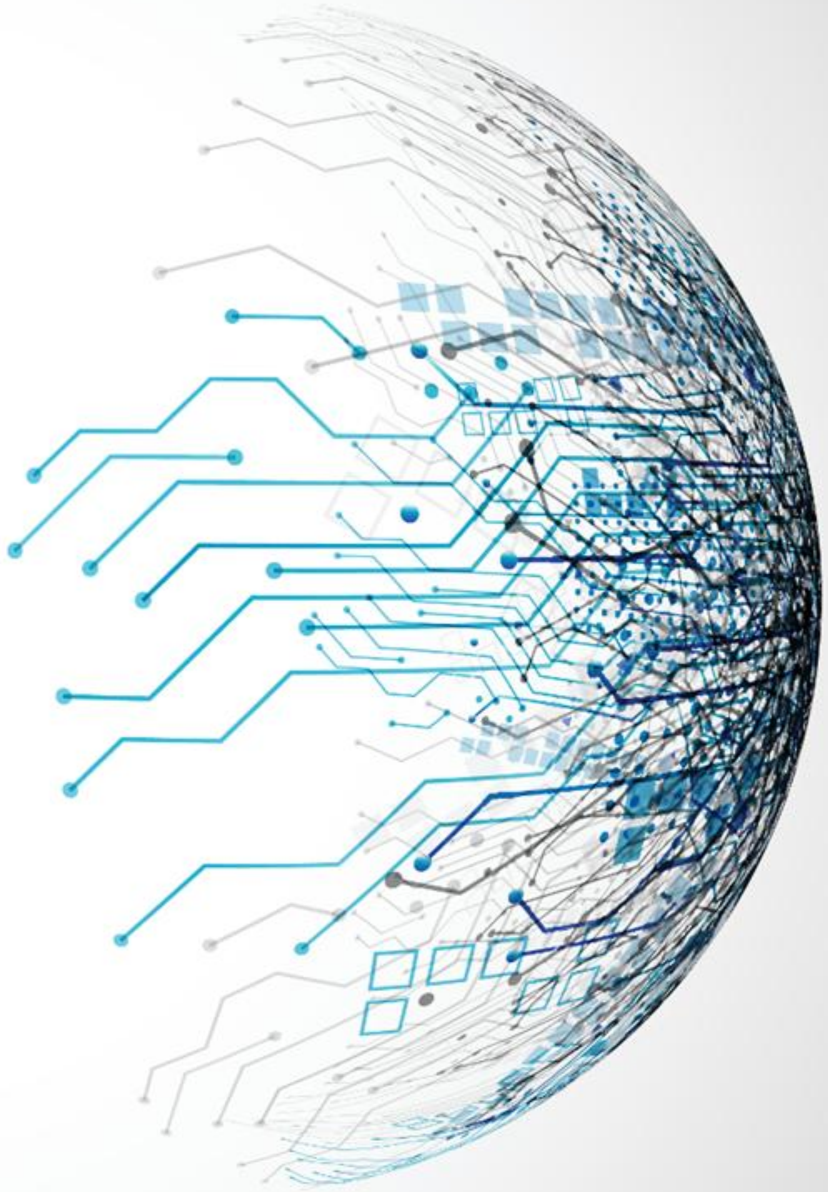
Deep Learning

Computer Vision

Dr. Mohammed Salah Al-Radhi

(slides by: Dr. Bálint Gyires-Tóth)



Copyright

Outline

1. Computer Vision
2. Data Annotation & Augmentation
3. Semantic Segmentation

Computer Vision

Motivation

- The human vision system is not designed to measure absolute values of light.
- It is designed to try to understand "what's there" in the world.

the human visual system usually guesses correctly. does it?

Computer Vision

Make computers understand images and videos.

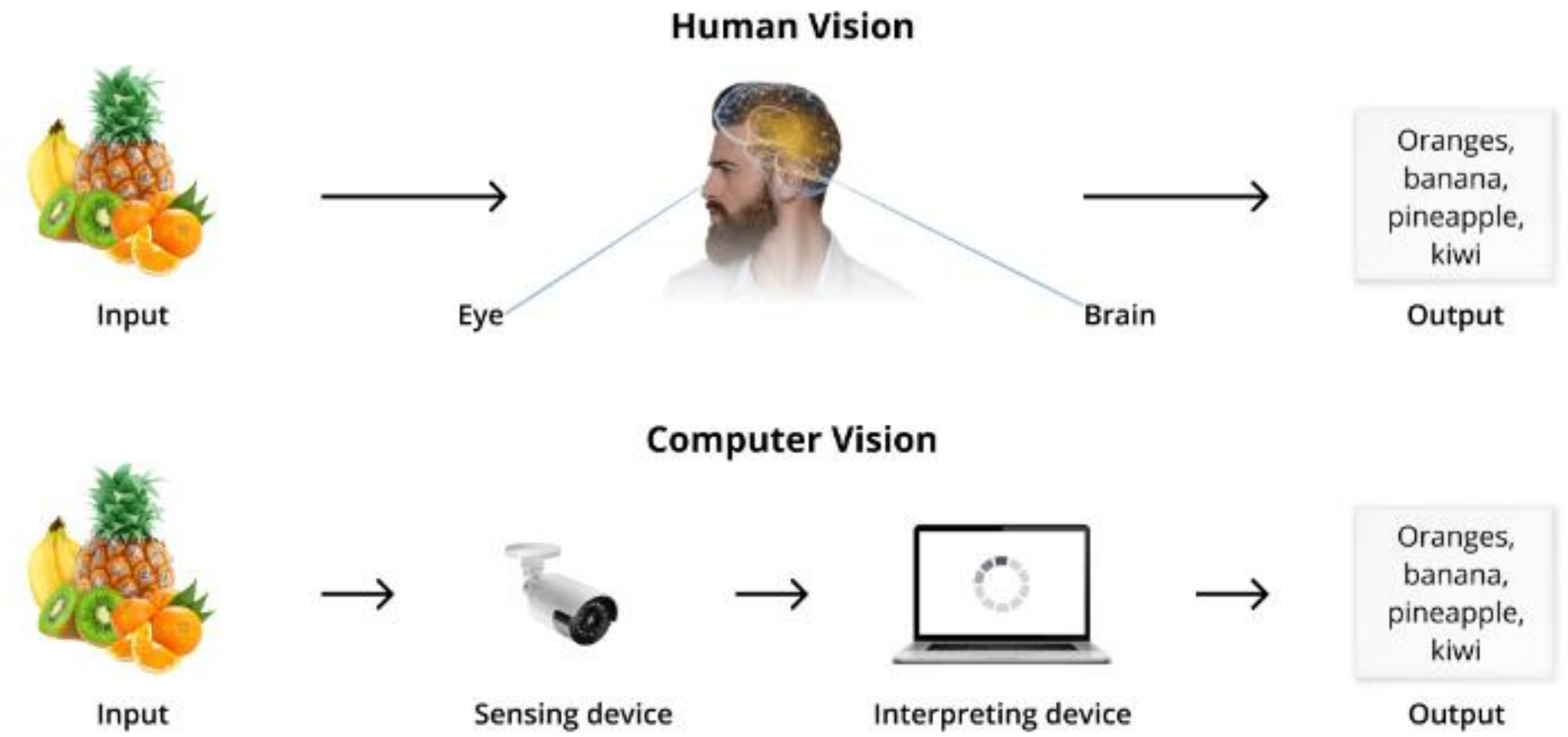
What kind of scene?

Where are the cars?

How far is the building?

…

How does CV work?



Computer vision vs human vision

What we see

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

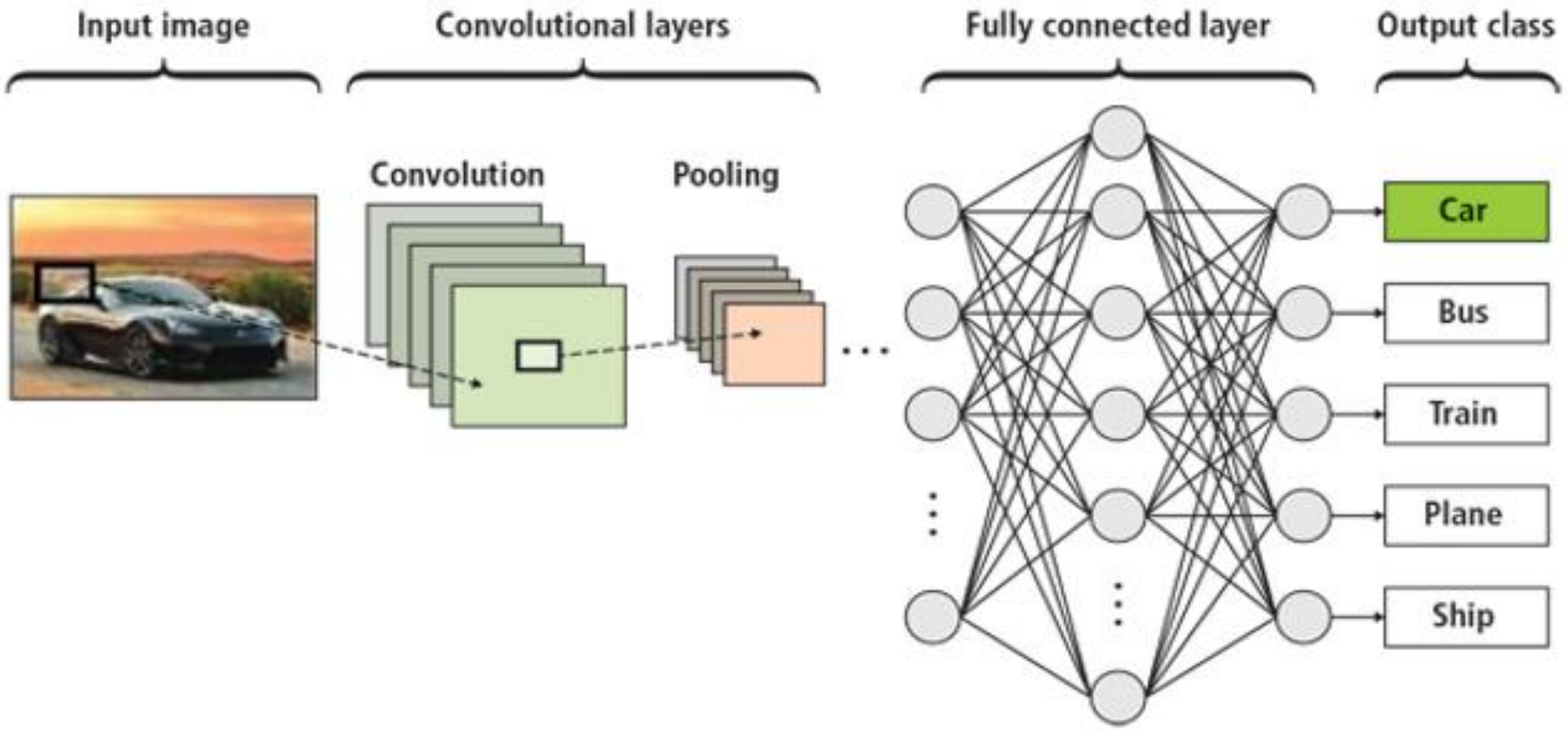
What a computer sees

How does CV work?

- CV analyzes images using CNN.
- CNNs create numerical representations of what is seen in the images.
- CNNs use convolutional layers (CL) to filter input data for useful information.
- Convolution involves combining input data with a convolution kernel to form a transformed feature map.
- CL modify filters based on learned parameters for specific tasks.
- CNNs adjust automatically to find the best features for a given task.
- CNNs differentiate between objects based on their shape for general object recognition tasks.
- CNNs differentiate between objects based on their color for specific tasks like bird recognition.
- CNNs understand that different object classes have different shapes.

How does CV work?

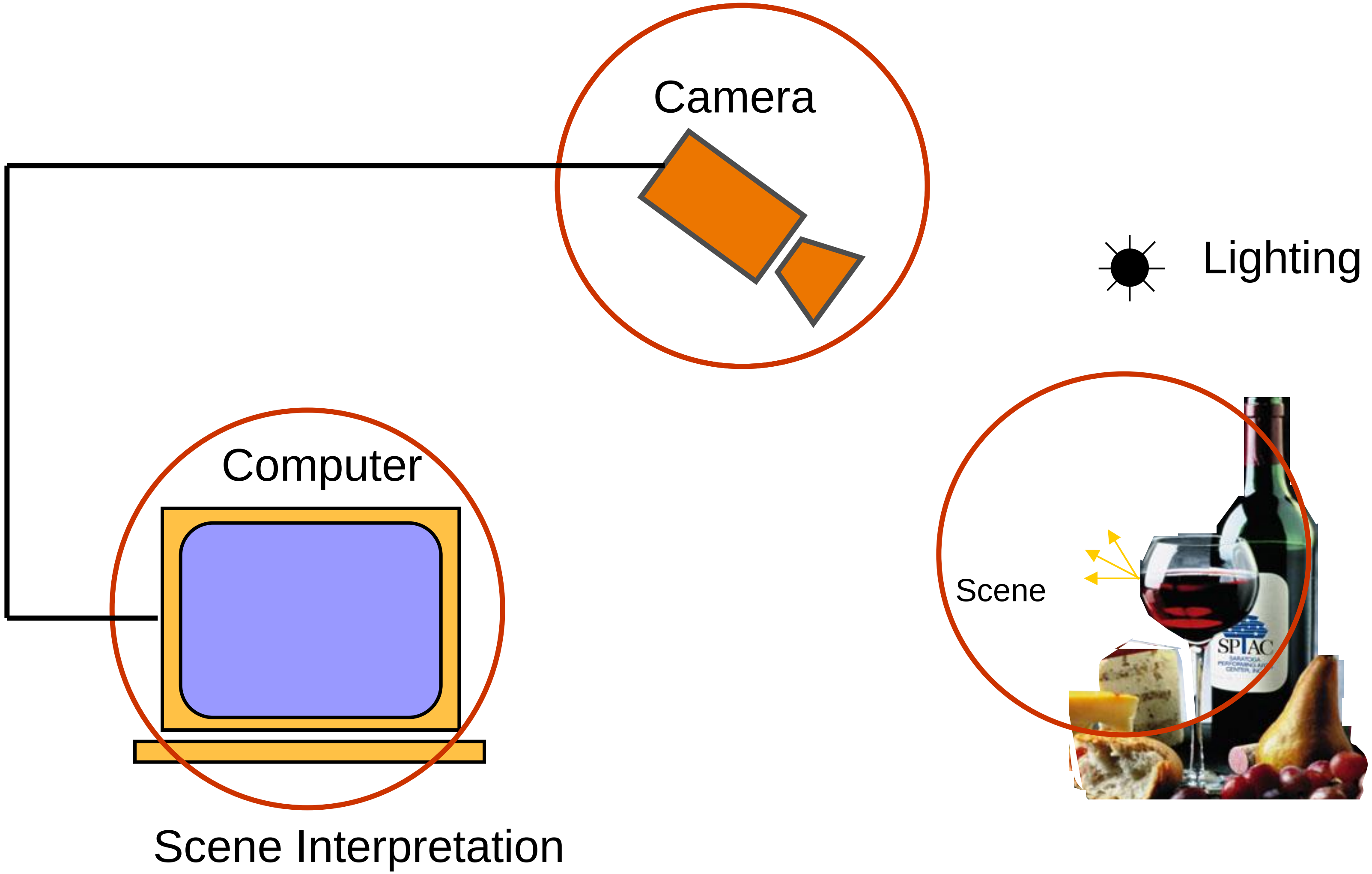
Computer vision can include specific training of CNNs for segmentation, classification, and detection using images and videos for data.



How does CV work?

Segmentation	Classification	Detection
Good at defining objects	Is it a cat or a dog?	Where does it exist in space?
Used in self-driving vehicles	Classifies with precision	Recognizes things for safety

Components of a computer vision system



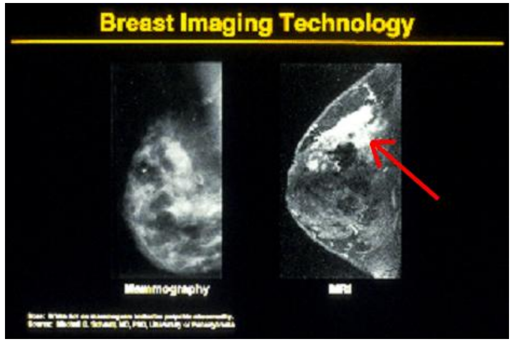
Why study it?

- Replicate human vision to allow a machine to see:
  - Central to that problem of Artificial Intelligence
  - Many industrial applications

- Gain insight into how we see:
  - Vision is explored extensively by neuroscientists to gain an understanding of how the brain operates (e.g. the Center for Neural Science at NYU)

Why computer vision matters

Safety

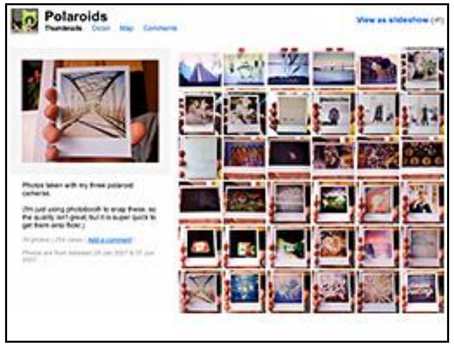


Health

Security

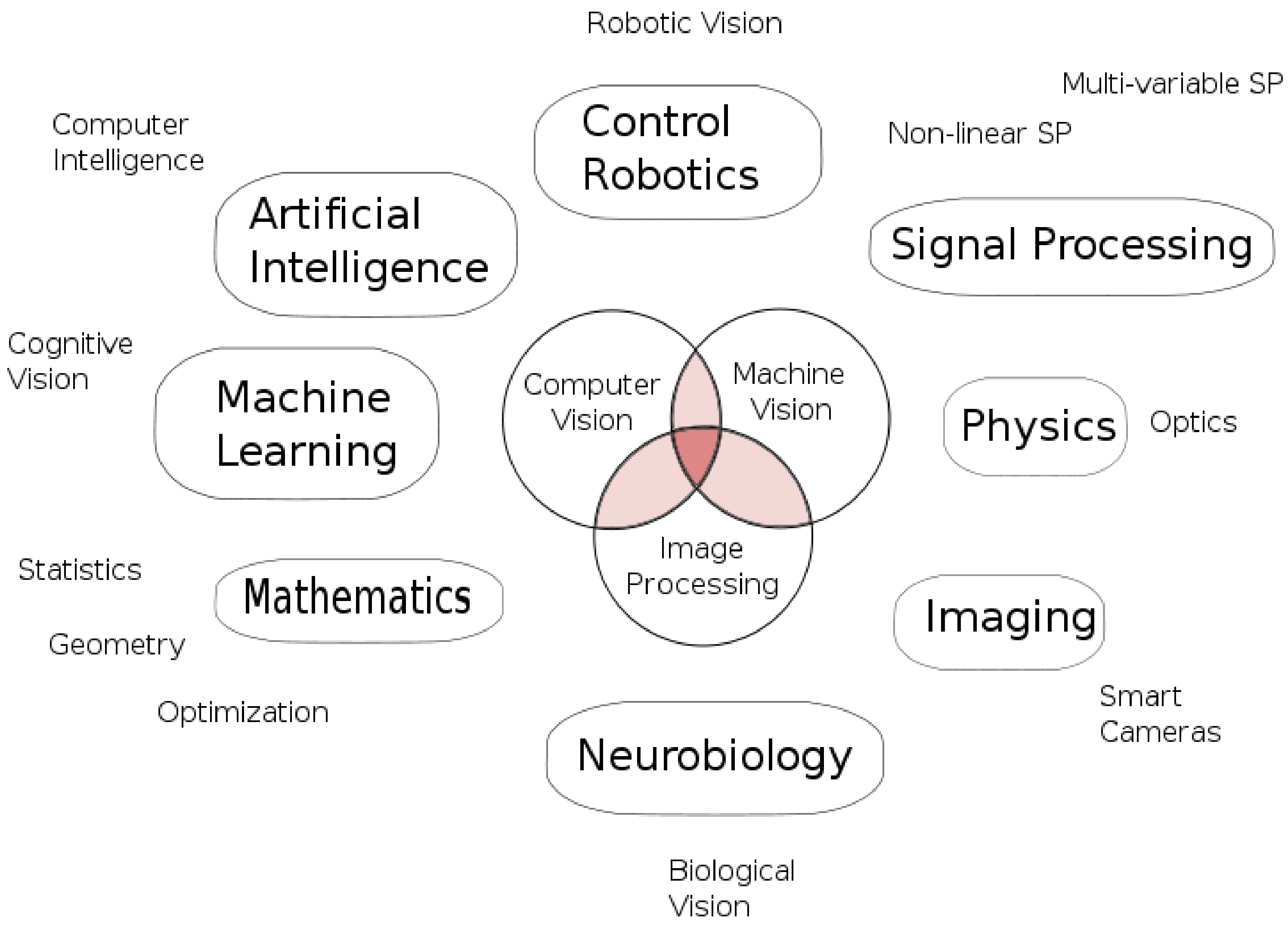
Comfort

Fun

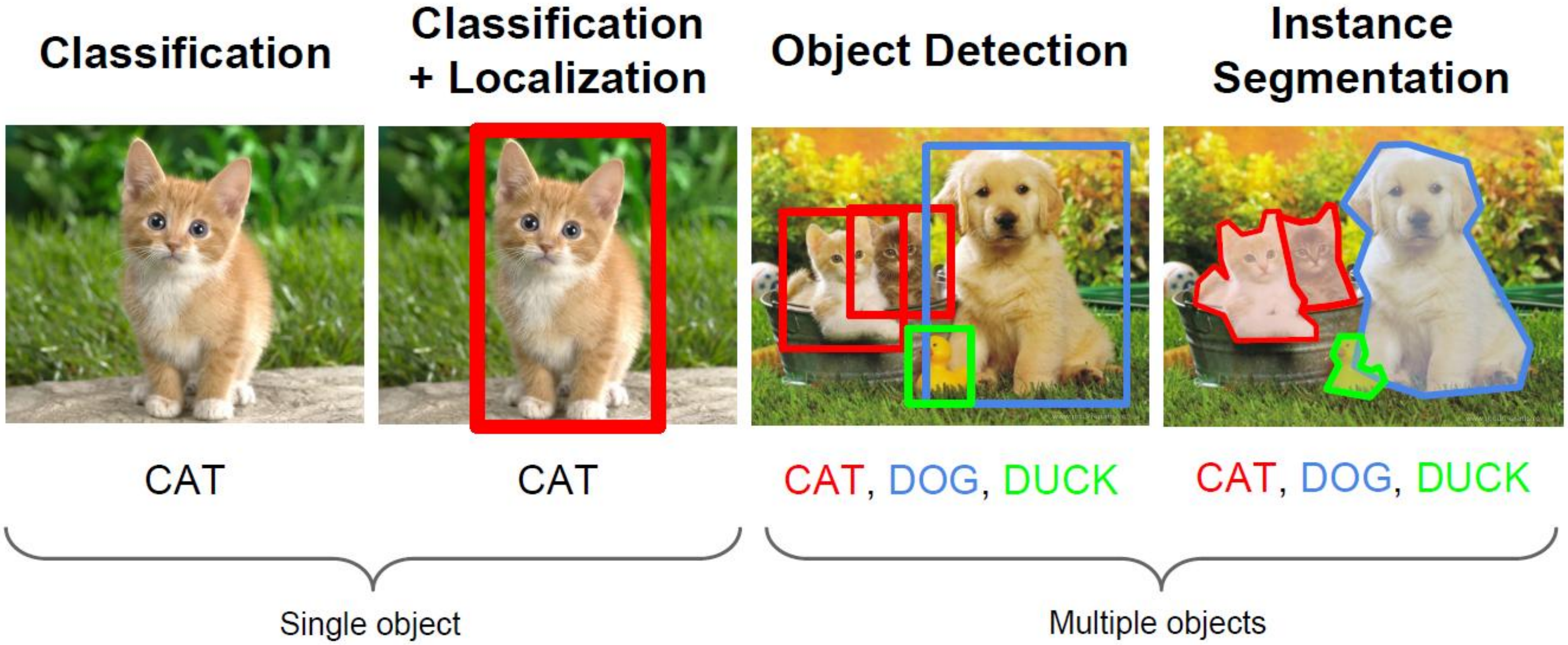


Access

Vision is multidisciplinary



Computer Vision Tasks



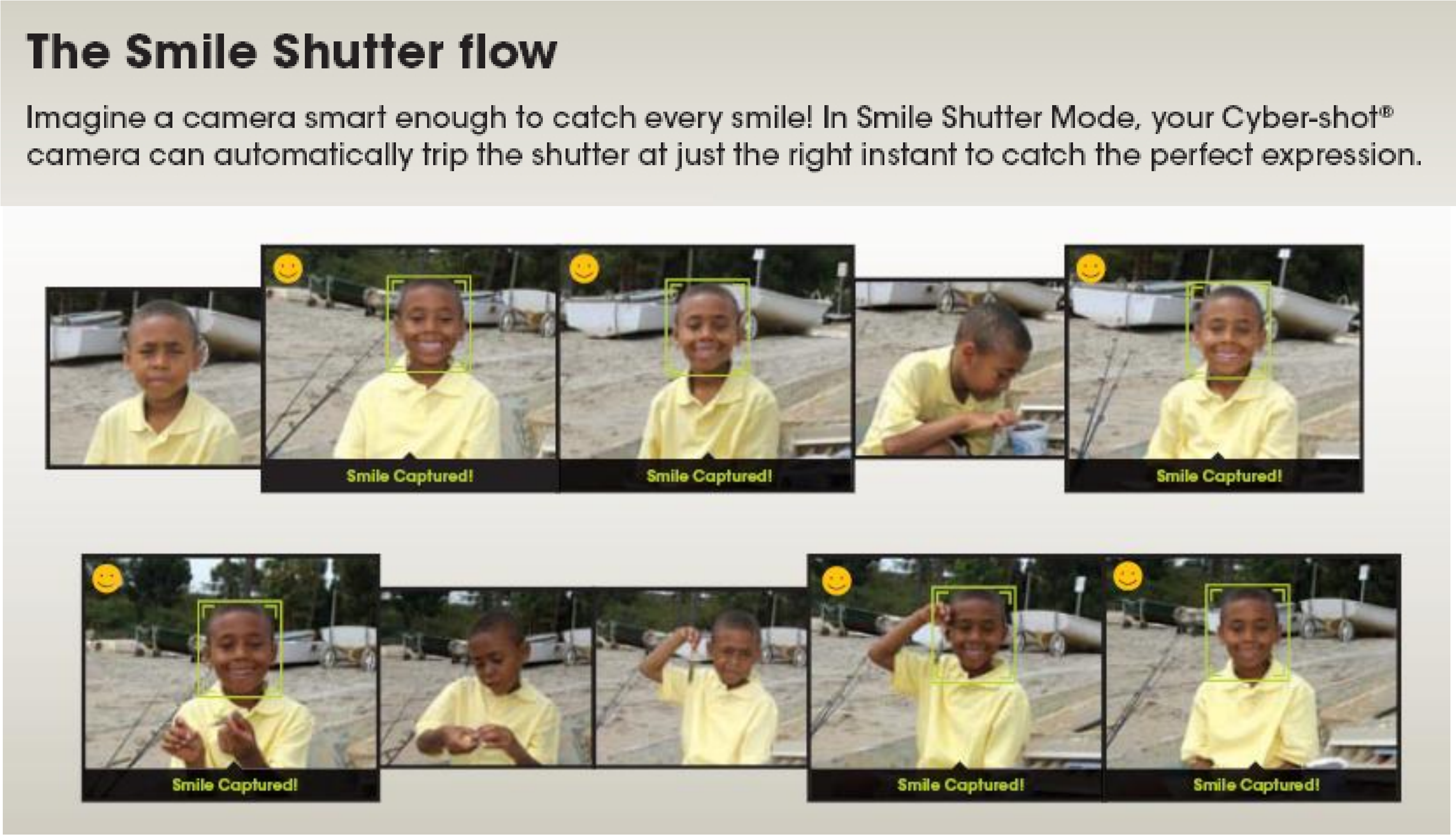
Picture from: Fie-Fei Li, Andrej Karpathy, Justin Johnson – Understanding and Visualizing CNNs

Applications of computer vision

Face detection

- Many new digital cameras now detect faces
  - Canon, Sony, Fuji, …

Smile detection



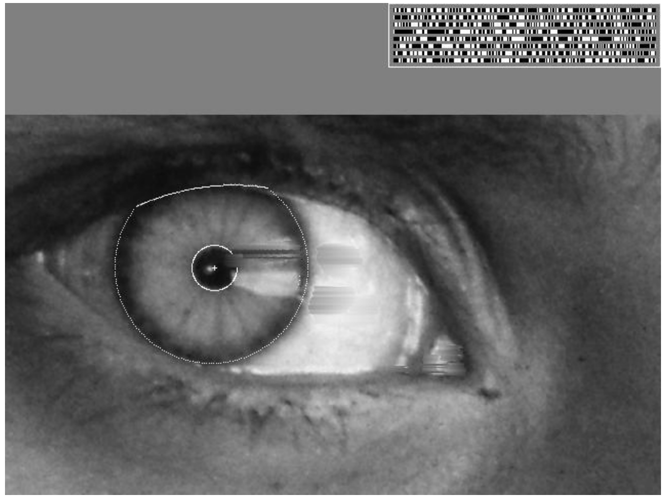
[Sony Cyber-shot® T70 Digital Still Camera]

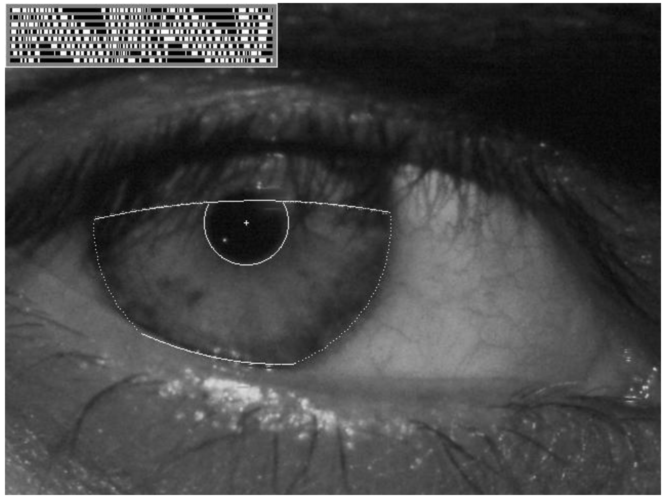
Object recognition (in supermarkets)

LaneHawk by EvolutionRobotics

“A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized, the cashier verifies the quantity of items that were found under the basket, and continues to close the transaction. The item can remain under the basket, and with LaneHawk,you are assured to get paid for it… “

Vision-based biometrics





Login without a password…

Fingerprint scanners on
many new laptops,
other devices

Face recognition systems now
beginning to appear more widely
http://www.sensiblevision.com/

Object recognition (in mobile phones)

[Point & Find](), [Nokia]() [Google Goggles]()

Special effects: Shape capture

The Matrix movies, ESC Entertainment, XYZRGB, NRC

Special effects: Motion capture

Pirates of the Carribean, Industrial Light and Magic

Sports

Sportvision first down line
Nice explanation on www.howstuffworks.com

http://www.sportvision.com/video.html

Smart cars

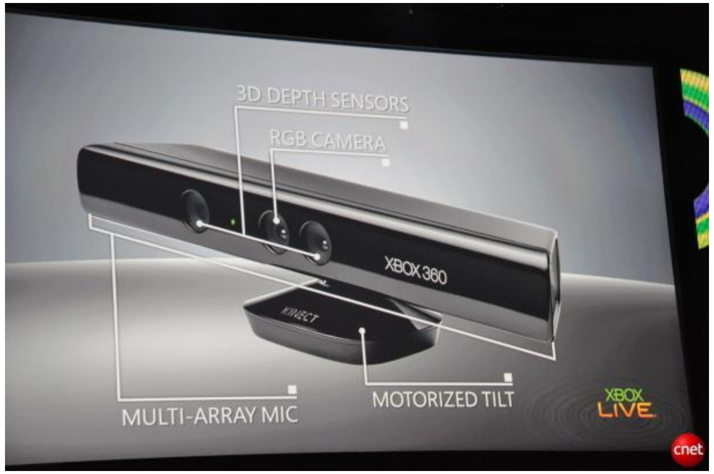
- Mobileye
  - Vision systems currently in many car models

Google cars

http://www.nytimes.com/2010/10/10/science/10google.html?ref=artificialintelligence

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
- 3D tracking, reconstruction, and interaction: http://research.microsoft.com/en-us/projects/surfacerecon/default.aspx



Vision in space

NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

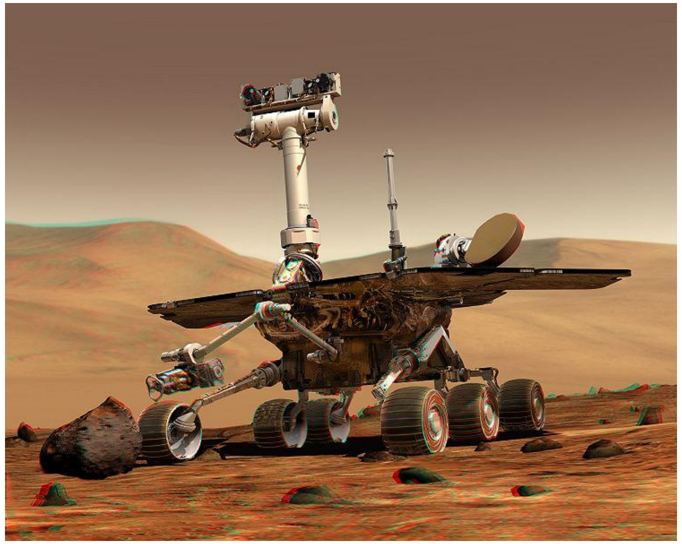
Vision systems (JPL) used for several tasks

- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking

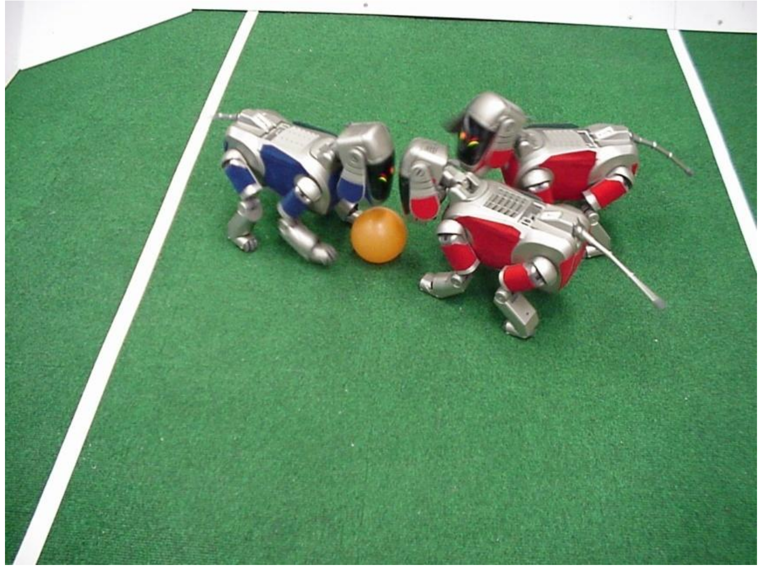
Industrial robots

Vision-guided robots position nut runners on wheels

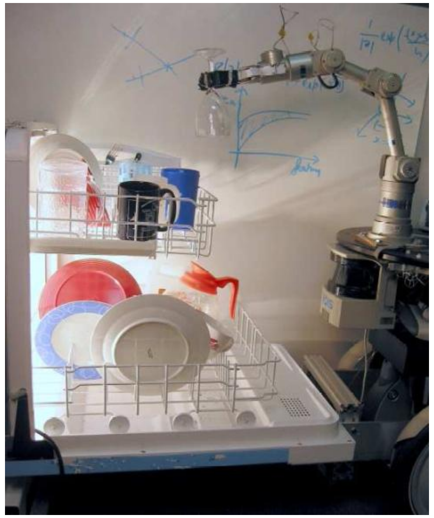
Mobile robots



NASA’s Mars Spirit Rover
http://en.wikipedia.org/wiki/Spirit_rover

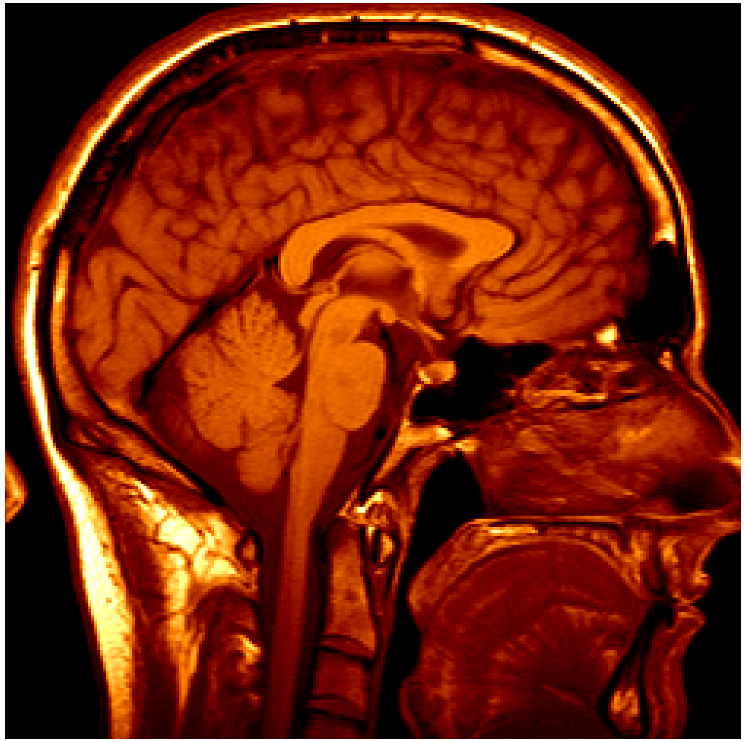


http://www.robocup.org/



Saxena et al.
2008
STAIR at Stanford

Medical imaging



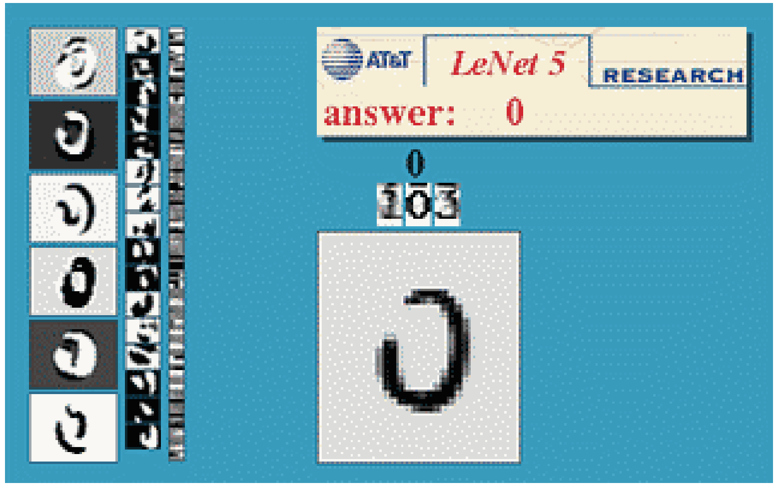
3D imaging
MRI, CT

Image guided surgery
Grimson et al., MIT

Optical character recognition (OCR)

Technology to convert scanned docs to text

- If you have a scanner, it probably came with OCR software



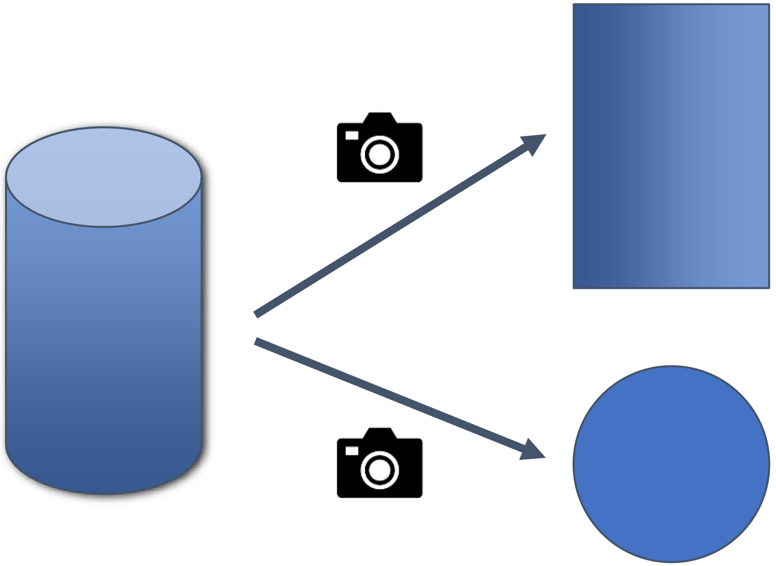
Digit recognition, AT&T labs
http://www.research.att.com/~yann/

License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

Challenges in CV

- Semantic gap
  - RGB array or jpeg -> useful representation?
  - Pixels -> high-level understanding
  - Priors, other modalities
- Data dimensionality
- Ambiguity
- Infinite 3D reprojections

```
[[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0]
 [0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0]
 [0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
```



Challenges: viewpoint variation

Michelangelo 1475-1564

slide credit: Fei-Fei, Fergus & Torralba

Challenges: illumination

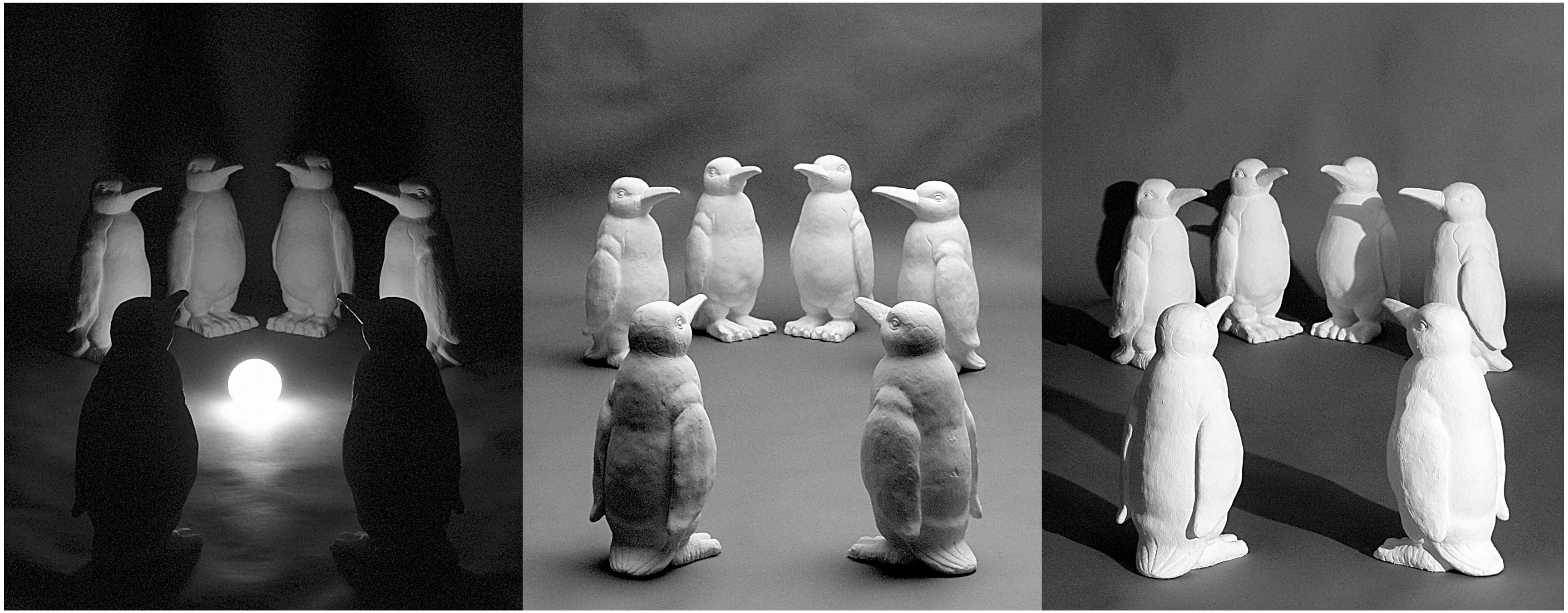
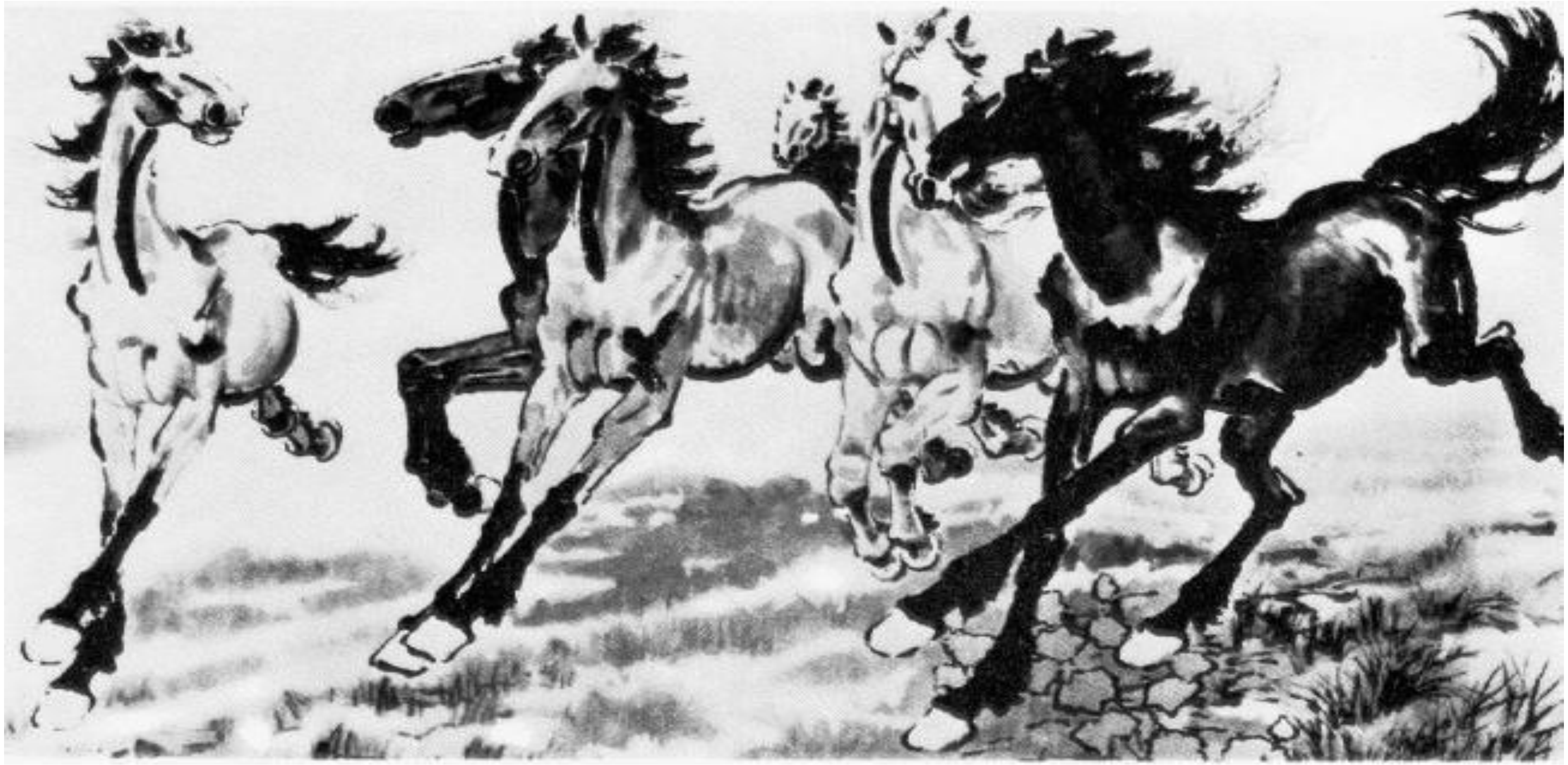


image credit: J. Koenderink

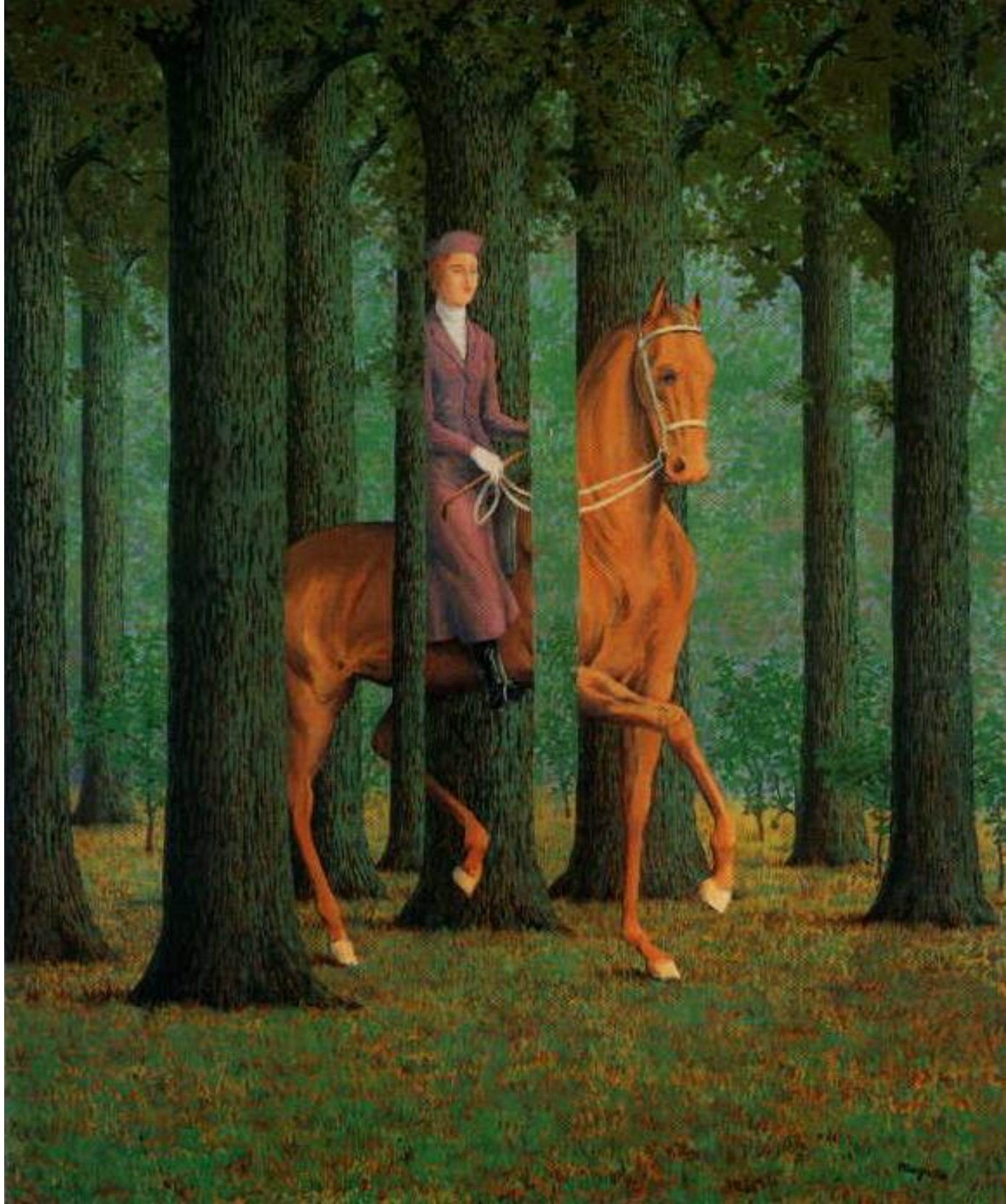
Challenges: scale

Challenges: deformation



Xu, Beihong 1943

Challenges: occlusion



Magritte, 1957

Challenges: background clutter

Chihuahua or Muffin?

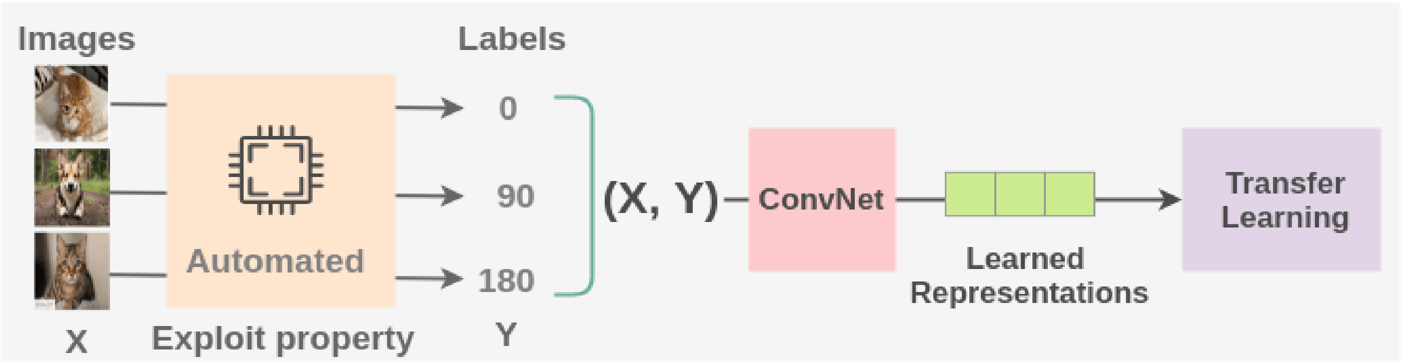
Self-Supervised Learning in Vision

Supervised vs Unsupervised Learning

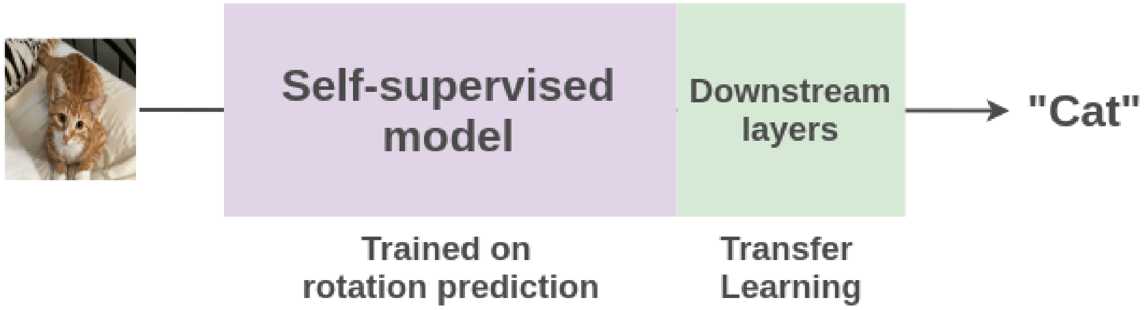
- ***Supervised learning*** – learning with labeled data
  - Approach: collect a large dataset, manually label the data, train a model, deploy
  - It is the dominant form of ML at present
  - Learned feature representations on large datasets are often transferred via pre-trained models to smaller domain-specific datasets
- ***Unsupervised learning*** – learning with unlabeled data
  - Approach: discover patterns in data either via clustering similar instances, or density estimation, or dimensionality reduction …
- ***Self-supervised learning*** – representation learning with unlabeled data
  - Learn useful feature representations from unlabeled data through pretext tasks
  - The term “self-supervised” refers to creating its own supervision (i.e., without supervision, without labels)
  - Self-supervised learning is one category of unsupervised learning

Self-Supervised Learning

- Self-supervised learning example
  - ***Pretext task***: train a model to predict the rotation degree of rotated images with cats and dogs (we can collect million of images from internet, labeling is not required)



  - ***Downstream task***: use transfer learning and fine-tune the learned model from the pretext task for classification of cats vs dogs with very few labeled examples



Picture from: Amit Chaudhary – The Illustrated Self-Supervised Learning

Self-Supervised Learning

- Why self-supervised learning?
  - Creating labeled datasets for each task is an expensive, time-consuming, tedious task
    - Requires hiring human labelers, preparing labeling manuals, creating GUIs, creating storage pipelines, etc.
    - High quality annotations in certain domains can be particularly expensive (e.g., medicine)
  - Self-supervised learning takes advantage of the vast amount of unlabeled data on the internet (images, videos, text)
    - Rich discriminative features can be obtained by training models without actual labels
  - Self-supervised learning can potentially generalize better because we learn more about the world
- ***Challenges*** for self-supervised learning
  - How to select a suitable pretext task for an application
  - There is no gold standard for comparison of learned feature representations
  - Selecting a suitable loss functions, since there is no single objective as the test set accuracy in supervised learning

Self-Supervised Learning

- Self-supervised learning versus unsupervised learning
  - Self-supervised learning (SSL)
    - Aims to extract useful feature representations from raw unlabeled data through ***pretext tasks***
    - Apply the feature representation to improve the performance of ***downstream tasks***
  - Unsupervised learning
    - Discover patterns in unlabeled data, e.g., for clustering or dimensionality reduction
  - Note also that the term “self-supervised learning” is sometimes used interchangeably with “unsupervised learning”
- Self-supervised learning versus transfer learning
  - Transfer learning is often implemented in a supervised manner
    - E.g., learn features from a labeled ImageNet, and transfer the features to a smaller dataset
  - SSL is a type of transfer learning approach implemented in an unsupervised manner
- Self-supervised learning versus data augmentation
  - Data augmentation is often used as a regularization method in supervised learning
  - In SSL, image rotation of shifting are used for feature learning in raw unlabeled data

BatchNorm & ResNets

Normalization vs Batch Normalization

- **Normalization** is applied at the input before data is passed to the network

- **Batch normalization** takes place within the network, specifically within hidden layers.

Batch normalization is used to address issues like exploding gradients and can help improve the training of deep neural networks.

Batch Normalization

- ***Batch normalization layers*** act similar to the data preprocessing steps mentioned earlier
  - They calculate the mean μ and variance σ of a batch of input data, and normalize the data *x* to a zero mean and unit variance
  - I.e., $\hat{x} = \frac{x-\mu}{\sigma}$
- BatchNorm layers alleviate the problems of proper initialization of the parameters and hyper-parameters
  - Result in faster convergence training, allow larger learning rates
  - Reduce the internal covariate shift
- BatchNorm layers are inserted immediately after convolutional layers or fully-connected layers, and before activation layers
  - They are very common with convolutional NNs

Batch Normalization

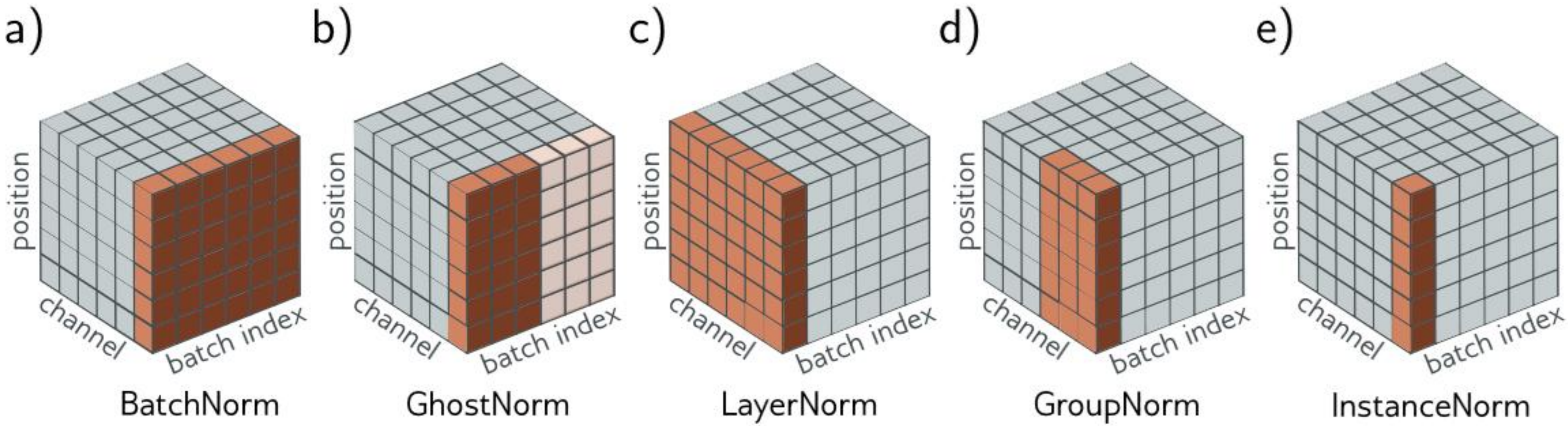
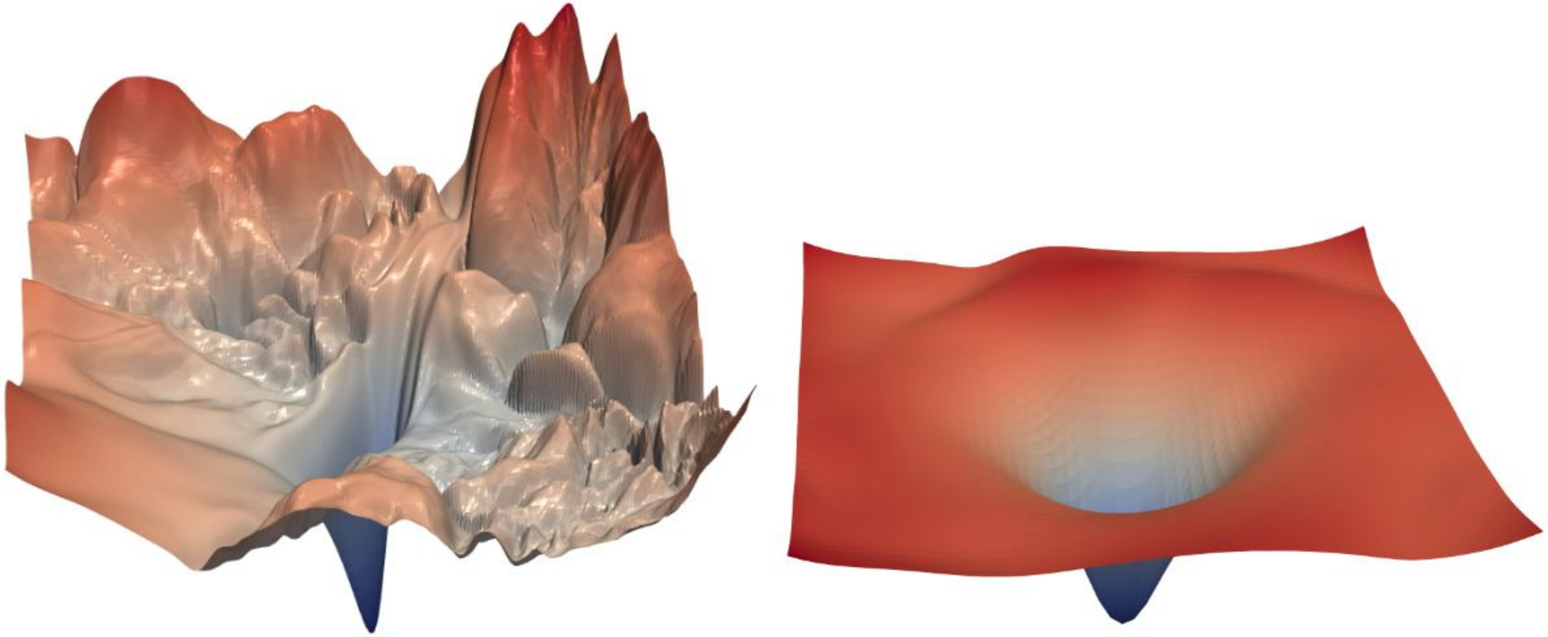


Figure source: [Simon J.D. Prince (2023). Understanding Deep Learning. MIT Press.]

Residual / skip connections - Why?



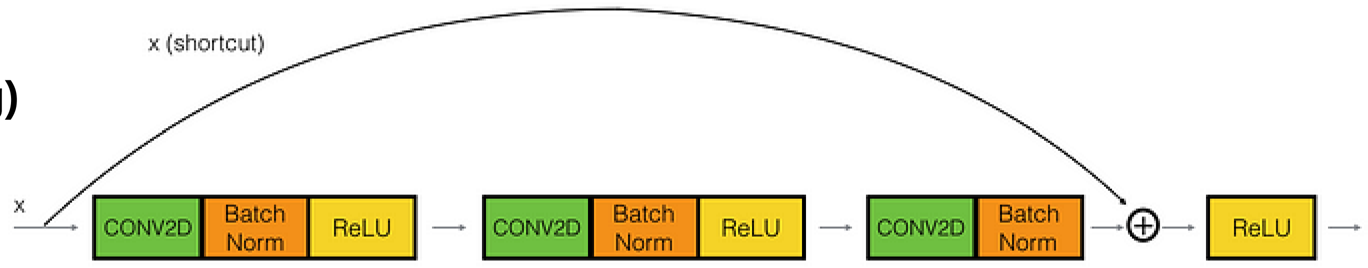
(a) without skip connections (b) with skip connections

Figure source: Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. Neural Information Processing Systems, 31, 6391–6401

Skip connection VS Residual connections

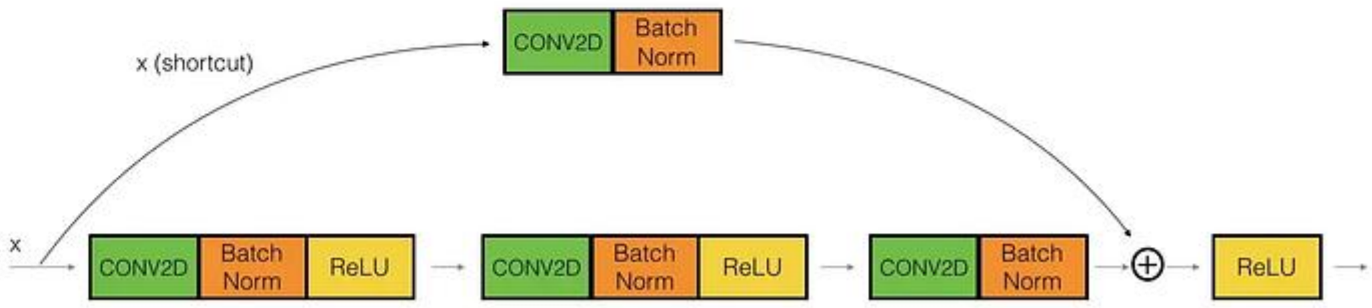
Skip Connection (Skip Connection or Identity Mapping)

- the input to the layer is simply added or concatenated to the output of the skipped layer.



Residual Connection (Residual Block or ResNet)

- the input to the block is added to the output of the block, which is the result of passing the input through one or more layers.
- the output of a layer is not directly added to the input. Instead, it is the difference (residual) between the output and the input that is added to the input.



Data Annotation & Augmentation

What is …

Data Annotation?

- the process of adding tags or labels to data:
  - you can do this manually or automatically

Annotated Datasets?

- a dataset that has been labeled with information that machine learning algorithms can use:
  - use annotated datasets to train machine learning models

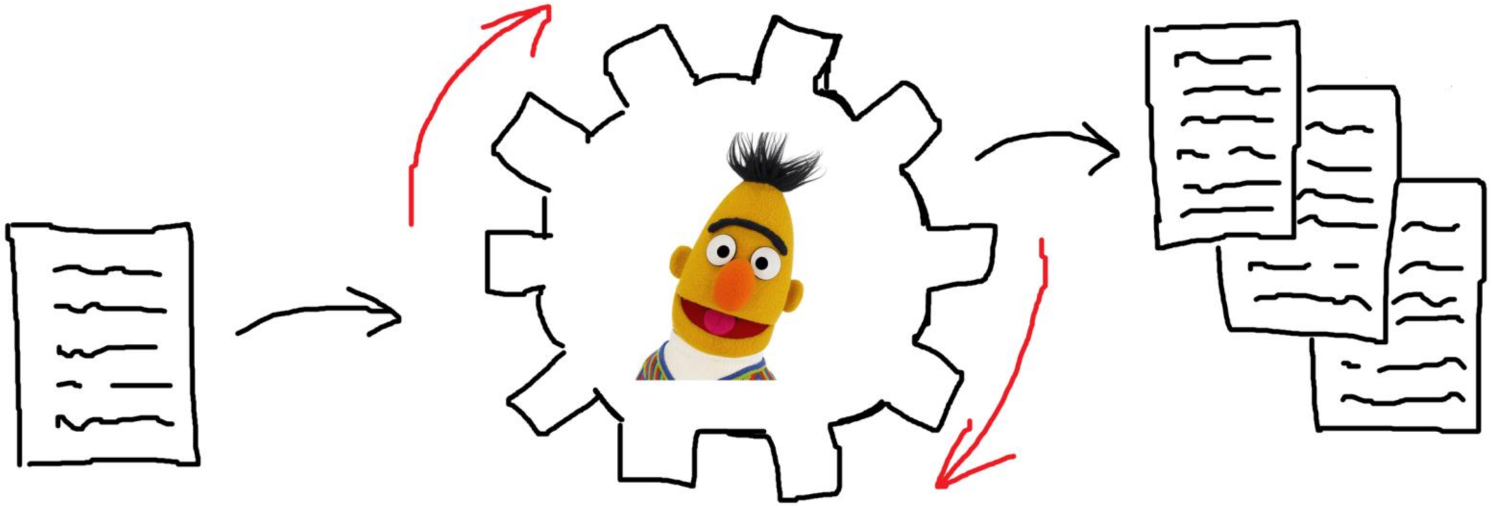
Types of Data Annotation?

Advantages:

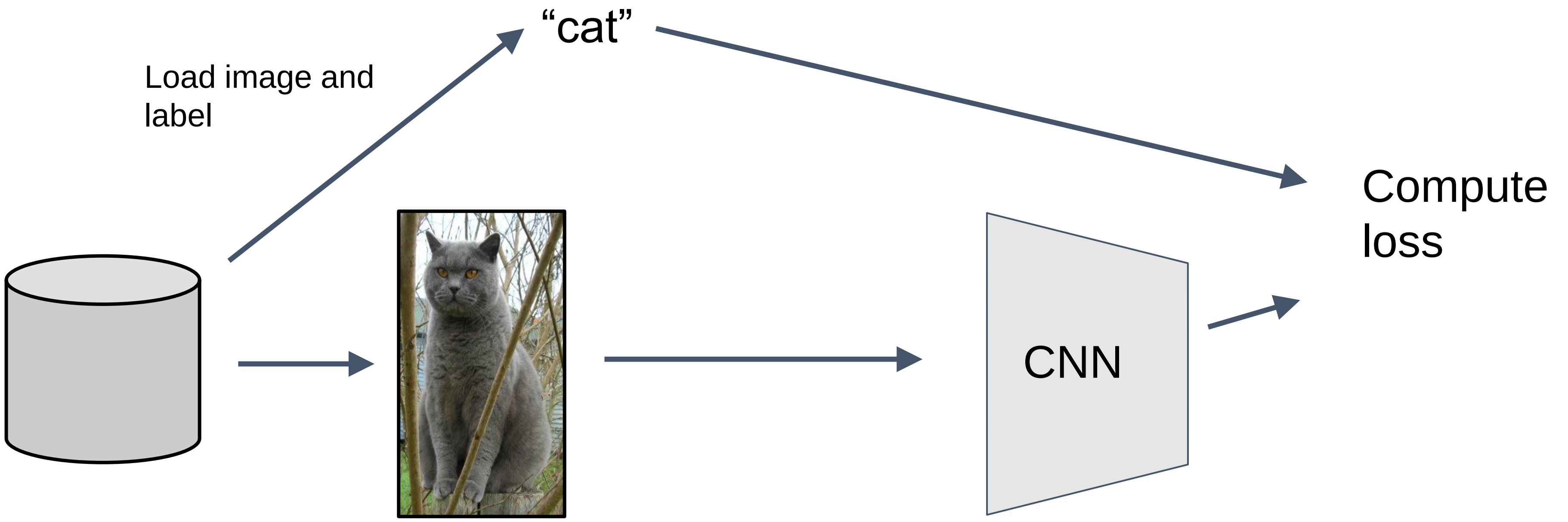
- Cost savings opportunities
- Higher quality of annotation work
- Better scalability
- Timely availability
- Mitigating internal bias
- Increased data security

What is Data Augmentation

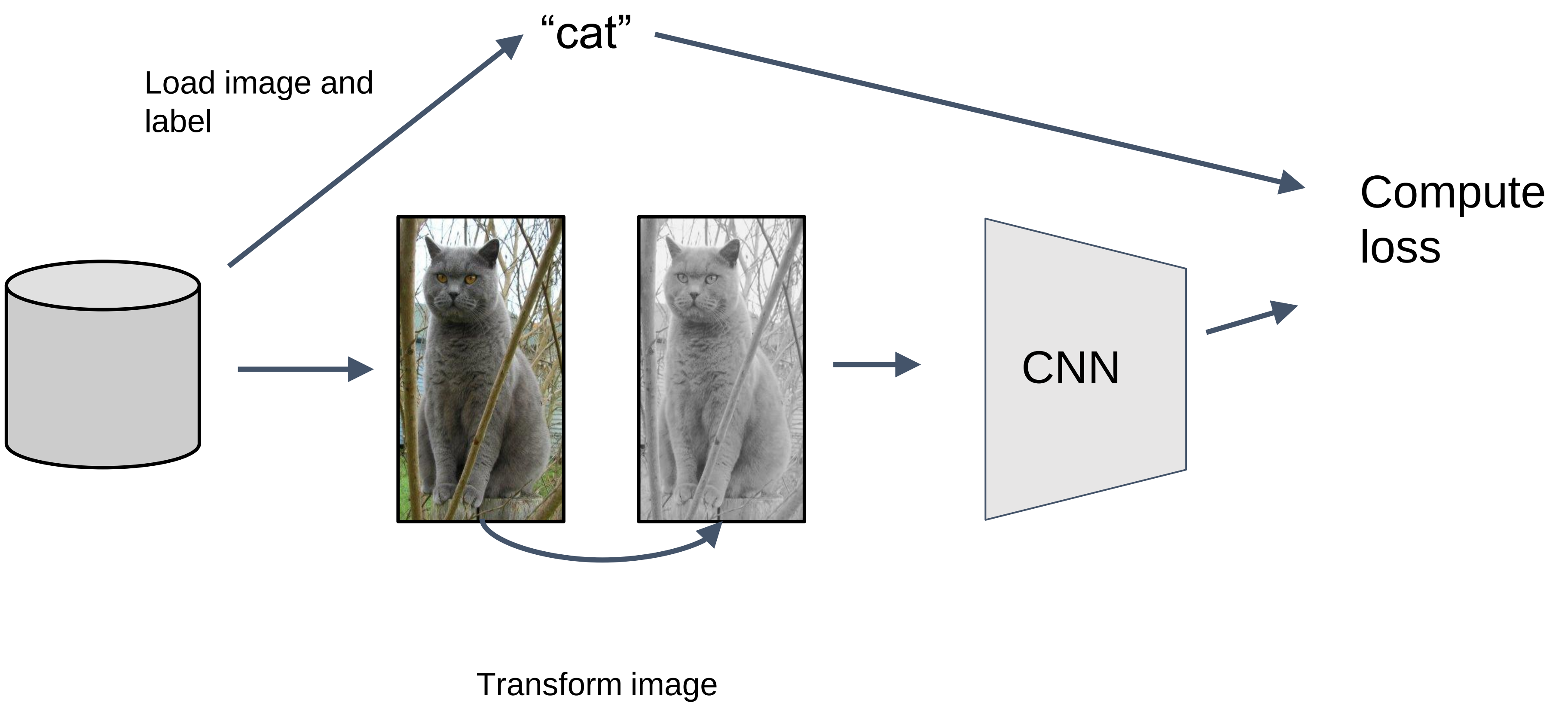
- It involves applying various transformations to existing data to create additional training examples, reduce overfitting, and improve model generalization.



Data Augmentation

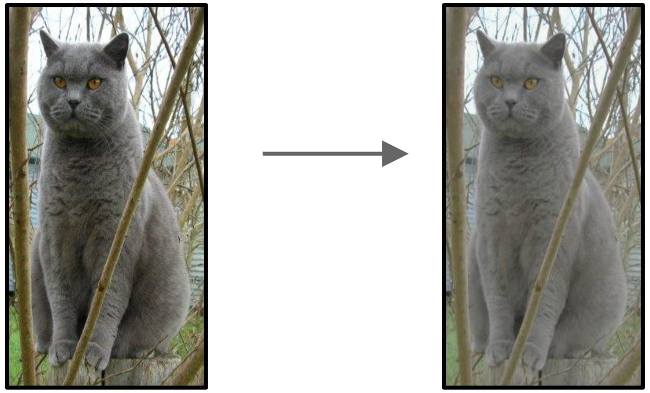


Data Augmentation

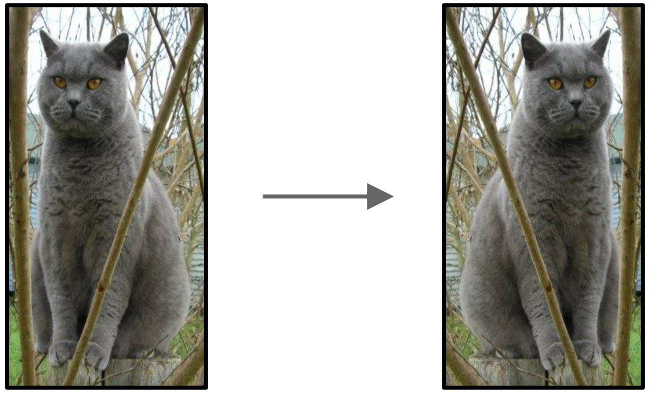


Data Augmentation

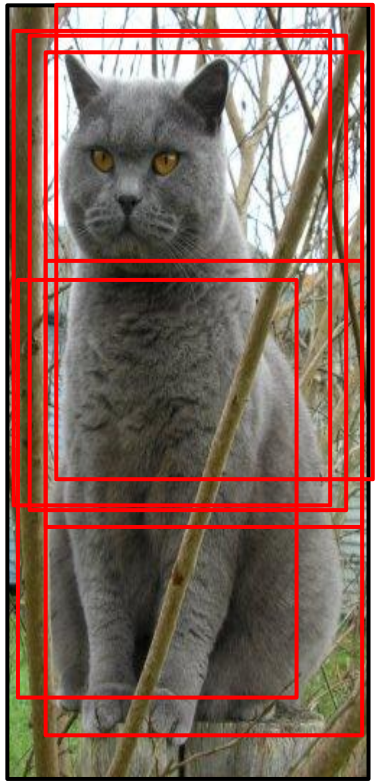
Color jitter



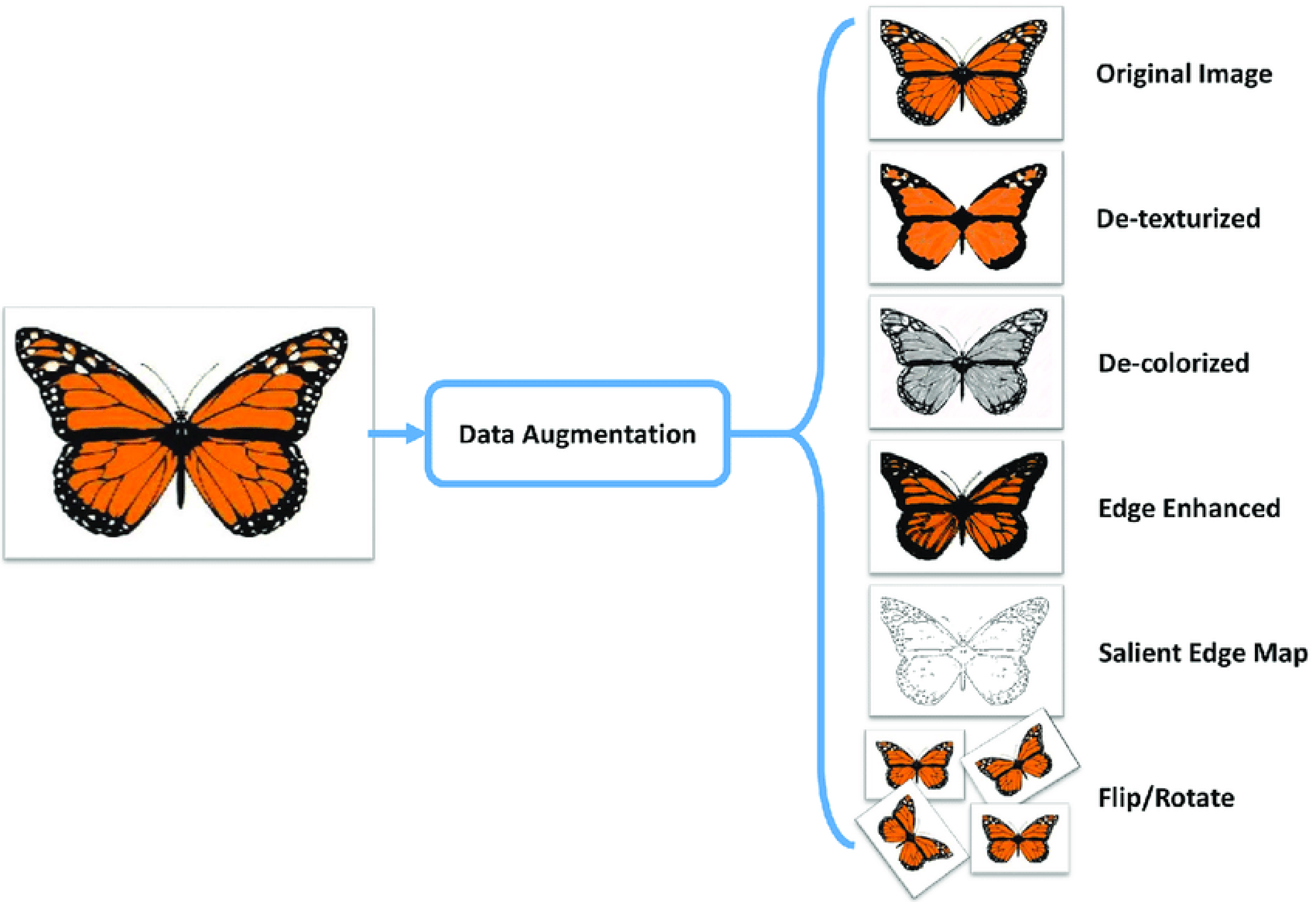
Horizontal flips



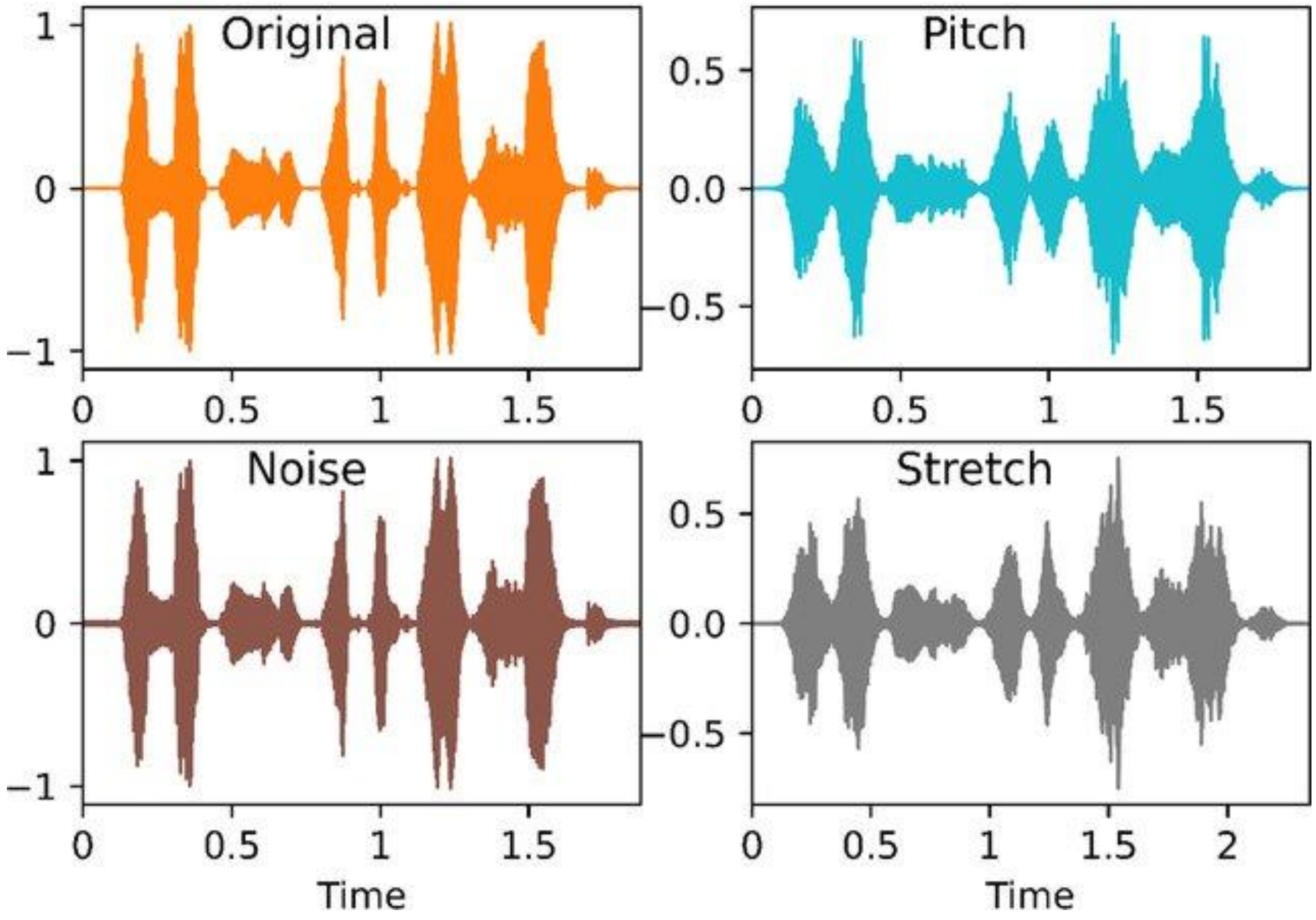
Random crops/scales



Data Augmentation



Data Augmentation



Data Augmentation

Simple to implement, use it

- especially useful for small datasets
- fits into framework of noise / marginalization

Be careful about performance measurements:

- test/train split ***before*** augmentation
- otherwise test data is an “easy” mod of training data

Semantic Segmentation

What is semantic segmentation?

- It is the operation of partitioning an image into a collection of connected sets of pixels.

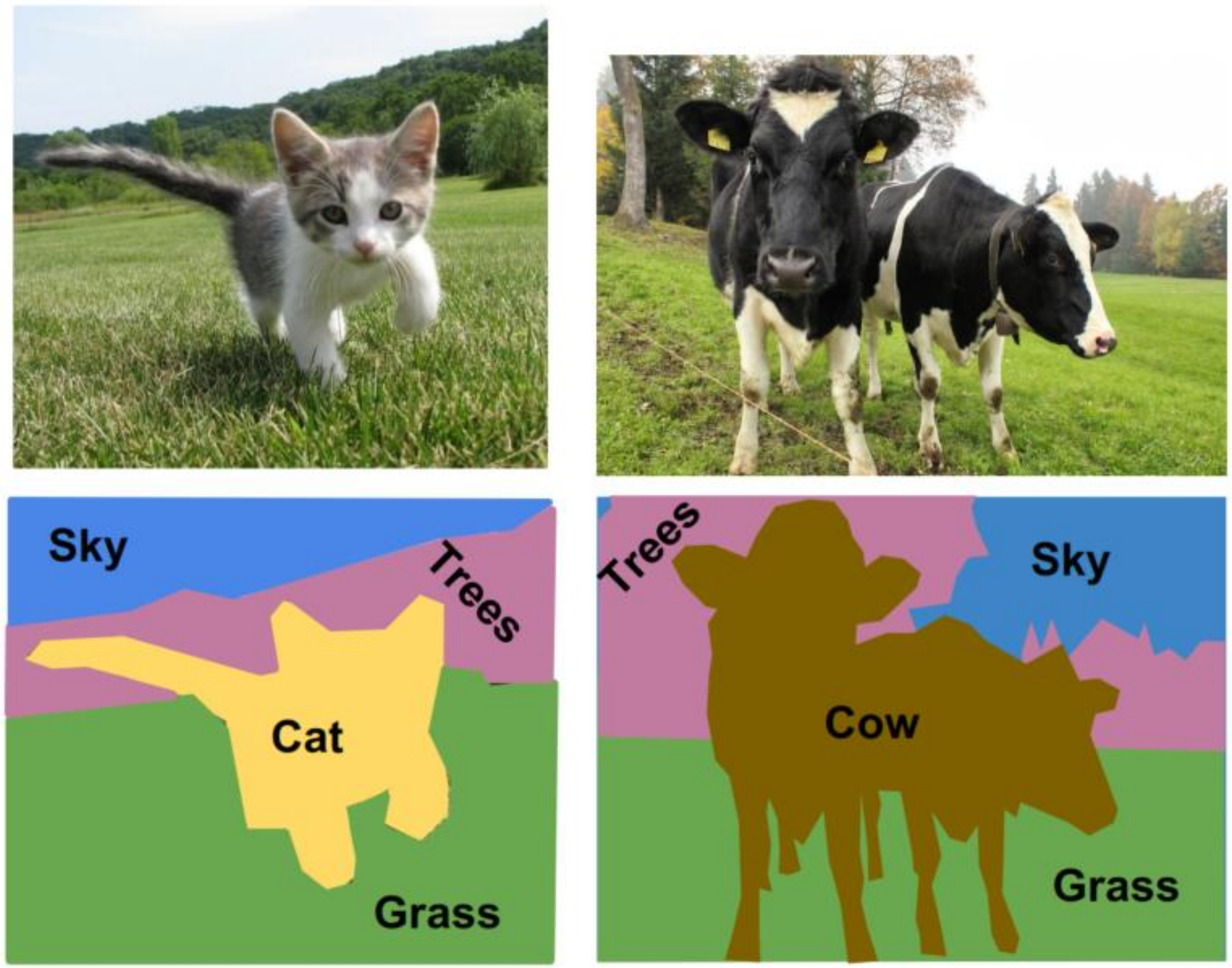
1. into regions, which usually cover the image

2. into linear structures, such as
   - line segments
   - curve segments

3. into 2D shapes, such as
   - circles
   - ellipses
   - ribbons (long, symmetric regions)

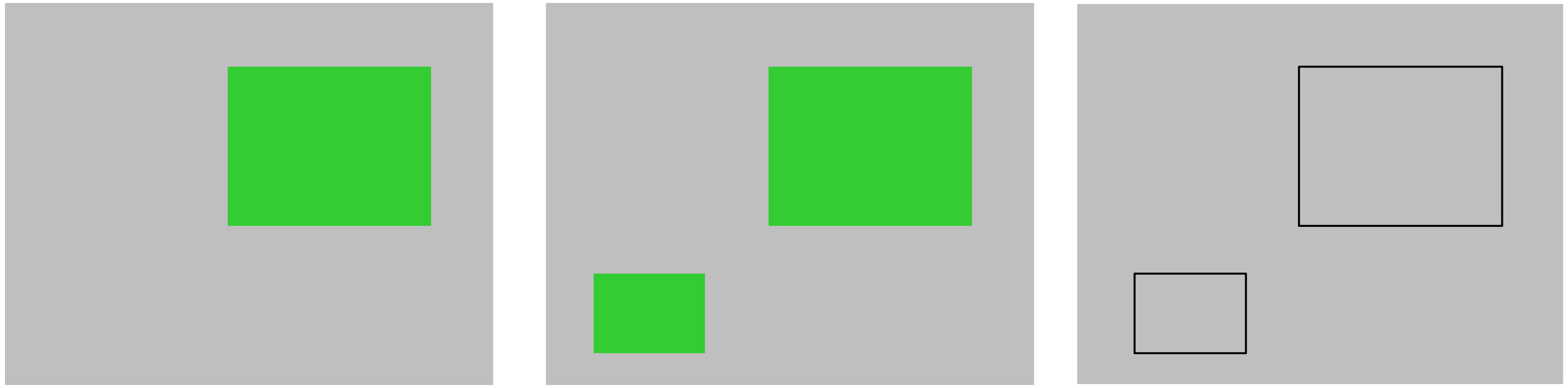
Semantic Segmentation

- Label each pixel in the image with a category label.
- Don’t differentiate instances, only care about pixels



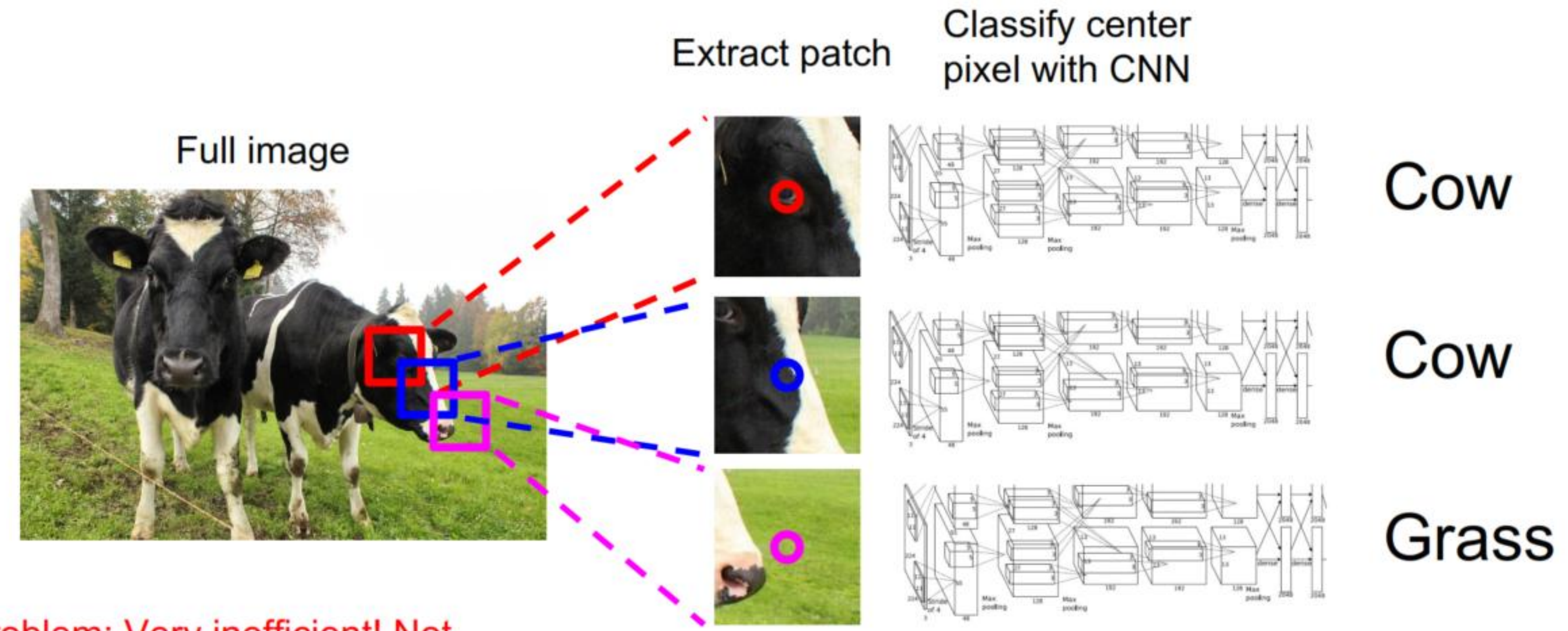
Semantic Segmentation: Issues

- How do we decide that two pixels are likely to belong to the same region?



- How many regions are there?

Semantic Segmentation Idea: Sliding Window

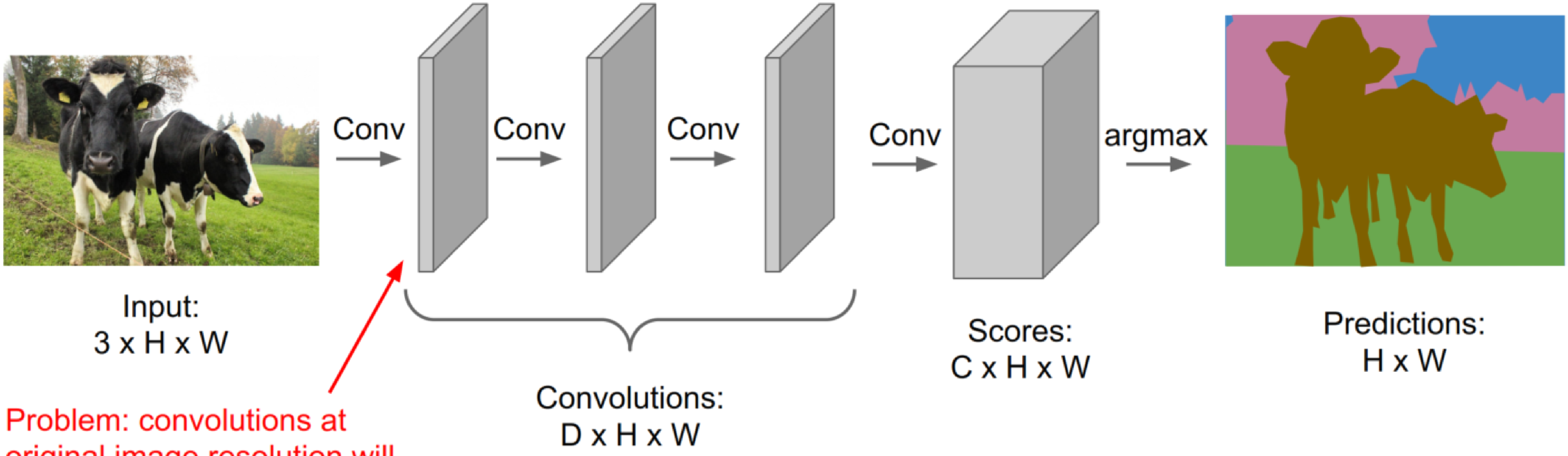


Problem: Very inefficient! Not reusing shared features between overlapping patches

Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation Idea: Fully Convolutional

Design a network as a bunch of convolutional layers
to make predictions for pixels all at once!



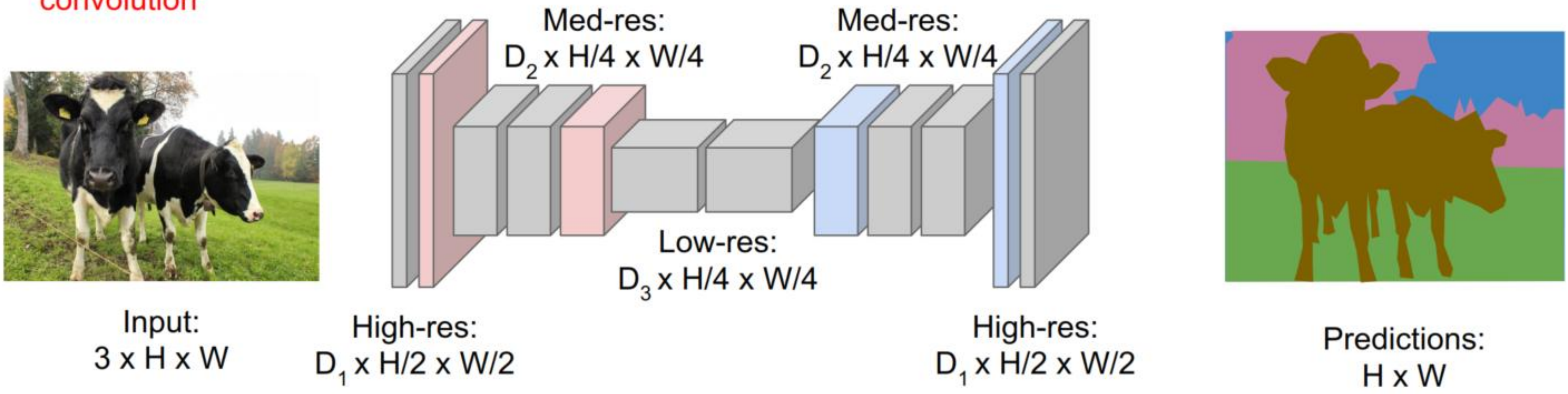
Problem: convolutions at
original image resolution will
be very expensive ...

Semantic Segmentation Idea: Fully Convolutional

Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling: ???



Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Applications of Semantic Segmentation

Autonomous Driving

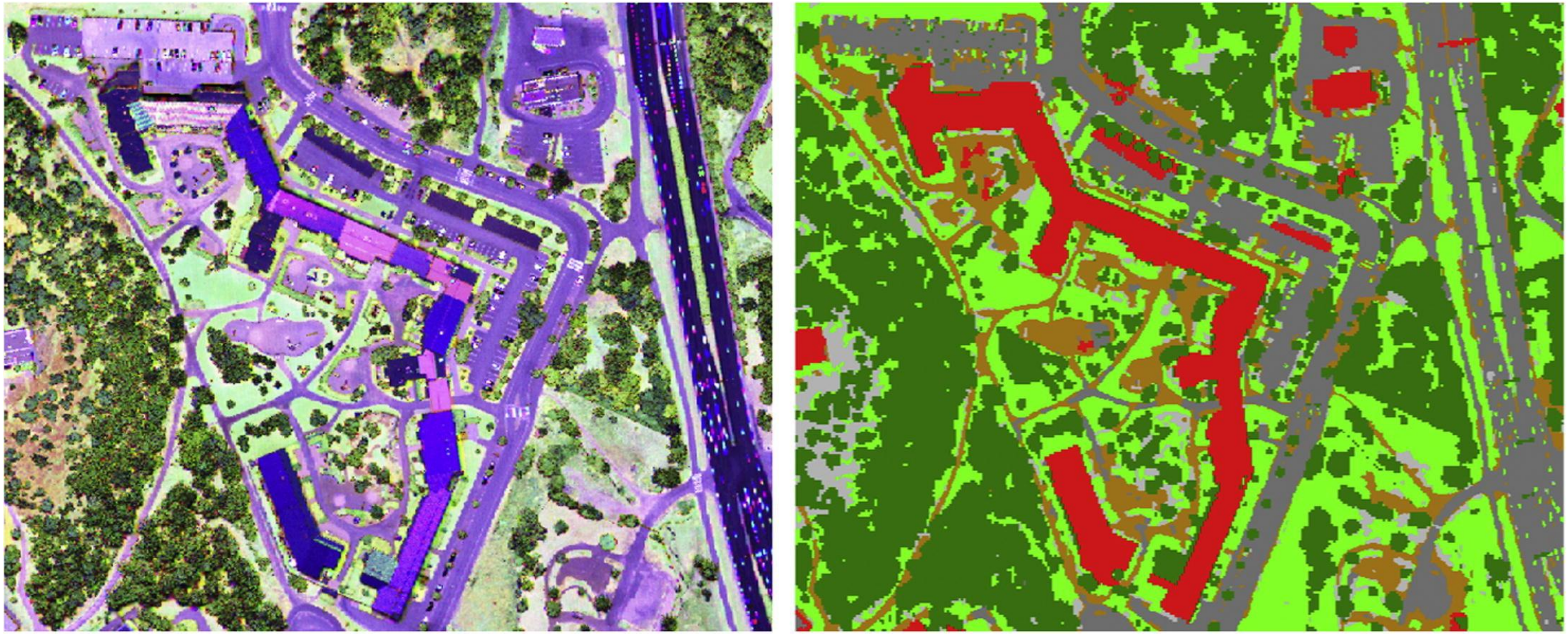
Facial Segmentation

Applications of Semantic Segmentation

Indoor Object Segmentation

Applications of Semantic Segmentation

Geo Land Sensing



Segmentation as clustering

- Cluster together (pixels, tokens, etc.) that belong together…
- Agglomerative clustering
  - attach closest to cluster it is closest to
  - repeat
- Divisive clustering
  - split cluster along best boundary
  - repeat
- Dendrograms
  - yield a picture of output as clustering process continues

Semantic Segmentation using Torchvision

https://youtu.be/doGyJokDoWM

Please, don’t forget to send feedback:

https://bit.ly/bme-dl

Thank you
for your attention

Dr. Mohammed Salah Al-Radhi
malradhi@tmit.bme.hu

(slides by: Dr. Bálint Gyires-Tóth)

29 October 2024