Multi-Omics Integration for Personalised Medicine

EMBL-EBI Course

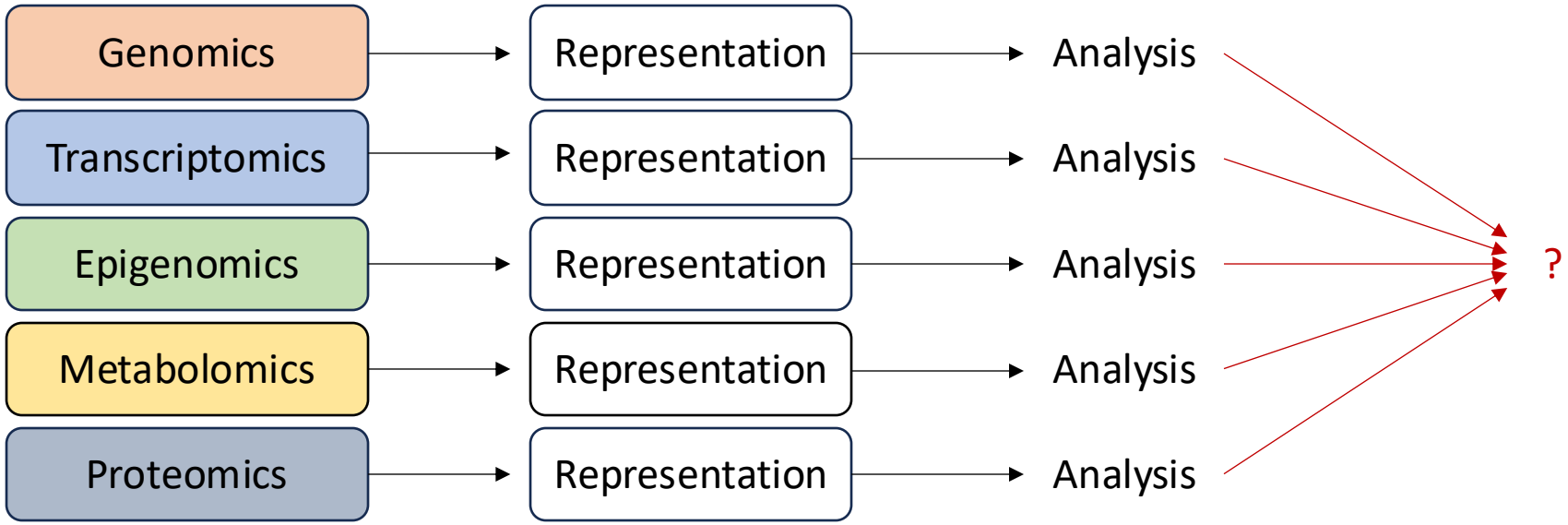
Introduction to Multi-Omics Data Integration and Visualisation

Group Project

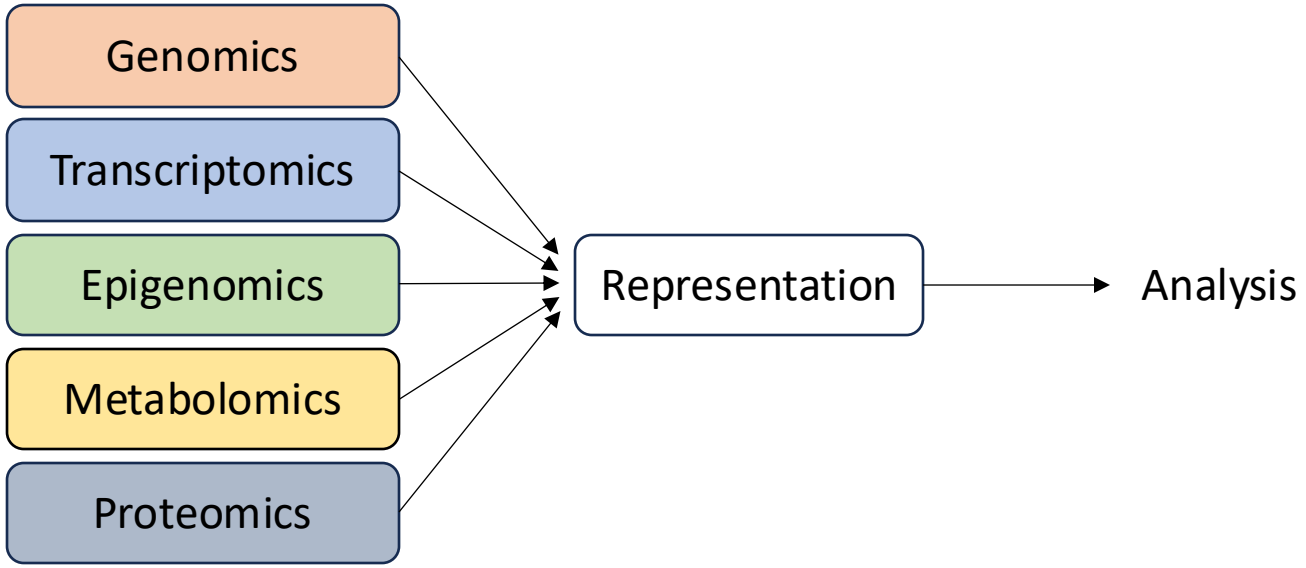
Project Introduction

What to expect

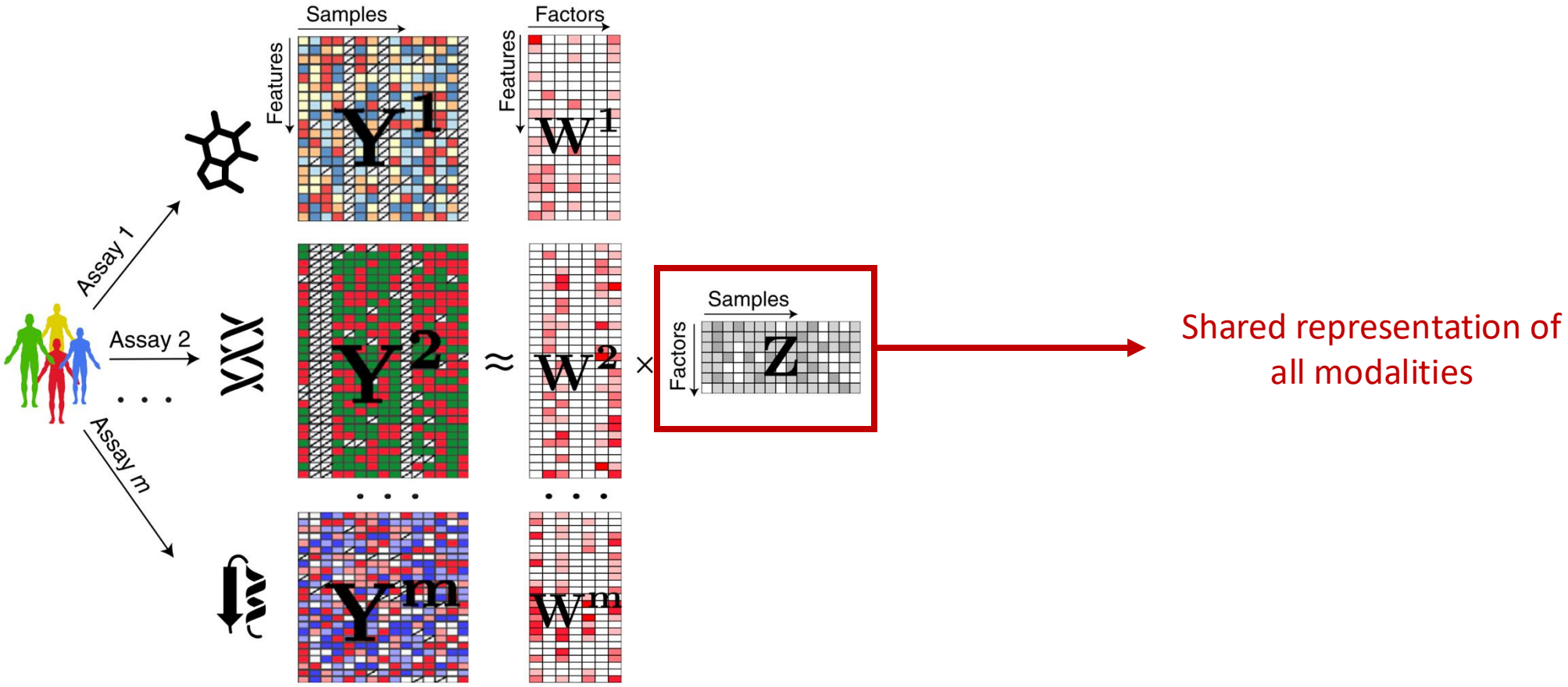
Integration of Multi-Omics Data



Integration of Multi-Omics Data

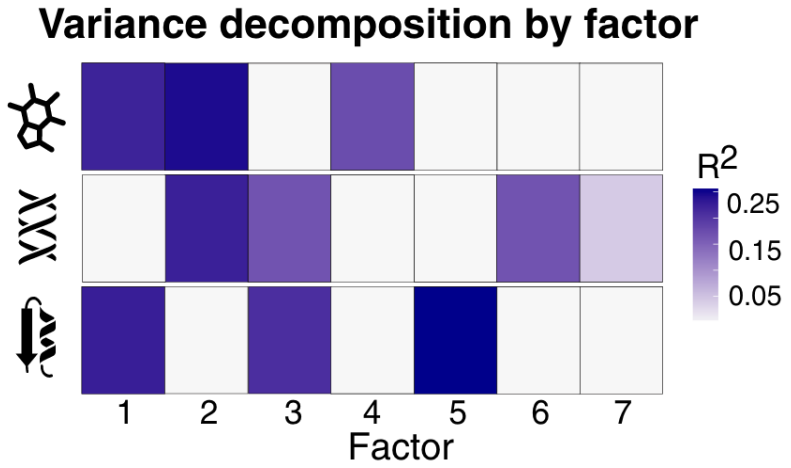


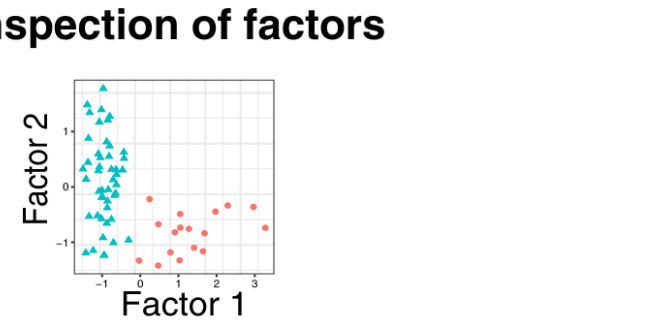
Multi-Omics Factor Analysis (MOFA)

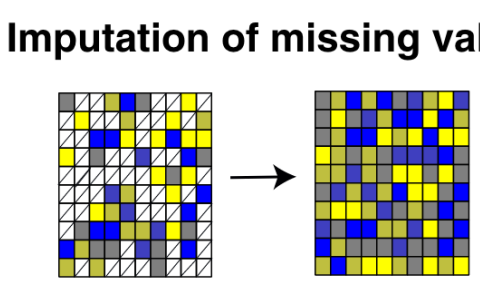


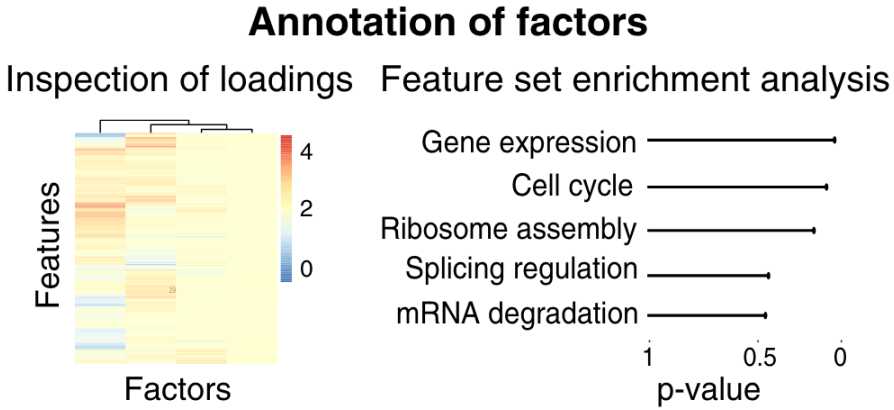
Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular systems biology* 14.6 (2018): e8124.

MOFA Downstream analysis









Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular systems biology* 14.6 (2018): e8124.

Application to a chronic lymphocytic leukemia (CLL) data set

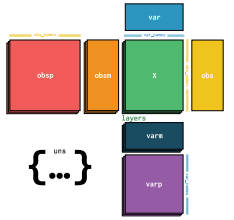
Somatic mutations	200 patients	69 mutation loci
Transcriptome	136 patients	5000 transcripts
DNA methylation	200 patients	200 methylation sites
Ex vivo drug response	200 patients	62 drugs (5 concentrations)

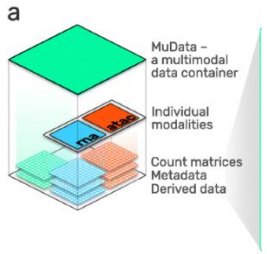
Patient metadata:

- Gender
- Age
- IC50 before treatment
- Treated after data collection
- Survival

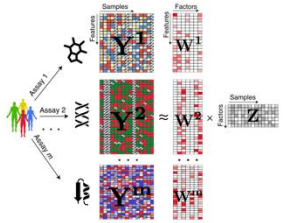
Project Overview

Tuesday Data handling and the CLL data set

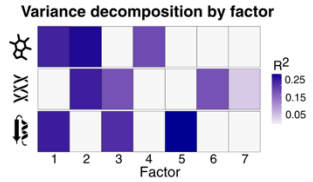




Wednesday Training a MOFA model



Thursday Downstream analysis

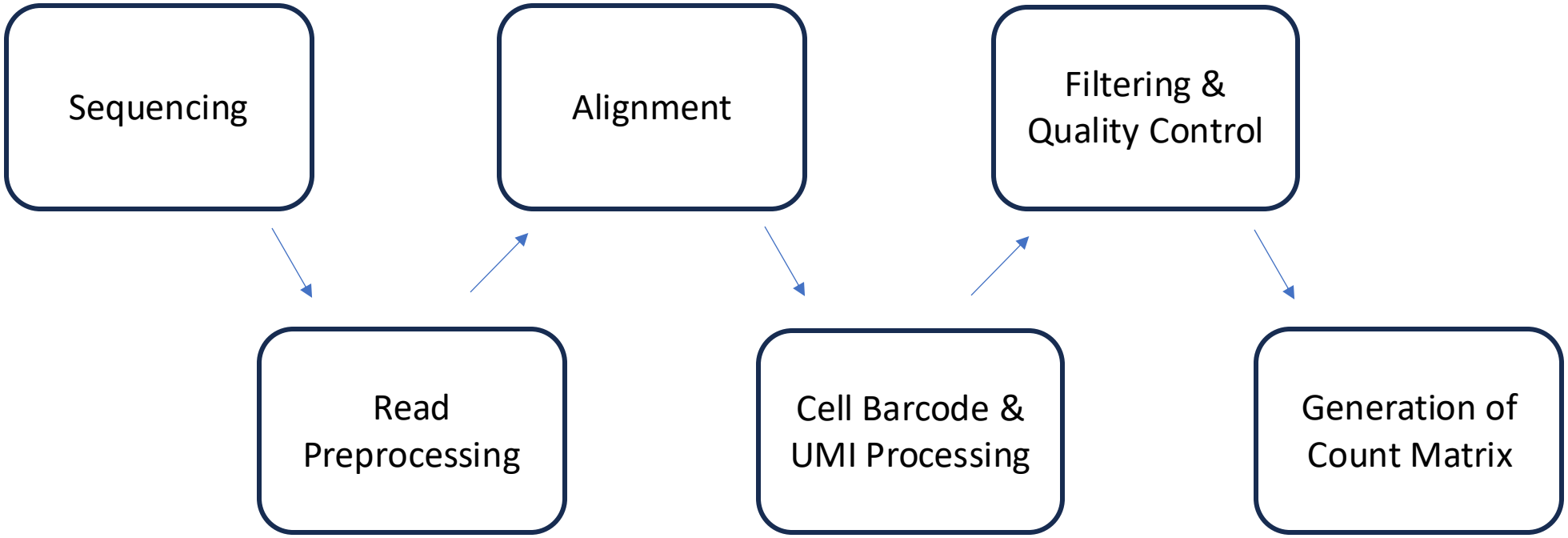


Friday Time for questions and presentation preparation

Data Handling

AnnData, MuData, and the CLL data set

From the Sequencing Machine to Count Matrices



How to Represent Count Matrices (in Python)

Array

(Numpy)

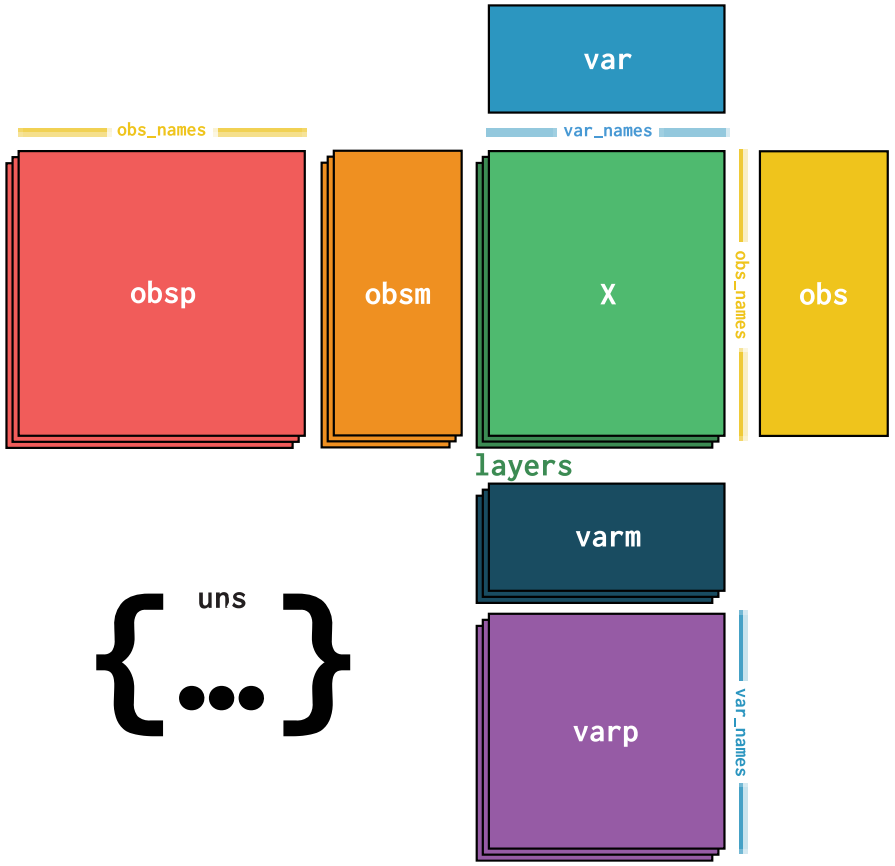
2	5	…	0
0	3	…	8
…	…	…	..
2	1	…	0

DataFrame

(Pandas)

	Cell 1	**Cell 2**	**…**	**Cell N**
Gene 1	2	5	…	0
Gene 2	0	3	…	8
…	…	…	…	..
Gene D	2	1	…	0

AnnData



https://anndata.readthedocs.io/

Virshup, Isaac, et al. "anndata: Annotated data." *BioRxiv* (2021): 2021-12.

How to Create Your Own AnnData Object

Manually

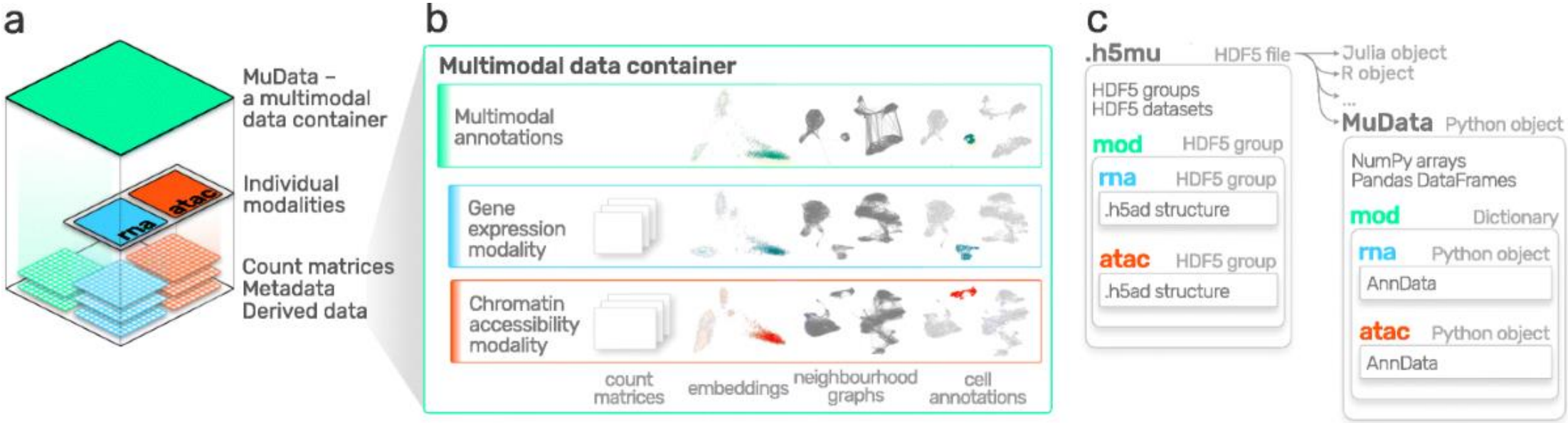
```
adata_X = np.array(...)
adata_obs = pd.DataFrame(...)
adata_var = pd.DataFrame(...)

adata = ad.AnnData(
  X=adata_X,
  obs=adata_obs,
  var=adata_var,
  )
```

With ScanPy

```
sc.read(...)
sc.read_10x_h5(...)
sc.read_10x_mtx(...)
sc.read_visium(...)
sc.read_h5ad(...)
sc.read_csv(...)
sc.read_excel(...)
sc.read_hdf(...)
sc.read_loom(...)
sc.read_mtx(...)
sc.read_text(...)
sc.read_umi_tools(...)
```

https://scanpy.readthedocs.io/en/1.10.x/api/reading.html

MuData



https://muon.readthedocs.io

Bredikhin, Danila, Ilia Kats, and Oliver Stegle. "MUON: multimodal omics analysis framework." *Genome biology* 23.1 (2022): 42.

The CLL Data Set

Somatic mutations	200 patients	69 mutation loci	Binary encoding (0 or 1)
Transcriptome	136 patients	5000 transcripts	Transformed counts
DNA methylation	200 patients	200 methylation sites	M-value
Ex vivo drug response	200 patients	62 drugs (5 concentrations)	Viability score (0 to 1)

Patient metadata:

- Gender
- Age
- IC50 before treatment
- Treated after data collection
- Survival

Factor Models and MOFA

An introduction

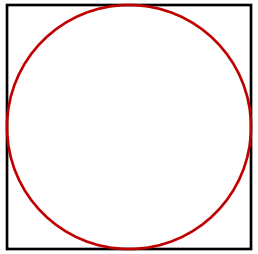
Omics Data is High-Dimensional

Modality	**Number of features / dimensions**
Proteome	e.g. 10 000 proteins
Transcriptome	e.g. 20 000 genes
Genome	e.g. 5 million SNPs
Epigenome	e.g. 20 million CpG sites

... often in just 100s to 1000s of cells / samples →

$$n_{dimensions} \gg n_{observations}$$

The Curse of Dimensionality

D = 1	D = 2	D = 3	D = ...
			
$\frac{1}{1} = 1$	$\frac{\pi \cdot 0.5^2}{1} \cong 0.78$	$\frac{\frac{4}{3}\pi \cdot 0.5^3}{1} \cong 0.52$	$\cong 0$

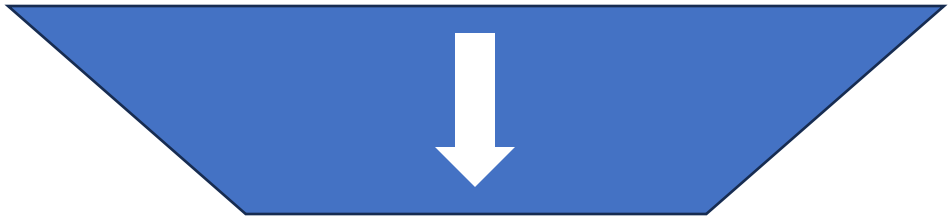
In very high dimensions...

- spheres around data points fill vanishingly small volumes
- it becomes difficult to establish relations between data points

→ High dimensions are typically not suitable for direct analysis

Dimensionality Reduction Methods

	Gene 1	Gene 2	Gene 3	...	Gene D
Cell 1	2	5	2	…	0
…	…	…	…	…	..
Cell N	2	1	0	…	0



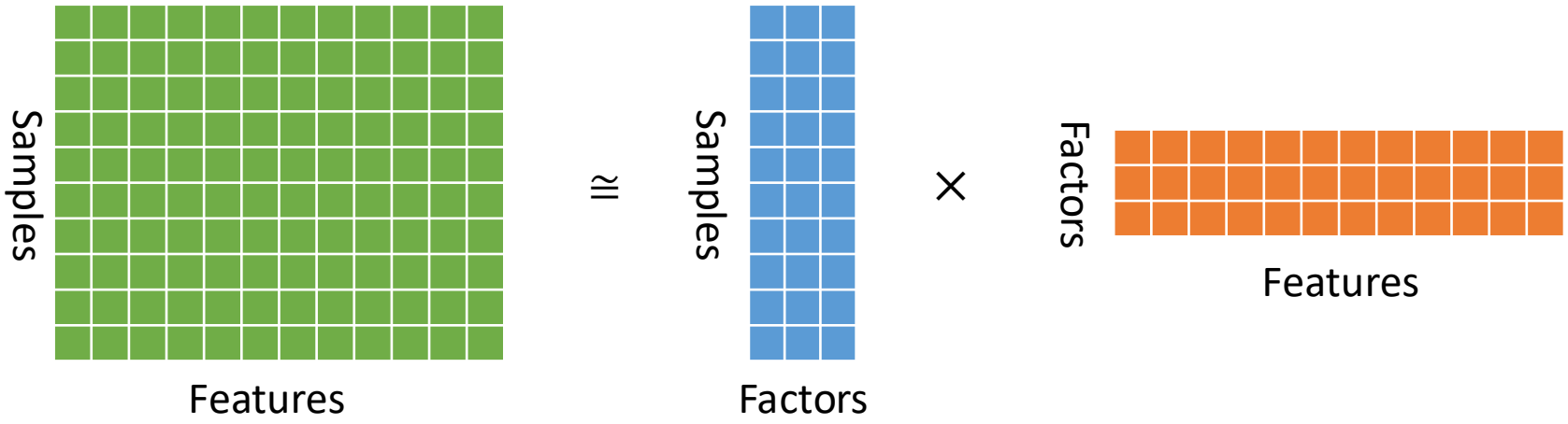
	Dim 1	Dim 2
Cell 1	0.4	-6.2
…	…	…
Cell N	0.0	9.1

Linear Methods
Principal Component Analysis (PCA)
Independent Component Analysis (ICA)
Latent Dirichlet Allocation (LDA)
Factor Analysis (FA)
Non-Negative Matrix Factorization (NMF)
...

Non-Linear Methods
(Variational) Autoencoder (VAE)
Deep Matrix Factorization
t-SNE
UMAP
Spectral Embedding
...

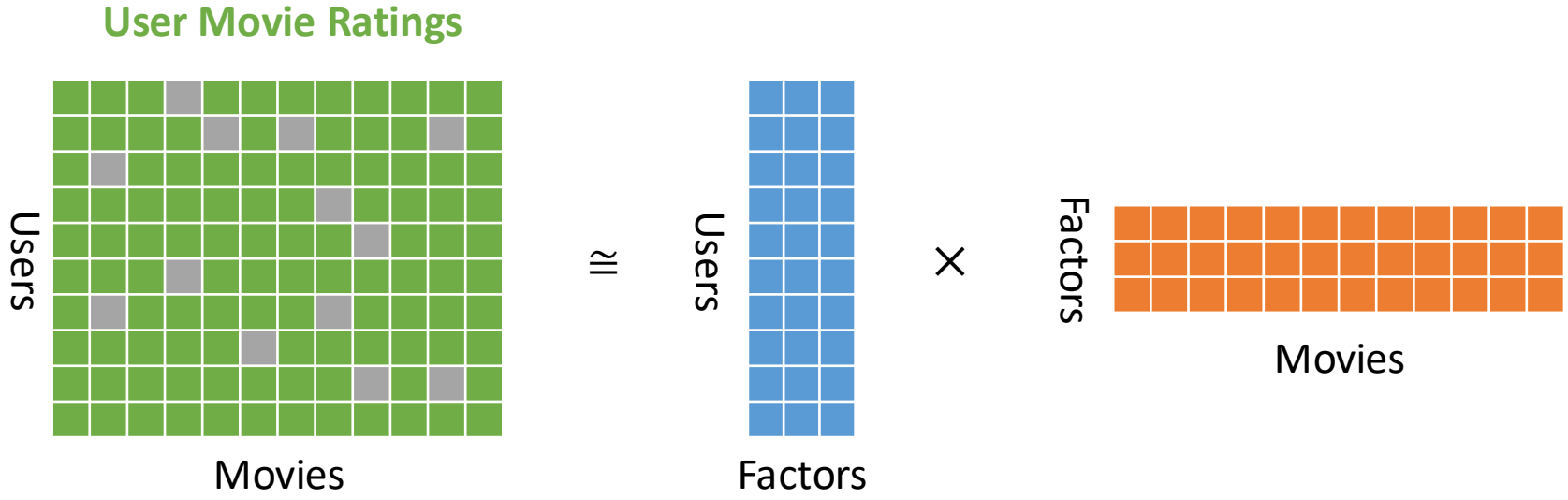
What is a Factor Model Intuitively?

- **Factors** can be seen as **meta-features** that summarise the behaviour of groups of features
- The reduced **data** is represented as **factor scores** (a matrix of dimensions $n_{samples} \times n_{factors}$
- Factors are linked to all the original features via **factor loadings** (a matrix of dimensions $n_{factors} \times n_{features}$)



An Example: Movie Recommendations

- A streaming service has access to the **star ratings** its users have given for different **movies**.
- The service wants to know **how much a user would like another movie** to provide better recommendations
- What could the factors represent in this situation? Do they always represent something “real”?
- What about positive and negative factor scores and loadings?
- How could movie ratings be predicted?



What is a Factor Model Mathematically?

- Factor scores and loadings are called **latent variables**
- Given the **observed data**, the goal is to **infer** the latent variables

Matrix Factorisation

$$y_{nd} \cong \sum_{k=1}^{K} z_{nk} w_{kd}$$

$Y \in \mathbb{R}^{NxD}$ Observed data
$Z \in \mathbb{R}^{NxK}$ Factor scores
$W \in \mathbb{R}^{KxD}$ Factor loadings

Probabilistic Formulation

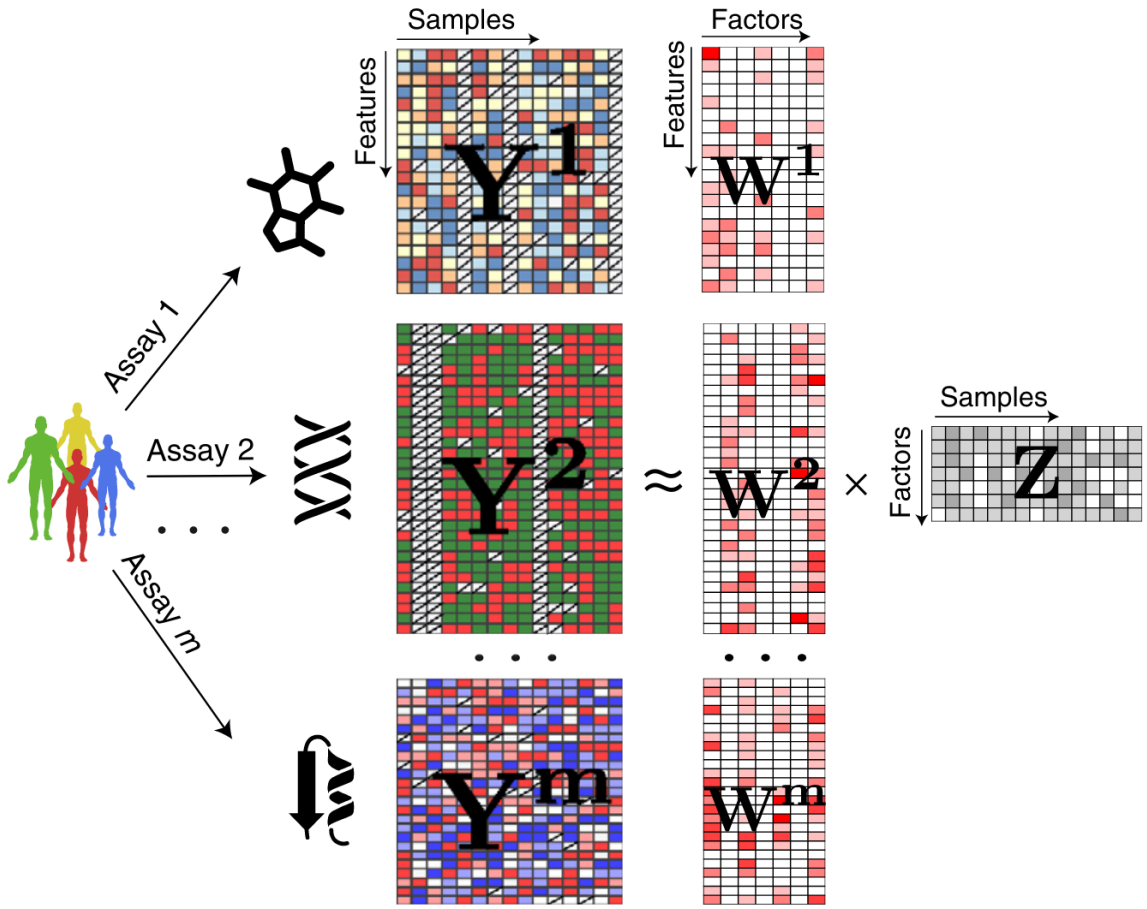
$$y_{nd} \sim \mathcal{N}(\mu_{nd}, \sigma_d^2)$$

$$\mu_{nd} = \sum_{k=1}^{K} z_{nk} w_{kd}$$

$$z_{nk} \sim p(z_{nk})$$

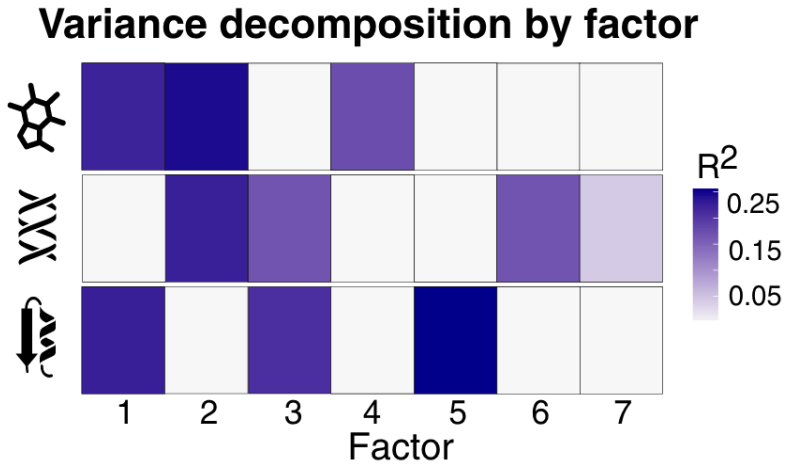
$$w_{kd} \sim p(w_{kd})$$

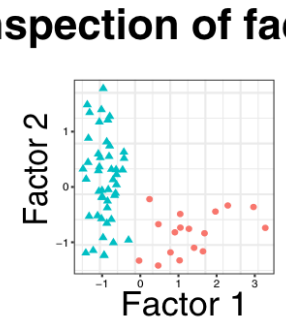
Multi-Omics Factor Analysis (MOFA)

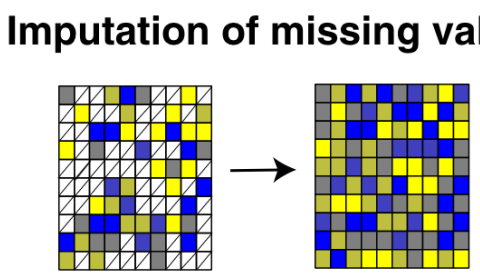


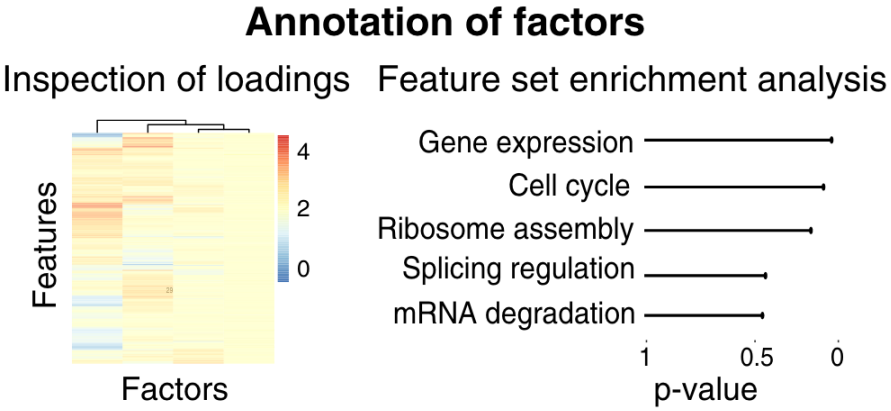
Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular systems biology* 14.6 (2018): e8124.

MOFA Downstream analysis









Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular systems biology* 14.6 (2018): e8124.

Gene Set Enrichment Analysis (GSEA)

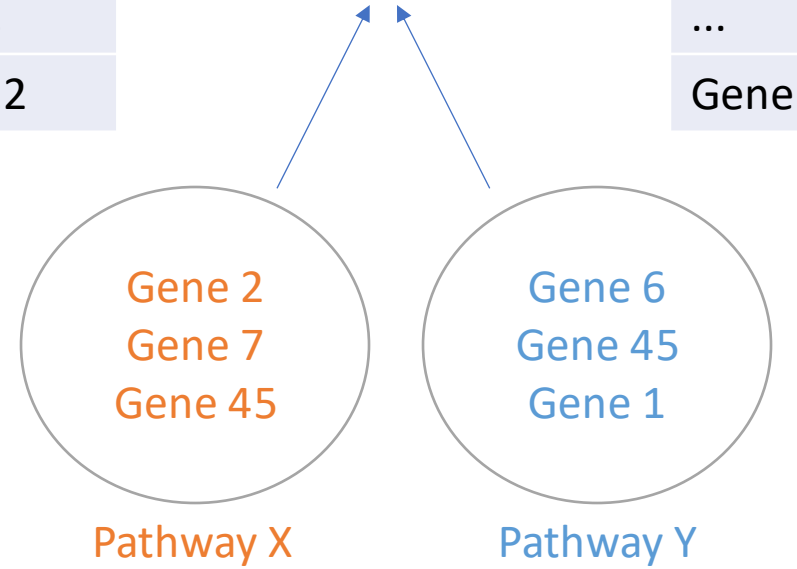
Gene 1	0.2
Gene 2	-5.1
Gene 3	1.1
…	…
Gene D	4.0

Sorting

Gene 2	-5.1
Gene 7	-4.9
Gene 45	-4.5
…	…
Gene 9	5.2

GSEA

Gene 2	-5.1
Gene 7	-4.9
Gene 45	-4.5
…	…
Gene 9	5.2



Bayesian Latent Variable Models

Generative Model
("Telling a story of the data")

Data → $y_{nd} \sim \mathcal{N}(\mu_{nd}, \sigma_d^2)$ ← Data Likelihood

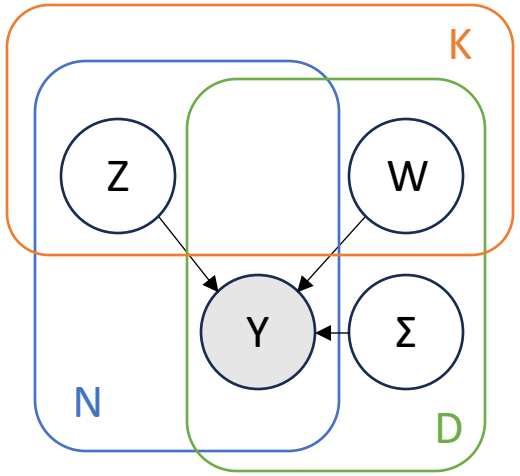
$$\mu_{nd} = \sum_{k=1}^{K} z_{nk} w_{dk}$$

$$w_{dk} \sim p(w_{dk})$$
$$z_{nk} \sim p(z_{nk})$$
$$\sigma_d \sim p(\sigma_d)$$

← Prior distributions

Latent variables

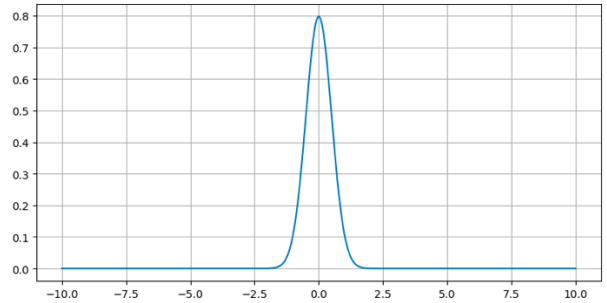
Graphical Model



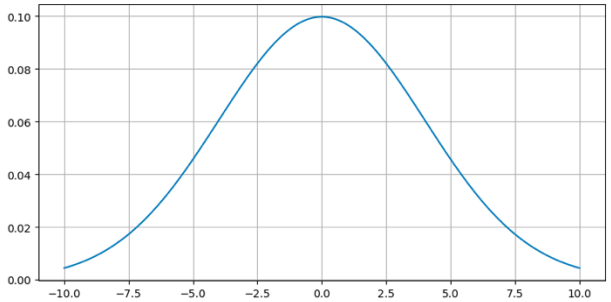
The Power of Prior Distributions

$$p(w_{dk}) = ?$$
$$p(z_{nk}) = ?$$
$$p(\sigma_d) = ?$$

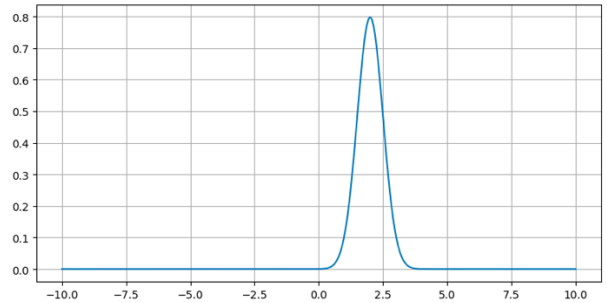
Prior distributions encode **a-priori assumptions about the variables** before seeing any data



”Values should be close to 0”



”Values could also be quite large”



”Values should (for whatever reason) be close to 2”

Bayes' Theorem

Thomas Bayes
1701 – 1761
Statistician and Presbyterian minister

Posterior **Likelihood** **Prior**

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

Marginal likelihood

Bayesian Inference

Given the prior distribution and data likelihood, the posterior should be easy to compute...

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{\cancel{p(B)}}$$

Computationally not tractable 😭

Solution: **Approximate inference**

$$q(A) \approx p(A \mid B)$$

$$\text{ELBO} = \ln(p(B)) - \text{D}_{KL}(q(A) \mid\mid p(A \mid B))$$