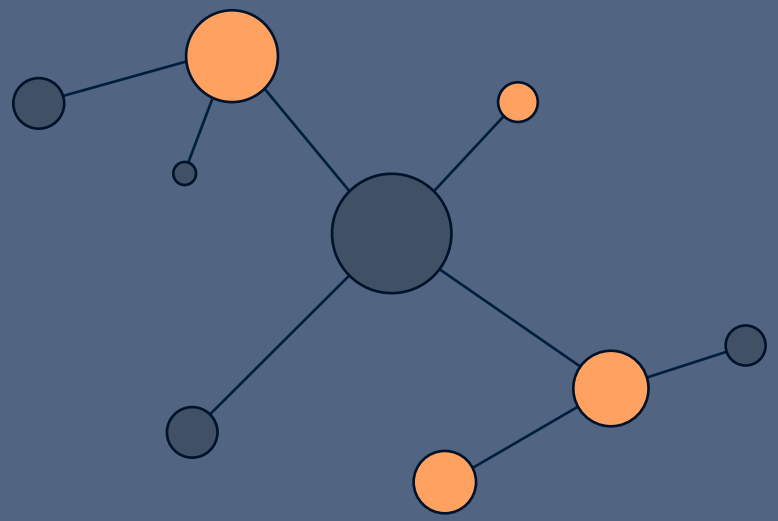
Unsupervised learning

By Jip de Kok



Maastricht UMC+

Introduction


Yay, big data!?

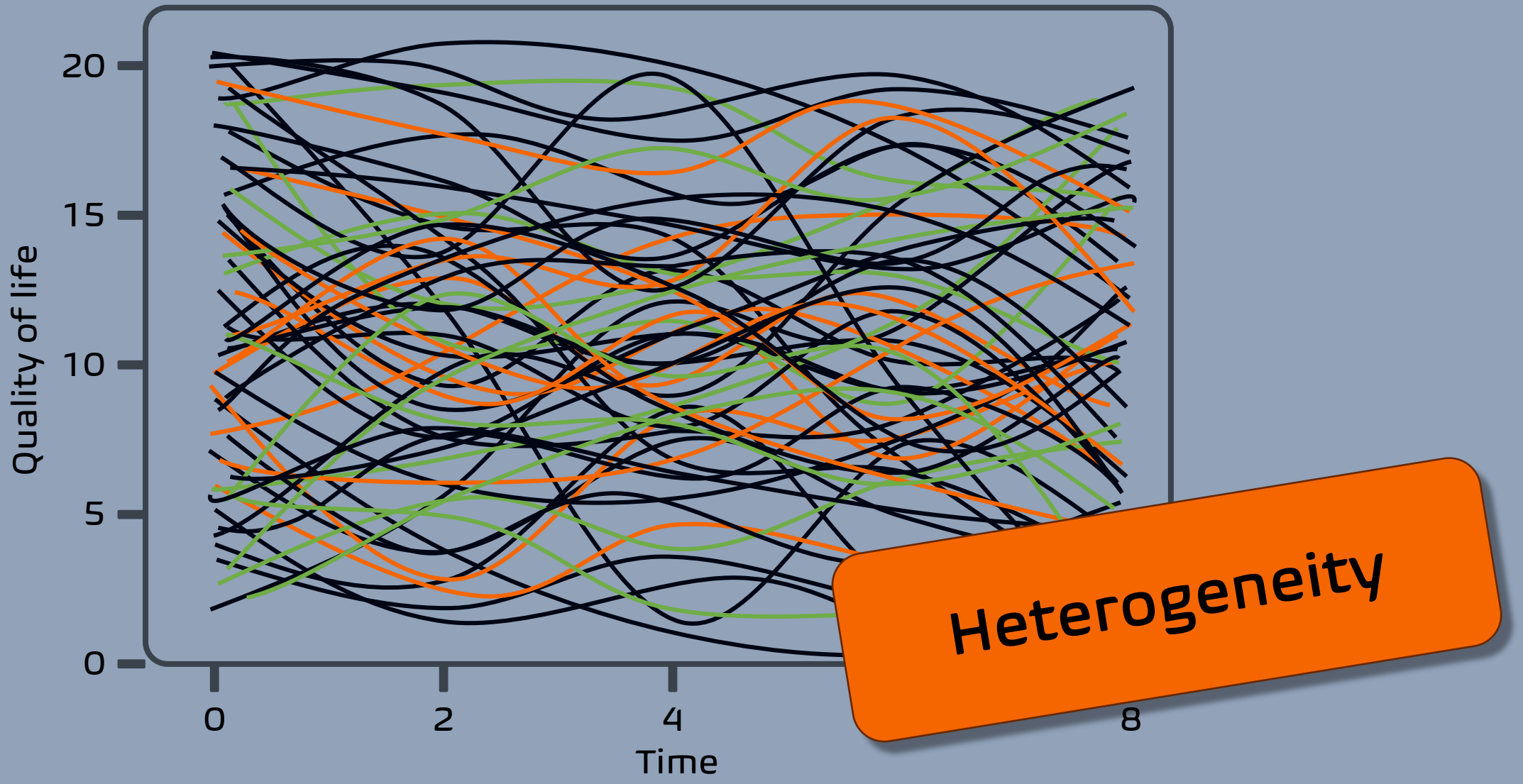
	Variable 1	Variable 2	...	Variable 10
Pt. 1	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 2	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 3	~~~~~	~~~~~	~~~~~	~~~~~
~~	~~~~~	~~~~~	~~~~~	~~~~~
~~	~~~~~	~~~~~	~~~~~	~~~~~
~~	~~~~~	~~~~~	~~~~~	~~~~~
~~	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 49,988	~~~~~	~~~~~	~~~~~	
Pt. 49,989	~~~~~	~~~~~		
Pt. 50,000	~~~~~	~~~~~		

Many observations

	Variable 1	Variable 2	...	Variable 50,000
Pt. 1	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 2	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 3	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 4	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 5	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 6	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 7	~~~~~	~~~~~	~~~~~	~~~~~
Pt. 8	~~~~~	~~~~~	~~~~~	
Pt. 9	~~~~~	~~~~~	~~	
Pt. 10	~~~~~	~~~~~	~~~	

Many variables
(high dimensionality)

PROBLEM:



Possible solution:

Unsupervised learning

Today's topics

1. Terminology
2. Introduction to clustering
3. Clustering algorithms
4. Dimensionality reduction

Learning goals

- Become familiar with Data Science terminology
- Have an intuitive understanding of clustering
- Know and understand the most familiar clustering algorithms
- Understand the purpose of dimensionality reduction
- Know and understand some dimensionality reduction techniques
- Be ready for the clustering tutorial!

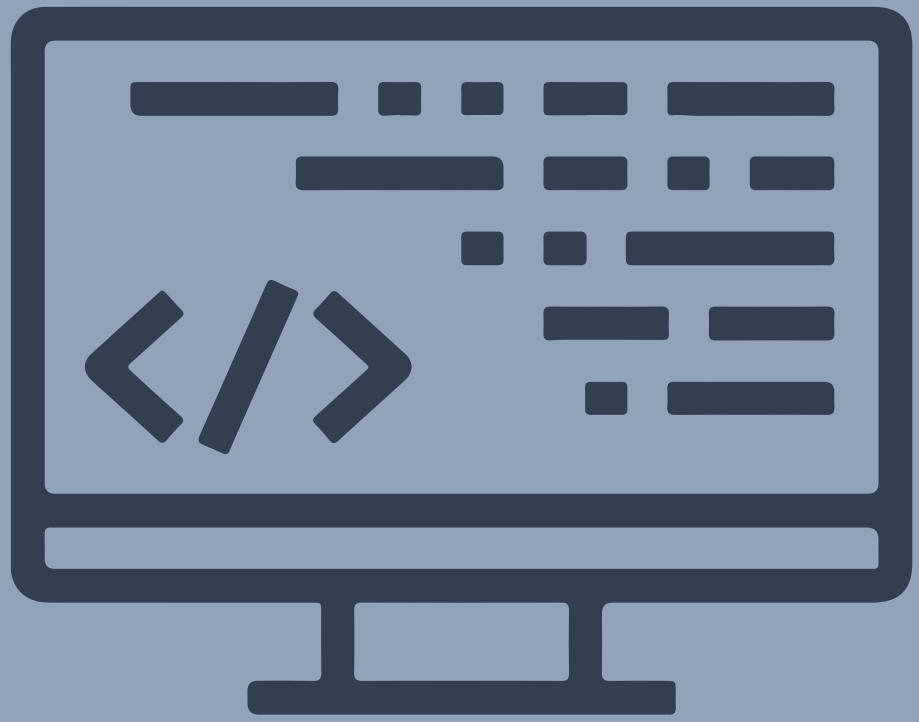
Today's topics

1. Terminology
2. Introduction to clustering
3. Clustering algorithms
4. Dimensionality reduction

Terminology

1. Programming
2. Artificial Intelligence
3. Machine Learning
4. Supervised learning
5. Unsupervised learning

Terminology



Programming

Programming

BMI < 18.5 → underweight
18.5 < BMI < 25 → healthy weight
25 < BMI < 30 → overweight
BMI > 30 → obesity

Model

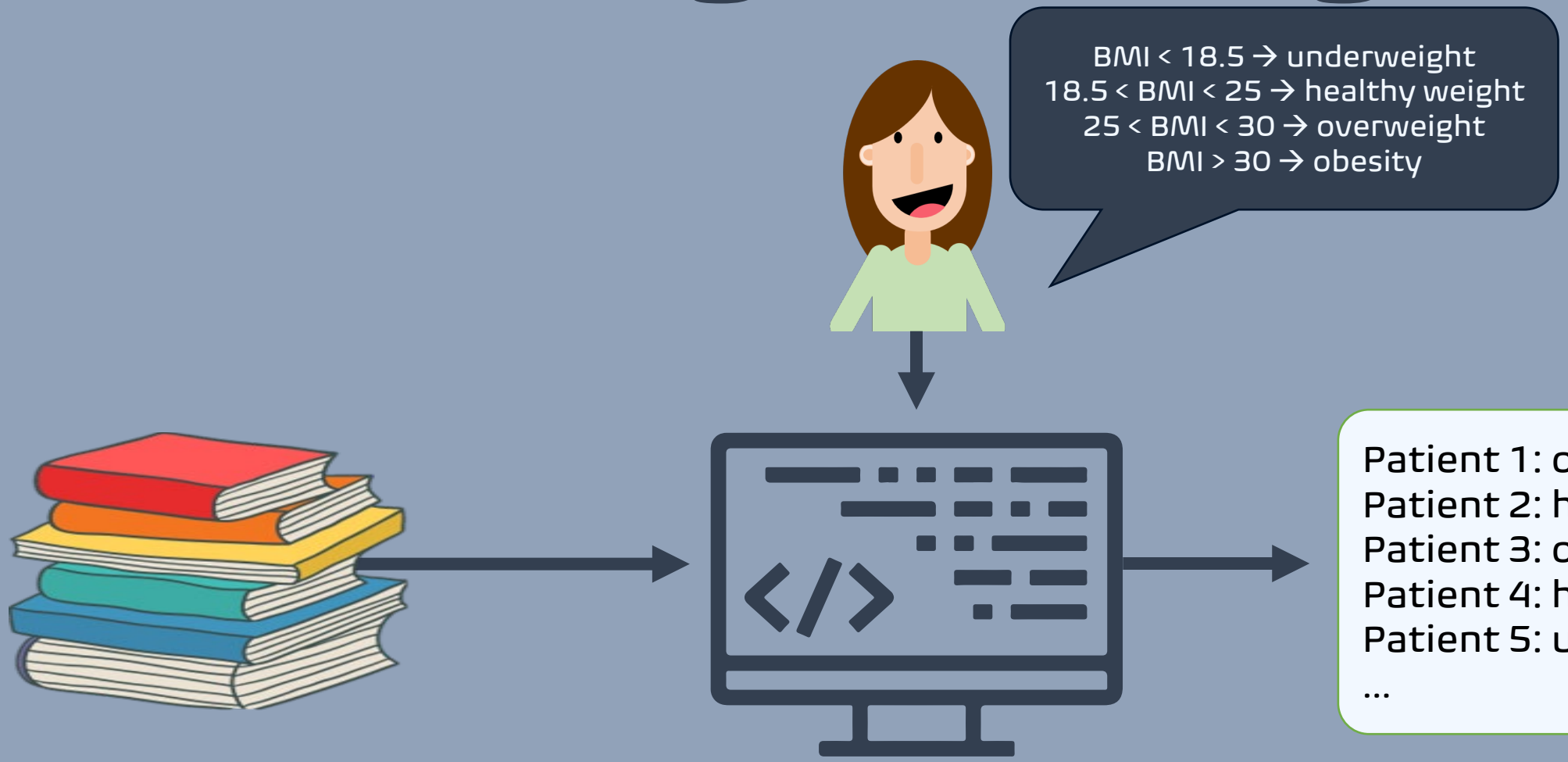
Programming

Input

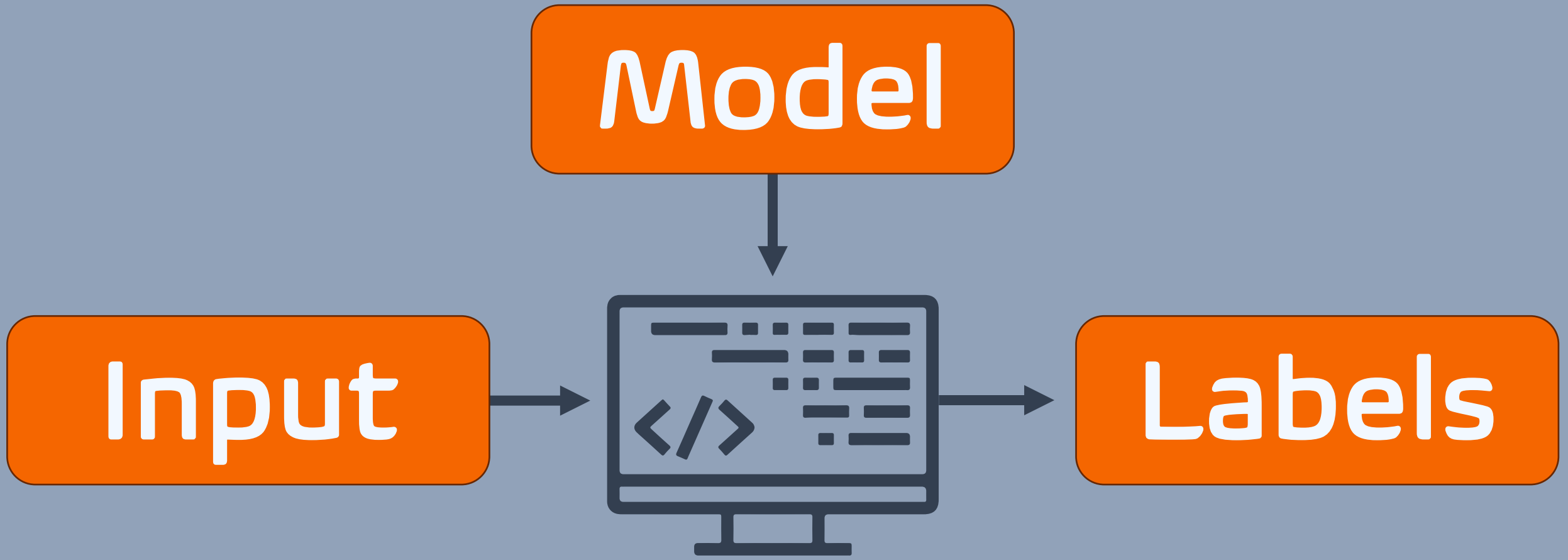
Programming

Terminology

Programming



Programming



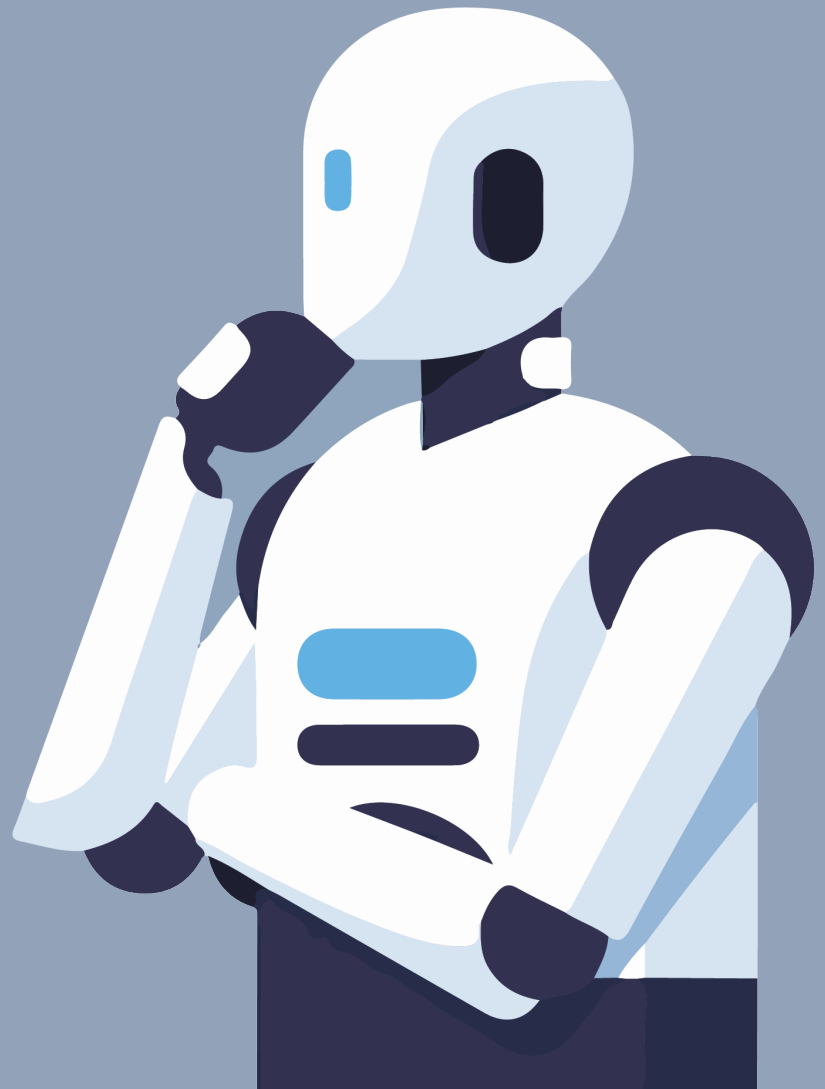
Terminology

1. ~~Programming~~
2. Artificial Intelligence
3. Machine Learning
4. Supervised learning
5. Unsupervised learning

Artificial Intelligence (AI)
vs.
Machine Learning (ML)

Artificial Intelligence (AI)

AI is an umbrella term that is broadly defined as the capability of a machine to imitate intelligent human behaviour, such as learning and problem-solving.



Machine Learning (ML)

The field of study that gives computers the ability to learn without explicitly being programmed.

Terminology

1. ~~Programming~~
2. ~~Artificial Intelligence~~
3. ~~Machine Learning~~
4. Supervised learning
5. Unsupervised learning

Terminology

Sounds familiar?

Supervised
vs.
Unsupervised

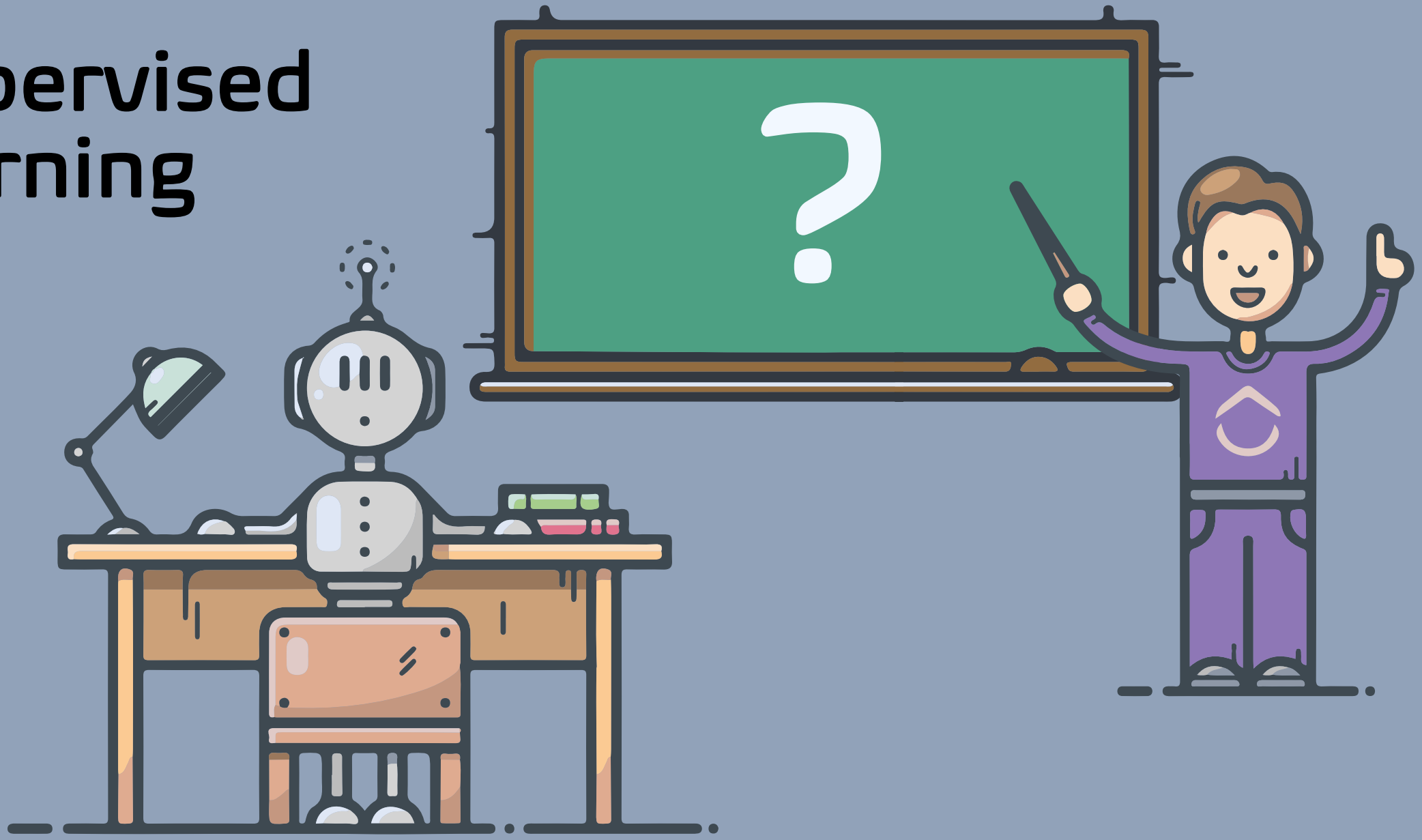
Terminology

Supervised

Unsupervised

Terminology

Supervised learning



Terminology

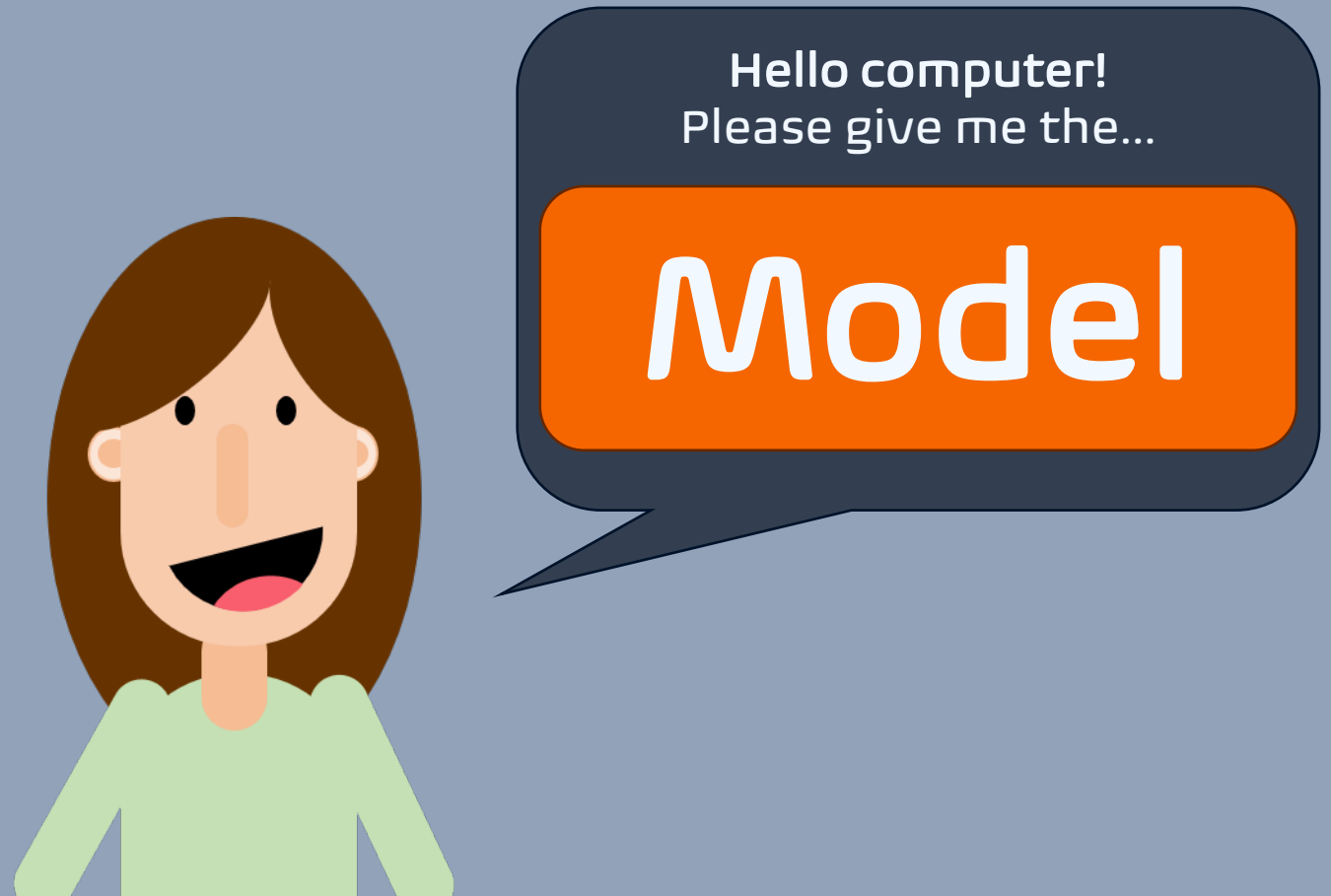
Input:

Patient 1: BMI = 14
Patient 2: BMI = 22
Patient 3: BMI = 28
Patient 4: BMI 31
...
...
...
...
...
...
...
Patient 50,000: BMI = 26

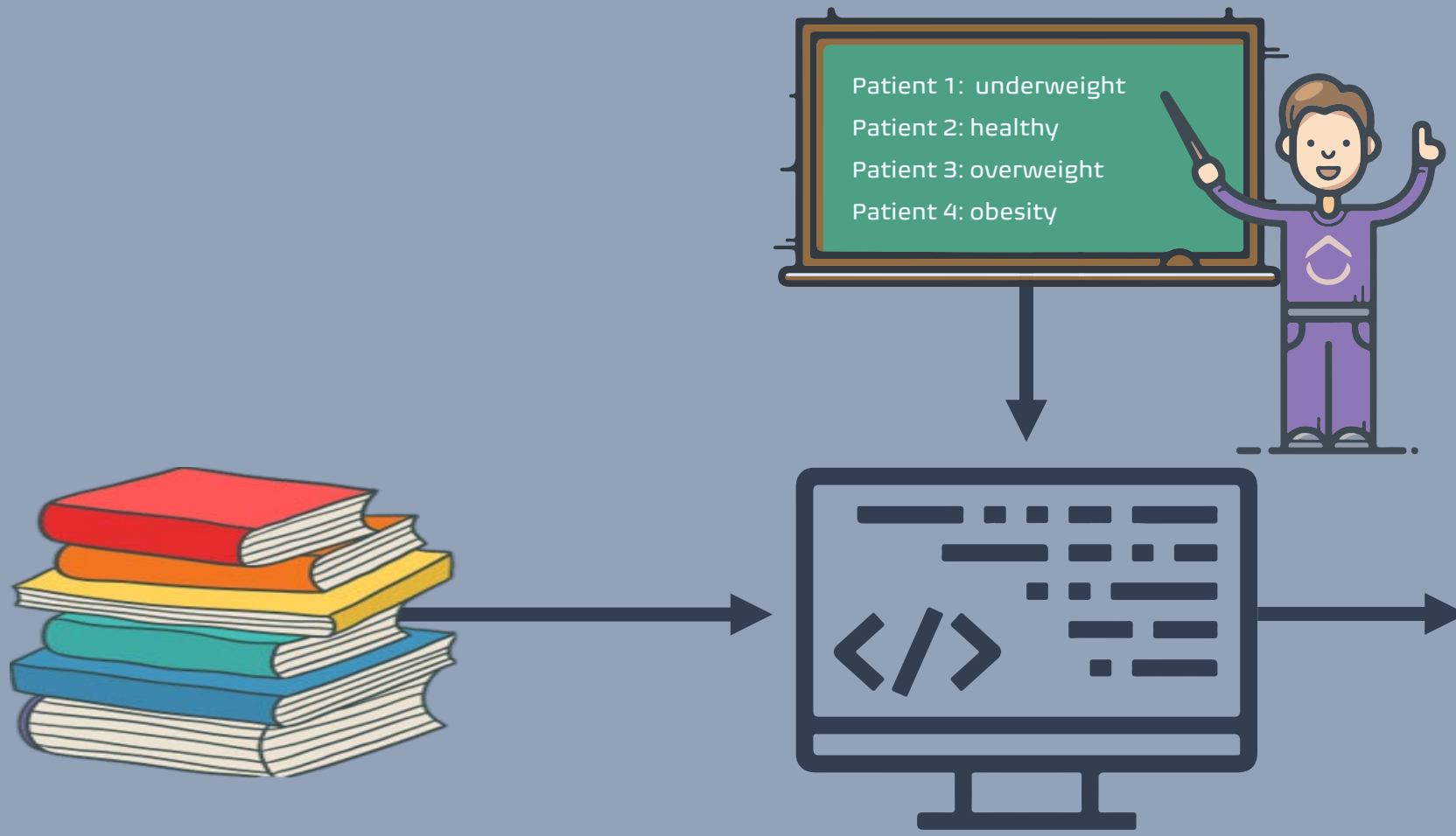
Labels:

Patient 1: underweight
Patient 2: healthy weight
Patient 3: overweight
Patient 4: obesity
...
...
...
...
...
...
...
Patient 50,000: BMI = overweight

Supervised learning

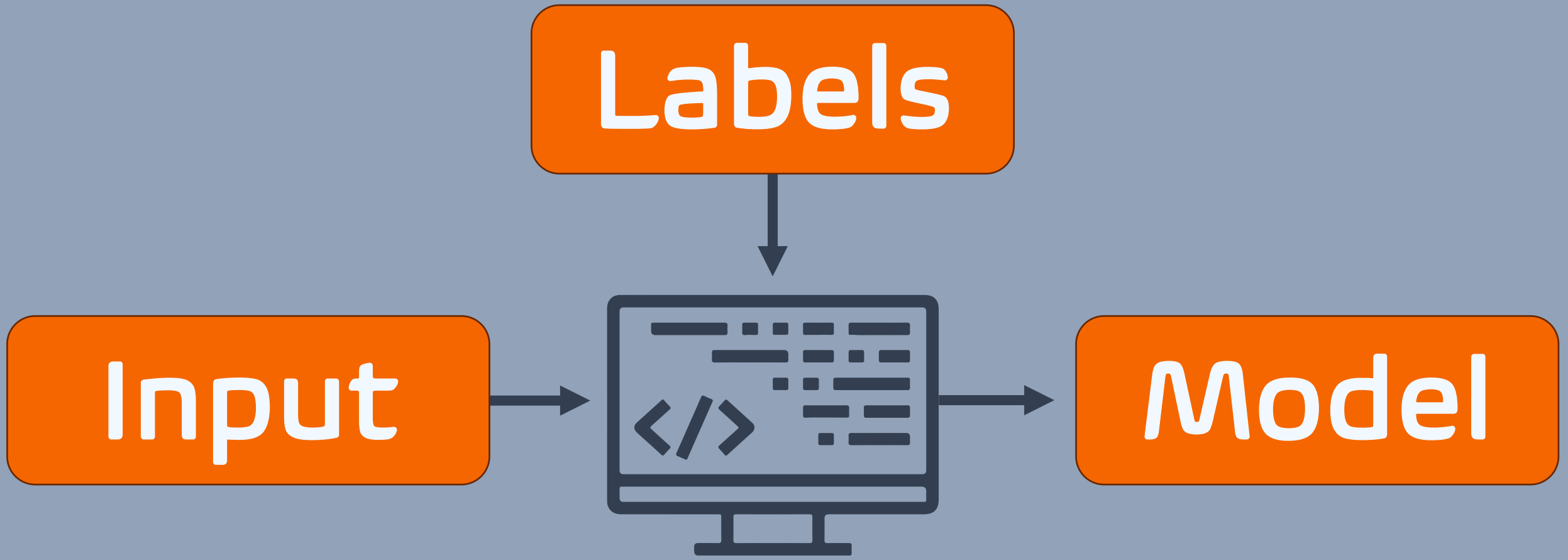


Supervised learning

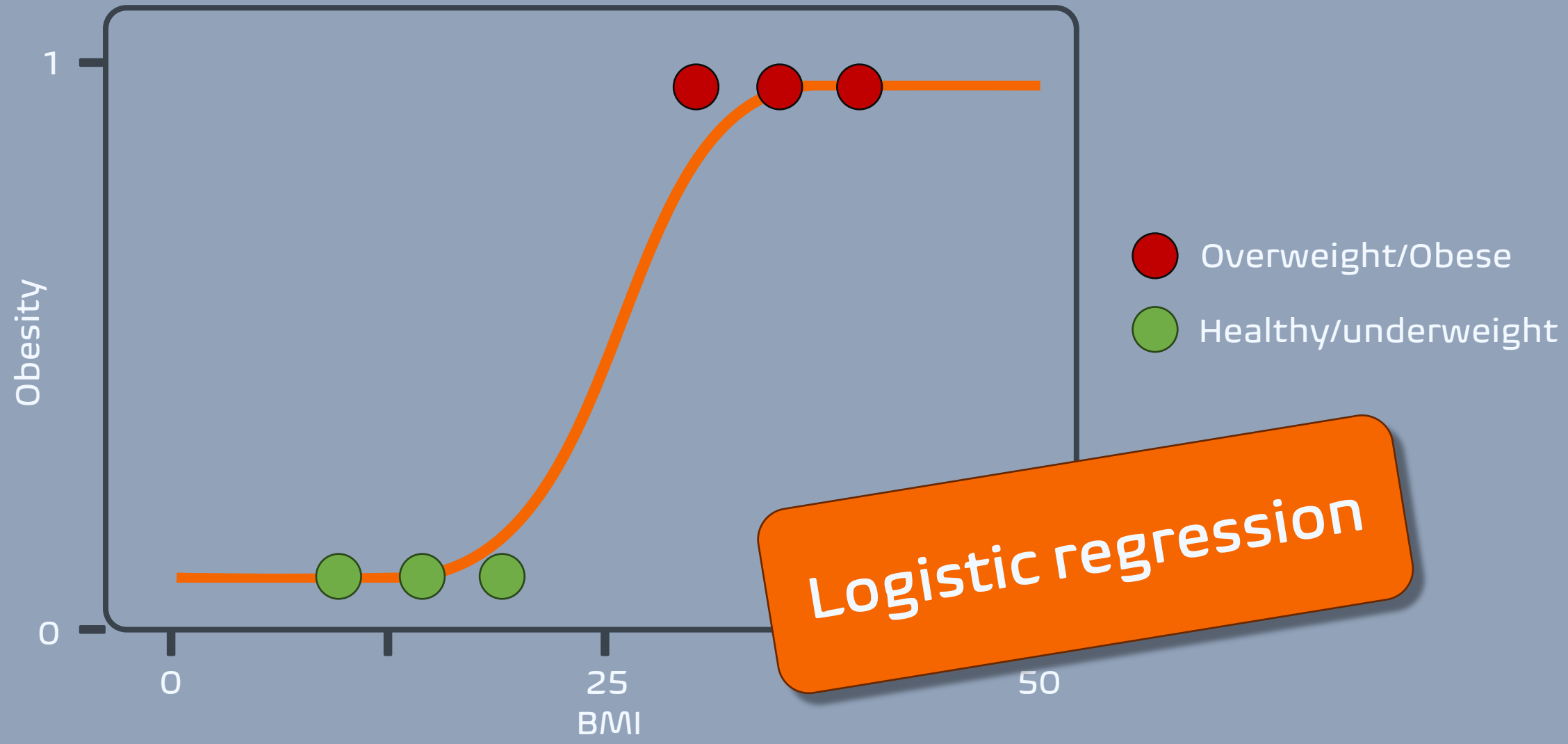


BMI < 18.5 = underweight
18.5 < BMI < 25 = healthy
25 < BMI < 30 = overweight
BMI > 30 = obesity

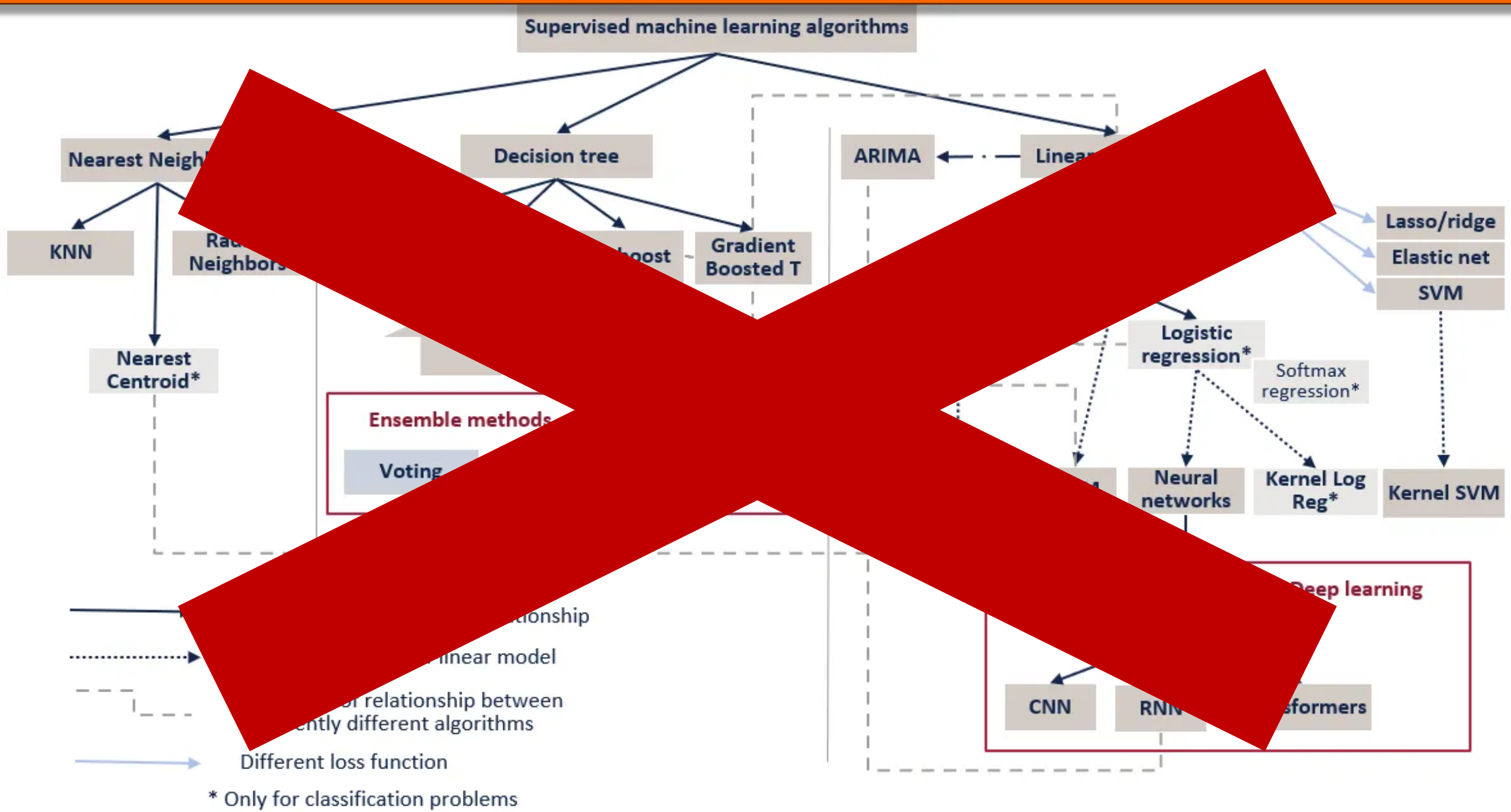
Supervised learning



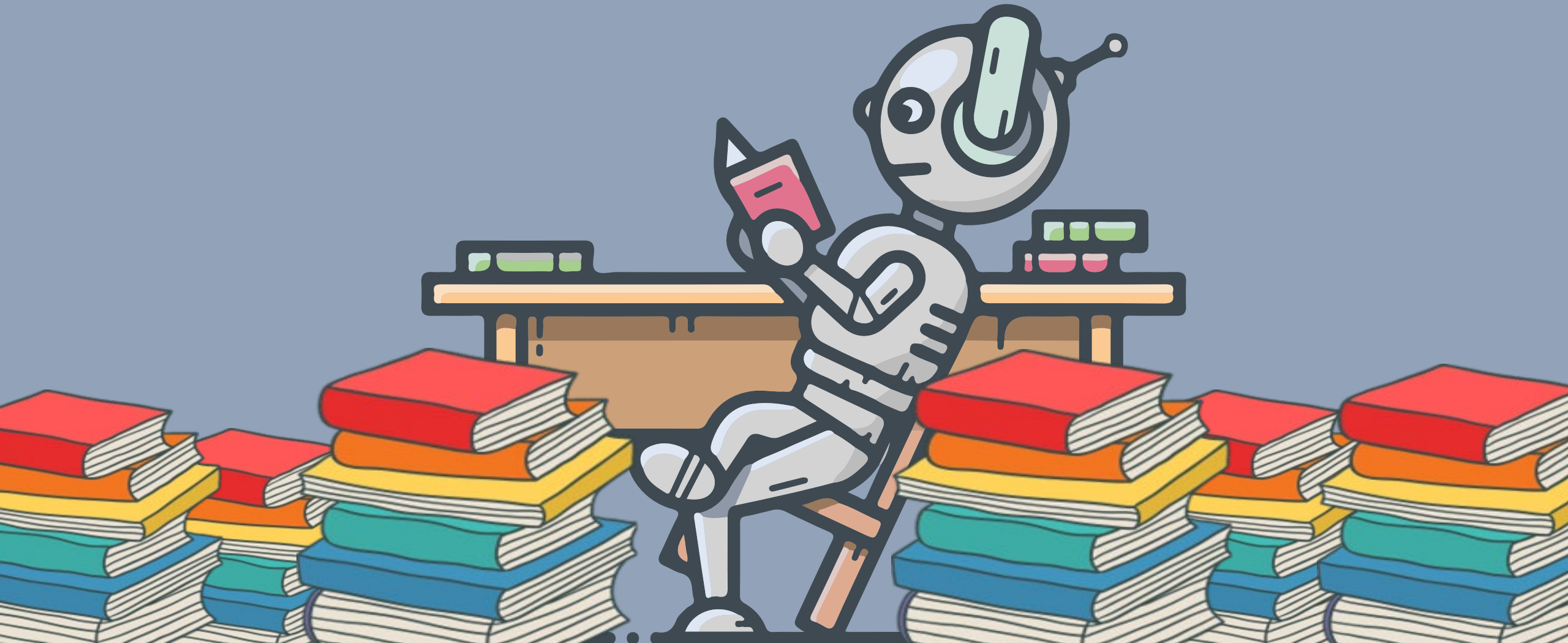
Supervised learning



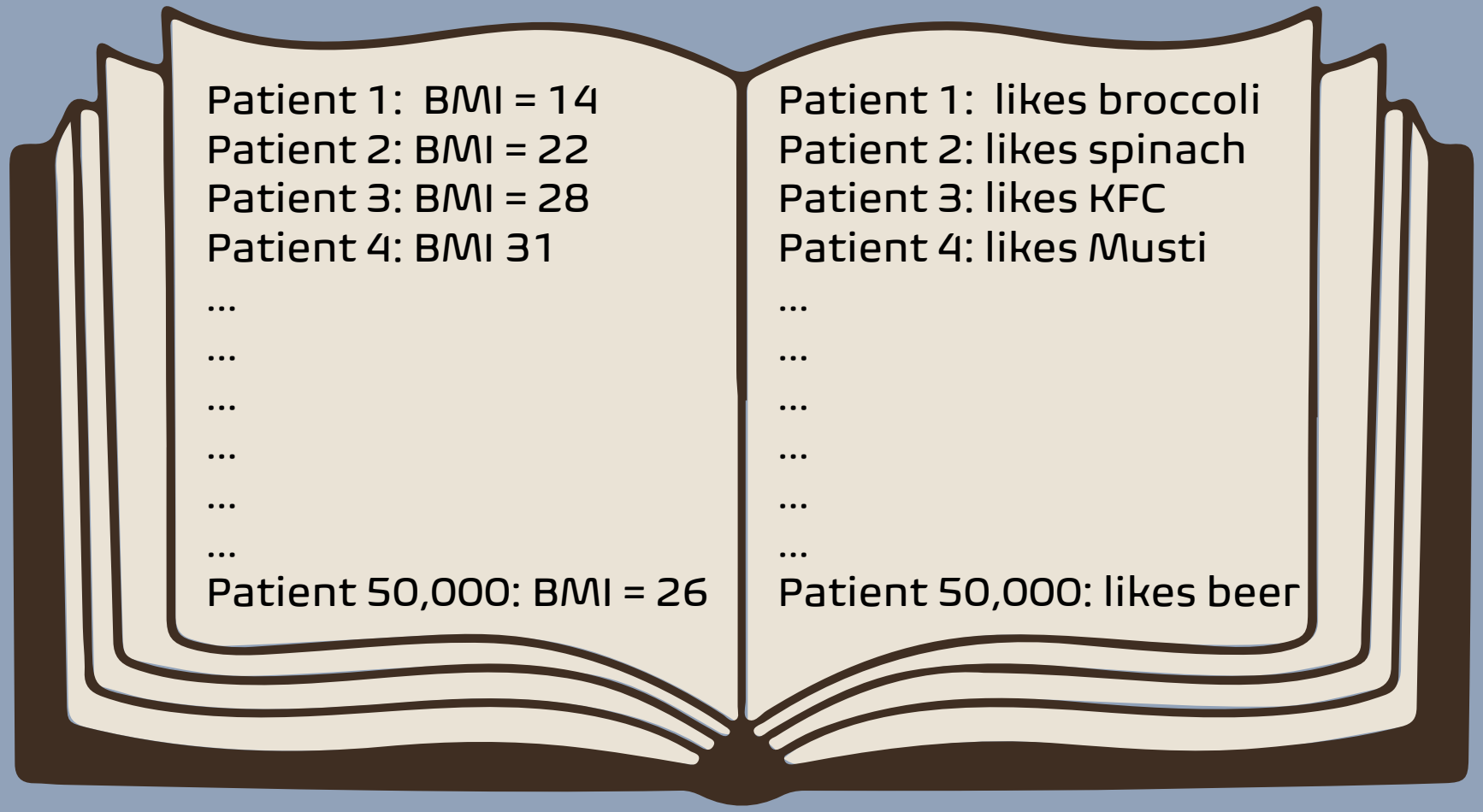
Terminology

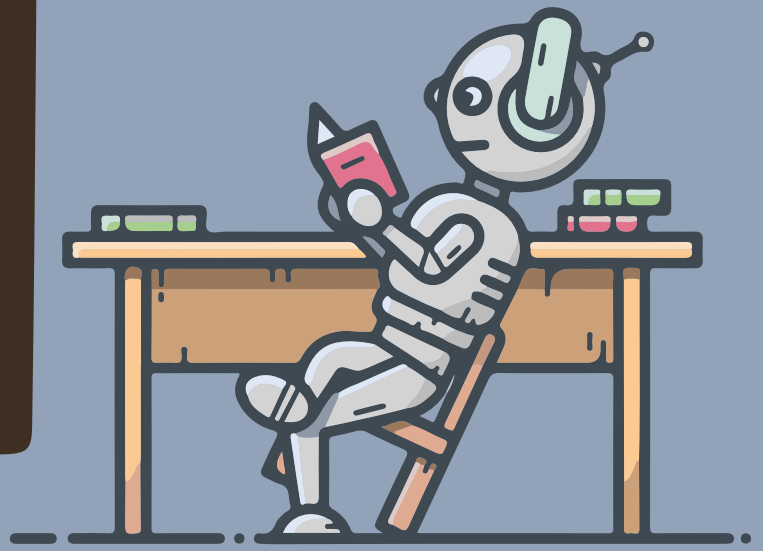


Unsupervised learning

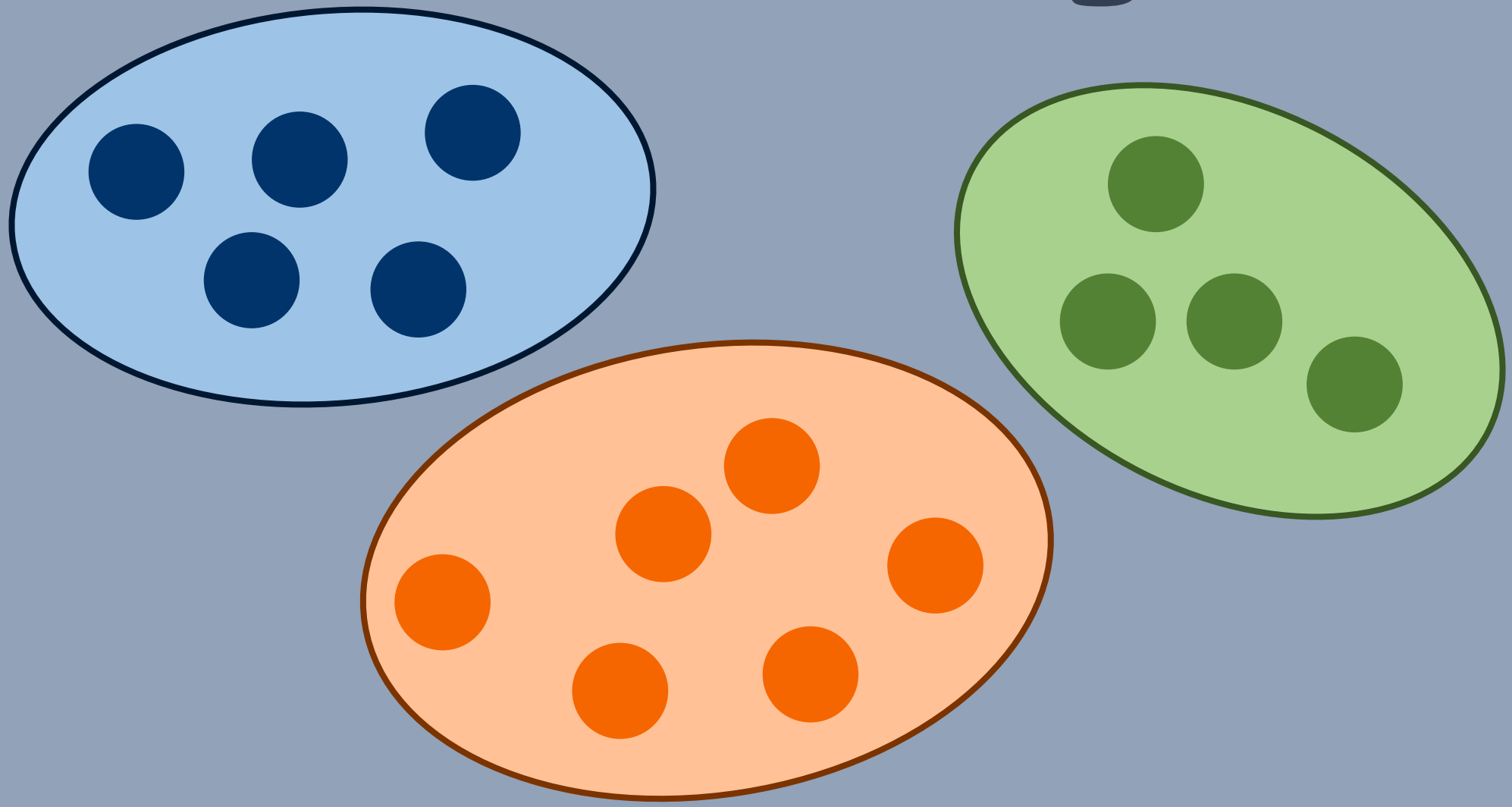


Unsupervised learning



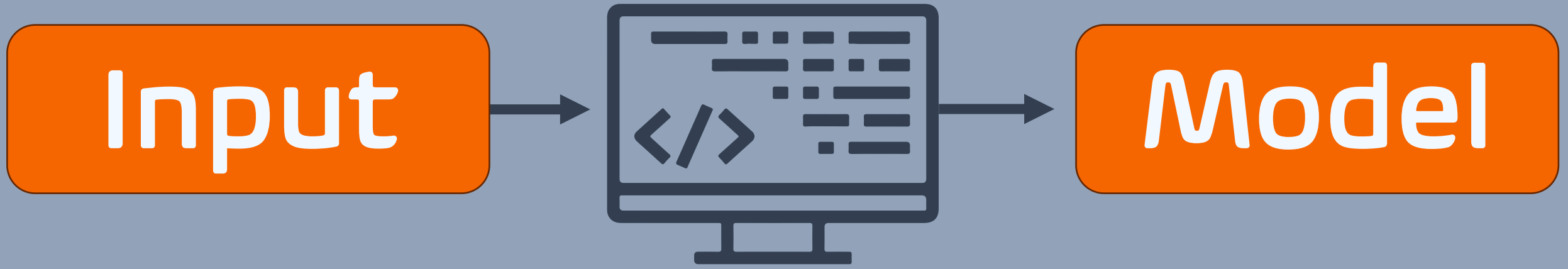


Clustering



Association rules

Unsupervised learning

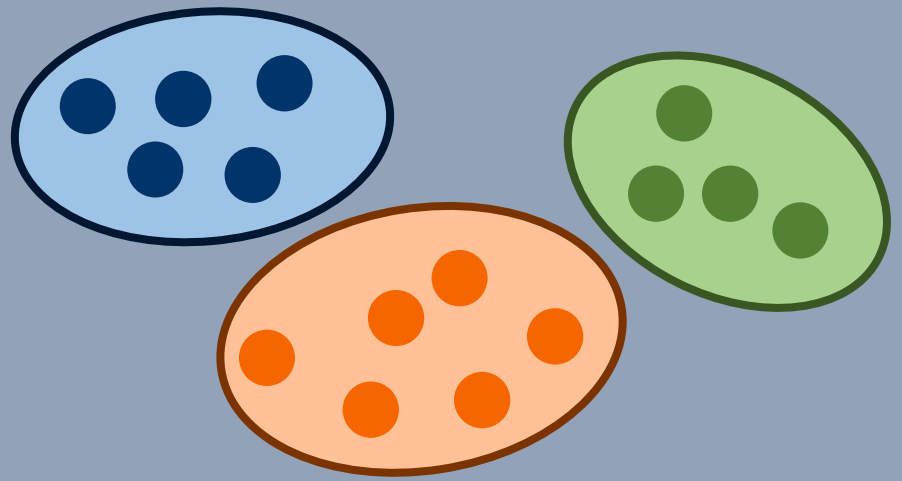


Today's topics

1. ~~Terminology~~
2. **Introduction to clustering**
3. Clustering algorithms
4. Dimensionality reduction

What is clustering?

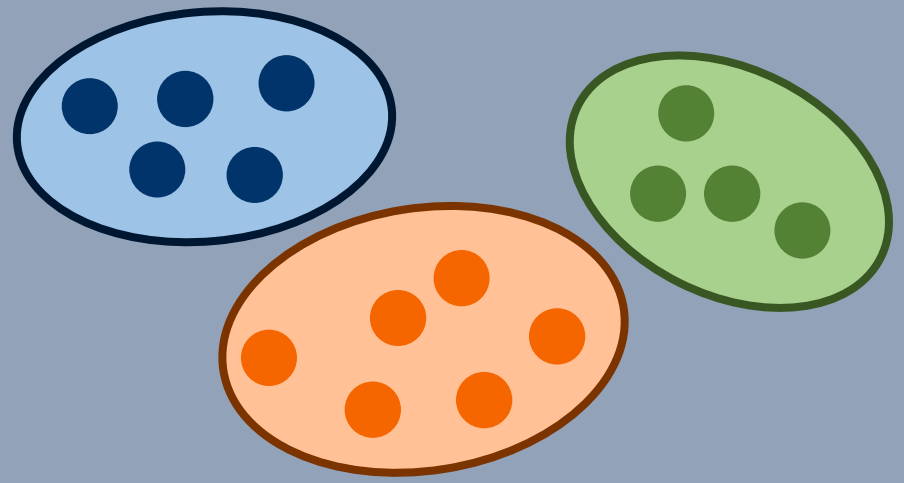
"Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)." – Wikipedia 2023



Why would one cluster?

- Exploratory data analysis
  - Are there groups in my data?
  - Are there outliers?

- Hypothesis generation
  - Do clusters differ significantly?
  - Do hypotheses hold for all clusters?

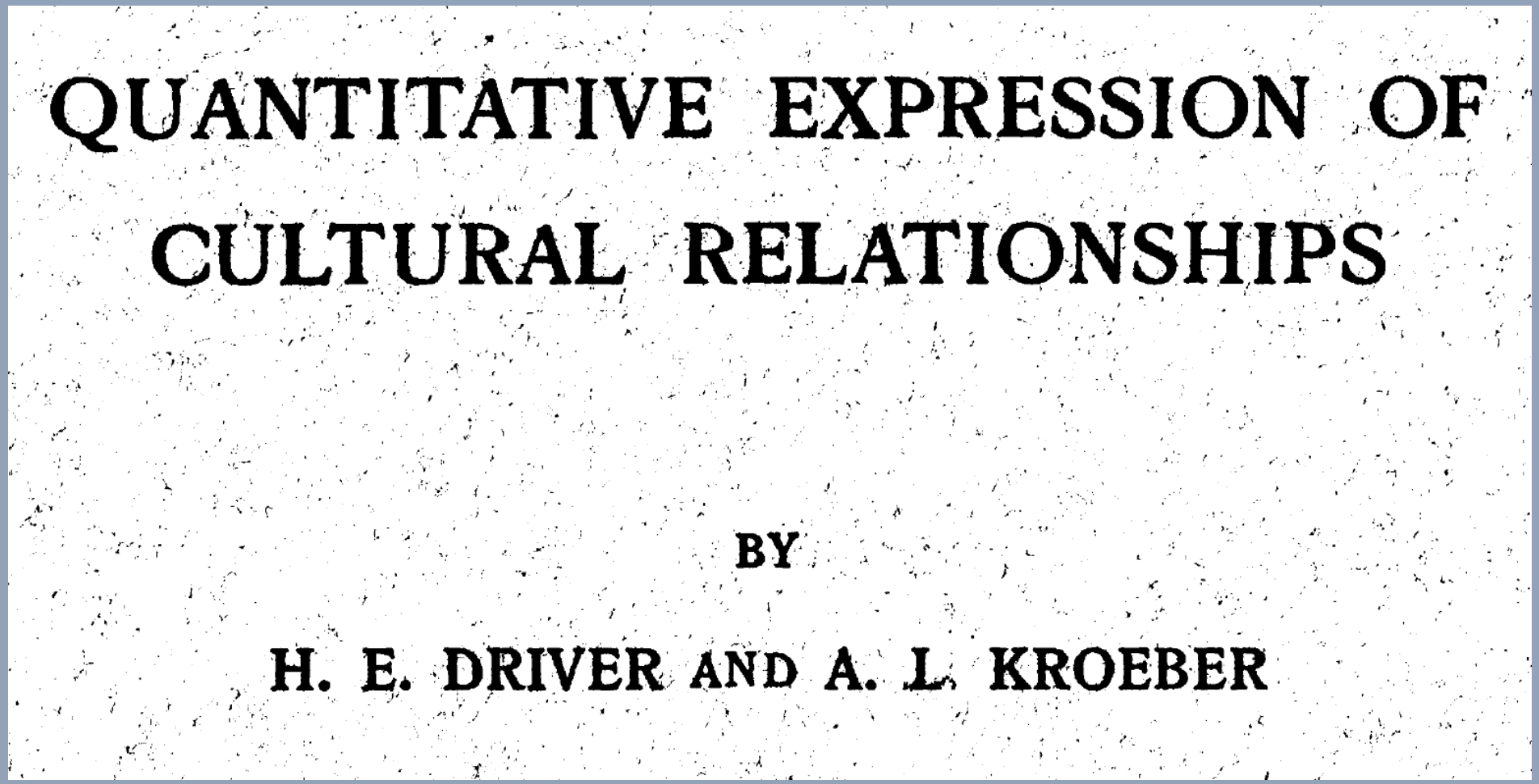


An example


Let's travel back in time

Introduction to clustering

1932



QUANTITATIVE EXPRESSION OF CULTURAL RELATIONSHIPS

BY

H. E. DRIVER AND A. L. KROEBER

Introduction to clustering

Hidatsa

Assiniboin

Ponca

Introduction to clustering

	Hidatsa	Assiniboin	Ponca
Single feather	1	0	1
Wooden shield	0	1	0
Spears	0	1	0
Robe of deer fur	1	1	1
Painted houses	0	1	1
Elk antler spoons	0	1	1
Tattoos	1	1	1

Which tribes are most similar?

Let's calculate similarity

What is wrong?

- Similarity between tribe A & B =

$$\frac{number\ of\ overlapping\ present\ traits\ between\ tribe\ A\ \&\ B}{Total\ number\ of\ present\ traits\ in\ tribe\ A}$$

	Hidatsa	Assiniboin	Ponca
Hidatsa	1	0.67	1
Assiniboin	0.33	1	0.67
Ponca	0.60	0.8	1

Let's calculate similarity

- Similarity between tribe A & B =

$$\left(\frac{overlap\ A\ \&\ B}{number\ of\ traits\ in\ A} + \frac{overlap\ A\ \&\ B}{number\ of\ traits\ in\ B}\right)/2$$

	Hidatsa	Assiniboin	Ponca
Hidatsa	1	0.5	0.8
Assiniboin	0.5	1	0.73
Ponca	0.8	0.73	1

Let's calculate similarity

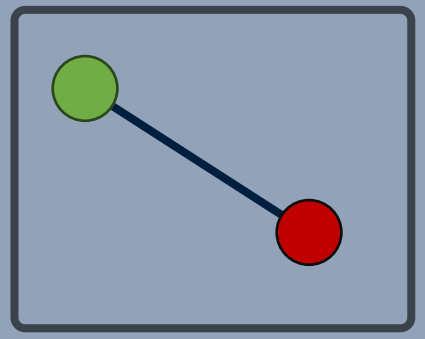
- Similarity between tribe A & B =

$$\left(\frac{overlap\ A\ \&\ B}{number\ of\ traits\ in\ A} + \frac{overlap\ A\ \&\ B}{number\ of\ traits\ in\ B}\right)/2$$

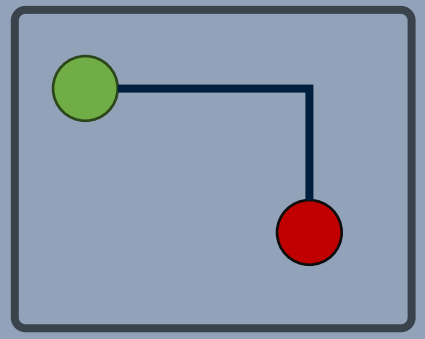
	Hidatsa	Assiniboin	Ponca
Hidatsa			
Assiniboin	0.5		
Ponca	0.8	0.73	

Introduction to clustering

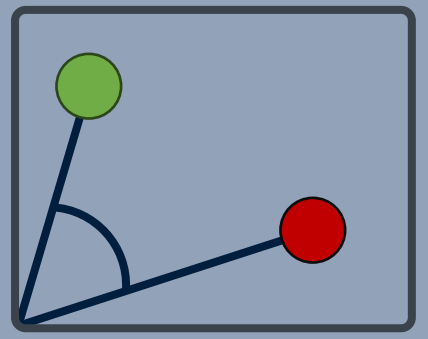
Euclidean



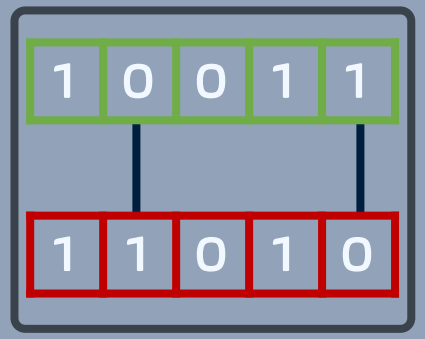
Manhatten



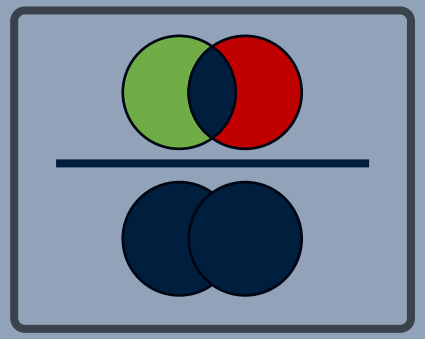
Cosine



Hamming



Jaccard



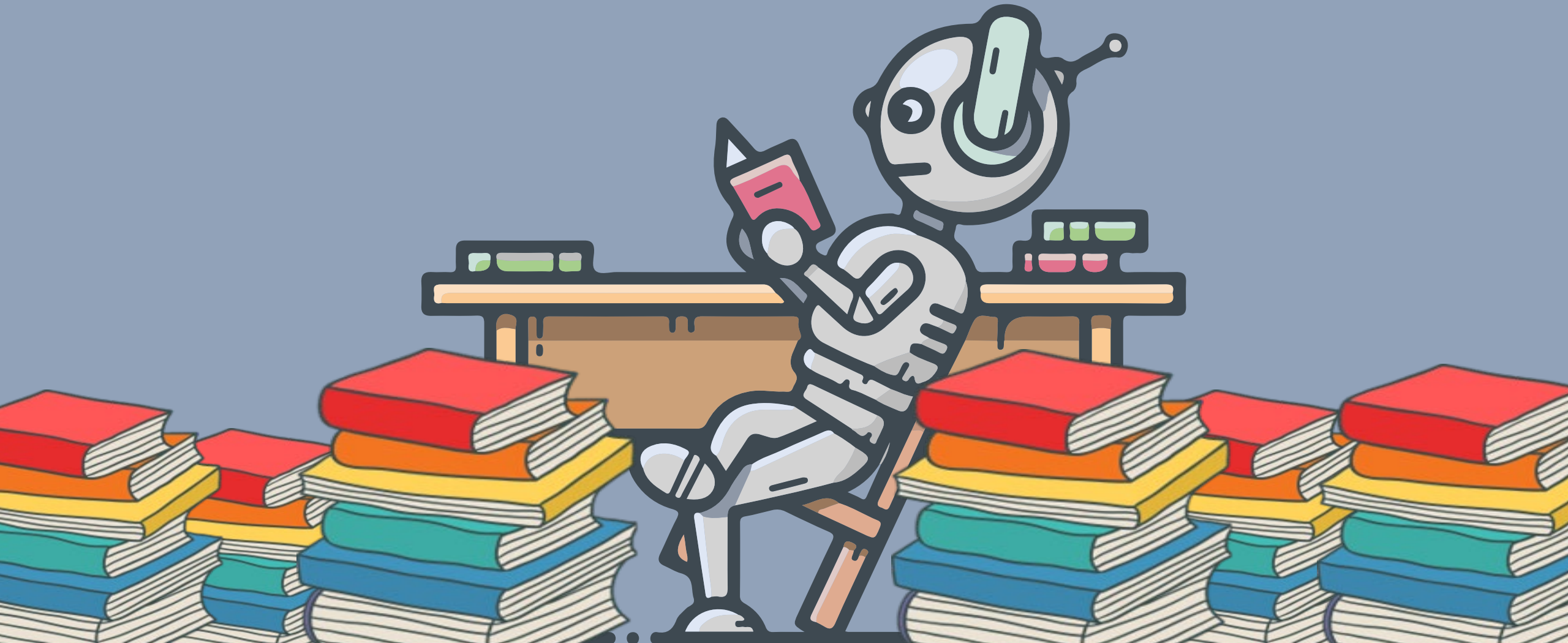
And many more...

Let's calculate Hamming distance

	Hidatsa	4	Assiniboin	5	Ponca
Single feather	1	←	0	→	1
Wooden shield	0	←	1	→	0
Spears	0	←	1	→	0
Robe of deer fur	1		1		1
Painted houses	1		1	→	0
Elk antler spoons	0	←	1		1
Tattoos	1		1	→	0

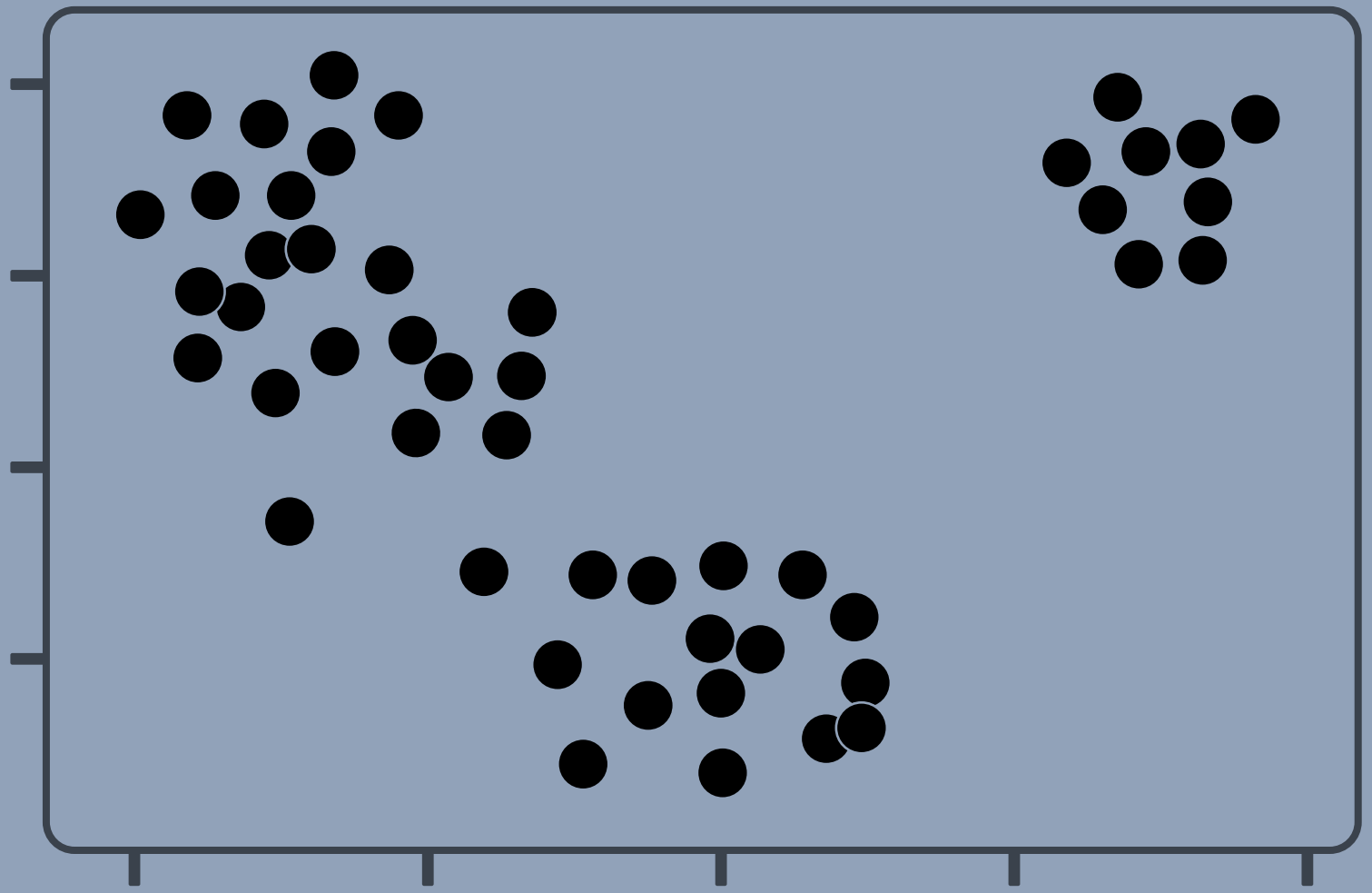
Introduction to clustering

How does the computer see clusters?



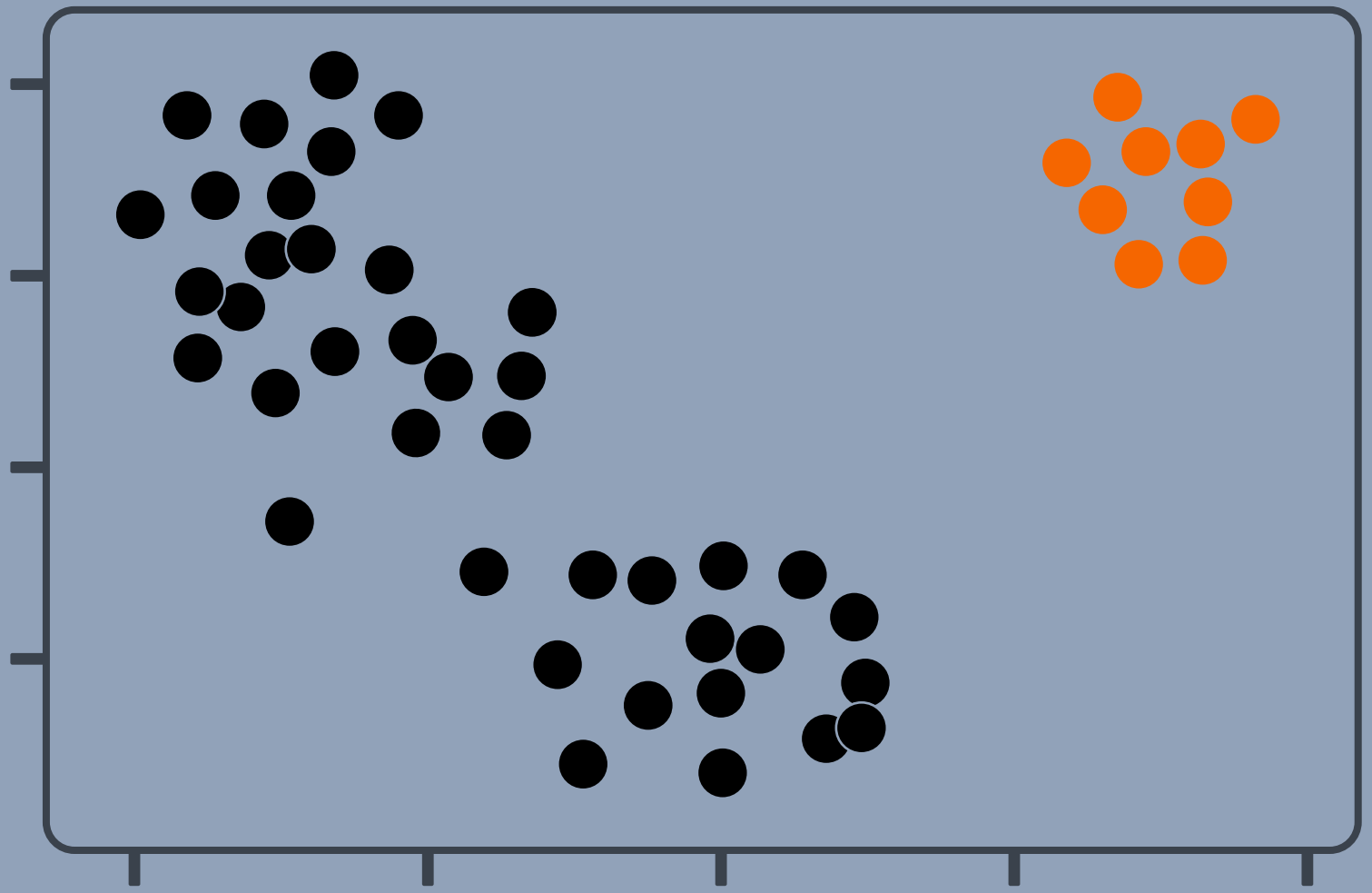
Introduction to clustering

How many clusters do you see?



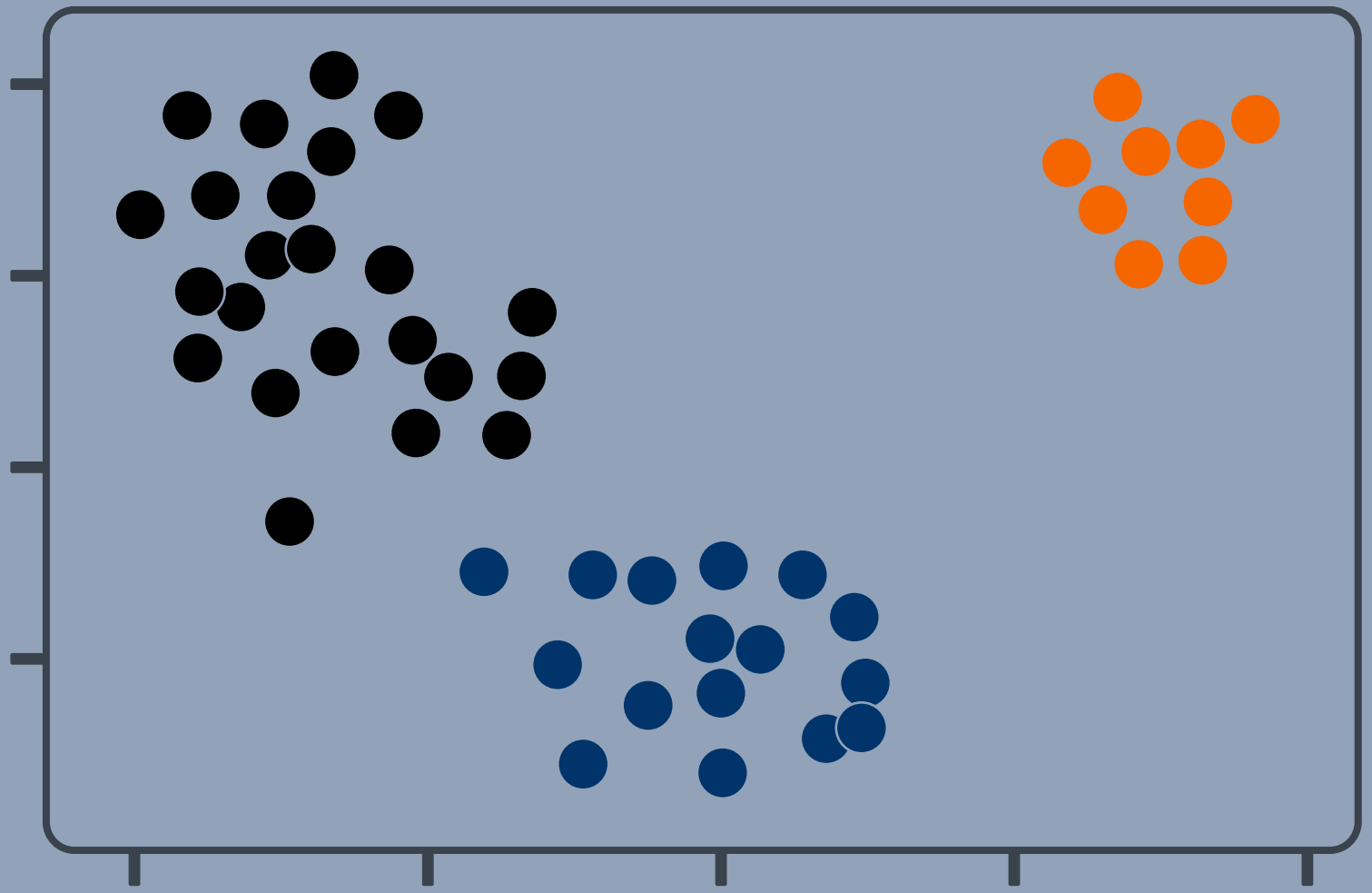
Introduction to clustering

How many clusters do you see?



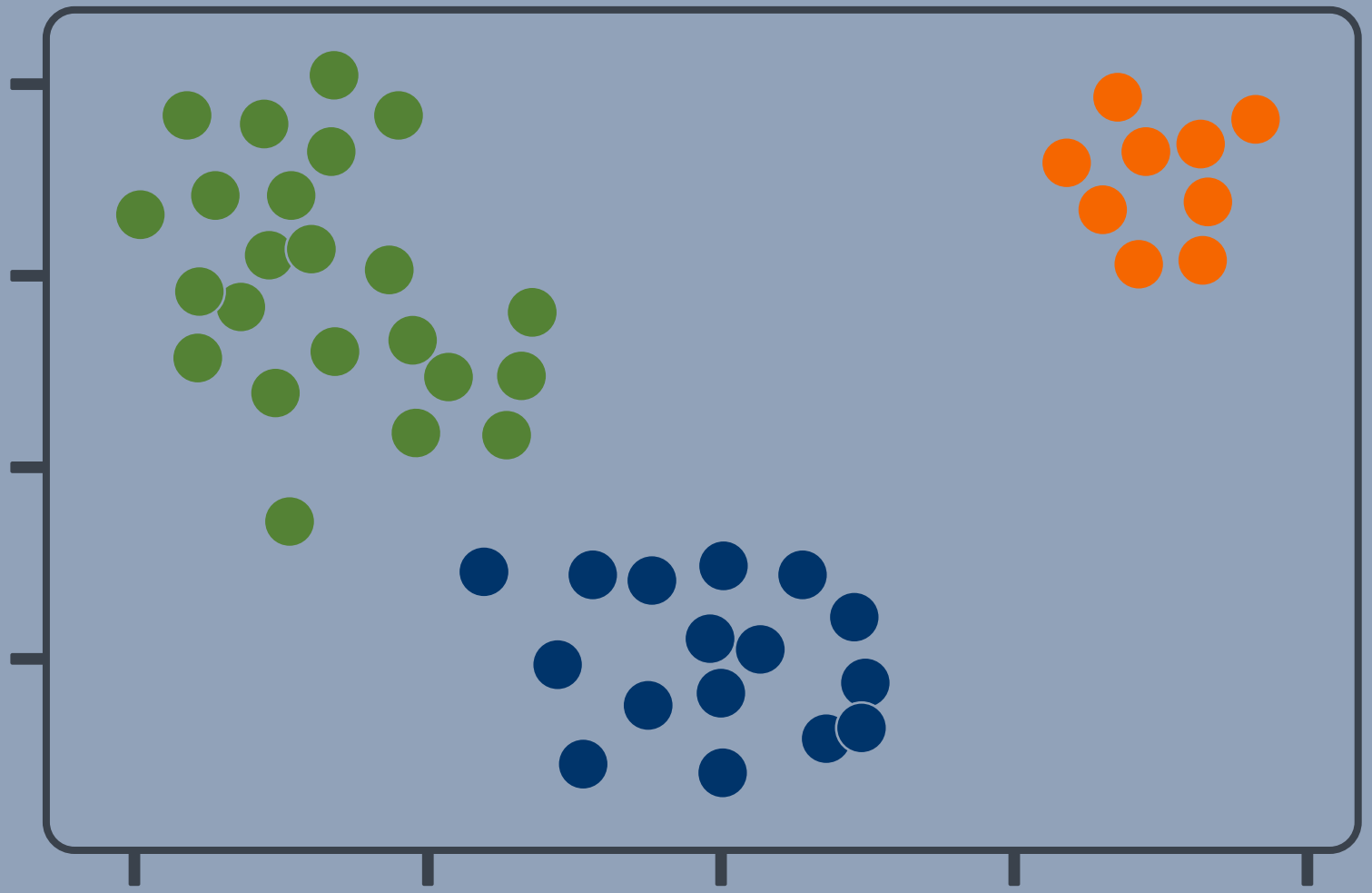
Introduction to clustering

How many clusters do you see?



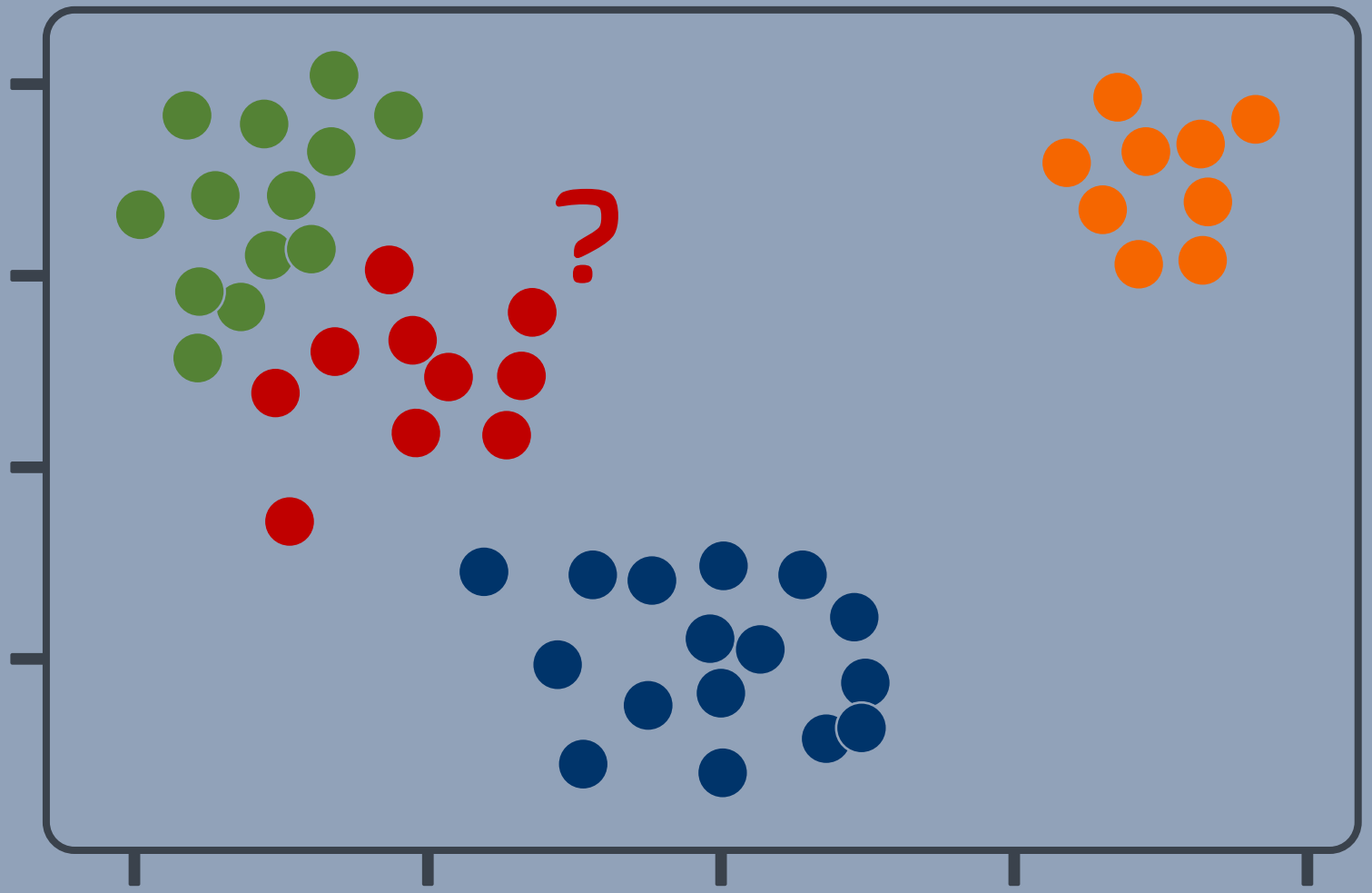
Introduction to clustering

How many clusters do you see?



Introduction to clustering

How many clusters do you see?

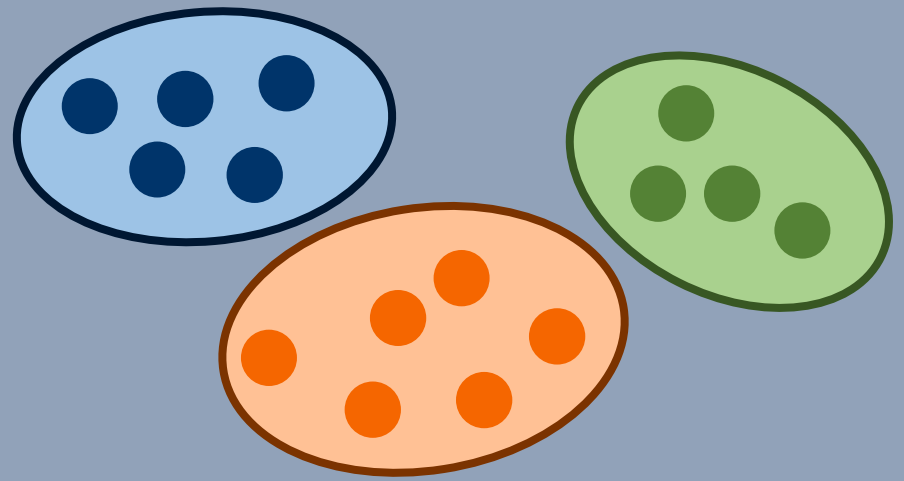


Today's topics

1. ~~Terminology~~
2. ~~Introduction to clustering~~
3. **Clustering algorithms**
4. Dimensionality reduction

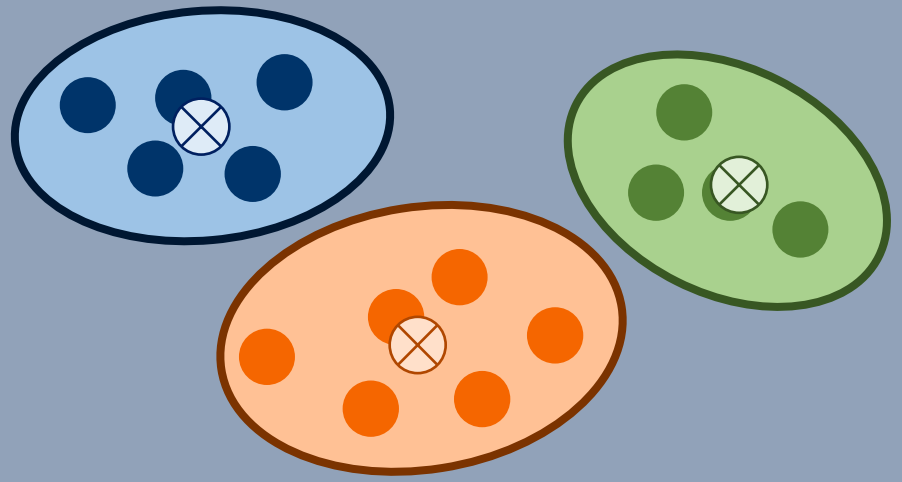
Types of clustering algorithms

1. Centroid-based clustering
2. Density-based clustering
3. Hierarchical clustering

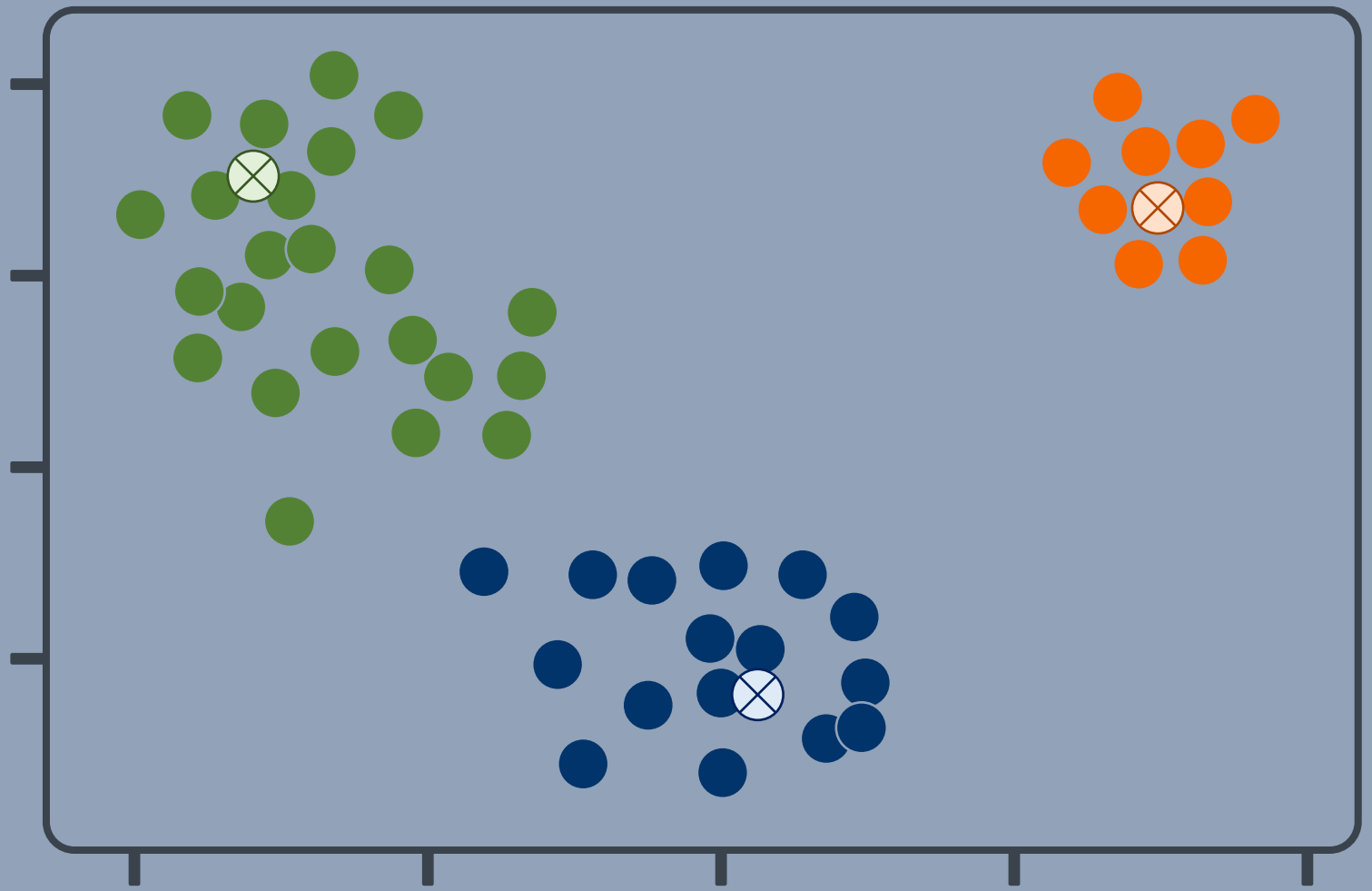


Centroid-based clustering

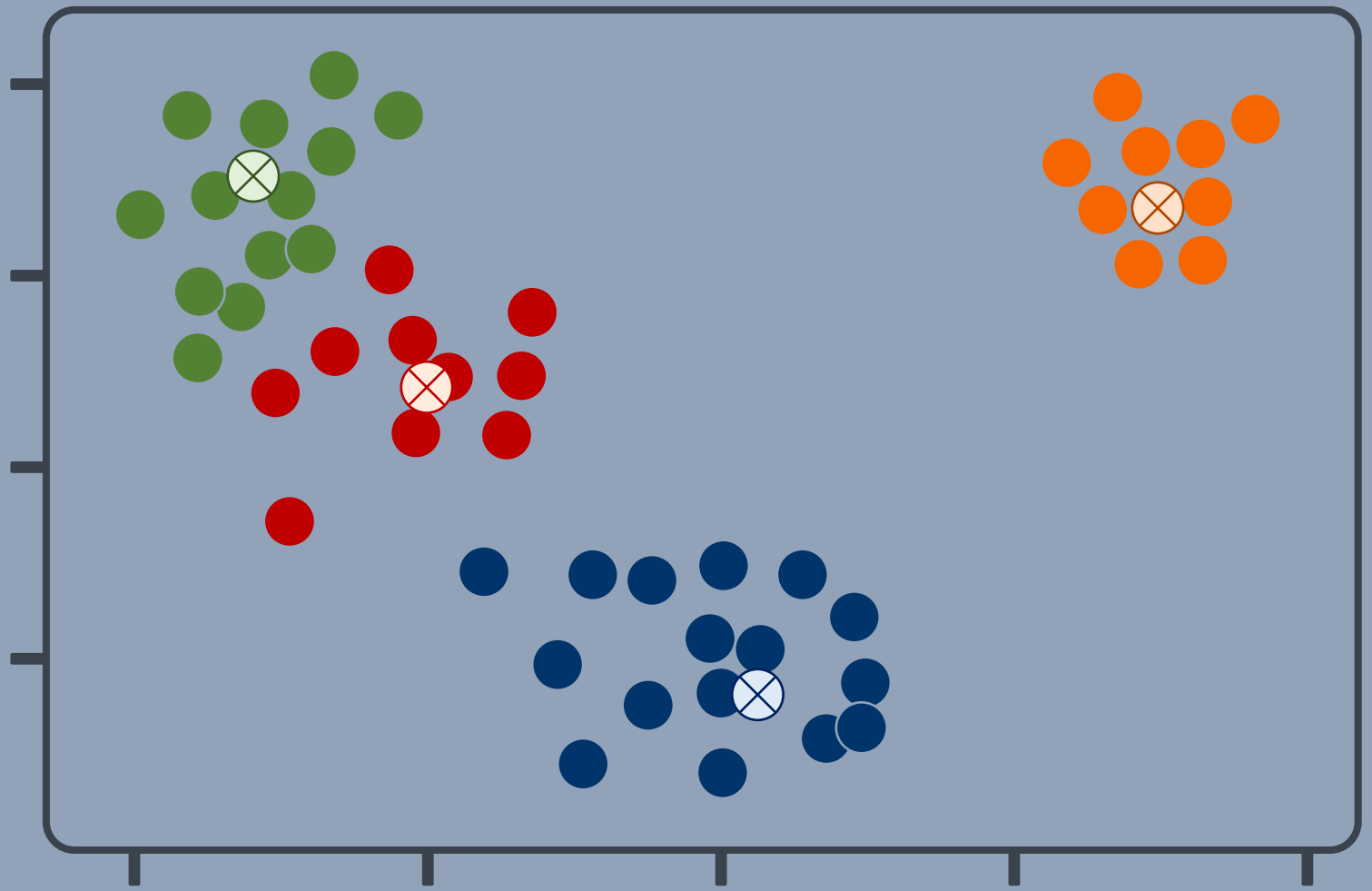
- Identifies cluster centroids
- Fixed number of clusters
- Assigns samples to nearest centroid



Centroid-based clustering

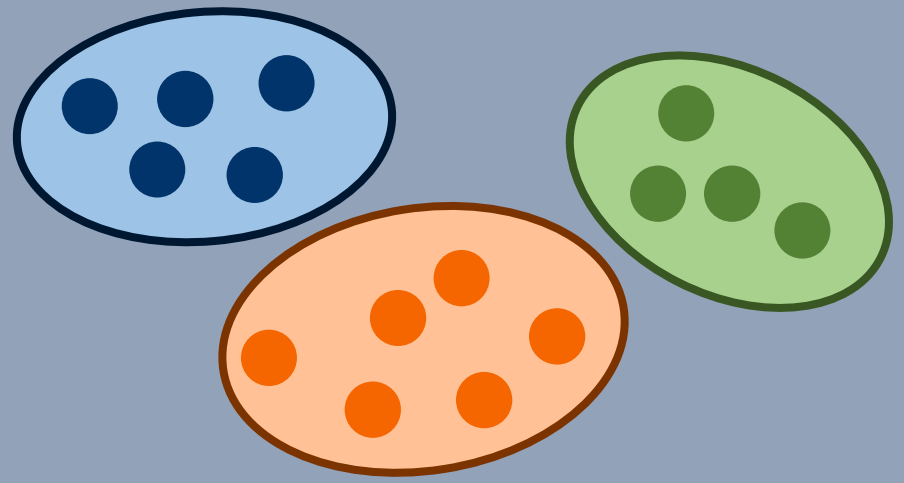


Centroid-based clustering

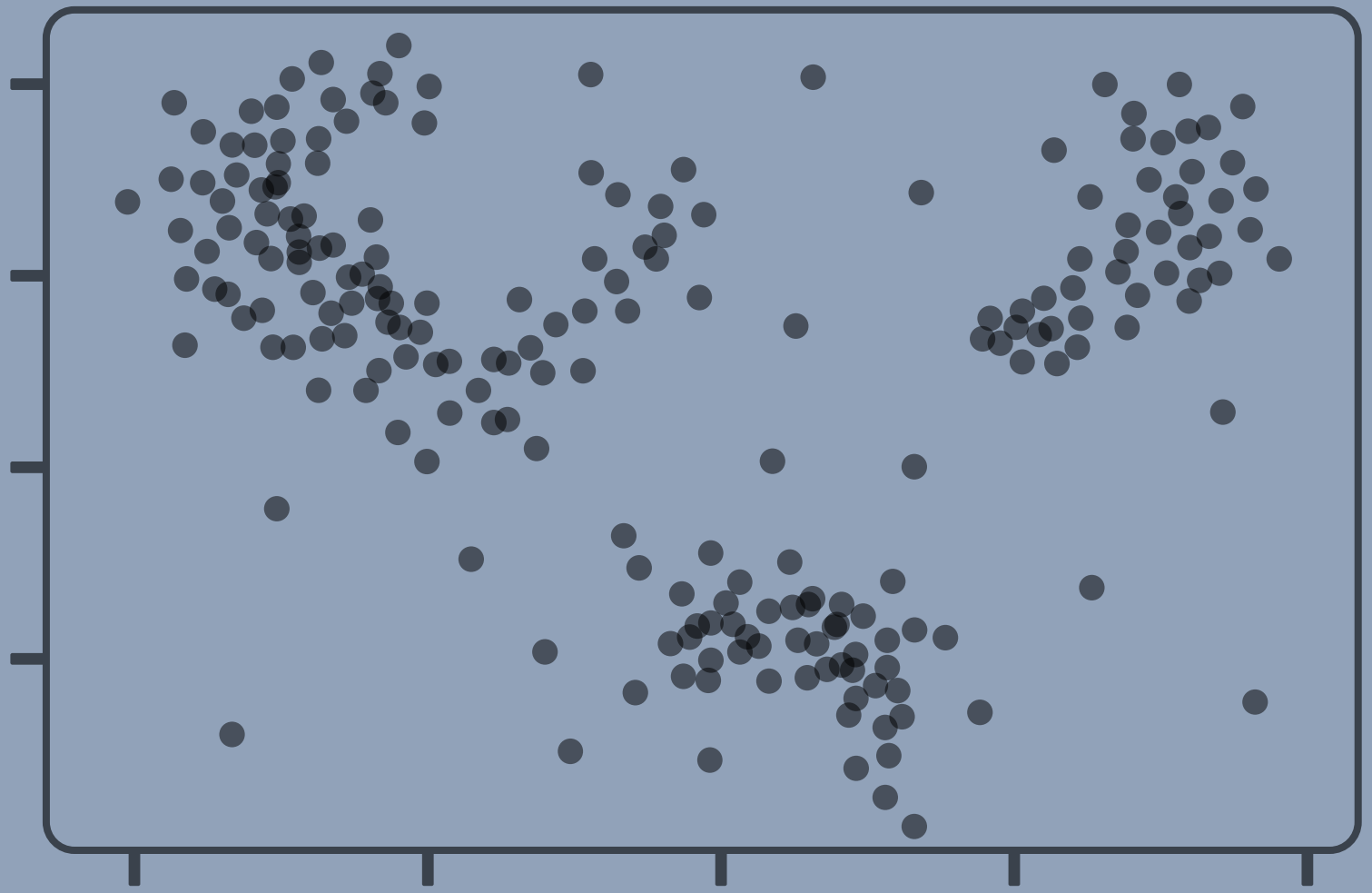


Density-based clustering

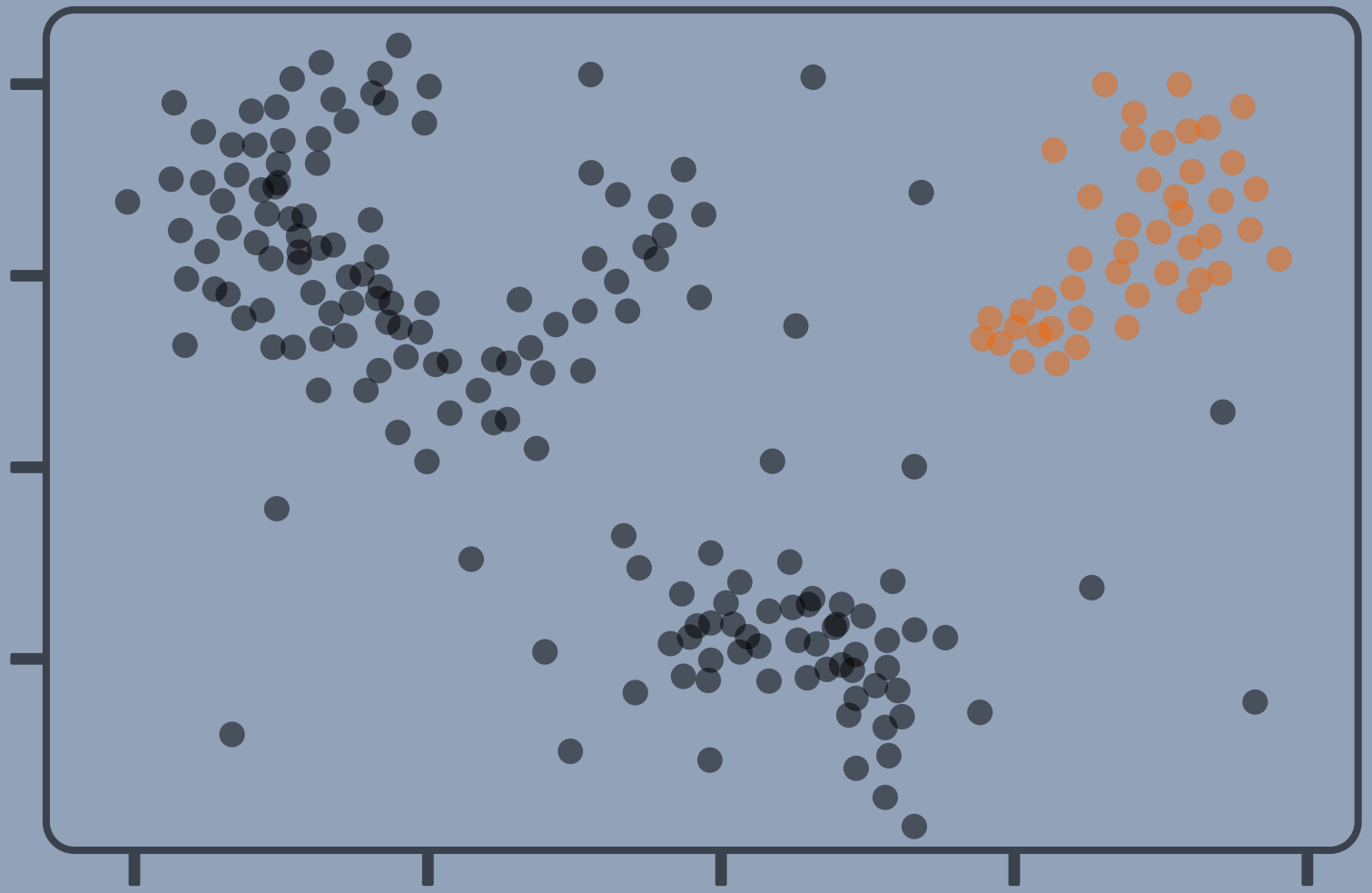
- Identifies dense areas
- Flexible number of clusters
- Excludes outliers/noise
- Cluster can take any shape



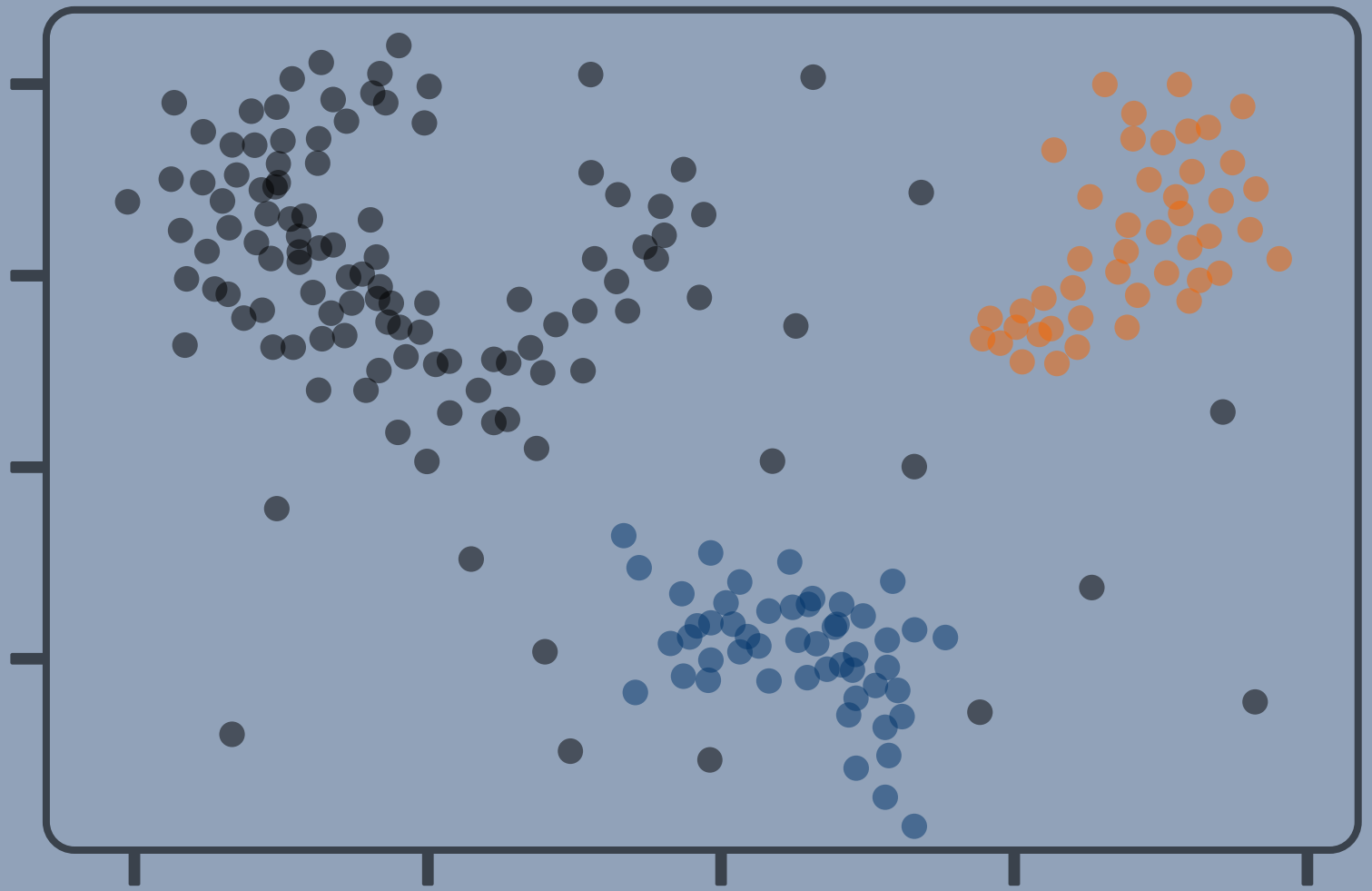
Density-based clustering



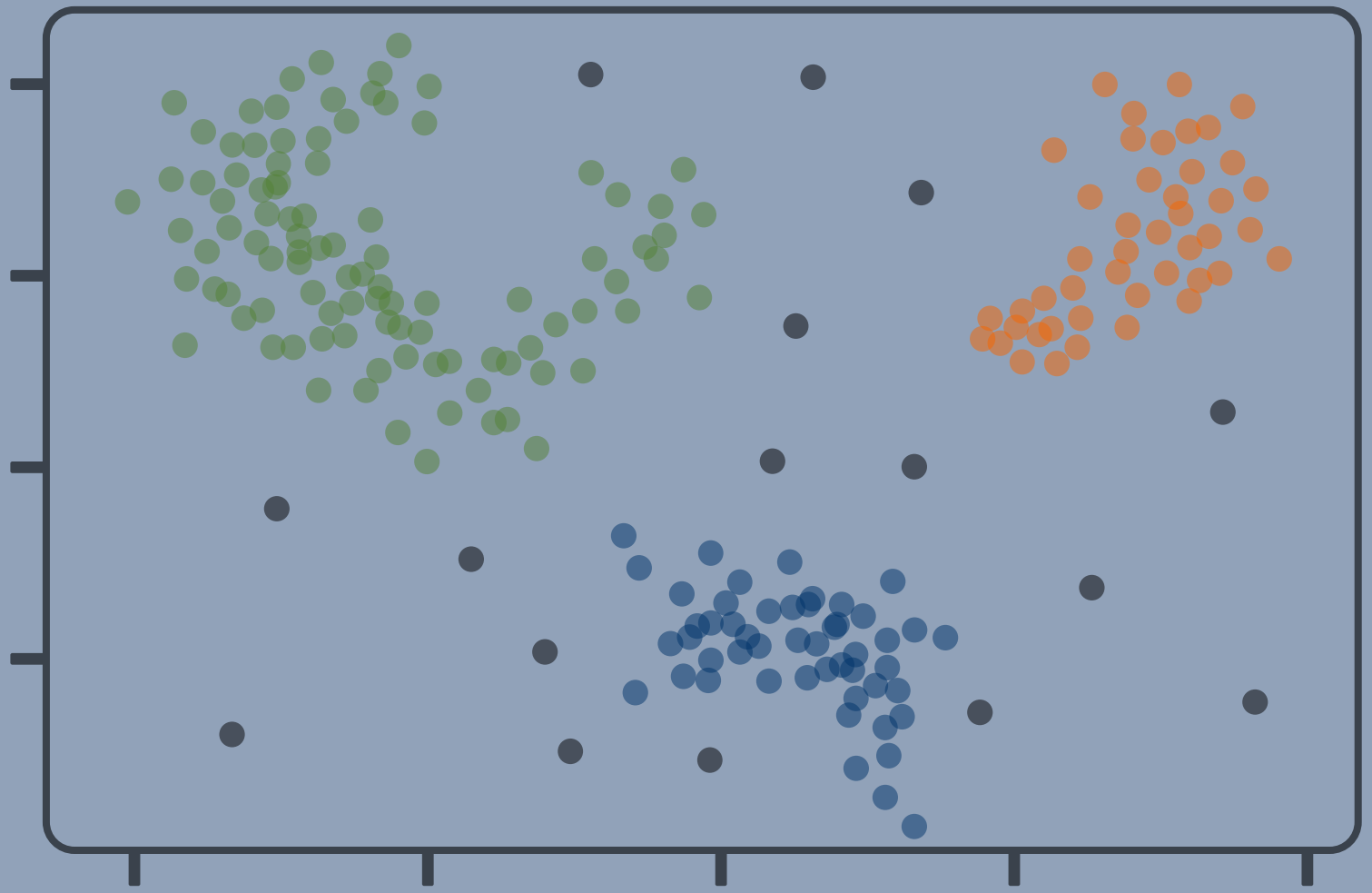
Density-based clustering



Density-based clustering

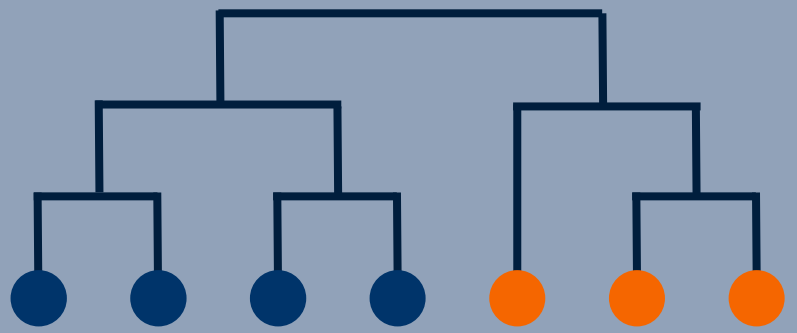


Density-based clustering

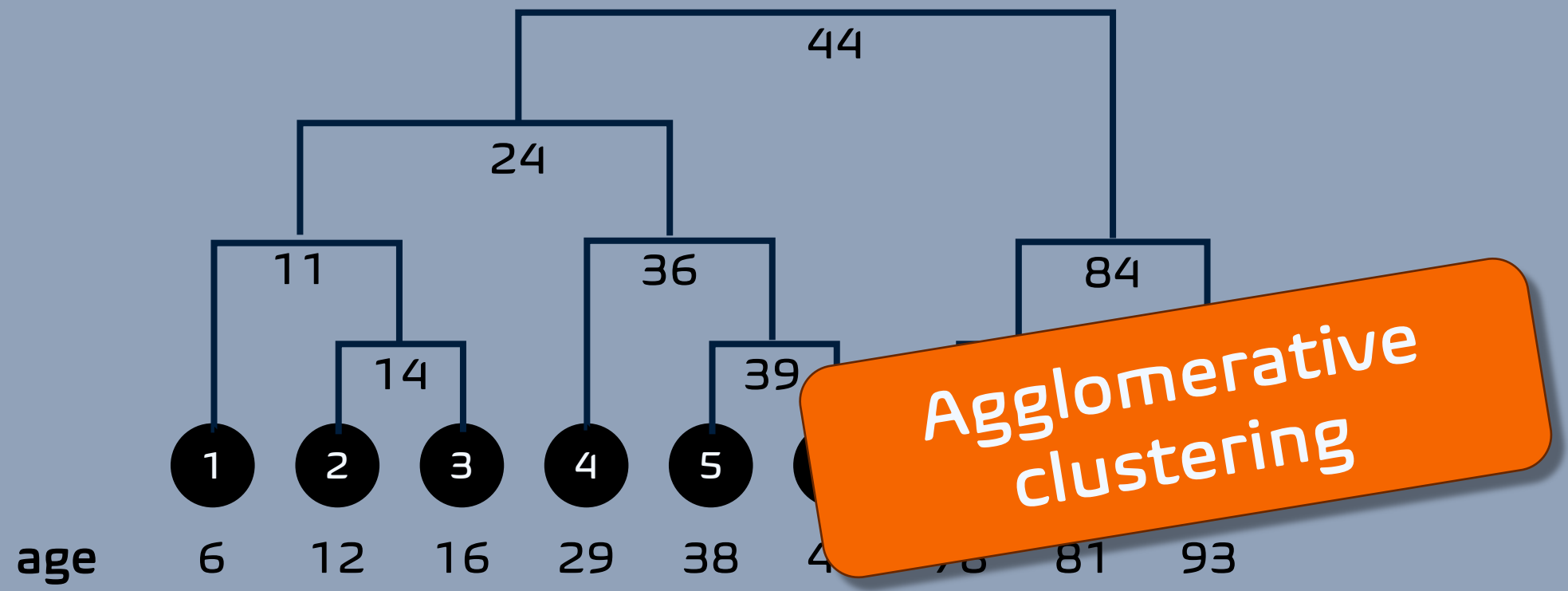


Hierarchical clustering

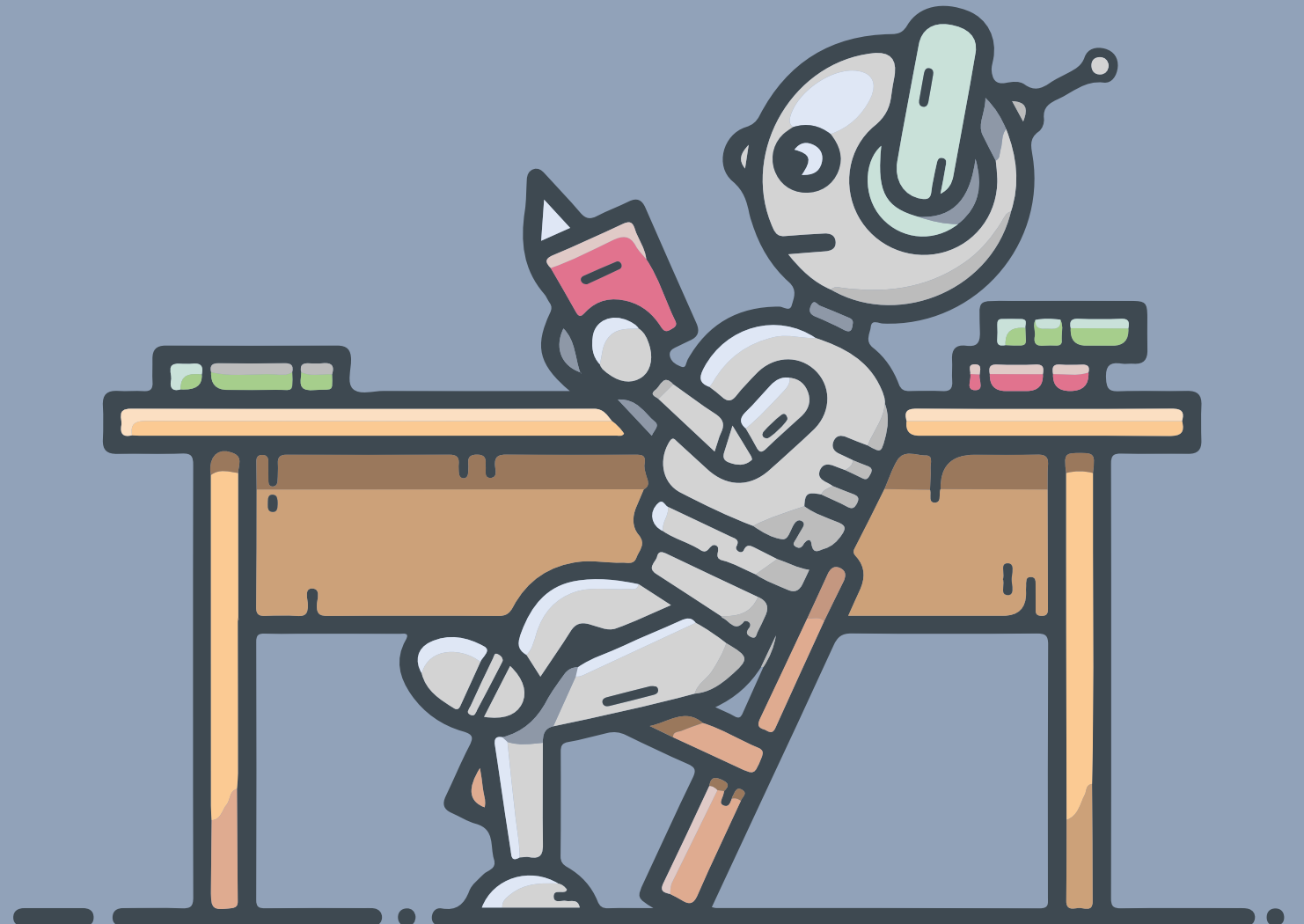
- Builds a hierarchy of similarity
- Flexible number of clusters



Hierarchical clustering



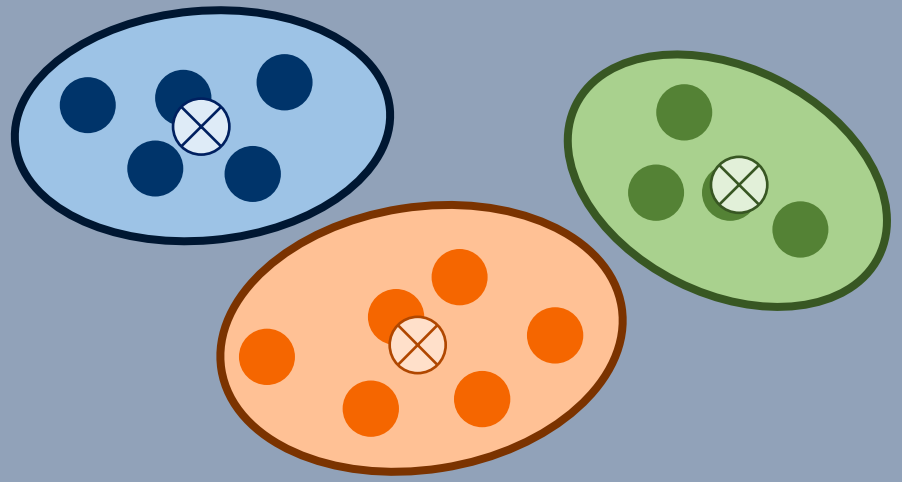
Time for a break!



Let’s look at some clustering models

K-means clustering

- The most commonly used clustering algorithm
- Centroid-based clustering model

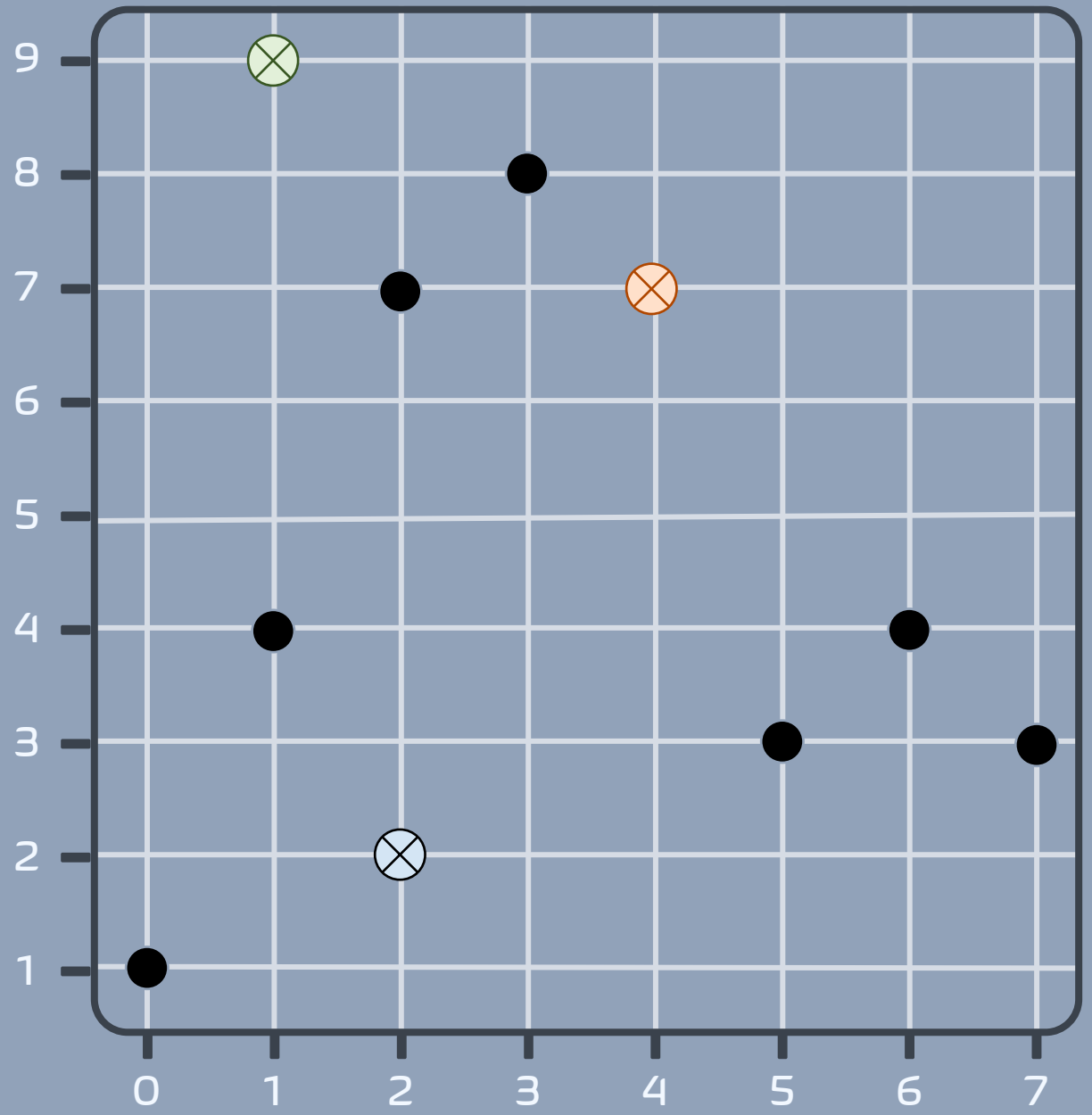


K-means clustering

- The initial centroids:
  - C1 = Pt. 3 = (1, 9)
  - C2 = Pt. 4 = (2, 2)
  - C3 = Pt. 7 = (4, 7)

	Variable 1	Variable 2
Pt. 1	0	1
Pt. 2	1	4
Pt. 3	1	9
Pt. 4	2	2
Pt. 5	2	7
Pt. 6	3	8
Pt. 7	4	7
Pt. 8	5	3
Pt. 9	6	4
Pt. 10	7	3

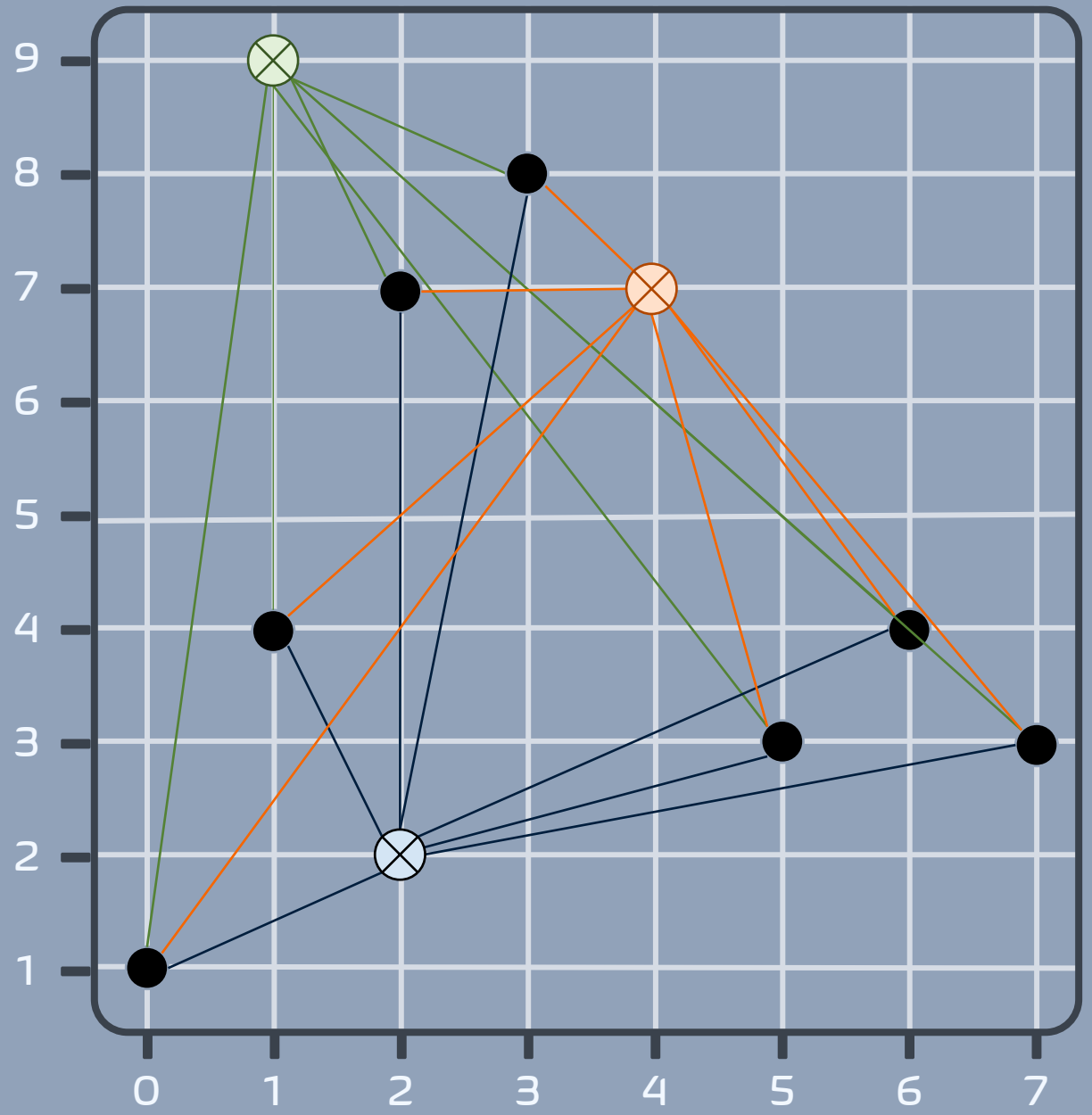
Clustering algorithms



K-means clustering

	Variable 1	Variable 2
Pt. 1	0	1
Pt. 2	1	4
Pt. 3	1	9
Pt. 4	2	2
Pt. 5	2	7
Pt. 6	3	8
Pt. 7	4	7
Pt. 8	5	3
Pt. 9	6	4
Pt. 10	7	3

Clustering algorithms



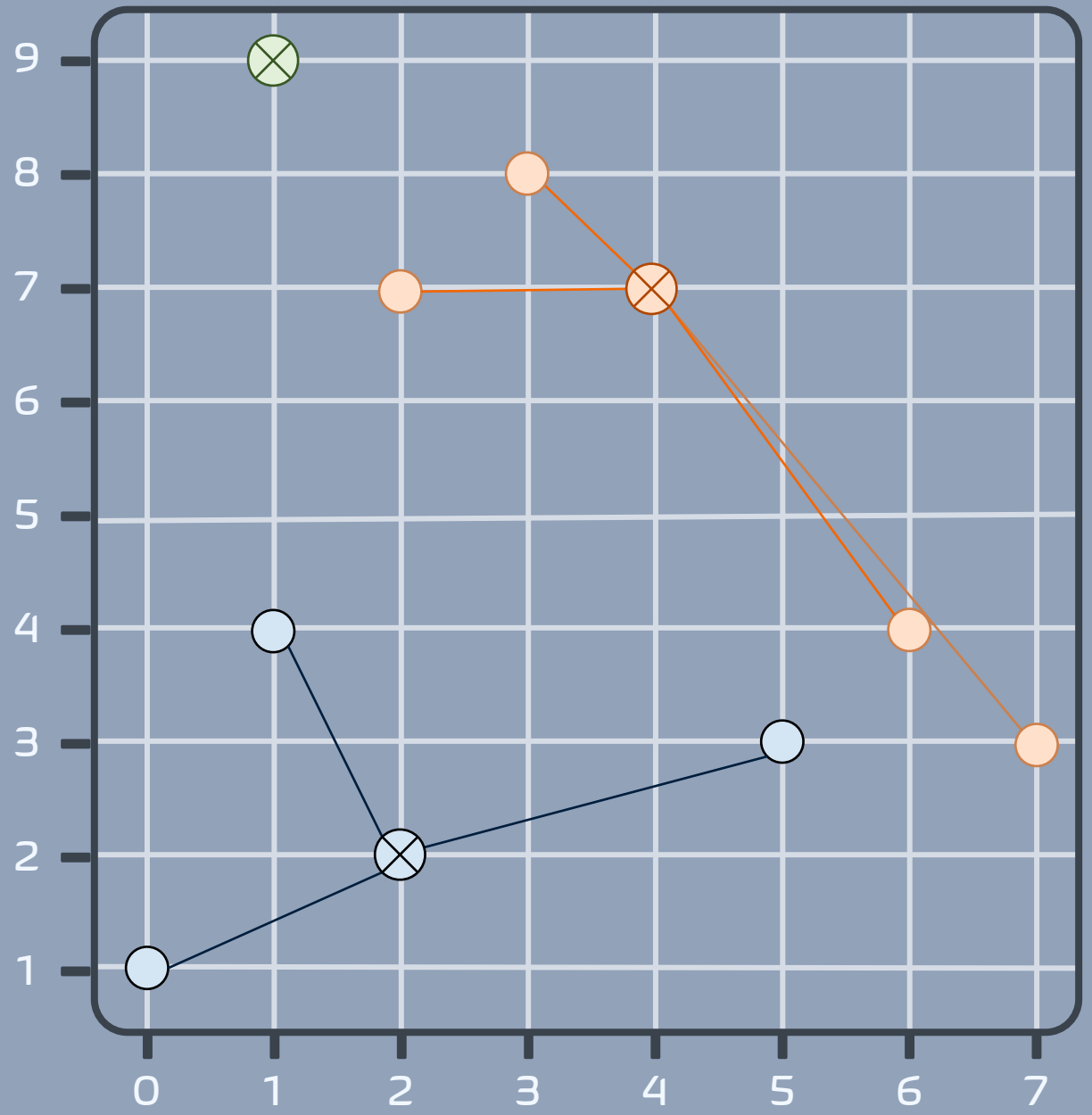
K-means clustering

	Variable 1	Variable 2
Pt. 1	0	1
Pt. 2	1	4
Pt. 3	1	9
Pt. 4	2	2
Pt. 5	2	7
Pt. 6	3	8
Pt. 7	4	7
Pt. 8	5	3
Pt. 9	6	4
Pt. 10	7	3

Clustering algorithms

i	x_1	x_2	distance to c_1 = (1,9)	distance to c_2 = (2,2)	distance to c_3 = (4,7)
1	0	1	$(0-1)^2 + (1-9)^2 = 65$	$(0-2)^2 + (1-2)^2 = 5$	$(0-4)^2 + (1-7)^2 = 52$
2	1	4	$(1-1)^2 + (4-9)^2 = 25$	$(1-2)^2 + (4-2)^2 = 5$	$(1-4)^2 + (4-7)^2 = 18$
3	1	9	$(1-1)^2 + (9-9)^2 = 0$	$(1-2)^2 + (9-2)^2 = 50$	$(1-4)^2 + (9-7)^2 = 13$
4	2	2	$(2-1)^2 + (2-9)^2 = 50$	$(2-2)^2 + (2-2)^2 = 0$	$(2-4)^2 + (2-7)^2 = 29$
5	2	7	$(2-1)^2 + (7-9)^2 = 5$	$(2-2)^2 + (7-2)^2 = 25$	$(2-4)^2 + (7-7)^2 = 4$
6	3	8	$(3-1)^2 + (8-9)^2 = 5$	$(3-2)^2 + (8-2)^2 = 37$	$(3-4)^2 + (8-7)^2 = 2$
7	4	7	$(4-1)^2 + (7-9)^2 = 13$	$(4-2)^2 + (7-2)^2 = 29$	$(4-4)^2 + (7-7)^2 = 0$
8	5	3	$(5-1)^2 + (3-9)^2 = 52$	$(5-2)^2 + (3-2)^2 = 10$	$(5-4)^2 + (3-7)^2 = 17$
9	6	4	$(6-1)^2 + (4-9)^2 = 50$	$(6-2)^2 + (4-2)^2 = 20$	$(6-4)^2 + (4-7)^2 = 13$
10	7	3	$(7-1)^2 + (3-9)^2 = 72$	$(7-2)^2 + (3-2)^2 = 26$	$(7-4)^2 + (3-7)^2 = 25$

Clustering algorithms



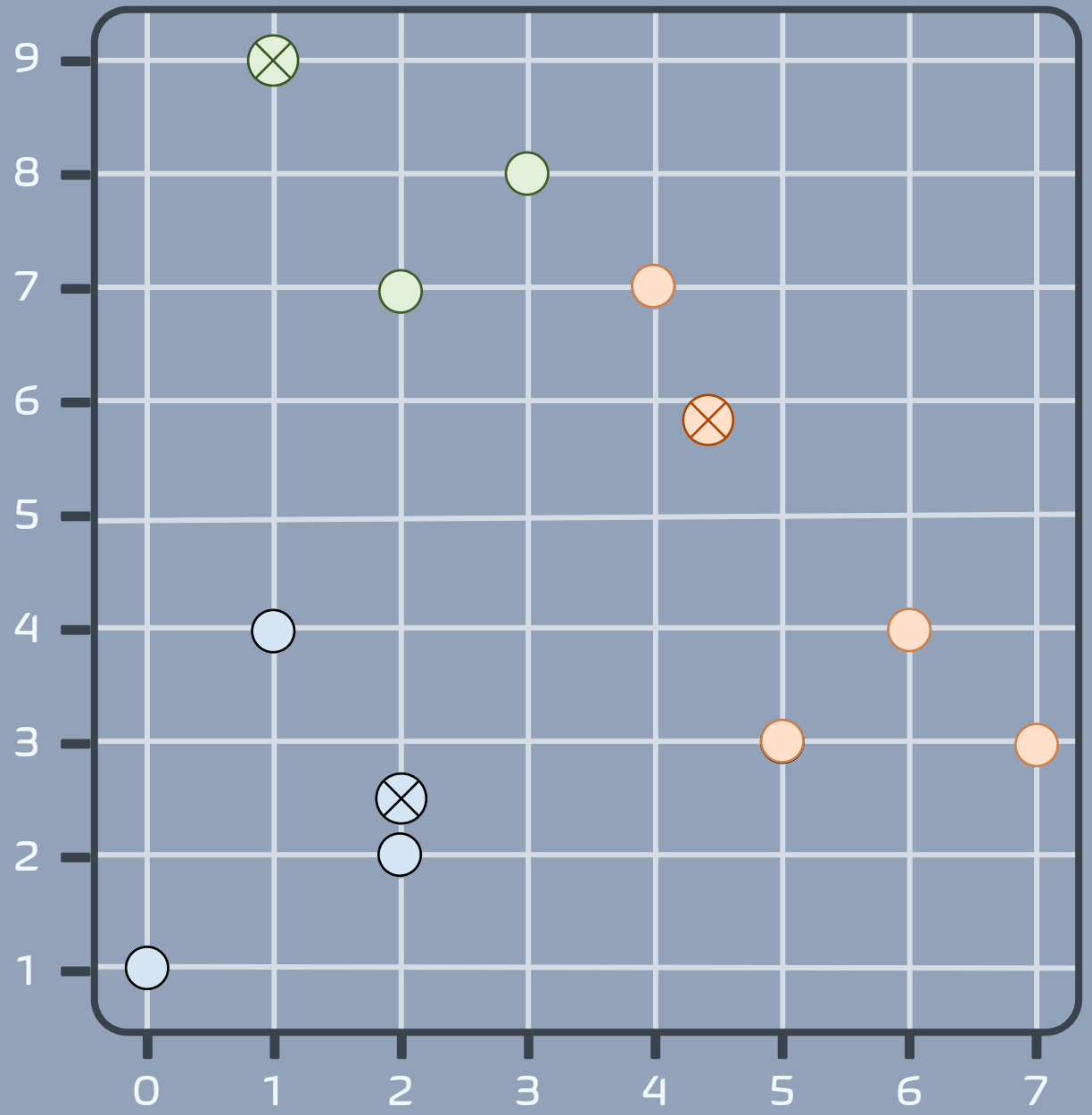
K-means clustering

	Variable 1	Variable 2
Pt. 1	0	1
Pt. 2	1	4
Pt. 3	1	9
Pt. 4	2	2
Pt. 5	2	7
Pt. 6	3	8
Pt. 7	4	7
Pt. 8	5	3
Pt. 9	6	4
Pt. 10	7	3

Clustering algorithms

i	x_1	x_2	distance to c_1 = (1,9)	distance to c_2 = (2,2.5)	distance to c_3 = (4.4,5.8)
1	0	1	$(0-1)^2 + (1-9)^2 = 65$	$(0-2)^2 + (1-2.5)^2 = 6.25$	$(0-4)^2 + (1-7)^2 = 42.4$
2	1	4	$(1-1)^2 + (4-9)^2 = 25$	$(1-2)^2 + (4-2.5)^2 = 3.25$	$(1-4)^2 + (4-7)^2 = 14.8$
3	1	9	$(1-1)^2 + (9-9)^2 = 0$	$(1-2)^2 + (9-2.5)^2 = 43.25$	$(1-4)^2 + (9-7)^2 = 21.8$
4	2	2	$(2-1)^2 + (2-9)^2 = 50$	$(2-2)^2 + (2-2.5)^2 = 0.25$	$(2-4)^2 + (2-7)^2 = 20.2$
5	2	7	$(2-1)^2 + (7-9)^2 = 5$	$(2-2)^2 + (7-2.5)^2 = 20.25$	$(2-4)^2 + (7-7)^2 = 7.2$
6	3	8	$(3-1)^2 + (8-9)^2 = 5$	$(3-2)^2 + (8-2.5)^2 = 31.25$	$(3-4)^2 + (8-7)^2 = 6.8$
7	4	7	$(4-1)^2 + (7-9)^2 = 13$	$(4-2)^2 + (7-2.5)^2 = 24.25$	$(4-4)^2 + (7-7)^2 = 1.6$
8	5	3	$(5-1)^2 + (3-9)^2 = 52$	$(5-2)^2 + (3-2.5)^2 = 9.25$	$(5-4)^2 + (3-7)^2 = 8.2$
9	6	4	$(6-1)^2 + (4-9)^2 = 50$	$(6-2)^2 + (4-2.5)^2 = 18.25$	$(6-4)^2 + (4-7)^2 = 5.8$
10	7	3	$(7-1)^2 + (3-9)^2 = 72$	$(7-2)^2 + (3-2.5)^2 = 25.25$	$(7-4)^2 + (3-7)^2 = 14.6$

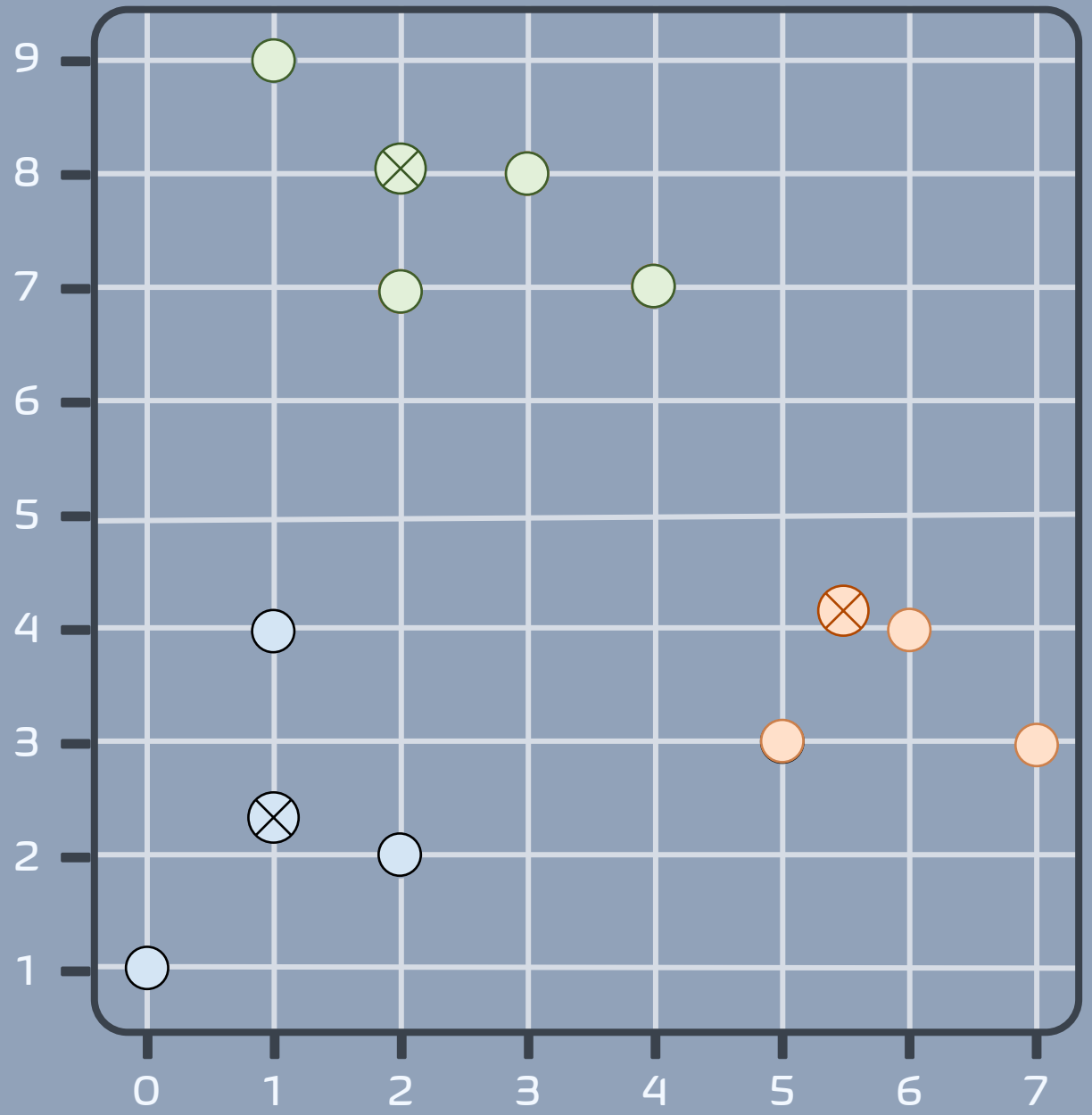
Clustering algorithms



K-means clustering

	Variable 1	Variable 2
Pt. 1	0	1
Pt. 2	1	4
Pt. 3	1	9
Pt. 4	2	2
Pt. 5	2	7
Pt. 6	3	8
Pt. 7	4	7
Pt. 8	5	3
Pt. 9	6	4
Pt. 10	7	3

Clustering algorithms



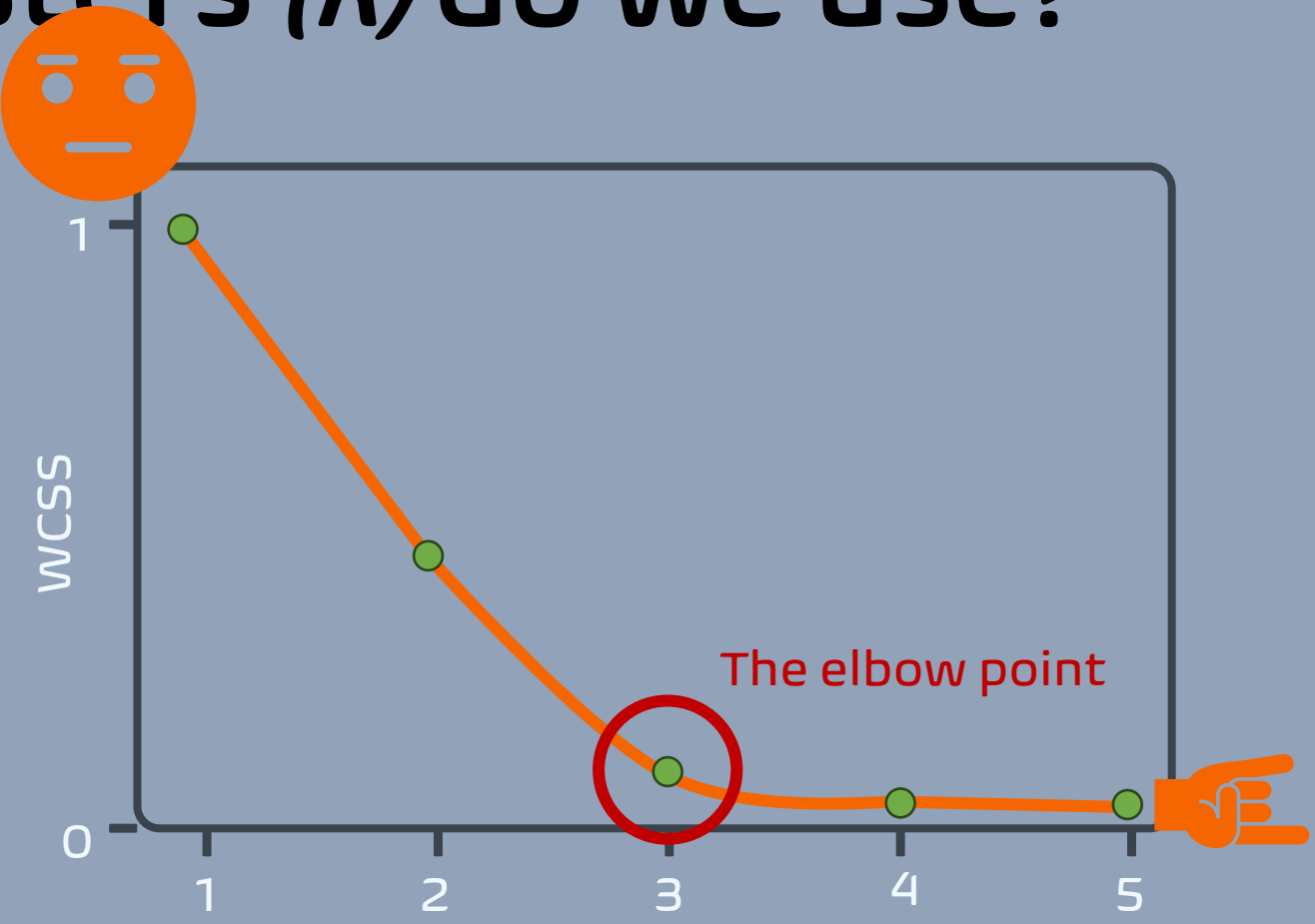
K-means clustering

	Variable 1	Variable 2
Pt. 1	0	1
Pt. 2	1	4
Pt. 3	1	9
Pt. 4	2	2
Pt. 5	2	7
Pt. 6	3	8
Pt. 7	4	7
Pt. 8	5	3
Pt. 9	6	4
Pt. 10	7	3

So how many clusters *(k)* do we use?

The elbow method:

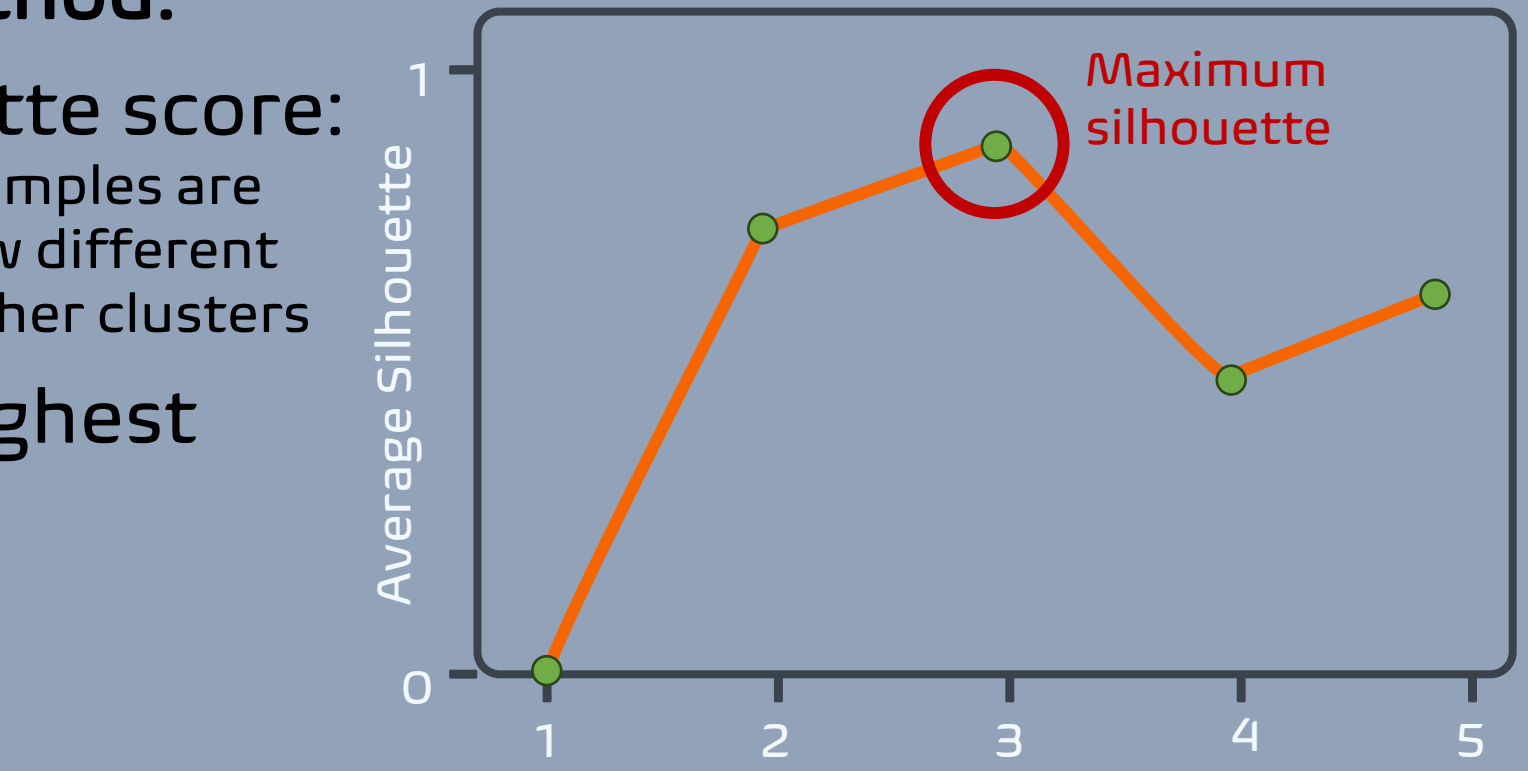
- Compute Sum of Squared Distances within clusters (WCSS) for different values of *k*
- Pick the 'elbow point'



So how many clusters *(k)* do we use?

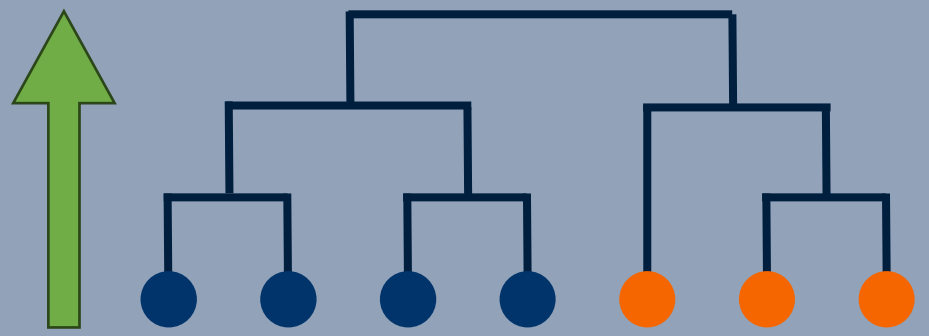
The silhouette method:

- Compute Silhouette score:
  describes how similar samples are within a cluster, and how different they to samples from other clusters
- Pick *k* with the highest silhouete



Agglomerative clustering

- Hierarchical clustering method
- Bottom-up approach: each patient starts in its own cluster

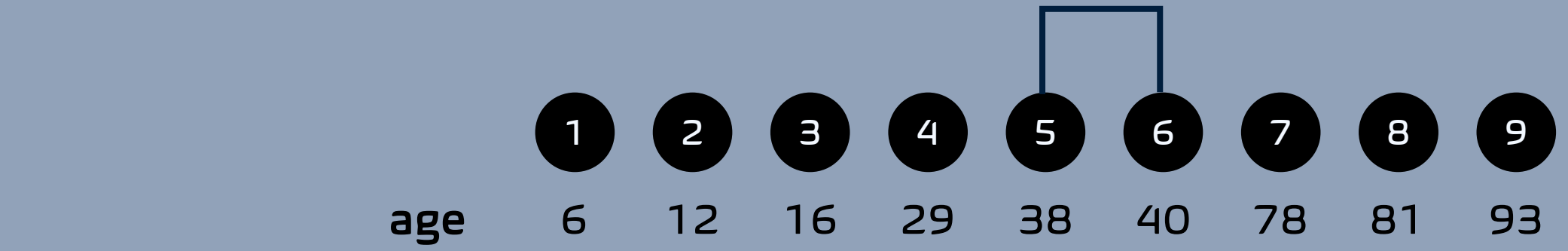


Clustering algorithms

Agglomerative clustering

1. Compute distance matrix (the distance between all points)
2. Combine closest patients

	1	2	3	4	5	6	7	8	9
1	0	6	10	...	...	...	...	...	...
2	6	0	4	...	...	...	...	...	...
3	10	4	0	...	...	...	...	...	...
4	...	...	...	0	...	...	...	...	...
5	...	...	...	...	0	2	...	...	...
6	...	...	...	...	2	0	...	...	...
7	...	...	...	...	...	...	0	...	...
8	...	...	...	...	...	...	...	0	...
9	...	...	...	...	...	...	...	...	0

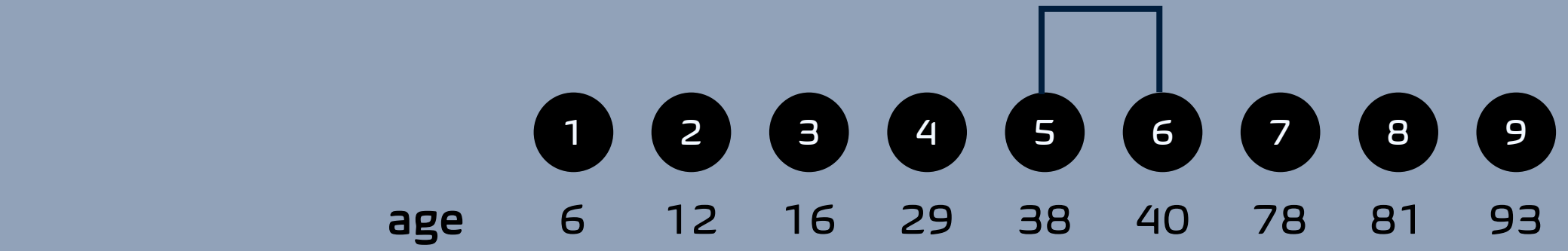


Clustering algorithms

Agglomerative clustering

1. Compute distance matrix (the distance between all points)
2. Combine closest patients
3. Repeat step 1 & 2

	1	2	3	4	5/6	7	8	9
1	0	6	10	…	?	…	…	…
2	6	0	4	…	?	…	…	…
3	10	4	0	…	?	…	…	…
4	…	…	…	0	?	…	…	…
5/6	?	?	?	?	0	?	?	…?
7	…	…	…	…	?	0	…	…
8	…	…	…	…	?	…	0	…
9	…	…	…	…	?	…	…	0

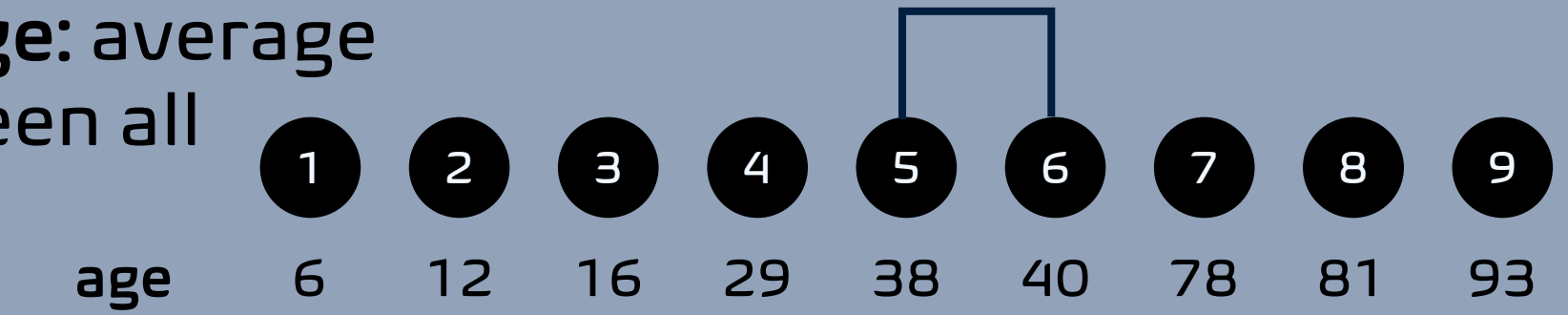


Clustering algorithms

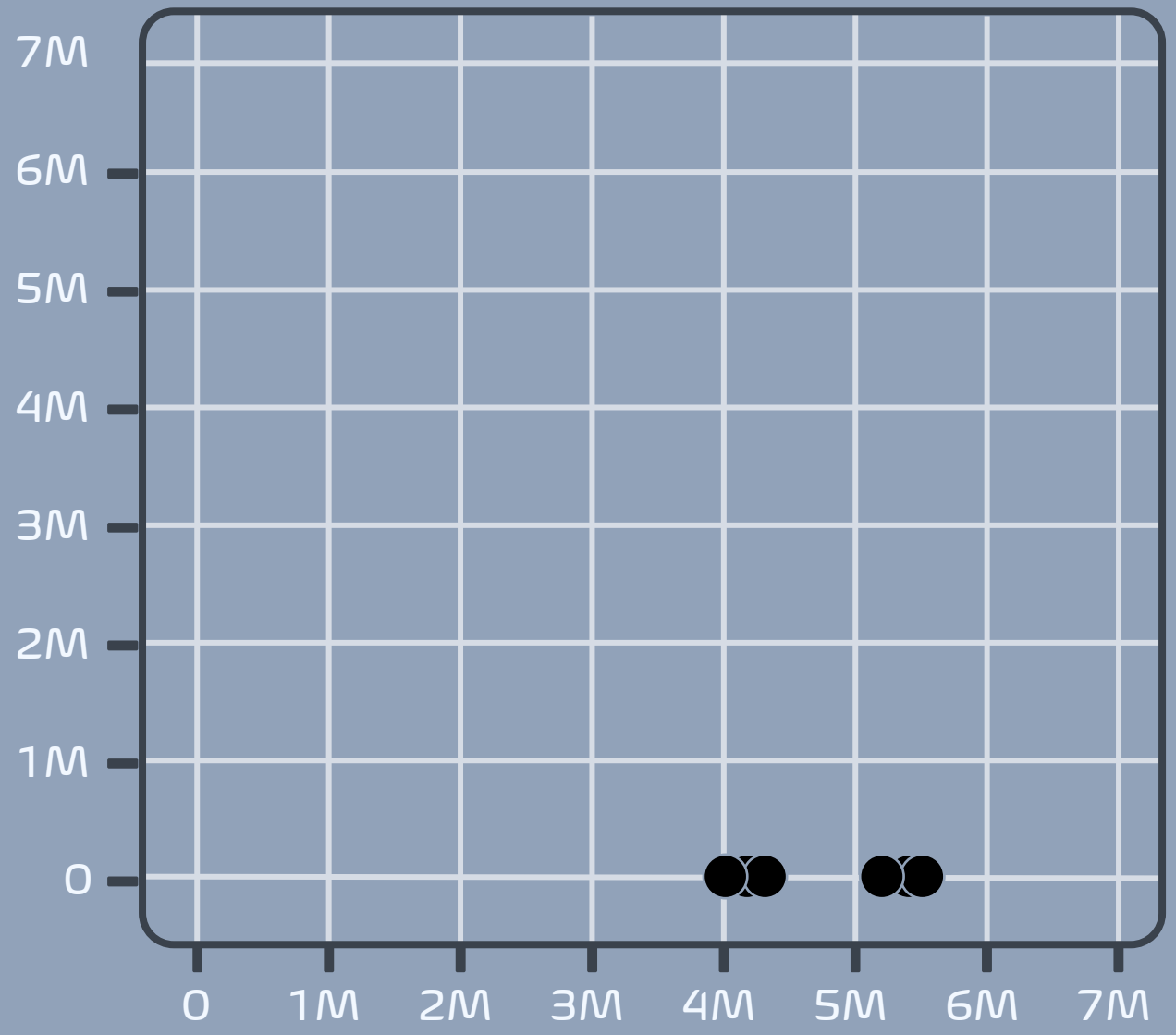
Distance between clusters

1. **Single-linkage:** distance between closest patients
2. **Complete-linkage:** distance between farthest patients
3. **Average-linkage:** average distance between all patients

	1	2	3	4	5/6	7	8	9
1	0	6	10	...	?	...	...	...
2	6	0	4	...	?	...	...	...
3	10	4	0	...	?	...	...	...
4	...	...	...	0	?	...	...	...
5/6	?	?	?	?	0	?	?	...?
7	...	...	...	...	?	0	...	...
8	...	...	...	...	?	...	0	...
9	...	...	...	...	?	...	...	0



Clustering algorithms



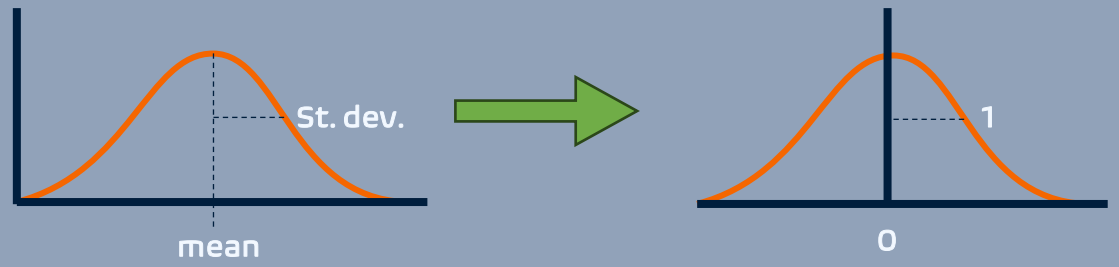
What's wrong?

	BMI	Red blood cells (cells/mcL)
Pt. 1	15	4350000
Pt. 2	33	4120000
Pt. 3	22	5480000
Pt. 4	26	4020000
Pt. 5		5110000
		320000

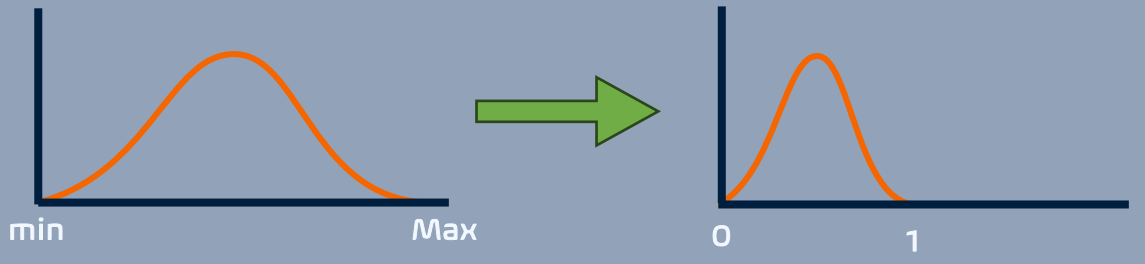
1 BMI = 1 cell/ml ??

Data scaling

1. **Standardisation** : subtract the mean and divide by standard deviation



2. **Min-max scaling**: Set minimum to zero, and maximum to one.



Terminology confusion

1. Standardisation = Z-standardisation = Z-score normalisation

2. Min-max scaling = normalisation = min-max normalisation = rescaling

Terminology confusion

1. **Data scaling:** the process of transforming the values of a variable such that they are within a specific range

2. **Normalisation:** the process of transforming the distribution of a variable to a normal distribution

3. **Normalisation:** the process of scaling individual samples to have unit norm.

Terminology confusion

1. **Data scaling:** the process of transforming the values of a variable such that they are within a specific range

2. **Normalisation:** the process of transforming the distribution of a variable to a normal distribution

3. ~~**Normalisation:** the process of scaling individual samples to have unit norm.~~

Take-home message

Clustering is contextual

- Numbers alone cannot tell the full story
- Always describe your clusters. What do they mean?

We’re almost there!

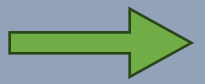
Today's topics

1. ~~Terminology~~
2. ~~Introduction to clustering~~
3. ~~Clustering algorithms~~
4. Dimensionality reduction

Dimensionality reduction

- **High-dimensional data:** many variables
- **What is the issue with high dimensional data?**
  - Difficult to understand
  - High computational load
  - Multicollinearity
  - noisy

	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f	g	h
Pt. 1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
Pt. 2	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	1	4
Pt. 3	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	1	9
Pt. 4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Pt. 5	2	7	2	7	2	7	2	7	2	7	2	7	2	7	2	2	7
Pt. 6	3	8	3	8	3	8	3	8	3	8	3	8	3	8	3	3	8
Pt. 7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	4	7
Pt. 8	5	3	5	3	5	3	5	3	5	3	5	3	5	3	5	5	3
Pt. 9	6	4	6	4	6	4	6	4	6	4	6	4	6	4	6	6	4
Pt. 10	7	3	7	3	7	3	7	3	7	3	7	3	7	3	7	7	3



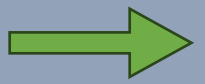
	1	2	3	4	5
Pt. 1	0	1	0	1	0
Pt. 2	1	4	1	4	1
Pt. 3	1	9	1	9	1
Pt. 4	2	2	2	2	2
Pt. 5	2	7	2	7	2
Pt. 6	3	8	3	8	3
Pt. 7	4	7	4	7	4
Pt. 8	5	3	5	3	5
Pt. 9	6	4	6	4	6
Pt. 10	7	3	7	3	7

Dimensionality reduction

- Two types:
  1. Feature selection
  2. Feature extraction

- Remember: feature = variable (mostly)

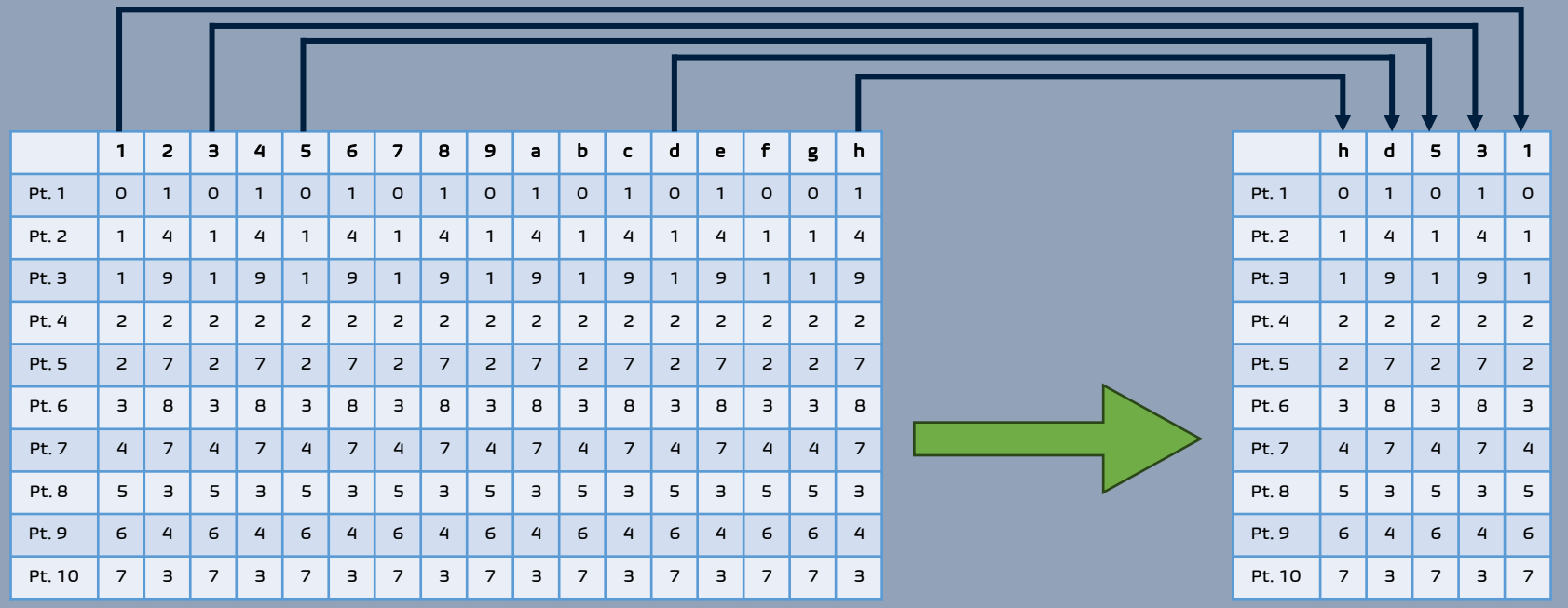
	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f	g	h
Pt. 1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
Pt. 2	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	1	4
Pt. 3	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	1	9
Pt. 4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Pt. 5	2	7	2	7	2	7	2	7	2	7	2	7	2	7	2	2	7
Pt. 6	3	8	3	8	3	8	3	8	3	8	3	8	3	8	3	3	8
Pt. 7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	4	7
Pt. 8	5	3	5	3	5	3	5	3	5	3	5	3	5	3	5	5	3
Pt. 9	6	4	6	4	6	4	6	4	6	4	6	4	6	4	6	6	4
Pt. 10	7	3	7	3	7	3	7	3	7	3	7	3	7	3	7	7	3



	1	2	3	4	5
Pt. 1	0	1	0	1	0
Pt. 2	1	4	1	4	1
Pt. 3	1	9	1	9	1
Pt. 4	2	2	2	2	2
Pt. 5	2	7	2	7	2
Pt. 6	3	8	3	8	3
Pt. 7	4	7	4	7	4
Pt. 8	5	3	5	3	5
Pt. 9	6	4	6	4	6
Pt. 10	7	3	7	3	7

Feature selection

- Filter out specific input variables
  - Automatic
  - Manual



	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f	g	h
Pt. 1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
Pt. 2	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	1	4
Pt. 3	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	1	9
Pt. 4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Pt. 5	2	7	2	7	2	7	2	7	2	7	2	7	2	7	2	2	7
Pt. 6	3	8	3	8	3	8	3	8	3	8	3	8	3	8	3	3	8
Pt. 7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	4	7
Pt. 8	5	3	5	3	5	3	5	3	5	3	5	3	5	3	5	5	3
Pt. 9	6	4	6	4	6	4	6	4	6	4	6	4	6	4	6	6	4
Pt. 10	7	3	7	3	7	3	7	3	7	3	7	3	7	3	7	7	3

	h	d	5	3	1
Pt. 1	0	1	0	1	0
Pt. 2	1	4	1	4	1
Pt. 3	1	9	1	9	1
Pt. 4	2	2	2	2	2
Pt. 5	2	7	2	7	2
Pt. 6	3	8	3	8	3
Pt. 7	4	7	4	7	4
Pt. 8	5	3	5	3	5
Pt. 9	6	4	6	4	6
Pt. 10	7	3	7	3	7

Feature extraction

- Transforming data to a lower dimensional feature space
- PCA is a popular example

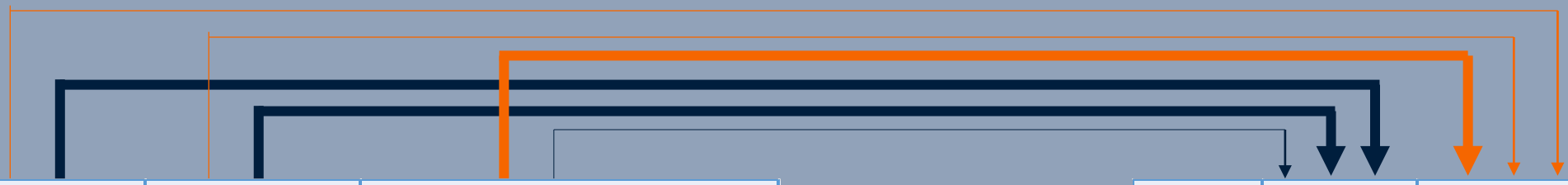
	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f	g	h
Pt. 1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
Pt. 2	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	1	4
Pt. 3	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	1	9
Pt. 4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Pt. 5	2	7	2	7	2	7	2	7	2	7	2	7	2	7	2	2	7
Pt. 6	3	8	3	8	3	8	3	8	3	8	3	8	3	8	3	3	8
Pt. 7	4	7	4	7	4	7	4	7	4	7	4	7	4	7	4	4	7
Pt. 8	5	3	5	3	5	3	5	3	5	3	5	3	5	3	5	5	3
Pt. 9	6	4	6	4	6	4	6	4	6	4	6	4	6	4	6	6	4
Pt. 10	7	3	7	3	7	3	7	3	7	3	7	3	7	3	7	7	3

	1	2	3	4	5
Pt. 1	0	1	0	1	0
Pt. 2	1	4	1	4	1
Pt. 3	1	9	1	9	1
Pt. 4	2	2	2	2	2
Pt. 5	2	7	2	7	2
Pt. 6	3	8	3	8	3
Pt. 7	4	7	4	7	4
Pt. 8	5	3	5	3	5
Pt. 9	6	4	6	4	6
Pt. 10	7	3	7	3	7

Principal Component Analysis (PCA)

- PCA summarises the input variables in principal components
- Principal components are linear combinations of the input variables
- Principal components are order on how much data (variance) they summarise

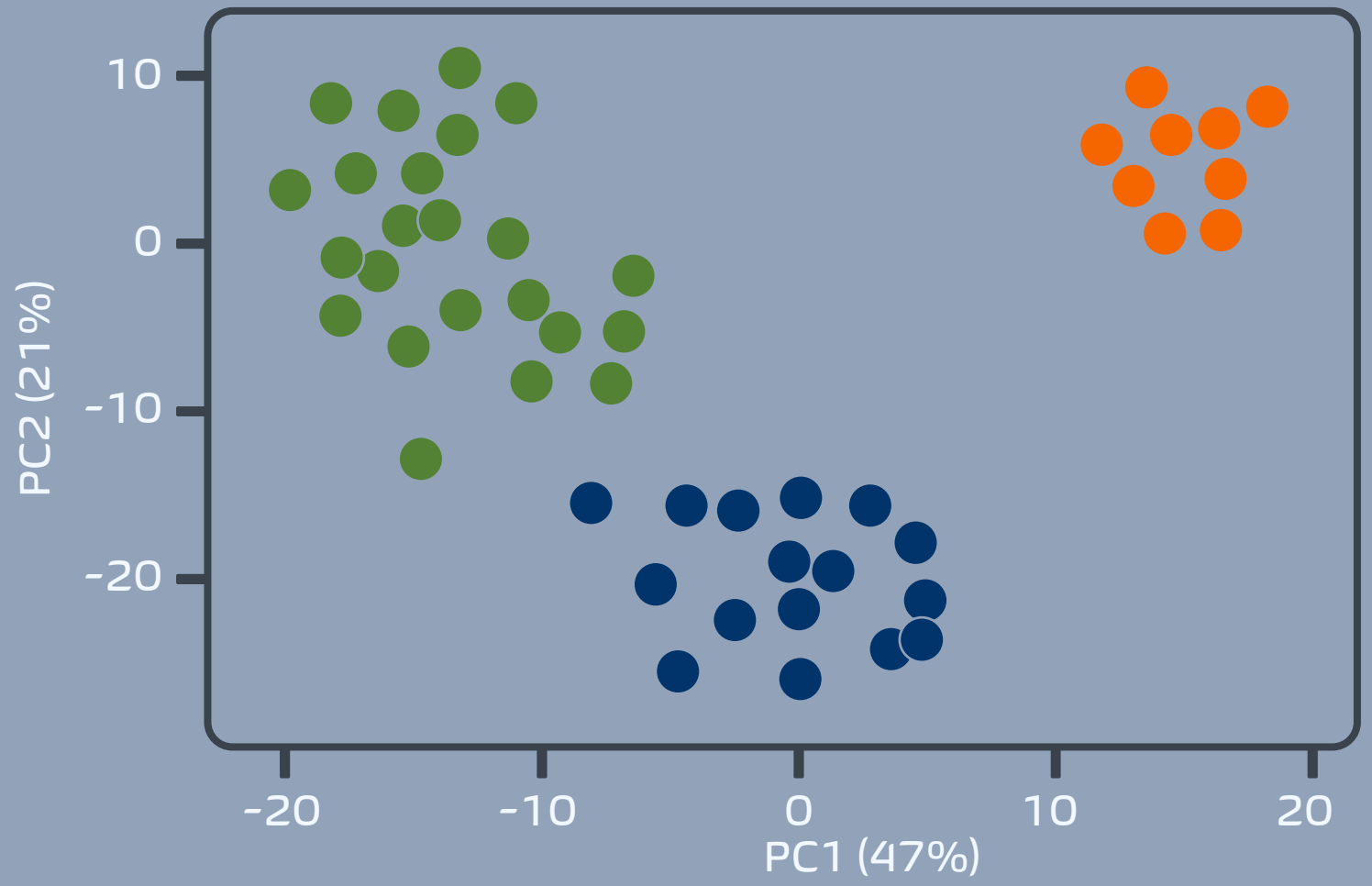
Principal Component Analysis (PCA)



	BMI	Weight	Red blood cells
Pt. 1	-1.4	-1.34	-0.59
Pt. 2	0.39	0.2	-0.95
Pt. 3	-0.7	-0.7	1.16
Pt. 4	-0.3	-0.39	-1.11
Pt. 5	1.19	1.29	0.58
Pt. 6	0.89	0.93	0.91

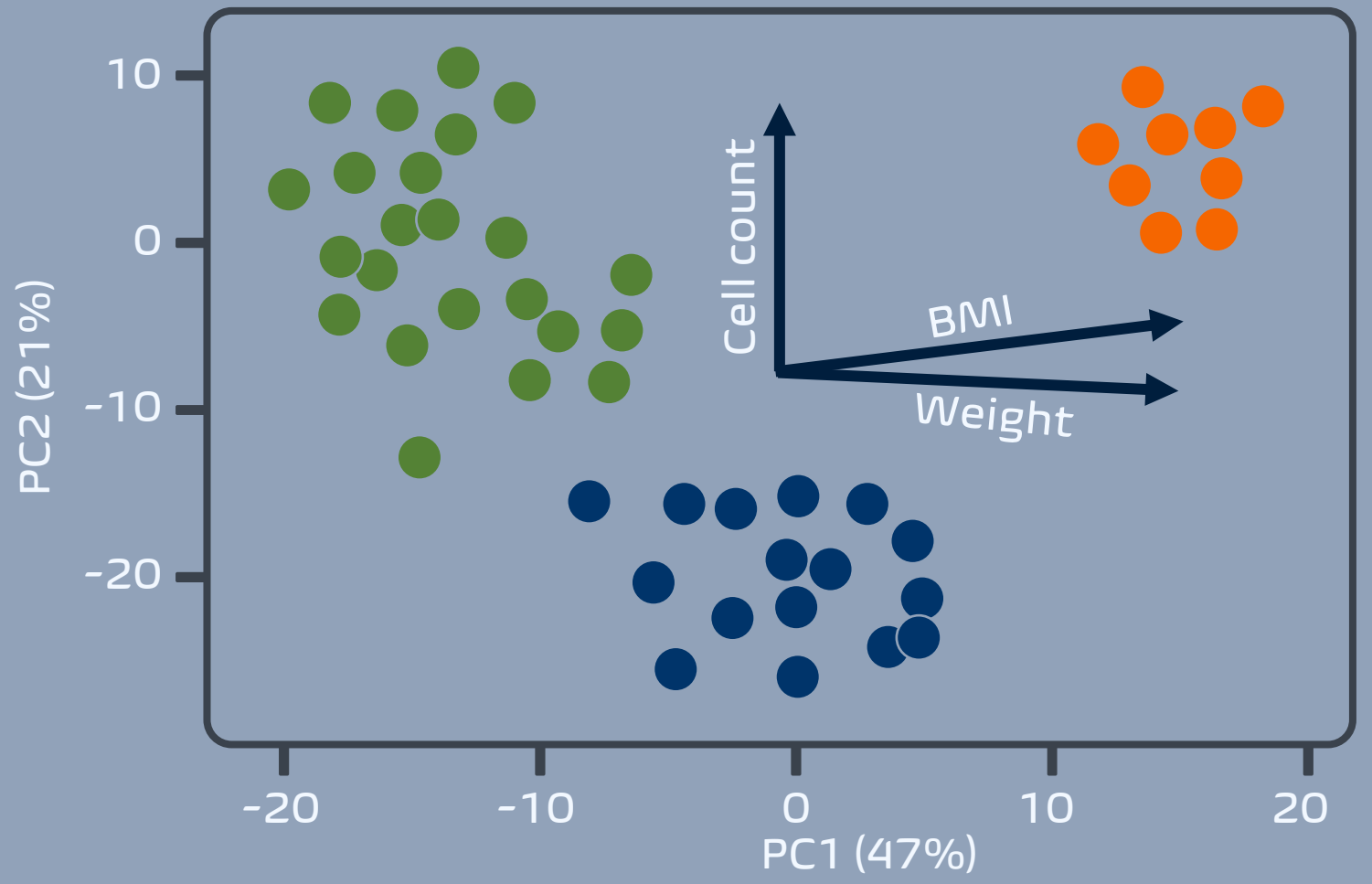
	PC1	PC2
Pt. 1	-1.4	0.6
Pt. 2	0.28	-0.92
Pt. 3	-0.7	1.14
Pt. 4	-0.4	-1.06
Pt. 5	1.25	0.61
Pt. 6	0.91	0.91

PCA score plot



Dimensionality reduction

Biplot



Wrap-up of the lecture

Learning goals

- ✓ Become familiar with Data Science terminology
- ✓ Have an intuitive understanding of clustering
- ✓ Know and understand the most familiar clustering algorithms
- ✓ Understand the purpose of dimensionality reduction
- ✓ Know and understand some dimensionality reduction techniques
- ✓ Be ready for the clustering tutorial!

Questions?