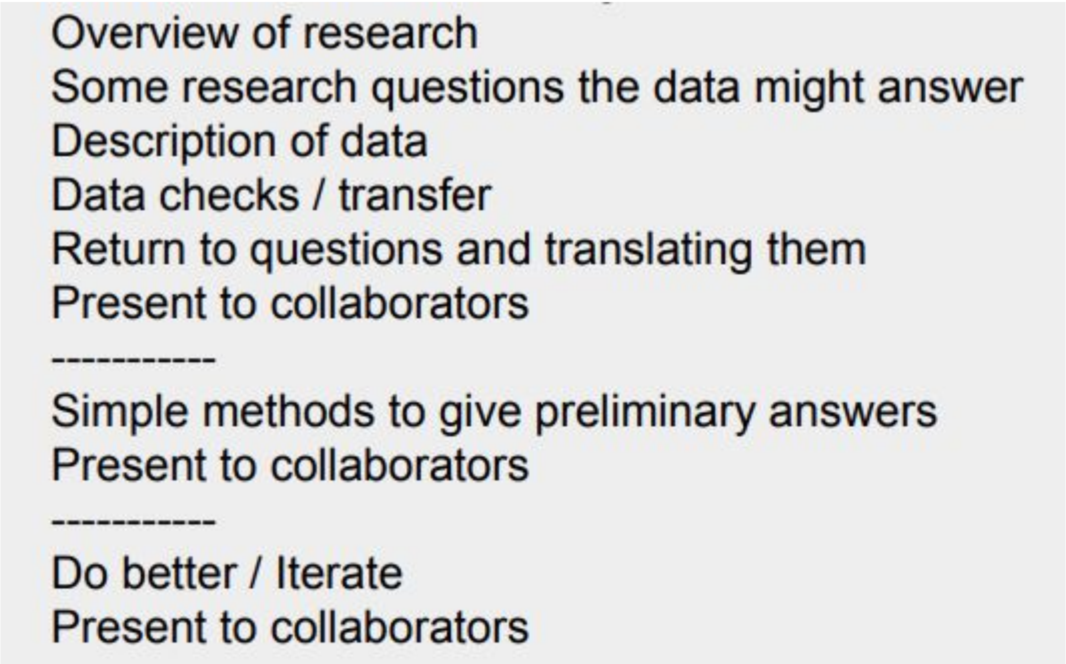


Propose a detailed pipeline to solve the problem

- Evaluate
- Diagnose
- Iterate!



Typical steps of applied data analysis



Overview of research
Some research questions the data might answer
Description of data
Data checks / transfer
Return to questions and translating them
Present to collaborators

Simple methods to give preliminary answers
Present to collaborators

Do better / Iterate
Present to collaborators

Any thoughts?...