FAIRNESS RISK MEASURES

FAIRNESS RISK MEASURES

Robert C. Williamson

Australian National University

Aditya Menon

Google AI

LOSS FUNCTIONS – OUTCOME CONTINGENT UTILITIES

LOSS FUNCTIONS – OUTCOME CONTINGENT UTILITIES

- Wald's abstraction: a loss function

$$\ell : Y \times A \to \mathbb{R}_+ \cup \{+\infty\} =: \overline{\mathbb{R}}$$

Label space — Action space

LOSS FUNCTIONS – OUTCOME CONTINGENT UTILITIES

- Wald's abstraction: a loss function

$$\ell : \underbrace{Y}_{\text{Label space}} \times \underbrace{A}_{\text{Action space}} \to \mathbb{R}_+ \cup \{+\infty\} =: \overline{\mathbb{R}}$$

- $a \mapsto \ell(y, a)$ is an outcome contingent utility

LOSS FUNCTIONS – OUTCOME CONTINGENT UTILITIES

- Wald's abstraction: a loss function

$$\ell : \underbrace{Y}_{\text{Label space}} \times \underbrace{A}_{\text{Action space}} \to \mathbb{R}_+ \cup \{+\infty\} =: \overline{\mathbb{R}}$$

- $a \mapsto \ell(y, a)$ is an outcome contingent utility

- Learning goal: expected risk minimisation

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

LOSS FUNCTIONS – OUTCOME CONTINGENT UTILITIES

- Wald's abstraction: a loss function

$$\ell : \underbrace{Y}_{\text{Label space}} \times \underbrace{A}_{\text{Action space}} \to \mathbb{R}_+ \cup \{+\infty\} =: \overline{\mathbb{R}}$$

- $a \mapsto \ell(y, a)$ is an outcome contingent utility

- Learning goal: expected risk minimisation

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

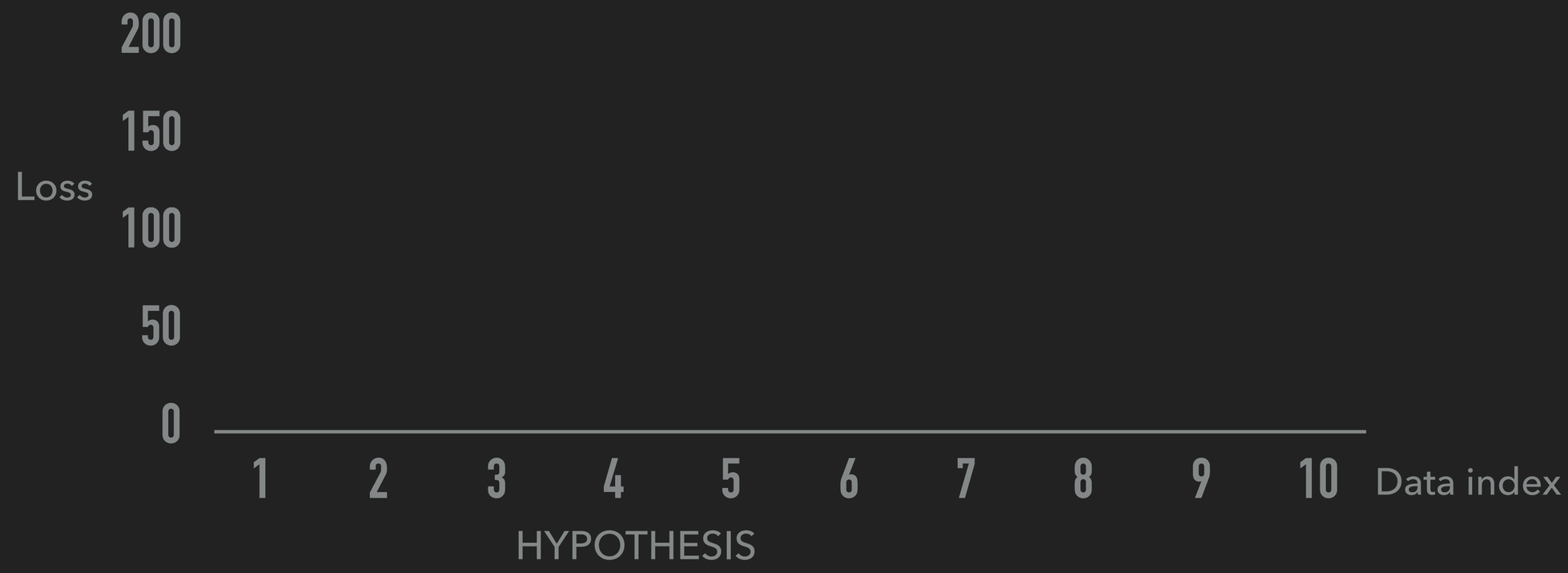
- In practice: empirical risk minimisation

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

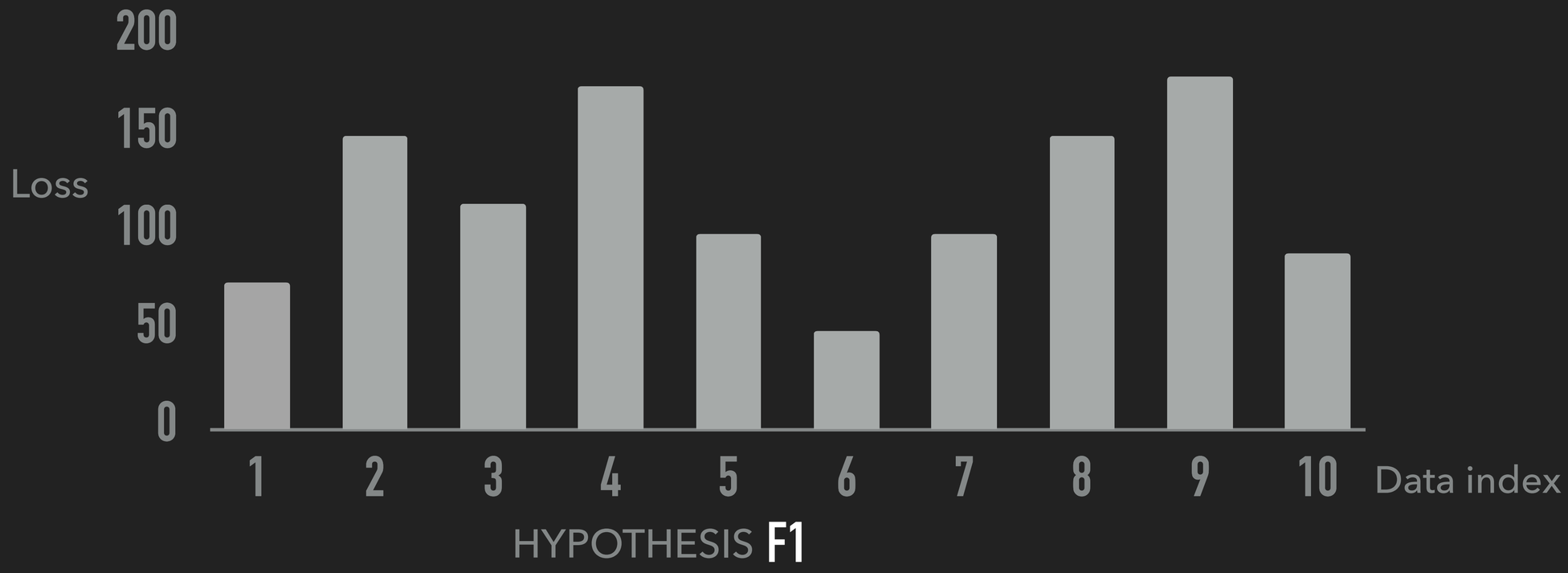
$$= \min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, f(x_i))$$

MINIMISING EMPIRICAL RISK

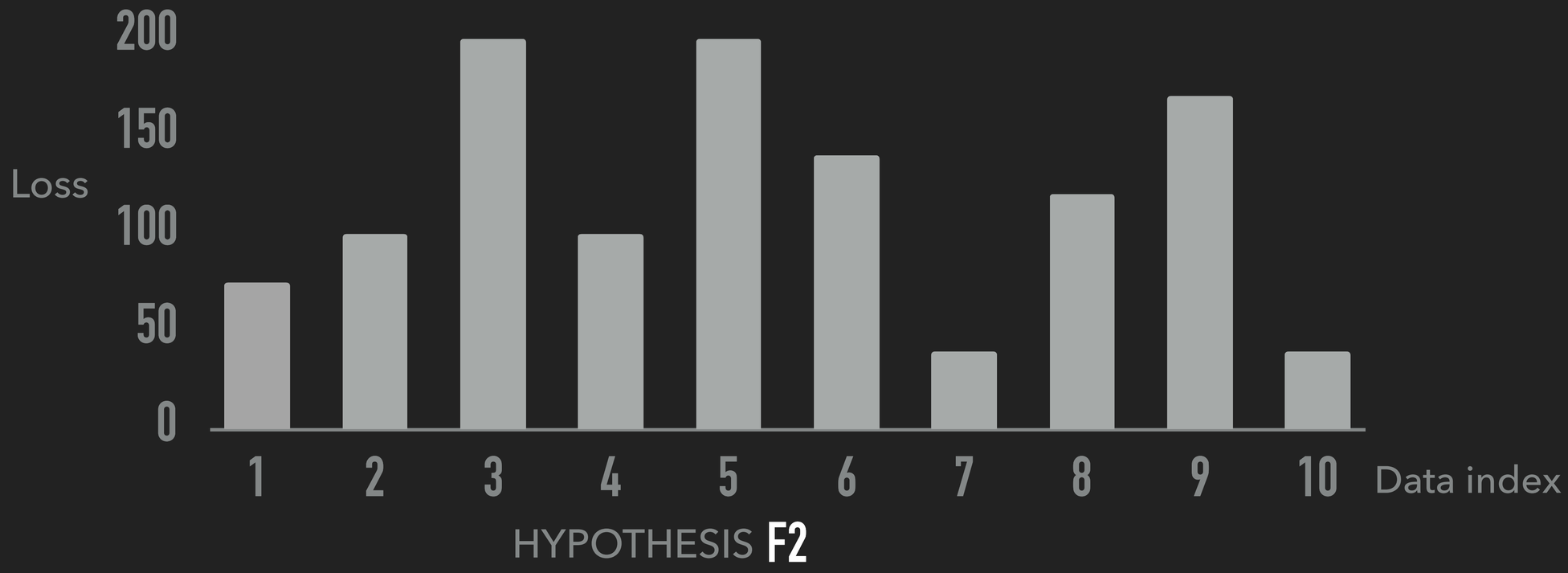
MINIMISING EMPIRICAL RISK



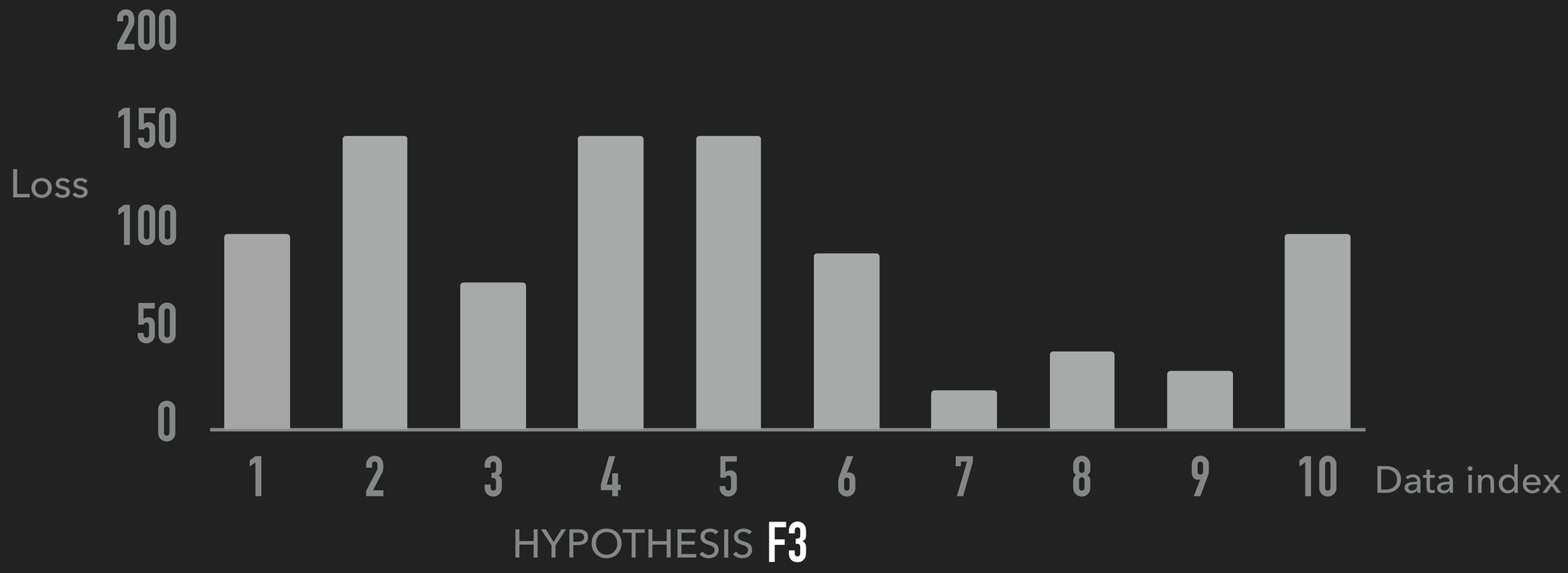
MINIMISING EMPIRICAL RISK



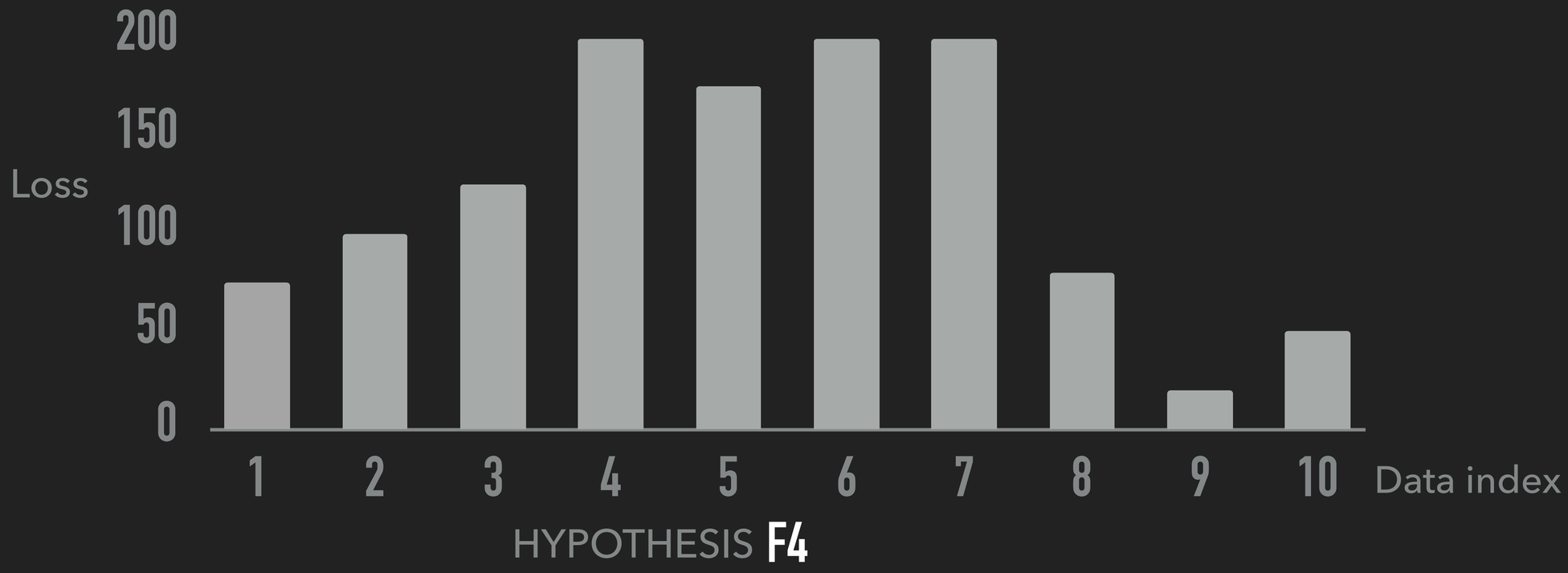
MINIMISING EMPIRICAL RISK



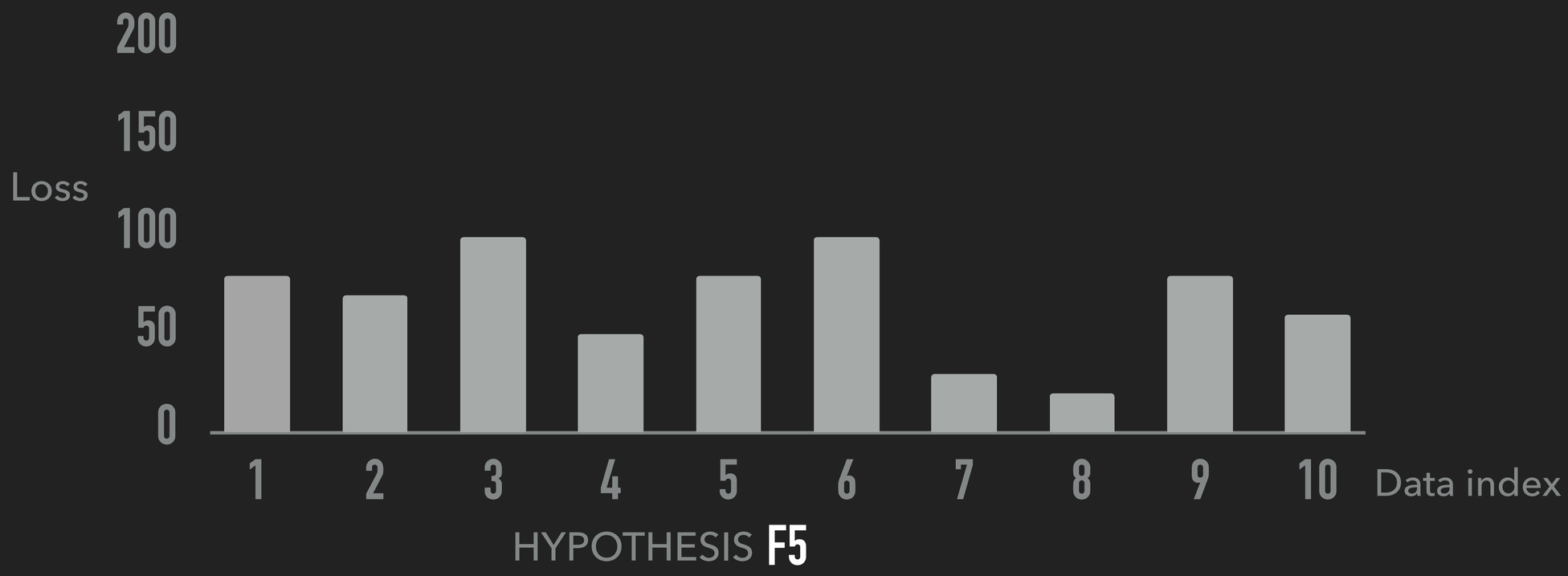
MINIMISING EMPIRICAL RISK



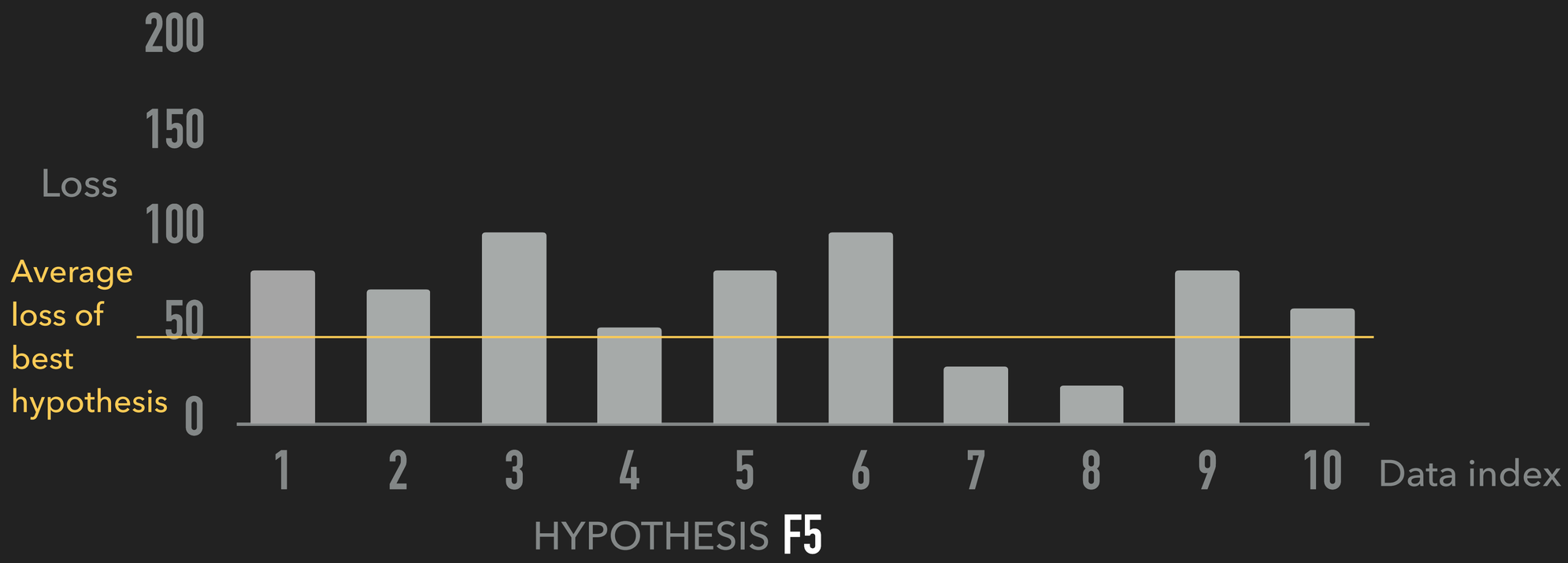
MINIMISING EMPIRICAL RISK



MINIMISING EMPIRICAL RISK

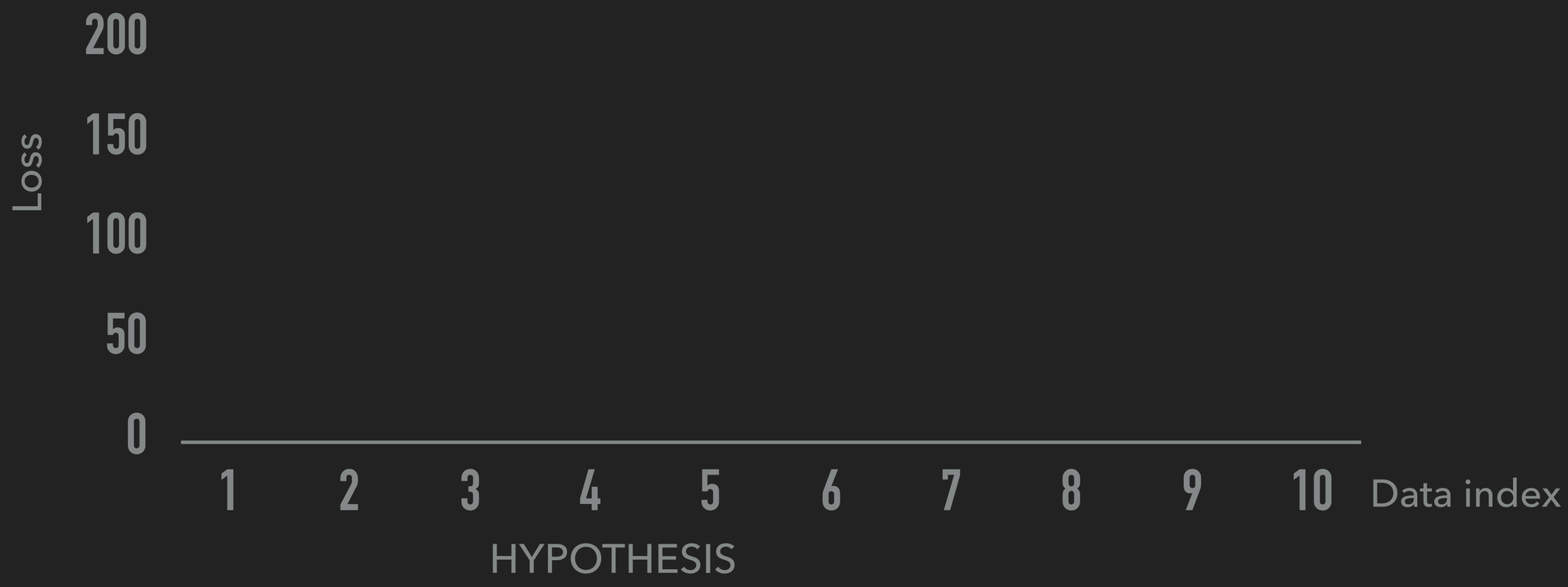


MINIMISING EMPIRICAL RISK

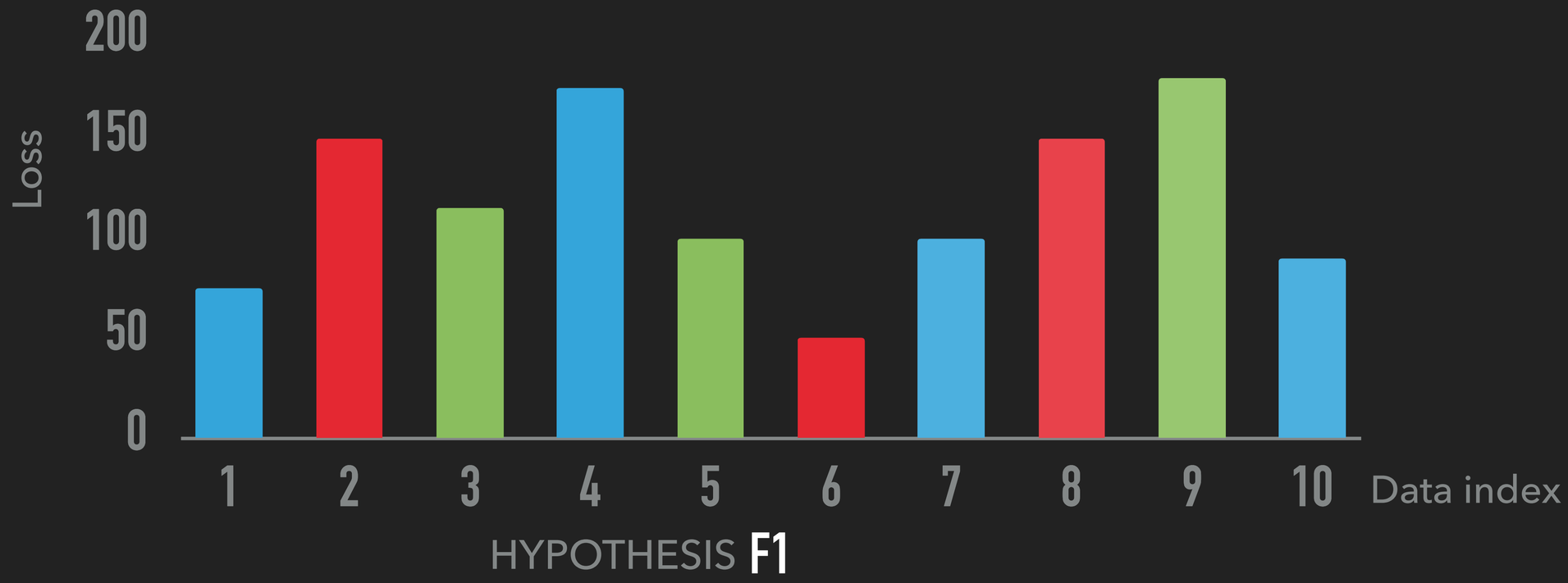


MINIMISING EMPIRICAL RISK WITH SENSITIVE ATTRIBUTES VISIBLE

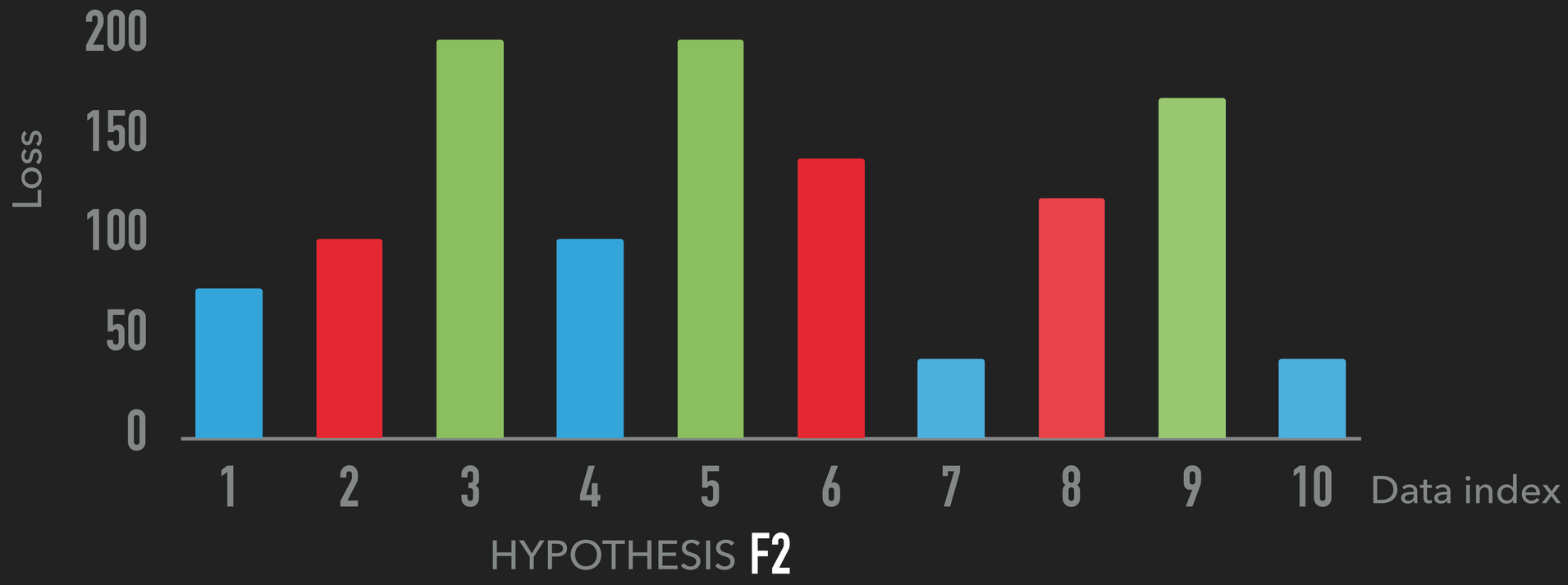
MINIMISING EMPIRICAL RISK WITH SENSITIVE ATTRIBUTES VISIBLE



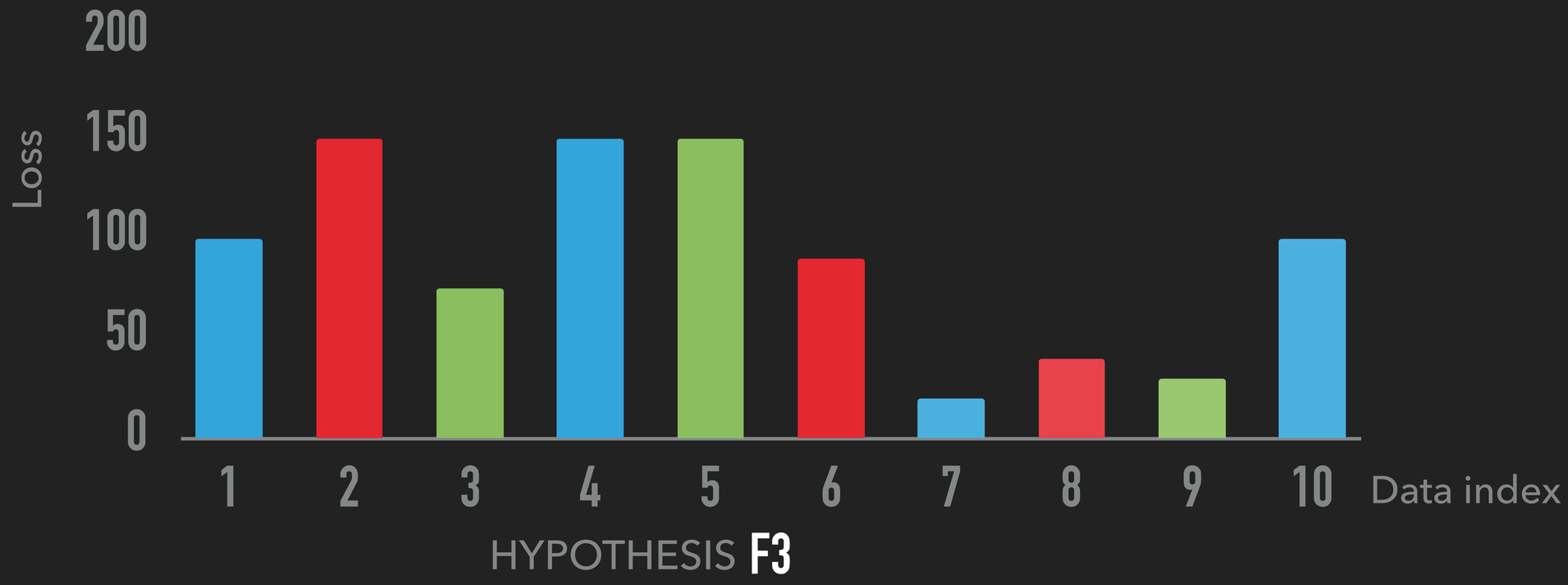
MINIMISING EMPIRICAL RISK WITH SENSITIVE ATTRIBUTES VISIBLE



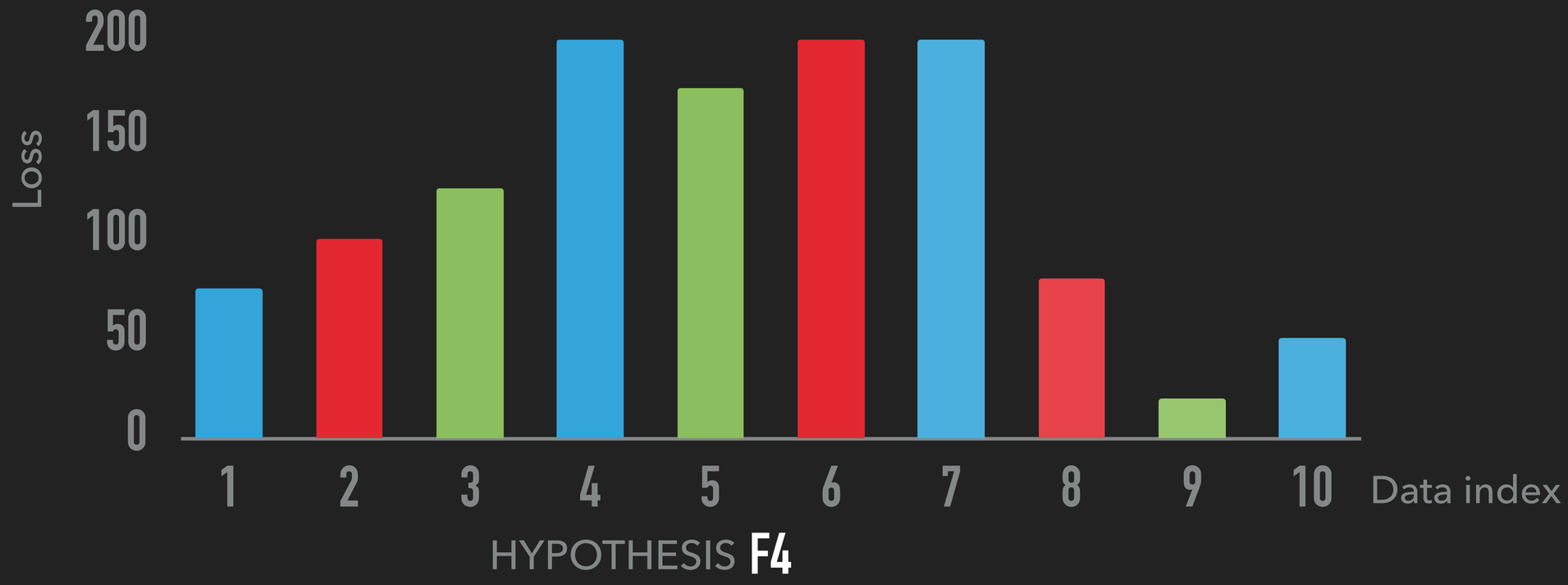
MINIMISING EMPIRICAL RISK WITH SENSITIVE ATTRIBUTES VISIBLE



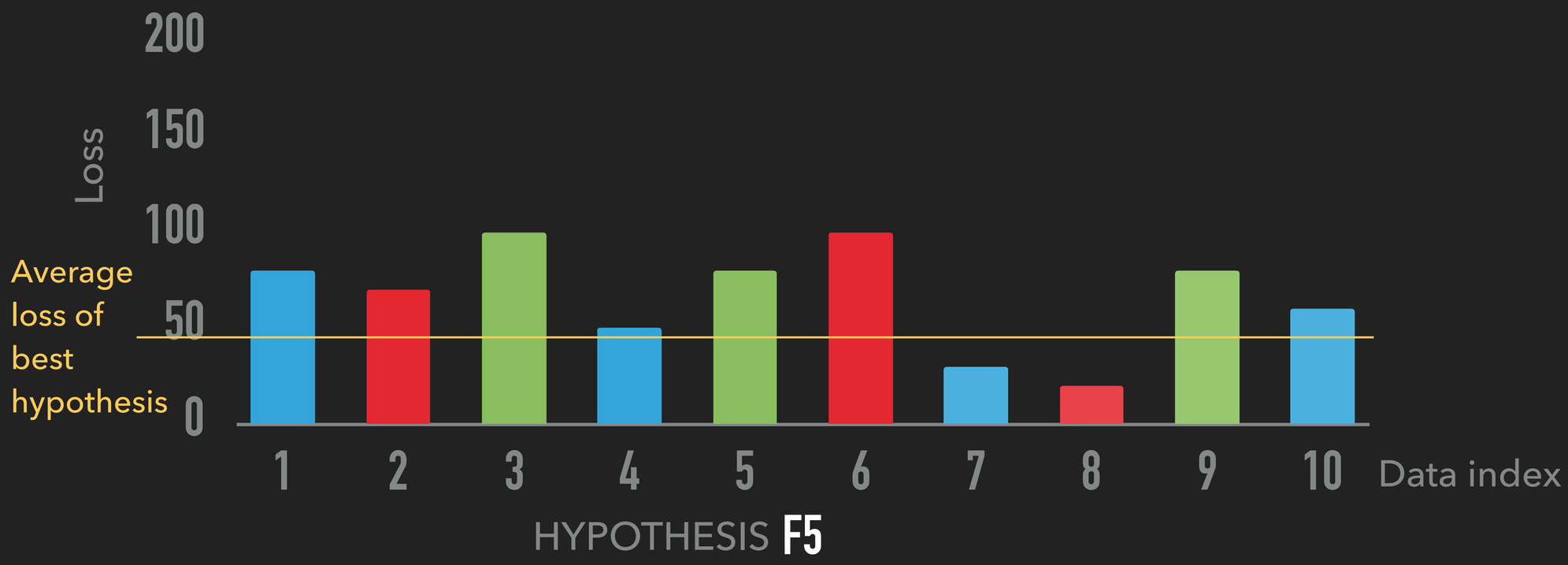
MINIMISING EMPIRICAL RISK WITH SENSITIVE ATTRIBUTES VISIBLE



MINIMISING EMPIRICAL RISK WITH SENSITIVE ATTRIBUTES VISIBLE

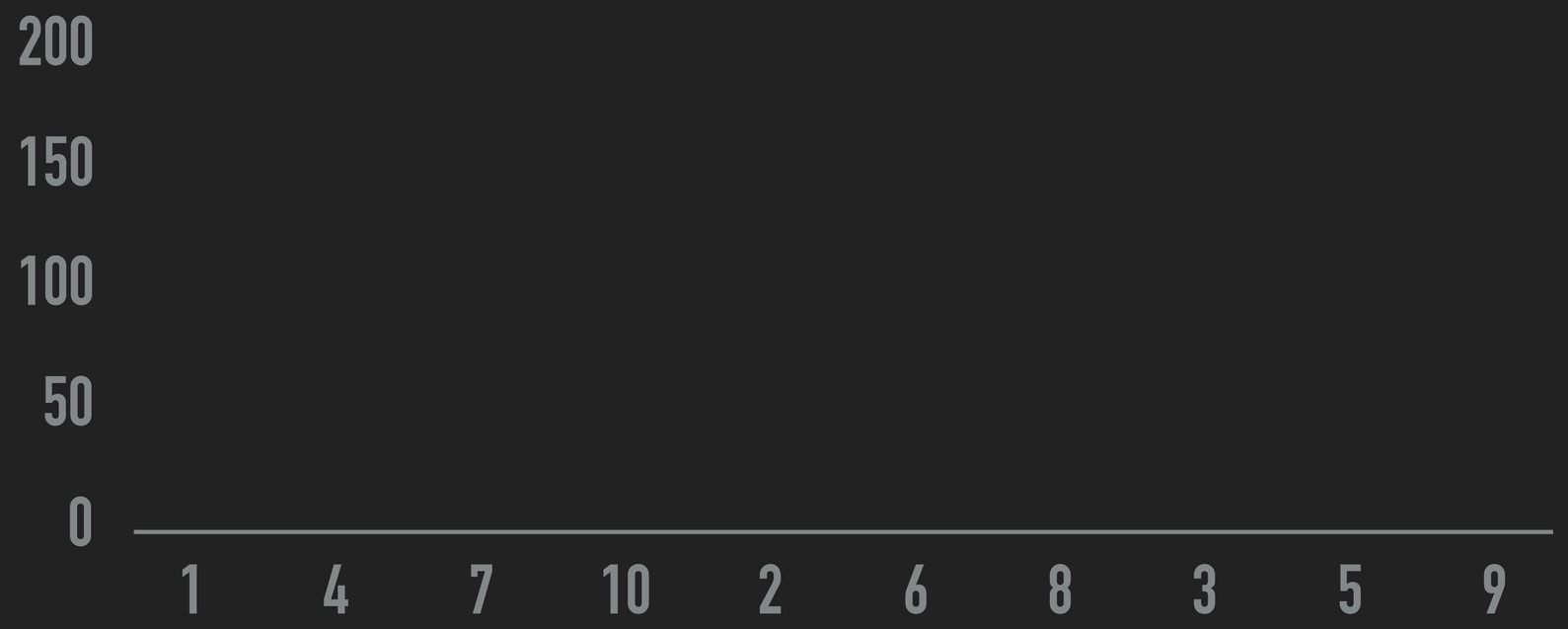


MINIMISING EMPIRICAL RISK WITH SENSITIVE ATTRIBUTES VISIBLE

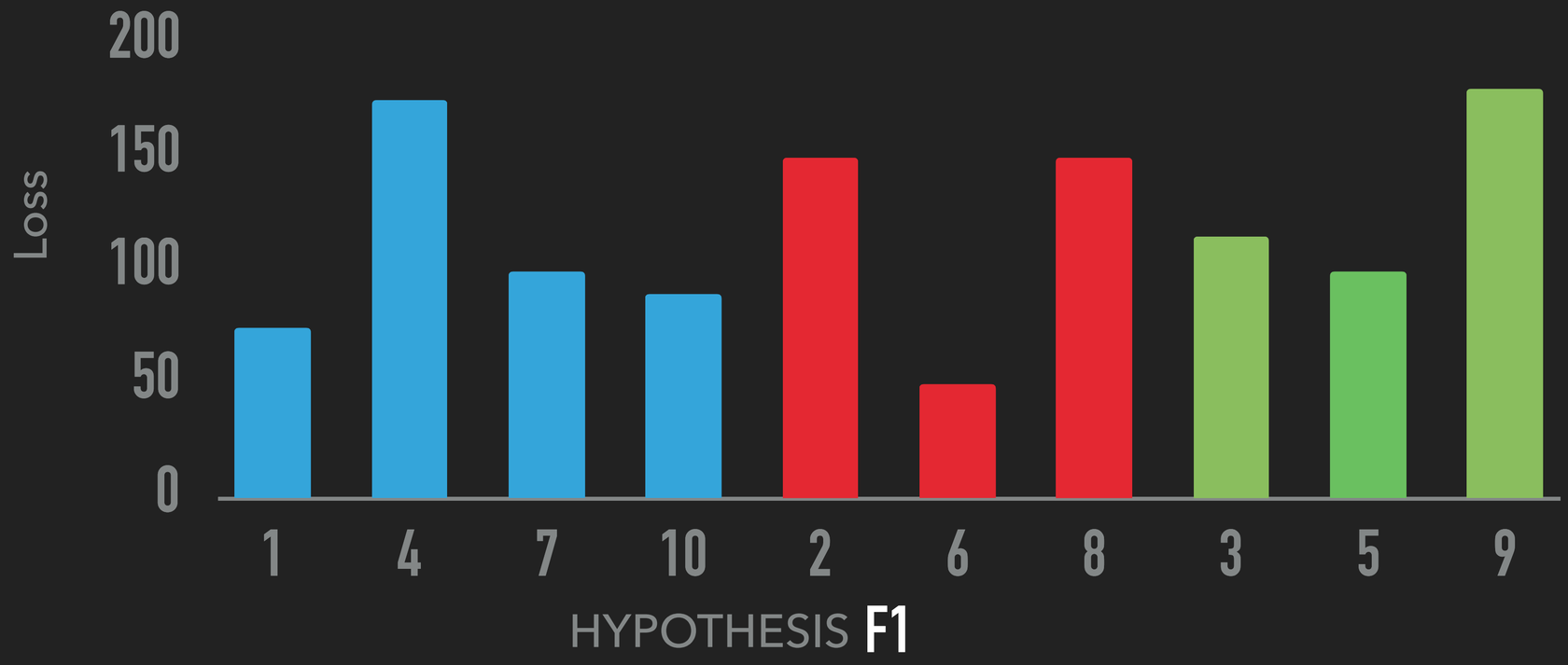


MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE

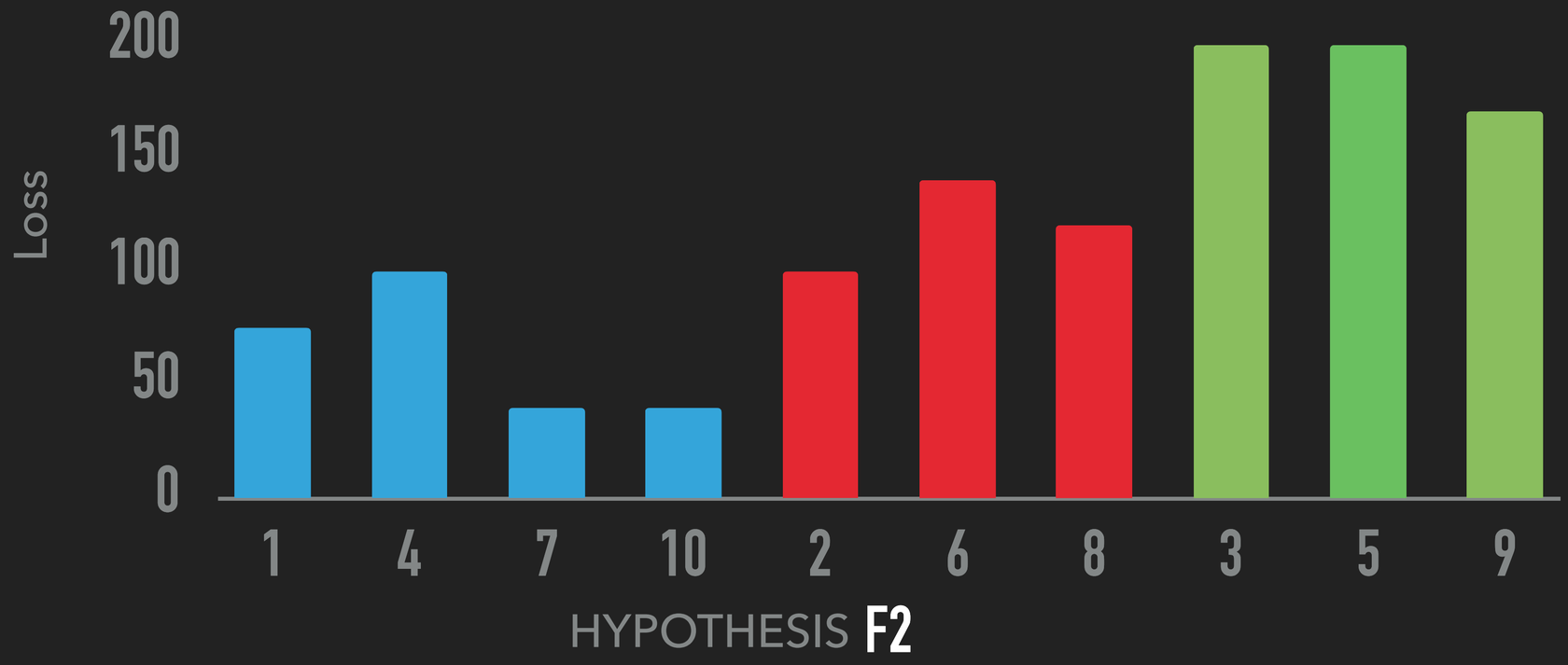
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



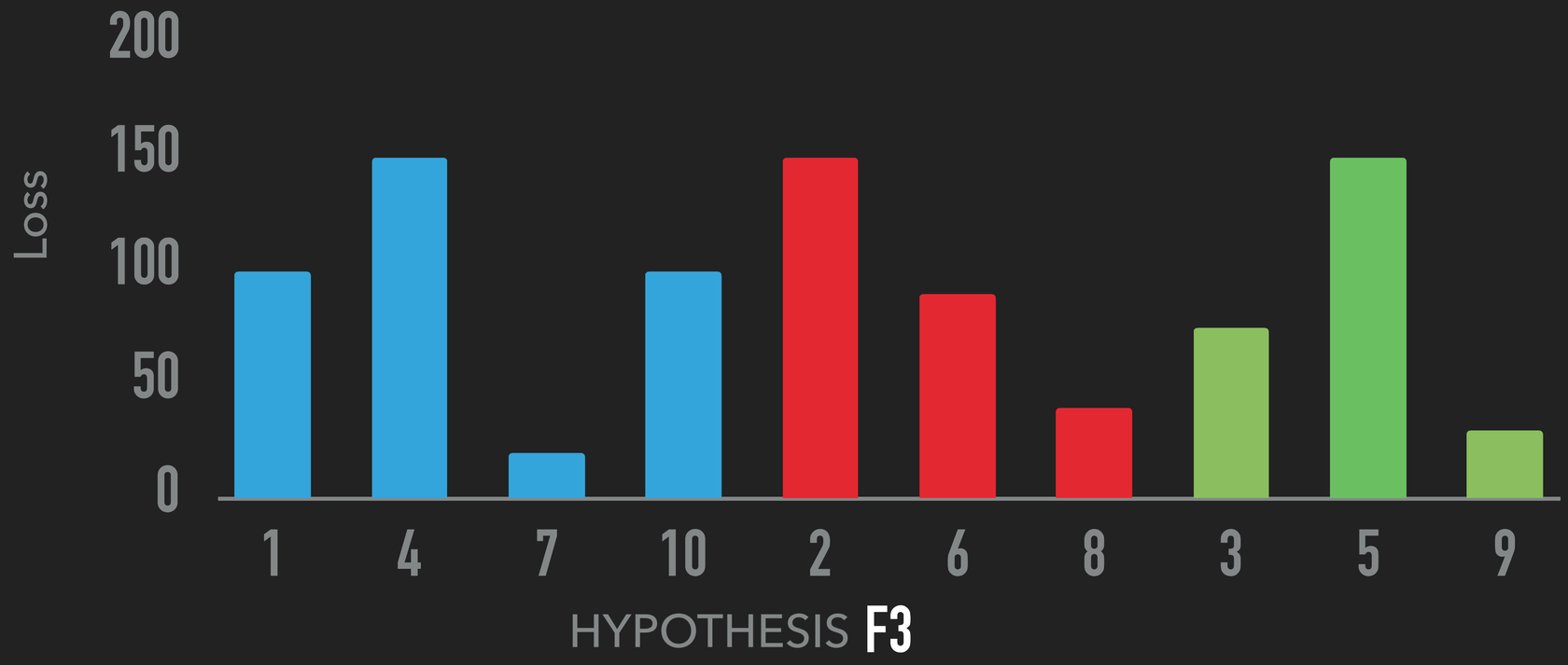
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



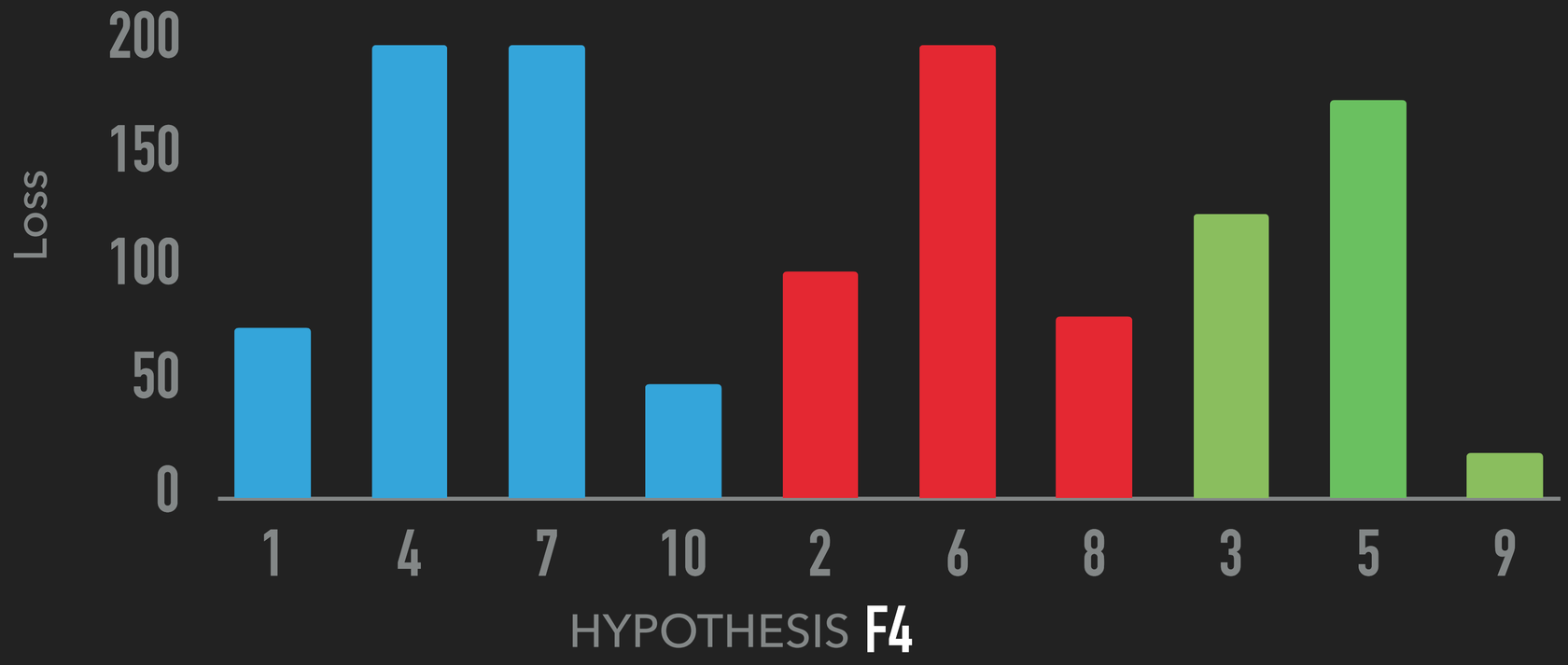
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



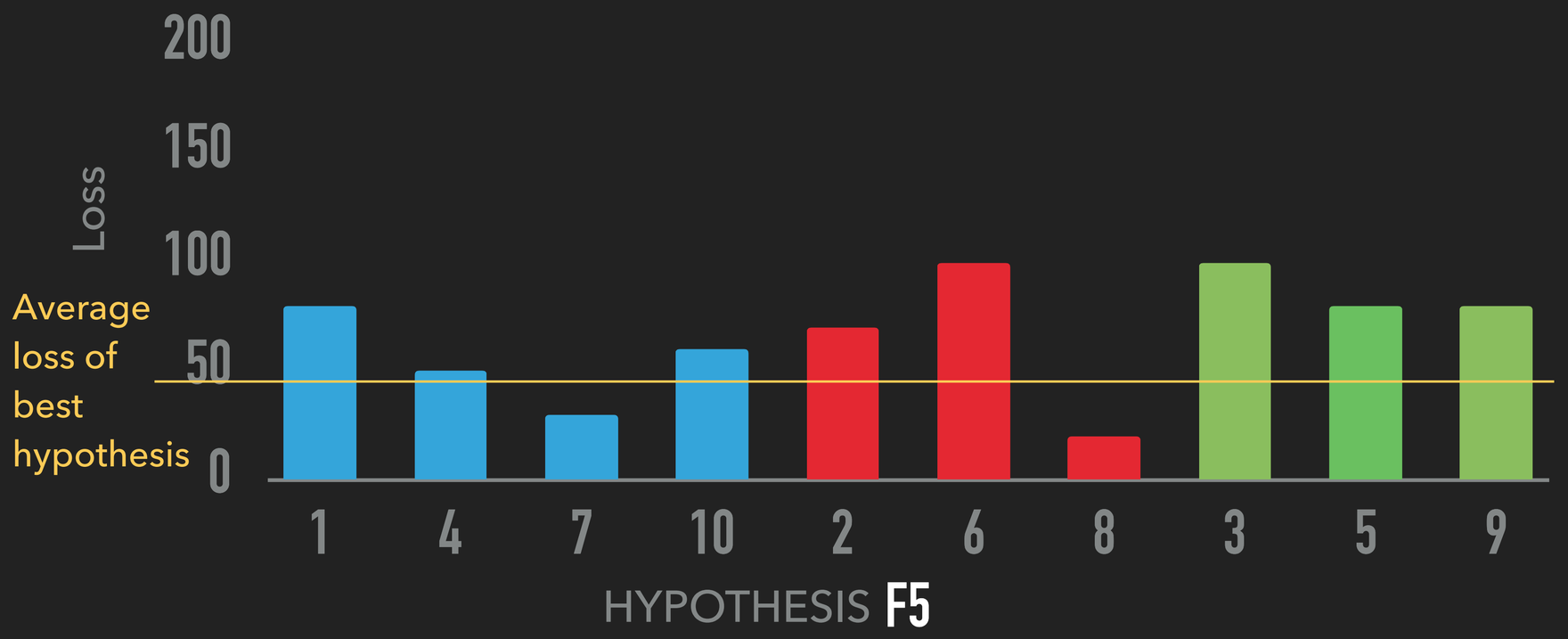
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



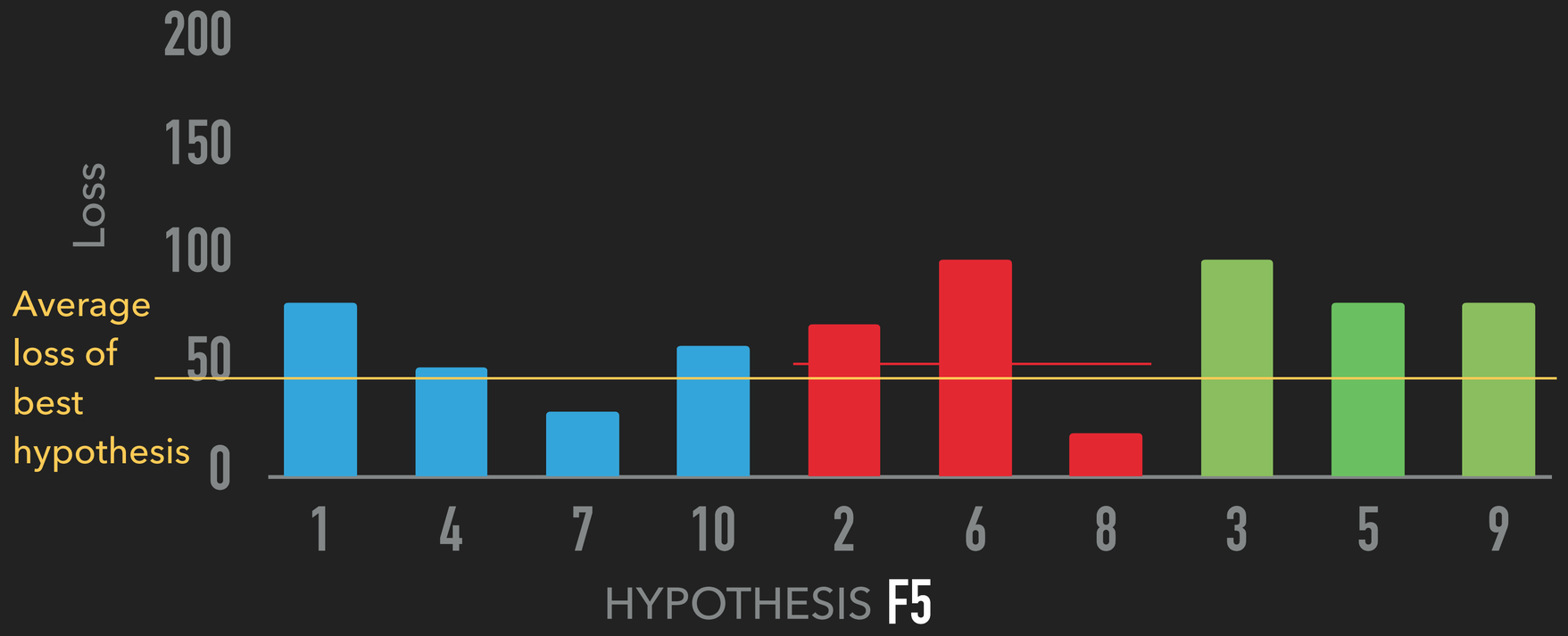
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



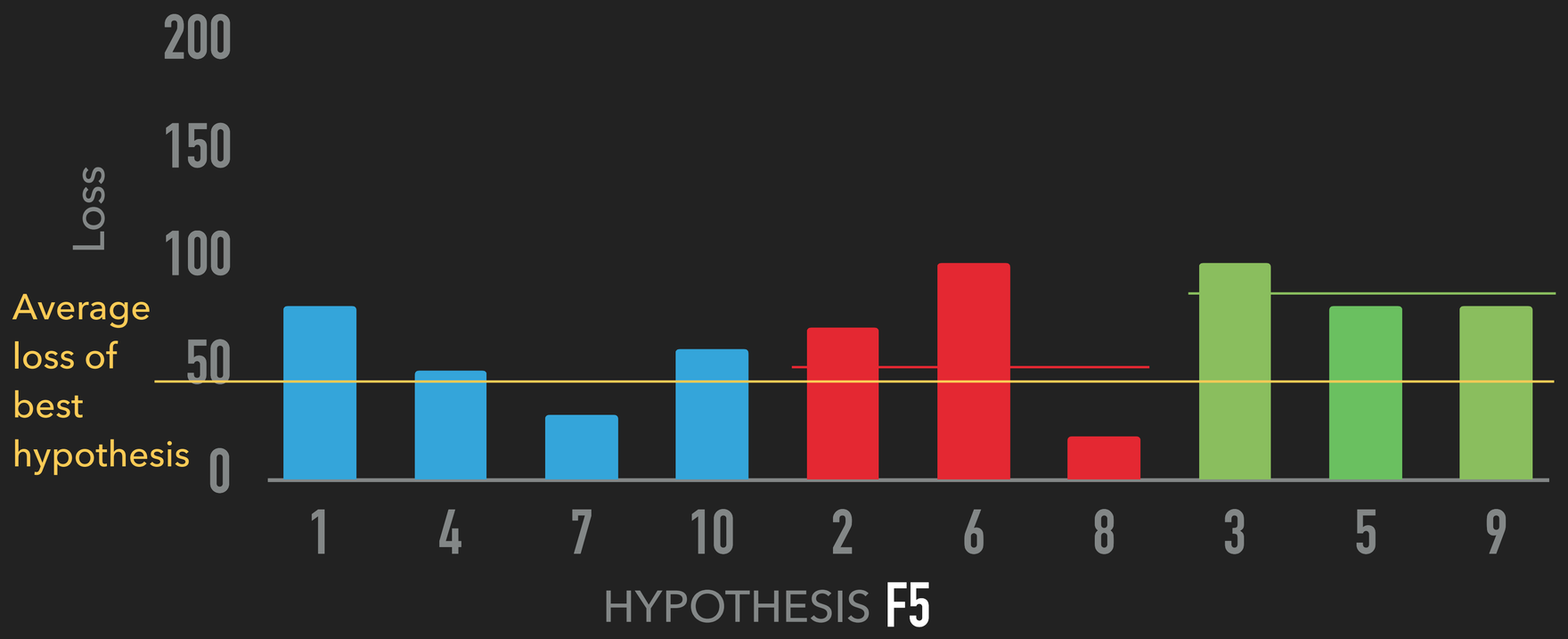
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



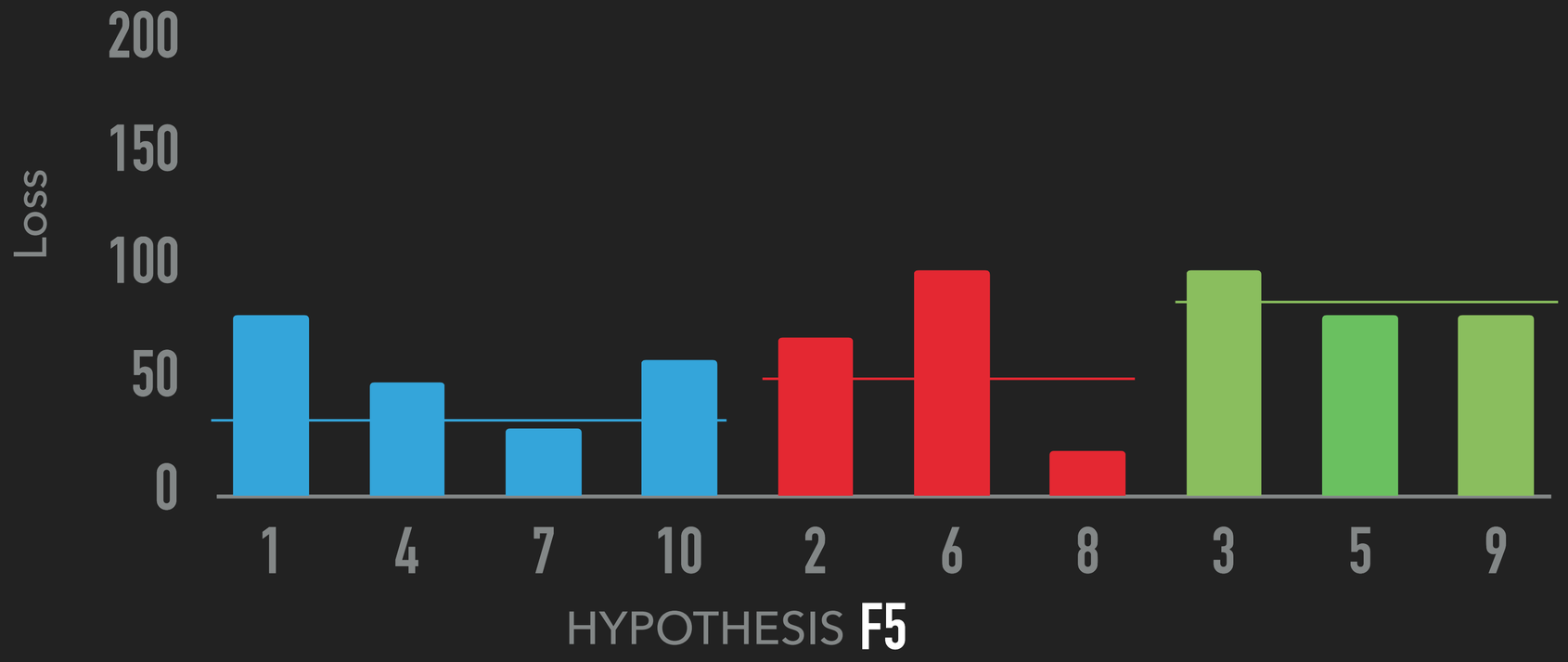
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



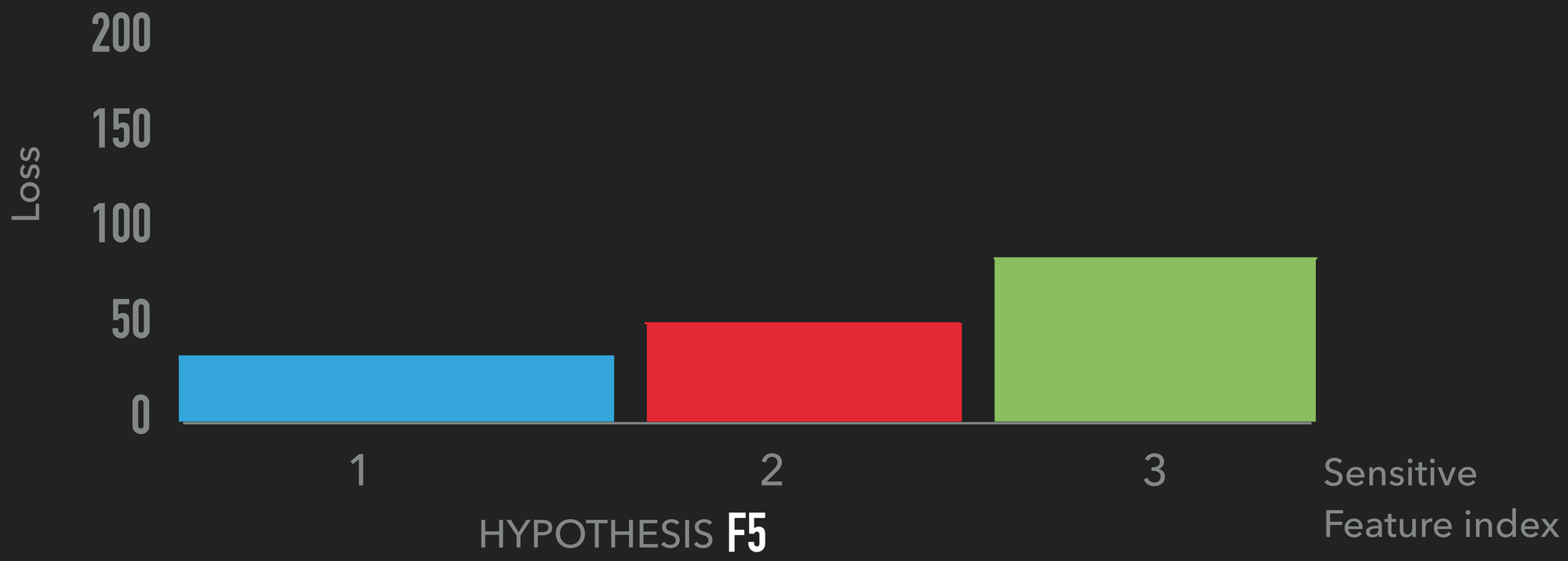
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



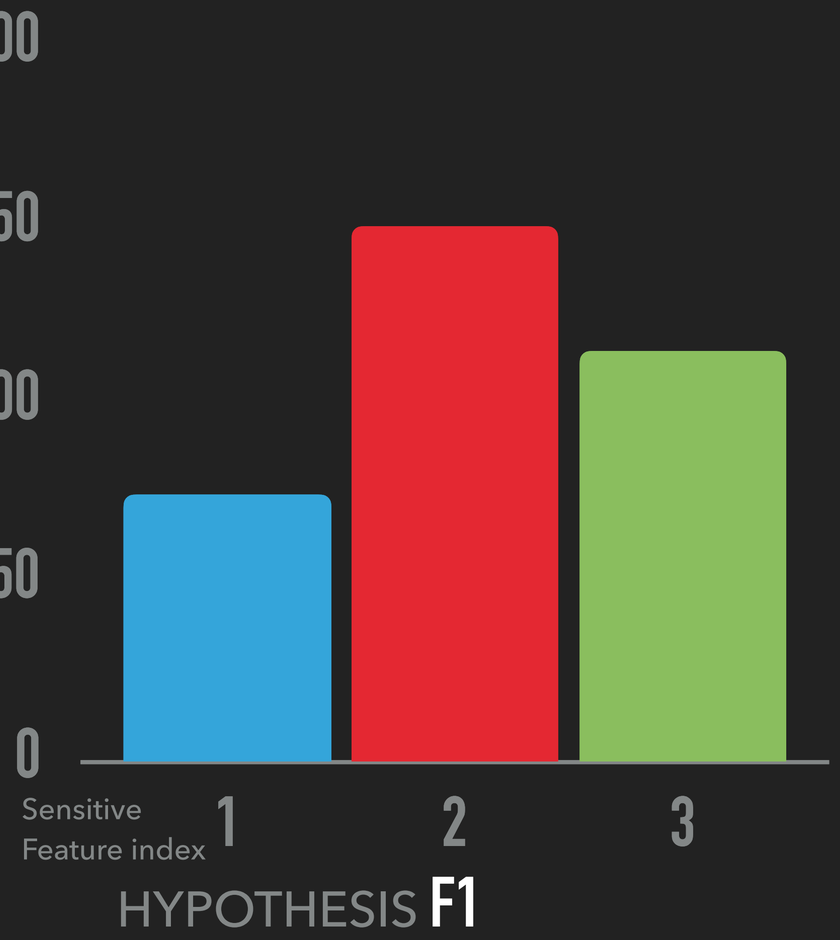
MINIMISING EMPIRICAL RISK REINDEXING PER SENSITIVE ATTRIBUTE



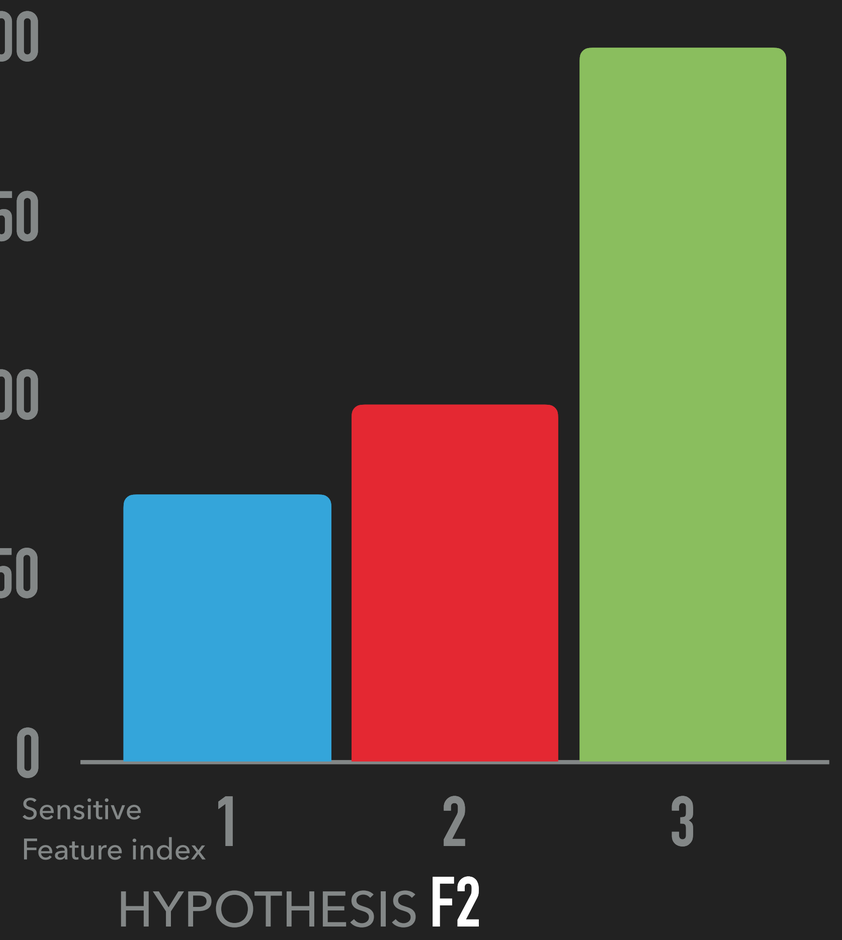
MINIMISING AGGREGATED EMPIRICAL RISK

MINIMISING AGGREGATED EMPIRICAL RISK

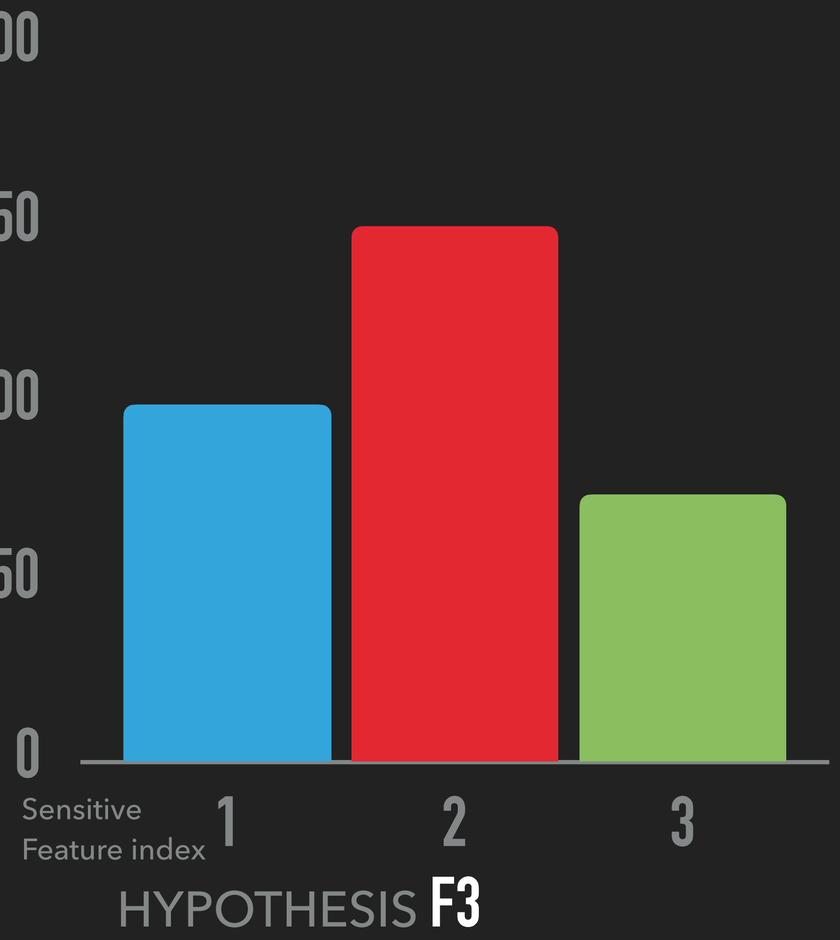
MINIMISING AGGREGATED EMPIRICAL RISK



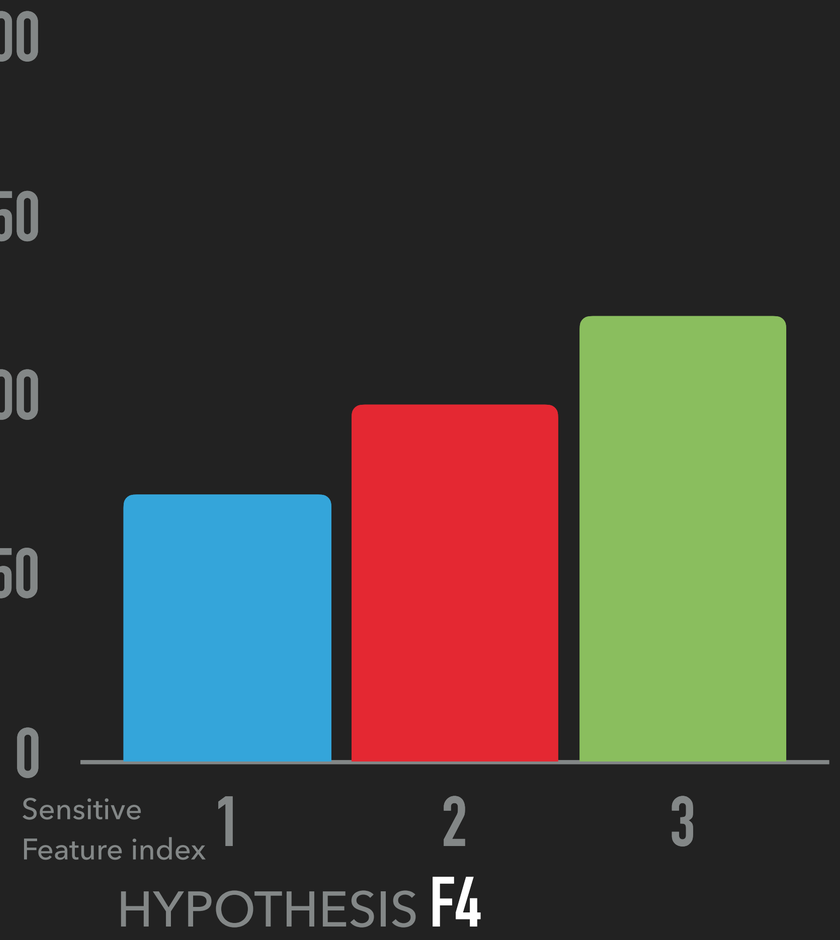
MINIMISING AGGREGATED EMPIRICAL RISK



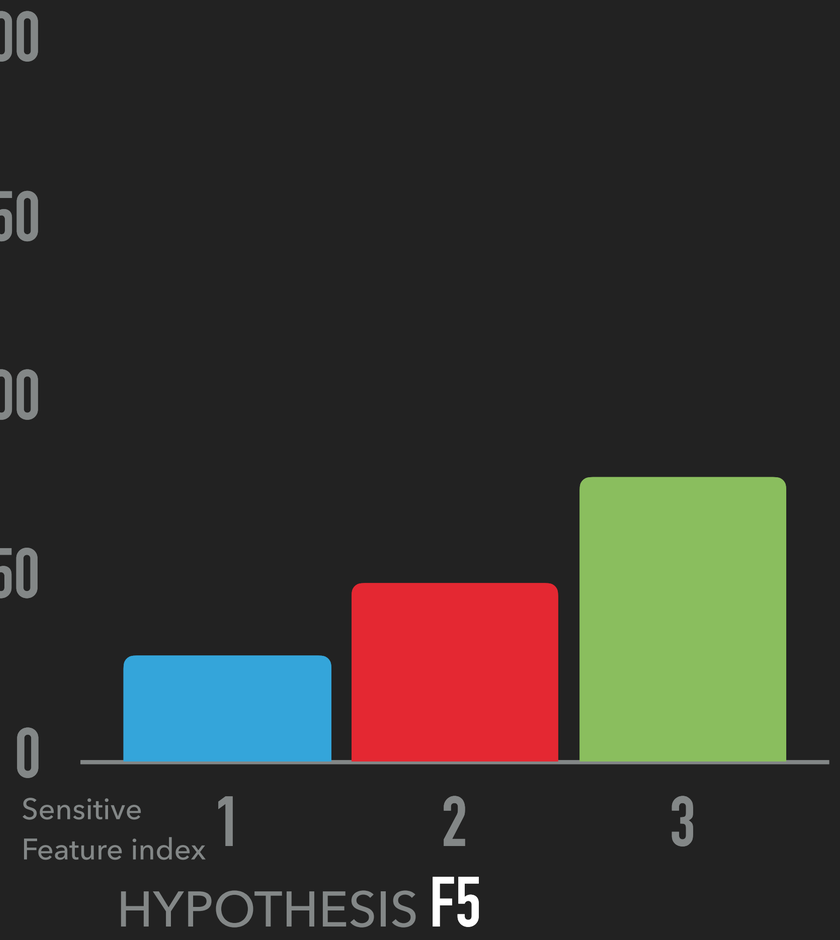
MINIMISING AGGREGATED EMPIRICAL RISK



MINIMISING AGGREGATED EMPIRICAL RISK

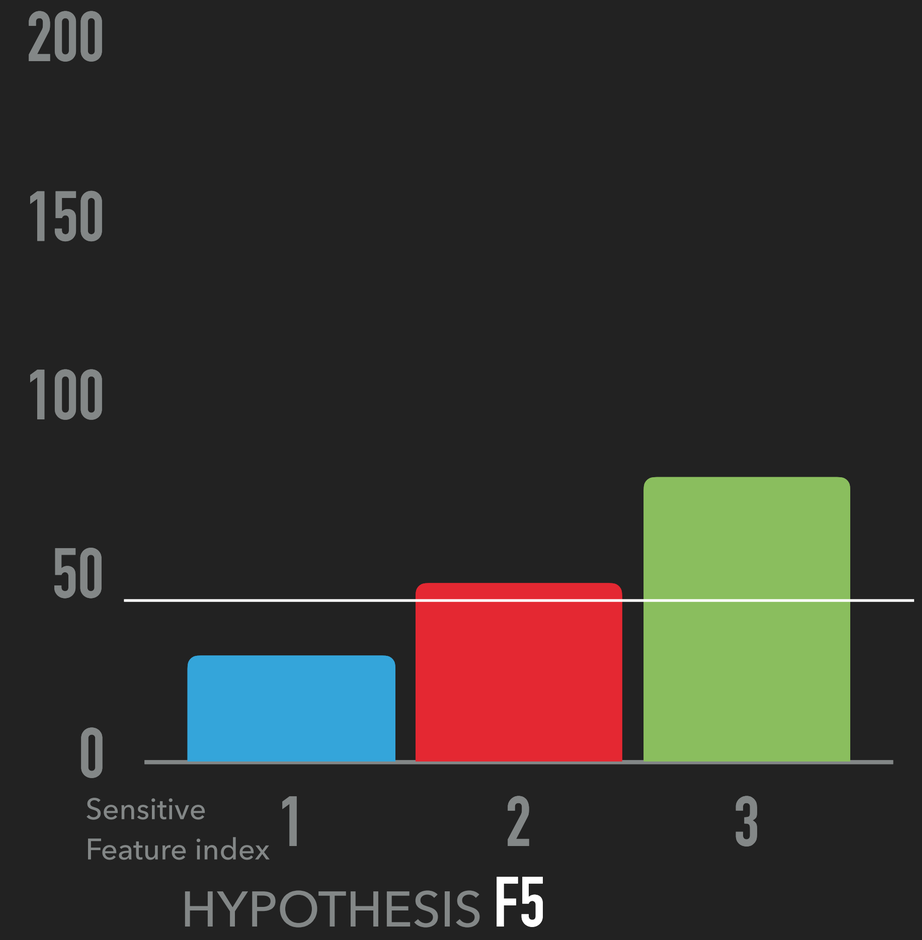


MINIMISING AGGREGATED EMPIRICAL RISK



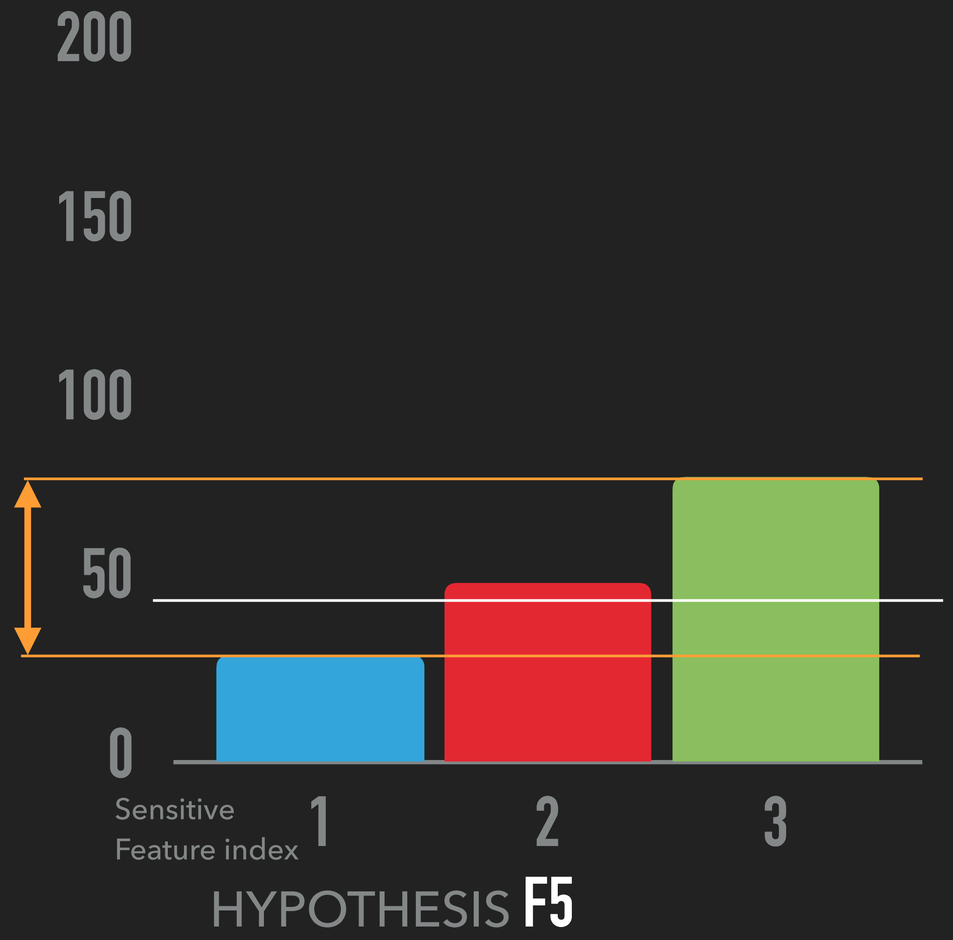
MINIMISING AGGREGATED EMPIRICAL RISK

- Standard problem: minimise average risk



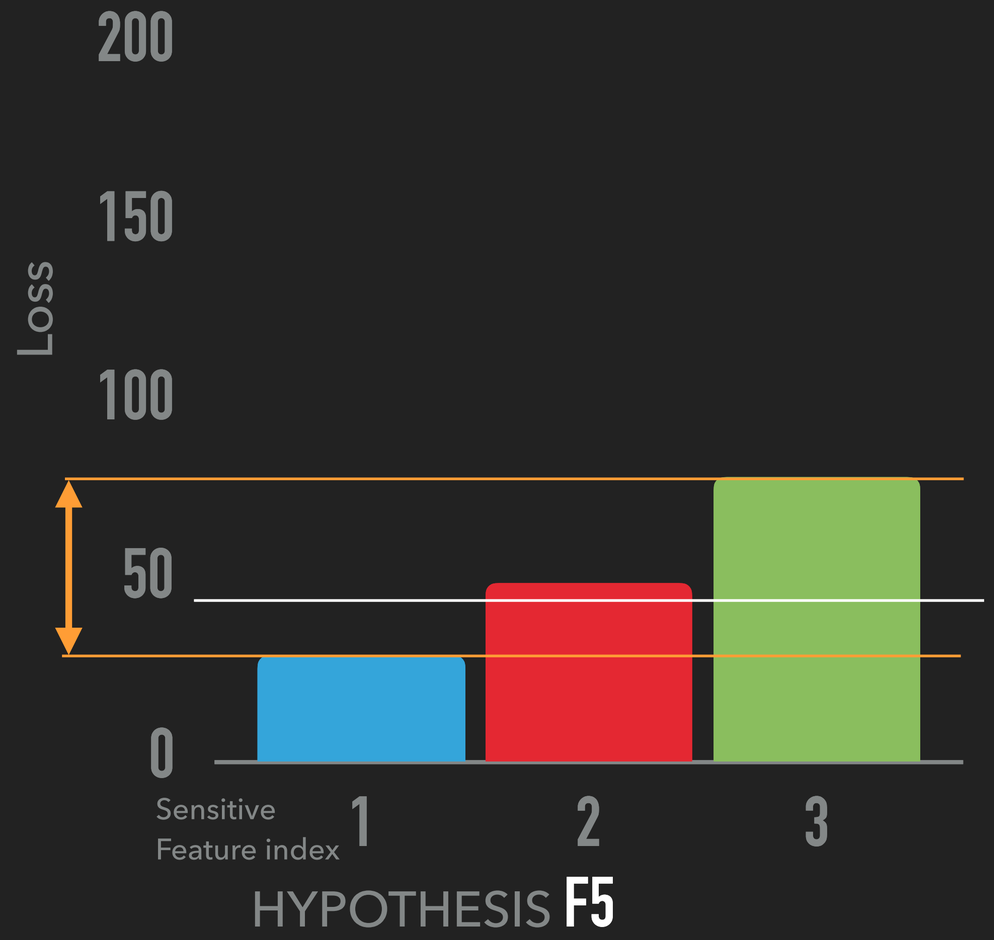
MINIMISING AGGREGATED EMPIRICAL RISK

- Standard problem: minimise average risk
- Equity problem: also take account of variation



MINIMISING AGGREGATED EMPIRICAL RISK

- Standard problem: minimise average risk
- Equity problem: also take account of variation
- Fairness problem: mixture of both



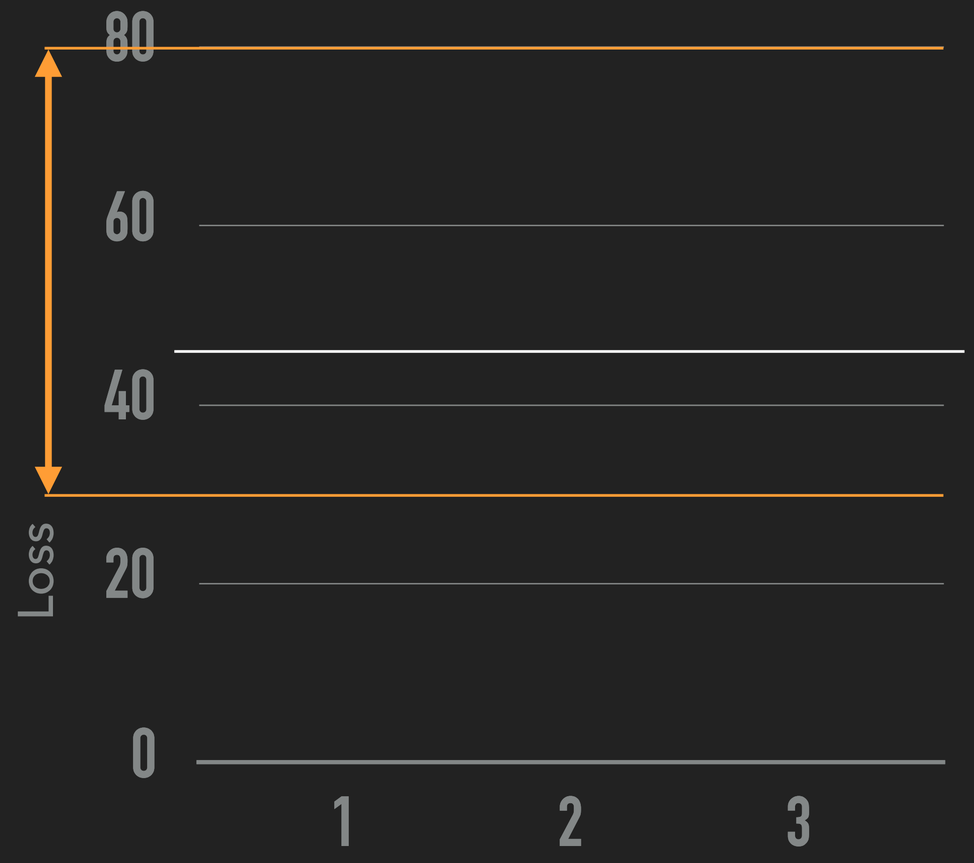
MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

- Trade off low deviation against higher average

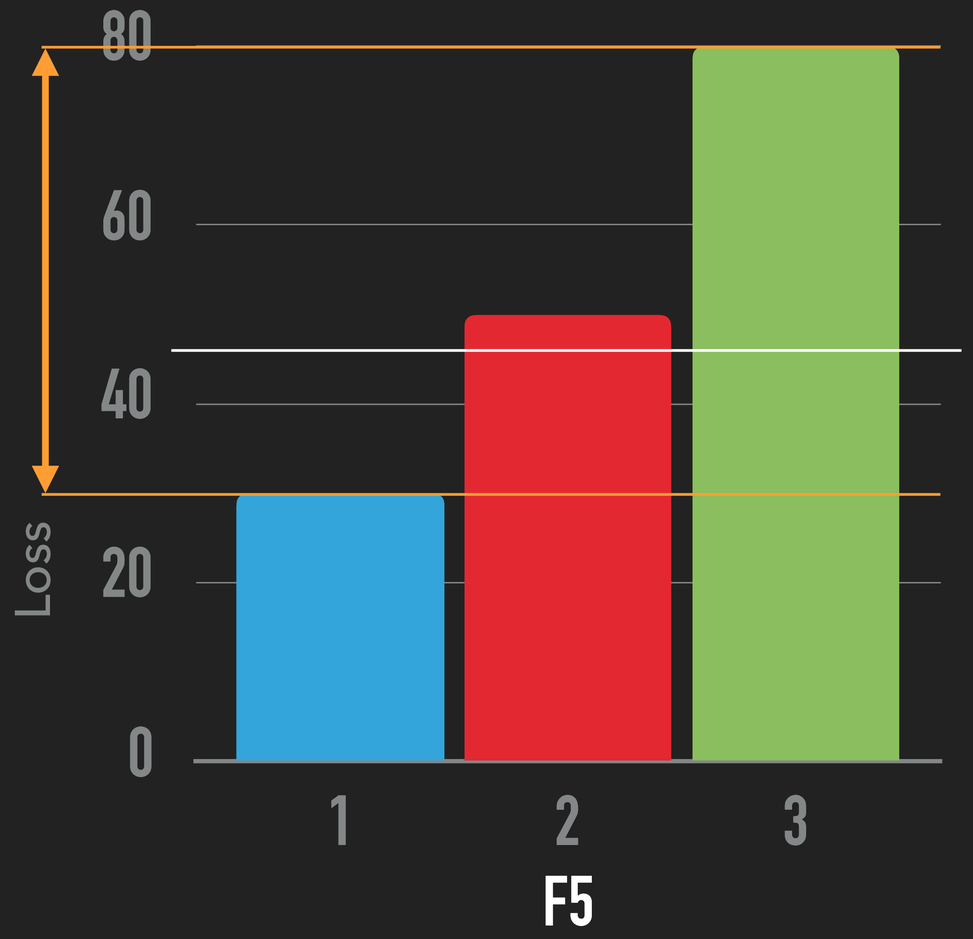
MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

- Trade off low deviation against higher average



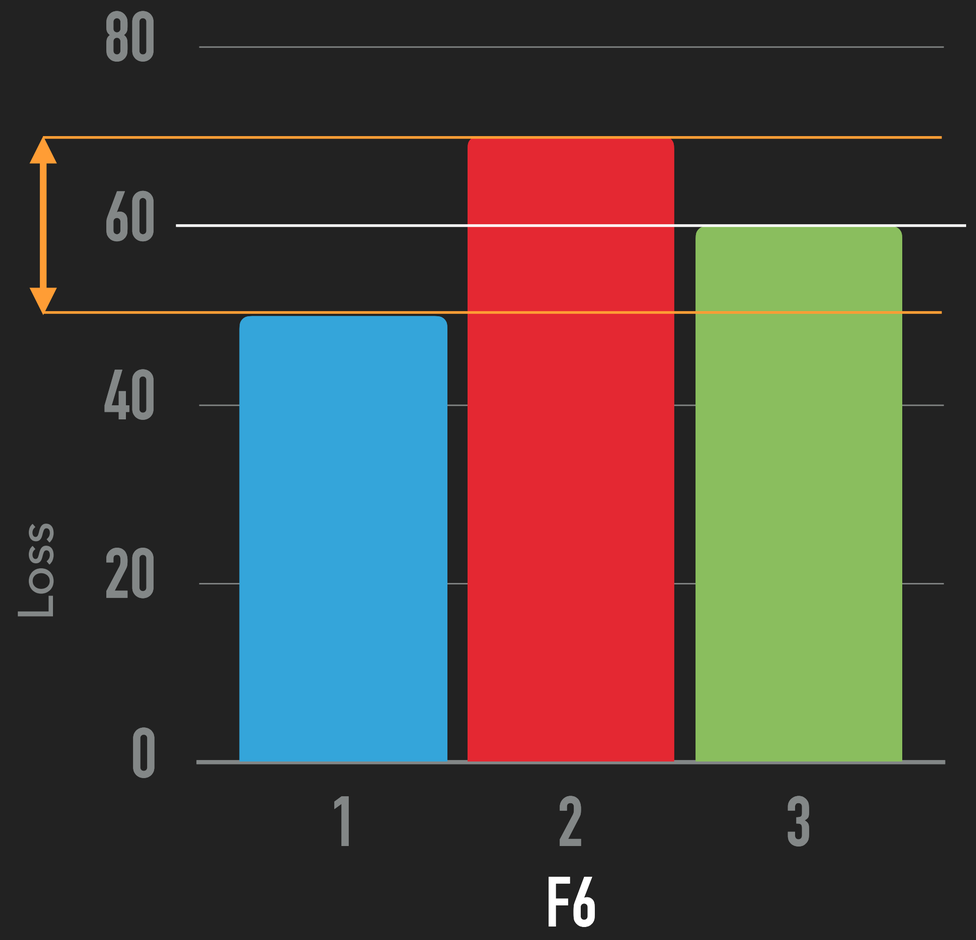
MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

- Trade off low deviation against higher average



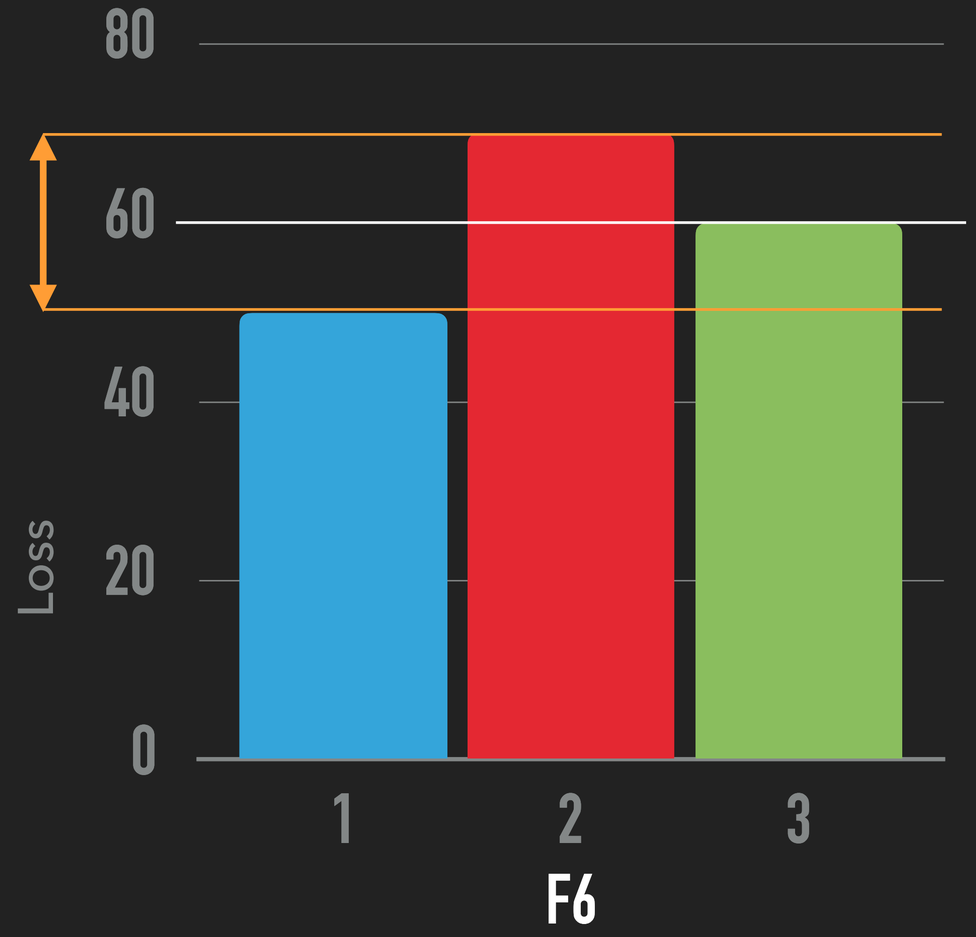
MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

- Trade off low deviation against higher average



MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

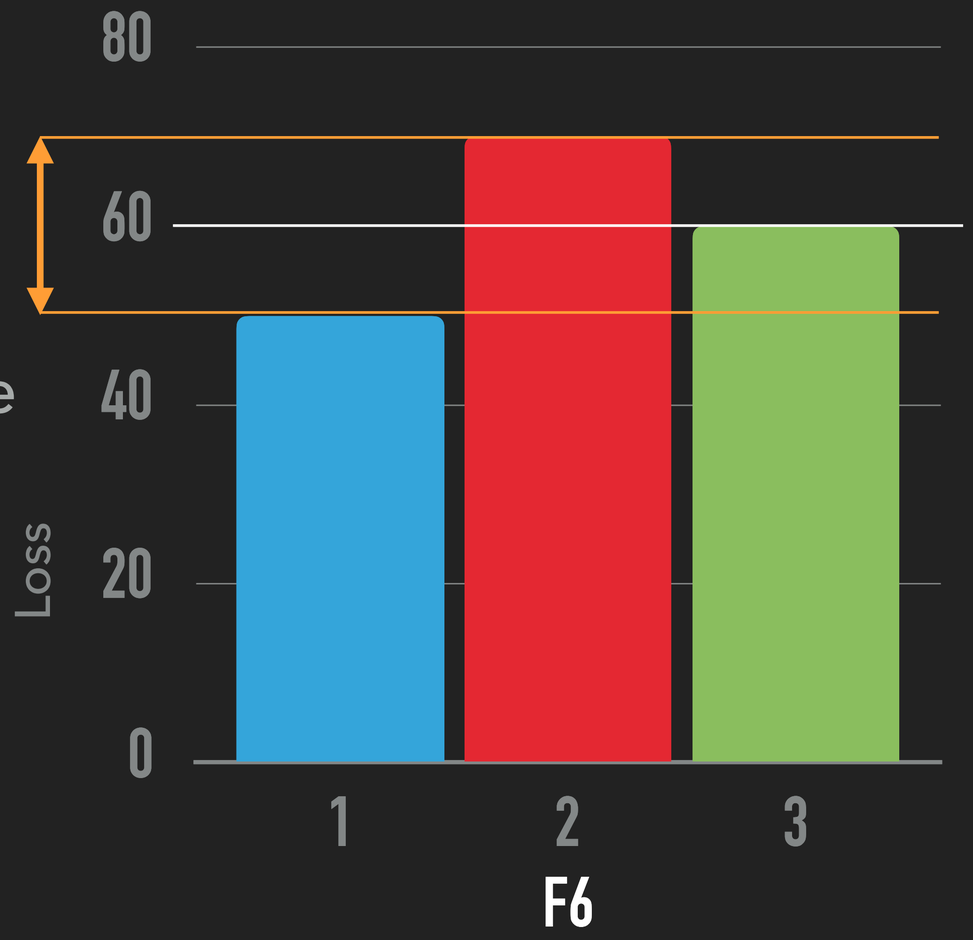
- Trade off low deviation against higher average
- Let $S = \{1,2,3\}$ be the sensitive feature space



MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

- Trade off low deviation against higher average
- Let $S = \{1,2,3\}$ be the sensitive feature space
- For $f \in \mathcal{F}$ let $\mathsf{R}_f \colon S \to \mathbb{R}$ be a r.v. (taking S as the sample space, with a uniform base measure)

$$\mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})}\left[\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s\right]$$



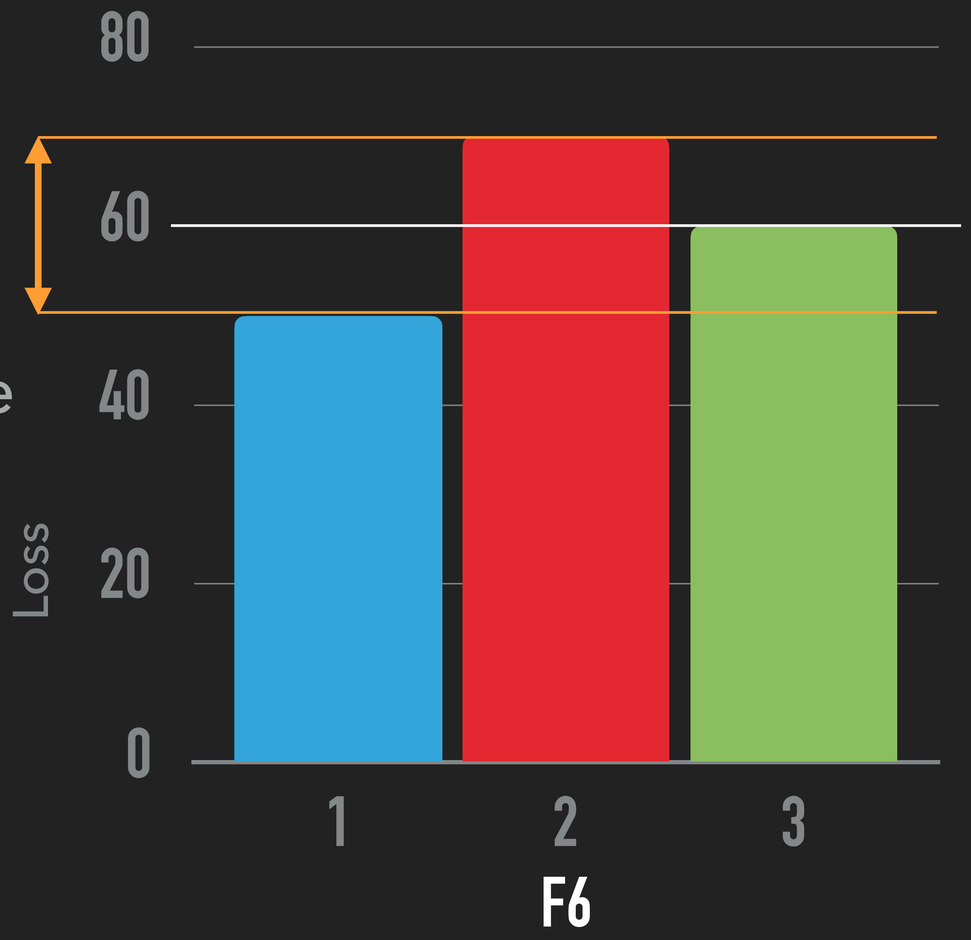
MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

- Trade off low deviation against higher average
- Let $S = \{1,2,3\}$ be the sensitive feature space
- For $f \in \mathcal{F}$ let $\mathsf{R}_f \colon S \to \mathbb{R}$ be a r.v. (taking S as the sample space, with a uniform base measure)

$$\mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})}\left[\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s\right]$$

- Standard ERM:

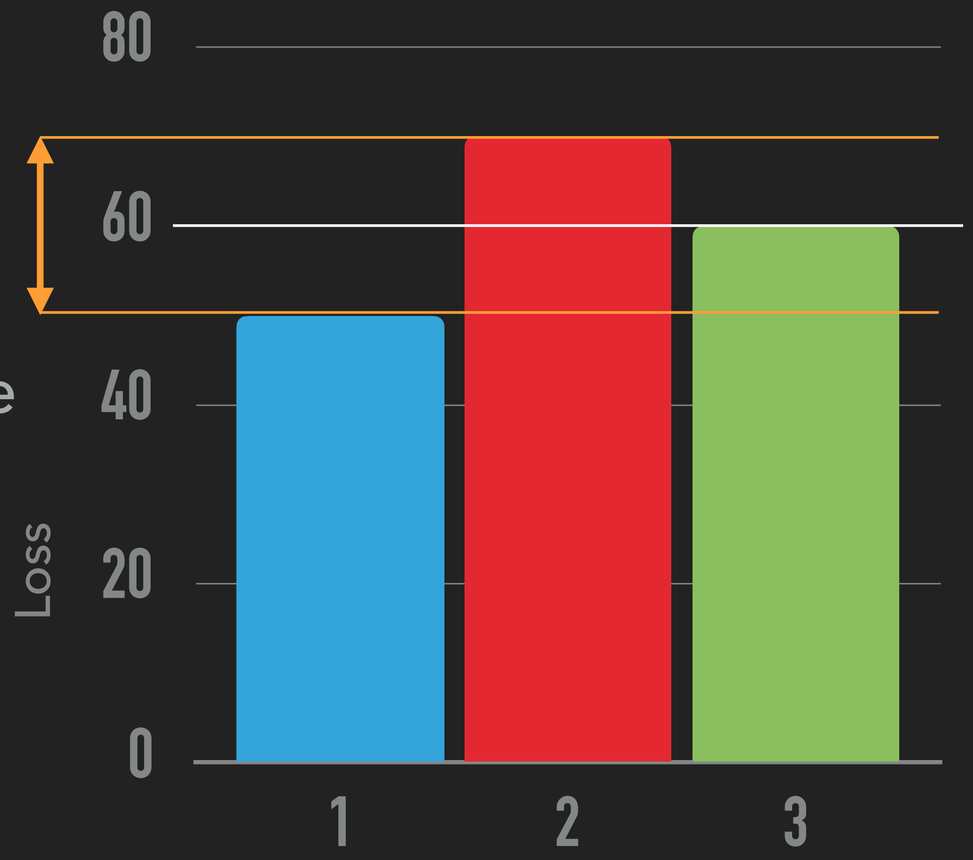
$$\min_{f \in \mathcal{F}} \mathbb{E}(\mathsf{R}_f)$$



MINIMISING AGGREGATED EMPIRICAL RISK AND DEVIATION

- Trade off low deviation against higher average
- Let $S = \{1,2,3\}$ be the sensitive feature space
- For $f \in \mathcal{F}$ let $\mathsf{R}_f \colon S \to \mathbb{R}$ be a r.v. (taking S as the sample space, with a uniform base measure)

$$\mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})}\left[\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s\right]$$



- Standard ERM: $\min_{f \in \mathcal{F}} \mathbb{E}(\mathsf{R}_f)$
- Fairness Augmented ERM: $\min_{f \in \mathcal{F}} \mathbb{E}(\mathsf{R}_f) + \mathscr{D}(\mathsf{R}_f) = \min_{f \in \mathcal{F}} \mathscr{R}(\mathsf{R}_f)$

FAIRNESS RISK MEASURES ARE “REGULAR MEASURES OF RISK”

FAIRNESS RISK MEASURES ARE "REGULAR MEASURES OF RISK"

- Instead can start with axioms for $\mathscr{R}$

Paper lists and justifies them

FAIRNESS RISK MEASURES ARE "REGULAR MEASURES OF RISK"

- Instead can start with axioms for $\mathscr{R}$ Paper lists and justifies them
- Then show that such fairness risk measures are "regular measures of risk"
  - (In fact they are "coherent measures of risk")

FAIRNESS RISK MEASURES ARE "REGULAR MEASURES OF RISK"

- Instead can start with axioms for $\mathscr{R}$ Paper lists and justifies them
- Then show that such fairness risk measures are "regular measures of risk"
  - (In fact they are "coherent measures of risk")
- Such measures can always be written as

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

Fairness risk measure

Deviation measure

FAIRNESS RISK MEASURES ARE "REGULAR MEASURES OF RISK"

- Instead can start with axioms for $\mathscr{R}$ Paper lists and justifies them
- Then show that such fairness risk measures are "regular measures of risk"
  - (In fact they are "coherent measures of risk")
- Such measures can always be written as

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

Fairness risk measure

Deviation measure

- Here $\mathscr{D}$ is a "regular measure of deviation"
  (i.e. convex, positively homogeneous, zero only when R is constant, and lower semicontinuous)

EXAMPLE RISK MEASURE AND CORRESPONDING DEVIATION MEASURE

EXAMPLE RISK MEASURE AND CORRESPONDING DEVIATION MEASURE

$$\mathscr{R}_{\mathrm{Q},\alpha}(Z) = \mathsf{CVaR}_{\alpha}(Z) \qquad \mathscr{D}_{\mathrm{Q},\alpha}(Z) = \mathsf{CVaR}_{\alpha}(Z - \mathbb{E}(Z))$$

EXAMPLE RISK MEASURE AND CORRESPONDING DEVIATION MEASURE

$$\mathscr{R}_{\mathrm{Q},\alpha}(\mathsf{Z}) = \mathsf{CVaR}_{\alpha}(\mathsf{Z}) \qquad \mathscr{D}_{\mathrm{Q},\alpha}(\mathsf{Z}) = \mathsf{CVaR}_{\alpha}(\mathsf{Z} - \mathbb{E}(\mathsf{Z}))$$

- CVaR is the "Conditional Value at Risk".
- When Z is continuous random variable:

$$\mathsf{CVaR}_{\alpha}(\mathsf{Z}) = \mathbb{E}(\mathsf{Z} \,|\, \mathsf{Z} \geq q_{\alpha}(\mathsf{Z}))$$

  - where $q_{\alpha}(\mathsf{Z})$ is the αth quantile of Z

EXAMPLE RISK MEASURE AND CORRESPONDING DEVIATION MEASURE

$$\mathscr{R}_{\mathrm{Q},\alpha}(\mathrm{Z}) = \mathrm{CVaR}_{\alpha}(\mathrm{Z}) \qquad \mathscr{D}_{\mathrm{Q},\alpha}(\mathrm{Z}) = \mathrm{CVaR}_{\alpha}(\mathrm{Z} - \mathbb{E}(\mathrm{Z}))$$

- CVaR is the "Conditional Value at Risk".
- When Z is continuous random variable:

$$\mathrm{CVaR}_{\alpha}(\mathrm{Z}) = \mathbb{E}(\mathrm{Z} \mid \mathrm{Z} \geq q_{\alpha}(\mathrm{Z}))$$

  - where $q_{\alpha}(\mathrm{Z})$ is the αth quantile of Z
- Have $\mathrm{CVaR}_{0}(\mathrm{Z}) = \mathbb{E}(\mathrm{Z})$ and $\mathrm{CVaR}_{1}(\mathrm{Z}) = \max(\mathrm{Z})$

EXAMPLE RISK MEASURE AND CORRESPONDING DEVIATION MEASURE

$$\mathscr{R}_{\mathrm{Q},\alpha}(\mathsf{Z}) = \mathsf{CVaR}_{\alpha}(\mathsf{Z}) \qquad \mathscr{D}_{\mathrm{Q},\alpha}(\mathsf{Z}) = \mathsf{CVaR}_{\alpha}(\mathsf{Z} - \mathbb{E}(\mathsf{Z}))$$

- CVaR is the "Conditional Value at Risk".
- When Z is continuous random variable:

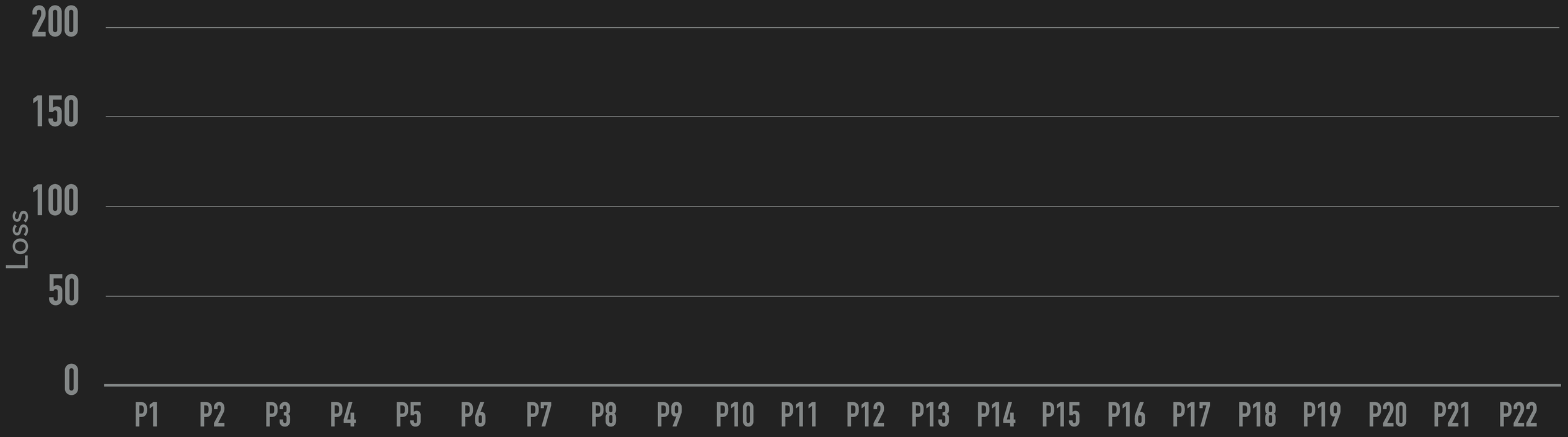
$$\mathsf{CVaR}_{\alpha}(\mathsf{Z}) = \mathbb{E}(\mathsf{Z} \,|\, \mathsf{Z} \geq q_{\alpha}(\mathsf{Z}))$$

  - where $q_{\alpha}(\mathsf{Z})$ is the αth quantile of Z
- Have $\mathsf{CVaR}_0(\mathsf{Z}) = \mathbb{E}(\mathsf{Z})$ and $\mathsf{CVaR}_1(\mathsf{Z}) = \max(\mathsf{Z})$
- Fairness objective becomes (see paper, eq (26)):

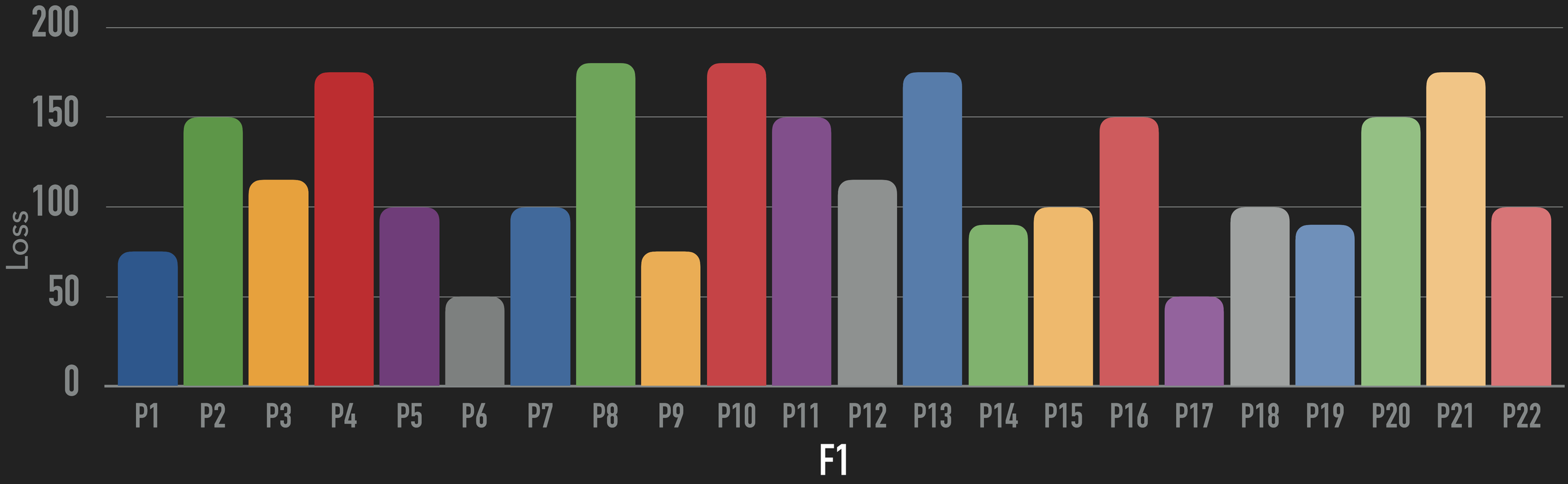
$$\min_{f \in \mathscr{F}, \rho \in \mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha} \cdot \mathbb{E}[\mathsf{L}(f) - \rho]_+ \right\}.$$

AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!

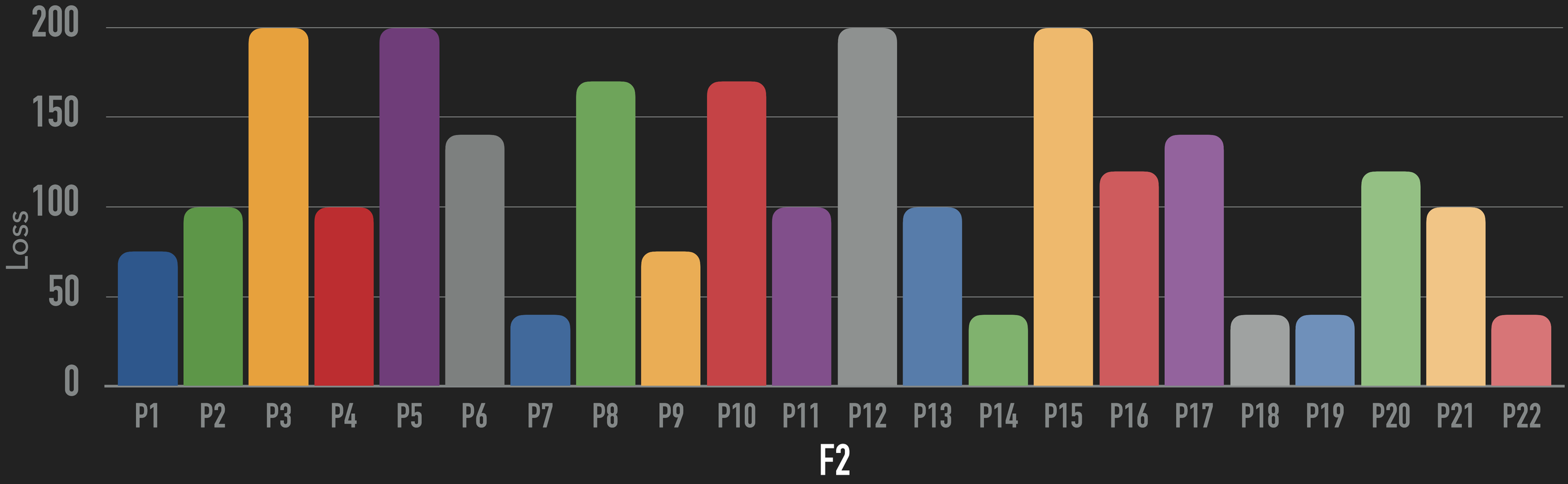
AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



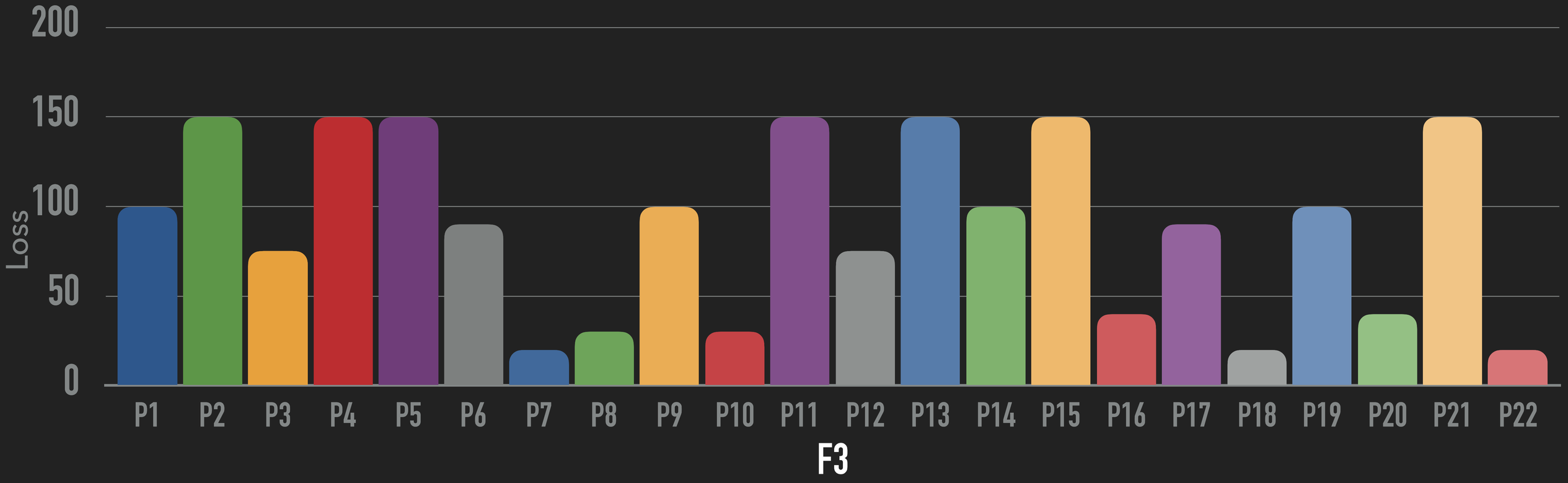
AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



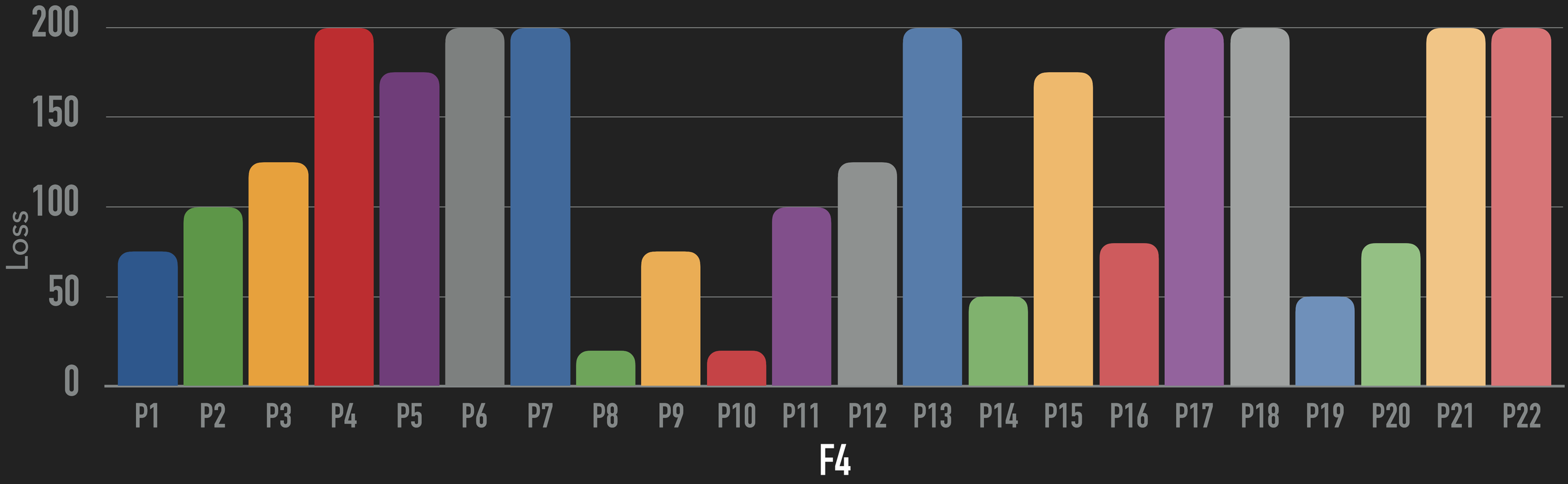
AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



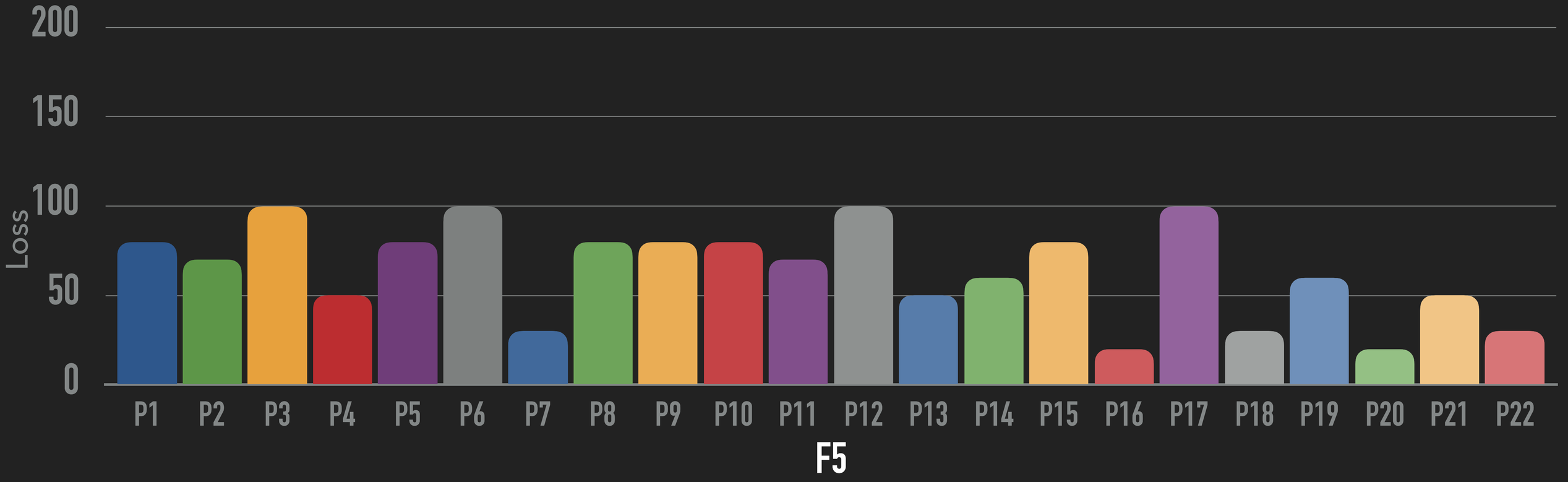
AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



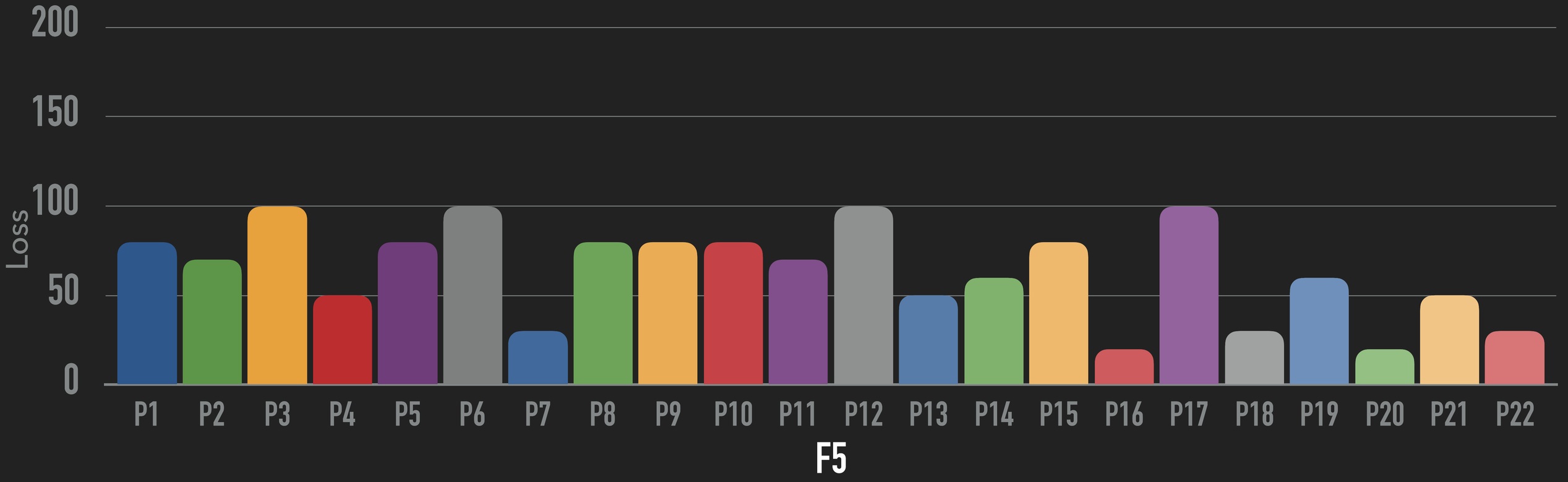
AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!

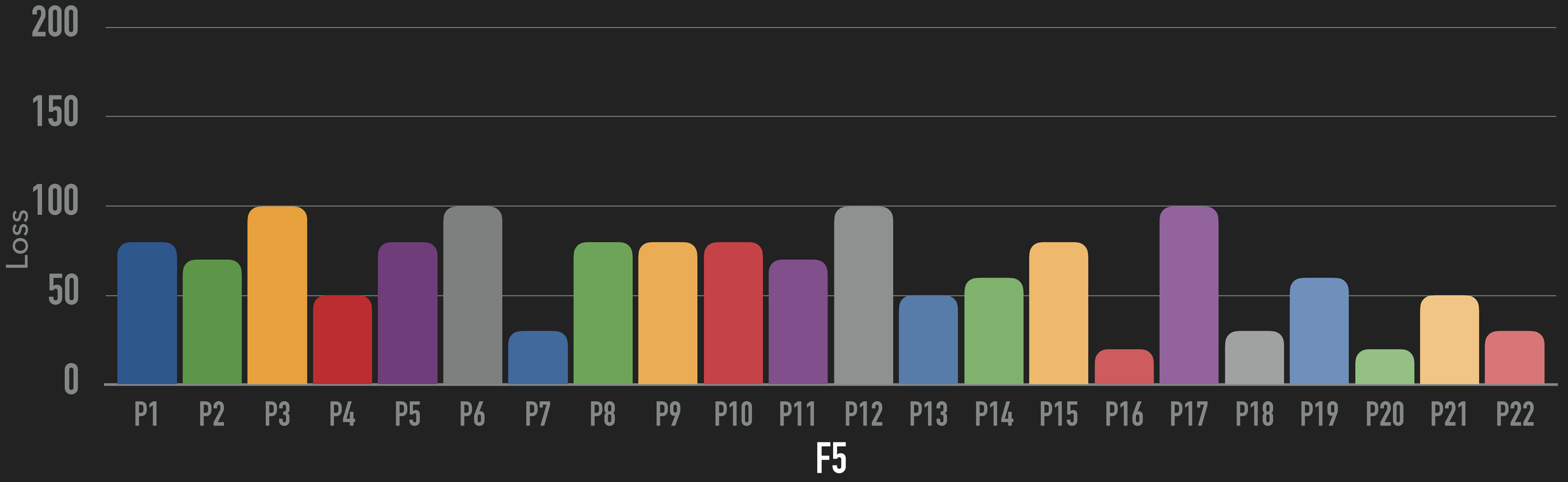


AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



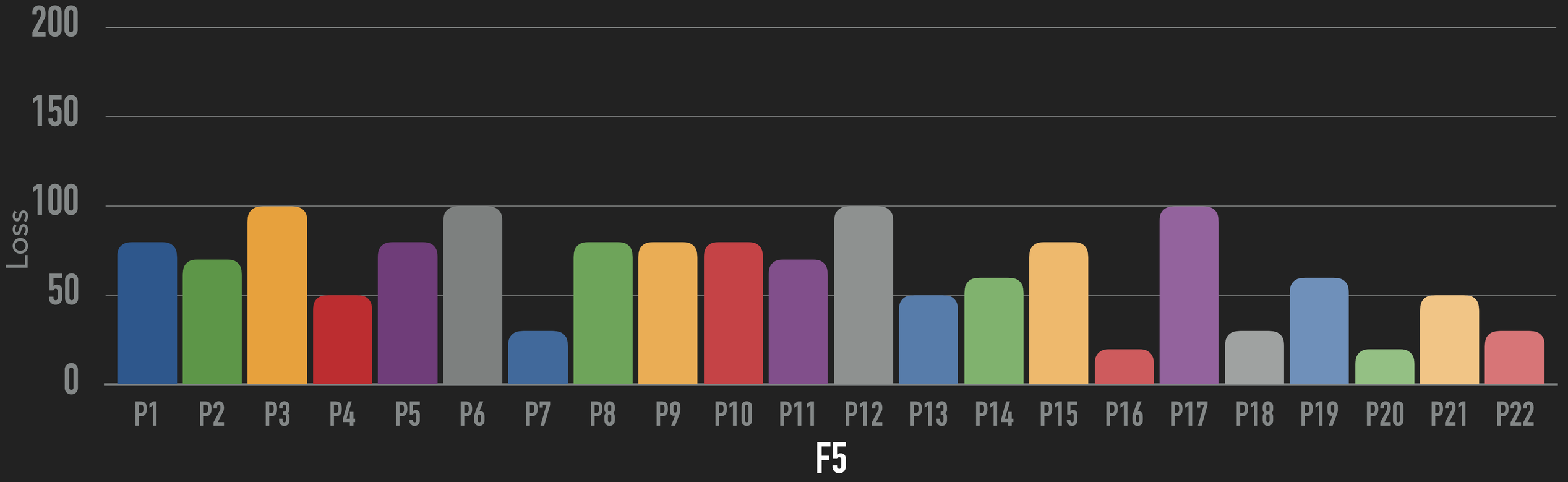
- Consistent with the principle that fundamental moral unit is the individual person

AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



- Consistent with the principle that fundamental moral unit is the individual person
- Avoids headaches with group boundaries and multiple group membership

AN INTERESTING LIMITING CASE – EACH PERSON IS THEIR OWN CATEGORY!



- Consistent with the principle that fundamental moral unit is the individual person
- Avoids headaches with group boundaries and multiple group membership
- Fairness risk measures automatically extend to this case (trivially)

CONCLUSION

CONCLUSION

$$\min_{f \in \mathcal{F}} \mathcal{R}(\mathsf{R}_f) \qquad \mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})} [\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s]$$

- New and general approach to fairness in ML problems

CONCLUSION

$$\min_{f \in \mathcal{F}} \mathcal{R}(\mathsf{R}_f) \qquad \mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})} [\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s]$$

- New and general approach to fairness in ML problems
- Fairness only depends upon **losses**, not **predictions**

CONCLUSION

$$\min_{f \in \mathcal{F}} \mathcal{R}(\mathsf{R}_f) \qquad \mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})} \left[\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s \right]$$

- New and general approach to fairness in ML problems
- Fairness only depends upon **losses**, not **predictions**
- Fairness risk measures are **symmetric coherent measures of risk**

CONCLUSION

$$\min_{f \in \mathcal{F}} \mathcal{R}(\mathsf{R}_f) \qquad \mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})} \left[\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s \right]$$

- New and general approach to fairness in ML problems
- Fairness only depends upon **losses**, not **predictions**
- Fairness risk measures are **symmetric coherent measures of risk**
- Close connection to **measures of inequality** (see appendix)

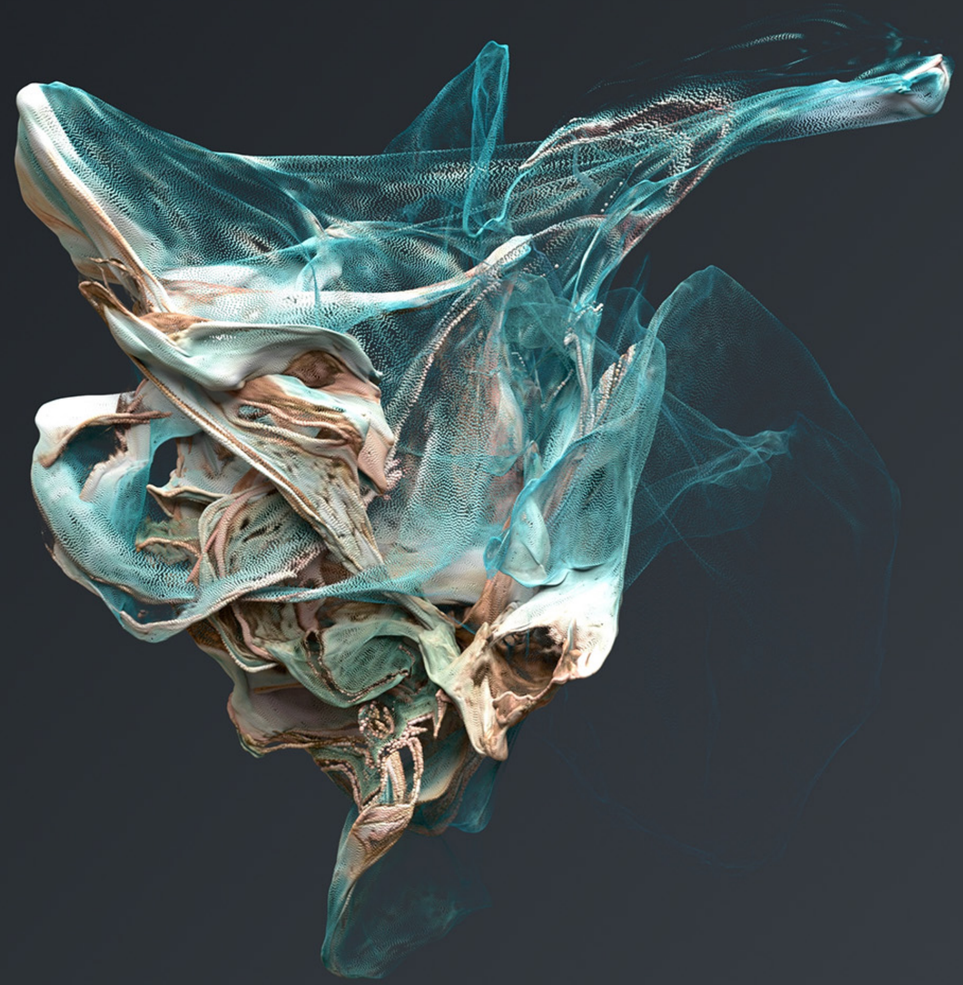
CONCLUSION

$$\min_{f \in \mathcal{F}} \mathcal{R}(\mathsf{R}_f) \qquad \mathsf{R}_f \colon S \ni s \mapsto \mathbb{E}_{(\mathsf{X},\mathsf{Y})} [\ell(\mathsf{Y}, f(\mathsf{X})) \,|\, \mathsf{S} = s]$$

- New and general approach to fairness in ML problems
- Fairness only depends upon **losses**, not **predictions**
- Fairness risk measures are **symmetric coherent measures of risk**
- Close connection to **measures of inequality** (see appendix)
- Computationally **tractable**; related to **SVM**! (see paper / poster for **experiments**)

Humanising Machine Intelligence

Machine Learning Postdoc position available



hmi.anu.edu.au