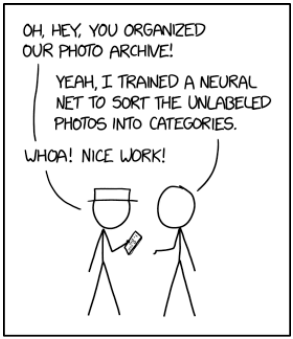
Classification & Data Decomposition

Mini-Lecture: Sorting, Grouping, and Compressing Data

Methods and Algorithms

MSc Data Science

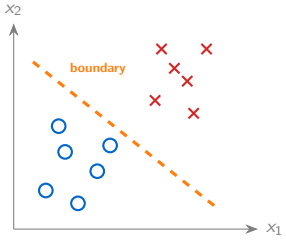
The Neural Net Knows Best



XKCD #2173 by Randall Munroe (CC BY-NC 2.5)

Classification: Assigning Labels

- Predict a **discrete class**: $y \in \{0, 1, \ldots, K\}$
- Binary: default / no-default; Multiclass: sector label
- A **decision boundary** separates classes in feature space
- Full treatment in L02 (logistic regression) and L04 (random forests)



Classification is the most common ML task in banking — credit scoring, fraud, compliance.

How Classifiers Decide

- **Logistic regression**: estimates $P(y{=}1 \mid \mathbf{X})$, threshold at 0.5 (L02)
- **KNN**: majority vote among K nearest neighbours (L03)
- **Decision tree**: recursive if-then splits on features (L04)
- Each algorithm creates a *different* decision boundary shape

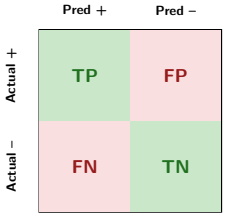
Choosing a Classifier

No single method dominates all problems — the "best" classifier depends on data size, feature types, and interpretability requirements.

This is definition-level orientation — full formulas and derivations come in L02–L04.

Evaluating Classifiers

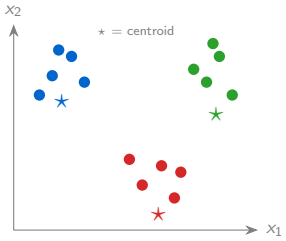
- **Accuracy** can be misleading: 99% on 1% fraud = useless model
- **Confusion matrix**: TP, FP, TN, FN
- Precision = $\frac{TP}{TP+FP}$ (of predicted positives, how many correct?)
- Recall = $\frac{TP}{TP+FN}$ (of actual positives, how many found?)



In fraud detection, recall matters most — missing a fraud case is far costlier than a false alarm.

Clustering: Grouping Without Labels

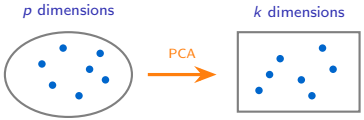
- Partition n observations into K groups based on **similarity**
- **K-Means**: assign to nearest centroid, update centroids, repeat (definition only — full algorithm in L03)
- Evaluate with **inertia** (within-cluster sum of squares) and **silhouette score**
- Finance: segment retail banking customers by spending behaviour



K-Means is the most widely used clustering algorithm — simple, fast, and surprisingly effective.

Dimensionality Reduction

- High-dimensional data ($p \gg 3$) is hard to visualize and model
- **PCA**: project onto directions of maximum variance (definition only — full derivation in L05)
- **t-SNE**: preserve local neighbourhoods in 2D (L05)
- Goal: reduce p features to $k \ll p$ while retaining information



PCA is the workhorse of dimensionality reduction — it connects directly to eigenvalues from P01.

The Decomposition Perspective

- **Decomposition** = breaking a complex signal into simpler components
- **SVD**: $\mathbf{X} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (truncated to rank k)
- Factor analysis: discover **latent factors** driving observed correlations
- Finance: stock return = market factor + sector factor + idiosyncratic noise

$$\underbrace{\mathbf{X}}_{n\times p} \approx \underbrace{\mathbf{U}}_{n\times k}\ \underbrace{\mathbf{\Sigma}}_{k\times k}\ \underbrace{\mathbf{V}^\top}_{k\times p} \qquad (k \ll p)$$

Why It Matters

PCA is a special case of SVD applied to centred data — L05 builds on this foundation.

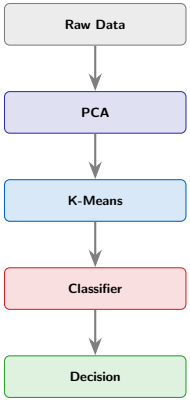
Decomposition unifies PCA, factor models, and matrix approximation under one framework.

Finance: Putting It All Together

- **Classification**: credit scoring (L02), fraud detection (L04)
- **Clustering**: customer segmentation, market regime detection (L03)
- **Decomposition**: PCA on 50 stock returns → 3 risk factors (L05)

Combined Workflow

Decompose → Cluster → Classify → Decide



This pipeline appears throughout L01–L06 — each lecture fills in one piece of the puzzle.

Summary: Classification, Clustering, Decomposition

1. **Classification** assigns discrete labels — evaluate with precision and recall, not just accuracy
2. **Clustering** (K-Means) groups data without labels — full algorithm in L03
3. **Dimensionality reduction** (PCA, t-SNE) compresses features — derivation in L05
4. Finance **combines all three**: decompose → cluster → classify → decide

You Are Ready

With P01–P03 complete, you have the vocabulary and intuition for L01–L06.

These three concepts — classify, cluster, decompose — are the verbs of applied machine learning.