

# Clustering / Unsupervised Representation Learning

COMP3314 — Week 8

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!

# Unsupervised Learning

Supervised Learning

Labelled data with  
guidance

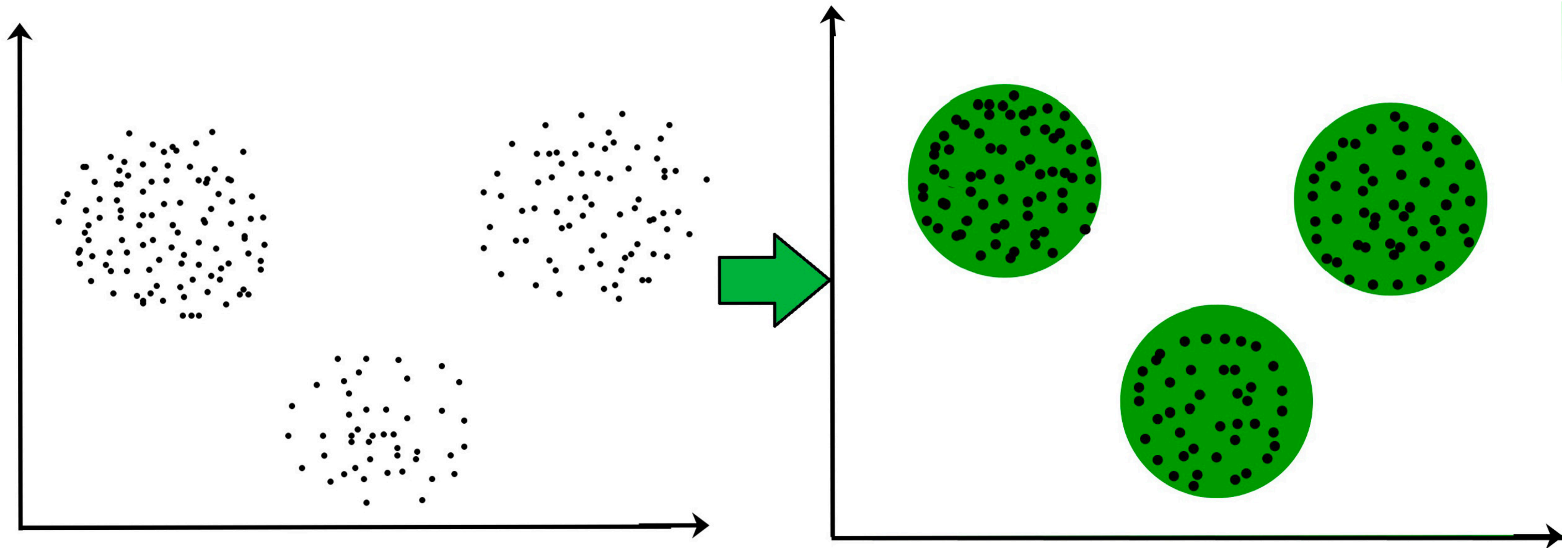
Unsupervised  
Learning

No labelled without  
guidance

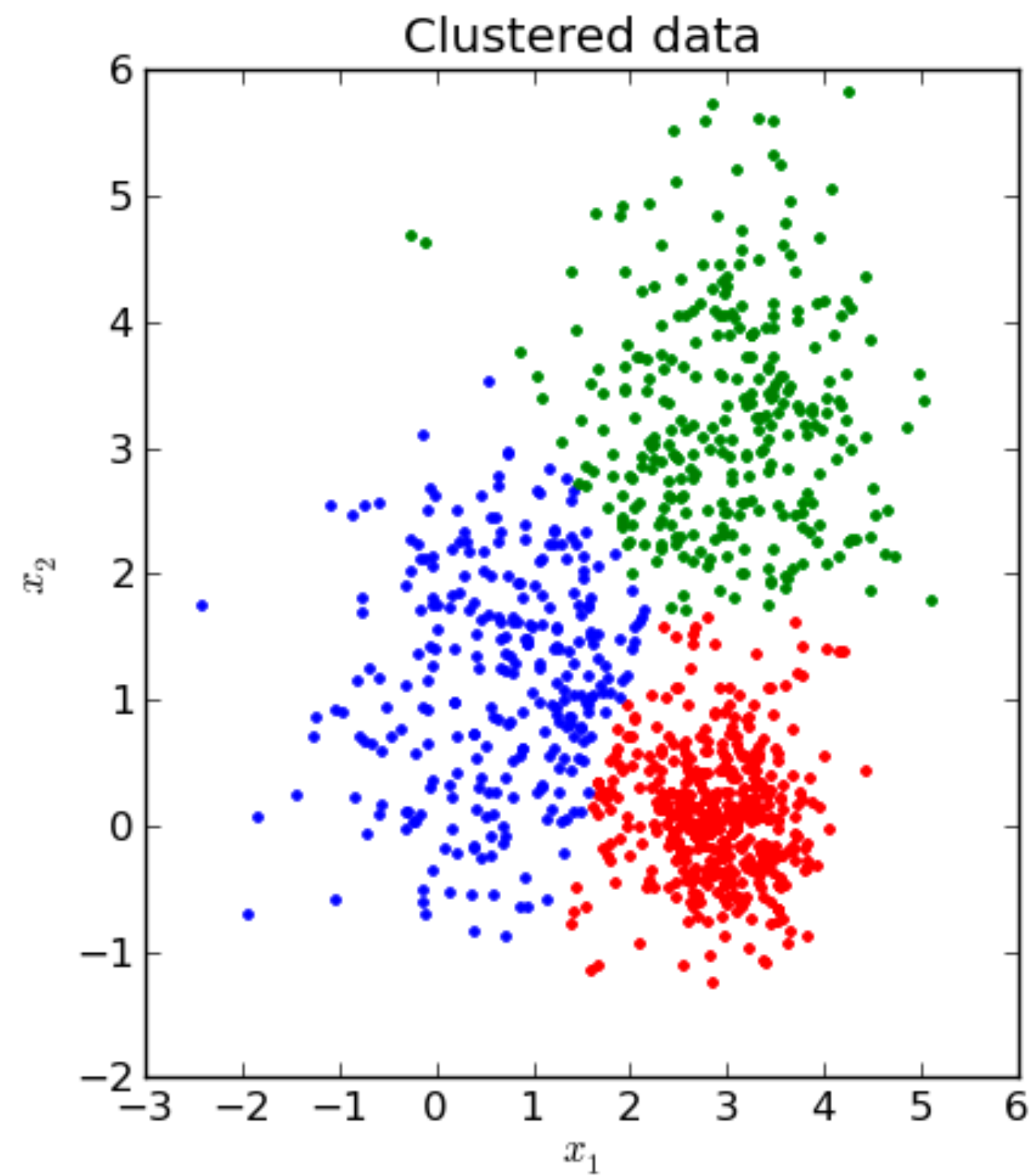
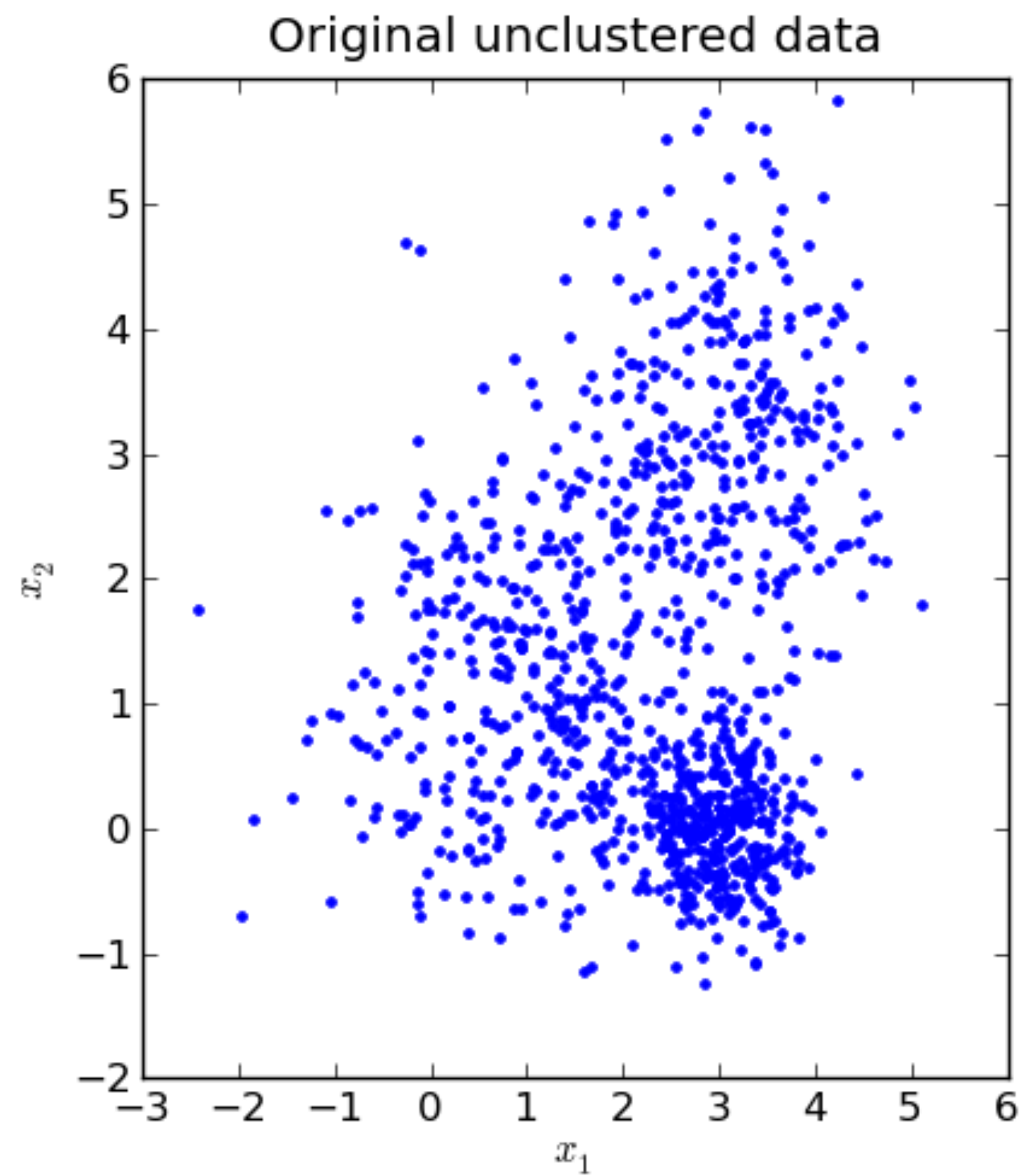
Reinforcement  
Learning

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

# Clustering



# Clustering



Is it even a well-defined machine learning problem?

# Evaluating Clustering Methods

It is hard to evaluate the quality of the output of any given method.

**Intuition:** assign points that are similar to the same cluster, and to ensure that points that are dissimilar are in different clusters.

Rely on external form of data, where we have labels for each object.



$$\text{purity} \triangleq \sum_i \frac{N_i}{N} p_i$$

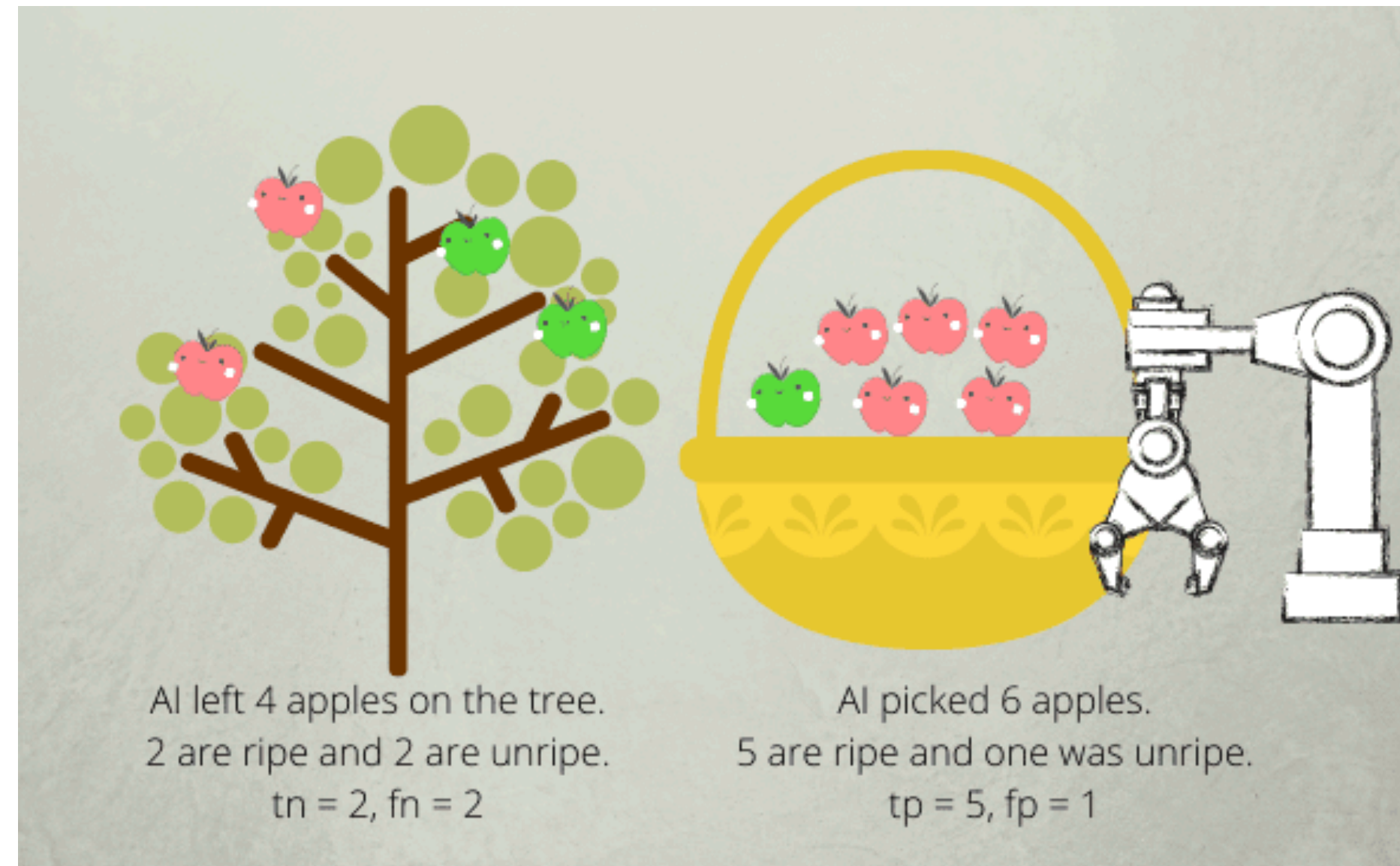
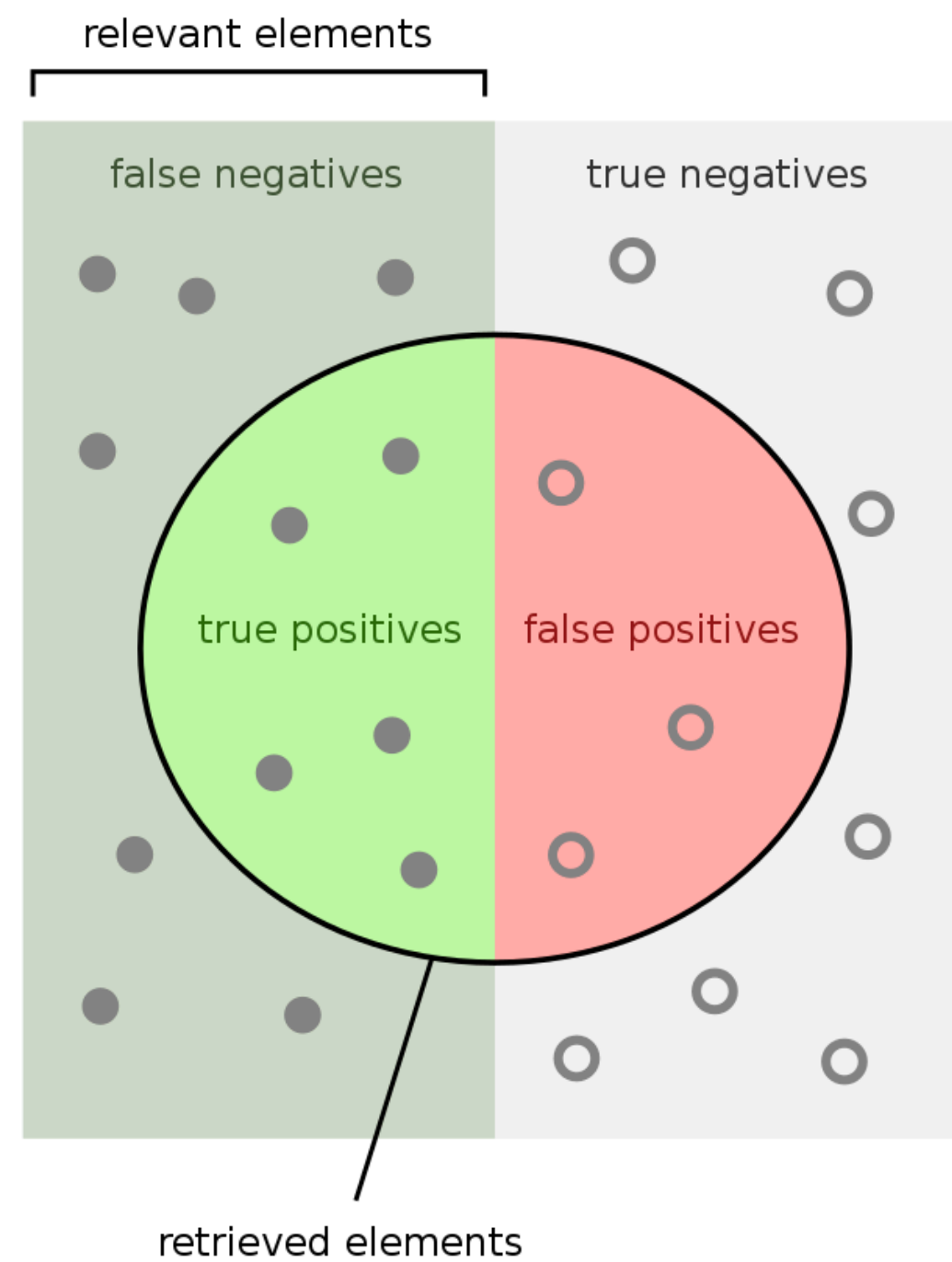
↑

$$\frac{6}{17} \frac{5}{6} + \frac{6}{17} \frac{4}{6} + \frac{5}{17} \frac{3}{5} = \frac{5 + 4 + 3}{17} = 0.71$$

dominant class

range? Why it can be bad?

# F-Score



<https://deepai.org/machine-learning-glossary-and-terms/f-score>

How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2tp}{2tp + fp + fn}$$

# Evaluating Clustering Methods

Rand index

$$R \triangleq \frac{TP + TN}{TP + FP + FN + TN}$$

	same cluster	different cluster
same label	TP	FN
different label	FP	TN



$$TP = C(5,2) + C(4,2) + C(3,2) + C(2,2) = 20$$

$$FP = 5*1 + 4*2 + 1*1 + 2*3 = 5 + 8 + 1 + 6 = 20$$

$$R = (20 + 72) / (20 + 20 + 24 + 72) = 0.68$$

# K Means Clustering

Iterative over:

Assignment:

Assume there are K cluster centers  $\mu_k \in \mathbb{R}^D$

Each data point  $\mathbf{x}_n \in \mathbb{R}^D$  is assigned to the closest center

$$z_n^* = \arg \min_k \|\mathbf{x}_n - \mu_k\|_2^2$$

Multiple  
initialization!!

Compute the centers:

Given the assignments of  $\mathbf{x}_n$  to  $z_n = k$

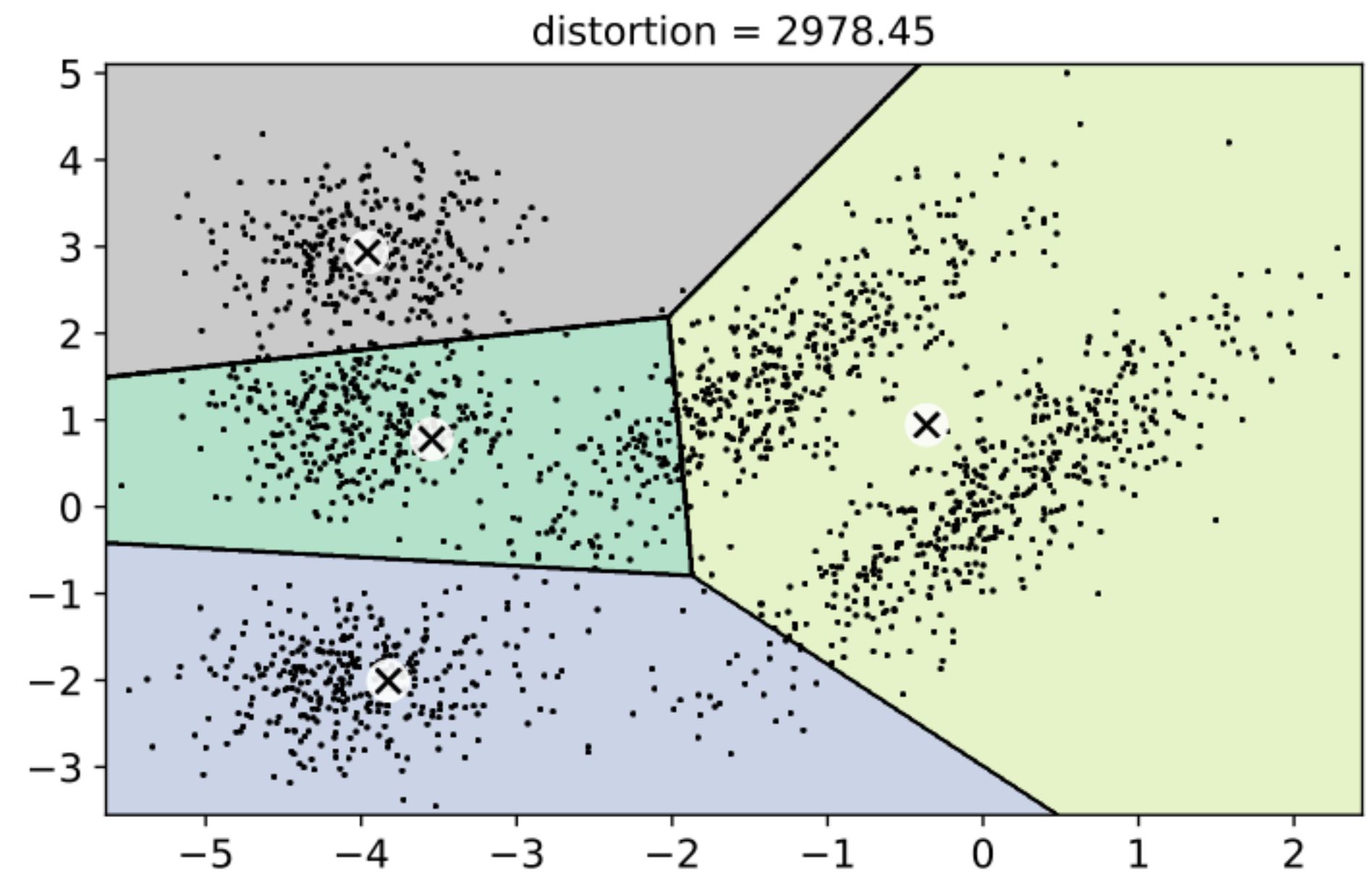
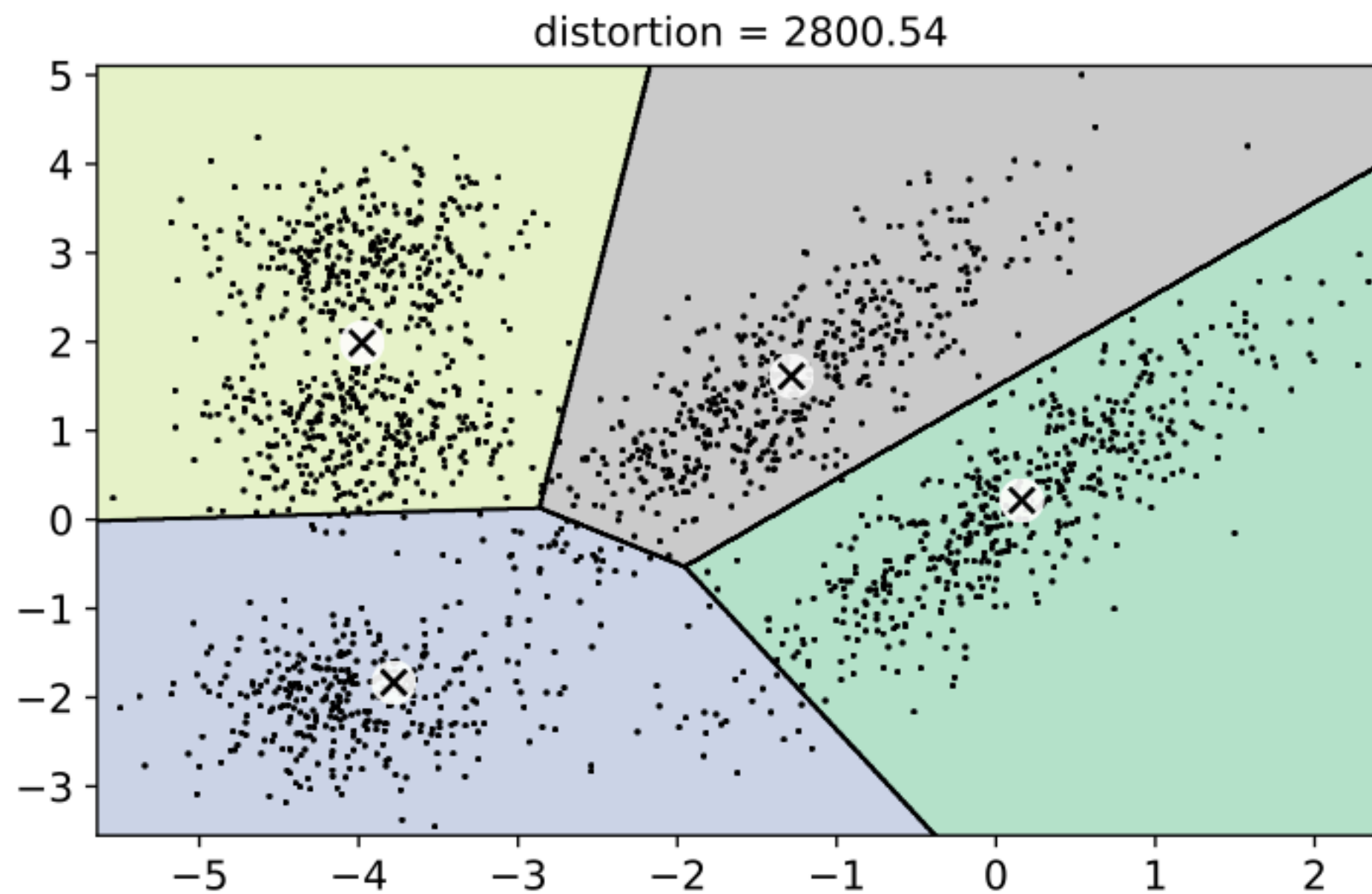
$$\mu_k = \frac{1}{N_k} \sum_{n: z_n = k} \mathbf{x}_n$$



# K Means Clustering

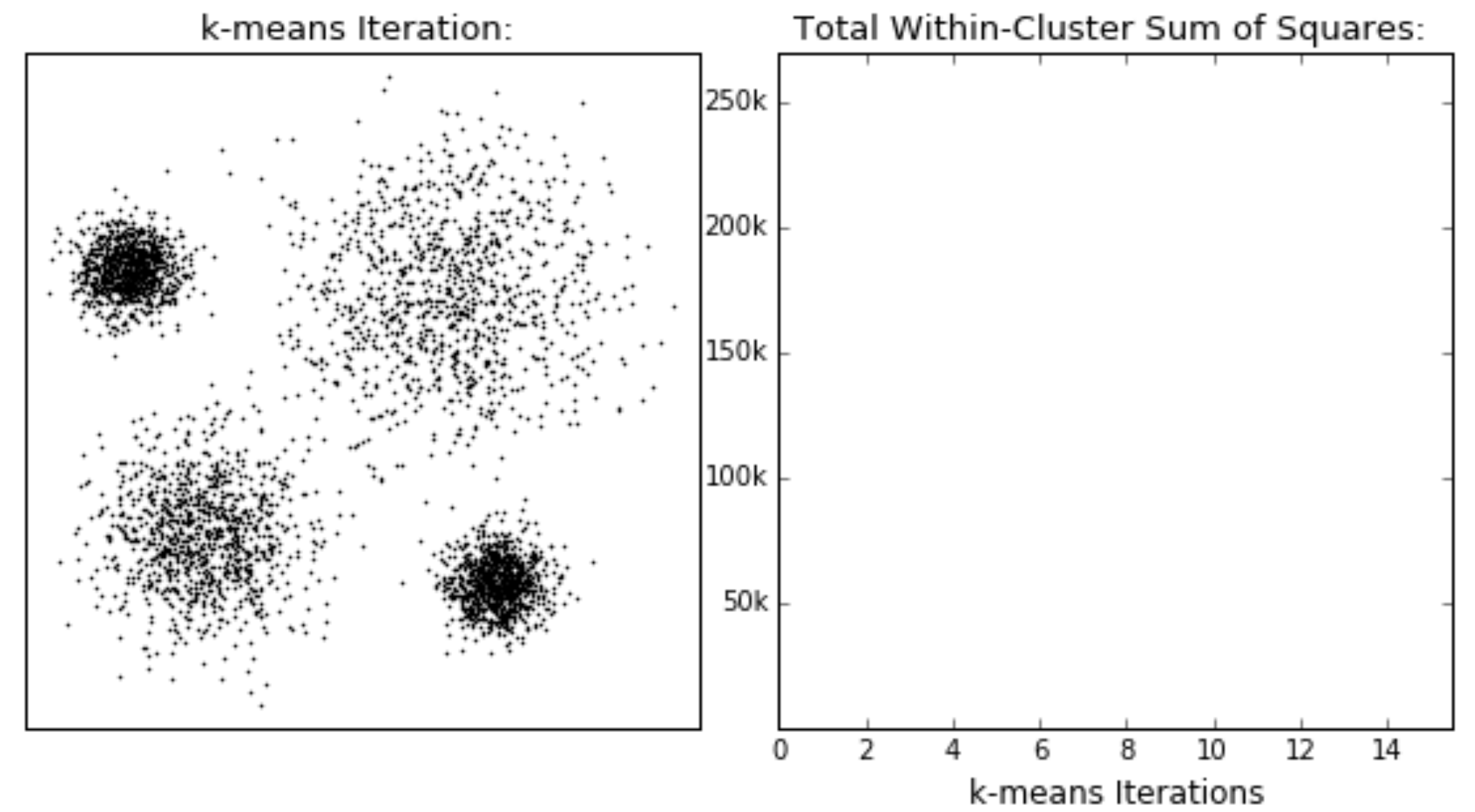
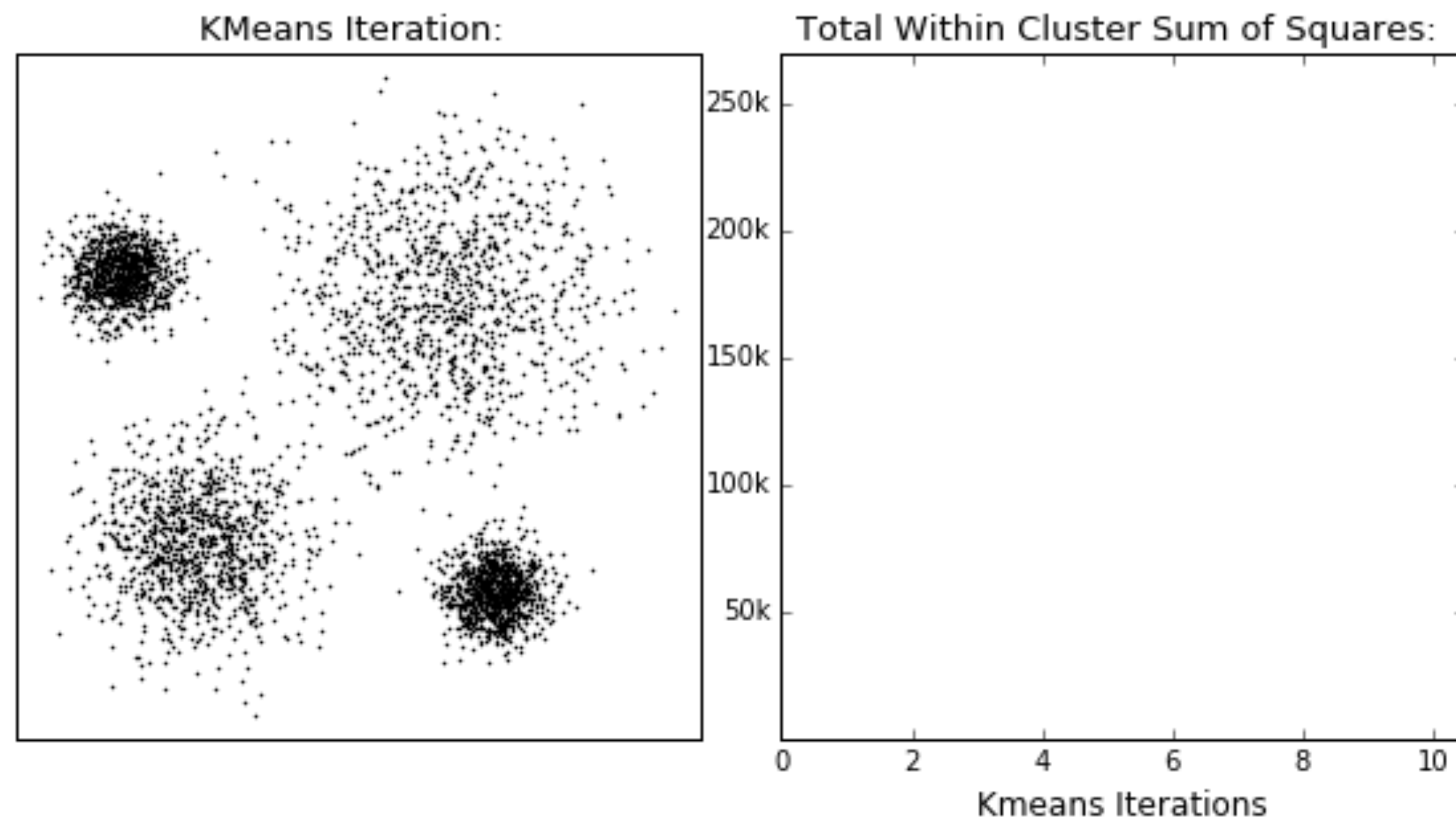
Finding a local minimum of **distortion**

$$J(\mathbf{M}, \mathbf{Z}) = \sum_{n=1}^N \|\mathbf{x}_n - \boldsymbol{\mu}_{z_n}\|^2 = \|\mathbf{X} - \mathbf{Z}\mathbf{M}^T\|_F^2$$

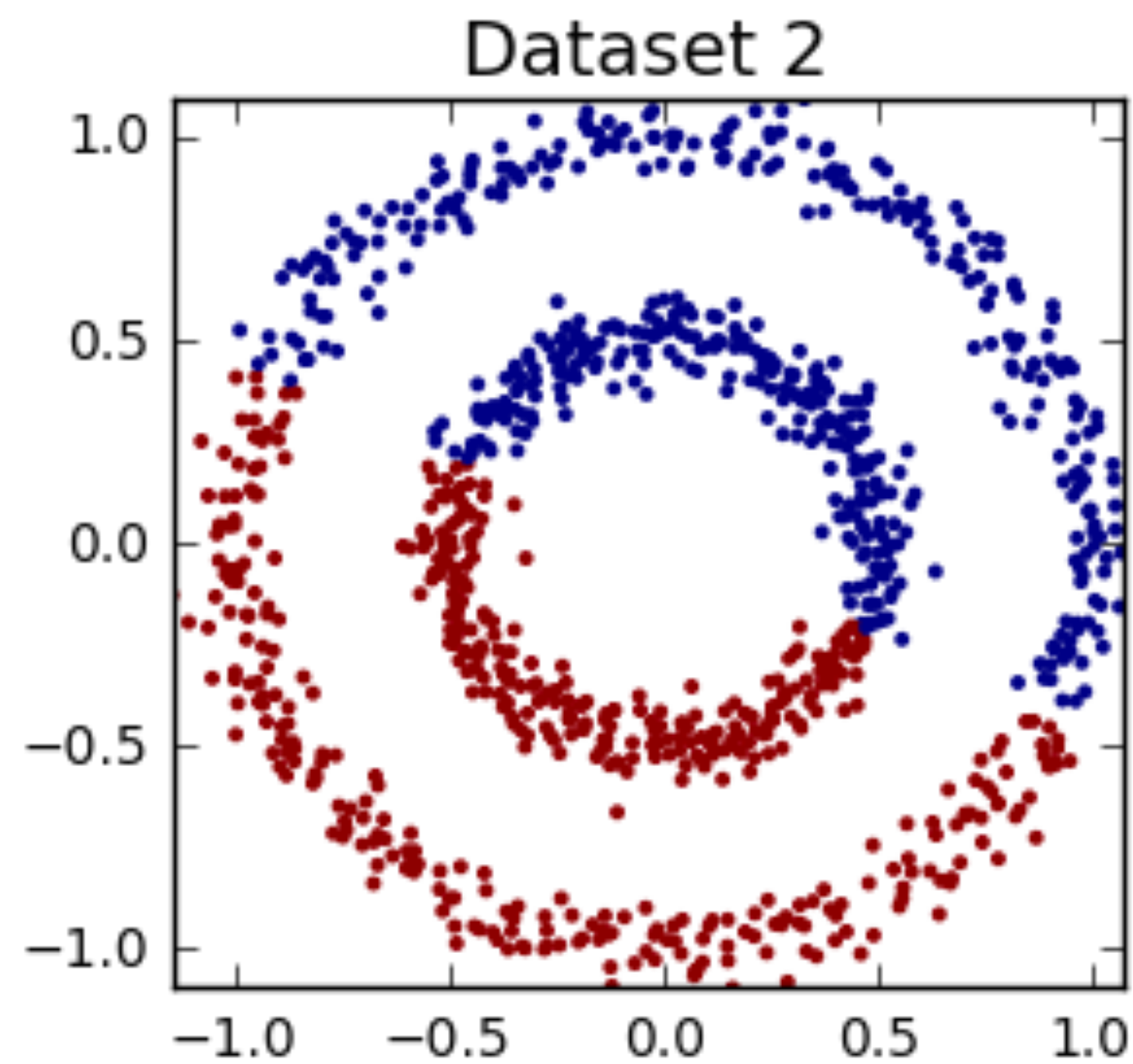
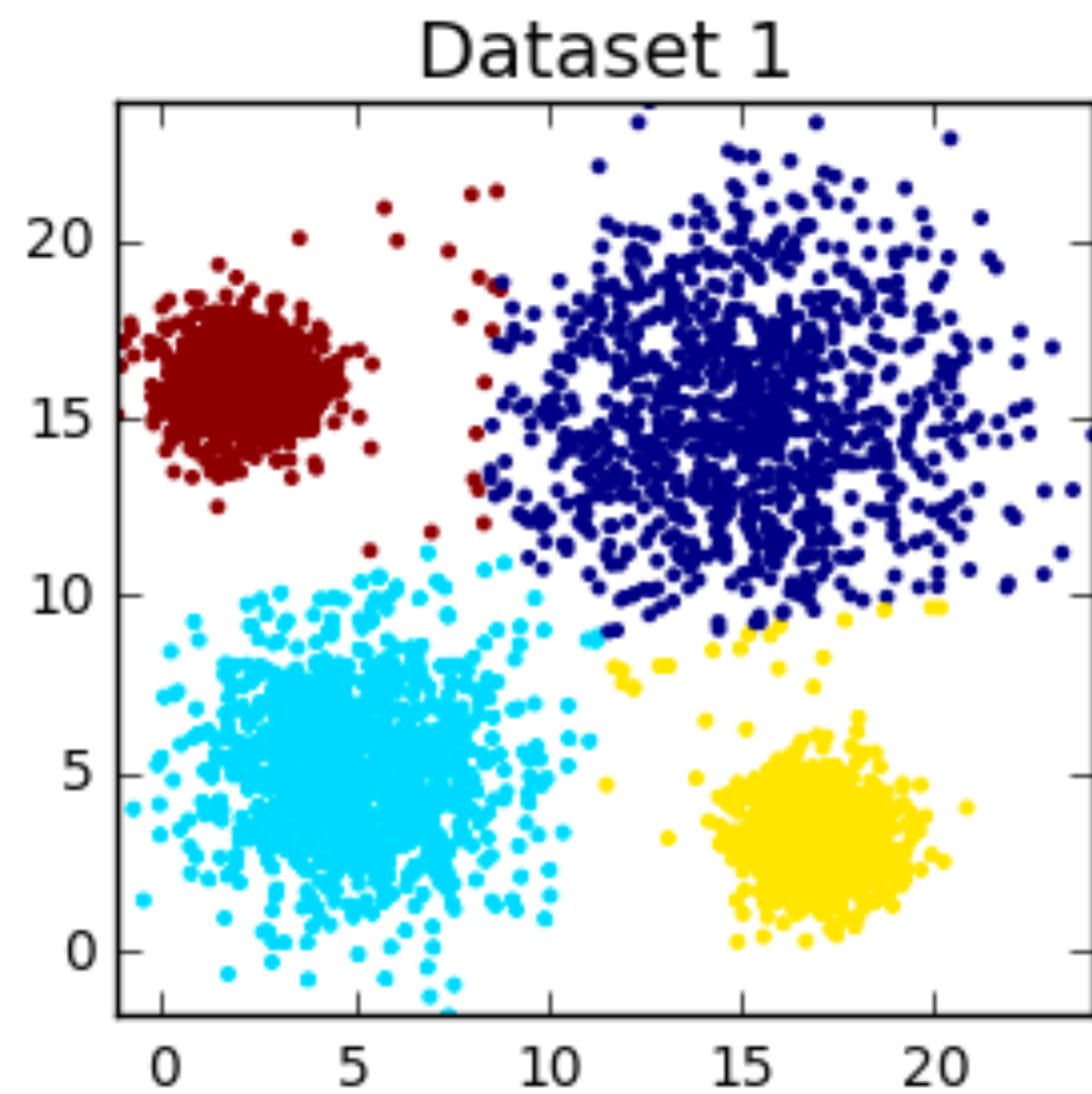


The resulting clustering is sensitive to the initialization!

# K Means Clustering



# K Means Clustering



[https://en.wikipedia.org/wiki/Spectral\\_clustering](https://en.wikipedia.org/wiki/Spectral_clustering)

<https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/>

# K-means++ Algorithm

K-means is optimizing a non-convex objective, and hence needs to be initialized carefully.

**Intuition:** Pick the centers sequentially so as to try to “cover” the data.

At iteration  $t$ , we pick the next cluster center to be  $\mathbf{x}_n$  with probability:

$$p(\boldsymbol{\mu}_t = \mathbf{x}_n) = \frac{D_{t-1}(\mathbf{x}_n)}{\sum_{n'=1}^N D_{t-1}(\mathbf{x}_{n'})} \quad \text{farthest point clustering}$$

where

$$D_t(\mathbf{x}) = \min_{k=1}^{t-1} \|\mathbf{x} - \boldsymbol{\mu}_k\|_2^2 \quad \text{squared distance to the closest existing centroid}$$

# K-means++ Algorithm

K-means is optimizing a non-convex objective, and hence needs to be initialized carefully.

**Intuition:** Pick the centers sequentially so as to try to “cover” the data.

1. Randomly select the first centroid from the data points.
2. For each data point compute its distance from the nearest, previously chosen centroid.
3. Select the next centroid from the data points such that the probability of choosing a point as centroid is directly proportional to its distance from the nearest, previously chosen centroid. (i.e. the point having maximum distance from the nearest centroid is most likely to be selected next as a centroid)
4. Repeat steps 2 and 3 until  $k$  centroids have been sampled

# K-medoids Algorithm

Estimate each cluster center  $\mu_k$  by choosing the data example  $x_n \in \mathcal{X}$  whose average dissimilarity to all other points in that cluster is minimal.

- (1) More robust to outliers
- (2) Can be applied to data that does not live in  $\mathbb{R}^D$ , where the average is not well-defined

---

**Algorithm 12:** K-medoids algorithm

---

- 1 Initialize  $m_{1:K}$  as a random subset of size  $K$  from  $\{1, \dots, N\}$ ;
  - 2 **repeat**
  - 3      $z_n = \operatorname{argmin}_k d(n, m_k)$  for  $n = 1 : N$ ;
  - 4      $m_k = \operatorname{argmin}_{n:z_n=k} \sum_{n':z_{n'}=k} d(n, n')$  for  $k = 1 : K$ ;
  - 5 **until** converged;
- 

For each cluster  $k$ , look at all the points currently assigned to that cluster, and then set  $m_k$  to be the index of the media of that set.

For each point  $n$ , assign it to its closest medoid.  $z_n = \operatorname{argmin}_k D(n, k)$

# Clustering Using Mixture Models

Gaussian mixture model (GMM)

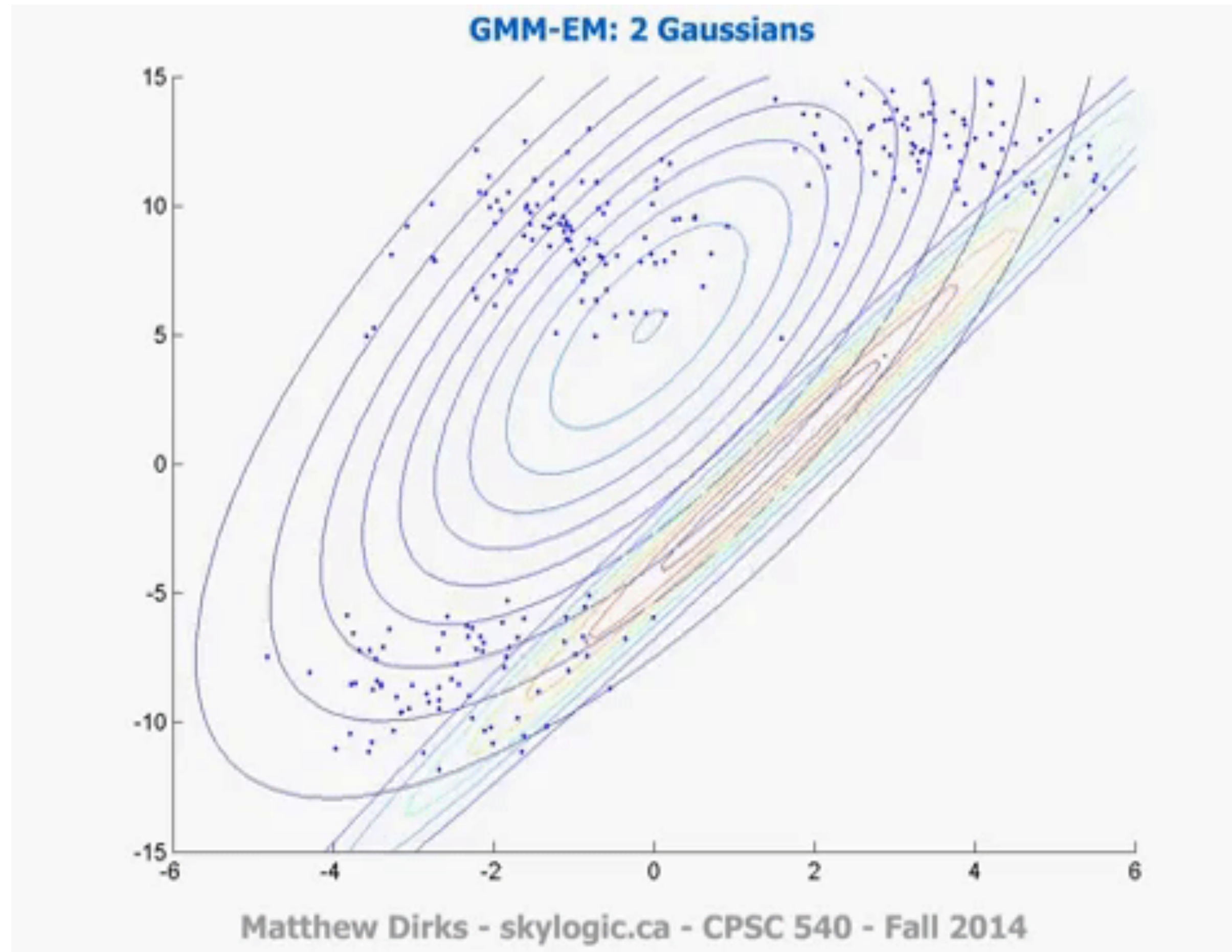
$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

We can use Bayes rule to compute the responsibility (posterior membership probability) of cluster  $k$  for data point  $\mathbf{x}_n$

$$r_{nk} \triangleq p(z_n = k|\mathbf{x}_n, \boldsymbol{\theta}) = \frac{p(z_n = k|\boldsymbol{\theta})p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta})}{\sum_{k'=1}^K p(z_n = k'|\boldsymbol{\theta})p(\mathbf{x}_n|z_n = k', \boldsymbol{\theta})}$$

EM algorithm — similar to the one used in K-means

# Clustering Using Mixture Models





# K-means and EM

Iterative over:

Assignment (E-step):

Assume there are  $K$  cluster centers  $\boldsymbol{\mu}_k \in \mathbb{R}^D$

Each data point  $\boldsymbol{x}_n \in \mathbb{R}^D$  is assigned to the closest center

$$z_n^* = \arg \min_k \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|_2^2$$

Compute the centers (M-Step):

Given the assignments of  $\boldsymbol{x}_n$  to  $z_n = k$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n: z_n = k} \boldsymbol{x}_n$$

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Hard instead of soft

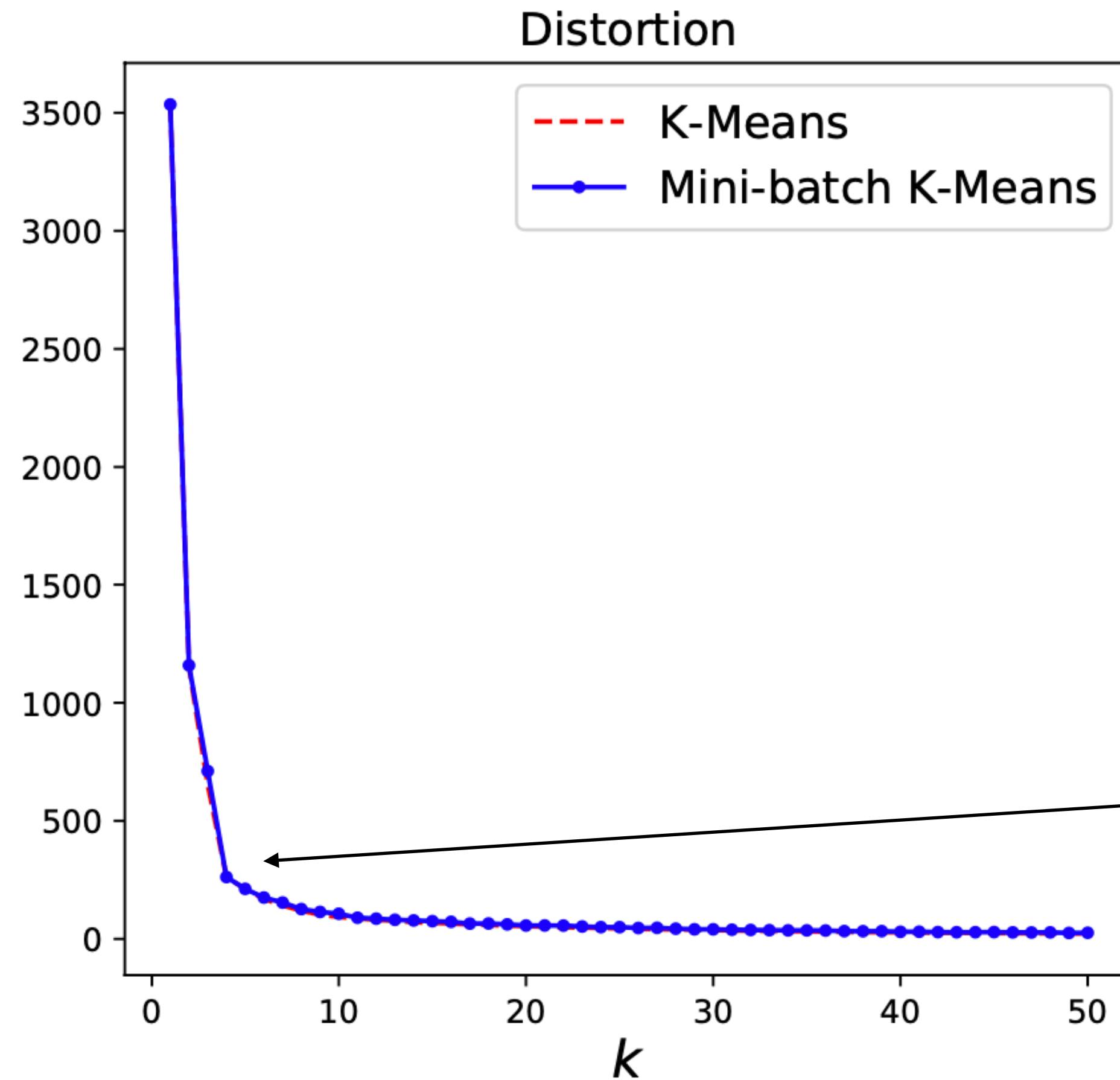
$$z_n^* = \operatorname{argmax}_k r_{nk}$$

Simplify

$$\boldsymbol{\Sigma}_k = \mathbf{I}$$

$$\pi_k = 1/K$$

# Choosing the Number of Clusters K



Minimize the distortion on a validation set?

$$\text{err}(\mathcal{D}_{\text{valid}}, K) = \frac{1}{|\mathcal{D}_{\text{valid}}|} \sum_{n \in \mathcal{D}_{\text{valid}}} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|_2^2$$

Can the distortion figure tell some information?

Elbow!

# Choosing the Number of Clusters K

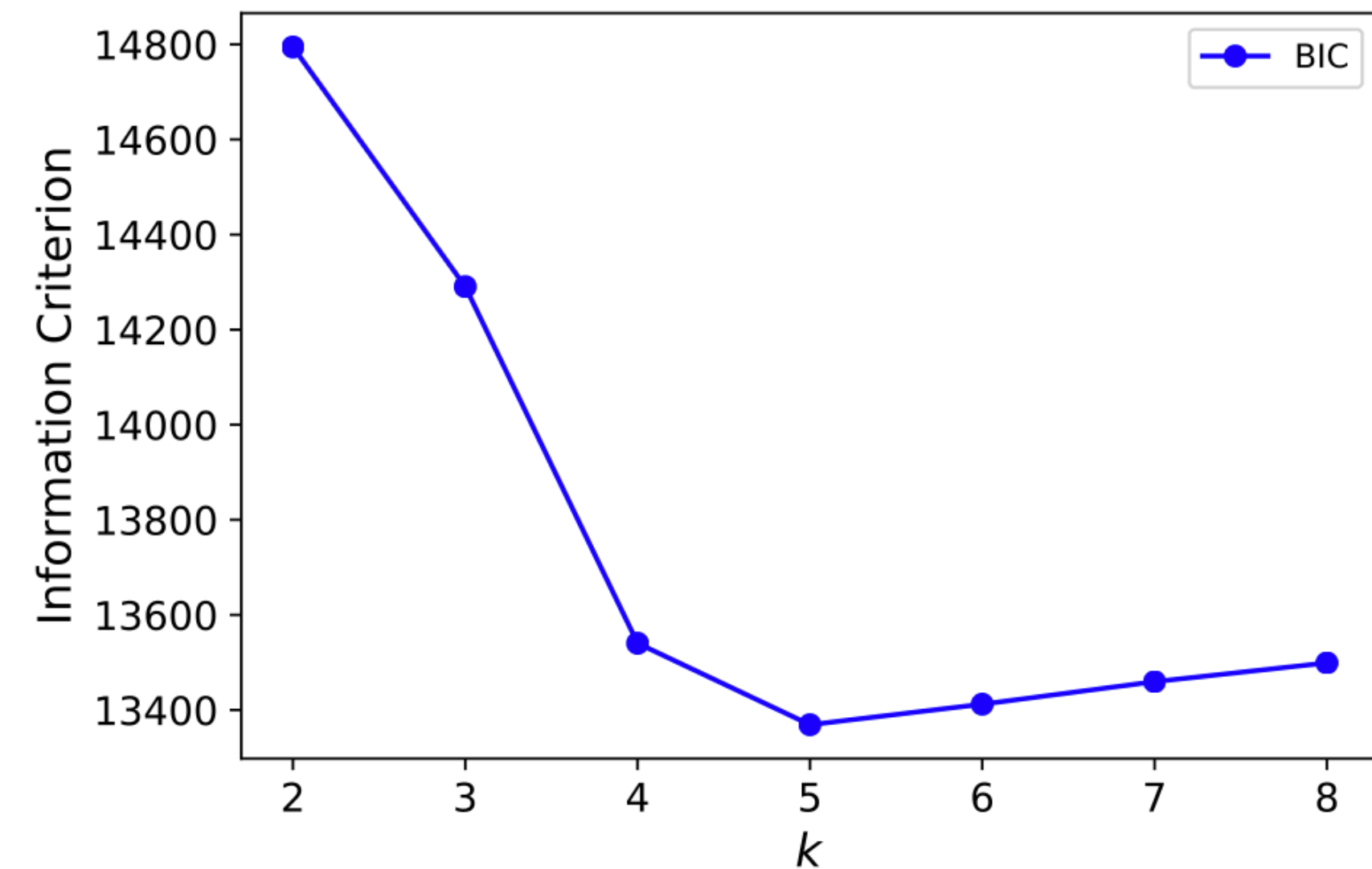
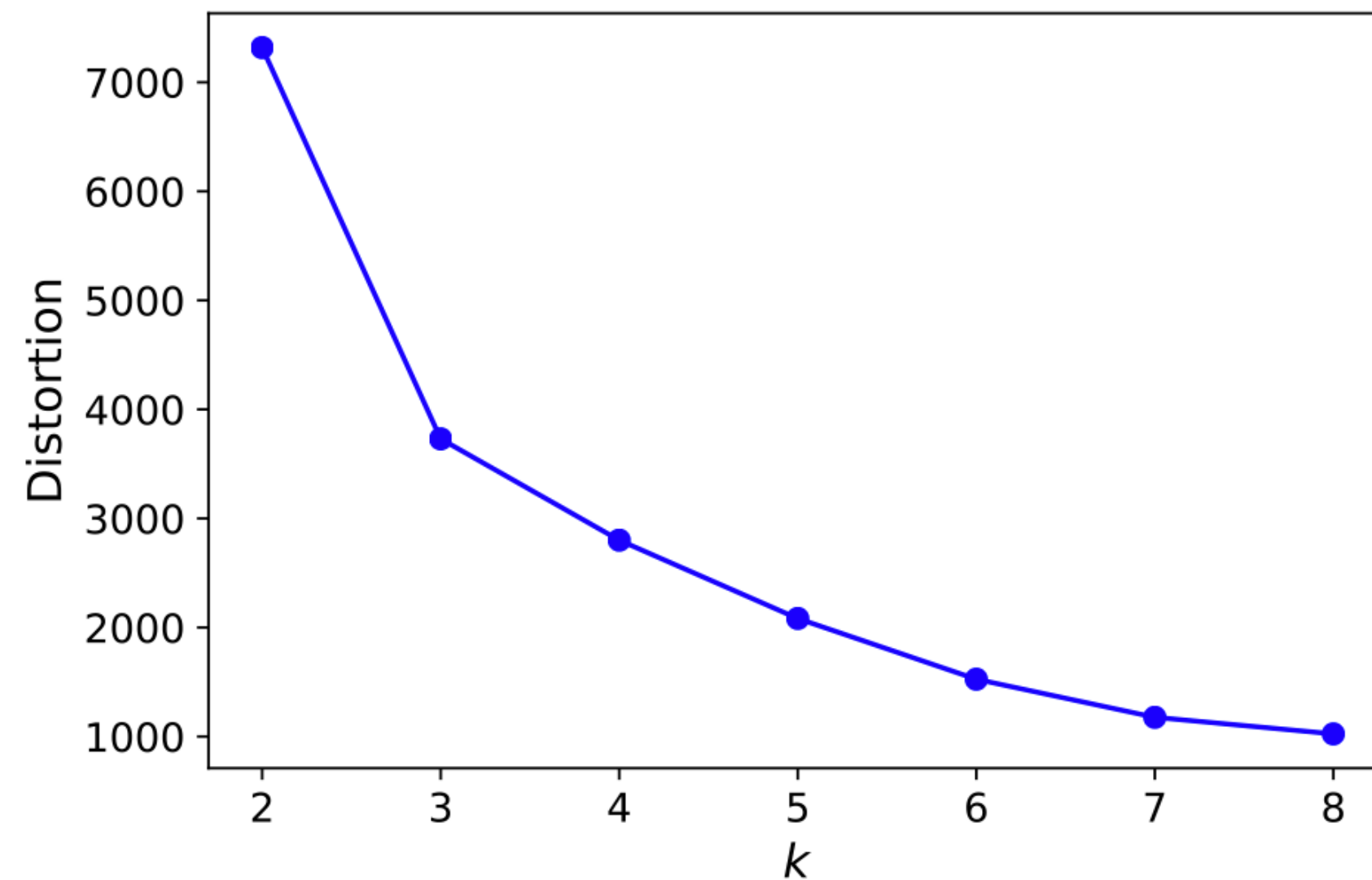
Maximizing the marginal likelihood

$$\text{BIC}(K) = \log p(\mathcal{D} | \hat{\theta}_k) - \frac{D_K}{2} \log(N)$$

BIC: Bayesian Information Criterion

MLE, fits K Gaussian to the clusters

Number of parameters in a model with K clusters



# Vector Quantization (Optional)

A very simple approach to performing lossy compression of some real-valued vectors.

Replace each real-valued vector  $\mathbf{x}_n$  with a discrete symbol  $z_n \in \{1, \dots, K\}$

$$\text{encode}(\mathbf{x}_n) = \arg \min_k \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

The quality of a codebook by computing the reconstruction error or distortion it induces:

$$J \triangleq \frac{1}{N} \sum_{n=1}^N \|\mathbf{x}_n - \text{decode}(\text{encode}(\mathbf{x}_n))\|^2 = \frac{1}{N} \sum_{n=1}^N \|\mathbf{x}_n - \boldsymbol{\mu}_{z_n}\|^2$$

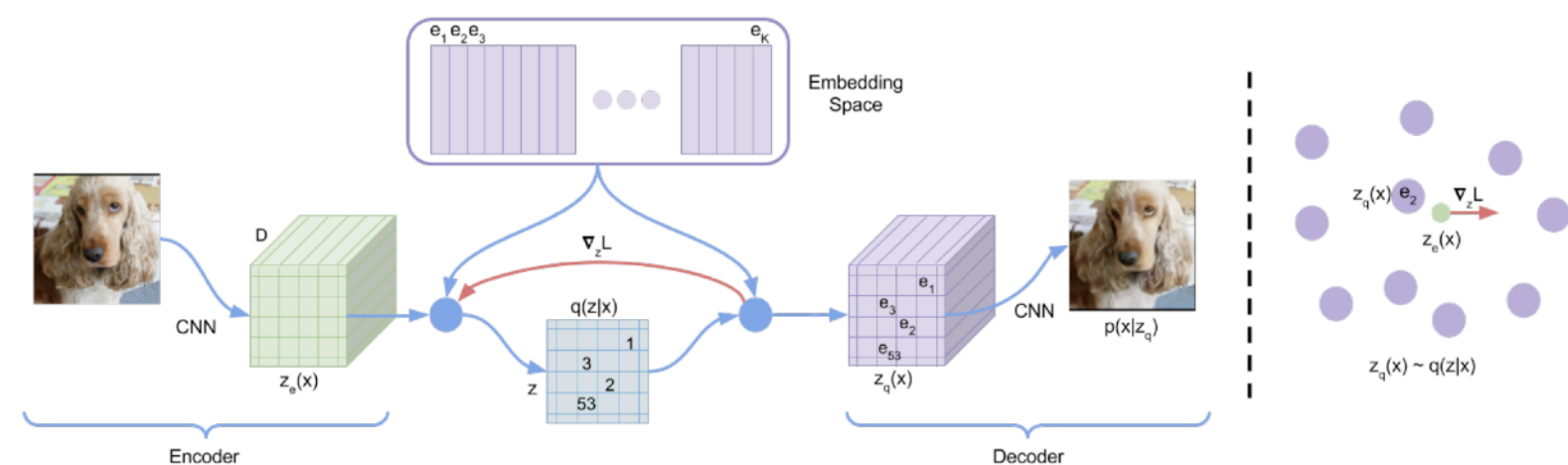
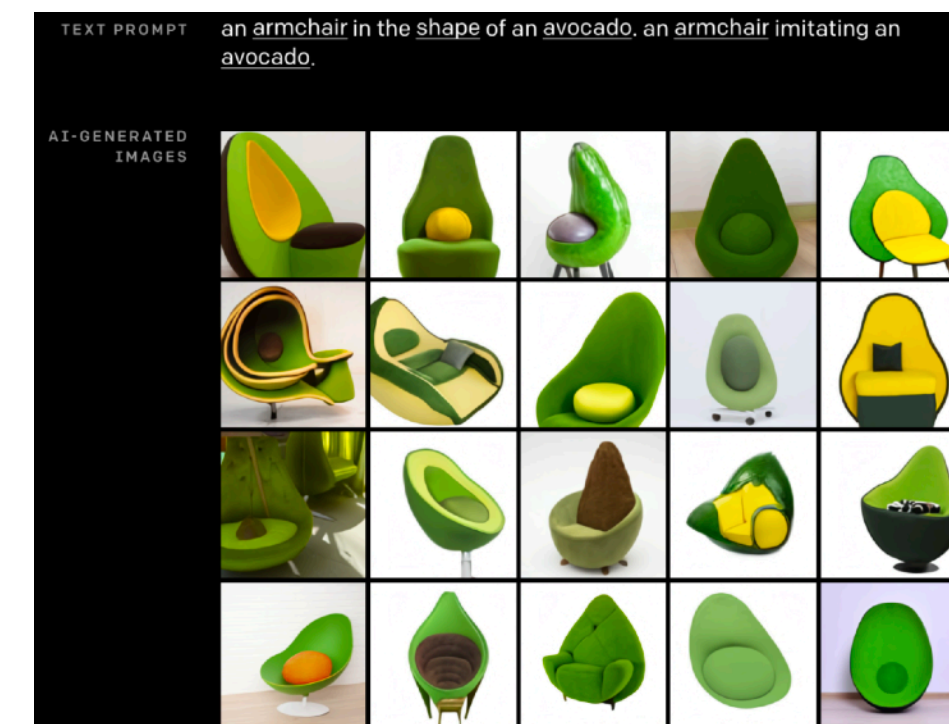
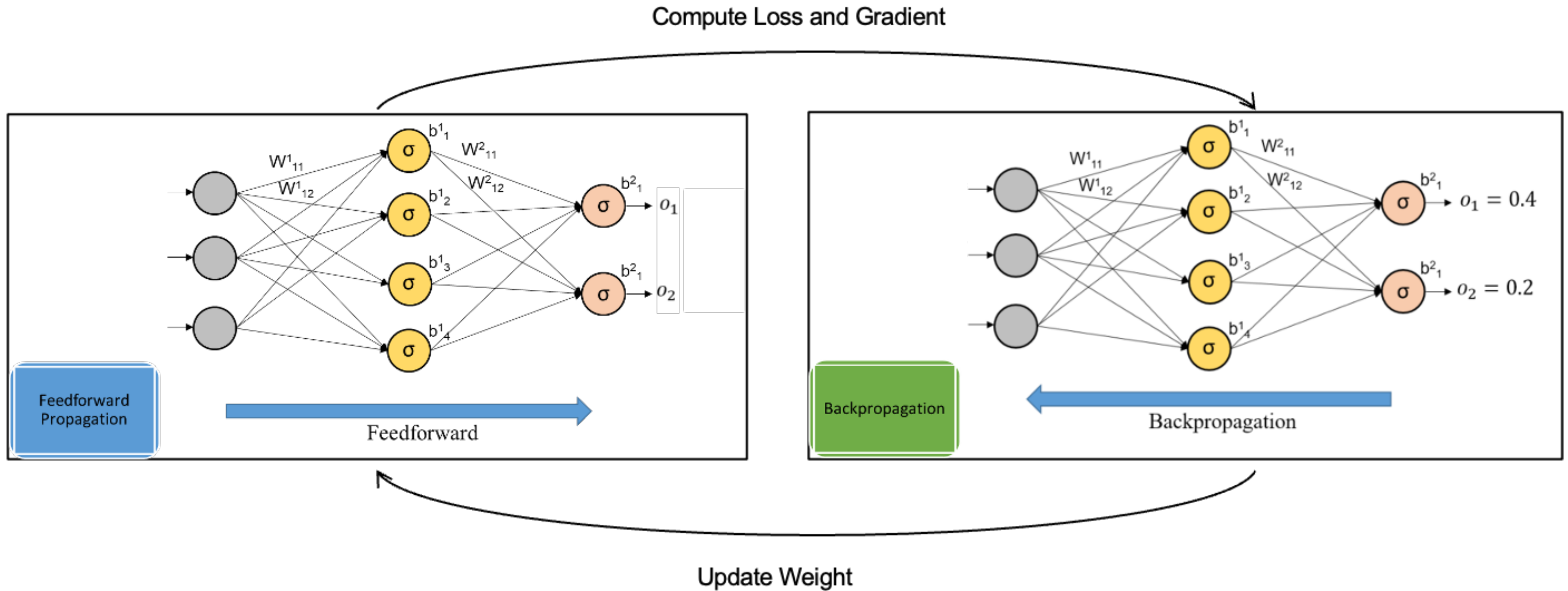


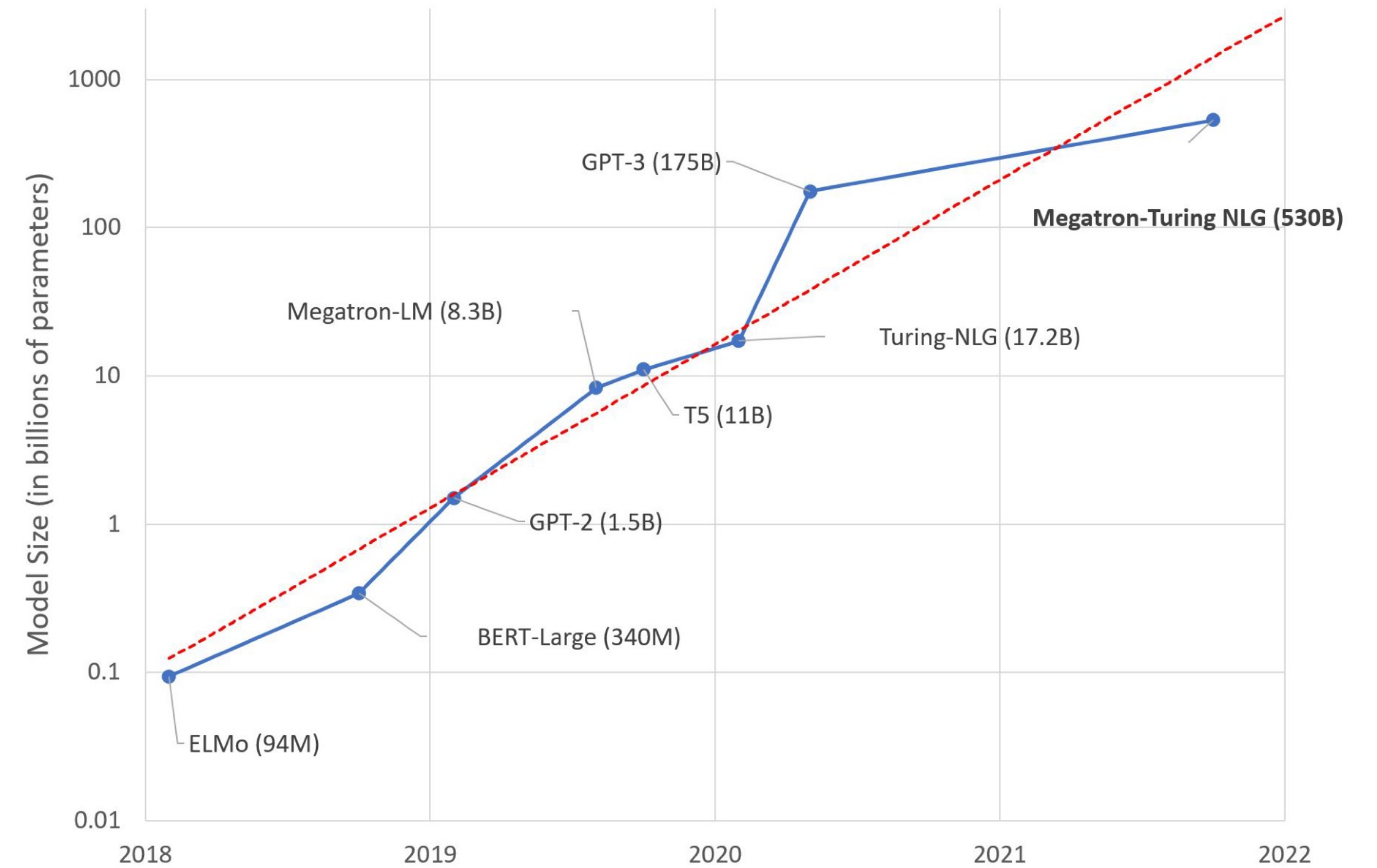
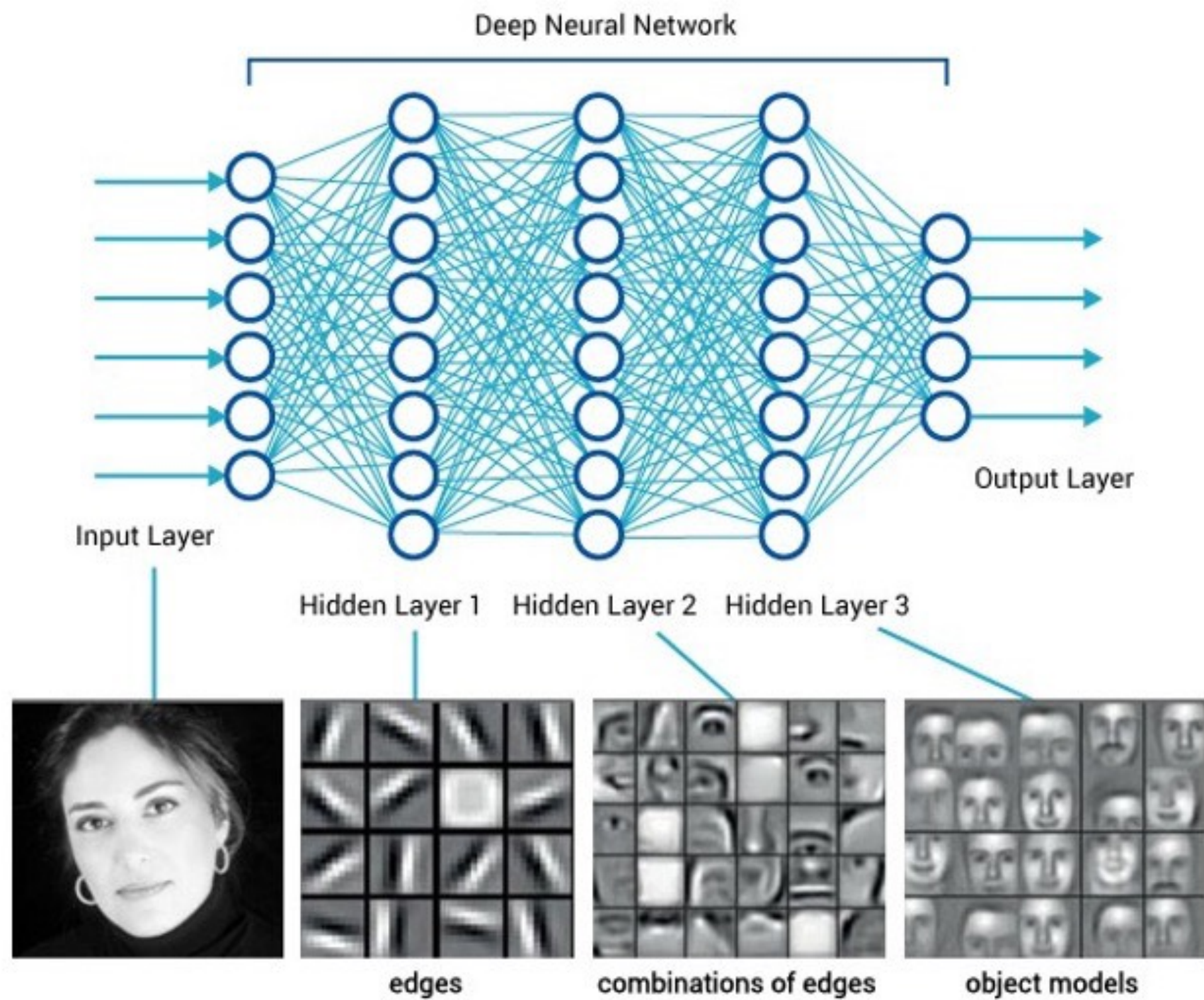
Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder  $z(x)$  is mapped to the nearest point  $e_2$ . The gradient  $\nabla_z L$  (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.



# Recap: Feedforward Neural Network

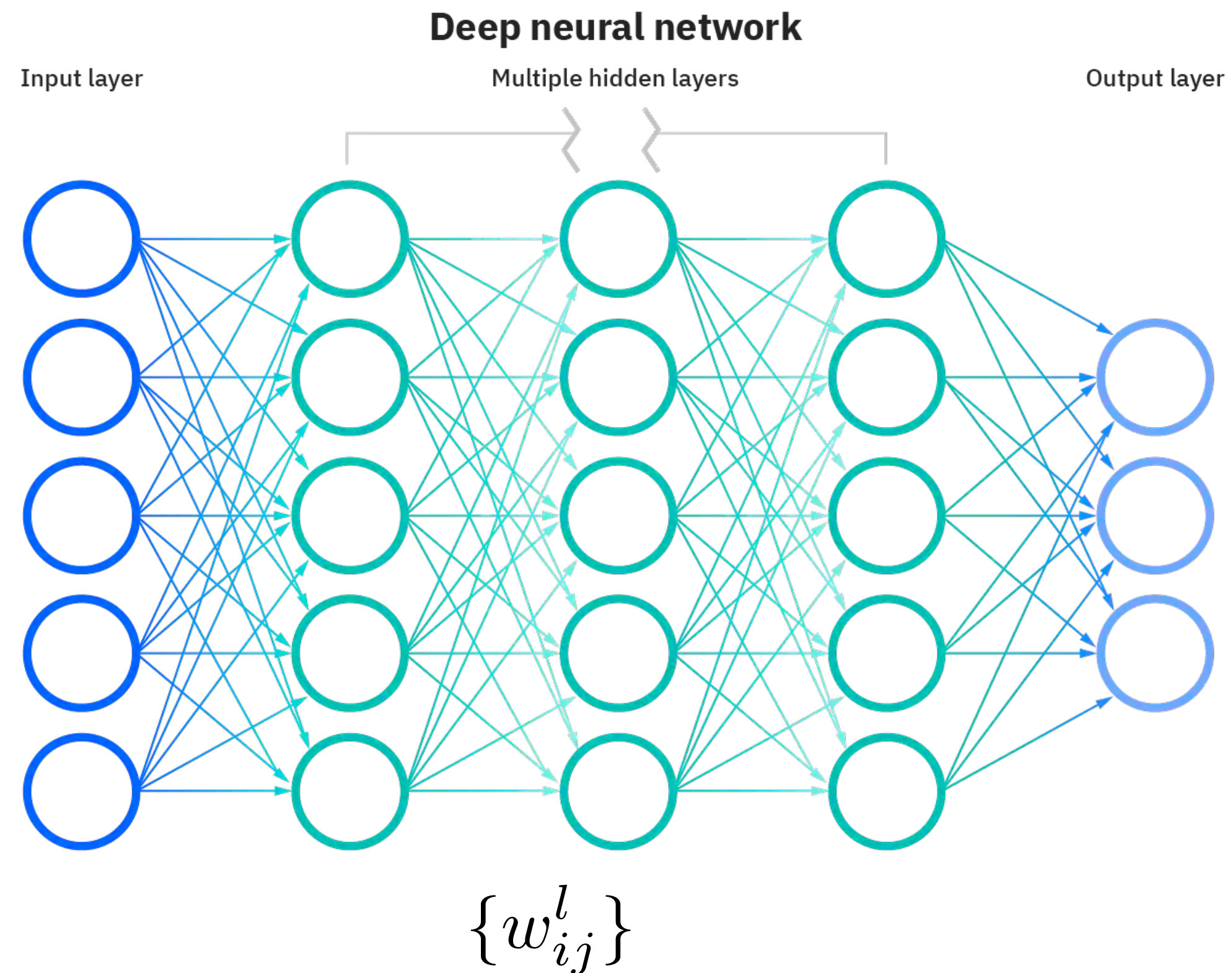


# Recap: Larger Networks



Microsoft Research Blog. Oct 6, 2021.

# What do we get from this?



Case #1:



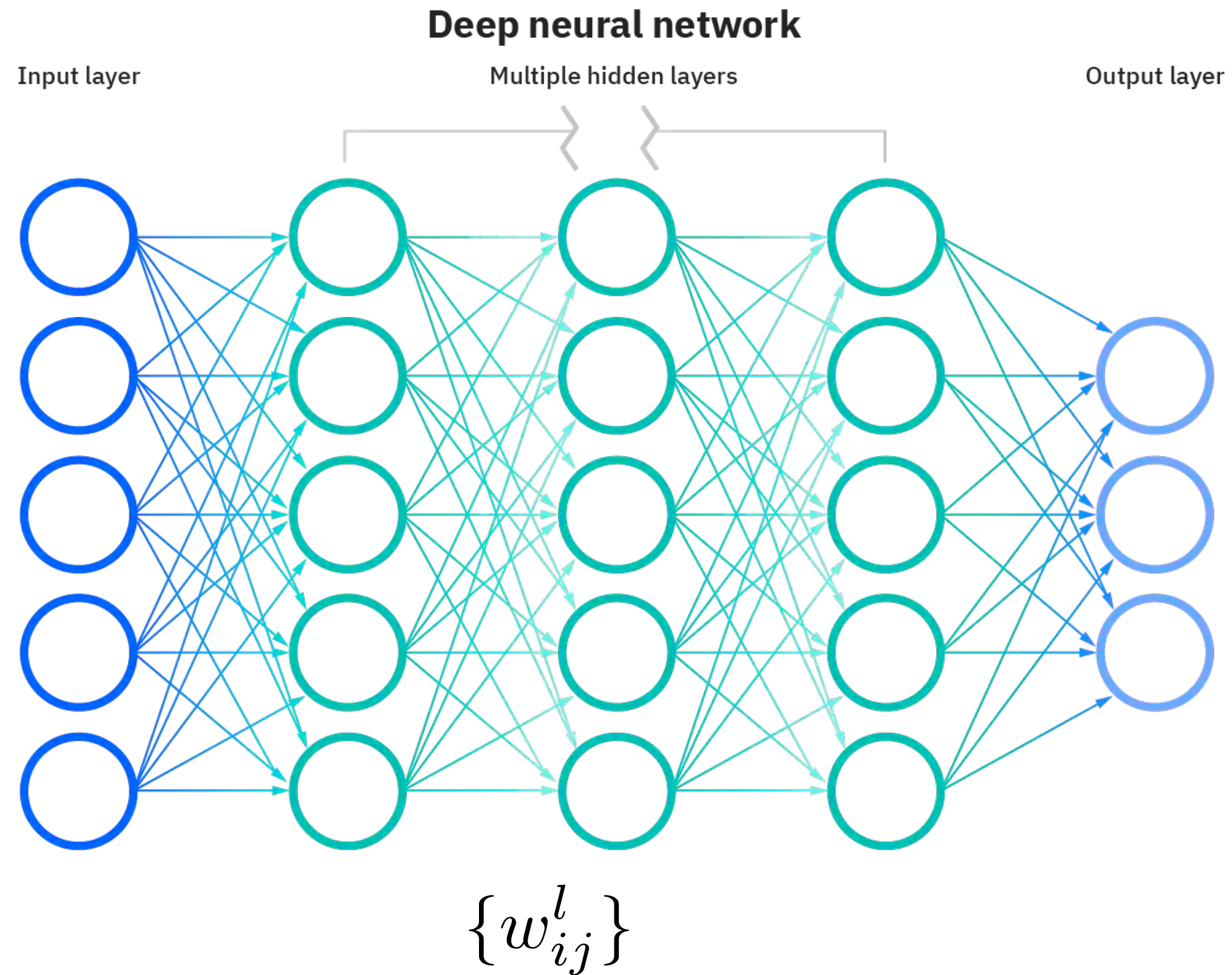
$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

Results:

$$f_{\theta}(\mathbf{x}) \rightarrow \mathbf{y}$$

“product”

# What do we get from this?



Case #2:



$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

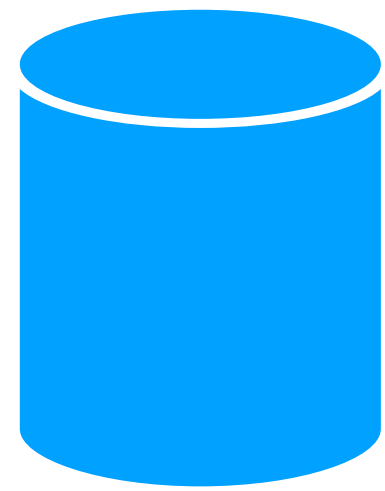
Results:

$$\{w_{ij}^l\}$$

“product”



# “Unsupervised” Representation Learning



$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$



In supervised learning setting — annotation

design “unsupervised” learning tasks

**original sentence:**

I wish you all the very best

**unsupervised learning task:**

output:    very     $\mathbf{y}$



input:    I wish you all the <mask> best     $\mathbf{x}$

# “Unsupervised” Representation Learning

original sentence:

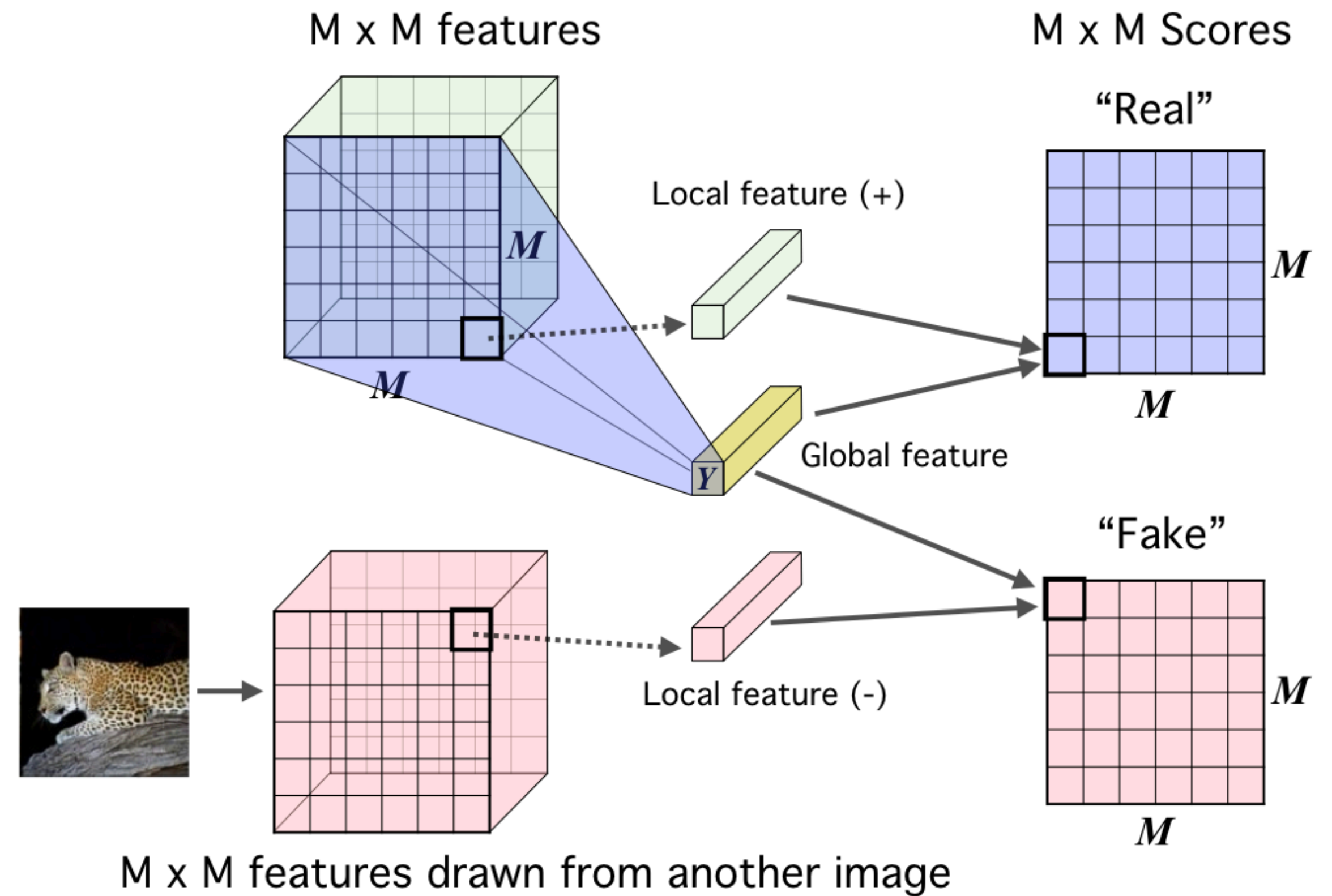
I wish you all the very best

unsupervised learning task:

input: I wish      output: you  
input: I wish you      output: all  
input: I wish you all      output: the  
⋮

Language Modeling Task

contrastive learning tasks:



# The Outcome of Unsupervised Representation Learning

$$f_{\theta}(x) \rightarrow y$$

input: I wish you all output: the

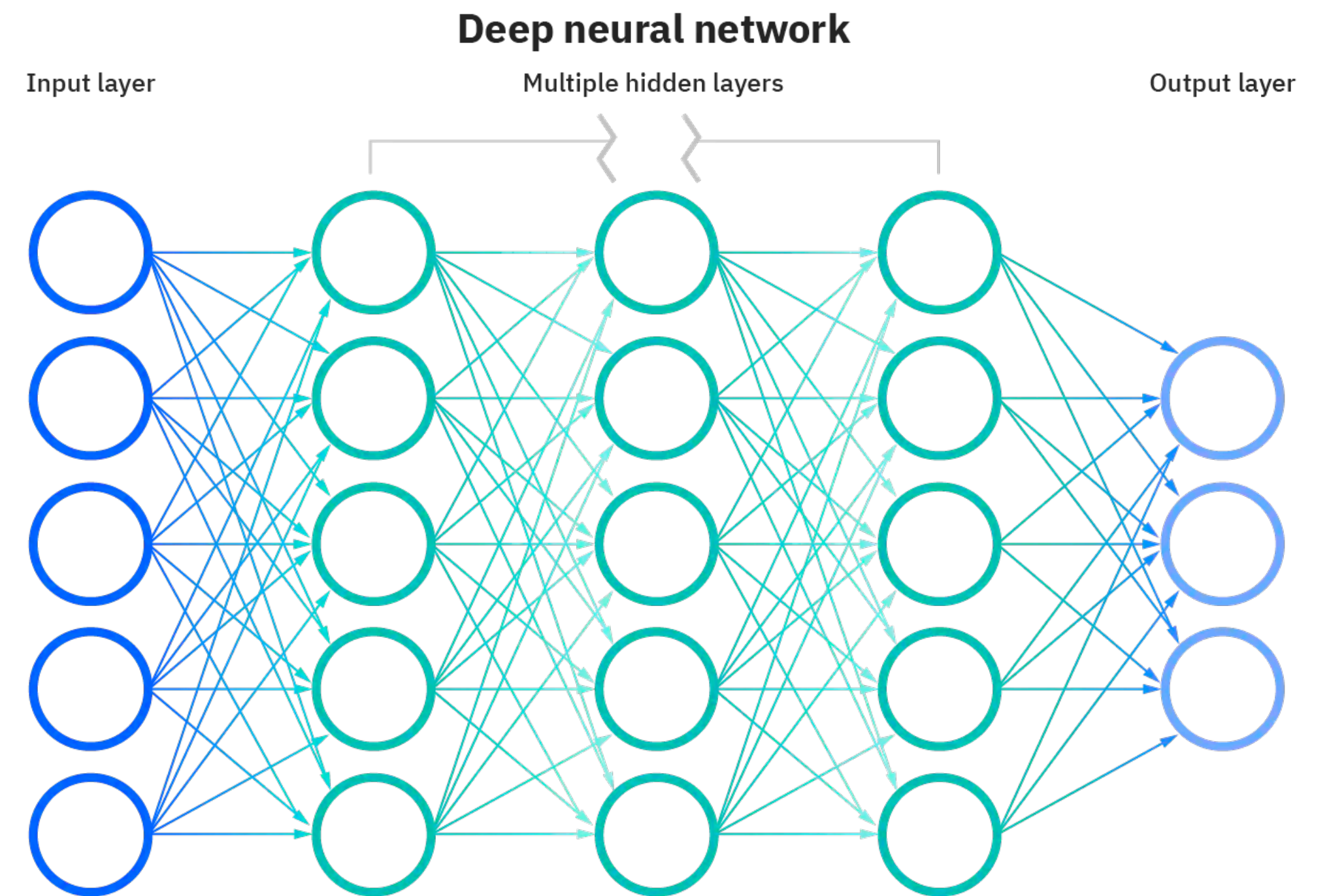
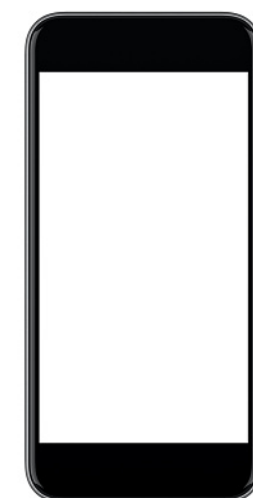
$\theta$

You need to do a better job understanding me.

Noted.

Yeah, make a note of that.

Here's your note:



$$\{w_{ij}^l\}$$

$$g_{\{\theta, \theta'\}}(x) \rightarrow y \quad \text{sentiment analysis?}$$

# Why this is useful?

1 y the irradiated and refrigerated chicken. Acceptance of radiopasteurization  
2 torehouse". Glendora dropped a chicken and a flurry of feathers, and went  
3 will specialize in steaks, chops, chicken and prime beef as well as Tom's fa  
4 ard as the one concerned with the chicken and the egg. Which came first? Is  
5 he millions of buffalo and prairie chicken and the endless seas of grass that  
6 "! "Come on, there's some cold chicken and we'll see what else". They wen  
7 ves to extend the storage life of chicken at a low cost of about 0.5 cent per  
8 CHICKEN CADILLAC# Use one 6-ounce chicken breast for each guest. Salt and pe  
9 ion juice, to about half cover the chicken breasts. Bake slowly at least one-  
10 d, in butter. Sprinkle over top of chicken breasts. Serve each breast on a th  
11 around, they had a hard time". #CHICKEN CADILLAC# Use one 6-ounce chicken  
12 successful, and the shelf life of chicken can be extended to a month or more  
13 ay from making a cake, building a chicken coop, or producing a book, to found  
14 , they decided, but a deck full of chicken coops and pigpens was hardly suita  
15 im. "Johnny insisted on cooking a chicken dinner in my honor- he's always bee  
16 nutes. Kid Ory, the trombonist chicken farmer, is also one of the solid a  
17 y Johnson reaching around the wire chicken fencing, which half covered the tr  
18 yes glittering behind dull silver chicken fencing. "That was Tee-wah I was t  
19 wine in the pot roast or that the chicken had been marinated in brandy, and  
20 yed this same game and called it "Chicken". He could not go through the f  
21 f the Mexicans hiding in a little chicken house had passed through his head,  
22 I'll never forget him cleaning the chicken in the tub". A story, no doubt  
23 . Organ meats such as beef and chicken liver, tongue and heart are planne  
24 p. "Miss Sarah, I can't cut up no chicken. Miss Maude say she won't". Aga  
25 pot. "What is it"? he asked. "Chicken", Mose said, and theatrically licke  
26 im"? Adam shook his head. "Chicken", Mose said. She was a child too m

Where is "chicken" ?

I want to eat some fruits today.  
Therefore, I will eat <mask> tonight.

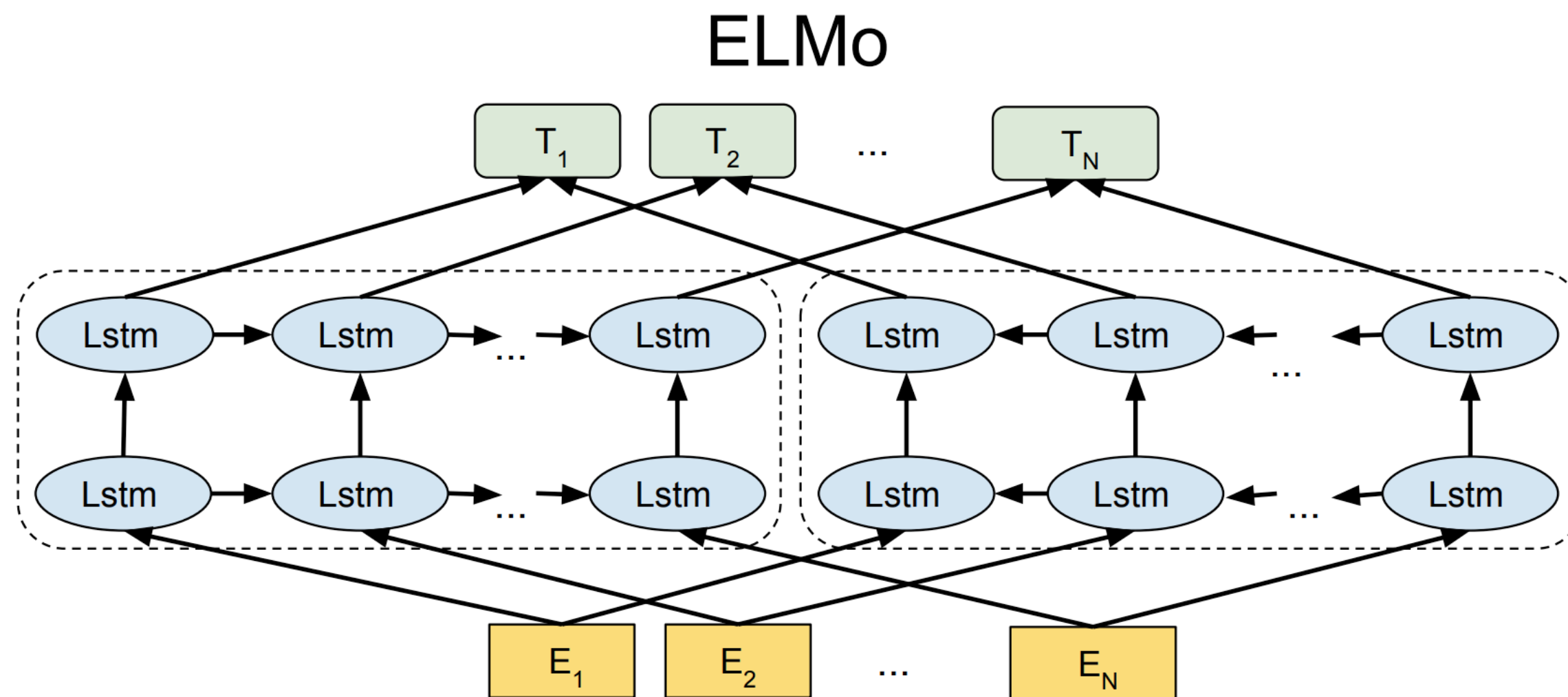
apple  
watermelon

$\theta$  representation -> semantic meaning

“transfer”

useful in other tasks

# ELMo (Embeddings from Language Models)



$$\sum_{k=1}^N ( \log p(t_k | t_1, \dots, t_{k-1}; \Theta_x, \vec{\Theta}_{LSTM}, \Theta_s) \\ + \log p(t_k | t_{k+1}, \dots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) )$$

Bidirectional Language Model

# T5 (Text-to-Text Transfer Transformer)

Multi-task learning – using one set of parameters for many different tasks

