Data Preparation

Data Mining: Seminar 3

Dr. Thomas Price

AI Academy

NC STATE UNIVERSITY

Review Exercises

AI Academy

Review: Attribute Types

Table 1: Students Dataset

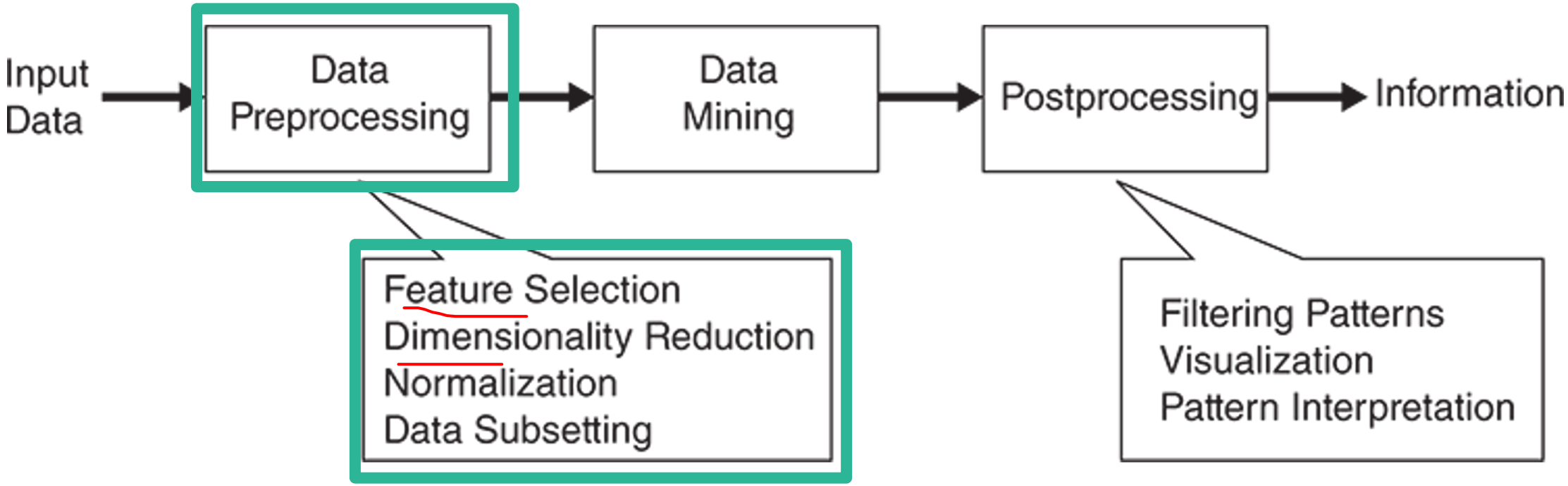
Course	StudentID	GroupID	# of Teammates	Grade	Letter
STAT 501	001	G11	3	92.1	A-
STAT 505	002	G13	3	89.2	B+
STAT 511	005	G02	2	93.6	A
CS 516	007	S03	2	95.0	A
CS 522	202	S03	3	85.3	B
CS 589	203	G02	2	82.4	B-
PSY 501	003	G06	3	78.2	C+
PSY 505	003	S02	3	86.7	B
PSY 516	391	S07	3	93.1	A
PSY 530	226	G08	2	96.2	A

Sampling

Lesson

AI Academy

The Data Mining Pipeline



Data Preprocessing

- **Sampling**
- Feature subset selection
- Dimensionality Reduction
- Feature creation
- Discretization and binarization
- Attribute Transformation

Case Study: Risks in Pregnancy

- *DARPA study of a large sample of pregnant woman who visited military hospitals*
- **Question:** What factors reduce risks during pregnancy?
- **Conclusion:** The factor with the largest impact is a pregnant woman being single!

  Statistically significant!
- **What is wrong?**

Case Study: Risks in Pregnancy

- **Problem**: Two main types of women in military hospitals:
  - Women **in the armed forces**, likely to be very fit and healthy.
    - May or may not be married.
  - Women **married** to people in the armed forces.
    - Almost all are married.

Data Biases

Watch out for data biases

- Try to understand the data source
- Sample data used for analysis should represent data we want to understand

Results (conclusions) derived from biased data do not hold in general!

- It is very easy to derive "unexpected" results when data used for analysis and learning are biased.

Case Study: Risks in Pregnancy

- **Population:** All pregnant woman
- **Sample**: Women at a military hospital
    - Does the sample represent the population?

Why sample at all?

Why Sample?

- **Obtaining** *all relevant data* is too expensive and time consuming.
  - It's often impossible!

- **Processing** *all relevant data* is too slow and difficult.

Sampling

- **Sampling**: Limiting analysis to only a subset of data objects in a population
  - It can occur as data is collected (to save time/effort) or after (to reduce/focus data)

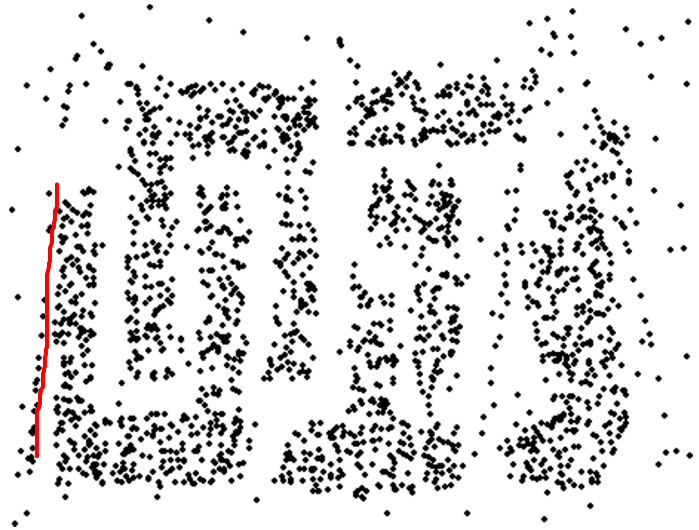
Principles of Effective Sampling

The sample must be representative of the whole population to avoid **bias** in the sample.

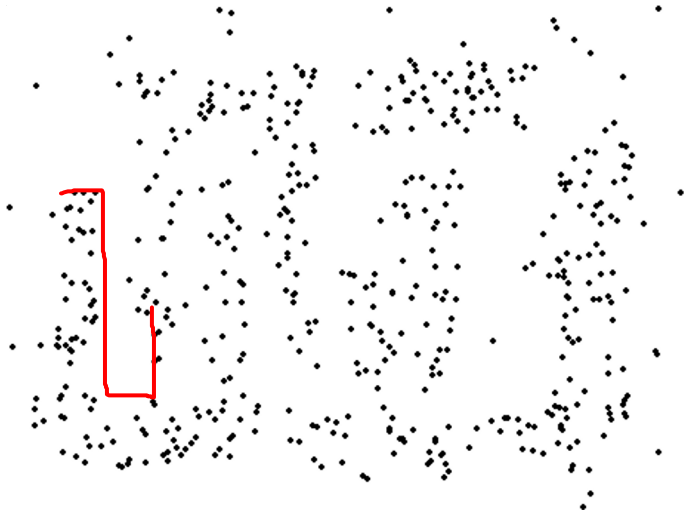
A sample is representative if it has approximately the same properties (of interest) as the original set of data

Sample Size

8000 points



2000 Points



500 Points

To be representative, a sample must be large enough to represent the patterns in the population.

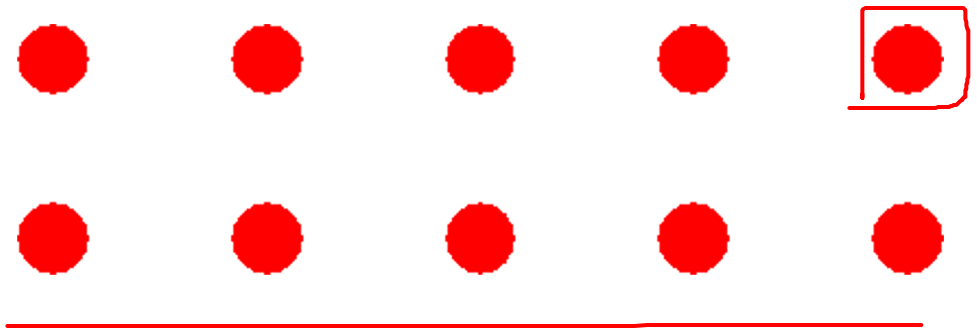
Simple Random Sampling

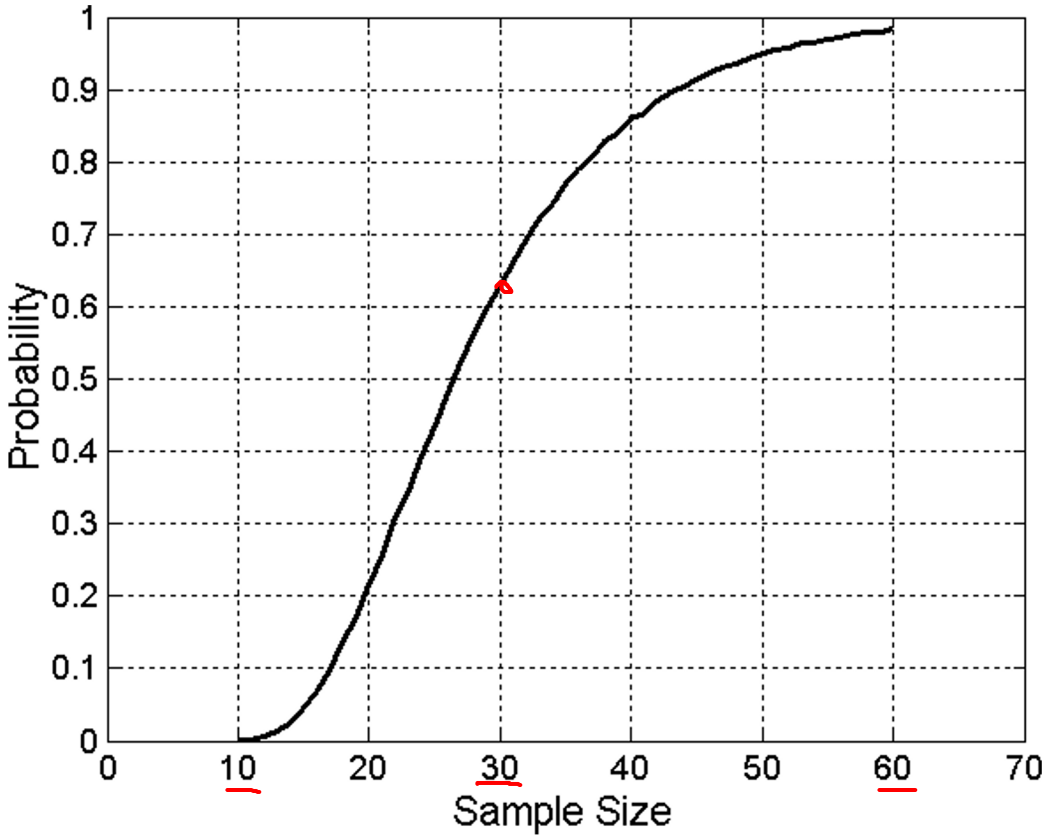
- Each item has the same probability of being selected
  - Sampling without replacement
  - Sampling with replacement
- Similar results with large enough population

- **Limitation**: If some relevant objects are rare (e.g. 1%), random sampling may not be representative

Sample Size

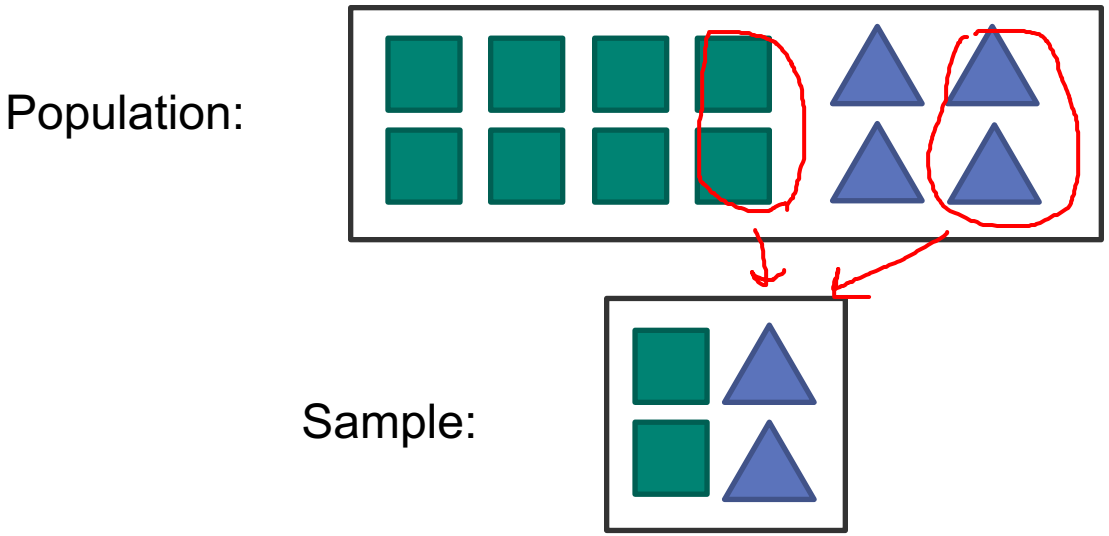
What sample size is necessary to get at least one object from each of 10 groups?





Stratified Sampling

- Draw the same number of objects from each group.
- Ensures all groups are represented.



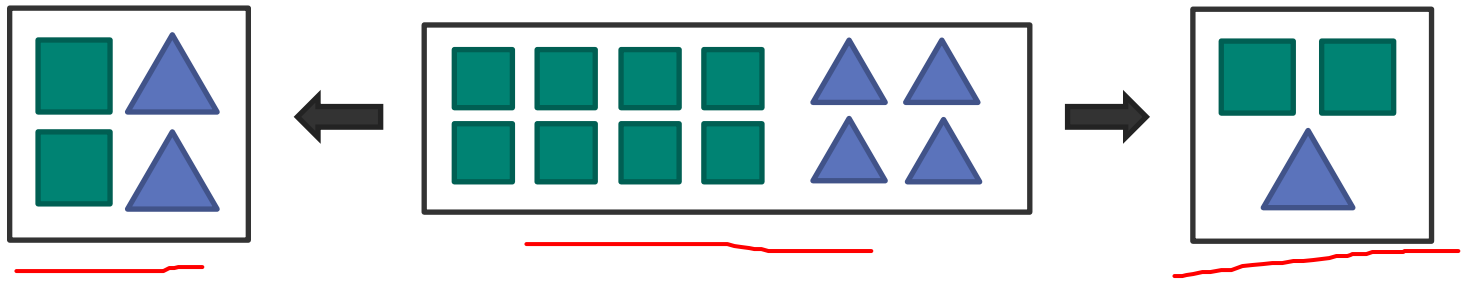
Stratified Sampling Variations

Same number from each group.

- e.g. 50 from Iceland, 50 from China.

Number sampled is proportional to group size.

- e.g. 1 from Iceland, 99 from China.



Progressive Sampling

- Start with a small sample
- Increase the sample size until the size is sufficient
- Eliminates the need to determine the sample size
- Requires a way to evaluate the sample and judge if it's large enough

Lessons from the Stock Market: Backtest Biases

Goal: Analyze stocks from 1920-2020 to predict stock trends for 2021-2025

- **Backtest:** Test a strategy on historical data to predict how well it will do in the future

Challenges learning from historical data:

- **Survivorship Bias:** Only analyzed stocks that survived until today
- **Lookahead Bias:** We encoded rules into our model based on our modern knowledge

Learning Objectives: Sampling

You now should be able to:

Sampling: Explain the notion of sampling and compare the pros and cons of different sampling methods

AI Academy
NC STATE

AI Academy

Sampling

Exercises

AI Academy

Practice Question: Sampling

(On Moodle): Match the sampling method with the most appropriate situation to use it.