University of Pittsburgh | School of Public Health

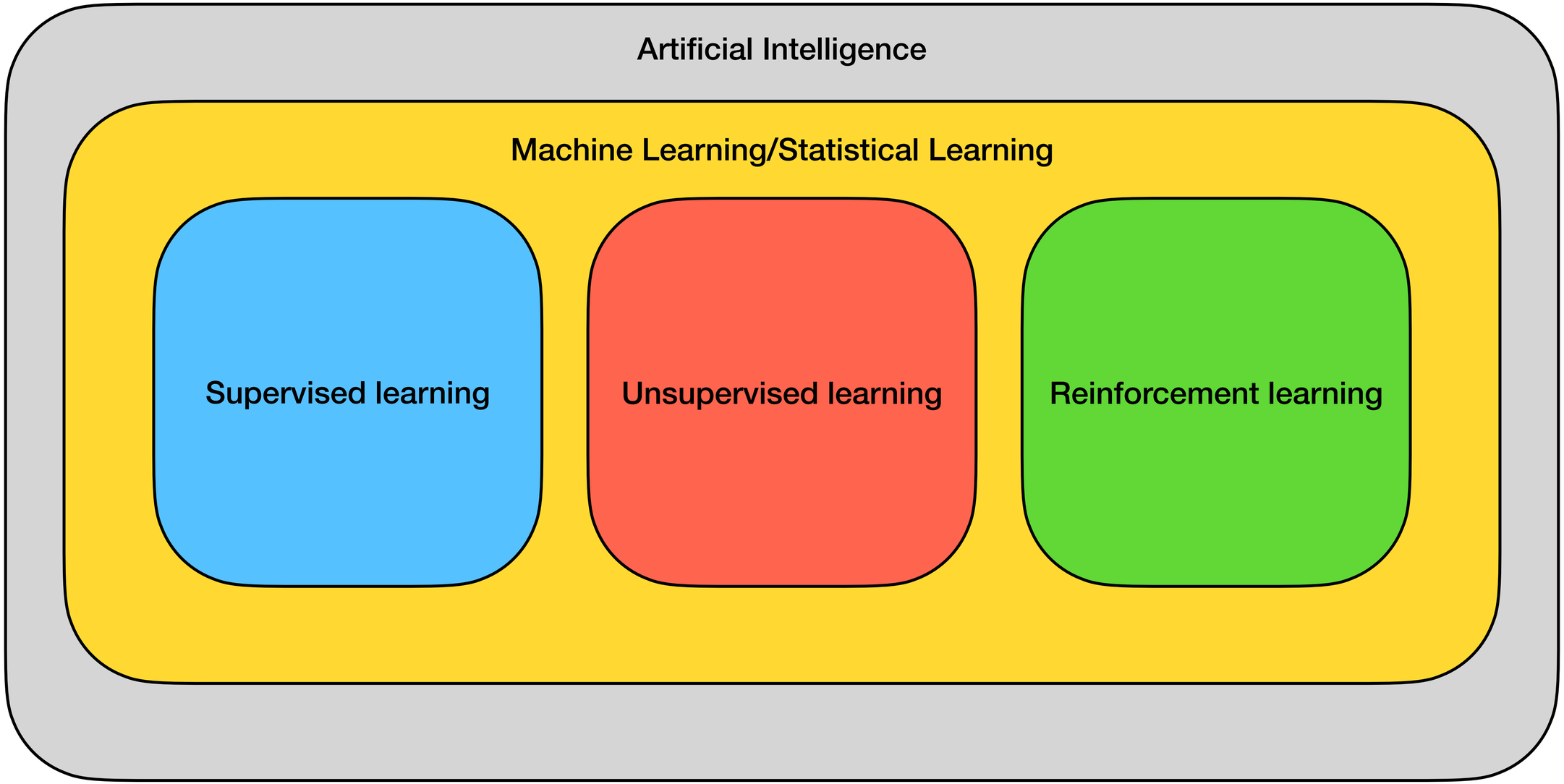
Introduction to
- supervised learning
- unsupervised learning

BIOST 2155

Class 1
August 29th, 2025

More data, more problems…

Statistical learning problems in health sciences



Four biomedical analyses problems

More data, more problems (continued…)

1. Prediction of brain volume from gestational age
2. Diagnosis of heart failure using ultrasound imaging
3. Patient health profile study using wearable devices
4. Detecting breast cancer from biopsy of cell nuclei

Four biomedical analyses problems

More data, more problems (continued…)

1. **Prediction of brain volume from gestational age**
2. Diagnosis of heart failure using ultrasound imaging
3. Patient health profile study using wearable devices
4. Detecting breast cancer from biopsy of cell nuclei

1. Prediction of brain volume from gestational age

How can we leverage machine learning to understand and predict brain development in preterm neonates?

Preterm birth is a significant global health issue. When a baby is born, gestational age helps the healthcare team understand if the baby is premature (born before 37 weeks), full-term, or post-term (born after 42 weeks). This information is vital for assessing the baby's health and any potential needs.

By using a dataset of neonatal brain volumes and their corresponding gestational ages, we can train a model to predict brain volume.

While gestational age is a primary indicator of a newborn's development, there is considerable variability in brain volume among infants born at the same gestational age.

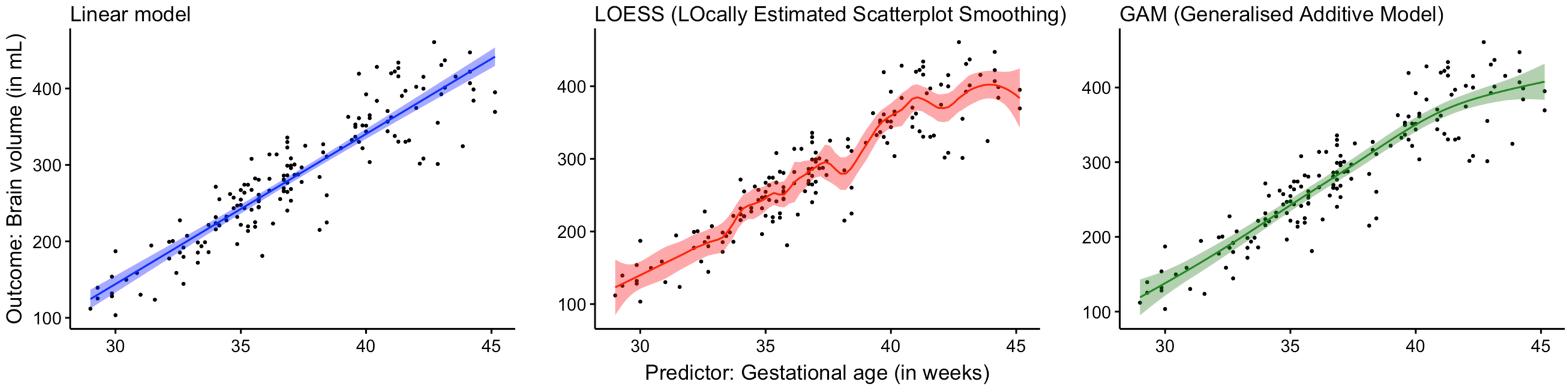
Why do we care?

Early Detection: A predictive model can help clinicians identify neonates with brain volumes that deviate from the expected range for their age, flagging them for closer monitoring and potential early intervention.

Quantitative Tracking: It provides a quantitative and objective way to track brain growth and development in this vulnerable population.

Research Insights: This approach can contribute to a deeper understanding of the factors influencing early-life brain development.

1. Prediction of brain volume from gestational age



While it correctly shows a positive correlation (brain volume increases with gestational age), it fails to capture the slight S-shaped curve present in the data's growth pattern.

Follows the data very closely, capturing local variations and fluctuations.

However, this curve might be overfitting the noise in the data.

Captures the non-linear trend in the data without the excessive fluctuations of the LOESS model. It represents a good balance, illustrating the natural, curved pattern of brain growth over time.

Four biomedical analyses problems

More data, more problems (continued…)

1. Prediction of brain volume from gestational age
2. **Diagnosis of heart failure using ultrasound imaging**
3. Patient health profile study using wearable devices
4. Detecting breast cancer from biopsy of cell nuclei

2. Diagnosis of heart failure using ultrasound imaging

Use ML to predict cardiac health of patients based on their measured heart function.

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the US.

- In 2023, 1 in every 3 deaths was tied to cardiovascular disease.
- ENORMOUS financial burden: overall healthcare cost in the US in 2020-21 was ~$420bn.

https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

By using a dataset of measured heart function features and heart failure outcome (yes/no), we can associate functional features with health outcome.

Ejection Fraction (EF): percentage of blood that the left ventricle (the heart's main pumping chamber) pumps out with each contraction.

A lower EF is a classic sign of a weakened heart muscle, often associated with heart failure.

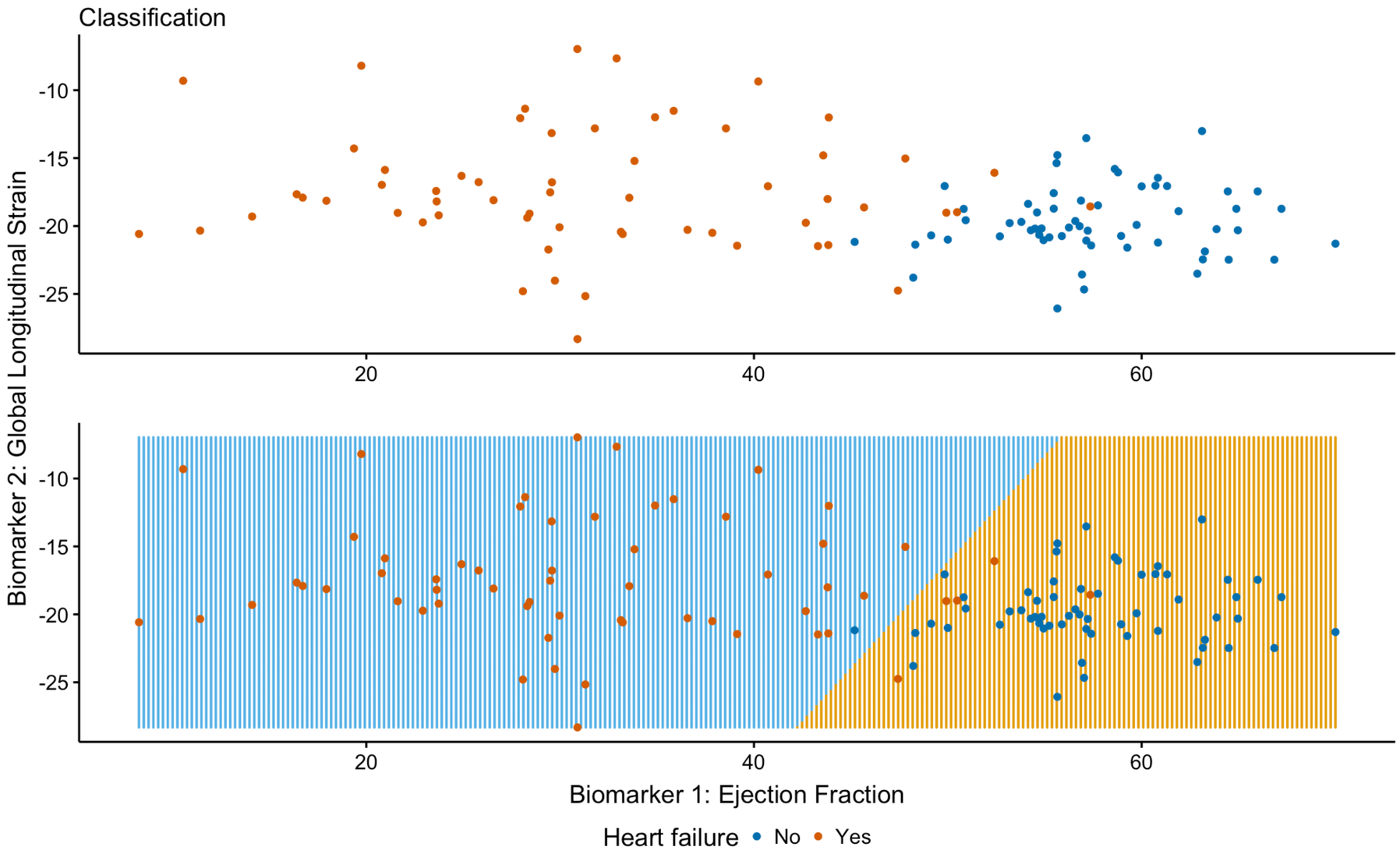
Global Longitudinal Strain (GLS): measures the percentage of deformation (shortening) of the heart muscle tissue in the longitudinal direction during a heartbeat.

GLS can detect subtle changes in heart function even when EF is still in the normal range, making it a valuable tool for early detection.

Why do we care?

Early Detection.
Personalized treatment.
Improved Quality-Adjusted Life Years, better patient outcomes.

2. Diagnosis of heart failure using ultrasound imaging



Healthy patients are clustered in the top right, characterized by a higher EF (typically > 45%) and a lower GLS (typically < -15%).

Patients with heart failure are grouped in the bottom left, showing a lower EF (< 45%) and higher GLS.

Classification model will identify a boundary separating healthy individuals and those with heart failure based on EF and GLS measurements.

Four biomedical analyses problems

More data, more problems (continued…)

1. Prediction of brain volume from gestational age
2. Diagnosis of heart failure using ultrasound imaging
3. **Patient health profile study using wearable devices**
4. Detecting breast cancer from biopsy of cell nuclei

3. Patient health profile study using wearable devices

Use ML to identify patient health profiles based on their routine clinical measurements.

Context: Chronic conditions related to obesity and high blood pressure are a massive public health burden, leading to diseases like heart attack, stroke, and type 2 diabetes.
Challenge: Patients often have multiple risk factors. A manual review of every patient's data is inefficient for large-scale population health management.

By using a dataset of common patient measurements, we can group individuals and associate their combined metrics with different levels of health risk.

Body Mass Index (BMI): A measure of body fat calculated from height and weight.
A higher BMI is a well-established indicator for a range of chronic health issues, serving as a foundational metric for general health.

Systolic Blood Pressure (SBP): Measures the pressure in your arteries when your heart beats.
Elevated SBP is a direct and critical risk factor for cardiovascular events.

Why do we care?

Risk Stratification.
Targeted Interventions.
Proactive Healthcare.

3. Patient health profile study using wearable devices

Group 1 (**High-Risk**): High BMI (e.g., 33-38) and high SBP (e.g., 150-165).
Group 2 (**Healthy**): Lower BMI (e.g., 20-24) and normal SBP (e.g., 110-125).
Group 3 (**At-Risk**): Moderate BMI (e.g., 27-31) and elevated SBP (e.g., 130-145).

Four biomedical analyses problems

More data, more problems (continued…)

1. Prediction of brain volume from gestational age
2. Diagnosis of heart failure using ultrasound imaging
3. Patient health profile study using wearable devices
4. **Detecting breast cancer from biopsy of cell nuclei**

4. Detecting breast cancer from biopsy of cell nuclei

Use ML to visualize high-dimensional tumor data to identify separable groups.

Context: Breast cancer diagnosis relies on analyzing over 30 quantitative features from tumor images, such as cell size, shape, and texture.
Challenge: Manually analyzing 30+ interacting features for every patient is impossible. We cannot "see" the data in 30 dimensions to spot hidden patterns.

By applying PCA to a dataset of tumor measurements, we can distill complex tumor-related information into a simple 2D map.

Original Features: Measurements like mean radius, mean texture, and mean smoothness describe the tumor.

Principal Components (PCs): PCA combines the original 30+ features into two new, powerful summary variables that capture the most important information.

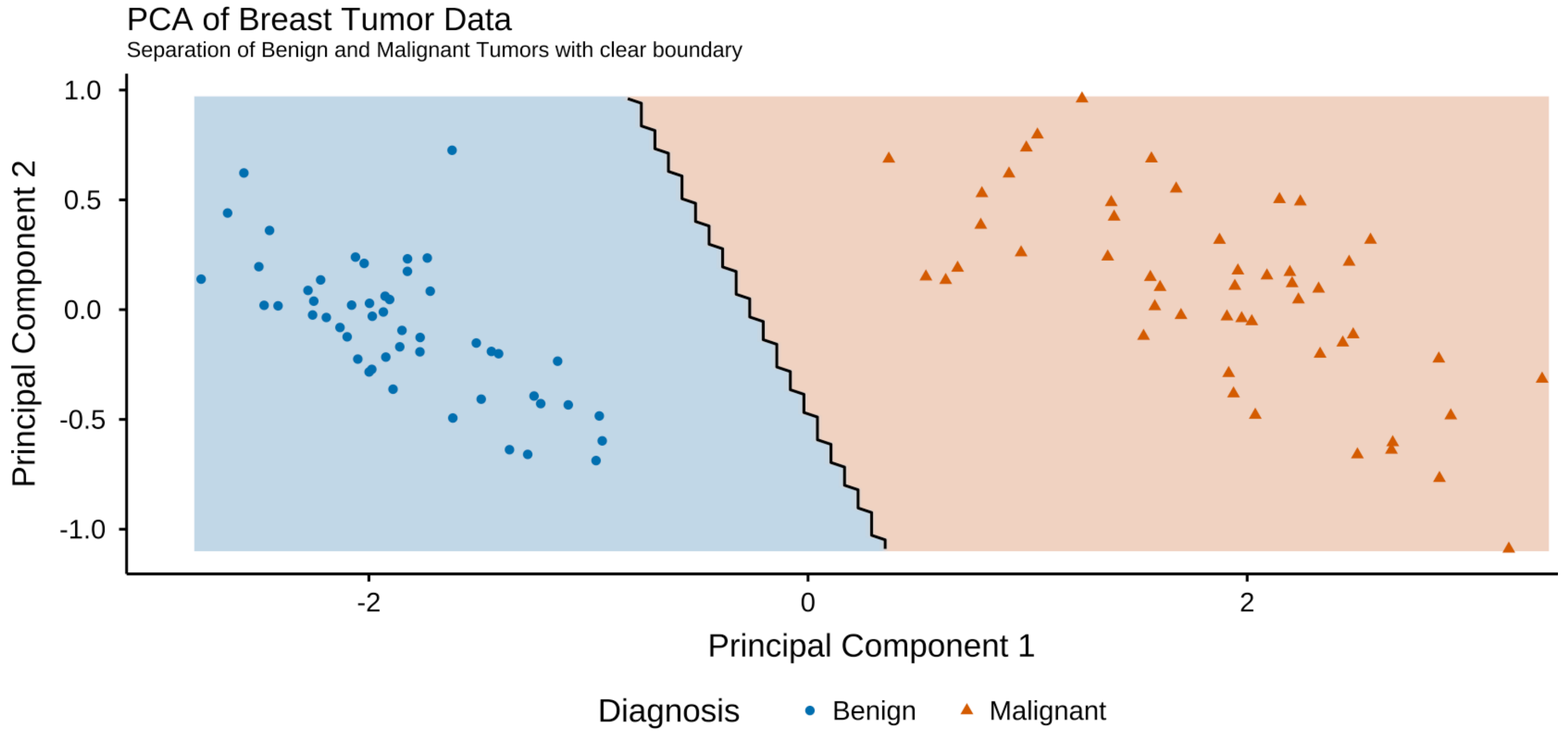
Why do we care?

Rapid Pattern Discovery.
Confirms Model Feasibility.
Data-Driven Insights.

4. Detecting breast cancer from biopsy of cell nuclei

	Benign (N=50)	Malignant (N=50)	Overall (N=100)
radius			
Mean (SD)	9.99 (0.935)	20.1 (1.69)	15.0 (5.26)
Median [Min, Max]	9.99 [7.97, 11.9]	20.1 [16.5, 23.6]	14.2 [7.97, 23.6]
texture			
Mean (SD)	15.2 (1.94)	22.2 (2.84)	18.7 (4.27)
Median [Min, Max]	14.8 [11.1, 19.3]	22.3 [16.8, 29.2]	18.2 [11.1, 29.2]
perimeter			
Mean (SD)	59.8 (4.74)	120 (8.61)	89.9 (31.1)
Median [Min, Max]	59.8 [50.4, 68.9]	121 [98.9, 138]	83.9 [50.4, 138]
area			
Mean (SD)	303 (92.2)	999 (125)	651 (366)
Median [Min, Max]	308 [117, 513]	1020 [740, 1250]	626 [117, 1250]

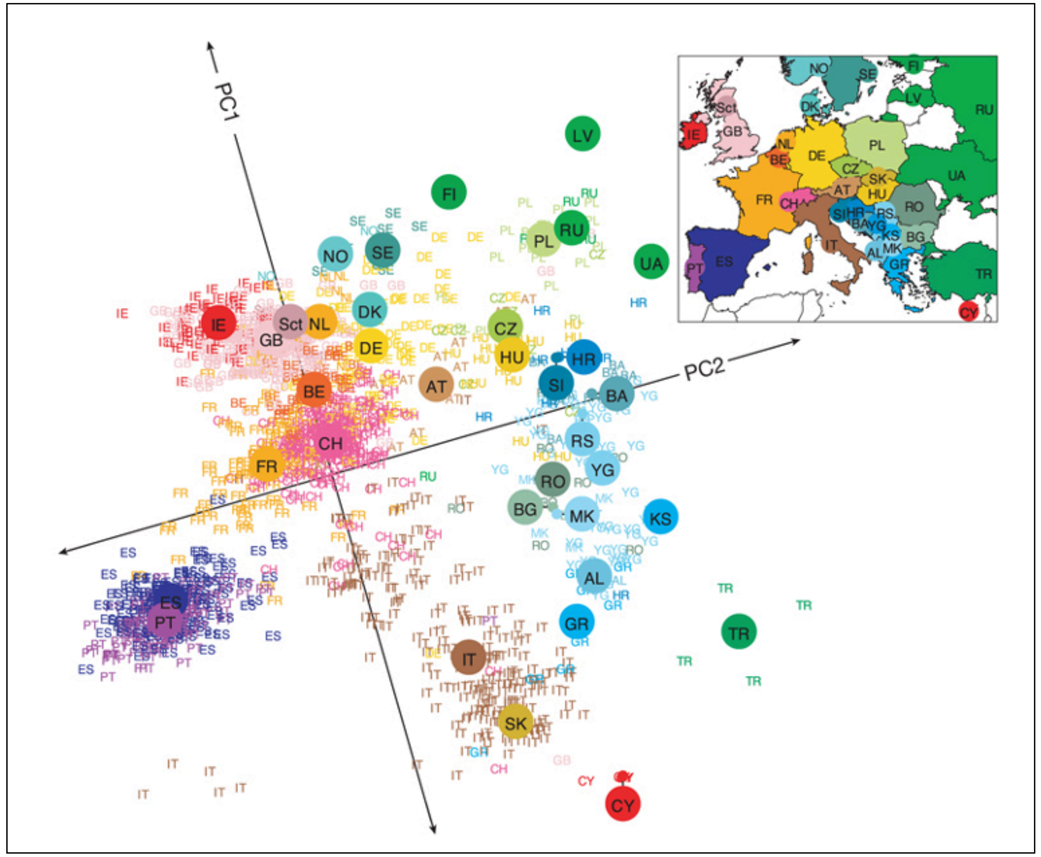
→



PCA can take a complex dataset with many measurements and distill it into a simple, powerful visualization.

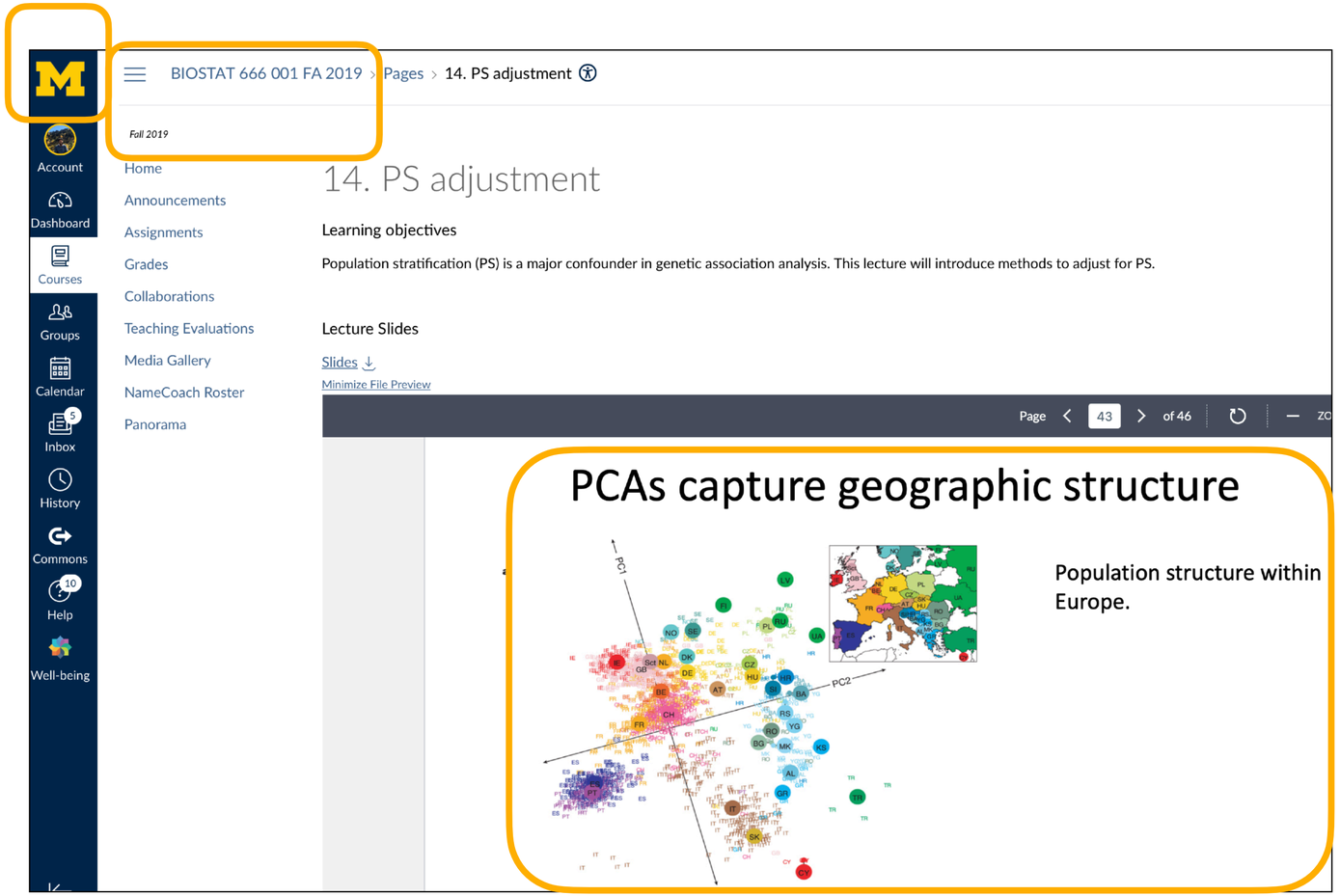
We see a clear underlying pattern that separates **benign** and **malignant** tumors, confirming that the data is highly suitable for building an accurate predictive machine learning model.

Prehistoric days: a story from 2019…



Novembre, J., Johnson, T., Bryc, K. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008). https://doi.org/10.1038/nature07331

A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small colored labels represent individuals and large colored points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe.



Math notation (🤢)

1. # of study units: n
2. Measurements on study units: samples
   A. Must have feature vector $\mathbf{x} = (x_1, \ldots, x_D)^T$
   B. May/may not have label/outcome: y
   C. Outcome present: supervised
   D. Outcome absent: unsupervised
3. Data from sample: $\{(y_i, x_{1i}, \ldots, x_{Di})^T\}_{i=1}^n$

Output Input

$$\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad \begin{pmatrix} x_{11} \cdots x_{D1} \\ \cdots \\ \cdots \\ \cdots \\ x_{n1} \cdots x_{nD} \end{pmatrix}$$

Rows: study units

Columns: features

Math notation (🤢)

1. Mathematical model f: take feature as input and produce output

$$\hat{\gamma} = f(\mathbf{x})$$

   A. Regression $\hat{\gamma}$: target value
   B. Classification and clustering $\hat{\gamma}$: label
   C. Dimension reduction $\hat{\gamma}$: transformed feature vector

2. Specify form of the model f using θ: $\hat{\gamma} = f(\mathbf{x}, \theta)$
3. Training a model = finding a ‘good’ θ

4. What is/are good θ?
   A. Minimize loss function $\mathscr{L}$ for the given data: $\hat{\theta} = \underset{\theta}{\text{argmin}}\ \mathscr{L}(\mathbf{X}, \mathbf{y}, \theta)$

Supervised learning overview

1. Both output Y and input $\mathbf{x}$ are present.
    1. In the regression problem, Y is quantitative (e.g price, blood pressure)
    2. In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).

2. On the basis of the data we would like to get f that can
    1. Accurately predict unseen Y for a given $\mathbf{x}$
    2. Explain how varying $\mathbf{x}$ might affect Y

Unsupervised learning overview

1. Only input $\mathbf{x}$ are present.
    1. No outcome variable; just a set of features measured on a set of samples

2. Objective is more fuzzy
    1. Find groups of samples that behave similarly
    2. Find features that behave similarly
    3. Find groups of features with the most variation

3. Different from supervised learning; can be useful as a pre-processing step for supervised learning

Mapping what we have learnt so far…

	Features (input)	Targets/ labels (output)	Task type	ML type	
Prediction of brain volume from gestational age	✅	✅	Regression	**Supervised learning**	
Classification of heart failure using ultrasound or magnetic resonance imaging	✅	✅	Classification		
Patient health profiles using wearable device data	✅	🚫	Clustering	**Unsupervised learning**	
Compressing information from 30 features to study similarities/differences in benign and malignant breast tumors	✅	🚫	Dimensionality reduction		

Mapping what we have learnt so far…

	Features (input)	Targets/ labels (output)	Task type	ML type	Metric
Prediction of brain volume from gestational age	✅	✅	Regression	**Supervised learning**	*RMSE, R-squared*
Classification of heart failure using ultrasound or magnetic resonance imaging	✅	✅	Classification		*Accuracy, Precision, Recall, AUC*
Patient health profiles using wearable device data	✅	🚫	Clustering	**Unsupervised learning**	*Silhouette score*
Compressing information from 30 features to study similarities/differences in benign and malignant breast tumors	✅	🚫	Dimensionality reduction		*Explained Variance*

Artificial intelligence 🤝 Machine learning 🤝 Statistical learning

Machine learning: learn from data and predict outcome(s).

Statistical modeling: Formalization of relationships between variables in the form of equations.

Statistical learning: Learn from data and predict outcome(s) while understanding relationships between variables

There is **A LOT** of overlap — both fields focus on supervised and unsupervised problems:

Emphasizes model interpretability, precision and uncertainty

Greater emphasis on large scale applications and prediction accuracy.

But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”

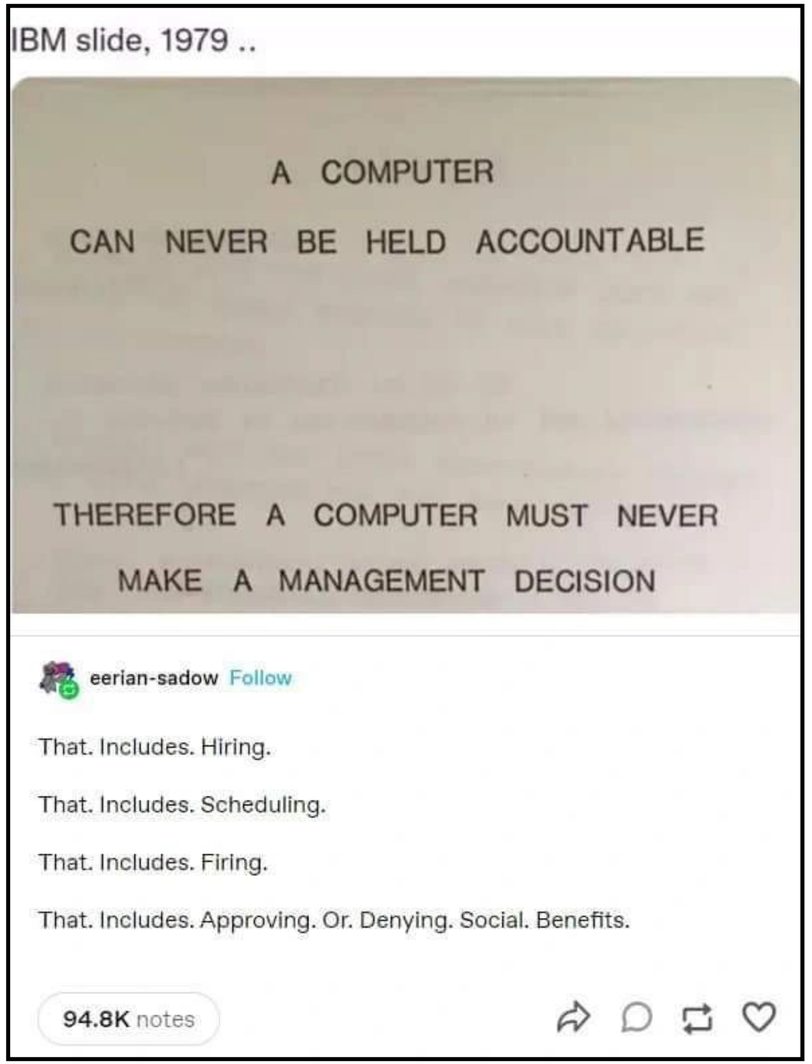
Ethics in AI

Irresponsible to proceed without discussing the ethical implications of artificial intelligence

Bias and fairness: our best models are (**often**) a reflection of our worst selves.

1. A model that predicts salary levels for individuals based on historical data will probably predict that women be paid less than men.

2. An AI system for super-resolving face images made non-white people look more white.



Deep learning systems make decisions, but we do not usually know exactly how or based on what information

What lies ahead…

Preliminaries	
08/29	**Introduction**: supervised learning (regression and classification); unsupervised learning (clustering and dimensionality reduction).
09/05	**Training statistical learning models**: under- and over-fitting; training, validation, and testing; cross-validation; bootstrap.
Supervised learning I: Regression	
09/12	**Regression (I):** regularization using subset selection; shrinkage.
09/19	**Regression (II):** non-linear regression; splines; kernel trick-based.
Supervised learning II: Classification	
09/26	**Classification (I):** basics; generative models for classification.
10/03	**Classification (II):** support vector machines.
Unsupervised learning I: Clustering	
10/17	**Clustering (I):** k-means; cluster evaluation.
10/24	**Clustering (II):** Gaussian mixture models.
Unsupervised learning II: Dimensionality reduction	
10/31	**Dimensionality reduction (I):** principal component analysis.
11/07	**Dimensionality reduction (II):** independent component analysis.
More statistical learning tools	
11/14	**Decision trees.**
11/21	**Ensemble learning.**
12/05	**Case studies.**

Housekeeping

Course logistics, textbooks, coding support…

Instructor: Soumik (show-mick) Purkayastha
soumik@pitt.edu
A740 Public Health
Office Hours: Fri 11.00 AM - 12.00 noon

TA: Hao Wang
haw291@pitt.edu
A724A Public Health
Office Hours: Mon 11:30 - 12:30 PM

TA: Jessica Shao
yus221@pitt.edu
A724A Public Health
Office Hours: Wed 12:00 - 1:00PM

- No class on 10/10 (Fall Break)
- No class on 11/28 (Thanksgiving Break)
- Project proposal due on 10/03.
- Project presentations on 12/05. Reports due on 12/12.

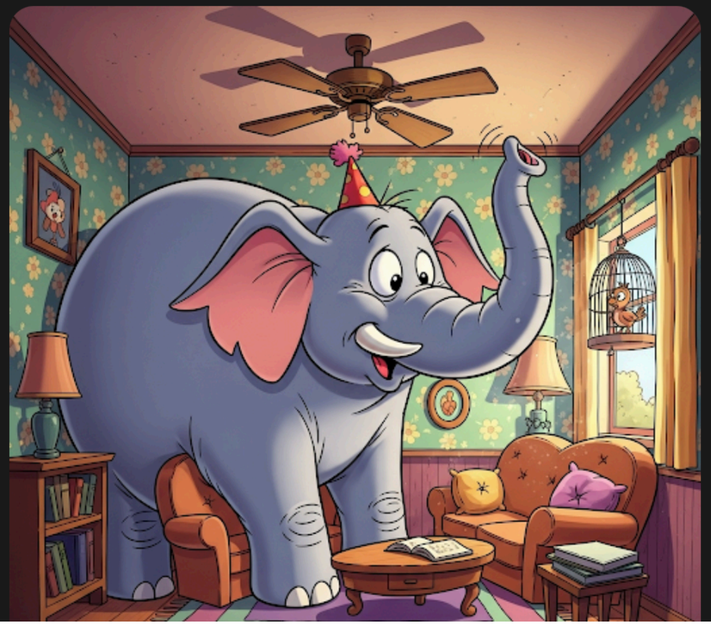
Grading (tentative)
- Homework: 3 problem sets, 25 points each.
- One final project: 25 points.

Delays in submission will be penalized unless instructor is informed in advance.

Hey Gemini, I'm going to talk to students about the use of AI for my course, which is the elephant in the room -- give me a cartoon that shows an elephant in a room.

TIP: Ask Gemini to refine images
With image generation in Gemini, your imagination is the limit. If what you see doesn't quite match what you had in mind, try adding more details to the prompt. The more specific you are, the better Gemini can create images that reflect your vision.

Show thinking



Housekeeping

Course logistics, textbooks, coding support…

- Will contain all class slides, code, datasets, etc…
  - Canvas: https://canvas.pitt.edu/courses/324229
  - Github repo: https://github.com/soumikp/2025_sl4hds/tree/main