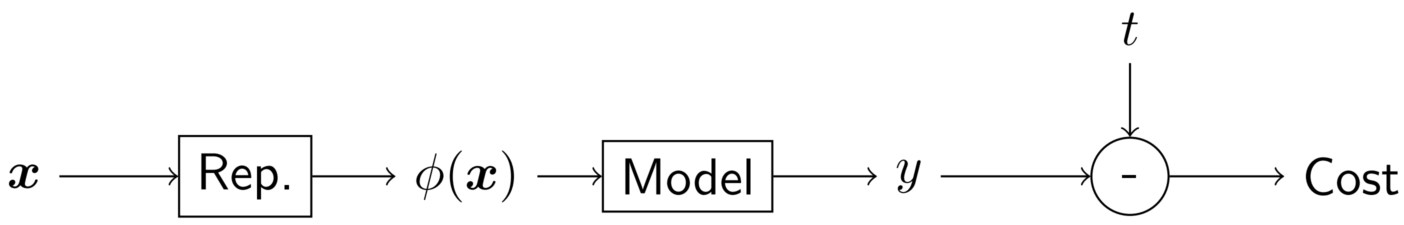
Chapter 1

Introduction

Machine Learning

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



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

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

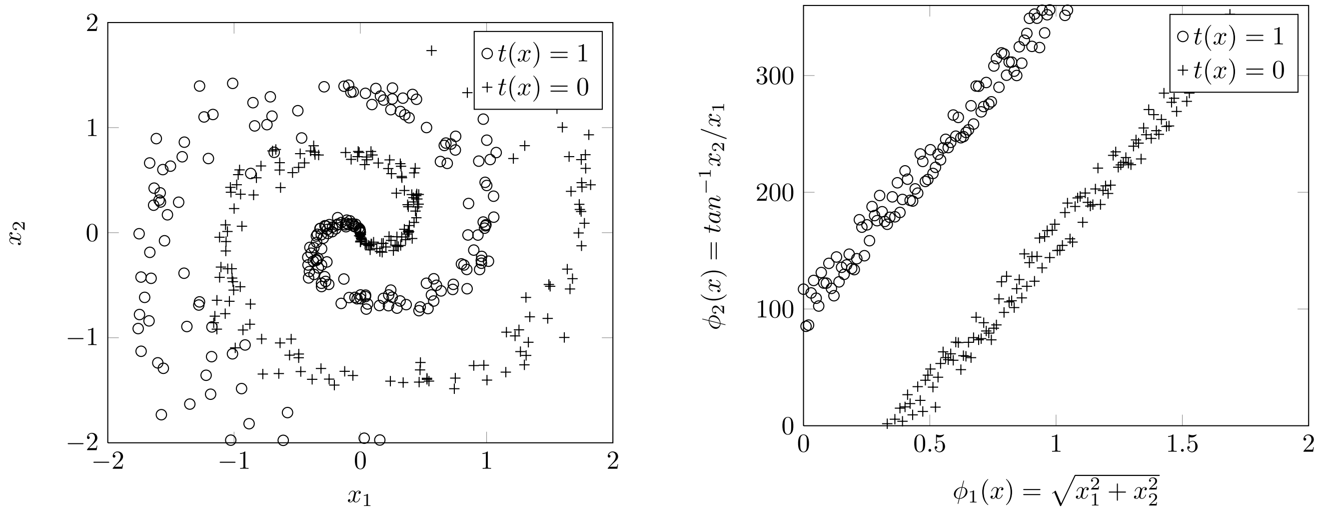
  - $t \in \{0, 1\}$: ground-truth result associated with input $\boldsymbol{x}$

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

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

Data Representation, $\phi(\boldsymbol{x})$

- 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

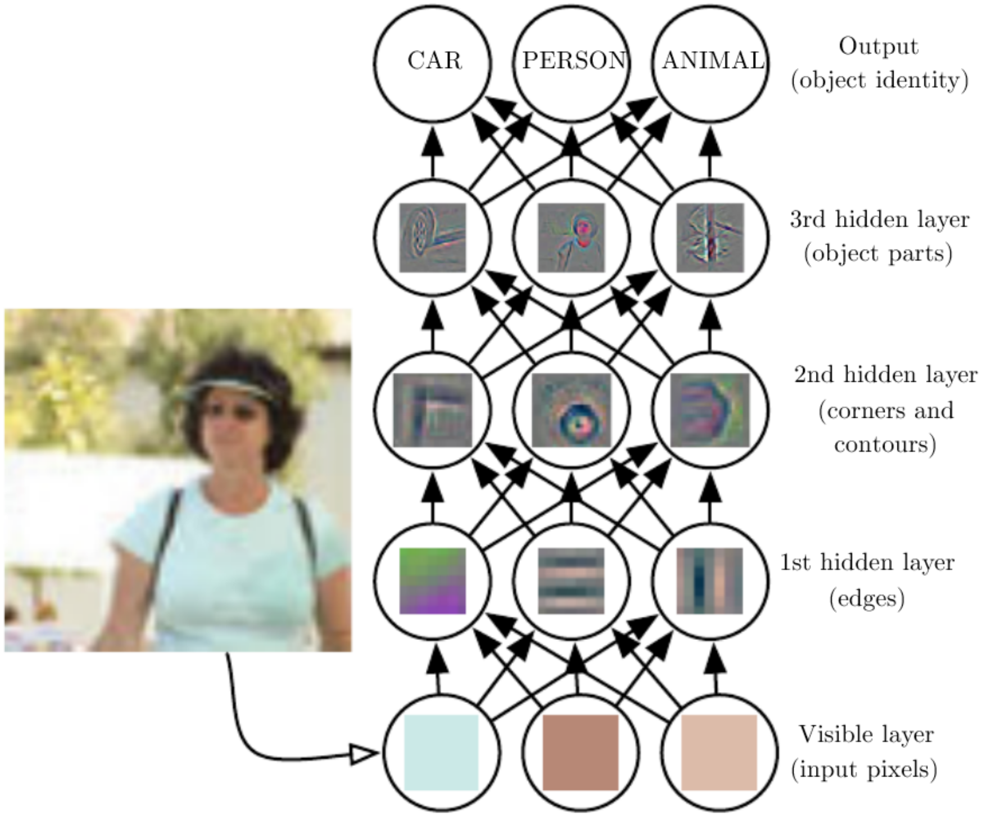
$$f_{\boldsymbol{w},\boldsymbol{\theta_n},\boldsymbol{\theta_{n-1}},\cdots,\boldsymbol{\theta_1}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T \underbrace{\phi_{\boldsymbol{\theta_n}}(\phi_{\boldsymbol{\theta_{n-1}}}(\phi_{\boldsymbol{\theta_{n-2}}}(\cdots\phi_{\boldsymbol{\theta_1}}(\boldsymbol{x}))))}_{\text{Hierarchy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

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

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

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

Example: Feedforward Deep Networks



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

$$\uparrow$$

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

$$\uparrow$$

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

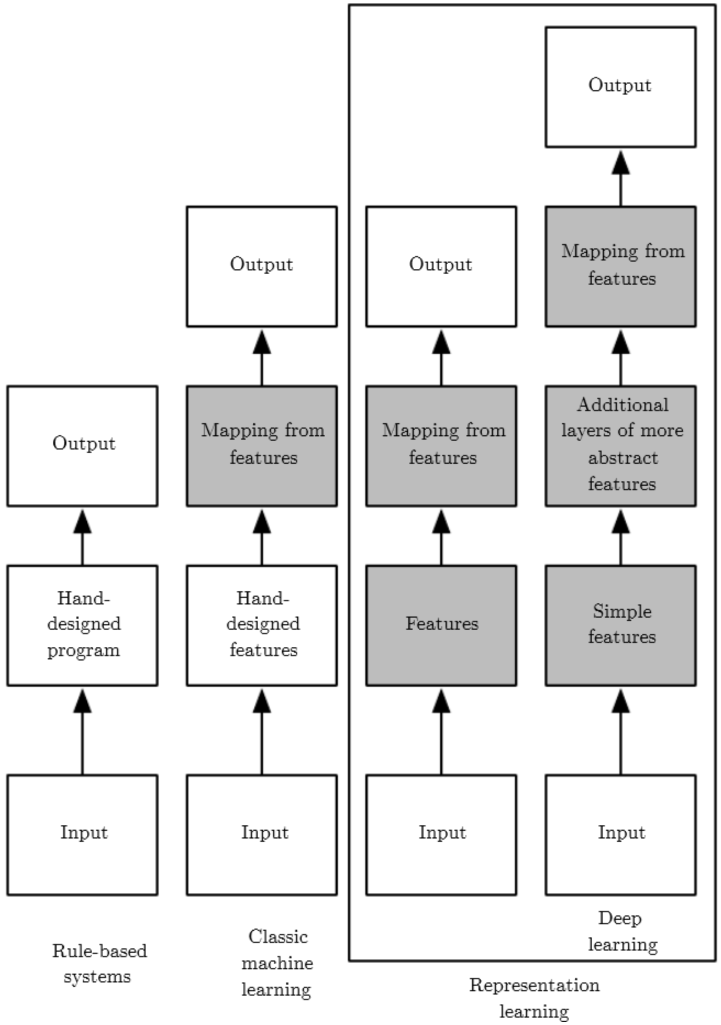
$$\uparrow$$

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

$$\uparrow$$

$$\boldsymbol{x}$$

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); Variational Autoencoder (Kingma et al., 2014); Generative Adversarial Networks (Goodfellow et al., 2014)