SINTEF

Digital Akademi

UNSUPERVISED LEARNING 101

Katarzyna Michałowska

Outline

1. Unsupervised vs. supervised ML
2. Clustering
3. Dimensionality reduction
4. Anomaly detection
5. *Generative models

Supervised ML

- “Having a teacher”
- The goal is to **make predictions**

“cat”

“dog”

$2.5 million

Unsupervised ML

- “Learning by observation”
- The goal is to **learn patterns**

$\neq$

SINTEF

CLUSTERING

Clustering

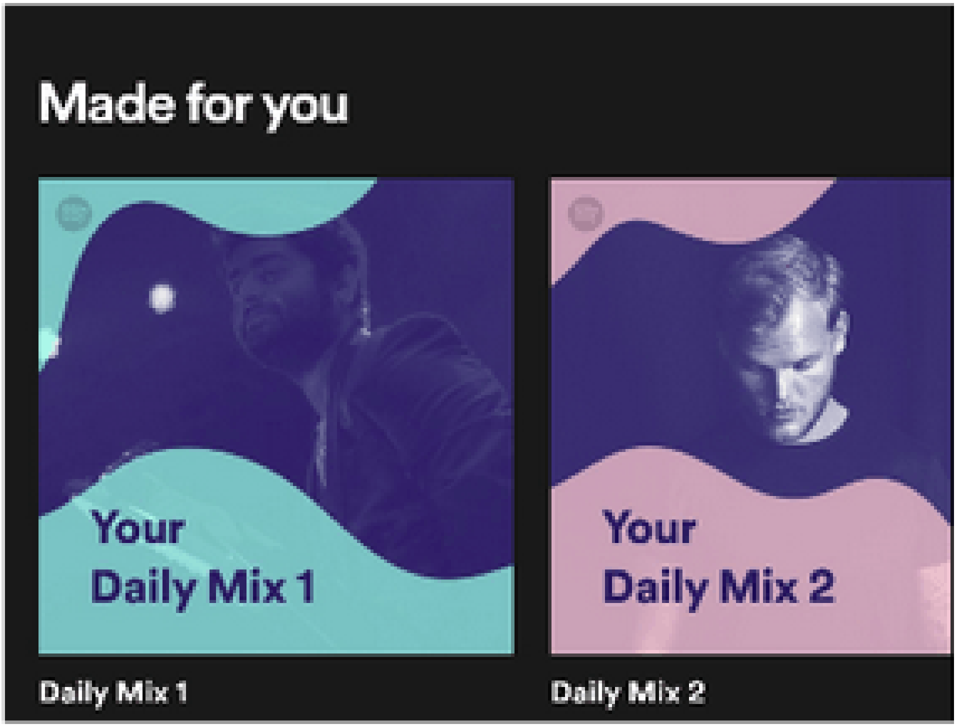
Grouping objects:

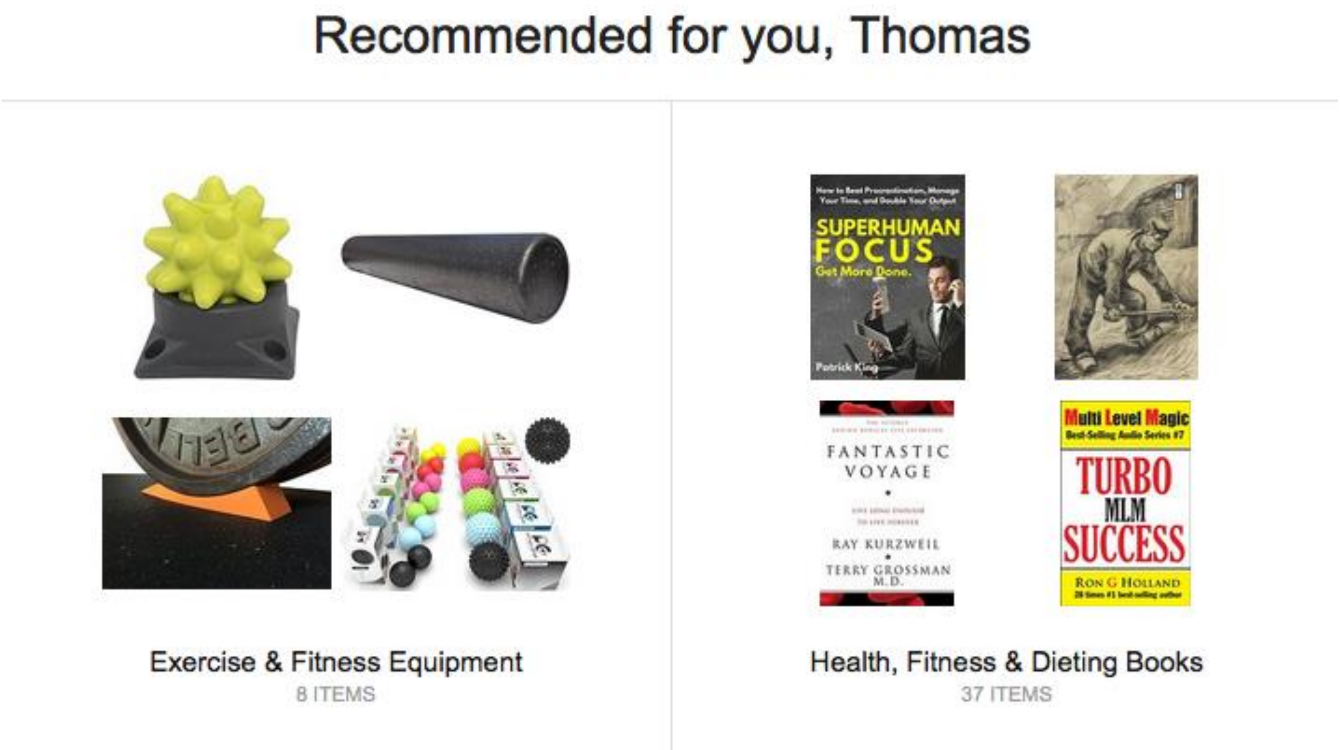
- **Similar** objects go to the same group
- **Dissimilar** objects are **separated** into different groups

(a) (b) (c)

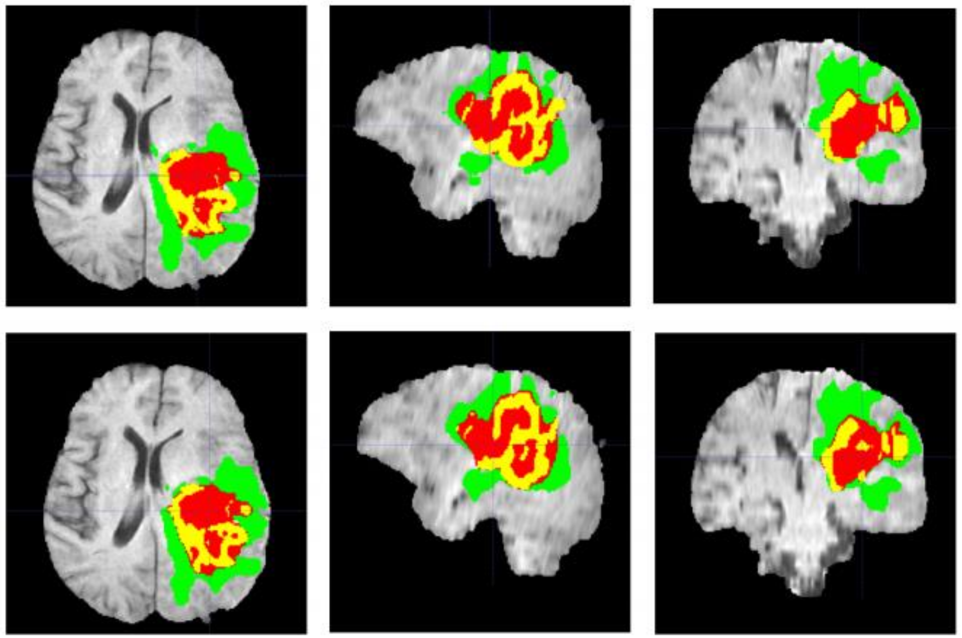
Source: J. Hu, J. Pei, Subspace multi-clustering: a review, 2017.

Example: Recommender systems





Example: Image segmentation



SINTEF

Feature engineering

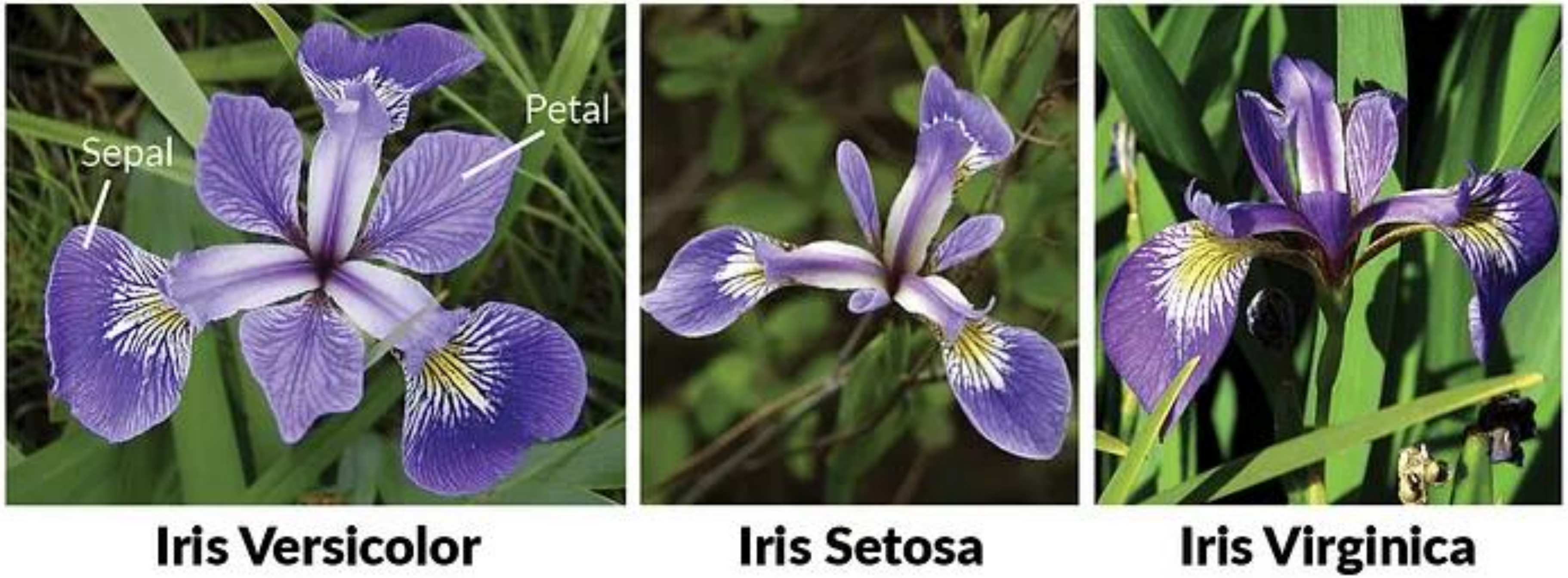


Image credit: https://medium.com/data-science/k-medoids-clustering-on-iris-data-set-1931bf781e05

Feature engineering

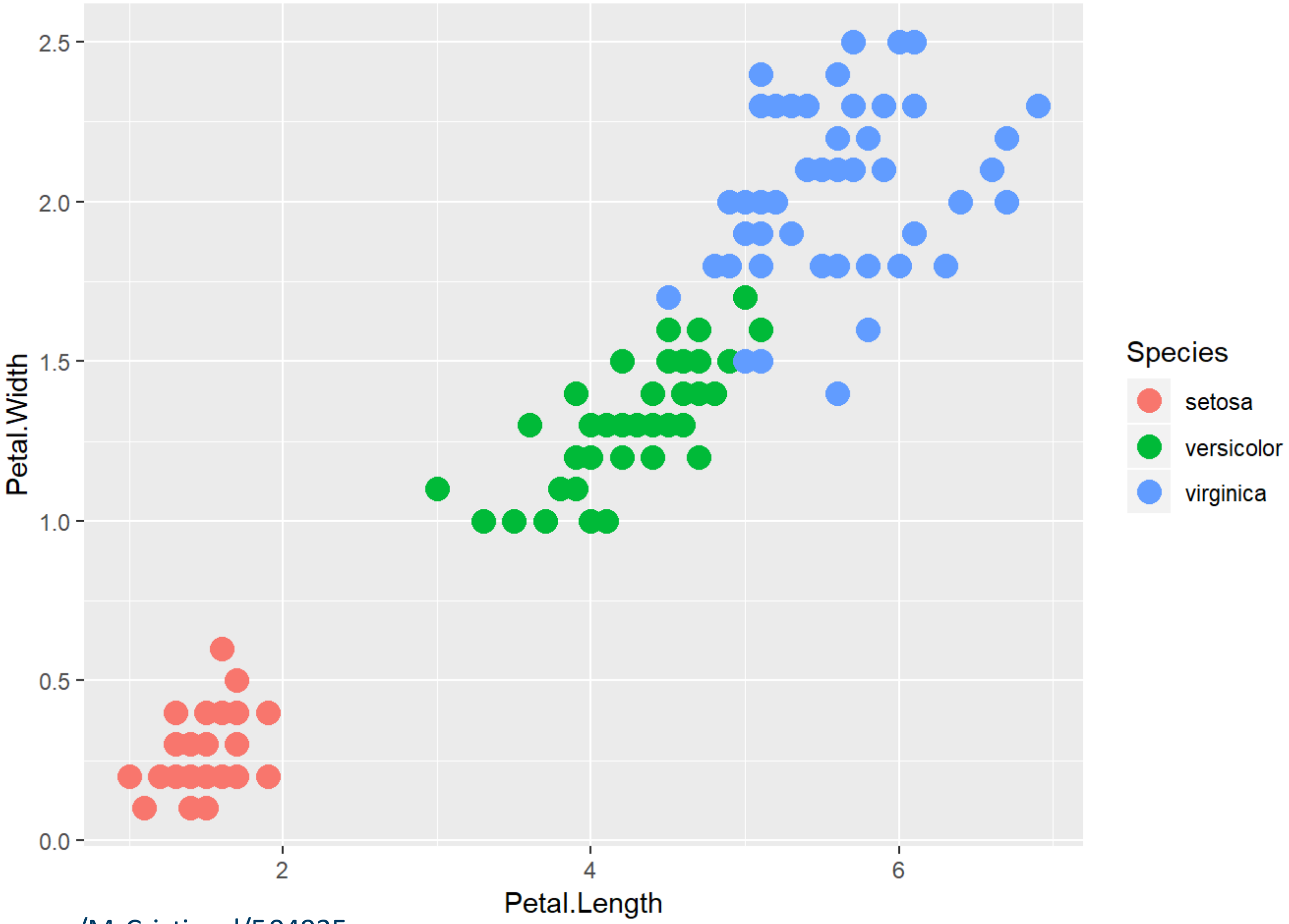
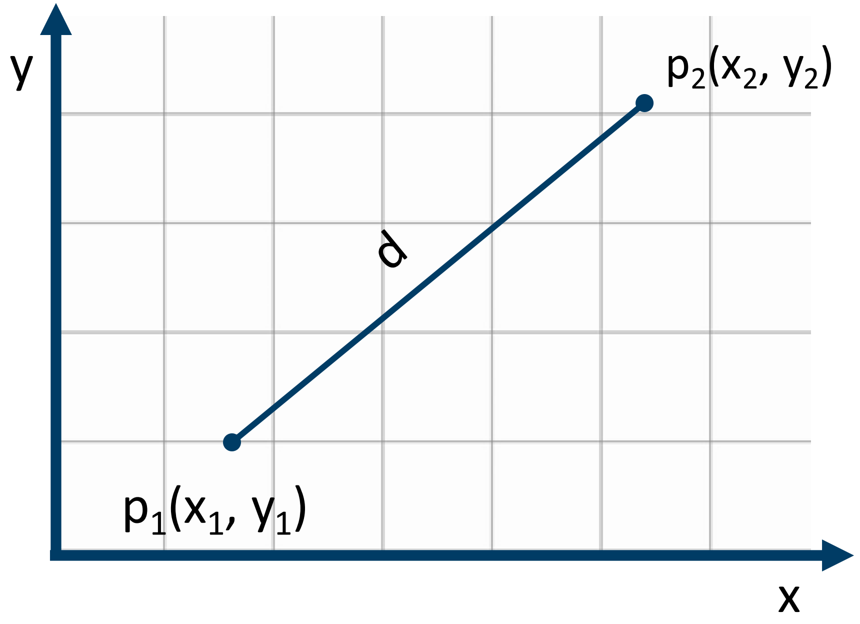


Image credits: https://rpubs.com/MrCristianrl/504935

Multivariate similarity

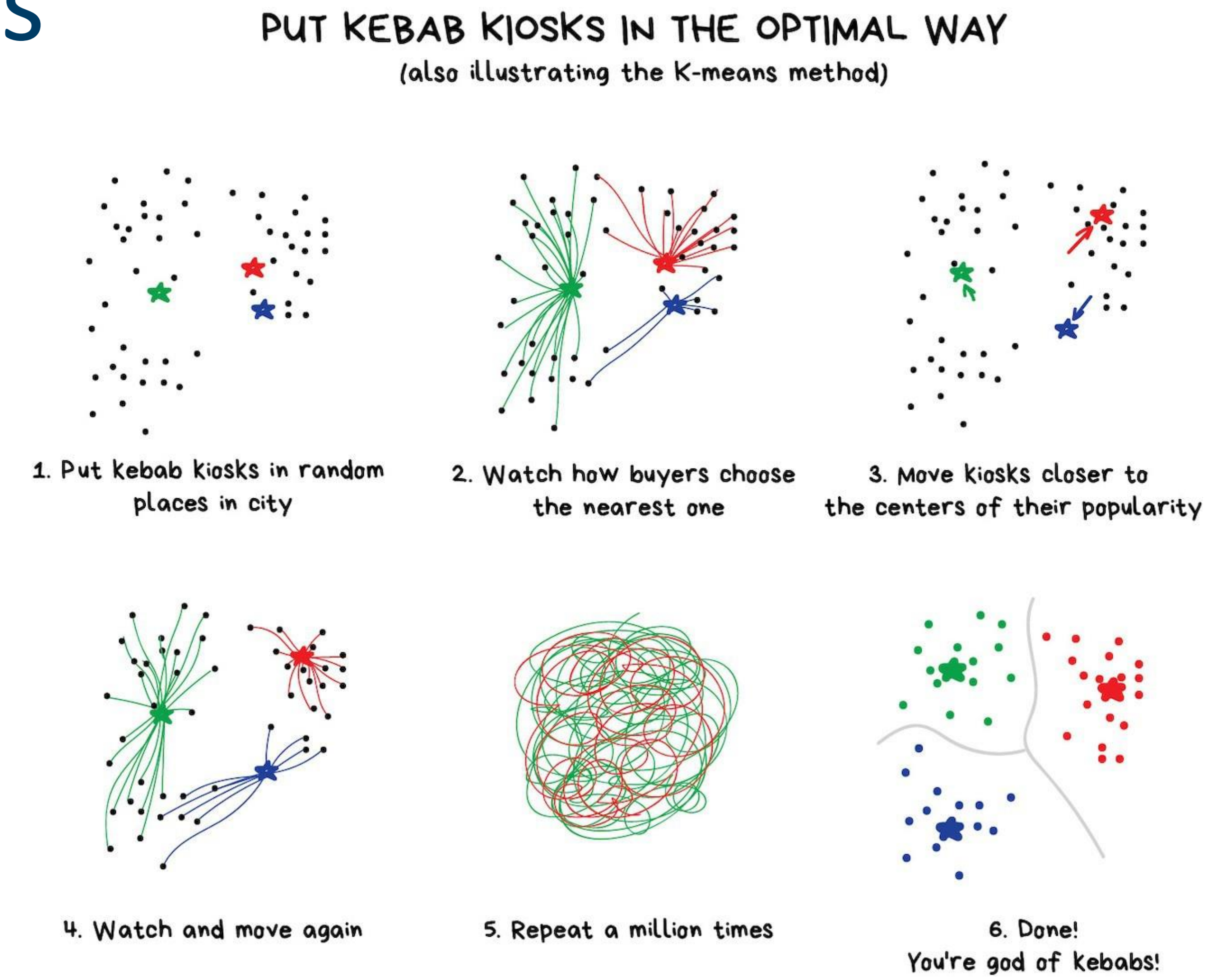
- Distance or correlation metric
- Scaling is important!
  - Petal width (cm)
  - Petal length (mm)

Most common: Euclidean distance



$$d = \sqrt{(x_2 - x_1) + (y_2 - y_1) + \cdots}$$

K-means



kiosk = cluster centroid
buyer = observation
(x,y) position of a buyer = features describing an observation

Image credits: vas3k.com/blog/machine_learning/?ref=hn

SINTEF

K-means

- Parameters: K = nr clusters
- Spherical clusters of similar size
- All points are in clusters

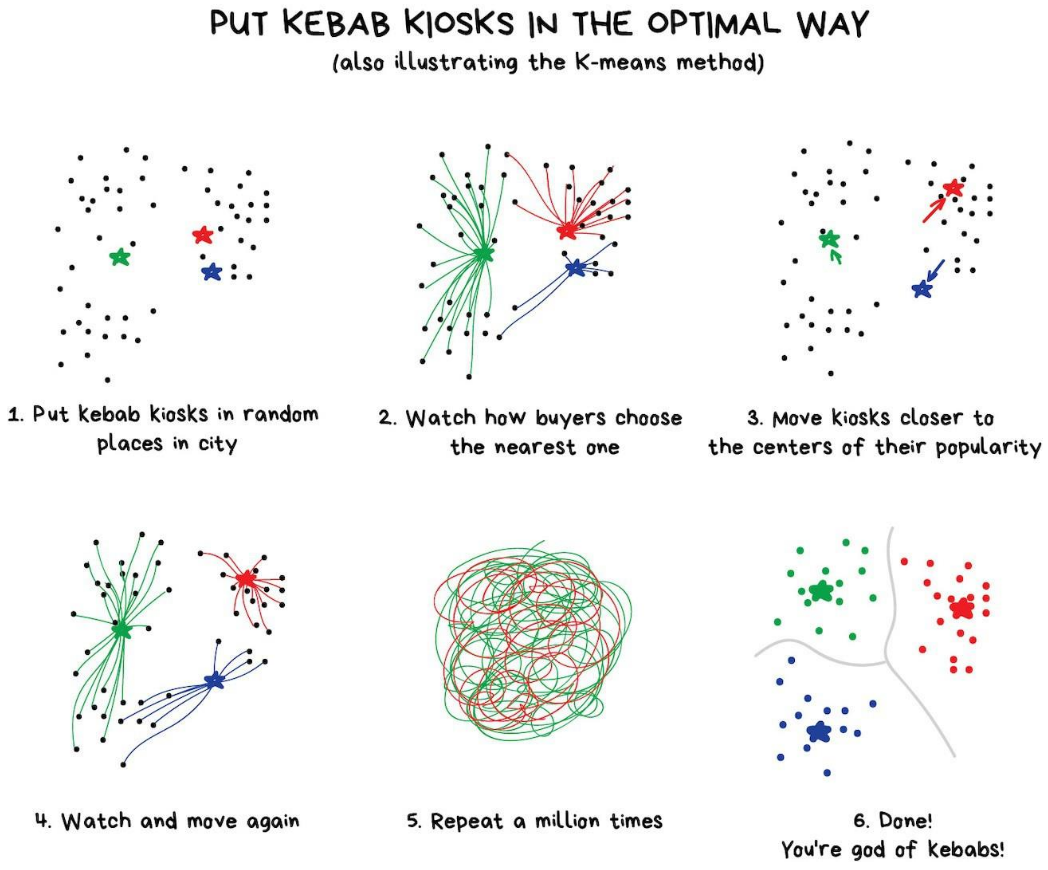
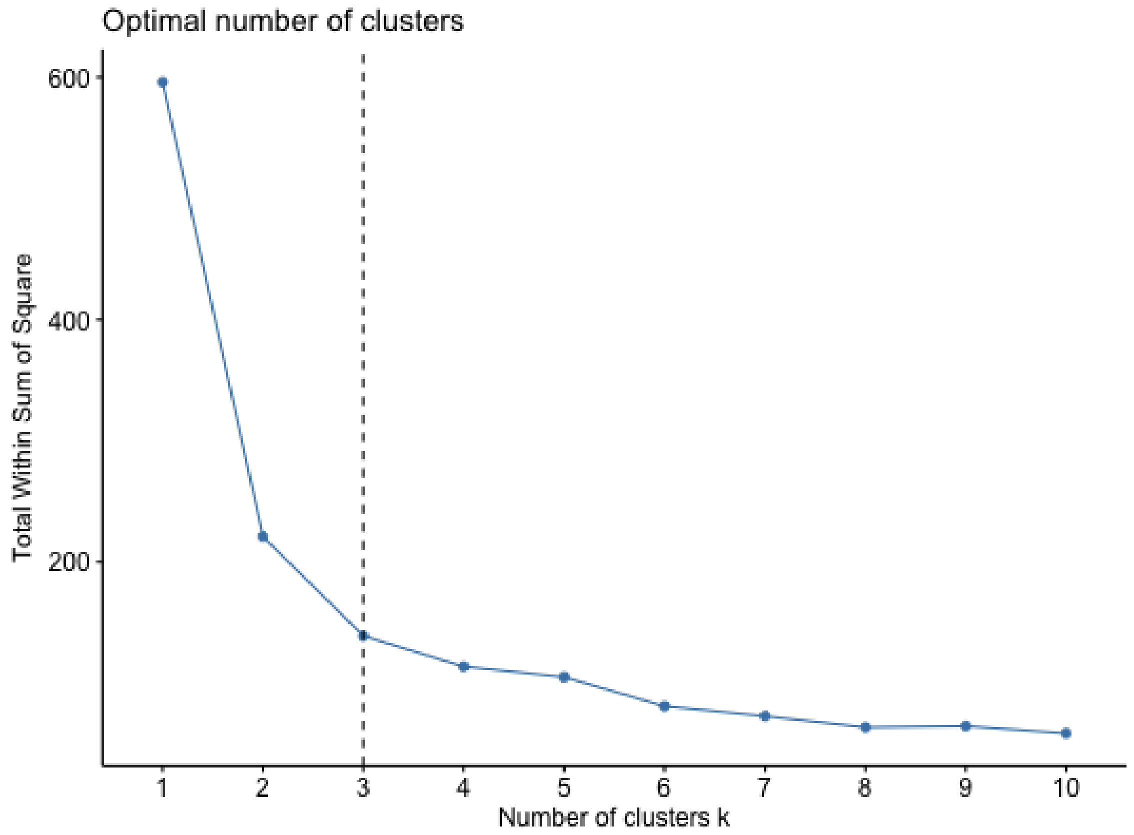


Image credits: vas3k.com/blog/machine_learning/?ref=hn

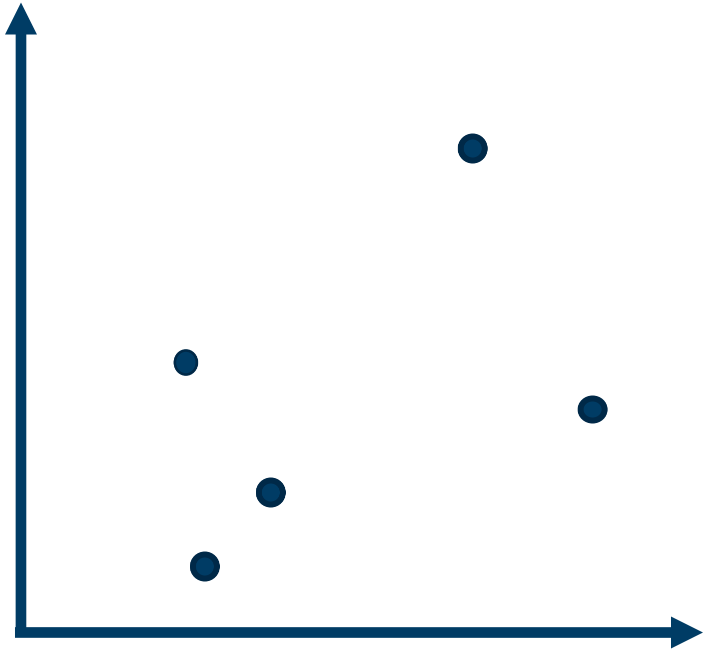
Elbow method

- Choosing number of clusters K
- How tight are the clusters?
- Measures within-cluster sum of squares, i.e., how close the points in a cluster are to its centroid



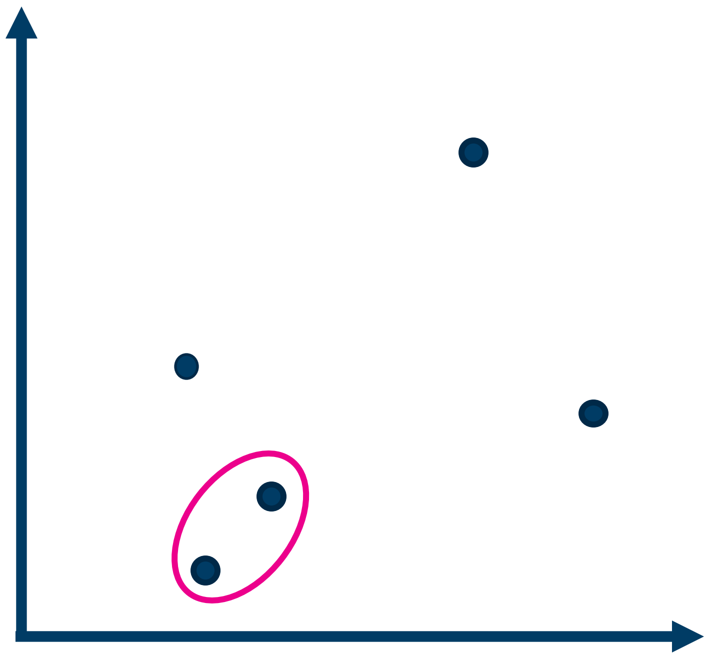
SINTEF

Hierarchical agglomerative clustering



Step 1
Each point is an individual cluster
Calculate a matrix of all distances

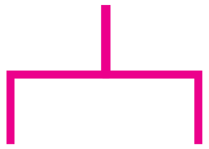
Hierarchical agglomerative clustering



Step 2

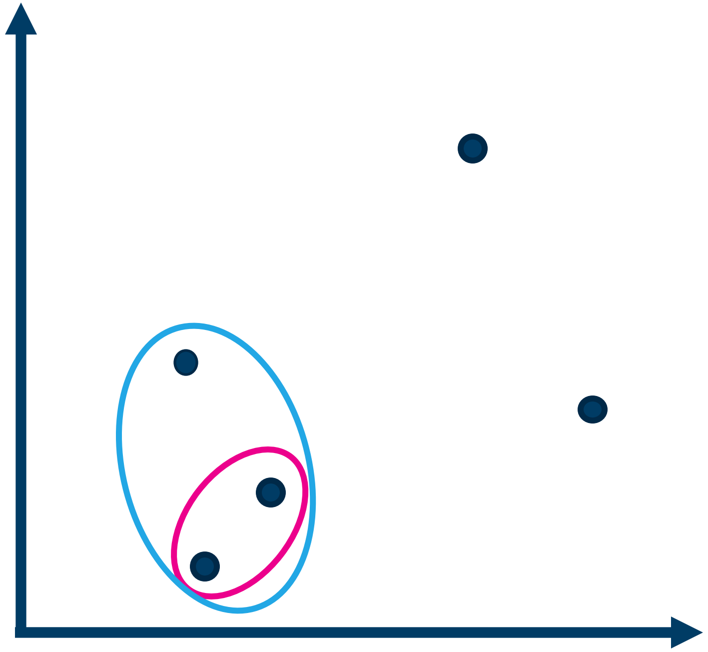
Cluster two closest points

Recalculate the distances of points to this new cluster

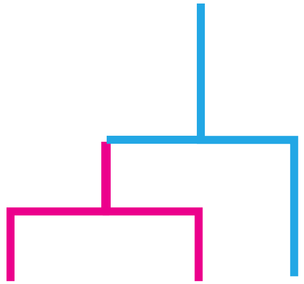


SINTEF

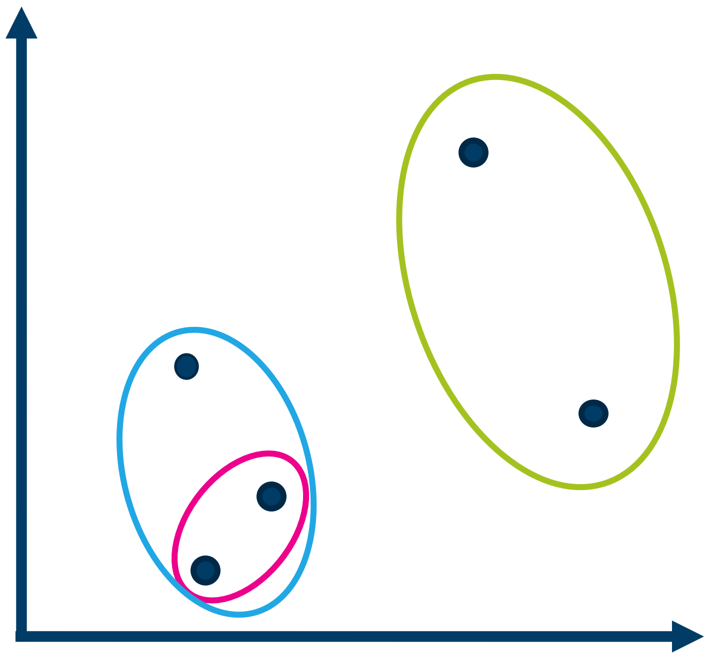
Hierarchical agglomerative clustering



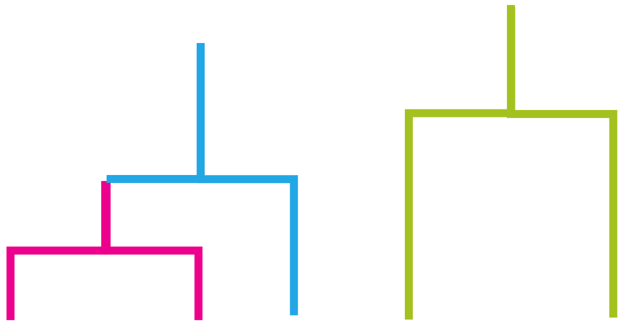
Step 3
Repeat



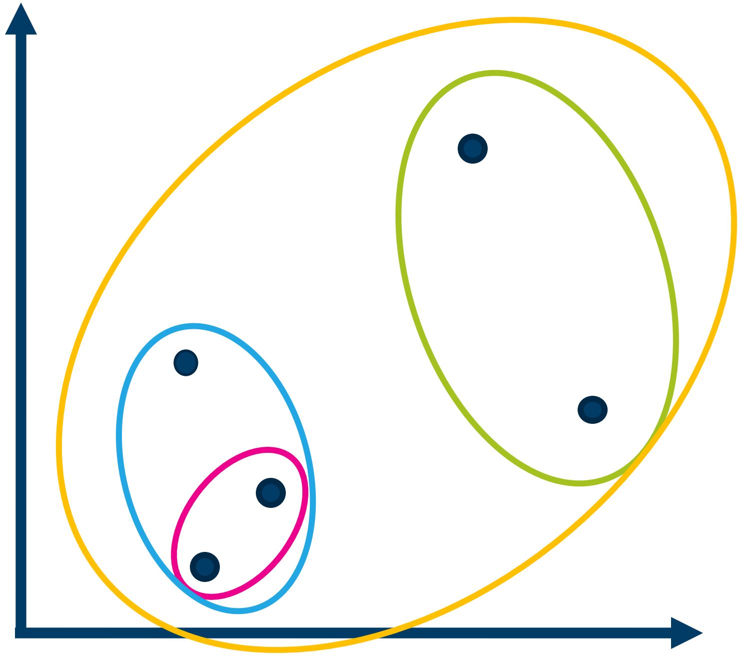
Hierarchical agglomerative clustering

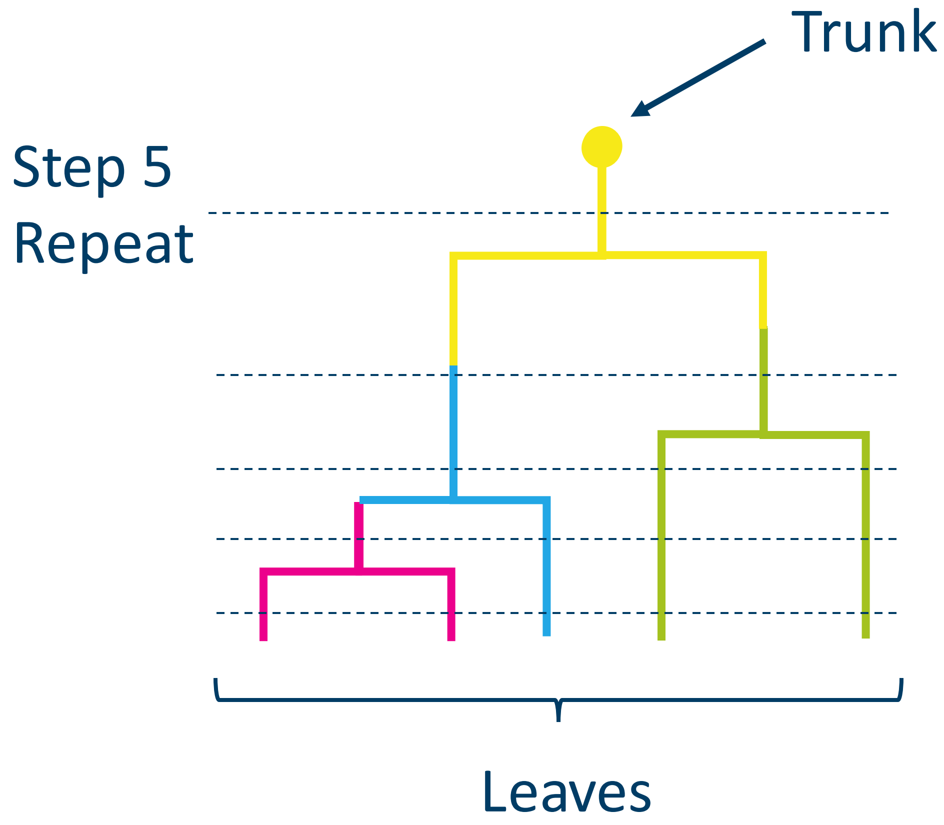


Step 4
Repeat



Hierarchical agglomerative clustering





SINTEF

Hierarchical clustering

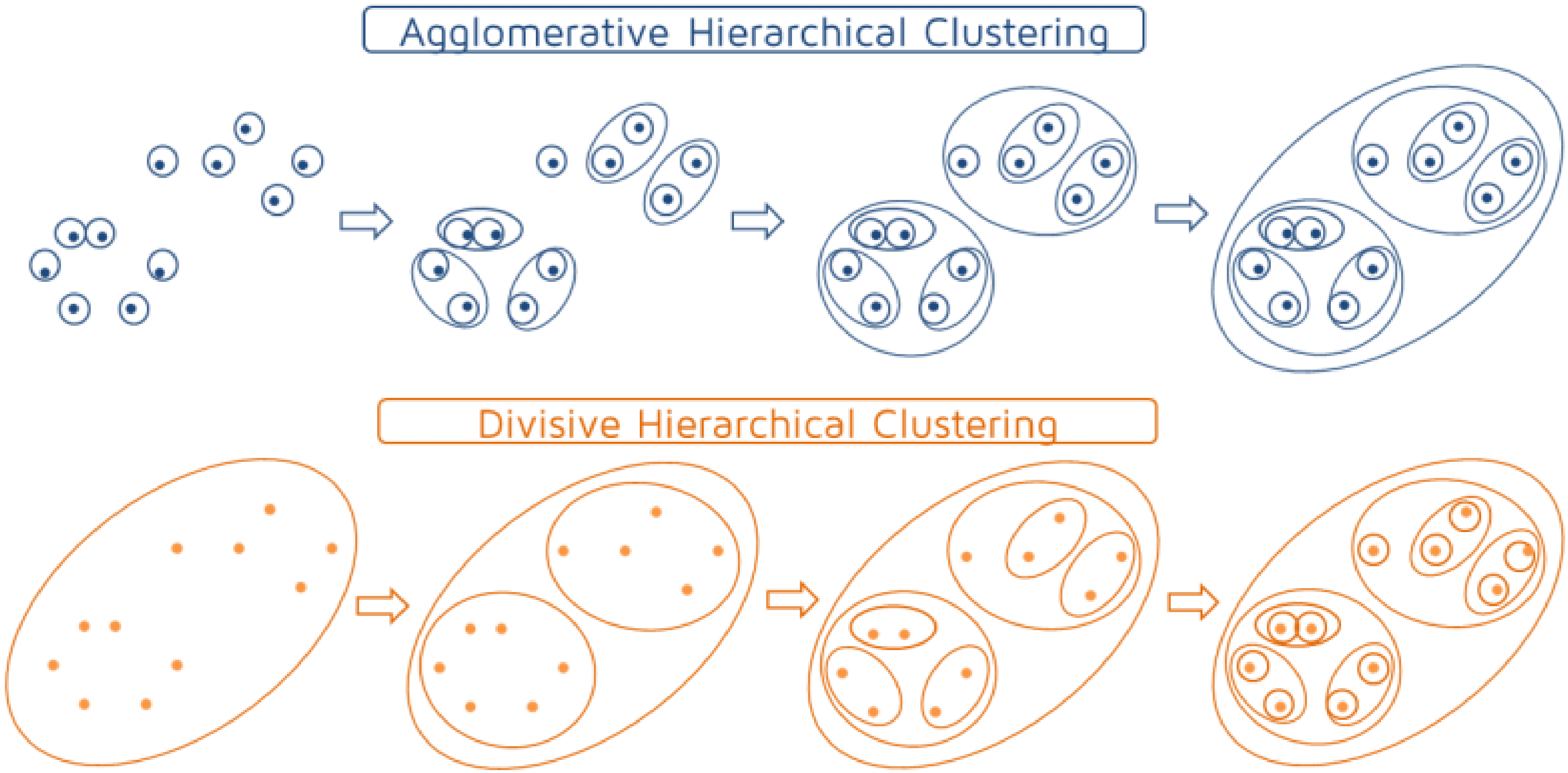


Image credits: https://quantdare.com/hierarchical-clustering/

Hierarchical clustering

- All points are in clusters
- Flexible cluster number and shapes

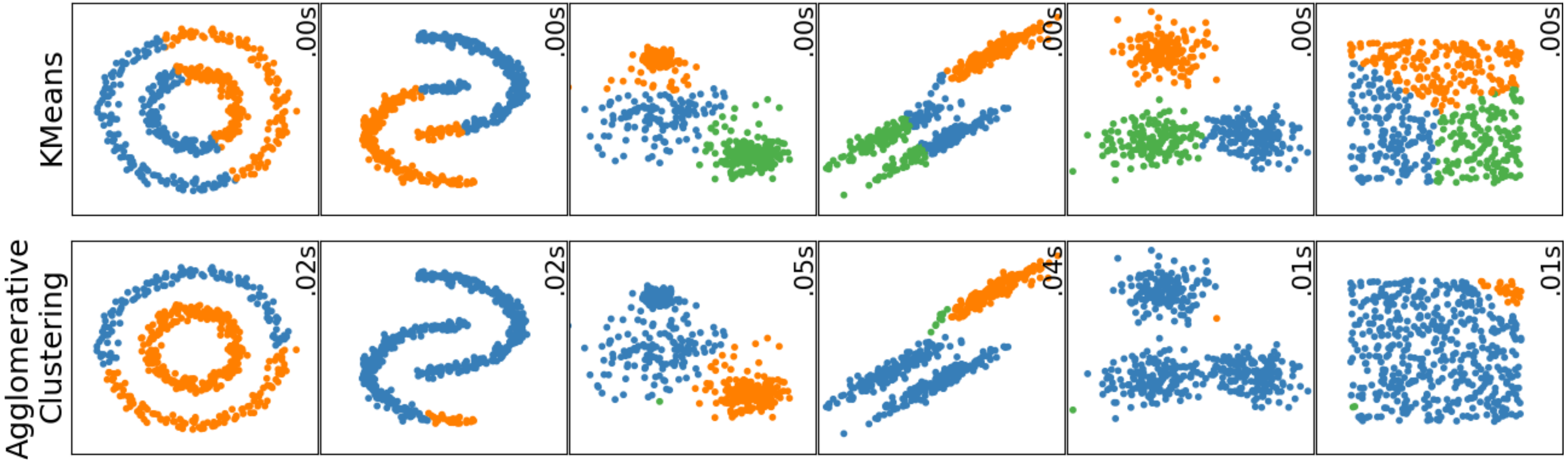


 Image credits: https://scikit-learn.org/stable/modules/clustering.html

DBSCAN: Density-based clustering

- Parameters: ε = radius, min_samples

1. Compute ε neighborhood for every sample.
2. Identify *core points* that have at least min_samples in their neighborhood.
3. Expand the groups as long as you find new core points.

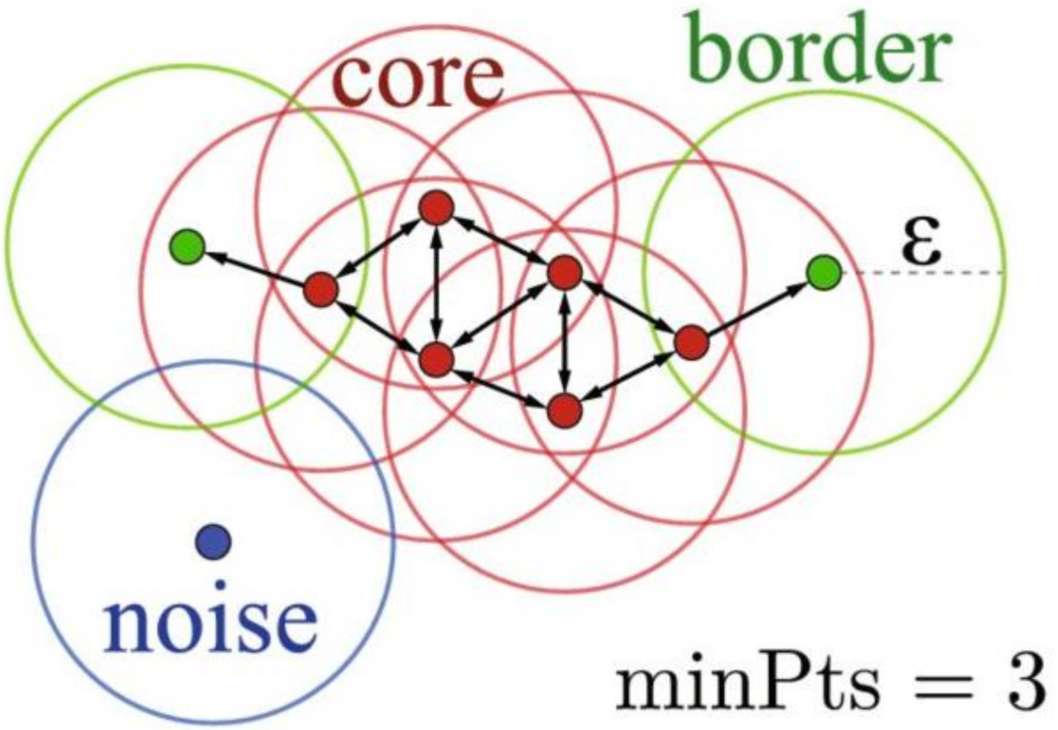


Image credits: Mehle et al., In-line recognition of agglomerated pharmaceutical pellets with density-based clustering and convolutional neural network, 2017.

DBSCAN

- Flexible cluster number, size and shape
- Can identify outliers

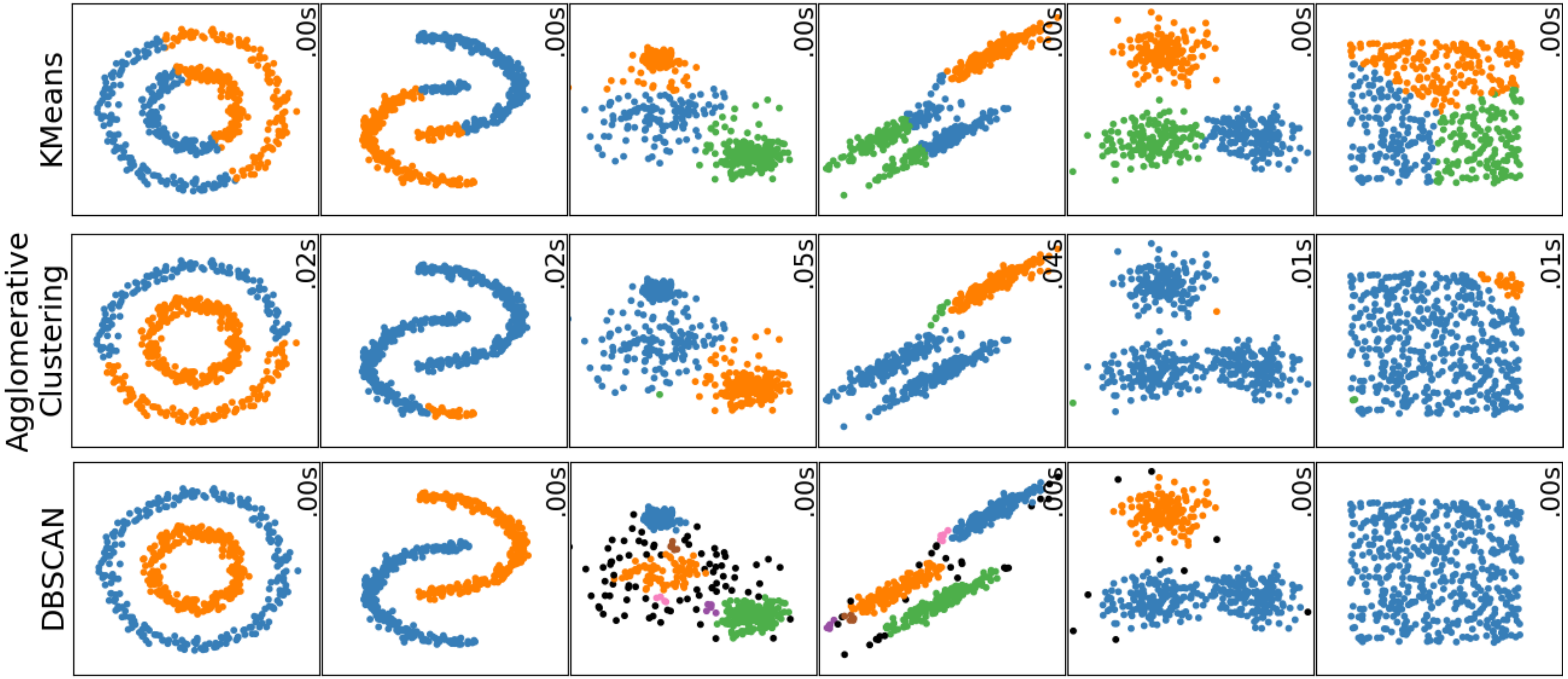


Image credits: https://scikit-learn.org/stable/modules/clustering.html

SINTEF

HDBSCAN: Hierarchical DBSCAN

- Hierarchy of density levels instead of a fixed density based on ε

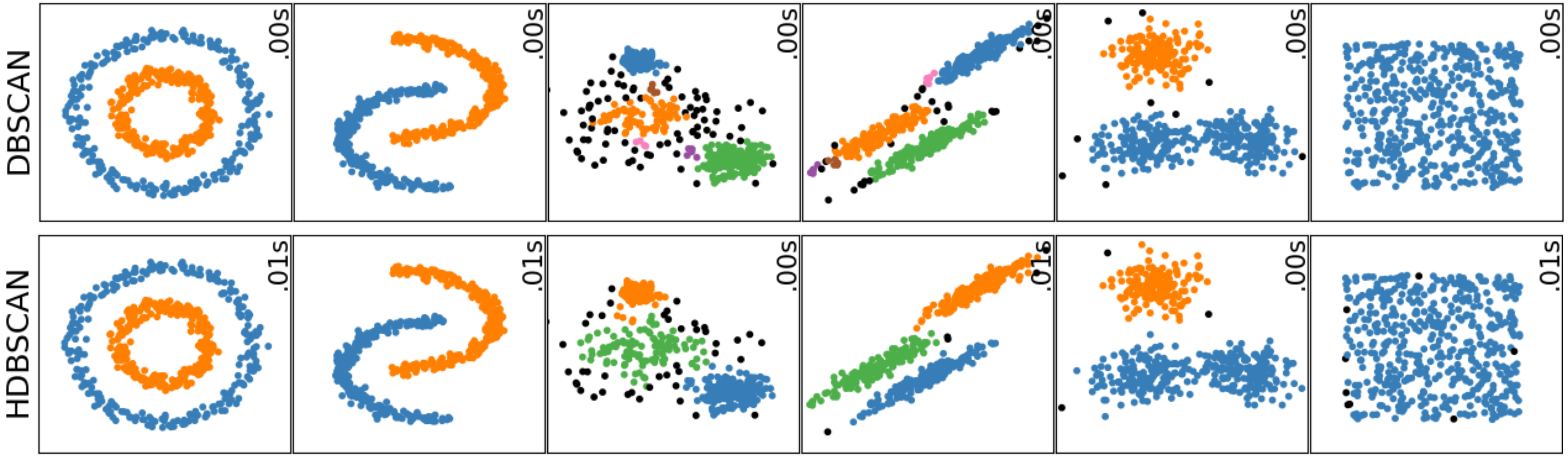


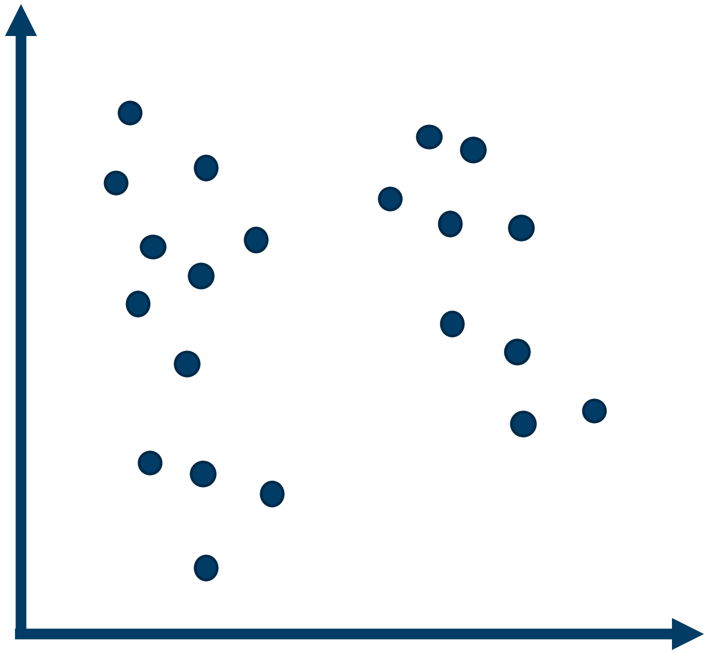
Image credits: https://scikit-learn.org/stable/modules/clustering.html

SINTEF

Optimal number of clusters

The 'best' number of clusters depends on the application

- Geography: Continent, country, state, city, neighbourhood
- Customers: Demographic group, behavioral segments, personalized micro-segments
- Machine: Type, type + operating mode (idle, active...)



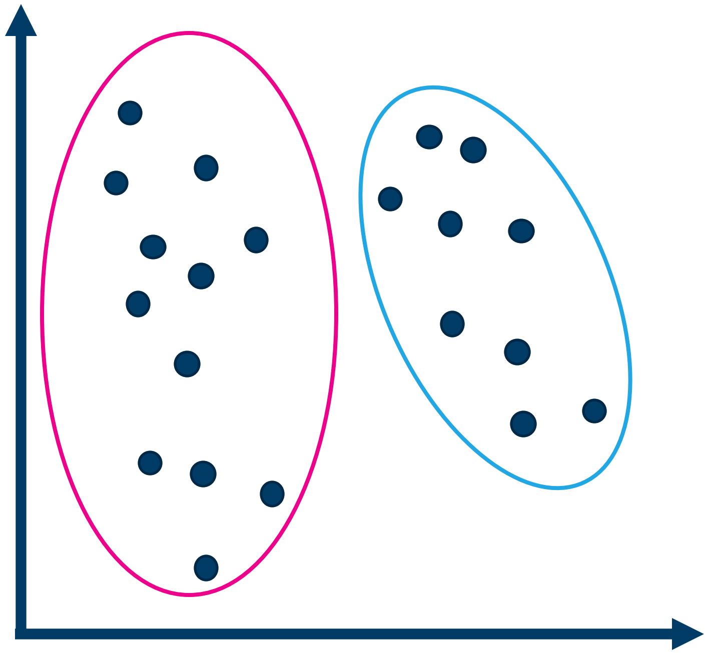
SINTEF

Optimal number of clusters

The 'best' number of clusters depends on the application

Think about:

- Geographical regions of different sizes
- Taxonomic families
- Etc.

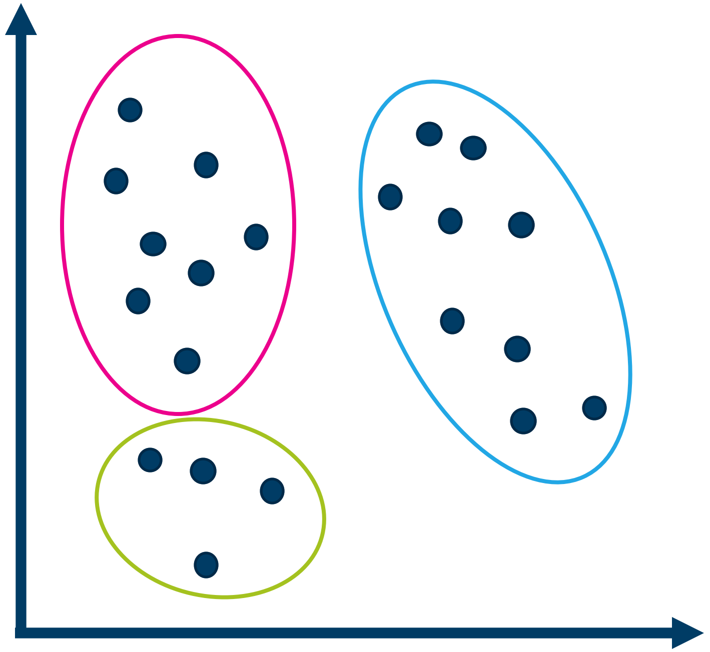


Optimal number of clusters

The 'best' number of clusters depends on the application

Think about:

- Geographical regions of different sizes
- Taxonomic families
- Etc.



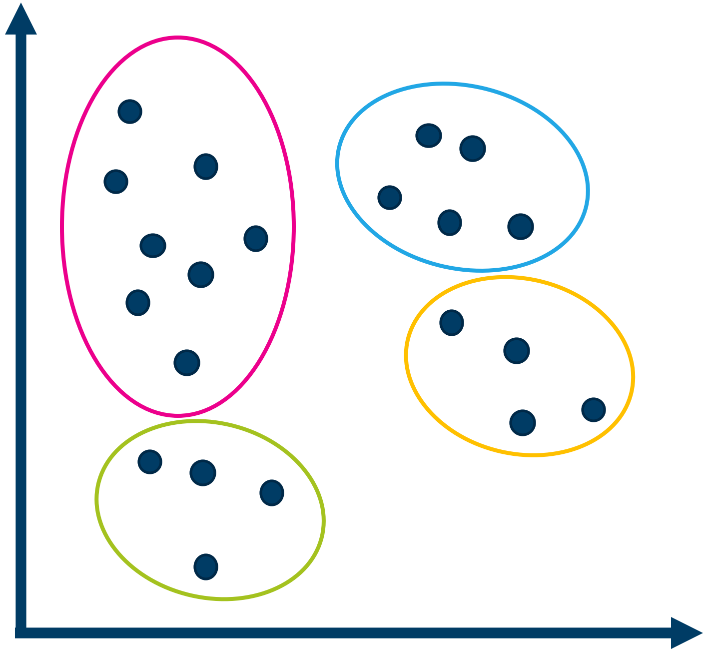
SINTEF

Optimal number of clusters

The 'best' number of clusters depends on the application

Think about:

- Geographical regions of different sizes
- Taxonomic families
- Etc.



Curse of dimensionality

- In high dimensions, all points are similarly far
- Distances lose meaning — clustering becomes unreliable

Solution: Feature selection or **dimensionality reduction**

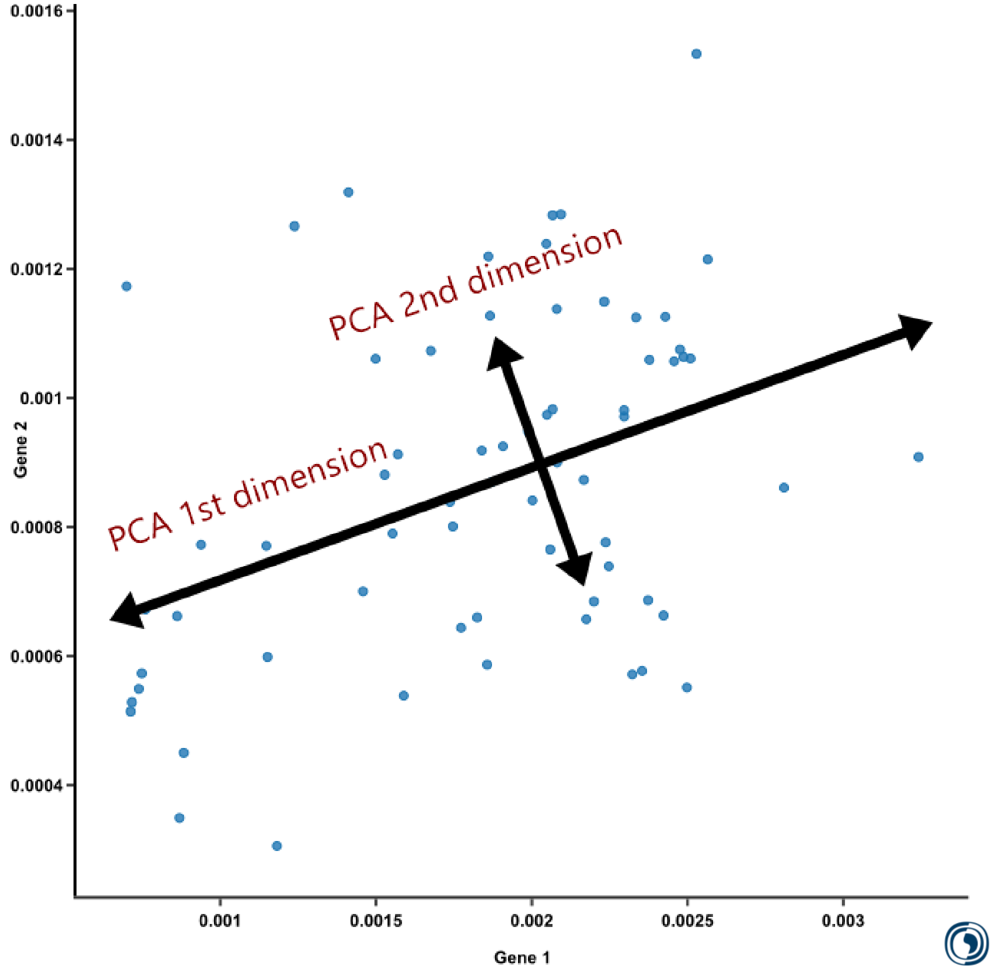
DIMENSIONALITY REDUCTION

Dimensionality reduction

- Transform data to a **lower-dimensional space**
- Preserve the most **important structure** or **patterns**
- Remove **noise** and **redundancy**

Principal component analysis (PCA)

- Compute how each feature varies **with each other** (covariance)
- PCA finds new axes that are:
  - **Uncorrelated** (orthogonal directions)
  - Ordered by how much **variance** they capture
- PCA is a linear projection



SINTEF

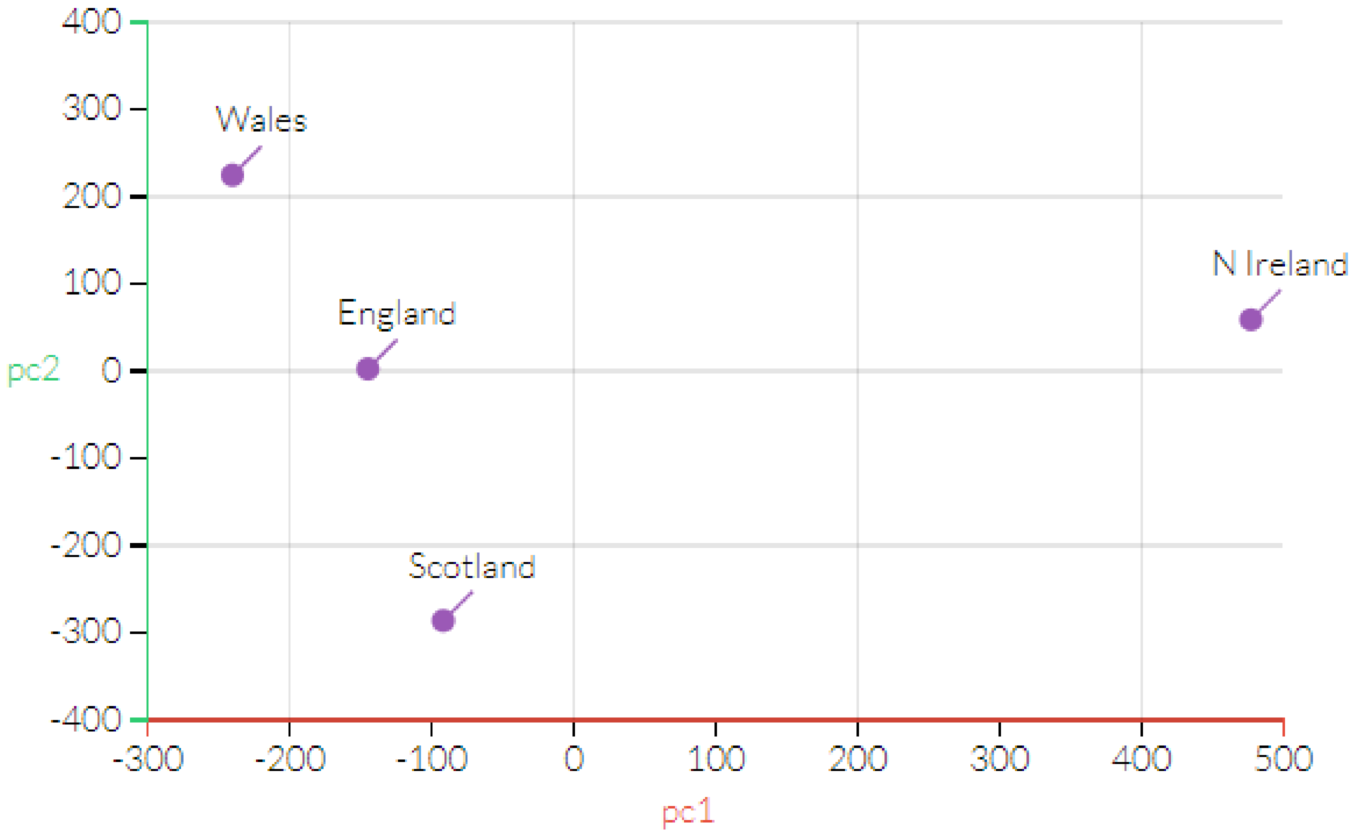
Example: Food habits across countries

Plot: Average consumption of 17 types of food in grams per person per week for every country in the UK.

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

setosa.io/ev/principal-component-analysis/

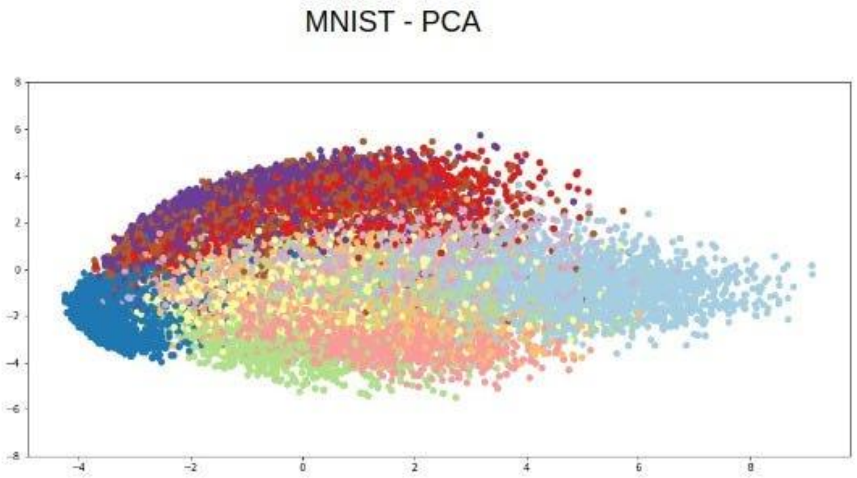
Principal component analysis

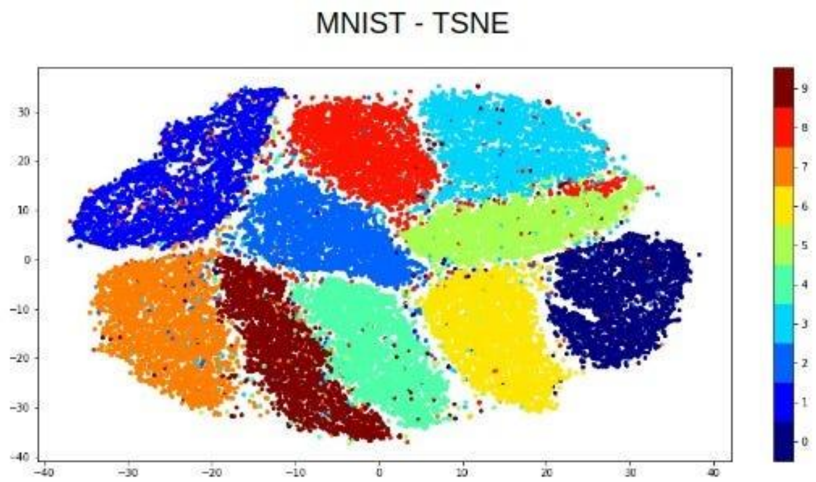


setosa.io/ev/principal-component-analysis/

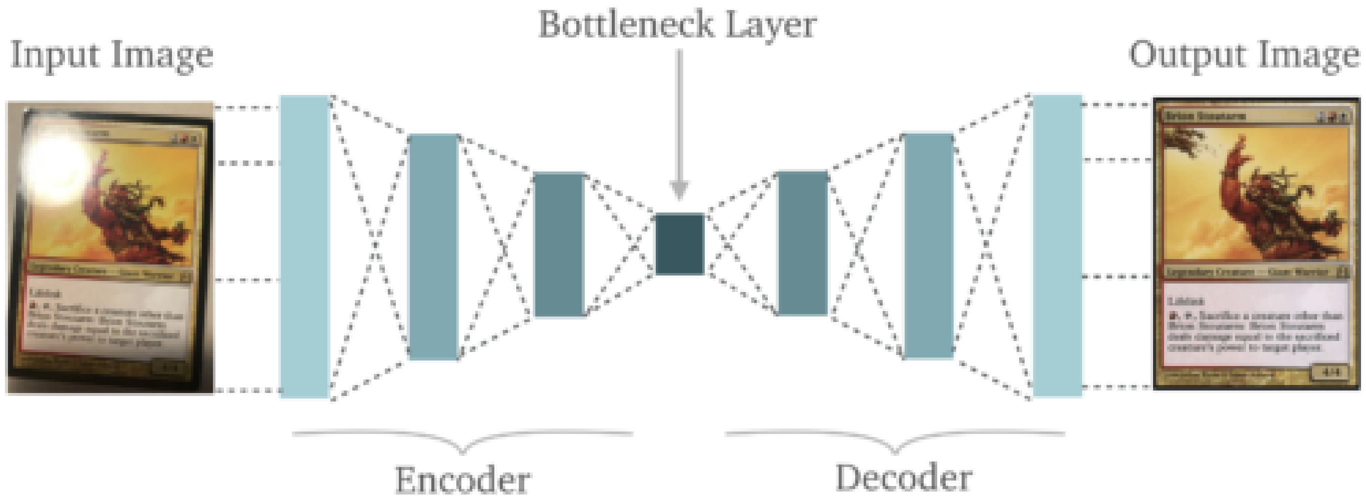
t-SNE

- Non-linear projection to a lower dimension
- Create new axes where similar points are close to each other:
- Used mainly for visualization - cannot be reapplied to new data





Autoencoders



- Bottleneck forces compressed representation
- Input = output
- Repetitive patterns across data samples
- Bonus: denoising

ANOMALY DETECTION

SINTEF

Anomaly detection

- Identify **deviations from usual behavior**: data points that differ significantly from the majority.
- Also called outliers, novelties, or anomalies.
- Anomalies can be errors, threats, or rare events.

Examples

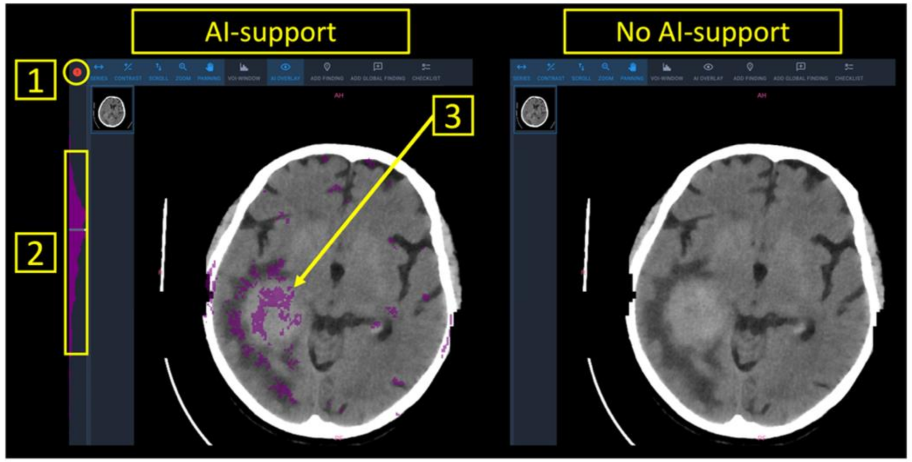
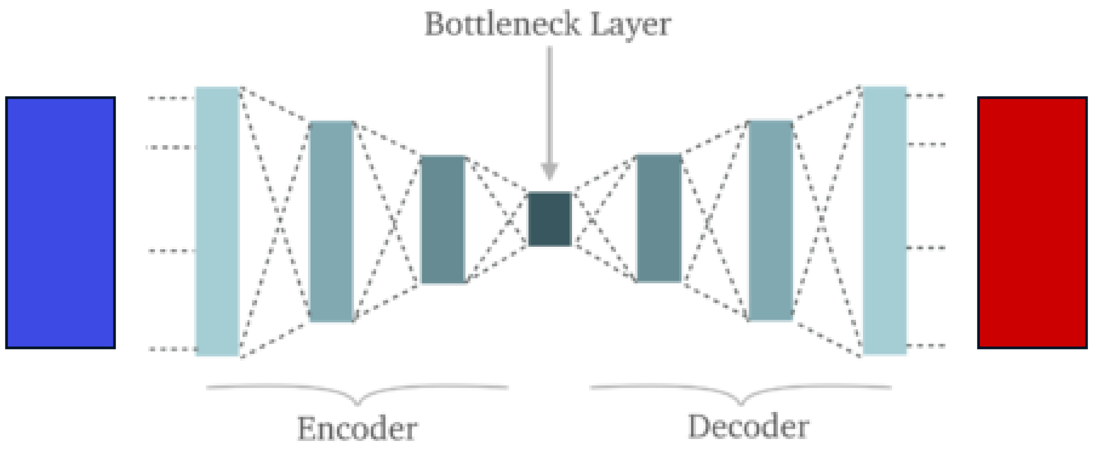
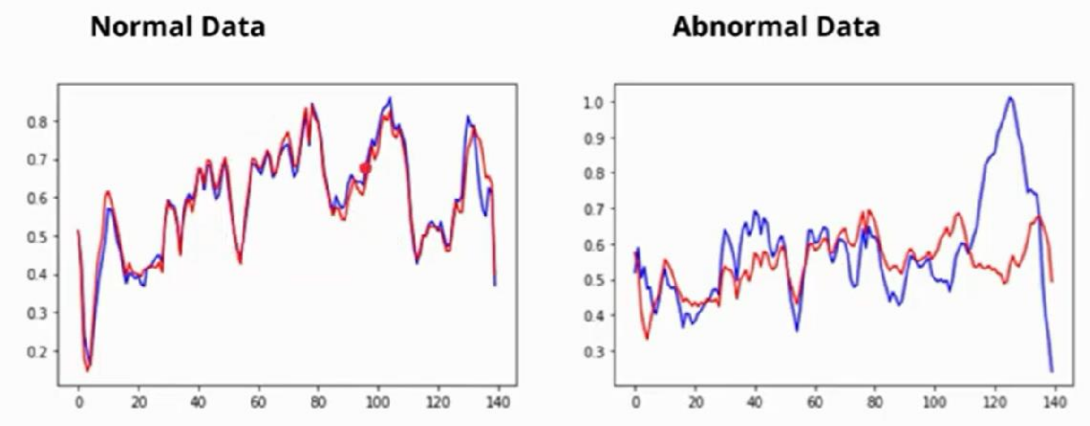


Image credits: Finck et al., Faster and Better: How Anomaly Detection Can Accelerate and Improve Reporting of Head Computed Tomography, 2022.

Autoencoder

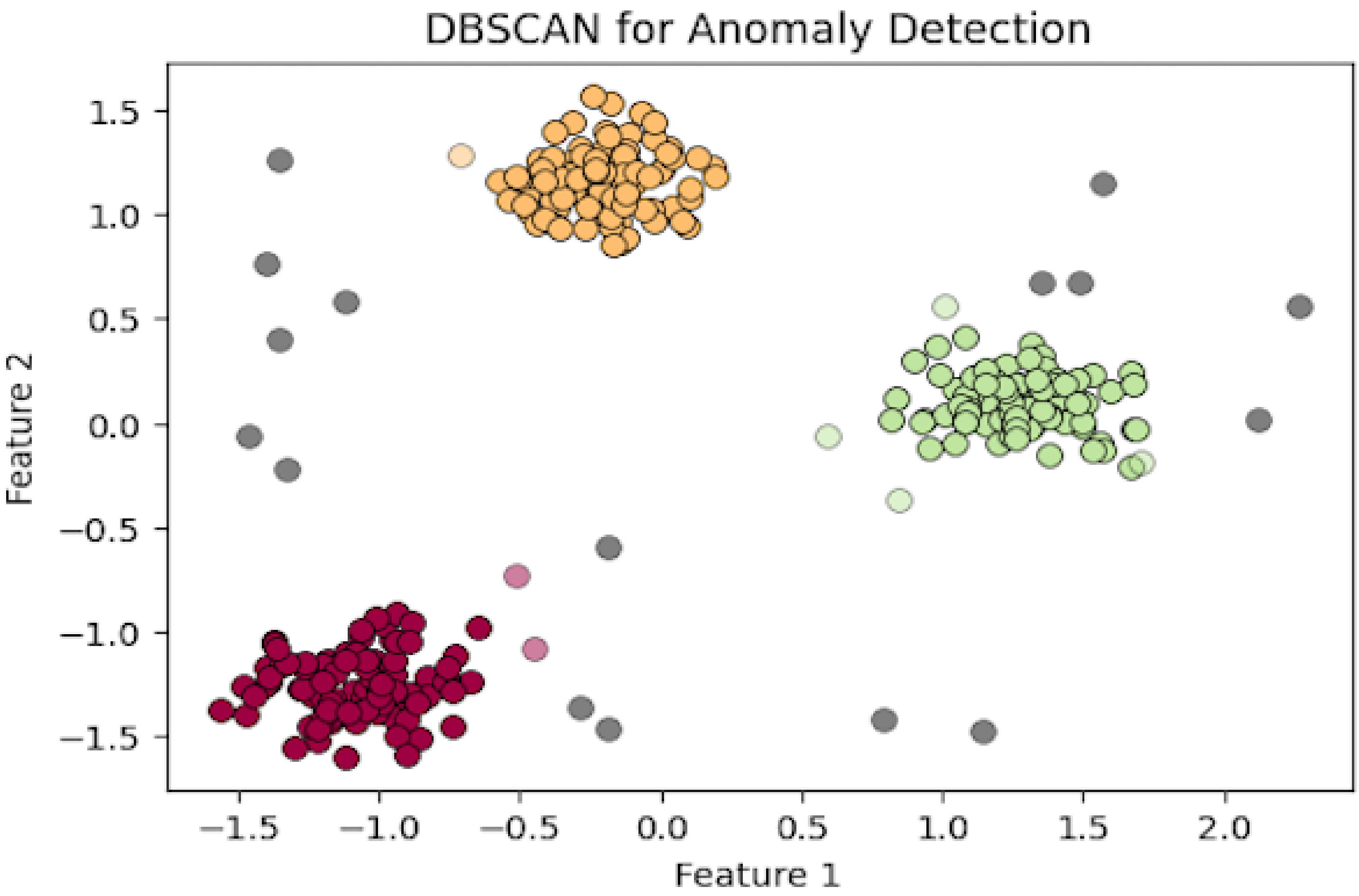
- Train autoencoder on normal behavior only
- Monitor reconstruction error to detect anomalies





Reconstruction vs. signal

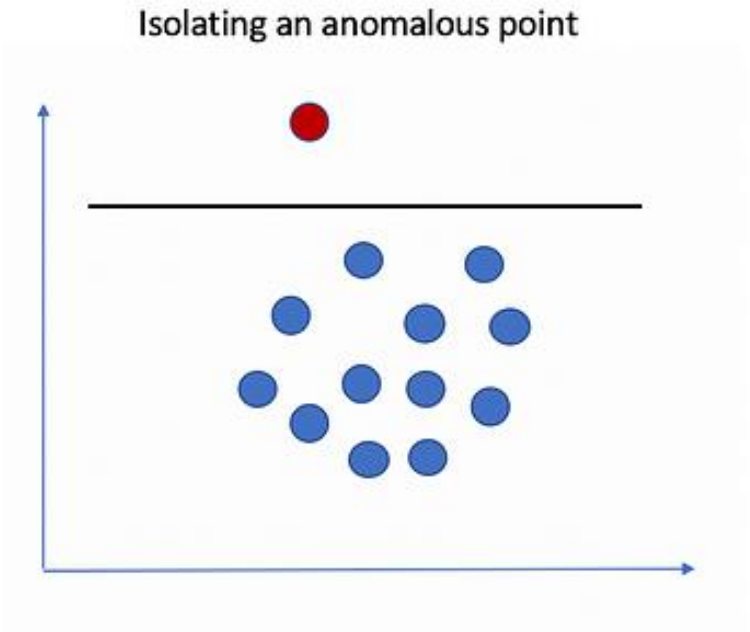
Clustering

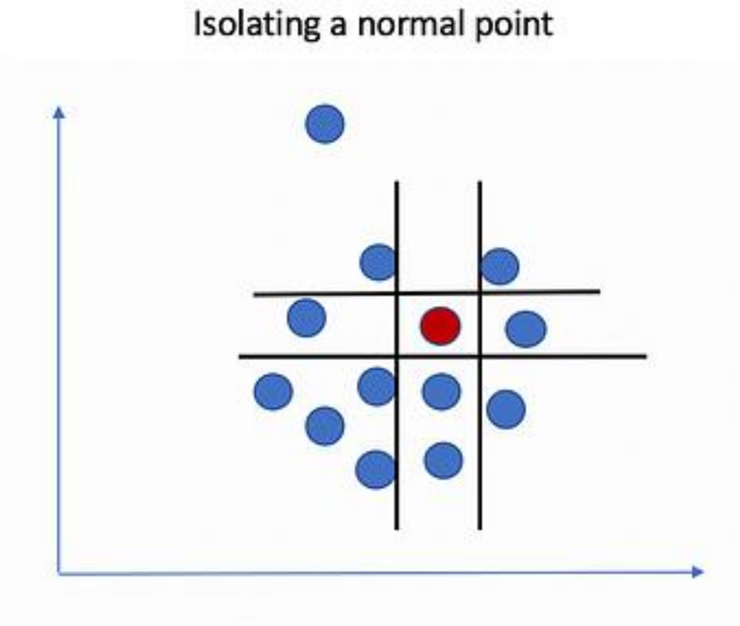


Credits: https://blog.gopenai.com/unsupervised-fraud-detection-algorithms-detecting-the-unknown-9075b4447c68

Isolation forest

1. Randomly select a feature and split point in the data.
2. Build trees that isolate individual points.
3. The fewer splits are needed to isolate a point, the more likely its anomalous.





Credits: https://blog.gopenai.com/unsupervised-fraud-detection-algorithms-detecting-the-unknown-9075b4447c68

Feature engineering is important!

1. Capture normal behavior:
   - User shops more on weekends than weekdays
   - A machine heats up gradually during long operation
2. Highlight irregular patterns
   - A sudden purchase in a foreign country at 3 a.m.
   - A machine suddenly vibrates at high frequency after stable operation
3. Ignore irrelevant variance:
   - Minor fluctuations in transaction amount (e.g., €24.97 vs €25.00)
   - Sensor readings affected by ambient temperature changes during the day

GENERATIVE MODELS

Generative adversarial networks

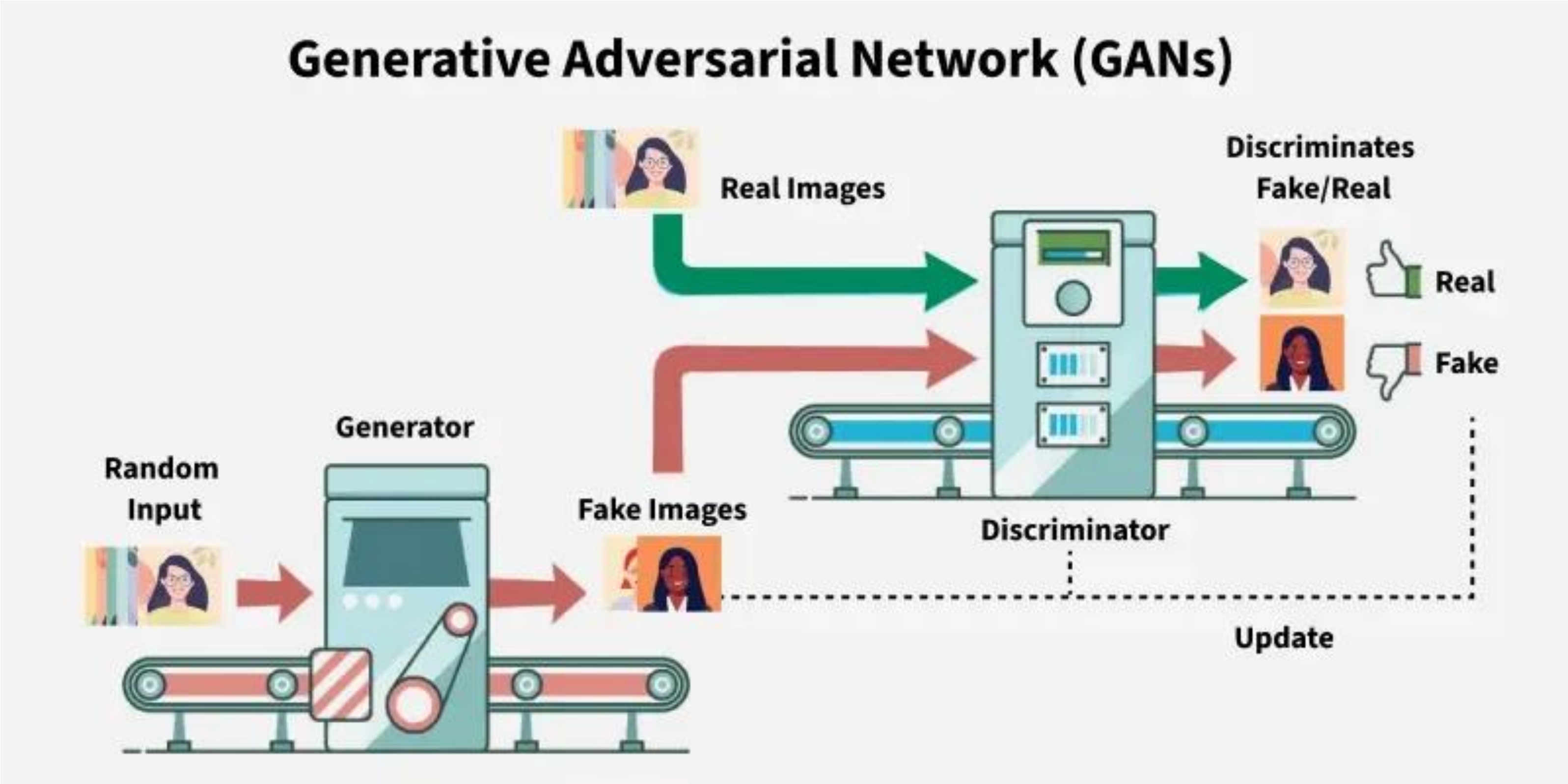


Image credit: https://www.geeksforgeeks.org/generative-adversarial-network-gan/

Example: Cycle GAN

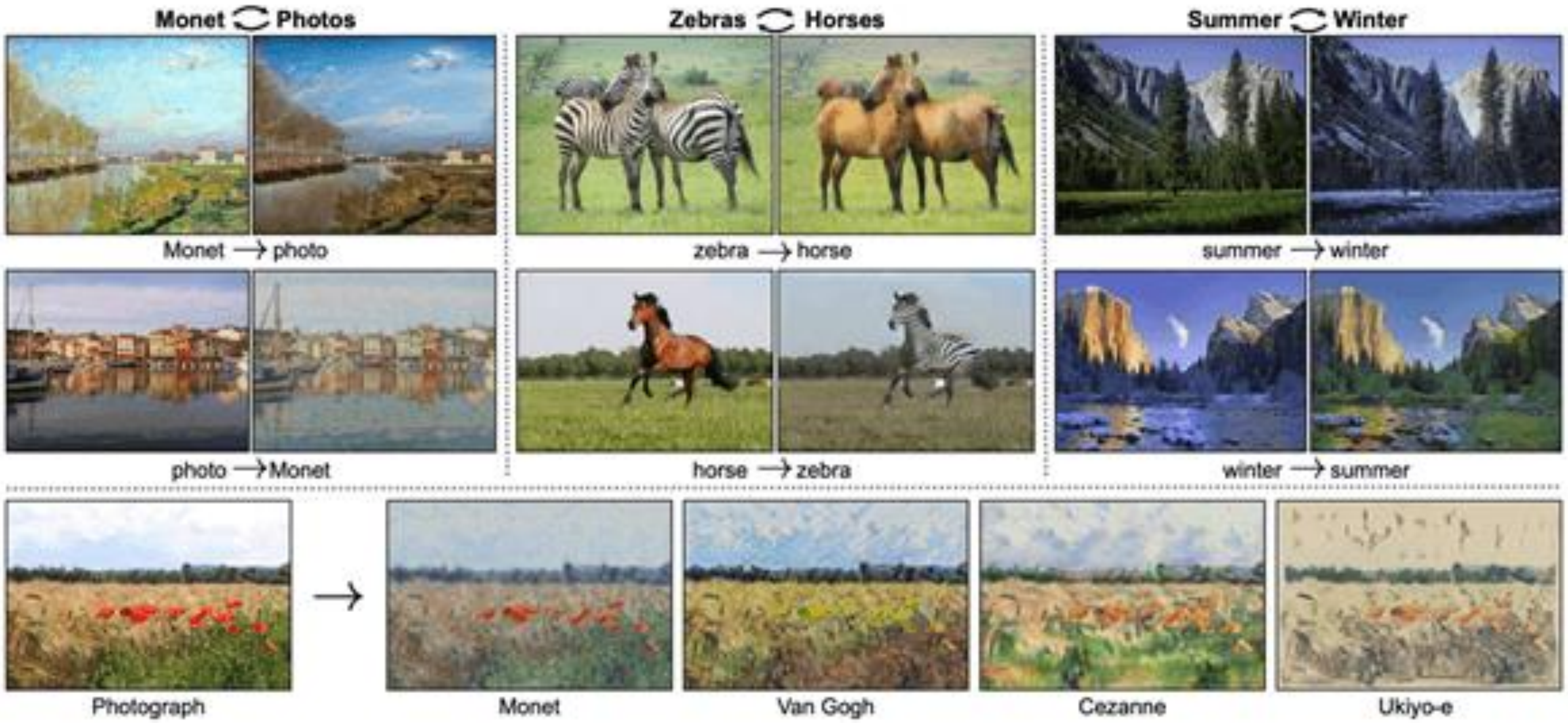


Image credits: https://medium.com/imagescv/what-is-cyclegan-and-how-to-use-it-2bfc772e6195

Example: StyleGAN

SINTEF

Example: text-2-image

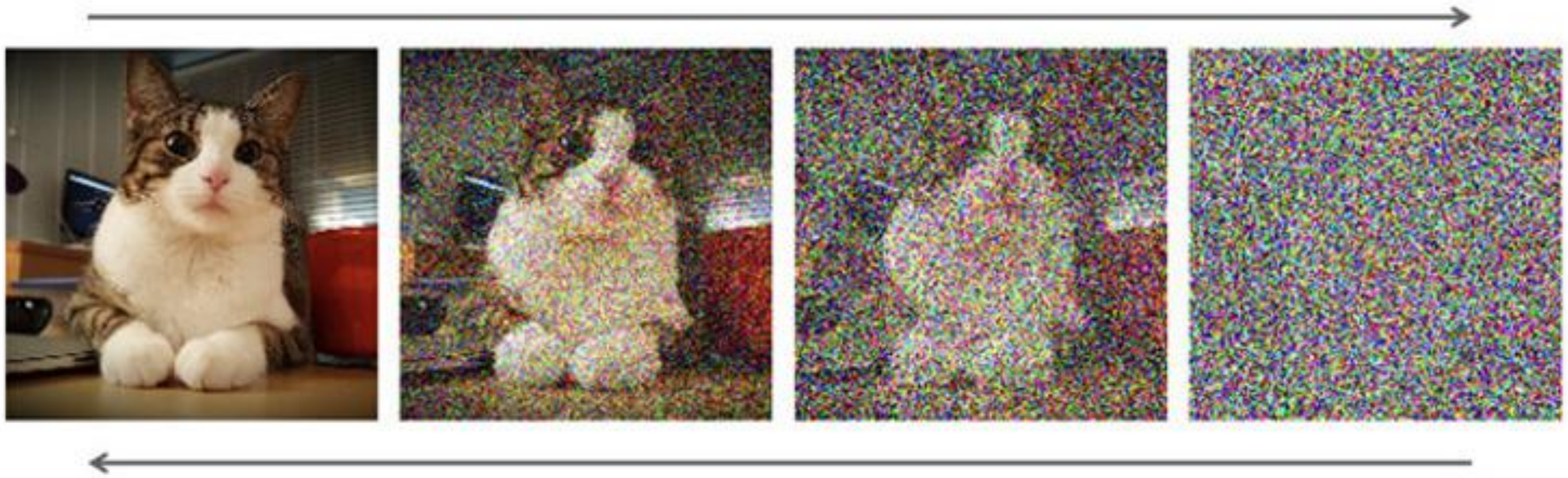
this flower has a lot of small round pink petals

GAN

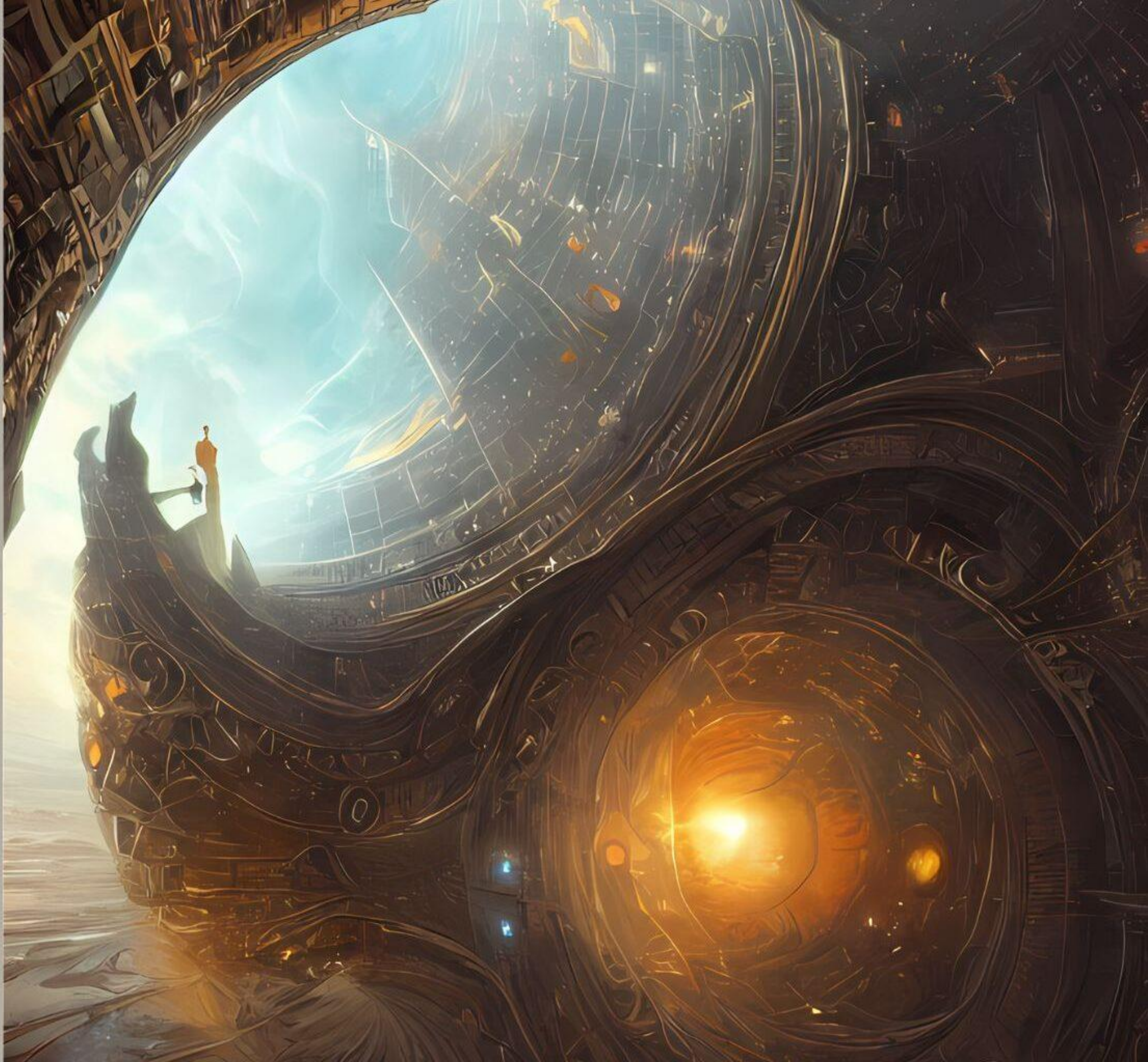
SINTEF

Stable diffusion

- ”Self-supervised”, “weakly supervised” if with text
- Pairing between text and image is weakly defined and naturally occuring



SINTEF

Links

https://stablediffusionweb.com/

https://www.youtube.com/watch?v=kSLJriaOumA&t=66s&ab_channel=TeroKarrasFI

https://deepnote.com/workspace/signe-riemer-sorensen-1db03745-dff4-4089-bda3-d4c62598d5e0/project/0cea47dd-126c-401e-9faa-acb08ae1a21f/notebook/DigitalAcademyIntroductionML-unsupervised-app-fb12dbe4b1744ea7bfd960c3a125f578

SINTEF

Teknologi for et bedre samfunn