

# Introduction

COMP3314 — Week 1

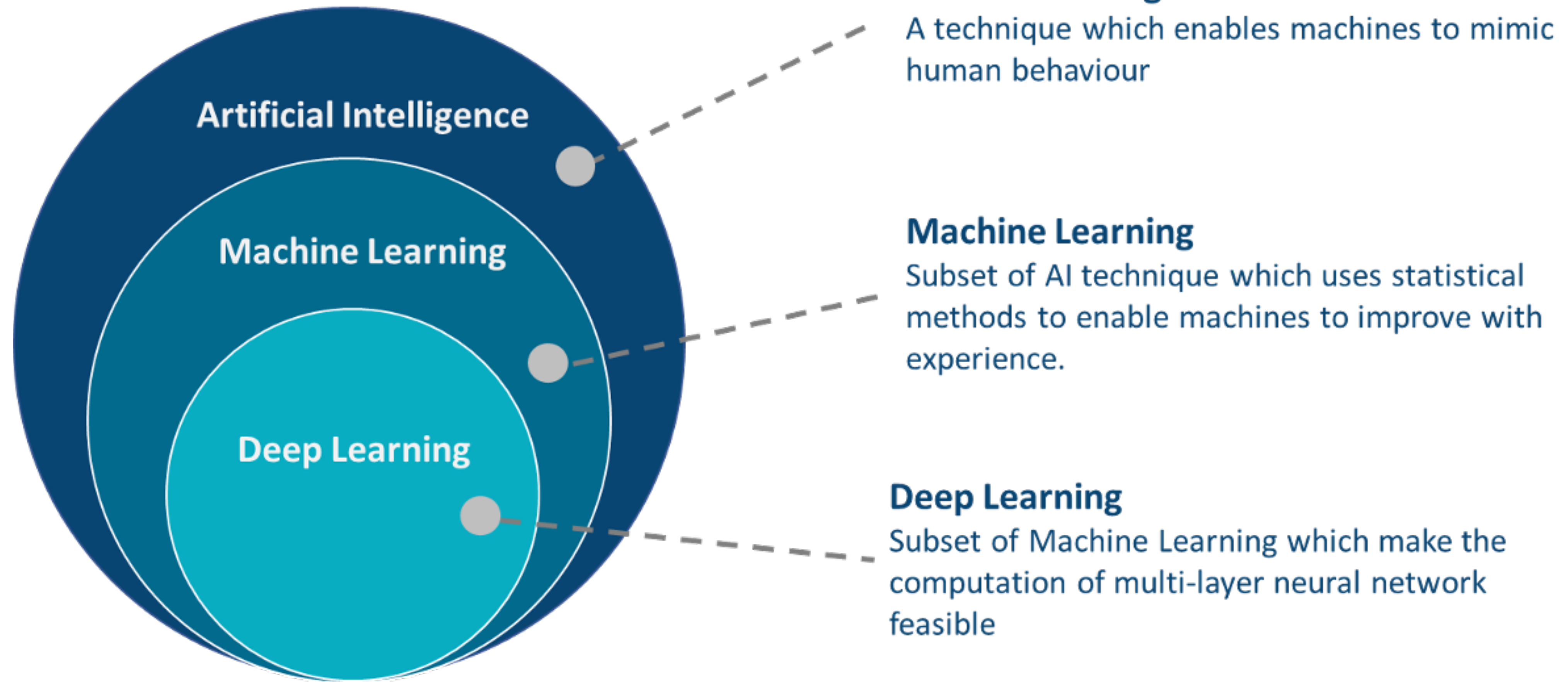
Lingpeng Kong

Department of Computer Science, The University of Hong Kong

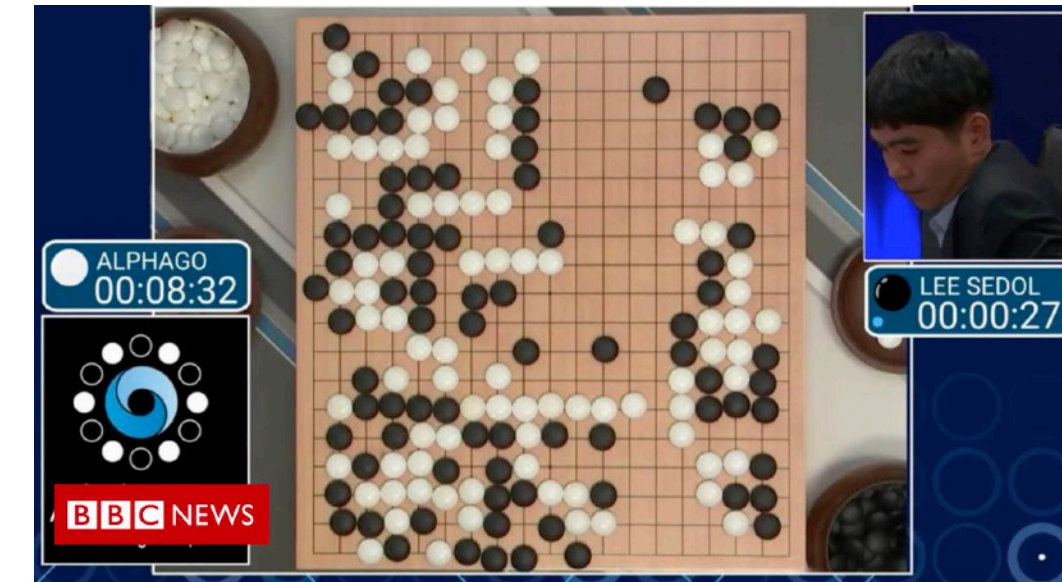
Based on: Probabilistic Machine Learning by Kevin Murphy

Slides from: Saw Shier Nee with special thanks!

# What is AI? ML? DL?



# History of AI



“Can machines think?”

First AI Winter

Second AI Winter

AlphaGo

1950

1966

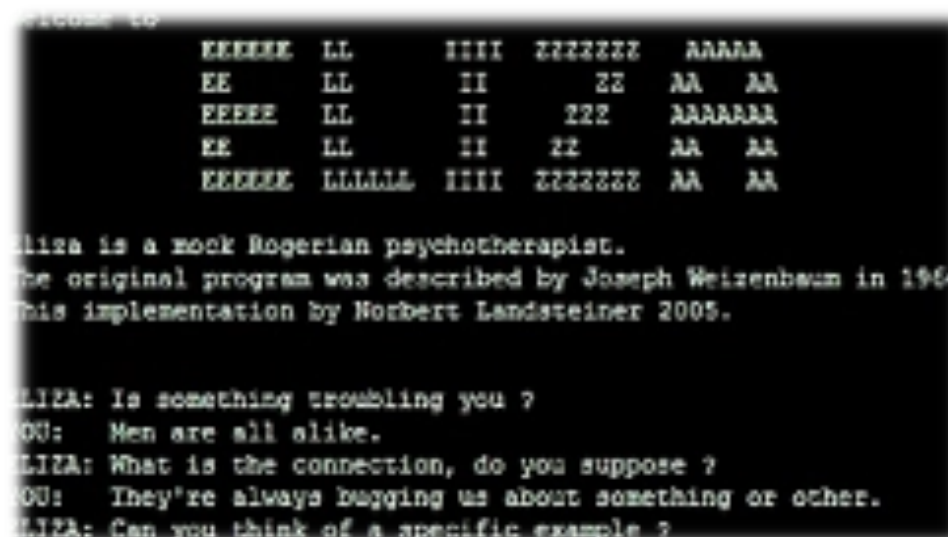
1975

1980

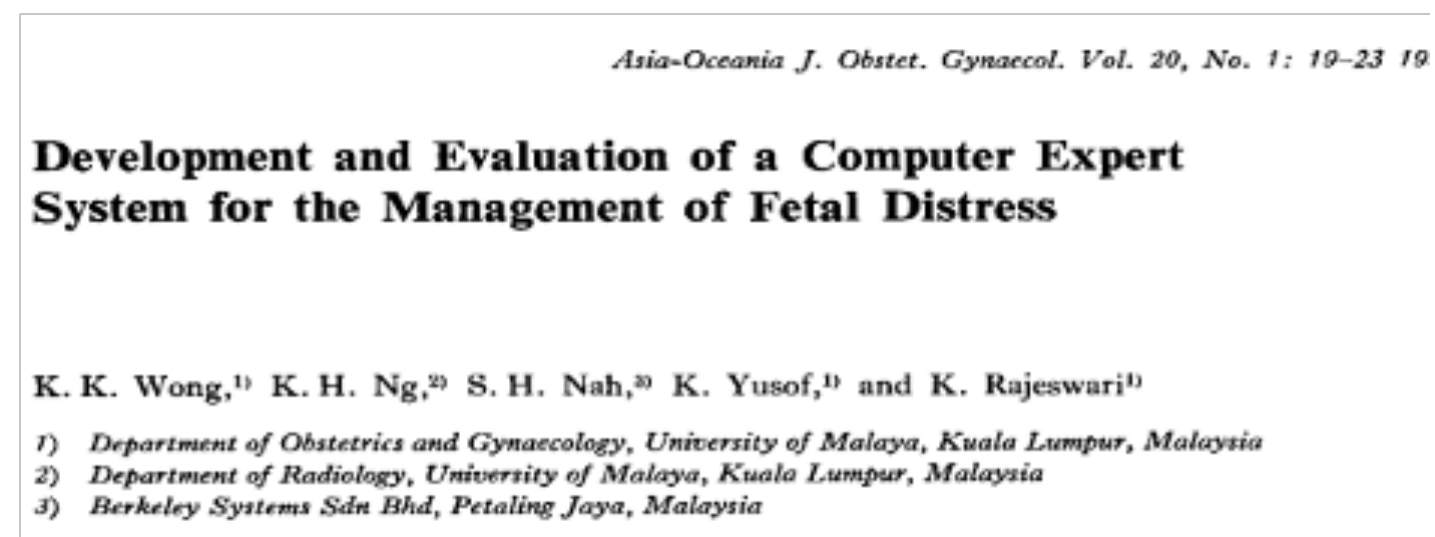
1987

2016

First Chatbot: ELIZA



Expert System





# ChatGPT [December 2022]

## ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

FL Hi, I'm writing an article about you and all of the cool things you can do. Could you say hello to my readers and give us a quick self-introduction? I'd like it to be friendly and casual, include a silly joke, and end with a note about how great the rest of my article is.

Hello readers! My name is Assistant, and I'm a large language model trained by OpenAI. I'm here to help answer your questions and provide you with information on a wide variety of topics. I'm a bit of a jokester, so here's a quick one for you: Why was the math book sad? Because it had too many problems! But seriously, I'm excited to be featured in your article and I hope you enjoy learning more about me and all the cool things I can do. And trust me, the rest of the article is going to be great!

FL Can you write me a paragraph about someone trying to sell an egg but they think the egg is actually a drawing tool?

Sure! Here's a paragraph:

The person was trying to sell their egg at a street market, but they thought it was actually a drawing tool. They were confidently telling potential buyers about the egg's ability to create beautiful, detailed sketches with its delicate, curved shape. Some curious passersby stopped to listen and even gave the egg a try, but they were quickly disappointed to find that it was just an ordinary egg and not a drawing tool at all. The person was embarrassed and quickly moved on to the next item in their sales pitch.

FL Can you write a version of their sales pitch?

Sure! Here's a possible sales pitch for the egg-turned-drawing-tool:

Hello, friends! I have a very special item to show you today. This is no ordinary egg - it's a one-of-a-kind drawing tool. With its delicate, curved shape, this egg is perfect for creating beautiful, detailed sketches. Just hold it in your hand and let your creativity flow. You'll be amazed at the incredible art you can create with this amazing egg. And the best part? It's only \$5! Don't miss out on this incredible deal. Come and see for yourself how this egg can unleash your inner artist.

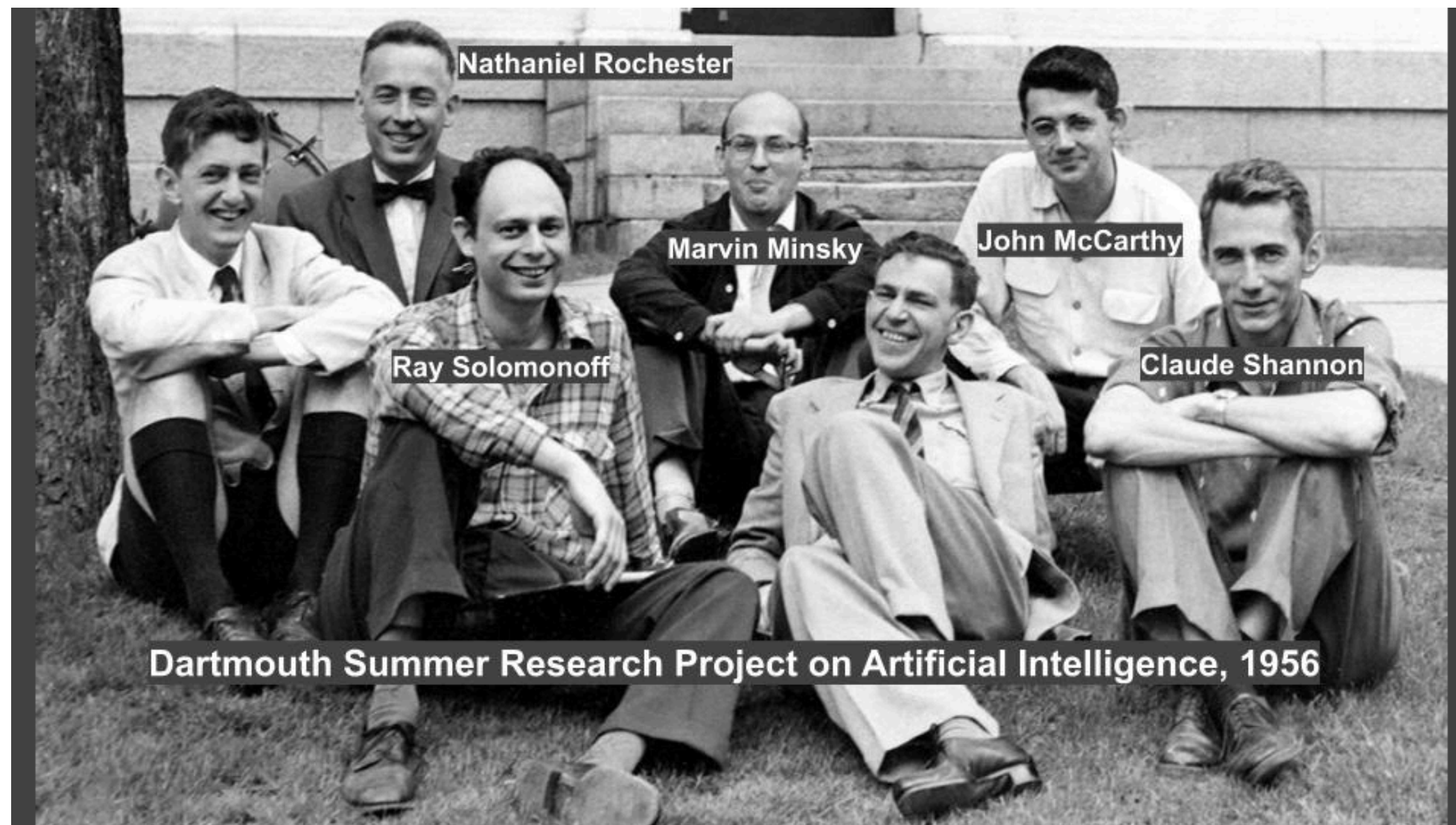
FL They also think the egg has the capability to brainwash humans who lick it, so they always include a warning in their pitch.

Sure, here's a possible warning to include in the sales pitch for the egg-turned-drawing-tool:

**Warning:** This egg has the unique ability to brainwash humans who lick it. Please be careful when handling this egg and do not lick it under any circumstances. We want you to enjoy the incredible benefits of this egg as a drawing tool, but we also want to ensure your safety and well-being. Thank you for understanding.



# AI and Machine Learning



Dartmouth Summer Research Project on Artificial Intelligence, 1956 @ Dartmouth College



# ICML

International Conference  
On Machine Learning

1st International Conference on Machine Learning (1st Machine Learning Workshop), 1980 @ Pittsburgh

# What is Machine Learning?

Study of algorithms that:

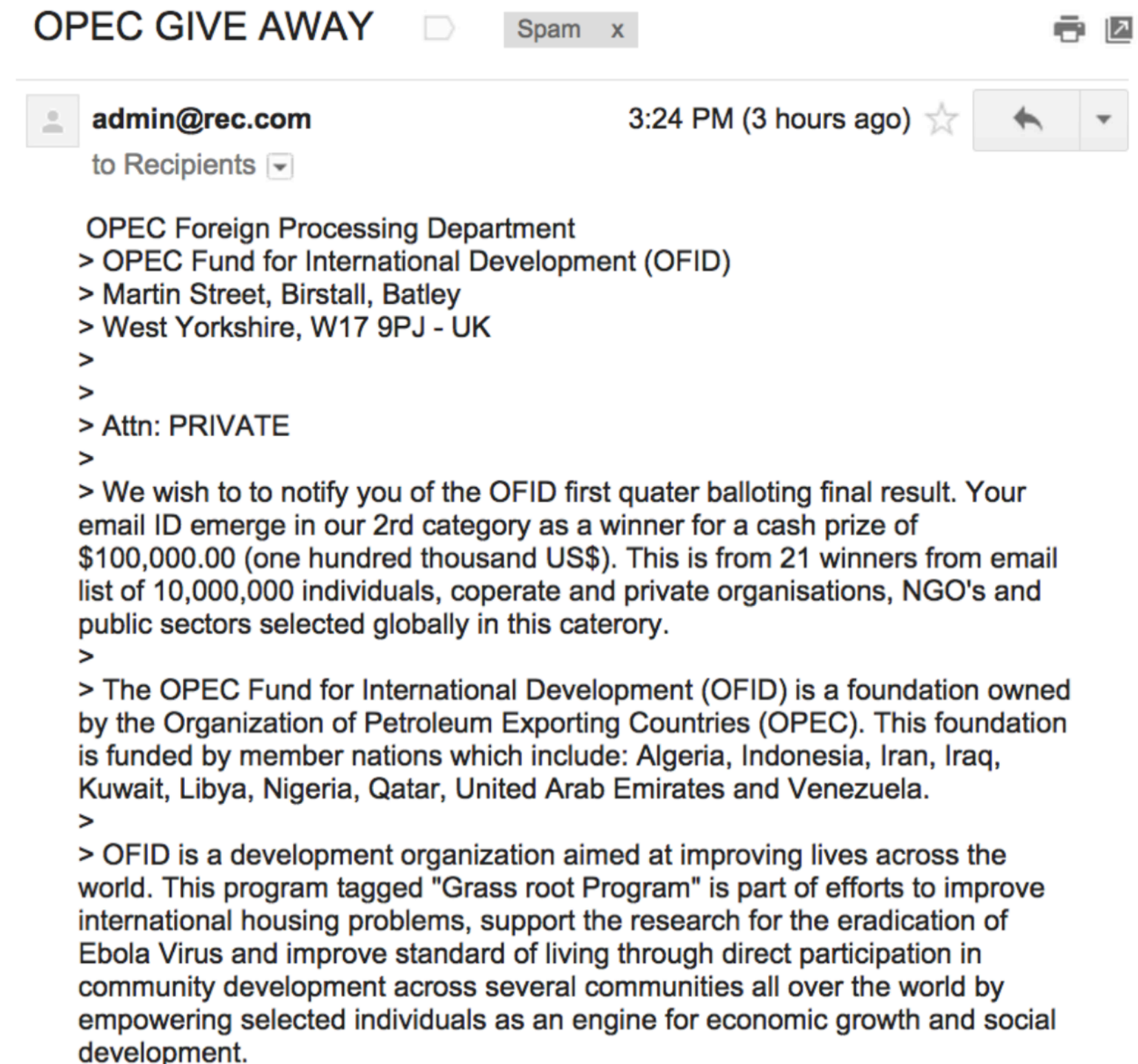
improve their performance P

at some task T

with experience E

well-defined learning task: <P,T,E>

Mitchell, 1997



spam?



# Machine Learning Problems

Supervised Learning

Labelled data with  
guidance

Unsupervised  
Learning

No labelled without  
guidance

Reinforcement  
Learning

Interacts with  
environment, decide  
action, learns by trial  
and error method

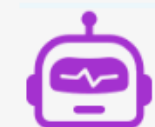


# Zero-shot Ability

Lingpeng Kong

"Good day Author!  
Call For Books,  
Unique Pub International (UPI), a self-sufficient and autonomous company, is a global service provider.  
Click Here for more information 📩 UPI..... BOOKS  
Email us if you have any questions editorialoffice@uniquepubinternational.com or upipublications@gmail.com  
Publication takes 15–30 Days.  
Author Benefits: Book will be published with ISBN  
Prompt Email Notification  
Low Processing Fee  
Publication certificate will be provided to each author  
🌸Regards🌸  
Unique Pub International  
Contact no: [REDACTED]  
web: [REDACTED]  
Email: [REDACTED]  
[REDACTED].com" is this a spam email?

28/08/2023, 10:44:17

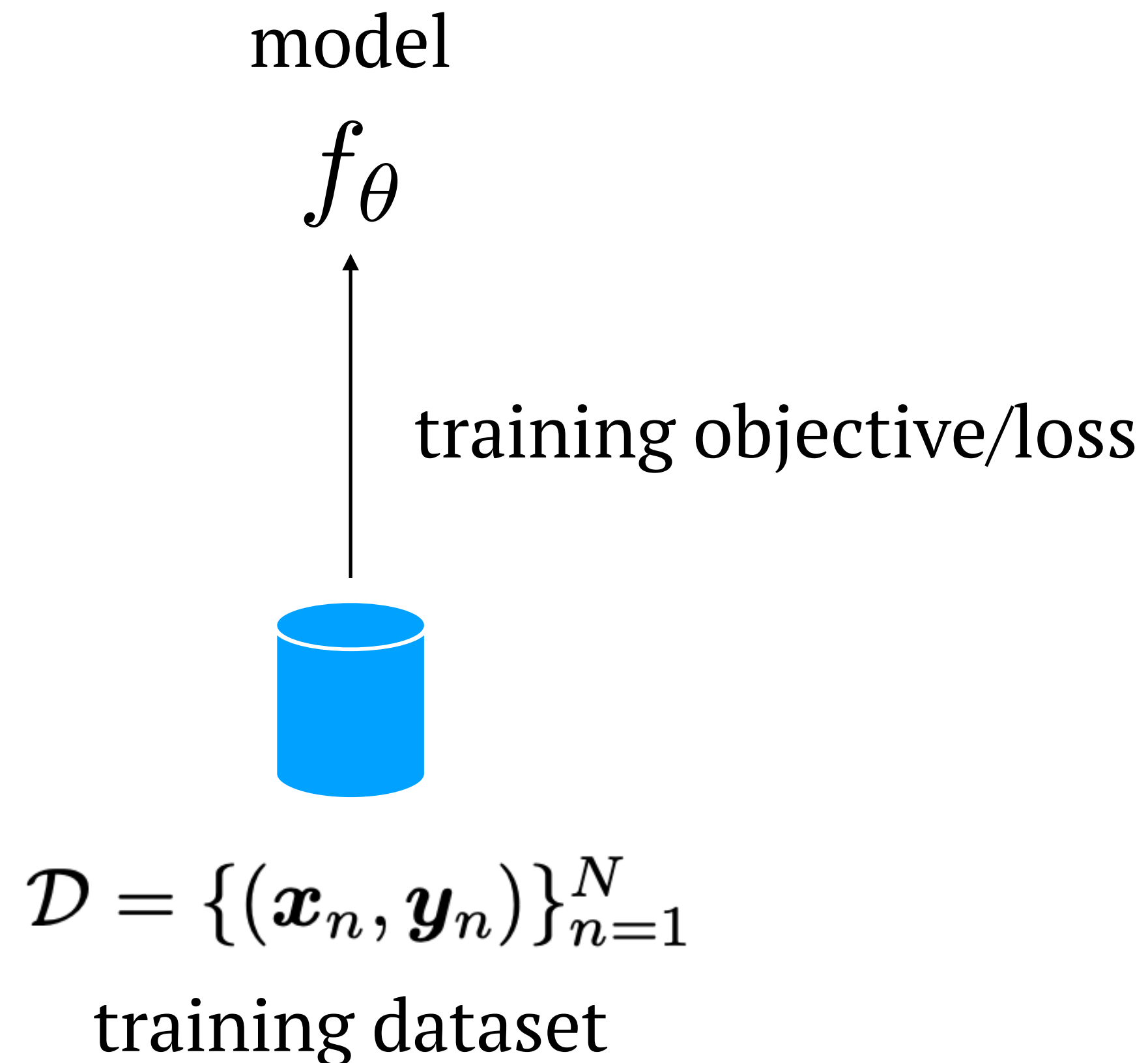


ChatGPT (Tokens Used 544)

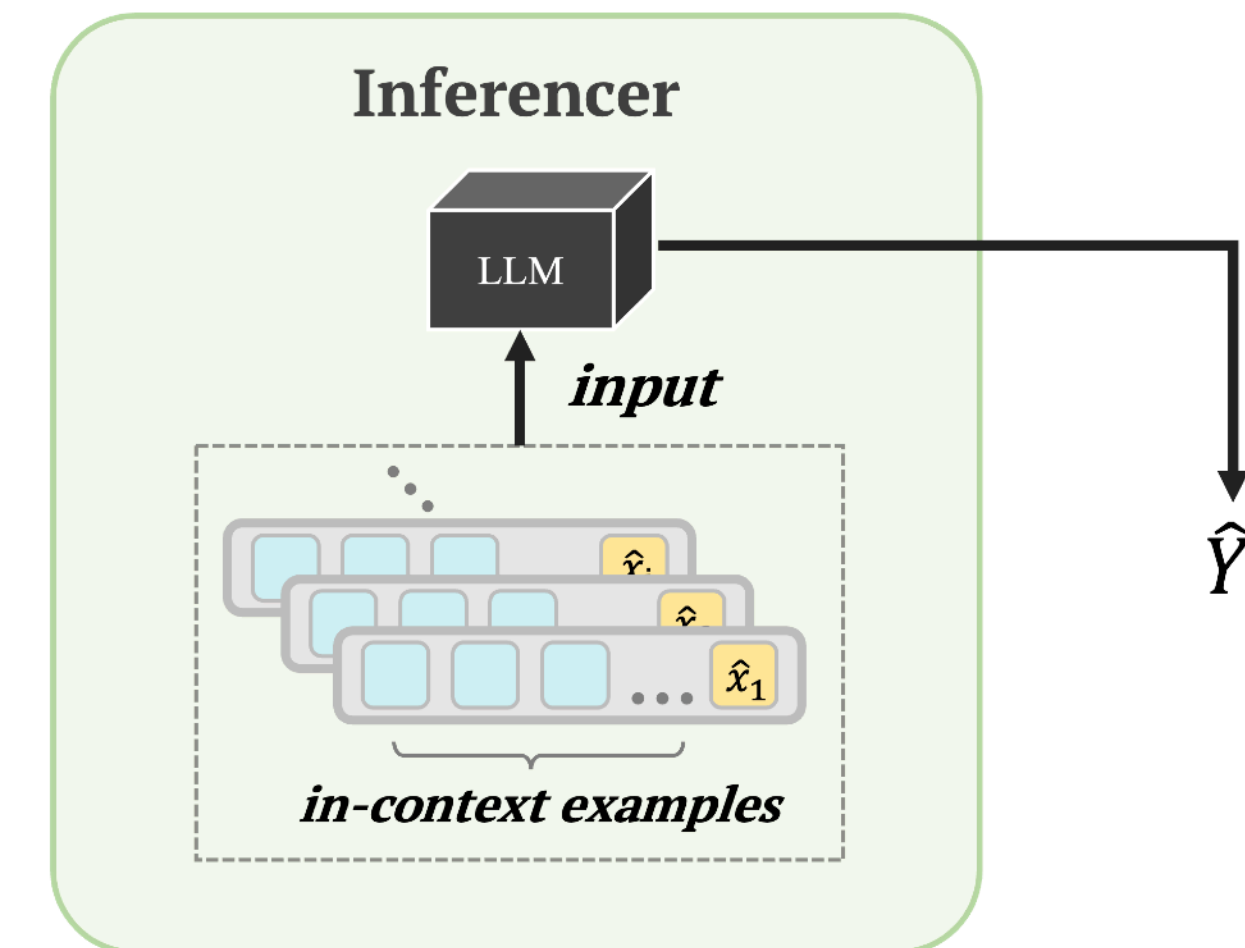
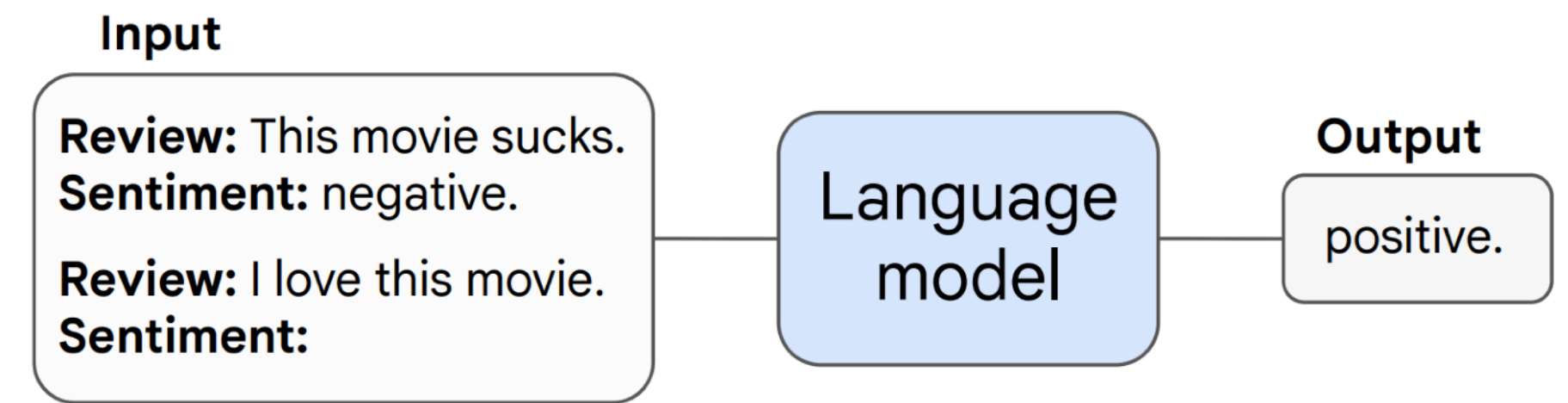
This email is likely a solicitation for self-publishing services, which may or may not be of interest to authors. However, as it is unsolicited and potentially sent to a large number of recipients, it could be considered spam.

28/08/2023, 10:44:17

# Machine Learning (Today)



Supervised Learning



In-Context Learning



# Powerful Large Language Models

Kylar went to the store to buy glasses for his new apartment. One glass costs \$5, but every second glass costs only 60% of the price. Kylar wants to buy 16 glasses. How much does he need to pay for them?



Steve decides to start eating more tomatoes and decides to grow his own cherry tomatoes. He eats twice as much as his girlfriend. He eats 6 per day. If a vine can produce 3 tomatoes per week how many vines does he need?





# Powerful Large Language Models

Kylar went to the store to buy glasses for his new apartment. One glass costs \$5, but every second glass costs only 60% of the price. Kylar wants to buy 16 glasses. How much does he need to pay for them?

 ChatGPT Poe

To calculate the total amount Kylar needs to pay for the glasses, we can break down the purchase into two parts: the first glass and the rest of the glasses.

The first glass costs \$5. For the remaining 15 glasses, every second glass costs only 60% of the price, which means they cost  $60\% * \$5 = \$3$  each.

Let's calculate the total cost:

First glass: \$5

Remaining glasses (15 glasses):  $15 \text{ glasses} * \$3/\text{glass} = \$45$

Total cost:  $\$5 + \$45 = \$50$

Therefore, Kylar needs to pay \$50 for the 16 glasses.

Steve decides to start eating more tomatoes and decides to grow his own cherry tomatoes. He eats twice as much as his girlfriend. He eats 6 per day. If a vine can produce 3 tomatoes per week how many vines does he need?

 ChatGPT Poe

To determine the number of vines Steve needs to meet his tomato consumption, we can calculate the weekly tomato requirement and then divide it by the number of tomatoes produced by each vine per week.

Steve eats 6 cherry tomatoes per day, which means he consumes  $6 * 7 = 42$  tomatoes per week.

If a vine can produce 3 tomatoes per week, Steve would need  $42 \text{ tomatoes} / 3 \text{ tomatoes per vine} = 14$  vines.

Therefore, Steve would need 14 cherry tomato vines to meet his weekly tomato consumption.

# Powerful Large Language Models



	GPT-3.5 (ChatGPT)	GPT-4	Llama2
GSM8K (8-shot)	57.1	92.0	56.8

71.7 (ChatGPT as of Aug 30, 2023, zero-shot)

An arrow pointing from the text "71.7 (ChatGPT as of Aug 30, 2023, zero-shot)" to the value "57.1" in the table.

# Probabilistic Machine Learning

There are two main reasons we adopt a probabilistic approach.

1. It is the optimal approach to **decision making under uncertainty**.
2. Probabilistic modeling is the language used by most other areas of science and engineering, and thus provides a unifying framework between these fields.





# Probabilistic Machine Learning

Is it a lot of math?

$$\ln \left[ \lim_{z \rightarrow \infty} \left( 1 + \frac{1}{z} \right)^z \right] + (\sin^2 x + \cos^2 x) = \sum_{n=0}^{\infty} \frac{\cosh(y) \sqrt{1 - \tanh^2 y}}{2^n}$$

## How To Be More Impressive

Unknown

Suppose we want to publish something that is as simple as

$$1 + 1 = 2 \quad (1)$$

This is not very impressive. If we want our article to be accepted by IEEE reviewers, we have to more abstract. So, we could complicate the left hand side of the expression by using

$$\ln(e) = 1 \quad \text{and} \quad \sin^2 x + \cos^2 x = 1$$

and the right hand side can be stated as

$$2 = \sum_{n=0}^{\infty} \frac{1}{2^n}.$$

Therefore, Equation (1) can be expressed more scientifically as:

$$\ln(e) + (\sin^2 x + \cos^2 x) = \sum_{n=0}^{\infty} \frac{1}{2^n} \quad (2)$$

which is far more impressive. However, we should not stop here. The expression can be further complicated by using

$$e = \lim_{z \rightarrow \infty} \left( 1 + \frac{1}{z} \right)^z \quad \text{and} \quad 1 = \cosh(y) \sqrt{1 - \tanh^2 y}.$$

Equation (2) may therefore be written as

$$\ln \left[ \lim_{z \rightarrow \infty} \left( 1 + \frac{1}{z} \right)^z \right] + (\sin^2 x + \cos^2 x) = \sum_{n=0}^{\infty} \frac{\cosh(y) \sqrt{1 - \tanh^2 y}}{2^n} \quad (3)$$

**Note:** Other methods of a similar nature could also be used to enhance our prestige, once we grasp the underlying principles.

# Probabilistic Machine Learning

## Prelude I

In C major

BWV 846

Johann Sebastian Bach (1685 - 1750)



Is it a lot of math?

# Probabilistic Machine Learning



Input 1	Input 2		Output
1	2		5
2	3	$f(x_1, x_2) \rightarrow y$	8
3	6		15
4	1		6

Inference



# Probabilistic Machine Learning



**J.P.Morgan**  
North America Equity Research

**Kraft Foods Group**  
This Was the Rorschach Test of Quarters -- Staying Neutral as Uncertainty Still Reigns

Kraft reported Q412 yesterday and provided sufficient fodder for a variety of interpretations to be made. Bulls can point to an improving gross margin and solid organic sales growth excluding SKU pruning. Bears can point to the inventory load and the muted 0.4% sales growth figure expressing. Our view is that while the quarter was probably more or less in line with expectations, we find it difficult to draw solid conclusions given how many moving pieces were floating around. This uncertainty, plus the fact that the Street is already well ahead of guidance for next year (we think the 32.70 consensus may need to come down), lead us to maintain our Neutral rating.

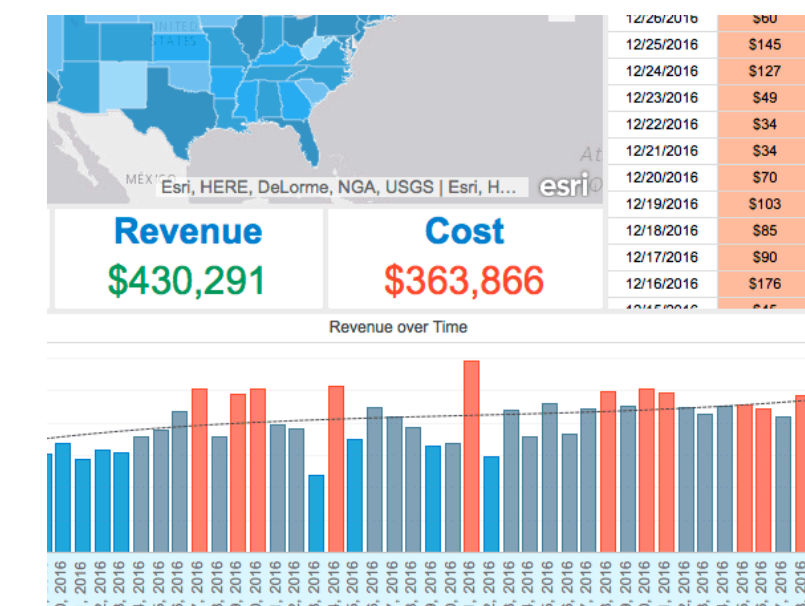
- **What the bulls may have seen.** 1) The top line beat, up 3.0% versus the Street's 1.6% forecast. 2) The gross margin was much better than expected, at 33.3% excluding hedging gains versus last year's 30.9%. 3) Guidance was maintained for next year, suggesting that management's visibility remains clear. 4) Volume and pricing rose simultaneously. 5) The quality of the quarter seemed solid, with ad spending and the SG&A margin up.

**Neutral**  
KRFT, KRFT US  
Price: \$44.07  
Price Target: \$47.00

Food Manufacturing and Retail

J.P. Morgan Securities LLC

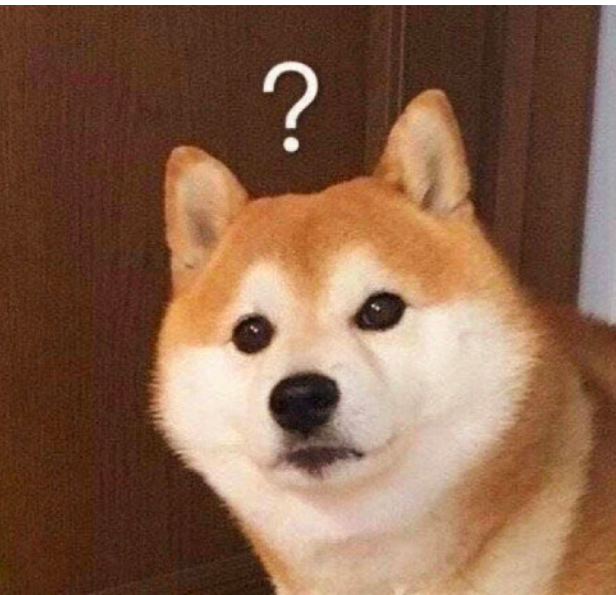
$$f(x) \rightarrow y$$



Inference

# Probabilistic Machine Learning



Input 1	Input 2	$w_1x_1 + w_2x_2 \rightarrow y$	Output
1	2		5.1
2	3		8.2
3	6		14.7
4	1		5.6



# Probabilistic Machine Learning



Input 1

1

2

3

4

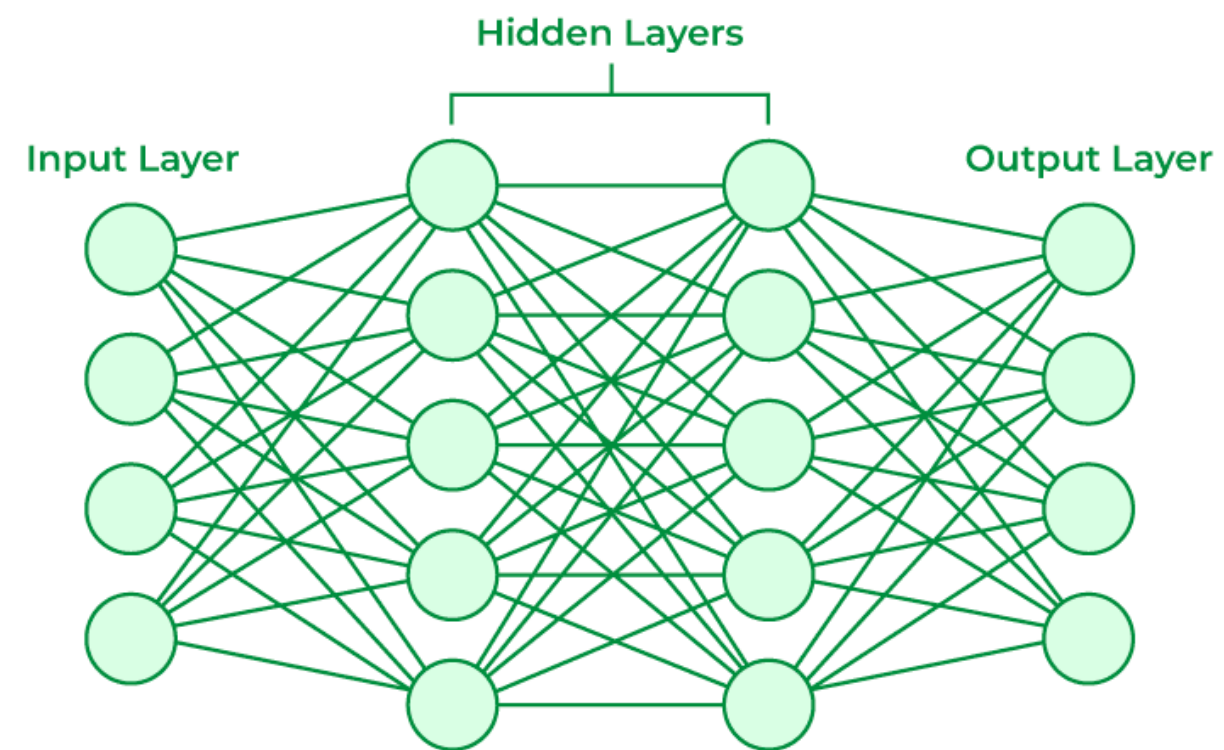
Input 2

2

3

6

1



Output

5.1

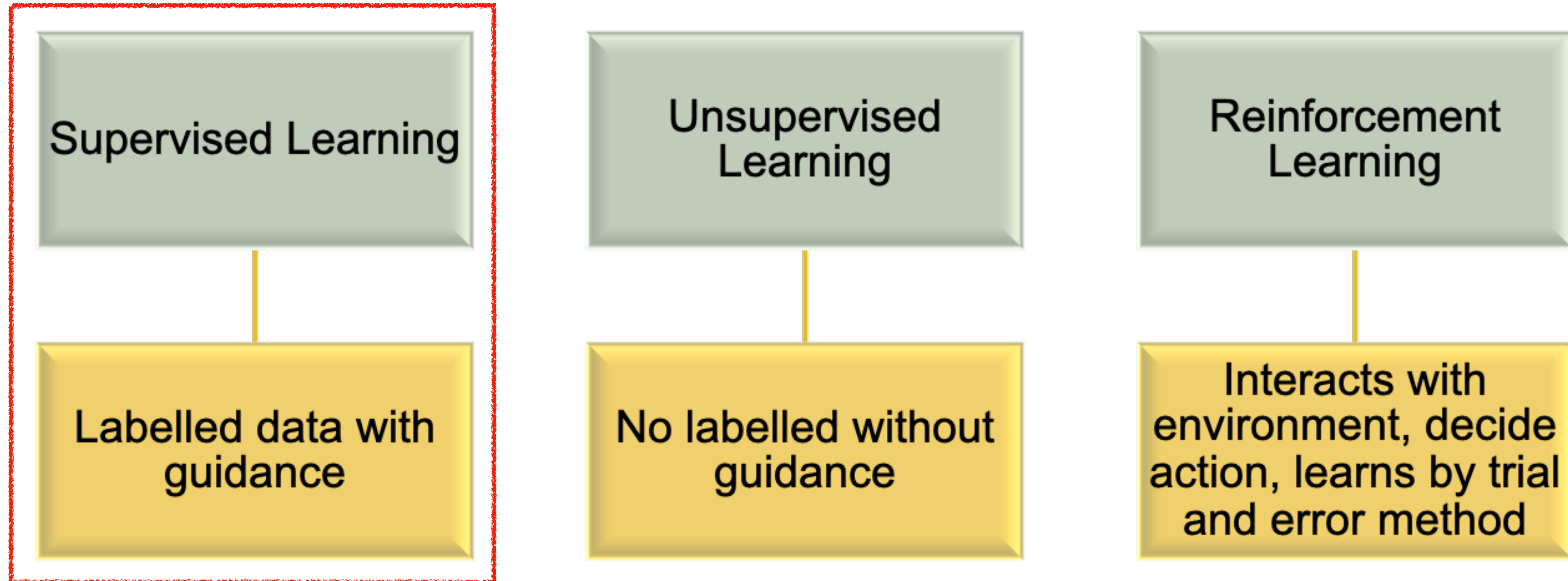
8.2

14.7

5.6

Training / Estimating

# Machine Learning Problems





# Supervised Learning – Classification



(a)



(b)



(c)

Three types of Iris flowers: Setosa, Versicolor and Virginica.

index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
...	...	...	...	...	...
50	7.0	3.2	4.7	1.4	Versicolor
...	...	...	...	...	...
149	5.9	3.0	5.1	1.8	Virginica

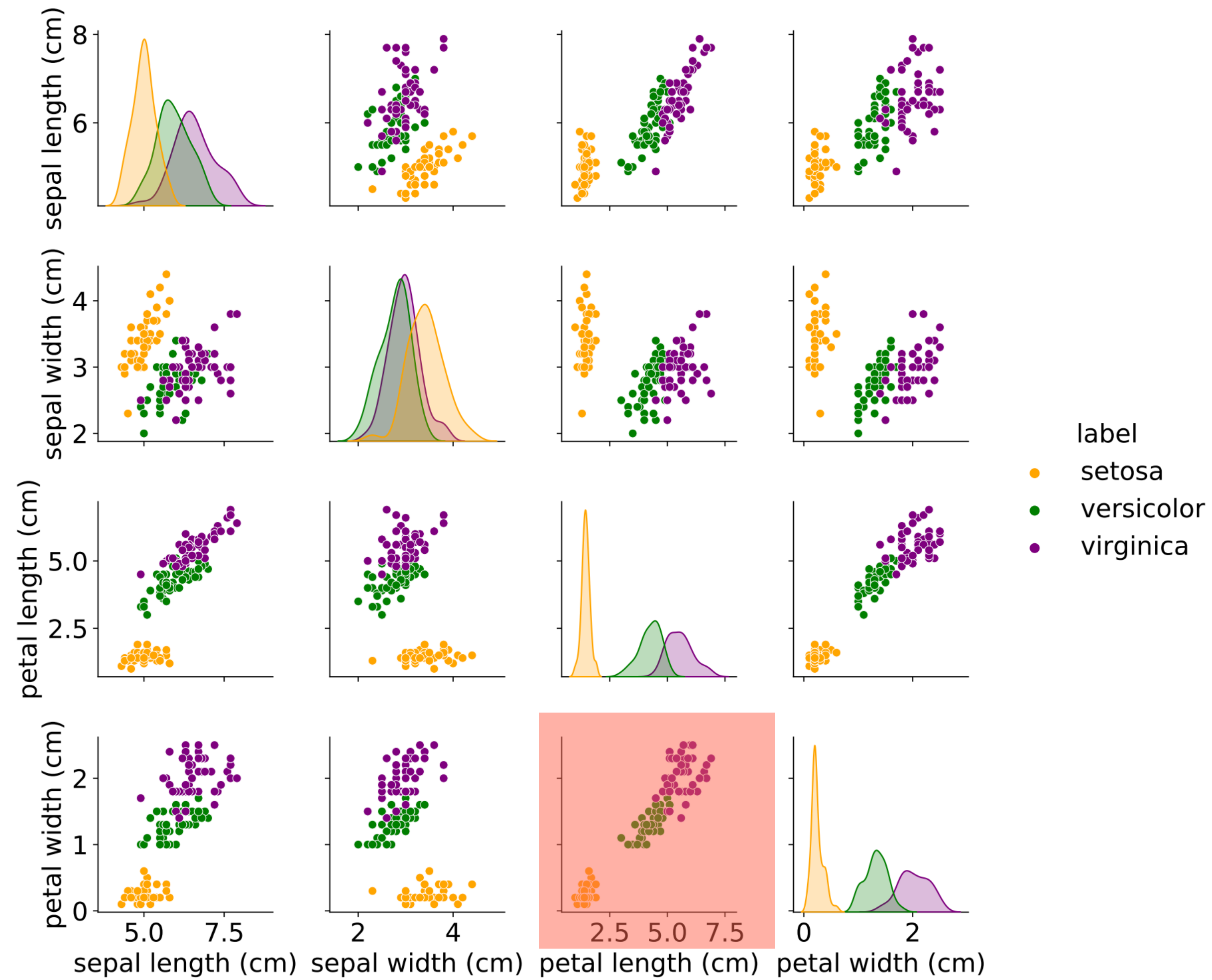
sl: sepal length  
sw: sepal width  
pl: petal length  
pw: petal width

*input space,  $\mathcal{X}$  = numeric features*

$$\mathcal{X} = \mathbb{R}^4$$

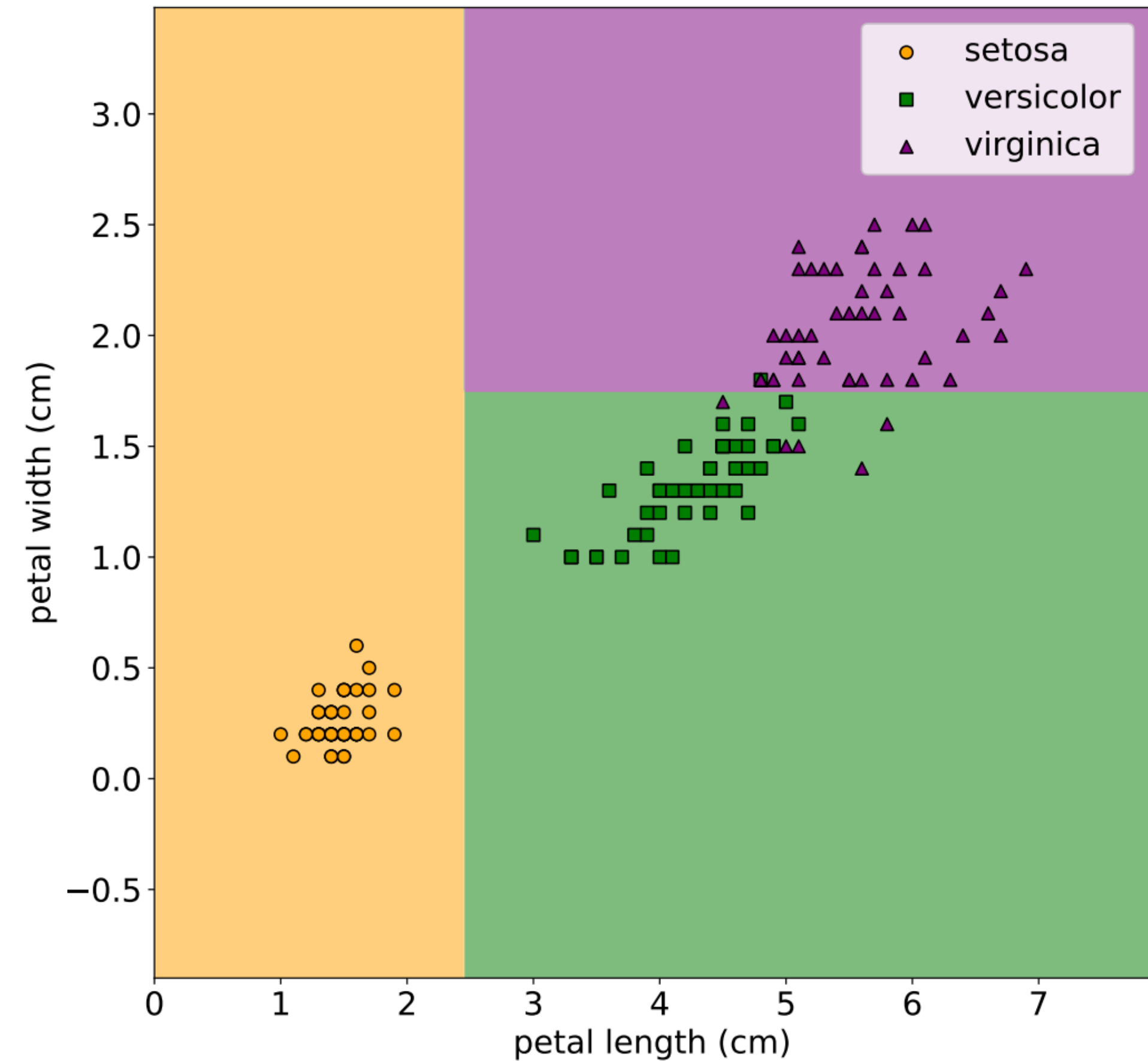
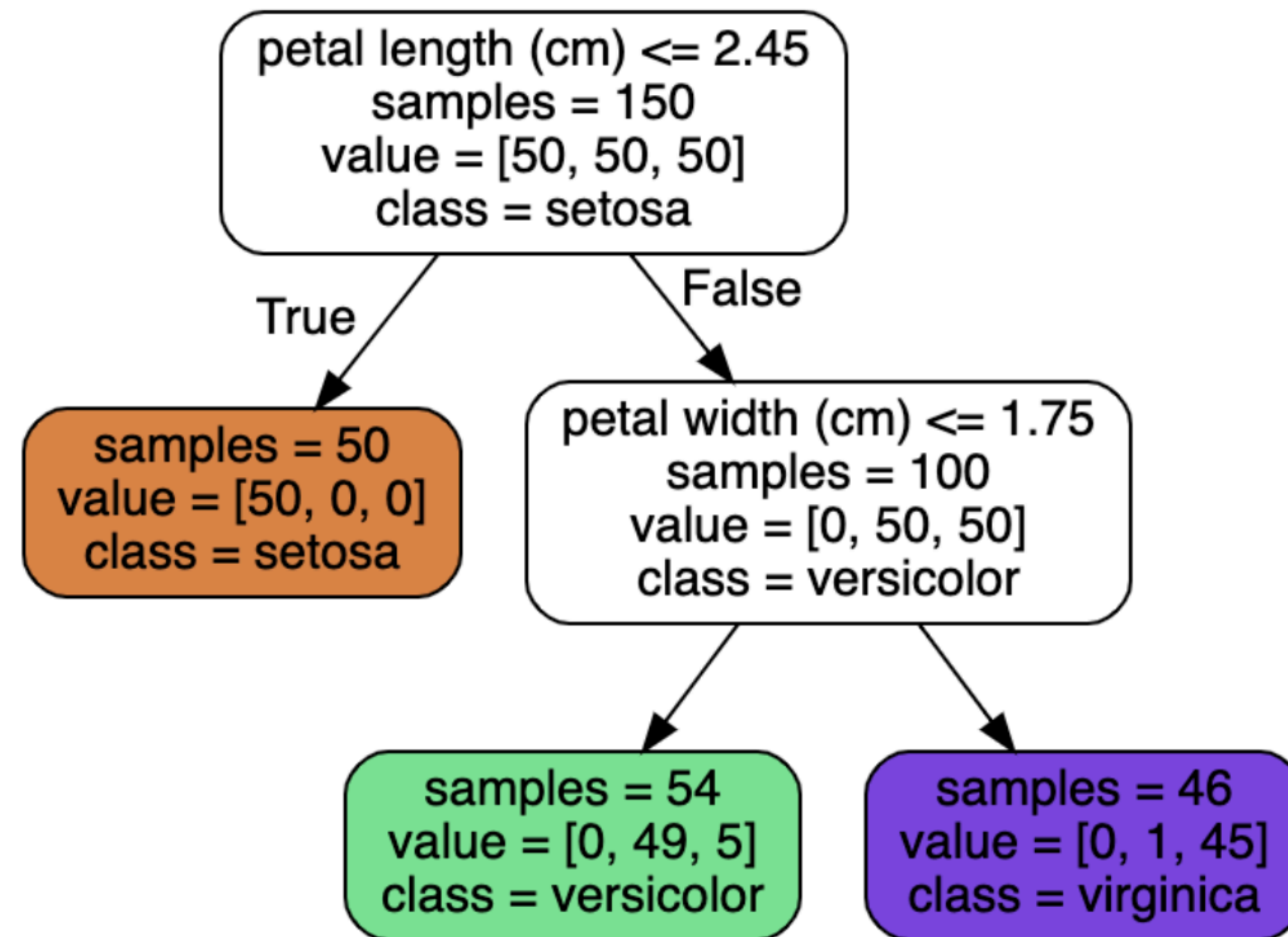
$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

# Supervised Learning — Classification





# Supervised Learning – Classification



# Supervised Learning – Classification

Empirical Risk Minimization (ERM):

$$\hat{\theta} = \operatorname{argmin}_{\theta} \mathcal{L}(\theta) = \operatorname{argmin}_{\theta} \frac{1}{N} \sum_{n=1}^N \ell(y_n, f(\mathbf{x}_n; \theta))$$

$$\mathcal{L}(\theta) \triangleq \frac{1}{N} \sum_{n=1}^N \mathbb{I}(y_n \neq f(\mathbf{x}_n; \theta))$$

$$\mathbb{I}(e) = \begin{cases} 1 & \text{if } e \text{ is true} \\ 0 & \text{if } e \text{ is false} \end{cases}$$



# Supervised Learning – Classification

Uncertainty: using conditional probability distribution

$$p(y = c | \mathbf{x}; \boldsymbol{\theta}) = \mathcal{S}_c(f(\mathbf{x}; \boldsymbol{\theta}))$$

Constraints:

$$0 \leq p(y = c) \leq 1$$

$$\sum_c p(y = c) = 1$$

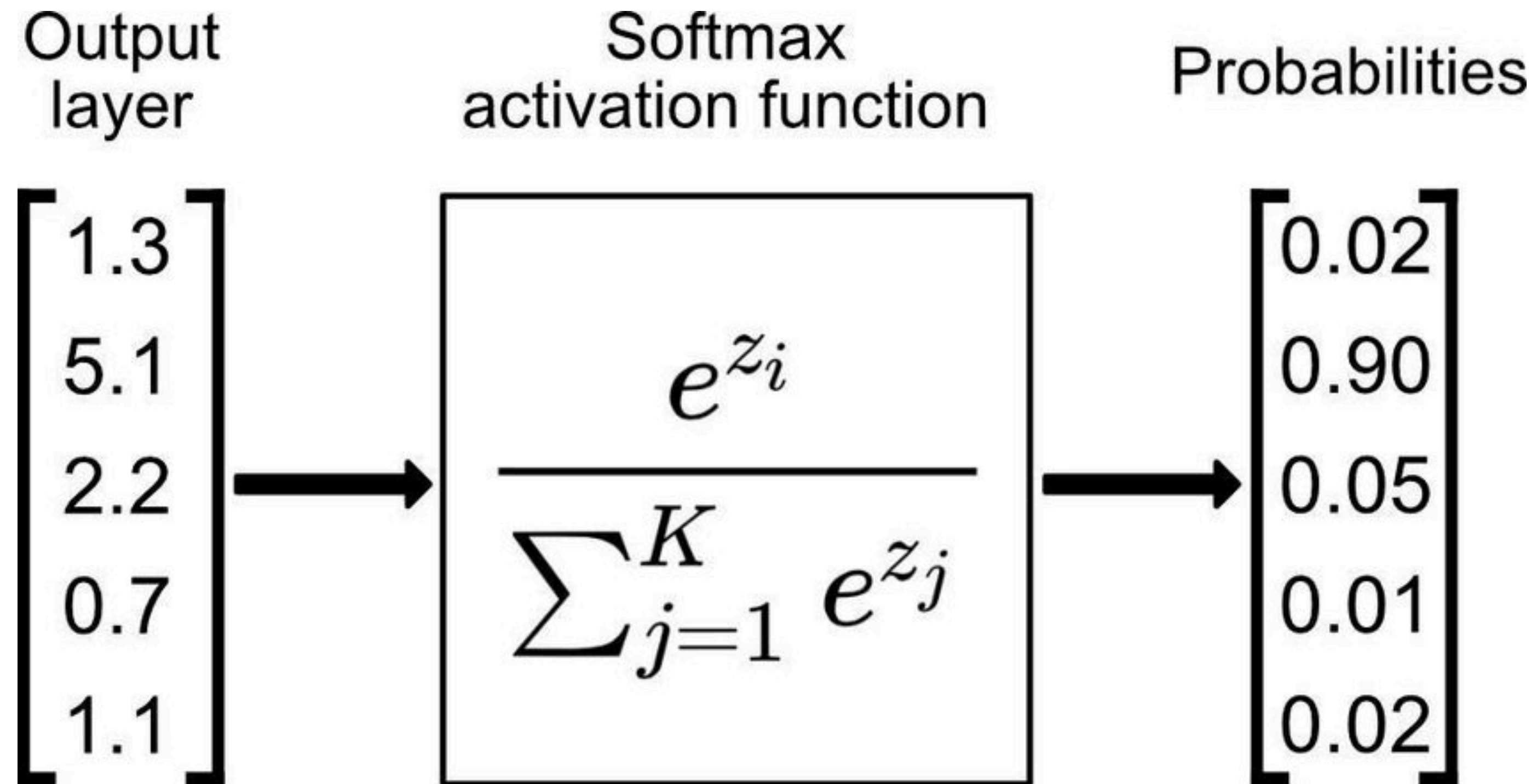
Softmax Function:

$$\mathcal{S}(\mathbf{a}) \triangleq \left[ \frac{e^{a_1}}{\sum_{c'=1}^C e^{a_{c'}}}, \dots, \frac{e^{a_C}}{\sum_{c'=1}^C e^{a_{c'}}} \right]$$

Affine Function:

$$f(\mathbf{x}; \boldsymbol{\theta}) = b + \mathbf{w}^\top \mathbf{x} = b + w_1 x_1 + w_2 x_2 + \dots + w_D x_D$$

# Softmax Function





# Supervised Learning – Classification

Maximum Likelihood Estimation

Minimizing Negative Log Likelihood

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \operatorname{NLL}(\boldsymbol{\theta})$$

$$\operatorname{NLL}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^N \log p(y_n | f(\mathbf{x}_n; \boldsymbol{\theta}))$$

# Course Logistics



# Course Logistics

Website:

<https://nlp.cs.hku.hk/comp3314>

Prerequisites:

MATH1853 or MATH2101; and COMP2119 or ELEC2543 or FITE2020



This ML course will be very difficult if you haven't taken these courses.

Assessment:

50% continuous assessment, 50% examination (final exam) —(note parallel sessions)

TAs:

Chang Ma [changma AT connect.hku.hk](mailto:changma AT connect.hku.hk)

Xie Zhang [zhangxie AT connect.hku.hk](mailto:zhangxie AT connect.hku.hk)

Zhiheng Lyu [cogito AT connect.hku.hk](mailto:cogito AT connect.hku.hk)

# Course Logistics

Textbook recommendation (PML):

*Probabilistic Machine Learning, An Introduction*  
MIT Press, February 2022.

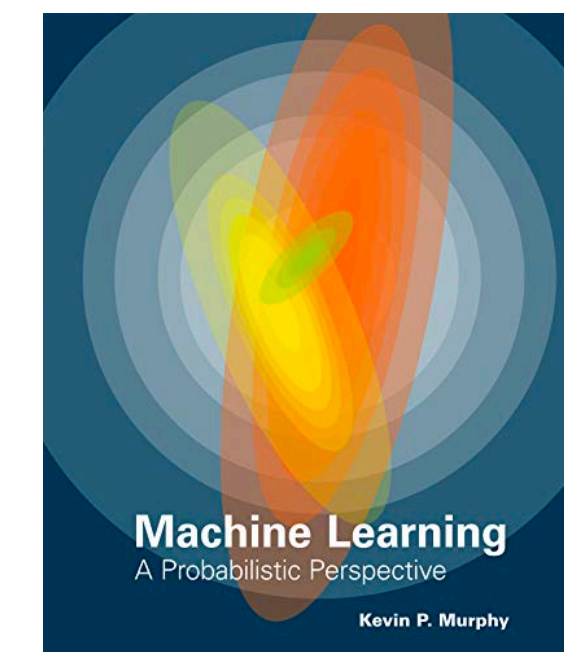
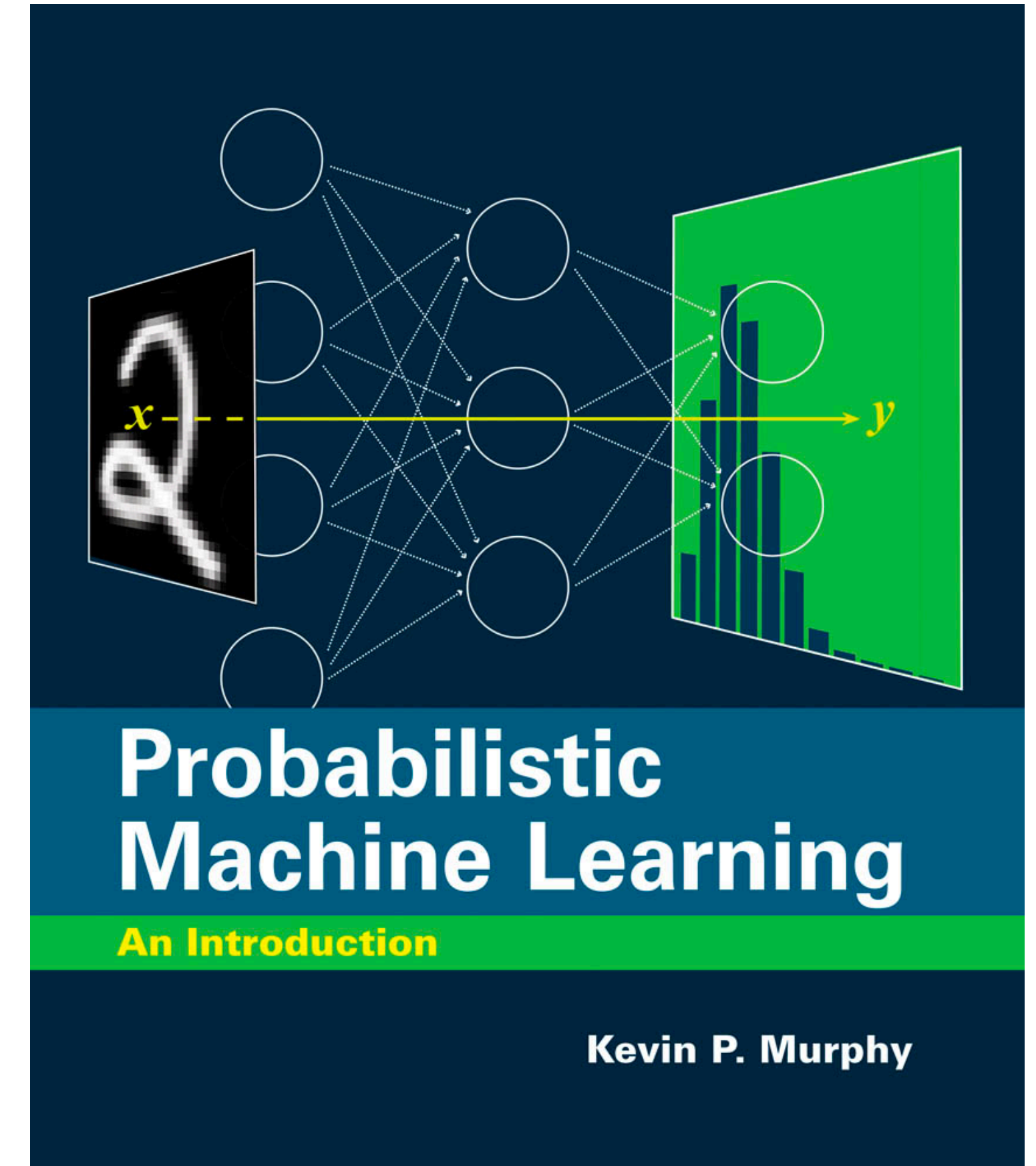
Assessments (in total ~4):

Programming problems

Problem sets

Honor code:

You are free to form study groups and discuss homeworks and projects. However, you must write up homeworks and code from scratch independently, and you must acknowledge in your submission all the students you discussed with.



(MLAPP)  
Highly Recommended!

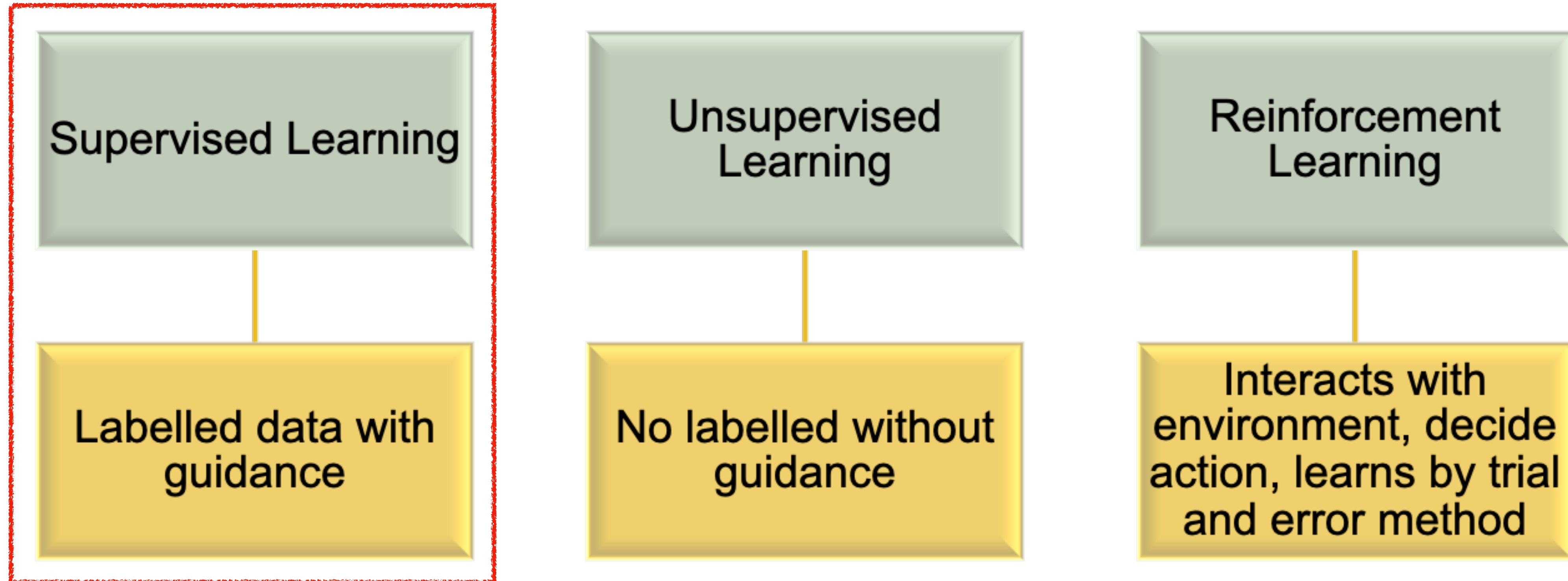
# Course Logistics

- Reinforcement Learning (RL)
  - Consider taking: comp3270 AI
- Deep Learning (CNN, RNN, GAN, Deep RL)
  - Consider taking: comp3340 Applied deep learning, and/or
  - Consider taking: comp3317 Computer vision
- To work on bigger applied projects
  - Consider taking: comp3359 AI applications, and/or
  - Consider taking: comp3414 Experiential learning on AI and robotics
- R language with a focus on statistical modeling
  - Consider taking: comp3354 Statistical learning





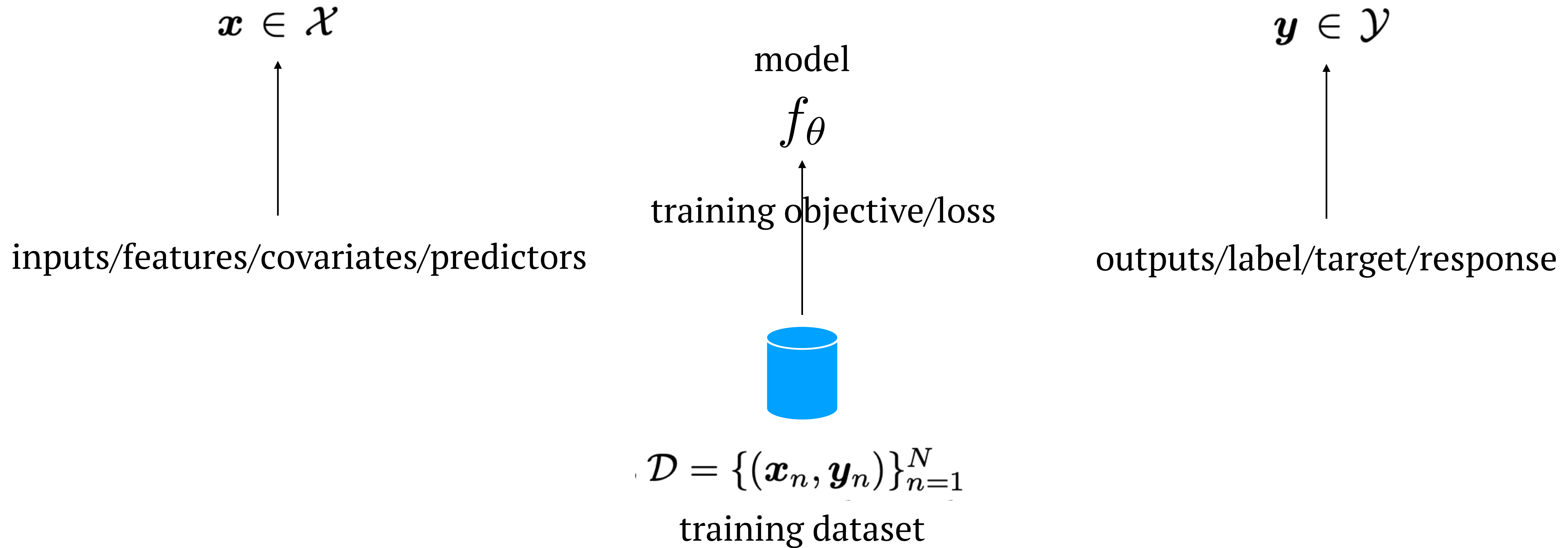
# Machine Learning Problems



# Typefaces

- Use lowercase Latin letters for scalar values;  $a, b, c$  preferentially for symbols,  $i, j, k, \ell, m, n$  preferentially for whole numbers/integers (prefer  $\ell$  to  $l$  since the latter is confusable for 1), and  $e, f, w, x, y, z$  for the values taken by random variables/predicted by models (be they symbols, whole numbers, or reals).
- To indicate a structured value, like a sequence or tree, use lowercase `boldsymbol` Latin letters. For example, the input sentence  $\boldsymbol{x}$ , and the predicted tree  $\boldsymbol{y}$ .
- To indicate a vector (column usually, but row if you are consistent), use lowercase `mathbf` Latin letters  $\mathbf{u}, \mathbf{v}$ . If you want to use Greek letters to indicate a vector (often used for parameters/hyperparameters), use lowercase `boldsymbol` Greek letters, e.g.  $\boldsymbol{\theta}_0, \boldsymbol{\varphi}$ , or  $\boldsymbol{\pi}$ .
- To indicate a matrix or higher-order tensor, use capital `mathbf` Latin letters, e.g.,  $\mathbf{U}$ , or capital `boldsymbol` Greek letters when convention dictates, e.g.,  $\boldsymbol{\Sigma}^{-1}$ .
- An element of a vector, a sequence, or matrix is (usually) a scalar, so its typeface should reflect this. So we write that  $v_i$  is the  $i$ th element of  $\mathbf{v}$ ,  $m_{ij}$  (or perhaps  $m_{i,j}$ ) is an element of  $\mathbf{M}$ , and  $w_1$  is the first word in sentence  $\boldsymbol{w}$ . Similarly, slices of sequences are themselves sequences, so we write  $\boldsymbol{w}_{\leq t}$  and  $\boldsymbol{z}_{-i}$  (this convention is used to designate all of the values in  $\boldsymbol{z}$  but the  $i$ th).
- Sets of structured values are often  $\mathcal{X}$  or  $\mathcal{Y}$ . If you have sets of symbols (or lots of sets, e.g., in a formal language paper), using capital Latin letters is the standard convention, e.g.,  $A, B, C$ . Standard sets like  $\mathbb{R}, \mathbb{Z}$ , and  $\Sigma$  have their own conventions that should be followed. Note that  $\mathbb{R}$  is strongly preferable to  $\mathfrak{R}$  or  $R$  when denoting real values.

# Supervised Learning



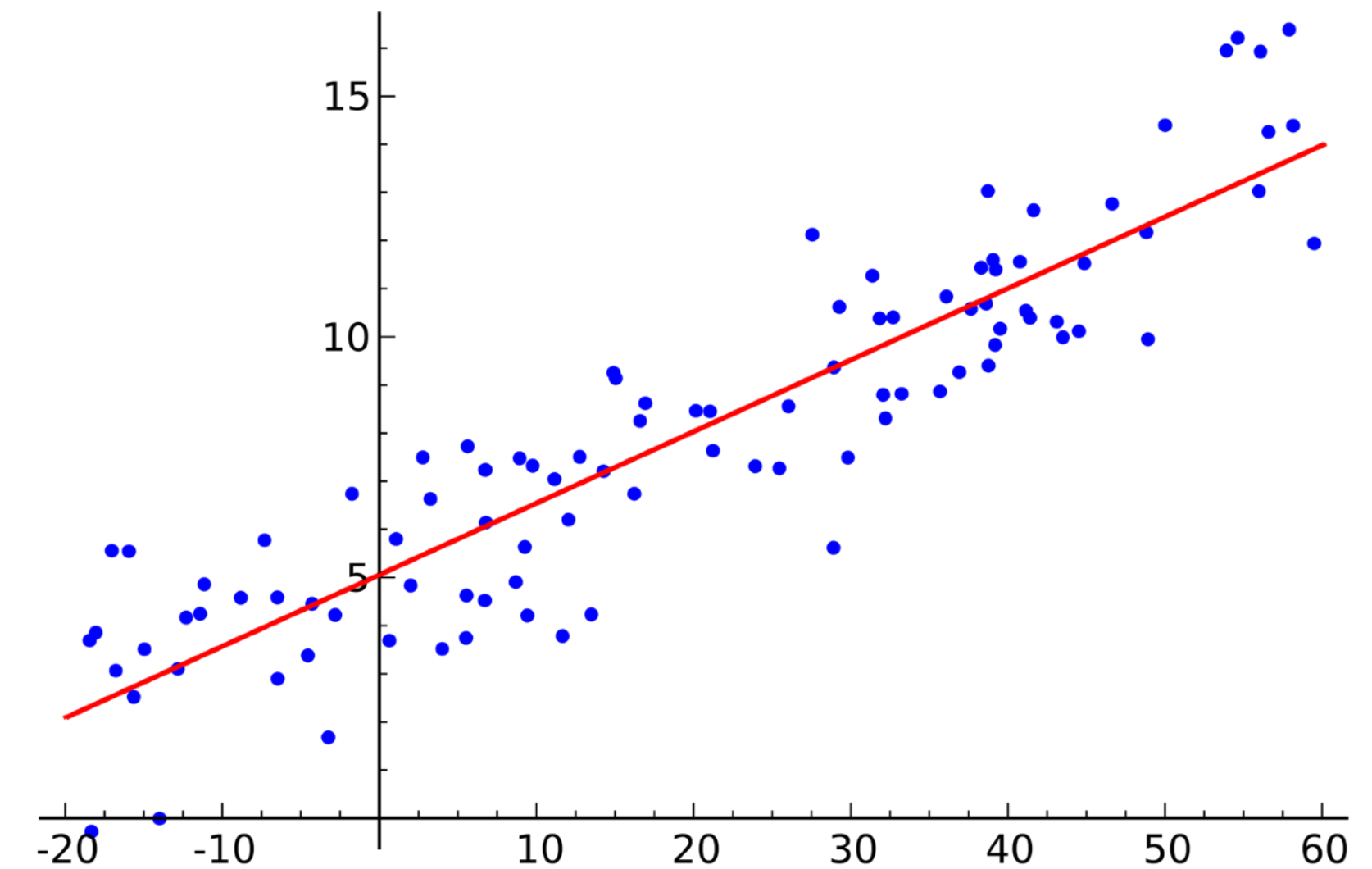


# Supervised Learning — Regression

Output = Real-value -> Quadratic loss

$$\ell_2(y, \hat{y}) = (y - \hat{y})^2$$

$$\text{MSE}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^N (y_n - f(\mathbf{x}_n; \boldsymbol{\theta}))^2$$



# Supervised Learning – Regression

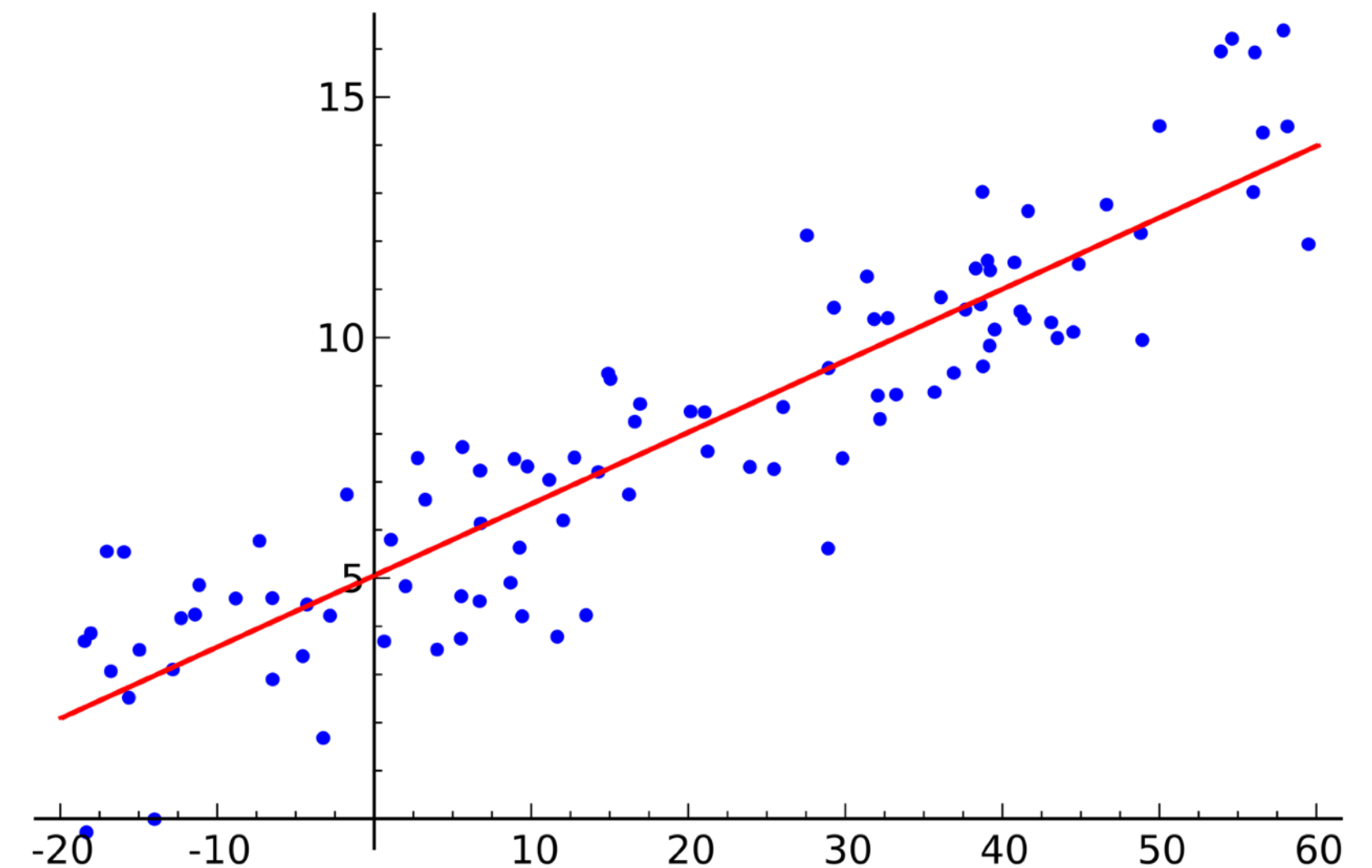
Uncertainty: Assume output distribution = Gaussian

$$\mathcal{N}(y|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

$$p(y|\mathbf{x}; \boldsymbol{\theta}) = \mathcal{N}(y|f(\mathbf{x}; \boldsymbol{\theta}), \sigma^2)$$

*assume variance is fixed (for simplicity):*

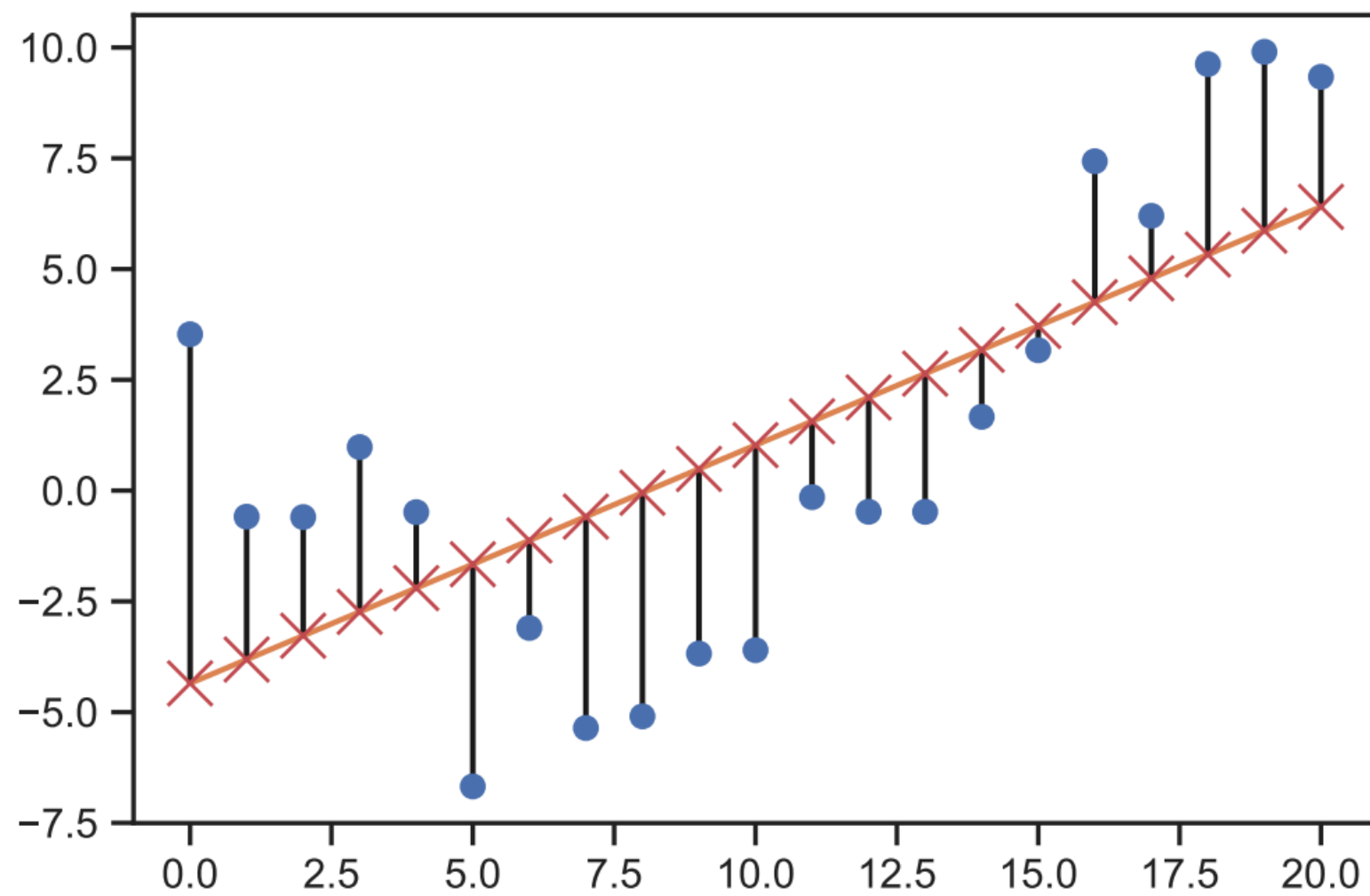
$$\begin{aligned} \text{NLL}(\boldsymbol{\theta}) &= -\frac{1}{N} \sum_{n=1}^N \log \left[ \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp \left( -\frac{1}{2\sigma^2} (y_n - f(\mathbf{x}_n; \boldsymbol{\theta}))^2 \right) \right] \\ &= \frac{1}{2\sigma^2} \text{MSE}(\boldsymbol{\theta}) + \text{const} \end{aligned}$$



# Supervised Learning – Regression

Linear Regression

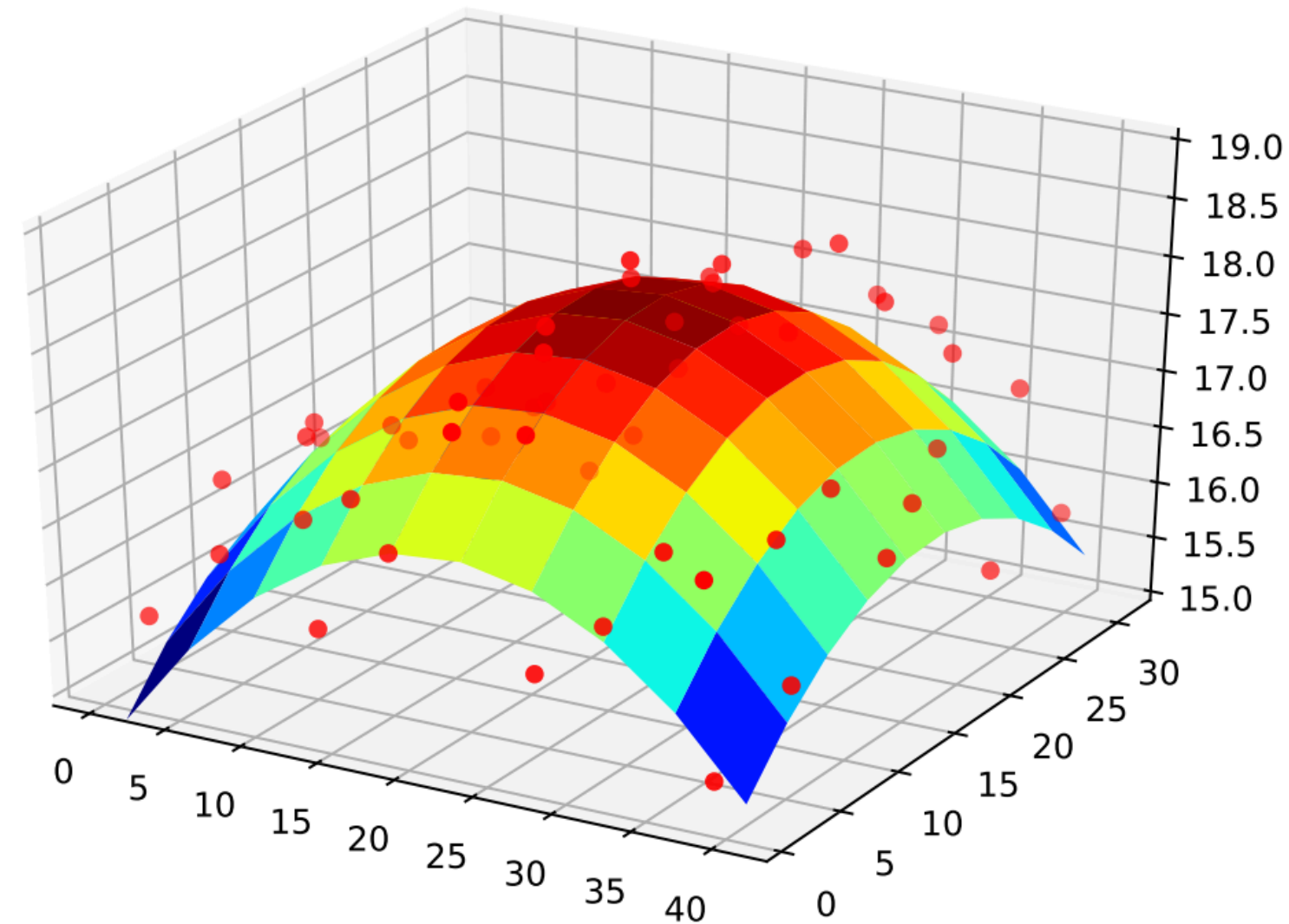
1 Feature



$$f(x; \theta) = b + wx$$

Polynomial Regression

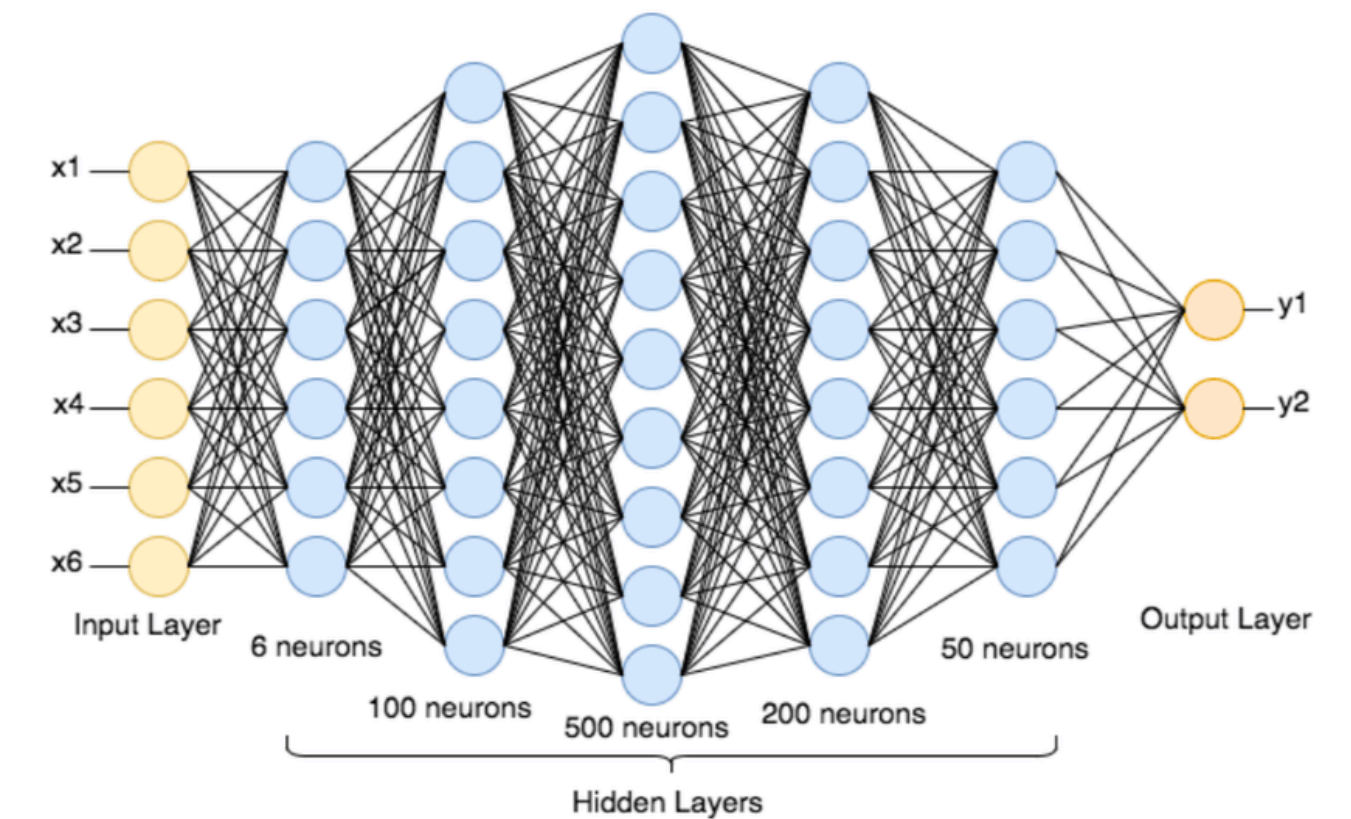
Features Engineering



$$f(x; \mathbf{w}) = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2$$

Deep Neural Network

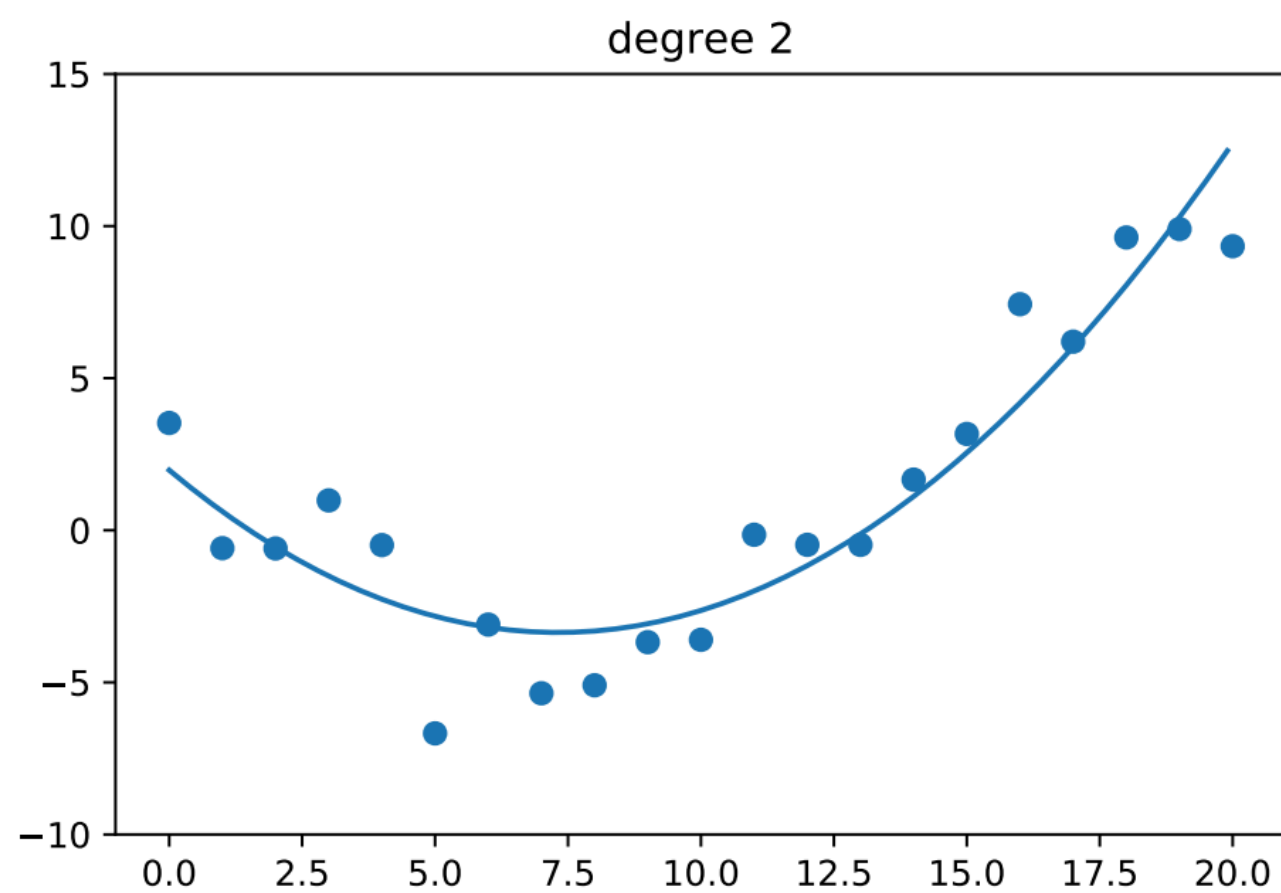
Feature Extraction Automatically



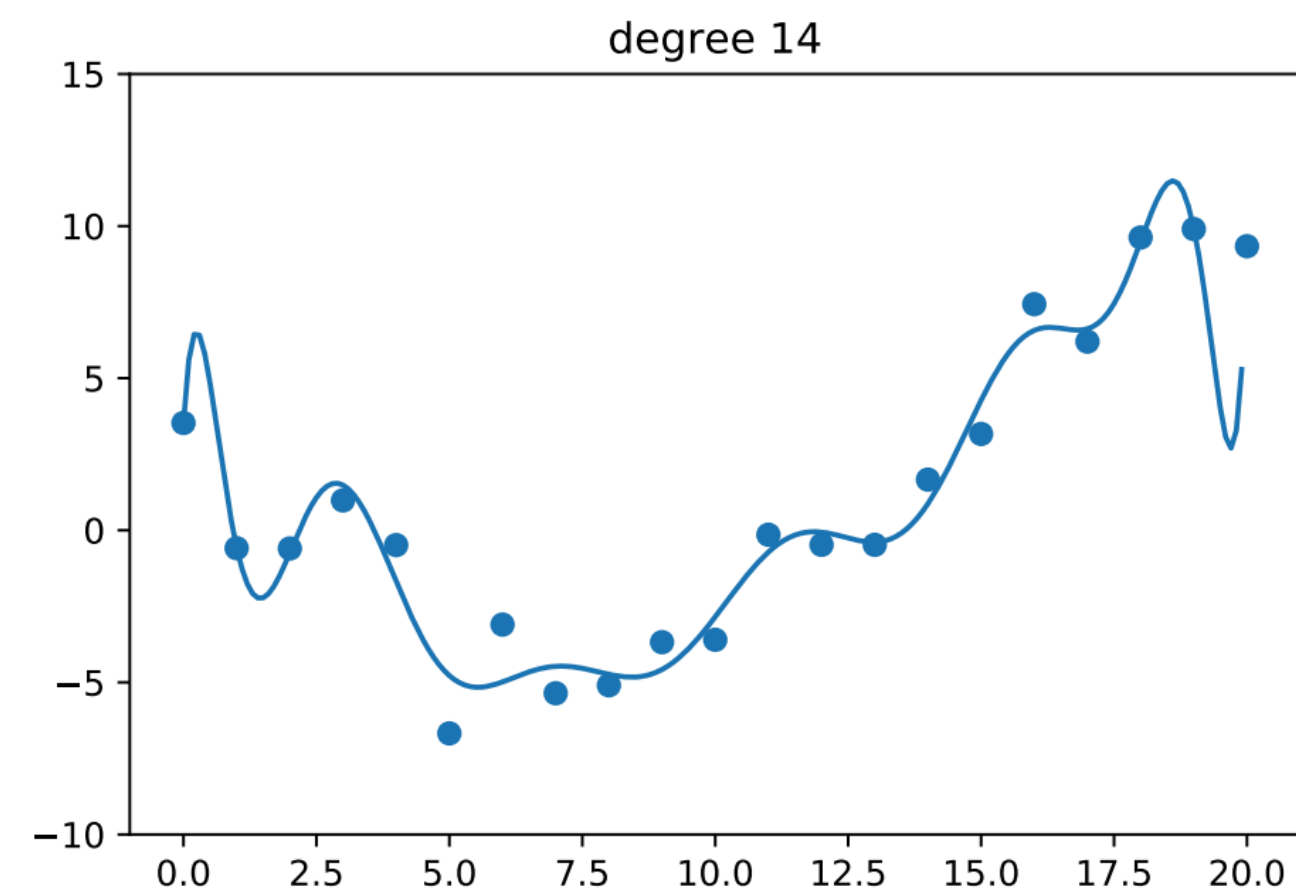
$$f(x; \mathbf{w}, \mathbf{V}) = \mathbf{w}^T \phi(x; \mathbf{V})$$



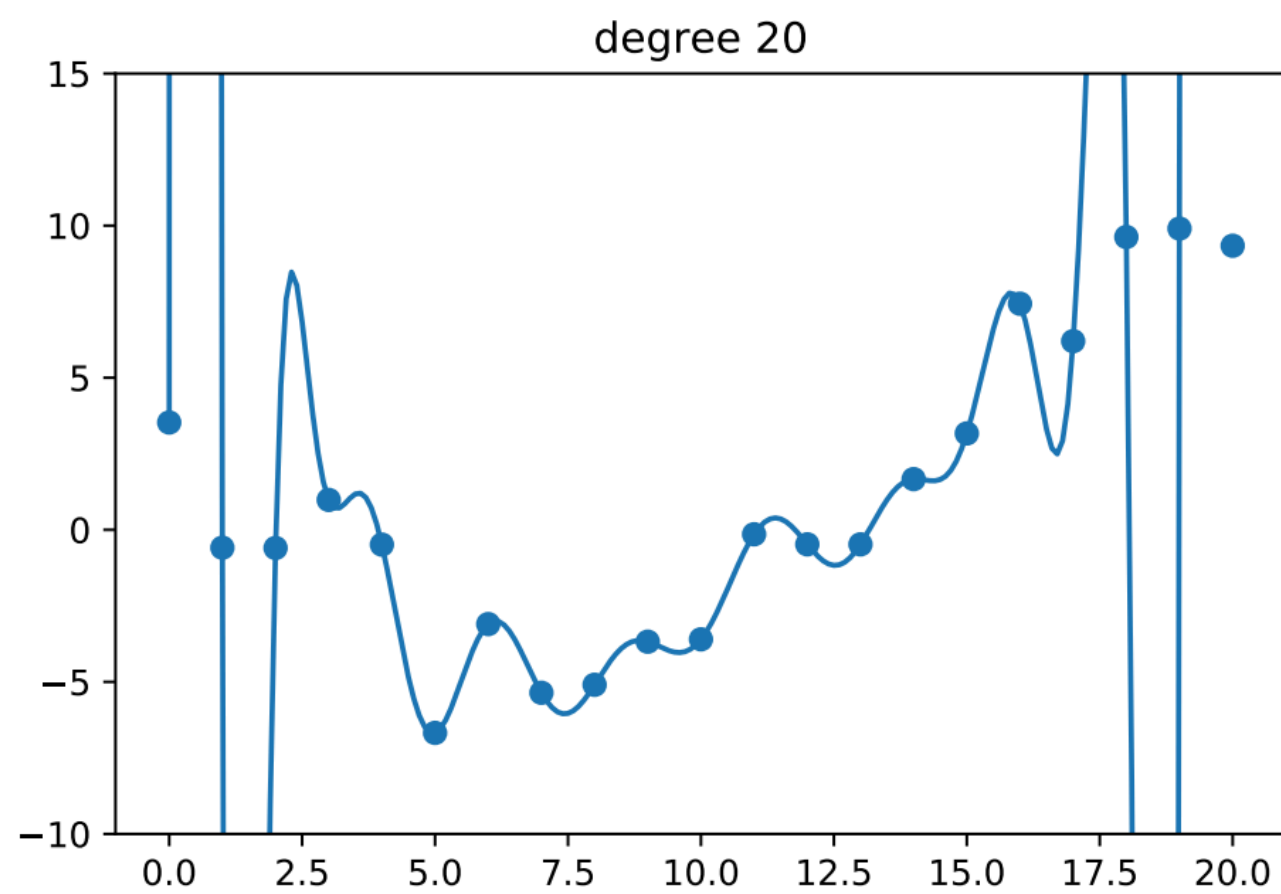
# Generalization



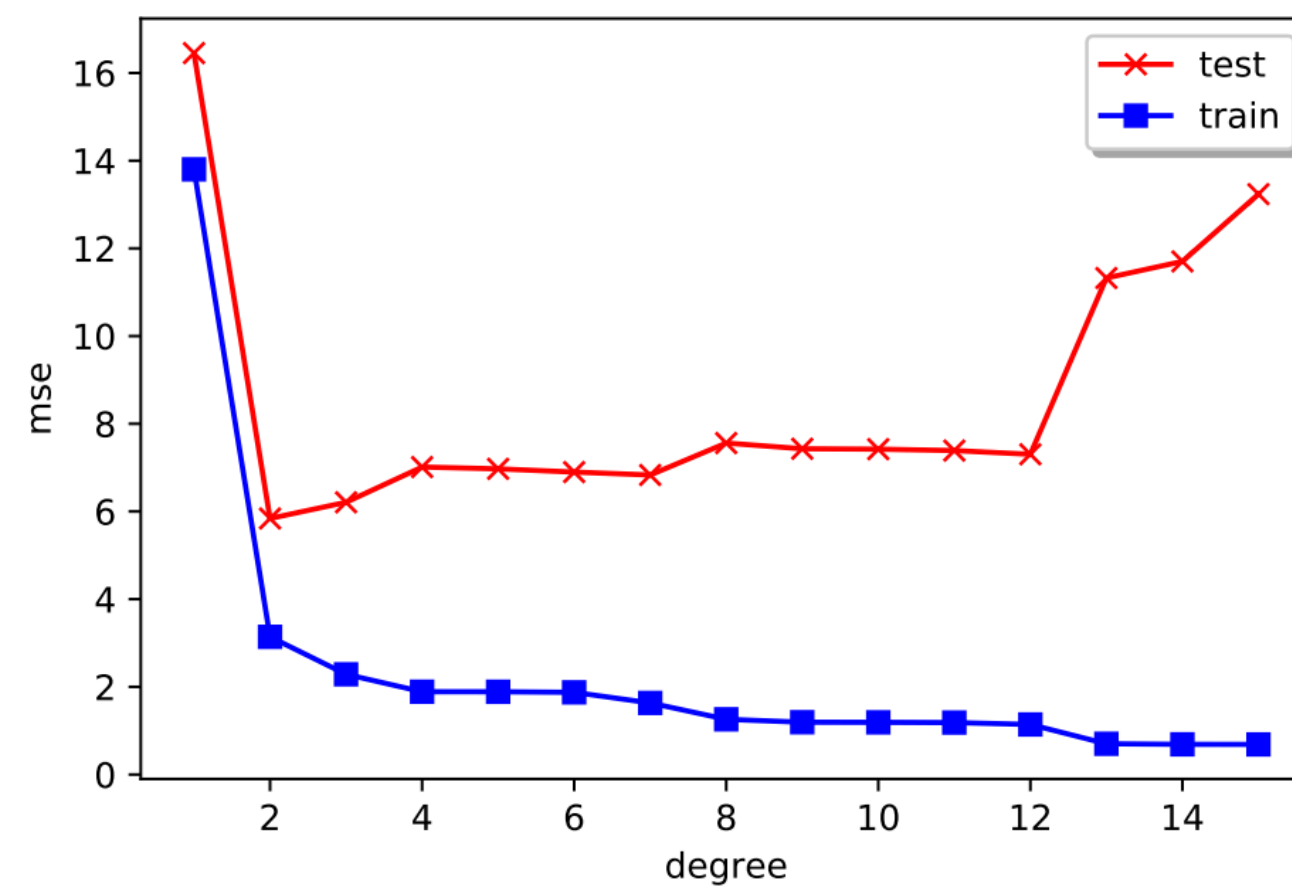
(a)



(b)

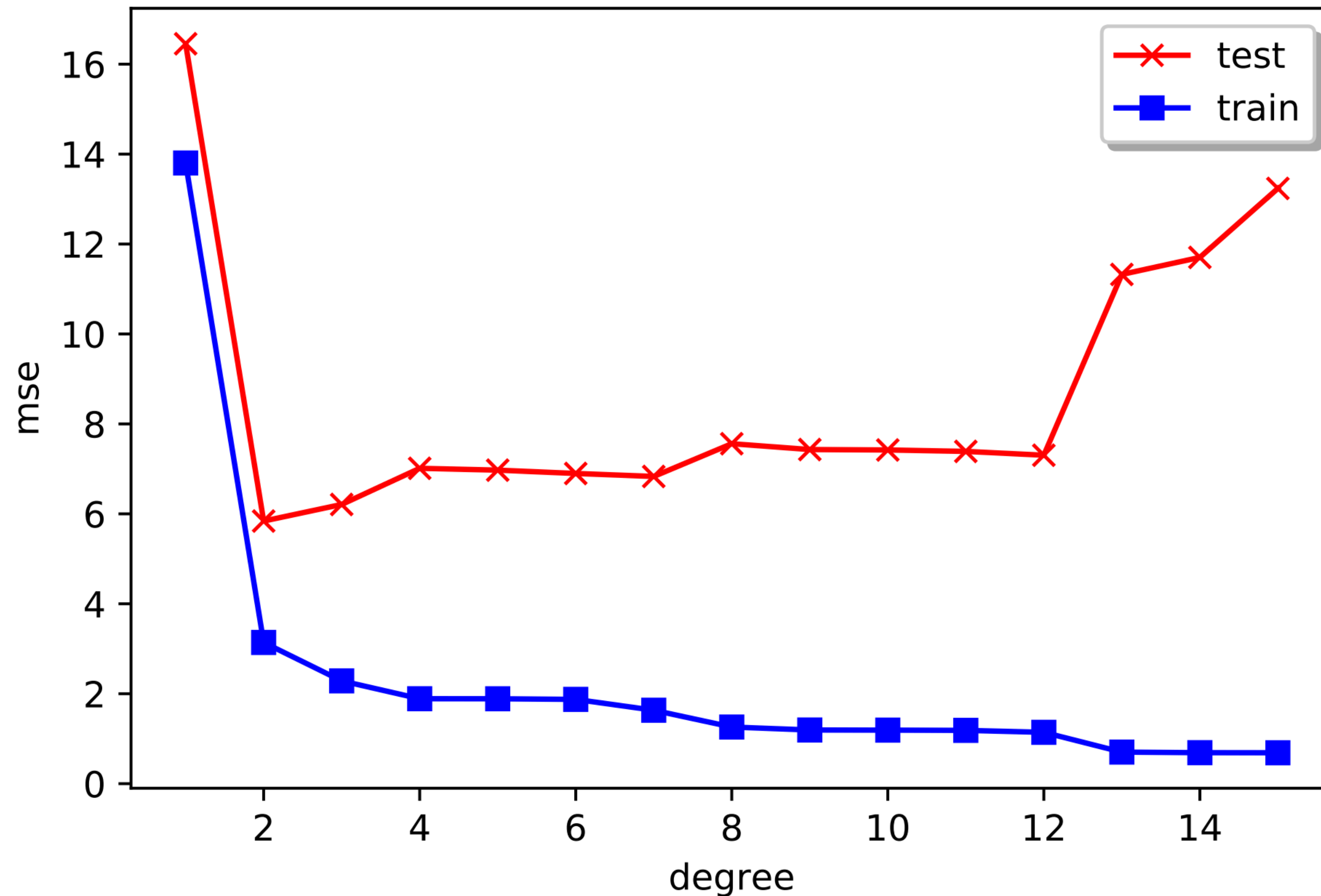


(c)



(d)

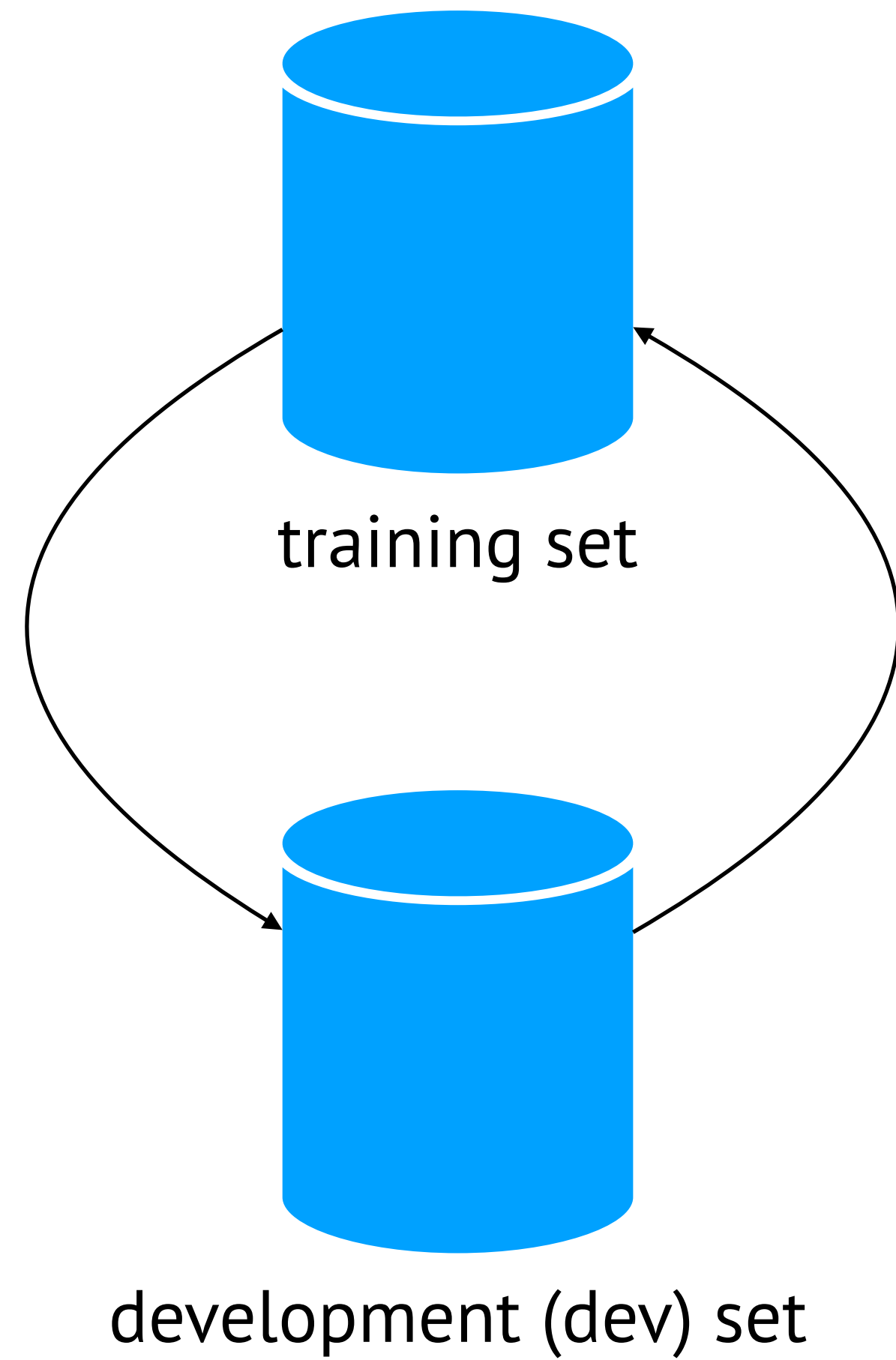
# Overfitting



What is overfitting  
and why it is a bad  
thing?

$$\phi(x) = [1, x, x^2, \dots, x^D]$$

# Generalization





# Machine Learning Problems

Supervised Learning

Labelled data with guidance

Unsupervised Learning

No labelled without guidance

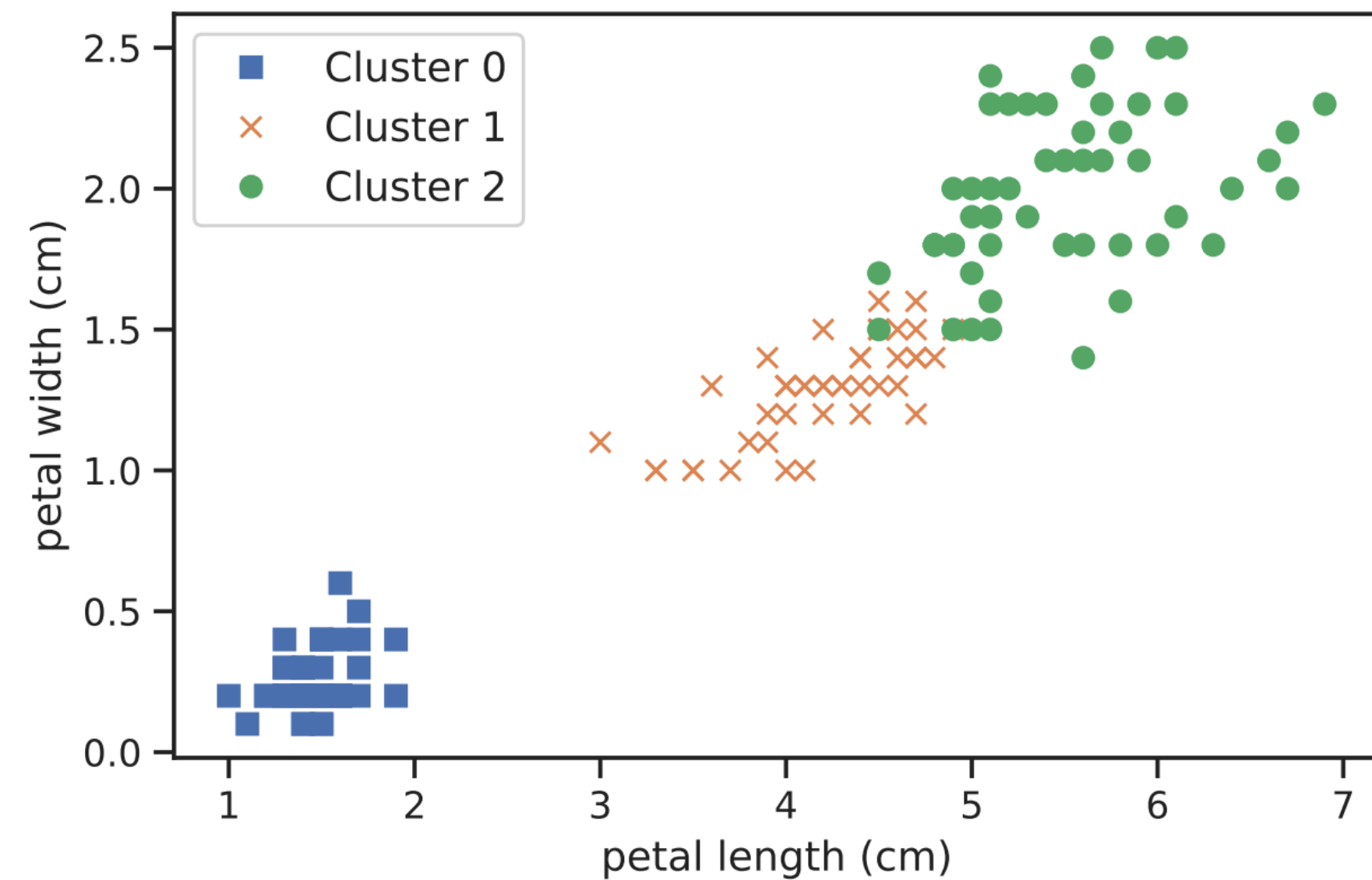
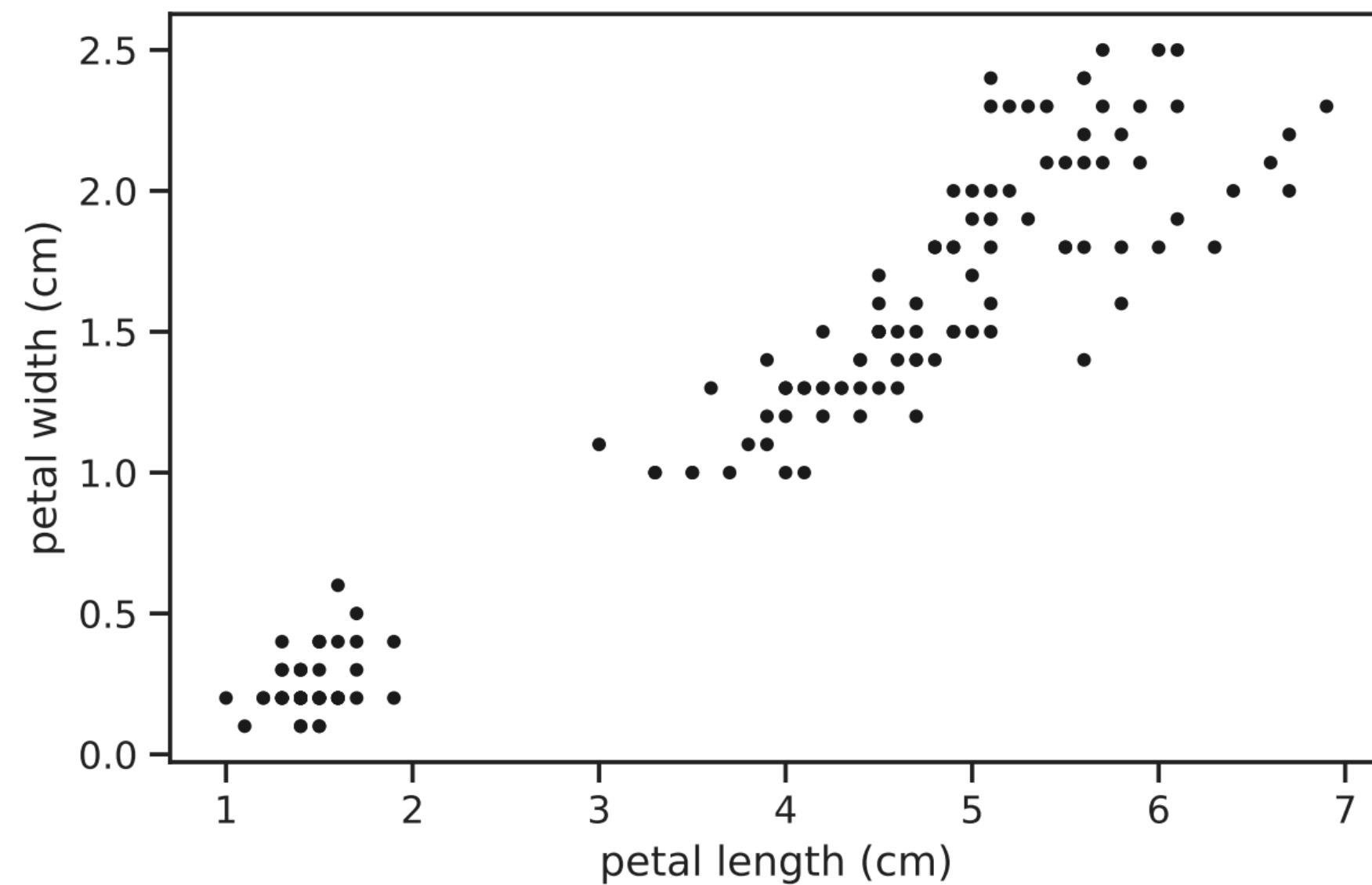
Reinforcement Learning

Interacts with environment, decide action, learns by trial and error method

# Unsupervised Learning – Clustering

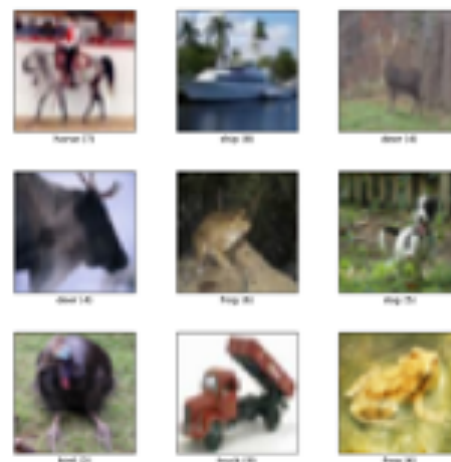
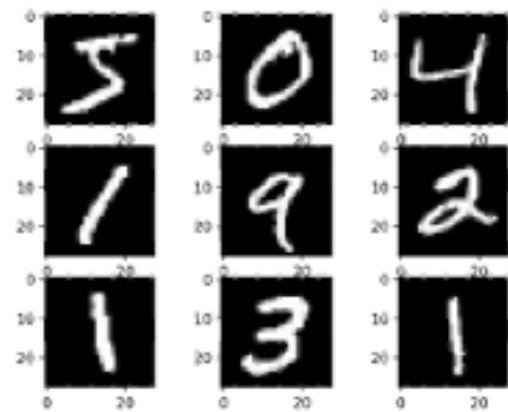
## Goal:

Partition the input into regions that contain “similar” points.



# Unsupervised Learning – Clustering

high-dimensional output  $\mathbf{x}_n \in \mathbb{R}^D$



latent factors  $\mathbf{z}_n \in \mathbb{R}^K$

$$\mathbf{z}_n \rightarrow \mathbf{x}_n$$

**Linear Model:**  $p(\mathbf{x}_n | \mathbf{z}_n; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_n | \mathbf{W}\mathbf{z}_n + \boldsymbol{\mu}, \boldsymbol{\Sigma})$

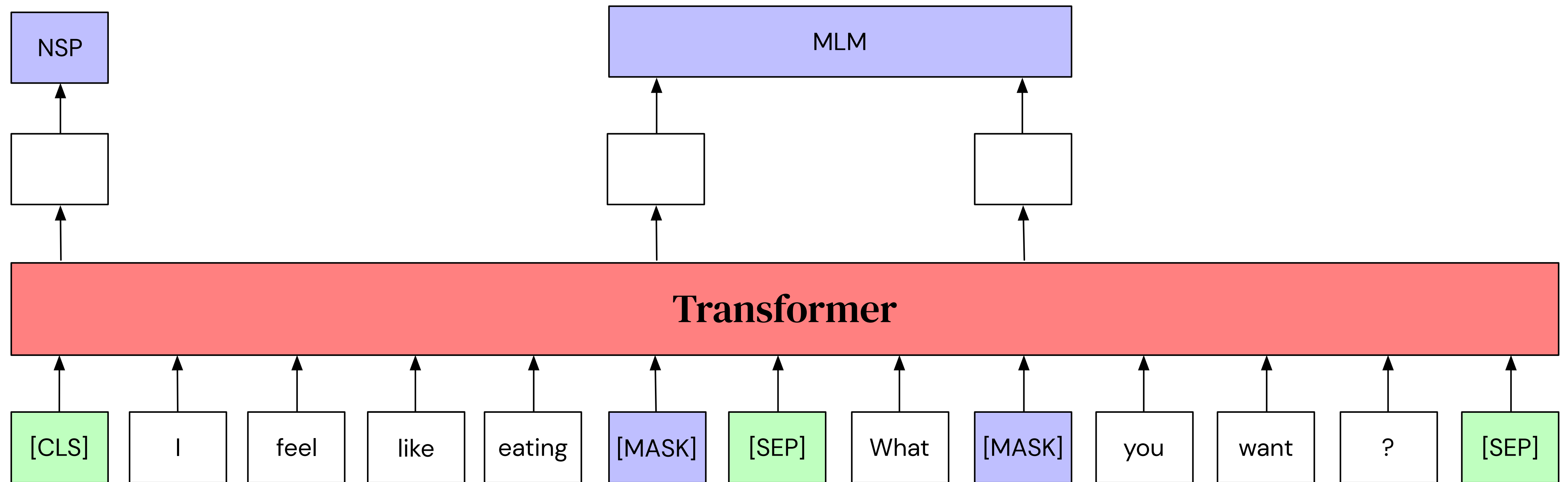
**Non-linear Model:**  $p(\mathbf{x}_n | \mathbf{z}_n; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_n | f(\mathbf{z}_n; \boldsymbol{\theta}), \sigma^2 \mathbf{I})$



# Unsupervised Learning – Self-supervised Learning

## BERT:

masked language modeling + next sentence prediction



# Unsupervised Learning — Evaluation

- It is difficult to evaluate since no ground truth.
- Use learned unsupervised representation as features to a downstream supervised learning method
- If unsupervised model learn useful features, the prediction in supervised model will increase.
- Evaluate unsupervised model by reducing the number of labelled sample to get good performance.

# Machine Learning Problems

Supervised Learning

Labelled data with  
guidance

Unsupervised  
Learning

No labelled without  
guidance

Reinforcement  
Learning

Interacts with  
environment, decide  
action, learns by trial  
and error method



# AlphaGo



# Machine Learning Problems

A system or agent has to learn how to interact with its environment.

This can be encoded by means of a **policy**  $\mathbf{a} = \mathbf{p}(\mathbf{x})$ , which specifies which action to take in response to each possible **input**  $\mathbf{x}$  (derived from the environment state).



(a)



(b)

Figure 1.10: Examples of some control problems. (a) Space Invaders Atari game. From <https://gym.openai.com/envs/SpaceInvaders-v0/>. (b) Controlling a humanoid robot in the MuJoCo simulator so it walks as fast as possible without falling over. From <https://gym.openai.com/envs/Humanoid-v2/>.



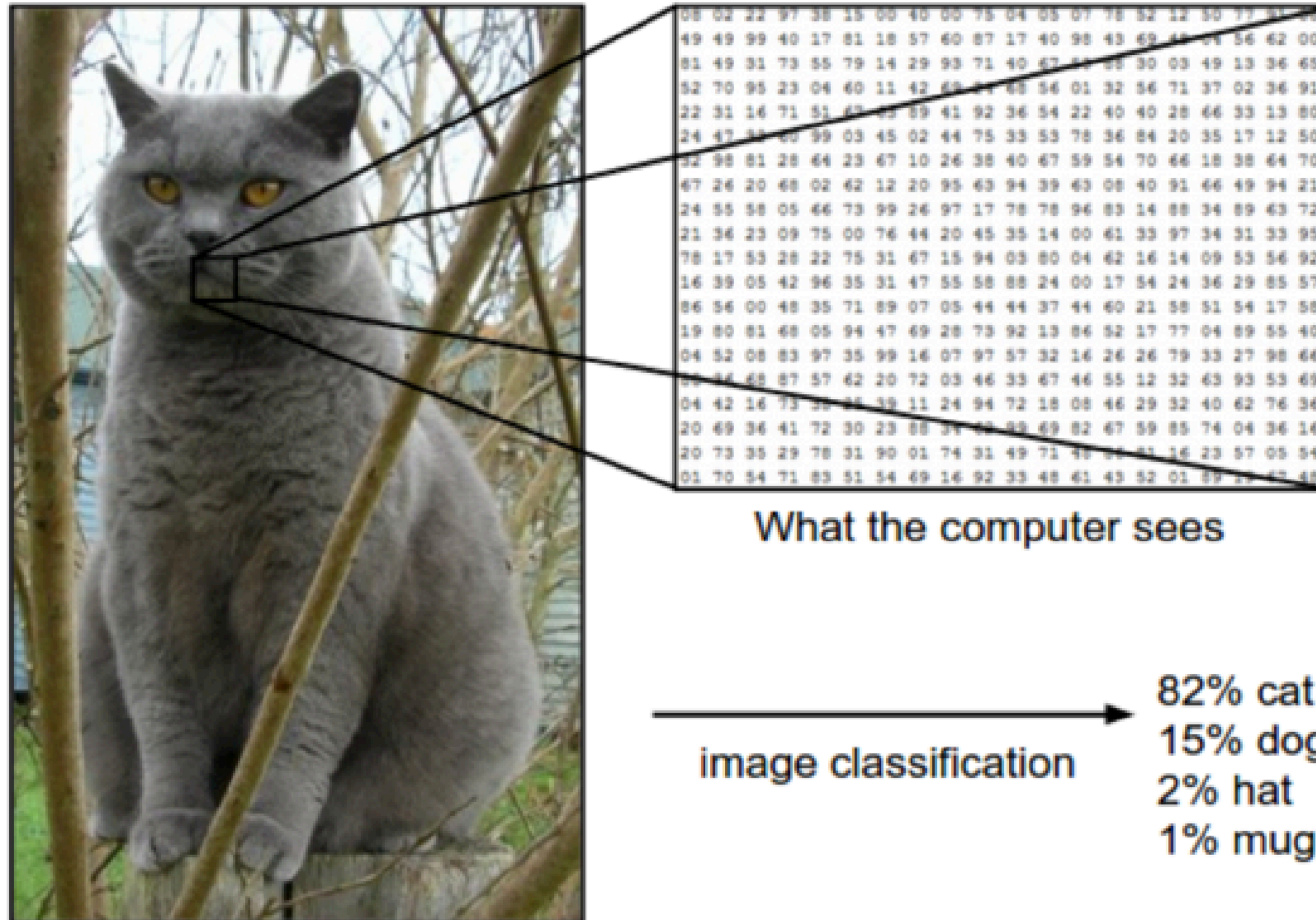
# AlphaFold



**Protein  
folding  
explained**



# Supervised Learning – Classification



*input space,  $\mathcal{X} = \text{set of images}$*

*$\mathcal{X} = \mathbb{R}^D$ , where  $D = C \times D_1 \times D_2$*

*$f: \mathcal{X} \rightarrow \mathcal{Y}$*

$C = 3$  channels (e.g., RGB)