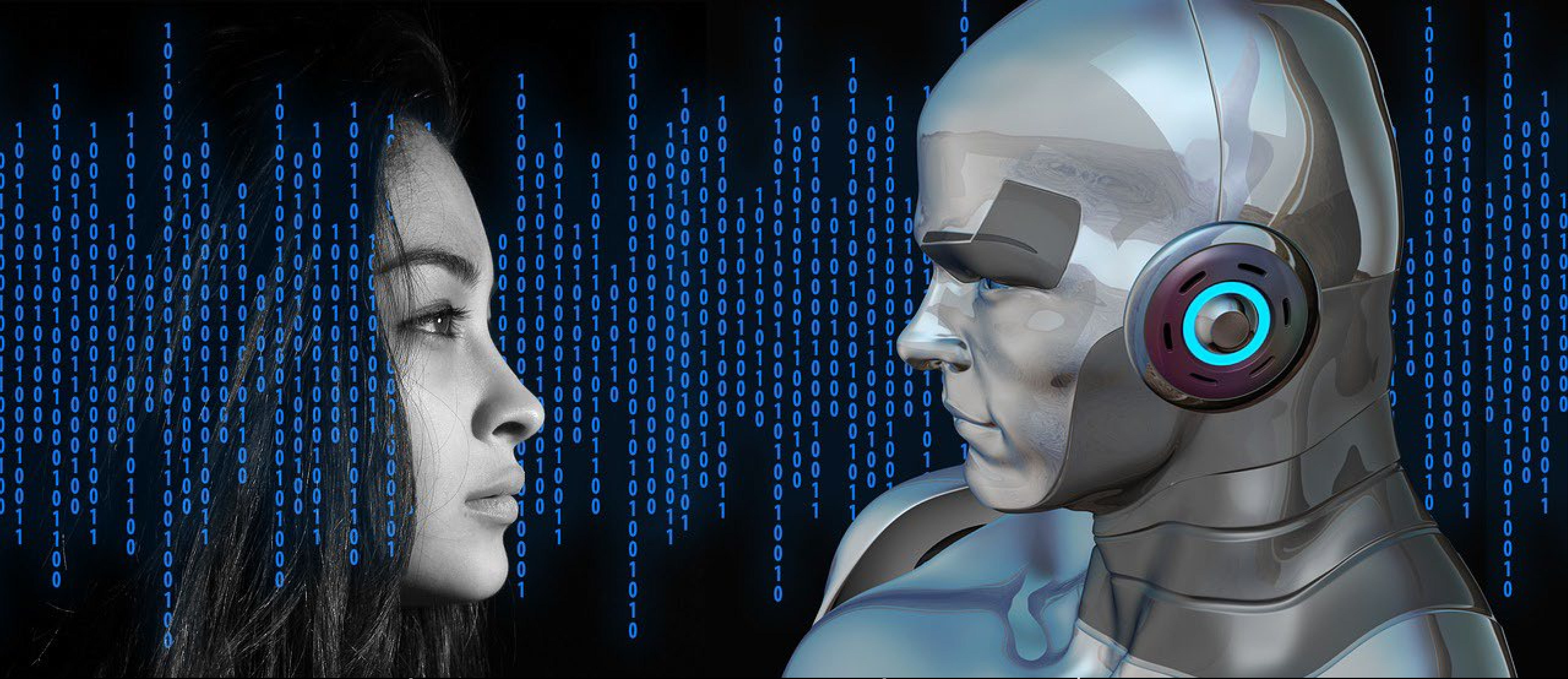


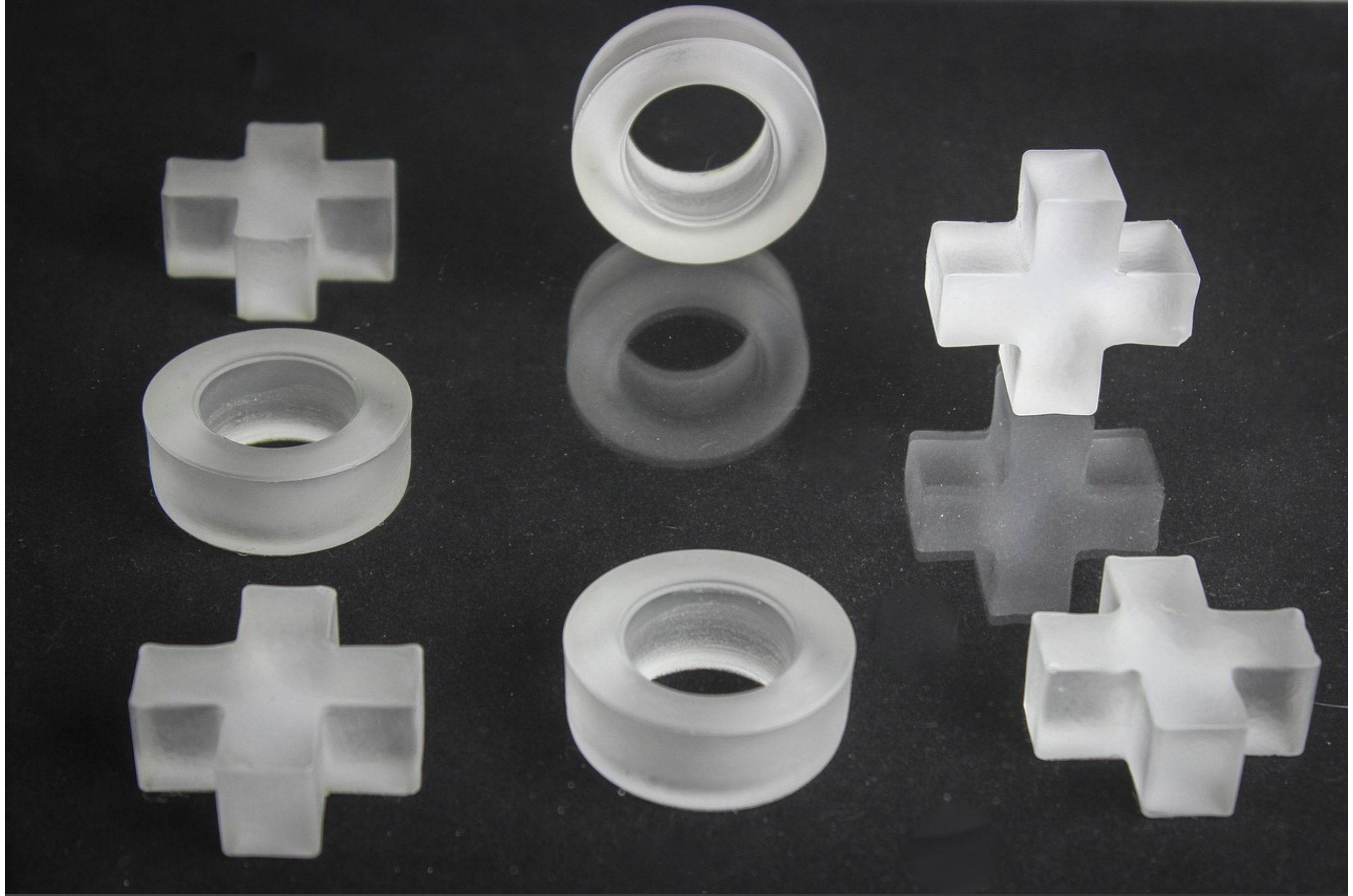
The AI Data Revolution
Doing More with Less Data Labeling

David C. Anastasiu
Assistant Professor
San Jose Staté University

(Starting Fall 2019,
Santa Clara University)

According to Miriam-Webster:

1. A branch of computer science dealing with the simulation of intelligent behavior in computers.
2. The capability of a machine to imitate intelligent human behavior.



What is AI?

Types of AI

Narrow

Pattern Recognition

Prediction

General

Can think like a human

Augmented

Can help a human think

Why Care About AI?

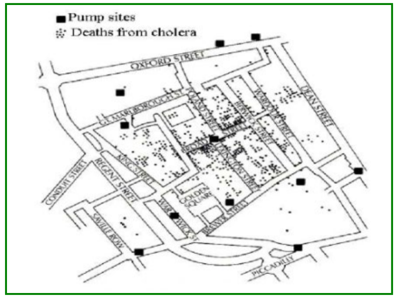
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies

- New mantra
  - Gather whatever data you can whenever and wherever possible.

- Machine learning is becoming ubiquitous in society
  - Here's a movie you might like (recommender systems)
  - You should try this product (advertising)
  - Found a shortcut that will save you 10 minutes (maps)
  - Jessica is a friend suggestion for you (social media)
  - Caution! Vehicle approaching in right lane (drive assist / self-driving cars)

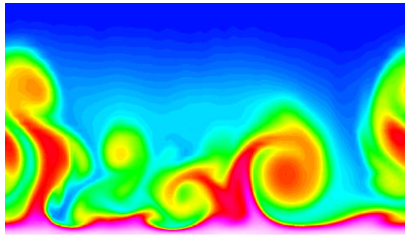
Homeland Security

Business Data



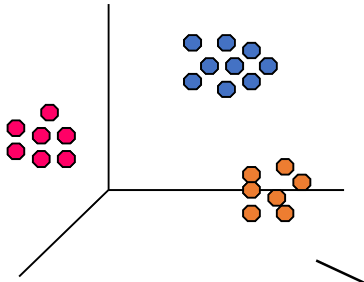
Geo-spatial data

Sensor Networks

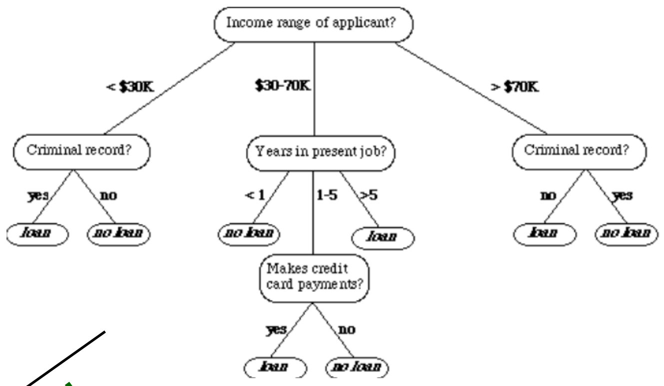


Computational Simulations

Machine Learning Tasks



Clustering



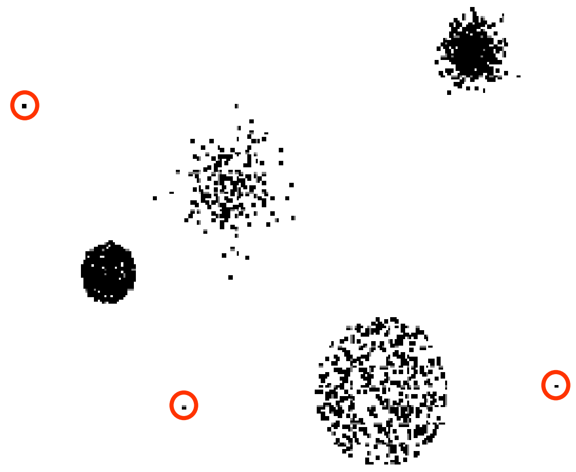
Predictive Modeling

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules

Anomaly Detection



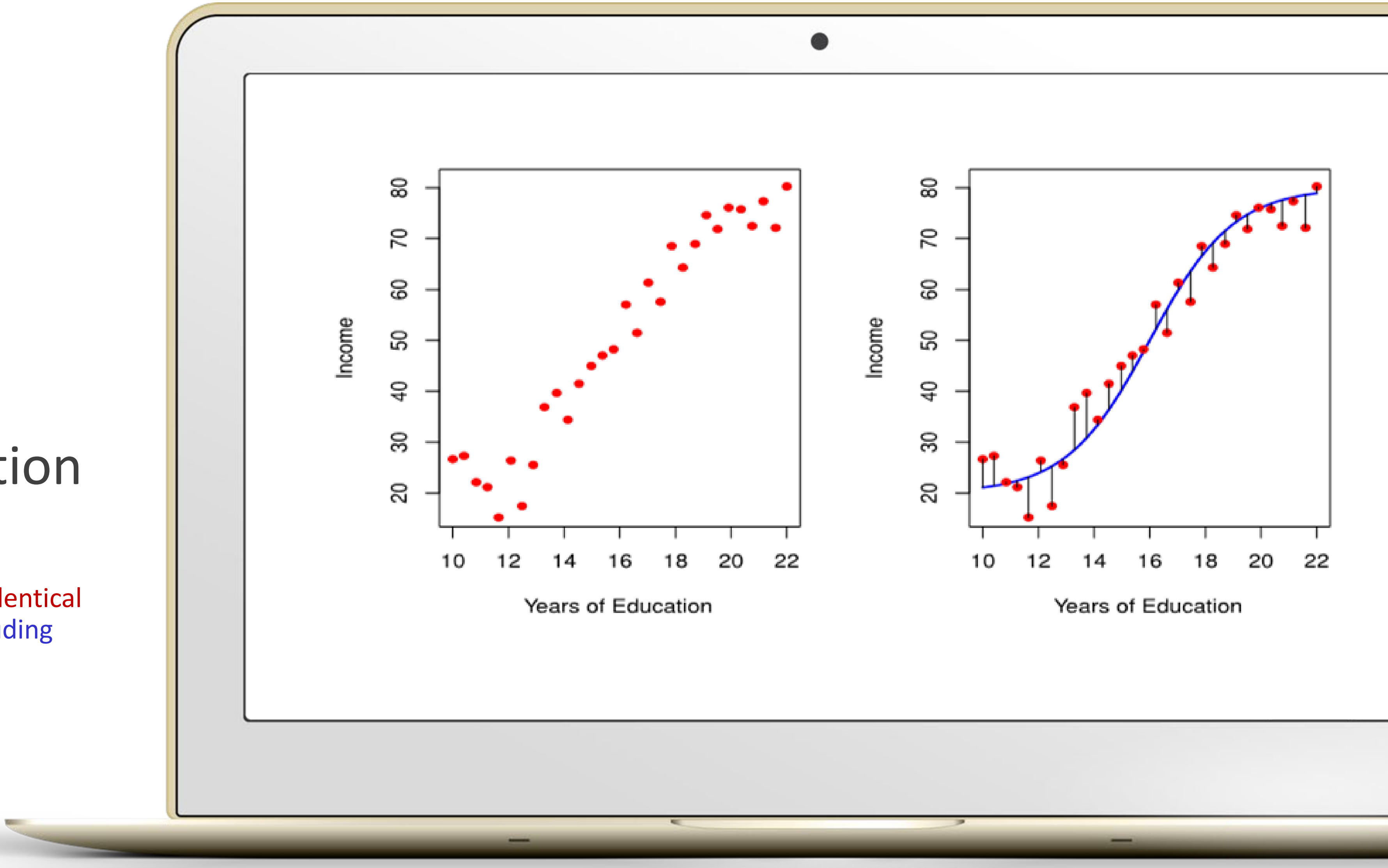
Supervised Learning

Learn a function

$$Y = f(X) + \epsilon$$

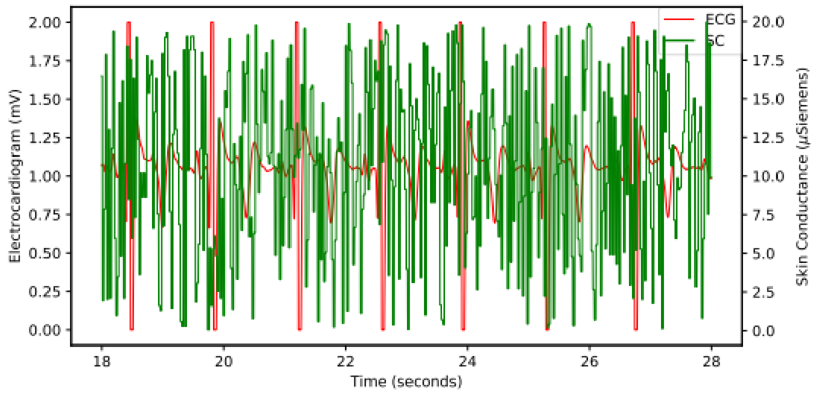
Fundamental Assumption of Learning:

- The distribution of training examples is identical to the distribution of test examples (including future unseen examples).
- Training examples must be sufficiently representative of (future) test data.



General Supervision

- Time series analysis problem
- Data collected during a sensory challenge protocol (SCP) in which the reactions to eight stimuli were observed.
- Based on electrocardiogram (ECG) and skin conductance.
- Multivariate time series /w 2M+ samples for each subject.



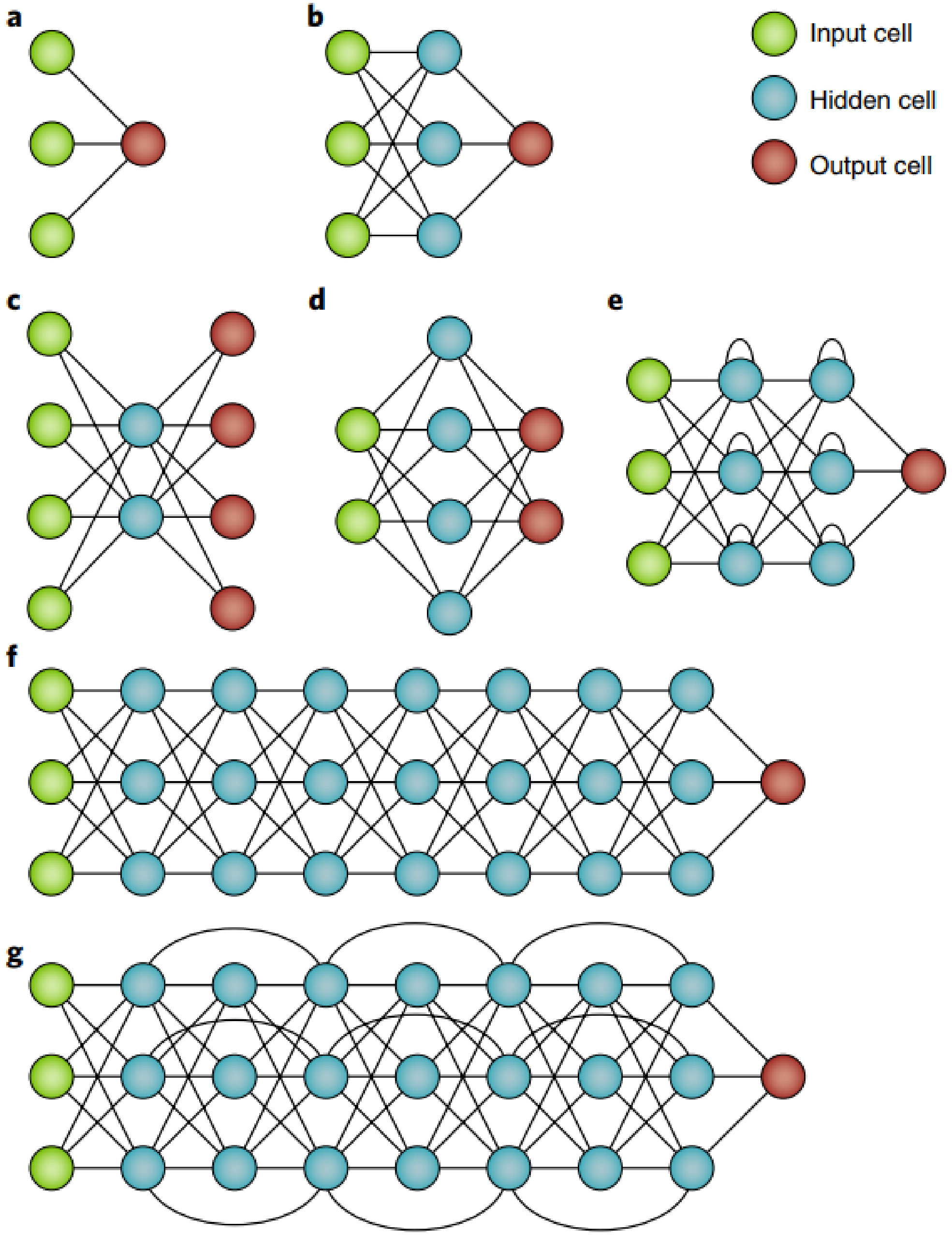
Example:

Autism Spectrum Prediction

- /w Dr. Megan C. Chang
- Student: Manika Kapoor

Deep Learning

- A subfield of machine learning.
- Uses artificial neural networks (ANNs) with many layers for pattern discovery.
- Building block: Perceptron
  - $\boldsymbol{w} \cdot \boldsymbol{x} + b > 0$
- DNNs can approximate infinite functions
- Needs sufficient labeled input
  - Avoid overfitting



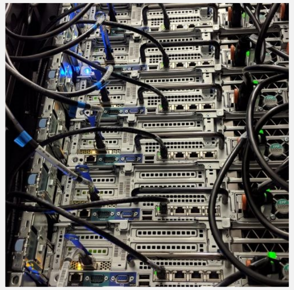
Credit: Figure adapted from Yu et al., Artificial Intelligence in healthcare, Nature Biomedical Engineering, VOL 2,OCTOBER 2018, 719–731, https://www.nature.com/articles/s41551-018-0305-z.pdf

What Made Deep Learning Possible

Many Core Hardware

GPUs, TPUs

Thousands of cores, really good at dense matrix operations.



Distributed Computing

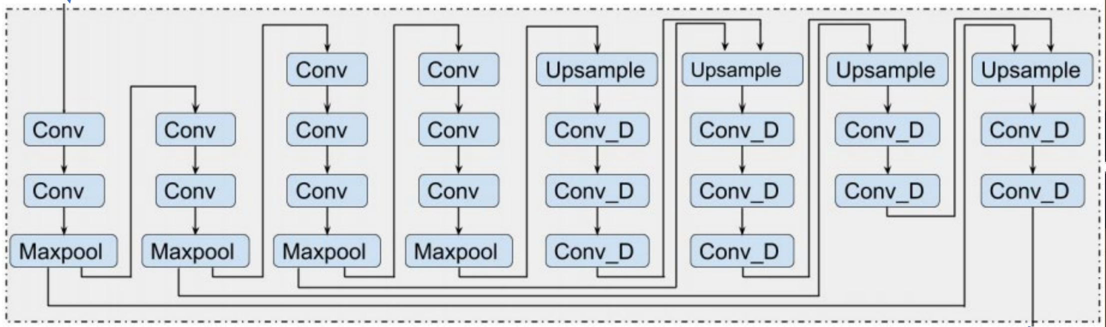
Supercomputing
Shared-nothing Computing

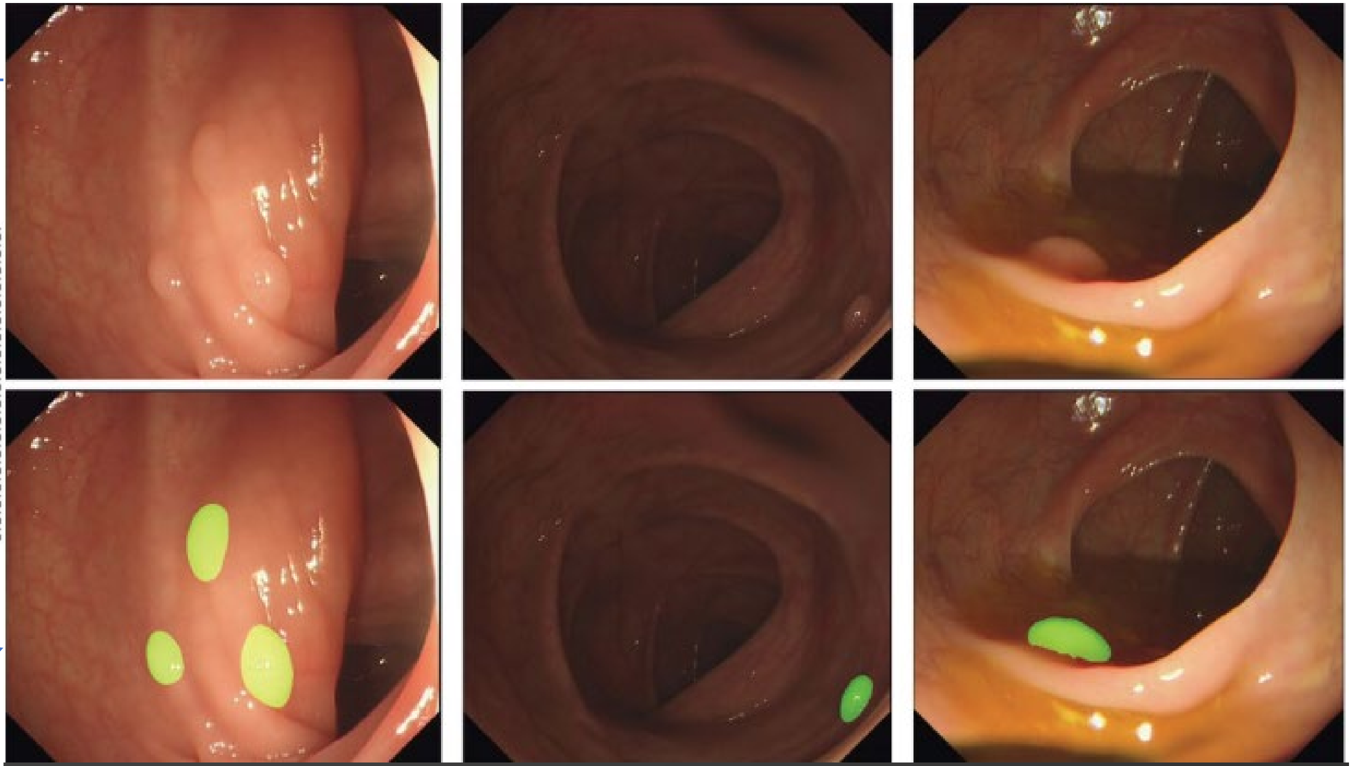
Split the work among many systems.

Lots of Labelled Data

Mechanical Turk
Label Generation

Split the work among many humans, or be clever about creating labels



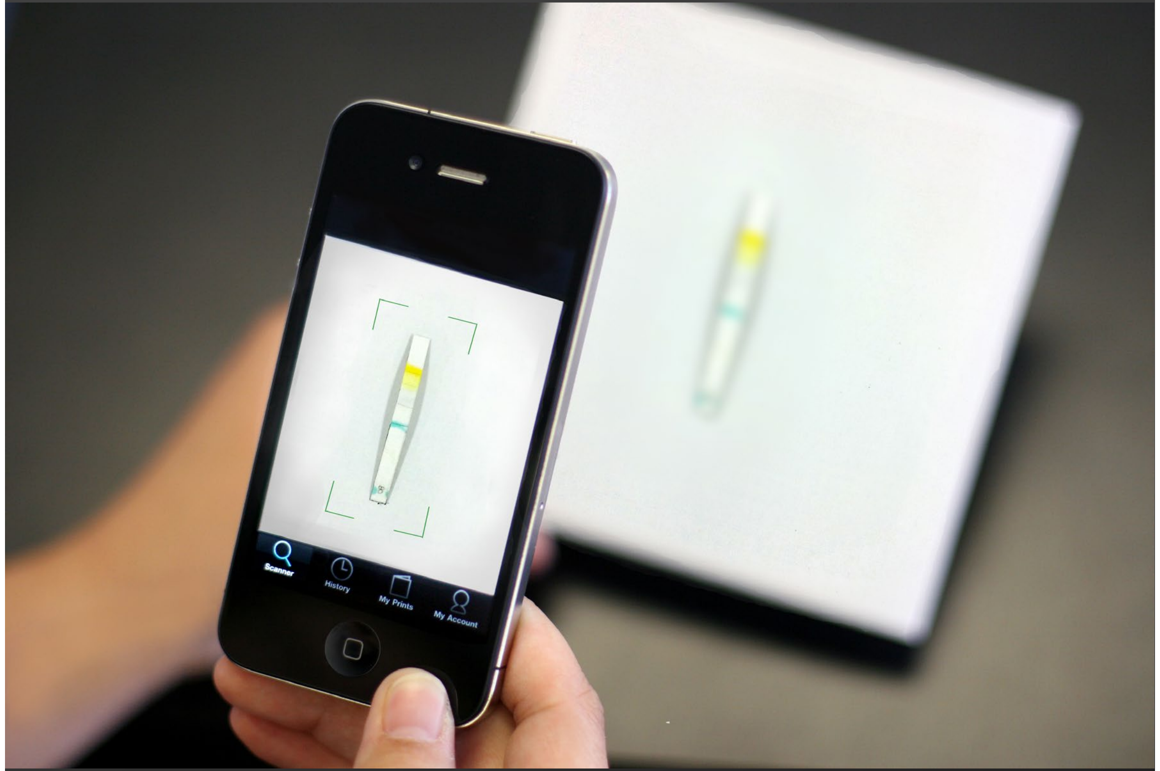


Deep Learning

- Convolutional Neural Network (CNN)
- SegNet Architecture
- Compute the probability of each pixel belonging to a polyp

Example: **Real-time detection of polyps during colonoscopy using machine learning.**

Credit: Figures adapted from Wang et al., Nature Biomedical Engineering, volume 2, pages741–748 (2018) https://www.nature.com/articles/s41551-018-0315-x.pdf

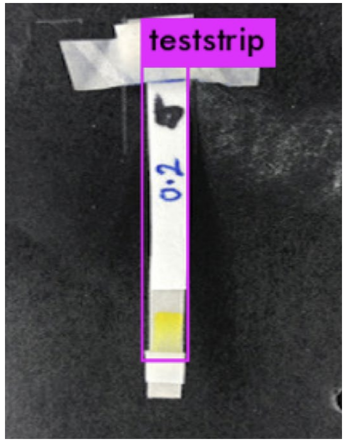


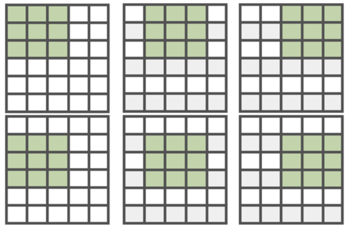
Example:
Kidney Health Monitoring

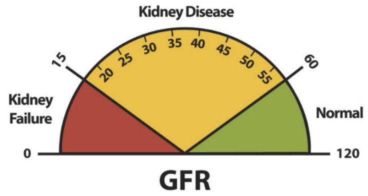
/w Dr. Alessandro Bellofiore

Students: Rathna Ramesh, Ragwa Elsayed

Borrowed Supervision
(Transfer Learning)







Localization

Transfer Learning

Deep Learning – YOLO

Alternatives

Feature Extraction

Color-based features:

RGB, Color Histogram, Gradient Histogram

Prediction

Creatinine level (regression)

Kidney health level (classification)

General/Borrowed Supervision

2017 AI City Challenge

- Collaborative annotation
- Over 150,000 annotations from 80 h video
- Localization and classification

Thomas Tang & team - UW

Charles MacKay

2018 AI City Challenge

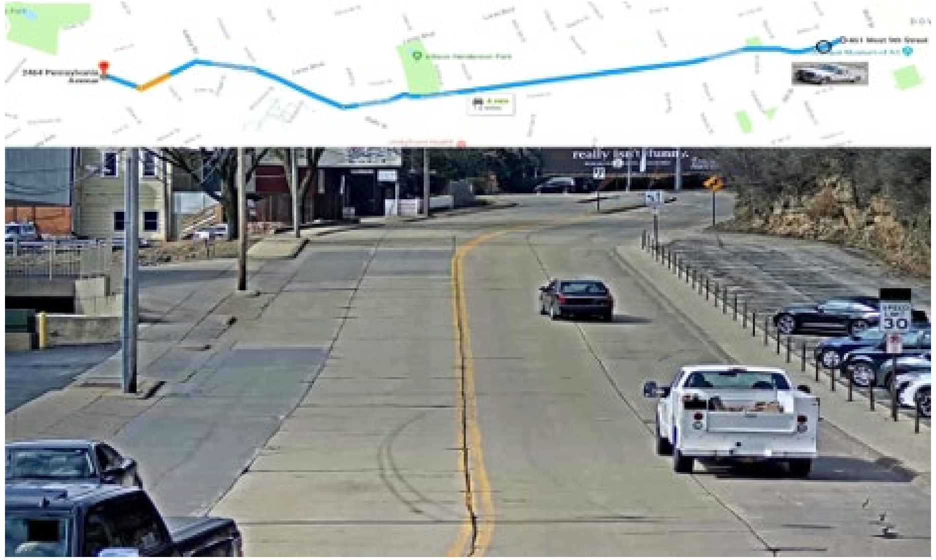
- GPS-based annotation
- 27 videos, 3 locations
- Speed estimation, anomaly detection, re-identification and tracking

Semi/Aided Supervision

2019 AI City Challenge

Multi-car multi-video tracking

- Applications in corridor-level traffic flow optimization
- Videos from 40 cameras within 6 km^2 of a city (3 hours)
- Hybrid models (single-camera tracking, deep-learning based Re-ID, camera linking)



Vehicle Re-ID

- Applications in vehicle counting, traffic volume estimation
- 56K vehicle images, 1K queries
- Visual features from CNNs + semantic features from travelling direction and vehicle type classification

2019 AI City Challenge

Multi-car multi-video tracking

- Applications in corridor-level traffic flow optimization
- Videos from 40 cameras within 6 km^2 of a city (3 hours)
- Hybrid models (single-camera tracking, deep-learning based Re-ID, camera linking)

Rank	Team ID	Team name (and paper)	IDF1
1	21	U. Washington IPL [12]	0.7059
2	49	DiDi Global [20]	0.6865
3	12	BUPT Traffic Brain [10]	0.6653
5	97	Australian National U. [11]	0.6519
6	59	Baidu ZeroOne [33]	0.5987
9	104	Shanghai Tech. U. [40]	0.3369
10	52	CUNY NPU [9]	0.2850
17	79	NCCU-UAlbany [7]	0.1634
18	64	GRAPH@FIT [32]	0.0664
19	43	U. Autonoma de Madrid [23]	0.0566

Rank	Team ID	Team name (and paper)	rank-K mAP
1	59	Baidu ZeroOne [33]	0.8554
2	21	U. Washington IPL [14]	0.7917
3	97	Australian National U. [24]	0.7589
4	4	U. Tech. Sydney [42]	0.7560
5	12	BUPT Traffic Brain [10]	0.7302
8	5	U. Maryland RC [16]	0.6078
13	27	INRIA STARS [8]	0.5344
18	24	National Taiwan U. [22]	0.4998
19	40	Huawei AI Brandits [2]	0.4631
23	52	CUNY-NPU [9]	0.4096
25	113	VNU HCMUS [27]	0.4008
36	26	SYSU ISENET [13]	0.3503
45	64	GRAPH@FIT [32]	0.3157
50	79	NCCU-UAlbany [7]	0.2965
51	63	Queen Mary U. London [15]	0.2928
54	46	Siemens Bangalore [17]	0.2766
60	43	U. Autonoma de Madrid [23]	0.2505

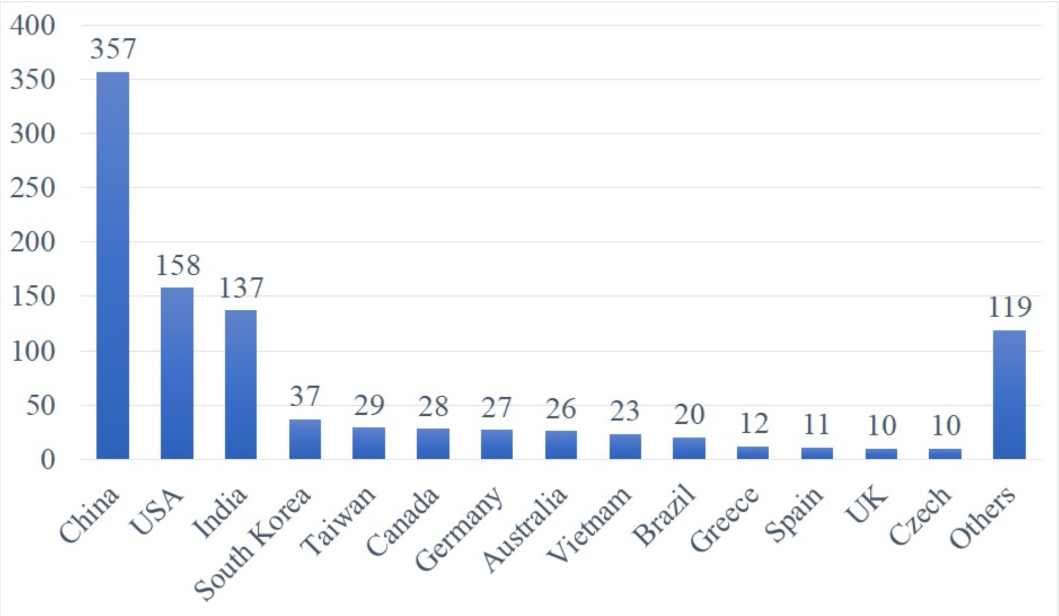
Vehicle Re-ID

- Applications in vehicle counting, traffic volume estimation
- 56K vehicle images, 1K queries
- Visual features from CNNs + semantic features from travelling direction and vehicle type classification

2019 AI City Challenge

Traffic Anomaly Detection

- Speed up emergency response
- 50 h of highway videos from Iowa
- Foreground segmentation + spatio-temporal anomaly detection



Challenge Stats

- 334 teams (1,004 researchers) signed up
- Competitive evaluation
- 129 track submissions (1,337 solution submissions) from 96 teams (355 researchers)
- CVPR 2019 Workshop

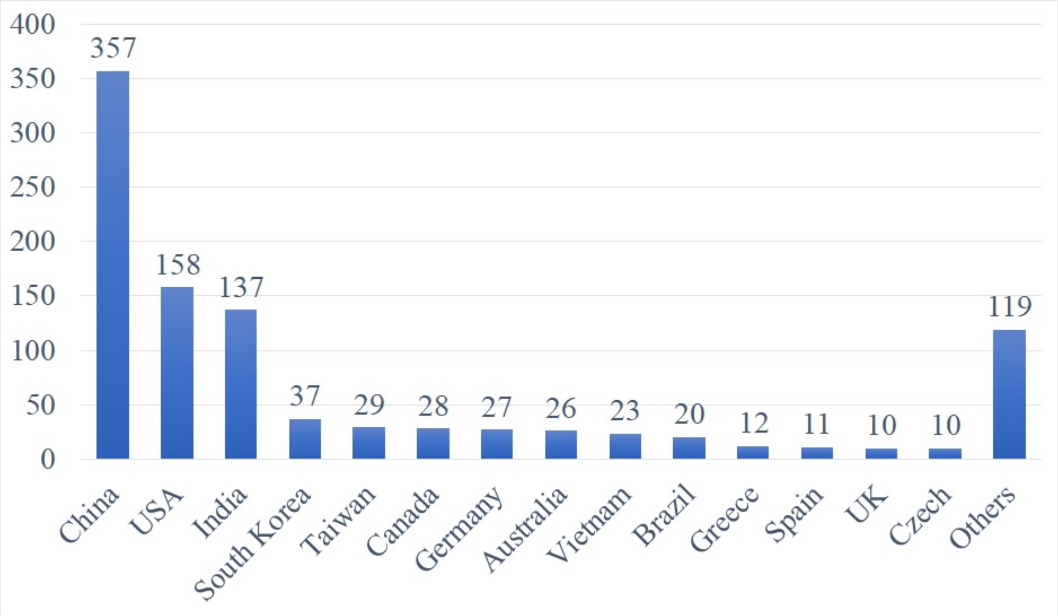
Check out AI City Challenge 2020

2019 AI City Challenge

Traffic Anomaly Detection

- Speed up emergency response
- 50 h of highway videos from Iowa
- Foreground segmentation + spatio-temporal anomaly detection

Rank	Team ID	Team name (and paper)	S_3
1	12	BUPT Traffic Brain [3]	0.9534
2	21	U. Washington IPL [39]	0.9362
6	79	NCCU-UAlbany [7]	0.6997
7	48	BUPT MCPRL [41]	0.6585
8	113	VNU HCMUS [27]	0.6129
12	65	Trivandrum CET CV [31]	0.3636
16	61	MNIT Vision Intelligence [6]	0.2641
17	5	U. Maryland RC [16]	0.2207



Challenge Stats

- 334 teams (1,004 researchers) signed up
- Competitive evaluation
- 129 track submissions (1,337 solution submissions) from 96 teams (355 researchers)
- CVPR 2019 Workshop

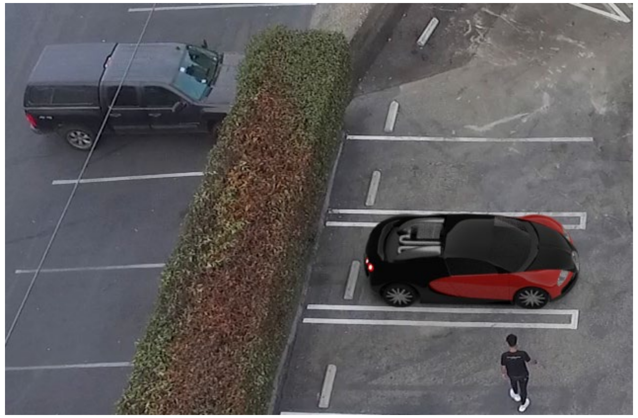
Check out AI City Challenge 2020

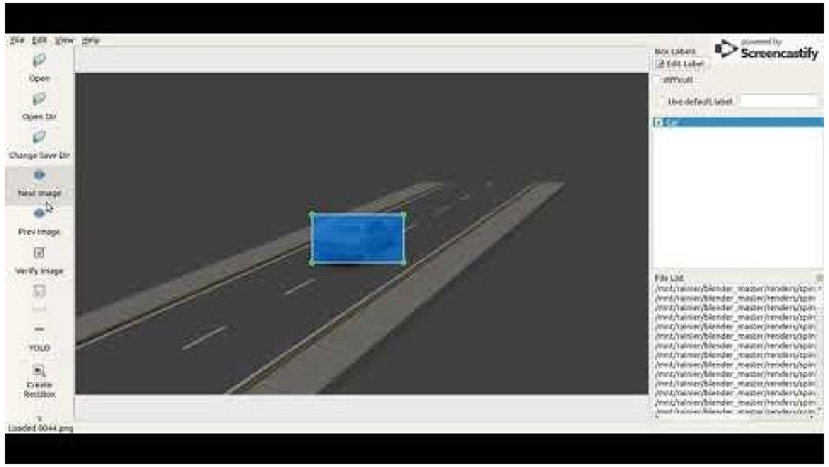
Generated Supervision

Charles MacKay

CGI-based Labelling

- Blender-generated objects w/ motion
- Bounding box automatically generated
- GAN-based smoothing to improve blend

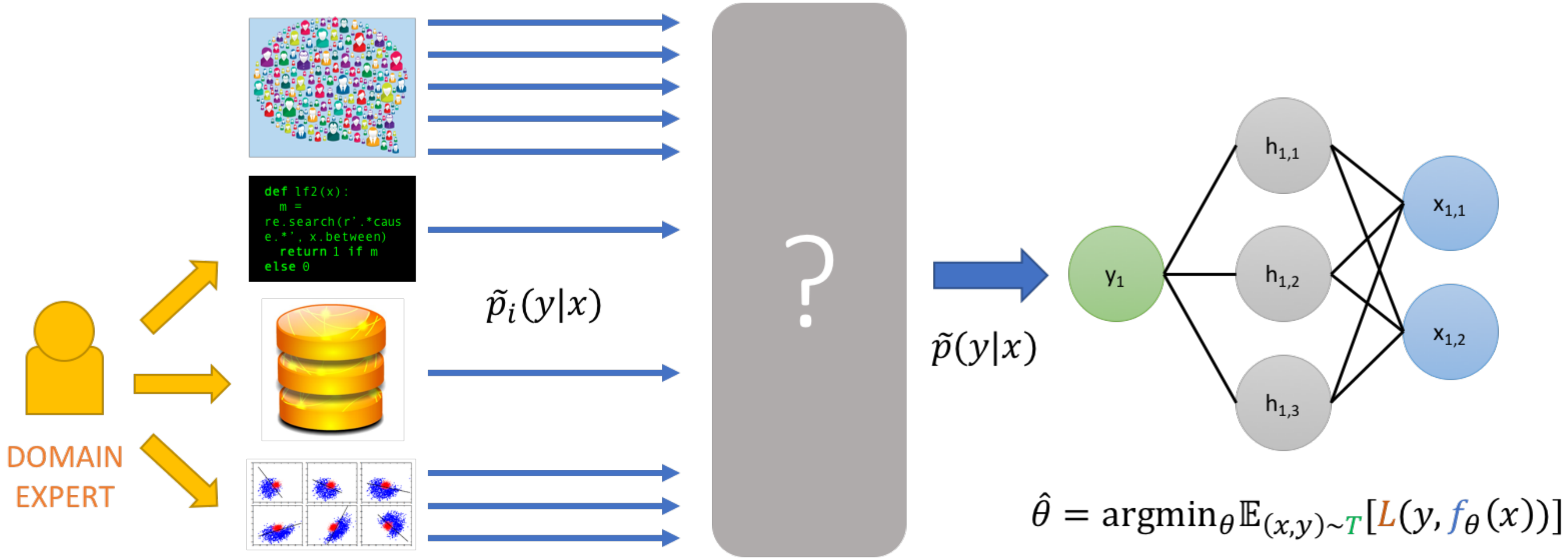




Anomaly Detection in Expense Reports

- Receipt localization dataset generation
- Small representative set of receipts
- GAN-based model for background
- Generate millions of receipts & combinations

Guided Supervision



Example Weak Supervision Sources

Technical Challenge: Integrating & Modeling Diverse Sources

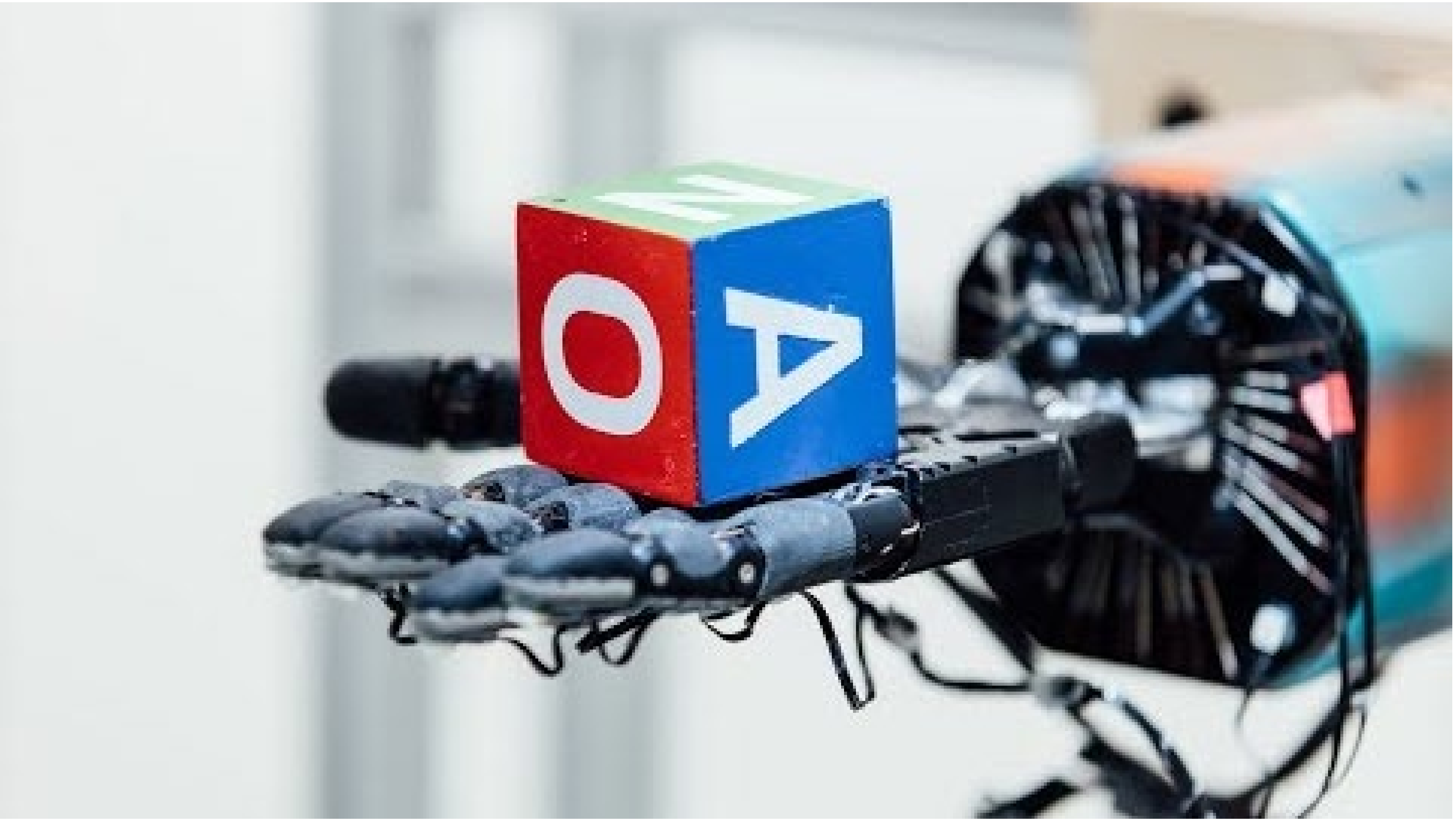
Use Weak Supervision to Train End Model

- Weak Supervision Work by Alex Ratner & Chris Ré at Stanford

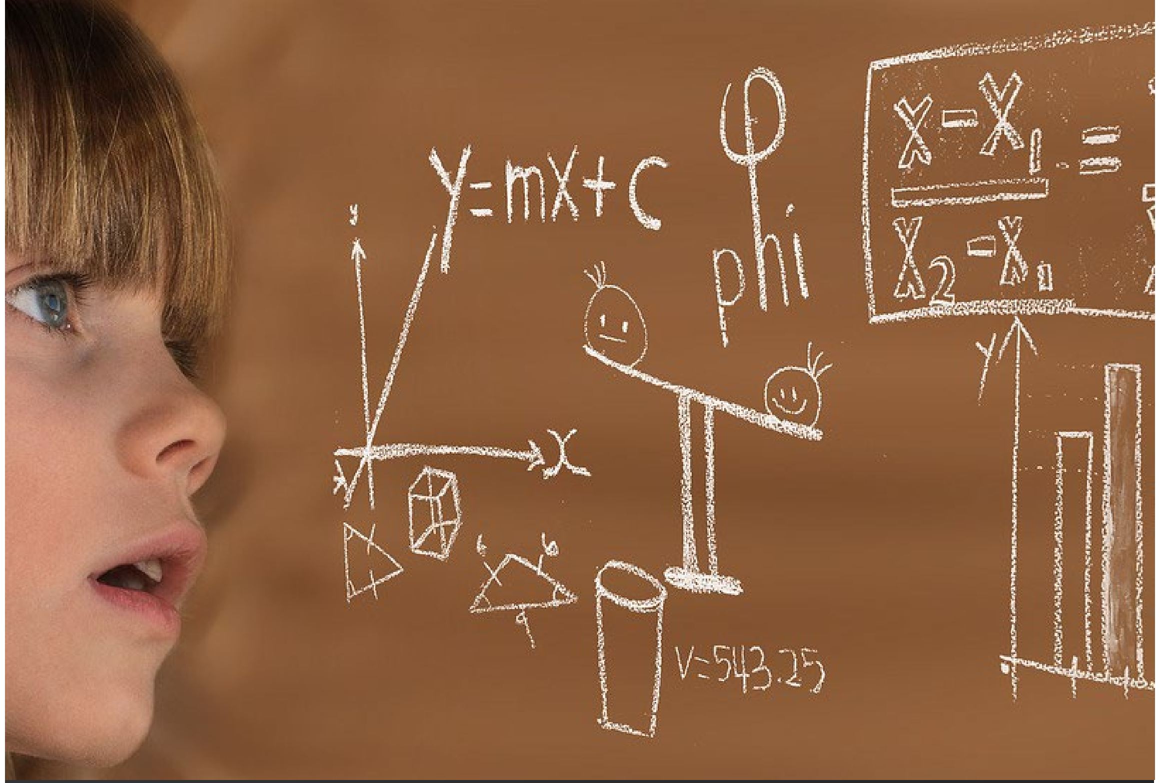
https://hazyresearch.github.io/snorkel/blog/ws_blog_post.html

- Provide inexact rules covering many (but not all) cases
- Learn how to combine weak predictors into a strong one

Taught Supervision



https://youtu.be/jwSbzNHGflM?t=28

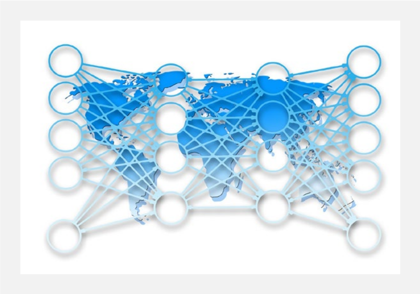


Few Labels?

No worries…

Label	**Borrow**	**Prioritize**
Crowd-Source Label Generation	Transfer Learning	Semi-Supervised Learning
Generate	**Guide**	**Teach**
Data + Label Generation	Weak Supervision	Reinforcement Learning

Challenges of Deep Learning



Data Driven Behaviour

GPUs, TPUs

What is the path that will be taken at inference time?

Input Bias

(Un)intended consequences

Who’s being left out?

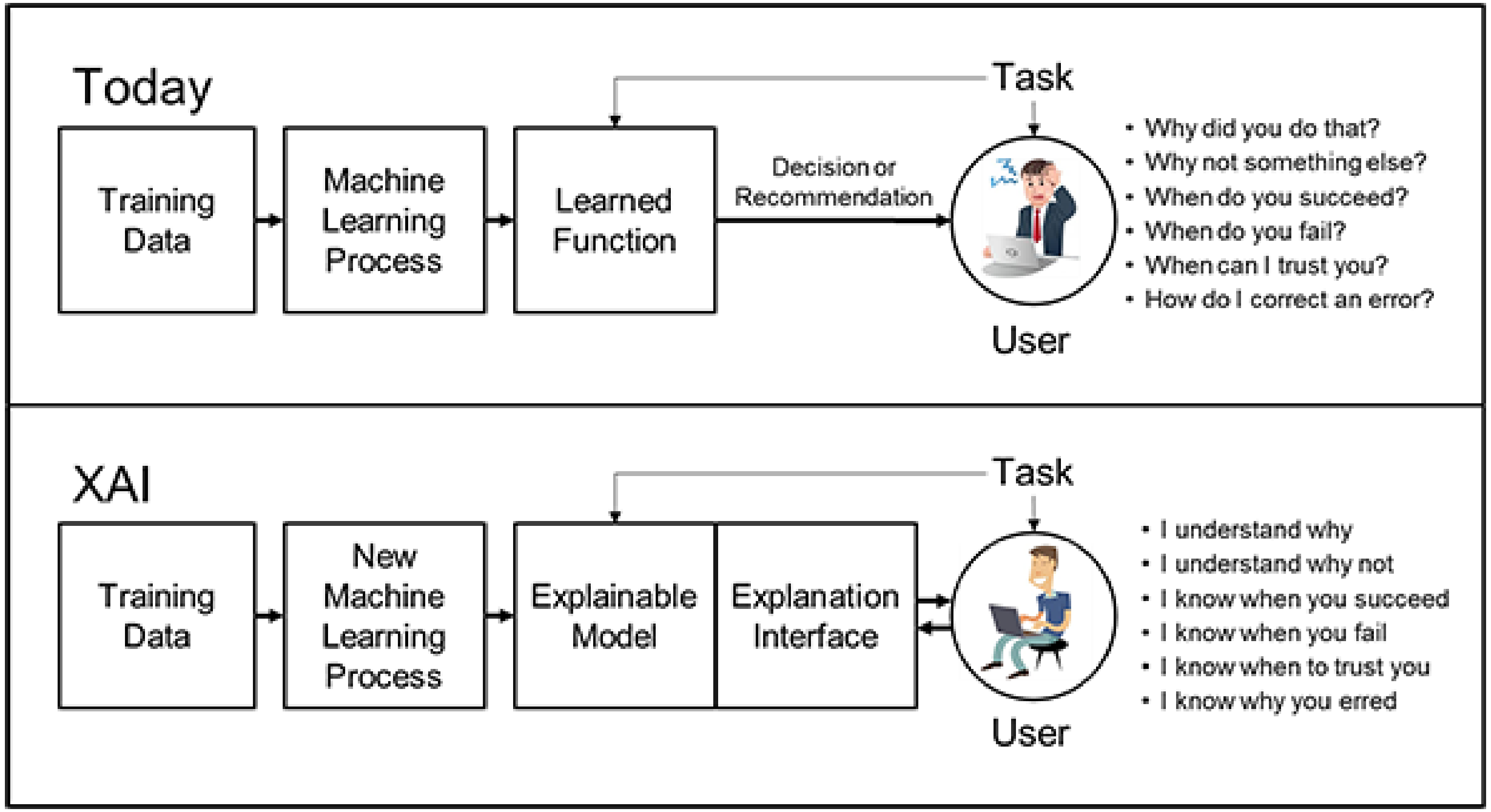
XAI

Explainable AI

Understand why the model made the decision

Explainable AI

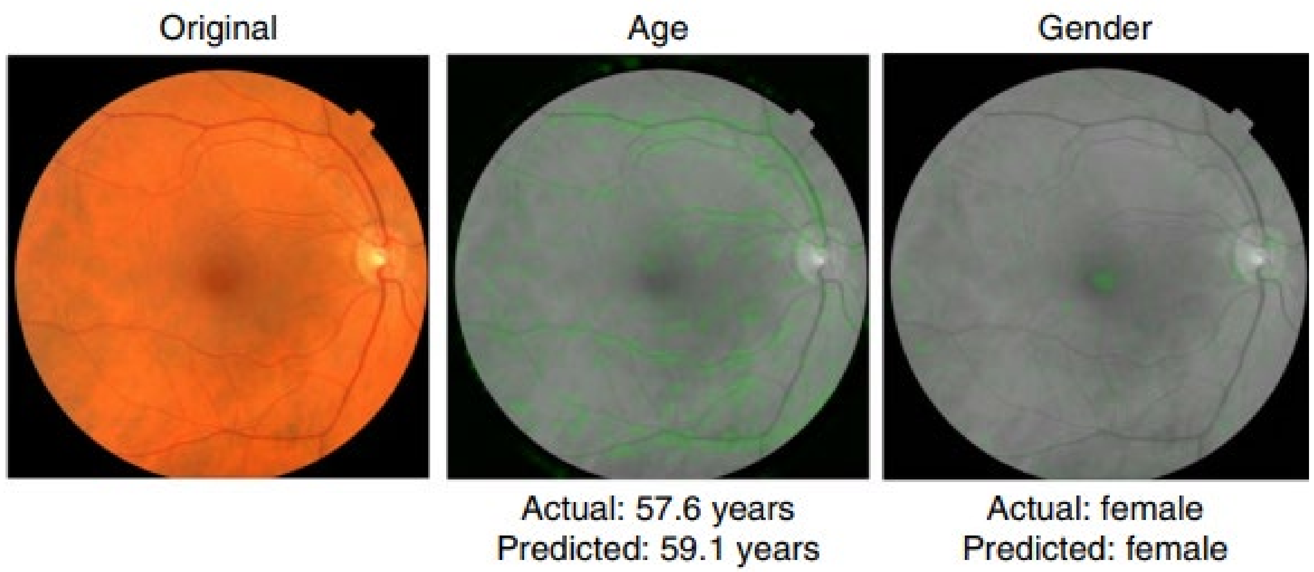
- Critical requirement for AI in many contexts beyond healthcare
  - DARPA/DoD priority
  - Already required in the General Data Protection Right (GDPR)
- RISE: randomized input sampling for explanation of black-box models



https://www.darpa.mil/program/explainable-artificial-intelligence

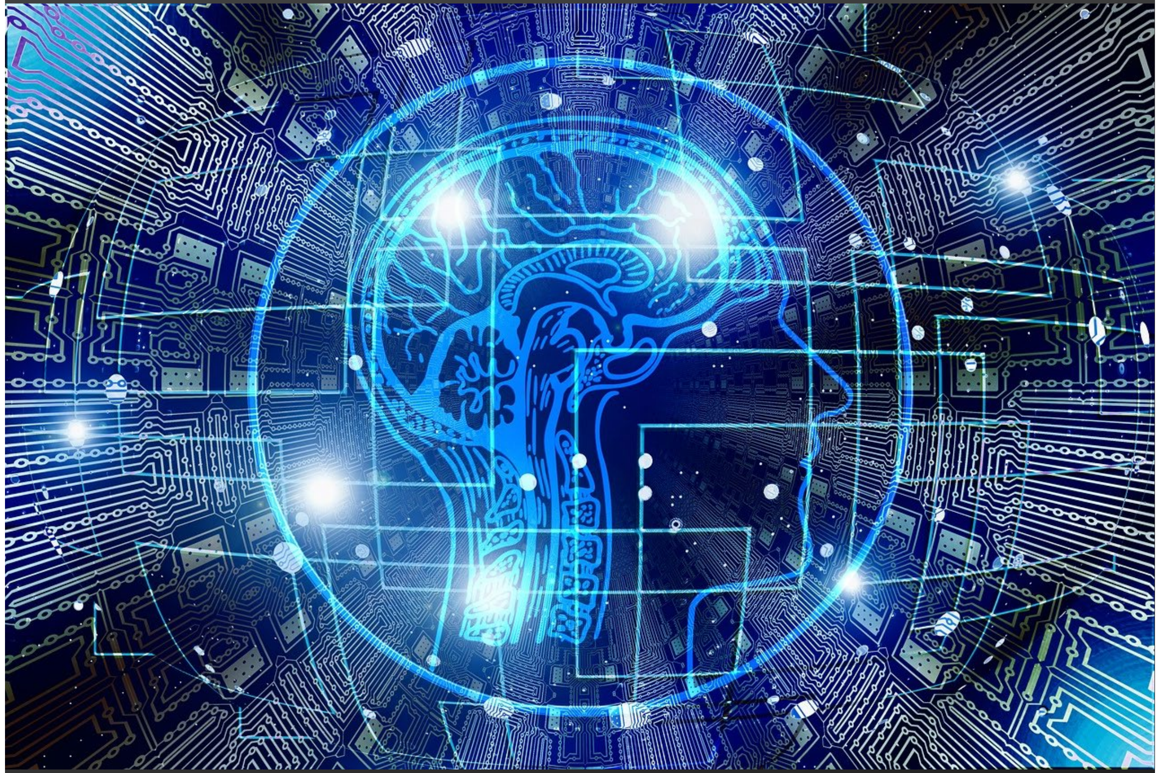
Deep Learning

- Convolutional Neural Network (CNN)
- Improved prediction of cardiovascular risk factors
- Predicted side-factors:
    - Age, gender, smoking, diabetic, BMI
- “Soft attention” used to point out the salient pixels



Example: **Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning**

Credit: Figures adapted from Poplin1 et al., Nature Biomedical Engineering, volume 2, pages 158–164 (2018) https://www.nature.com/articles/s41551-018-0195-0.pdf



Conclusions

- AI is here to stay (this time)
- Will permeate CS/Engineering (and many other fields)
- Learn machine learning
- Or partner with a data scientist