CS6375: Machine Learning

Gautam Kunapuli

Introduction to Machine Learning

THE UNIVERSITY OF TEXAS AT DALLAS
Erik Jonsson School of Engineering and Computer Science

Learning

Herbert Simon: "Learning is any process by which a system improves performance from experience."

A Collaborator: "We need machine learning because we like being lazy. i.e., let the machines learn to do what we do"

Bill Gates: If we invented a breakthrough in artificial intelligence so machines can learn, that is worth 10 Microsofts.

Machine Learning seeks to:

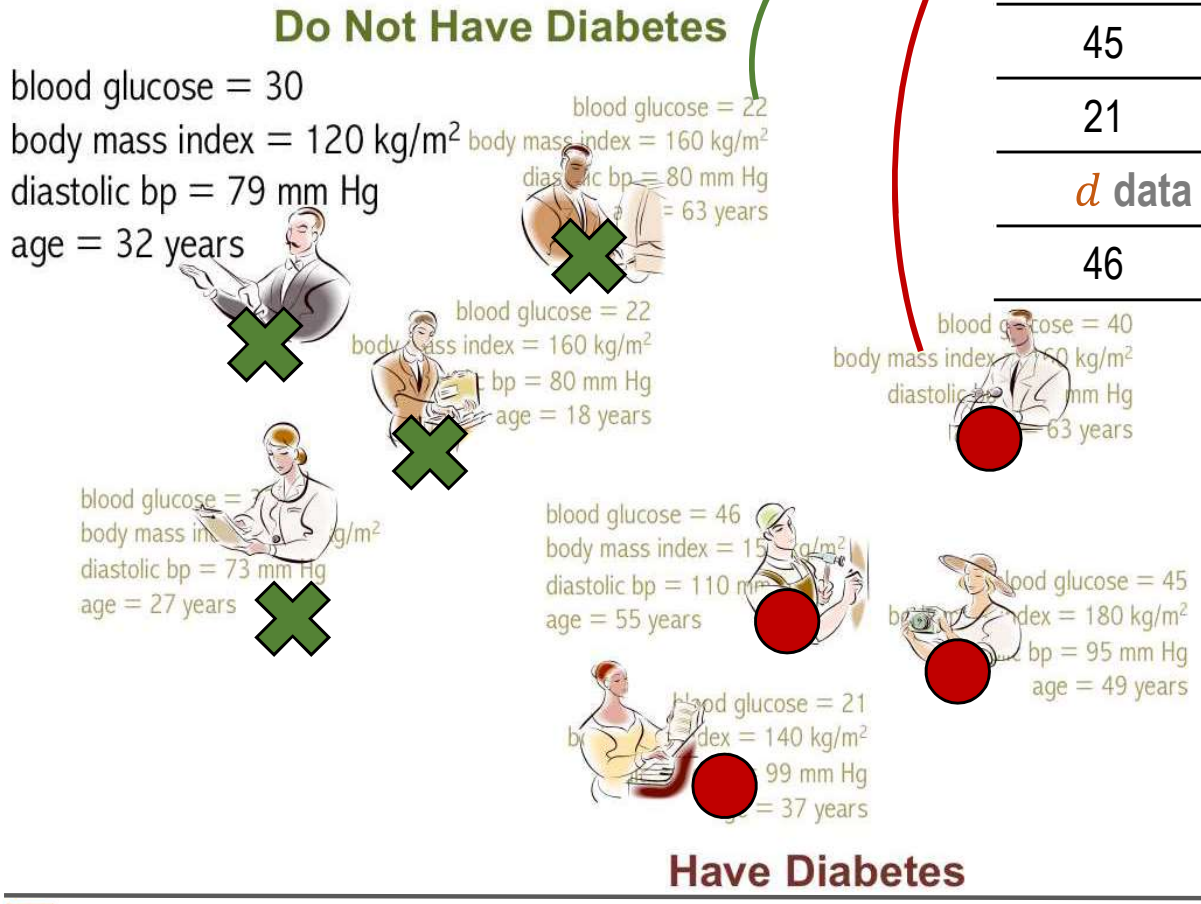
- Develop systems that can **automatically adapt** and **customize** themselves to **individual** users;
- **Discover** new **knowledge** from large databases.

Machine Learning Terminology

Example: **Diabetes Diagnosis from Medical Records.** You are developing a clinical decision-support system to **diagnose diabetes** from patient health records.

blood glucose	body mass idx	diastolic blood pr.	age	Diabetes?
30	120	79	32	NO
22	160	80	63	NO
40	160	93	63	YES
22	160	80	18	NO
45	180	95	49	YES
21	140	99	37	YES
d **data features** (aka attributes, variables)				NO
46	153	110	55	YES

n data examples



Each patient is an **example** (or a **data point**) the machine can learn from

Data: **patient information** (with **patient features** describing the patient: *vitals, health history, demographics, lab test results etc.*) from patient health records

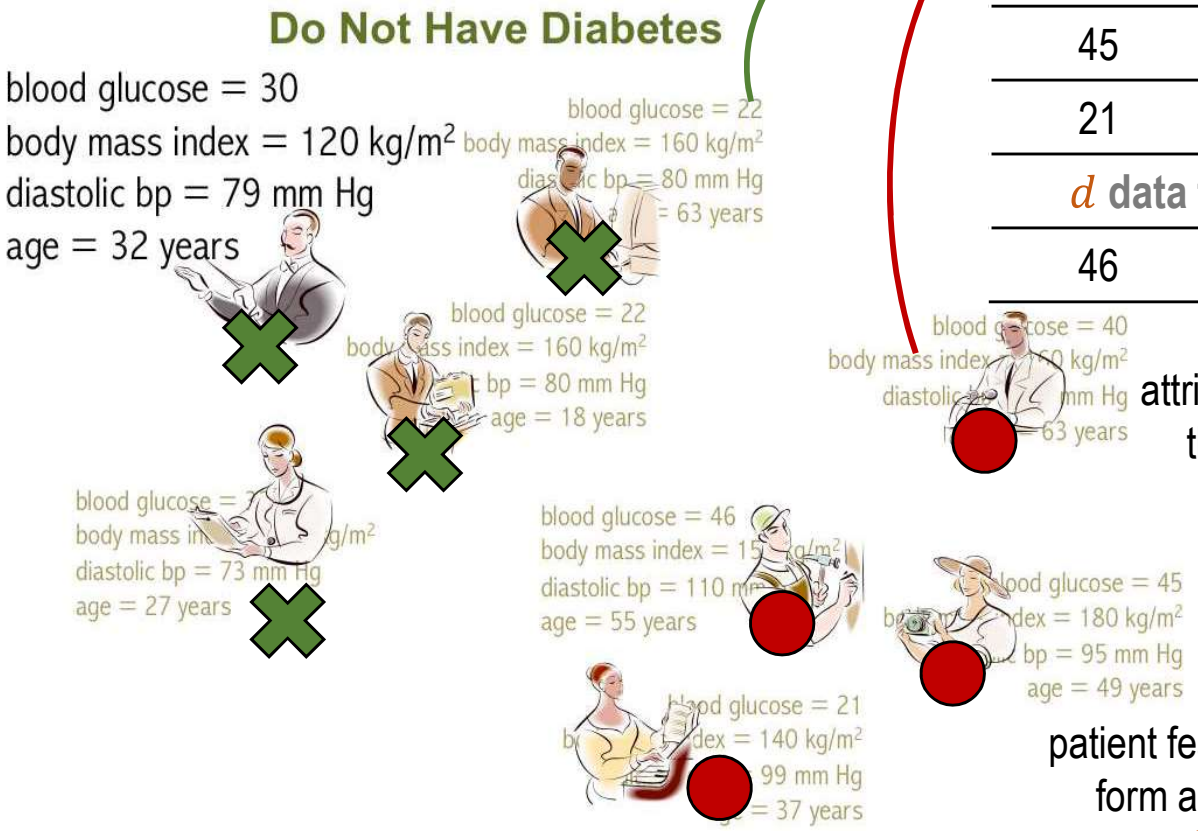
Label: **patient diagnosis** (does the patient have diabetes, **YES**/**NO**?)

Machine Learning Terminology

Example: **Diabetes Diagnosis from Medical Records.** You are developing a clinical decision-support system to **diagnose diabetes** from patient health records.

blood glucose	body mass idx	diastolic blood pr.	age	Diabetes?
30	120	79	32	NO
22	160	80	63	NO
40	160	93	63	YES
22	160	80	18	NO
45	180	95	49	YES
21	140	99	37	YES
d **data features** (aka attributes, variables)				NO
46	153	110	55	YES

n data examples

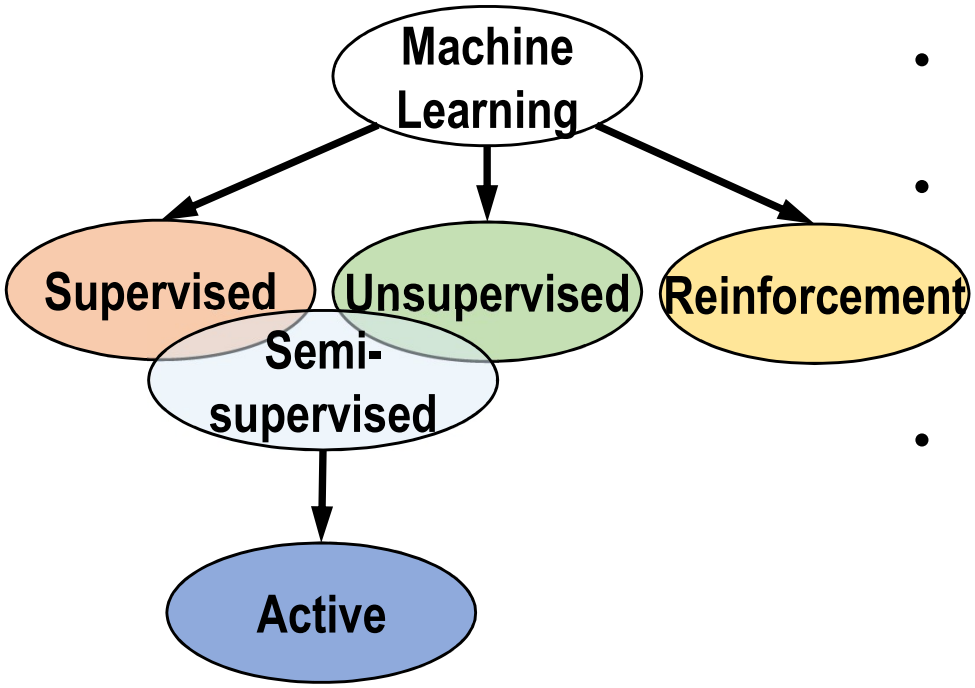


attributes and descriptors for each patient are the **features** or **independent variables**
for the i^{th} patient, the k^{th} feature is denoted x_i^k

the diagnosis or the **prediction** is the target or the (training) **label**
for the i^{th} patient, denoted y_i

patient features are collected to form a (training) **example**
for the i^{th} patient, denoted $\boldsymbol{x}_i$

Types of Learning



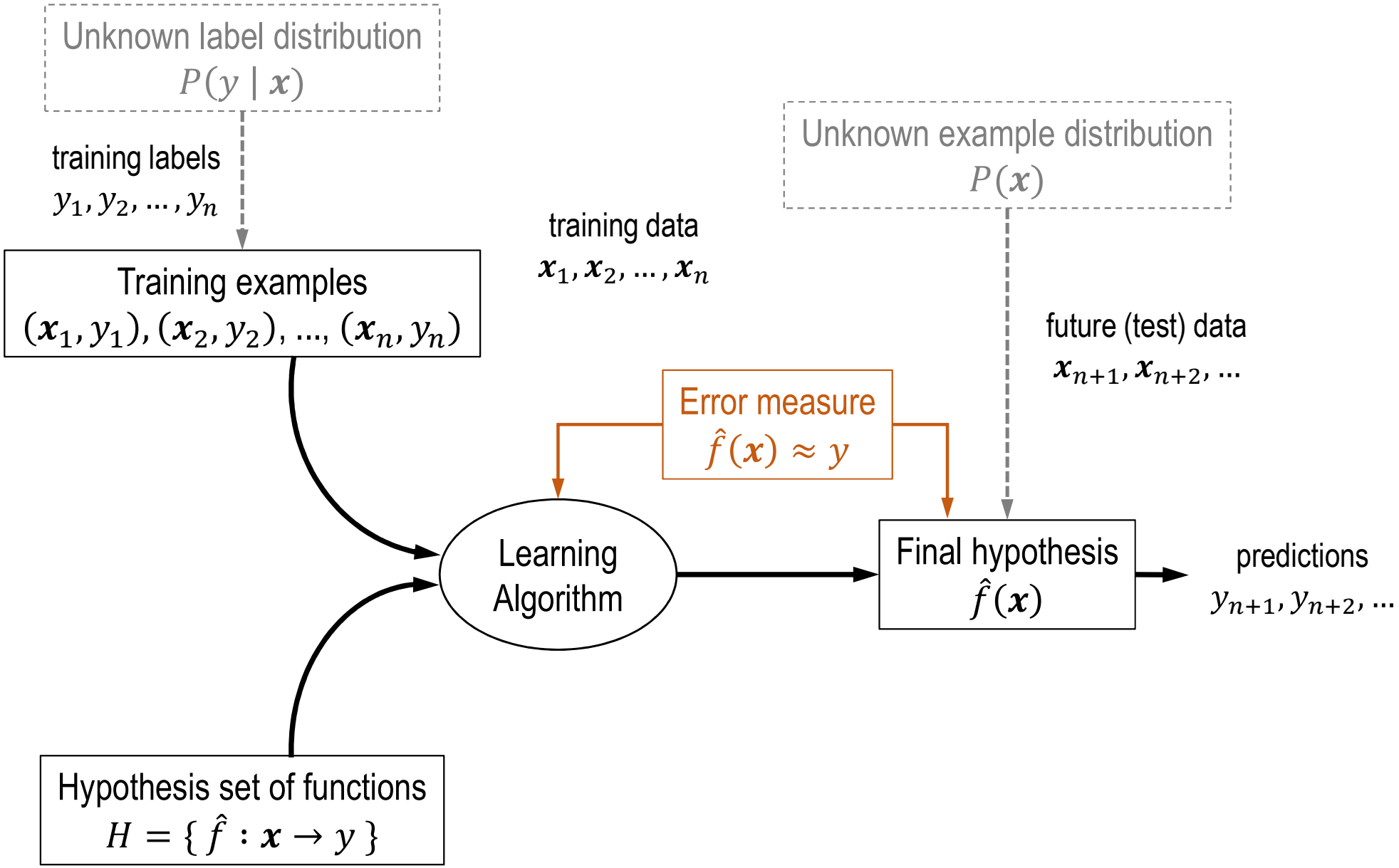
- **Supervised learning**
  - training data includes desired output
- **Unsupervised learning**
  - training data does not include desired output
- **Semi-supervised learning**
  - some training data comes with desired output
- **Active learning**
  - semi-supervised learning where machine learner can ask for the correct outputs for specific data points
- **Reinforcement learning**
  - Machine learner interacts with the world via allowable actions that change the state of the world and result in rewards
  - learner attempts to maximize rewards through repeated trial and error (*practice makes perfect*)

Supervised Learning: Examples

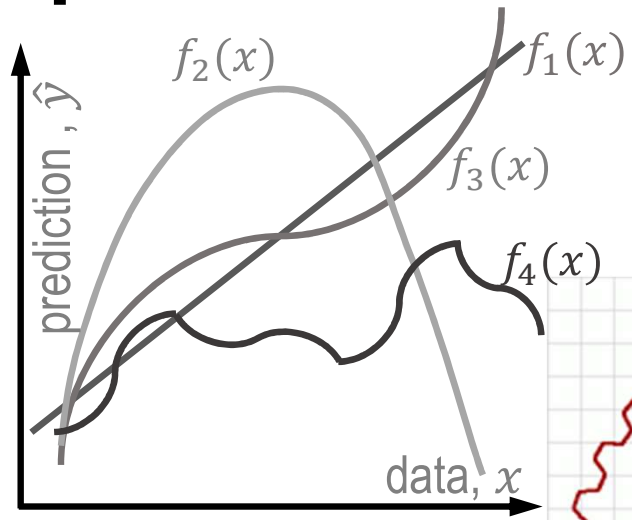
Given: **Labeled** training examples $(\boldsymbol{x}_i, y_i)_{i=1}^n$ for some task,
Find: Model $f(\boldsymbol{x})$ that can **predict** $y_i = f(\boldsymbol{x}_i)$

- **Situations where there is no human expert**
  - x: bond graph of a new molecule
  - f(x): predicted binding strength to AIDS protease molecule
- **Situations where humans can perform a task but can't describe how they do it**
  - x: picture of a hand-written character
  - f(x): ascii code of the character
- **Situations where the desired function is changing frequently**
  - x: description of stock prices and trades for last 10 days
  - f(x): recommended stock transactions
- **Situations where each user needs a customized function *f***
  - x: incoming email message
  - f(x): importance score for presenting to the user

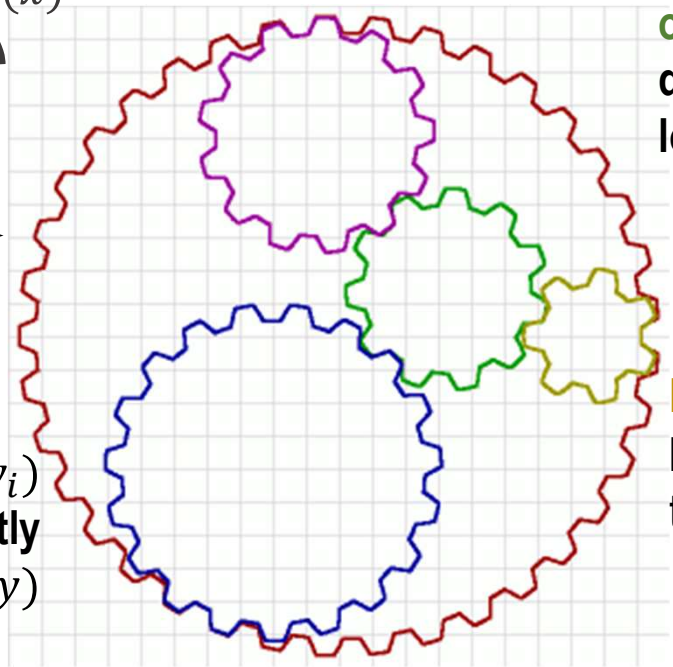
Supervised Learning: General Setup



Supervised Learning: General Setup

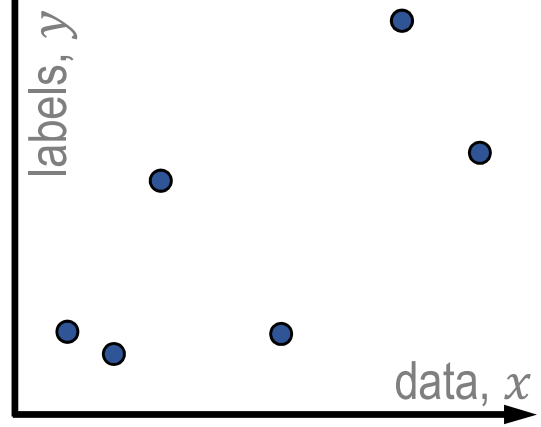


hypothesis class: the set of functions **allowed to model** the data and make predictions $y \approx \hat{f}(x)$

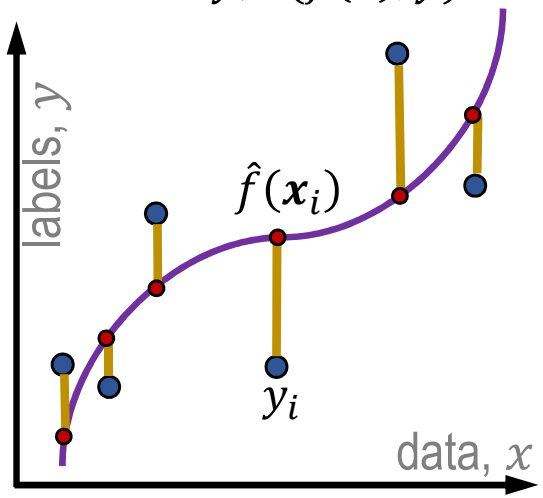


optimization algorithm: to find the **best quality hypothesis** (with the **smallest loss**) as measured on the **training data**

training data (x_i, y_i) **identically and independently** sampled from unknown $P(x, y)$



loss function: to measure the **quality** of how well the model's predictions $\hat{f}(x)$ fit the true labels y; $L(f(x), y)$



Key Questions

How can we formulate practical problems using machine learning?

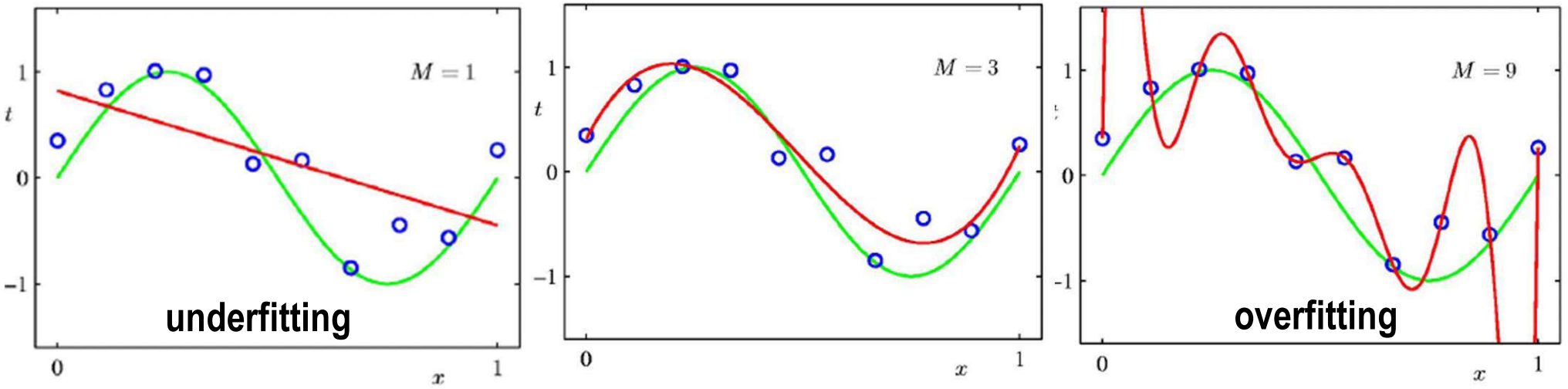
- What are **good hypothesis spaces**?
  - Selecting the best model for the data
- What are **good loss functions**?
  - Selecting a good measure for the quality of the **fit**
- What **algorithms work** on these spaces?
  - Selecting the most efficient optimization algorithms
- How can we **generalize** to unseen points (i.e., **future data**)?
  - Ensuring we predict well (overfitting vs. underfitting)
- How can we **trust** our results?
  - Best machine learning practices during data gathering, preprocessing, model selection, model evaluation

Three Learning Principles: 1. Occam's Razor

The simplest model that fits the data is also the most plausible.

smaller hypothesis space produces simpler models, but this does **not** necessarily fit the data sufficiently

larger hypothesis space always decreases the loss function, but this does **not** necessarily mean better predictive performance



less complex worse fit ← → *more complex better fit*

pick the model carefully

Y. S. Abu-Mostafa, M. Magdon Ismail, H.-T. Lin, *Learning from Data: A Short Course*, Chap. 5

Three Learning Principles: 2. Sampling Bias

If the data is sampled in a biased way, learning will produce a similarly biased outcome.

Example: Suppose a biologist wants to estimate the average size of fish in a pond. The biologist collects data with a fishing net: scoop out a few specimens, measure their length, and compute their average.

Bias:

- Fishing net's mesh size is too large
  - *fish smaller than mesh size not sampled*
- Fishing net can't reach bottom of the pond
  - *very large fish at the bottom of the pond not sampled*

pick the data carefully

Three Learning Principles: 3. Data Dredging

If you torture the data long enough, it will confess.

Data Dredging, data snooping: testing **huge numbers of hypotheses** about a single data set exhaustively

When **enough** hypotheses (machine-learning models) are tested, **eventually some** will be **statistically significant**

- misleading, because nearly every data set contains some **spurious correlations**
- if not cautious, MLers using data mining techniques can be easily misled by these results.

handle the data carefully

Smith GD, Ebrahim S. Data dredging, bias, or confounding : They can all get you into the BMJ and the Friday papers. BMJ : British Medical Journal. 2002; 325(7378):1437-1438. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124898/