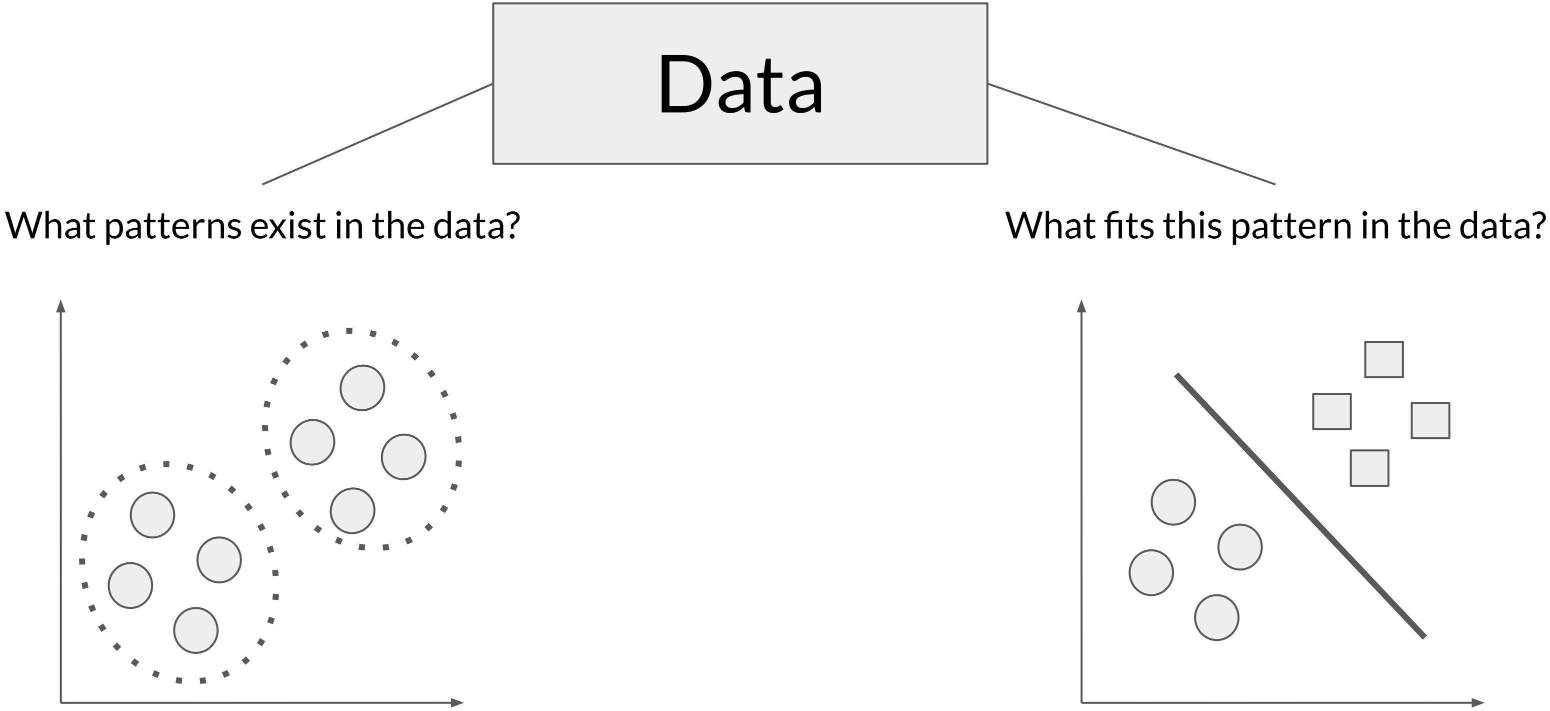
Introduction to Machine Learning

The CCDL

Machine Learning, what is it?

Having a computer program learn to perform a task (like predicting an outcome) from data, rather than programming explicit instructions

Classes of machine learning algorithms



Data

What patterns exist in the data?

What fits this pattern in the data?

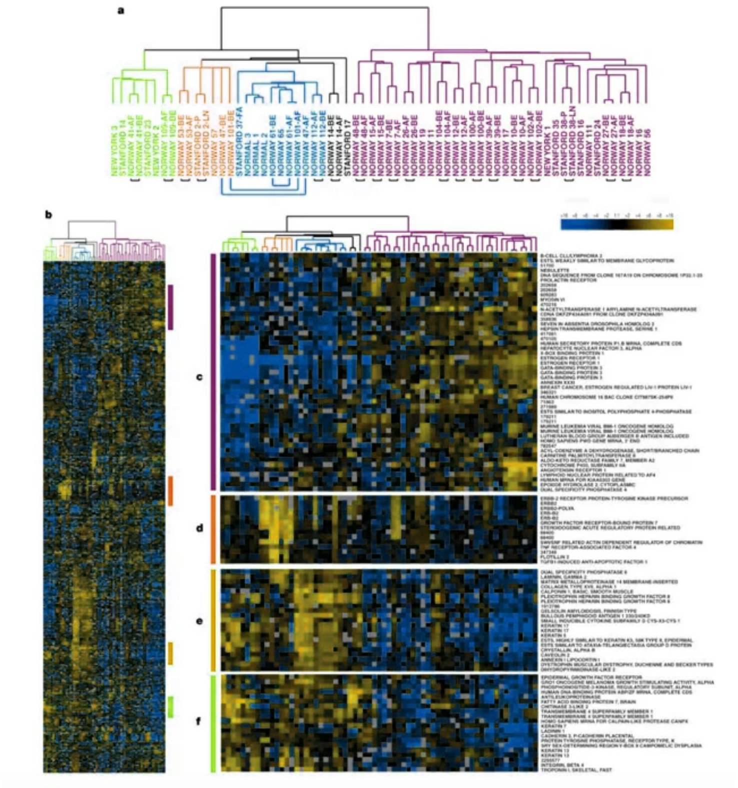
UNSUPERVISED

SUPERVISED

We’ll focus on **unsupervised** learning

- Which samples are most similar to one another?
- How many groups of samples exist in my data?
- What patterns of gene expression exist in my data? How do the genes vary together?

Why do we care about sample-sample relationships?



If we want to find groups of samples with different underlying molecular processes, we will want to know about how stable those groupings are going forward.

Perou et al. “Molecular portraits of human breast tumors.” *Nature.* 2000.

Consensus clustering can help us understand the stability of groupings

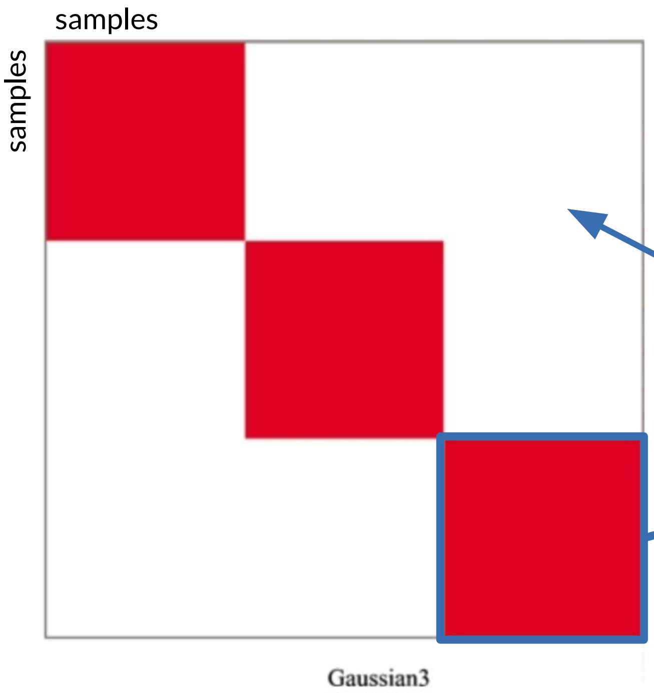
Idea: We can identify how robust clusters are (and how many there are) by resampling the original data and applying clustering over multiple runs to find the agreement or *consensus*.

The *consensus index* tells us how often two samples are clustered together (0 = never, 1 = always)

Monti et al. *Machine Learning*. 2003.

Consensus clustering: simulated data with three groups

Consensus matrix for *k* = 3



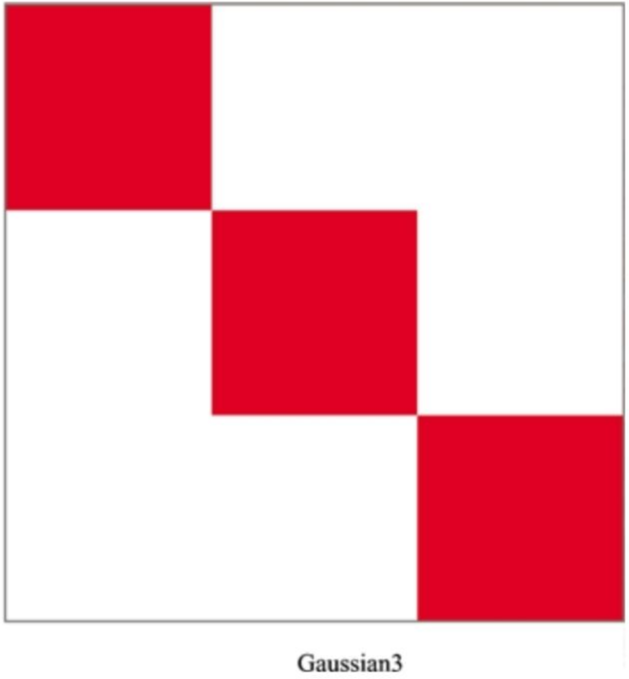
The values in this matrix are the *consensus indices* for pairs of samples.

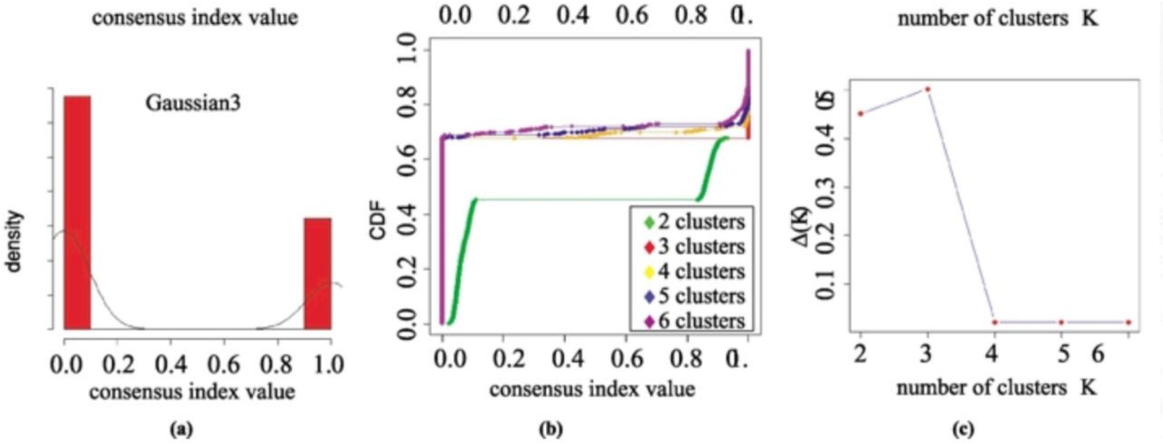
And they never group with these samples

These samples always group together

[Monti et al. *Machine Learning*. 2003.]

Consensus clustering: simulated data with three groups





Monti et al. *Machine Learning*. 2003.

Let’s look at notebooks 01 and 02

We’ll focus on **unsupervised** learning

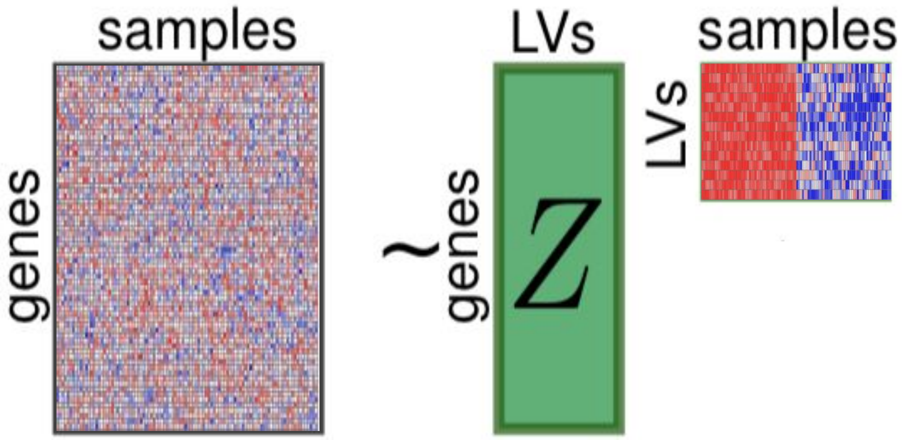
- Which samples are most similar to one another?
- How many groups of samples exist in my data?
- What patterns of gene expression exist in my data? How do the genes vary together?

Why analyze data at the pathway- or pattern-level rather than the gene-level?

Small, coordinated changes in an entire pathway are likely to be biologically significant. However, these are unlikely to be detected by differential expression analysis. (Subramanian et al. *PNAS.* 2005.).

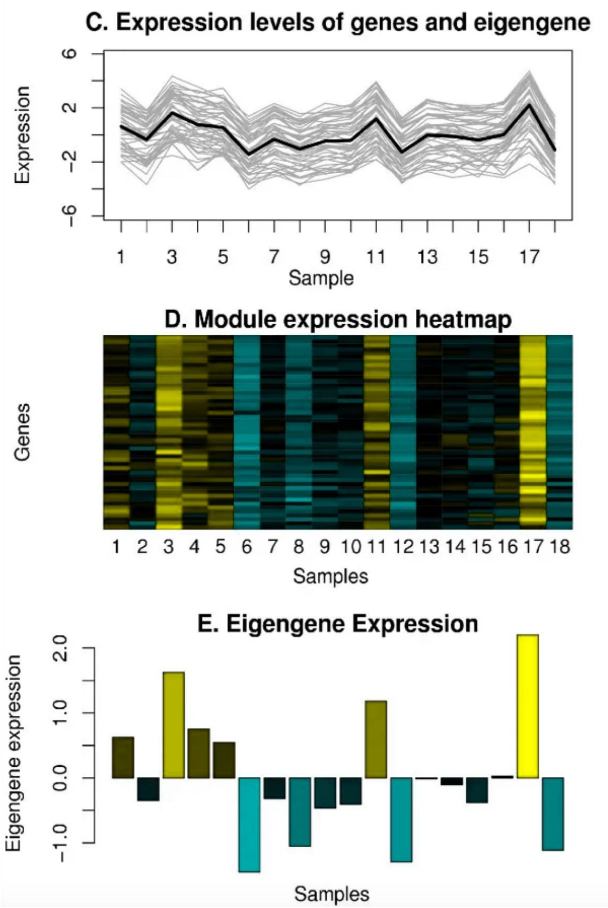
We can also **learn patterns**, which may correspond biological pathways, **from the data themselves**.

Pathway-Level Information ExtractoR (PLIER)



PLIER learns “eigengene-like” latent variables (e.g. patterns that are combinations of genes’ expression) and some of them will correspond to pathways that it takes as input (e.g., KEGG)

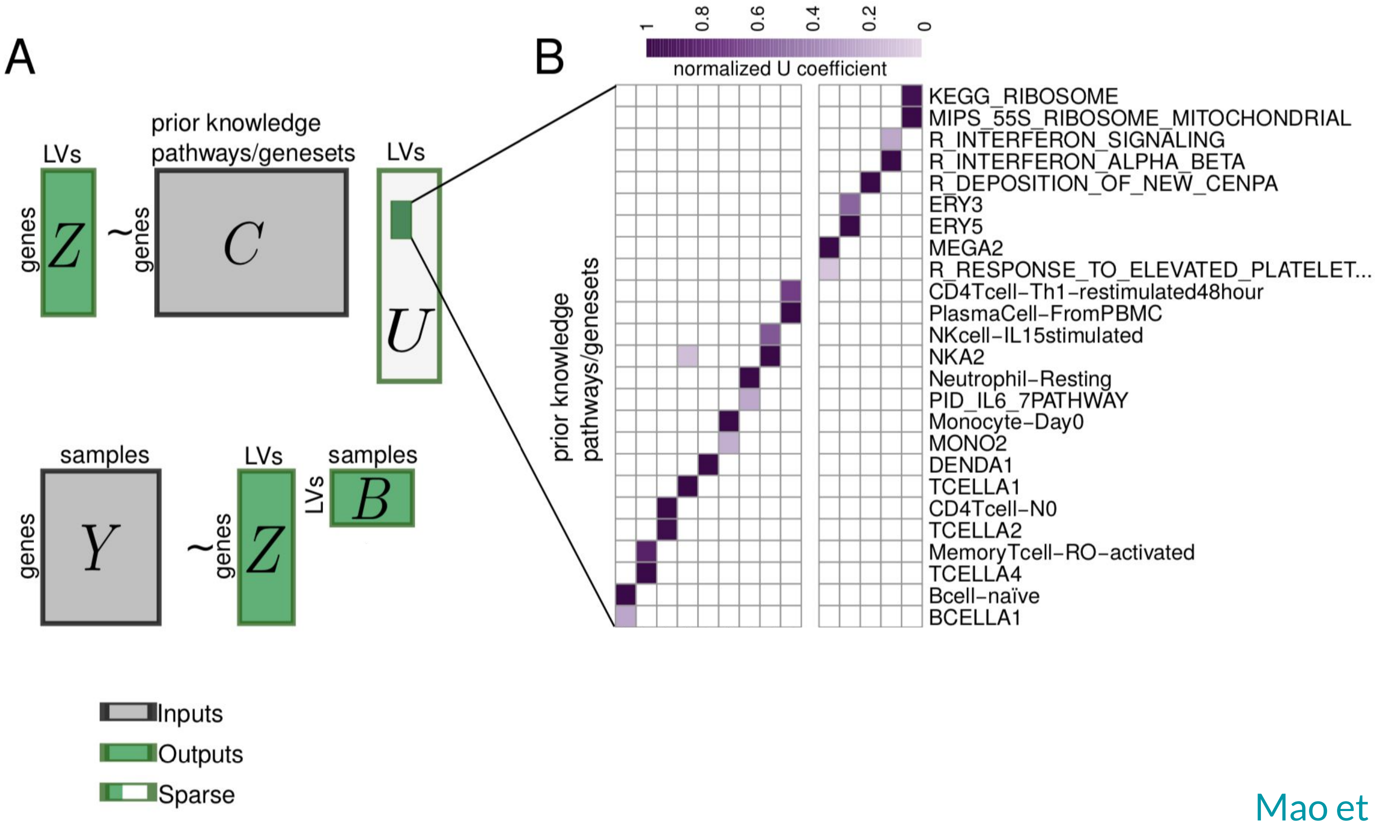
Mao et al. *Nature Methods*. 2019.



“Eigengene-like”

[Langfelder and Horvath. *BMC Systems Biology.* 2007.](#)

A model that is helpful for biological discovery will have latent variables that map to few, hopefully related input pathways



Mao et al. *bioRxiv*. 2017.

How does this compare to what we’ve seen so far?

Like the pathway analysis modules we covered yesterday, gene sets are provided to PLIER.

PLIER is similar to Gene Set Variation Analysis (GSVA) in that it is unsupervised – you do not need to do a two group comparison ahead of time – and the output is a pattern by sample matrix.

PLIER is designed to align the LVs it constructs with the relevant input gene sets that the data supports, whereas other methods will use all gene sets you provide as input. It also compares favorably to other immune infiltrate estimate methods, but it’s not limited to that – it does other pathways/gene sets, too!

Let’s look at notebooks 03 and 04