

Workshop: Data Analysis Strategies for Single-Cell Image-Based Profiling:

Day 3: Single-cell data analysis

Gisele Miranda

**KTH Royal Institute of
Technology**

**Division of Computational
Science & Technology, EECS**

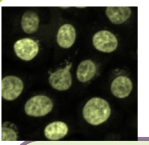


gmirand@kth.se

Workshop overview



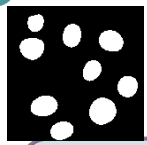
1



Introduction to BioImage Analysis

- ❖ Basics of Digital Images
- ❖ File formats & Metadata
- ❖ Fund. Image Processing: filtering, denoising, etc
- ❖ Image Quality and reproducibility
- ❖ Hands-on session using Fiji

2



Cell Segmentation and Feature extraction

- ❖ Overview of segmentation techniques
- ❖ Introduction to machine and deep learning-based segmentation
- ❖ Feature extraction: measurements
- ❖ Exporting single-cell datasets for analysis

3



Single Cell Data Analysis

- ❖ Overview of high-content data structure
- ❖ Dimensionality reduction: PCA, LDA, t-SNE, UMAP
- ❖ Clustering methods: K-means, GMM
- ❖ Hands-on session with single-cell datasets

4



Daily Schedule

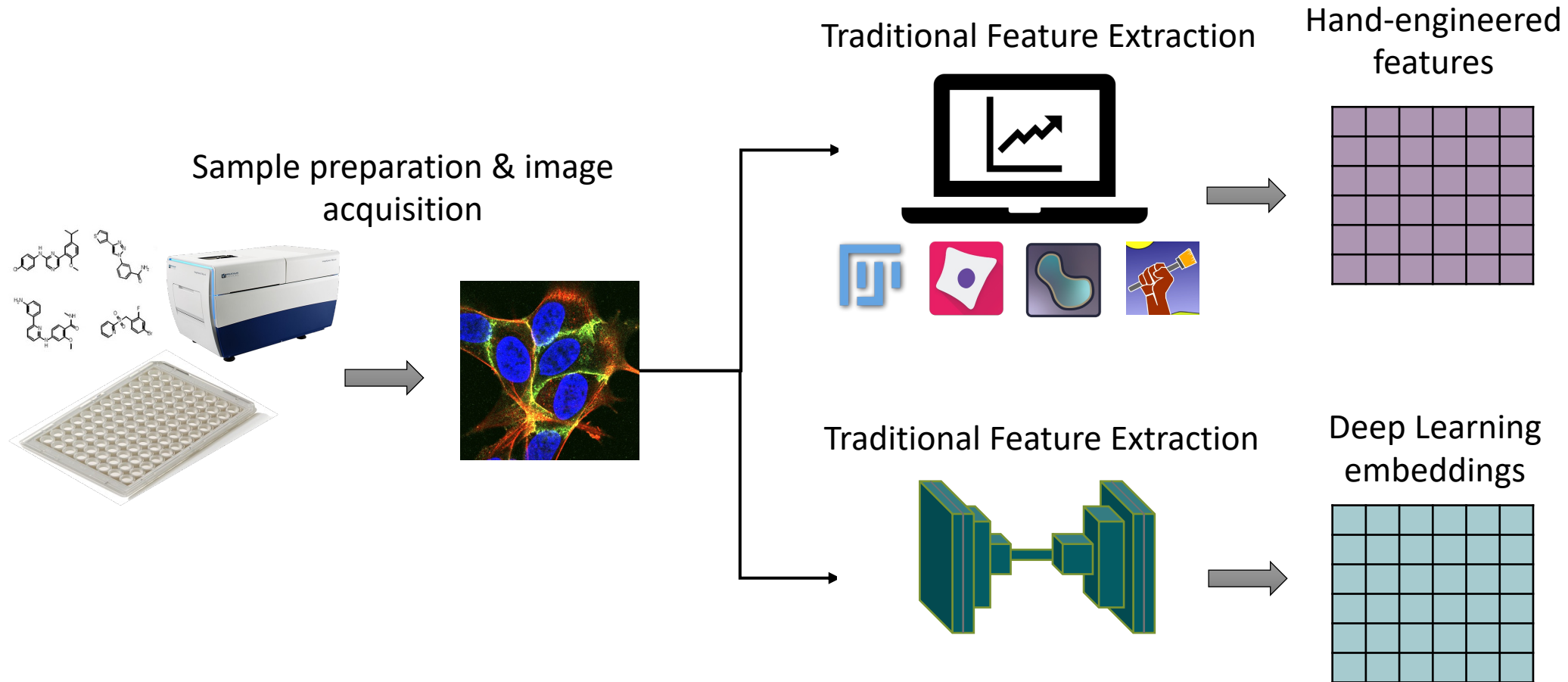
- 🕒 **09:00 – 10:20:** First Session
- ☕ **10:20 – 10:40:** Break
- 🕒 **10:40 – 12:00:** Second Session
- 🍴 **12:00 – 13:30:** Lunch Break
- 🕒 **13:30 – 14:30:** Third Session
- ☕ **14:30 – 14:45:** Break
- 🕒 **14:45 – 16:00:** Fourth Session

High-content screening



- Automated microscopy + quantitative image analysis
 - ✓ Thousand of cells, many channels, large-scale data
- Assays to measure cellular morphology under different perturbations
 - Used in drug discovery, phenotyping, functional genomics
- Challenge Analysis
 - ✓ High-dimensional data (hundreds of features)
 - ✓ Noise reduction, classification based on cell morphology

High-content screening pipeline



High-content screening pipeline



- Aggregation Level
 - Single-cell: retains heterogeneity, ideal for deep profiling
 - Population-level: averages features across cells per well

High-content screening pipeline



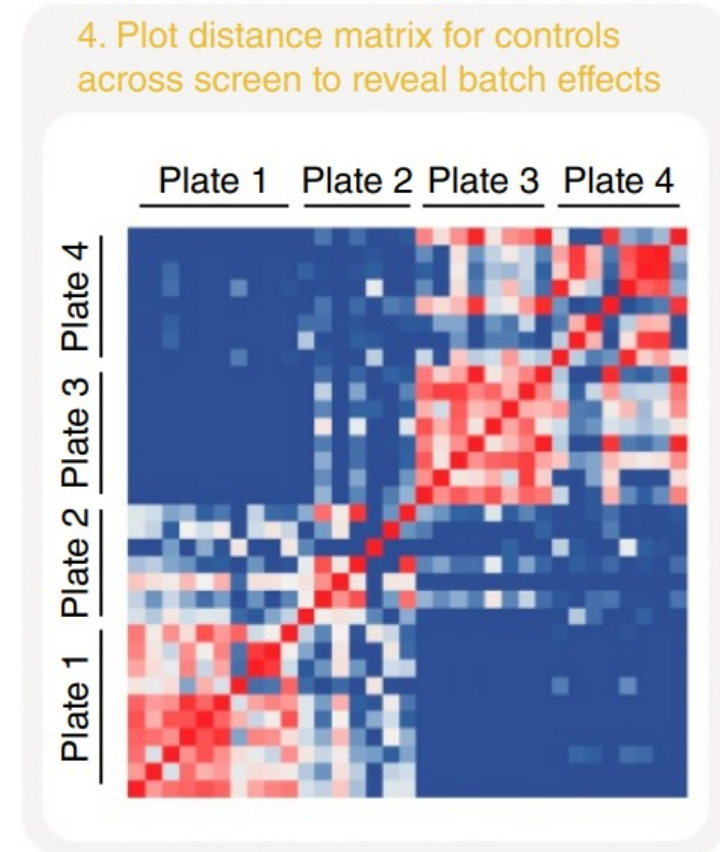
- Normalization:
 - Adjust feature scales and reduce intra-plate and inter-plate variability (e.g., intensity differences, edge effects, illumination differences).
 - Examples: z-score

Method	Description	Use Case
Z-score	Subtract mean, divide by std dev (typically control wells)	Normalize across wells or plates
Robust Z-score	Median-centered, divide by MAD (median absolute deviation)	Resistant to outliers
Min-Max Scaling	Rescale features between 0–1	Rare; less robust to outliers
Plate-level Median Normalization	Subtract plate median or DMSO median	Used to align different plates

Batch Effect



- Unwanted variation from experimental conditions like different imaging times, microscope settings, or reagent lots
- High variance explained by batch metadata (e.g., plate, well, site)
- Samples cluster by plate instead of treatment



Caicedo, Juan C., et al. "Data-analysis strategies for image-based cell profiling." *Nature methods* 14.9 (2017): 849-863.

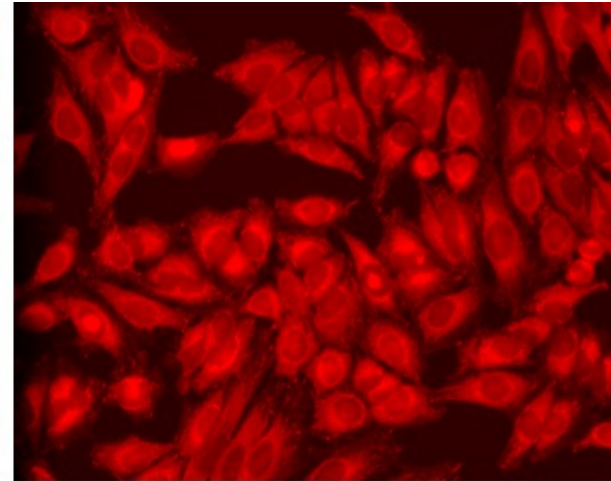
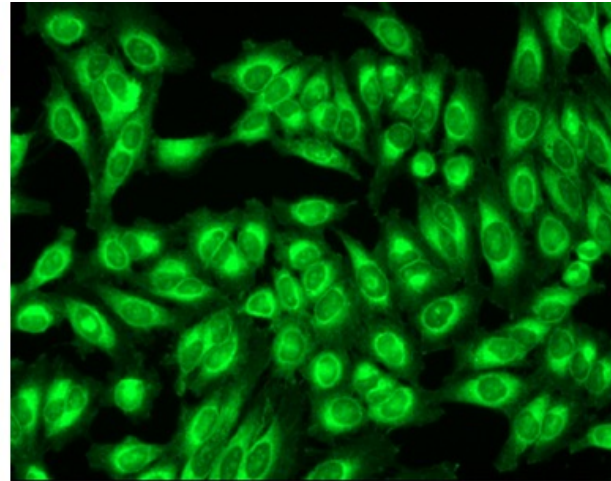
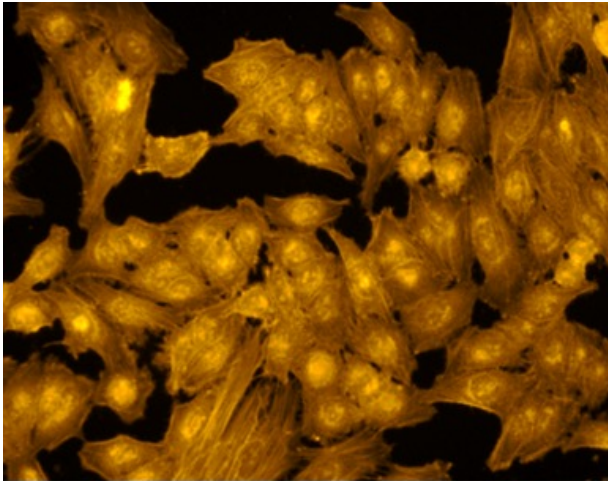
Feature selection



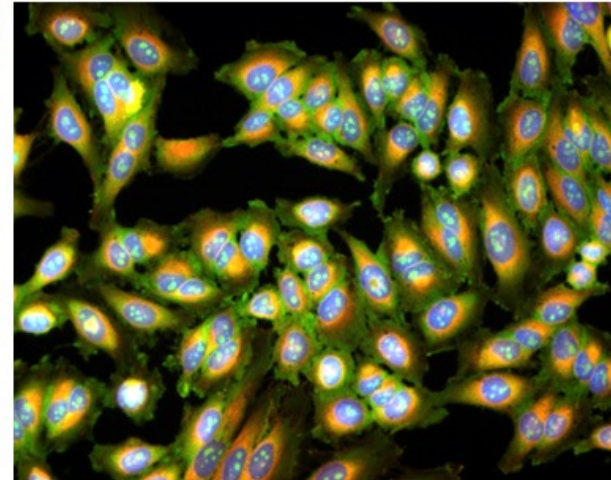
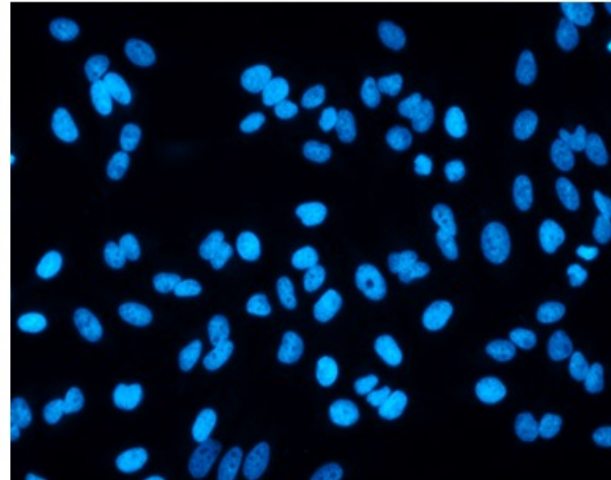
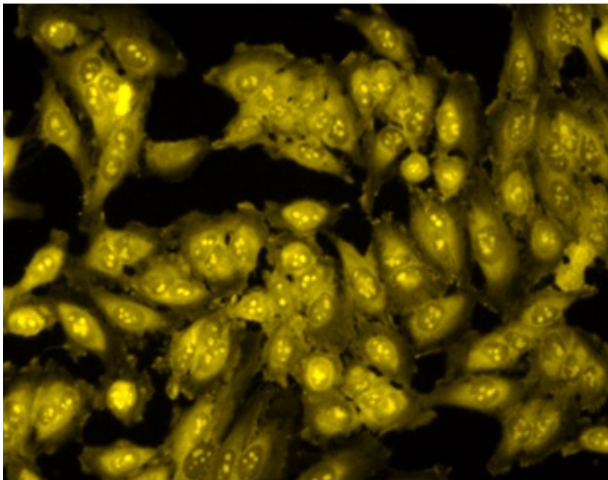
- Reduce the number of irrelevant or redundant features and reduce computational cost for analysis.

Method	Description	Example
Variance Thresholding	Remove features with low variance	<0.01 or 5th percentile
Correlation Filtering	Drop highly correlated features (e.g., >0.9)	Keep one of each pair
Blocklist Removal	Drop known noisy/uninformative features	e.g., metadata artifacts
Biological Feature Curation	Keep only shape or texture features	Manual subset

Cell Painting Assay



- ✓ Nuclei
- ✓ Cytoplasm
- ✓ ER
- ✓ Golgi
- ✓ Mitochondria



JUMP-CP dataset



- Consortium: Broad Institute, NIH, pharma companies
- Goal: reference dataset for image-based profiling
 - 116,000 chemical perturbations (i.e. small-molecule treatments)
 - ~8,000 CRISPR knockout perturbations (loss-of-function genetic perturbations)
 - ~12,600 ORF overexpression perturbations (gain-of-function genetic perturbations)



Dataset Overview

Dataset Collection

This collection comprises 4 datasets:

- Principal dataset ([cpg0016](#)): 116k chemical and ~22k gene perturbations, split across 12 data-generating centers using human U2OS osteosarcoma cells. This includes JUMP-ORF, JUMP-CRISPR, and JUMP-compounds
- 3 pilot datasets testing:
 - Different perturbation conditions ([cpg0000-jump-pilot](#) , including different cell types)
 - Staining conditions ([cpg0001-cellpainting-protocol](#))
 - Microscopes ([cpg0002-jump-scope](#))

Design of the Dataset

Cell Line Selection

- We chose U2OS (osteosarcoma) cells for our major data production work because phenotypes are equally or more visible than the few other lines we've tested and there is existing data in this cell type (namely, [cpg0012-wawer-bioactivecompoundprofiling](#))

Available chemicals in JUMP

For a given compound, find its JUMP id or its most similar equivalent. It may take a few seconds to load.

The similarity is calculated using the Jaccard index between the query's PubChem fingerprints and all the JUMP compounds available on PubChem.

On this page

[Available chemicals in JUMP](#)

 [Edit this page](#)

[View source](#)

[Report an issue](#)

Other Formats

 [Jupyter](#)

Find your compound in JUMP


This tool uses Pubchem fingerprints to associate any Pubchem-available compound to its closest JUMP analog.

Instructions:

1. Submit your compound and identifier type (e.g., 'aspirin' and 'Common Name', 'BSYNRYMUTXBXSQ-UHFFFAOYSA-N' and 'InChIKey')
2. Copy the top choice (either InChIKey or Metadata_JCP2022) and use it on broad.io/compound.
3. If your compound is found on PubChem, you will be shown the top matches in JUMP

Identifier

Identifier type

Common Name 

Submit

https://broadinstitute.github.io/jump_hub/explanations/data_description.html

JUMP-CP dataset



- We have selected 30 sample images of five distinct compounds to be used in this workshop:

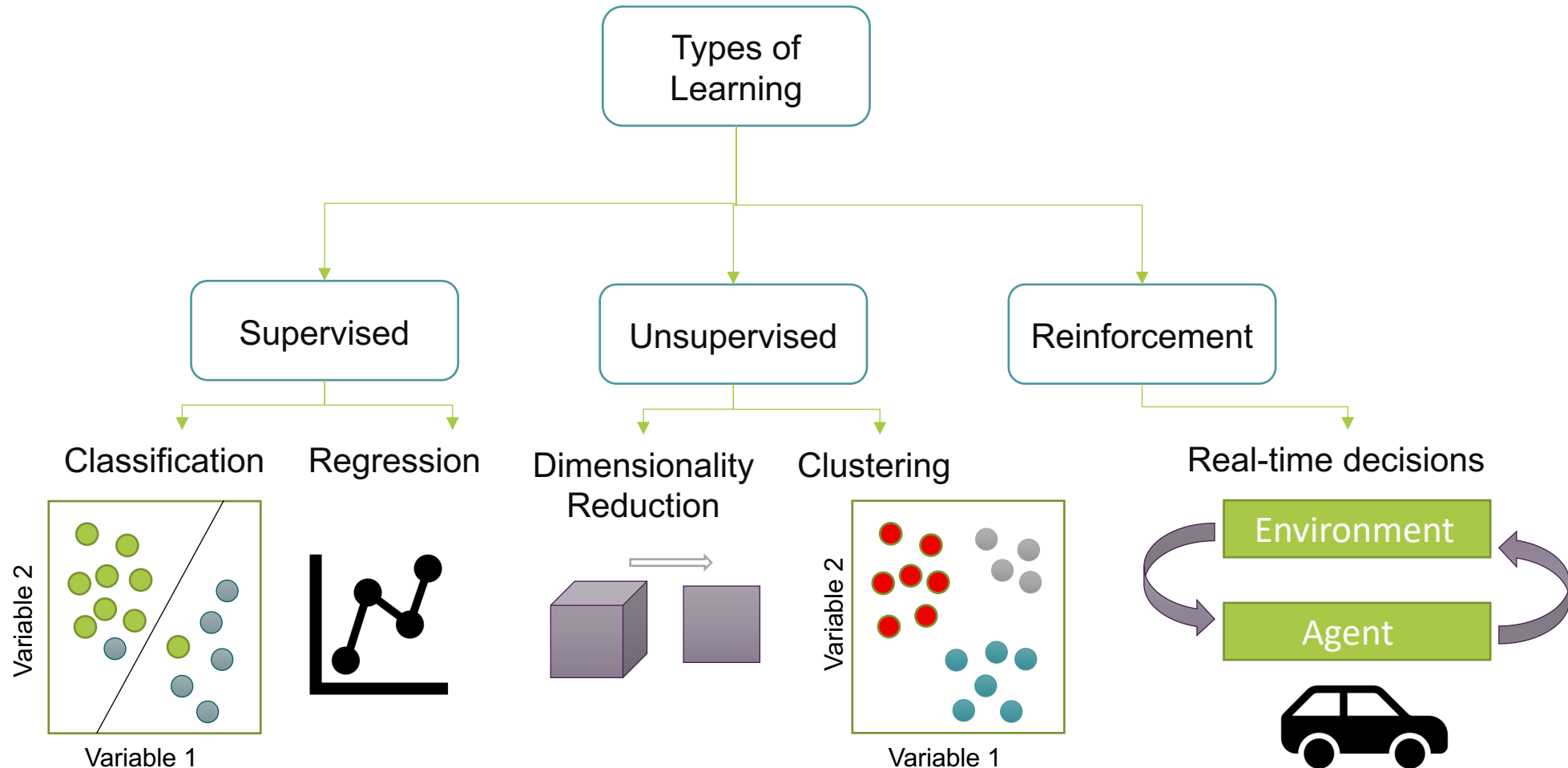
Compound	Mechanism of Action	Distinct from
Brefeldin A	Inhibits ER-to-Golgi protein transport	Most others
Etoposide	Inhibits topoisomerase II, leading to DNA double-strand breaks (antineoplastic agent)	Brefeldin A, Rapamycin
Nocodazole	Binds tubulin to inhibit microtubule polymerization (prevent microtubule formation)	Nocodazole, Staurosporine
Rapamycin	Inhibits mTOR (mammalian target of rapamycin) pathway - immunosuppressant drug	Rapamycin
Staurosporine	Broad-spectrum protein kinase inhibitor, induces apoptosis	Etoposide, Brefeldin A

Handling High-Dimensional Data

Types of Learning



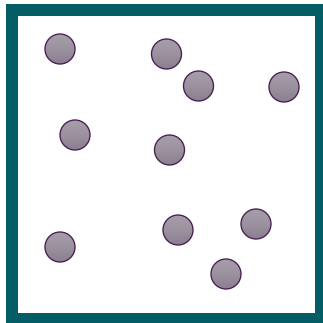
- Machine learning tasks



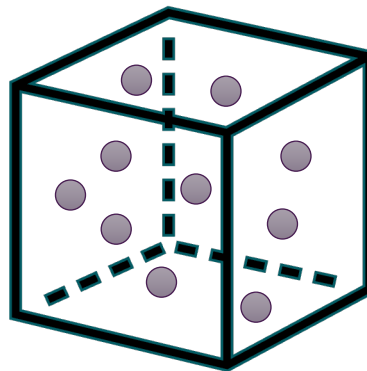
Why High-Dimensional Data is Challenging?



- In Cell Painting assays, each cell is described by 1,000 -1,500+ features, a very high-dimensional space.
- Key challenges:
 - Curse of Dimensionality: as dimensions increase, distance metrics become less meaningful, sparsity
 - Visualization becomes impossible, more redundant/noisy features, computational burden



2D



3D



High-
dimensional?

Dimensionality Reduction



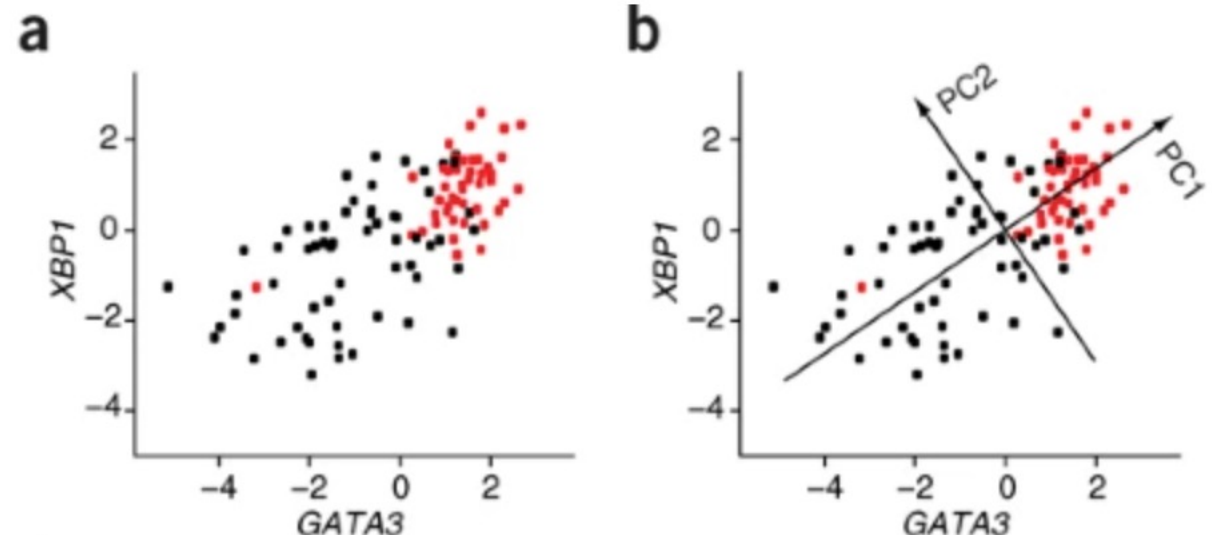
- Motivation:
 - Pattern visualization (e.g., clusters, gradients)
 - Remove redundant or irrelevant features
 - Compress data while preserving key biological variation
 - Improve performance of clustering and ML algorithms

Principal Component Analysis (PCA)



- Transforms high-dimensional data into a lower-dimensional space while preserving as much variance (information) as possible.
- A dataset with many correlated variables is transformed into a smaller set of uncorrelated variables called principal components

PCA is like rotating the feature space to reveal the directions where your data varies the most.



Principal Component Analysis (PCA)



- Given a data matrix:

$$\boxed{X} \in \mathbb{R}^{n \times p}$$

Data matrix

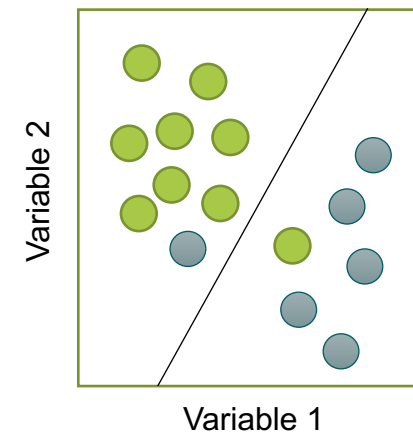
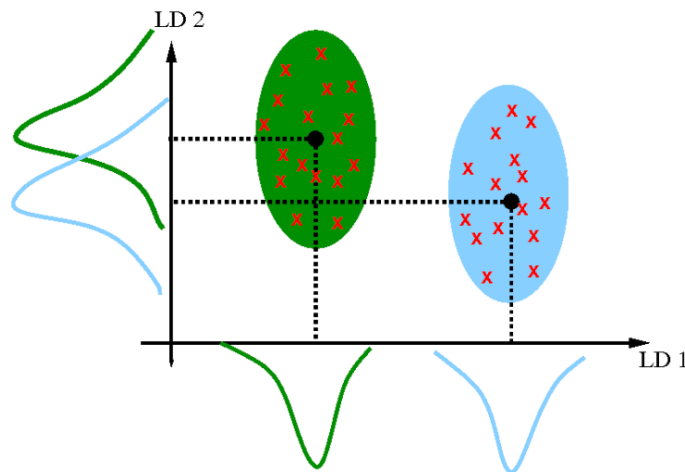
n: number of samples
p: number of features

1. Center the data (subtract mean): $X_{\text{centered}} = X - \mu$
2. Compute covariance matrix: $\Sigma = \frac{1}{n-1} X_{\text{centered}}^T X_{\text{centered}}$
3. Compute eigenvalues and eigenvectors: $\Sigma w_i = \lambda_i w_i$
4. Sort eigenvectors by eigenvalue
5. Project data onto principal components: $Z = X_{\text{centered}} W_k$

Linear Discriminant Analysis (LDA)



- LDA is a **supervised** dimensionality reduction technique that projects high-dimensional data into a lower-dimensional space while maximizing class separability.
- Difference between PCA and LDA
 - **PCA** focuses on directions of maximum variance and **LDA** focuses on directions that best separate known classes (e.g., treatment labels, mechanisms of action).



Linear Discriminant Analysis (LDA)



- Given a data matrix and the corresponding class labels y :

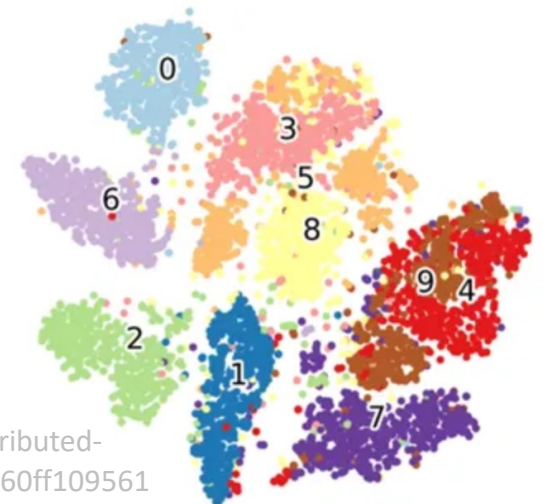
$$X \in \mathbb{R}^{n \times p} \quad y \in \{1, \dots, C\}$$

1. Compute the class means (μ_c) and global mean (μ):
2. Compute the within-class scatter matrix:
$$S_W = \sum_{c=1}^C \sum_{x_i \in c} (x_i - \mu_c)(x_i - \mu_c)^T$$
3. Compute the between-class scatter matrix:
$$S_B = \sum_{c=1}^C n_c (\mu_c - \mu)(\mu_c - \mu)^T$$
4. Solve the generalized eigenvalue problem:
$$S_W^{-1} S_B w_i = \lambda_i w_i$$
5. Select the top k discriminant directions
6. Project data into LDA space: $Z = XW_k$

t-SNE (t-distributed Stochastic Neighbor Embedding)



- Nonlinear dimensionality reduction technique that maps high-dimensional data into 2D or 3D space, preserving local structure
- Focused on capturing the relationships between nearby points
 - Compute pairwise similarities (probability that i and j are neighbors) using a Gaussian kernel
 - Create a Map in 2D or 3D
 - Adjust the Map to Match High-D Relationships



UMAP (Uniform Manifold Approximation and Projection)

1. Build a graph of the data (**high-D space**)
 - For each point, find its nearest neighbors
 - Build a weighted graph where edge strength reflects similarity
2. Build a Similar Graph in 2D or 3D (**low-D space**)
 - Create a graph in 2D where the neighbor relationships are as similar as possible to the high-D graph
3. Optimize the Layout
 - Use an algorithm that pulls similar points together and pushes dissimilar points apart, balancing local and global structure

Practical Exercises



- Now, you will work with the *data_analysis.ipynb* notebook. Follow the instructions on the notebook to read a csv data file and run the different dimensionality reduction techniques discussed in this section
- Refer to slide 13 for more information about the selected JUMP-CP compounds

Practical Exercises



Why did LDA work best?

- Class separation dominates the variance
- PCA captures directions of maximum variance, whether or not they separate classes.
- If PCA mixes the classes, it suggests that overall variance is not aligned with class structure.
- In contrast, LDA ignores irrelevant variance and focuses only on between-class vs. within-class spread.

Clustering Methods

Why Cluster Single-Cell Morphology Data?

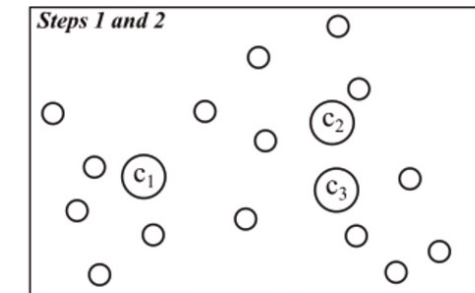


- Why Cluster Single-Cell Morphology Data?
 - Discover subpopulations of cells
 - Identify compound-induced phenotypes without knowing MoA
 - Reveal outliers or heterogeneity within a compound's effect
- Clustering is an unsupervised learning technique
 - Groups data points based on similarity
 - Unlike LDA, clustering doesn't use labels

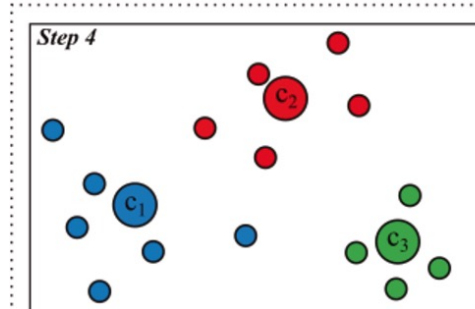
K-Means



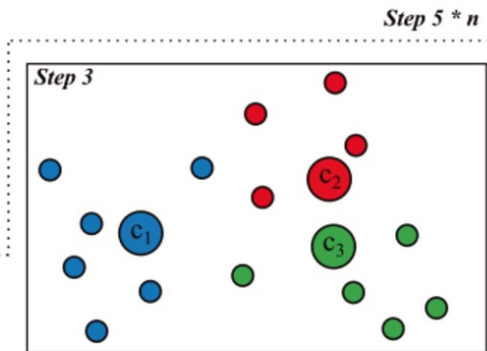
- Input:
 - Data points $X = \{x_1, x_2, \dots, x_n\}$
 - Number of clusters k
- Output:
 - Cluster assignments for each x_i
 - Centroid positions $\mu_1, \mu_2, \dots, \mu_k$
- 1. Initialize: select k data points as initial centroids μ_1, \dots, μ_k
- 2. Repeat until convergence (or max iterations):
 - For each point x_i :
 - Assign x_i to the nearest centroid (based on Euclidean distance)
 - For each cluster j :
 - Update centroid $\mu_j = \text{mean of all points assigned to cluster } j$
- 3. Return:
 - Final cluster assignments



Definition of the clusters and their centroids



Recalculation of the centroids and reassignment of the labels



Assignment of the observations

*Step 5 * n*

K-Means

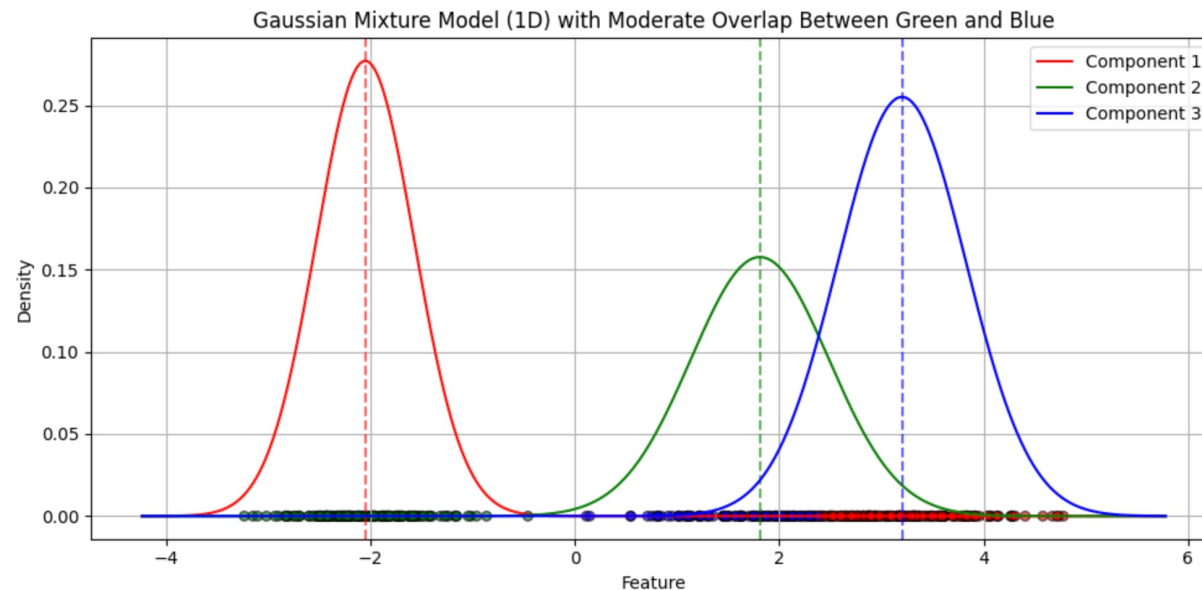


- Assumes spherical, equally sized clusters
- Sensitive to initialization
- Sensitive to outliers: outliers can pull centroids away from true cluster centers
- Hard assignment (no uncertainty)

Gaussian Mixture Models (GMMs)



- Probabilistic clustering method
- Unlike K-means, GMM performs soft clustering:
 - Each data point is assigned a probability of belonging to each cluster.



Gaussian Mixture Models (GMMs)



- Each cluster is modeled as a multivariate normal distribution with its own:
 - Mean (μ), the center
 - Covariance (Σ), the shape and orientation
 - Weight (π), the proportion of points in the cluster
- It is more flexible than K-means because it can model elliptical clusters, overlapping clusters, and clusters with different sizes or densities.

DBSCAN

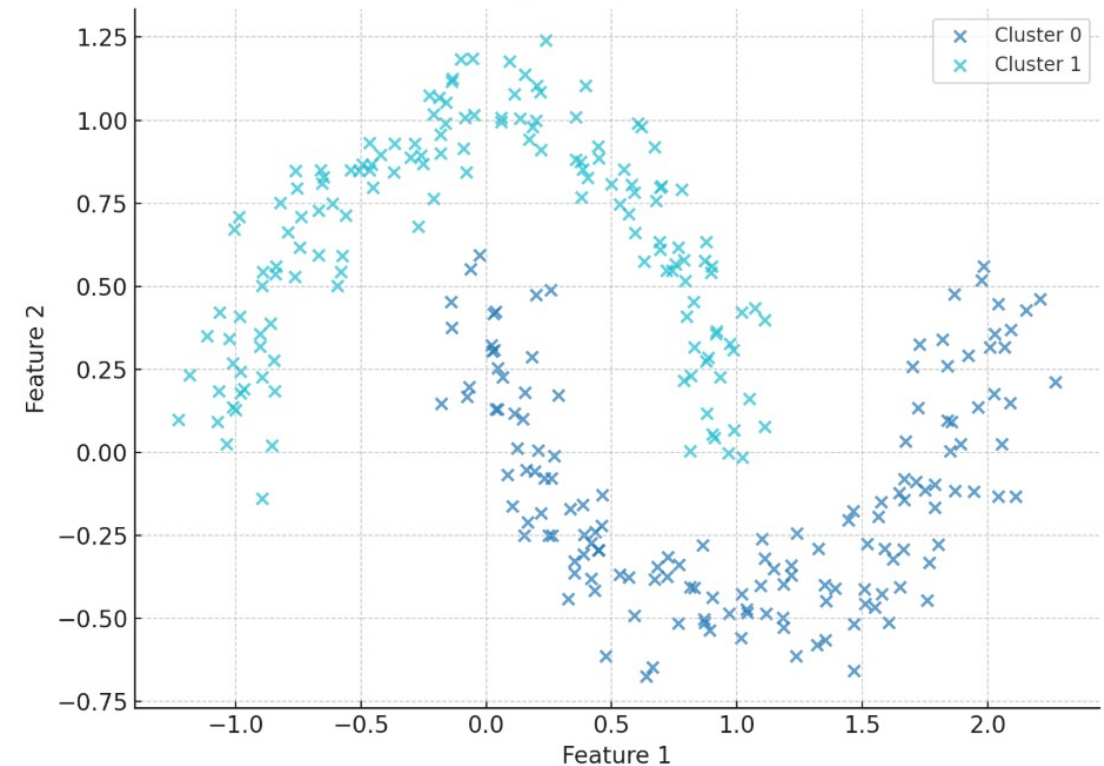


- Density-based spatial clustering of applications with noise
- Clusters are dense regions of points separated by areas of low density
- No need to pre-specify number of clusters
- Can identify arbitrary-shaped clusters and outliers (noise)

DBSCAN



- How It Works:
 - Choose two parameters:
 - ϵ (epsilon): radius of neighborhood
 - minPts: minimum number of points to form a dense region
 - For each point:
 - If $\geq \text{minPts}$ points are within ϵ , then mark as core point
 - If reachable from a core point, assign to that cluster
 - Else, mark as noise



Comparing methods



Feature	K-means	GMM	DBSCAN
Cluster shape	Spherical	Elliptical / Gaussian	Arbitrary
Cluster assignment	Hard (1 cluster per point)	Soft (probabilities across clusters)	Hard (or noise)
Number of clusters (k)	Must be specified	Must be specified	Not required
Handles outliers?	No	No	Yes (labels them as noise)
Initialization sensitive?	Yes	Yes	No
Interpretable?	Very	Somewhat	Can be complex

Evaluating clustering results



- Two Evaluation Scenarios:
 - Labels available: external evaluation (compare with true classes)
 - No labels (typical case): internal evaluation (use structure only)

Metric	Type	What It Measures	Range / Interpretation
Silhouette Score	Internal	Cohesion vs. separation (within vs. between clusters)	-1 to 1 (higher = better)
Davies-Bouldin Index	Internal	Cluster similarity based on dispersion and separation	≥ 0 (lower = better)
Calinski-Harabasz Index	Internal	Ratio of between- to within-cluster variance	Higher = better
Adjusted Rand Index (ARI)	External	Agreement between predicted and true labels	-1 to 1 (1 = perfect match)
Normalized Mutual Info (NMI)	External	Shared information between clusters and labels	0 to 1 (1 = perfect match)
Purity	External	Proportion of dominant class in each cluster	0 to 1 (1 = pure clusters)

Practical Exercises



- You will continue working with the *data_analysis.ipynb* notebook. Follow the instructions on the notebook to read a csv data file and run the different clustering techniques discussed in this section
- Refer to slide 13 for more information about the selected JUMP-CP compounds