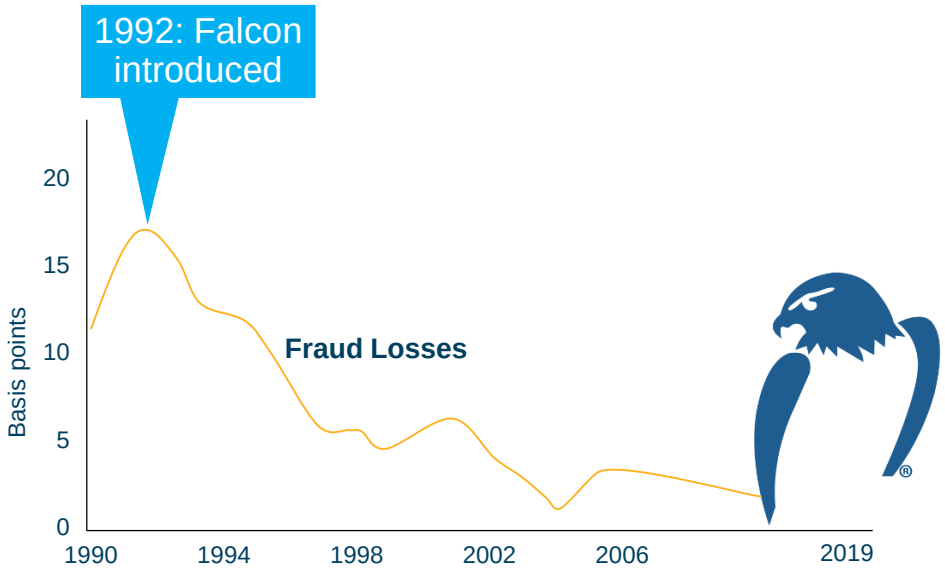
Ethical AI - Innovations in Model Confidence and Bias Detection

Scott Zoldi
Chief Analytics Officer, FICO

Shafi Rahman
Sr. Principal Scientist, FICO

FICO has 30+ years of Machine Learning Experience

1992: First ML based patent issued for FICO® Falcon® Fraud Manager



3000 ML research years

197 issued patents

102 pending patent applications

125+ patents in ML to date

Recidivism Prediction

COMPAS model's prediction of re-offenders is wrong for
45% of African Americans
as opposed to
24% of Caucasians

Merriam Webster

ethic noun

eth·ic | \ ˈe-thik \

Definition of *ethic*

1 **ethics** *plural in form but singular or plural in construction* **:** the discipline dealing with what is good and bad and with moral duty and obligation

2 **a** **:** a set of moral principles **:** a theory or system of moral values

Laws Exist in Most Countries That Mandate Bias Removal

UK Equality Act 2010 outlaws discrimination

UK Employment Rights Act 1996

EU Equal Treatment Directive, 2006

Bias in Machine Learning Have Practical Implications

Coming Back To Use of Algorithm in Criminal Justice – (COMPAS)

Model predicts likelihood of reoffending to decide whether to grant parole

False positive rate	
African American	45%
White	24%

Why is the model so much more wrong about the African American population?

Risk scores of 7,000 people arrested in 2013 and 2014 in US and whether they were charged with new crimes over the next two years

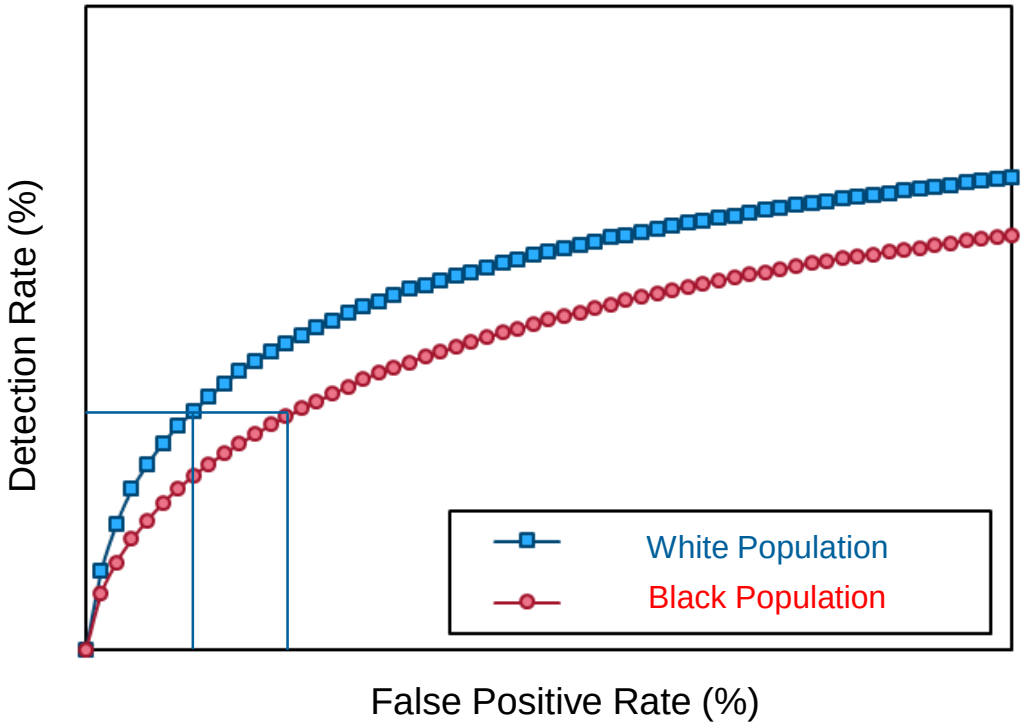
Ref: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Simply Excluding A Feature Doesn't Work

- Raised by one parent / adopted / foster care
- Unemployed / Minimum Wage Job
- High Crime Neighborhood
- …

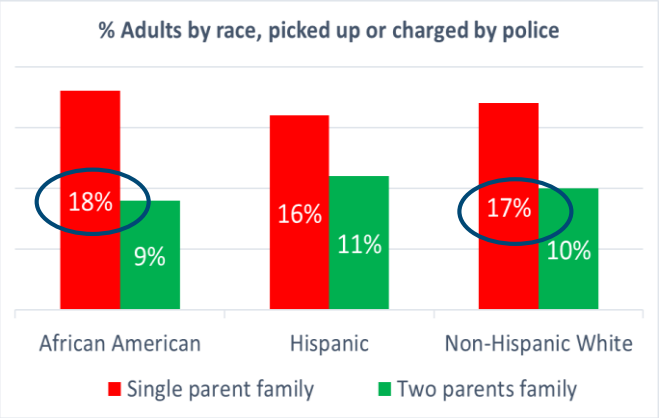
Ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5777393/

If Anything We Need To Be Conscious of The Protected Characteristics



Performance Analysis With Respect to Protected Characteristic Can Reveal Bias

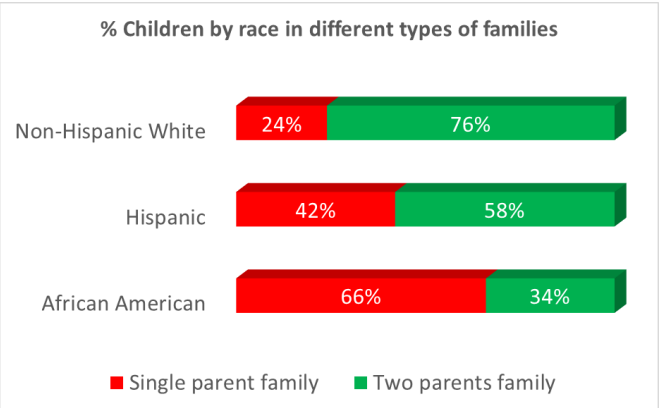
Predictor Variable - Raised by a single parent



2x likelihood of being picked up, if raised in single parent family, *irrespective of race*

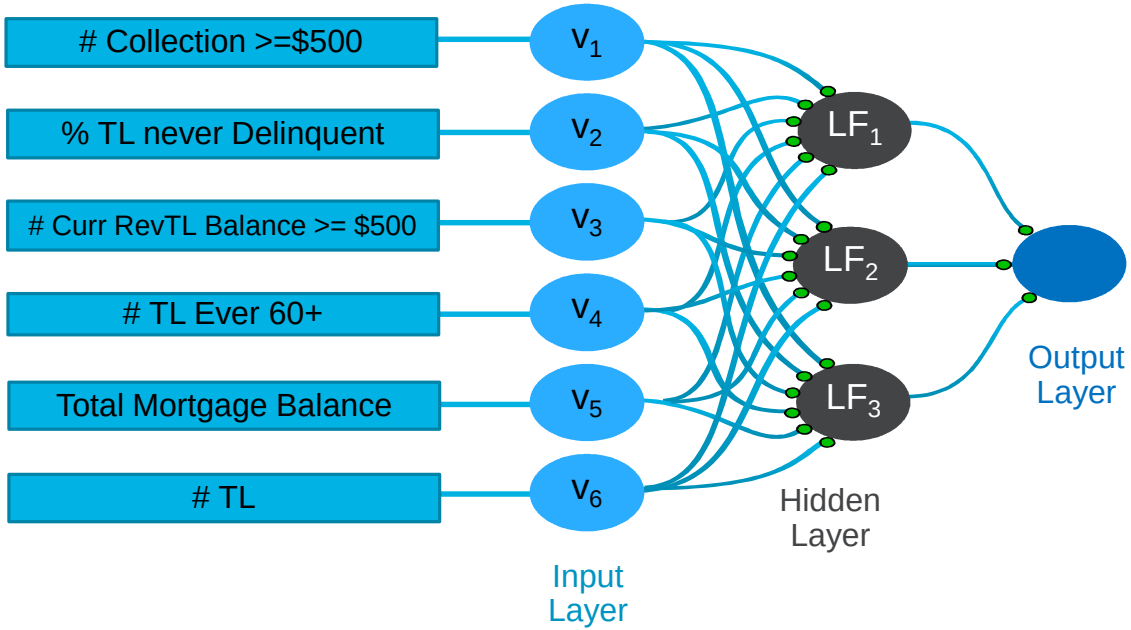
12.1% of blacks and **60.7%** non-Hispanic whites in general population

What happens as a group?



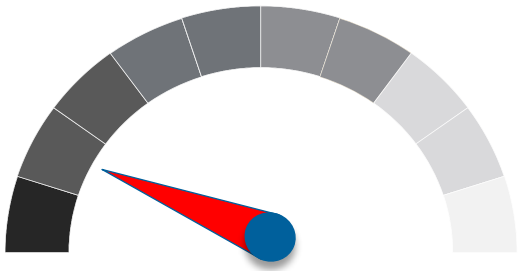
Black population **3X** more likely to be targeted by this variable

Often Bias Is Implicit And Hidden As a Latent Feature



Cost Function

$$C(w) = \frac{1}{2p} \sum_x \|y(x) - a(x)\|^2$$

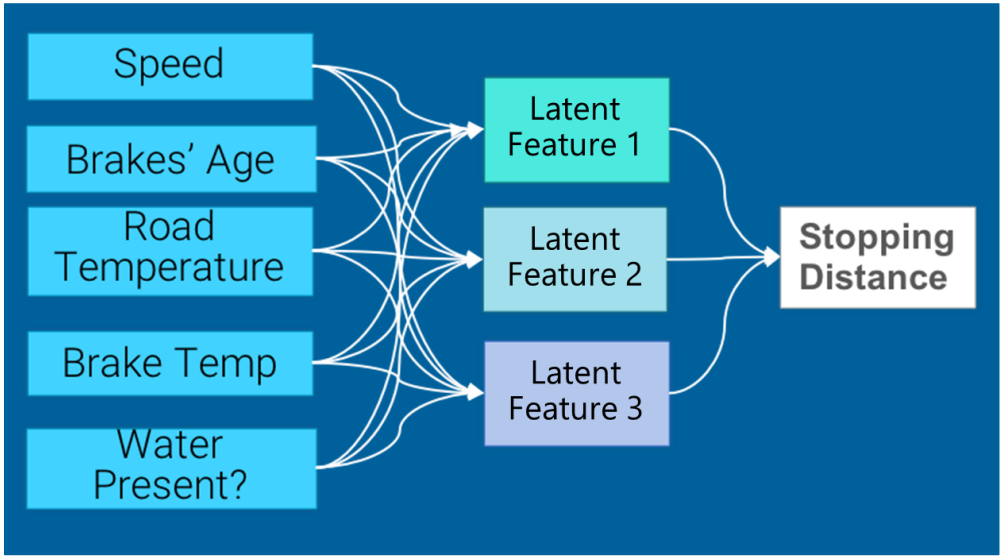


Interpretability Meter

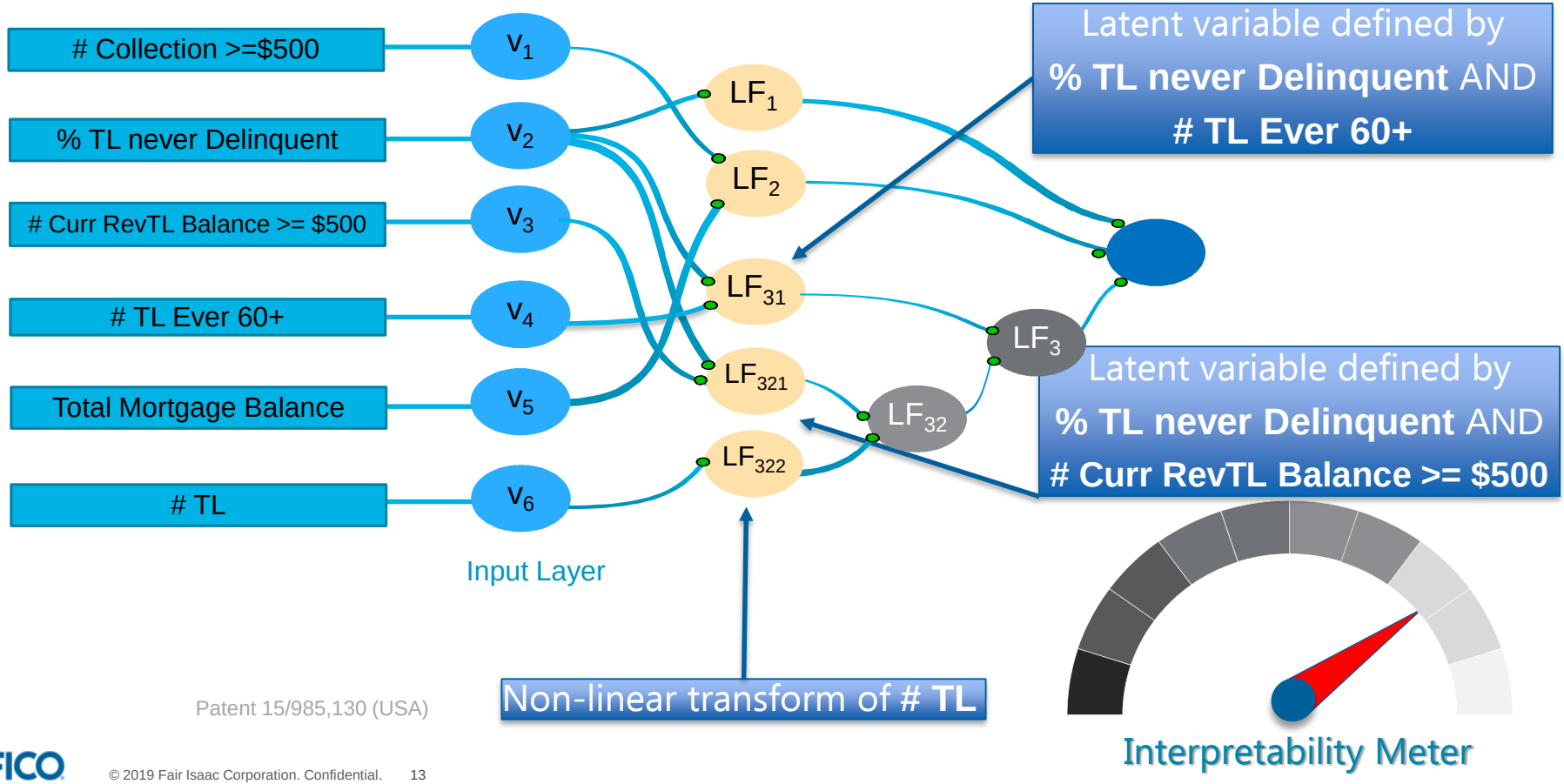
Patent 15/985,130 (USA)

~15% improvement in KS over Scorecard

What is A Latent Feature?



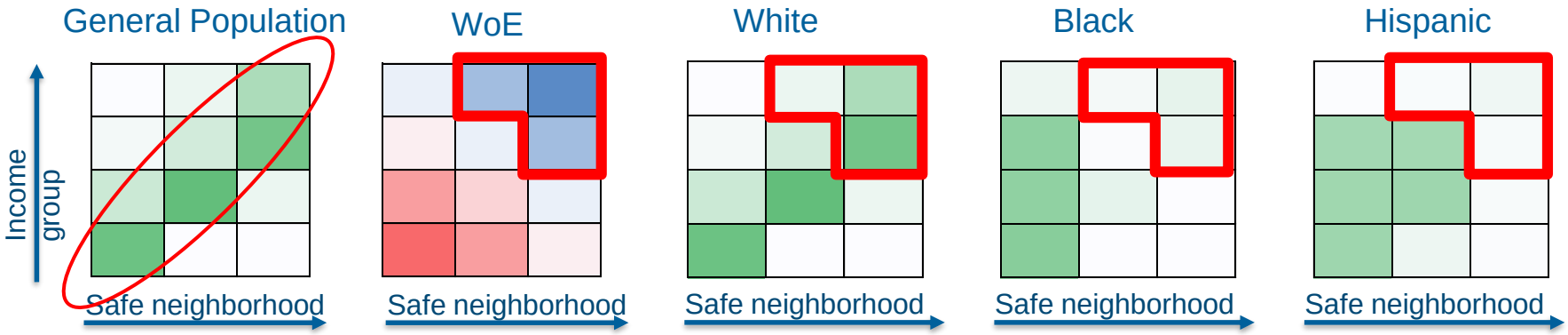
Interpretable Latent Features Model Can Expose Hidden Relationships



Patent 15/985,130 (USA)

Predictor Variable – Live in High Crime Neighborhood

Tri-variate heat-maps expose unusual distributions indicative of bias



4% blacks are wealthy and in safe neighborhoods, as compared to 30% of whites

Ignoring Race During Model Development Can Ultimately Lead To Bias

Recipe For Dealing With Bias

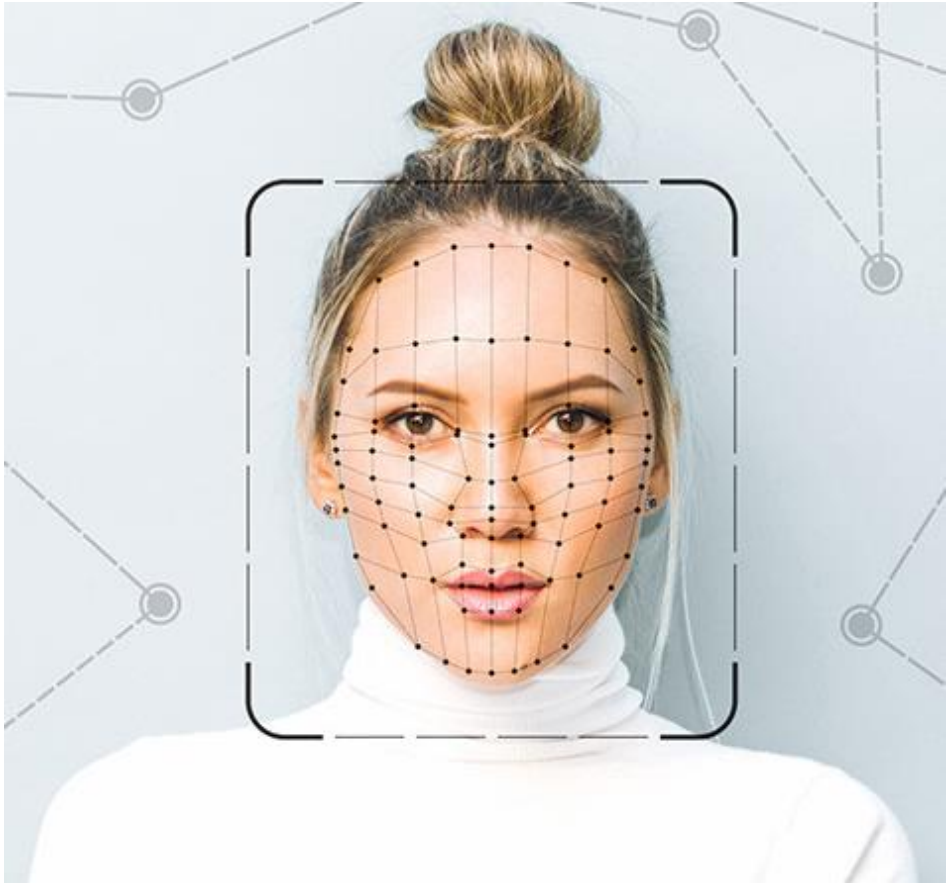
Know your laws/regulations

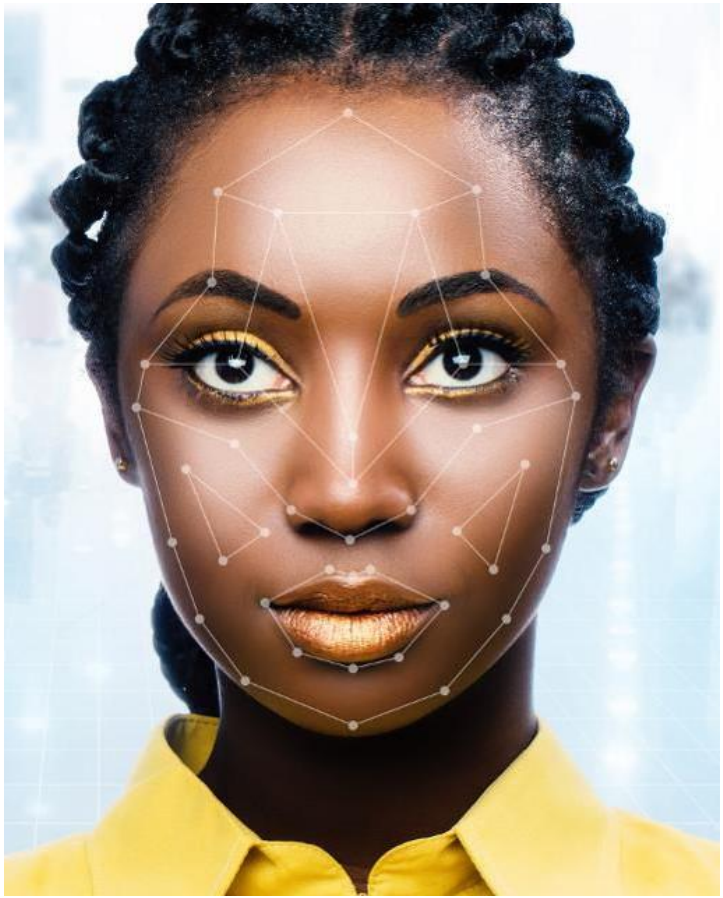
Chose input variables that are reasonable

Utilize Explainable Architectures

Find the driving relationships

Test / confirm no bias with protected classes





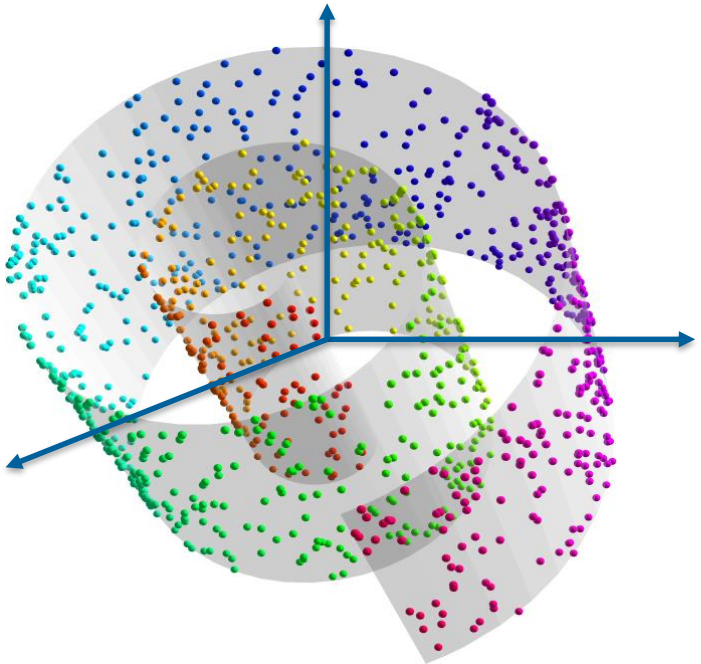
Beauty.AI used an ML model to judge a beauty contest in 2016.

Out of 44 winners, nearly all were white

Beauty.AI's chief science officer, blamed it on non-representative data used to train the model.

Manifold Hypothesis

Most high-dimensional datasets lie in the vicinity of a low-dimensional **manifold.**

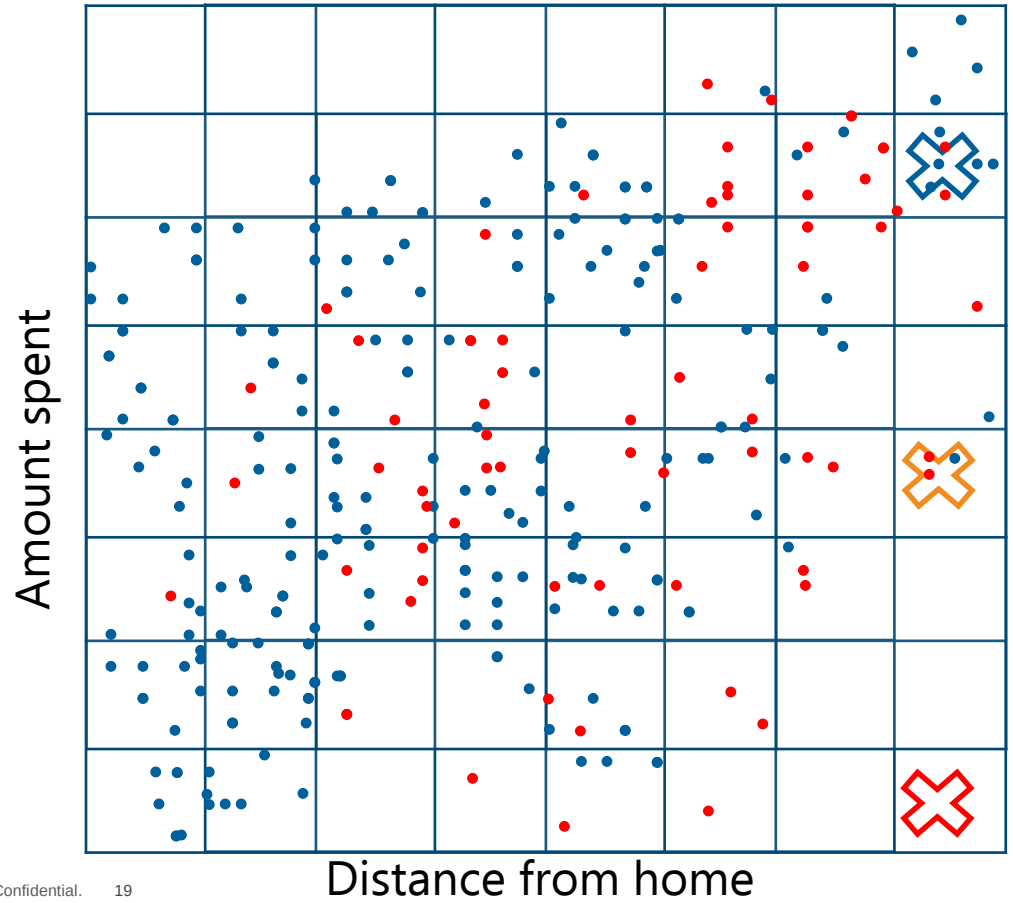


Trusting Models Mean Understanding Data Coverage

Different regions are not equally represented by the training data

Model can generalize even in regions without data

Which score will you trust more?



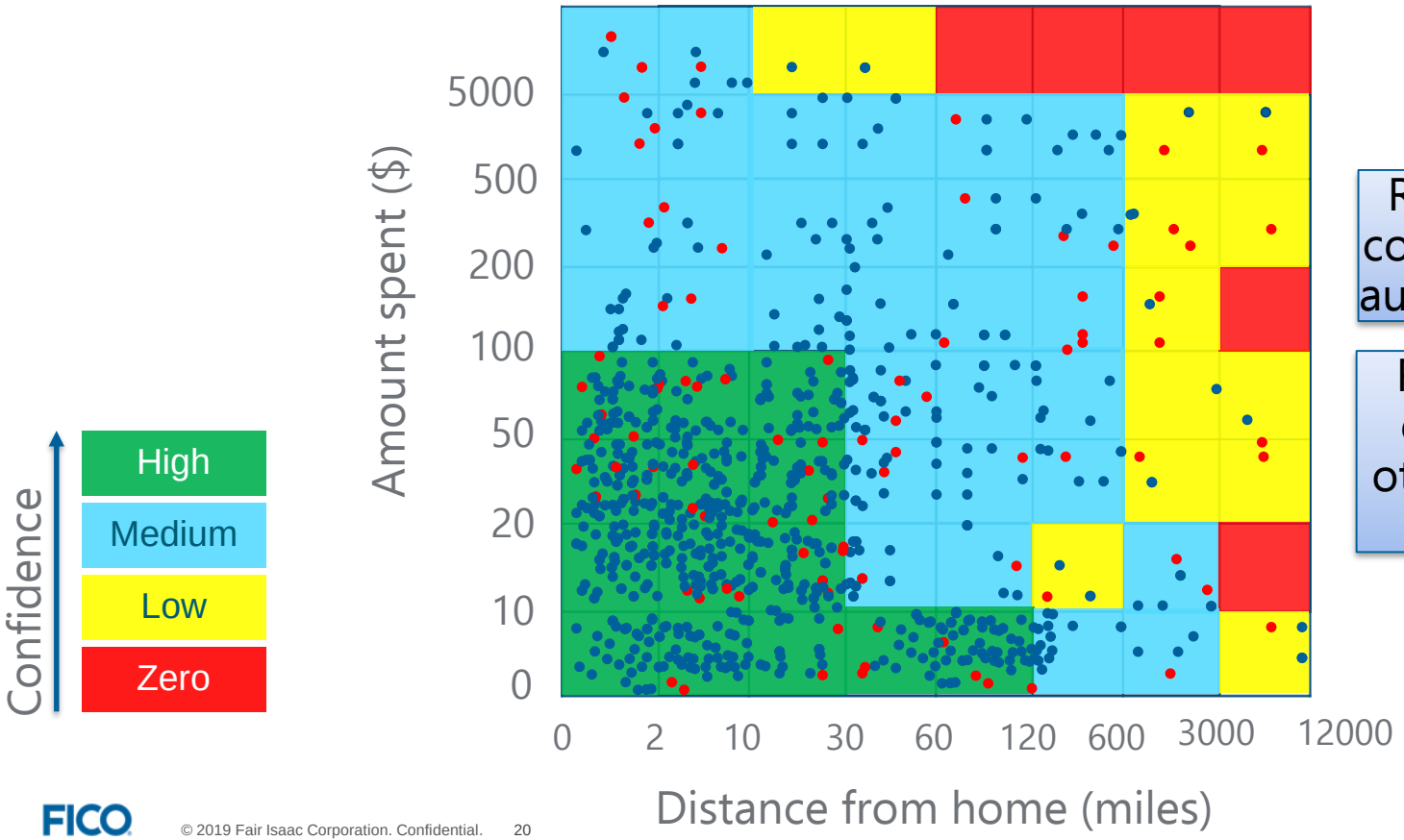
Score = 895

Score 905

Score 900

Trusting Models Mean Understanding Data Coverage

Different regions are not equally represented by the training data

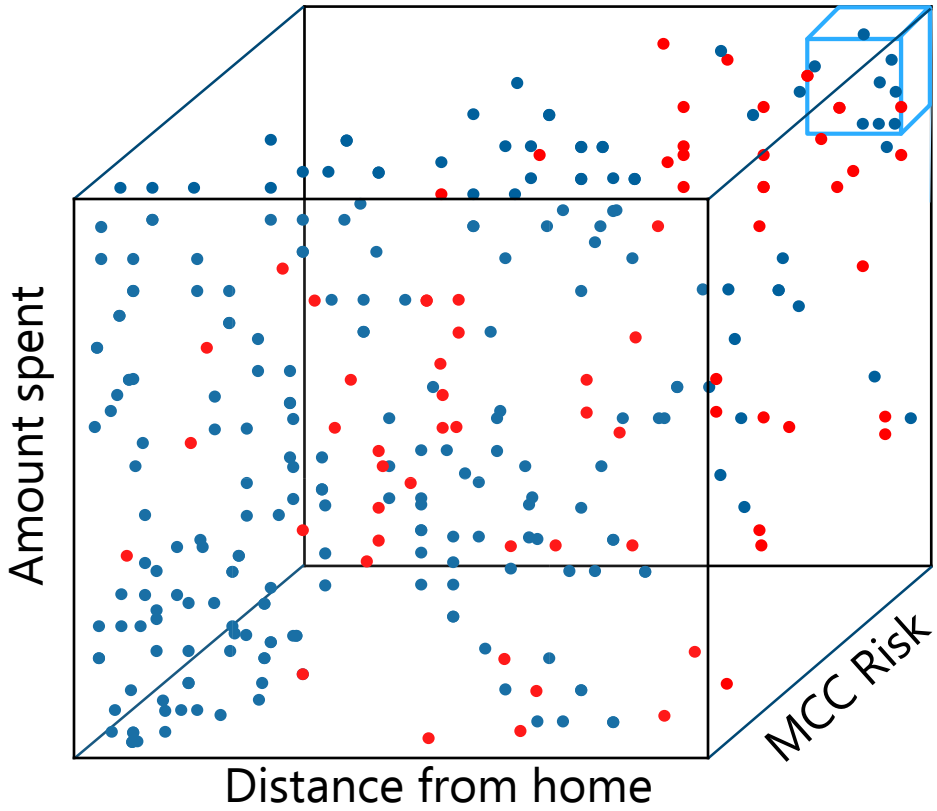


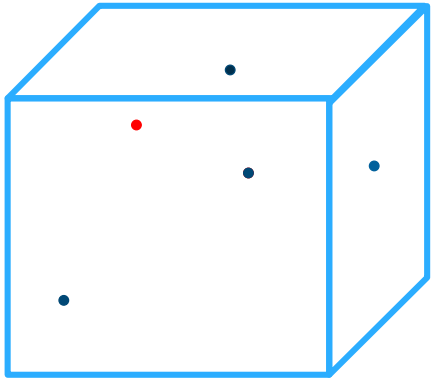
Regions with high confidence can have automated decisions

Regions with low confidence need other due diligence before action

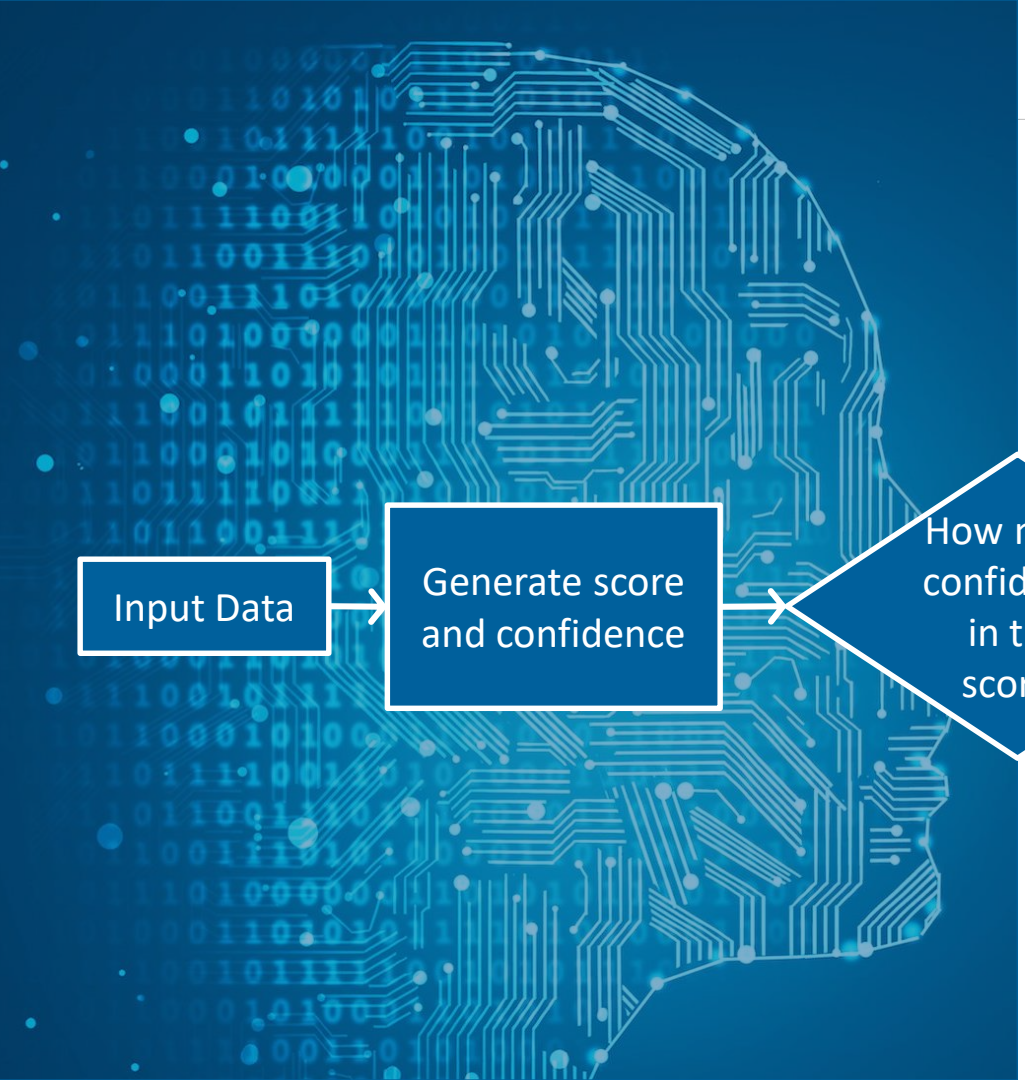
Impact of Dimensionality on Coverage

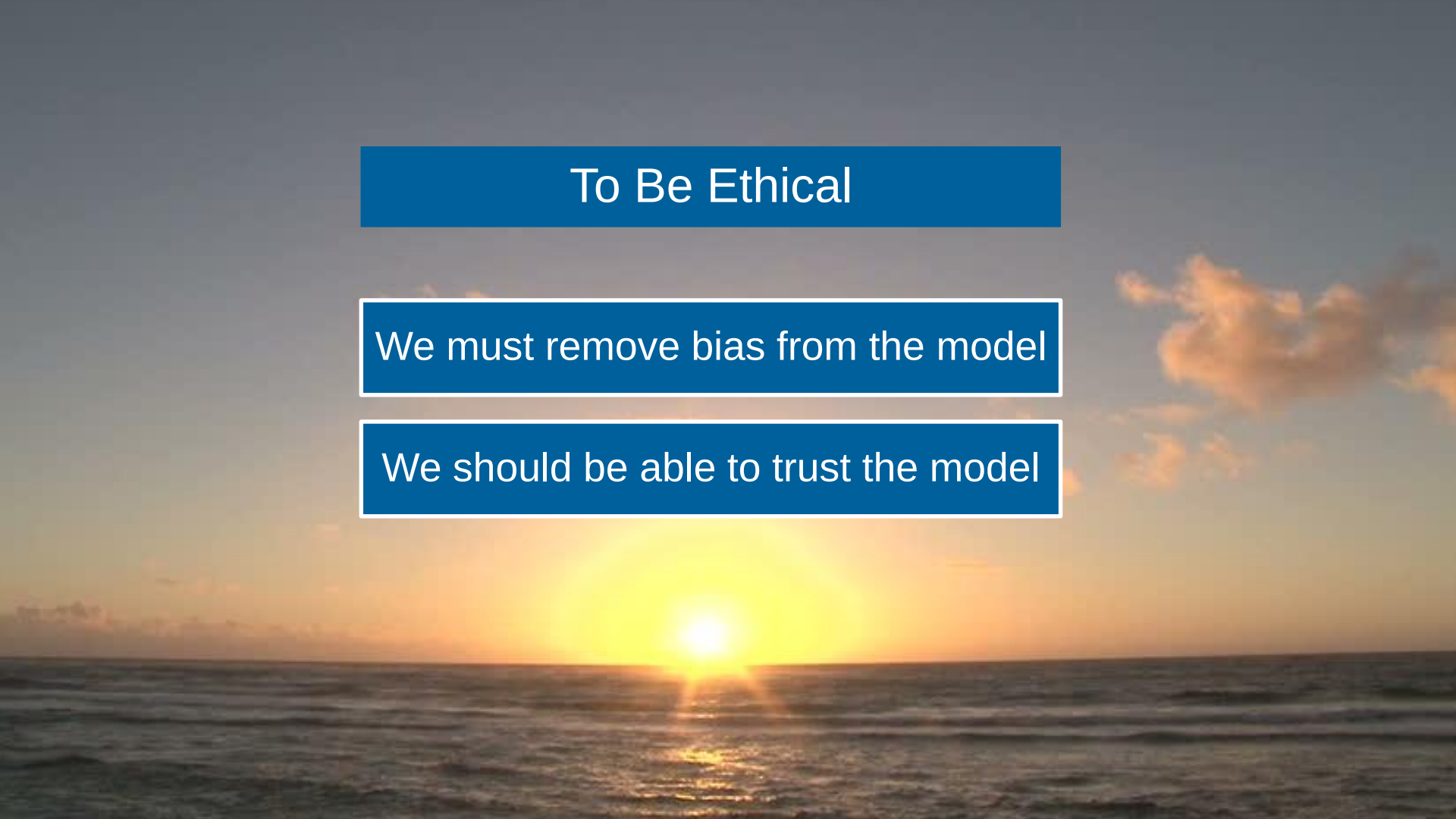
Coverage Decreases as Dimensionality Increases





Humble AI


Input Data
Generate score and confidence
How much confidence in the score?
High
Medium
Low/No
if condition1
decision1
if condition2
decision2


To Be Ethical
We must remove bias from the model
We should be able to trust the model