Module 5A: Unsupervised Learning

A smattering of options: PCA, permutations, bootstrap

1. Review of Hypothesis testing

- Review χ^2 goodness of fit tests: assumptions, how it works, demonstrate use with any distribution (it is a non-parametric method)
  - χ^2 contingency test: a specialized type of χ^2 goodness of fit test with the Ho: two variables are **independent** (tested using probability multiplication rule)
- Reviewed Z scores:

$$\frac{Signal}{Noise} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

2. Student's t tests:

- Different versions (one sample, paired, two sample), with different assumptions
- Same principle as Z score <u>but replace **σ** with **s**</u>, adds uncertainty, the t-distribution is wider than Z

$$\frac{\boldsymbol{Signal}}{\boldsymbol{Noise}} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- Nonparametric versions that can be used when assumptions are not met
- Welch's approximate t test (still parametric – assumes normal distribution) when variances between the two populations are wildly different; adjusts degrees of freedom downward

3. ANOVA & Correlation

4. Regression

1. Review of Hypothesis testing

2. Student's t tests

3. ANOVA & Correlation

- Extension of the two-sample t test when >2 populations are compared
- Allows accurate α (no inflation)
- How well does the model of "belonging to a particular treatment group" explain the total variation?
- You are allocating variation: **between group variation,** and **within group** (**stochastic) variation**

$$\frac{Signal}{Noise} = \frac{MS_{groups}}{MS_{error}}$$

- Post-hoc testing to determine **WHICH group(s)** has/have significantly different population means (μ)
- (Pearson) Linear Correlation:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{Covariance(X,Y)}{s_x s_y}$$

4. Regression

1. **Review of Hypothesis testing**
2. **Student's t tests**
3. **ANOVA & Correlation**
   - (Pearson) Linear Correlation:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{Co\,\mathrm{var}\,iance(X,Y)}{s_x s_y}$$

4. **Regression**
   - Independent (explanatory) variable impacts dependent (response) variable
   - Many different types that have slightly different assumptions
   - Typical: Homoscedasticity and normally distributed Y around each X_i
   - Note structural similarities between slope and correlation

$$\bar{Y} = a + \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}\bar{X}$$

   - Brief overview of General Linear models to emphasize the similarity: stepwise comparison of the full model to a model with one variable (or interaction) removed to see if there is a statistically significant improvement in fit.

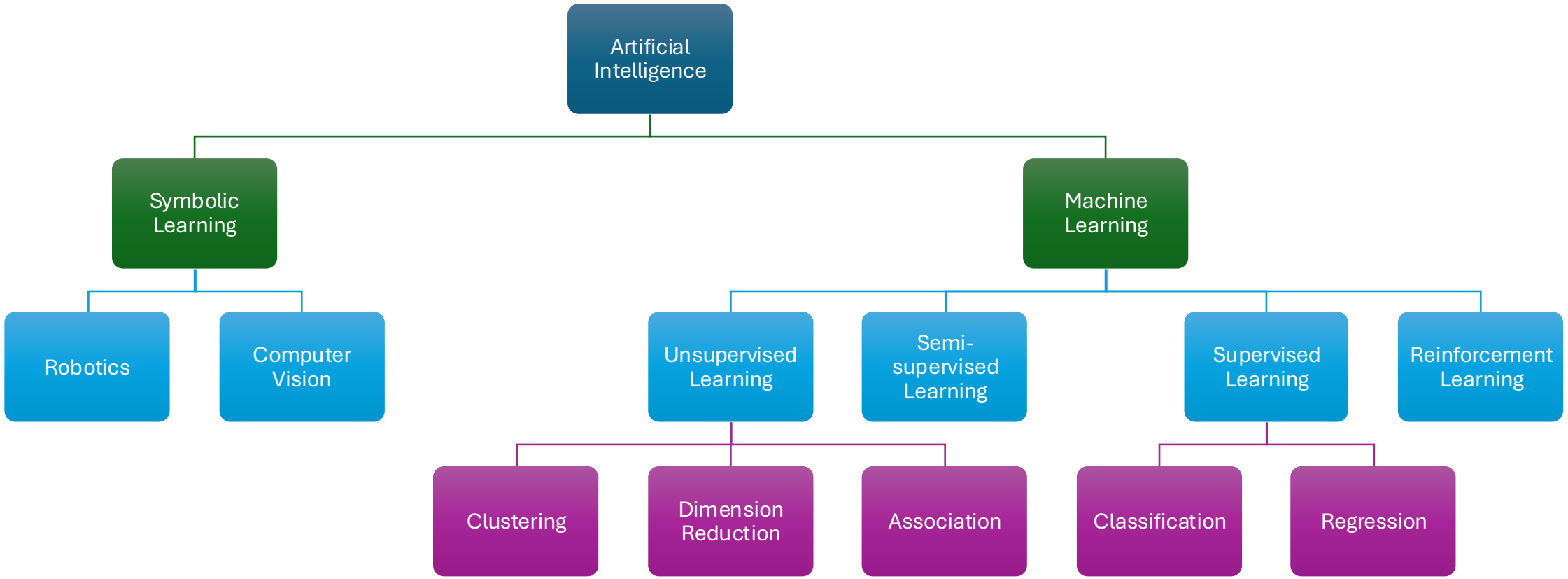
Agenda:

- Sometimes parametric methods are not powerful enough
- What is **unsupervised learning?**

Finding hidden patterns in data without labels

1. Clustering: K-means
   - Uses: candidate miRNA targets, gene expression data
2. Dimension Reduction: PCA, Discriminant Analysis
   - single-cell analysis, gene expression; population structure

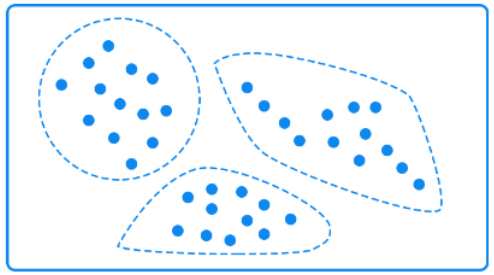
- Survey of computational methods: bootstrapping, permutation, and simulation


Artificial Intelligence
Symbolic Learning
Machine Learning
Robotics
Computer Vision
Unsupervised Learning
Semi-supervised Learning
Supervised Learning
Reinforcement Learning
Clustering
Dimension Reduction
Association
Classification
Regression

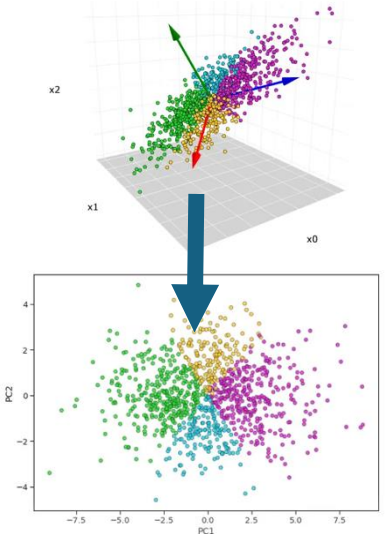
Machine Learning

Unsupervised Learning

Clustering



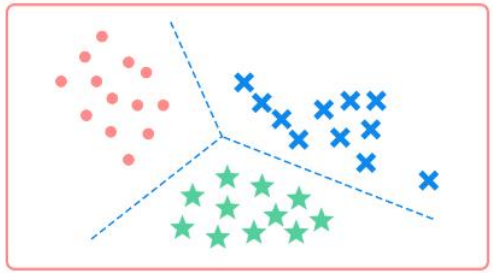
Dimension Reduction

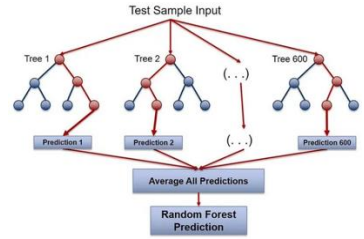


Association

Supervised Learning

Classification





Regression