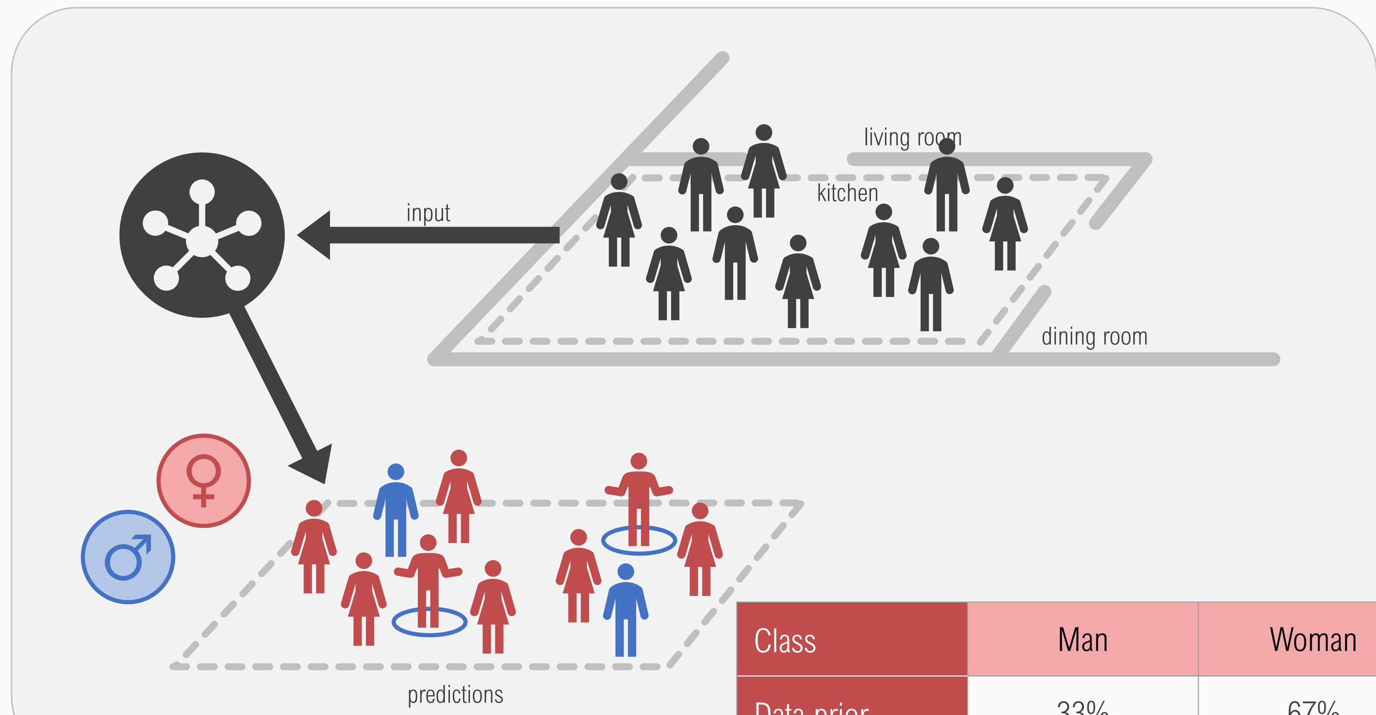
Feature-wise Bias Amplification

Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, Anupam Datta

Carnegie Mellon University
School of Computer Science

What is Bias Amplification?

A model exhibits *bias amplification* if the prior distribution of the model's predictions does not match that of the data.



Class	Man	Woman
Data prior	33%	67%
Pred. prior	16%	84%

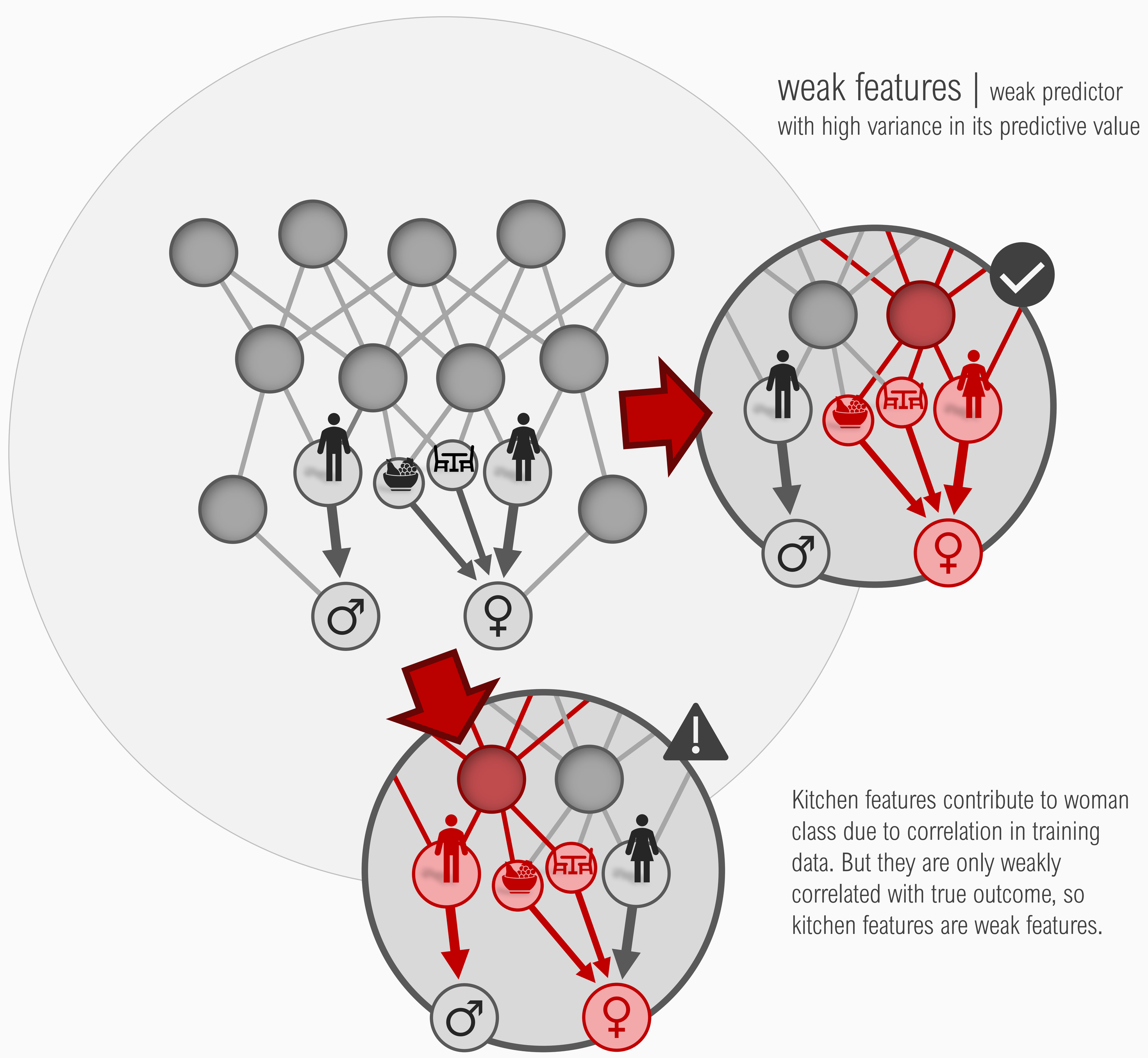
[1] Zhao et al. "Men also Like Shopping: Reducing Gender Bias Amplification Using Corpus-level Constraints"

Bias Amplification | Let $\mathcal{D}$ be a distribution over features, x, and labels, y. Let h_S be a binary classifier trained on $S \sim \mathcal{D}^n$. The *bias amplification* of h_S on $\mathcal{D}$ is

$$B_{\mathcal{D}}(h_S) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h_S(x) - y]$$

Hypothesis: Overreliance on Weak Features

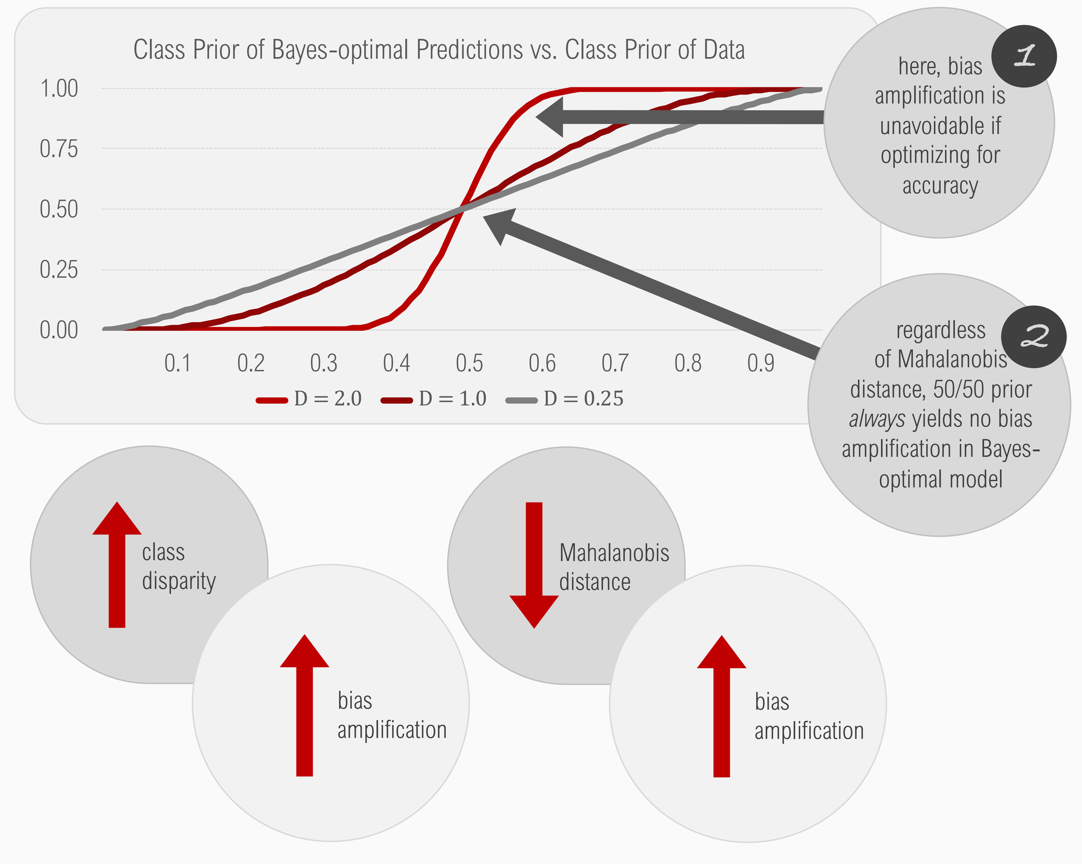
weak features | weak predictor with high variance in its predictive value

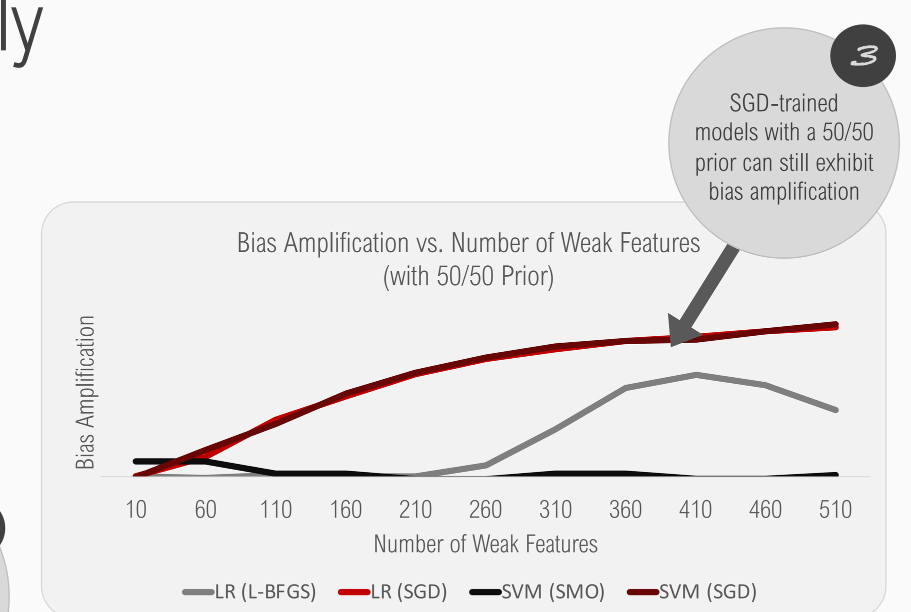


Kitchen features contribute to woman class due to correlation in training data. But they are only weakly correlated with true outcome, so kitchen features are weak features.

SGD Amplifies Bias — Unnecessarily

In the setting of Gaussian naïve-Bayes data, the bias of the Bayes-optimal classifier is a function of the *Mahalanobis distance* between the classes and the class prior of the data.





Implication | this bias amplification exhibited by SGD is *preventable* without decreasing accuracy

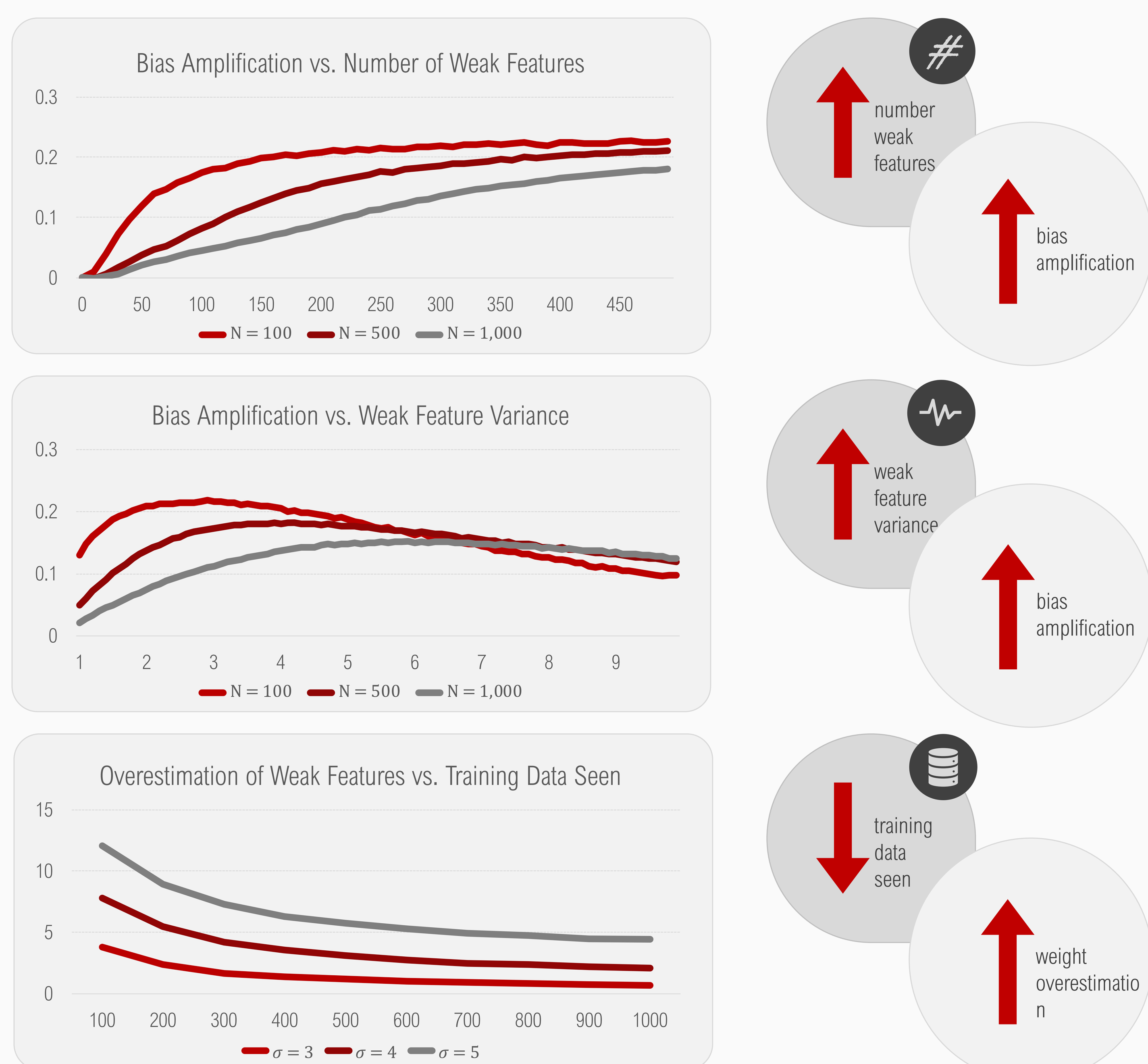
Feature-Wise Bias Amplification

manifestation | a model trained with SGD will *overestimate* the weak features for the task, and thus over-predict the class with more weak features

overestimation | putting undue weight (in linear models) or influence (in deep models) on a feature

three main factors:

- number of weak features
- variance of weak features
- amount of training data



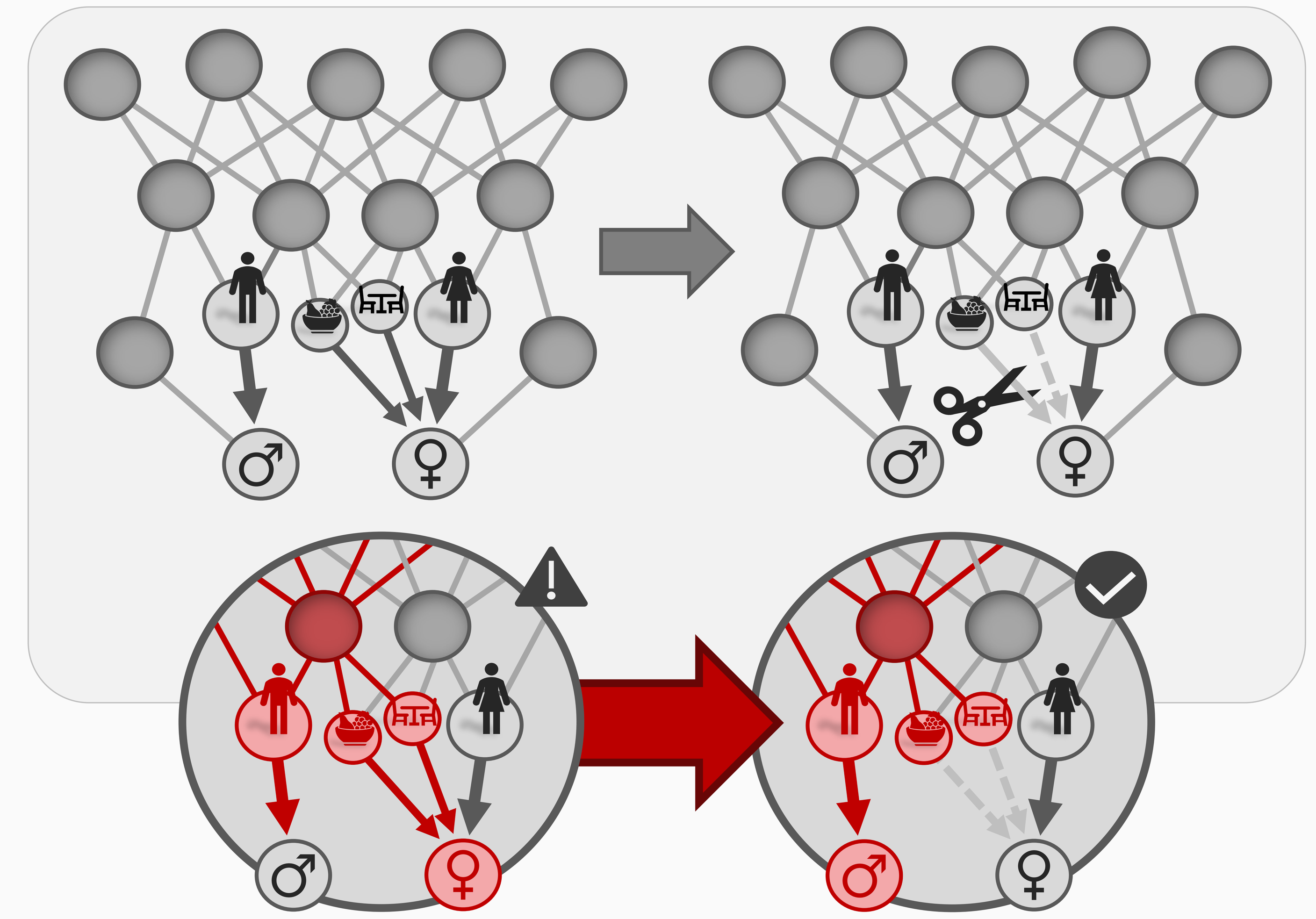
intuition | sum of the weights from many overestimated weak features overpowers those of strong features

intuition | a feature with higher variance is a weaker predictor—the weaker the predictor, the more the model overestimates but the less it uses the feature overall

intuition | more data brings model closer to convergence, at which point the model learns appropriate weights

How do we Fix Bias Amplification?

basic idea | fix bias amplification via targeted feature selection



Influence | Let $s = \langle g, h\rangle$ be a *slice* of deep network, f, such that $f = g \circ h$, and let P be a distribution over internal points, $z = h(x)$. Then the *internal influence* of feature z_j on class, y, is

$$\chi_j^y(g \circ h, P) = \int_{z \in h(\mathcal{X})} \left[\frac{\partial g_y}{\partial z_j}\right]_z P(z)dz$$

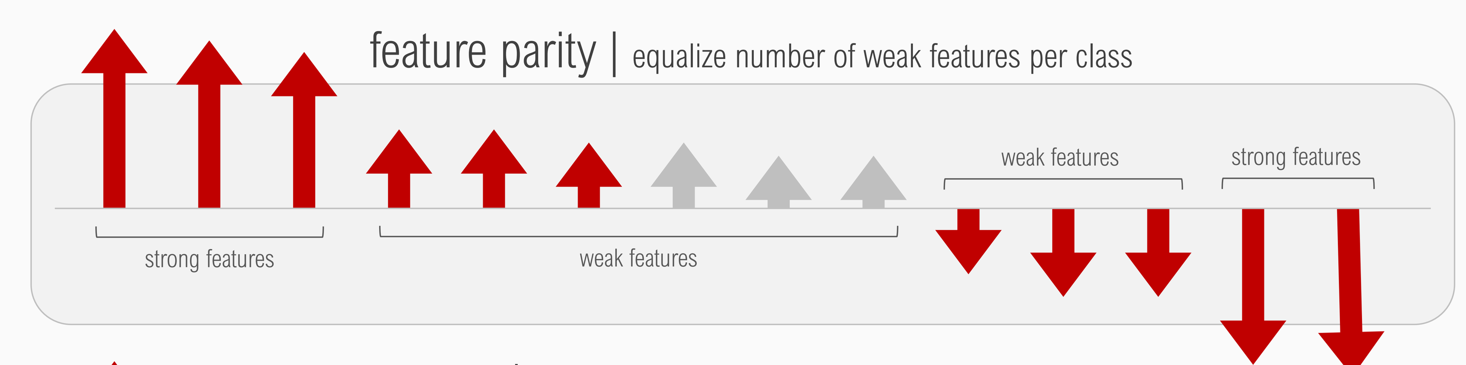
[2] Leino et al. "Influence-directed Explanations for Convolutional Neural Networks"

Experts | Let F_α be the set of the α most influential neurons towards class 1, let F_β be the set of the β most influential neurons towards class 0, and let $\mathcal{L}_S$ be the 0-1 loss on training set S. Then the *expert binary classifier* is $g_{\beta^*}^{\alpha^*}$, where

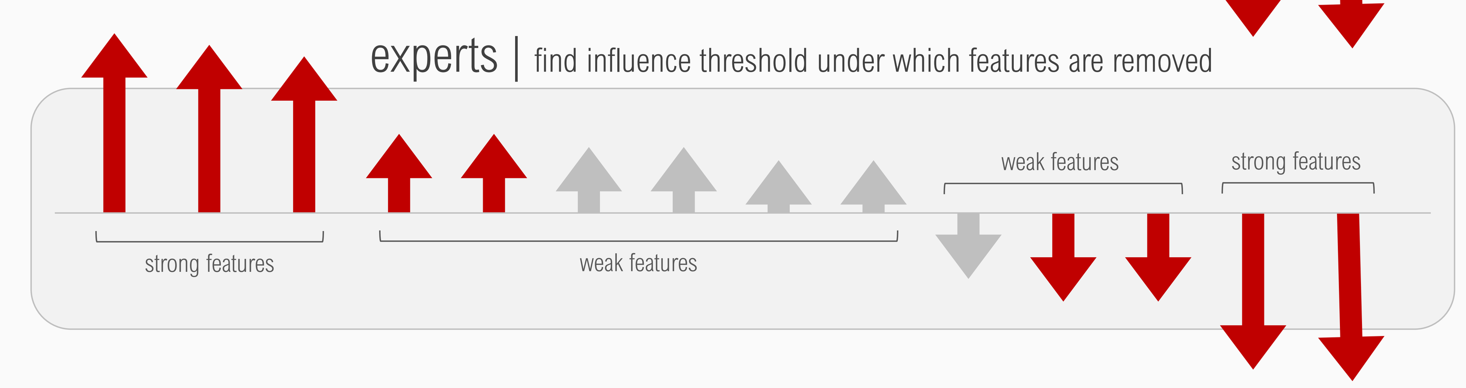
$$m_{\beta\, j}^{\alpha} = \mathbb{I}(j \in F_\alpha \cup F_\beta) \qquad g_\beta^\alpha(z) = g(m_\beta^\alpha z)$$

$$\alpha^*, \beta^* = \operatorname*{argmin}_{\alpha,\beta} |B_{\mathcal{D}}(g_\beta^\alpha)| \text{ subject to } \mathcal{L}_S(g_\beta^\alpha) \le \mathcal{L}(g)$$

feature parity | equalize number of weak features per class



experts | find influence threshold under which features are removed



You Can Have Your High Accuracy and Low Bias, Too