Introduction to Data Analytics

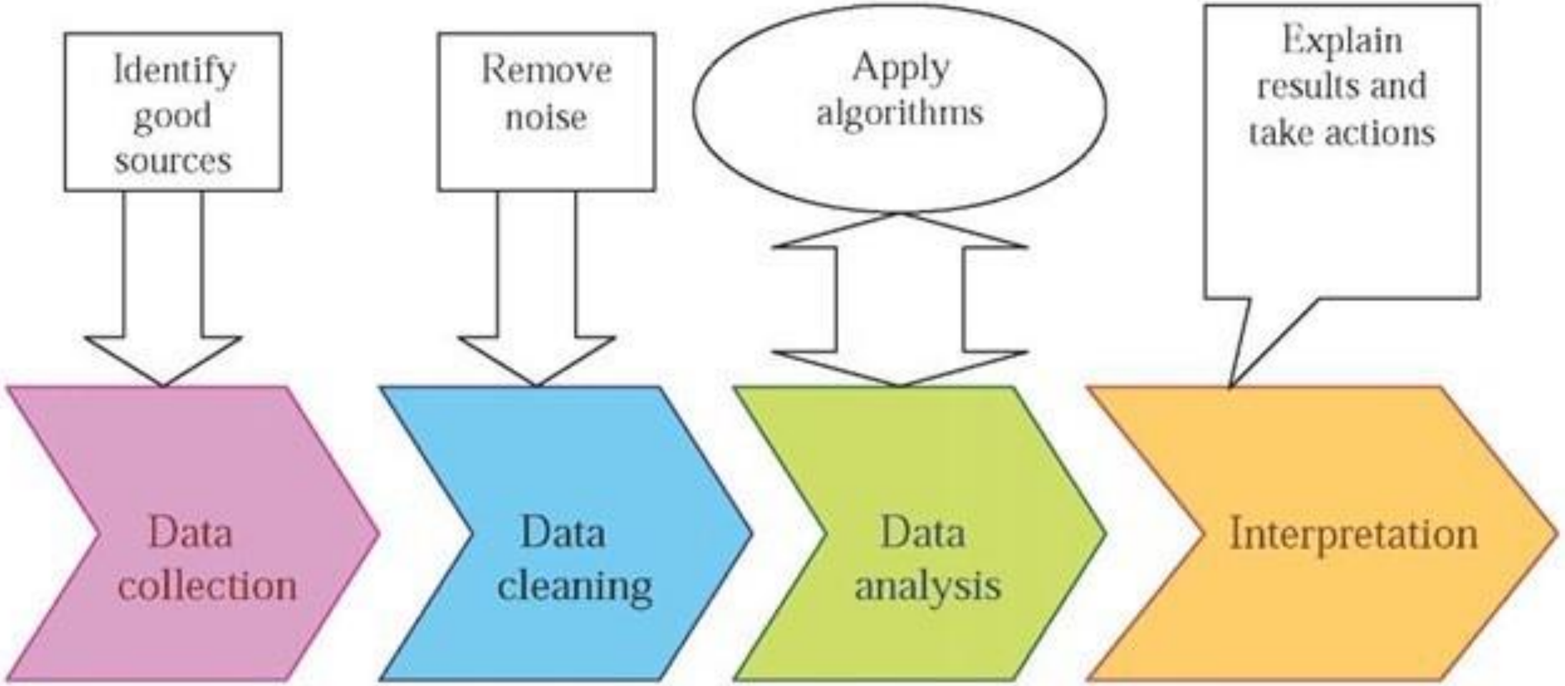
Topics of this session

- Data munging
- Scraping
- Sampling
- Cleaning

Recap of Measures of Data Quality

- **Accuracy:** correct or wrong, accurate or not
- **Completeness:** not recorded, unavailable
- **Consistency:** some modified but some not, dangling
- **Timeliness:** timely update?

Steps involved in Data Analysis



Data Munging

- Required for improving the quality of gathered data
- Involves **cleaning** and **transformation** of messy data available with us
- Also referred as **Data Wrangling**

Data Munging

Before Data Munging

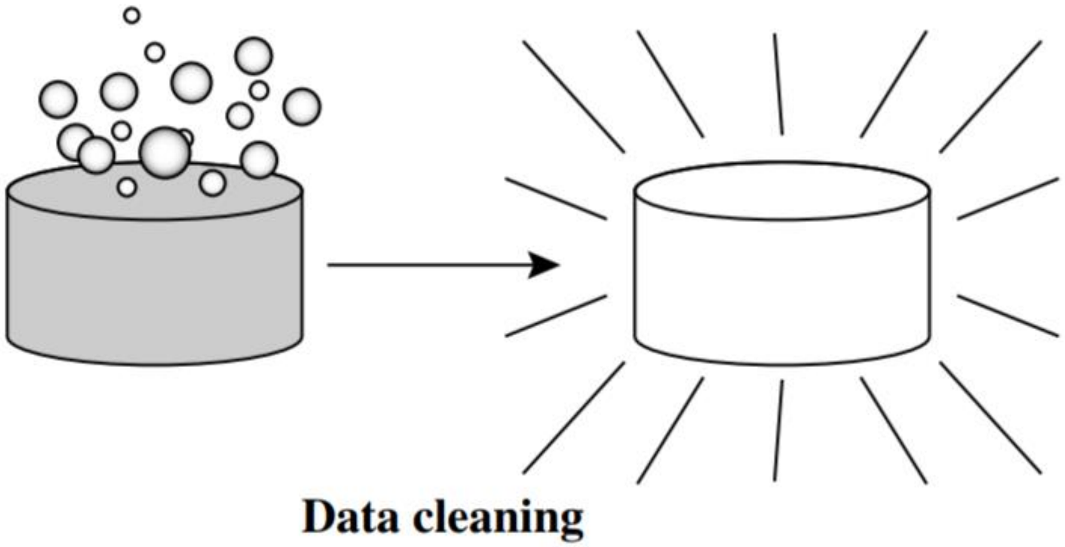
After Data Munging

Reason for noise in data

- Data in the Real World Is Dirty

- Lots of potentially incorrect data
  - Faulty instruments
  - Human or computer error
  - Transmission error

Data Cleaning



Some examples for noisy data

- **Incomplete:** lacking attribute values, lacking certain attribute values of interest, or containing only aggregate data

  Ex: Occupation=" " (missing data)

- **Noisy:** containing noise, errors, or outliers

  Ex: Salary="−10" (an error)

Some examples for noisy data

- **Inconsistent:** containing discrepancies in codes or names, discrepancy between duplicate records

Ex:

1. Age="42", Birthday="03/07/2010"
2. Was rating "1, 2, 3", now rating "A, B, C"

- **Intentional:** disguised missing data

  Ex: Jan. 1 as everyone's birthday?

Data Cleaning

- Fill in missing values
- Smooth noisy data
- Identify or remove outliers
- Resolve inconsistencies

Data Cleaning (Dealing with Missing Values)

- Ignore the tuple
- Fill in the missing value manually
- Fill in it automatically with
  - a global constant
  - the attribute mean
  - the attribute mean for all samples belonging to the same class
  - the most probable value

Data Cleaning (Dealing with Noise)

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions

Data Cleaning (Removing Outliers)

- Clustering
  - detect and remove outliers

Data Cleaning (Dealing with inconsistencies)

- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Transformation

- Transformation is mapping the Data to a new space

Ex:

  - Fourier Transform
  - Wavelet Transform

Data Transformation

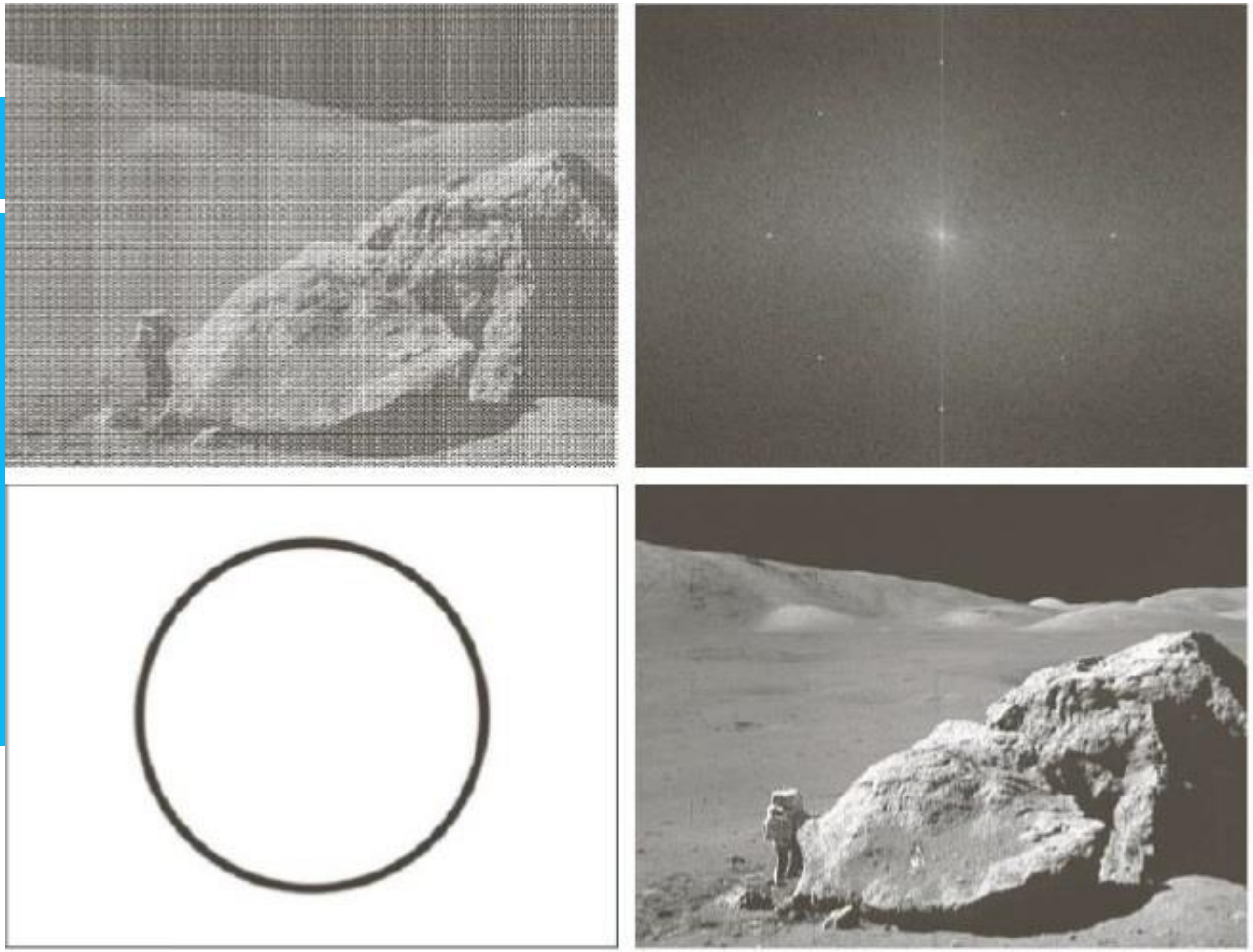
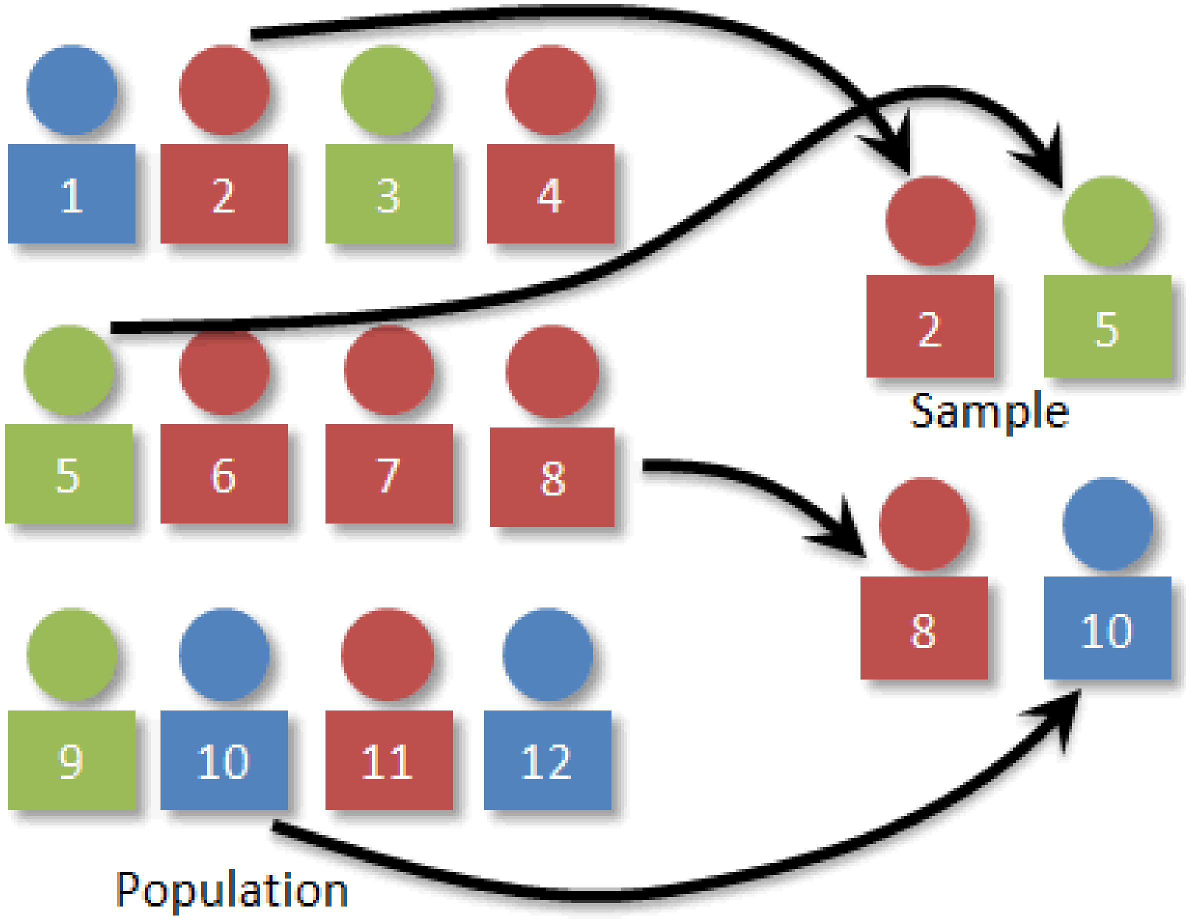


FIGURE 2.40 (a) Image corrupted by sinusoidal interference. (b) Magnitude of the Fourier transform showing the bursts of energy responsible for the interference. (c) Mask used to eliminate the energy bursts. (d) Result of computing the inverse of the modified Fourier transform. (Original image courtesy of NASA.)

Data Sampling

Sampling: obtaining a small sample *s* to represent the whole data set *N*



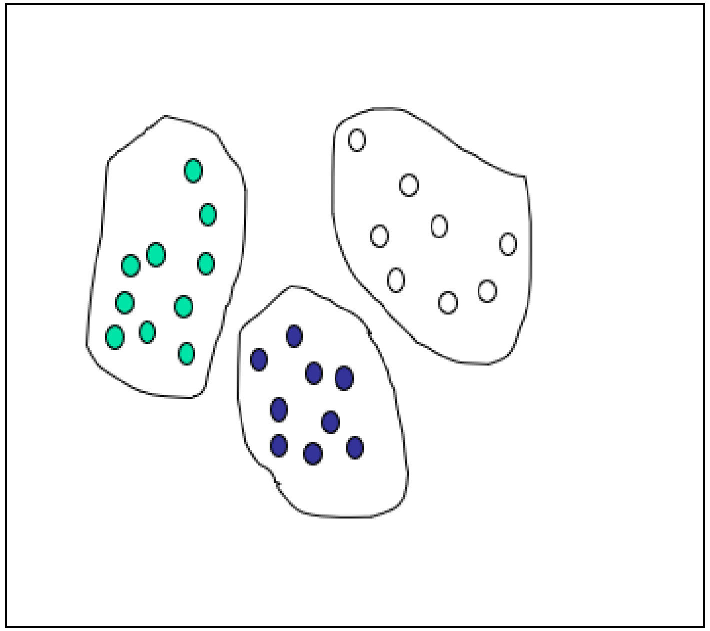
Types of Sampling

- **Simple random sampling:**
  - There is an equal probability of selecting any particular item
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

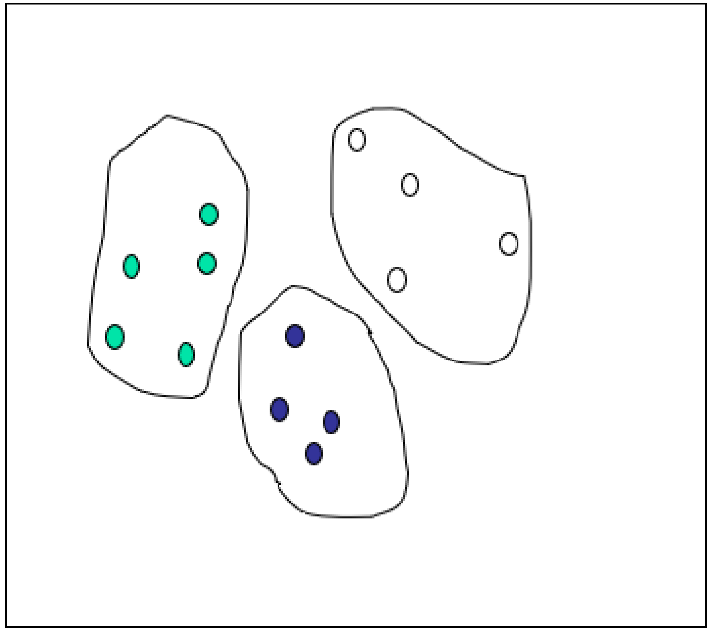
Types of Sampling

Stratified Sampling

Raw Data



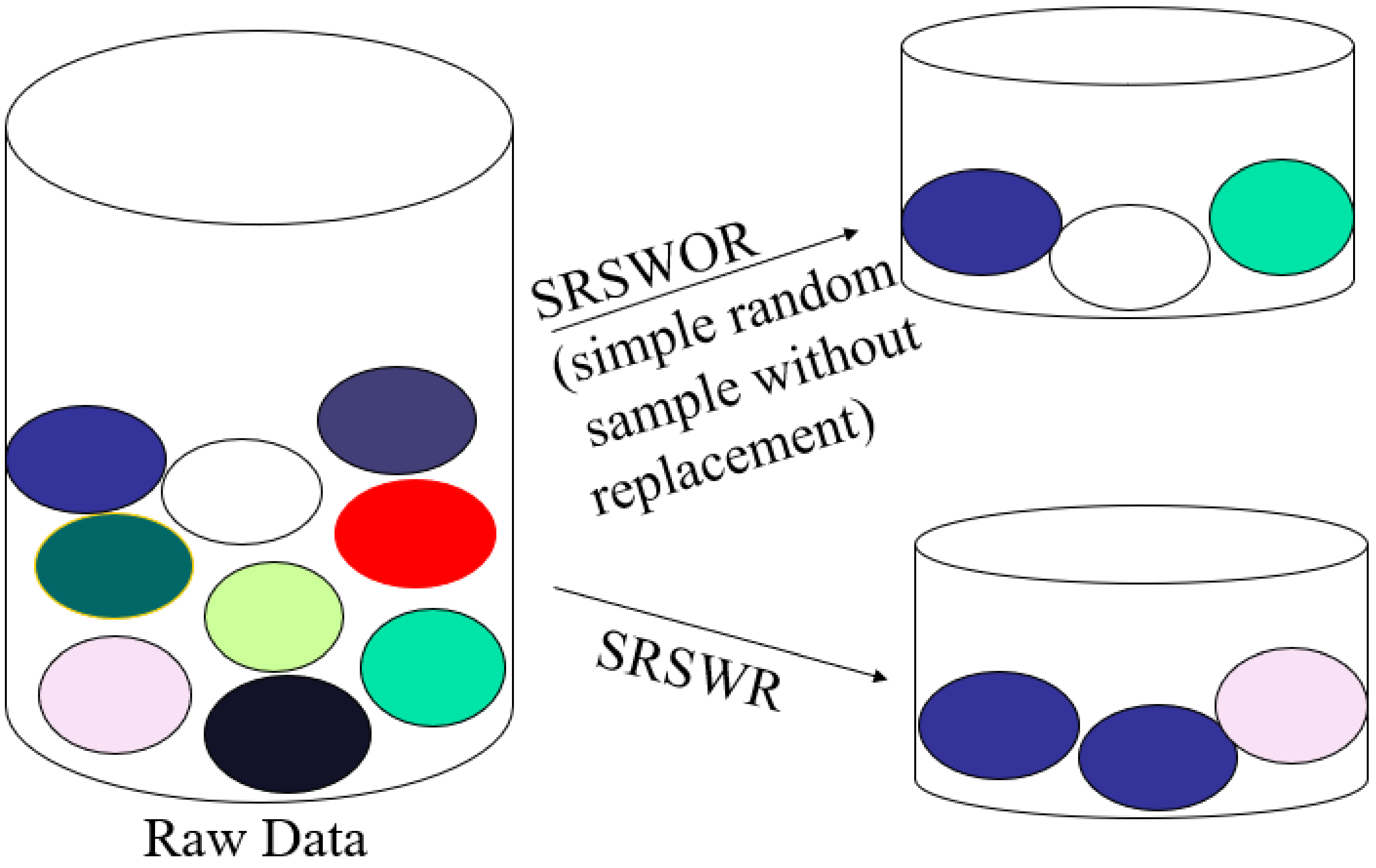
Cluster/Stratified Sample



Types of Sampling

- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population

Types of Sampling


SRSWOR
(simple random
sample without
replacement)
SRSWR
Raw Data