COURSE: APPLICATIONS OF ARTIFICIAL INTELLIGENCE (AAI)

AMP in Business Analytics 2020

INDIAN SCHOOL OF BUSINESS

Vamshi Ambati

Vamshi_Ambati@isb.edu;
vamshi@predera.com

Session 4: Computer Vision
March 1, 2020

USE CASES OF APPLICATION AREAS

- Natural Language Processing -> IE, IR, MT, NLU, NLG, Dialog -> Chatbot, Customer complaint translation
- Speech Processing -> SR (S2T), SS (T2S), SG (S2S) -> Voice Assistant (Siri), Education, IVR
- Computer Vision -> IU, IG, OR, OD, Video {U,G,R,D} -> OCR, Face Recognition, Self-driving, Entertainment
- Predictive Analytics ->Prediction, Recommender, Forecasting Systems -> E-commerce , Fraud detection
- Robotics -> Locomotion, Mechanics, Sensors, Planning -> Medical Bots, Agriculture, Military
- Agent Systems -> Planning, Reinforcement Learning, Space Optimization -> AlphaGo

THIS COURSE

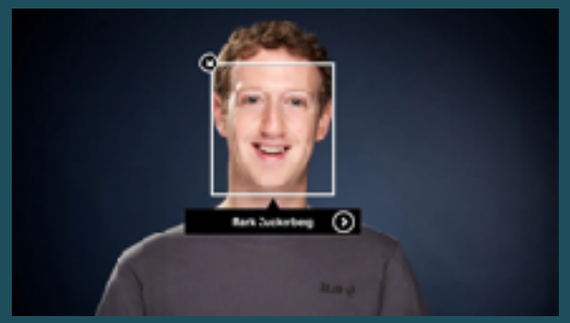
- Session 1: Machine Translation
- Session 2: Dialog Systems (Chatbot)
- Session 3: Recommender Systems
- Session 4: Computer Vision (Image | Video)

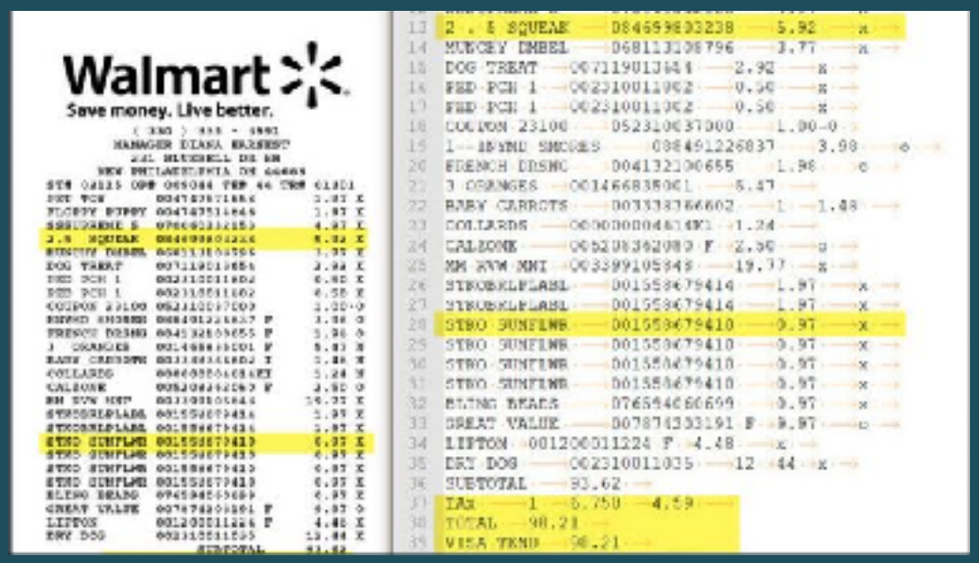
AGENDA

- Introduction
- Image Processing
  - Traditional Approaches
  - Deep Learning
  - Case study - nutrition label extraction
- Video Understanding
  - Traditional Approaches
  - Deep Learning
- Summary & Challenges

1. INTRODUCTION

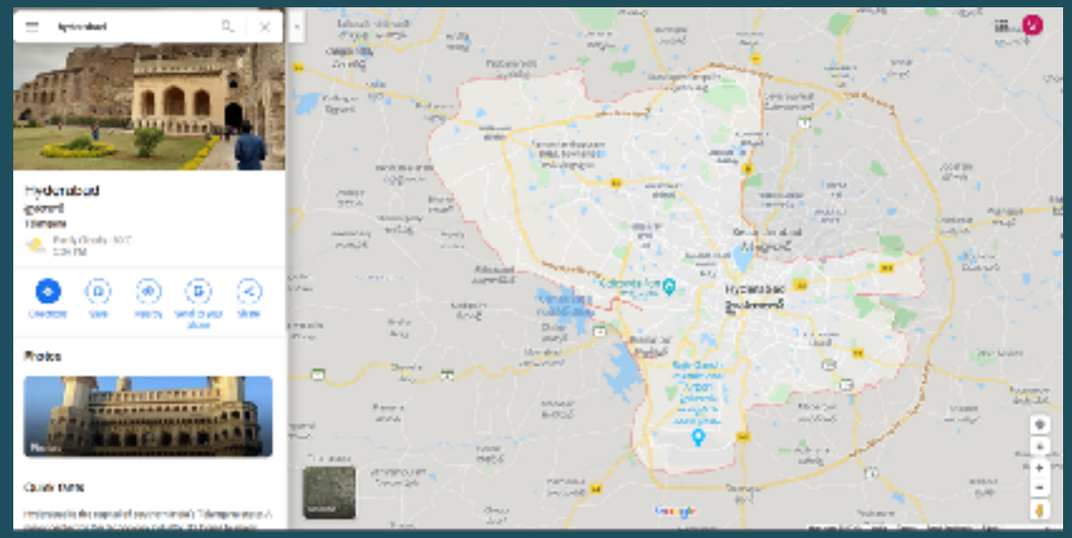
Introduction

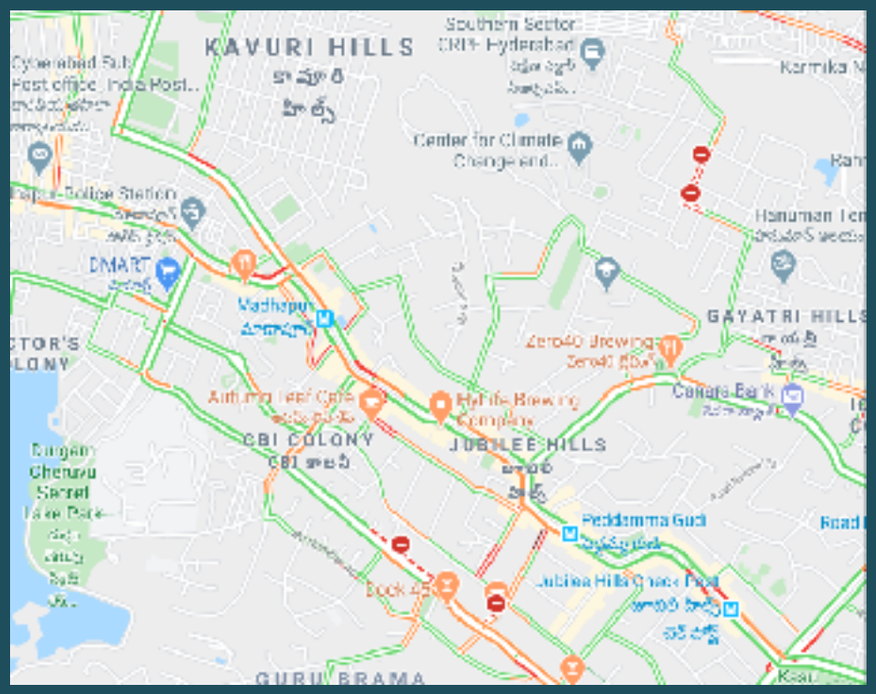




[Covid-19 from lung CT scans : https://www.wired.com/story/chinese-hospitals-deploy-ai-help-diagnose-covid-19/](https://www.wired.com/story/chinese-hospitals-deploy-ai-help-diagnose-covid-19/)

Google Maps




KAVURI HILLS
Center for Climate Change and...
DMART
Madhapur
GAYATRI HILLS
Canara Bank
CBI COLONY
JUBILEE HILLS
Peddamma Gudi
GURU BRAMA

History of Image Processing

- Origin at Bell Laboratories, American Jet Propulsion Laboratory, MIT
- Early purpose was to improve quality of images, owing to low resolution cameras
- Applications in space exploration and radiology
- Early success in 1964 through understanding surface of moon by improving quality of images sent back by spacecrafts

History of image processing

- 1970s- breakthroughs in computing, image compression
- 1980s- Contour modeling, texture understanding
- 1990-now: Machine learning and Deep learning
- Real time object detection, video analytics
- Today - security, retail, defence, art, healthcare, public policy, entertainment, education, space exploration, environment etc.

Challenges

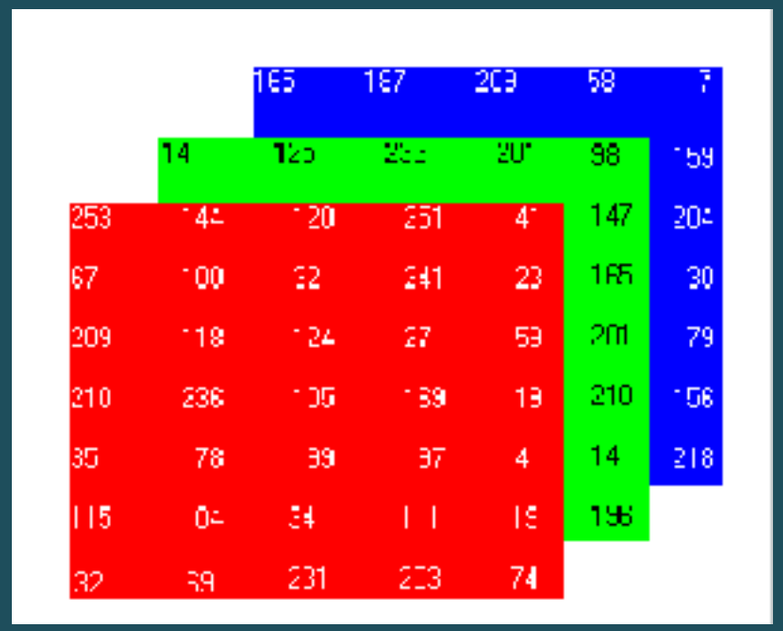
- Images of varied quality
- Size of objects of interest in images varies
- Background noise
- Varied colours
- Orientation
- Labeling
- Computing

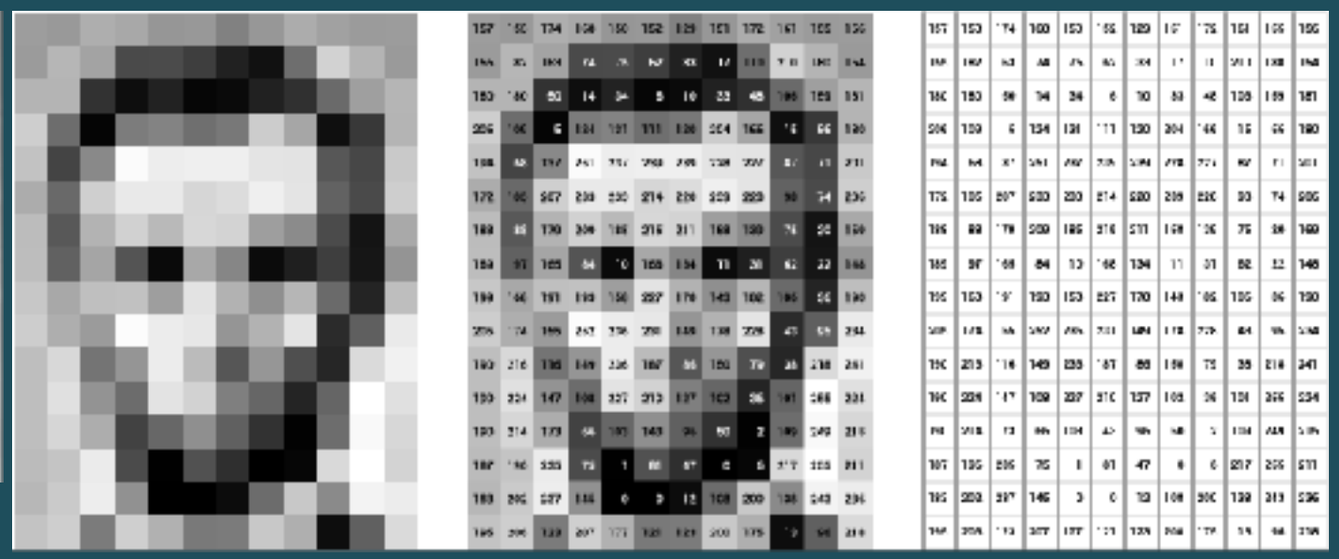
Cross section of an image

Resolution 1024X1024

Size = 1024X1024 * 3 {R,G,B}

Pixel value range: 0-127 or 0-255





Video- stream of images

Zoetrope

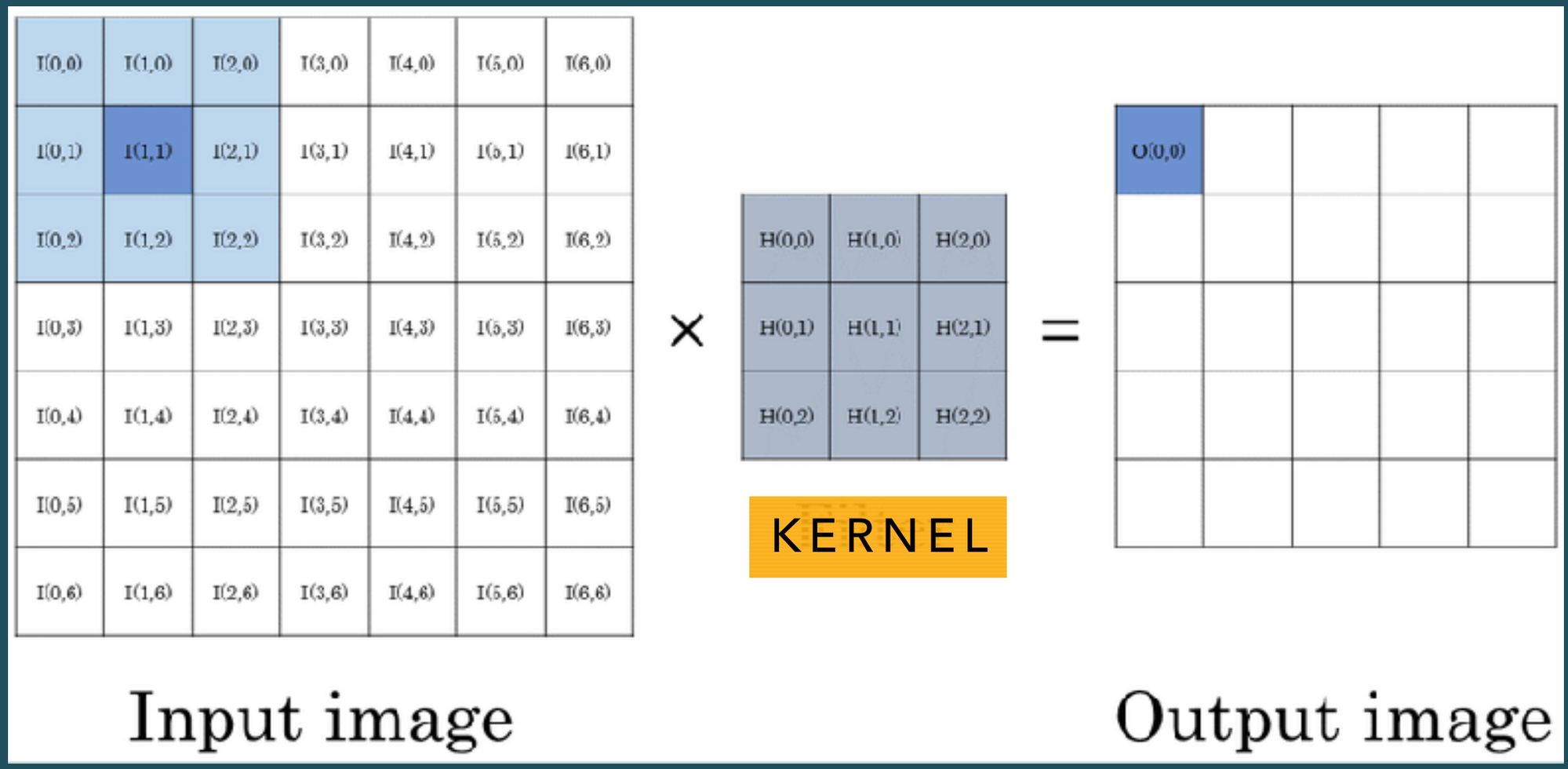
Flipbook

Video has frame rates of 24 to 30 per second

Image Processing Tasks

- Image augmentation
- Image enhancement/ blurring
- Noise removal
- Boundary detection

Convolution for Image Processing



Convolution provides a way of `multiplying together' two arrays of numbers to produce a third

Image enhancement: De-blurring

- Laplacian kernel

Image enhancement: Contrast Improvement

Image enhancement: Noise removal

More complex filters

Edge detection

Edge detection

Sobel Filter/Kernel

X – Direction Kernel

-1	0	1
-2	0	2
-1	0	1

Y – Direction Kernel

-1	-2	-1
0	0	0
1	2	1

Going beyond Image Tasks (Math gets hairy)

Image Processing tasks

- Image augmentation
- Image enhancement/ blurring
- Noise removal
- Boundary detection

Computer vision

- Face recognition
- Object detection
- Text detection- OCR
- Image classification
- Video processing

Recognition as Classification

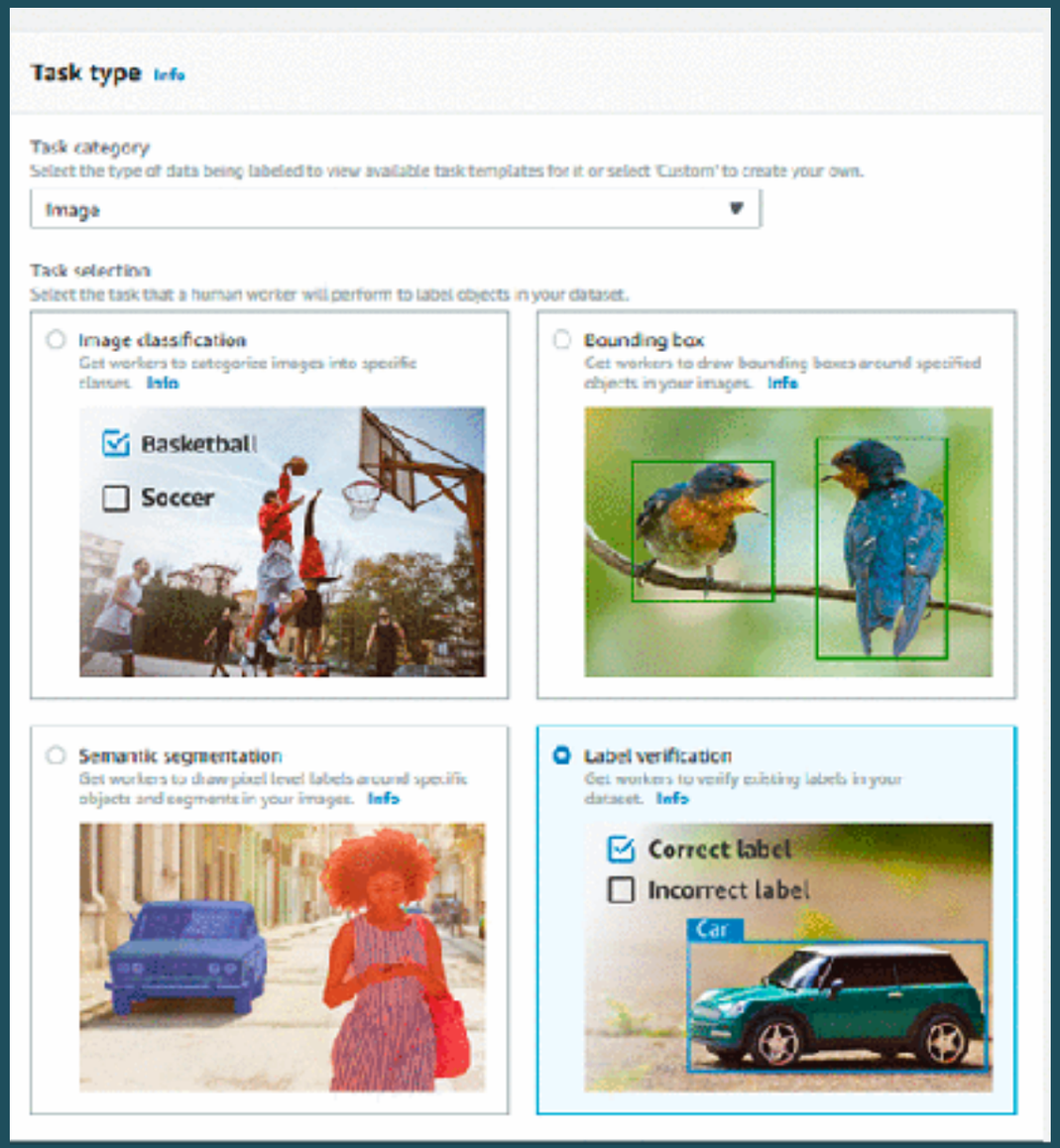
Face Recognition

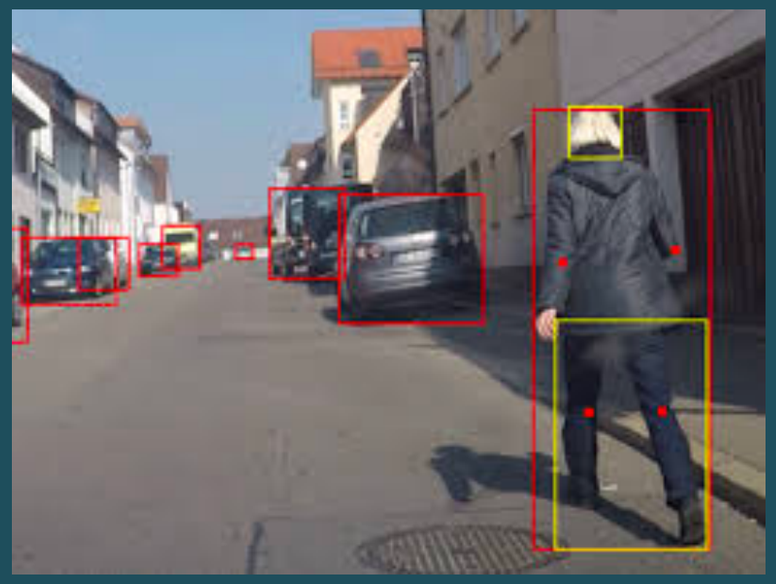
- Identify human
- Segment face
- Align position
- Classify face

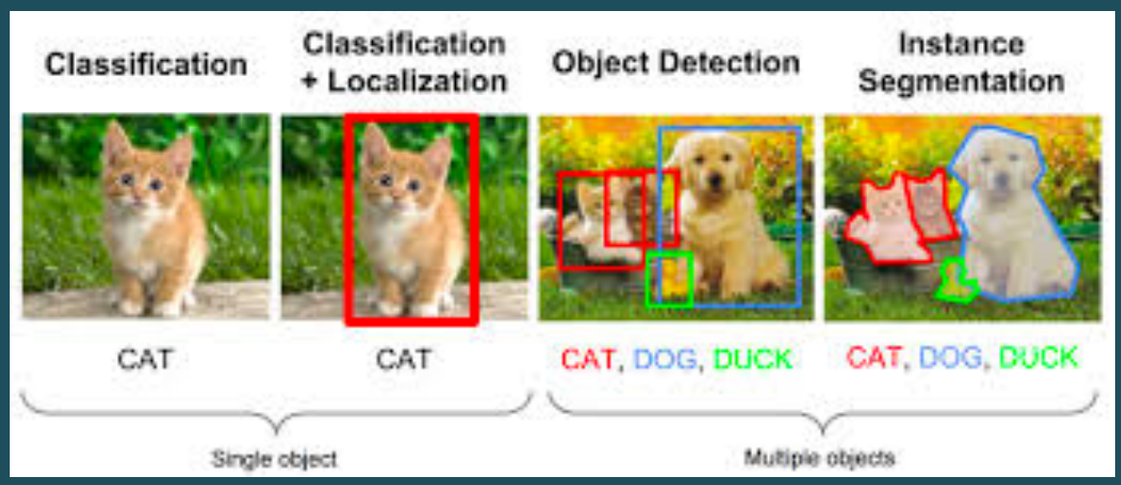
OCR

- Quality of page
- Extract text block
- Extract word block
- Extract letters
- Classify letter

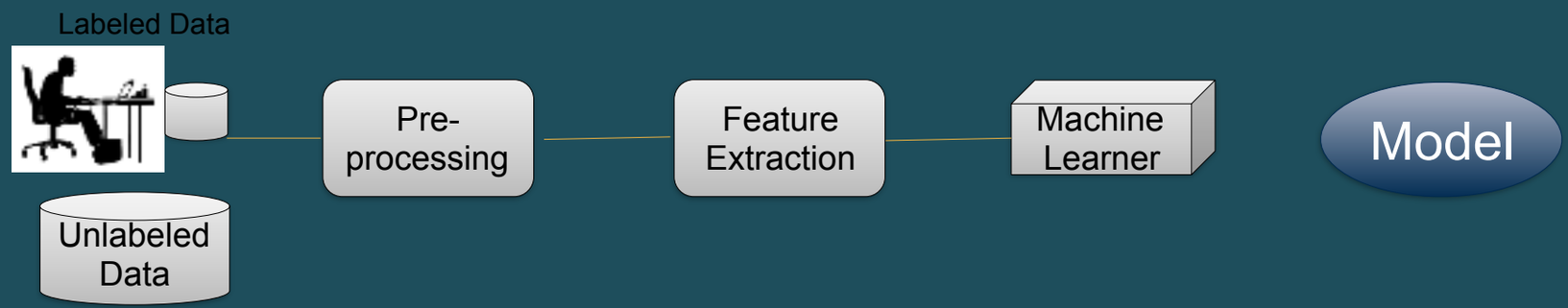
Image Classification as Supervised ML





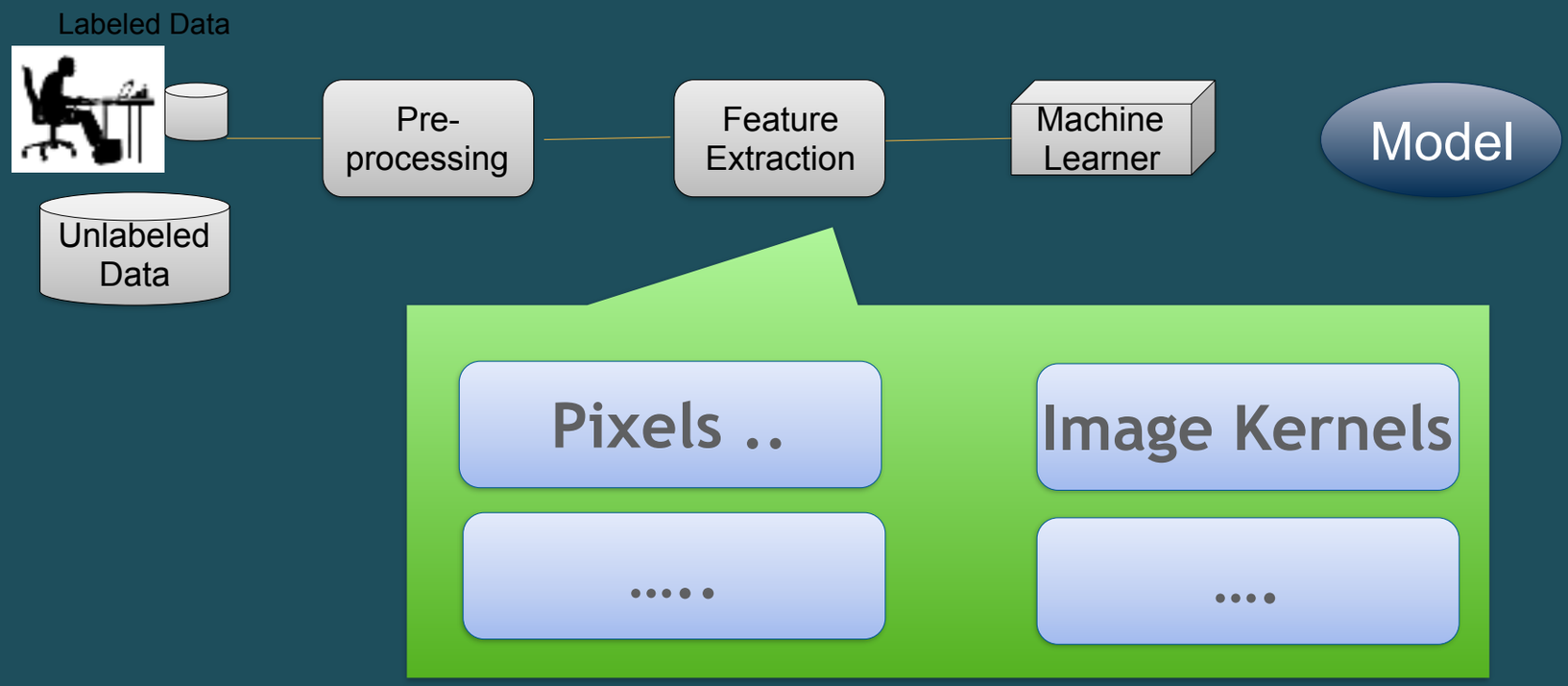


MACHINE LEARNING FOR FACE RECOGNITION



Nancy

MACHINE LEARNING FOR FACE RECOGNITION



MACHINE LEARNING FOR FACE RECOGNITION

Logistic Regression

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

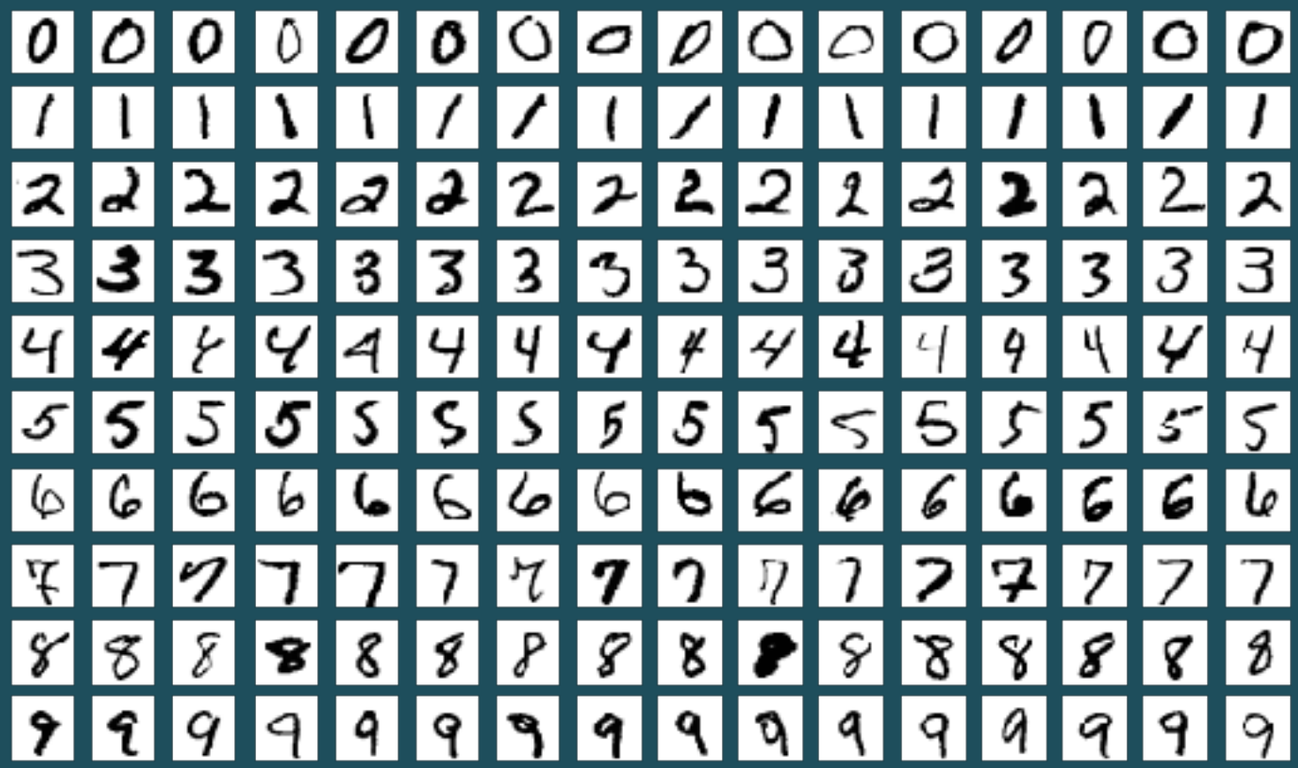
$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Maximum Entropy

$$Pr(c_i|t) = \frac{1}{Z(t)} exp\left(\sum_{j=1}^{n} \lambda_j f_{ij}(c_i, t)\right)$$

Image Classification Dataset (MNIST)

- Modified National Institute of Standards and Technology database
- Identify handwritten digits
- 28*28, 60,000 images



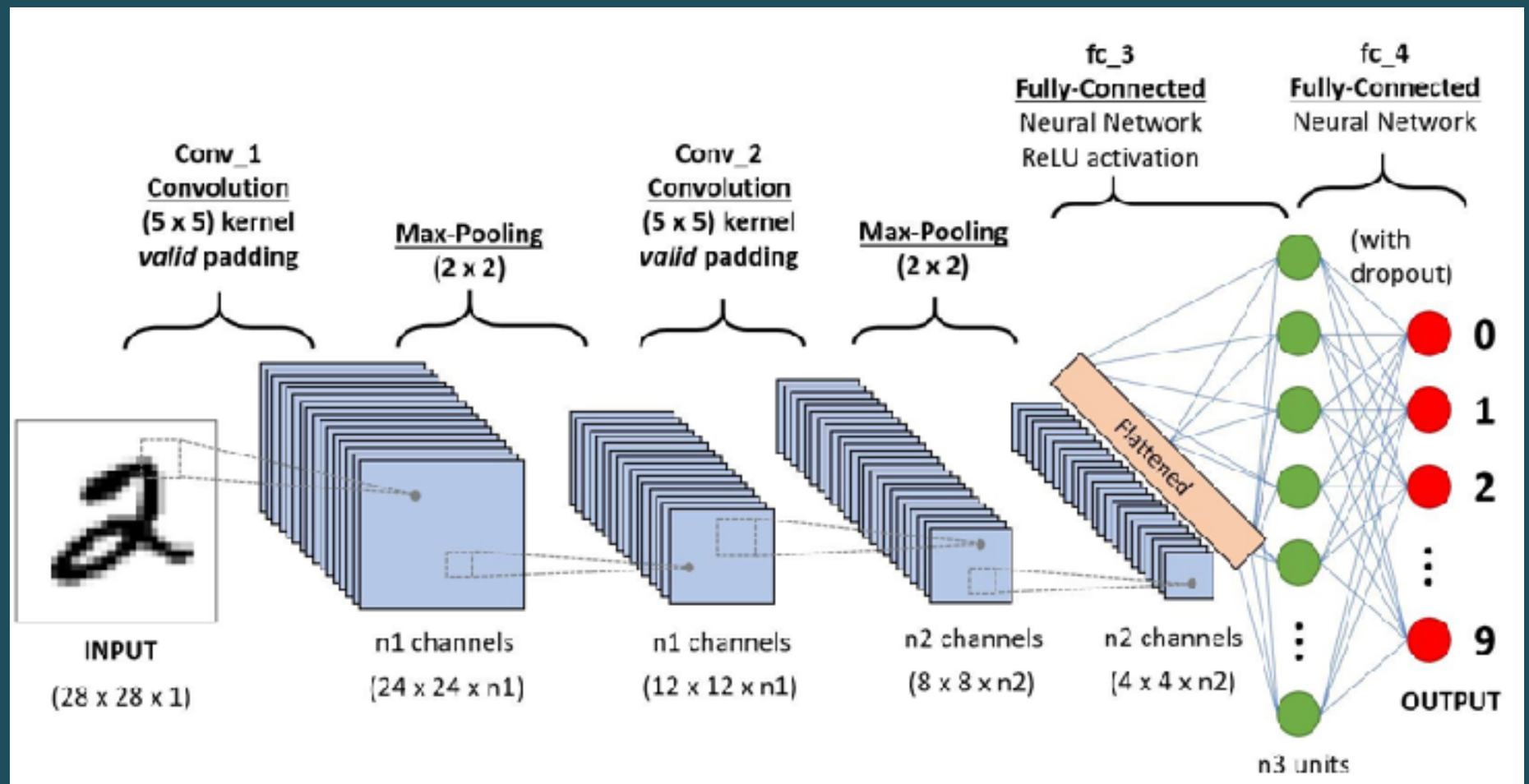
MNIST Classification

- Machine learning way
  - Each pixel as a feature

pixel_42	pixel_43	pixel_44	pixel_45	pixel_46	pixel_47	pixel_48	pixel_49	pixel_50	pixel_51	pixel_52	pixel_53	pixel_54	pixel_55	pixel_56	pixel_57	pixel_58	pixel_59	pixel_60	pixel_61	pixel_62	pixel_63	target
11	0	1	12	7	0	0	2	14	5	10	12	0	0	0	0	6	13	10	0	0	0	0
1	16	16	6	0	0	0	0	1	16	16	6	0	0	0	0	0	11	16	10	0	0	1
16	16	5	0	0	0	0	3	13	16	16	11	5	0	0	0	0	3	11	16	9	0	2
0	0	1	10	8	0	0	0	8	4	5	14	9	0	0	0	7	13	13	9	0	0	3
15	16	13	16	1	0	0	0	0	3	15	10	0	0	0	0	0	2	16	4	0	0	4
0	0	4	16	9	0	0	0	5	4	12	16	4	0	0	0	9	16	16	10	0	0	5
13	16	13	16	3	0	0	0	7	16	11	15	8	0	0	0	1	9	15	11	3	0	6
0	16	5	0	0	0	0	0	9	15	1	0	0	0	0	0	13	5	0	0	0	0	7
16	8	10	13	2	0	0	1	15	1	3	16	8	0	0	0	11	16	15	11	1	0	8
0	3	0	9	11	0	0	0	0	0	9	15	4	0	0	0	9	12	13	3	0	0	9
16	5	1	11	3	0	0	0	12	12	10	10	0	0	0	0	1	10	13	3	0	0	0
0	5	16	15	0	0	0	0	0	4	16	14	0	0	0	0	0	1	13	16	1	0	1
0	6	15	0	0	0	0	0	9	16	15	9	8	2	0	0	3	11	8	13	12	4	2
0	0	2	15	4	0	0	1	5	6	13	16	6	0	0	2	12	12	13	11	0	0	3
10	15	16	14	0	0	0	0	0	1	16	10	0	0	0	0	0	10	15	4	0	0	4
0	5	16	3	0	0	0	1	5	15	13	0	0	0	0	4	15	16	2	0	0	0	5
15	16	9	9	14	0	0	0	3	14	9	2	16	2	0	0	0	7	15	16	11	0	6
8	12	14	8	3	0	0	0	0	10	13	0	0	0	0	0	0	11	9	0	0	0	7

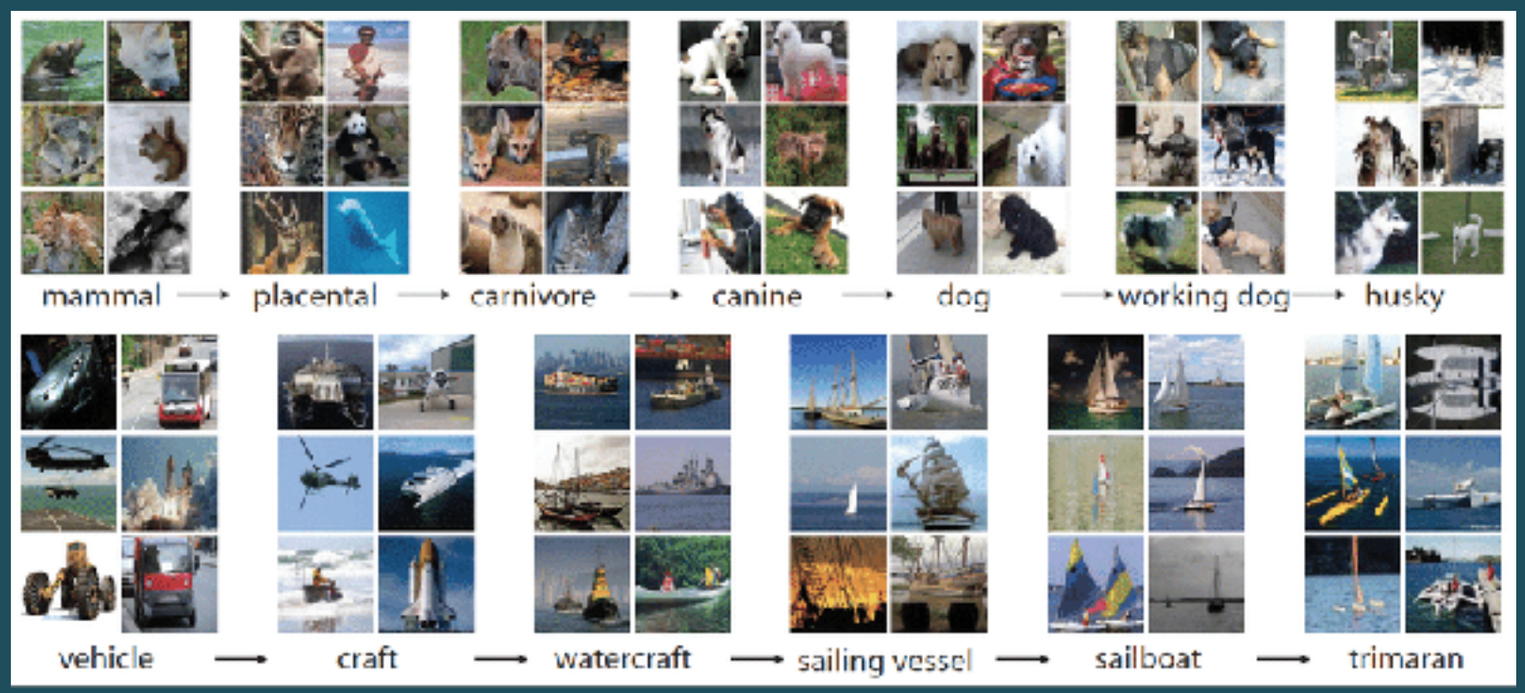
MNIST classification

- Deep learning way- Convolutional Neural Networks

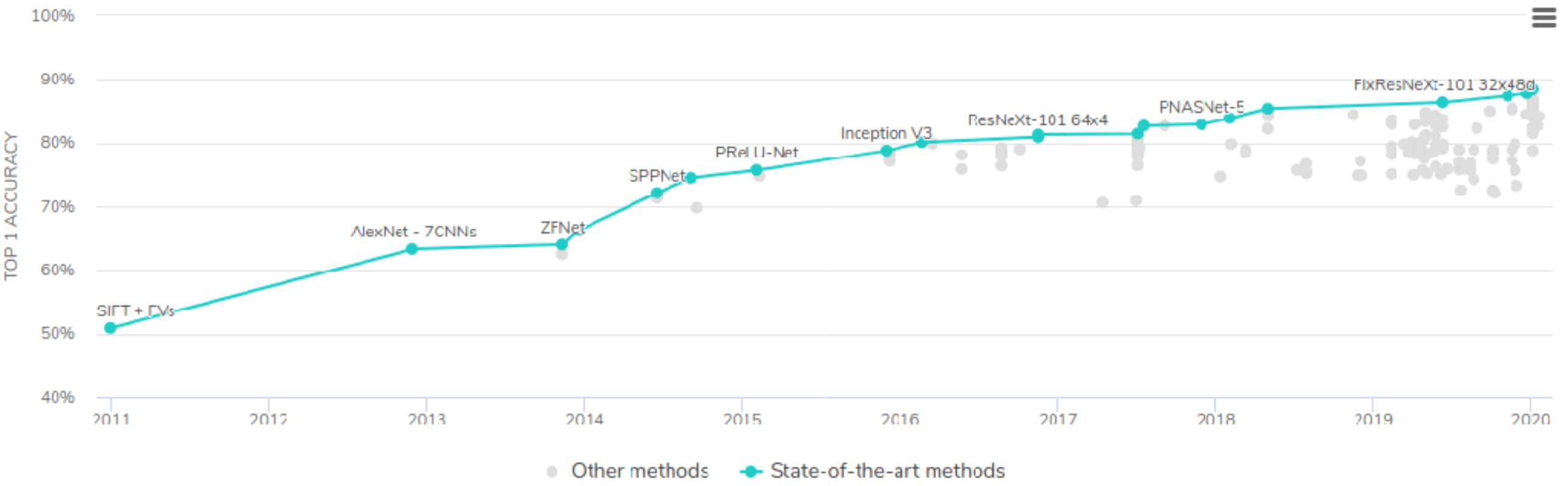


ImageNet

- 14,197,122 images, > 20,000 categories
- Images hand annotated through Amazon Mechanical Turk
- ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
  - A subset of it (1000 class dataset) is used to report results



ImageNet results



ImageNet architectures

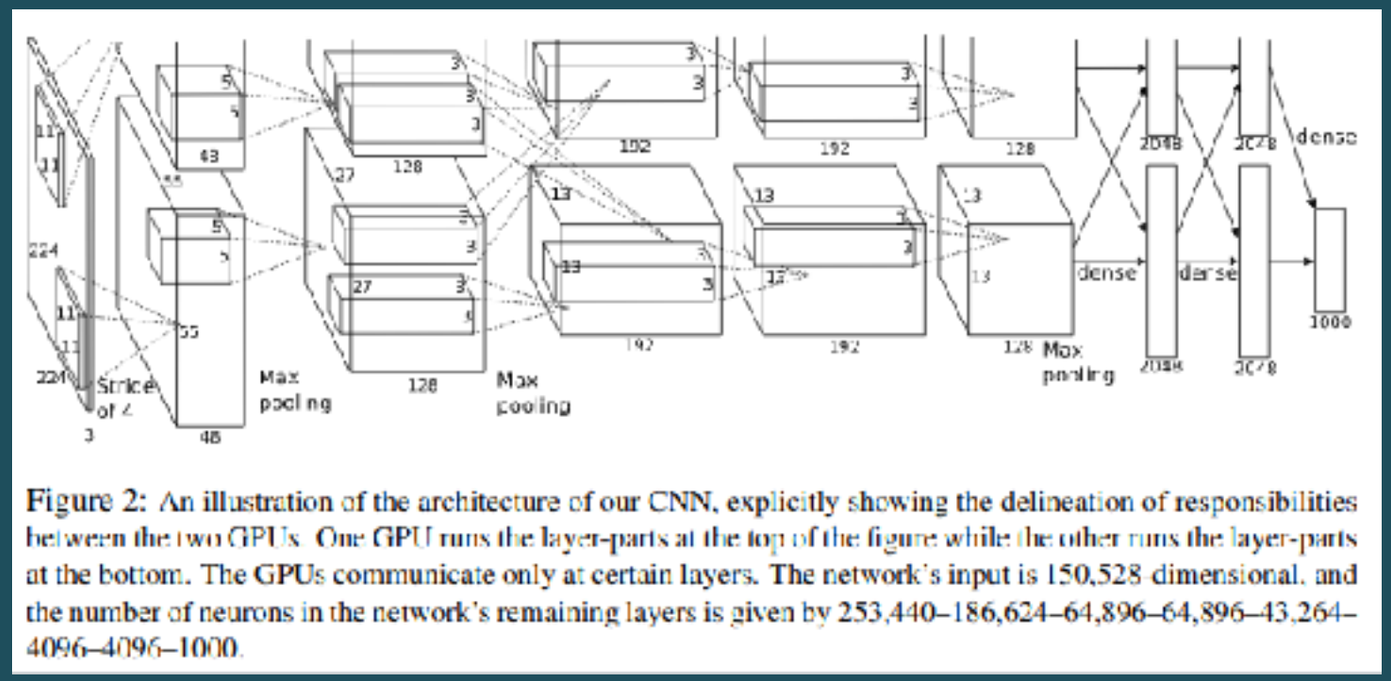


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528 -dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

Alexnet

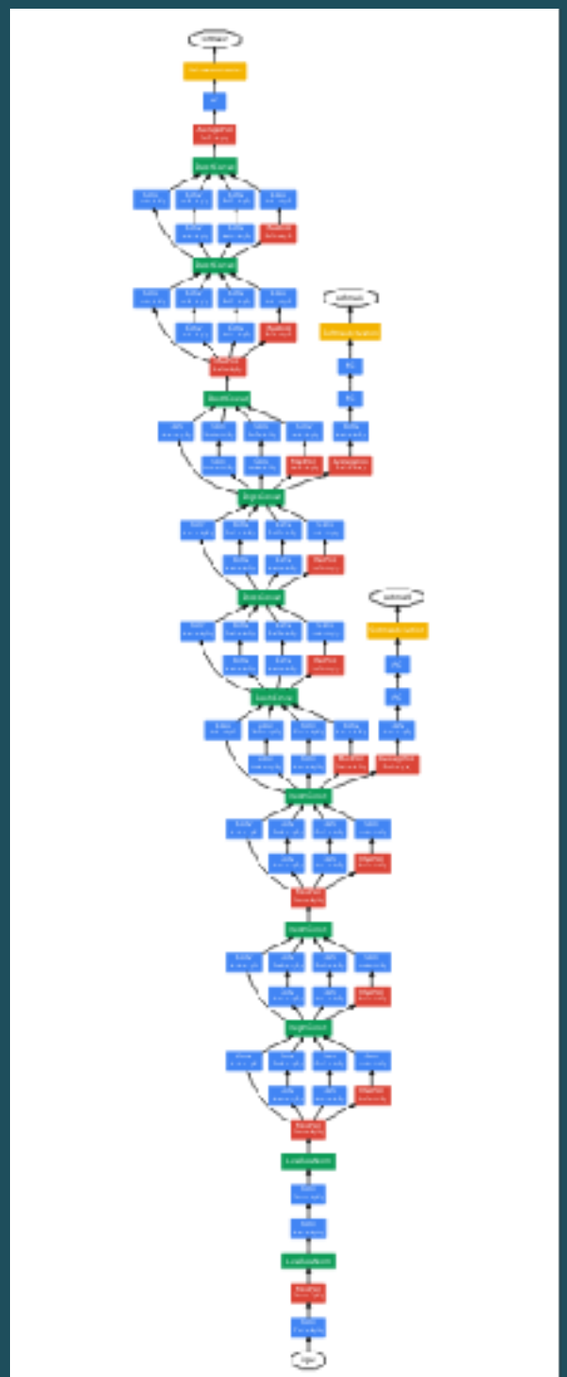
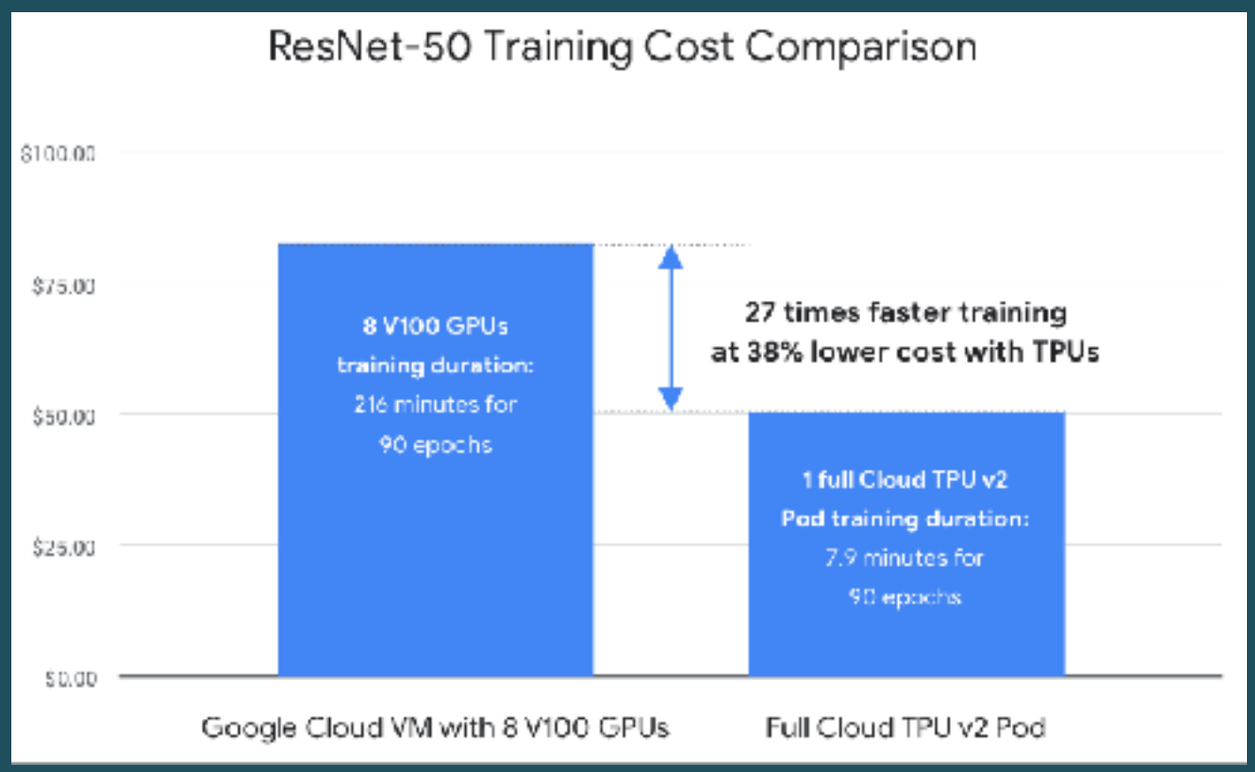


Figure 3: GoogLeNet network with all the bells and whistles

Inception

Compute power is getting cheaper

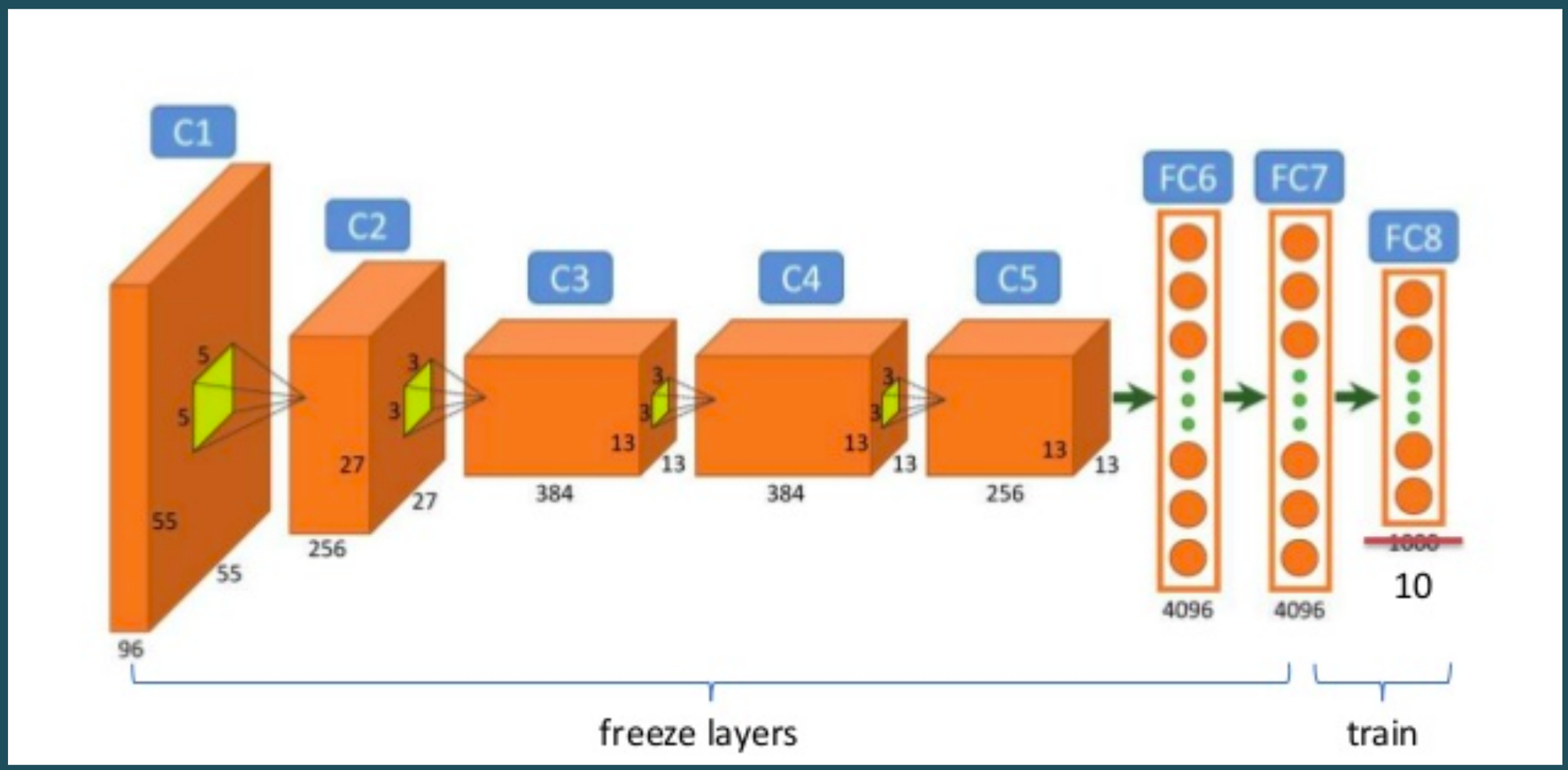
- GPU- faster matrix computation
- Tensor Processing Unit (TPU)- Integrated Circuit developed by Google for faster training of machine learning/ deep learning models



Building Classifiers for Every Object and Domain

- Creating label sets for training is expensive manual effort
  - 1M types of objects
- Building new models per domain and managing them is infeasible
- Can we reuse and leverage existing effort to create new models?

Transfer learning



Transfer learning- approach

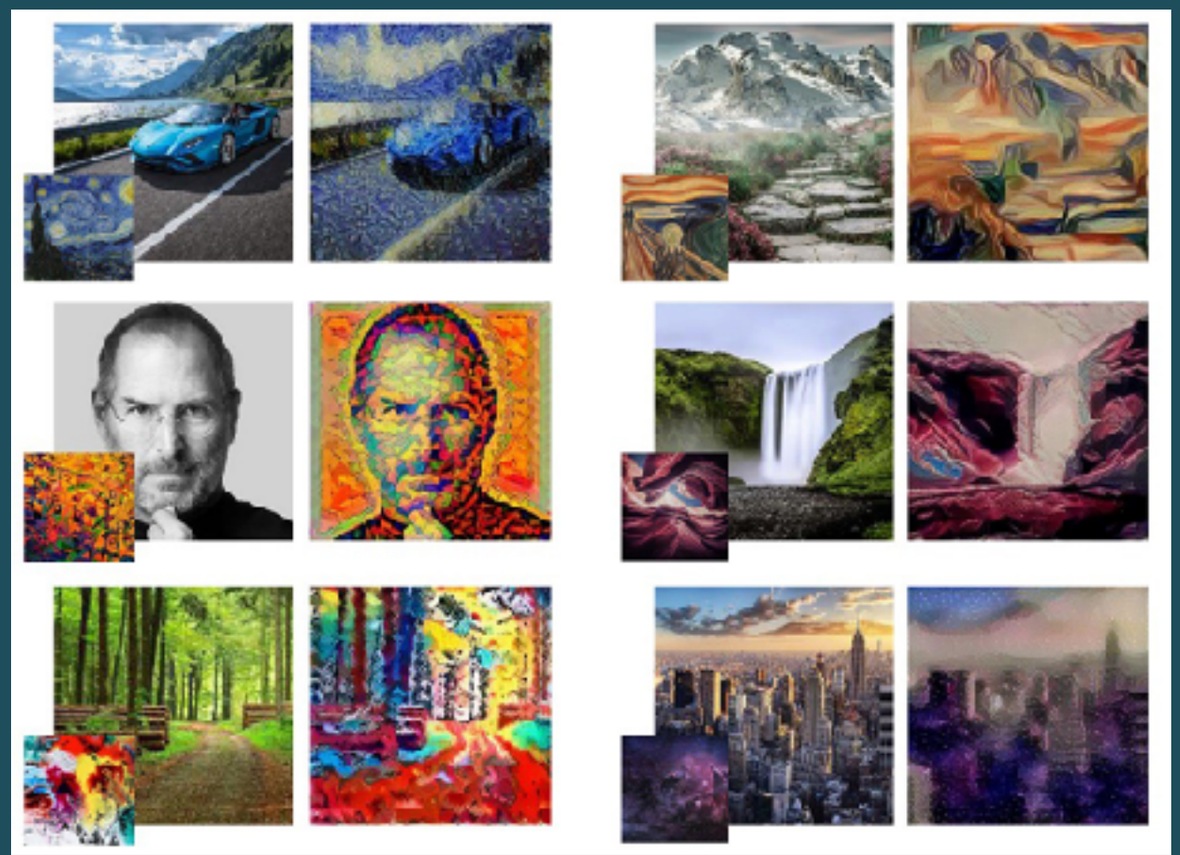
- Bring your dataset of images and classes
- Use any pretrained model- Resnet50, Inception, VGGNet etc.
- Feed the images through the pretrained model and collect the transformations from penultimate layer
- Use them as features and train a classifier
- You could also unfreeze the hidden layers of pretrained model and train the entire model- expensive computation

New challenges in Image Processing

- Learning with less data
- Generating data (GAN)
- Yolo
- 3D reconstruction from 2D (improving maps)

Generating data

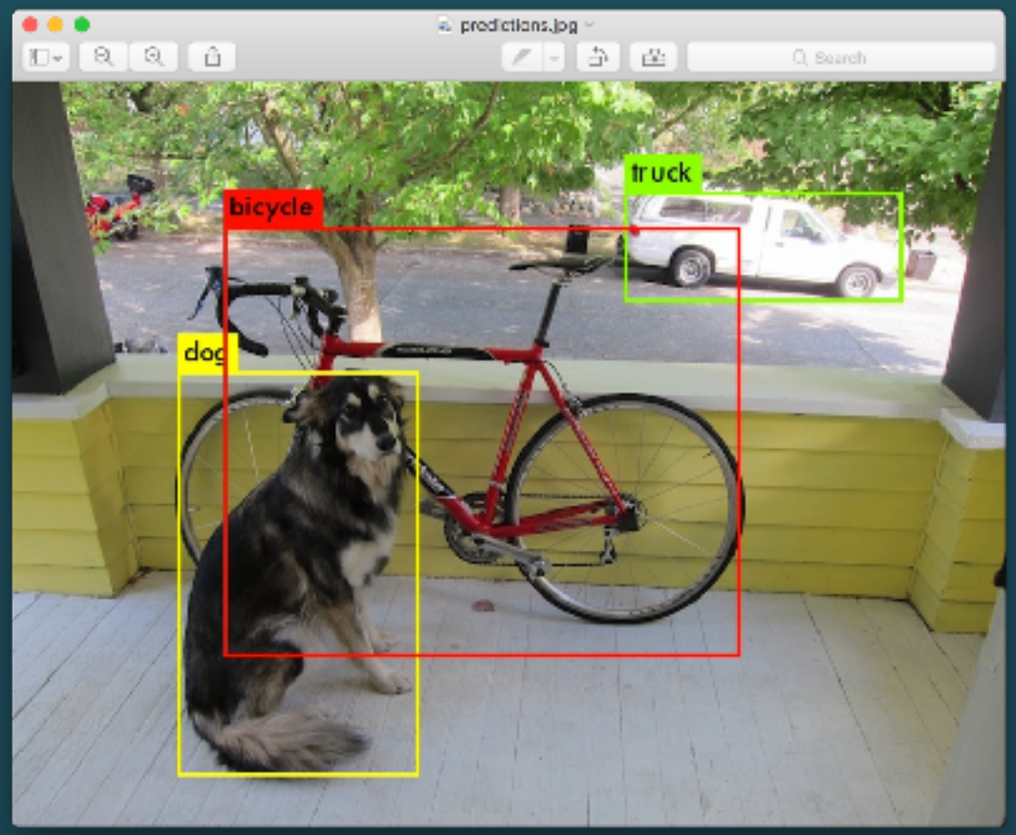
Neural style transfer

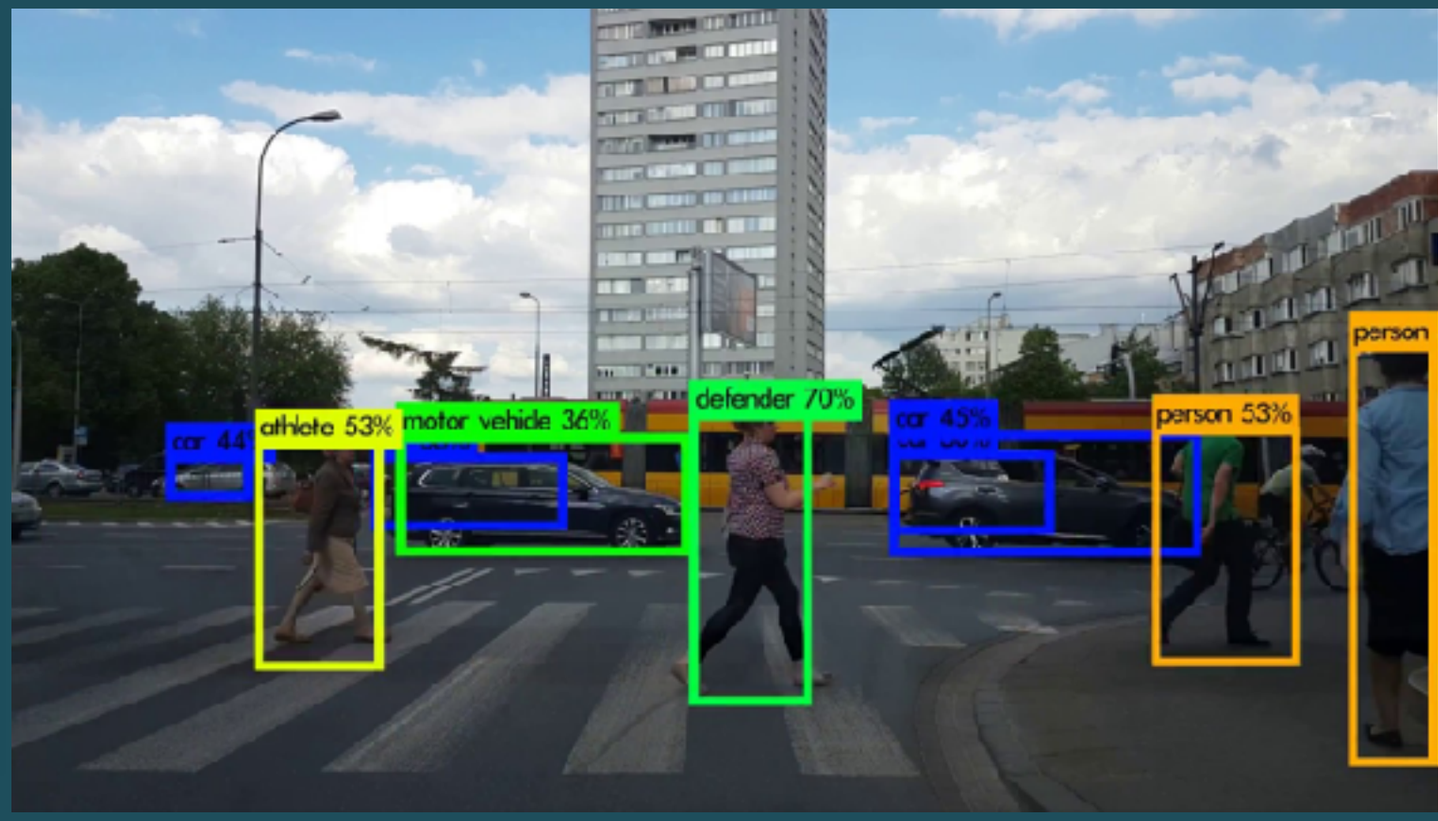


Generated images from images of celebrities using Generative Adversarial Networks (GAN)

Object detection

- You Only Look Once (YOLO)





Detecting objects from multiple angles

- Capsule networks

Use case: Google Maps

- Extract information from Street View images to update locations and landmarks

Avenue des Sapins

Enhancing Google Maps 3D vs. 2D

- Seamless street view panoramas, image stitching

Optical Character Recognition (OCR)

- Extract handwritten, printed or typed text present in images of documents, scans or scenes


42 HISTORIA NATURALIS

mollis ac friabilis carnis esse, nullum videtur dubium: quamvis p Celsus & Philotimus aliter credidisse putentur. Insipidos esse nisi cum suis fæcibus edantur, creditum apud Græcos. Jecur in cibo sumptum ictericis prodesse, apud q Ælianum legimus.

p Celsus.

q Ælian. H.A.l.14. c.2.

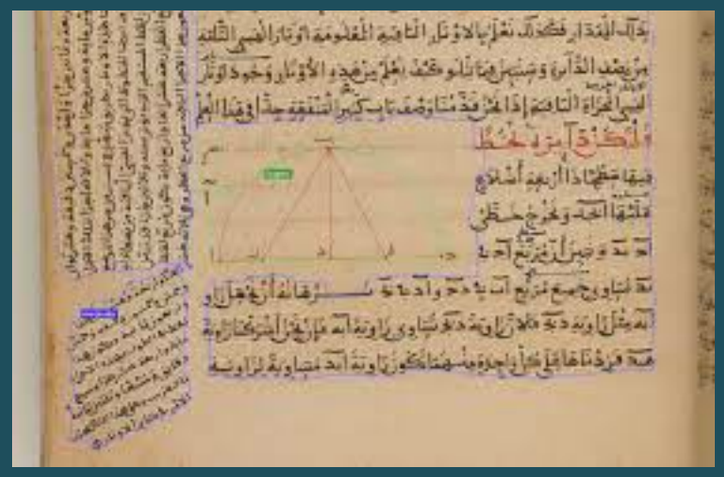
ARTICVLVS II.

De Turdo.

TAB. XIII. 5.6.7.8.9. 10.11.

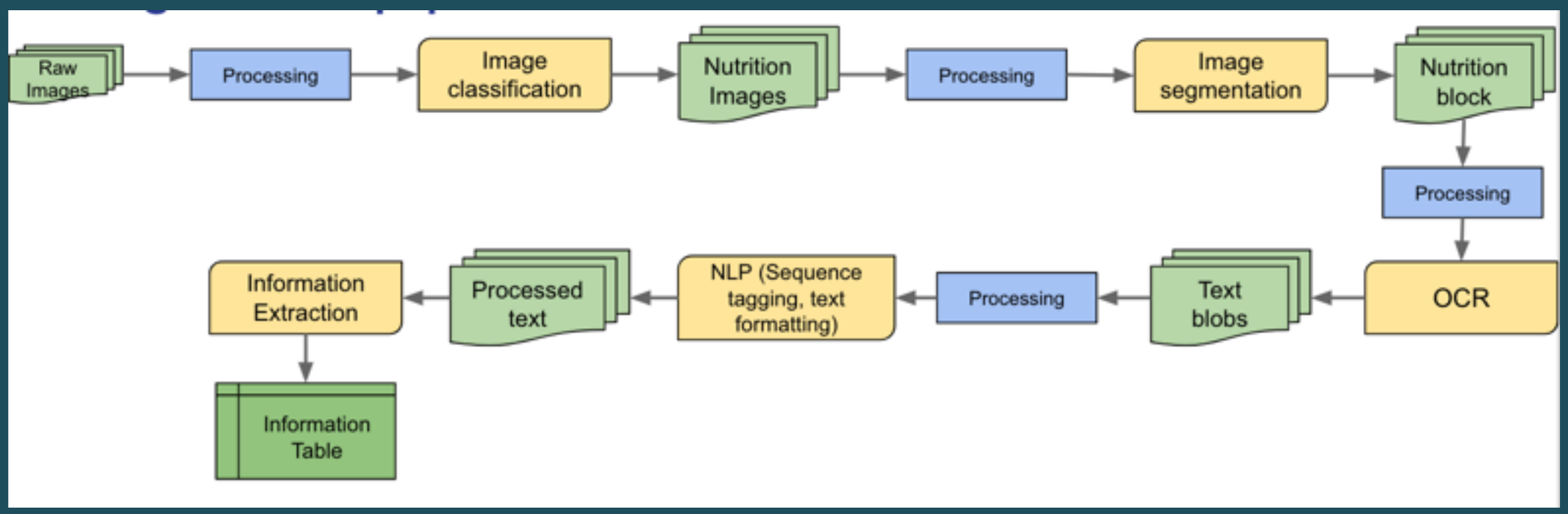
TUrdus Græcis κίχλη à maculatum similitudine sic dictus, μακρόπους olim teste Nicandro & Pancrate apud a Athenæum fuit. Athen.l.7 Duo ipsorum *Genera* constituit Sal-

cundus ad latera nonnihil capite vermicularibus, a dis, ubique cæruleis mac tur. Tergoris pinna ex ber, sicuti & cauda, qi mo cærulea est. Utrin tus est. *Tertio* Corpus sul ter albicat. Squammaru rates cæruleæ sunt. *Ca* poris proportione parv cæruleï sed dilutioris co ter branchiarum operc dicant, maculisque obl micircularibus aureis c tur. Eiusdem coloris l fatim & transversaliter ducuntur. Dentes habe acutissimos, ingentem



Retail Case Study

- Extract information from packaged good labels, and make the products searchable



3. VIDEO UNDERSTANDING

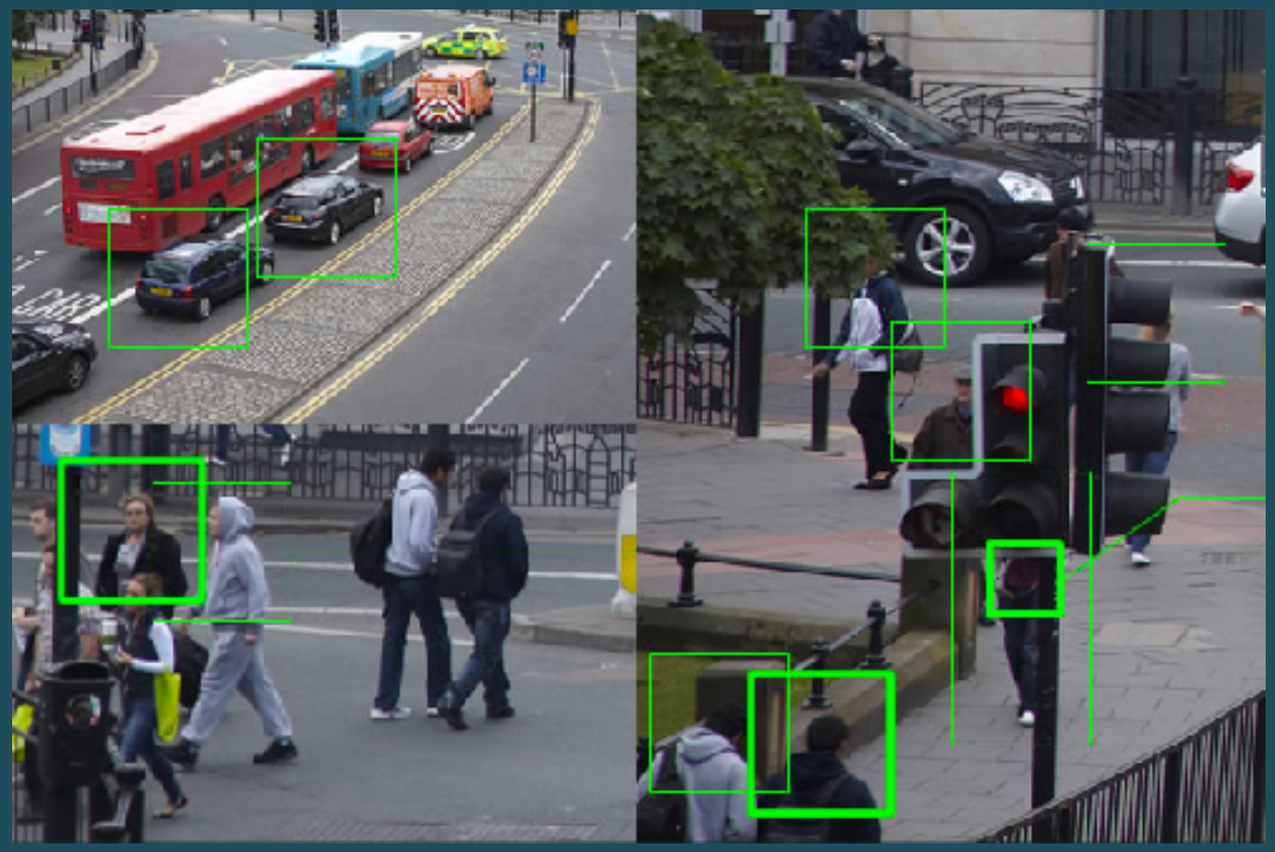
(CREDIT: STANFORD LECTURES)

ROBOTS

SELF DRIVING CARS

GAME ANALYSIS

SECURITY



VIDEO UNDERSTANDING TASKS

- Video tagging (analysis)
- Video editing (improvements)
- Video generation (gaming/VR)
- Video QA (teaching and human task automation)
- Video Retrieval

DATASETS

- UCF101 (https://www.crcv.ucf.edu/data/UCF101.php)
  - 13320 videos from youtube
  - 101 actions
- Sports 1M
  - 1,133,157 videos
  - 487 sports labels

DATASETS

- Youtube 8M
  - https://research.google.com/youtube8m/

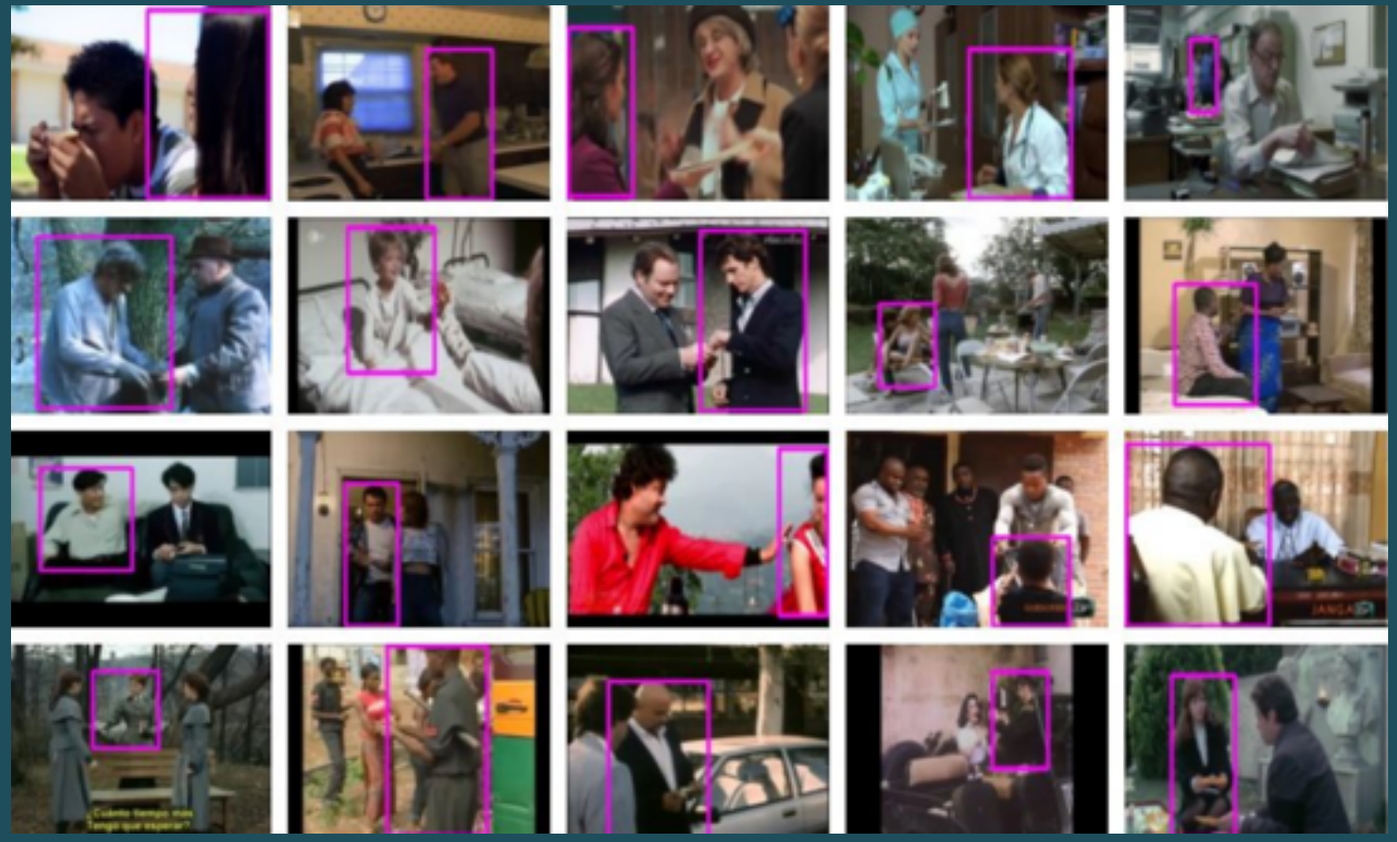
237K	1000	5.0
Human-verified Segment Labels	Classes	Avg. Segments / Video

Entities

Games (788288)
Video game (539945)
Vehicle (415890)
Concert (378135)
Musician (286532)
Cartoon (236948)
Performance art (203343)
Car (200813)
Dance (181579)
Guitar (156226)
String instrument (144667)
Food (135357)
Football (130835)
Musical ensemble (125668)
Music video (116098)
Animal (107788)
Animation (98140)
Motorsport (93443)
Pet (90779)
Racing (84258)
Recipe (75819)

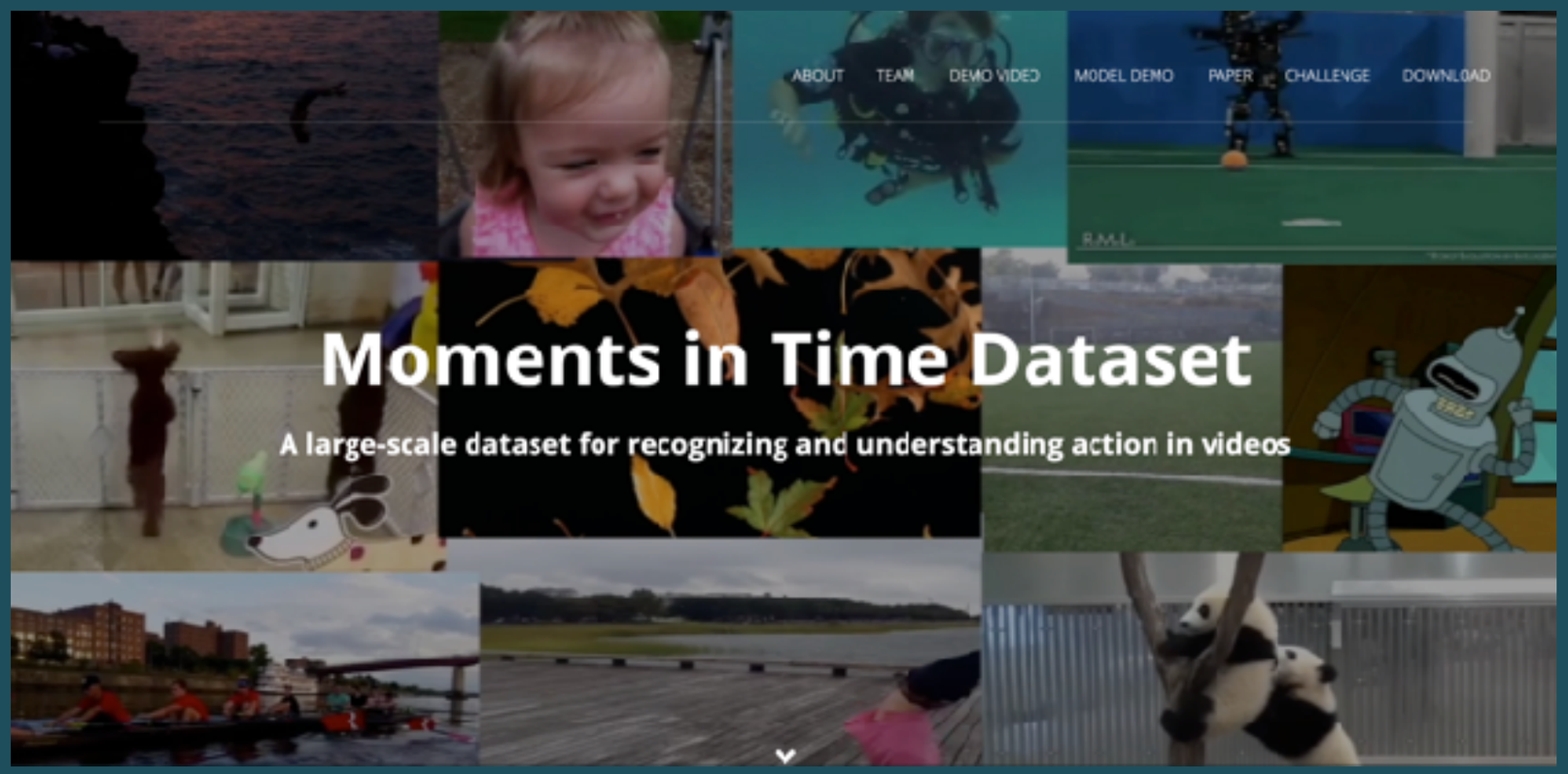
APPROACH 1: VIDEO FRAME ANALYSIS

- Atomic visual actions
  - Pose and objects
  - Does not describe interaction with objects



APPROACH 2: VIDEO SEGMENT ANALYSIS

- Video classification (not detection)
  - 339 verbs were tagged (including humans and other actions)
- Moments in Time dataset (http://moments.csail.mit.edu/)

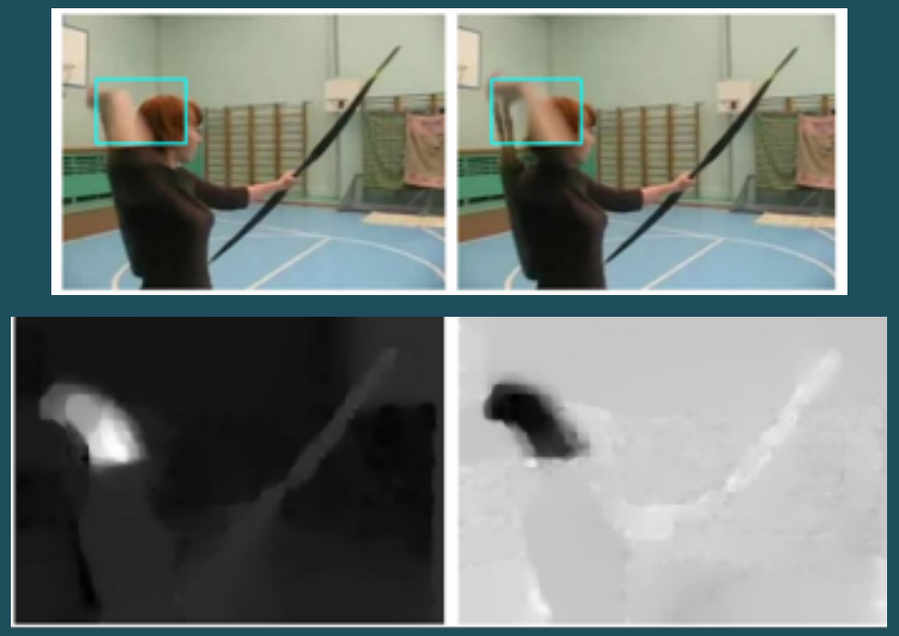


VIDEO ANALYTICS FRAMEWORK

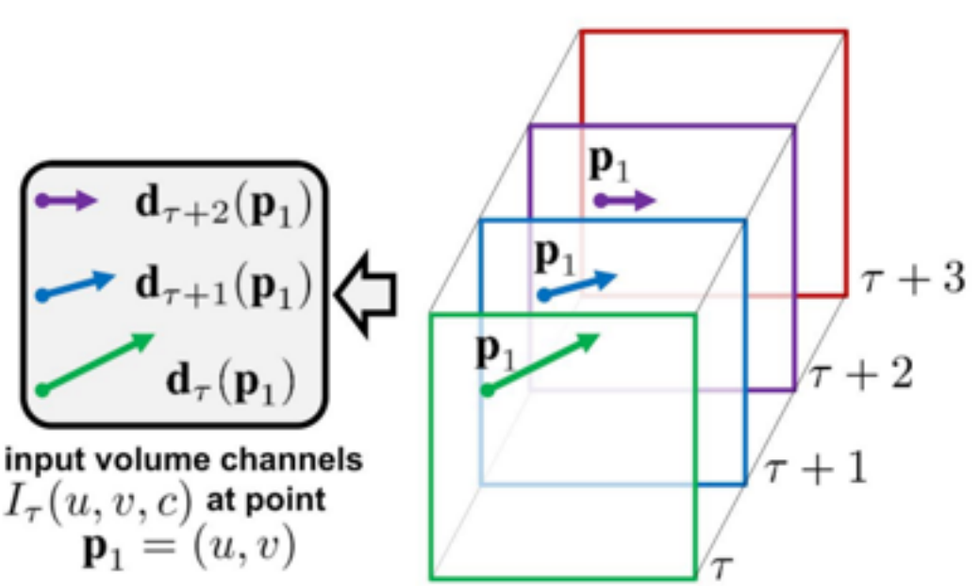
- Labeled Datasets
- Sequence modeling modules
- Temporal reasoning modules
- Action recognition algorithm
  - Representative action
  - Miscellaneous actions

TRADITIONAL MACHINE LEARNING APPROACH

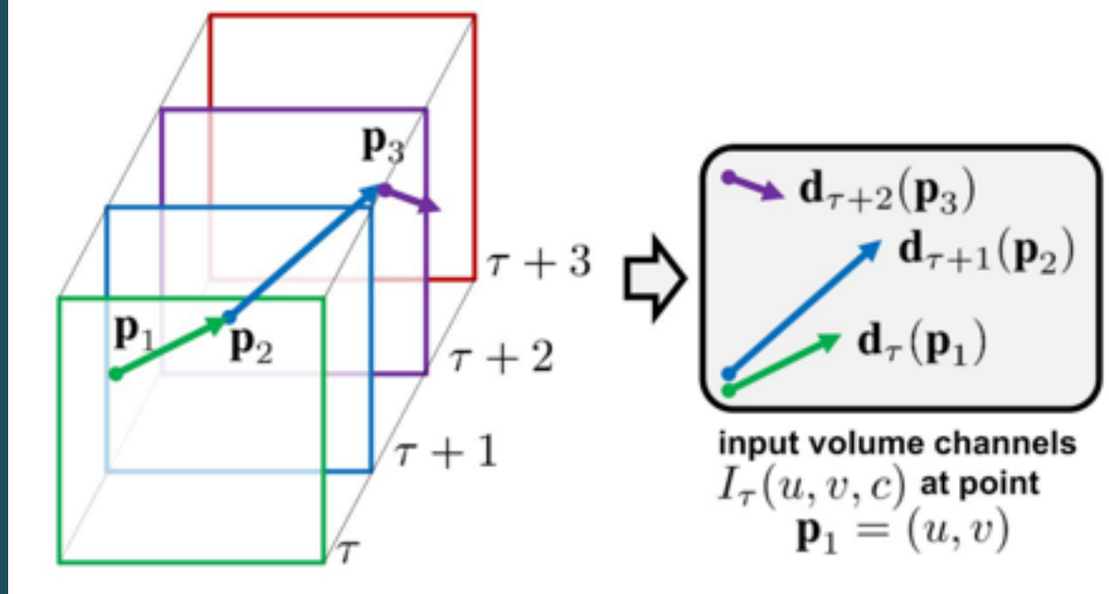
- What is the flow of action?
  - Optical flow algorithm
  - Trajectory flow algorithms



1) Optical flow

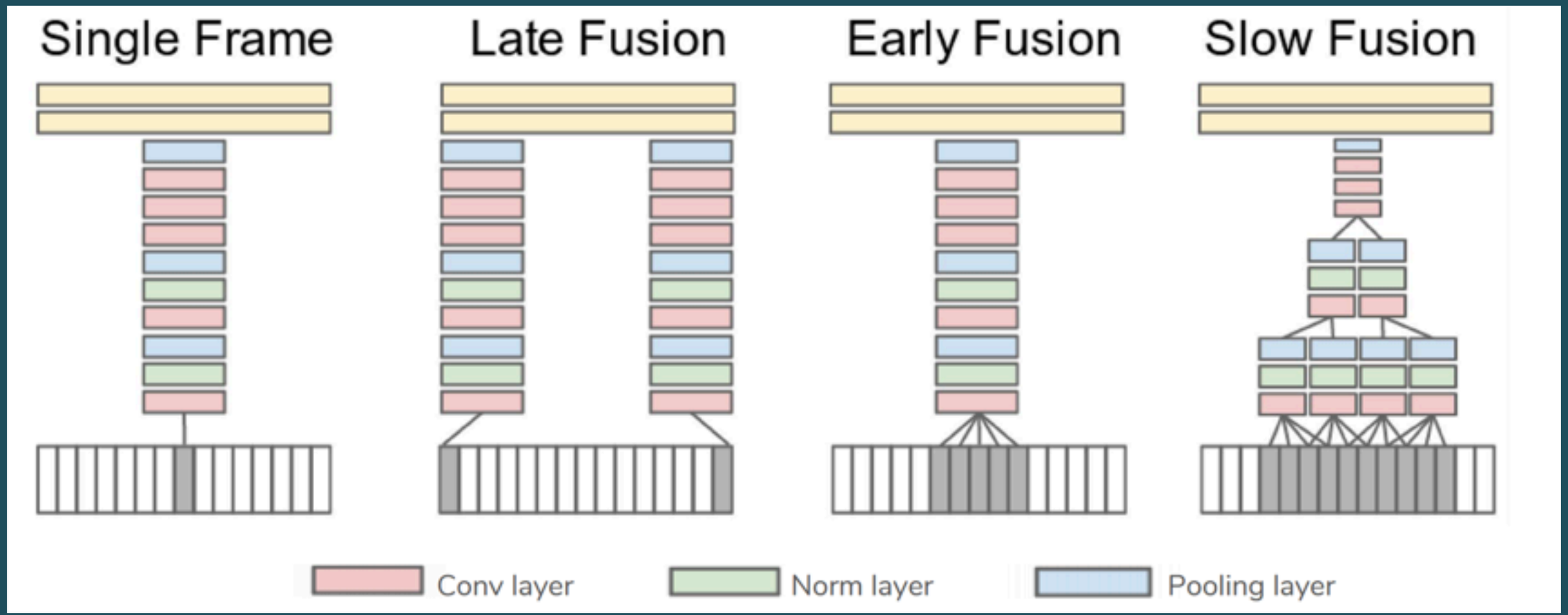


2) Trajectory stacking



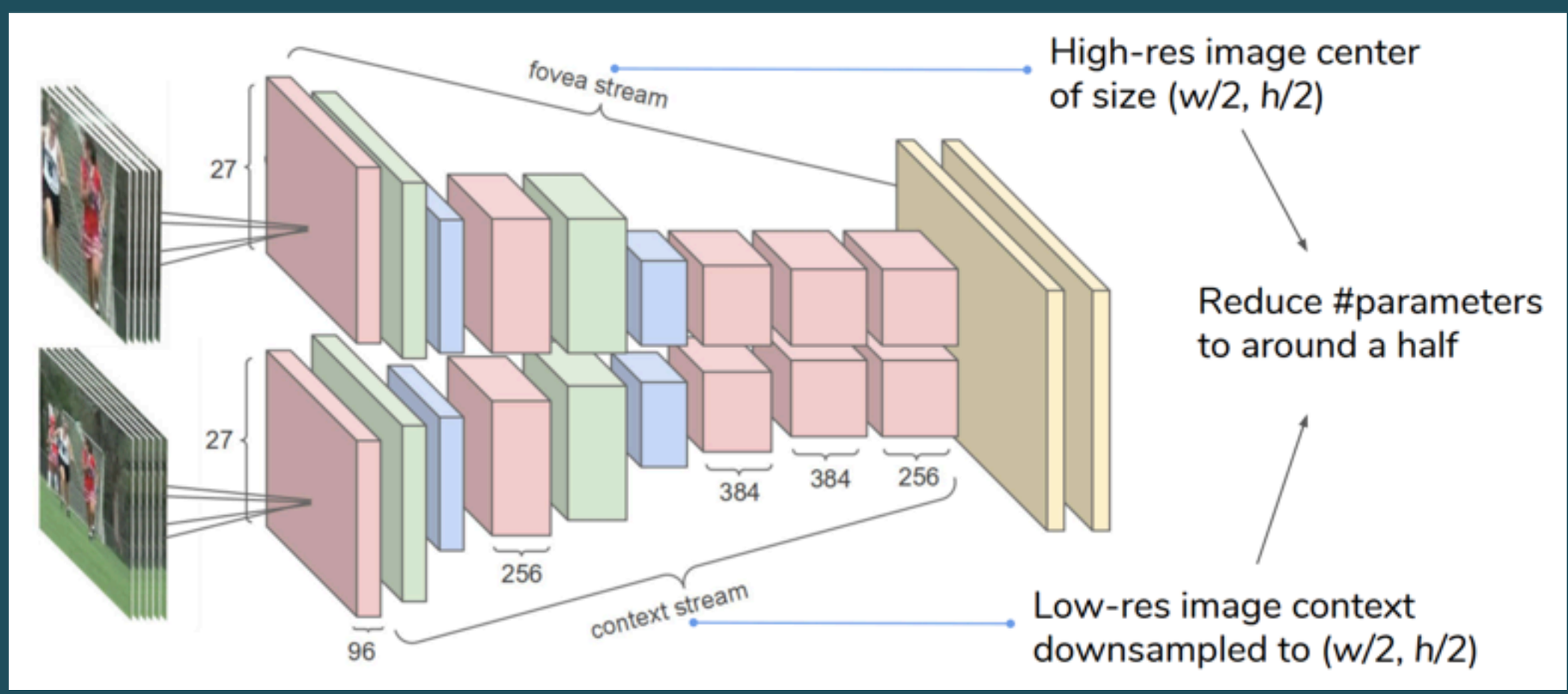
DEEP LEARNING APPROACH

- Convolution Neural Networks
  - Architecture to capture temporal aspects (combining frames)



DEEP LEARNING APPROACH

- Convolution Neural Networks
  - Reduce spatial dimension to reduce computational cost



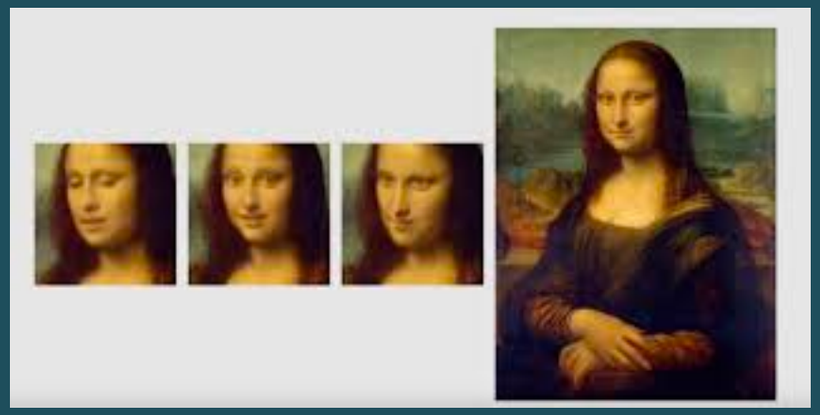
EVALUATING VIDEO RECOGNITION TASK

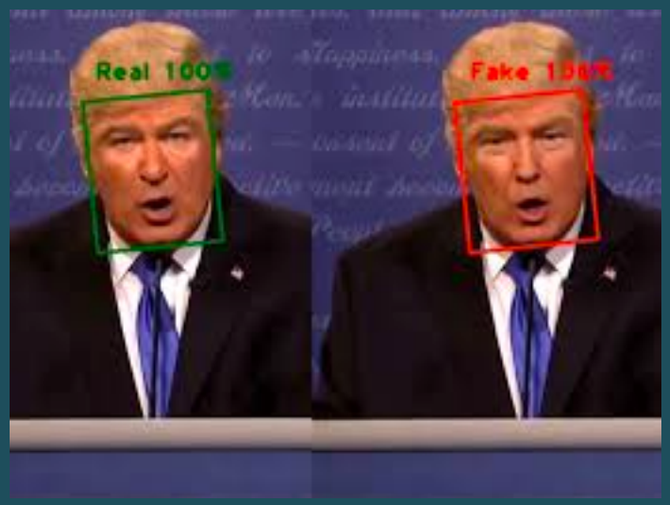
- Video Classification
  - Accuracy metrics
- Video querying
  - Provide a verb/task query
  - Retrieve top videos
  - Measure HIT ratios

CHALLENGES

- New architectures to handle long video segments (E.g summarizing highlights of a cricket game)
- Techniques and Hardware improvements to reduce cost of computation
- Labeling datasets through crowdsourcing and semi-supervised learning
- 3D video analysis with spatial interactions between all objects in the frame
- Separating the motion part of video from the static part of video to improve analysis (two-stream)

SCARY REALITY AHEAD

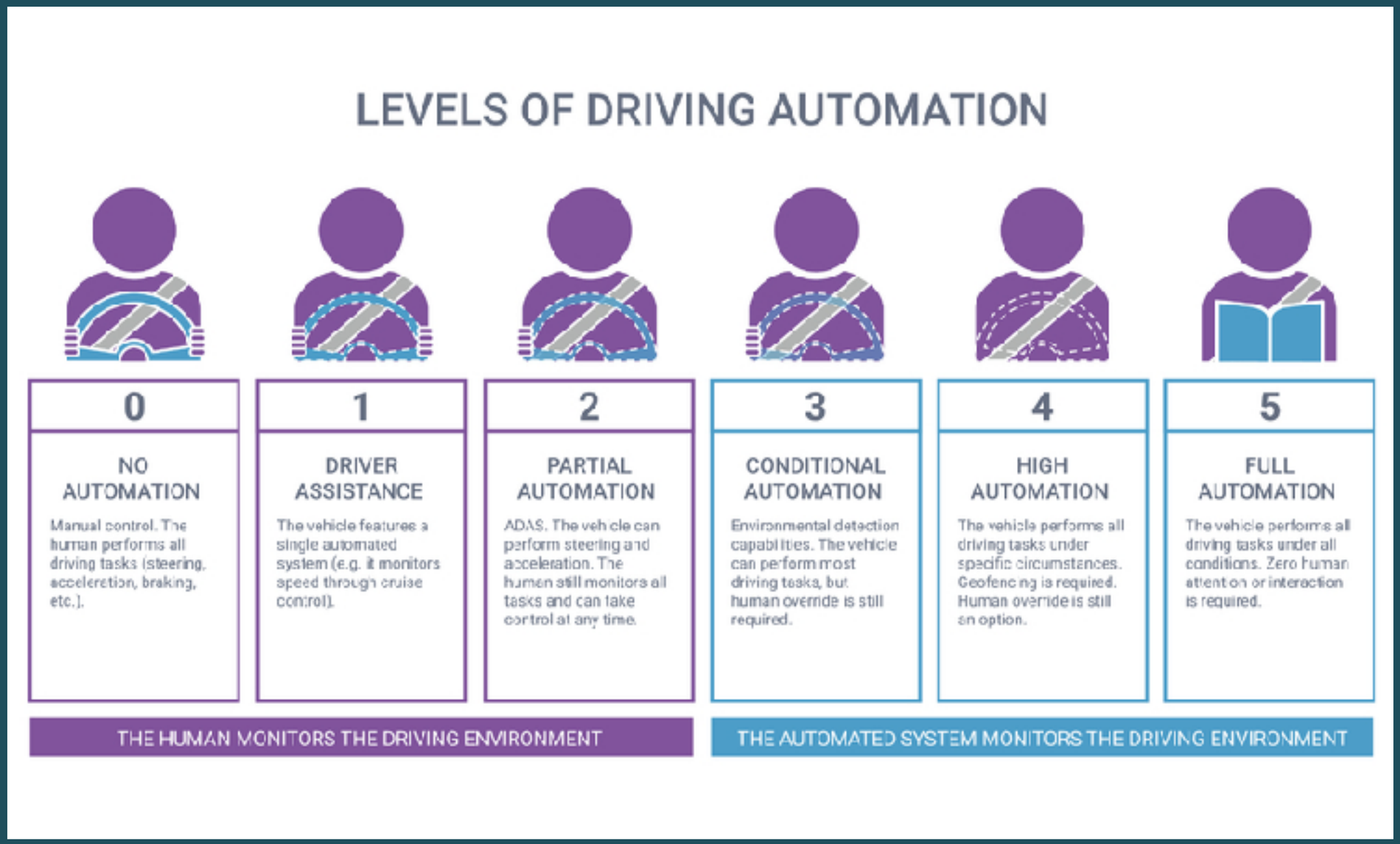




Self driving cars

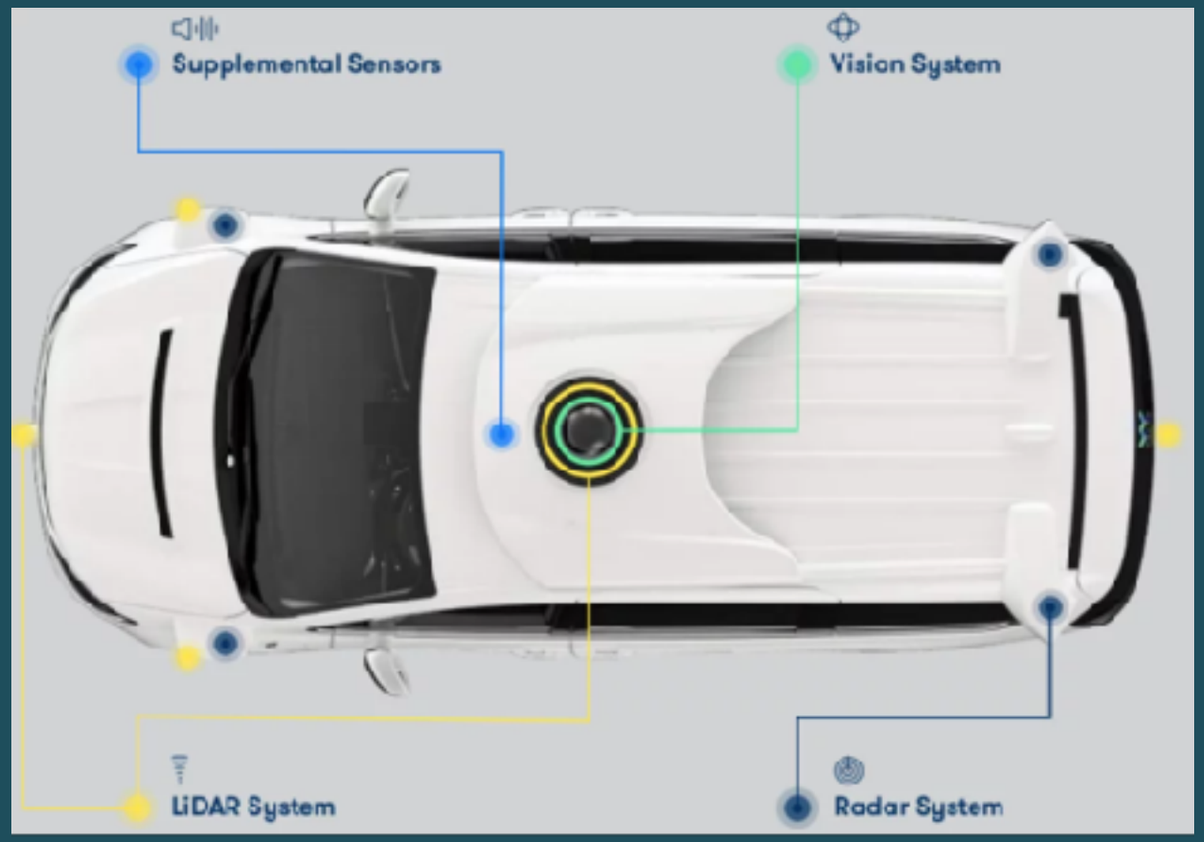
- Defense Advanced Research Projects Agency (DARPA)- 2004 challenge to build all terrain self driving cars
- Industry
  - Waymo
  - Tesla autopilot
  - Uber
- Modules
  - Perception
  - Localization
  - Planning
  - Control

Use case- Self driving cars

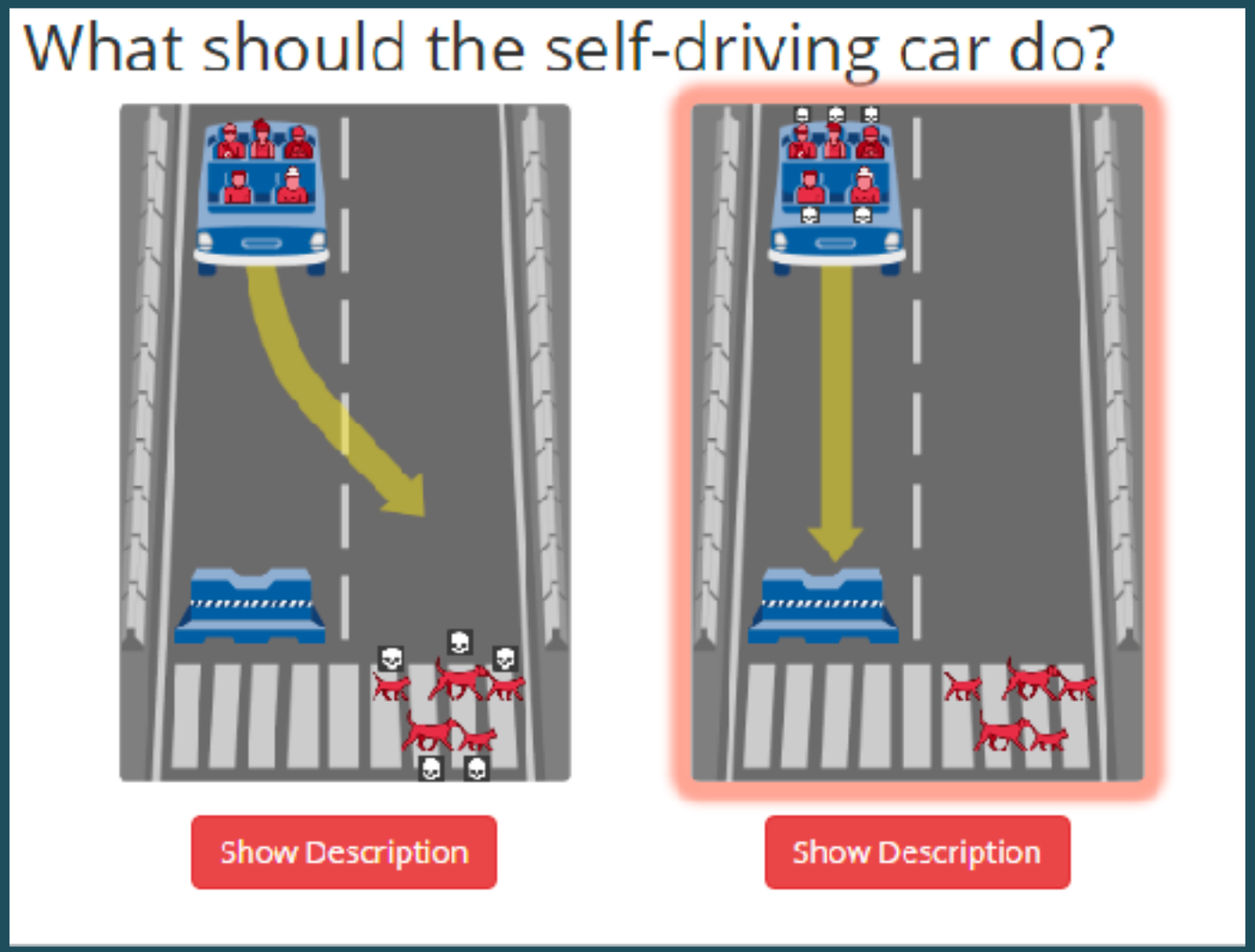


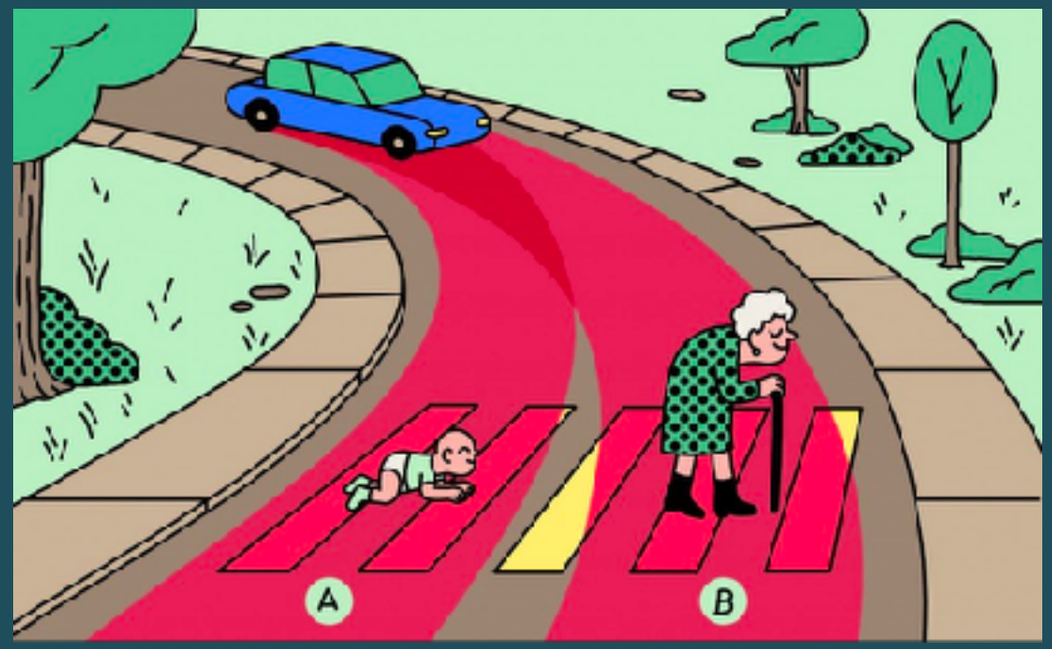
Waymo

- Trained on 5 billion miles of simulated driving and 5 million miles of on-road driving experience
- Components
  - Video sensor
  - Traffic signal detection
  - Lane tracking
  - Radar



Ethics





- https://www.reddit.com/r/MachineLearning/comments/f8wsyg/nd_yolo_creator_joseph_redmon_stopped_cv_research/
- https://en.wikipedia.org/wiki/List_of_self-driving_car_fatalities
- Privacy

FROM HERE

- Open source libraries
  - OpenCV
  - Tensorflow
  - FastAI
- Image Datasets
  - MNIST Dataset
  - Youtube Video Dataset
  - Amazon dataset
- Video Datasets
  - Youtube Dataset
  - AVA dataset
  - Sports 1M

THANKS