CS4980-003: Introduction to Deep Learning with Python

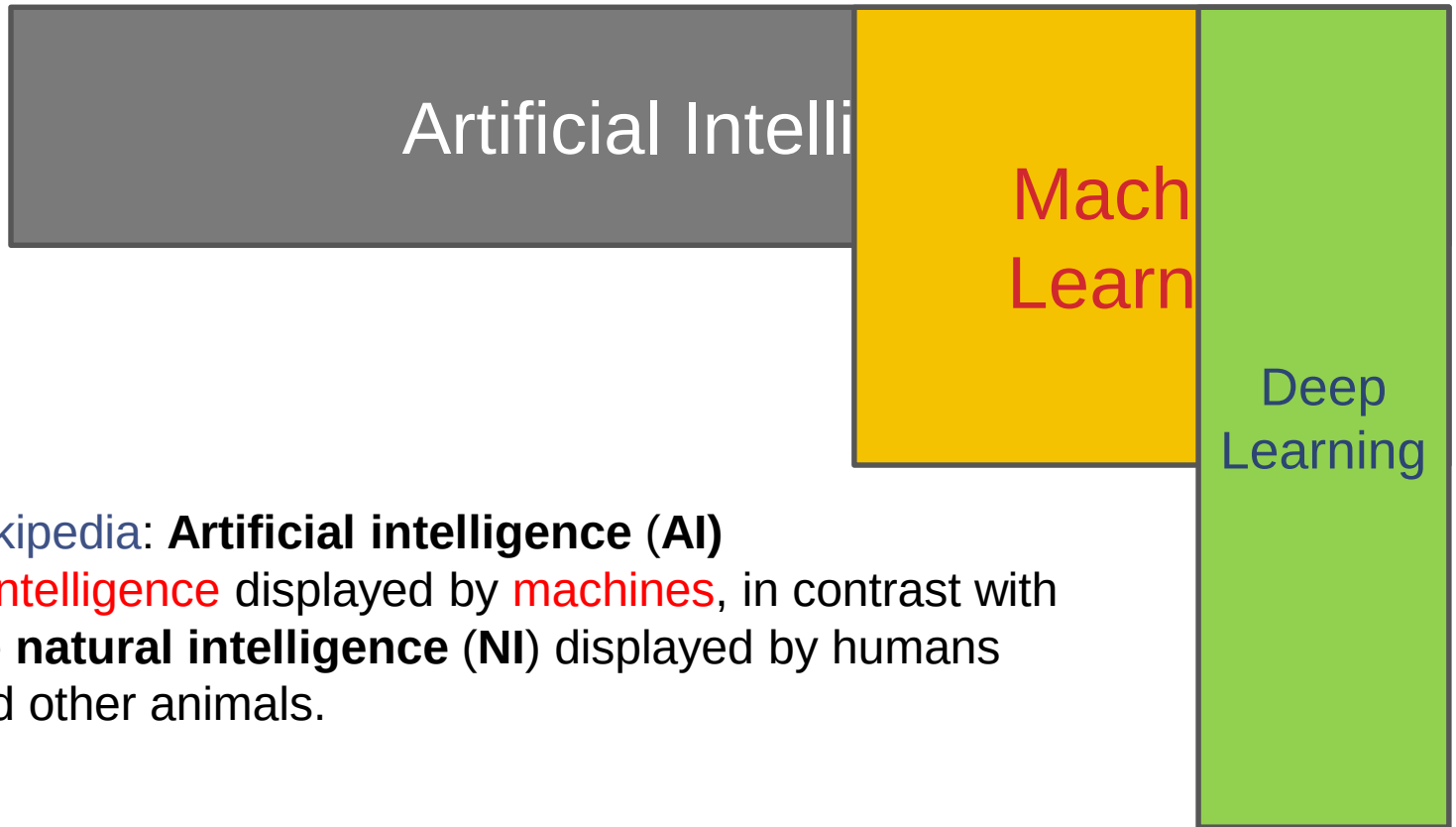
Hantao Zhang

Spring 2018

What is Deep Learning ?

- Deep Learning is a Machine Learning method using Deep Neural Networks.

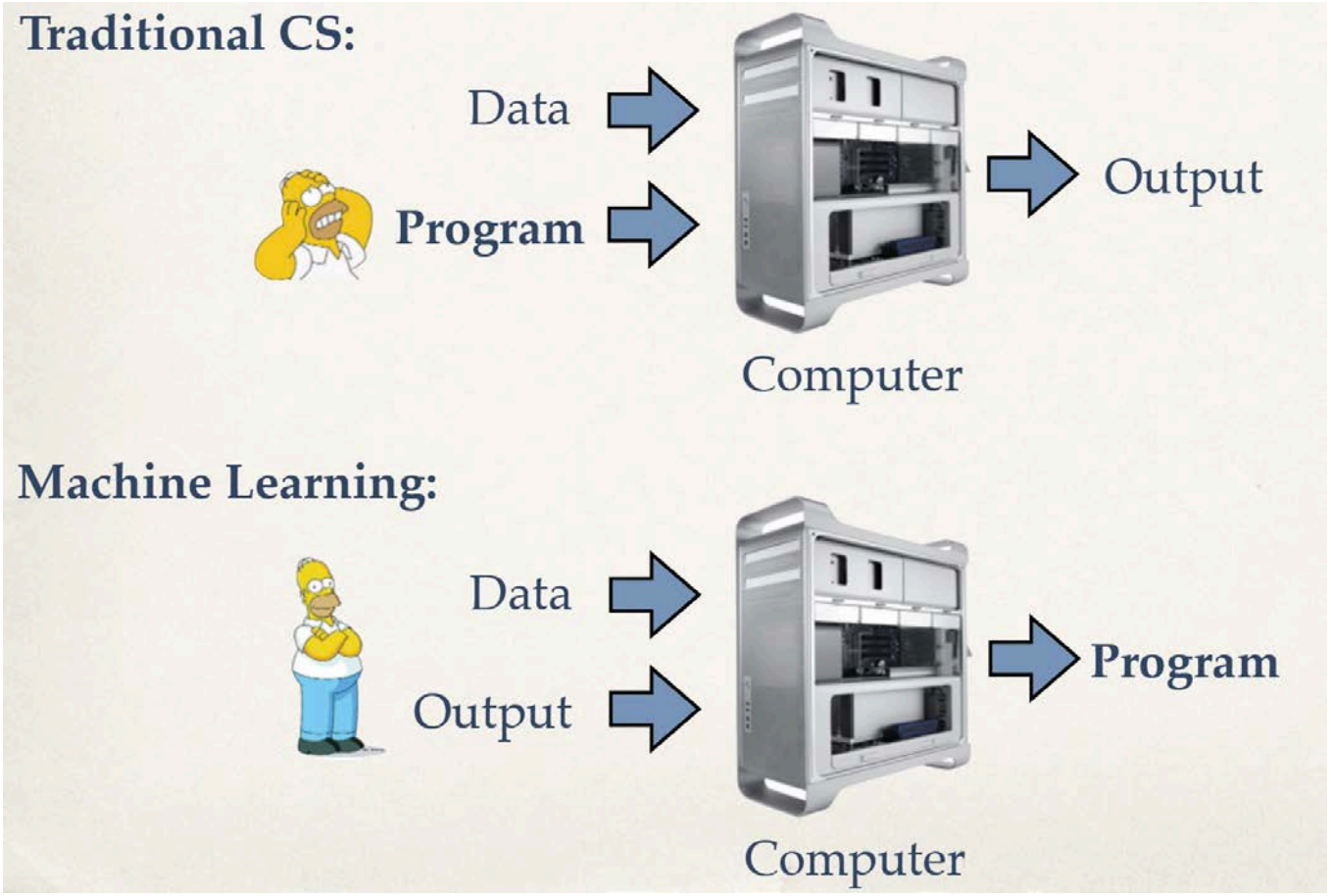
- Deep Learning as a course in Computer Science:



Wikipedia: **Artificial intelligence** (**AI**)
is intelligence displayed by machines, in contrast with the **natural intelligence** (**NI**) displayed by humans and other animals.

What is machine learning?

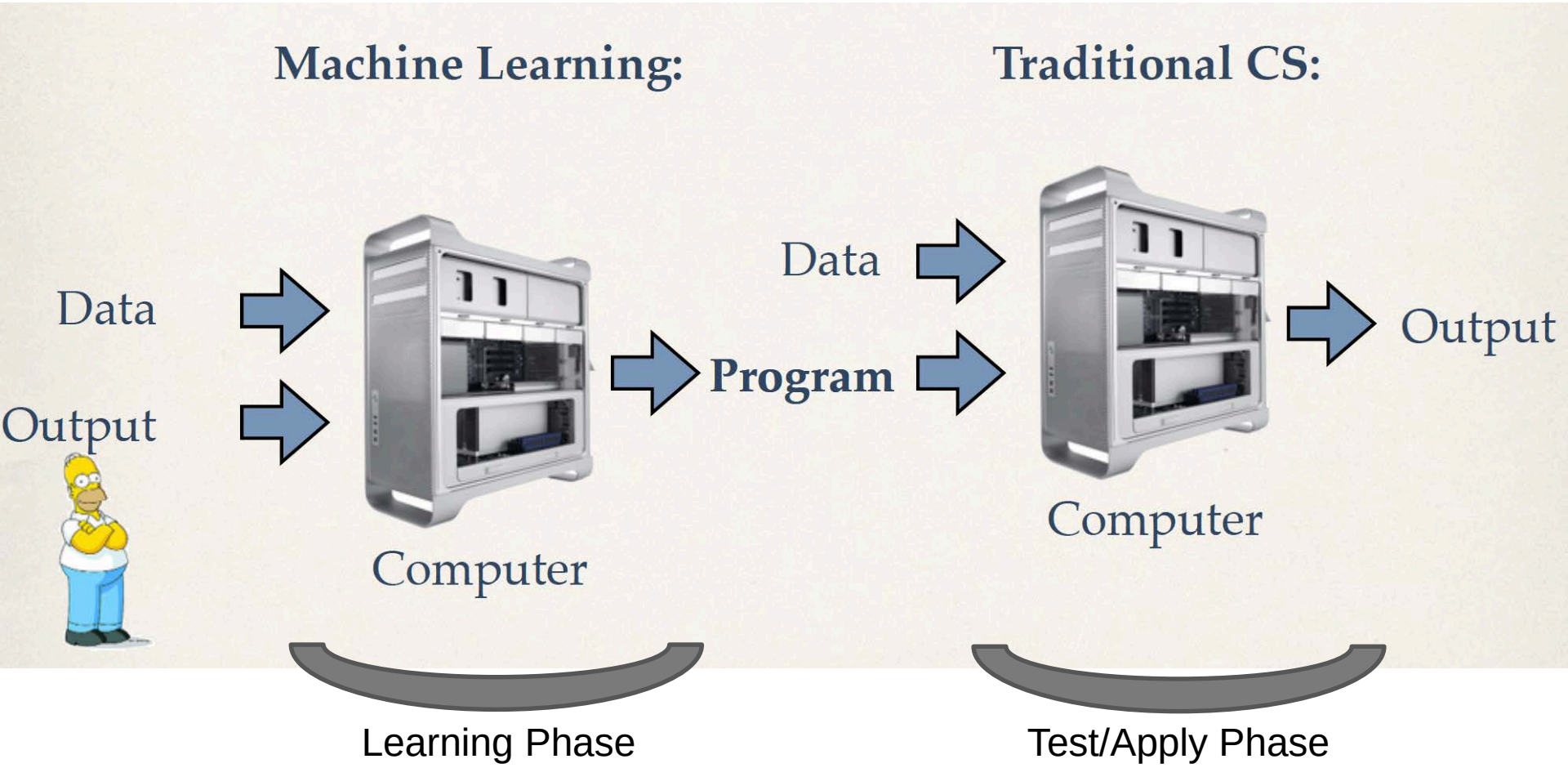
- Use of Computers:



picture courtesy: kilian weinberger

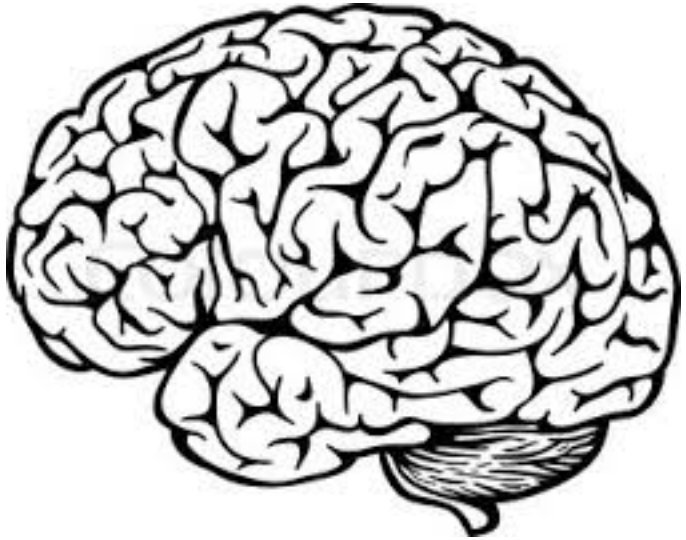
What is machine learning?

- Use of Computers:



How do people learn things?

- Memorization
  - Accumulation of individual facts Limited by
  - time to observe and memory to store the facts



How do people learn things?

- Generalization
  - Deduce new facts from old ones

Chairs

Chair?

How do people learn things?

- Generalization
  - Deduce new facts from old ones
  - limited by the number of facts

Chairs

Chair?

What is Machine Learning

- Tom M. Mitchell (CMU)
  - A computer program is said to learn from experience E with respect to some task T and performance measure P, if it
  - improves performance P at task T with experience E
- Example: Computer Checkers
  - T (task): play checkers
  - P (performance): probability to win
  - E (experience): play the game with itself many times

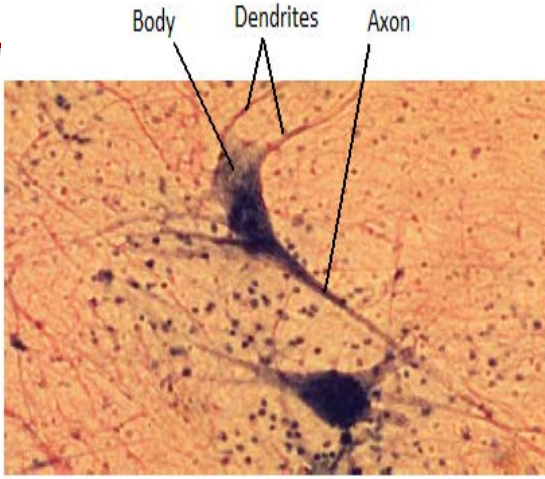
What is Machine Learning

- Memorization
  - If we regard the learned program as a multi-variable polynomial function, then coefficients in the polynomial are modified and stored during the learning phase.
- Generalization
  - The desired property
  - Overfitting: the opposite side of generalization. The learned program works well only for the examples used during learning, not for general examples.

Categorization of Machine Learning

- Based on tasks
  - regression (e.g. predict house price)
  - classification (e.g. predict spam)
  - clustering, ranking, etc.
- Based on availability of labels on inputs (desired output)
  - yes: supervised learning (classification, regression)
  - no: unsupervised learning (clustering)
  - partially: semi-supervised learning (classification, regression)
  - late after multiple moves: reinforcement learning (games)

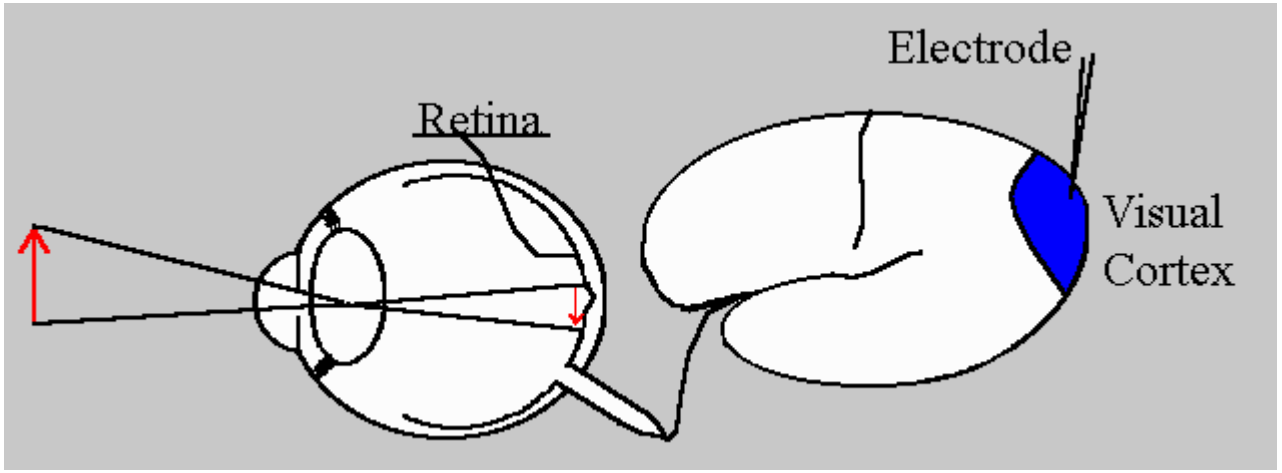
Artificial Neural Network



- Try to mimic biological neural network, such as the brain.
- The brain has approximately 100 billion neurons, which communicate through electro-chemical signals.
- Each neuron has thousands of connections with other neurons, constantly receiving incoming signals to reach the cell body.
- If the resulting sum of the signals surpasses a certain threshold, a response is sent through the axon.

How does human brain work?

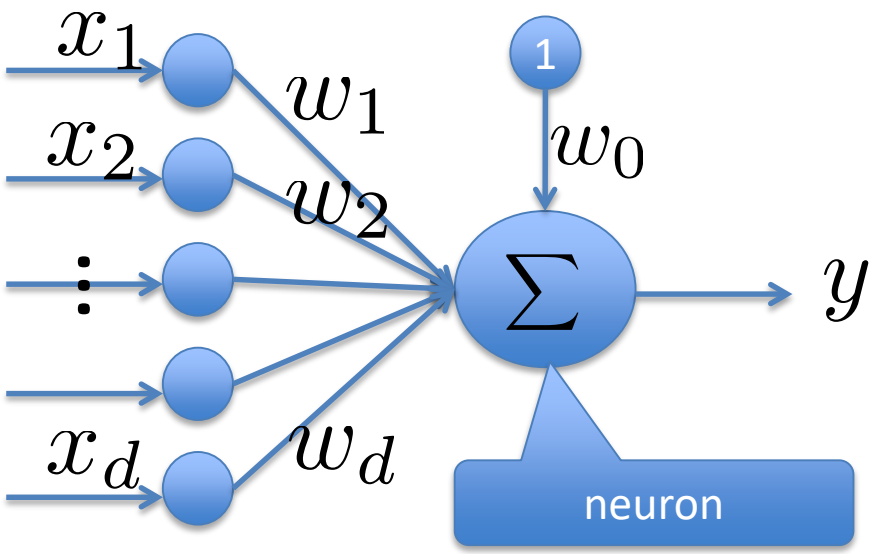
- it is very complicated (even the experts do not understand completely)



Artificial Neuron (Perceptron)

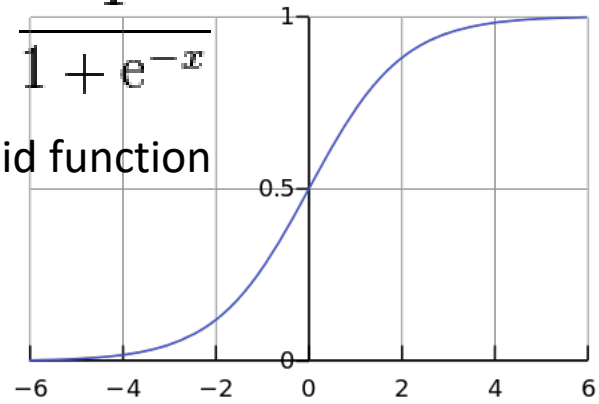
$\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T\mathbf{x} = w_0x_0 + {}_1x_1 + w_2x_2 + \ldots + w_dx_d$

where $x_0 = 1$



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

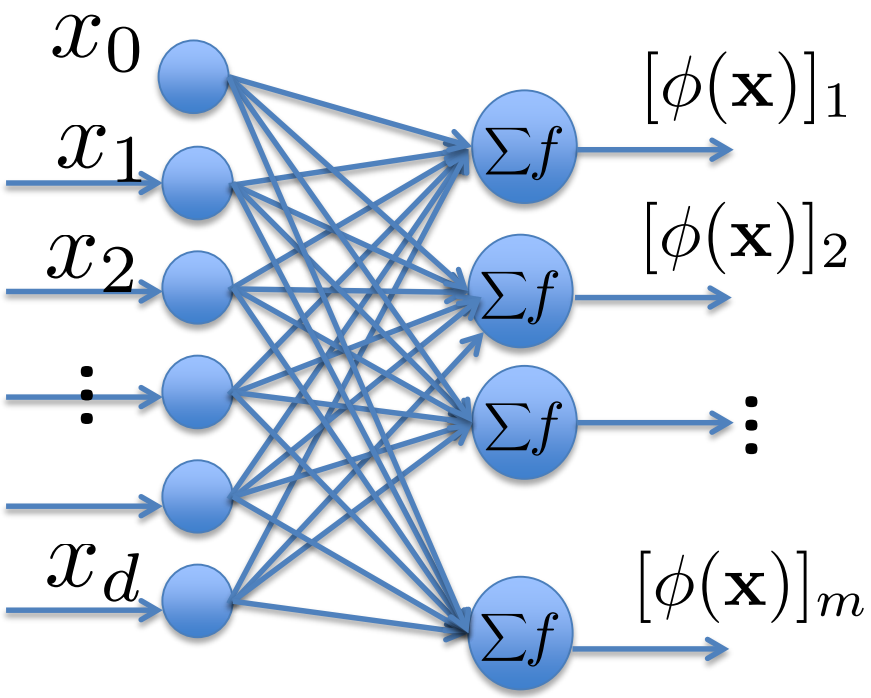
Sigmoid function



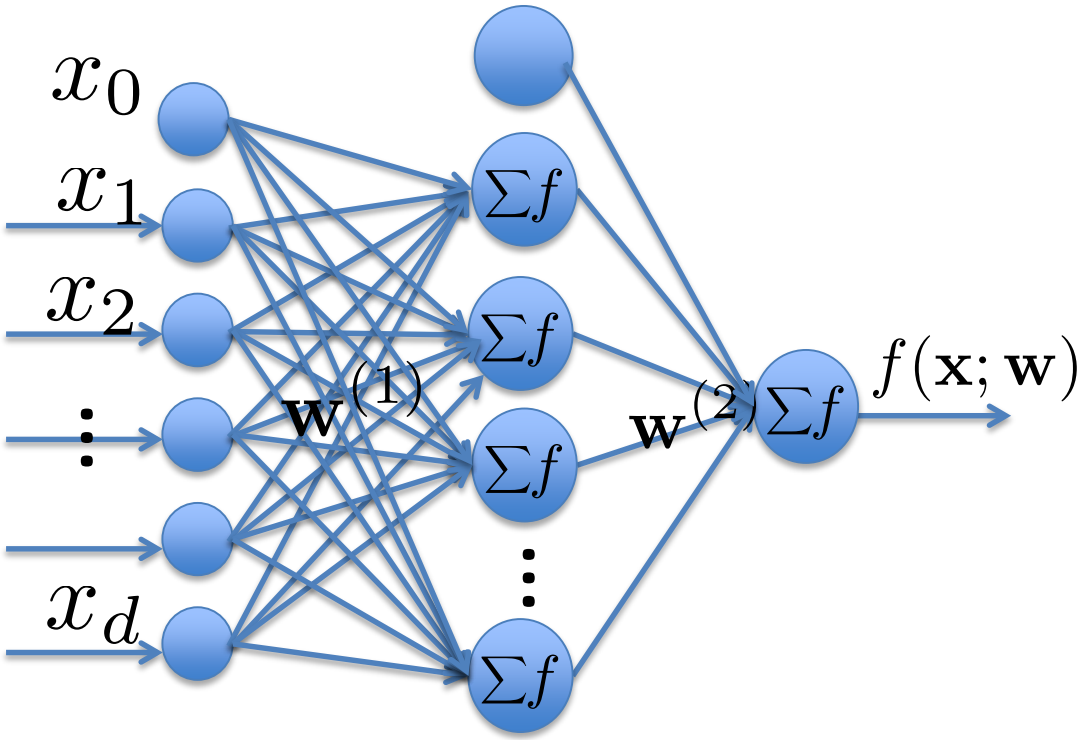
Logistic regression: $\Pr(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top\mathbf{x}) = \dfrac{1}{1 + \exp(-(\mathbf{w}^\top\mathbf{x}))}$

Building-block of Multi-layer NN

- $[\phi(\mathbf{x})]_j = h(\mathbf{x}^\top \mathbf{w}_j^{(1)} + w_{j0}^{(1)})$

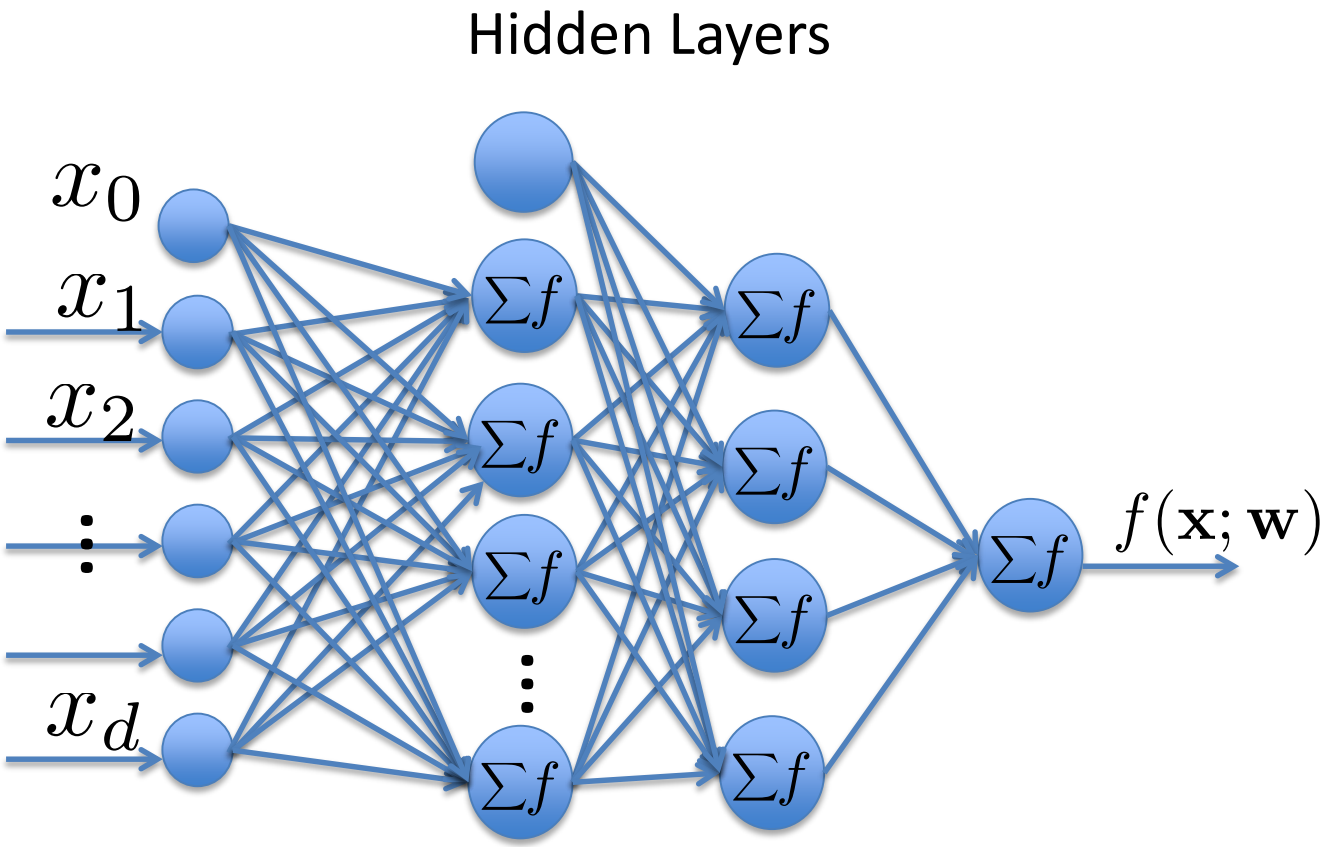


Two-layer NN

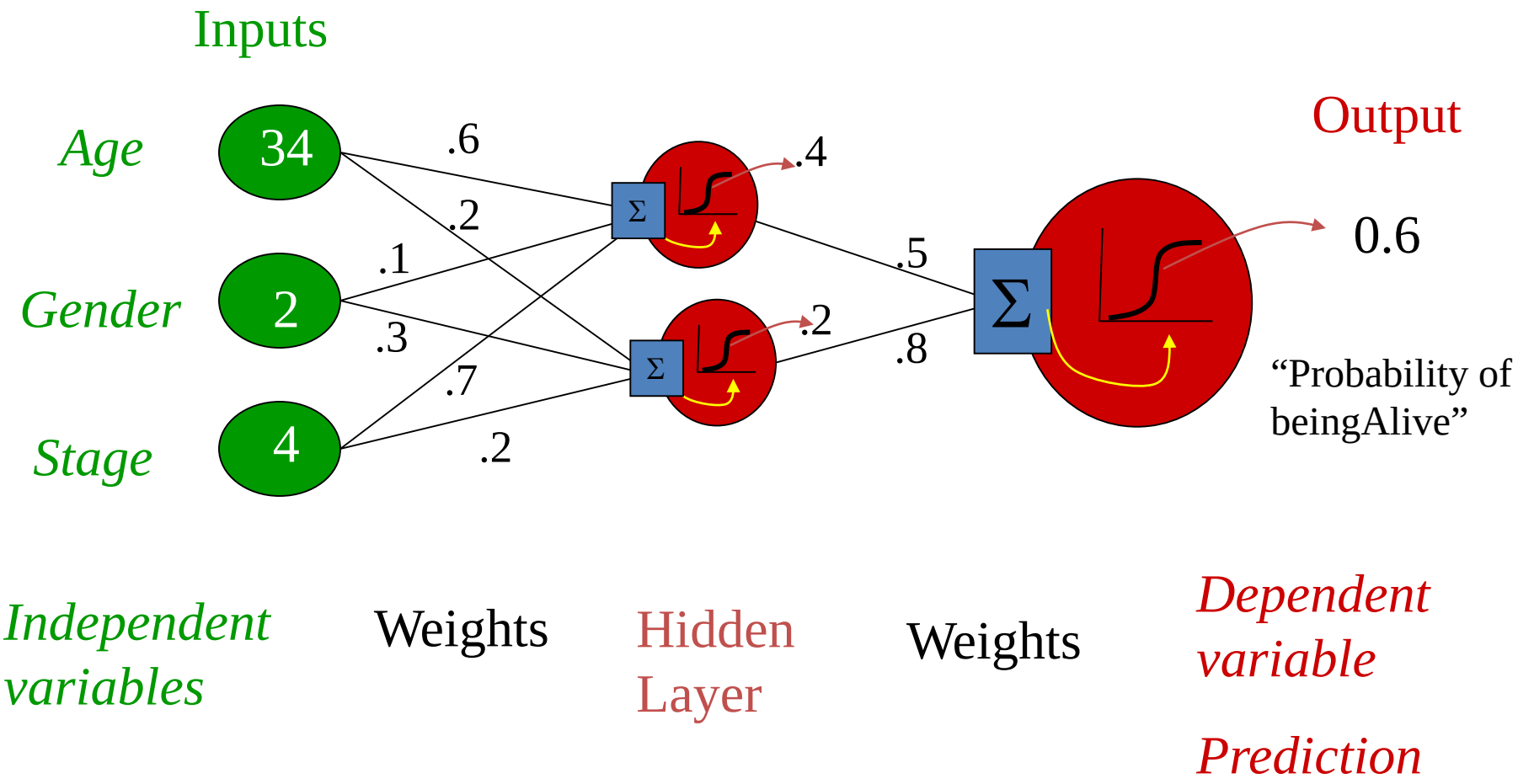


$$f(\mathbf{x};\mathbf{w}) = \sigma\left(\phi(\mathbf{x};\mathbf{w}^{(1)})^\top \mathbf{w}^{(2)} + w_0^{(2)}\right)$$

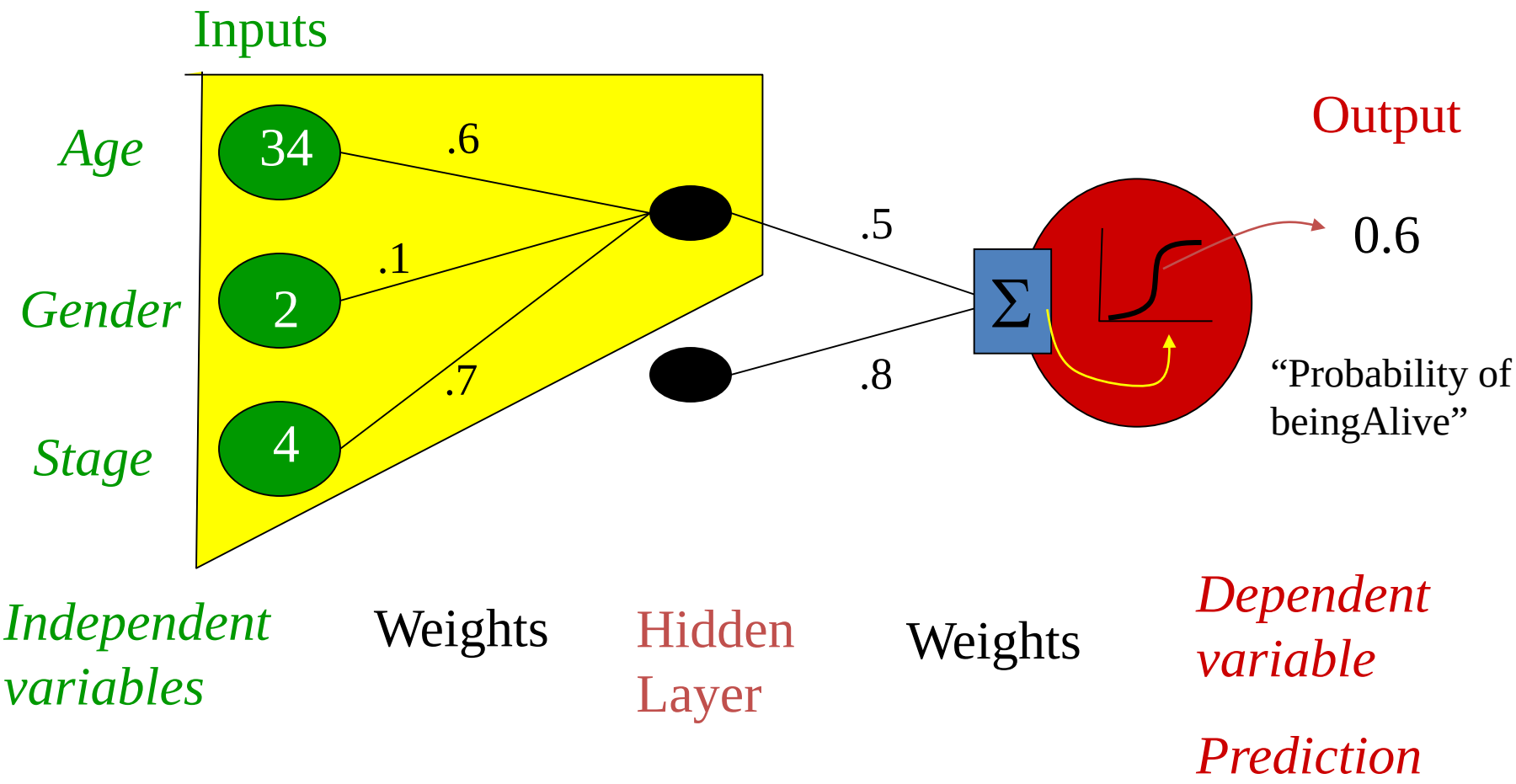
Multi-layer NN



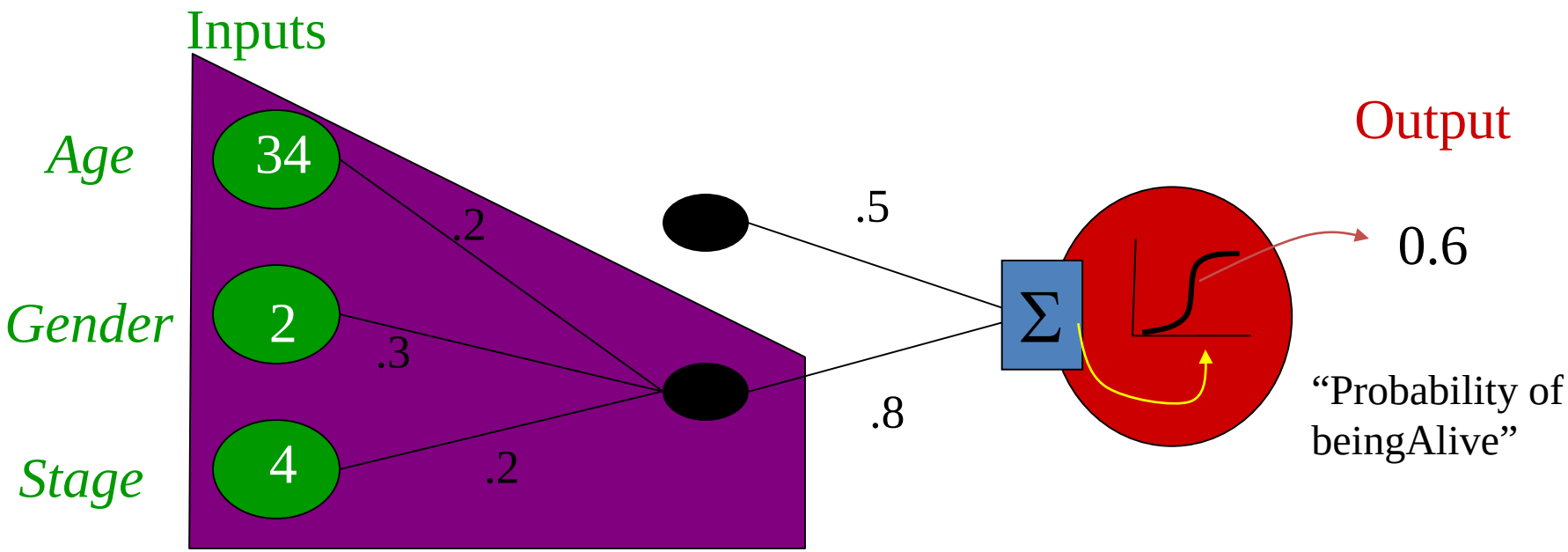
Example of 2-layer Neural Network



Feed Forward Computation



Feed Forward Computation



Independent variables

Weights

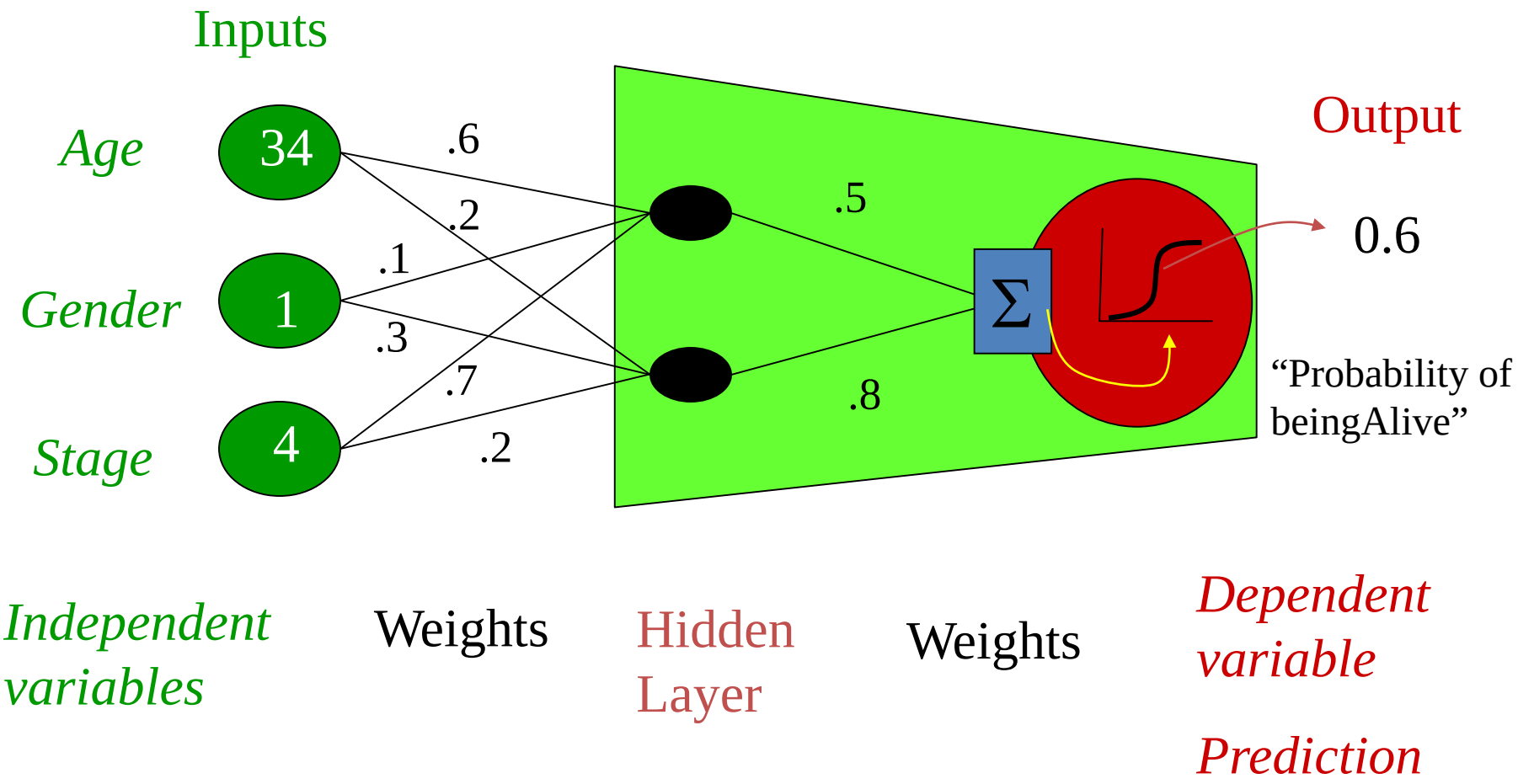
Hidden Layer

Weights

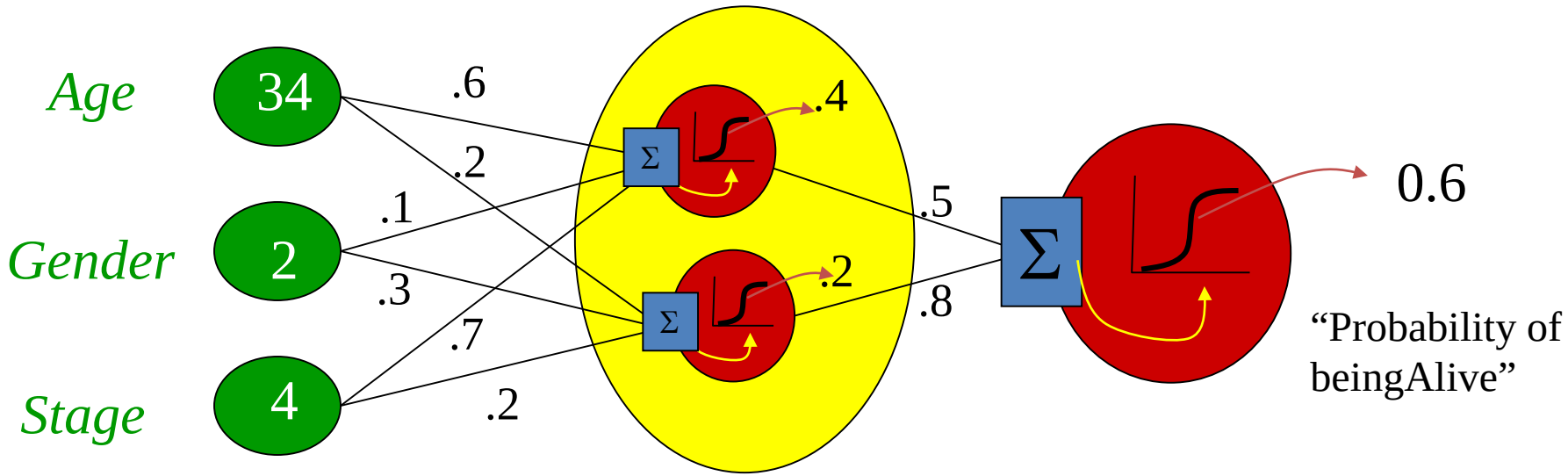
Dependent variable

Prediction

Feed Forward Computation



Not really, no target for hidden units...



Independent variables

Weights

Hidden Layer

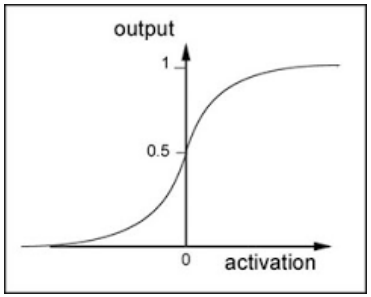
Weights

Dependent variable

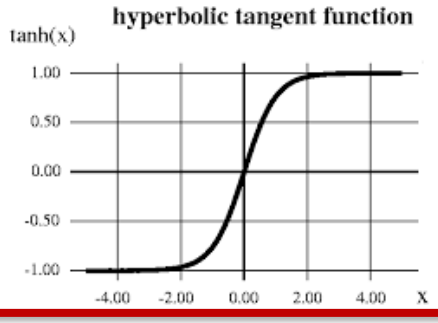
Prediction

Activation Function (hidden layers)

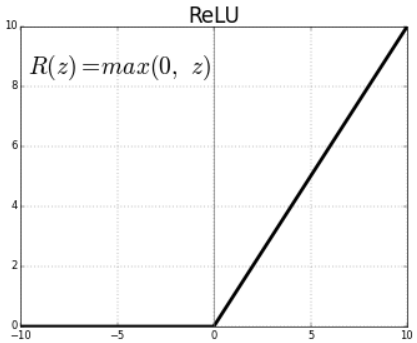
- sigmoid



- tanh



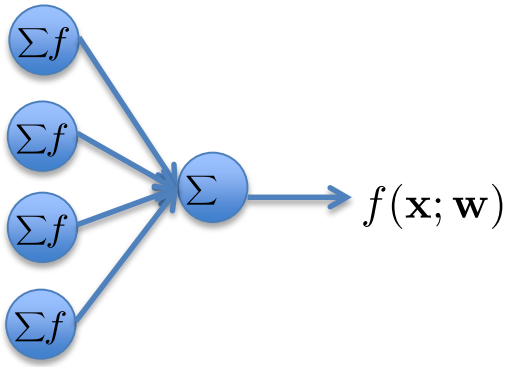
- rectified linear



https://en.wikipedia.org/wiki/Activation_function

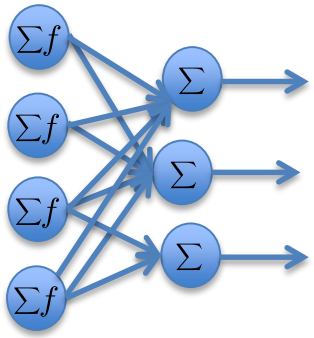
Activation Function (output layer)

- identity (regression)



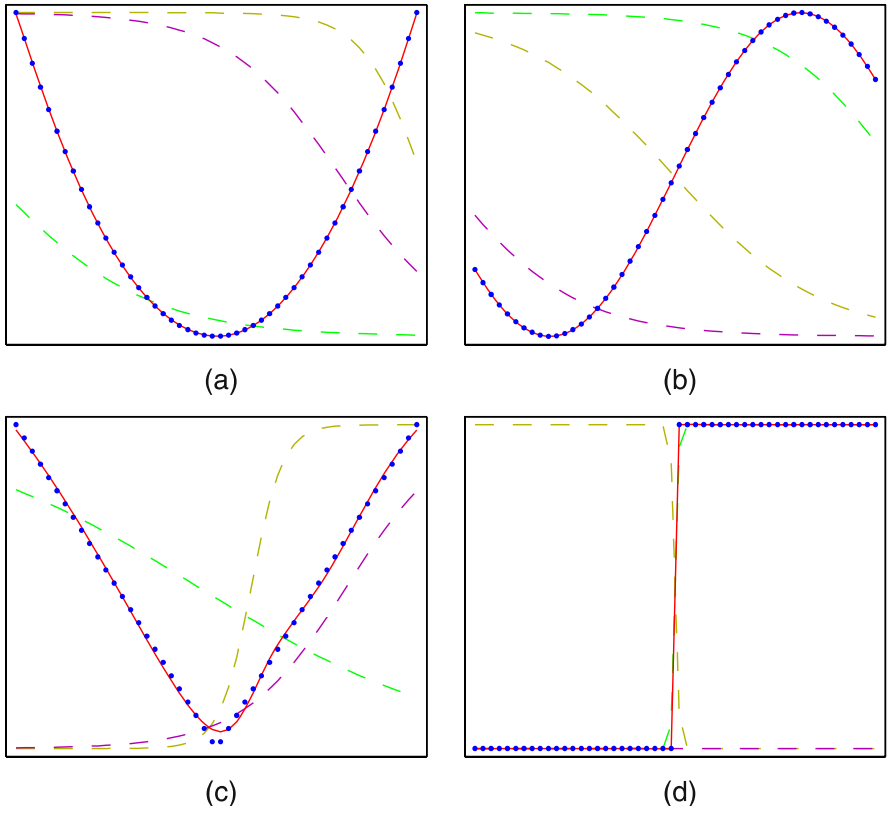
- sigmoid (logistic regression)

- softmax (multi-class classification)



Why Neural Network?

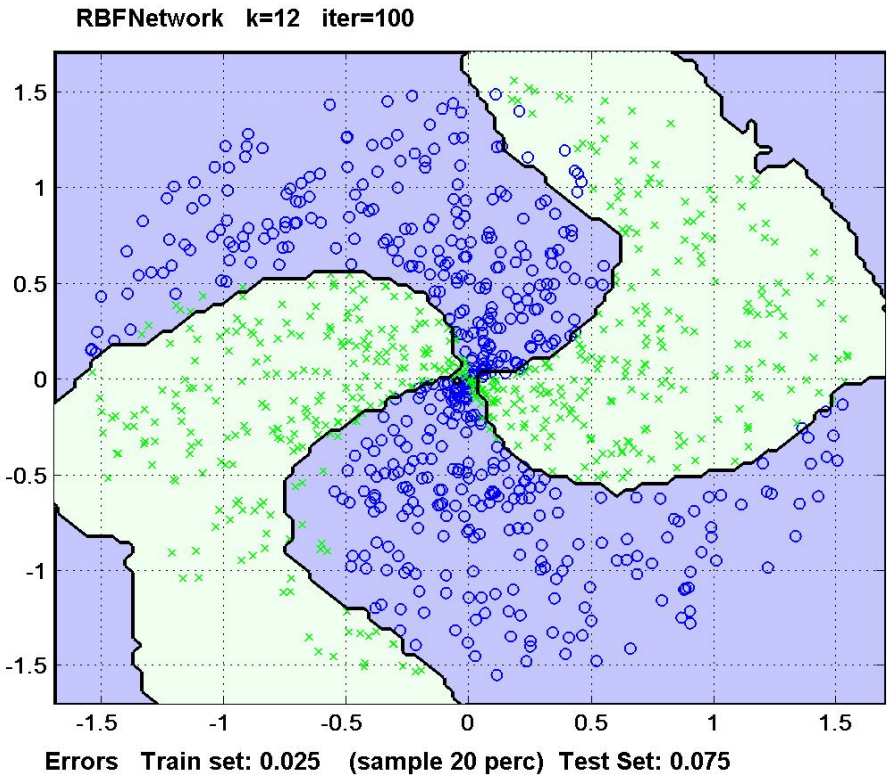
- universal approximators
  - Can achieve any non-linear function on a compact input domain



3 hidden units NN with tanh activation functions

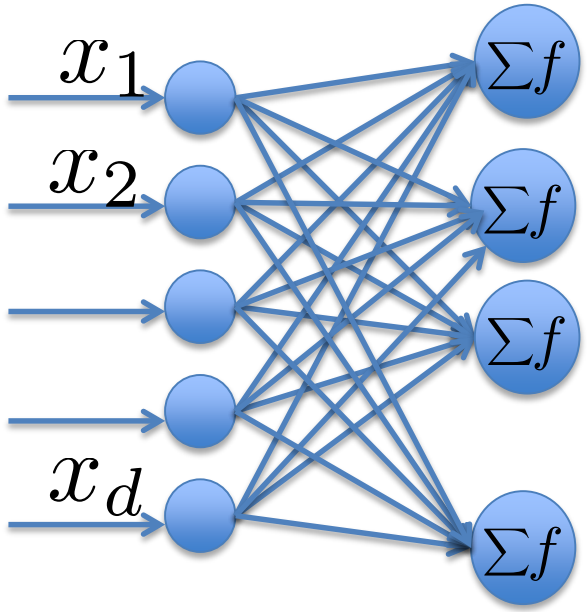
Why Neural Network?

- Can achieve any decision boundary

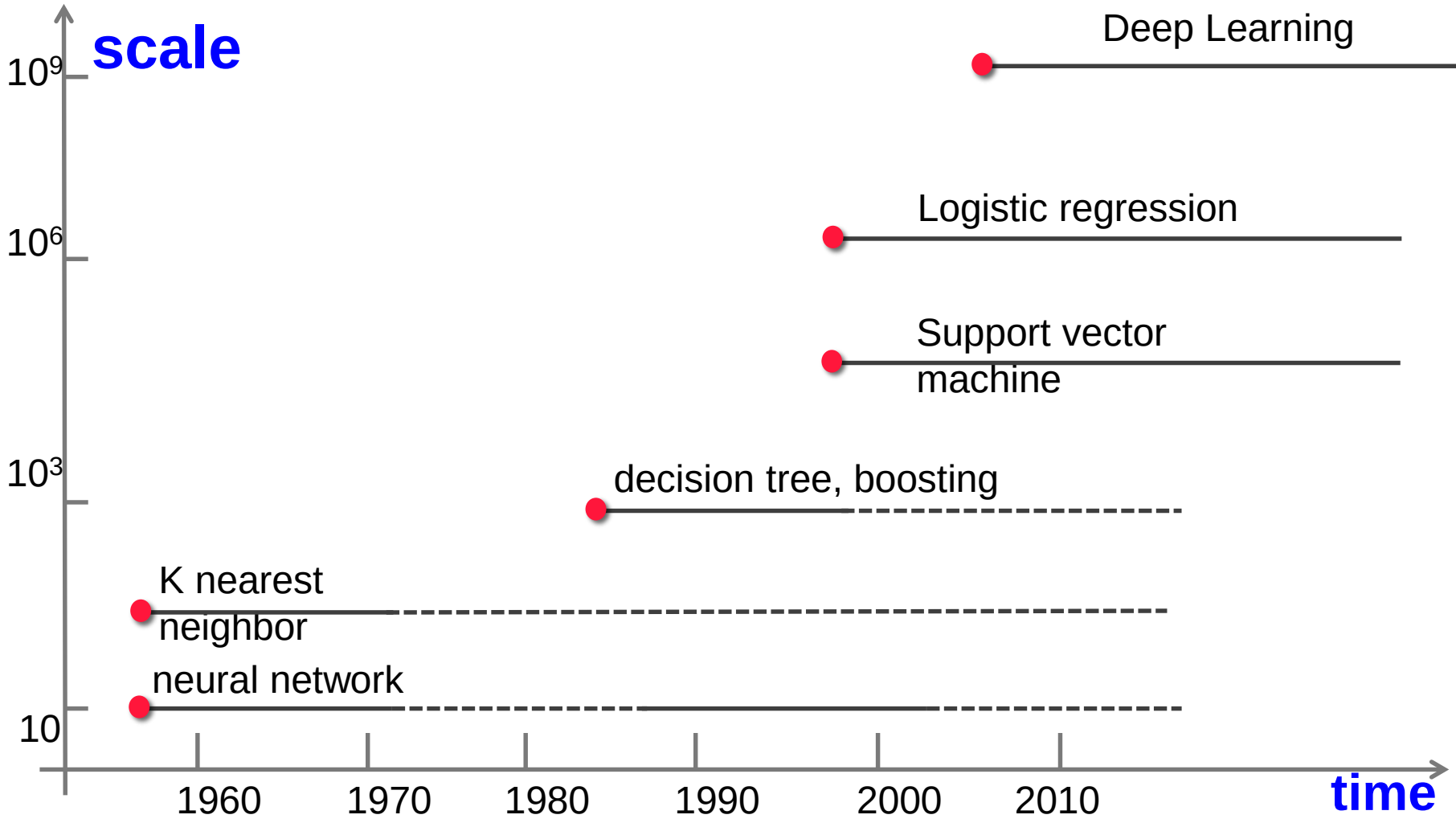


Why Neural Network?

- Can be trained using
  - (Stochastic) Gradient Descent (SGD)
  - Back Propagation
- Perceptron can be completely trained (for generalized linear functions only)

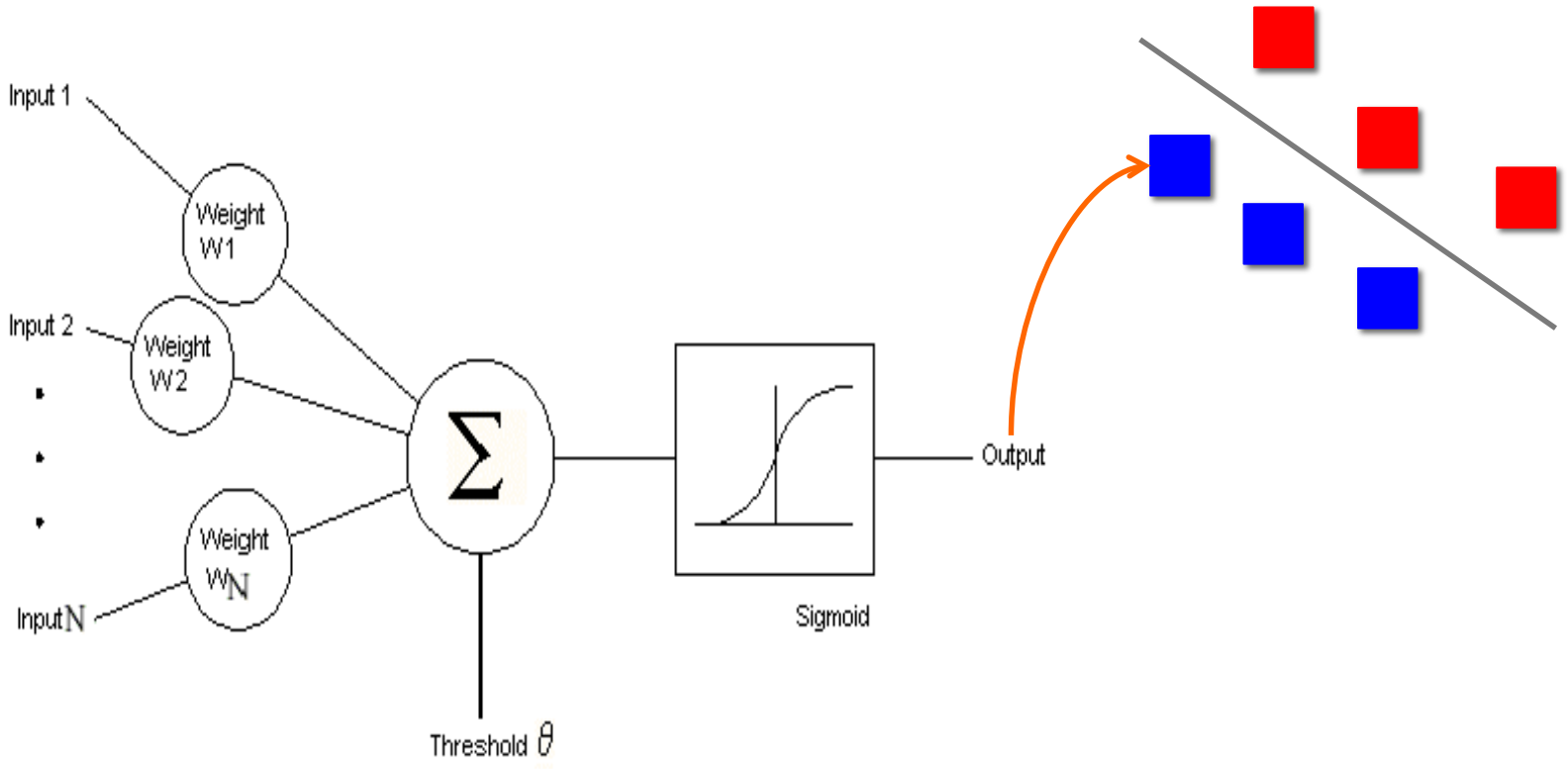


History of Machine Learning



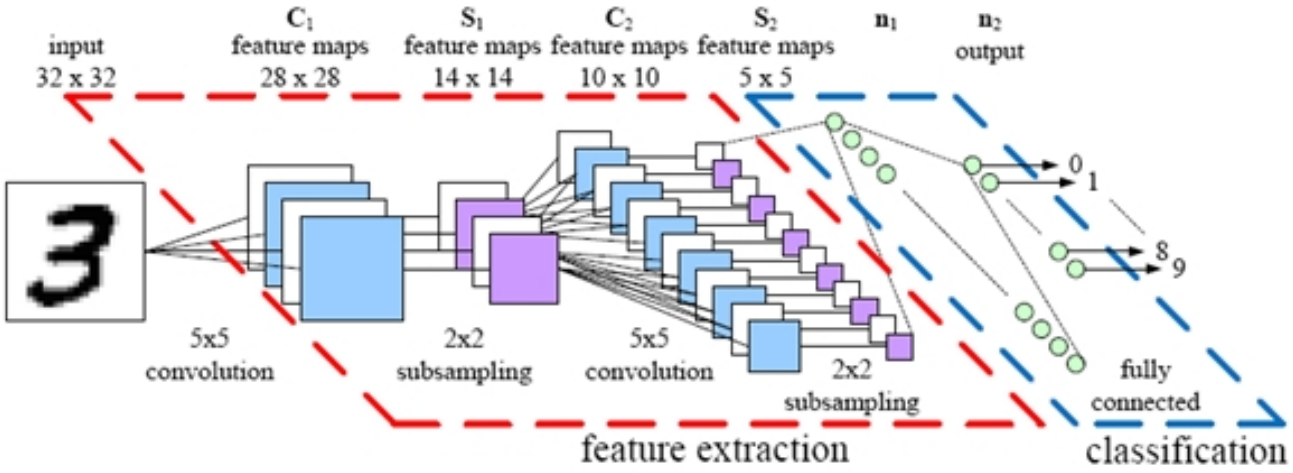
Earlier Days

- One Layer Neural Networks (1960 – 1970)
  - Perceptron (Frank Rosenblatt)



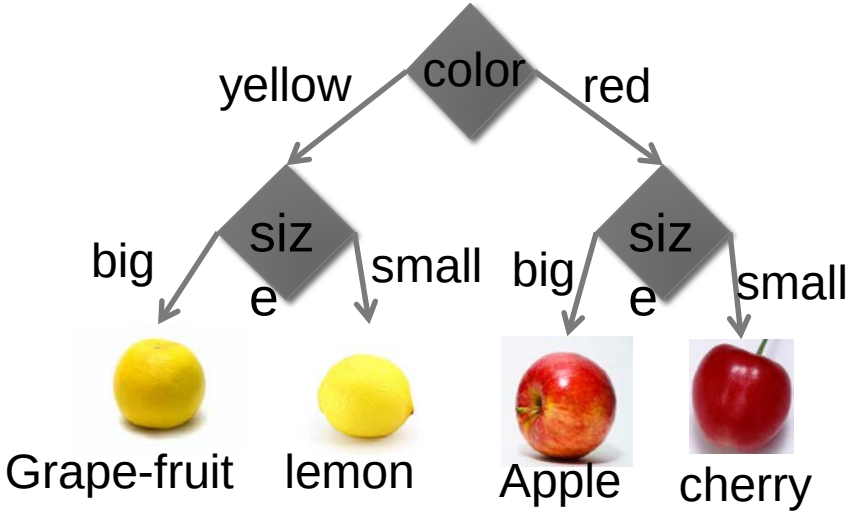
1980 – 2000

- Multiple Layers Neural Network
  - Several layers
  - Application: digit recognition



1980 – 2000

- Decision tree
  - Rule based model
  - Easy to understand
  - Application: commercial systems (credit risk analysis)



Since Mid-90’s

- Support Vector Machine
  - Vladimir N. Vapnik
  - widely studied
- Logistic Regression (for classification)
  - Application: internet

Since 2000, “Deep Learning” appears

- Around 1998, LeCun of NYU applied gradient-based learning to the idea of convolutional neural network and obtained good results on images and pattern recognition.
- Around 2006, Hinton of Toronto introduced the idea of unsupervised pretraining, and deep belief nets. The idea was to train a simple 2-layer unsupervised model like a restricted Boltzman machine, freeze all other parameters, and train *just the parameters* for one new layer. You would keep adding and training layers until you had a deep network.
- Around 2008, Bengio of Montreal used the idea of Autoencoder for unsupervised pretraining for deep neural networks.
- The three published together an Nature article on deep learning in 2015.

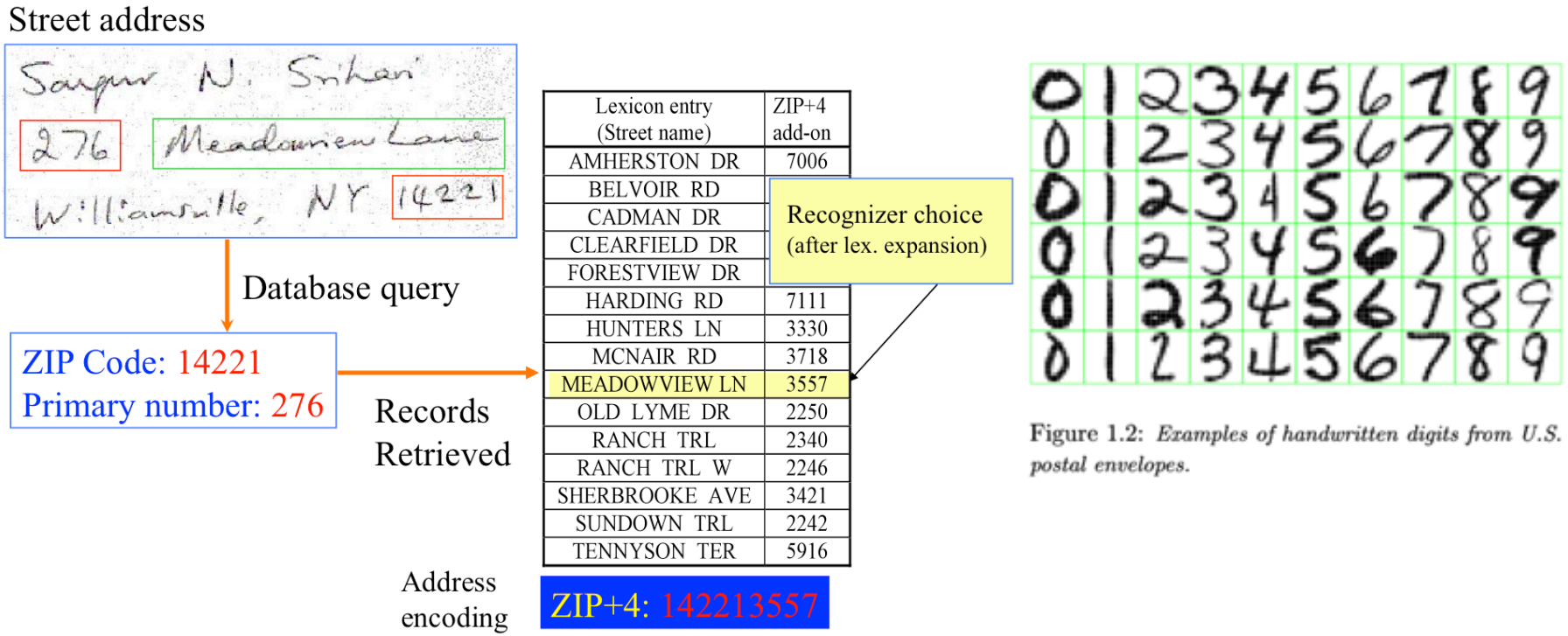
Regression

- Predict the examination score of students
  - study the effectiveness of schools

- Training Data
  - student-dependent features:
    - hours of study
    - gender, ethnic group
  - school-dependent features:
    - percentage of students eligible for free school meals
    - school gender, school denomination

Classification

- Handwritten Recognition
  - Postal address recognition
  - more than 95% of handwritten mail is sorted automatically

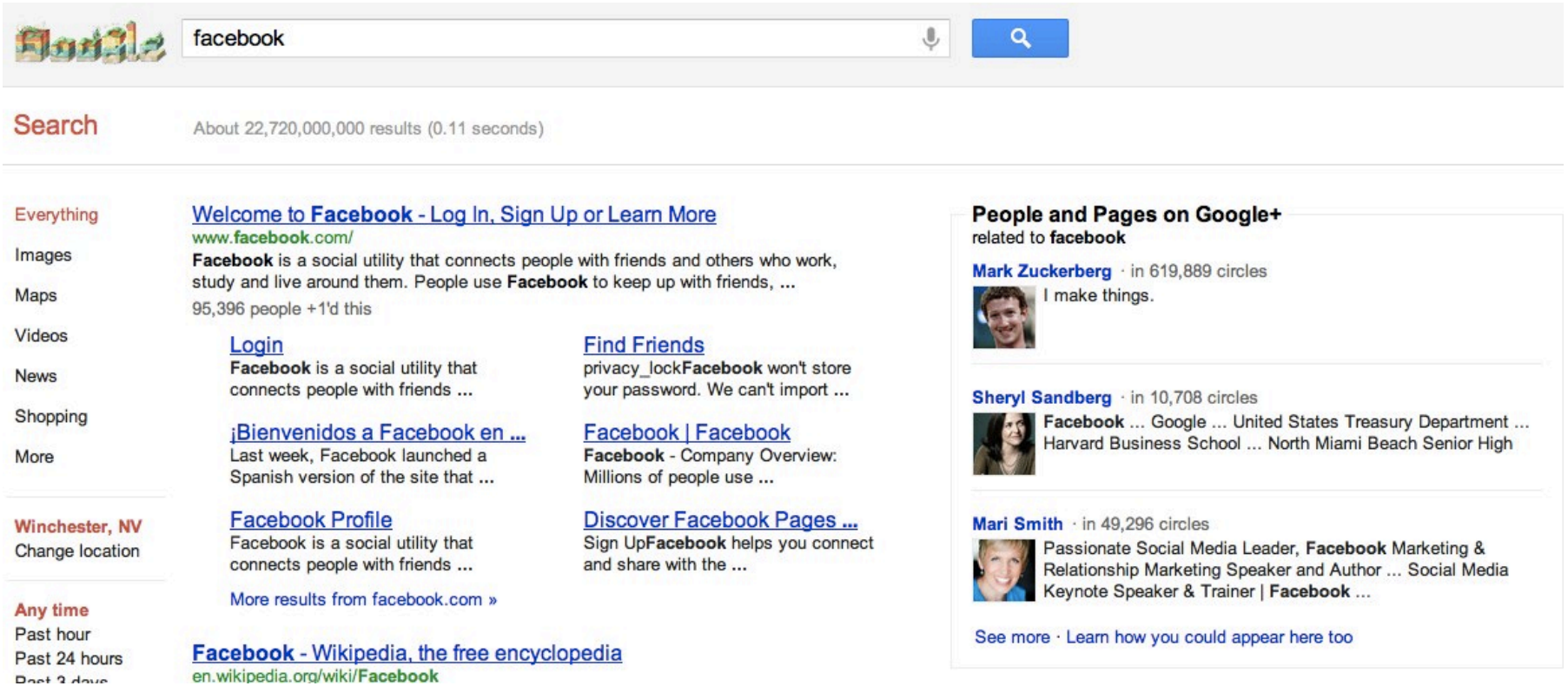


Lexicon entry (Street name)	ZIP+4 add-on
AMHERSTON DR	7006
BELVOIR RD	
CADMAN DR	
CLEARFIELD DR	
FORESTVIEW DR	
HARDING RD	7111
HUNTERS LN	3330
MCNAIR RD	3718
MEADOWVIEW LN	3557
OLD LYME DR	2250
RANCH TRL	2340
RANCH TRL W	2246
SHERBROOKE AVE	3421
SUNDOWN TRL	2242
TENNYSON TER	5916

Figure 1.2: *Examples of handwritten digits from U.S. postal envelopes.*

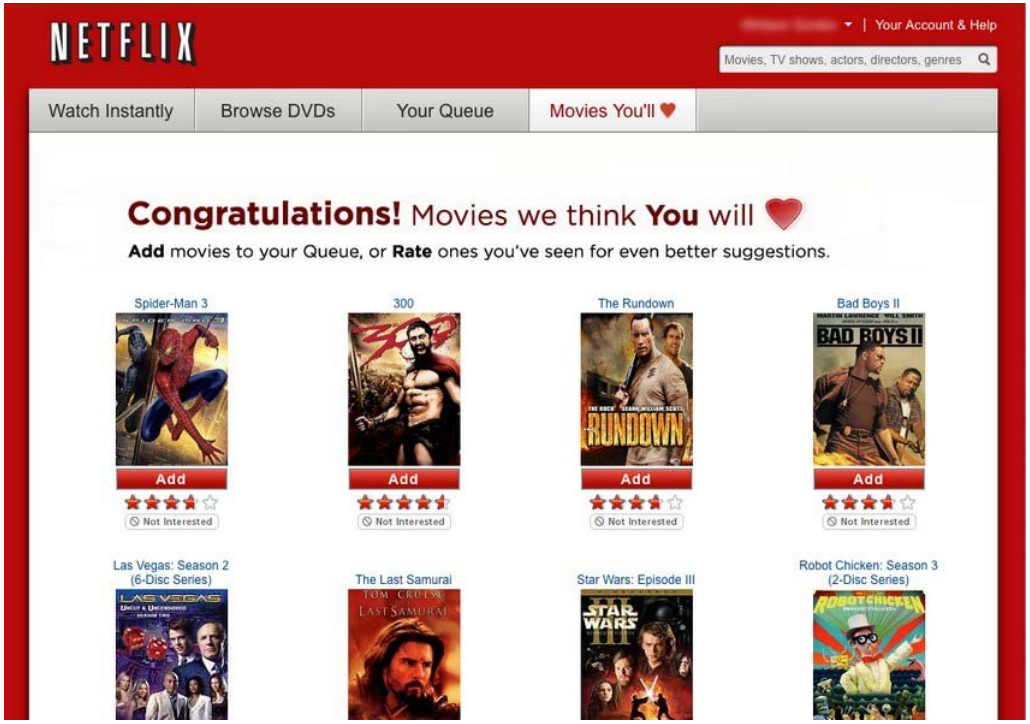
Ranking

- Google Search
  - given a query, find relevant webpages and rank in order



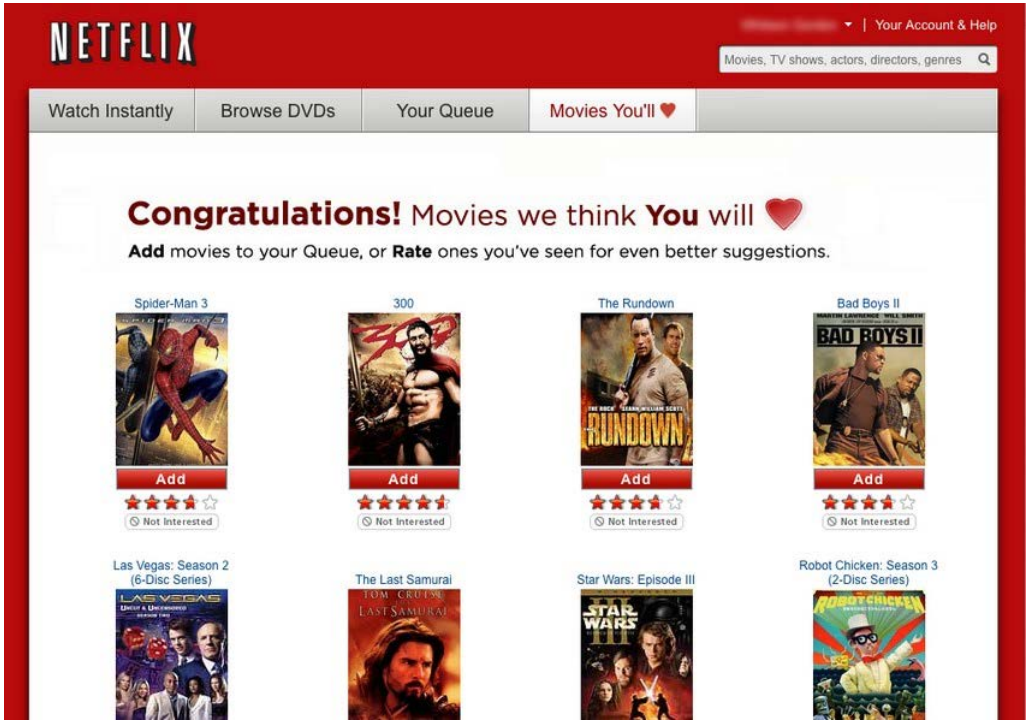
Recommendation

- Netflix movie recommendation: predict users' ratings
  - given users' information, watching history, etc, recommend most-likely to watch movies
  - Netflix Prize: $1,000,000, 2006 – 2009



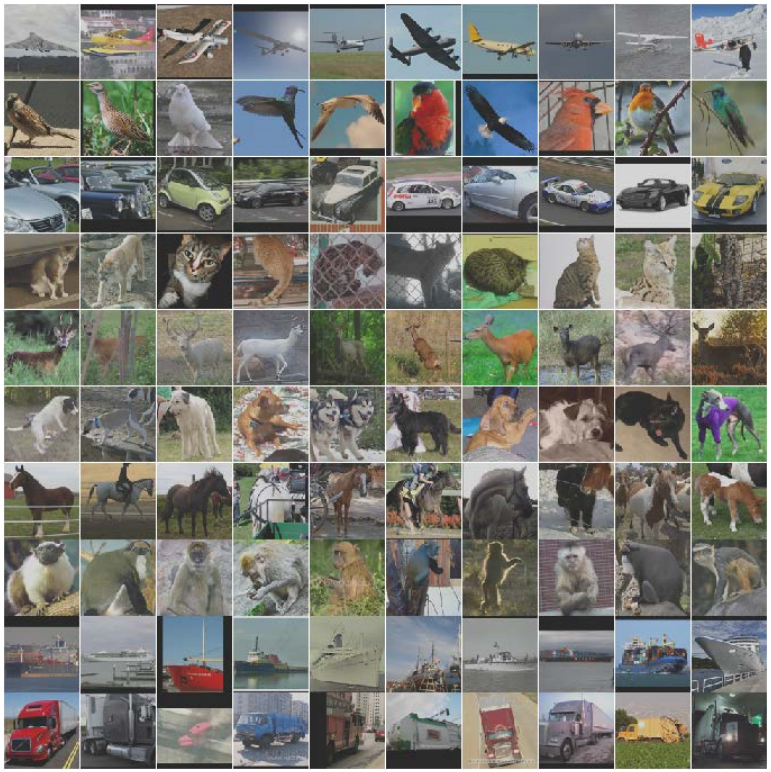
Recommendation

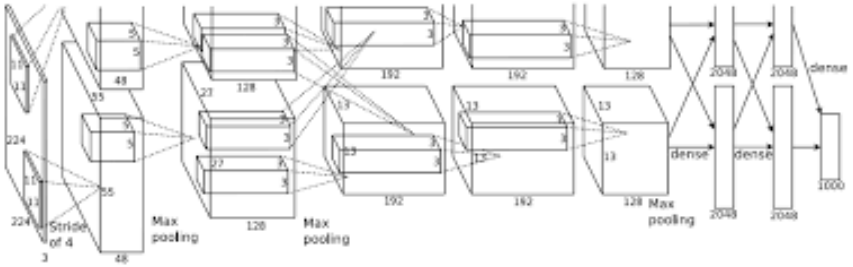
- Netflix movie recommendation: predict users' ratings
  - given users' information, watching history, etc., recommend most-likely to watch movies
  - Netflix Prize: $1,000,000, 2006 – 2009

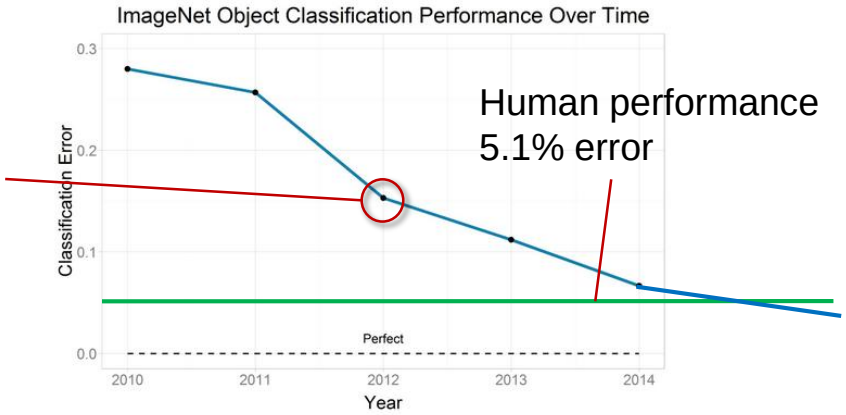


Milestones: Image Classification

Convolutional NNs: AlexNet (2012): trained on 200 GB of ImageNet Data

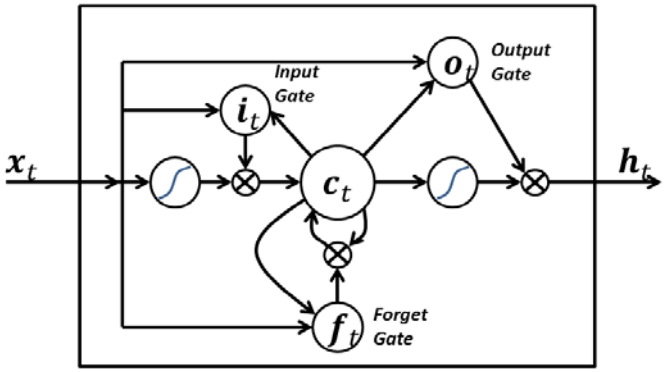


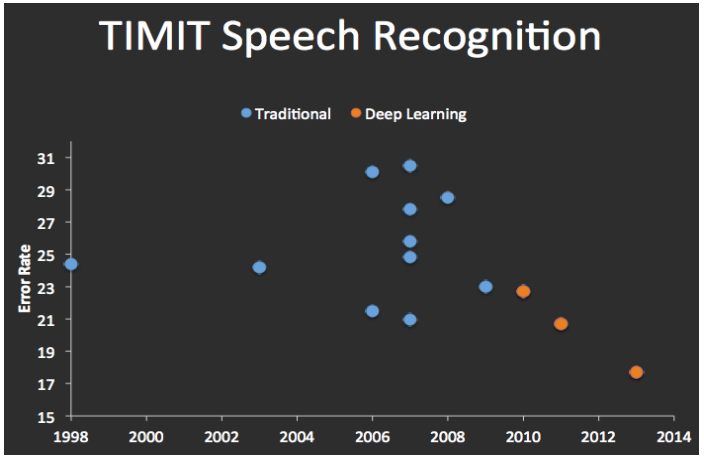


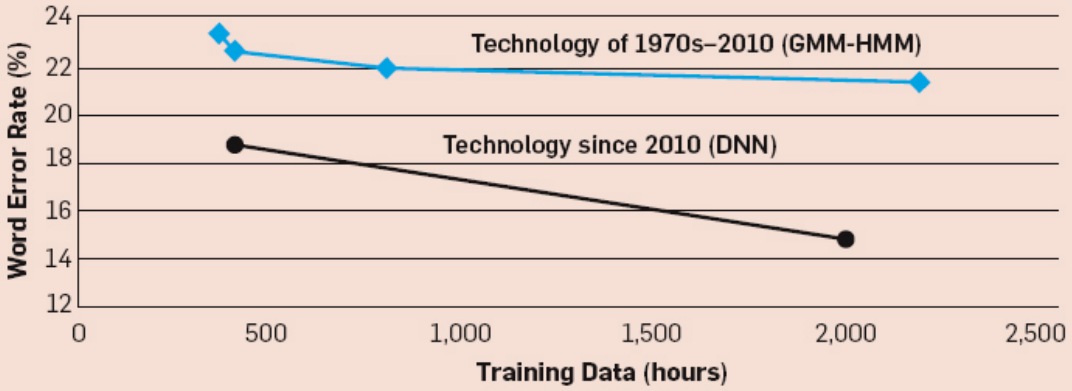


Milestones: Speech Recognition

Recurrent Nets: LSTMs (1997):





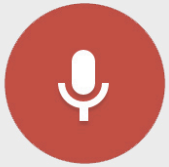


Speech Recognition

- SIRI

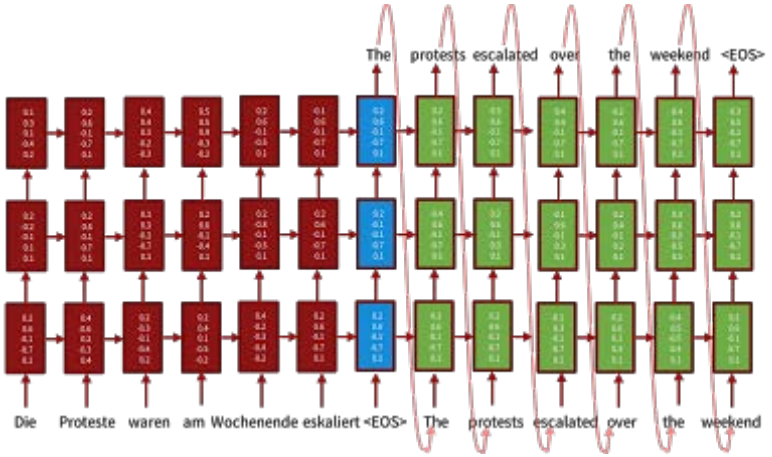
- Echo

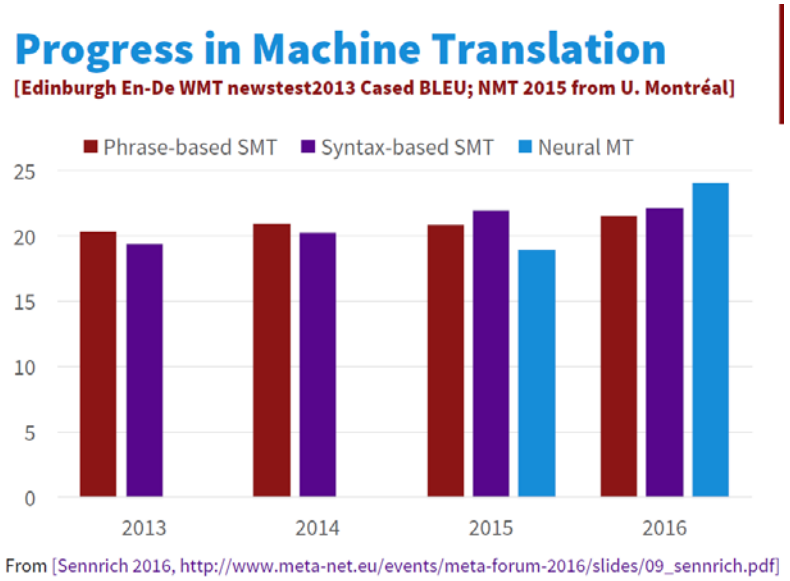
- OK Google



Milestones: Natural Language Processing

Sequence-to-sequence models with LSTMs and attention:





From [Sennrich 2016, http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf]

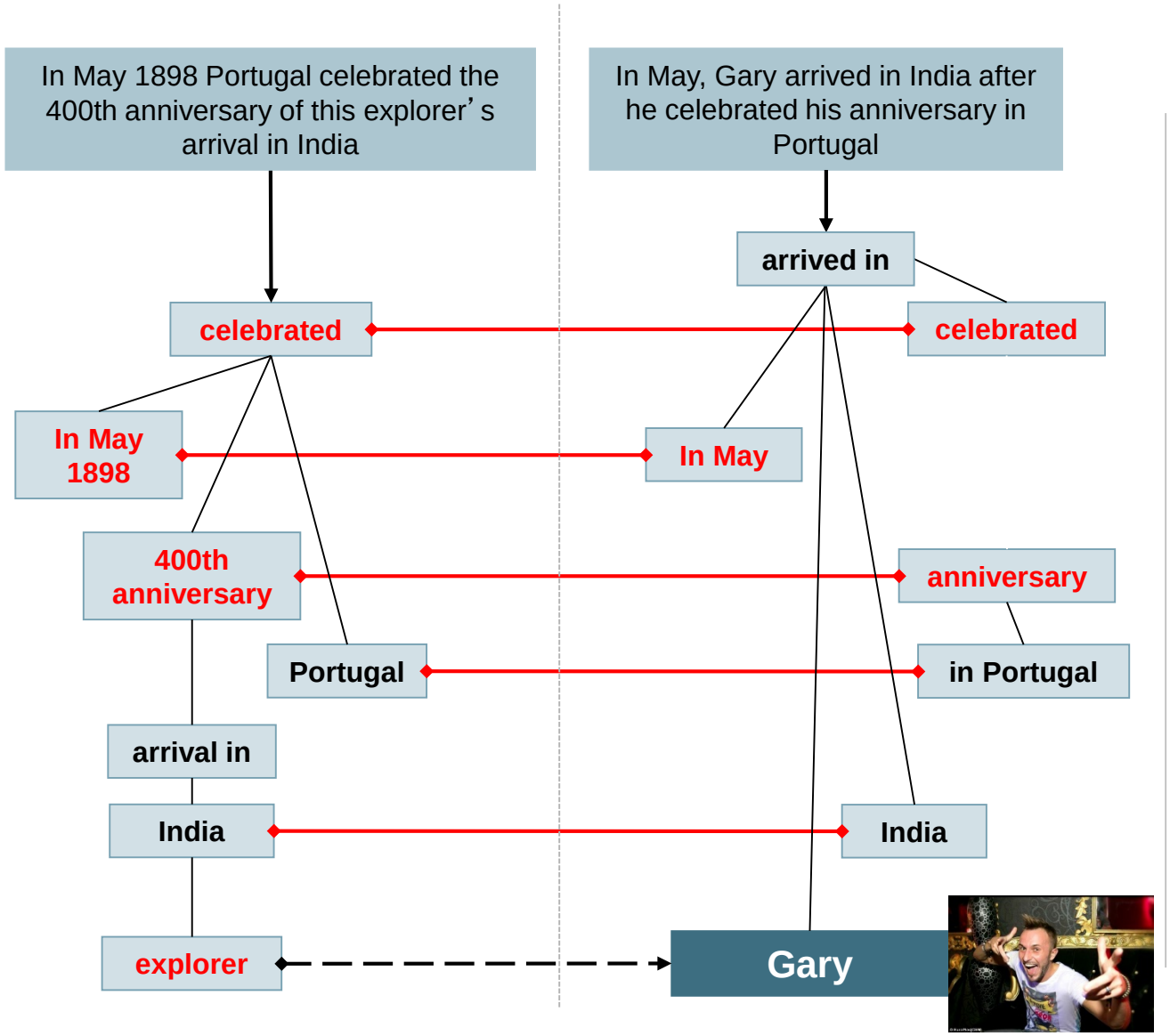
Source Luong, Cho, Manning ACL Tutorial 2016.

Watson won against two all-time Jeopardy champions

February 14-16, 2011

IBM

Answering complex questions requires more than keyword evidence



Legend

- Keyword "Hit"
- Reference Text
- Answer
- Red Text — Weak evidence

This evidence suggests "Gary" is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence

IBM Research Project (2006 –)

Jeopardy! Grand Challenge (Feb 2011)

Watson for Healthcare (Aug 2011 –)

Watson for Financial Services (Mar 2012 –)

Watson Ecosysteem (2014–)

R&D

Demonstration

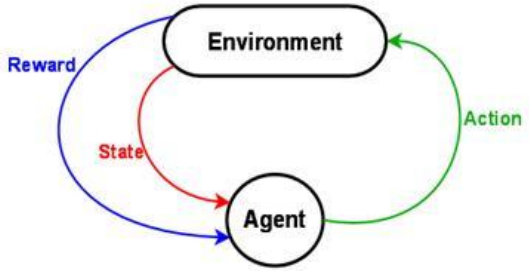
Commercialization

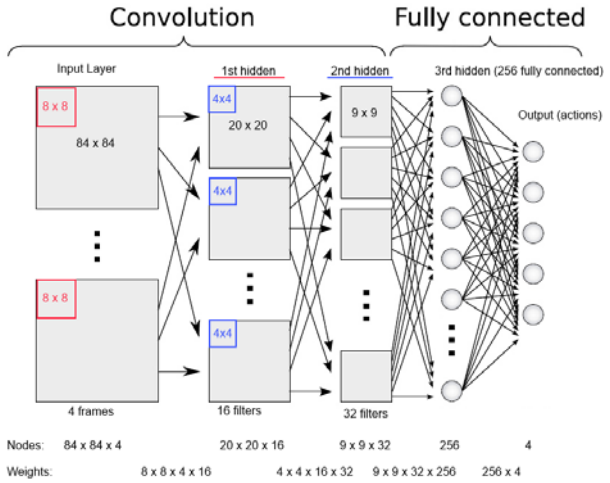
Expansion

Cross-industry Applications

Milestones: Deep Reinforcement Learning

In 2013, Deep Mind's arcade player bests human expert on six Atari Games. Acquired by Google in 2014,.





In 2016-17, Google's alphaGo defeats former world champion Lee Sedol, and all world's top Go players.

January 27, 2016

Nature **550**, 354–359 (19 October 2017)

Mastering the game of Go without human knowledge

After three days of training using 4 TPU (Tensor Processing Unit), a machine can beat any human expert on the Go game.
https://deepmind.com/blog/alphago-zero-learning-scratch/

ALPHAGO ZERO CHEAT SHEET

The training pipeline for AlphaGo Zero consists of three stages, executed in parallel

SELF PLAY

Create a 'training set'

The best current player plays 25,000 games against itself

See MCTS section to understand how AlphaGo Zero selects each move

At each move, the following information is stored

The game state
(see 'What is a Game State section')

π

The search probabilities
(From the MCTS)

The winner
(+1 if this player won, -1 if this player lost - added once the game has finished)

RETRAIN NETWORK

Optimise the network weights

A TRAINING LOOP

Sample a mini-batch of 2048 positions from the last 500,000 games

Retrain the current neural network on these positions
– The game states are the input (see 'Deep Neural Network Architecture')

Loss function
Compares predictions from the neural network with the search probabilities and actual winner

PREDICTIONS **p** **v** — Cross-entropy + Mean-squared error + Regularisation — π  ACTUAL

After every 1,000 training loops, evaluate the network

EVALUATE NETWORK

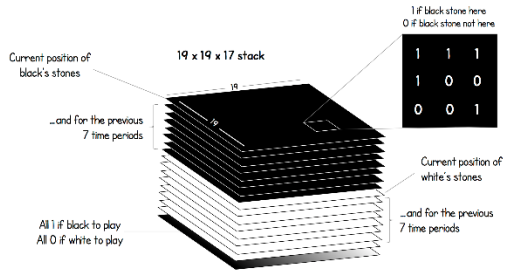
Test to see if the new network is stronger

Play 400 games between the latest neural network and the current best neural network

Both players use MCTS to select their moves, with their respective neural networks to evaluate leaf nodes

Latest player must win 55% of games to be declared the new best player

WHAT IS A 'GAME STATE'



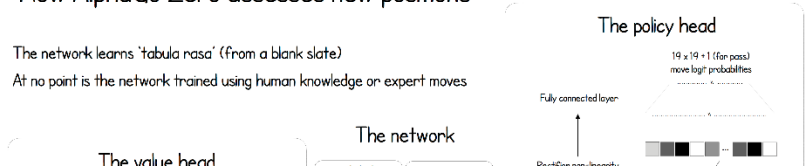
This stack is the input to the deep neural network

THE DEEP NEURAL NETWORK ARCHITECTURE

How AlphaGo Zero assesses new positions

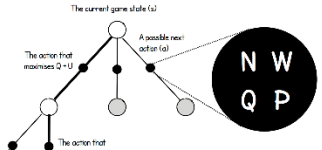
The network learns 'tabula rasa' (from a blank slate)

At no point is the network trained using human knowledge or expert moves



MONTE CARLO TREE SEARCH (MCTS)

How AlphaGo Zero chooses its next move



Each potential action from a game state stores four numbers:

N The number of times action a has been taken from state s

W The total value of the next state

Q The mean value of the next state

P The prior probability of selecting

First, run the following simulation 1,600 times...

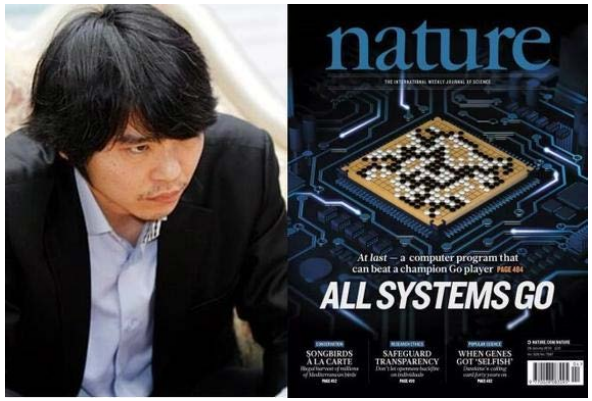
Start at the root node of the tree (the current game state)

1. Choose the action that maximises...

$Q + U$

Why is AlphaGo Significant for AI?

Three Major Milestones for AI

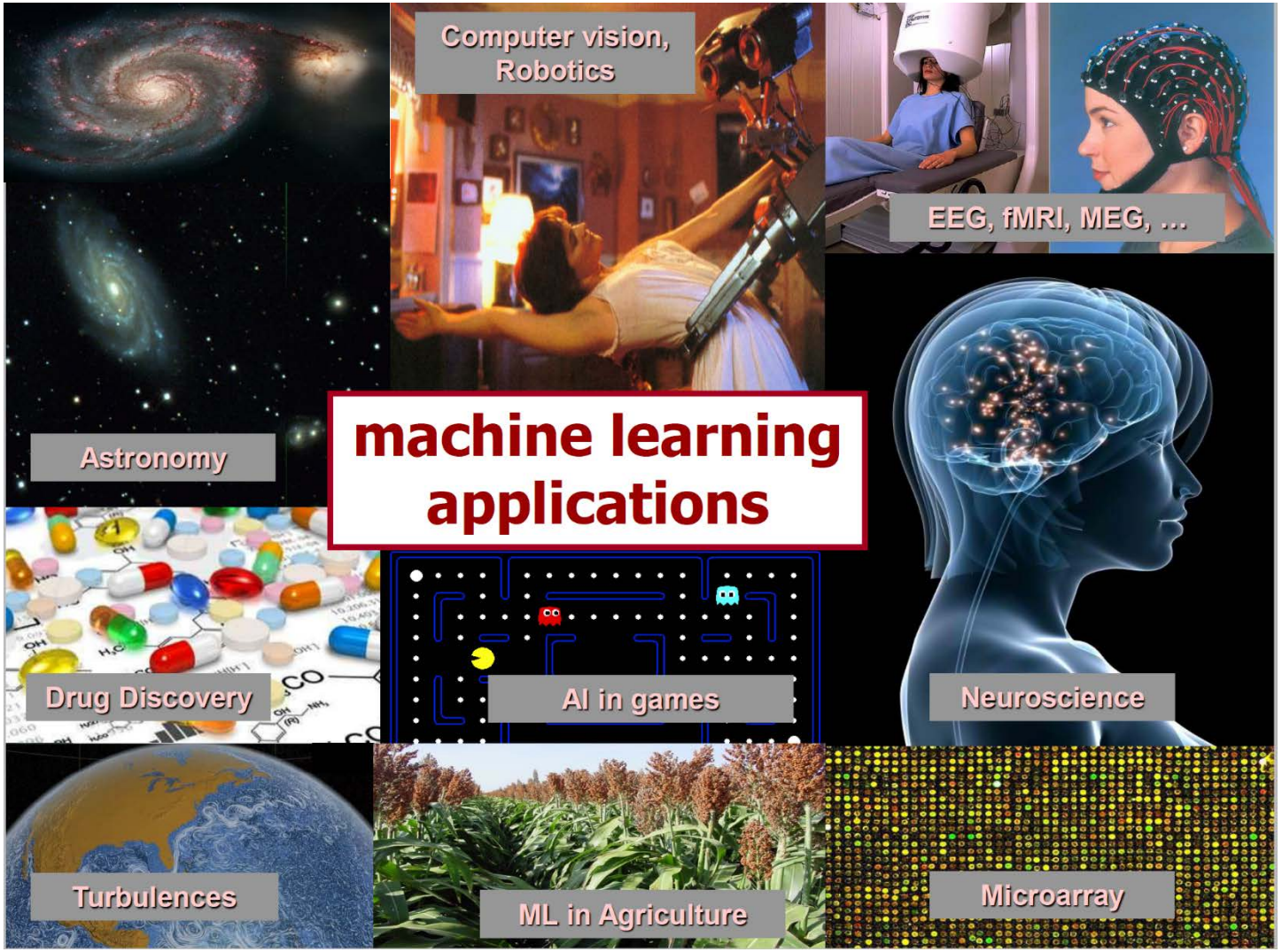


Chess: IBM Deep Blue — 1997

IBM WATSON — 2011

Google AlphaGo — 2016

More



picture courtesy : Barnabas Poczos

Deep Learning: Hype or Hope?

Hype: (n) “extravagant or intensive publicity or promotion”

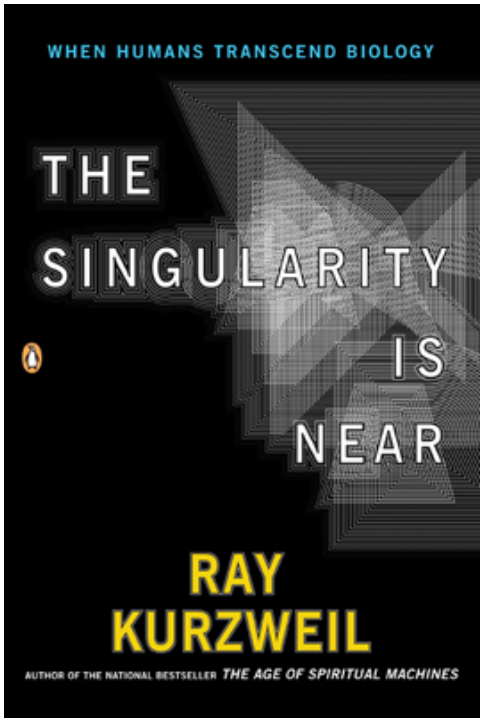
Hope: (n) “expectation of fulfillment or success”

Hype: The Singularity Is Near

A 2005 top-seller book by inventor and futurist Ray Kurzweil.

Central thesis

- Kurzweil predicts an exponential increase in technologies, like computers, nanotechnology genetics, robotics and artificial intelligence.
- The Singularity is the point when machine intelligence will be infinitely more powerful than all human intelligence combined.
- Kurzweil predicts, by the early 2030s the amount of non-biological computation will exceed the "capacity of all living biological human intelligence".
- "I set the date for the Singularity—representing a profound and disruptive transformation in human capability—as 2045".

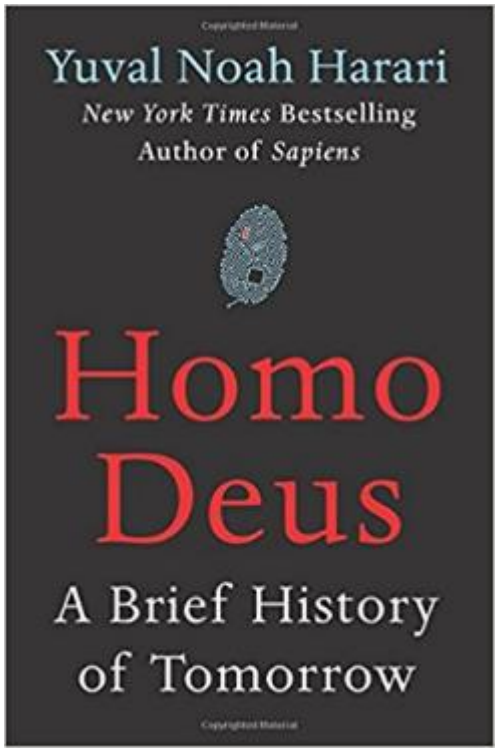


Homo Deus: A Brief History of Tomorrow

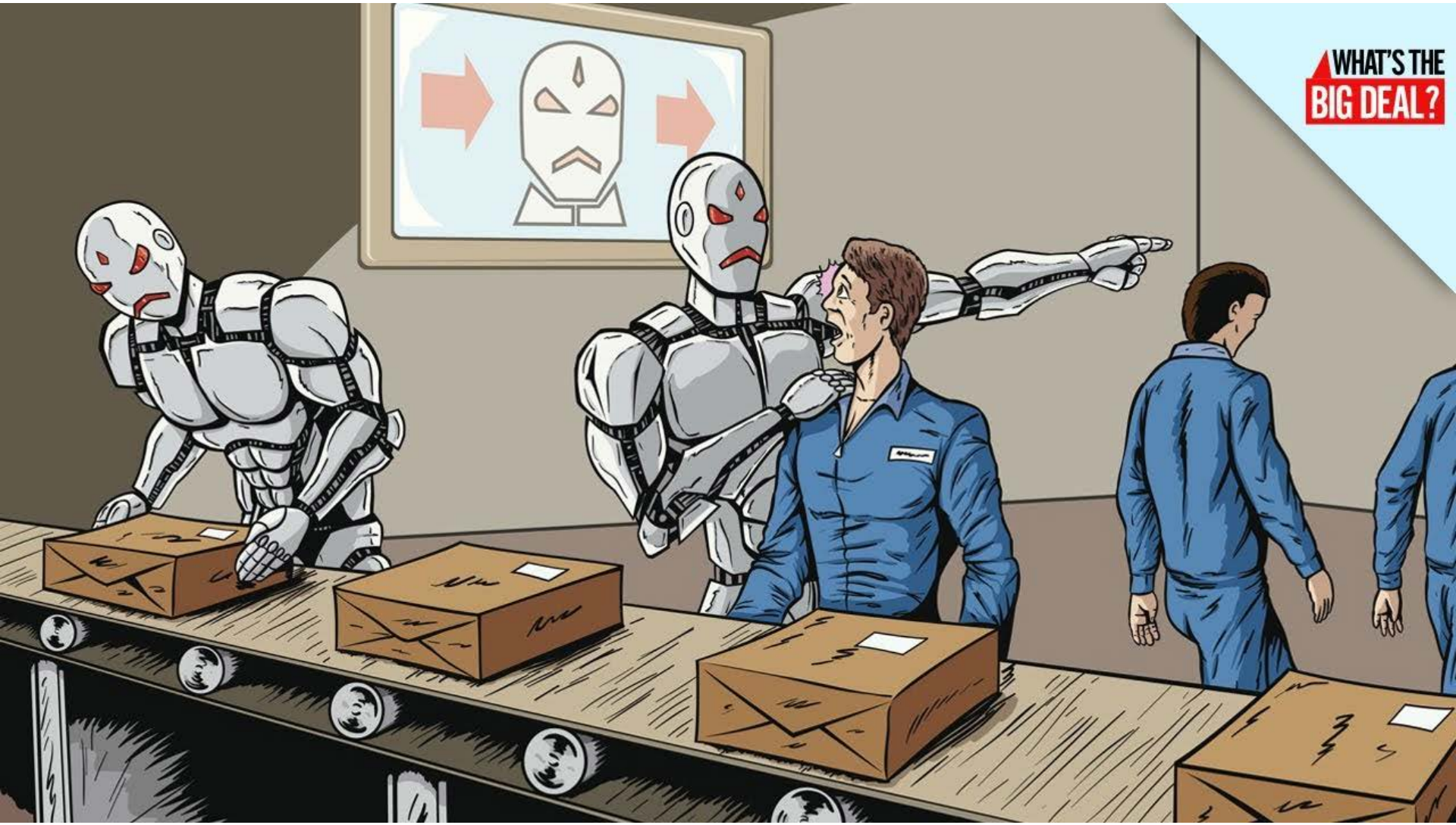
A 2016 top seller book by Historian Yuval Noah Harari

Central thesis:

- Organisms are algorithms, and as such homo sapiens (today's human) may not be dominant in a universe where dataism becomes the paradigm.
- Computers will do much better than organisms. Many professions will be out-of-date and labors become less worth.
- Harari believes that humanism may push humans to search for immortality, happiness, and power.
- Harari suggests the possibility of the replacement of humankind with a super-man, i.e. "homo deus", endowed with abilities such as eternal life.



Will AI Steal Our Jobs?



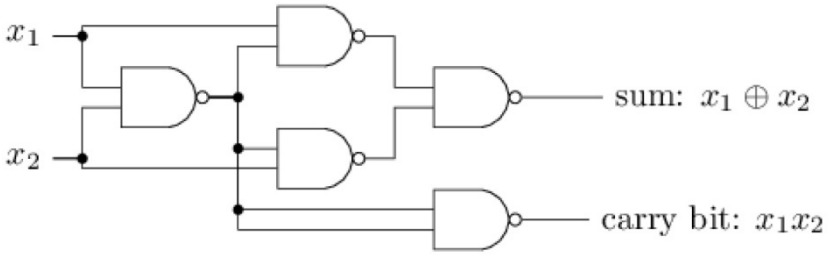
Why the success of DNNs is surprising

- The most successful DNN training algorithm is a version of gradient descent which will only find local optima. In other words, it's a greedy algorithm.
- Greedy algorithms are even more limited in what they can represent and how well they learn.
- If a problem has a greedy solution, it's regarded as an "easy" problem.

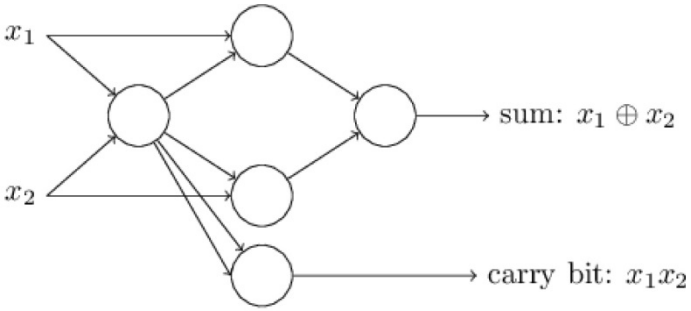
Why the success of DNNs is surprising

- In graphical models, values in a network represent random variables, and have a clear meaning. The network structure encodes dependency information, i.e. you can represent rich models.

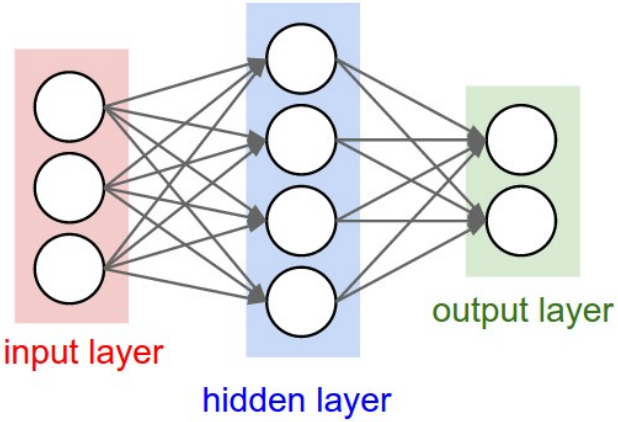
(1)



(2)



(3)

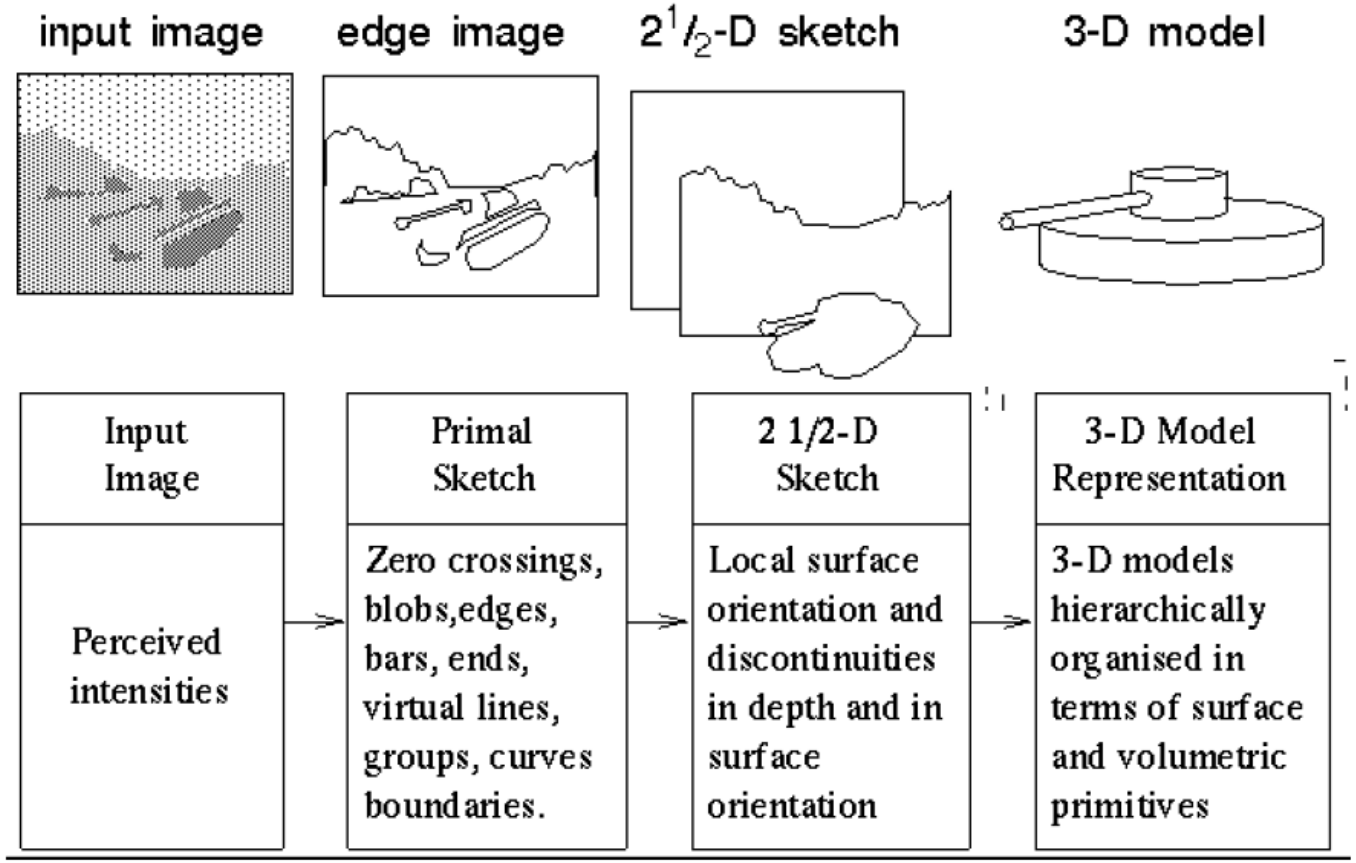


1. A half-adder
2. NN with the same structure
3. Actual NN

- In a NN, node activations encode nothing in particular, and the network structure only encodes (trivially) how they derive from each other.

Why the success of DNNs is ~~surprising~~ obvious

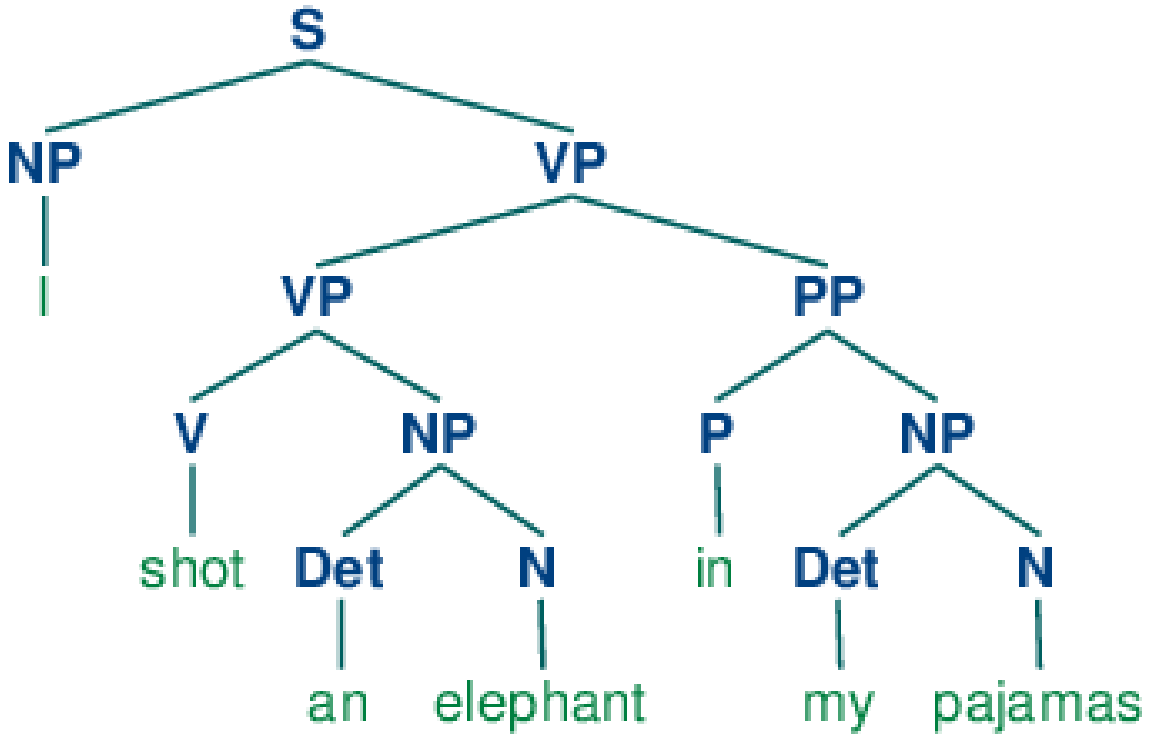
- Hierarchical representations are ubiquitous in AI. Computer vision:



Stages of Visual Representation, David Marr,

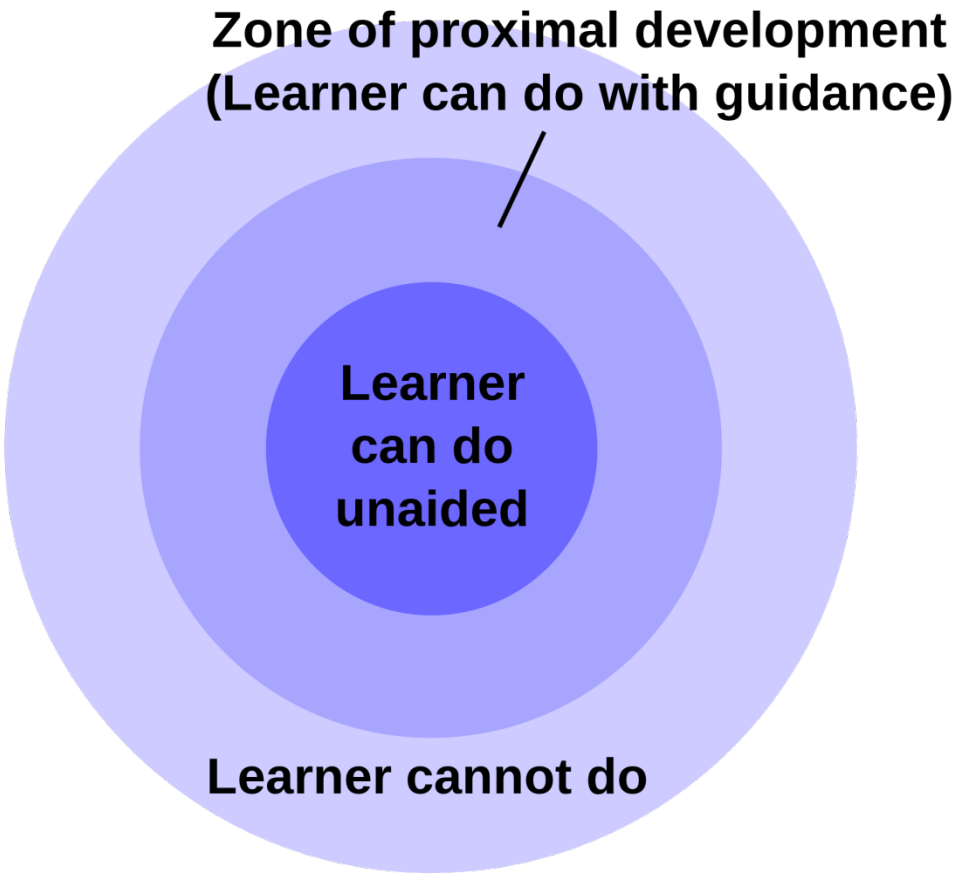
Why the success of DNNs is ~~surprising~~ obvious

- Natural language:



Why the success of DNNs is ~~surprising~~ obvious

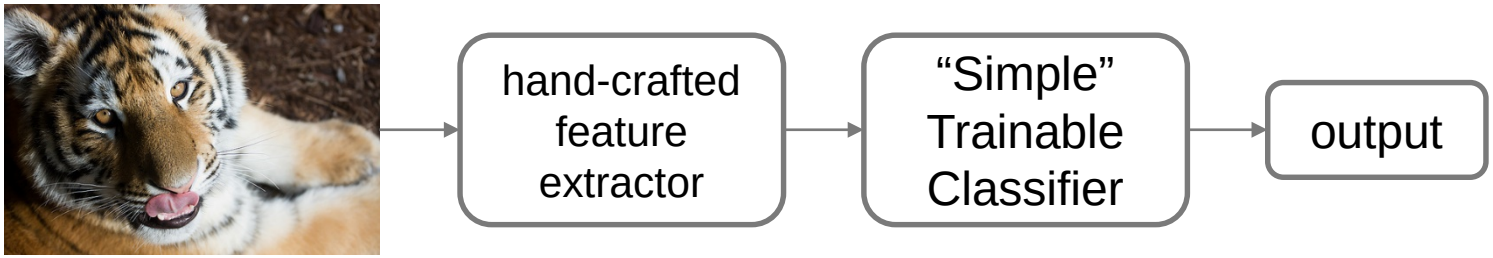
- Human Learning: is deeply layered.



Deep expertise

Pattern Recognition: Traditional Approach

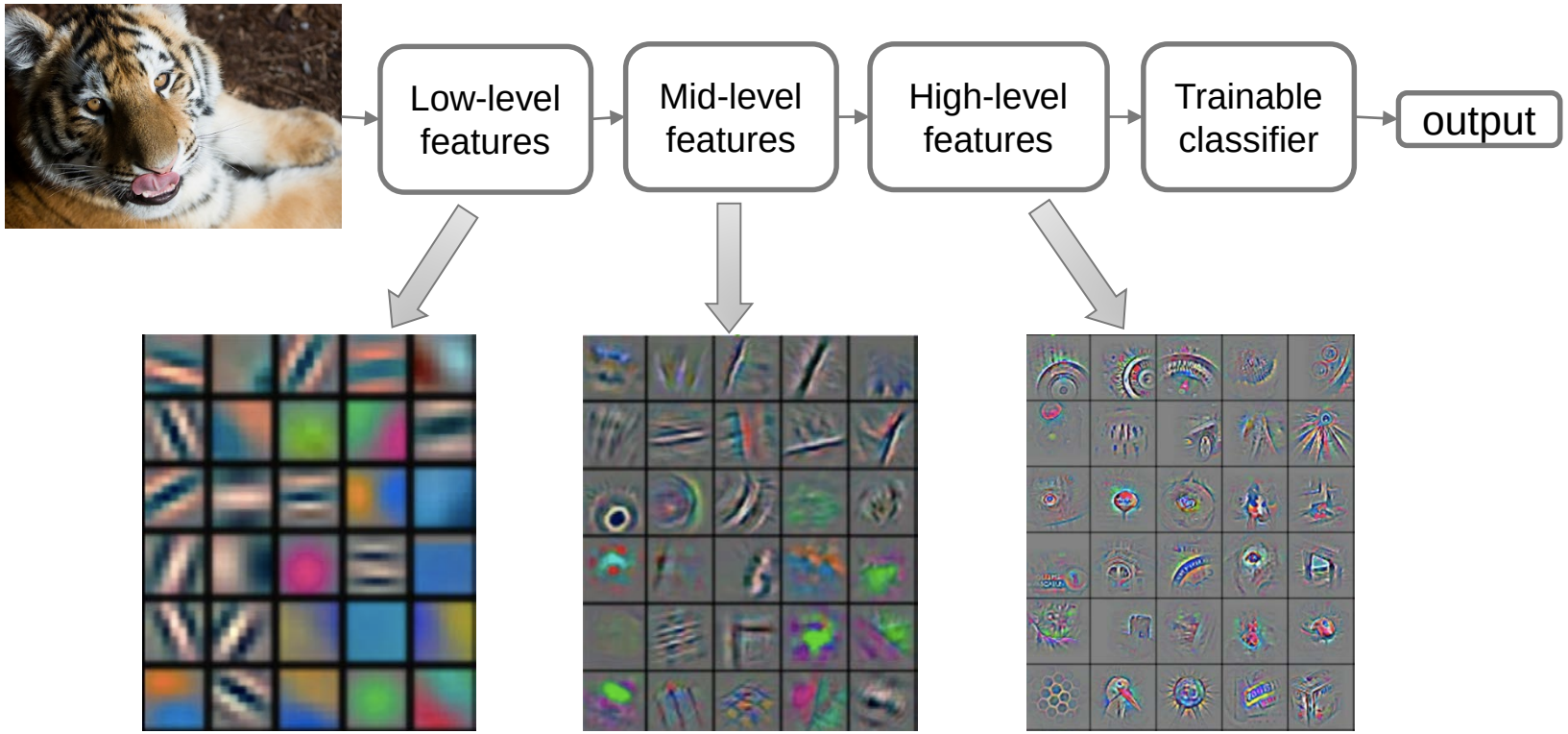
- Traditional pattern recognition models use hand-crafted features and relatively simple trainable classifier.



- This approach has the following limitations:
  - It is very tedious and costly to develop hand-crafted features
  - The hand-crafted features are usually highly dependents on one application, and cannot be transferred easily to other applications

Pattern Recognition: Deep Learning

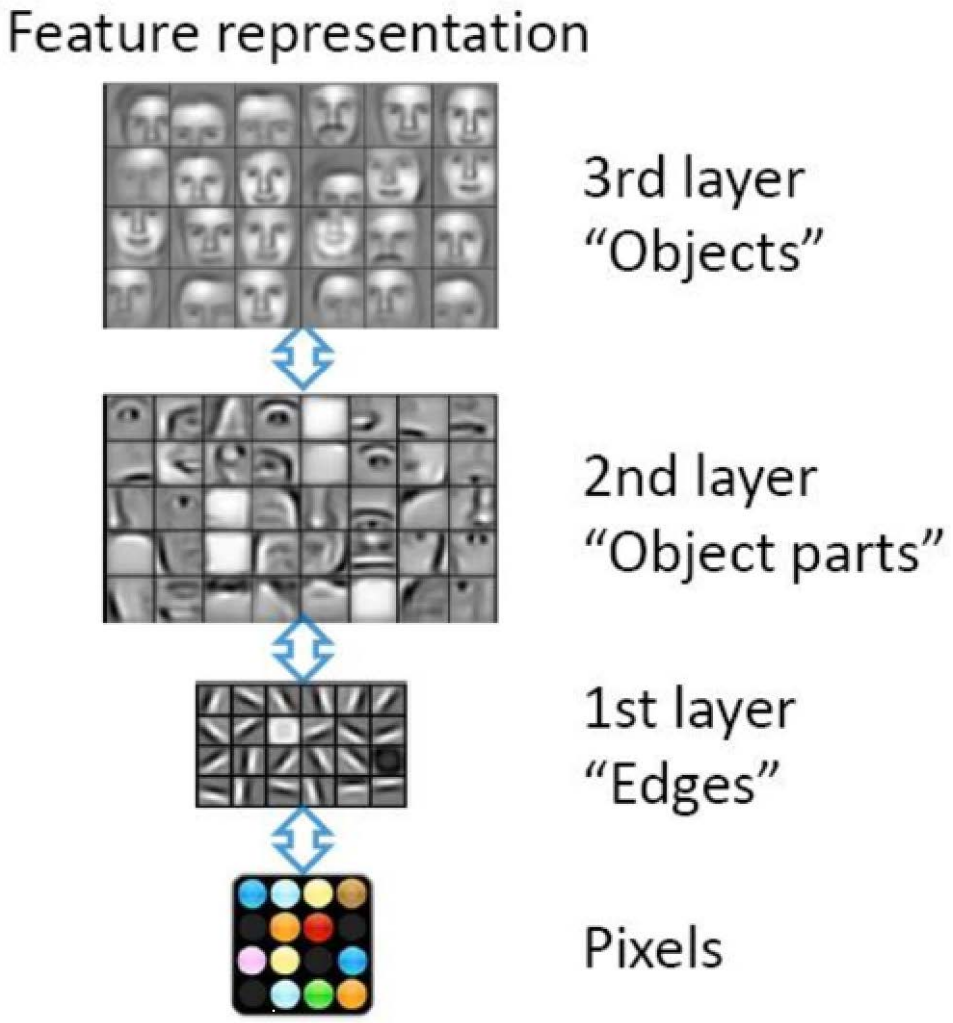
- Deep learning (= deep structure representation learning) seeks to learn rich hierarchical representations (i.e. features) automatically through multiple stage of feature learning process.



Feature visualization of convolutional net trained on ImageNet (Zeiler and Fergus, 2013)

Different Levels of Abstraction

- We don't know the "right" levels of abstraction
- So let the model figure it out!
- Deep Network can build up increasingly higher levels of abstraction



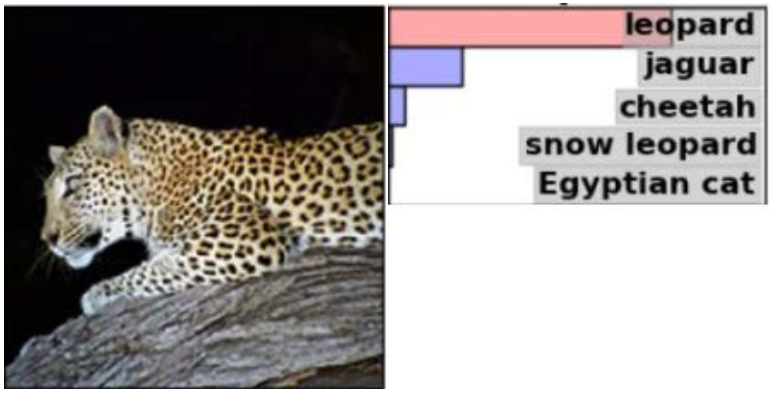
Example from Honglak Lee (NIPS 2010)

Pattern Recognition:

What has changes in 20 years?

- **In 1996:**
  - Small images (*e.g.*, 10x10)
  - Few classes (< 100)
  - Small network (< 4 layers)
  - Small data (< 50K images)

- **In 2016:**
  - Large images (256x256)
  - Many classes (> 1K)
  - Deep net (> 100 kerrosta)
  - Large data (> 1M)



Pattern Recognition:

Net Depth Evolution Since 2012

ILSVRC Image Recognition Task:

- 1.2 million images
- 1 000 categories

(Prior to 2012: error 25.7 %)

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

8 layers

16 layers

22 layers

152 layers
(> 1200 layers tested)

- 2015 winner: MSRA (error 3.57%)
- 2016 winner: Trimps-Soushen (2.99 %)

Learning about Deep Neural Networks

Yann Lecun quote: DNNs require: “an interplay between intuitive insights, theoretical modeling, practical implementations, empirical studies, and scientific analyses”

i.e. there isn’t a framework or core set of principles to explain everything (c.f. graphical models for machine learning).

We try to cover the ground in Lecun’s quote.

Deep Learning Libraries

- TensorFlow (Python, Google)
- Theano (Python, University of Montreal)
- Keras (Python wrapper, call TensorFlow and Theano)
- Caffe (C++, Cuda, Berkeley, good for research)
- Torch (FaceBook)
- Mxnet (Amazon)