INTRODUCTION TO MACHINE LEARNING – PART II

SCHEDULE

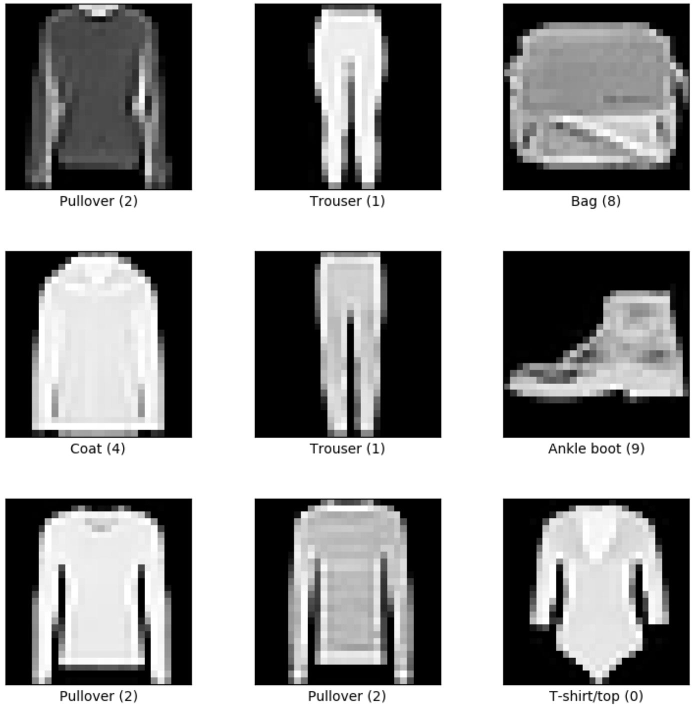
Time	Activity
10:00 – 10:40	Recap from last time, examples of ML-driven research in the humanities. Reflection exercise: Use of ML in your own field and potential ML tasks.
10:40 – 11:00	Introduction to unsupervised machine learning
11:00-11:15	Break
11:15 – 12:00	Exercise 3: Topic Discovery
12:00 – 12:30	Lunch break
12:30 – 13:30	Exercise 3 cont.
13:30 – 14:00	Reflection on exercises, discussion, perspectives

RECAP FROM LAST TIME

TRADITIONAL ALGORITHMS vs MACHINE LEARNING

Rule based tasks

Stable relations between input data and output labels



WHAT IS A GOOD ML PROBLEM?

A ”good” ML problem has:

- A stable and predictable relation between input and output data.
- Strong and well-curated data that exemplify the problem with as much breadth as possible.
- Concrete, "objective" and measurable criteria of success for the training algorithm.
- A way to test the model in realistic situations.
- <u>No better alternative solutions.</u>

WHAT IS MACHINE LEARNING?

Definition: *A computer program is said to* ***learn*** *from experience* E *with respect to some class of tasks* T *and performance measure* P*, if its performance at tasks in* T*, as measured by* P*, improves with experience* E*.*

Tom M. Mitchell (1997)

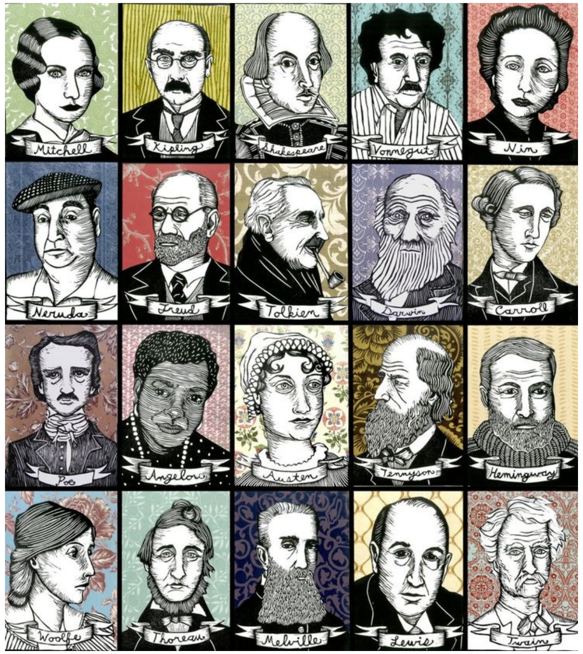
Translation: Machine Learning creates programs, that *measurably improve* at a *concrete task* given *increased experience*.

Mitchell, T. M. (1997). Machine learning. *Burr Ridge, IL: McGraw Hill*, *45*(37), 870-877.

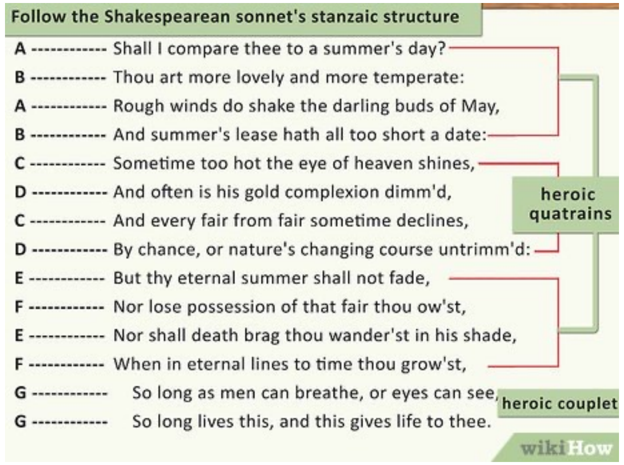
DISCUSSION – WHAT IS THE DATA AND WHAT ARE THE LABELS? IS IT A GOOD ML PROBLEM?

Archaeology:
Classification of pottery and stone tools

Literary studies:
Authorship attribution



Poetics
Identification of sonnets



CONFUSION MATRIX

Is an accuracy of 99% always good?

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

UNSUPERVISED LEARNING

WHAT DOES SUPERVISION MEAN AND WHAT IF IT ISN'T THERE?

Recap: Supervised Learning (Predictive Modeling)

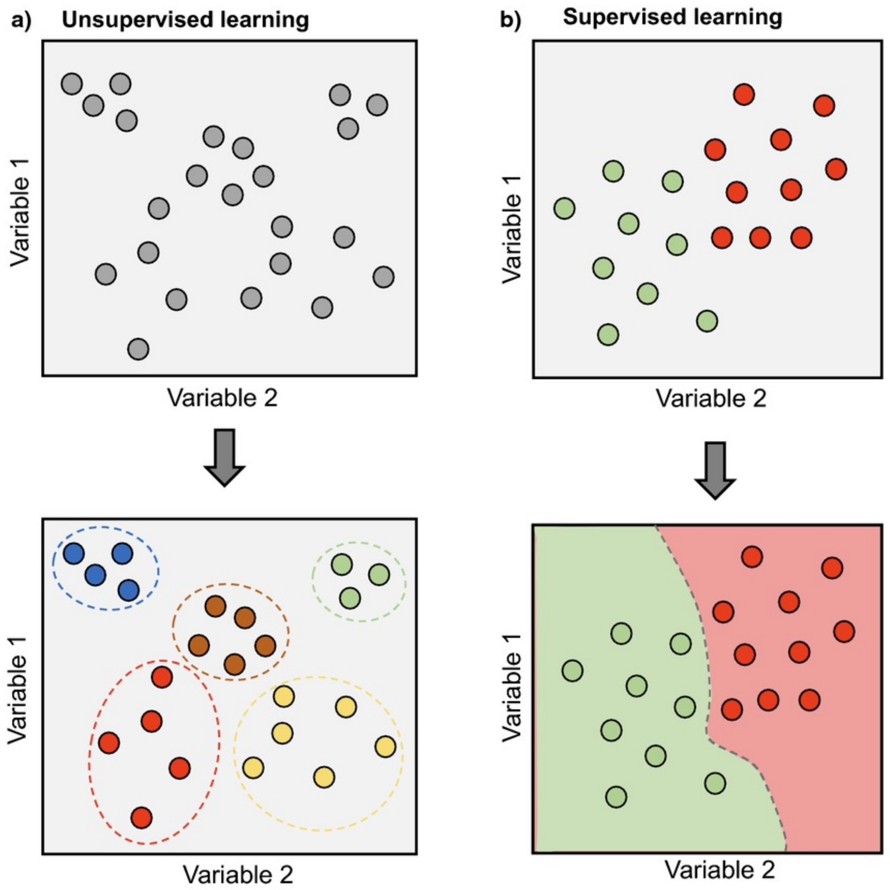
- Data comes as pairs: Input (X) and Target (Y).
- Goal: Map X→Y by minimizing error.
- Success is directly measurable: Did we predict the correct label?

Unsupervised Learning (Data Exploration)

- Data is just Input (X). There is no Y.
- **The Challenge:** Without a label to guide us, the algorithm doesn't know what is "right."
- **The Goal:** Find "interesting structure," "patterns," or "simplified representations" inherent in the data geometry.

The Role of Inductive Bias:

- Since we lack a ground truth, the **inductive bias** is even more critical. We must choose an algorithm whose assumptions (biases) match our understanding of the data.
- *We* define what "similarity" means; the algorithm just optimizes for it.



PARTITION-BASED CLUSTERING

The Algorithm: K-Means

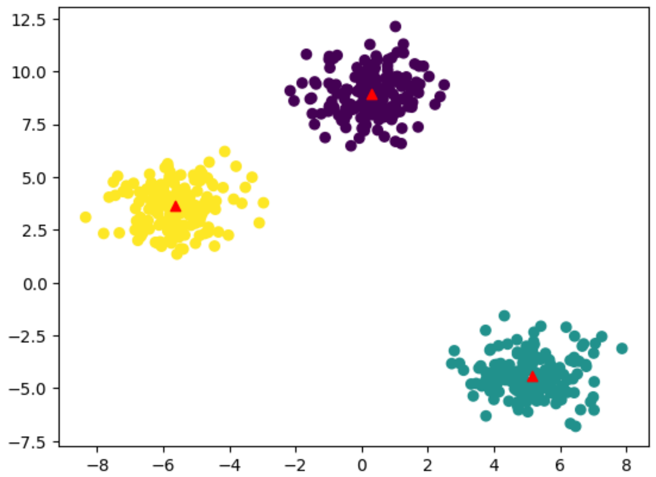
- Iteratively moves k center-points ("centroids") to the average position of their nearest neighbors.

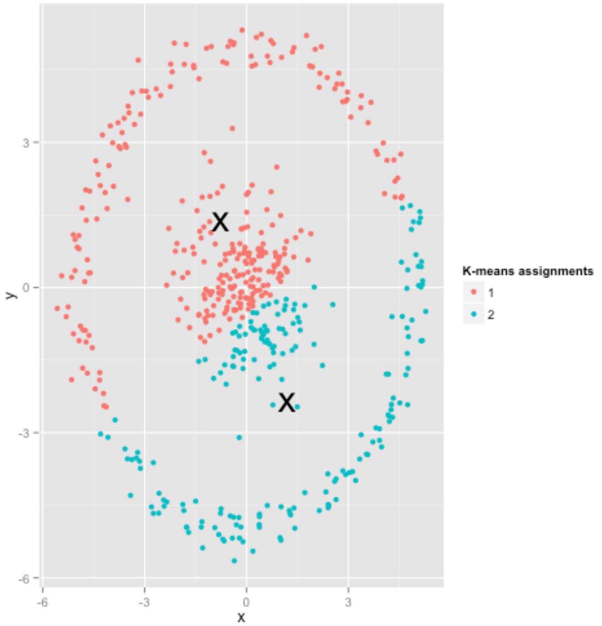
Inductive Bias (Assumptions):

- **Spherical Clusters:** Assumes clusters are round blobs.
- **Equal Variance:** Assumes clusters are roughly the same size and density.
- **Hard Assignment:** Every data point belongs 100% to exactly one cluster.

Implications:

- Efficient and simple, but fails if data is shaped like "moons," "rings," or is elongated.
- It forces the world into Voronoi cells—it partitions space rather than finding connected densities.





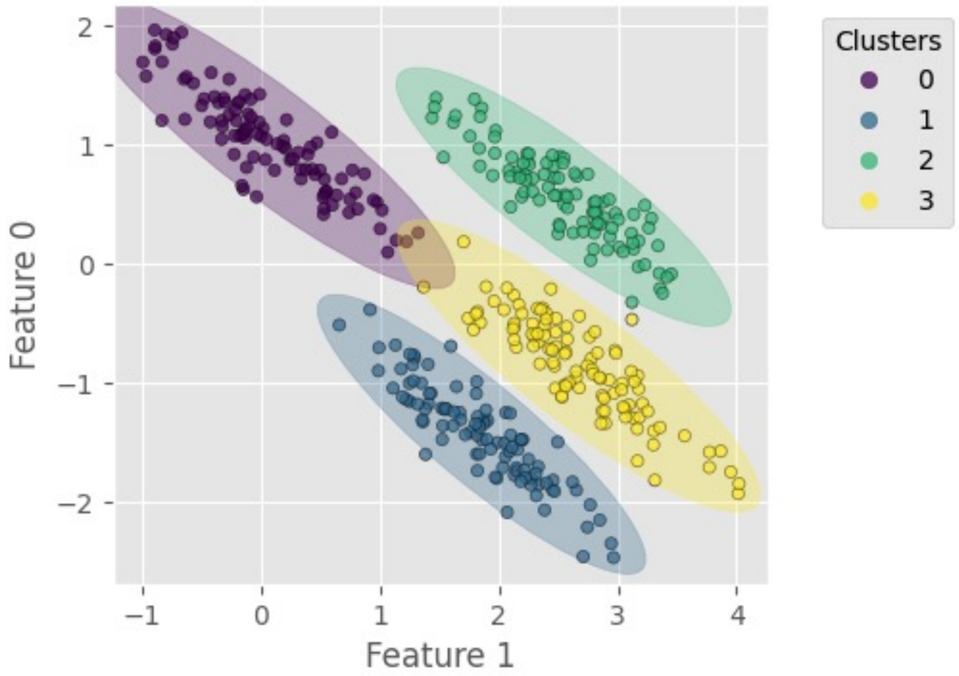
PROBABILISTIC CLUSTERING

The Algorithm: Gaussian Mixture Models (GMM)

- Instead of hard boundaries, we model data as a mixture of statistical distributions.
- **Soft Clustering:** A point isn't just "Cluster A." It might be "80% Cluster A, 20% Cluster B." Captures uncertainty/ambiguity.

The Inductive Bias:

- **Elliptical Shapes:** Unlike K-Means (circles only), GMMs can stretch to fit elongated data (covariance).
- **Gaussian Assumption:** Assumes the data was generated by combining Normal distributions.



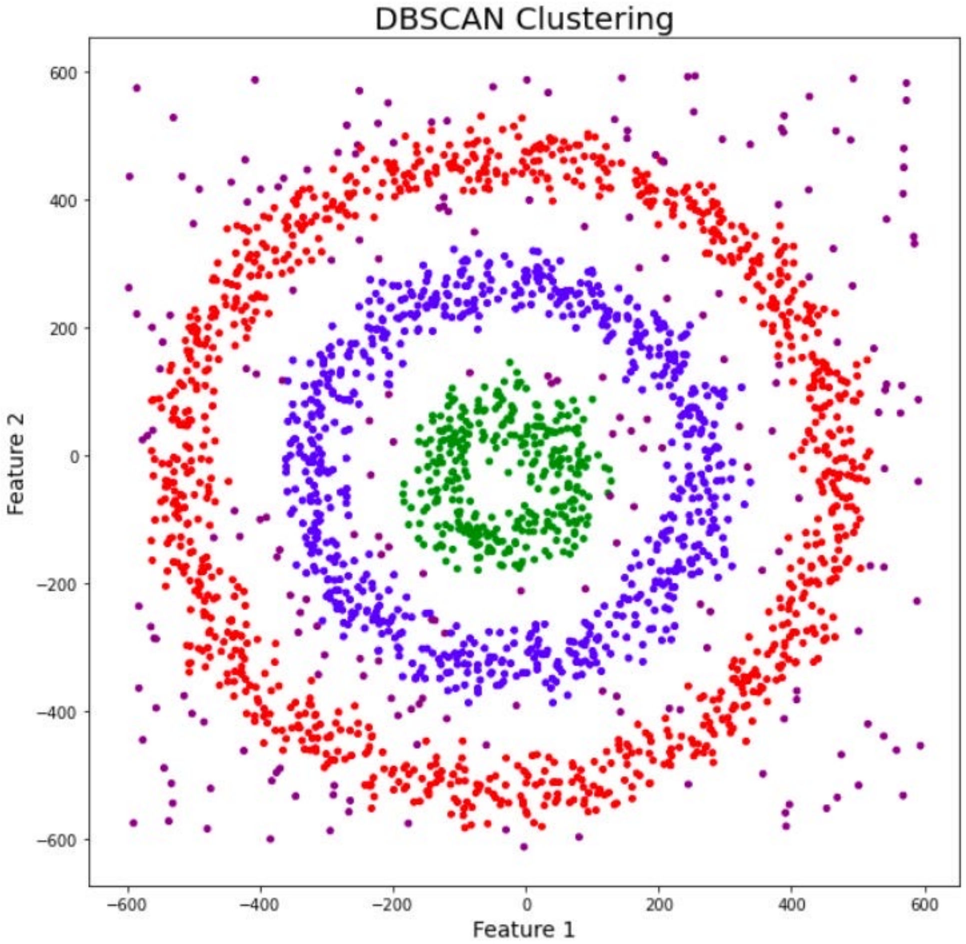
DENSITY-BASED CLUSTERING

The Algorithm: DBSCAN / HDBSCAN

- Clusters are "areas of high density separated by areas of low density."
- **Crucial difference:** You do not need to choose k (number of clusters) in advance. The data dictates the number of clusters.

The Inductive Bias:

- **Connectedness:** Things that are packed close together belong together, regardless of the overall shape.
- **Noise Handling:** Assumes that data in low-density regions is "noise" or "background" and should not be clustered.



OUTLIERS

What is an Outlier?

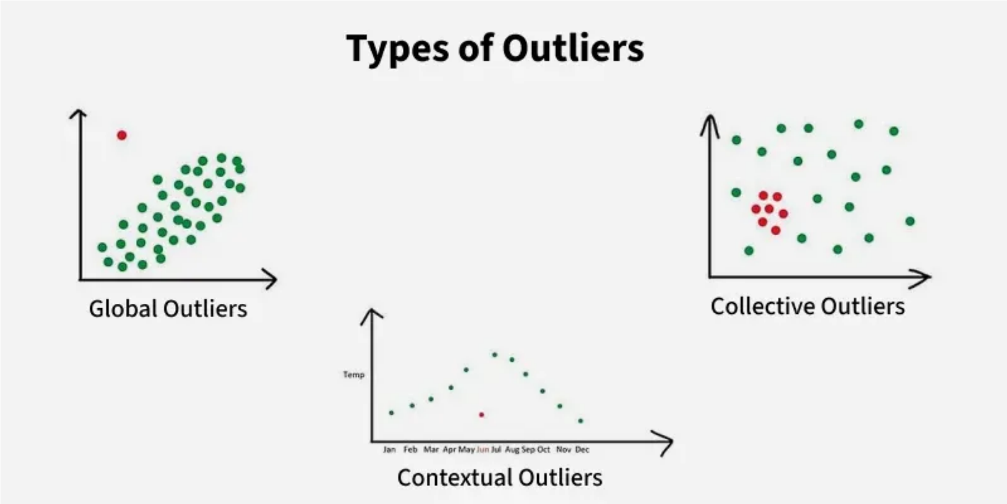
- A data point that deviates significantly from the underlying structure of the data.

Two Perspectives:

- **Nuisance:** Noise/Errors (e.g., sensor glitches, bad OCR in text) that confuse the model.
- **The Goal:** The "needle in the haystack" (e.g., credit card fraud, rare disease, scientific anomaly).

How ML defines "Normal" (Bias):

- **Distance:** "You are too far from the centroid" (K-Means).
- **Density:** "You are in a low-density region" (DBSCAN).
- **Reconstruction:** "I cannot compress and reproduce you accurately" (Autoencoders).



DIMENSIONALITY REDUCTION & VISUALIZATION

The Problem:

- We cannot visualize 768-dimensional text embeddings or 1024-pixel image vectors.
- How do we see the "structure" we are looking for?

The Algorithms:

- **PCA (Principal Component Analysis):** Preserves global variance. Good for simple structure. Linearly projects data to flat surfaces.
- **t-SNE / UMAP:** Preserves local neighborhoods. Good for complex structure. "Unfolds" the data manifold to keep similar points close.