Introduction to Machine Learning (and Notation)

David I. Inouye

Friday, September 2, 2022

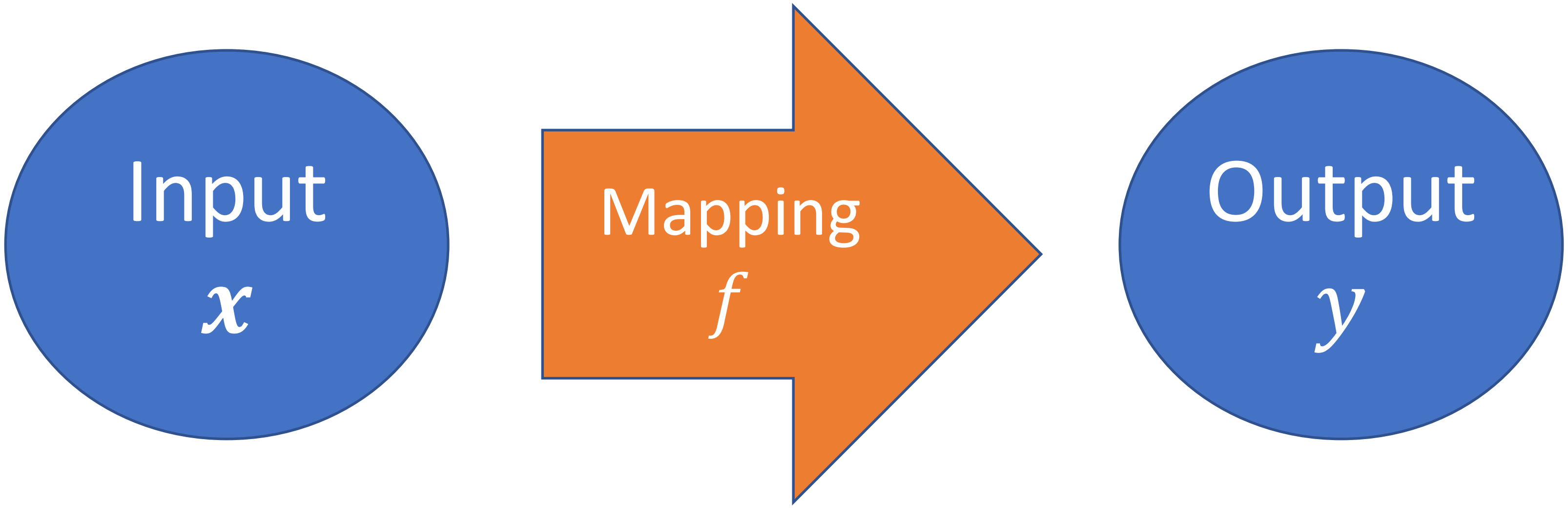
Outline

- Supervised learning
  - Regression
  - Classification

- Unsupervised learning
  - Dimensionality reduction (PCA)
  - Clustering
  - Generative models

- Other key concepts
  - Generalization
  - Curse of dimensionality
  - No free lunch theorem

The goal of supervised learning is to estimate a mapping (or function) between input and output

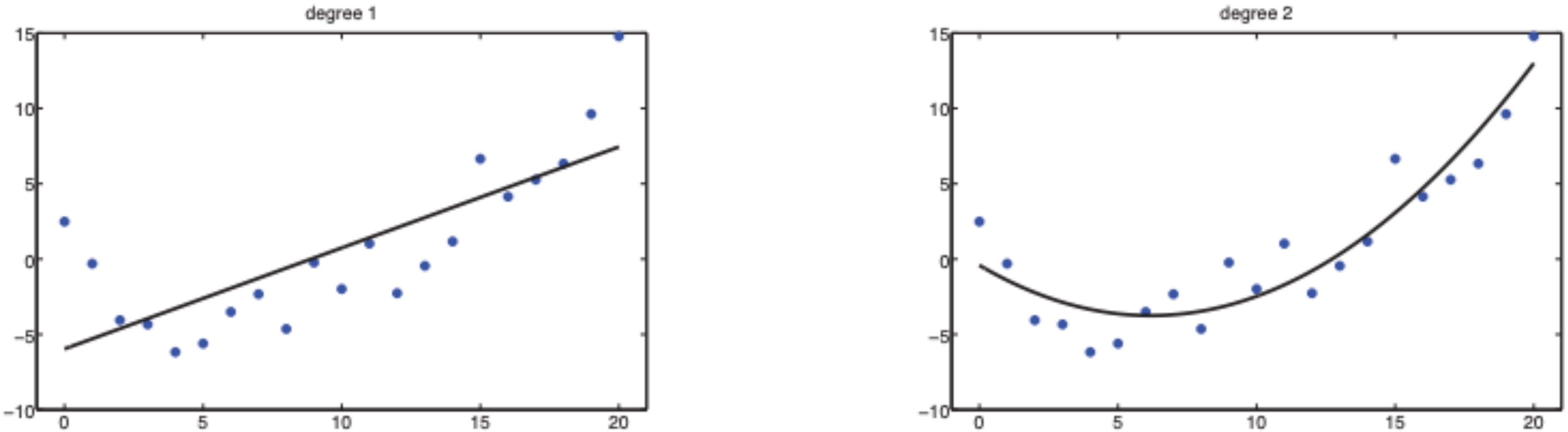


The goal of supervised learning is to estimate a mapping (or function) between input and output *given only input-output examples*

The set of input-output pairs is called a training set, denoted by $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$

- Input $\boldsymbol{x}_i$
  - Called features (ML), attributes, or covariates (Stats). Sometimes just variables.
  - Can be numeric, categorical, discrete, or nominal.
  - Examples
    - [height, weight, age, gender]
    - $[x_1, x_2, \cdots, x_d]$ – A *d*-dimensional vector of numbers
    - Image
    - Email message
- Output y_i
  - Called output, response, or target (or label)
  - Real-valued/numeric output: e.g., $y_i \in \mathcal{R}$
  - Categorical, discrete, or nominal output: y_i from *finite* set, i.e., $y_i \in \{1,2,\cdots,c\}$

If the output y_i is numeric, then the problem is known as <u>regression</u>



NOTE: Input x does not have to be numeric. Only the output y must be numeric.

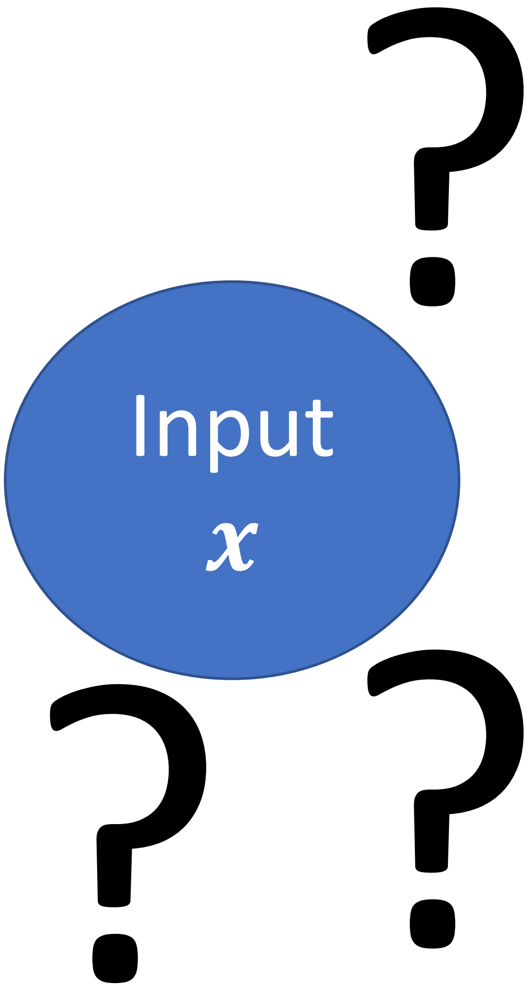
- Given height x_i, predict age y_i
- Predict GPA given SAT score
- Predict SAT score given GPA
- Predict GRE given SAT and GPA

If output is categorical, then the problem is known as classification

- Given height x, predict "male" ($y = 0$) or "female" ($y = 1$)

- Given salary x_1 and mortgage payment x_2, predict defaulting on loan ("yes" or "no")

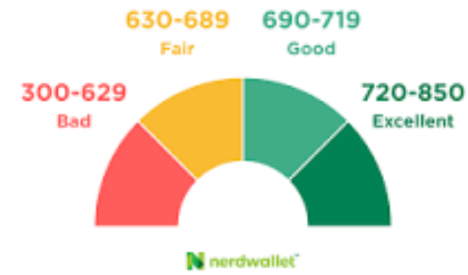
The goal of unsupervised learning is to model or understand the input data directly



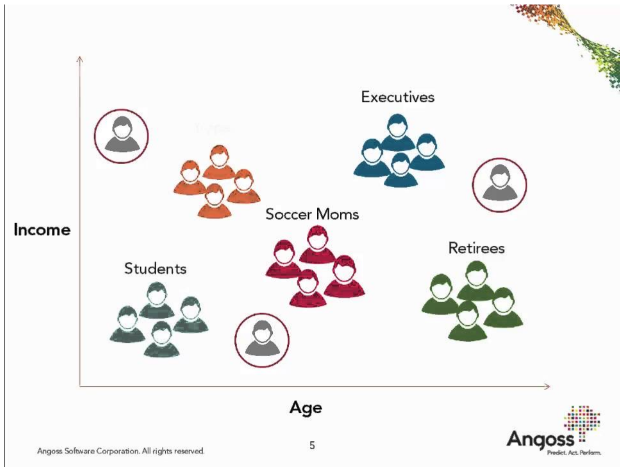
- Dimensionality reduction
- Clustering
- Generative models

"What I cannot create I do not understand"
– Richard Feynman

In unsupervised learning, the training set is only a set of input values $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$

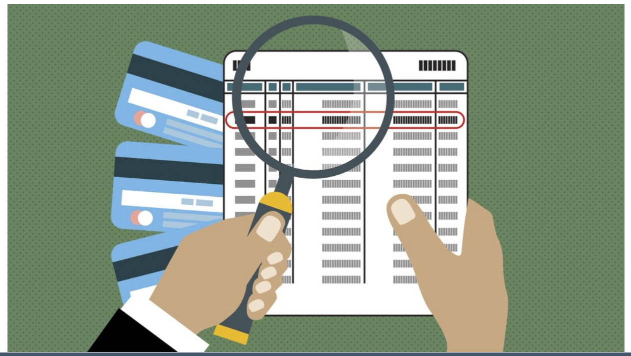


- [Dimensionality reduction] Estimate a single number that summarizes all variables of wealth (e.g. credit score)

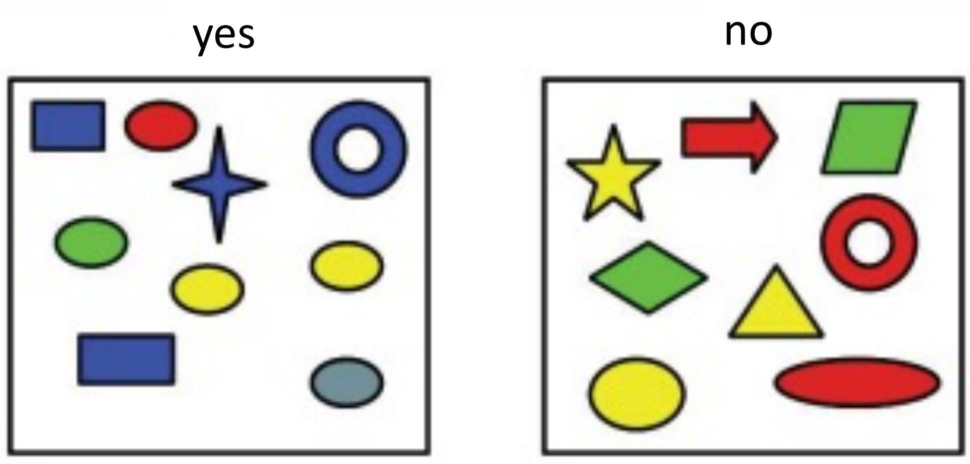


- [Clustering] Estimate natural groups of customers

- [Generative Models] Estimate the distribution of normal transactions to detect fraud (anomalies)



Given this dataset, should we use supervised or unsupervised learning?



d features/attributes/covariates

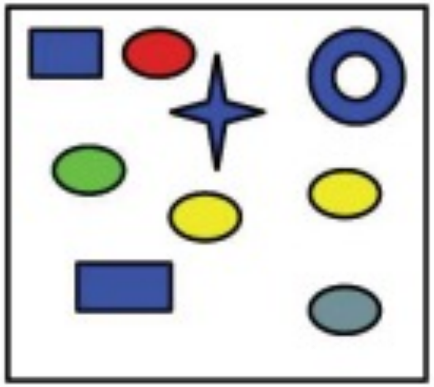
n samples/ observations/ examples

Color	Shape	Size (cm)	Is it good?
Blue	Square	10	yes
Red	Ellipse	2.4	yes
Red	Ellipse	20.7	no

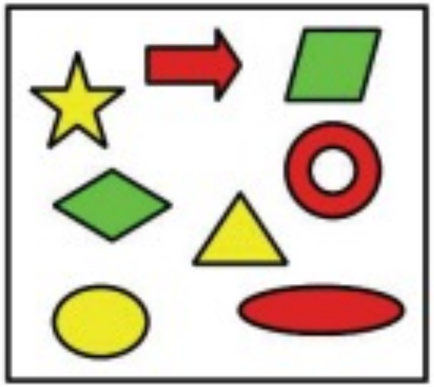
Adapted from Machine Learning: A Probabilistic Perspective, Ch. 1, Kevin P. Murphy, 2012.

The dataset cannot determine the task, rather the context determines the task

yes



no



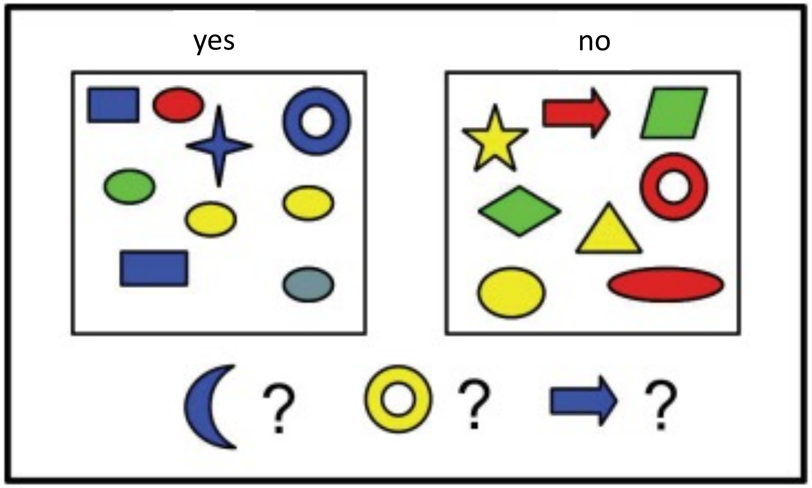
d features/attributes/covariates

n samples/ observations/ examples

Color	Shape	Size (cm)	Is it good?
Blue	Square	10	yes
Red	Ellipse	2.4	yes
Red	Ellipse	20.7	no

Adapted from Machine Learning: A Probabilistic Perspective, Ch. 1, Kevin P. Murphy, 2012.

<u>Generalization</u> *beyond* the training set is the main goal of learning



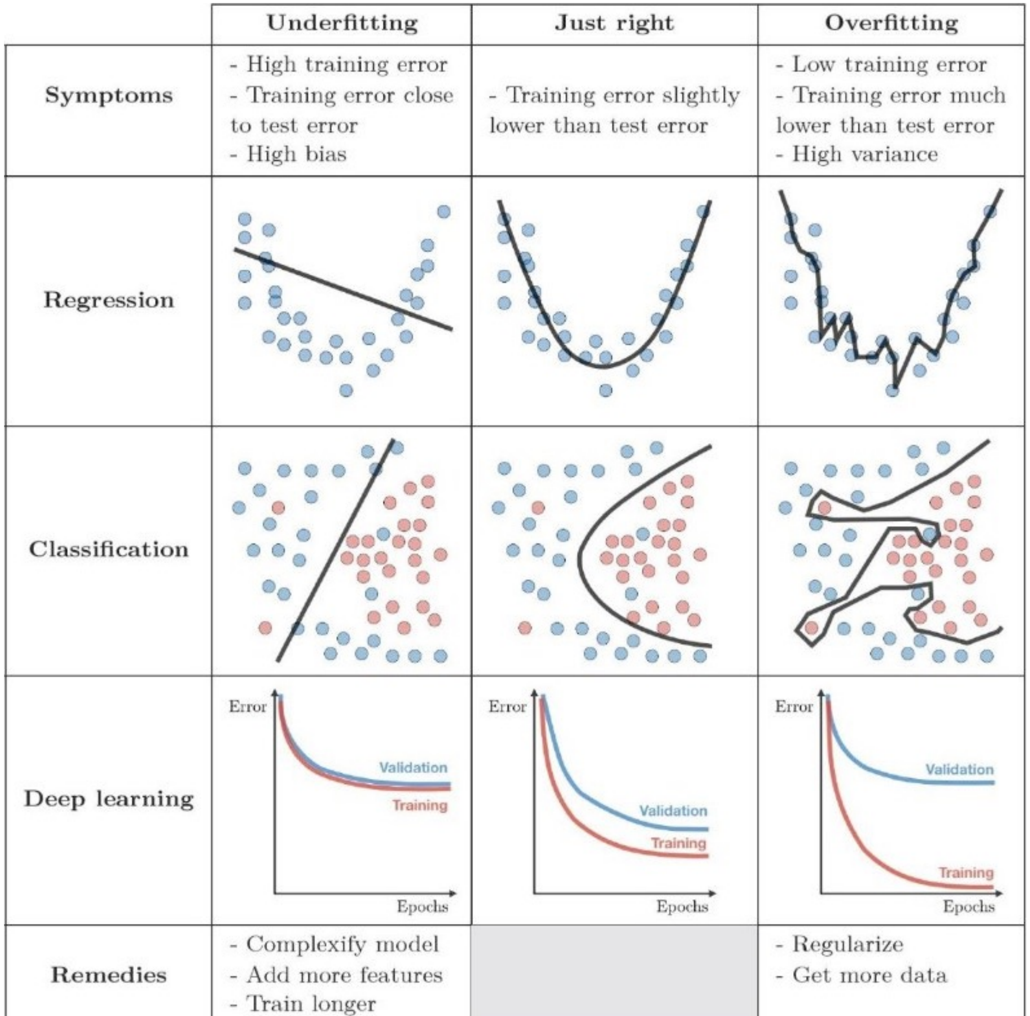
d features/attributes/covariates

n samples/ observations/ examples

	Color	Shape	Size (cm)	Is it good?	
x_1	Blue	Square	10	yes	y_1
x_2	Red	Ellipse	2.4	yes	y_2
	Red	Ellipse	20.7	no	

Example from Machine Learning: A Probabilistic Perspective, Ch. 1, Kevin P. Murphy, 2012.

<u>Generalization</u> *beyond* the training set is the main goal of learning



Original source for figure unknown.

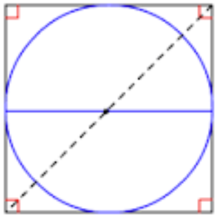
What does generalization look like for *unsupervised learning*?

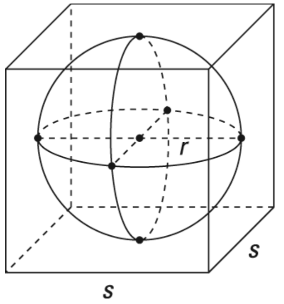
- Generalization in dimensionality reduction

- Generalization in generative models can be understood through the view of log likelihood.

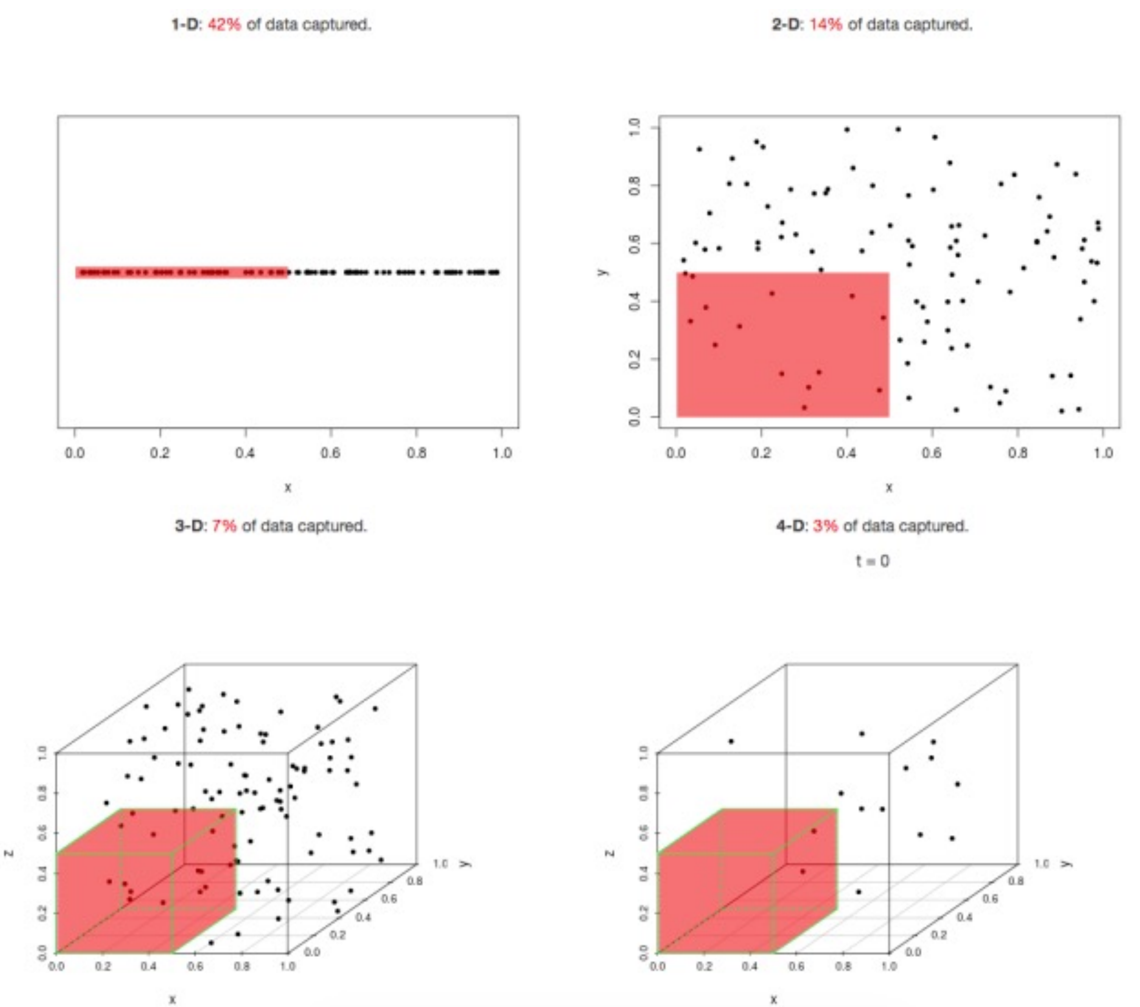
The <u>curse of dimensionality</u> is *unintuitive*
Example: Most space is in the "corners"

- Ratio between unit hypersphere to unit hypercube
  - 1D : $2/2 = 1$
  - 2D : $\frac{\pi}{4} = 0.7854$
  - 3D : $\frac{\frac{4}{3}\pi}{8} = 0.5238$
  - d-dimensions: $V_d(r) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} r^d$
    - Thus, for 10-D: 2.55/2^10 = 2.55/1024 = 0.00249





The curse of dimensionality is *unintuitive*
The number of points in ½ cube is very small



https://eranraviv.com/curse-of-dimensionality/

The <u>curse of dimensionality</u> is ***unintuitive***
Example: Need edge length to be 0.9 to capture 1/2 data samples in 10 dimensions

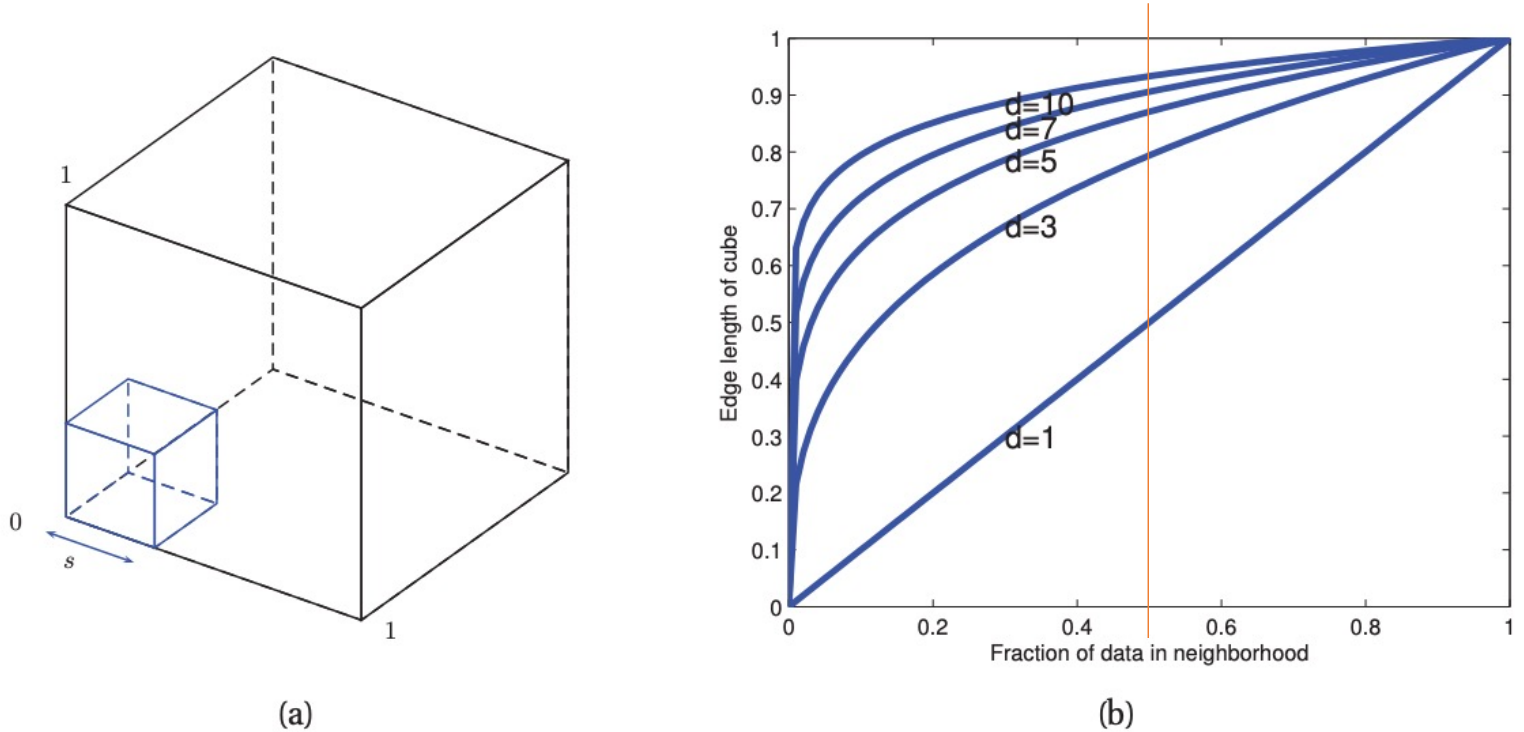
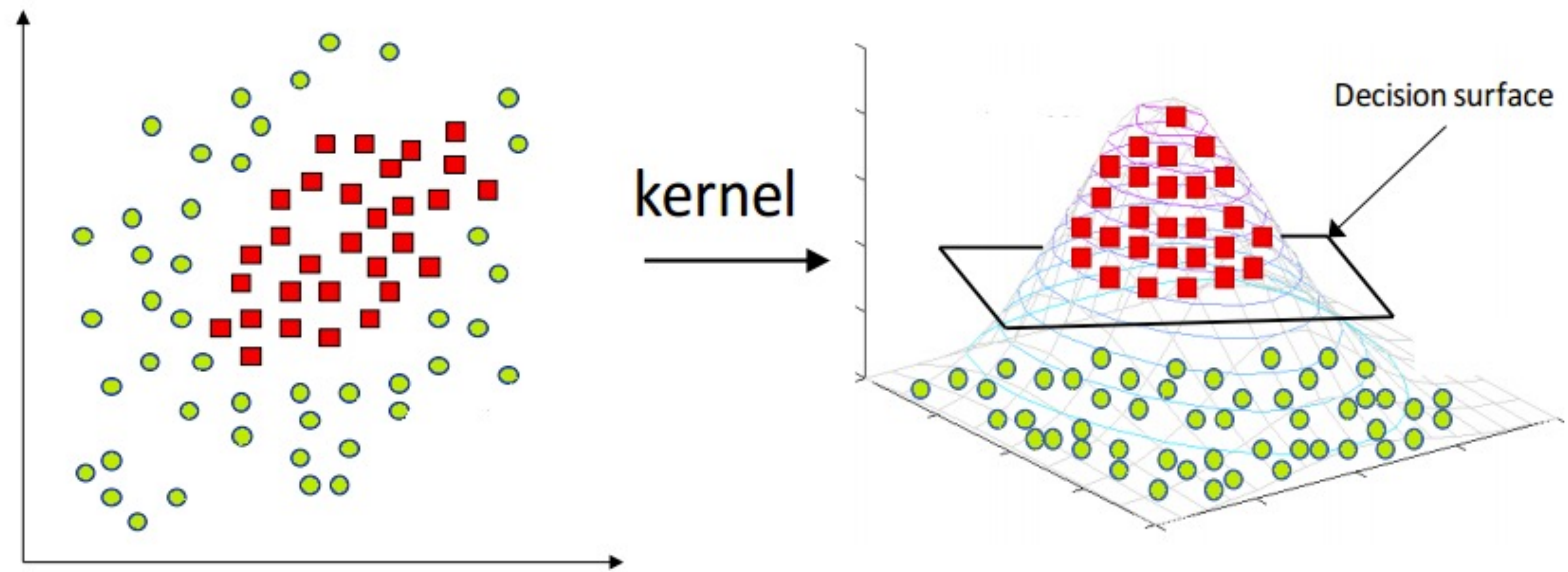


Figure 1.16 Illustration of the curse of dimensionality. (a) We embed a small cube of side s inside a larger unit cube. (b) We plot the edge length of a cube needed to cover a given volume of the unit cube as a function of the number of dimensions. Based on Figure 2.6 from (Hastie et al. 2009). Figure generated by `curseDimensionality`.

From *Machine Learning: A Probabilistic Perspective*, Kevin Murphy, 2012.

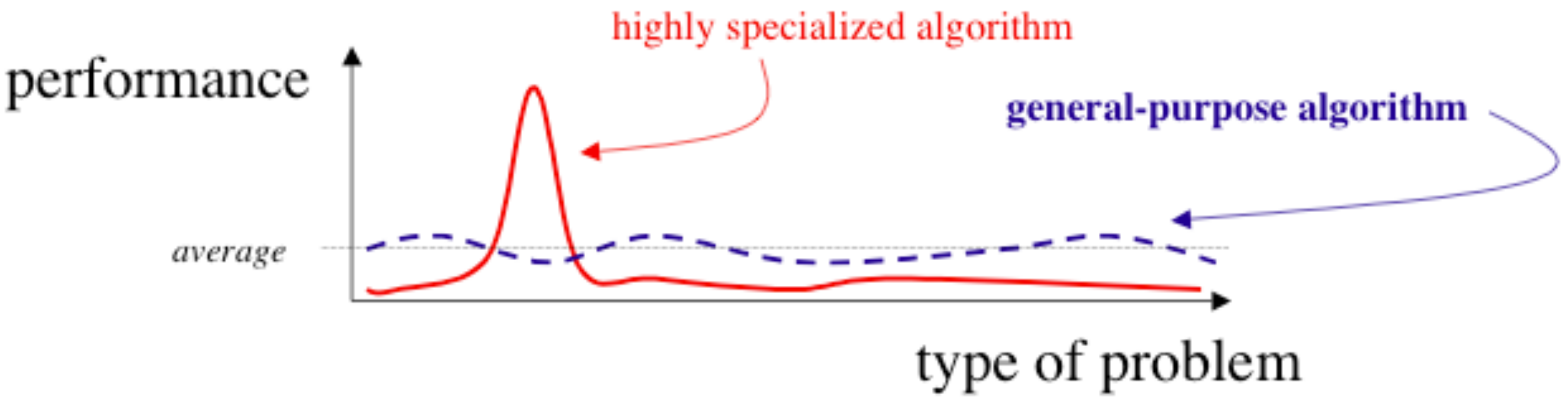
The “blessing” of dimensionality
(more data generally doesn’t hurt if you can ignore)



https://www.hackerearth.com/blog/developers/simple-tutorial-svm-parameter-tuning-python-r/

No Free Lunch Theorem
(“All models are wrong, but some models are useful.”*)

- All models are approximations
- All models make assumptions
- Assumptions are never perfect



* George Box (Box and Draper 1987, page 424).