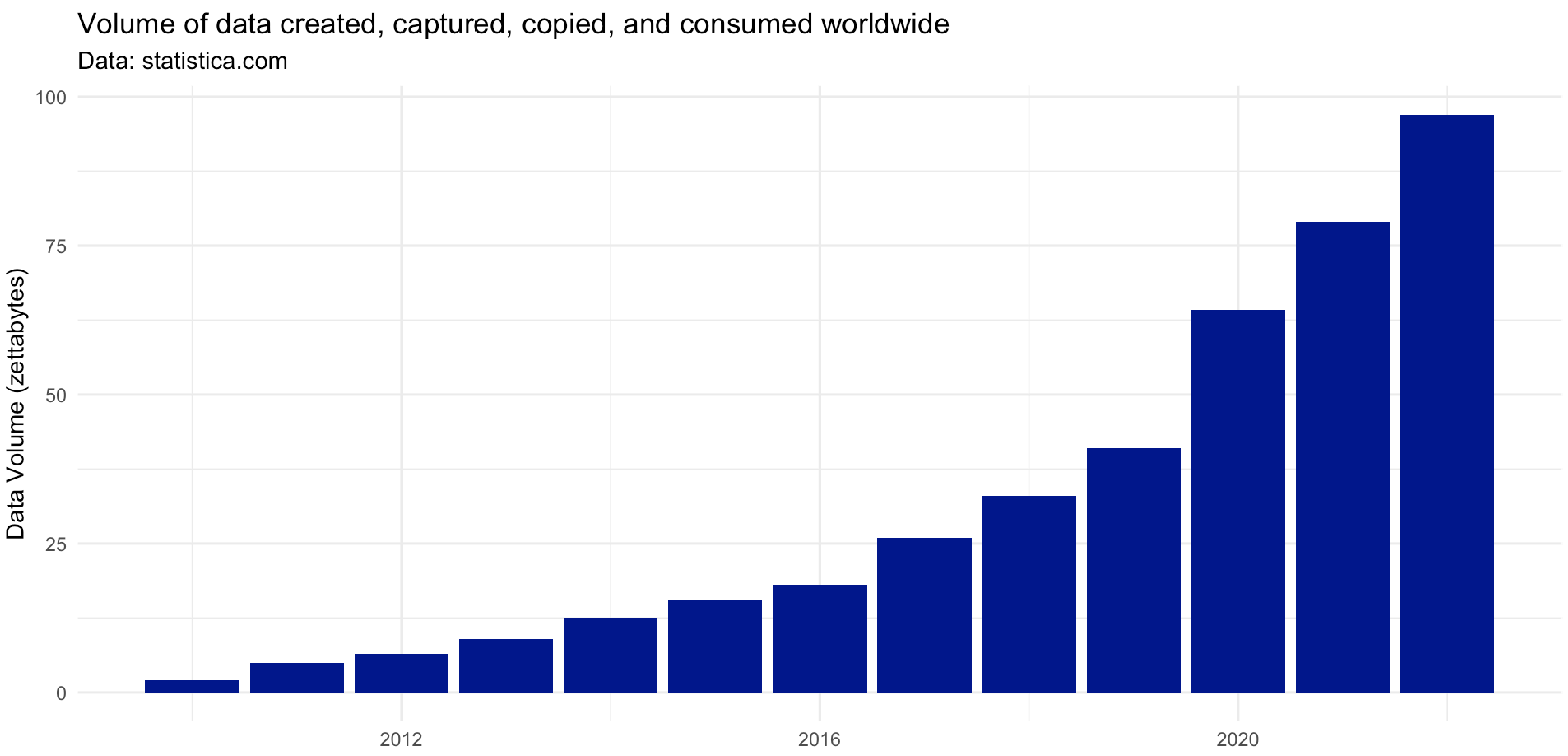
Fairness

Ethical Data Science

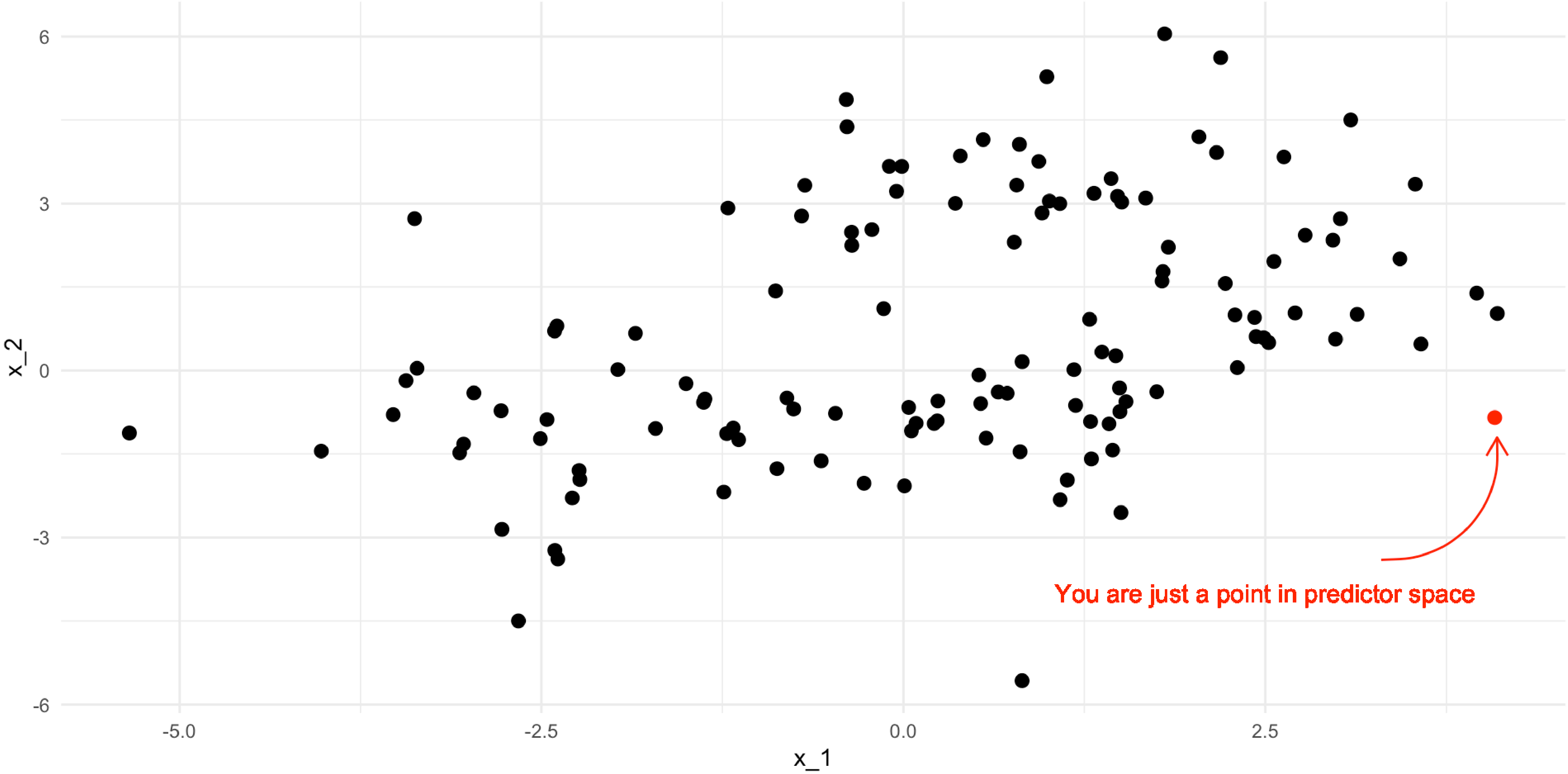
Dr Zak Varty

Fairness and the Data Revolution

Fairness and the Data Revolution

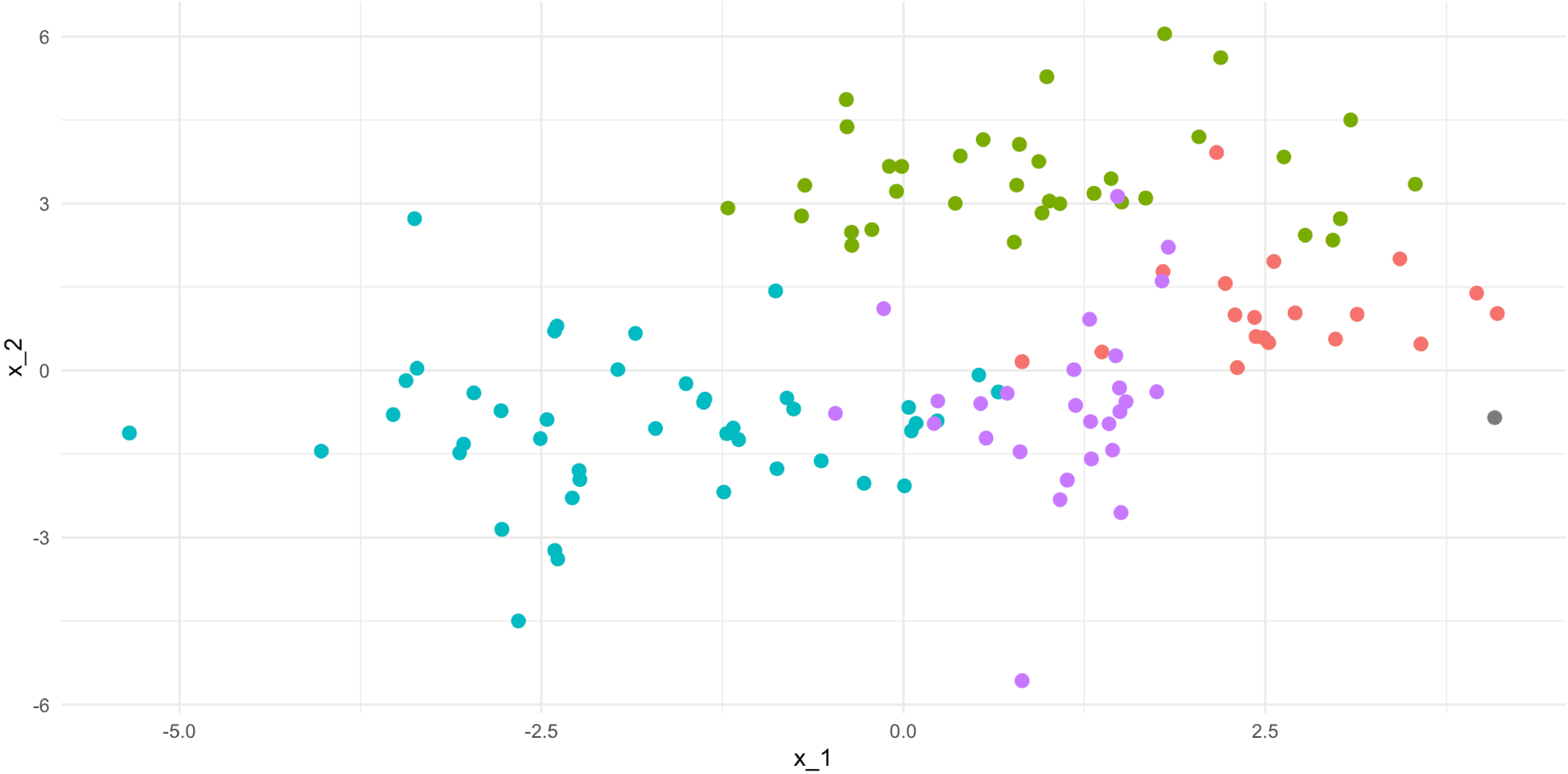


You are your Data



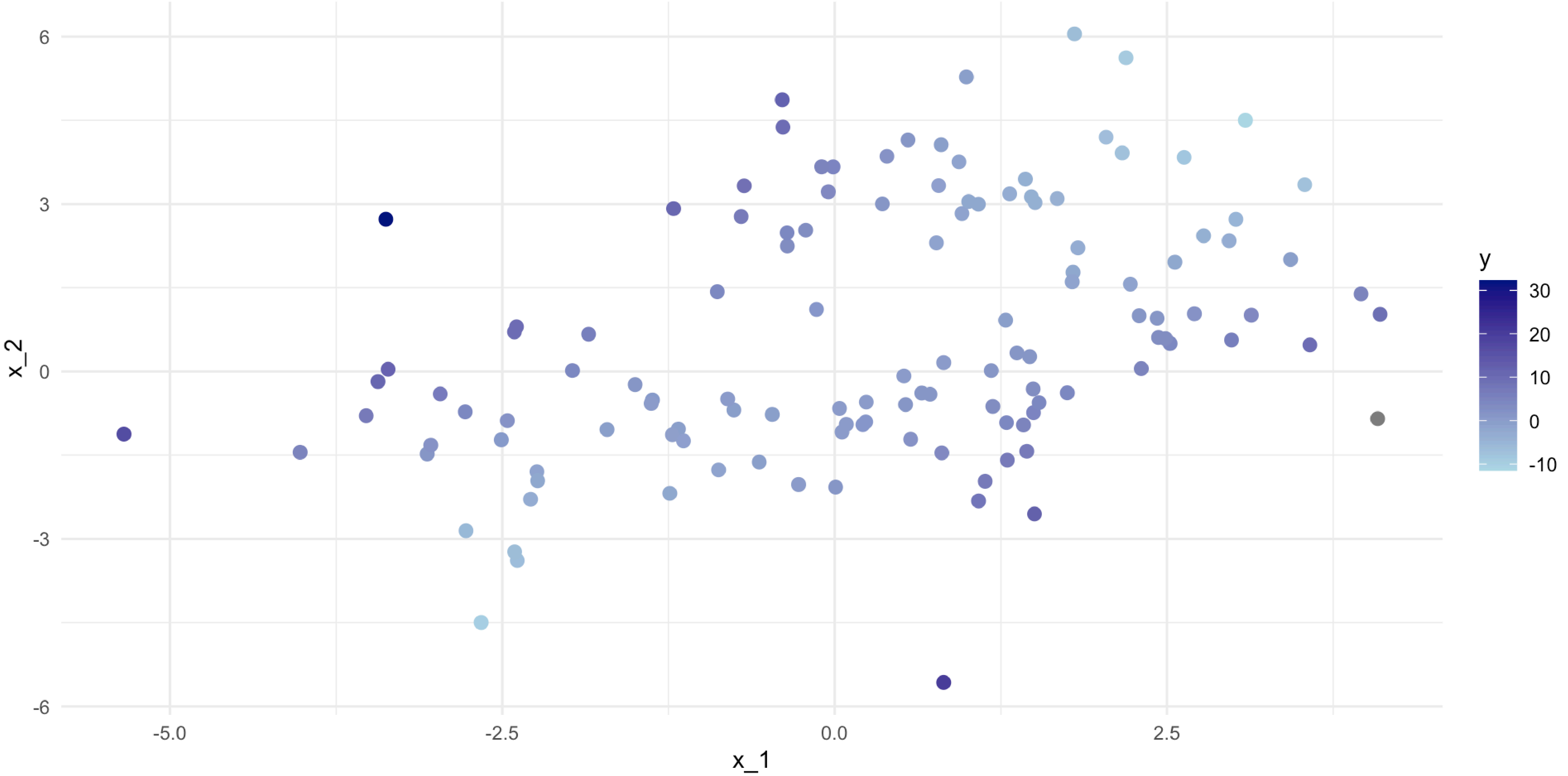
MATH-70076 Effective Data Science

You are Your Data: Clustering



MATH-70076
Effective
Data
Science

You are Your Data: Prediction



Forbidden Predictors

Protected Characteristics under the Equality Act (2010)

- age
- gender reassignment
- being married or in a civil partnership
- being pregnant or on maternity leave
- disability
- race including colour, nationality, ethnic or national origin
- religion or belief
- sex
- sexual orientation

Measuring Fairness

- Mapping from human to mathematical concept, many measures of fairness.
- Binary outcome $Y \in \{0, 1\}$.
- Binary Prediction $\hat{Y} \in \{0, 1\}$.
- Protected attribute A takes values in $\mathcal{A} = \{a_1, \dots, a_k\}$.

Demographic Parity

> The probability of predicting a ‘positive’ outcome is the same for all groups.

$$\mathbb{P}(\hat{Y} = 1 | A = a_i) = \mathbb{P}(\hat{Y} = 1 | A = a_j), \quad \text{for all} \;\; i, j \in \mathcal{A}.$$

Equal Opportunity

> Among those who have a true ‘positive’ outcome, the probability of predicting a ‘positive’ outcome is the same for all groups.

$$\mathbb{P}(\hat{Y} = 1 | A = a_i, Y = 1) = \mathbb{P}(\hat{Y} = 1 | A = a_j, Y = 1), \quad \text{for all } \ i, j \in \mathcal{A}.$$

MATH-70076 Effective Data Science

Equal Odds

> Among those who have a true ‘positive’ outcome, the probability of predicting a ‘positive’ outcome is the same for all groups.
>
> AND
>
> Among those who have a true ‘negative’ outcome, the probability of predicting a ‘negative’ outcome is the same for all groups.

$$\mathbb{P}(\hat{Y} = y | A = a_i, Y = y) = \mathbb{P}(\hat{Y} = y | A = a_j, Y = y), \quad \text{for all} \;\; y \in \{0, 1\} \;\; \text{and} \;\; i,$$

Predictive Parity

> The probability of a true 'positive' outcome for people who were predicted a 'positive' outcome is equal across groups.

$$\mathbb{P}(Y = 1 | \hat{Y} = 1, A = a_i) = \mathbb{P}(Y_1 = 1 | \hat{Y} = 1, A = a_j) \text{ for all } i, j \in \mathcal{A}.$$

This is all a bit much

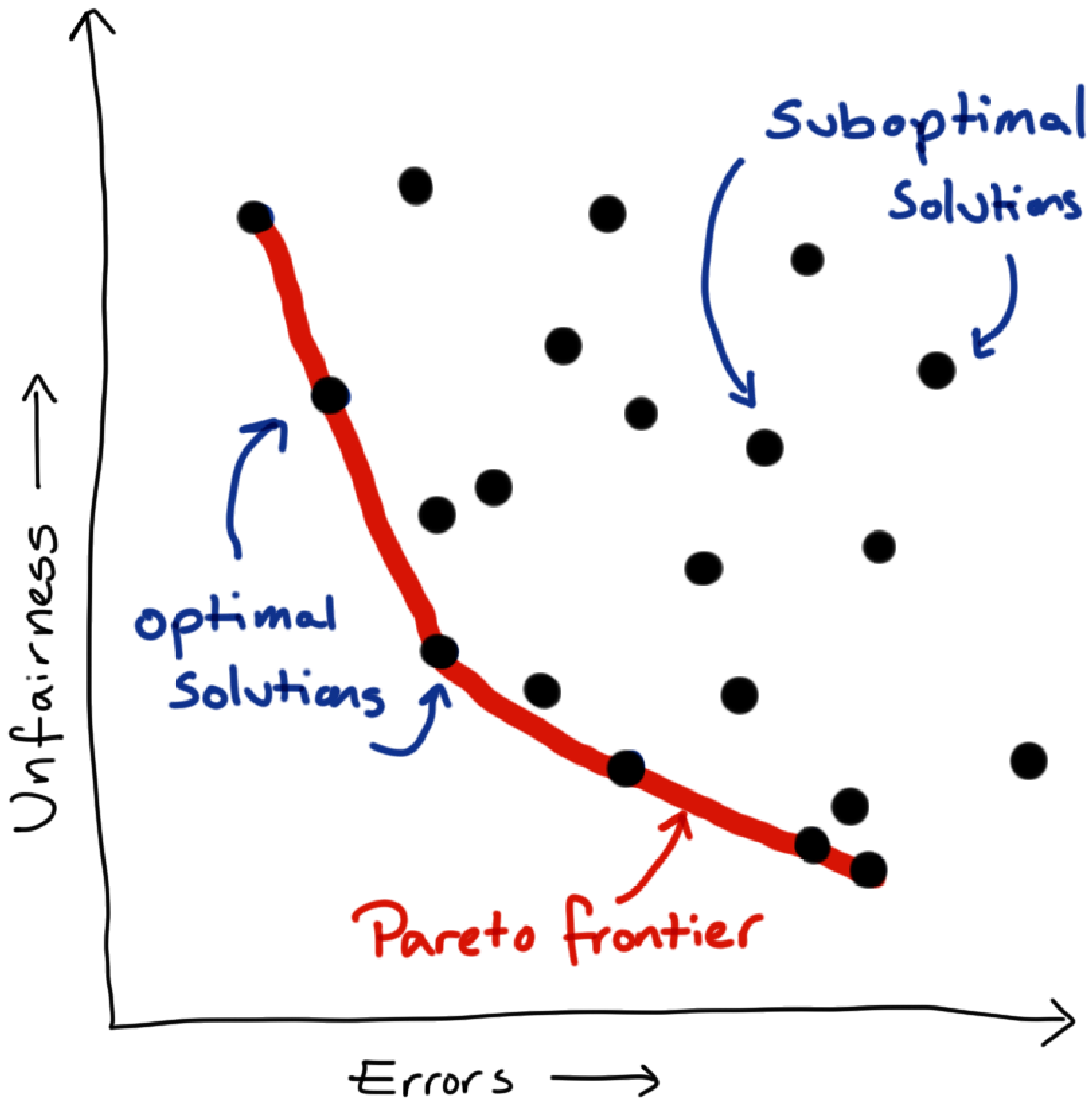
- Even in this simple case there are so many ways you can consider fairness.
- Some metrics rely on knowing the true outcome.
- Sampling issues: inference or tolerance bounds.
- Conditional probability is hard.

Modelling Fairly

- Multi-objective optimisation ill-defined

$$L = w_1 * \text{fit} + w_2 * \text{fairness}$$

- Moving target: how to pick weights?



MATH-70076 Effective Data Science

Other Approaches to Fairness

- **Minority Groups:** Re-weight in loss function or up-sample.
- **Historical Bias:** Forgetting factor to down-weight older observations.
- **Feedback loops:** need direct intervention.
- Meta-modelling one way of doing this.

Wrapping Up

- Optimising for predictive accuracy alone can lead to unjust models.
- Many measures of fairness
- Can implement fairness by constructing appropriate loss functions
- No universal answers, but an exciting area of ongoing research.