Topic 4: Trustworthy Vision

Some Factors Affecting Trust in Deep Learning

- **Models Complexity - Non-Decomposability into Simple Components**
  - Explainability
  - Interpretability
- **Social Discrimination and Data/Model Misrepresentations**
  - Disparate Treatment (e.g. Social Biases in Datasets)
  - Disparate Impact (e.g. Discriminative Outcomes)
- **Unreliable Inference even to Minor Input Disruptions**
  - Adversarial Examples

Impact of Stakeholders on Explainable AI (XAI)

How do diverse stakeholders perceive about neural networks?

- Decision Maker
  - Use predictions as recommendations to make appropriate judgements
  - e.g. doctors trying to diagnose patients
  - Cares about global explanations as well as local explanations
- Affected User
  - Analyze their inputs in retrospect to change the future outcome
  - e.g. patients
  - Cares only about local explanations
- Regulator
  - Ensures the model is safe and compliant with
  - e.g. government official trying to validate the model
  - Cares about both global explanations and local explanations
- Data Scientist
  - Improve model performance
  - e.g. some of you in the future!

Types of Explainable AI (XAI)

Local Explanations: Explain predictions for a given input data point

- Saliency Maps
- Class Activation Maps (CAM)
- Grad-CAM

Global Explanations: Explain the overall model

- ?

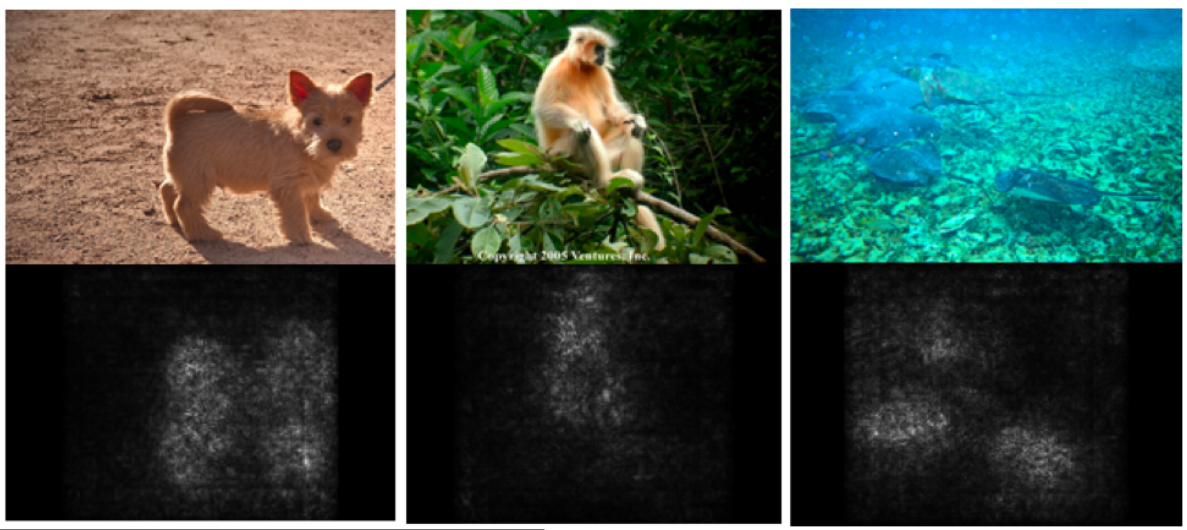
Saliency Maps[1]

- Consider input image I_0 of size $m \times n$, and a class c
- Highly non-linear class score function $S_c(I)$ in deep NNs $\Rightarrow$

  Approximate $S_c(I)$ with a linear function in the neighborbood of I_0 using Taylor's expansion:

$$S_c(I) \approx w^T I + b, \text{ where } w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0} \text{ can be found via backprop.}$$

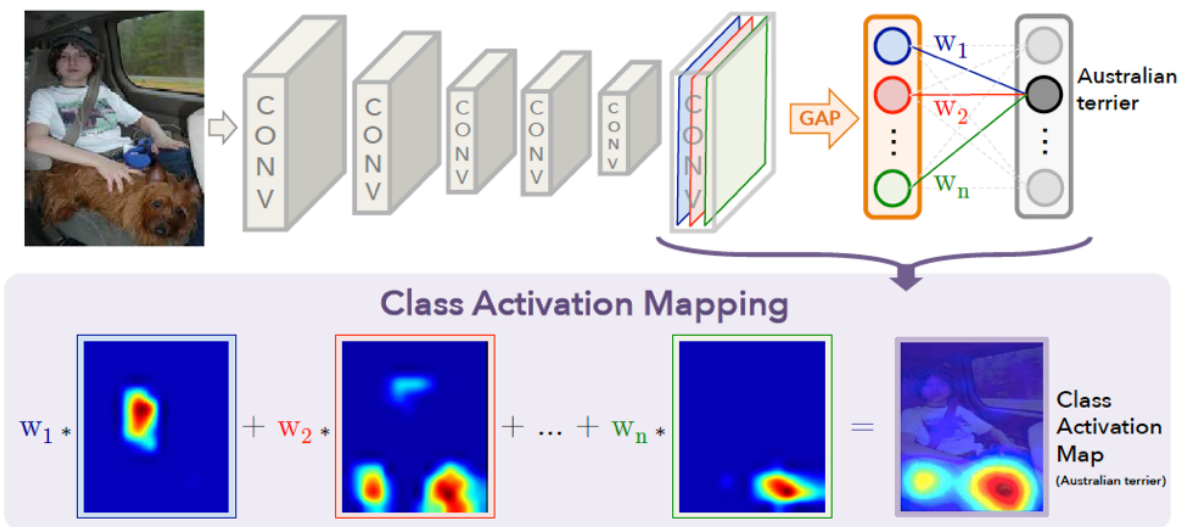
- Saliency Map $M_{i,j} = |w_{h(i,j)}|$, where $h(i,j)$ is the index in w that corresponds to $(i,j)^{th}$ pixel in I_0.
- Multi-channel images $\Rightarrow M_{i,j} = \max_c |w_{h(i,j,c)}|$
- Also, a regression problem to produce images that maximize a given class score



[1] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," ArXiv:1312.6034, 2013.

Class Activation Maps (CAM[2])

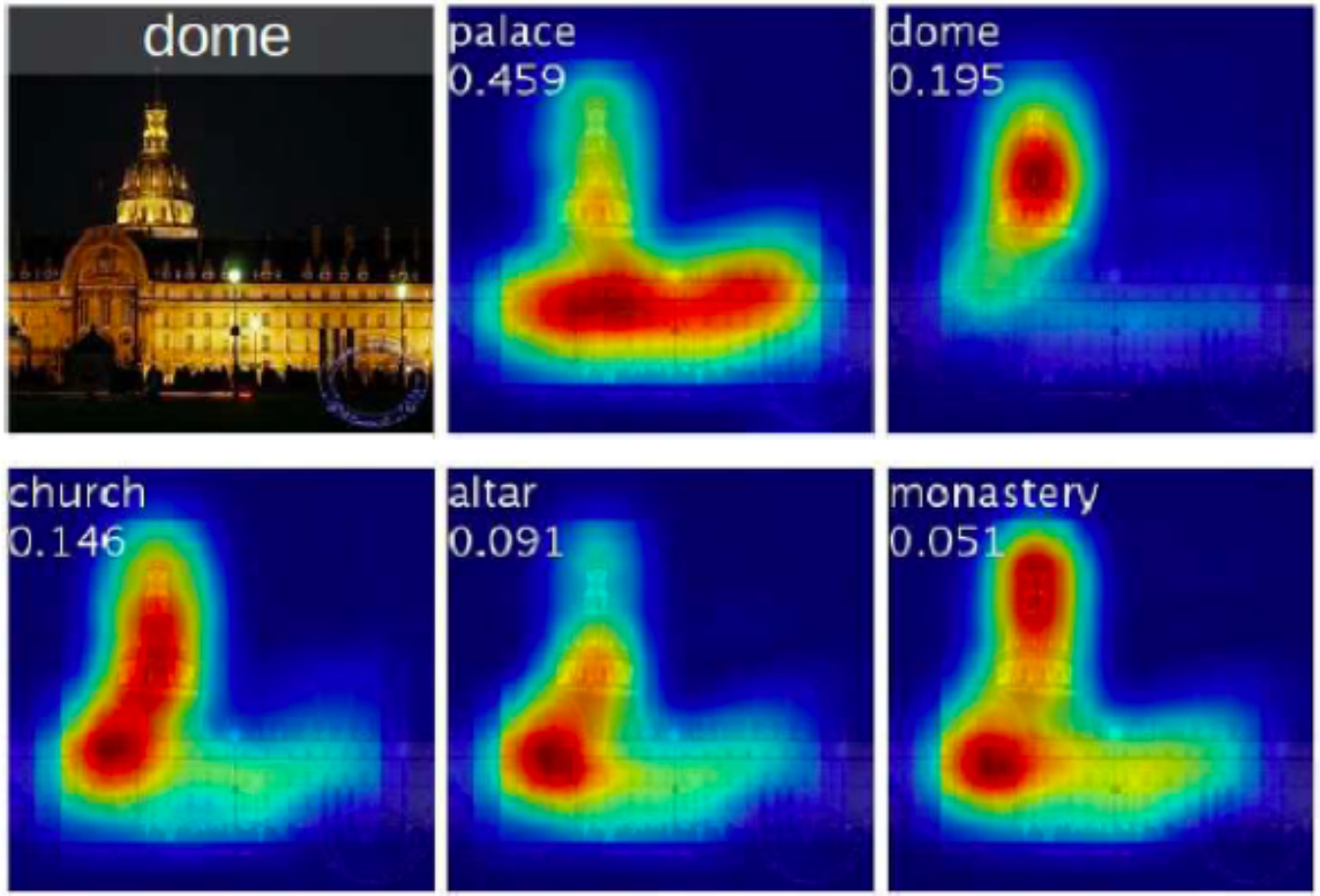
- Conv. layers are natural object detectors $\Rightarrow$ *Global average pooling* (GAP) instead of FC layers
- Let $f_k(x, y)$ denote activation of unit k at location (x, y).
- Result of GAP at unit k: $F_k = \frac{1}{Z} \sum_{x,y} f_k(x, y)$
- Class score: $S_c = \sum_k w_k^c F_k$ (ignore bias term) $\Rightarrow$ Softmax output: $\frac{\exp(S_c)}{\sum_c \exp(S_c)}$
- CAM: $M_c(x, y) = \sum_k w_k^c f_k(x, y) \Rightarrow S_c = \frac{1}{Z} \sum_{x,y} M_c(x, y)$
- Need to retrain the NN for weights w_k^c
- Upscale CAM to input size.



[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921-2929, 2016.

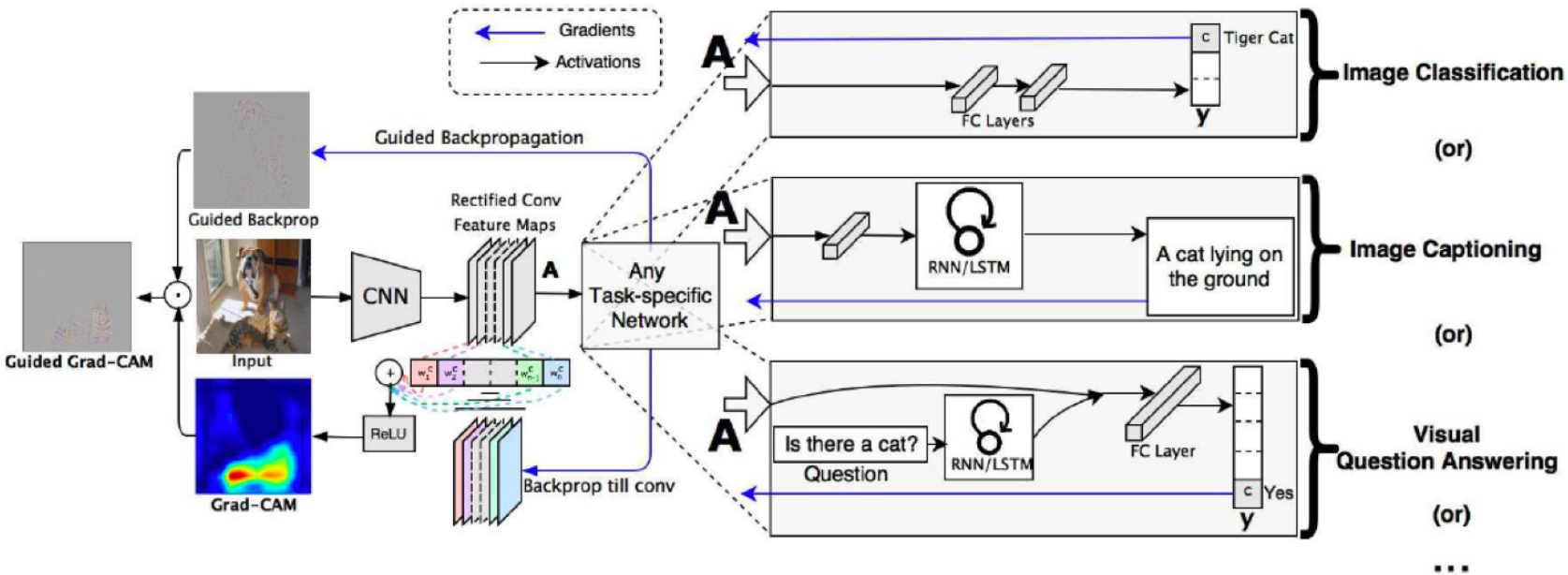
CAM (cont...)

- Example of CAMs generated from top-5 predicted categories
- Note that the *dome* class activates the upper round portion, while *palace* activates the lower flat portion of the compound.



Grad-CAM[3]

- Class score: $S_c = \sum_k w_k^c F_k$, where $F_k = \frac{1}{Z} \sum_{x,y} f_k(x,y)$
- CAM: $M_c(x,y) = \sum_k w_k^c f_k(x,y)$
- $$w_k^c = \frac{\partial S_c}{\partial F_k} = \frac{\partial S_c}{\partial f_k(x,y)} \left(\frac{\partial F_k}{\partial f_k(x,y)} \right)^{-1} = Z \cdot \frac{\partial S_c}{\partial f_k(x,y)}$$
- $$\sum_{x,y} w_k^c = Z \cdot \sum_{x,y} \frac{\partial S_c}{\partial f_k(x,y)} \quad \Rightarrow \quad w_k^c = \sum_{x,y} \frac{\partial S_c}{\partial f_k(x,y)}$$ (No need to retrain!)
- Grad-CAM: $\tilde{M}_c(x,y) = ReLU\left[M_c(x,y)\right]$

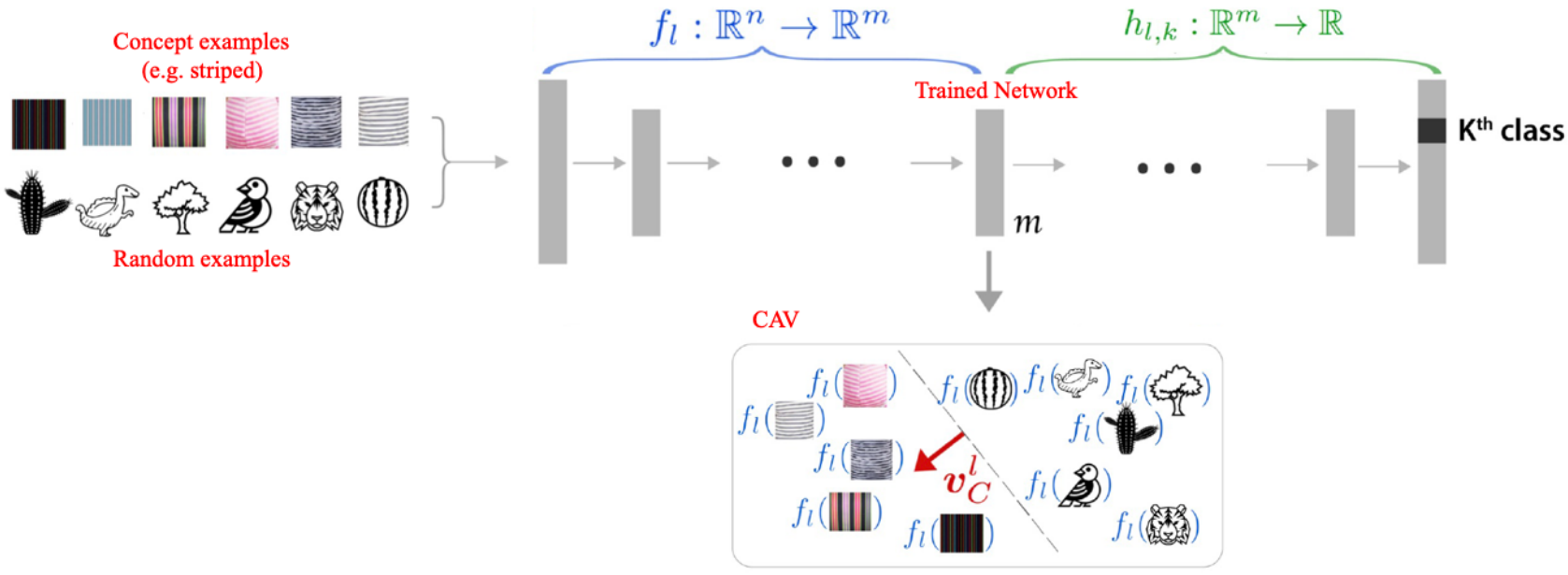


[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017.

Interpretable AI

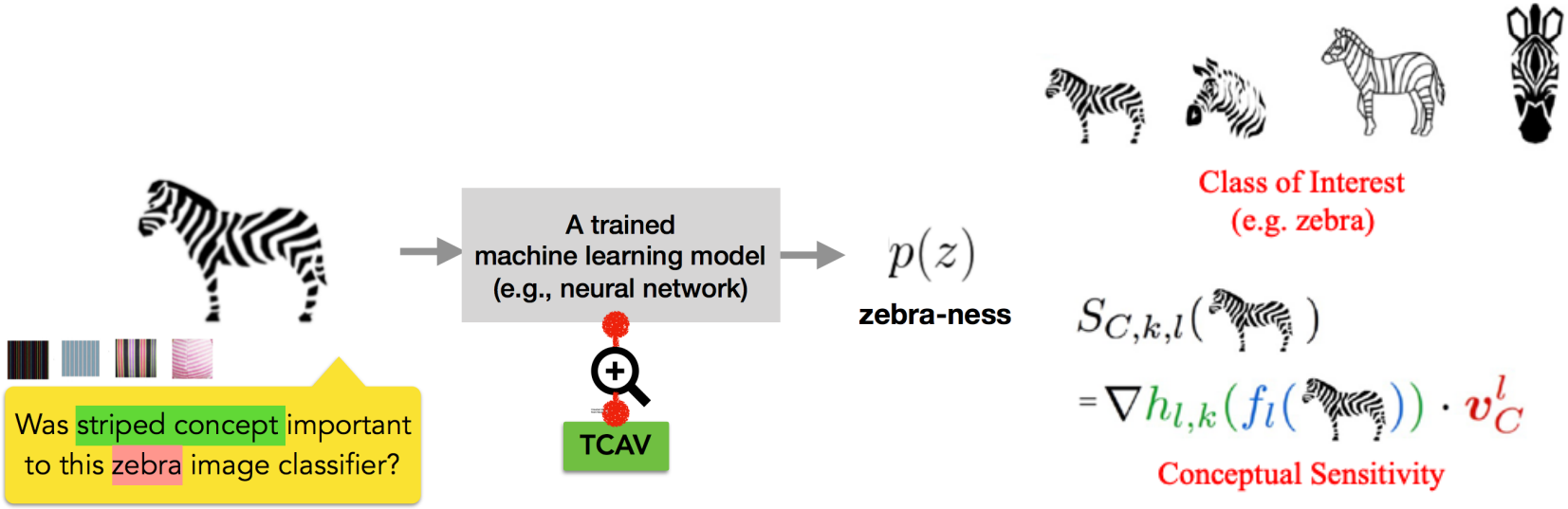
- Concept Activation Vectors (CAV)
- Uncertainty Quantification and Bayesian Neural Networks

Concept Activation Vectors (CAV)



- $f_l(x)$ takes input x and outputs layer l activations $a \in \mathbb{R}^M$.
- $h_{l,k}(a)$ takes layer l activation a and outputs the class-k logit $\in \mathbb{R}$.
- Given a user-defined concept C, let
  - P_C denote the set of images that positively represent the concept
  - N_C denote the set of images that negatively represent the concept
  - $A_P = \{f_l(x) | x \in P_C\}$, $A_N = \{f_l(x) | x \in N_C\}$
- Train a linear classifier to find a hyperplane with normal $v_C^l \in \mathbb{R}^M$ (CAV) that separates A_P and A_N.

Testing with CAV (TCAV[4])

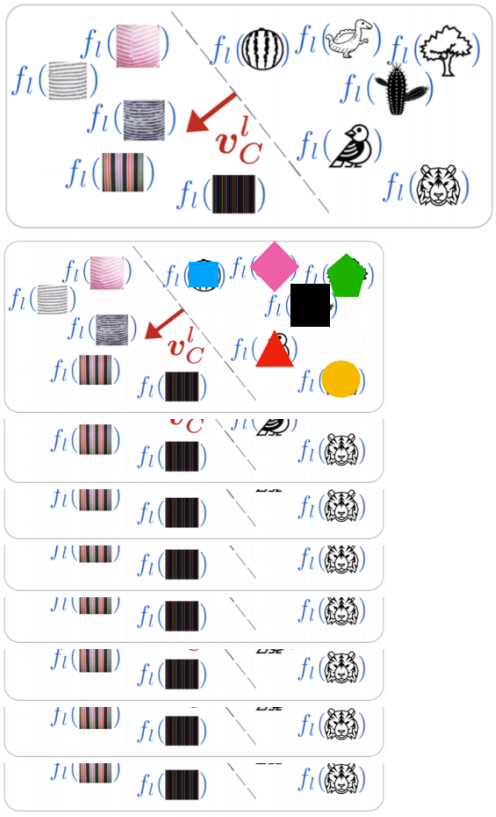


- **Conceptual Sensitivity:** A directional derivative $S_{C,k,l}(x)$ that measures the sensitivity of logit output to change in CAV.
- In saliency maps, we compute the gradient wrt input pixels instead.
- **TCAV:** Aggregate per-input conceptual sensitivity over a class k

$$TCAV_{C,k,l} = \frac{|\{x \in \mathcal{X}_k | S_{C,k,l}(x) > 0\}|}{|\mathcal{X}_k|}, \text{ where } \mathcal{X}_k \text{ denotes all inputs for class } k.$$

[4] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," In *International Conference on Machine Learning (ICML)*, pp. 2668-2677, 2018.

Statistical Significance of CAV



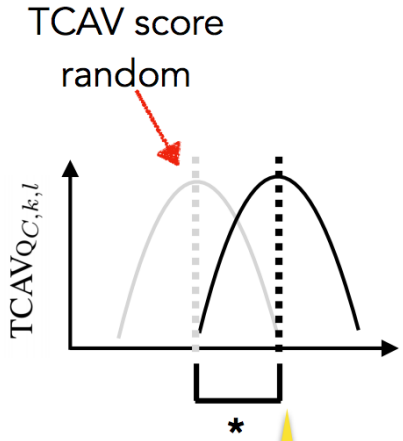
Zebra

$\rightarrow \text{TCAVQ}_{C,k,l}$:

$\rightarrow \text{TCAVQ}_{C,k,l}$:

$\rightarrow \text{TCAVQ}_{C,k,l}$:

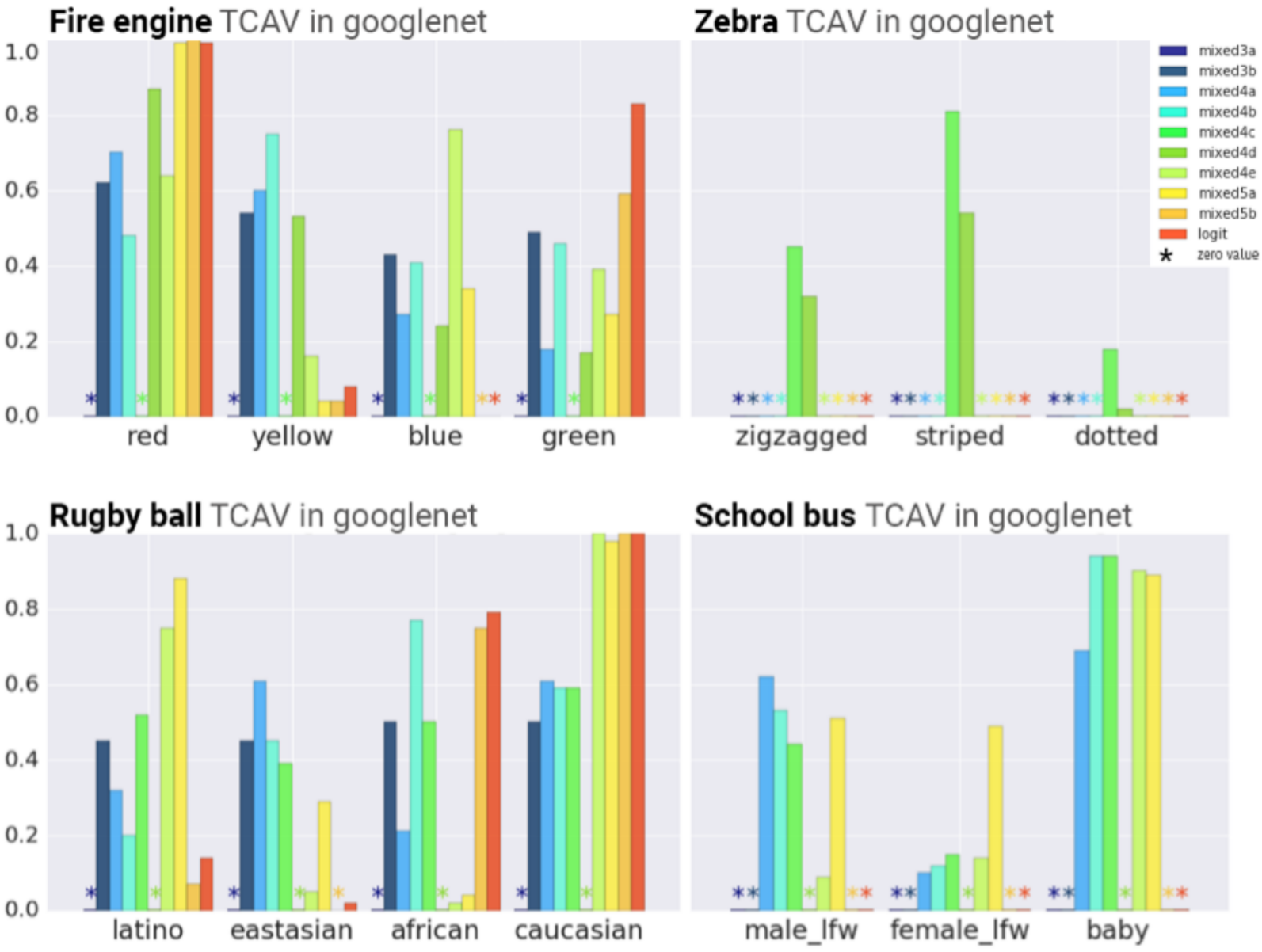
$\rightarrow \text{TCAVQ}_{C,k,l}$:



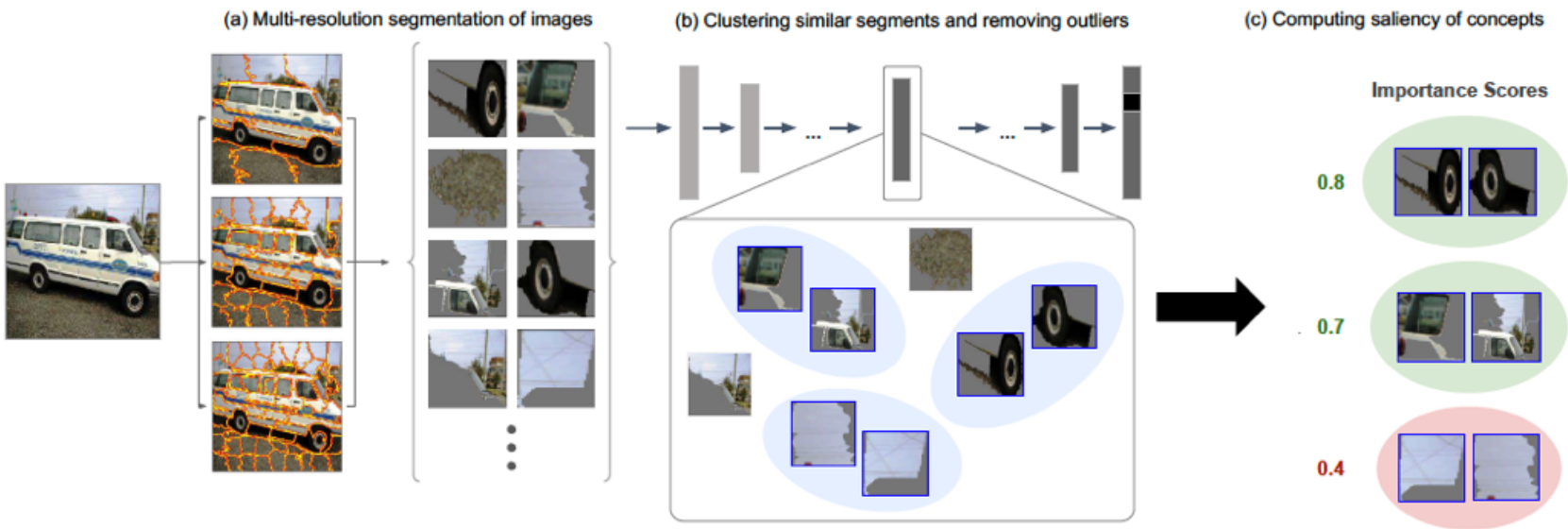
Check the distribution of $\text{TCAVQ}_{C,k,l}$ is statistically different from random using t-test

- Note: TCAV is very sensitive to low-quality random CAV.
- Compute TCAVs T times using different N_C sets to obtain $\{TCAV_{C,k,l}^{(i)}\}_{i=1}^{T}$
- Perform two-sided t-test.

Example: TCAV on GoogLeNet



Automatic Concept-Based Explanations (ACE[5])



Desired Properties of Concept-Based Explanation:

- ***Meaningfulness:*** Examples need to be semantically meaningful on its own. Also, multiple individuals should associate similar meaning to the same concept. (e.g. a group of pixels that contains a specific texture/object)
- ***Coherency:*** xamples need to be perceptually similar to each other, but also different from examples of other concepts.
- ***Importance:*** The concept's presence is necessary for the true prediction

[5] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. "Towards Automatic Concept-Based Explanations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

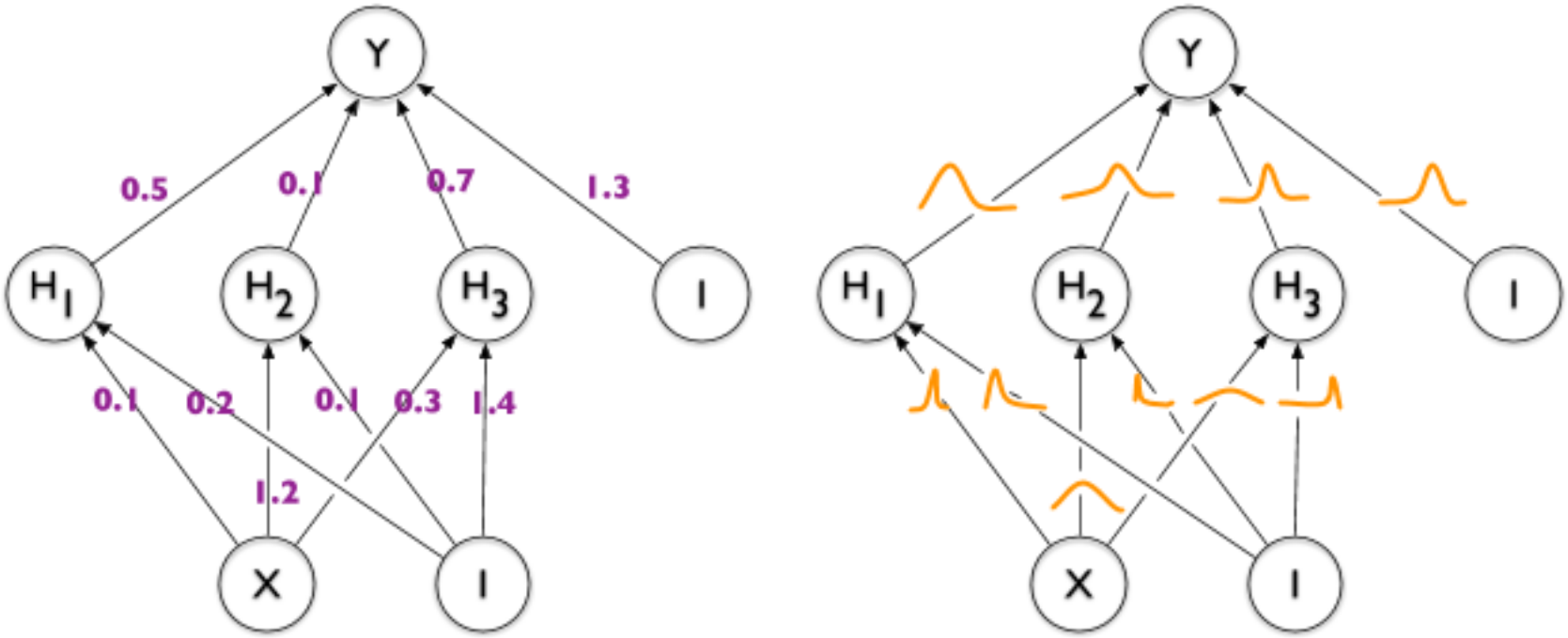
Uncertainty Quantification

Two types of uncertainty:

- **Aleatoric Uncertainty:** Confidence in input data
  - High when input data is noisy
  - Cannot be reduced by adding more data
  - Can be estimated using likelihood methods using neural networks
- **Epistemic Uncertainty:** Confidence in Prediction
  - High when training data is small
  - Can be reduced by adding more data
  - Very difficult to estimate (Knowing when the model does not know the answer)

Solution to Epistemic Uncertainty: Bayesian Neural Networks

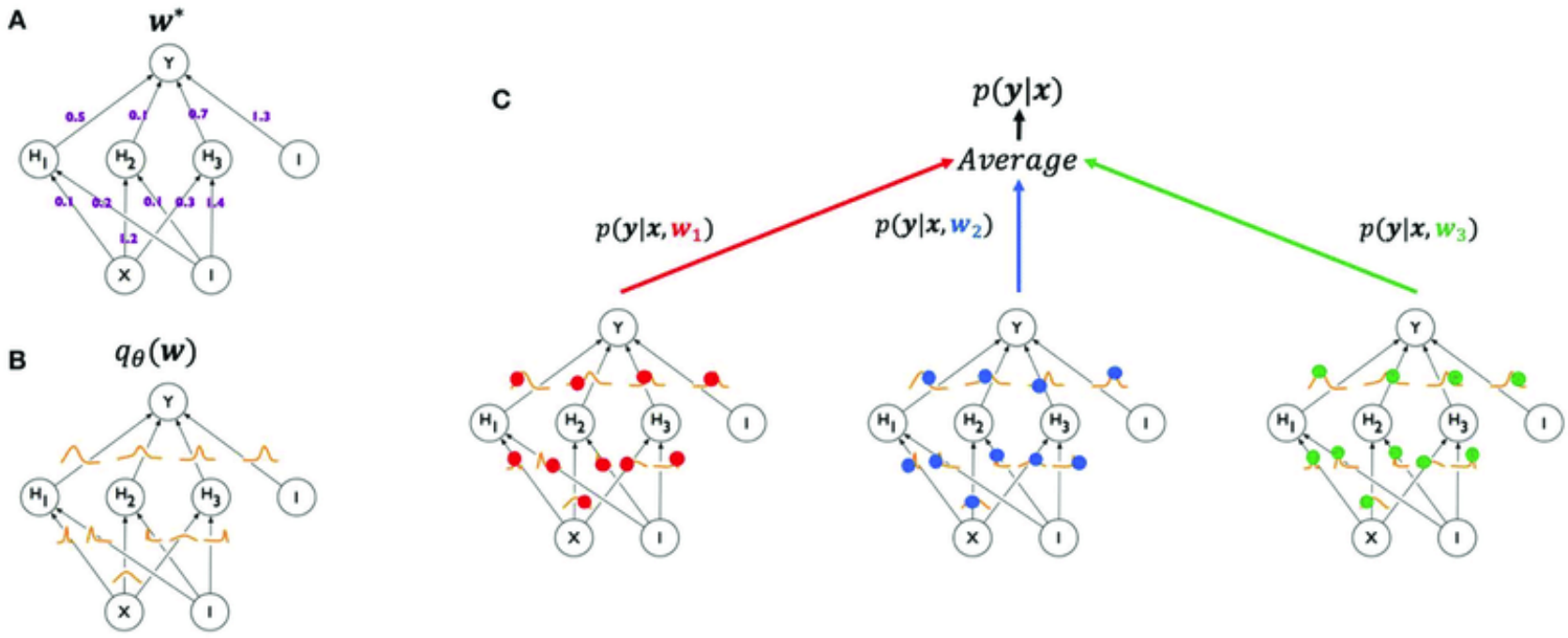
Bayesian Neural Networks (BNNs)



- Train weight distributions, as opposed to just weights as in traditional NNs.
- Assume a prior distribution for weights $p(\mathbb{W})$, and a dataset $(\boldsymbol{X}, \boldsymbol{Y})$.
- Use Bayes' rule to update weight distribution via computing its posterior:

$$p(\mathbb{W}|\boldsymbol{X}, \boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \mathbb{W}) \cdot p(\mathbb{W})}{p(\boldsymbol{Y}|\boldsymbol{X})}$$

Emulating BNNs through Monte-Carlo Sampling[7]



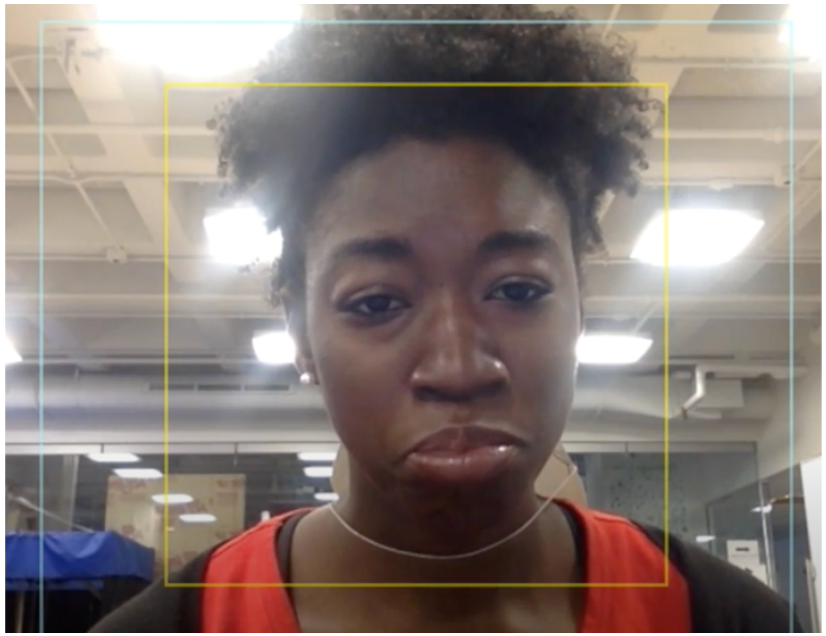
- Sample weights from the trained distribution of weights several times
- Compute the average logit probability at the output of each class
- Similar approaches: Use Dropout[6] in Testing Phase to capture epistemic uncertainty

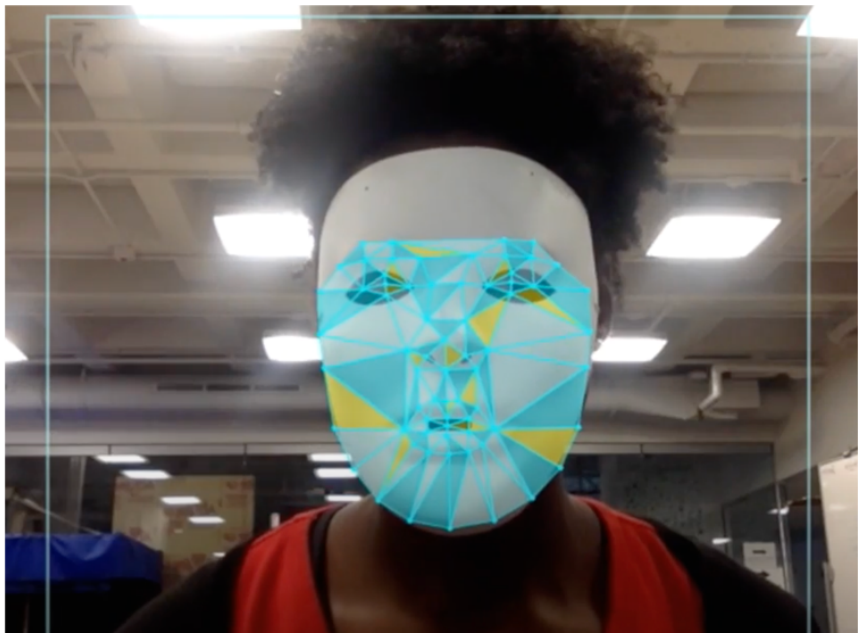
---

[6] Y. Gal, and Z. Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," In *International Conference on Machine Learning (ICML)*, pp. 1050-1059, PMLR, 2016.

[7] B. Lakshminarayanan, A. Pritzel, C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances of Neural Information Processing Systems (NeurIPS)*, 2017.

ML Meets Social Sciences: Social Discrimination in CV[8]

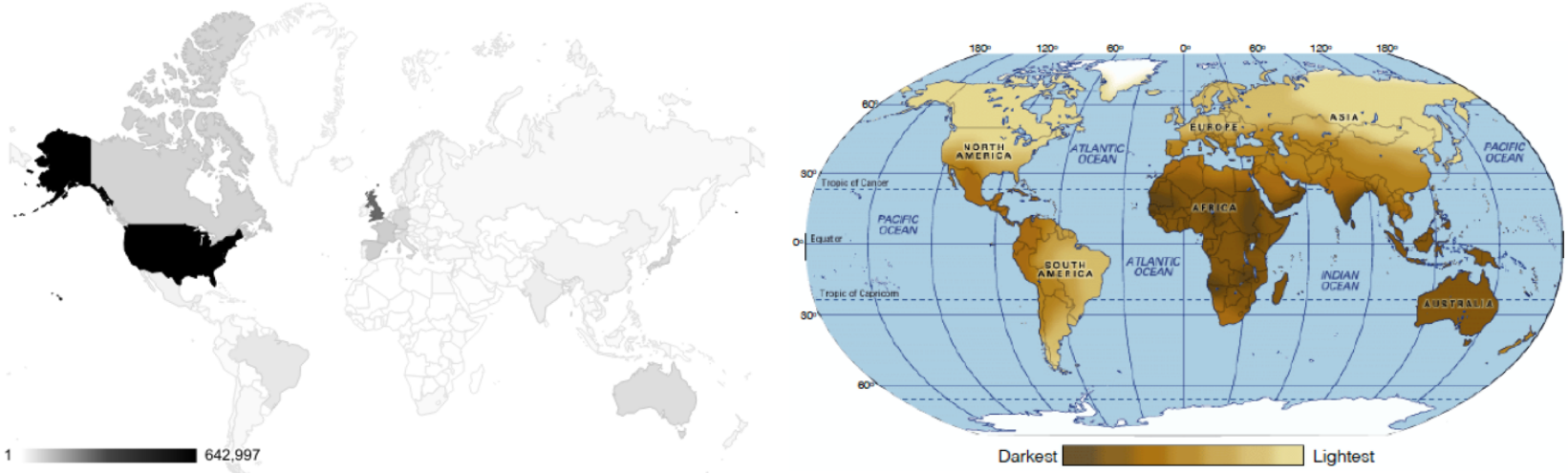




- ML Designer's Perspective: Represent the data accurately
- Affected User's Perspective: **Is this algorithm fair? Am I (or, my group) being affected negatively by this solution?**
- Social biases observed in hindsight, after deployment
- Significant legal concern...
- Who is accountable for such flaws?

[8] J. Buolamwini, "The Coded Gaze: Unmasking Algorithmic Bias," YouTube: `https://www.youtube.com/watch?v=162VzSzzoPs`

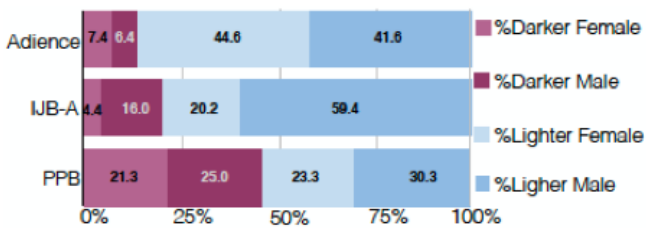
Racial Discrimination in Commercial Gender Classification[9]



Data Distribution vs. Skin Color Distribution

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Property	PPB	IJB-A	Adience
Release Year	2017	2015	2014
#Subjects	1270	500	2284
Avg. IPD	63 pixels	-	-
BBox Size	141 (avg)	≥36	-
IM Width	160-590	-	816
IM Height	213-886	-	816



[9] J. Buolamwini, and T. Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," In *Conference on Fairness, Accountability and Transparency (FAT)*, pp. 77-91, PMLR, 2018.

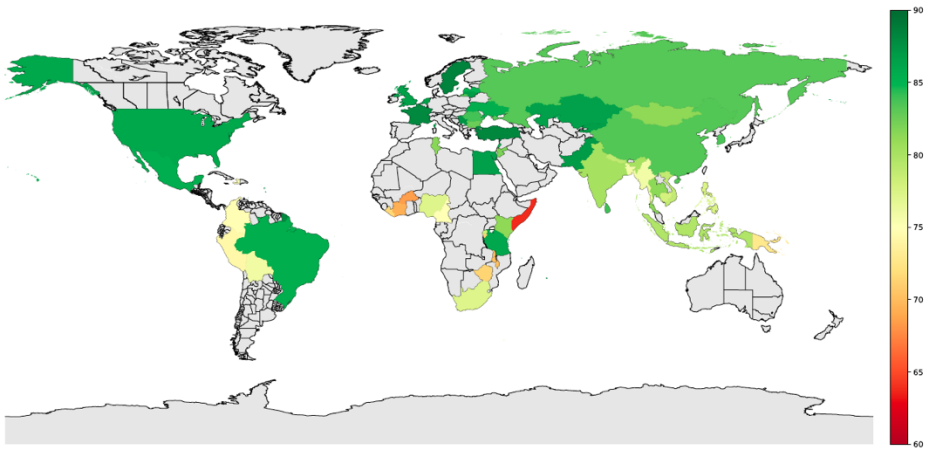
Income-Based Discrimination in Object Detection[10]

Ground truth: Soap **Nepal, 288 $/month**

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave



[10] T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, "Does Object Recognition Work for Everyone?," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 52-59. 2019.

Biases in Pedestrian Detection

Racial Discrimination in Pedestrian Detection[11]

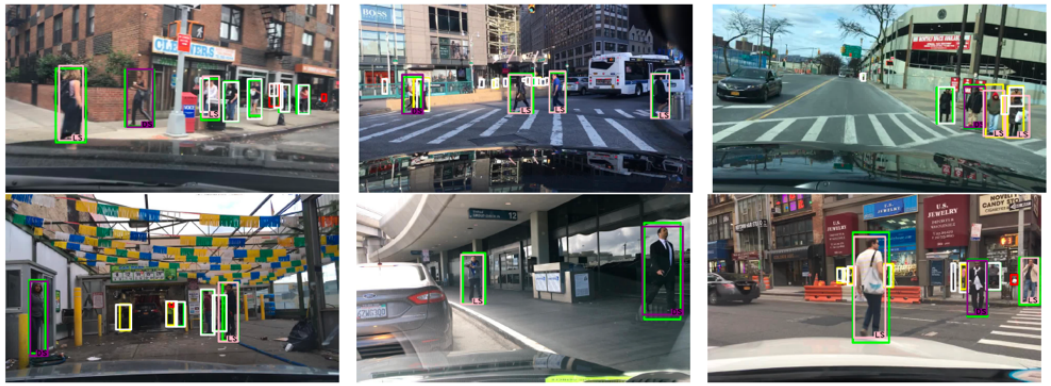
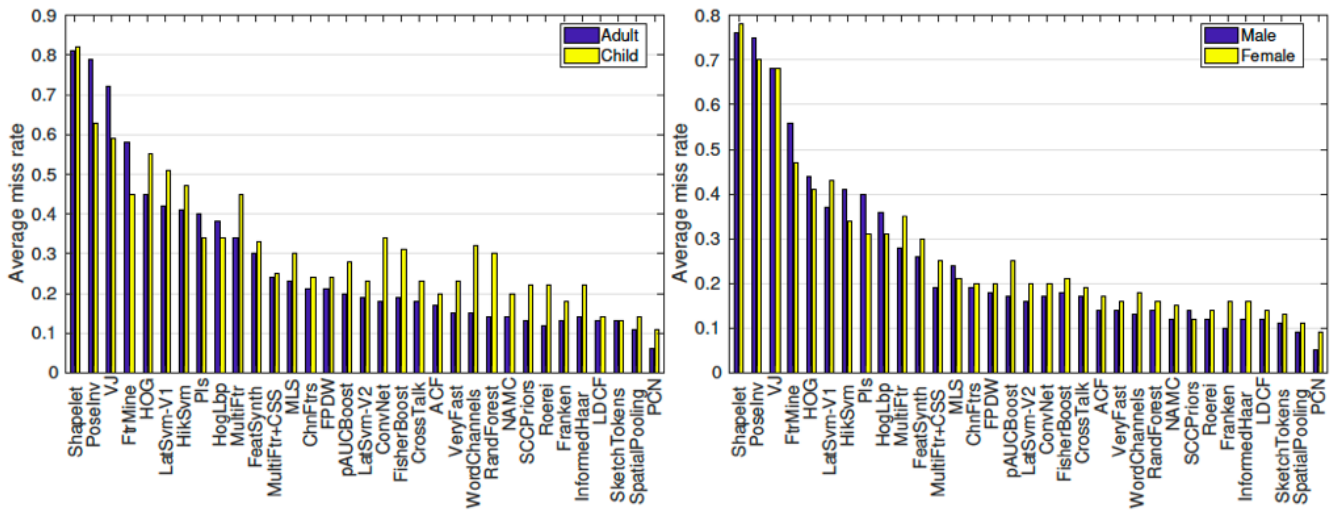


Figure 5. Example detections from Faster R-CNN using the R-50-FPN backbone, trained on BDD100K. For reference, the ground truth annotations for LS and DS are pink and purple respectively. Yellow boxes correspond to true positives under the AP_{50} metric and false positives under the AP_{75} metric. Green boxes correspond to true positives under the AP_{75} metric. All the predictions shown are greater than an 85% confidence threshold.

Model	Backbone	AP (%)		AP_{50} (%)		AP_{75} (%)	
		LS	DS	LS	DS	LS	DS
Faster R-CNN	R-50-C4	**62.0**	57.1	93.5	**94.1**	**72.9**	60.7
Faster R-CNN	R-50-FPN	**61.2**	58.0	93.5	**95.8**	**70.9**	59.4
Faster R-CNN	R-101-FPN	**64.2**	60.8	**95.7**	95.3	**77.0**	63.2
Faster R-CNN	X-101-32x8d-FPN	**64.7**	59.4	**95.7**	92.8	**75.9**	68.9
Mask R-CNN	R-50-C4	**63.4**	58.2	94.0	**95.0**	**74.1**	58.7
Mask R-CNN	R-50-FPN	**63.2**	56.3	**95.3**	94.7	**75.6**	56.1
Mask R-CNN	R-101-FPN	**64.3**	58.5	95.3	**95.5**	**73.5**	59.1
Mask R-CNN	X-101-32x8d-FPN	**64.9**	61.4	**96.1**	95.0	**76.6**	70.5
Average		**63.5**	58.7	**94.9**	94.8	**74.6**	62.1

Table 5. Average precision on BDD100K validation set with occluded individuals removed for models trained using MS COCO.

Age/Gender Based Discrimination in Pedestrian Detection[12]

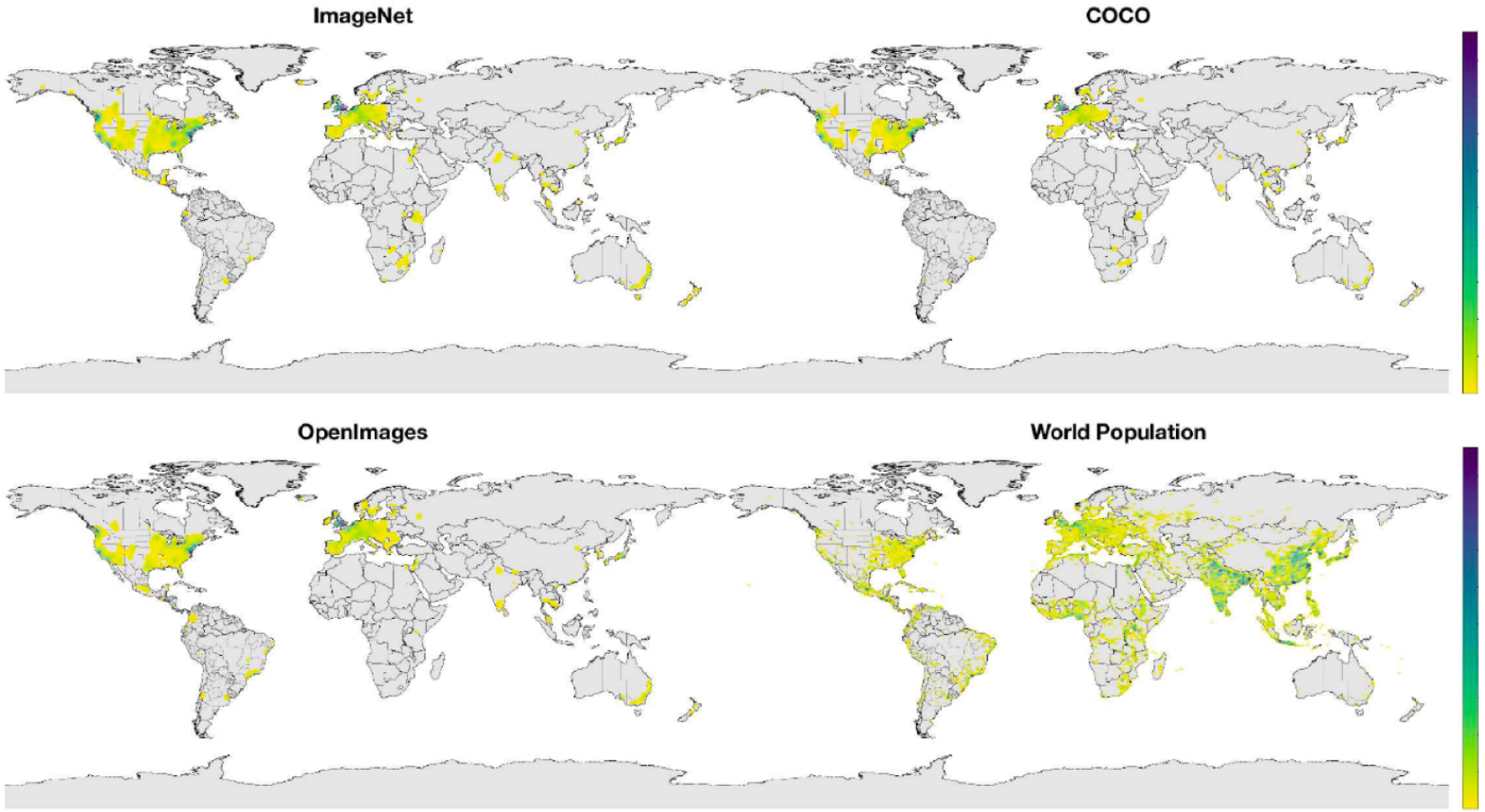


[11] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive Inequity in Object Detection," ArXiv:1902.11097, 2019.

[12] M. Brandao, "Age and Gender Bias in Pedestrian Detection Algorithms," in *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision (FATE/CV) in CVPR-2019, Available: ArXiv: 1906.10490.*

What Causes Social Discrimination? [13]

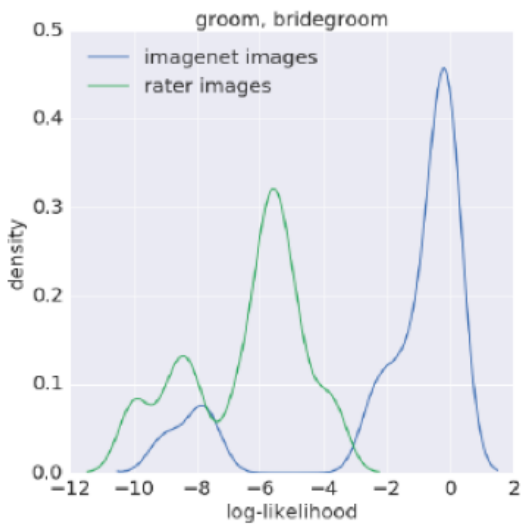
Reason 1: Imbalanced Data Collection

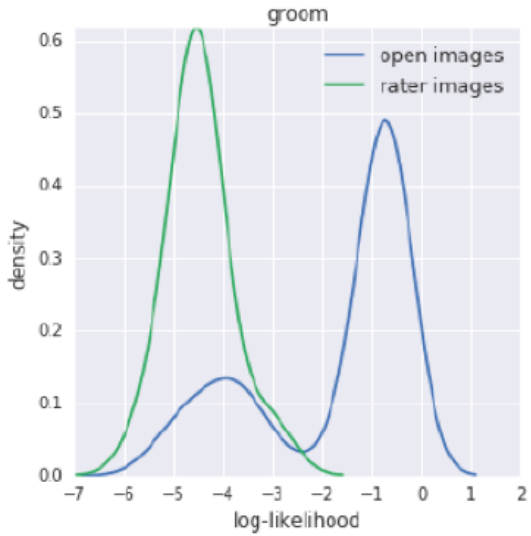


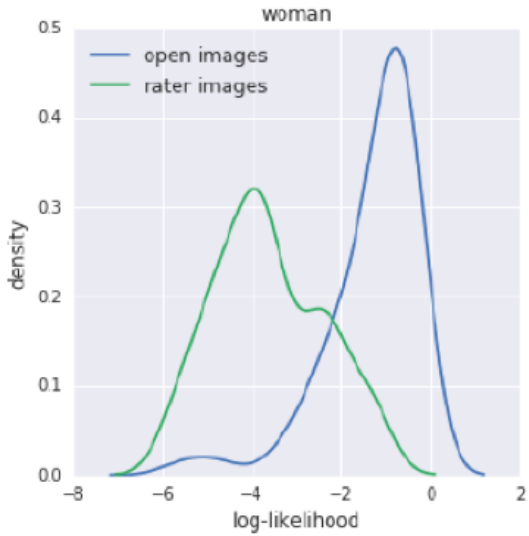
[13] T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, "Does Object Recognition Work for Everyone?," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 52-59. 2019.

What Causes Social Discrimination? [14]

Reason 2: Biased Selection of Crowd Labelers







[14] T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, "Does Object Recognition Work for Everyone?," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 52-59. 2019.

Evaluating Discrimination using Fairness Notions[15]

Sources: Systemic vs. Statistical

- **Systemic Discrimination:** Preference towards affected users who are similar to the decision maker. (e.g. job hiring)
- **Statistical Discrimination:** Using average group statistics to judge an individual (e.g. kidney matching)

Types of Discrimination: Direct vs. Indirect

- **Direct Discrimination:** Use of protected attributes of individuals to explicitly result in non-favorable outcomes (e.g. Admission decisions based on gender)
- **Indirect Discrimination:** Although seemingly neutral, protected groups still get to be treated unjustly due to implicit biases (e.g. gentrification and gerrymandering)

Types of Discrimination: Treatment vs. Impact

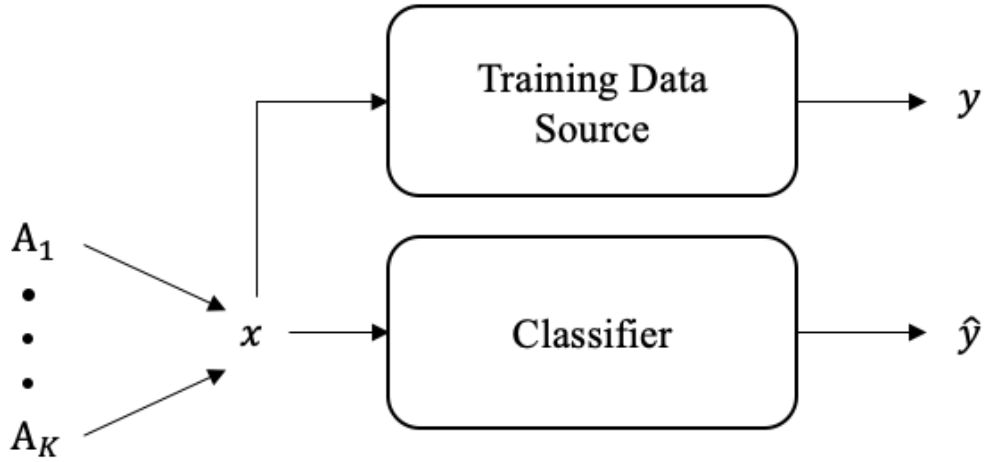
- **Disparate Impact** $\Rightarrow$ *Obtain different impact at outcomes for different groups*
  - Group Fairness
- **Disparate Treatment** $\Rightarrow$ *Treat individuals differently*
  - Individual Fairness

[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1-35, 2021.

Group Fairness[16]

Setting:

- Data tuple: (X, Y)
- Groups: $A_1, \cdots, A_K$
- System Prediction: $\hat{Y}$



Notion	Intuition	Formal Definition
Statistical Parity	$\hat{Y} \perp A$	$\mathbb{P}(\hat{Y}\|A_i) = \mathbb{P}(\hat{Y}\|A_j),\ \forall\, i \neq j.$
Equalized Odds	$\hat{Y} \perp A \mid Y$	$\mathbb{P}(\hat{Y}\|A_i, Y=y) = \mathbb{P}(\hat{Y}\|A_j, Y=y),\ \forall\, i \neq j, y \in \mathcal{Y}.$
Equal Opportunity	$\hat{Y} \perp A \mid Y = y$	$\mathbb{P}(\hat{Y}\|A_i, Y=y) = \mathbb{P}(\hat{Y}\|A_j, Y=y),\ \forall\, i \neq j.$
Calibration	$Y \perp A \mid \hat{Y}$	$\mathbb{P}(Y\|A_i, \hat{Y}=y) = \mathbb{P}(Y\|A_j, \hat{Y}=y),\ \forall\, i \neq j, y \in \mathcal{Y}.$
Predictive Parity	$Y \perp A \mid \hat{Y} = y$	$\mathbb{P}(Y\|A_i, \hat{Y}=y) = \mathbb{P}(Y\|A_j, \hat{Y}=y),\ \forall\, i \neq j.$

[16] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

Fundamental Tradeoff[17,18] Amongst Group Fairness Notions

True Label / Decision	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Claim 1

Given a prevalence (base rates) p, positive predictive value PPV, false positive and false negative rates (FPR, FNR), we have

$$FPR = \frac{p}{1-p} \cdot \frac{1-PPV}{PPV}(1-FNR)$$

- Assume prevalence p is non-identical across groups. Let PPV is same across groups.
- Then, FPR and FNR cannot be equal across groups.
- Alternatively, need identical p, or perfect prediction.

[17] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big data*, vol. 5, no. 2, pp. 153-163, 2017.

[18] J. Kleinberg, M. Raghavan, and S. Mullainathan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS)*, 2017.

Individual Fairness

- **Fairness through Unawareness:** No protected attributes are explicitly used in the decision-making process

- **Fairness through Awareness**[19]**:** Similar individuals w.r.t. a similarity metric (satisfies non-negativity and symmetry) defined for a particular task should receive a similar outcome.

  Formally, a stochastic classifier $C : \mathcal{X} \to \Delta(\mathcal{Y})$ is IF if it is (D, d)-Lipschitz, i.e.,

  $$D\Big(C(x_1), C(x_2)\Big) \leq d(x_1, x_2), \text{ for all } x_1, x_2 \in \mathcal{X}.$$

- **Probably Approximate-Metric Fairness**[20]**:**

  $$\mathbb{P}_{x_1, x_2 \sim \mathcal{P}}\left[D\Big(C(x_1), C(x_2)\Big) \leq d(x_1, x_2) + \delta\right] \leq \epsilon.$$

- **Counterfactual Fairness**[21]**:** A decision is fair if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group. In other words,

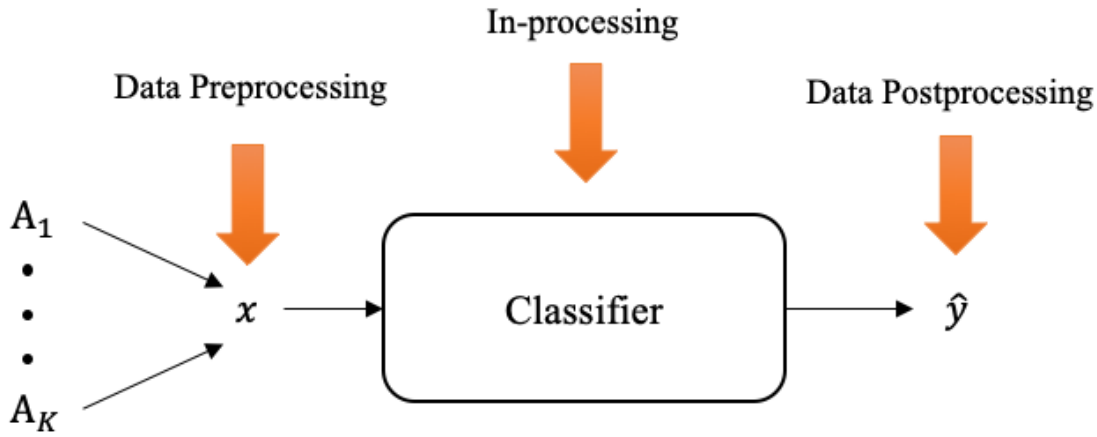
  $$\mathbb{P}(\hat{Y}_{\mathcal{A} \leftarrow \mathcal{A}_i} | X, \mathcal{A}) \; = \; \mathbb{P}(\hat{Y}_{\mathcal{A} \leftarrow \mathcal{A}_j} | X, \mathcal{A}), \text{ for all } i \neq j.$$

[19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. "Fairness Through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214-226. 2012.

[20] G. Yona, and G. Rothblum, "Probably Approximately Metric-Fair Learning." in *International Conference on Machine Learning (ICML)*, pp. 5680-5688, PMLR, 2018.

[21] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

Mitigating Social Discrimination[24]: Preprocessing



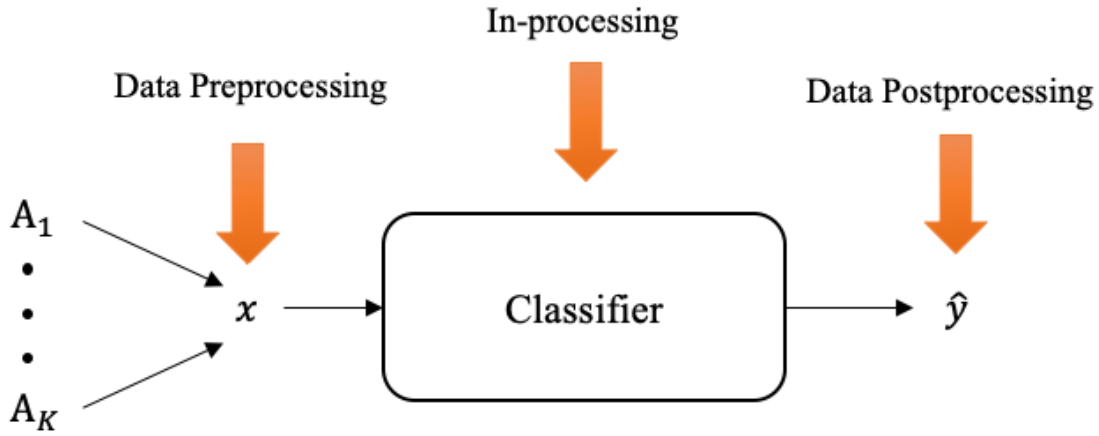
- Blinding
- Sampling: Sub-Group Analysis (Fairness under Composition does not do well[22]!)
- Debiasing: Removal of Causal Dependencies, Transformation[23], Label Perturbation
- Reweighting

[22] C. Dwork, and C. Ilvento, "Fairness Under Composition," ArXiv Preprint:1806.06122, 2018.

[23] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. "Optimized Pre-Processing for Discrimination Prevention," in *Advances in neural information processing systems (NeurIPS)*, vol. 30, 2017.

[24] S. Caton, and C. Haas, "Fairness in Machine Learning: A Survey," ArXiv Preprint:2010.04053, 2020.

Mitigating Social Discrimination[29]: In-Processing



- Constrained Optimization and/or Regularization[25,26]
- Adversarial Learning[27]
- Bandits[28] (for reinforcement learning)

---

[25] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. "A Reductions Approach to Fair Classification." In *International Conference on Machine Learning (ICML)*, pp. 60-69, PMLR, 2018.

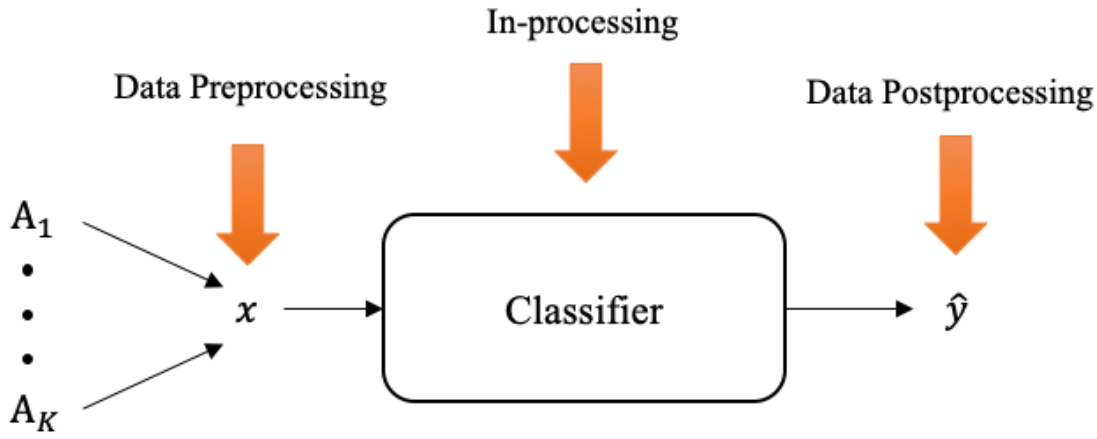
[26] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. "Fairness-Aware Classifier with Prejudice Remover Regularizer." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35-50, 2012.

[27] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness GAN: Generating Datasets with Fairness Properties using a Generative Adversarial Network," *IBM Journal of Research and Development*, vol. 63, no. 4/5, 2019.

[28] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in Learning: Classic and Contextual Bandits," In *NeurIPS*, 2016.

[29] S. Caton, and C. Haas, "Fairness in Machine Learning: A Survey," ArXiv Preprint:2010.04053, 2020.

Mitigating Social Discrimination[33]: Postprocessing



- Calibration through Output Randomization
  - Provably gives sub-optimal results[30,31]
- Thresholding of Posterior Probabilities[32]

---

[30] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning Non-Discriminatory Predictors," In *Conference on Learning Theory (COLT)*, pp. 1920-1953, PMLR, 2017.

[31] A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, and A. Pentland, "Active Fairness in Algorithmic Decision Making," In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 77-83. 2019.

[32] F. Kamiran, A. Karim, and X. Zhang, "Decision Theory for Discrimination-Aware Classification," In *IEEE 12th International Conference on Data Mining (ICDM)*, pp. 924-929, 2012.

[33] S. Caton, and C. Haas, "Fairness in Machine Learning: A Survey," ArXiv Preprint:2010.04053, 2020.

Unique Challenges in Computer Vision Applications

- Protected attribute information is typically not provided.
  - e.g. Skin tone (Fitzpatrick scale[34]), race, gender...
- Labels can be manually collected - expensive process
- Labels can be automatically generated – prone to errors
  - Accountability issues arise!
- Very few datasets benchmarked for fairness research (e.g. CelebA-Skewed)
- Relatively unexplored in vision[35]

[34] T. B. Fitzpatrick, "The Validity and Practicality of Sun-Reactive Skin Types I through VI," *Arch Dermatol.*, vol. 124, no. 6, pp. 869-871, 1988.

[35] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8919-8928, 2020.

Commercial Toolkits for Measuring and Mitigating Discrimination[36]

Project	Features
AIF360 [20]	Set of tools that provides several pre-, in-, and post-processing approaches for binary classification as well as several pre-implemented datasets that are commonly used in Fairness research
Fairlean[9]	Implements several parity-based fairness measures and algorithms [129, 5, 6] for binary classification and regression as well as a dashboard to visualize disparity in accuracy and parity.
Aequitas [235]	Open source bias audit toolkit. Focuses on standard ML metrics and their evaluation for different subgroups of a protective attribute.
Responsibly [190]	Provides datasets, metrics, and algorithms to measure and mitigate bias in classification as well as NLP (bias in word embeddings).
Fairness[10]	Tool that provides commonly used fairness metrics (e.g., statistical parity, equalized odds) for R projects.
FairTest [264]	Generic framework that provides measures and statistical tests to detect unwanted associations between the output of an algorithm and a sensitive attribute.
Fairness Measures[11]	Project that considers quantitative definitions of discrimination in classification and ranking scenarios. Provides datasets, measures, and algorithms (for ranking) that investigate fairness.
Audit AI[12]	Implements various statistical significance tests to detect discrimination between groups and bias from standard machine learning procedures.
Dataset Nutrition Label [134]	Generates qualitative and quantitative measures and descriptions of dataset health to assess the quality of a dataset used for training and building ML models.
ML Fairness Gym	Part of Google's Open AI project, a simulation toolkit to study long-run impacts of ML decisions.[13] Analyzes how algorithms that take fairness into consideration change the underlying data (previous classifications) over time (see e.g. [187, 92, 94, 135, 204]).

[36] S. Caton, and C. Haas, "Fairness in Machine Learning: A Survey," ArXiv Preprint:2010.04053, 2020.

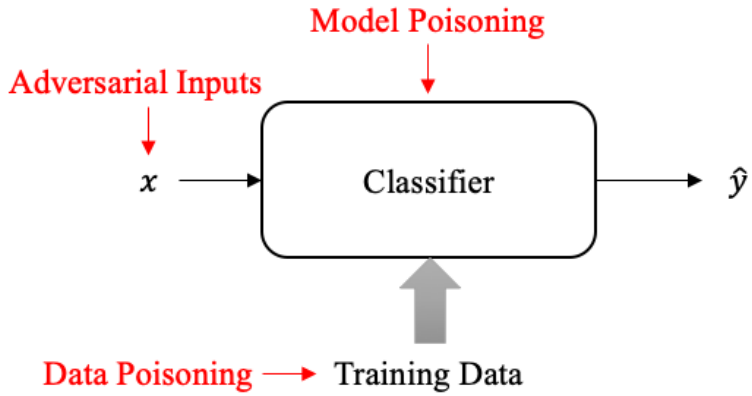
Fairness Dilemmas[37]

- Fairness vs. Model Performance
- (Dis)agreement and Incompatibility of Fairness Notions
- Tensions with Context and Policy
- Democratisation of ML vs. Fairness Skills Gap

[37] S. Caton, and C. Haas, ”Fairness in Machine Learning: A Survey,” ArXiv Preprint:2010.04053, 2020.

Robust Machine Learning[38]

- Security in Machine Learning
  - White-Box vs. Black-Box Attacks
  - Data Poisoning vs. Model Poisoning
  - **Adversarial Inputs**

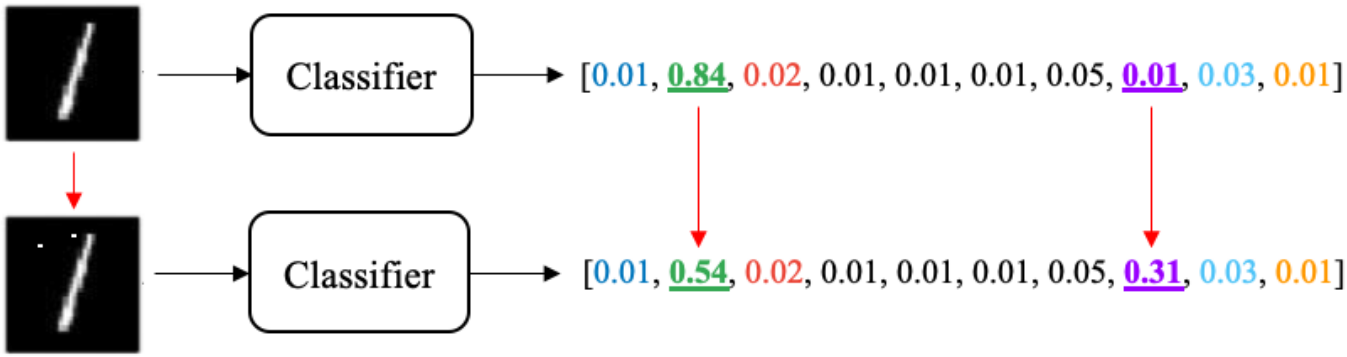


- Privacy in Machine Learning
  - Data exfiltration vs. Model exfiltration

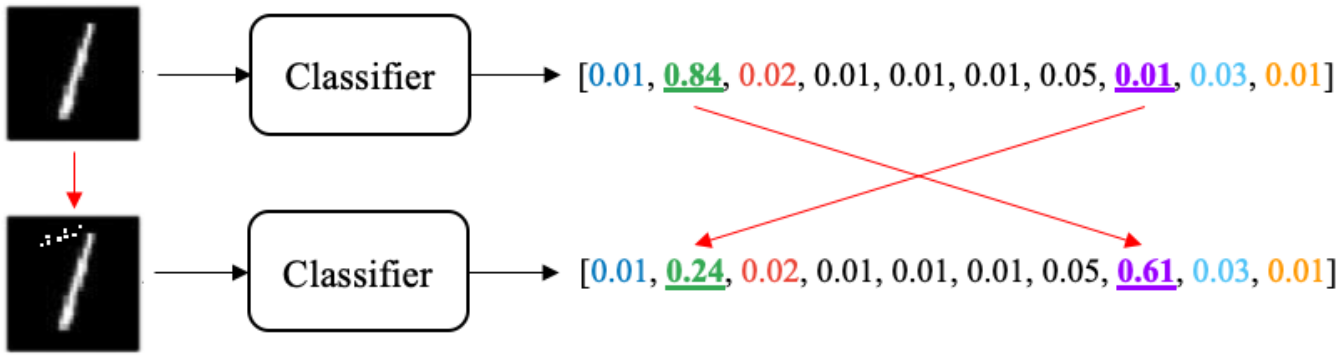
[38] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman. "Towards the Science of Security and Privacy in Machine Learning," ArXiv Preprint:1611.03814, 2016.

Security in Machine Learning: Adversarial Threat Models[39]

- Confidence Reduction



- Misclassification

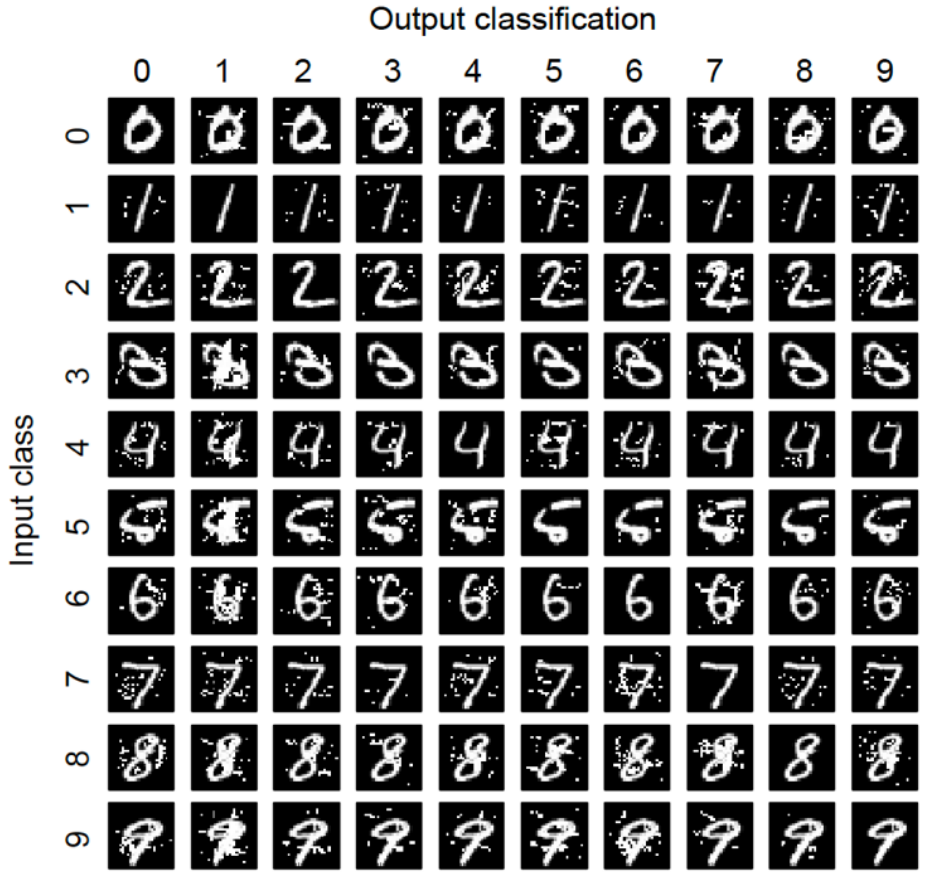


[39] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372-387, 2016.

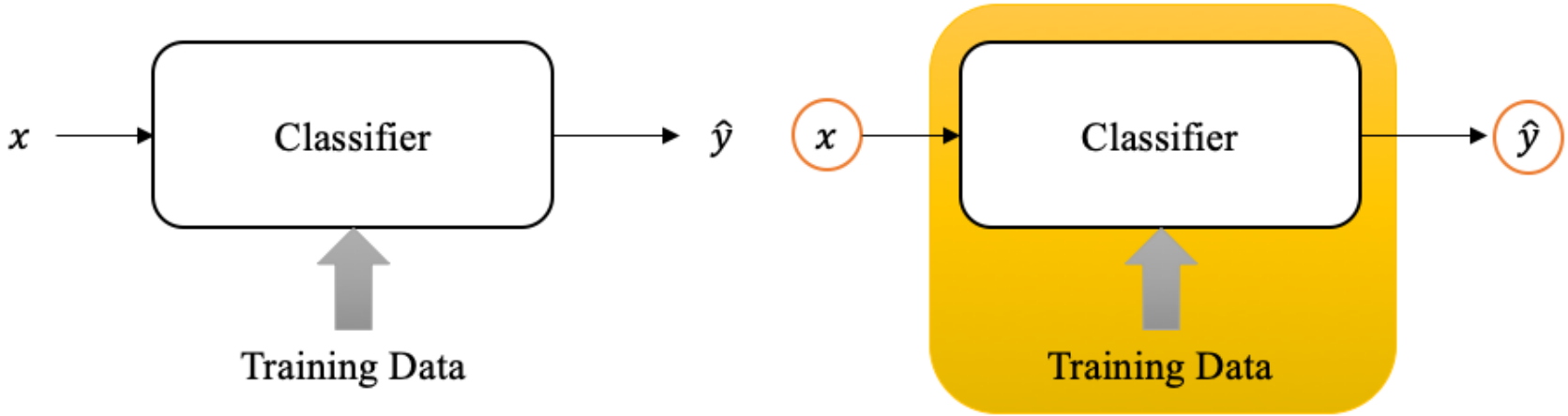
Adversarial Threat Models (cont...)

- Targeted Misclassification

- Source/Target Misclassification



Adversarial Capabilities at the Attacker



White-Box Attacks:

- Architecture + Training Data
- Architecture
- Training Data

Black-Box Attacks:

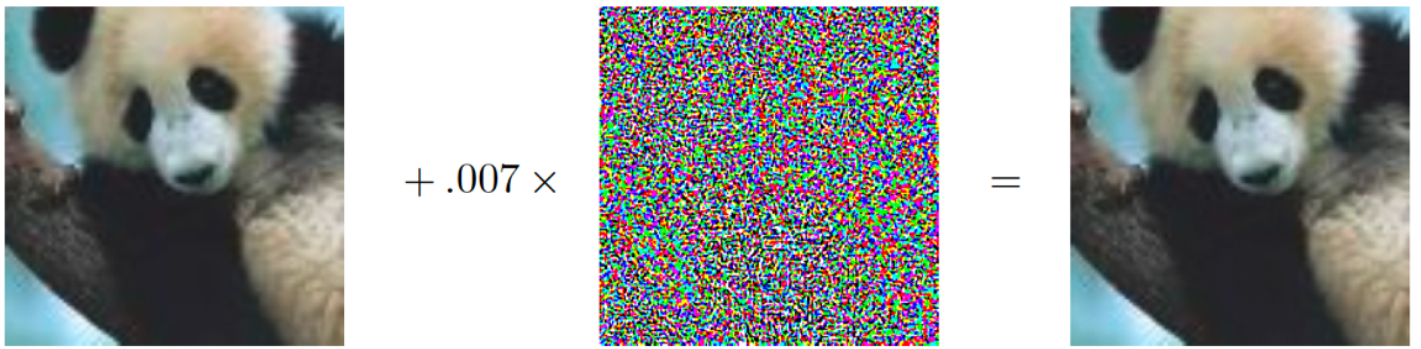
- Oracle (Input-Output pairs from ML model)
- Input Samples

White-Box Adversarial Samples[41]

Notation: Let

- x denote the input, $F(x)$ denote the class score (output) of NN.
- x^{adv} denote the adversarial input.
- ℓ_∞***-Bounded Adversaries:*** Given some budget ϵ, $||x^{adv} - x||_\infty \leq \epsilon$.
- $J(\hat{y}, y)$ is the loss function.

Fast Gradient Sign Method (FGSM)[40]***:*** $x_{FGSM} = x + \epsilon \cdot \text{sign}\Big(\nabla_x J(h(x), y)\Big)$



[40] I. Goodfellow, J. Shlens, C. Szegedy, "Explaining and Harnessing Adversarial Examples," *International Conference on Learning Representations (ICLR)*, 2015.

[41] A. Kurakin, D. Boneh, F. Tramèr, I. Goodfellow, N. Papernot, P. McDaniel, "Ensemble Adversarial Training: Attacks and Defenses," in *International Conference on Learning Representations (ICLR)*, 2018.

White-Box Adversarial Samples[43]

Single-Step Least-Likely Class Method[42] (LL): Let $y_{LL} = \arg\min\{h(x)\}$

$$x_{LL} = x - \epsilon \cdot \text{sign}\Big(\nabla_x J(f(x), y_{LL})\Big)$$

Iterative Attack[42] (I-FGSM/I-LL): Iteratively apply FGSM, or Step-LL k times with step size $\alpha \geq \epsilon/k$, and project each step onto ℓ_∞-ball of norm ϵ around x.

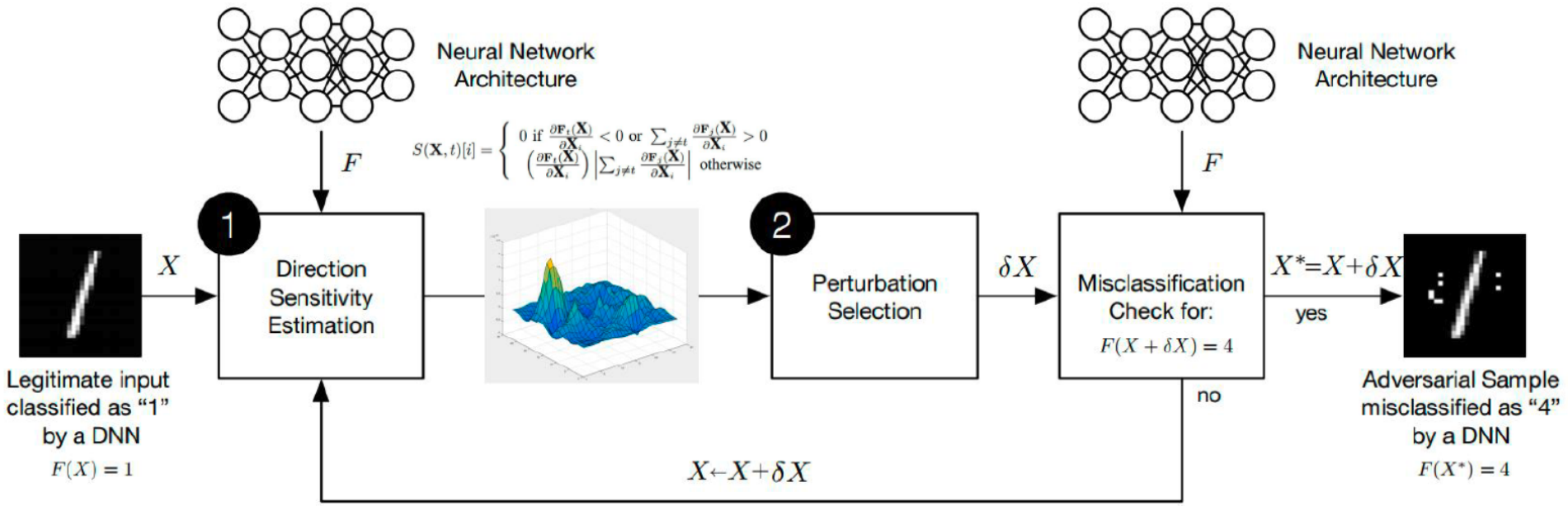
[42] A. Kurakin, I. J. Goodfellow, S. Bengio, "Adversarial Machine Learning at Scale," *International Conference on Learning Representations (ICLR)*, 2017.

[43] A. Kurakin, D. Boneh, F. Tramèr, I. Goodfellow, N. Papernot, P. McDaniel, "Ensemble Adversarial Training: Attacks and Defenses," in *International Conference on Learning Representations (ICLR)*, 2018.

White-Box Adversarial Samples[44] (cont.)

Jacobian-Based Saliency Map Approach (JSMA):

1. Compute forward derivative of neural network (saliency Map)
2. Construct adversarial saliency maps
3. Modify samples



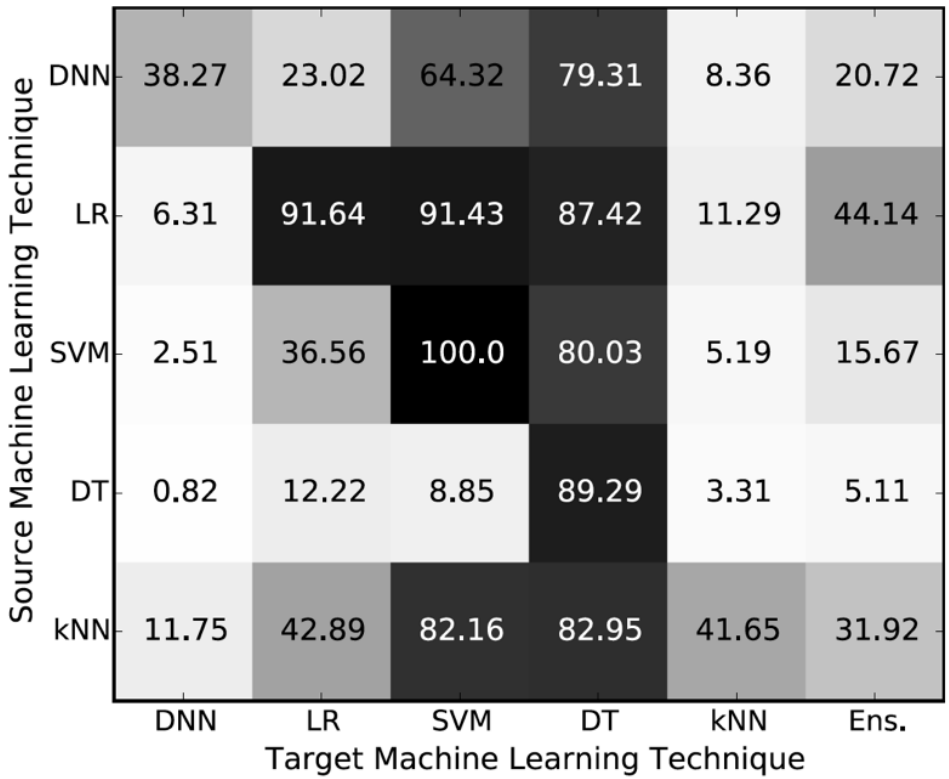
[44] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372-387, 2016.

Black-Box Adversarial Inputs and Transferability[45]

Need to alleviate

- Lack of knowledge about model
- Lack of knowledge about training data

*Take advantage of **transferability** of adversarial examples across ML models.*



Query the ML system using synthetic inputs and generate synthetic training data!

[45] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: From Phenomena to Black-Box Attacks using Adversarial Samples," ArXiv Preprint:1605.07277, 2016.

Substitute DNNs for Black-Box Adversarial Inputs[46]

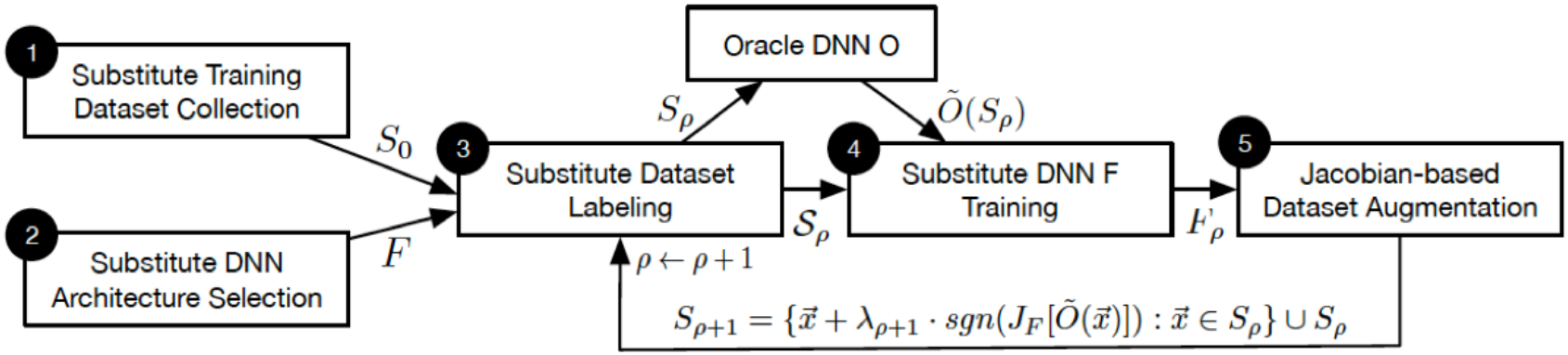


Figure 3: **Training of the substitute DNN** F: the attacker (1) collects an initial substitute training set S_0 and (2) selects an architecture F. Using oracle $\tilde{O}$, the attacker (3) labels S_0 and (4) trains substitute F. After (5) Jacobian-based dataset augmentation, steps (3) through (5) are repeated for several substitute epochs ρ.

Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	Deep Learning	6,400	84.24%
amazon web services	Logistic Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)

[46] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," in *Proceedings of ACM Asia Conference on Computer and Communications Security (ASIA CCS'17)*, pp. 506–519, 2017.

Defending Against Adversarial Inputs

- ***Adversarial Training***[47] ***:*** Train against adversarial examples
- ***Defensive Distillation***[48] ***:*** Train on model scores instead of hard labels, to smoothen out gradients.

Both approaches typically fail to mitigate black-box models due to ***gradient masking****!*

Resourceful Blog: cleverhans.io

Also comes with a Python library of all adversarial threat models and benchmark robust models.

---

[47] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," In *International Conference on Learning Representations (ICLR)*, 2018.

[48] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks," in *37th IEEE Symposium on Security & Privacy (SSP)*, San Jose, CA, 2016.