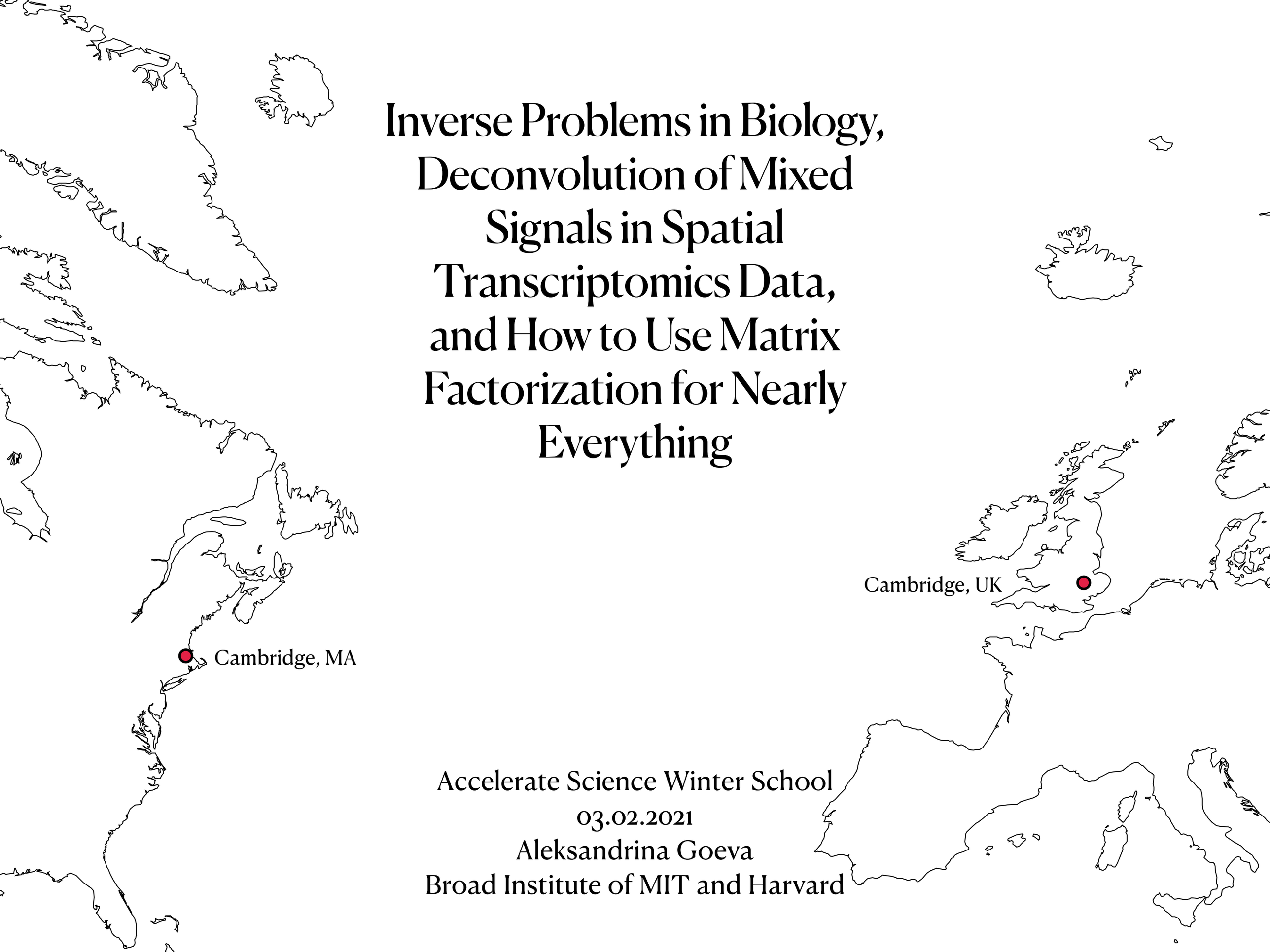
Inverse Problems in Biology, Deconvolution of Mixed Signals in Spatial Transcriptomics Data, and How to Use Matrix Factorization for Nearly Everything

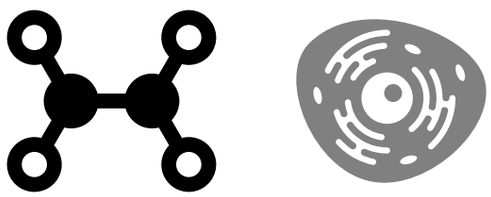


Accelerate Science Winter School
03.02.2021

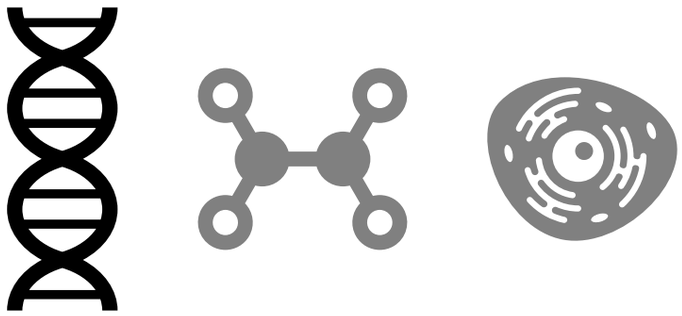
Aleksandrina Goeva
Broad Institute of MIT and Harvard

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

Cells are made out of **molecules**.



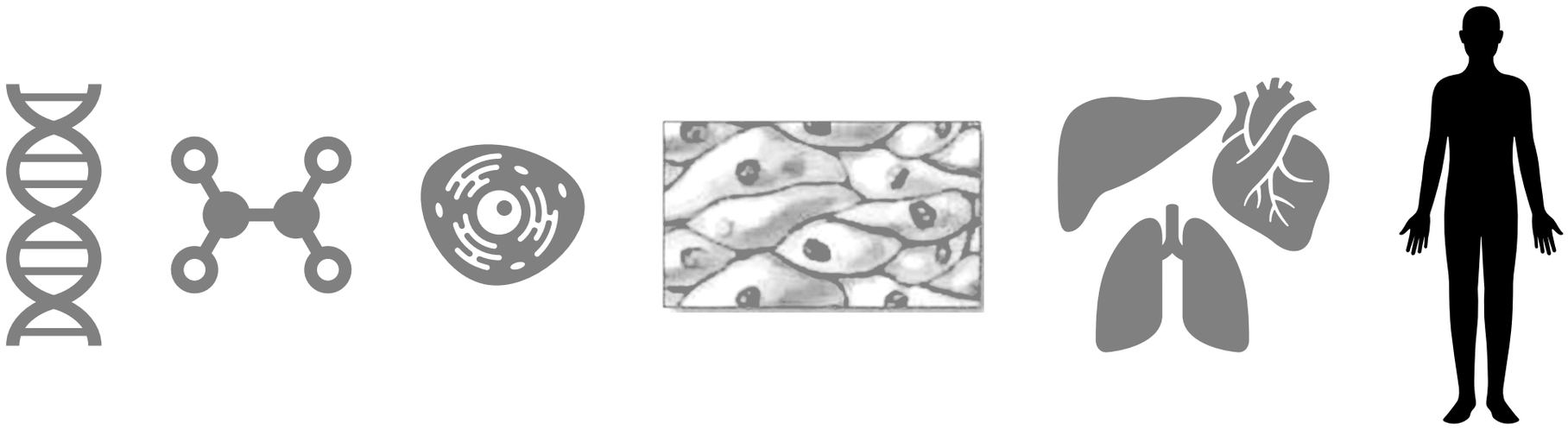
The molecules are encoded by **genes** (or made out of gene products).



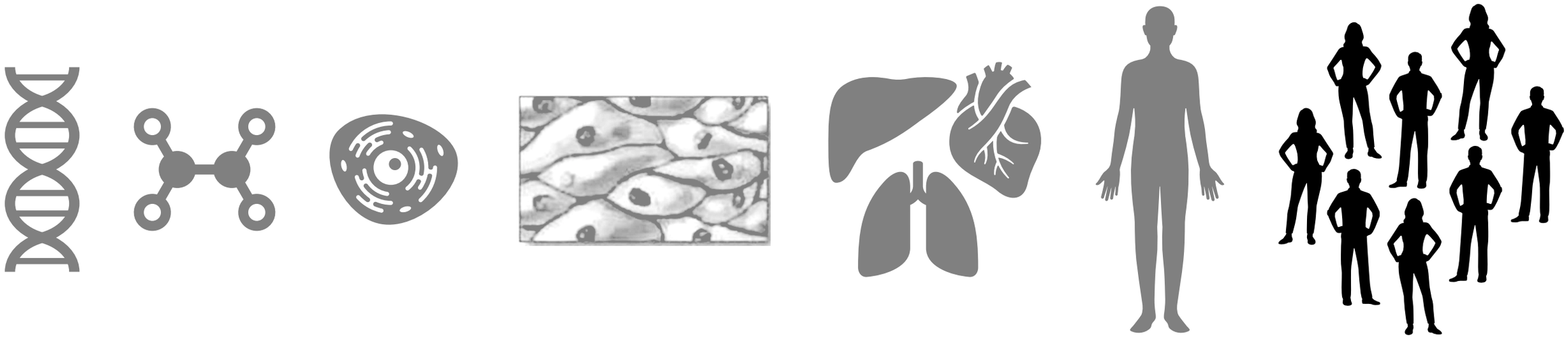
Cells interact with each other to make **tissues**.

Tissues form **organs.**

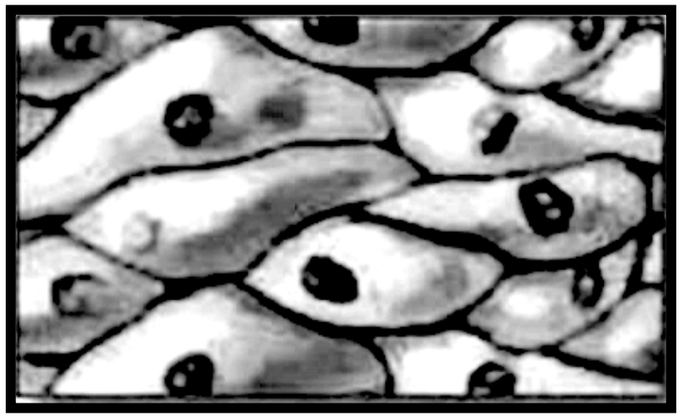
Organs together account for **organisms.**



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



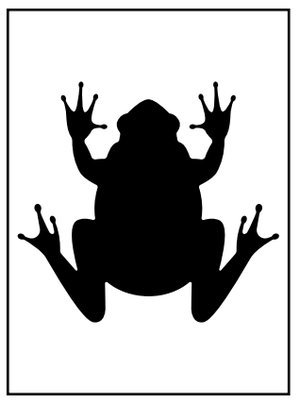
Today we will focus on **cells** and **tissues**.

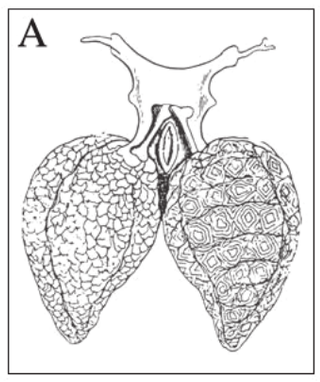


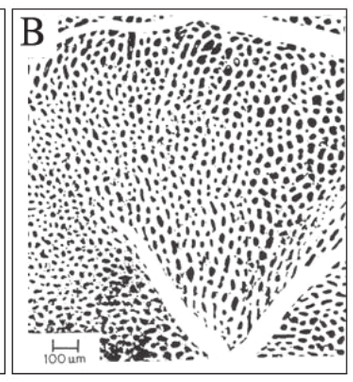
How do people collect **data** from tissues and cells?

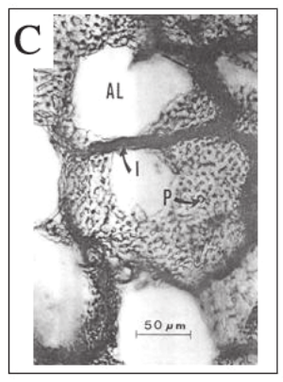
17th century

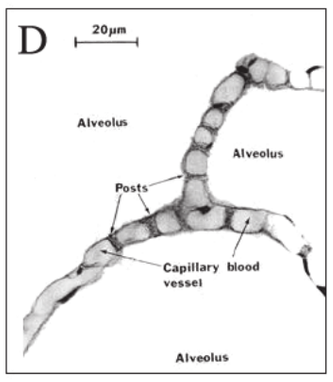
Using a **microscope** to look at parts of animals.









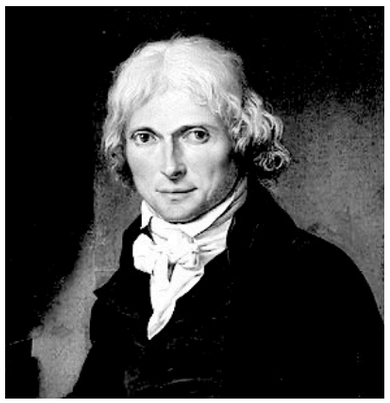


Timeline

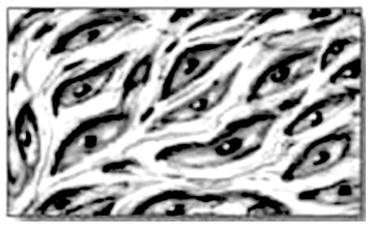
17th century

19th century

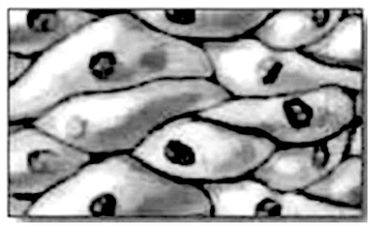
The birth of the terms **tissue** and **histology**.



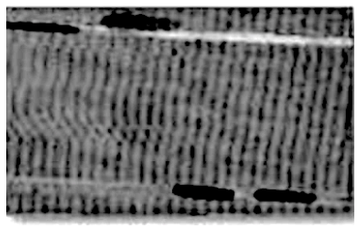
1801
21 elementary tissues



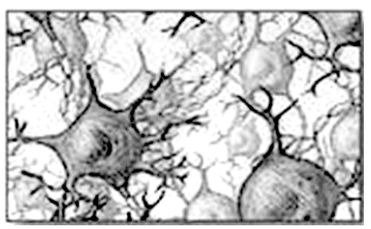
connective



epithelial



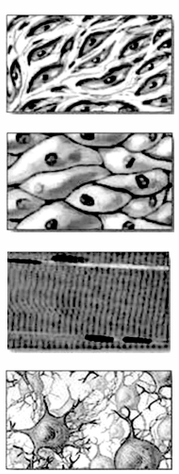
muscle



nervous

1857 - present
4 tissue types

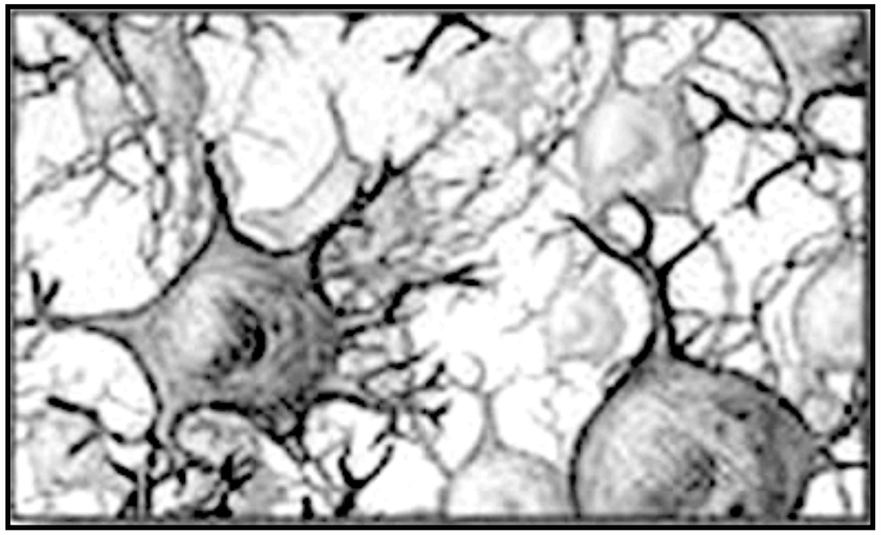
Timeline



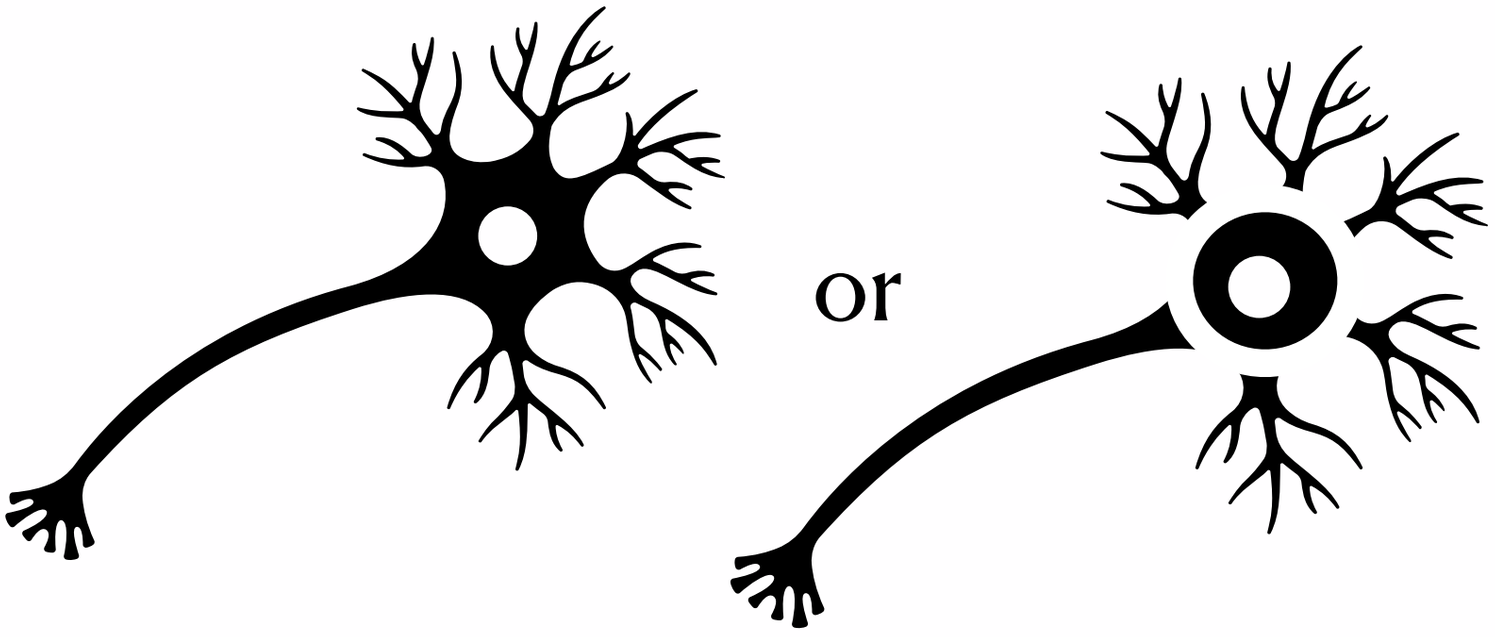
17th century

19th century

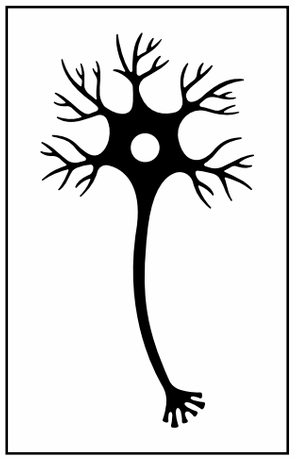
Neuron doctrine

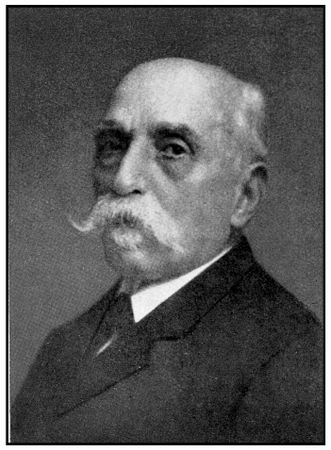


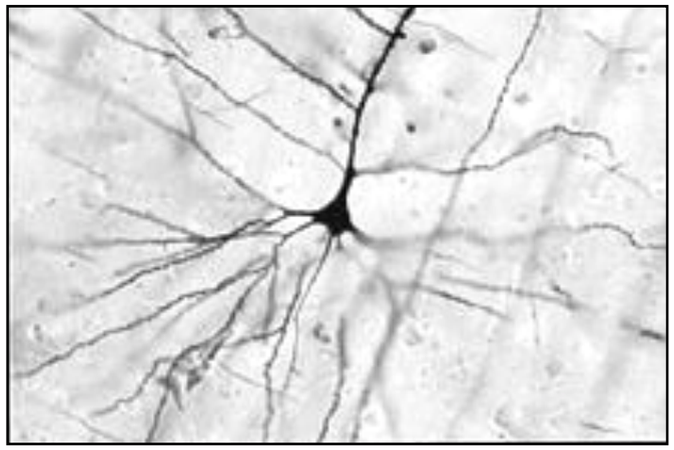
Neuron doctrine

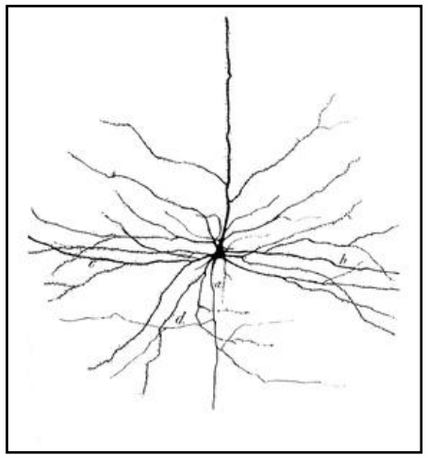


Neuron doctrine

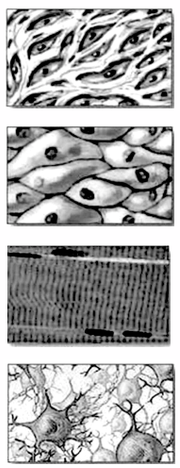


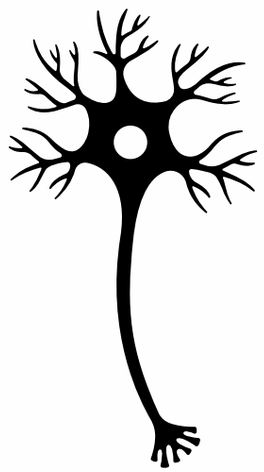






Timeline



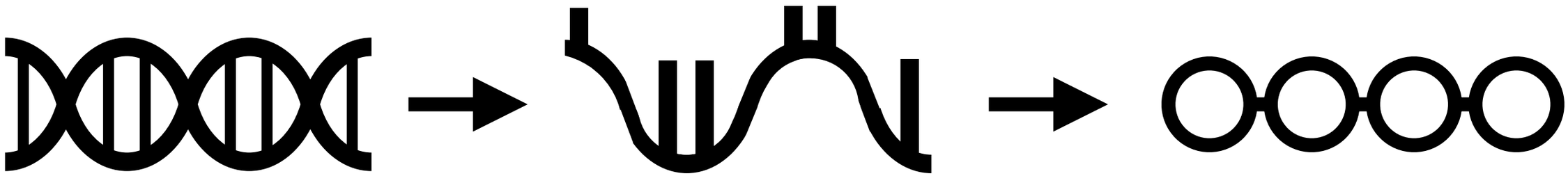


17th century

19th century

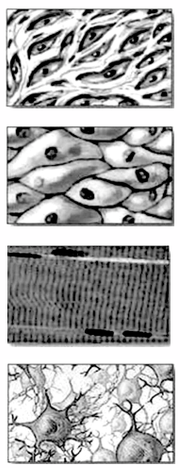
1906

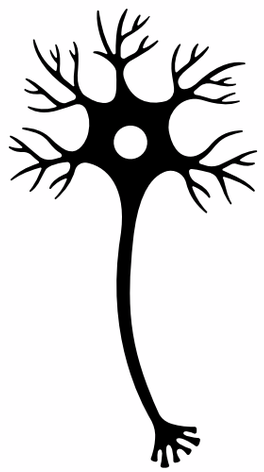
The central dogma.



DNA **mRNA** protein

Timeline



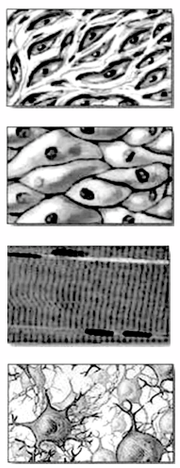


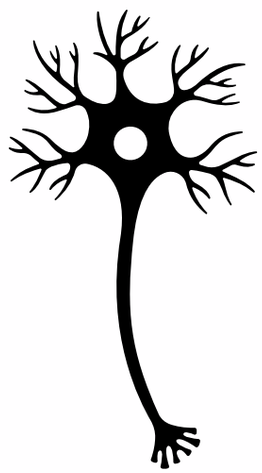
17th century

19th century

1906

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





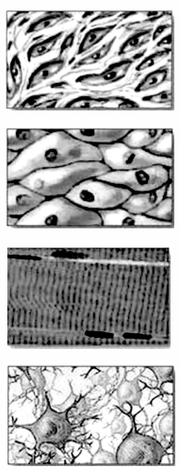
17th century

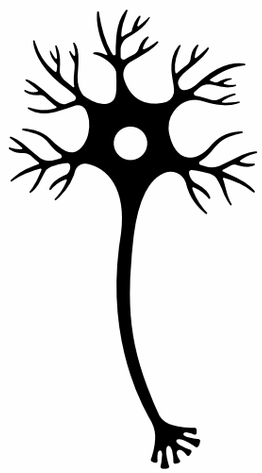
19th century

1906

1995

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





17th century

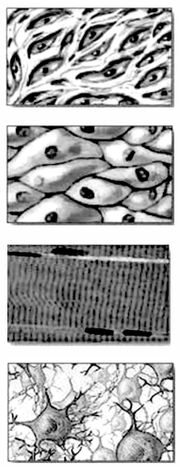
19th century

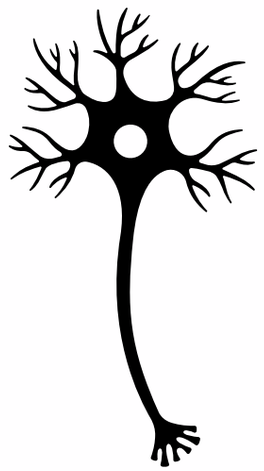
1906

1995

2006

Single-cell RNAseq





17th century

19th century

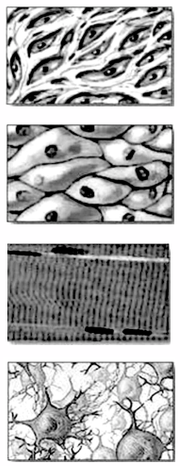
1906

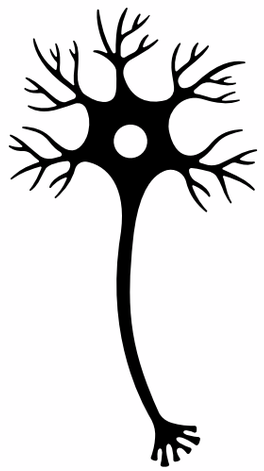
1995

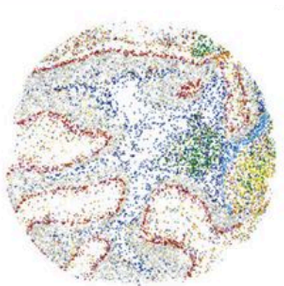
2006

2015

Spatial transcriptomics







17th century

19th century

1906

1995

2006

2015

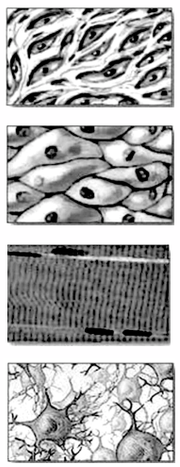
2019

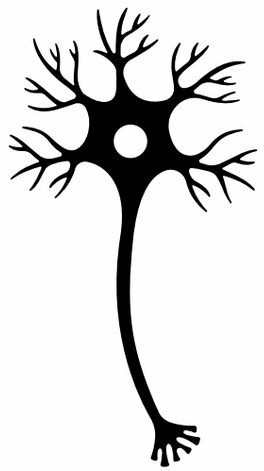
nature methods

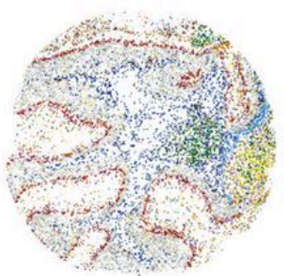
Technology Feature | Published: 06 January 2021

Method of the Year: spatially resolved transcriptomics

Timeline







17th century

19th century

1906

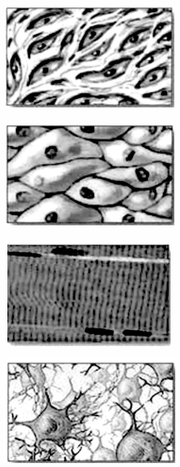
1995

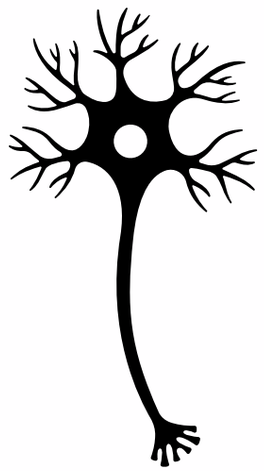
2006

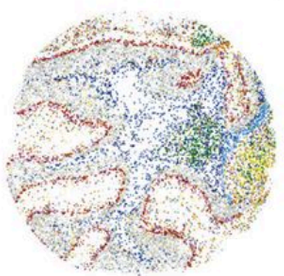
2015

2019

Timeline







17th century

19th century

1906

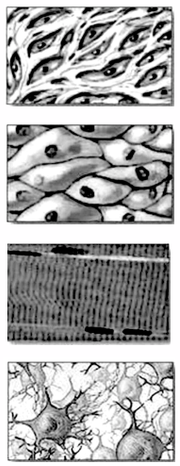
1995

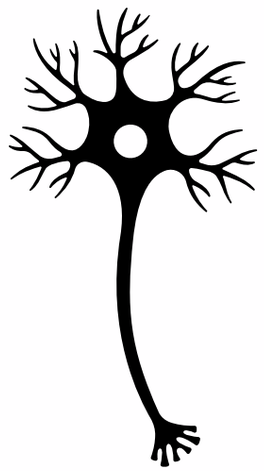
2006

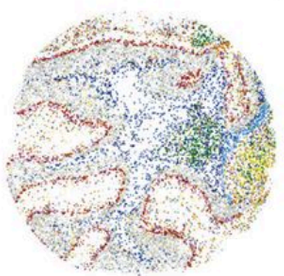
2015

2019

Timeline







17th century

19th century

1906

1995

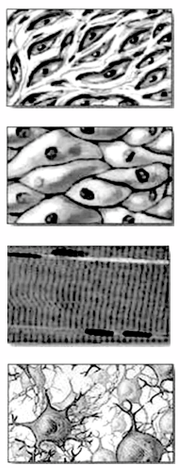
2006

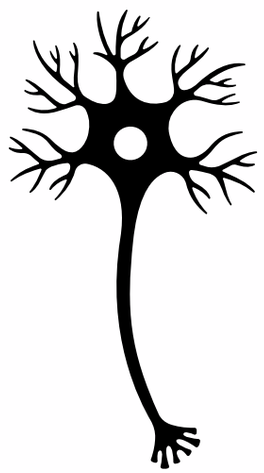
2015

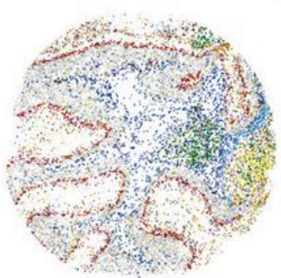
2019

Sofia, Bulgaria

Timeline







17th century

19th century

1906

1995

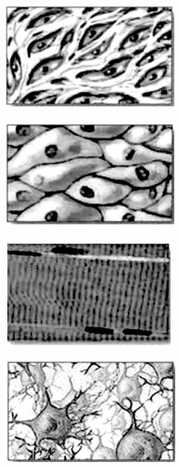
2006

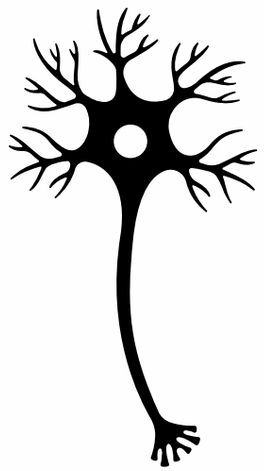
2015

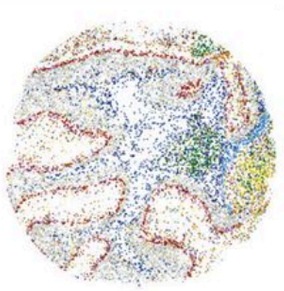
2019

Sofia, Bulgaria

Timeline







17th century

19th century

1906

1995

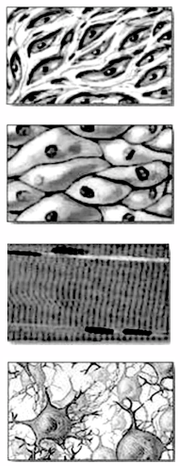
2006

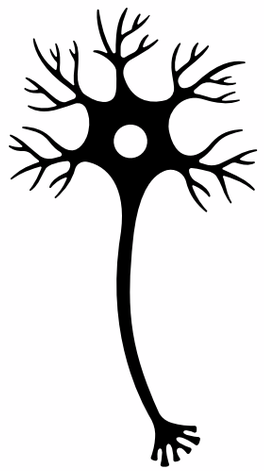
2015

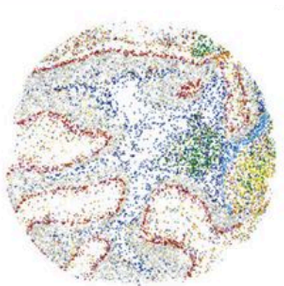
2019

Sofia, Bulgaria

Timeline







17th century

19th century

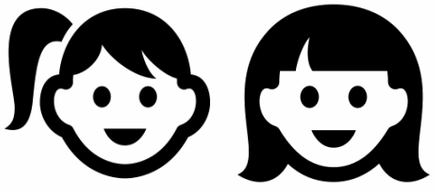
1906

1995

2006

2015

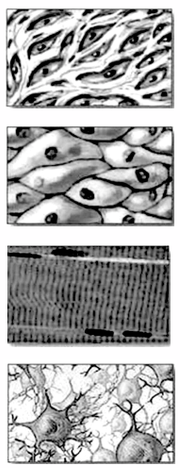
2019

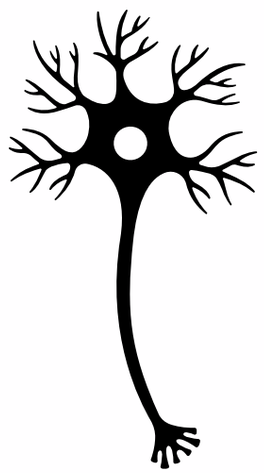


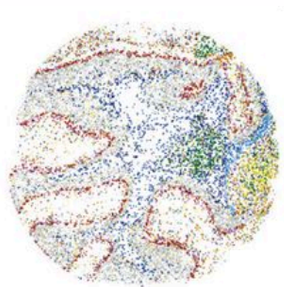
Sofia, Bulgaria

Boston, MA

Timeline







17th century

19th century

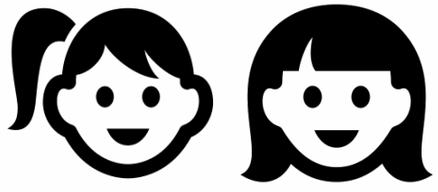
1906

1995

2006

2015

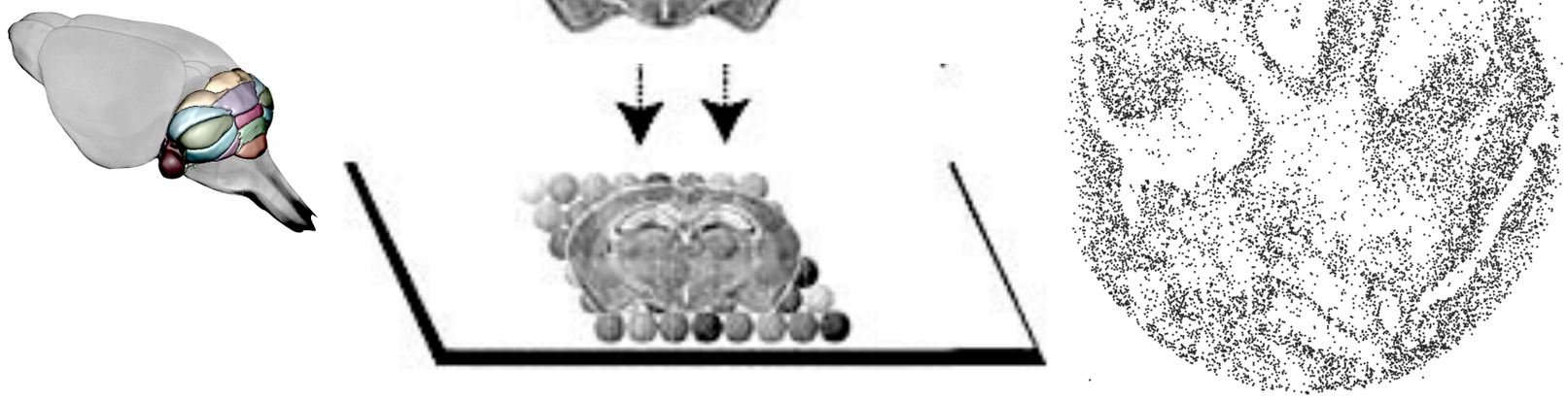
2019

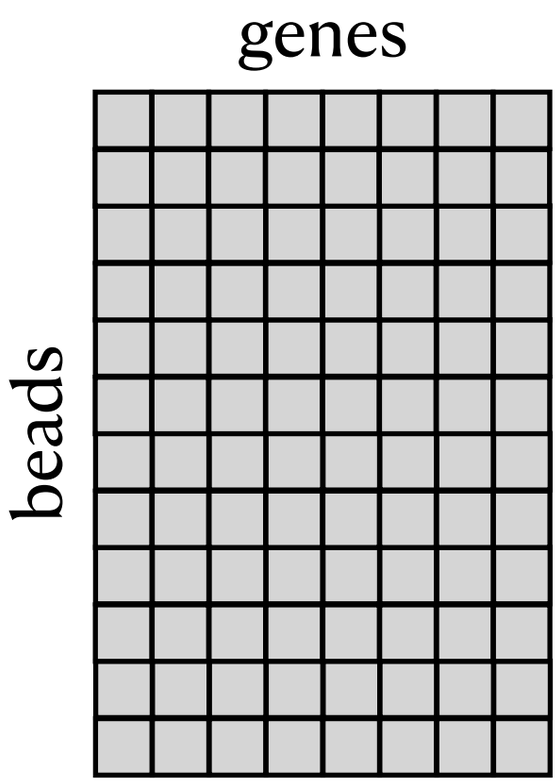


Sofia, Bulgaria

Boston, MA

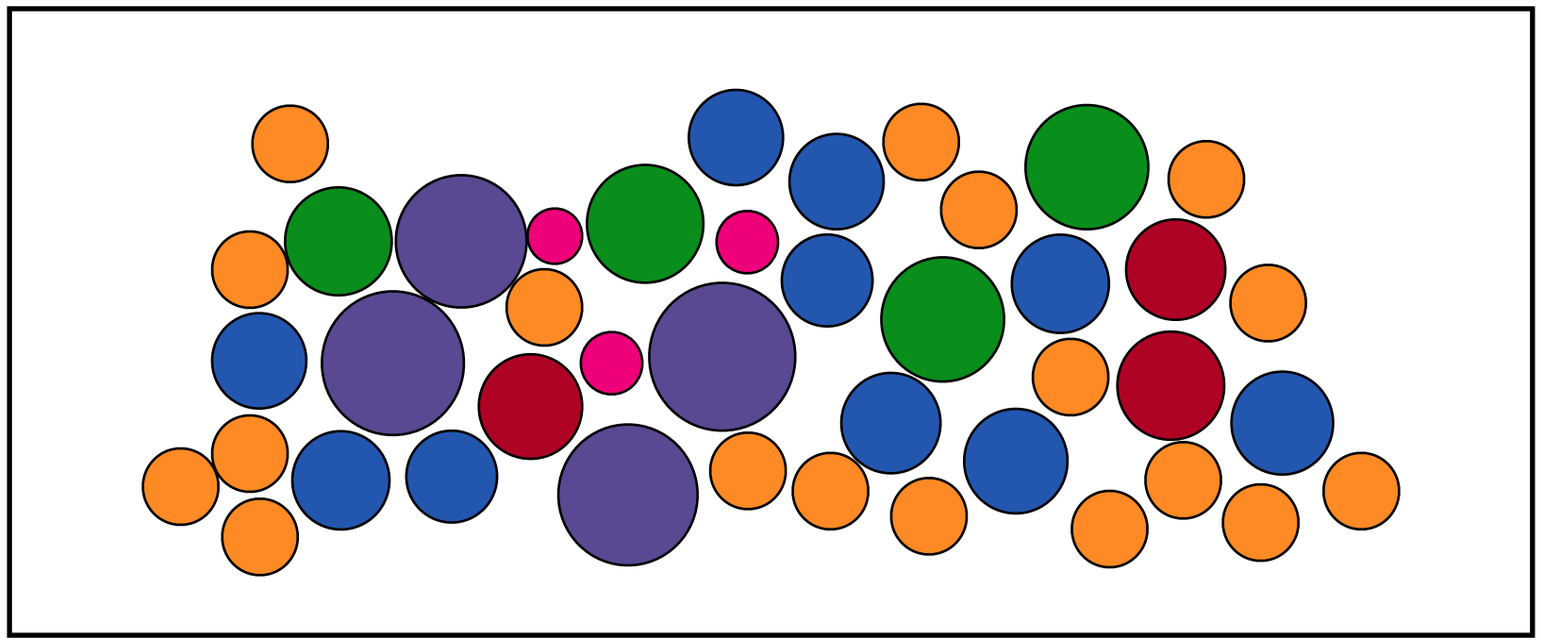
Spatial Transcriptomics Data





matrix of counts
of mRNA abundance
at each spatial probe

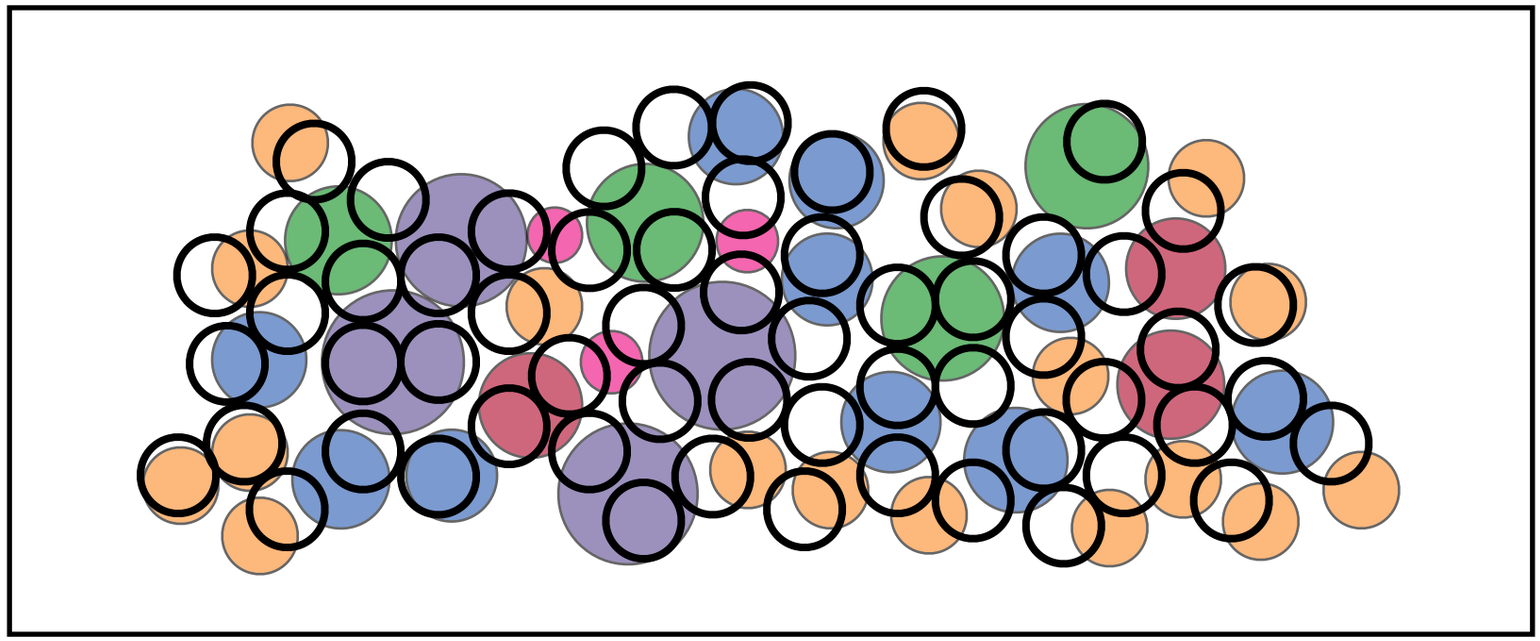
The tissue contains **multiple cell types.**



Cells in the tissue

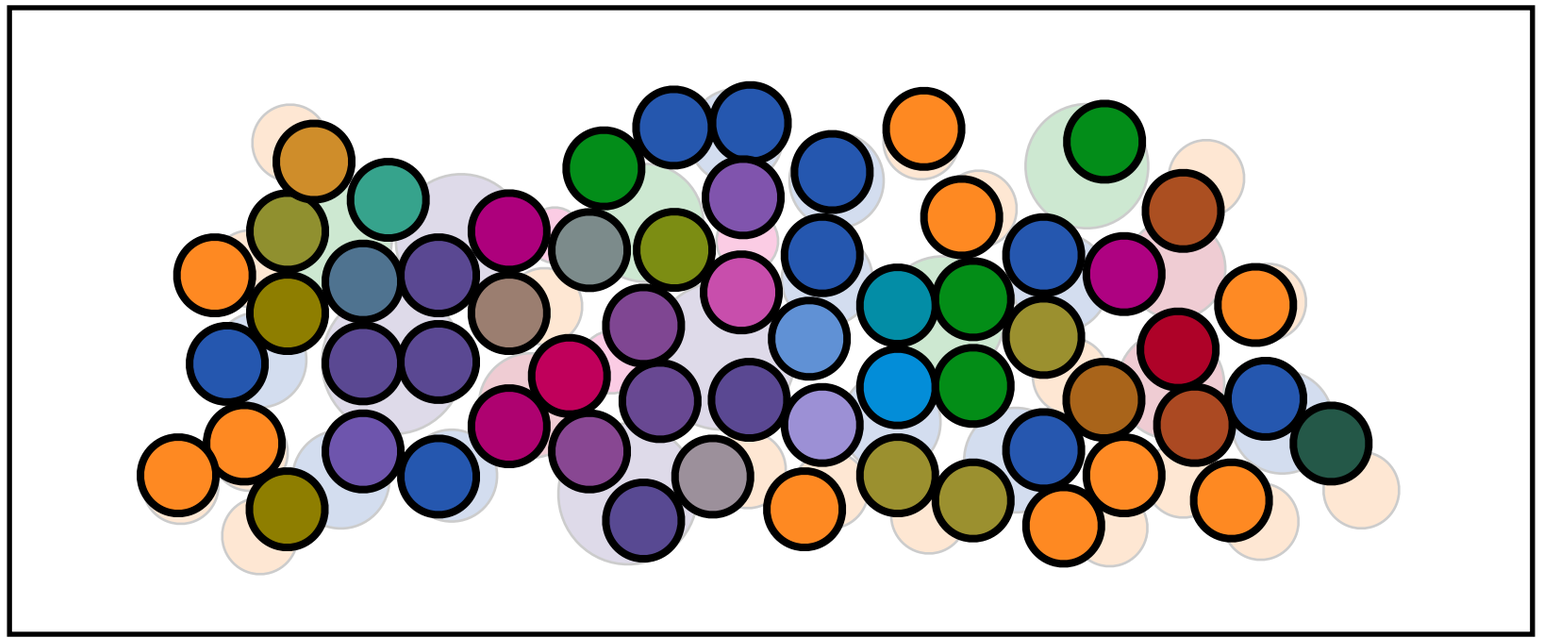
Beads densely cover the plane.

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



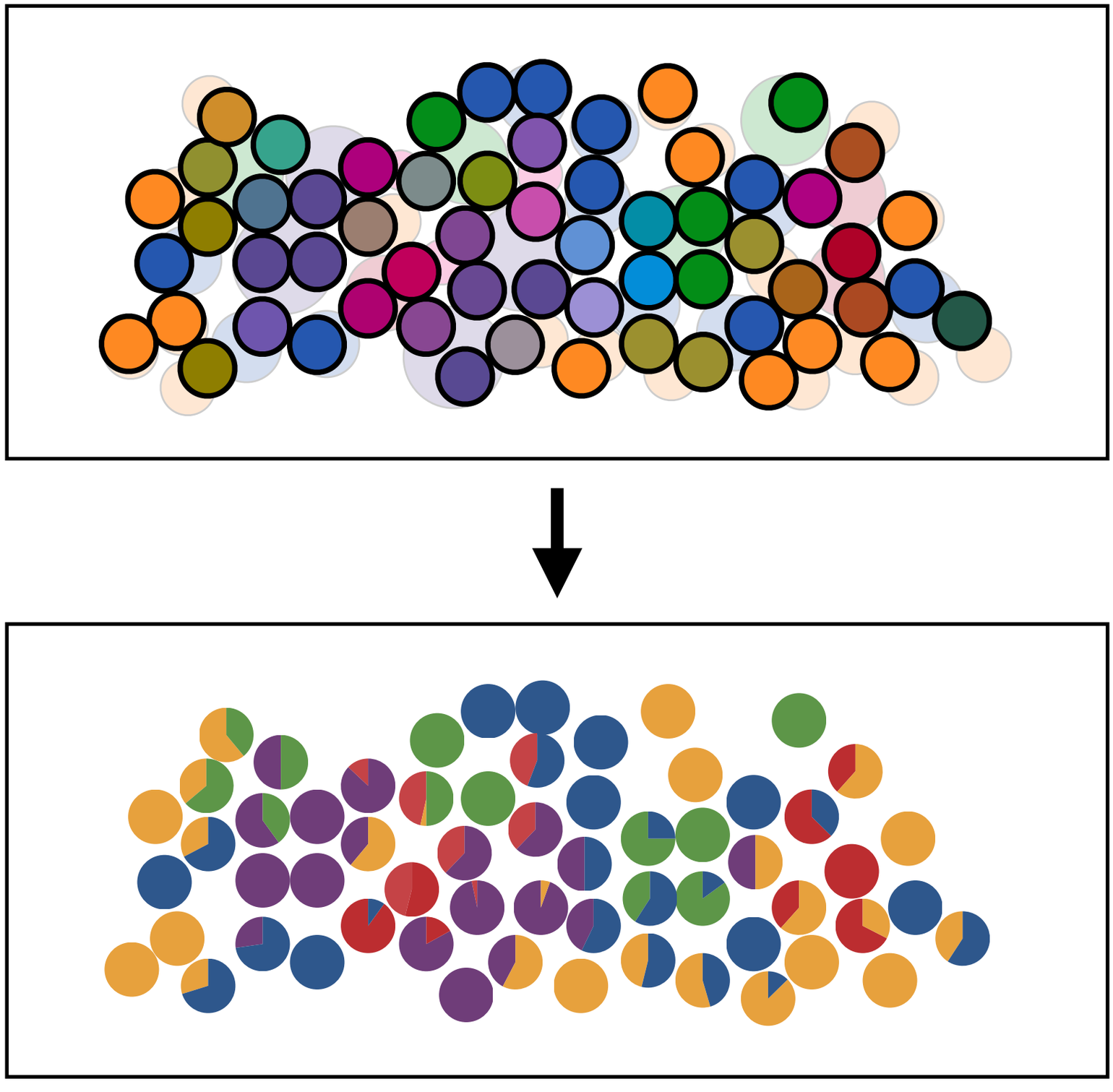
Locations of Slide-seq beads

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

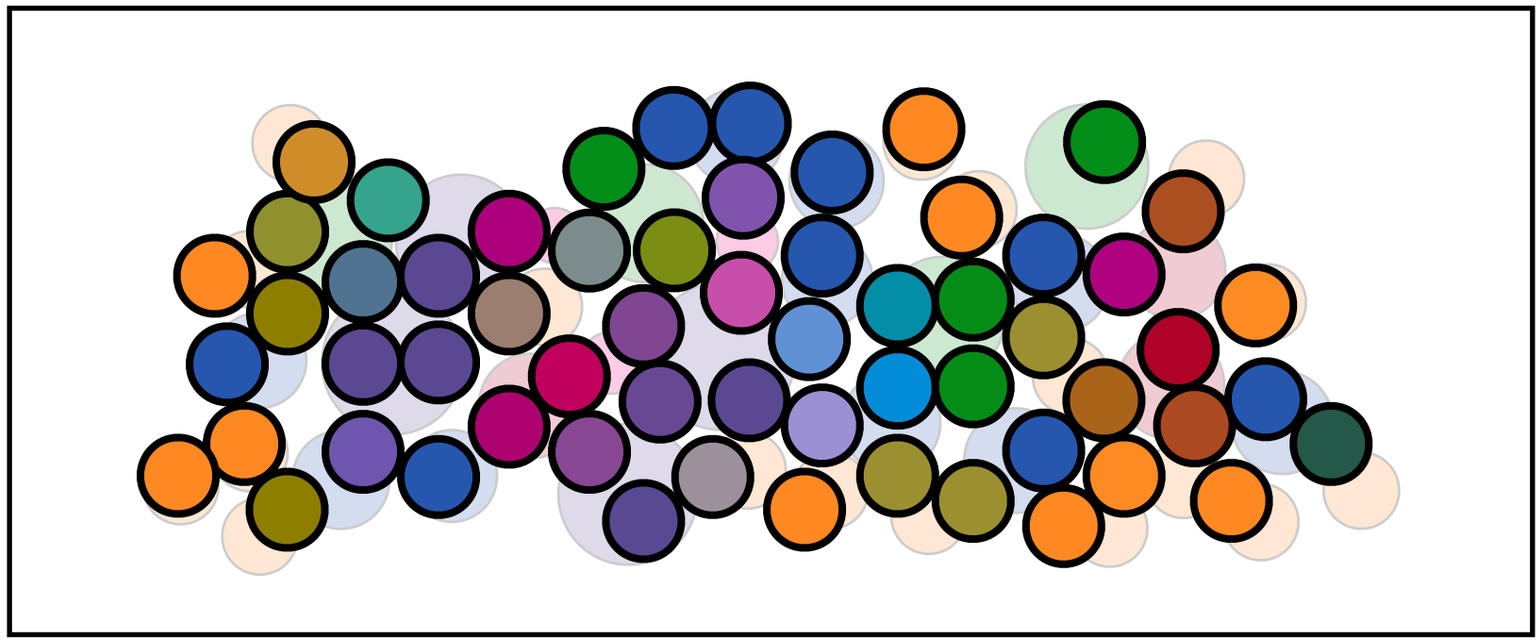


Slide-seq measurements

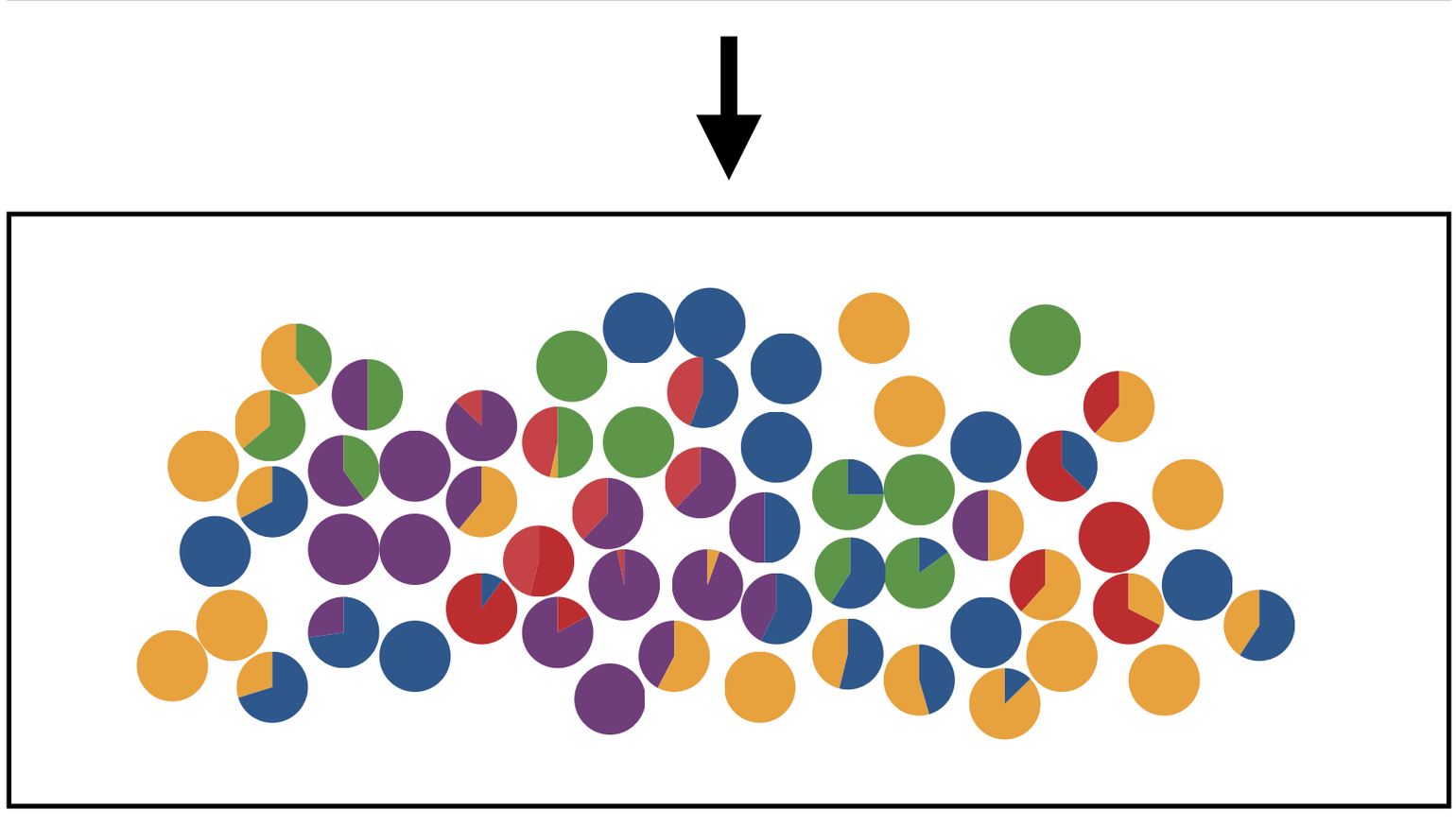
Given the Slide-seq observations, what are the cell types and the mixtures made out of?



Inverse problem

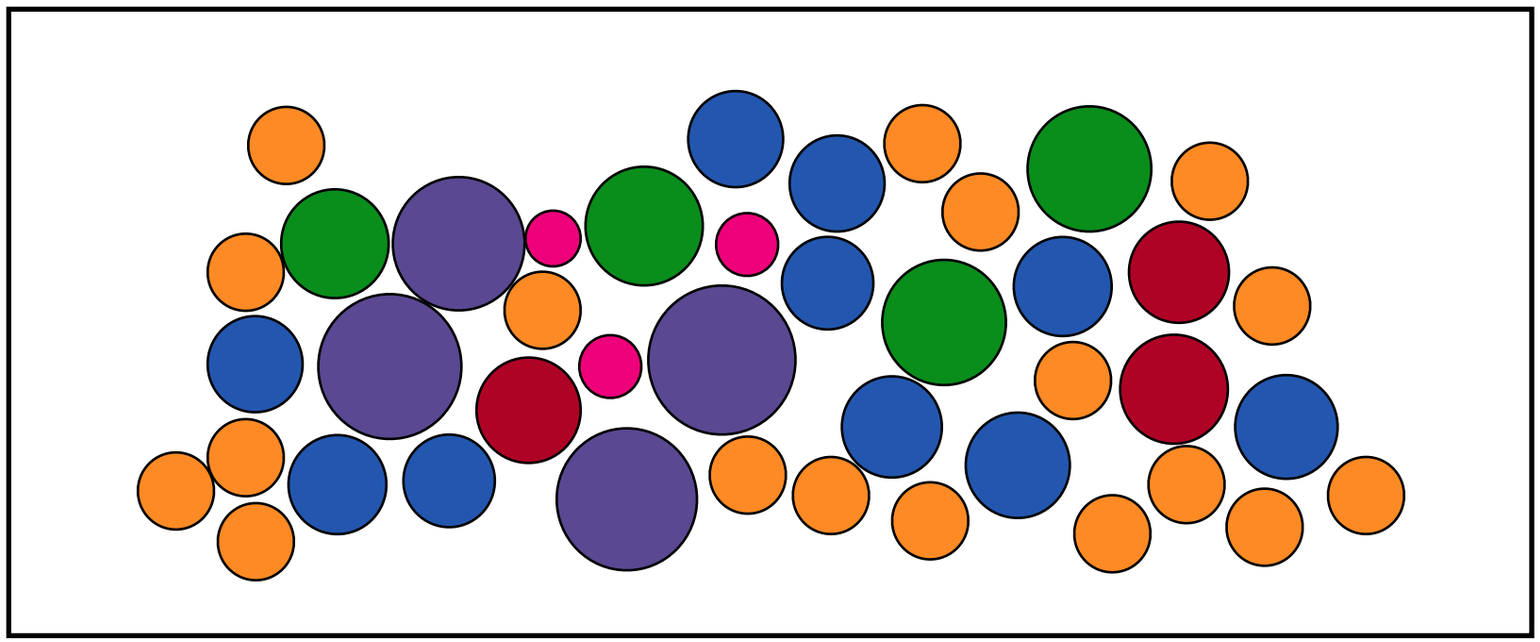


observations

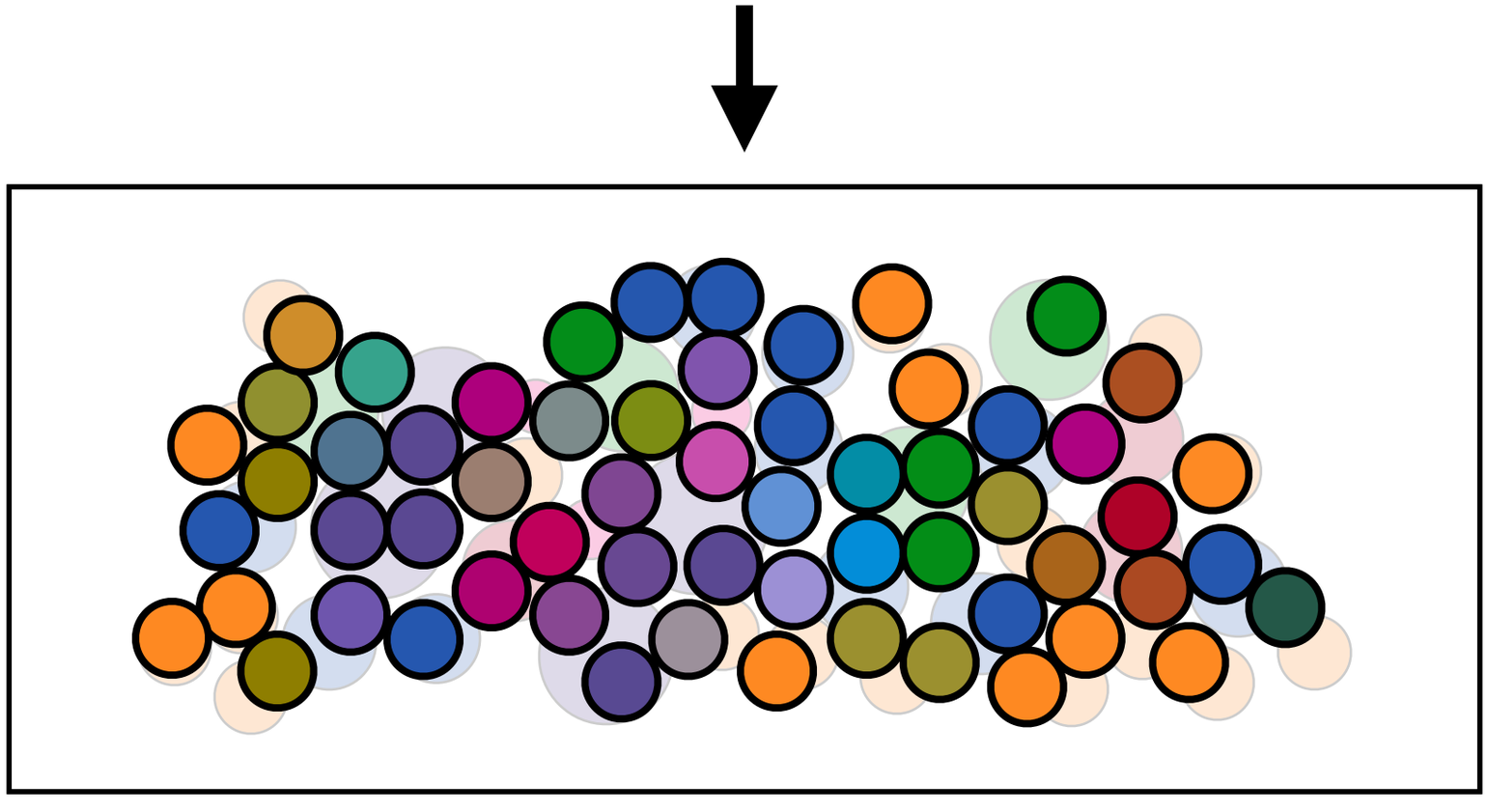


underlying truth

Forward problem



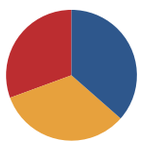
underlying truth



observations

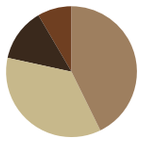
What is the **true mixture**?

=



or

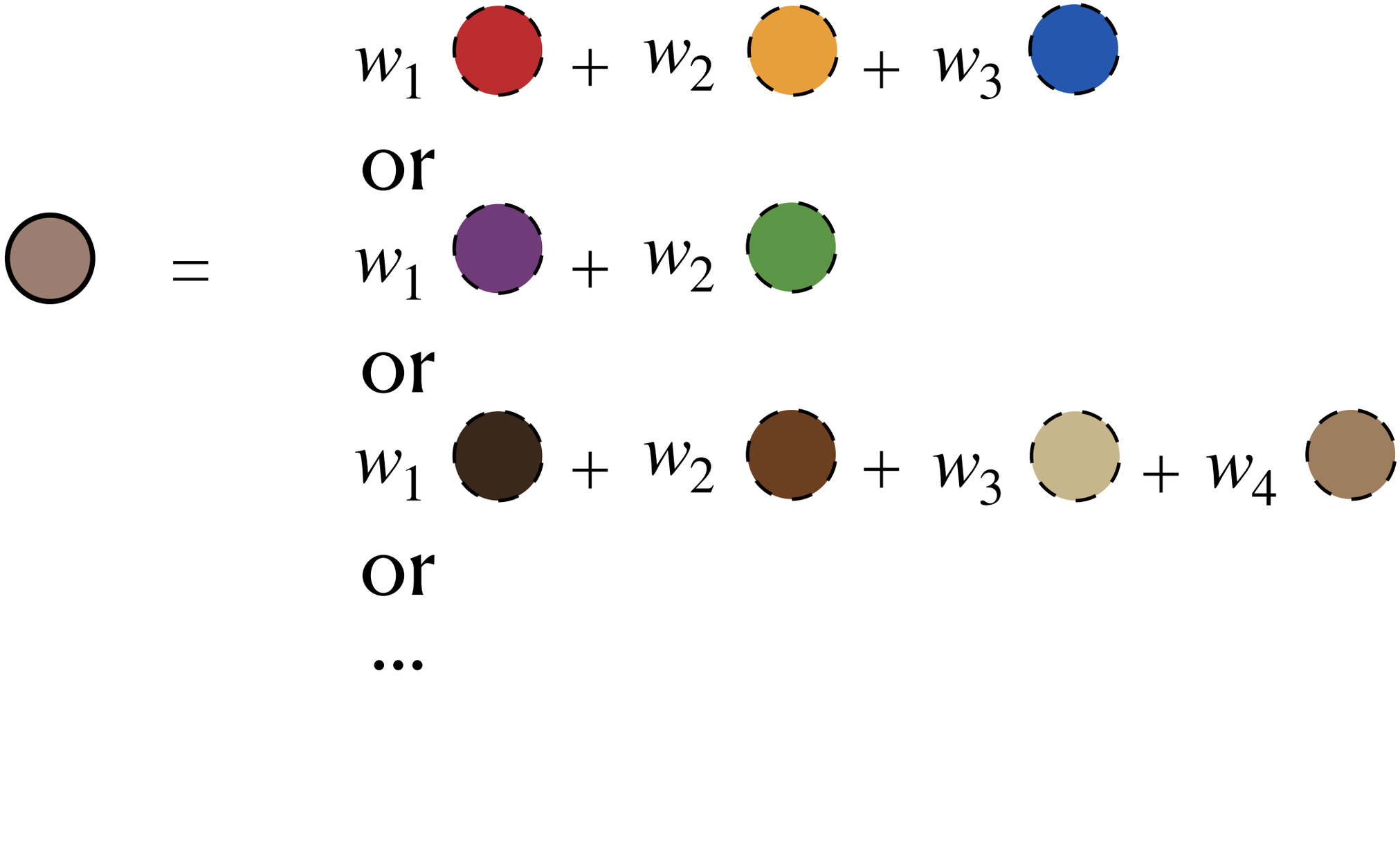
or



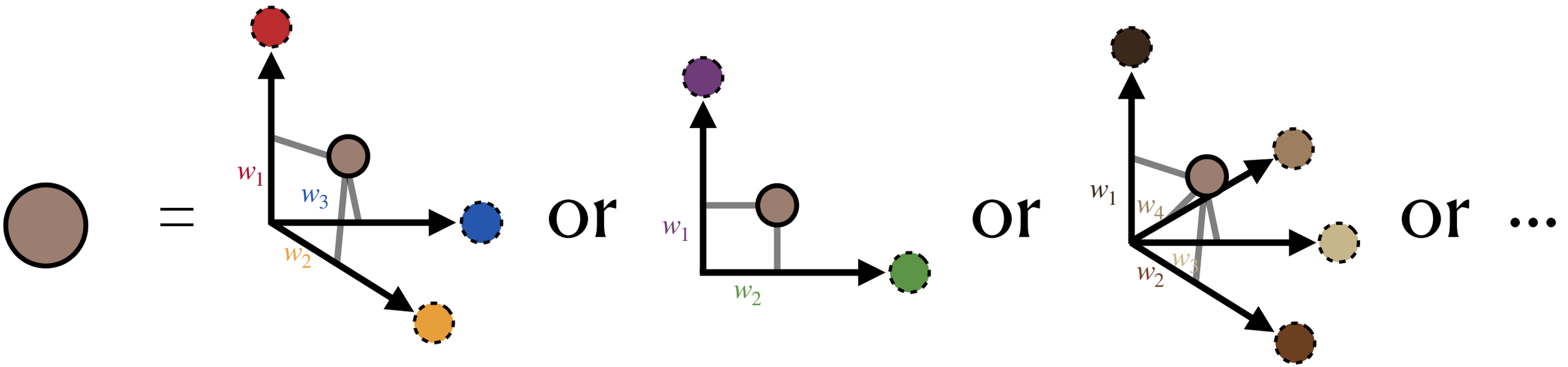
or

...

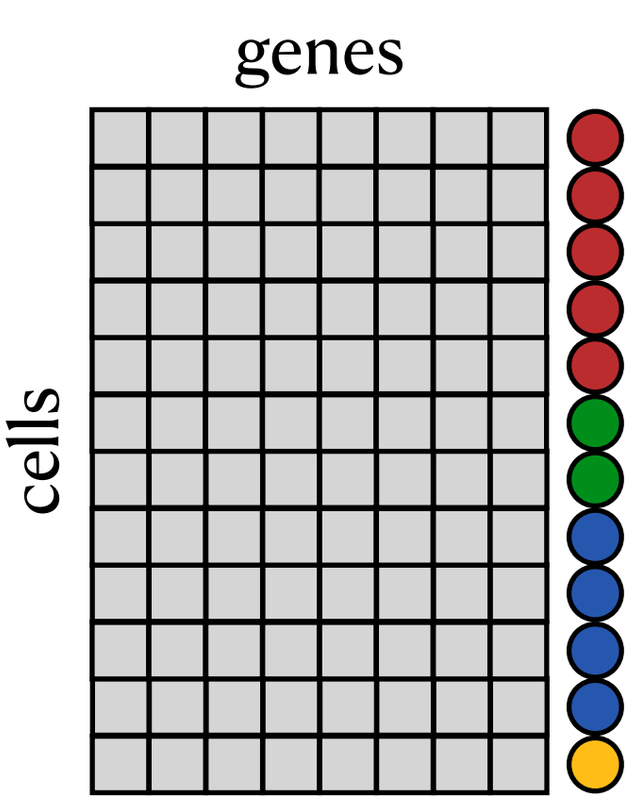
The same thing expressed **algebraically**.



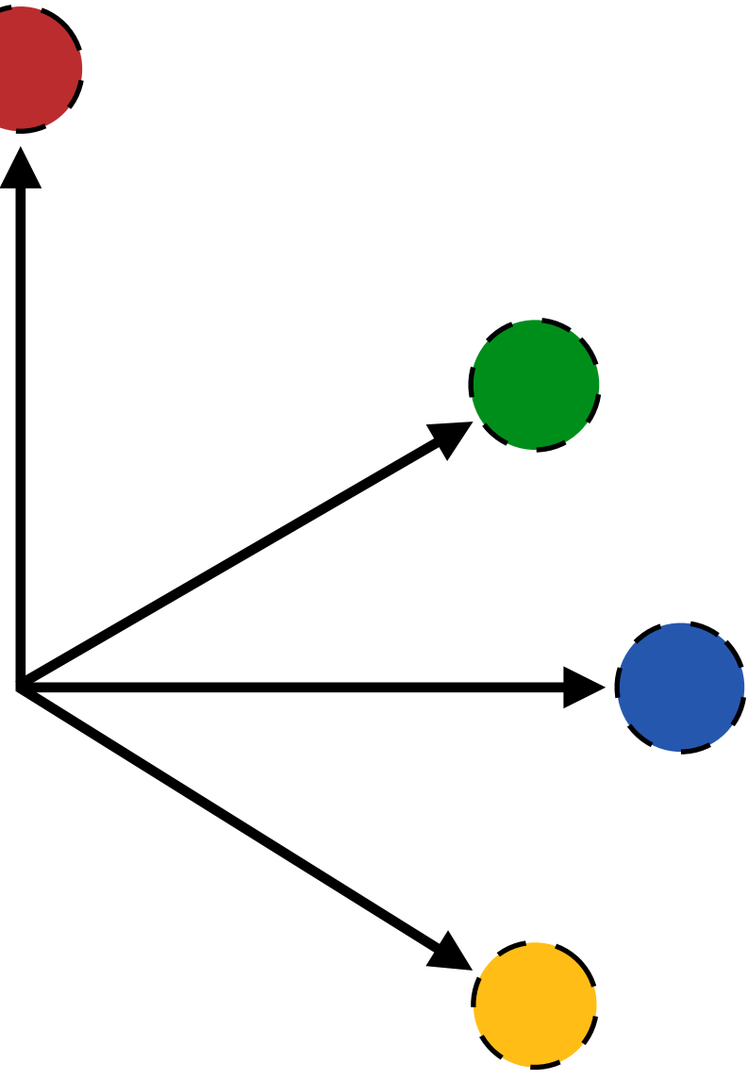
The same thing expressed **geometrically**.



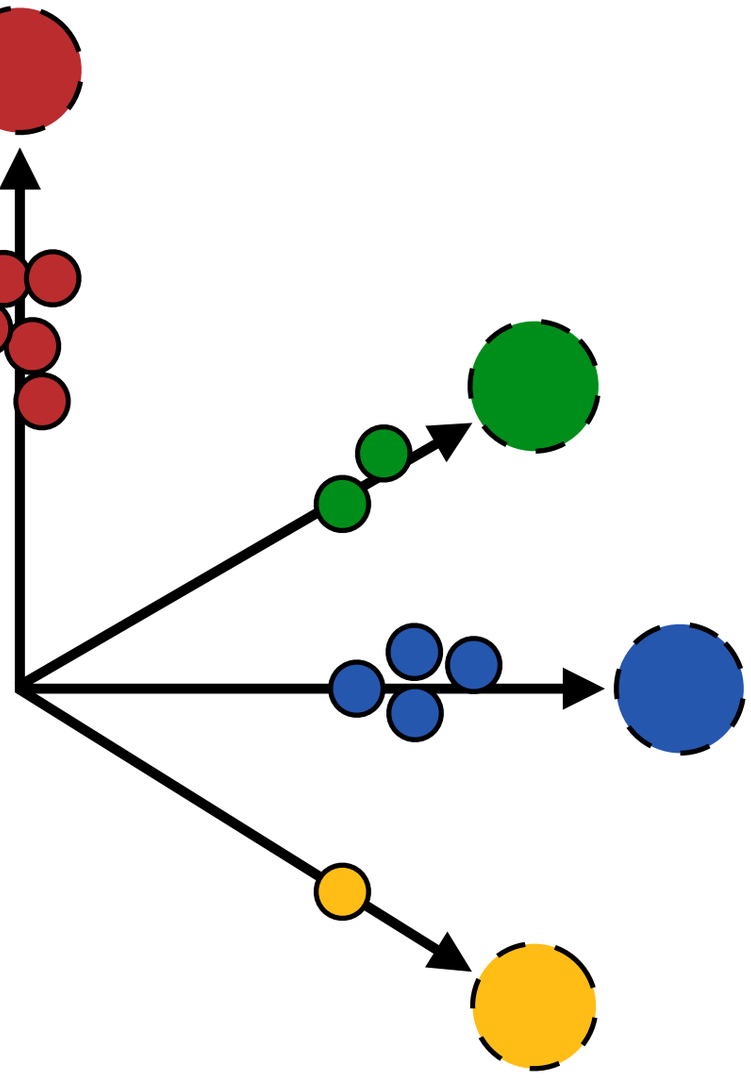
Leverage **prior knowledge** to define a **basis**.



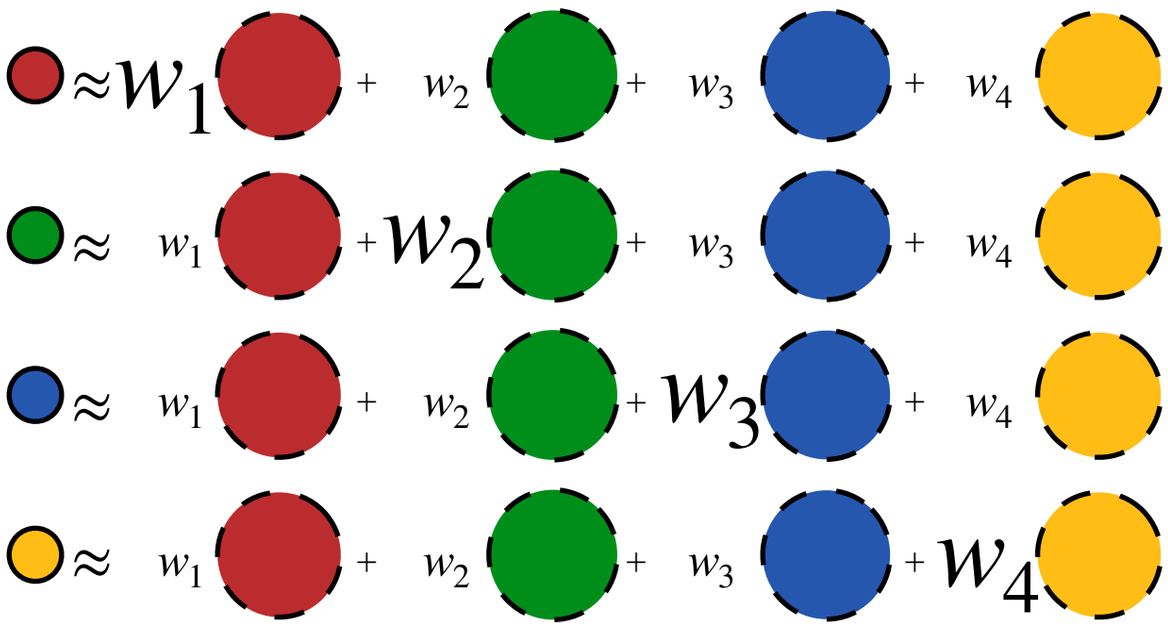
matrix of counts from
an **annotated reference**
single-cell RNAseq data

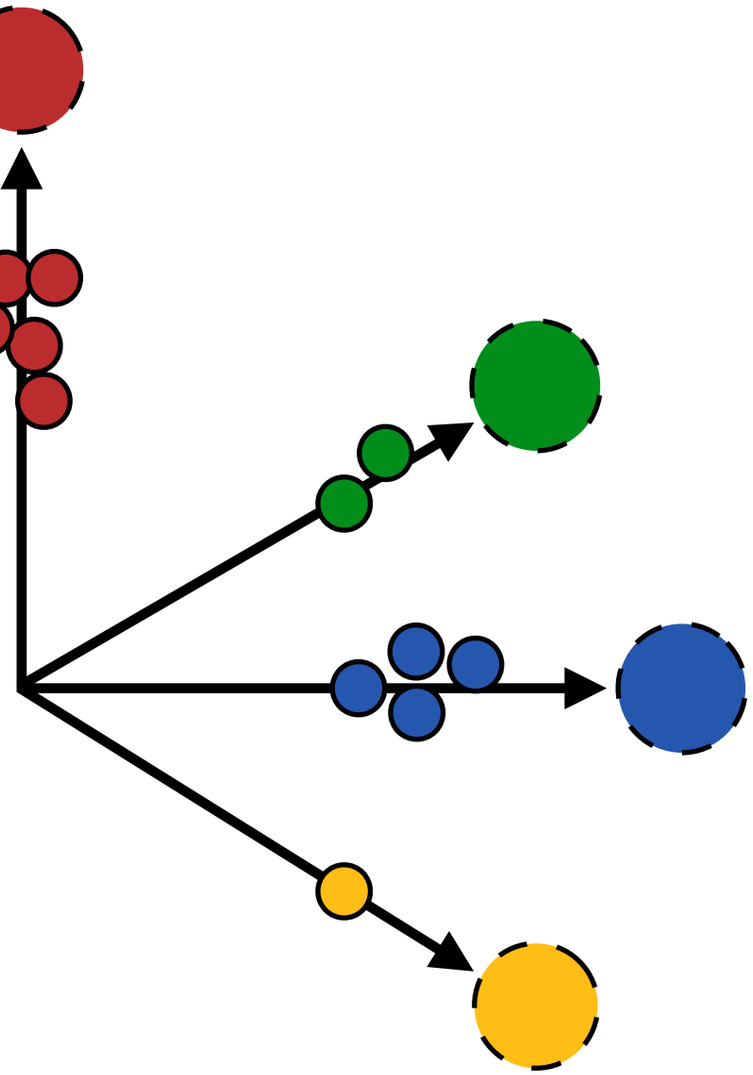


What is a **good basis**?

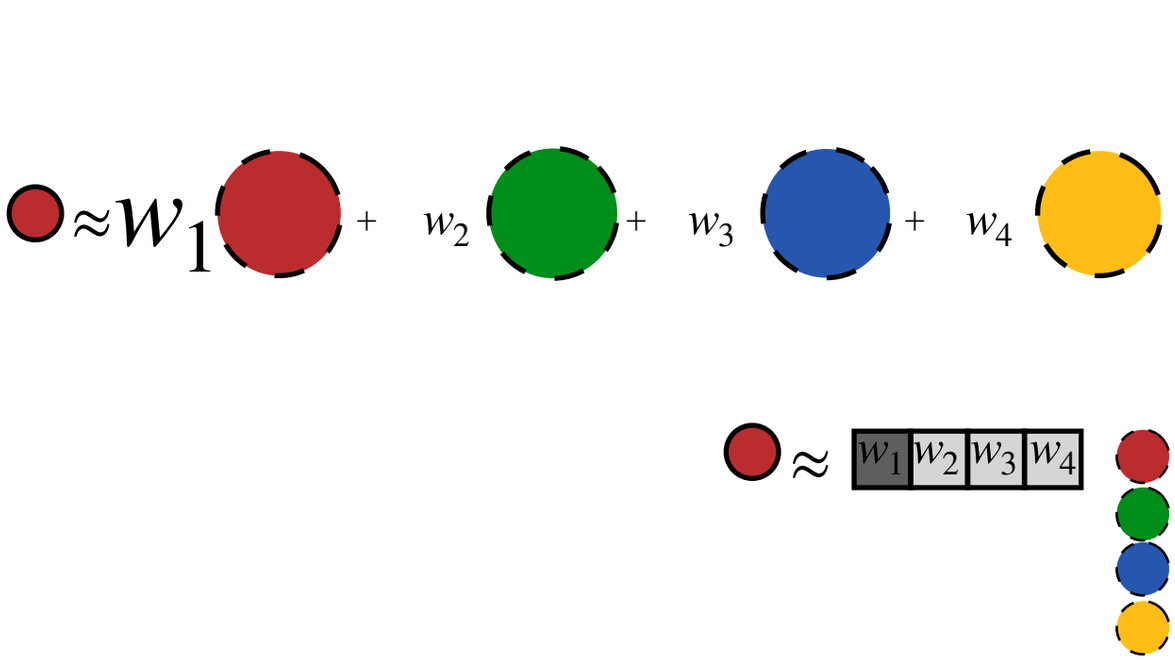


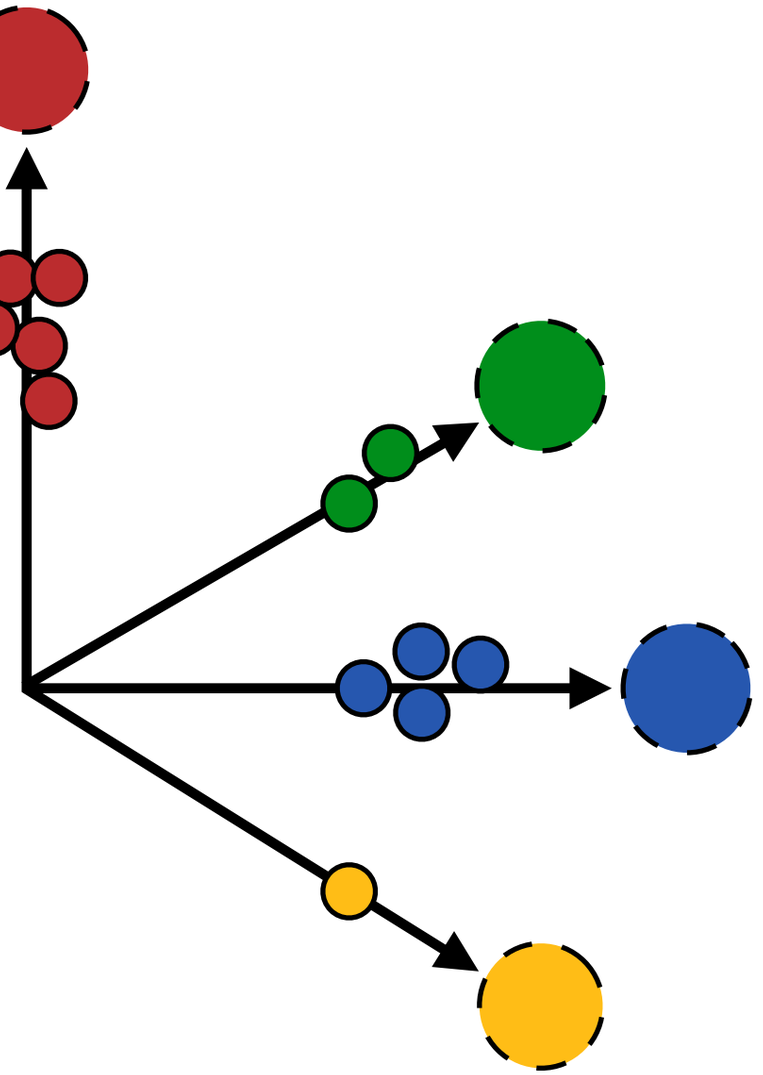
Written **algebraically**.



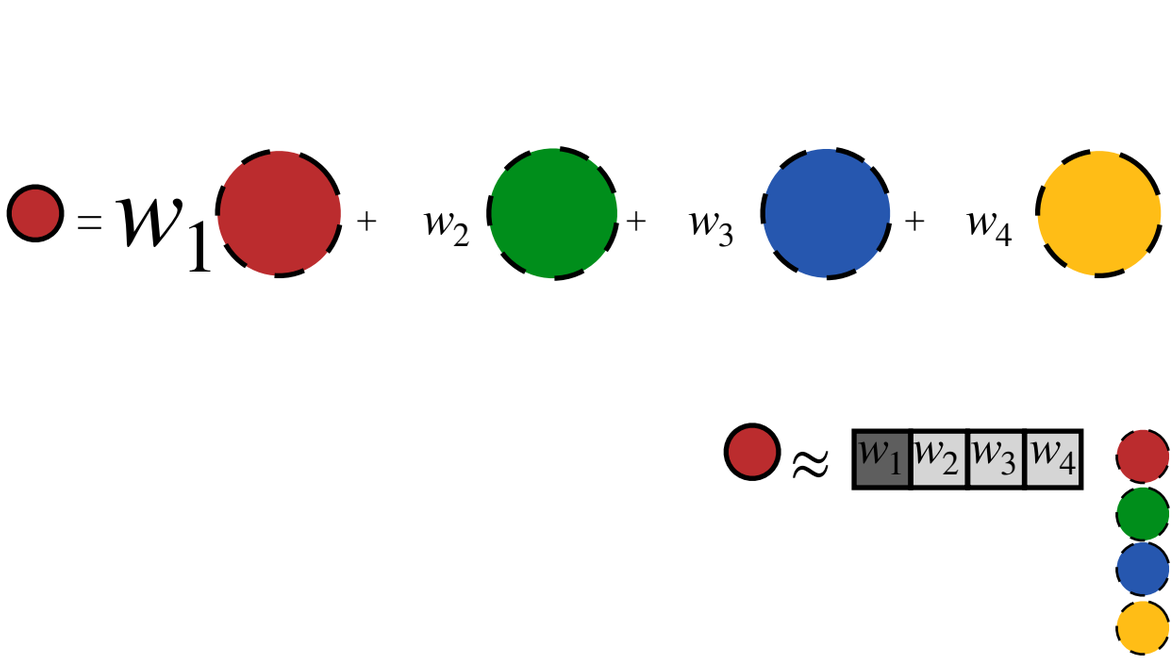


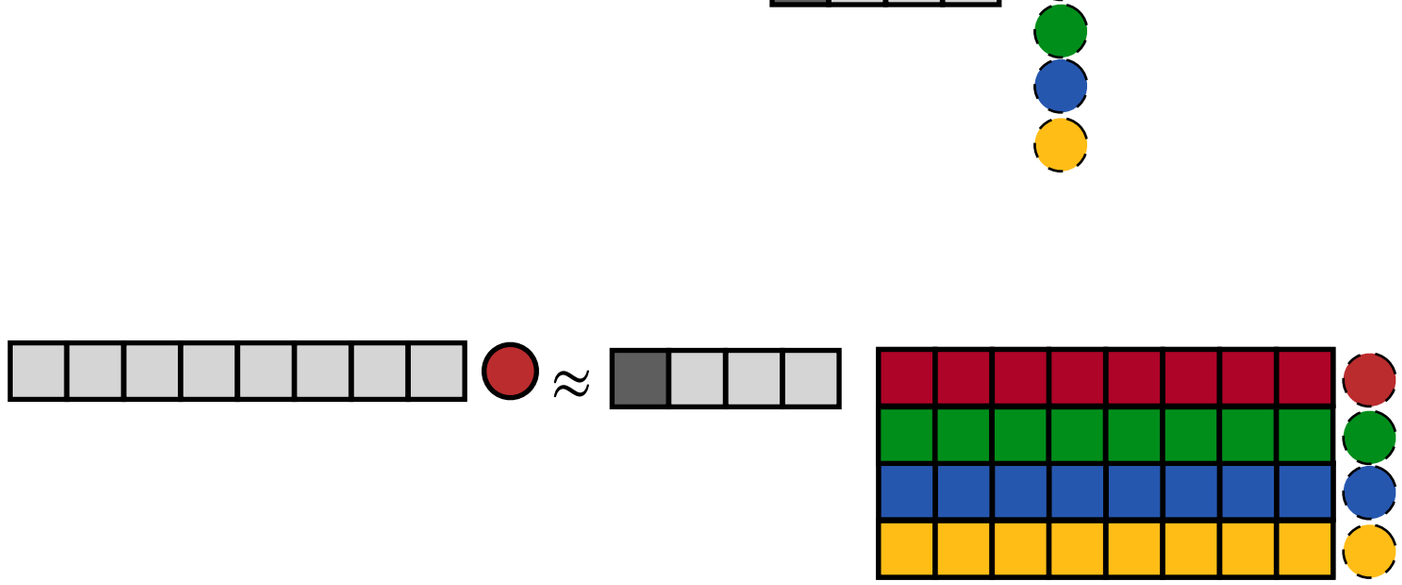
Written as a **vector multiplication.**

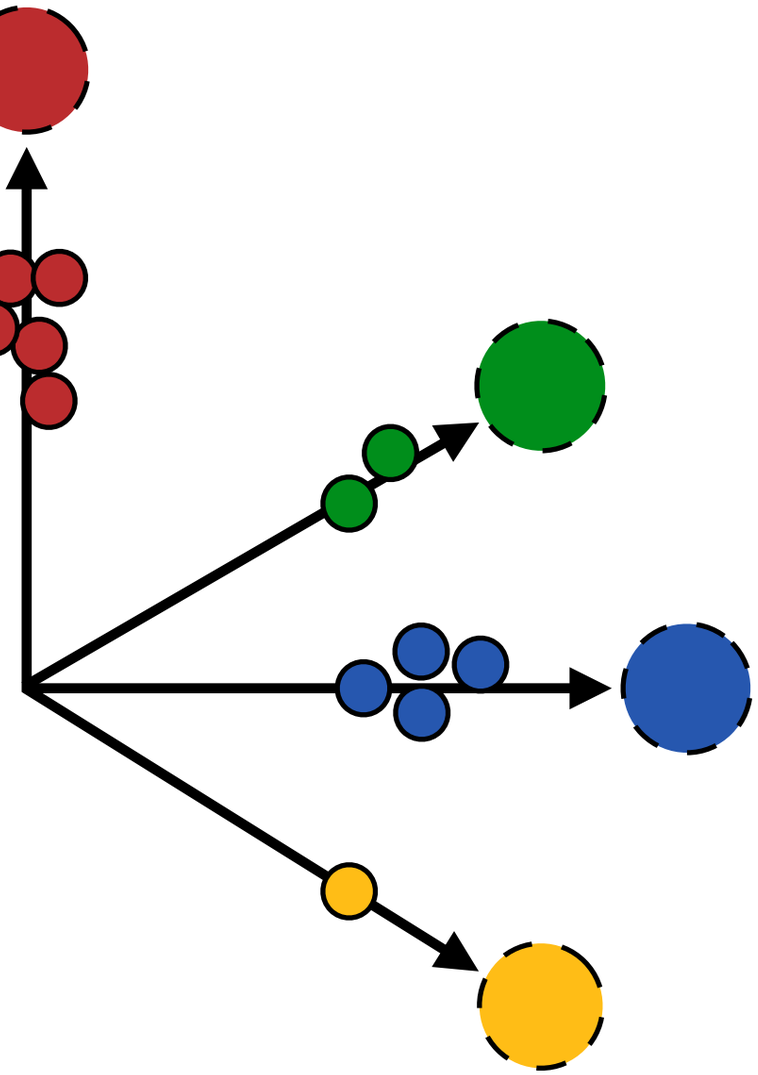




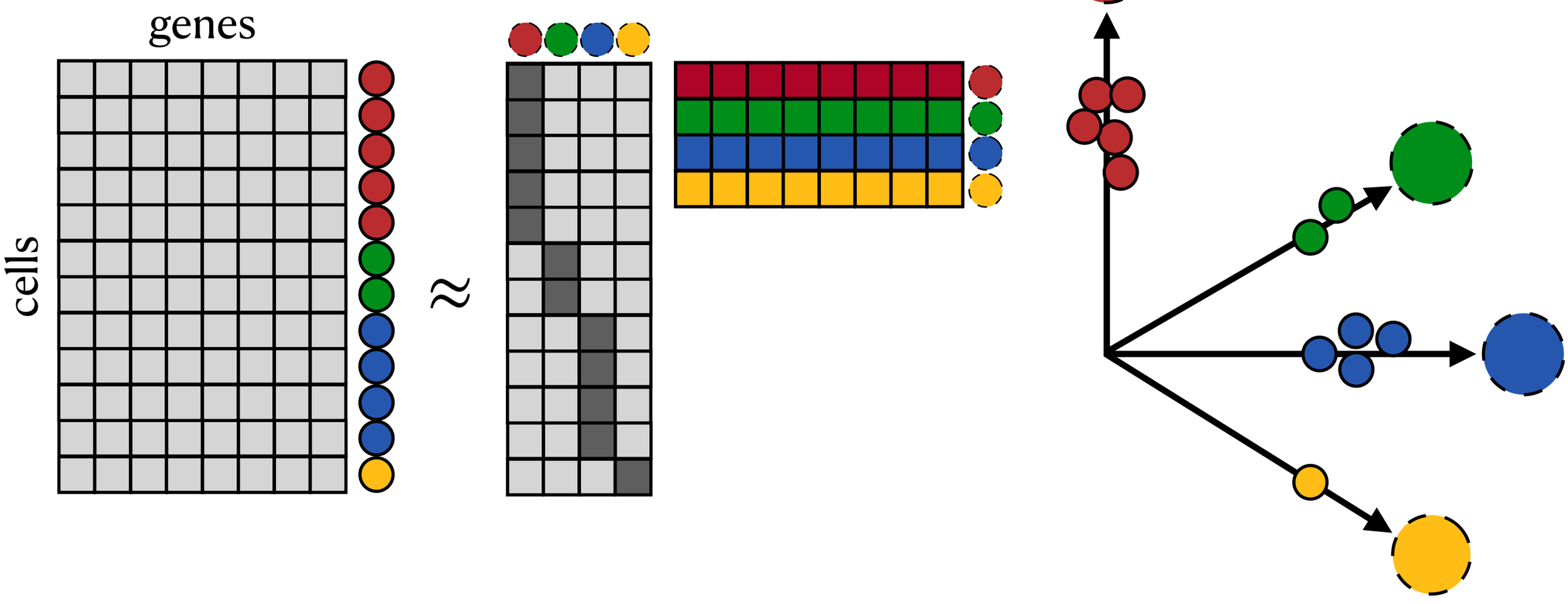
Written as a **vector multiplication.**



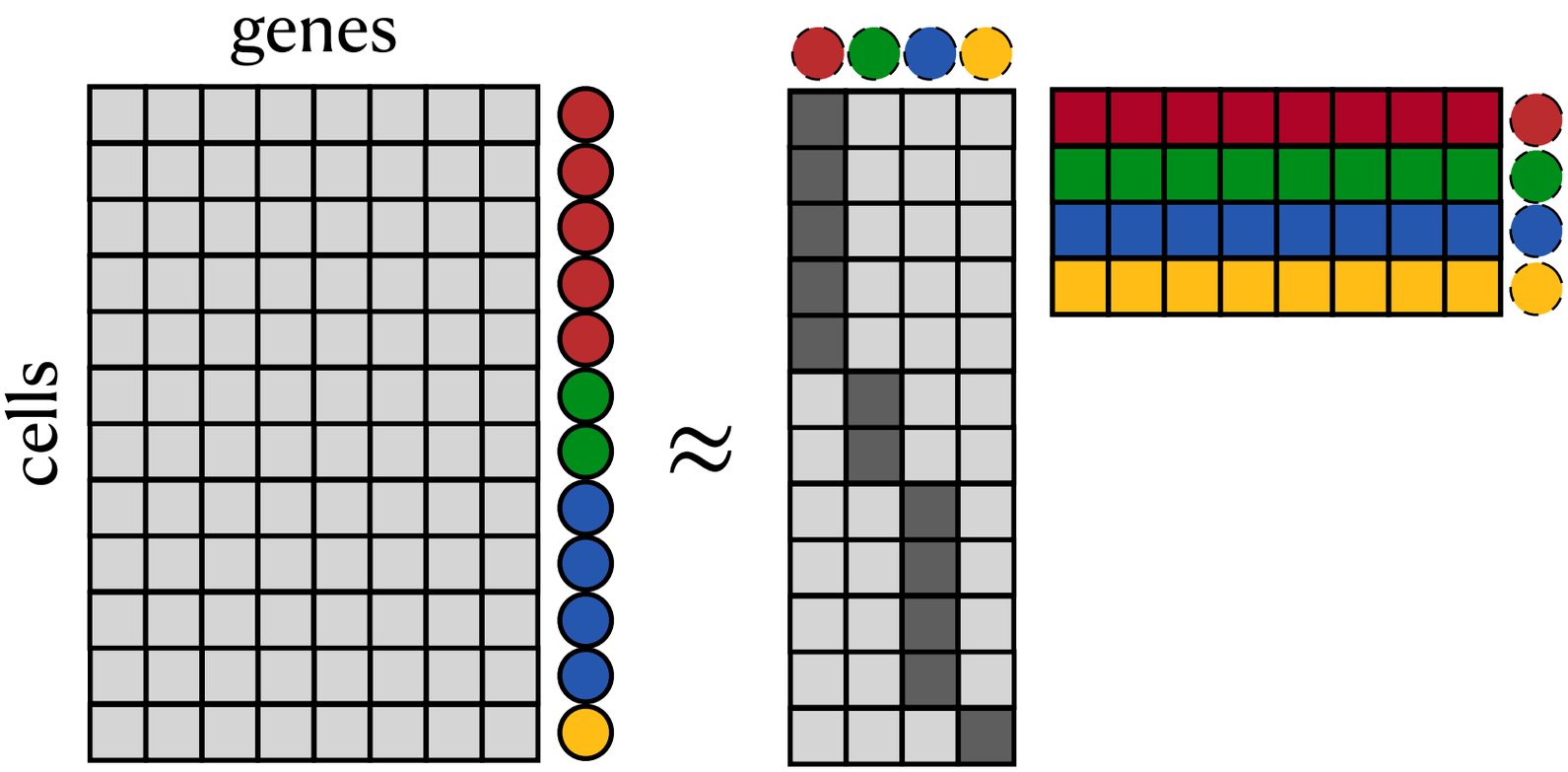




Matrix factorization.

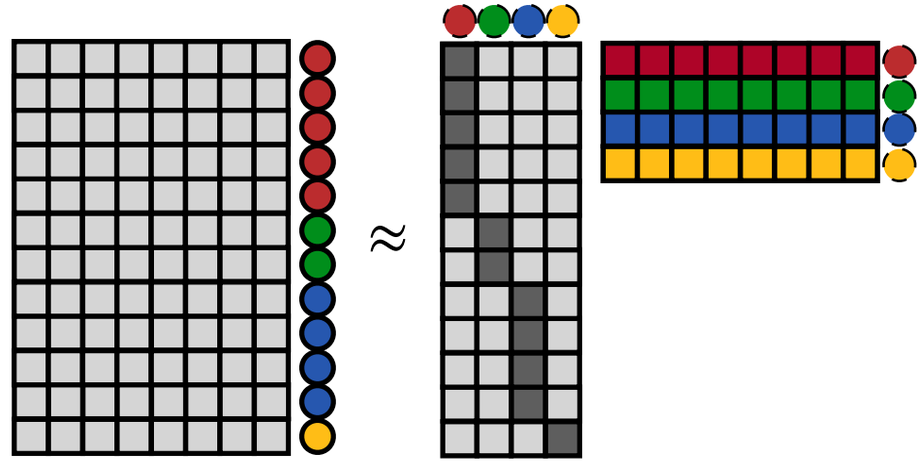


Linear algebra notation.



$$X \approx WH$$

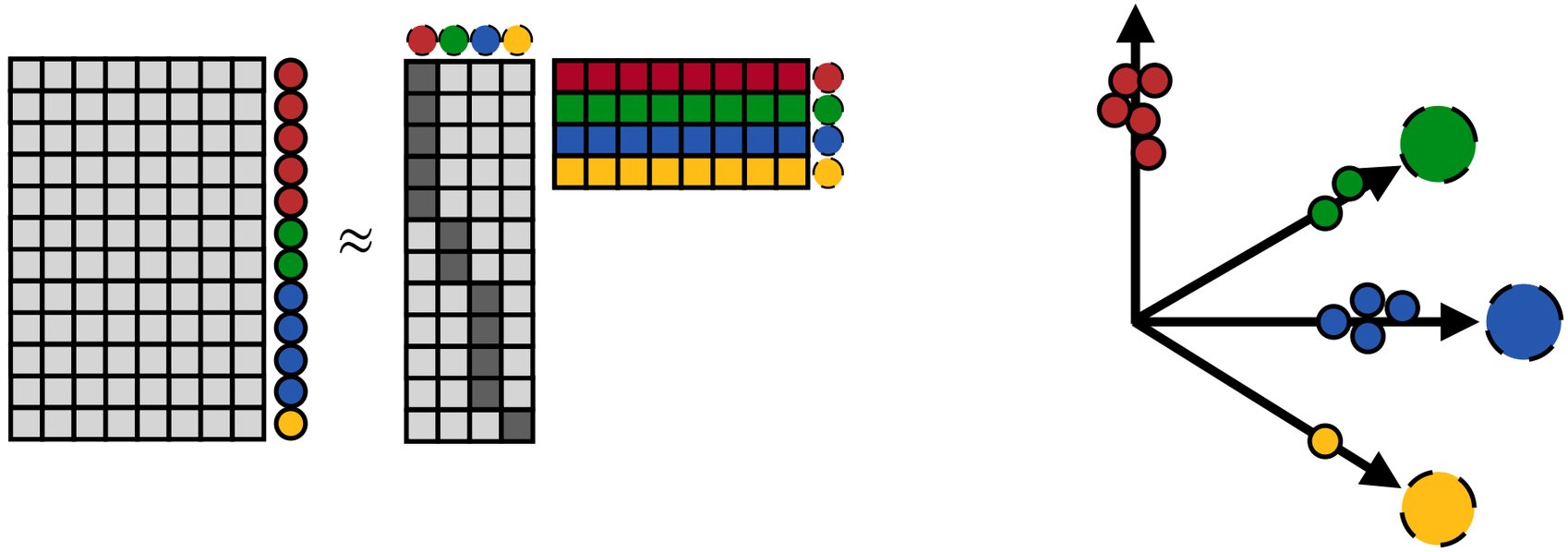
How to find W and H?



$$X \approx WH$$

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

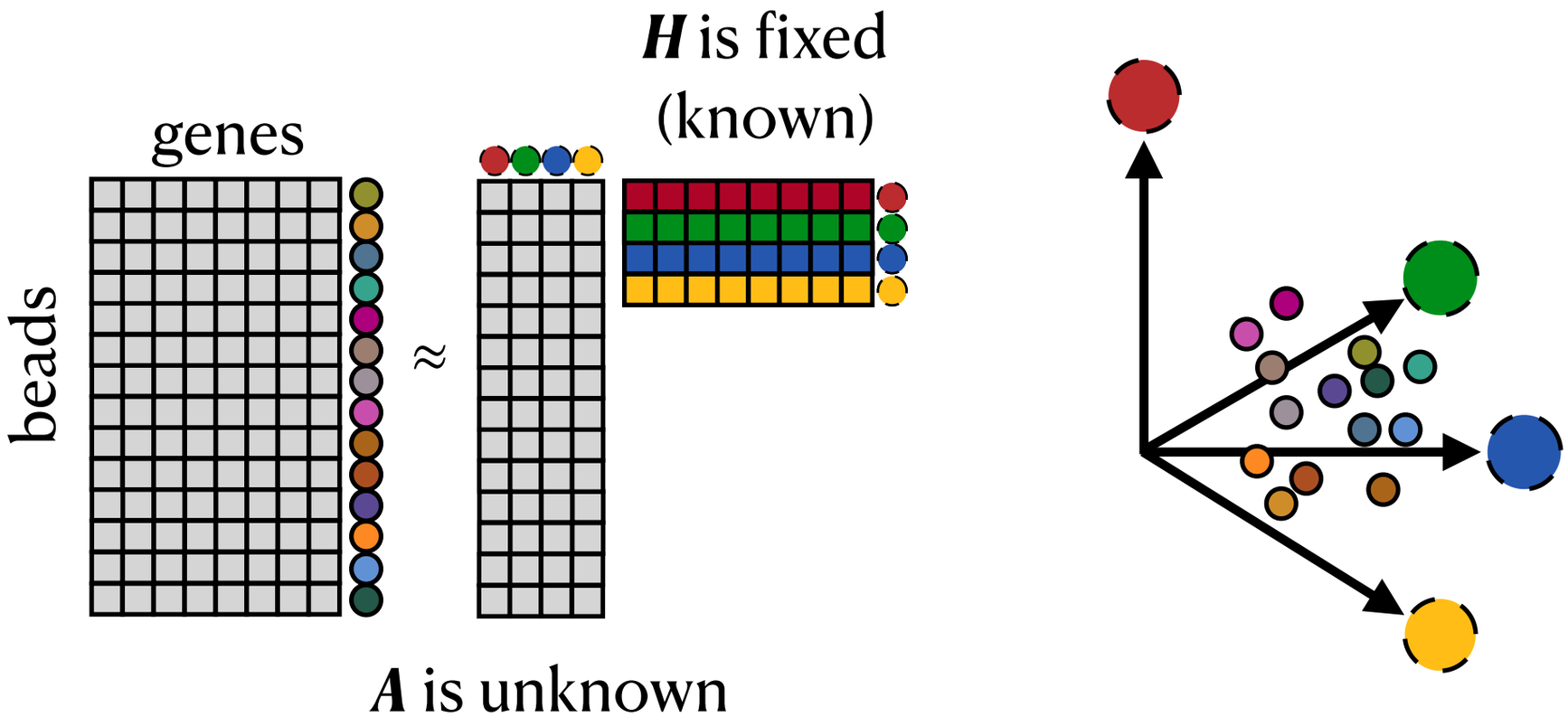
Non-negative Matrix Factorization (**NMF**).



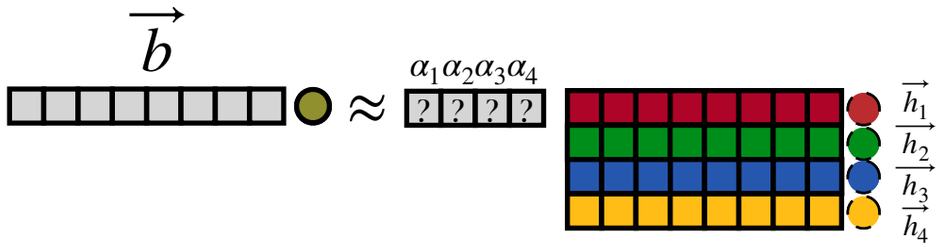
$$W, H = argmin \, ||X - WH||^2$$

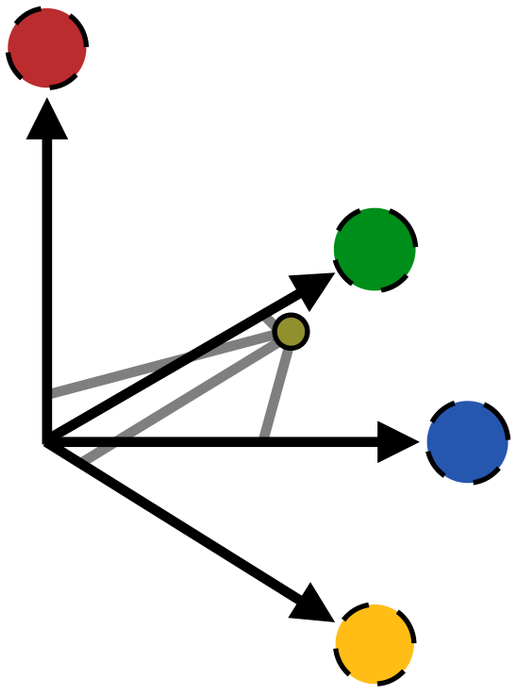
such that $W, H \geq 0$

Deconvolution of the mixed spatial beads.

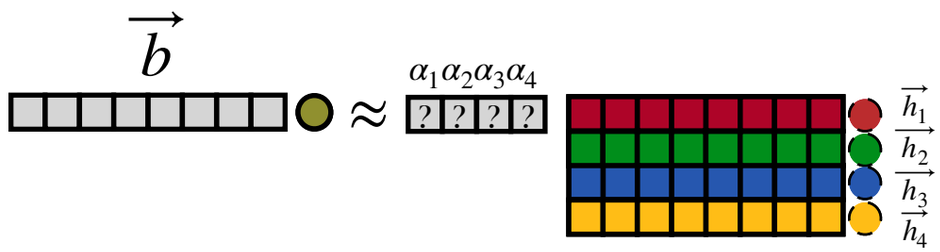


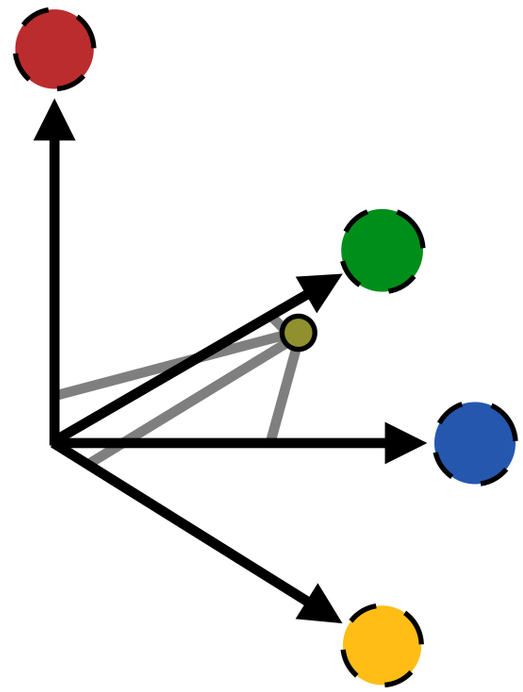
Finding the weights is doing **projection.**





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





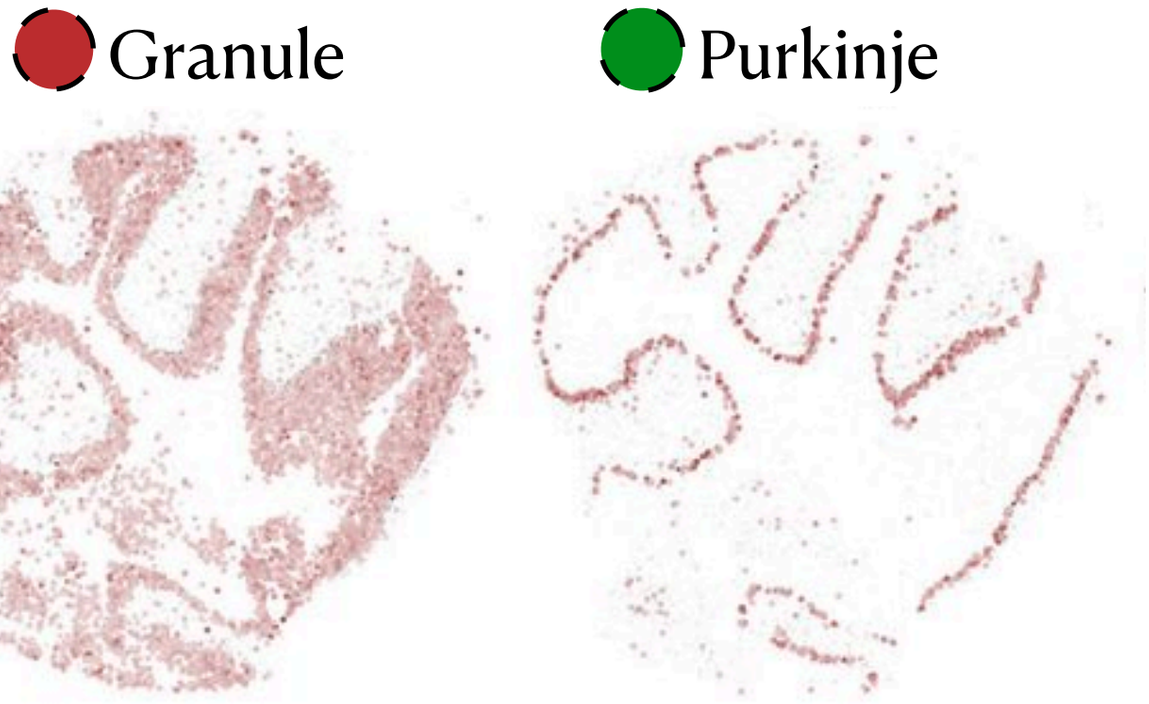
$$b \approx \alpha_1 . \overrightarrow{h_1} + \alpha_2 . \overrightarrow{h_2} + \alpha_3 . \overrightarrow{h_3} + \alpha_4 . \overrightarrow{h_4}$$

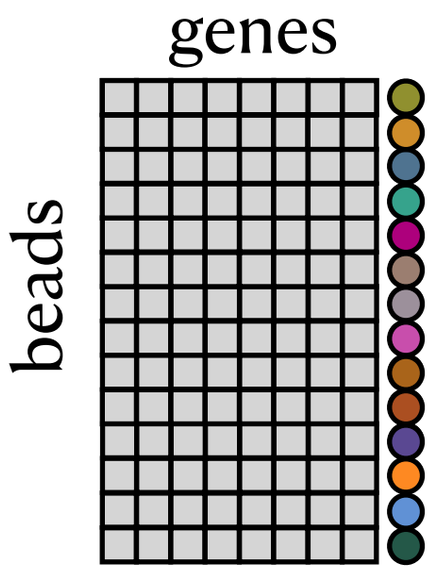
$$b \approx \overrightarrow{\alpha}^T H$$

$$B \approx AH$$

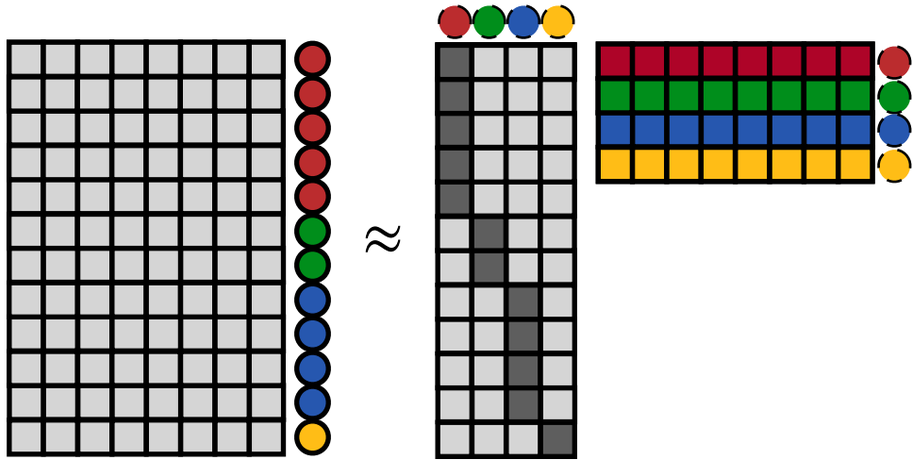
$$A = \underset{\text{such that } A \geq 0}{argmin} ||B - AH||^2$$

Did it work? **Validation?**





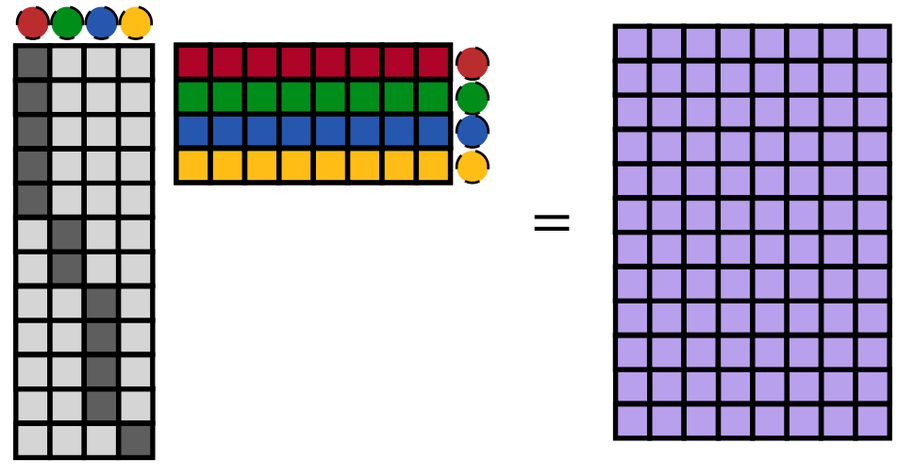
Let's take a fresh look at our model.



$$X \approx WH$$

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

such that $W, H \geq 0$

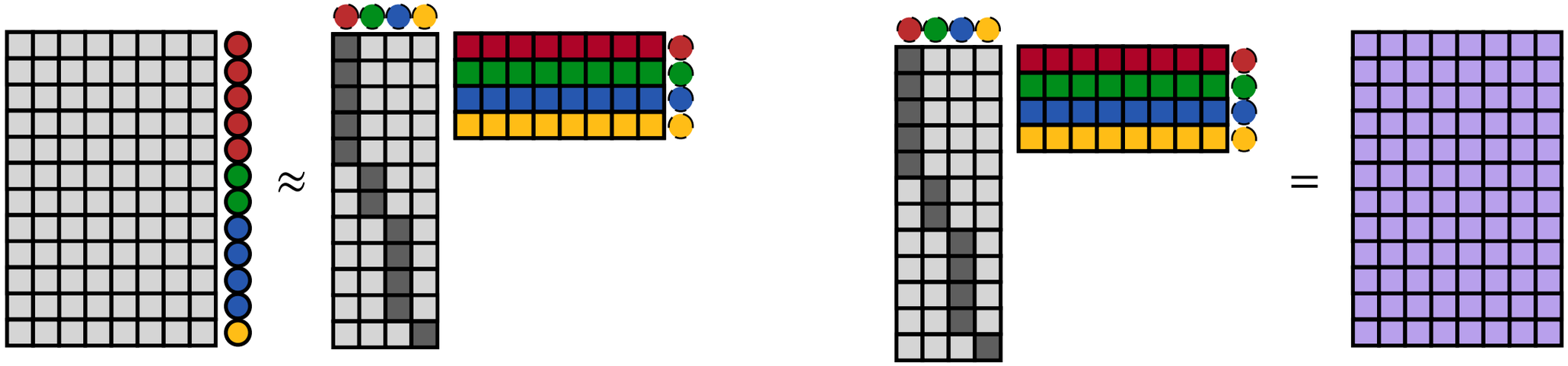


$$WH := \tilde{X}$$

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

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

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

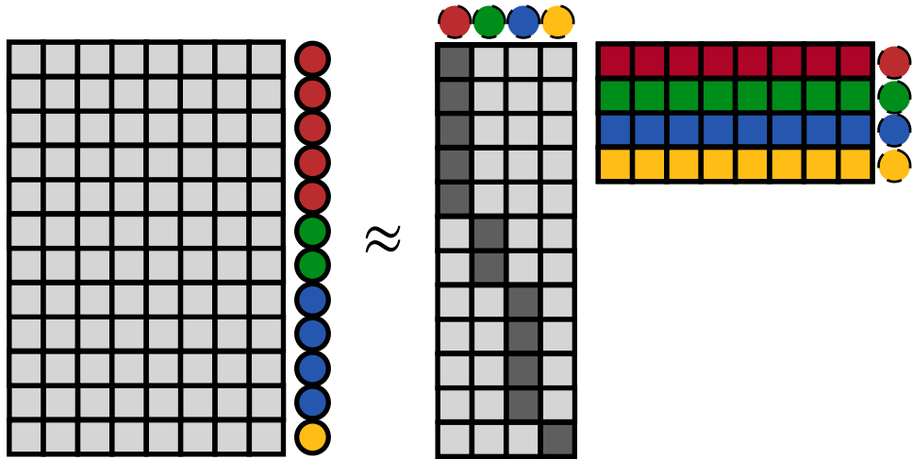


$$X \approx WH \qquad\qquad WH := \tilde{X}$$

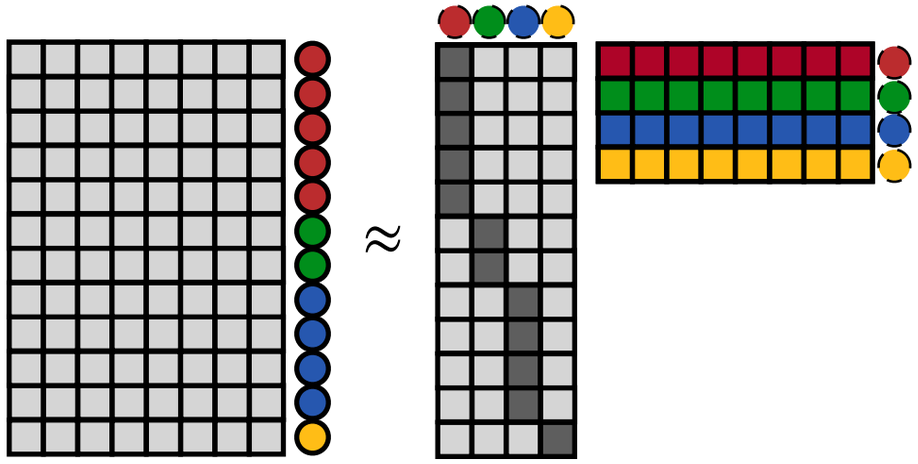
$$W, H = \underset{\text{such that } W,H \geq 0}{argmin} ||X - WH||^2 = \underset{\text{such that } \tilde{X} = WH \text{ and } W,H \geq 0}{argmin} ||X - \tilde{X}||^2$$

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

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



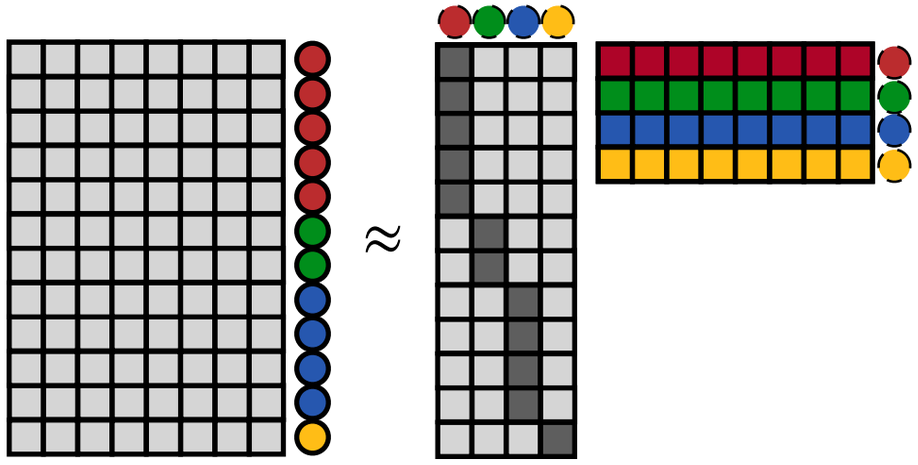
But is this a math model, or a stats model?



a **bilinear** model

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

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



a **bilinear** model

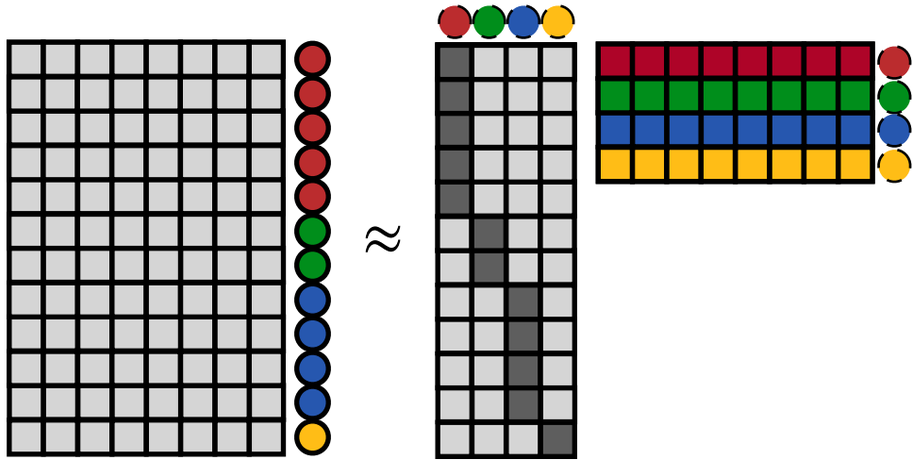
$$\tilde{X} = argmin\,||X - \tilde{X}||^2$$

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

maximum likelihood estimator

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

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



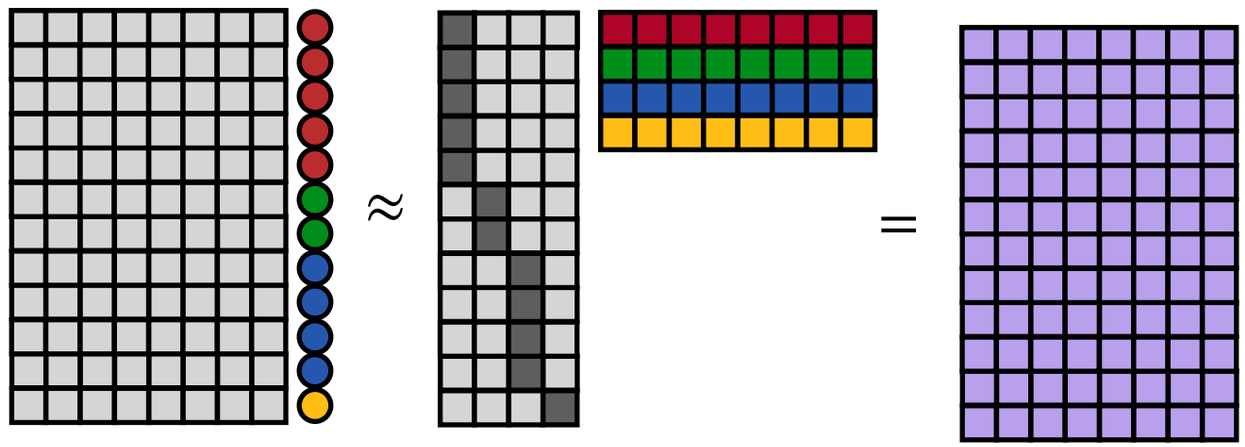
a **bilinear** model

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

maximum likelihood estimator

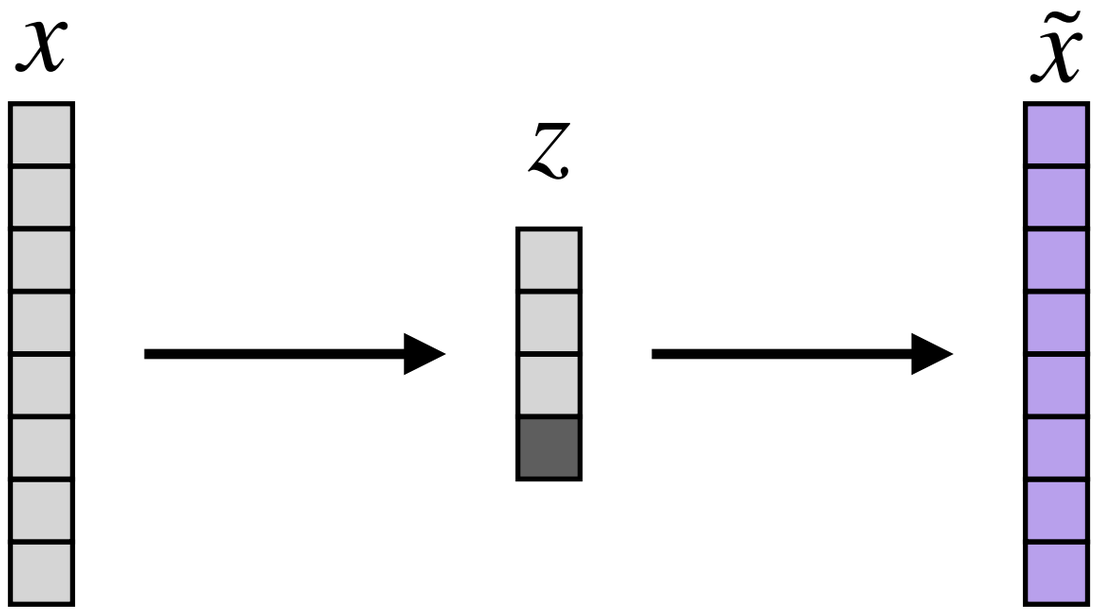
when $\overrightarrow{X_{ij}} \sim \text{Poisson}(e^{\tilde{X}_{ij}})$

Focus on one observation.



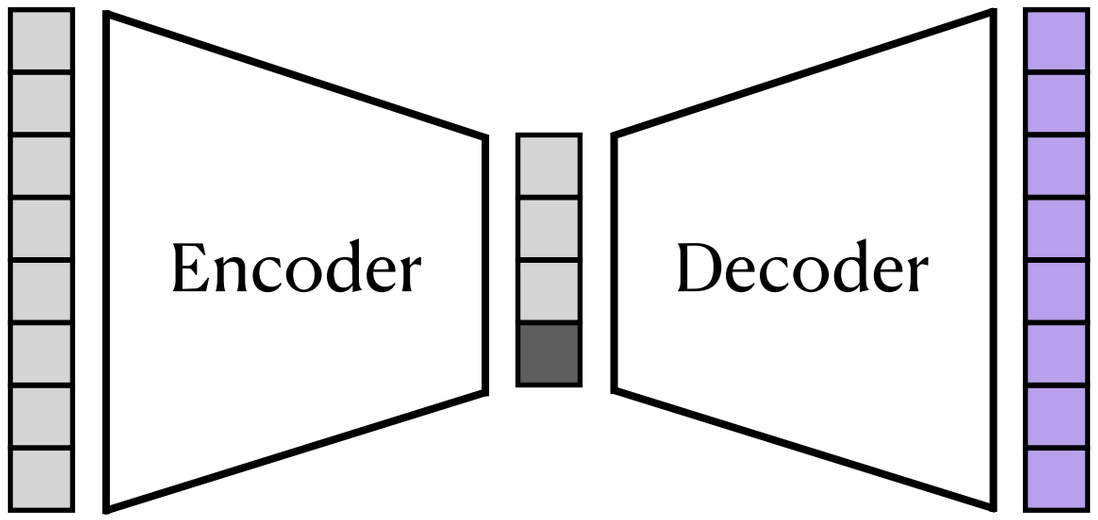
$$X \approx \tilde{X}$$

Dimensionality Reduction.



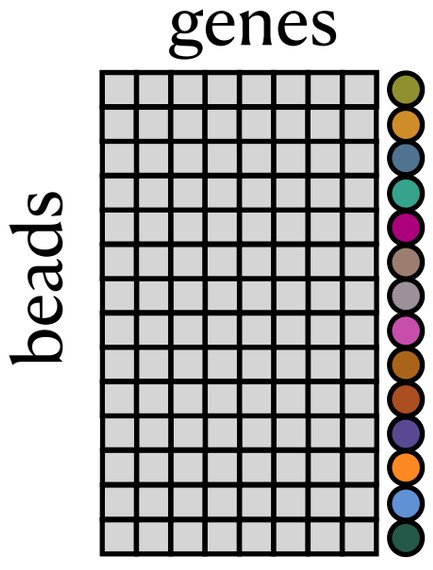
Representation Learning.

Dimensionality Reduction.
Representation Learning.



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

Which approach should we use?



Granule

Purkinje

Take-home messages:

- Incorporate prior domain knowledge into your models.
- Always try a simple baseline first.

Take-home messages:

- Incorporate prior domain knowledge into your models.
- Always try a simple baseline first.
- It is not always about finding the model that improves performance on a metric.
- Continual iteration between a team of people (including domain scientists, ML, and software engineers) is necessary.

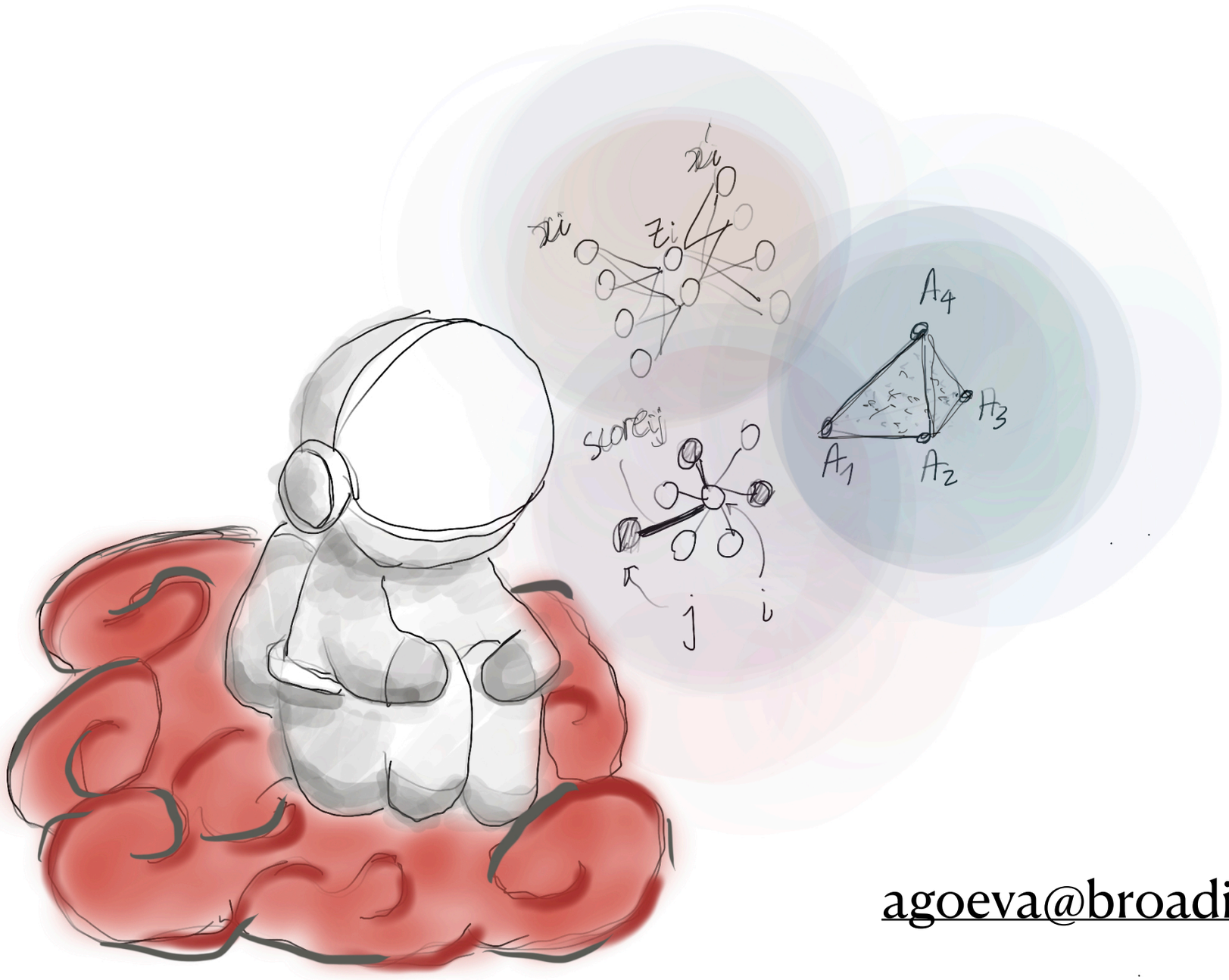
Take-home messages:

- Incorporate prior domain knowledge into your models.
- Always try a simple baseline first.
- It is not always about finding the model that improves performance on a metric.
- Continual iteration between a team of people (including domain scientists, ML, and software engineers) is necessary.
- **Disregard!**

  "You have to worry about your own work and ignore what everyone else is doing."

  - Richard Feynman, 1965

Be in touch!



agoeva@broadinstitute.org

References

- Aviv Regev's talk LMRL NeurIPS 2019
- Method of the year 2021: spatially resolved transcriptomics
- My NMF primer at MIA
- NMF + NNLS python code and tutorial on GitHub
- Slide-seq Science paper 2019
- Generalized Bilinear Models, Jeff Miller's MIA talk 2020
- Feynman: "Disregard!"