Chapter 1

Introduction

Machine Learning

- Acquiring knowledge by extracting *patterns* from *raw data*
- Example: To predict a person's wellness t from their MRI scan x by learning patterns from the medical records $\{x, t\}$ of some population

$$
\begin{tikzcd}
 & & & & & t \arrow[d] & \\
x \arrow[r] & \boxed{\text{Rep.}} \arrow[r] & \phi(x) \arrow[r] & \boxed{\text{Model}} \arrow[r] & y \arrow[r] & \text{-} \arrow[r] & \text{Cost}
\end{tikzcd}
$$

  - x: MRI scan
  - $\phi(x)$: data representation of MRI scan
  - $y \in (0, 1)$: model prediction with parameter w

$$y = f_w(\phi(x)) \triangleq \sigma(w^T \phi(x)), \text{ where } \sigma(x) = \frac{1}{1 + e^{-x}}$$

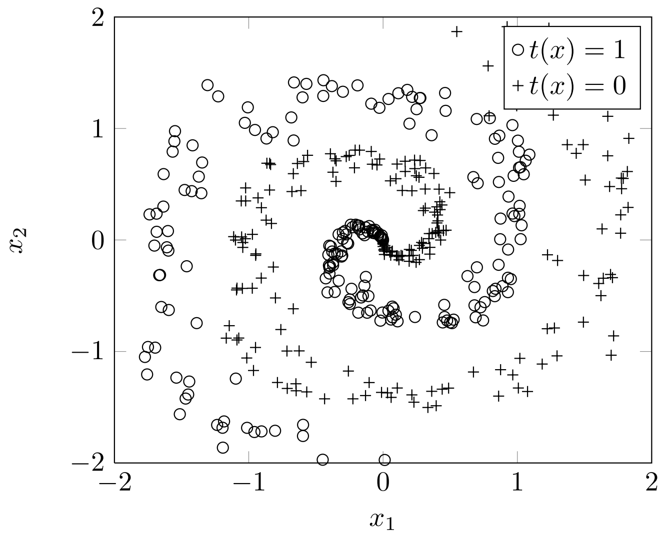
  - $t \in \{0, 1\}$: ground-truth result associated with input x

- Cost: some distance between y and t (e.g. $\|y - t\|_2^2$), which is to be minimized w.r.t. w over the $\{x, t\}$ pairs

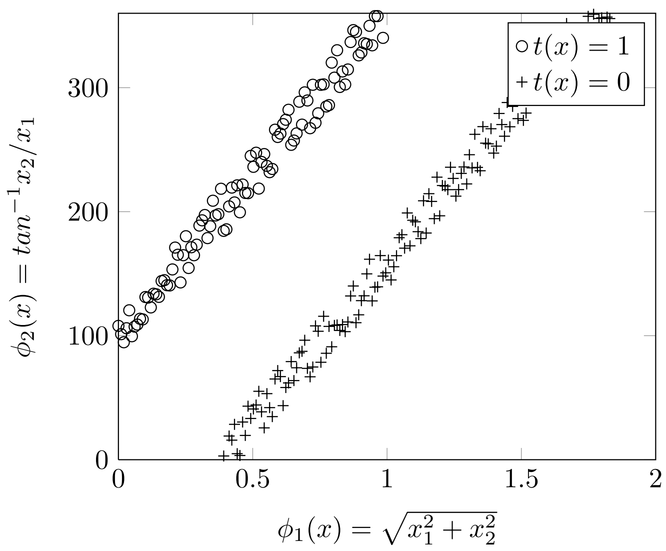
- Essentially, we want to find a function $f_w(\phi(x))$ to approximate $t(x)$
- In the present example, $f_w(\phi(x))$ bears a probabilistic interpretation of $p(t = 1|x; w)$
- The setting here is termed *supervised learning* as the ground-truth result t is given for each x

Data Representation, $\phi(x)$

- Data representation can critically determine the prediction performance



raw data domain



feature domain

- In classic machine learning, hand-designed features are usually used; for many tasks, it is however difficult to know what features should be used

Deep Learning

- A machine learning approach whose data representation is based on building up a *hierarchy of concepts*, with each concept defined through its relation to simpler concepts

- Using the previous example, this amounts to learning a function of the following form

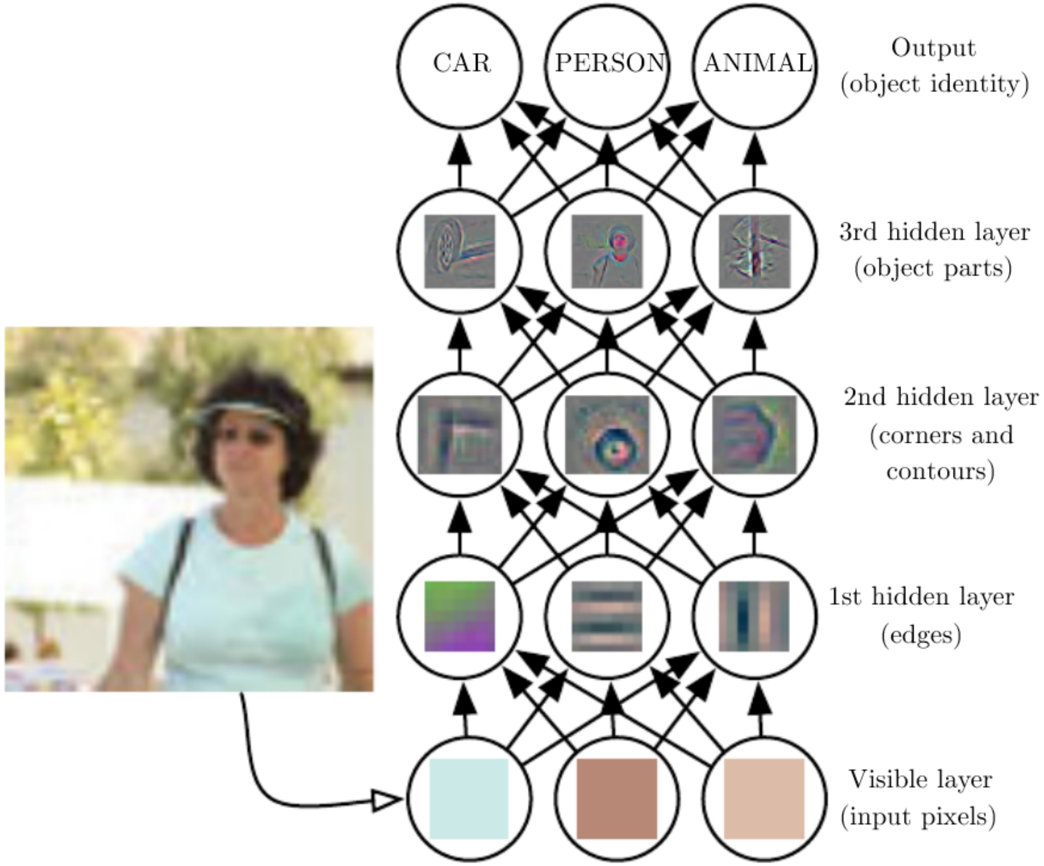
$$f_{w,\theta_n,\theta_{n-1},\cdots,\theta_1}(x) = \sigma(w^T \underbrace{\phi_{\theta_n}(\phi_{\theta_{n-1}}(\phi_{\theta_{n-2}}(\cdots\phi_{\theta_1}(x))))}_{\text{Hierarchy of concepts/features}})$$

  where $w, \theta_n, \theta_{n-1}, \cdots, \theta_1$ are model parameters

- $\phi_\theta(\cdot)$'s are generally vector-valued functions, e.g. $\phi_\theta(x) = \sigma(\theta x)$

- Such a deep model allows to construct a complicated function $f(x)$ from nested composition of simpler functions $\phi(\cdot)$'s

Example: Feedforward Deep Networks



$$\sigma(\phi_{\theta_3}(\phi_{\theta_2}(\phi_{\theta_1}(x))))$$

$$\uparrow$$

$$\phi_{\theta_3}(\phi_{\theta_2}(\phi_{\theta_1}(x)))$$

$$\uparrow$$

$$\phi_{\theta_2}(\phi_{\theta_1}(x))$$

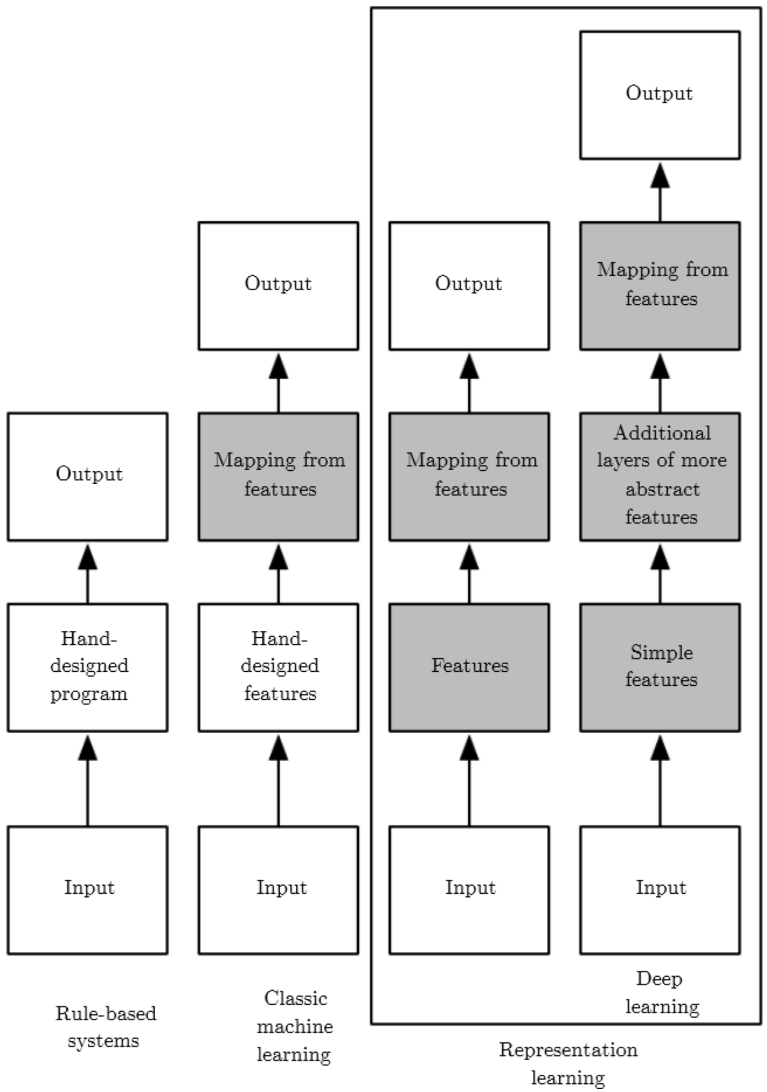
$$\uparrow$$

$$\phi_{\theta_1}(x)$$

$$\uparrow$$

$$x$$

Classic Machine Learning vs. Deep Learning

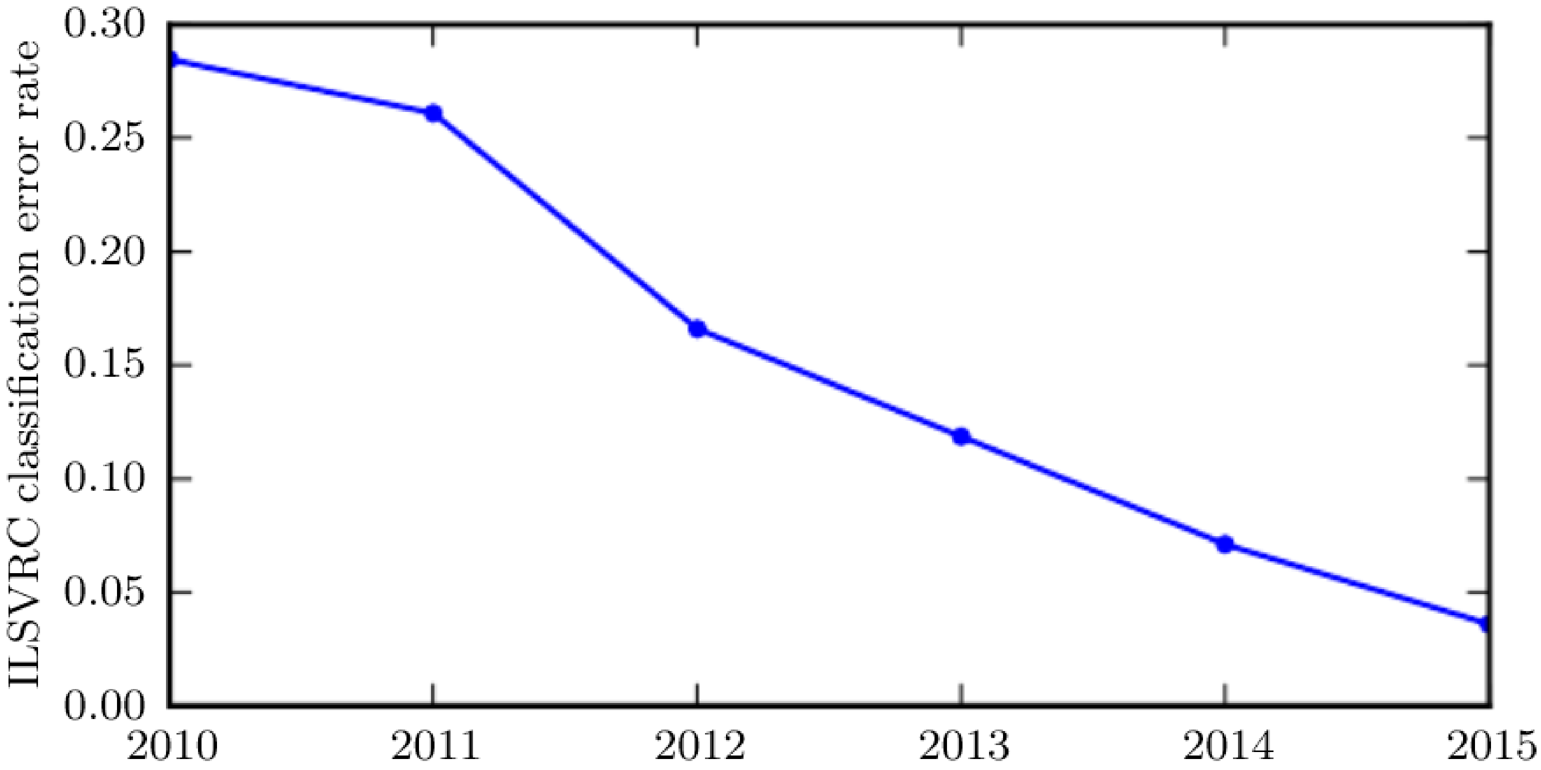


History of Deep Learning

- **Cybernetics** (1940s-1960s): Systems inspired by biological brains
  - Perceptron (Rosenblatt, 1958, 1960), Adaptive Linear Element, ADALINE (Widrow and Hoff, 1960)
- **Connectionism** (1980s-1990s): Connected simple computational units
  - Neocognition (Fukushima, 1980); Recurrent Neural Networks (Rumelhart et al., 1986); Convolutional Neural Networks (LeCun et al., 1998); Long Short-Term Memory (Hochreiter and Schmidhuber, 1997)
- **Deep Learning** (2006s-): Deeper networks and deep generative models
  - Deep Belief Networks (Hinton et al., 2006); Deep Boltzmann Machine (Salakhutdinov et al, 2009); Neural Turing Machine (Graves et al., 2014); Variational Autoencoder (Kingma et al.,2014); Generative Adversarial Networks (Goodfellow et al. 2014)

Recent Milestone

ImageNet Large Scale Visual Recognition Challenge



Top-5 error rate reduced from 26% to be less than 5%