Computer Vision 1

THEO GEVERS, SHAODI YOU, THOMAS MENSINK AND PASCAL METTES

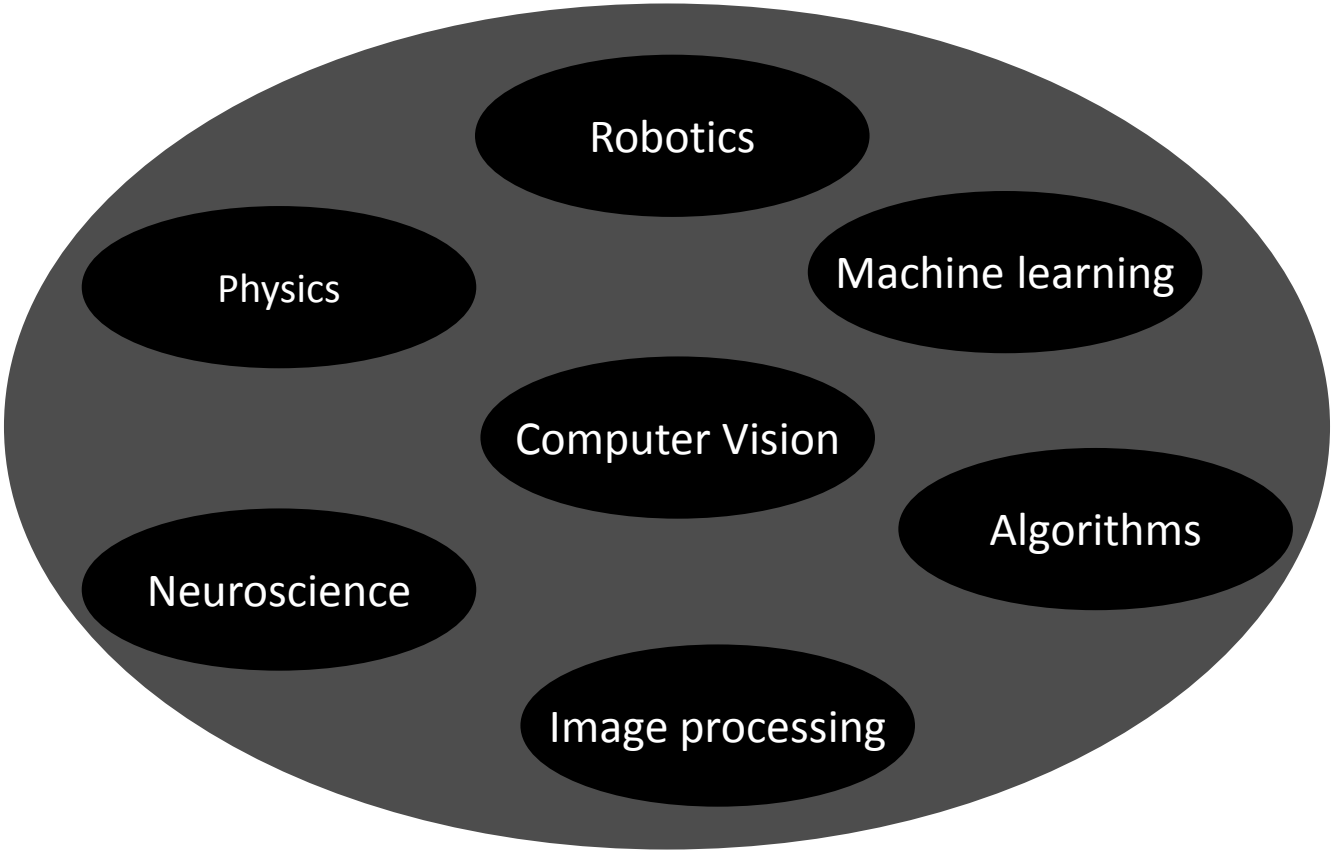
MASTER AI

UNIVERSITY OF AMSTERDAM

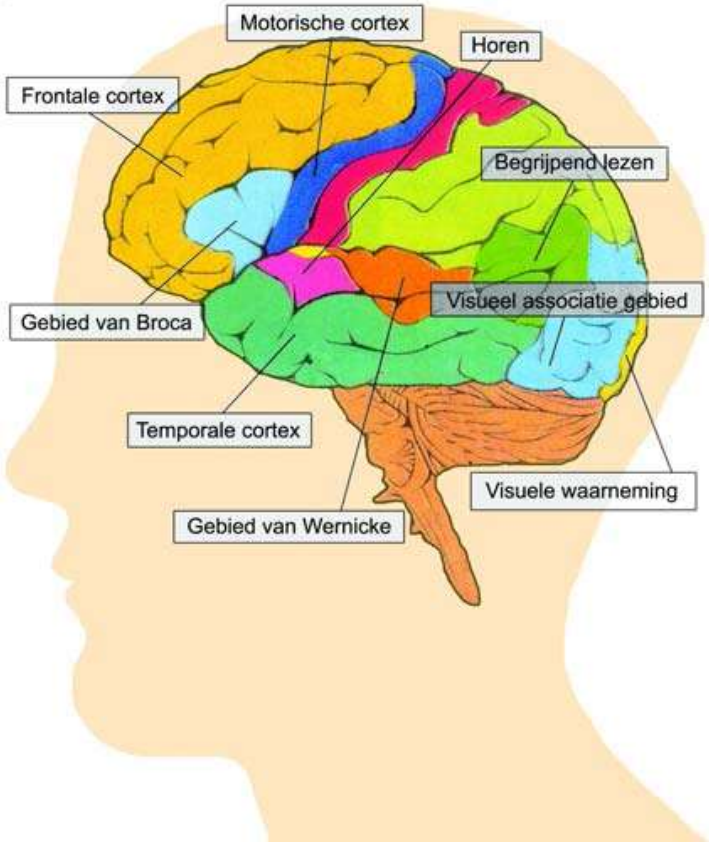
What is Computer Vision?

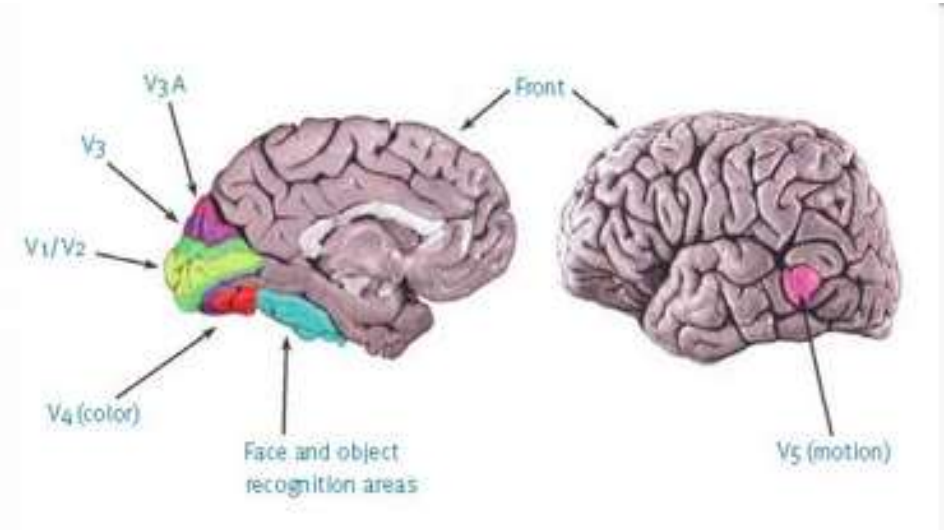
- Automatic understanding of images and video
  - Computing properties of the 3D world from visual data (**measurement**)
    E.g., Image processing, Physical modeling, Stereo, 3D reconstruction
  - Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. (**perception and interpretation**)
    E.g., Segmentation, Image Classification, Object Detection,
  - And of course many cross-overs

What is it related to?



Visual Cortex





Average number of images seen by a person:

3 images/second * 60 * 60 * 16 * 365 * 60

3.7 billion images

Images

YouTube has more than 1 billion videos

Google has more than 1.5 billion images

Facebook: more than 400 million image uploads per day

Instagram: more than 150 million image uploads per day

Course Organization

(more details on **canvas**)

This is a 6 credit course in 7 weeks. The course has lectures of basic **theory** based on the freely downloadable books http://szeliski.org/Book/ and http://www.deeplearningbook.org/, and related papers (all readings provided at *Canvas*). **Seminars** are given to practice the basic theory. Further, the course contains hands-on experience in the **practical** lab sessions.

Prerequisites

- Linear algebra, calculus, and probability
- Machine learning
- Matlab, data structures

Course Organization

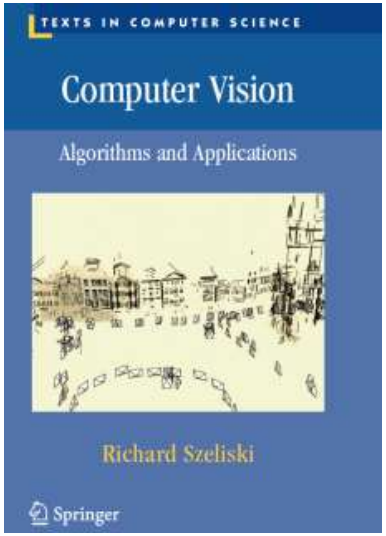
- Grades run from 1 to 10 (highest). Each part of the course should be scored **at least a 5.5**. The final grade is based on the following weighted parts:
- Exam: 50%
- Practical: 50%

Course Organization

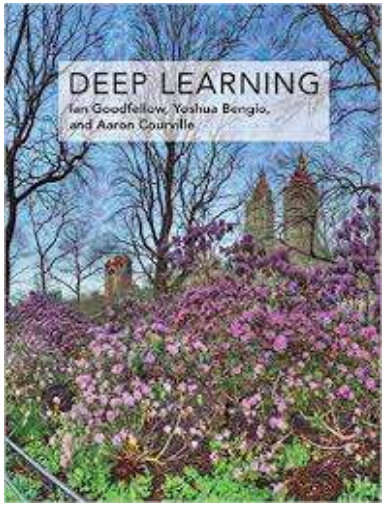
- **Exam, 50%**
- The questions will cover the basic theory on computer vision. The slides and exercises (seminars) will form the basis for the (closed book) exam. Read the book and papers for background knowledge but the slides and exercises are most important.
- **Focus on chapters**:
  - **Szeliski:** 1 + 2.1.1 + 2.1.2 + 2.1.5 + 2.2 + 2.3.2 + 2.3.3 + 3.1 + 3.2 + 3.3 + 4 + 8.4 + 14 (http://szeliski.org/Book/)
  - **Bengio:** 5.7 + 6 + 7.4 + 7.7 + 7.9 + 9 + 10.1 + 10.2 (intro, .2, .3) + 10.3 + 10.4 + 10.5 + 10.7 + 10.10 + 14.1 + 14.2 + 14.3 + 14.6 + 15.1 + 15.2 (http://www.deeplearningbook.org/)

Textbooks

- CV1 is based on "*Computer Vision: Algorithms and Applications*" by Richard Szeliski
  - Freely available for download from http://szeliski.org/Book/



- and "*Deep Learning*" by Ian Goodfellow, Yoshua Bengio and Aaron Courville
  - Freely available for download from http://www.deeplearningbook.org/

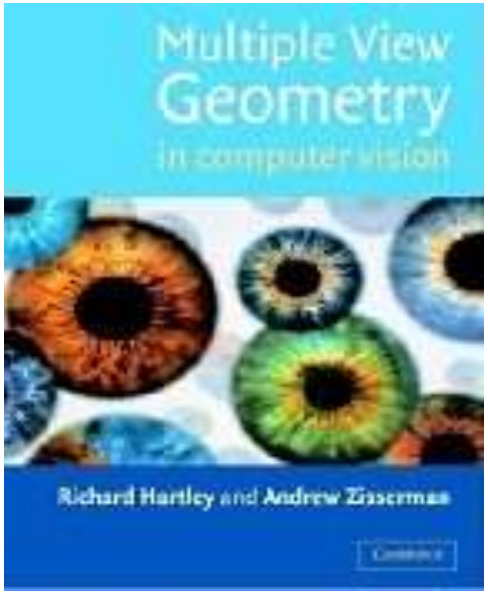


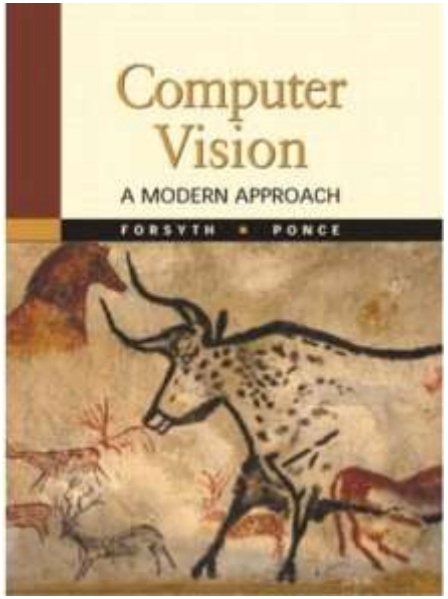
Background

- Two other useful books

  Forsyth, David A., and Ponce, J. Computer Vision: A Modern Approach, Prentice Hall, 2003.

  Hartley, R. and Zisserman, A. Multiple View Geometry in Computer Vision, Academic Press, 2002.



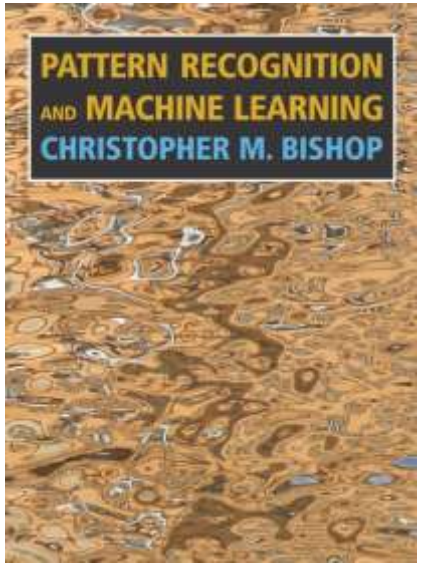


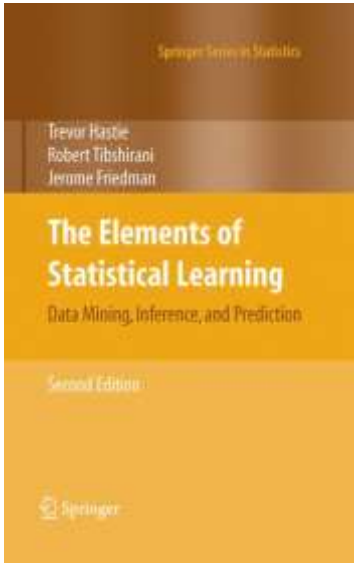
Even More Background and Tools

- Machine Learning books:

  Bishop, C. Pattern Recognition and Machine Learning, 2006.

  Hastie, T., Tribshirani R., and Friedman J, The Elements of Statistical Learning, Freely available for download: http://www-stat.stanford.edu/~tibs/ElemStatLearn/





- Software, programming tools

  Matlab, image processing toolbox,
  C++, openCV
  Python, PIL, numpy, scipy
  PRTools, Shogun, (machine learning toolboxes)

- Scientific papers:

  Top Journals: Transactions of Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, TIP, (Computer Vision and Image Understanding)
  Top Conferences: ICCV, ECCV, CVPR, NIPS, SIGGRAPH, BMVC, (ACM-Multimedia)
  http://www.cvpapers.com/

Tools and Tutorials

Tools

- VLFeat, open source implementations of Computer Vision algorithms Link
- OpenCV, open source Computer Vision framework Link
- Theano, Python Math Expression Library (Neural Network Optimization) Link
- Caffe, Neural Network Framework Link

Tutorials

- Neural Network Tutorial Link
- Matlab Tutorials
  - David Griffiths' Matlab notes Link
  - UCSD Computer Vision course Matlab introduction Link
- Camera Model Visualization Link

Lectures/Theory

- 02-09-2019, 17:00-19:00, H0.08, **Introduction** (*Szeliski 1*) – **Theo Gevers**
- 09-09-2019, 17:00-19:00, H0.08, **Image Formation** (*Szeliski: 2.1.1 + 2.1.2 + 2.1.5 + 2.2 + 2.3.2 + 2.3.3*) – **Theo Gevers/Shaodi You**
- 16-09-2019, 17:00-19:00, H0.08, **Image Processing** (*Szeliski: 3.1 + 3.2 + 3.3 + 4 + 8.4*) – **Shaodi You**
- 23-09-2019, 17:00-19:00, H0.08, **Object Recognition** (*Szeliski: 14; Bengio 5.7)* – **Shaodi You**
- 30-09-2019, 17:00-19:00, H0.08, **Deep Learning** (*Szeliski: 3.2; Bengio: 6 + 7.4 + 7.7 + 7.9 + 9*) – **Thomas Mensink**
- 07-10-2019, 17:00-19:00, H0.08, **Deep Video** (Bengio: 10.1 + 10.2 (intro, .2, .3) + 10.3 + 10.4 + 10.5 + 10.7 + 10.10) – **Pascal Mettes**
- 14-10-2019, 17:00-19:00, H0.08, **Applications** (*Szeliski: 12.6.2 + 12.6.3 + 12.2.4; Bengio: 14.1 + 14.2 + 14.3 + 14.6 + 15.1 + 15.2)*
- 25-10-2019, 13:00-16:00, **Written Exam**

Seminars/Exercises

- 10-09-2019, 13:00-15:00, C1.110
- 17-09-2019, 13:00-15:00, C1.110
- 24-09-2019, 13:00-15:00, C1.110
- 01-10-2019, 13:00-15:00, C1.110
- 18-10-2019, 13:00-15:00, C1.110
- 25-10-2019, 13:00-15:00, C1.110

Practical Assignments

- **Practical, 50%**
- You will implement 2 methods for object recognition. Each week you do a part. The idea is to combine computer vision modules to have a real working system in the end. In the first weeks you will do separate modules. Then, you have to combine the modules to get the final system and write a small report about.

Lab Session

9 TA's: Anil Baslamisli, Yunlu Chen, Partha Das, Rick Groenendijk, Jian Han, Mert Kilickaya, Hoang An Le, William Thong, Wei Zeng

- 05-09-2019, 17:00-19:00 **Introduction to MatLab (optional)**
- 12-09-2019, 17:00-19:00 **Photometric Stereo & Color**
- 19-09-2019, 17:00-19:00 **Neighborhood Processing: Gabor & Gaussian Filters**
- 26-09-2019, 17:00-19:00 **Harris Corner Detector, Optical Flow and Feature Tracking**
- 03-10-2019, 17:00-19:00 **Image Alignment and Stitching / Final Project**
- 10-10-2019, 17:00-19:00 **Final Project**
- 17-10-2019, 17:00-19:00 **Final Project**

Today's Class (Szeliski Chapter 1)

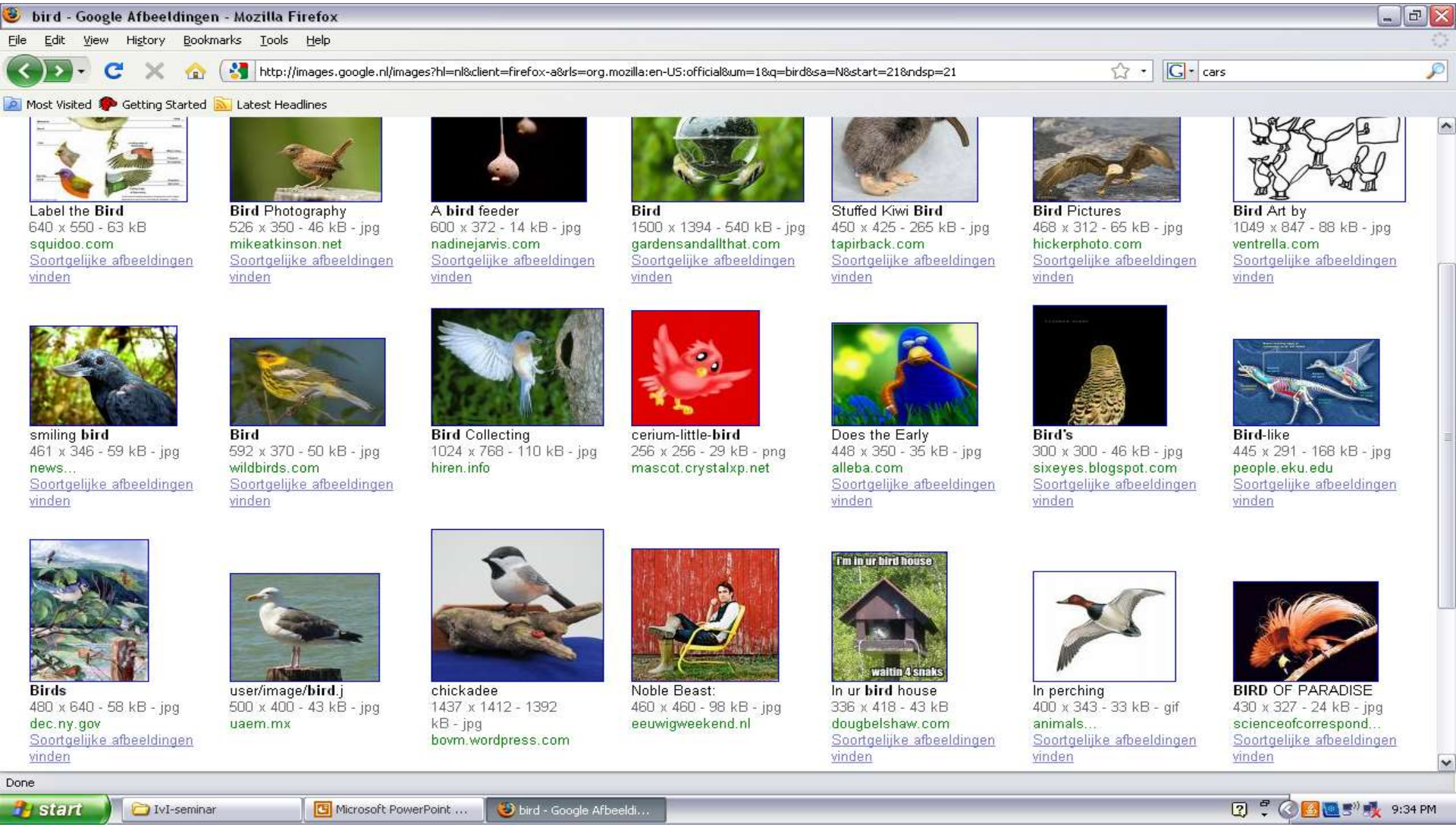
- Specifics of this course
- Introduction to Computer Vision
  - What are Computer Vision Systems
  - Computer vision and machine learning
- Image search

Image and Video Access

- Today, there are billions of images on the Internet and in collections such as FaceBook and Flickr.

- Suppose I want to find pictures of birds, humans, cars, boats or videos of explosion, violence etc

Google Image Search – Bird



Video Retrieval

Given a shot from a video…
… is some semantic *concept* present in that shot?

Example concepts:

- Airplane
- Building
- **Car**
- Crowd
- **Desert**
- **Explosion**
- **Outdoor**
- People
- **Vehicle**
- **Violence**

Object/Scene Categories


Aircraft


Animal


Boat


Building


Bus


Car


Chart


Corp. leader


Court


Crowd


Desert


Entertainment


Explosion


Face


Flag USA


Gov. leader


Map


Meeting


Military


Mountain


Natural disaster


Office


Outdoor


People


People marching


Police / security


Prisoner


Screen


Sky


Sports


Studio


Truck

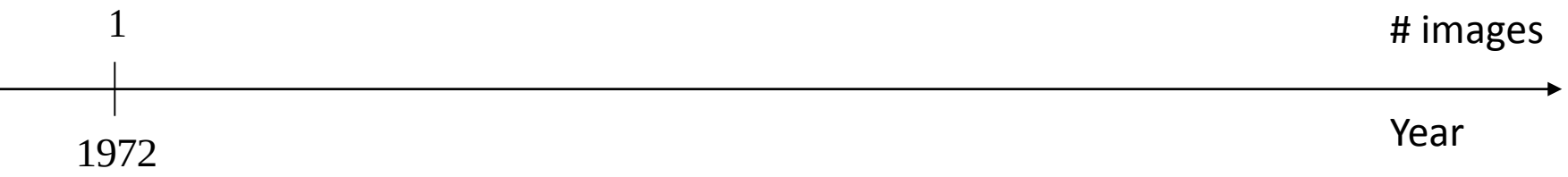

Urban


Vegetation

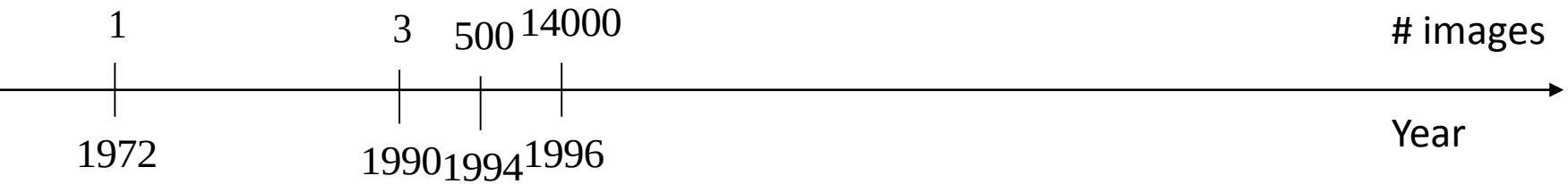

Vehicle


Violence

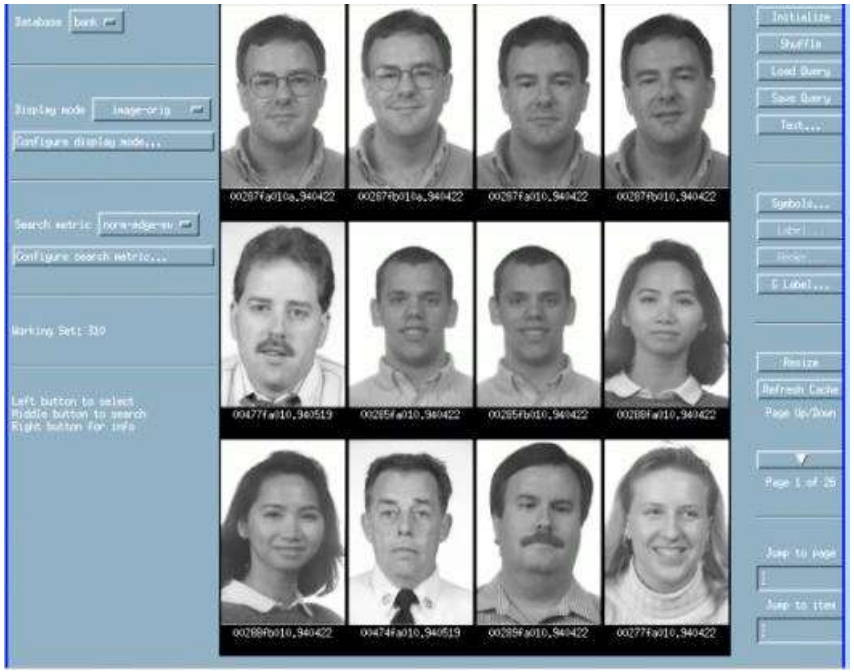
Datasets in perspective

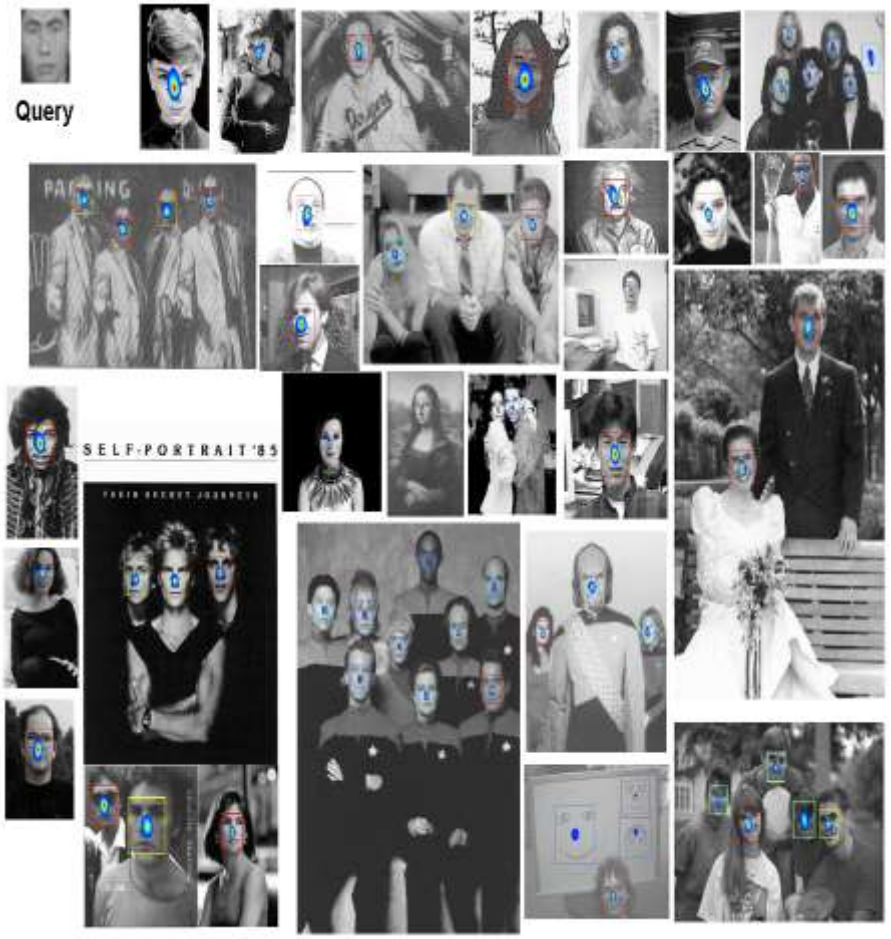


Lena (1972)

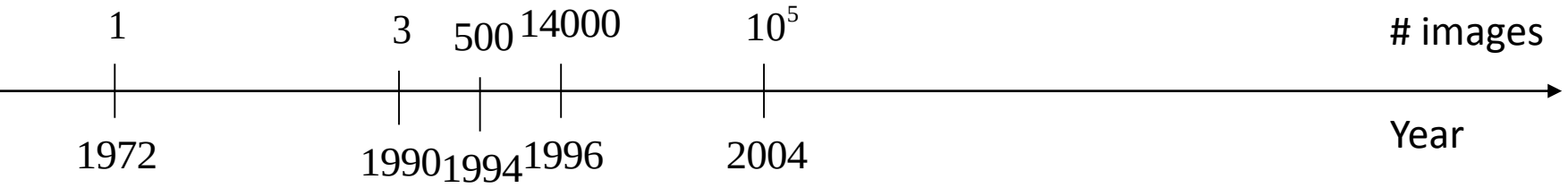


DARPA Faces (1996)





In 1996 DARPA released
14000 images,
from over 1000 individuals.

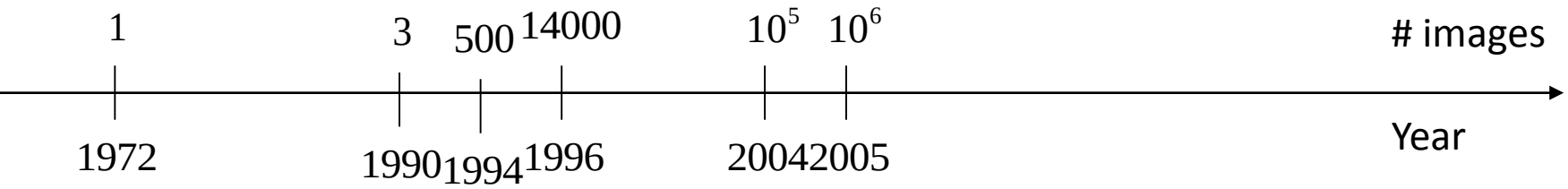


Caltech 101 and 256 (100,000)

Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007

101 categories. About 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003



TRECVID and PASCAL VOC competition (2005-2009)

- 86 hours of video from TRECVID 2005
- Shot segmentation available: 43.907 shots
- Ground truth available from Mediamill Challenge

- The goal of VOC challenge is to recognize objects from a number of visual object classes in realistic scenes
- The twenty object classes are:
  - *Person:* person
  - *Animal:* bird, cat, cow, dog, horse, sheep
  - *Vehicle:* aeroplane, bicycle, boat, bus, car, motorbike, train
  - *Indoor:* bottle, chair, dining table, potted plant, sofa, tv/monitor.

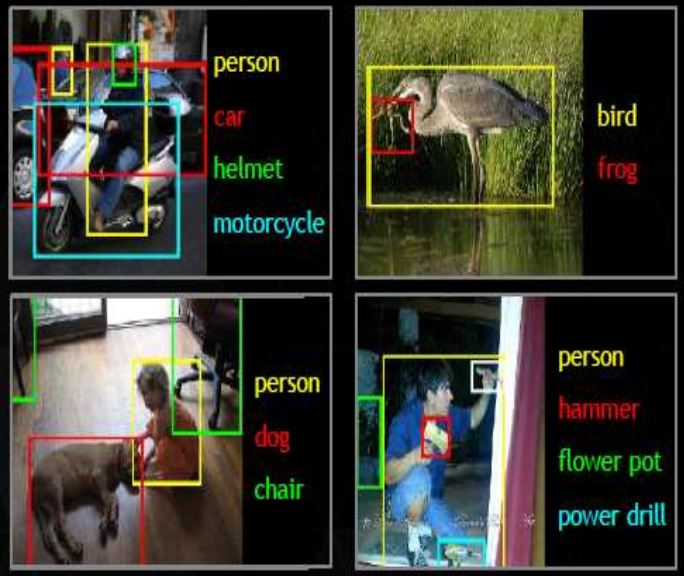
ImageNet

Image Recognition Challenge

1.2M training images • 1000 object categories

Hosted by

IMAGENET



Labeling

Labeling to get a Ph.D.

Just labeling

Labeling for fun

Labeling for money

1 cent per image

Task: Label one object in this image

Image and Video Formation

Challenges

- Viewpoint variation

- Illumination change

- Orientation and scale

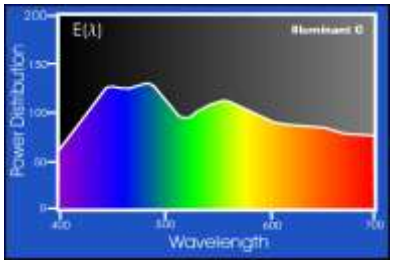
- Occlusion

- Clutter

- Appearance change

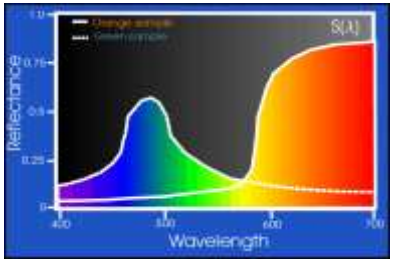
What makes an image

Light source



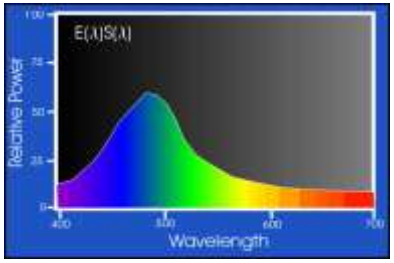
$e(\lambda)$

Object



$\rho(\lambda)$

Sensor



$e(\lambda)\rho(\lambda)$

$$R = \int_{\lambda} e(\lambda)\rho(\lambda) f_R(\lambda) d\lambda, \quad G = \int_{\lambda} e(\lambda)\rho(\lambda) f_G(\lambda) d\lambda, \quad B = \int_{\lambda} e(\lambda)\rho(\lambda) f_B(\lambda) d\lambda$$

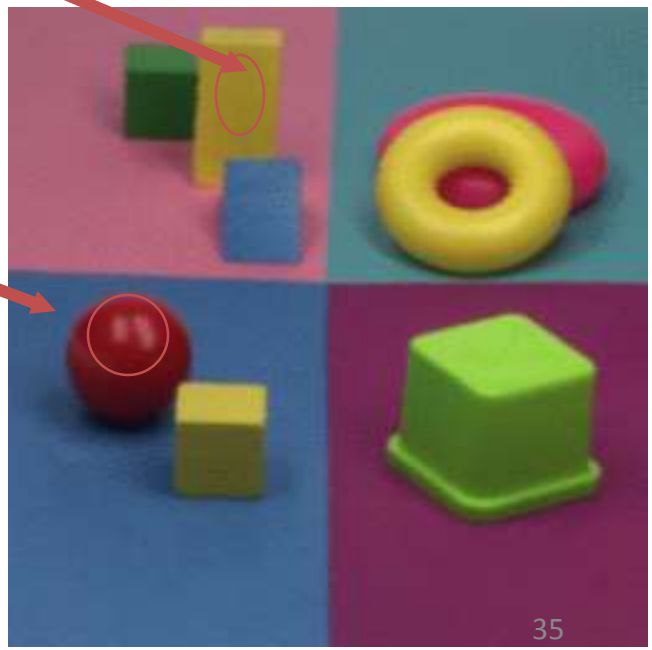
Image Formation

Reflection model:

body = $m_b(\vec{n}, \vec{s}) \int_\lambda f_C(\lambda) e(\lambda) c_b(\lambda) d\lambda \quad +$

surface = $m_s(\vec{n}, \vec{s}, \vec{v}) \int_\lambda f_C(\lambda) e(\lambda) c_s(\lambda) d\lambda$

for {R,G,B} giving an R-, B-, G-sensor response



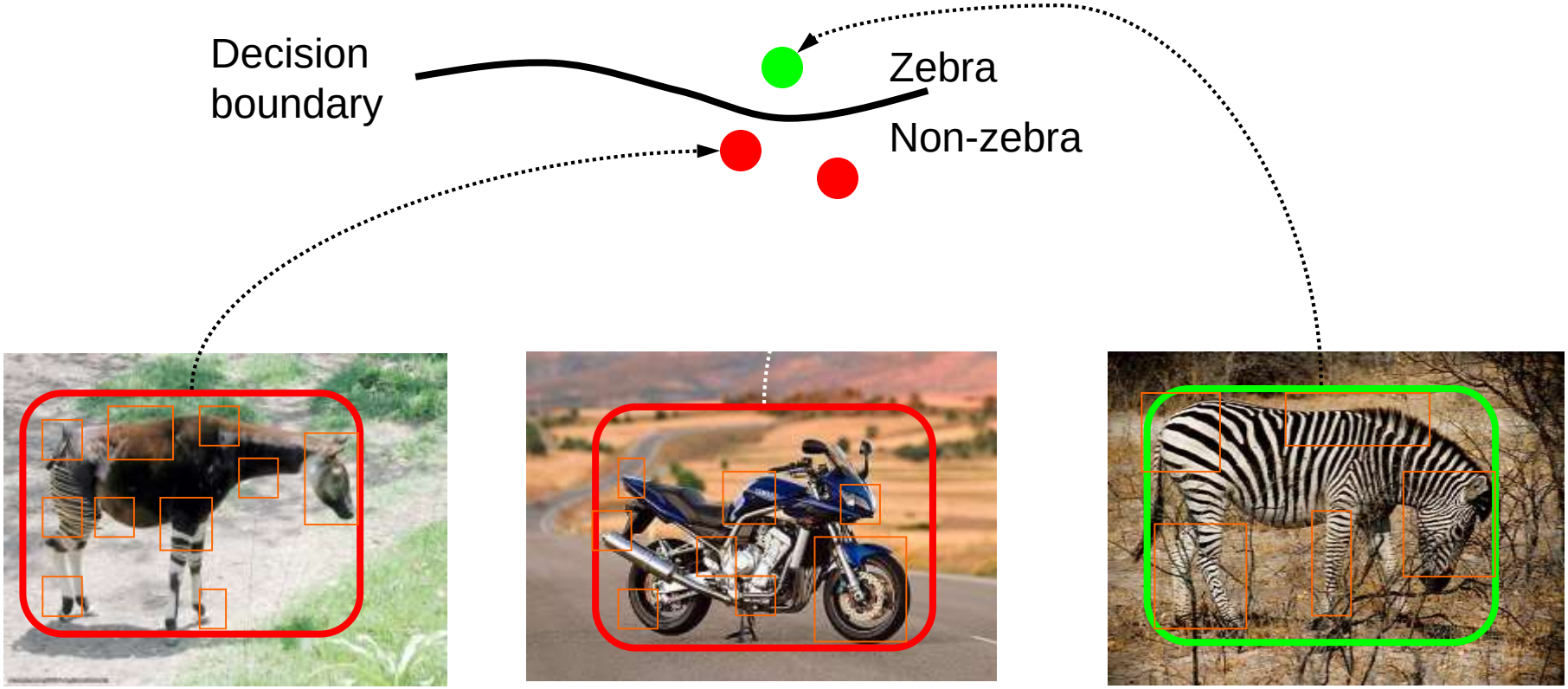
Reflection Model

$$C = m_b(\vec{n}, \vec{s}) \int_\lambda f_C(\lambda) e(\lambda) c_b(\lambda) d\lambda + m_s(\vec{n}, \vec{s}, \vec{v}) \int_\lambda f_C(\lambda) e(\lambda) c_s(\lambda) d\lambda$$

$c_b(\lambda)$	surface albedo	viewpoint invariant
$e(\lambda)$	illumination	scene dependent
$\vec{n}$	object surface normal	object shape variant
$\vec{s}$	illumination direction	scene dependent
$\vec{v}$	viewer's direction	viewpoint variant
$f_C(\lambda)$	sensor sensitivity	scene dependent

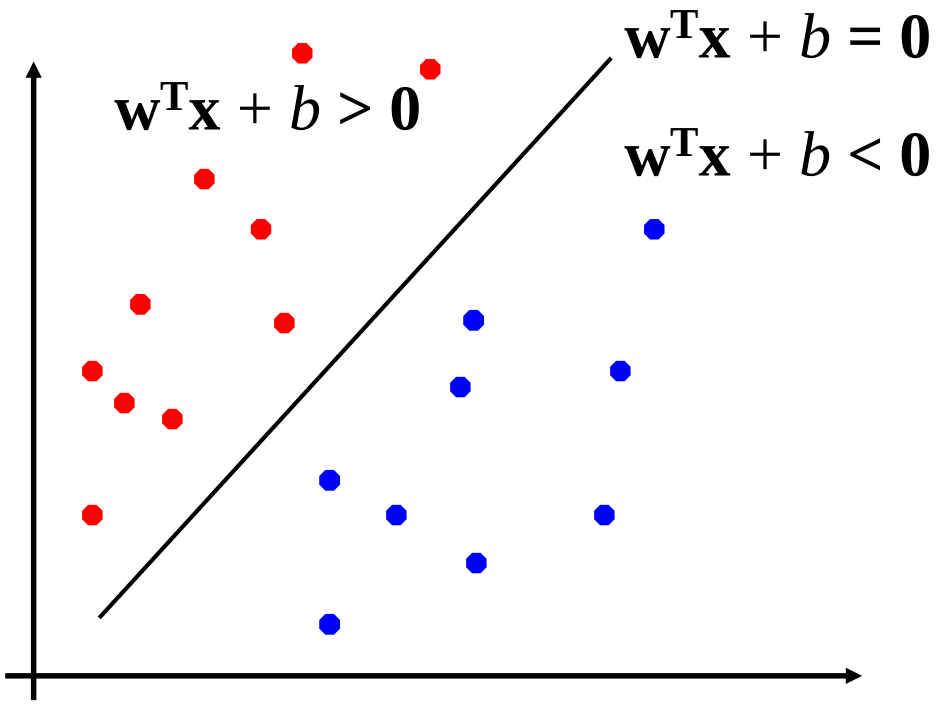
Machine learning

- Direct modeling of $\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}$



Linear Separators

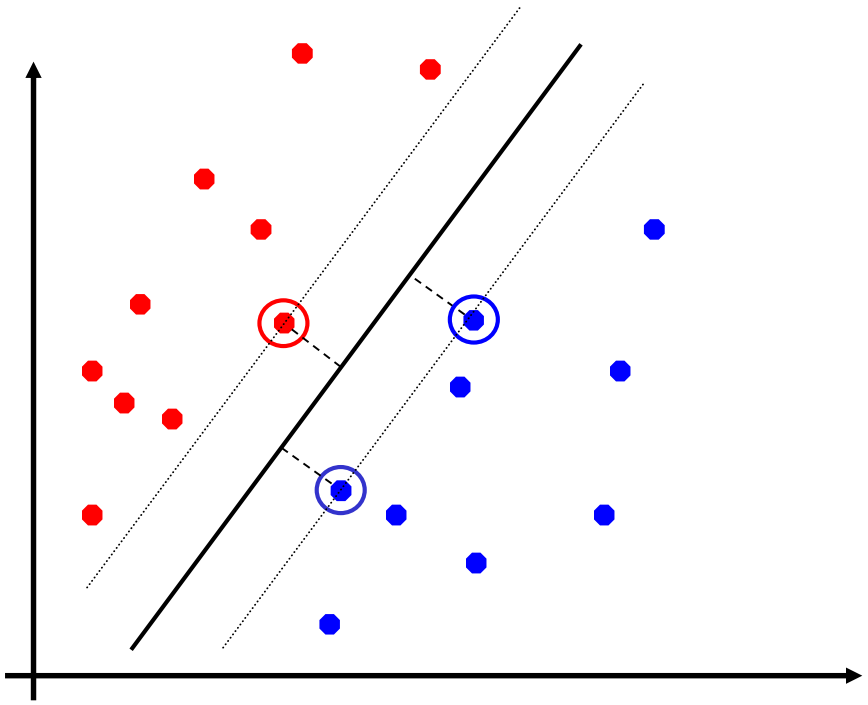
Binary classification can be viewed as the task of separating classes in feature space:



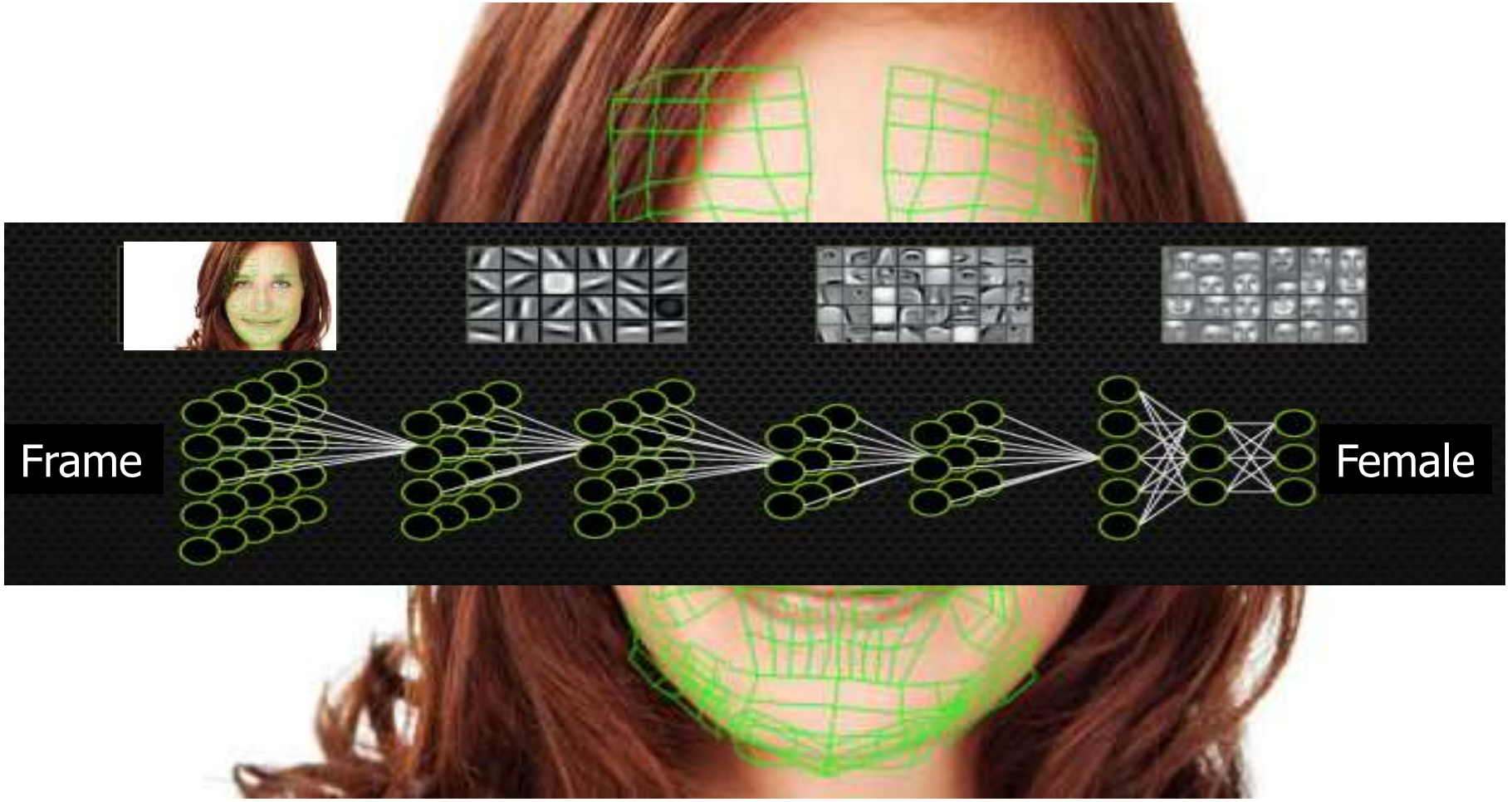
$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$$

Maximum Margin Classification: SVM

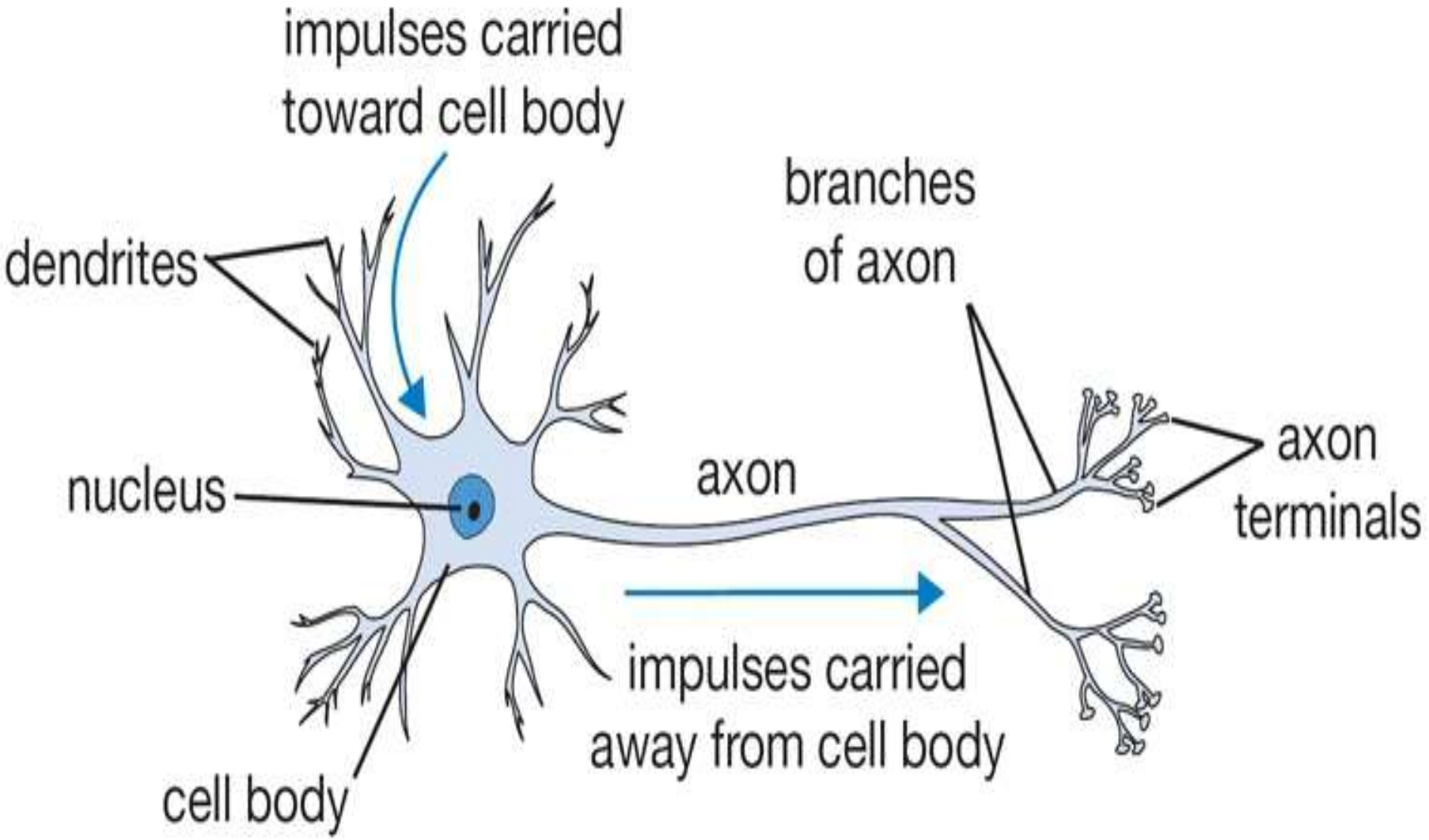
- Maximizing the margin is good.
- Implies that only support vectors matter; other training examples are ignorable.



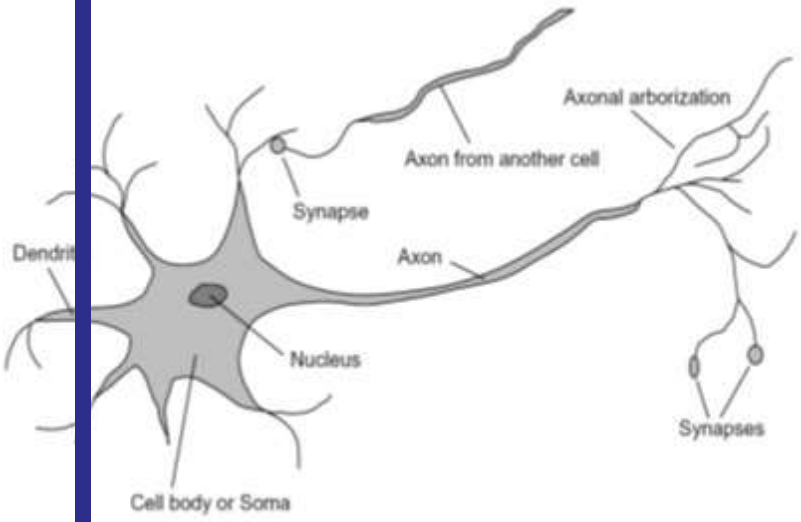
Deep Learning: Convolutional Neural Networks



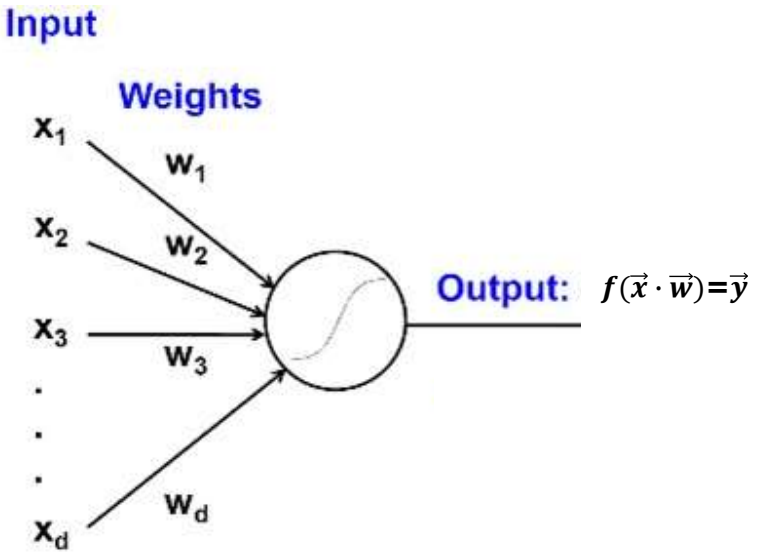
The Neuron


impulses carried
toward cell body
dendrites
branches
of axon
nucleus
axon
axon
terminals
impulses carried
away from cell body
cell body

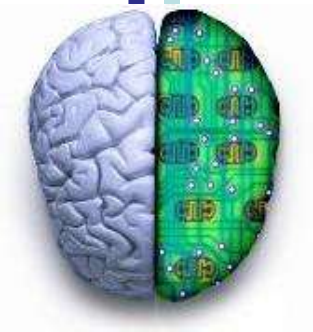
Artificial Model (1943 McCulloch/Pitts)



A biological neuron



An artificial neuron (Perceptron)
- a linear classifier

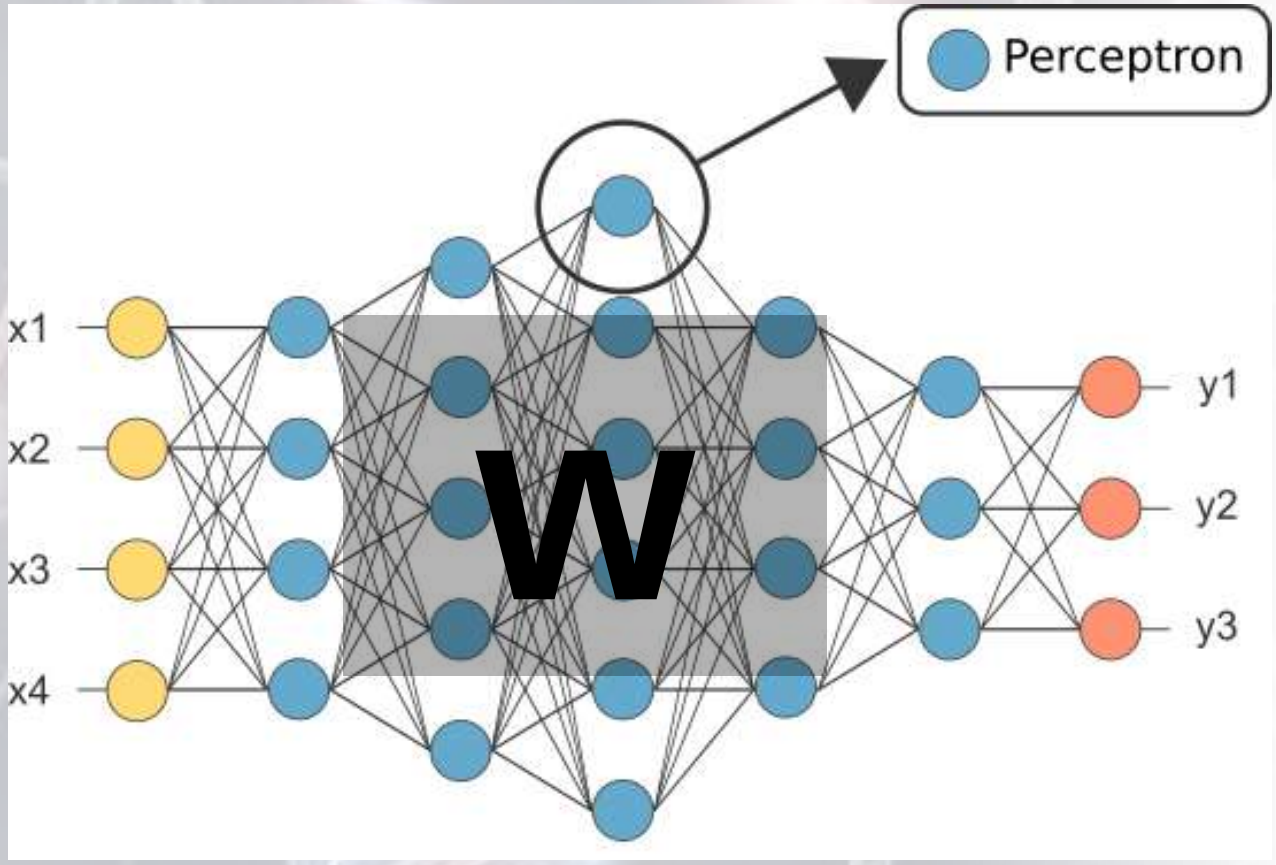


Feed-forward Neural Networks

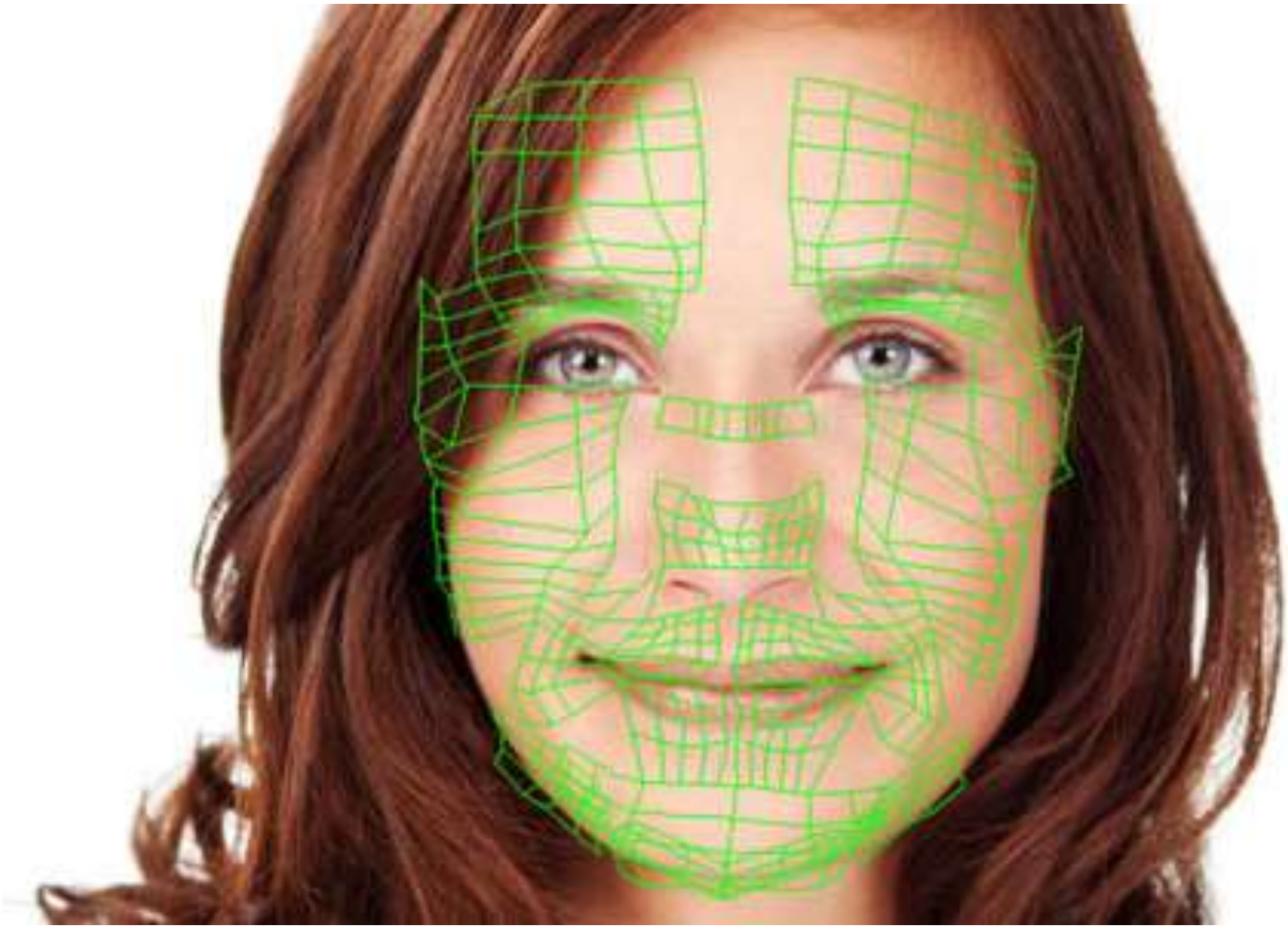
$$f(\vec{x} \cdot \vec{w}) = \vec{y}$$

Inputs: $\vec{x} = x_1, x_1, \ldots, x_n$

Ouputs: $\vec{y} = y_1, y_1, \ldots, y_m$



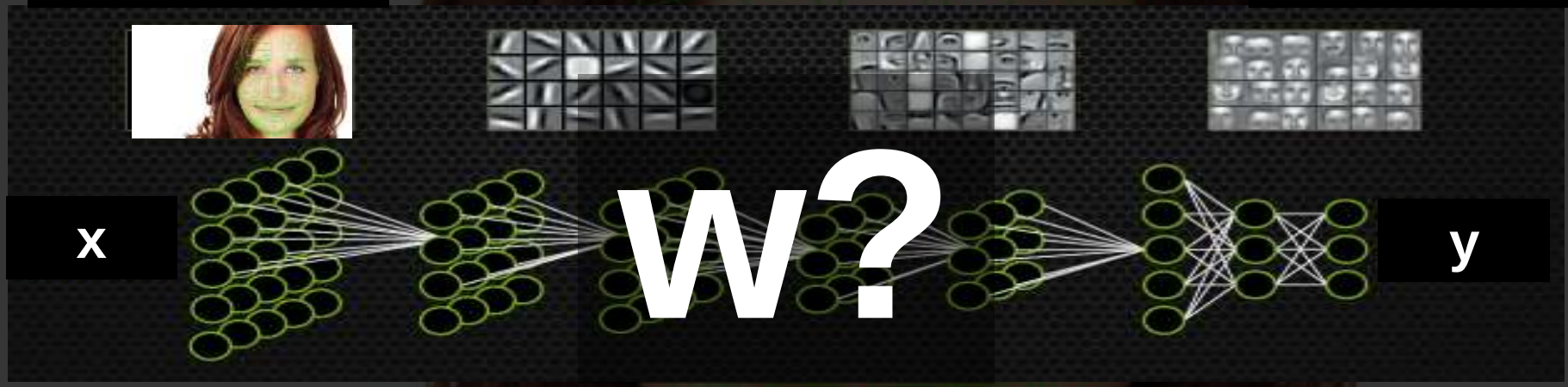
Face Analysis by Deep Learning



Face Analysis by Deep Learning

x = image

y = “Happy”

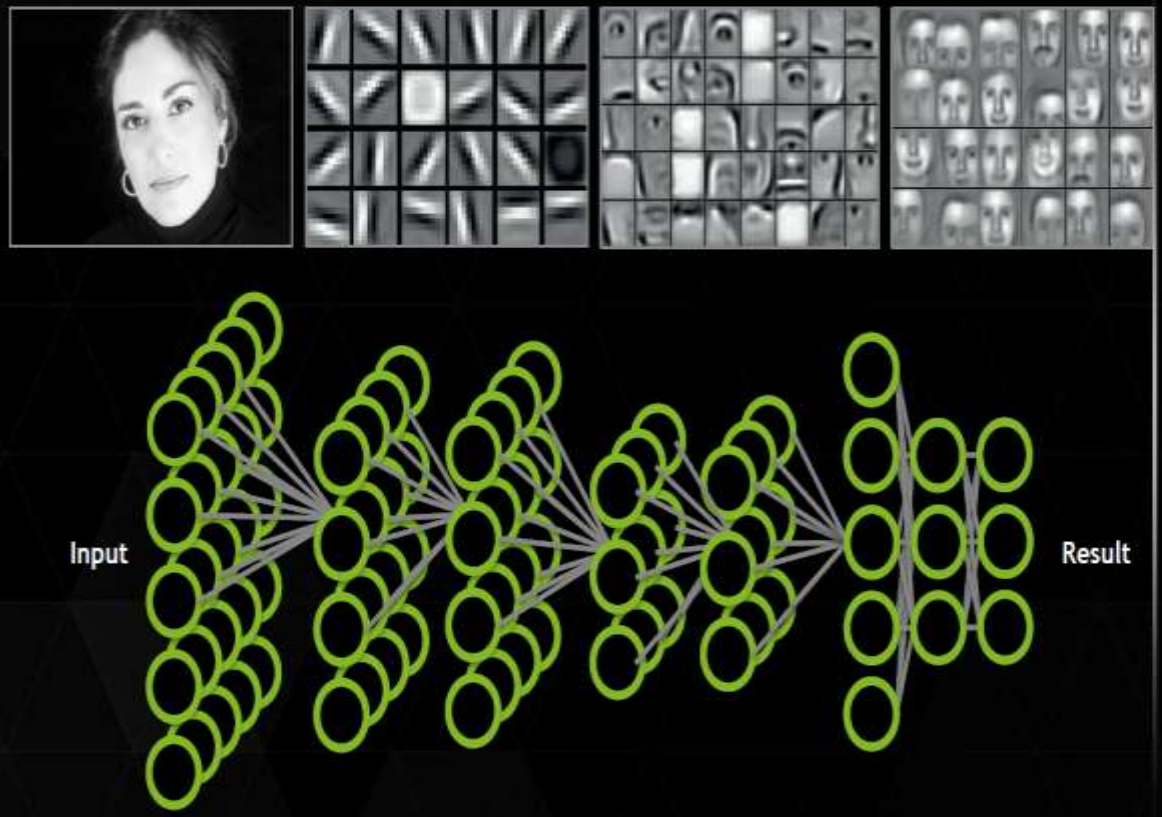


$$f(\vec{x} \cdot \vec{w}) = \vec{y}$$

Machine Learning: Deep Learning

$$f(\vec{x} \cdot \vec{w}) = \vec{y}$$

(Source: NVIDIA)

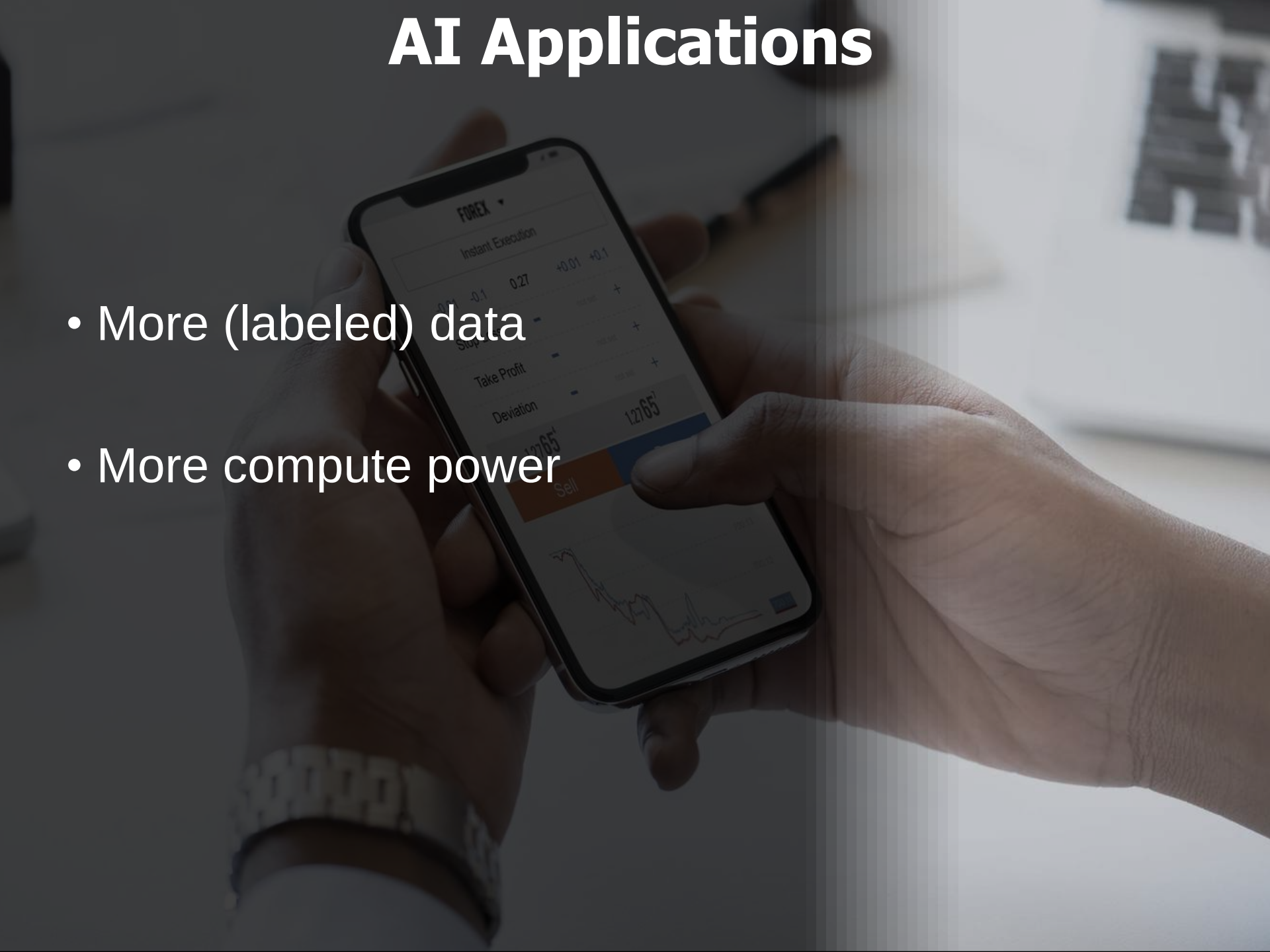


Largest network today:

- > 1000 layers
- > 1 billion parameters
- Datasets > 10 million images

- The human brain has a trillion parameters – that only 1.000 times more than a computer

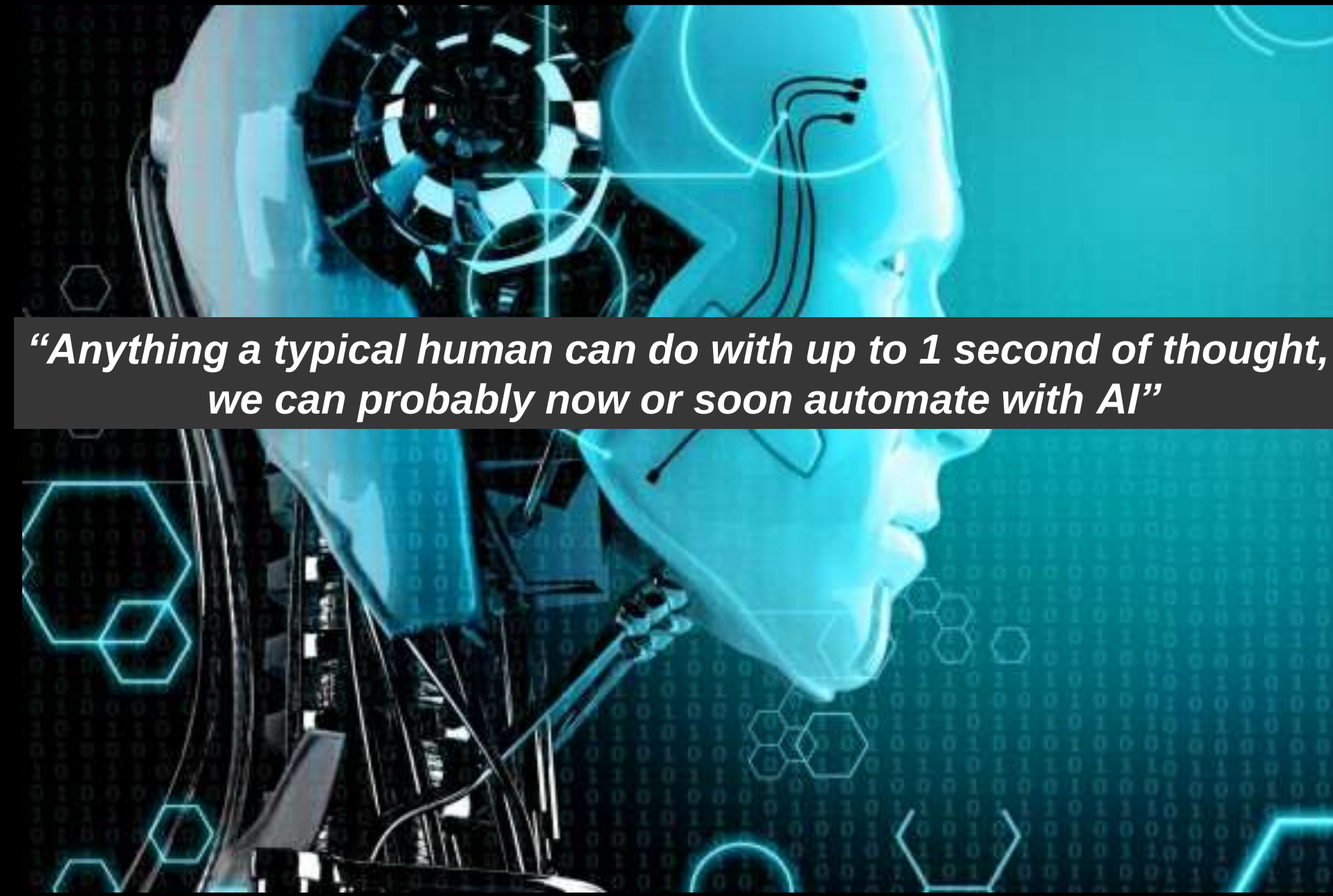
AI Applications



- More (labeled) data
- More compute power

All-in-One Deep Learning Super Computer

(Source: NVIDIA)


“Anything a typical human can do with up to 1 second of thought, we can probably now or soon automate with AI”

Artificial Intelligence

- Can computers recognize objects?
- What about faces and emotion recognition?
- Can computers be empathetic?
- When do computer become self-conscious ?
- What about super-intelligence?
- What are the moral implications?
- Job and societal impact?
- Etc etc.

Object Recognition

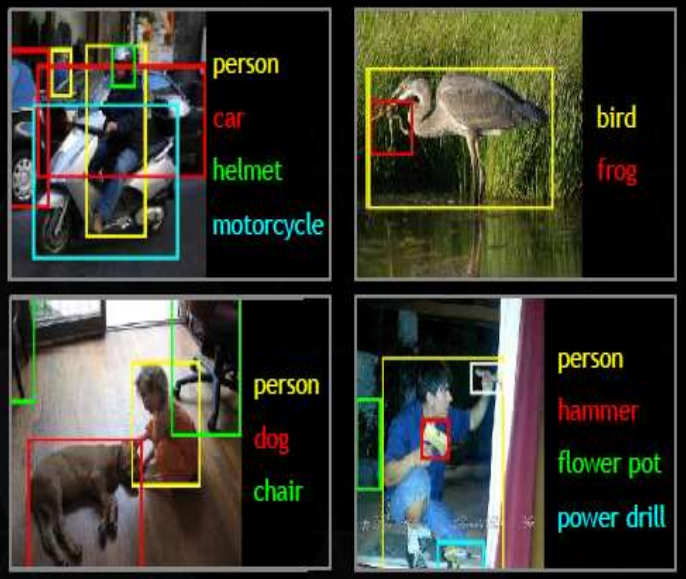
ImageNet

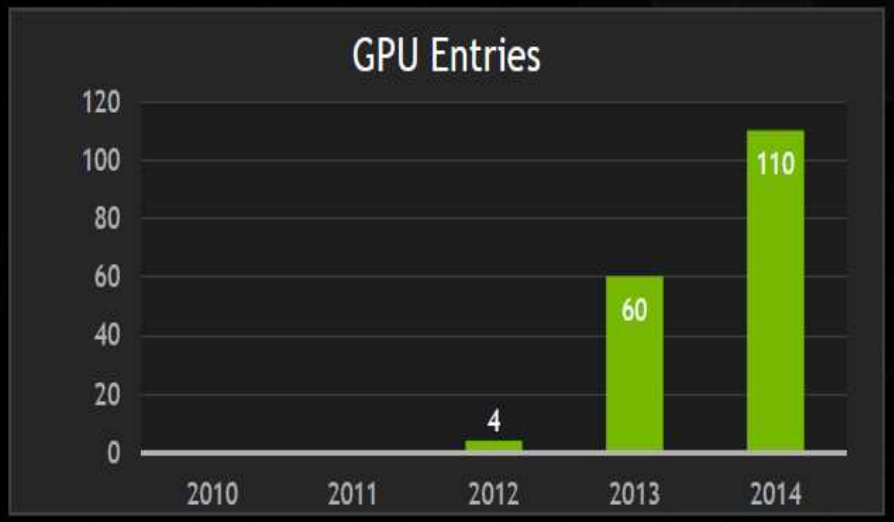
Image Recognition Challenge

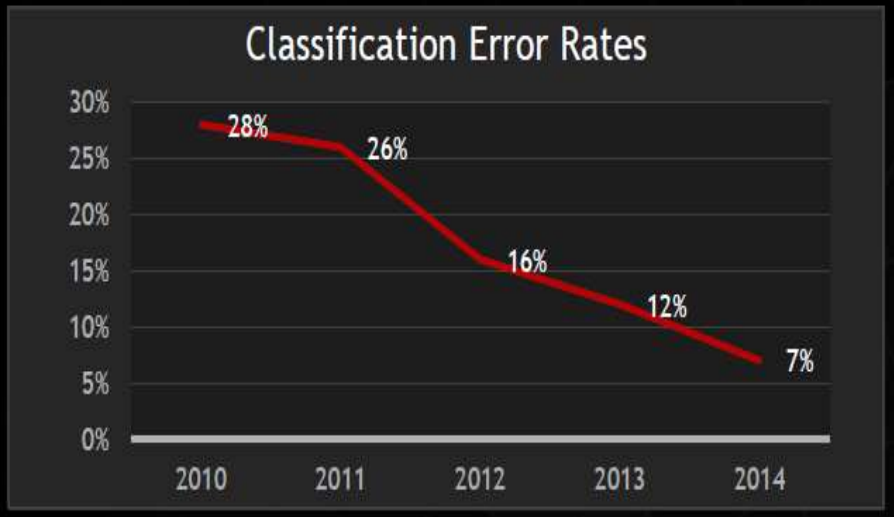
1.2M training images 1000 object categories

Hosted by

IMAGENET



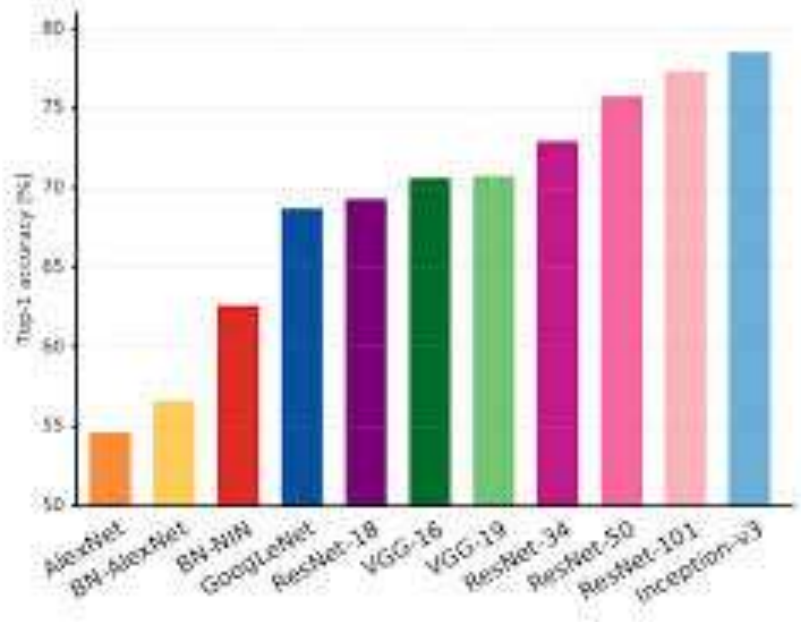


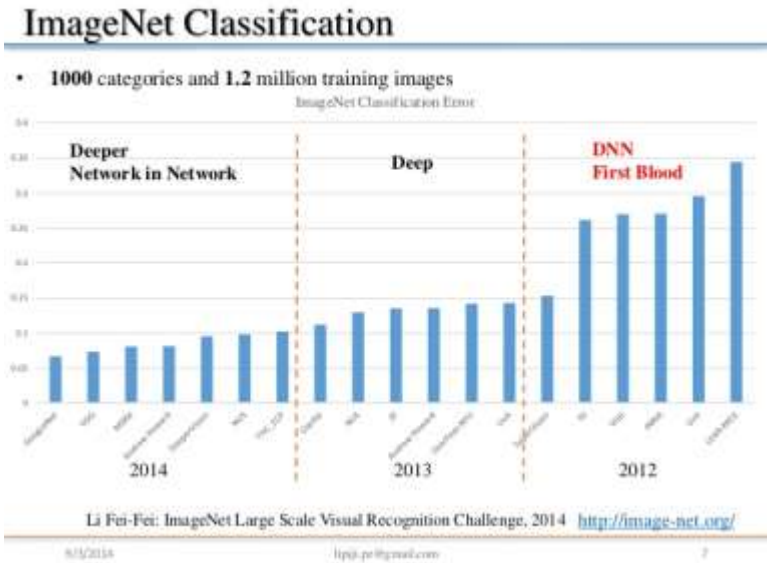


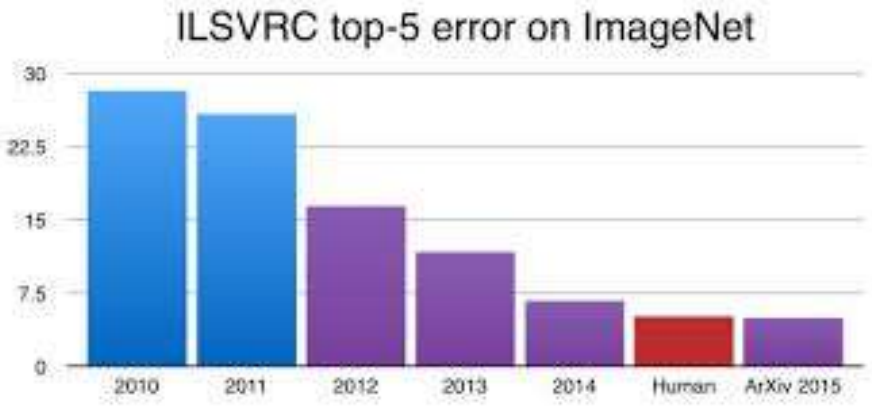
4.9% (Google, Microsoft) human performance in 2015

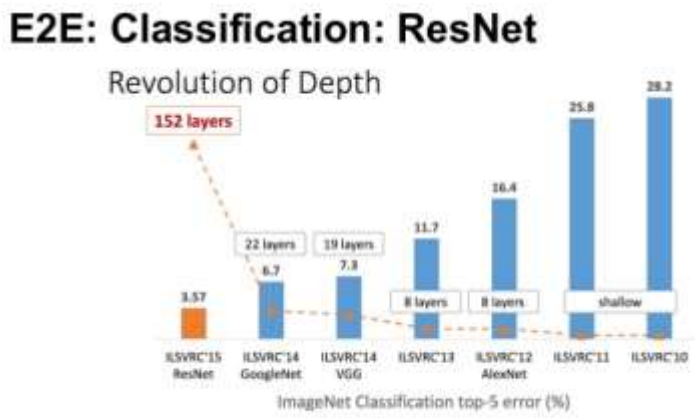
(Source: NVIDIA)

ImageNet Challenge 2012-2016





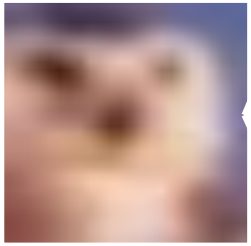




Feature Extraction and Object Recognition

SIFT

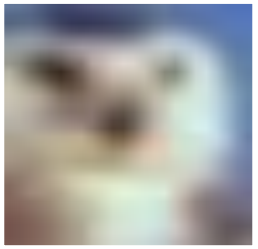
Feature Detection



Shape Description

SIFT

Color Description

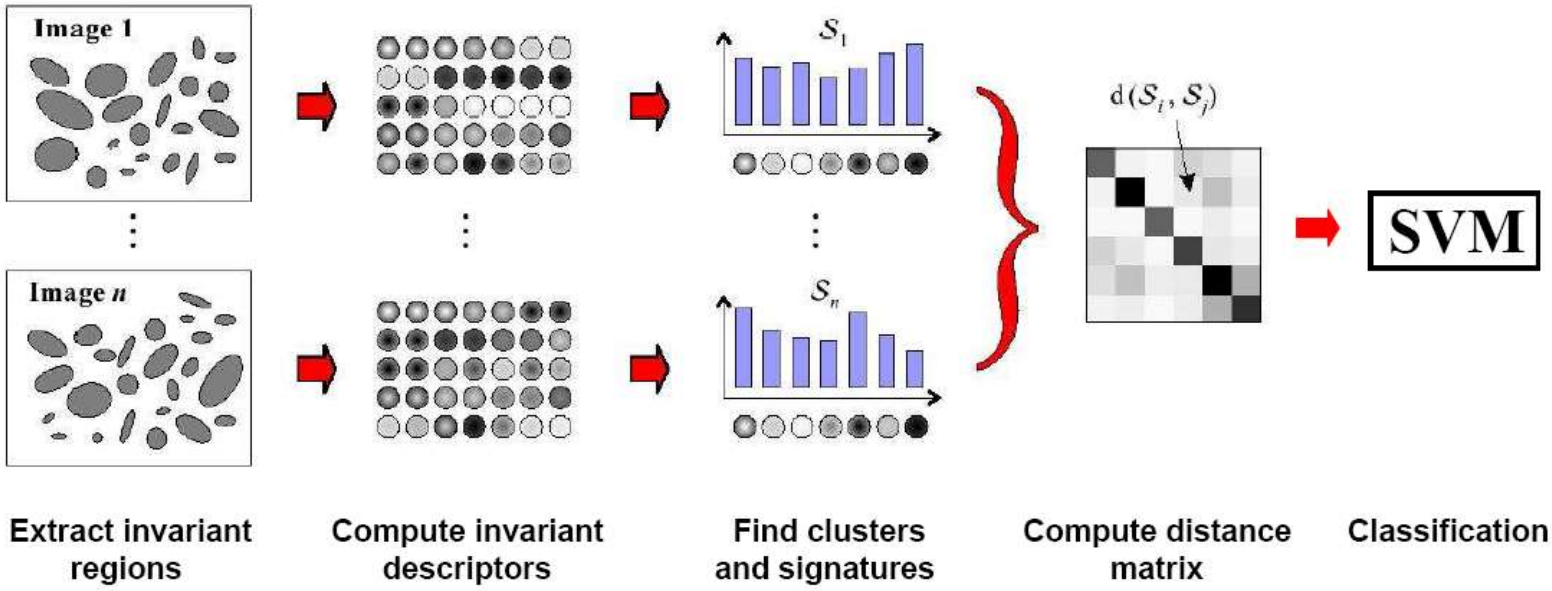


rgb

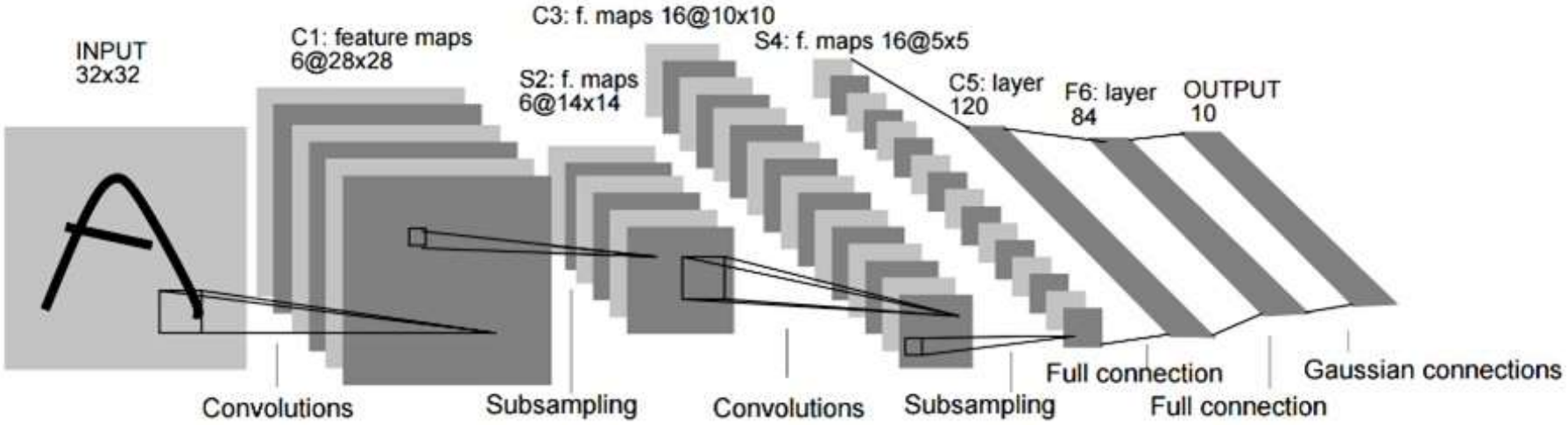
Invariance and Learning

Learn from the content-based image retrieval field: Zhang (2005) achieved state of the art performance using local features:

- invariant region detectors
- SIFT descriptor



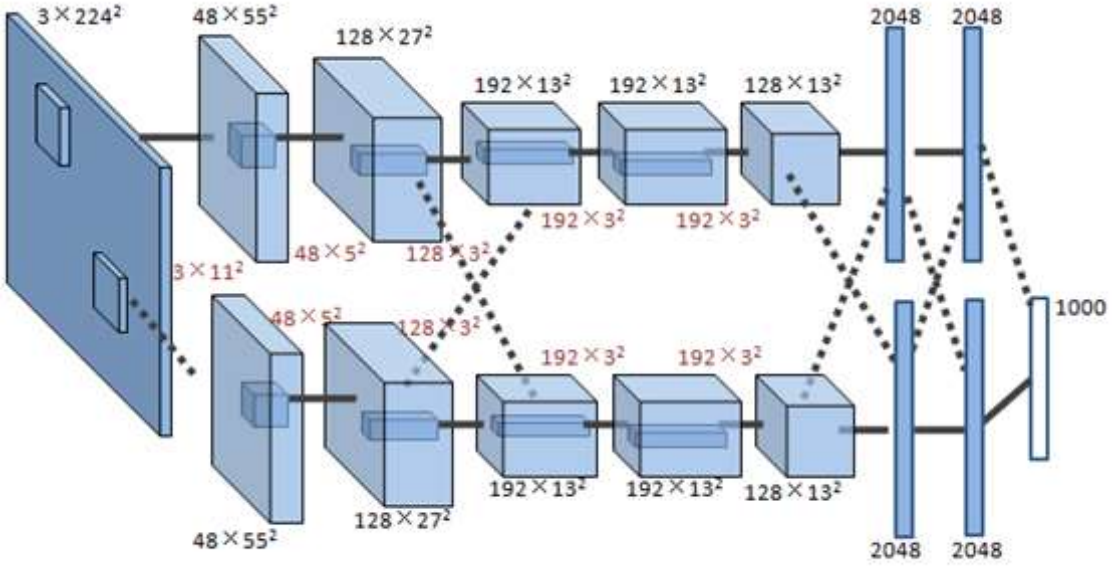
LeNet [LeCun et al. 1998]

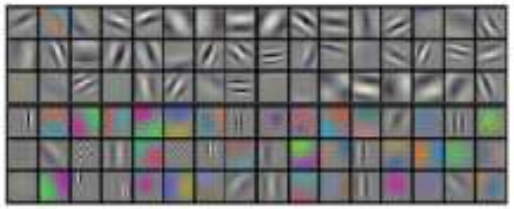


LeNet-1 from 1993

Gradient-based learning applied to document recognition [LeCun, Bottou, Bengio, Haffner 1998]

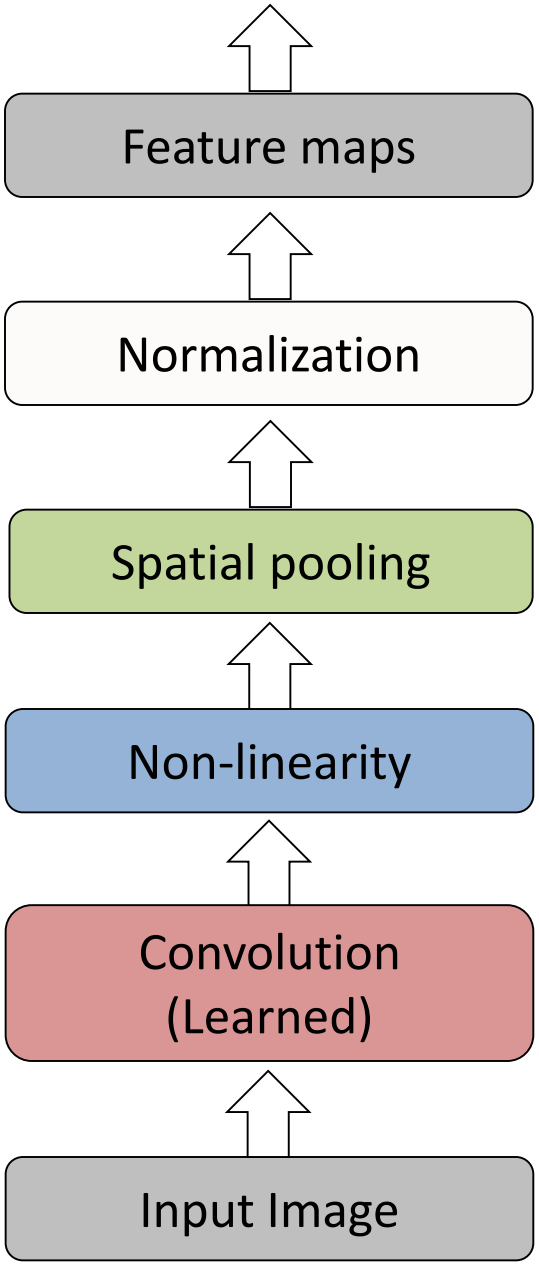
AlexNet


3×224²
48×55²
128×27²
192×13²
192×13²
128×13²
2048
2048
192×3²
192×3²
48×5²
128×3²
3×11²
48×5²
128×3²
192×3²
192×3²
1000
192×13²
192×13²
128×13²
48×55²
128×27²
2048
2048



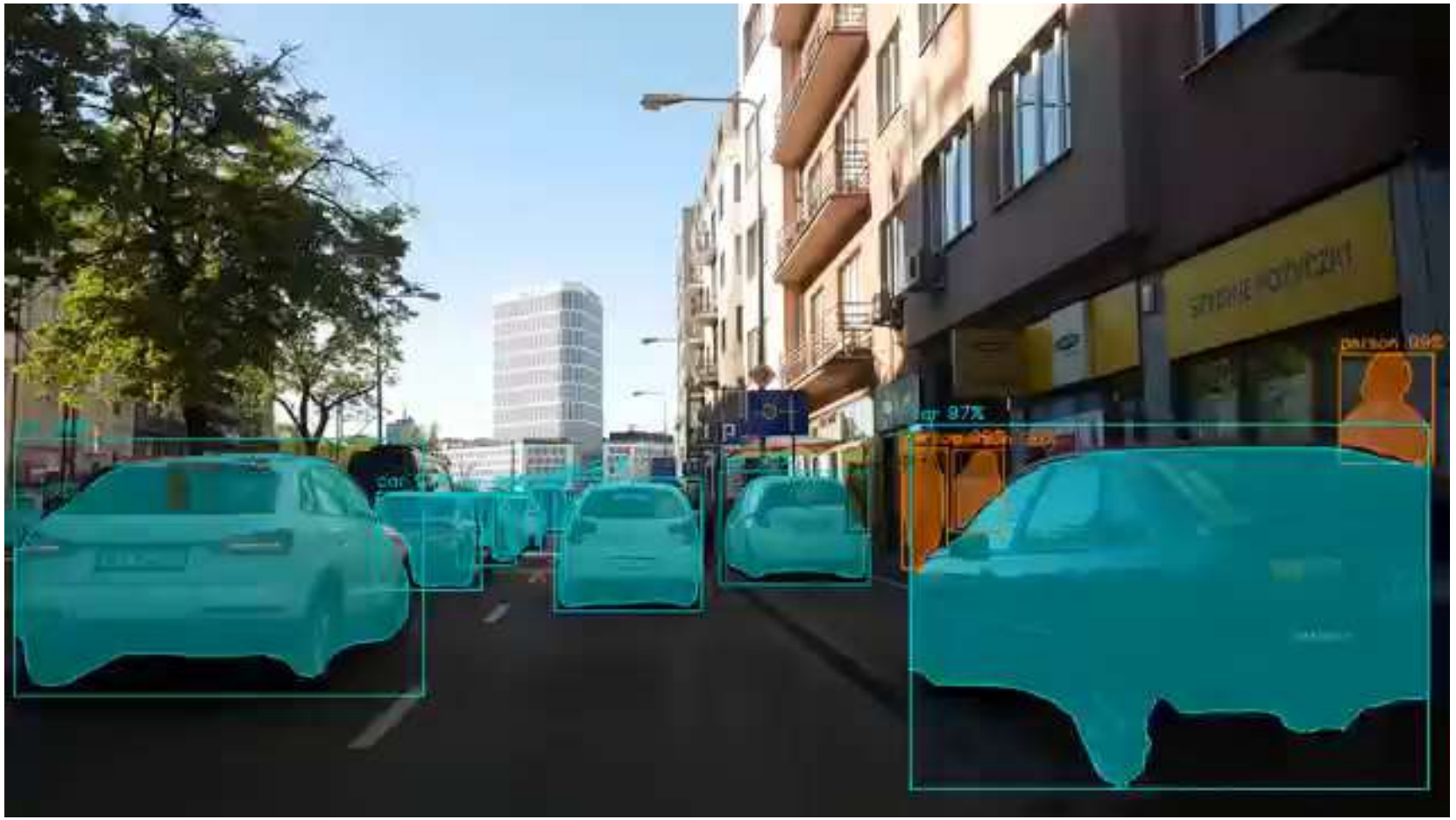

mite
container ship
motor scooter
leopard
grille
mushroom
cherry
Madagascar cat

Convolutional Neural Networks



slide credit: S. Lazebnik

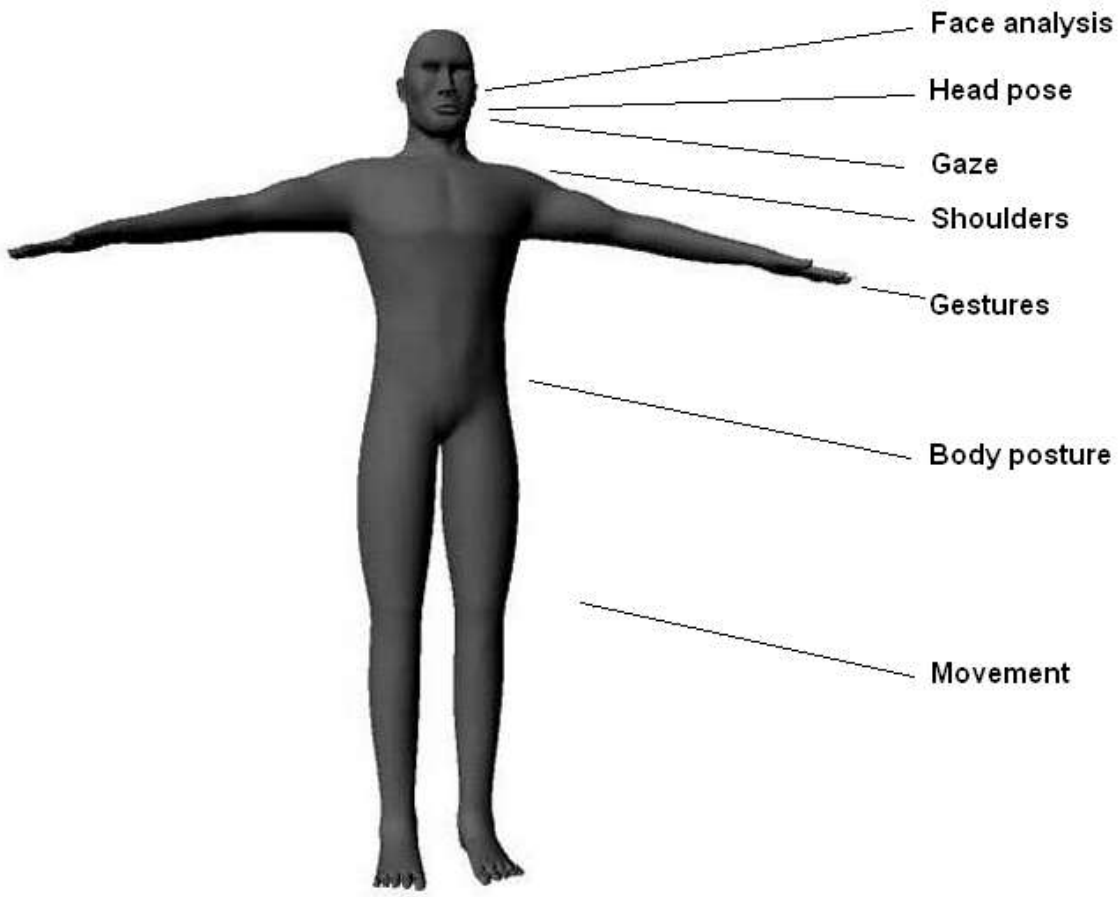
Semantic Segmentation and Tracking



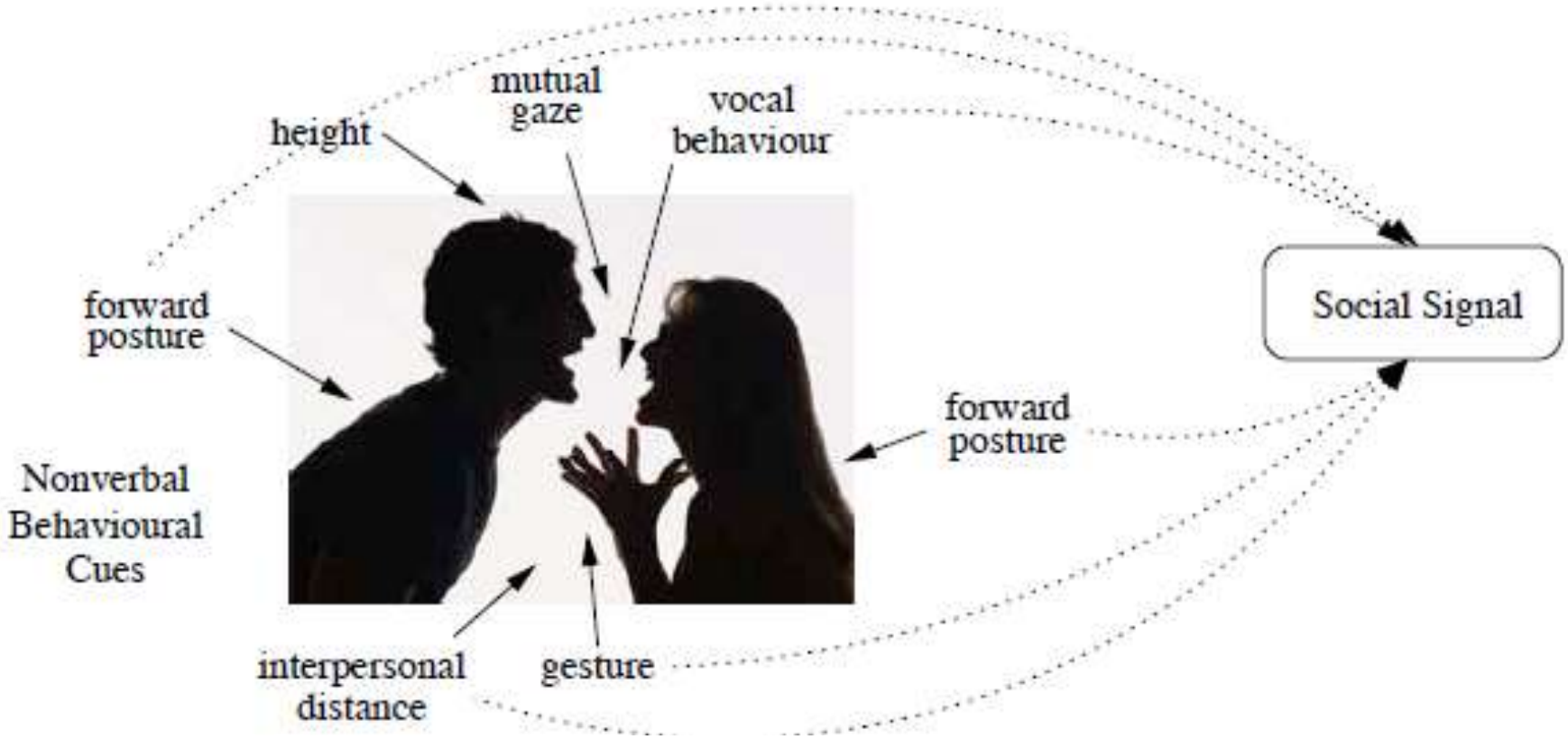
Human behaviour analysis

Body Language

Visual analysis of the human body



Activity Recognition




Buffy
Joyce
Dawn
Buffy
Buffy
Giles
Other
Tara
Willow
Willow
Buffy
Dawn
Willow
Tara
Willow
Willow
Buffy
Dawn
Tara
Buffy
Willow
Tara
Xander
Anya
Buffy
Riley
Spike

Emotion Recognition

Basic Emotions

Action Units

AU 1
Inner brow raise

AU 2
Outer brow raise

AU 4
Brow lower

AU 6
Cheek raise

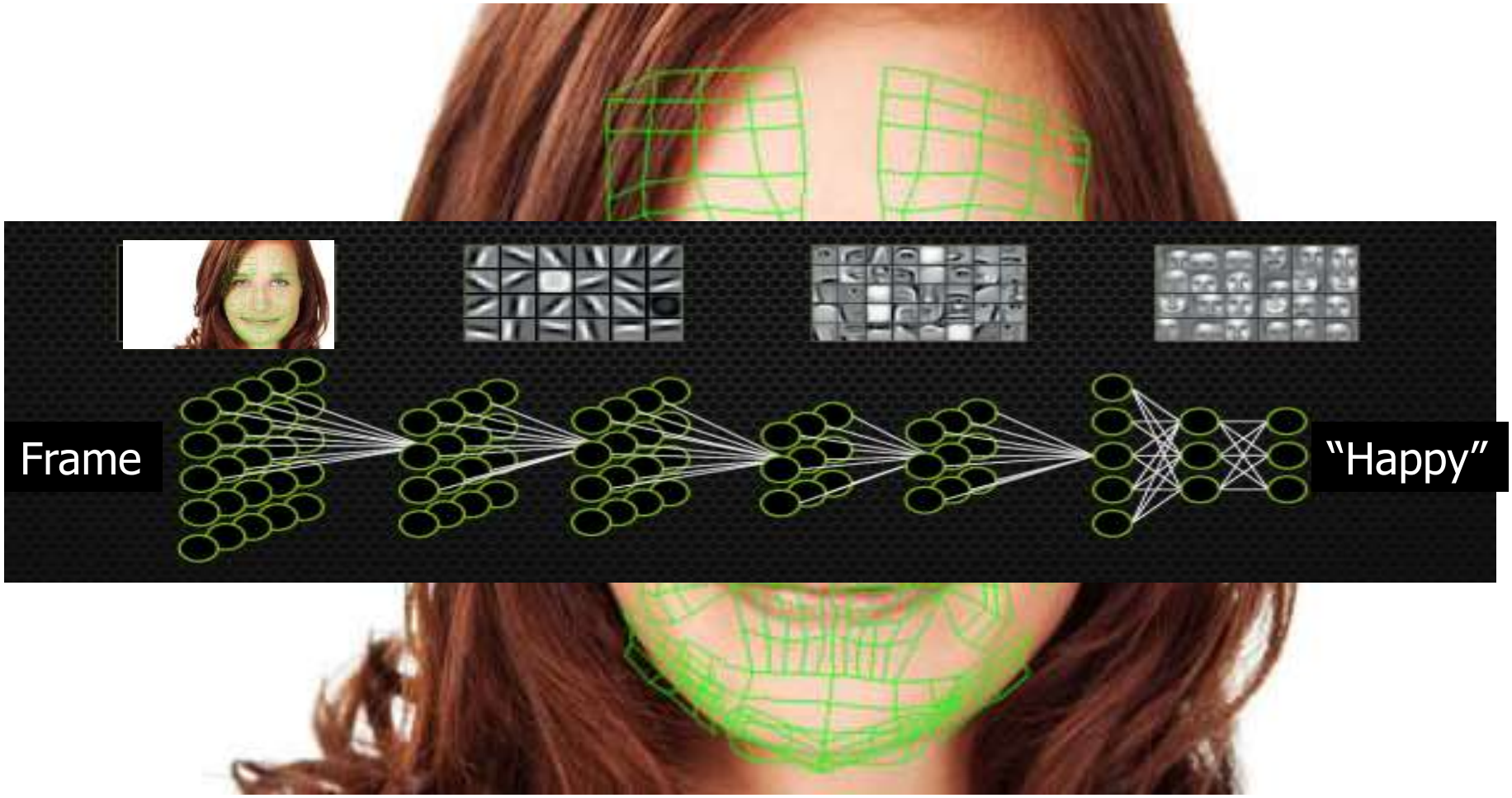
AU 9
Nose wrinkler

AU 12
Lip corner pull

AU 15
Lip corner depress

AU 20
Lip strecher

Face Analysis by Deep Learning



What is going on?

Face Analysis

The Synthesis of 3D Faces

Blanz and Thomas Vetter
Max-Planck-Institut fur biologische Kybernetik,
Tubingen, Germany

Consumer Applications

Face Analysis

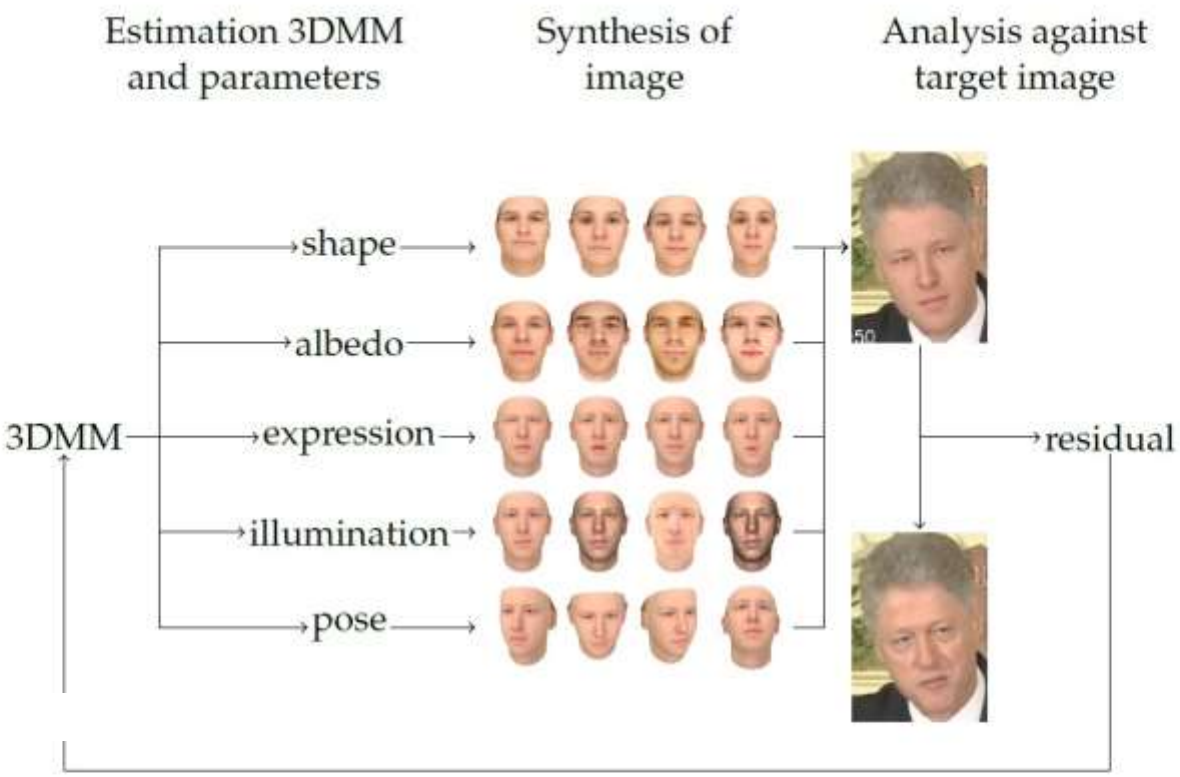
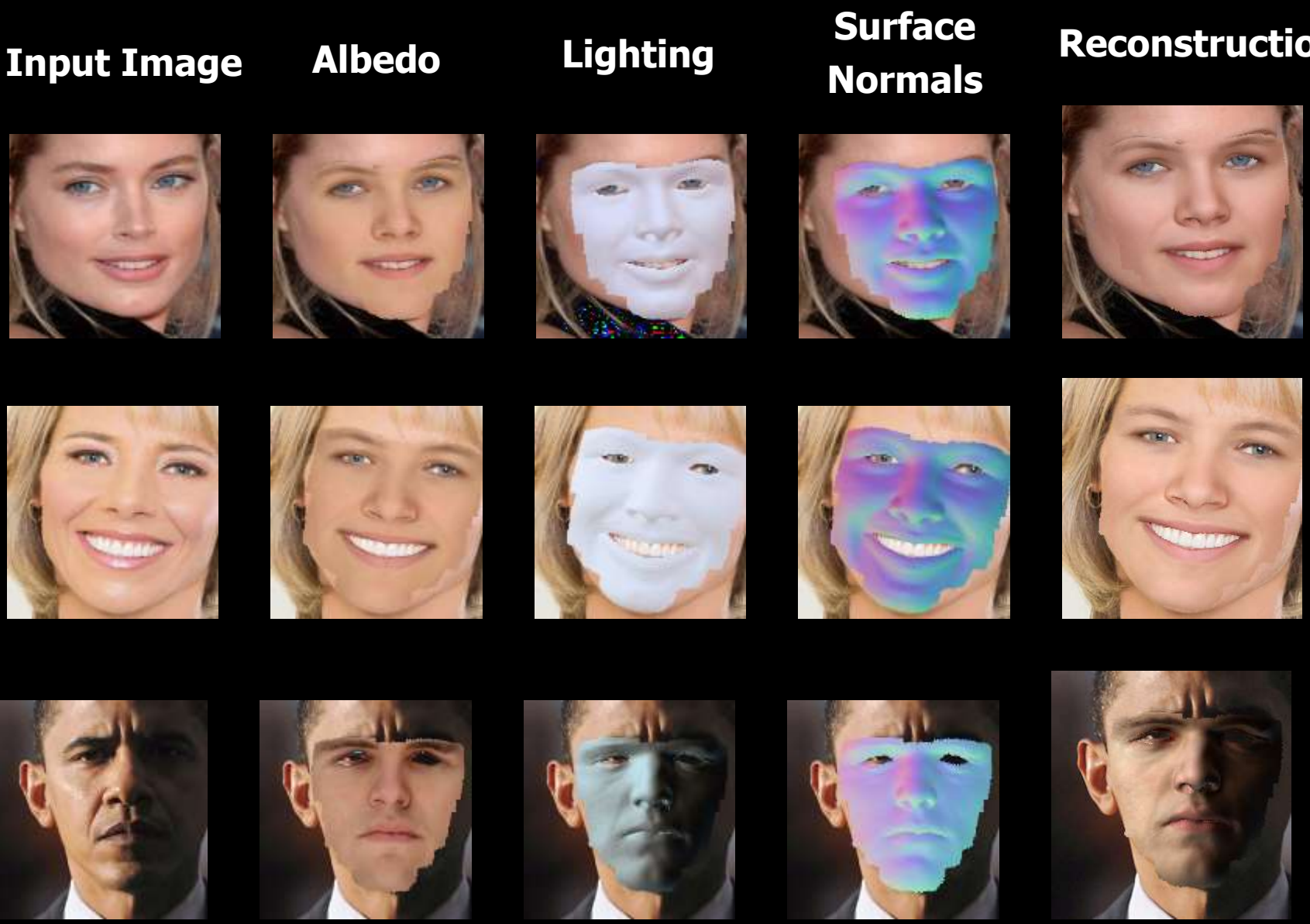
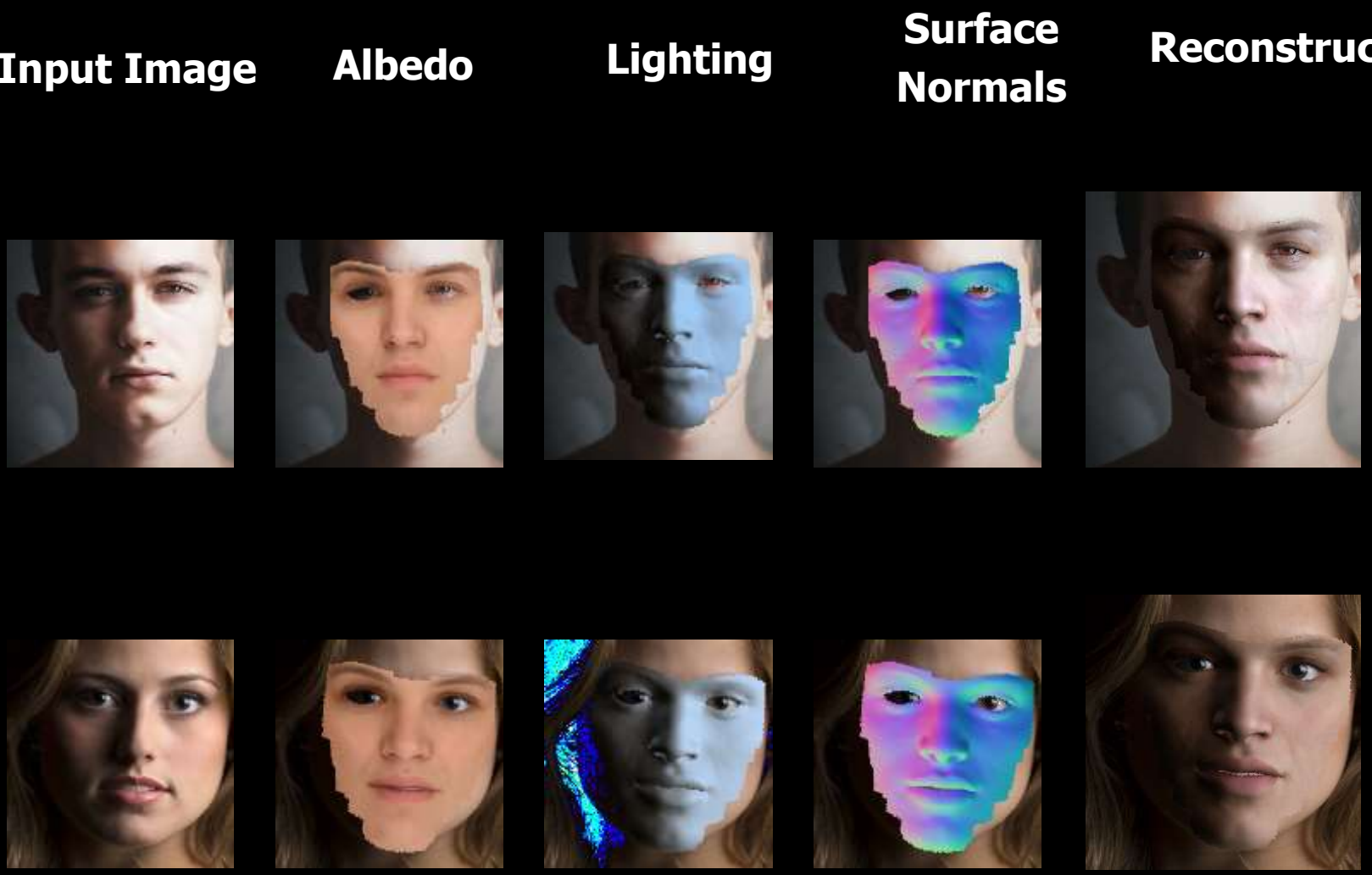


FIGURE 1.1: Example of 3DMM fitting following an analysis-by-synthesis approach. From left to right: The 3D Morphable Model, prediction of parameters for the different components of the optimization problem and rendered results, analyses against target image, and feedback loop.

A.I.: Deep Convolution Learning



A.I.: Deep Convolution Learning



Deep Convolution Learning: Illumination

Face2Face: Real-time Face Capture and Reenactment of RGB Videos

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt
Matthias Nießner (University of Erlangen-Nuremberg, Max-Planck-Institute for Informatics, Stanford University)

Computer Vision:
Many more Applications

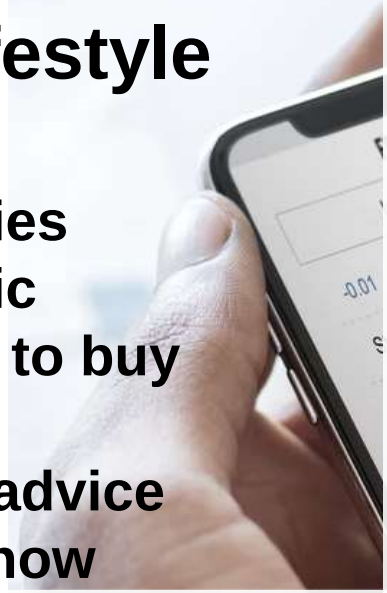
Healthcare

Diagnose pneumonia
Diagnose skin cancers
Predict autism
Predict heart attacks and strokes
Diagnose cataracts
Generate oncology treatment plans
Identify diabetic retinopathy
Identify malaria parasites in blood
Predict Alzheimer's disease
Predict high blood pressure
Predict schizophrenia
Predict sleep apnea
Detect falls in the home
Diagnose prostate cancer
Make precise incisions
Predict hospital readmissions
Screen for cervical cancer
Suture a wound

Applications

Home & Lifestyle

- Recommend movies
- Recommend music
- Recommend stuff to buy
- Adjust your lights
- Give you fashion advice
- Guess who you know
- Invent recipes
- Learn your weekly shop
- Mow your lawn
- Optimize your heating
- Reduce your water bills
- Vacuum your floors
- Buy stuff on your behalf
- Control your entire house
- Cook your meals



Creative

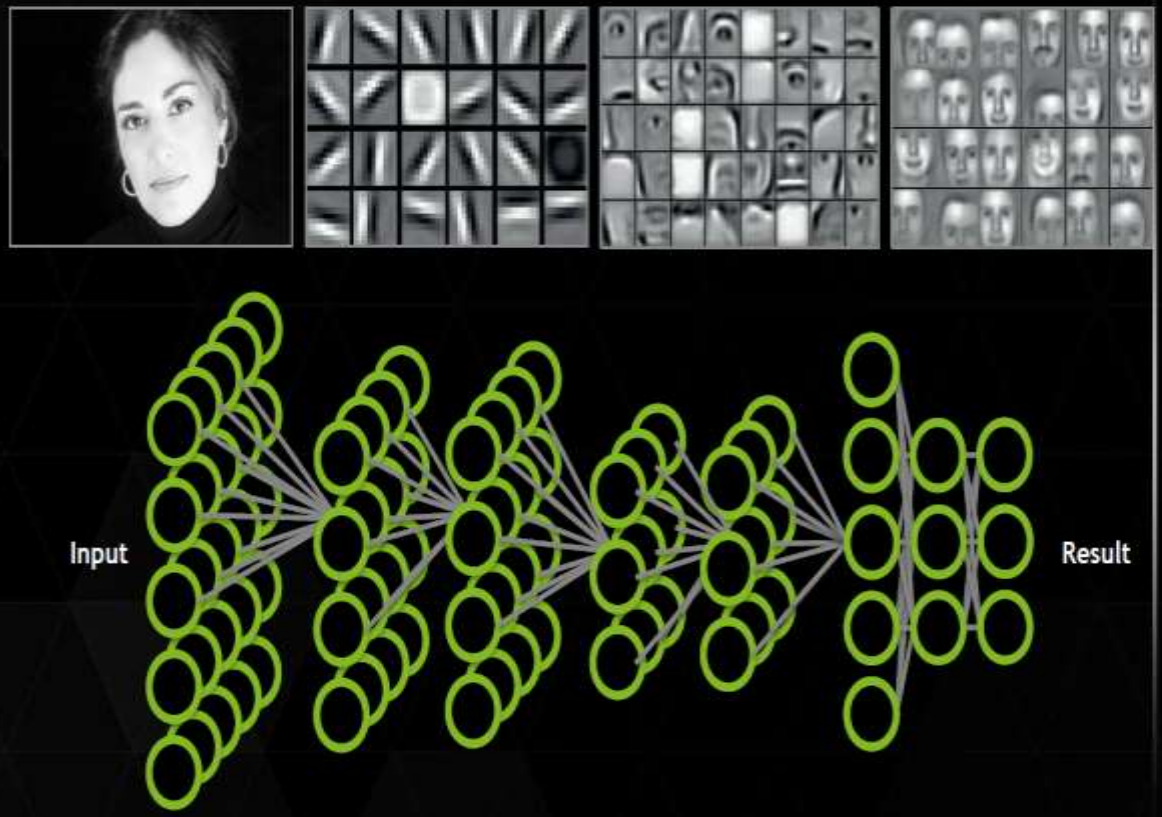
- Fake a video
- Mimic famous artists
- Spot forged artworks
- Compose classical music
- Copy your handwriting
- Design logos
- Direct a panel show
- Draw creepy pictures
- Edit photos
- Generate photorealistic faces
- Mix like a DJ
- Recognize doodles
- Compose pop music
- Write a film
- Write a novel

Style Transfer

Machine Learning: Deep Learning

$$f(\vec{x} \cdot \vec{w}) = \vec{y}$$

(Source: NVIDIA)

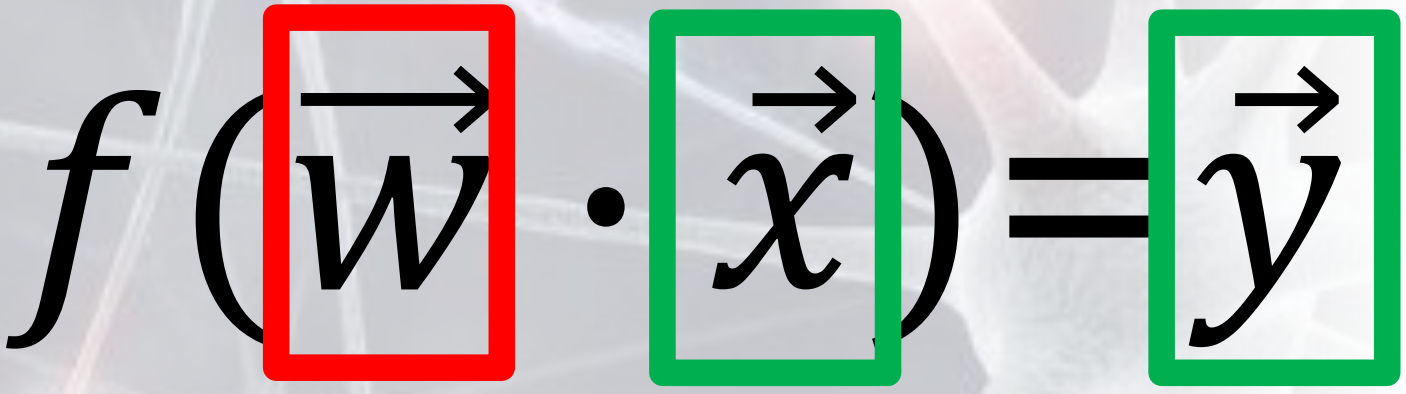


Largest network today:

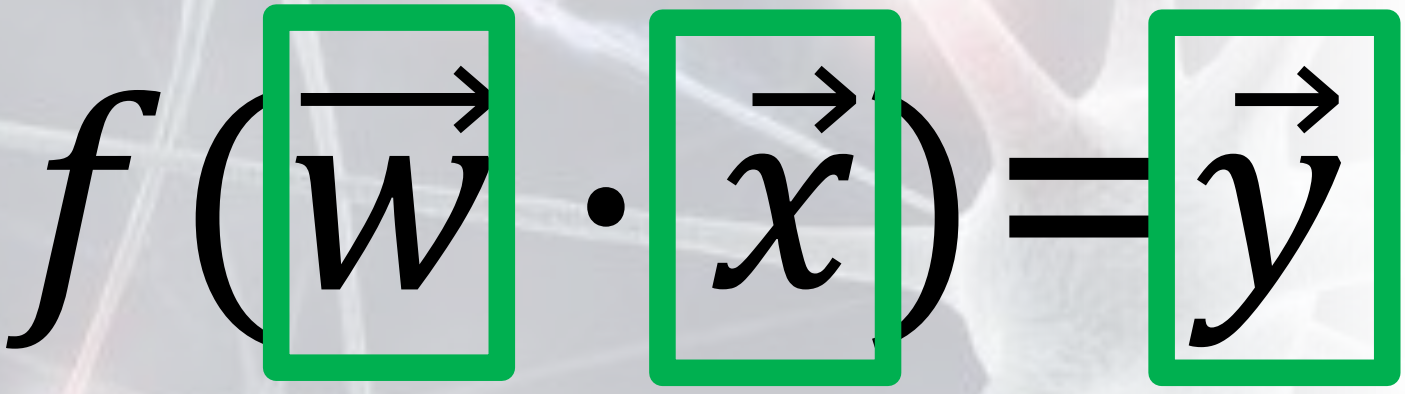
- > 1000 layers
- > 1 billion parameters
- Datasets > 10 million images

- The human brain has a trillion parameters – that only 1.000 times more than a computer

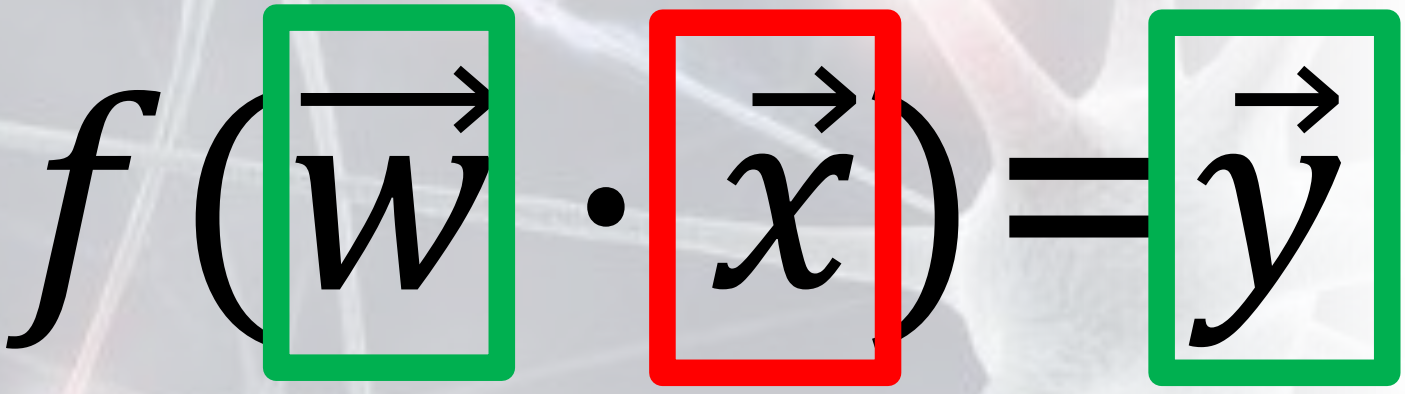
Generative AI Models



Generative AI Models

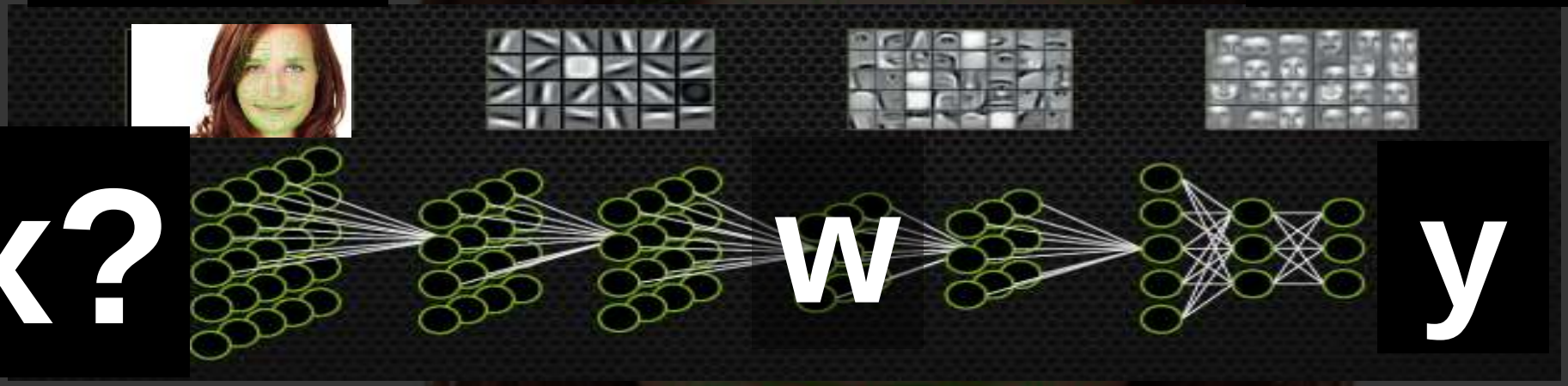


Generative AI Models



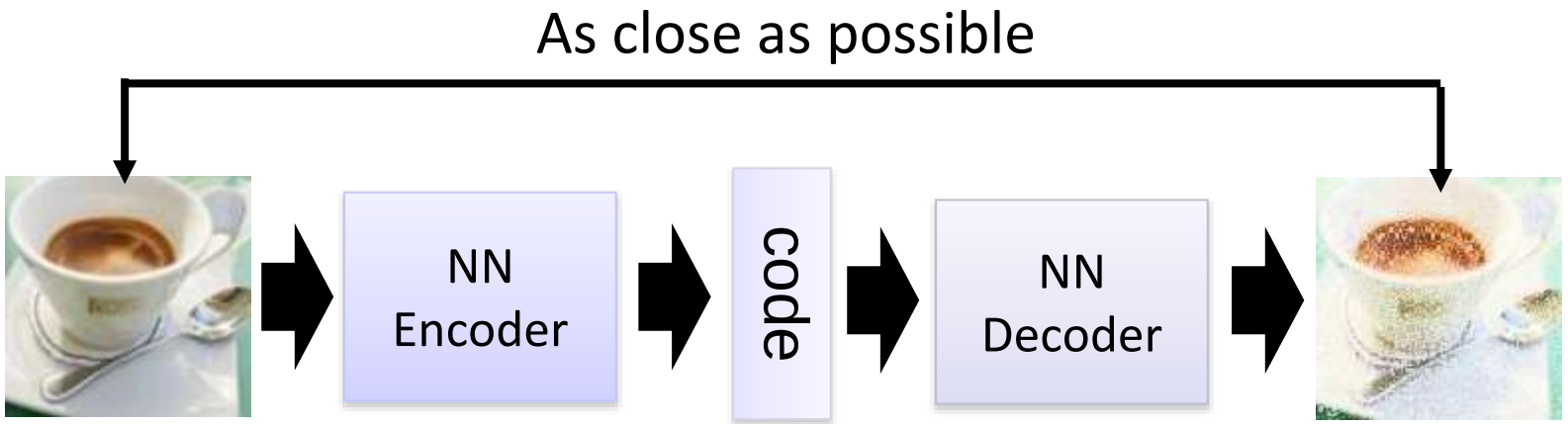
x = image?

y = “Happy”

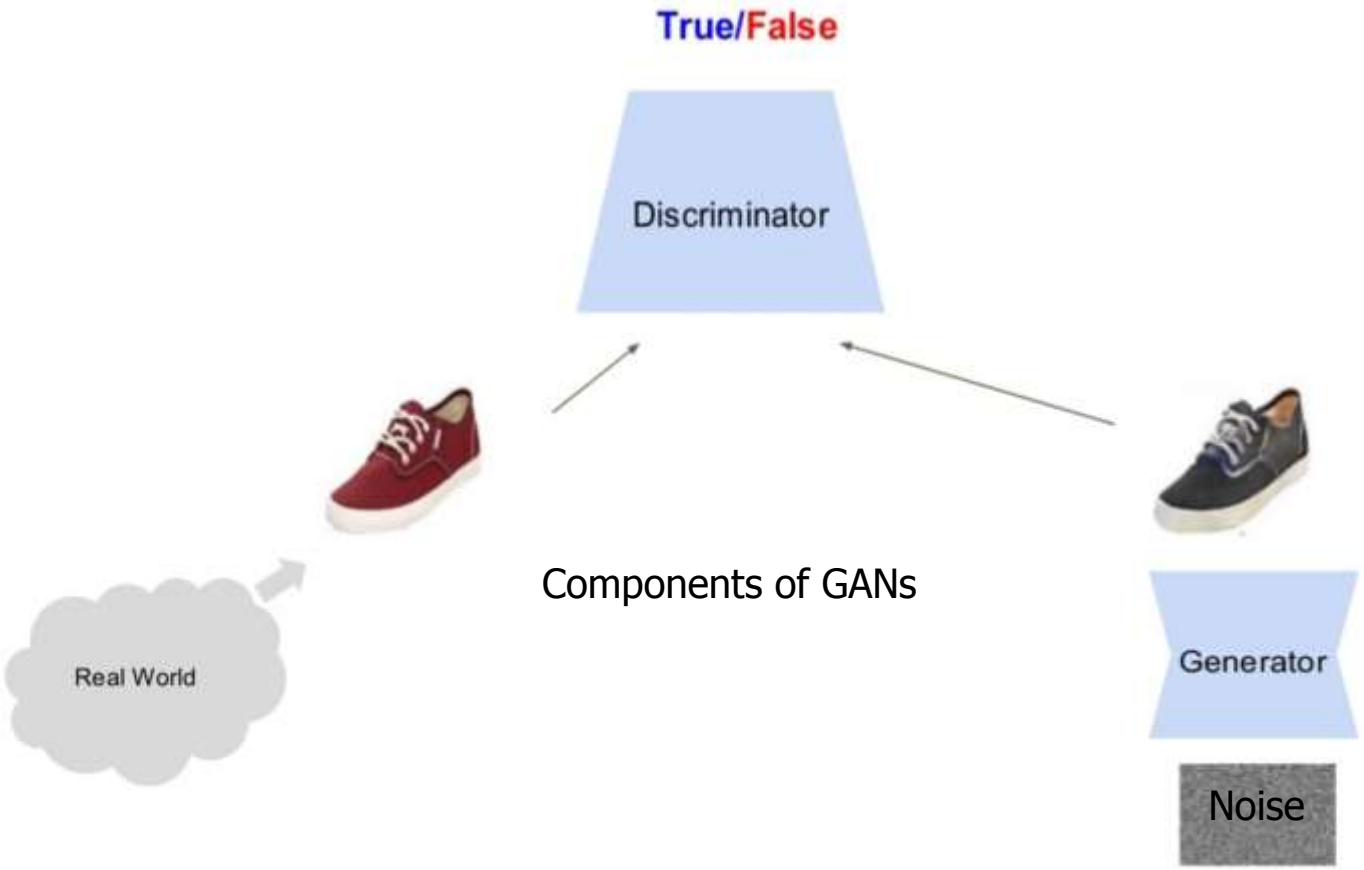


$$f(\vec{x} \cdot \vec{w}) = \vec{y}$$

Autoencoder

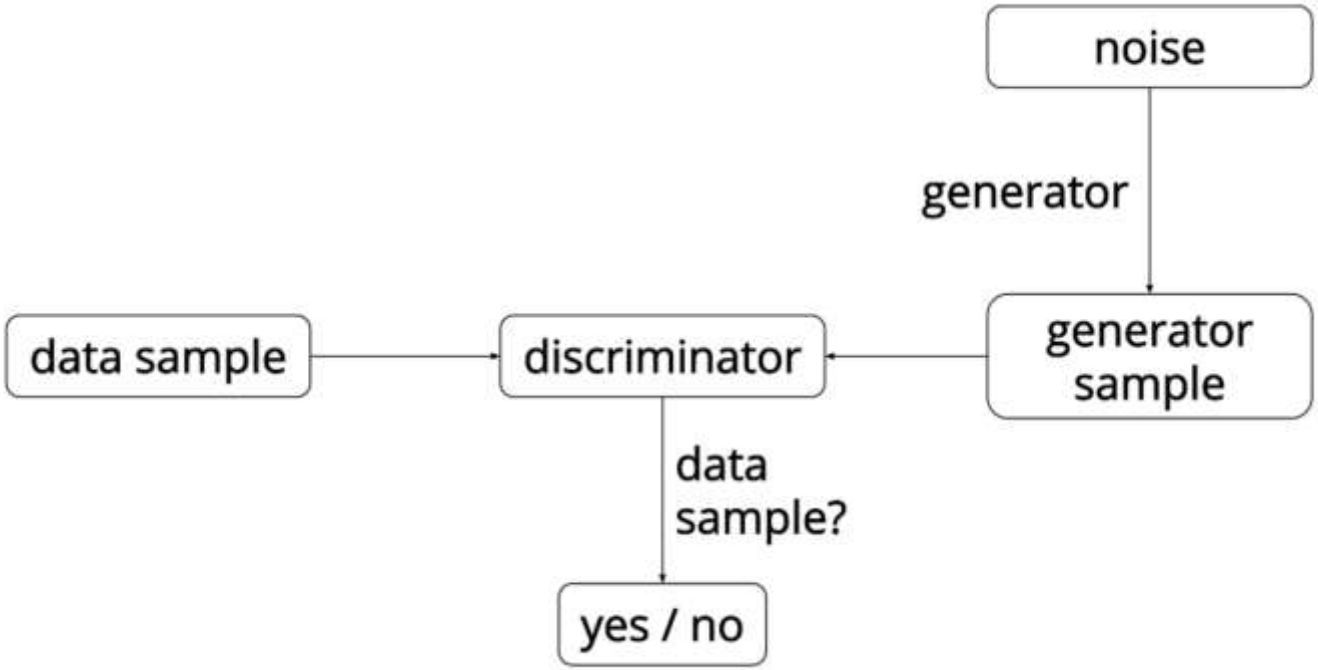


GAN



Slide credit – Victor Garcia

GAN



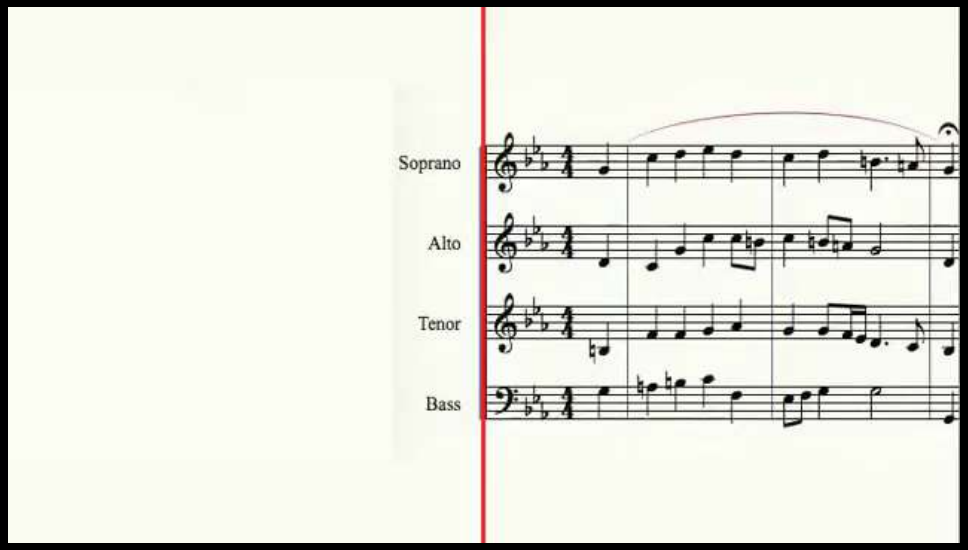
Overview of GANs
Source: https://ishmaelbelghazi.github.io/ALI

Art: Generative AI (GANs)

Edmond de Belamy
from La Famille de Belamy
(USD 432,500)

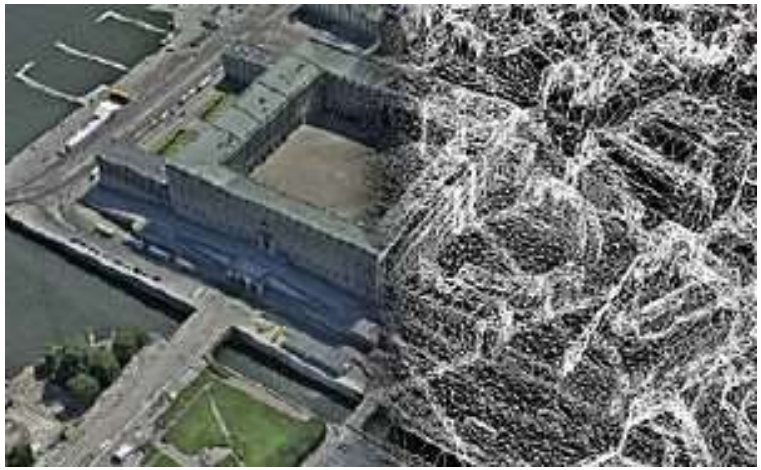
AI expert Robbie Barrat fed a neural network thousands of naked portraits

Can You Tell the Difference Between JS Bach and AI Bach?



Botnik published a 2018 Coachella Lineup poster composed entirely of performer names generated by neural networks.

3D from Images



Building Rome in a Day: Agarwal et al. 2009

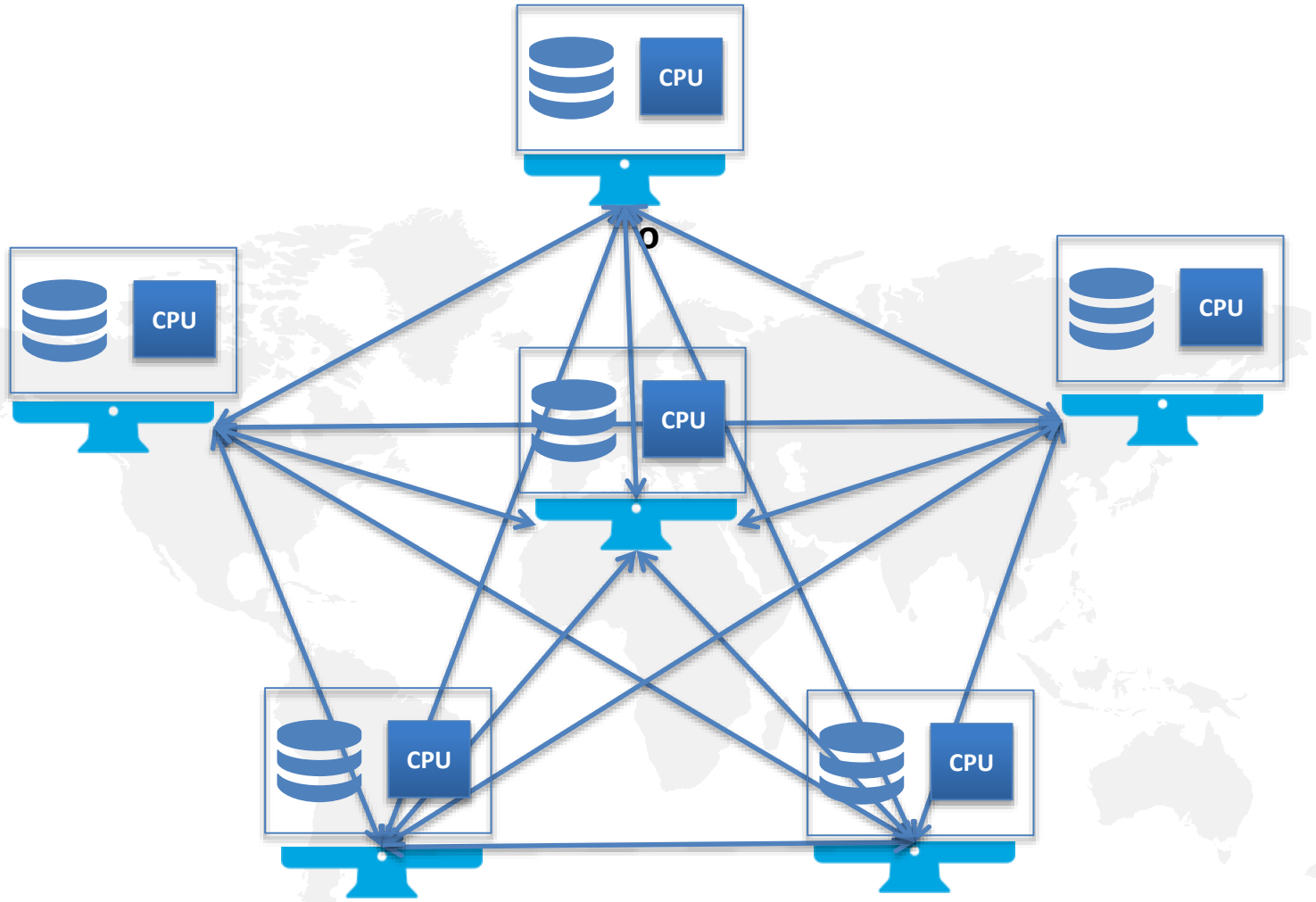
Human shape capture

Human Shape Capture

Interactive Games: Kinect

- Object Recognition: http://www.youtube.com/watch?feature=iv&v=fQ59dXOo63o
- Mario: http://www.youtube.com/watch?v=8CTJL5lUjHg
- 3D: http://www.youtube.com/watch?v=7QrnwoO1-8A
- Robot: http://www.youtube.com/watch?v=w8BmgtMKFbY

Internet of Things: 5G Network



Business Opportunities

- The number of workers in AI (deep learning) has grown at an exponential rate
- AI breaks new ground in almost every domain
- AI has revolutionized the industry and has been highly successful in solving real-world problems
- Many new businesses: surveillance, medicine, art, music, autonomous car driving, big data analysis and many more!
- New leading companies using large scale deep learning: Google, Microsoft, Facebook, Apple, Amazon, IBM, Baidu, NVIDIA, and Alibaba
- 1.8 million jobs will be eliminated by 2020, but 2.3 million new jobs will be created by then (Gartner).
- 2020 will be a pivotal year in AI-related employment dynamics, as AI will become a positive job motivator