CSC580: Principles of Data Science

Bias and Fairness

Prof. Jason Pacheco

Data Science Ethics

The movement to hold AI accountable gains more steam

First-in-US NYC law requires algorithms used in hiring to be "audited" for bias.

KHARI JOHNSON, WIRED.COM – 12/5/2021, 6:10 AM

ars technica

[Article Link]

As Data Science / AI / ML become more standard,
we need to address fairness and ethics…

State of Michigan's mistake led to man filing bankruptcy

Paul Egan Detroit Free Press

The secret bias hidden in mortgage-approval algorithms

By EMMANUEL MARTINEZ and LAUREN KIRCHNER/The Markup August 25, 2021

Senators Question Regulators About Tenant Screening Oversight

Facebook's race-blind
around hate speech
e expense of Black
documents show

- Washington Post

ExamSoft's remote bar exam sparks privacy and facial recognition concerns

- Venture Beat

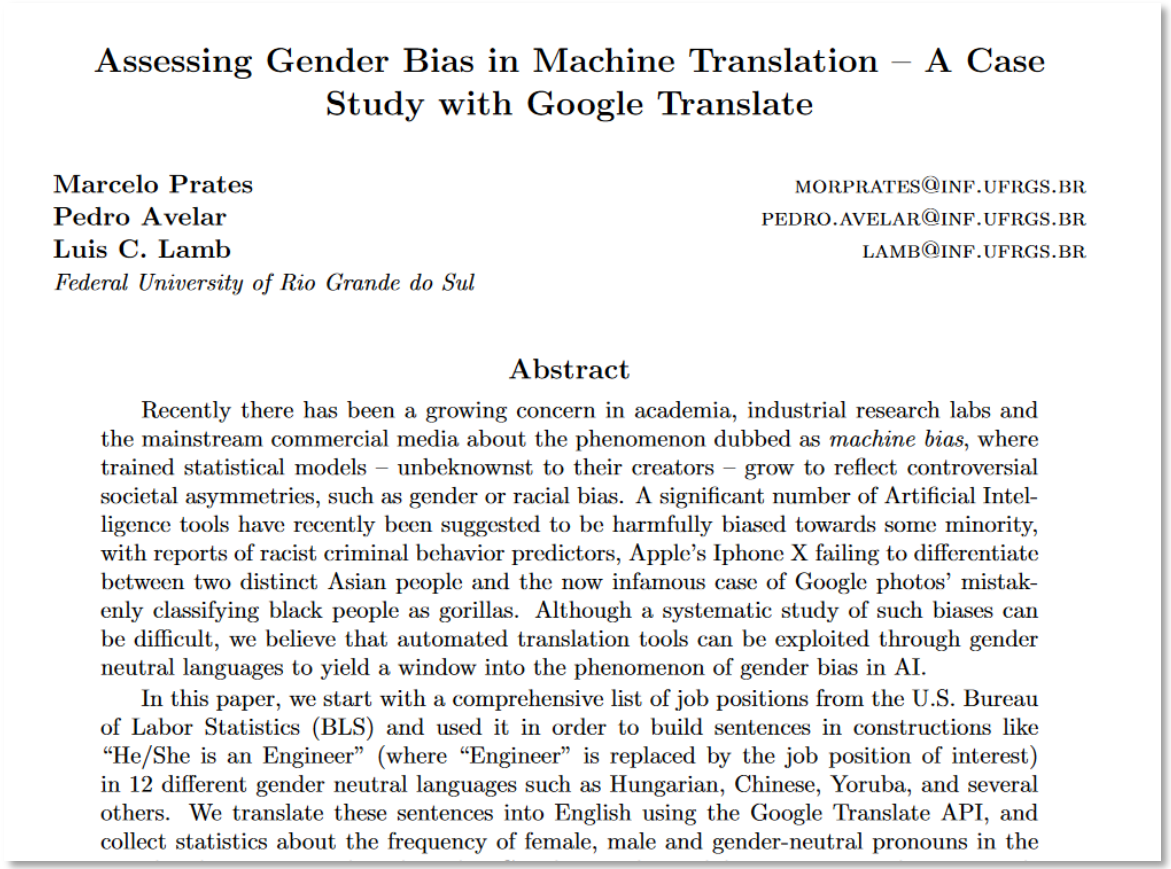
Data Science Ethics

A real-live example of dataset bias…

https://translate.google.com/

Exhibits gender bias in many languages…

…largely the result of using highly-parameterized neural networks with inadequate training data



Assessing Gender Bias in Machine Translation – A Case Study with Google Translate

Marcelo Prates MORPRATES@INF.UFRGS.BR
Pedro Avelar PEDRO.AVELAR@INF.UFRGS.BR
Luis C. Lamb LAMB@INF.UFRGS.BR
Federal University of Rio Grande do Sul

Abstract

Recently there has been a growing concern in academia, industrial research labs and the mainstream commercial media about the phenomenon dubbed as *machine bias*, where trained statistical models – unbeknownst to their creators – grow to reflect controversial societal asymmetries, such as gender or racial bias. A significant number of Artificial Intelligence tools have recently been suggested to be harmfully biased towards some minority, with reports of racist criminal behavior predictors, Apple's Iphone X failing to differentiate between two distinct Asian people and the now infamous case of Google photos' mistakenly classifying black people as gorillas. Although a systematic study of such biases can be difficult, we believe that automated translation tools can be exploited through gender neutral languages to yield a window into the phenomenon of gender bias in AI.

In this paper, we start with a comprehensive list of job positions from the U.S. Bureau of Labor Statistics (BLS) and used it in order to build sentences in constructions like "He/She is an Engineer" (where "Engineer" is replaced by the job position of interest) in 12 different gender neutral languages such as Hungarian, Chinese, Yoruba, and several others. We translate these sentences into English using the Google Translate API, and collect statistics about the frequency of female, male and gender-neutral pronouns in the

Machine Learning / AI Ethics

- NYC adopted law requiring audits of algorithms used in hiring
- White house proposes an AI bill of rights to disclose when AI makes decisions with societal impact
- EU lawmakers require inspection of AI deemed high-risk
- Analysis of automated hiring software found to be biased to appearance, software program used to create resume, accent, or whether applicants have a bookshelf in the background
- Photo ID software works well for white men—black women, not so much

Machine Learning / AI Ethics

Aspects of ethics include…

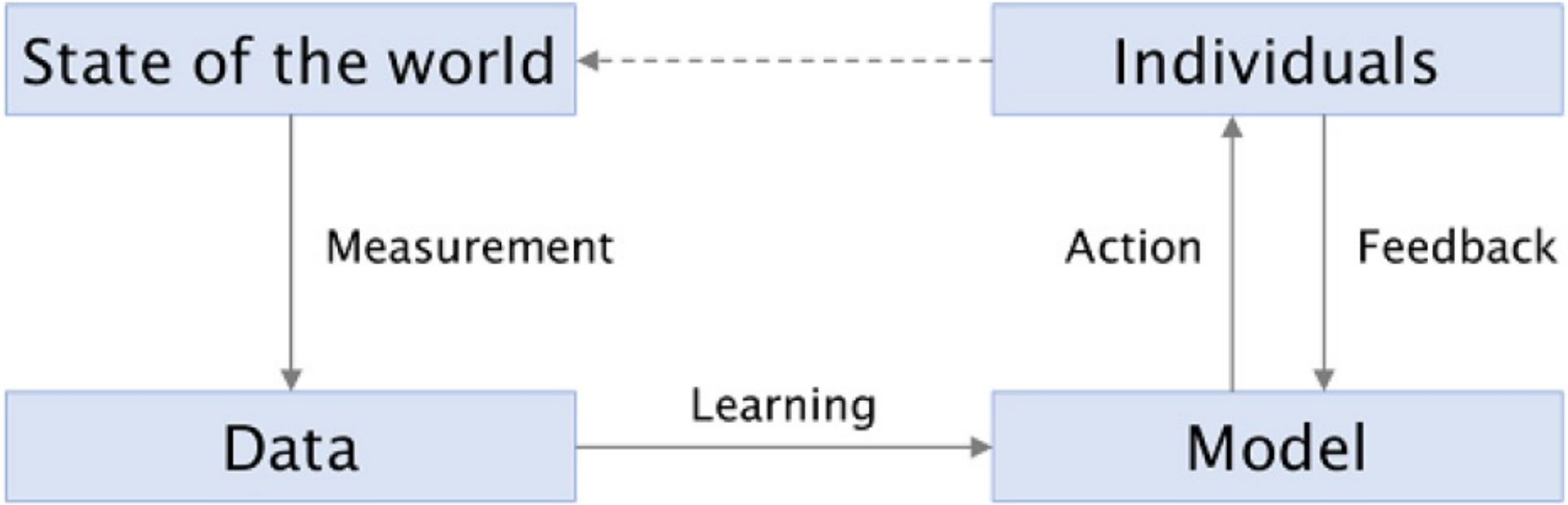
Security Who has access to the data?

Privacy Can data be used to identify individuals?

Fairness Are predictions biased across groups?

Transparency Do users know what they are consenting to? Are model decisions interpretable?

Impacts of Machine Learning / AI



It is rare for data science to *not* involve people in some way

- Of the top 30 recent Kaggle competitions, 14 involve making decisions that directly affect people
- An additional 5 have obvious indirect affect on people
- Only 9 had no obvious impact on people

[Source: Brocas et al. “Fairness and ML”]

Sampling Bias

Occurs if data are collected in a way that some members of the population have lower/higher probability of being sampled than others

Sometimes is unavoidable (e.g. not all members are equally accessible) but it must then be corrected for

Example We conduct a poll by randomly calling numbers in a phone book. People that have less time are less likely to respond. Called *non-response bias.*

Common Types of Sampling Bias

Self-selection Possible whenever members (typically people) under study have control over whether to participate. E.g. online or phone-in poll—user can choose whether to initiate participation.

Exclusion Results from excluding certain groups from the sample. E.g. excluding groups that move in or out of a study area during follow-up.

Survivorship Only *surviving* subjects are selected. Here "surviving" is a loose definition, non-survivors may simple fall out of view. E.g. using record of current companies as indicator of economic climate.

Survivorship Only *surviving* subjects are selected. Here "surviving" is a loose definition, non-survivors may simple fall out of view. E.g. using record of current companies as indicator of economic climate.

Example of Bias in a Simple Random Sample

Simple random sample (SRS) is least prone to bias, but not always…

> You want to study procrastination and social anxiety levels in undergraduate students at your university using a simple random sample. You assign a number to every student in the research participant database from 1 to 1500 and use a random number generator to select 120 numbers.

What is the cause of bias in this simple random sample?

https://www.scribbr.com/methodology/sampling-bias/

Example of Bias in a Simple Random Sample

Simple random sample (SRS) is least prone to bias, but not always…

You want to study procrastination and social anxiety levels in undergraduate students at your university using a simple random sample. You assign a number to every student in the research participant database from 1 to 1500 and use a random number generator to select 120 numbers.

Although you used a random sample, not every member of your target population –undergraduate students at your university – had a chance of being selected. Your sample misses anyone who did not sign up to be contacted about participating in research. This may bias your sample towards people who have less social anxiety and are more willing to participate in research.

https://www.scribbr.com/methodology/sampling-bias/

Sampling Methods

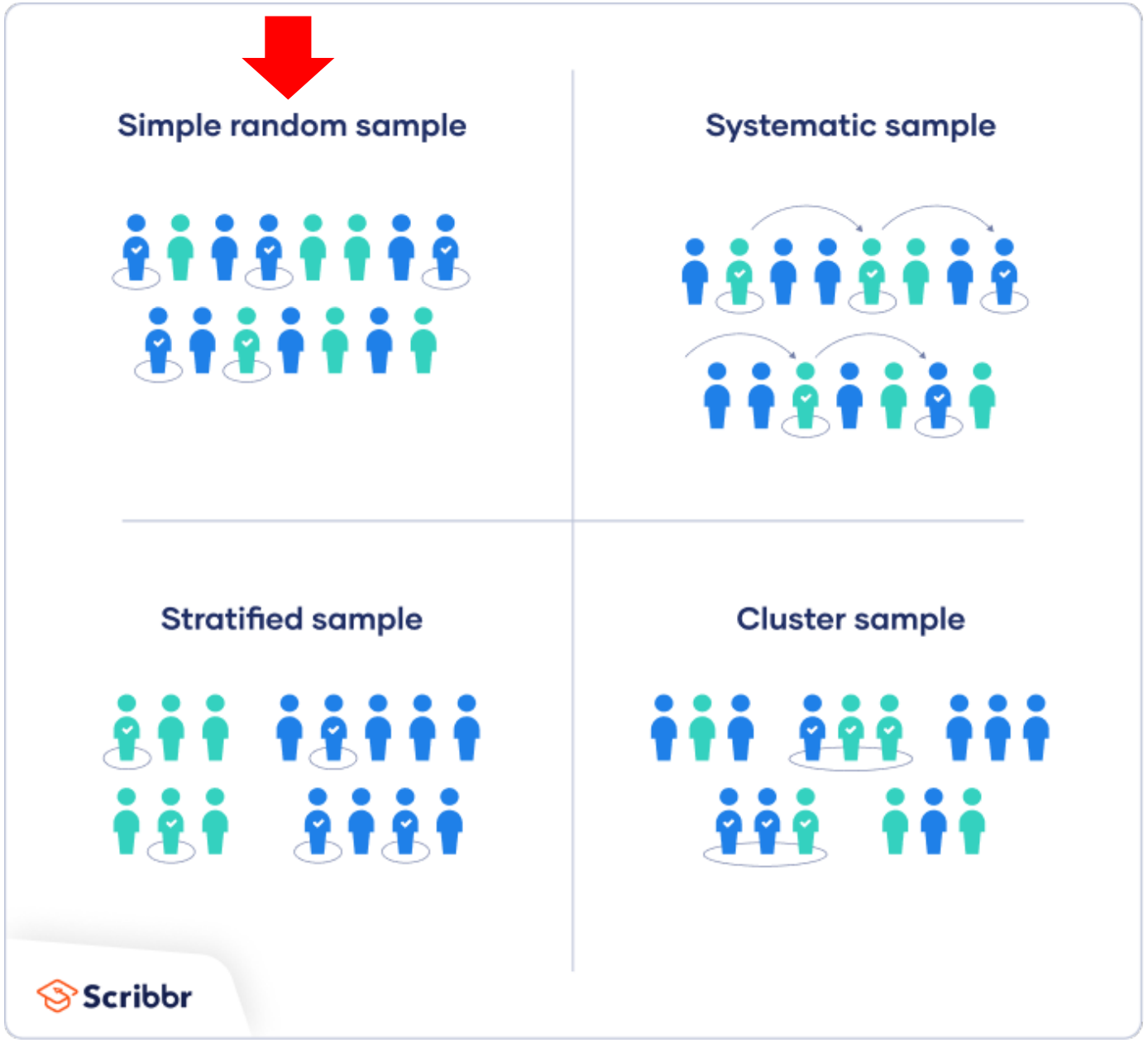
Sampling must be conducted properly, to avoid sample bias

Two primary types of sampling…

Probability Sampling Random selection allowing strong statistical inferences about the population

Non-Probability Sampling Based on convenience or other criteria to easily collect data (but no random sampling)

Probability Sampling



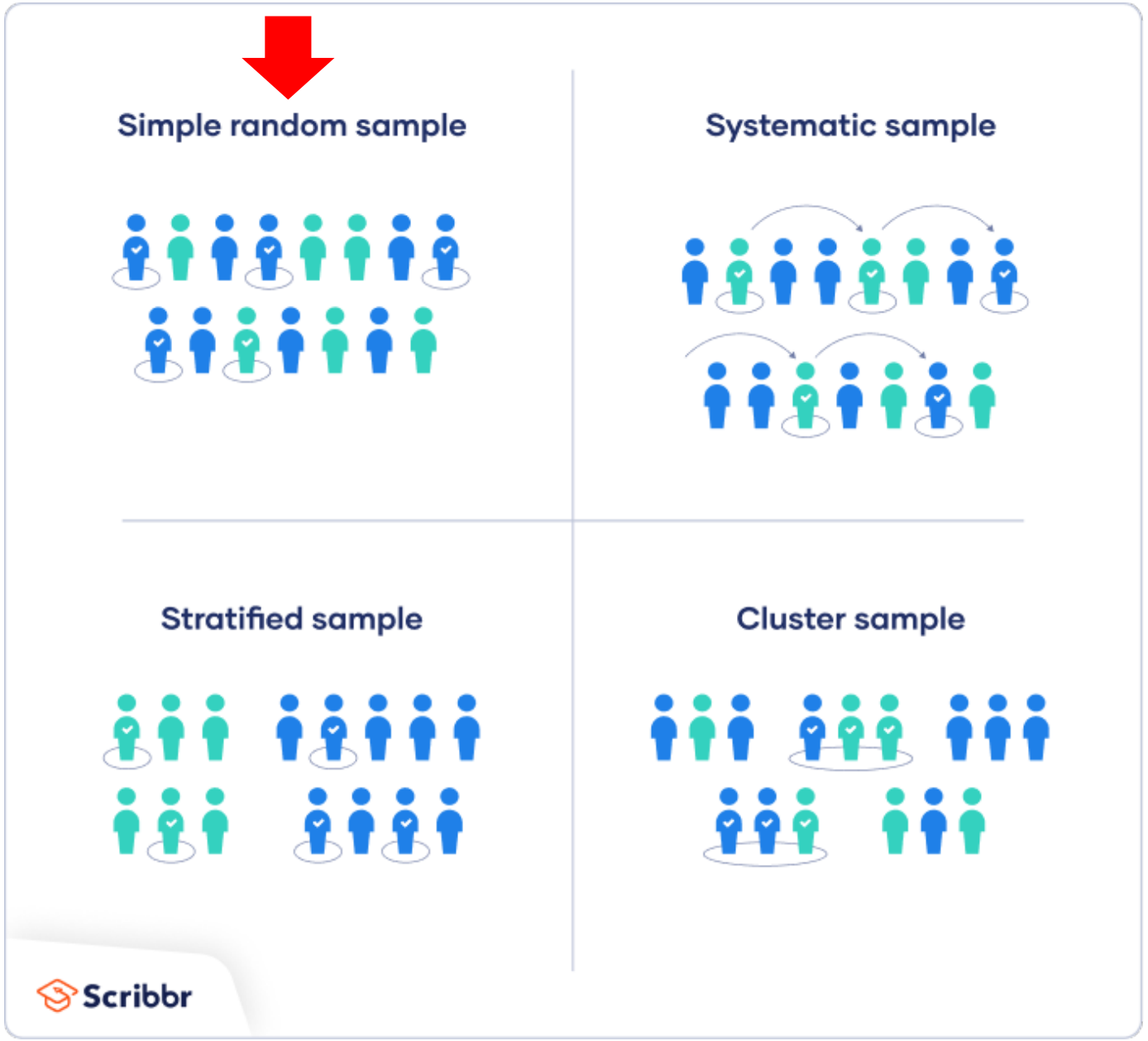
Simple Random Sample (SRS)

Each member of the population has the *same chance* of being selected (i.e. uniform over the population)

Example : American Community Survey (ACS)

Each year the US Census Bureau use simple random sampling to select individuals in the US. They follow those individuals for 1 year to draw conclusions about the US population as a whole.

Probability Sampling

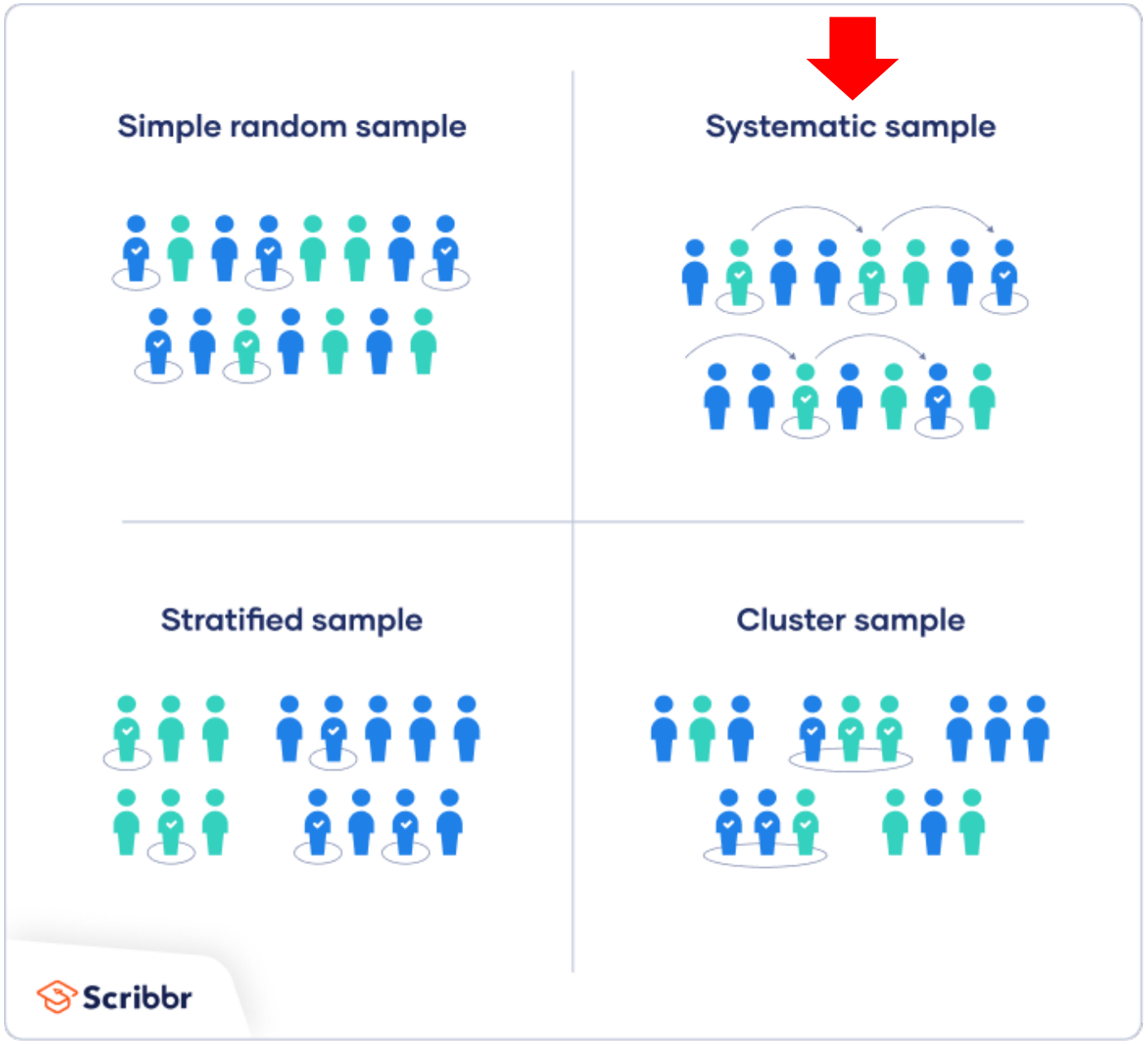


Simple Random Sample (SRS)

Each member of the population has the *same chance* of being selected (i.e. uniform over the population)

- Most straightforward probability sampling method
- Impractical unless you have a complete list of every member of population

Probability Sampling



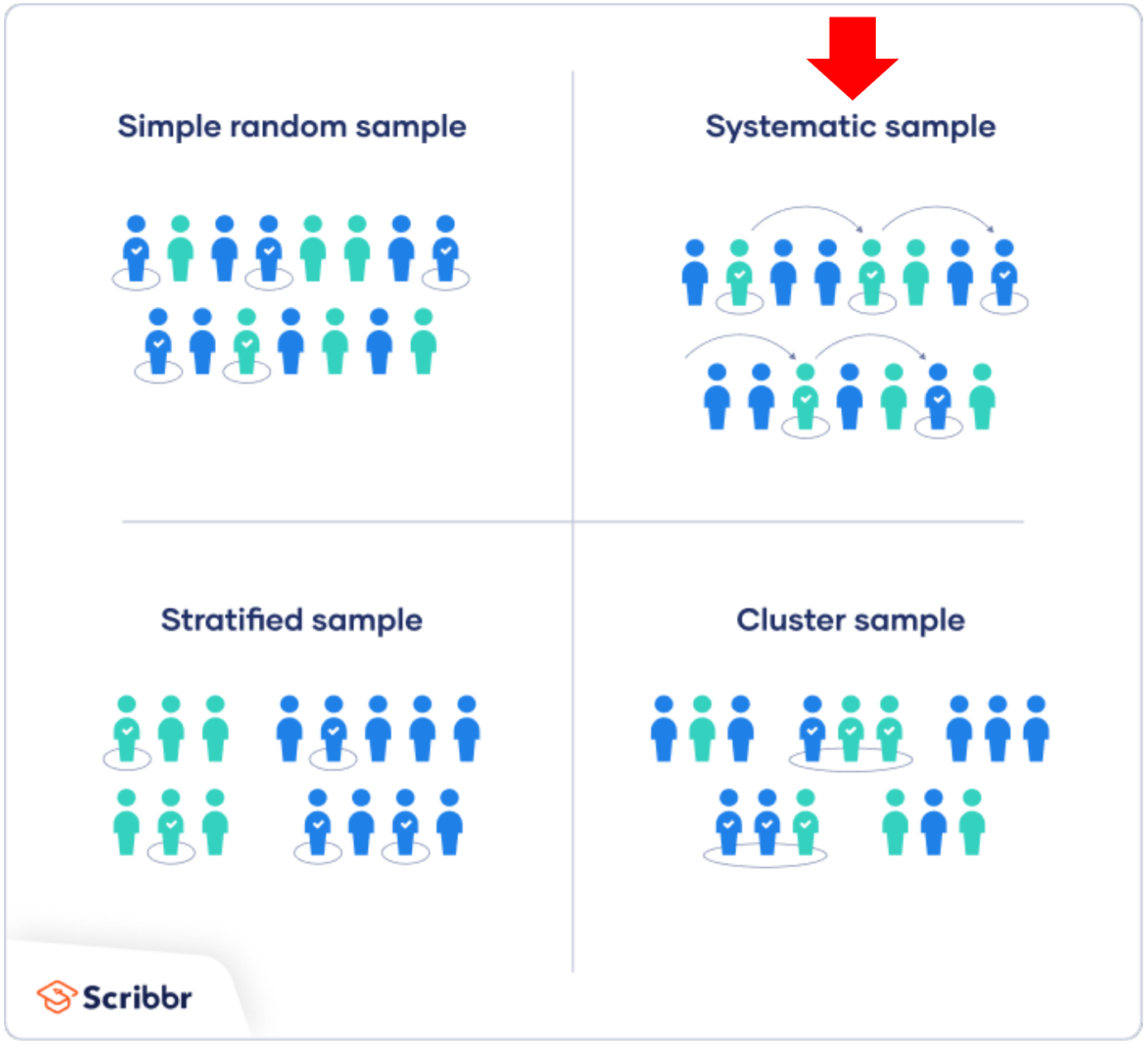
Systematic Sample

Select members of population at a regular interval, determined in advance

Example You own a grocery store and want to study customer satisfaction. You ask *every 20th customer* at checkout about their level of satisfaction.

Note We cannot itemize the whole population in this example, so SRS is not possible.

Probability Sampling



Systematic Sample

Select members of population at a regular interval, determined in advance

- Imitates SRS but is easier in practice
- Can even do systematic sampling when you can't access the entire population in advance
- **Do not** use when population is ordered

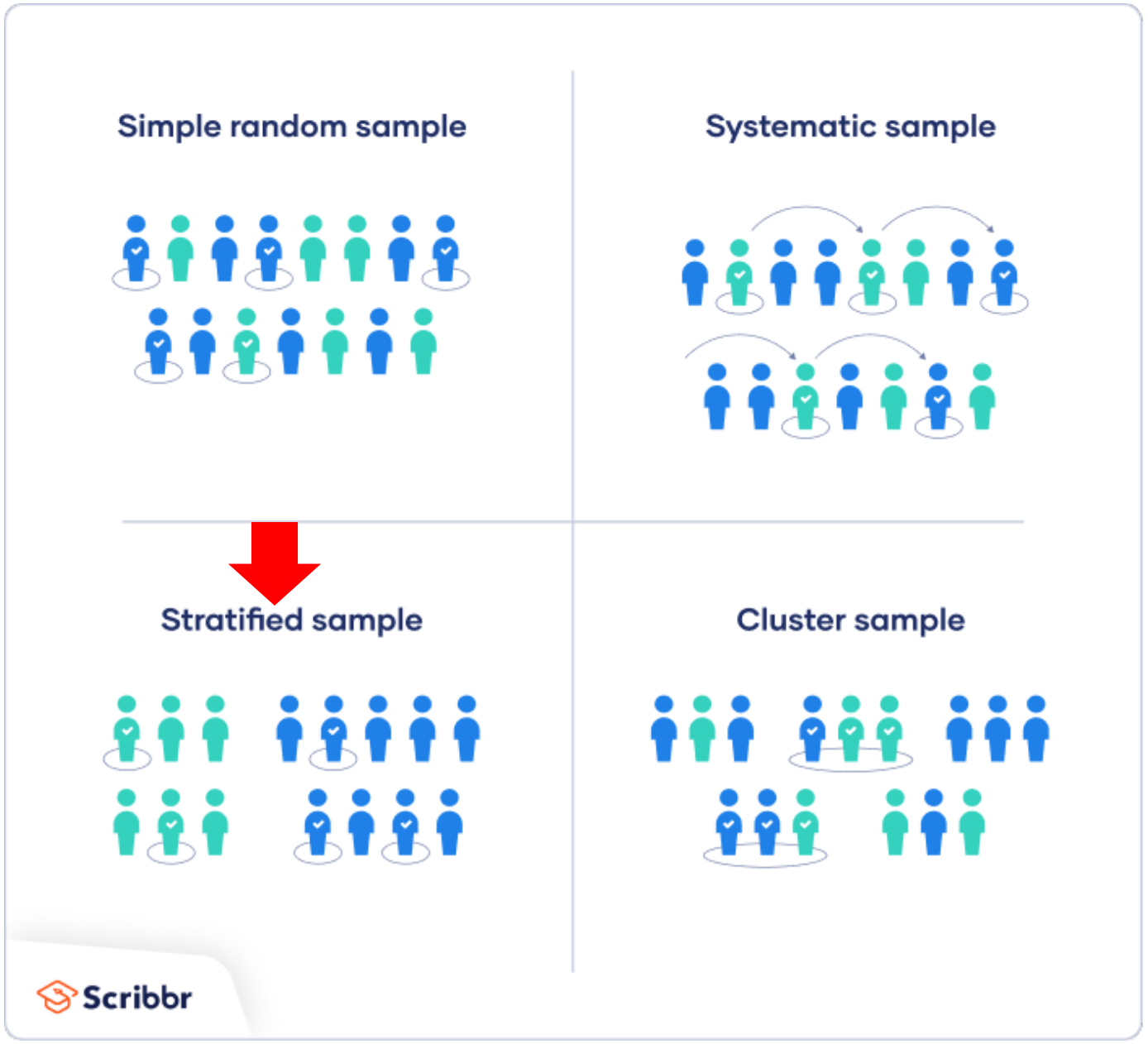
Simple vs. Systematic Random Sample

Consider a school with 1,000 students and suppose we want to select 100 for data collection…

Simple Random Sample Place all names in a bucket and draw 100 randomly. Each student has a 10% chance of being selected.

Systematic Random Sample Any systematic pattern may produce a random sample. One possibility: assume students have ID numbers from 1 to 1,000 and we choose a random starting point (e.g. 533). We pick every 10^{th} name thereafter, in a circular fashion (i.e. wraparound at 1,000). Note: students {3,13,23,…,993} have nonzero probability of being selected, whereas students outside this set have zero probability.

Probability Sampling

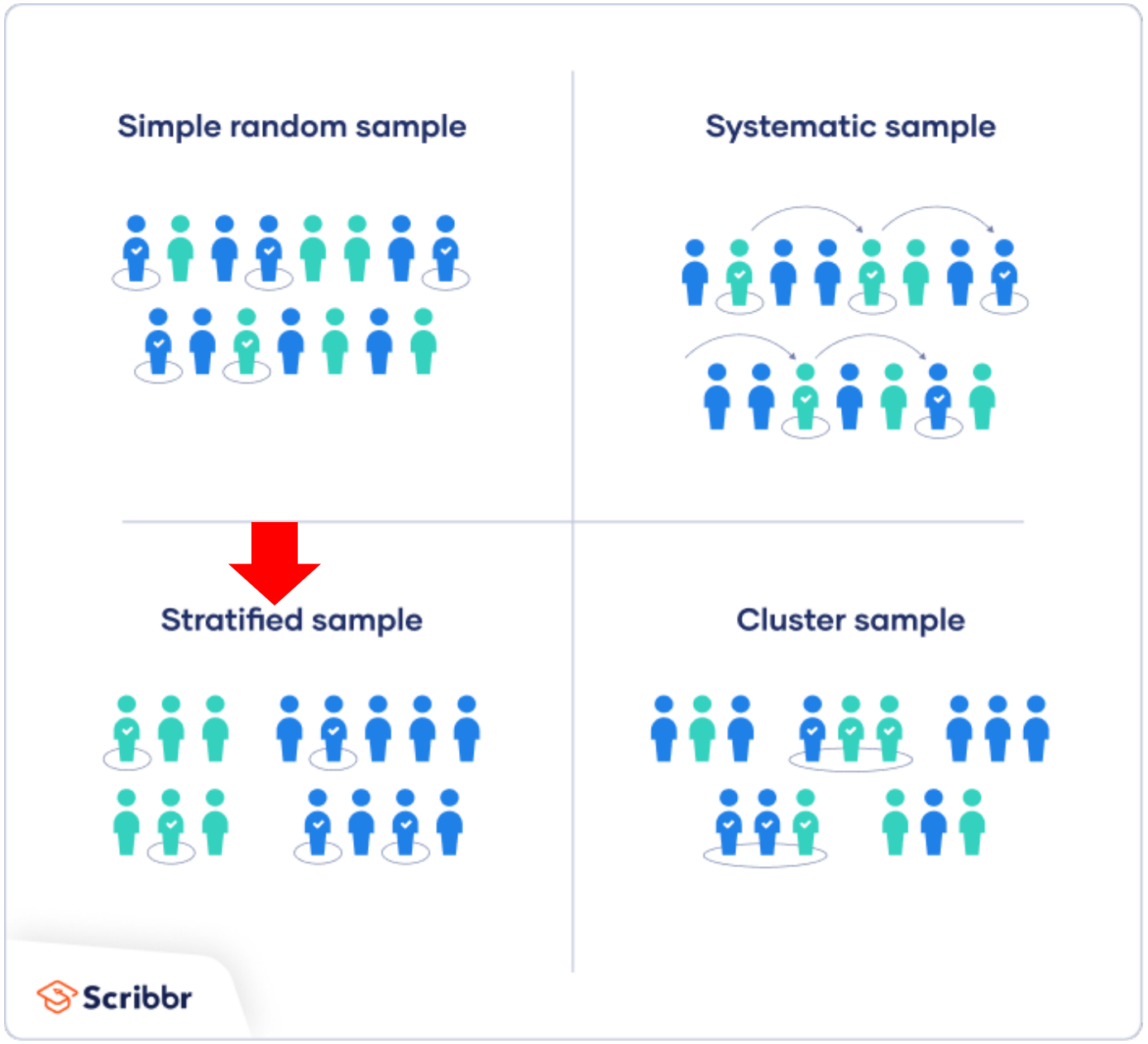


Stratified Sample

Divide population into *homogeneous* subpopulations (strata). Probability sample the strata.

Example We wish to solicit opinions of UA CS freshman, but they are about 18% women. SRS will fail to capture adequate proportion of women. We divide into women / non-women and perform SRS within each group.

Probability Sampling

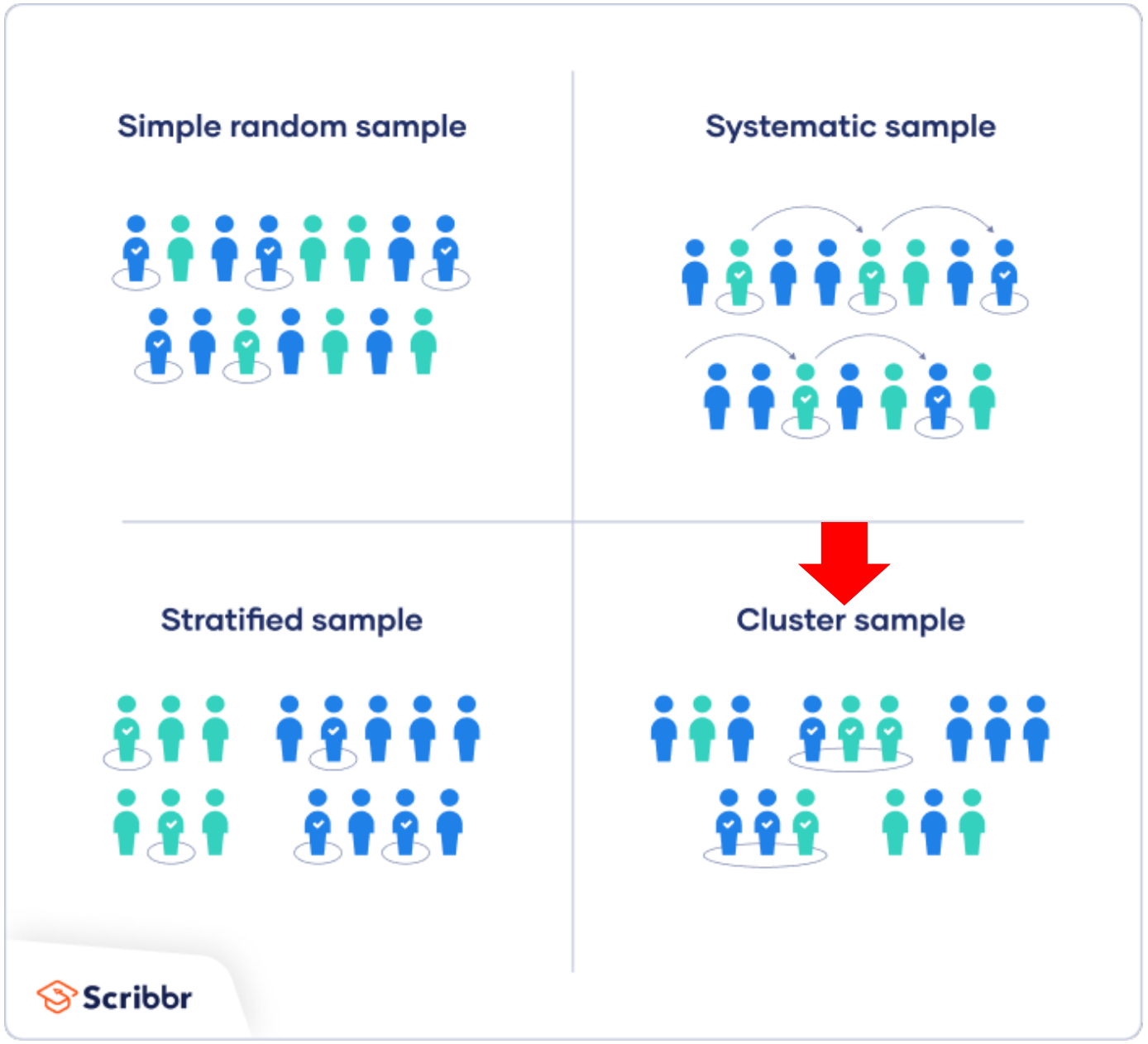


Stratified Sample

Divide population into *homogeneous* subpopulations (strata). Probability sample the strata.

- Use when population is diverse and want to accurately capture characteristic of each group
- Ensures similar variance across subgroups
- Lowers overall variance in the population

Probability Sampling

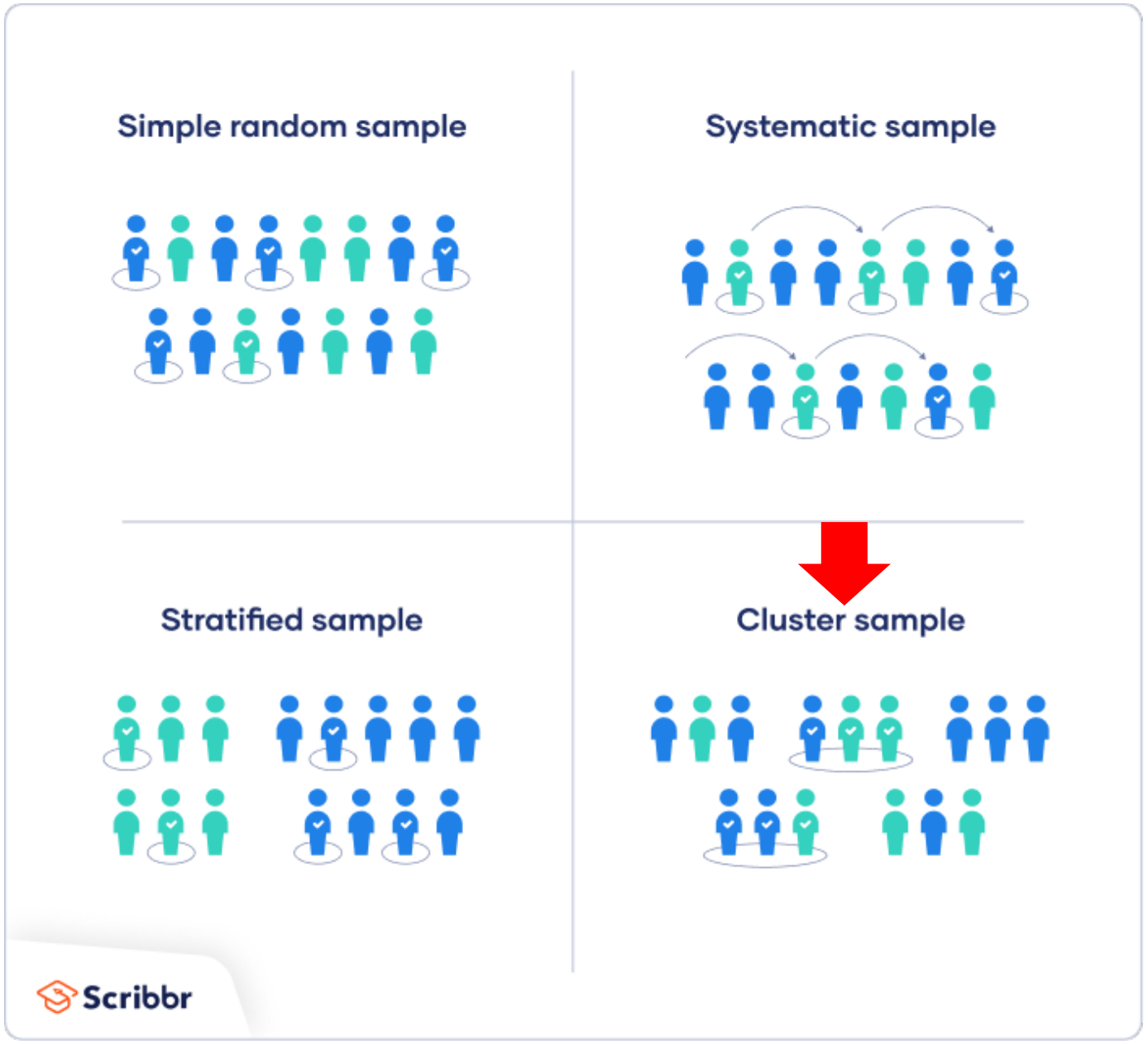


Cluster Sample

Divide population into subgroups (clusters). Randomly select entire clusters.

Example We wish to study the average reading level of *all 7th graders in the city* (population). Create a list of all schools (clusters) then randomly select a subset of schools and test every student.

Probability Sampling

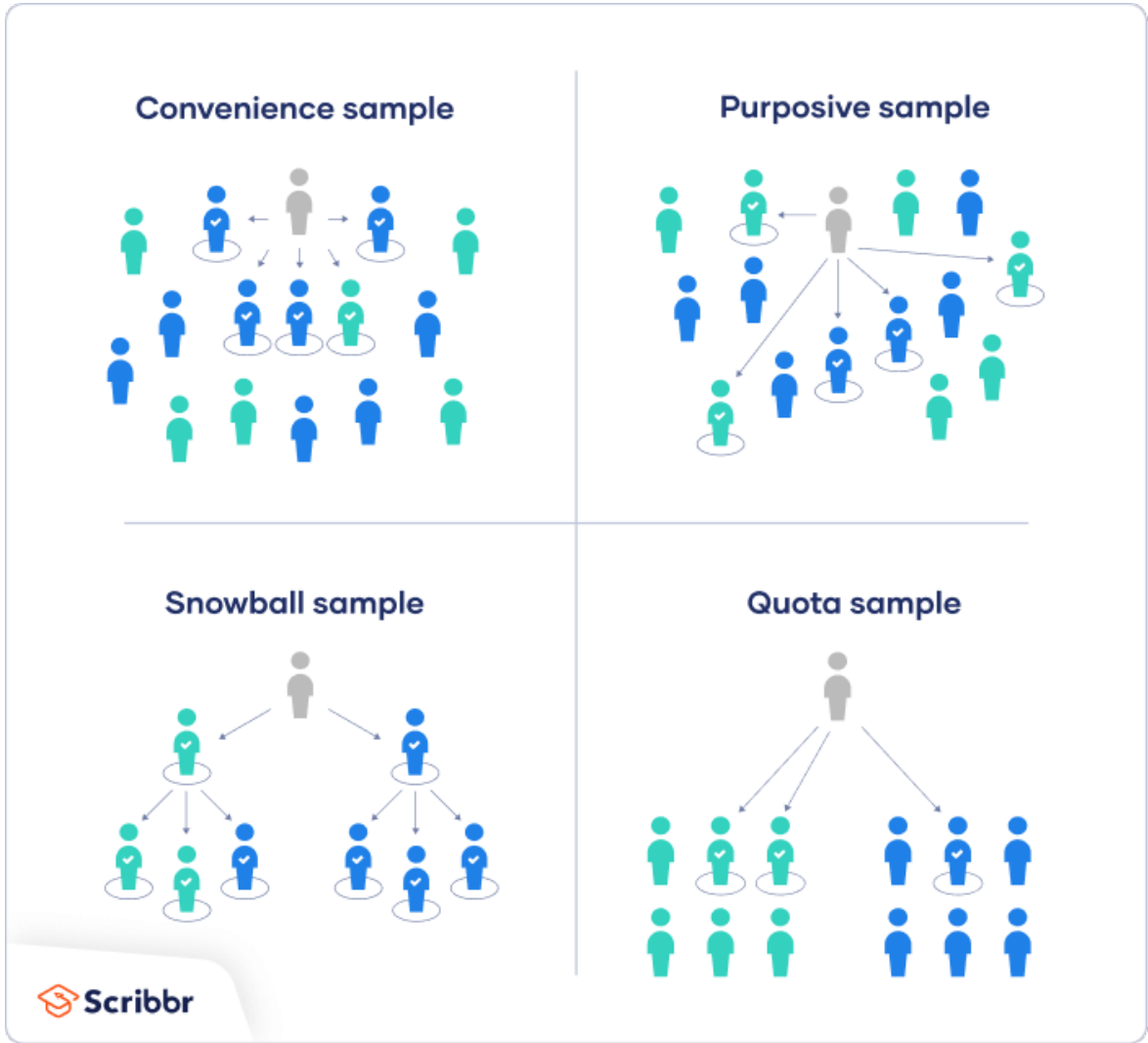


Cluster Sample

Divide population into subgroups (clusters). Randomly select entire clusters.

- This is *single-stage* cluster sampling
- *Multi-stage* avoids sampling every member of a group
- Related to stratified sampling, but groups are not homogeneous

Non-Probability Sampling



Easier to access data, but higher risk of *sample bias* compared to probability sampling

Usually used to perform *qualitative research* (e.g. gathering student opinions, experiences, etc.)

We will not focus on these, but you should be aware if your data are from non-probability methods

Data Science Fairness

Fairness issues can arise from biases in the data…

- Are there observable biases in the data?
- Can we correct for them?

- Differences in the distributions of training / test data?
- Can we detect these differences and avoid / correct them?

Training data reflect disparities, distortions, and biases from the real world and measurement process…

For each model a data scientist should ask… Does learning the model preserve, mitigate, or exacerbate these disparities?

Example Machine translation “She is a doctor” reverse translates to “He is a doctor” in many languages due to data biases.

Data Science Fairness

Example We are building a system to screen mortgage applications. Suppose we collect training data from two demographic groups: 85% White and 15% Black

- Predictive accuracy on the held-out validation set is 95%
- Only 5% error
- Should we sign off on the system as good?

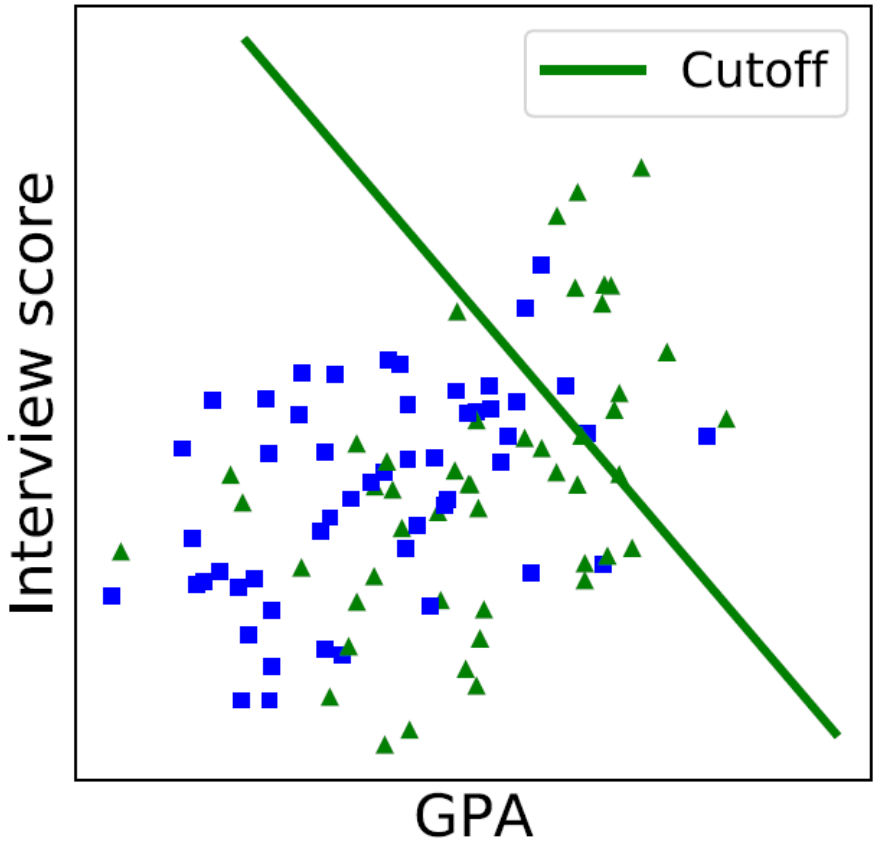
With 5% total error can have up to 66% error on the underrepresented group. Need to report error by each group and aim for 95% accuracy in any group.

Data Science Fairness

Example You are building a system for college admissions based on GPA and interview score (obviously a toy example)

- Fit a least squares regression model
- Model does not account for two demographic groups (blue / green)
- Does this make it fair? (fairness-as-blindness)

Admission rate much lower for blue cohort

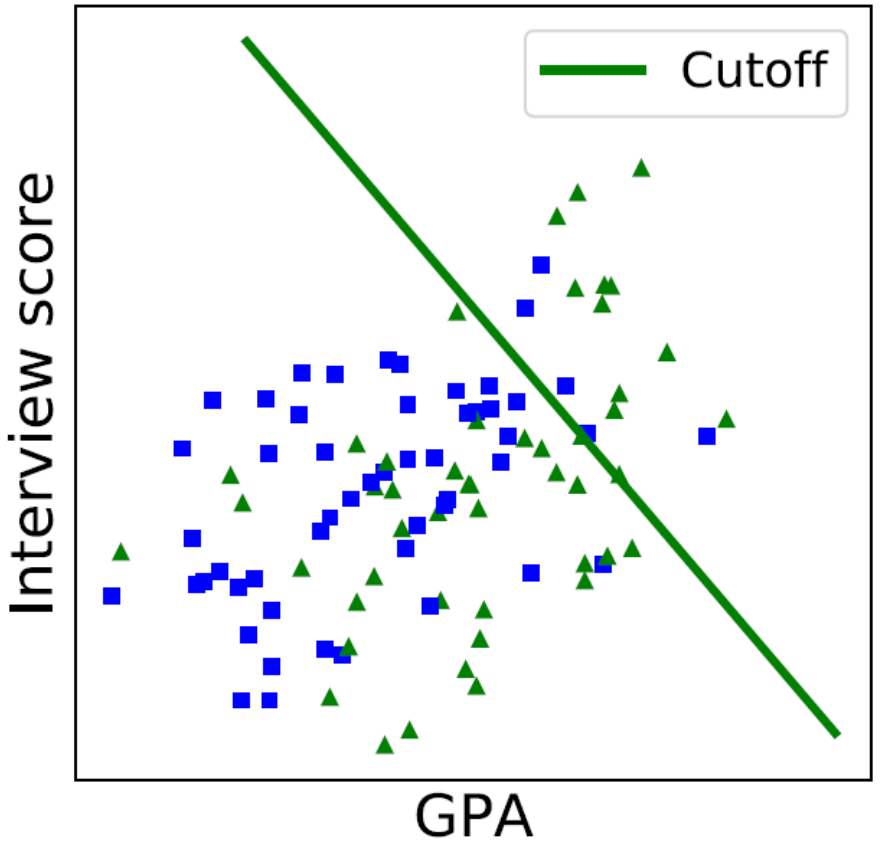


[Source: Brocas et al. “Fairness and ML”]

Data Science Fairness

How to address this behavior?

- GPA correlates with group—omit it as a predictor?
  - Would dramatically impact accuracy
- Pick separate cutoffs (fit separate model) for each group
  - No longer blind to demographics
  - What is the goal for picking cutoffs? Same admission rates?
- Could optimize for diversity among selected candidates
  - Measuring similarity is non-trivial



[Source: Brocas et al. “Fairness and ML”]

Classification Fairness Criteria

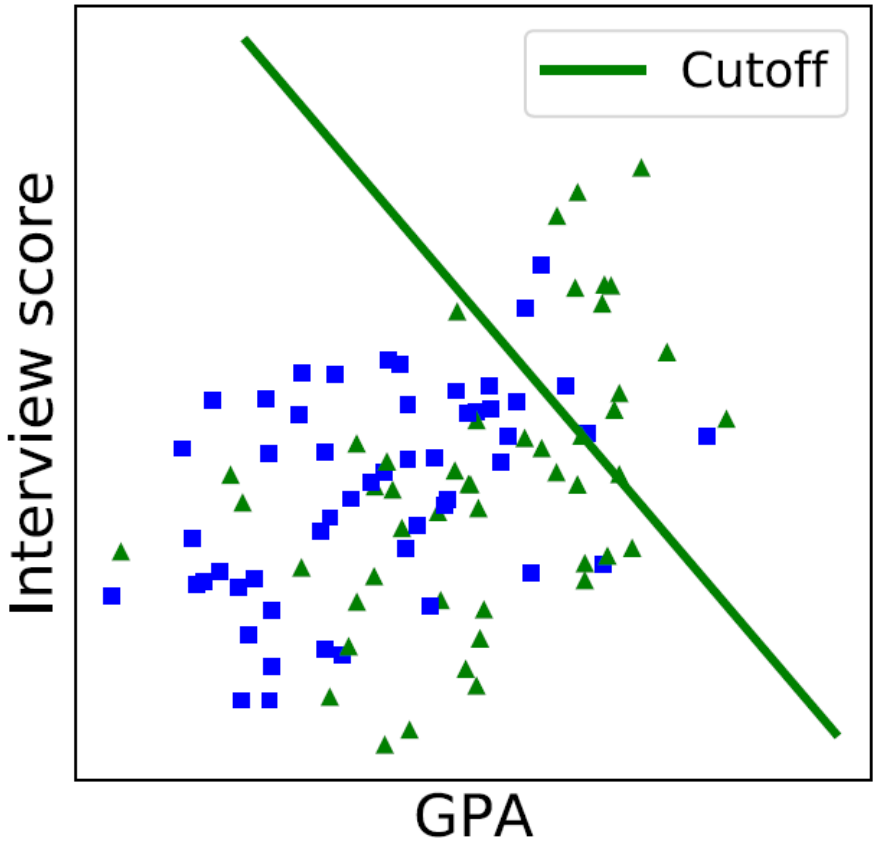
Let *A* be a sensitive attribute, target variable *Y*, and classifier prediction *R.*

Example In our admissions case,

A : Demographic group

R : Prediction of admission

Y : Actual acceptance outcome



Classification Fairness Criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Independence The prediction and attribute are independent

Example The probability of predicting admission doesn't differ across demographic groups,

$$P(R \mid A = a) = P(R \mid A = b)$$

Demographic parity, statistical parity, group fairness, disparate impact

Classification Fairness Criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Separation Score and attribute are conditionally independent, given the classifier decision

Example There is no relationship between prediction and attribute within accepted / non-accepted groups,

$$P(R \mid Y = 1, A = a) = P(R \mid Y = 1, A = b)$$

$$P(R \mid Y = 0, A = a) = P(R \mid Y = 0, A = b)$$

Classification Fairness Criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Sufficiency Outcome and attribute are independent given the model prediction

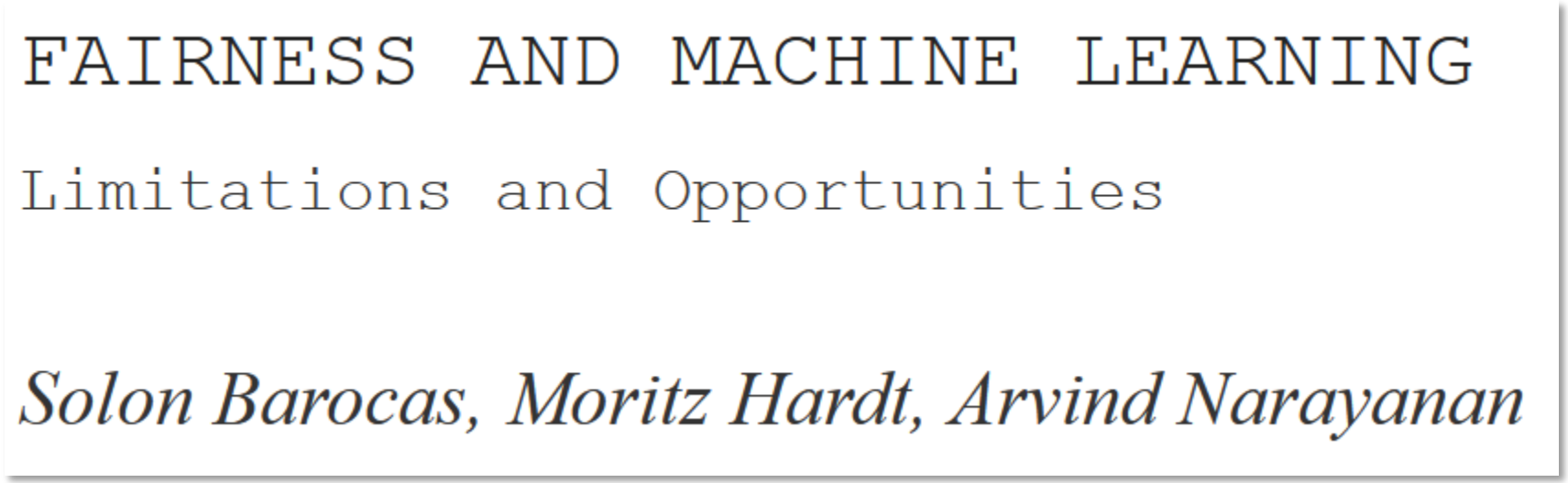
Example There is no relationship between whether someone is admitted and their demographic group within predictions

$$P(Y \mid R = 1, A = a) = P(Y \mid R = 1, A = b)$$

$$P(Y \mid R = 0, A = a) = P(Y \mid R = 0, A = b)$$

Data Science Fairness

In short… there is a lot to say on ethics and fairness… and much can be quantified rigorously…



https://fairmlbook.org/