Human-Centered Artificial Intelligence

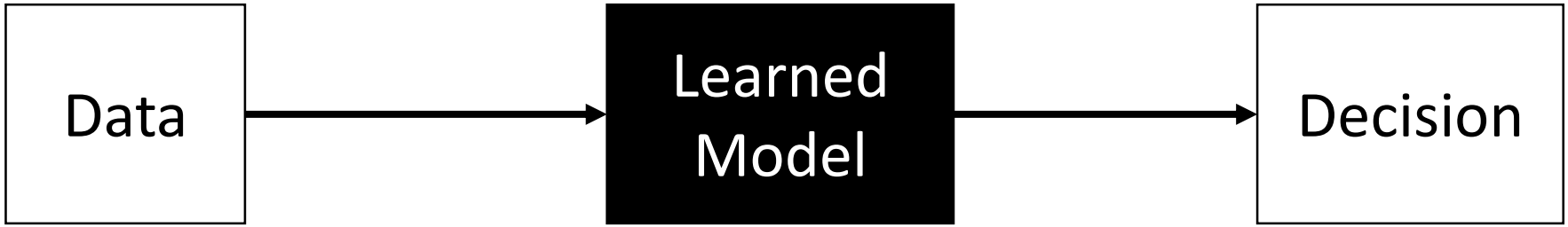
Human-Centered AI in the 21st Century

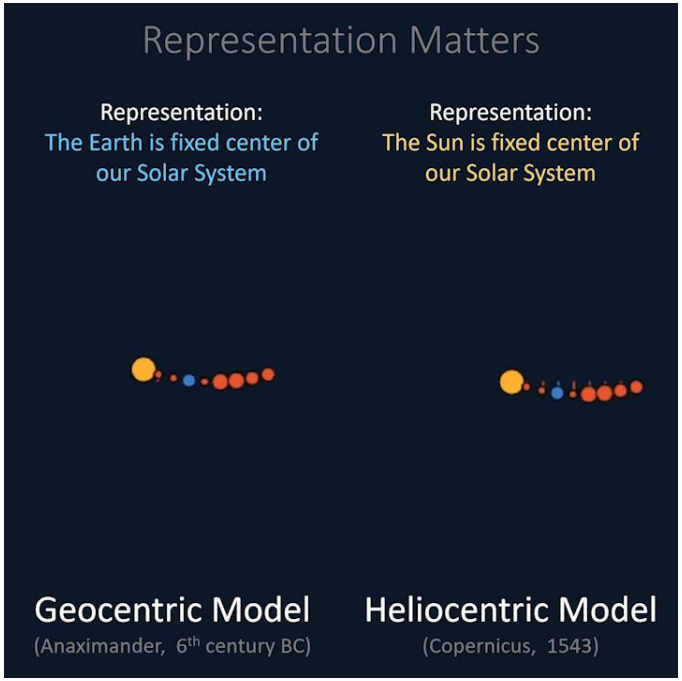
- **Prediction:** AI systems will become more & more learning-based
- **Corollary:** Smarter AI is achieved through:
  - Better machine teaching (optimizing data annotation) ← human-centered
  - Better machine learning (optimizing learning algorithms) ← current focus

- **Ethical and safety implications of learning-based AI systems:**
  - AI will not be provably safe → Human supervision is required
  - AI will not be provable fair → Human supervision is required
  - AI will not be perfectly explainable → Human supervision is required

- **Solution:** Human-Centered AI
  - Deep integration of humans into the data annotation process
  - Deep integration of humans into real-world operation

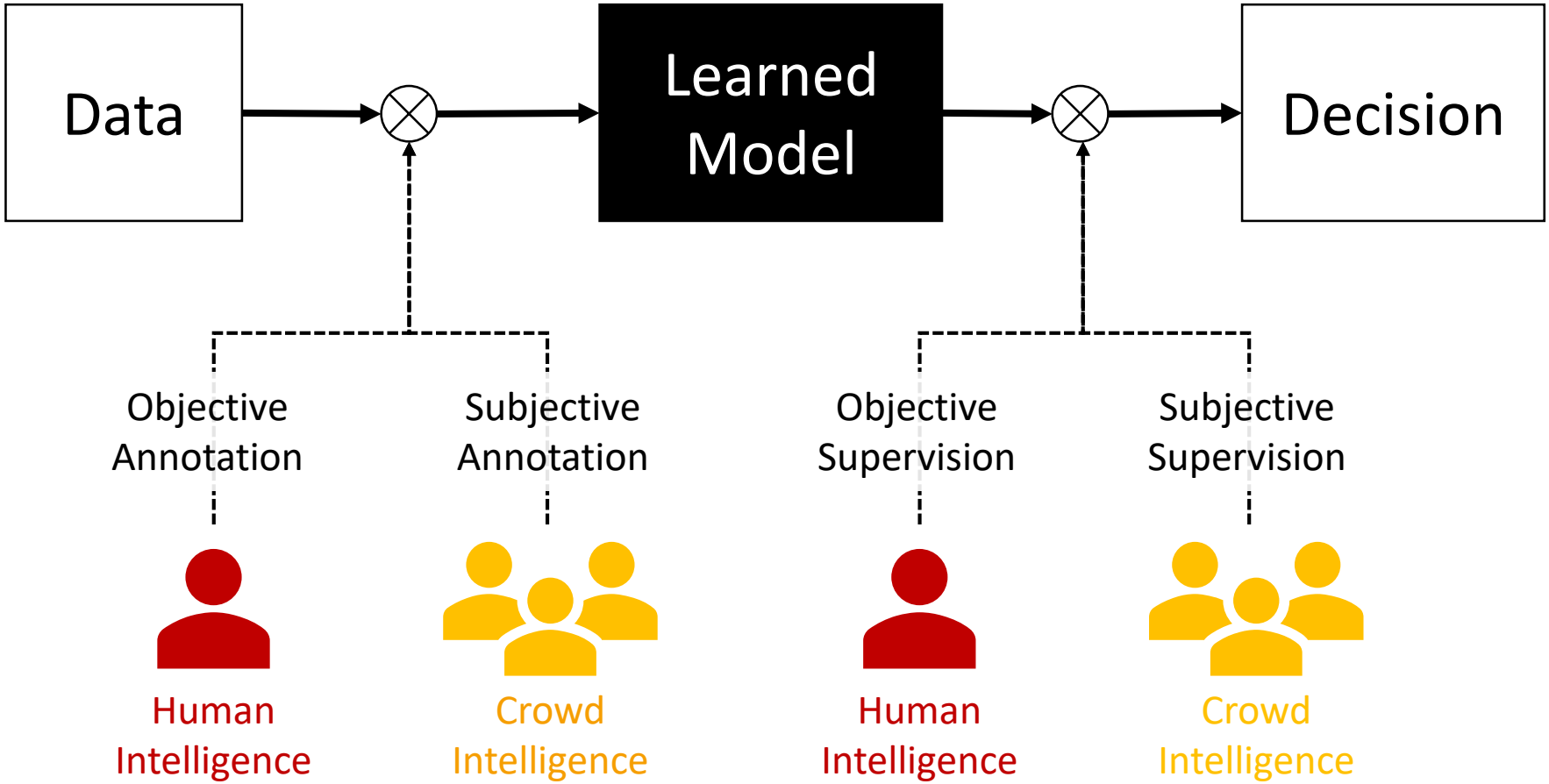
Deep Learning with Human Out of the Loop





Objective and Subjective Annotation/Supervision for

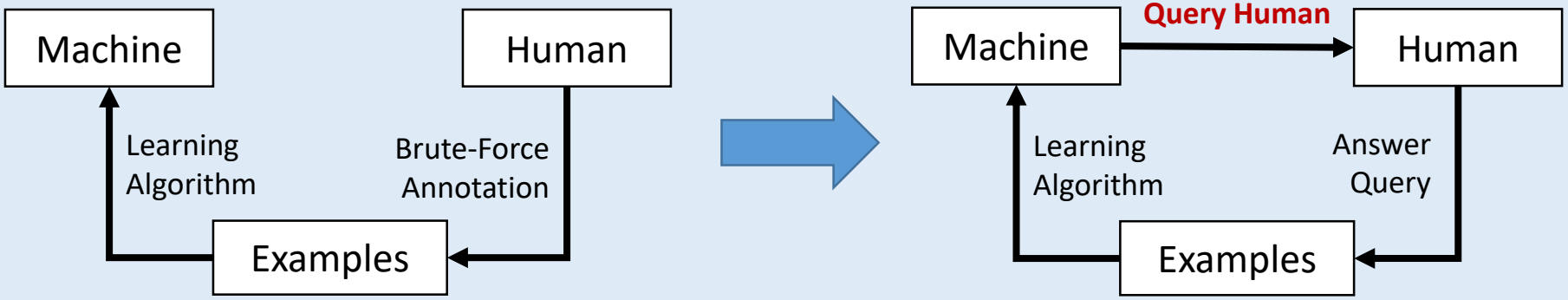
Deep Learning with Human in the Loop



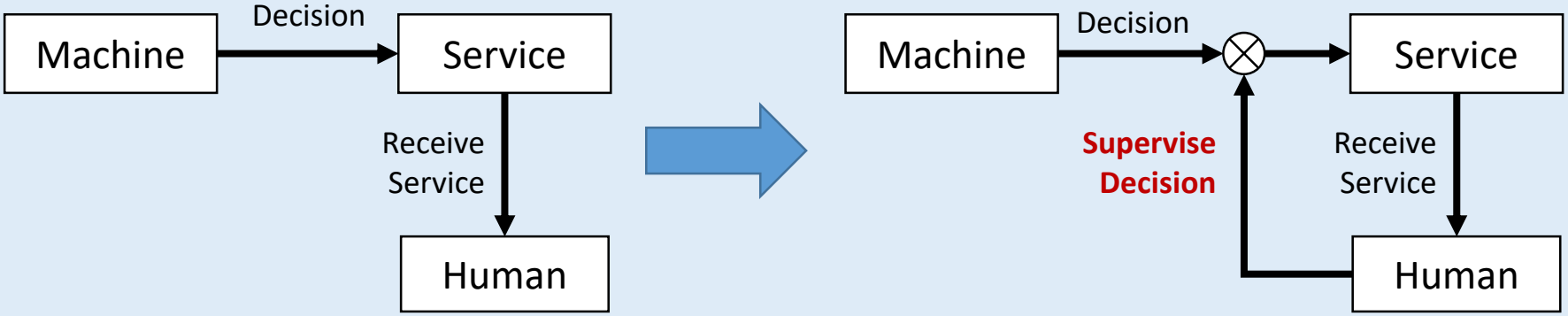
Human-Centered AI:

Integrate Human Intelligence* into the loop of
(1) training and **(2)** real-world operation of AI Systems

Training Process



Real-World Operation



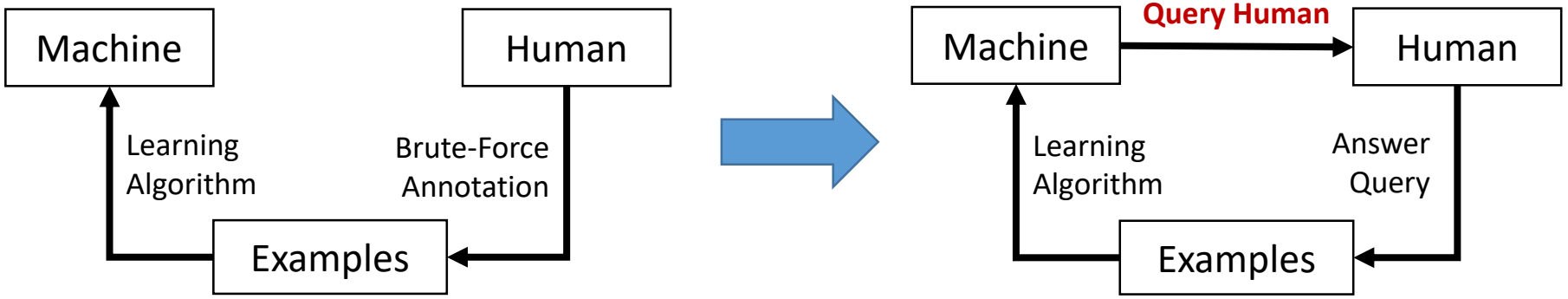
* The "better angels" of human intelligence (as Abraham Lincoln said in addressing the common force that unites us)

Human-Centered AI Grand Challenges

- Human-Centered AI during Learning Phase
  - Machine Teaching:
    Methods for efficient supervised learning
    (Improve annotation and learning algorithms)
  - Human-in-the-Loop Reward Engineering:
    Encoding human values into learning process
- Human-Centered AI during Real-World Operation
  - Human Sensing:
    Methods for perceiving the human state (physical, mental, social)
  - Human-Robot Interaction Experience:
    Methods for an immersive, meaningful interaction
  - AI Safety:
    Methods for effective supervision of machines (ethics & safety)

Machine Teaching and Efficient Supervised Learning

- Near-Term Directions of Research:
  - **Annotation:** Learn by asking for a few examples instead of being given a LOT of examples
  - **Algorithms:** Active learning, data augmentation, one shot learning, zero shot learning, transfer learning, self-play.
- Example Grand Challenges:
  - Win COCO Object Detection Challenge training only on Wikipedia (text, images, & ability to comment)
  - Achieve 0.3% error on MNIST using only 1 example of each digit



Human-in-the-Loop Loss/Reward Function

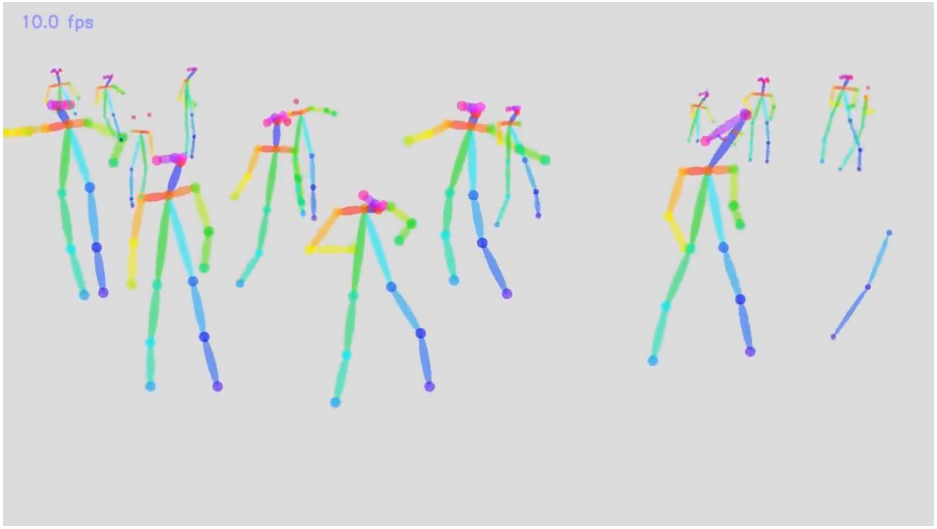
- Near-Term Directions of Research:
  - Reward engineering and tuning by human
  - Incorporate human subjective values into learning process
- Example Grand Challenge:
  - Artificial intelligence based representative democracy (replace congress with a recommender system)

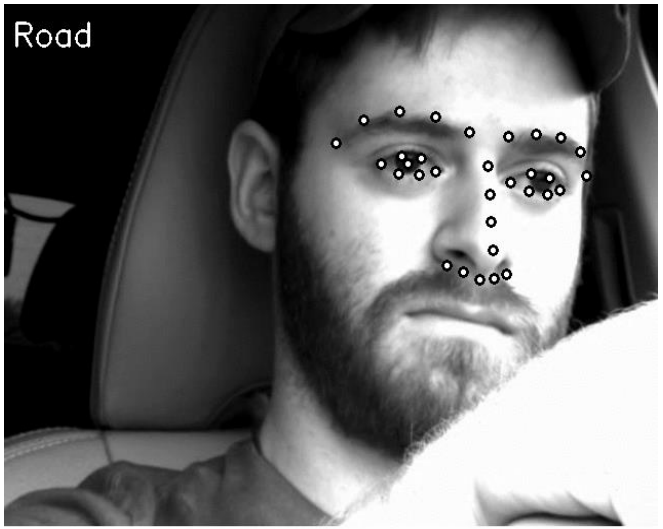
Human

AI (Deep RL Agent)

Human Sensing

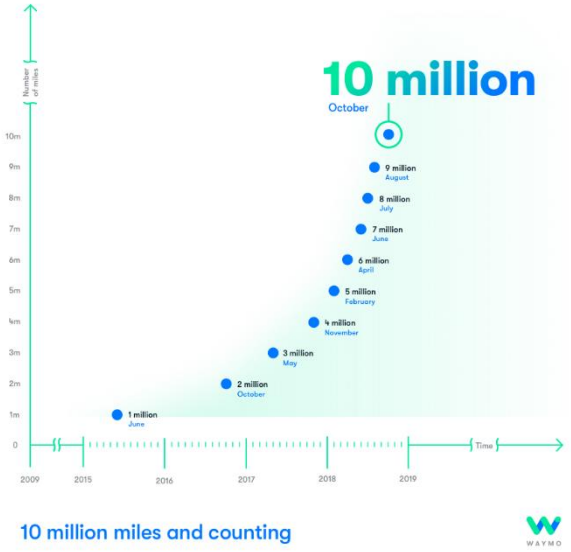
- Near-Term Directions of Research:
  - Perception of physical, mental, social state & context from video/audio
  - Long term personalization
- Example Grand Challenge:
  - Perceive if a person wants company or to be left alone with 95% accuracy after 30 days of interaction

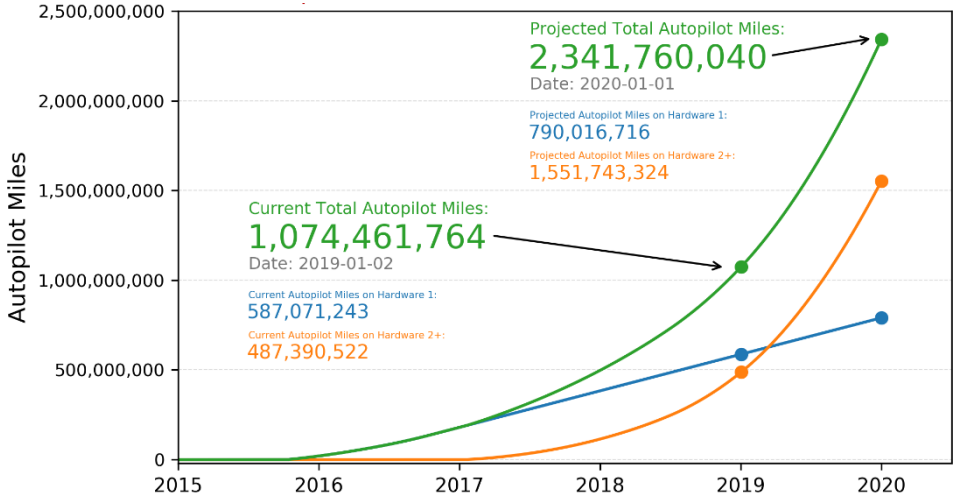




Human-Robot Interaction Experience

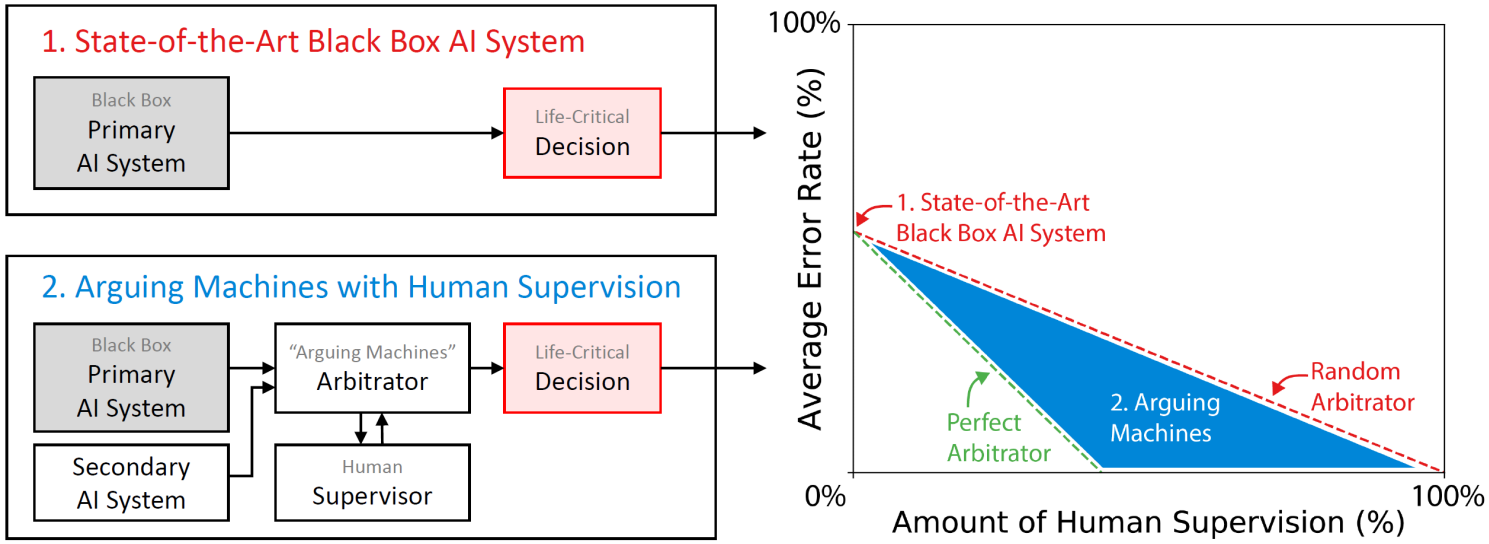
- Near-Term Directions of Research:
  - Collaborative interaction that is fulfilling and valuable to both parties
- Example Grand Challenges:
  - 100 billion miles of semi-autonomous driving
  - Alexa Prize Socialbot Grand Challenge





AI Safety

- Near-Term Directions of Research:
  - Human supervision of AI decisions in terms of safety and ethics
  - Avoid catastrophic actions in exploration or unintended consequences of reward function
- Example Grand Challenges:
  - Uncertainty estimate matches (within 5%) the error rate on large-scale image classification problem (that contains examples outside the distribution of the training set)



- Human-Centered AI during Learning Phase
  - Machine Teaching:
    Methods for efficient supervised learning
    (Improve annotation and learning algorithms)
  - Human-in-the-Loop Reward Engineering:
    Encoding human values into learning process
- Human-Centered AI during Real-World Operation
  - Human Sensing:
    Methods for perceiving the human state (physical, mental, social)
  - Human-Robot Interaction Experience:
    Methods for an immersive, meaningful interaction
  - AI Safety:
    Methods for effective supervision of machines (ethics & safety)

Series of lectures on aspects of the above will be released on:

deeplearning.mit.edu

Deep Learning for Understanding the Human

- General Applications
  - Face Detection
  - Gaze Estimation
  - Face Recognition
  - Activity Recognition
  - Emotion Recognition
  - Body Pose Estimation
  - Gesture Recognition
  - Speech Recognition
  - Recommendation Systems
  - Natural Language Understanding
  - Dialogue Systems
  - Emotion Recognition
- Special Applications
  - Glance Classification
  - Cognitive Load Estimation
  - Human Vision Simulation

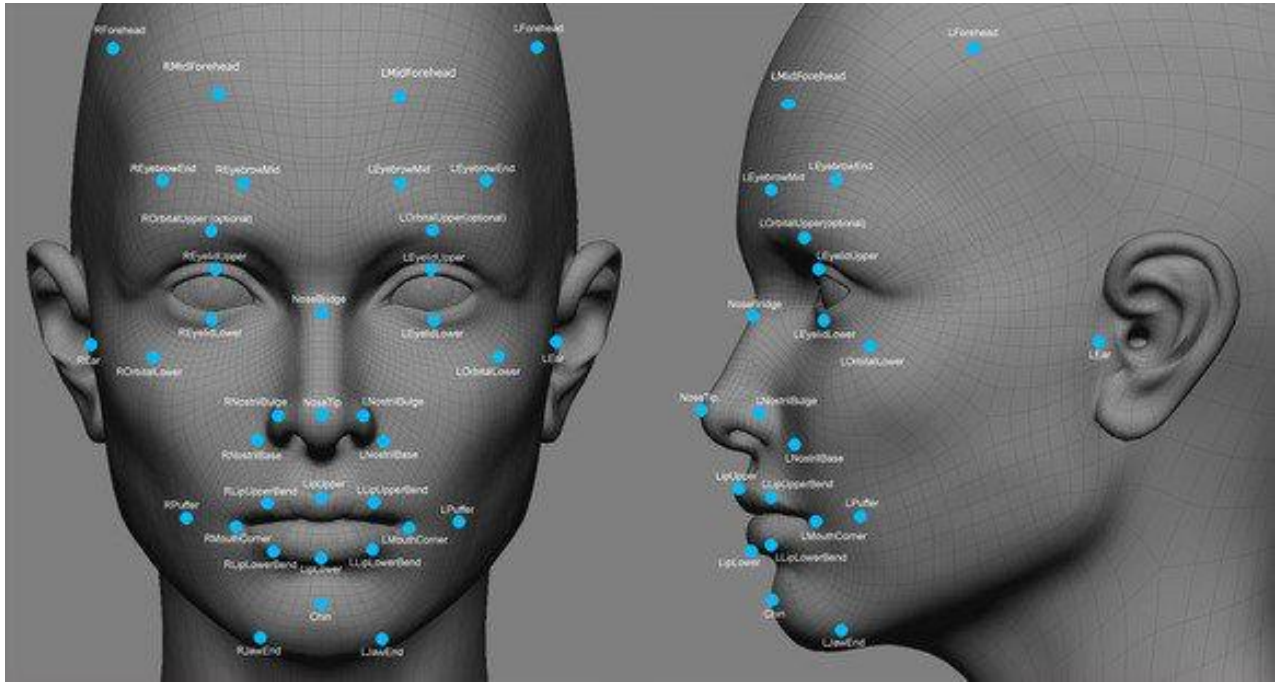
Face Recognition

Section Structure

- Motivation, Description, Current and Future Impact
- Paper 1: “Old School” Seminal Work
- Paper 2: Early Progress in the Field
- Paper 3: Recent Breakthrough
- Paper 4: State-of-the-Art
- Paper 5: Possible Future Direction
- Open Problems and Future Research

Skipping historical long historical context for today, will return in lecture dedicated to the topic.

Description and Motivation



- **What is it?**
  Recognize and identify different human faces
- **Why is it important?**
  Identification (e.g. FaceID), social media and photography, security and policing, …

Face Recognition: Why It's Hard

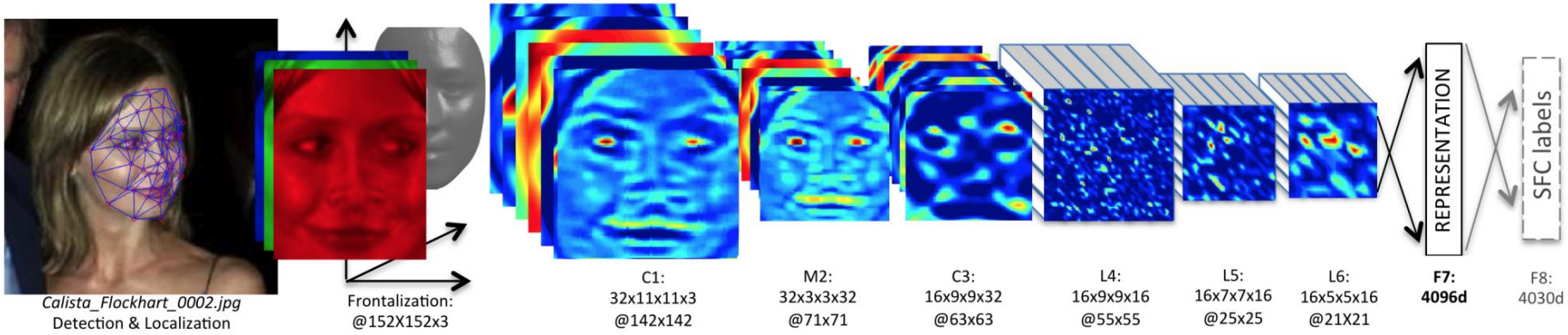
- As a subset of image recognition, it also suffers from common vision problems such as illumination, occlusion, scale
- **Visual similarity:** Faces can be very similar
- **Data:** Although a lot of face data is available, it's hard to get a lot of data from the same identity
- **Variation**: Recognition should be invariant to glasses, hair-style, makeup, aging, beard, glasses, aging, weight gain/loss…
- **Accuracy:** Requirement of high-accuracy for application related to security

Future Impact

- **Future Impact:** User-friendly identification method.
- **Utopia:** Use your face as passport, ID card, credit card, everything, national security is enhanced by public camera to detect criminals.
- **Dystopia:** No privacy, people are watched anytime, your face is used to make a robot just identical to you.
- **Middle path:** FaceID to unlock your phone.

Paper 3: Recent Breakthrough

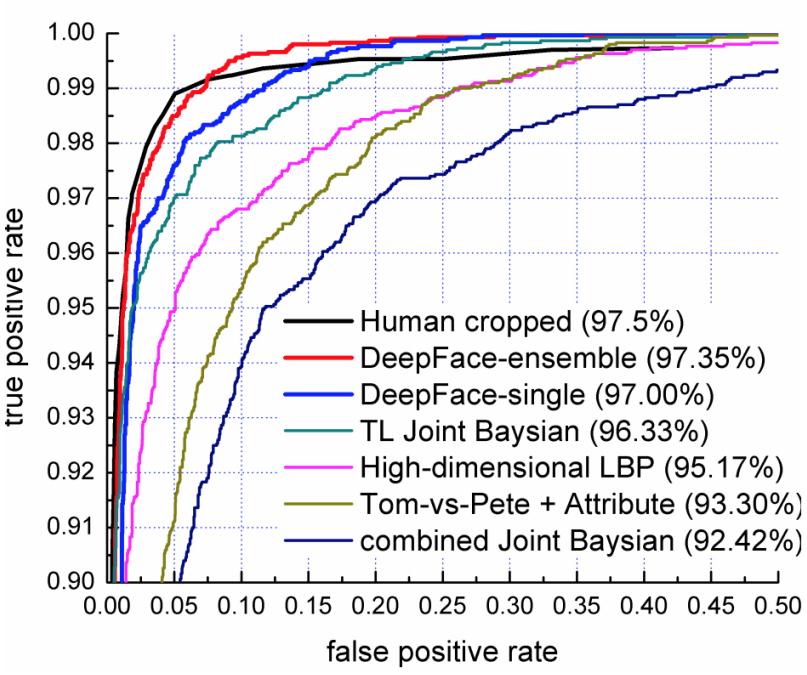
2014. CVPR. Taigman, Yang, Ranzato, & Wolf (Facebook)
"DeepFace: Closing the Gap to Human-Level Performance in Face Verification"

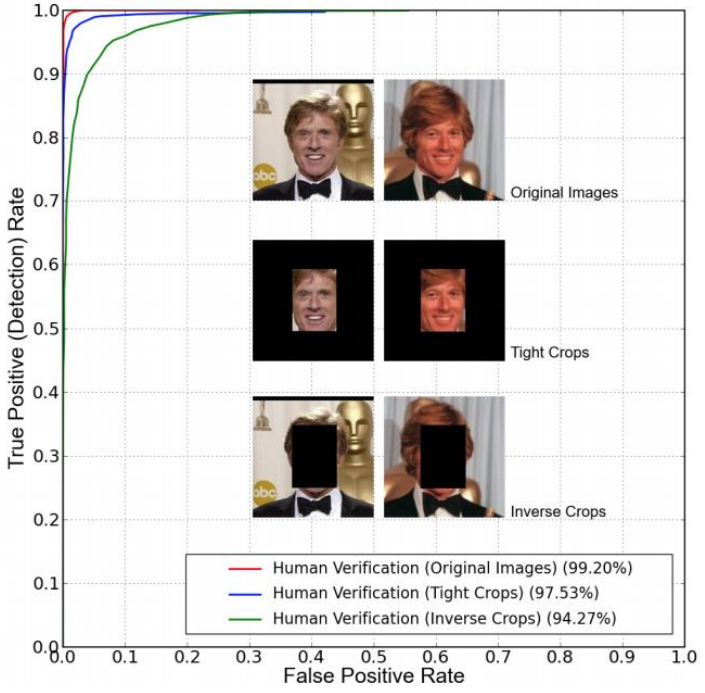


Key Idea: Development of an effective deep neural net (DNN) architecture for face recognition.

Paper 3: Recent Breakthrough

- Deep learning method that leverages a very large dataset of faces in order to obtain a face representation that generalizes well.
- Reaching near human-performance on the Labeled Faces in the Wild benchmark (LFW).





Paper 4: State-of-the-Art

2015. CVPR. Schroff, Kalenichenko, & Philbin (Google)
"FaceNet: A Unified Embedding for Face Recognition and Clustering"

Figure 1. **Illumination and Pose invariance.** Pose and illumination have been a long standing problem in face recognition. This figure shows the output distances of FaceNet between pairs of faces of the same and a different person in different pose and illumination combinations. A distance of 0.0 means the faces are identical, 4.0 corresponds to the opposite spectrum, two different identities. You can see that a threshold of 1.1 would classify every pair correctly.

Key Idea: Deep architecture that learns to optimize feature representation itself directly.

Paper 4: State-of-the-Art

- FaceNet directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity.
- Implement face recognition, verification and clustering at once using FaceNet embeddings as feature vectors that take only 128-bytes per face image.

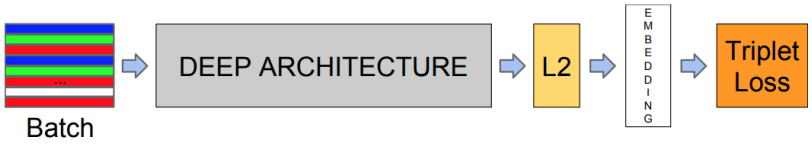


Figure 2. **Model structure.** Our network consists of a batch input layer and a deep CNN followed by L_2 normalization, which results in the face embedding. This is followed by the triplet loss during training.

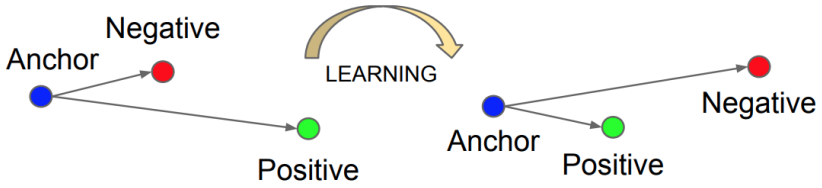


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

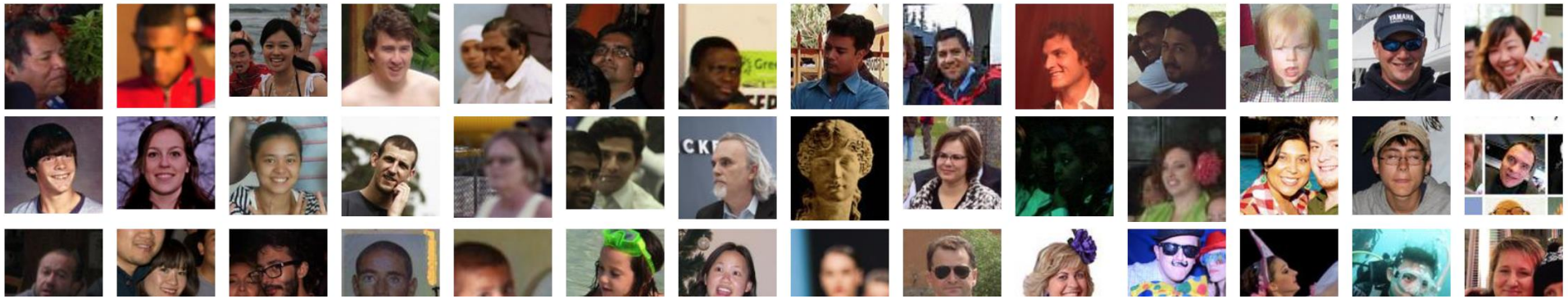
Paper 5: Possible Future Direction

2017. CVPR. Nech, & Kemelmacher

"Level Playing Field for Million Scale Face Recognition"

All that glisters is not gold
Often have you heard that told.

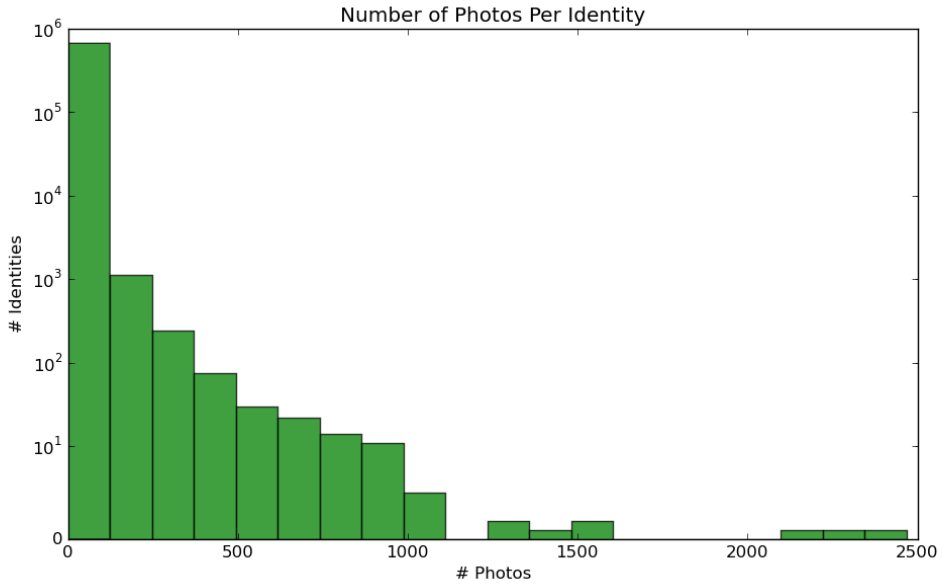
William Shakespeare



Key Idea: A public large-scale set with 672K identities and 4.7M photos, with the goal to level playing field for face recognition.

Paper 5: Possible Future Direction

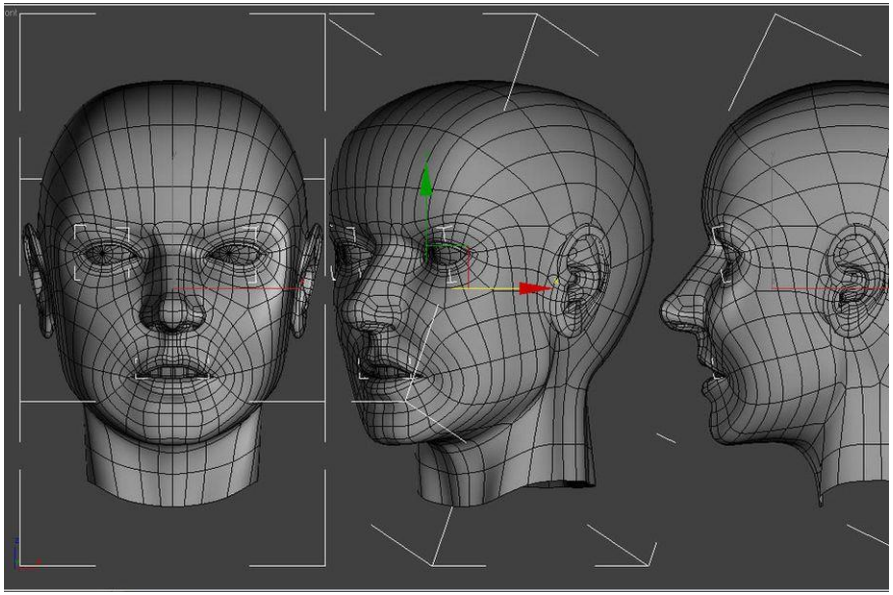
- With deep learning, the performance of face recognition systems can be boosted by large-scale datasets, however when tested at the million-scale exhibits dramatic variation in accuracies across the different algorithms.
- Large-scale deep learning algorithms need to be tested with larger-scale benchmarks.



Select Topics Not Covered (of Many)

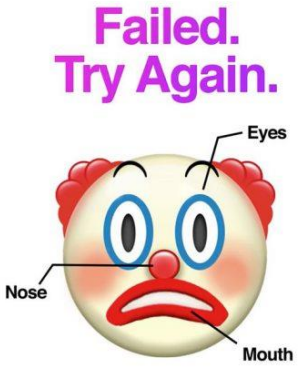
- Facial expression analysis

- 3D face recognition



Open Problems and Future Research

- Requirement of accurate face detection: Most of the current systems assume a bounding box of face is given.
- Black-box deep model: Uninterpretable recognition models make it hard to analyze the performance of the algorithms, which means it's not yet trustful.
- Privacy: Data-driven approach leads companies to label and use large-scale personal face data, which may cause privacy issue.



Activity Recognition

Section Structure

- Motivation, Description, Current and Future Impact
- Paper 1: “Old School” Seminal Work
- Paper 2: Early Progress in the Field
- Paper 3: Recent Breakthrough
- Paper 4: State-of-the-Art
- Paper 5: Possible Future Direction
- Open Problems and Future Research

Skipping historical long historical context for today, will return in lecture dedicated to the topic.

Description and Motivation

- **What is it?**
  Classify different human activities.
- **Why is it important?**
  Starting point of quantitatively modeling the semantic meaning of human motion.

Activity Recognition: Why It's Hard

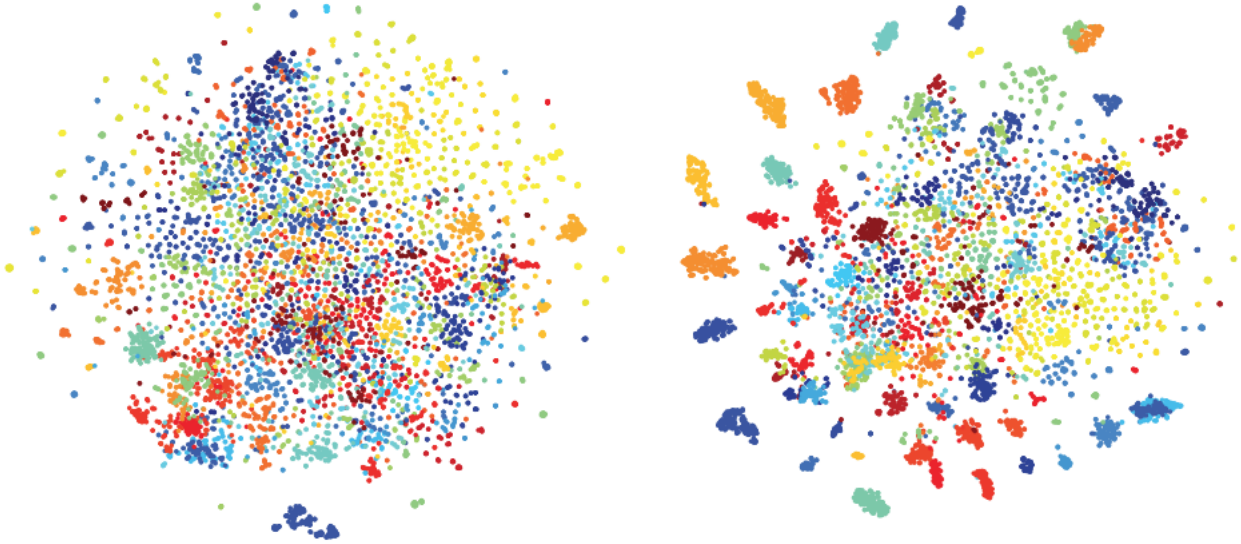
- All the problems in image recognition is the same in video, such as illumination, occlusion, scale, …
- **Complexity:** Magnitude larger data to process than image recognition.
- **Motion:** Hard to quantify, noise caused by background motion.
- **Camera:** Viewpoint changes, blurry, …
- **Ambiguity:** Activity is more various in kind and style.

Future Impact

- **Future Impact:** Describe the world through time and predict.
- **Utopia:** All the human activities can be recognized through camera, preventing crime, danger, robots can interact with humans.
- **Dystopia:** All the human activities can be reproduced by robots, forming a robot society and sever the collaborative relationships between robots and humans.
- **Middle path:** Find useful information in massive video data.

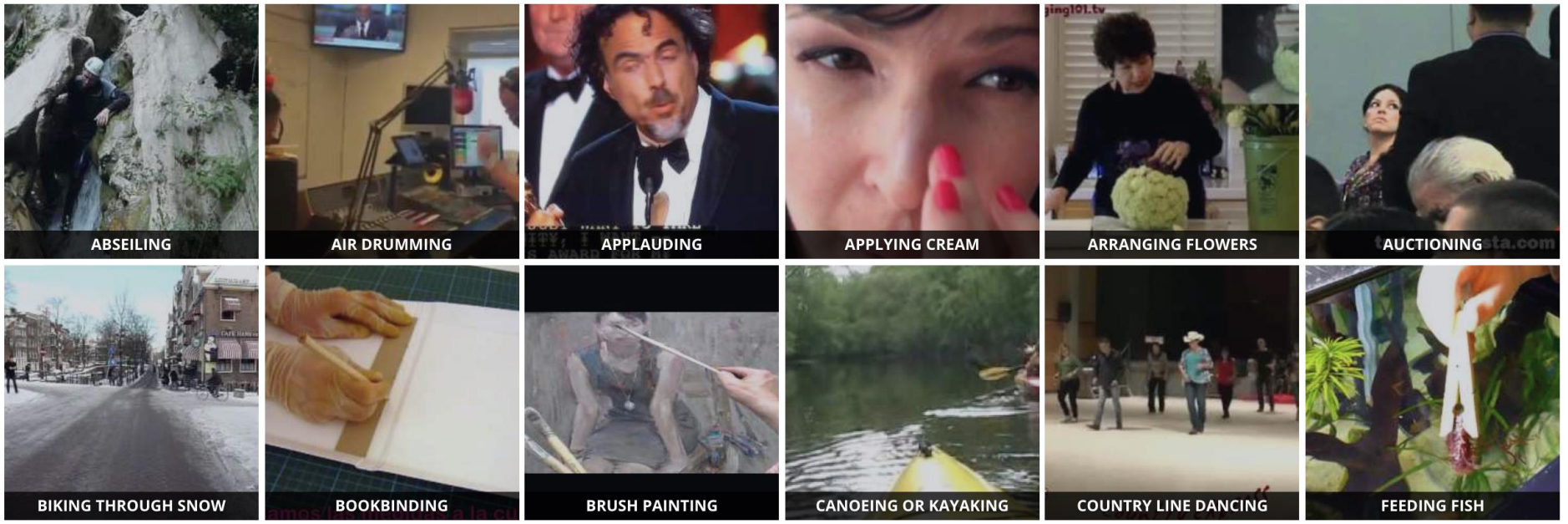
Paper 3: Recent Breakthrough

- Use deep learning method to learn video feature representation end-to-end.
- Shows the benefit of using spatio-temporal feature in the field of action recognition, rather than using single-image feature.



Paper 4: State-of-the-Art

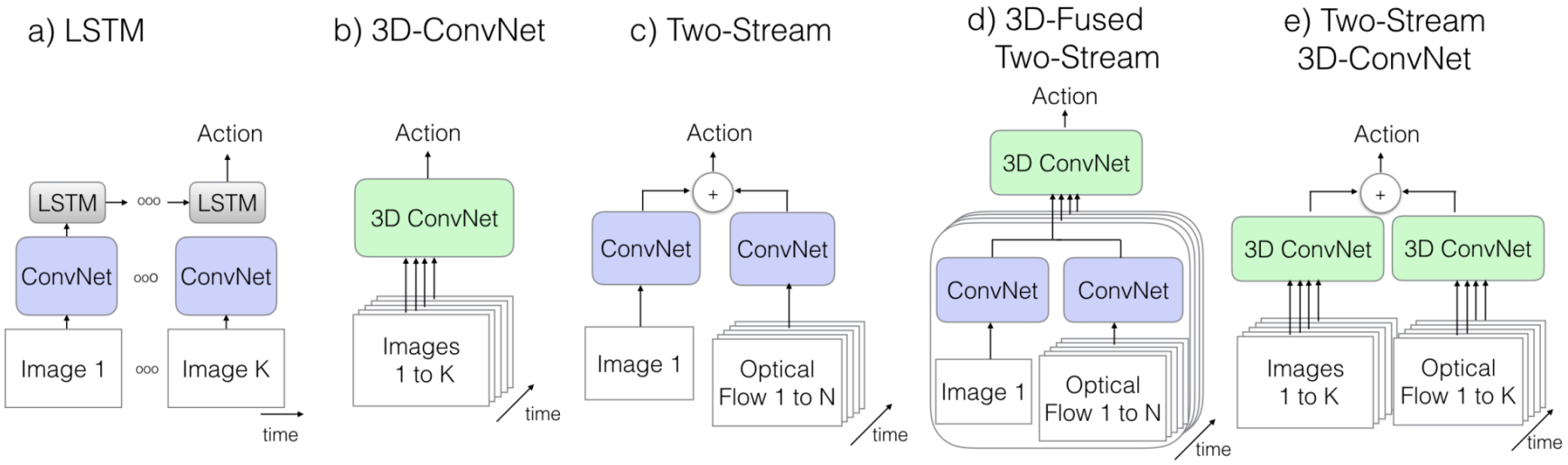
2017. CVPR. Carreira, & Zisserman (DeepMind)
"Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset "



Key Idea: Two-Stream 3D ConvNets with a large video dataset.

Paper 4: State-of-the-Art

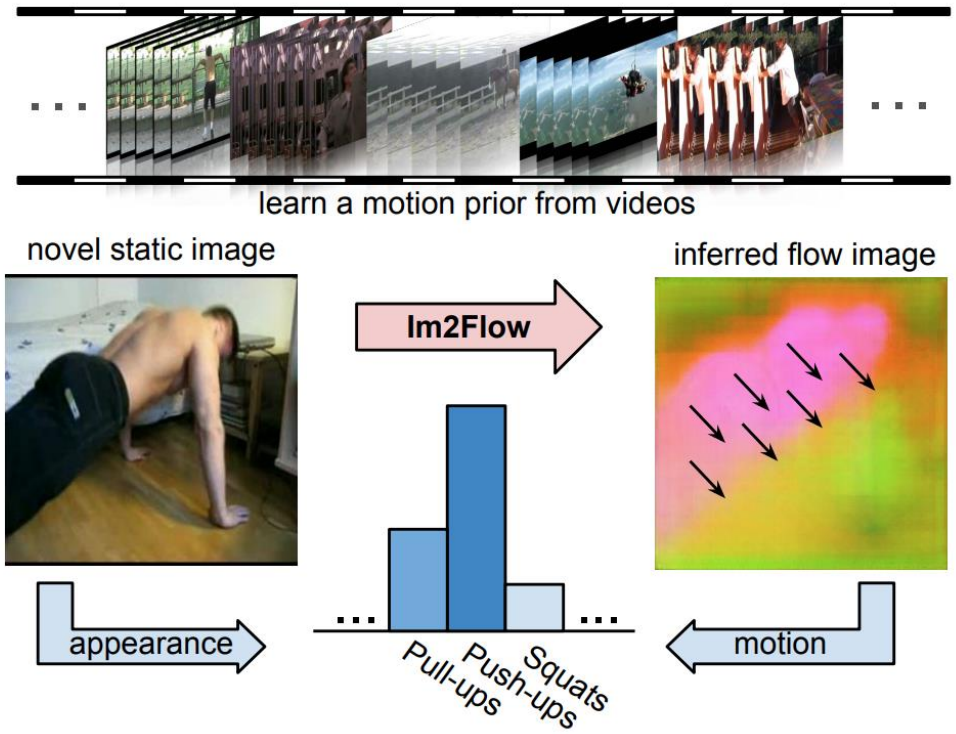
- Introduce a new Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D.
- Considering both spatiotemporal feature and motion feature (optical flow), training end-to-end on large-scale video dataset.



Paper 5: Possible Future Direction

2018. CVPR. Gao, Xiong, & Grauman (UT Austin)

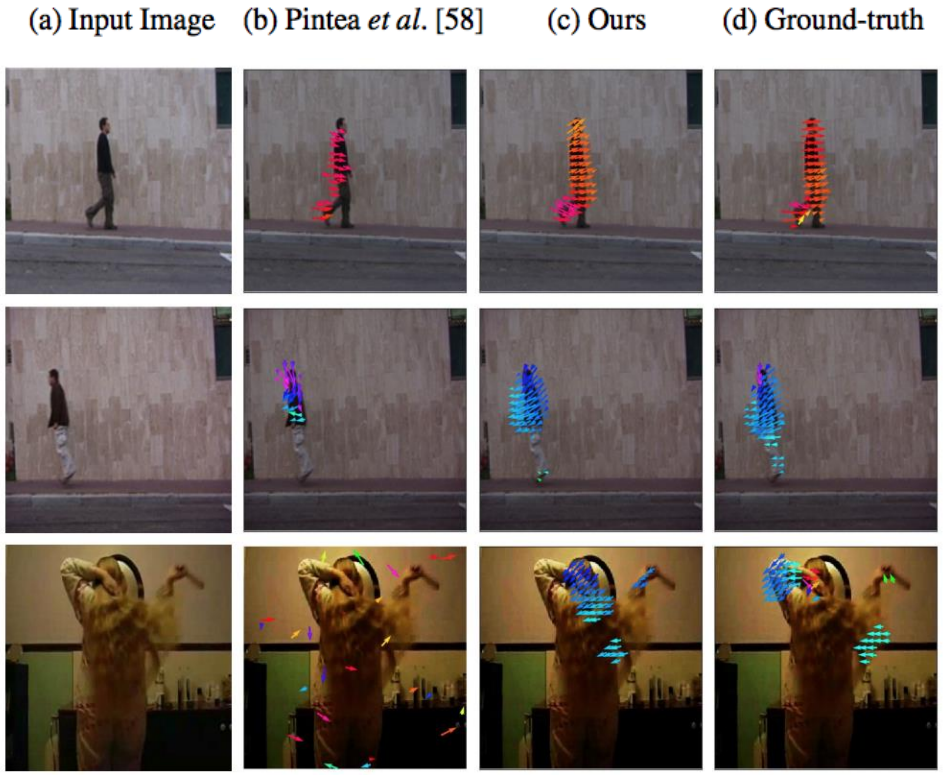
"Im2Flow: Motion Hallucination from Static Images for Action Recognition"

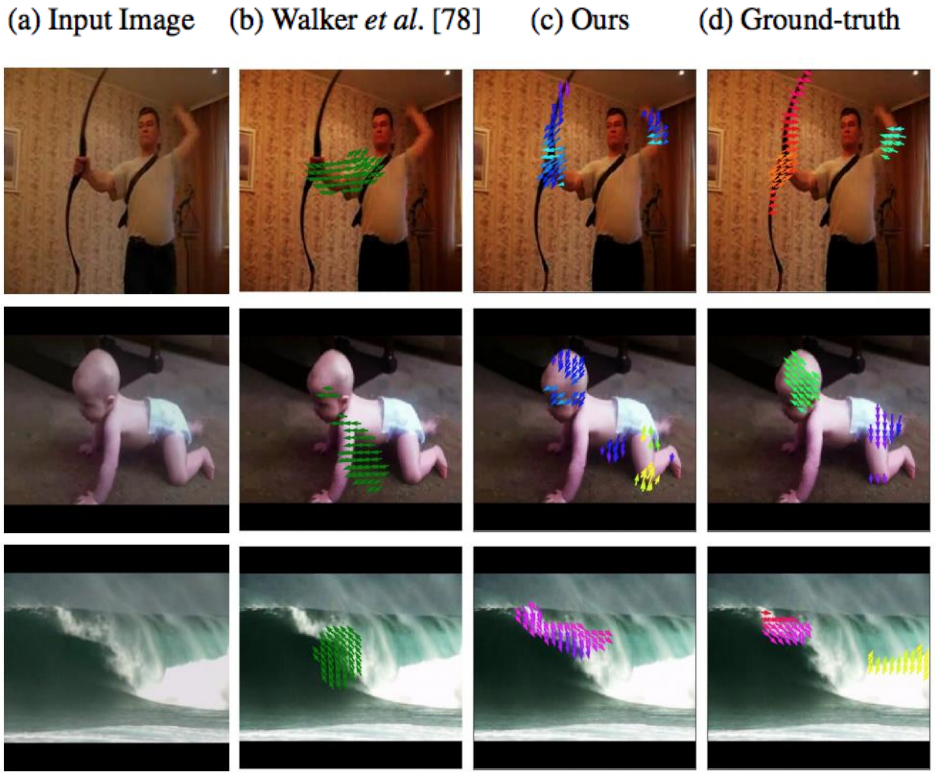


Key Idea: Action recognition with visual anticipation.

Paper 5: Possible Future Direction

- Hallucinate the motion implied by a single snapshot and then use it as an auxiliary cue for static-image action recognition

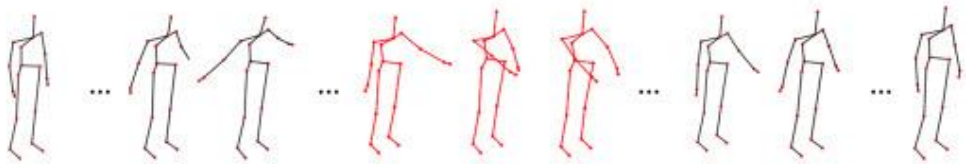




Select Topics Not Covered (of Many)

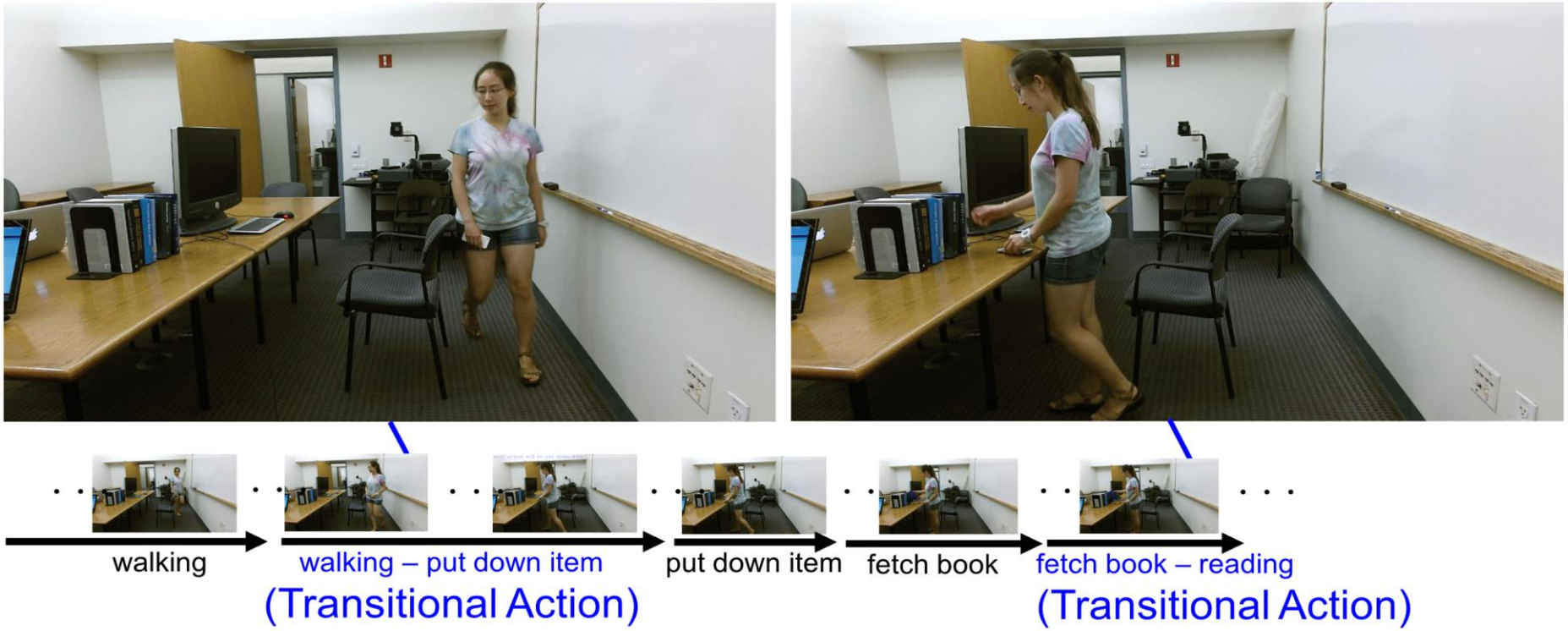
- Action detection/localization/segmentation in long, untrimmed videos

- Skeleton-base action recognition



Open Problems and Future Research

- Complexity reduction: How much information do we need to model the video?
- Action Prediction: How much are we able to predict future actions?



Body Pose Estimation

Section Structure

- Motivation, Description, Current and Future Impact
- Paper 1: “Old School” Seminal Work
- Paper 2: Early Progress in the Field
- Paper 3: Recent Breakthrough
- Paper 4: State-of-the-Art
- Paper 5: Possible Future Direction
- Open Problems and Future Research

Skipping historical long historical context for today, will return in lecture dedicated to the topic.

Description and Motivation

What is it?

Infer the pose of an articulated body, i.e. a skeleton.

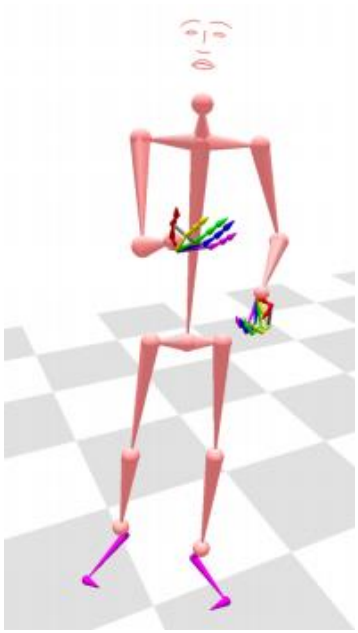
Why is it important?

Animation, body-language, activity recognition. It is a useful, compressed representation of the human body.

Body Pose Estimation: Why It's Hard

Part Occlusion

High dimensional



- High Dimensional Optimization Problem (e.g., up to 6 DOF per joint)
- Usual computer vision culprits: Scale, pose, occlusion, expression, makeup, illumination

Future Impact

Key Idea: Enable us to measure the geometry, position, and motion of the human body.

Utopia: Motion capture technology (sports, CGI, video games) and human-robot interaction.

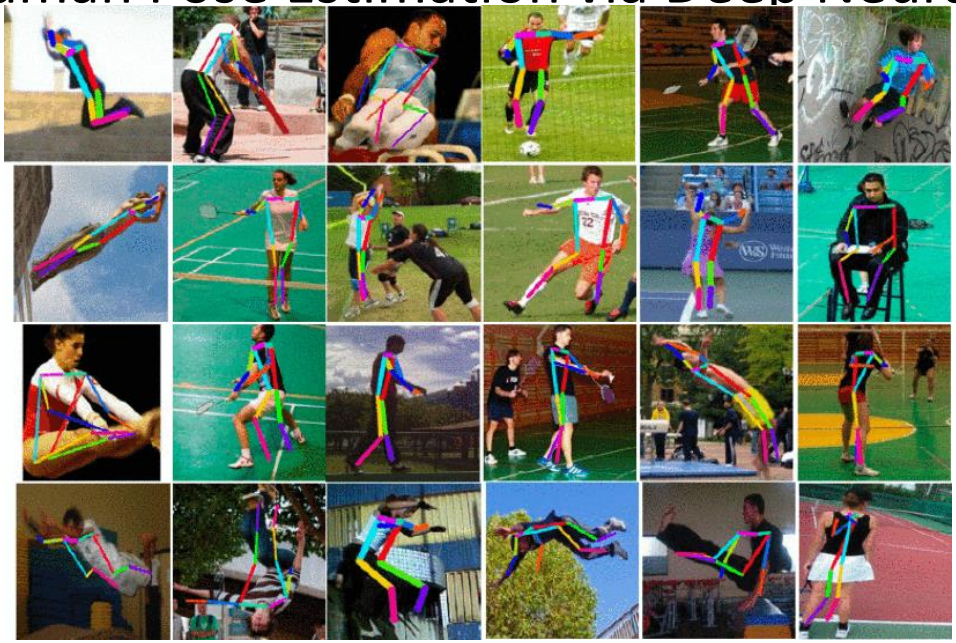
Dystopia: Combat robots.

Middel path: HCI that is aware of the human body in space.

Paper 3: Recent Breakthrough

2014. CVPR. Toshev; Szegedy. (Google)

"DeepPose: Human Pose Estimation via Deep Neural Networks"



Key Idea: Neural Networks enable **holistic human pose estimation**. Localize joint locations with the full context of the scene.

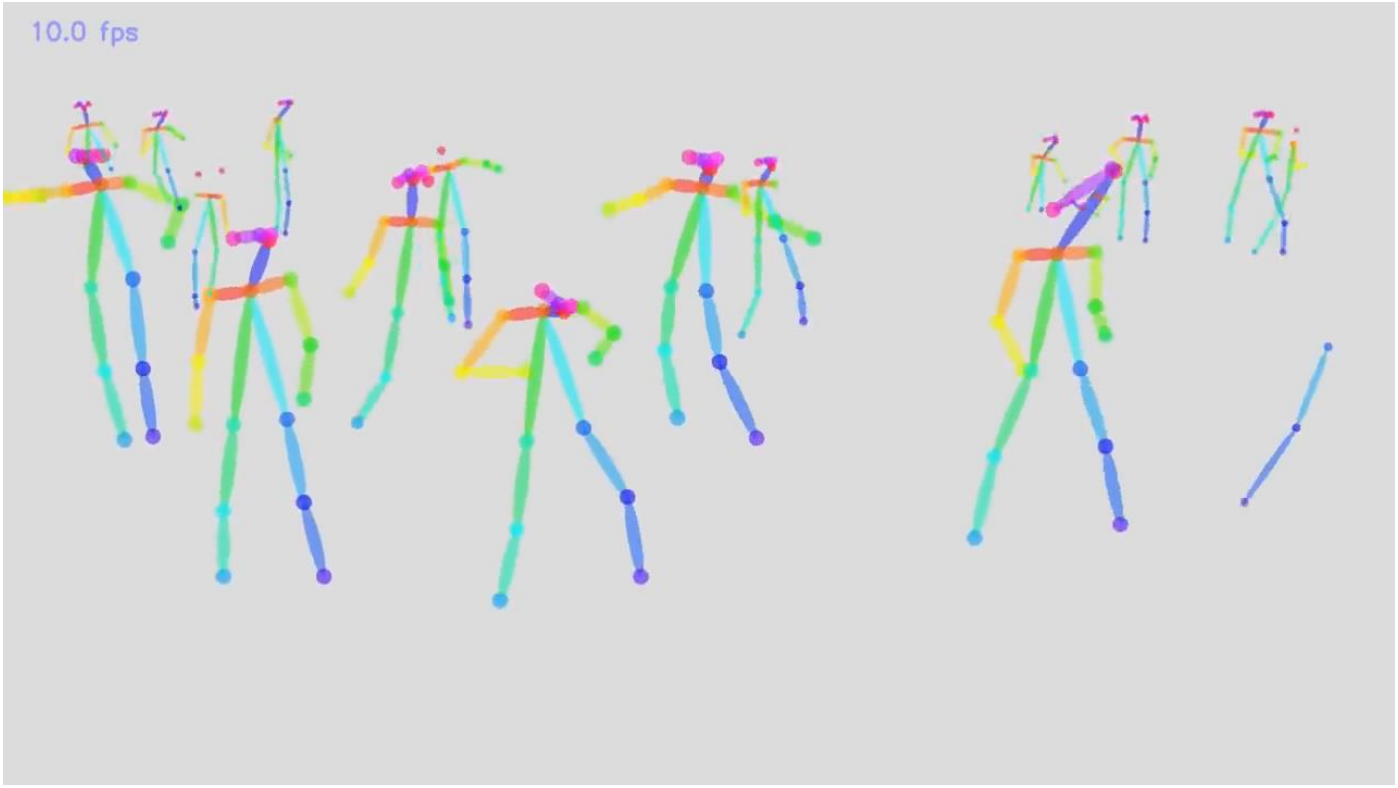
Paper 3: Recent Breakthrough

- Previous work used hand crafted features.
- Features for part detection can be learned.

Paper 4: State-of-the-art

2017. CVPR. Cao; Simon; Wei; Sheikh. (CMU)

"Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields"

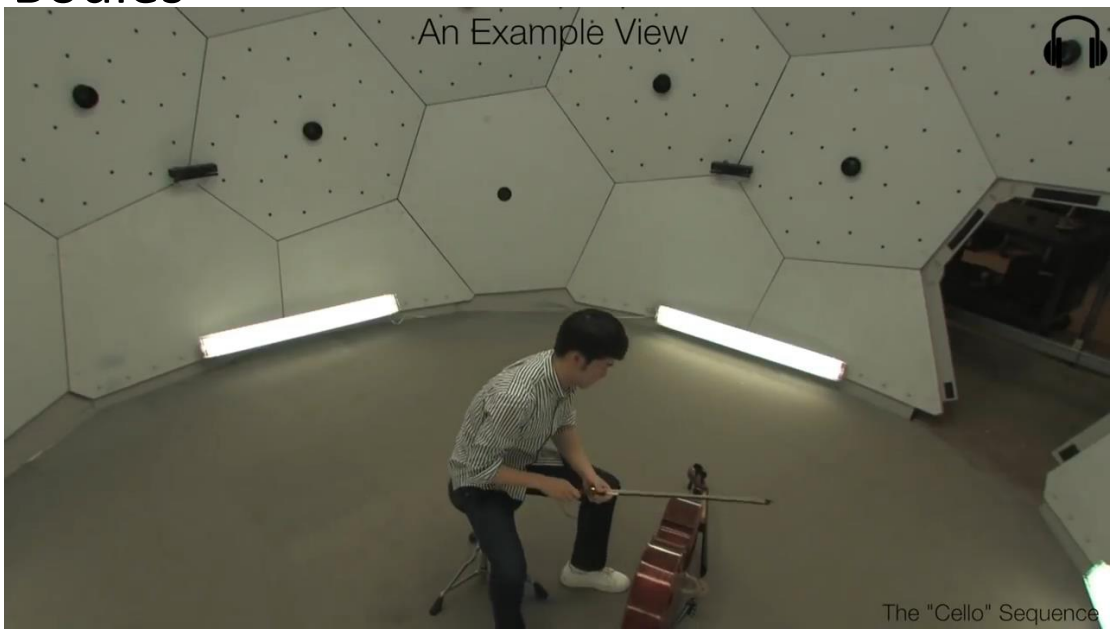


Key Idea: A bottom-up approach using Part Affinity Fields (PAFs).

Paper 5: Possible Future Direction

2018. CVPR. Joo; Simon; Wei; Sheikh.

"Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies"

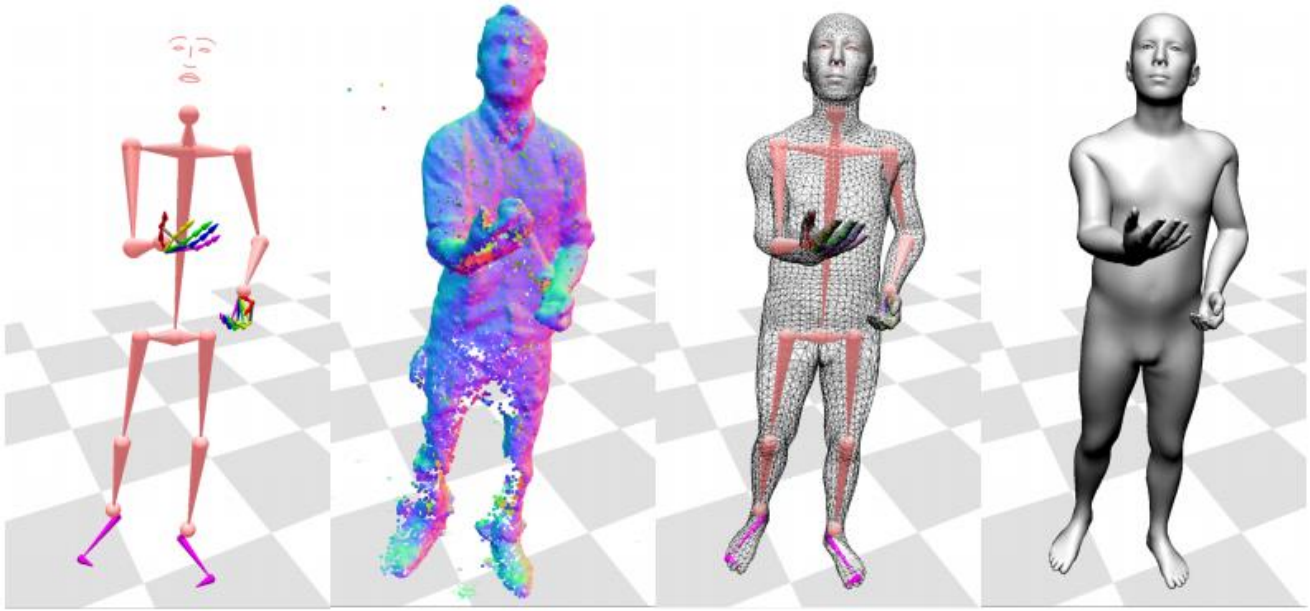


Key Idea: Deformable models capture more information.

- Deformation model = pose + shape of individual + orientation in space
- PCA reduces dimensionality of the model.

Paper 5: Future

- Mesh models capture nuanced expressions.
- Previous work focuses only on articulated bodies.
- Enable research to finely measure the human body, expressions, motion.



Open Problems and Future Research

- Include physical dynamics of human motion into current CNN methods.

Gesture Recognition

Section Structure

- Motivation, Description, Current and Future Impact
- Paper 1: “Old School” Seminal Work
- Paper 2: Early Progress in the Field
- Paper 3: Recent Breakthrough
- Paper 4: State-of-the-Art
- Paper 5: Possible Future Direction
- Open Problems and Future Research

Skipping historical long historical context for today, will return in lecture dedicated to the topic.

Description and Motivation

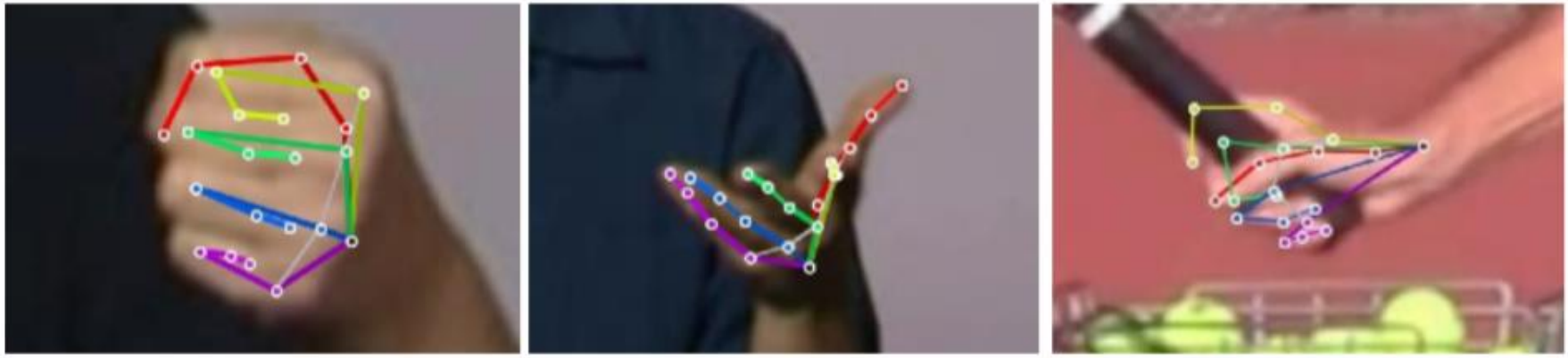
What is it?

Interpreting non-verbal communication.

Why is it important?

UI design. Natural mode of communication

Gesture Recognition: Why It's Hard



(a) Articulation (b) Viewpoint (c) Object

- Gestures can be temporal
- Gestures are often occluded
- Gestures can involve other objects such as tools
- Hands have a complex articulation

Future Impact

Key Idea: Enable us to communicate with computers using the motion of our bodies..

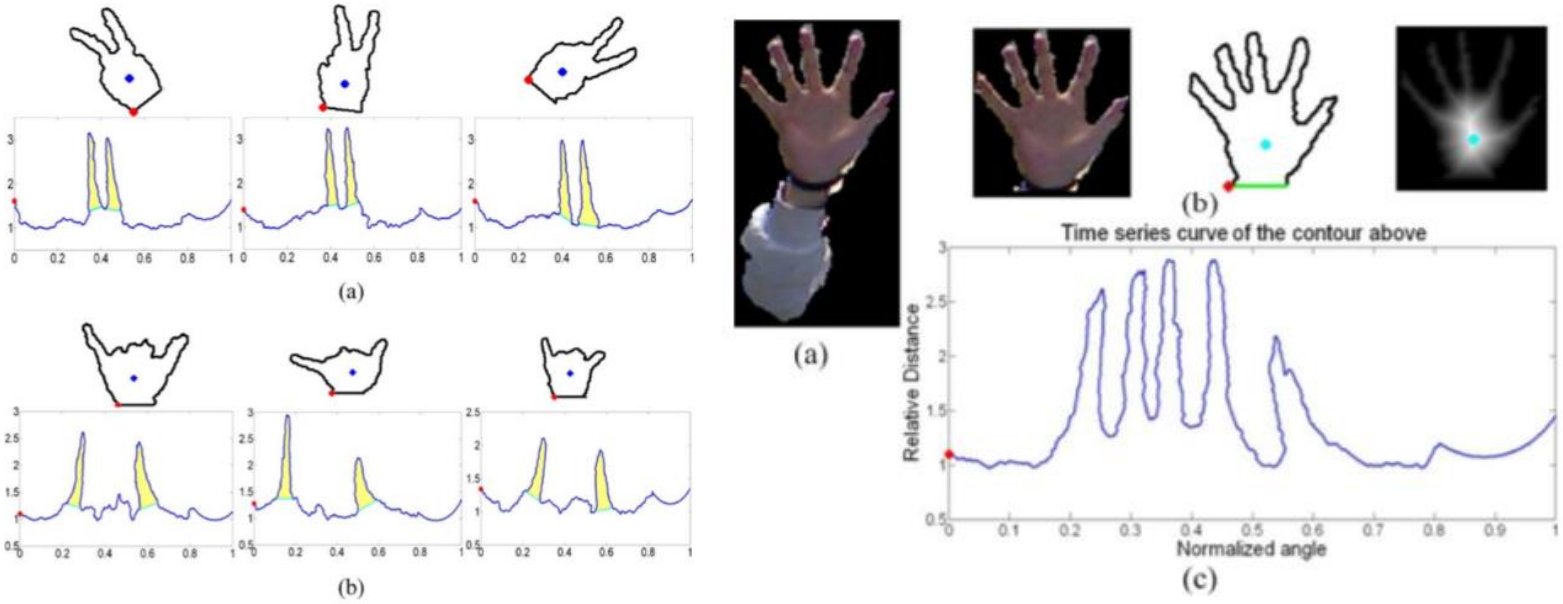
Utopia: Awesome UI. Improved accessibility for sign language users. Multi-modal input.

Dystopia: Terribly misused UI.

Middle path: HCI which can interpret body language.

Paper 3: Recent Breakthrough

2013. IEEE. Ren; Yuan; Meng; Zhang "Robust Part-Based Hand Gesture Recognition Using Kinect Sensor"

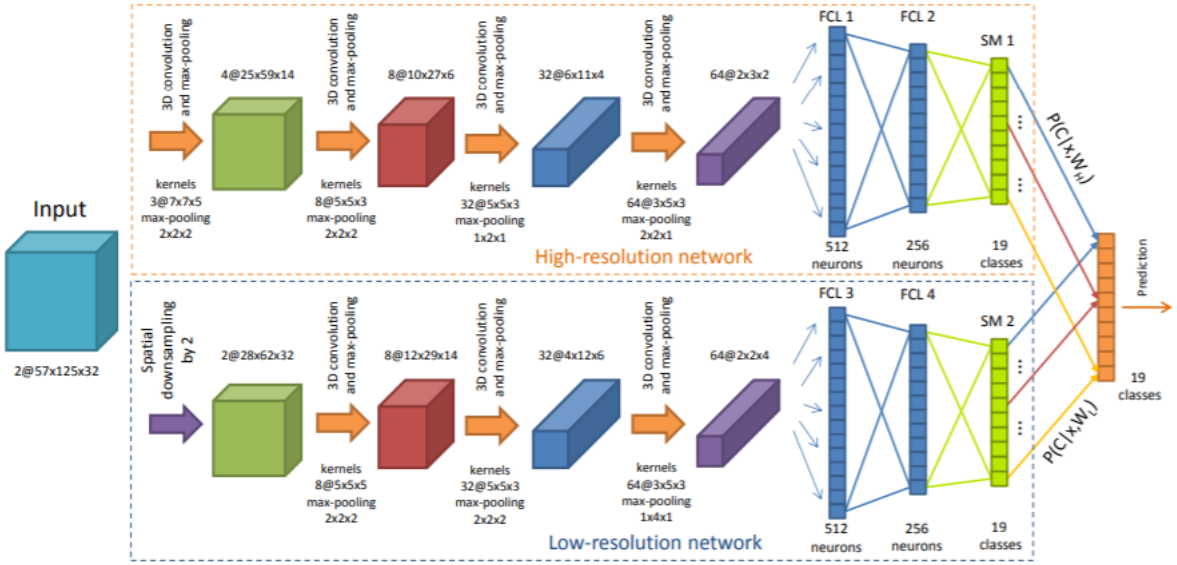


Key Idea: Novel distance metric, tailored to hands

- Depth sensor + Finger-Earth Mover's Distance

Paper 4: State-of-the-art

2015. CVPR. Molchanov; Gupta; Kim; Kautz. (NVIDIA)
"Hand Gesture Recognition with 3D Convolutional Neural Networks"

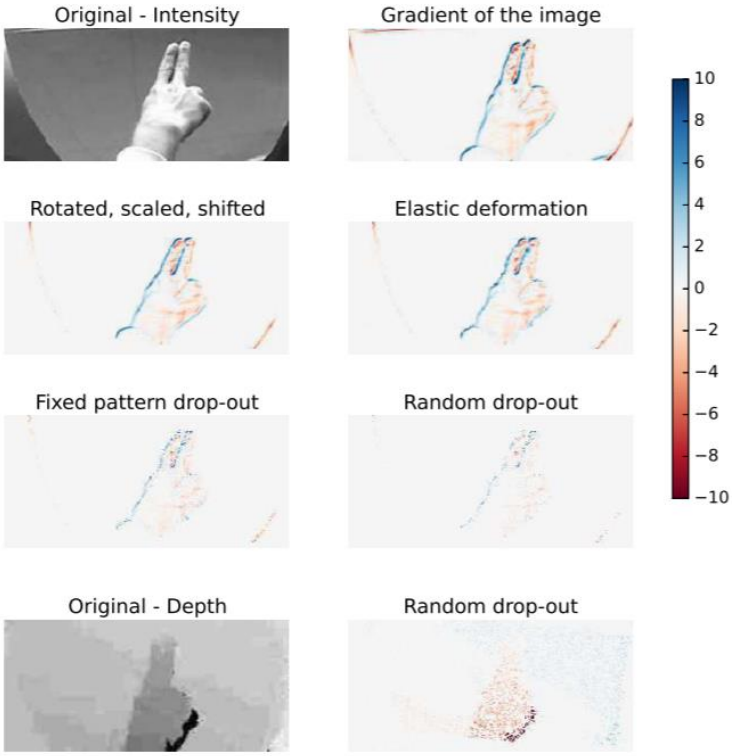


Key Idea: Spatio-temporal features can be learned with 3D convolutions

- CNNs can combine sensor data, e.g. depth + video
- Augment data by adding noise to avoid overfitting

Paper 4: State-of-the-art

- Previous work used hand crafted features. Difficult to combine sensor streams.
- CNNs can learn features automatically and combine sensors.

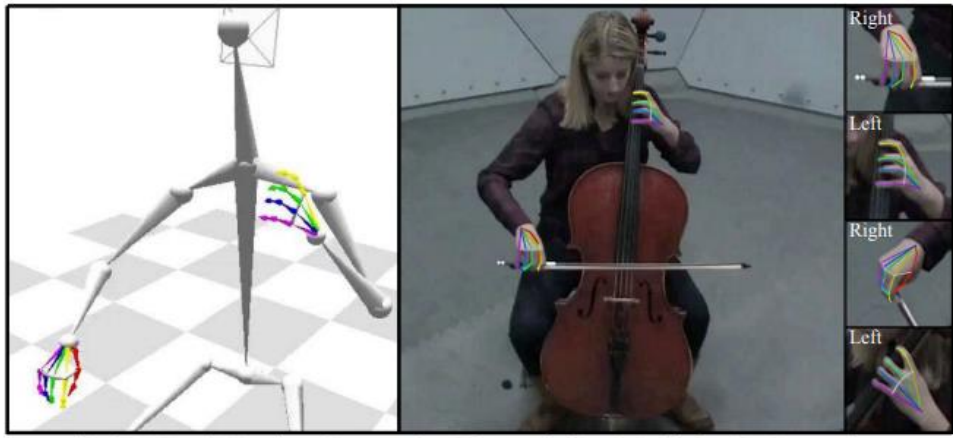


Paper 5: Possible Future Direction

2017. CVPR. Simon; Joo; Mathews; Sheikh. (CMU)

"Hand Keypoint Detection in Single Images using Multiview Bootstrapping"

(a) Realtime 2D Hand Detection on YouTube and Webcam Videos



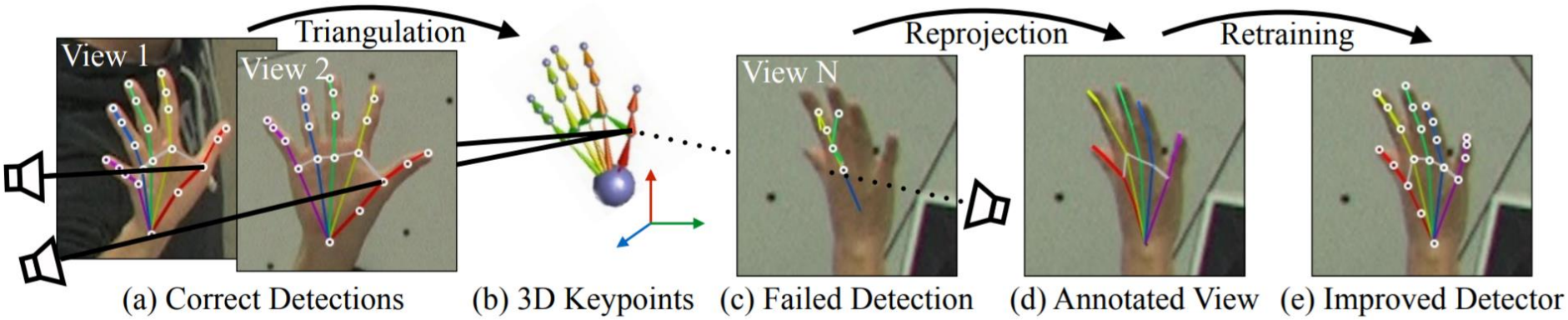
(b) 3D Hand Motion Capture by Triangulating Multiple 2D Detections

Key Idea: Data is everything. Bootstrap from multiple views.

- Multiple-camera system: easy view points, hard viewpoints
- Coincident localizations are likely good.

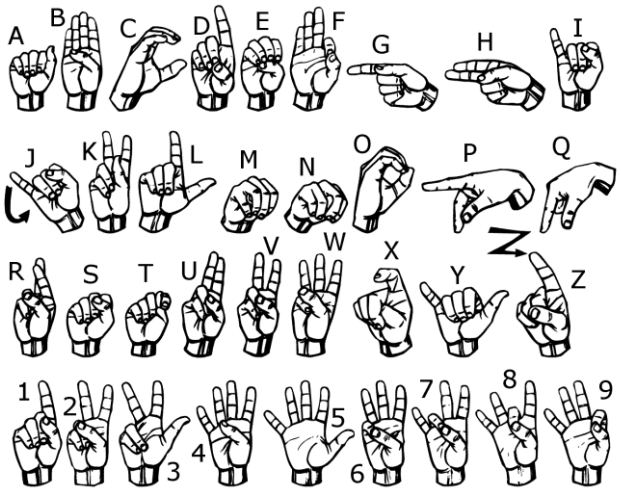
Paper 5: Possible Future Direction

- Robust hand pose detection from vision alone.
- Previous work used small datasets.
- Data is the fuel and fire of machine learning.

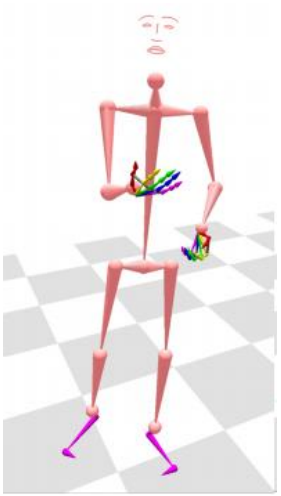


Select Topics Not Covered (of Many)

- Sign Language



- Other gesture modalities



Open Problems and Future Research

- Full-body gestures
- Natural body language understanding
- One shot learning (custom user gestures)

Speech Recognition

Section Structure

- Motivation, Description, Current and Future Impact
- Paper 1: “Old School” Seminal Work
- Paper 2: Early Progress in the Field
- Paper 3: Recent Breakthrough
- Paper 4: State-of-the-Art
- Paper 5: Possible Future Direction
- Open Problems and Future Research

Skipping historical long historical context for today, will return in lecture dedicated to the topic.

Description and Motivation

- **What is it?**
  Translation of spoken language into text by computers.
- **Why is it important?**
  Voice user interfaces such as voice dialing (e.g. "Call home"), speech-to-text processing (e.g., word processors or emails), …

Speech Recognition: Why It's Hard

- Human speech has huge amount of variations which occur while pronouncing a word.
- When the scale of vocabulary increases, we need more training data and the model becomes more computational expensive.

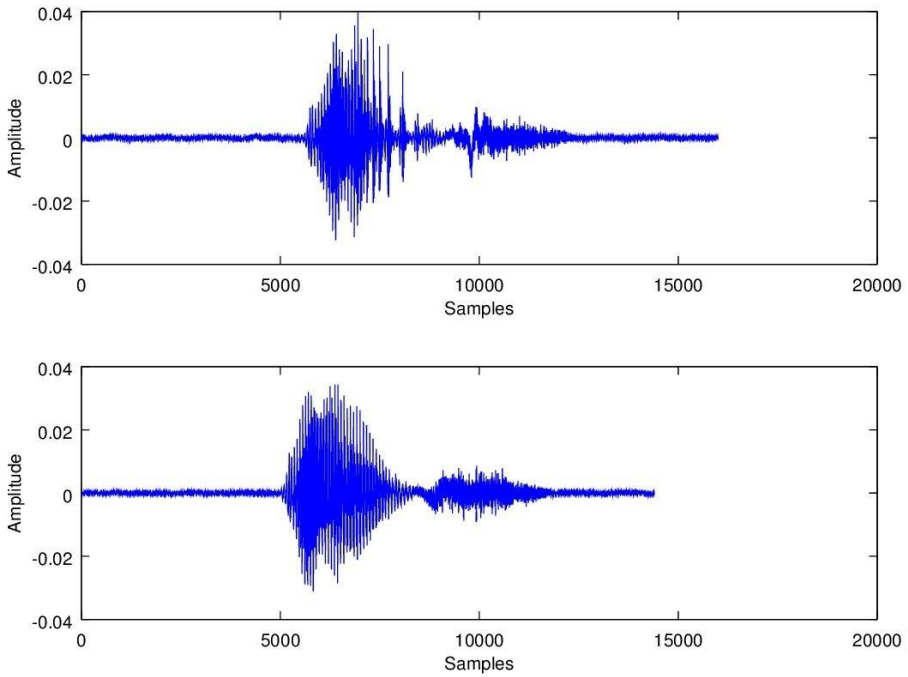
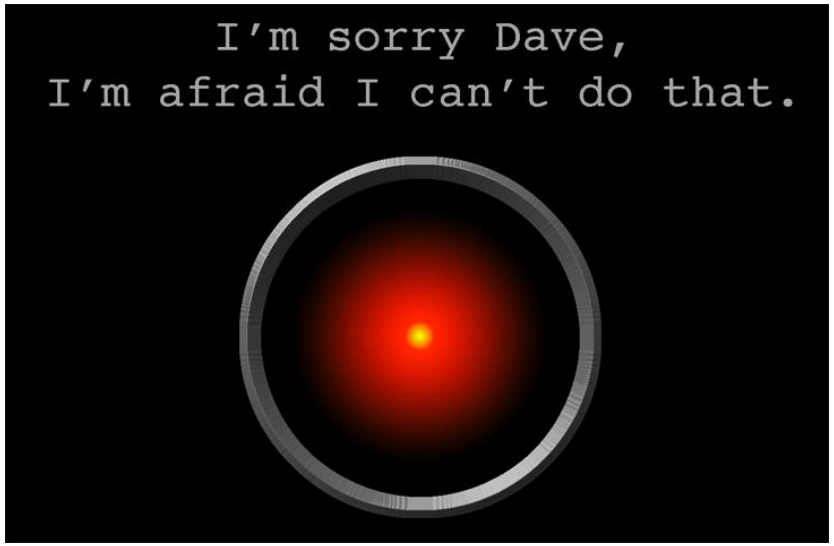


Fig. Comparison of two different recording of the word "Yes" in the time domain.

Future Impact

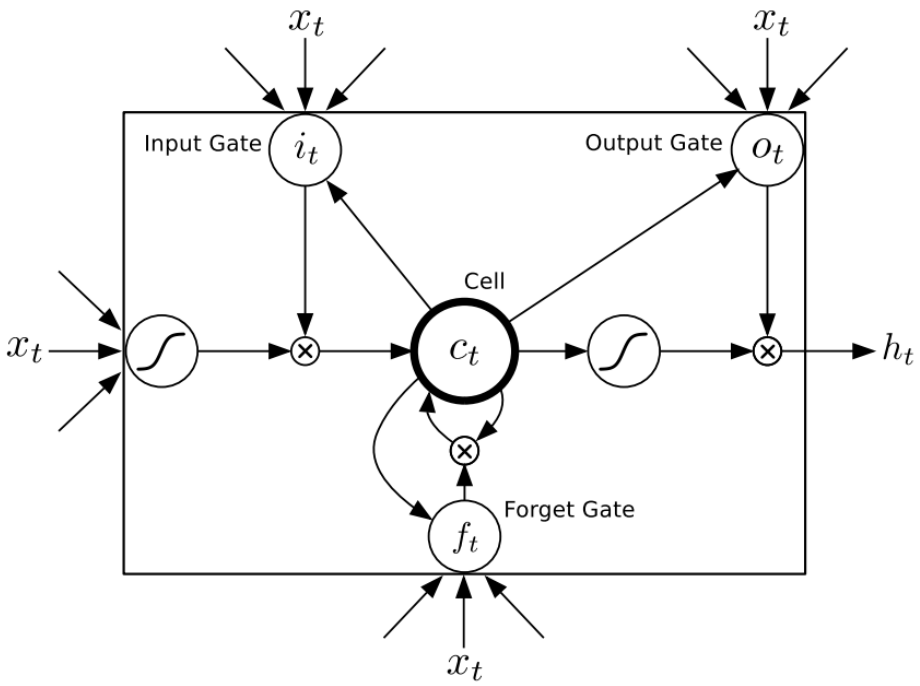


- **Future Impact:** Voice is becoming an interface of its own.
- **Utopia:** Ask Alexa everything. Conversational interfaces are everywhere.
- **Dystopia:** "Personal assistants" listen to everything and is misused by hackers, companies, or governments.
- **Middle path:** Ask Alexa to make an espresso.

Paper 3: Recent Breakthrough

2013. ICASSP. Graves, Mohamed, & Hinton

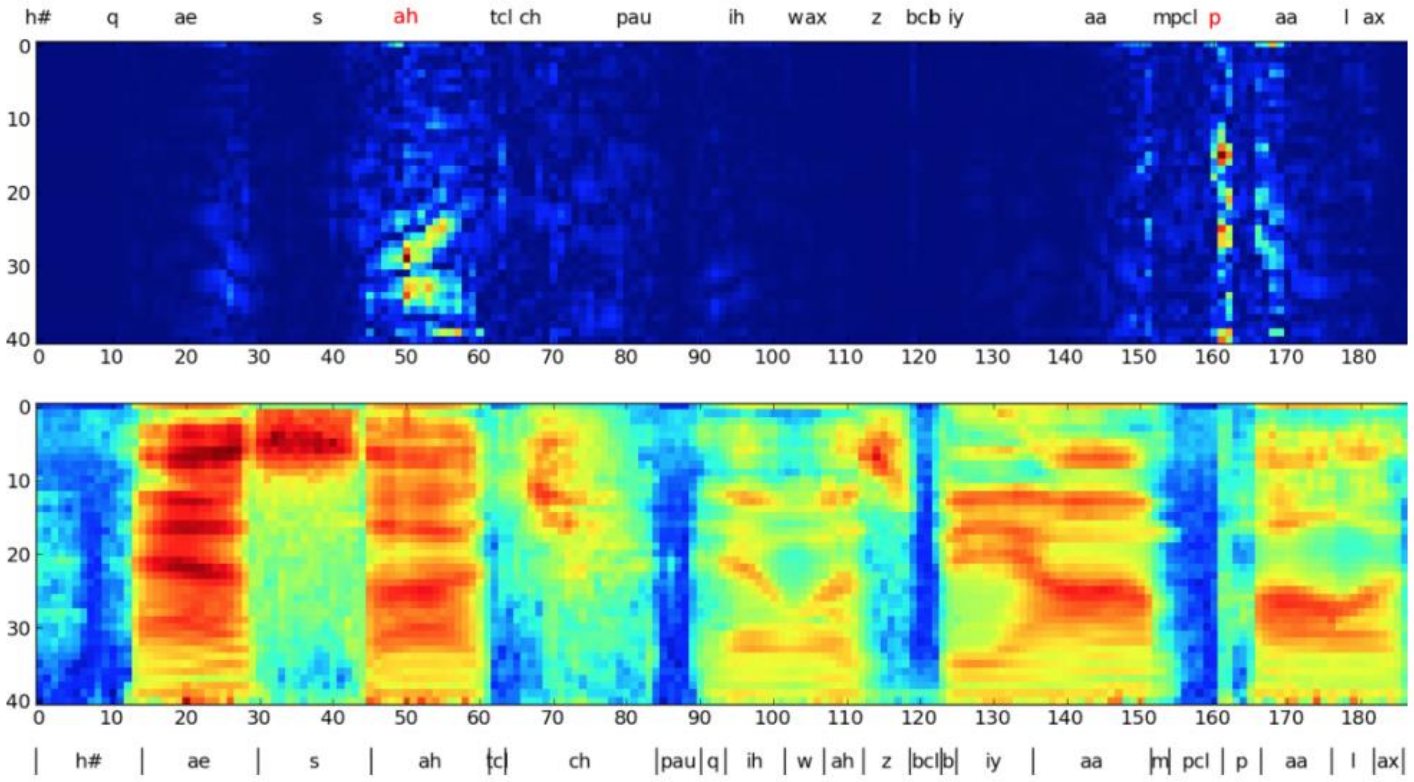
"Speech Recognition with Deep Recurrent Neural Networks"



Key Idea: Deep Long Short-term Memory RNNs with Connectionist Temporal Classification algorithm.

Paper 3: Recent Breakthrough

- The combination of deep, bidirectional Long Short-term Memory RNNs with end-to-end training and weight noise gives state-of-the-art results in phoneme recognition.



Paper 4: State-of-the-Art

2018. ICASSP. Chiu et al. (Google)
"State-of-the-art Speech Recognition with Sequence-to-sequence Models"

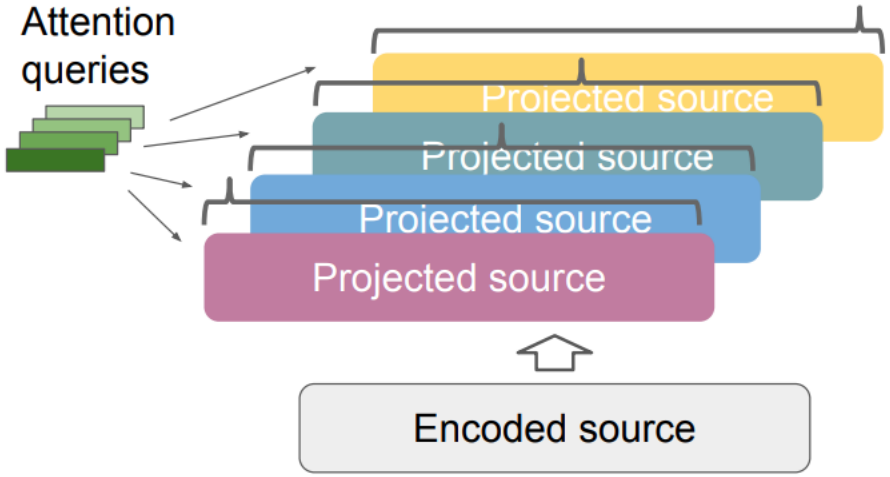


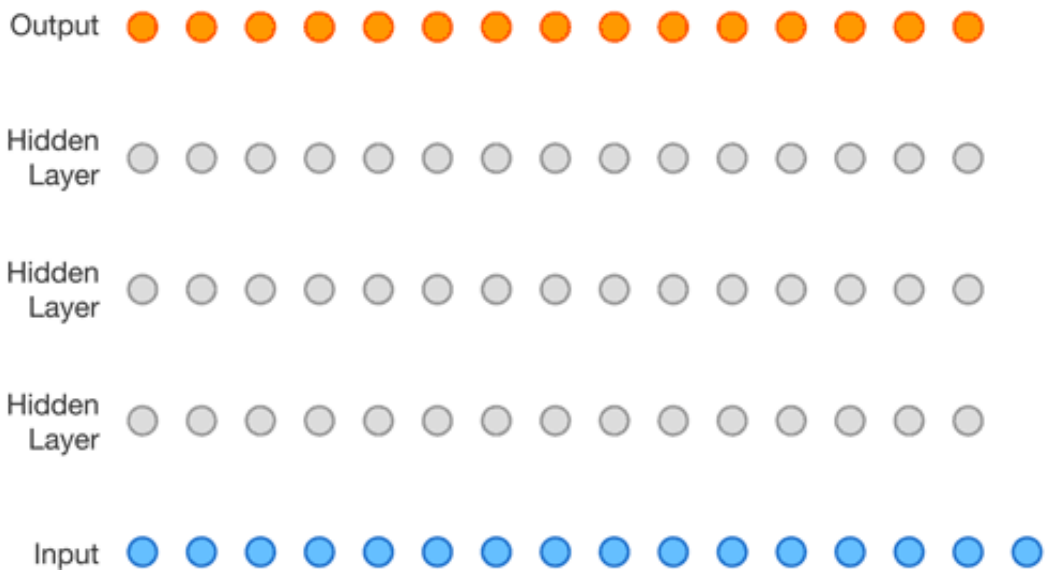
Fig. 2: Multi-headed attention mechanism.

Key Idea: Attention-based deep model for sequence-to-sequence speech recognition.

Paper 5: Possible Future Direction

2016. arXiv. Oord et al. (DeepMind)

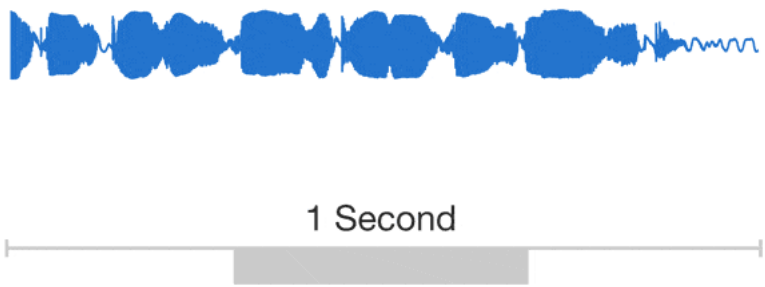
"WaveNet: A Generative Model for Raw Audio"



Key Idea: A deep neural network for generating raw audio waveforms.

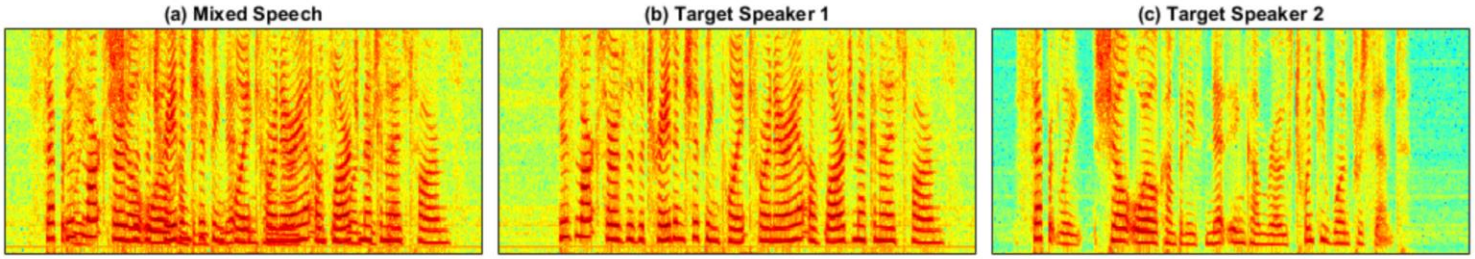
Paper 5: Possible Future Direction

- Operates directly at the waveform level without extra information or constraints.
- Although designed as a generative model, it can straightforwardly be adapted to discriminative audio tasks such as speech recognition.
- WaveNets can be conditioned on other inputs in a global (e.g. speaker identity) or local way (e.g. linguistic features).

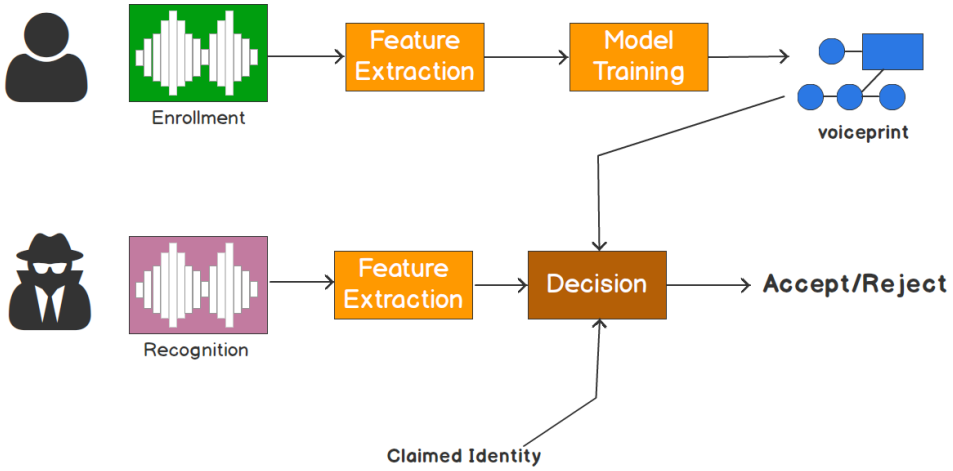


Select Topics Not Covered (of Many)

- Speech Separation



- Speaker Verification



Open Problems and Future Research

- Making and measuring progress: manually transcribing data is at about same word error rate as system.
- Noise-robustness.
- Speaker / accent / channel variability.
- Streaming fashion: Most current systems have the limitation that the entire utterance must be seen by the encoder, before any labels can be decoded.

Recommendation Systems

Section Structure

- Motivation, Description, Current and Future Impact
- Paper 1: “Old School” Seminal Work
- Paper 2: Early Progress in the Field
- Paper 3: Recent Breakthrough
- Paper 4: State-of-the-Art
- Paper 5: Possible Future Direction
- Open Problems and Future Research

Skipping historical long historical context for today, will return in lecture dedicated to the topic.

Description and Motivation

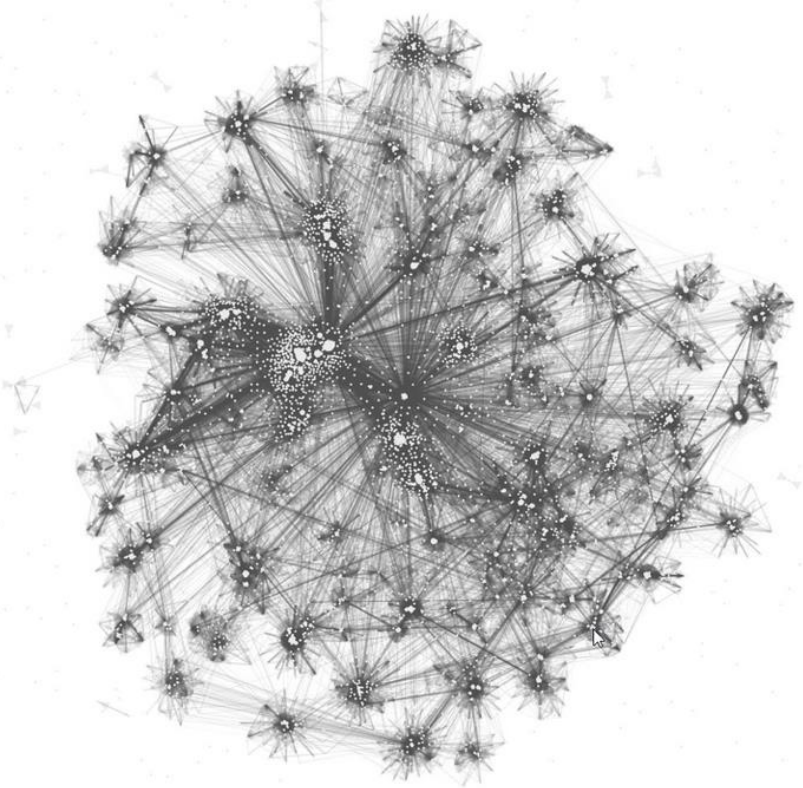
What is it?

Predicting user preferences.

Why is it important?

Our options exceed our resources for exploring.

Recommendation Systems: Why It's Hard



- Scale
- Sparsity
- Noisy data (unobservable external factors)
- Cold Starts (new content)
- Feature engineering (even with deep learning)

Future Impact

Key Idea: Systems which are customized for each user.

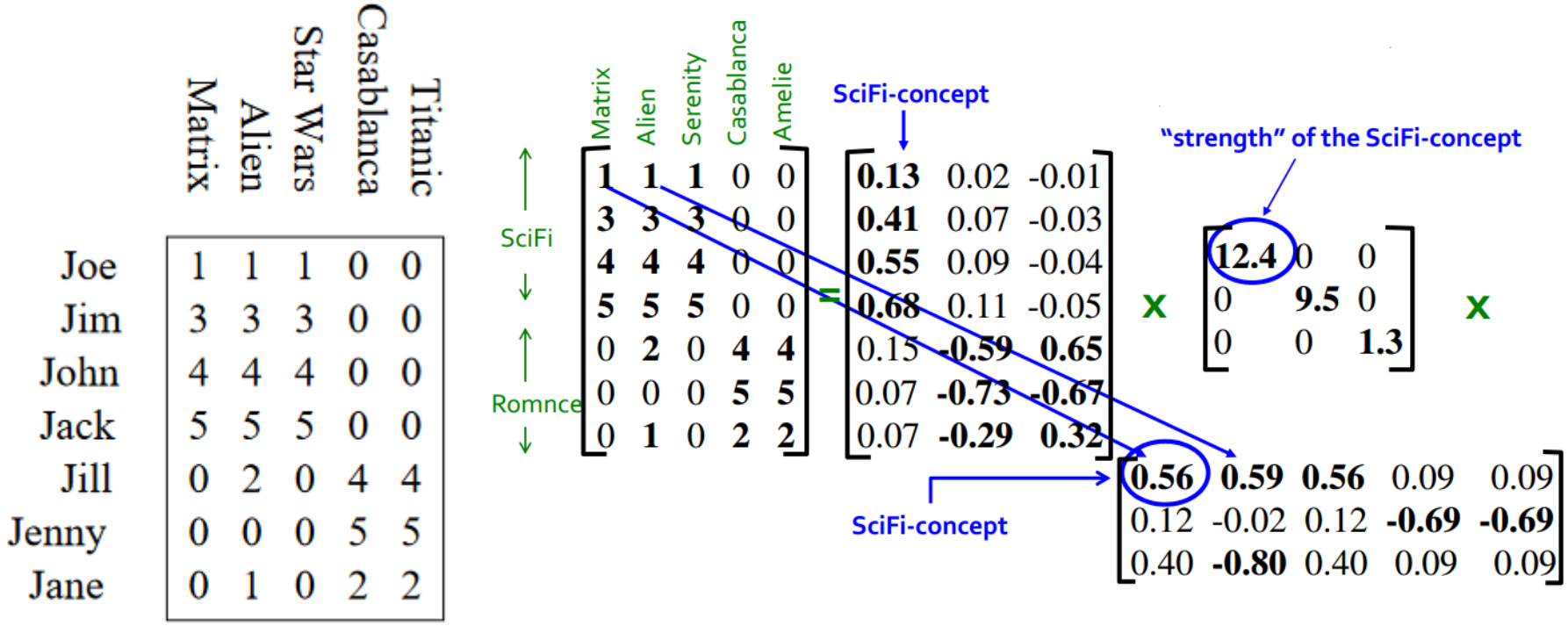
Utopia: AI that know us better than we do and helps us.

Dystopia: AI that know us better than we do and manipulates us.

Middle path: HCI helps us manage large decision spaces.

Paper 3a: Recent Breakthroughs

2002. ICCIS. Sarwar; Karypis; Konstan; Reidl.
"Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems"

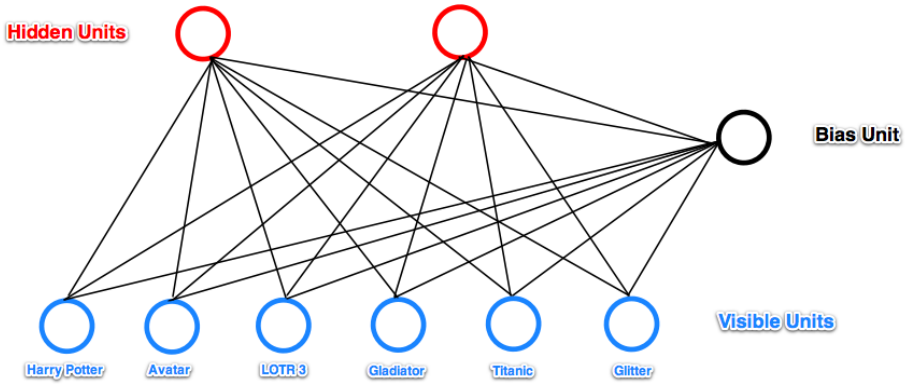


Key Idea: Low dimensional representation by matrix factorization

- Best low rank linear approximation of data
- Vectors can represent genres and kinds of users

Paper 3b: Recent Breakthroughs

2007. ICML. Salakhutdinov; Mnih; Hinton. (UToronto)
"Restricted Boltzmann Machines for Collaborative Filtering"



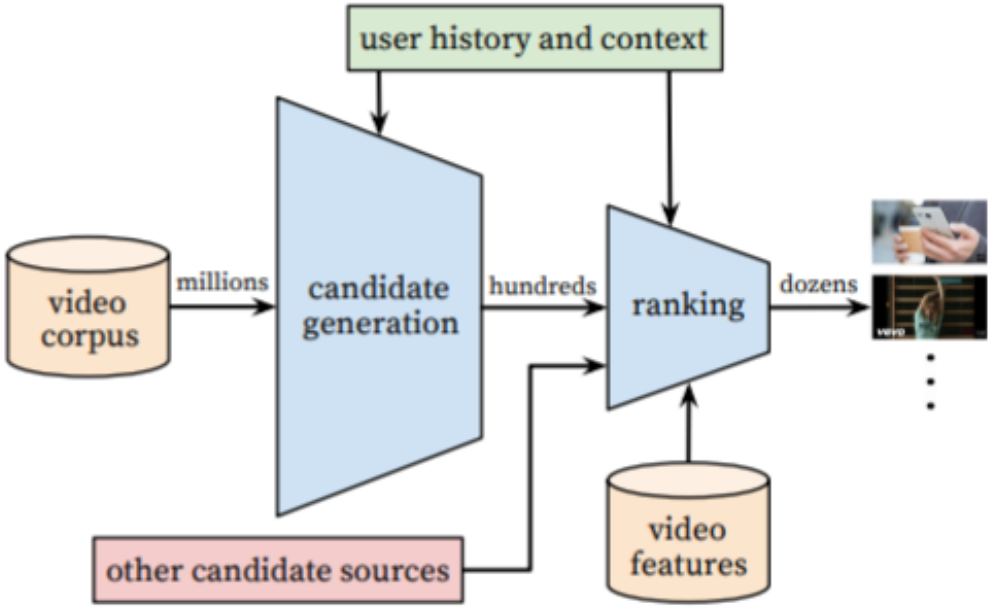
	Bias Unit	Hidden 1	Hidden 2
Harry Potter	-0.82602559	-7.08986885	4.96606654
Avatar	-1.84023877	-5.18354129	2.27197472
LOTR 3	3.92321075	2.51720193	4.11061383
Gladiator	0.10316995	6.74833901	-4.00505343
Titanic	-0.97646029	3.25474524	-5.59606865
Glitter	-4.44685751	-2.81563804	-2.91540988

Key Idea: RBMs learn a probability distribution over your data

- Learn latent variables which can better explain data (genres)
- RBMs outperform SVD
- RBMs in top Netflix Prize solutions
- Reconstruct missing data, i.e. unrated items

Paper 4: State-of-the-art

2016. AMC. Covington; Adams; Sargin. (Google)
"Deep Neural Networks for YouTube Recommendations"

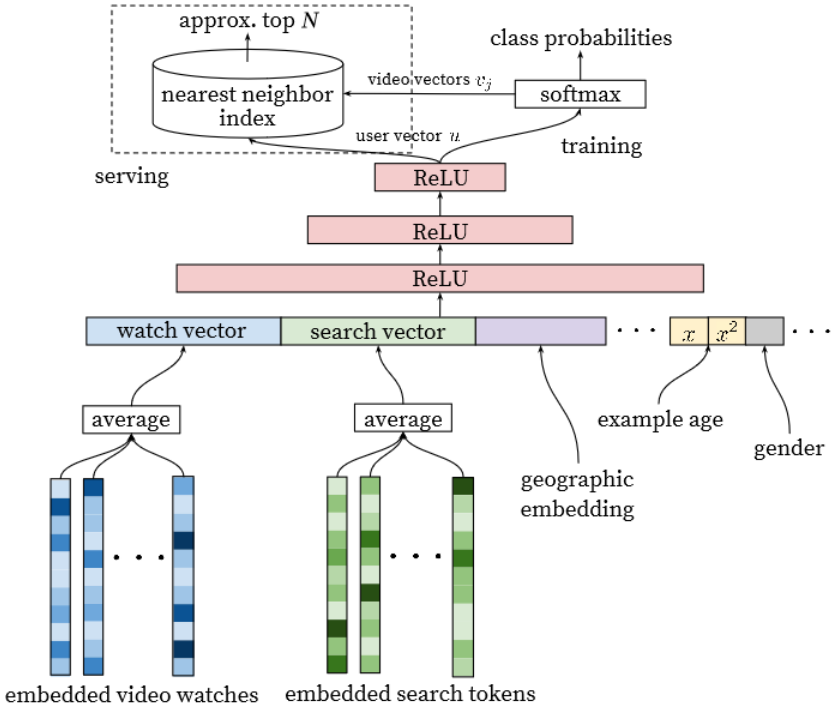


Key Idea: Predict the *next watched* video for a user

- Generate candidate videos by learning an embedding
- Predict a rank by incorporating content features
- Expected watch time vs. clicks

Paper 4: State-of-the-art

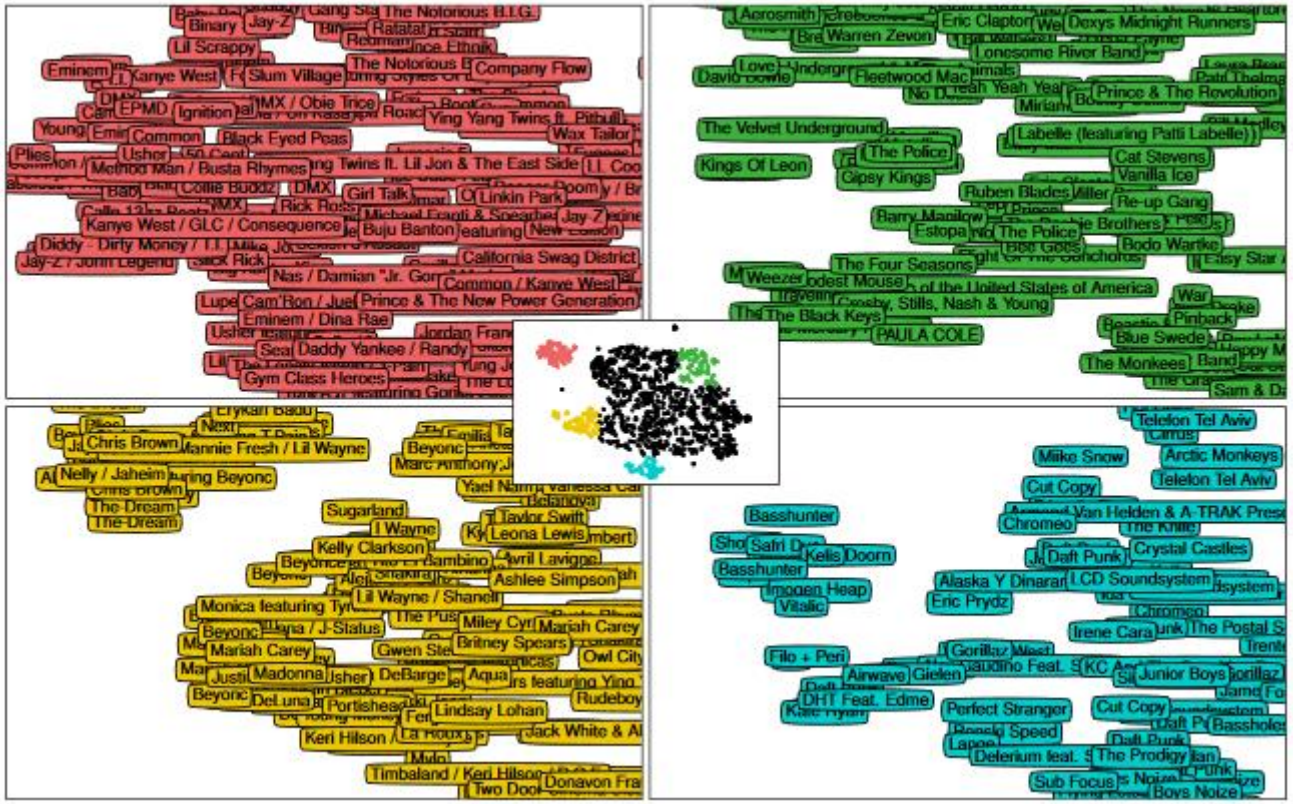
- Deep learning for non-linear embedding's
- Massive scale: 1B parameters 100B data points.
- Content-based raking.



Paper 5: Possible Future Direction

2013. NIPS. Oord; Dieleman; Schrauwen. (Ghent University)

"Deep content-based music recommendation"



Key Idea: Learn latent factors from song content

Select Topics Not Covered (of Many)

- Large scale data
  - Spark, Hadoop, etc.
  - Data collection
- Privacy Concerns
  - Differential Privacy
- Selecting performance metrics
  - View time vs. Clicks
- Baseline rating estimations

Deep Learning for Understanding the Human

- General Applications
  - Face Detection
  - Gaze Estimation
  - Face Recognition
  - Activity Recognition
  - Emotion Recognition
  - Body Pose Estimation
  - Gesture Recognition
  - Speech Recognition
  - Recommendation Systems
  - Natural Language Understanding
  - Dialogue Systems
  - Emotion Recognition
- Special Applications
  - Glance Classification
  - Cognitive Load Estimation
  - Human Vision Simulation

- Human-Centered AI during Learning Phase
  - Machine Teaching:
    Methods for efficient supervised learning
    (Improve annotation and learning algorithms)
  - Human-in-the-Loop Reward Engineering:
    Encoding human values into learning process
- Human-Centered AI during Real-World Operation
  - Human Sensing:
    Methods for perceiving the human state (physical, mental, social)
  - Human-Robot Interaction Experience:
    Methods for an immersive, meaningful interaction
  - AI Safety:
    Methods for effective supervision of machines (ethics & safety)

Series of lectures on aspects of the above will be released on:

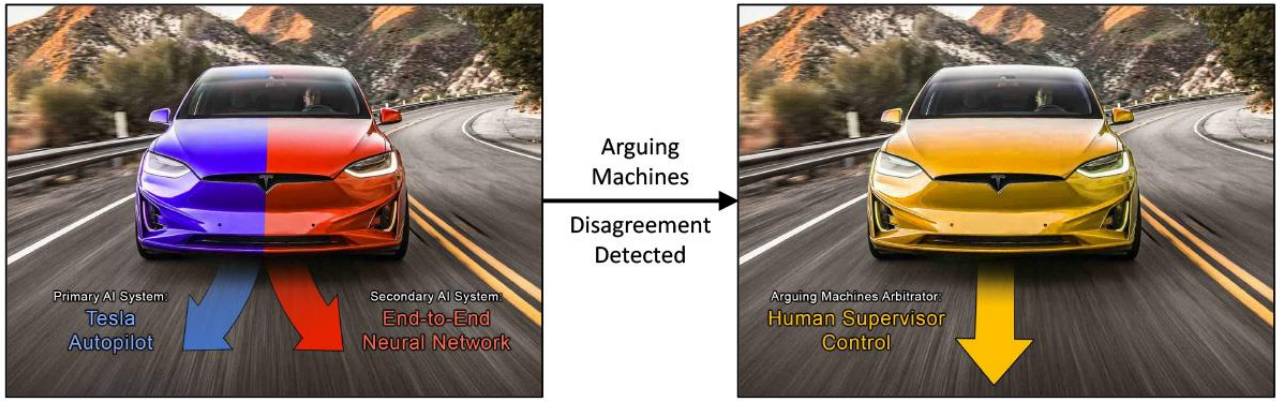
deeplearning.mit.edu

- Human-Centered AI during Learning Phase
  - Machine Teaching:
    Methods for efficient supervised learning
    (Improve annotation and learning algorithms)
  - Human-in-the-Loop Reward Engineering:
    Encoding human values into learning process
- Human-Centered AI during Real-World Operation
  - Human Sensing:
    Methods for perceiving the human state (physical, mental, social)
  - Human-Robot Interaction Experience:
    Methods for an immersive, meaningful interaction
  - AI Safety:
    Methods for effective supervision of machines (ethics & safety)

Series of lectures on aspects of the above will be released on:

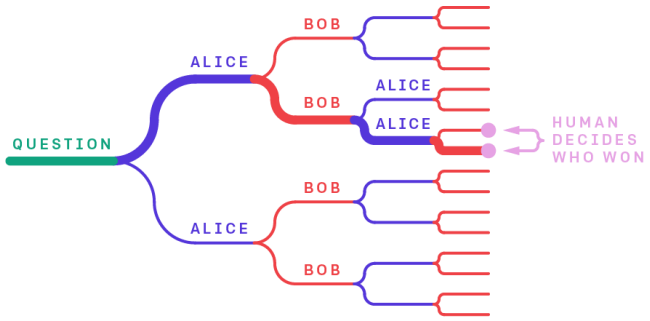
deeplearning.mit.edu

Arguing Machines + Safety via Debate

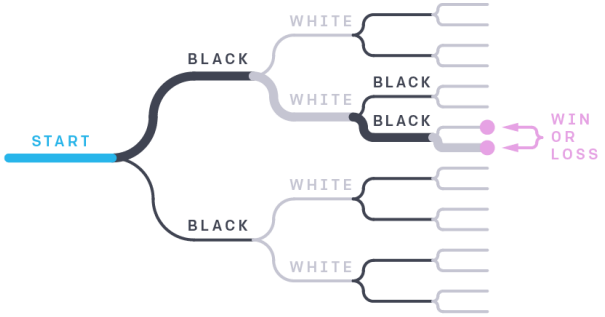


OpenAI

Tree of all possible debates

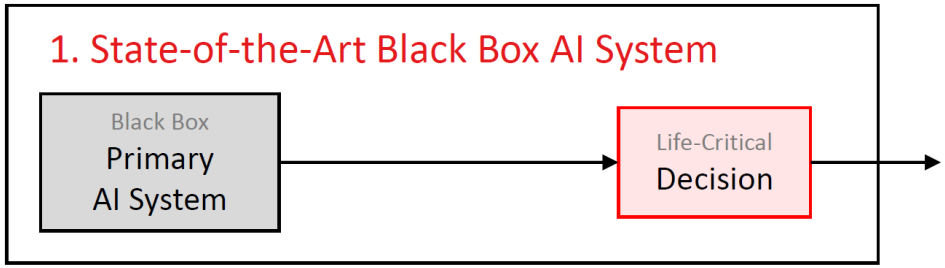


Tree of all possible Go moves

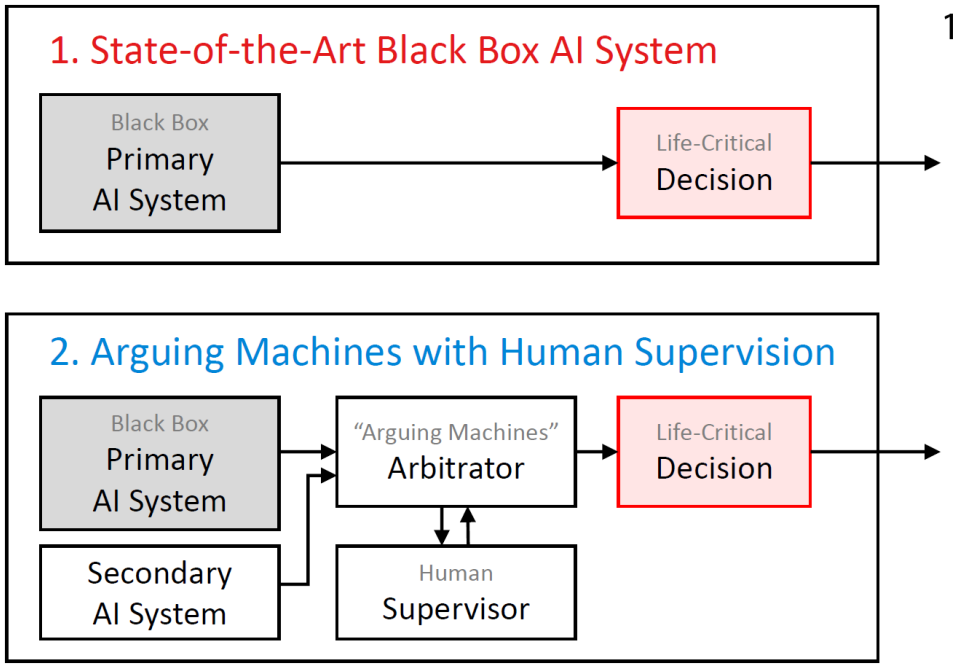


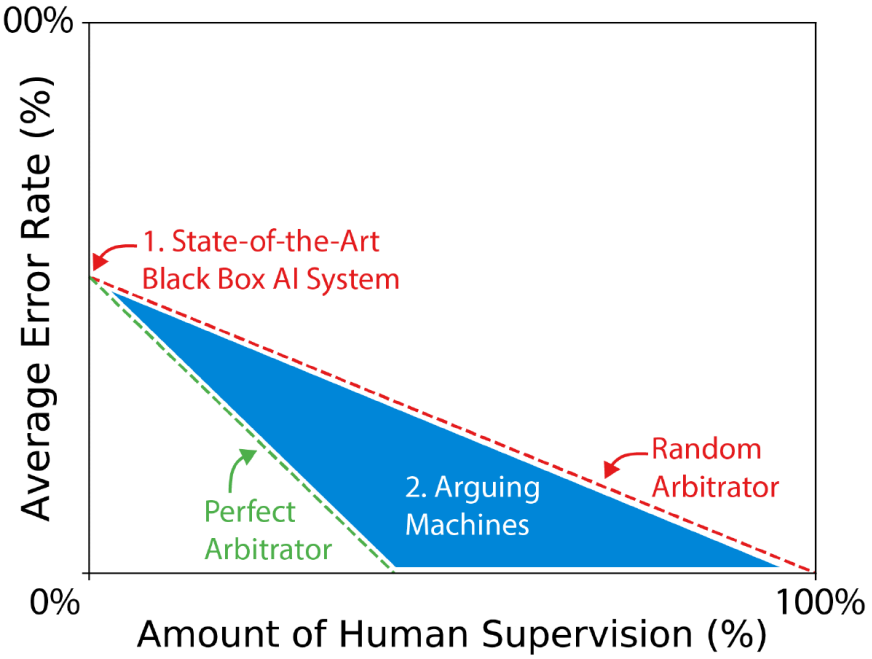
For the full list of references visit:
https://hcai.mit.edu/references [346, 347]

Human Supervision of AI Systems that Make Life-Critical Decision

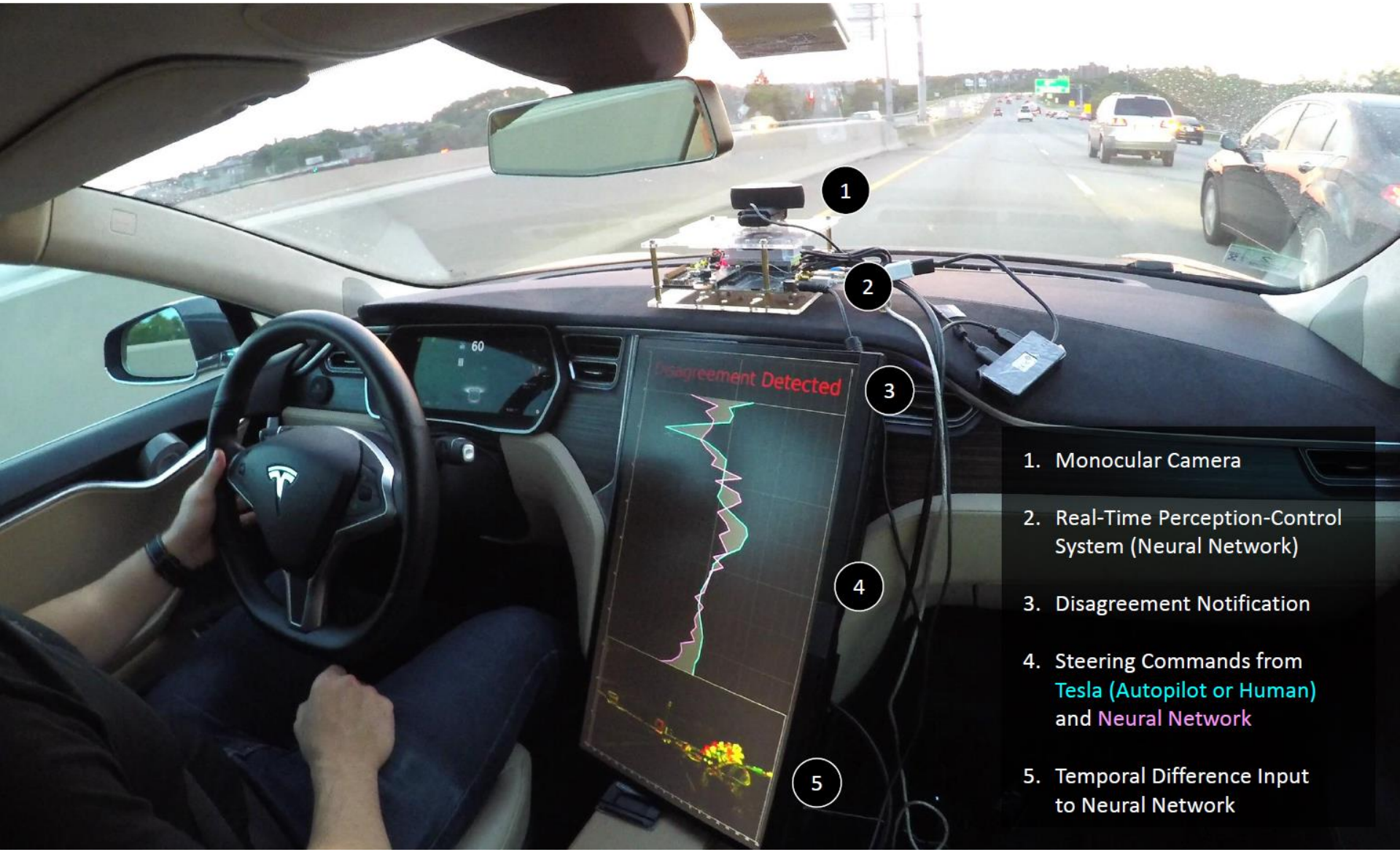


Human Supervision of AI Systems that Make Life-Critical Decision



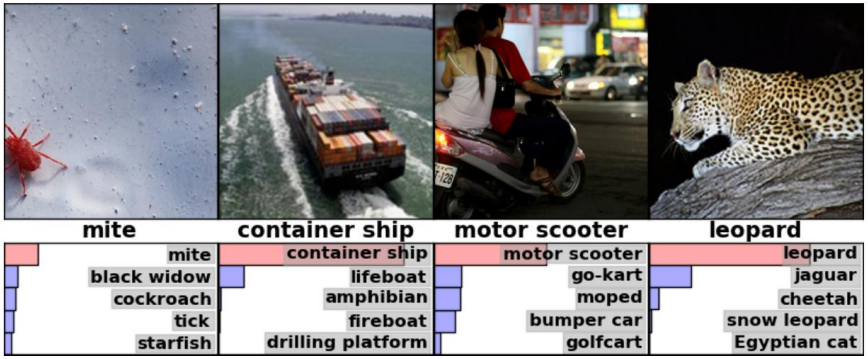


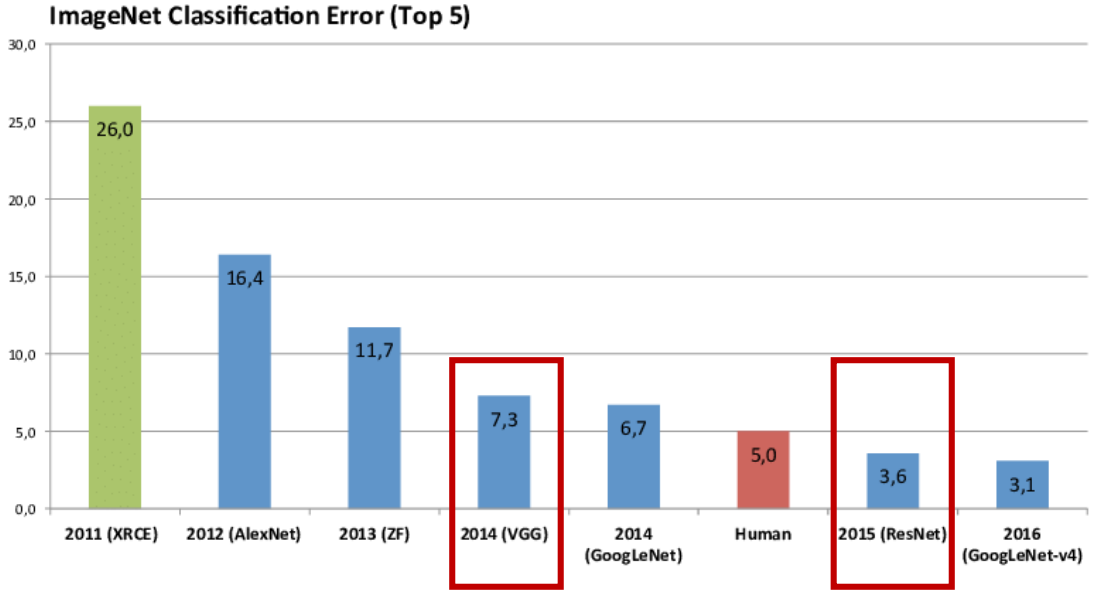
Arguing Machines:
Tesla Autopilot and End-to-End Neural Network



Arguing Machines:
Tesla Autopilot and End-to-End Neural Network

Arguing Machines: Image Classification





Arguing Machines: Image Classification

Method	Top-1 Error (%)	Top-5 Error (%)
ResNet-50 (primary system)	25.2	8.0
VGG-16 (secondary system)	29.0	10.1
Ensemble: ResNet-50, VGG-16	24.4	7.8
Random Arbitrator	19.3	6.2
Arguing Machines	**10.7**	**2.8**

Adding human supervision via arguing machines we:

Reduce error from 8.0% to 2.8%

Ground Truth:

Wine bottle

Arguing Machines:

Disagree

ResNet Prediction:

0.93 - Paper towel

0.03 - Toilet tissue

0.03 - Bath towel

VGG Prediction:

0.25 - Seat belt

0.10 - Paper towel

0.08 - Syringe

Ground Truth:

Mailbox

Arguing Machines:

Disagree

ResNet Prediction:

0.21 - Garbage truck

0.14 - Tow truck

0.14 - Steam locomotive

VGG Prediction:

0.41 - Traffic light

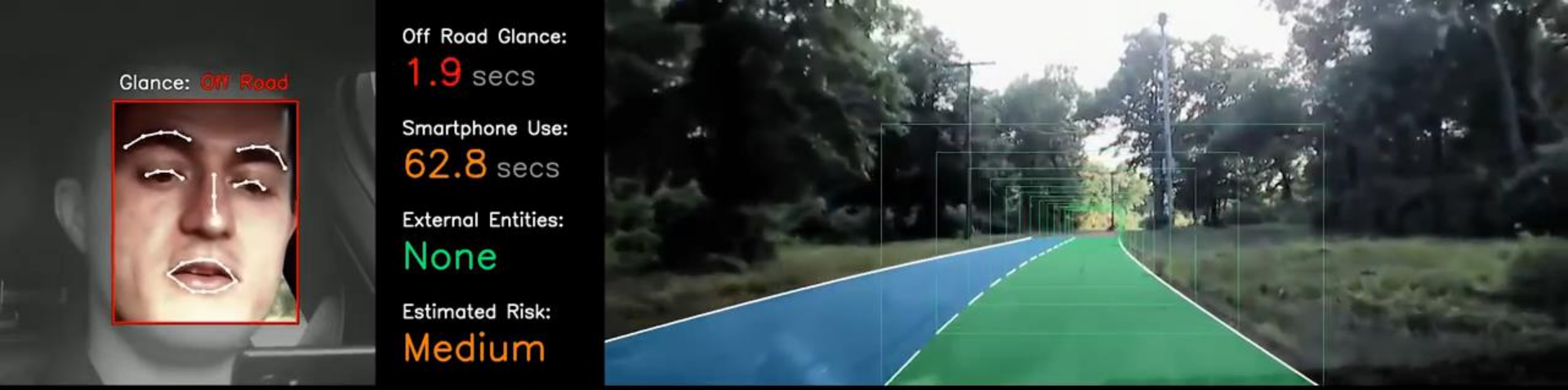
0.37 - Mailbox

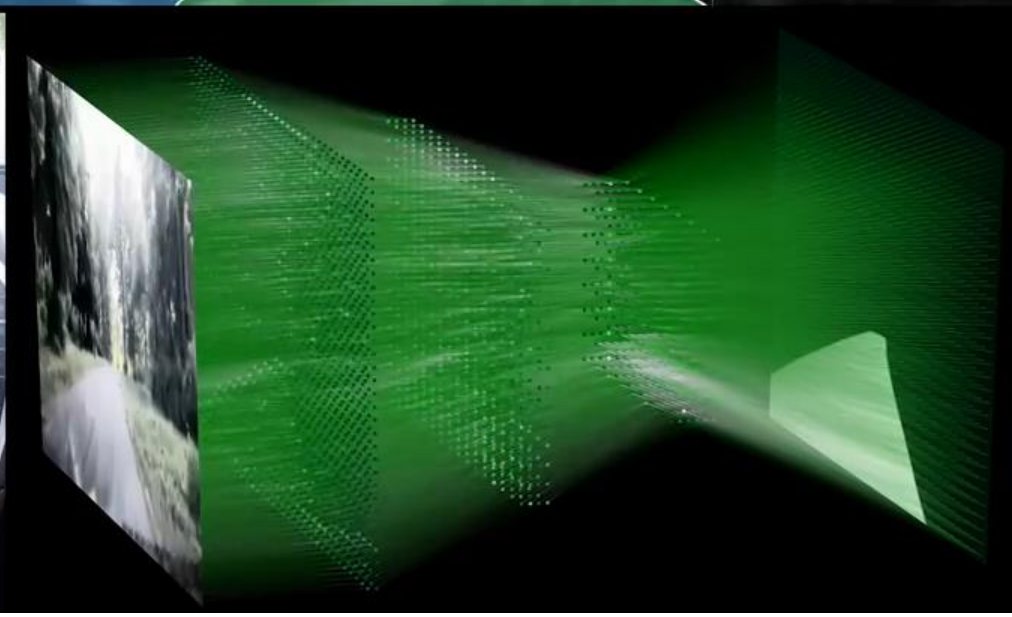
0.09 - Street sign

- Human-Centered AI during Learning Phase
  - Machine Teaching:
    Methods for efficient supervised learning
    (Improve annotation and learning algorithms)
  - Human-in-the-Loop Reward Engineering:
    Encoding human values into learning process
- Human-Centered AI during Real-World Operation
  - Human Sensing:
    Methods for perceiving the human state (physical, mental, social)
  - **Human-Robot Interaction Experience:**
    Methods for an immersive, meaningful interaction
  - AI Safety:
    Methods for effective supervision of machines (ethics & safety)

Series of lectures on aspects of the above will be released on:

deeplearning.mit.edu





YouTube

Playing Guitar in MIT Autonomous Vehicle (Driver...

MIT Human-Centered Autonomous Vehicle

Arguing Machines: Tesla Autopilot vs Neural Network

https://hcai.mit.edu/hcav

Human-Centered Autonomous Vehicle Systems: Principles of Effective Shared Autonomy

Lex Fridman
Massachusetts Institute of Technology (MIT)
fridman@mit.edu

Principle 1
Shared Autonomy:
Beyond Levels of Automation

Principle 2
Learn from Data:
Machine Learning at Every Level

Principle 3
Human Sensing:
Multi-Modal Understanding of the Human

Principle 4
Shared Perception-Control:
A Second Pair of Eyes and Hands

Principle 5
Deep Personalization:
Human Inside the Machine

Principle 6
Imperfect by Design:
Flaws Are Features

Principle 7
System-Level Experience:
Greater Than the Sum of Its Parts

- Human-Centered AI during Learning Phase
  - Machine Teaching:
    Methods for efficient supervised learning
    (Improve annotation and learning algorithms)
  - Human-in-the-Loop Reward Engineering:
    Encoding human values into learning process
- Human-Centered AI during Real-World Operation
  - Human Sensing:
    Methods for perceiving the human state (physical, mental, social)
  - Human-Robot Interaction Experience:
    Methods for an immersive, meaningful interaction
  - AI Safety:
    Methods for effective supervision of machines (ethics & safety)

Series of lectures on aspects of the above will be released on:

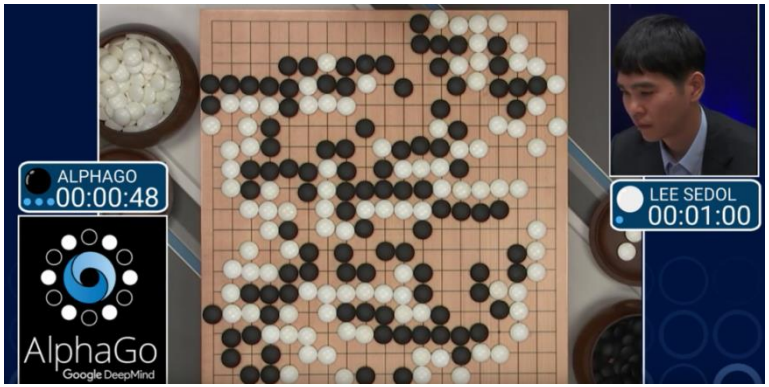
deeplearning.mit.edu

Scalability Requirement for Human-Centered AI:

Learn from Humans while Being Useful to Humans

- **Parasitism:** One organism benefits at the cost of another.
  - **In AI:** Models learn at the cost of human labor (brute-force annotation).

- **Symbiosis:** Both organisms benefit.
  - **In AI:** Models learn as a side effect of interacting with humans.



Human-Centered Artificial Intelligence

Human-Centered
Artificial Intelligence

Neuroscience
Cognitive Science
Psychology
Sociology
Game Theory
Human Factors
Design & Ergonomics
Natural Language Processing
Human Robot Interaction
Human Computer Interaction
Deep Learning
Reinforcement Learning
Robotics

Thank You

Website:

deeplearning.mit.edu

- Videos and slides will be posted online
- Code will be posted on GitHub: https://github.com/lexfridman/mit-deep-learning