Intro to Machine Learning

PHYS 6260

Slides borrowed from Stanford CS229, Spring 2020, modified slightly

Announcements:

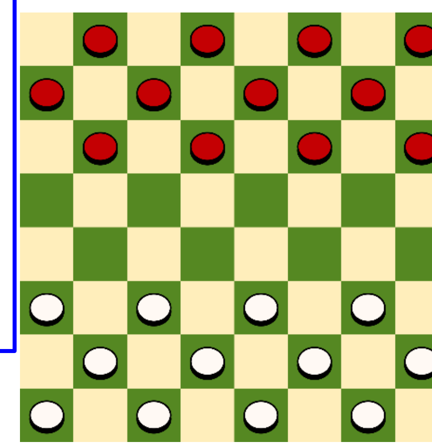
- Project Proposal: Due Friday 3/7

Definition of Machine Learning

Arthur Samuel (1959): Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

A. L. Samuel*

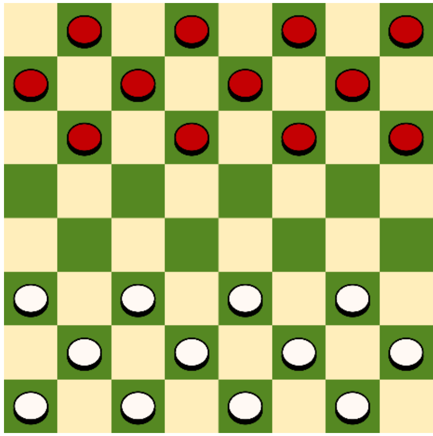
Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress



Photos from Wikipedia

Definition of Machine Learning

Tom Mitchell (1998): a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.



Experience (data): games played by the program (with itself)

Performance measure: winning rate

Image from Tom Mitchell's homepage

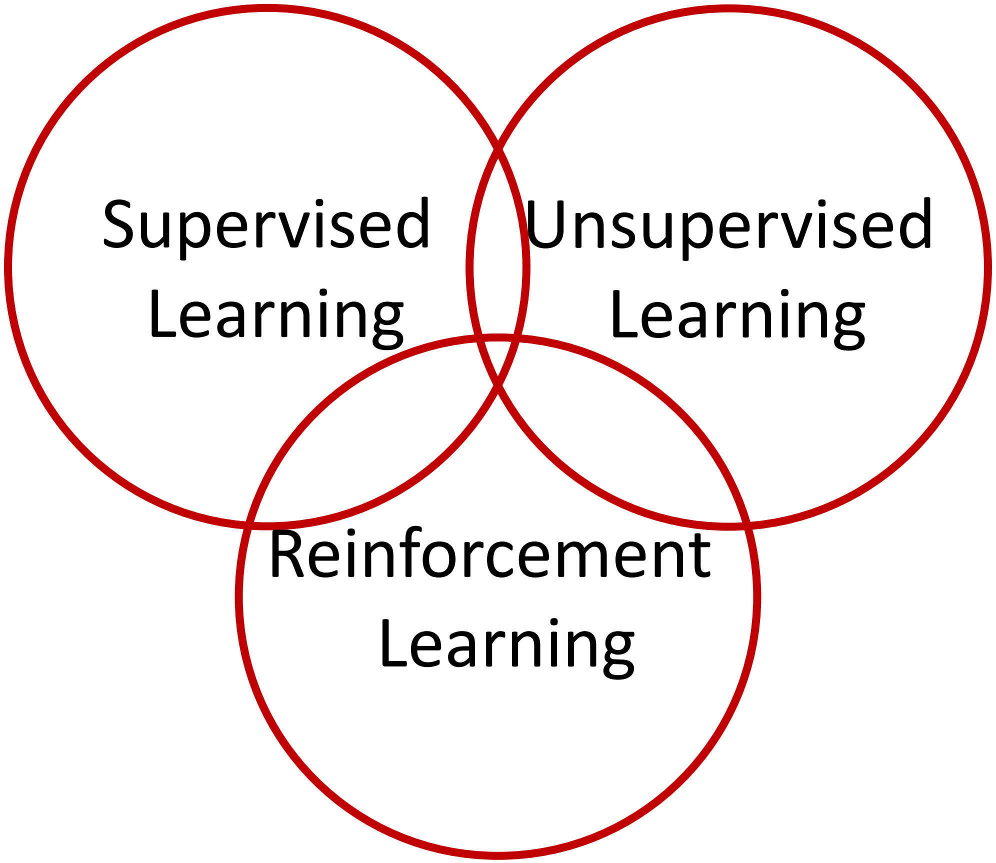
Taxonomy of Machine Learning (A Simplistic View Based on Tasks)

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Taxonomy of Machine Learning
(A Simplistic View Based on Tasks)



can also be viewed as tools/methods

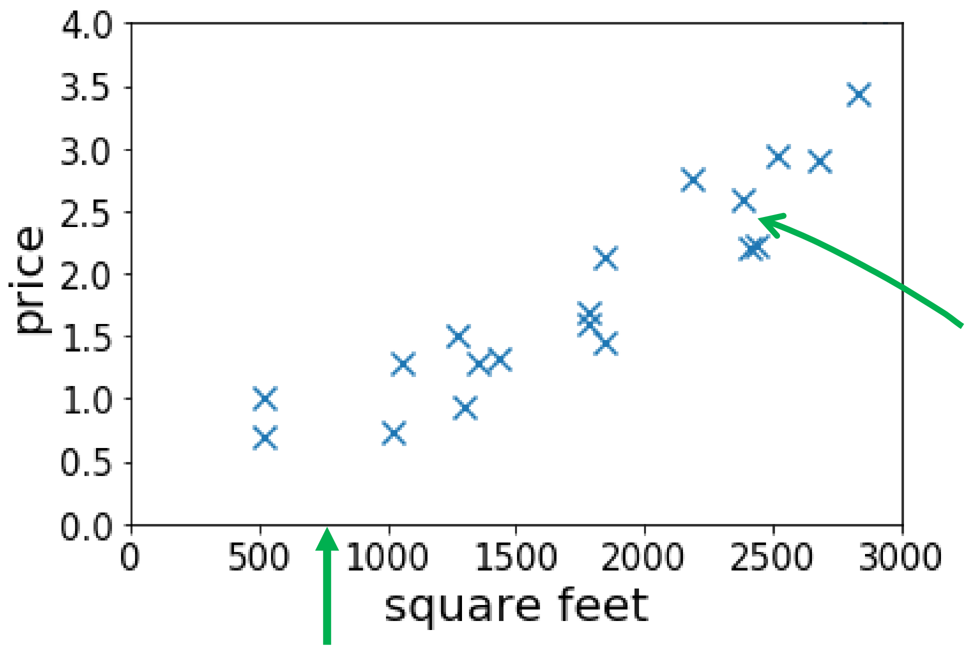
Supervised Learning

Housing Price Prediction

- Given: a dataset that contains n samples

$$\left(x^{(1)}, y^{(1)}\right), \ldots \left(x^{(n)}, y^{(n)}\right)$$

- Task: if a residence has x square feet, predict its price?



15th sample
$\left(x^{(15)}, y^{(15)}\right)$

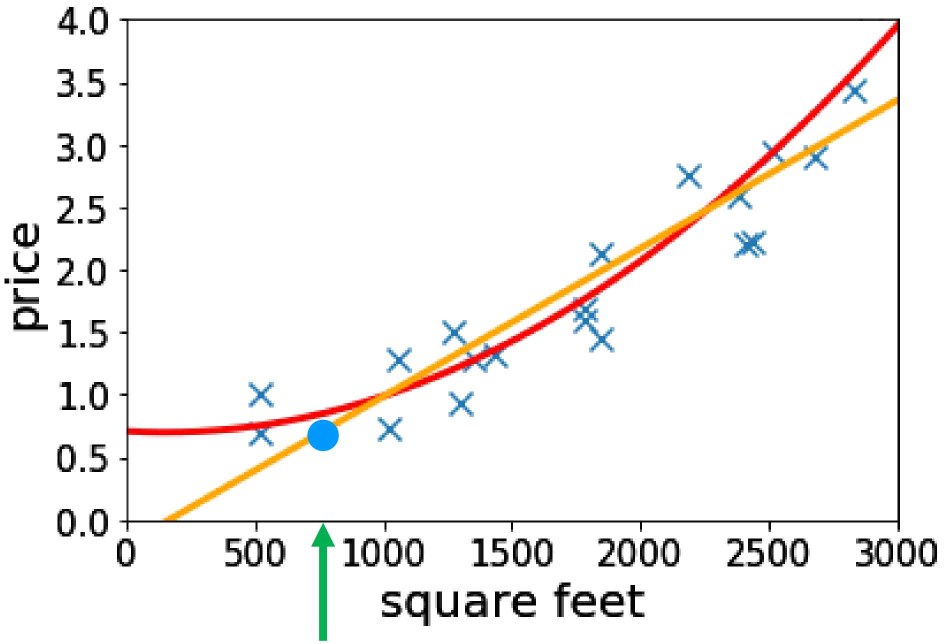
$x = 800$
$y = ?$

Housing Price Prediction

- Given: a dataset that contains n samples

$$\left(x^{(1)}, y^{(1)}\right), \dots \left(x^{(n)}, y^{(n)}\right)$$

- Task: if a residence has x square feet, predict its price?



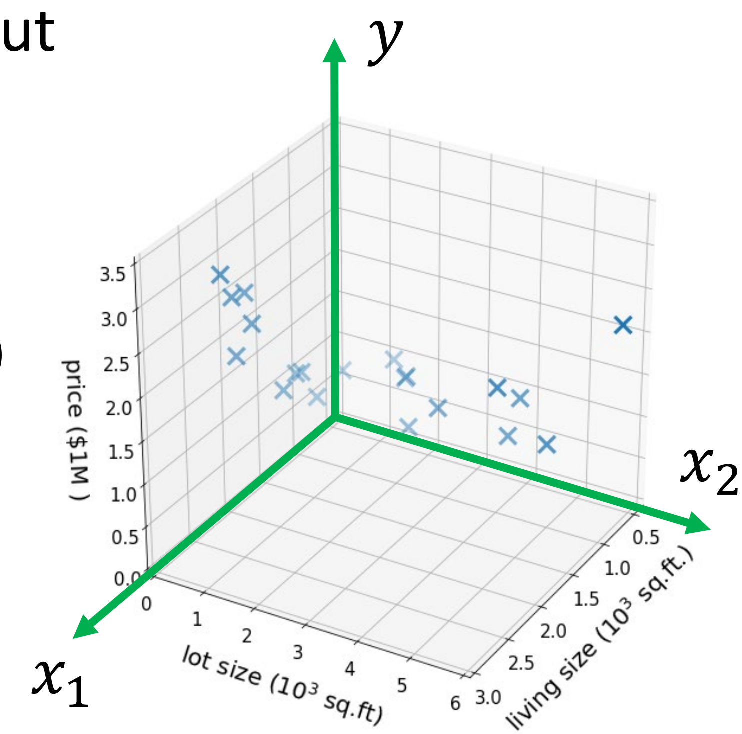
$x = 800$

$y = ?$

More Features

- Suppose we also know the lot size
- Task: find a function that maps

$$\underbrace{\text{(size, lot size)}}_{\substack{\text{features/input} \\ x \in \mathbb{R}^2}} \quad \rightarrow \quad \underbrace{\text{price}}_{\substack{\text{label/output} \\ y \in \mathbb{R}}}$$



- Dataset: $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$

where $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}\right)$

- "Supervision" refers to $y^{(1)}, \ldots, y^{(n)}$

High-dimensional Features

- $x \in \mathbb{R}^d$ for large d
- E.g.,

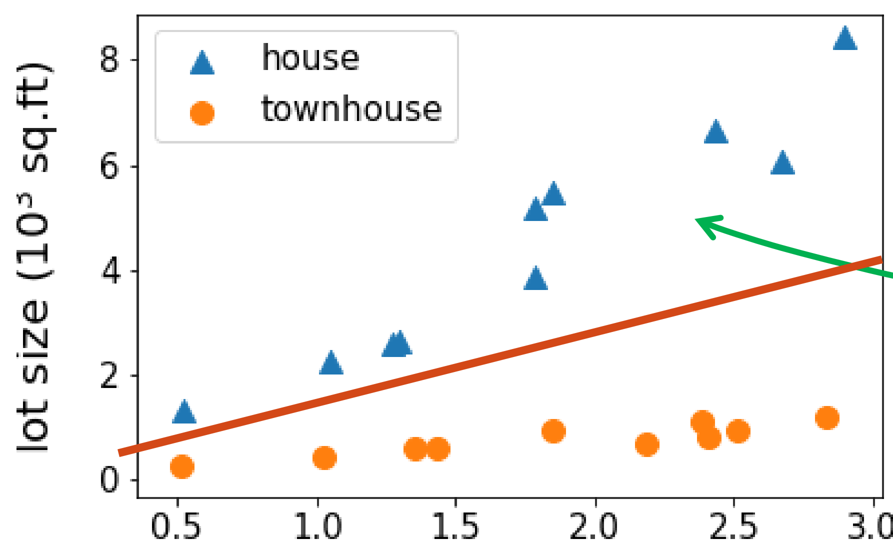
$$x = \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{array}\right] \begin{array}{l} \text{--- living size} \\ \text{--- lot size} \\ \text{--- \# floors} \\ \text{--- condition} \\ \text{--- zip code} \\ \qquad \vdots \\ \end{array} \longrightarrow \quad y \text{ --- price}$$

- Can include infinite dimensional features; select features based on the data

Regression vs Classification

- regression: if $y \in \mathbb{R}$ is a continuous variable
  - e.g., price prediction
- classification: the label is a discrete variable
  - e.g., the task of predicting the types of residence

(size, lot size) $\rightarrow$ house or townhouse?



$y =$ house or townhouse?

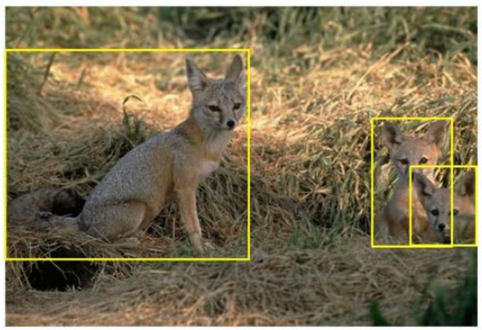
Supervised Learning in Computer Vision

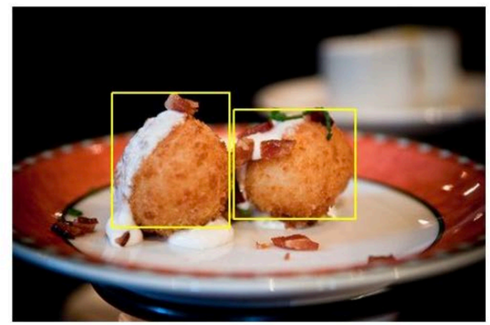
- Image Classification
  - $x =$ raw pixels of the image, $y =$ the main object

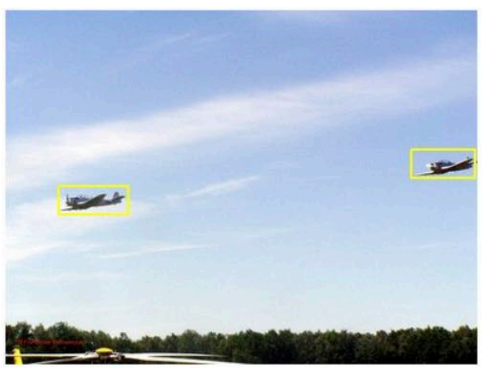
ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

Supervised Learning in Computer Vision

- Object localization and detection
  - x = raw pixels of the image, y = the bounding boxes


kit fox


croquette

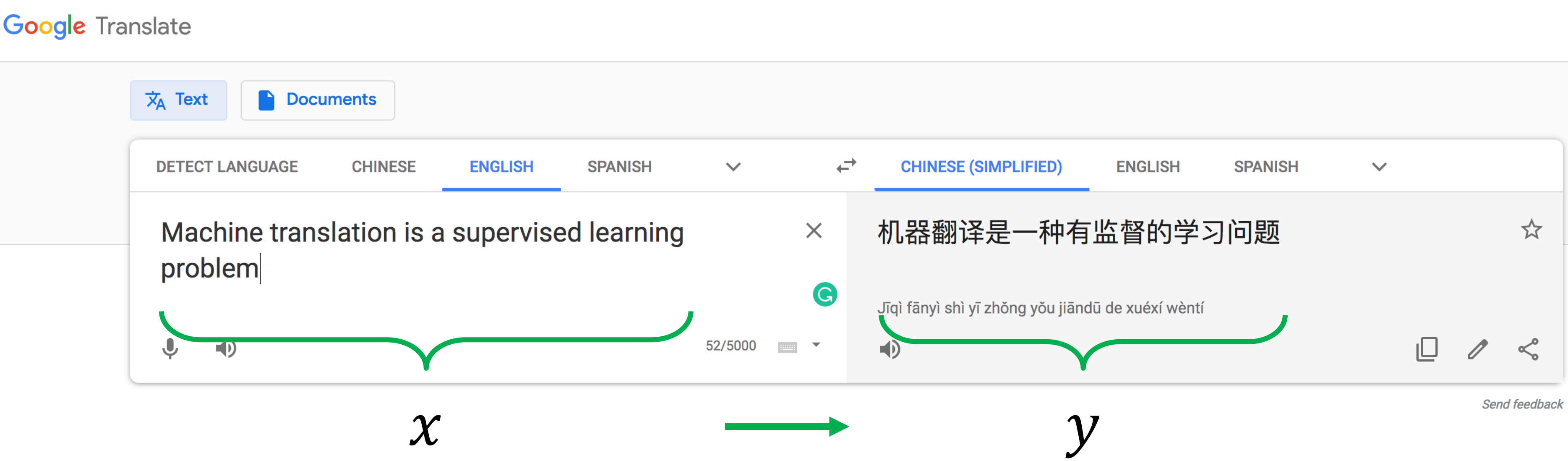

airplane


frog

ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

Supervised Learning in Natural Language Processing

- Machine translation

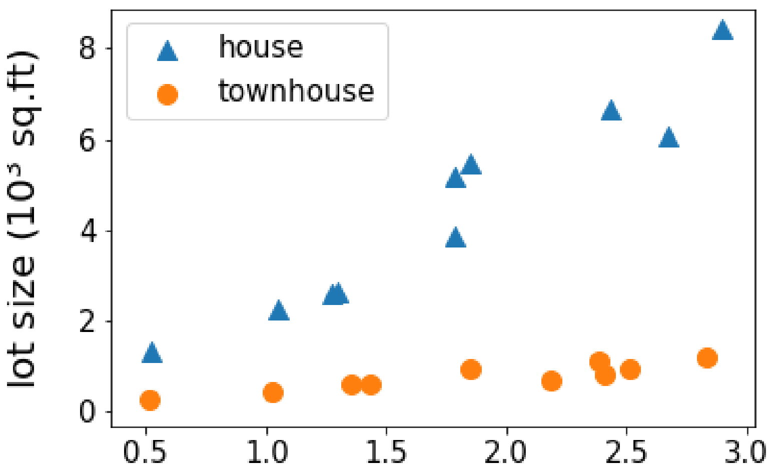


Unsupervised Learning

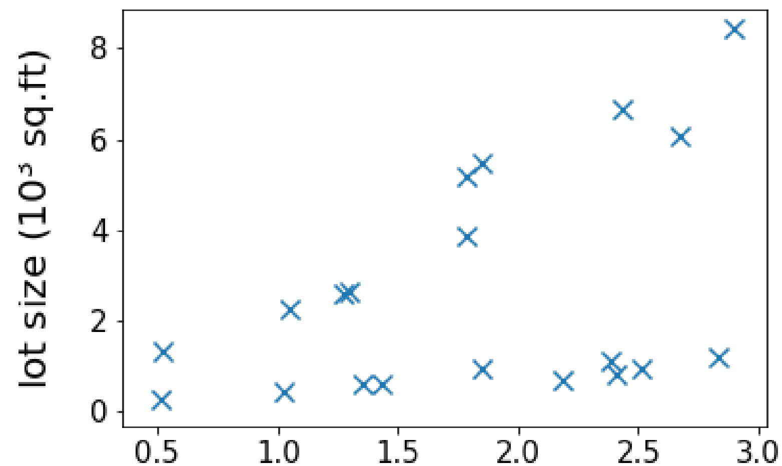
Unsupervised Learning

- Dataset contains no labels: $x^{(1)}, \ldots x^{(n)}$
- Goal (vaguely-posed): to find interesting structures in the data

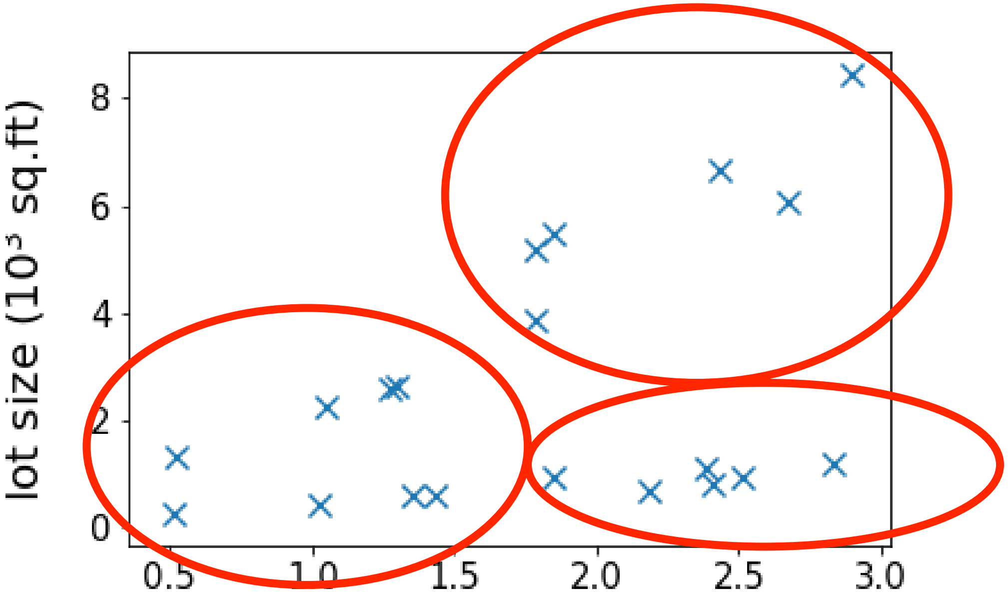
supervised



unsupervised

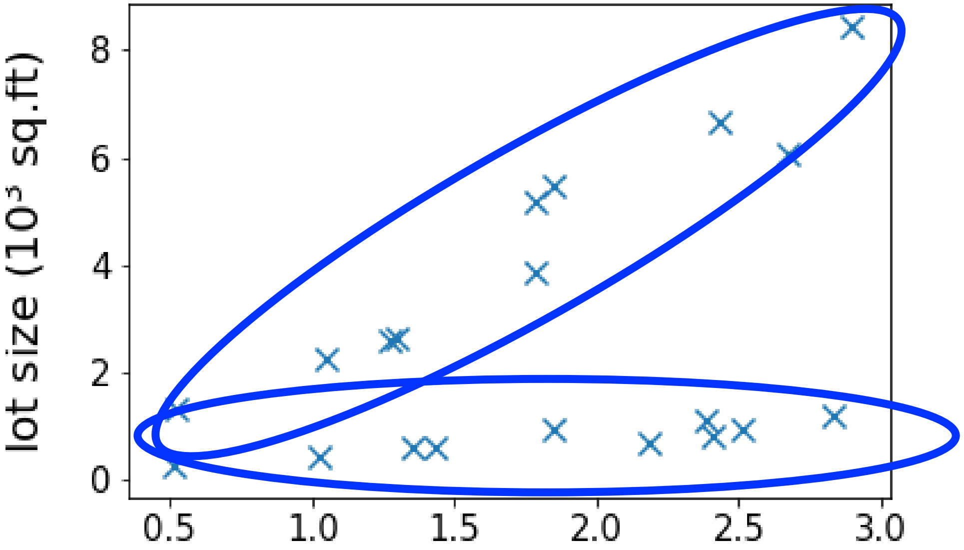


Clustering

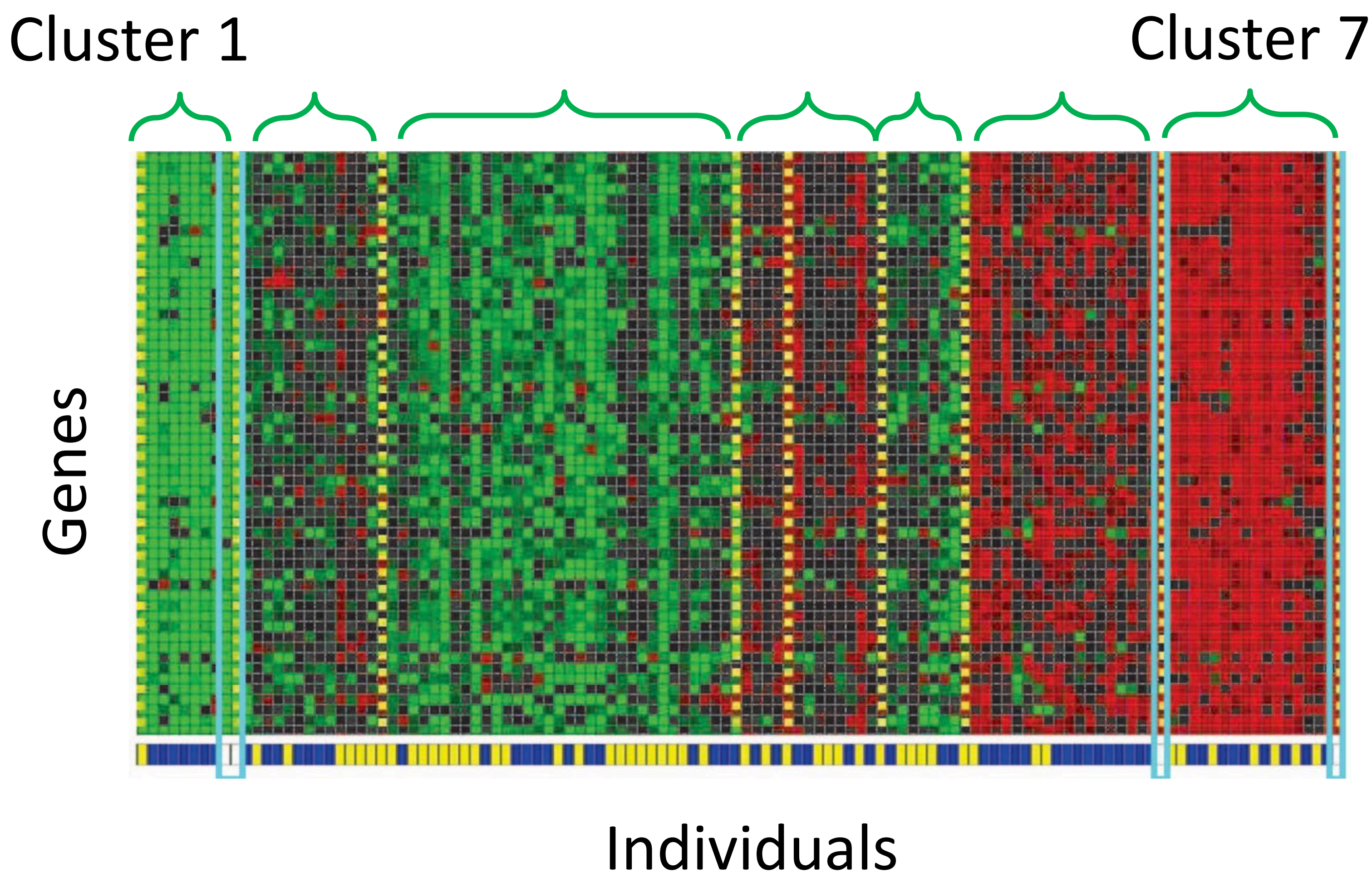


Clustering

- k-mean clustering, mixture of Gaussians

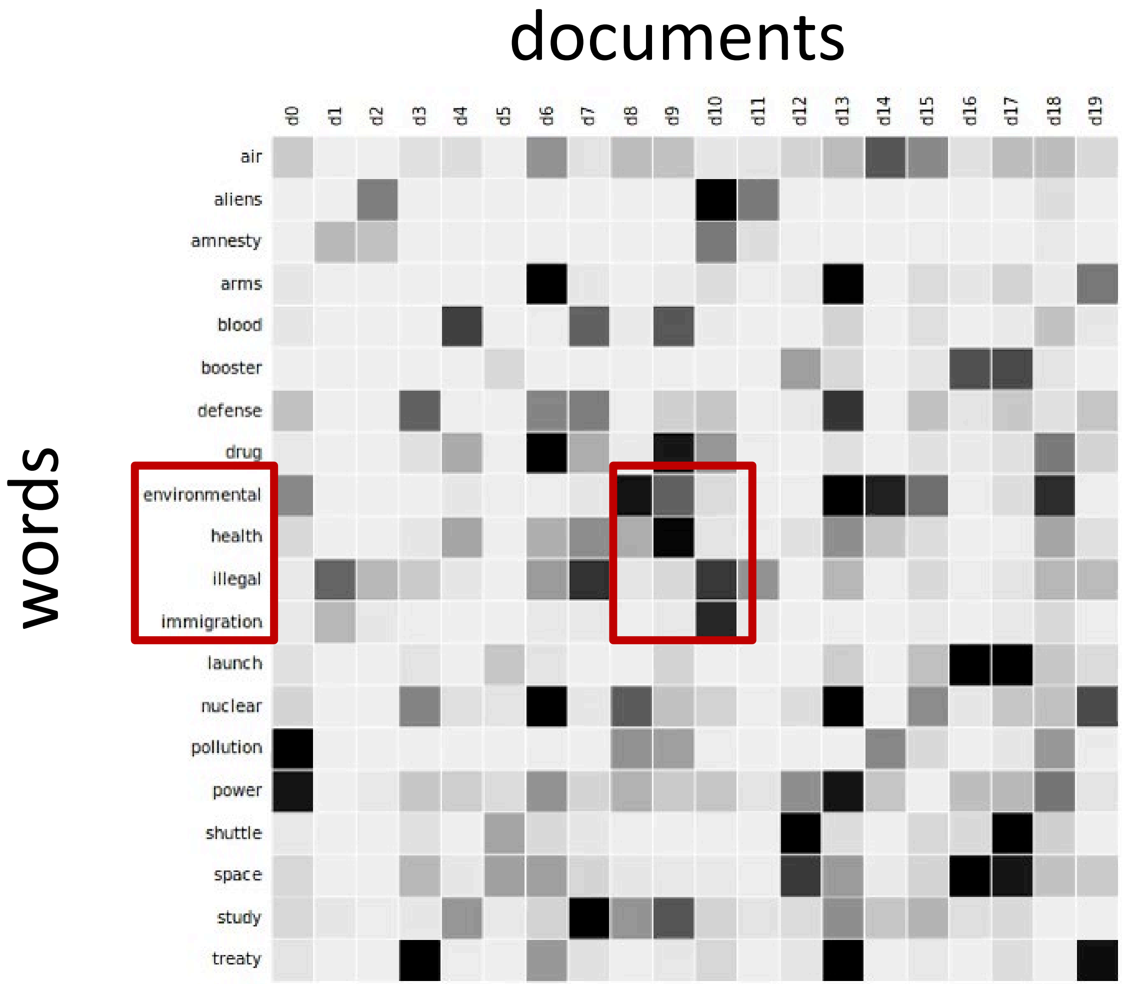


Clustering Genes



Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. [Su-In Lee, Dana Pe'er, Aimee M. Dudley, George M. Church and Daphne Koller. '06]

Latent Semantic Analysis (LSA)



- See: Principal component analysis (tools used in LSA)

Image credit: https://commons.wikimedia.org/wiki/File:Topic_detection_in_a_document-word_matrix.gif

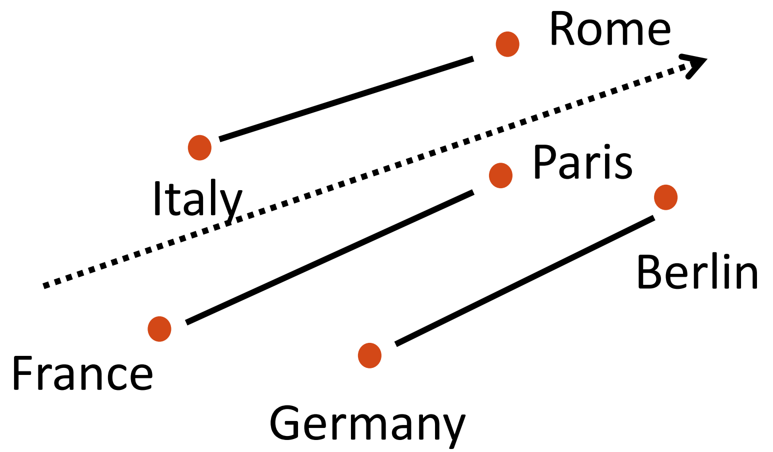
Word Embeddings

WIKIPEDIA
The Free ~~Encyclopedia~~
CORPUS

Unlabeled dataset

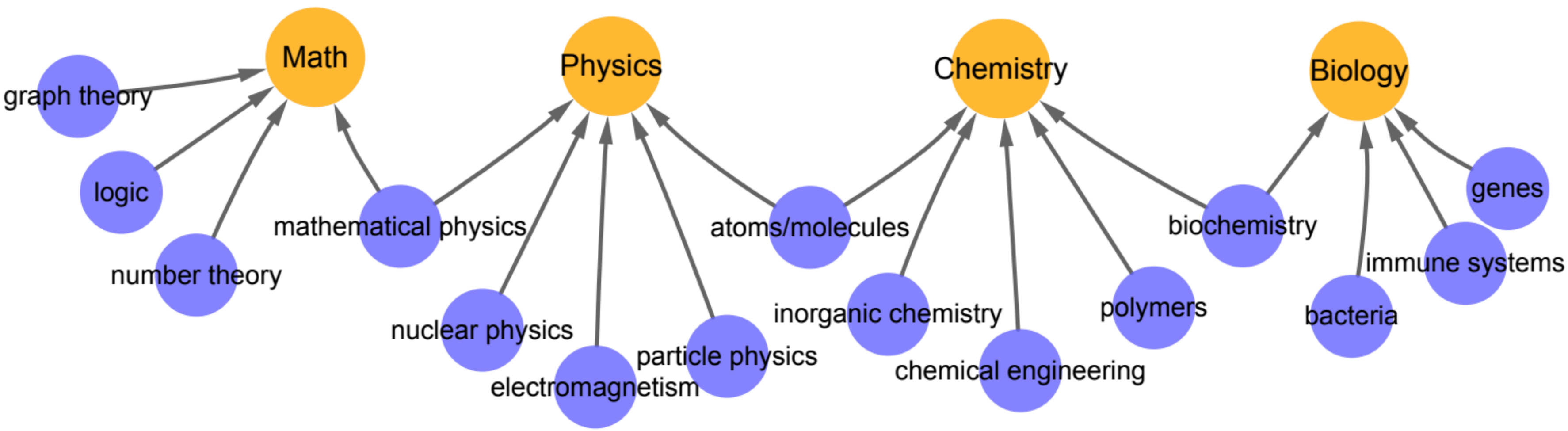
Represent words by vectors

- word $\xrightarrow{\text{encode}}$ vector
- relation $\xrightarrow{\text{encode}}$ direction



Word2vec [Mikolov et al'13]
GloVe [Pennington et al'14]

Clustering Words with Similar Meanings (Hierarchically)



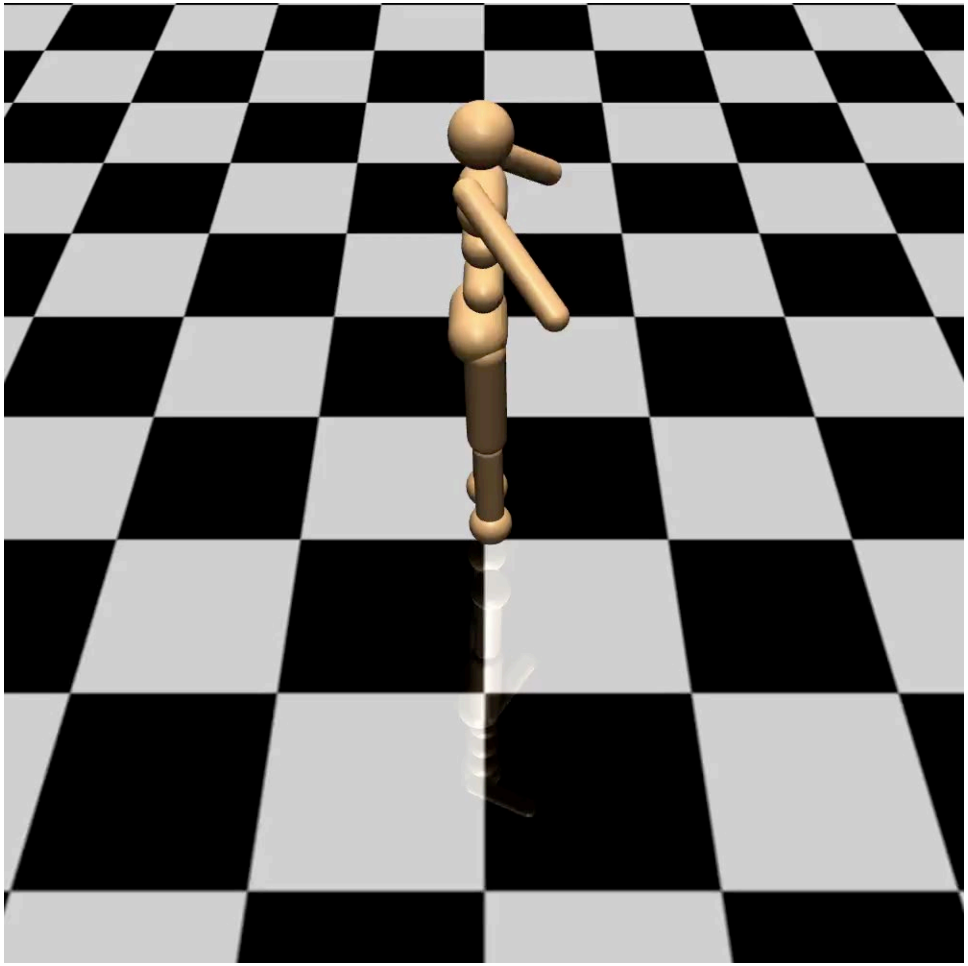
	logic deductive propositional semantics	graph subgraph bipartite vertex	boson massless particle higgs	polyester polypropylene resins epoxy	acids amino biosynthesis peptide
tag	*logic*	*graph theory*	*particle physics*	*polymer*	*biochemistry*

[Arora-Ge-Liang-M.-Risteski, TACL'17,18]

Reinforcement Learning

learning to walk to the right

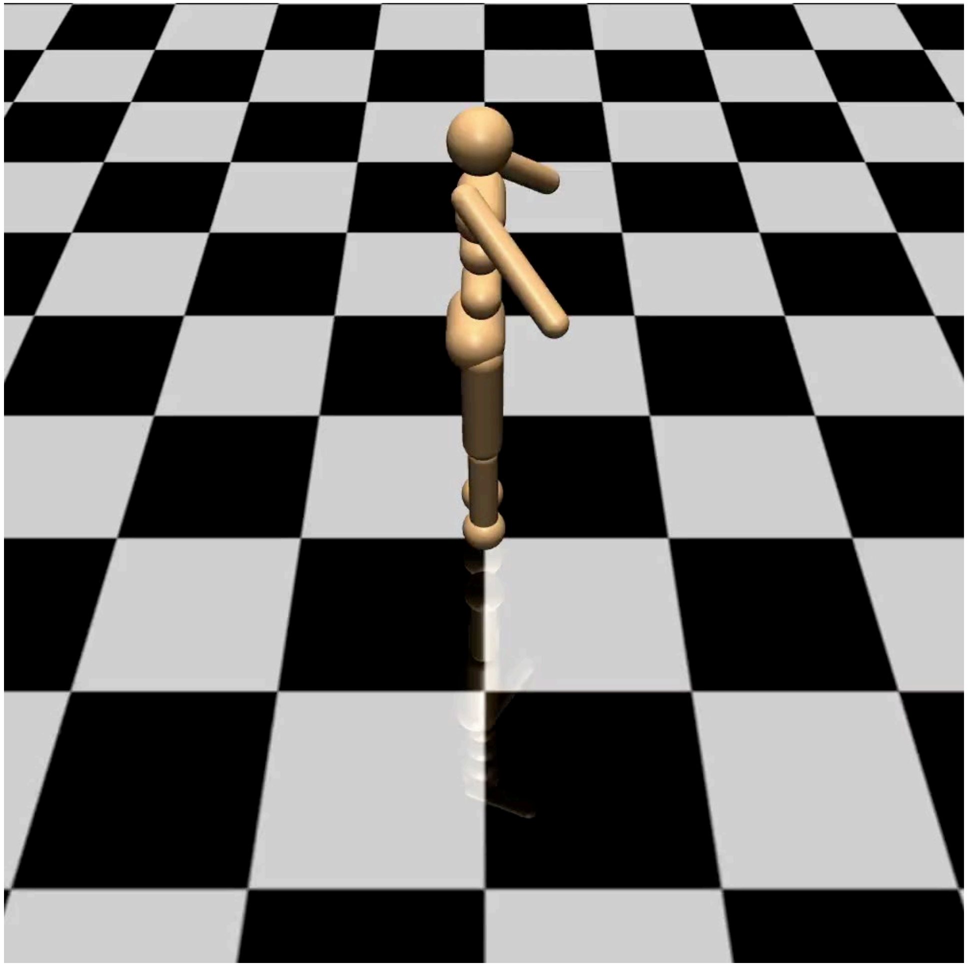
Iteration 10



[Luo-Xu-Li-Tian-Darrell-M.’18]

learning to walk to the right

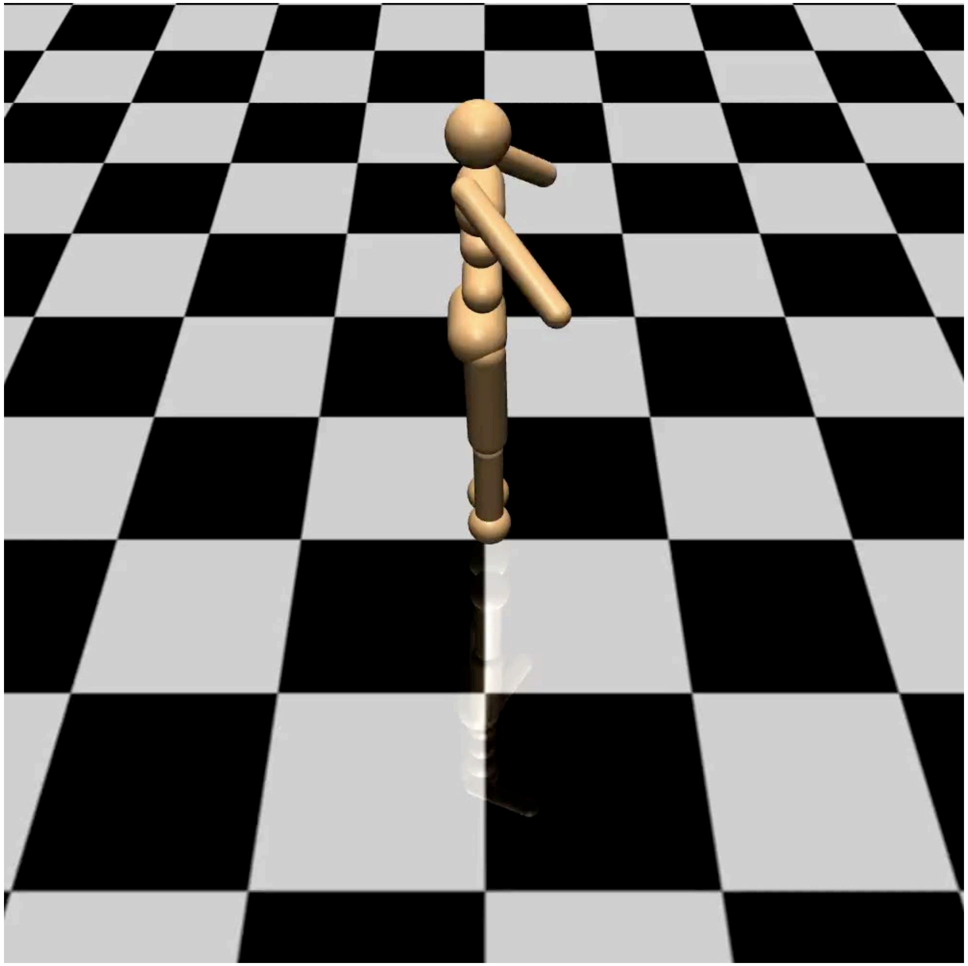
Iteration 20



[Luo-Xu-Li-Tian-Darrell-M.'18]

learning to walk to the right

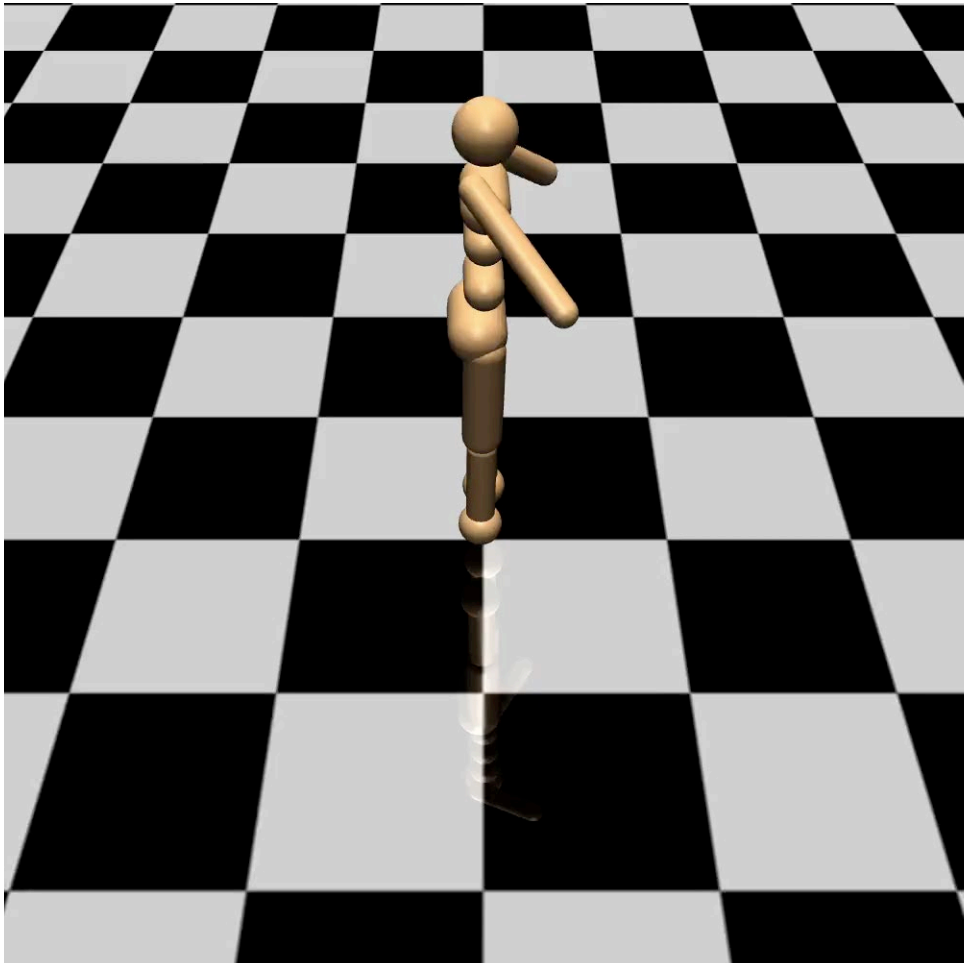
Iteration 80



[Luo-Xu-Li-Tian-Darrell-M.’18]

learning to walk to the right

Iteration 210

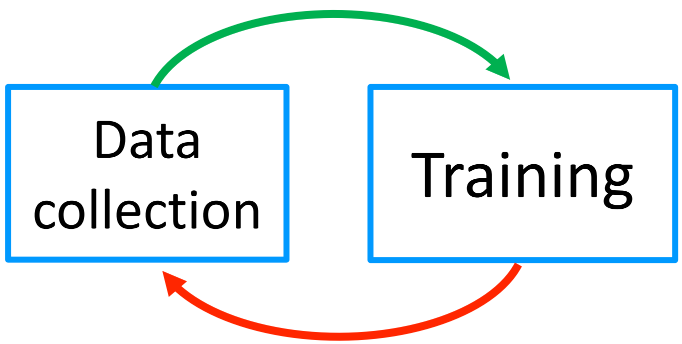


[Luo-Xu-Li-Tian-Darrell-M.'18]

Reinforcement Learning

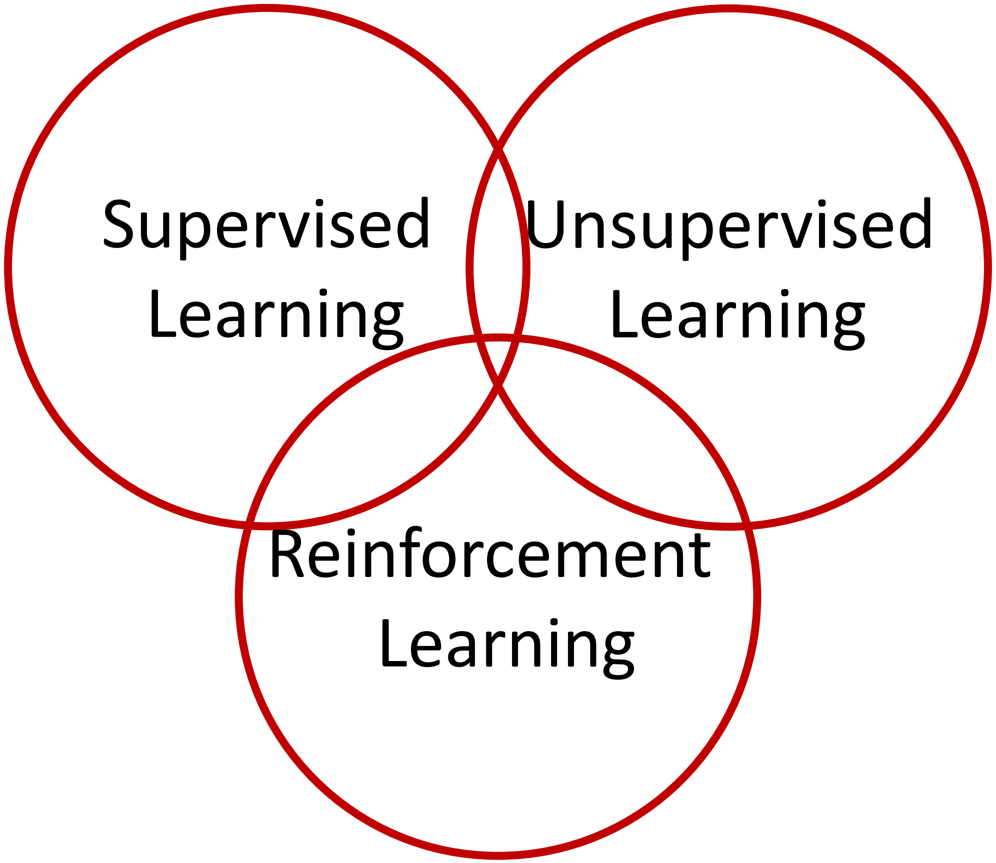
- The algorithm can collect data interactively

Try the strategy and collect feedbacks



Improve the strategy based on the feedbacks

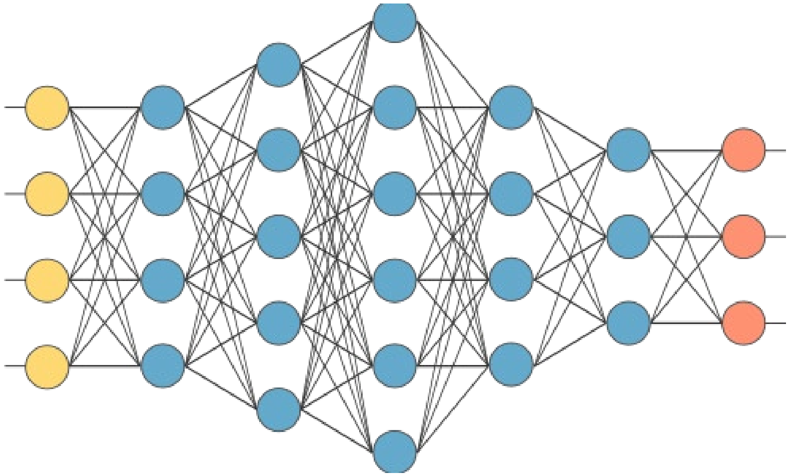
Taxonomy of Machine Learning
(A Simplistic View Based on Tasks)



can also be viewed as tools/methods

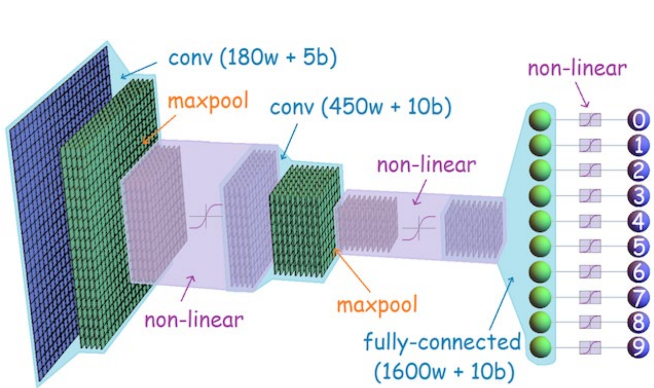
Other Tools/Topics in Machine Learning

- Deep learning basics

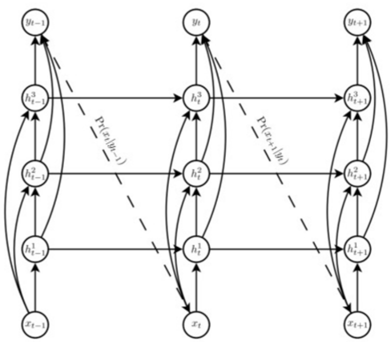


- Learning theory
  - Bias variance tradeoff
  - Feature selection
  - ML advice
- Broader aspects of ML
  - Robustness/fairness

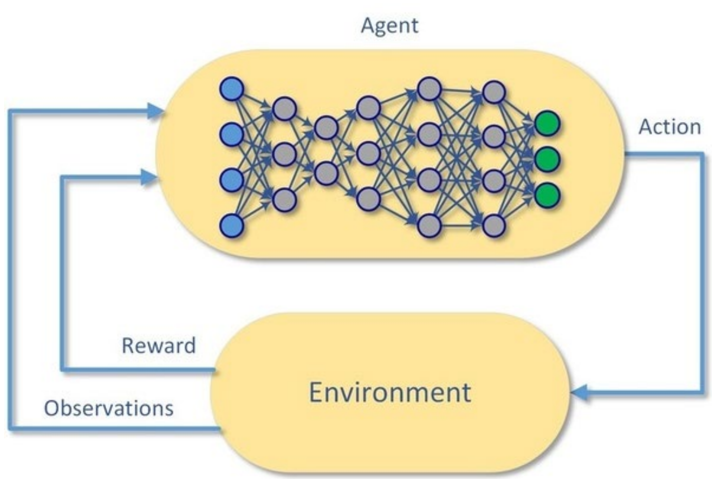
What are areas of deep learning?



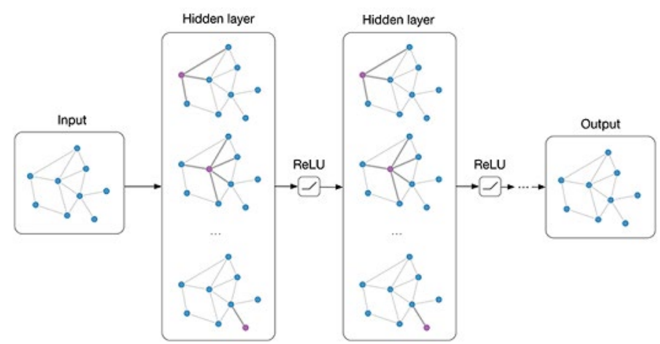
Convolutional NN Image



Recurrent NN
Time Series



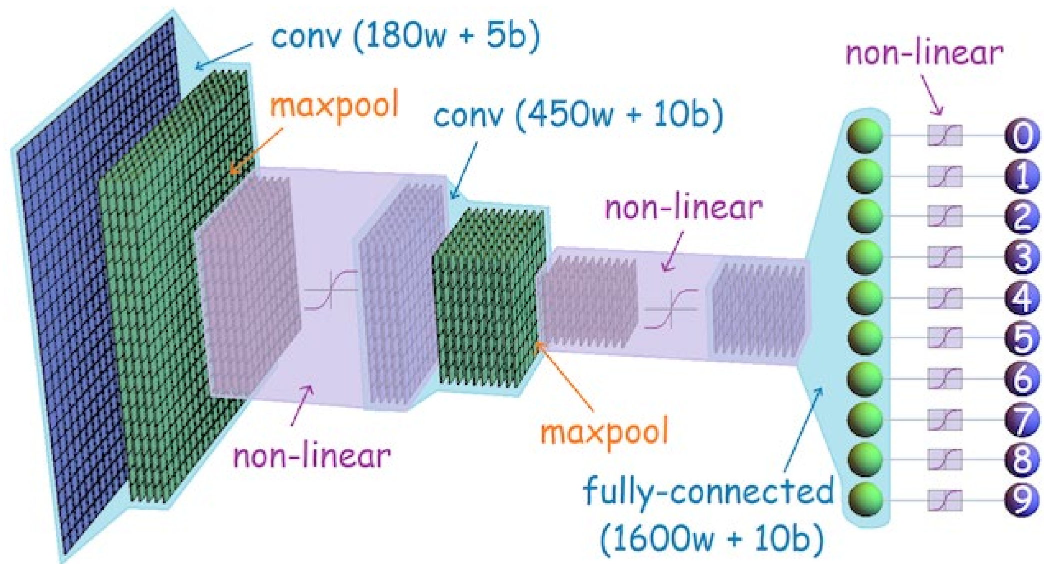
Deep RL
Control System



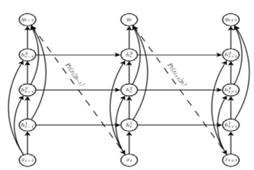
Graph NN
Networks/Relational

What are areas of deep learning?

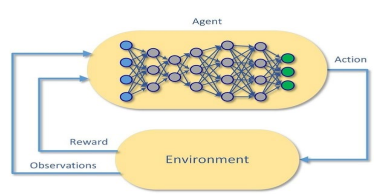
Convolutional Neural Network



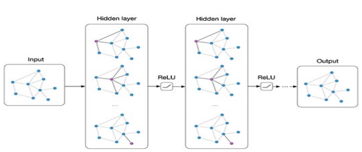
Recurrent NN



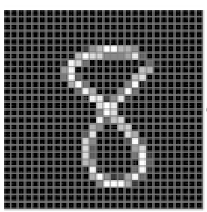
Deep RL



Graph NN

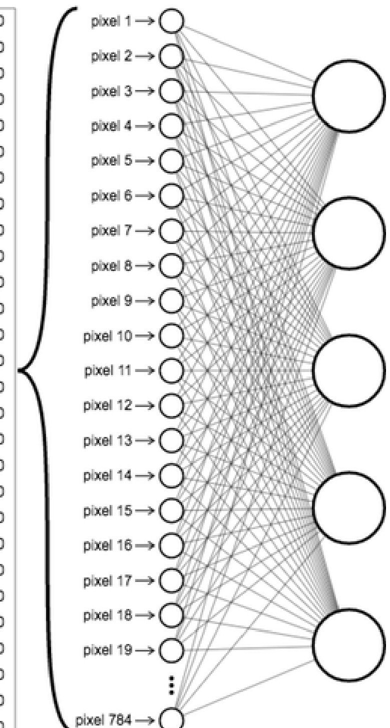


Let us look at images in detail



28 x 28
784 pixels

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	4	62	146	182	254	254	181	176	139	15	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	34	186	253	217	208	136	136	136	166	232	99	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	61	242	208	111	3	0	0	0	0	0	18	32	107	43	0	0	0	0	0	0
0	0	0	0	0	0	0	0	156	242	23	0	0	0	0	0	0	0	13	191	181	6	0	0	0	0	0	0
0	0	0	0	0	0	0	0	121	255	98	3	0	0	0	0	0	8	194	225	12	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	169	253	120	3	0	0	0	0	128	247	51	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	3	111	244	169	19	0	14	131	249	117	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	59	241	235	72	142	229	66	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	25	218	254	231	36	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	133	253	221	33	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	19	237	111	196	217	19	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	174	138	0	23	193	204	18	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	96	224	0	0	0	25	218	169	3	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	215	138	0	0	0	0	86	253	99	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	215	97	0	0	0	0	3	162	214	11	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	215	97	0	0	0	0	0	118	253	68	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	185	157	0	0	0	0	0	40	254	98	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	50	244	61	0	0	0	0	112	244	58	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	174	251	142	59	83	167	244	111	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	6	133	253	253	253	169	61	3	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

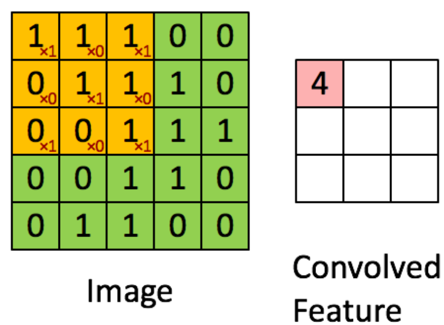


Filters

Why not extract features using filters?

Better, why not let the data dictate what filters to use?

Learnable filters!!



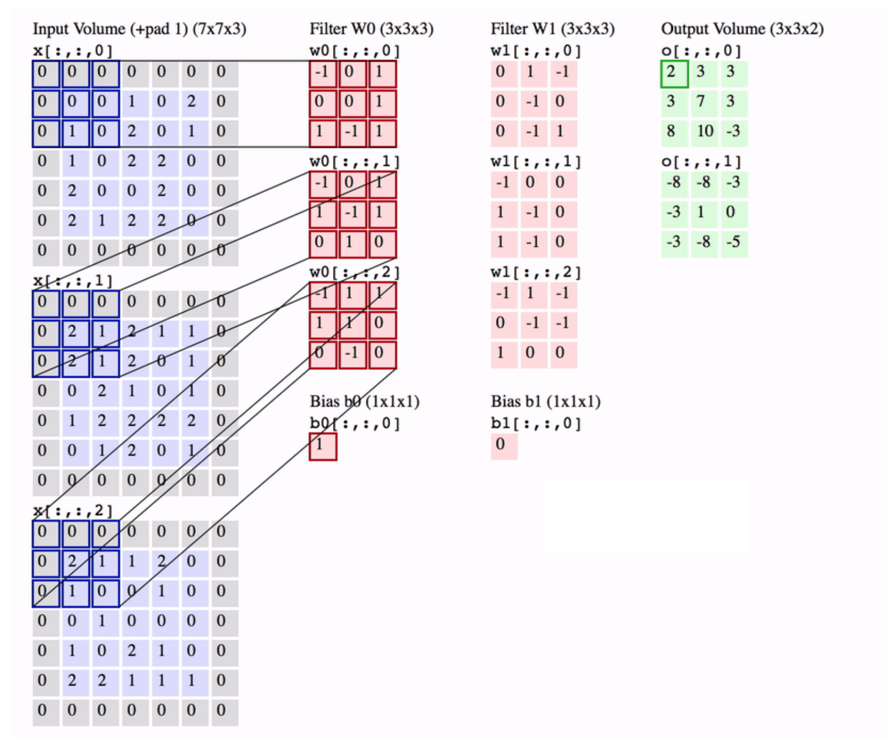
Convolution on multiple channels

Images are generally RGB !!

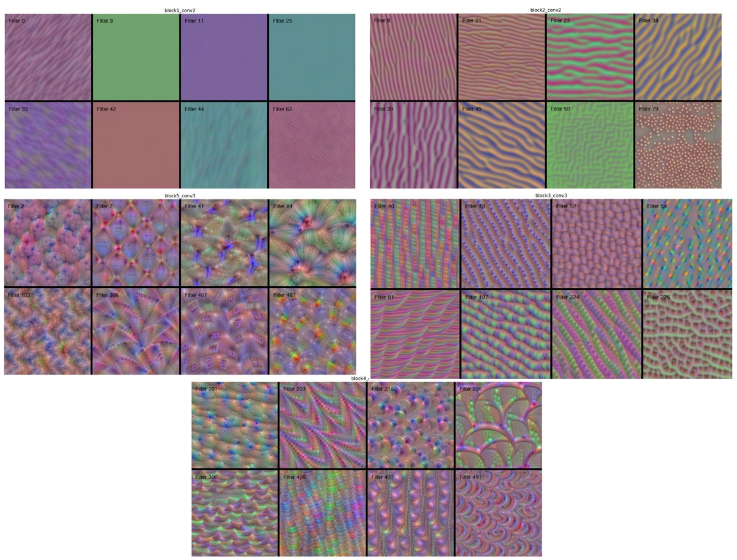
How would a filter work on a image with RGB channels?

The filter should also have 3 channels.

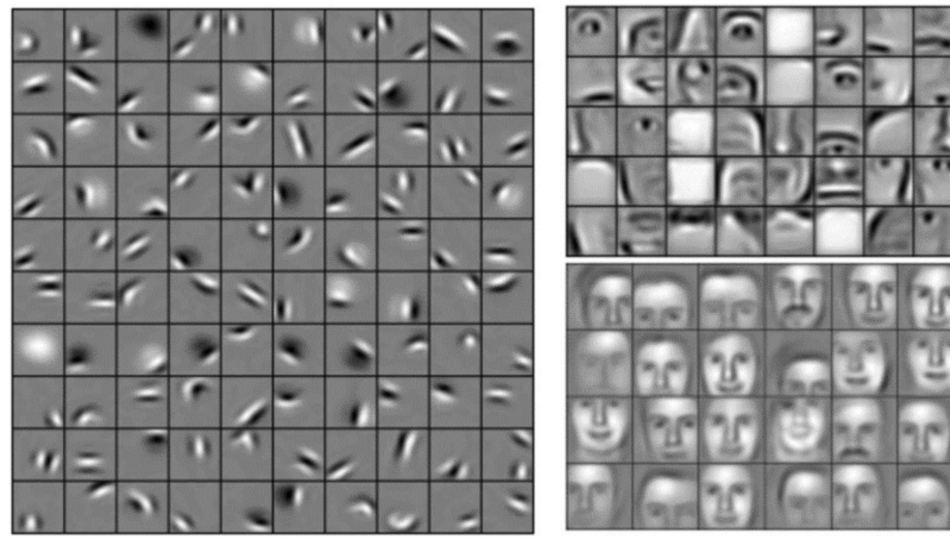
Now the output has a channel for every filter we have used.



Filters? Layers of filters?

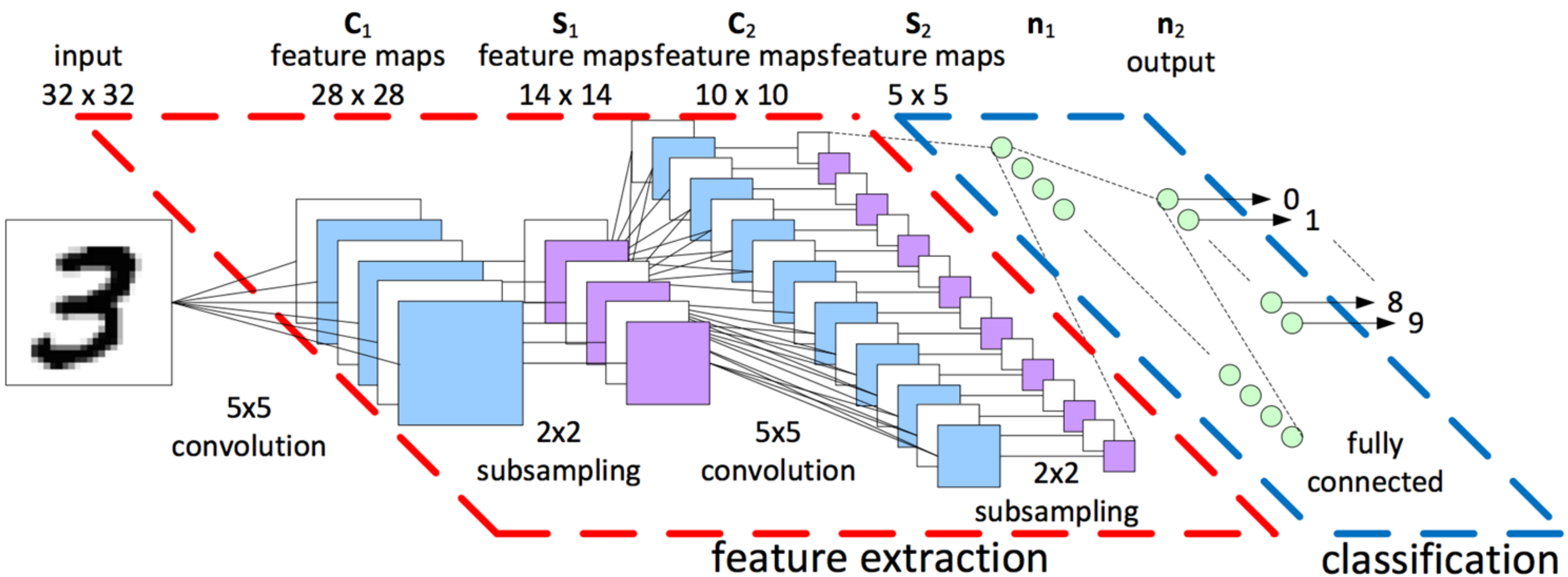


Images that maximize filter outputs at certain layers. We observe that the images get more complex as filters are situated deeper



How deeper layers can learn deeper embeddings. How an eye is made up of multiple curves and a face is made up of two eyes.

How do we use convolutions?



Let convolutions extract features and let normal cnn's decide on them.

Image credit: LeCun et al. (1998)

Convolution really is just a linear operation

$$\begin{pmatrix} x1 & x2 & x3 \\ x4 & x5 & x6 \\ x7 & x8 & x9 \end{pmatrix} * \begin{pmatrix} k1 & k2 \\ k3 & k4 \end{pmatrix}$$

$$\begin{pmatrix} k1 & k2 & 0 & k3 & k4 & 0 & 0 & 0 & 0 \\ 0 & k1 & k2 & 0 & k3 & k4 & 0 & 0 & 0 \\ 0 & 0 & 0 & k1 & k2 & 0 & k3 & k4 & 0 \\ 0 & 0 & 0 & 0 & k1 & k2 & 0 & k3 & k4 \end{pmatrix} . \begin{pmatrix} x1 \\ x2 \\ x3 \\ x4 \\ x5 \\ x6 \\ x7 \\ x8 \\ x9 \end{pmatrix}$$

$$\begin{pmatrix} k1\,x1 + k2\,x2 + k3\,x4 + k4\,x5 \\ k1\,x2 + k2\,x3 + k3\,x5 + k4\,x6 \\ k1\,x4 + k2\,x5 + k3\,x7 + k4\,x8 \\ k1\,x5 + k2\,x6 + k3\,x8 + k4\,x9 \end{pmatrix}$$

In fact convolution is a giant matrix multiplication.

We can expand the 2 dimensional image into a vector and the convolution operation into a matrix.