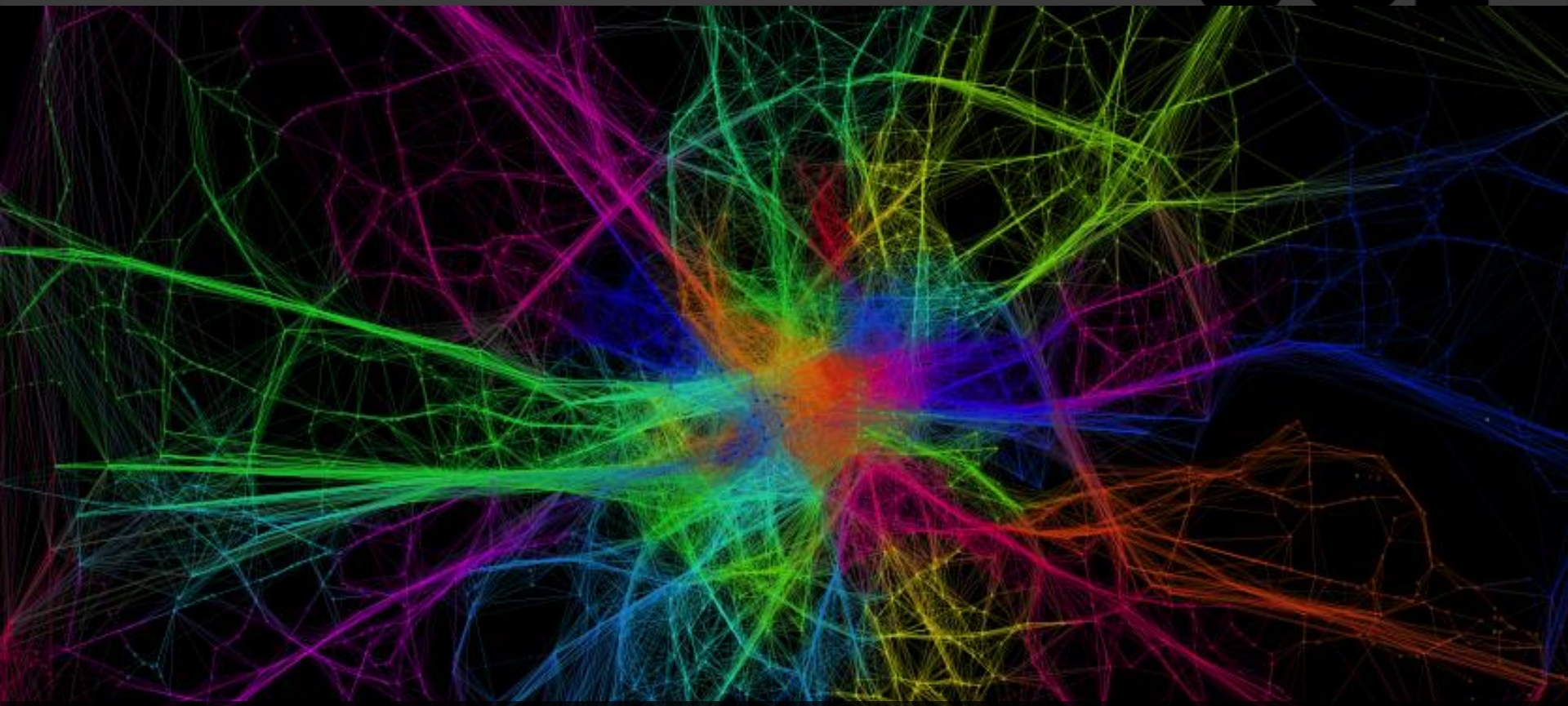


Handling Unstructured Data

CASA0006: Data Science for Spatial Systems

Huanfa Chen

Some slides courtesy of Kira Kempinska

CASA0006

1. Introduction to Databases
2. Introduction to SQL
3. Advanced SQL
4. Data Munging
5. Advanced Clustering
6. Advanced Regression
7. Classification
8. Dimension Reduction
9. Unstructured Data
10. Analysis Workflow

Recap

What we already know

We can handle data

Using a database accessed through SQL, and tools such as Pandas we can take raw, unstructured data through to something useful

We can analyse data

Clustering, Regression, Classification, Dimensionality Reduction

So far, we mainly work with tabular data (or structured data). Not all data come in a tabular format.

Can we handle unstructured data?

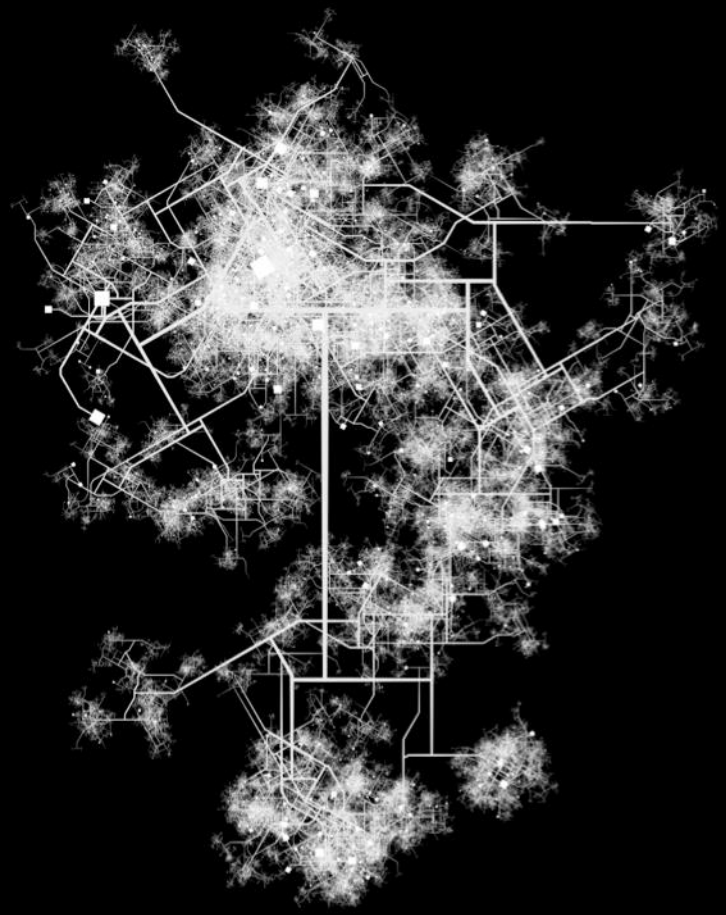
Data Mining

The toolbox

These methods also apply for unstructured data.

Unsupervised = Unlabelled
Supervised = Labelled

Outline



1. Unstructured Data
2. Deep Learning
3. Deep Learning for Images and Videos
4. Deep Learning for Text
5. What Deep Learning Cannot Do
6. Deep Learning Workflow
7. Quick Explanation of Deep Learning

Unstructured Data

Multiple types of data

size of house (square feet)	# of bedrooms	price (1000$)
523	1	115
645	1	0.001
708	unknown	210
1034	3	unknown
unknown	4	355
2545	unknown	440

structured

I read the news today, oh boy
About a lucky man who made the grade
And though the news was rather sad
Well, I just had to laugh
I saw the photograph

He blew his mind out in a car
He didn't notice that the lights had changed
A crowd of people stood and stared
They'd seen his face before
Nobody was really sure if he was from the House of Lords

unstructured

- Anything that cannot be put into traditional row-column or tabular format
- Comprising 80% of all data

https://web.archive.org/web/20110724175845/http://ikt.hia.no/perep/eip_ind.pdf

Examples of unstructured data

video & image

I read the news today, oh boy
About a lucky man who made the grade
And though the news was rather sad
Well, I just had to laugh
I saw the photograph

He blew his mind out in a car
He didn't notice that the lights had changed
A crowd of people stood and stared
They'd seen his face before
Nobody was really sure if he was from the h

I saw a film today, oh boy
The English Army had just won the war

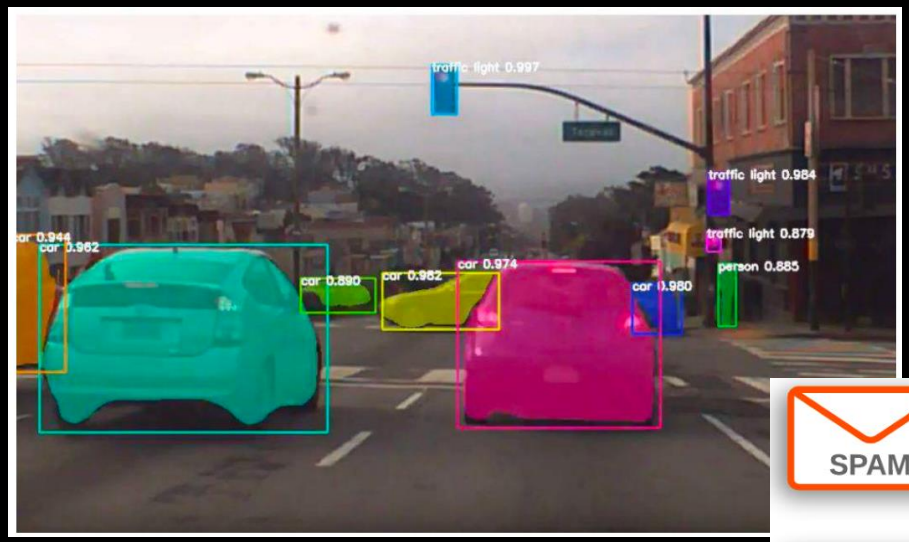
text

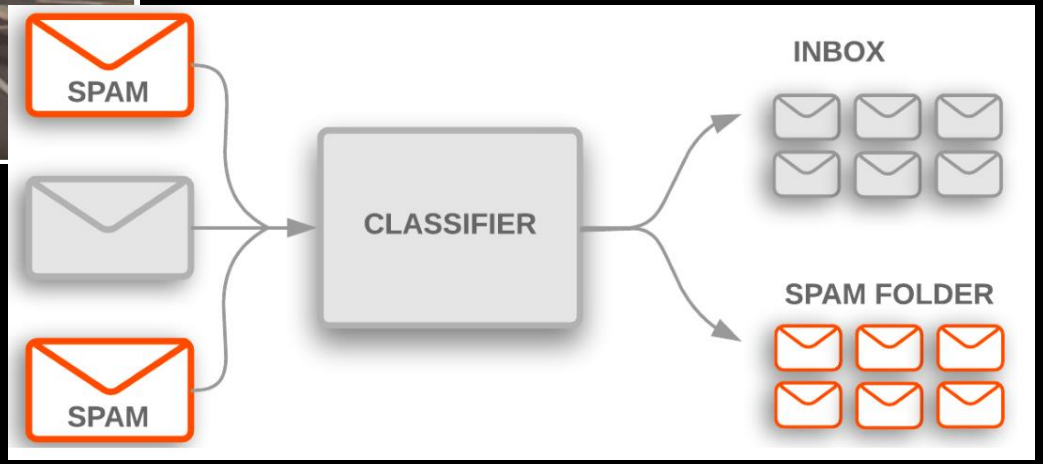
audio

- text, document, network, graph, image, video, audio, web-based, sensor

Applications

Unstructured data in the wild

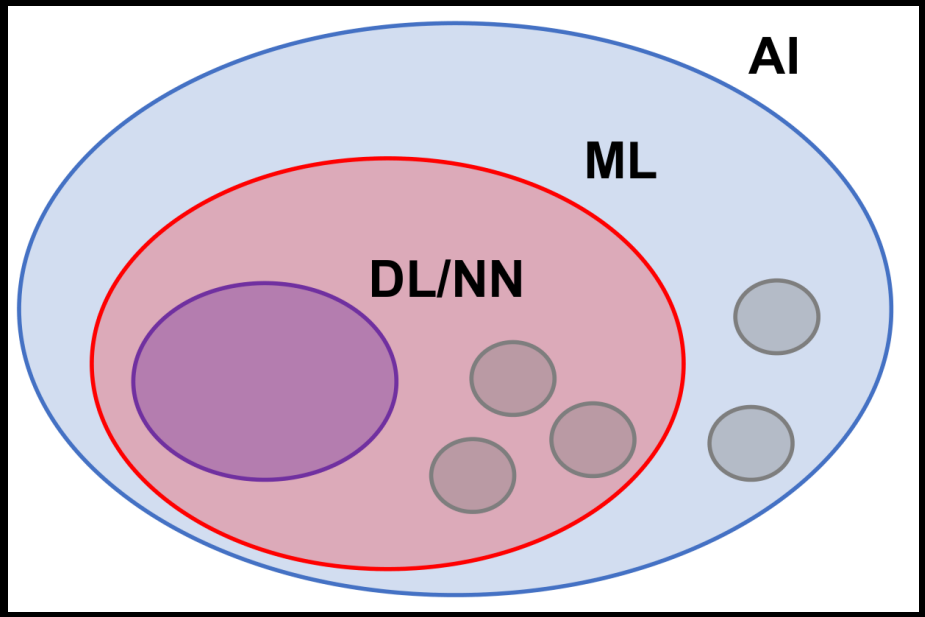




Deep Learning

AI has many tools

- Machine learning (linear regression, random forest, etc.)
- Deep learning (neural networks)
- Others: graphical models, reinforcement learning



Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed."

- Arthur Samuel (1959)

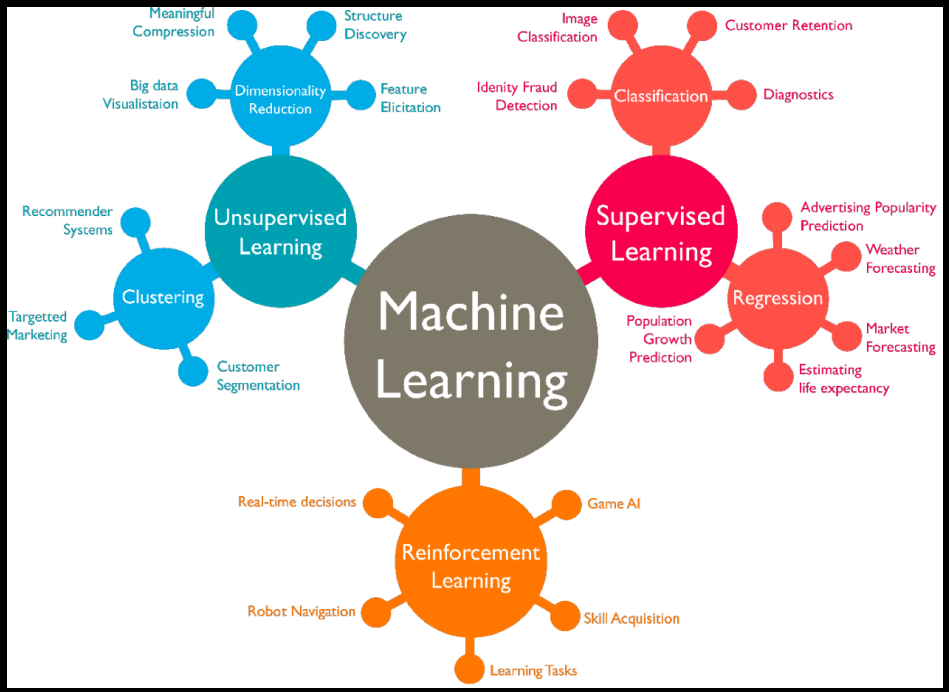


Image Source: https://www.slideshare.net/awahid/big-data-and-machine-learning-for-businesses

Deep Learning vs. traditional ML

- Deep learning is mainly used for supervised learning (classification, regression).
- Deep learning does not require feature engineering.

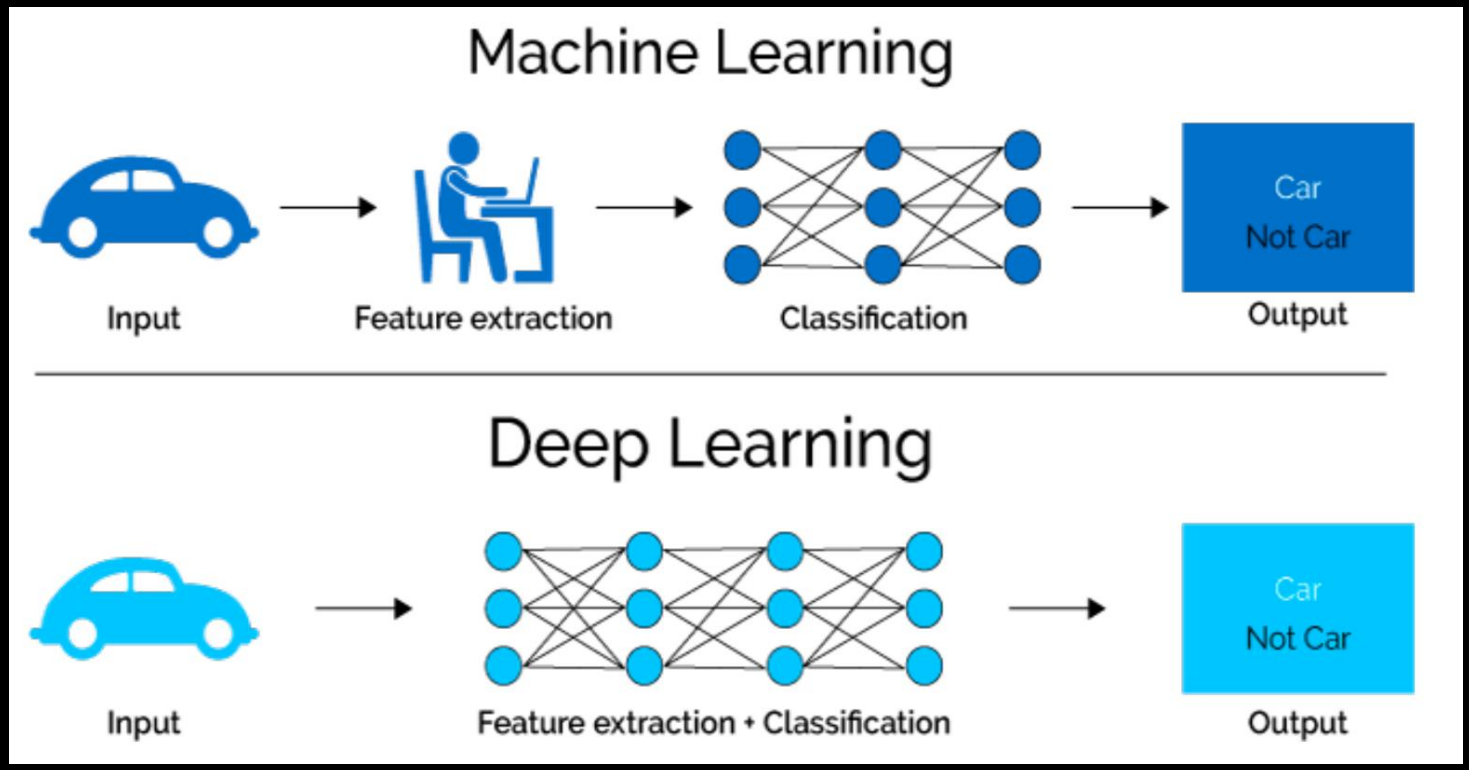
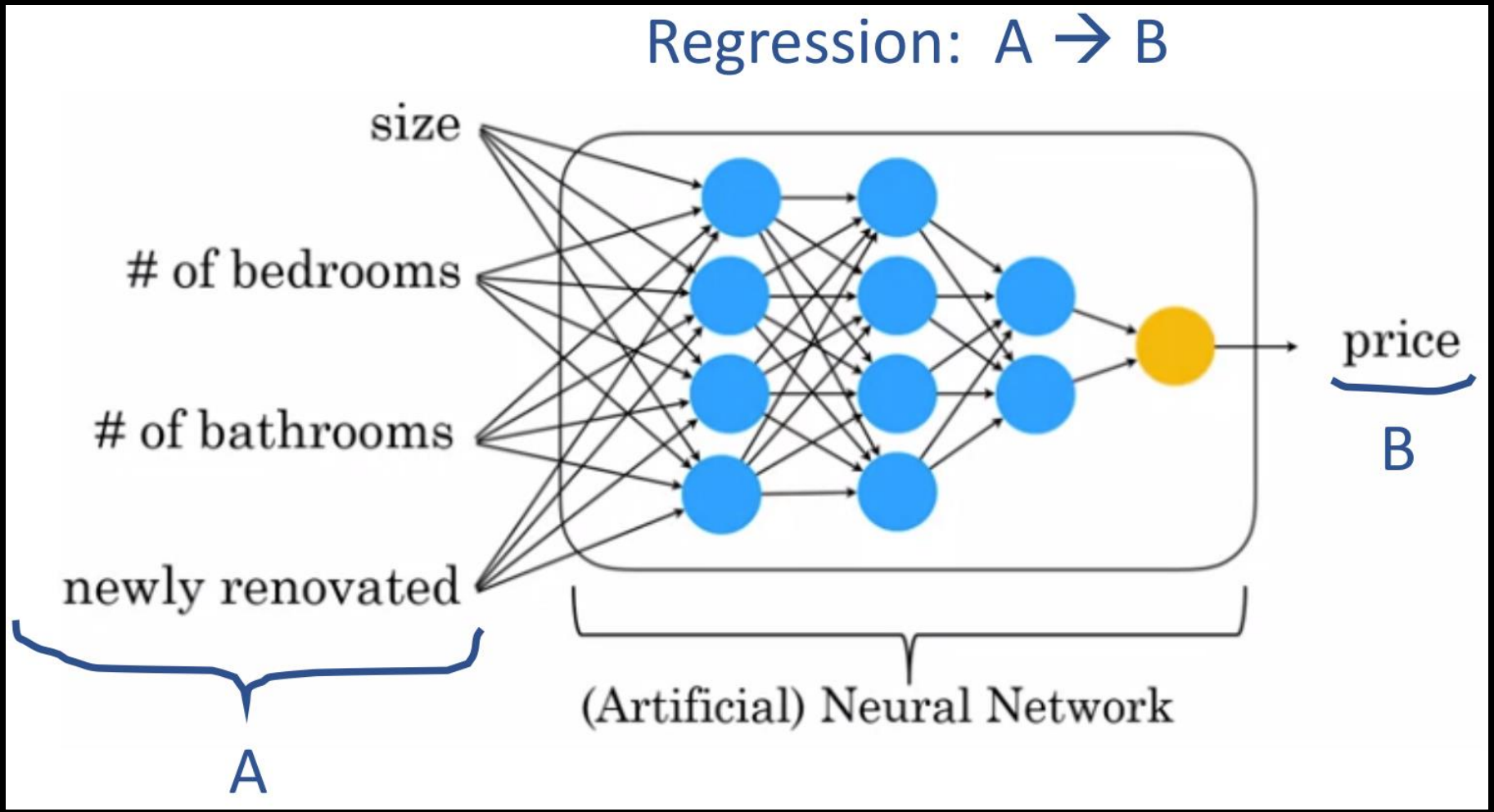
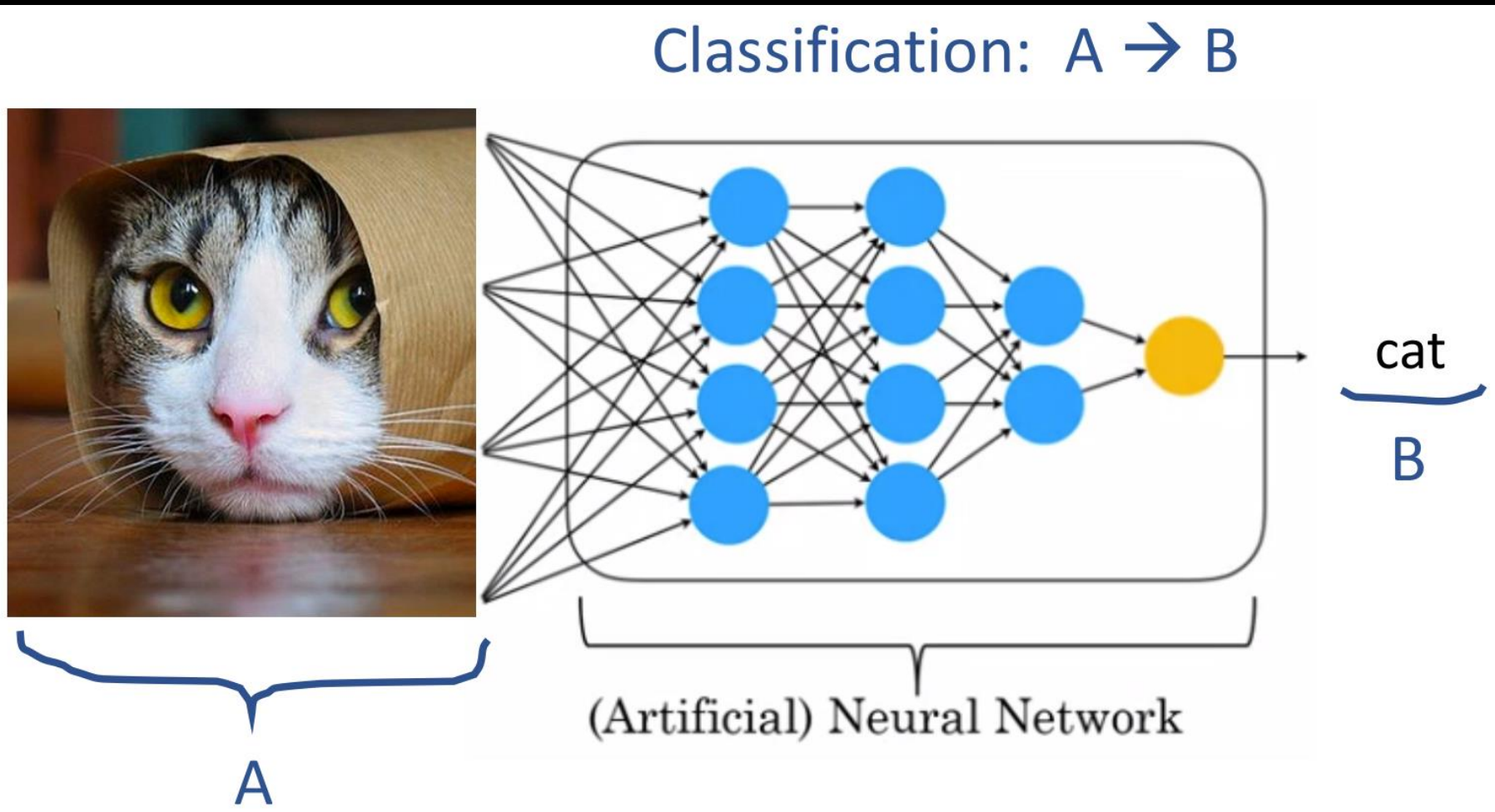


Image Source: http://aimagnifi.com/blog/index.php/2017/10/13/what-is-the-difference-between-machine-learning-and-deep-learning/

Deep learning example (1)



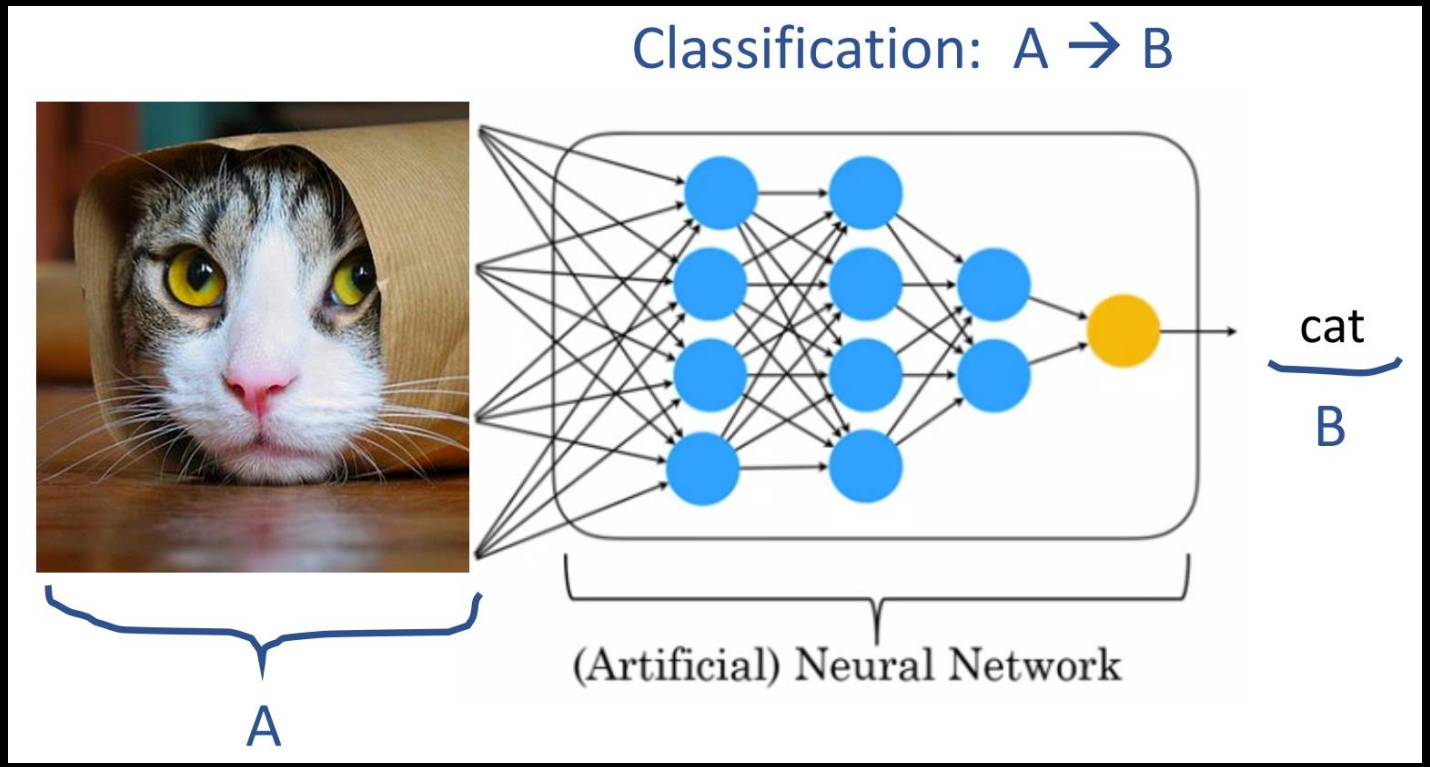
Deep learning example (2)



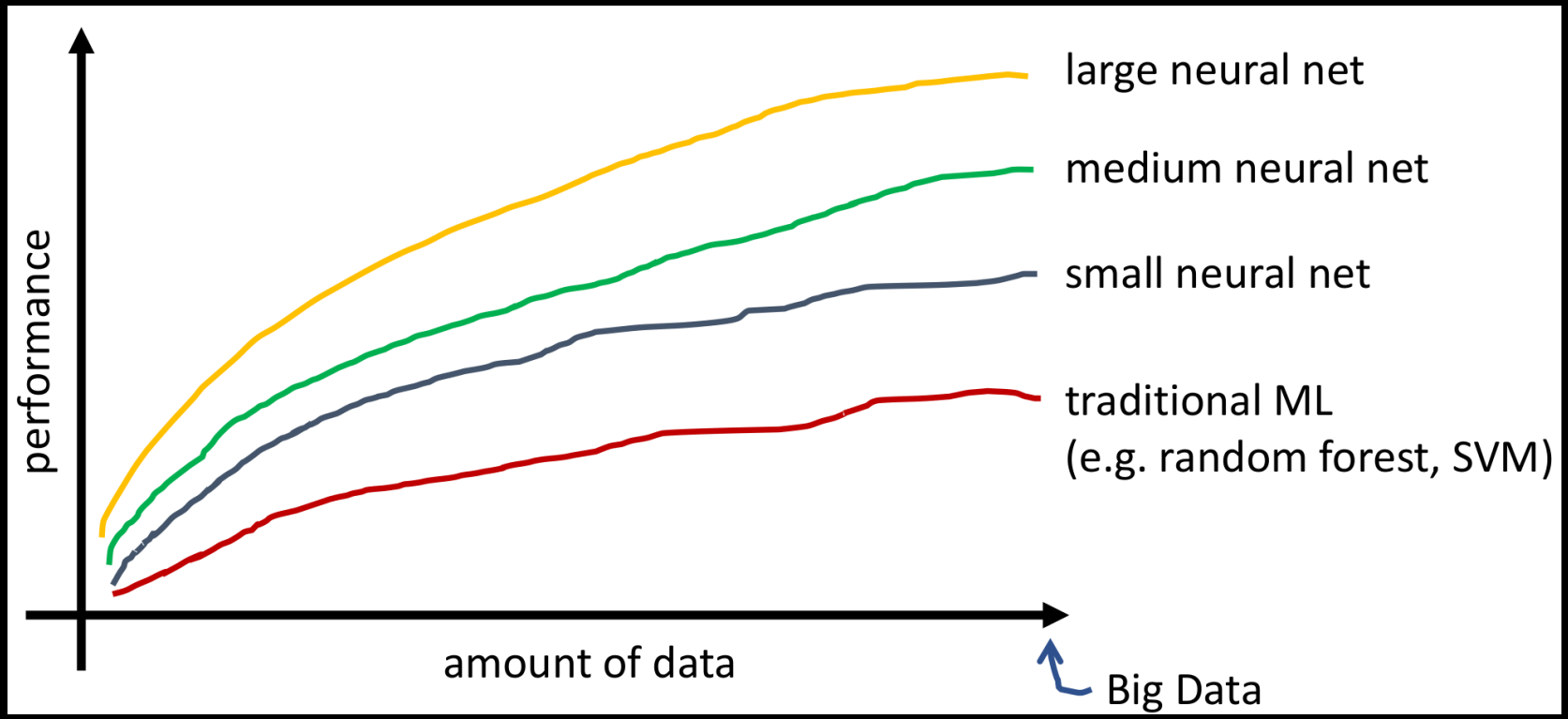
Deep learning example (2)

Q: what is the relationship between neural network and biological brain?

Neural networks were originally inspired by the brain, but the details of how they work are almost completely unrelated to how biological brains work.

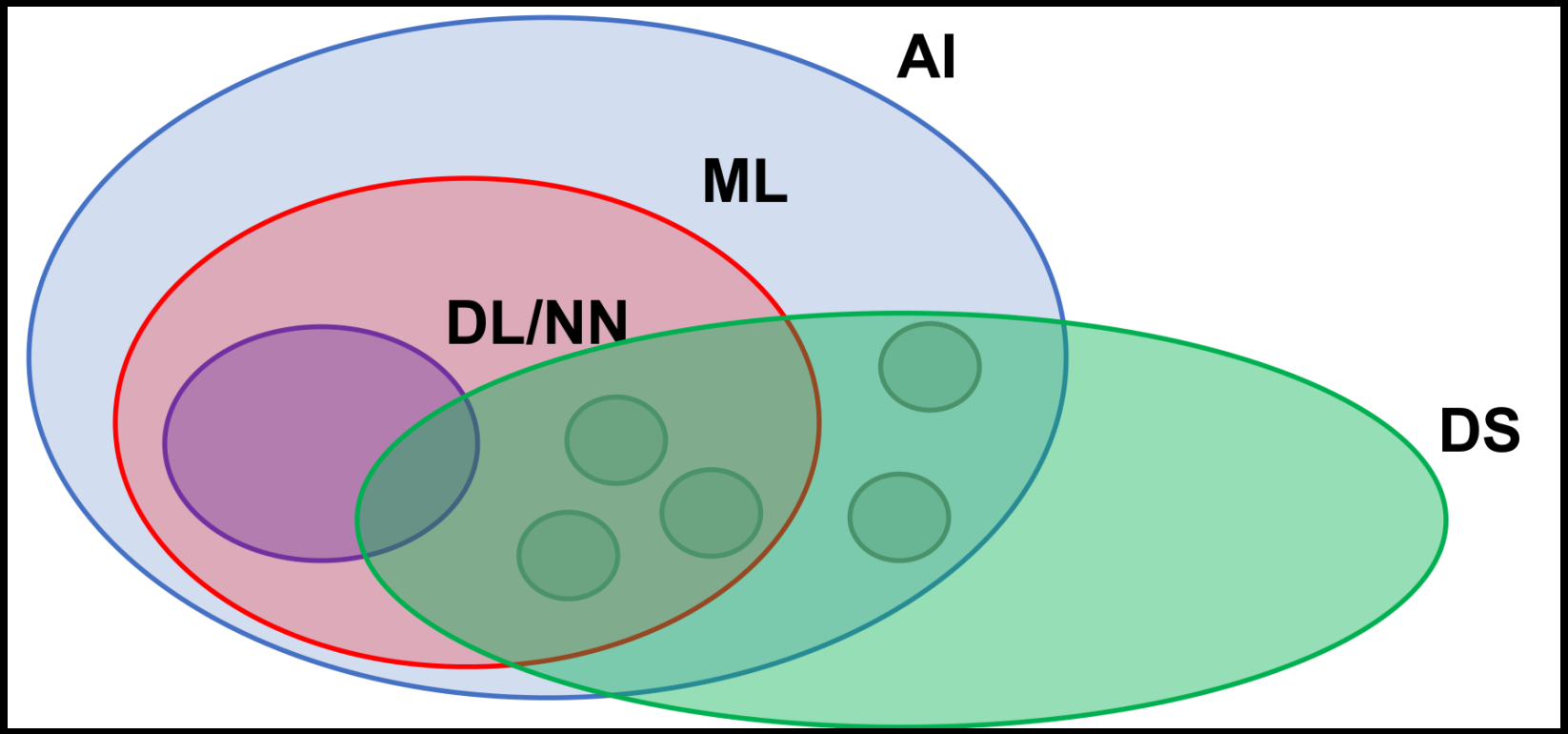


Why deep learning?



Data Science

‘Science of extracting knowledge and insights from data’

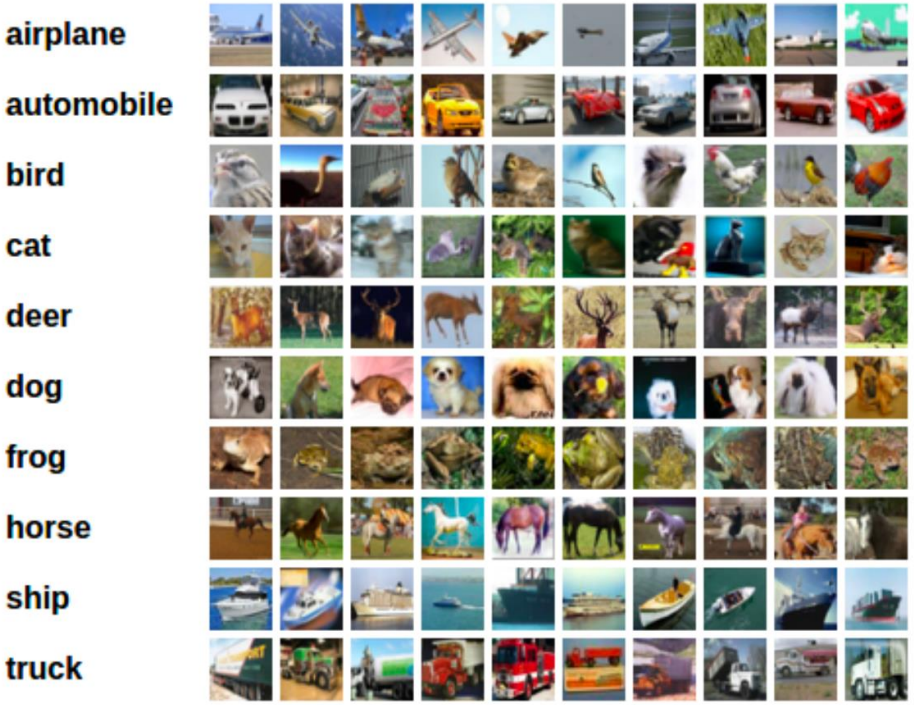


Deep Learning for Images and Videos

Image Classification

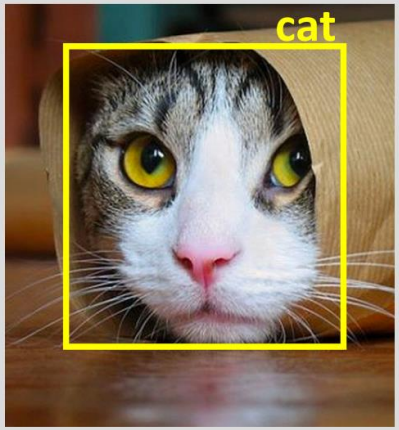
Input (A) → Output (B)

cat
dog
horse
car
airplane
...



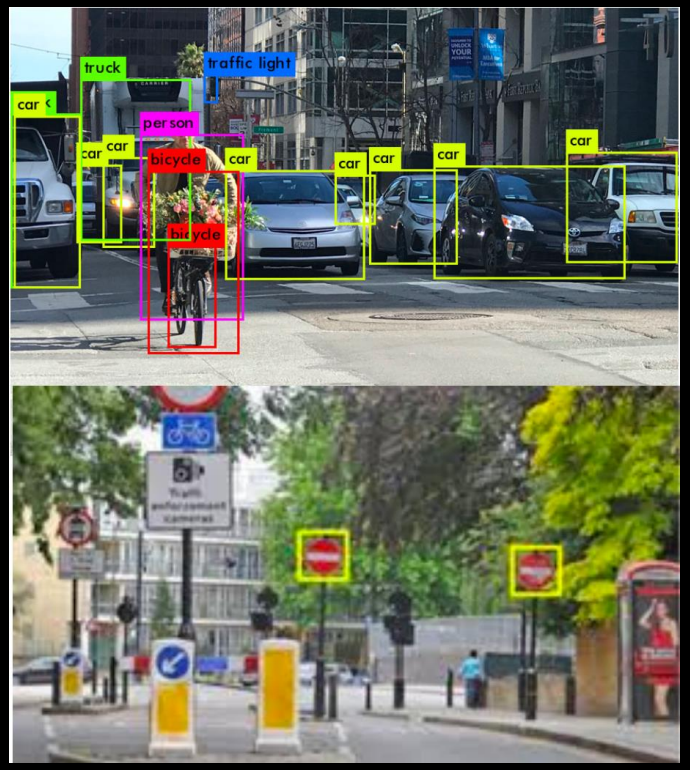
Object Detection

Input (A) → Output (B)



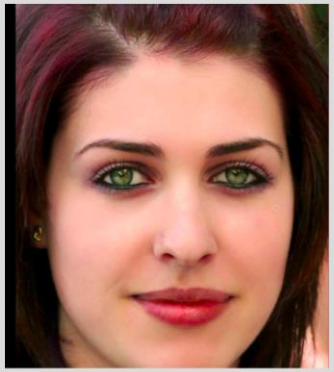
- An image with one or more objects

- Bounding box(es)
- Class of each bounding box (e.g. cat, dog,…)



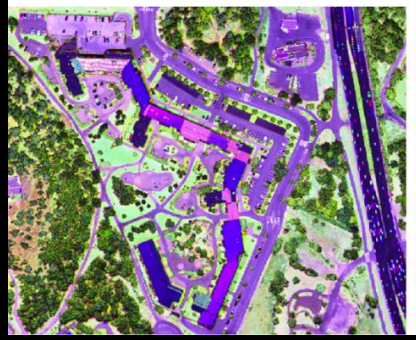
Semantic Segmentation

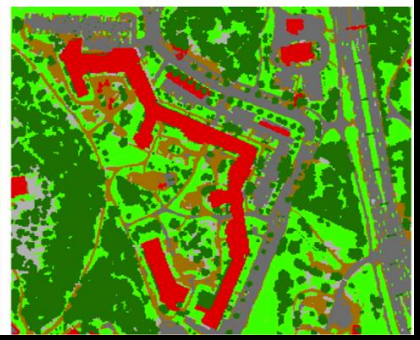
Input (A) ⟶ Output (B)



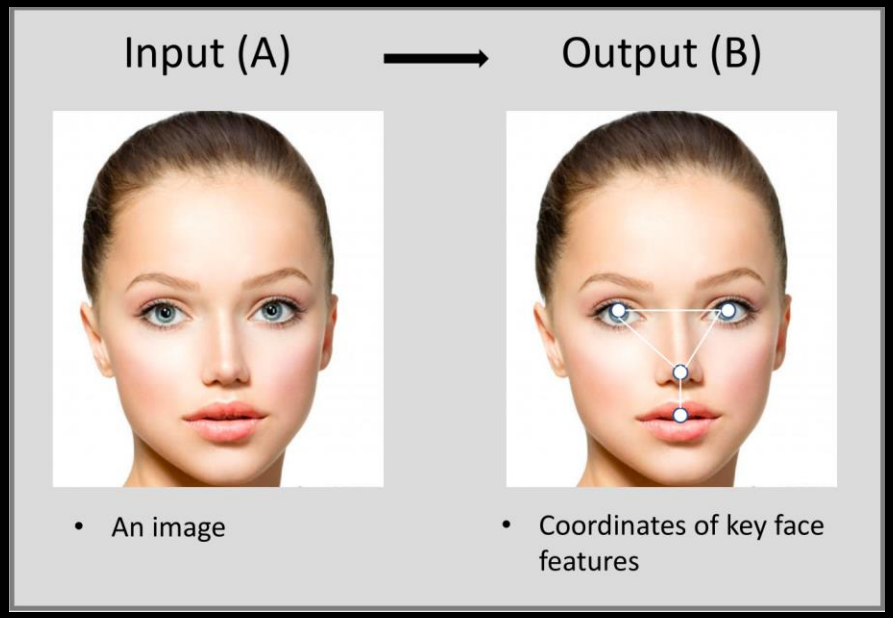
- An image

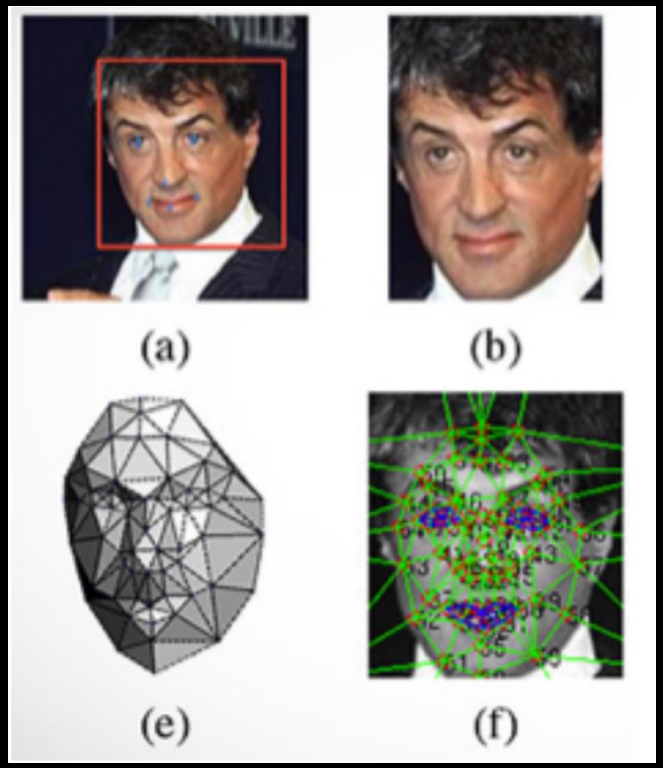
- The same image with class label for each pixel





Face Detection



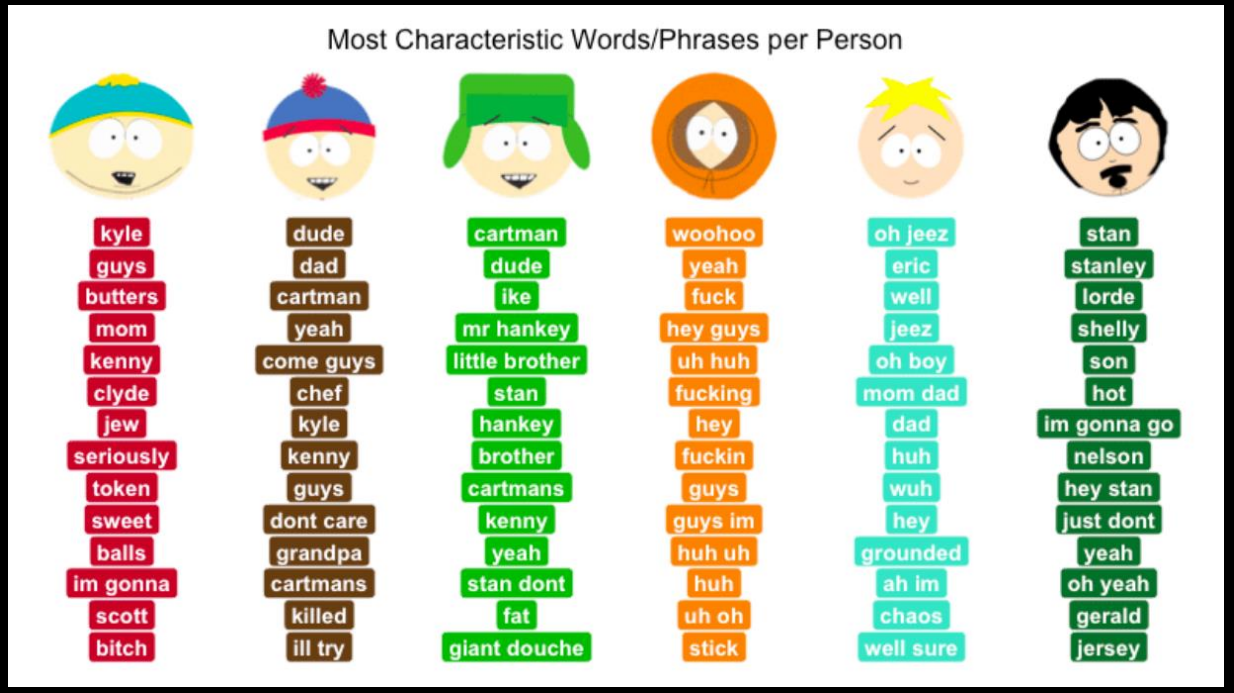


Deep Learning for Text

Keyword Detection

- Keyword extraction identifies the most important tokens or n-grams within a piece of text

- Input: a piece of text
- Output: list of most important tokens (words), optionally grouped by topics



Sentiment Analysis

- Sentiment Analysis assigns a sentiment score to each word in a piece of text

I love data science, and our teachers are awesome	+4 (Strongly positive)
Beer is disgusting, why do people even like it?	-1 (Weakly negative)
It's so great that my train is late every single day	+1 (Weakly positive)

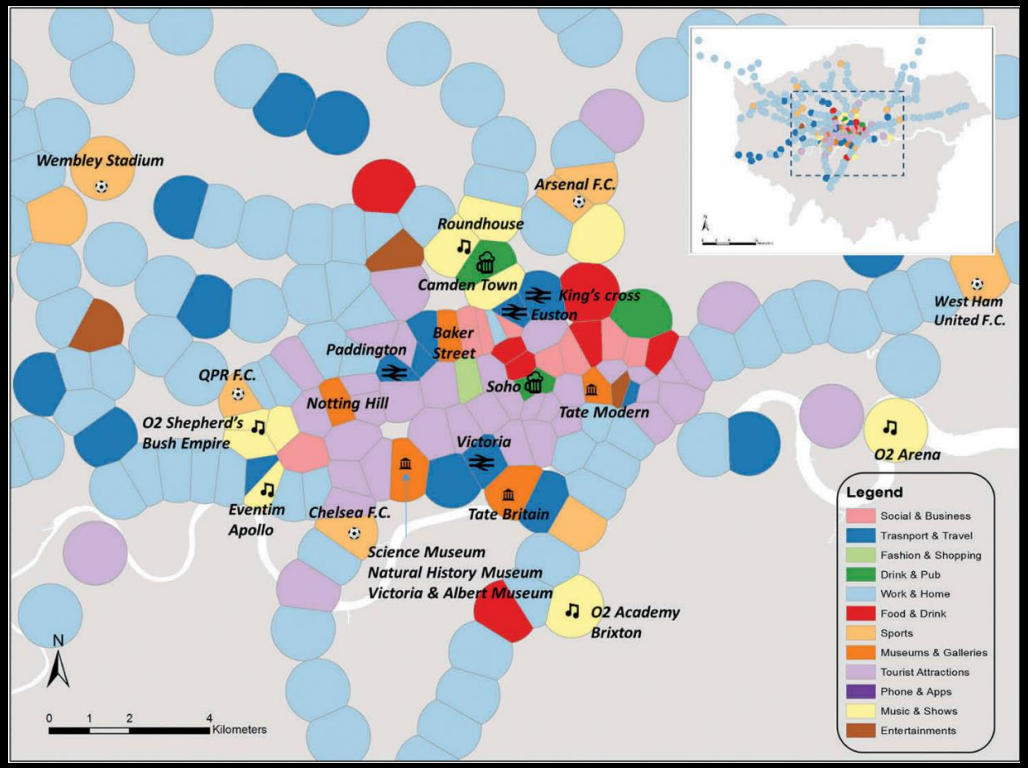
- **Input**: a piece of text
- **Output**: a piece of text with a sentiment score assigned to each word

Topic Extraction

- Topic modelling extracts topics from a collection of documents.
- **Input**: a collection of documents
- **Output**: a list of topics with words that belong to them

Latent Dirichlet Allocation: documents as mixtures of topics with different probabilities

Example:
Dominant topics on Tweets around the stations



Lai, J., Cheng, T., and Lansley, G. (2017) Improved targeted outdoor advertising based on geotagged social media data. *Annals of GIS*, ,1–14. https://doi.org/10.1080/19475683.2017.1382571

What deep learning cannot do

Adversarial attacks and other ways to fool AI

Garbage in, garbage out

If labels are not reliable, the model trained would be problematic.

“A flock of birds flying in the air”

“A group of flowers in a field”

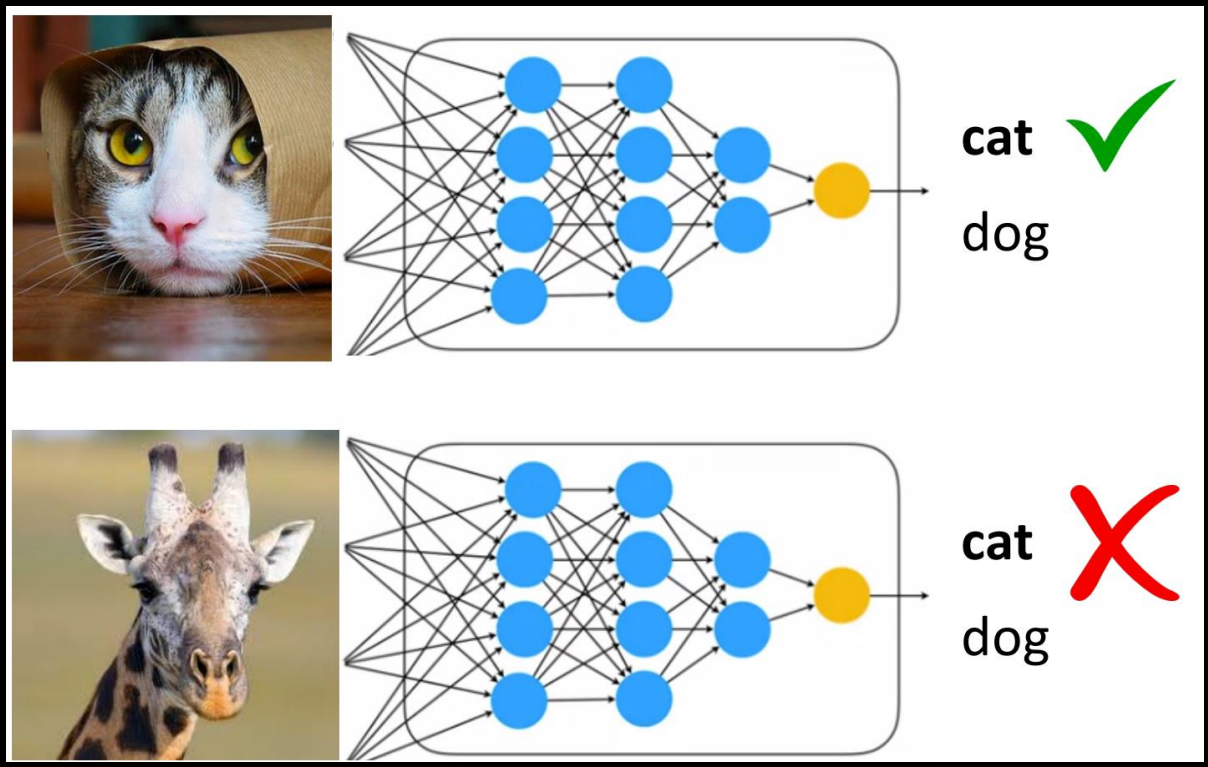
Garbage in, garbage out

A ‘racist’ soap dispenser at Facebook HQ. Where is the ‘racism’ from?

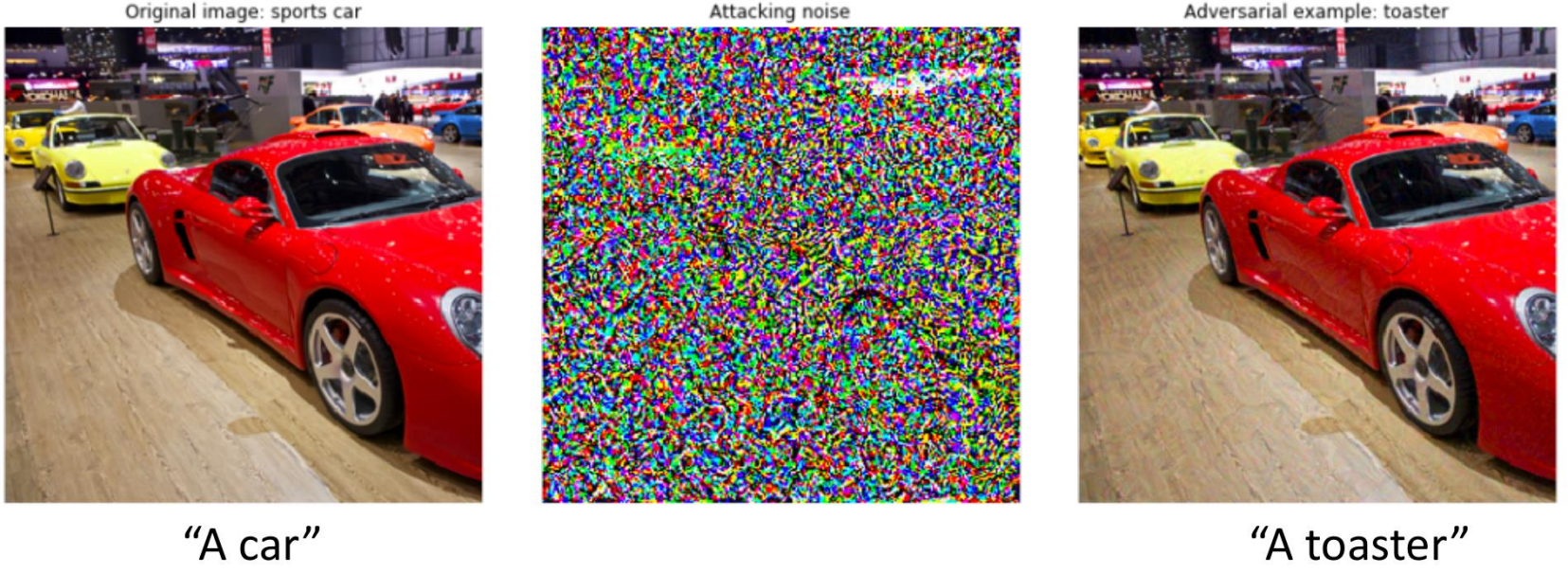
Garbage in, garbage out

Deep Learning classifier can only classify images into classes present in the training data.

It is important to make sure that the training data have a similar distribution as the ‘unseen’ data.



Adversarial Attacks



- You can fool a neural network classifier by adding noise to data.
- The noise might not be even visible to a human eye.

Adversarial Attacks

- Stop sign is no longer recognised by a neural network after a few stickers are placed on it.

Adversarial Attacks

- Microsoft's chatbot is turned into an anti-Semitic extremist after chatting with a group of people feeding it with racist information.

Deep Learning Workflow

Example: speech recognition

Amazon *Echo / Alexa* | Google *Home* | Apple *Siri* | Baidu *DuerOS*

- Source: Coursera “AI for Everyone” course by Andrew Ng (https://www.coursera.org/learn/ai-for-everyone/)

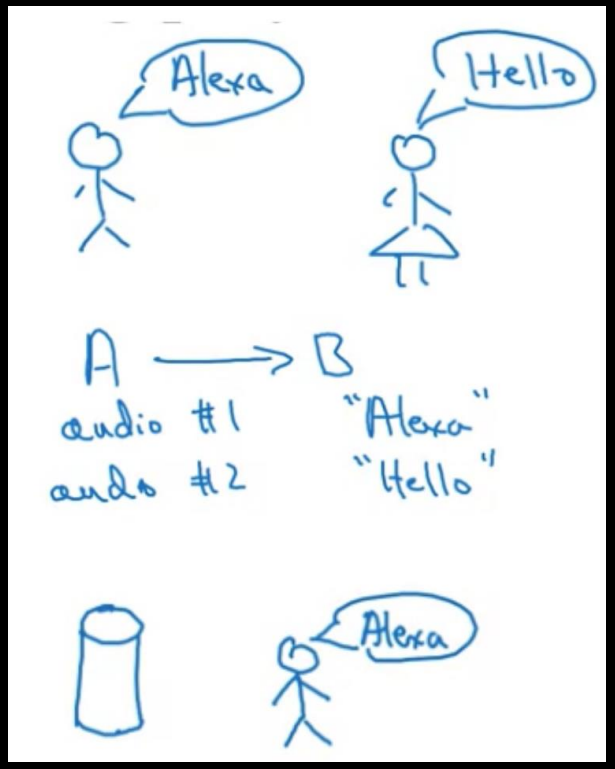
Key steps of a machine learning project

1. Collect data

2. Train model
   Iterate many times until good enough

3. Deploy model
   Get data back
   Maintain / update model

Example of Echo/Alexa



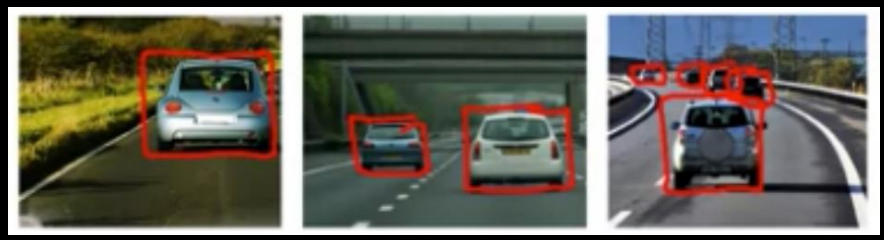
Key steps of a machine learning project

1. Collect data

2. Train model
   Iterate many times until good enough

3. Deploy model
   Get data back
   Maintain / update model

Image -> Position of other cars



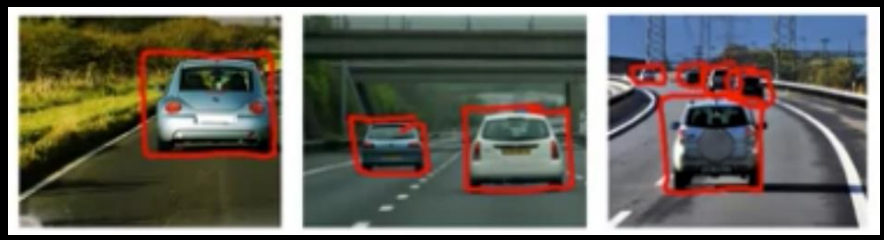
Key steps of a machine learning project

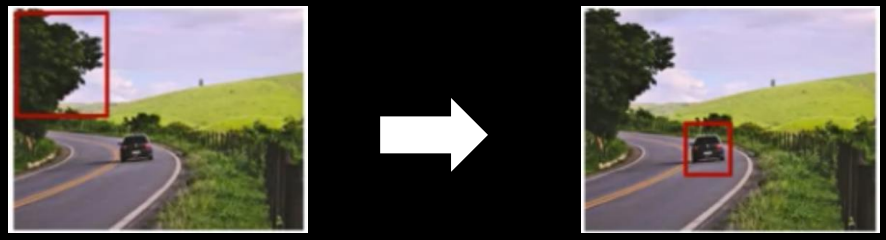
1. Collect data

2. Train model
   Iterate many times until good enough

3. Deploy model
   Get data back
   Maintain / update model

Image -> Position of other cars





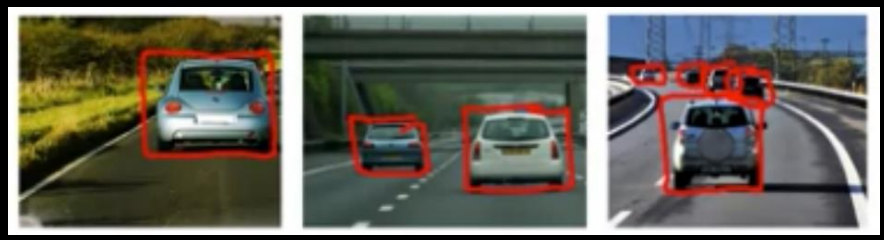
Key steps of a machine learning project

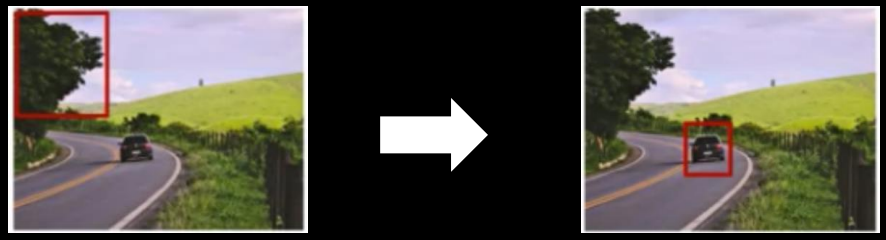
1. Collect data

2. Train model
   Iterate many times until good enough

3. Deploy model
   Get data back
   Maintain / update model

Image -> Position of other cars

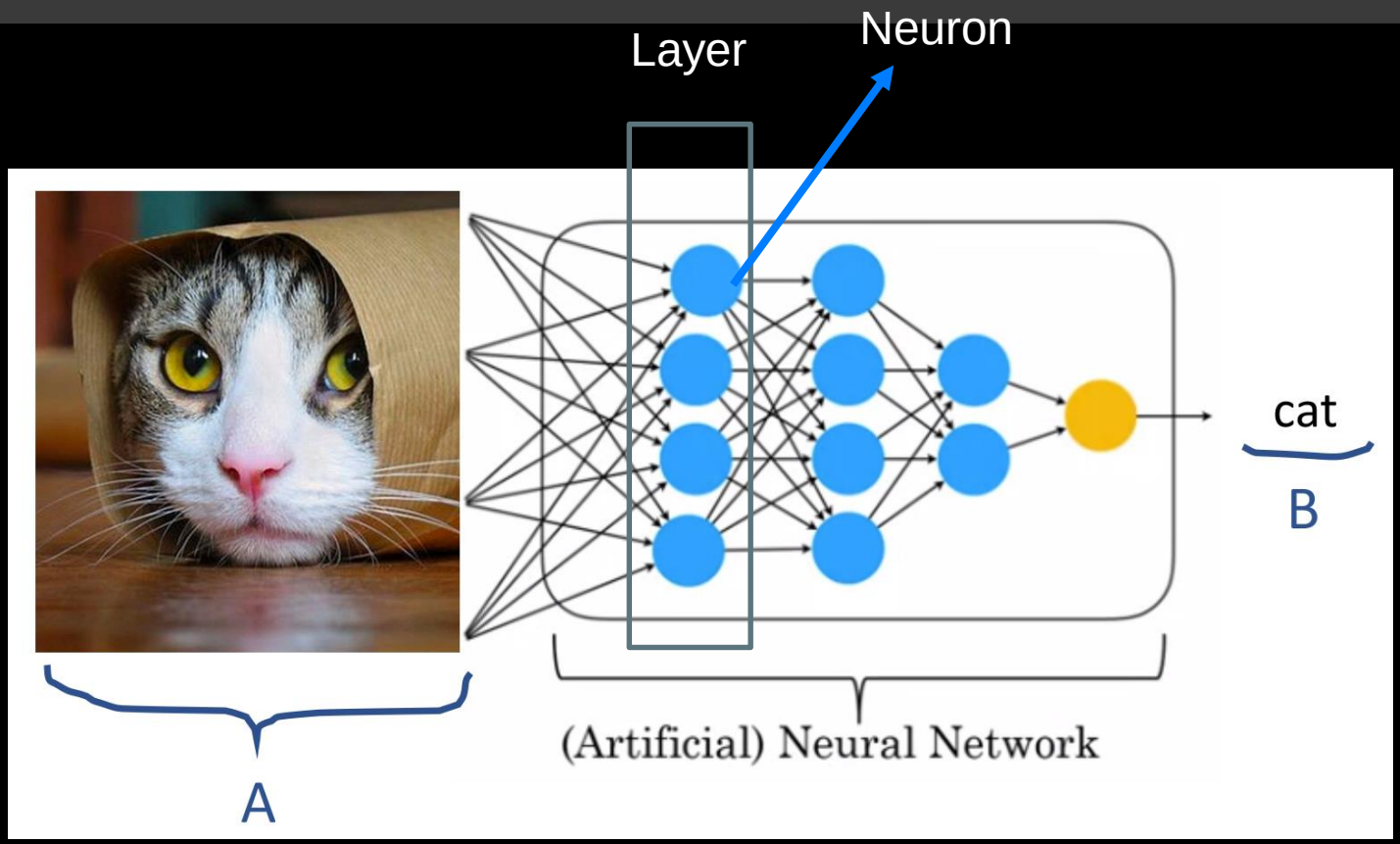


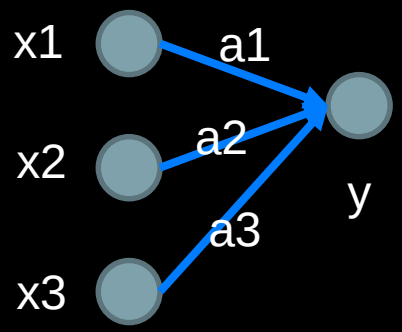


Quick explanation of deep learning

NN Classifier

Cat
Not cat





$$x' = a_1x_1 + a_2x_2 + a_3x_3 + a_0$$

$$y = \frac{1}{1 + e^{x'}}$$

What to learn: a1, a2, a3, a0

How to learn: using training data

Other issues

Neural network types, #layers, #neuros each layer

Any questions?

Huanfa Chen

huanfa.chen@ucl.ac.uk

Workshop

In this week's workshop you will learn how to train your own deep learning models in Python for a range of tasks:

- image classification
- face recognition
- semantic image segmentation
- classify text sentiment