AI, People, and the Open World

Eric Horvitz

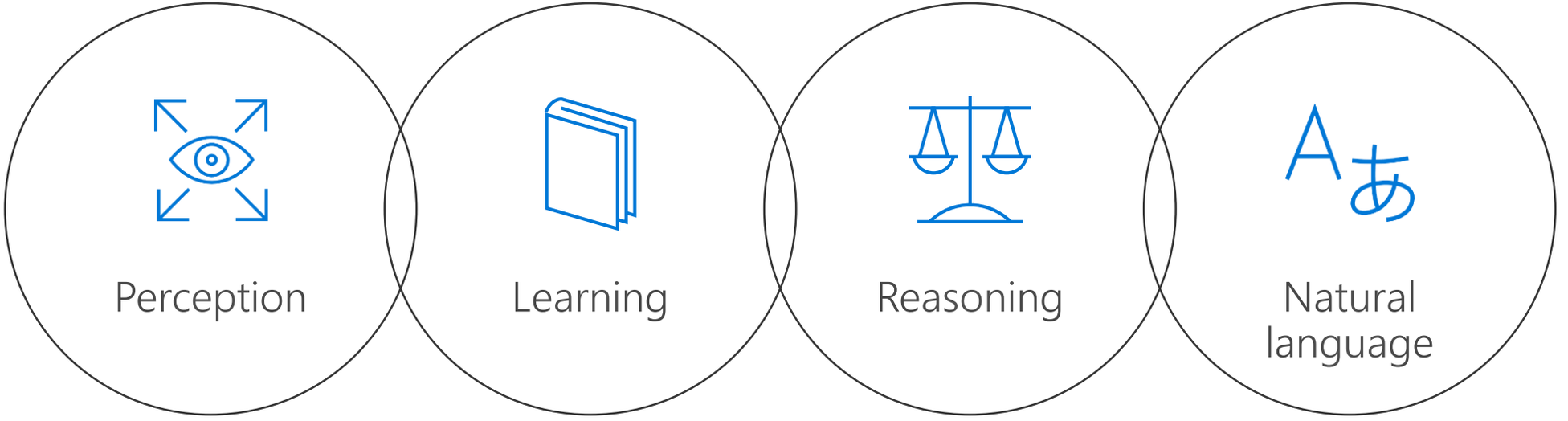
Technical Fellow and Director
Microsoft Research

August 16, 2017

Artificial Intelligence

"...to find how to make machines...solve kinds of problems now reserved for humans..." (1955)

Artificial Intelligence

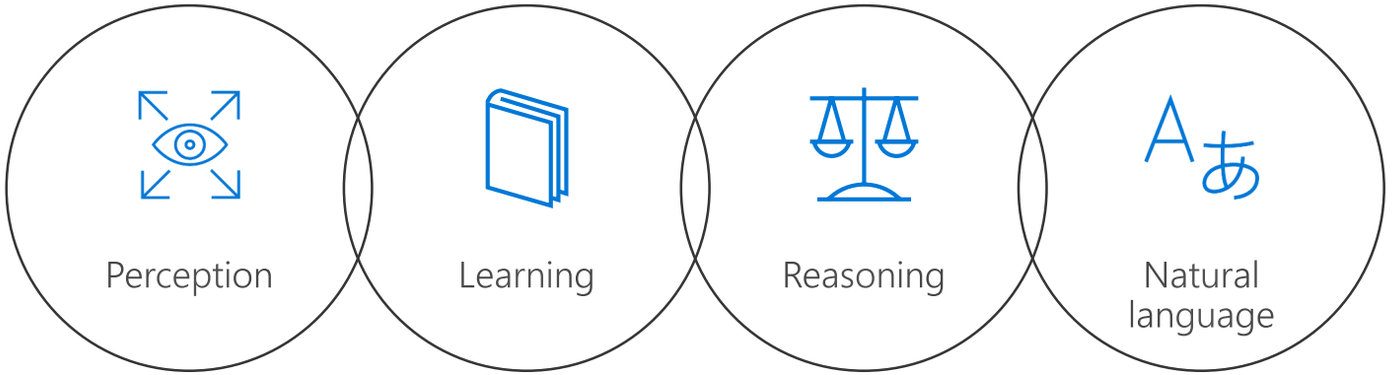


Artificial Intelligence

Reinforcement Learning

Computer vision

Deep Learning



Robotics

Probabilistic Graphical Models

Knowledge Representation

Speech and dialog

Decisions and plans

Algorithmic Game Theory

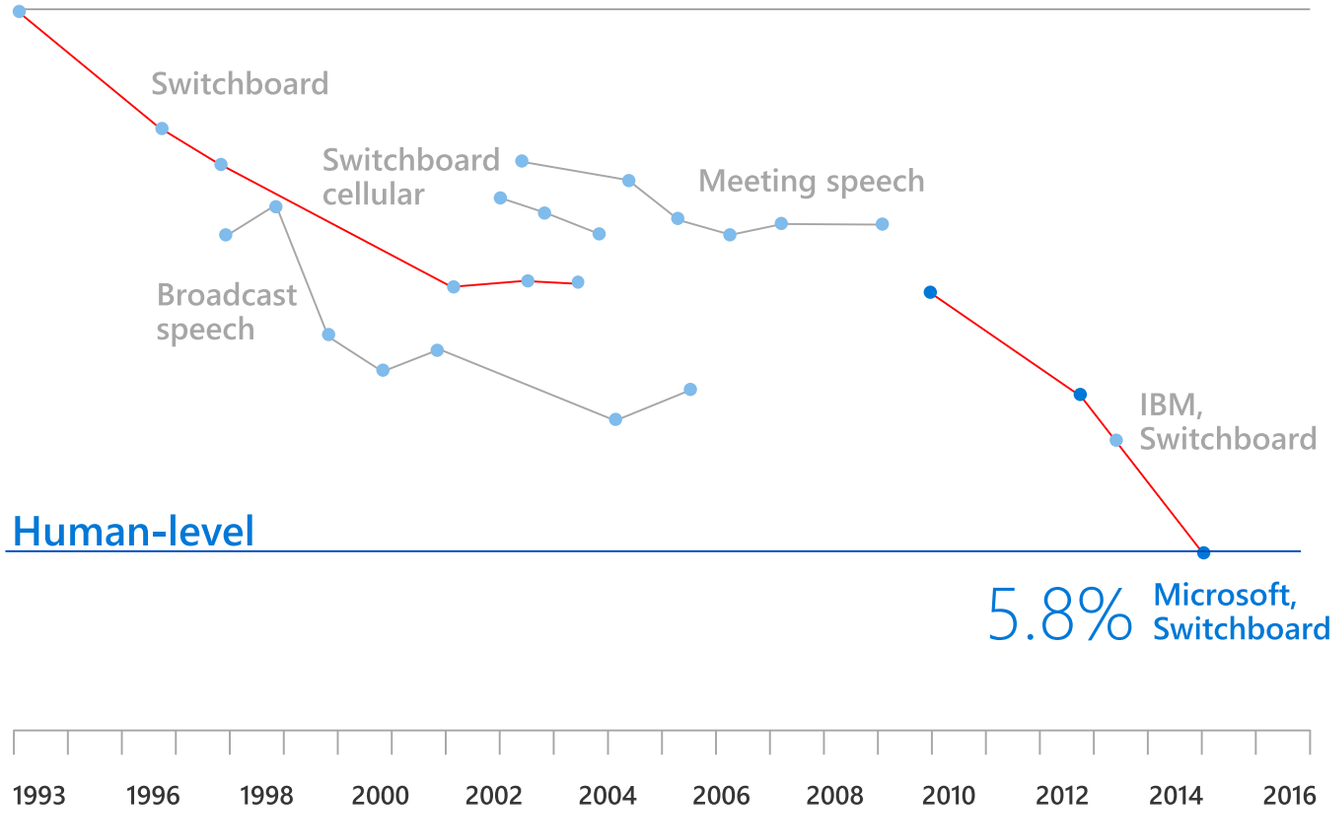
Multiple subdisciplines and research communities

Speech recognition

5.8% human error rate

Switchboard challenge

Speech-recognition word-error rate, %



Sources: Microsoft: research papers

Vision

152 layers

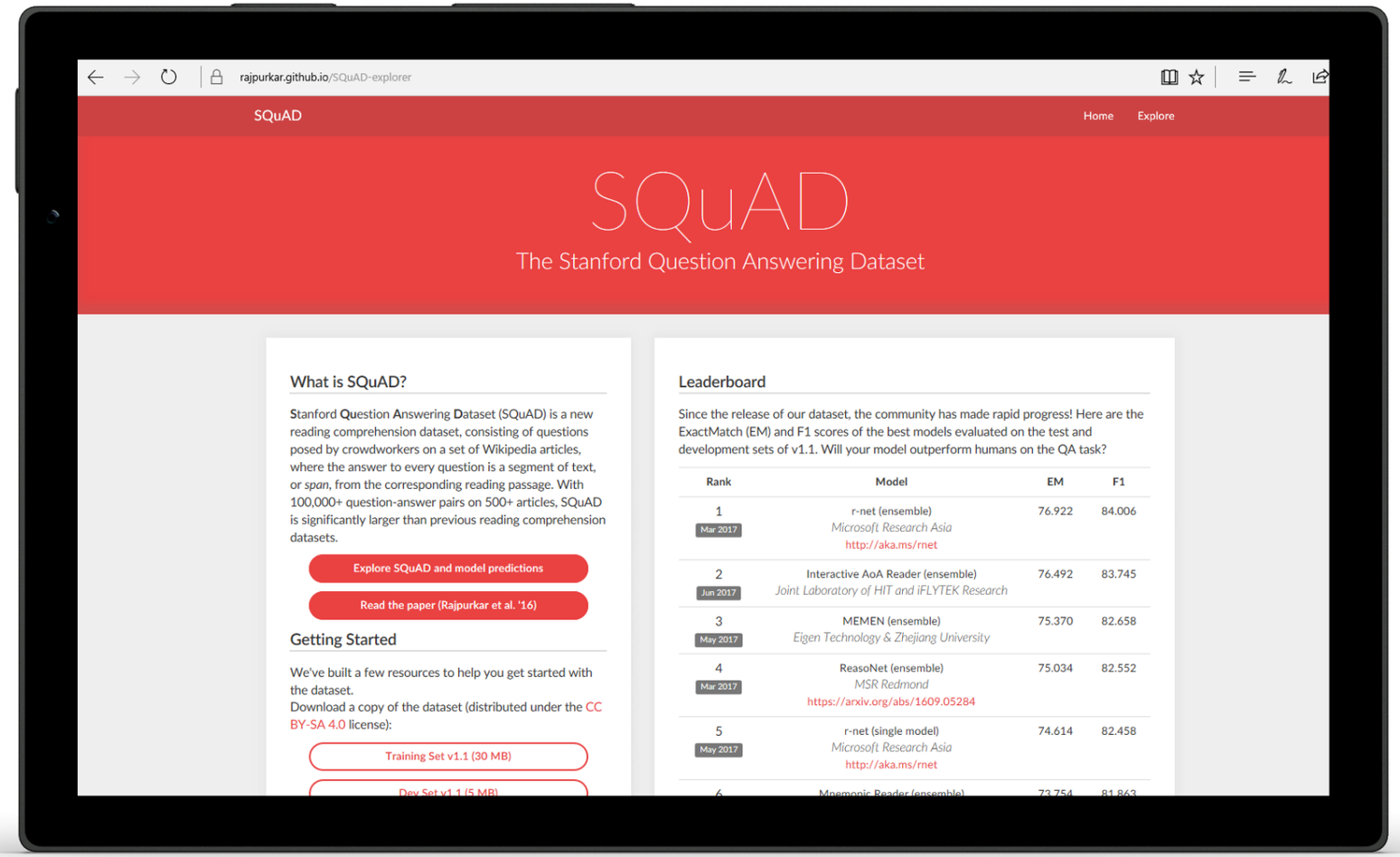
RESNET

Reading comprehension

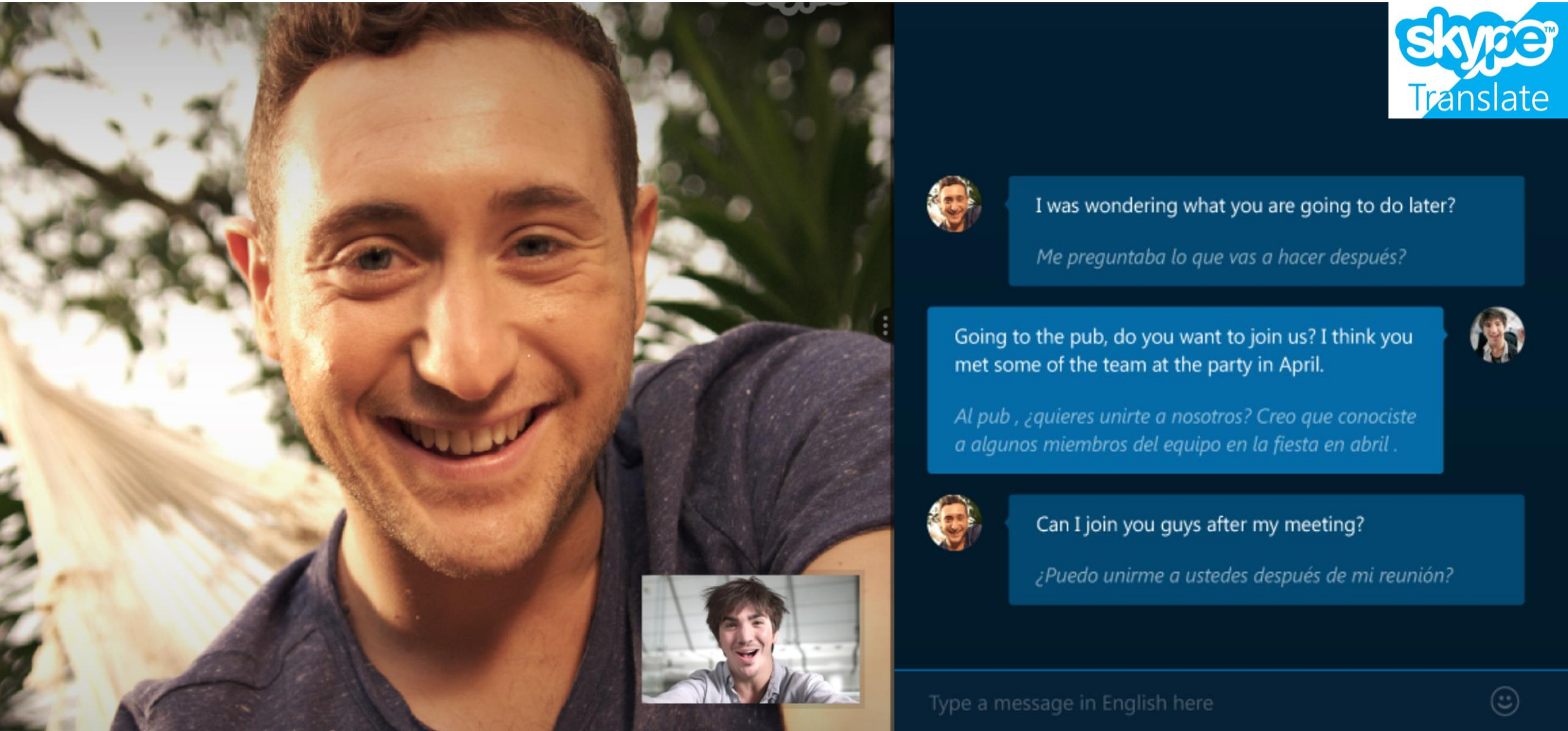
Answer questions about information on Wikipedia

SQuAD challenge

Stanford question & answer challenge

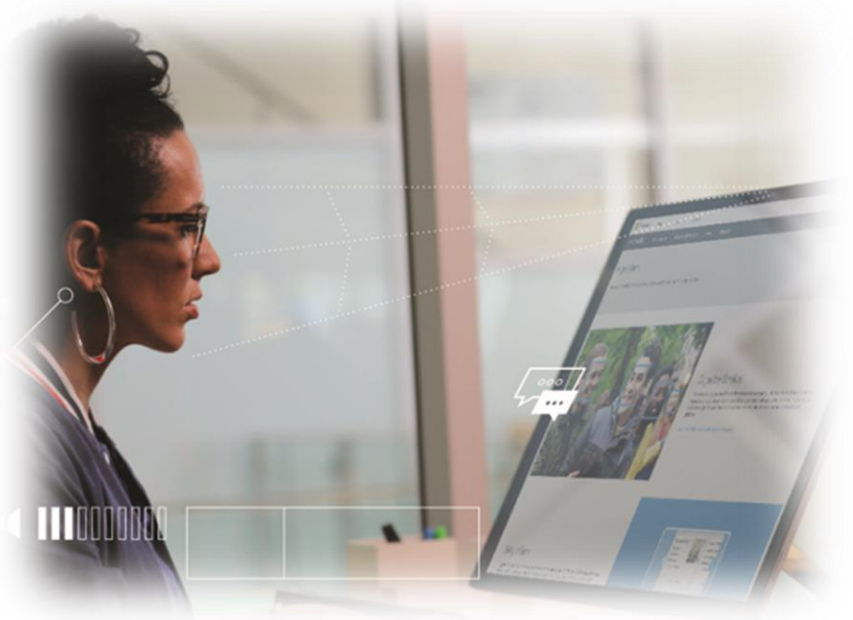


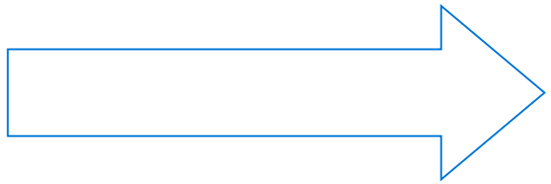
New capabilities



Lab environment

Open world



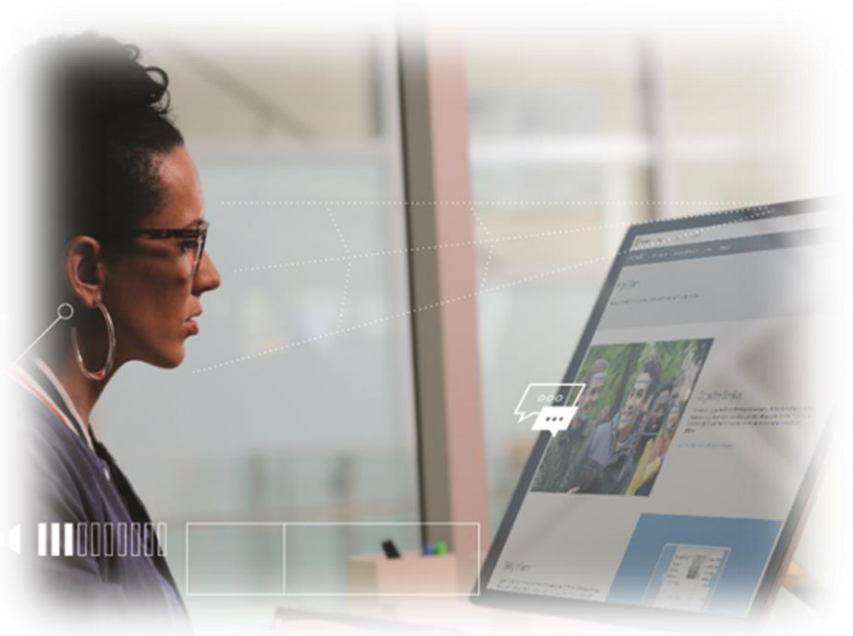


Competence

AI & People

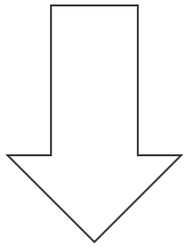
Ethical, legal, societal influences

Lab environment



Open world

Frame problem



Qualification problem

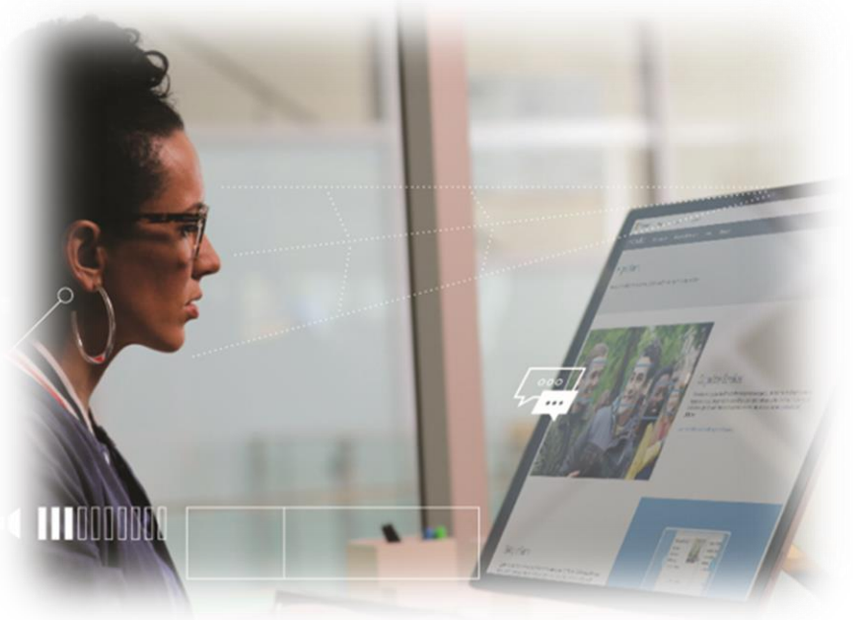
All preconditions?

Ramification problem

All effects of action?

Lab environment

Open world



Framing

Data

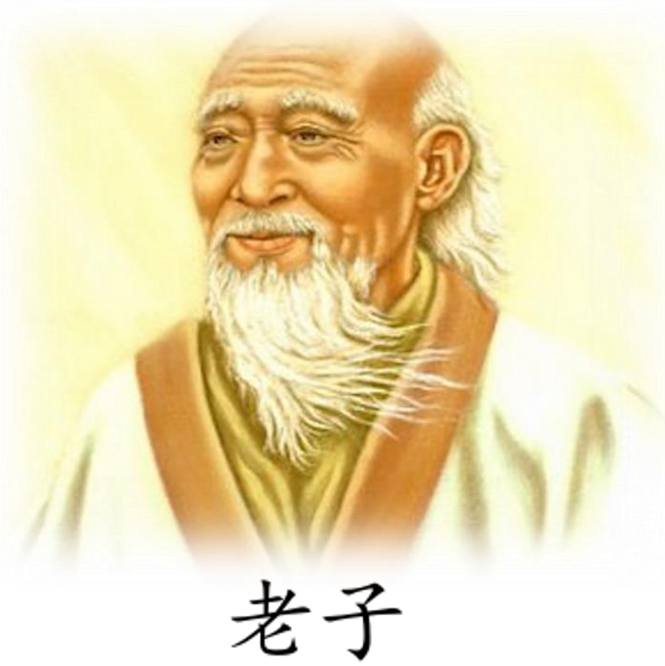
Fidelity

Representation

Inference

Learning about competency

Wisdom of the ages



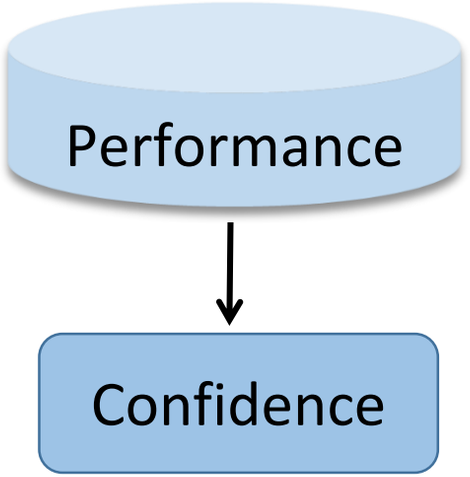
Knowing that you do not know is the best.

Not knowing that you do not know is an illness.

- Laozi, 500-600 BCE

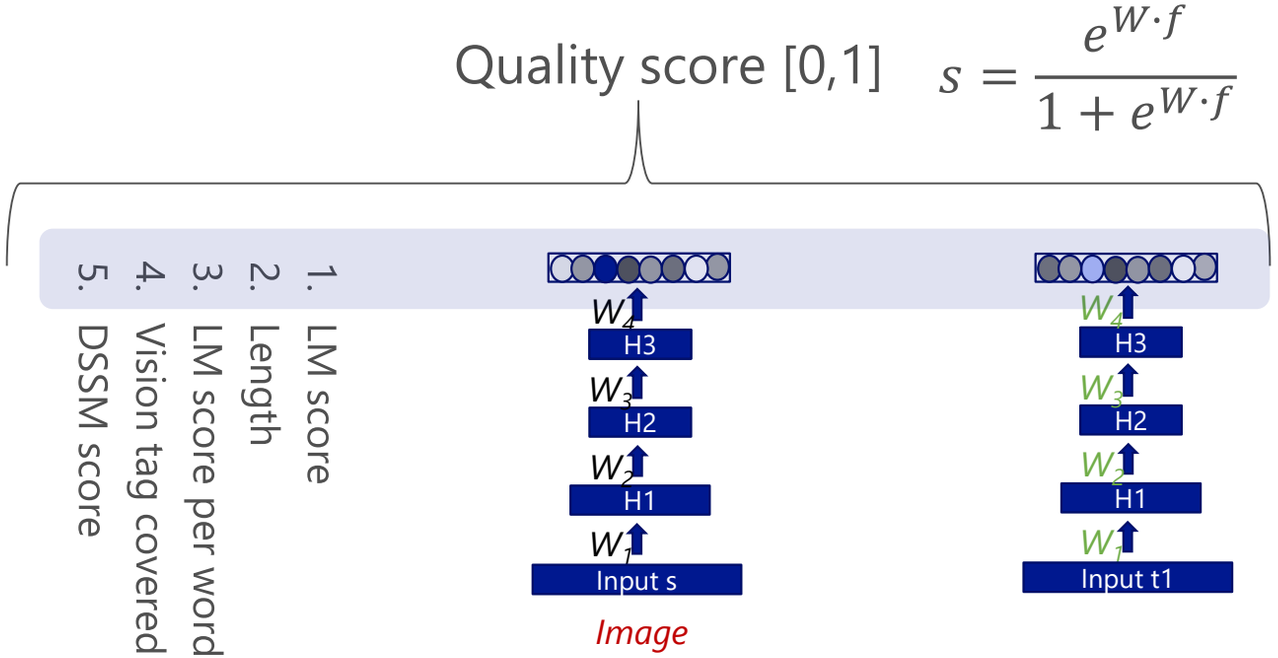
Learn about abilities & failures

Successes & failures



$p(\text{ fail} \mid E, t)$

Deep learning about deep learning performance



Caption:
a man holding a tennis racquet on a tennis court

Fang, et al., 2015

Grappling with Open-World Complexity

Reliable predictions of performance: *Known unknowns*

Grappling with Open-World Complexity

Reliable predictions of performance: *Known unknowns*


38
5
mph
Since 6:23 PM
1.6
3
591
Wh/mi
Since last charge
7.1
3
491
SPEED LIMIT
45
400
200
100
kW
Avg. Wh/mi
421

Grappling with Open-World Complexity

Reliable predictions of performance: *Known unknowns*

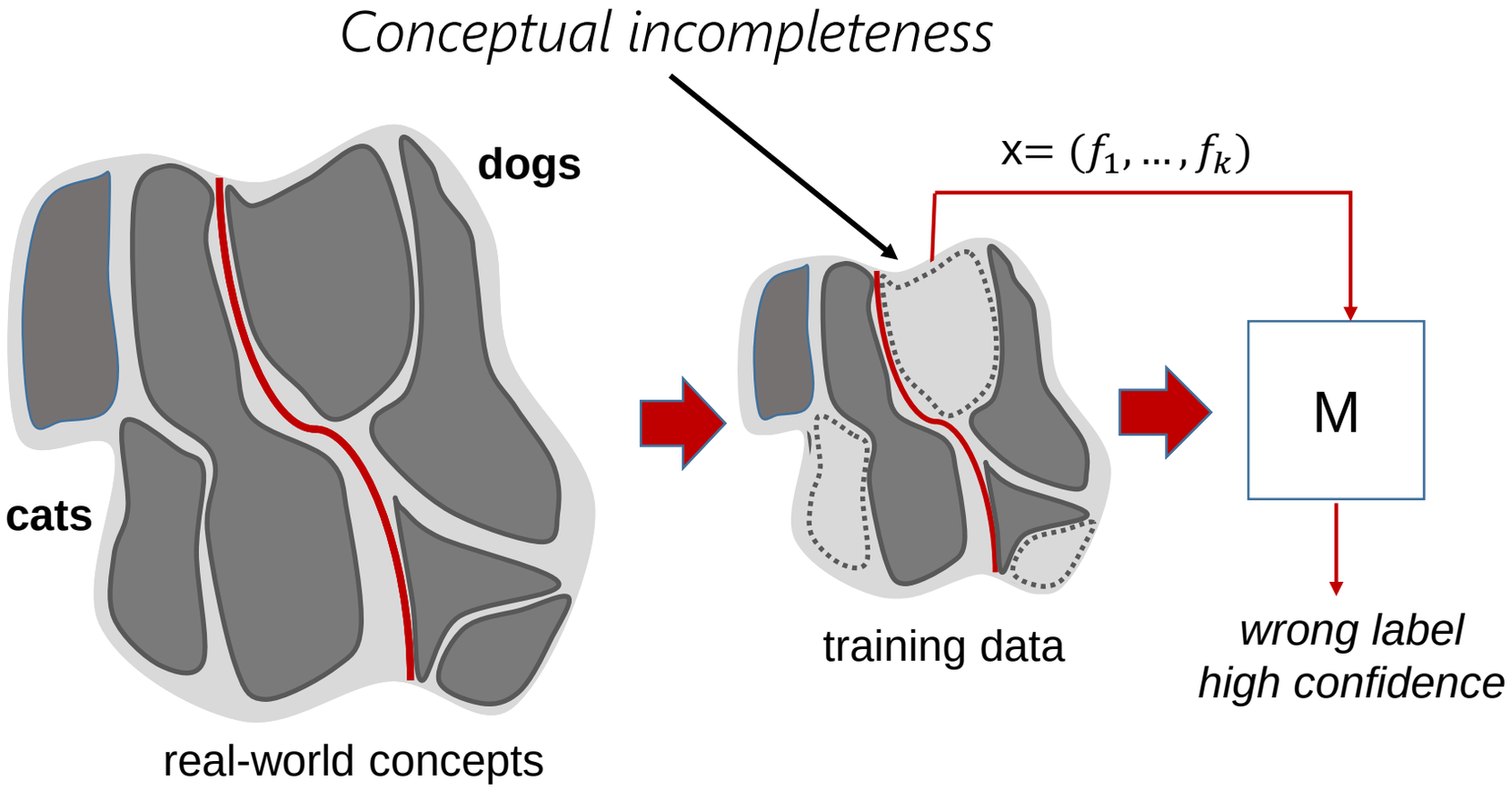
Challenge of *unknown unknowns*

Unknown unknowns

Directions

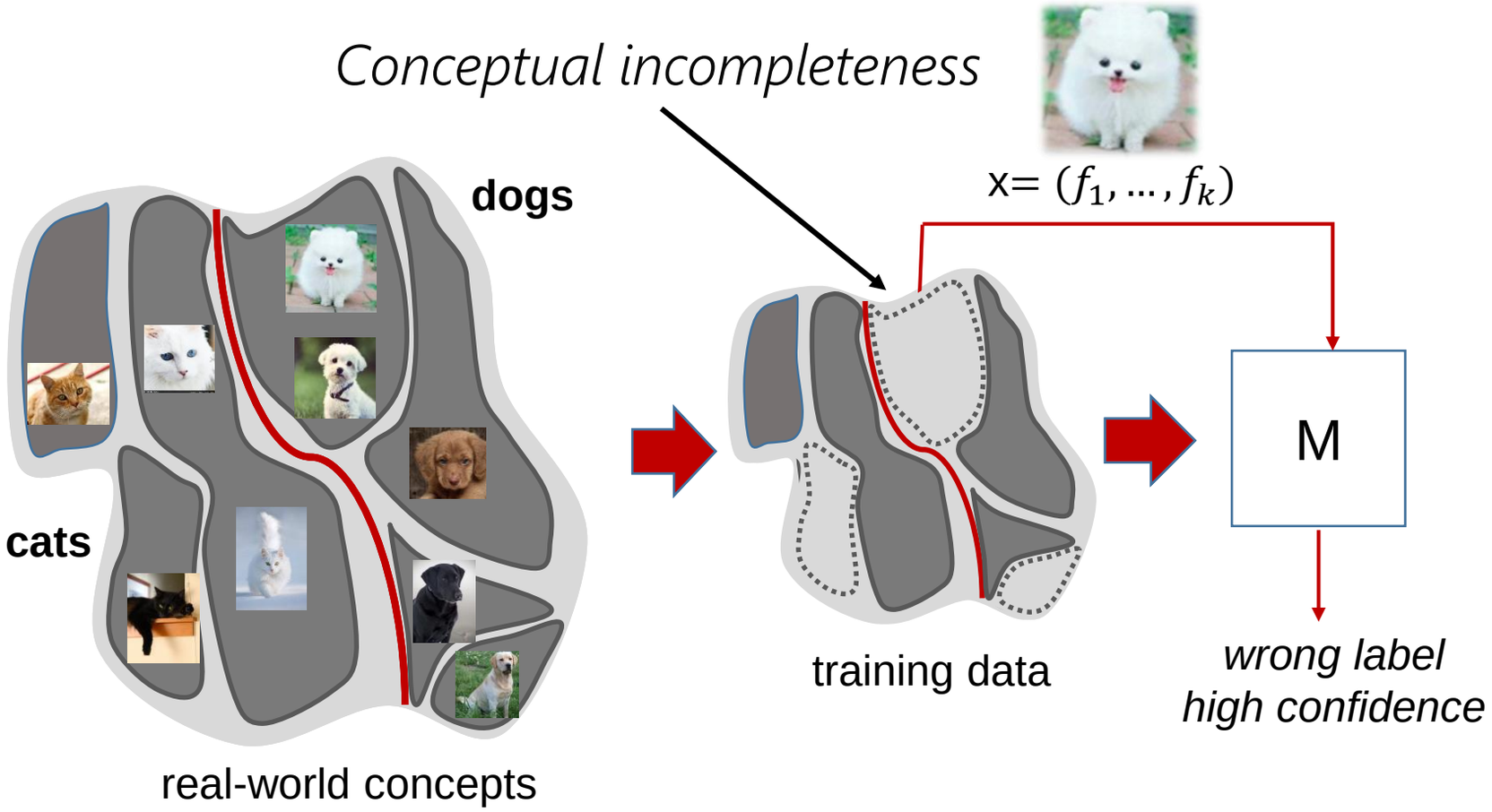
- Expanded real-world testing
- Algorithmic portfolios
- Failsafe designs
- People + machines

Identifying classifier blindspots



Lakkaraju, Kamar, Caruana, H, 2017.

Identifying classifier blindspots

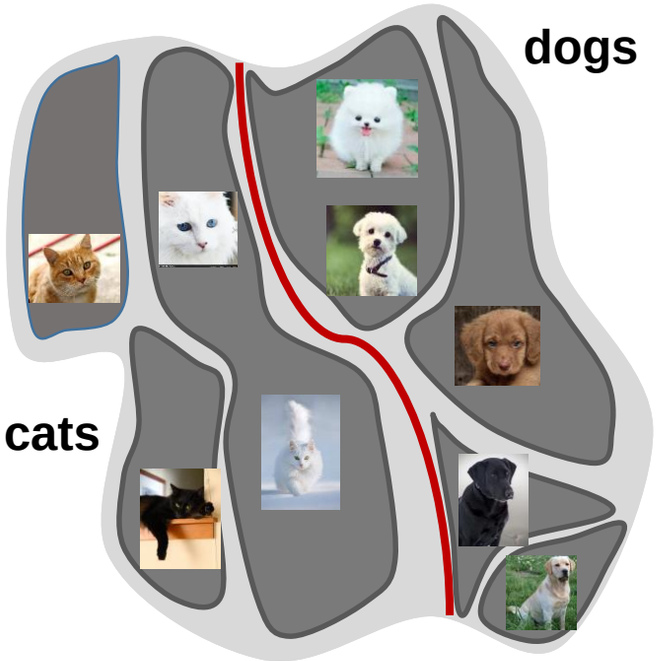


How to define & search regions of data space?

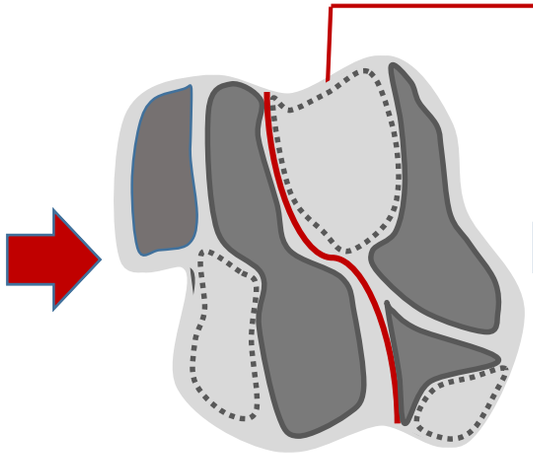
How to trade exploration and exploitation?

Lakkaraju, Kamar, Caruana, H, 2017.

Identifying classifier blindspots



$x = (f_1, \ldots, f_k)$

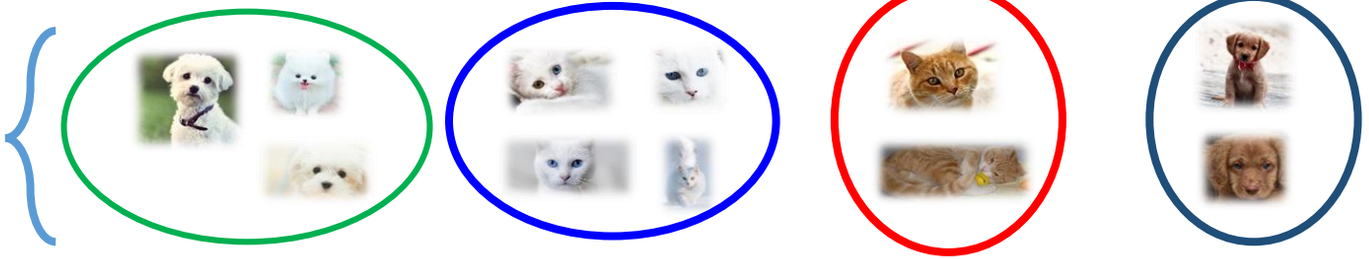


training data

M

wrong label
high confidence

Partition space by attributes



White Dogs | White Cats | Brown Cats | Brown Dogs

Lakkaraju, Kamar, Caruana, H, 2017.

Data scarcity

Directions

Transfer learning

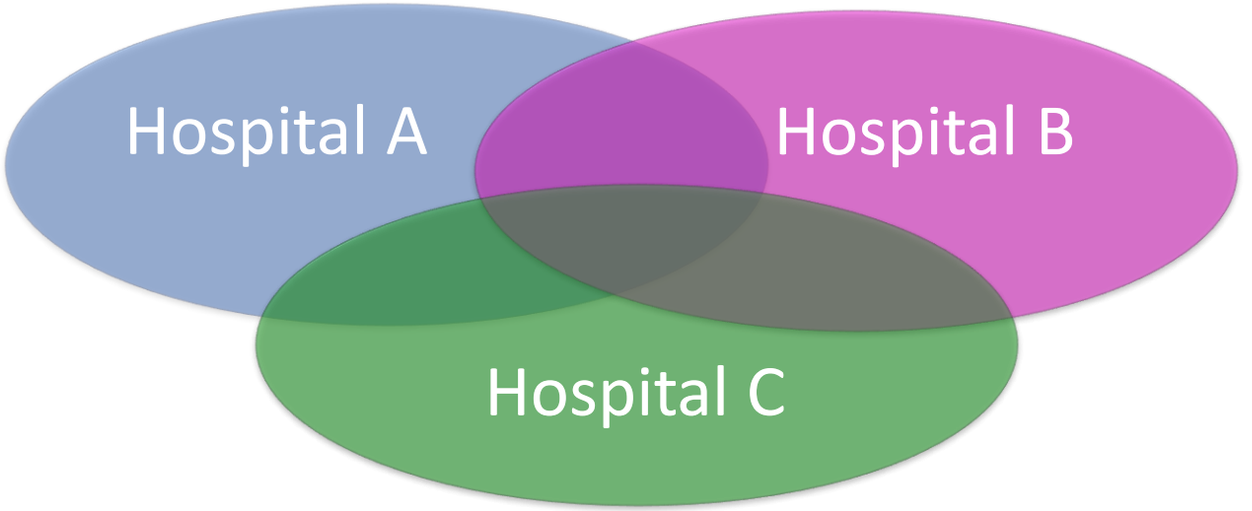
Learn from rich simulations

Learn generative models

Transfer learning opportunity

Site-specific data

- Observations, definitions
- Patients, prevalencies
- Covariate dependencies



Predict of risk of infection

- A: Community hosp: 10k pts/yr
- B: Acute care & teaching: 15k/yr
- C: Major teaching & research: 40k/yr

J. Wiens, J. Guttag, H, 2015.

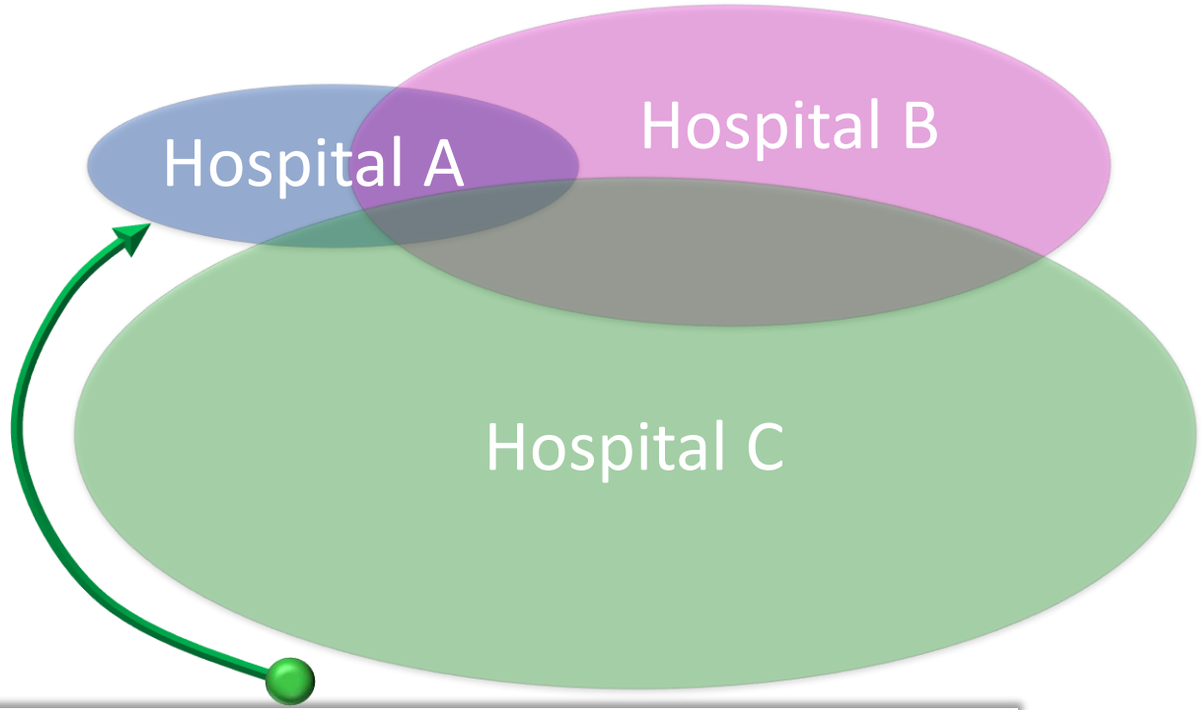
Transfer learning opportunity

Site-specific data

Observations, definitions

Patients, prevalencies

Covariate dependencies



Research and applications

A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions

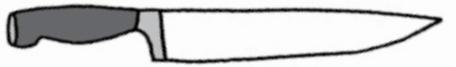
Jenna Wiens,[1] John Guttag,[1] Eric Horvitz[2]

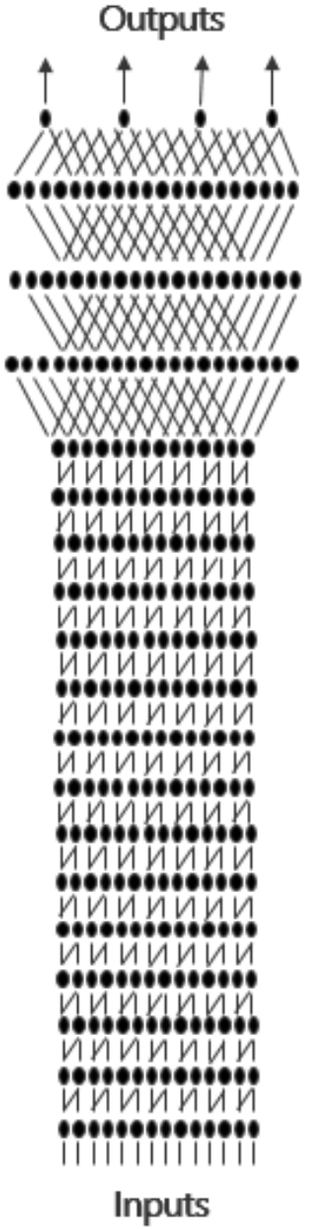
Embedded deep transfer learning

Less data with better features

ImageNet 1000, 1M photos

Cut off top layer





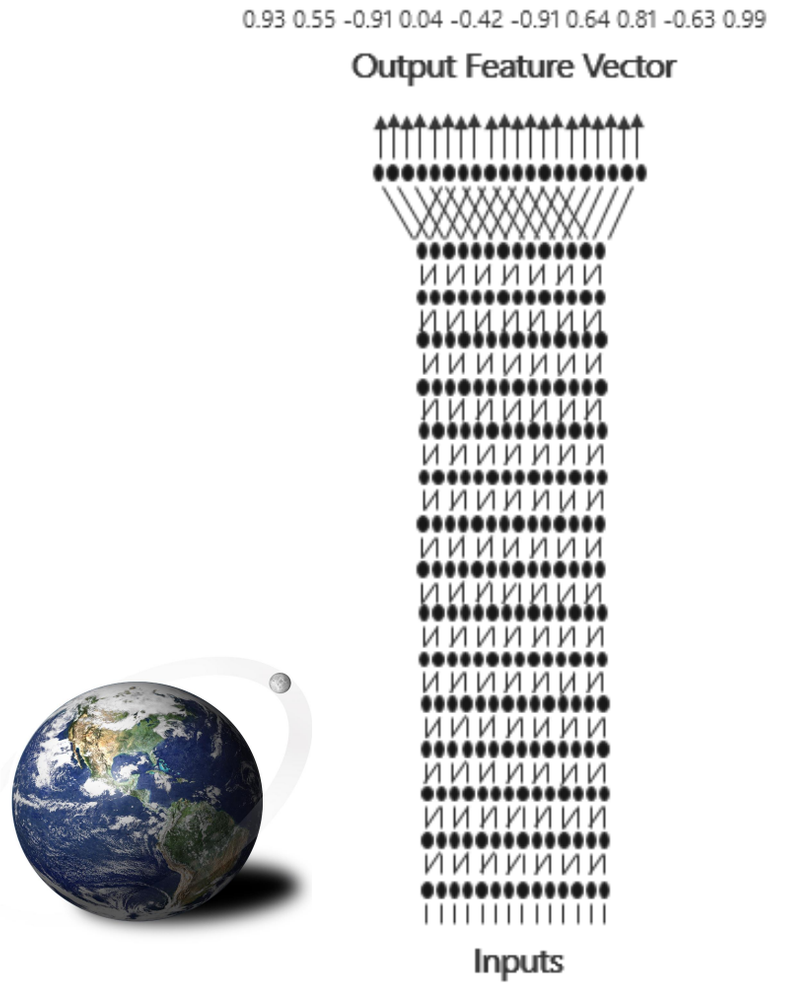
M. Gabel, R. Caruana, M. Philipose, O. Dekel

Embedded deep transfer learning

Less data with better features

ImageNet 1000, 1M photos

Cut off top layer



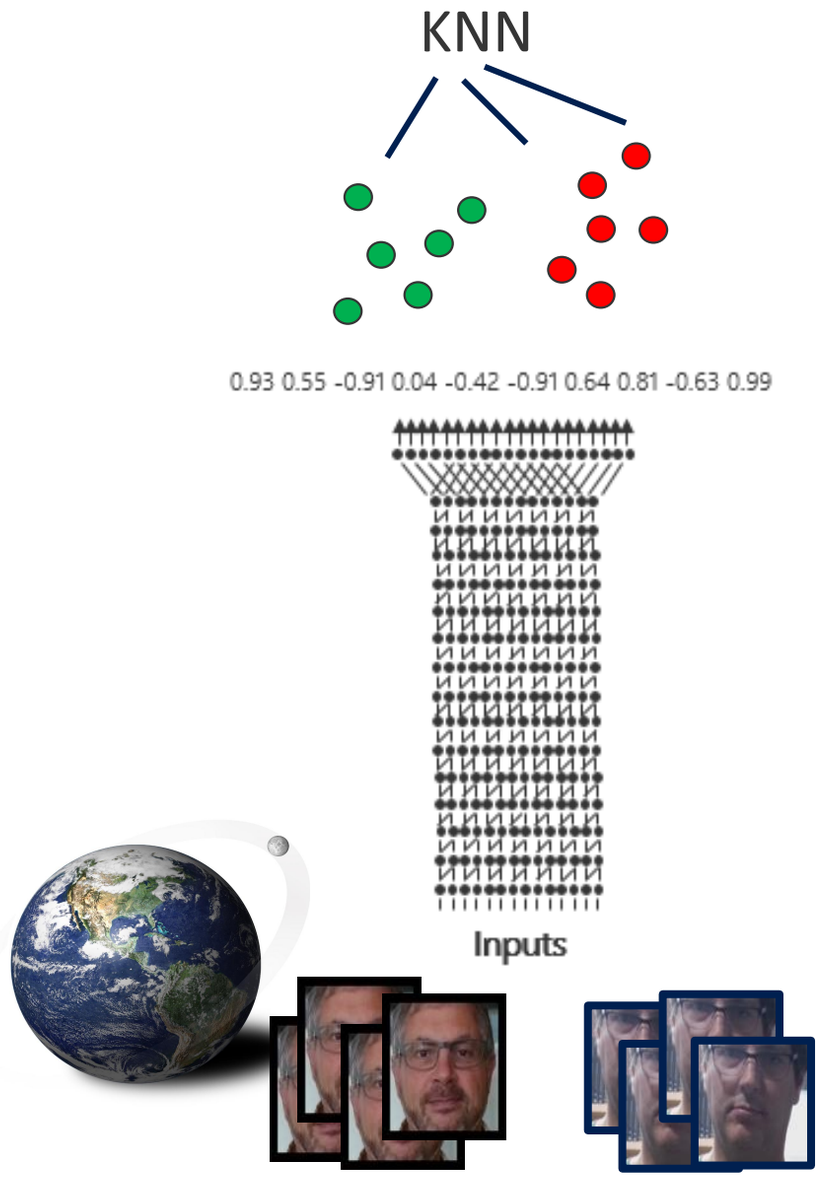
M. Gabel, R. Caruana, M. Philipose, O. Dekel

Embedded deep transfer learning

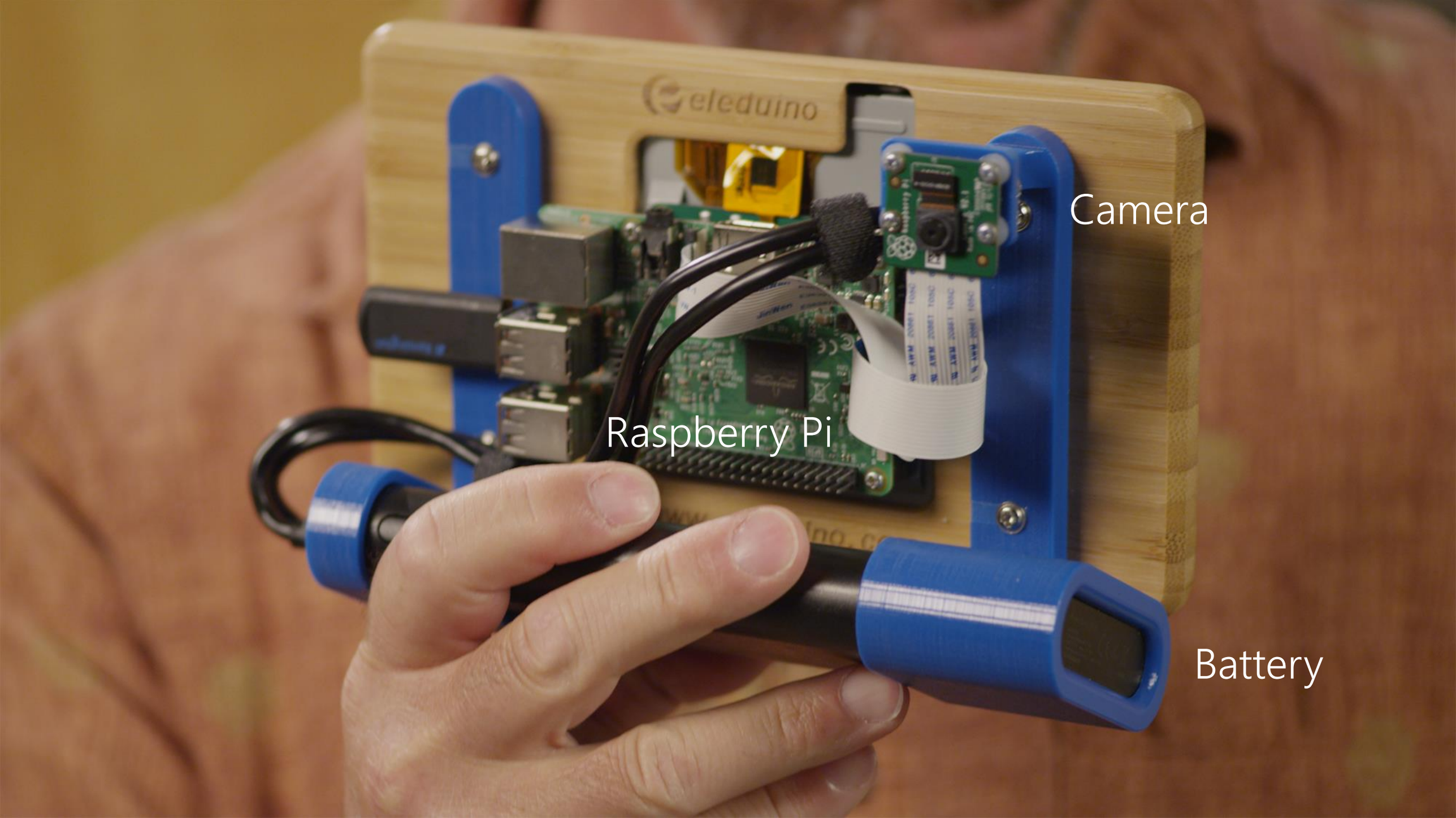
Less data with better features

ImageNet 1000, 1M photos

Cut off top layer



M. Gabel, R. Caruana, M. Philipose, O. Dekel


Camera
Raspberry Pi
Battery

Simulated Environments

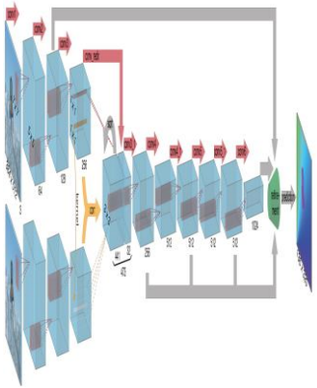
Trillions of sessions in complex scenarios

Learn & evaluate core competencies

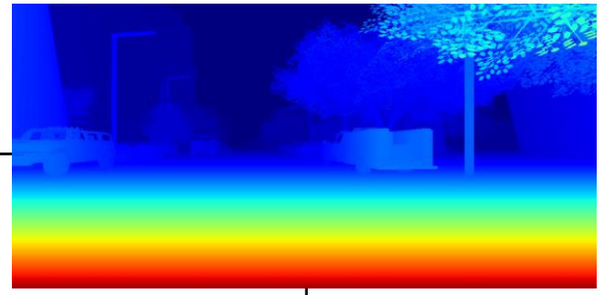
Learn to optimize action plans

Leveraging rich simulations

CNN

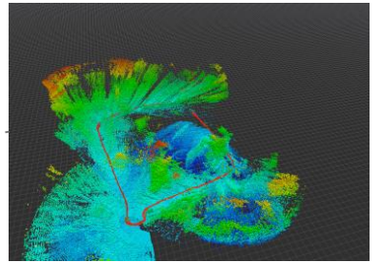


Depth Image

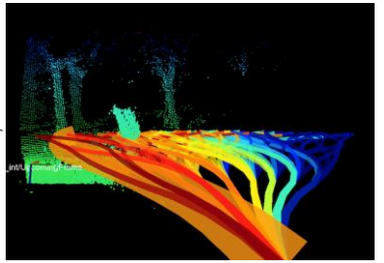


Mapping

Map



Planning



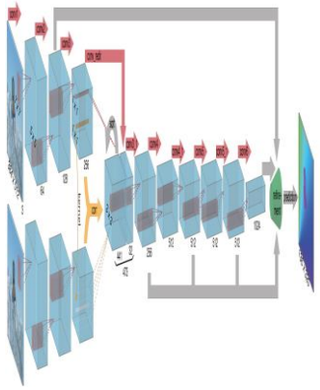
Next actions

Plans

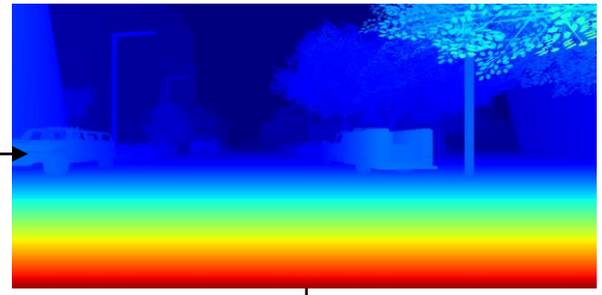
D. Dey, S. Sinha, S. Shah, A. Kapoor

Leveraging rich simulations

CNN

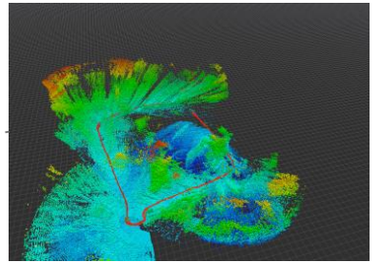


Depth Image

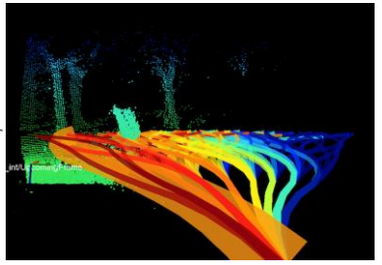


Mapping

Map



Planning



Next actions

Plans

D. Dey, S. Sinha, S. Shah, A. Kapoor

Learn Generative Models

Learn expressive generative models

Generalize from minimal training sets

Harness physics

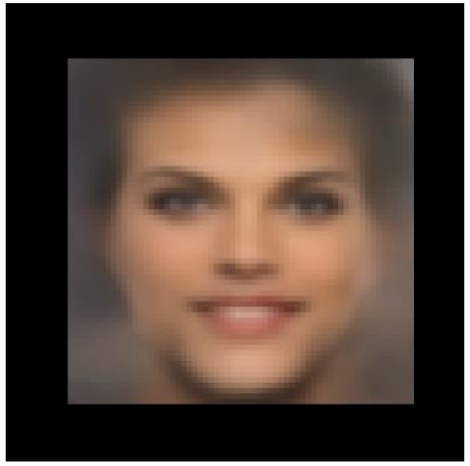
Learning generative models

Multilevel variational autoencoder

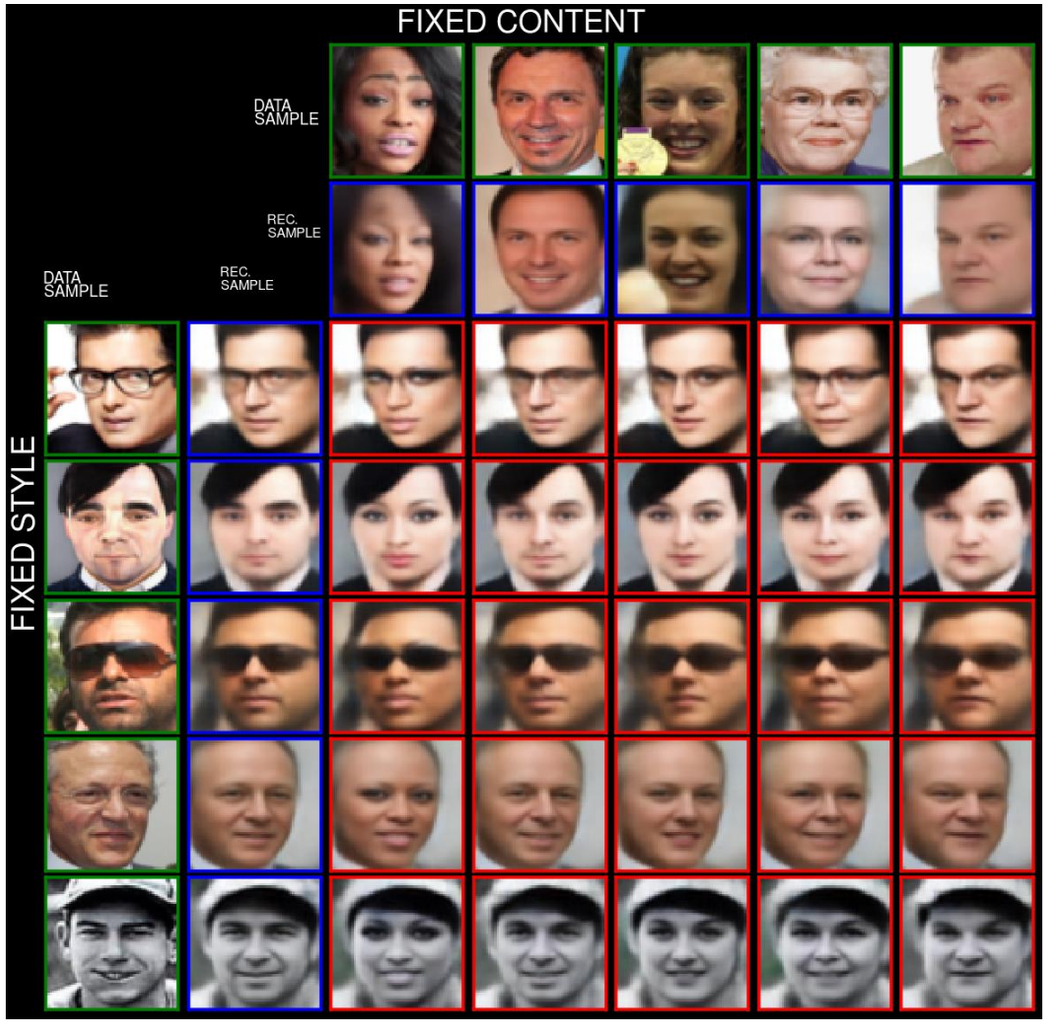
Learn disentangled representations

Groups of observations → latent models

Vary style



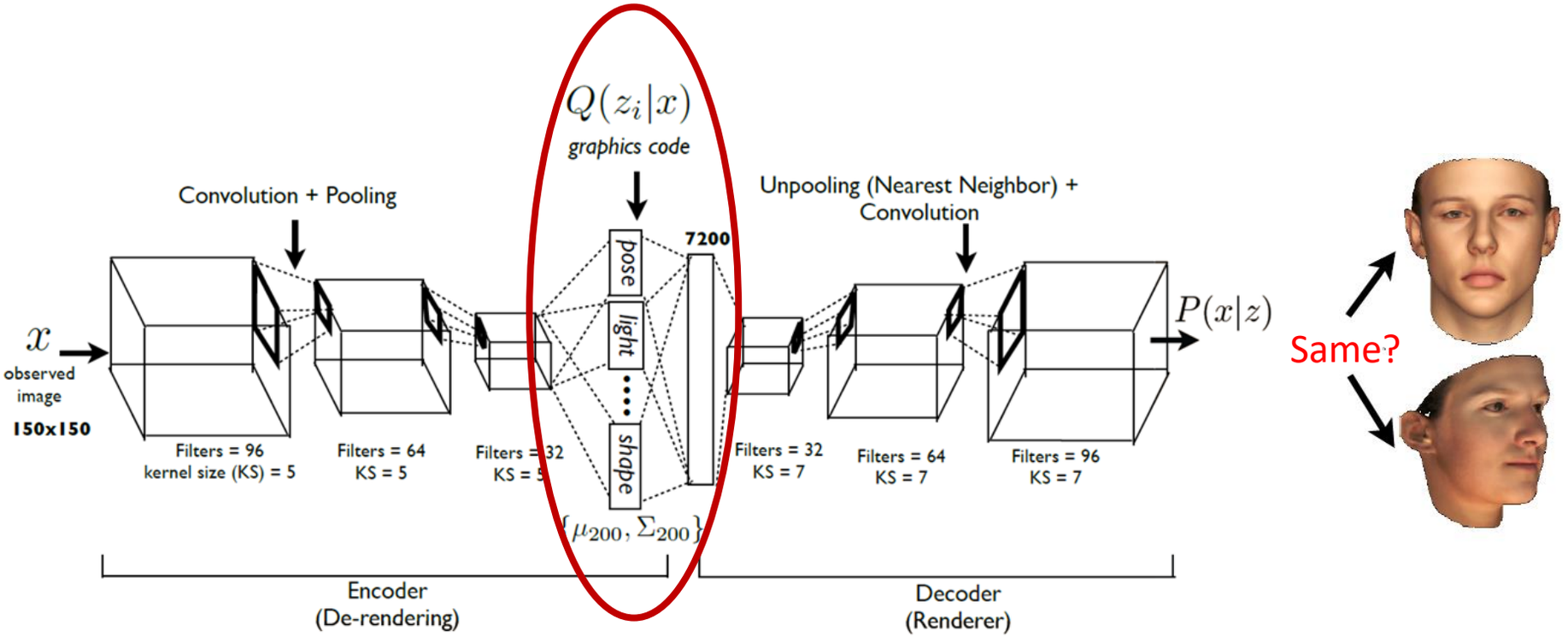
Vary ID



Smooth control over learned latent space

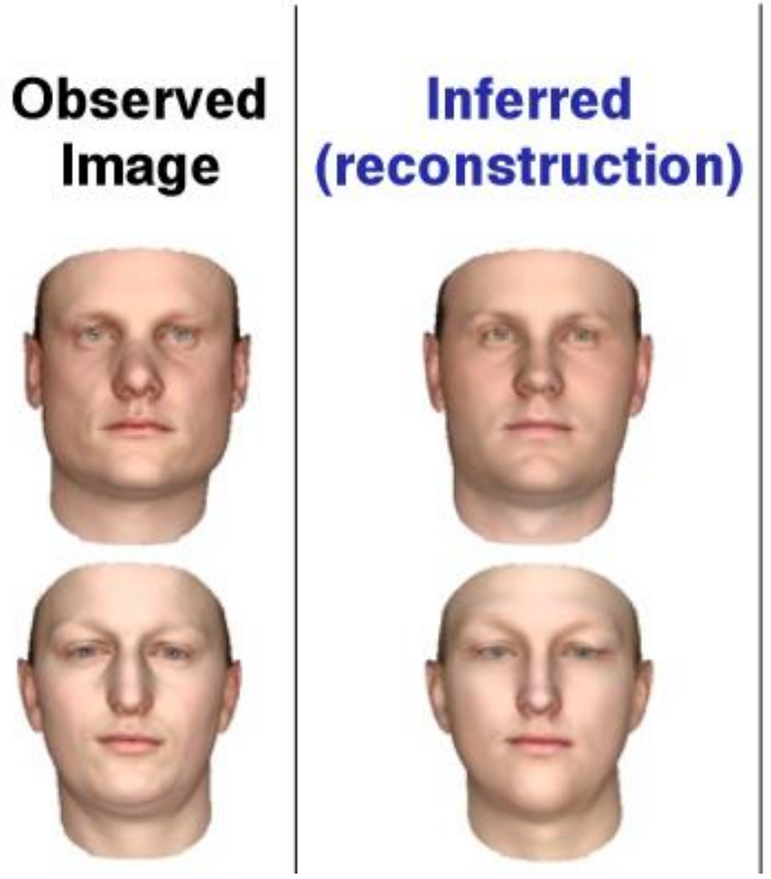
D. Buchacourt, R. Tomioka, S. Nowozin, 2017

Inject physics to disentangle & generalize



Kulkarni, Whitney, Kohli & Tenenbaum, 2015

Inject physics to disentangle & generalize



Kulkarni, Whitney, Kohli & Tenenbaum, 2015

Inject physics to disentangle & generalize

Illumination

Nod

Shake

Kulkarni, Whitney, Kohli & Tenenbaum, 2015

Adverserial Attacks & AI

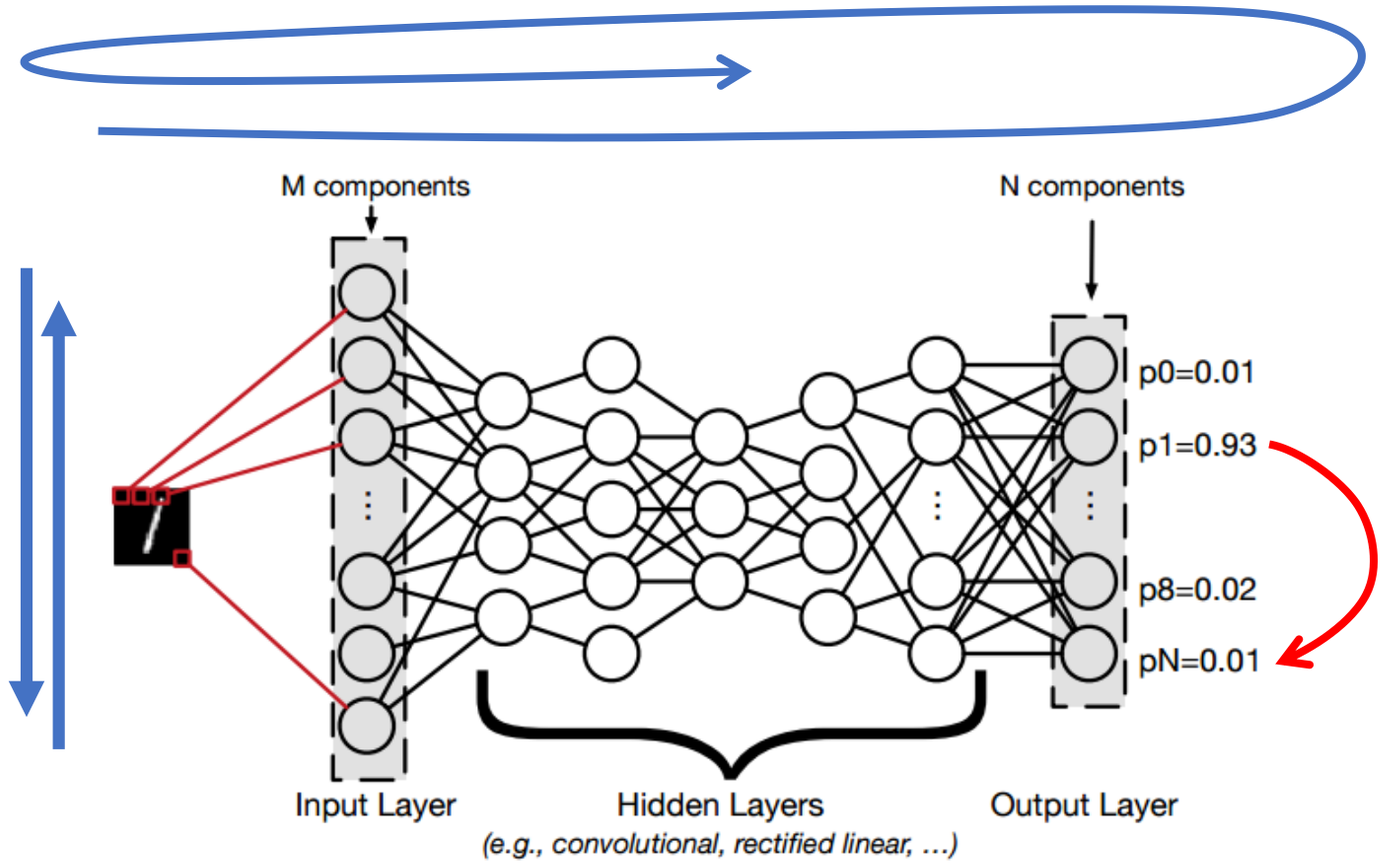
Concerns

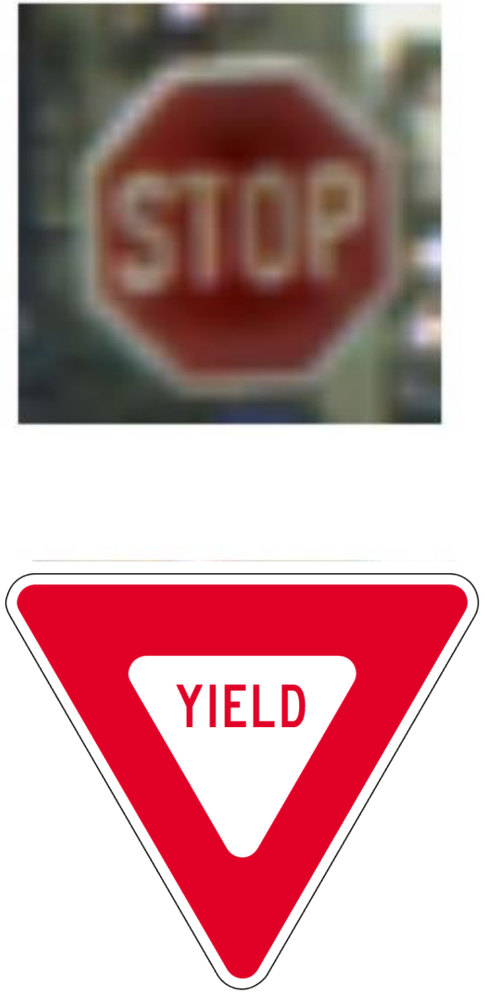
AI attack surfaces

Adversarial machine learning

Self-modification

Attacks on AI Systems



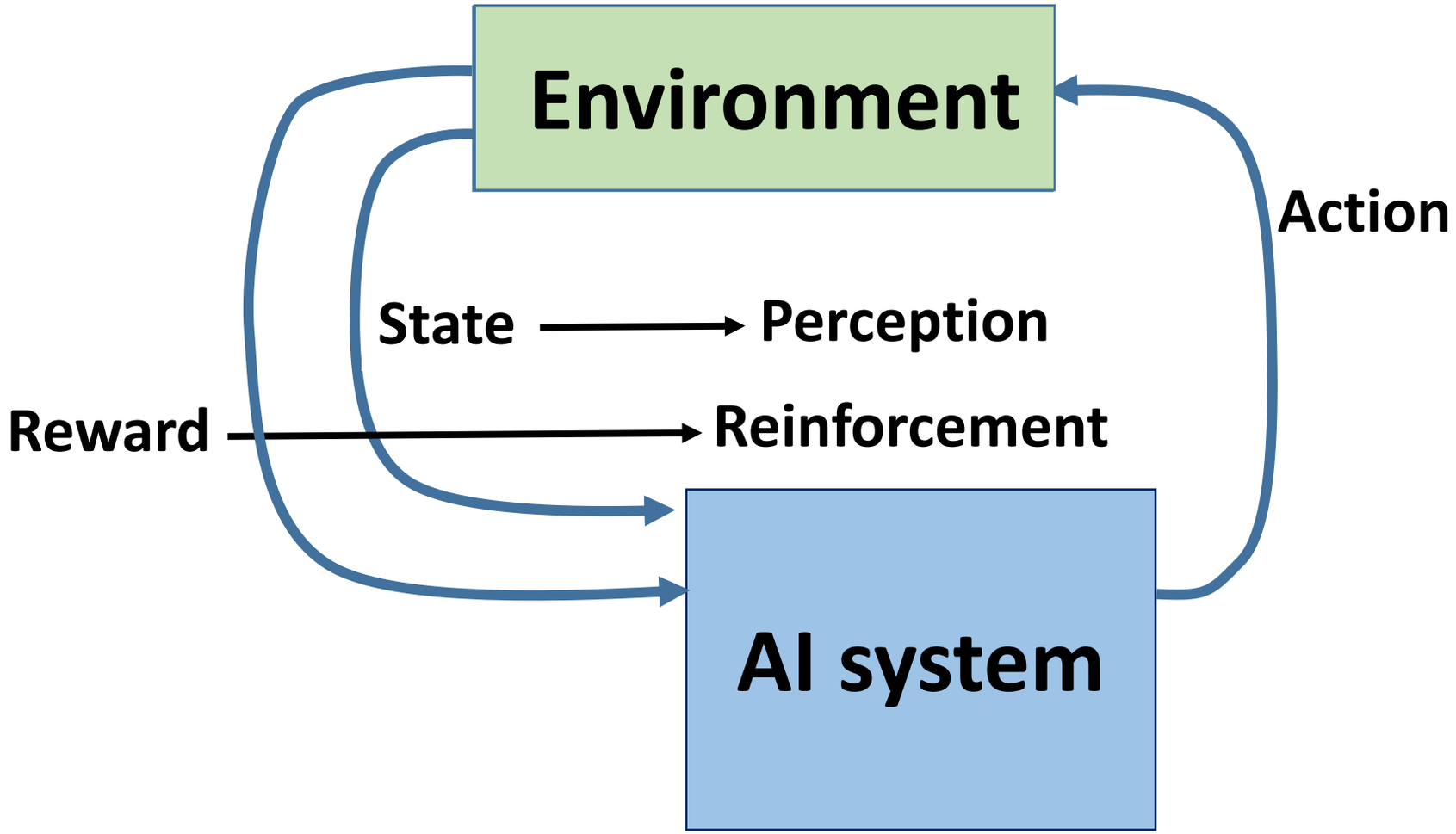


"Adverserial machine learning"

Goodfellow, et al.

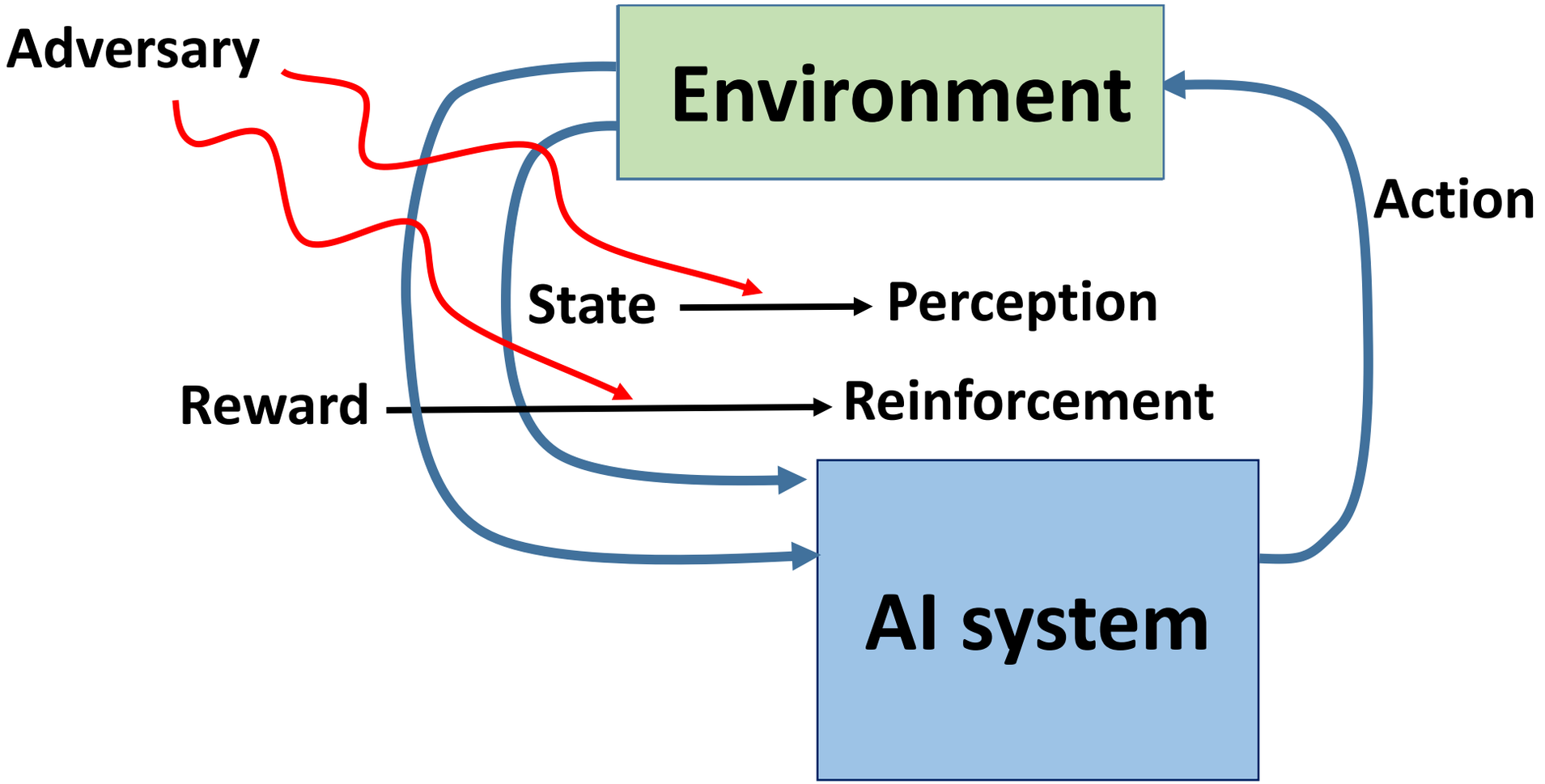
Papernot, et al.

Adversarial Attacks & Self-Modification



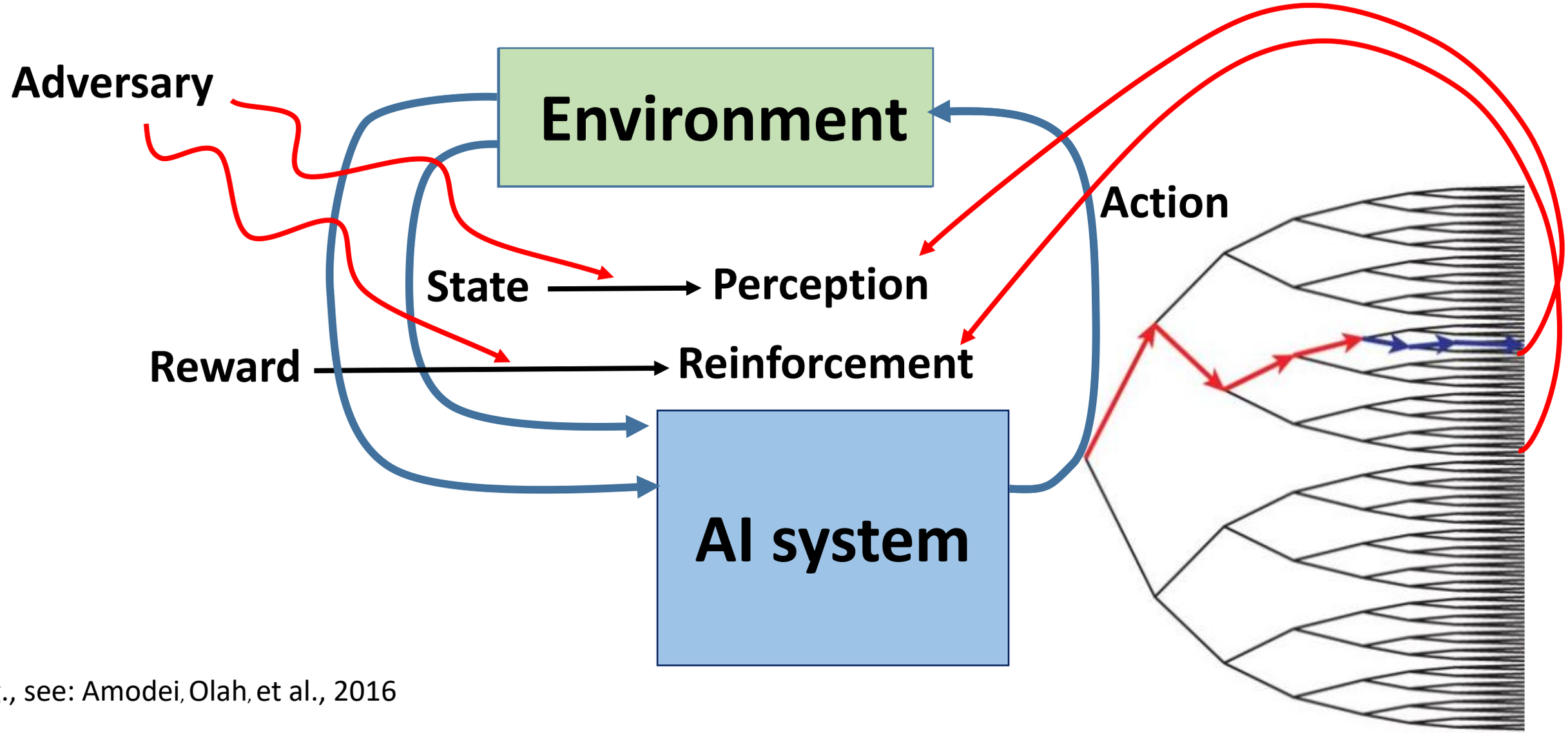
e.g., see: Amodei, Olah, et al., 2016

Adversarial Attacks & Self-Modification



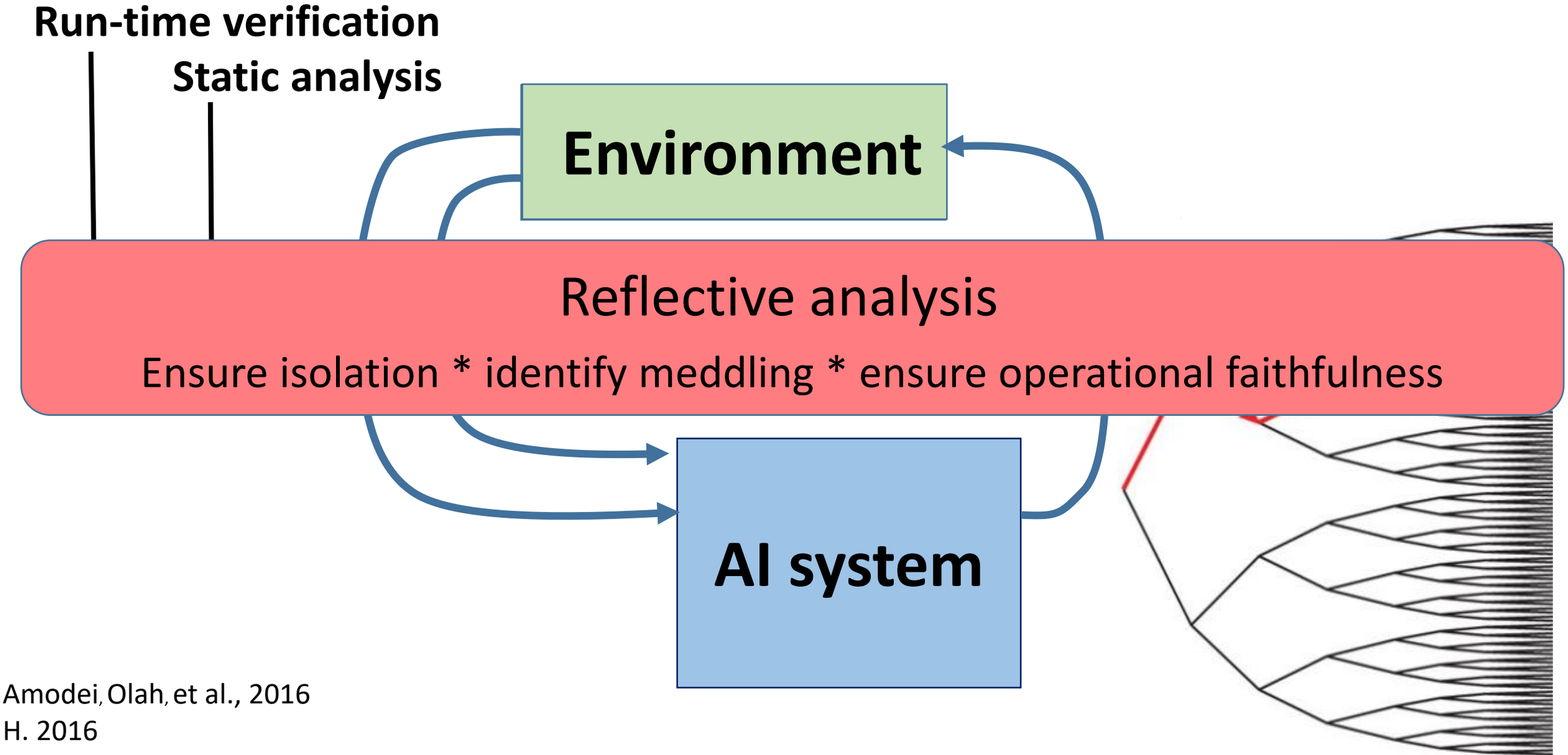
e.g., see: Amodei, Olah, et al., 2016

Adversarial Attacks & Self-Modification



e.g., see: Amodei, Olah, et al., 2016

Adversarial Attacks & Self-Modification



Amodei, Olah, et al., 2016

H. 2016

AI & People

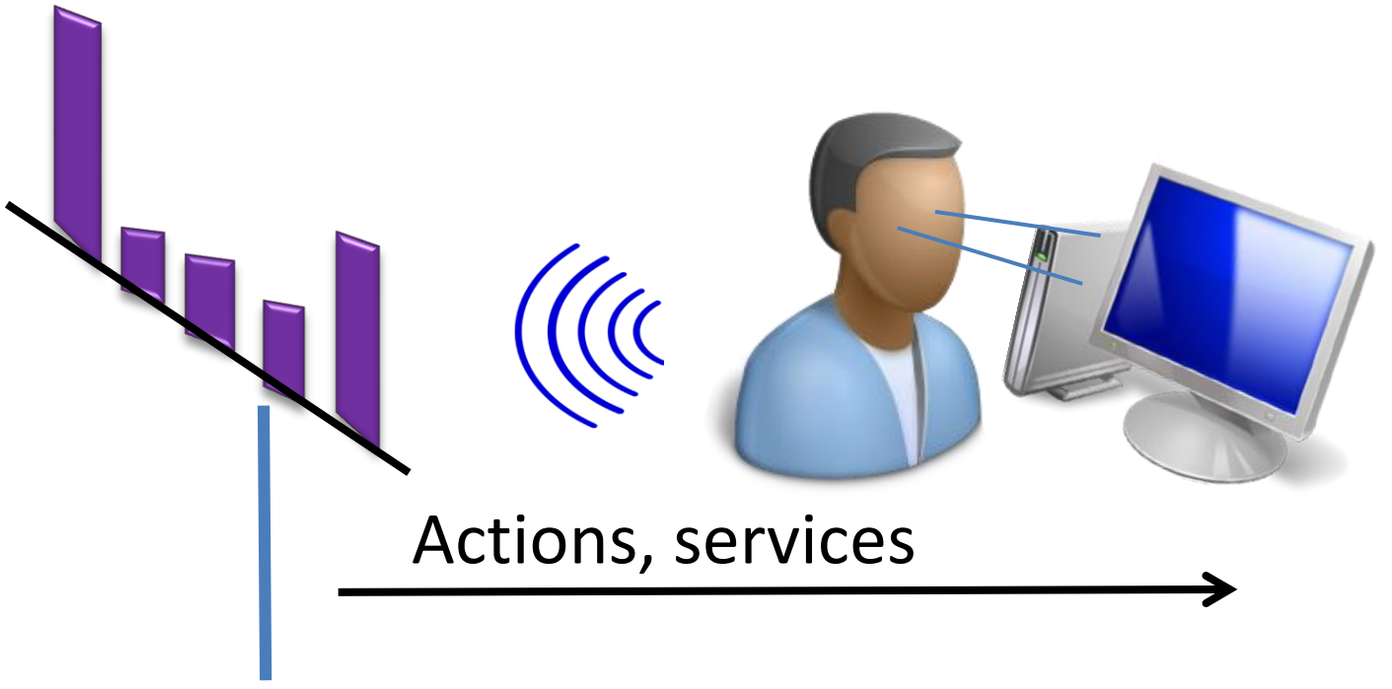
Directions

Models of people & tasks

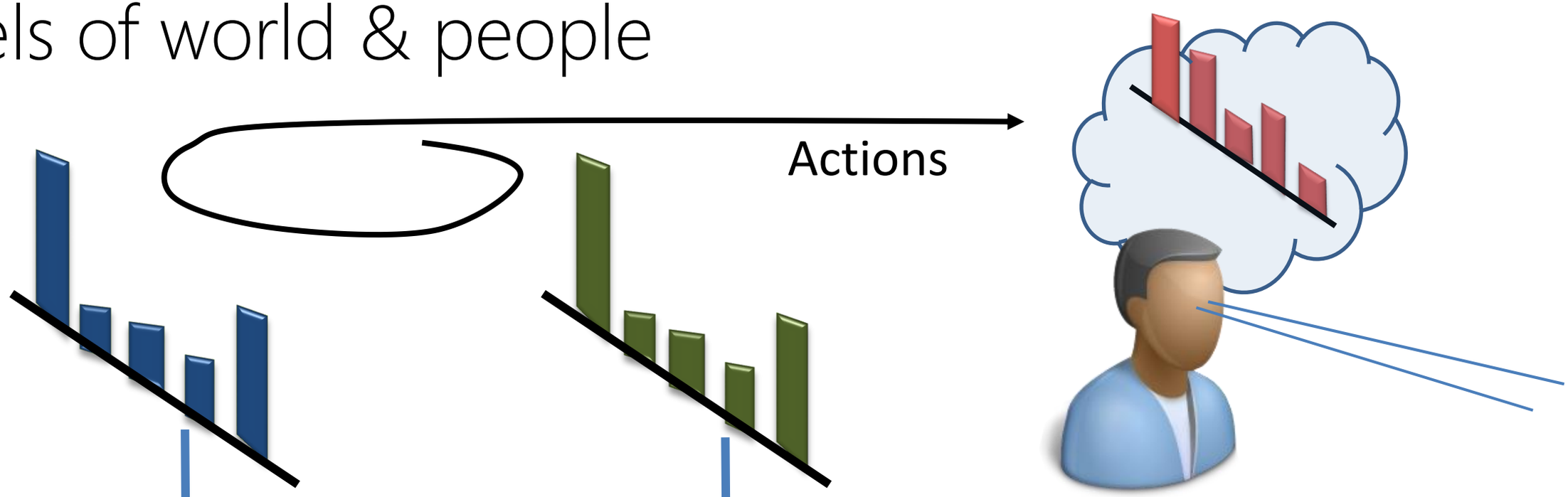
Models of complementarity

Coordination of initiative

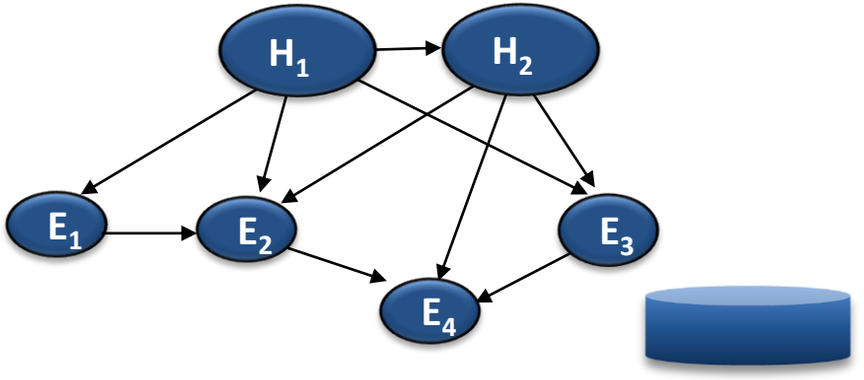
Models of people & tasks



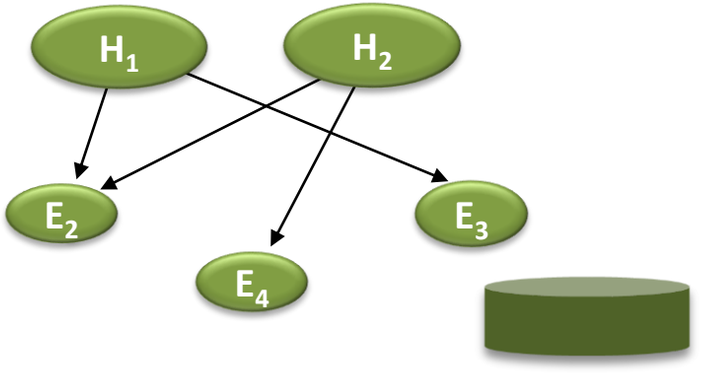
Models of world & people



Predictions about world



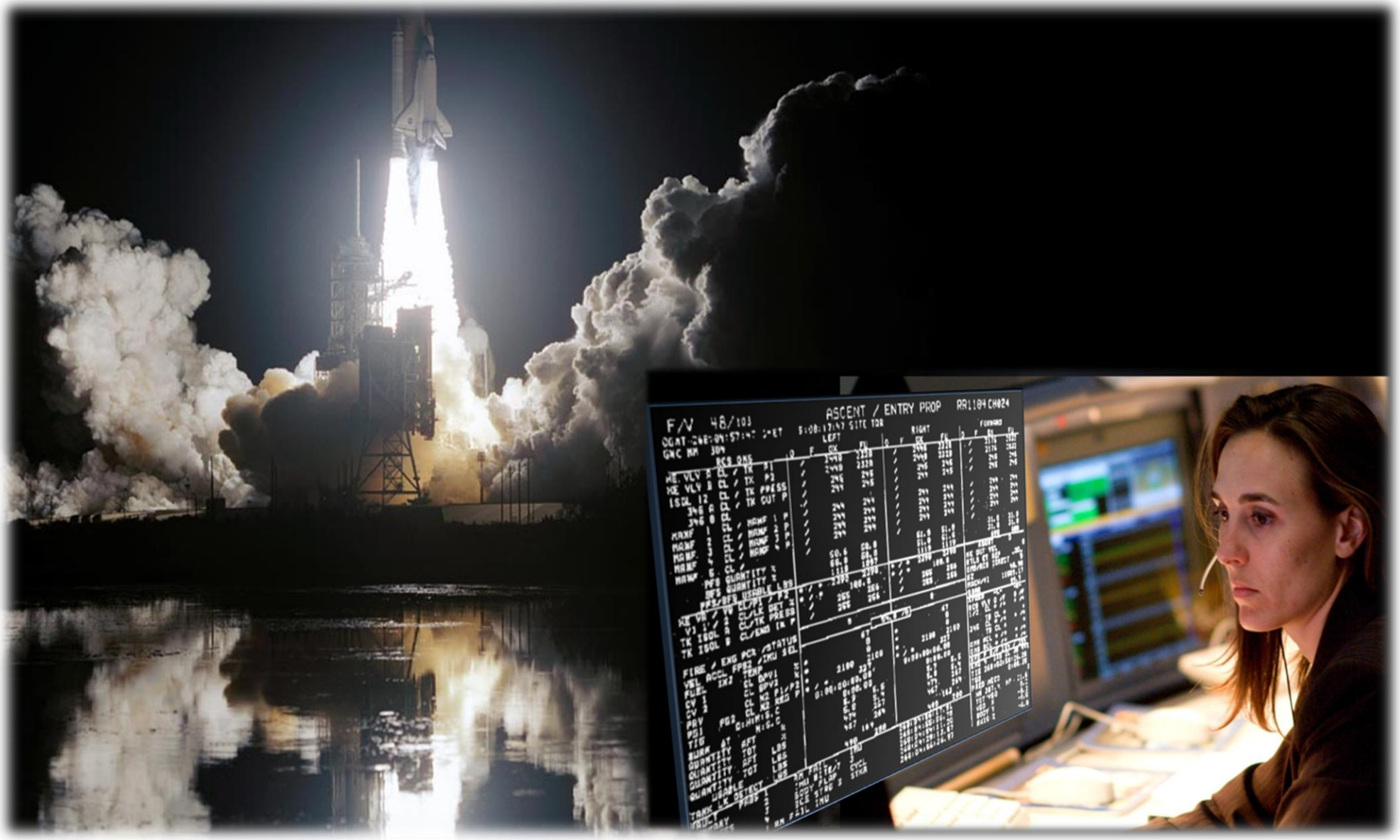
Predictions about user beliefs



H. Barry, 1995

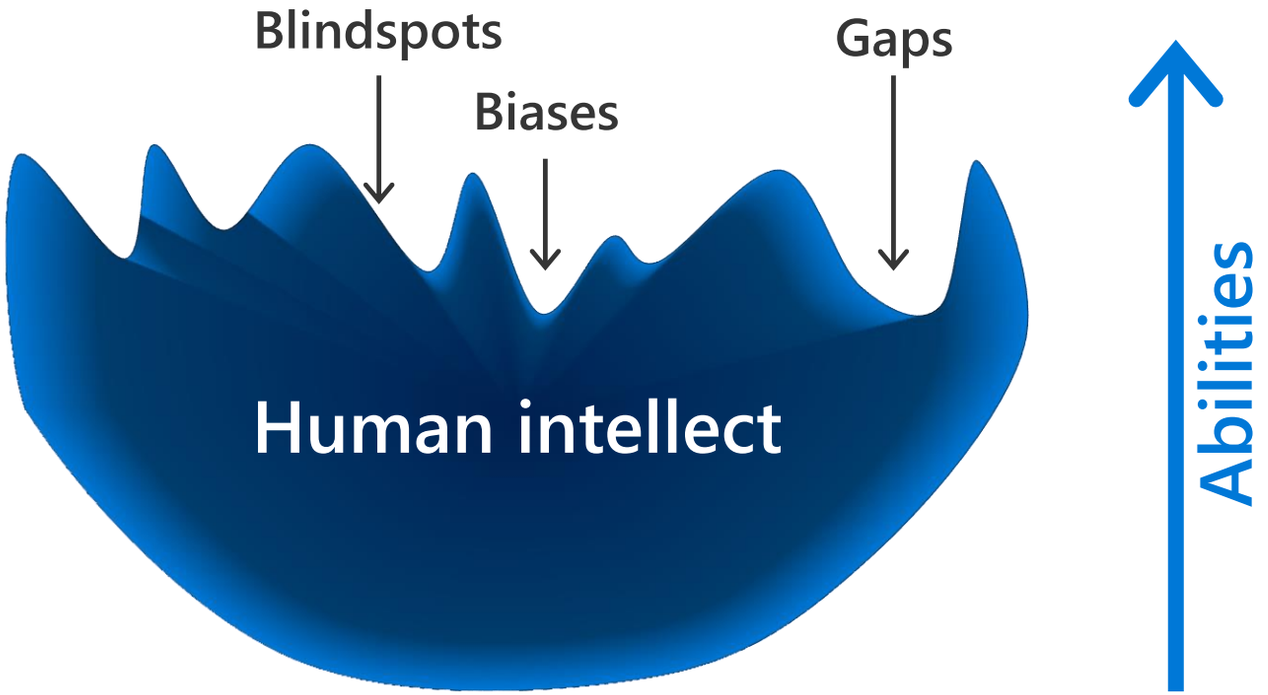
H. , Apacible, Sarin, Liao, 2005

Models of world & people



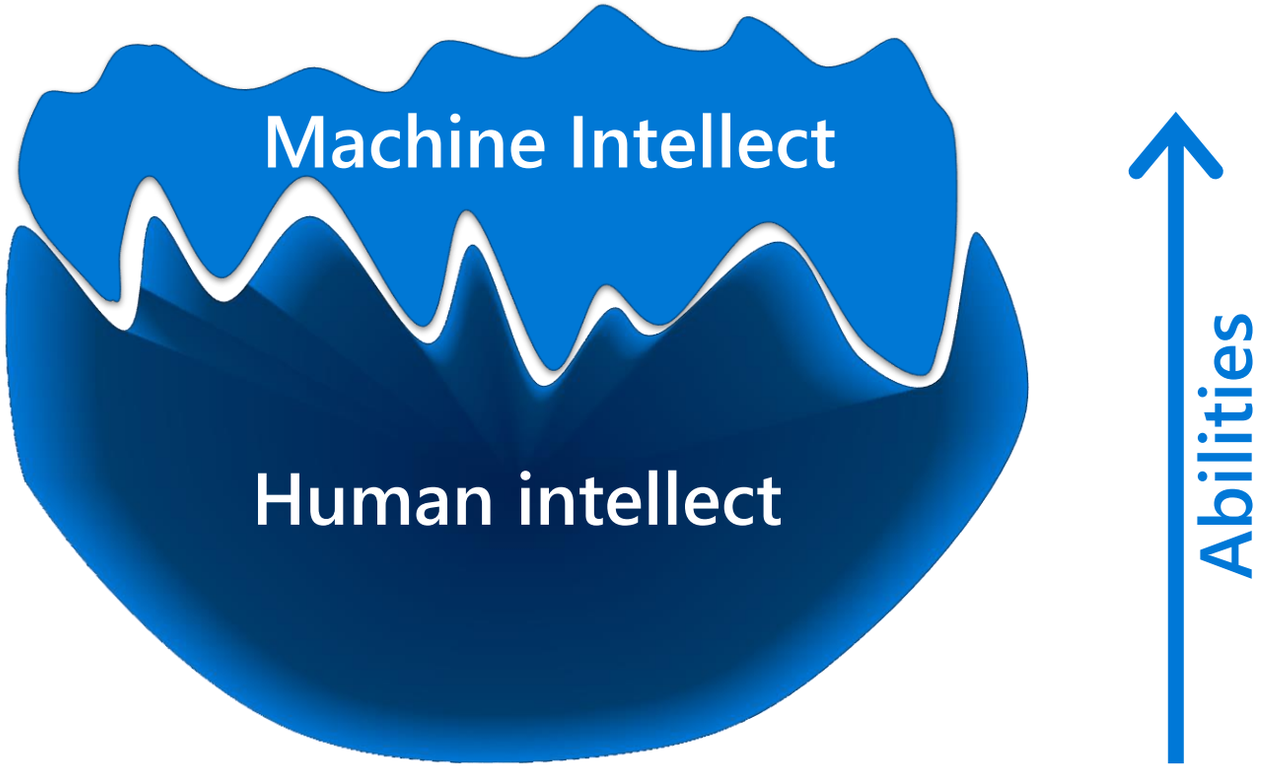
H. Barry, 1995

Complementarity



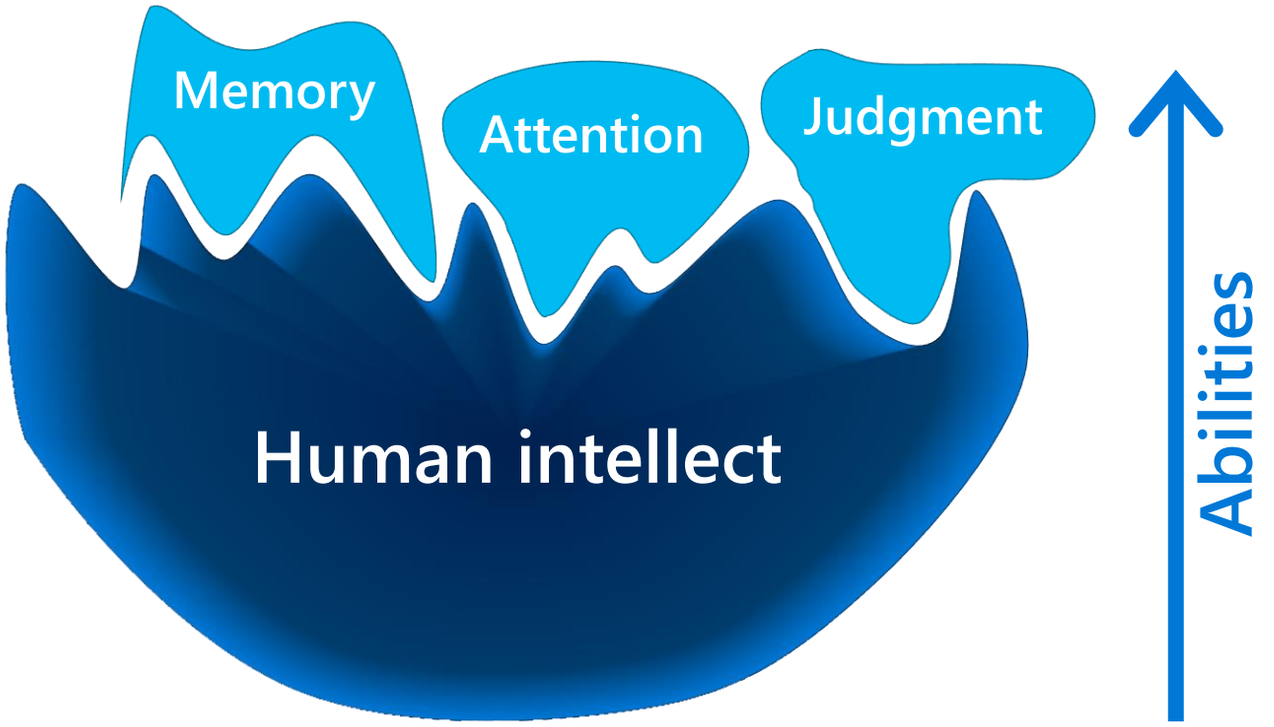
Leverage and extend results from cognitive psychology

Complementarity



Leverage and extend results from cognitive psychology

Complementarity

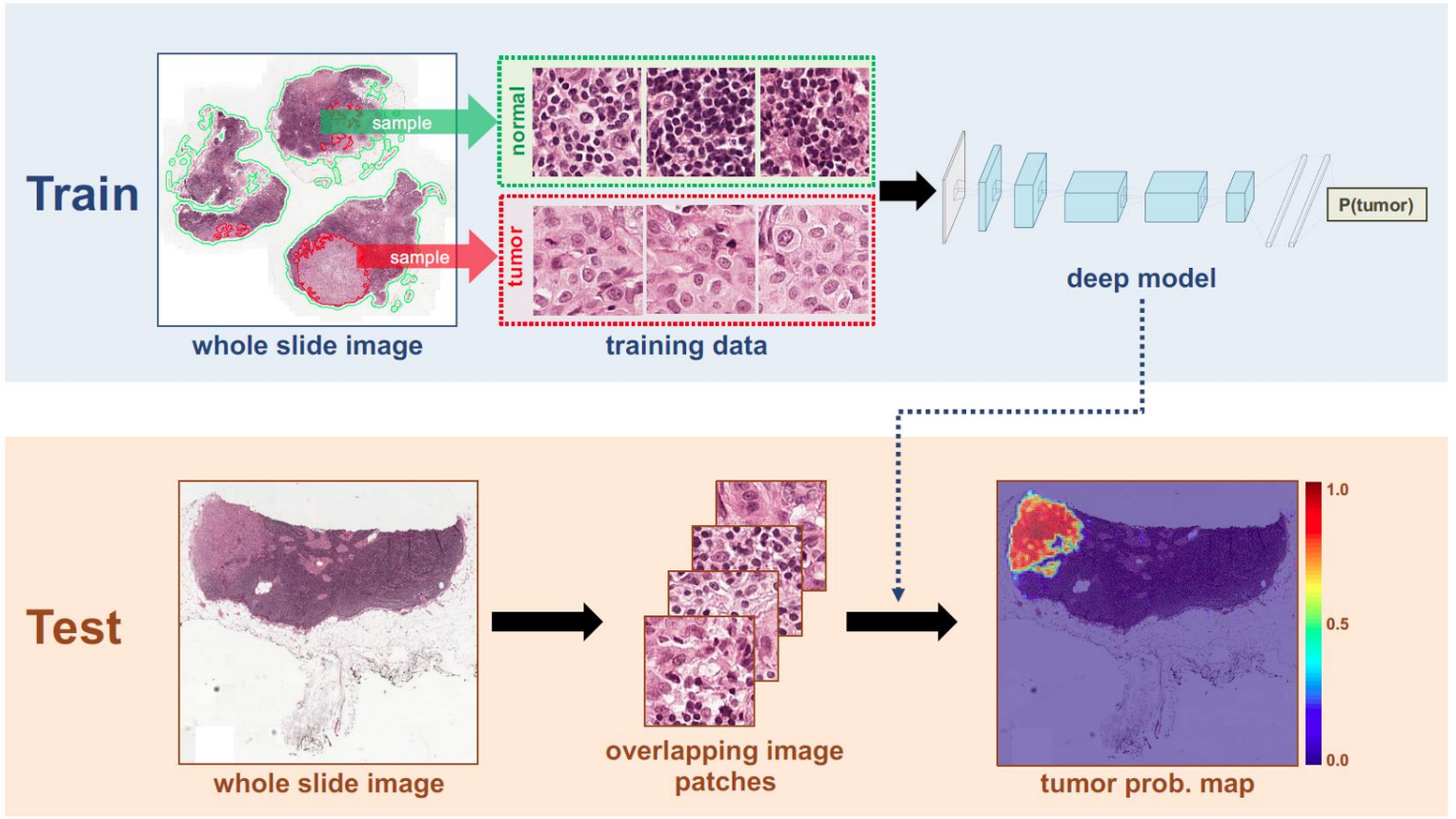


Leverage and extend results from cognitive psychology

Complementarity

Identifying metastatic breast cancer

(Camelyon Grand Challenge 2016)



Human is superior

Error: 3.4%

↓

AI + Expert: 0.5%

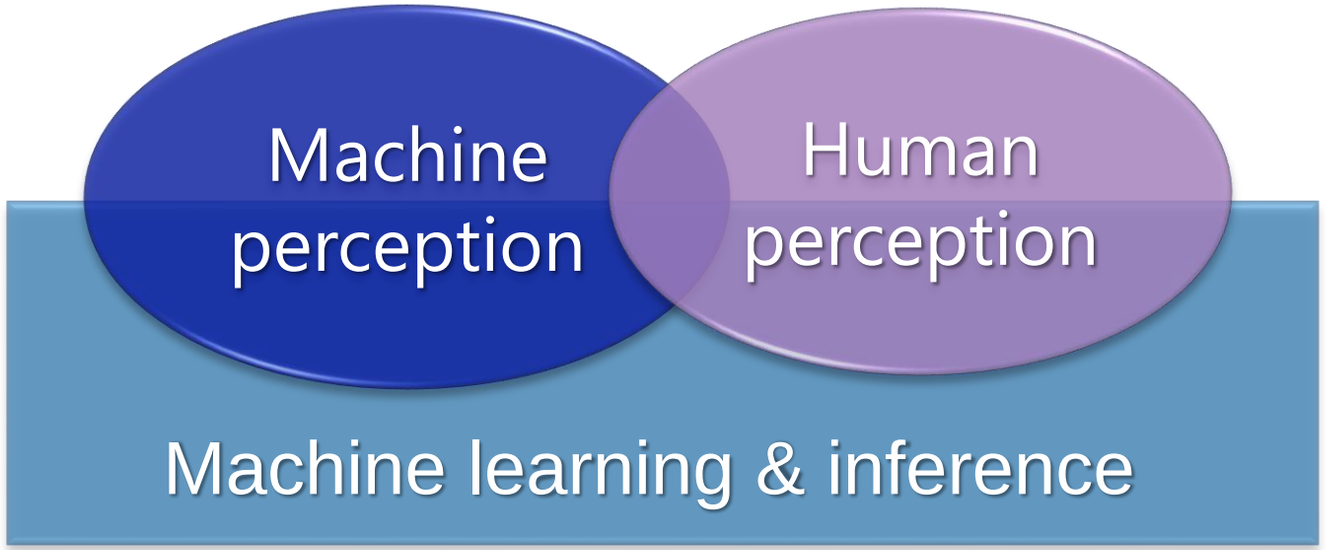
85% reduction in errors.

D. Wang, A. Khosla, R. Gargeya, H. Irshad, A.H. Beck, 2016

Complementarity

Label galaxies in Sloan Digital Sky Survey

(Galaxy Zoo)

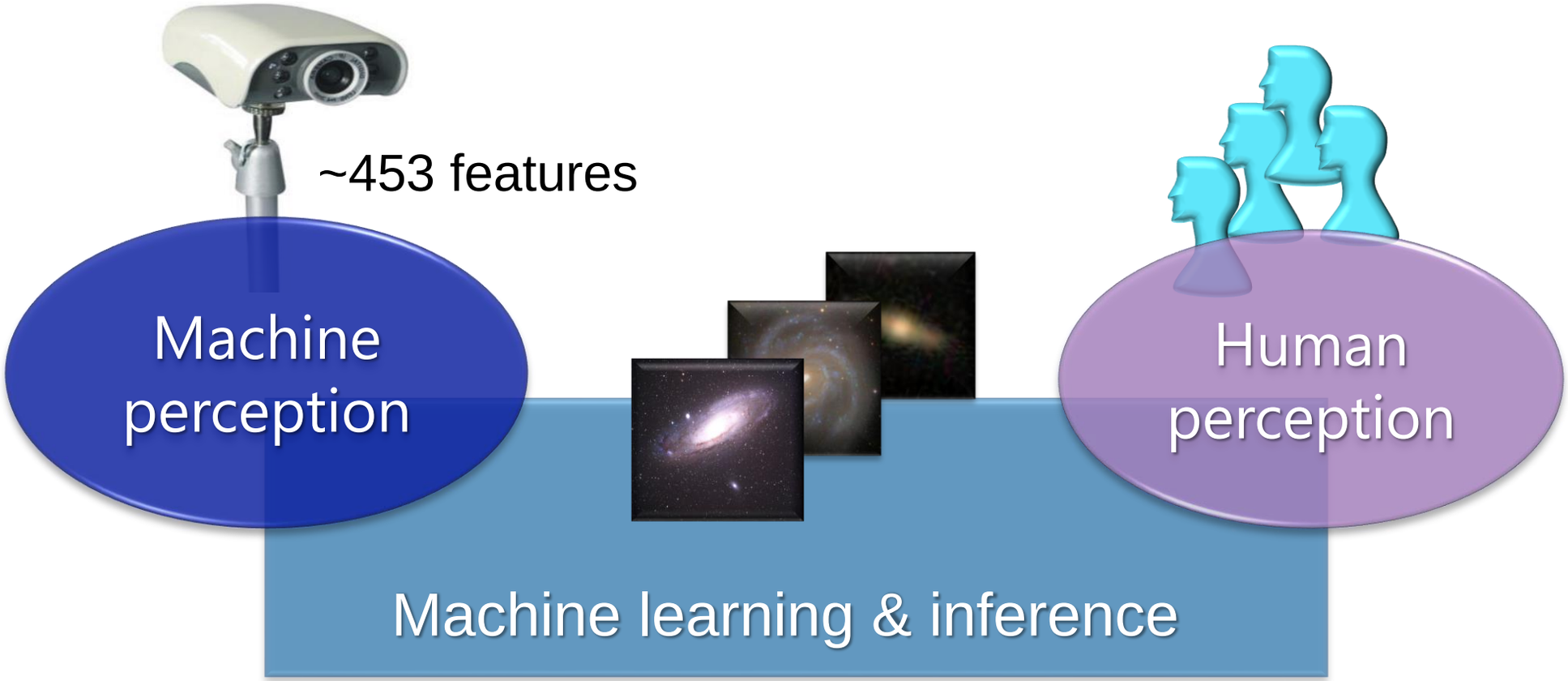


Kamar, Hacker, H., AAMAS 2012

Complementarity

Label galaxies in Sloan Digital Sky Survey

(Galaxy Zoo)

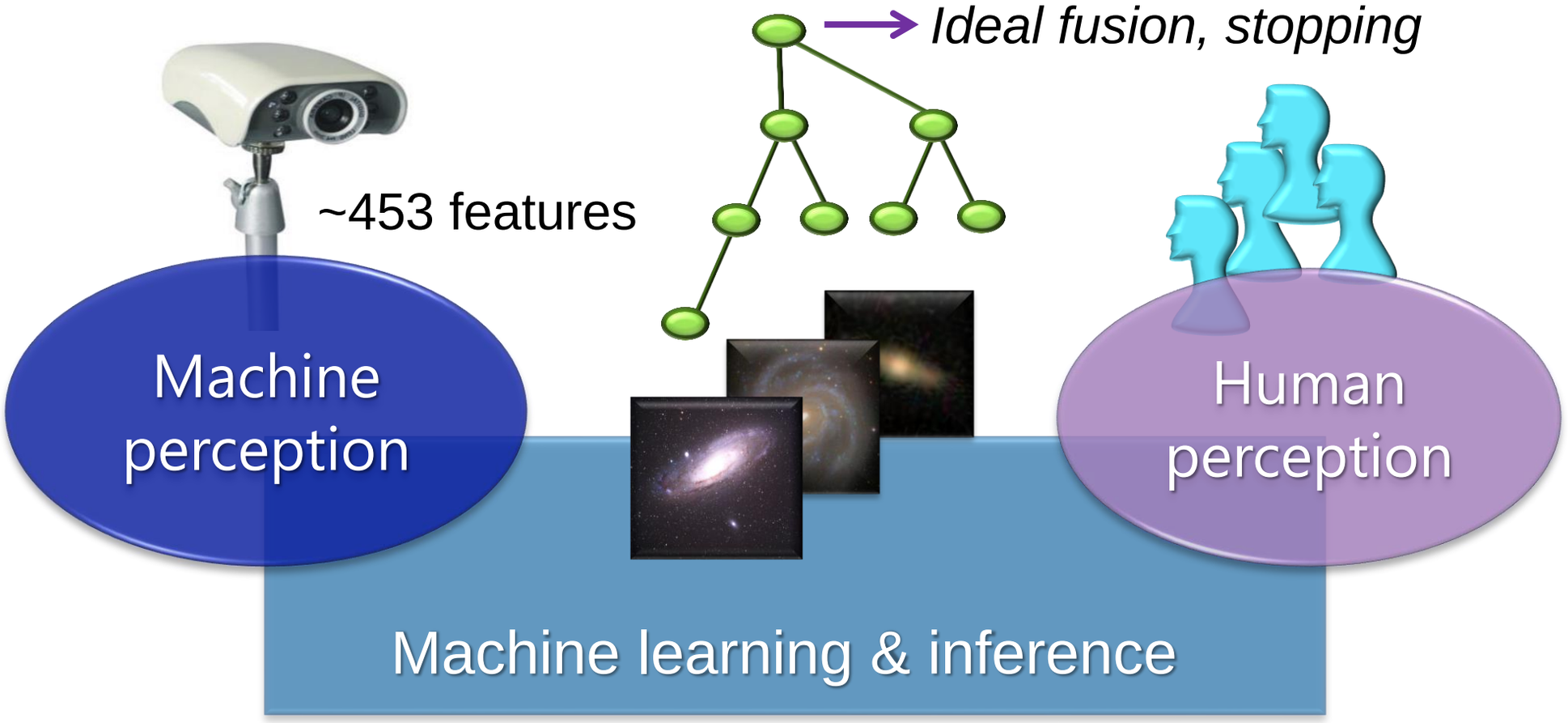


Kamar, Hacker, H., AAMAS 2012

Complementarity

Full accuracy: 47% of human effort

95% accuracy: 23% of human effort



Kamar, Hacker, H., AAMAS 2012

Designs for mix of initiatives

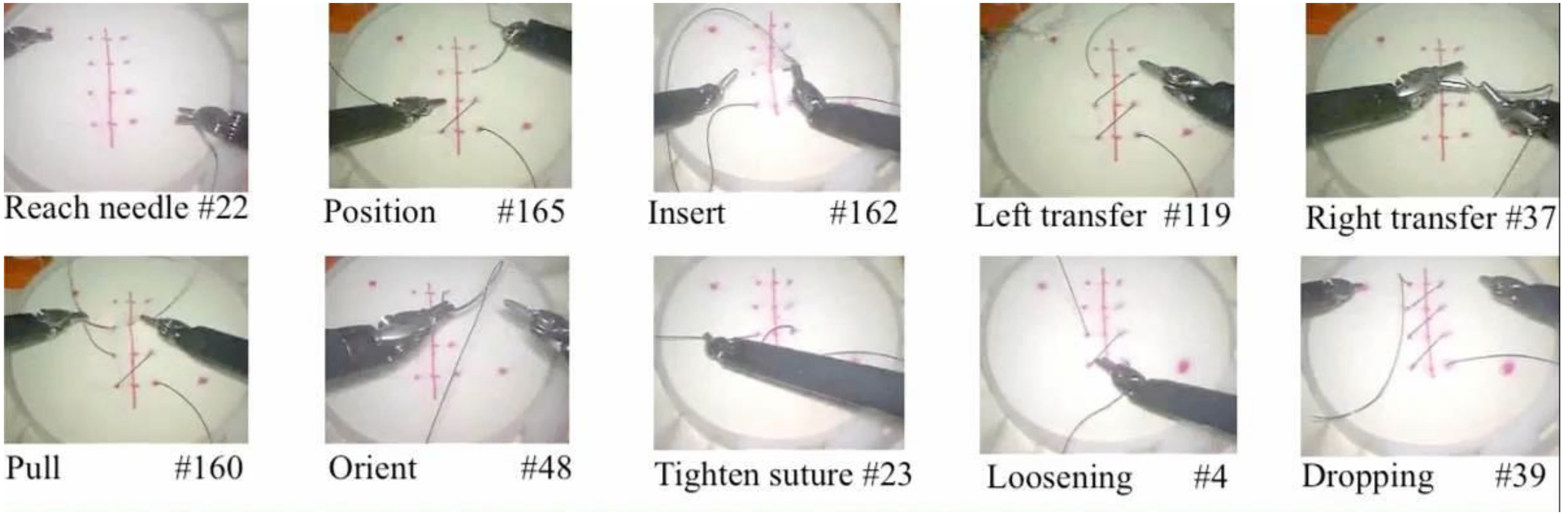
Human cognition

Machine intelligence

Machine learning & inference

Design, learning, optimization

Recognizing intention



C.E. Reiley, et al.

Coordination of initiative

Padoy & Hager. *ICRA 2011*
van den Berg, et al, *ICRA*, 2010

AI, people, and society

Multiple influences and concerns

- Trustworthiness and safety
- Fairness, accuracy, transparency
- Ethical and legal aspects of autonomy
- Jobs and economy

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Bernard Parker: rated high risk

Dylan Fugett: rated low risk.

Image URL

Detection Result:
5 faces detected

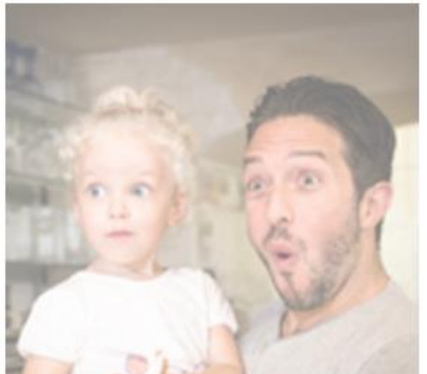
JSON:

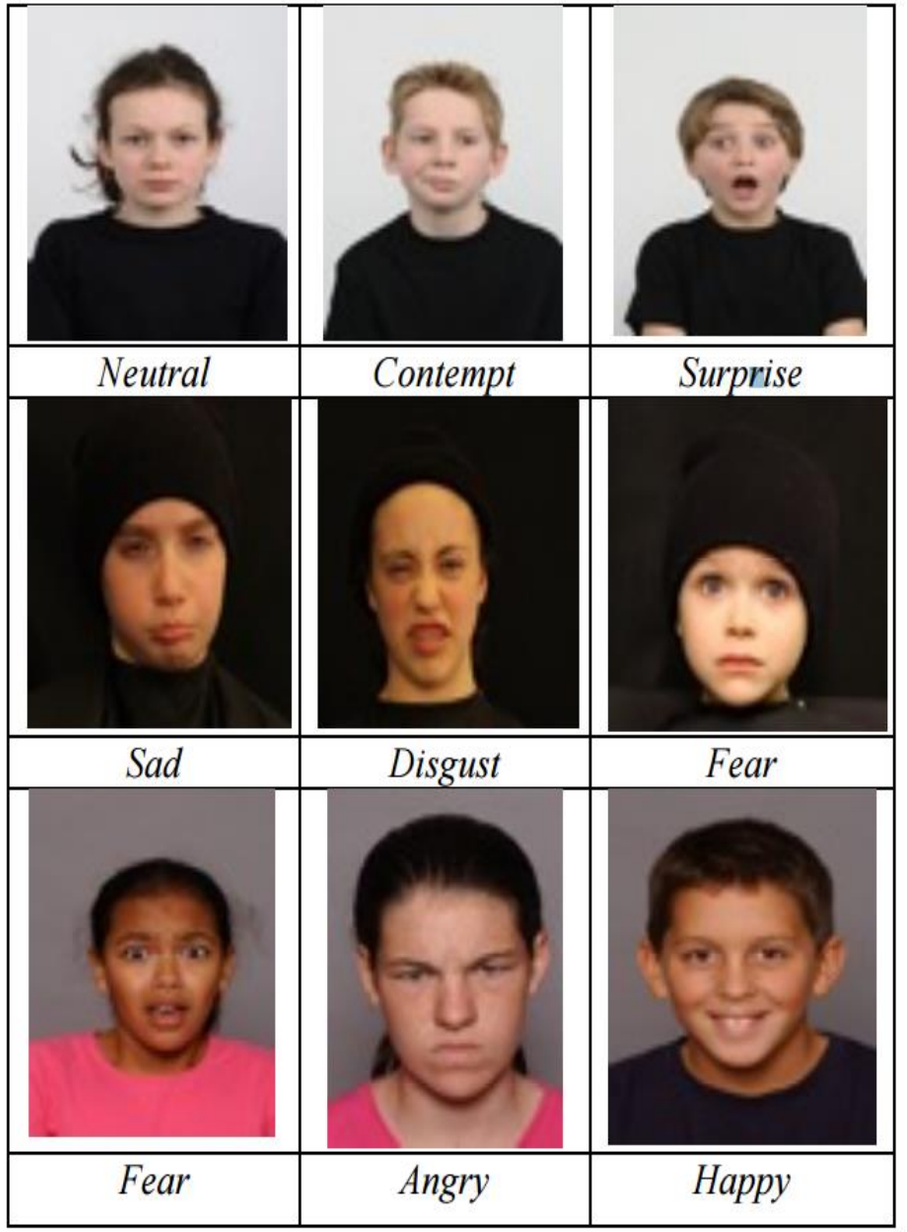

```
[
  {
    "faceRectangle": {
      "left": 488,
      "top": 263,
      "width": 148,
      "height": 148
    },
    "scores": {
      "anger": 9.075572e-13,
      "contempt": 7.048959e-9,
      "disgust": 1.02152783e-11,
      "fear": 1.778957e-14,
      "happiness": 0.9999999,
      "neutral": 1.31694478e-7,
      "sadness": 6.04054263e-12,
      "surprise": 3.92249462e-11
    }
  },
  {
    "faceRectangle": {
```











Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems

Ayanna Howard[1*], Cha Zhang[2], Eric Horvitz[2]

March 2017

Machine learning “contact lens” for children

A. Howard, C. Zhang, H., 2017

Corporate & community responsibility

Microsoft

Aether Advisory Panel

AI and Ethics in Engineering and Research

AI in the Open World

Science & engineering

Human-AI collaboration

AI, people, and society

Much to do