Supervised & Unsupervised Learning

Mini-Lecture: Two Paradigms of Machine Learning

Methods and Algorithms

MSc Data Science

When Is a Task “Easy”?

XKCD #1425 by Randall Munroe (CC BY-NC 2.5)

What Is Machine Learning?

- Learning patterns **from data**, not from explicit programming
- Three paradigms: **supervised**, **unsupervised**, reinforcement
- This course: supervised (L01–L04), unsupervised (L03, L05)
- Finance: predict defaults *vs.* segment customers

ML Paradigms

Supervised
(L01–L04)

Unsupervised
(L03, L05)

Reinforcement
(L06 intro)

Knowing which paradigm a problem belongs to is the first step in choosing an algorithm.

Supervised Learning: Learn from Labels

- Training data consists of (X, y) **pairs** — features and a known target
- Goal: learn a function $f : X \to y$ that **generalizes** to unseen data
- **Regression**: y is continuous (stock return, house price)
- **Classification**: y is discrete (default/no-default, sector label)

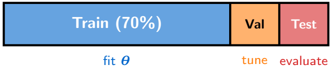
Core Equation

$$\hat{y} = f(\mathbf{x}; \boldsymbol{\theta}) + \varepsilon \qquad \text{where } \boldsymbol{\theta} \text{ is learned from data}$$

"Supervised" = the labels y supervise (guide) the learning process.

The ML Workflow: Train / Validate / Test

- Split data: **Train 70%** / Validation 15% / **Test 15%**
- Train = fit model parameters; Validate = tune hyperparameters; Test = final evaluation
- **Never** use test data during training — this is *data leakage*



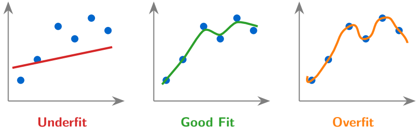
Golden Rule

The test set is opened exactly **once** — at the very end.

Data leakage is the most common source of over-optimistic results in finance ML papers.

Bias–Variance Trade-off

- **Underfitting** (high bias): model too simple, misses patterns
- **Overfitting** (high variance): model memorizes noise, fails on new data
- The **sweet spot** minimizes total error $= \text{Bias}^2 + \text{Variance}$
- Stock prediction: fitting 50 parameters to 60 data points $\Rightarrow$ overfitting



Every model selection decision in this course is a bias–variance trade-off in disguise.

Unsupervised Learning: Find Hidden Structure

- No labels — only feature matrix $\mathbf{X}$; the algorithm discovers **structure**
- **Clustering**: group similar observations (K-Means, hierarchical)
- **Dimensionality reduction**: compress p features to $k \ll p$ (PCA, t-SNE)
- Finance applications: segment retail customers, reduce 50 stock returns to 3 latent factors

Key Difference from Supervised

No "right answer" to evaluate against — success is measured by *coherence* (inertia, silhouette score) not prediction accuracy.

Unsupervised learning often serves as a preprocessing step before supervised models.

Key Terminology

Term	**Definition**
Features ($\mathbf{X}$)	Input variables (predictors, covariates, independent variables)
Target (y)	Output variable to predict (response, dependent variable)
Hyperparameter	Setting chosen *before* training (e.g. learning rate, K in KNN)
Cross-validation	Rotate train/val split k times for robust performance estimate
Metric	Quantitative measure of model quality (MSE, accuracy, AUC)

These five terms appear in every lecture — make sure you can define each from memory.

Finance: Supervised vs. Unsupervised

Supervised

- Credit scoring
- Return forecasting
- Fraud detection

Unsupervised

- Customer segmentation
- PCA risk factors
- Anomaly detection

In Practice

Most production systems **combine both**: cluster customers (unsupervised), then build a classifier per segment (supervised).

Real-world ML pipelines rarely use a single paradigm — hybrid approaches dominate.

Summary: Two Paradigms

1. **Supervised learning** requires labeled data (X, y) and predicts outcomes
2. **Unsupervised learning** finds structure in unlabeled data X
3. The ML workflow (**train / validate / test**) prevents overfitting and data leakage
4. Finance uses **both paradigms**: predict defaults, segment customers, reduce dimensions

Coming Up

P03: Classification & Data Decomposition — sorting, grouping, and compressing data.

Every algorithm in L01–L06 falls into one of these two paradigms — always ask "supervised or unsupervised?"