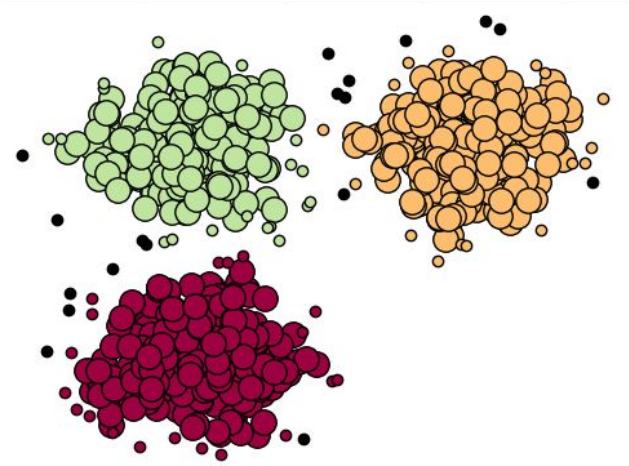
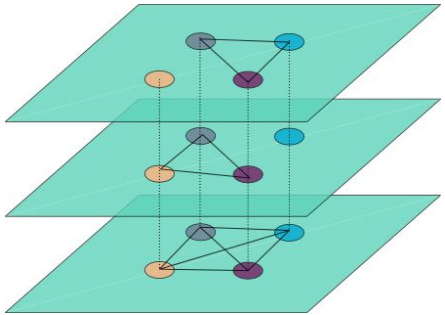


~~Clustering~~ Unsupervised Learning in Computational Biology

Karolina Sienkiewicz

Weill Cornell Medicine, Cornell University

sienkiewicz2k@gmail.com @sienkieee



Unsupervised learning

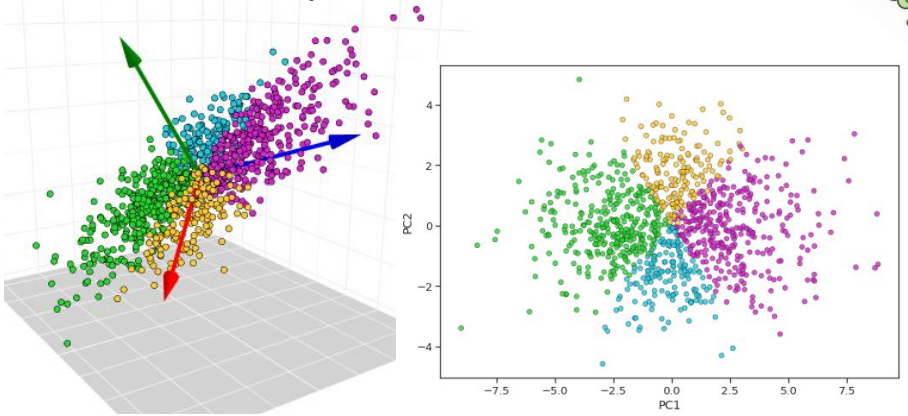
low-rank representation

clusters

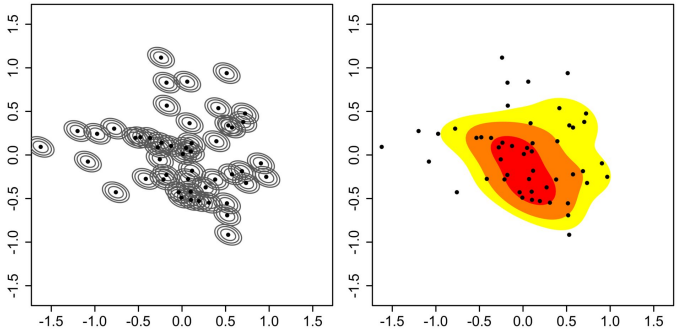
probability density estimation

$$P \approx P_{data}$$

model

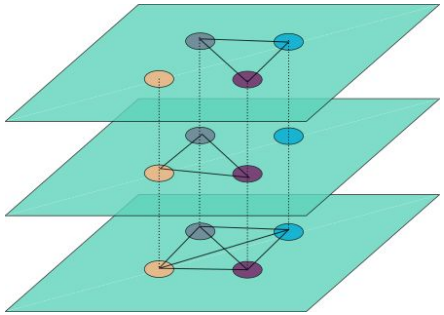


From: Casey Chang

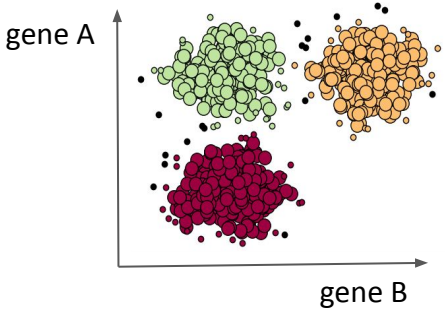


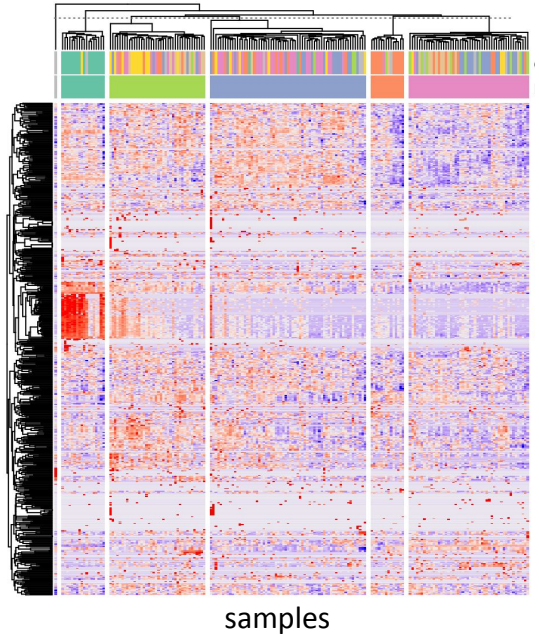
wikiwand.com/en/Multivariate_kernel_density_estimation

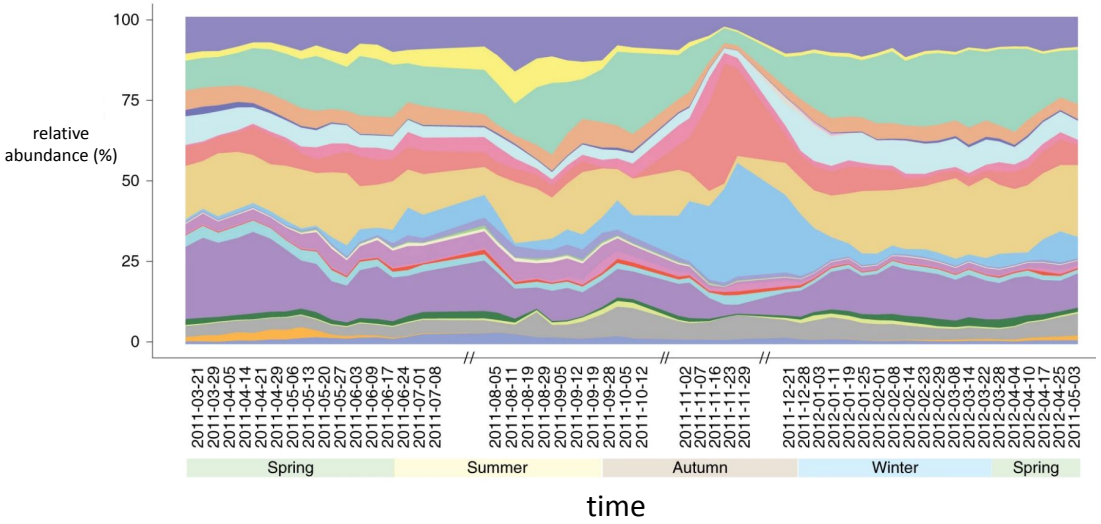
Biological datasets are complex



"omics"







PMID:
33139880

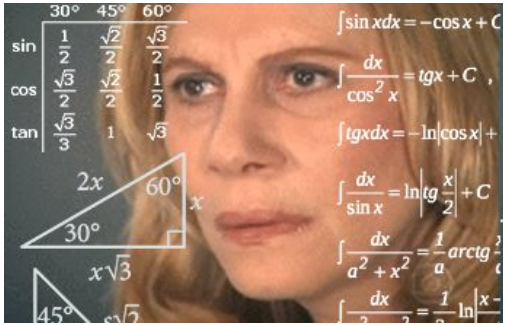
UNSUPERVISED LEARNING

Practical application #1:

Batch effect correction

Batch effect = non-biological variation

Why is there a batch effect in my data?

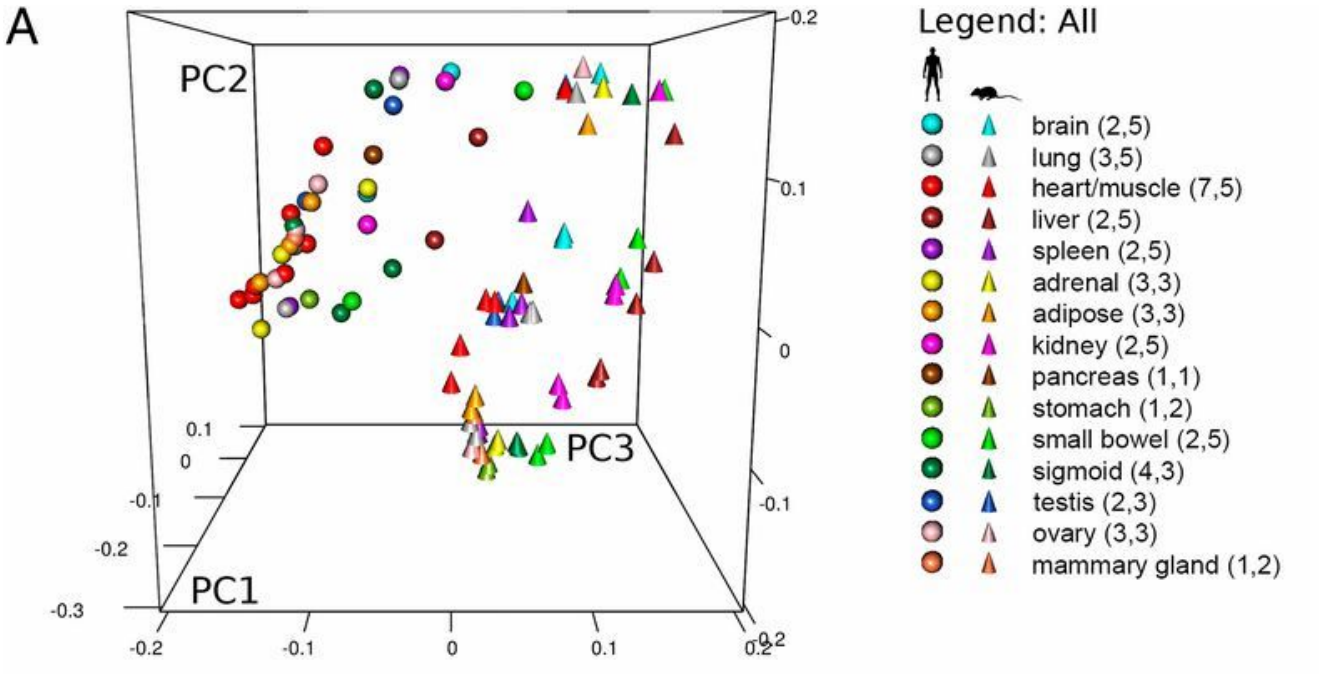


- Different days/months of the experiments
- Different reagents (enzymes, buffers)
- Different mice (from different companies)
- Different sequencers
- Lab protocol or experimenter/technicians

based on STAT115 2020 materials by Shirley Liu

Batch effect example

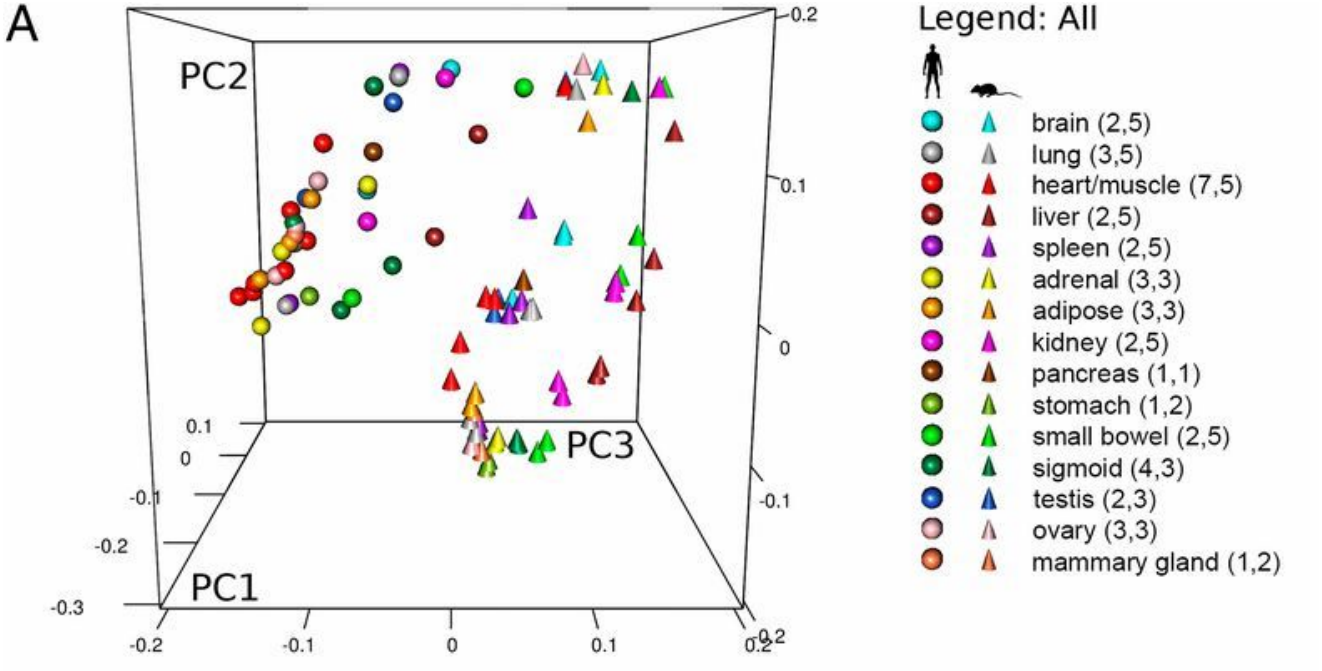
"Gene Expression Is More Similar Among Tissues Within a Species Than Between Corresponding Tissues of the Two Species"

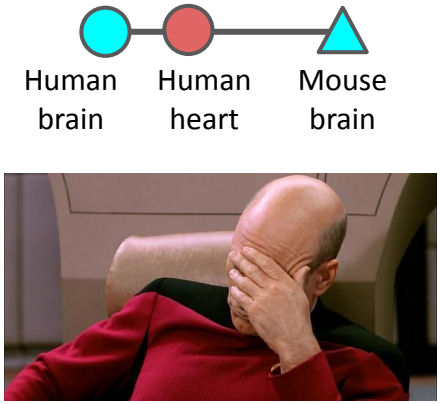


Lin et al (2014) *PNAS*

Batch effect example

"Gene Expression Is More Similar Among Tissues Within a Species Than Between Corresponding Tissues of the Two Species"





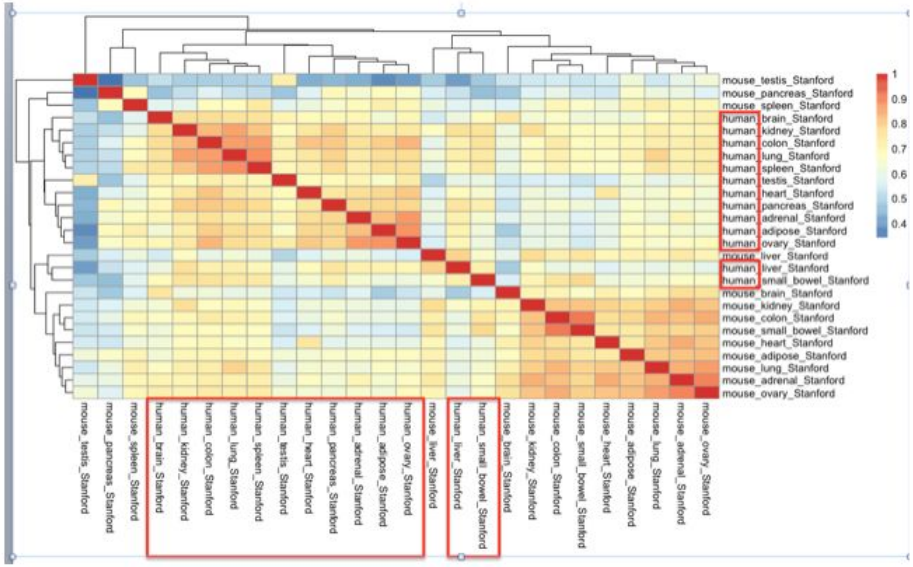
Lin et al (2014) *PNAS*

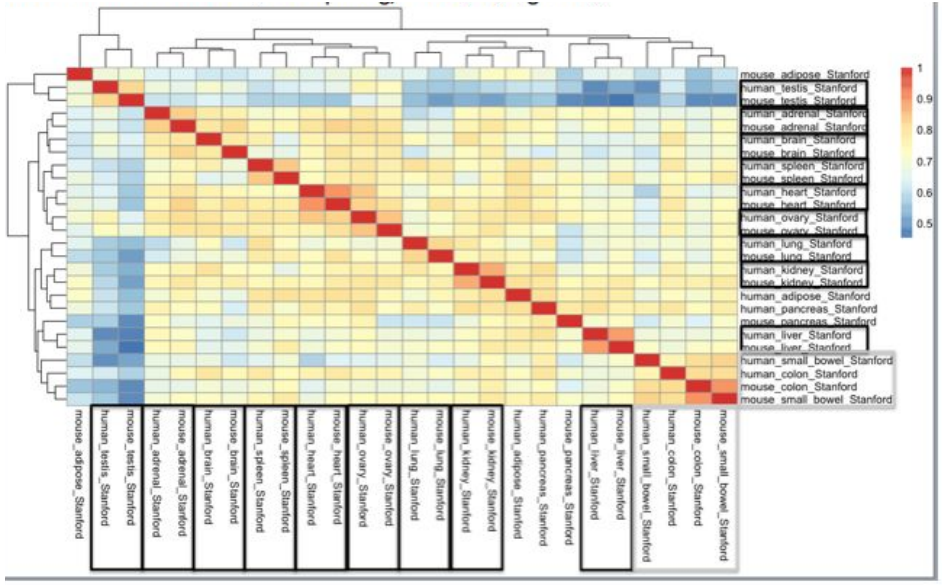
Yoav Gilad
@Y_Gilad

We reanalyzed the data from pnas.org/content/111/48... and found the following:

D87PMJN1 (run 253, lane 7)	D87PMJN1 (run 253, lane 8)	D4LHBFN1 (run 276, lane 4)	MONK (run 312, lane 6)	HWI-ST373 (run 375, lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	human
testis		pancreas		mouse

The original analysis in the paper, considering only the samples that were sequenced at Stanford (data cluster by species):



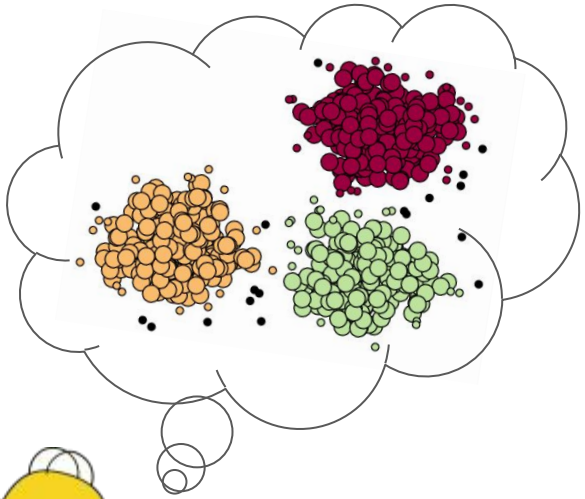


by Yoav Gilad & Orna Man

Good experimental design minimizes batch effects

consistent
pre-processing
of all samples

keeping track
of metadata



balancing
groups of interest

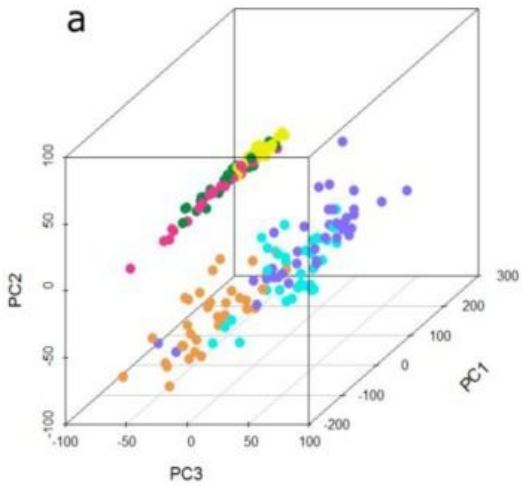
avoiding
“perfect confounding”

Combating batch effect

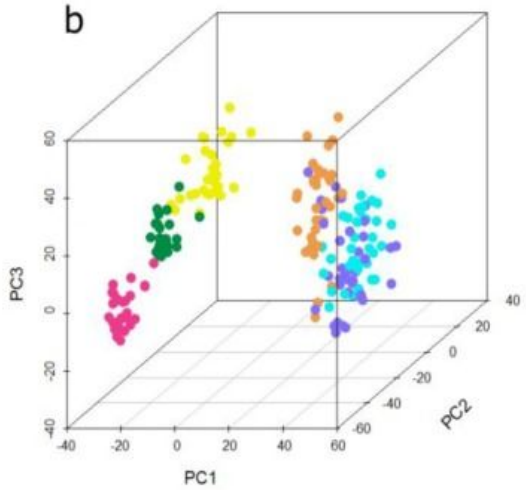
visualize/cluster → identify → correct

ComBat
Johnson et al (2007) *Biostatistics*
limma
complex (e.g. nested) batches

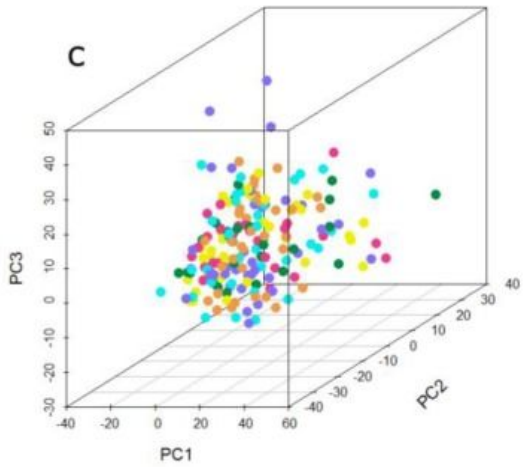
unnormalized data



normalized data



batch-corrected data



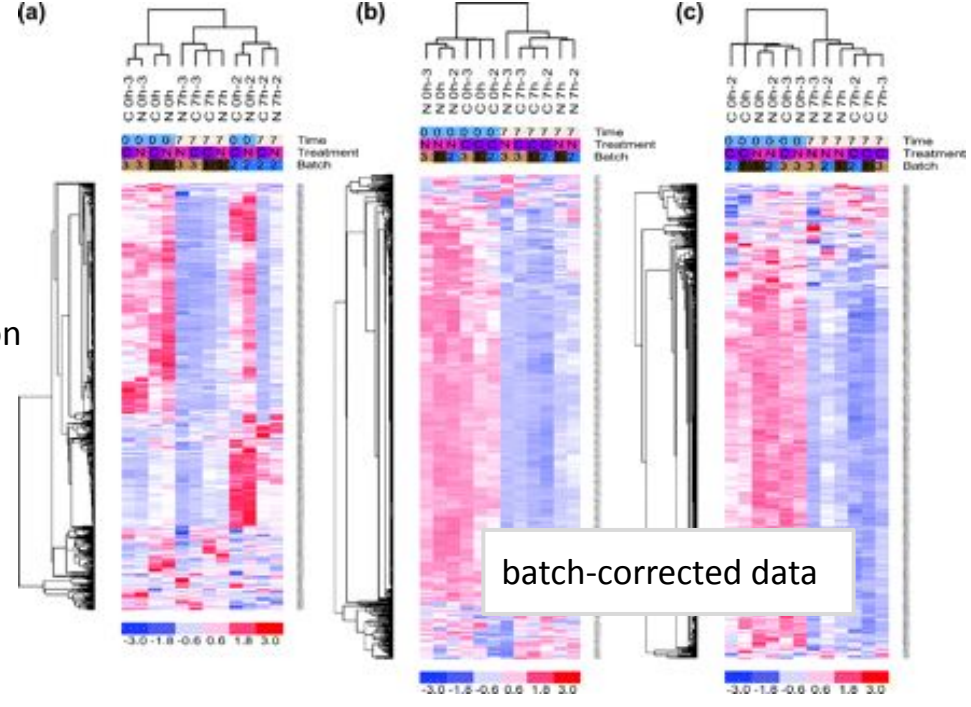
Brezina et al (2015) *Microarray*

Combating batch effect

visualize/cluster → identify → correct

ComBat
Johnson et al (2007) *Biostatistics*
limma
complex (e.g. nested) batches

before batch correction



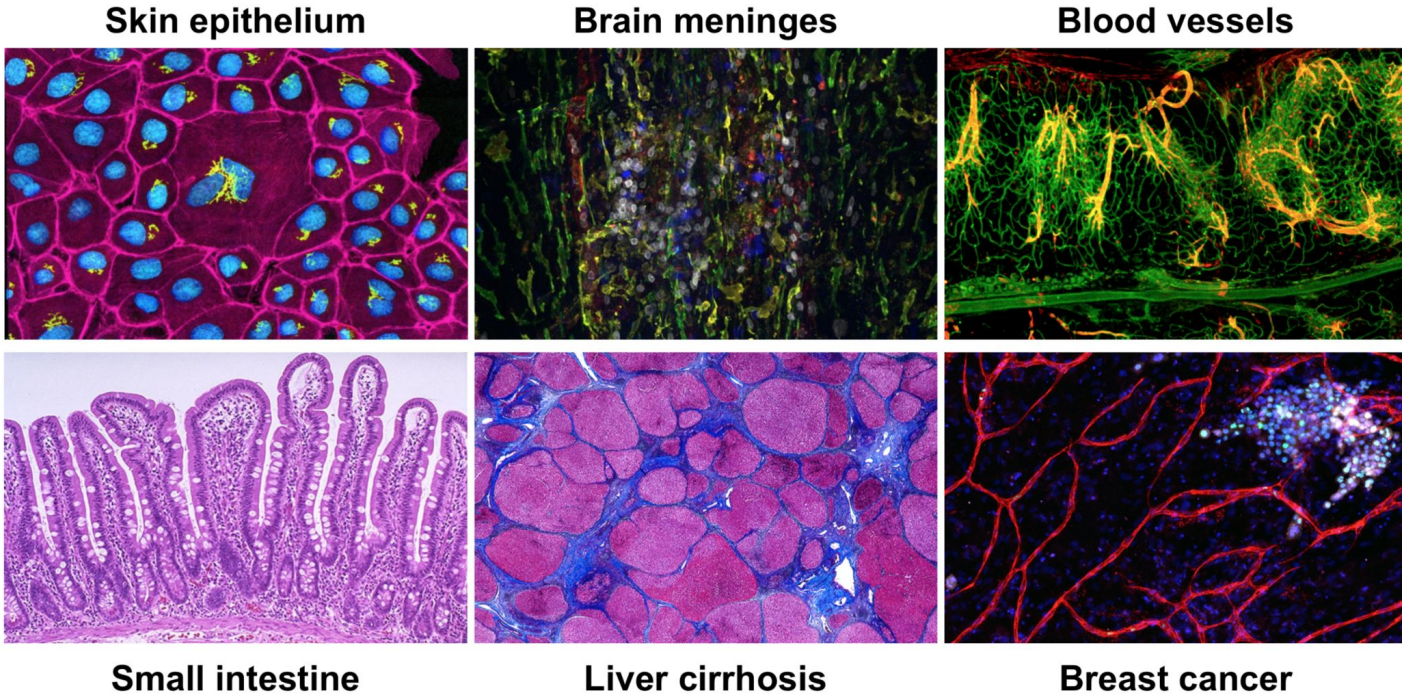
Johnson et al (2007) *Biostatistics*

UNSUPERVISED LEARNING

Practical application #2-4:

Single-cell gene expression analysis

Single-cell technologies - motivation

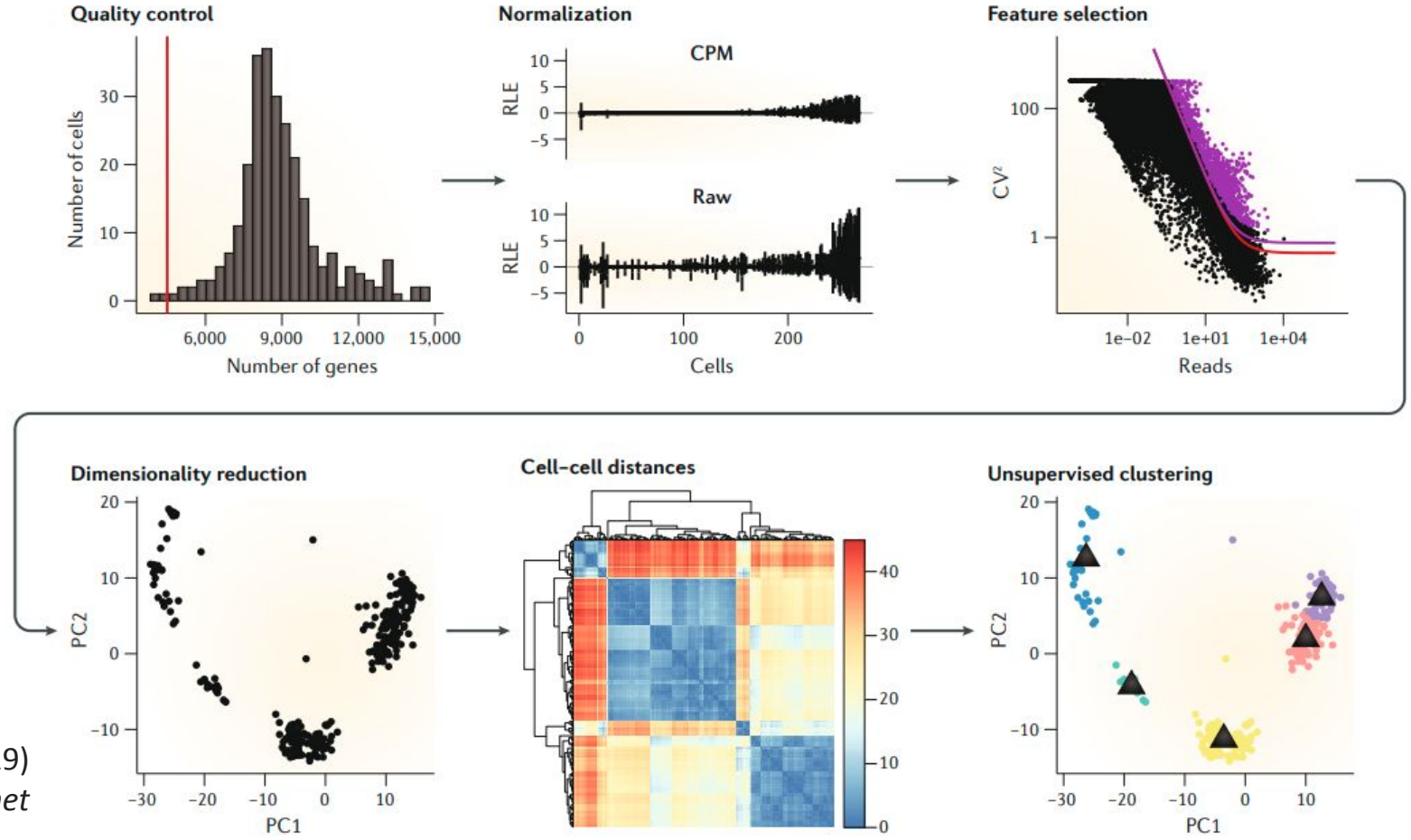


10x Genomics (2016) — >1.3 million cells

Tang et al. (2009) *Nat Methods* — 8 cells

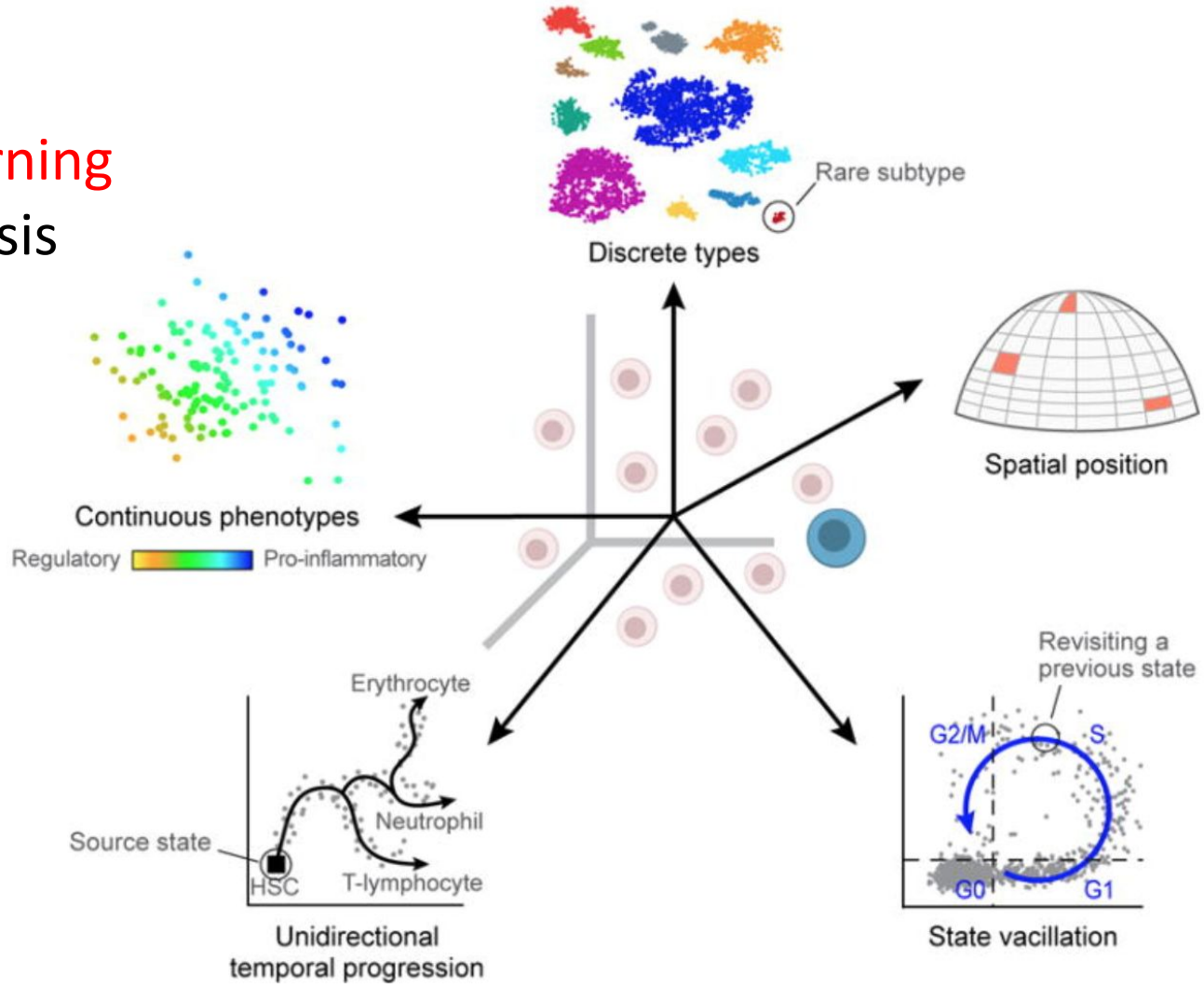
http://www.cell.com/pictureshow/skin | https://library.med.utah.edu/WebPath/webpath.html

Single-cell analysis common workflow



Kiselev et al (2019)
Nat Rev Genet

Many faces of Unsupervised Learning in single-cell analysis



Wagner et al (2016)
Nat Biotechnol

Single-cell technologies - technical considerations

cell-specific capture efficiency

amplification bias

library quality

Dropout in single-cell

stochastic transcription

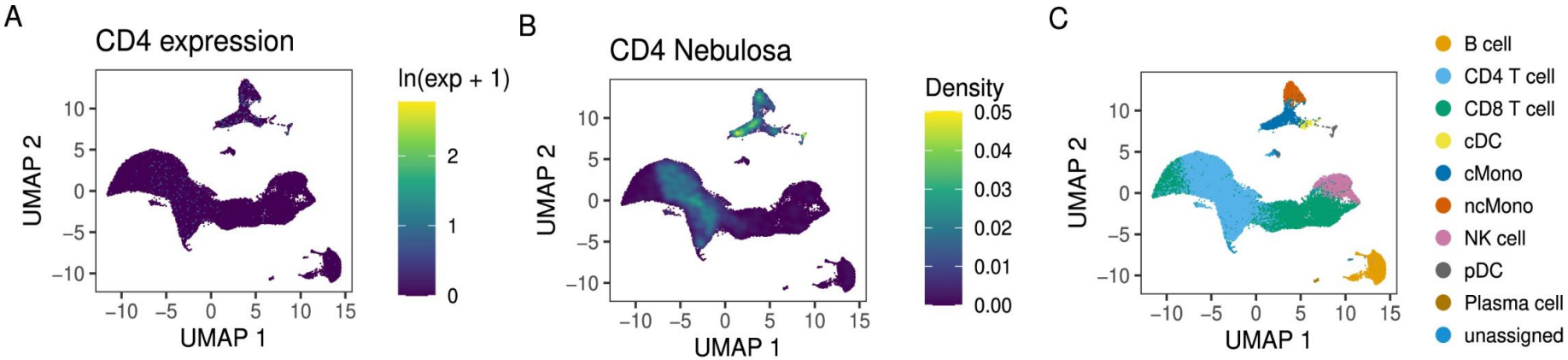
low abundance of transcripts

shallow sequence depth

Kernel density estimation example

Nebulosa
inference based on nearby cells

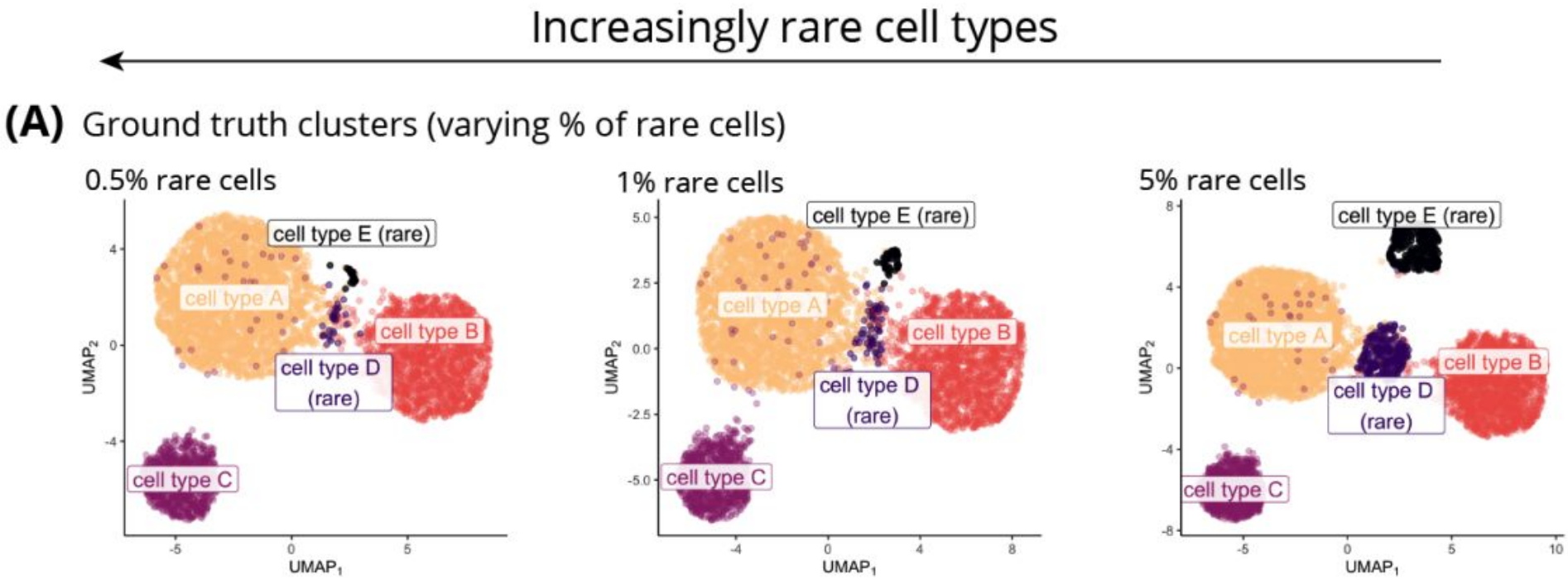
$$\hat{f}(x; H) = \frac{1}{n} \sum_{i=1}^{n} w_i K_H(x - X_i)$$



Alquicira-Hernandez & Powell (2021) *Bioinformatics*

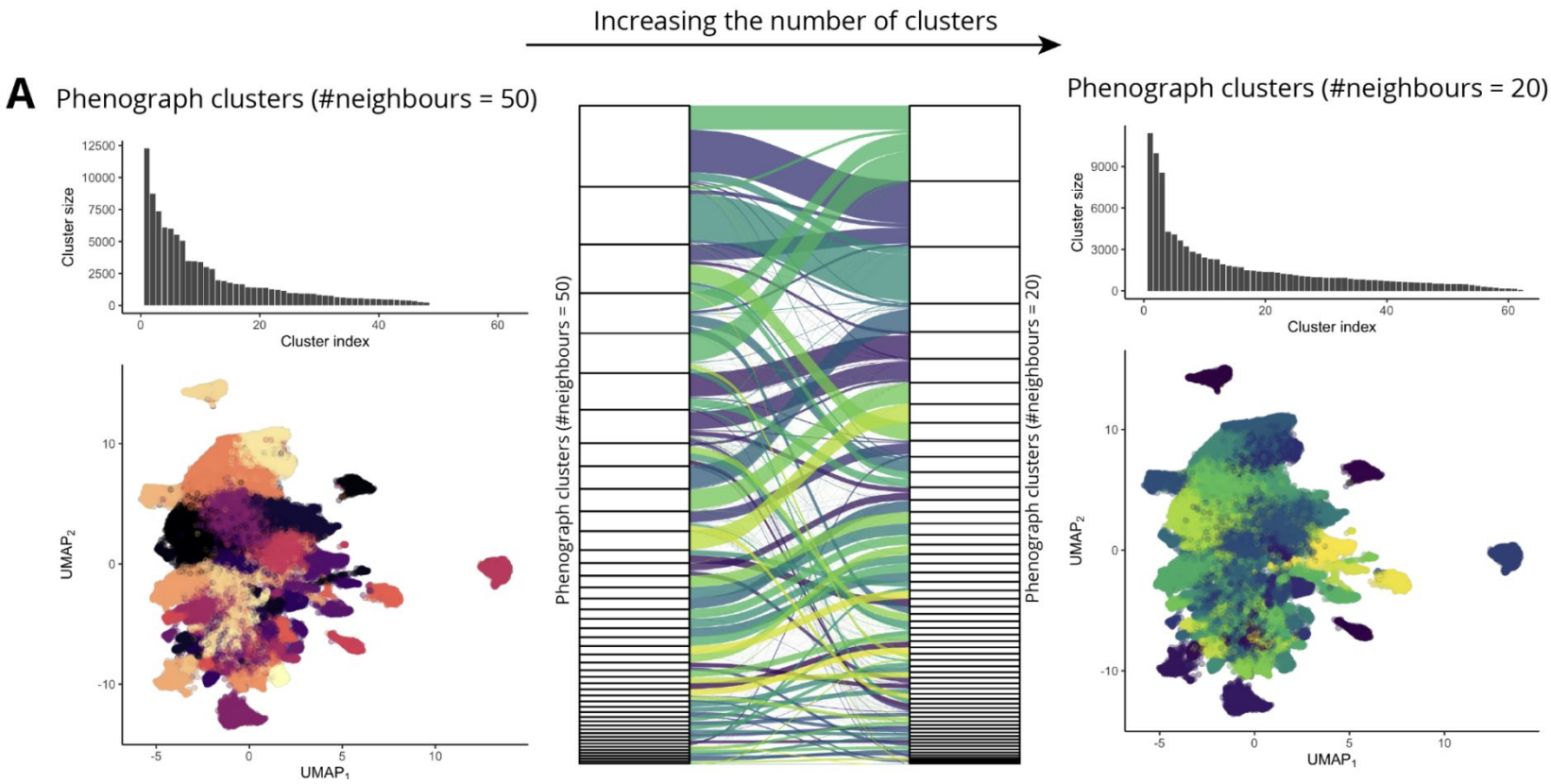
Identifying rare clusters

Detecting rare clusters is challenging

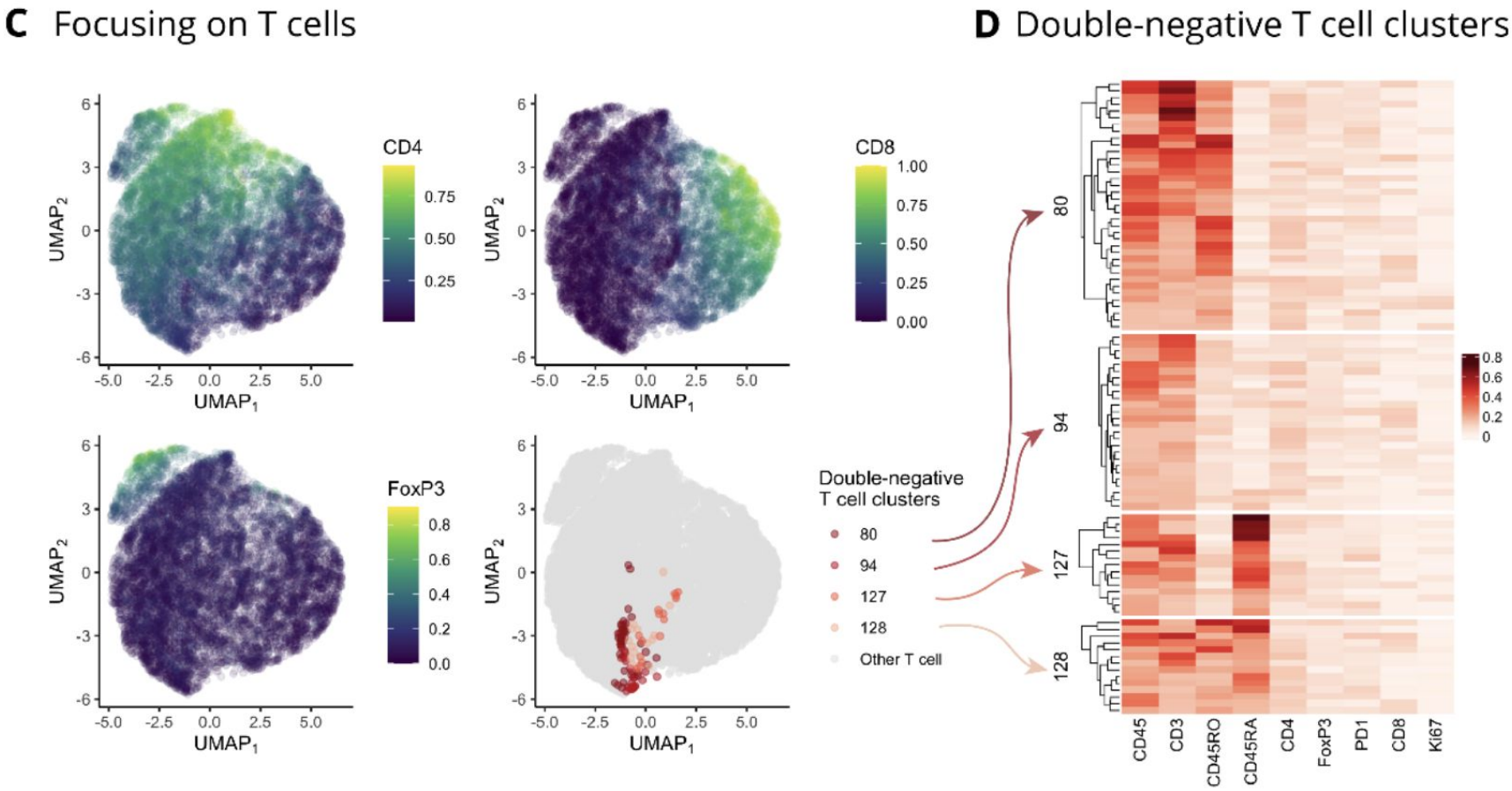


Rarity: Discovering rare cell populations from single-cell imaging data
https://www.biorxiv.org/content/10.1101/2022.07.15.500256v1

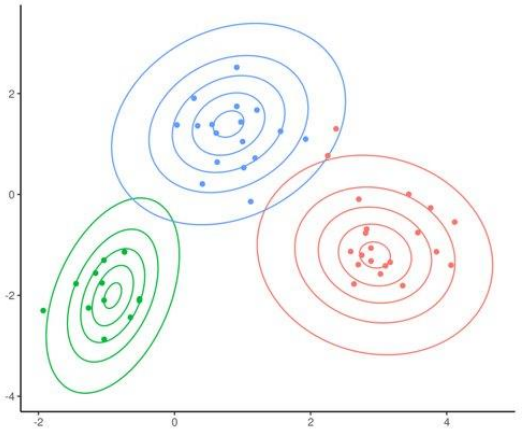
Can we simply increase *k* to increase granularity of clusters?



Example: discovering rare CD4- CD8- T cell clusters



The everlasting question:
How to choose “k” in clustering models



Can we circumvent choosing the number of clusters?

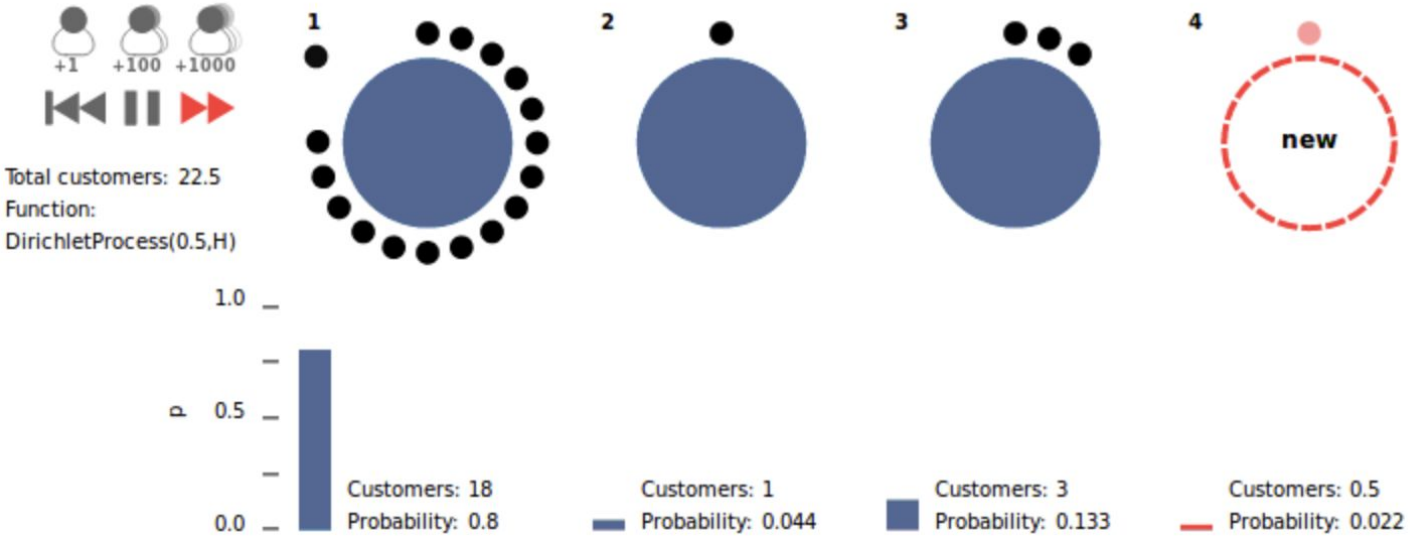
To some extent: there is an entire research field "**Bayesian nonparametrics**" trying to circumvent the issue by allowing a flexible (potentially infinite) number of clusters

Desirable properties of Bayesian nonparametric models:

- When new data comes in, ***new clusters*** should be ***allowed to appear*** (i.e. the number of clusters should not be fixed)
- We might want to capture ***uncertainty*** about the number of clusters

Flexible number of clusters with the Chinese Restaurant Process

We could use a Gaussian Mixture Model, where the number of clusters is specified by the Chinese Restaurant Process (thus allowed to be flexible)



See gif in https://commons.wikimedia.org/wiki/File:Chinese_Restaurant_Process_for_DP(0.5,H).gif

Probabilistic and non-linear extensions of PCA

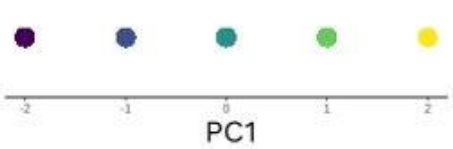
Probabilistic PCA: turning PCA into a generative model

PCA

Observed data

PCA

Low-dimensional latent representation

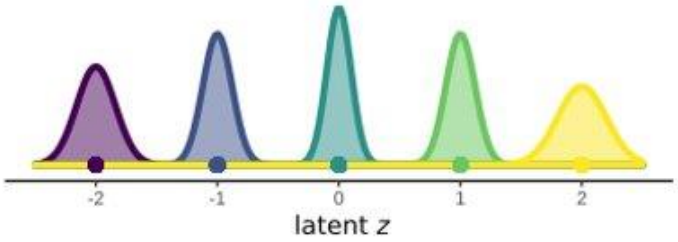


Probabilistic PCA

Observed data

Probabilistic PCA

Low-dimensional latent representation



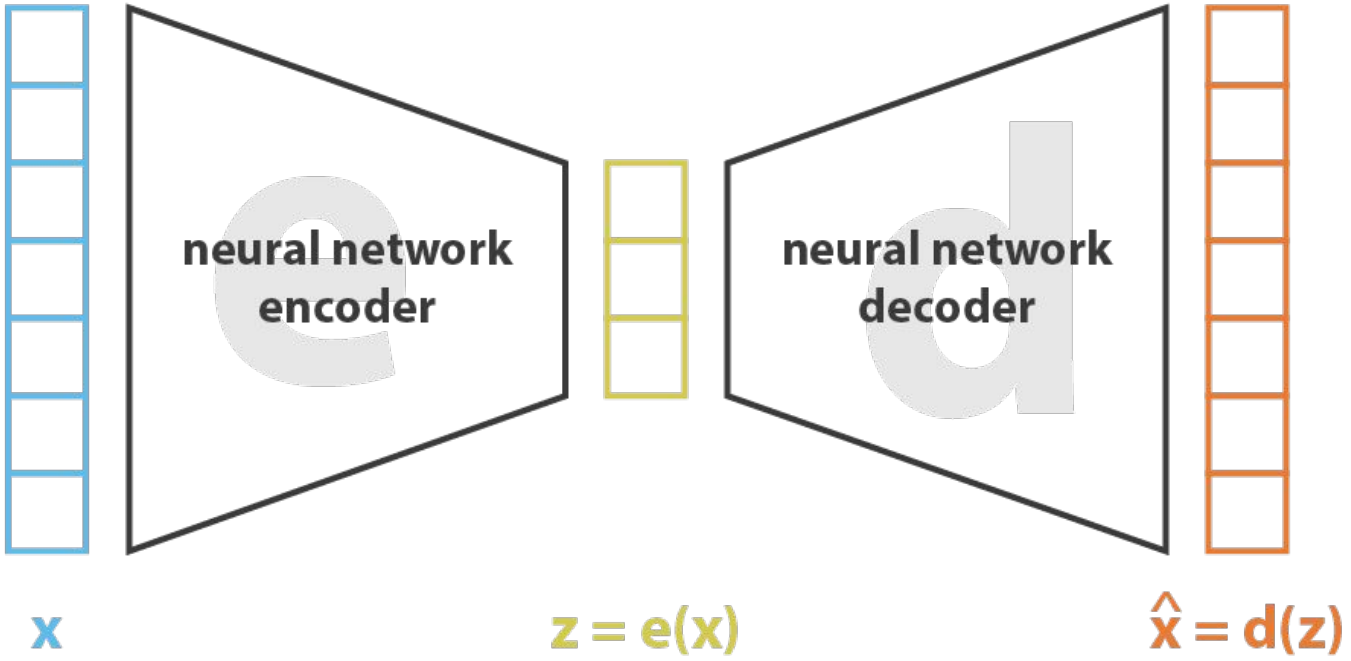
Advantages of Probabilistic PCA

- Interpretation as a ***generative*** model
- **Uncertainty** in latent space
- Can handle **missing data**
- Can "automatically" choose the latent space dimension
- Naturally extendable to **binary/count/categorical data**

Citation: Tipping and Bishop, 1999

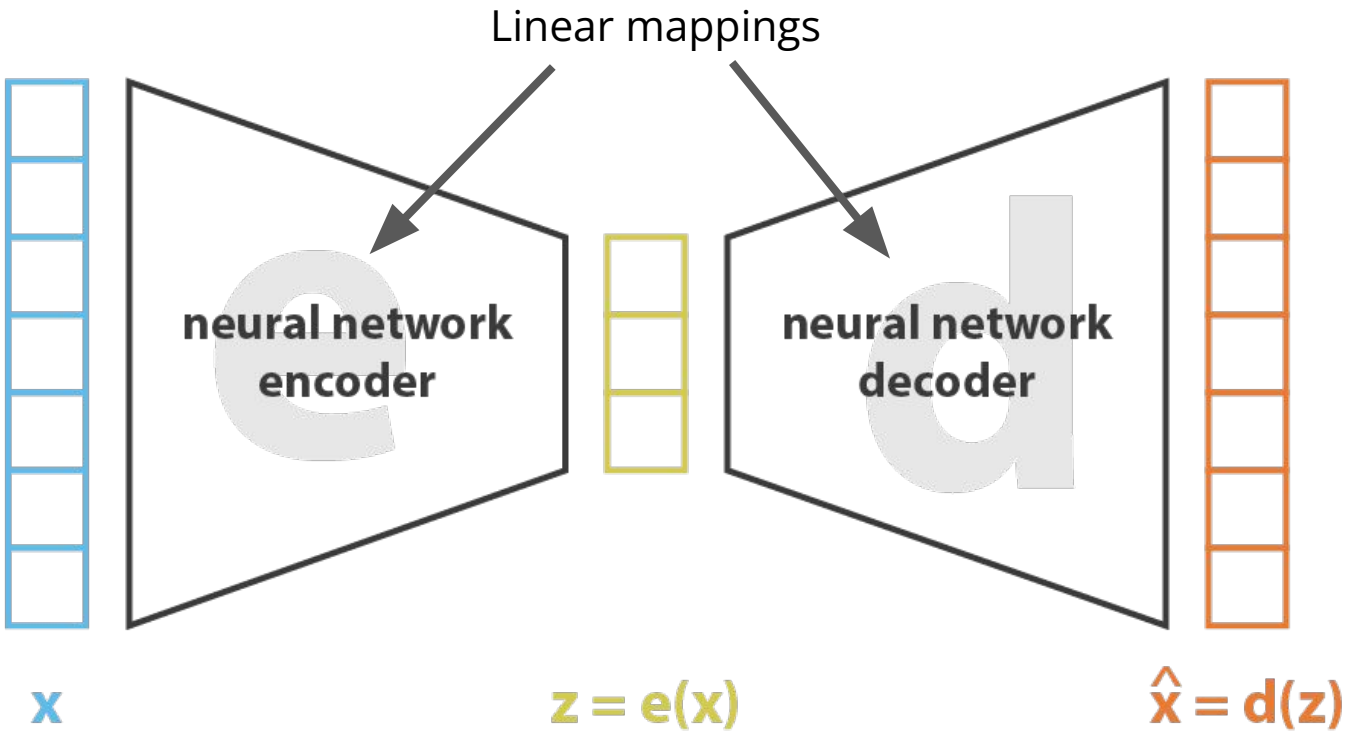
From PCA to Autoencoders

What is an autoencoder?

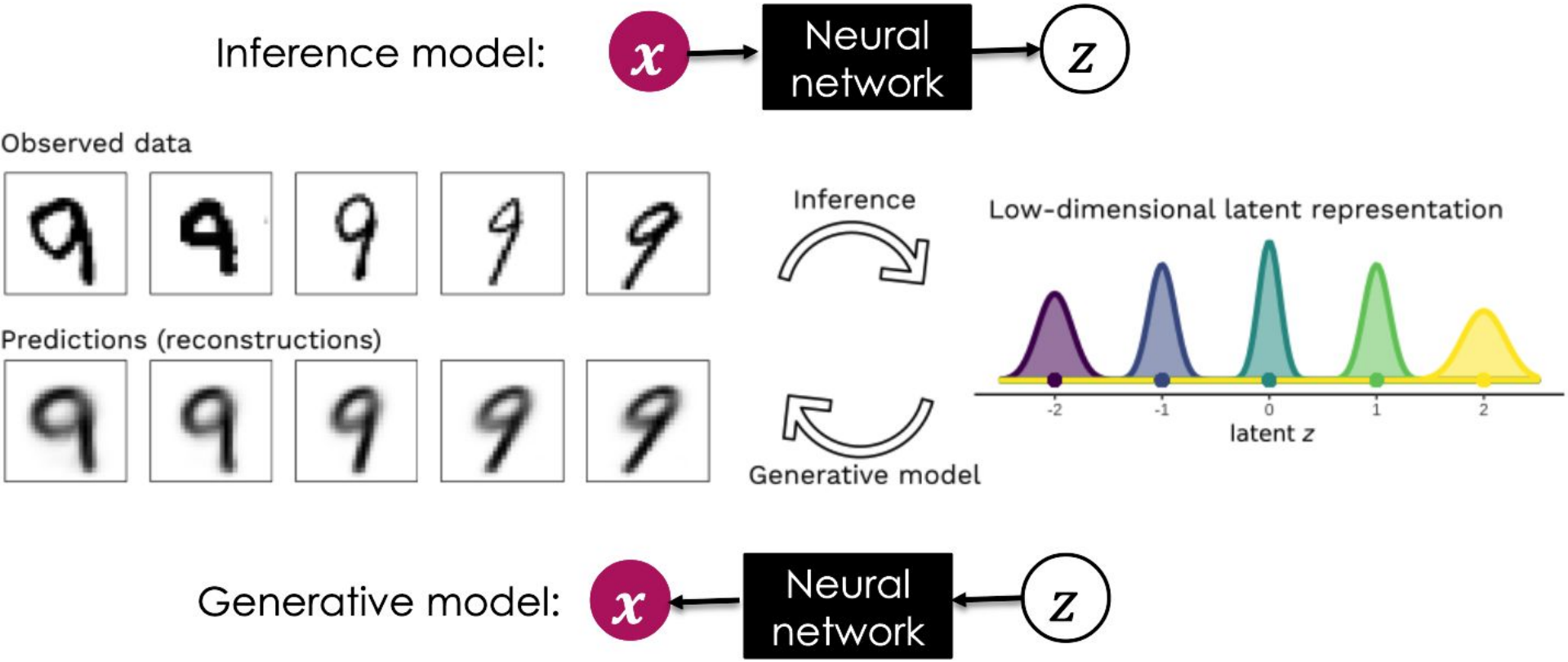


From PCA to Autoencoders

Linear autoencoder implements PCA

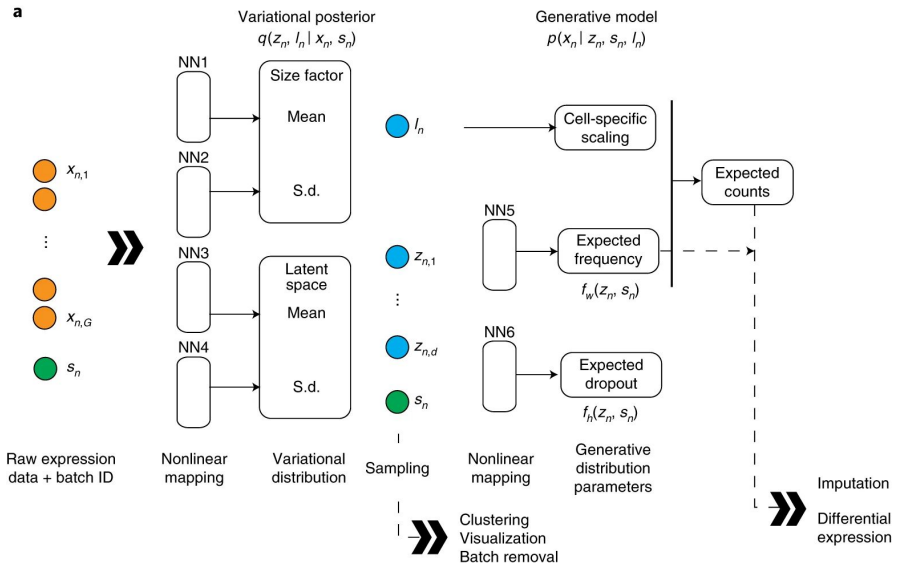


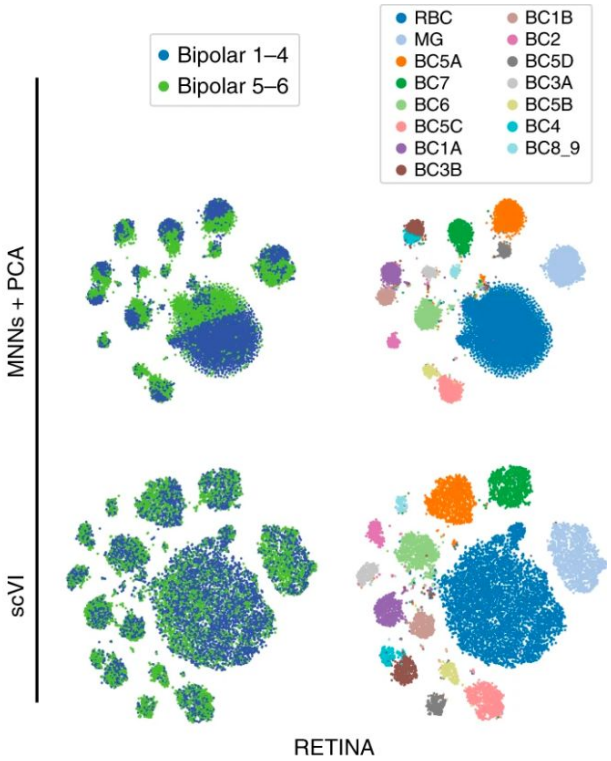
Variational Autoencoders are a non-linear extension of probabilistic PCA



Variational Autoencoders for single cell transcriptomics

Model setup/architecture:

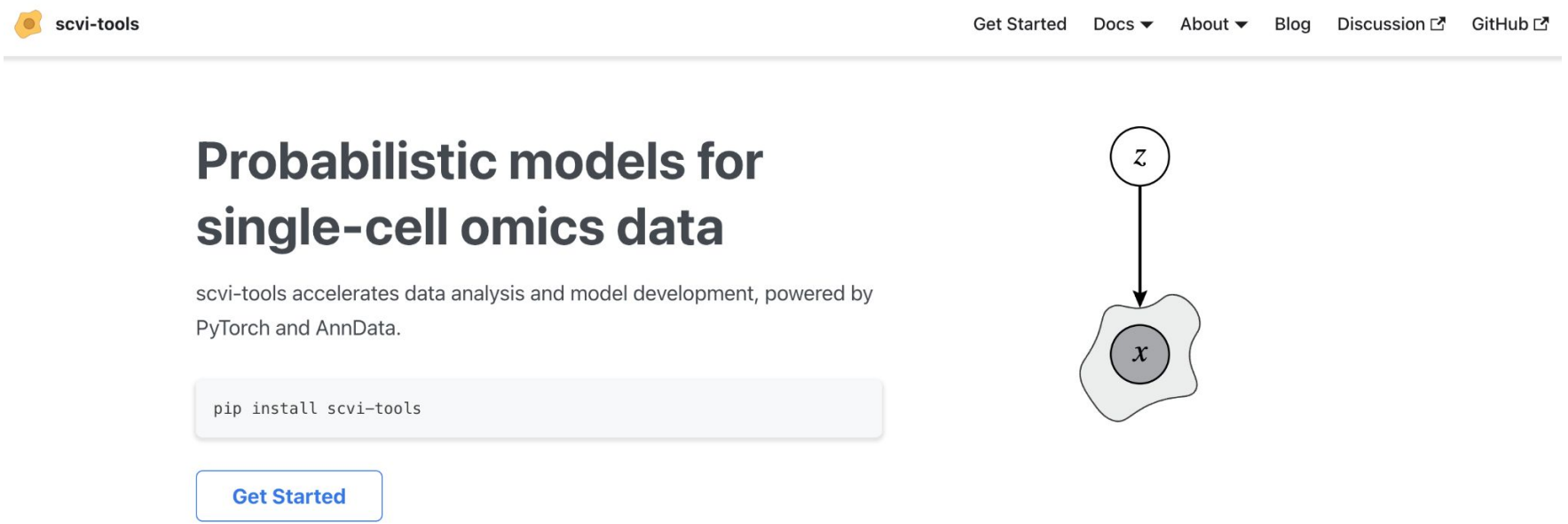




Deep generative modeling for single-cell transcriptomics https://www.nature.com/articles/s41592-018-0229-2

scVI-tools

VAE based methods can be useful for large single-cell omics data sets



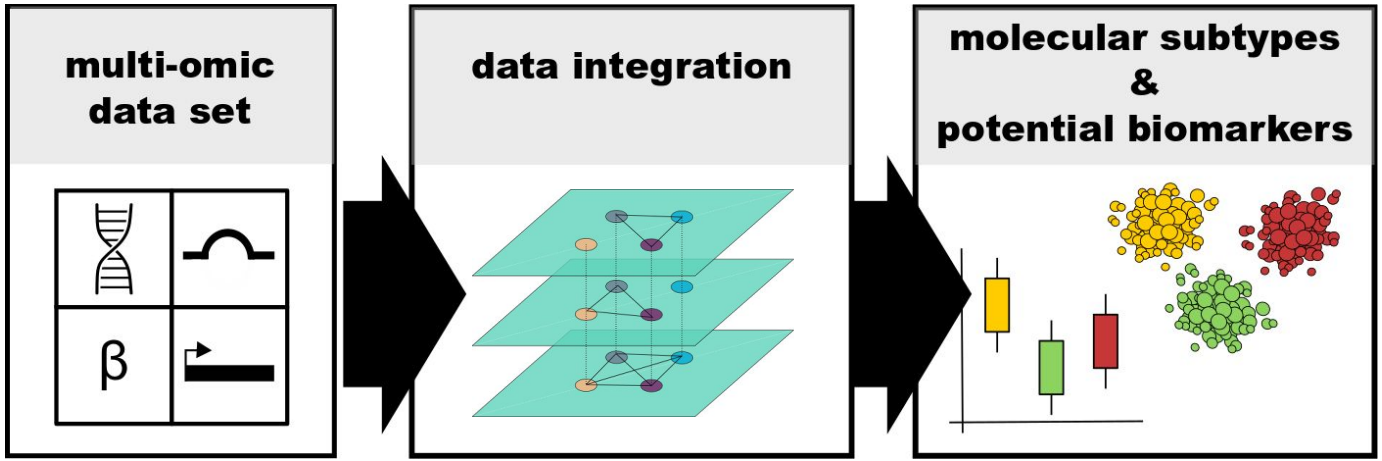
https://docs.scvi-tools.org/en/stable/tutorials/index.html

UNSUPERVISED LEARNING

Practical application #3:

Multi-omic data integration

Why are we integrating multi-omic datasets?



What are the challenges of multi-omic data integration?

What are the challenges of multi-omic data integration?

Data heterogeneity

Different number
of samples per data-type

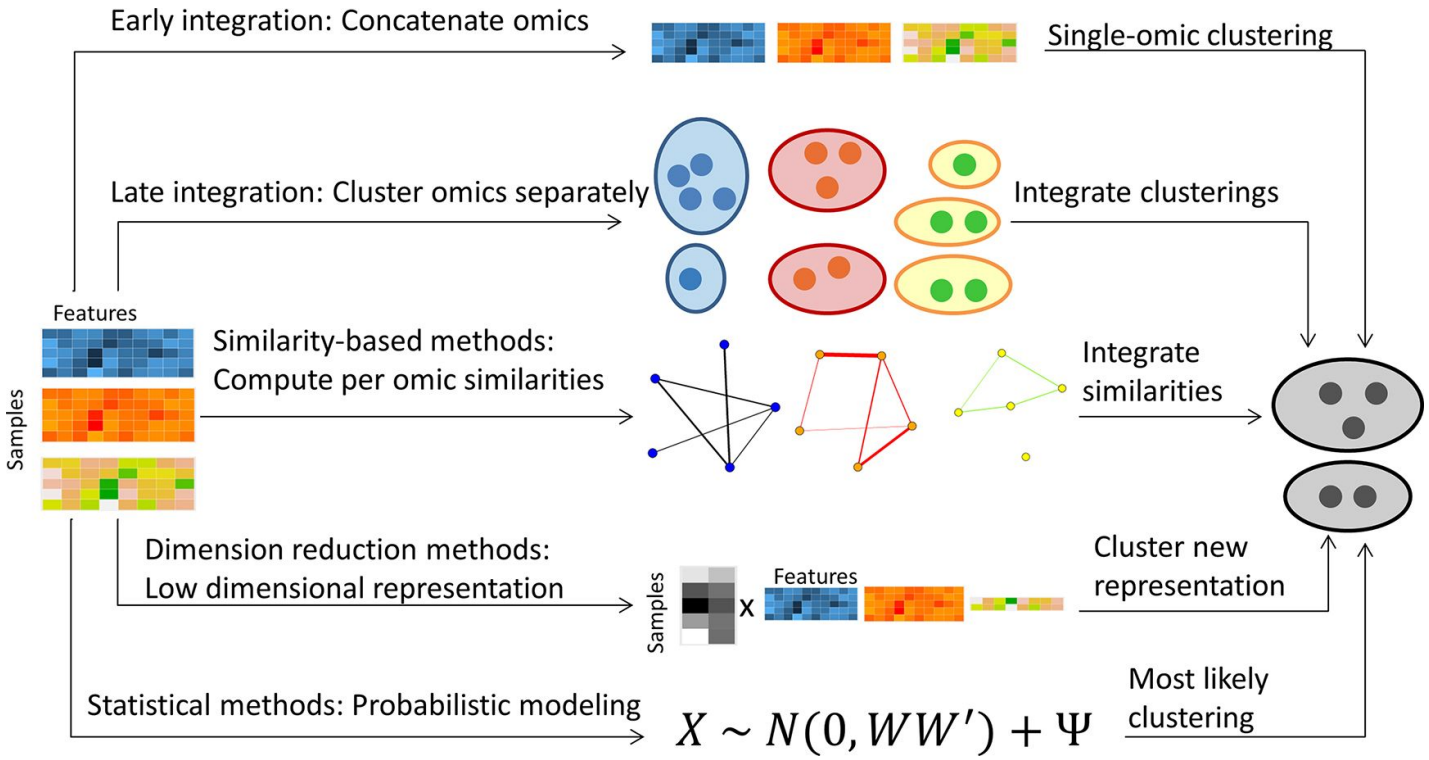
Normalization across
data types following
different distributions

Missing values
in feature matrices

Feature selection

Sparse datasets

Approaches to multi-omic data integration

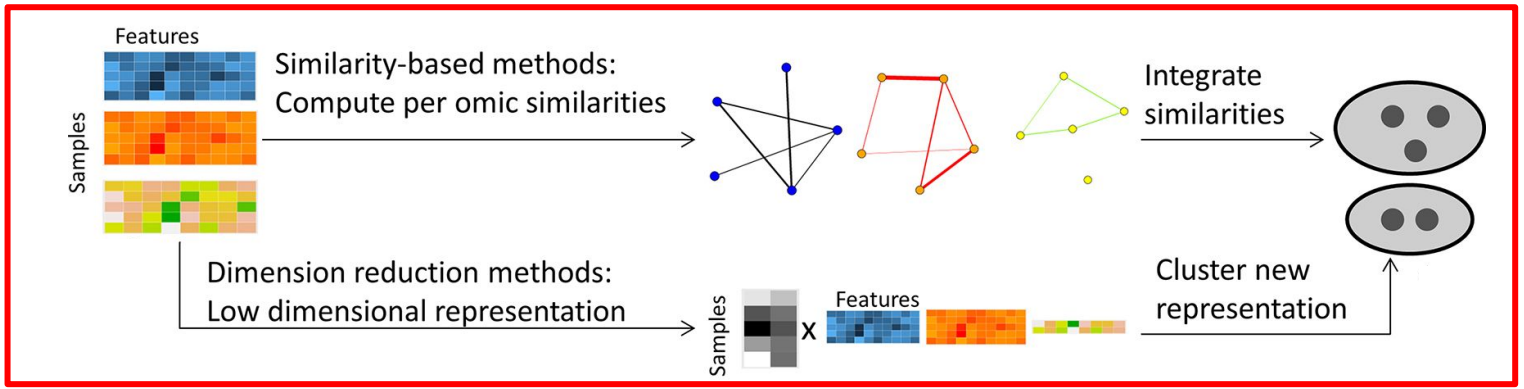


Rappoport & Shamir (2018) *Nucleic Acids Res.*

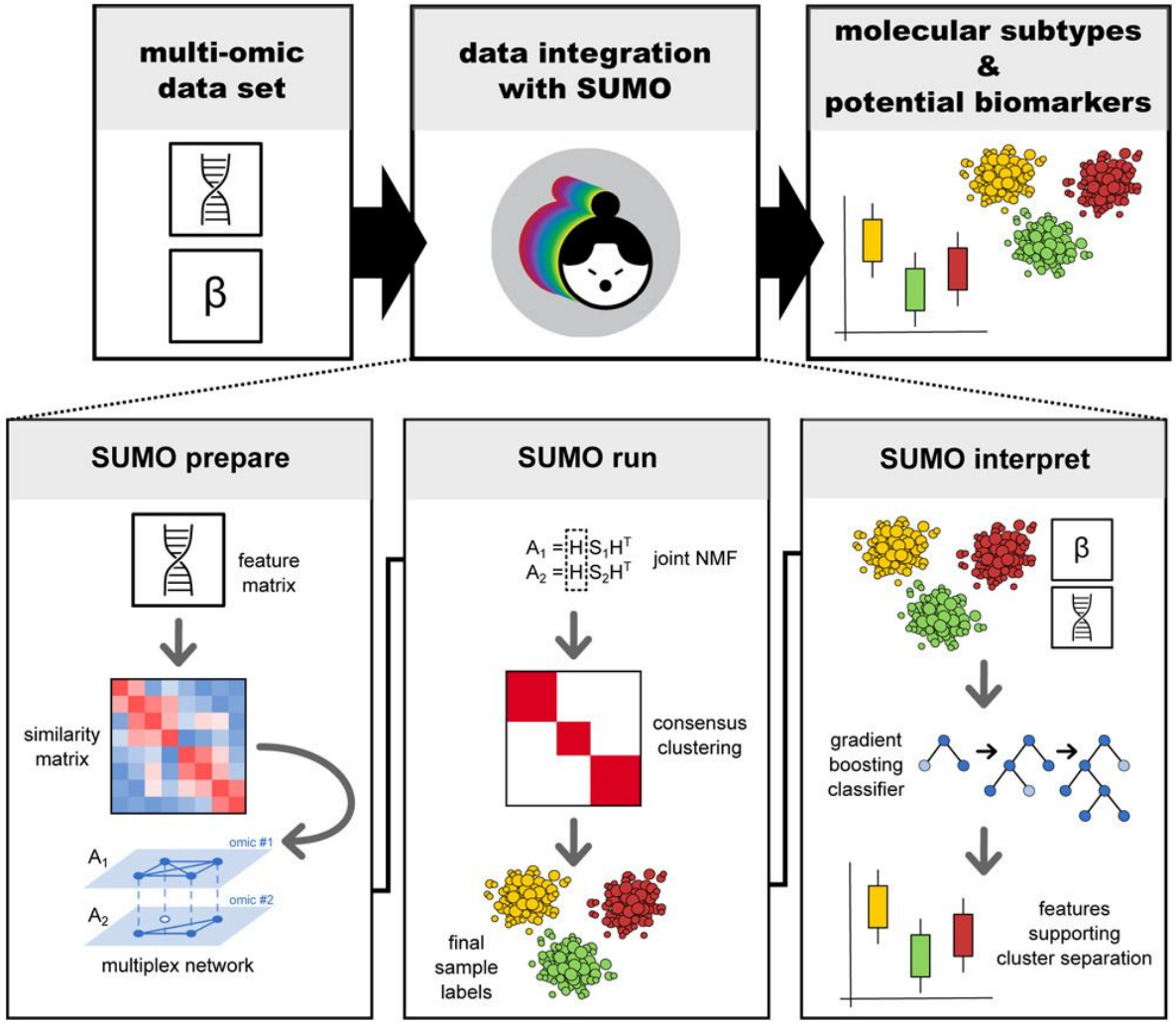
SUMO: **S**ubtyping **U**sing **M**ulti-**O**mic data

github.com/ratan-lab/sumo

pypi.org/project/python-sumo



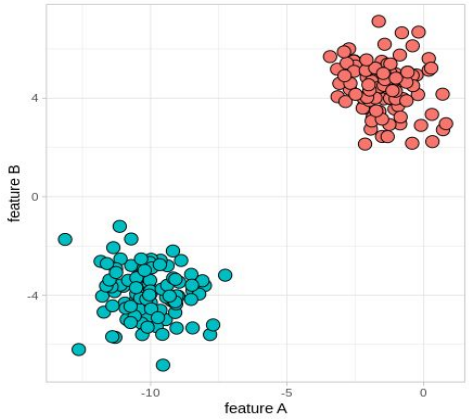
SUMO workflow

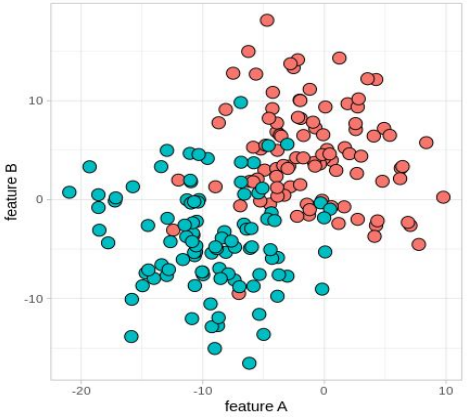


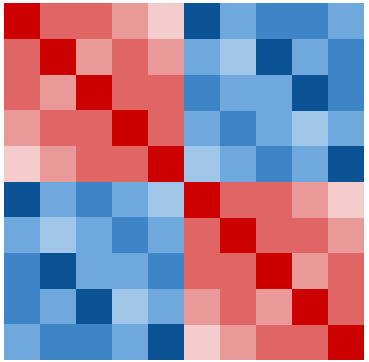
Sienkiewicz et al (2022)
Cell Reports Methods

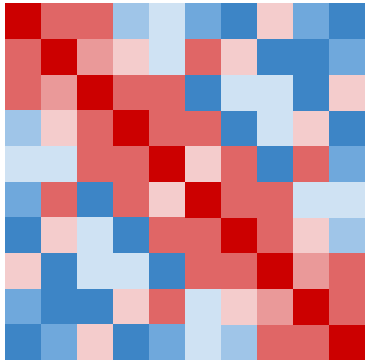
Similarity matrices - intuition

different data type requires different similarity metrics





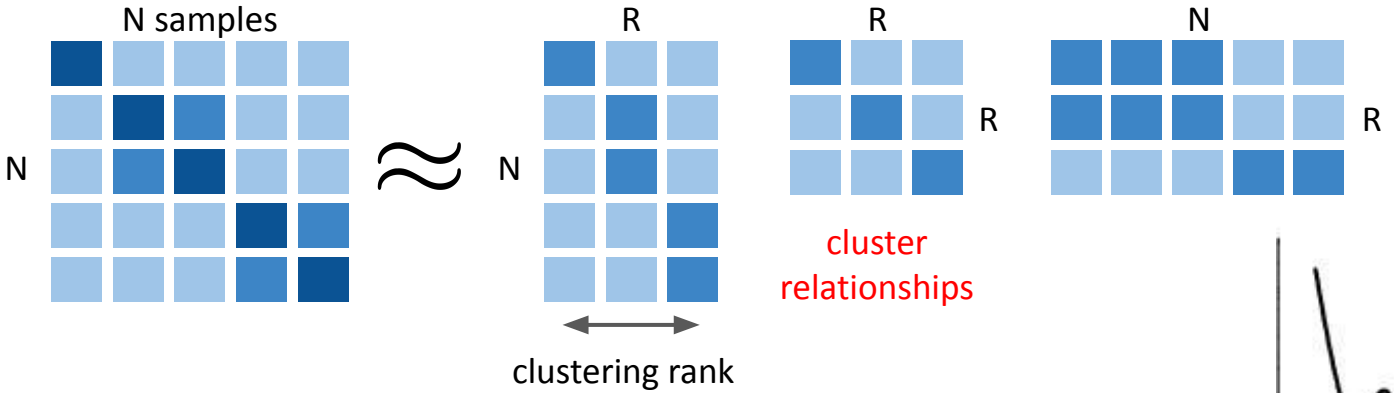


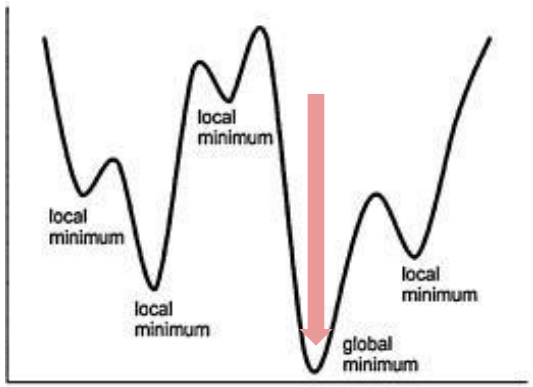


SUMO: Factorization

NMF	$X \approx WH$
Symmetric NMF	$A \approx HH^T$

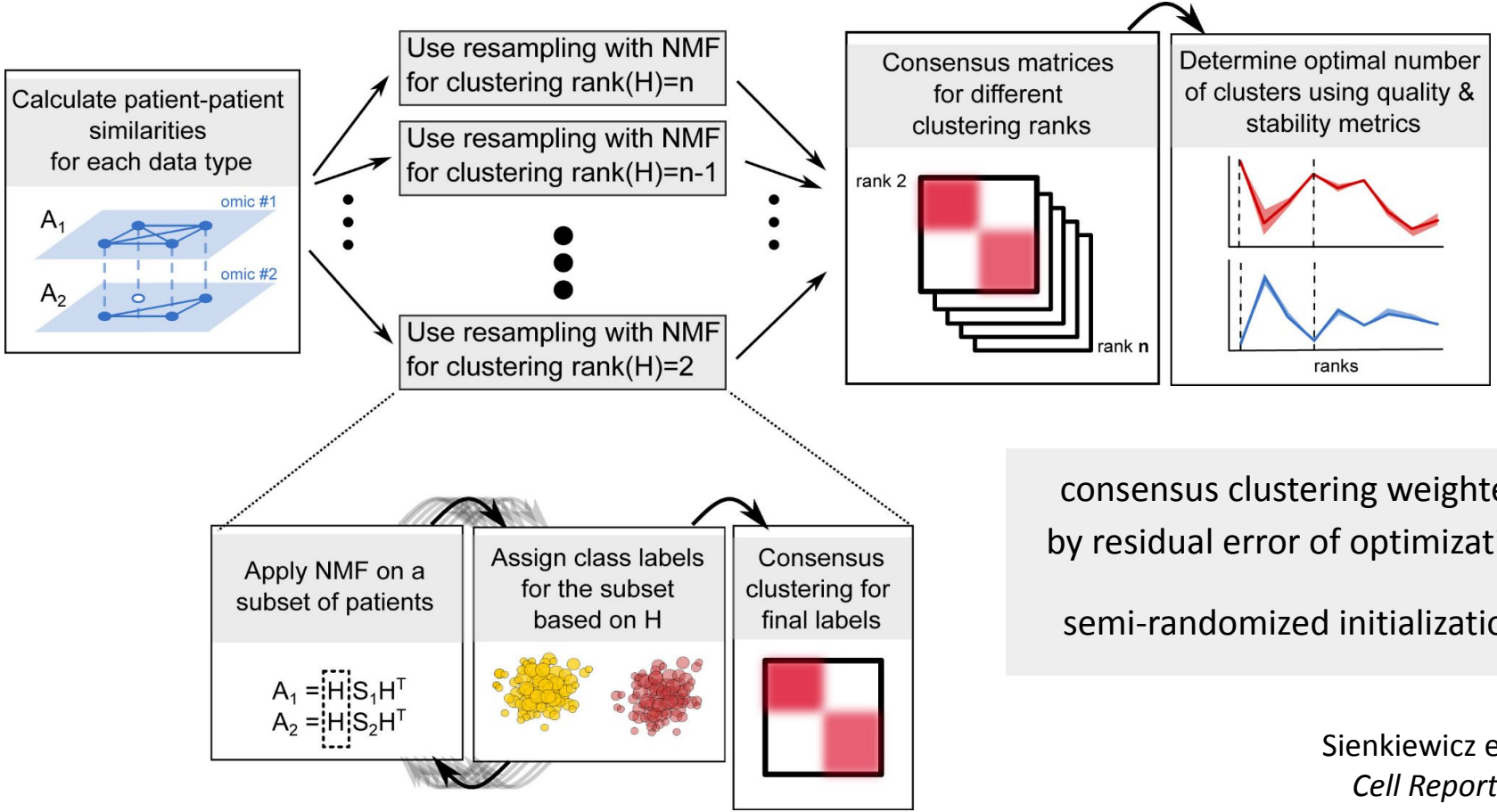
$$A^i \approx HS_iH^T$$





using gradient descent

SUMO: Re-sampling based approach

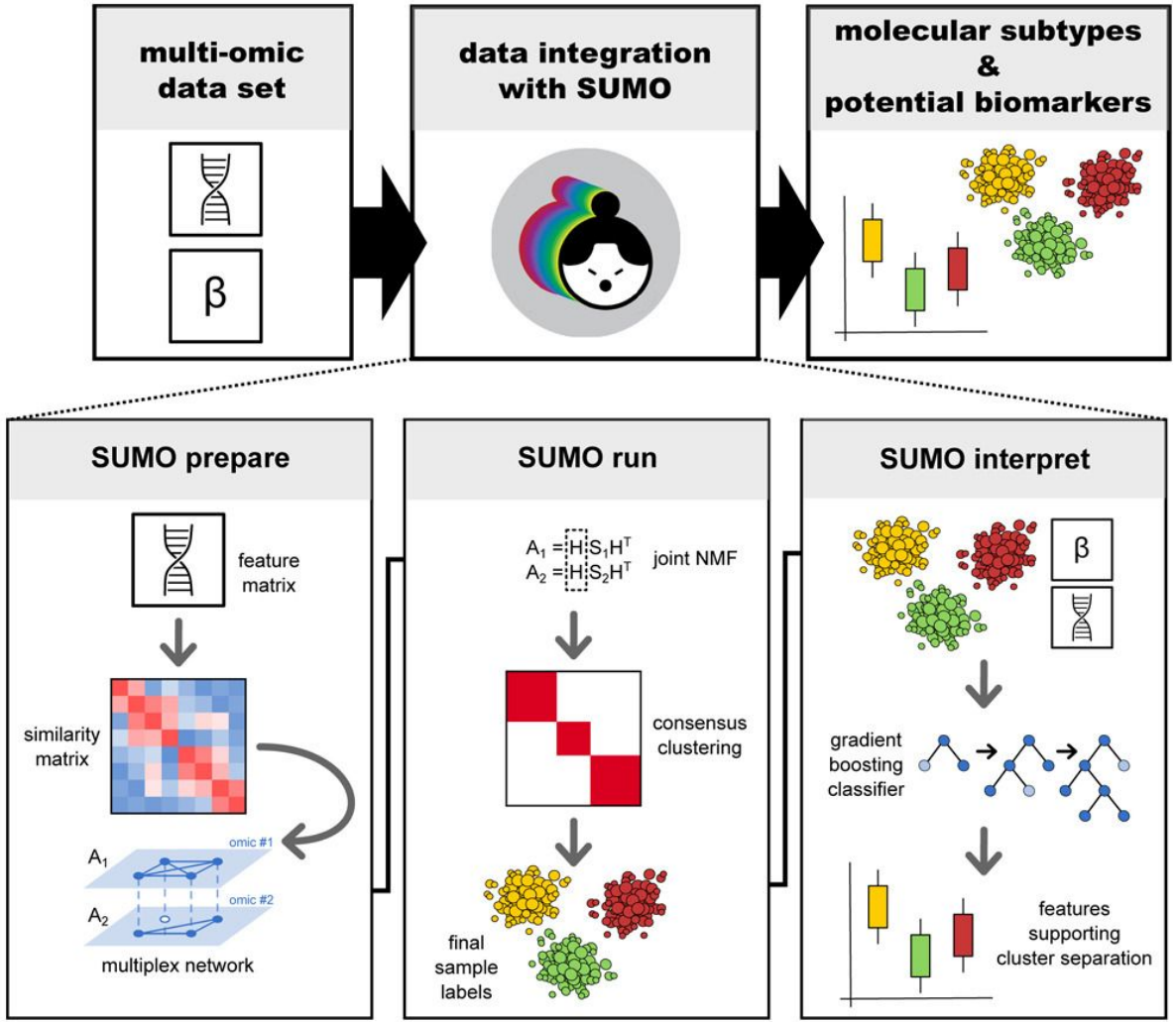


consensus clustering weighted by residual error of optimization

semi-randomized initialization

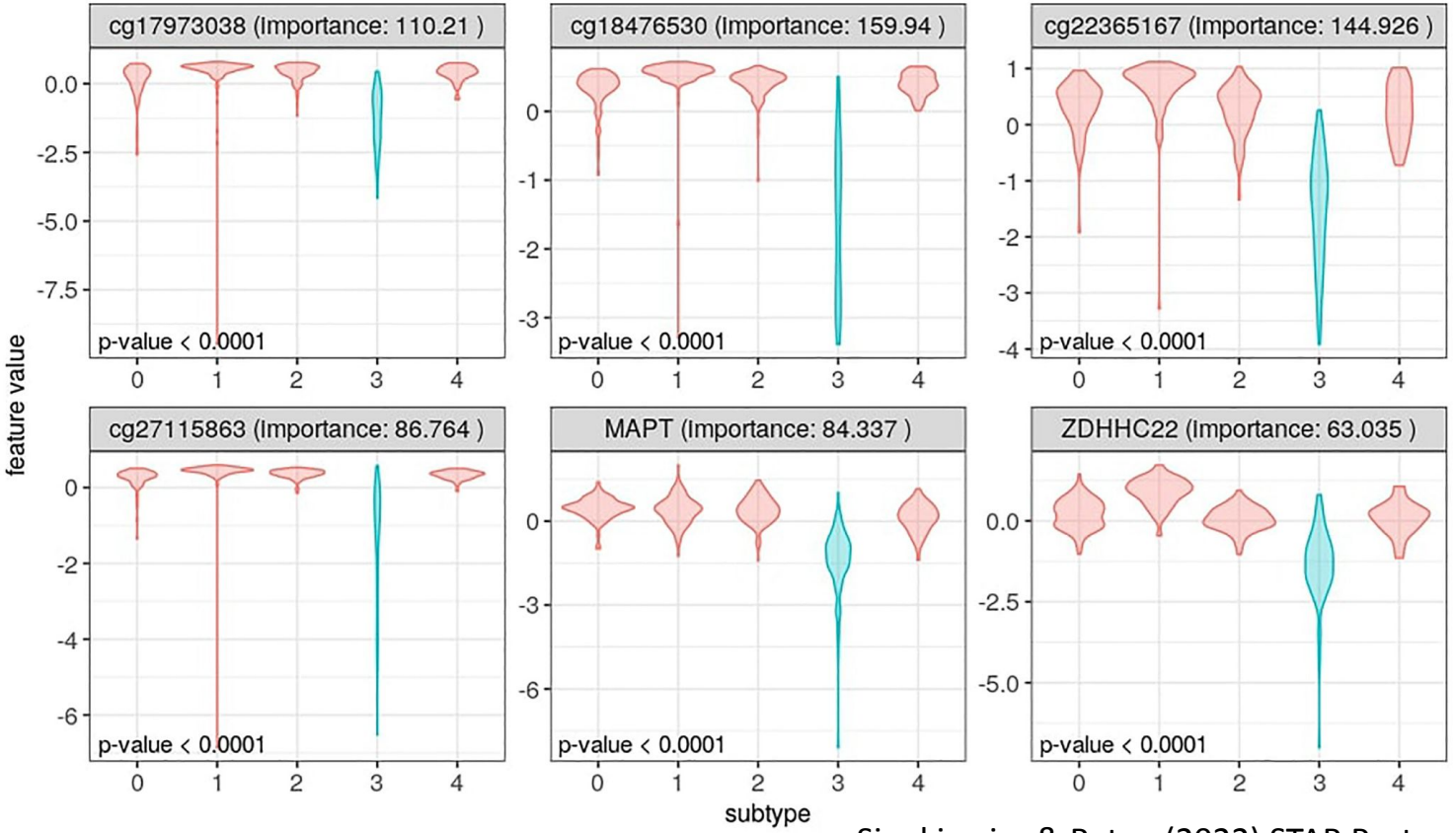
Sienkiewicz et al (2022)
Cell Reports Methods

SUMO workflow



Sienkiewicz et al (2022)
Cell Reports Methods

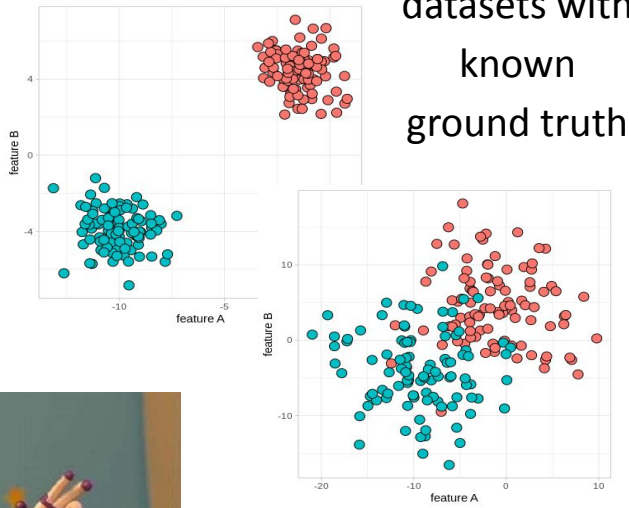
SUMO INTERPRET: downstream analysis



Sienkiewicz & Ratan (2022) STAR Protoc.

How do you test a new ML model? - a practical approach

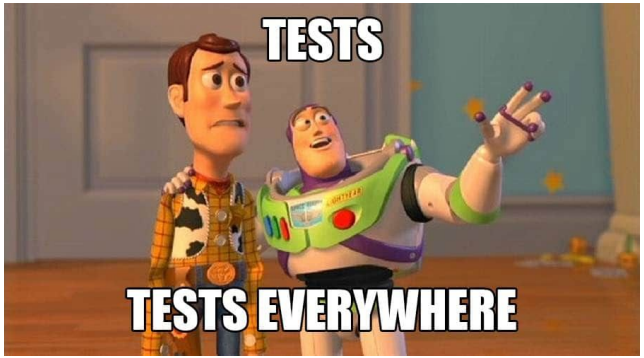
consult your goals



datasets with known ground truth

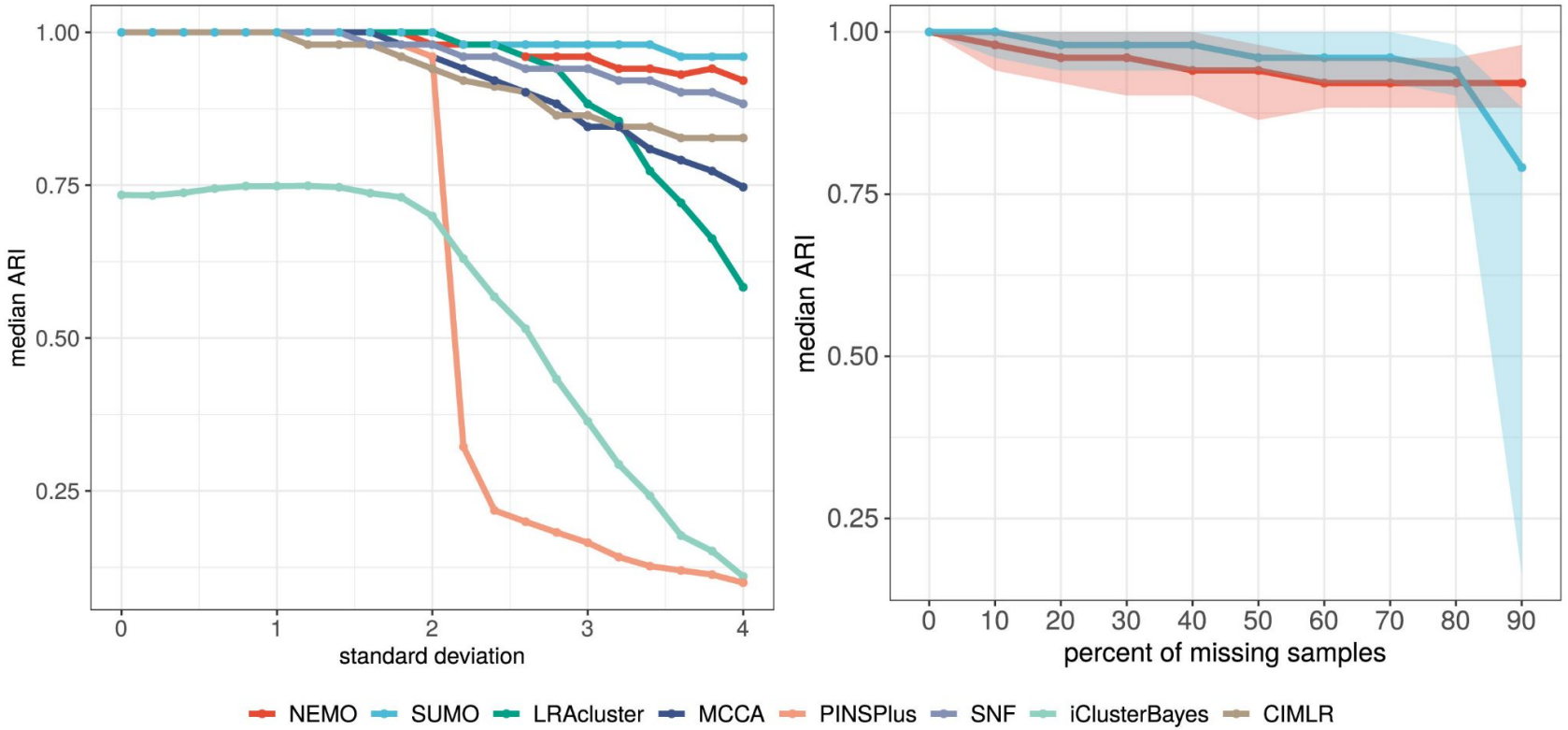
repeat each test

technical considerations (time & memory)



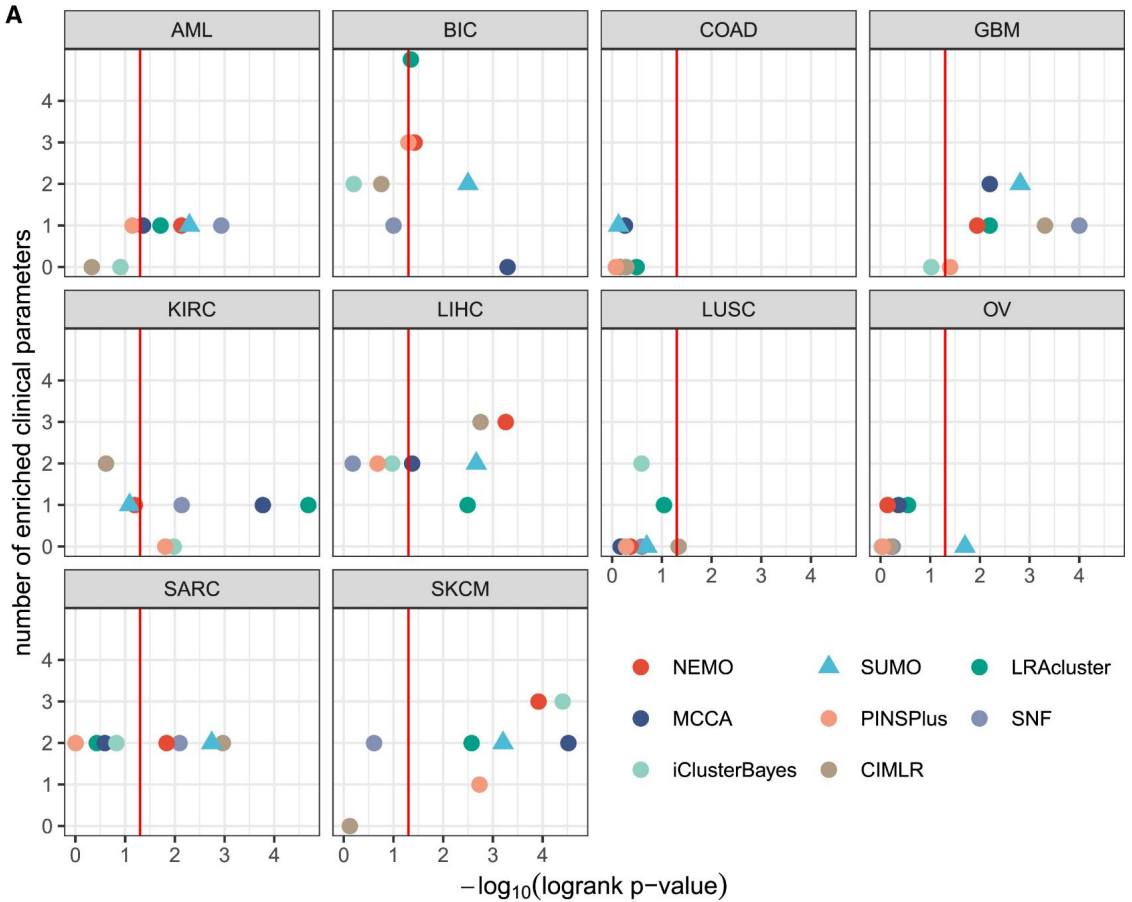
compare with competition

SUMO improves performance with noisy and incomplete data



Sienkiewicz et al (2022) *Cell Reports Methods*

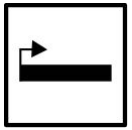
SUMO consistently identifies differential subtypes in TCGA data using an established benchmark



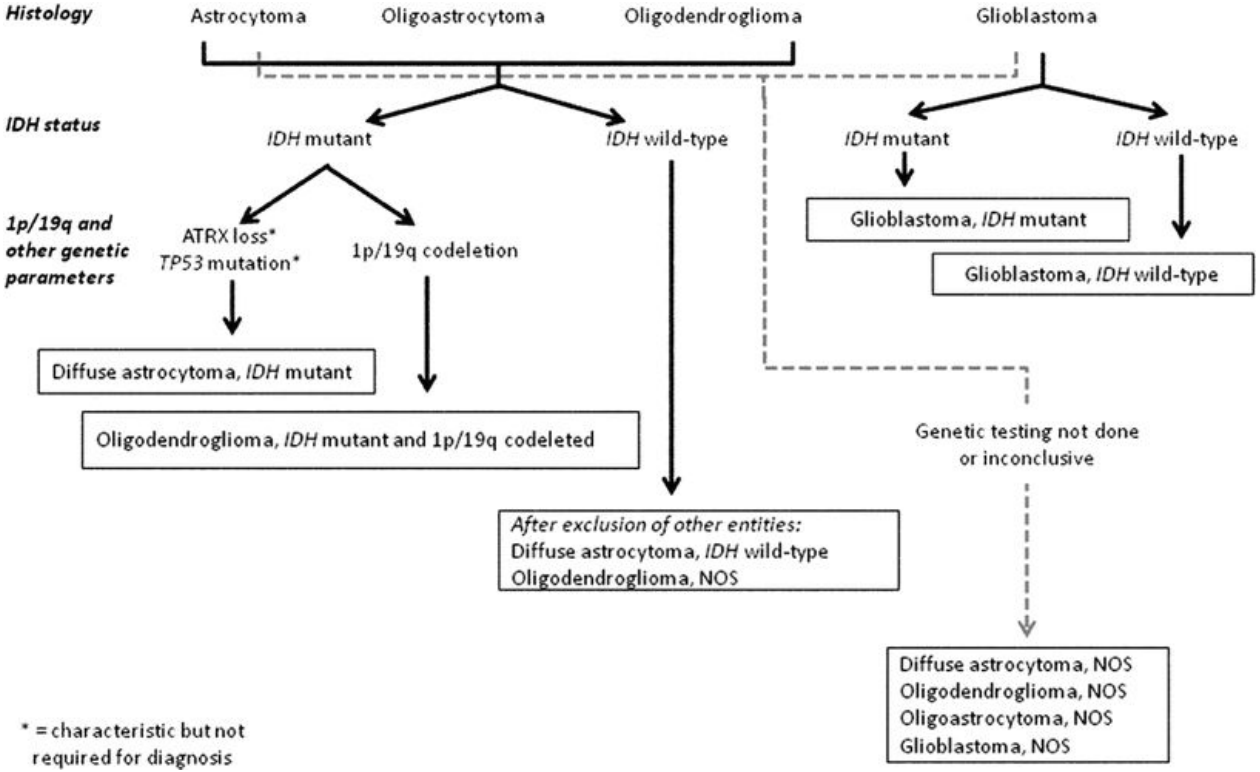
Sienkiewicz et al (2022) *Cell Reports Methods*

SUMO detects latent relationships between patients in Lower-Grade Gliomas

TCGA-LGG cohort
(a total of
556 primary tumors)

omic	# samples	tech
 RNA-seq	530	Illumina HiSeq
 DNA methylation	530	Illumina Infinium HumanMethylation450
 miRNA expression	524	Illumina HiSeq

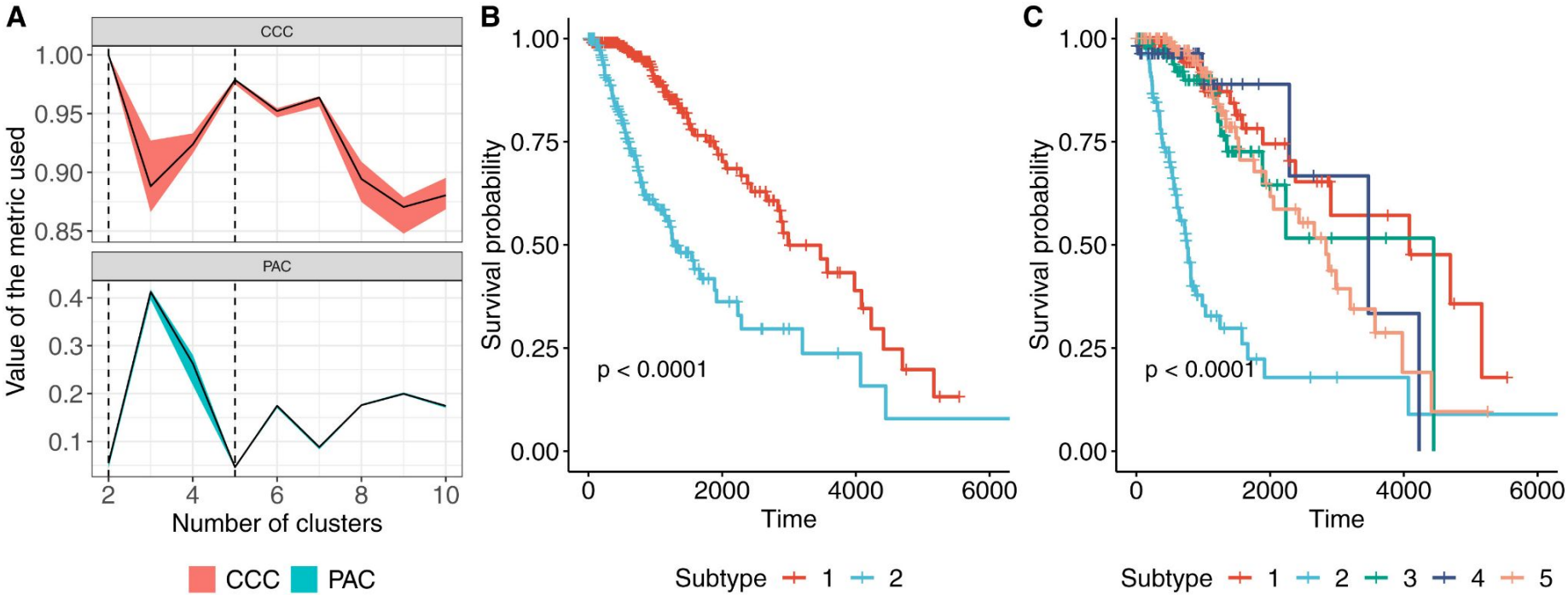
WHO guidelines for Low-Grade Gliomas



- Gliomas are neuroepithelial tumors originating from glial cells
- Postoperative RT remains the standard of care for adult grade 2/3
- ~80% have mutations in IDH1, 90% of those are R132H.
- ~50% have mutations in TP53, 40% have mutations in ATRX and 20% in CIC

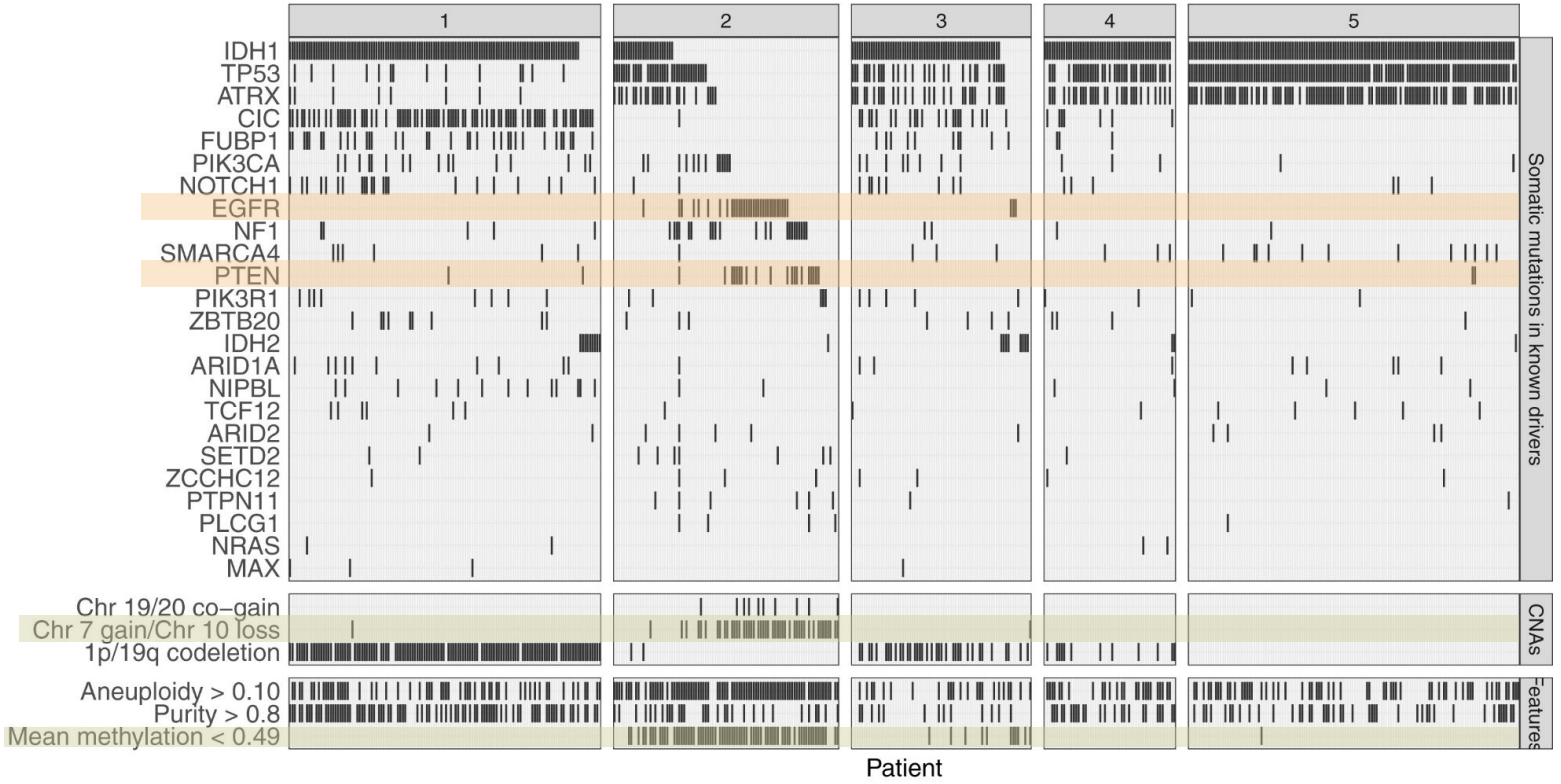
Source: The 2016 World Health Organization Classification of Tumors of the Central Nervous System (PMID: 27157931)

SUMO identifies a subtype of LGG with differential prognosis & GBM-like features



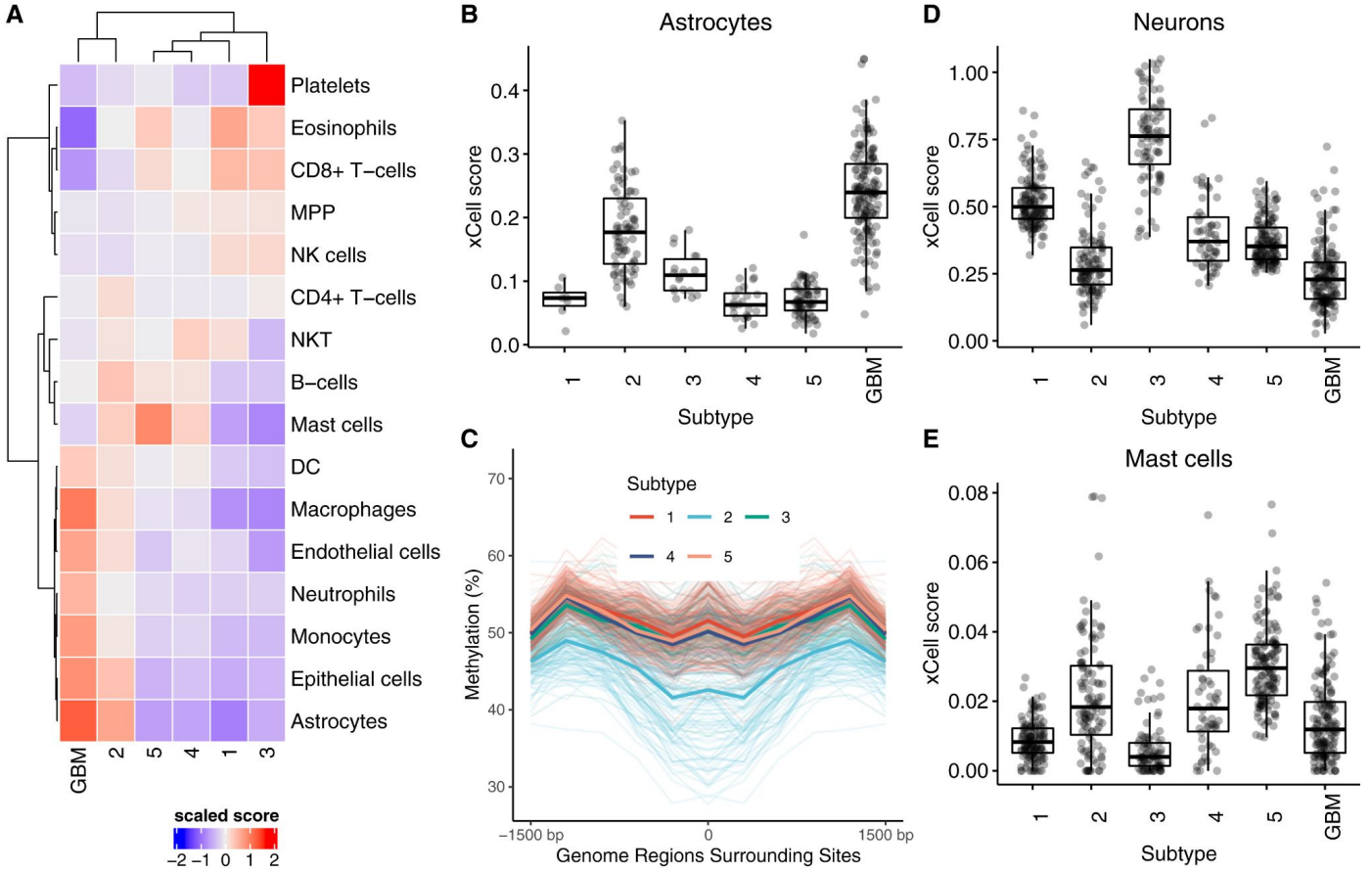
Sienkiewicz et al (2022) *Cell Reports Methods*

SUMO identifies a subtype of LGG with differential prognosis & GBM-like features



Sienkiewicz et al (2022) *Cell Reports Methods*

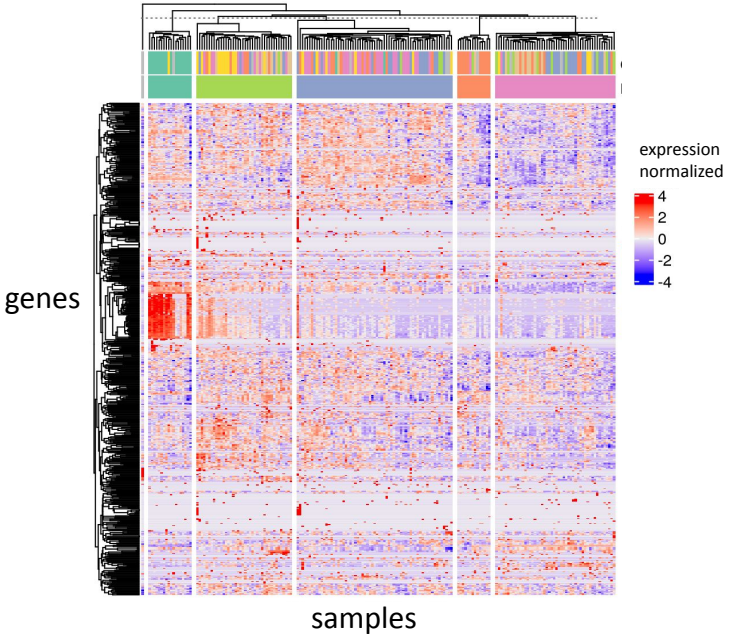
SUMO subtype 2 cellular profile is more similar to GBMs than other LGGs



Sienkiewicz et al (2022) *Cell Reports Methods*

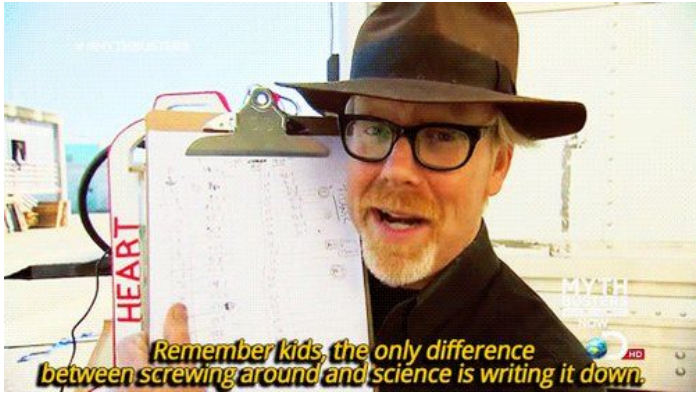
Conclusions

- Heatmaps are your friend



Conclusions

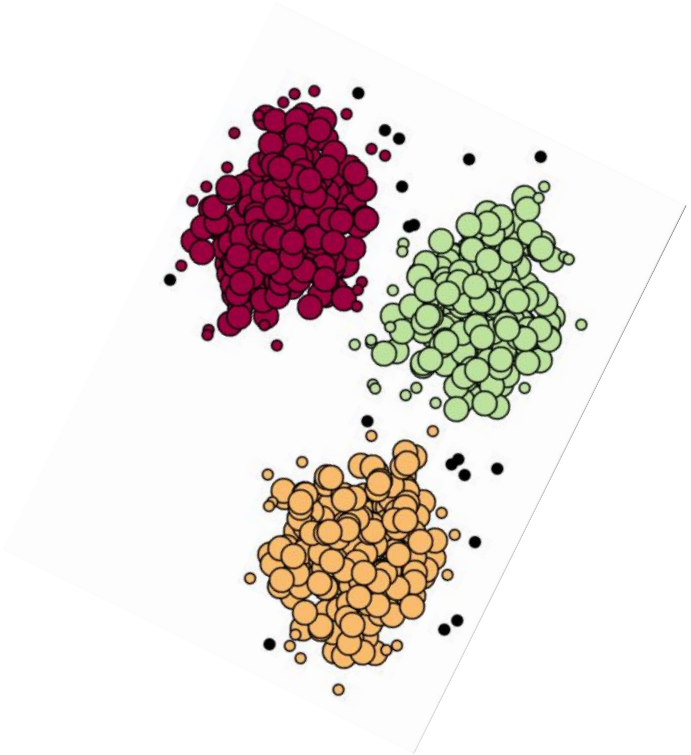
- Heatmaps are your friend
- Analysis-aware experimental design is essential
- Keep track of the metadata



Adam Savage

Conclusions

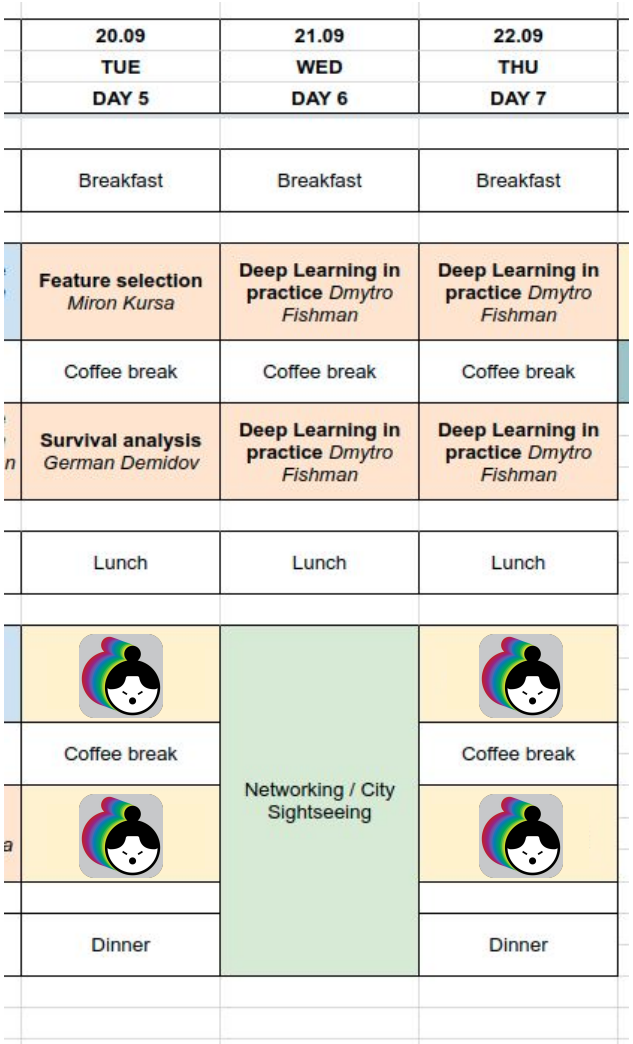
- Heatmaps are your friend
- Analysis-aware experimental design is essential
- Keep track of the metadata
- Each model has assumptions
- Don’t underestimate the value of simulated data



Conclusions

- Heatmaps are your friend
- Analysis-aware experimental design is essential
- Keep track of the metadata
- Each model has assumptions
- Don’t underestimate the value of simulated data

- If you are interested in multi-omic data integration, come to my hackathon!



Acknowledgements

SUMO TEAM

University of Virginia

Ajay Chatrath

John Lawson

Nathan Sheffield

Aakrosh Ratan

National University of Singapore

Jinyu Chen

Louxin Zhang

Project funding

University of Virginia

Center for Public Health Genomics

NIH National Institutes of Health

NIH NIGMS

National Research Foundation Singapore

Tri-Institutional PhD Program Computational Biology & Medicine

The Rockefeller University

Memorial Sloan Kettering Cancer Center

Weill Cornell Medicine

The Mason Lab