INTRODUCTION TO DEEP LEARNING

Dr. Hilman F. Pardede

Research Center for Informatics
Indonesian Institute of Sciences

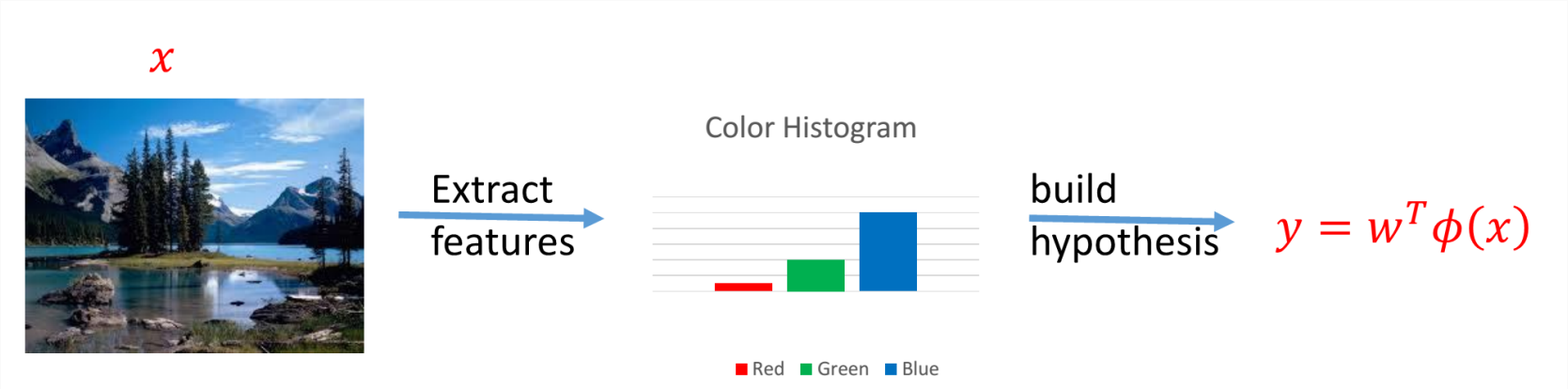
INTRO. TO MACHINE LEARNING

MACHINE LEARNING 1-2-3

- Collect data and extract features
- Build model: choose hypothesis class $\mathcal{H}$ and loss function l such as MMSE
- Optimization model by minimizing the empirical loss (such as gradient descent, maximum likelihood, etc)

BUILDING A GOOD ML METHODS

- ML methods depend heavily on the **representation** or **features** of the data given
- Many hand-designed features are customized for the designated data
- Finding such features are often not easy

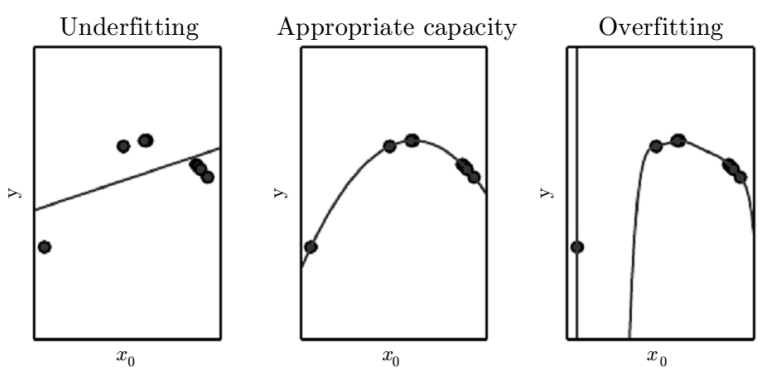


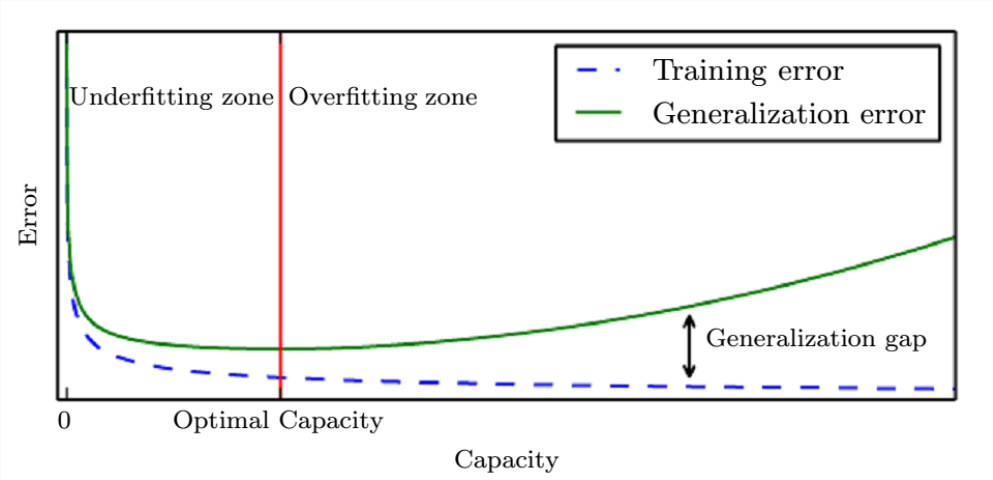
CAPACITY, OVERFITTING, AND UNDERFITTING

- The central challenge in machine learning is that we must perform well on new, previously unseen inputs – not just those on which our model was trained.
- The ability to perform well on previously unobserved inputs is called **generalization**. We want to make small training error but the gap between training error and testing error to be small

CAPACITY, OVERFITTING, AND UNDERFITTING

- Underfitting occurs when the model is not able to obtain a sufficiently low error value on the training set. Overfitting occurs when the gap between the training error and test error is too large.





THE NO FREE LUNCH THEOREM

- We only see limited set of examples. To logically infer a rule describing every member of a set, one must have information about every member of that set.
- No silver bullets: no machine learning algorithm is universally any better than any other.
- Our goal is to understand what kinds of distributions are relevant to the "real world" that an AI agent experiences, and what kinds of machine learning algorithms perform well on data drawn from the kinds of data generating distributions we care about.

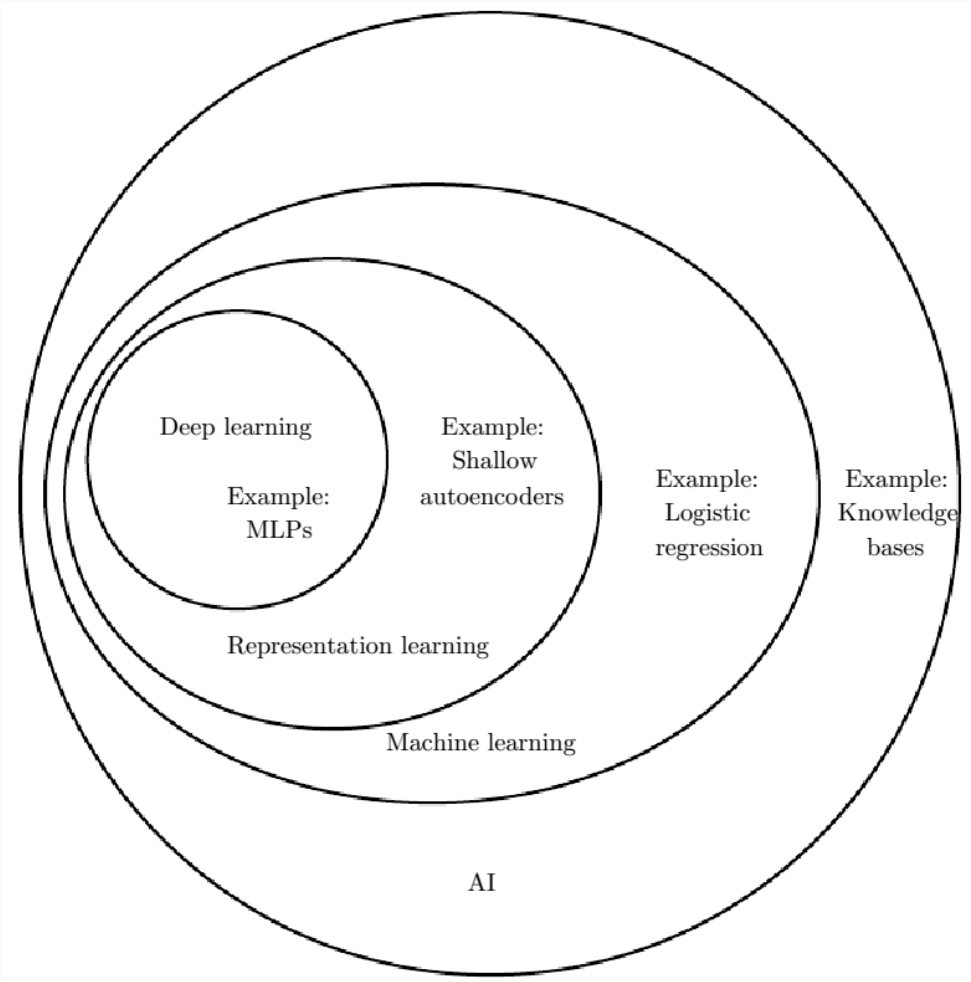
OVERCOMING OVERFITTING AND UNDERFITTING

- Regularization: Adding penalty to the cost function, allowing error to happen in some conditions
- Cross validation: Provide validation test and repeat the training several times with different validation set

INTRODUCTION TO DEEP LEARNING

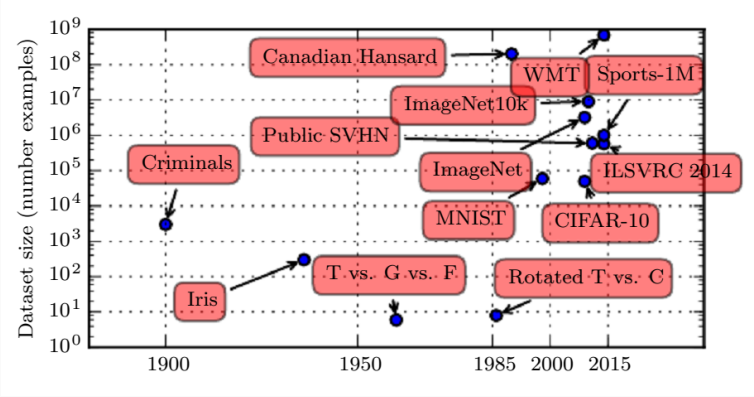
IN SHORT, WHAT IS DEEP LEARNING?

- A neural network with more hidden layer with **new algorithm for training many hidden layers**
- The human brain is a deep neural network, has many layers of neurons which acts as feature detectors, detecting more and more abstract features as you go up
- The idea of unsupervised learning for feature learning to find a better "intermediate features" (Experiments show DL seems to naturally learn them)



WHY IT IS SO POPULAR NOWADAYS

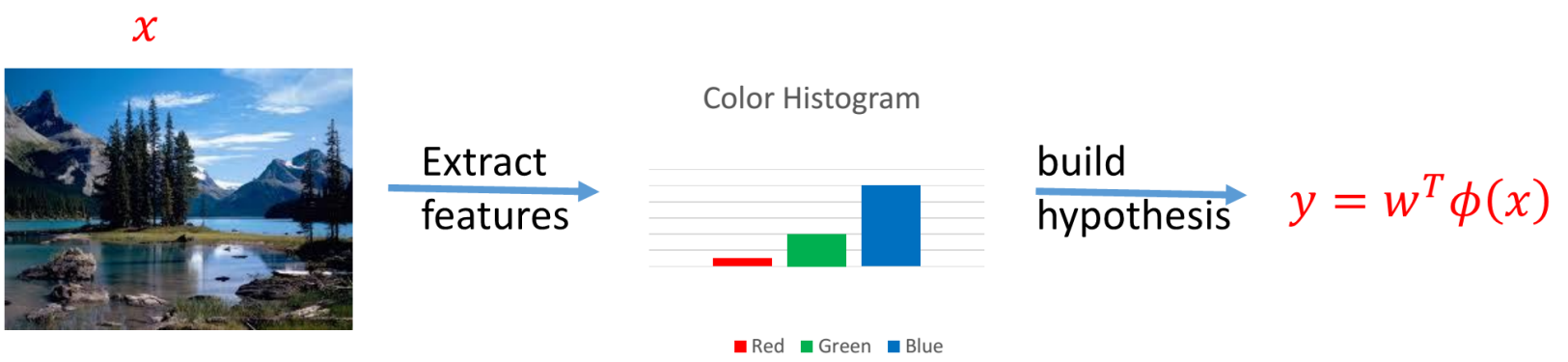
- It is effective
- Data are increasingly bigger (Big data), and deep learning provide a framework to use them easily (unsupervised learning)
- Computing power is increasing



modeling technique	#params [10^6]	WER Hub5'00-SWB	WER RT03S-FSH
GMM, 40 mix DT 309h SI	29.4	23.6	27.4
NN 1 hidden-layer×4634 units	43.6	26.0	29.4
+ 2×5 neighboring frames	45.1	22.4	25.7
DBN-DNN 7 hidden layers×2048 units	45.1	17.1	19.6
+ updated state alignment	45.1	16.4	18.6
+ sparsification	15.2 nz	16.1	18.5
GMM 72 mix DT 2000h SA	102.4	17.1	18.6

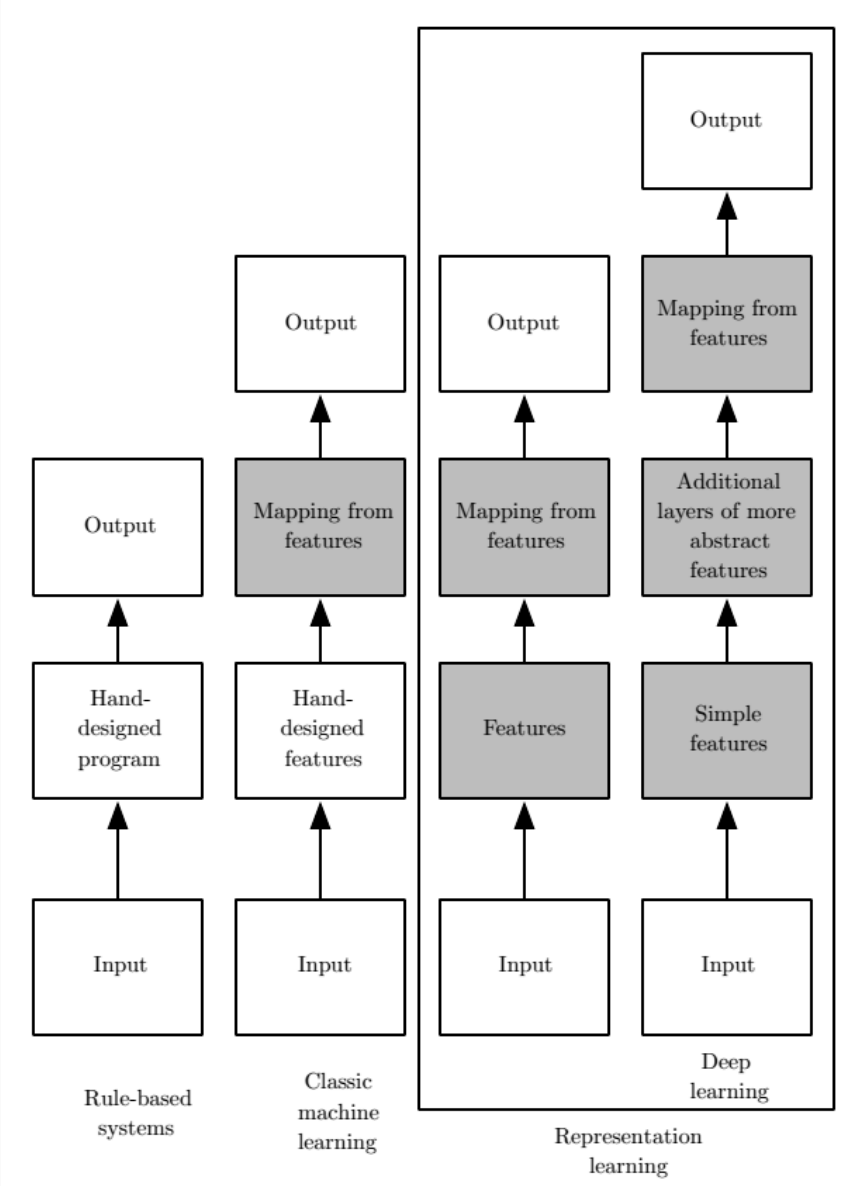
CONT'D

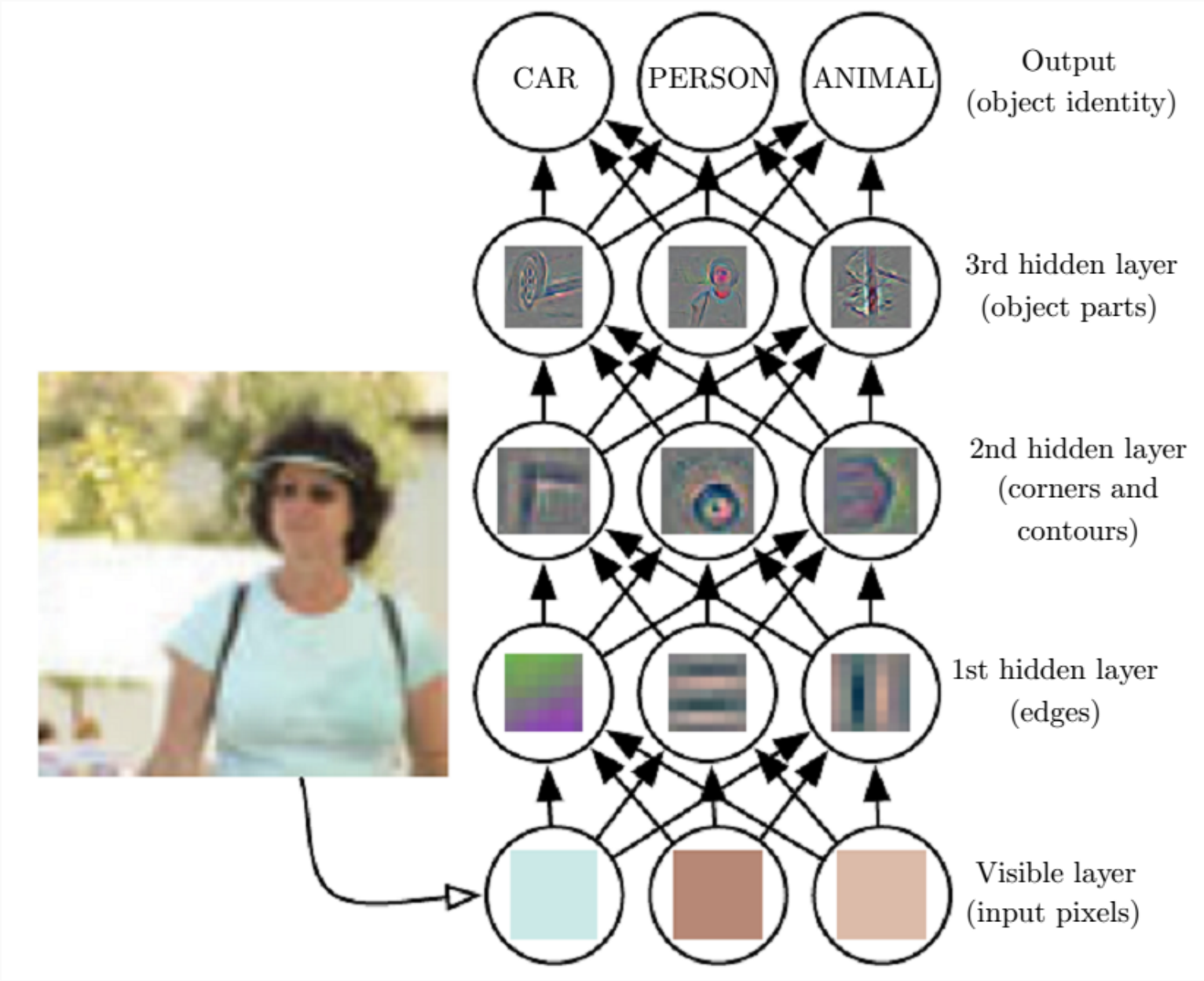
- One solution is to use ML to discover not only the mapping of features to output but also the features it self → **representation learning**
  - ex. autoencoder
- Real world data may be the results of many unobserved objects of forces in the physical world → abstraction or concepts that help us to make sense the variability of data
- Real world data such as image, video, speech, genetics information have many influencing factors that greatly increase the variation of the every single observation data



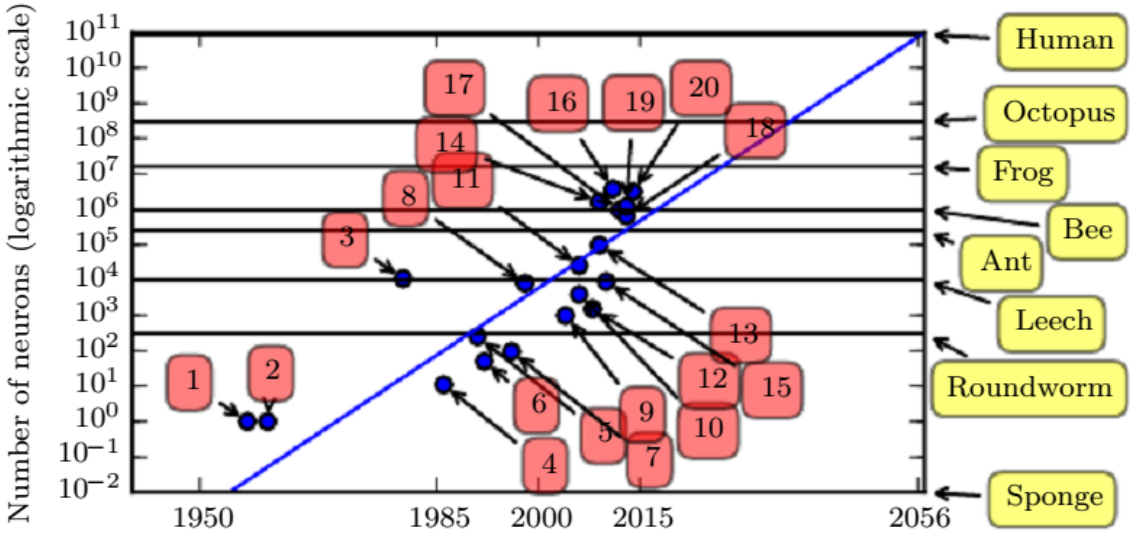
DEEP LEARNING: THE PHILOSOPHY

- Deep learning solves the representation learning by allowing the mapping of input (raw data) to other representations to capture the abstraction/the concept from the data using stacks of multilayer perceptrons.
- It allows the the computer to build complex concepts out of simpler concepts.
- It may be different on how human "see" the world, however it shows significant improvements over conventional ML methods




CAR
PERSON
ANIMAL
Output
(object identity)
3rd hidden layer
(object parts)
2nd hidden layer
(corners and
contours)
1st hidden layer
(edges)
Visible layer
(input pixels)

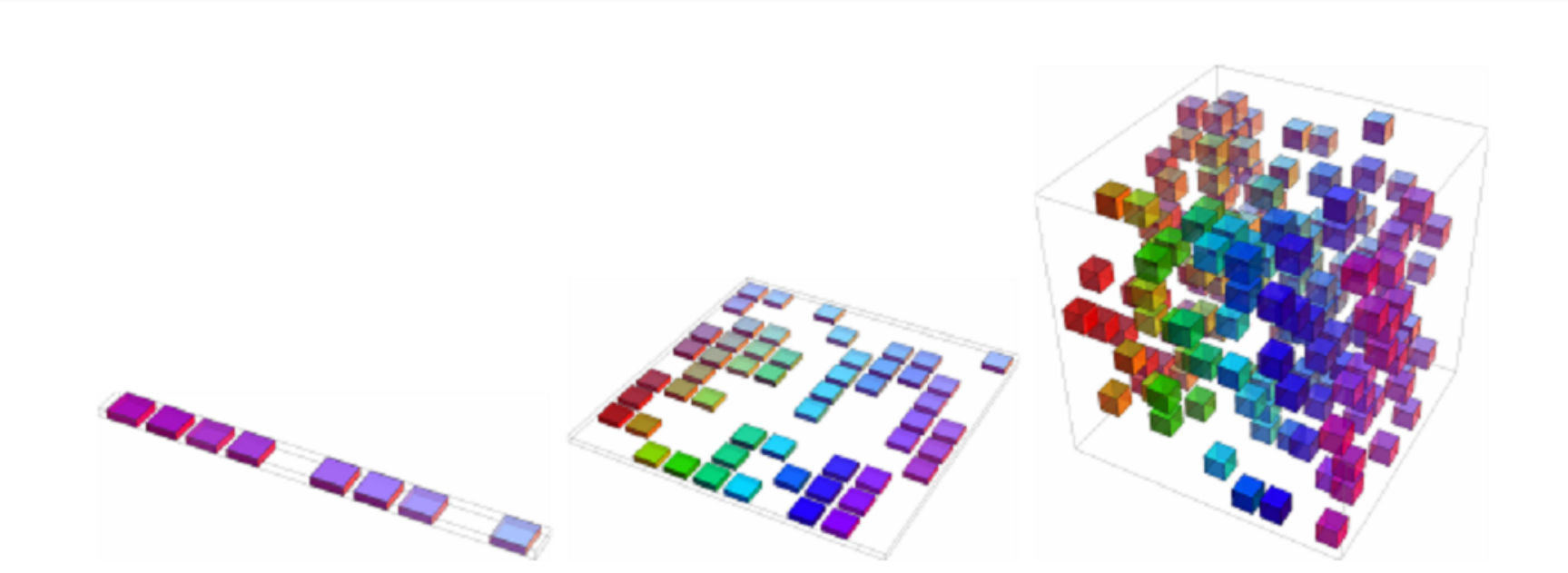
THE FUTURE OF DL



1. Perceptron (Rosenblatt, 1958, 1962)
2. Adaptive linear element (Widrow and Hoff, 1960)
3. Neocognitron (Fukushima, 1980)
4. Early back-propagation network (Rumelhart *et al.*, 1986b)
5. Recurrent neural network for speech recognition (Robinson and Fallside, 1991)
6. Multilayer perceptron for speech recognition (Bengio *et al.*, 1991)
7. Mean field sigmoid belief network (Saul *et al.*, 1996)
8. LeNet-5 (LeCun *et al.*, 1998b)
9. Echo state network (Jaeger and Haas, 2004)
10. Deep belief network (Hinton *et al.*, 2006)
11. GPU-accelerated convolutional network (Chellapilla *et al.*, 2006)
12. Deep Boltzmann machine (Salakhutdinov and Hinton, 2009a)
13. GPU-accelerated deep belief network (Raina *et al.*, 2009)
14. Unsupervised convolutional network (Jarrett *et al.*, 2009)
15. GPU-accelerated multilayer perceptron (Ciresan *et al.*, 2010)
16. OMP-1 network (Coates and Ng, 2011)
17. Distributed autoencoder (Le *et al.*, 2012)
18. Multi-GPU convolutional network (Krizhevsky *et al.*, 2012)
19. COTS HPC unsupervised convolutional network (Coates *et al.*, 2013)
20. GoogLeNet (Szegedy *et al.*, 2014a)

MOTIVATING DEEP LEARNING: THE CURSE OF DIMENSIONALITY

- Many machine learning problems become exceedingly difficult when the number of dimensions in the data is high.
- possible distinct configurations of a set of variables increases exponentially as the number of variables increases.



MOTIVATING DEEP LEARNING: LOCAL CONSTANCY AND SMOOTHNESS REGULARIZATION

- In order to generalize well, machine learning algorithms need to be guided by prior beliefs about what kind of function they should learn.
- In shallow architecture, we define the prior implicitly by choosing algorithms that are biased toward choosing some class of functions over another
- The smoothness prior or local constancy prior is widely used $\rightarrow$ the function we learn should not change very much within a small region.
- They fail to scale to the statistical challenges involved in solving AI level tasks.

MOTIVATING DEEP LEARNING: MANIFOLD LEARNING

- Manifold: a set of points, associated with a neighborhood around each point.
- Many machine learning problems seem hopeless if we expect the machine learning algorithm to learn functions with interesting variations across all dimensions since interesting inputs occur only along a collection of manifolds containing a small subset of points, with interesting variations in the output of the learned function occurring only along directions that lie on the manifold, or with interesting variations happening only when we move from one manifold to another.
- Data like images, sounds, or text have such characteristics