DATA PREPROCESSING

BEIYU LIN

TODAY'S TOPIC

- Data Preprocessing
- Exploratory Analysis
- Post-processing

THE DATA ANALYSIS PIPELINE

- Mining steps

Preprocessing: real data is noisy, incomplete and inconsistent.

- sampling, dimensionality reduction, feature selection.

Post-Processing: make data easy to interpret and useful to users

- statistical analysis of importance
- visualization.

DATA QUALITY

- Data quality problems:
  - Noise and outliers (example in yellow box)
  - Missing values (in red box)
  - Duplicate data (in green box)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

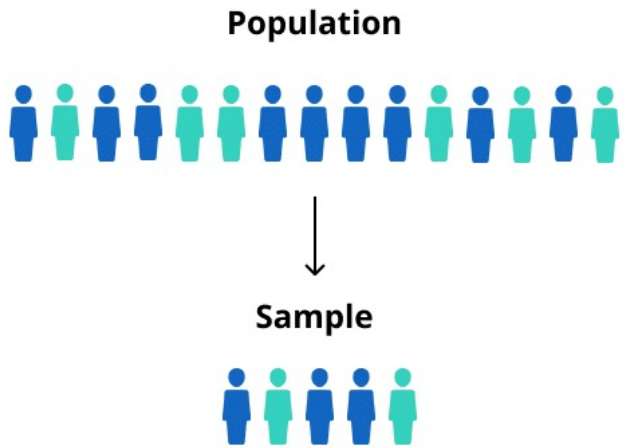
SAMPLING

- Why sampling:
  - obtaining the entire set of data of interest is too expensive or time consuming.
  - e.g., calculate the average height of a person in Las Vegas?

    population in 2019: 634,773
  - what fraction of tweets in a year contain the word "Greece"?

    300M tweets per day, if 100 characters on average, 86.5TB to store all tweets

SAMPLING …



- key principles for effective sampling:
  - using a sample that is representative to estimate the entire data sets
- What is a representative sample
  - if a sample has approximately the same property (of interest) as the original set of entire data
  - otherwise, the sample introduces some bias

- Question: what happens if we take a sample from UNLV to compute the average height of a person in Las Vegas?

TYPES OF SAMPLING

- Simple Random Sampling

  any particular item will be selected with an equal probability

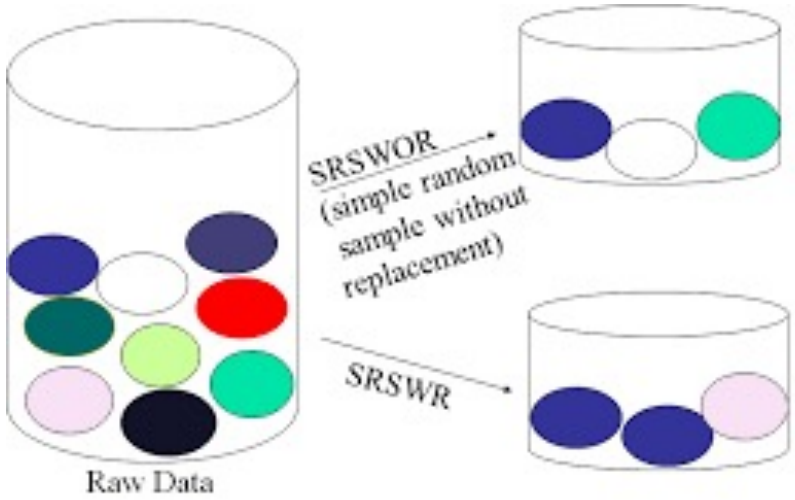
- Sampling without replacement

  when an item is selected, it is removed

- Sampling with replacement

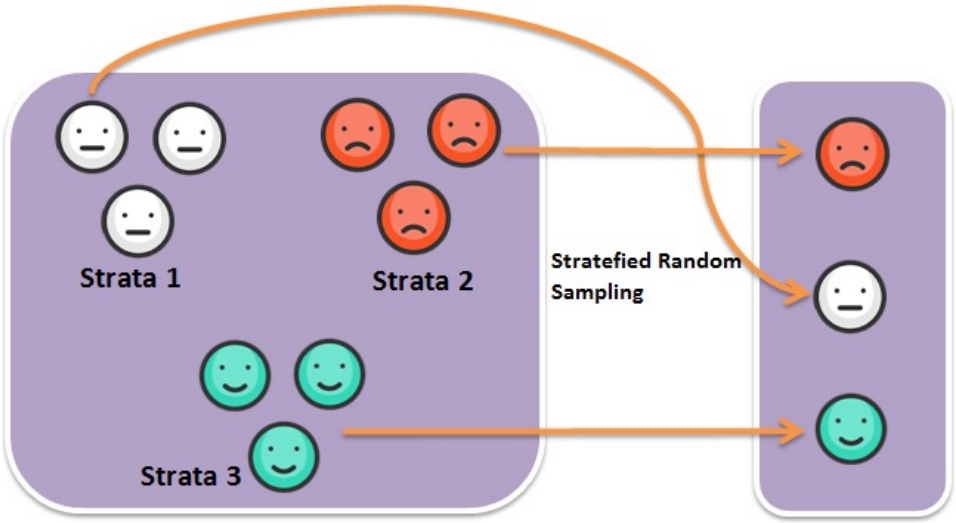
  items are not removed when they are selected for the sample

  (i.e., the same object can be picked up more than once.)

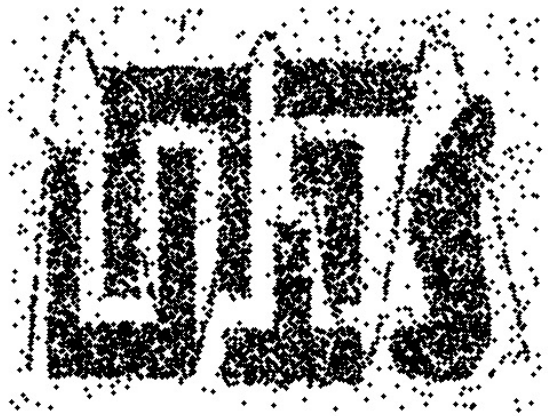


TYPES OF SAMPLING

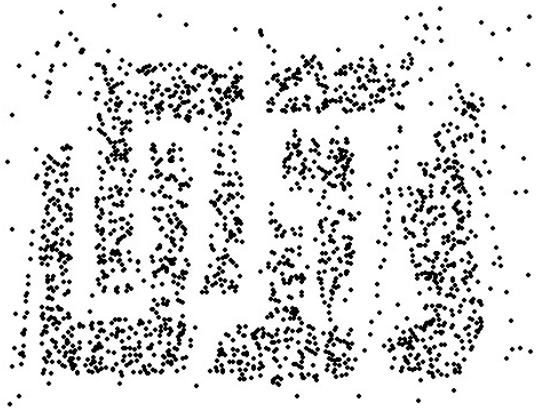
- Stratified sampling
  - Split the data into several groups; then draw random samples from each group.
  - Example: with 0.2% of transactions are fraudulent, how to understand the differences between legitimate and fraudulent credit card transactions?
  - If I select 500 transactions at random? What would happen?
  - Solution: sample 500 legitimate, 500 fraudulent transactions



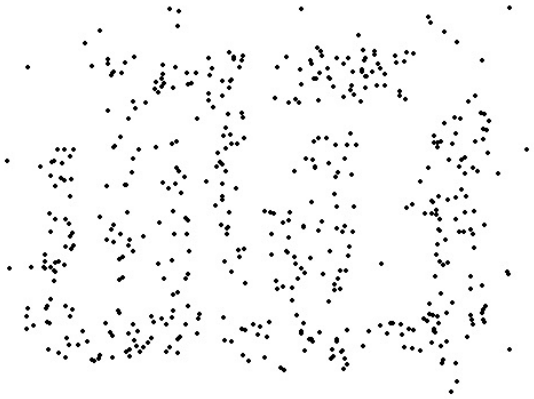
SAMPLE SIZE



8000 points



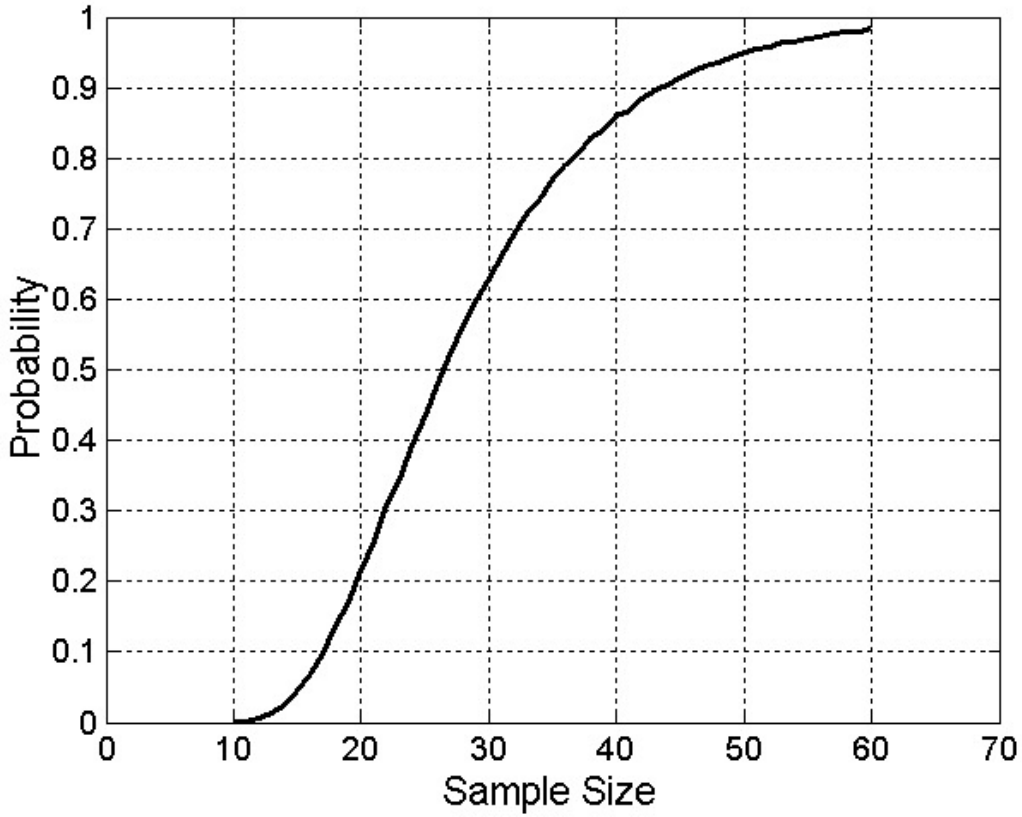
2000 Points



500 Points

SAMPLE SIZE

- **What sample size is necessary to get at least one object from each of 10 fruits.**



STREAMING DATA

Names in the phone book

Videos watched by a student in August

STREAMING DATA

SAMPLING FOR STREAMING DATA

- Task: efficiently return a random sample of 1,000 elements evenly / uniformly distributed from the original stream

Simple solution when the size of stream N is known.

1. generate random integers between 0 and N-1.
2. use the random integer as an index
3. retrieve the elements at those indices

SAMPLING FOR STREAMING DATA

- Task: efficiently return a random sample of 1,000 elements evenly / uniformly distributed from the original stream

What if:

1. the size of the stream N is unknown in advance

2. not enough memory to store the stream in memory

3. only keep a constant amount of integers in memory

RESERVOIR SAMPLING

1	9	2	1	8	…	7													

Notes in class

- Steps:

1. make a reservoir (array) of 1,000 elements (n = 1,000)
2. fill it with the first 1,000 elements in the stream.
3. process the ith element (i>1,000)

   choose to sample the ith item with probability 1,000/i

   at the end of processing that step, the 1,000 element in the array / reservoir are randomly sampled amongst the i elements .

RESERVOIR SAMPLING

- What is the probability of the n-th items to survive for N-n rounds?
  - $\left(1-\frac{1}{n+1}\right)\left(1-\frac{1}{n+2}\right)\cdots\left(1-\frac{1}{N}\right)$