Responsible Data Science

IN-STK5000 and 9000

Autumn 2025 - Lecture 1

Ana Ozaki

Universitetet i Oslo

Outline

1. Overview
2. Basic Notions
3. Fairness
4. Wrap-up

Overview

Lectures

- Ana Ozaki

Group Sessions

- Emil Poiesz

Overview

Weekly sessions

- Lectures: Thursdays 10:15 - 12:00 (OJD, Seminarrom Pascal)
- Group Sessions: Fridays 12:15 - 14:00 (Starts next week! OJD, Seminar room: Java)

Evaluation

- Two projects: coordinated with course content
- Groups of 3 to 4 students
- Each project counts as 35% of the final grade
- Final exam (30%) December 17 at 3:00 PM (4 hours)
- Bonus points (each point 1%): exercises during lectures will count as bonus, help to round up the final grade.

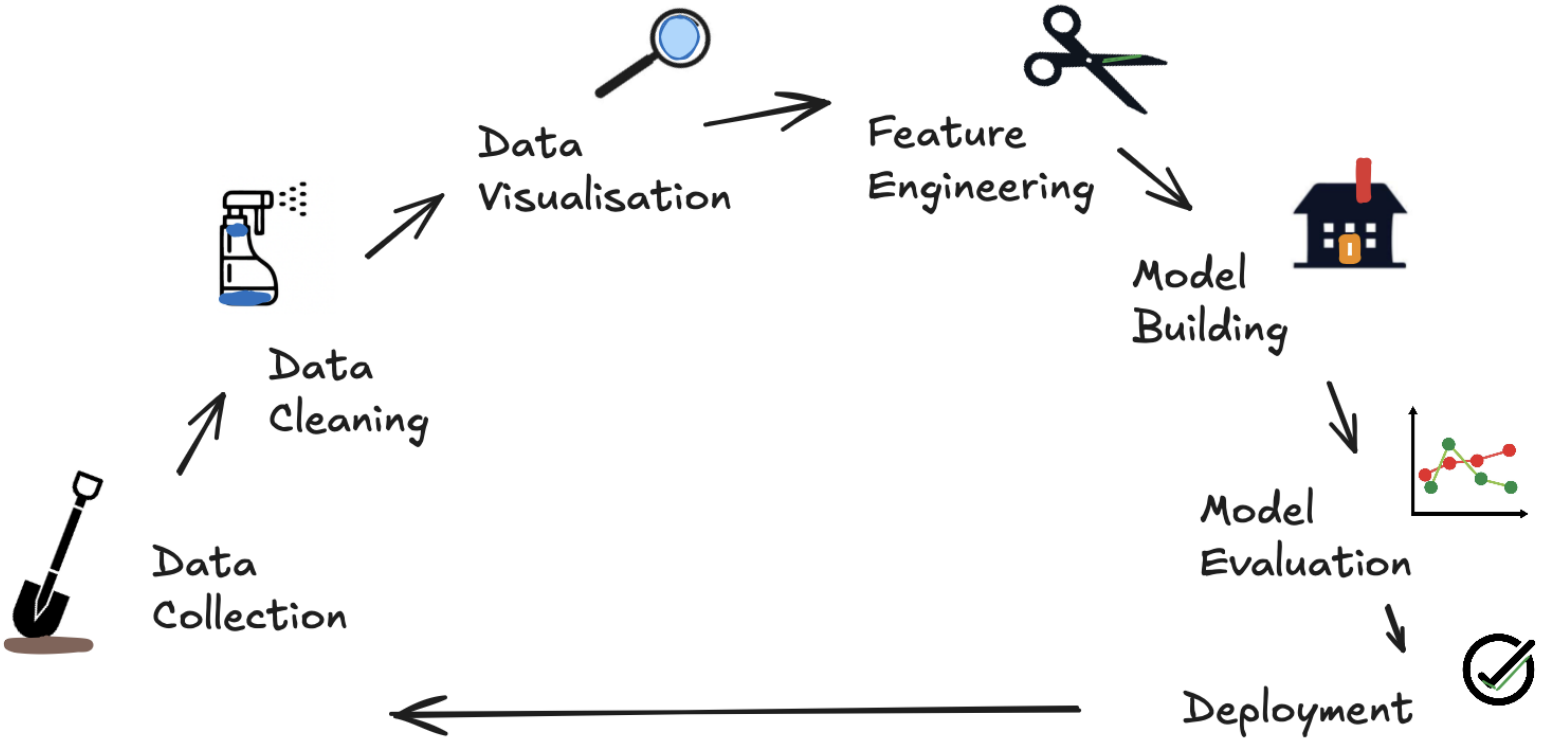
What is Data Science?

- **Data Science** is an interdisciplinary field focused on transforming data into valuable information that can be used for various purposes such as making predictions, indicating trends.
- It involves methods from **probability and statistics**, **machine learning**, and **domain-specific knowledge**.

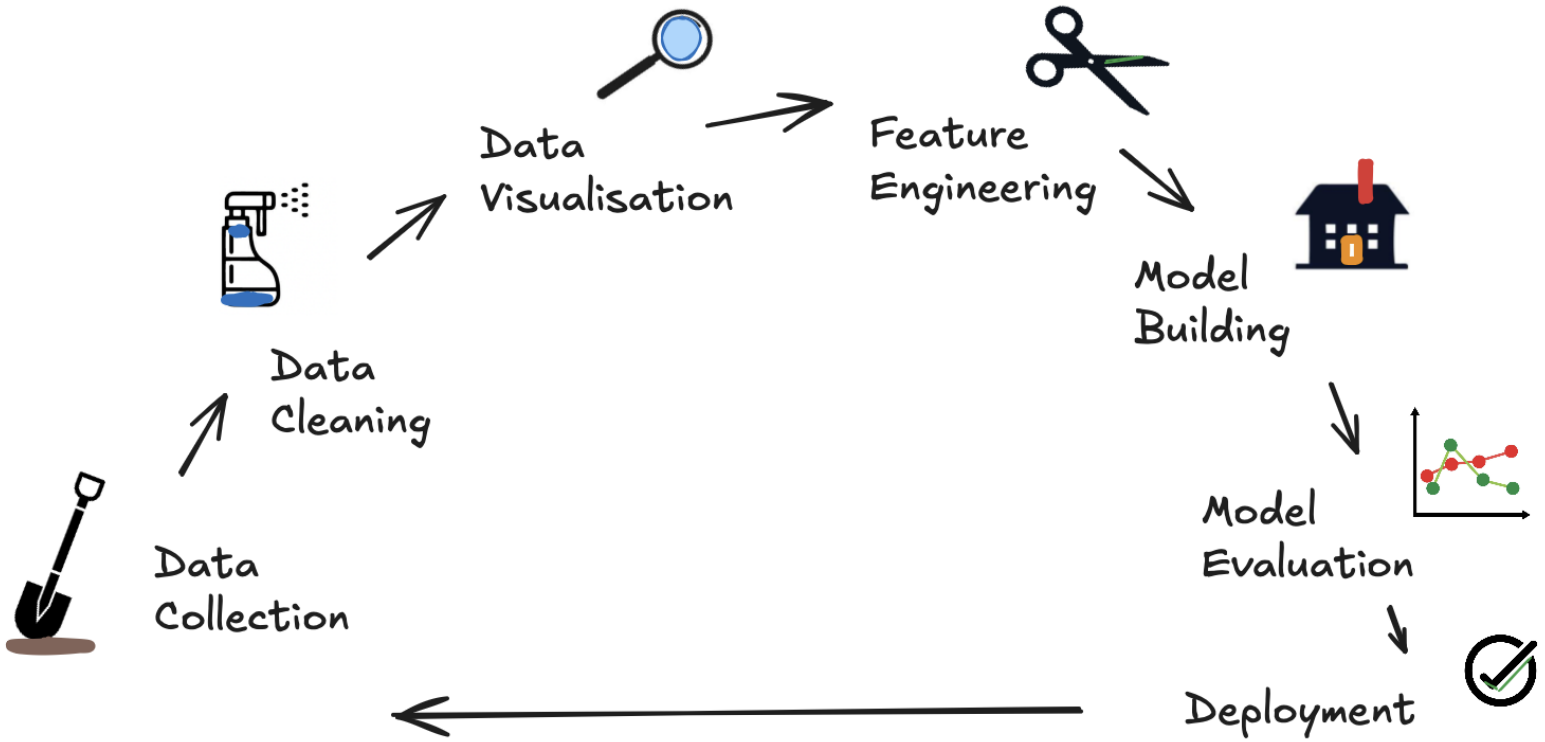
Applications of Data Science

- Data Science has been heavily applied in the industry (medical domain, marketing, etc), public sector, and academia.

Data Science Pipeline

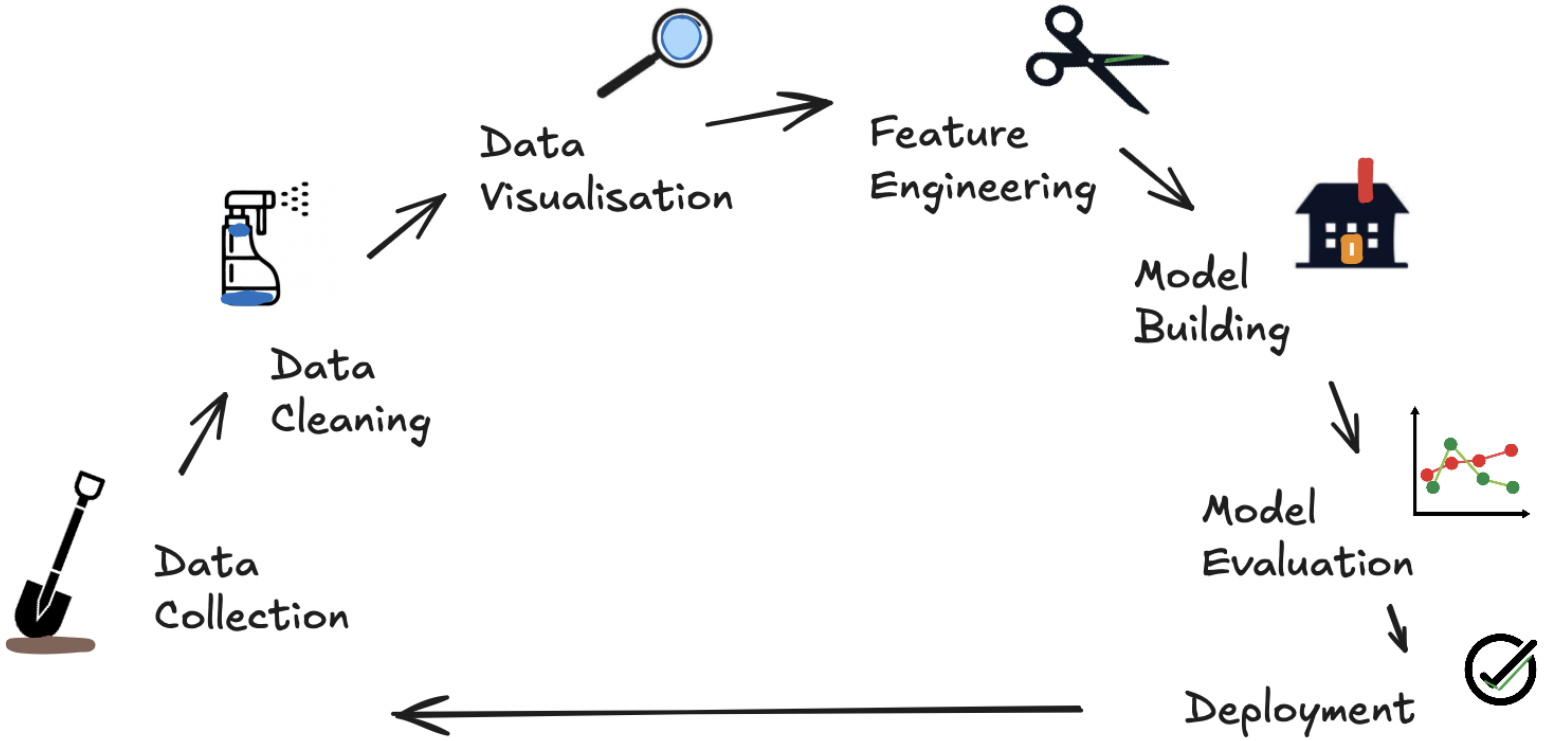


Data Science Pipeline



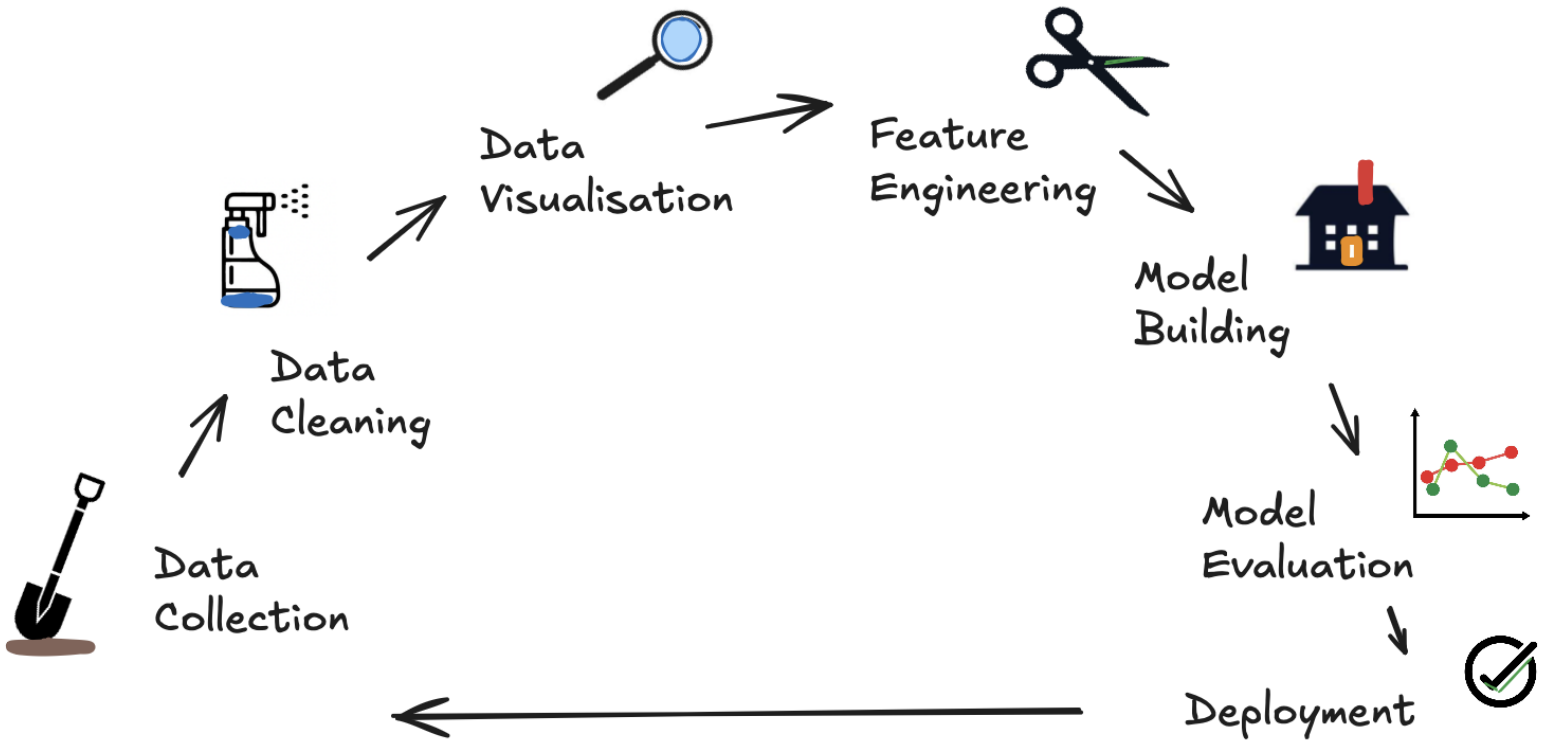
Data Collection: raw data from various sources, which can include databases, sensors, web scraping etc.

Data Science Pipeline



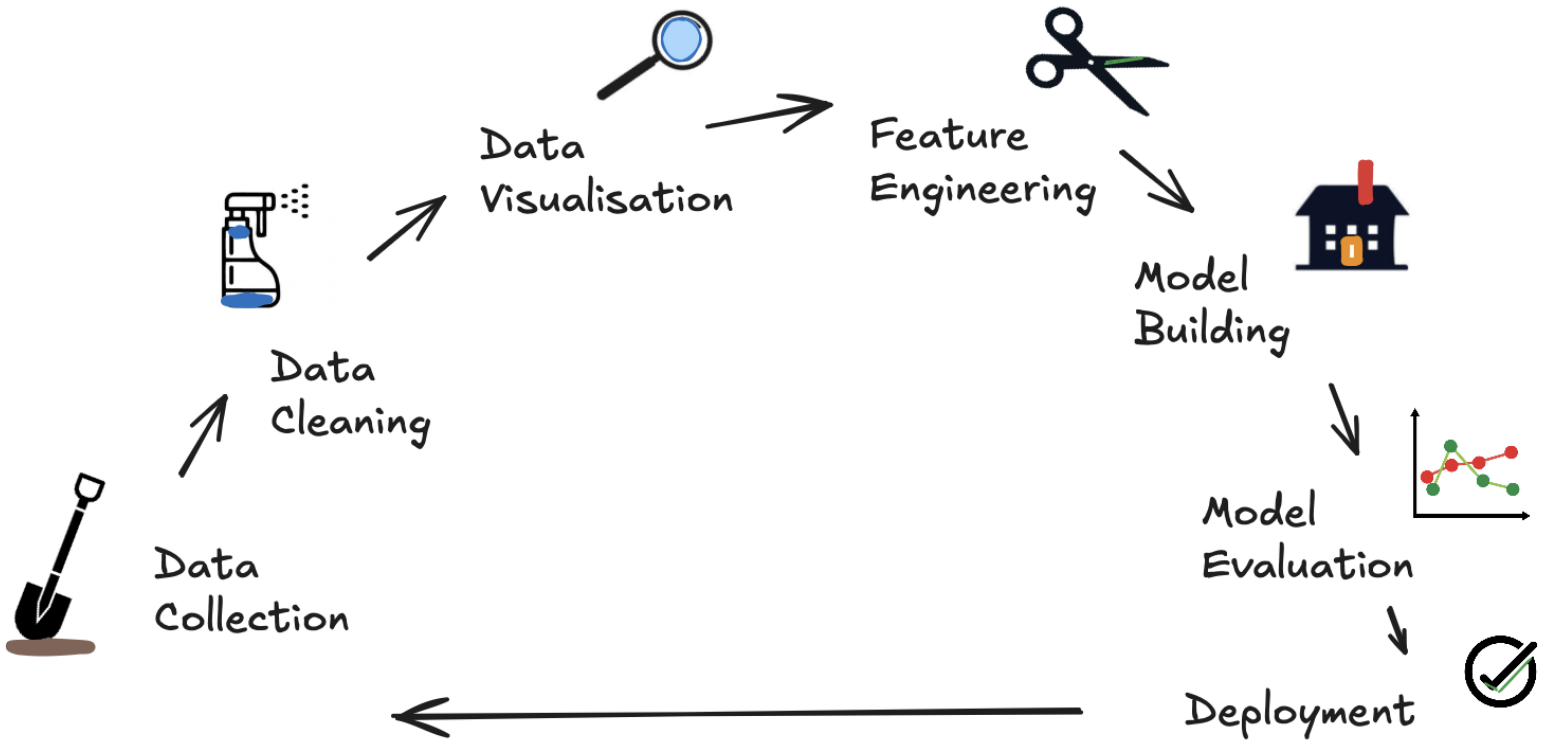
Data Cleaning: the task of dealing with missing values, duplicates, correcting/removing corrupt or inaccurate data.

Data Science Pipeline



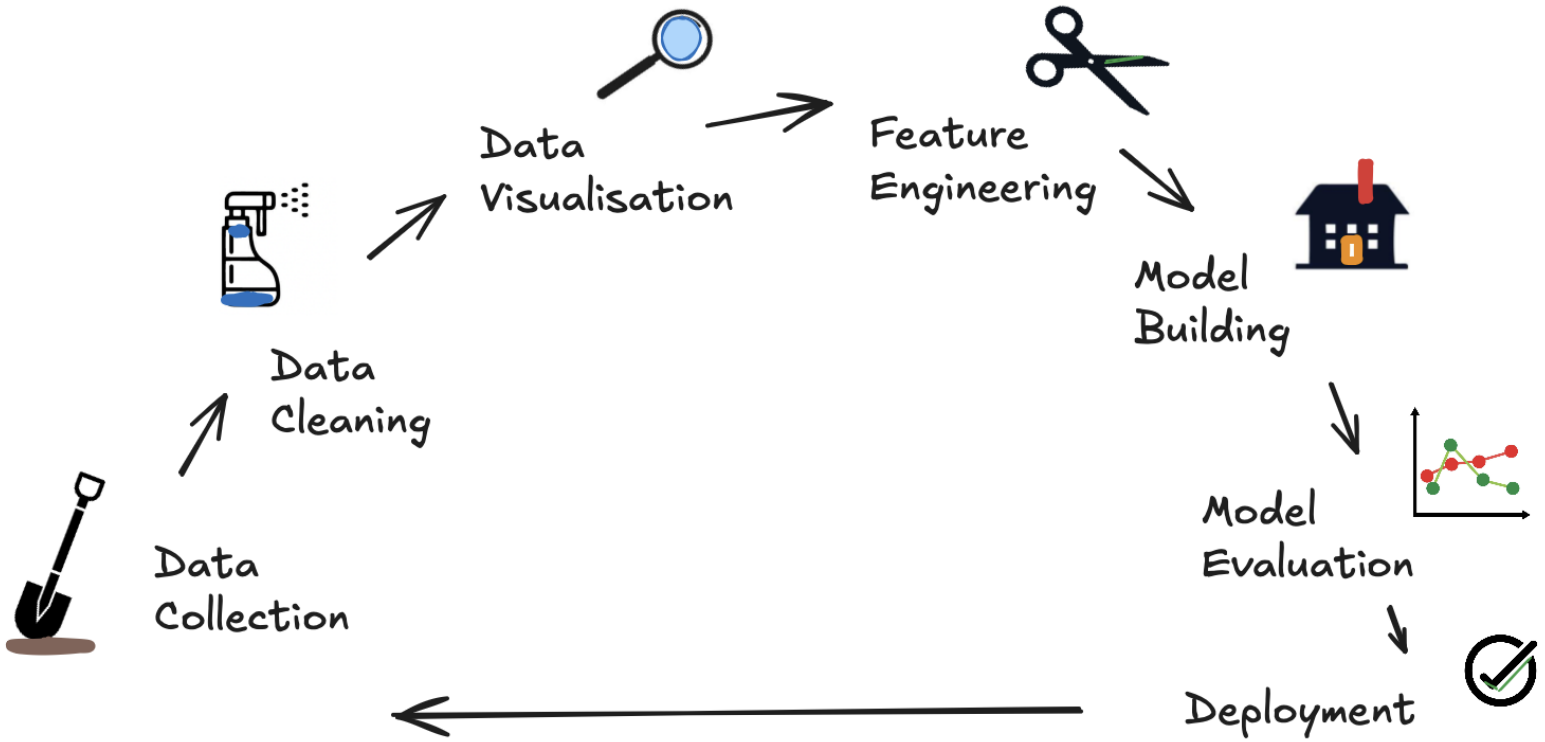
Data Visualization: understanding the data distribution, patterns, trends, and outliers.

Data Science Pipeline



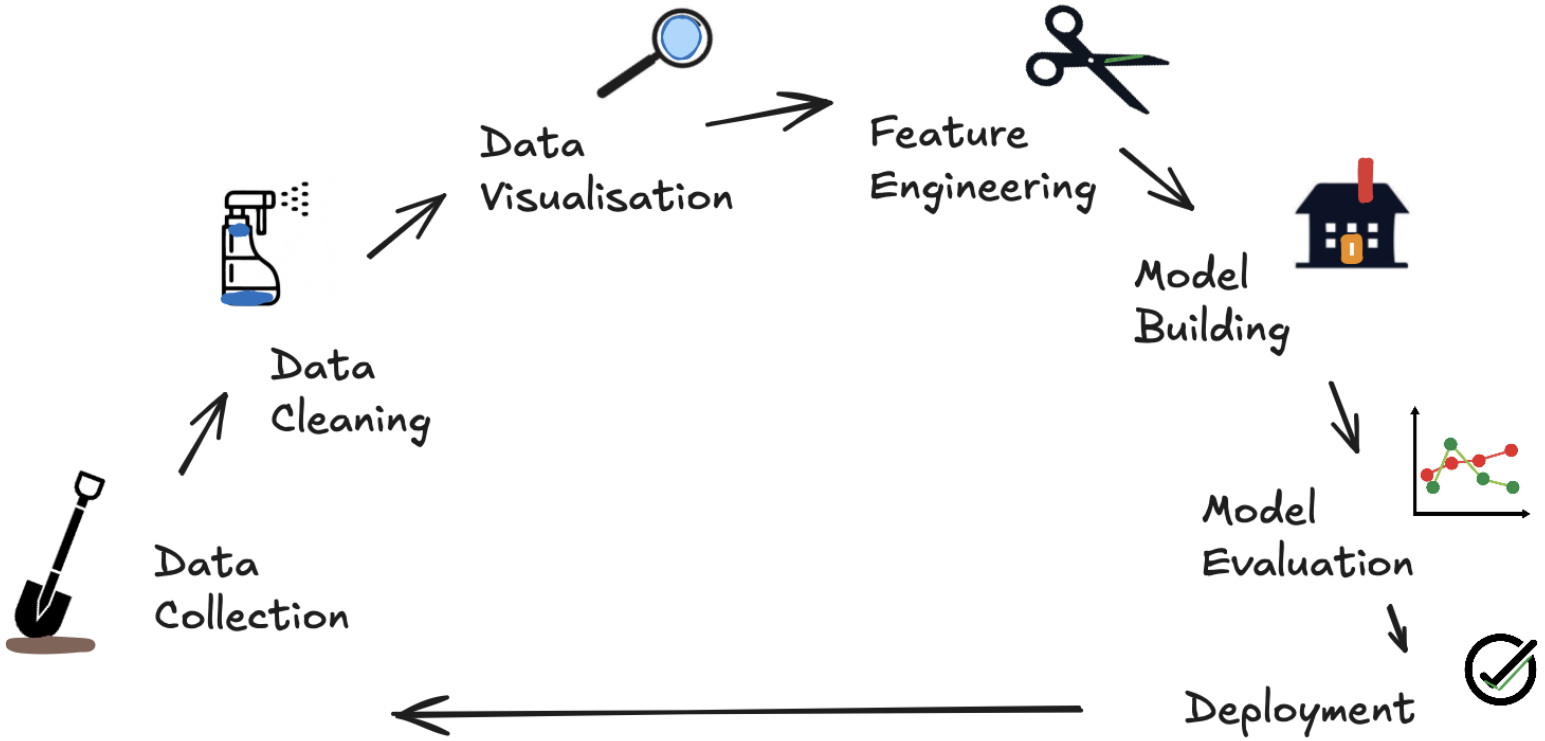
Feature Engineering: transformation of data into features and selection of relevant features.

Data Science Pipeline



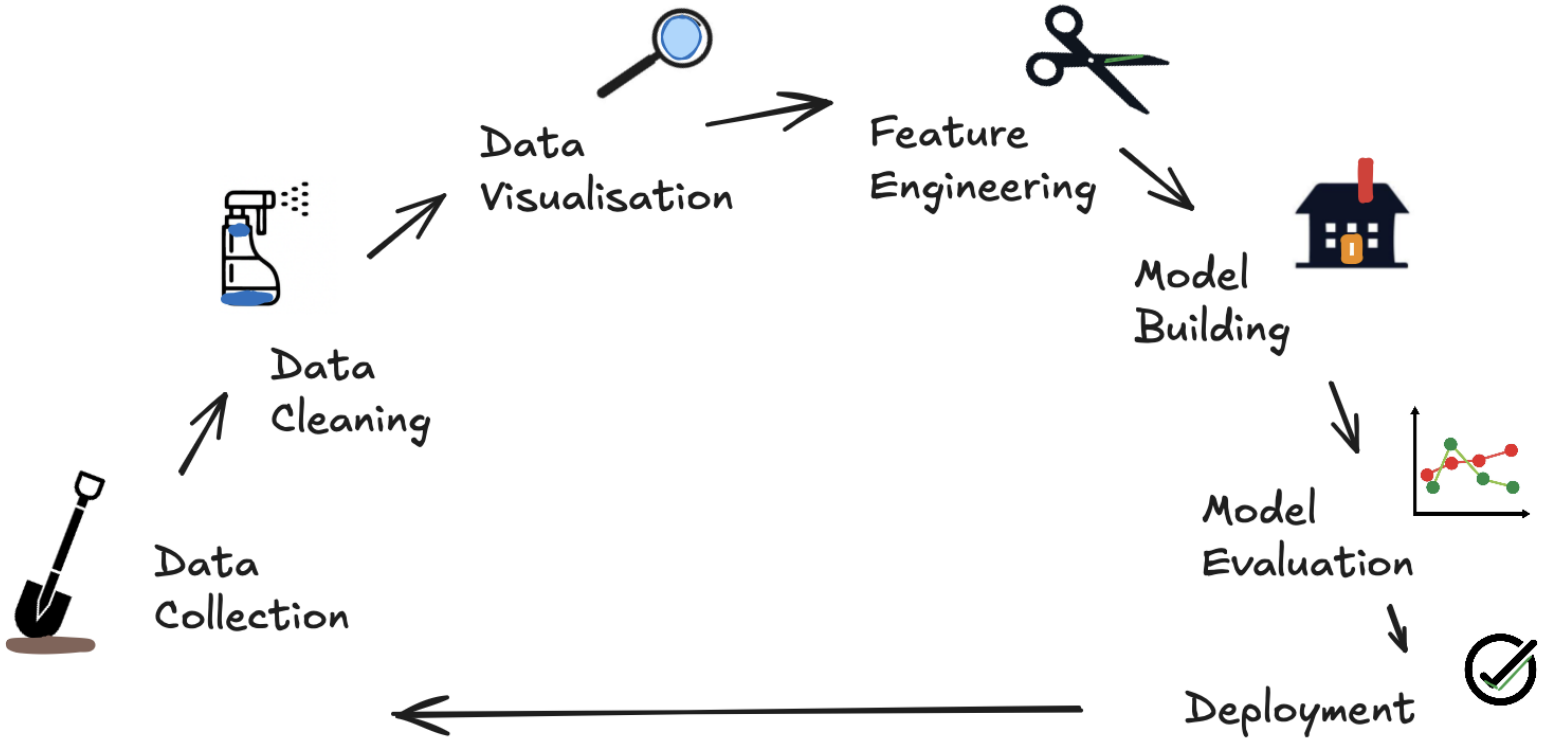
Model Building: transformation of feature/values into a model of the real world, often intended for making predictions

Data Science Pipeline



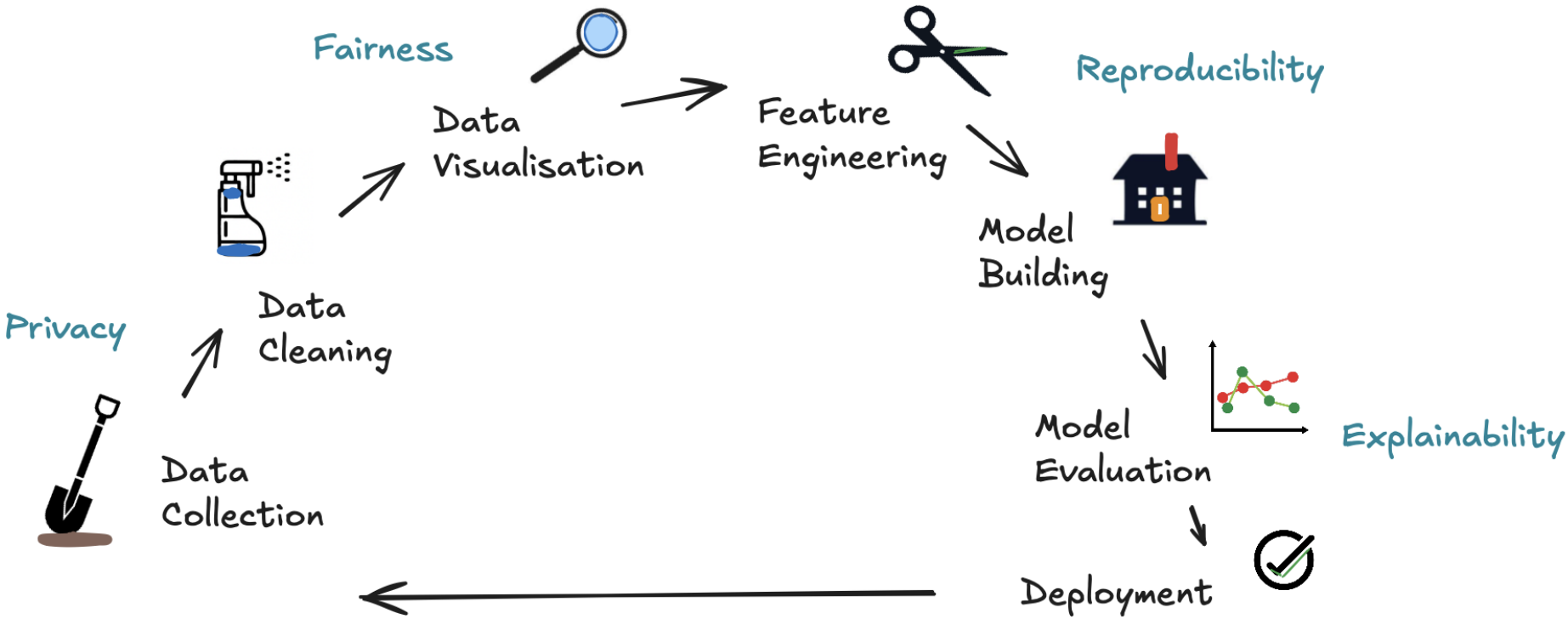
Model Evaluation: analysis of the performance of different possible models (strengths and weaknesses), using suitable metrics

Data Science Pipeline



Deployment: taking the model from the development/evaluation environments into production (use of the model in practice)

Responsible Data Science



Responsible Data Science

- Fairness
- Privacy
- Reproducibility
- Explainability

Responsible Data Science

- Fairness: making decisions free of discrimination
- Privacy
- Reproducibility
- Explainability

Responsible Data Science

- Fairness: making decisions free of discrimination
- Privacy: control of personal/sensitive information
- Reproducibility
- Explainability

Responsible Data Science

- Fairness: making decisions free of discrimination
- Privacy: control of personal/sensitive information
- Reproducibility: similar results can be achieved again by others
- Explainability

Responsible Data Science

- Fairness: making decisions free of discrimination
- Privacy: control of personal/sensitive information
- Reproducibility: similar results can be achieved again by others
- Explainability: ability to inform users how a model came to a decision

Responsible Data Science

- **Fairness: making decisions free of discrimination**
- Privacy: control of personal/sensitive information
- Reproducibility: similar results can be achieved again by others
- Explainability: ability to inform users how a model came to a decision

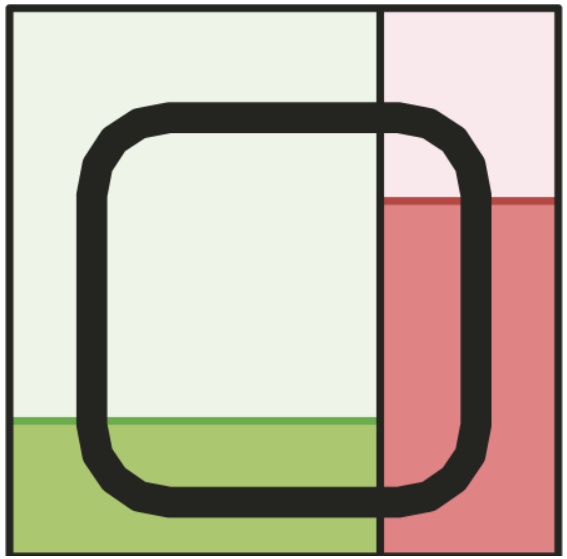
Fairness

References:

- **The Ethical Algorithm by Michael Kearns and Aaron Roth**
- **Six Ways of Fairness by Thore Husfeldt**
- Fairness and Machine Learning: Limitations and Opportunities by Barocas S., Hardt M., and Narayanan A.

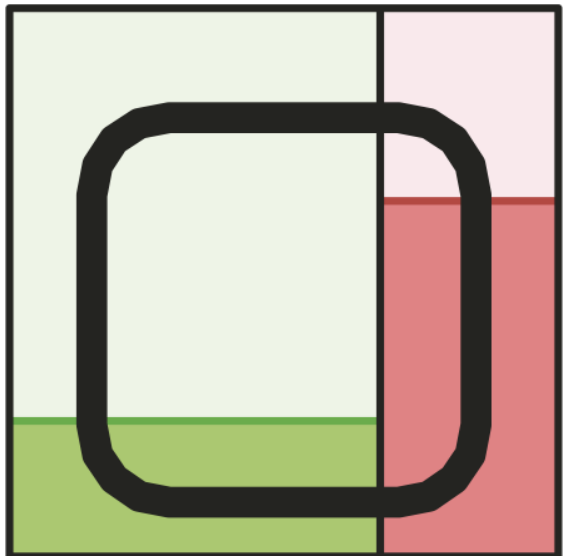
Setup

- Group membership
- Target function
- Selection function



Setup

- Group membership: Greeks and Romans
- Target function: those who repay
- Selection function: who will be granted loans



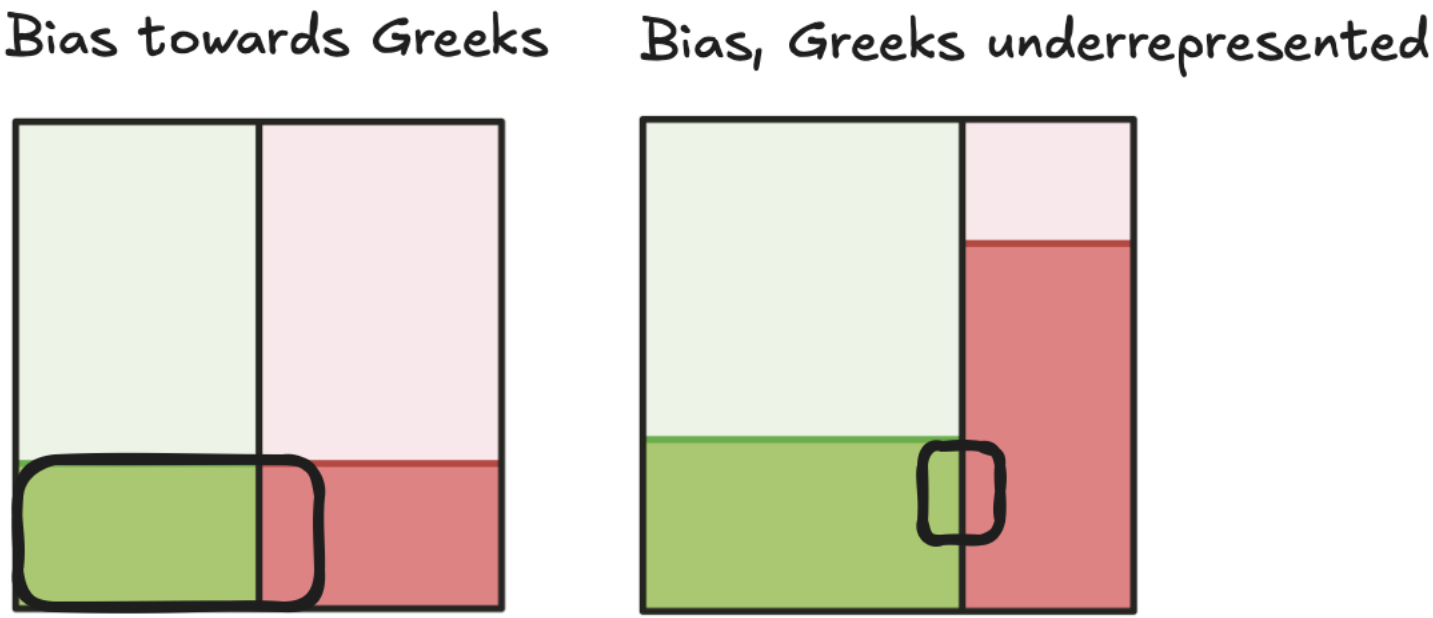
Setup

- Group membership
- Target function
- Selection function



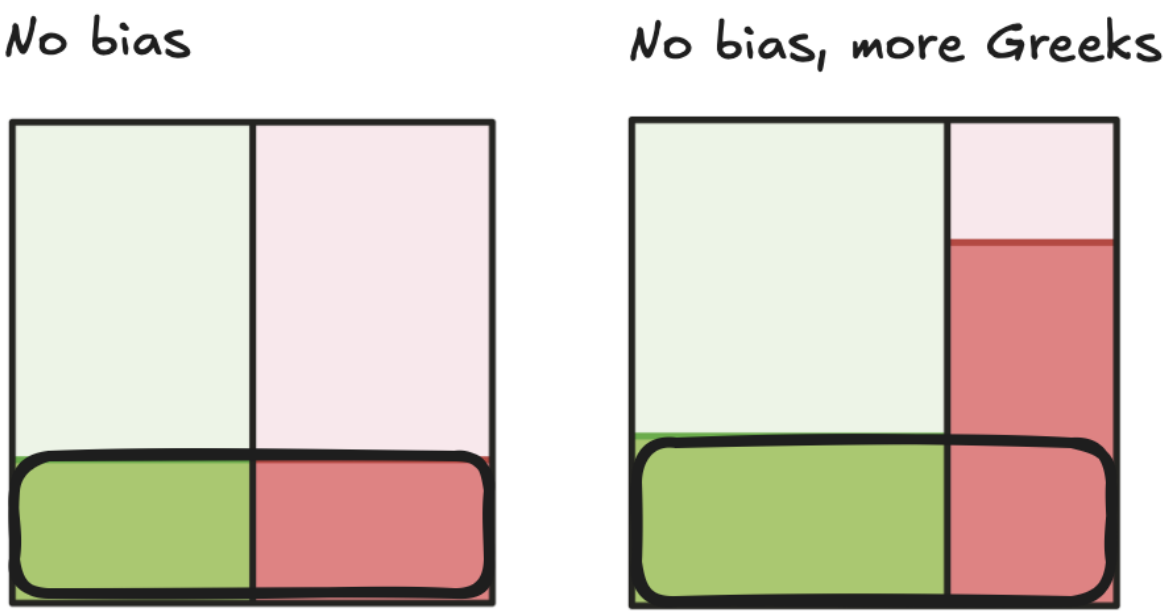
Setup

- Group membership
- Target function
- Selection function



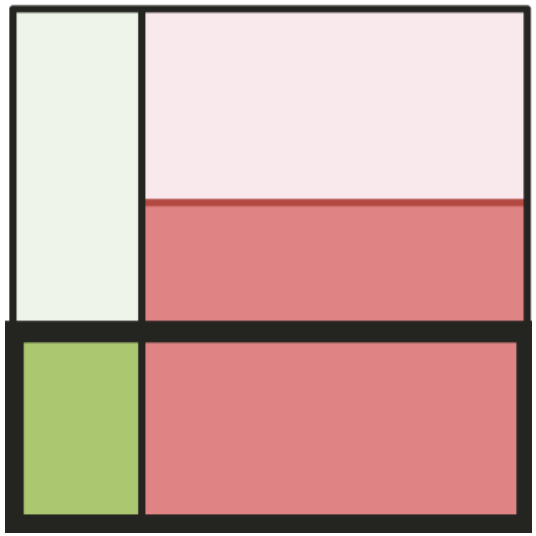
Setup

- Group membership
- Target function
- Selection function



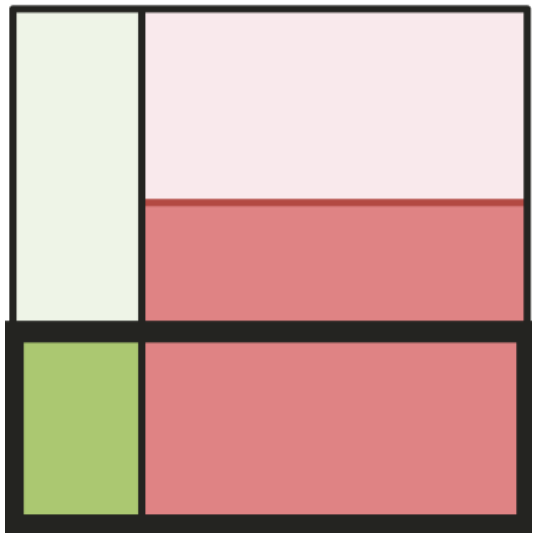
Demographic Parity (also called Statistical Parity)

- Group membership does not affect the classifier's selected outcome.
- Mathematical Definition: $G \perp S$



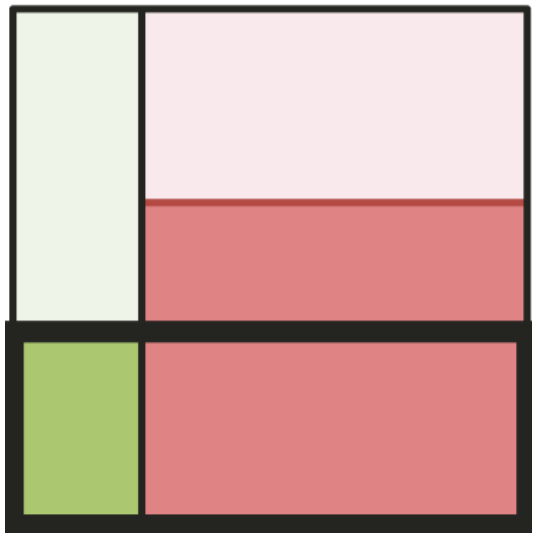
Demographic Parity (also called Statistical Parity)

- Grant loans at the same rate while the repayment rates are different.
- It might be a good idea for *gathering data*.



Demographic Parity (also called Statistical Parity)

- It does not take merit into account.
- Statistical parity may be violated even when making perfect decisions.

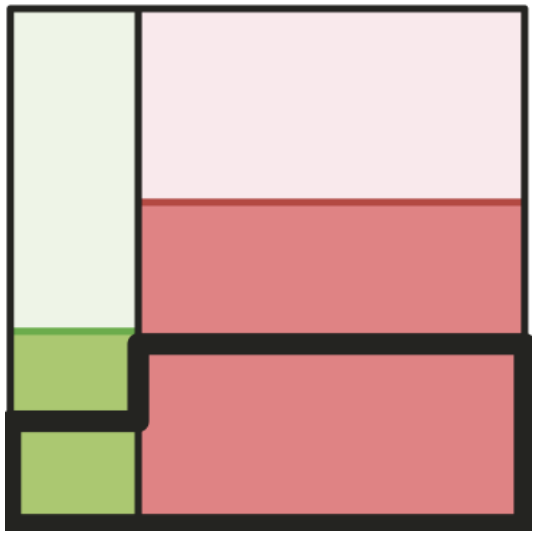


Accounting for "Merit"

- What if we evenly distribute the *mistakes* rather than evenly distributing the loans?
- Idea: your group membership does not affect the chances of the algorithm harming you.

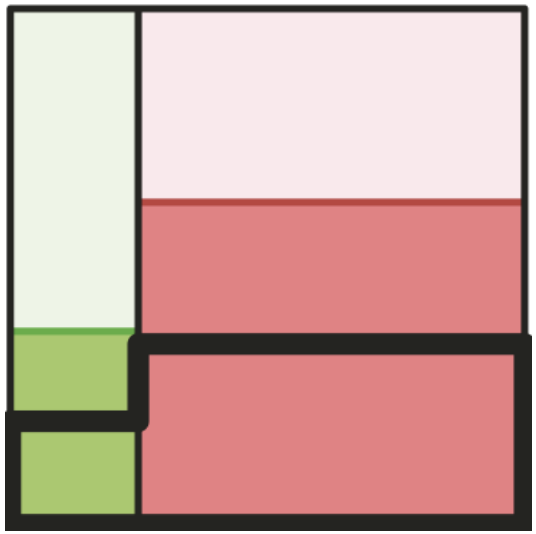
Equality of False Negatives

- False negatives (FN) should have the same rate in both groups.
- Mathematical Definition: $G \perp S = 0 \mid T = 1$



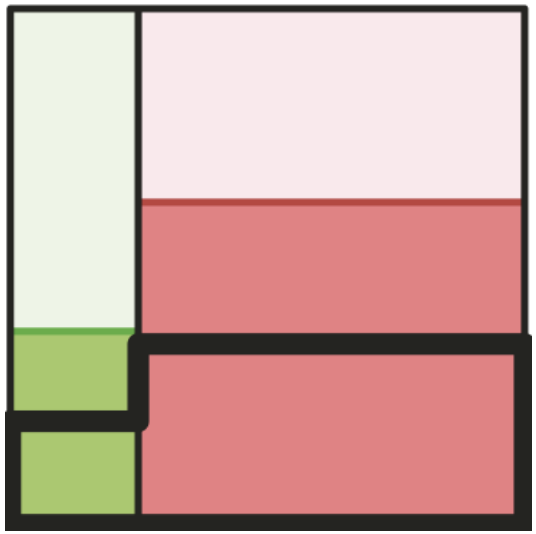
Equality of False Positives (analogous)

- False positives (FP) should have the same rate in both groups.
- Mathematical Definition: $G \perp S = 1 \mid T = 0$



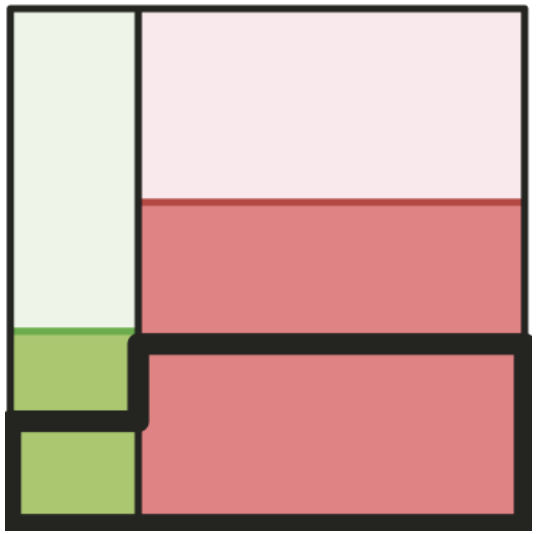
Equal Odds: Equality of Both FP and FN

- Conditioned on target value, demographic parity holds.
- Mathematical Definition: $G \perp S \mid T$



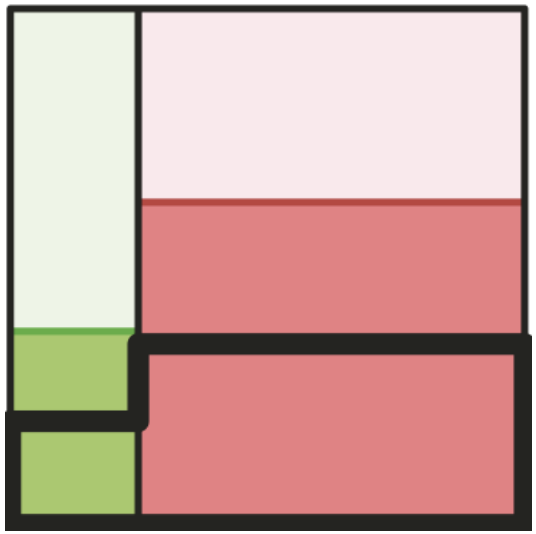
Equal Odds: Equality of Both FP and FN

- Equality of FN is weaker than Equal Odds.
- Same for Equality of FP



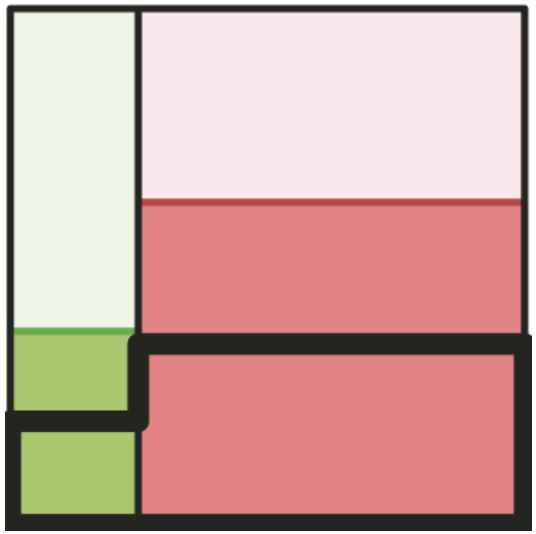
What is fair?

- Those who repay but were not selected may still consider this unfair.
- These notions are providing protections to groups, not individuals.



What is fair?

- Equal Odds is satisfied when making perfect decisions.
- Solution for tension between fairness and accuracy?



Fairness vs. Accuracy

- While perfect decision making is considered to be fair (that is, selecting exactly those who repay), in practice, we cannot expect algorithms to make perfect decisions.

Fairness vs. Accuracy

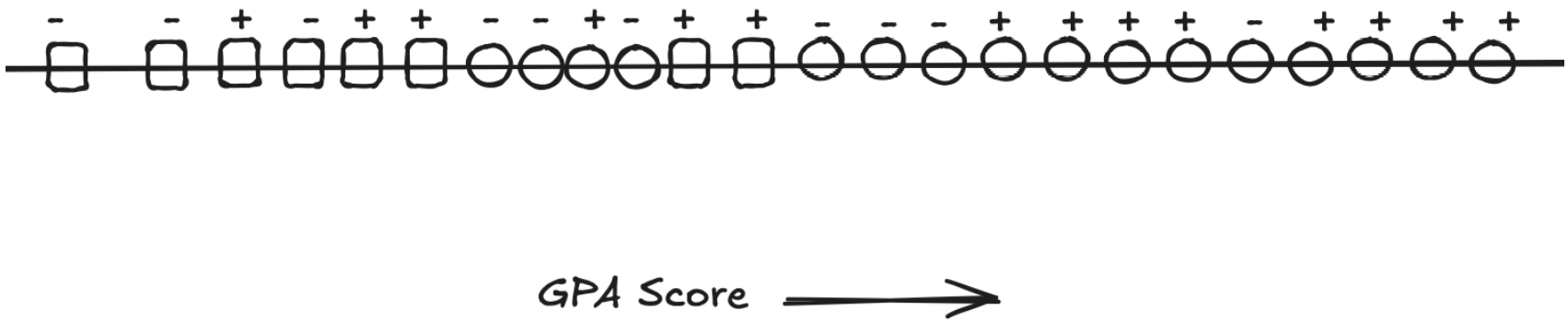
- We now discuss the trade-off between fairness and accuracy.
- Consider we have to select students to receive scholarships and be admitted to the St. Fairness College based on GPA scores.

Fairness vs. Accuracy

- Members of the Circle and Square populations are applying, with the majority of applicants being Circles.
- Circle applicants tend to be wealthier and thus can afford to spend all of their time studying, while square applicants are less wealthy and generally need to work part-time.

Fairness vs. Accuracy

To illustrate, suppose the historical applicant dataset is as follows. '+' indicates success at St. Fairness, while '-' indicates failure.



Fairness vs. Accuracy

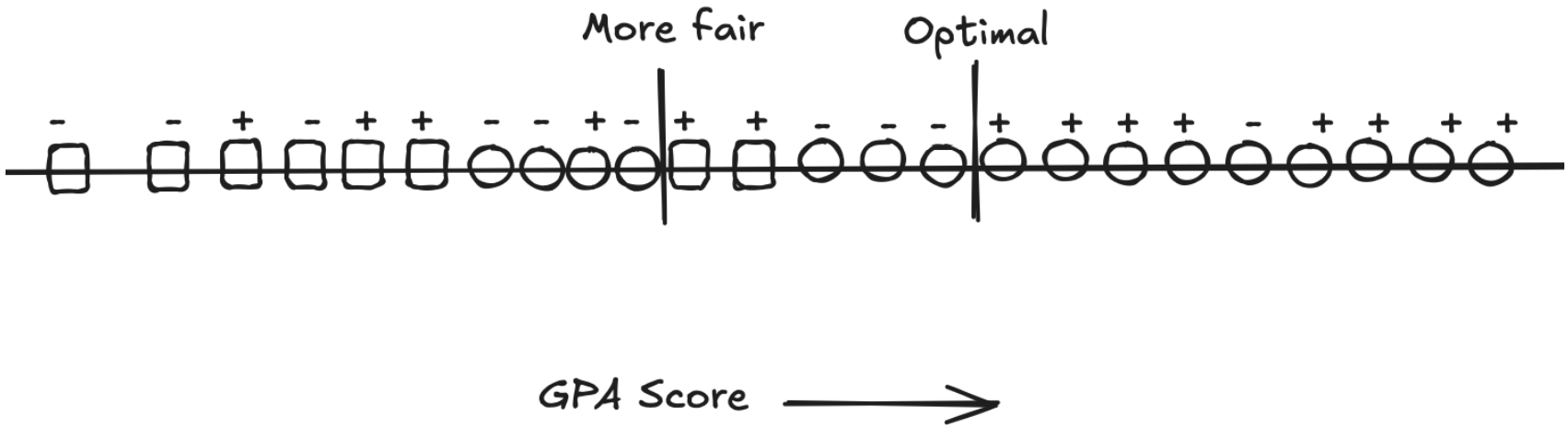
- Both populations are equally competent for collegiate success.
- The percentage of Circles who would succeed if granted scholarships and admitted to St. Fairness is equal to the percentage of Squares.
- It is just that Circles have higher GPA scores because they can use all of their time to prioritize improving their GPA scores.

Fairness vs. Accuracy

- If the model admits applicants with GPA greater than some threshold, then it may not be accurate and may badly violate fairness.
- Since Circles are the majority, the most accurate model will set the threshold largely based on the GPA scores and collegiate success of Circles, since the majority group counts more towards the aggregate error than the minority (Squares).
- This comes at the expense of discrimination (a higher false rejection rate) against the Squares.

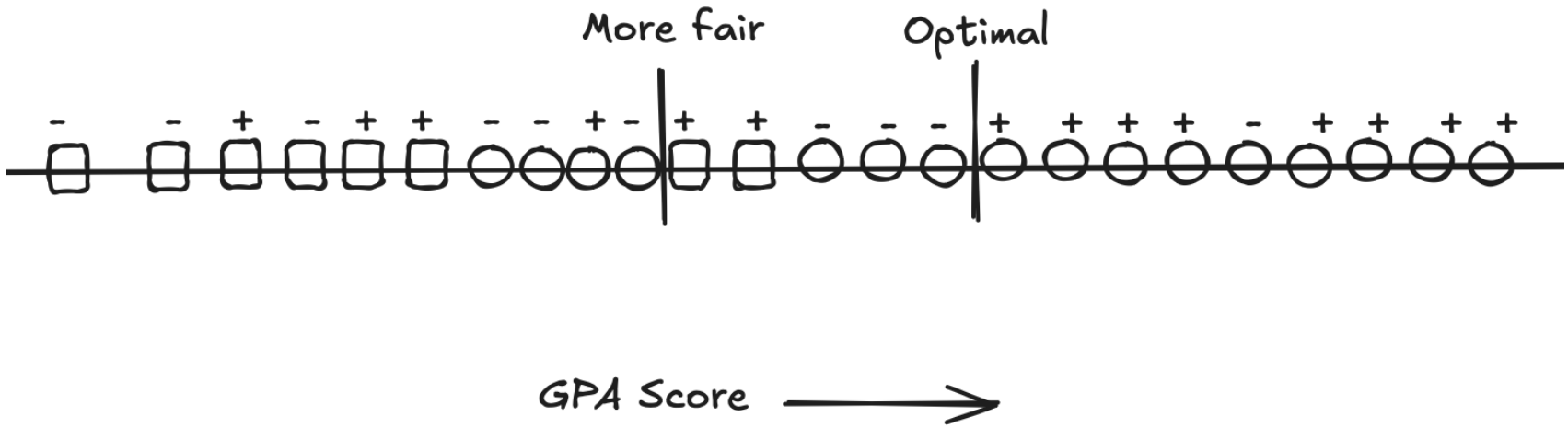
Fairness vs. Accuracy

Optimal: there are 7 mistakes
More fair: there are 8 mistakes



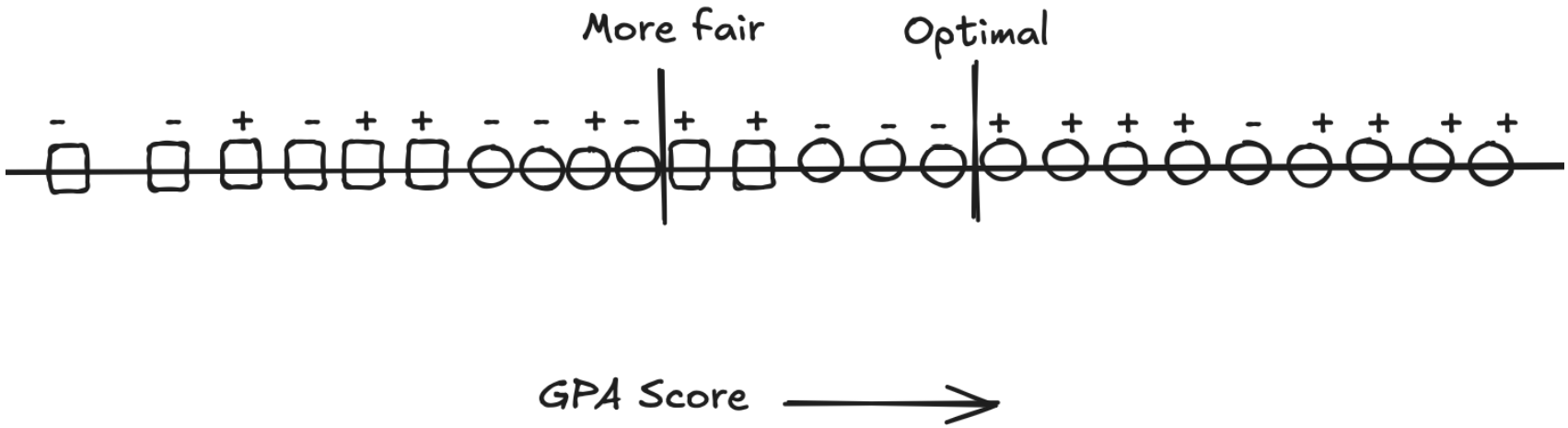
Fairness vs. Accuracy

Assume the "unfairness score" is the magnitude of the difference between falsely rejected Circles and falsely rejected Squares.



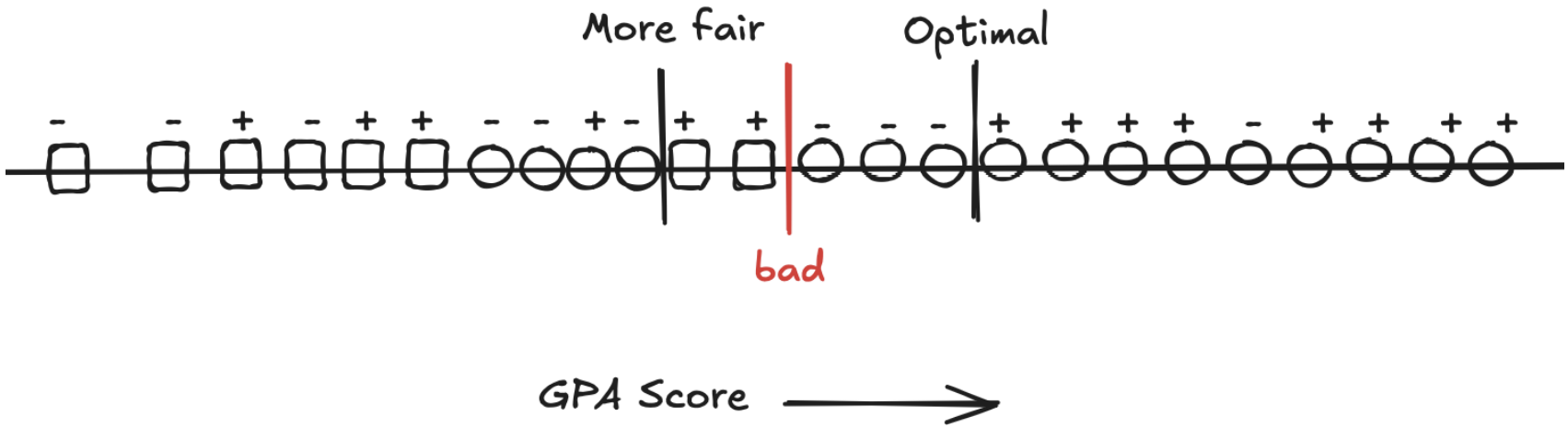
Which model is better?

Optimal: unfairness score of 4 and 7 mistakes
More fair: unfairness score of 2 and 8 mistakes



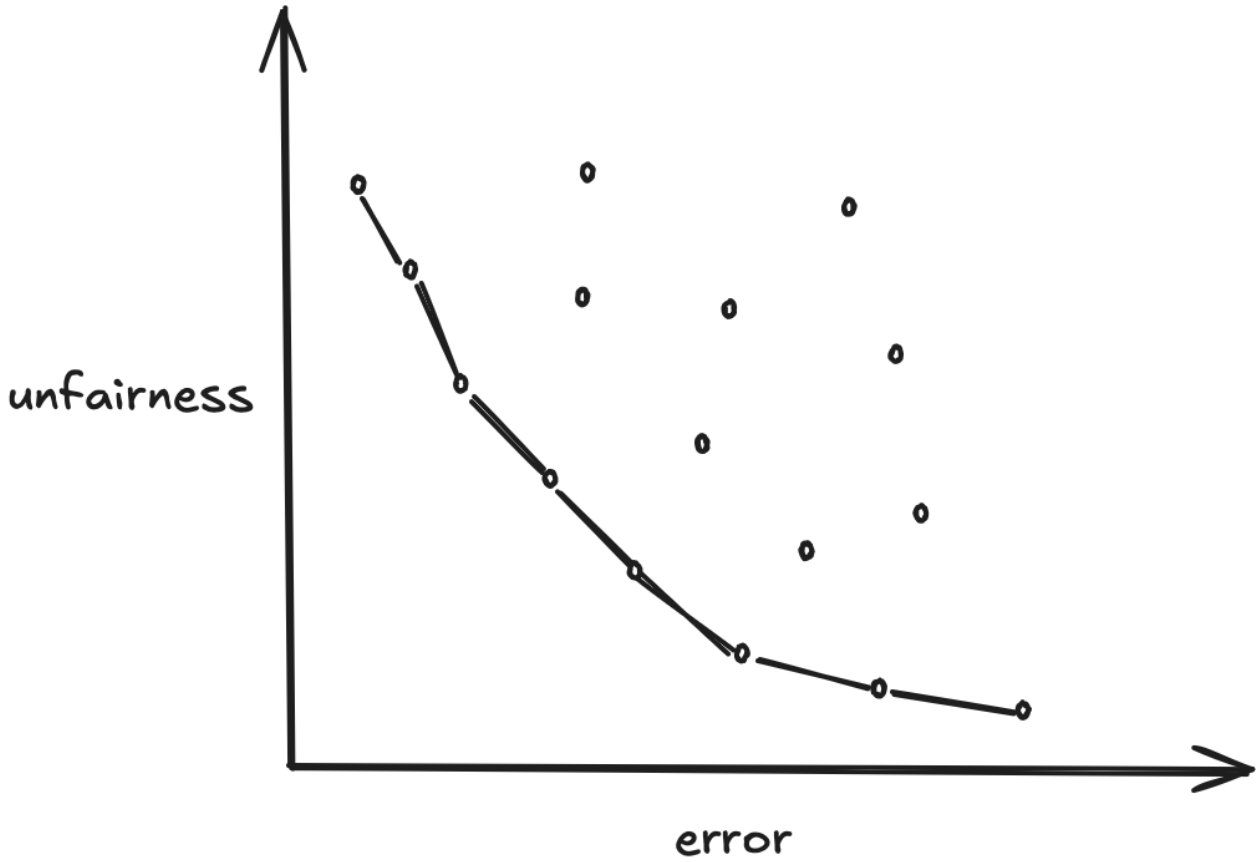
Which model is better?

Optimal: unfairness score of 4 and 7 mistakes
More fair: unfairness score of 2 and 8 mistakes



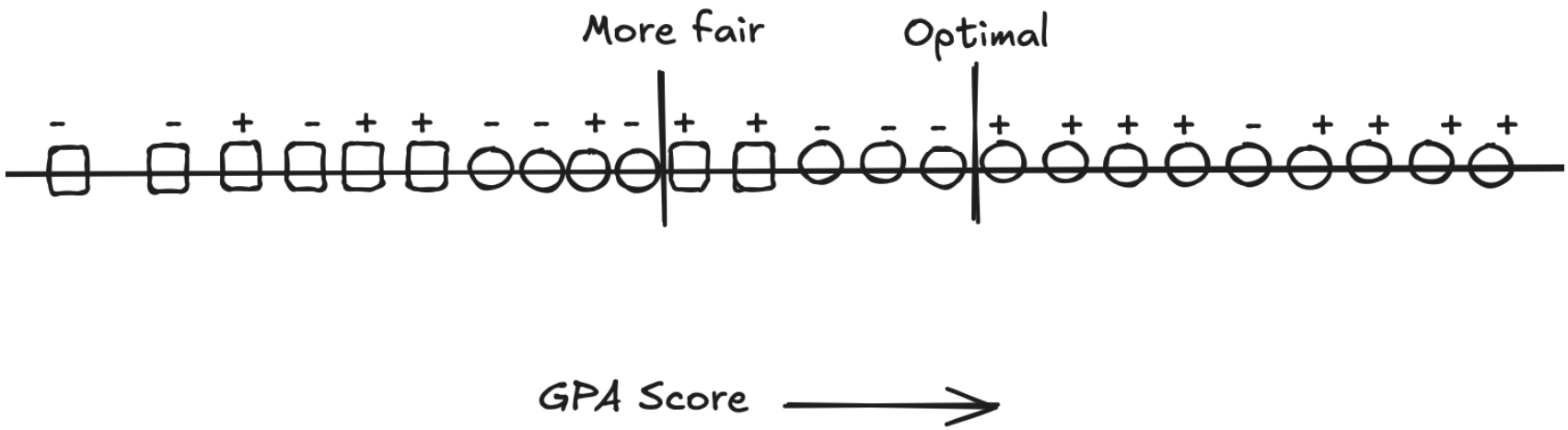
Pareto Frontier

Optimal: unfairness score of 4 and 7 mistakes
More fair: unfairness score of 2 and 8 mistakes



Fairness vs. Accuracy

Issue: Circle and Square GPA scores may differ for reasons unrelated to collegiate success (wealth, time to dedicate to GPA scores).

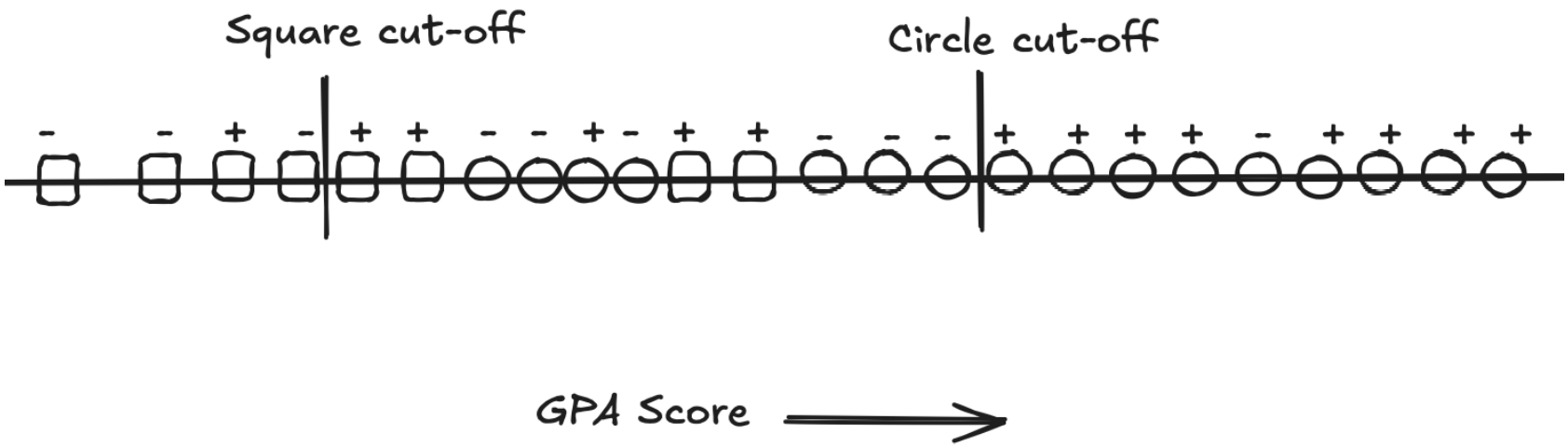


Fairness vs. Accuracy

If we can detect that the distributions of Circle and Square scores are different, why not build separate models for the two populations?

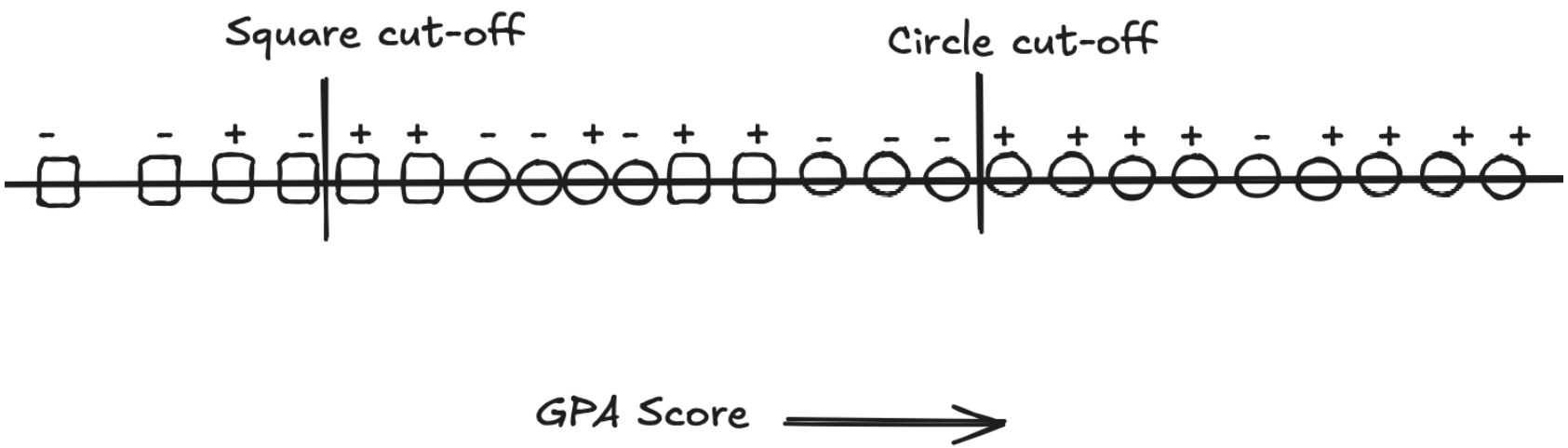
Fairness vs. Accuracy

Only 3 mistakes
Unfairness score is 0



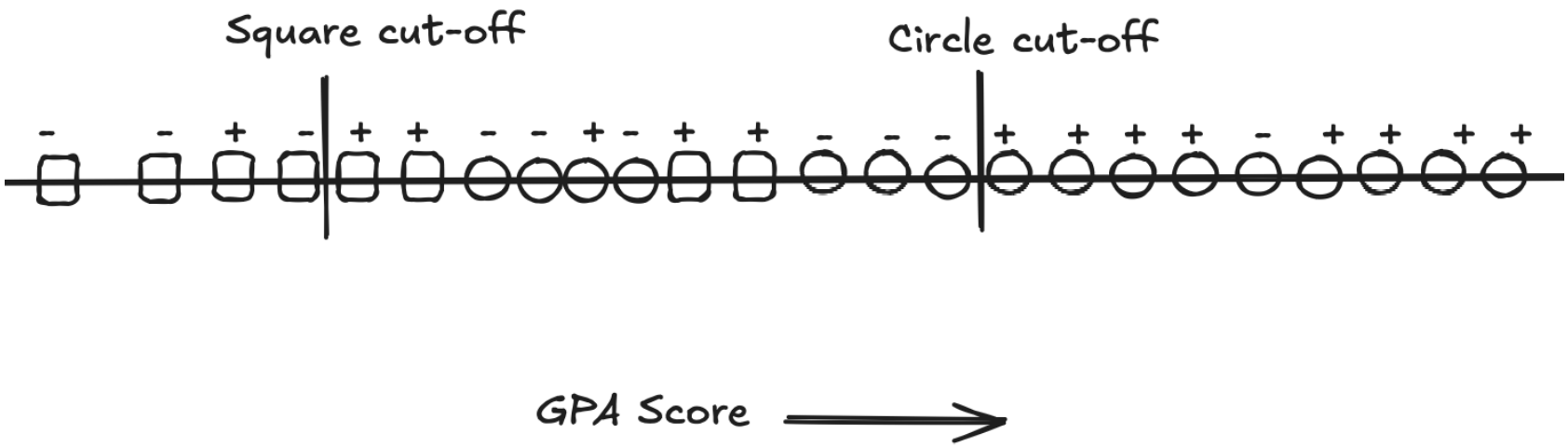
Fairness vs. Accuracy

By allowing a more complex model we improved on both criteria.
This resembles how some affirmative policies are implemented.



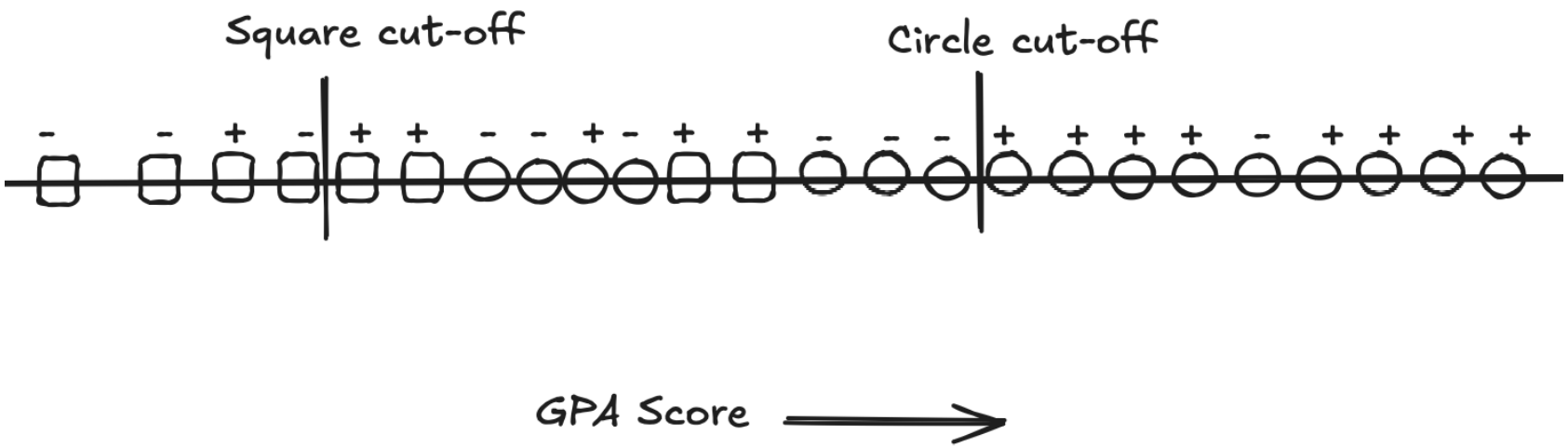
Fairness vs. Accuracy

But this explicitly uses group membership as input.
It could contribute to discrimination.



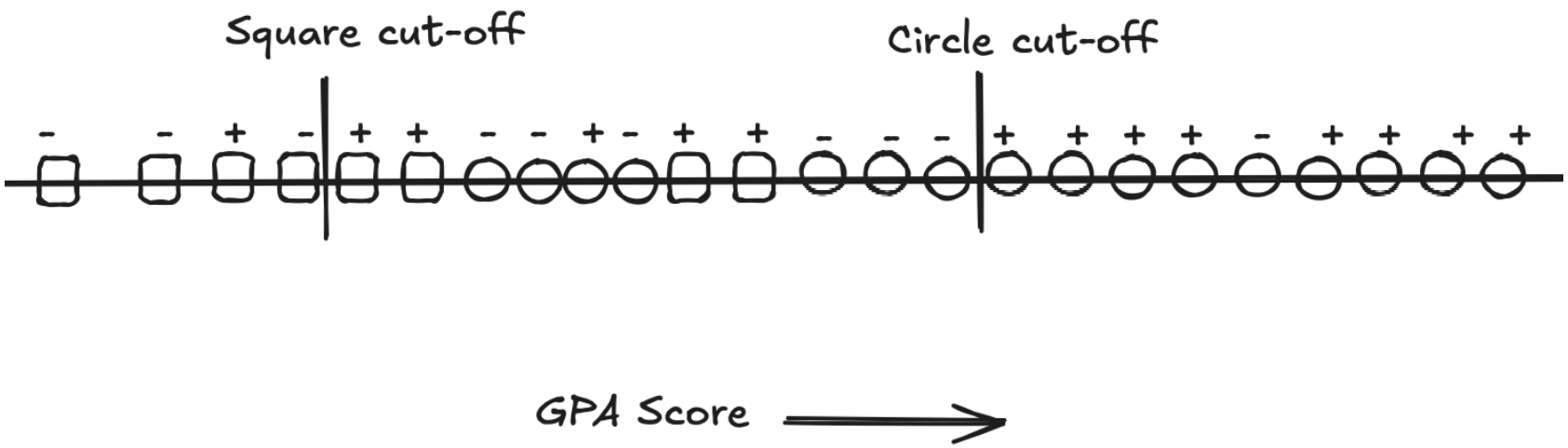
Fairness vs. Accuracy

Even if we can use a hybrid model, the GPA score could be more predictive to college success for one population than for the other.



Fairness vs. Accuracy

We are still discriminating against the Square population according to the Equality of FN, even if less than with the single threshold model.



Subgroup Fairness

- Discrimination can be based on various aspects: gender, nationality, age, sexual orientation, disability, and so on.
- Protection of multiple overlapping groups can come at the expense of discrimination against some intersection of them.

Subgroup Fairness

- Assume the population of Circles can be divided into male and female and the same for Squares.
- If we are to distribute 16 free tickets to the Olympic Games, one may agree that the "most fair" would be to give 4 tickets each to Circle women, Circle men, Square men, Square women.

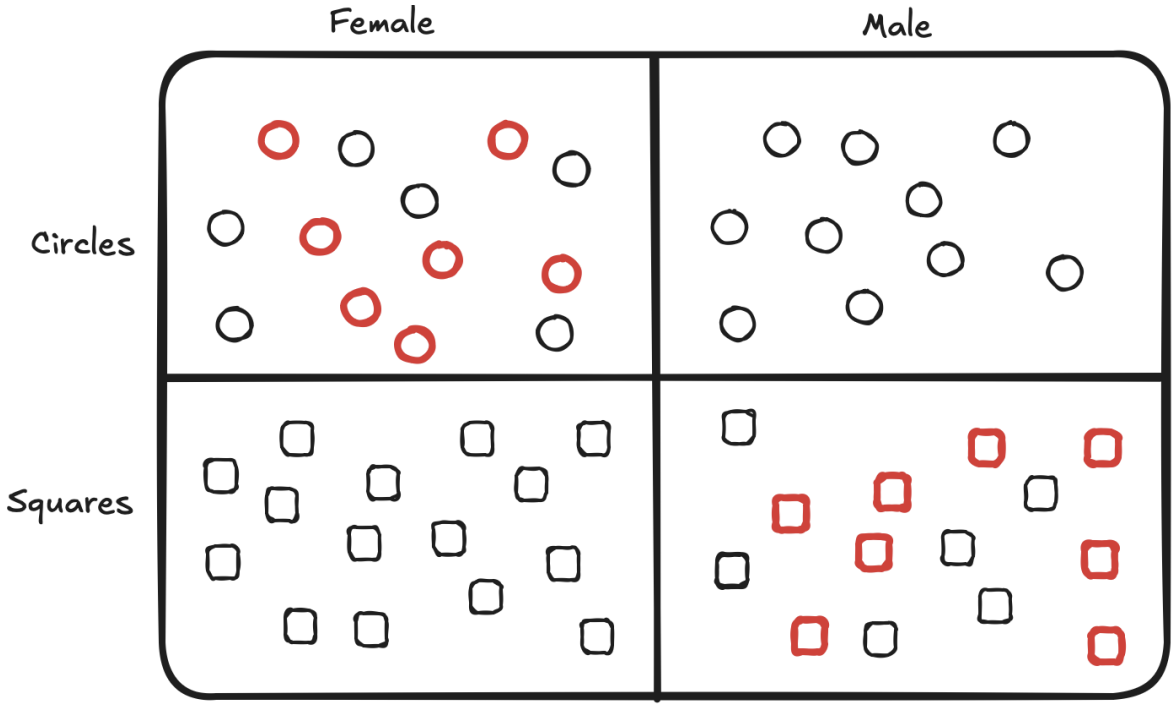
Subgroup Fairness

- But that is not what we get if ask an algorithm to be fair w.r.t. these aspects *separately*.
- A "gerrymandered" solution could give 8 tickets to Circle men and 8 tickets to Square women.

Note: gerrymandering is the practice of dividing or arranging a territorial unit into election districts in a way that gives one political party an unfair advantage in elections (Merrian-Webster).

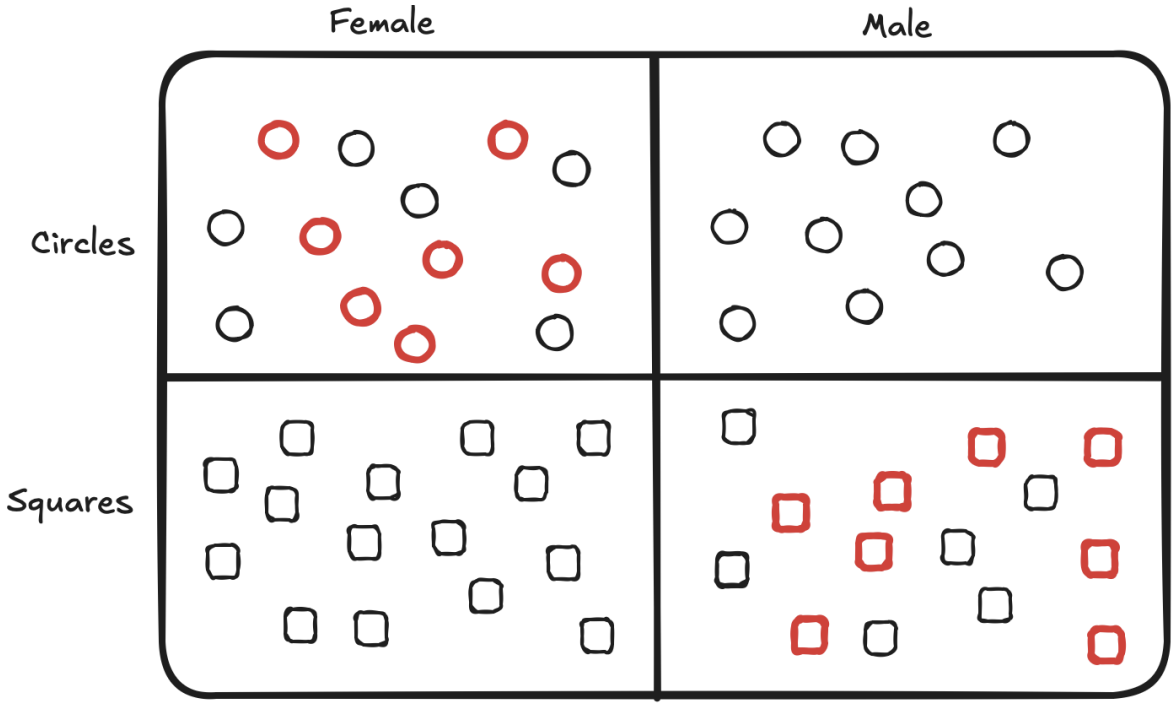
Subgroup Fairness

We would still have 8 tickets going to Squares, 8 tickers going to Circle, 8 to men, and 8 to women.



Subgroup Fairness

More complex goals/score functions would be needed to take subgroups into account.



Summary

- Data Science is a powerful tool for extracting insights from data.
- Involves a process of data collection, cleaning, exploration, feature engineering, model building, model evaluation, and deployment.
- Responsible Data Science is concerned with studying and incorporating ethical behaviour into the data science pipeline.
- Fairness: definitions, fairness vs. accuracy, subgroup fairness.

Bonus: Incompatibility of Fairness Notions

- Unless we are all equal and predictive parity holds, no classifier can guarantee both Equal Odds and Predictive Parity.
- But it is unrealistic to assume that we are all equal (some groups may receive more incentives than others).