

# DEEP LEARNING FOR VISUAL RECOGNITION

Lecture 1 – Introduction



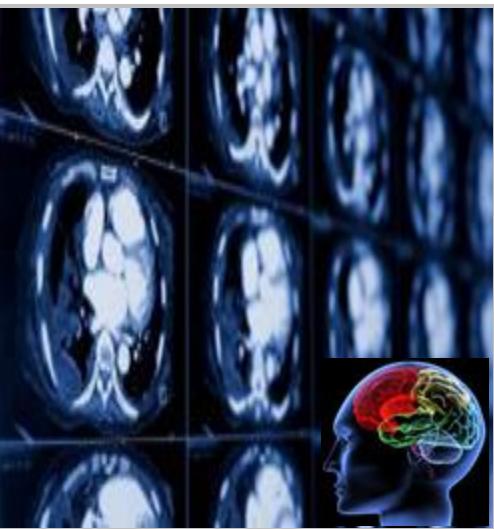
**Henrik Pedersen, PhD**  
Part-time lecturer  
Department of Computer Science  
Aarhus University  
[hpe@cs.au.dk](mailto:hpe@cs.au.dk)

# The Deep Learning Revolution

## Internet Services



## Medicine



## Media & Entertainment



## Security & Defense



## Autonomous Machines



- > Image/Video classification
- > Speech recognition
- > Chatbots

- > Cancer cell detection
- > Diabetic grading
- > Drug discovery

- > Video captioning
- > Content based search
- > Real time translation

- > Face recognition
- > Video surveillance
- > Cyber security

- > Pedestrian detection
- > Lane tracking
- > Recognize traffic signs

# Today's agenda

---

- Practical information
- Introduction to traditional computer vision
- Introduction to deep learning

# Practical information

---

# Overall structure of the course

---

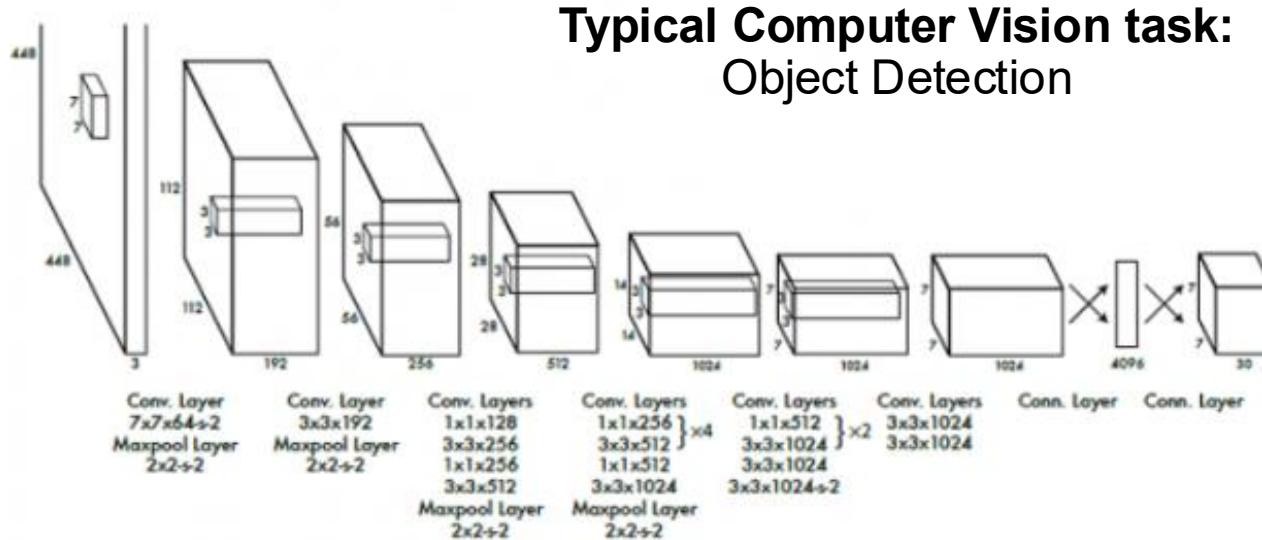
- **Lectures** on Mondays will cover deep learning theory **bottom-up** (theory first).
- **Programming exercises** on Thursdays will teach you deep learning **top-down** (practise first).
- You will learn *how* deep learning works before learning *why* it works.
- This will allow you to start working on your course project early in the semester.

# Learning goals

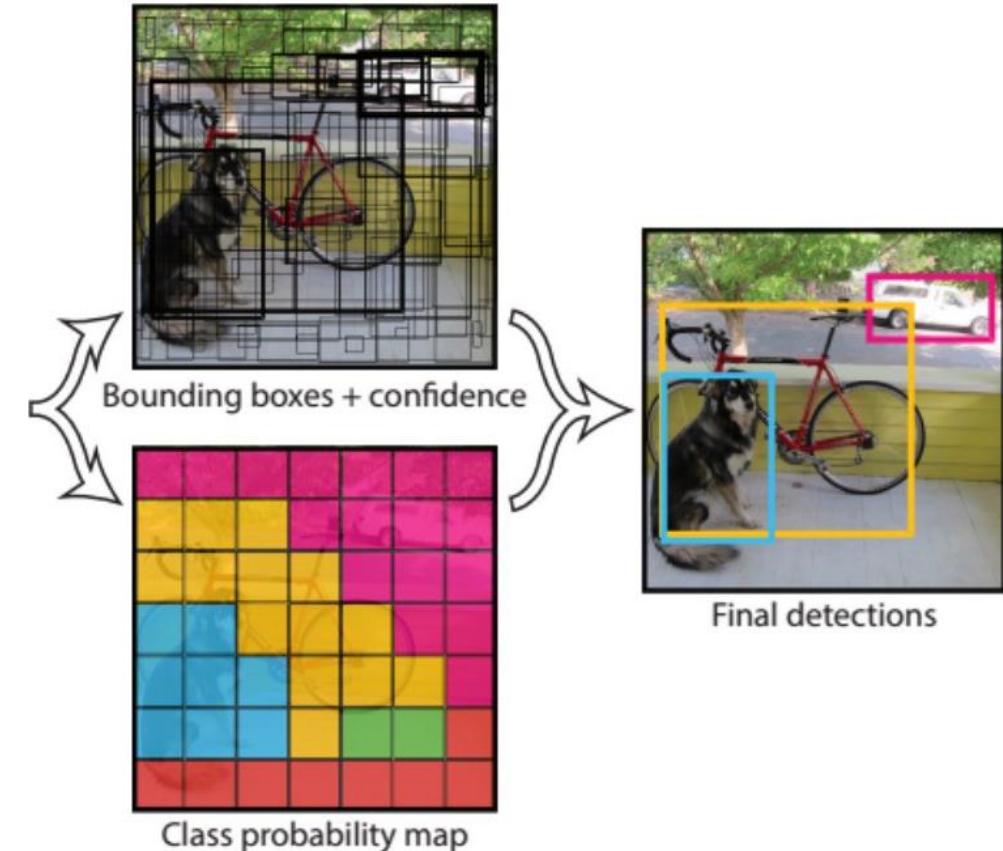
---

- Identify and describe visual recognition tasks that can be solved with deep learning.
- Describe and compare different neural networks architectures.
- Explain and compare techniques for training neural networks.
- Apply deep learning to standard visual recognition tasks and interpret the results.
- Design your own course project, implement it, perform experiments, analyse and relate the results to techniques and theories learned in class.
- Detailed learning goals are available [here](#)

# Learning goals



**Typical Computer Vision task:  
Object Detection**



- 1) Which neural network architecture(s) will solve the task?
- 2) What type of training data do you need to train the network?
- 3) Which training objective(s) are suitable for training the network?
- 4) Once trained, how do you evaluate the performance of the network?

# Resources

---

- Prerequisites
  - Basic calculus (you are expected to be able to understand the math in sections 1-3 [here](#))
  - Basic linear algebra (at the very least, you should be able to understand [chapter 2.1-2.3](#))
  - Book chapter covering preliminaries [here](#) (with code!)
- Brightspace
  - <https://brightspace.au.dk/d2l/home/183737>
  - Course schedule, syllabus, lecture slides, links to exercises, video recordings, etc.
- Slack
  - <https://dlau2025.slack.com> (invite link on Brightspace)
  - Group formation, questions about exercises, etc.
- GitHub repository (programming exercises)
  - <https://github.com/klaverhenrik/Deep-Learning-for-Visual-Recognition-2025>
- Google Colab (recommended tool to solve programming exercises)
  - <https://colab.research.google.com/>

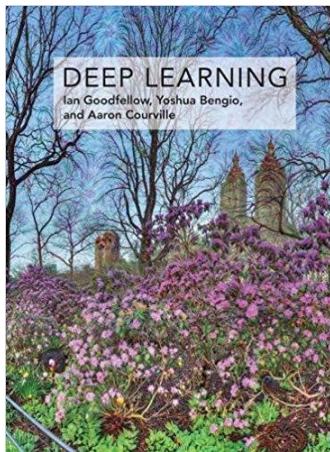
# Syllabus

---

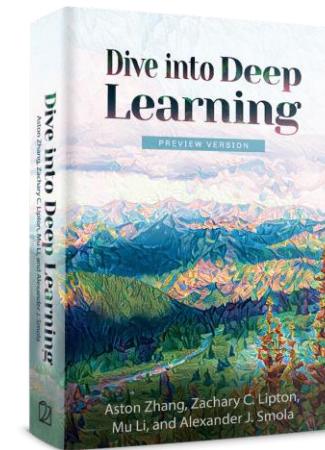
- Mostly lecture slides and research papers
- Freely available textbooks (optional)
- Lecture recordings will be uploaded at the end of the semester

<https://www.deeplearningbook.org/>

<https://github.com/janishar/mit-deep-learning-book-pdf>



<https://d2l.ai/>



# Course project

---

- Your project should involve pixels of visual data in some form.
- You must train a neural network model, evaluate it and perform experiments to improve its performance.
- When designing your project, keep the deep learning workflow in mind:
  - Acquire and prepare a dataset for deep learning
  - Set up an appropriate deep learning model (i.e., a neural network) to solve the task at hand
  - Train and test your model with the correct evaluation metrics
  - Perform motivated experiments to improve your model's performance
  - Explain your results
- Groups of 2 or 3 people (use dedicated slack channel to form groups).

# Exam

---

- 15 minutes oral exam without preparation (8 topics)
- Present on whiteboard (8 minutes)
  - Slides not allowed
- Questions relating to course theory (4 minutes)
- Grade reflects an overall assessment of project report and individual oral examination
  - Report weighs 50%
- Detailed learning goals are available [here](#)

# Questions?

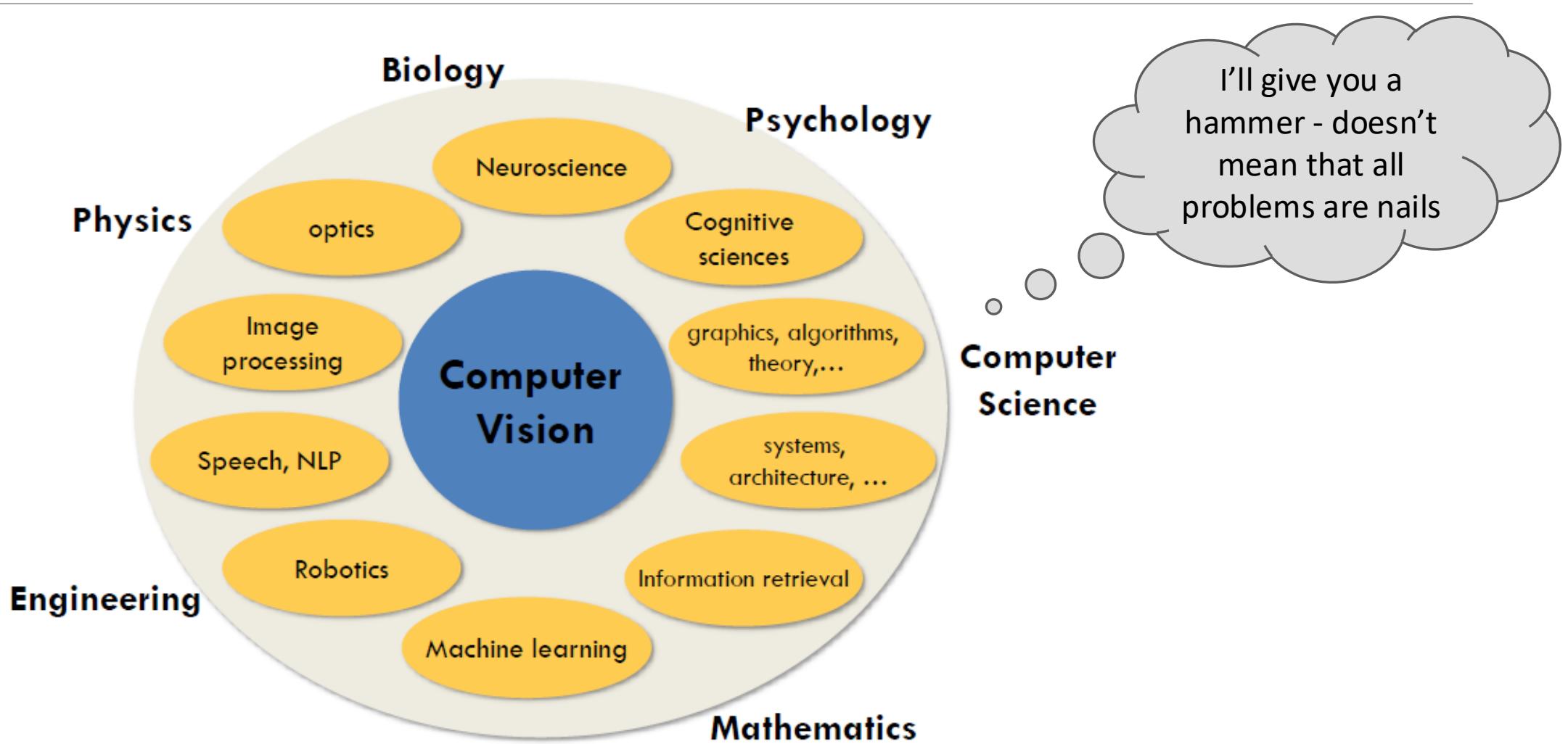
---

# Traditional computer vision

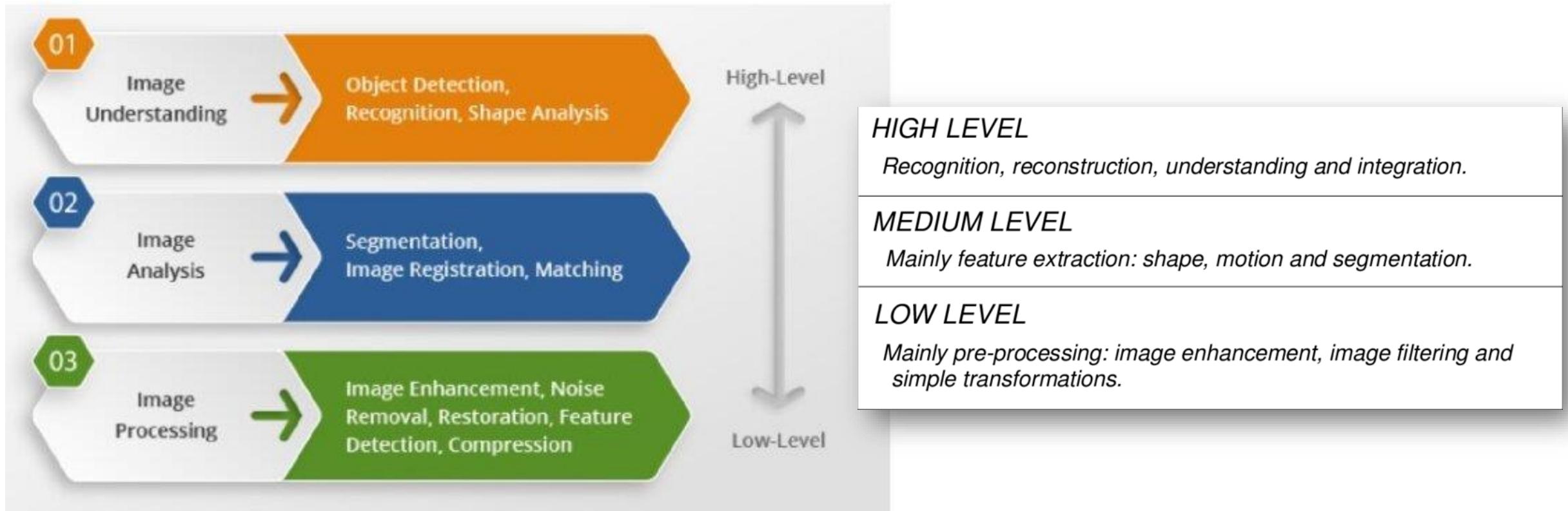
---

QUICK OVERVIEW

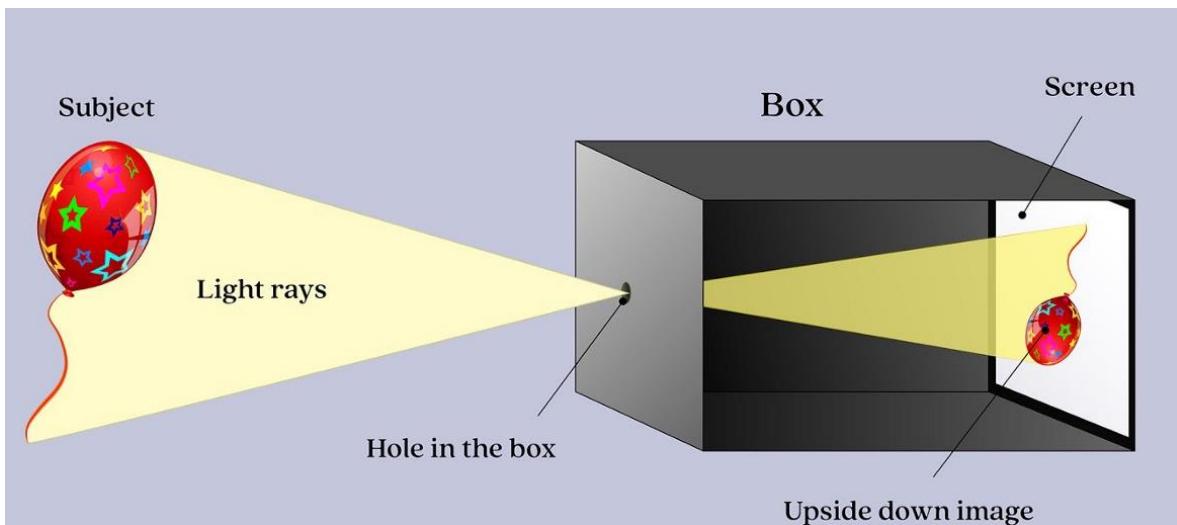
# An interdisciplinary field



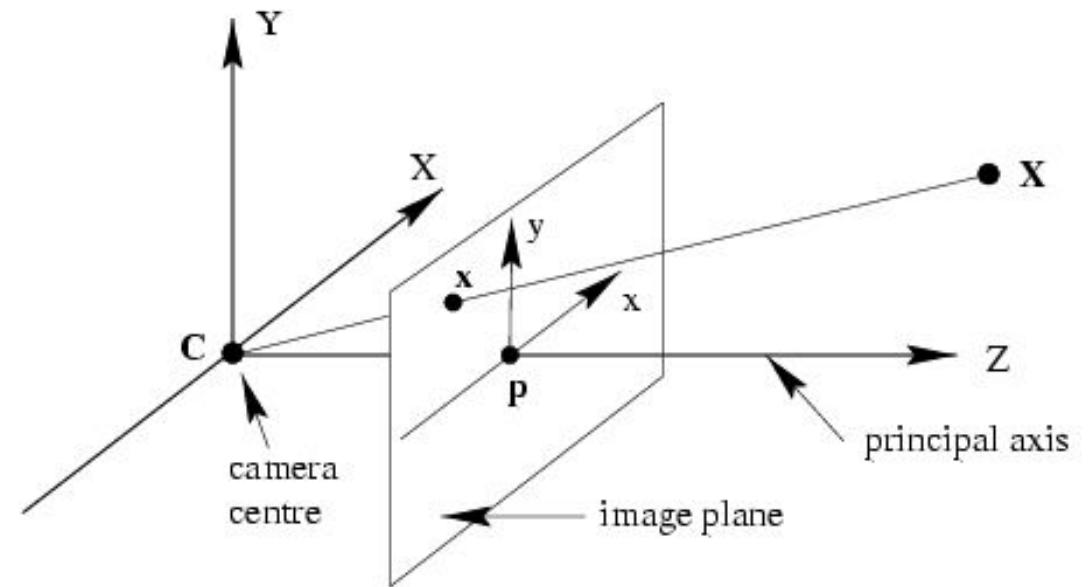
# Levels of computer vision



# The pinhole camera model



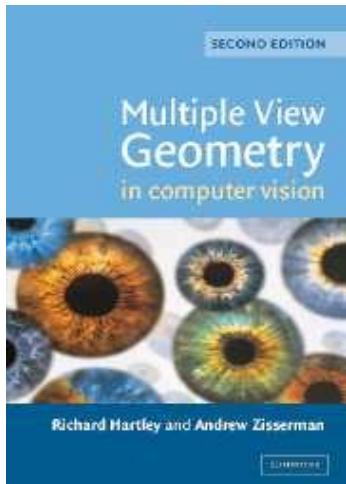
[https://en.wikipedia.org/wiki/Camera\\_obscura](https://en.wikipedia.org/wiki/Camera_obscura)



[https://docs.opencv.org/2.4/modules/calib3d/doc/camera\\_calibration\\_and\\_3d\\_reconstruction.html](https://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html)

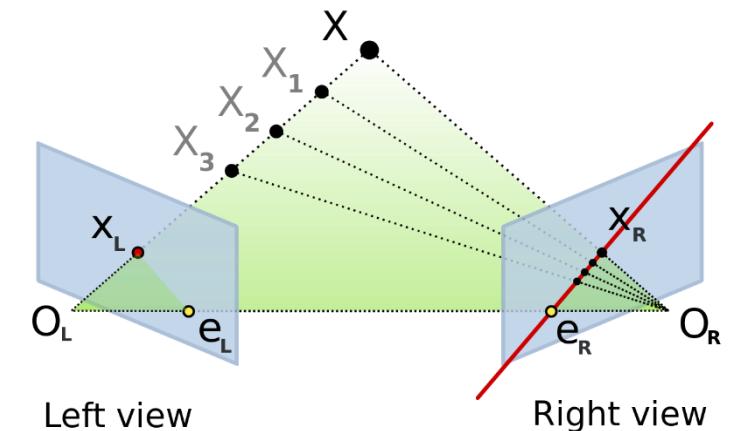
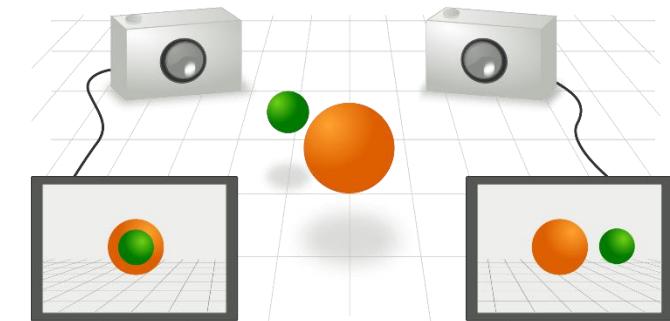
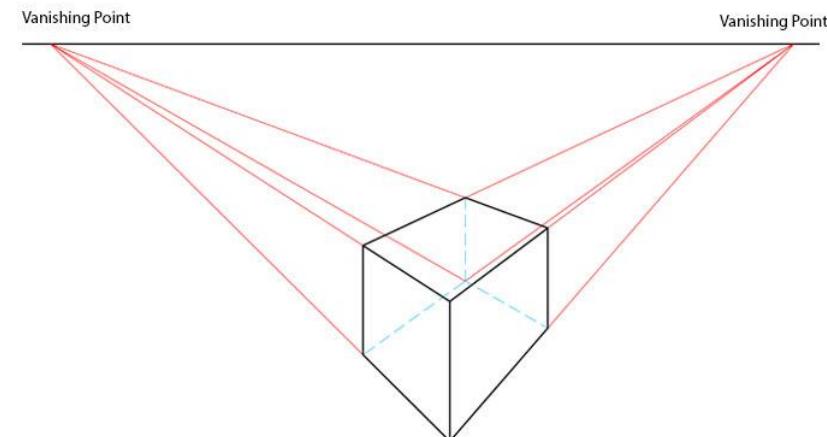
# Applications of the pinhole camera model

- Perspective drawing
- Stereo (epipolar geometry) and 3D reconstruction
- Pose estimation
- Image stitching (panoramas)



<http://www.robots.ox.ac.uk/~vgg/hzbook/>

[https://en.wikipedia.org/wiki/Perspective\\_\(graphical\)](https://en.wikipedia.org/wiki/Perspective_(graphical))



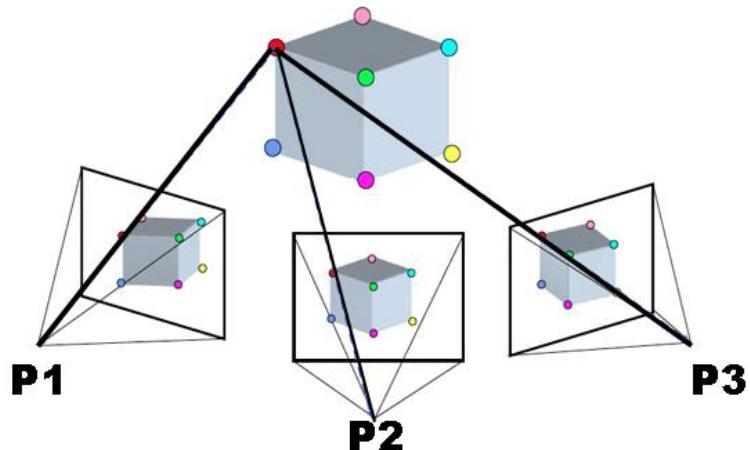
[https://en.wikipedia.org/wiki/Epipolar\\_geometry](https://en.wikipedia.org/wiki/Epipolar_geometry)

[https://en.wikipedia.org/wiki/3D\\_reconstruction](https://en.wikipedia.org/wiki/3D_reconstruction)

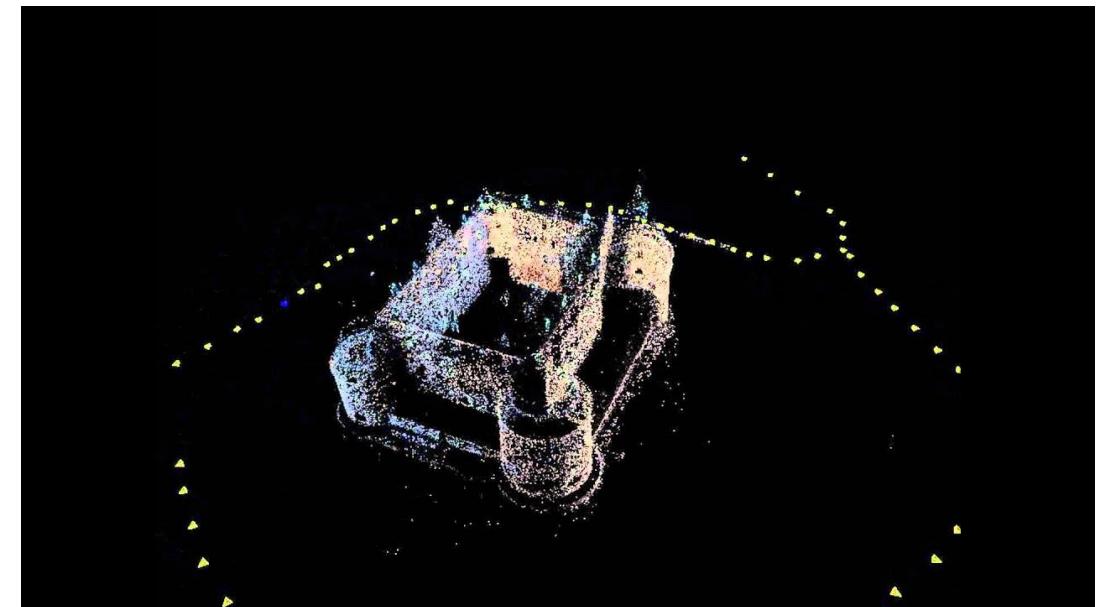
# 3D scene reconstruction

---

- Reconstruct a 3D scene from multiple images taken from different viewpoints
- Related topics: Structure-from-motion, Stereo, Camera calibration, Camera pose estimation



[https://es.wikipedia.org/wiki/Structure\\_from\\_motion](https://es.wikipedia.org/wiki/Structure_from_motion)

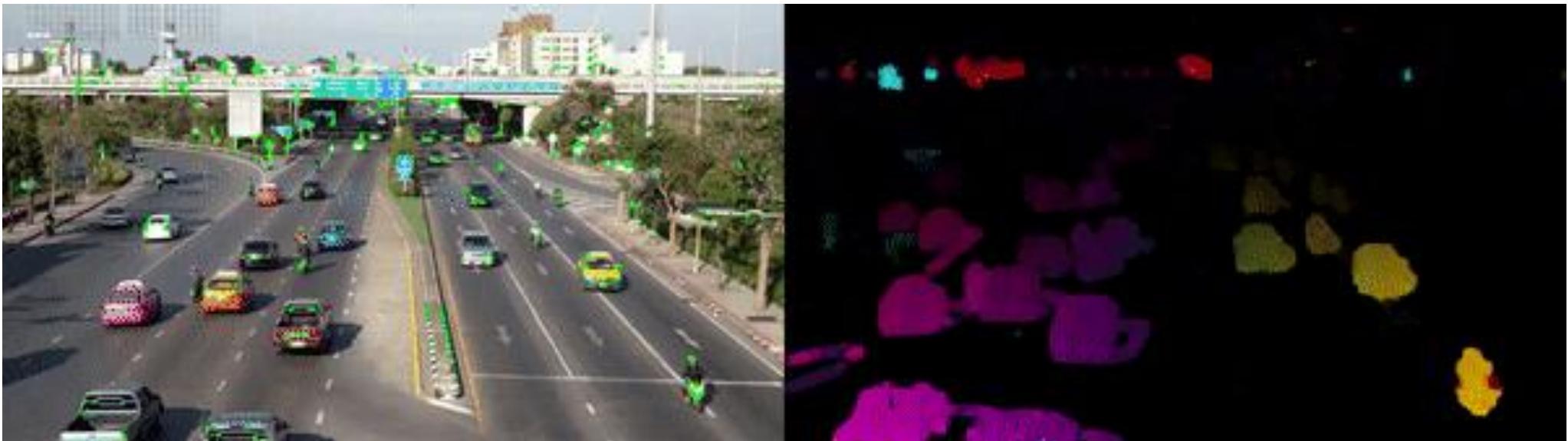


<https://www.youtube.com/watch?v=i7ierVkJya8>

# Object tracking

---

- Track moving objects in time series of images
- Related topics/synonyms: Motion estimation, Video tracking, Image registration, Optical flow



<https://blog.nanonets.com/optical-flow/>

# More examples

- Visit Papers With Code
- <https://paperswithcode.com/area/computer-vision>

Browse SoTA > Computer Vision

## Computer Vision

3136 benchmarks • 1036 tasks • 2116 datasets • 26981 papers with code

### Image Classification



#### Image Classification

348 benchmarks

2557 papers with code



#### Knowledge Distillation

3 benchmarks

618 papers with code



#### Few-Shot Image Classification

90 benchmarks

145 papers with code



#### OOD Detection

141 papers with code



#### Fine-Grained Image Classification

32 benchmarks

118 papers with code

[▶ See all 25 tasks](#)

### Semantic Segmentation



#### Semantic Segmentation

167 benchmarks

3068 papers with code



#### Tumor Segmentation

1 benchmark

133 papers with code



#### 3D Semantic Segmentation

11 benchmarks

103 papers with code



#### Panoptic Segmentation

13 benchmarks

99 papers with code



#### Weakly-Supervised Semantic Segmentation

3 benchmarks

84 papers with code

[▶ See all 20 tasks](#)

### Object Detection



#### Object Detection

242 benchmarks

2335 papers with code



#### 3D Object Detection

55 benchmarks

308 papers with code



#### RGB Salient Object Detection

25 benchmarks

87 papers with code



#### Real-Time Object Detection

9 benchmarks

81 papers with code



#### RGB-D Salient Object Detection

8 benchmarks

50 papers with code

[▶ See all 33 tasks](#)

# Towards image understanding

---

- **Example:** Image classification
- Given an image, tell which in a number of classes it belongs to
  - What type of object?
  - Quality control: OK or needs manual inspection?
  - View direction: frontal, side, top, bottom?

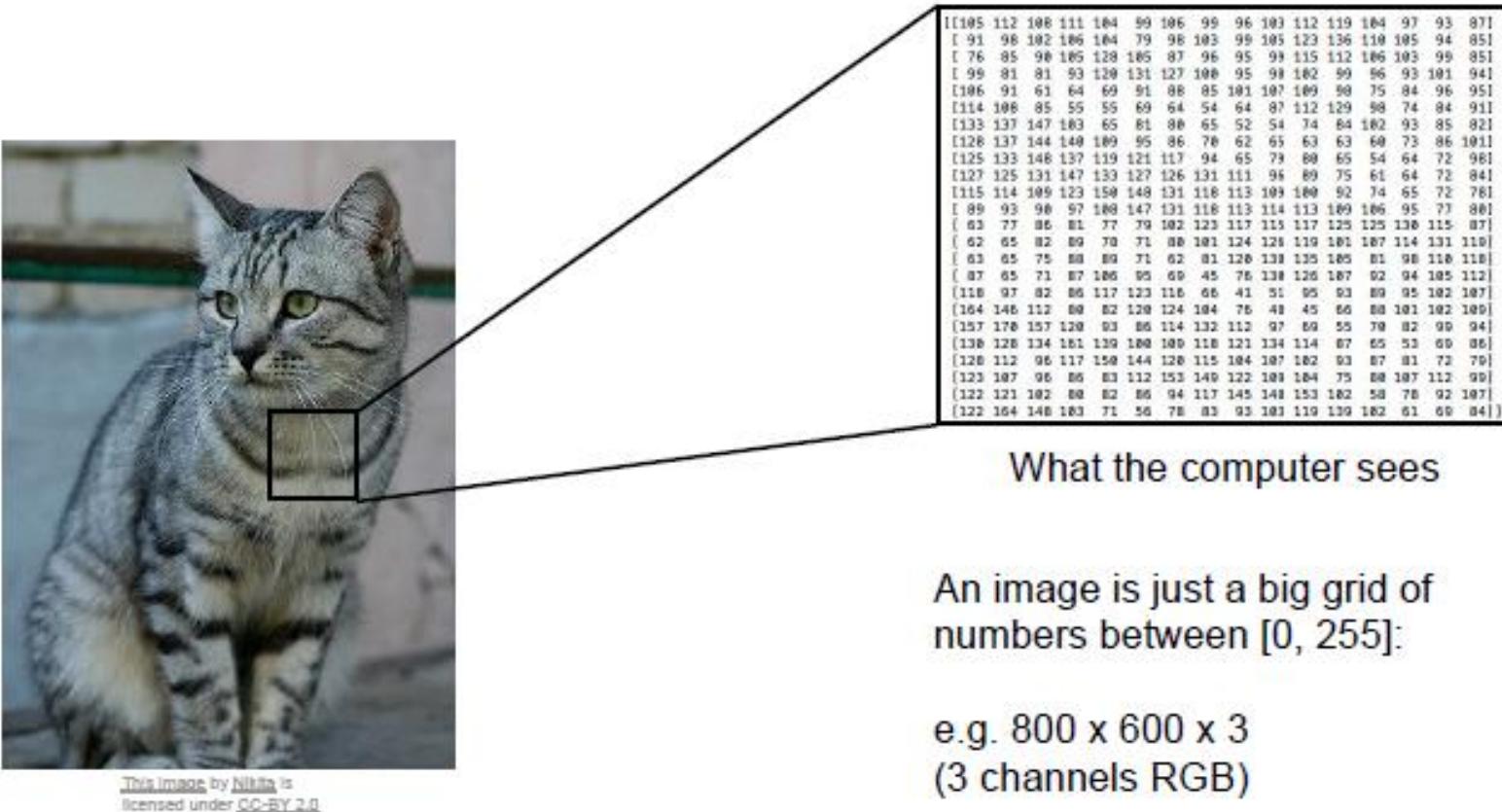


(assume given set of discrete labels)  
{dog, cat, truck, plane, ...}

→ cat

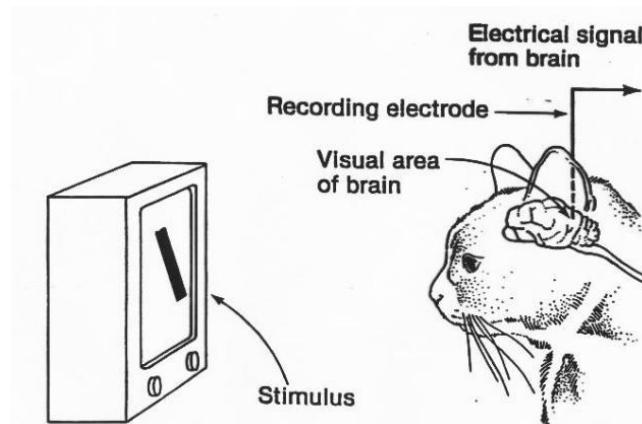
# Main problem – semantic gap

- Computers don't see the way humans do



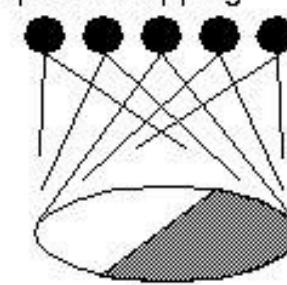
# Inspiration from neuroscience

- Famous experiment by Hubel and Weisel
- Feature hierarchy: Neurons along the visual pathway extract increasingly complex information from the pattern of light cast on the retina to construct and understand an image.
- Example: Edges → Corners → Boxes → Object parts → Whole objects

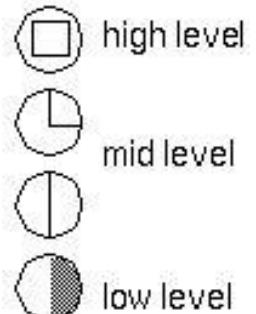
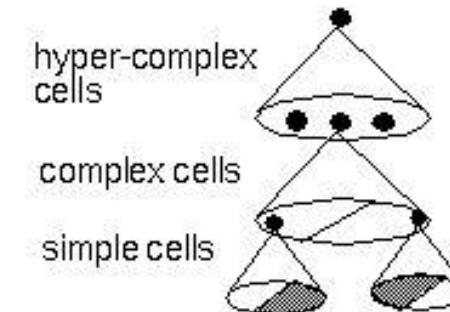


**Hubel & Weisel**

topographical mapping



**featural hierarchy**



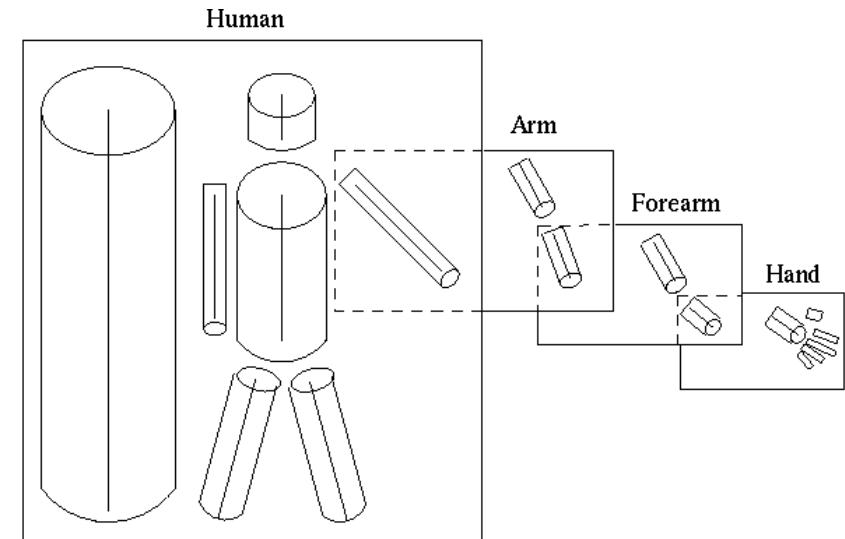
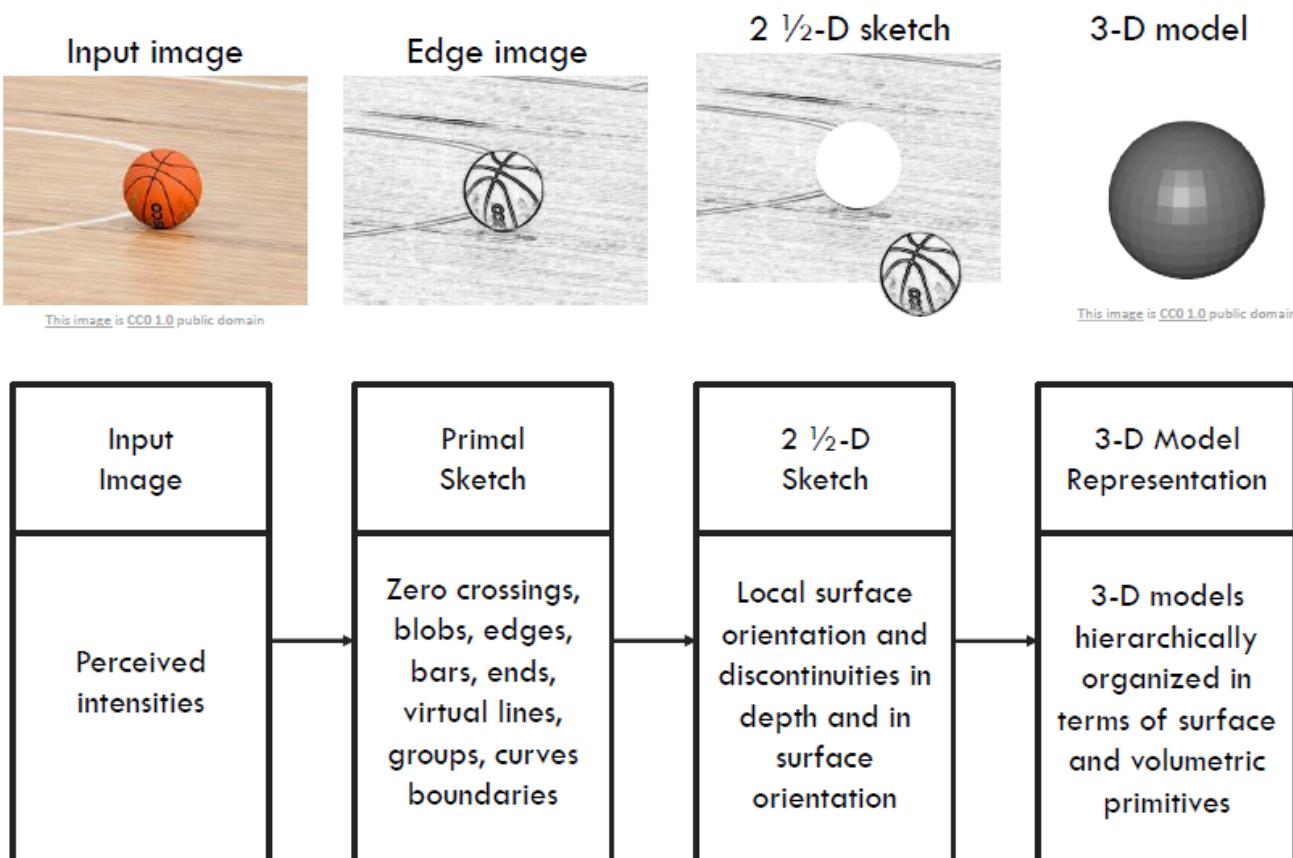
<https://goodpsychology.wordpress.com/2013/03/13/235/>

<https://www.youtube.com/watch?v=Cw5PKV9Rj3o>

<http://cns-alumni.bu.edu/~slehar/webstuff/pcave/hubel.html>

<https://knowingneurons.com/2014/10/29/hubel-and-wiesel-the-neural-basis-of-visual-perception/>

# Early computer vision models



David Marr, 1970s

*edges, bars, ends* and *blobs* were represented by a 5-tuple: (*type, position, orientation, scale, contrast*).

**All steps are hard-coded**

# Early computer vision models

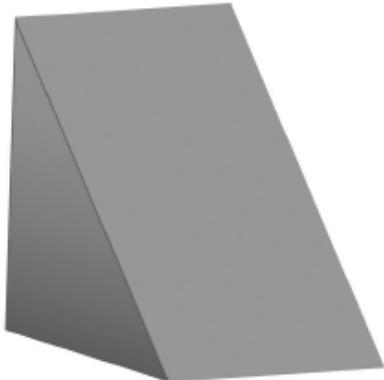


## Blocks World

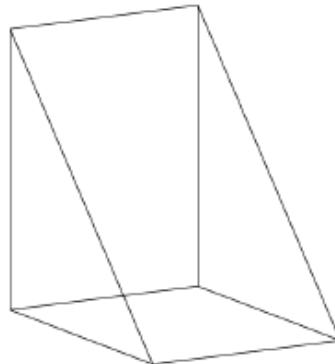
Lawrence Roberts

Massachusetts Institute of Technology, 1963

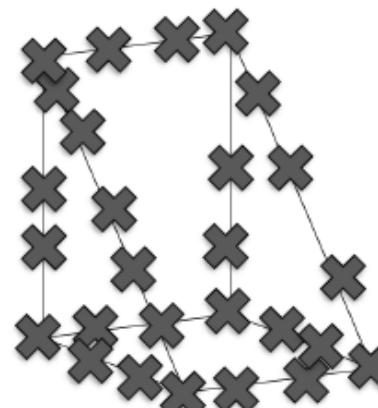
**Extracting 3D information about solid objects from 2D photographs of line drawings.**



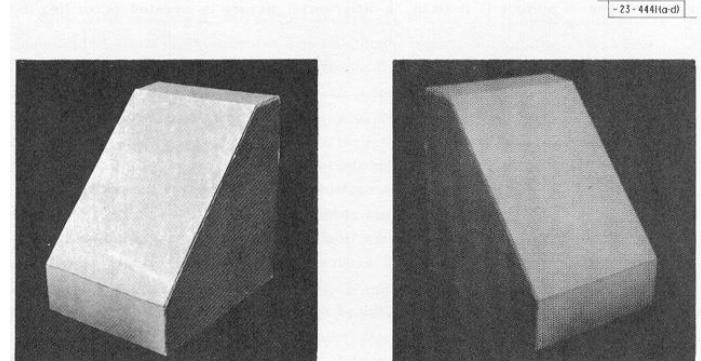
(a) Original picture



(b) Differentiated picture

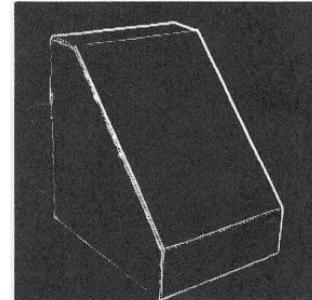


(c) Feature points selected

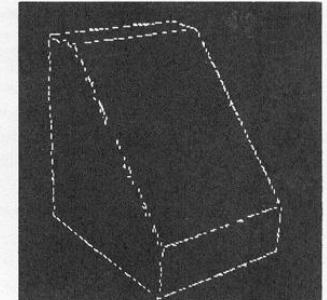


(a) Original picture.

(b) Computer display of picture  
(reflected by mistake).



(c) Differentiated picture.



(d) Feature points selected.

**Limited application in the real world**

# Challenges

---

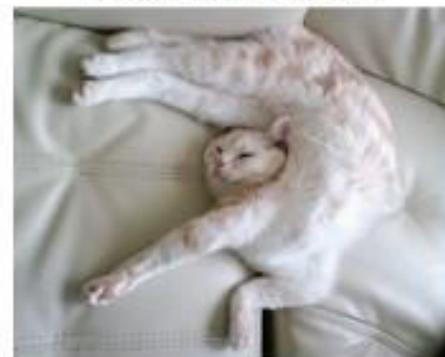
Viewpoint variation



Scale variation



Deformation



Occlusion



Illumination conditions



Background clutter



Intra-class variation



# Challenges

---

Viewpoint variation



Scale variation



Deformation



Occlusion

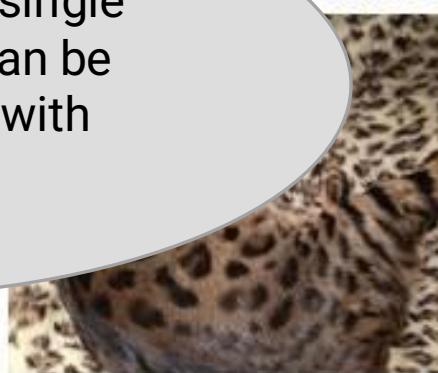


Illumination conditions



**Viewpoint variation.** A single instance of an object can be oriented in many ways with respect to the camera.

Clutter and clutter



Intra-class variation



# Challenges

Viewpoint variation



Illumination conditions



Scale variation



**Scale variation.** Visual classes often exhibit variation in their size (size in the real world, not only in terms of their extent in the image).



Background clutter



Intra-class variation



# Challenges

---

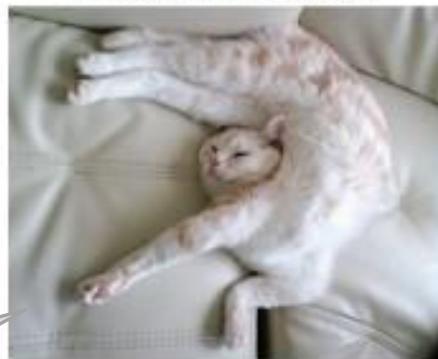
Viewpoint variation



Scale variation



Deformation



Occlusion



Illumination



**Deformation.** Many objects of interest are not rigid bodies and can be deformed in extreme ways.

Background clutter



Intra-class variation



# Challenges

Viewpoint variation



Scale variation



Deformation



Occlusion



Illumination conditions



**Occlusion.** The objects of interest can be occluded. Sometimes only a small portion of an object (as little as few pixels) could be visible.

Clutter



Intra-class variation



# Challenges

---

Viewpoint variation



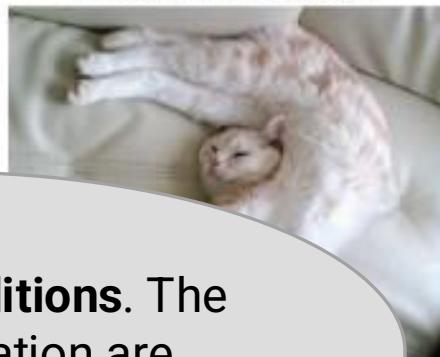
Illumination conditions



Scale variation



Deformation



Occlusion



**Illumination conditions.** The effects of illumination are drastic on the pixel level.



Intra-class variation



# Challenges

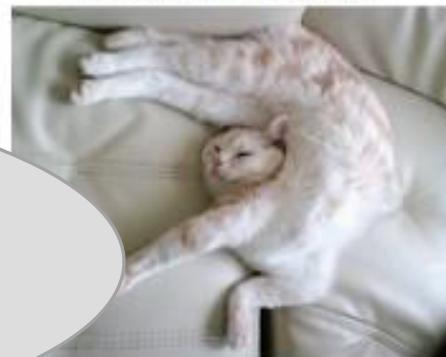
Viewpoint variation



Scale variation



Deformation



Occlusion



**Background clutter.** The objects of interest may *blend* into their environment, making them hard to identify.

Illumination conditions



Background clutter



Intra-class variation



# Challenges

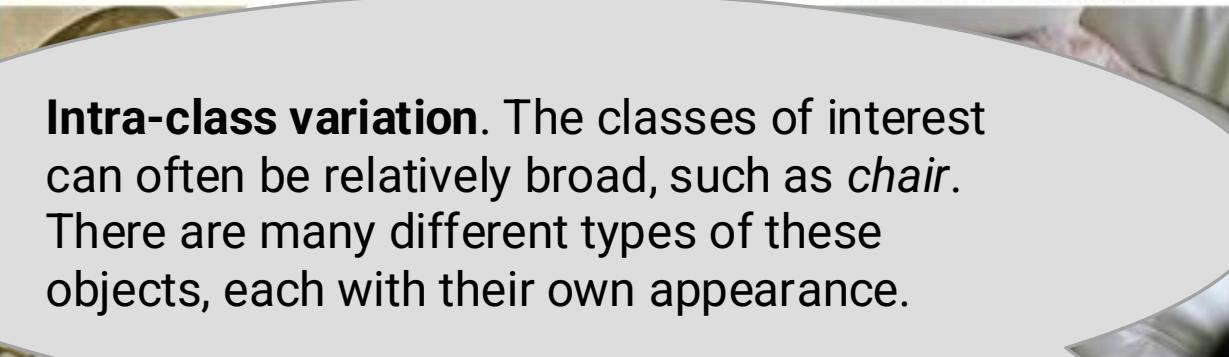
Viewpoint variation



Scale variation



Deformation



Occlusion

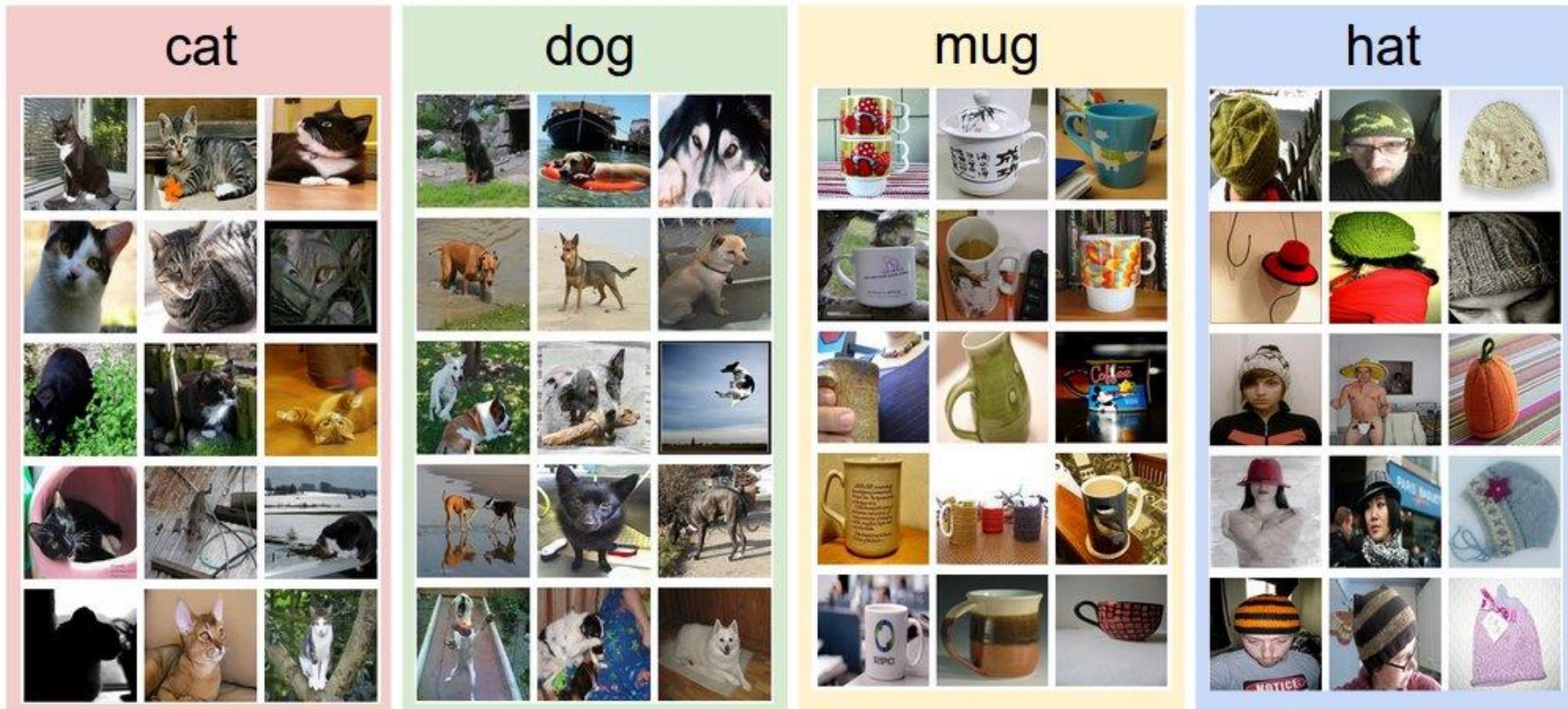


Illumination conditions



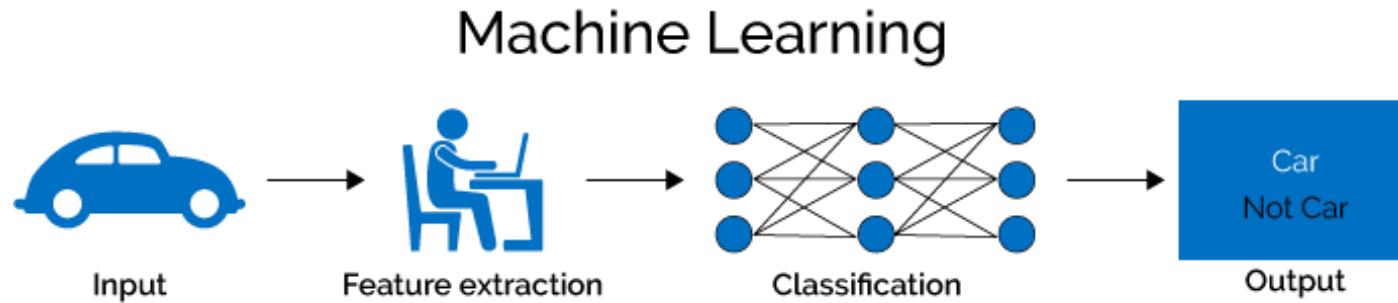
# Data-driven approach

- No obvious way to hard-code the algorithm for recognizing a cat, a dog, or anything else
- Our approach: Learn from a **training dataset** of labelled images.



# Basic idea

---



## Data-driven approach using traditional machine learning

- Extract simple features like edges and blobs from the input image
- Feed those features into a machine learning model
- Train the machine model to solve the task at hand

Note: Under the hood, the machine learning model learns to combine the simple features into more complex features (like in the brain).

# Image features

---

Edges = rapid changes in image intensity, color or texture

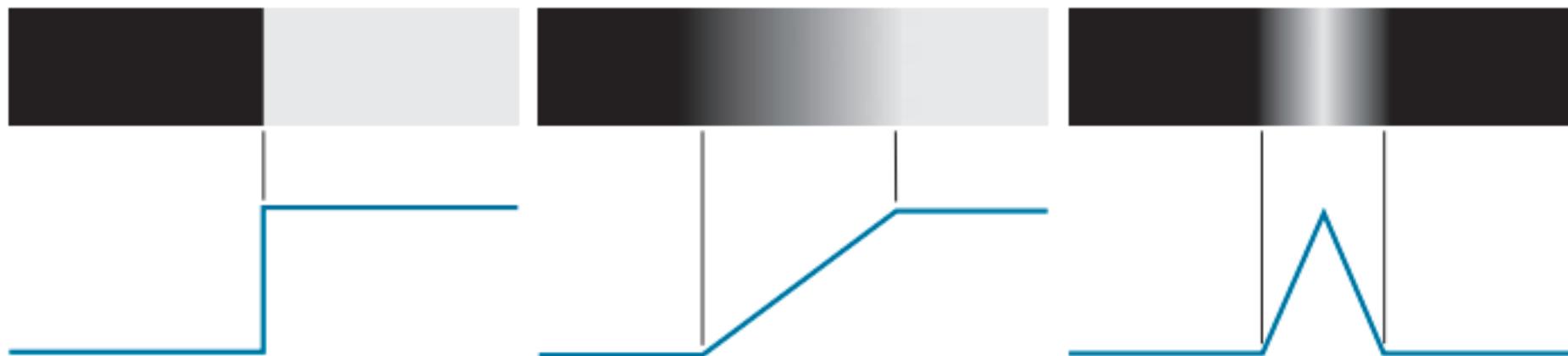


Source: [https://en.wikipedia.org/wiki/Edge\\_detection](https://en.wikipedia.org/wiki/Edge_detection)

# Image features

---

Edges = rapid changes in image intensity, color or texture

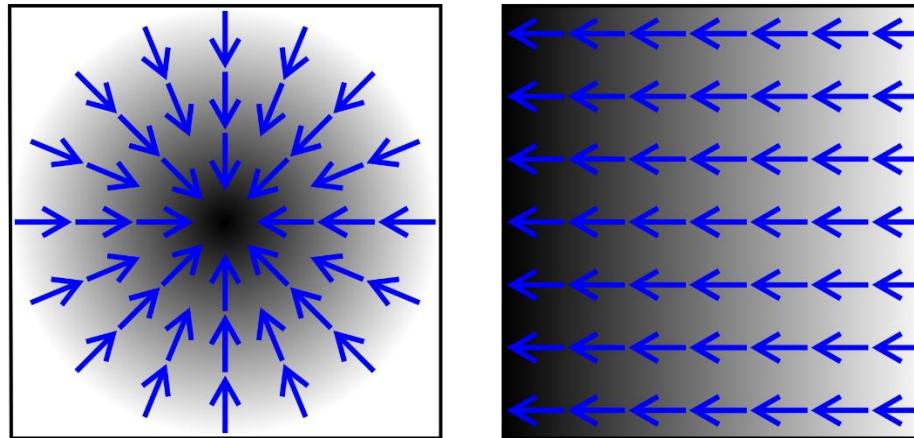


Edges can have different intensity profiles.  
How do we detect them?  
(Answer: convolution)

# Image features

---

Edges = rapid changes in image intensity, color or texture



[https://en.wikipedia.org/wiki/Image\\_gradient](https://en.wikipedia.org/wiki/Image_gradient)

**Note:** We can use image gradients to determine the direction of edges.

# Image features

## Histograms of oriented gradients (HoG)



Input  
example



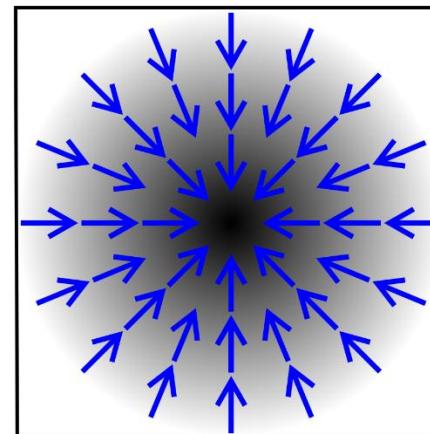
Average  
gradients



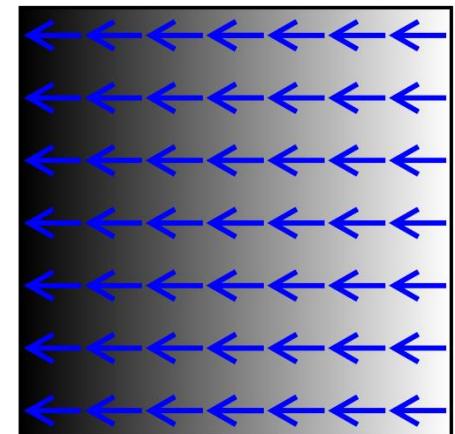
Weighted  
pos wts



Weighted  
neg wts



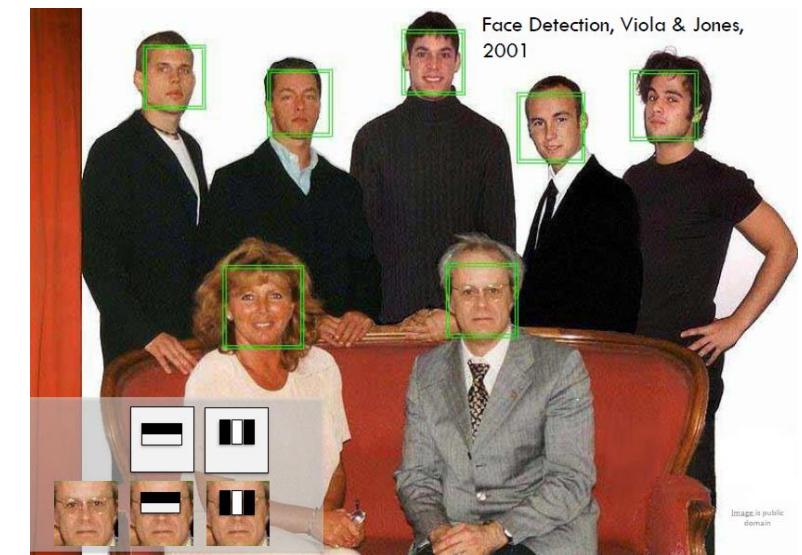
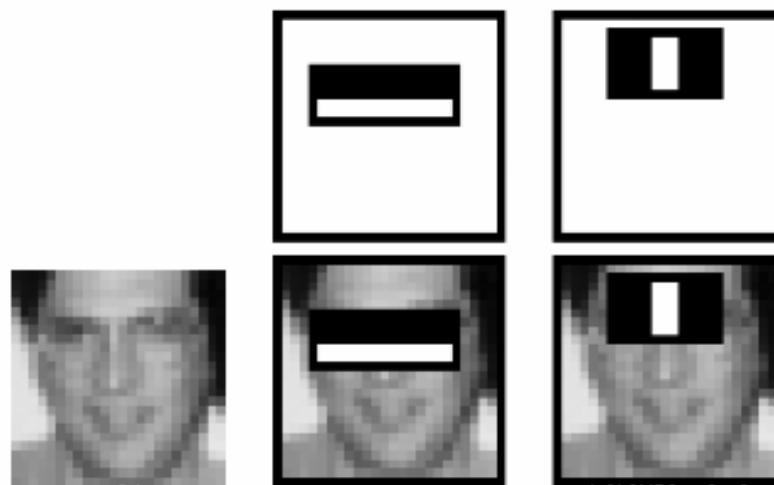
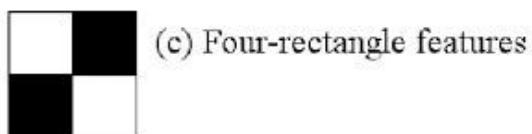
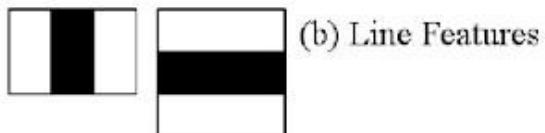
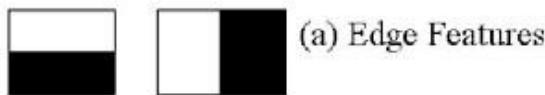
[https://en.wikipedia.org/wiki/Image\\_gradient](https://en.wikipedia.org/wiki/Image_gradient)



[http://vision.stanford.edu/teaching/cs231b\\_spring1213/papers/CVPR05\\_DalalTriggs.pdf](http://vision.stanford.edu/teaching/cs231b_spring1213/papers/CVPR05_DalalTriggs.pdf)

# Image features

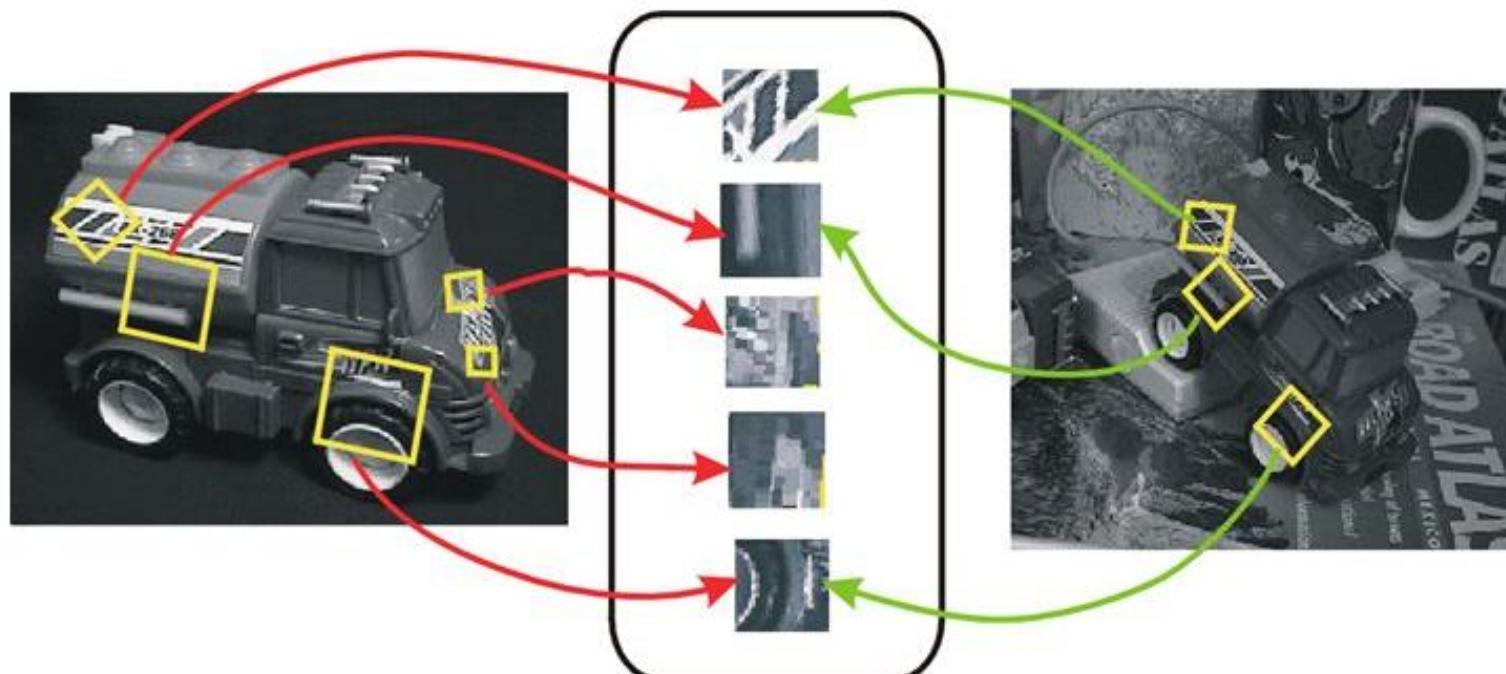
## More advanced structures (filters)



[https://docs.opencv.org/3.4.1/d7/d8b/tutorial\\_py\\_face\\_detection.html](https://docs.opencv.org/3.4.1/d7/d8b/tutorial_py_face_detection.html)

# Image features

Interest points and descriptors (e.g., SIFT)



<https://medium.com/software-incubator/introduction-to-sift-scale-invariant-feature-transform-65d7f3a72d40>

<https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf>

# Image features

---

Cluster similar pixels (e.g., similar color) into so-called super pixels

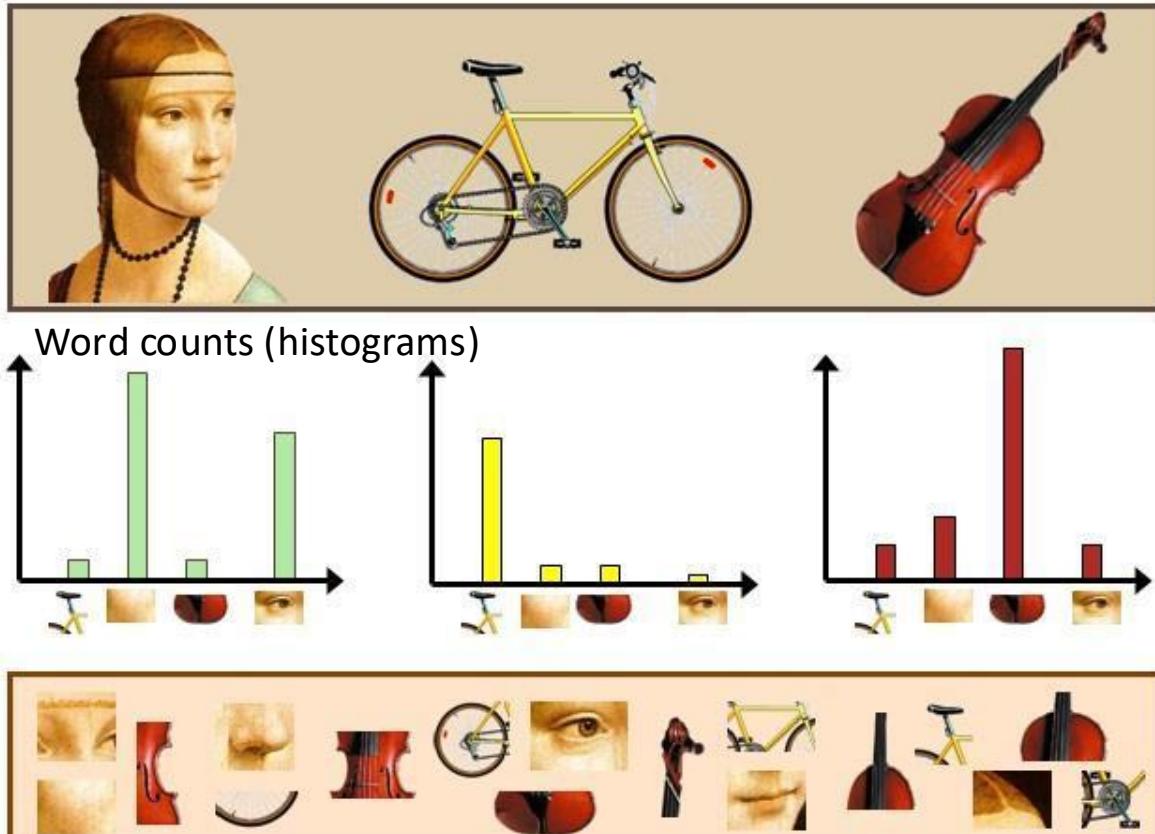


# Object detection and recognition

---

- Combine information from image features
- Two examples:
  - Bag of words
  - Part-based model

# Bag of words



Bag of visual words (= **image features**)

<https://prateekvjoshi.com/2014/08/17/image-classification-using-bag-of-words-model/>

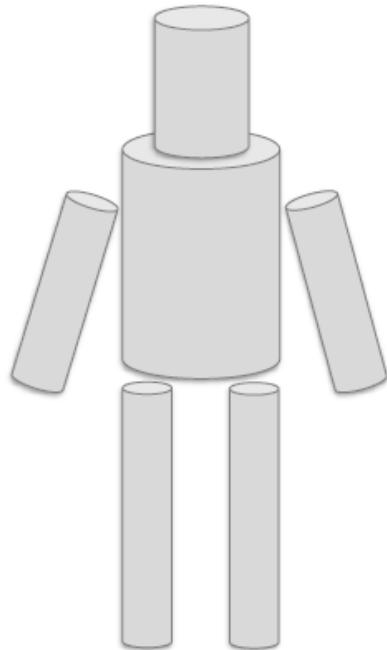
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach our brain from our eyes. For a long time it was believed that the visual image was formed in the visual centers in the brain. In 1959, in a movie showing the visual pathway, it was discovered that the visual image is formed in the retina. It was known that the visual system undergoes a complex analysis in the brain. Following the work of Hubel and Wiesel, who studied the visual pathway in the cat's brain, it was demonstrated that the message about the visual image falling on the retina undergoes a complex analysis in a system of nerve cells stored in columns. In this system each column has its specific function and is responsible for a specific detail in the pattern of the retinal image.

# Part-based models

---

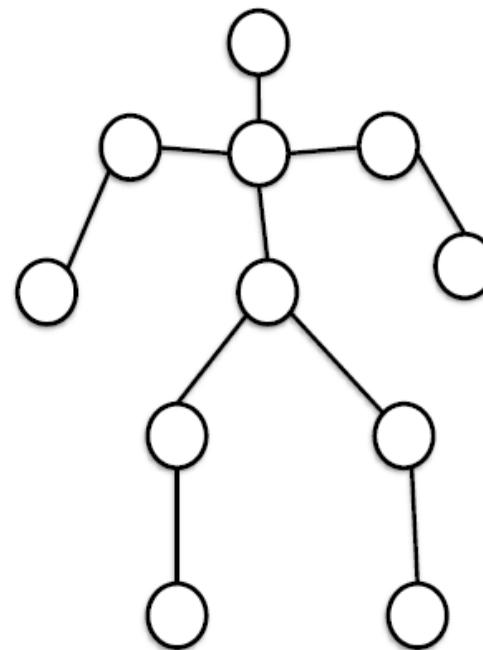
- Generalized Cylinder

Brooks & Binford, 1979



- Pictorial Structure

Fischler and Elschlager, 1973



Recall how the brain works:

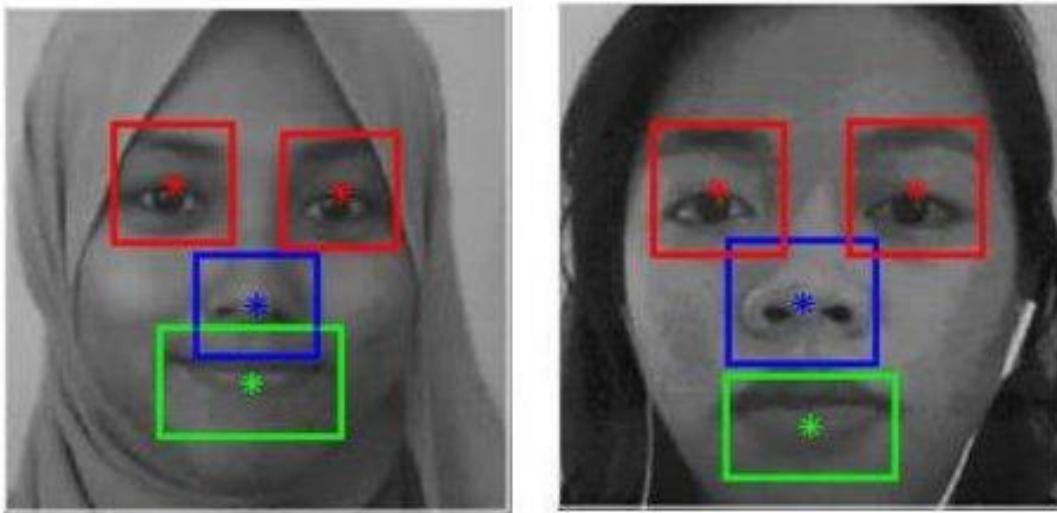
Look for simple features like edges and blobs, then combine into more complex features like corners.

Part-based models can be thought of as an implementation of this. Basic idea is to combine “image features” into object parts.

Object parts are then connected based on tree-like data structures (or graphs).

# Part-based models

---



## Model training:

- Learn from examples how individual parts look (e.g., eyes, nose, etc.)
- Learn from examples how parts are spatially organized (e.g., one eye on each side of the nose)

## Model inference:

- Look for candidate parts
- Is there a spatial configuration of the candidates that is consistent with the object you are looking for?

<http://rogerioferis.com/VisualRecognitionAndSearch2014/material/papers/DPMCVPR98.pdf>

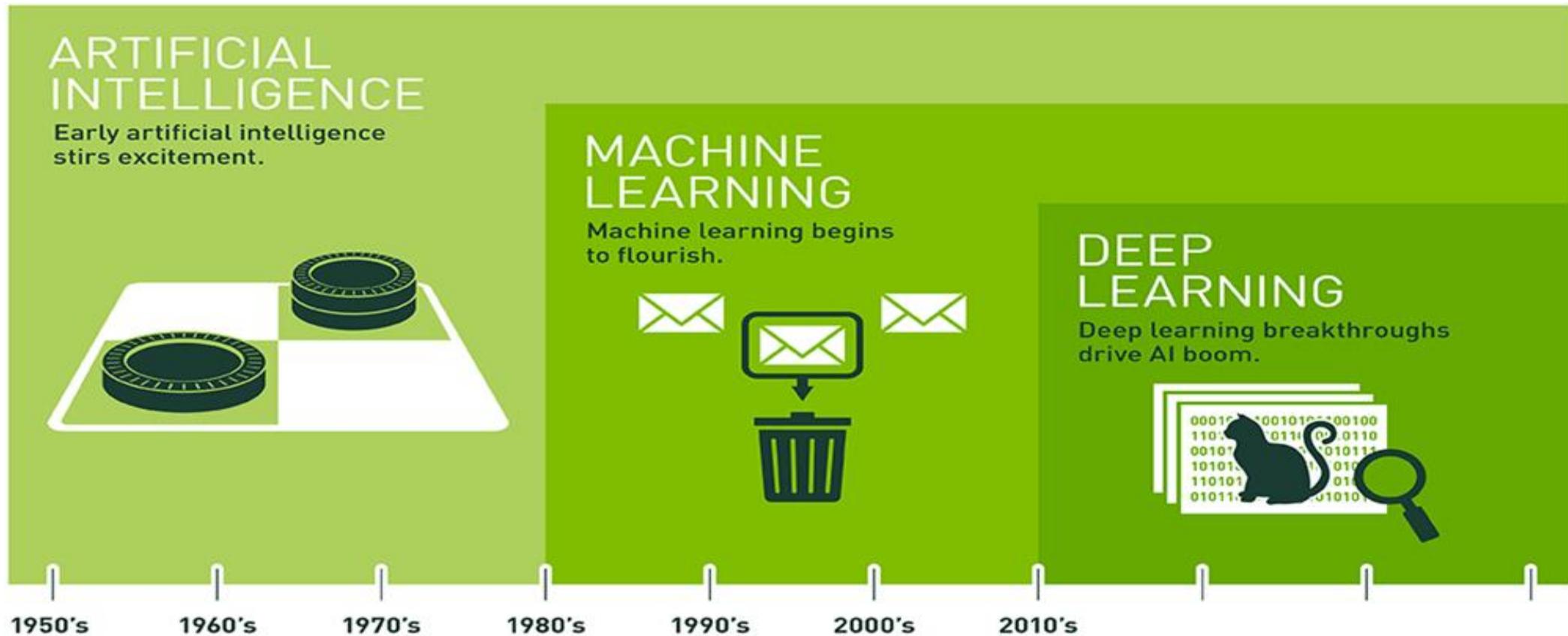
# Deep Learning

---

INTRODUCTION

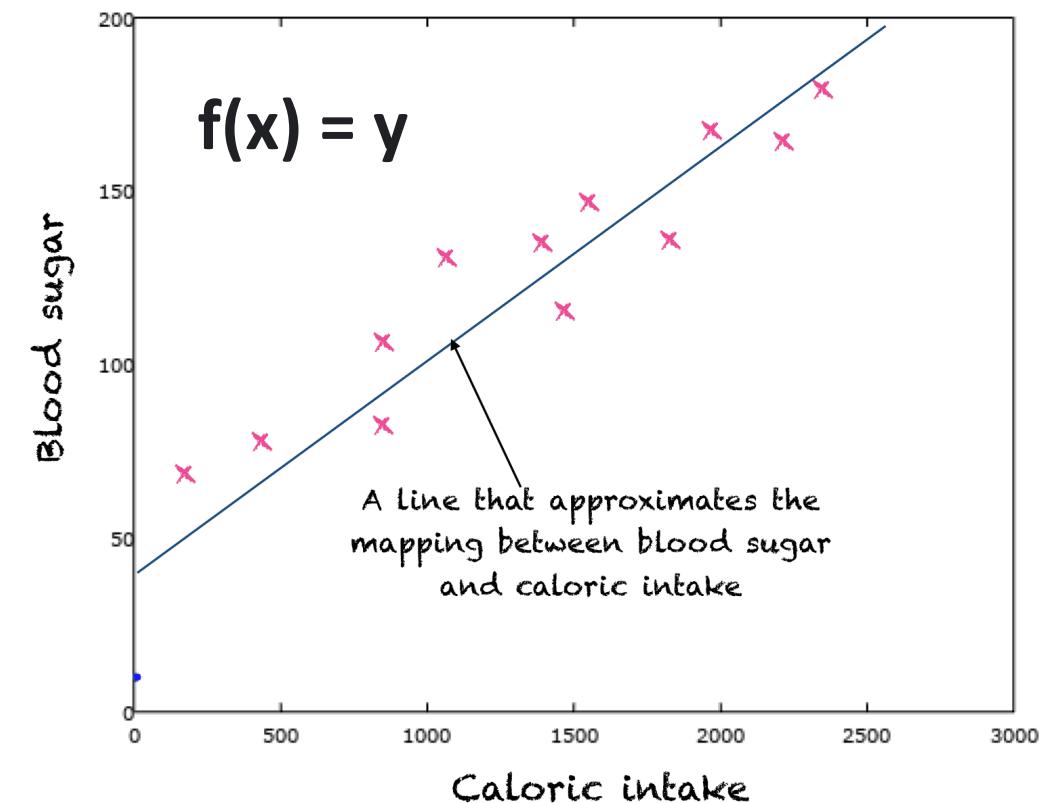
# Definitions

---



# Machine Learning in a nutshell

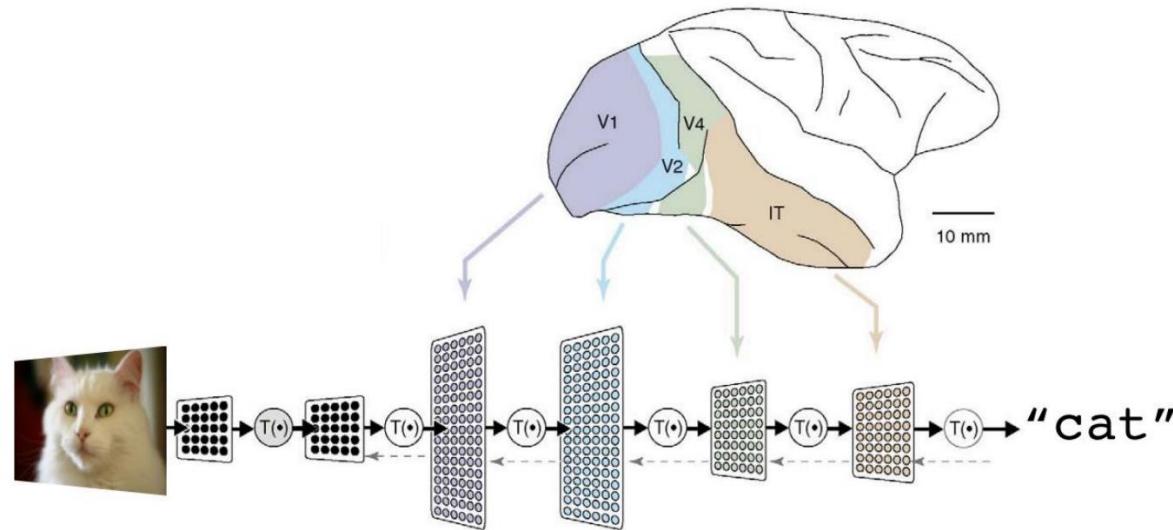
- Functions describe the world – everything you can think of is described by functions!
- A function is just a system of inputs ( $x$ ) and outputs ( $y$ ).
- If you know the function  $f$ , you can always calculate the correct output  $y$ , given any input  $x$ .
- Say we know some of the inputs ( $x$ ) and outputs ( $y$ ), but we don't know the function used to produce them.
- Is there a way to “reverse engineer” that function?
- If we could construct such a function, we could use it to calculate a  $y$ -value, given an  $x$ -value that is not in our original dataset.
- What we need is a function approximation or more generally, a function approximator.
- Fundamentally, machine learning models are function approximators.
- Neural networks are universal function approximators (i.e., they can approximate any function).



<https://machinelearningmastery.com/a-gentle-introduction-to-approximation>

# What is deep learning?

- Neural networks are machine learning algorithms inspired by the structure and function of the brain.
- They have been around since the 1970s, but the deep learning revolution started around 10 years ago mainly due to **Big data** (huge amounts of *labelled* data!) + **GPUs** (Graphics Processing Units).

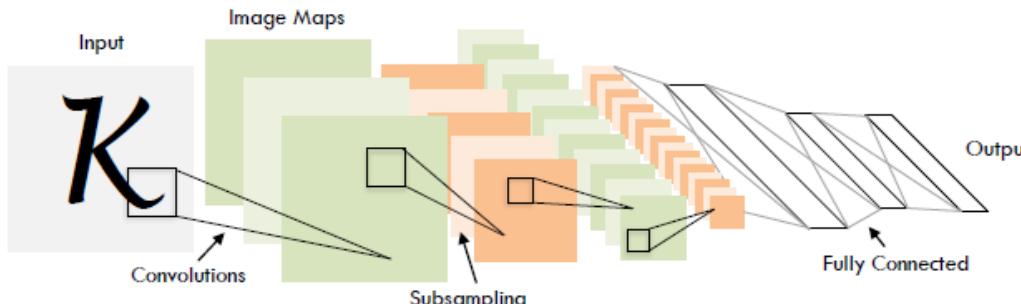


A deep neural network consists of a **hierarchy of layers**, whereby each layer **transforms the input data** into more abstract representations (e.g. edge -> nose -> face). The output layer combines those features to make predictions.

# Convolutional Neural Networks

1998

LeCun et al.



# of transistors



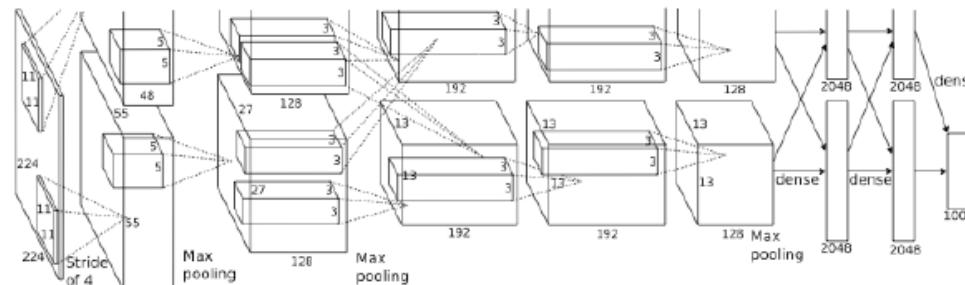
$10^6$

# of pixels used in training

$10^7$  NIST

2012

Krizhevsky et al.



# of transistors



$10^9$

GPUs



# of pixels used in training

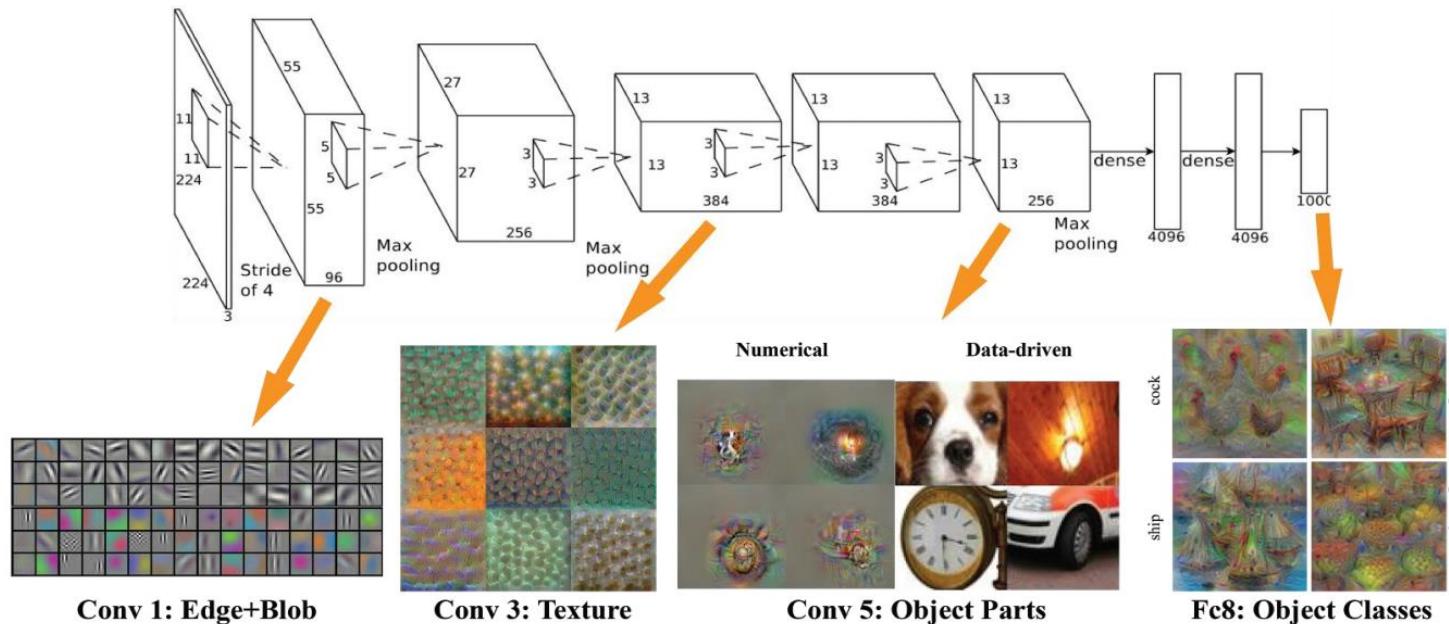
$10^{14}$  IMAGENET

Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

<http://yann.lecun.com/exdb/publis/pdf/lecun-98.pdf>

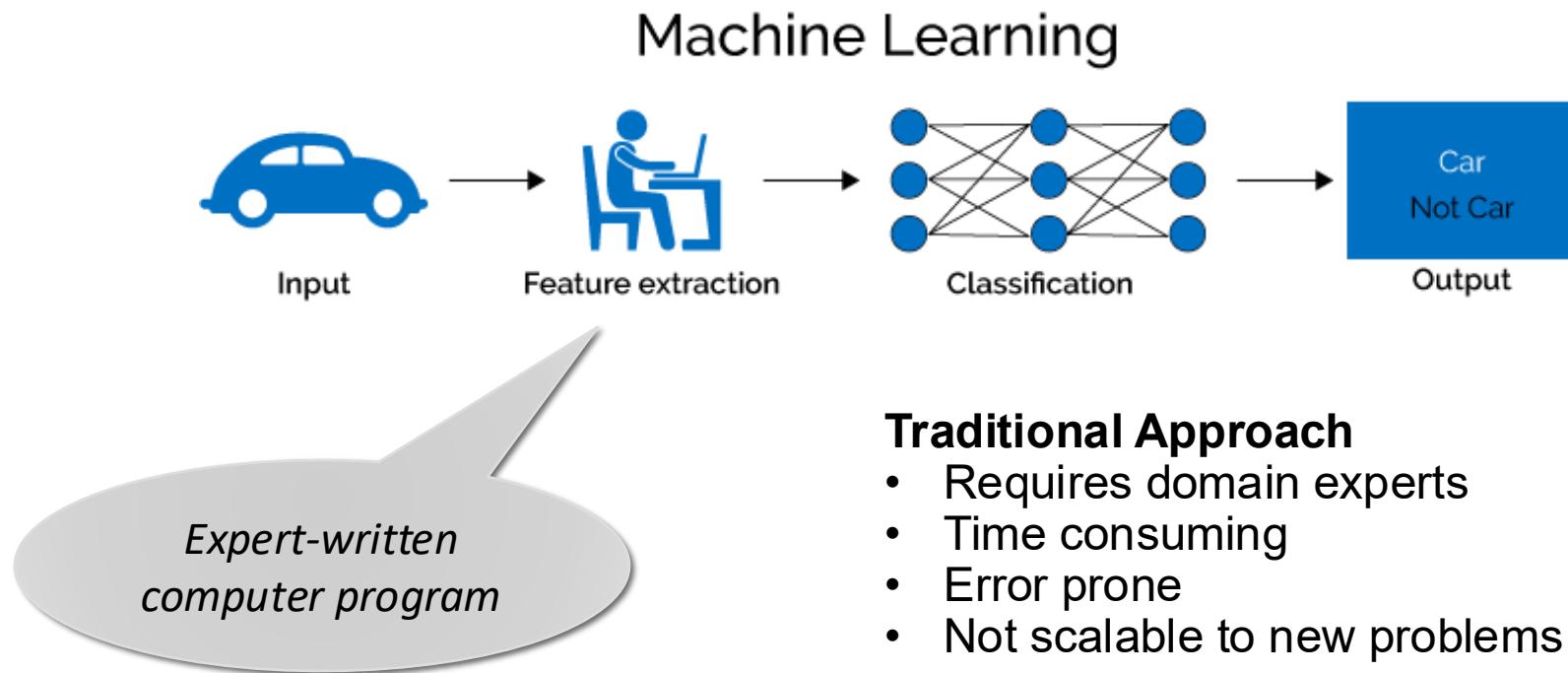
# The Big Bang of Deep Learning

- AlexNet is the name of a specific neural network architecture that won the ImageNet Large Scale Visual Recognition Challenge in 2012.
- The network achieved a top-5 error of 15.3%, more than 10.8 percentage points ahead of the runner up.



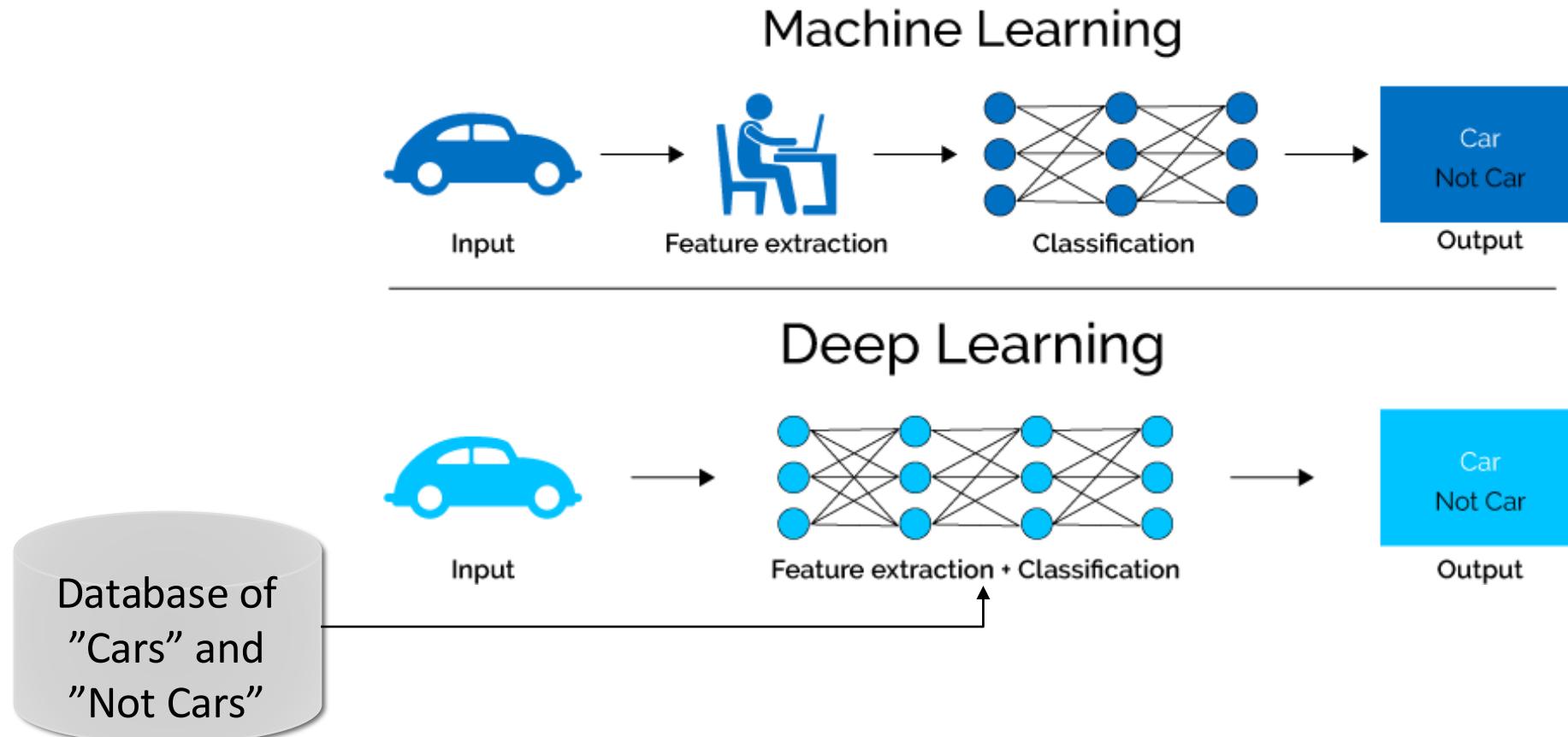
[http://www.image-net.org/papers/imagenet\\_cvpr09.pdf](http://www.image-net.org/papers/imagenet_cvpr09.pdf)

# Why is deep learning so powerful?



# Why is deep learning so powerful?

Solve complex problems without domain knowledge – you just need data!



# SOFTWARE 2.0

---



Andrej Karpathy  
Director of AI at Tesla

**Medium**  
Nov 11, 2017  
**Software 2.0**

*Neural networks are not just another classifier, they represent the beginning of a fundamental shift in how we write software. They are Software 2.0.*

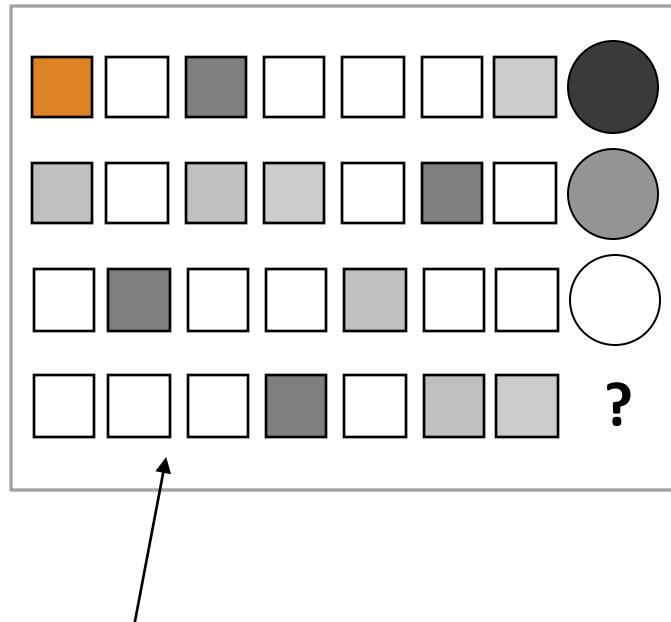
*The “classical stack” of Software 1.0 is what we’re all familiar with [...] It consists of explicit instructions to the computer written by a programmer.*

*In contrast, Software 2.0 is written in neural network weights. No human is involved in writing this code [...] Instead, we specify some constraints on the behavior of a desirable program (e.g., a dataset of input output pairs of examples) and use the computational resources at our disposal to search the program space for a program that satisfies the constraints.*

# Learning principle

---

Dataset

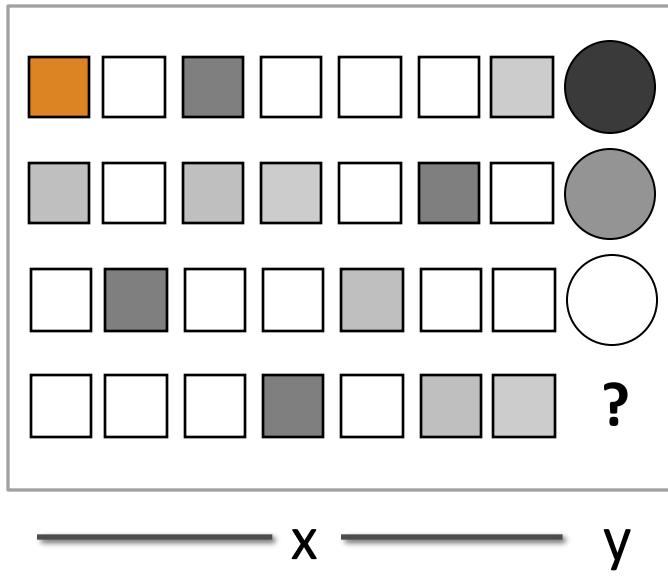


Each row represents a sample in our dataset.

The circles represent the **target output** (say labels like "Car" or "Not Car")

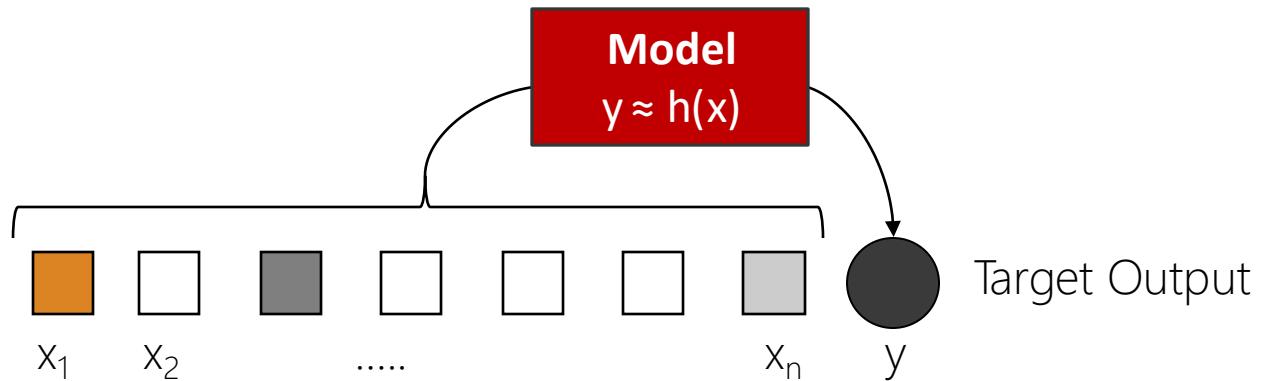
# Learning principle

Dataset



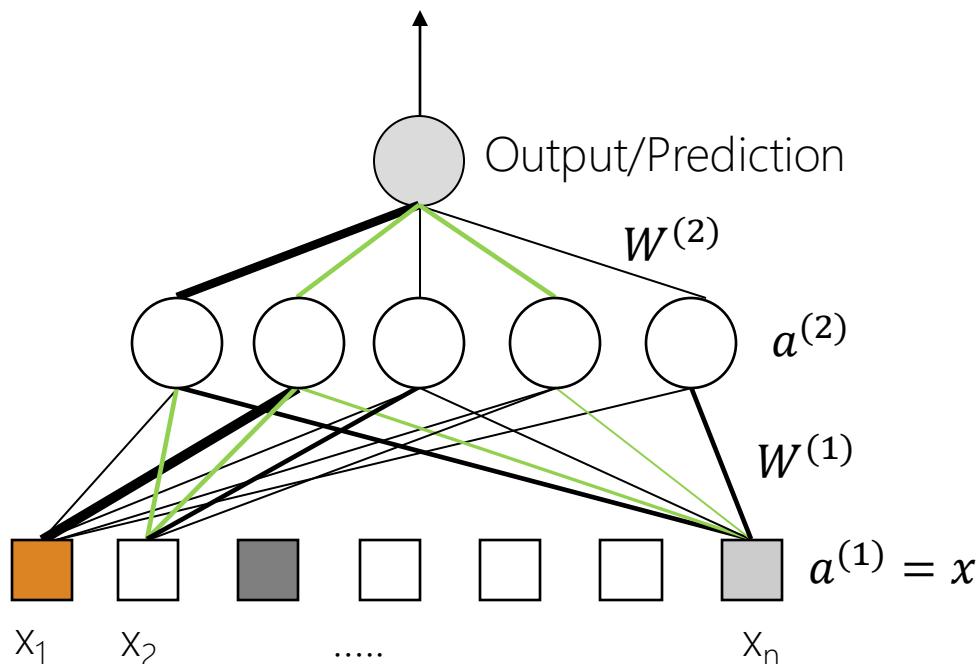
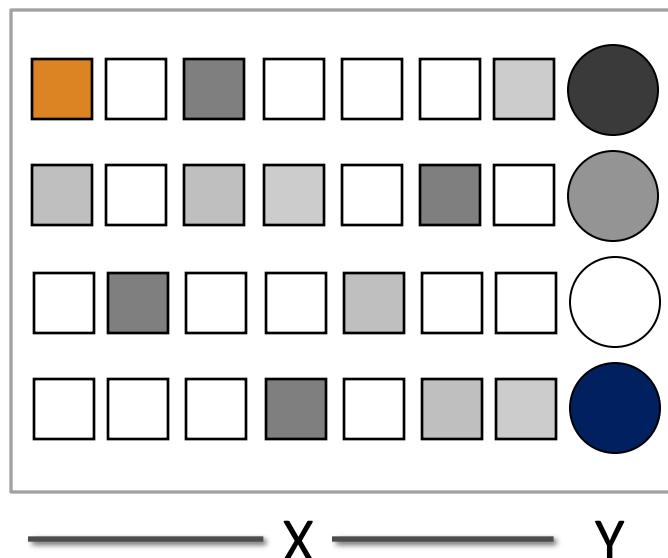
Goal:

- Learn a **model** that is good at predicting  $y$  from  $x$  across the entire training dataset
- A good/useful model generalizes to unseen data



# Learning principle

Dataset

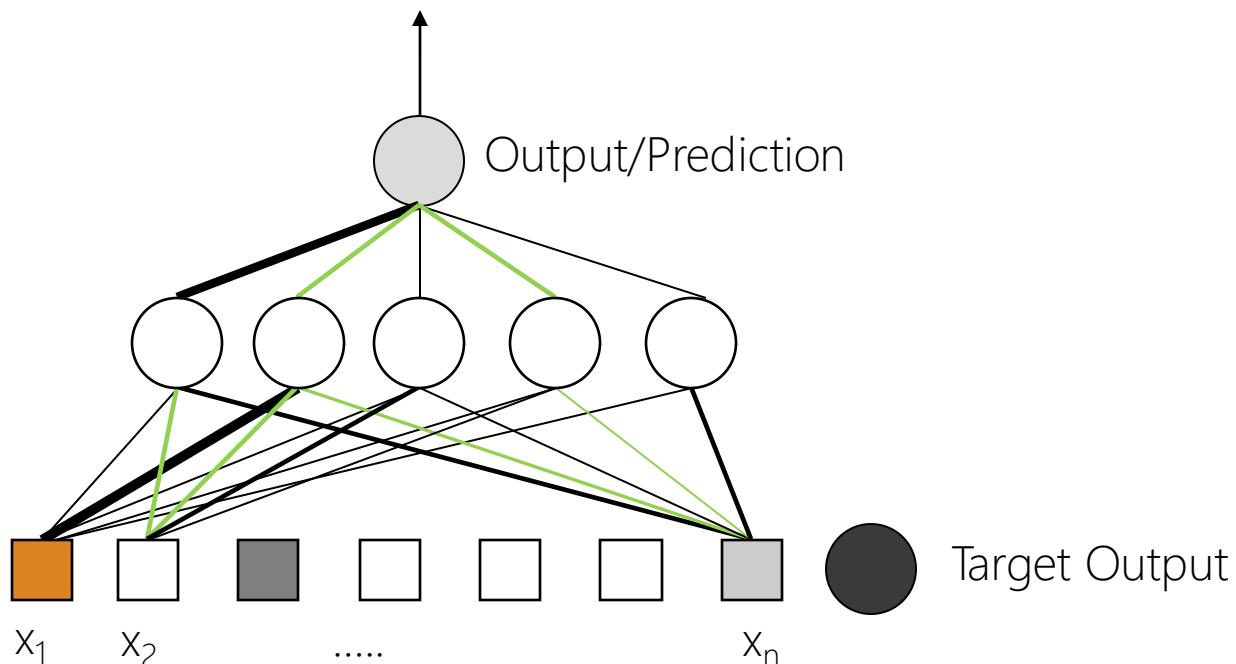
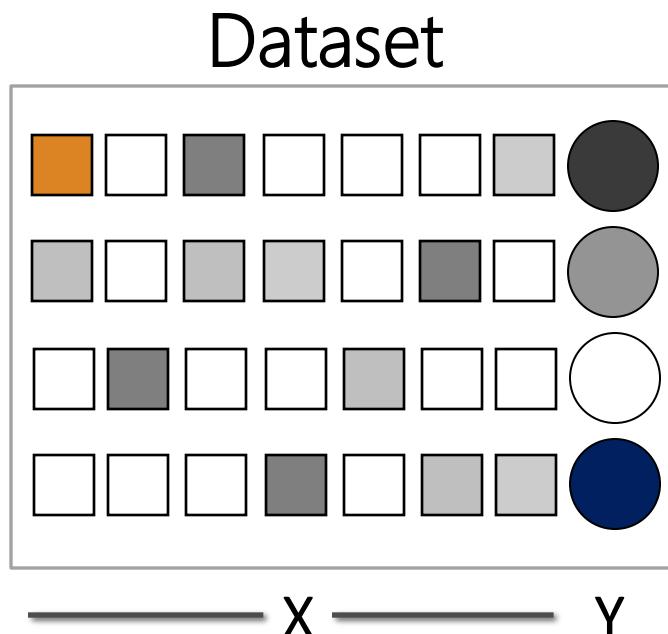


## Neural network:

- White circles are called **neurons**.
- Output/Prediction

$$\begin{aligned}a^{(1)} &= x \\z^{(j+1)} &= W^{(j)}a^{(j)} + b^{(j)} \\a^{(j+1)} &= \sigma(z^{(j+1)})\end{aligned}$$

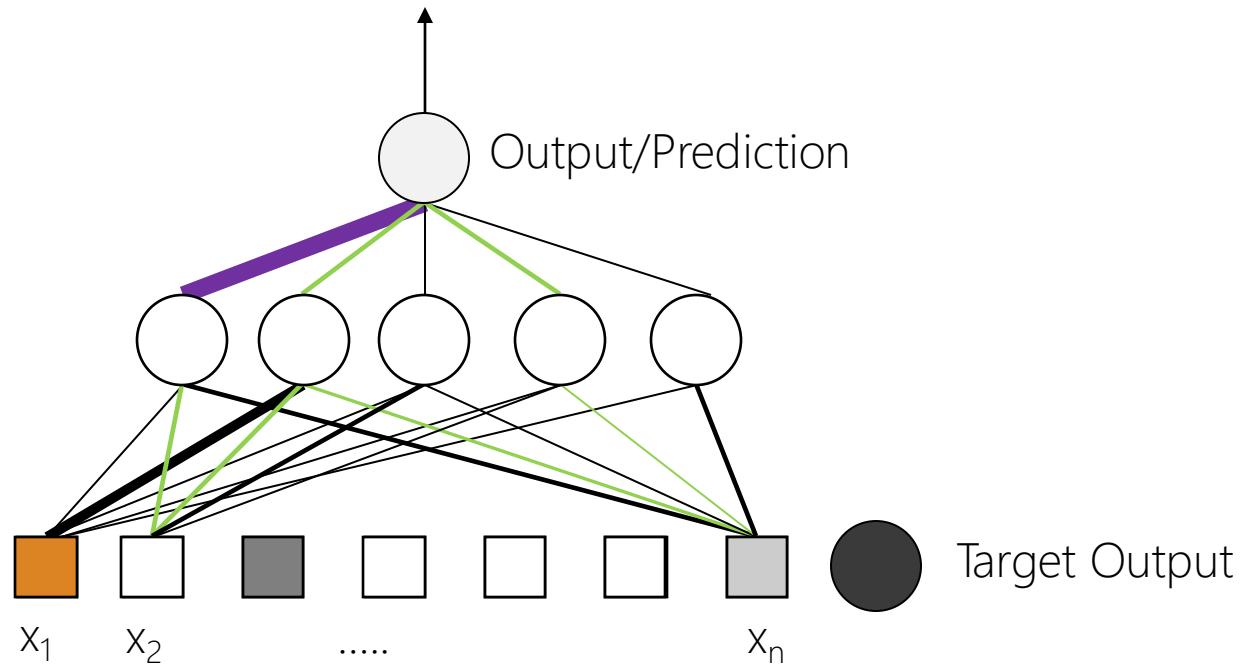
# Learning principle



# Learning principle

---

Adjusting the weights  
either increases or  
decreases the error.

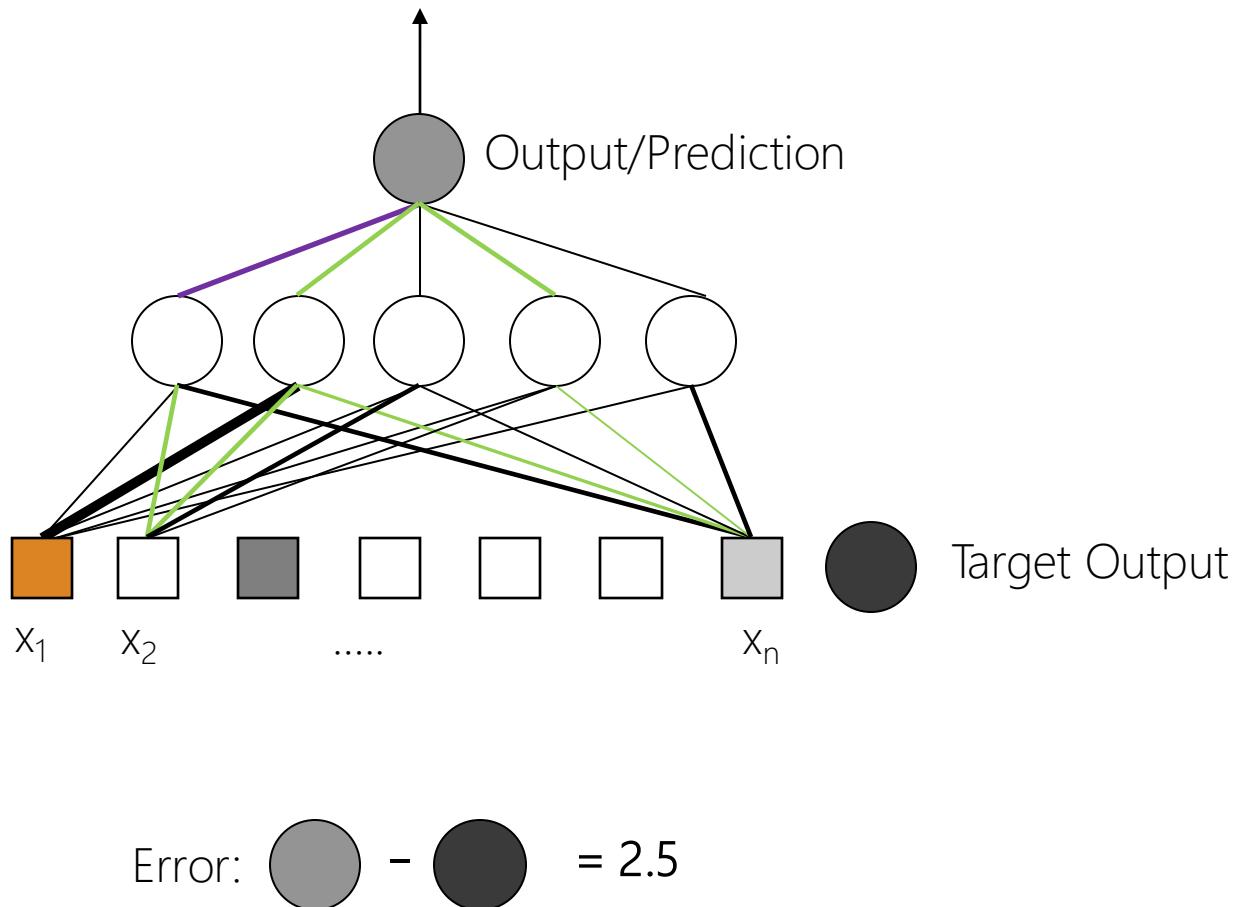
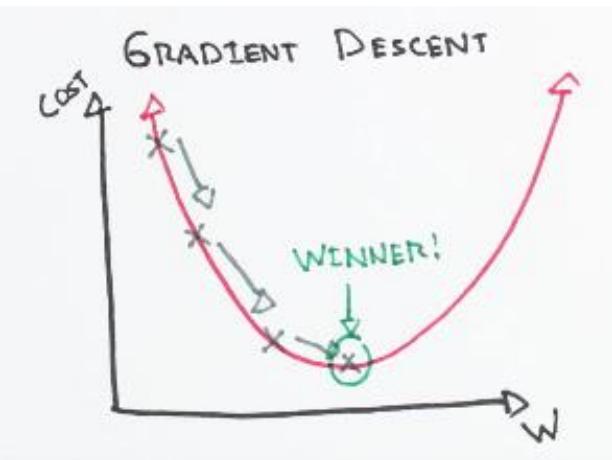


$$\text{Error: } \text{○} - \text{●} = 15$$

# Learning principle

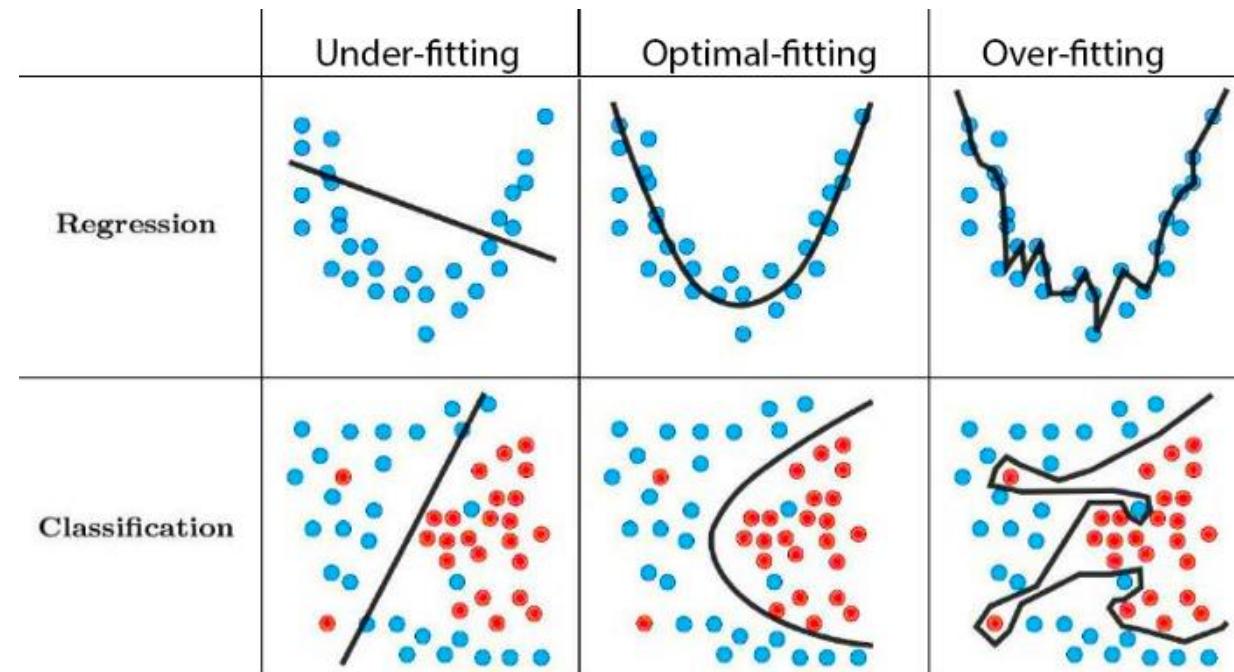
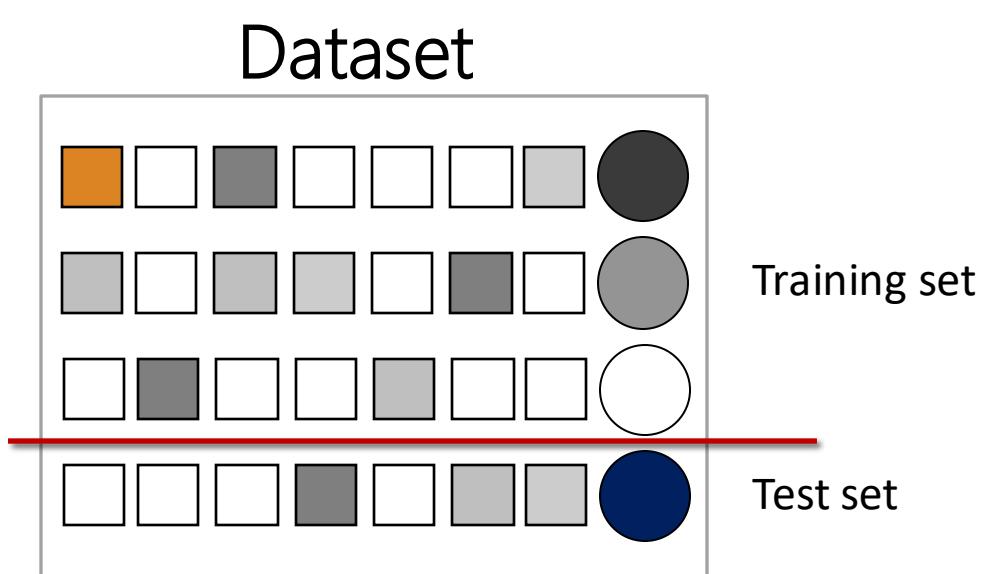
**Training** is the process of adjusting the weights such that the error is small for all samples in your dataset.

**Algorithm:** Gradient descent



# Evaluating model performance

- Neural networks are incredibly good at memorizing the training data, in which case our model most likely won't generalize to unseen data. This is called **overfitting**.
- To evaluate model performance (i.e., how well does it generalize to unseen data) and to check for overfitting, it is standard practise to train and test your model on separate subsets of the data.



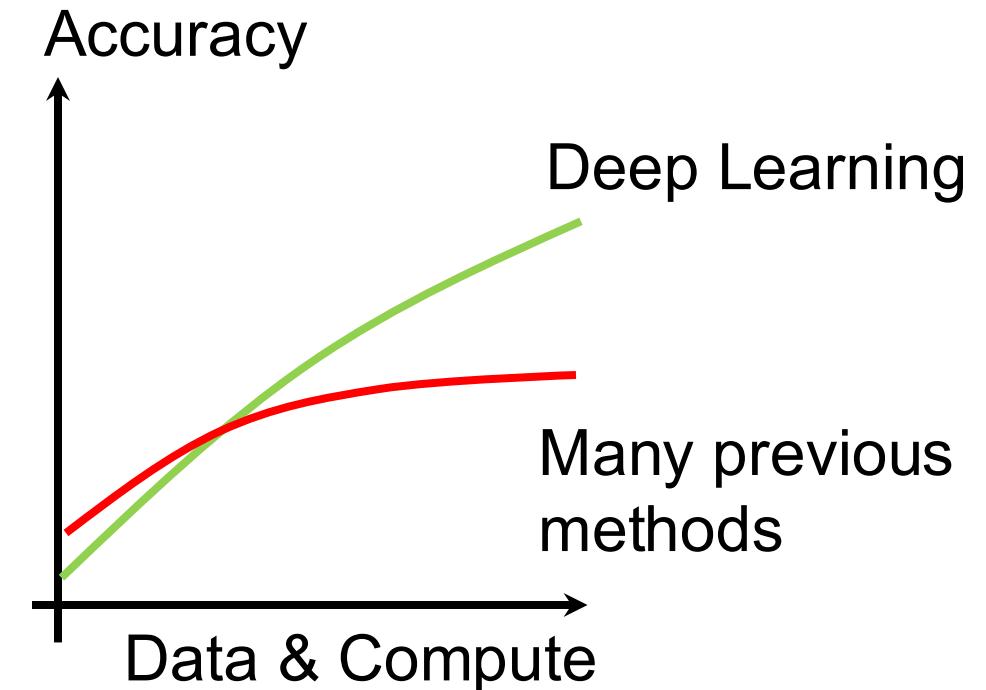
# HOW DOES IT WORK?

---

# Feature extraction using convolution

---

- The traditional approach of using hard-coded features + machine learning doesn't scale well.
- What separates deep learning from other machine learning techniques is that it is able to extract learnable features.
- How?



# Feature extraction using convolution

---

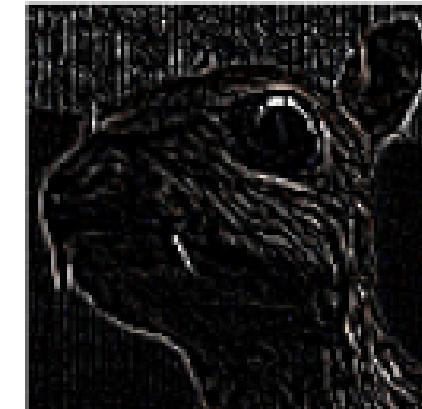
Input image



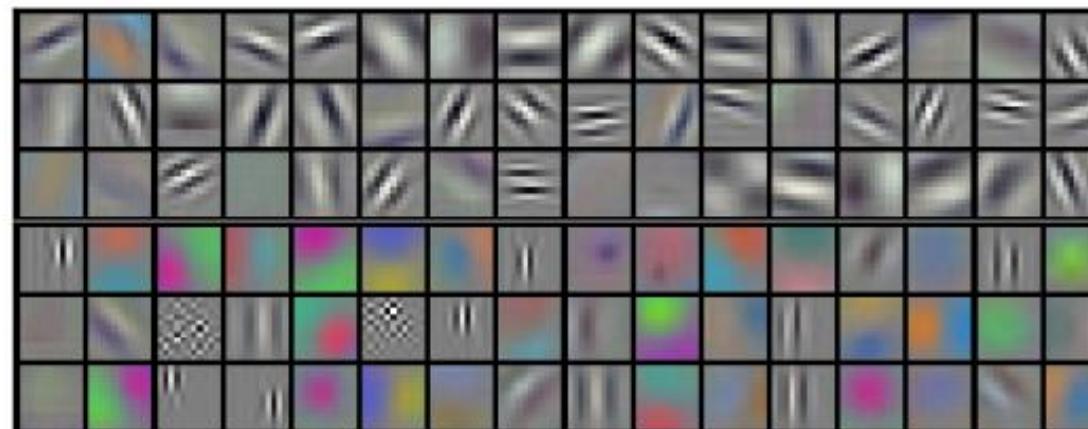
Convolution  
Kernel

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

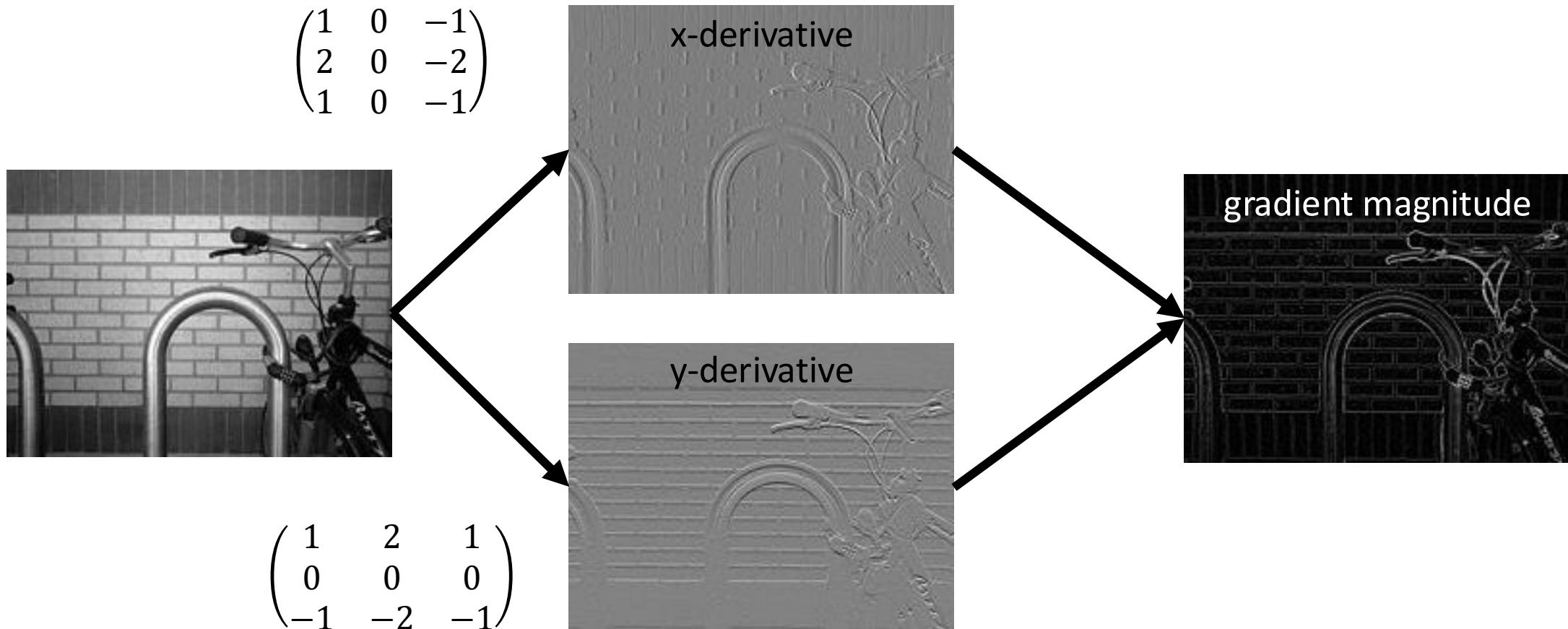
Feature map



Convolution kernels  
learned by a neural  
network.



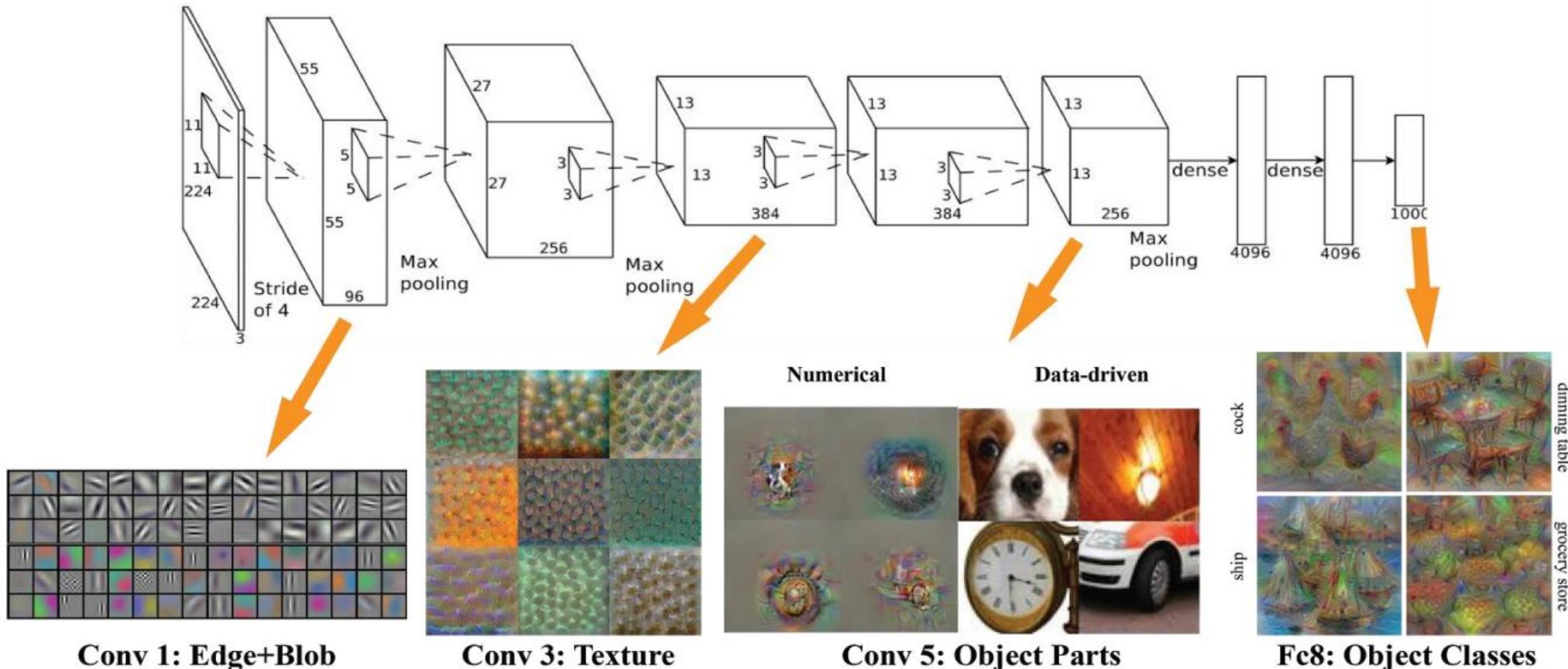
# Feature extraction using convolution



Sobel filters act as edge detectors.  
(formally they estimate the derivative with some built-in noise reduction – we will use them in Lab 1)

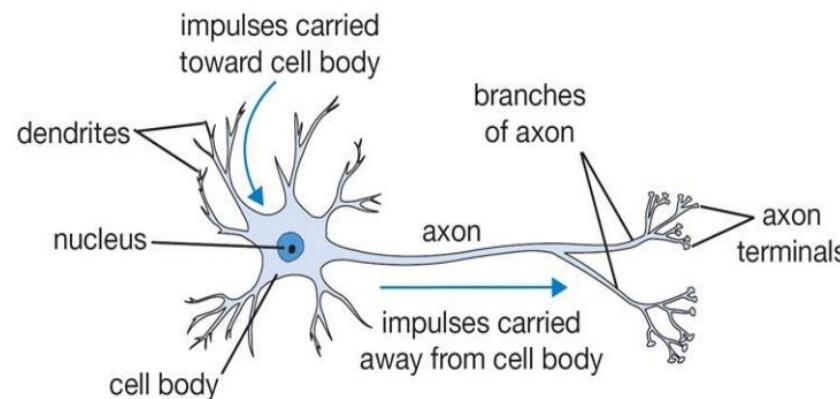
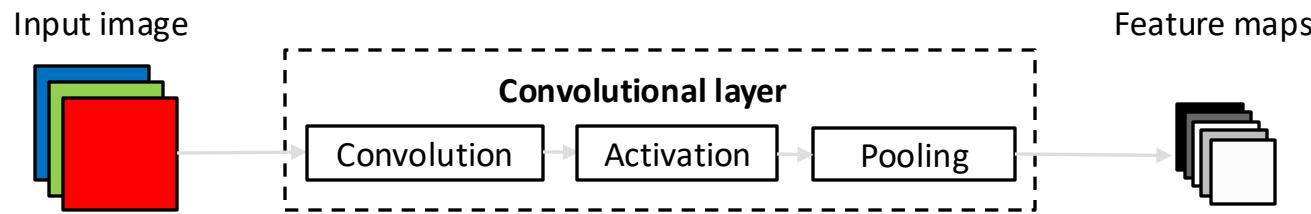
# Features of features

- Deep = features of features of ...



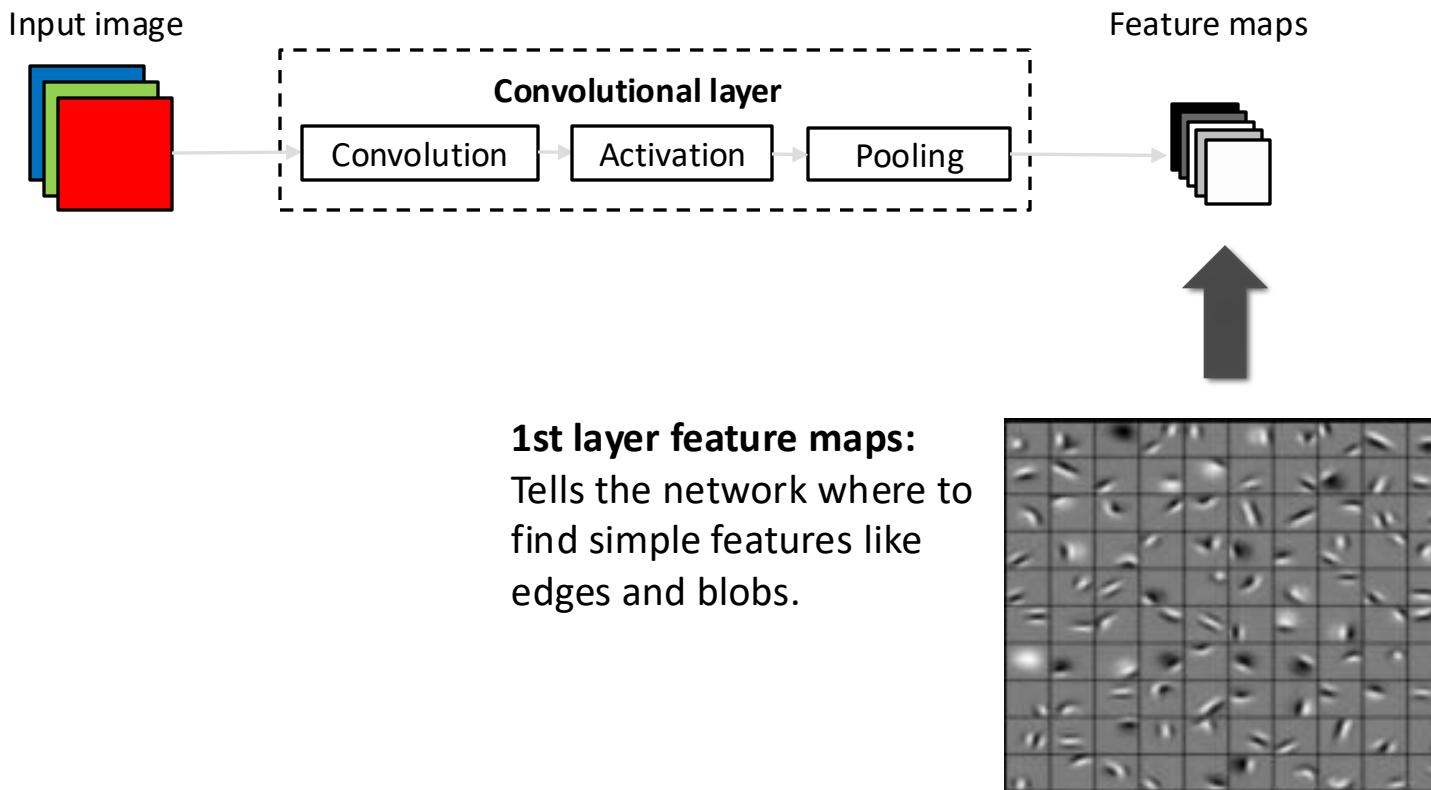
# Convolutional Neural Networks

---



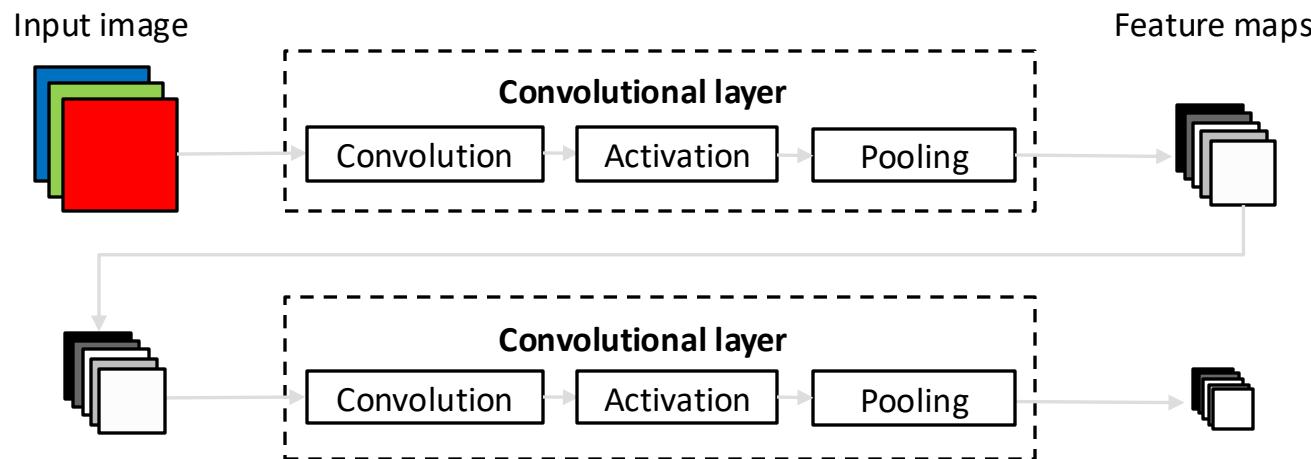
# Convolutional Neural Networks

---

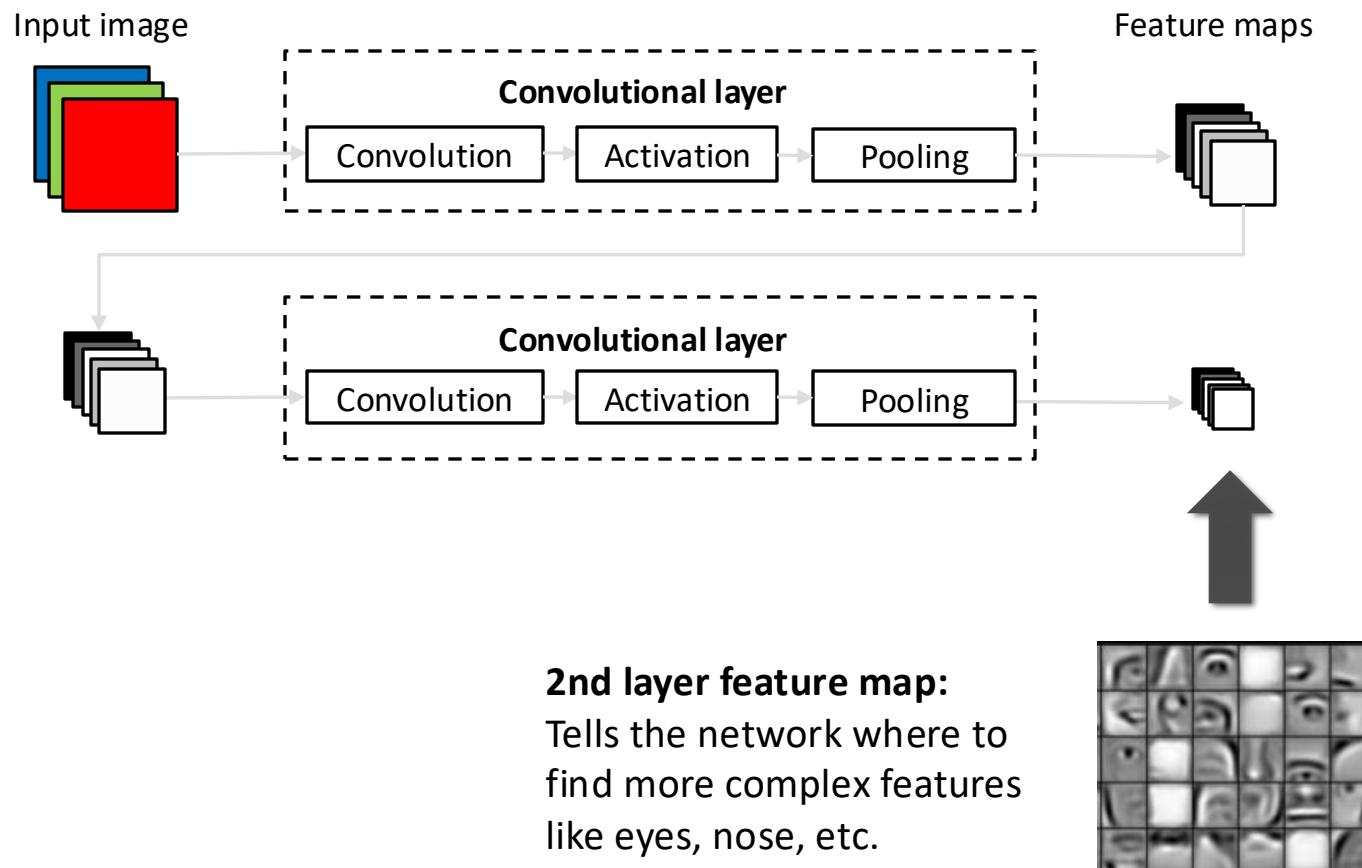


# Convolutional Neural Networks

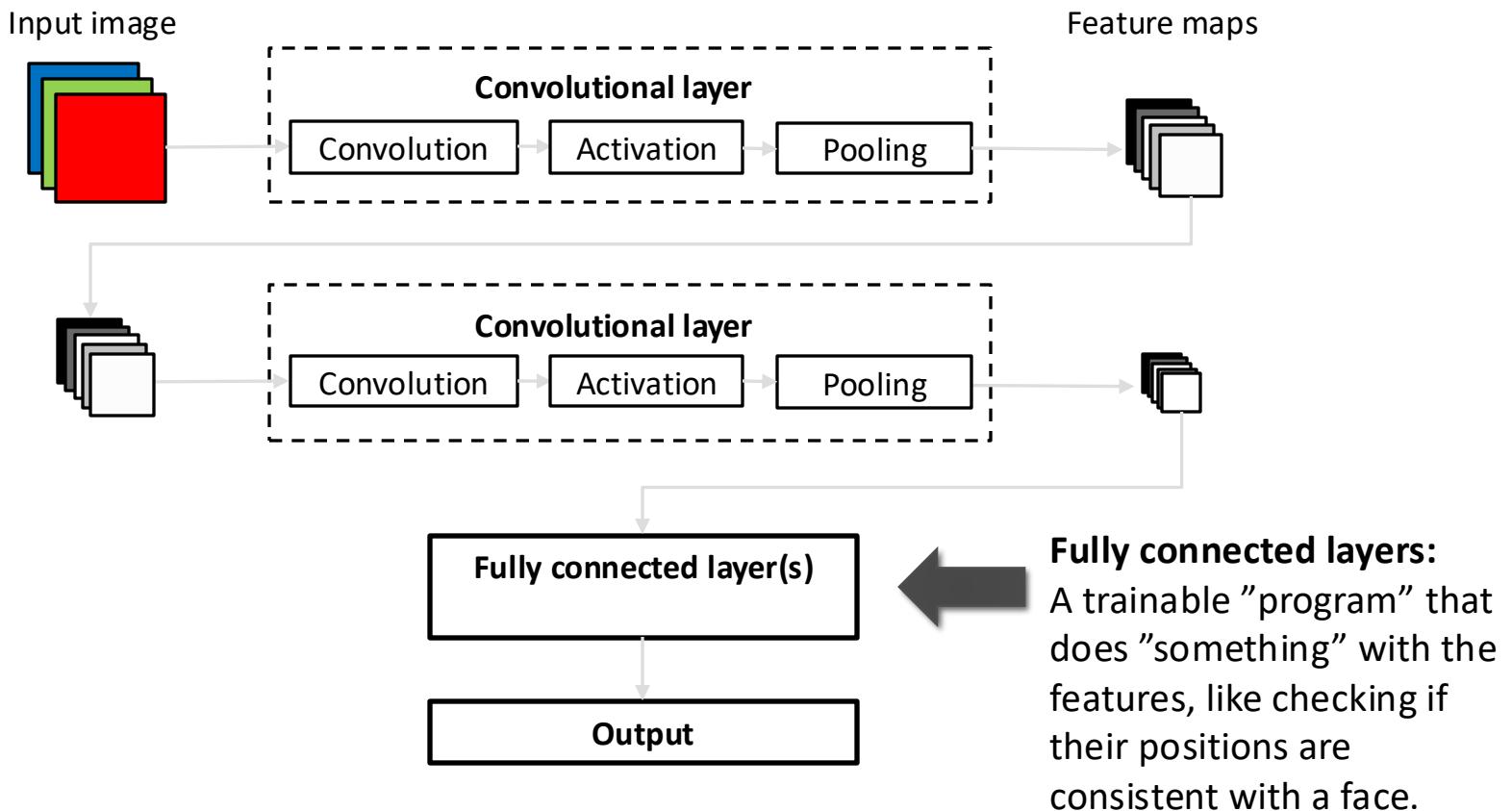
---



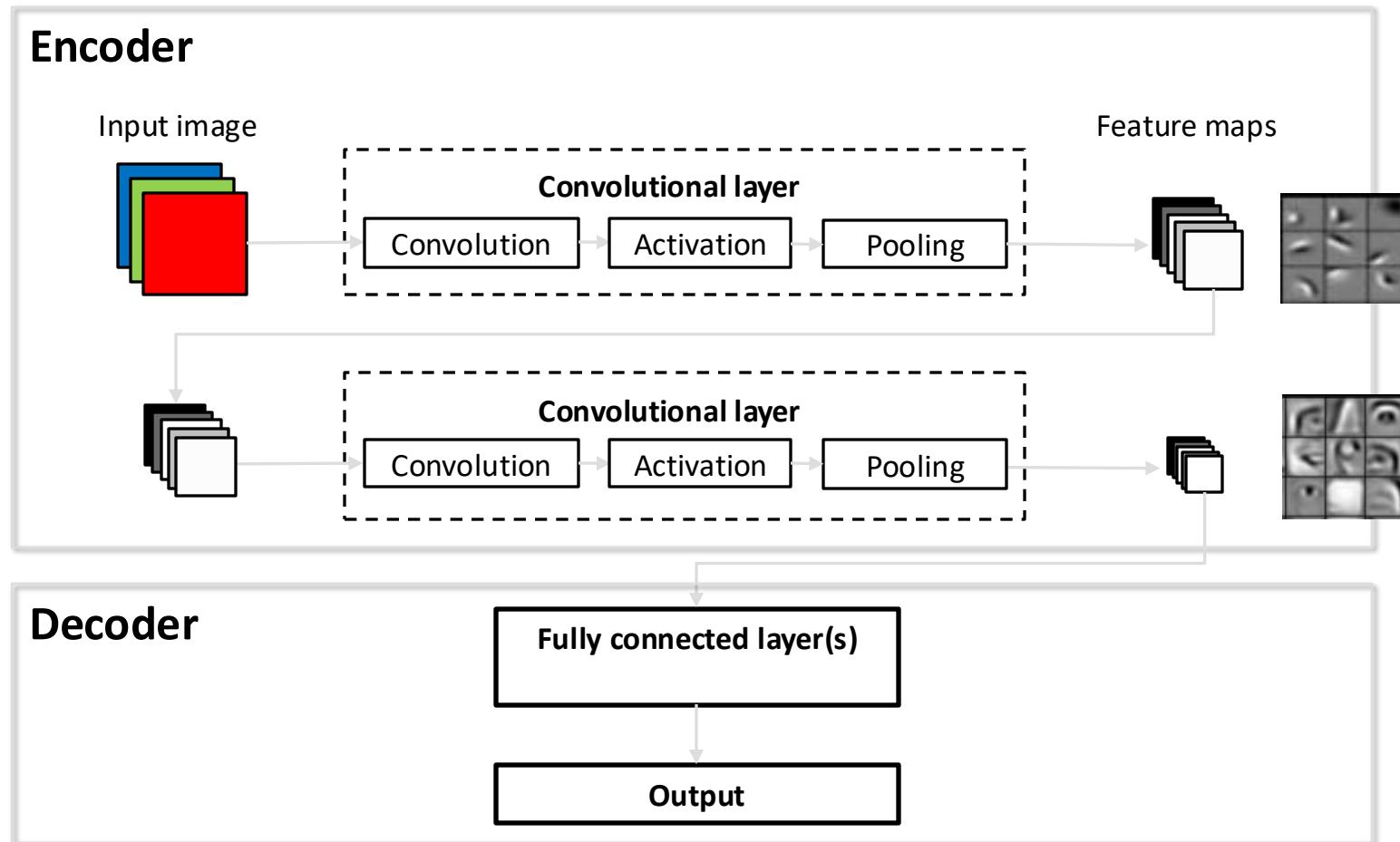
# Convolutional Neural Networks



# Convolutional Neural Networks



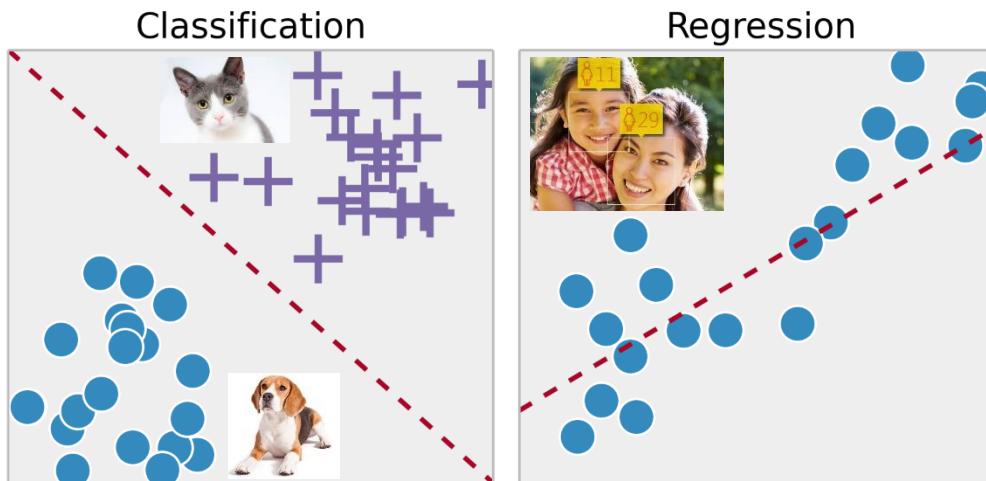
# Encoder/decoder perspective



# Encoder/decoder perspective

**Classifier**  
Uses features to distinguish between two or more classes.

**Output:**  
Discrete labels ("Dog" or "Cat")



**Regressor**  
Uses features to predict some functional relationship.

**Output:**  
Real numbers ("Age" of person in image)

**Decoder**

Fully connected layer(s)

Output

# Computer vision tasks

---

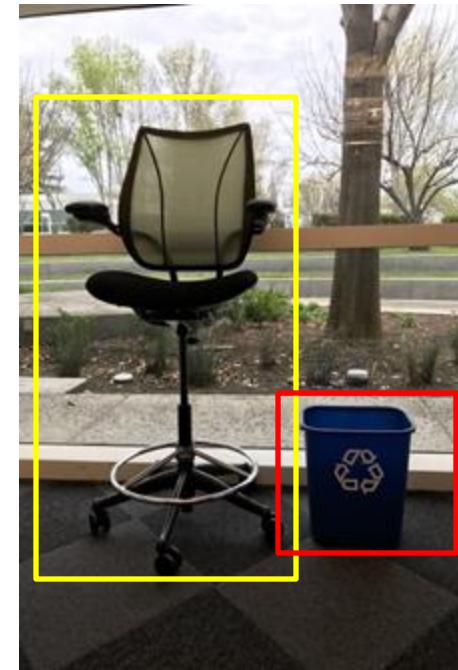
**Image  
Classification**



**Image  
Classification +  
Localization**



**Object Detection**



**Image  
Segmentation**



# Advanced computer vision tasks

Colorize gray-scale images

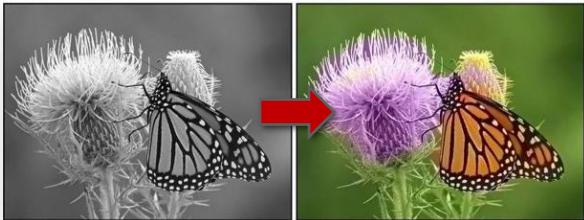
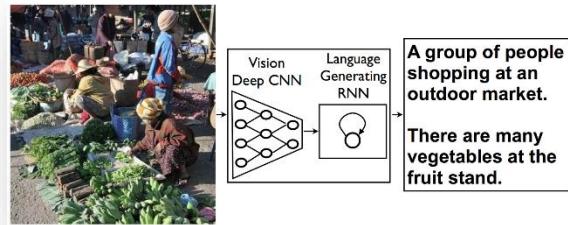
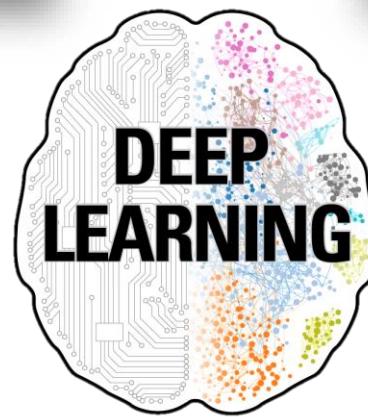


Image captioning



Turn horses into zebras



Turn images into Van Gogh paintings



"Dream" images of fake celebrities

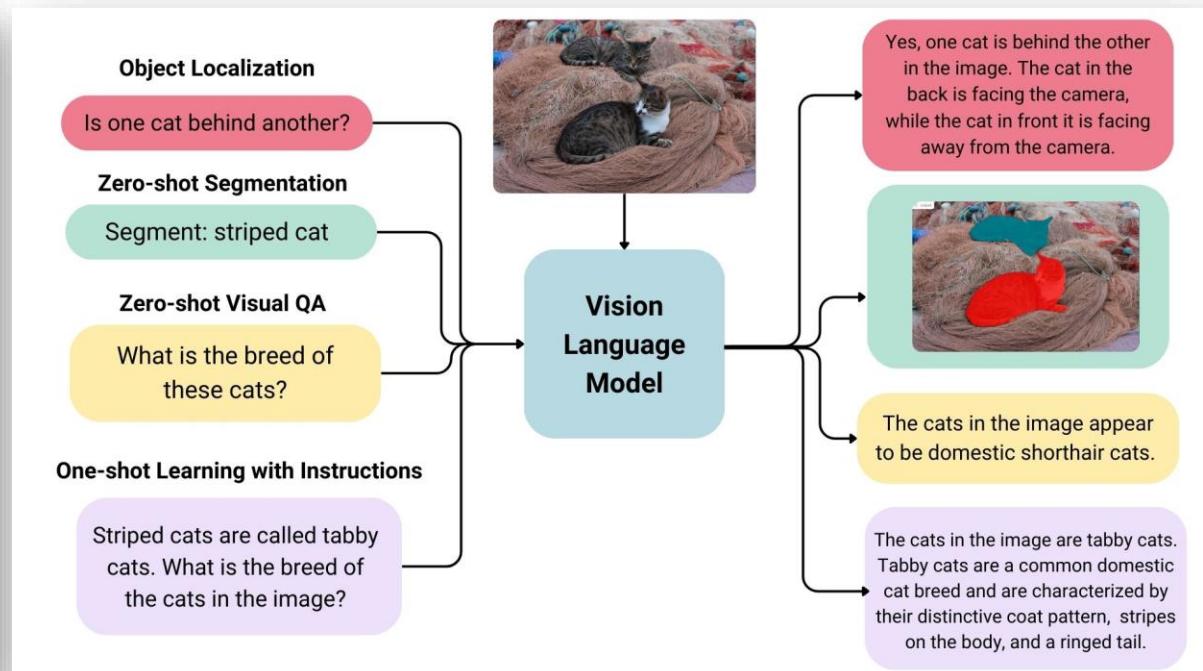
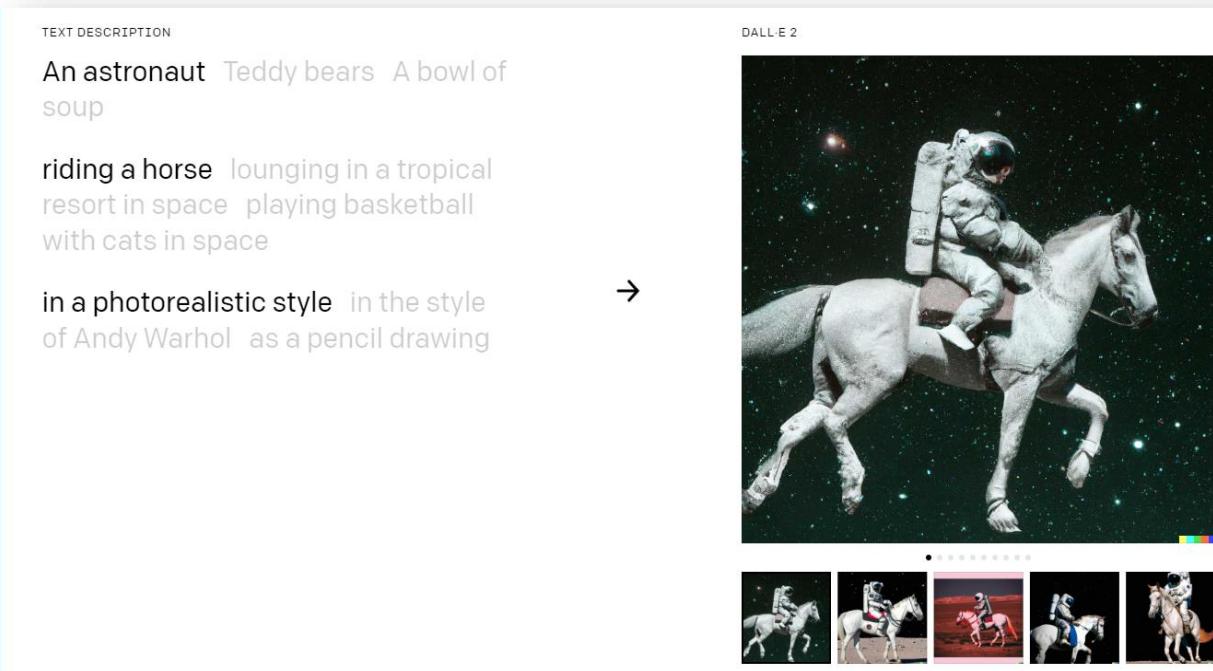


Detect human body pose



# Trend: Visual Language Models

- VLMs are models that can learn simultaneously from images and texts to tackle many tasks, from image generation ("Generative AI") to visual question answering and image captioning.



# Ways to inspire ideas for your project

---

- Pick a framework and go through some tutorials: PyTorch, Keras, ...
  - PyTorch tutorials: <https://pytorch.org/tutorials/index.html>
  - Keras examples: <https://keras.io/examples/>
  - Keras developer guide: <https://keras.io/guides/>
- Brainstorm useful X → Y mappings
- Experiment! Try something out, e.g., reproduce results from public GitHub repos.
- Papers with Code: <https://paperswithcode.com/area/computer-vision>

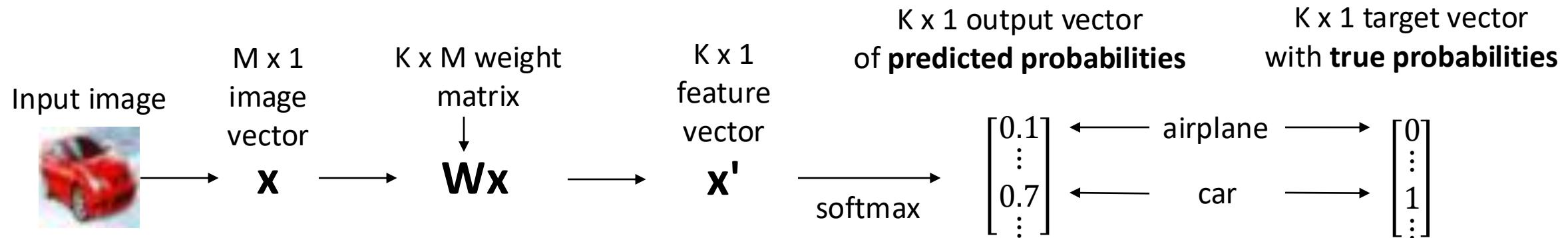
# Lecture plan

---

# Week 2: Machine Learning fundamentals

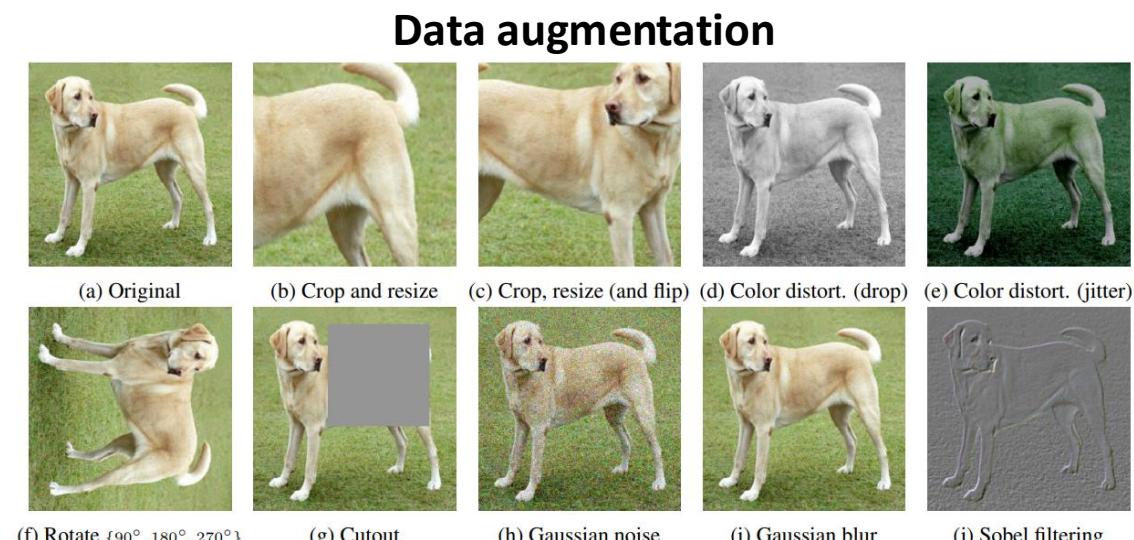
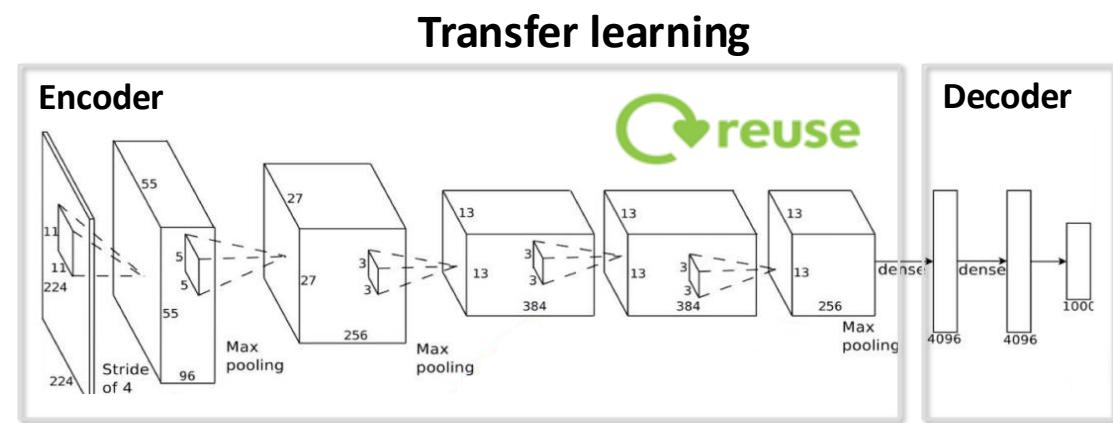


Rows of  $\mathbf{W}$

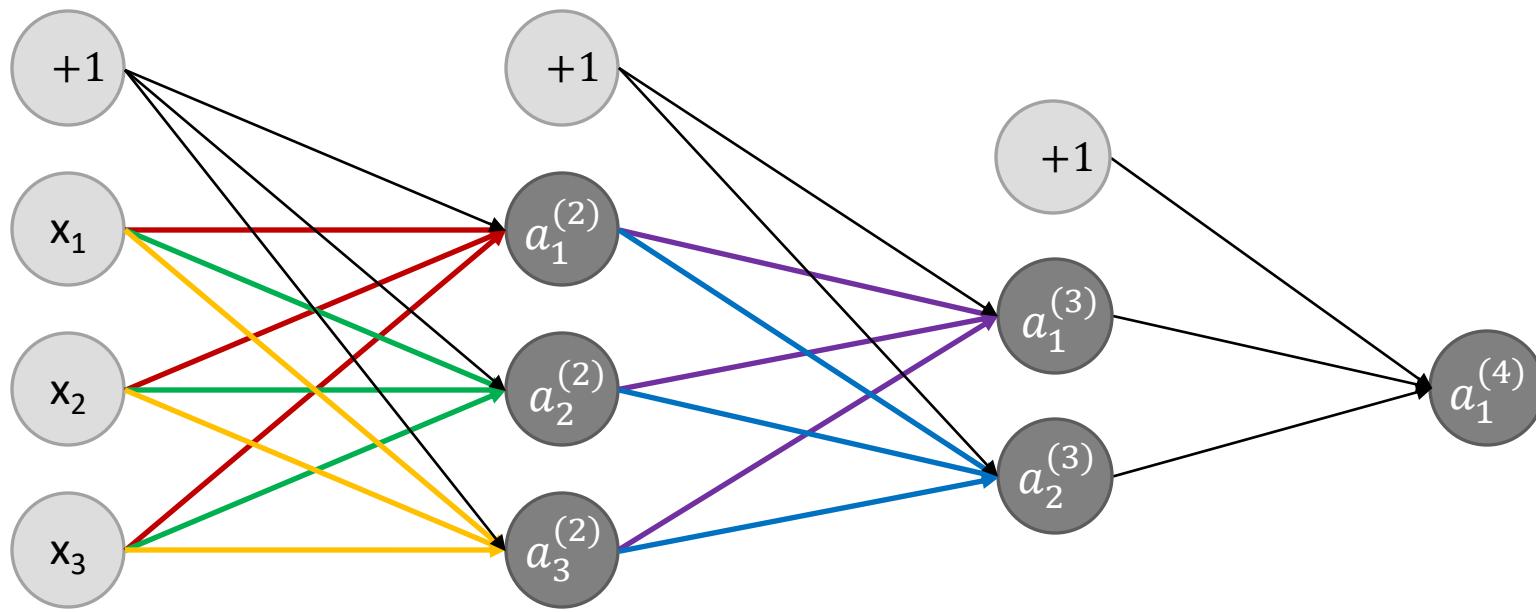


# Week 2: Practical info about the course project

- Task is to recognize three types of Russian tanks
  - BMD4, BMP2, and T72
- Dataset
  - Self-made from Google Image search
  - 500 images from each class
- Neural network architecture: ResNet50
- Training:
  - Transfer learning using ResNet50 backbone (encoder) pre-trained on ImageNet
  - Data augmentation
- Evaluation / analysis
  - Inspect loss curves
  - Confusion matrix
  - Qualitative assessment of mis-classified images
  - Heatmaps (GradCAM – covered in Lecture 10)

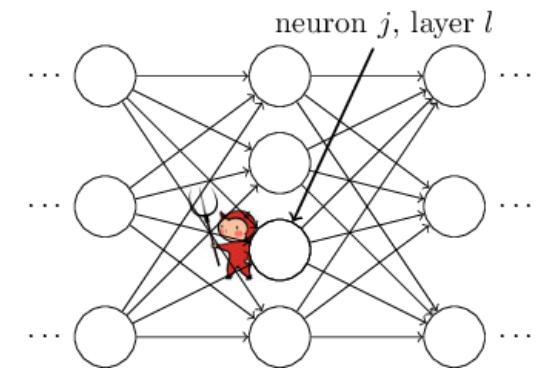


# Week 3: Traditional Neural Networks

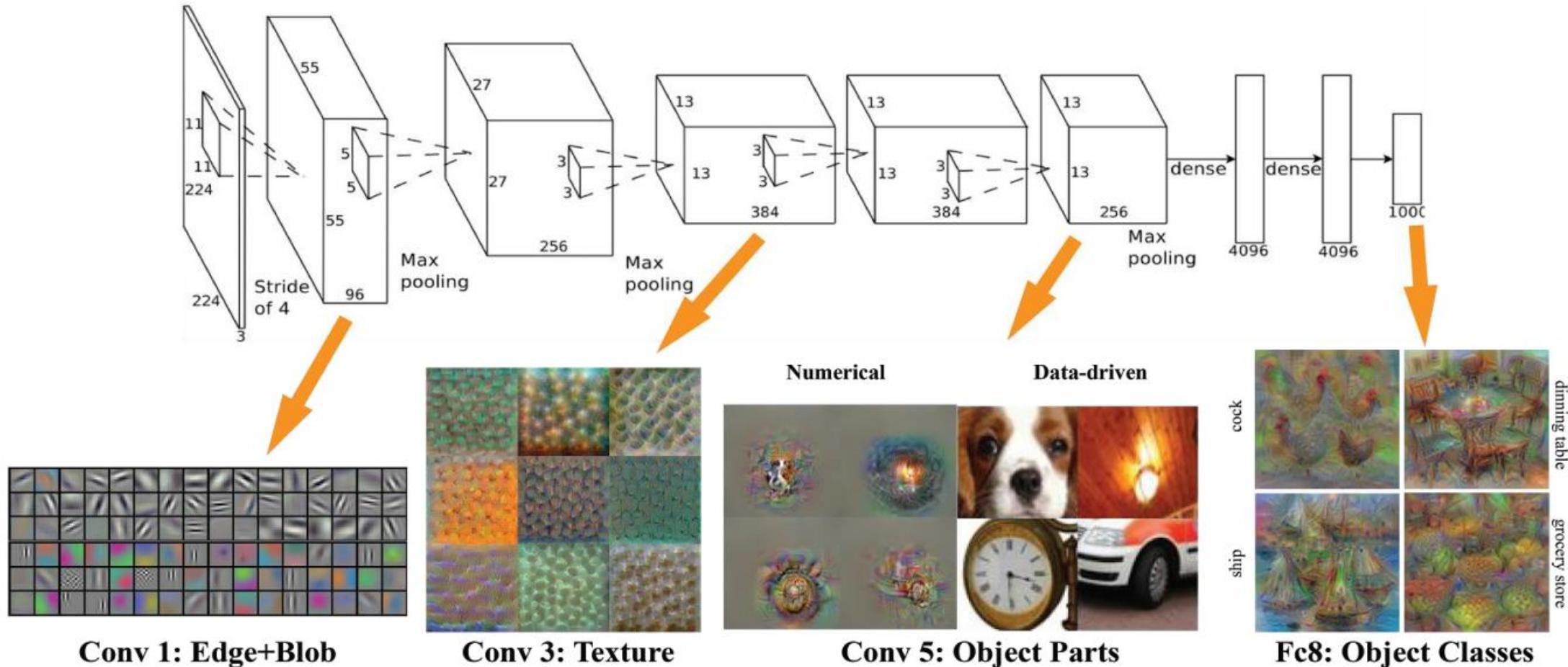


$$\begin{aligned}a^{(1)} &= x \\z^{(j+1)} &= W^{(j)}a^{(j)} + b^{(j)} \\a^{(j+1)} &= \sigma(z^{(j+1)})\end{aligned}$$

**Backpropagation  
algorithm**



# Week 4: Convolutional Neural Networks



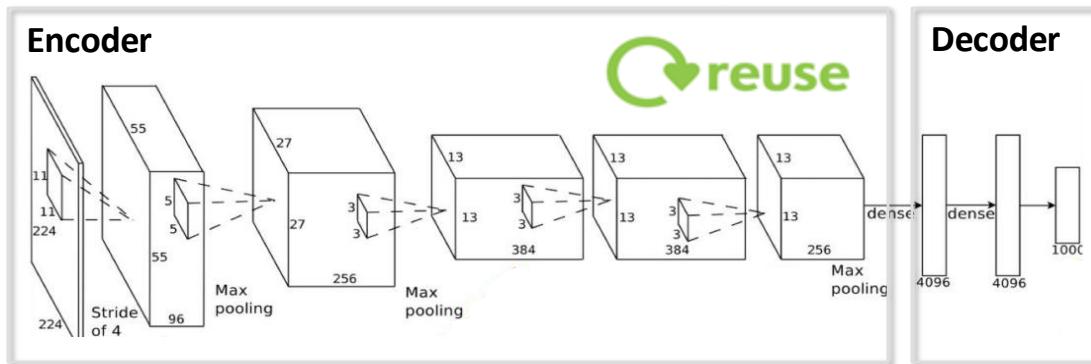
# Week 4: How to write a good report

---

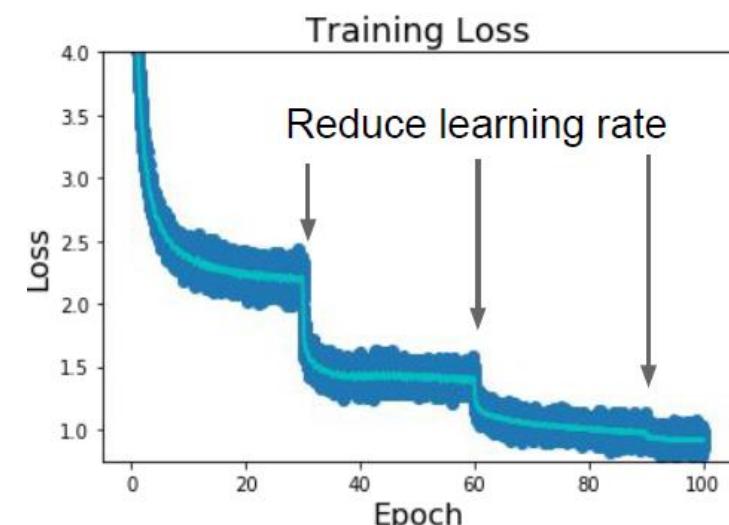
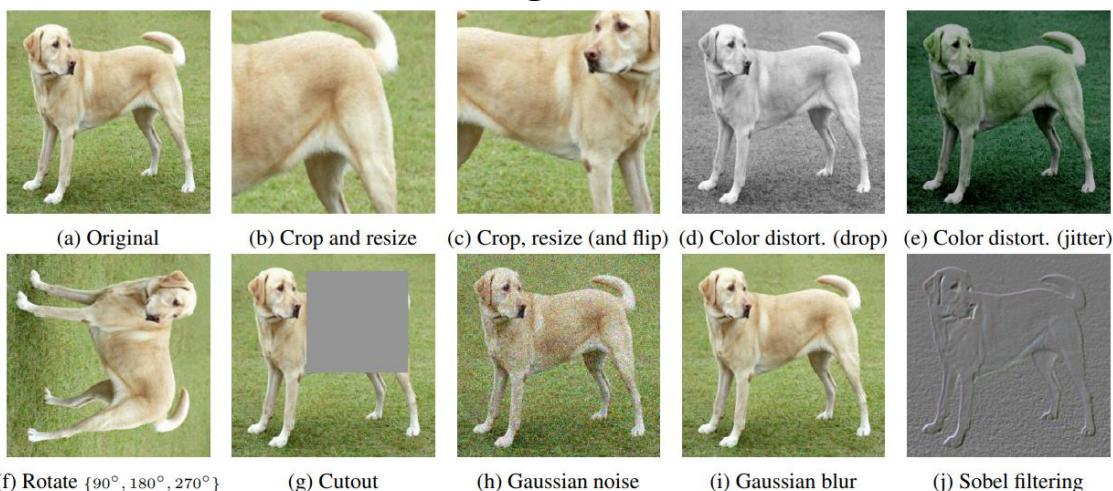
- Use [this arXiv template](#)
- Report should be structured like a research paper:
  - Abstract
  - Introduction (Includes motivation and objectives of your project. Be realistic and be specific!)
  - Related work (Summarize at least 2-3 key references. What makes your approach different?)
  - Methods (What did you do and how? Describe network architecture, data set, experiments, etc.)
  - Results (Objectively summarize your results and compare to related work or state-of-the-art)
  - Discussion (How well did you do? What worked? What didn't work?)
- Final report between 6-8 pages.

# Week 5+6: Training ConvNets

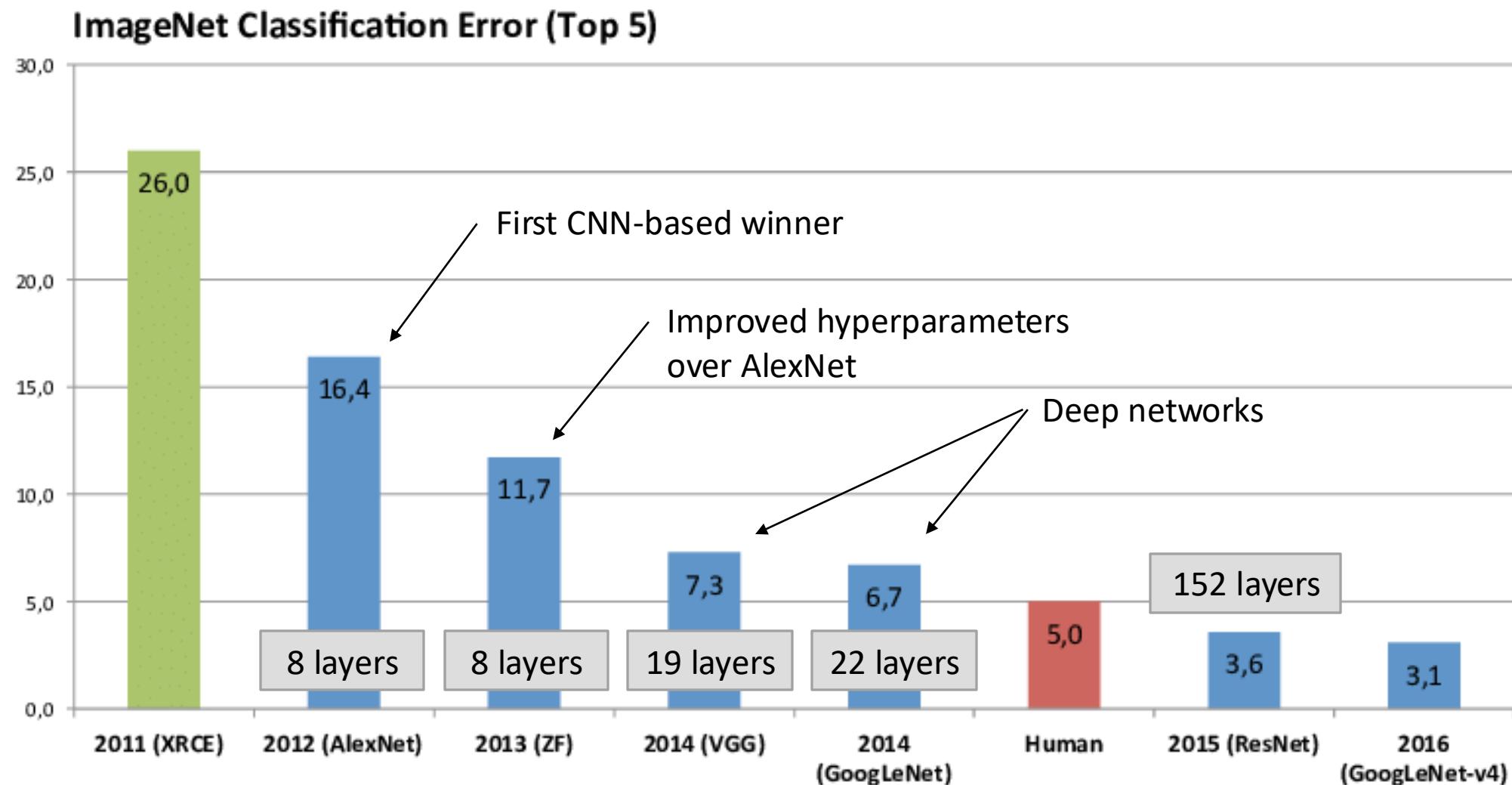
## Transfer learning



## Data augmentation



# Week 7: ConvNet architectures



# Week 8: Object detection and segmentation

Classification



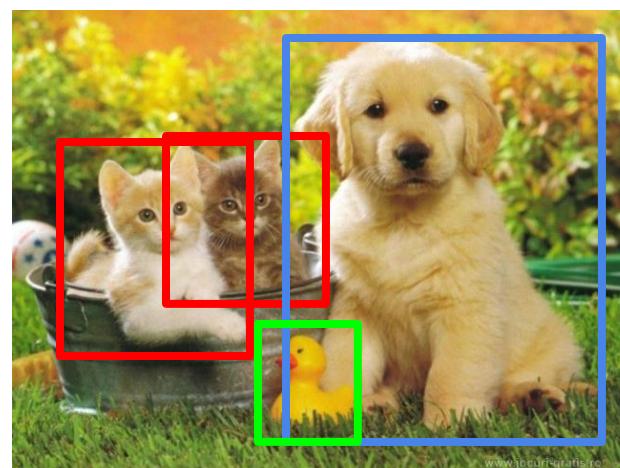
CAT

Classification  
+ Localization



CAT

Object Detection



CAT, DOG, DUCK

Image  
Segmentation



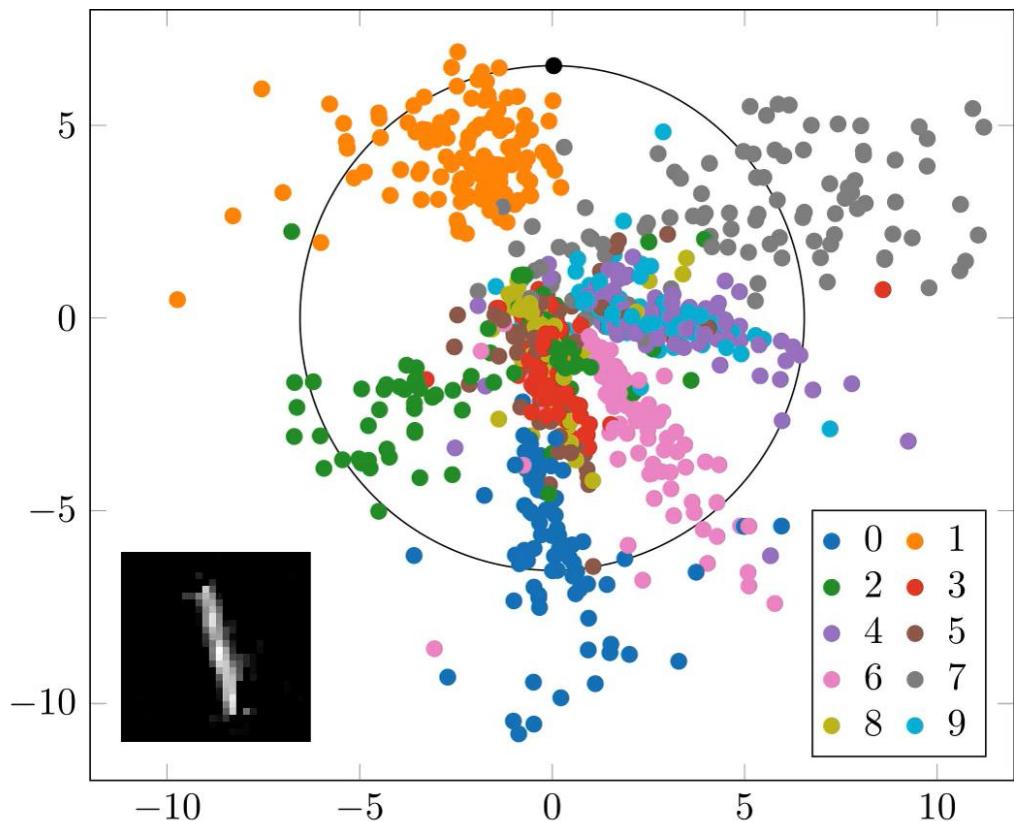
CAT, DOG, DUCK

Single object

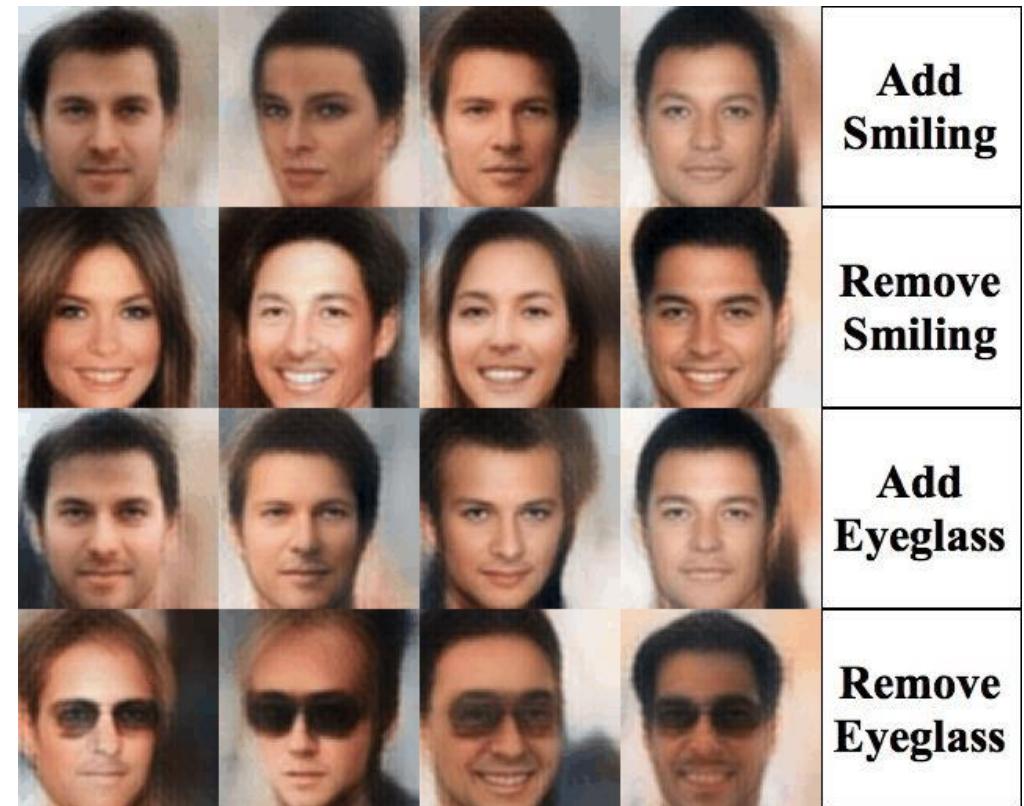
Multiple objects

# Week 9: Generative models

<https://gertjanvandenburg.com/blog/autoencoder/>

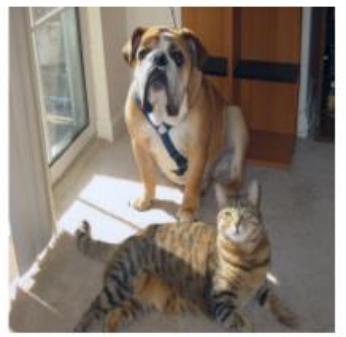


<https://github.com/houxianxu/DFC-VAE>



Variational Autoencoder on CelebA dataset

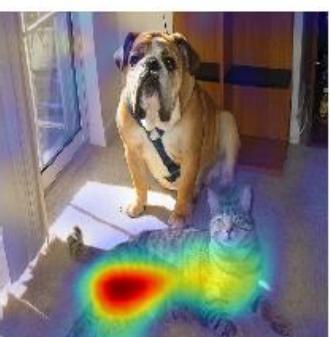
# Week 10: Visualizing and understanding ConvNets



(a) Original Image



(b) Guided Backprop 'Cat'



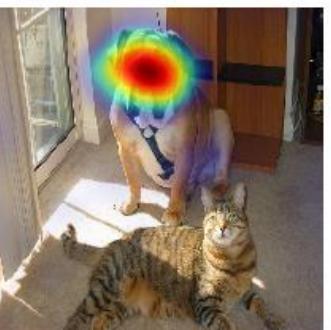
(c) Grad-CAM 'Cat'



(g) Original Image



(h) Guided Backprop 'Dog'



(i) Grad-CAM 'Dog'



A

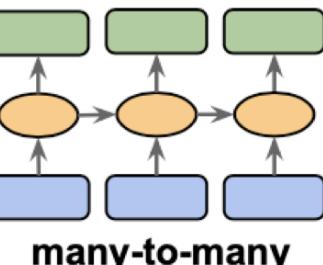
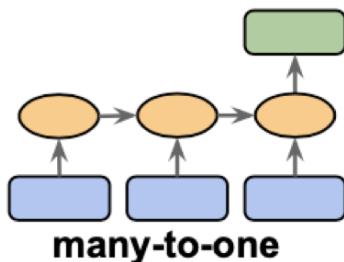


B

# Week 11: Sequence models

## Example

**Input:** Sentence  
**Output:** Sentiment



## Example

**Input:** Video  
**Output:** Human pose

## Example

**Input:** Image  
**Output:** Caption

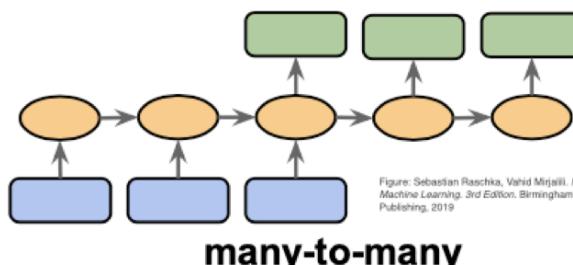
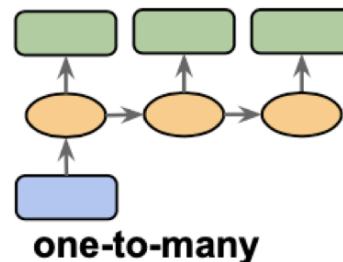
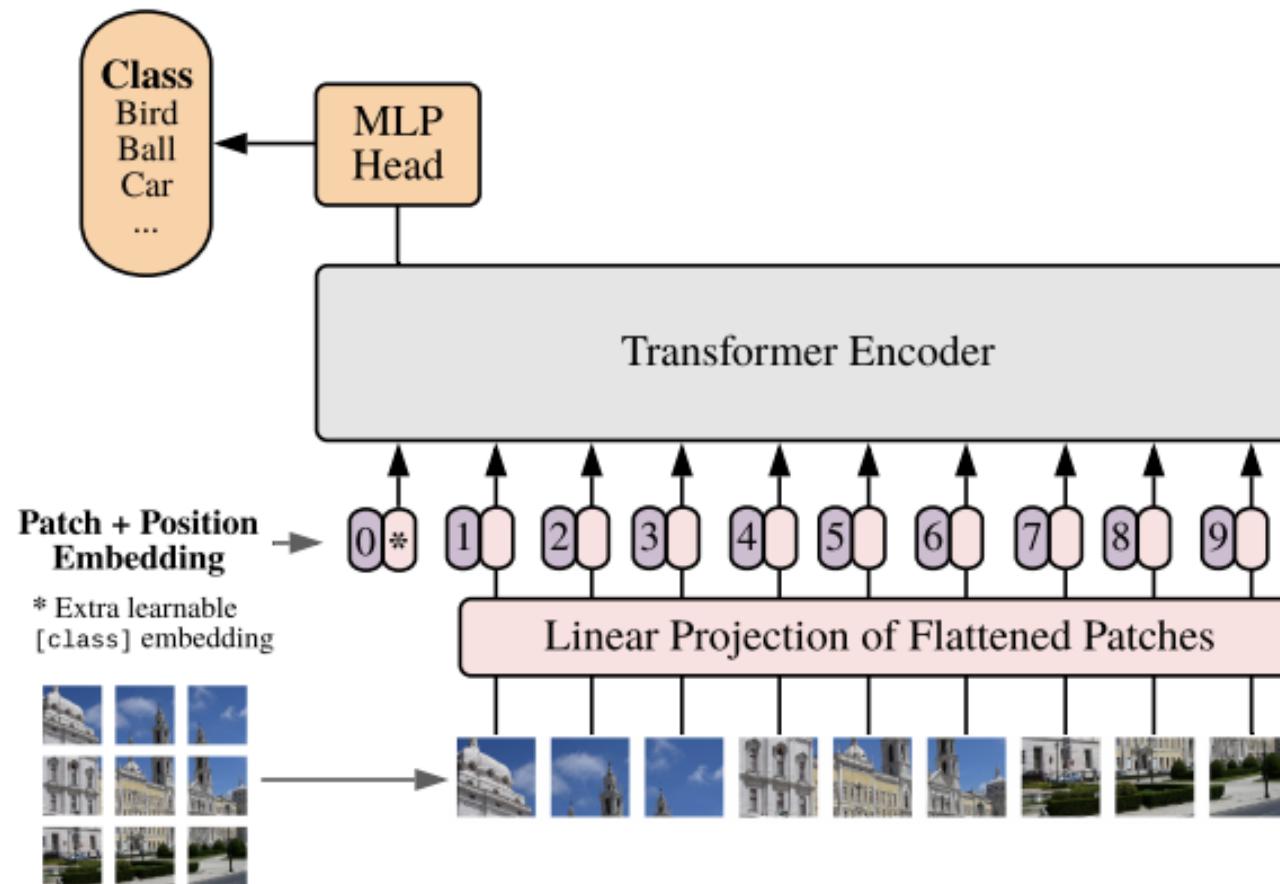


Figure: Sebastian Raschka, Vahid Mirjalili. Python Machine Learning, 3rd Edition. Birmingham, UK: Packt Publishing, 2019

## Example

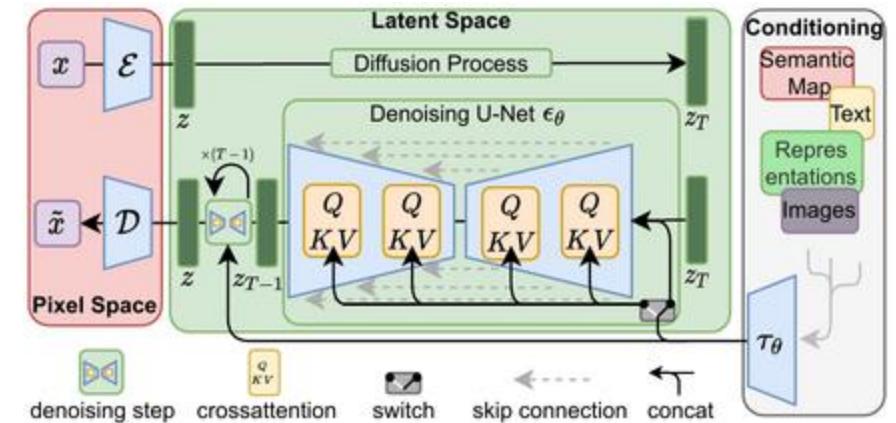
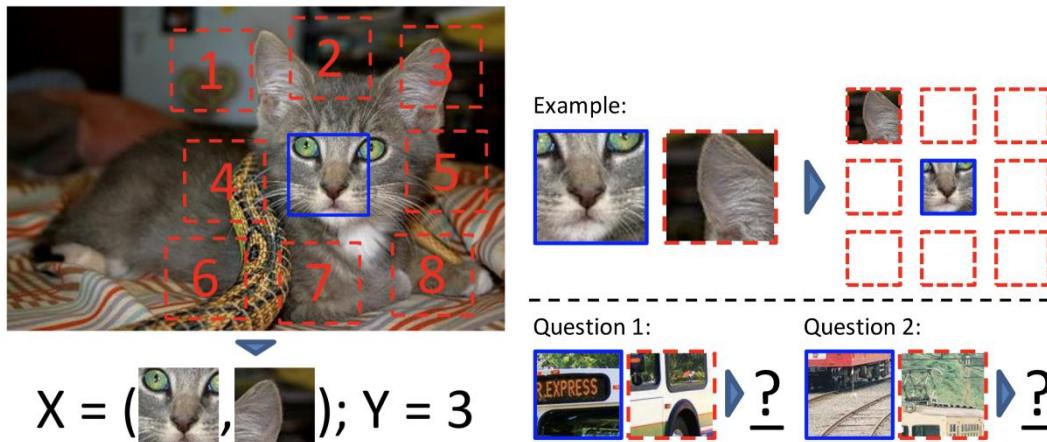
**Input:** English sentence  
**Output:** German translation

# Week 12: Vision Transformers



# Week 13: Advanced topics

- Self-supervised learning
- Diffusion models



# Week 14

---

- Practical info about the exam
- Guest lecture