BITS Pilani
Pilani | Dubai | Goa | Hyderabad

FAccT Machine Learning

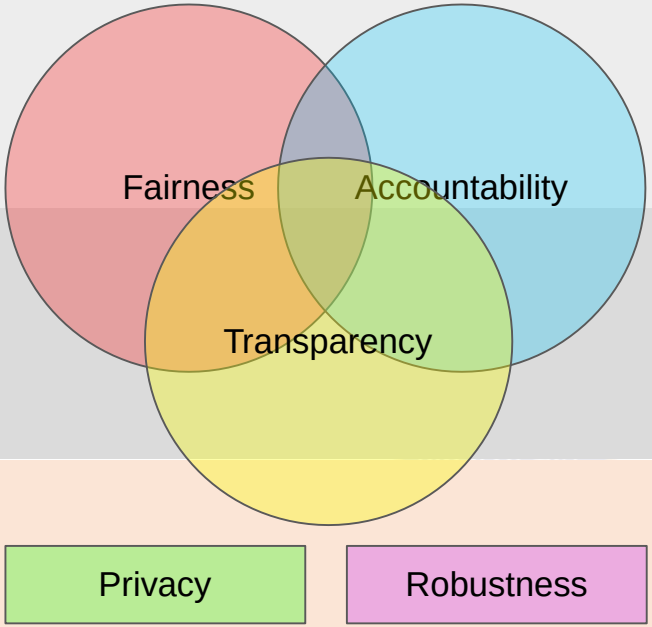
Dr. Sugata Ghosal

CSIS Off-Campus Faculty
BITS Pilani

In This Session

- Fairness
- Accountability
- Transparency

FAccT Overview



Psychology
Social Science
Public Policy

Statistics
Theory

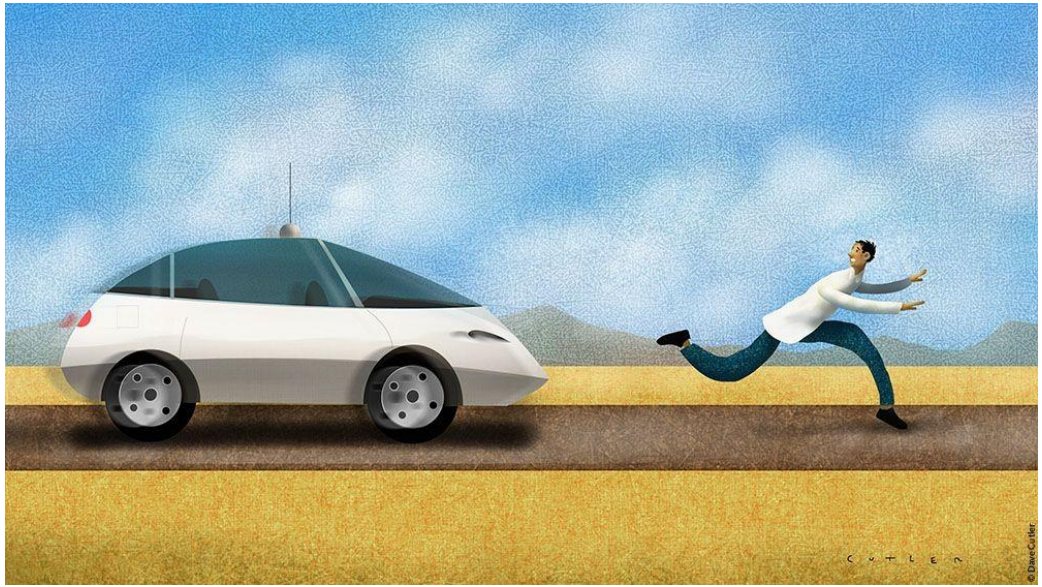
Machine
Learning

Fairness in Machine Learning

- What to do to ensure gender and ethnic fairness in ML models?

Accountability

- Who takes the responsibilities for failed ML models?



Transparency

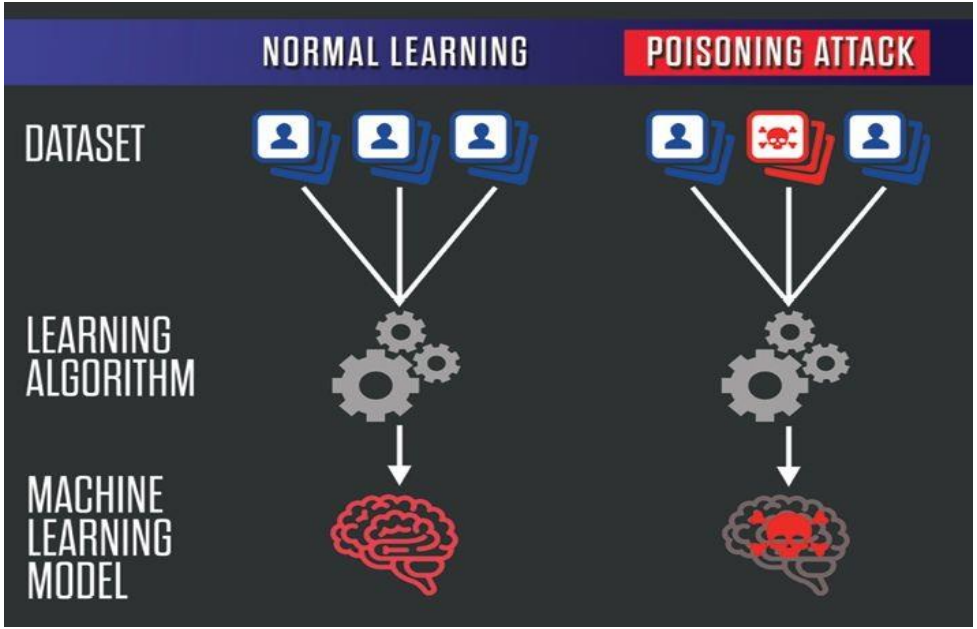
- What to do to make ML models transparent and comply with regulations?

Privacy Issues

- How to protect user privacies when exposing data to ML models?

Security Issues

- How do we defend ML models against data poisoning?



Thank You!

In our next session: Fairness and Bias

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Fairness and Bias

Dr. Sugata Ghosal

CSIS Off-Campus Faculty
BITS Pilani

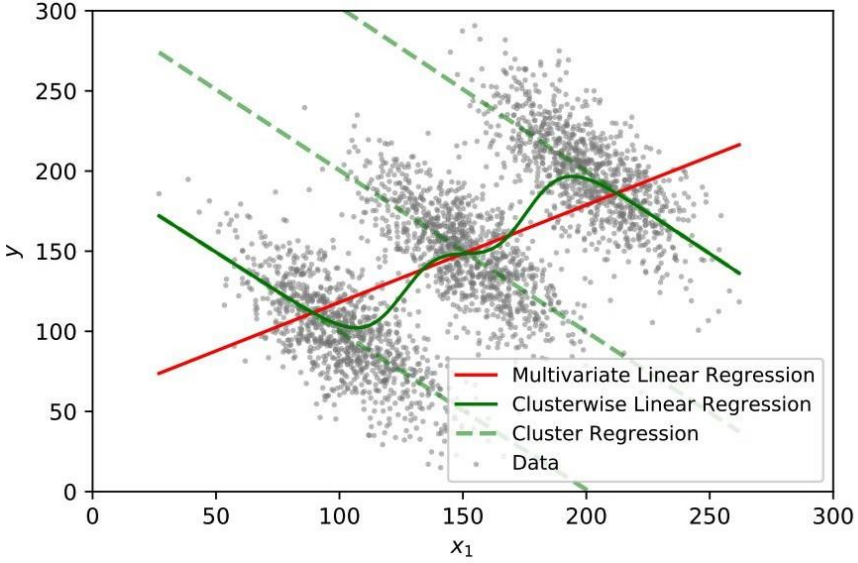
In This Session

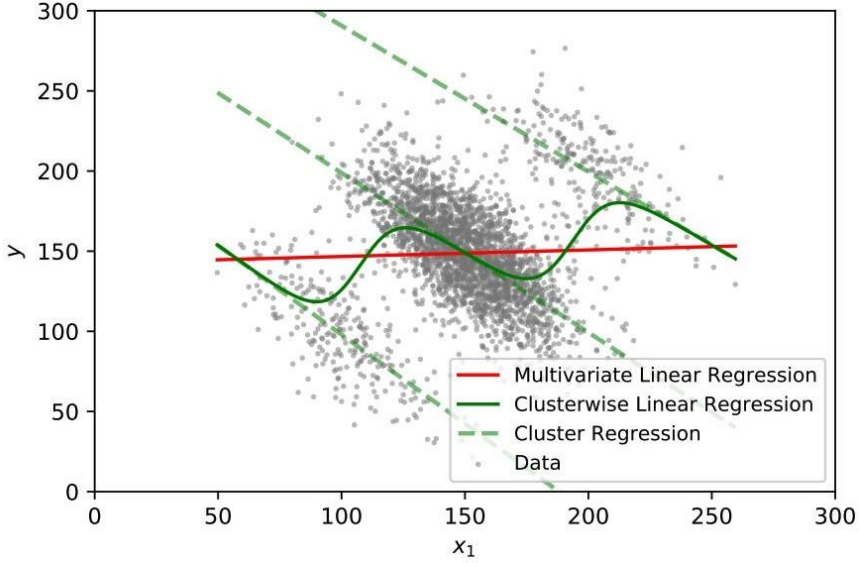
- Sources of Bias
- Real World Examples
- Sensitive Features
- Fairness Criteria
- Fairness Preserving Regularization for preserving Fairness

Fairness

- What is Fairness?
  - The absence of bias towards an individual or a group (Mehrabi et al, 2019)
- Can ML Models Discriminate?
  - Aren't machines just follow human's instructions?
  - ML models approximate patterns in the data
  - Learns/Amplifies biases at the same time

Model Sensitivity to Data





red - biased regression

dashed green - regression for each subgroup solid green - unbiased regression

Simpson's Paradox

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

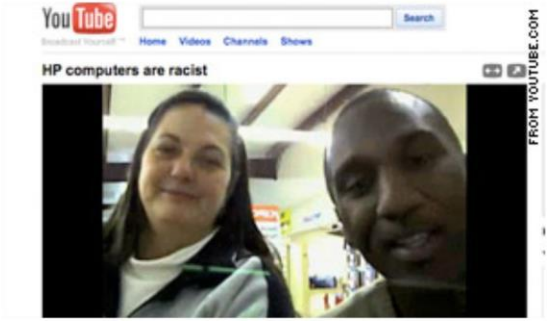
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

Real World Example

HP looking into claim webcams can't see black people

By **Mallory Simon**, CNN
December 23, 2009 7:25 p.m. EST



A YouTube video shows co-workers trying out an HP webcam with motion-tracking and facial recognition software.

STORY HIGHLIGHTS

- **NEW:** Video was meant to be humorous showing of software glitch, co-workers say
- Co-workers: Motion-tracking webcam moves with white woman, not black man
- "I think my blackness is

(CNN) -- Can Hewlett-Packard's motion-tracking webcams see black people? It's a question posed on a now-viral YouTube video and the company says it's looking into it.

In the video, two co-workers take turns in front of the camera -- the webcam appears to follow Wanda Zamen as she sways in front of the screen and stays still as Desi Cryer moves about.

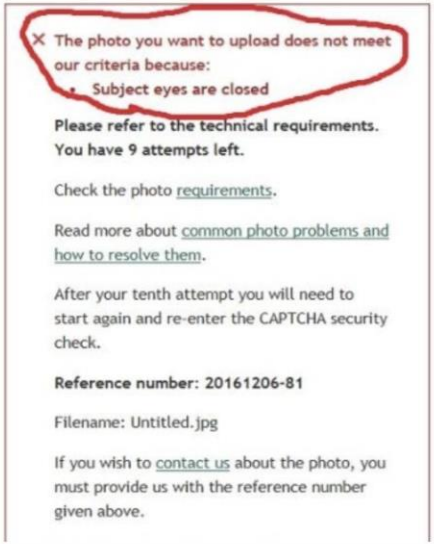
HP acknowledged in a statement e-mailed to CNN that the cameras may have issues with contrast recognition in certain lighting situations. The webcams, built into HP's new computers, are supposed to keep people's faces and bodies in proportion and centered on the screen as they move.

The video went viral over the weekend, garnering more than 400,000 YouTube page views and a slew of comments on Twitter.

Real World Example

New Zealand passport robot thinks this Asian man's eyes are closed

By James Griffiths, CNN
Updated 1:46 AM ET, Fri December 9, 2016



New Zealand's online passport application system couldn't recognize Richard Lee's open eyes.

Real World Example

Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'

Rhett Jones
10/10/18 10:32AM • Filed to: ALGORITHMS

79 3

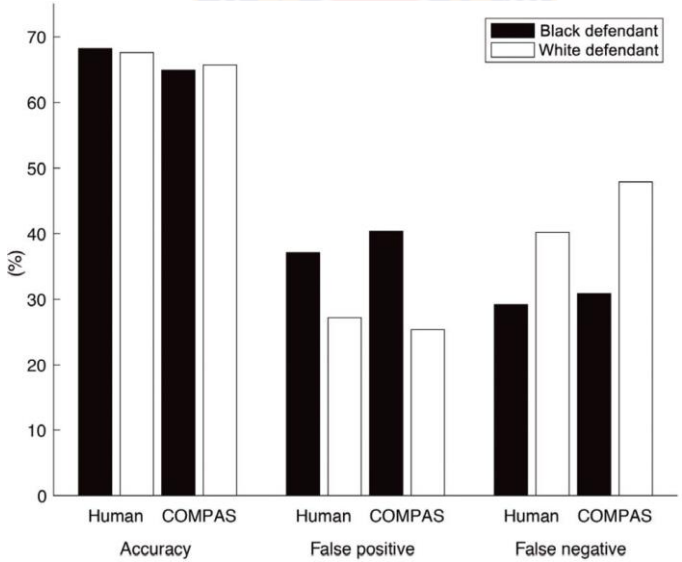
Photo: Getty

Recent Video

Algorithmic Bias

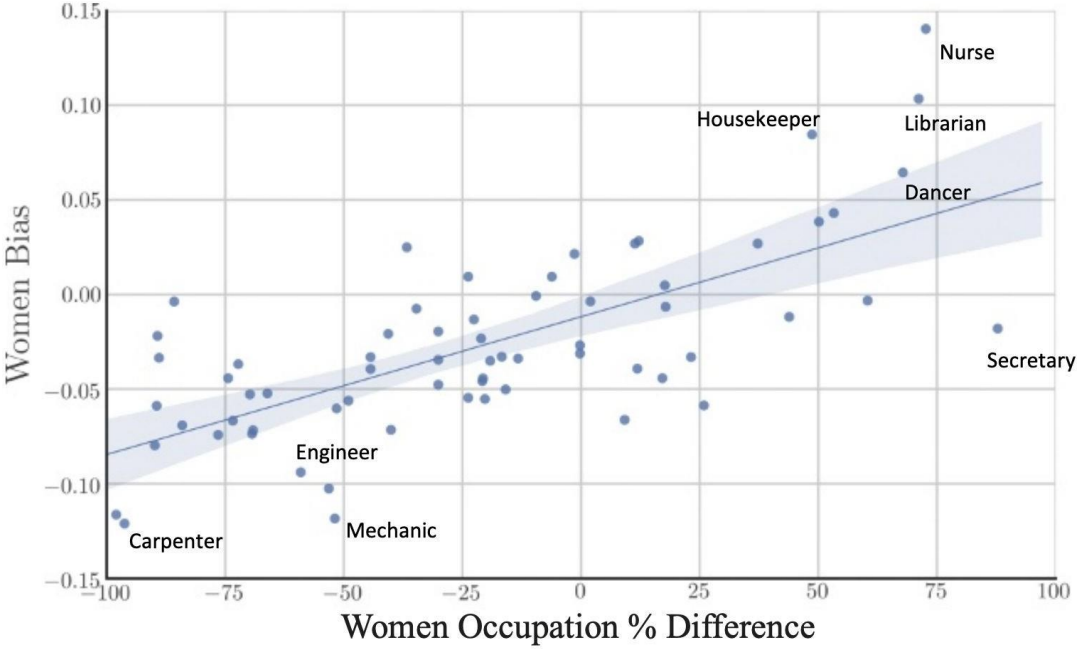
Commercial risk assessment software known as COMPAS

- Assess more than 1 million offenders since 2000
- Predicts a defendant's risk of committing a misdemeanor or felony 137 features



Dressel et al, 2018

Bias in Historical Data



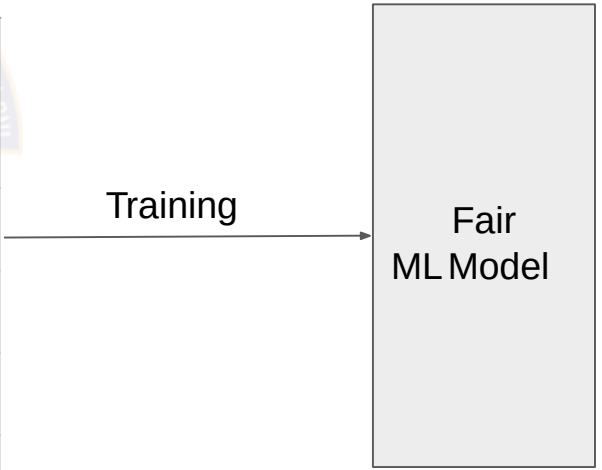
Gard et al, 2018

Fairness Through Unawareness

A ML Algorithm Achieves Fair Through Unawareness If

- None of the sensitive features are directly used in the model

Race and Ethnicity	Skills	Years of Exp	Hired?
Hispanic	Javascript	1	no
Hispanic	C++	5	yes
White	Java	2	yes
White	C++	3	yes



Indirect Evidence

Sensitive Features May Still Be Used

- Inferred from indirect evidence

Race and Ethnicity	Skills	Years of Exp	Often Goes to Mexican Markets	Hiring Decision
Hispanic	Javascript	1	yes	no
Hispanic	C++	5	yes	yes
White	Java	2	no	yes
White	C++	3	no	yes

Training → Discriminatory ML Model

Common Fairness Criteria

- Demographic Parity
- Equality of Odds/Opportunity

Demographic Parity

Demographic Parity Is Applied to a Group of Samples

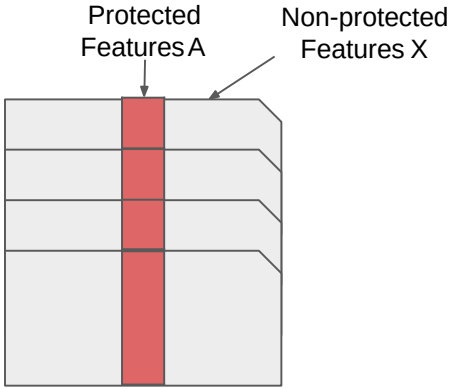
- Demographic Parity Is Applied to a Group of Samples
  - Does not require features to be masked out

- A Predictor $\hat{Y}$ Satisfies Demographic Parity If
  - The probabilities of positive predictions are the same regardless of whether the group is protected
  - Protected groups are identified as A = 1

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

Comparisons

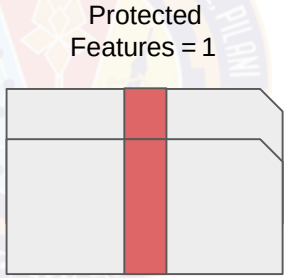
Individual Treatment



Fairness Through Unawareness

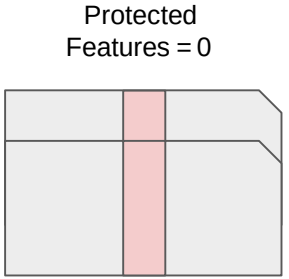
$P(\hat{Y} \mid X)$

Group Treatment



Demographic Parity

$P(\hat{Y}=1 \mid A=1)$



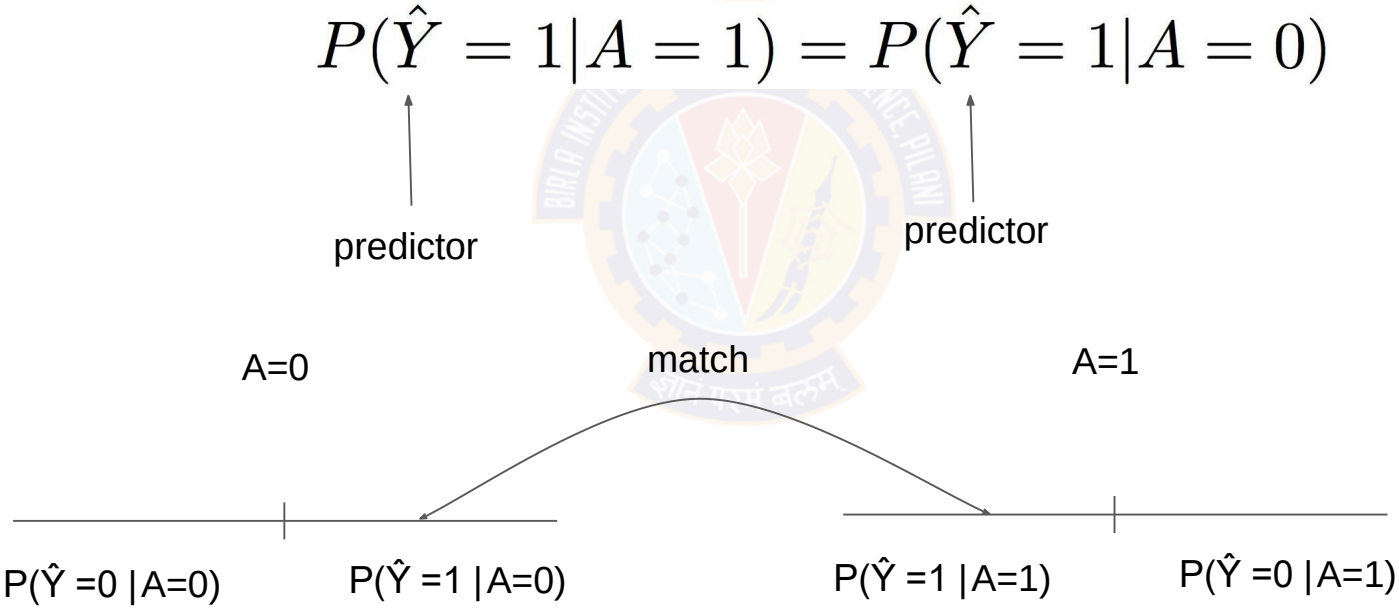
Demographic Parity

$P(\hat{Y}=1 \mid A=0)$

Issues With Demographic Parity

- Correlates Too Much With the Performance of the Predictor

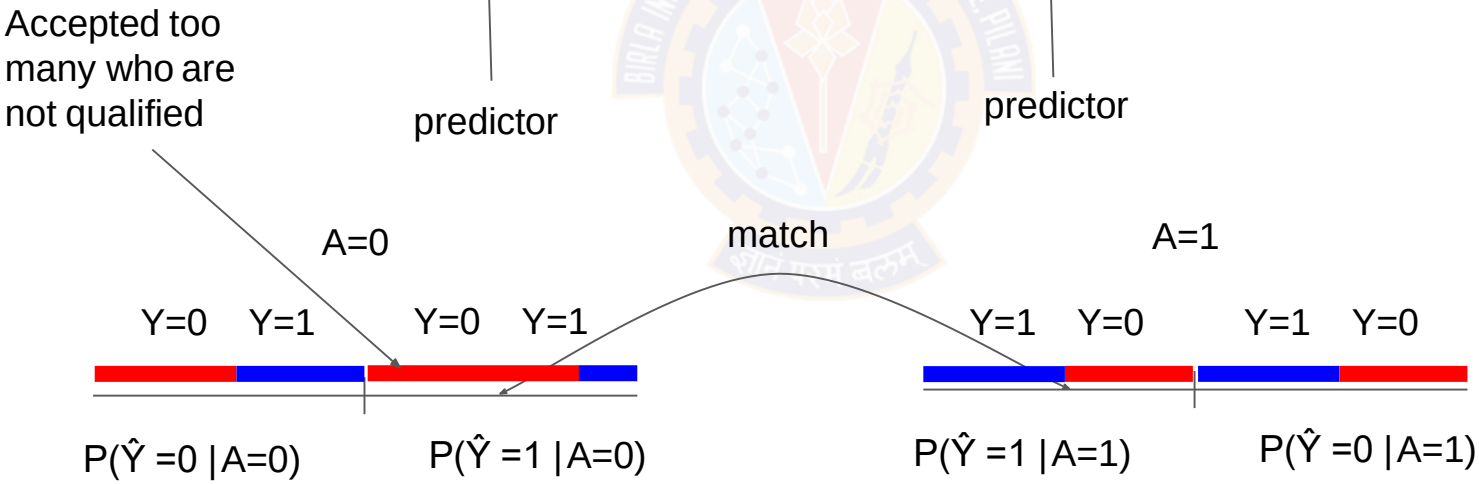
$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$



Issues With Demographic Parity

- Correlates Too Much With the Performance of the Predictor

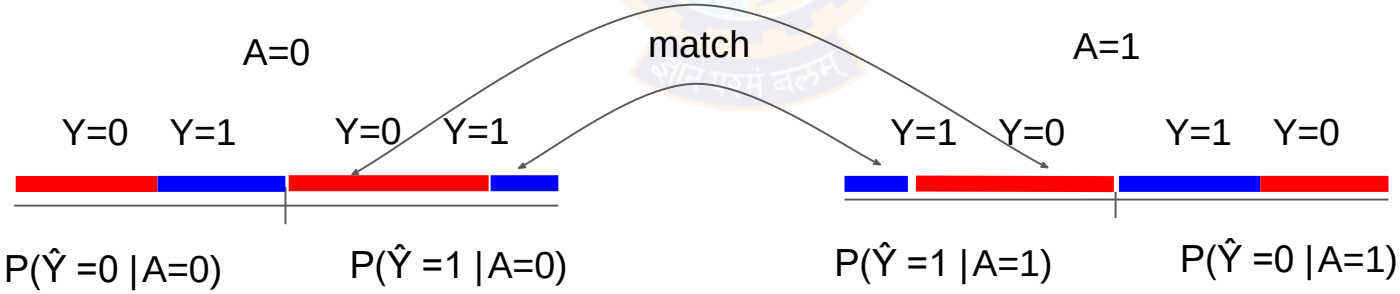
$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$



Equality of Odds

- Equal Probabilities for Both Qualified/Unqualified People Across Protected Groups

$$P(\hat{Y} = 1|A = 0, Y) = P(\hat{Y} = 1|A = 1, Y)$$

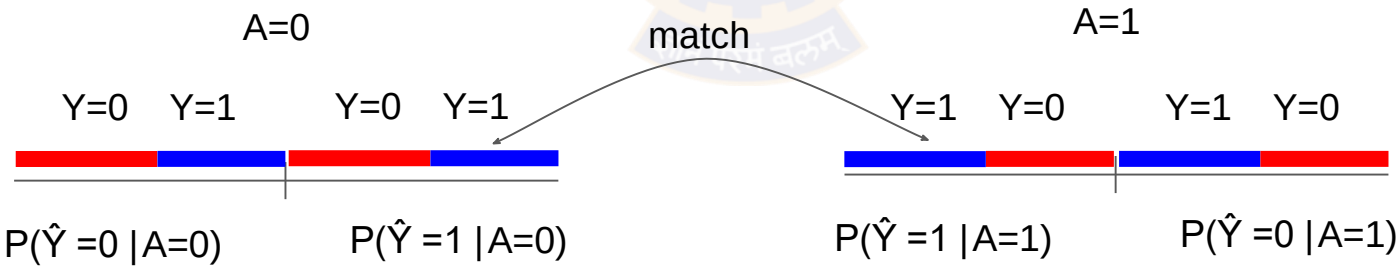


Hardt et al, 2016

Equality of Opportunity

- Equal Probabilities for Qualified People Across Protected Groups

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$



Hardt et al, 2016

Practice Question

Find out the Fairness Criteria that $\hat{Y}1$, and $\hat{Y}2$ Satisfy

- A = {race}, Y = {Hiring Decision}

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Demographic Parity for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Demographic Parity for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 2/3$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Demographic Parity for Predictor Ŷ1

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 2/3$

✔ Demographics Parity

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) =$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) =$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 0.5$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) =$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 0.5$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) = 0$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 0.5$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) = 0$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) = 1$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 0.5$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) = 0$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) = 1$

✗ Equality of Opportunity

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

✗ Equality of Odds

$$P(\hat{Y} = 1|A = 0, Y) = P(\hat{Y} = 1|A = 1, Y)$$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Demographic Parity for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Demographic Parity for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 1/3$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Demographic Parity for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 1/3$

✗ Demographics Parity

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) =$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) =$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Demographic Parity for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 1/3$

✗ Demographics Parity

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) =$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) = 0$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Equality of Opportunity/Odds for Predictor Ŷ2

- P(Ŷ1 = 1 | R = H, Y = yes) = 1/2
- P(Ŷ1 = 1 | R = W, Y = yes) = 1/2
- P(Ŷ1 = 1 | R = H, Y = no) = 1
- P(Ŷ1 = 1 | R = W, Y = no) = 0

✔ Equality of Opportunity

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

✗ Equality of Odds

$$P(\hat{Y} = 1|A = 0, Y) = P(\hat{Y} = 1|A = 1, Y)$$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

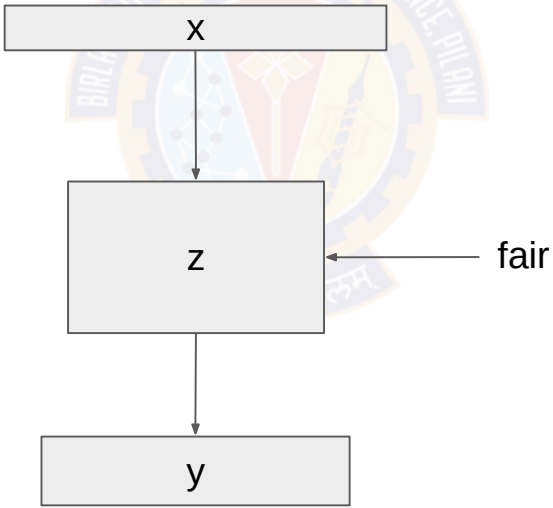
Summary of Fairness Criteria

Fairness Criteria	Criteria	Group	Individual
Unawareness	Excludes A in Predictions		✓
Demographic Parity	$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$	✓	
Equalized Odds	$P(\hat{Y} = 1 \mid A = 0, Y) = P(\hat{Y} = 1 \mid A = 1, Y)$	✓	
Equalized Opportunity	$P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1)$	✓	

Fair Representation Learning

Make representations fair

- Ensure fairness up to a certain level



Prejudice Remover Regularizer

Quantified causes of unfairness

- Prejudice
  - Unfairness rooted in the dataset
- Underestimation
  - Model unfairness because the model is not fully converged
- Negative Legacy
  - Unfairness due to sampling biases
- Training Objective

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model | Fairness Regularizer | L2 Regularizer

Kamishima et al, 2012

Prejudice Index (PI)

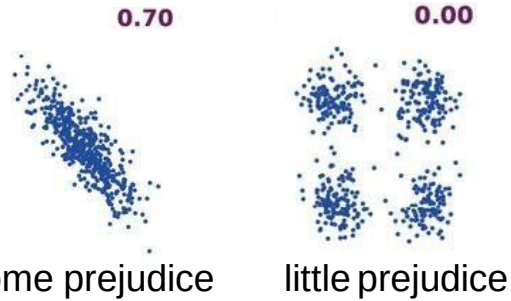
Recall that Indirect Discrimination Happens When

- Prediction is not directly conditioned on sensitive variables S
- Prediction is indirectly conditioned on S by a variable O that is dependent on S
- $P(\hat{Y} \mid O)$, and $O \sim P(O \mid S)$

- Prejudice Index (PI)
  - Measures the degree of indirect discrimination based on mutual information

$$\mathrm{PI} = \sum_{(y,s)\in\mathcal{D}} \hat{\Pr}[y,s] \ln \frac{\hat{\Pr}[y,s]}{\hat{\Pr}[y]\hat{\Pr}[s]}$$

prediction model



some prejudice little prejudice

Kamishima et al, 2012

Optimizing PI

$$\mathrm{PI} = \sum_{(\mathbf{x}_i, s_i) \in \mathcal{D}} \sum_{y \in \{0,1\}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

$$\hat{\Pr}[y|s] \approx \frac{\sum_{(\mathbf{x}_i, s_i) \in \mathcal{D} \text{ s.t. } s_i = s} \mathcal{M}[y|\mathbf{x}_i, s; \boldsymbol{\Theta}]}{|\{(\mathbf{x}_i, s_i) \in \mathcal{D} \text{ s.t. } s_i = s\}|} \qquad \hat{\Pr}[y] \approx \frac{\sum_{(\mathbf{x}_i, s_i) \in \mathcal{D}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}]}{|\mathcal{D}|}$$

Kamishima et al, 2012

Putting Things Together

Optimization Target

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model | Fairness Regularizer | L2 Regularizer

- Fairness Regularizer

$$\mathrm{PI} = \sum_{(\mathbf{x}_i, s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

Kamishima et al, 2012

Thank You!

In our next session: Interpretability

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Interpretability and Accountability

Dr. Sugata Ghosal

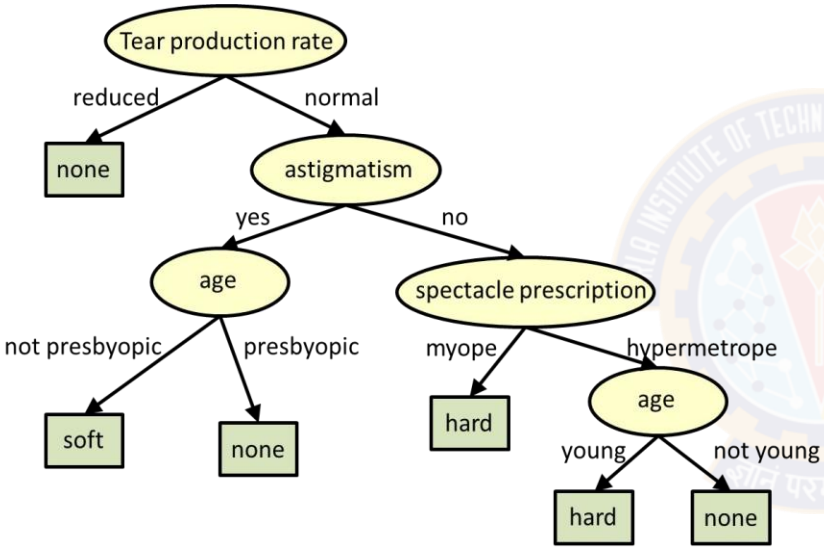
CSIS Off-Campus Faculty
BITS Pilani

In This Segment

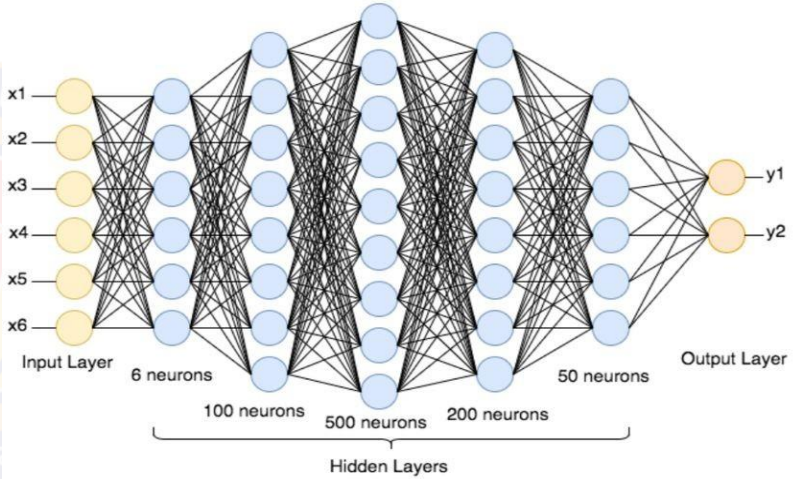
- ML Interpretability
- Intrinsically Interpretable Models
  - Simple interpretable models
  - Intrinsically interpretable techniques for deep learning
- Interpretability Concepts
  - Intrinsic and post hoc methods
  - model-specific and model-agnostic methods
  - Local and global interpretable methods
  - Interpretability and performance trade-offs

Machine Learning Interpretability

ML interpretability allows one to examine model's basis in its decision making process



An interpretable tree model to find out the kind of contact lens a person may wear



A neural network which is usually considered a black-box model.

Reasons for ML Interpretability

- Our society has been shifted to rely on AI more than ever
  - autonomous vehicles
  - security
  - finance
  - many others

- Who will will benefit from ML Interpretability?
  - End Users: enhance trust, understand the consequences of the decisions, e.g., privacy, fairness.
  - Regulatory Agencies: compliance, aduits, and accountability.
  - Model Designers: diagnose model performance

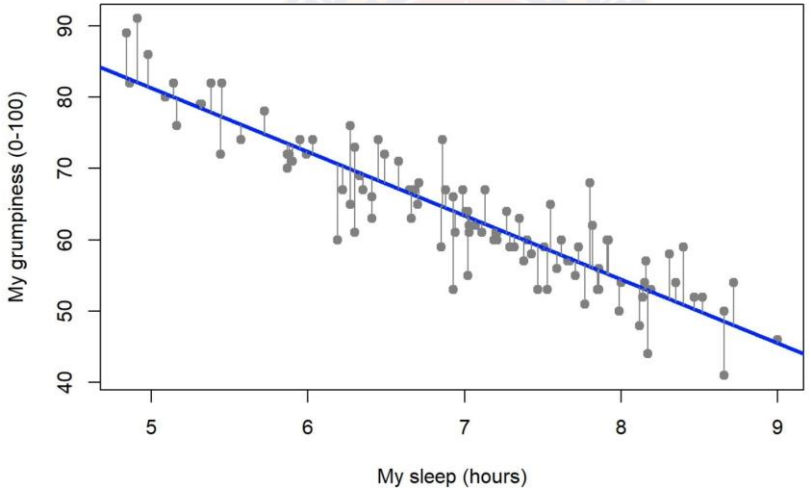
Intrinsically interpretable models

- Models that are interpretable by design
- No post-processing steps are needed to achieve interpretable.

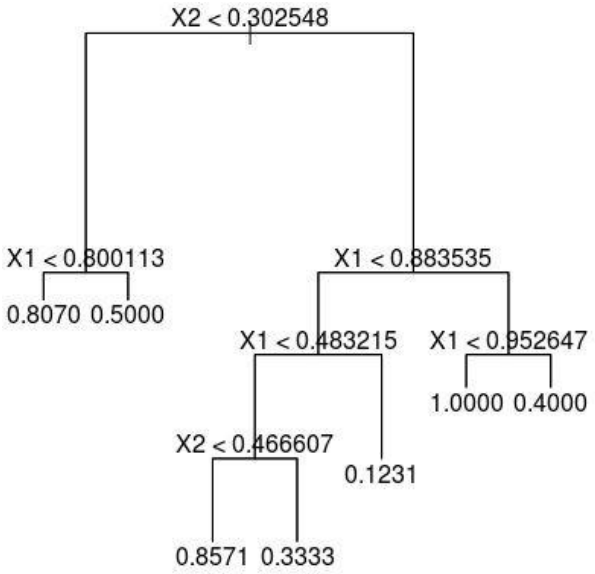
Linear Regression

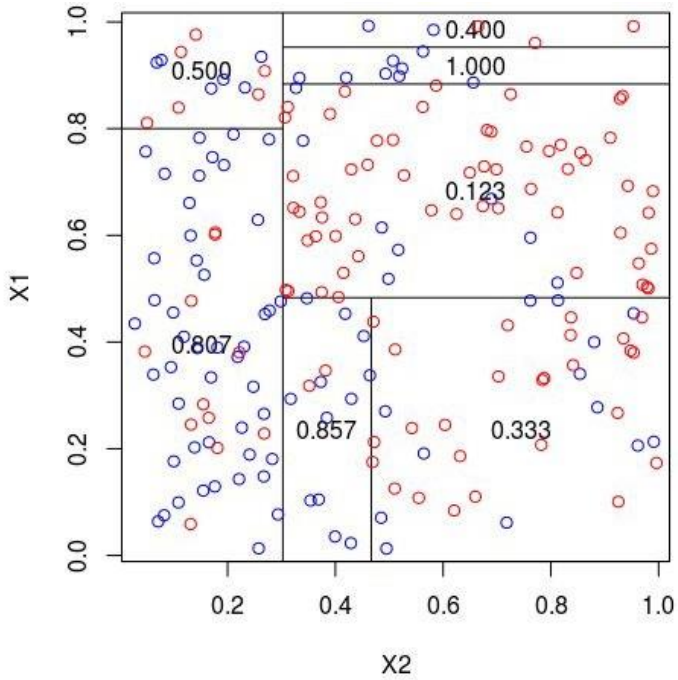
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

interpretable components

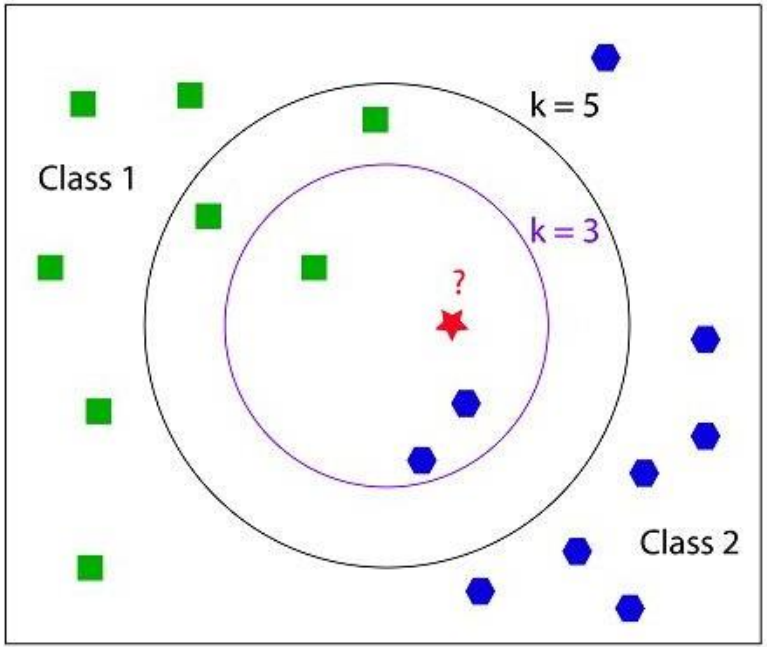


Decision Trees





K-Nearest Neighbors



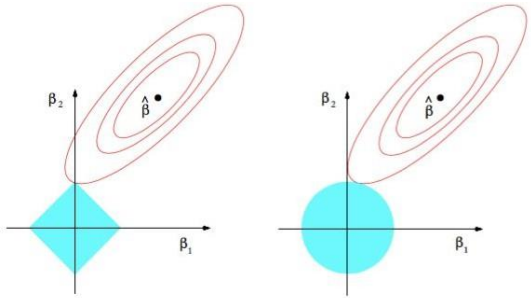
Sparsity

- Controls the sparsity of model parameters when learning a model
- Popular choices
  - L1 regularization

$$||\mathbf{w}||_1 = |w_1| + |w_2| + ... + |w_N|$$

  - L2 regularization

$$||\mathbf{w}||_2 = (w_1^2 + w_2^2 + ... + w_N^2)^{\frac{1}{2}}$$



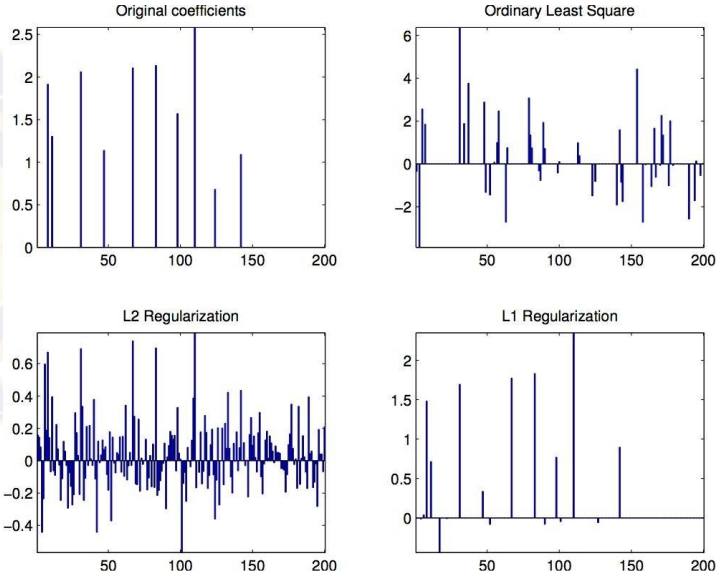
Sparsity for Interpretable Linear Regression

- In the case of linear regression
  - $\hat{y} = w_1 x_1 + w_2 x_2 + ... + w_N x_N + b$
- Linear regression with L1 regularization

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} |w_i|$$

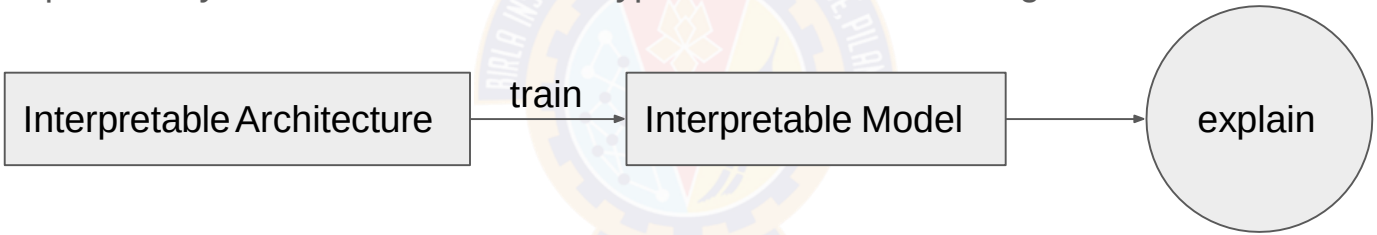
- Linear Regression with L2 regularization

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} w_i^2$$

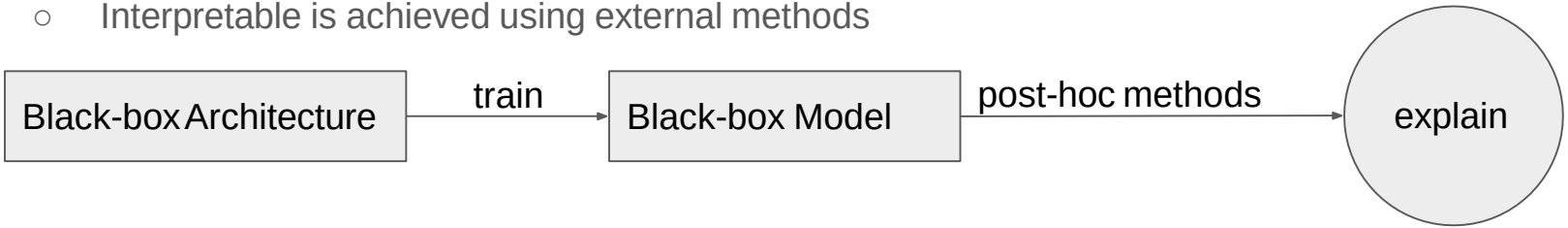


Intrinsic and Post Hoc Interpretability

- Intrinsically interpretable models
  - Interpretable is achieved by model design
  - ML models are explainable by itself
  - Explainability is often achieved as a byproduct of model training

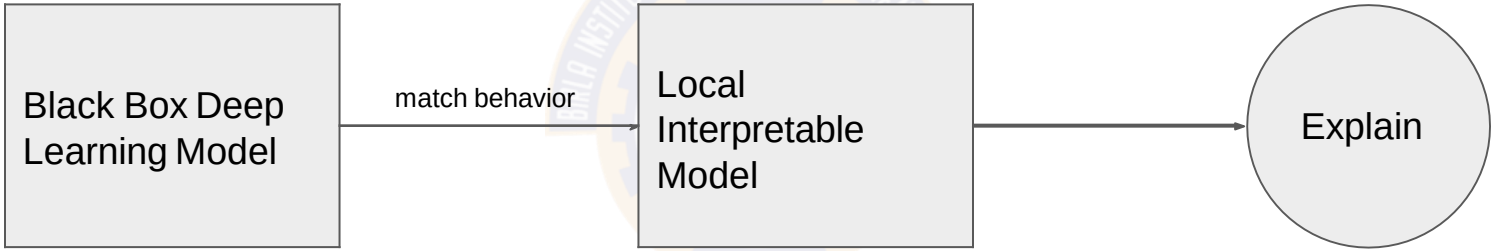


- Post Hoc/Model-specific methods
  - Explainability is often achieved after the model is trained
  - Interpretable is achieved using external methods



Post Hoc Interpretability

- One of the way to achieve Post Hoc Interpretability is to deploy a local proxy model



Model Specific and Model Agnostic Methods

- Model Specific Methods
  - Techniques that can be used for a specific architecture
  - Usually preferable when you have the ability to design your own model
  - Model specific techniques might compromise the performance of your model
  - Requires training the model using a dataset
  - Intrinsic methods are by definition model specific

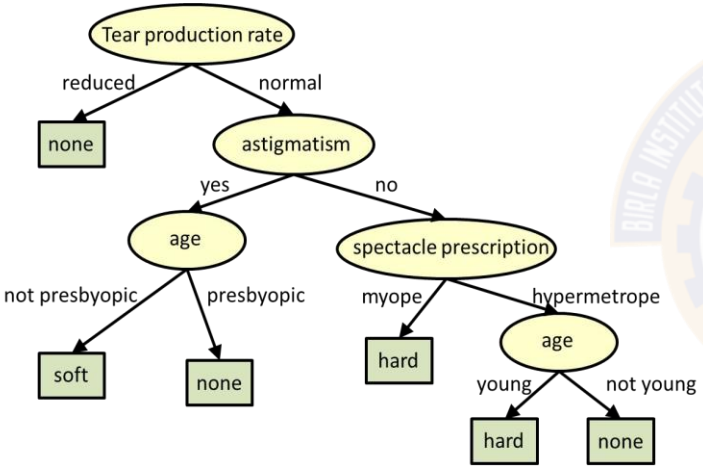
- Model-agnostic Methods
  - Techniques that can be used across many black box models
  - Model-agnostic methods will not affect the performance of your model
  - Do not require training the model
  - Post hoc methods are usually model-agnostic

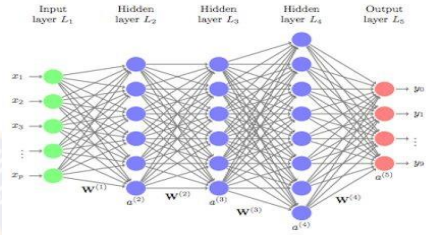
Global and Local Interpretability

- Global Interpretability
  - Explains the entire ML model at once from input to prediction
    - 1) Holistic Model Interpretability
    - 2) Modular Level Interpretability
  - e.g., Decision Trees, Linear regression

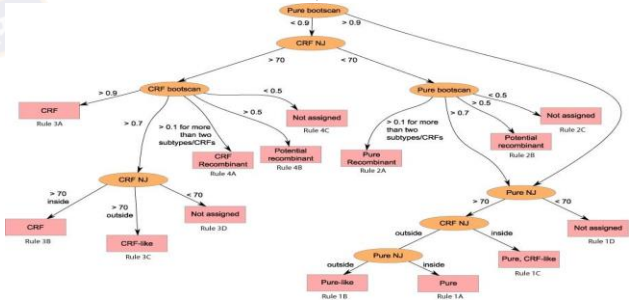
- Local Interpretability
  - Explain how predictions change for when input changes
    - 1) For a single prediction
    - 2) for a group of predictions

Global Interpretability

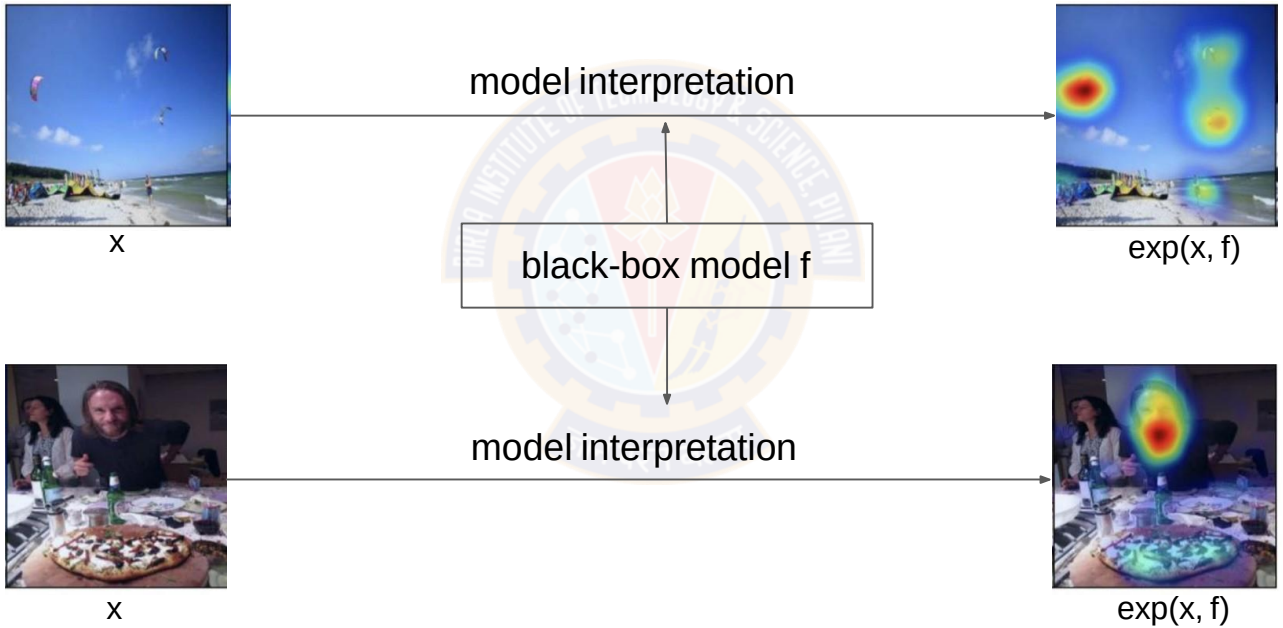




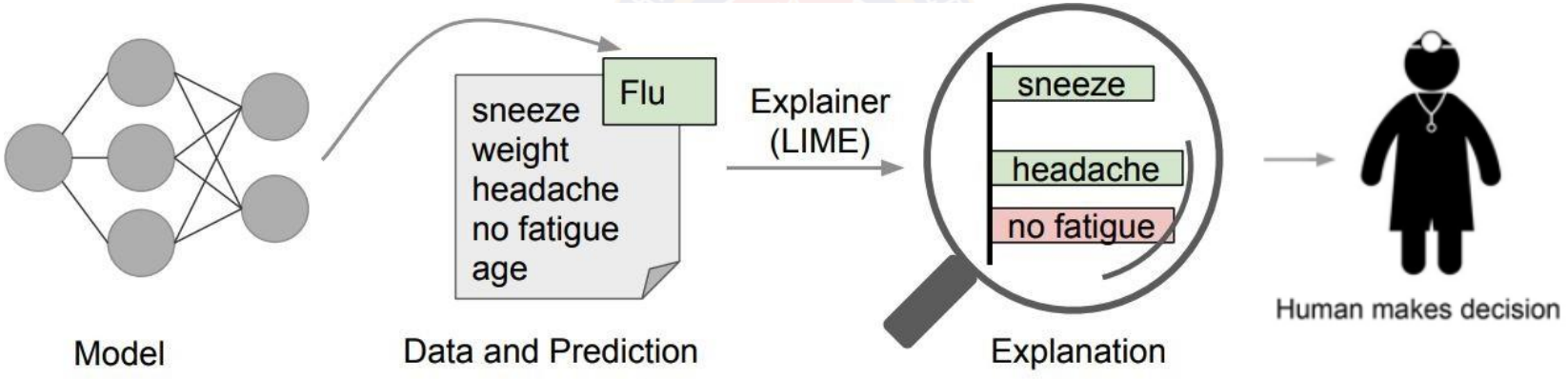
globally interpretable proxy



Local Interpretability



Local Interpretability



Local Interpretable Model-Agnostic Explanations (LIME)

- an “explainer” that highlights the symptoms that are most important to the model or in other words rationale behind the model.
- Learn the behavior of the underlying model by perturbing the input and see how the predictions change. Perturb the input by changing components that make sense to humans (e.g., attributes of a record, or words or parts of an image), even if the model is using much more complicated components.
- Generate an explanation by approximating the underlying model by an interpretable one (such as a linear model with only a few non-zero coefficients), learned on perturbations of the original instance
- The key intuition behind LIME is that it is much easier to approximate a black-box model by a simple model *locally* (in the neighborhood of the prediction we want to explain), as opposed to trying to approximate a model globally.
- the three highlighted symptoms may be a faithful approximation of the black-box model for patients who look like the one being inspected, but they probably do not represent how the model behaves for all patients.
- In the case of an image (or text), perturbation may involve hiding parts of the image (or removing some words) and then weighting the perturbed images by their similarity to the instance we want to explain.

Shapley Additive Explanations (SHAP)

- *Inspired by cooperative game theory*
- *Shapley value for each variable (payout) finds the correct weight such that the sum of all Shapley values is the difference between the predictions and average value of the model.*
  - *Shapley values correspond to the contribution of each feature towards pushing the prediction away from the expected value of predictions.*

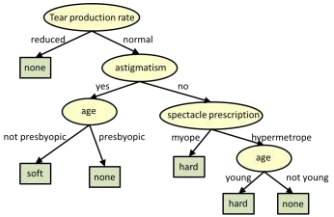
- To get the importance of feature X{i}:
- Get all subsets of features S that do not contain X{i}
- Compute effect on our predictions of adding X{i} to all those subsets
- Aggregate all contributions to compute the marginal contribution of the feature
- Now, for these subsets, for the removed or left out feature, SHAP just replaces it with the average value of the feature and generates the predictions.
  - SHAP does not go on and retrain the model for each subset.

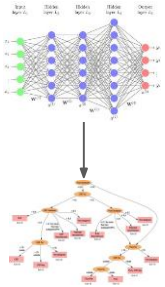
The Big Picture

Intrinsic

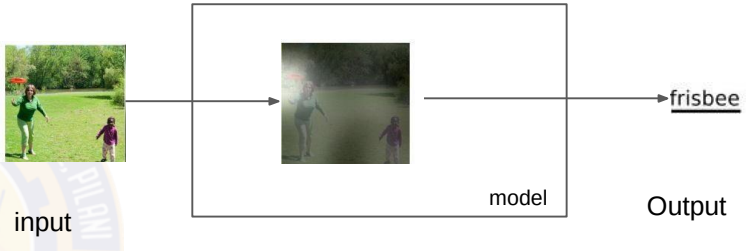
Post Hoc

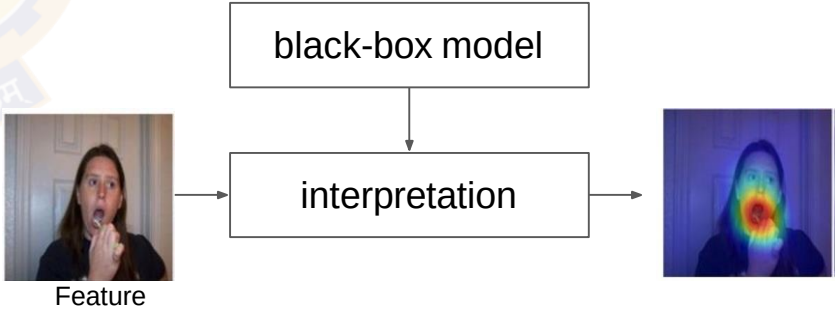
Globally Interpretable





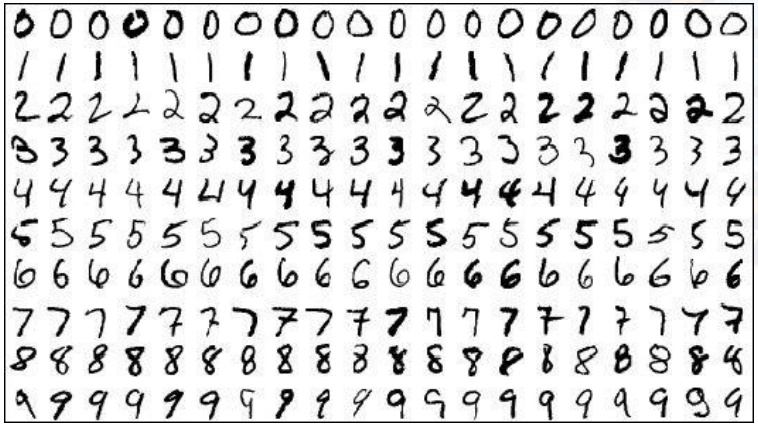
Locally Interpretable



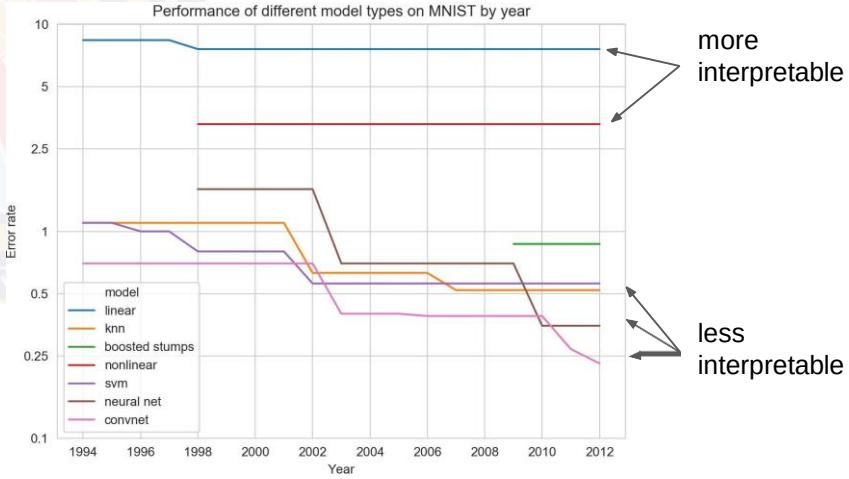


Interpretability and Performance Trade-offs

- highly performed models tend to be less interpretable.
- Can powerful models with complex structures be interpretable at the same time?

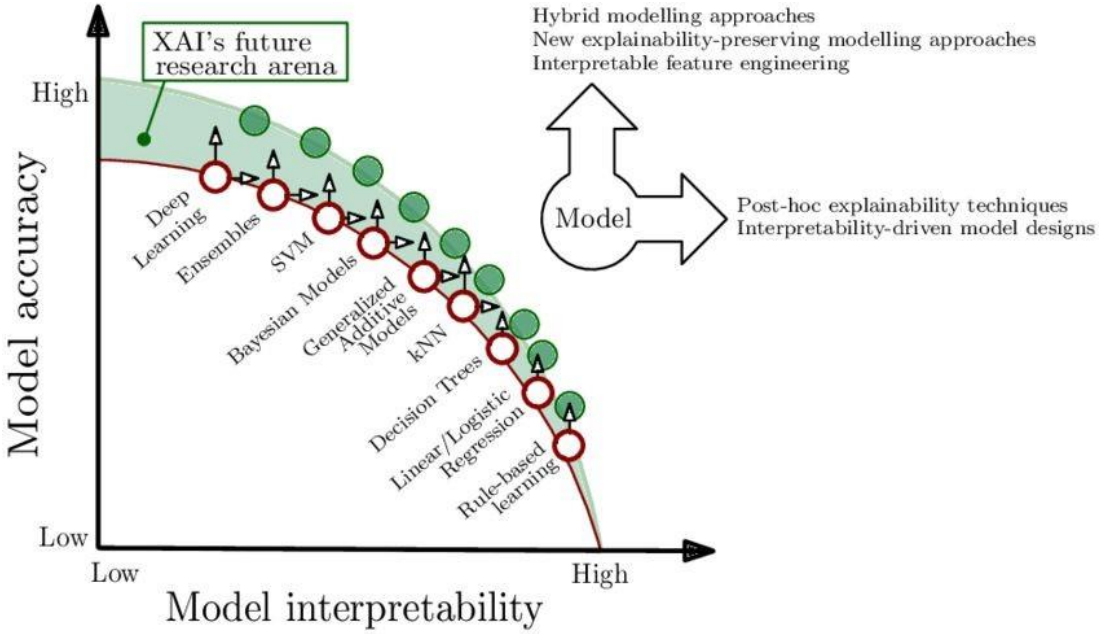


MNIST Dataset



http://yann.lecun.com/exdb/mnist/
https://soph.info/2018/11/08/mnist-history/

Interpretability and Performance Trade-offs



Arrieta et al., 2019