

BIOS226 - Topic 5 - Supervised Learning (Part 1)

Student Companion Guide

Foundations for Biological Model Evaluation

This companion follows the Part 1 lecture titles and explains each idea in plain language. Use it alongside the slides to build confidence before the Part 2 pipeline session.

From Exploration to Prediction

In exploratory analysis, we ask what structure is in the data (for example PCA and clustering). In supervised learning, we ask a different question: can we predict an outcome for a new patient from measured features?

- Exploration: pattern finding without outcome labels.
- Prediction: learning from labelled examples to support decisions.

In breast cancer, this move from description to prediction has had direct treatment impact (Paik et al., 2004; Sparano et al., 2018).

Aside: Oncotype DX in Plain Language

Oncotype DX is a lab test that measures expression of a small group of genes in a breast tumour sample.

- It gives a recurrence score to estimate how likely the cancer is to return.
- Clinicians use that score, with other clinical information, to discuss whether chemotherapy is likely to help.
- For many patients, this supports avoiding chemotherapy when expected benefit is low.

The evidence base developed in stages: early validation work (Paik et al., 2004), then large prospective studies refining who benefits from chemotherapy (Sparano et al., 2018; Kalinsky et al., 2021).

At global scale, the test has been used in more than 2 million patients across over 100 countries over about 22 years, according to company reporting, with an estimated 1.6 million people avoiding potentially unnecessary chemotherapy (Exact Sciences, 2026).

Unsupervised vs Supervised

Unsupervised methods group samples by similarity. Supervised methods learn relationships between features and known outcomes.

This distinction matters in cancer research: molecular classes can change how disease is defined and treated (Perou et al., 2000; Hoadley et al., 2018).

What Are X and Y?

Supervised learning uses:

- X : input features (for example SNPs, gene expression, clinical variables).
- Y : target outcome (for example disease class, risk, or response).

Polygenic risk scores are a good example: many small SNP effects can add up to meaningful risk differences (Khera et al., 2018).

Classification vs Regression

- Classification predicts categories (for example subtype A vs subtype B).
- Regression predicts numeric values (for example a response score).

In practice, teams often model probabilities and then choose thresholds for clinical decision-making.

What Is a Model?

A model is a mathematical mapping from X to Y . During training, it learns patterns from observed data. During validation, we test whether those patterns generalise to unseen data.

The key point for biology: a useful model is not just accurate on one dataset. It must also be robust across cohorts and measurement settings.

The High-Dimensional Problem ($p \gg n$)

In many biological studies, features (p) greatly outnumber samples (n). This increases the risk of overfitting.

- The model can learn noise instead of true biology.
- Internal performance may look strong but fail in external data.
- Feature signatures may not replicate.

That is why careful validation and leakage control are central to good practice.

Why Biology Is Harder

Biological datasets often include technical variation, batch effects, small sample sizes, and real biological heterogeneity.

So model quality depends on study design and evaluation discipline, not just algorithm choice.

Key Take-Home Messages

- Supervised learning supports decisions, not just predictions.
 - Always separate training logic from final testing logic.
 - In biology, validation design is as important as model selection.
 - Oncotype DX is a concrete example of supervised learning affecting care at global scale.
-

References

- Hoadley, K.A., et al. (2018) 'Cell-of-origin patterns dominate the molecular classification of 10,000 tumours from 33 types of cancer', *Cell*, 173(2), pp. 291–304.e6.
- Kalinsky, K., et al. (2021) '21-Gene assay to inform chemotherapy benefit in node-positive breast cancer', *New England Journal of Medicine*, 385(25), pp. 2336–2347.
- Khera, A.V., et al. (2018) 'Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations', *Nature Genetics*, 50(9), pp. 1219–1224.
- Paik, S., et al. (2004) 'A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer', *New England Journal of Medicine*, 351(27), pp. 2817–2826.

- Perou, C.M., et al. (2000) 'Molecular portraits of human breast tumours', *Nature*, 406(6797), pp. 747–752.
- Sparano, J.A., et al. (2018) 'Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer', *New England Journal of Medicine*, 379(2), pp. 111–121.
- Exact Sciences (2026) *Oncotype DX Breast Recurrence Score Test Surpasses 2 Million Patients Worldwide*. Available at: <https://investor.exactsciences.com/investor-relations/press-releases/press-release-details/2026/Oncotype-DX-Breast-Recurrence-Score-Test-Surpasses-2-Million-Patients-Worldwide/default.aspx> (Accessed 22 February 2026).