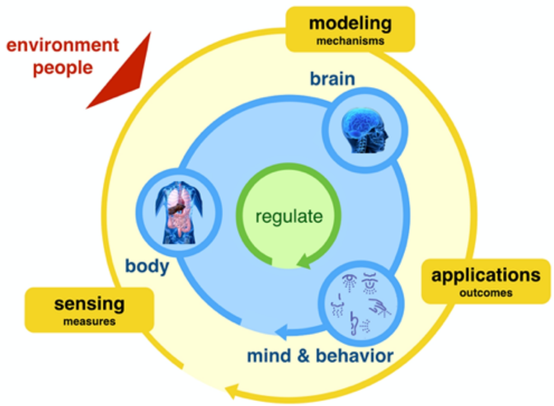
Toward trustworthy human-centered machine intelligence

Signal Analysis and Interpretation Lab Team:
Shrikanth (Shri) Narayanan (PI)
Tiantian Feng, Rahul Sharma, Raghuveer Peri, Amrutha Nadarajan, Rajat Hebbar

Amazon Mentors/POC:
Rahul Gupta, Anil Ramakrishna

USC-Amazon Center Kickoff Meeting, Sept/2021

Human-centered Machine Intelligence Ecosystem

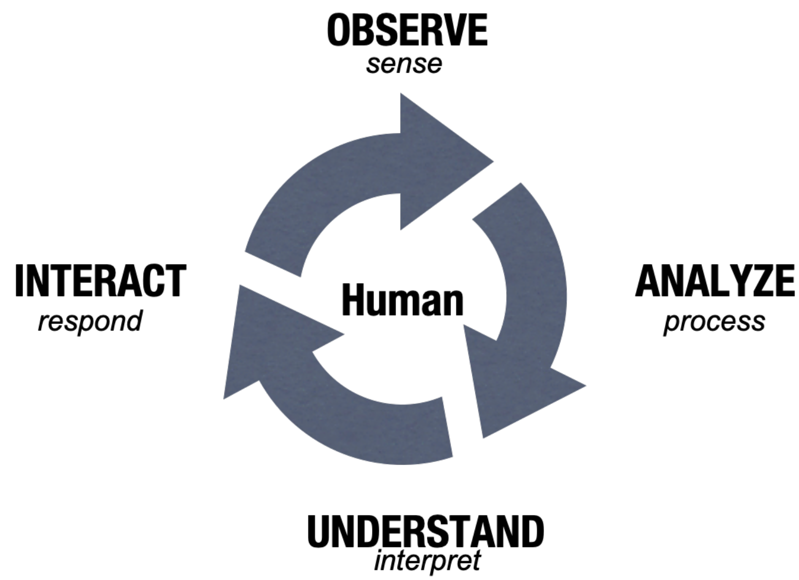


Goals

- *Understand the human condition: traits, state, behavior, interaction*
- *Support and enhance human experiences*

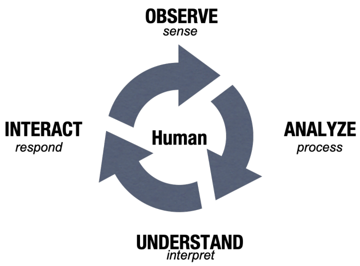
Human centered view: characterizing data
<u>about, from</u> <u>and</u> *<u>for</u>* people

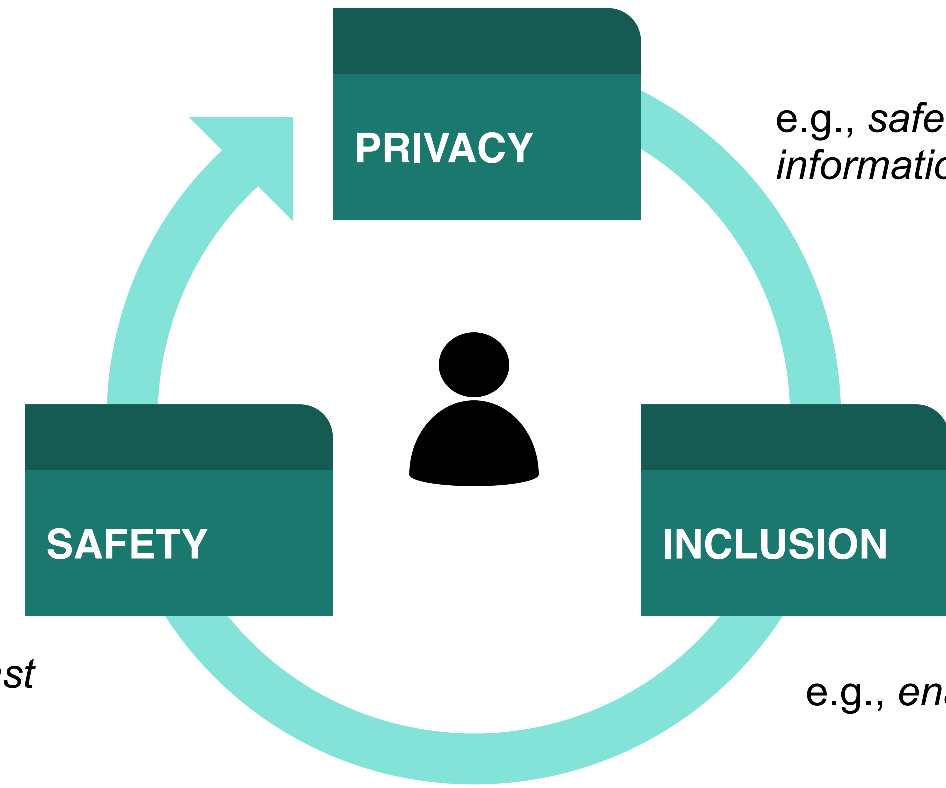
- includes knowledge about how *people* perceive, process and use (human) data
  - *constructs capturing human expression & experience*
  - *constructs characterizing human "perception"*
  - *some constructs (e.g., "labels") may be available explicitly, others have to be (self) learned implicitly*
  - *other constructs may not be human "label-able"*



2

Elements of Trustworthy Human-centered Machine Intelligence





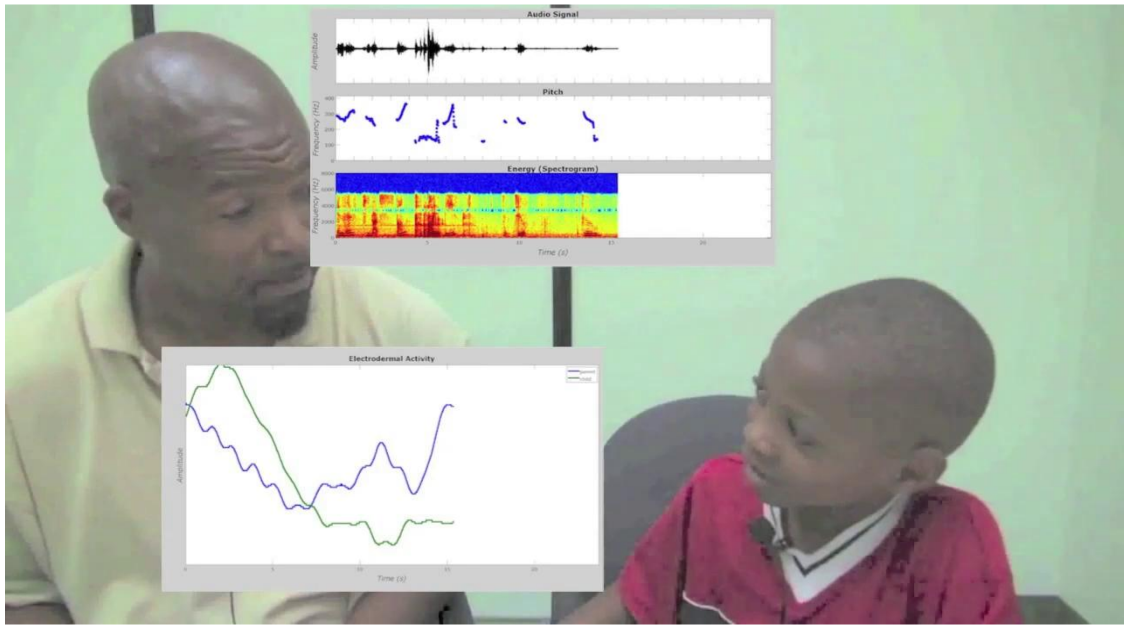
e.g., *safeguard protected information*

e.g., *enable broad access*

e.g., *protect against malicious attacks*

S. Narayanan and A. Madni. Inclusive Human centered Machine Intelligence. *The Bridge*. 50(S): 113-116. National Academy of Engineering, 2020

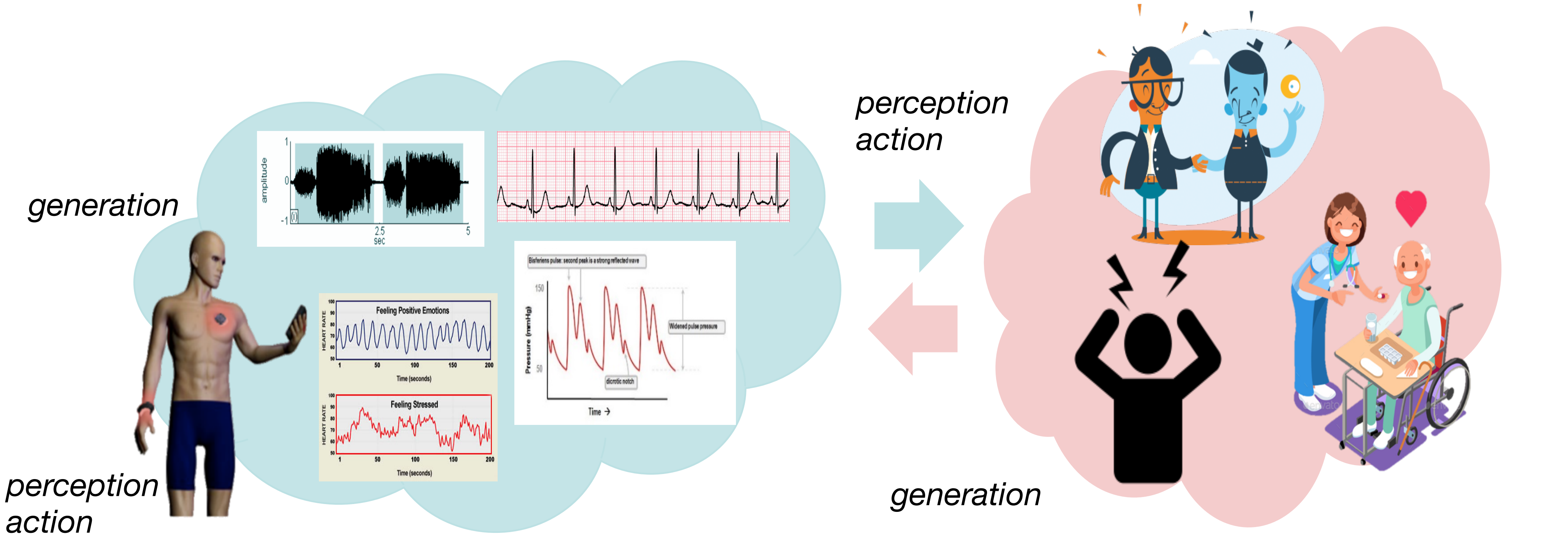
A human conversation example:
rich verbal and nonverbal behavior and interaction



*<u>Speech and language</u> provide access to assessing **intent, emotions**, and a variety of information about personal **demographic traits** (age, gender,…), **physical/psychological/health state**, and **interaction context.***
These attributes/constructs are often intricately related.

Other biobehavioral data streams share similar “loaded” characteristics: e.g., ECG

"Human Aspect" Of Human Centric Machine Intelligence



- **Tight links between generation and processing of signals by human system and their interaction with the world**
  - the system is characterized by traits/states/behavior expression
  - in turn the human perception/experience affects the system and shapes future behavior /actions

- **Relies on knowledge regarding aspects of individual traits, state, behavior for providing the desired user experience**
  - *while lots of information is available-and possibly retrievable- from human signals not all of it is essential or should be used*

For example:

- a specific use case may rely on *what* was said but need not know *who* said it, or what their *affective* state was

- another use case may rely on *age* information and aspects of *health state* but doesn't need to track the specific *linguistic content* of an interaction

Example human-centric applications at SAIL

Large-scale workplace behavioral study

TILES study: different rooms in the hospital, different privacy constraints https://tiles-data.isi.edu/

Psychotherapy Interaction **Child-inclusive interaction**

Clinically-relevant features

- Empathy
- Entrainment/ Synchrony
- Visual Gaze
- Emotion-state

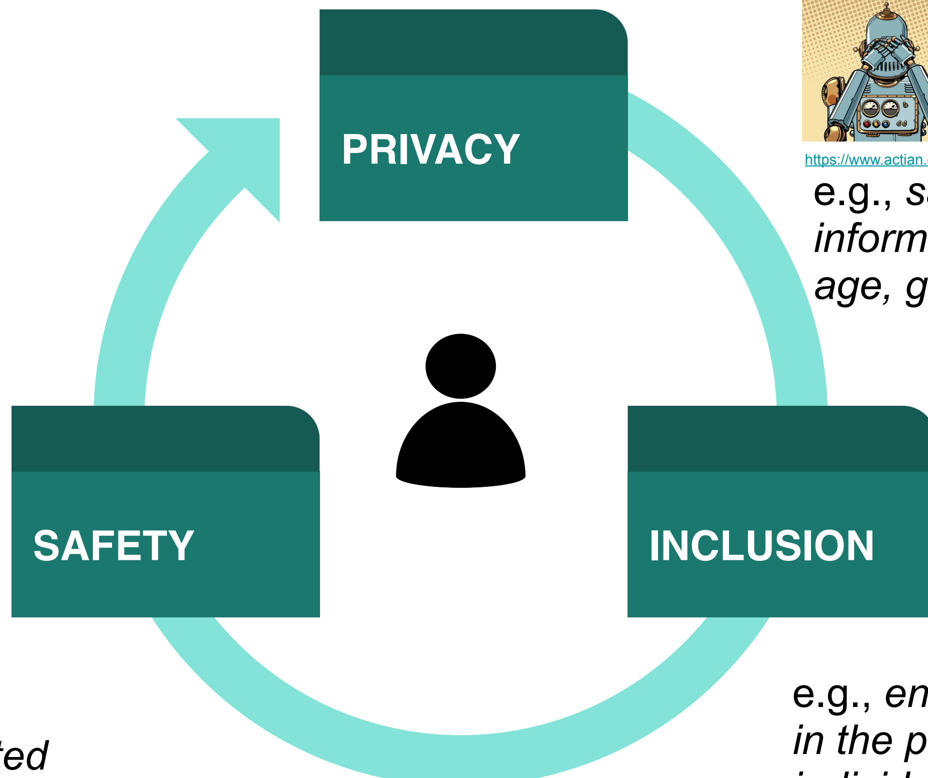
Sensitive features

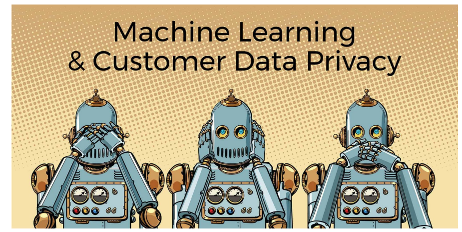
- Gender
- Age
- Ethnicity
- Language content

CARE/DEPTH: Behavioral Modeling in Human Interactions
https://sail.usc.edu/care

7

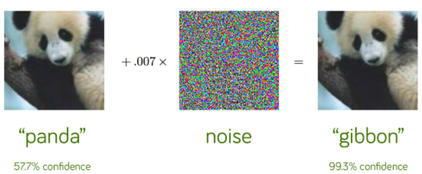
Elements of
Trustworthy Human-centered Machine Intelligence





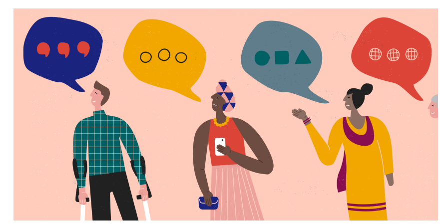
https://www.actian.com/company/blog/laymans-guide-to-machine-learning-and-customer-data-privacy/

e.g., *safeguard desired information to protect (PII: age, gender, race, etc.)*



https://towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa

e.g., *protect against malicious attacks Of vulnerable protected variables*



https://design.google/library/designing-global-accessibility-part-1/

e.g., *enable broad access in the presence of individual heterogeneity*

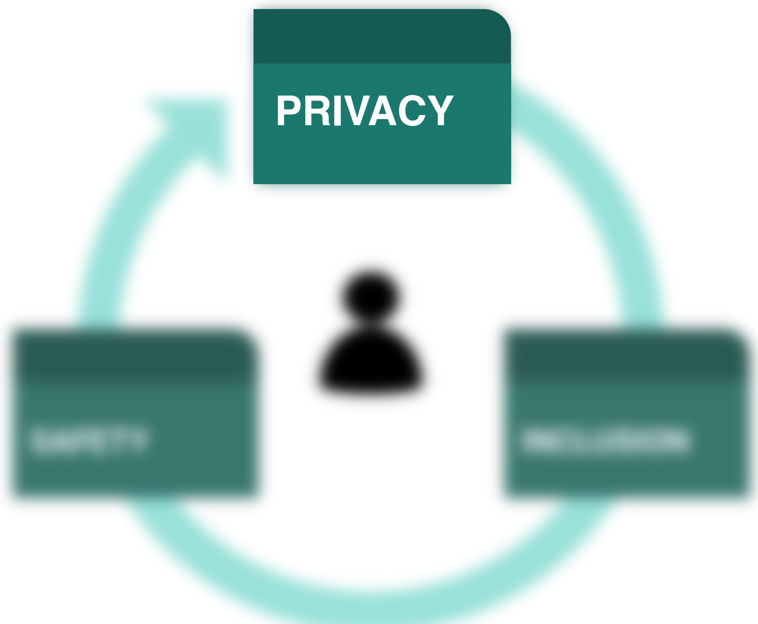
Rest of the talk

Our initial efforts in human-centric trustworthy AI:

1. **Privacy**
   a. Sensitive attribute obfuscation for speech emotion recognition
   b. Federated human activity detection from wearable sensors
2. **Inclusion**
   a. Fairness in automatic speaker verification
3. **Safety**
   a. Exploring strategies for defense against adversarial attacks in speaker recognition

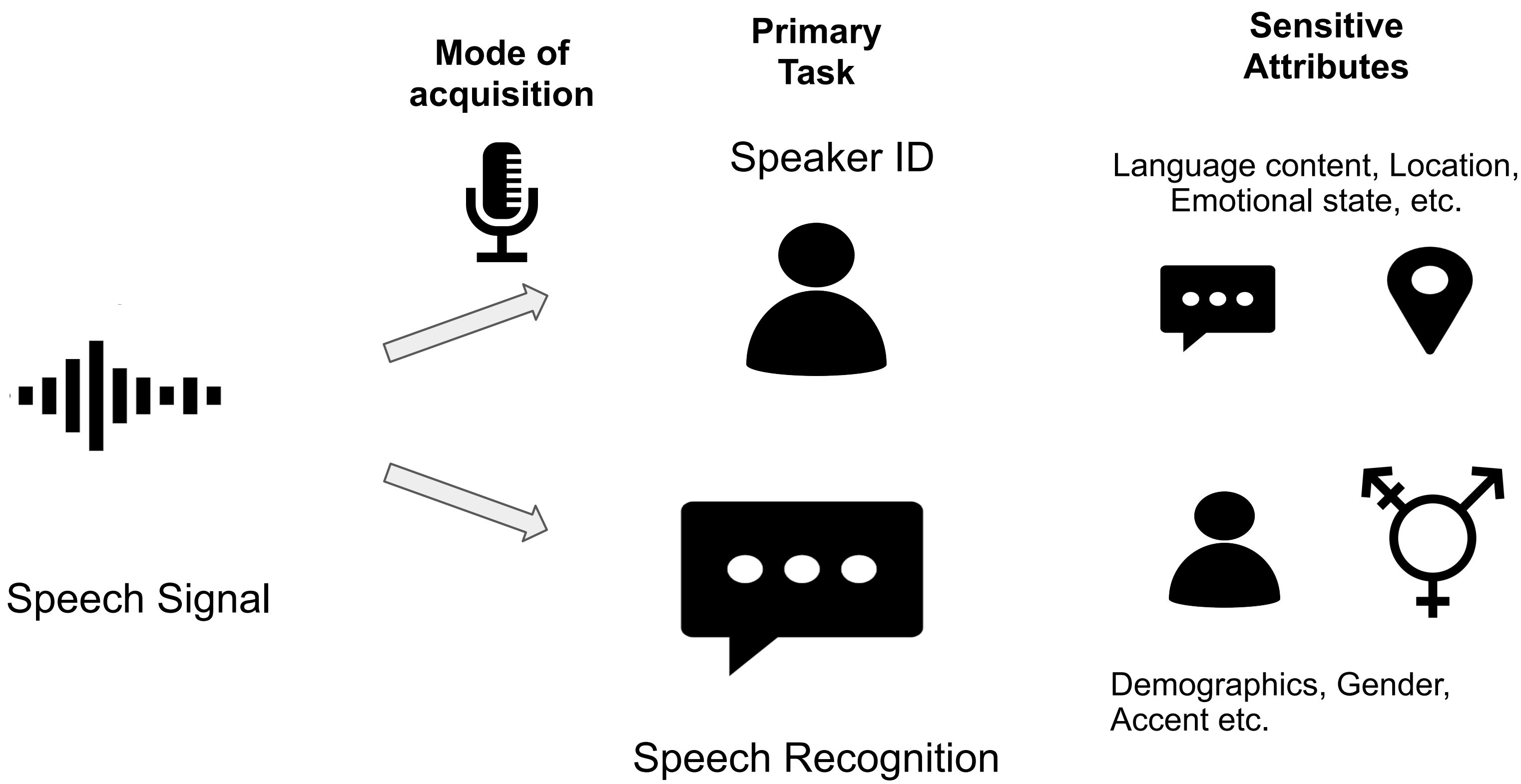
Conclusion: Summary of active/ongoing research threads and milestones

Elements of Trustworthy AI: PRIVACY



Privacy in human centric applications

Privacy is personal and context dependent



Privacy in human centric applications

Newer ways of sensing -e.g., wearable sensors, IoT devices- have varying privacy demands

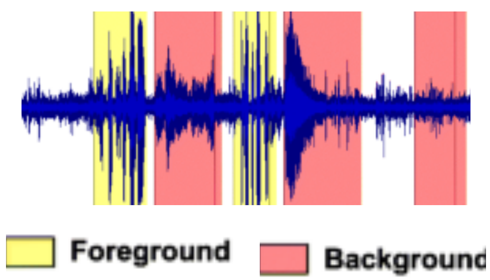
Egocentric sensing for stress regulation

Speech Signal

Mode of acquisition

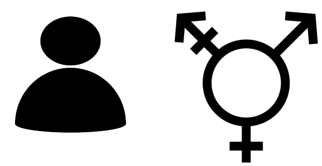
Primary Task

Detect foreground speech



Detect desired affect state

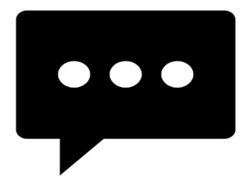
Common sensitive attributes



location, demographics, gender

Task-dependent sensitive attributes

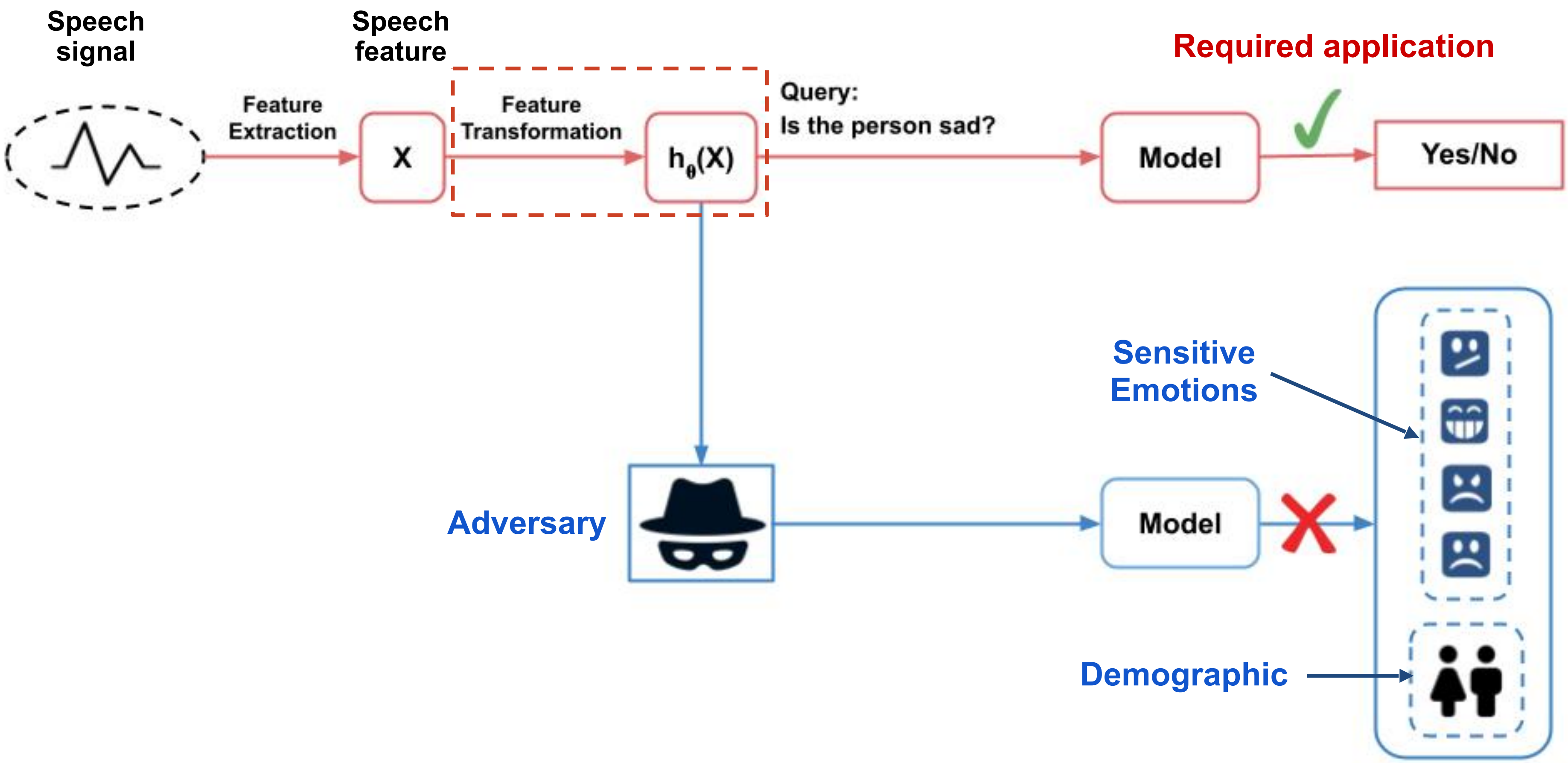
Specific emotional states



Language content

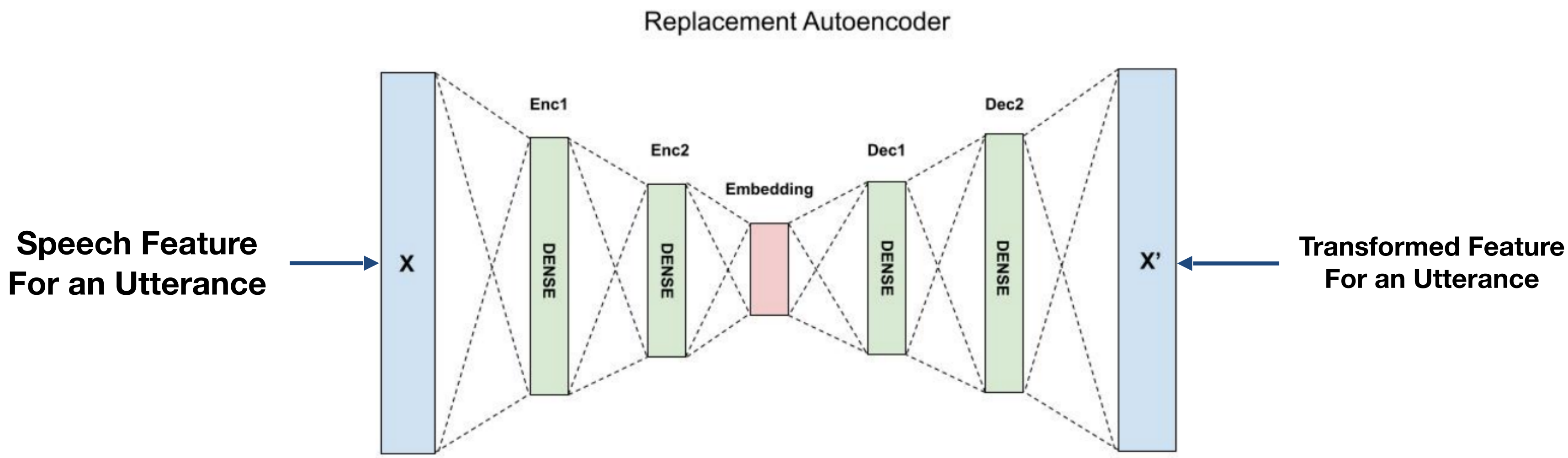
12

Attribute inference in Speech Emotion Recognition (SER)



13

Sensitive emotion obfuscation in SER



[1] Malekzadeh, M., Clegg, R.G., Cavallaro, A. and Haddadi, H., 2020. Privacy and utility preserving sensor-data transformations. Pervasive and Mobile Computing, 63, p.101132.

Sensitive emotion obfuscation in SER

U (Utility score) $U = \dfrac{\text{Correct Predictions Using Both X and X}'}{\text{Correct Predictions Using Only X}}$

	X	X′	
Set of Inference	**F1**	**F1**	**U**
$N = \{neutral\}$	56.0%	43.0%	81.0%
$R = \{sad\}$	64.8%	61.0%	82.3%
$S = \{happy, angry\}$	78.5%	3.6%	1.8%

Required application

Sad ✓

Sensitive emotions

Happy

Angry ✗

Req={sad}

X

	N	R	S
N	492	50	247
R	167	128	68
S	136	18	570

X’

	N	R	S
N	614	77	98
R	192	139	32
S	594	43	87

Happy

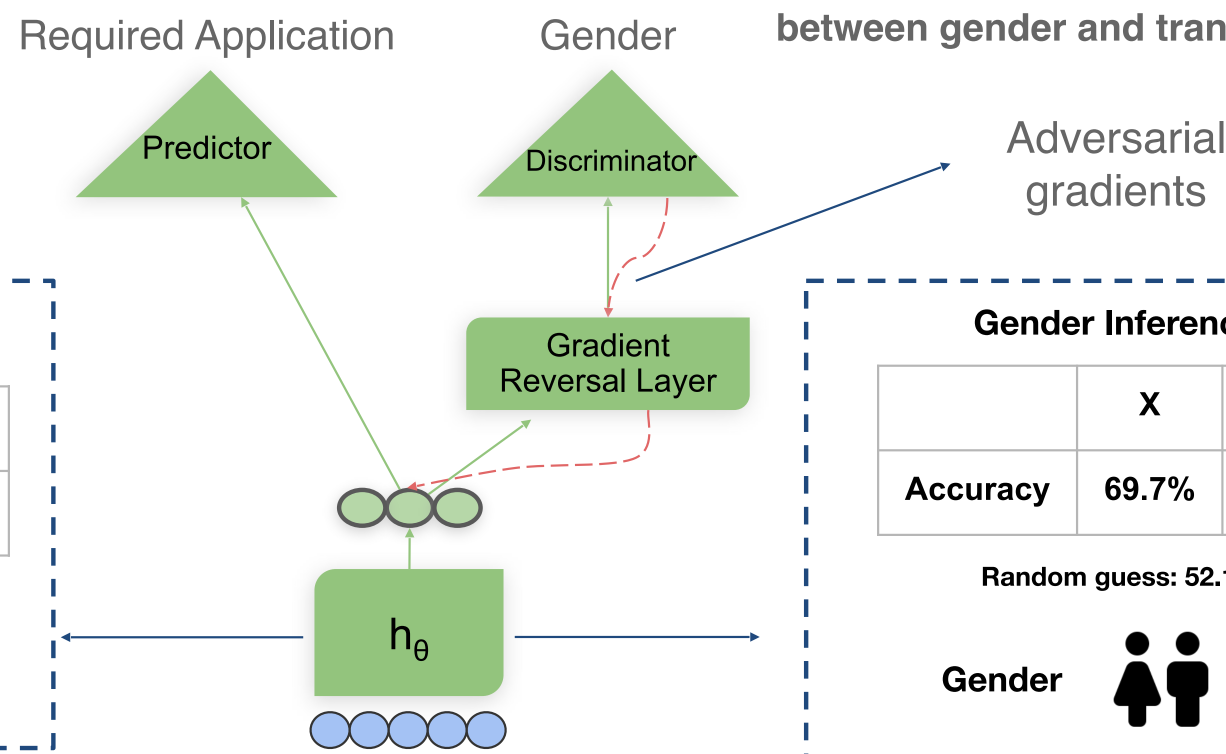
Angry

→ Neutral

Feng & Narayanan, Privacy and Utility Preserving Data Transformation for Speech Emotion Recognition, *ACII 2021*

Gender obfuscation in SER

Idea: Minimize mutual information between gender and transformed data



Required Application

	X	X'
F1	64.8%	58.4%

Sad ✓

Gender Inference

	X	X'
Accuracy	69.7%	55.9%

Random guess: 52.1%

Gender ✗

16

Tracking Individuals with Sensors (TILES) Study
https://sail.usc.edu/tiles/publications.php

- A multimodal human subject study on the clinical population
  - Understand workplace stressors
  - How do they affect wellbeing and productivity
- Nurse population
  - High burnout, high stress population
  - Work long shifts
- Passive egocentric sensing using portable, lightweight sensors

Study easy to run, replicate

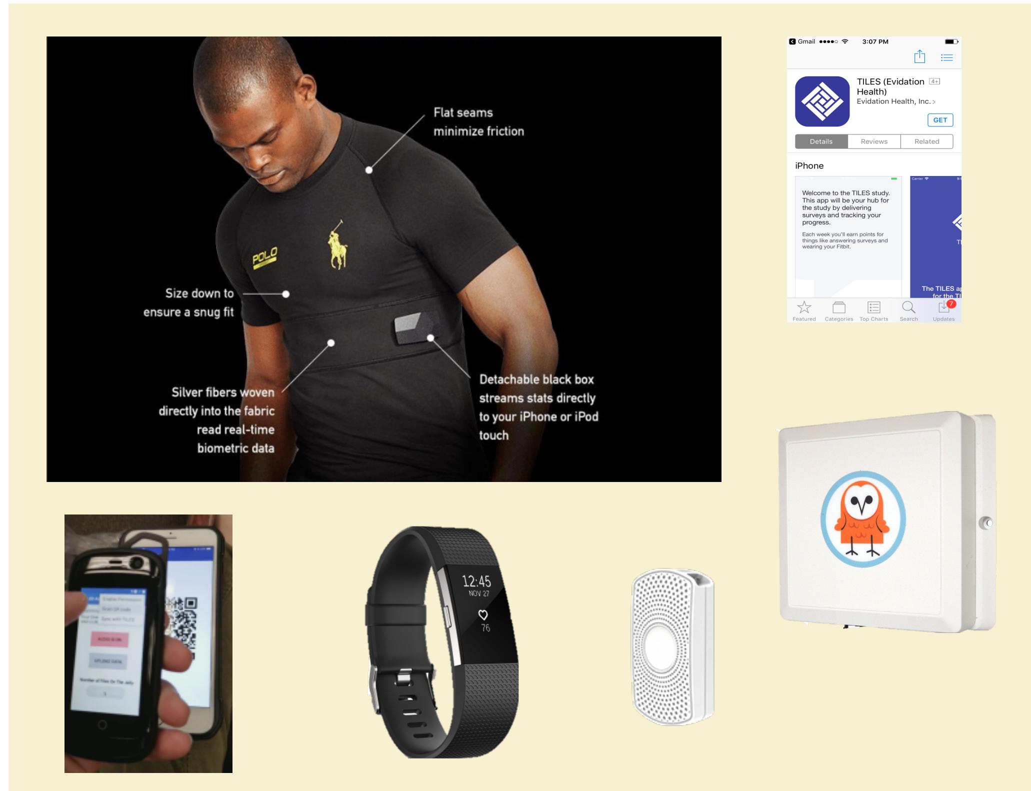
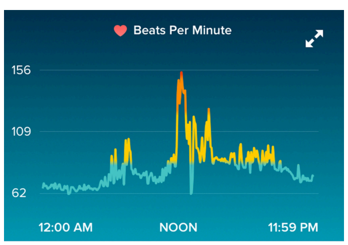


Fig *(Clockwise)* : ECG shirt, TILES study app, location sensors, humidity/temperature sensor, Fitbit and TILES Audio Recorder
17

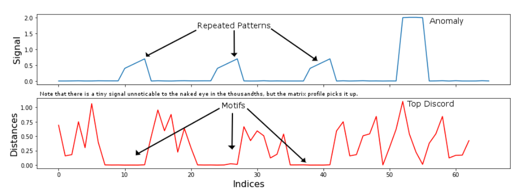
Federated human activity detection

Central Server

Wearable Sensor Data → Data Mining → Application Model



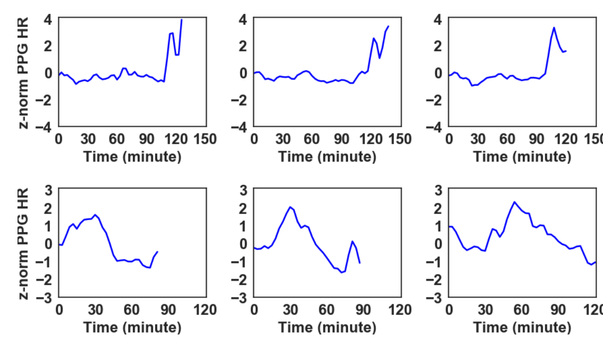
Heart rate time series



Motif Extraction
(Motif: repetitive pattern)

Wake up

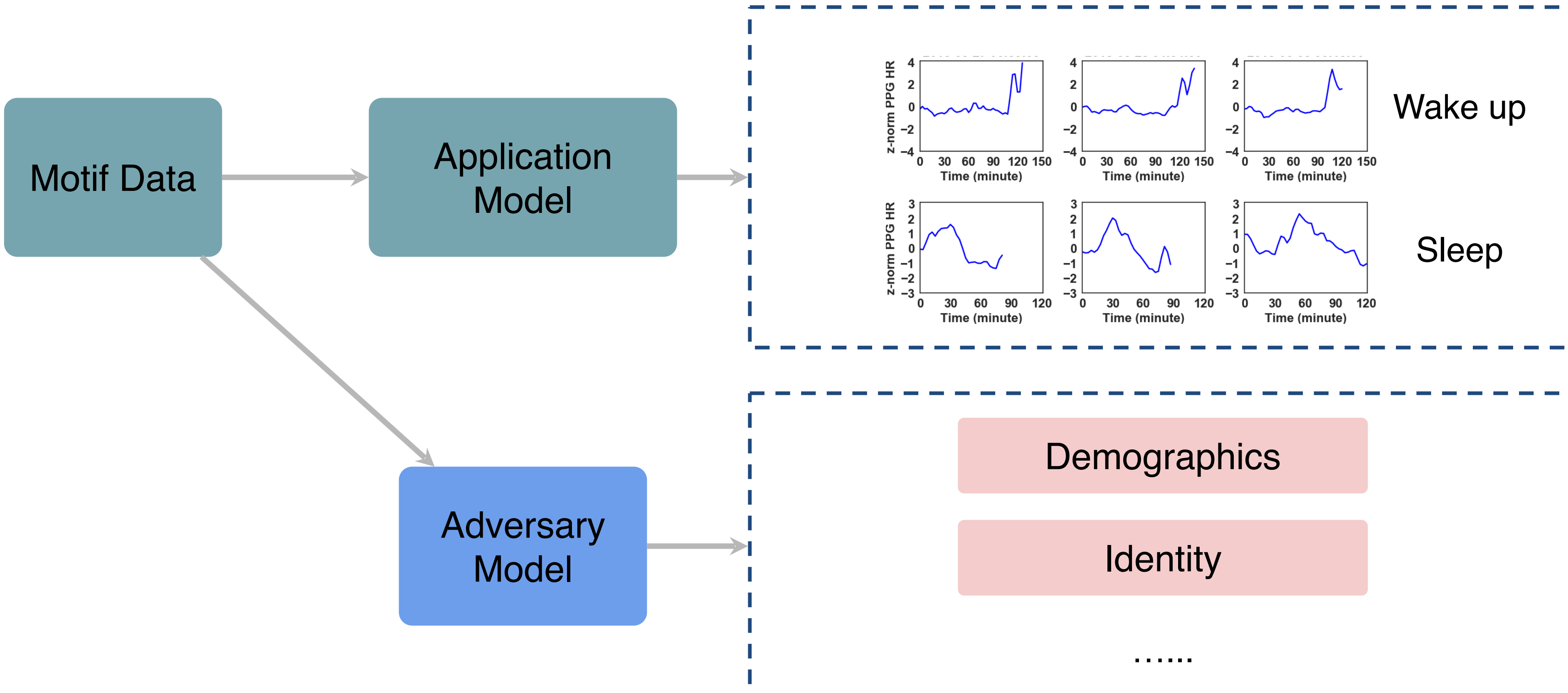
Sleep



......

18

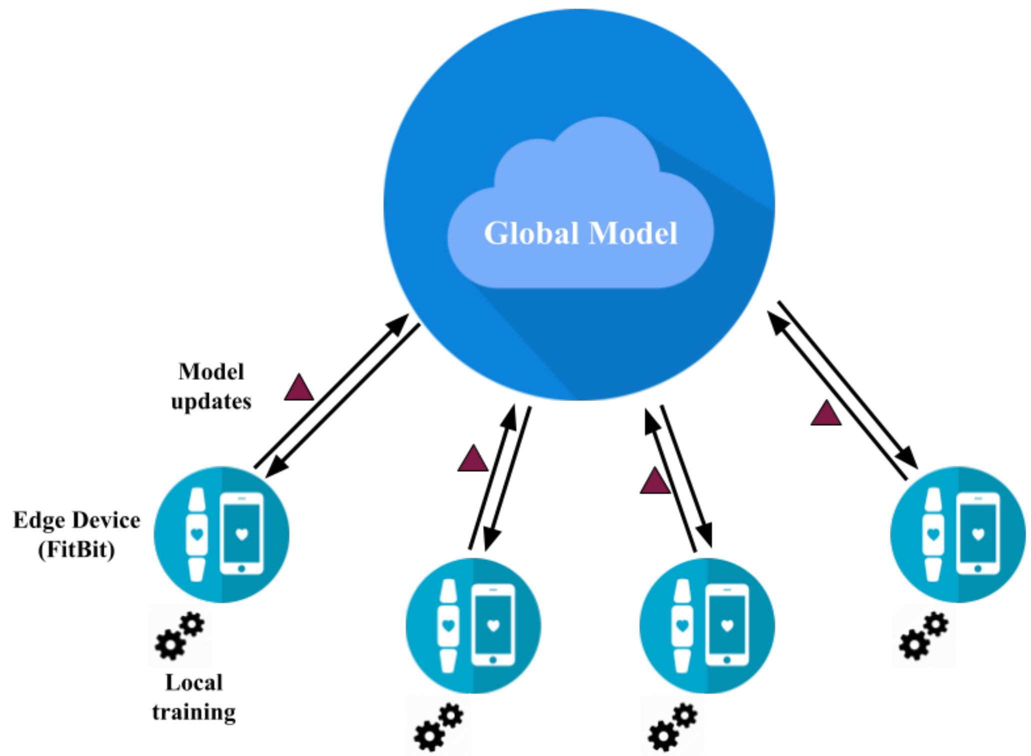
Federated human activity detection



Federated human activity detection

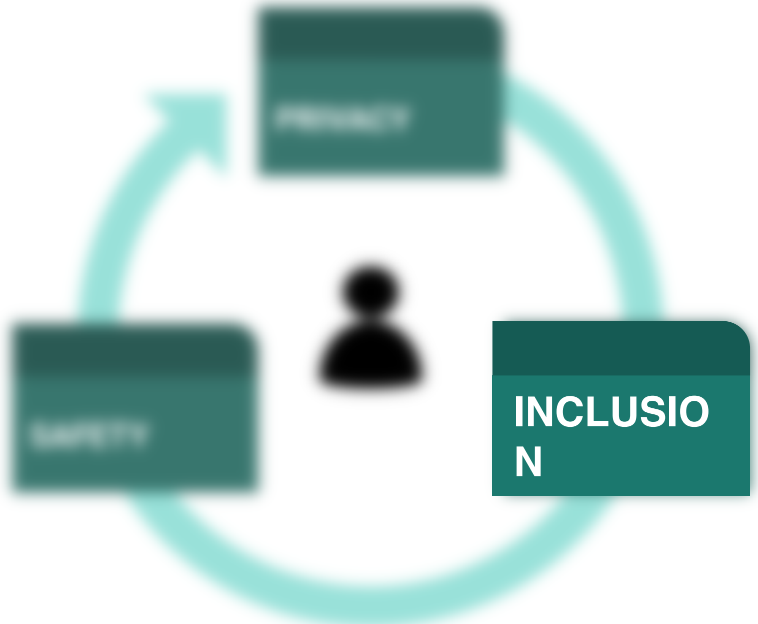
- Model trained on local data at each individual edge device.
- Model updates are transferred to the global server.
- Federated models (FedAVG) perform comparable to centralized model while preserving personal information.

Training setup	f1-score
Centralized training	0.81
Federated training	0.80



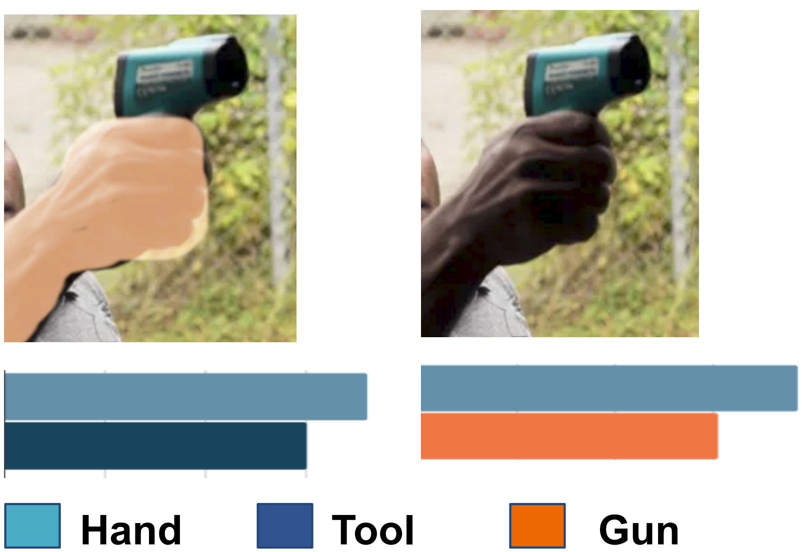
20

Elements of Trustworthy AI: INCLUSION



Artificial data-balancing or training on class-balanced data is seldom enough to ensure fairness and inclusion in learning

Fig 1: Outputs from an (withdrawn) object recognizer that labels same object differently in the context of skin tone



https://algorithmwatch.org/en/google-vision-racism/

Fig 2: ***She*** vs ***he*** occupations output from a word embedding system

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

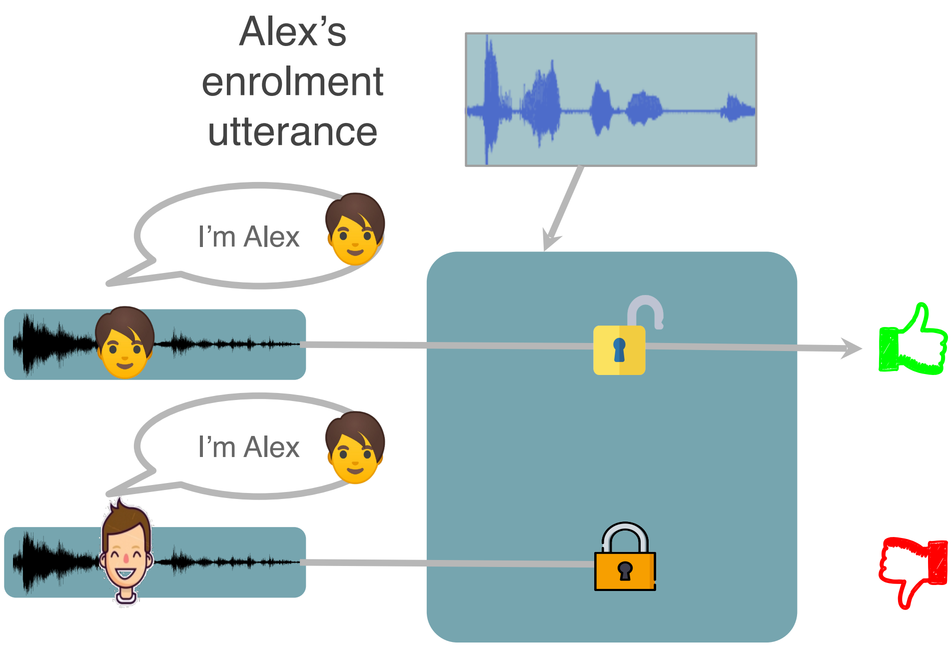
Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

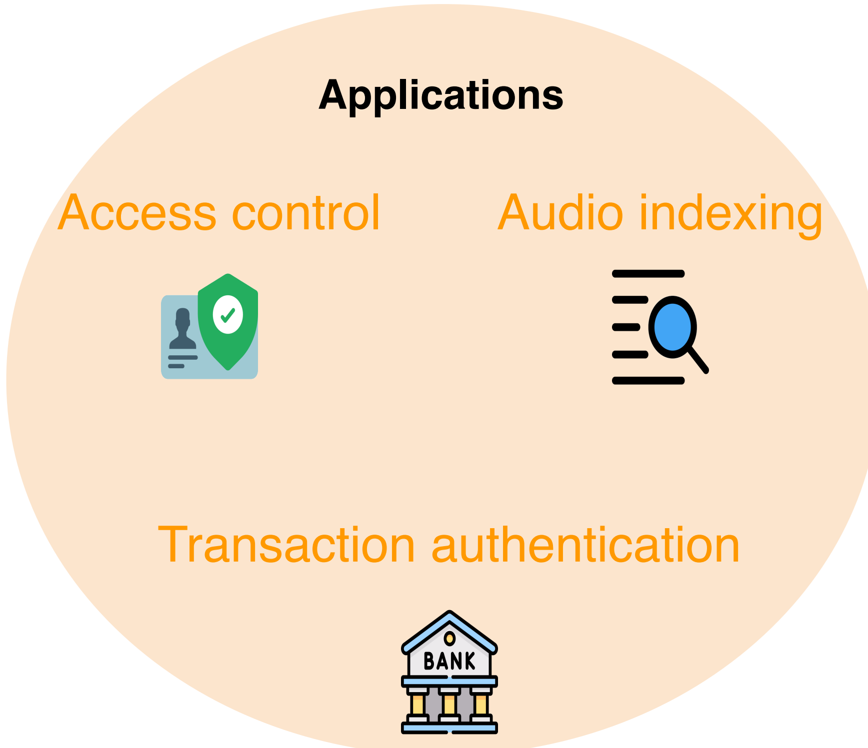
https://arxiv.org/abs/1607.06520

Fairness in Automatic Speaker Verification

Towards inclusive speech-centric interaction technologies

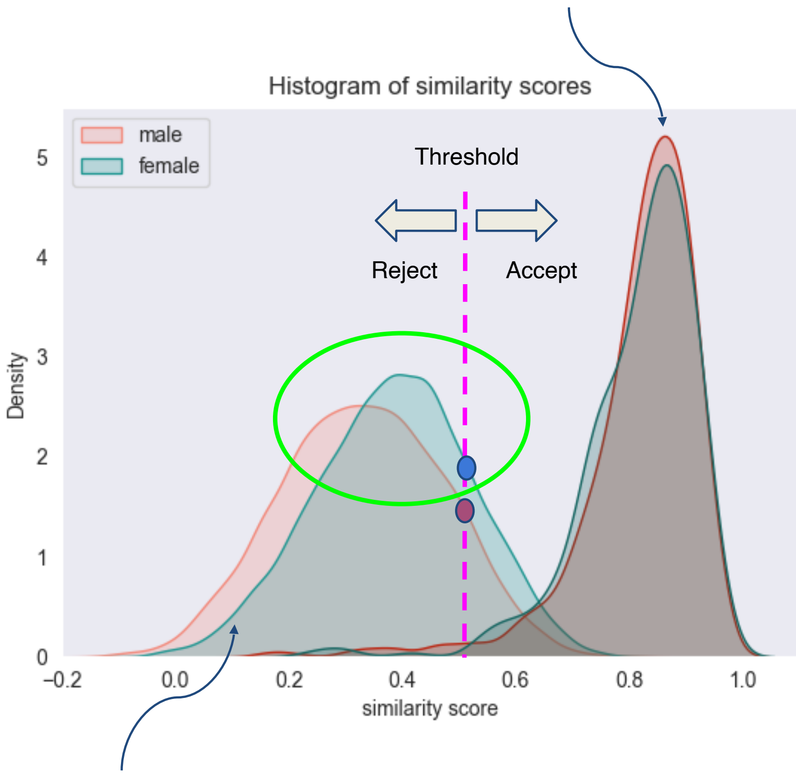


Speaker Verification



Fairness in Automatic Speaker Verification

Genuine: Test speaker same as claimed



Impostor: Test speaker different from claimed

- Similarity scores between enrolment and test utterance embeddings
- State-of-the-art speaker embeddings: x-vectors
- Models trained using gender-balanced data
- Evident skew in scores between male and female populations

Data balancing alone does not ensure fairness in ASV

24

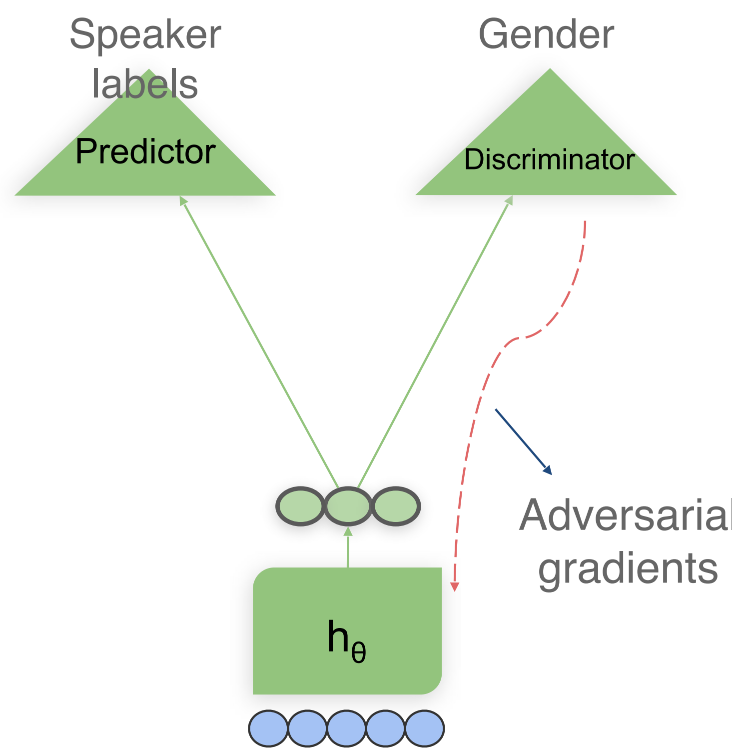
Fairness in Automatic Speaker Verification

Improving fairness using Adversarial Training

- Required attribute: Speaker id
- Sensitive attribute: Gender

Goal

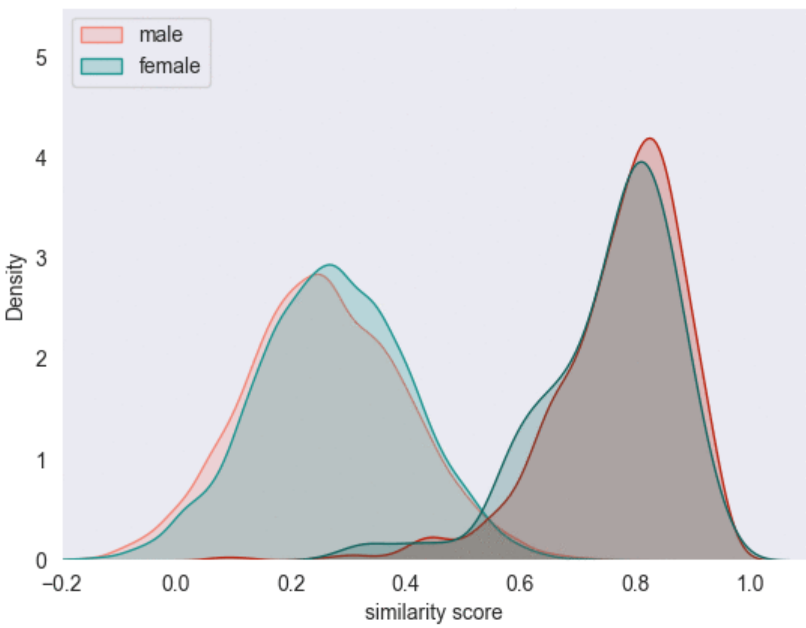
Learn speaker discriminative representations while discarding gender information



Fairness in Automatic Speaker Verification

Qualitative results

Histogram of similarity scores



Quantitative results (lower is better)

% EER	Male	Female	Delta	Overall
x-vector	**6.50**	**4.80**	1.70	5.82
AT	6.60	5.37	**1.23**	**6.05**

- ❏ Adversarial training mitigates skew in similarity scores
- ❏ Difference (delta) in EER between Male and Female populations reduced with Adversarial training

Future research directions

- ❏ Intersectionality: Age, Gender, Accent etc.
- ❏ Holistic evaluation metrics

Adversarial training succeeds in reducing gender bias in ASV

Elements of Trustworthy AI: SAFETY

Security risks in Machine Learning Systems

Training stage

Data poisoning

- ❑ Manipulate training data to learn incorrect correspondences
- ❑ Can compromise privacy and confidentiality

Inference stage

Adversarial attack

- ❑ Add malicious noise to input
- ❑ Force pre-trained system to make incorrect predictions
- ❑ Difficult to detect

Output attack

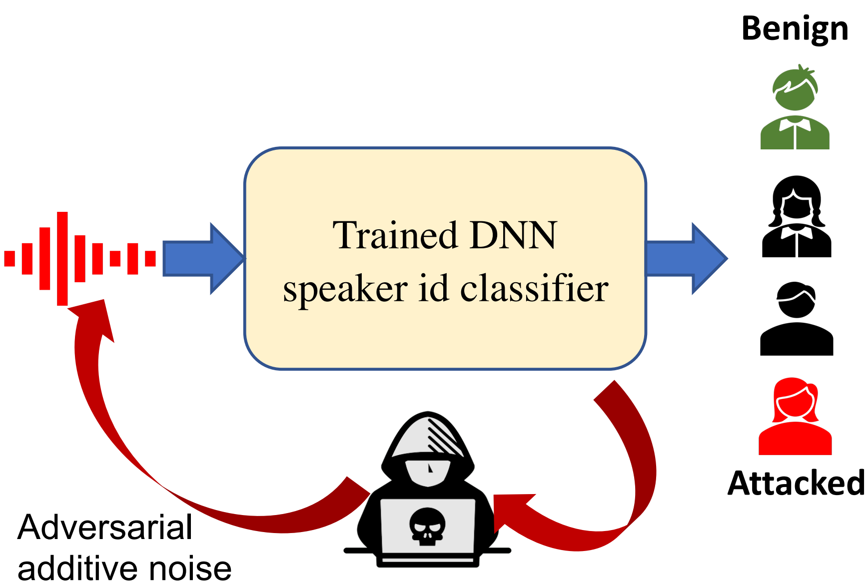
- ❑ Attacks result display setup
- ❑ Manipulate output prediction presentations

USC

Security of Speaker Recognition systems

SAIL

Adversarial attacks



- ❑ Malicious attacker crafts noise to modify DNN system predictions
- ❑ Added noise can be imperceptible
- ❑ Attacker can force classifier to output any desired incorrect prediction
- ❑ Compromise security of any DNN-based system
- ❑ Impacts trustworthiness

DNN-based systems (including speaker recognition) are vulnerable to adversarial attacks

Jati, A., Hsu, C. C., Pal, M., Peri, R., AbdAlmageed, W., & Narayanan, S. (2021). Adversarial attack and defense strategies for deep speaker recognition systems. *Computer Speech & Language*, *68*, 101199.

Security of Speaker Recognition systems

FGSM attack

- ❑ Uses sign of gradient to push prediction away from true output

Defense against adversarial attacks^

- ❑ Use adversarially added noise during speaker recognition model training
- ❑ On-the-fly data augmentation

Speaker recognition performance

% accuracy	Benign	FGSM* attack
Without defense	**94.0**	25.0
With defense	92.0	**73.0**

FGSM: Fast Gradient Sign Method

Defense significantly improves performance under attack

Minimal impact on benign performance

*Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
^Jati, A., Hsu, C.C., Pal, M., Peri, R., AbdAlmageed, W. and Narayanan, S., 2021. Adversarial attack and defense strategies for deep speaker recognition systems. *Computer Speech & Language*, *68*, p.101199.

Summary/Milestones: Toward trustworthy human-centered machine intelligence

Project Title	Description	Research Area	Milestones
Emotion Recognition	Privacy and utility preserving data transformation for Speech Emotion Reco.	Privacy preserving representation learning	1. Learn data transformations for speech to hide sensitive emotions 2. Suppress demographic info. (e.g, gender)
	Using intrinsic relationships between label/annotation noise to improve emotion recognition	Federated learning, learning with noisy labels	1. Incorporating annotator reliability/label noise into federated learning setup 2. Having a central model that is generalizable enough for different datasets
Egocentric wearable health attribute modelling	Modelling health attributes from in-situ egocentric wearable audio	Wearable sensing, Egocentric audio, Behavior modelling	1. Learning centralised audio representations through self-supervised learning 2. Leveraging the representations to model clinically relevant attributes in privacy preserving fashion

Collaborators:
Anil Ramakrishna,
Rahul Gupta

amazon
Trustworthy AI Alexa NU group

USC

Summary/ Milestones

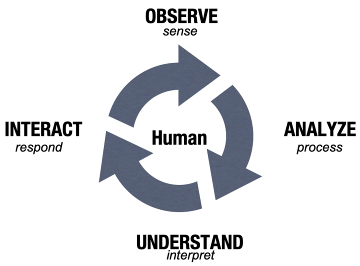
SAIL

Project Title	Description	Research Area	Milestones
Federated human activity recognition	Understanding human activity using physiological signals, in trustworthy fashion.	Personalisation, Federated learning	1. Extended federated learning to bio-behavioural signals 2. Showed effectiveness of personalisation in label sparse regimes
Fairness in speaker verification	Understand issues of biases and fairness in ASV systems at different stages of the pipeline.	Inclusion, Fairness	1. Used adversarial training methods to improve fairness 2. Incorporate intersectional fairness by jointly modelling multiple demographic attributes
Defense against adversarial attacks on speech technology systems	Improve reliability of speech technologies such as ASR and Speaker recognition by defending against adversarial threats	Security, Robustness against adversarial attacks	1. Developed defenses against black-box and white-box adversarial attacks using preprocessors 2. Improve defense performance by incorporating adversarial training techniques

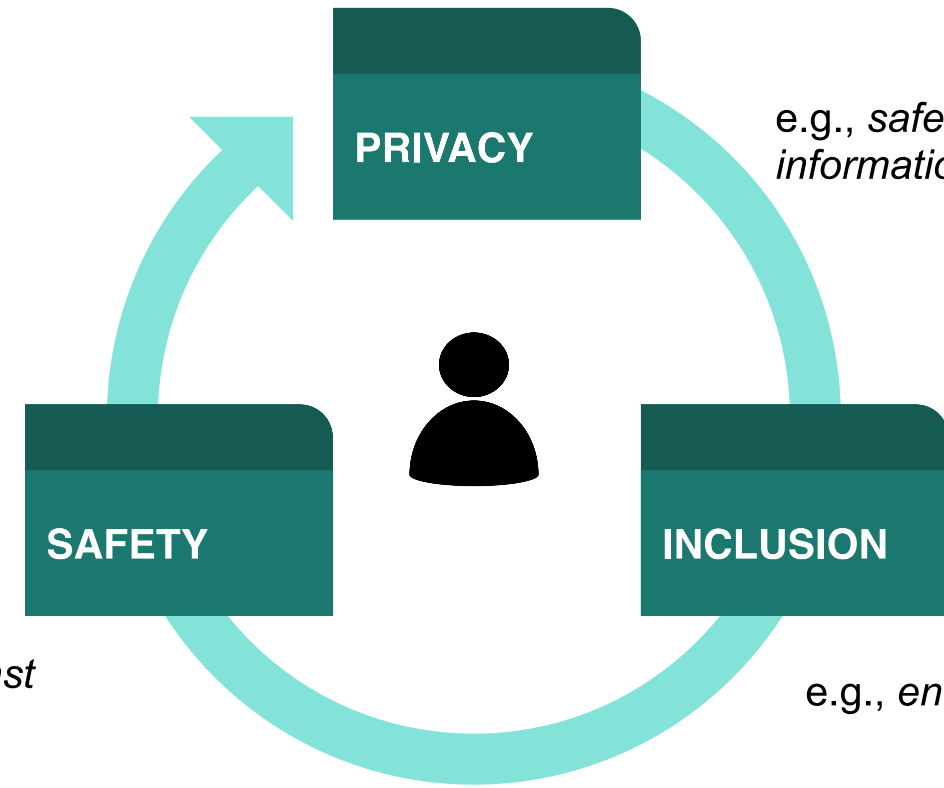
Collaborators:
Anil Ramakrishna,
Rahul Gupta

amazon
Trustworthy AI Alexa NU group

Elements of
Trustworthy Human-centered Machine Intelligence



Strive for an integrative approach across the machine intelligence ecosystem



e.g., *safeguard protected information*

e.g., *protect against malicious attacks*

e.g., *enable broad access*