Advanced machine learning

MTH793P 2024

Omer Bobrowski, QMUL

UNSUPERVISED LEARNING

Unsupervised Learning

So far, all our machine learning problems were based on training data pairs

$$S := \left\{(x_i, y_i) \text{ iid } \sim \mathscr{D}\right\}_{i=1}^{s}$$

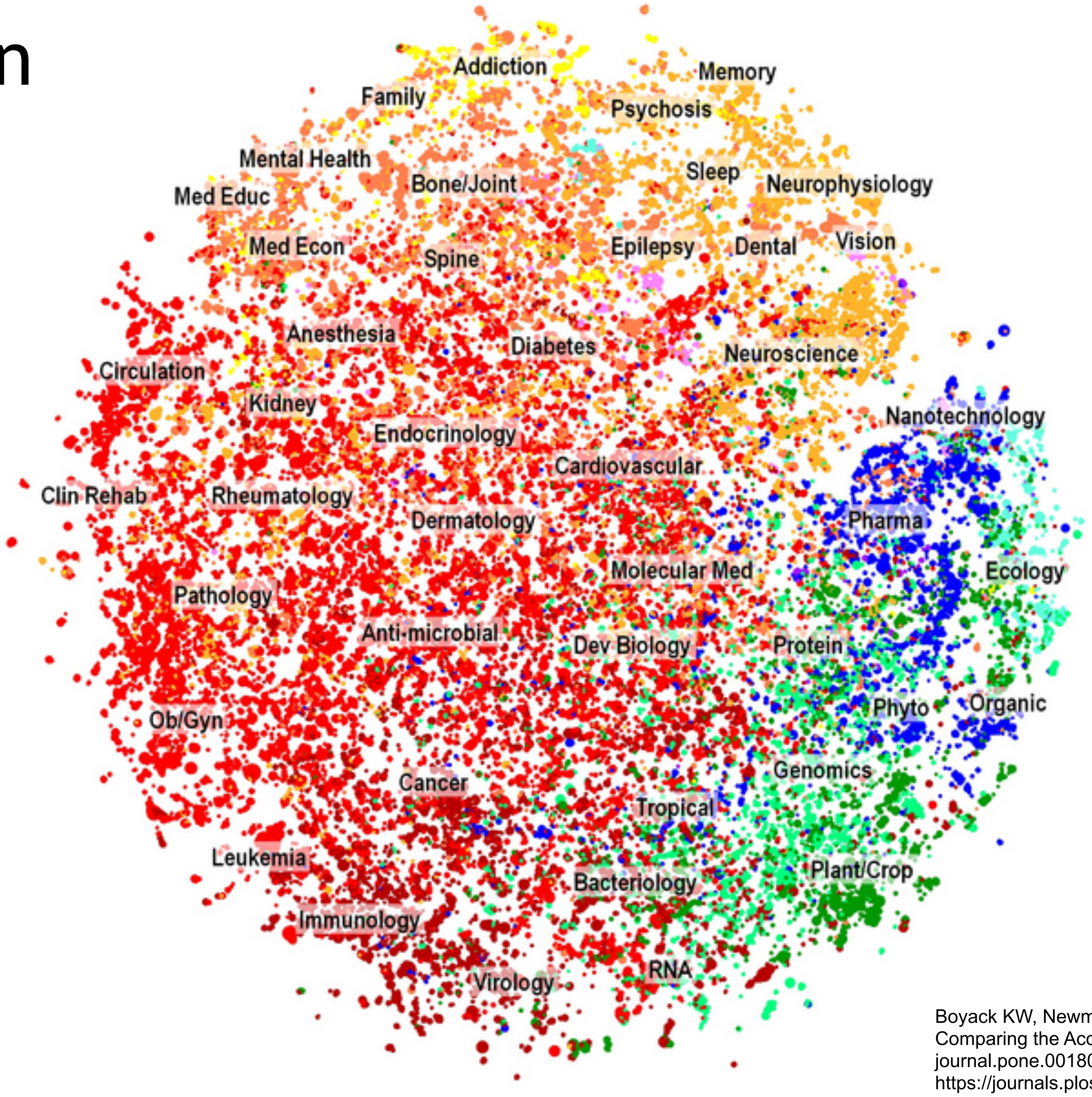
When given pairs of input **and** output data, we speak of *supervised learning*

If we are only given input data $\{x_i\}_{i=1}^{s}$, we speak of *unsupervised learning*

CLUSTERING

Unsupervised Learning

Clustering of two million biomedical publications



Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, et al. (2011) Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLOS ONE 6(3): e18029. https://doi.org/10.1371/journal.pone.0018029
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018029

Clustering

What is clustering?

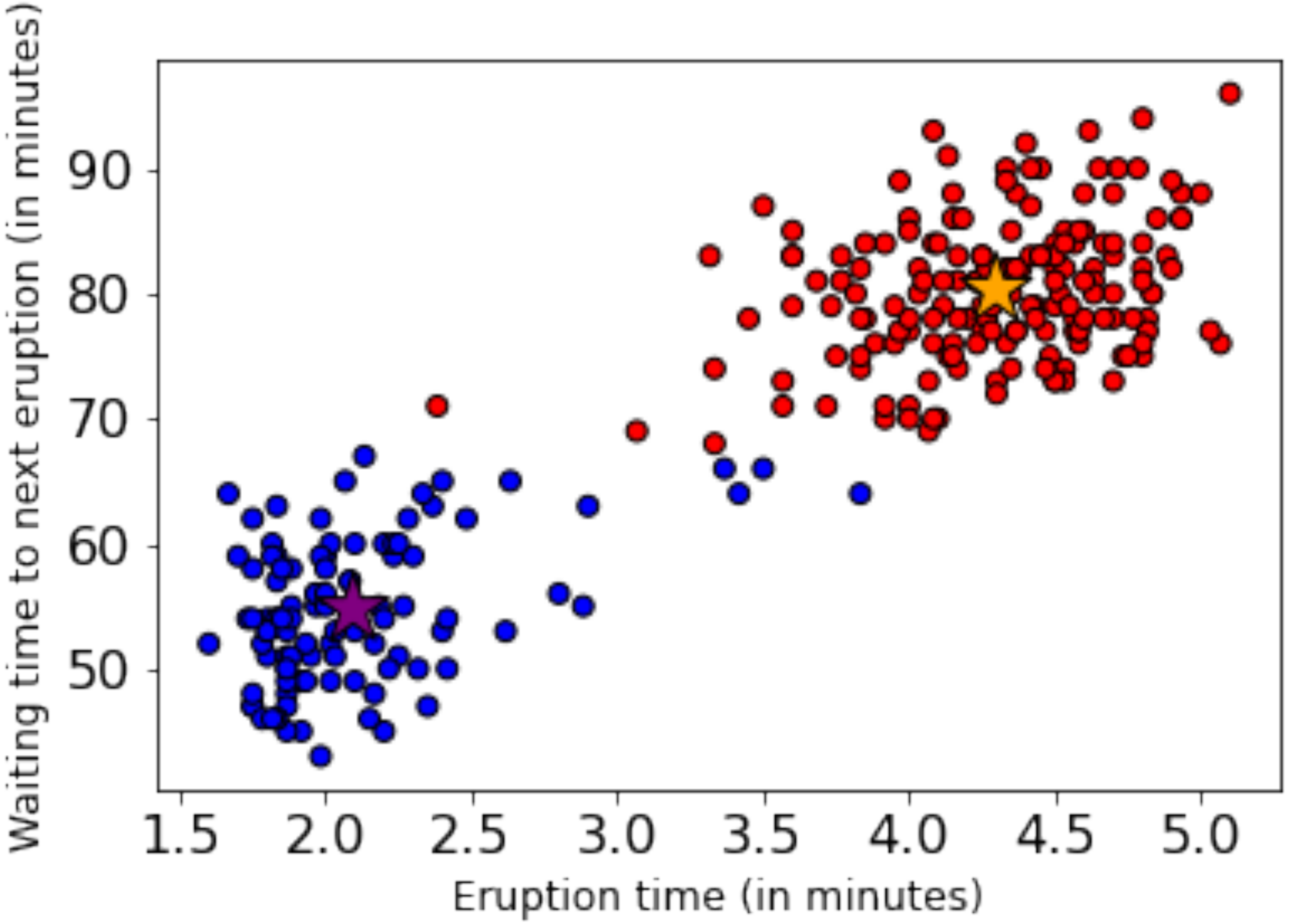
- The organisation of unlabelled data into groups with data that are similar; those groups are called *clusters*

- Each cluster is a collection of data (points) that are **similar** to the other data points in the same cluster, and **dissimilar** to the data points in the other clusters

Clustering

What is clustering?

Downloaded from National Park Service Photo Gallery



Old Faithful dataset

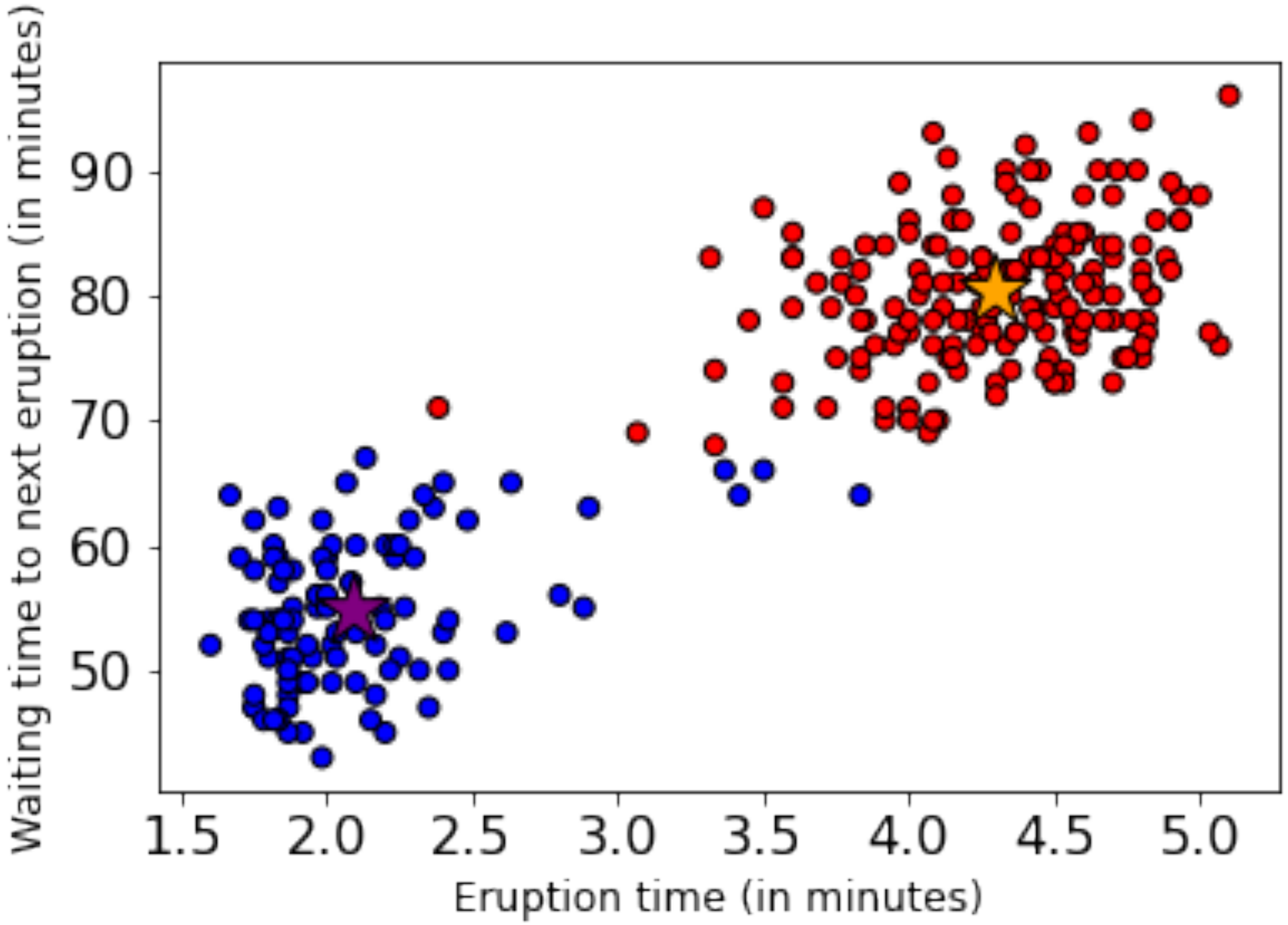
Clustering

Uses of Clustering:

1. Deduce about an individual given information about others in the same cluster (e.g., recommendations)

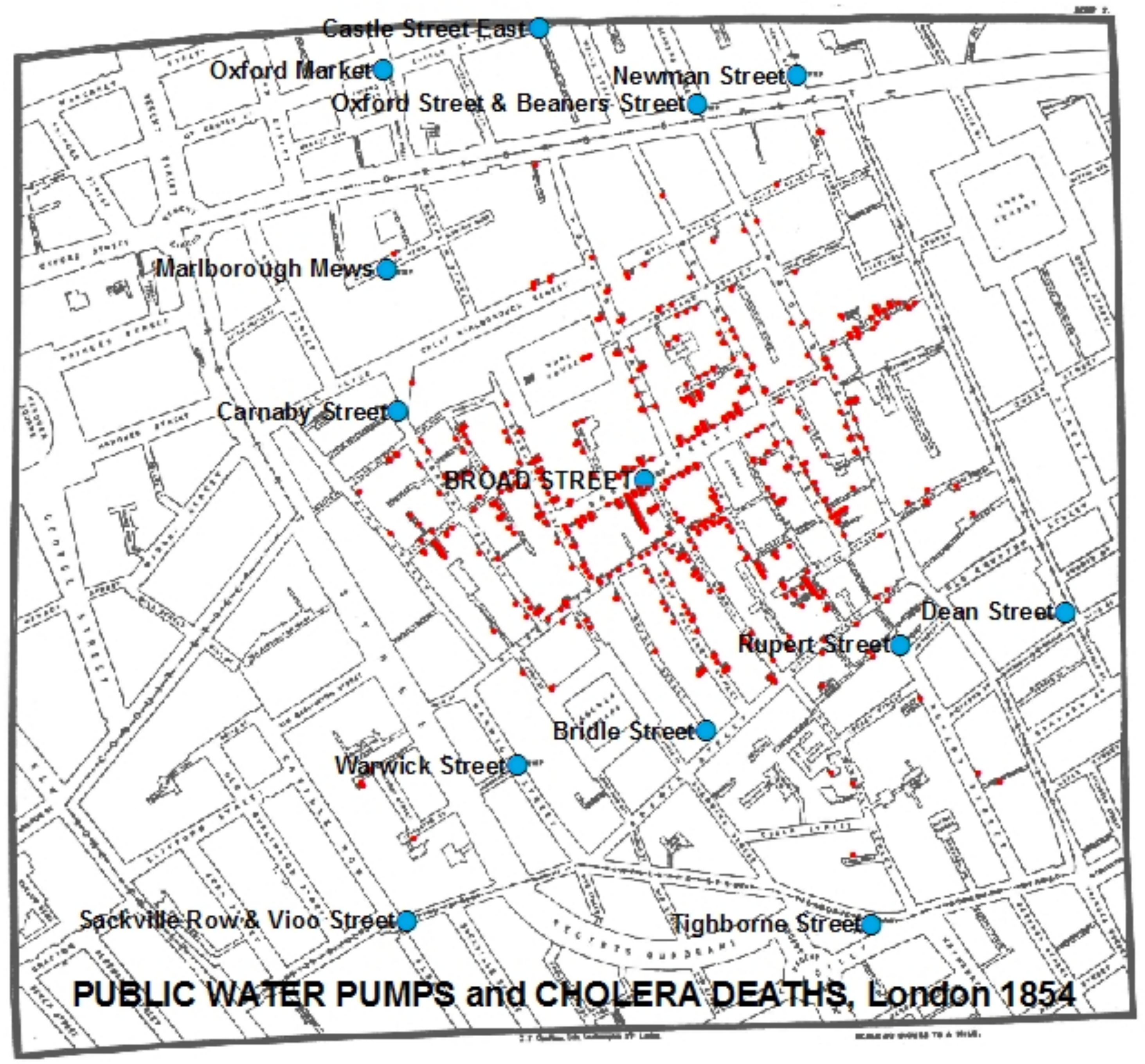
2. Apply different actions to different clusters (e.g., medical treatment)

3. Data compression (lossy)

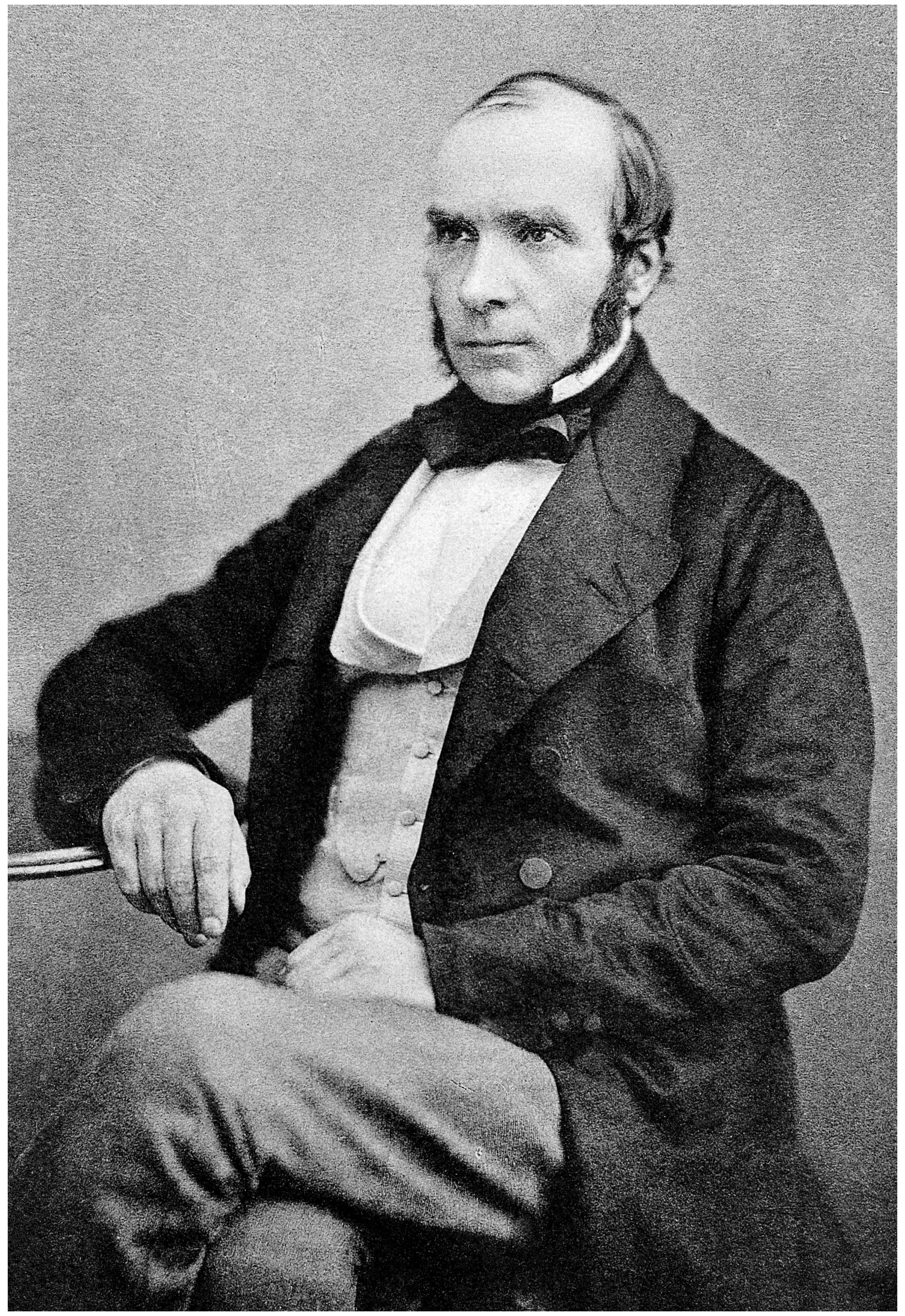


Old Faithful dataset

X



1854 Broad Street cholera outbreak



John Snow. ©Wikimedia commons

©Wikimedia commons

Clustering

What do we need for clustering?

- Proximity measure (either similarity or or dissimilarity measure)
- Clustering problem formulation (e.g. optimisation problem)
- Algorithm to solve the optimisation (and clustering) problem
- Criterion to evaluate a clustering

Clustering

What do we need for clustering?

Proximity measure (either similarity or or dissimilarity measure)

Examples

$$d(x,y) = \|x-y\| = \sqrt{\sum_{i=1}^{s} |x_i - y_i|^2}$$

Euclidean distance

$$d(x,y) = \|x-y\|_1 = \sum_{i=1}^{s} |x_i - y_i|$$

Manhattan distance

$$d(x,y) = \|x-y\|_p = \sqrt[p]{\sum_{i=1}^{s} |x_i - y_i|^p}$$

Minkowski distance

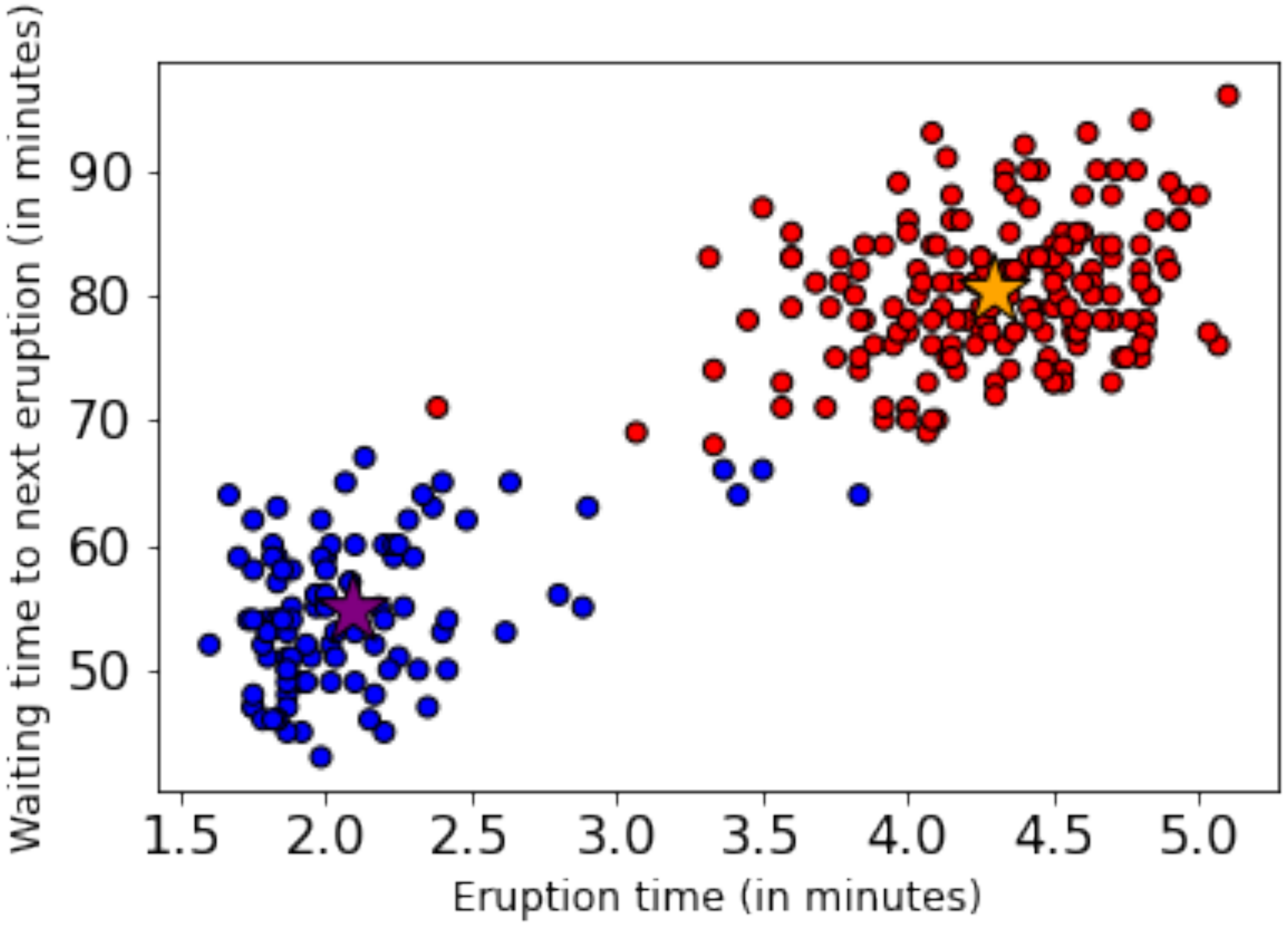
Clustering

What do we need for clustering?

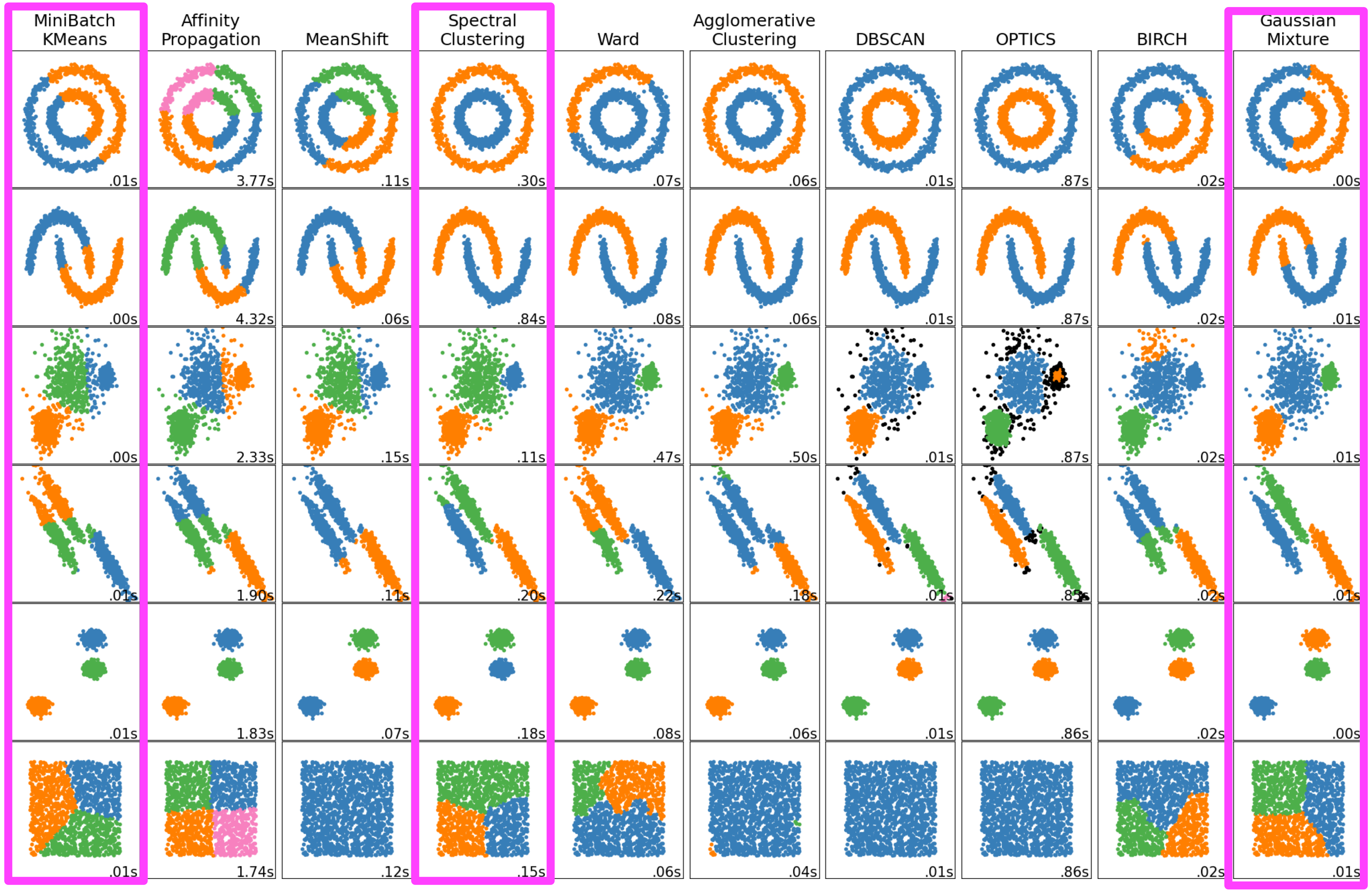
Criterion to evaluate a clustering:

1. Intra-cluster cohesion (compactness)

2. Inter-cluster separation (isolation)



Various Clustering Algorithms



Taken from scikit-learn python package documentation