CAI 4104/6108 – Machine Learning Engineering:

Interpretable ML & Fairness

Prof. Vincent Bindschaedler

Spring 2024

Administrivia

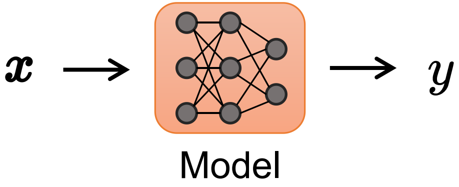
- Final Exam
  - When: May 2, 2024 — 7:30AM to 9:30AM
  - Where: Online (Canvas + **Honorlock**)
  - Note:
    - The CAI4104 and CAI6108 exams will be (slightly) different
  - Format:
    - Some Short answer questions (may include multiple choice)
    - Some multi-part problems

- **Sample Final Exam** (Practice Questions) — Live on Canvas
  - Please use it to prepare but do **not** overfit to it
  - It will close at 6:30am the day of the final (so there is no confusion)

Administrivia

- **Course Evaluation**
  - Help us improve the course: complete evaluation by **April 26**
  - Access the evaluation form:
    - Canvas: click on GatorEvals (left navigation panel)
    - or: https://ufl.bluera.com/ufl/
  - Optional and anonymous

Reminder: Interpretable ML



- Most ML models are black-box in terms of their decisions
  - You feed an input x, you get an output y
  - Why did we get output y (and not some $y' \neq y$)?

- We want human-understandable explanations
  - How?

Reminder: Taxonomy of Techniques

- Techniques for explaining processing
  - Explain how the model processed the data
  - “How did the model produce this output from this input?”
    - E.g.: proxy models, saliency mapping, etc.

- Techniques for explaining representations
  - Explain how the model represents information that influences the decision
  - “What information is represented by the model?”
    - E.g.: studying layers/neurons of a neural network

- Explanation-producing systems
  - We can design and use models themselves that produce simple interpretations of their behavior
    - E.g.: attention networks

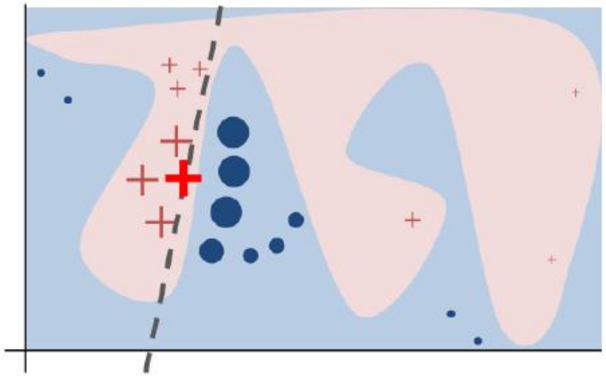
Reminder: Explaining Processing

- Proxy models techniques
  - Idea: create (train) a *proxy model* that behaves similarly but is easier to explain
    - Linear proxy models (e.g., LIME)
    - Decision trees
    - Rule extraction (e.g., if-then rules, MofN)

- Saliency mapping
  - Identify and highlight the salient features
    - Shows a small portion that is the most relevant
  - Typically applied to the image domain, but can be used more generally

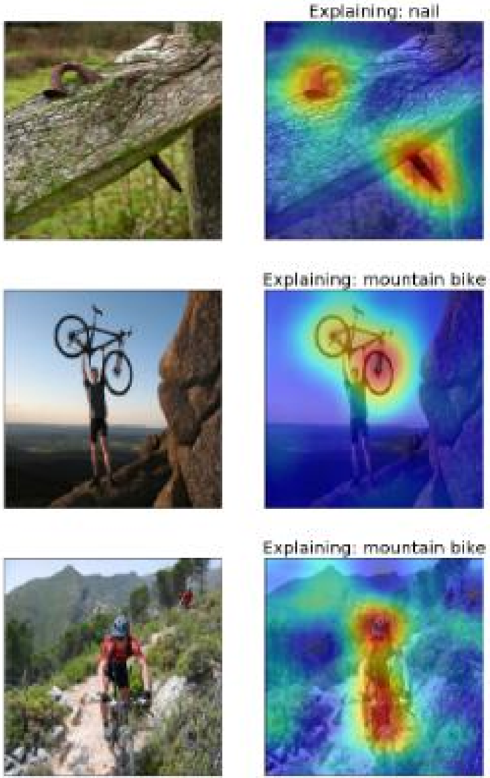
Reminder: LIME

- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "*Why should I trust you?: Explaining the predictions of any classifier*." KDD, 2016
  - Given an instance x
  - Idea: approximate the behavior of the model in a neighborhood of x using a **proxy model**
  - Choice for the proxy model:
    - Linear model, decision tree, falling rule list



- Decision function: blue-pink background
- Instance being explained: bold red cross
- Learned explanation: dashed line

Reminder: Saliency Maps



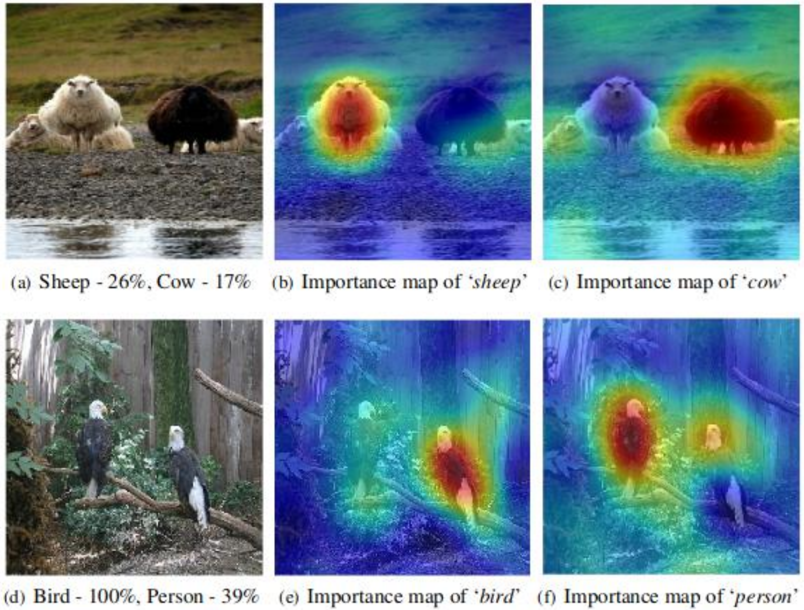


Figure 1: Our proposed RISE approach can explain why a black-box model (here, ResNet50) makes classification decisions by generating a pixel importance map for each decision (redder is more important). For the top image, it reveals that the model only recognizes the white sheep and confuses the black one with a cow; for the bottom image it confuses parts of birds with a person. (Images taken from the PASCAL VOC dataset.)

Source: Petsiuk, Das, and Saenko. "RISE: Randomized input sampling for explanation of black-box models." BMVC, 2018.

Explaining Representations

- Focus on neural networks
  - Internal representations are hard to grasp

- Layers-focused analysis
  - Pick a layer: what information does this layer contain?
  - Razavian et al. CVPR, 2014: internal layers of image classification networks can be used for other tasks!
    - E.g.: the feature vector can be reused directly for tasks such as classifying other species of birds
  - This is the insight used for transfer learning / use pre-trained models

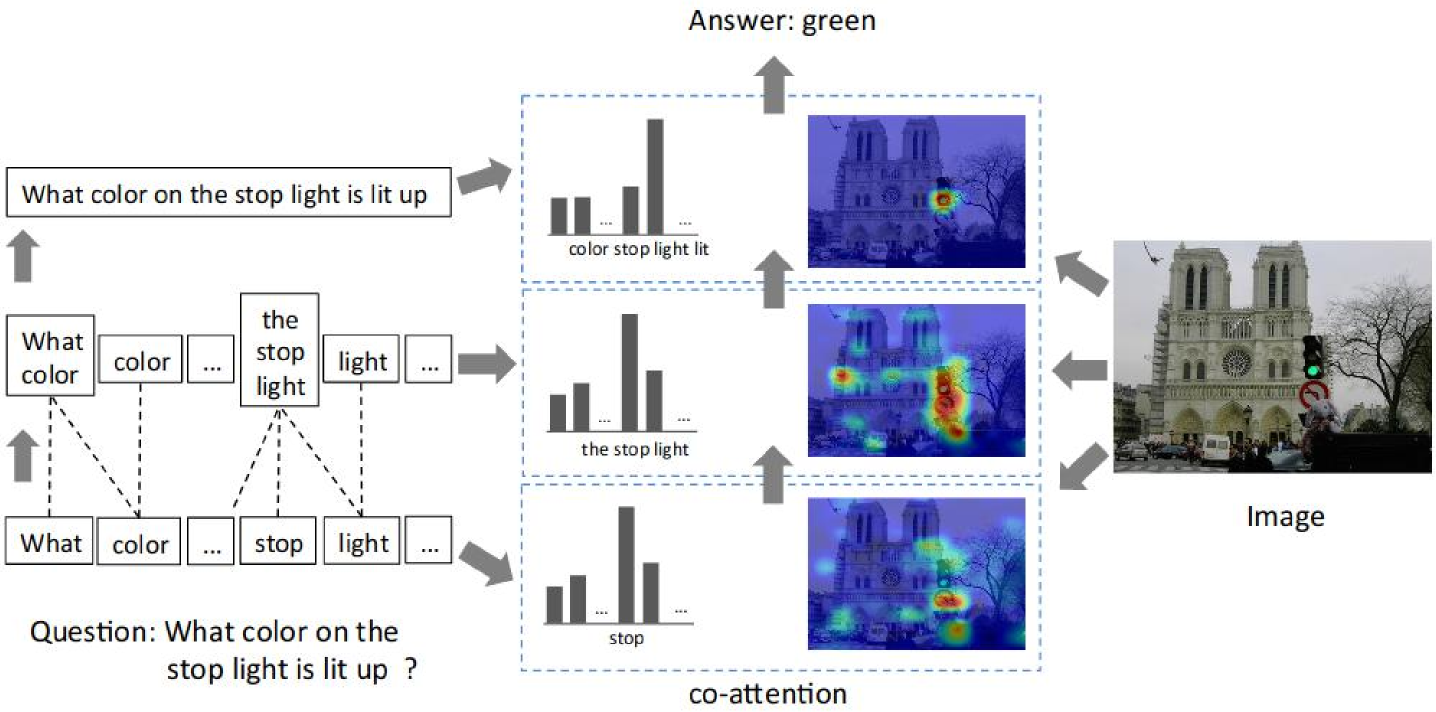
- Neurons-focused analysis
  - Pick a single neuron: what information is in this neuron?

Producing Explanations

- Attention Networks
  - Attention mechanisms allows the network to focus on a subset of features (or inputs)
  - Directly reveal information that passing through the neural net
    - This can serve as a form of explanation
- Disentangled Representations
  - Neural networks can learn entangled representations/latent factors
  - To provide explanations we may force the network to learn disentangled representation
    - E.g.: auto-encoders
- Explanation Synthesis
  - We can design neural nets so they produce explanations directly

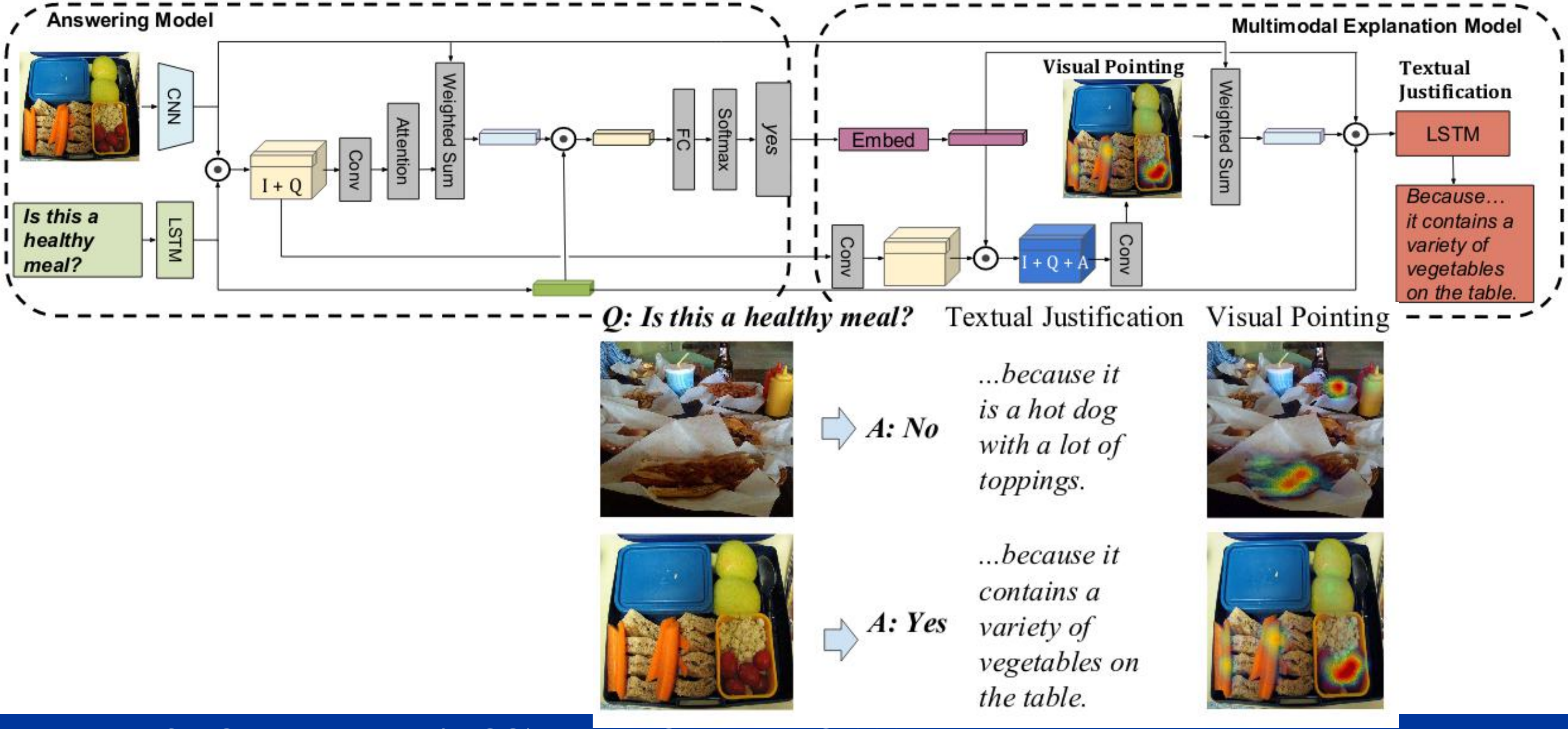
Attention Example

- E.g.: Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering." NIPS, 2016.



Explanation Synthesis

- E.g.: Huk Park et al. "Multimodal explanations: Justifying decisions and pointing to the evidence." CVPR, 2018.



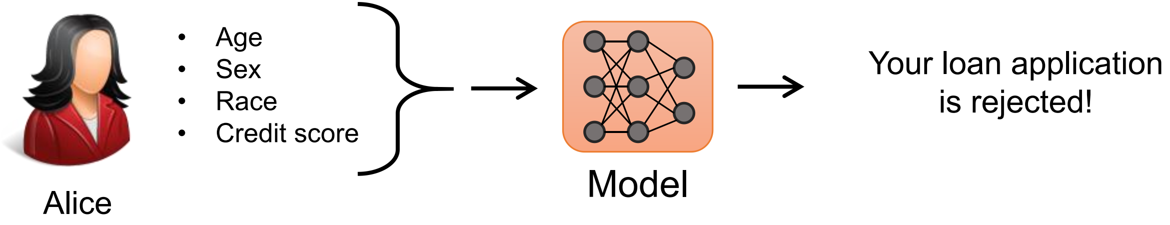
CAI 4104/6108 – Machine Learning Engineering:

Fairness

Prof. Vincent Bindschaedler

Spring 2024

Interpretable ML vs. Fairness



Interpretable ML

E.g.: why was Alice's loan denied? if her credit score was higher, would she have been approved?

Fairness

E.g.: Was Alice discriminated against?

What is Fairness?

- Definition (Merriam-Webster):
  - *the quality or state of being* fair
  - *especially :* fair *or impartial treatment : lack of favoritism toward one side or another*

- Definition (Cambridge Vocabulary):
  - *the quality of treating people equally or in a way that is right or reasonable*

- Definition (dictionary.com):
  - *the state, condition, or quality of being* fair*, or free from bias* or *injustice; evenhandedness*

Fairness & Discrimination in Law

- Discrimination (Merriam Webster)
  - 1: (a) prejudiced or prejudicial outlook, action, or treatment
    - // racial discrimination
  - 1: (b) the act, practice, or an instance of discriminating categorically rather than individually
  - 2: the quality or power of finely *distinguishing*

- (United Nation) Universal Declaration of Human Rights
  - Protected characteristic/group
    - race, color, sex, language, religion, political (or other opinion), national or social origin, property, birth (or other status)

- US Employment Discrimination Law
  - Protected class status (aka "*sensitive attribute*")
    - race, age (≥40), religion, sex (incl. pregnancy, sexual orientation, gender identity), disability, national origin, or genetic information (GINA — 2008)

Fairness & Discrimination in Law

- Is it ever legal to discriminate?
  - Yes, for example: *Hodgson v. Greyhound Lines, Inc.*, 499 F.2d 850 (7th Cir. 1974)
    - Greyhound Lines, Inc refused to hire bus driver applicants if they were over 35
      - Cited reason: passenger safety (statistically sound information)
    - Called a *Bona Fide Occupational Qualification*
  - Legal ≠ Moral

- Should we discriminate if statistical information justifies it?
  - Car rental for drivers less than 25 [drivers under 25 are riskier (statistically)]
    - You have to pay an additional Young Driver Fee / Under-25 Fee
    - Restrictions on the kind of car you can rent (no luxury cars or passenger vans)
  - Is it okay to discriminate in this case?

Discussion: What is Fair?

- Should we discriminate if statistical information justifies it?
  - Fairness in Car Rental vs. Likelihood of Accident
  - Let's look at some statistical information
    - Females cause fewer accidents that men (especially ≤ 30)
    - Should we discriminate based on this information?

Table 2
Single-car fatal and non-fatal crash counts by driver age, gender, and time of day in Great Britain, 2002–2012.

	Fatal crash counts			Nonfatal crash counts		
	Day	Evening	Night	Day	Evening	Night
Males						
17–20	16	10	44	2,081	938	2,251
21–29	35	16	66	4,101	1,347	2,616
30–39	40	16	35	4,702	1,218	1,637
40–49	28	10	19	4,251	948	1,112
50–59	17	7	12	3,086	591	578
60–69	12	2	5	1,682	286	234
70+	19	2	2	1,188	155	101
Females						
17–20	3	1	6	949	342	568
21–29	6	2	7	2,069	539	610
30–39	6	2	4	2,228	426	324
40–49	6	1	2	1,831	333	215
50–59	3	1	1	1,079	182	108
60–69	2	0	0	553	79	40
70+	6	0	0	474	50	22

Source: Regev et al. "Crash risk by driver age, gender, and time of day using a new exposure methodology." Journal of safety research 66, 2018.

- Car insurance
  - Factors that impact cost:
    - Credit score
    - Education level (get a PhD, save on car insurance!)
    - Marital status
  - Is that okay?

Sources of ML Unfairness

- Why worry about ML fairness?
  - Learning algorithms are objective; models are trained from data and don't have biases (like humans do)
- Models trained can be unfair. Why?
- Sources of unfairness in ML
  - Data
    - Training data could reflect some historical bias
    - Training data might not be representative of the population
  - Preprocessing
    - Data collection and cleaning
    - Feature engineering may aggregate, summarize, or select features that lead to unfairness
  - Model selection
    - Choice of model could be unfair
    - Model could have learned to “ignore” or make suboptimal predictions for some protected group

ML Fairness: Ingredients

- Metrics
  - Need metrics to quantify fairness
  - Tons of metrics have been proposed (e.g., statistical parity, predictive equality, rawlsian fairness, etc.)
    - *There is little to no consensus*
    - They are largely **incompatible** with each other

- Techniques
  - Detection: determine whether a model is fair
  - Fairness-aware:
    - Preprocessing — modify training data so any model trained on it will be fair
    - Algorithm modification — change an existing algorithm to make it fair no matter the training data
    - Postprocessing — take the output of any model (possibly an unfair one) and make it fair

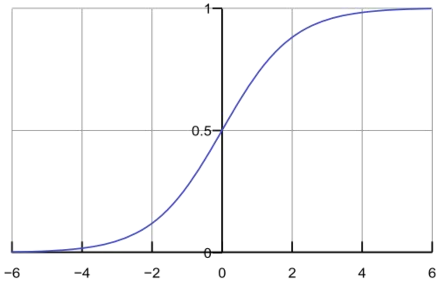
- *Note: ensuring fairness may have a negative impact on accuracy*

Fairness Notions & Metrics

- Many metrics have been proposed
  - anti-classification, statistical parity, conditional statistical parity, individual fairness, counterfactual fairness, rawlsian fairness, odds ratio, etc...
  - Categories
    - Individual fairness: e.g., similar individuals fairness
    - Group fairness: e.g., statistical parity, disparate impact, calibration, etc.
  - Most definitions are incompatible
    - Though some metrics are correlated
  - There is no concensus on a “right” definition

Thinking about Fairness

- Suppose we have a classifier to decide whether to approve a loan
  - Binary classification (labels $\in \{0, +1\}$) — loan denied (0), loan approved (1)
  - Features:
    - age, sex, race, credit score, yearly income
  - Logistic regression: $h_\theta(\boldsymbol{x}) = 1 / [1 + \exp\{-(\boldsymbol{w}\,\boldsymbol{x} + b)\}]$
    - $h_\theta(\boldsymbol{x}) = 1/(1+e^{-z})$ where $z = \boldsymbol{w}\,\boldsymbol{x} + b$



source: wikipedia

  - How do we know if it's fair?
    - Look at the weights/coefficients for protected attributes (age, sex, race) and see if they are ≠ 0

Fairness Notions & Metrics

- Setup
  - Binary classification (labels $\in \{0, 1\}$)
    - E.g.: loan denied (0), loan approved (1)
  - Classifier h trained on some dataset
  - Each data point $\boldsymbol{x}$ contains (unprotected) attributes $\boldsymbol{x}_u$ and protected attributes ($\boldsymbol{x}_p$)
    - Often we consider the single protected attribute binary case $\boldsymbol{x}_p$=0 vs $\boldsymbol{x}_p$= 1
  - Data points from the training set have an associated label y

- Notation
  - Prediction: $h(\boldsymbol{x}) = h((\boldsymbol{x}_u, \boldsymbol{x}_p)) = y'$

Fairness Notion: Anti-Classification

- Anti-classification (aka “Fairness through Unawareness”)
  - Idea: protected attributes are not explicitly used to make decisions (not part of the input)
    - E.g.: we can remove protected attributes (at training time or prediction time)
  - More formally — for all $\boldsymbol{x}$, $\boldsymbol{x}'$ such that $\boldsymbol{x}_u = \boldsymbol{x}_u'$: $h(\boldsymbol{x}) = h(\boldsymbol{x}')$
  - Does that solve the problem?
    - No because some (other) non-protected attribute could be used as a proxy
    - For example: geographic information (neighboorhood) could be used as a proxy for race (“redlining”)

- Should ML even use protected attributes? We could just avoid collecting the data
  - Example: classifier for loan applications — should sex be included? what about age?
  - If we can't collect them, we can't use them to enforce (or check) fairness

Fairness Notion: Statistical Parity

- Statistical Parity (aka "Demographic Parity")
  - Idea: proportion of positive predictions should be equal for each group
  - More formally — $\Pr\{y'=1 \mid x_p=0\} = \Pr\{y'=1 \mid x_p=1\}$

- Is this a good metric? What do you think?
  - Criticism: different groups could have very different base rates (rates of y=1)
    - E.g.: young drivers cause more car accidents

- We can also consider false positive parity instead
  - $\Pr\{y'=1 \mid y=0, x_p=0\} = \Pr\{y'=1 \mid y=0, x_p=1\}$
    - Idea: when false positives results in a high cost to the protected group, we want parity

Fairness Notion: Individual Fairness

- Individual Fairness (aka “Fairness through Awareness”)
  - Dwork et al. "*Fairness through awareness*." ITCS, 2012.
  - Idea: model should give similar predictions to similar individuals
  - More formally: given distance metric $d()$
    - If individuals are similar, i.e., $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is small, then predictions are similar: $h(\boldsymbol{x}_i) \approx h(\boldsymbol{x}_j)$

- Is this a good metric? What do you think?
  - Criticism: it depends on the distance metric $d()$ and how it behaves with respect to protected attributes

Incompatibility of Fairness Notions

- Researchers have proposed lots of fairness metrics
  - Quite a few may seem reasonable
  - Are they compatible?
    - Unfortunately: **No!**
  - For details, see:
    - Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv, 2016.
    - Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." arXiv, 2016.
    - Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." Big data 5, 2017.
    - Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in criminal justice risk assessments: The state of the art." Sociological Methods & Research, 2018.

Ensuring Fairness

- Suppose we decide on some notion of fairness
  - How do we ensure that our predictions conform to it?
- Fairness-aware techniques:
  - Preprocessing: modify training data so any model trained on it will be fair
  - Algorithm modification: change an existing algorithm to make it fair no matter the training data
  - Postprocessing: take the output of any model (possibly an unfair one) and make it fair

- Example: threshold adjustment
  - Postprocessing technique
  - Idea: adjust the threshold for the decision for each protected group (i.e., use a per-group threshold)
  - More formally — Let $x = (x_u, x_p)$ and consider binary classification
    - If $x_p = 0$ then $y' = 1$ if $h(x) \geq t_0$, 0 otherwise
    - If $x_p = 1$ then $y' = 1$ if $h(x) \geq t_1$, 0 otherwise $(t_1 \neq t_0)$
  - Is this fair?

Next Time

- Monday (4/22): Lecture
- Upcoming:
  - Project due 4/24