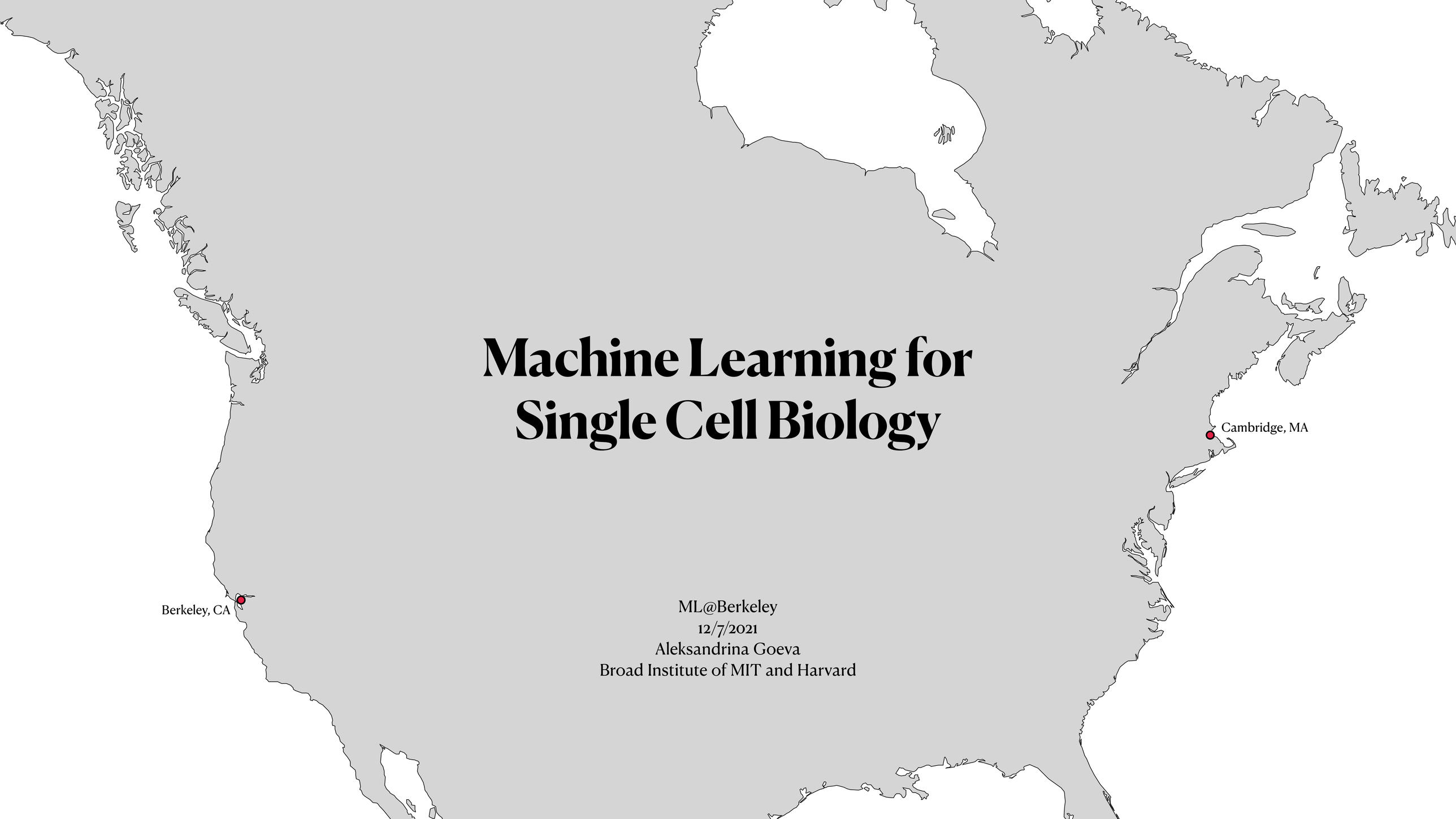
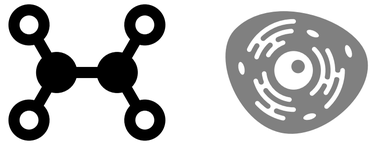


Machine Learning for Single Cell Biology

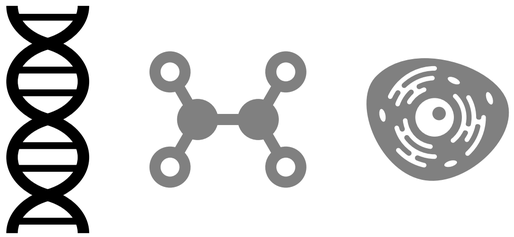
ML@Berkeley
12/7/2021
Aleksandrina Goeva
Broad Institute of MIT and Harvard

The **cell** is the fundamental unit of life.

Cells are made out of **molecules**.

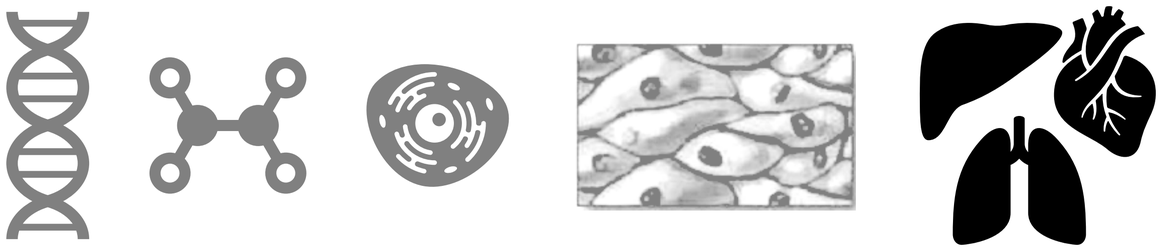


The molecules are encoded by **genes**
(or made out of gene products).

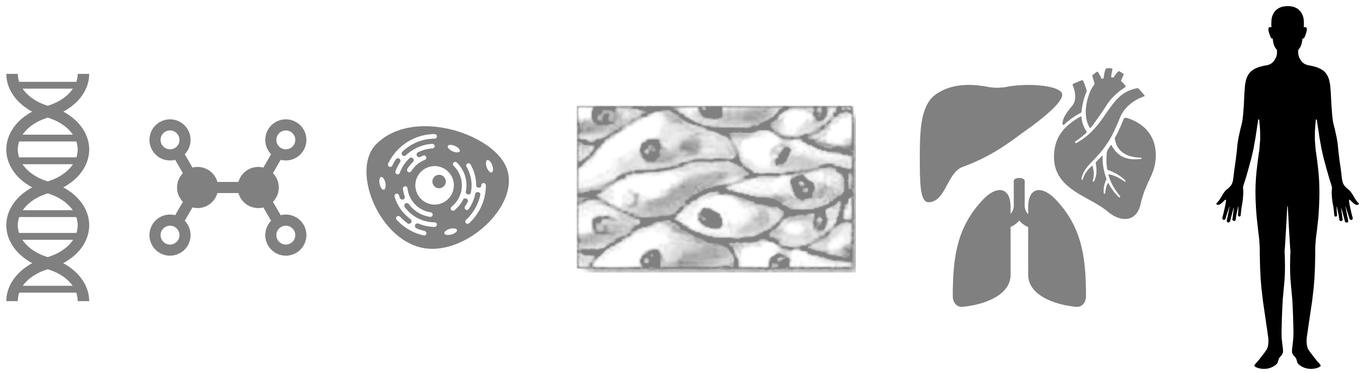


Cells interact with each other to make **tissues**.

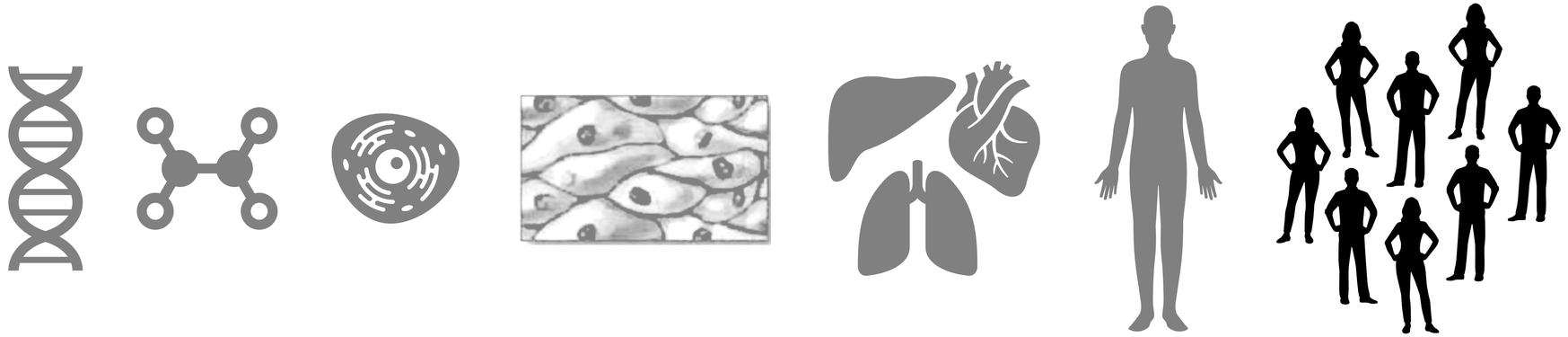
Tissues form **organs.**



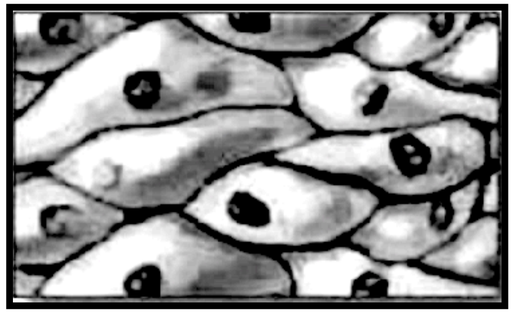
Organs together account for **organisms.**



Organisms together make **populations** and **ecosystems**.



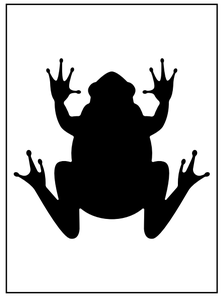
Today we will focus on **cells** and **tissues**.

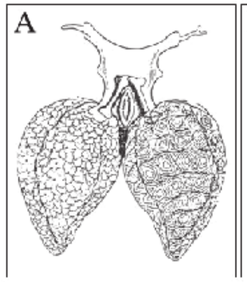


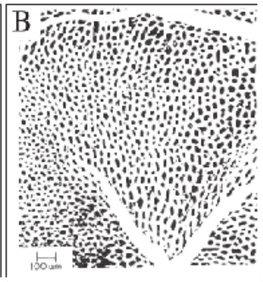
How do people collect **data** from tissues and cells?

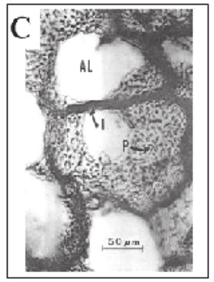
17th century

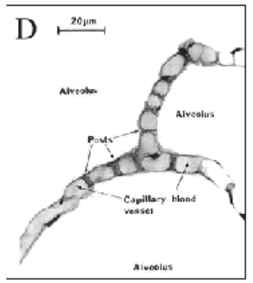
Using a **microscope** to look at parts of animals.











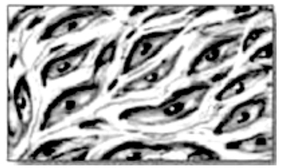
Timeline

17th century

19th century

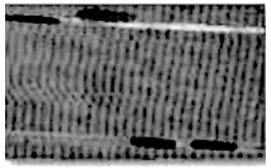
The birth of the terms **tissue** and **histology**.

1801
21 elementary tissues

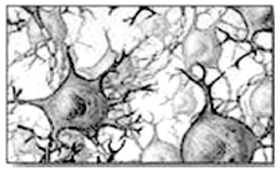


connective

epithelial



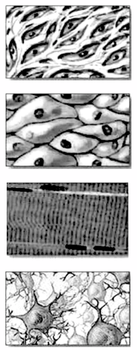
muscle



nervous

1857 - present
4 tissue types

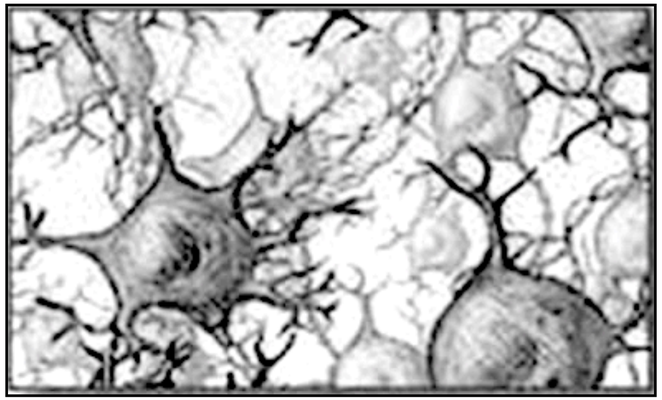
Timeline



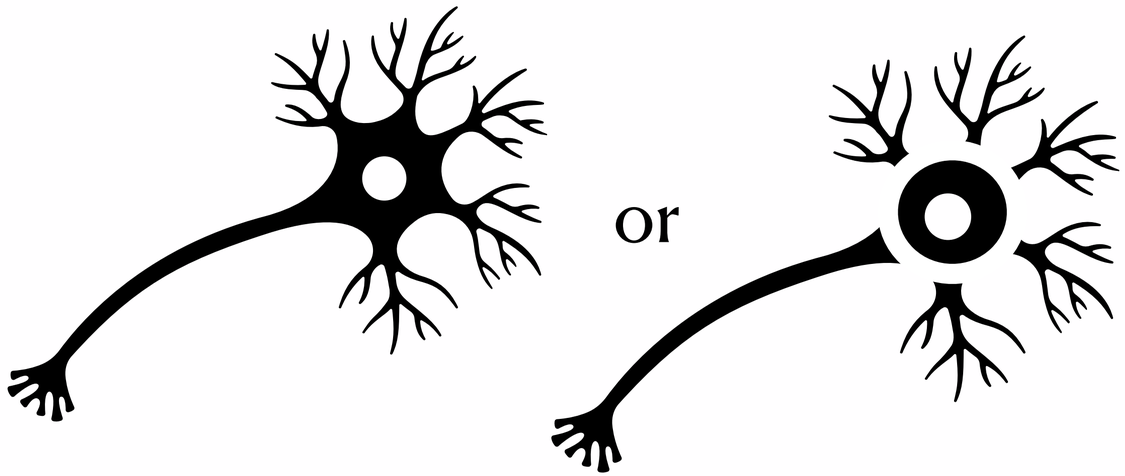
17th century

19th century

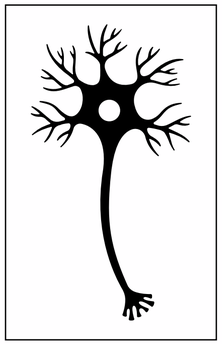
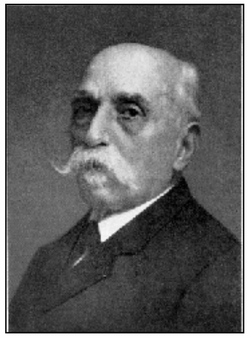
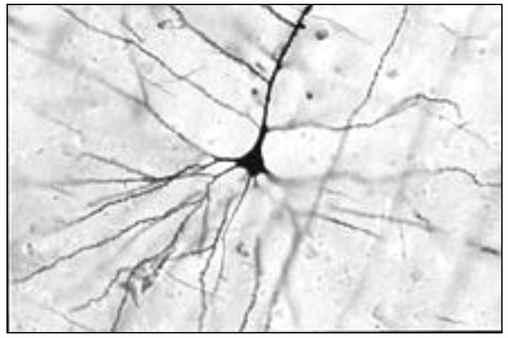
Neuron doctrine

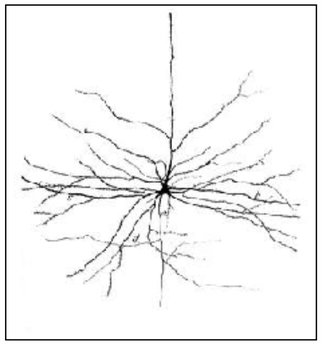


Neuron doctrine

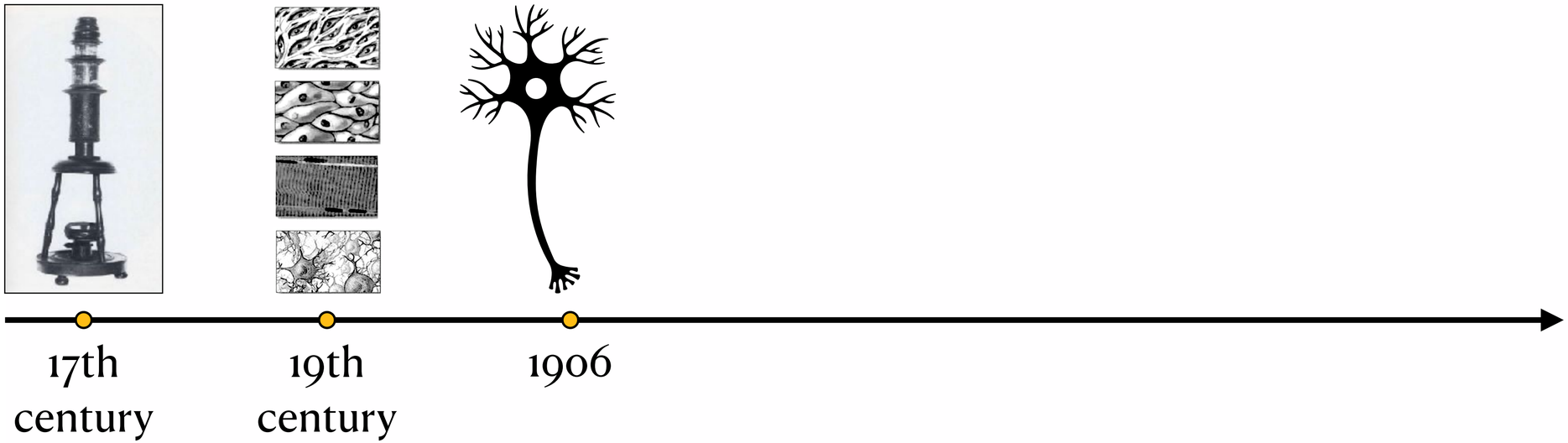


Neuron doctrine

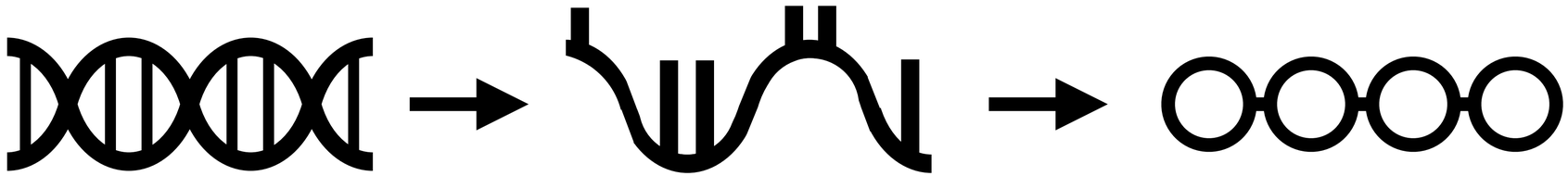






Timeline

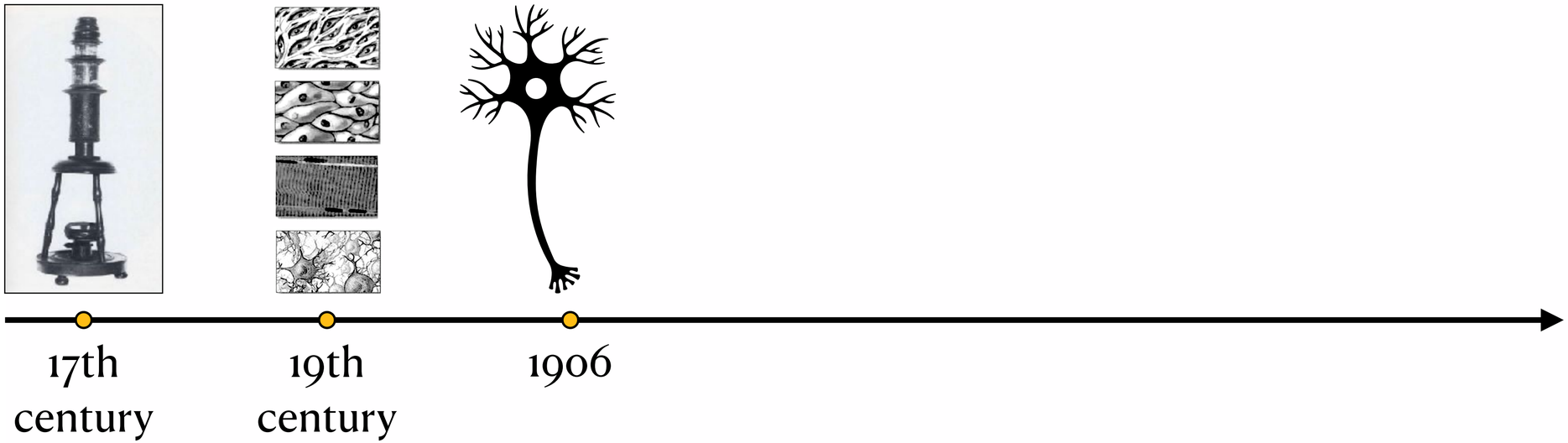


The central dogma.



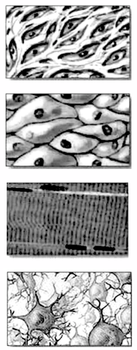
DNA **mRNA** protein

Timeline

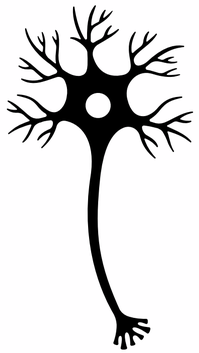


Microarrays measure the transcripts of **many genes** from a **bulk** sample.

17th century



19th century



1906

1995

Next Generation Sequencing allows **transcriptome-wide** measurements from a **bulk** sample.

Single-cell RNAseq

Spatial transcriptomics

17th century

19th century

1906

1995

2006

2015

2019

nature methods

Technology Feature | Published: 06 January 2021

Method of the Year: spatially resolved transcriptomics

Open Problems in Single-Cell Analysis

Benchmarking formalized challenges in single-cell analysis

Computational biology is undergoing a revolution. Recent advances in microfluidic technology enable high-throughput and high-dimensional of individual cells at unprecedented scale. But there's a catch.

Single-cell analysis is *hard*.

Timeline

17th century

19th century

1906

1995

2006

2015

2019

Timeline

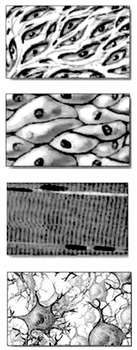

GeneChip
Affymetrix
17th century
19th century
1906
1995
2006
2015
2019

Timeline

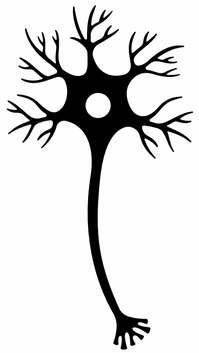
Sofia, Bulgaria

Timeline

17th century



19th century

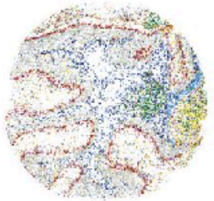


1906

1995

2006

2015

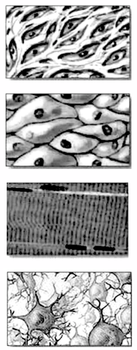


2019

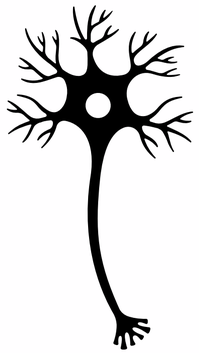
Sofia, Bulgaria

Timeline

17th century



19th century

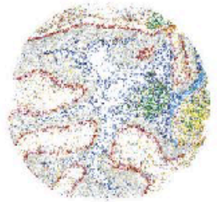


1906

1995

2006

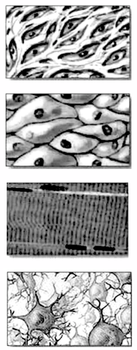
2015

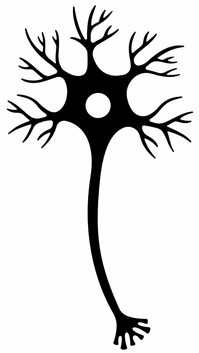


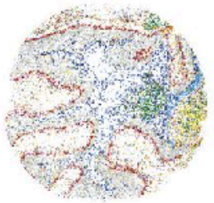
2019

Sofia, Bulgaria

Timeline







17th century

19th century

1906

1995

2006

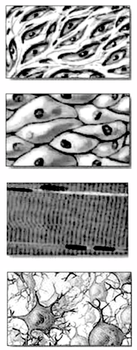
2015

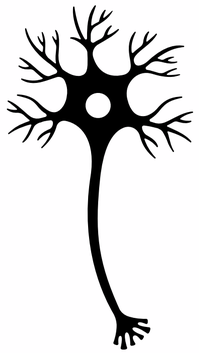
2019

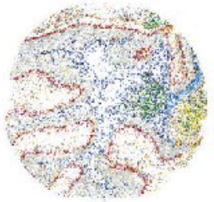
Sofia, Bulgaria

Boston, MA

Timeline







17th century

19th century

1906

1995

2006

2015

2019

Sofia, Bulgaria

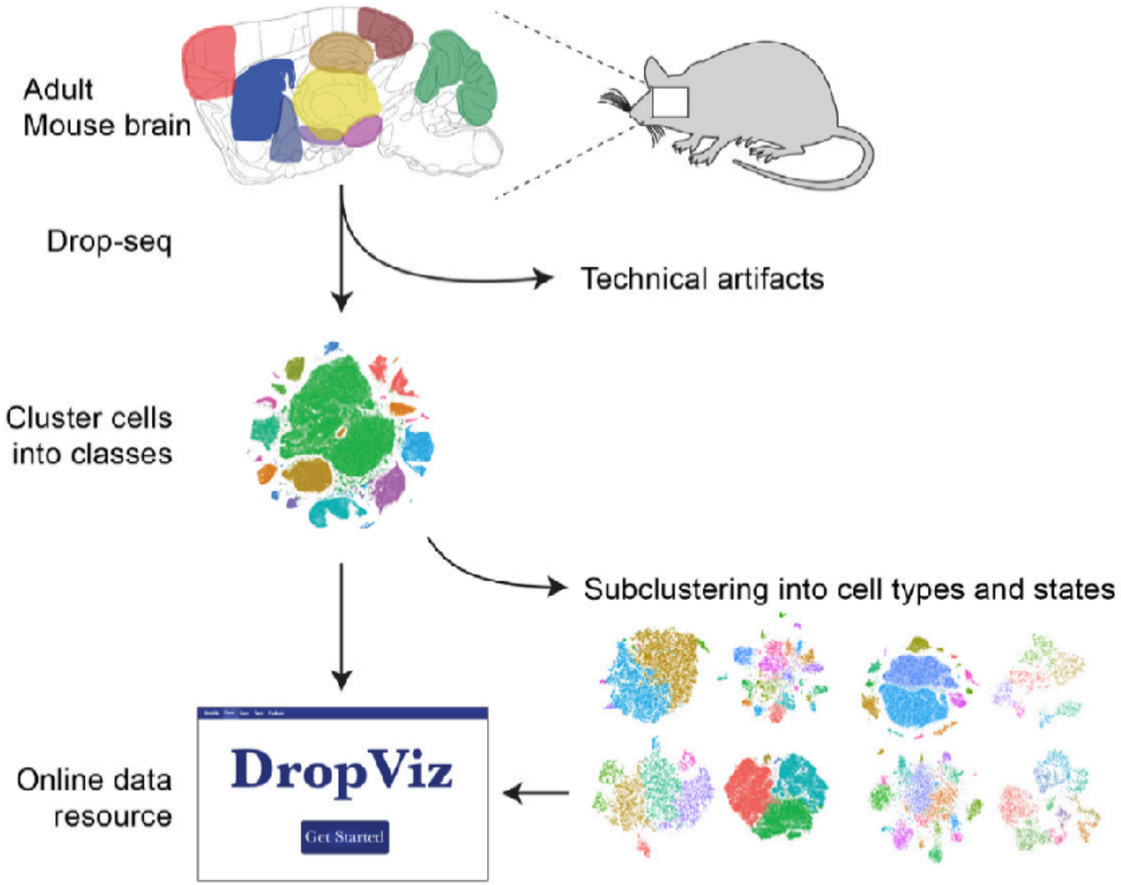
Boston, MA

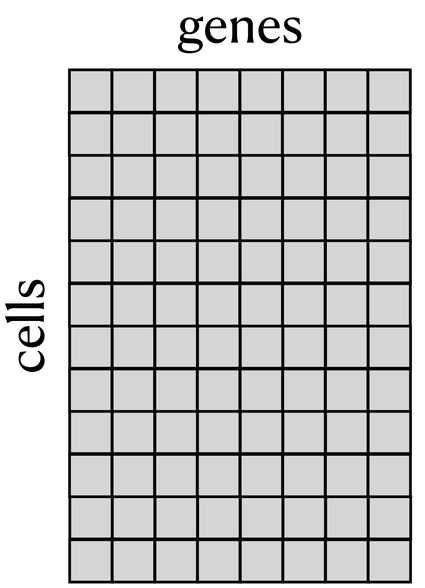
We would like to use this data to answer the following questions:

1. What are the different types of cells?
2. Where are they spatially in the tissue?
3. How do interactions with their neighbors shape who they are?
4. What are the cellular differences between health and disease?

1. What are the different types of cells in a tissue?

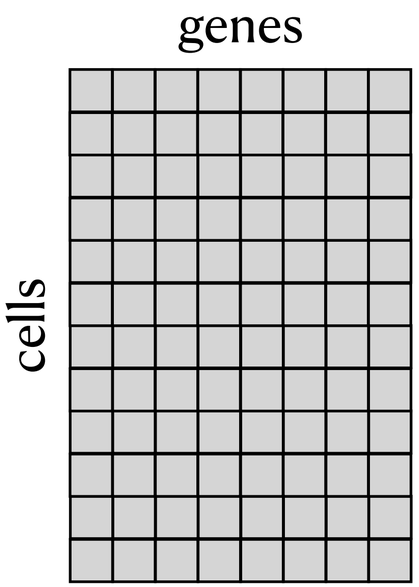
Single-cell RNA-seq Data



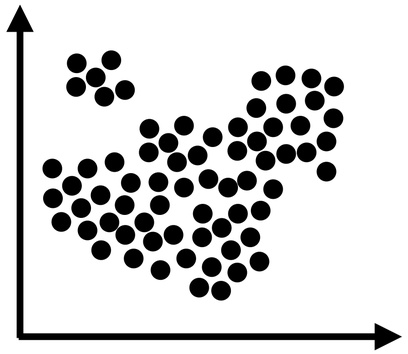


matrix of counts
of mRNA abundance
in each cell

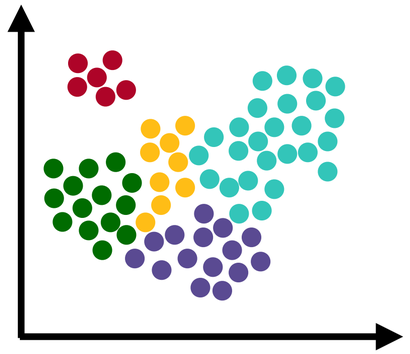
Analysis workflow



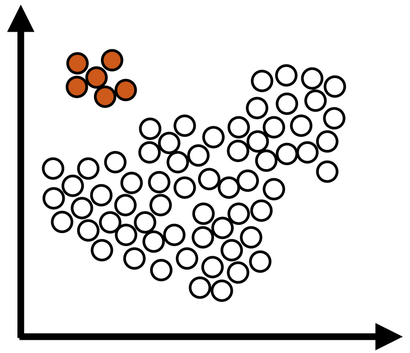
matrix of counts
of mRNA abundance
in each cell



feature selection
dimensionality reduction



clustering



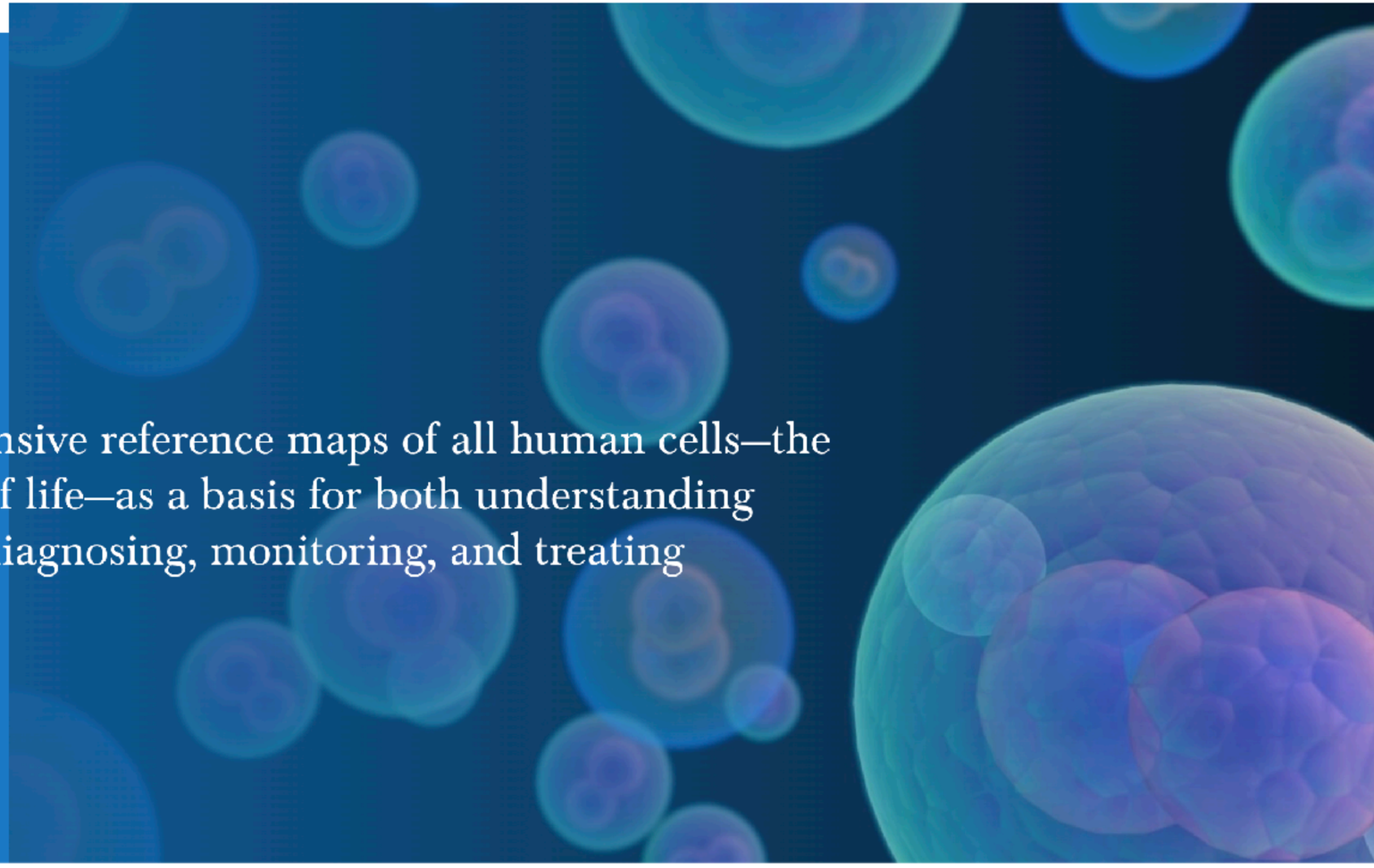
plotting known
marker genes
to annotate clusters

HUMAN CELL ATLAS

Home | About HCA | COVID-19 | Research | News | Publications | Data/DCP | Resources | Join/Contact

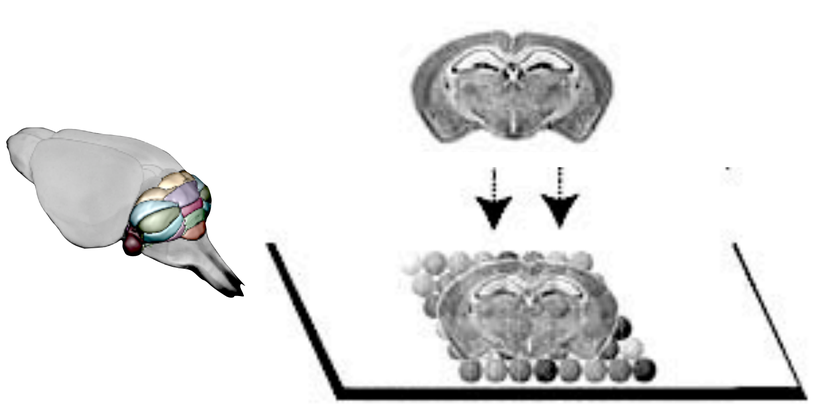
MISSION

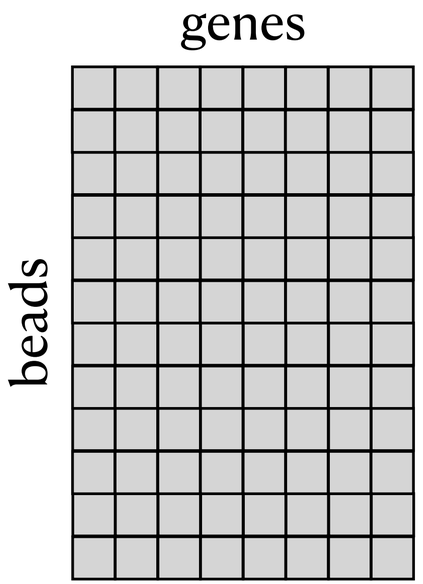
To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.



2. Where are the different types of cells in space?

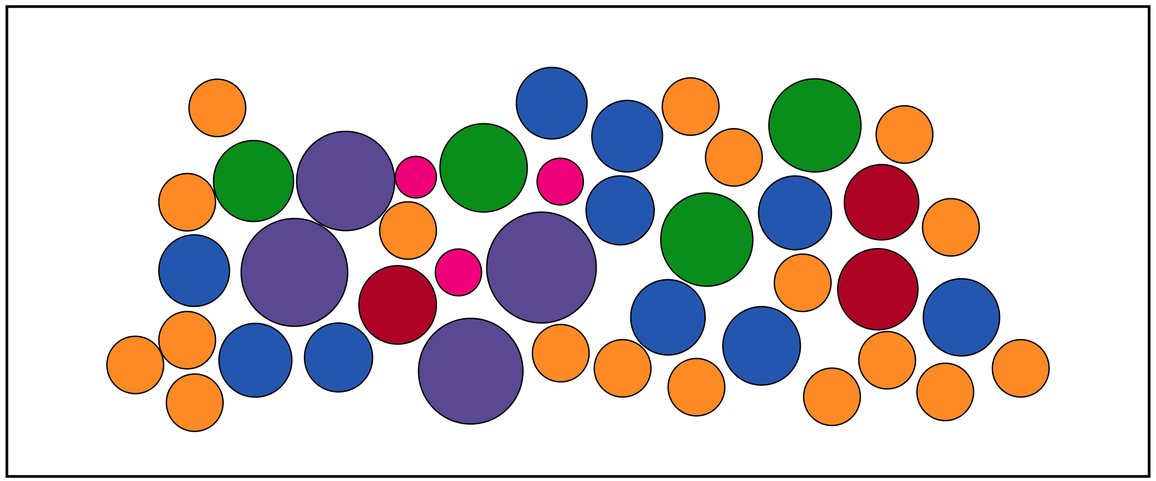
Spatial Transcriptomics Data





matrix of counts
of mRNA abundance
at each spatial probe

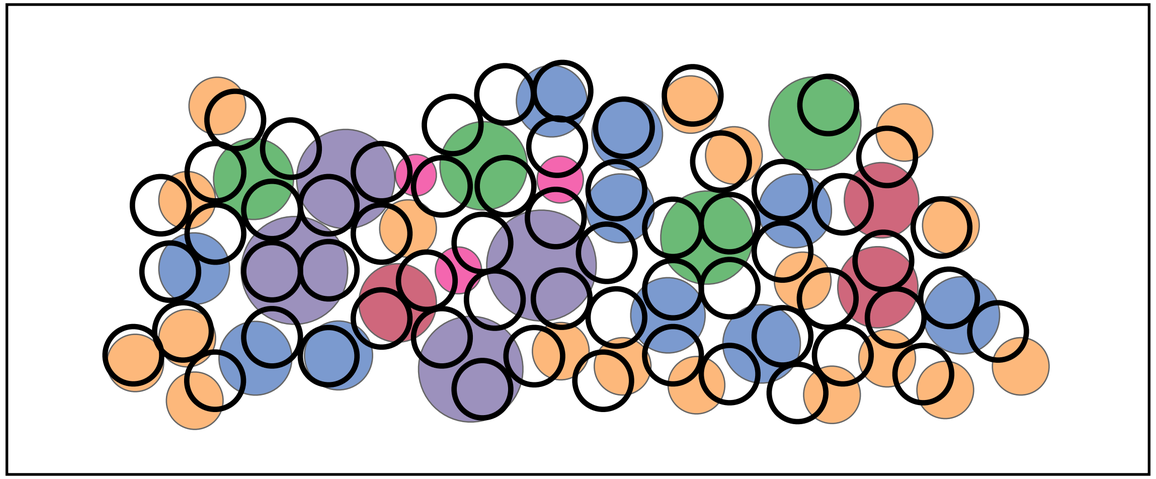
The tissue contains **multiple cell types.**



Cells in the tissue

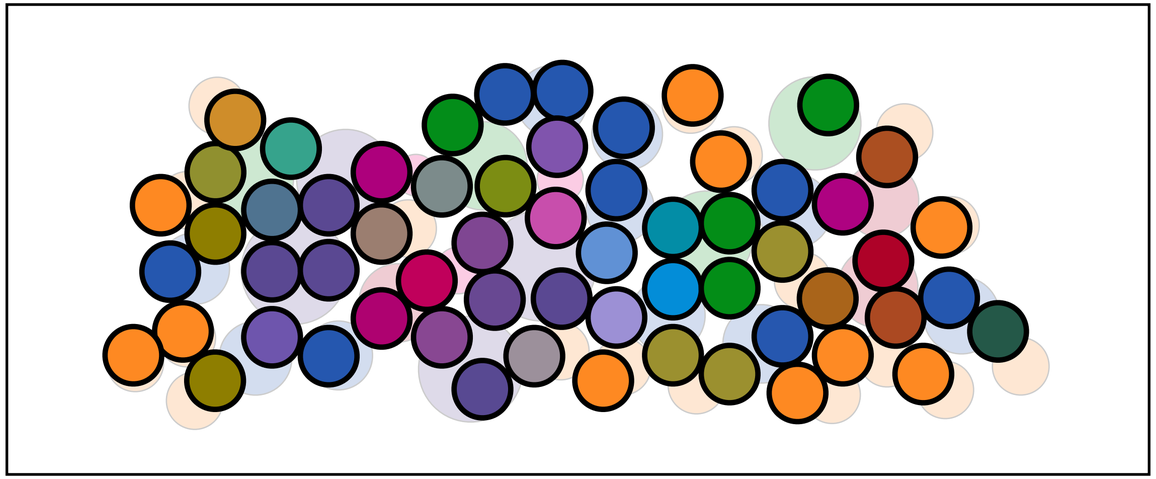
Beads densely cover the plane.

Although the beads are approximately as small as the cells in the tissue, they are not necessarily centered on top of individual cells.



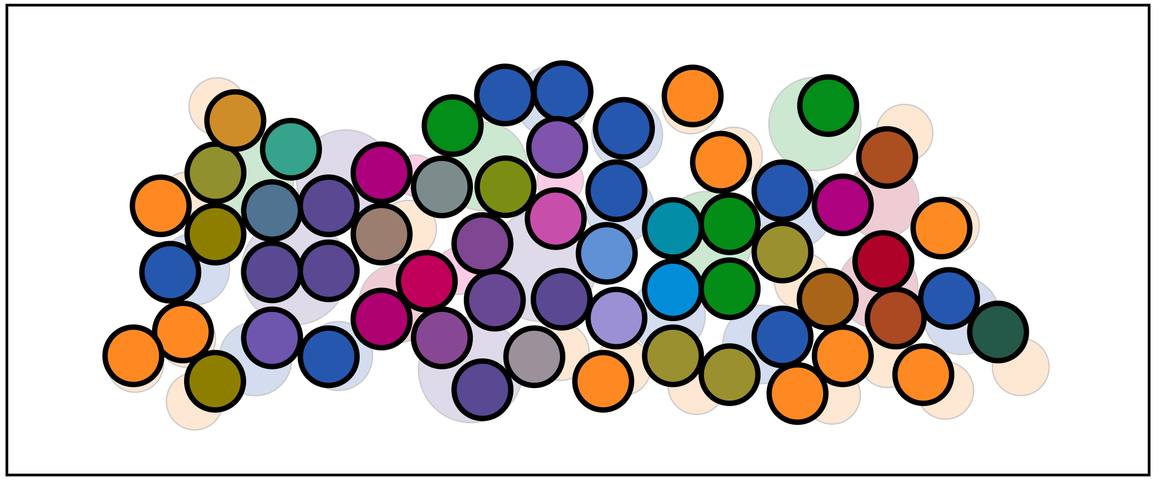
Locations of Slide-seq beads

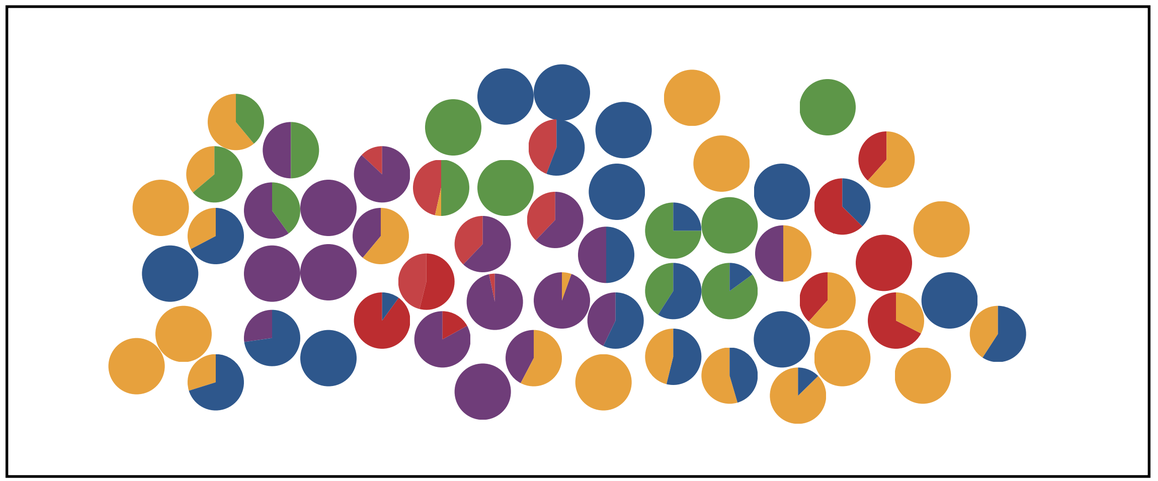
Each bead is a **mixture** of multiple cell types.



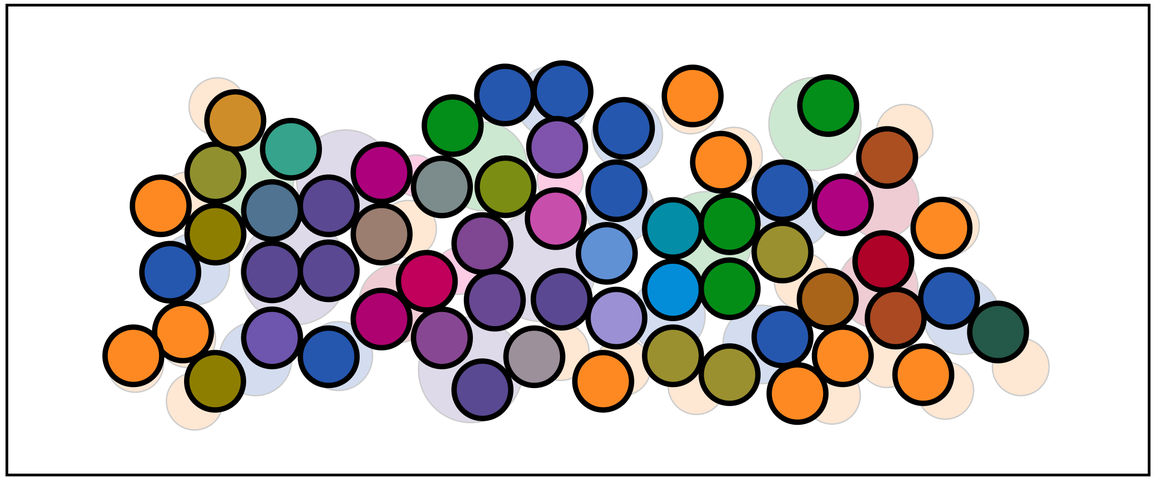
Slide-seq measurements

Given the Slide-seq observations,
what are the cell types and the mixtures made out of?

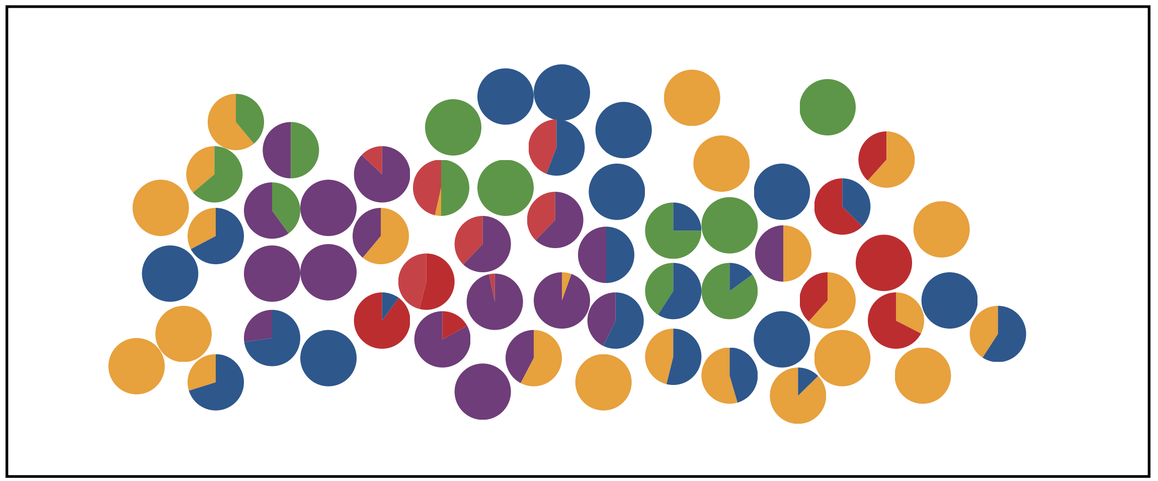




Inverse problem

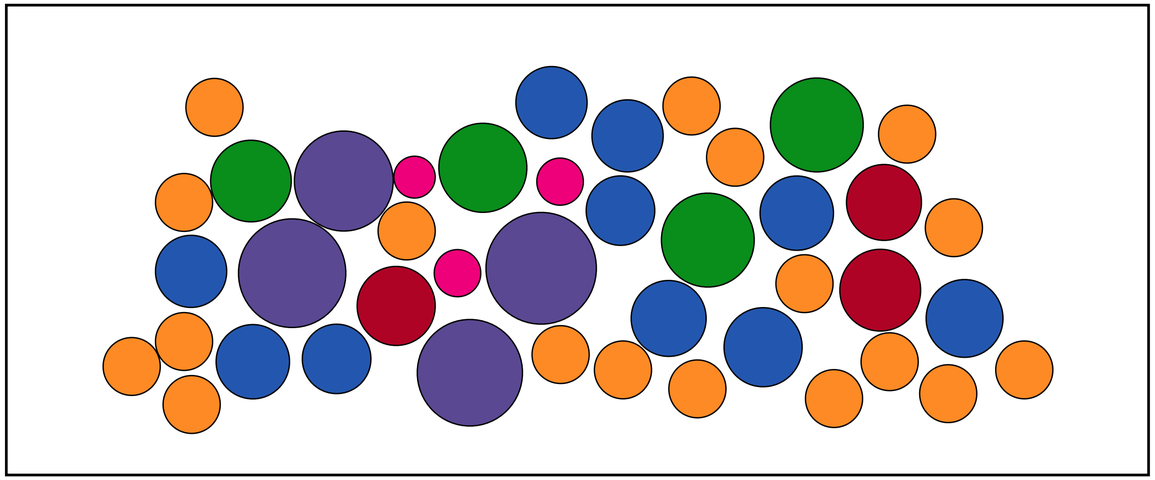


observations

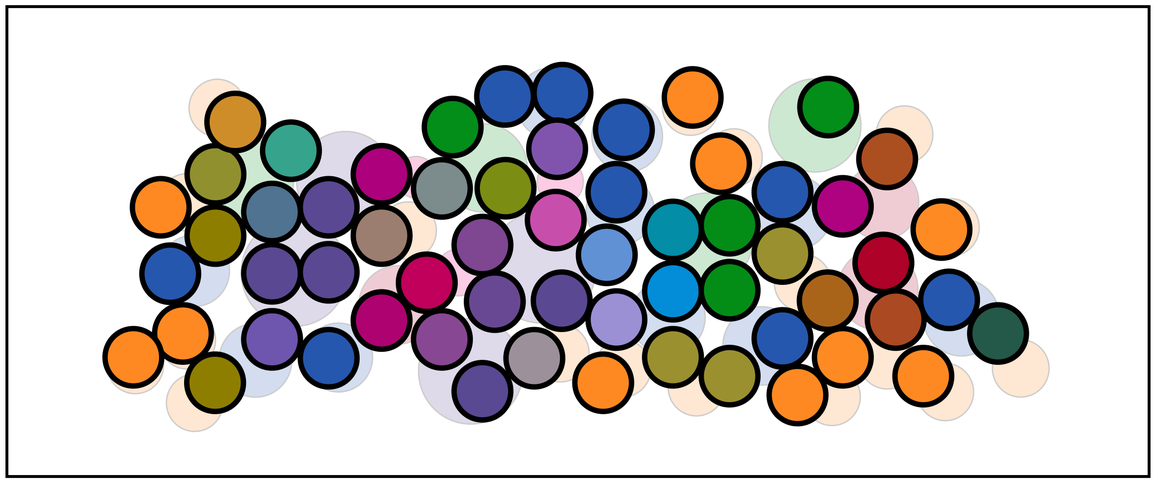


underlying
truth

Forward problem

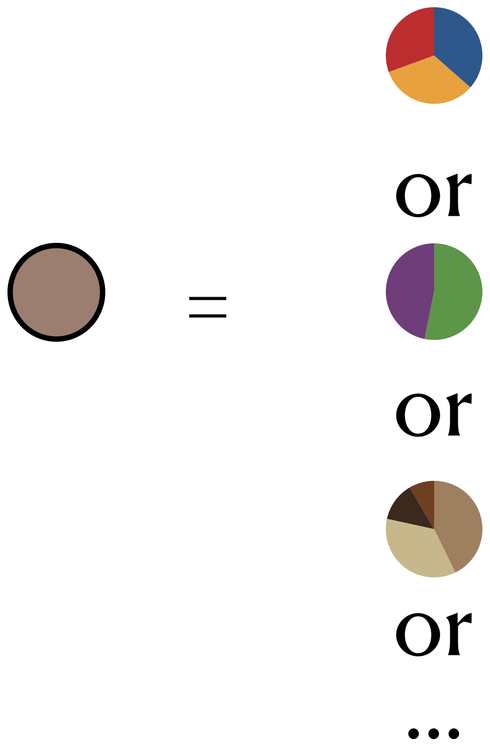


underlying truth

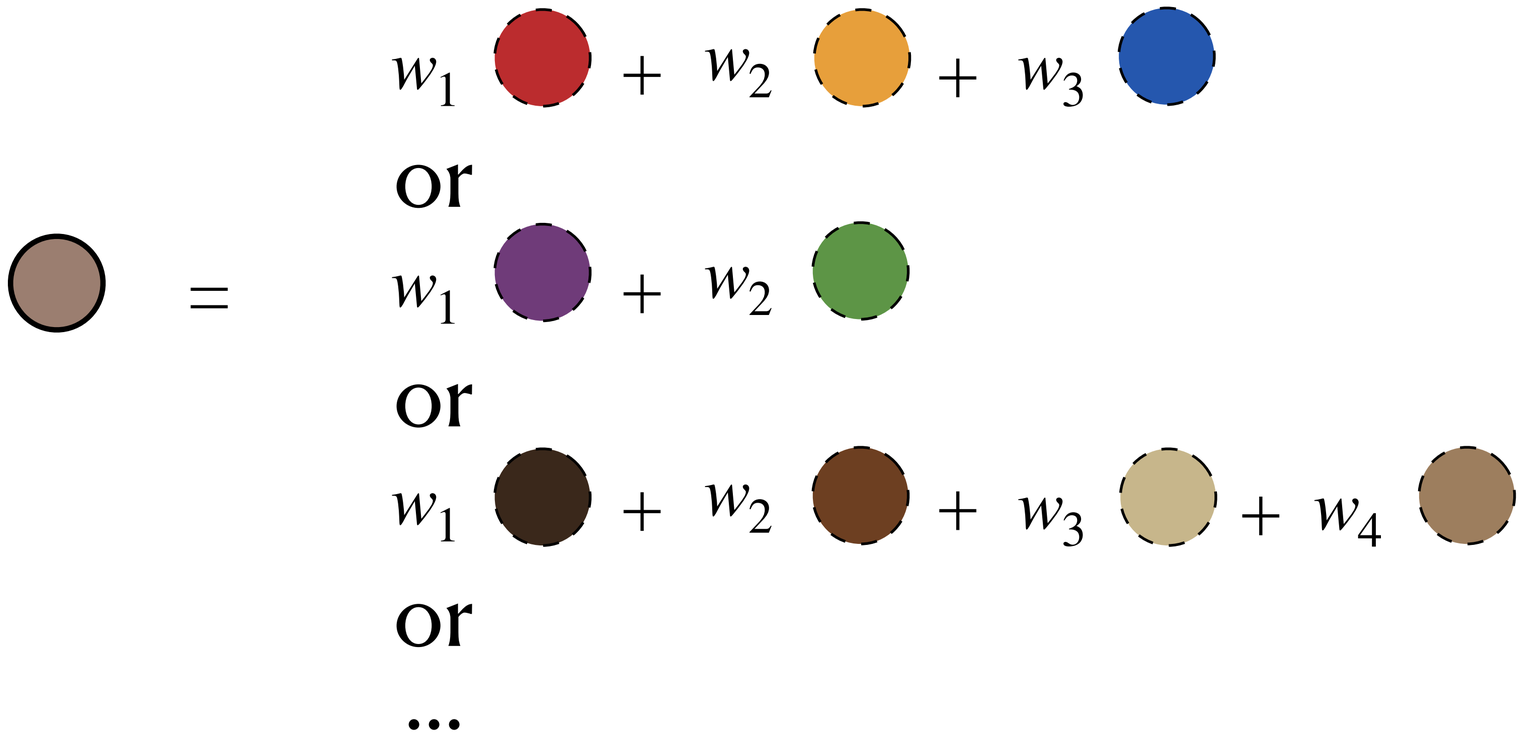


observations

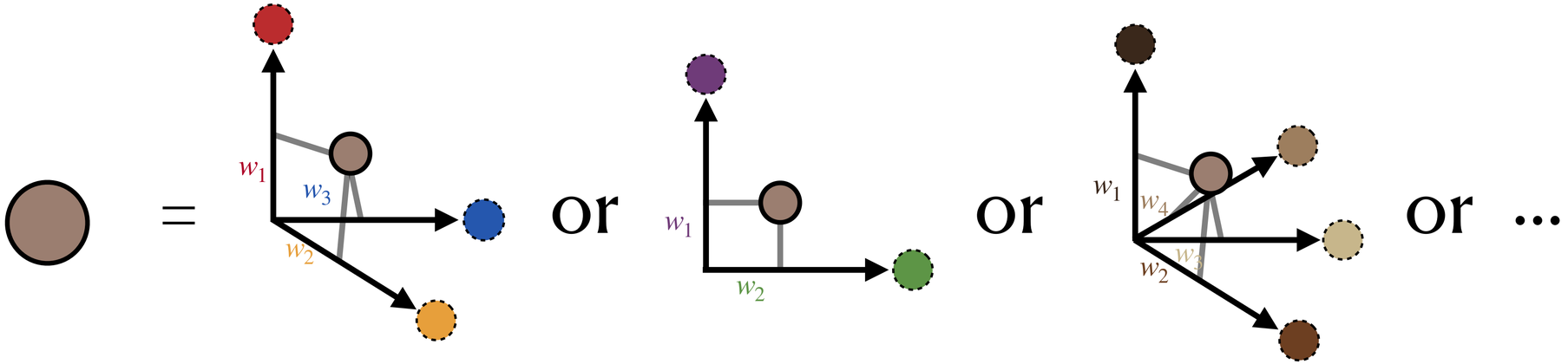
What is the **true mixture**?



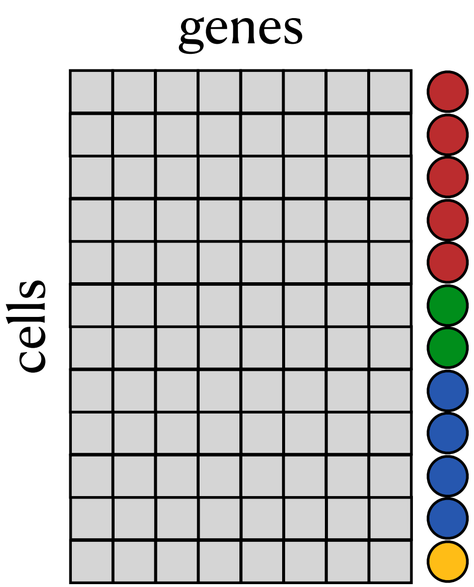
The same thing expressed **algebraically**.



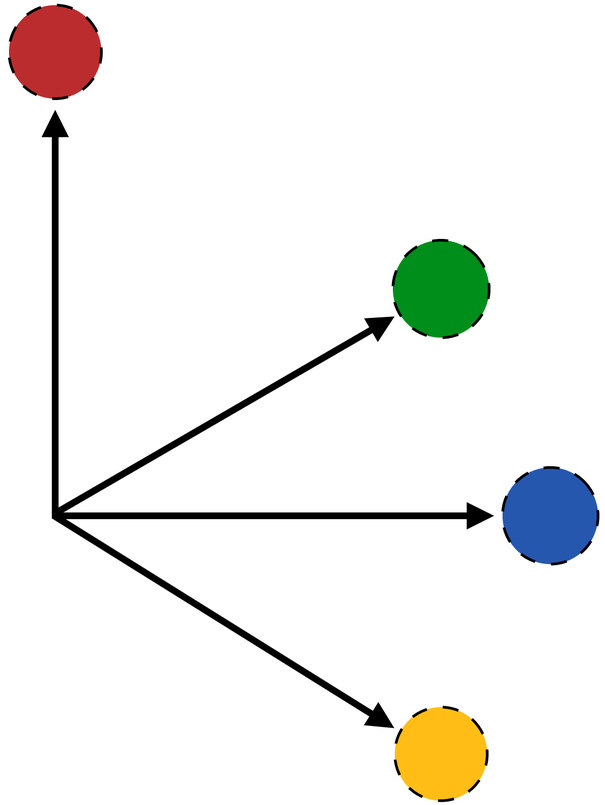
The same thing expressed **geometrically**.



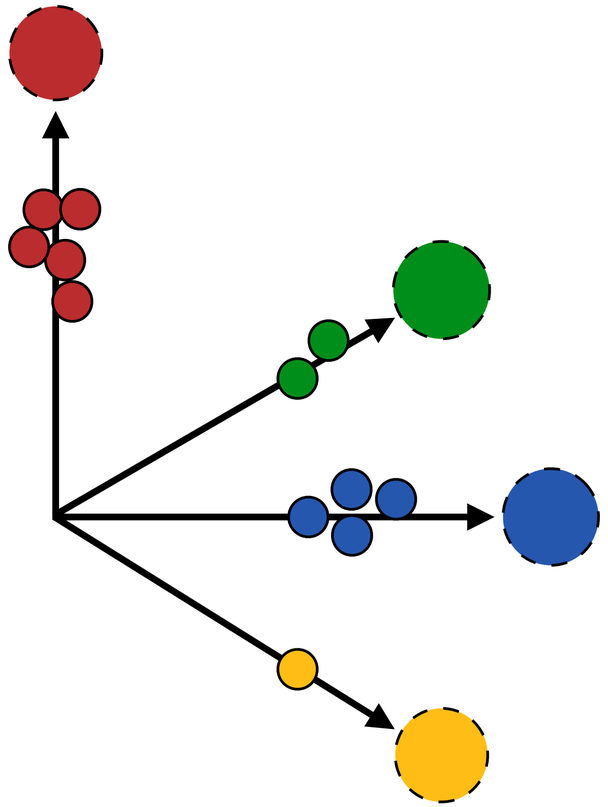
Leverage **prior knowledge** to define a **basis**.



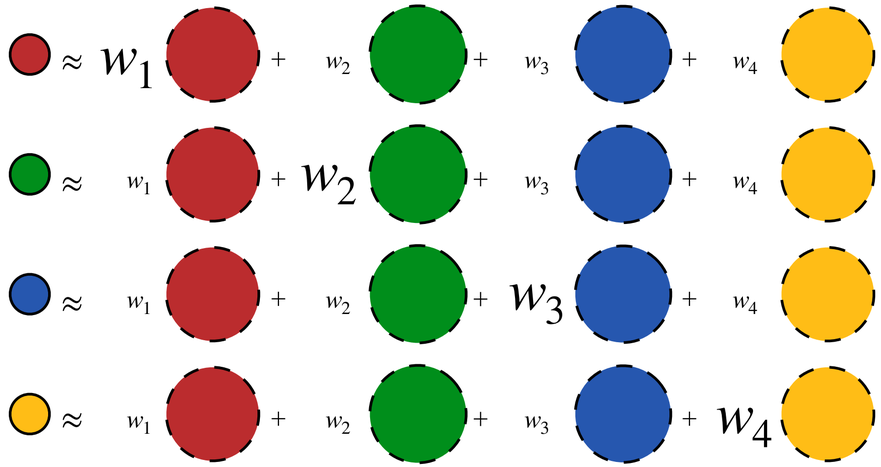
matrix of counts from
an **annotated reference**
single-cell RNAseq data

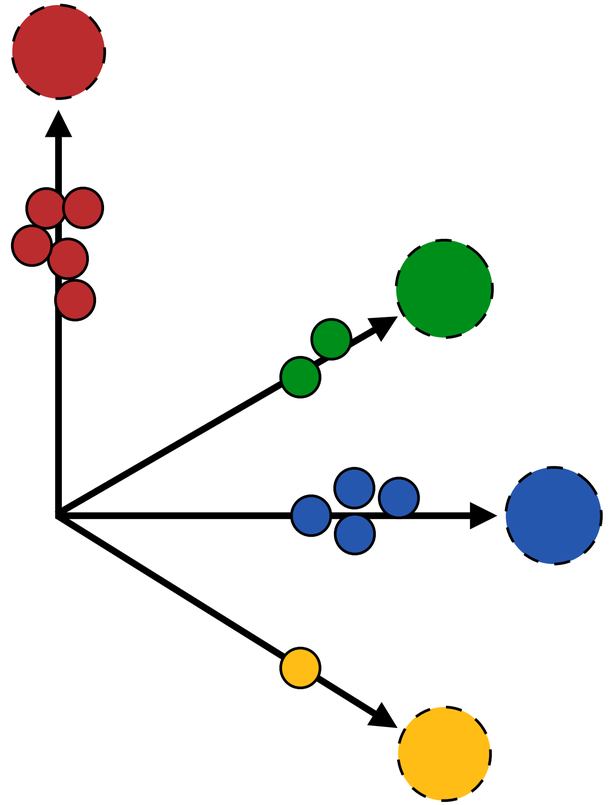


What is a **good basis**?

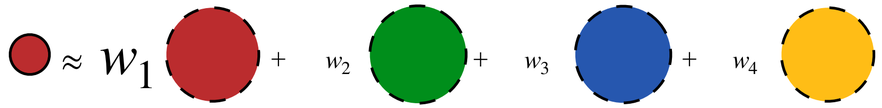


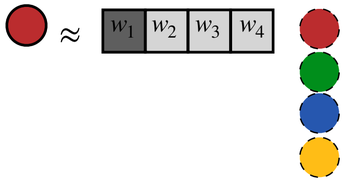
Written **algebraically**.

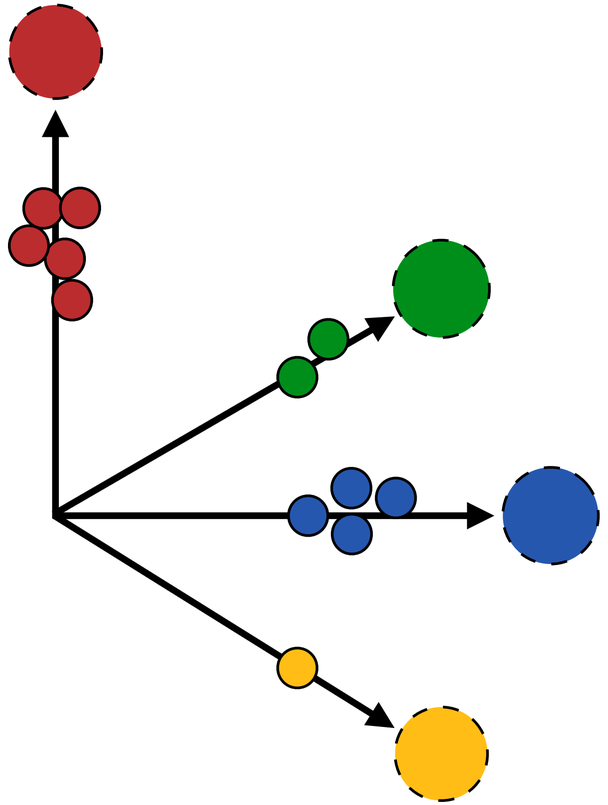




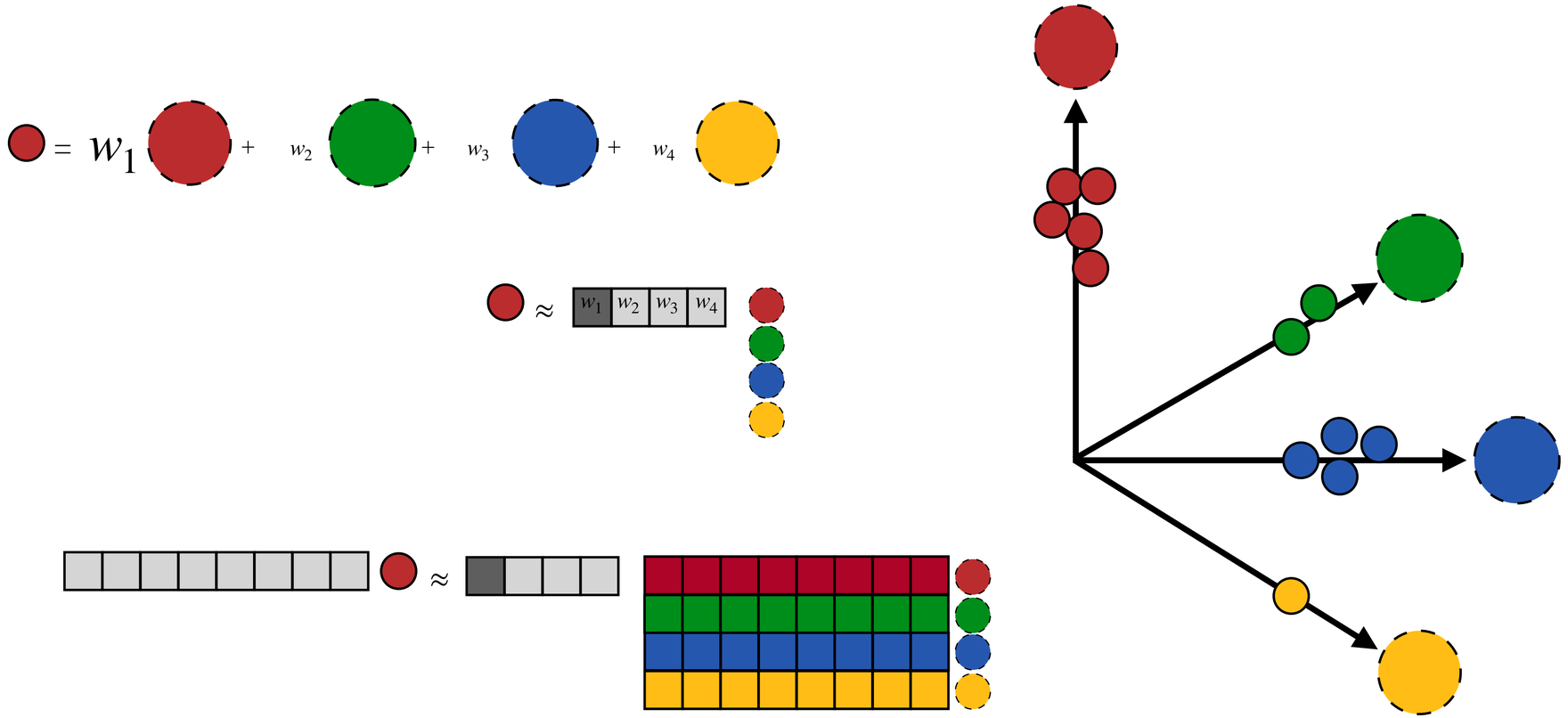
Written as a **vector multiplication**.



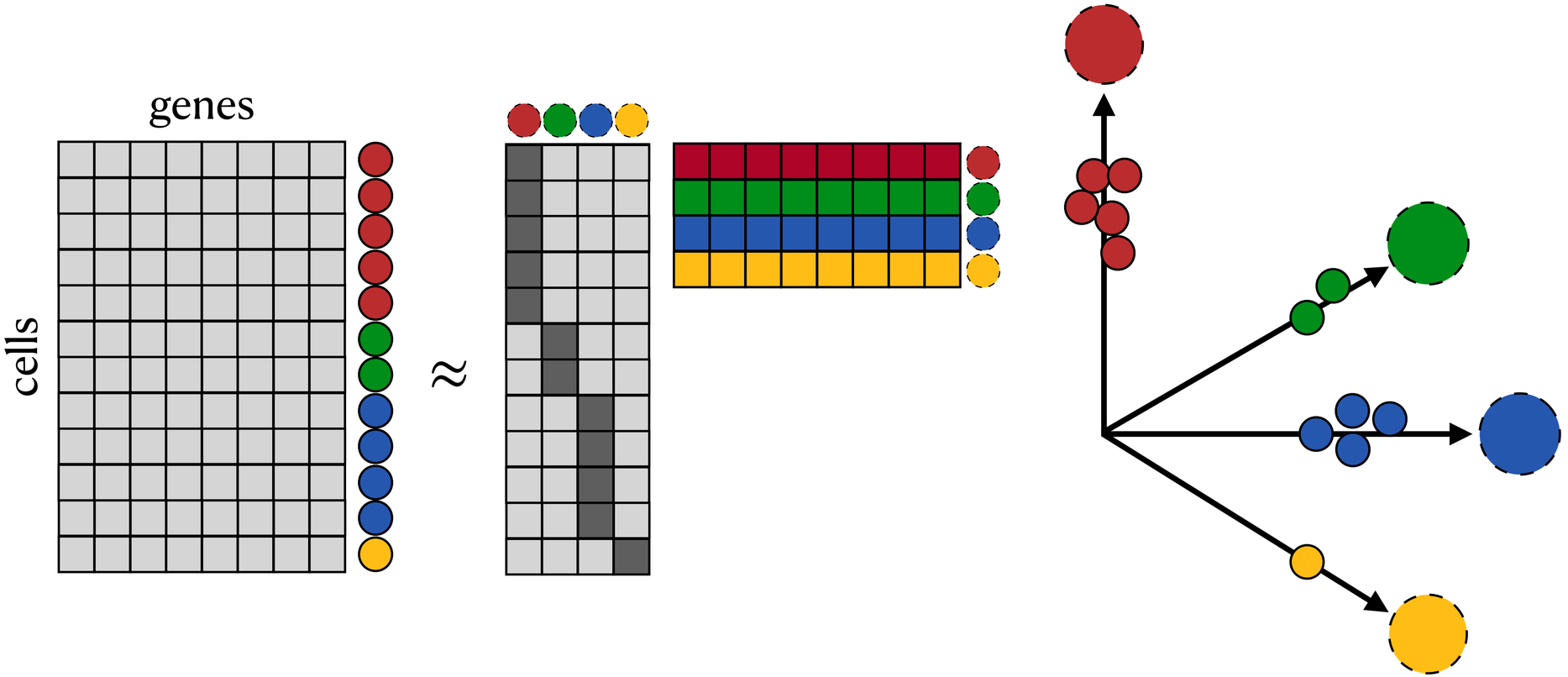




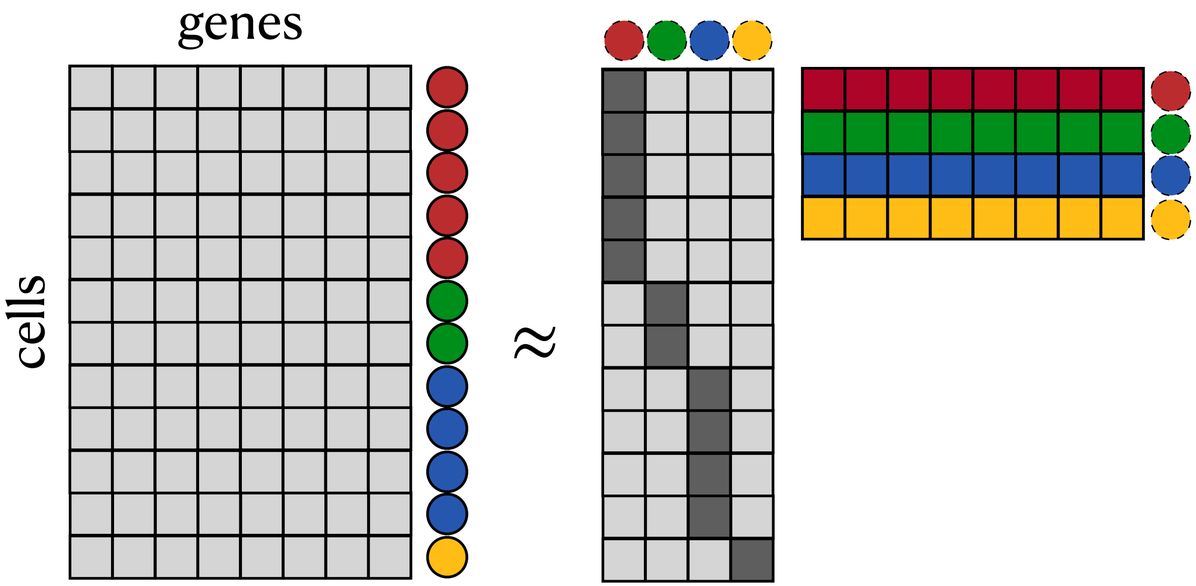
Written as a **vector multiplication**.



Matrix factorization.

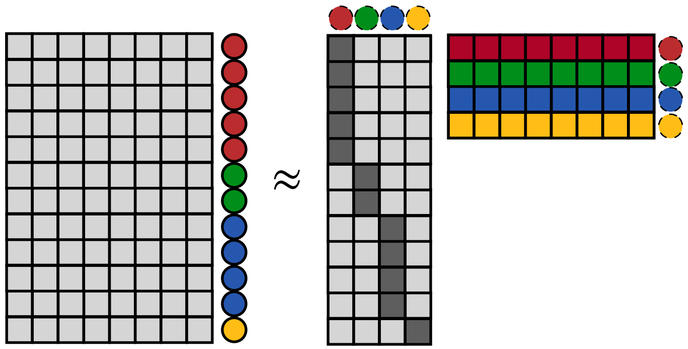


Linear algebra notation.



$$X \approx WH$$

How to find $\boldsymbol{W}$ and $\boldsymbol{H}$?

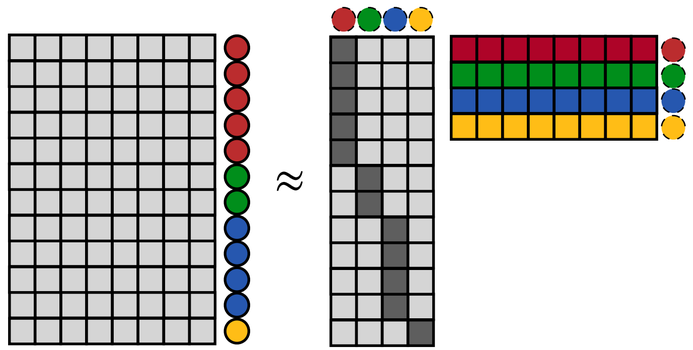


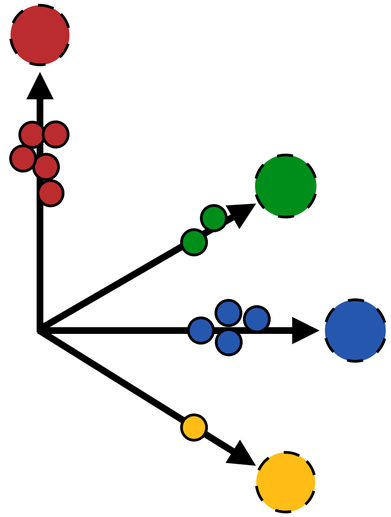
$$X \approx WH$$

$$W, H = argmin \, ||X - WH||^2$$

such that $W, H \geq 0$

Non-negative Matrix Factorization (**NMF**).

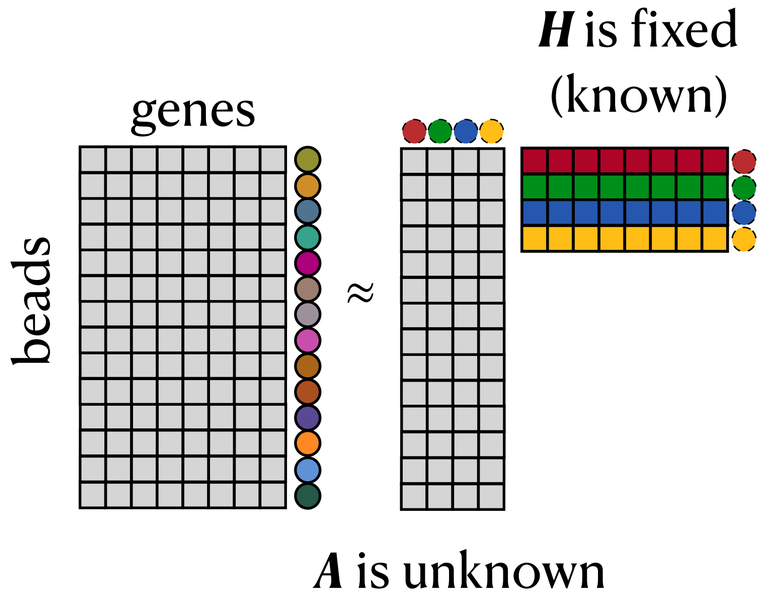


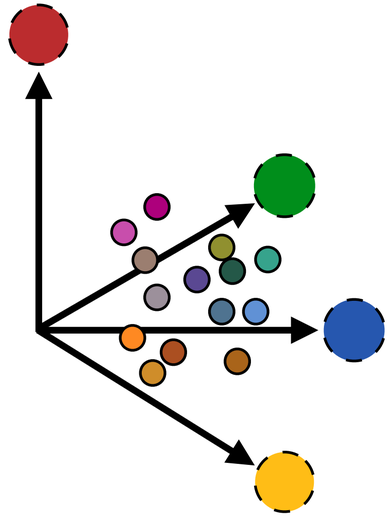


$$W, H = argmin \, ||X - WH||^2$$

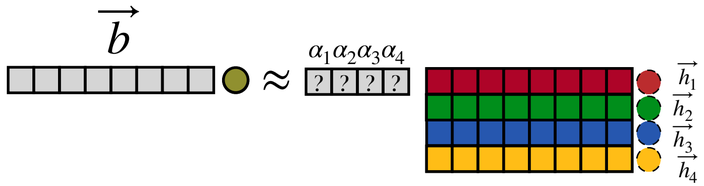
such that $W, H \geq 0$

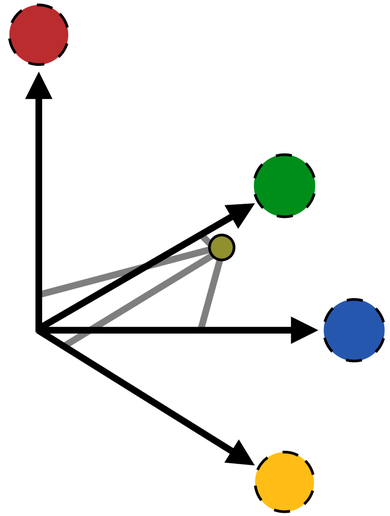
Deconvolution of the mixed spatial beads.



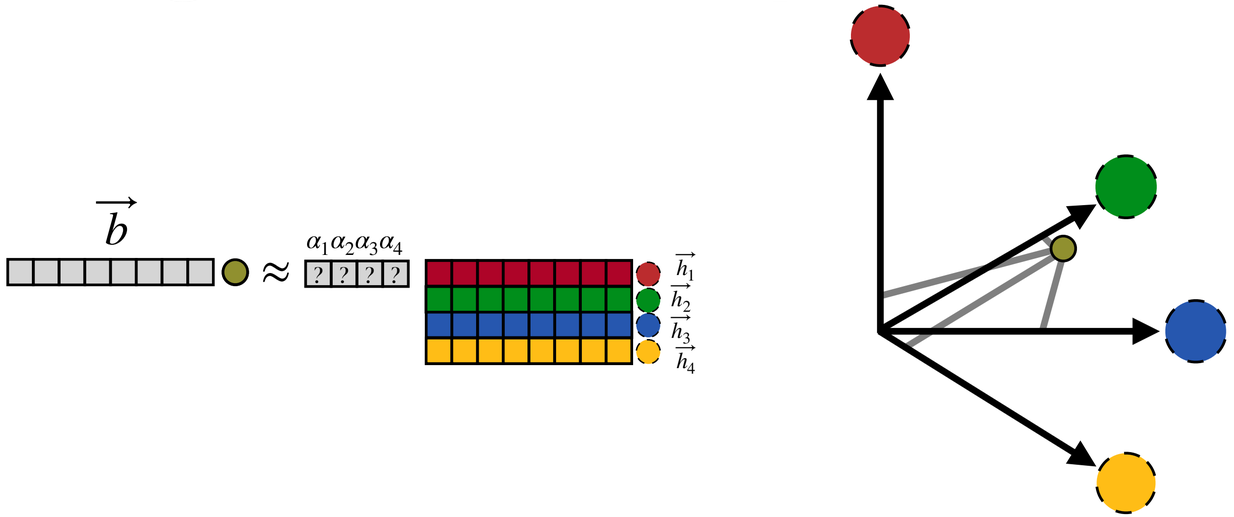


Finding the weights is doing **projection.**





Non-Negative Least Squares (**NNLS**).



$$b \approx \alpha_1 . \overrightarrow{h_1} + \alpha_2 . \overrightarrow{h_2} + \alpha_3 . \overrightarrow{h_3} + \alpha_4 . \overrightarrow{h_4}$$

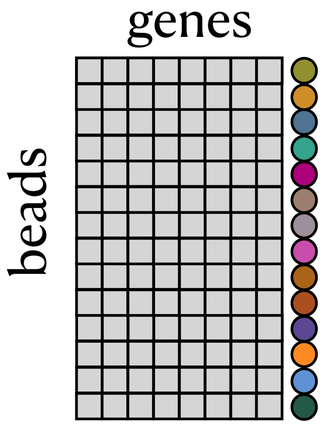
$$b \approx \overrightarrow{\alpha}^T H$$

$$B \approx AH$$

$$A = argmin \, ||B - AH||^2$$

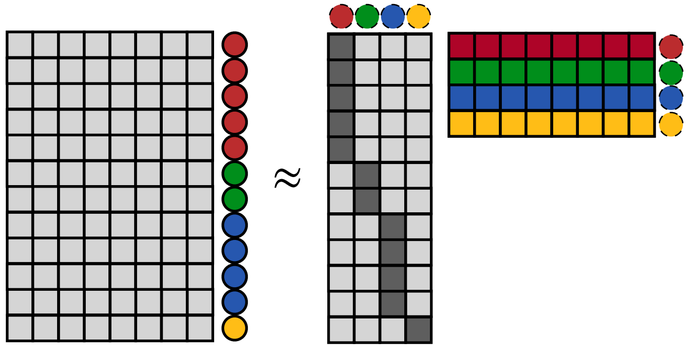
such that $A \geq 0$

Did it work? **Validation?**

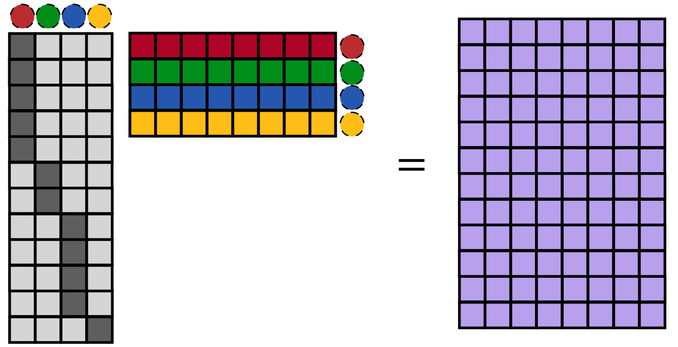


Granule Purkinje

Let's take a fresh look at our model.



$$X \approx WH$$

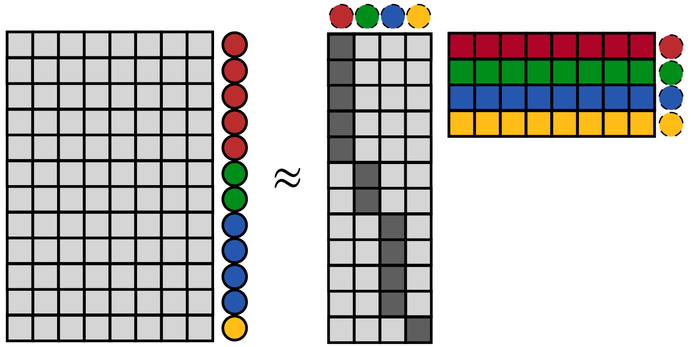


$$WH := \tilde{X}$$

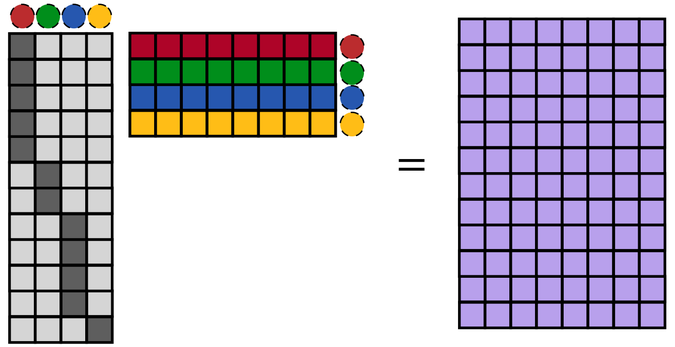
$$W, H = argmin\,||X - WH||^2 = argmin\,||X - \tilde{X}||$$

such that $W, H \geq 0$ such that $\tilde{X} = WH$ and $W, H \geq 0$

Without the constraints, it is exactly **PCA**!



$$X \approx WH$$

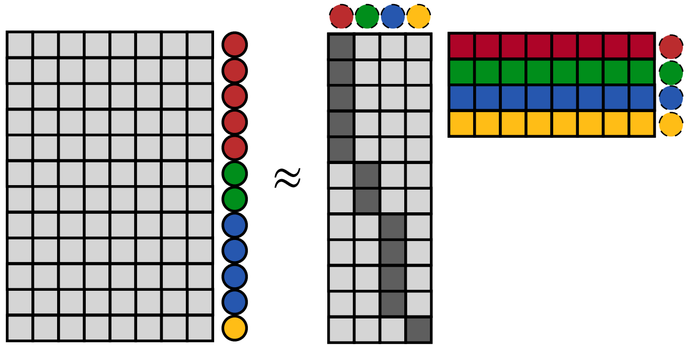


$$WH := \tilde{X}$$

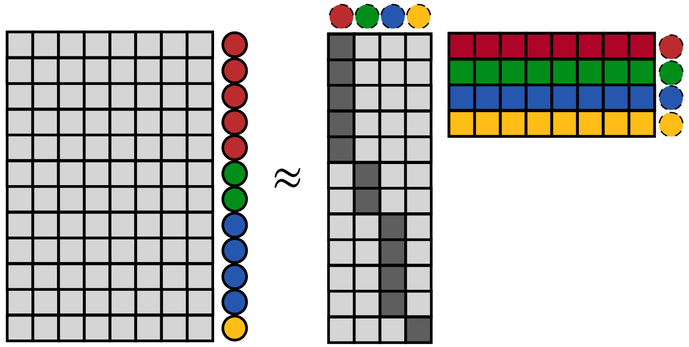
$$W, H = \underset{\text{such that } W,H \geq 0}{argmin} ||X - WH||^2 = \underset{\text{such that } \tilde{X} = WH \text{ and } W,H \geq 0}{argmin} ||X - \tilde{X}||^2$$

$$\tilde{X} = argmin ||X - \tilde{X}||^2 \quad \textbf{PCA!}$$

But is this a math model, or a stats model?



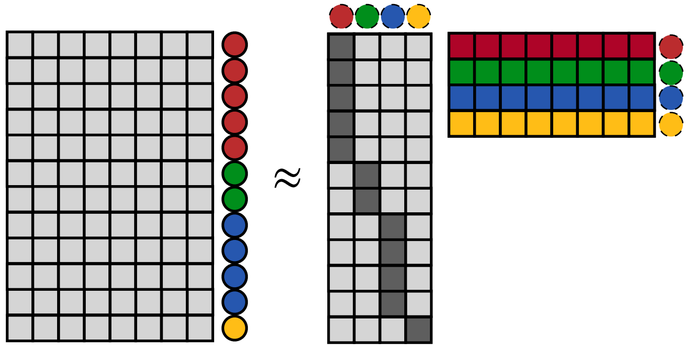
But is this a math model, or a stats model?



a **bilinear** model

$$\tilde{X} = argmin\,||X - \tilde{X}||^2$$

But is this a math model, or a stats model?



a **bilinear** model

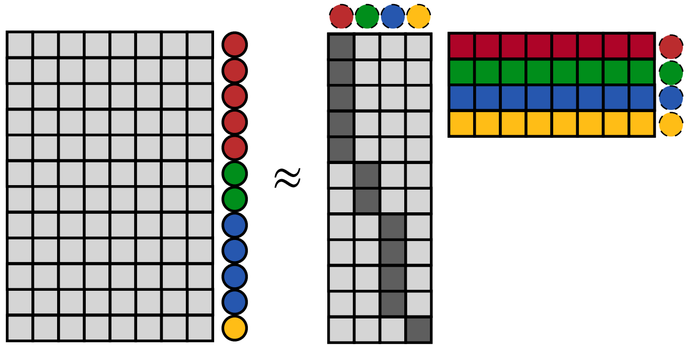
$$\tilde{X} = argmin\,||X - \tilde{X}||^2$$

$\tilde{X} = argmin\,||X - \tilde{X}||^2$ is the

maximum likelihood estimator

when $X_{ij} \sim \mathcal{N}(\tilde{X}_{ij}, \sigma^2)$

What if we want a **different likelihood**?



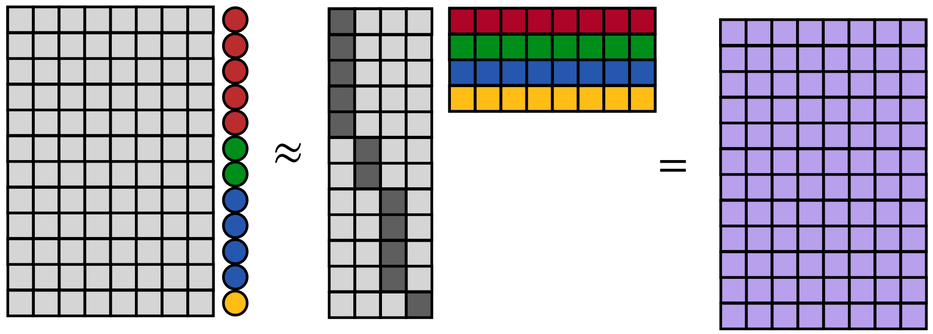
a **bilinear** model

$$\tilde{X}_{ij} = argmin \sum e^{\tilde{X}_{ij}} - X_{ij}\tilde{X}_{ij}$$ is the

maximum likelihood estimator

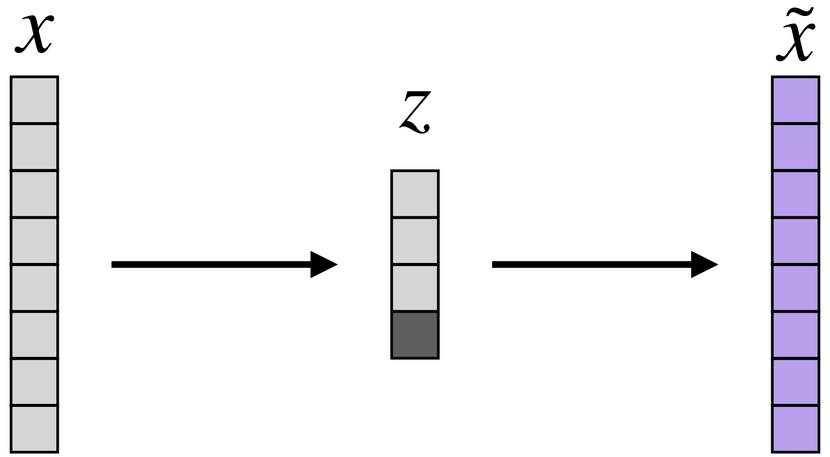
when $X_{ij} \sim \text{Poisson}(e^{\tilde{X}_{ij}})$

Focus on one observation.



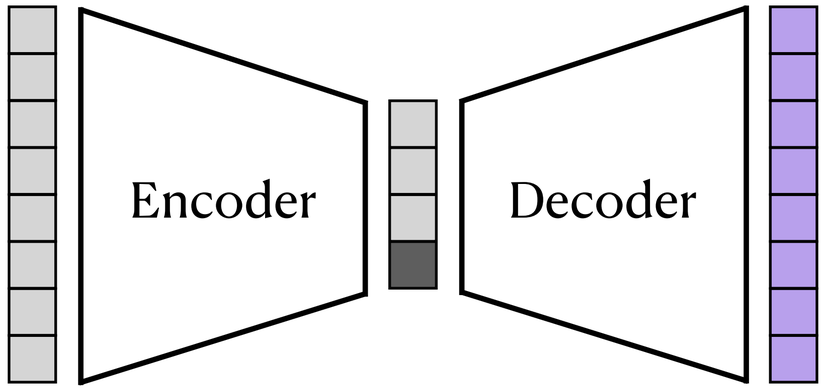
$$X \approx \tilde{X}$$

Dimensionality Reduction.



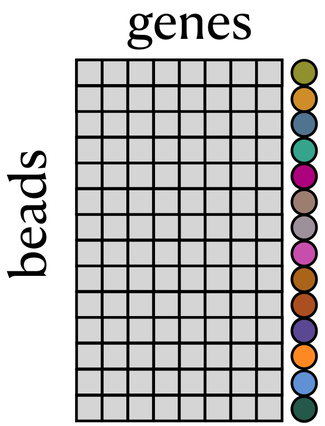
Representation Learning.

Dimensionality Reduction.
Representation Learning.



$$min\ \mathrm{d}(X, \tilde{X})$$

Which approach should we use?

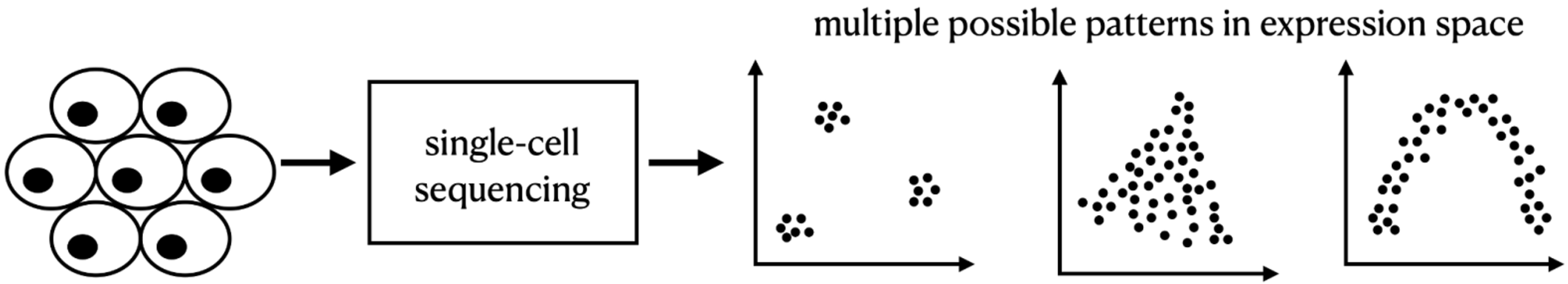


Granule

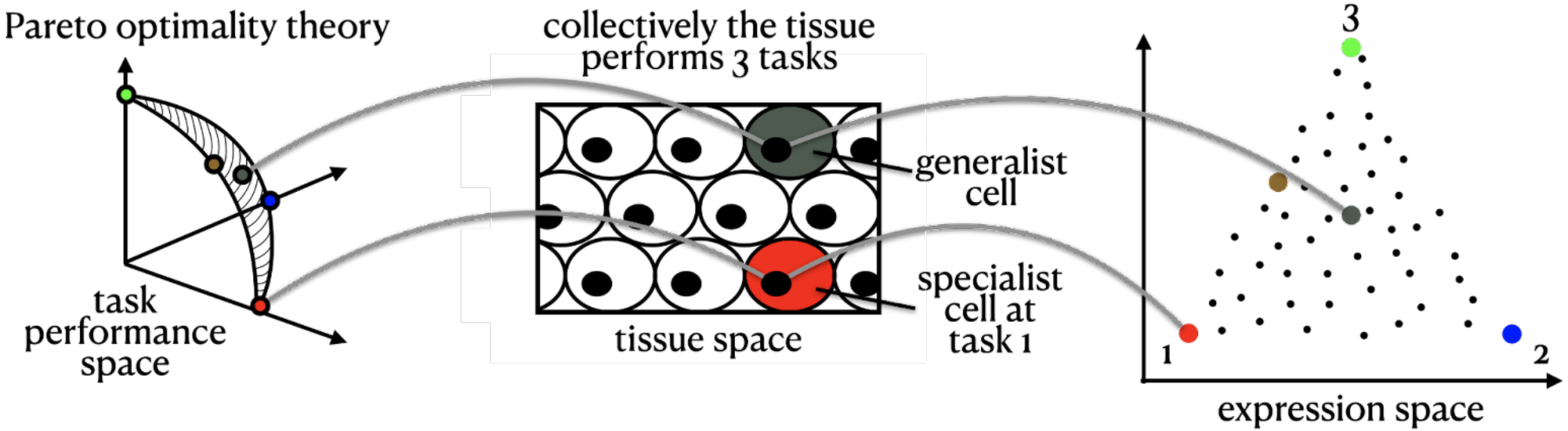
Purkinje

3. What is the role of local interactions in shaping single cell gene expression?

Single-cell RNA sequencing data in expression space reveals low-dimensional structure with various patterns.



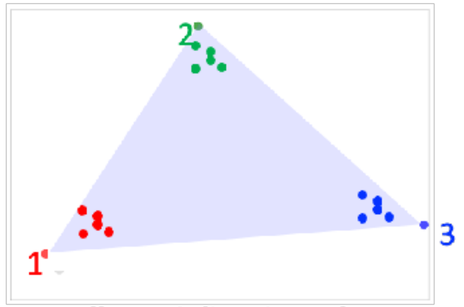
Maximal performance in multiple tasks under limited budget — the cells for a polytope in expression space.



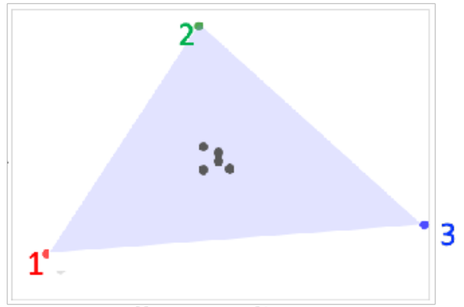
Shoval et al. Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space (2012)
Adler et al. Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type (2019)

Cells work as a collective to optimize tissue performance.

Mathematically, the optimal solution is either all specialists, or all generalists.

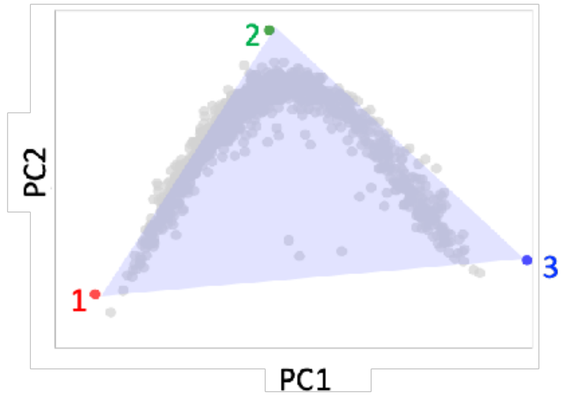


All specialists in tasks

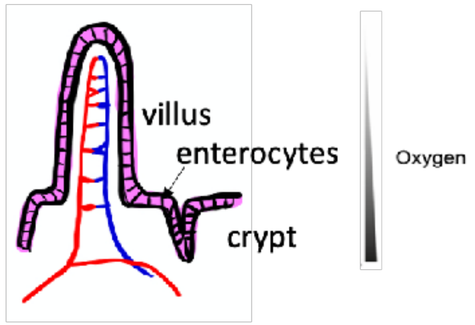


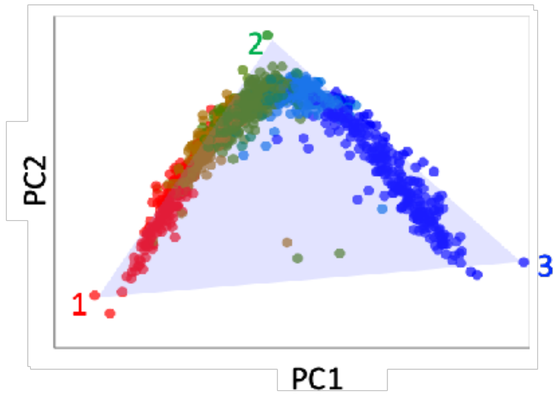
All generalists

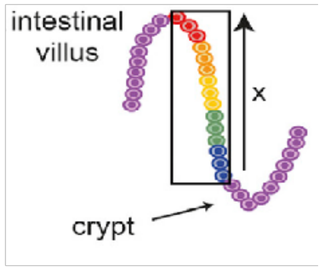
But in real data, we often observe a continuum of expression. What could it be due to?



A continuum in expression can be due to a spatial gradient in the tissue.





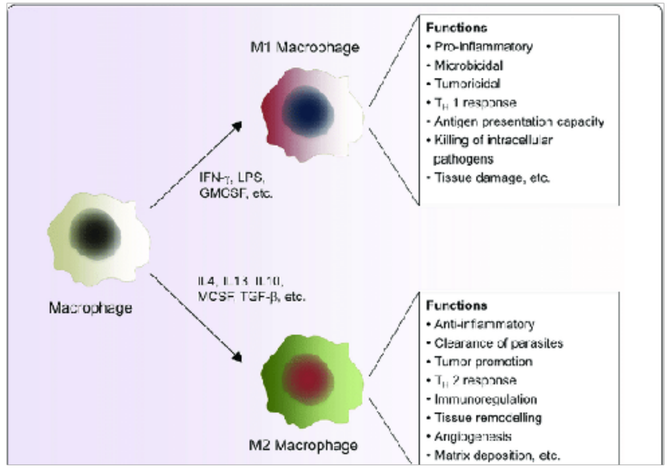


Adler et al. Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type (2019)

Local cell-cell interactions play a role in determining cell function.

"On-demand" functions

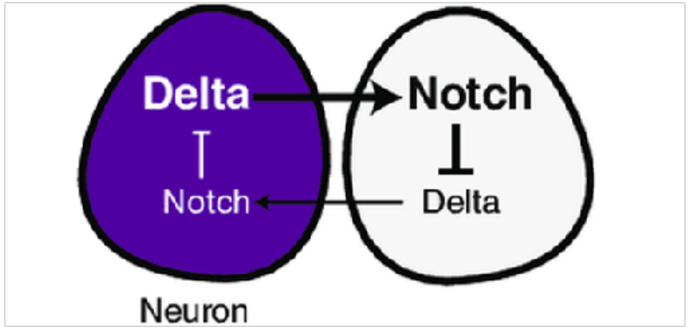
Ex: Macrophages pro- and anti-inflammatory signal-dependent response



Saqib et al. Phytochemicals as modulators of M1-M2 macrophages in inflammation (2018)

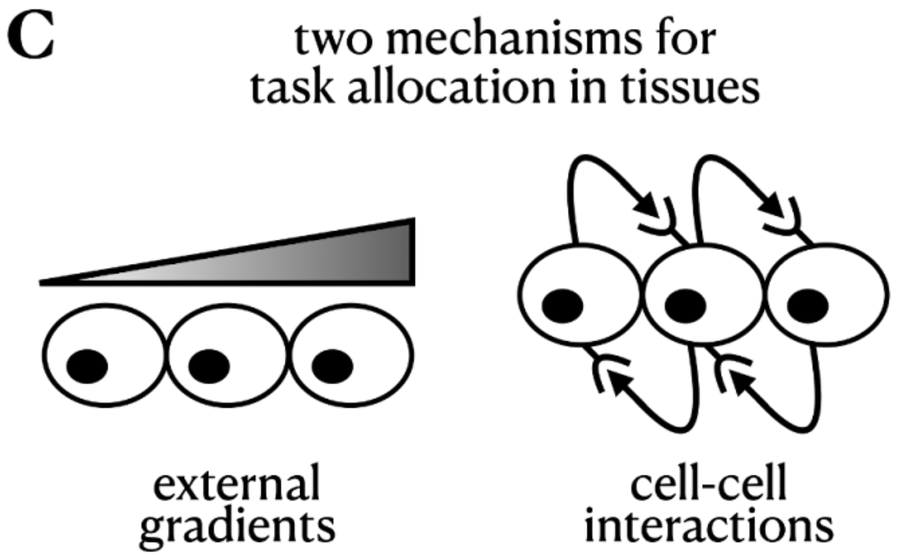
Cell differentiation

Ex: Delta-Notch lateral inhibition of proneural expression



Formosa-Jordan et al. Lateral inhibition and neurogenesis: Novel aspects in motion (2013)

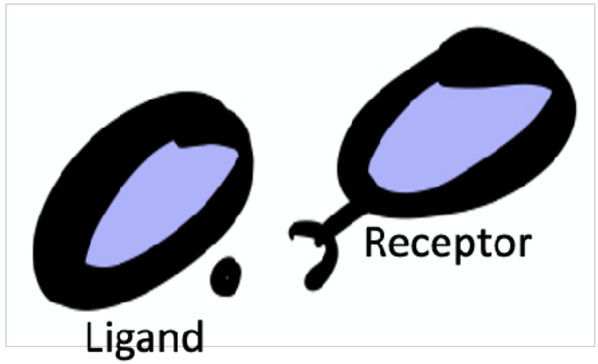
We consider two underlying mechanisms driving the allocation of cells in the tissue.



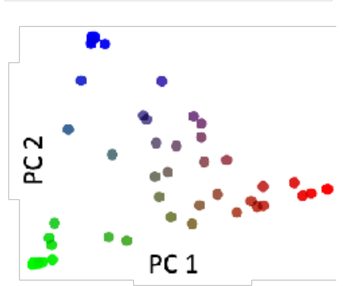
What is the effect of cell-cell interactions on the shape of the gene expression space?

Framework:

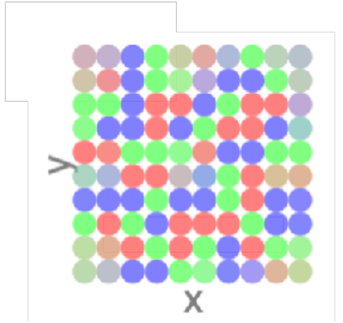
1. Maximal performance in multiple tasks under limited budget
2. Cells work as a collective to optimize tissue performance



Expression space



Tissue space



What is the effect of cell-cell interactions on the shape of the gene expression space?

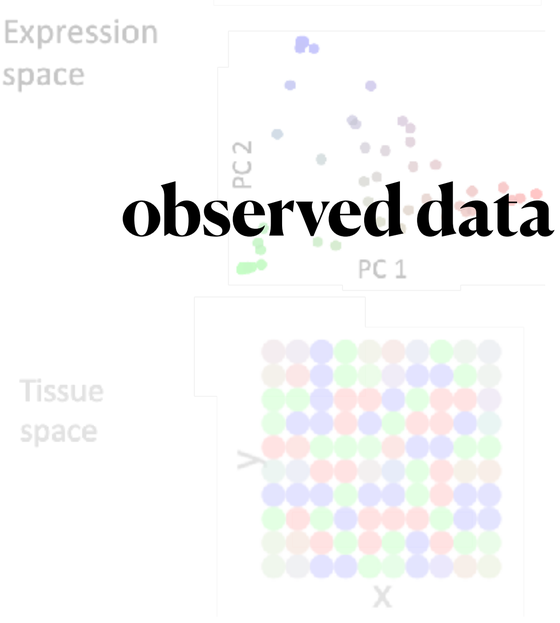
forward problem

Framework:

1. Maximal performance in multiple tasks under limited budget
2. Cells work as a collective to optimize tissue performance

underlying mechanism

observed data



What is the effect of cell-cell interactions on the shape of the gene expression space?

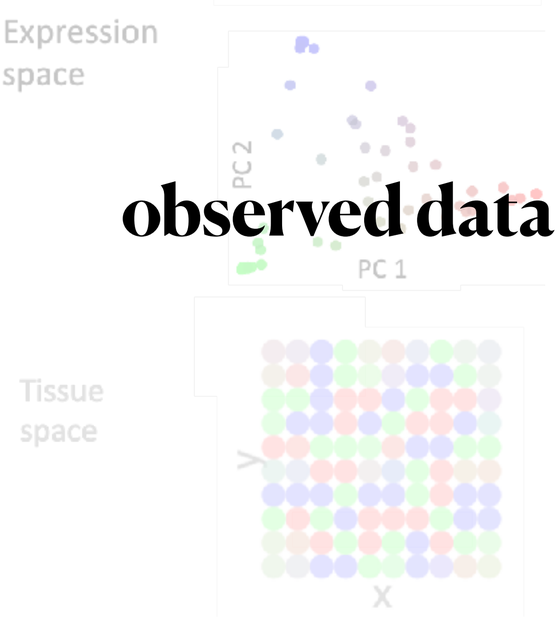
forward problem

Framework:

1. Maximal performance in multiple tasks under limited budget
2. Cells work as a collective to optimize tissue performance

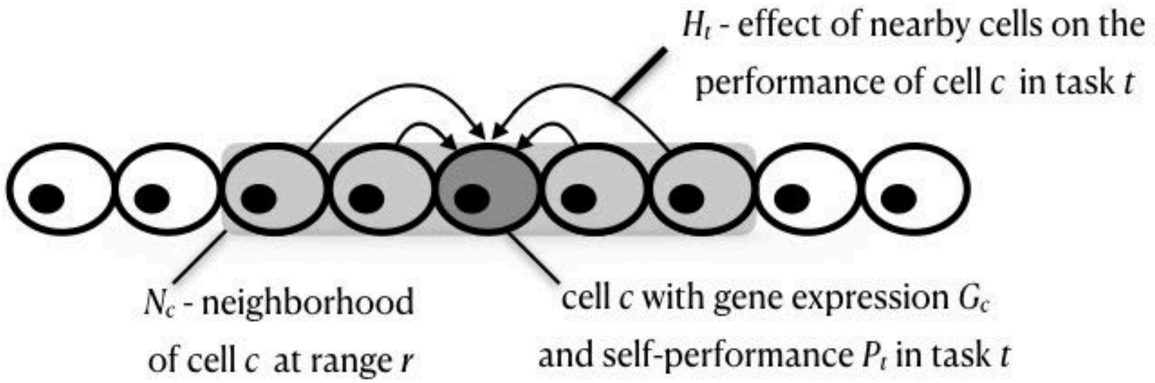
underlying mechanism

observed data



inverse problem

Modeling interactions — an example.



total performance function

$$F = \prod_{t \in \{tasks\}} \left(\sum_{c \in \{cells\}} H_t\left(\{G_i\}_{i \in N_c}\right) P_t(G_c) \right)$$

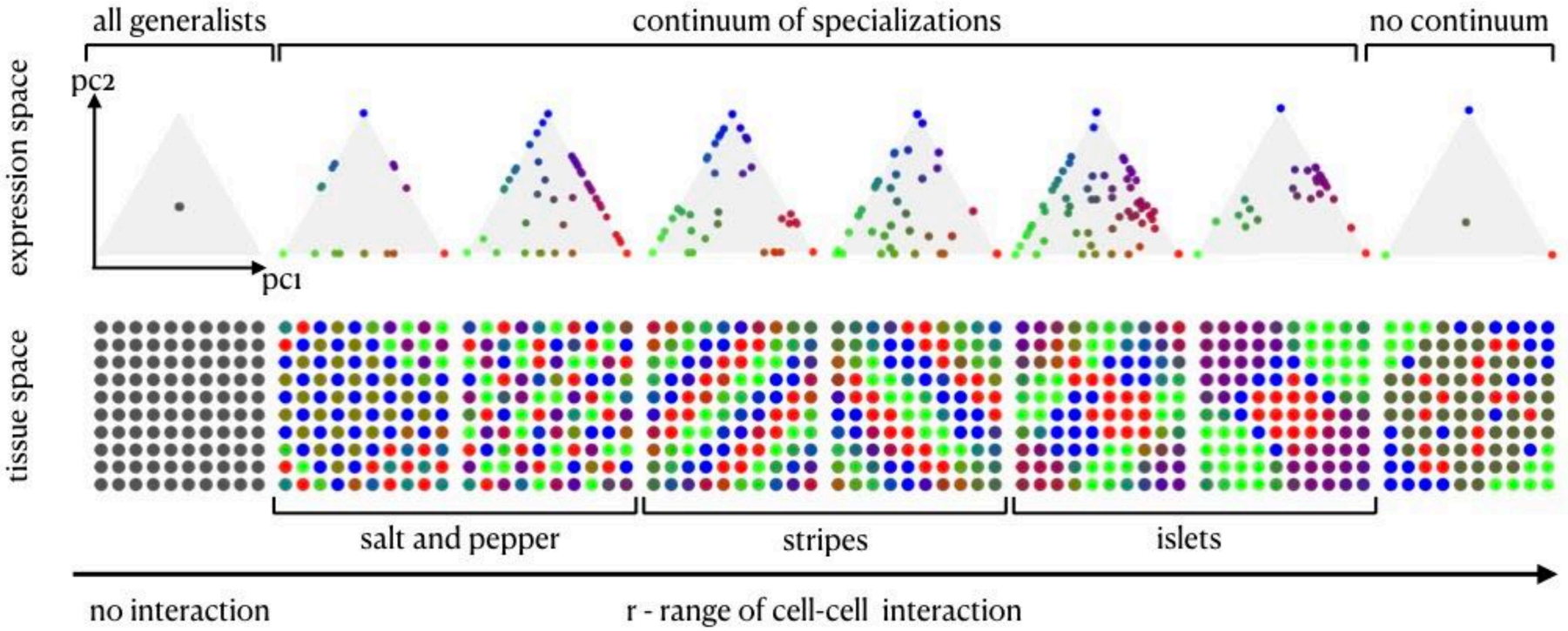
H_t

$avg(P_t(G_i)_{i \in N_c})$

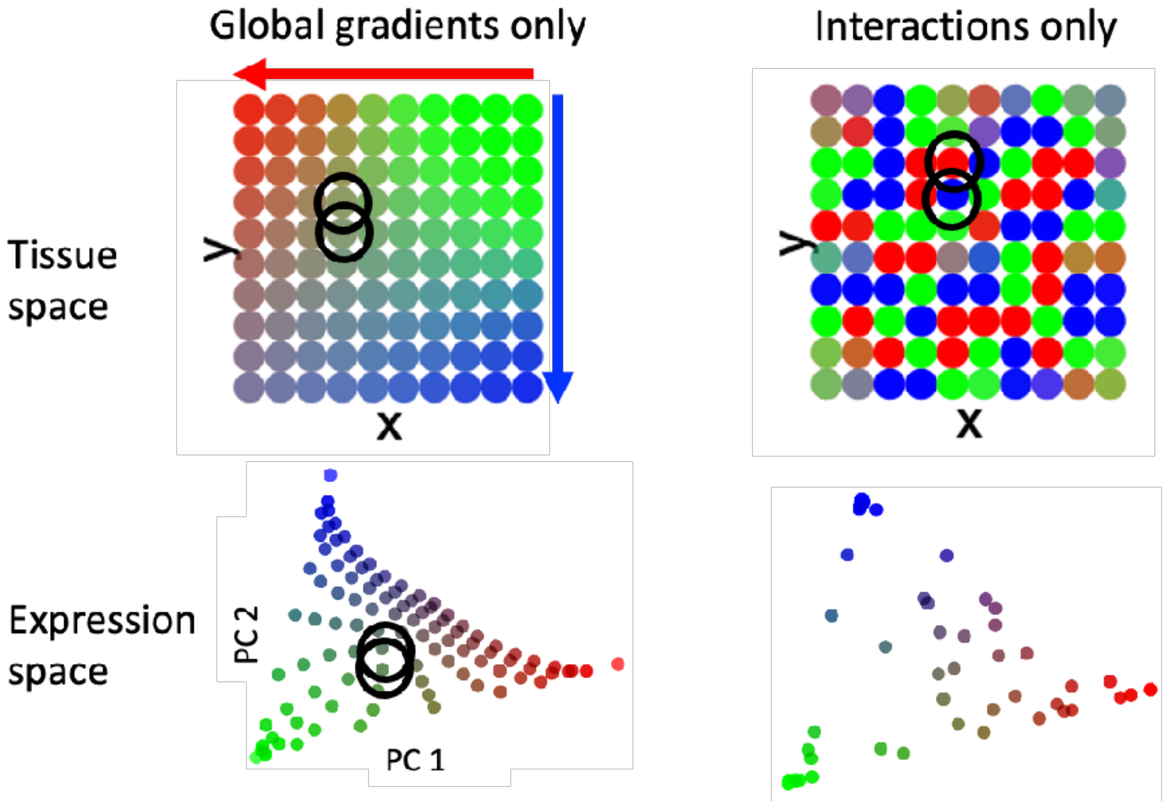
P_t

$\|G_c - G_t^*\|$

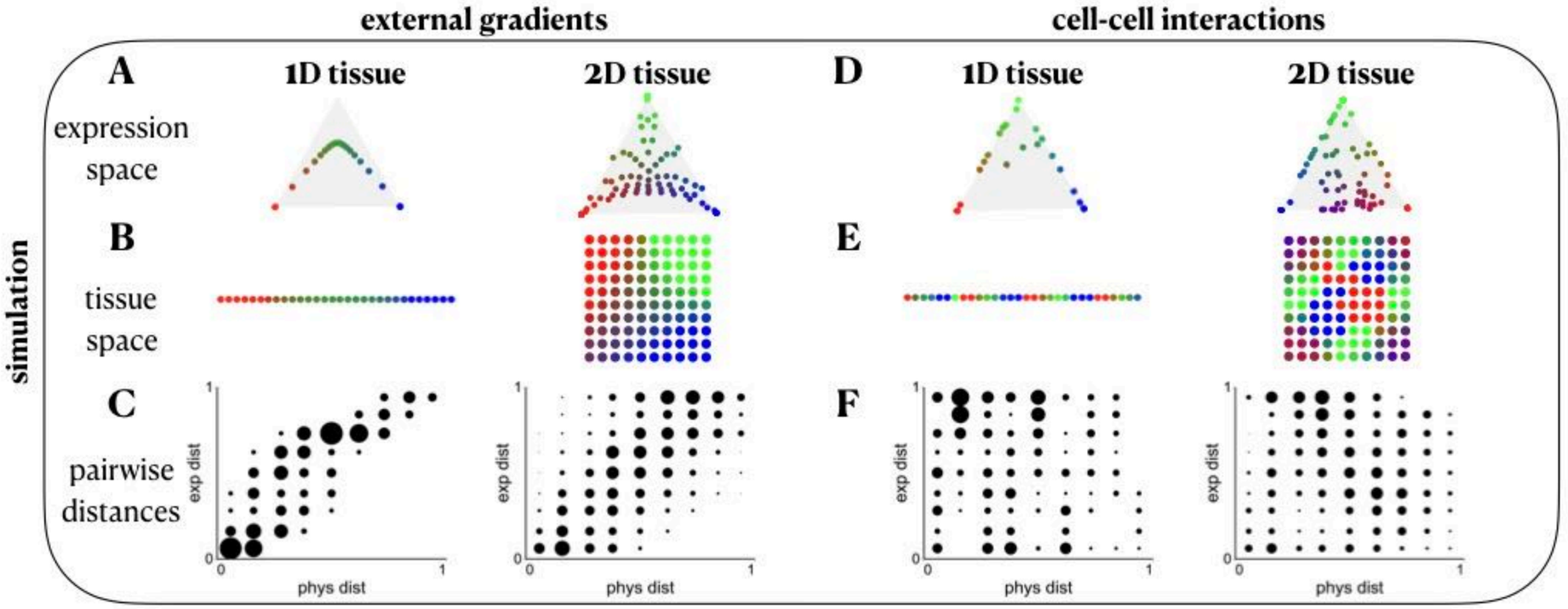
A variety of tissue and expression patterns emerge from the Pareto optimality framework with cell-cell communication mechanisms.



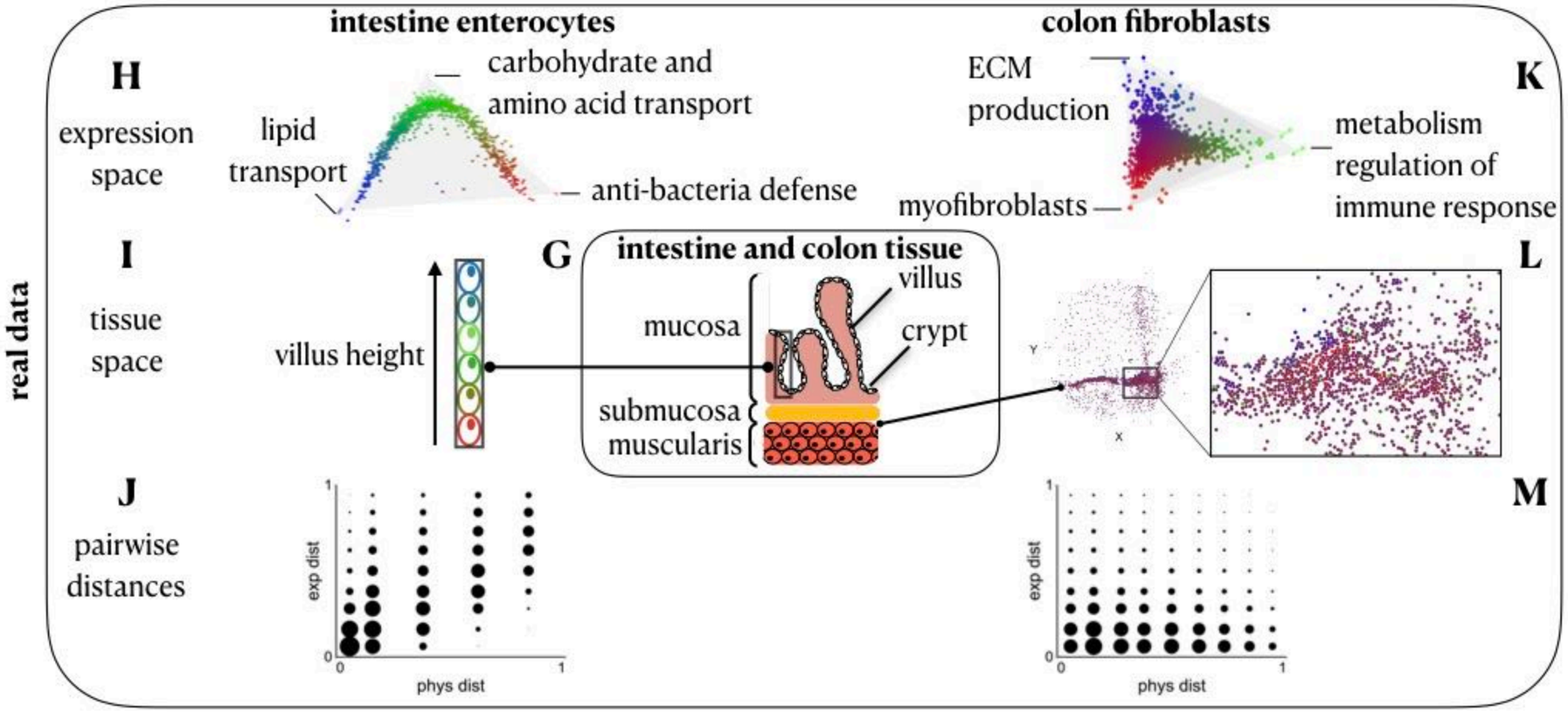
Difference between continuum due to global spatial tissue gradient and local cell-cell interactions:



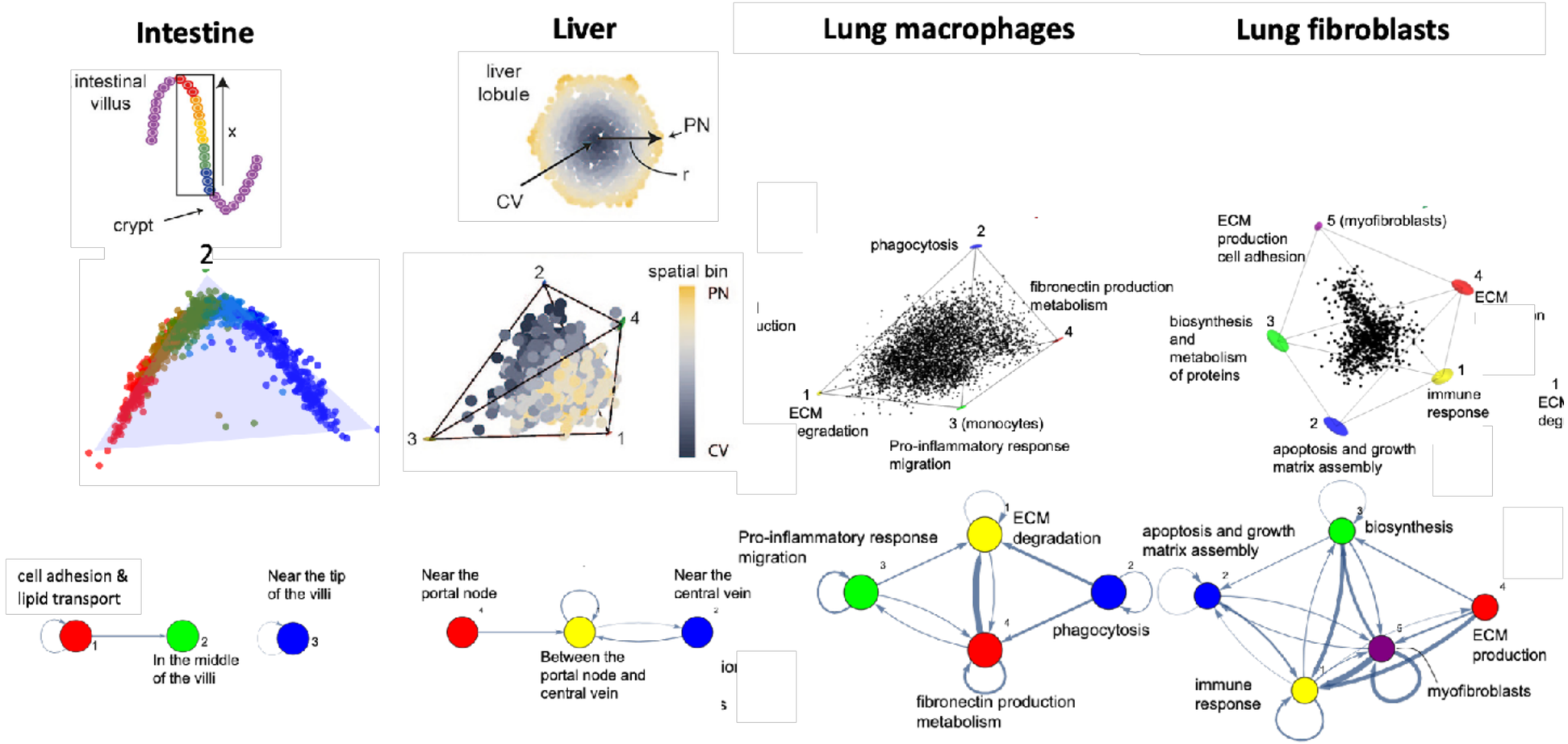
Distinct spatial patterns emerge from external gradients and cell-cell interactions.



Distinct spatial patterns emerge from external gradients and cell-cell interactions.



Interactions between archetypes:



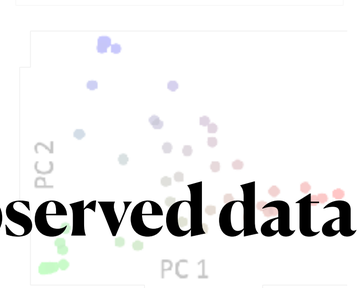
How to adress the inverse problem:

Simulating from the model

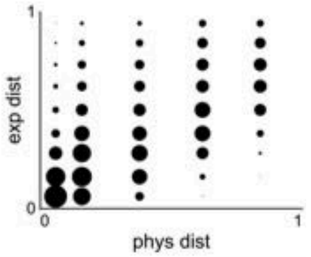
Framework:

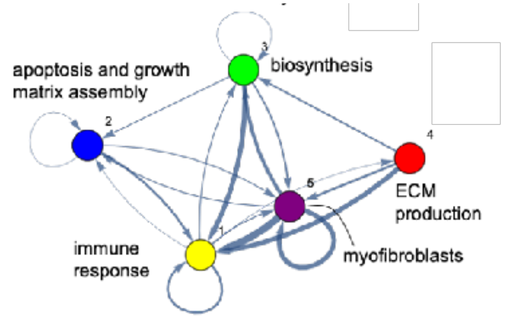
1. Maximal performance in multiple tasks under limited budget
2. Cells work as a collective to optimize tissue performance

underlying mechanism → Expression space



Tissue space → **observed data**

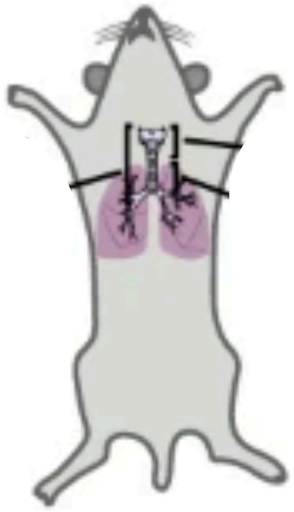


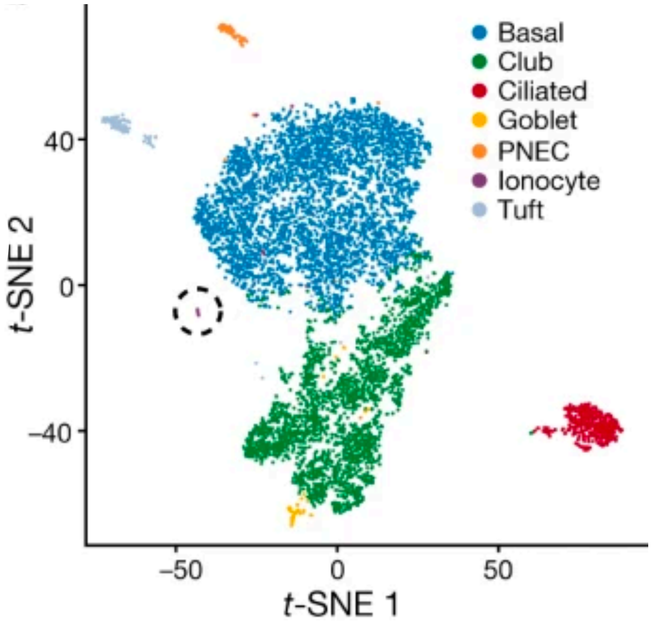


4. What are the cellular differences between health and disease?

Single cell atlases of healthy tissues can identify disease relevant cell types.

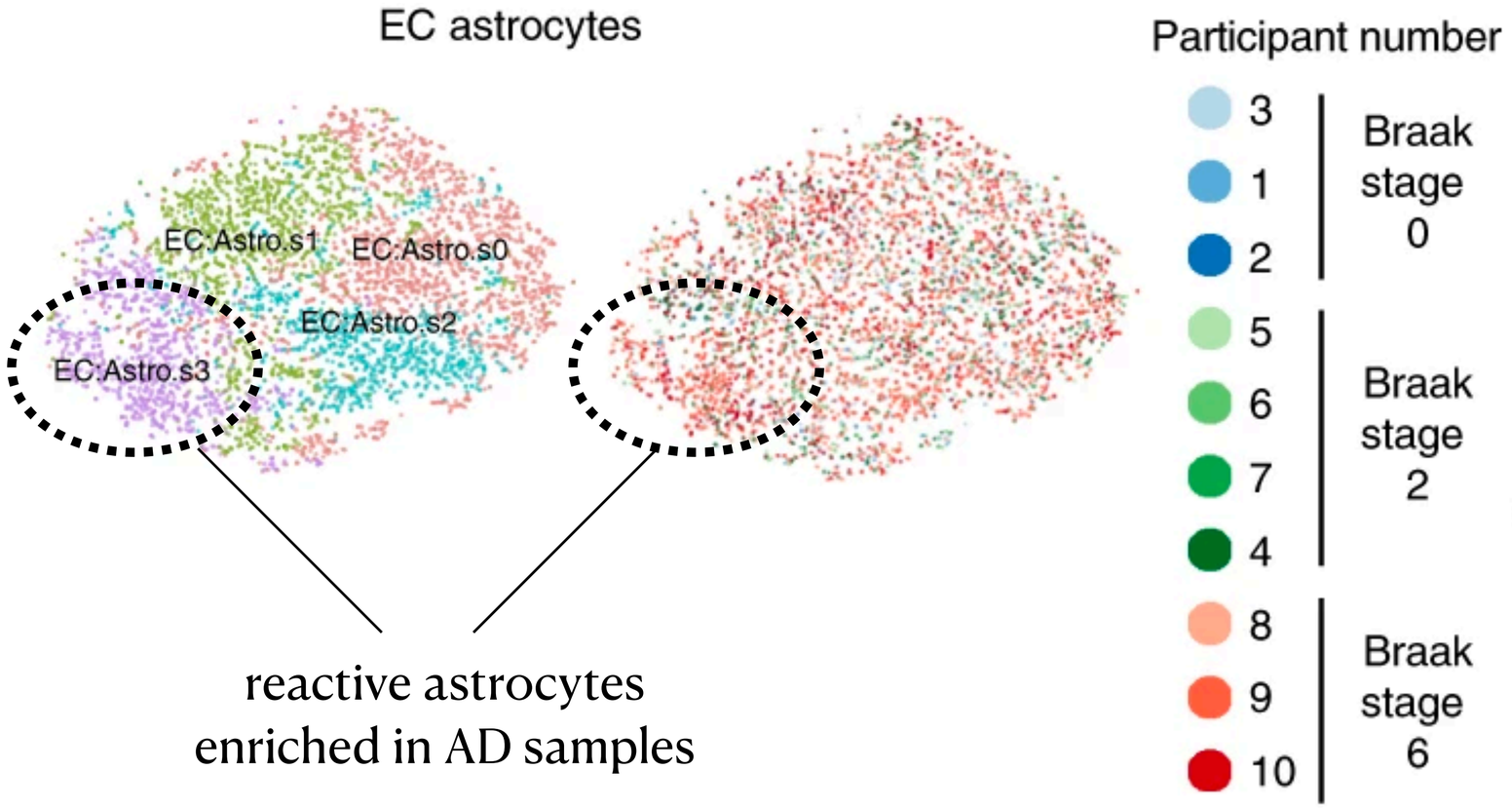
- Mouse airway atlas identifies a novel rare cell type.
- Pulmonary ionocytes play a key role in cystic fibrosis.





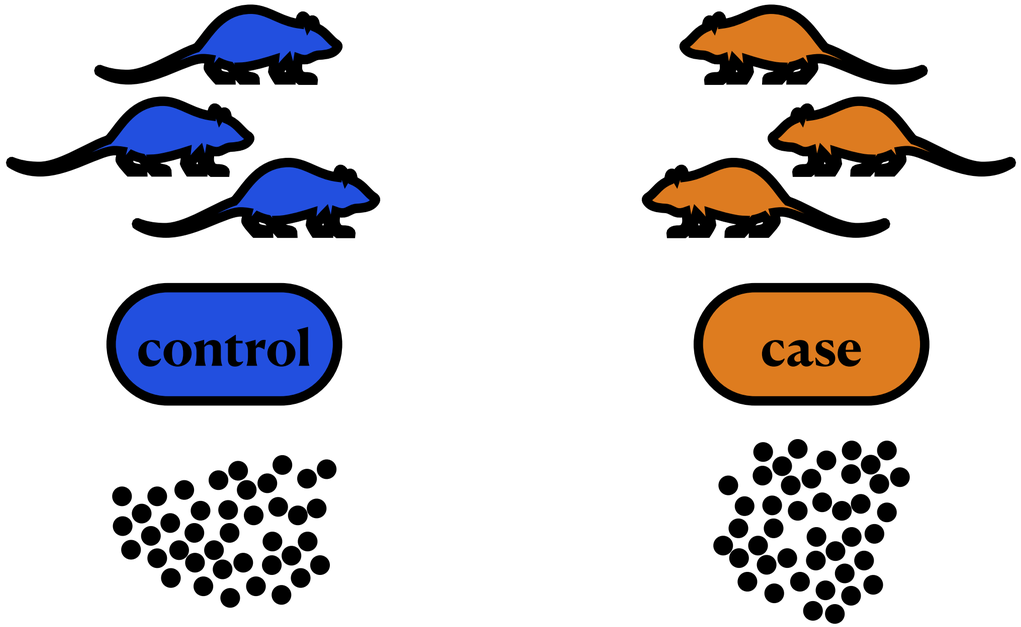
Montoro, D.T., Haber, A.L., Biton, M. et al. Nature (2018).

Designing case-control single cell experiments to discover disease relevant subpopulations of cells.

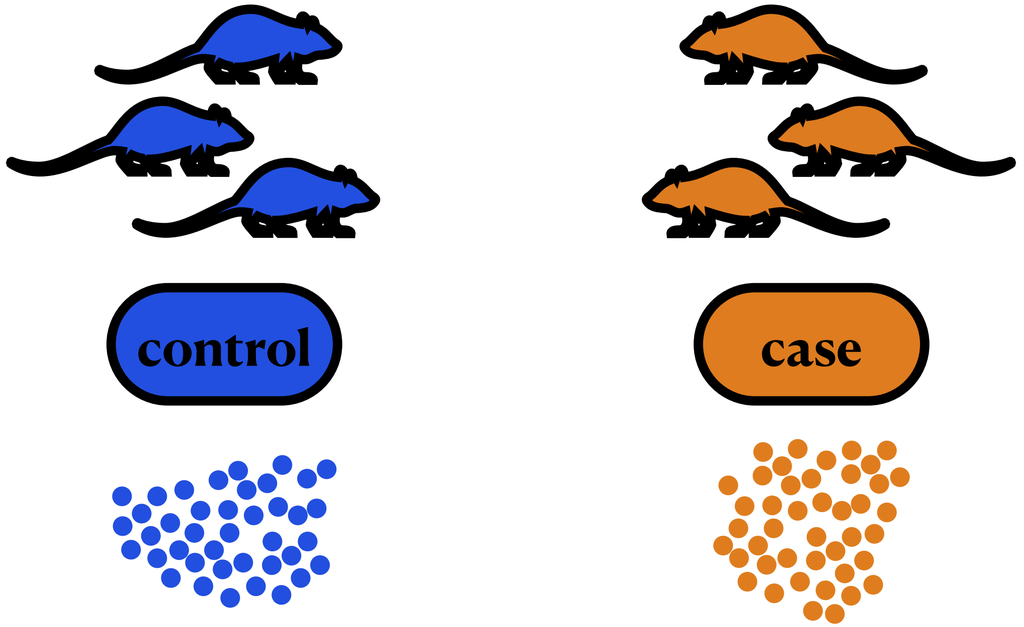


Leng, K., Li, E., Eser, R. et al. Nat Neurosci (2021)

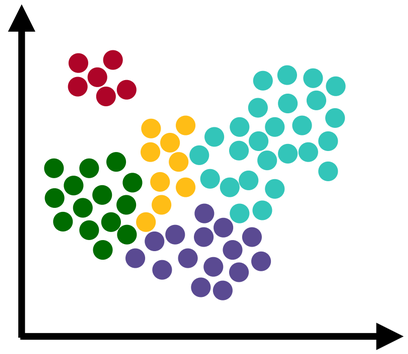
Case-control single cell experiment workflow:



Transfer the case-control sample labels onto the cells.

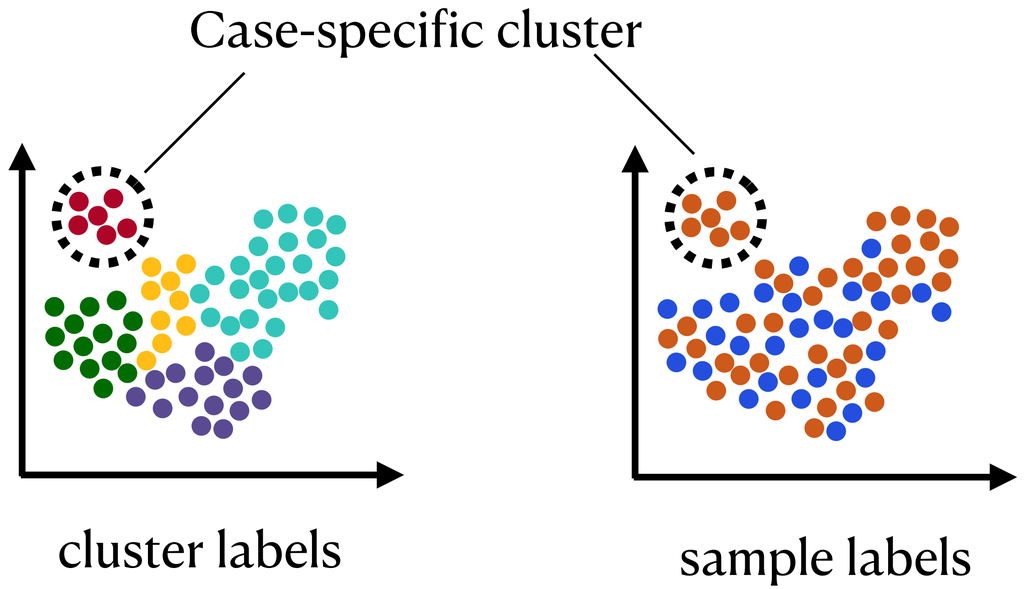


Perform standard clustering analysis.



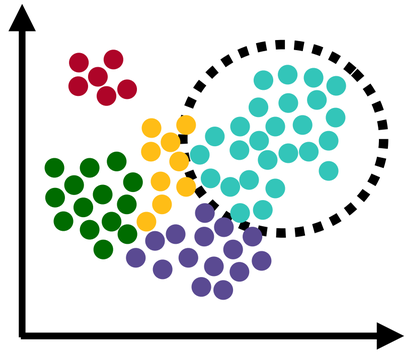
cluster labels

Sometimes there are case-specific clusters.

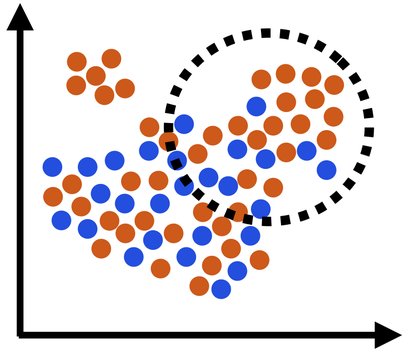


- A case-specific cluster represents a disease-specific population of cells.

For each of the other clusters we look for differences between cases and controls.



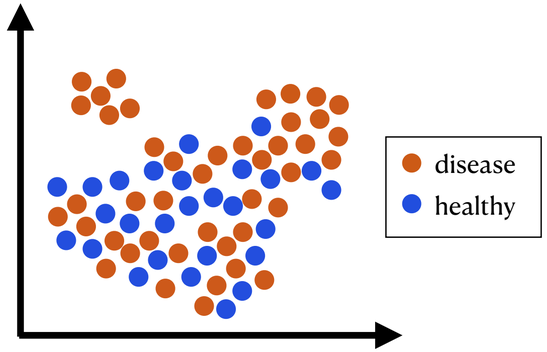
cluster labels



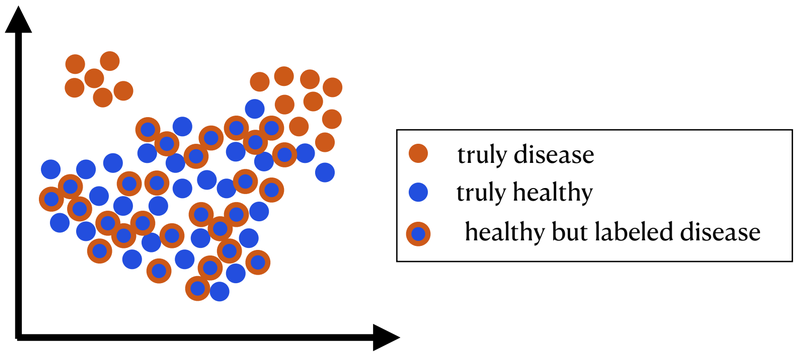
sample labels

- Perform DE using the sample labels in each cluster to identify disease-related differences.
- But for DE to work, the case-control labels have to be accurate.

All of the cells from control samples are healthy, but not all of the cells from the case samples are necessarily affected by the disease.

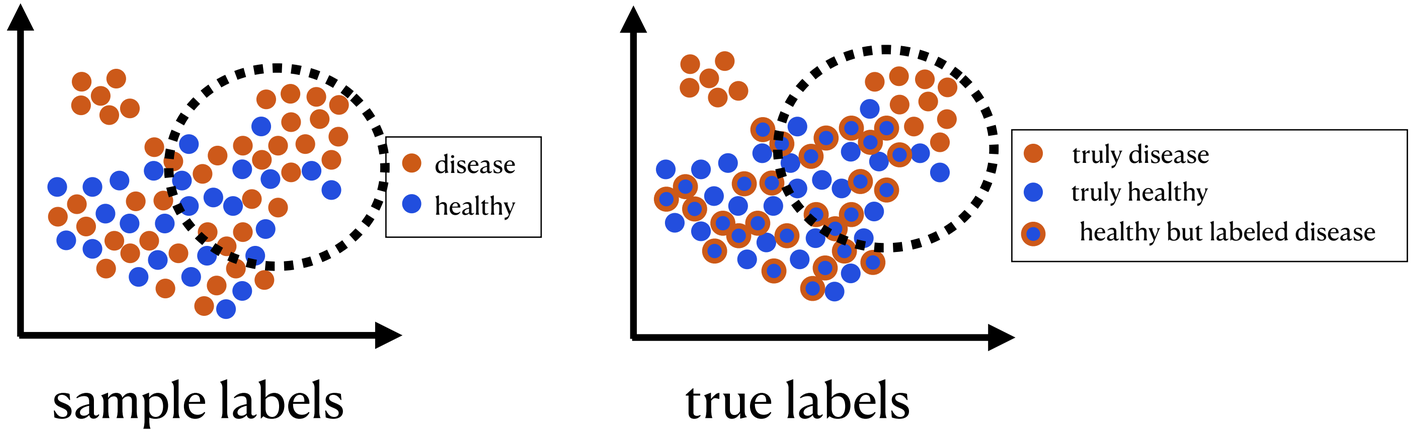


sample labels



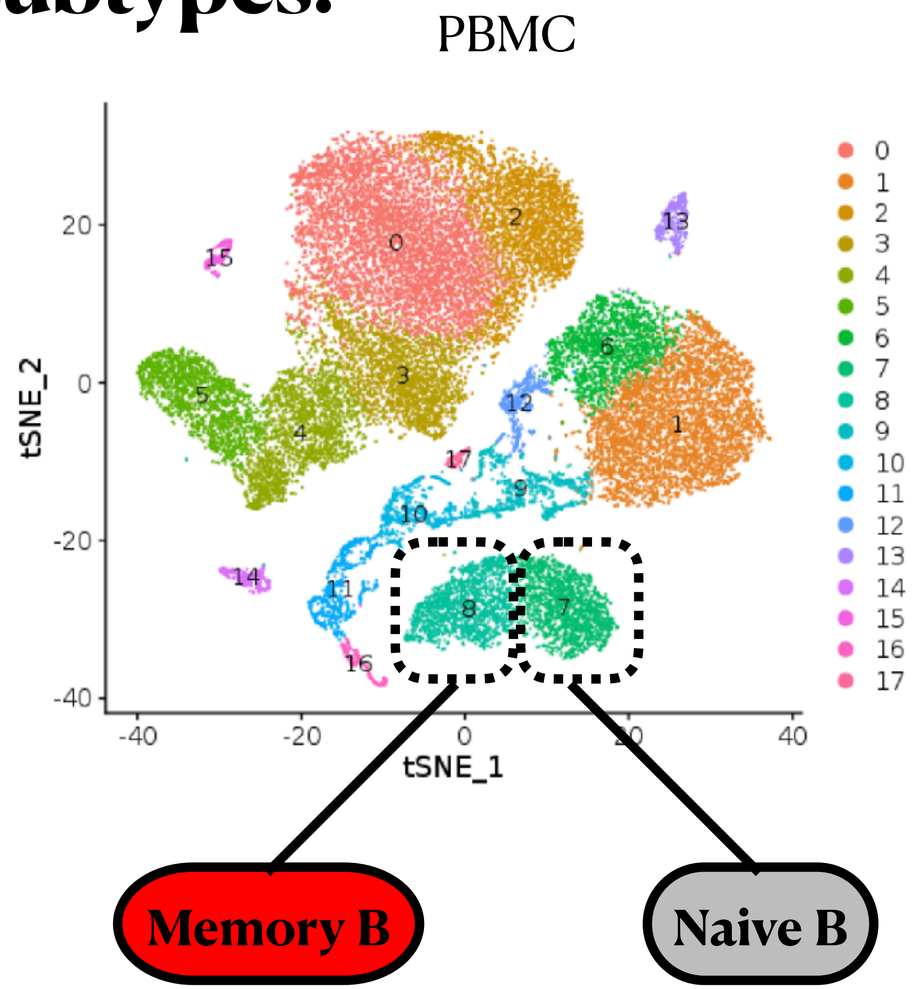
true labels

All of the cells from control samples are healthy, but not all of the cells from the case samples are necessarily affected by the disease.



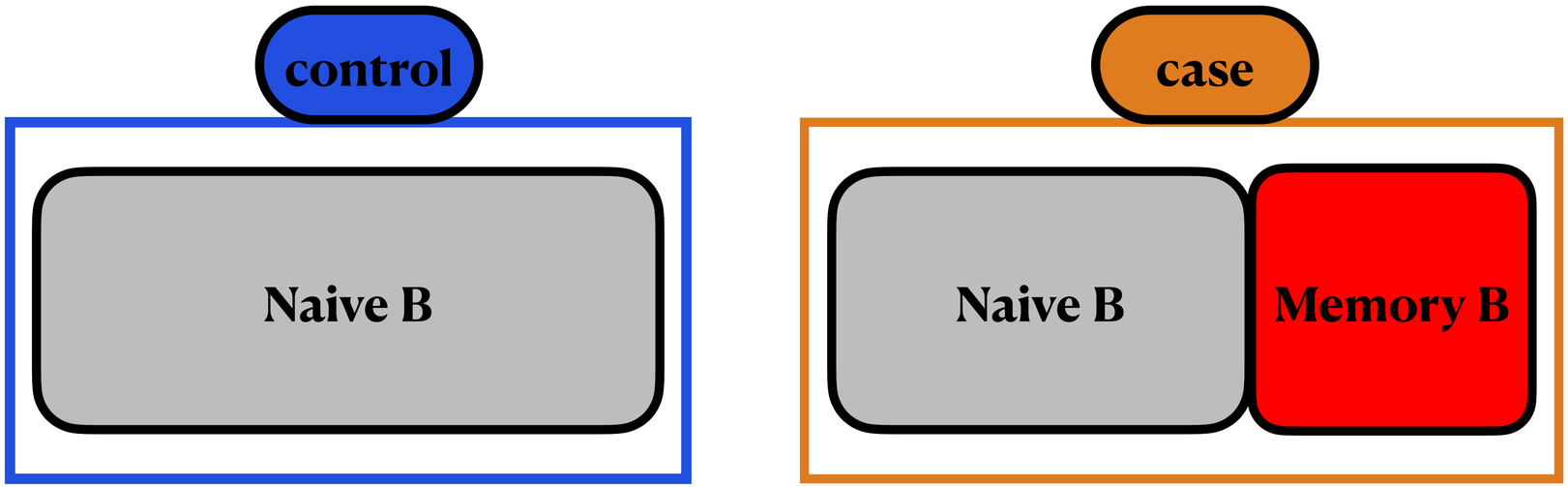
DE might fail to detect the disease signature because the signal is diluted.

Constructing ground truth by combining two cell subtypes.

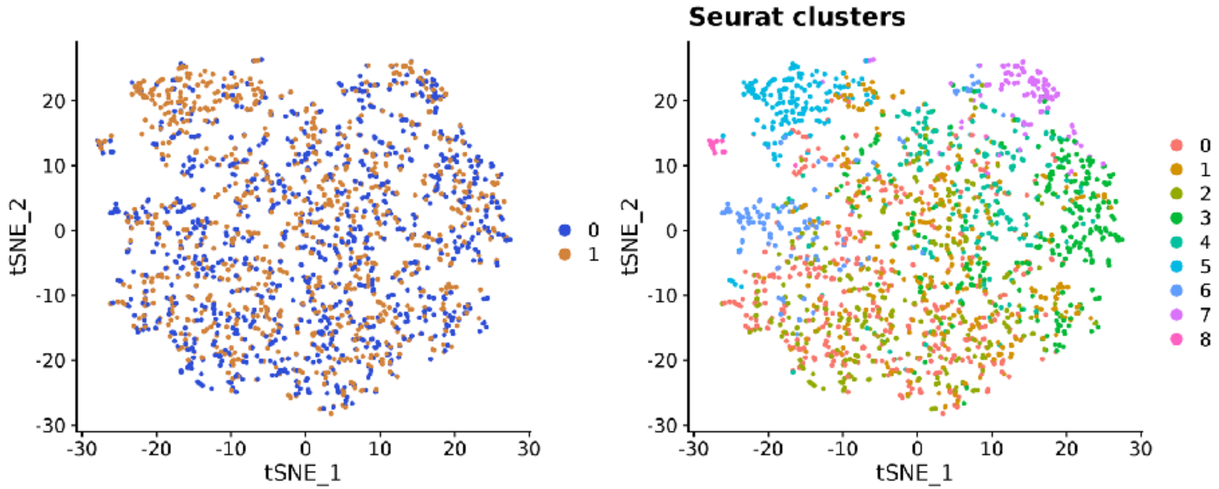


Stuart*, Butler* et al, Cell (2019)

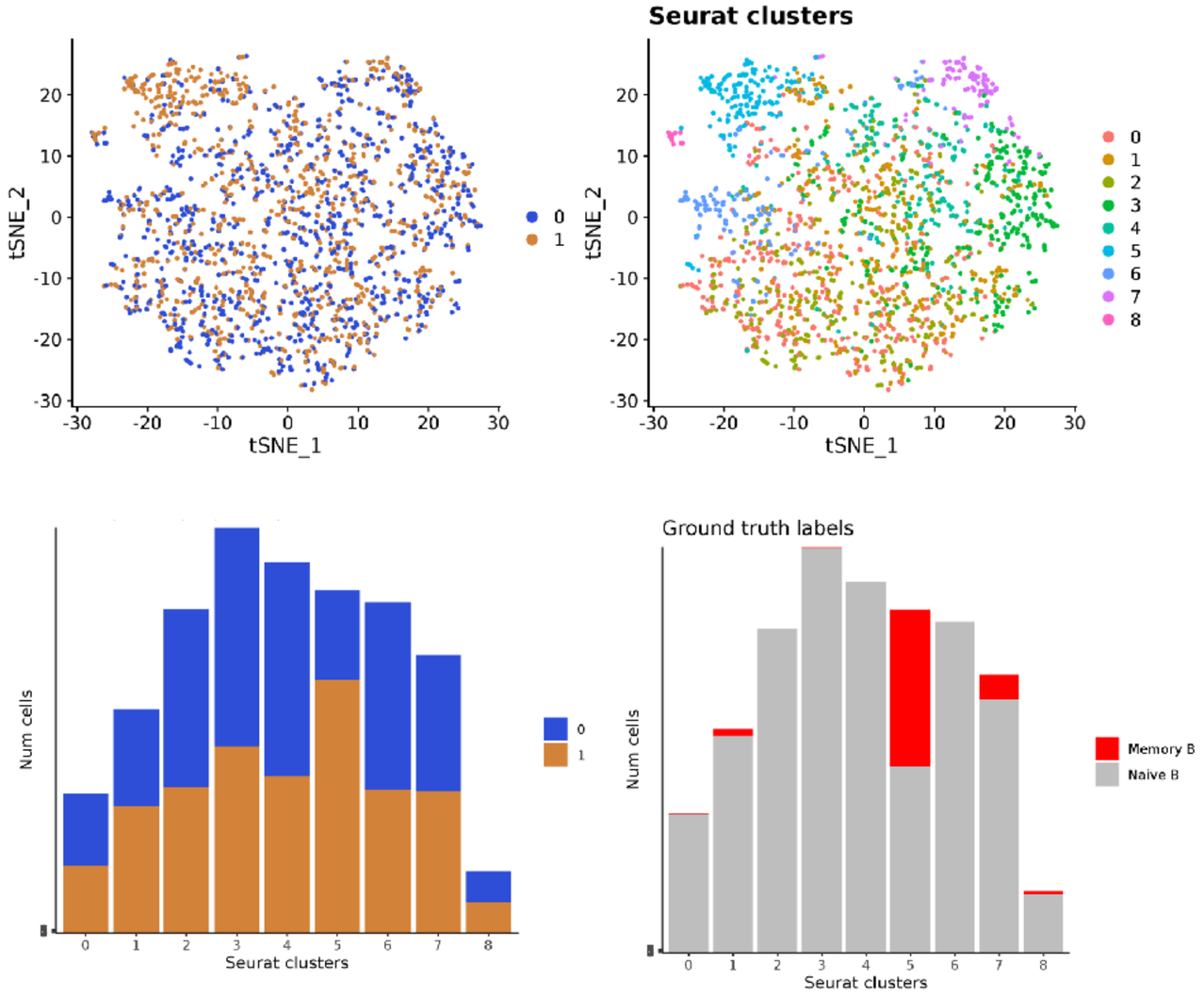
The control consists of one cell type, while the case is a mixture.



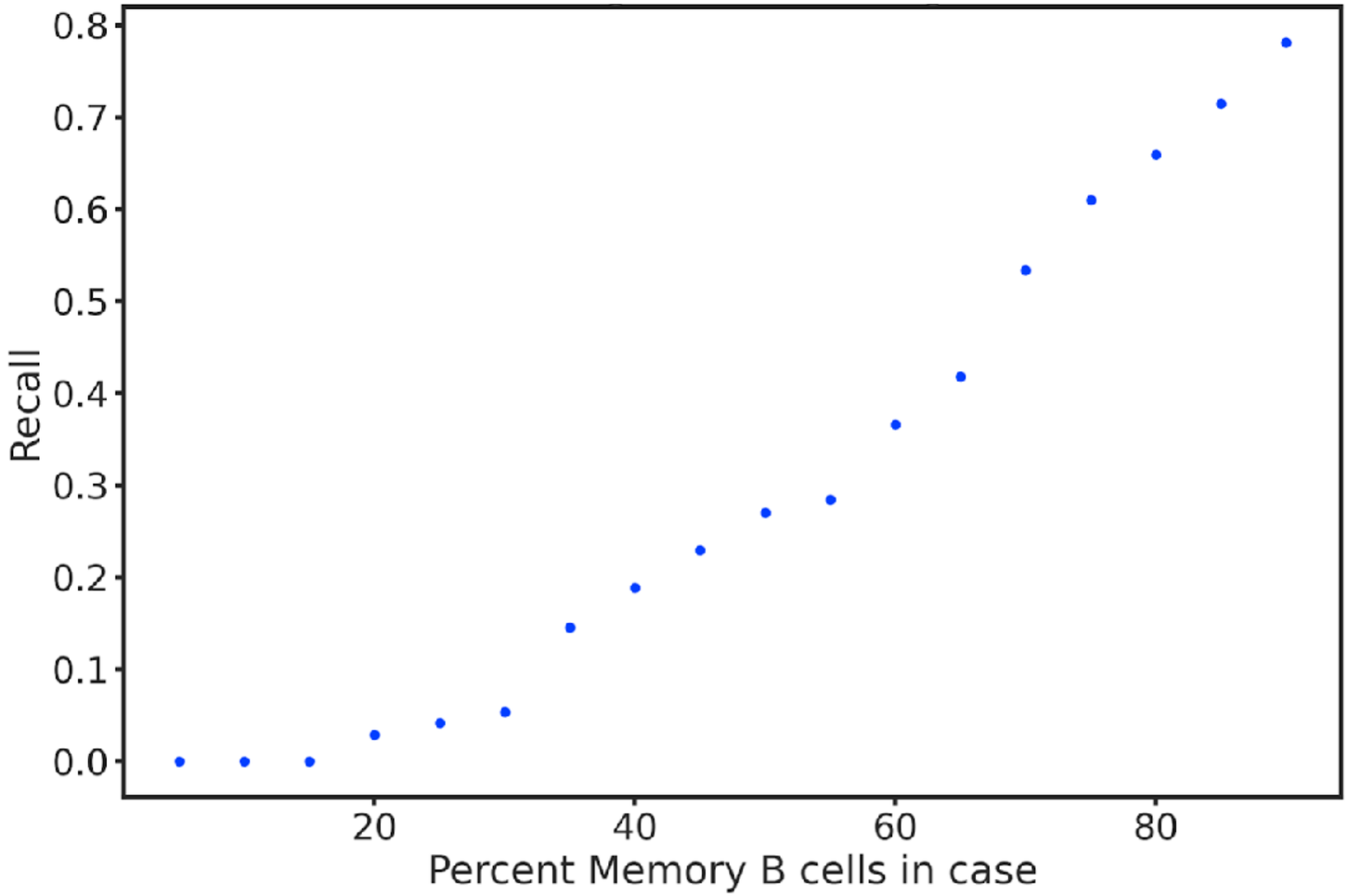
Standard analysis pipeline fails to separate out the Memory B cells.



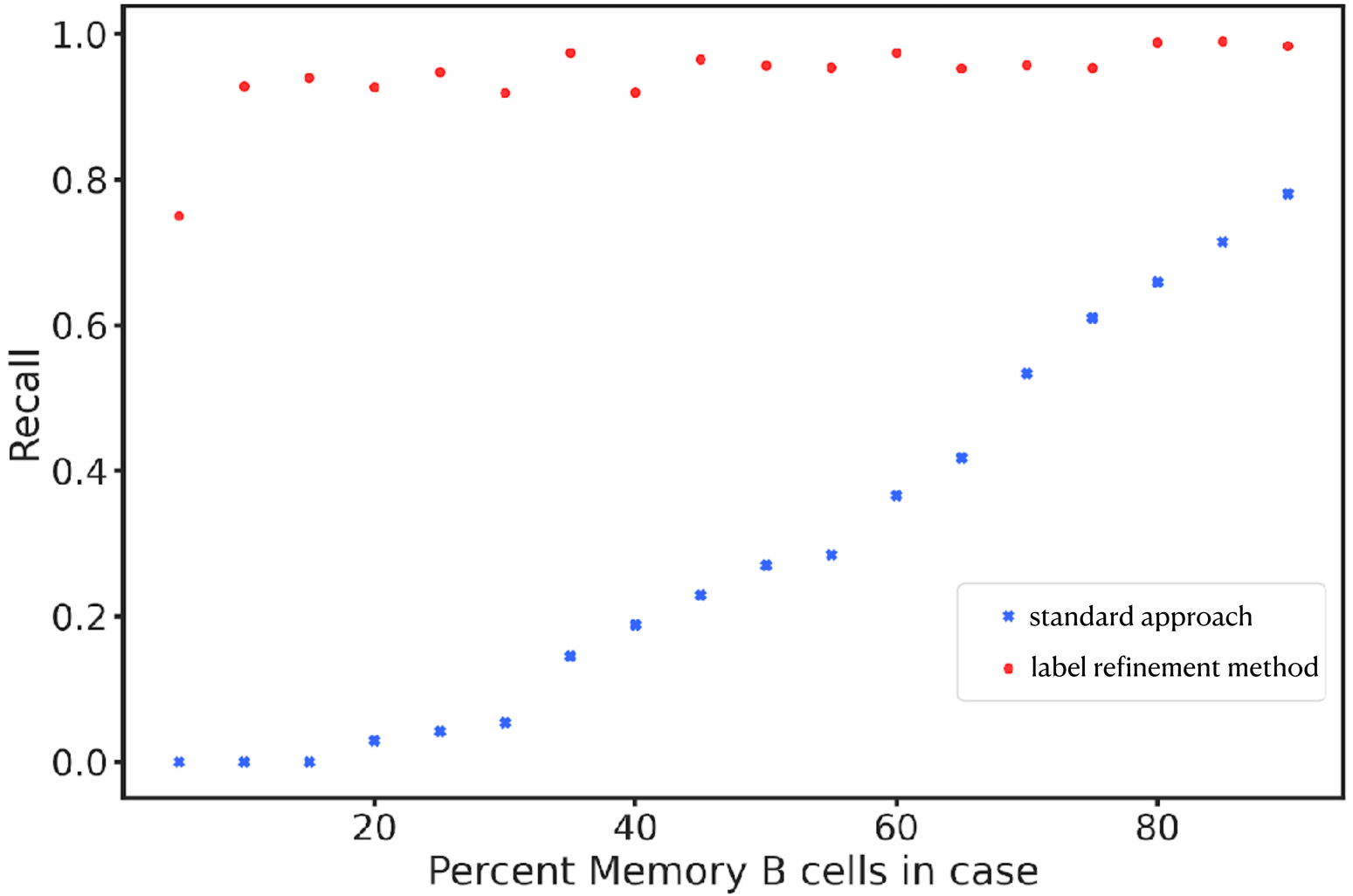
Standard analysis pipeline fails to separate out the Memory B cells.



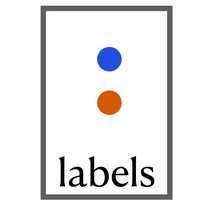
DE testing with the sample labels fails to detect the true DE genes, especially when the proportion of affected cells is small.

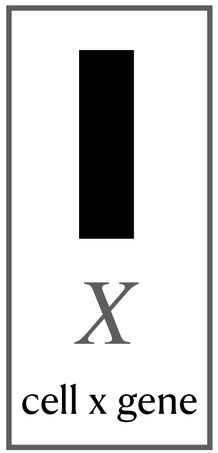


We have developed a method that improves the recall of ground truth DE genes.

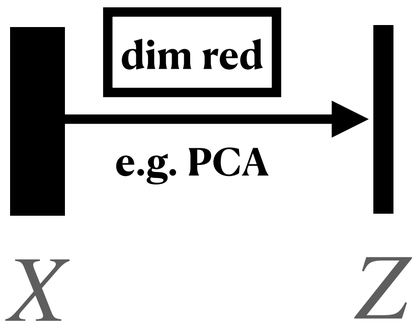


Our method combines partially correct sample labels with gene expression profiles to refine the labels.

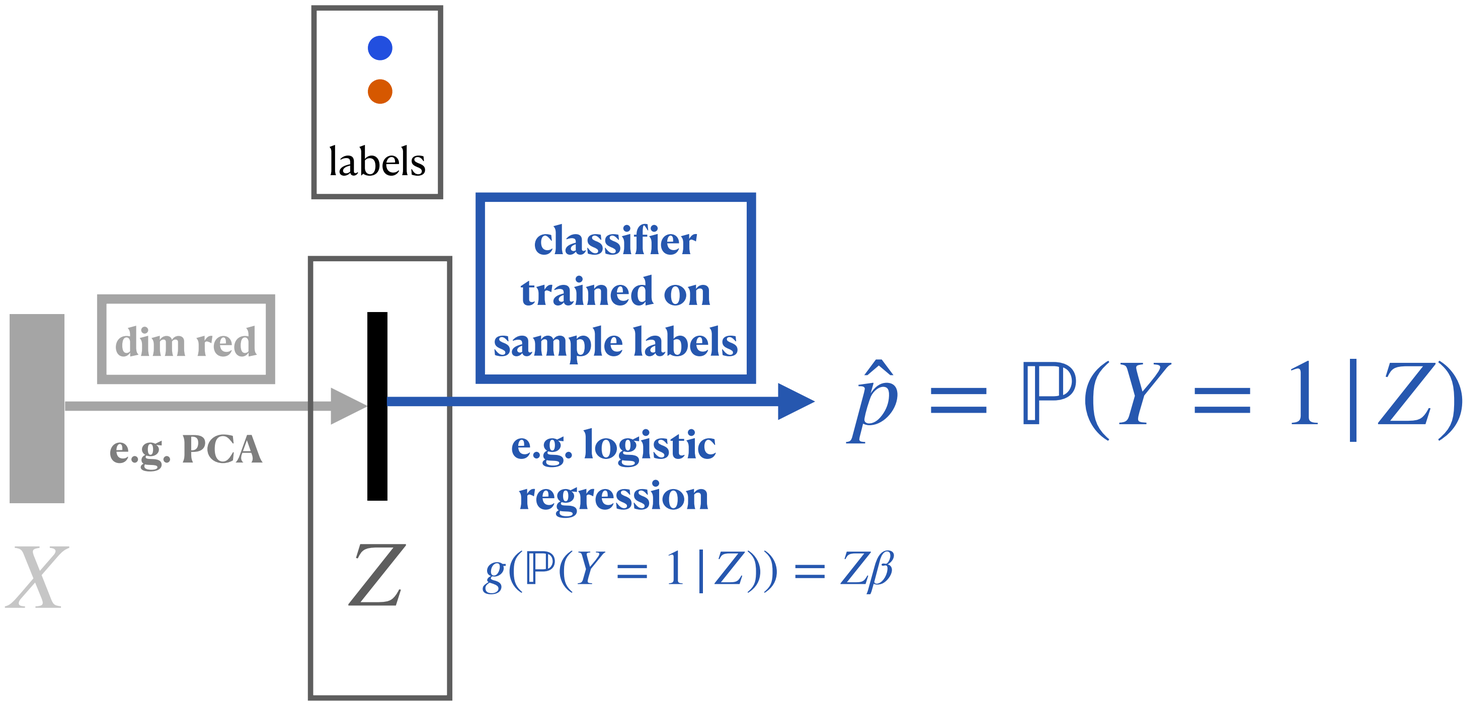




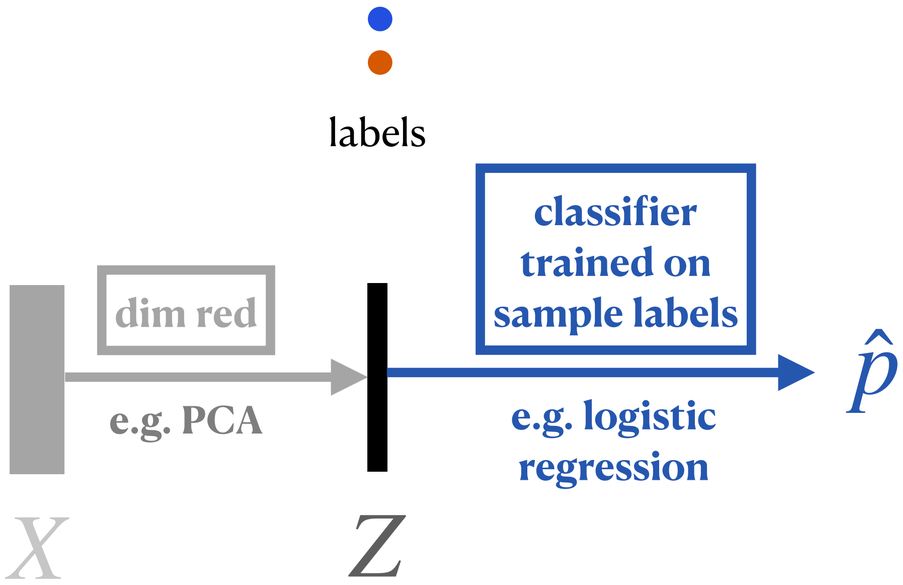
Perform dimensionality reduction.

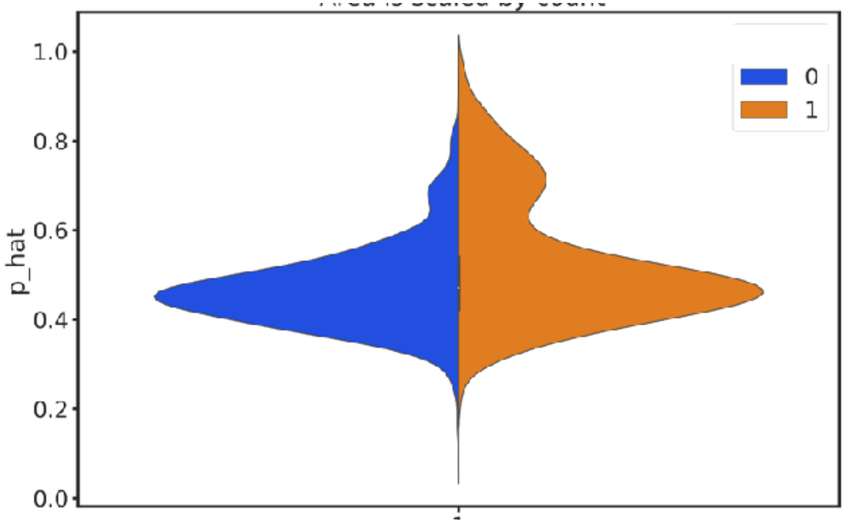


Train a classifier on the sample labels.

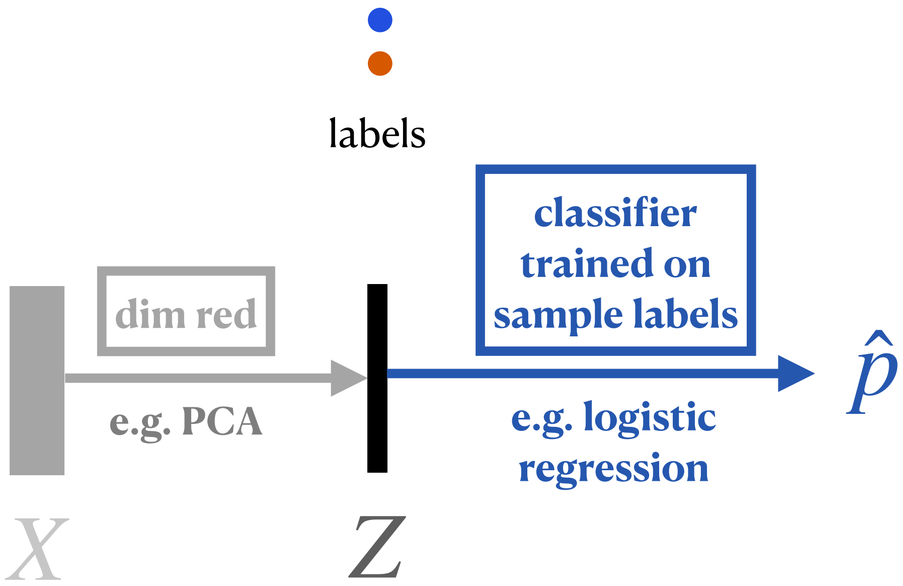


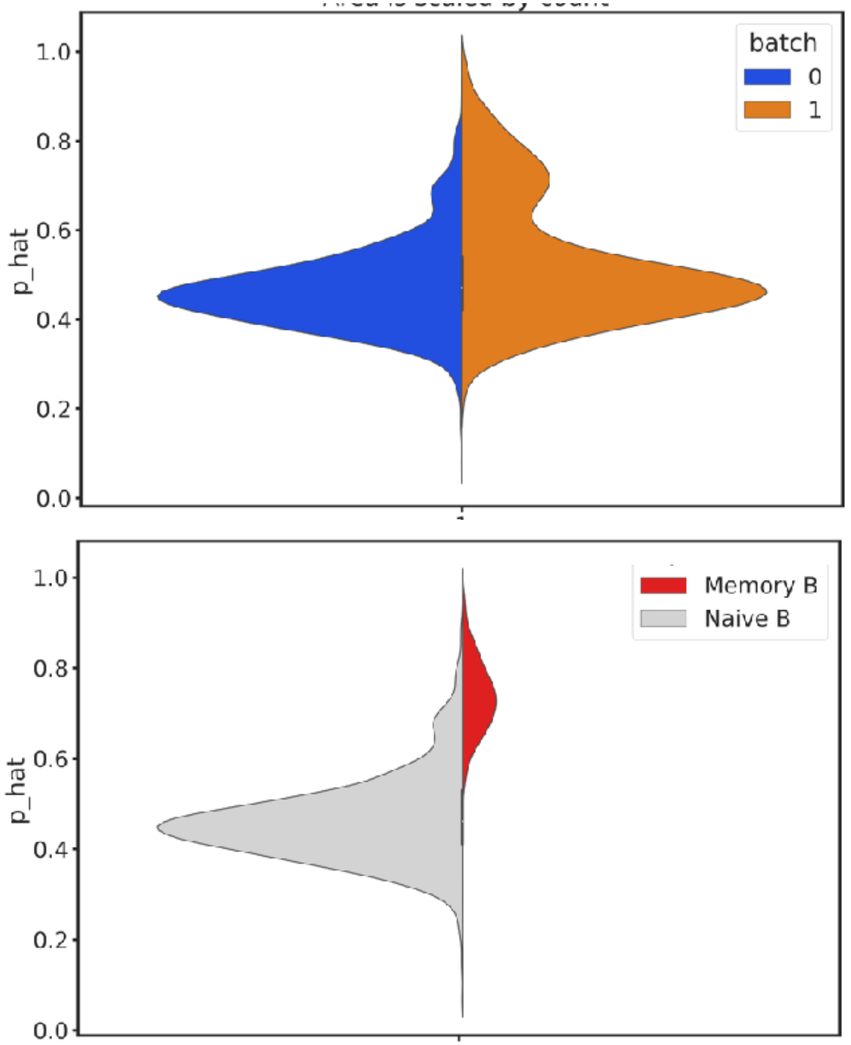
The predictions from a logistic regression trained on the sample labels well approximate the true labels.



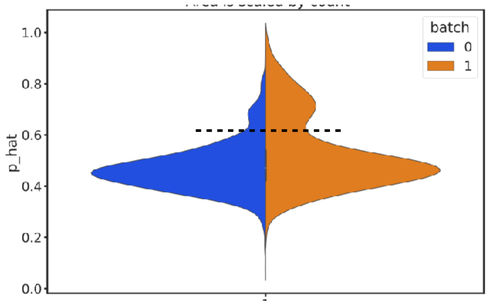


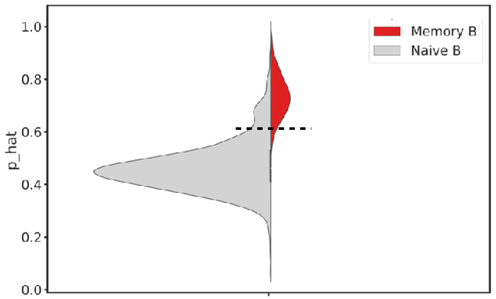
The predictions from a logistic regression trained on the sample labels well approximate the true labels.

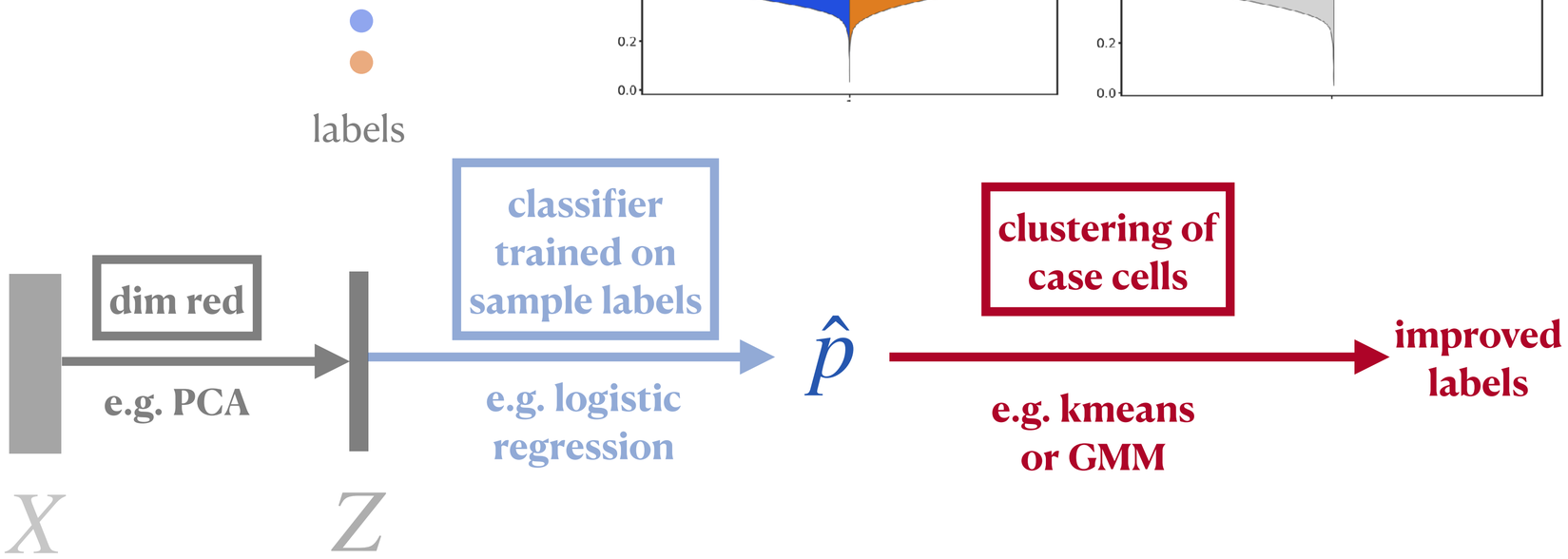




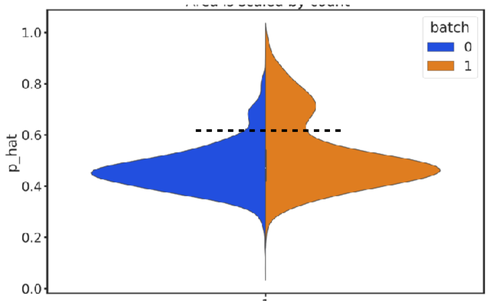
Cluster the predicted probability of being affected to refine the labels.

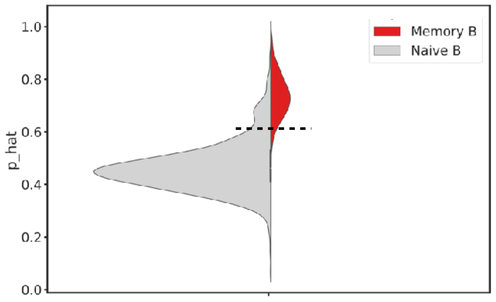


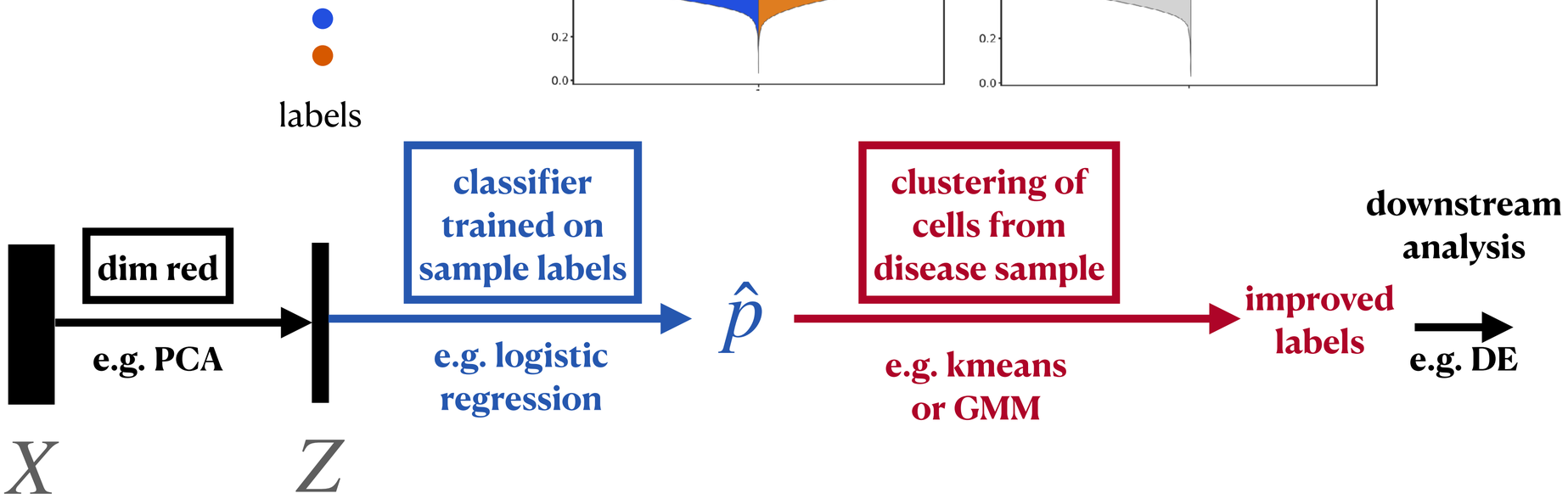




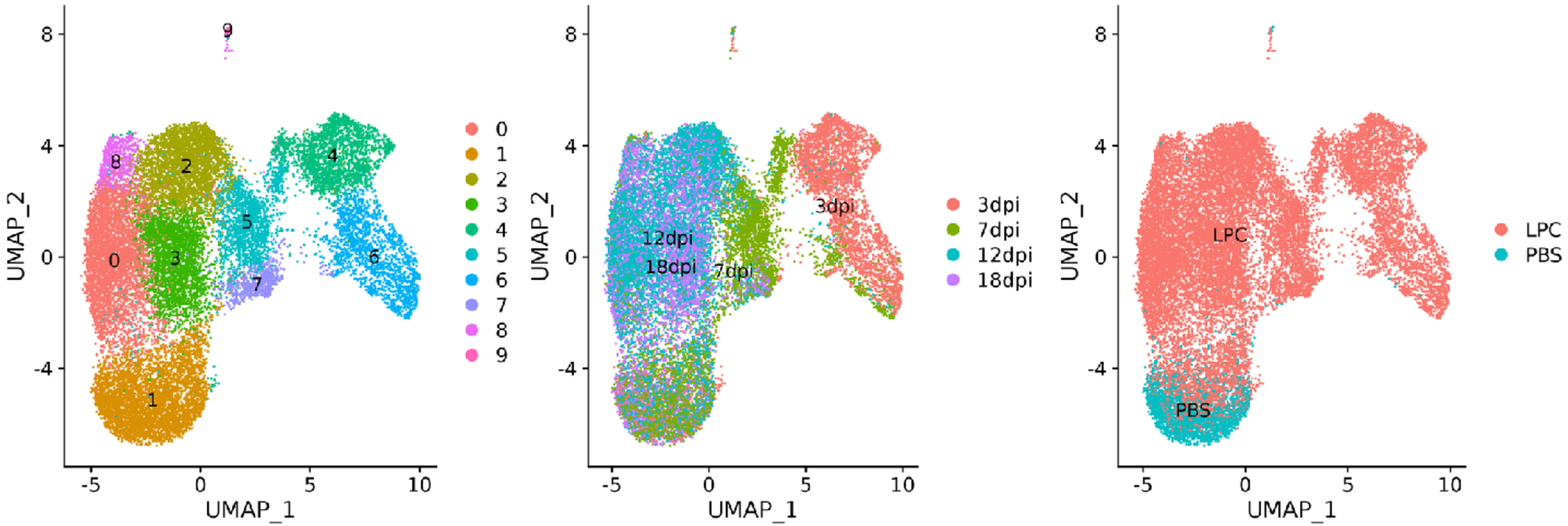
The refined labels well approximate the true labels and can be used in downstream analysis.



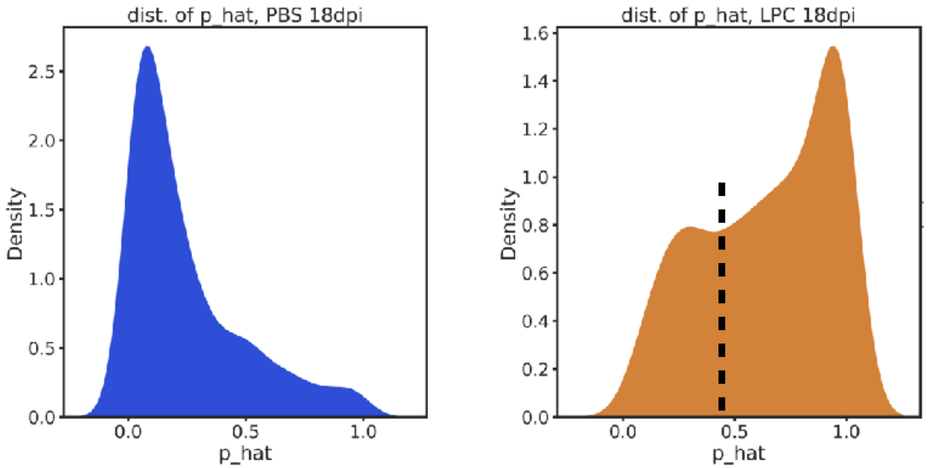


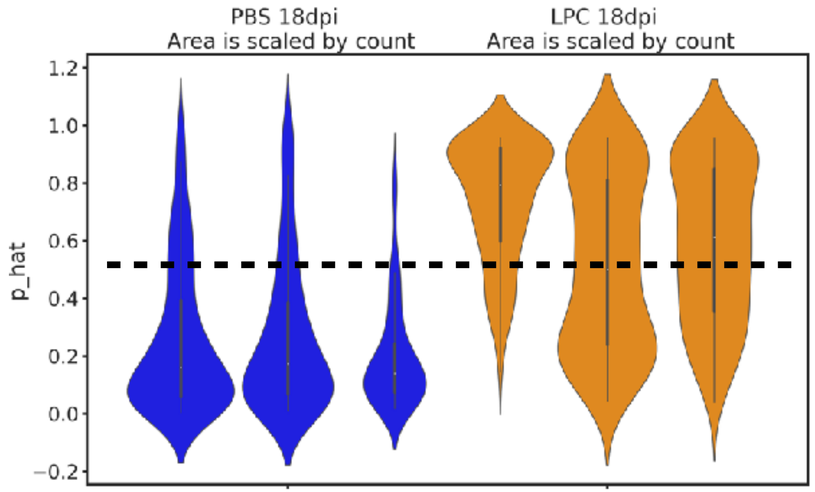


Applying our method to microglia from a case-control demyelination experiment.

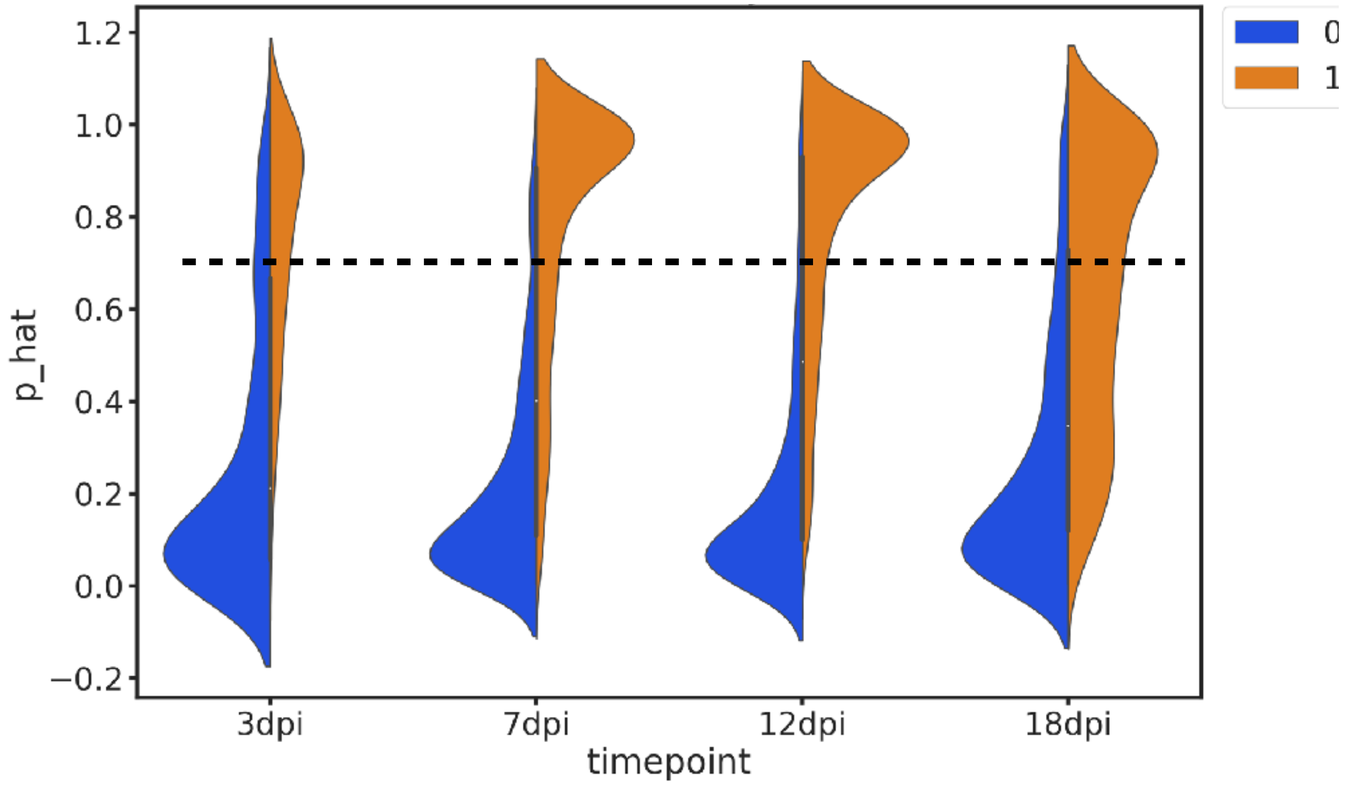


Our method splits the microglia from demyelination condition into activated and unactivated.



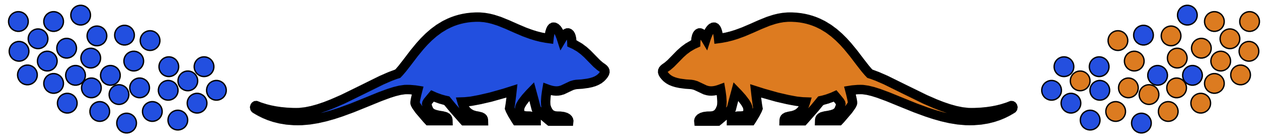


Our method reveals a changing proportion of activated to unactivated homeostatic microglia over the course of demyelination and repair.



A tale of two mice:

- This method generates hypotheses about **which the truly affected cells are** in a case-control experiment.
- The **refined labels improve downstream tasks**, such as DE.
- Applications **beyond the health-deasese context**: e.g. sex differences, perturbation experiments.
- We also output **soft labels** and can capture a **gradient** of the effect of the condition.



Acknowledgements

> Cell. 2018 Aug 9;174(4):1015-1030.e16. doi: 10.1016/j.cell.2018.07.028.

Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain

Arpiar Saunders [1], Evan Z Macosko [2], Alec Wysoker [3], Melissa Goldman [3], Fenna M Krienen [3], Heather de Rivera [3], Elizabeth Bien [3], Matthew Baum [3], Laura Bortolin [3], Shuyu Wang [4], Aleksandrina Goeva [4], James Nemesh [3], Nolan Kamitaki [3], Sara Brumbaugh [3], David Kulp [3], Steven A McCarroll [5]

Emergence of Division of Labor in Tissues through Cell Interactions and Spatial Cues

Miri Adler[1*], Noa Moriel[2*], Aleksandrina Goeva[1*], Inbal Avraham Davidi[1], Evan Macosko[1,3], Aviv Regev[1,4,5], Ruslan Medzhitov[6] and Mor Nitzan[2,7,8]

Science

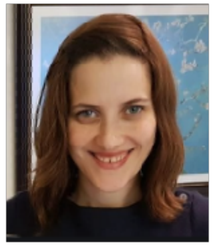
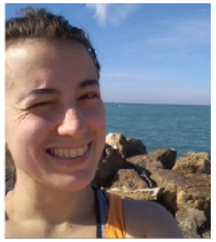
Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution

SAMUEL G. RODRIQUES, ROBERT R. STICKELS, ALEKSANDRINA GOEVA, CARLY A. MARTIN, EVAN MURRAY, CHARLES R. VANDERBURG, JOSHUA WELCH, LINLIN M. CHEN, FEI CHEN, AND EVAN Z. MACOSKO

Improved marker detection through label refinement in case-control single-cell RNA-seq studies

Aleksandrina Goeva[1], Michael-John Dolan[1], and Evan Macosko[1,2]




BROADIGNITE