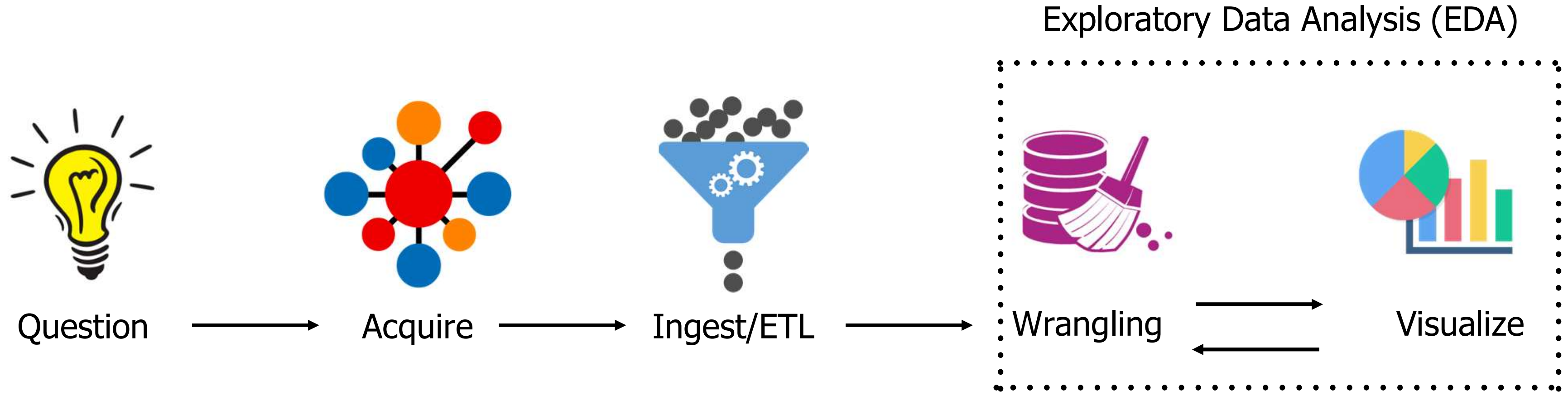
DICE
ANALYTICS

PYTHON & DATA SCIENCE WORKSHOP

https://www.facebook.com/diceanalytics/

1)

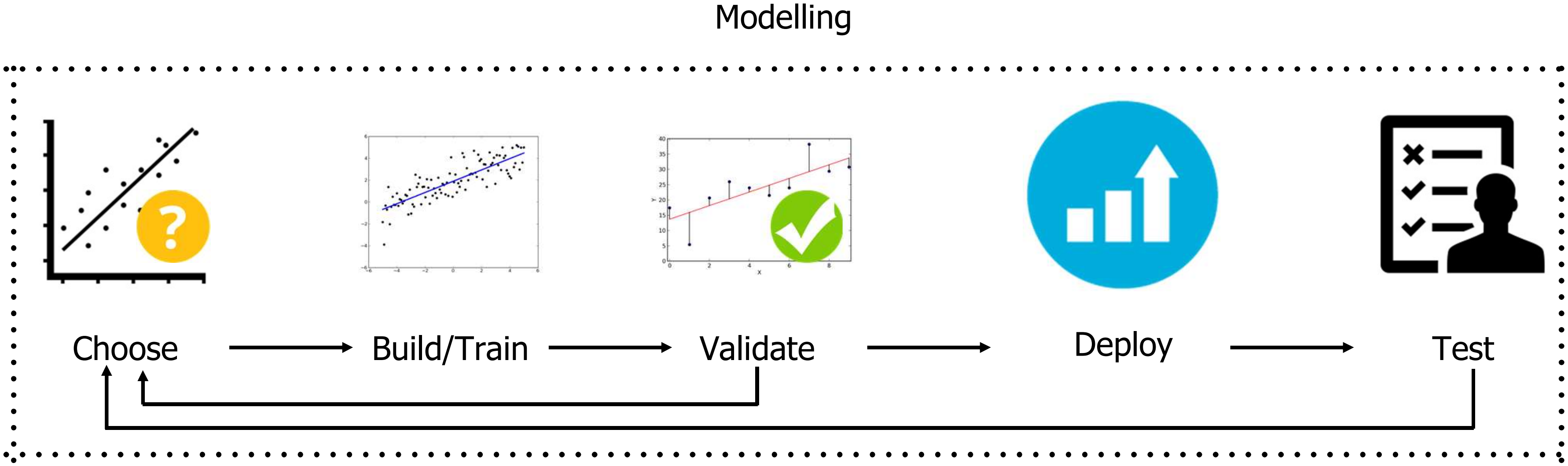


Exploratory Data Analysis (EDA)

Question → Acquire → Ingest/ETL → Wrangling ⇄ Visualize

2)

Modelling

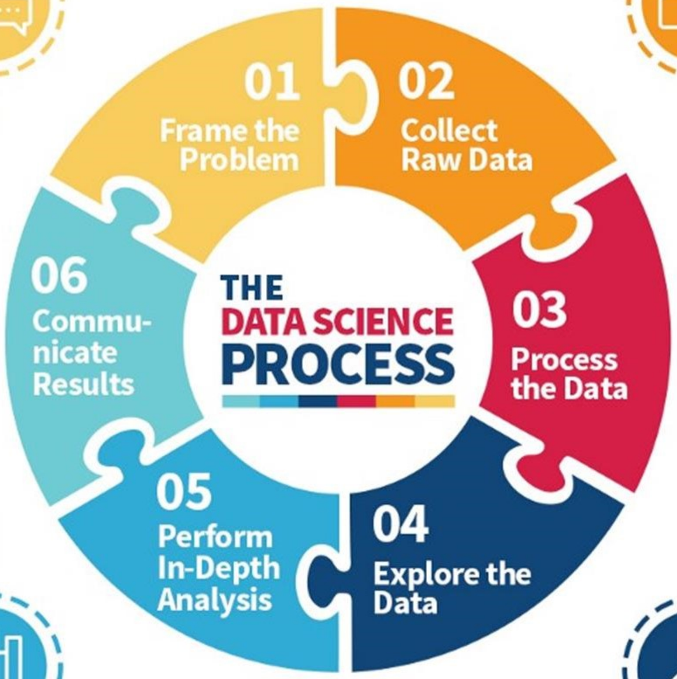


Choose → Build/Train → Validate → Deploy → Test

3)

DICE ANALYTICS

Data Science Deconstructed



Ask a Lot of Questions

- Translate an ambiguous request into a concrete, well-defined problem
- Identify business priorities & strategy decisions that will influence your work

Identify All Available Datasets

- Web, internal/ external databases, etc.

Extract Data Into Usable Format

- .csv, .json, .xml, etc.

Examine Data at a High-Level

- Understand every column; identify errors, missing values & corrupt records

Clean the data

- Throw away, replace, and/or filter corrupt /error prone / missing values

Play Around With the Data

- Split, segment, & plot the data in different ways

Identify Patterns & Extract Features

- Use statistics to identify & test significant variables

Create a Predictive Model

- Use feature vectors from step #4

Evaluate & Refine Model

- Perhaps return to step #2, 3, or 4

Identify Business Insights

- Return back to the business problem

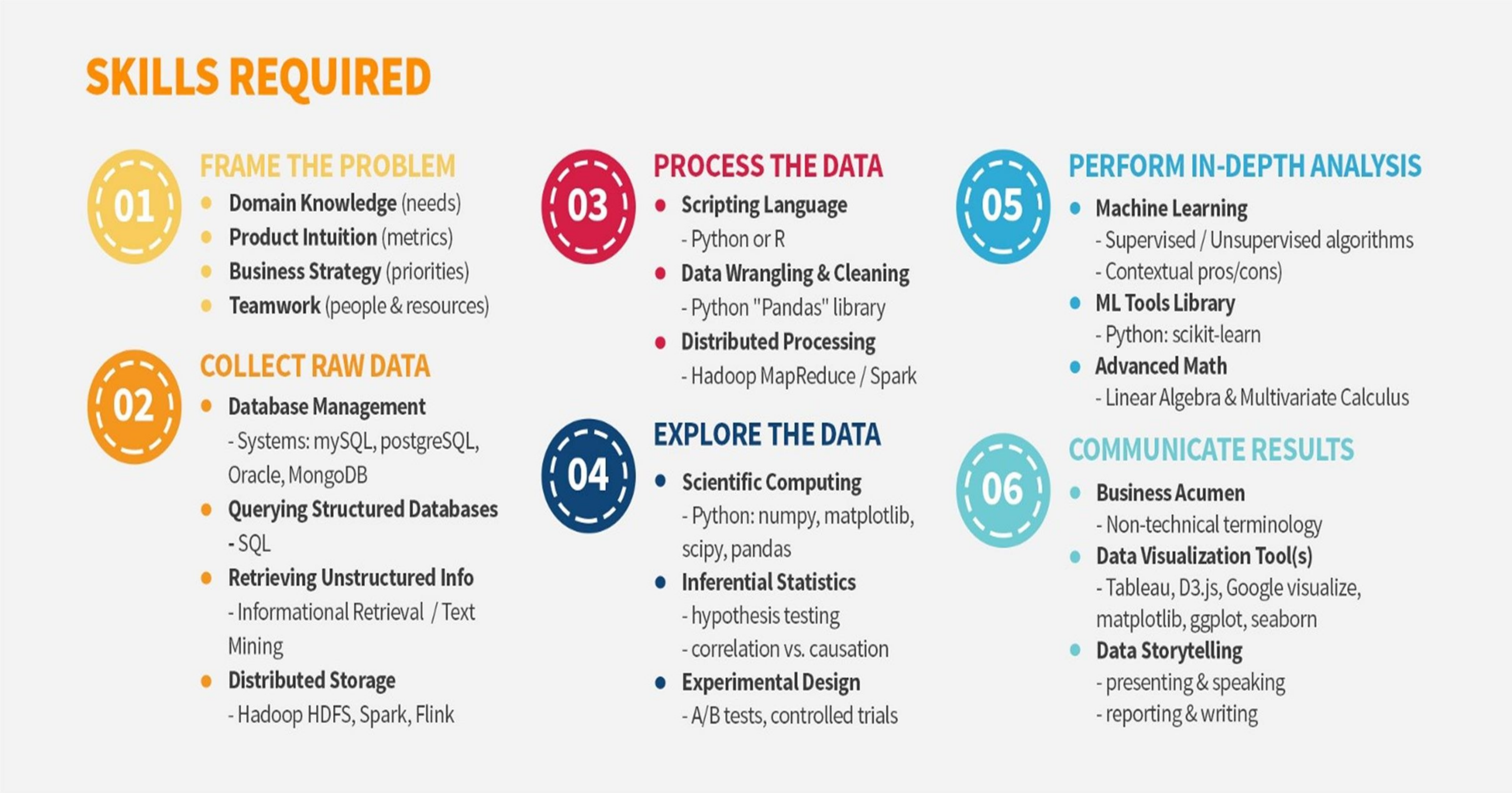
Visualize Your Findings

- Keep it simple & priority-driven

Tell a Clear & Actionable Story

- Effectively communicate to non-technical audiences

DICE ANALYTICS



SKILLS REQUIRED

01 FRAME THE PROBLEM

- **Domain Knowledge** (needs)
- **Product Intuition** (metrics)
- **Business Strategy** (priorities)
- **Teamwork** (people & resources)

02 COLLECT RAW DATA

- **Database Management**
  - Systems: mySQL, postgreSQL, Oracle, MongoDB
- **Querying Structured Databases**
  - SQL
- **Retrieving Unstructured Info**
  - Informational Retrieval / Text Mining
- **Distributed Storage**
  - Hadoop HDFS, Spark, Flink

03 PROCESS THE DATA

- **Scripting Language**
  - Python or R
- **Data Wrangling & Cleaning**
  - Python "Pandas" library
- **Distributed Processing**
  - Hadoop MapReduce / Spark

04 EXPLORE THE DATA

- **Scientific Computing**
  - Python: numpy, matplotlib, scipy, pandas
- **Inferential Statistics**
  - hypothesis testing
  - correlation vs. causation
- **Experimental Design**
  - A/B tests, controlled trials

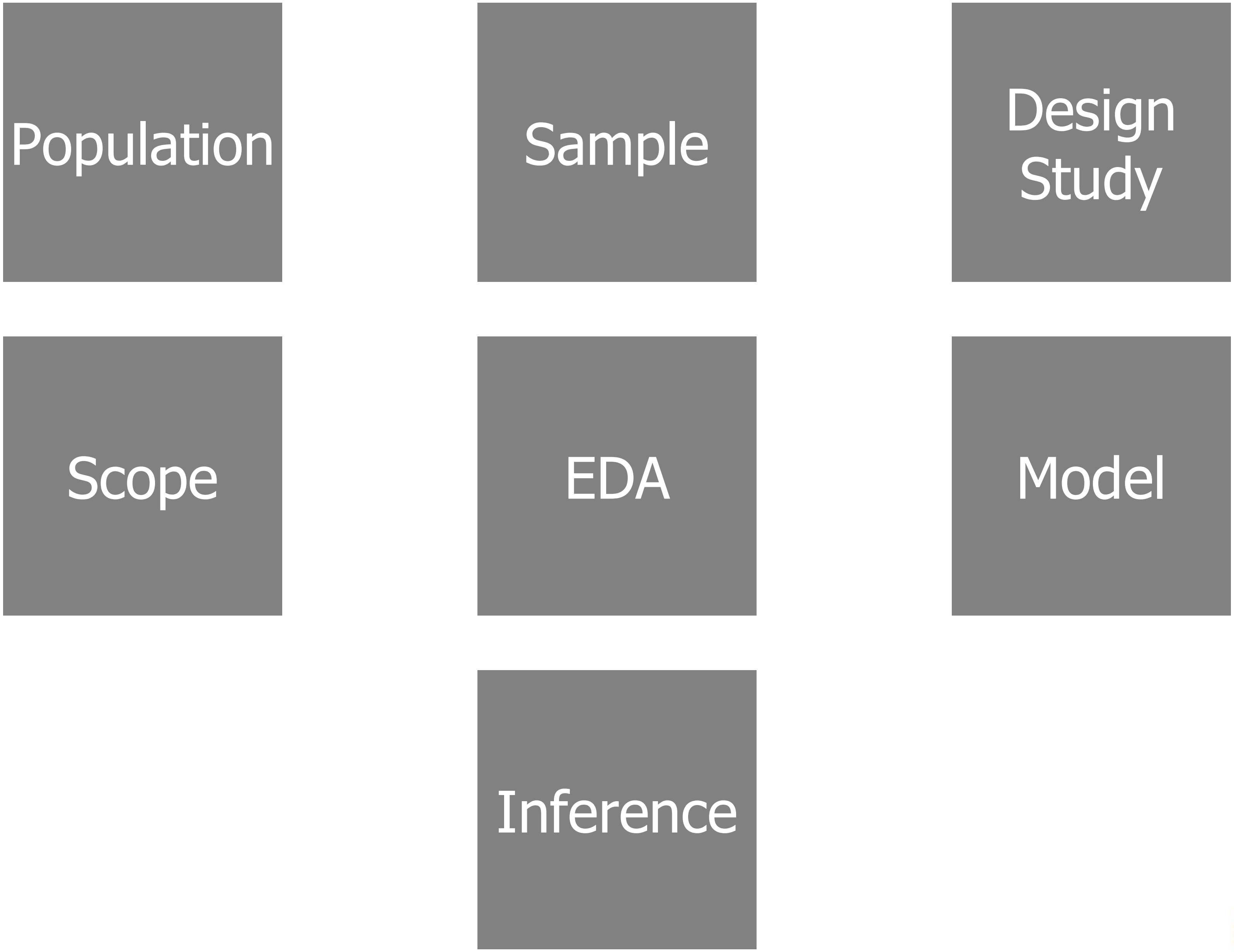
05 PERFORM IN-DEPTH ANALYSIS

- **Machine Learning**
  - Supervised / Unsupervised algorithms
  - Contextual pros/cons)
- **ML Tools Library**
  - Python: scikit-learn
- **Advanced Math**
  - Linear Algebra & Multivariate Calculus

06 COMMUNICATE RESULTS

- **Business Acumen**
  - Non-technical terminology
- **Data Visualization Tool(s)**
  - Tableau, D3.js, Google visualize, matplotlib, ggplot, seaborn
- **Data Storytelling**
  - presenting & speaking
  - reporting & writing

Statistical Flow



Population

Sample

Design Study

Scope

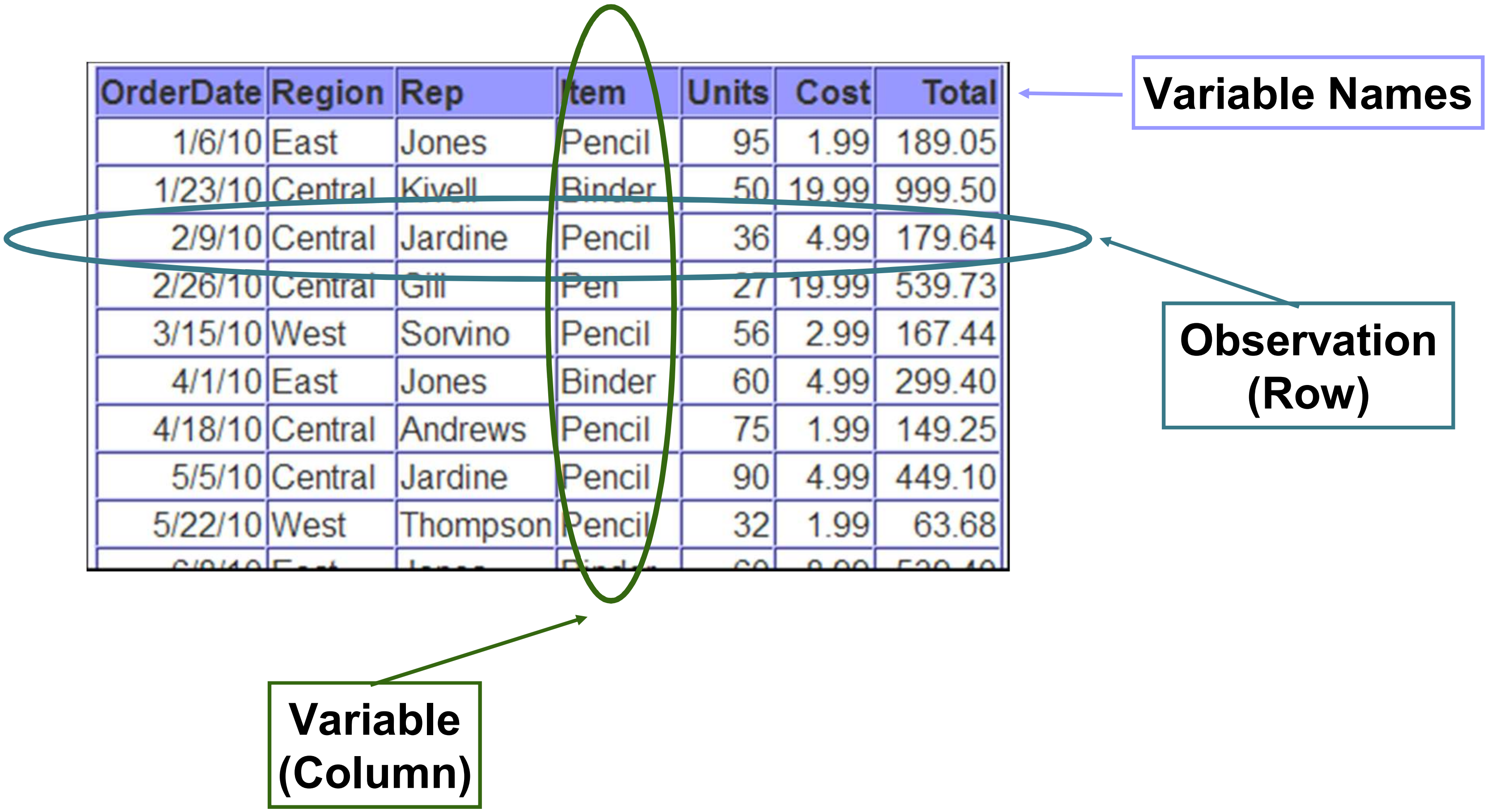
EDA

Model

Inference

Data Organization

Data is stored in the form of a ***Data Matrix***



OrderDate	Region	Rep	Item	Units	Cost	Total
1/6/10	East	Jones	Pencil	95	1.99	189.05
1/23/10	Central	Kivell	Binder	50	19.99	999.50
2/9/10	Central	Jardine	Pencil	36	4.99	179.64
2/26/10	Central	Gill	Pen	27	19.99	539.73
3/15/10	West	Sorvino	Pencil	56	2.99	167.44
4/1/10	East	Jones	Binder	60	4.99	299.40
4/18/10	Central	Andrews	Pencil	75	1.99	149.25
5/5/10	Central	Jardine	Pencil	90	4.99	449.10
5/22/10	West	Thompson	Pencil	32	1.99	63.68
6/8/10	East	Jones	Binder	60	8.99	539.40

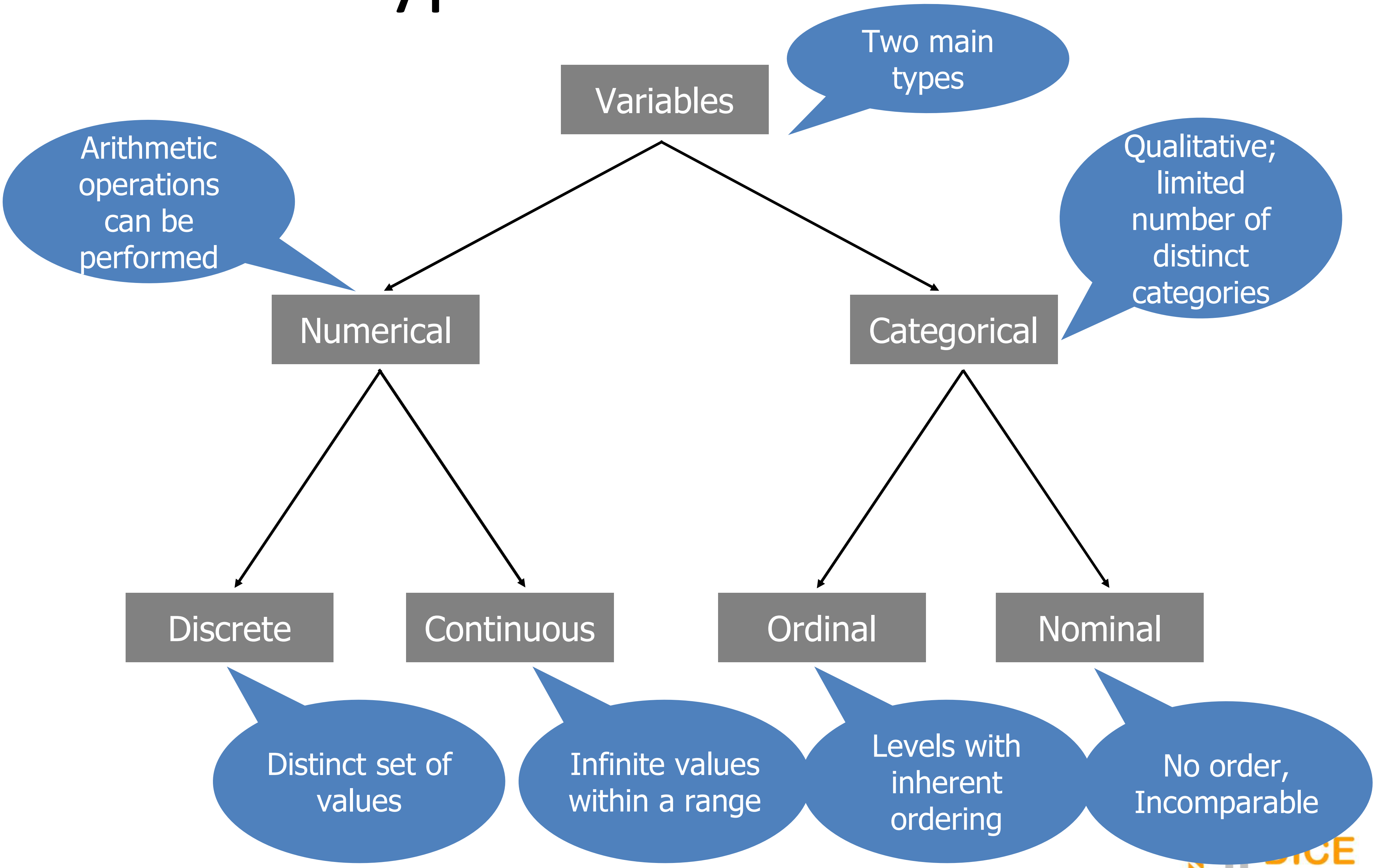
Variable Names

Observation (Row)

Variable (Column)

DICE ANALYTICS

Types of Variables



Variables

Two main types

Numerical

Arithmetic operations can be performed

Categorical

Qualitative; limited number of distinct categories

Discrete

Distinct set of values

Continuous

Infinite values within a range

Ordinal

Levels with inherent ordering

Nominal

No order, Incomparable

Types of Variables

http://www.statisticshowto.com/types-variables/

https://statistics.laerd.com/statistical-guides/types-of-variable.php

Types of Variables

- *Response Variable*: It is the focus of a question in a study or experiment. It is the variable we want to predict or observe. It is the dependent variable.

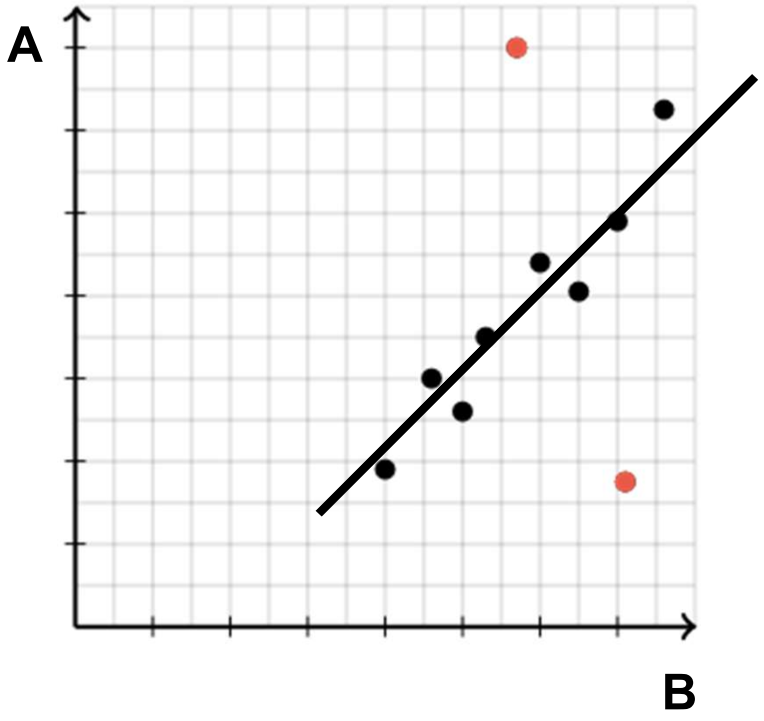
- *Explanatory Variable*: It is the variable on whom the response variable depends, or the variable which ‘explains’ the response variable. It is assumed to be independent variable.

Relationship b/w Variables

- Two variables that show connection with each other are called *Associated/Correlated (Dependent)*

- Two variables that do not show connection with each other are called *Independent*

- An observation that is away that is not close to majority of data is called *Outlier*



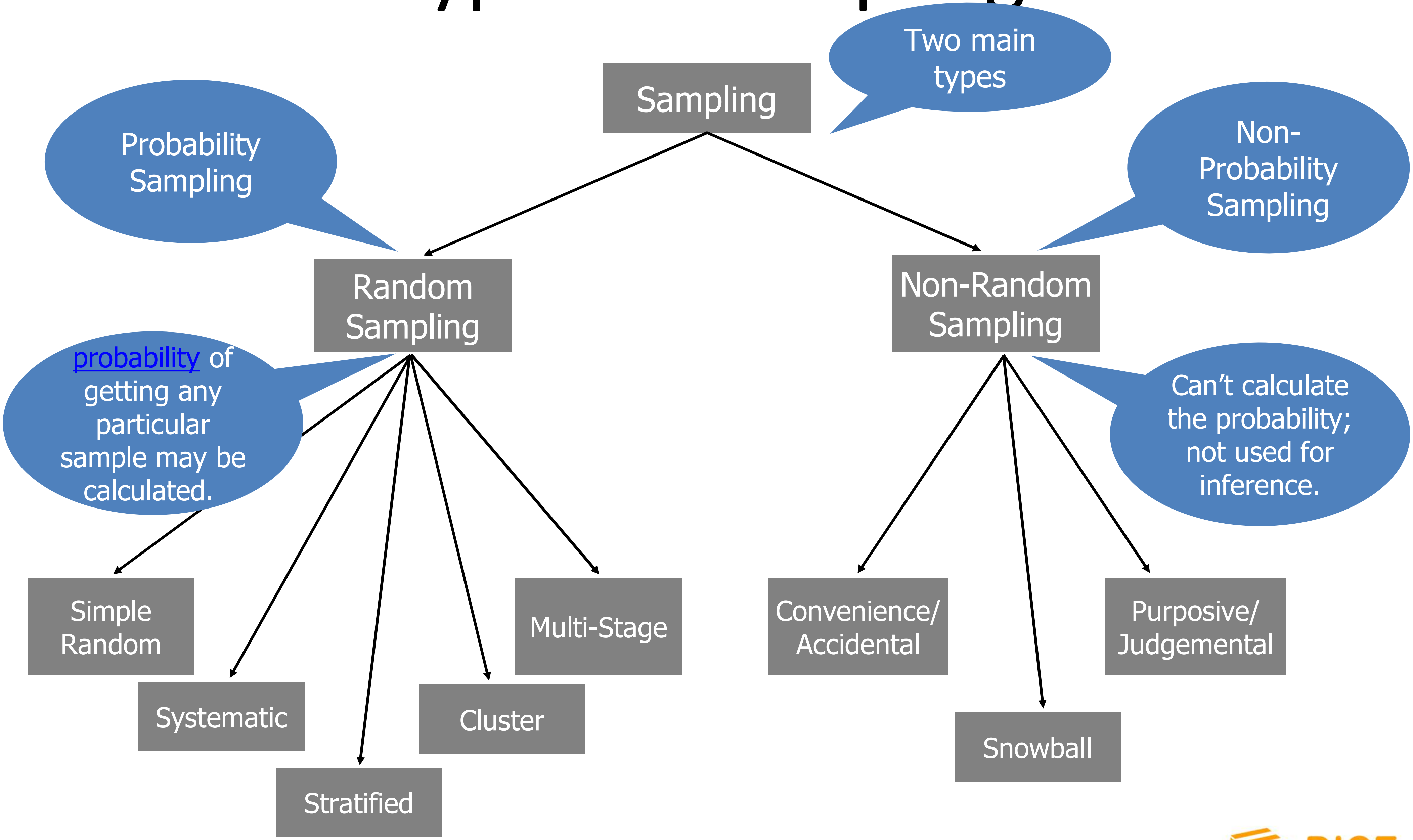
Sampling

Census vs Sample

- *Census*: A **census** is a study of every unit, everyone or everything, in a population. It is known as a complete enumeration, which means a complete count.
- Census not mostly possible: time-consuming, expensive, population hardly still, etc.

- *Sample*: A **sample** is a subset of units in a population, selected to represent all units in a population of interest.

Types of Sampling



Sampling

Two main types

Random Sampling

Probability Sampling

Non-Random Sampling

Non-Probability Sampling

probability of getting any particular sample may be calculated.

Simple Random

Systematic

Stratified

Cluster

Multi-Stage

Can't calculate the probability; not used for inference.

Convenience/ Accidental

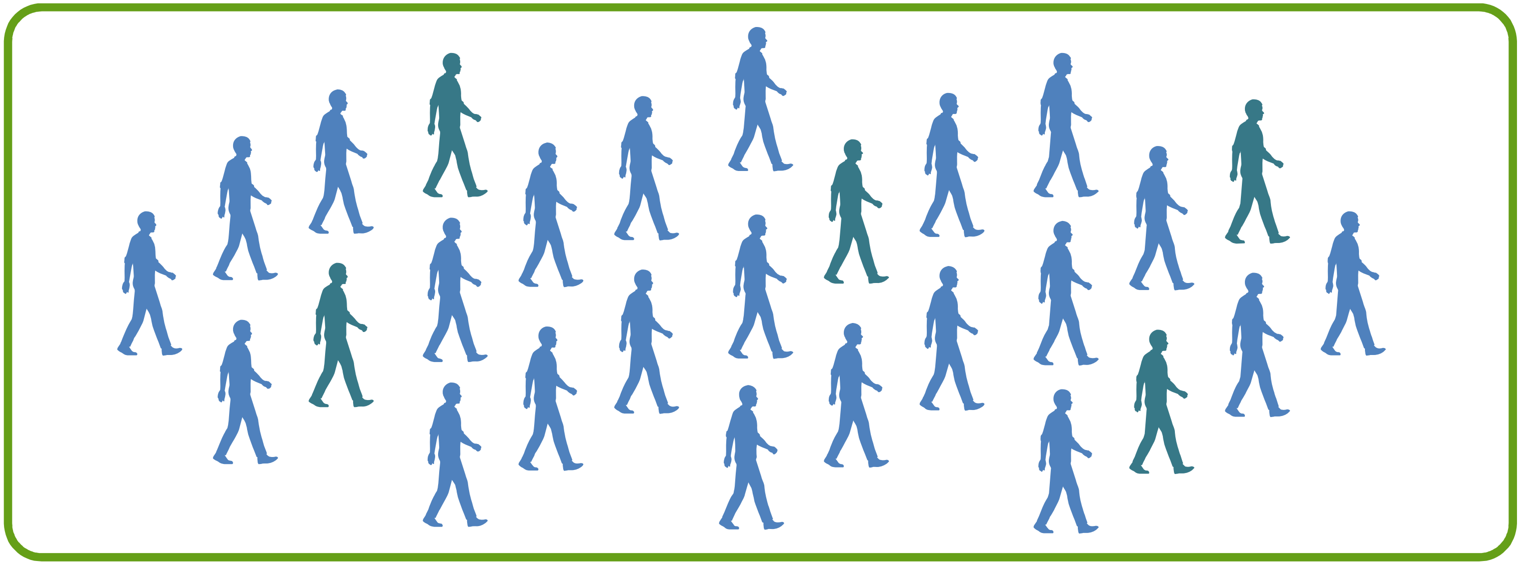
Snowball

Purposive/ Judgemental

Random Sampling

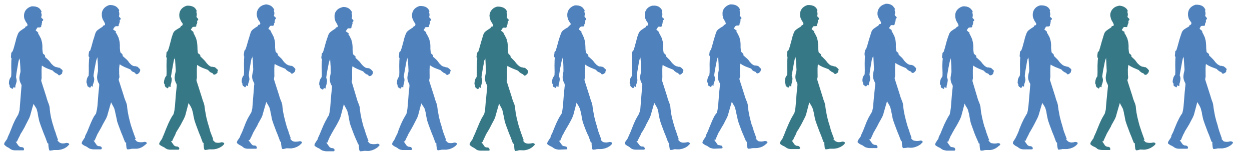
Simple Random Sampling (SRS)

- Select ***n*** observations randomly from entire population
- Each observation is likely to be selected



Systematic Sampling

- Arrange the population according to some ordering
- Start randomly and select every **k^{th}** observation

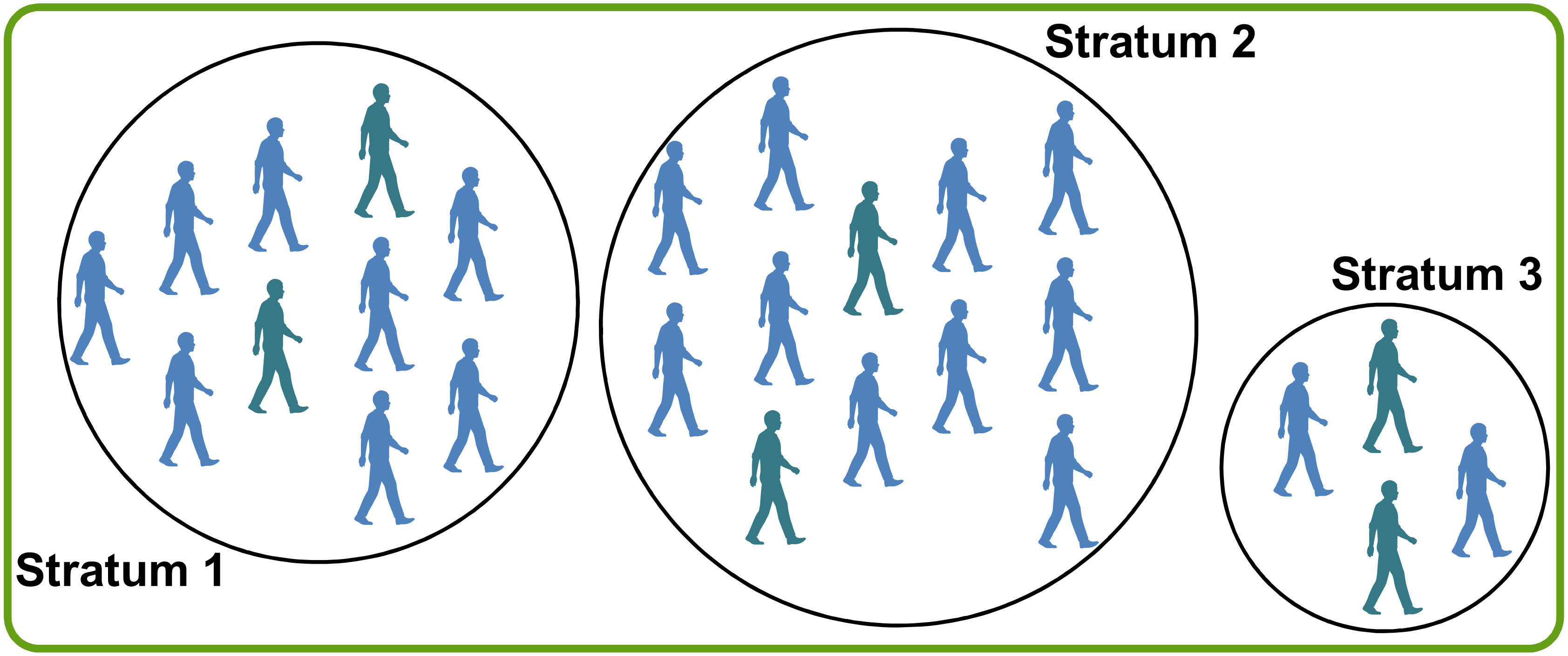


K = 4

Stratified Sampling

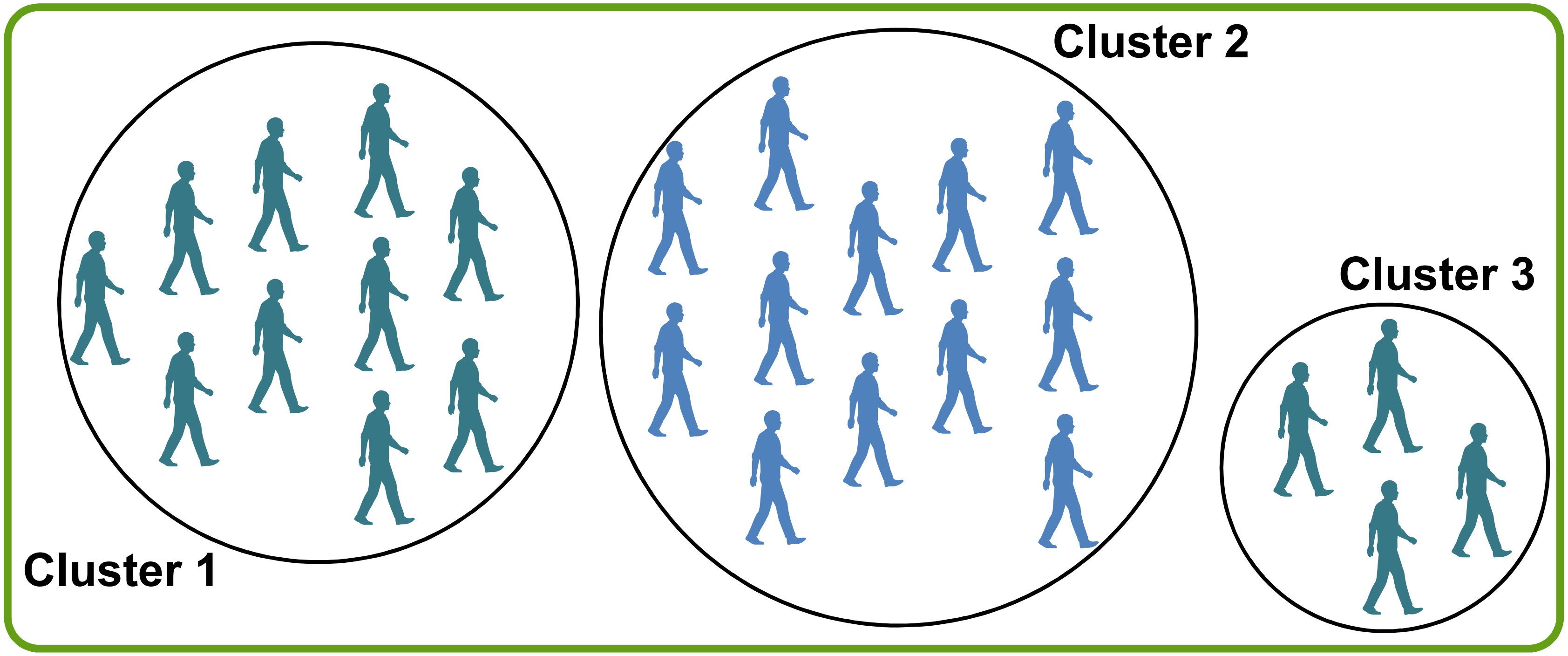
- Divide population in homogenous groups called ***strata***

- Do Simple Random Sampling (SRS) from each stratum



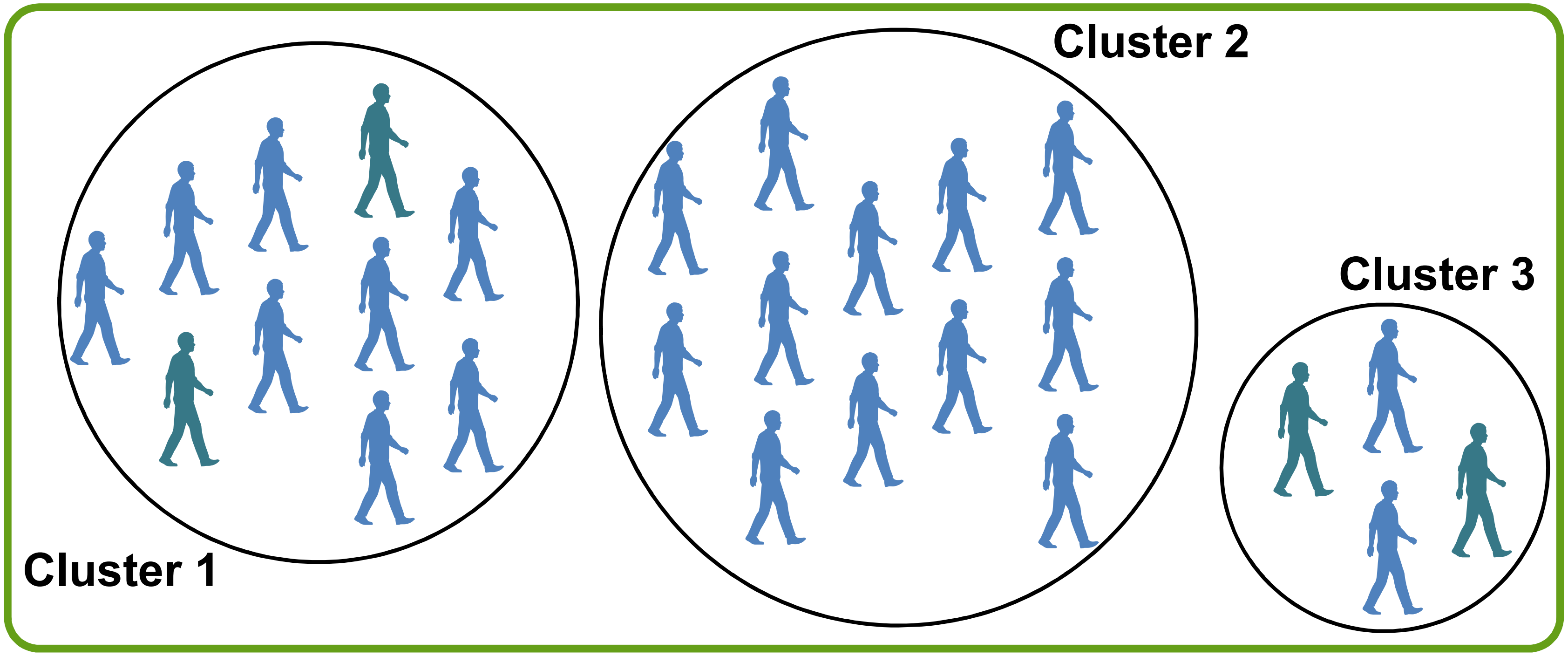
Cluster Sampling

- Divide population in heterogenous groups called *clusters*
- Randomly Sample ***k*** clusters; and sample all observations within those clusters



Multi-Stage Sampling

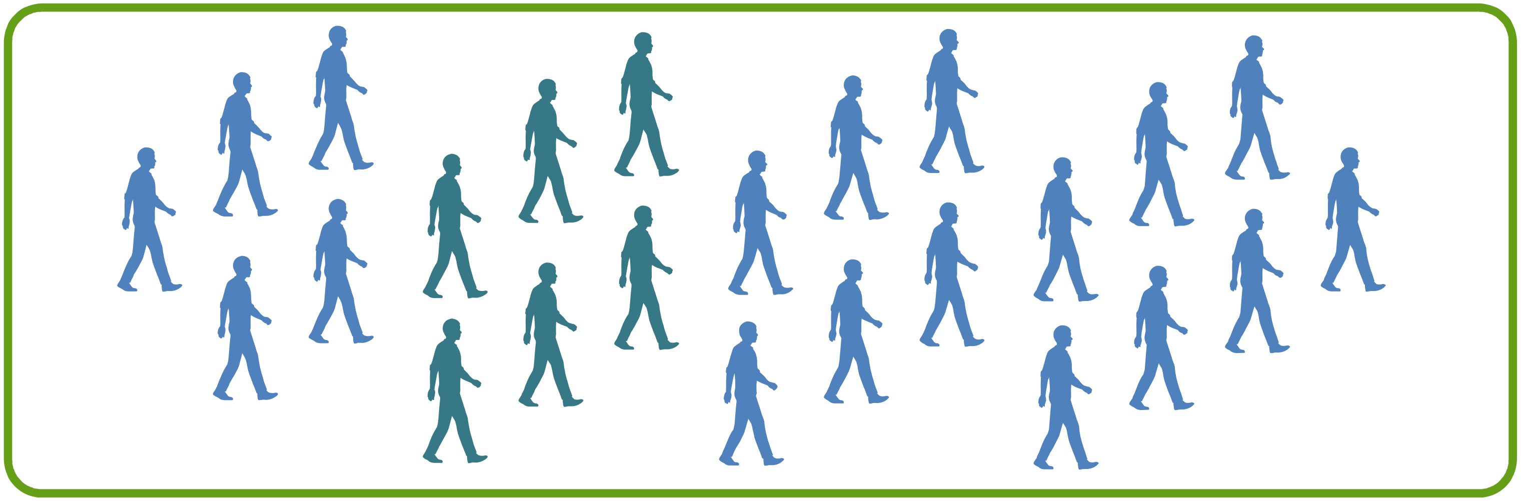
- Divide population in heterogenous groups called ***clusters***
- Randomly Sample ***k*** clusters; and do SRS within those clusters



Non-Random Sampling

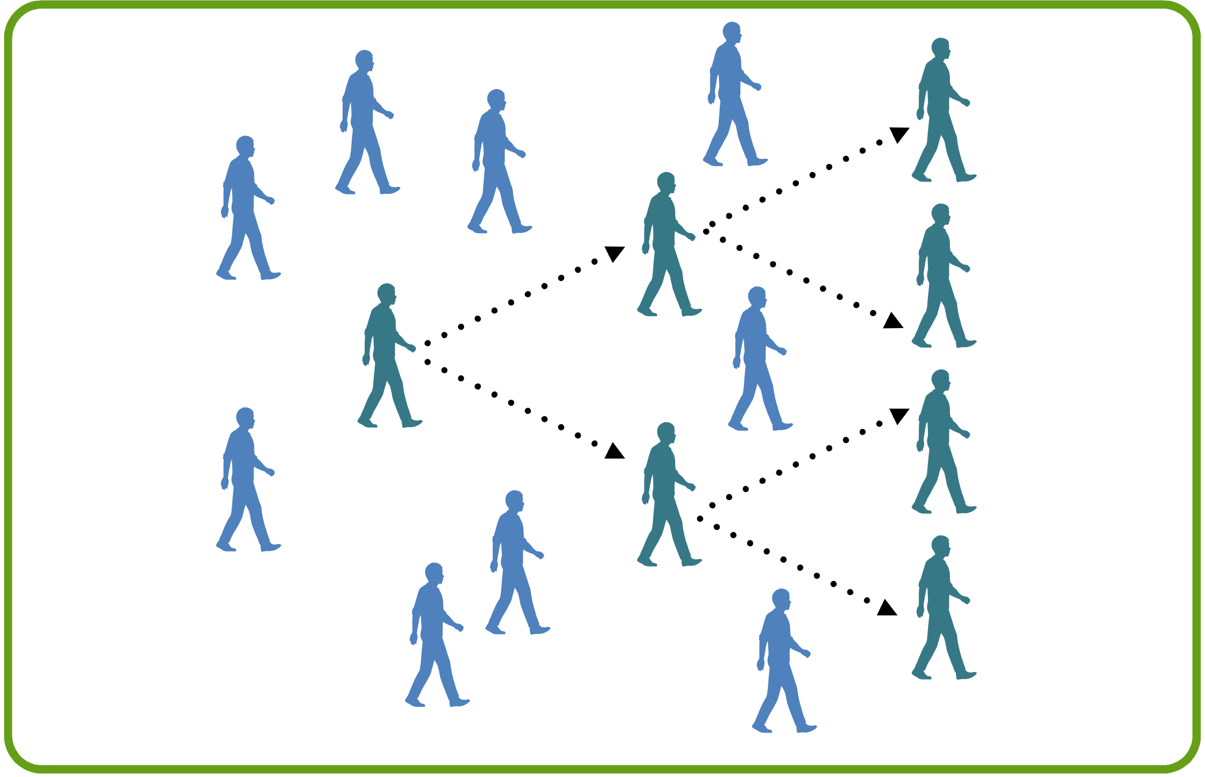
Convenience/Accidental Sampling

- Members of the population are chosen based on their relative ease of access.
- To sample friends, co-workers, or shoppers at a single mall, are all examples of convenience sampling.
- Such samples are biased because researchers may unconsciously approach some kinds of respondents and avoid others (Lucas 2014a), and respondents who volunteer for a study may differ in unknown but important ways from others (Wiederman 1999).



Snowball Sampling

- The first respondent refers an acquaintance. The friend also refers a friend, and so on.
- Such samples are biased because they give people with more social connections an unknown but higher chance of selection (Berg 2006), but lead to higher response rates.



Purposive/Judgmental Sampling

- The researcher chooses the sample based on who they think would be appropriate for the study.
- This is used primarily when there is a limited number of people that have expertise in the area being researched, or when the interest of the research is on a specific field or a small group.

Sampling Bias vs Selection Bias

- *Sampling Bias*: A **bias** in which a **sample** is collected in such a way that some members of the intended population are less likely to be included than others; occurs when you choose your sample which is the 1st step of a research.

- *Selection Bias*: A **bias** introduced by the **selection** of individuals, groups or data for analysis in such a way that proper randomisation is not achieved; occurs when you select which subject goes to the control group and which to the treatment group.

DICE ANALYTICS

Sources of Sampling Bias

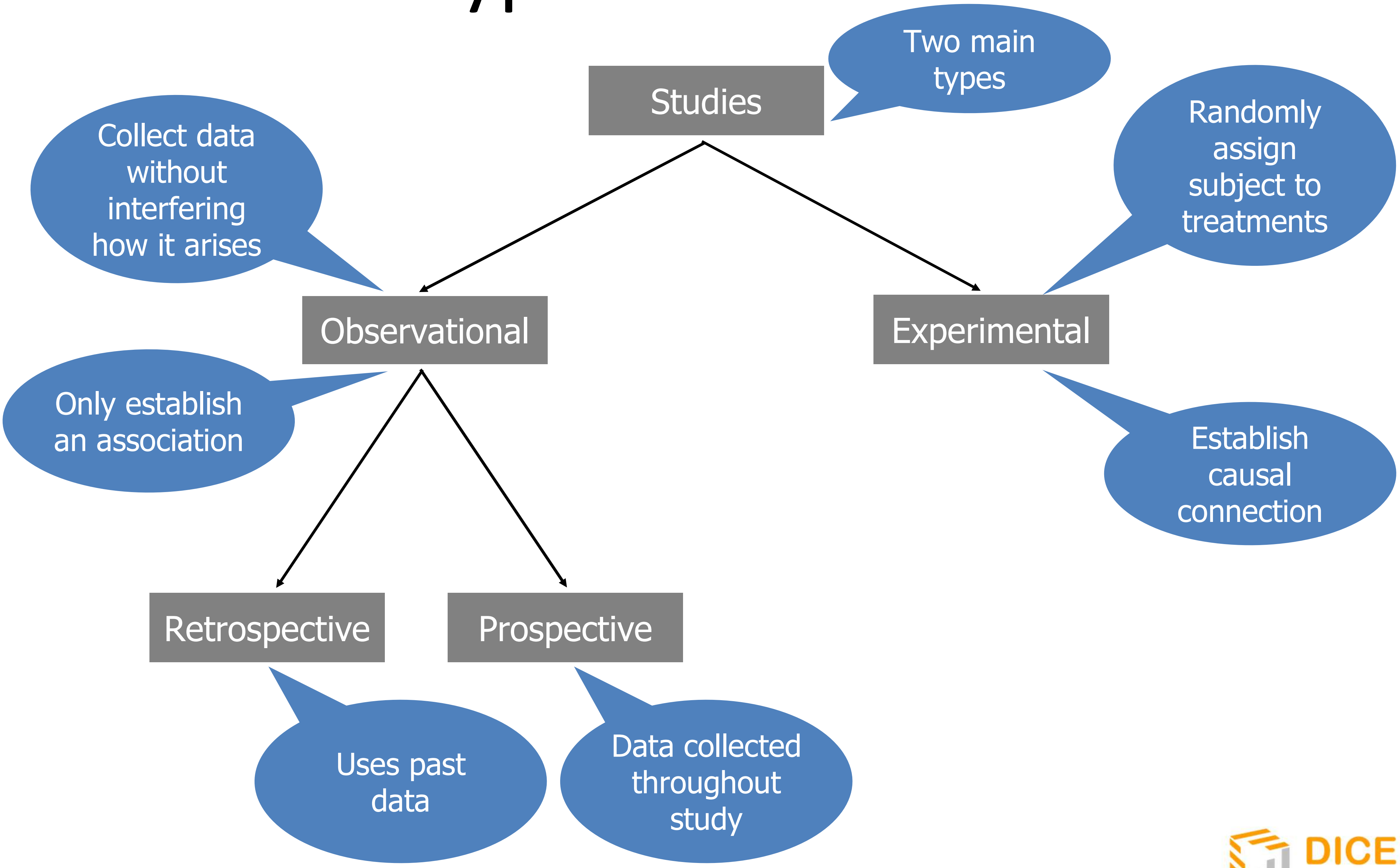
- *Convenience Sample:* Easily accessible people more likely to be included in the sample.

- *Non-Response*: If only particular type(s) of randomly sampled people respond to survey.

- *Voluntary Response*: Happens when sample consists of people who volunteered to respond because they are opinionated.

Study Design

Types of Studies



Studies

Two main types

Observational

Collect data without interfering how it arises

Only establish an association

Experimental

Randomly assign subject to treatments

Establish causal connection

Retrospective

Uses past data

Prospective

Data collected throughout study

DICE ANALYTICS

Correlation vs Causation

- *Correlation*: It describes the mutual relationship or connection between an independent and dependent variable.

- *Causation*: Causation, also known as cause and effect, is when an observed event or action (independent variable) appears to have caused a second event or action (dependent variable).

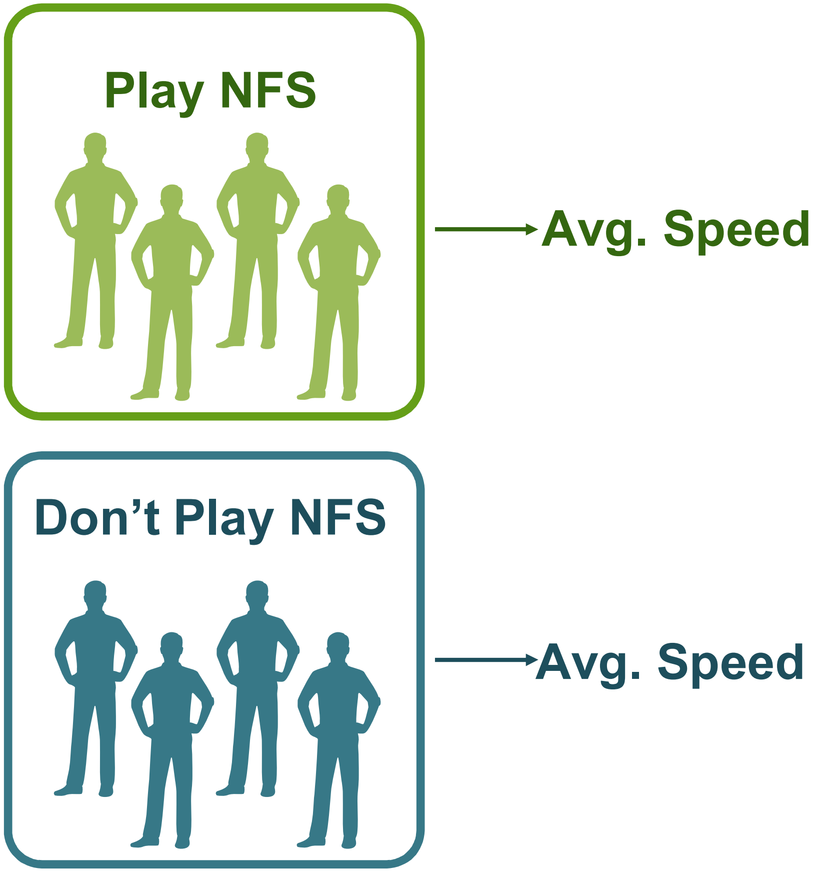
Correlation does not imply Causation!

Random Assignment

Find out relationship between ***playing NFS*** and actual ***driving speed*** of a person

Random Assignment

Observational Study



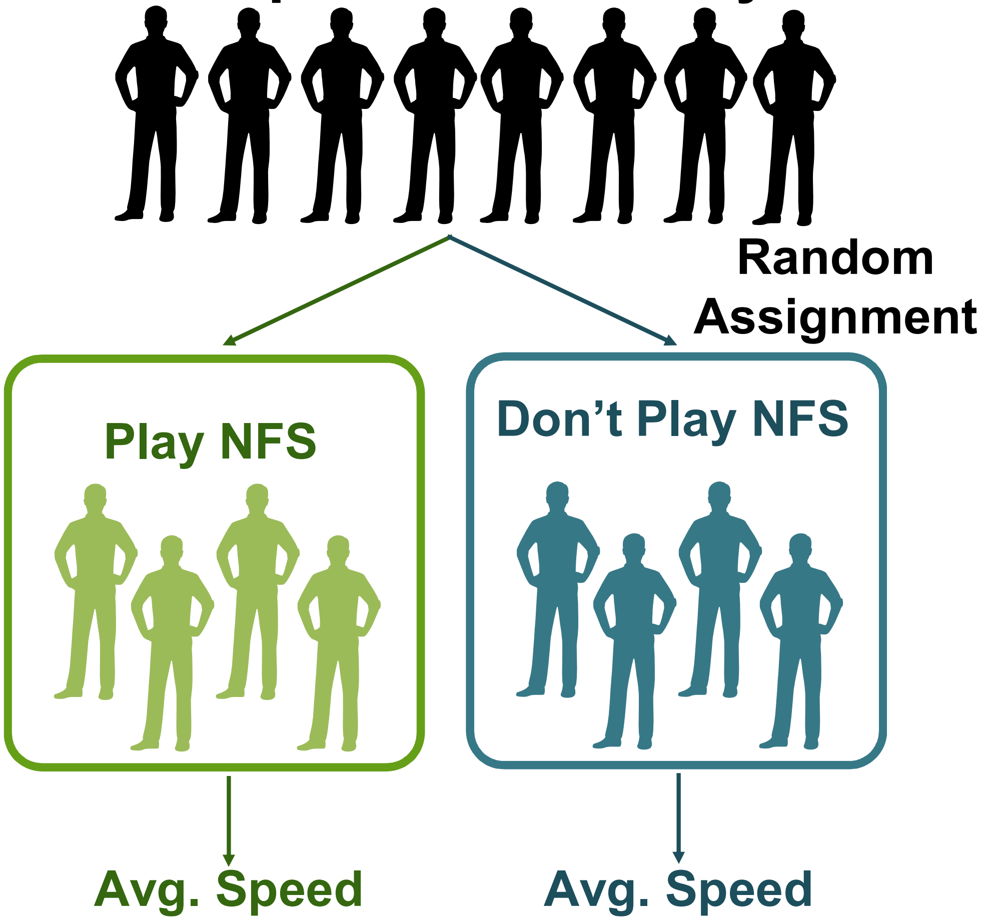
Play NFS

Avg. Speed

Don’t Play NFS

Avg. Speed

Experimental Study



Random Assignment

Play NFS

Don’t Play NFS

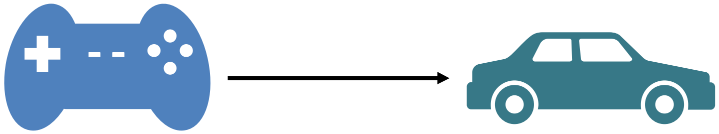
Avg. Speed

Avg. Speed

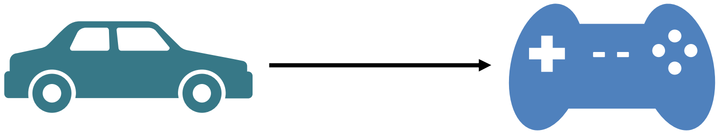
(No) Random Assignment

In Case of Observational Study:

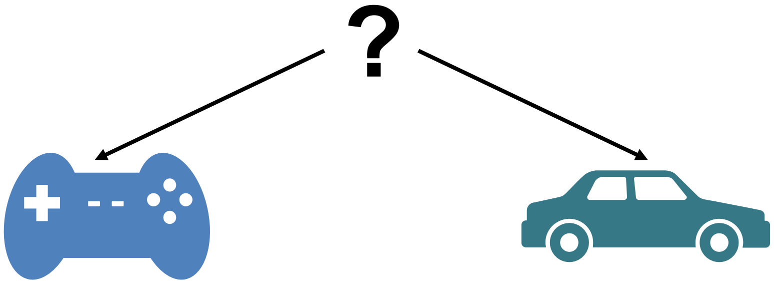
- Playing NFS causes the person to drive faster



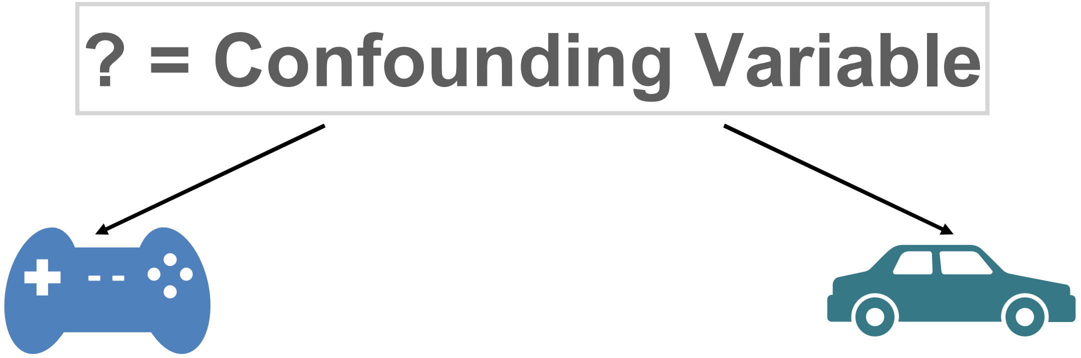
- Driving faster causes the person to play NFS



- A third variable is responsible for both these variables



Confounding Variable



? = Confounding Variable

Extraneous variables that affect both the explanatory and response variable, and make it look like there is a relationship (association/dependence) between them are called Confounding Variables.

Maybe because they are very rich!

Principles of Experimental Design

Control
Compare treatment of interest to a control group

Randomize
Randomly assign subjects to treatments

Replicate
Collect a sufficiently large sample;
or replicate entire study

Block
Block for variables known or suspected to affect the outcome

Blocking

Find out effect of ***Hi-Octane*** on car's ***Speed***

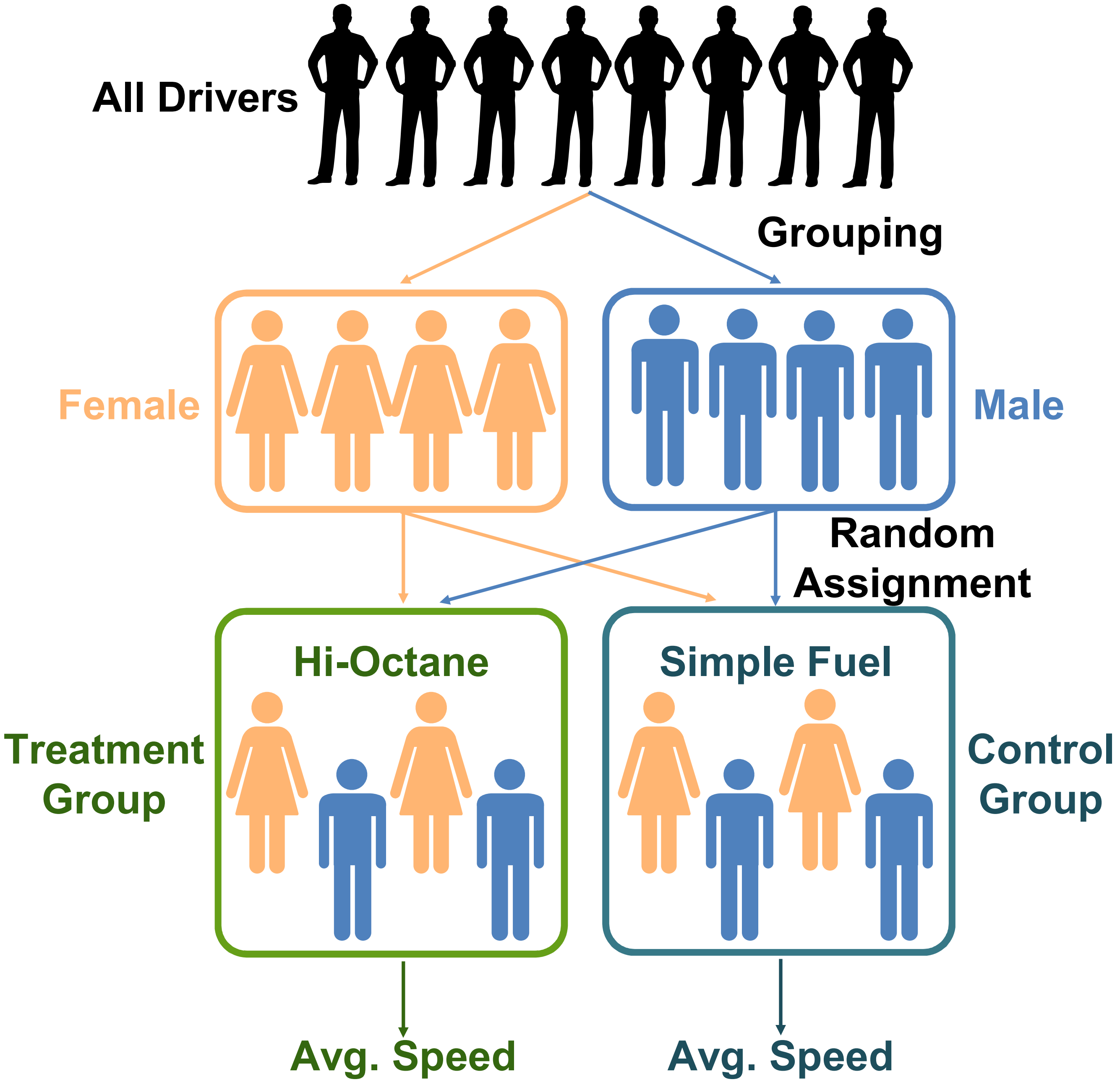
Blocking

- Control Group: Simple fuel
- Treatment Group: Hi-Octane

Male drivers may ride faster than Female drivers!

- Gender is blocking variable
- Need to ‘*block*’ Male status
- Divide sample to Male and Female groups (just like we do in Stratified Sampling)
- Randomly assign Male and Female drivers to control and treatment group, ensuring equal representation in both groups
- Now they cancel out the effect of gender, so we can say the difference in speed is solely because of Hi-Octane

Blocking



All Drivers

Grouping

Female

Male

Random Assignment

Hi-Octane

Simple Fuel

Treatment Group

Control Group

Avg. Speed

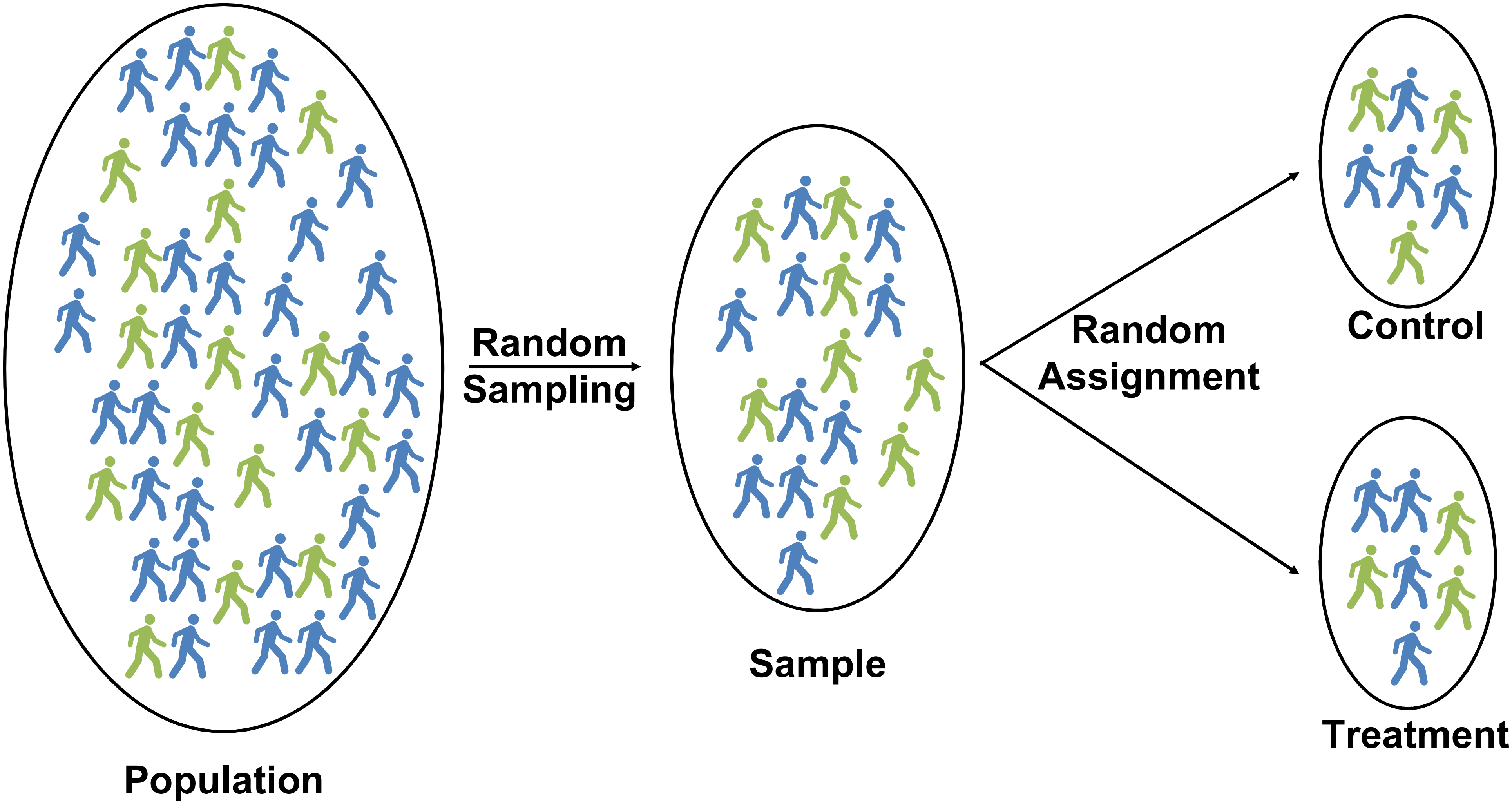
Avg. Speed

DICE ANALYTICS

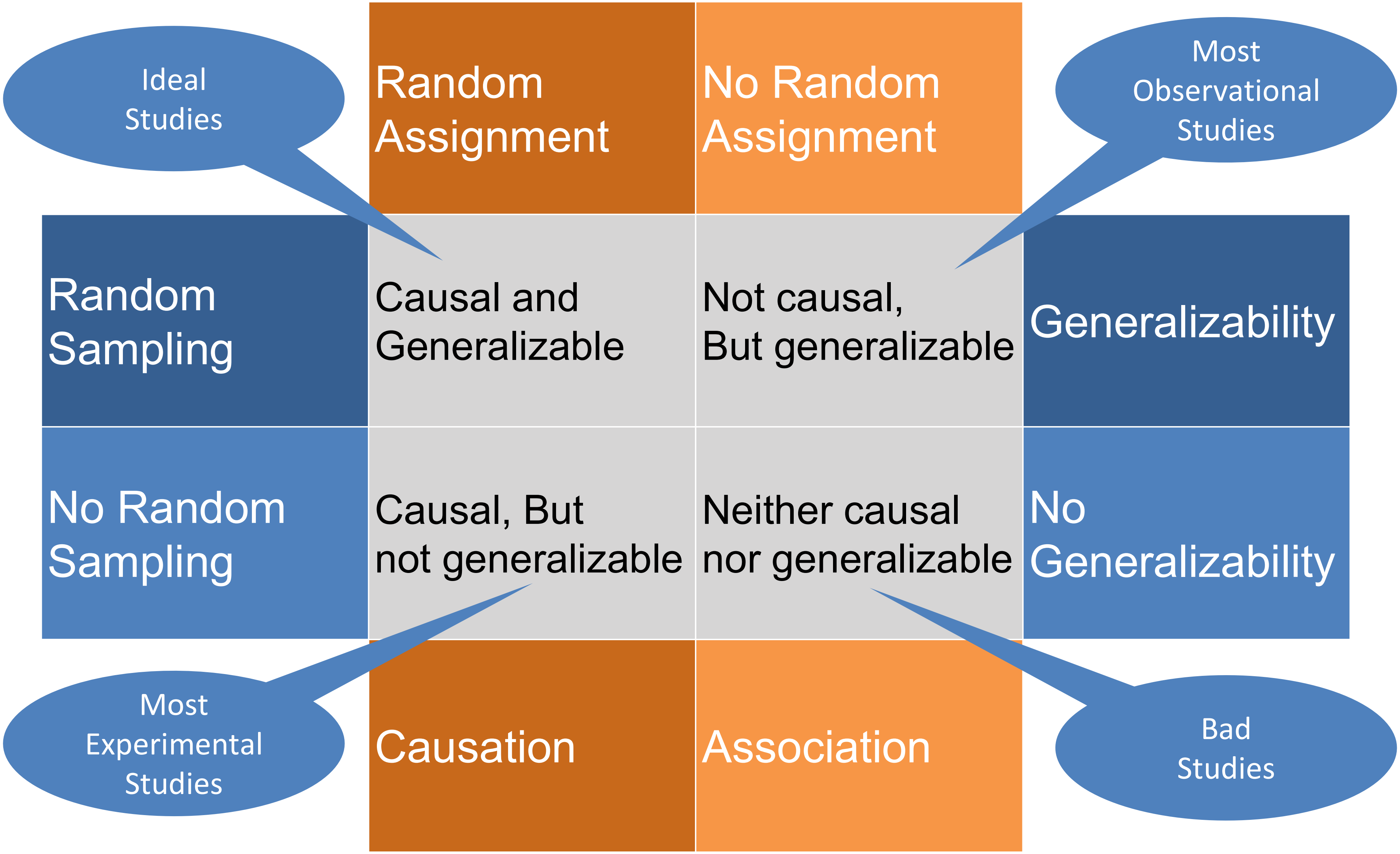
Blocking Variable vs Explanatory Variable

- *Explanatory Variables (factors)*: Conditions we can impose on experimental units.
- *Blocking Variables*: Characteristics that experimental units come with, that we would like to control for.
- Blocking is like Stratifying:
  -> blocking during random assignment
  -> stratifying during random sampling

Random Sampling vs Random Assignment



Scope



	Random Assignment	No Random Assignment	
Random Sampling	Causal and Generalizable	Not causal, But generalizable	Generalizability
No Random Sampling	Causal, But not generalizable	Neither causal nor generalizable	No Generalizability
	Causation	Association	

Ideal Studies

Most Observational Studies

Most Experimental Studies

Bad Studies

DICE ANALYTICS