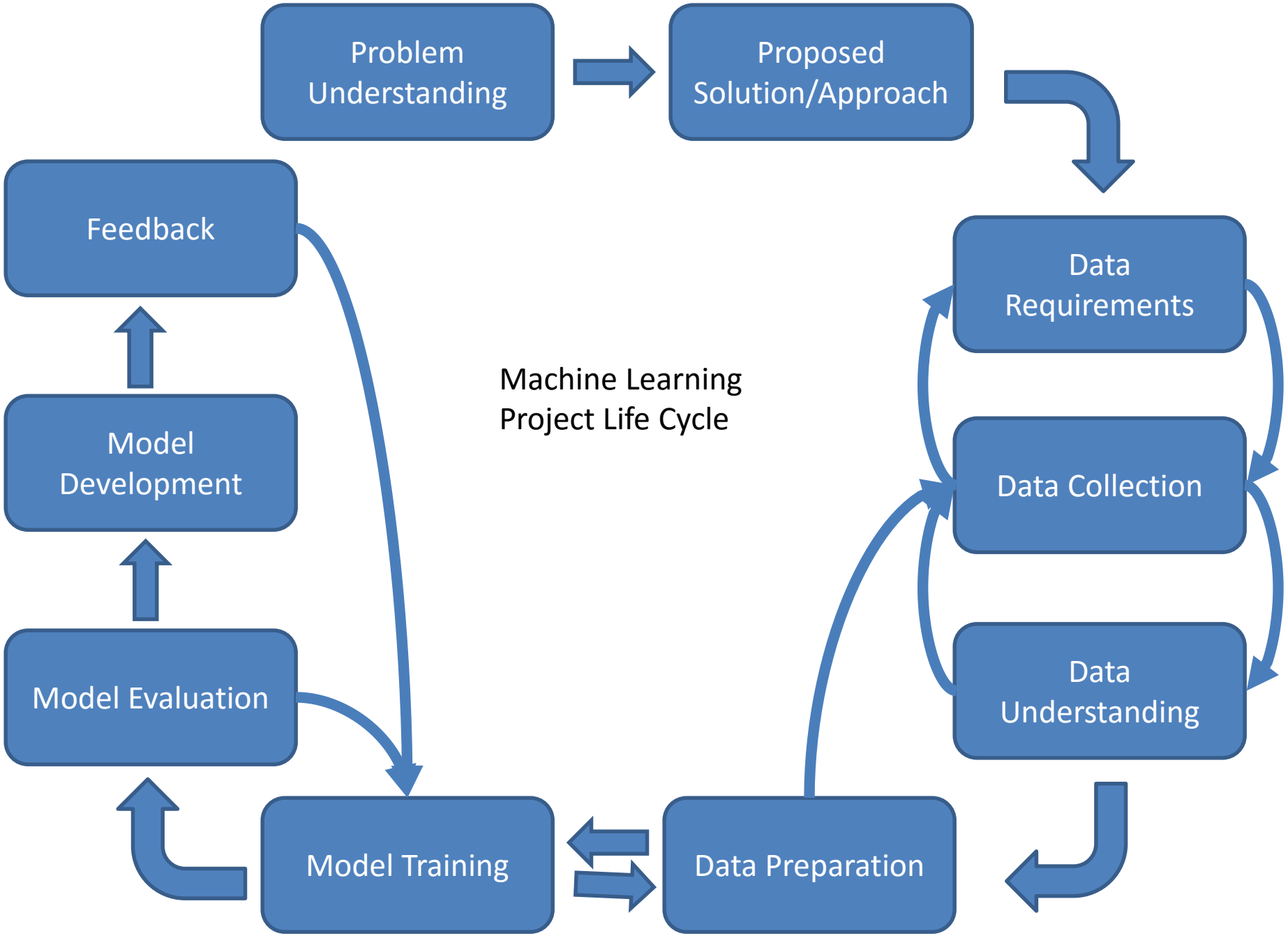
Data Requirements & Collection

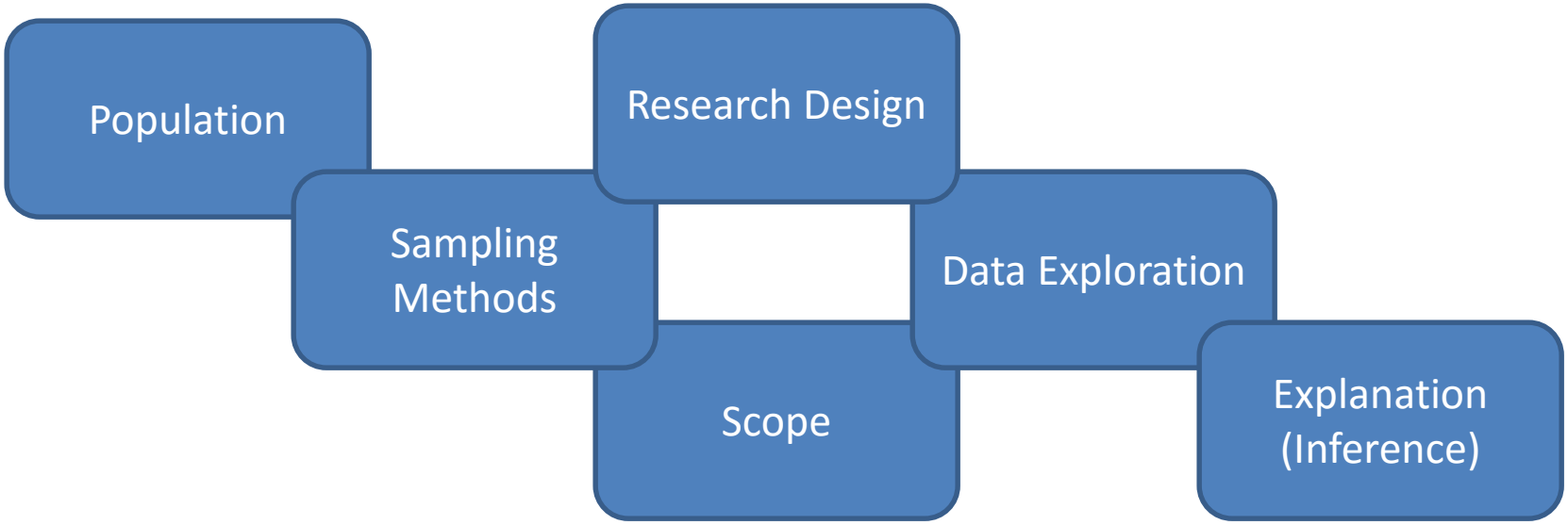
Dr. Uzair Ahmad
Data Science Lab,
Ryerson University


Problem Understanding
Proposed Solution/Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Model Training
Model Evaluation
Model Development
Feedback
Machine Learning Project Life Cycle

Population & Sample

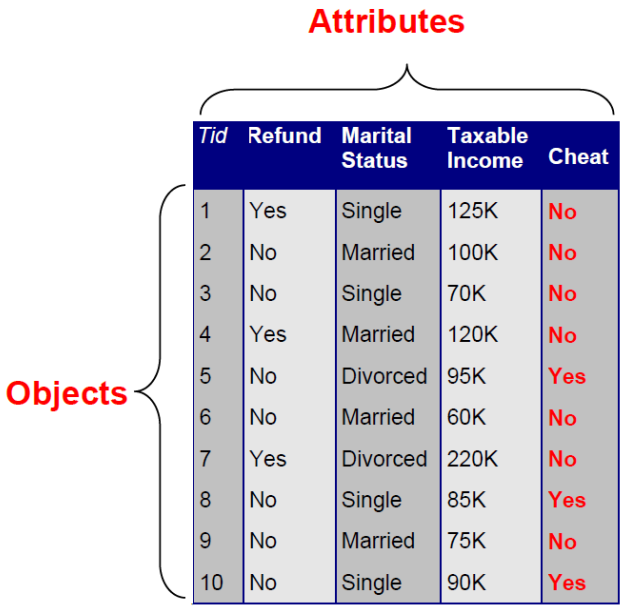
- **Discovering Associations (Research Question)**
  - Does expensive vehicle owners are involved in violation of traffic laws ?
- **Population**
  - Everybody that owns a vehicle in Neverland
- **Sample**
  - Traffic Police record at local police office
- **Inference Generalization**
  - Correlation vs Causation
  - Everybody that owns a vehicle in Neverland

Population & Sample



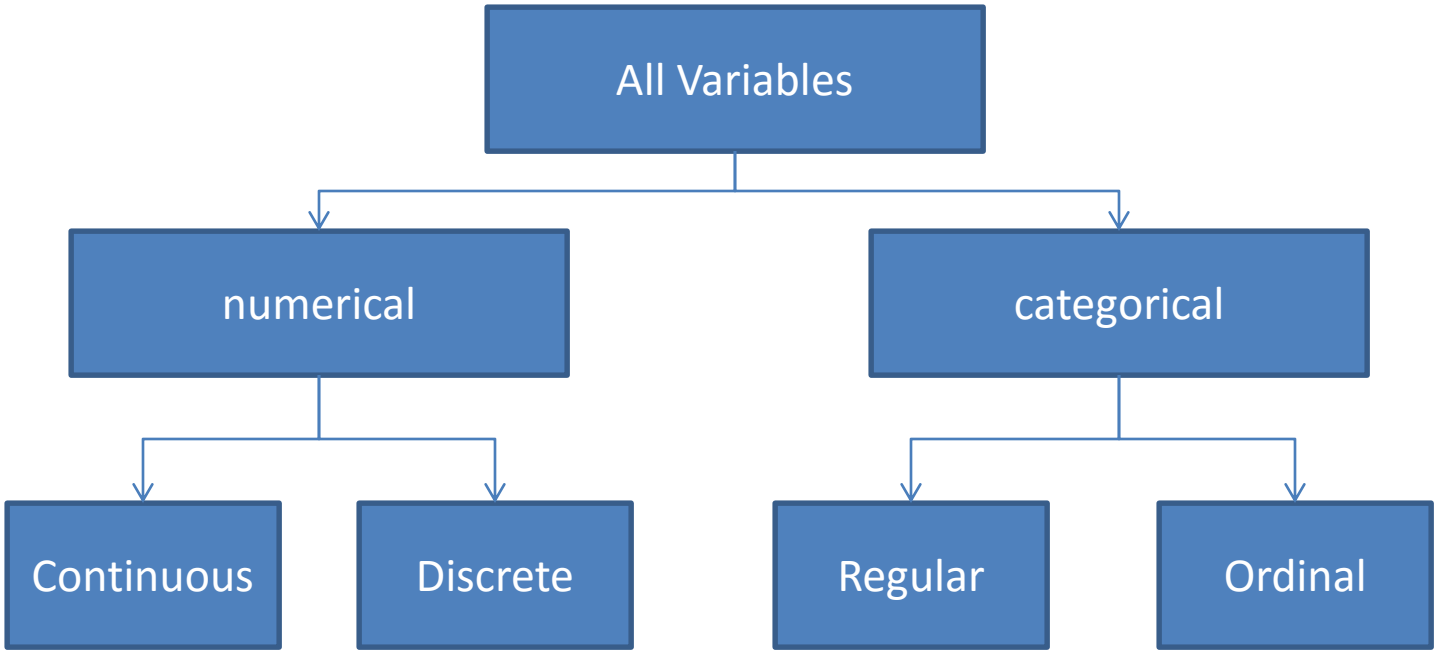
Data

- Data Matrices
- Data Observations, Cases, Objects
- Data Variables, Attributes
- Relationship in variables



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	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attributes, Variables



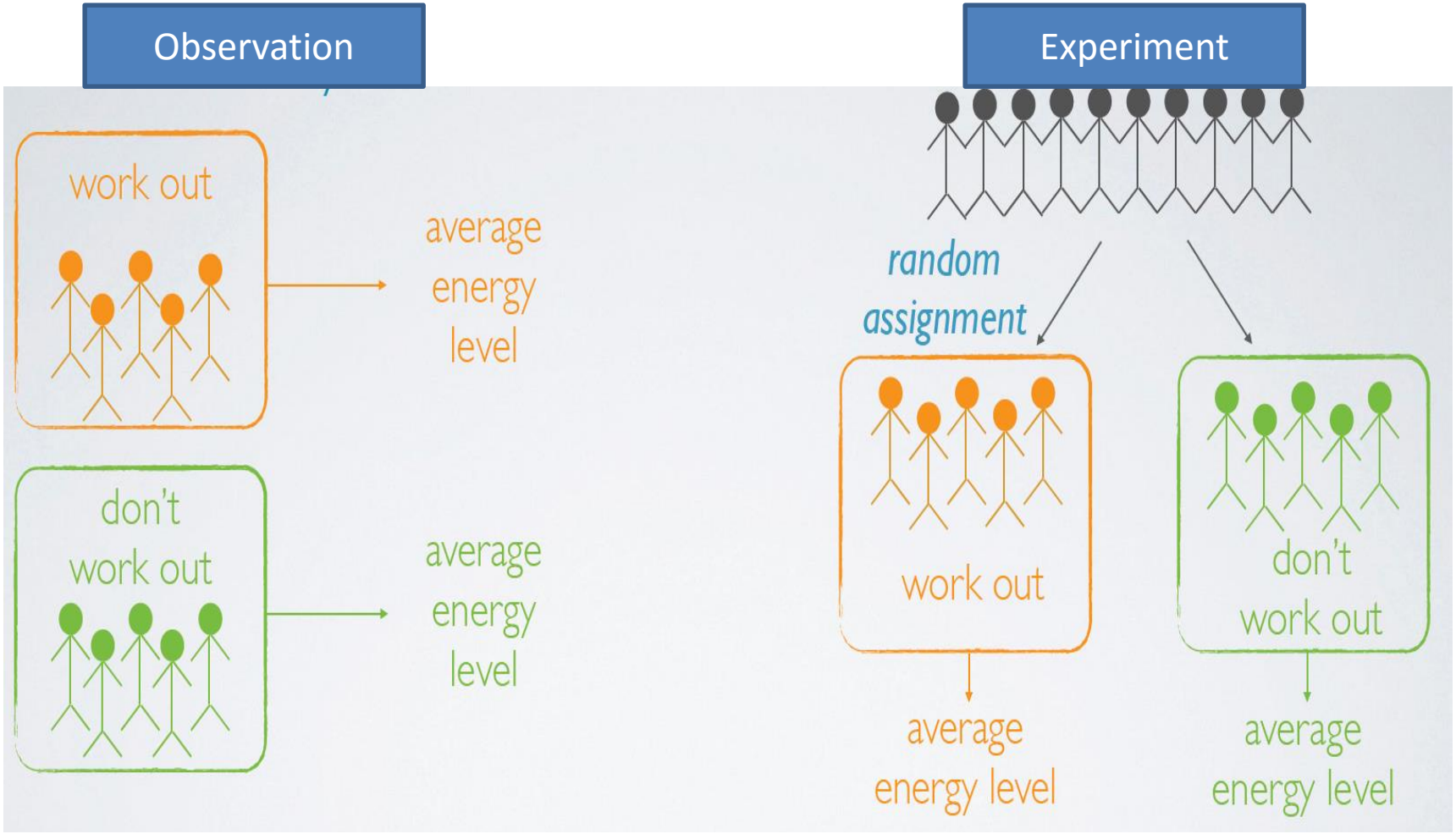
Relationships between Variables

- Dependent (Associated) Variables
  - Positive
  - Negative
- Independent Variables

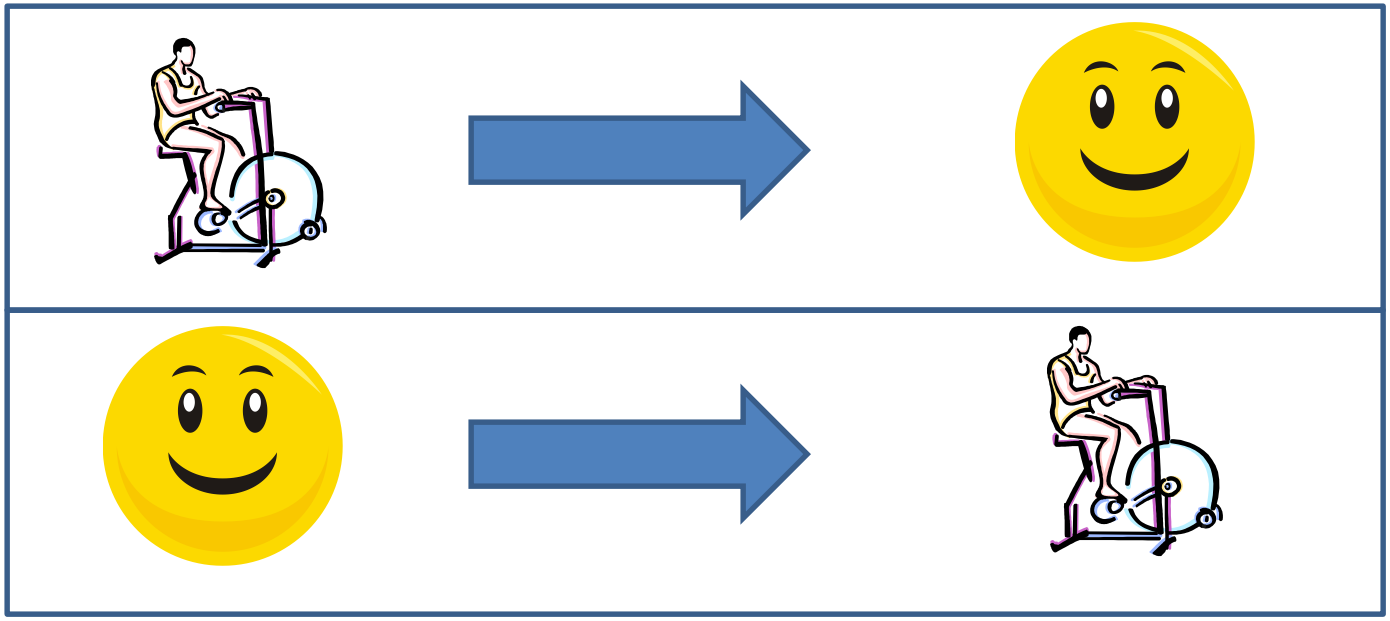
Data Collection Process

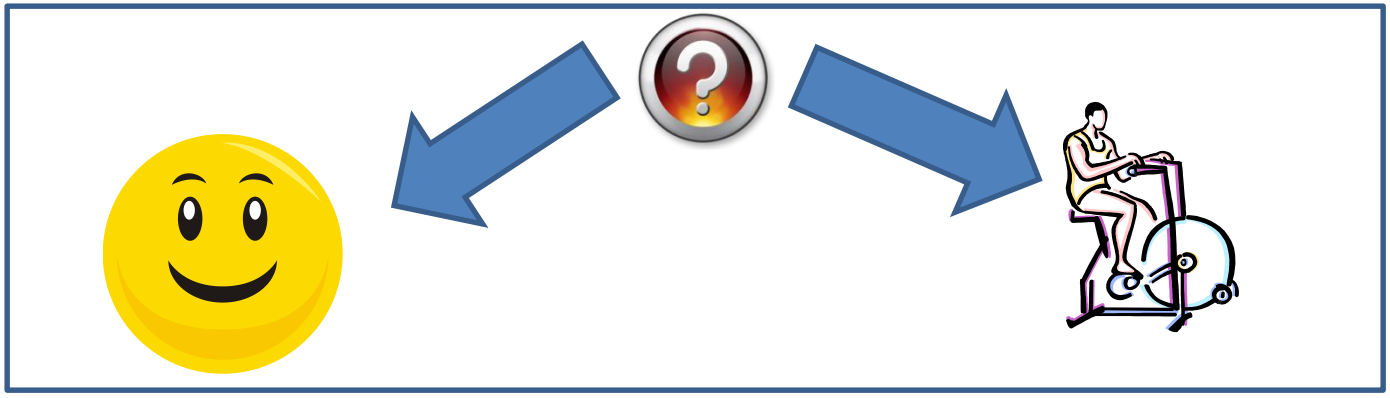
- Observation
  - Do not interfere how data is generated
  - Retrospective: Past data
  - Prospective: Current Data
- Experiment
  - Controlling the data generation
  - Random assignment of subjects to various conditions

Data Collection Process



Explanation of Relationships



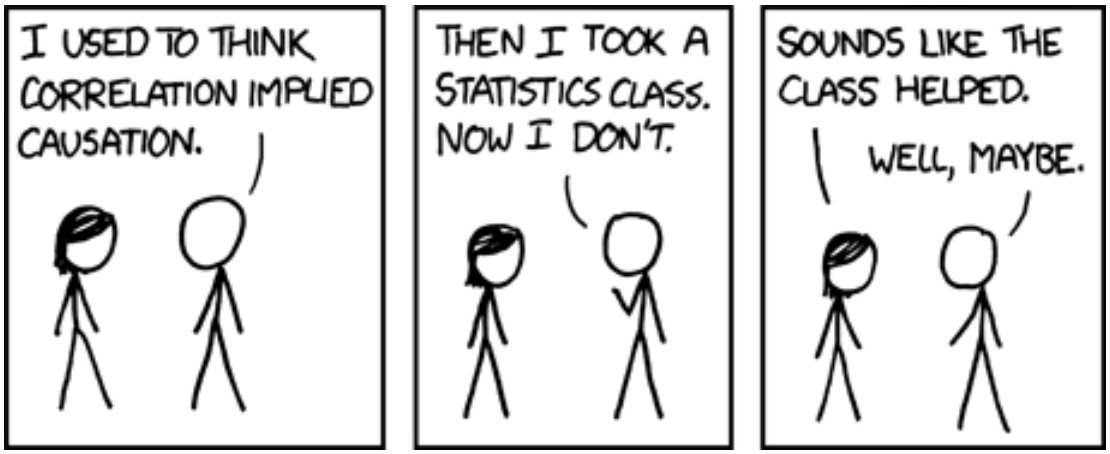


Confounding Variables

- Extraneous variables that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them are called confounding variables.

Explanation of Relationships

Correlation does not imply causation…



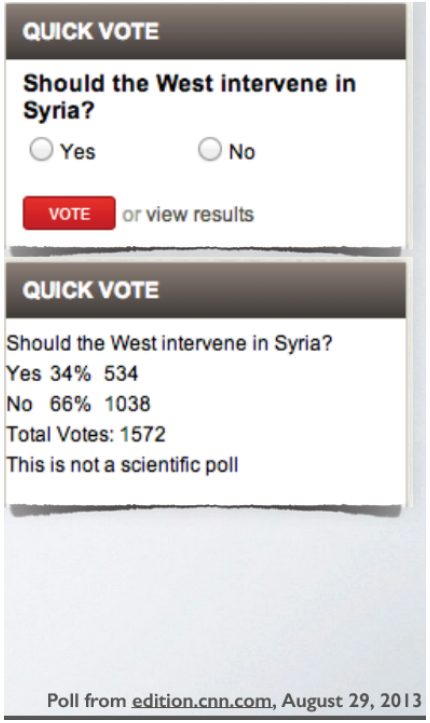
http://xkcd.com/552/

Sampling vs Census

- Census
  - Expensive: Time & Resources
  - Not everybody is accessible or willing to share data
  - Populations keep changing
- Sampling
  - Tasting your food (spoonful or the whole meal)
  - Representativeness

Sampling Bias

- Convenience Bias
  - Polling from neighborhood
- Non-Response
  - Data from only a fraction of randomly selected population
- Voluntary Response
  - Data from people with specific interests in outcome



- A Bank is considering updates to their credit card policies randomly samples 1000 of their credit card holders to survey on the phone. The phone calls are made during business hours, therefore there is a lower rate of responses from members who work during these hours. What type of bias is this indicative of?

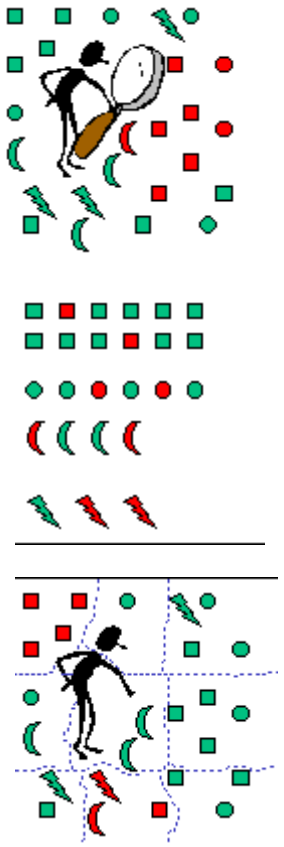
  - convenience sample
  - non-response
  - voluntary response
  - none of the above

- Samba Bank is considering updates to their credit card policies randomly samples 1000 of their credit card holders to survey on the phone. The phone calls are made during business hours, therefore there is a lower rate of responses from members who work during these hours. What type of bias is this indicative of?

  × convenience sample
  √ non-response
  × voluntary response
  × none of the above

Sampling Methods

- **simple random sample**
  - Randomly select cases from the population, such that each case is equally likely to be selected
- **stratified sample**
  - Divide the population into homogenous strata (groups) on the basis of common characteristic
  - randomly sample from within each stratum (group).
- **cluster sample**
  - Divide the population clusters (typically on the basis of geography),
  - Not necessarily homogeneous within themselves
  - randomly sample a few clusters, and then randomly sample from within these clusters.



- A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the **least effective?**

  - Simple random sampling
  - Cluster sampling, where each cluster is a neighborhood
  - Stratified sampling, where each stratum is a neighborhood type (Villas, Apartments, Markets)

- A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

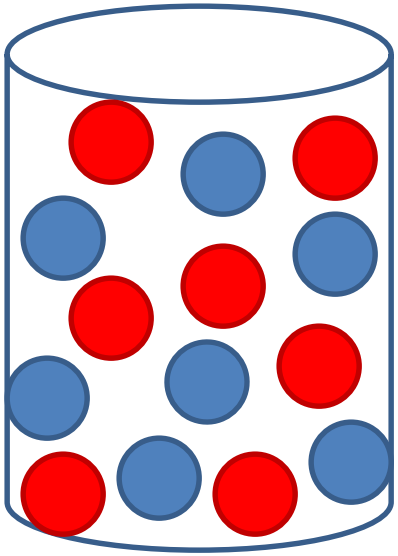
  √ Simple random sampling

  – Stratified sampling, where each stratum is a neighborhood type (Villas, Apartments, Markets)

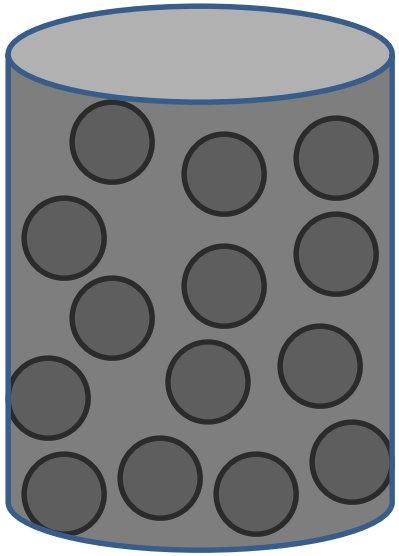
  – Cluster sampling, where each cluster is a neighborhood

Draw a Red Ball.

Known Distribution
50% Red, 50% Blue

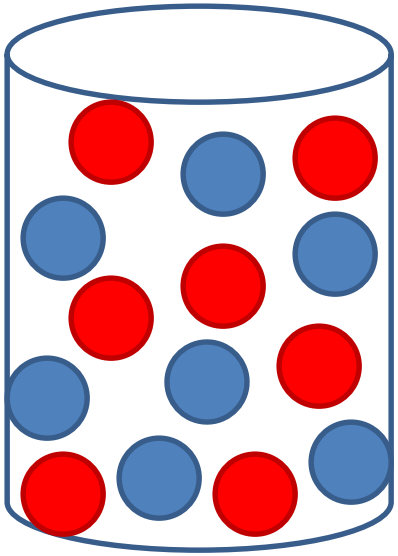


Unknown Distribution
My be 10% Red, 90% Blue

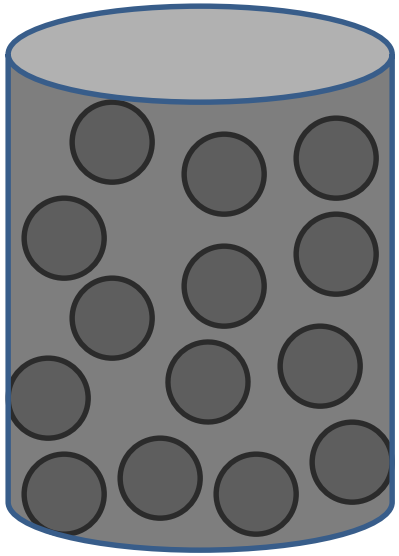


Draw a Blue Ball.

Known Distribution
50% Red, 50% Blue

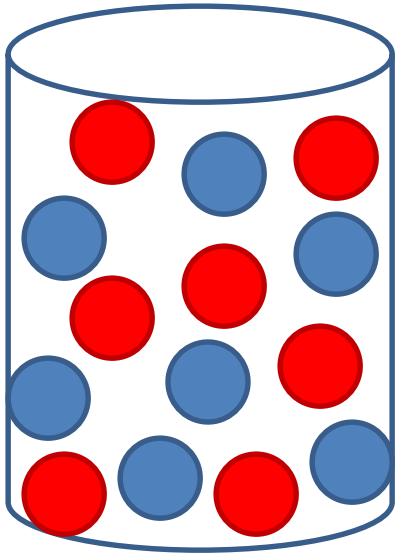


Unknown Distribution
My be 10% Red, 90% Blue or
90% Blue and 10 % Red



- Why did you choose the same bucket for both draws ?

Known Distribution
50% Red, 50% Blue



Principles of Experimental Design

- **Control**: Compare treatment of interest to a control group.
- **Randomize**: Randomly assign subjects to treatments.
- **Replicate**: Within a study, replicate by collecting a sufficiently large sample, or replicate the entire study.
- **Block**: If there are variables that are known or suspected to affect the response variable,
  - Group subjects into blocks based on these variables,
  - Randomize cases within each block to treatment groups.

Principles of Experimental Design

- We would like to design an experiment to investigate if Expensive Vehicle make you drive faster:
  - Treatment Group: Expensive Car
  - Control Group: Inexpensive Car
- It is suspected that Expensive Vehicles might affect married and single drivers differently,
  - therefore we block for married status:
    - Divide the sample to married and single
    - Randomly assign married and single drivers to treatment and control groups
  - Now Both are equally represented in the resulting treatment and control groups

Blocking vs Explanatory Variables

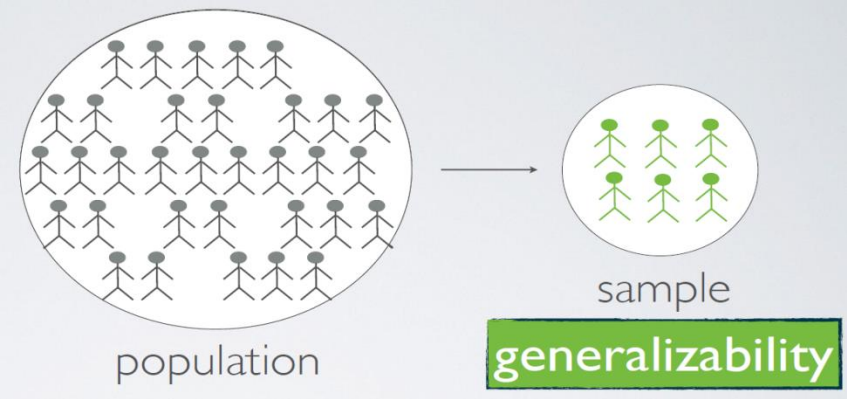
- Explanatory variables (also sometimes called factors) are conditions we can impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with, that we would like to control for.
  - Blocking is like stratifying, except that it is used in experimental settings when randomly assigning, as opposed to when sampling.

More Terminology

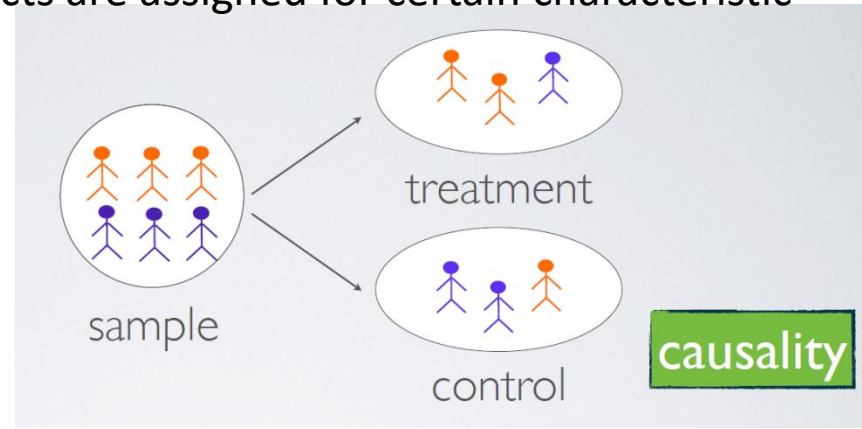
- **placebo**: fake treatment, often used as the control group for medical studies
- **placebo effect**: experimental units showing improvement simply because they believe they are receiving a special treatment
- **blinding**: when experimental units do not know whether they are in the control or treatment group
- **double-blind**: when both the experimental units and the researchers do not know who is in the control and who is in the treatment group

Random Sampling vs Assignment

Random Sampling occurs when subjects are selected fro a study



Random Assignment occurs only in experimental settings,
- subjects are assigned for certain characteristic



Sampling happens first
and then assignment happens second.

Random Sampling vs Assignment