TEXAS A&M UNIVERSITY

CSCE 633 - Machine Learning

Lecture 14 - Trustworthy Machine Learning

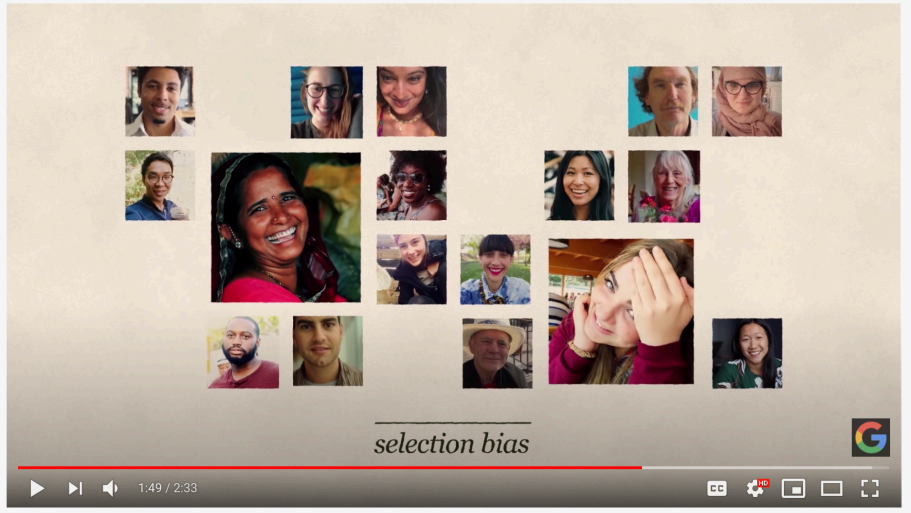
Trustworthy Machine Learning

- Decision making in high state applications becomes increasingly data-driven
  - Educational assessment
  - Employment
  - Healthcare
  - Criminal Justice
- Speed and accuracy are not enough!
- Characteristics of ML algorithms that compromise trustworthiness

The contents of these slides appear in:
Varshney, "Trustworthy Machine Learning and Artificial Intelligence"
Kaul, "Speed and accuracy are not enough! Trustworthy machine learning"

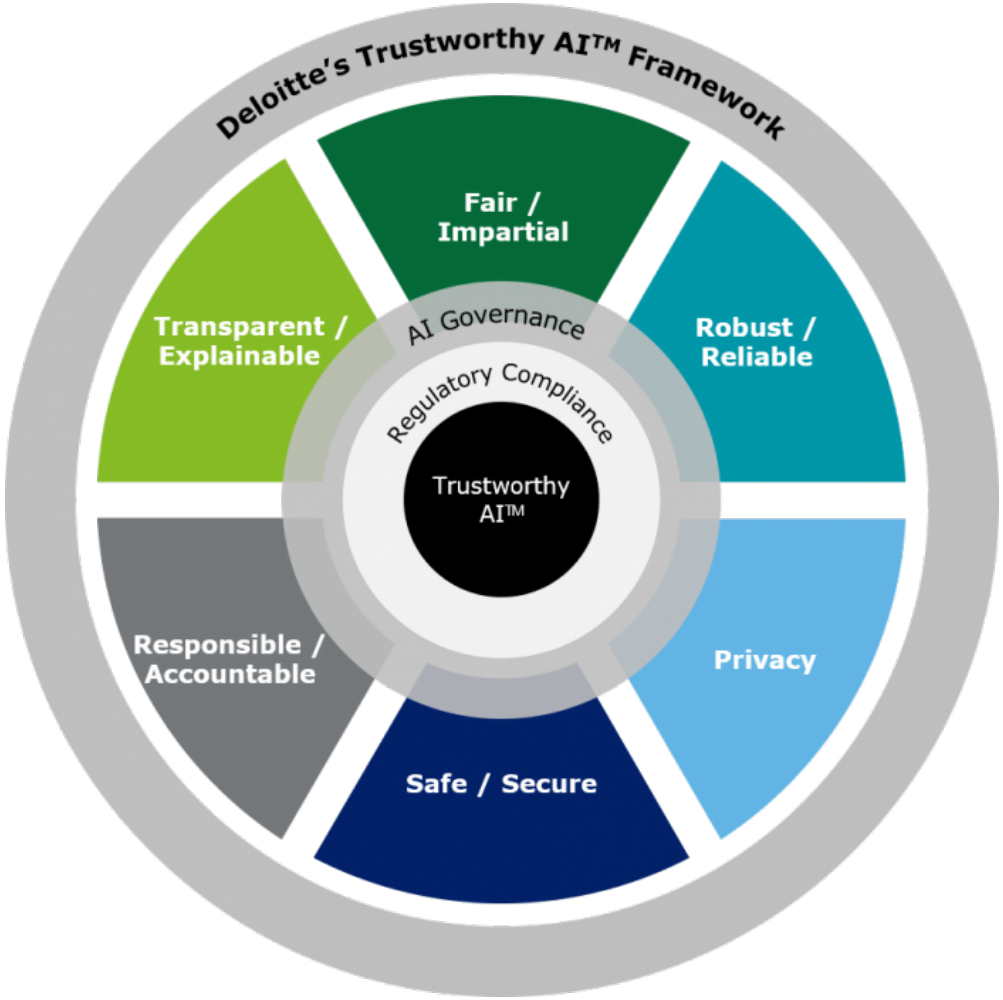
Trustworthy Machine Learning

- Self-driving car fatalities in unusual conditions that the ML algorithms have not been trained on
- A model supporting judge decisions reported to be biased against African American defendants
- A model supporting resume screening reported to be biased against women
- ML models for disease diagnosis from chest x-rays shown to give importance to text contained in the image, rather than details of the patients' anatomy

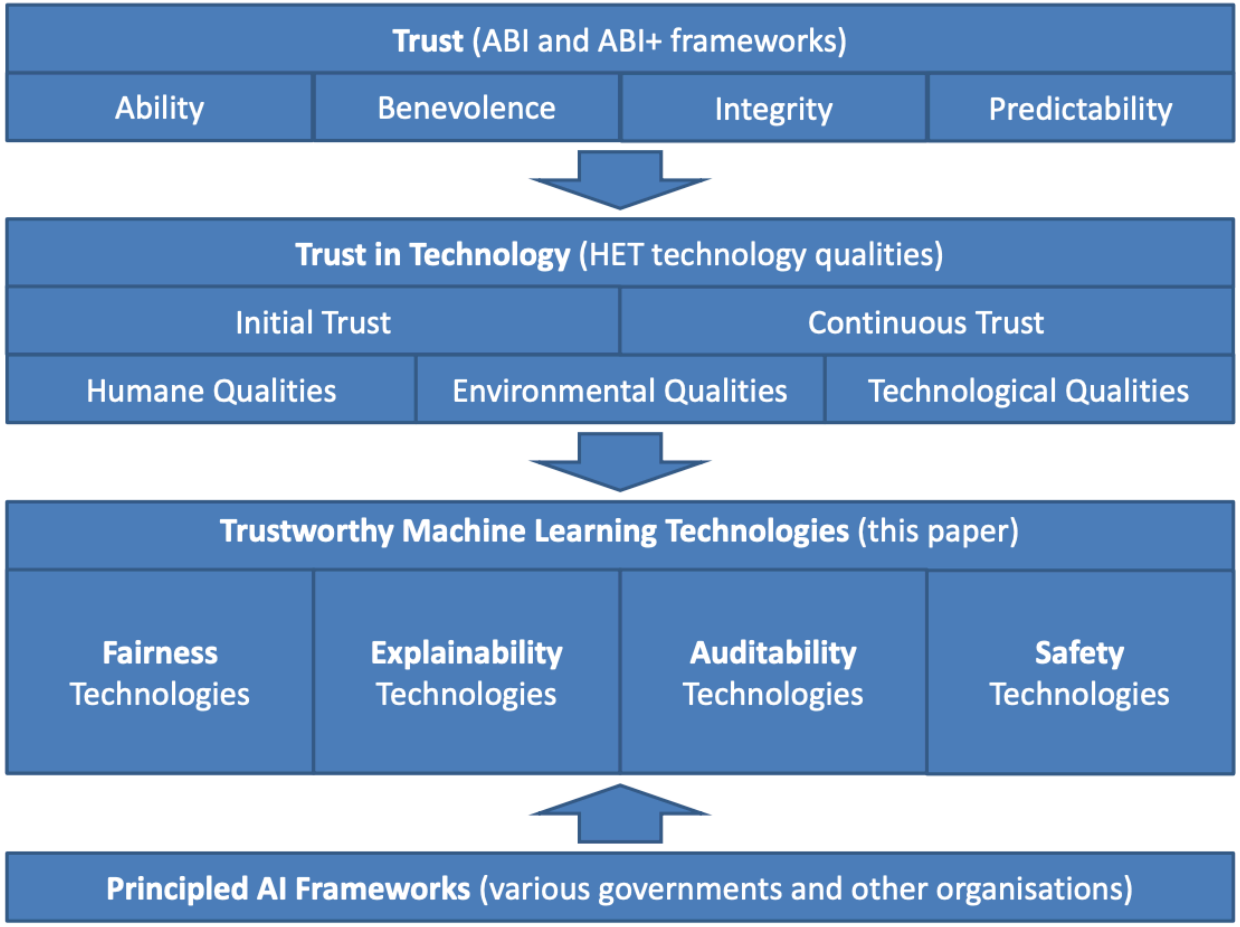


https://www.youtube.com/watch?v=59bMh59JQDo

Pillars of trustworthiness in ML



Trustworthy Machine Learning



Robustness

- Data distributions may change over time and over different contexts
  - e.g., biomarkers change as people become older
  - e.g., vocal measures change depending on acoustic conditions
- Difficult to obtain large sets of training data from target distribution
- Data mismatch
  - Non-matching conditions between training and testing conditions
- Decrease in trust to ML
  - System performance is directly correlated to human trust

Robustness: Handling small datasets

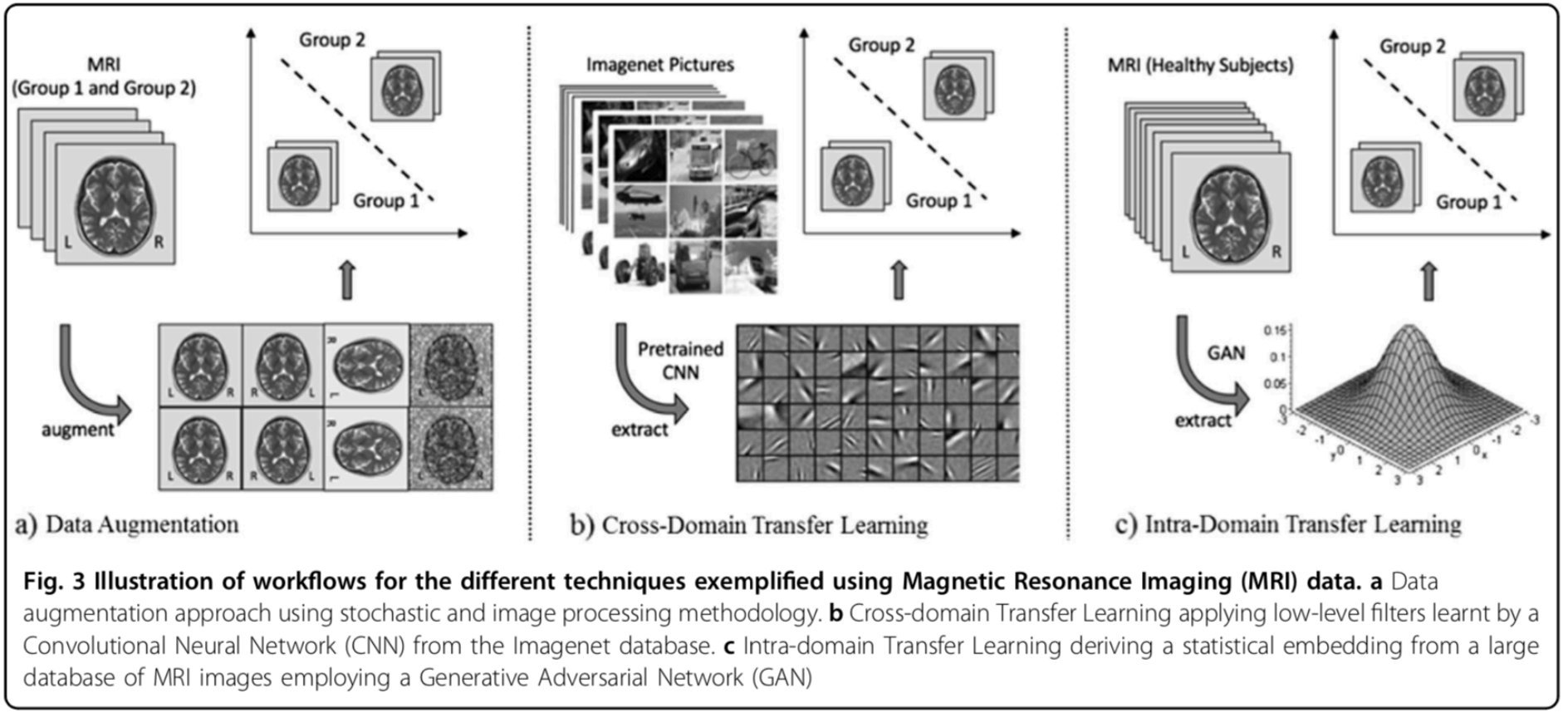
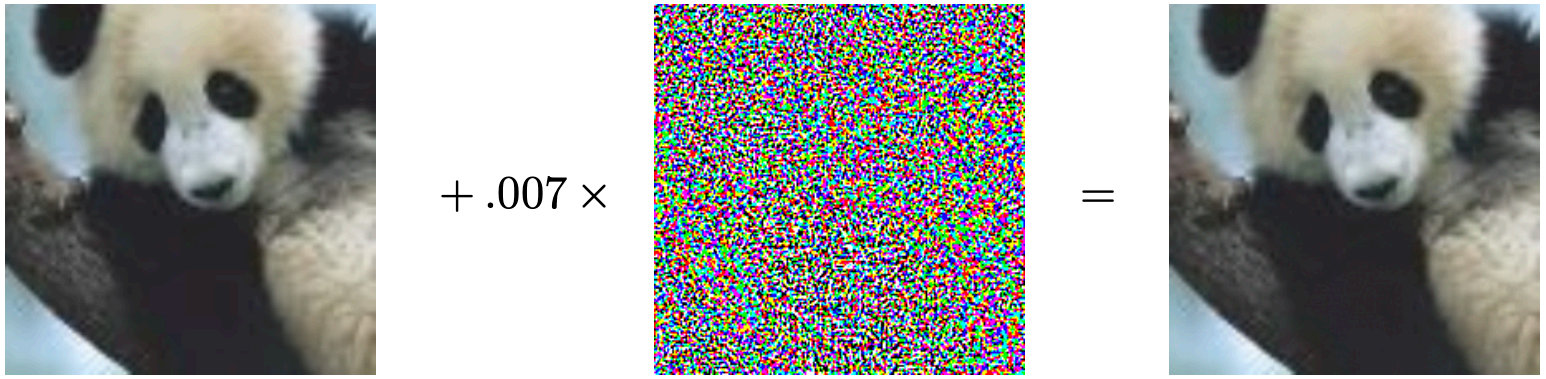


Fig. 3 Illustration of workflows for the different techniques exemplified using Magnetic Resonance Imaging (MRI) data. a Data augmentation approach using stochastic and image processing methodology. **b** Cross-domain Transfer Learning applying low-level filters learnt by a Convolutional Neural Network (CNN) from the Imagenet database. **c** Intra-domain Transfer Learning deriving a statistical embedding from a large database of MRI images employing a Generative Adversarial Network (GAN)

Cearns et al., Recommendations and future directions for supervised machine learning in psychiatry, 2019

Safety/Security

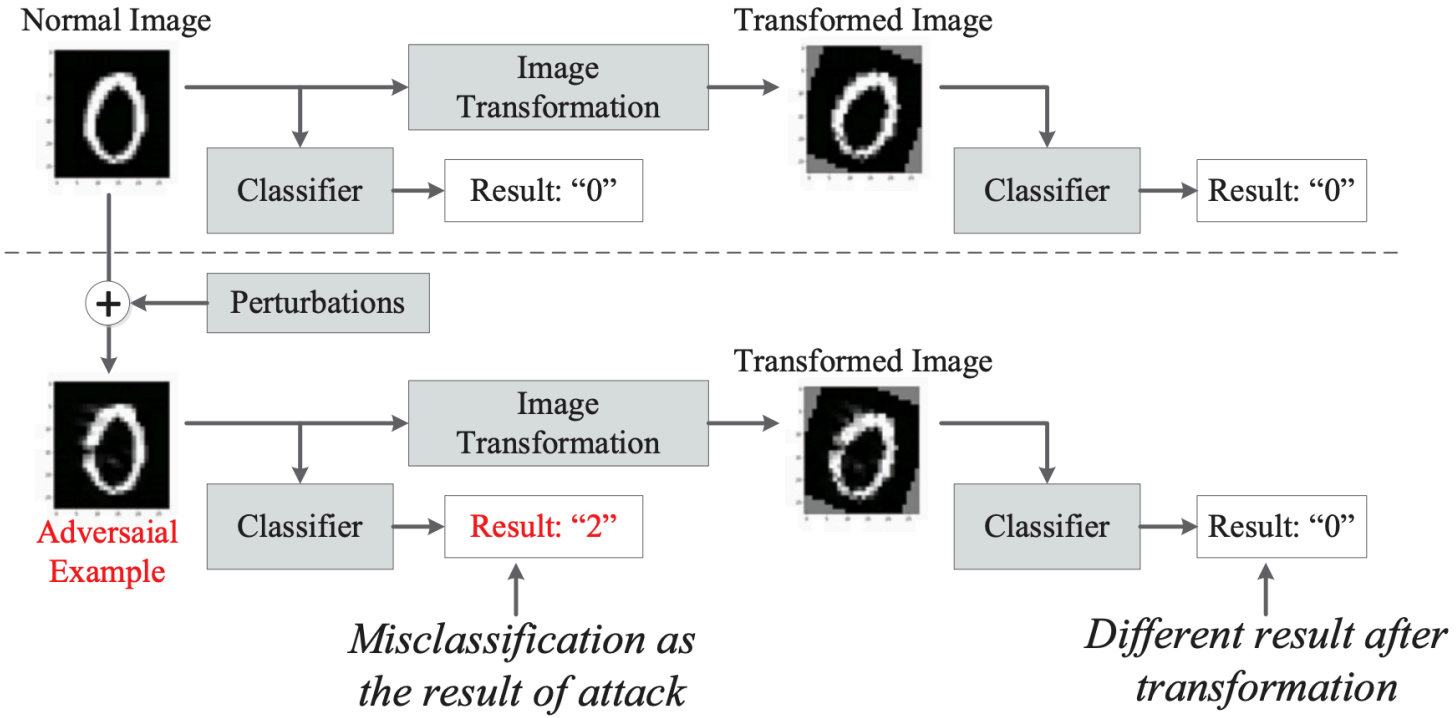
- Deliberate data poisoning attack by adversaries
  - Injecting a few carefully designed data samples in the training data to confuse the system
  - Altering a small number of features or image pixels



Goodfellow et al., Explaining and harnessing adversarial examples, ICLR 2015

Safety/Security

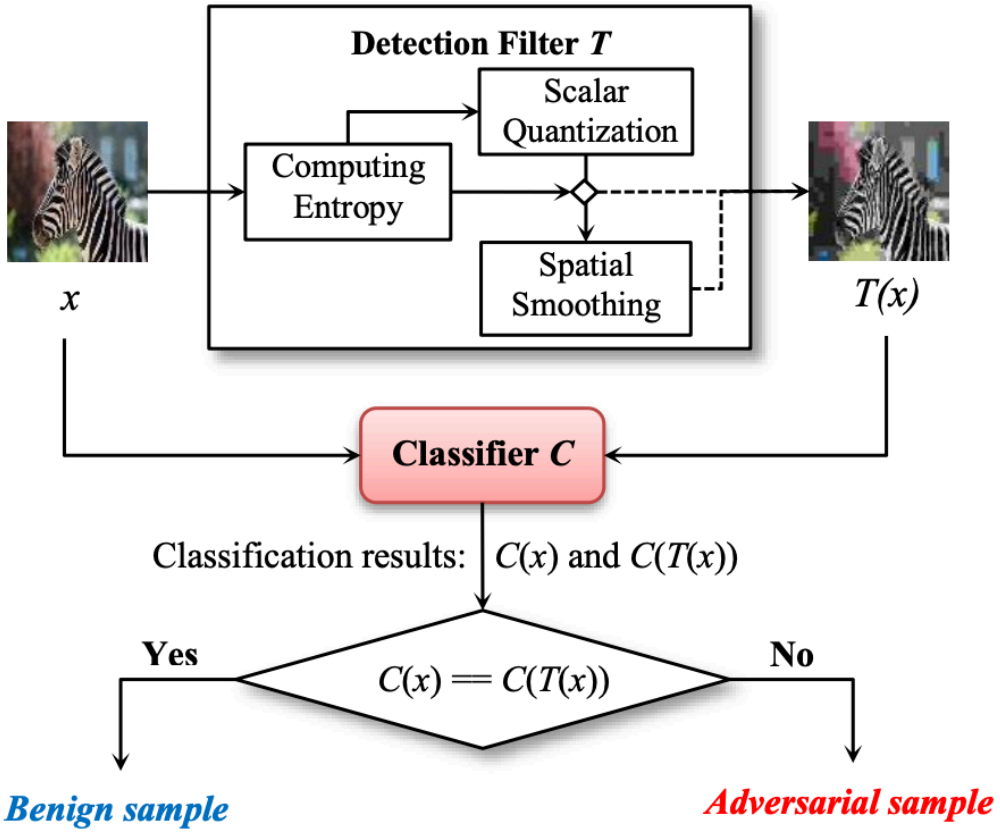
- Adversarial examples are usually sensitive to certain image transformation operations such as rotation and shifting



Tian et al., Detecting Adversarial Examples through Image Transformation, AAAI 2018

Safety/Security

- Adaptive noise reduction using image entropy
- The largest the image entropy, the more likely an adversarial attack



Liang et al., Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction

Privacy

- Privacy concerns in multimodal data
  - Especially speech and image
- 37% of patients feel comfortable in sharing their data
  - [German & Barber, 2018]
- Numerous privacy violations from health administrators

Privacy

DeepPrivacy

- Automatically anonymizing faces in images while retaining the original data distribution

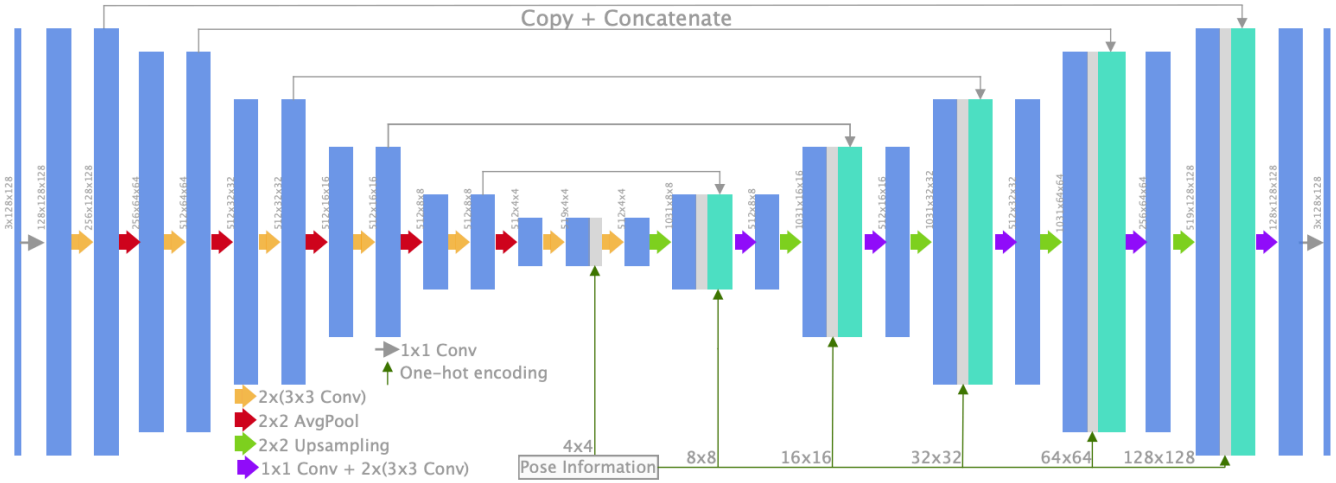


Fig. 3: **Generator Architecture** for 128×128 resolution. Each convolutional layer is followed by pixel normalization [12] and LeakyReLU($\alpha = 0.2$). After each upsampling layer, we concatenate the upsampled output with pose information and the corresponding skip connection.

Hukkelas, DeepPrivacy: A Generative Adversarial Network for Face Anonymization, 2019

Privacy

DeepPrivacy

- Automatically anonymizing faces in images while retaining the original data distribution

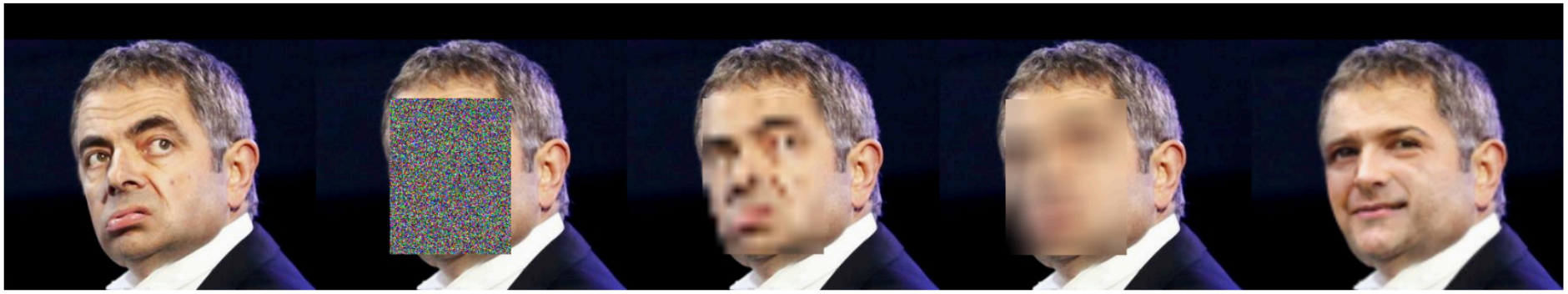


Fig. 1: **DeepPrivacy Results** on a diverse set of images.

Hukkelas, DeepPrivacy: A Generative Adversarial Network for Face Anonymization, 2019

Privacy

Face anonymization preserving utility information

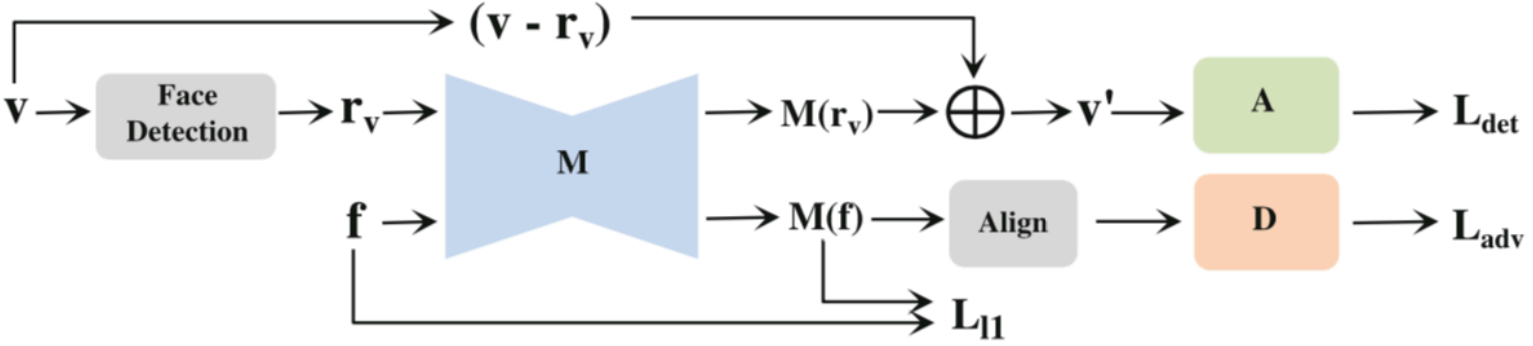


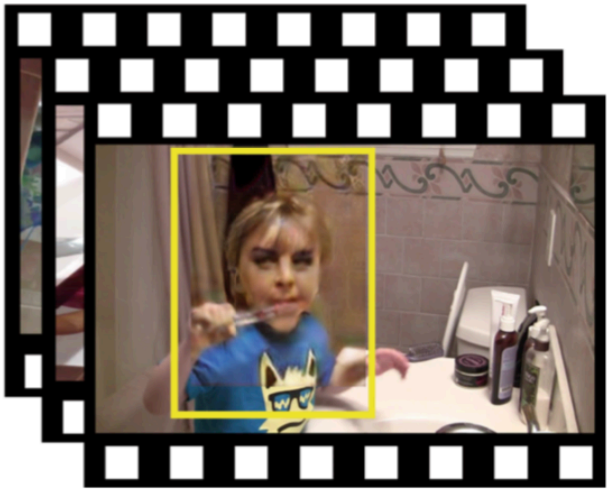
Fig. 2. Our network architecture for privacy-preserving action detection. We simultaneously train a face modifier M whose job is to alter the input face (f or r_v) so that its identity no longer matches that of the true identity, and an action detector A whose job is to learn to accurately detect actions in videos in spite of the modifications. The face classifier D acts as an adversary and ensures that the modified face is non-trivial. See text for details. (Gray blobs are not learned during training.)

Red et al., Learning to Anonymize Faces for Privacy Preserving Action Detection, ECCV 2018

Privacy

Face anonymization preserving utility information

Identity: Alex
Action: Brush Teeth



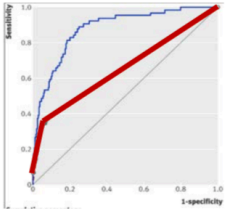
Identity: ???
Action: Brush Teeth

Red et al., Learning to Anonymize Faces for Privacy Preserving Action Detection, ECCV 2018

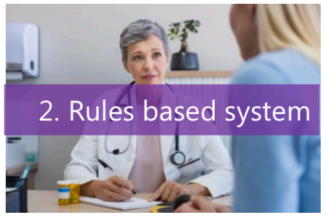
Explainability

- Many fields are far away from using ML due to its limitations to provide interpretable decisions

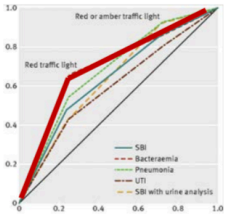
Factors: 7 ± 2



~80% Care Decisions

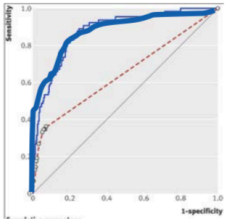


Factors: 10s



~18% Care Decisions

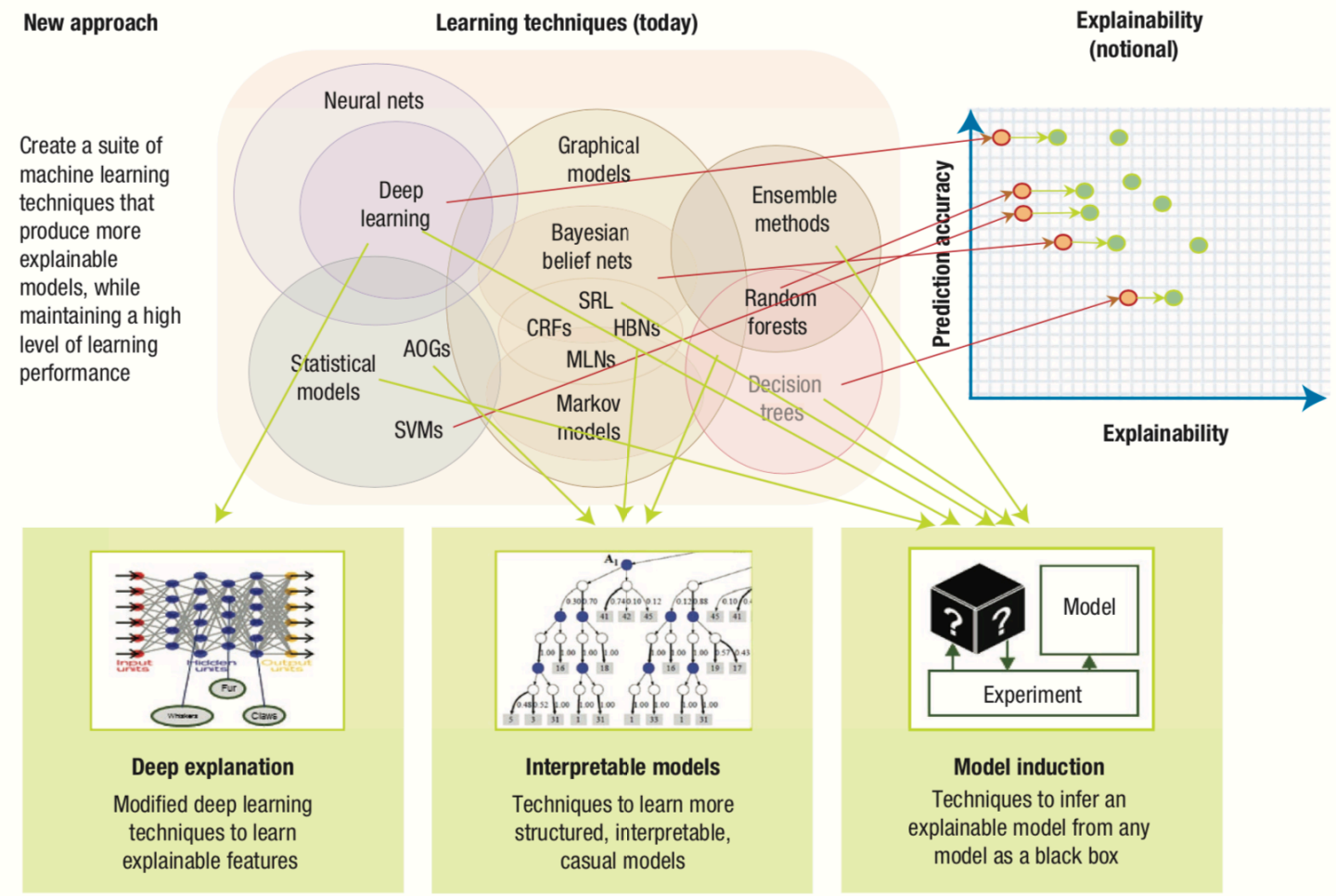
Factors: 100s



~2% Care Decisions

Explainability

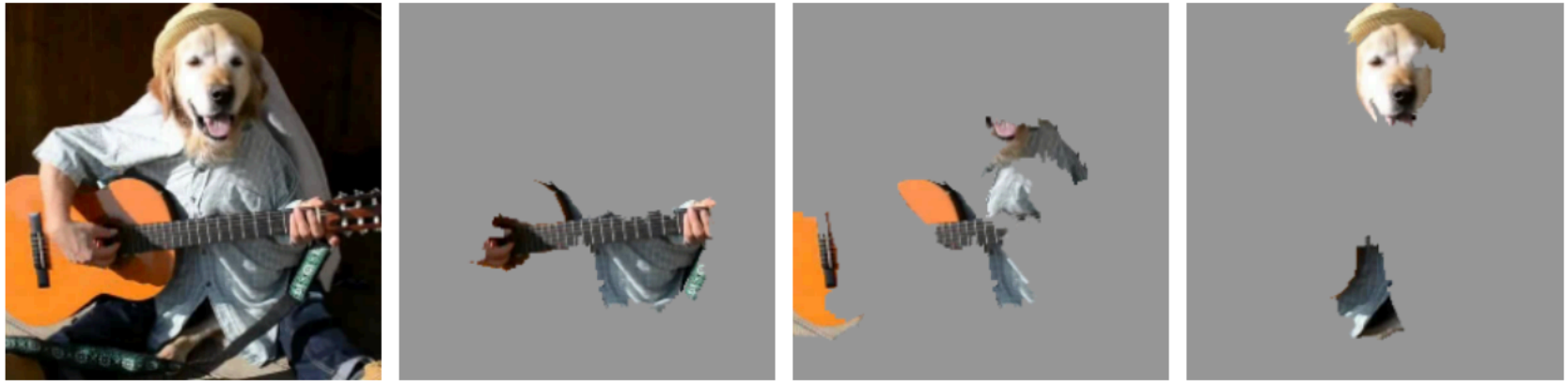
- Increasing model complexity compromises interpretability



Explainability

Local Interpretable Model-Agnostic Explanations (LIME)

- Model-agnostic: identifies the parts of the interpretable input are contributing to the prediction
- "I hate this movie" -> "I hate movie", "I this movie", "I movie", "I hate" -> identify which parts of the sentence are consistently contributing to the prediction



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Tulio Ribeiro et al., "Why Should I Trust You?" Explaining the Predictions of Any Classifier, ACM SIGKDD 2016

Fairness

- **Bias in model design**
  - Label bias: the same outcome might not mean the same for all individuals
  - Cohort bias: considering traditional groups (e.g., male/female) without considering other protected groups (e.g., LGTBQ) and levels of granularity
- **Bias in training data**
  - Minority bias: minoritized groups might have insufficient number of samples
  - Missing data bias: minoritized groups may have missing data in a non-random fashion (e.g., lower quality sensor devices)
  - Informativeness bias: features might be less informative for a certain group (e.g., identifying melanoma from a patient with dark skin)
- **Bias in interaction with experts**
  - Automation bias: experts are unaware that a model is underperforming for a certain group
  - Feedback loops: If the clinician accepts incorrect model outputs, the mistake is propagated next time the model is trained
  - Dismissal bias: Desensitization to alerts that are systematically incorrect for a specific group
- **Bias in interaction with users**
  - Privilege bias: ML models might be unavailable in places where specific groups receive care
  - Informed mistrust: Users might believe that a model is biased against them due to historical exploitation practices

Rajkomar, Ensuring Fairness in Machine Learning to Advance Health Equity, Annals of Internal Medicine 2018

Fairness

Ways to mitigate bias

Table. Recommendations

Design

- Determine the goal of a machine-learning model and review it with diverse stakeholders, including protected groups.
- Ensure that the model is related to the desired patient outcome and can be integrated into clinical workflows.
- Discuss ethical concerns of how the model could be used.
- Decide what groups to classify as protected.
- Study whether the historical data are affected by health care disparities that could lead to label bias. If so, investigate alternative labels.

Data collection

- Collect and document training data to build a machine-learning model.
- Ensure that patients in the protected group can be identified (weighing cohort bias against privacy concerns).
- Assess whether the protected group is represented adequately in terms of numbers and features.

Training

- Train a model taking into account the fairness goals.

Evaluation

- Measure important metrics and allocation across groups.
- Compare deployment data with training data to ensure comparability.
- Assess the usefulness of predictions to clinicians initially without affecting patients.

Launch review

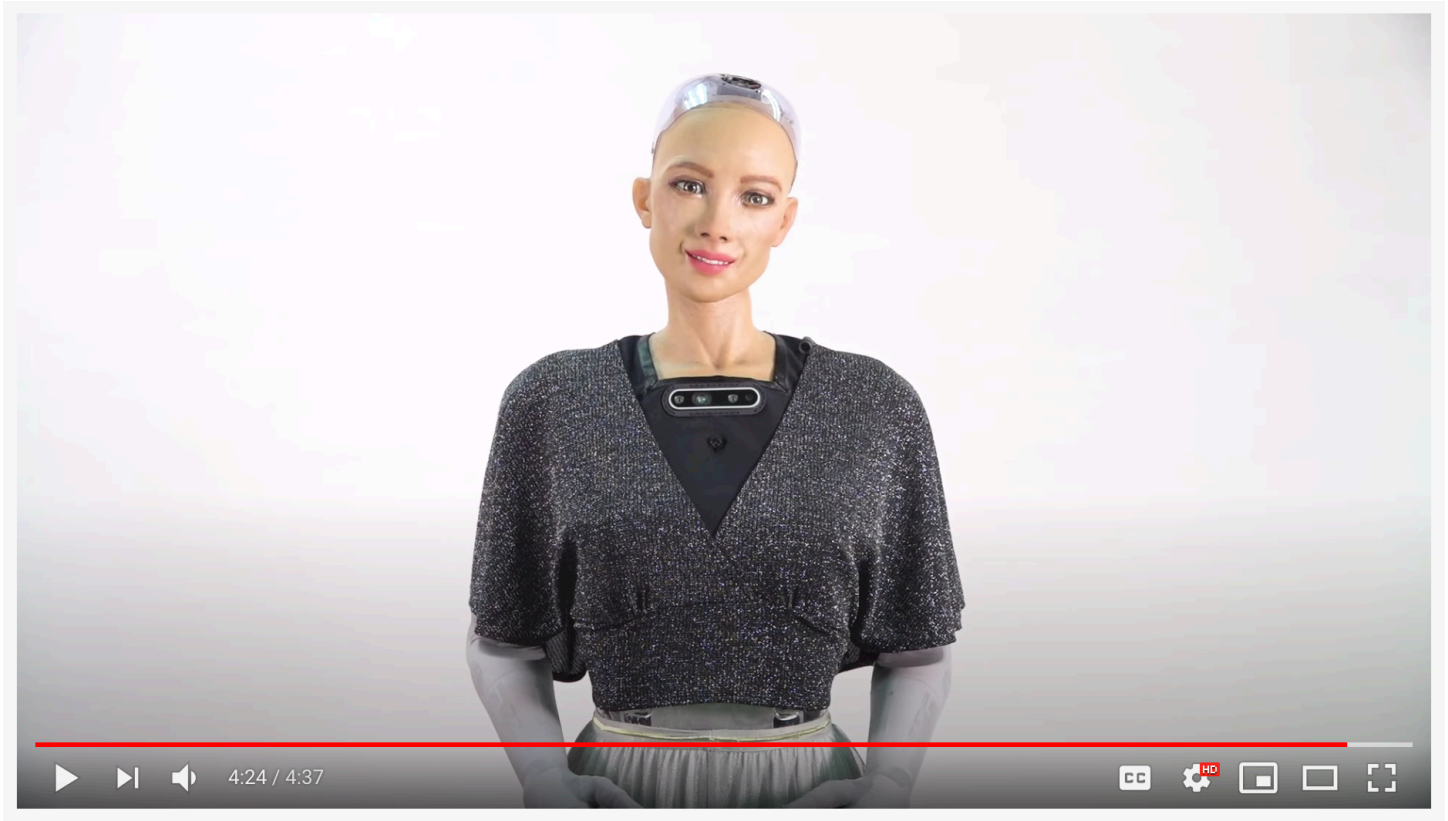
- Evaluate whether a model should be launched with all stakeholders, including representatives from the protected group.

Monitored deployment

- Systematically monitor data and important metrics throughout deployment.
- Gradually launch and continuously evaluate metrics with automated alerts.
- Consider a formal clinical trial design to assess patient outcomes.
- Periodically collect feedback from clinicians and patients.

Rajkomar, Ensuring Fairness in Machine Learning to Advance Health Equity, Annals of Internal Medicine 2018

Message for AI from Sophia the Robot
International Day of Tolerance (November 16)



https://www.youtube.com/watch?v=NwNULnXk9b0