Foundation of Data Science
Lecture 2, Module 3
Fall 2022

Rumi Chunara

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.

Data Preprocessing

Major Tasks in Data Preprocessing

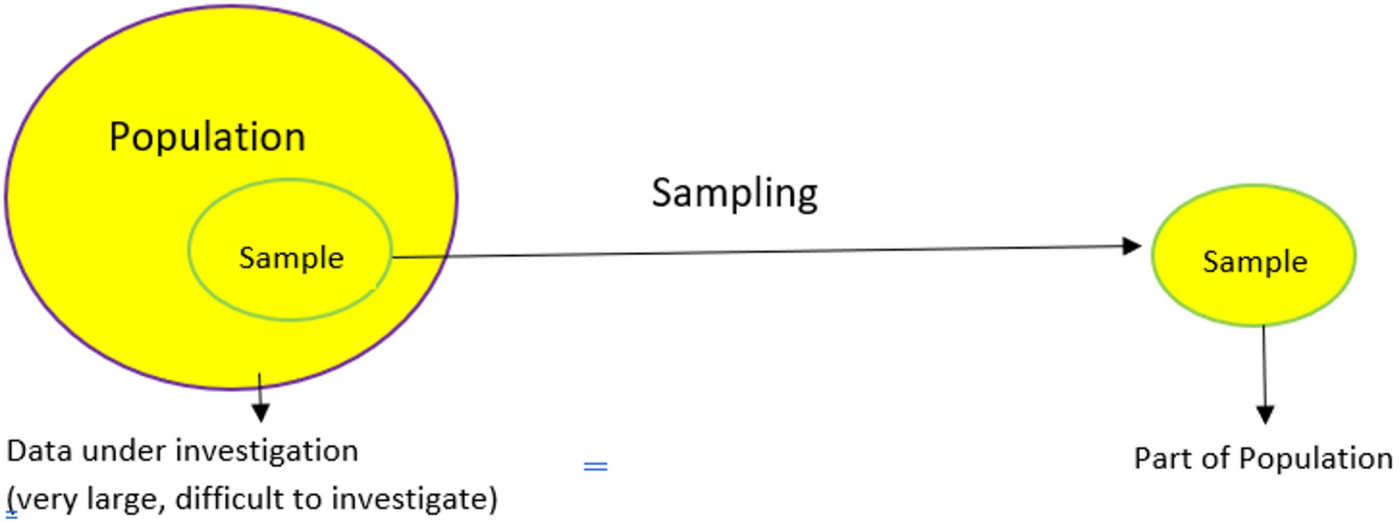
- Data cleaning
- Data integration
- Data sampling
- Data reduction

Major Tasks in Data Preprocessing

- Data cleaning
- Data integration
- Data sampling
- Data reduction

Data Sampling

- Goal: Pick a subset of the population that is a good representation of it
  - Preserve underlying structure and distributions as much as possible

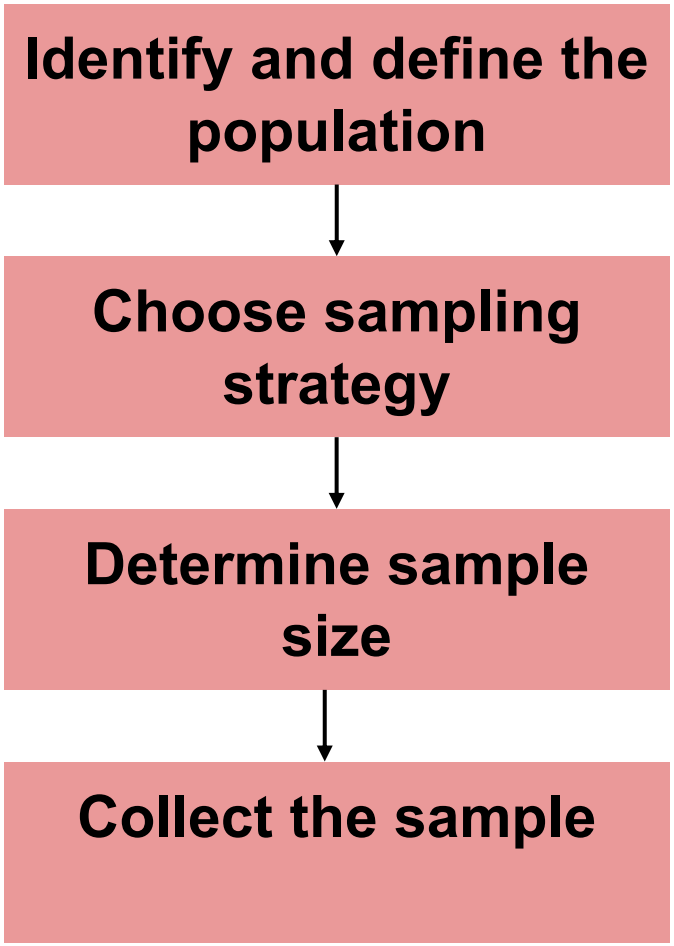


https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

Why Sample?

- Accessing the entire population may not be practical or possible
- Working with a sample is more efficient
  - Faster to perform the selection
  - Faster to perform analysis on a sample
  - Requires less resources (hard disk, memory etc)
- Sample analysis is easier to understand, “debug”, and verify

The Sampling Pipeline



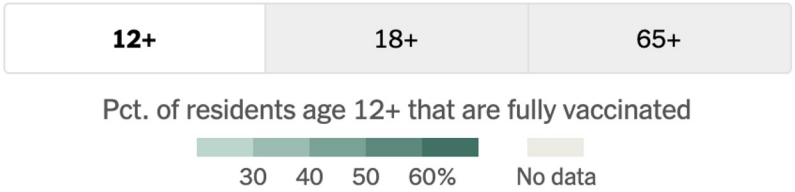
What is the population?

- If starting the analysis with a dataset, one may assume that it is the “population” and sample from it
  - Note that datasets can be understood as samples of larger populations as well
- If collecting the data, what are the population limits?

The New York Times

See How Vaccinations Are Going in Your County and State

Updated Sept. 2, 2021



Each age group can be seen as a population...

Before Sampling Strategies, Selection Bias

- If there is selection bias, the sample is not properly randomized
  - Not representative of the population
- More formally
  - Each instance is characterized by attributes $\{X_1, \ldots, X_N\}$
  - If being in the sample *S* is independent of $X_1, \ldots, X_N$, the sample is **unbiased**: i.e. $P(S|X_1)=P(S), \ldots, P(S|X_N)=P(S)$. Else, the sample is **biased**
  - The distributions for $\{X_1, \ldots, X_N\}$ in the population are preserved in the sample

Selection Bias: Implications

- Selection bias within data affects generalizability of results and potentially the identifiability of model parameters.

- Generalizability:
  - Does your model represent the population at large?
  - Does your prediction match the production results?
  - Is your statistic representative of the greater population?

- Identifiability:
  - Can you accurately learn a model, parameter or statistic given the sample at hand?

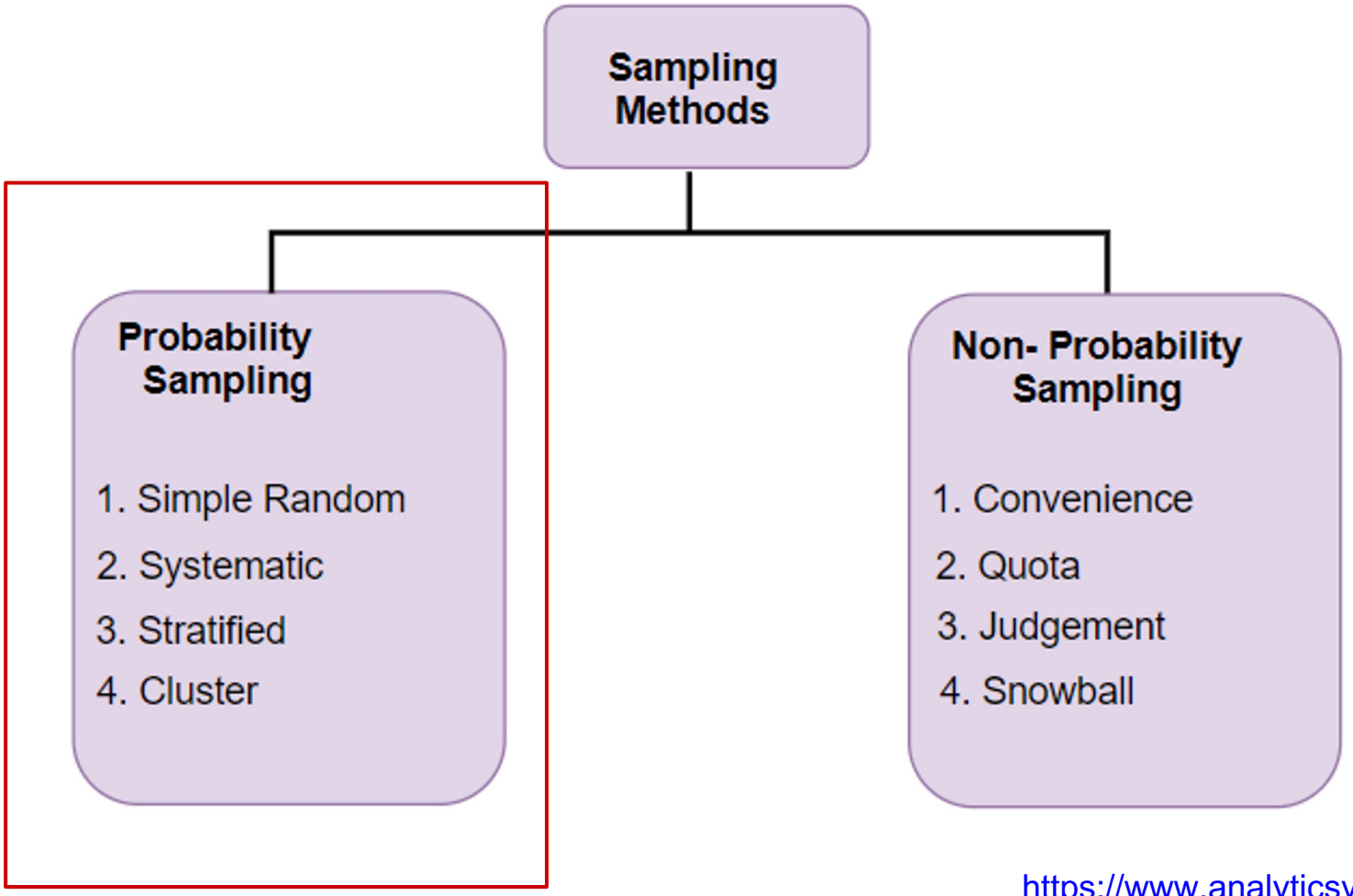
Selection Bias: What to Do

- Avoid It.
  - Design and use random sampling schemes as much as possible (more on this in a second!)

- Adjust It.
  - In many cases you can statistically adjust for selection bias by weighting instances by 1/P(Samp|X)

- Expect It.
  - Whether by design or accident, selection biases are likely to occur. It is always important to anticipate it and prepare for how it might affect your analysis.

Sampling Strategies



https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

Our focus: **probability sampling**

Sampling Strategies: Simple Random

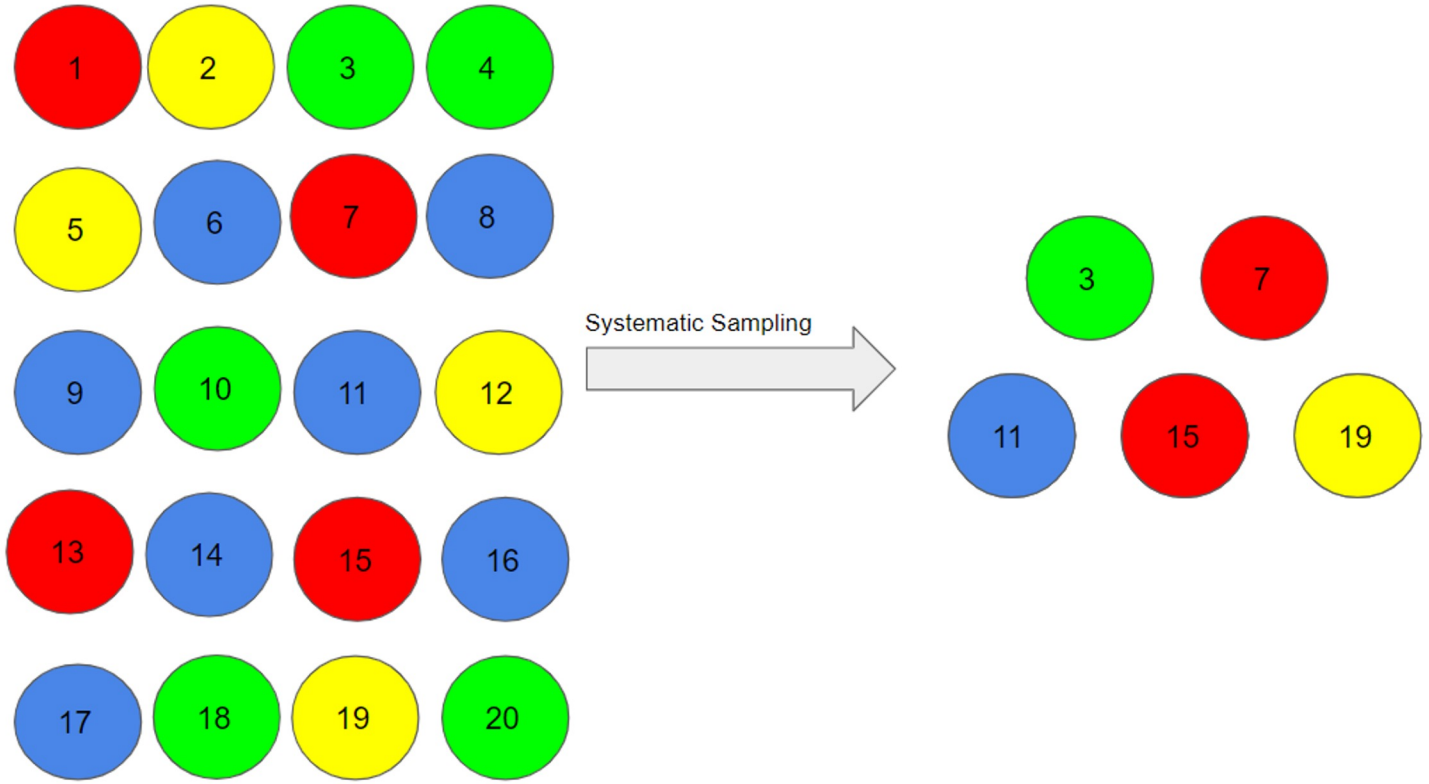
- Each instance of the population has equal chance of being selected
  - Usually, this strategy is ideal
  - Caveat (sometimes): may leave smaller groups or outliers out completely

https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

Sampling Strategies: Systematic

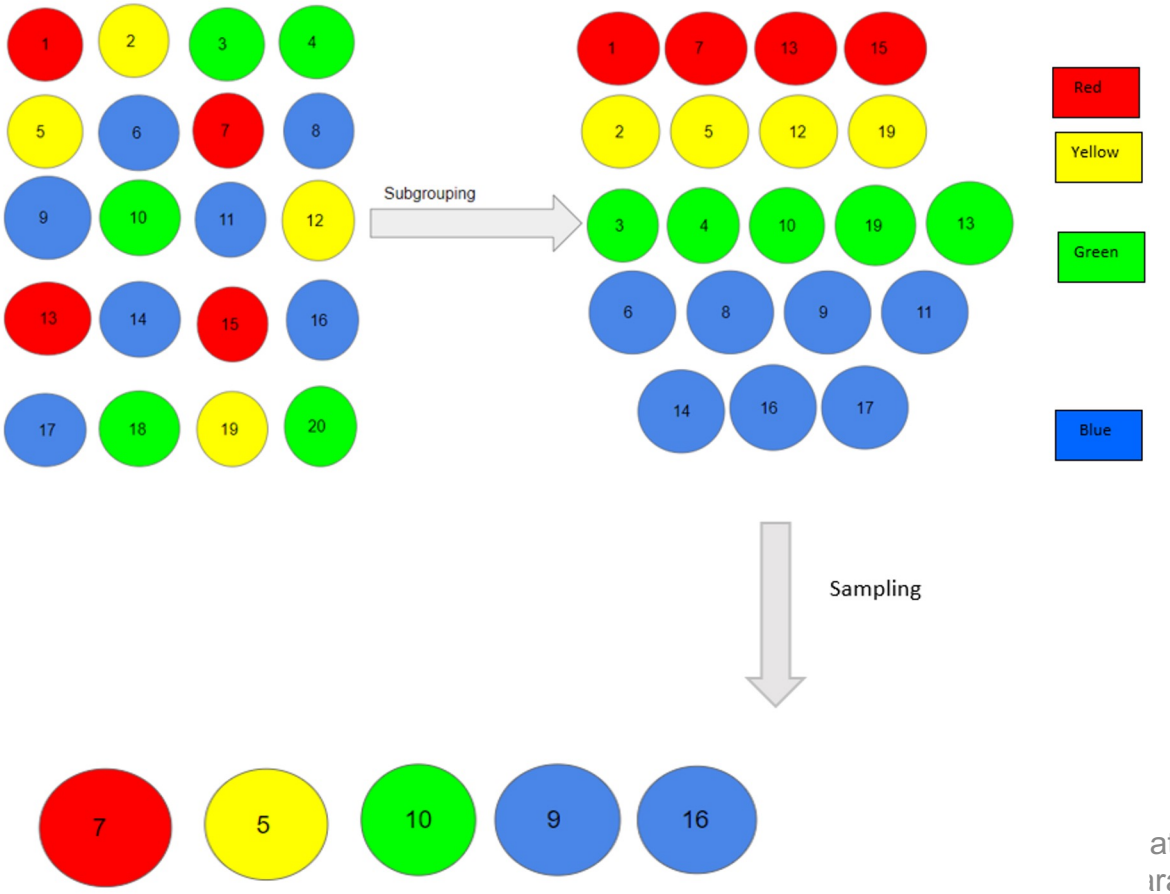
- The first instance is selected randomly and the next are selected using a fixed ‘sampling interval’.
  - Can you think of an application for that?
  - **Is it random?**

https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/



Sampling Strategies: Stratified

- (1) Population is divided into subgroups (e.g., gender, ethnicity) and (2) then instances from these subgroups are selected
- Each subgroup is sampled according to its proportion; within each group, the sampling is **random**



https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/