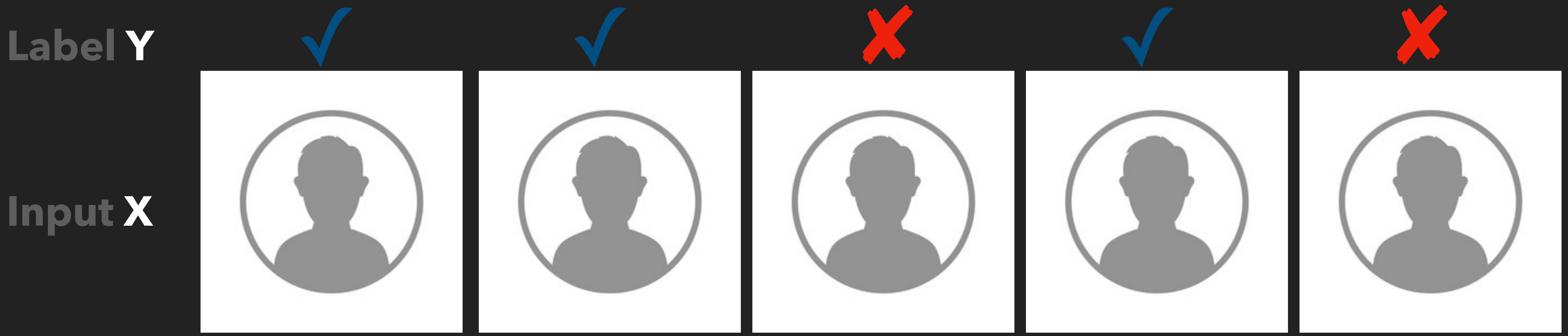
Fair Risk Aggregation

With Bob Williamson (~~ANU~~ Tübingen) and Aditya Krishna Menon (Microsoft, NY)

LEARNING BINARY CLASSIFIERS

Learn a classifier which has maximal accuracy of predicting a target label

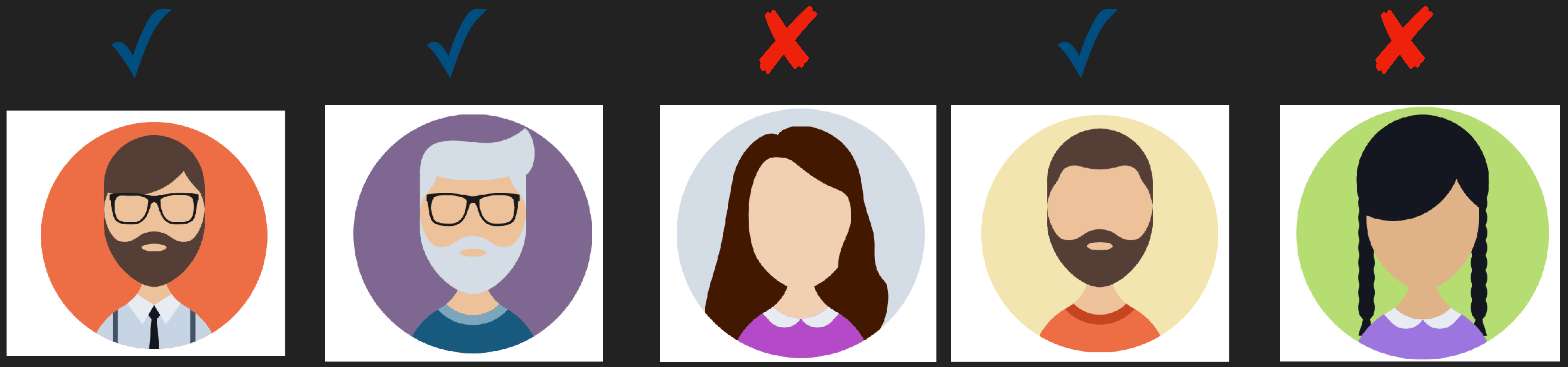


LEARNING FAIR BINARY CLASSIFIERS

Learn a classifier which has maximal accuracy of predicting a target label, **and does not discriminate on some sensitive feature**

Label Y

Input X



Input set X; Label set Y; Sensitive feature S

SET UP

- A feature set X and a label set Y
- A predictor $f : X \to A$ for some action set A
- A class of predictors $\mathscr{F} \subset A^X$
- A probability distribution P over X
- A sensitive feature S is a partition of X

Review of previous approaches

PERFECT FAIRNESS

- Demographic Parity $A \perp\!\!\!\perp S$
  - i.e. $P(A = a \text{ and } S = s) = P(A = a) \cdot P(S = s) \quad \forall a, s$
  - Knowledge of $A = f(X)$ provides no knowledge of the sensitive feature S

- Equalised Odds $A \perp\!\!\!\perp S \mid Y$
  - $P((A = a \text{ and } S = s) \mid Y = y) = P(A = a \mid Y = y) \cdot P(S = s \mid Y = y) \quad \forall a, s, y$
    i.e.
  - Given knowledge of the true label Y, knowledge of the predictions A provides no knowledge of the sensitive feature S.

APPROXIMATE PERFECT FAIRNESS

Can almost never attain perfect fairness: need a **measure** of imperfection; allows trade-offs. Only really done when $S = A = \{1,2\}$ i.e. binary features

- Mean-Difference Score:

$$\mathrm{MD}(f) = |P(\mathsf{A} = 2 \,|\, \mathsf{S} = 2) - P(\mathsf{A} = 2 \,|\, \mathsf{S} = 1)|$$

- Disparate Impact Factor:

$$\mathrm{DI}(f) := \frac{\mathbb{P}(\mathsf{A} = 2 \,|\, \mathsf{S} = 2)}{\mathbb{P}(\mathsf{A} = 2 \,|\, \mathsf{S} = 1)} \wedge \frac{\mathbb{P}(\mathsf{A} = 2 \,|\, \mathsf{S} = 1)}{\mathbb{P}(\mathsf{A} = 2 \,|\, \mathsf{S} = 2)}$$

BEYOND BINARY FEATURES

- Several attempts in the literature…
- Obvious: $\sup_{a,s,s'} |\mathbb{P}(\mathsf{A} = a \,|\, \mathsf{S} = s) - \mathbb{P}(\mathsf{A} = a \,|\, \mathsf{S} = s')|$.
- Combinatorial explosion for large S ☹️
- Use **mutual information** to measure independence

$$\mathrm{MI}(\mathsf{A};\mathsf{S}) = \mathrm{KL}(\mathbb{P}_{\mathsf{AS}} \,\|\, \mathbb{P}_{\mathsf{A}} \cdot \mathbb{P}_{\mathsf{S}})$$

- Approximate fairness becomes $\mathrm{MI}(\mathsf{A};\mathsf{S}) < \epsilon$
- Choice of KL divergence is arbitrary, so use $\mathbb{I}_\varphi(\mu \,\|\, \nu) := \mathbb{E}_\mu\left[\varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right]$

THE TROUBLE WITH INDEPENDENCE BASED APPROACHES TO FAIRNESS

- Suppose A and S are finite

$$\mathrm{MI}(\mathrm{A};\mathrm{S}) = \mathrm{KL}(\mathbb{P}_{\mathrm{AS}} \,||\, \mathbb{P}_{\mathrm{A}} \cdot \mathbb{P}_{\mathrm{S}})$$

$$= \sum_{a,s} \mathbb{P}_{AS}((a,s)) \cdot \log \frac{\mathbb{P}_{AS}(a,s)}{\mathbb{P}_A(a) \cdot \mathbb{P}_S(s)}$$

One needs sample from

1. the marginal distribution of the predictions $\mathbb{P}_A$

2. the conditional distribution $\mathbb{P}_{A|S=s}$ for all sensitive feature value

Plus these constrains, the optimization might not be convex

- A Decision Theoretical Approach

LOSS FUNCTION

- A **loss function** $\ell : Y \times A \rightarrow \mathbb{R}_{\geq 0}$ measuring the disagreement between target label and its prediction.

EXAMPLES OF THE LOSS FUNCTIONS

- Given a sample $Z = \{(x_1, y_1), \cdots, (x_n, y_n)\} \in (X \times Y)^n$ and a predictor h,

★ 0-1 loss function is defined: for all $(x, y) \in Z$,

$$\ell_{0-1}(y, h(x)) := \begin{cases} 0 & \text{if } \ h(x) = y \\ 1 & \text{if } \ h(x) \neq y \end{cases}$$

✤ Square loss function is defined: for all $(x, y) \in Z$,

$$\ell_{sq}(y, h(x)) := (h(x) - y)^2$$

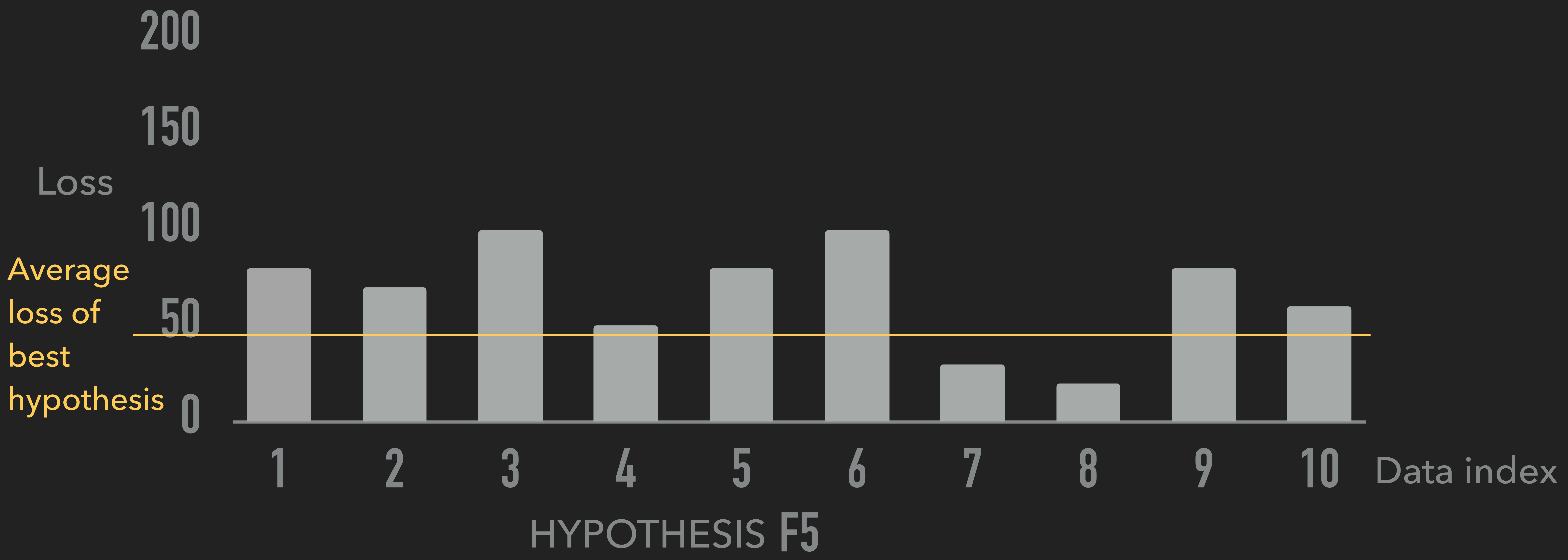
LEARNING GOAL

- Learning goal: Expected loss minimisation

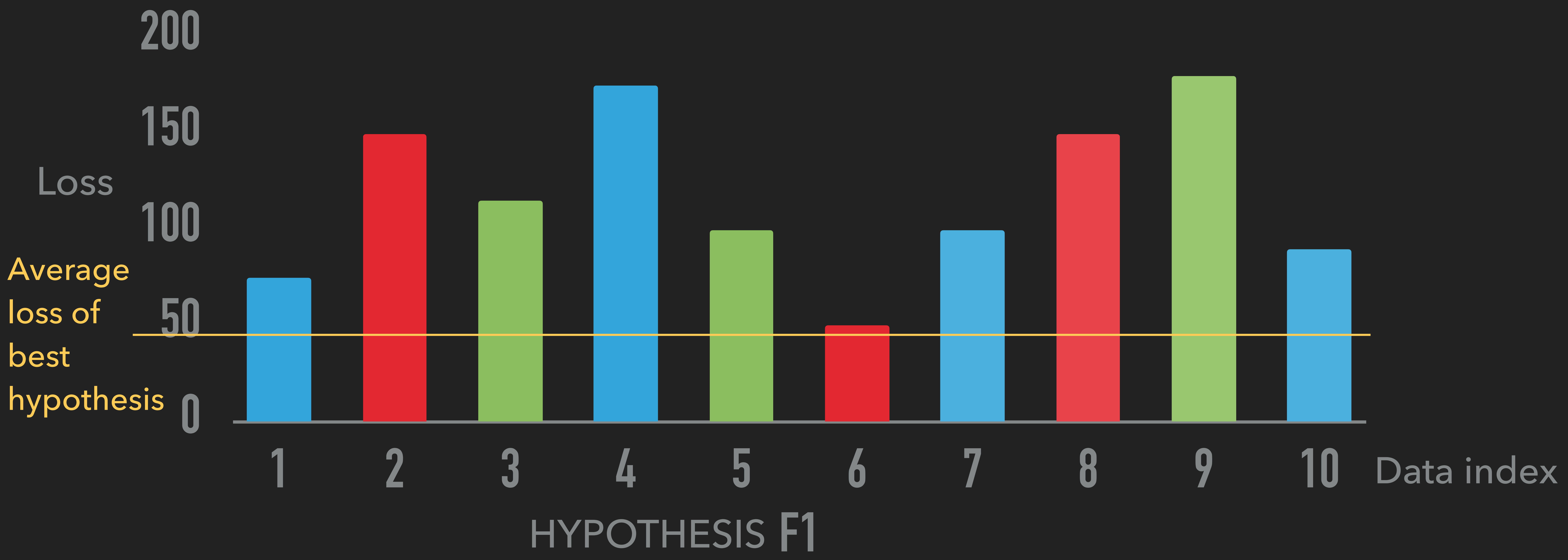
$$\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)\sim P}\,\ell(Y, f(X))$$

$$= \min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, f(x_i))$$

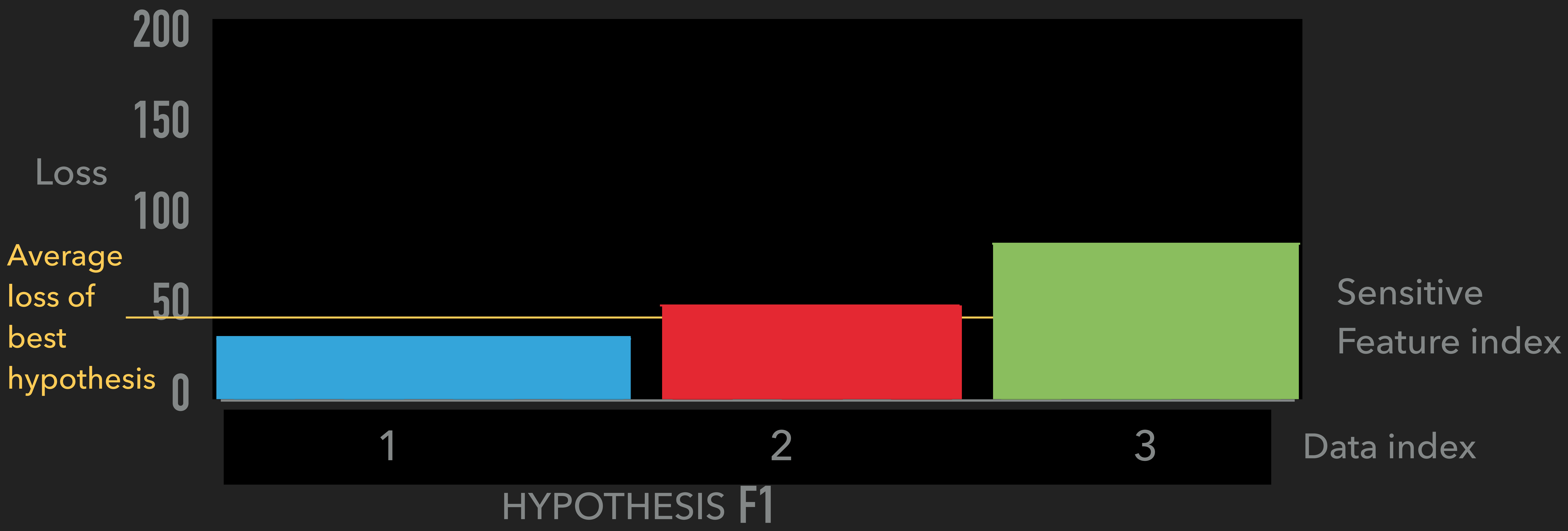
MINIMISING EXPECTED LOSS



MINIMISING EXPECTED LOSS WITH SENSITIVE ATTRIBUTES VISIBLE

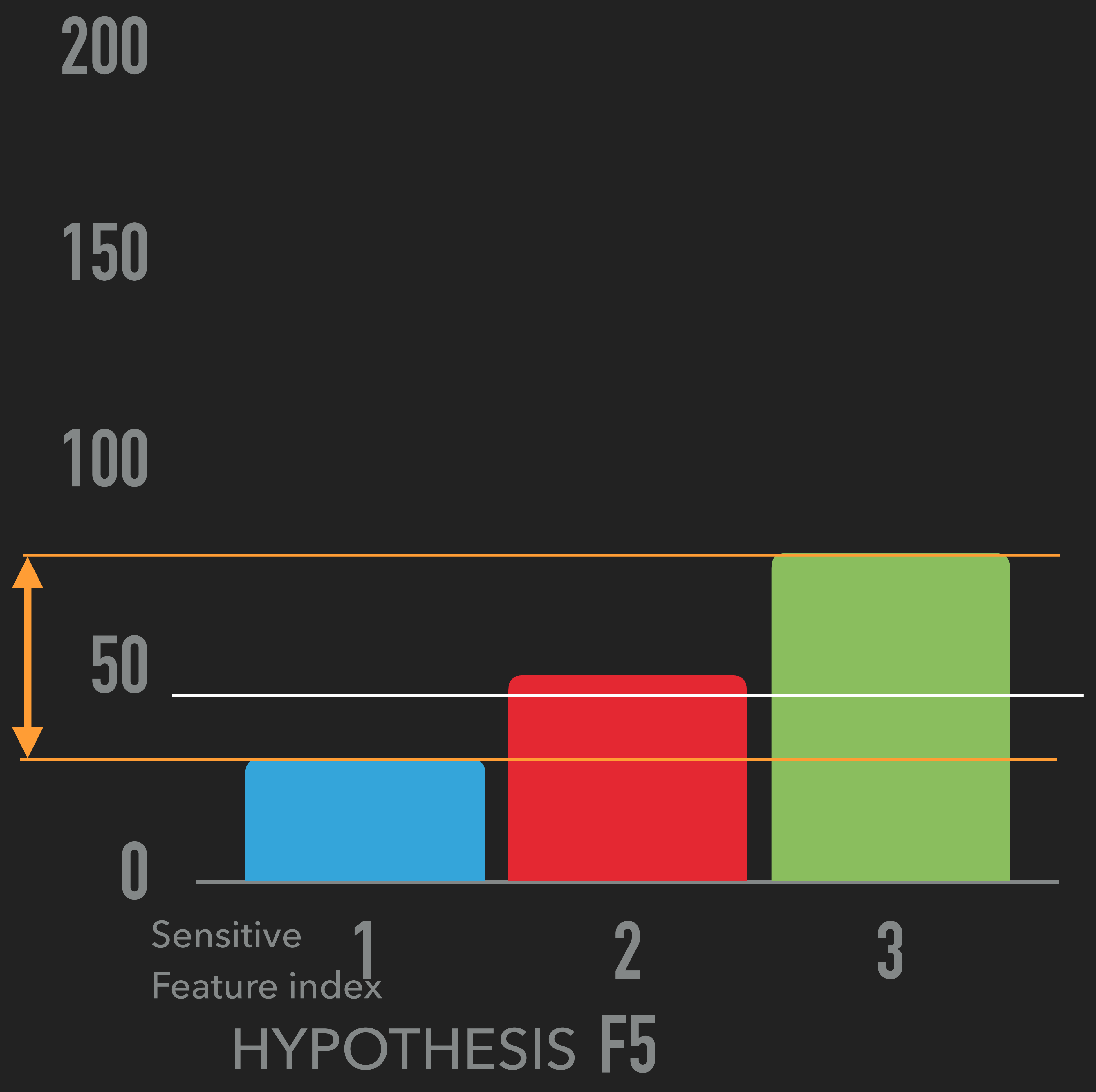


MINIMISING EXPECTED LOSS WITH SENSITIVE ATTRIBUTES VISIBLE



MINIMISING AGGREGATED EXPECTED LOSS

- Standard problem: minimise average loss
- Fairness problem: also take account of variation
- Overall problem: mixture of both



MINIMISING AGGREGATED EXPECTATION AND DEVIATION

- Given the sensitive feature $S = \{S_1, \cdots, S_n\}$, for all $S_i \in S$.

- Consider $\mathrm{R}_f : S \to \mathbb{R}$ viewed as a random variable

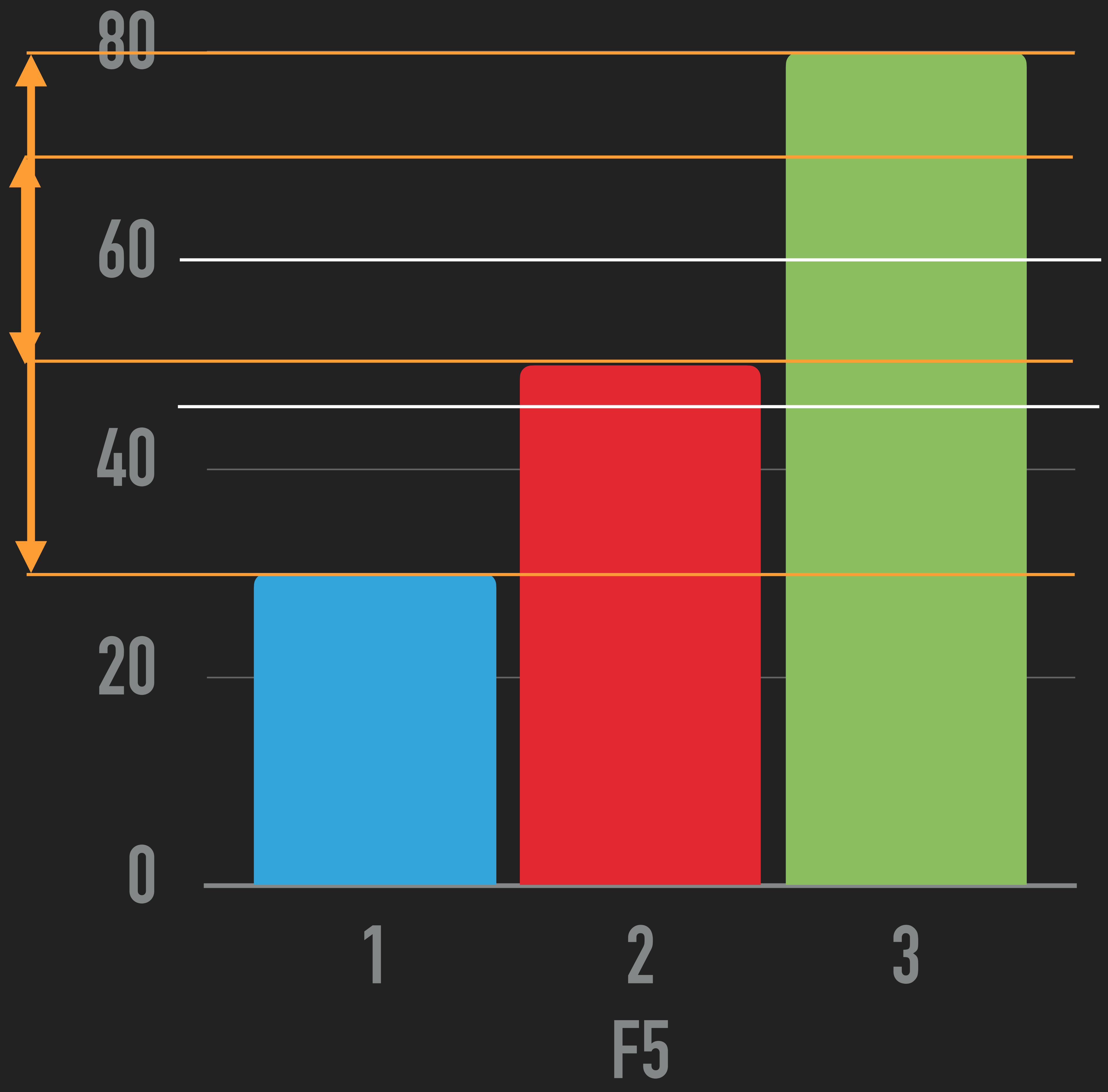
$$\mathrm{R}_f^S : S \ni S_i \mapsto \mathbb{E}_P[\ell(\mathrm{Y}, f(\mathrm{X})) \,|\, \mathrm{S} = S_i]$$

- Standard ERM:

$$\min_{f \in \mathscr{F}} \mathbb{E}(\mathrm{R}^S{}_f)$$

- Fairness Augmented ERM:

$$\min_{f \in \mathscr{F}} \mathbb{E}(\mathrm{R}^S{}_f) + \mathscr{D}(\mathrm{R}^S{}_f) = \min_{f \in \mathscr{F}} \mathscr{R}(\mathrm{R}^S{}_f)$$



WHAT AXIOM SHOULD $\mathscr{R}$ SATISFIES?

- **Positive homogeneity** is desirable, but not essential
  - It provides an invariance to scaling the loss, and simplifies the maths
- **Covexity and Monotonicity** are desirable, these two ensures the overall objective function remains convex
- **Translation invariance** means if we replace ℓ by $\ell + C$ we have not changed our measure of fairness
- **Law invariance** ensures the risk only depends on the distribution, not the indexing

WHAT AXIOM SHOULD $\mathscr{R}$ SATISFIES?

- **Aversity** ensures deviation from perfect fairness is penalised

$$E(\ell) \leq \mathscr{R}(\ell)$$

- **Lower semicontinuity** is a technical assumption that avoids problems with limits

WHAT IS $\mathscr{R}$ IN DECISION THEORY?

-Maxmin with probabilistic sophistication

FAIR RISK AGGREGATOR

- **We can always write**

$$\mathscr{R}(\mathrm{R}) = \mathbb{E}(\mathrm{R}) + \mathscr{D}(\mathrm{R})$$

Fair risk aggregator

Deviation measure

- **the corresponding fairness/ inequality measure**

$$I(f) = 1 - \frac{\mathscr{R}(f)}{\mathbb{E}(f)}$$

TAKING INTO ACCOUNT THE FAIRNESS WITHIN SENSITIVE FEATURES

- Given the sensitive feature $S = \{S_1, \cdots, S_n\}$, for all $S_i \in S$
- The random variable becomes

$$\mathrm{R}_f^S : S \ni S_i \mapsto \mathscr{R}_{s_i}(\ell(\mathrm{Y}, f(\mathrm{X})) \cdot 1_{S_i})$$

- Fairness Augmented ERM:

$$\min_{f \in \mathscr{F}} \mathscr{R}(\mathrm{R}^S{}_f) = \min_{f \in \mathscr{F}} \mathscr{R}^*(\mathrm{R}_f)$$

A IMPOSSIBILITY RESULT

- *can not ensure same aggregator independent of choice of Sensitive Feature -* **choice of** *Sensitive Feature* **essentially matters**

FORMAL THEOREM

- Aggregators are all -maxmin with linear u and strictly monotonic.
- We use ℓ to represent the random variable (lottery/prospect) over X.
- For any two the sensitive features $S = \{S_1, \cdots, S_n\}$ and $P = \{P_1, \cdots, P_n\}$,

✤ the random variables are

$$\mathrm{R}_f^S : S \ni S_i \mapsto \mathcal{R}_{S_i}(\ell \cdot 1_{S_i}) \text{ and } \mathrm{R}_f^P : P \ni P_i \mapsto \mathcal{R}_{P_i}(\ell \cdot 1_{P_i})$$

$$\mathcal{R}(\mathrm{R}^S{}_f) = \mathcal{R}(\mathrm{R}^P{}_f) \text{ for all } \ell.$$

✤ If and only if all aggregators are expectations.

FURTHER WORK OF THIS PROJECT DONE

CONCLUSION

- New approach to fairness in ML problems:
  - Key principle: *never actually "touch" the prediction … just its loss* ("strongly typed ML!")
  - Based on decision theory : Interpretations can be known from DT.