STATISTICS FOR DATA SCIENCE

Sampling Methods

D. Uma
Department of Computer Science and Engineering
umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

Sampling Methods

D. Uma
Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

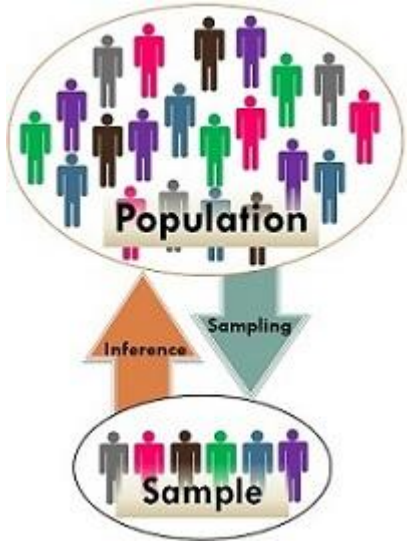
Recap of the Last Session....

Data

Population

Sample

Sampling or Sampling Process



Source: 365datascience.com

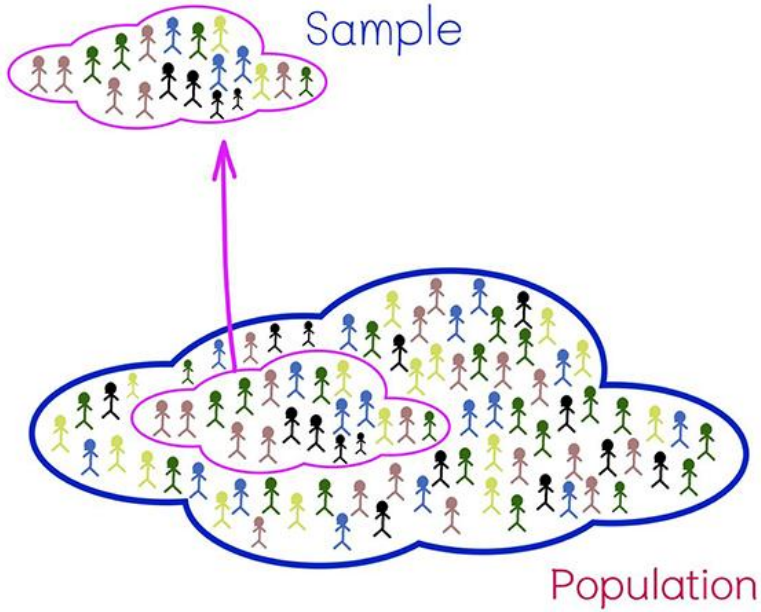
STATISTICS FOR DATA SCIENCE

What are sampling methods?

In a statistical study, sampling methods refer to how we select members from the population to be in the study.

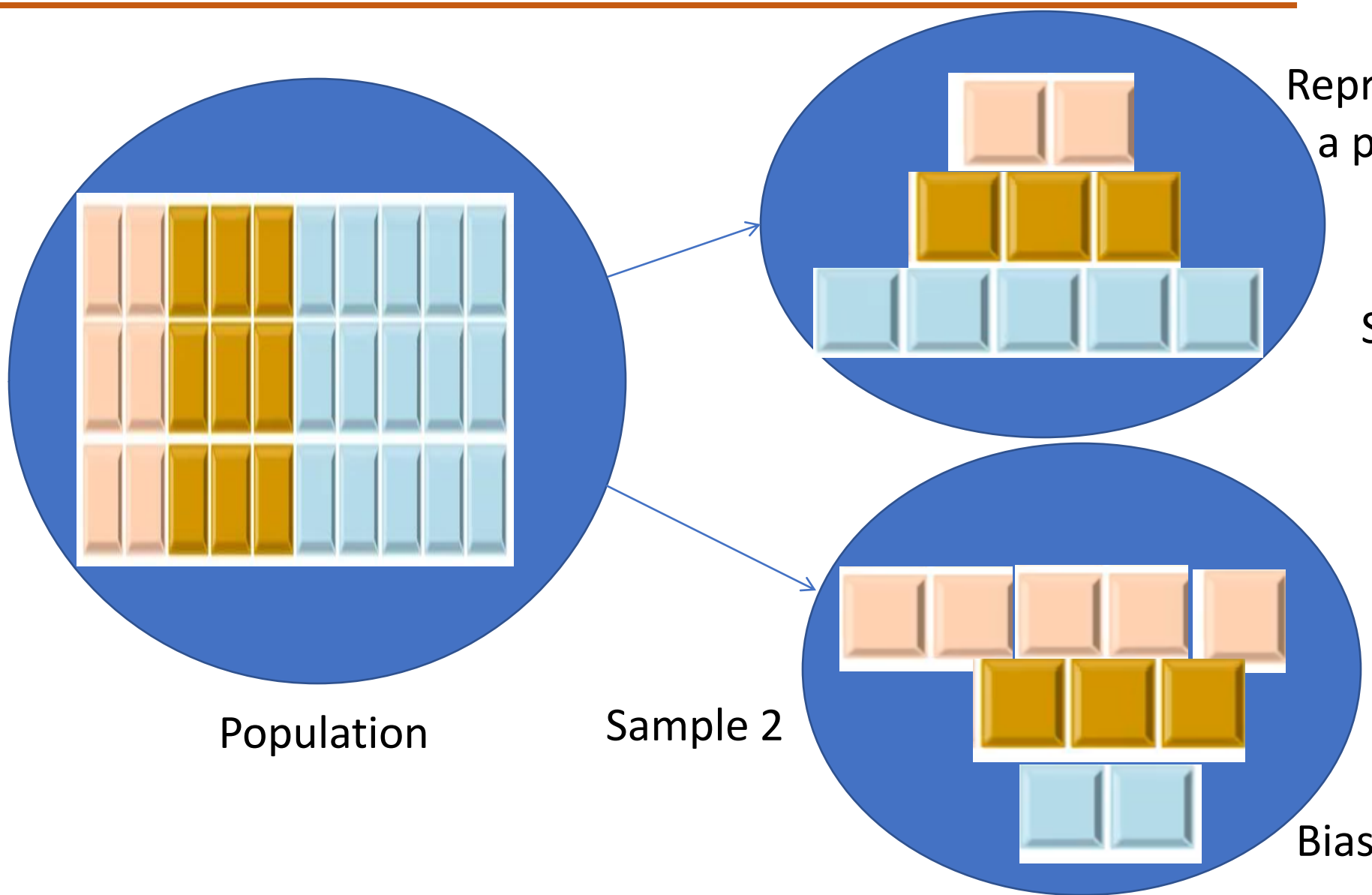
If a sample isn't randomly selected, it will probably be biased in some way and the data may not be representative of the population.

There are many ways to select a sample—some good and some bad.



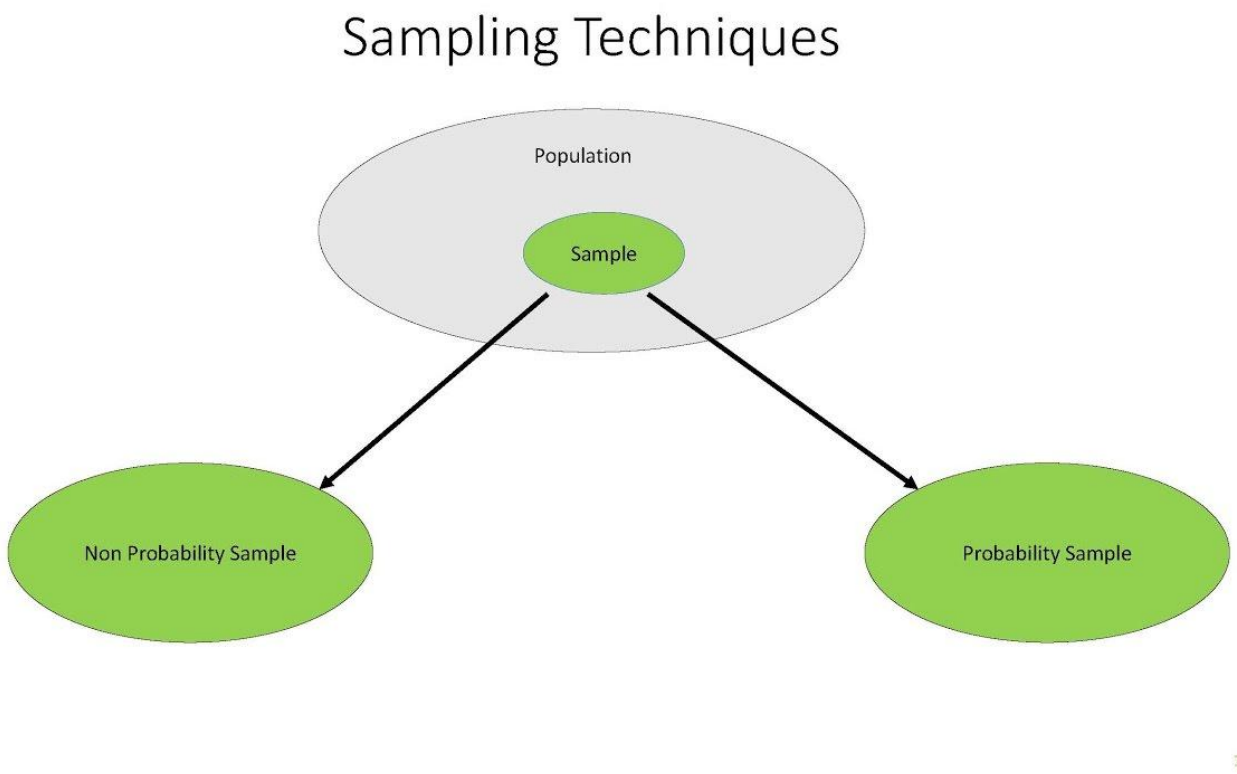
Source: blog.masterofproject.com , analytics-magazine.org

Representative and biased samples



STATISTICS FOR DATA SCIENCE

Types of sampling



Source: www.youtube.com

STATISTICS FOR DATA SCIENCE

Characteristics of a Probability Sampling

Every unit of the population has the same probability of being included in the sample.

A chance mechanism is used in the selection process.

Eliminates bias in the selection process.

This is also known as Random sampling.

Source: quotesgram.com

Characteristics of a Non-Probability Sampling

Every unit of the population does not have the same probability of being included in the sample.

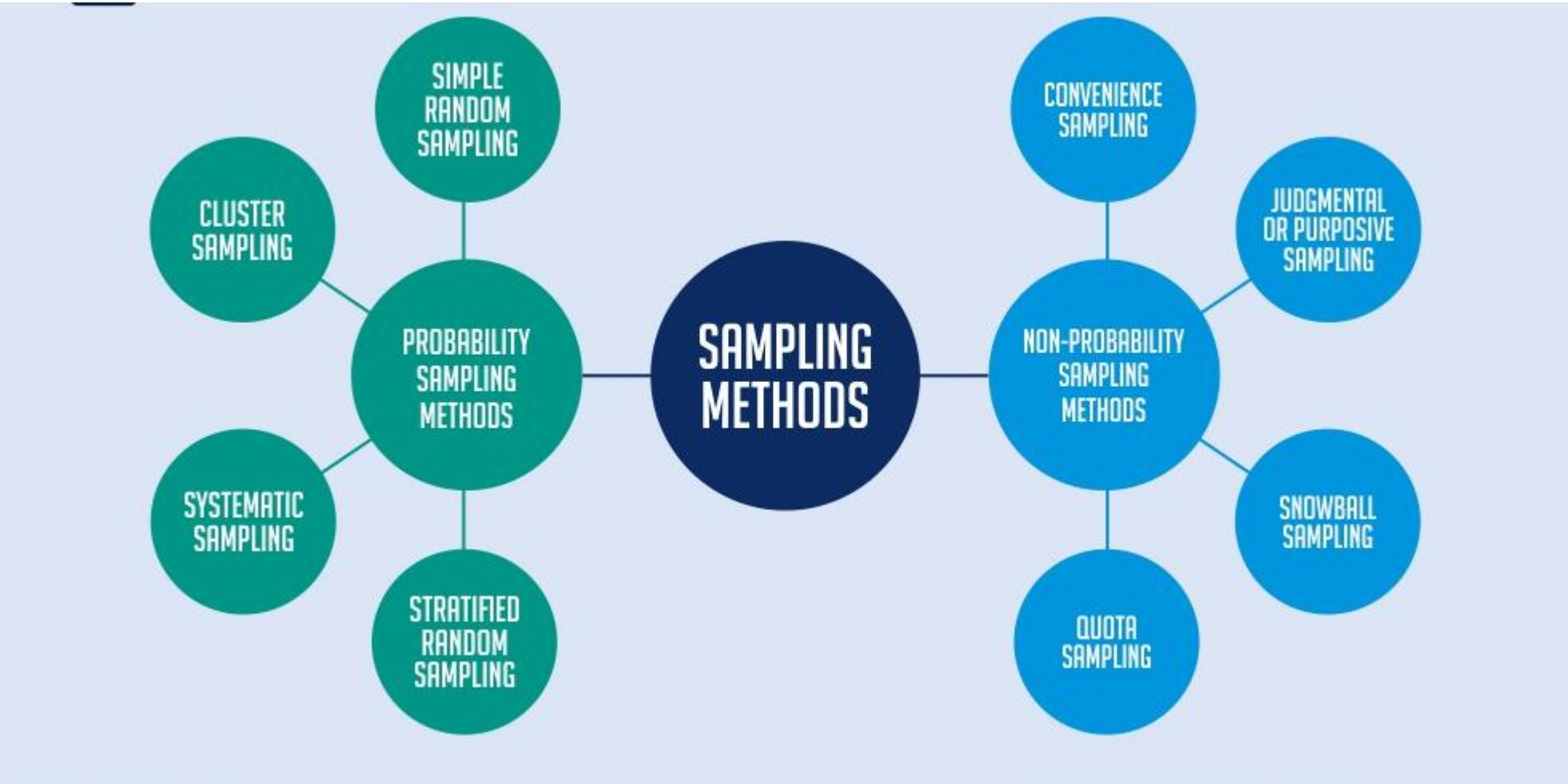
It opens up the selection bias.

Not an appropriate data collection methods for most of the statistical analysis.

Also known as non-random sampling.

STATISTICS FOR DATA SCIENCE

Types of Sampling Methods



Source: www.questionpro.com

What type of Sampling?

A teacher puts students' names in a hat and chooses without looking to get a sample of students.

Every member has an equal chance of being included in the sample.

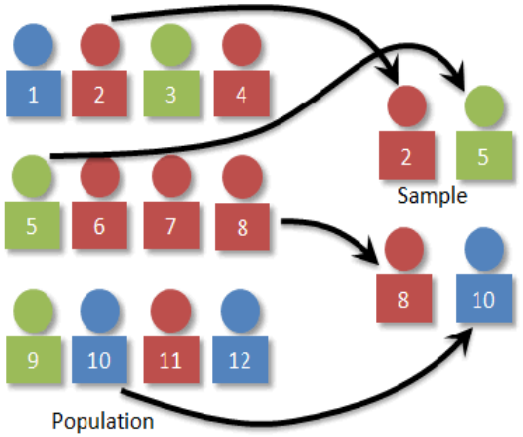
Why it's good: Random samples are usually fairly representative since they don't favor certain members.

Source: ccelearn.csus.edu

STATISTICS FOR DATA SCIENCE

Simple Random Sampling

Purpose : Random and Representative Sample from a Larger Population.
When to Use : Best to use when population is small and
produces better representative.



Key Thing: All members of a population has an equal probability.

How : Assign numbers to members of population & select randomly.
Small Population: Manual Lottery Method.
Larger Population : System Generated Number.

Advantages: Easier, Better representativeness, Low sampling error,
No prior information is required.

Disadvantages: Biased sample, Does not reflect proportionate representation, Difficult when population is larger.

Source Image: www.quora.com

STATISTICS FOR DATA SCIENCE

What type of Sampling?

A student council surveys 50 students by getting random samples of 25 juniors and 25 seniors.

Why it's good: This sample guarantees that members from each group will be represented in the sample, so this sampling method is good when we want some members from every group.

Source Image: slideshare.net

STATISTICS FOR DATA SCIENCE

Stratified Random Sampling

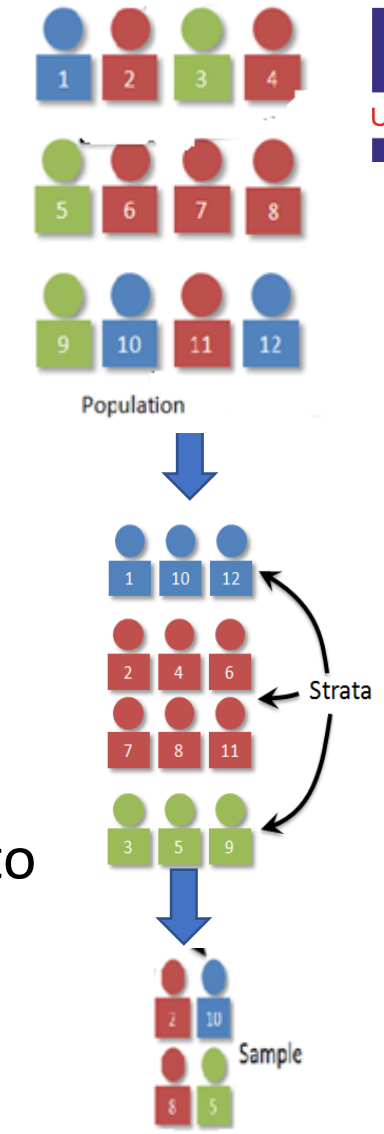
Purpose : Unbiased Random Sample from a Larger Population.
When to Use : Population proportion should be reflected in sample.

Key Thing: Sample proportion is same as Population proportion, Strata is homogeneous.

How : Divide the population into Strata or Groups.
Criteria : Gender, Hair Color, Eye Color, Salary, Designation, Age etc.
Selection : Simple Random Sampling approach to Strata.

Advantages: Enhancement of representativeness of a sample, Easy to carry out, Higher statistical efficiency.

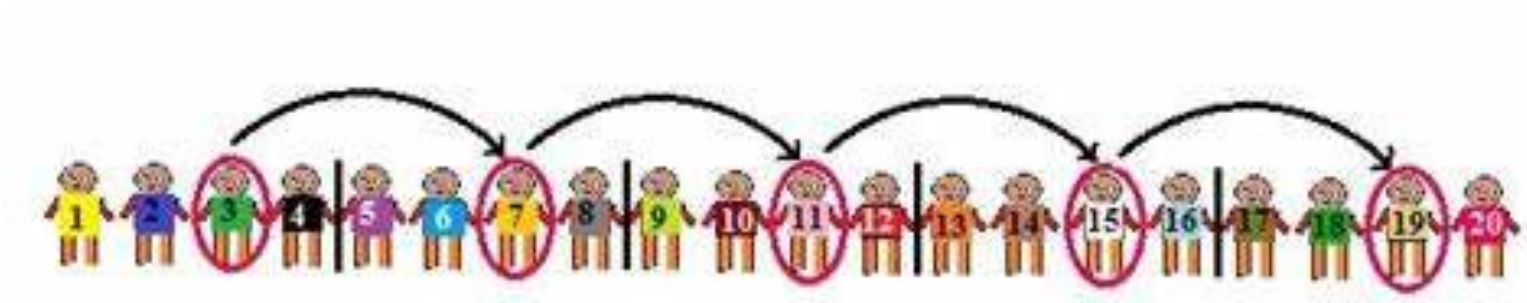
Disadvantages: Classification error, time consuming, expensive.



Source Image:www.quora.com

What type of Sampling?

Example—A principal takes an alphabetized list of student names and picks a random starting point. Every 4th student is selected to take a survey.



Why it's good: This sample includes every 4th person. But, first person is chosen randomly.

Source Image: eduation-savvy.blogspot.com

STATISTICS FOR DATA SCIENCE

Systematic Sampling

When to Use : When project budget is tight and less time to complete.

Key Thing: Find the kth value to select every kth member.

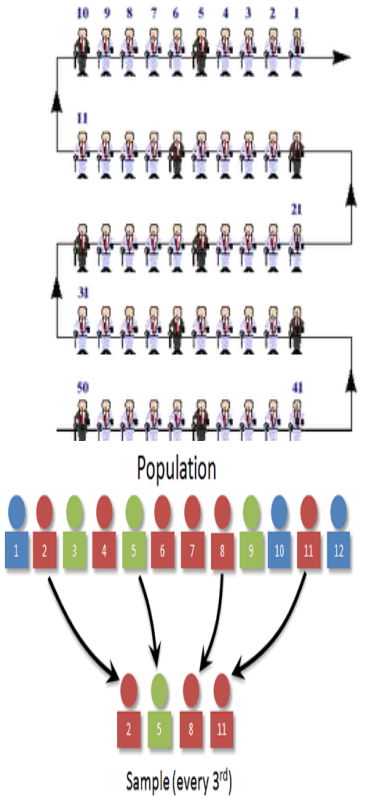
$$k = N / n$$

How: Assign numbers to each population member.

Selection : Randomly select first person and then select every kth person.

Advantages: Easy to select, Sample evenly spread over entire reference population, cost effective.

Disadvantages: Sample may be biased, Each element does not have equal chance, Ignorance of all elements between two kth element.



Source Images: slideshare.net, www.quora.com

What type of Sampling?

An airline company wants to survey its customers one day, so they randomly select 5 flights that day and survey every passenger on those flights.

Why it's good: This sample gets every member from some of the groups, so it's good when each group reflects the population as a whole.

Source Image: www.etihad.com

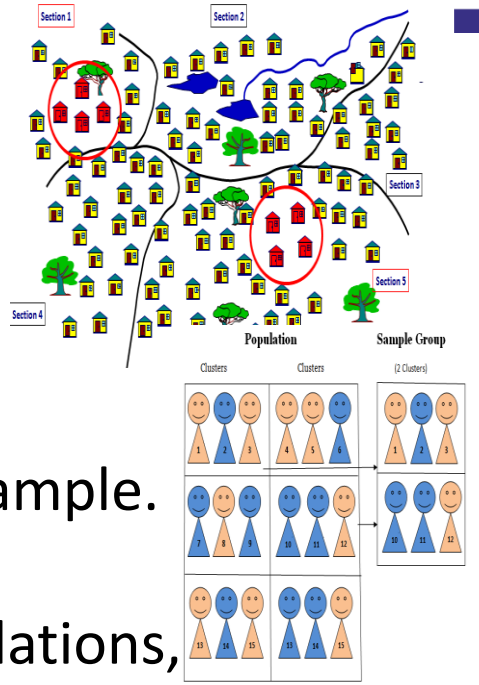
STATISTICS FOR DATA SCIENCE

Cluster Sampling

When to Use : When population is already broken up into groups(clusters).

Key Thing: Heterogeneous members in each group.

How: Population is divided into non-overlapping areas(clusters).
Each cluster is a miniature or microcosm of a population.



Selection : Clusters are selected randomly and all elements are included or elements are chosen using simple random sample.

Advantages: More convenient for geographically dispersed populations, Less travel cost, Simplified administration of the survey.

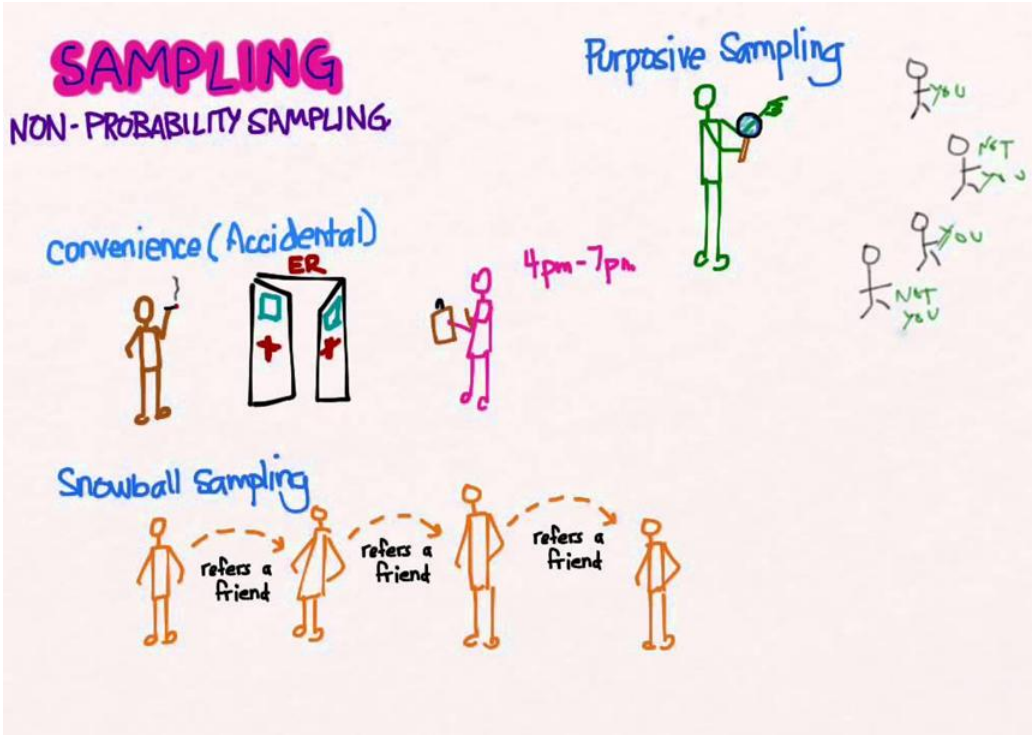
Disadvantages: Statistically less efficient, Sampling error is higher, problems are higher than simple random sampling.

Source Images: slideshare.net,

STATISTICS FOR DATA SCIENCE

Non-random Sampling Methods

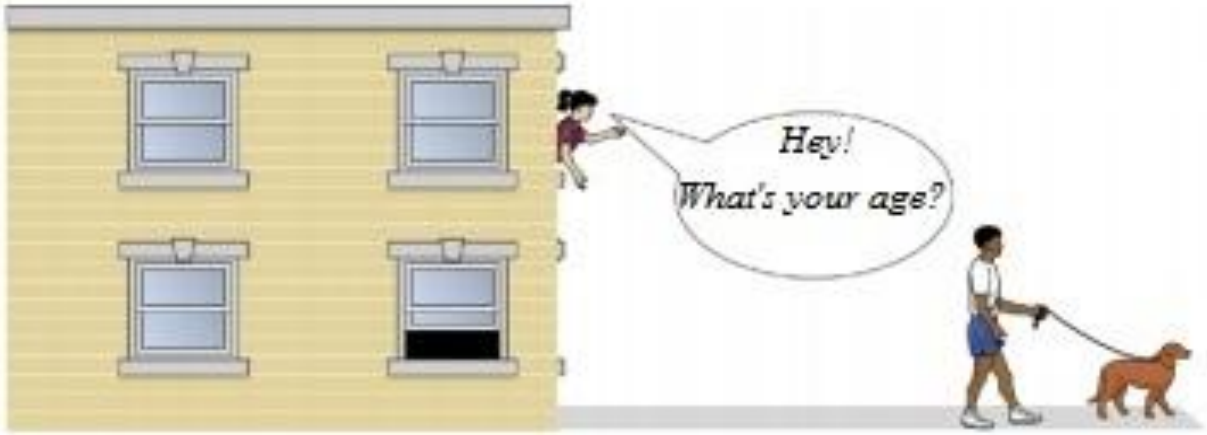
- Convenience
- Judgmental/Purposive
- Quota
- Snowball



Source Image: www.youtube.com

What type of Sampling?

A researcher polls people as they walk by on the street.



Bad ways to sample: The researcher chooses a sample that is readily available in some non-random way.

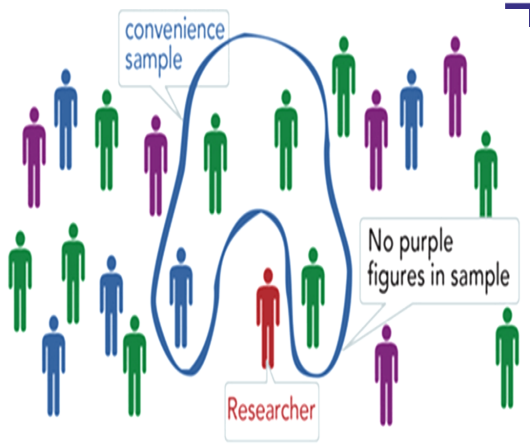
Why it's probably biased: The location and time of day and other factors may produce a biased sample of people.

Source Image: matcmath.org

STATISTICS FOR DATA SCIENCE

Convenience Sampling

When to Use : When population is not clearly defined or sampling unit is not clear or complete source list is not available.



Key Thing: Subjects for a study are easily available within the proximity of the researcher.

How: It is done at the “convenience” of the researcher.

Selection : Whichever individuals are easiest to reach and they are convenient.

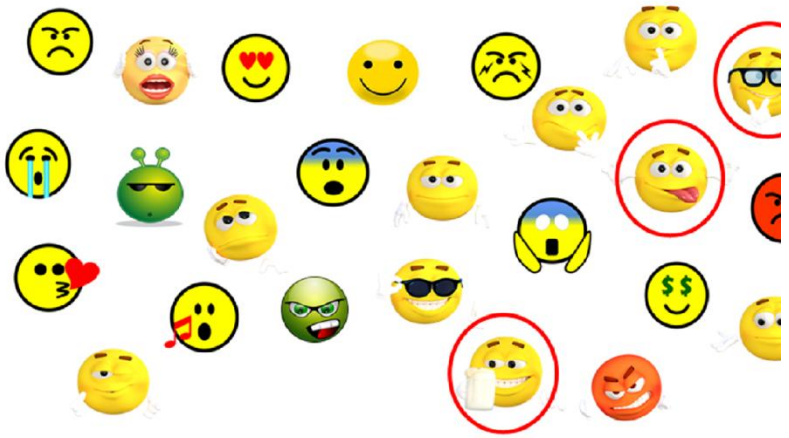
Advantages: Ease of availability, saves time, money, Useful in pilot study.

Disadvantages: Biased sample, Sampling errors, Results can’t be generalized.

Source Images: slideshare.net, www.quora.com

What type of Sampling?

People who owns the qualities expected by the researcher.



The researcher uses their own knowledge and judgement to include or exclude people in the sample.

Source Image: researchforevidence.fhi360.org

STATISTICS FOR DATA SCIENCE

Judgmental Sampling

When to Use : The sample is selected based upon judgment.
Also, the researcher must be confident that the chosen sample is truly representative of the entire population.

Key Thing: The researcher selects a sample based on experience or knowledge of the group to be sampled.

How: Basis of the researcher's knowledge and judgment.

Selection : People who owns the qualities expected by the researcher.

Advantages: Consumes minimum time, real-time results.

Disadvantages: Selection bias, Selection of proper sample size.

Source Images: slideshare.net, www.quora.com

STATISTICS FOR DATA SCIENCE

What type of Sampling?

An interviewer may be told to sample 200 females and 300 males between the age of 45 and 50.

Sample the people till the quota is met.

Source Image: climatwecommunication.yale.edu

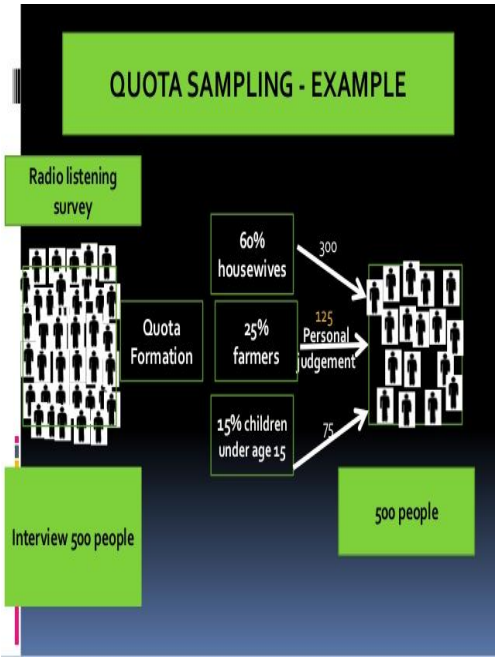
STATISTICS FOR DATA SCIENCE

Quota Sampling

When to Use : If a study aims to investigate a trait or a characteristic of a certain subgroup, this type of sampling is the ideal technique.

Key Thing: Sample elements are selected until the quota controls are satisfied.

How: First identify the stratums and their proportions as they are represented in the population.



Advantages: When the respondent refuses to cooperate, he may be replaced by another person who is ready to furnish information, Less expensive, speedy, Convenience in execution.

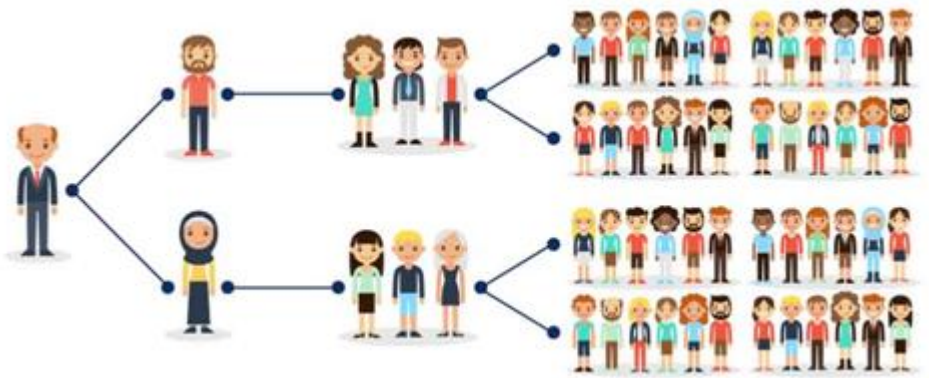
Disadvantages: Bias, lack of valuable data.

Note: It is the non-probability equivalent of stratified sampling.

Source Image: www.slideshare.net

What type of Sampling?

Survey the people who are involved in illegal activities.



Identify an initial object.

Source Image: www.questionpro.com

STATISTICS FOR DATA SCIENCE

Snowball Sampling

When to Use : When the desired sample characteristic is rare.

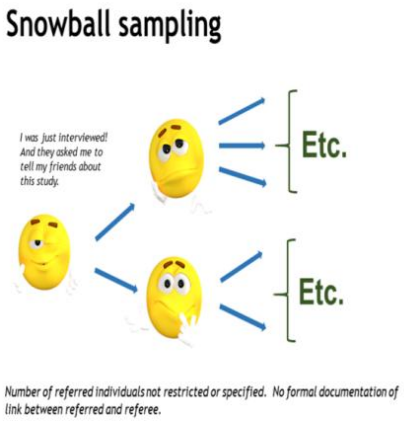
Key Thing: Research starts with a key person and introduce the next one to become a chain. It may be extremely difficult or cost prohibitive to locate respondents in these situations.

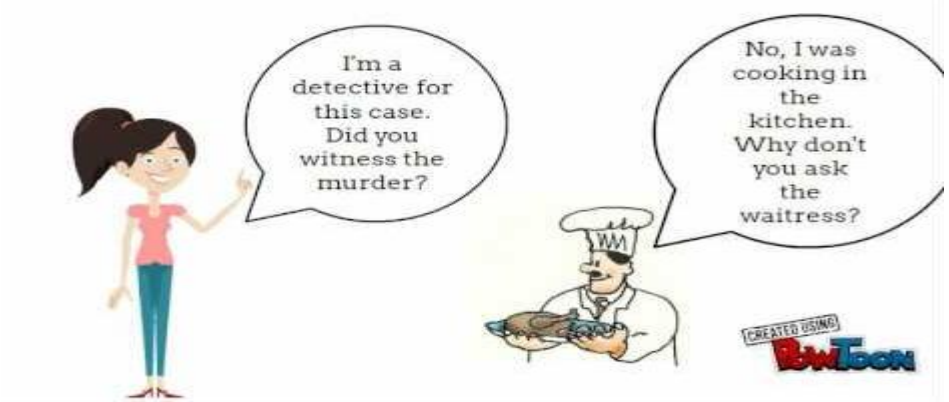
How: Identify an initial subject and ask these people to identify others.

Selection : This technique relies on referrals from initial subjects to generate additional subjects.

Advantage: Lowers search cost.

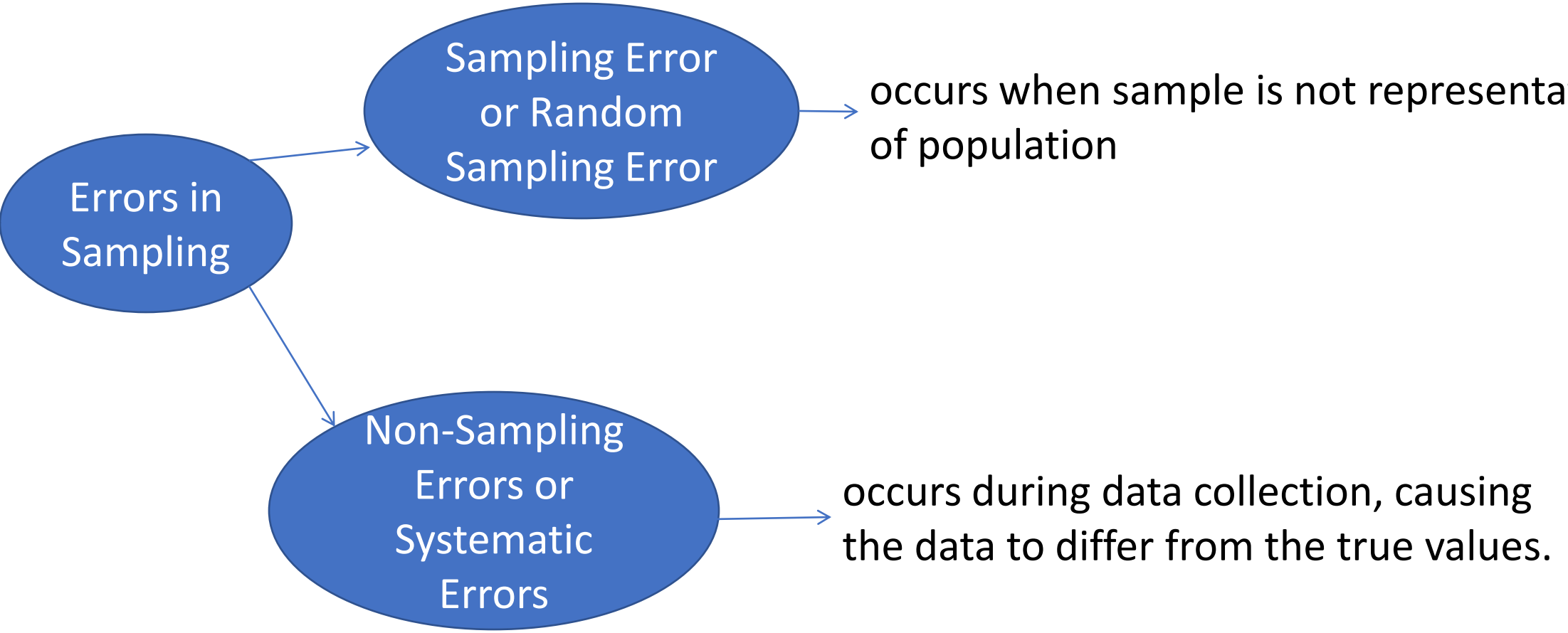
Disadvantage: Introduces bias.





Source Image: researchforevidence.fhi360.org

Errors in Sampling



STATISTICS FOR DATA SCIENCE

Errors in Sampling

- Data from nonrandom samples are not appropriate for analysis by inferential statistical methods.

- **Sampling Error** occurs when the sample is not representative of the population.

- **Non-sampling Errors**
  - Missing Data, Recording, Data Entry, and Analysis Errors
  - Poorly conceived concepts , unclear definitions, and defective questionnaires
  - Response errors occur when people so not know, will not say, or overstate in their answers

THANK YOU

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335