BILD 62

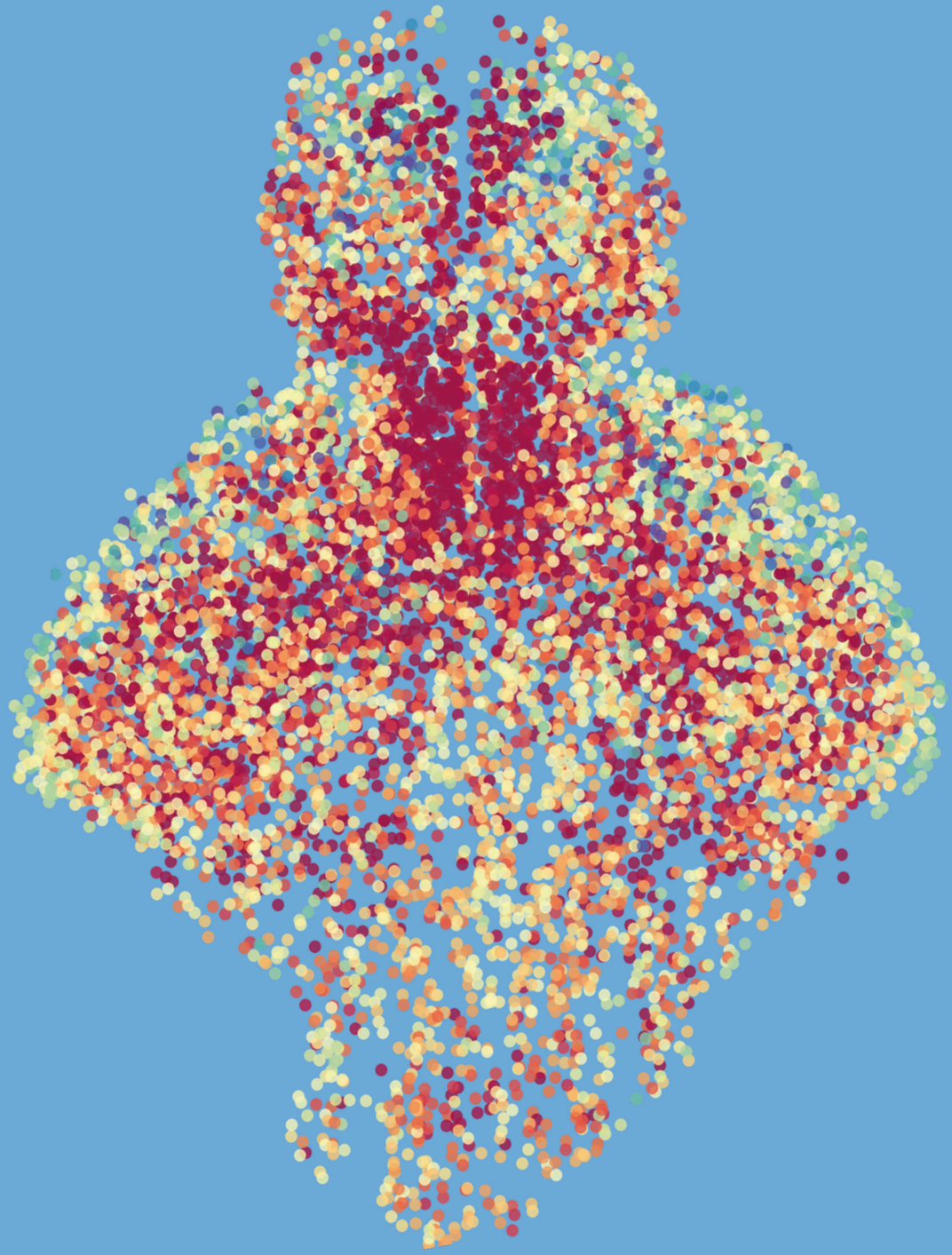
Identifying Celltypes Using Single Cell RNAseq

Dominic Burrows, PhD

UCSD Cog Sci

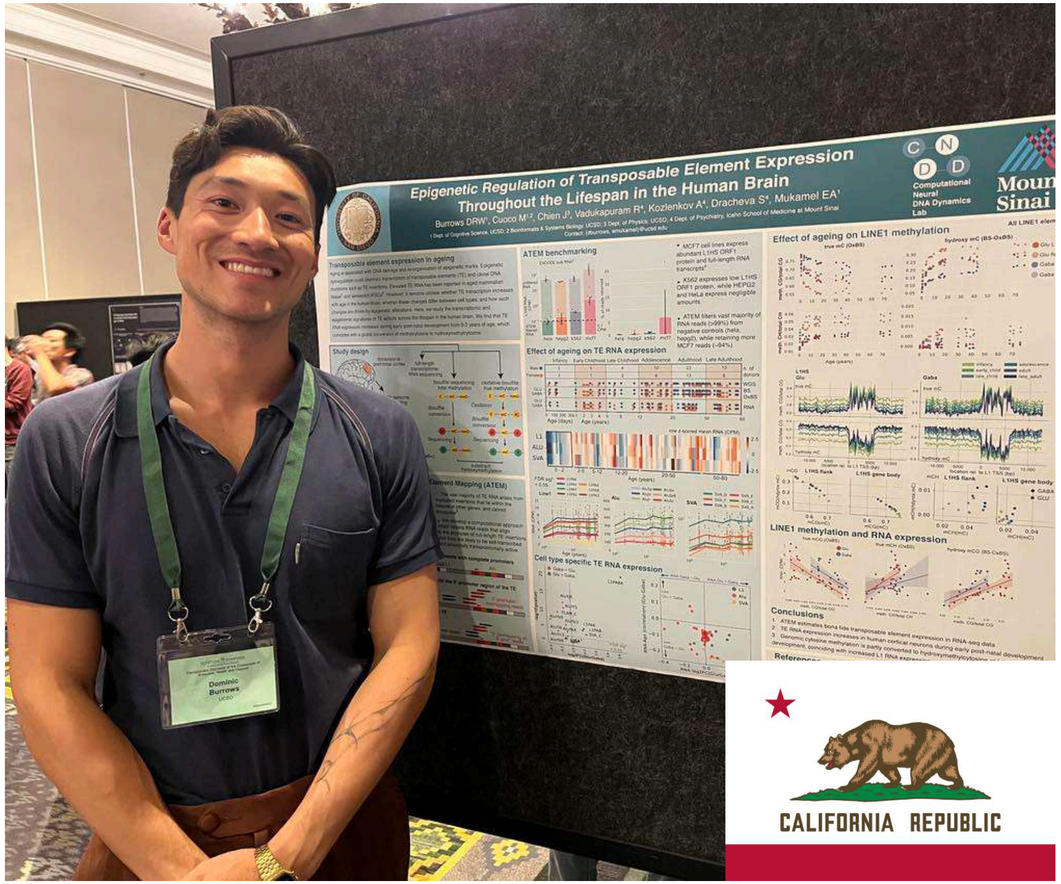
dburrows@ucsd.edu

@dominicburrows_



Hi, I’m Dominic!

PhD in Computational & Systems Neuroscience at King’s College London



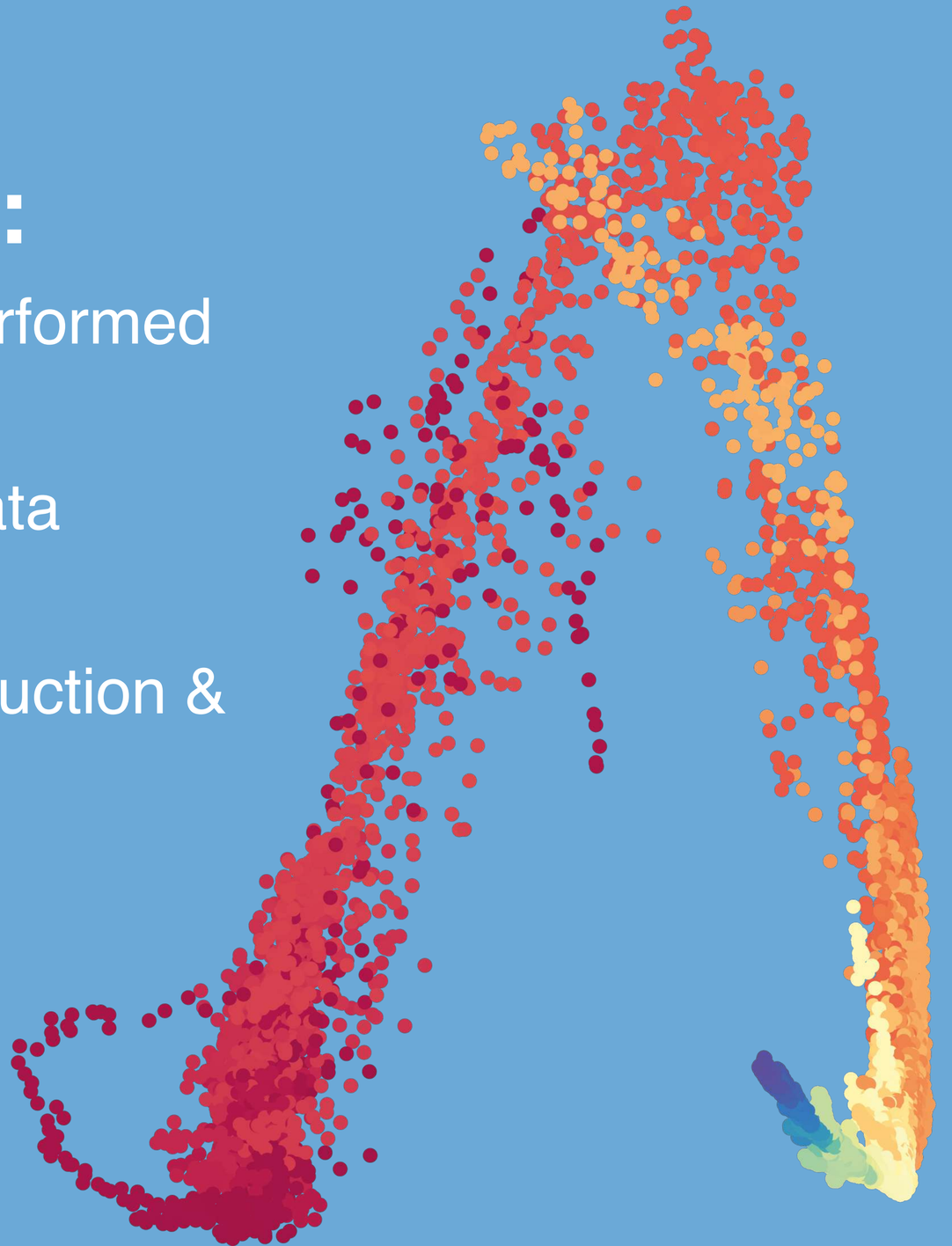
Lecturer in Data Science & Postdoctoral scholar in Computational Neuroscience at UCSD

Today’s learning objectives:

- Understand how single cell RNA seq is performed

- Understand how to process scRNA seq data

- Understand how to use dimensionality reduction & clustering to identify celltypes

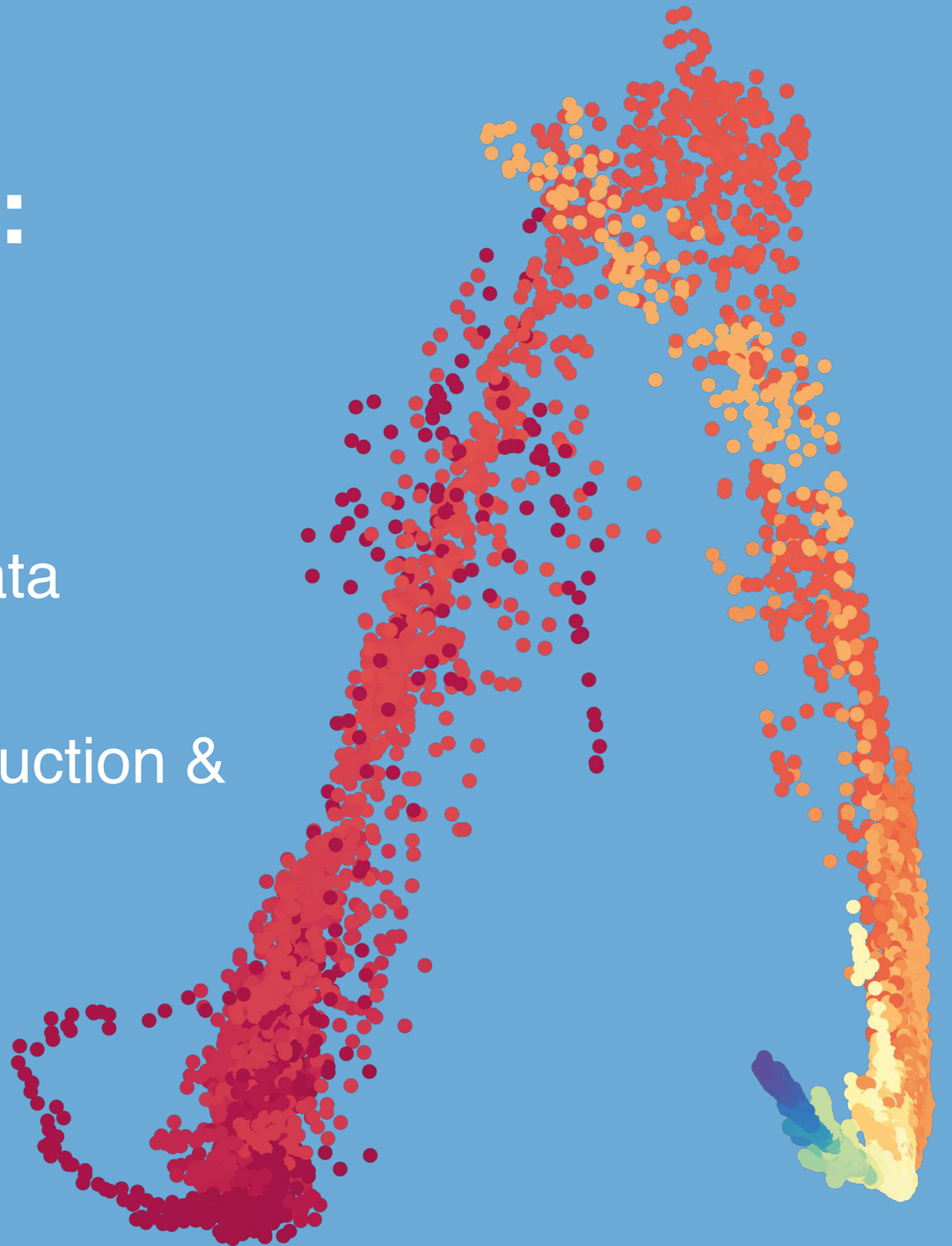


Today’s learning objectives:

- **Understand how single cell RNA seq is performed**

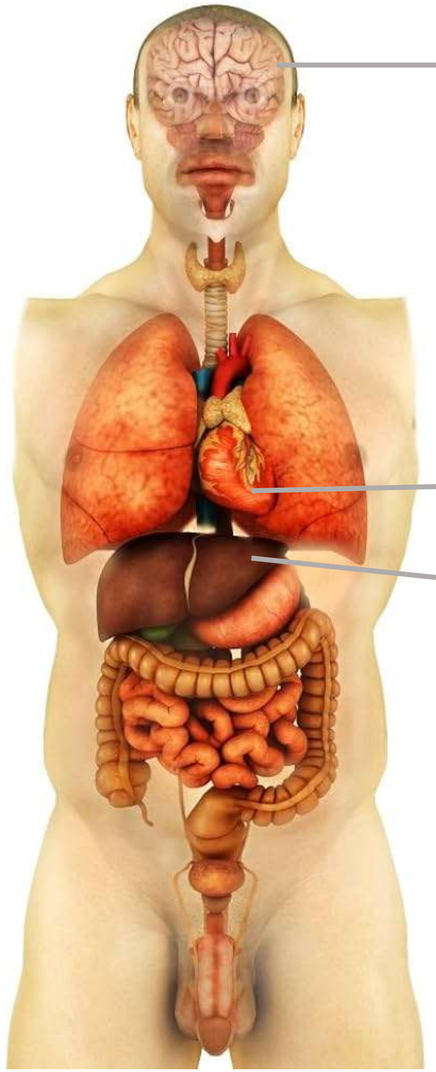
- Understand how to process scRNA seq data

- Understand how to use dimensionality reduction & clustering to identify celltypes

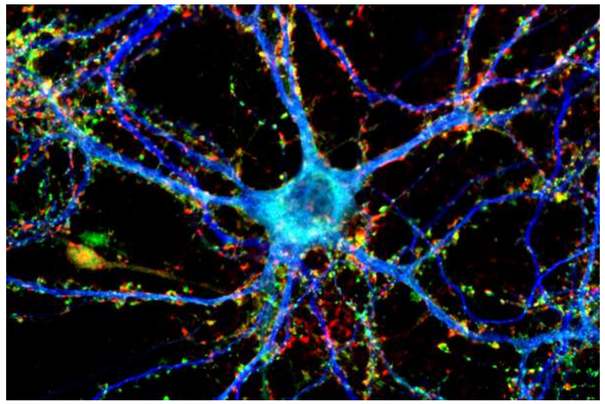


Setup coding environment!

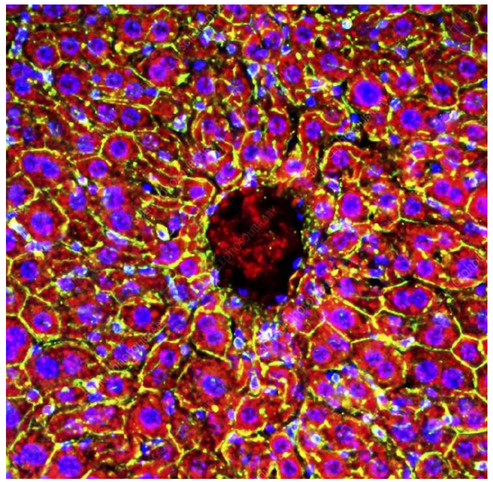
What are you made of?



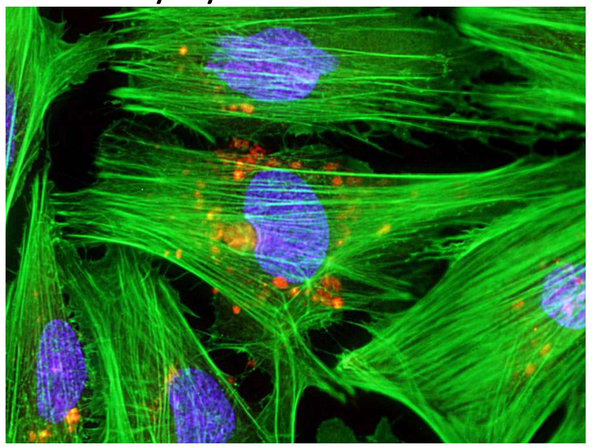
Neurons



Liver endothelial cells



Cardiomyocytes



We are made up of ~10^{14} cells, coming from approximately 200 distinct celltypes

What are you made of?

Me, 31 years ago

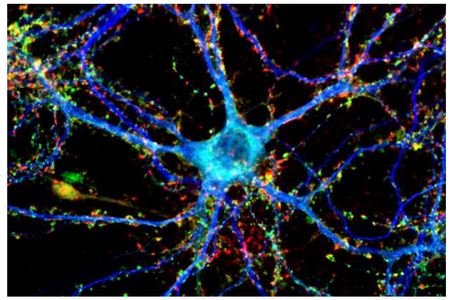
We are made up of ~10^{14} cells, coming from approximately 200 distinct celltypes

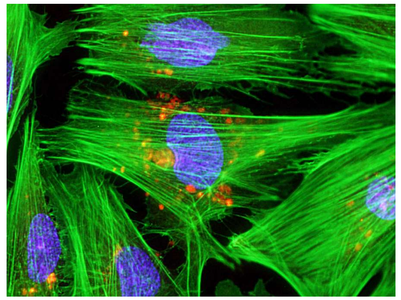
But we start life as only 1 cell!

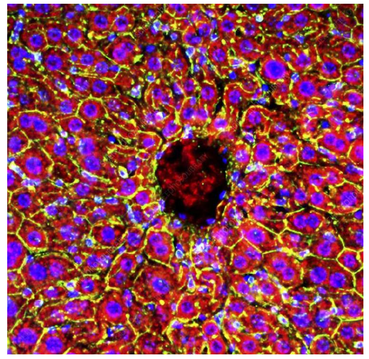
What are you made of?

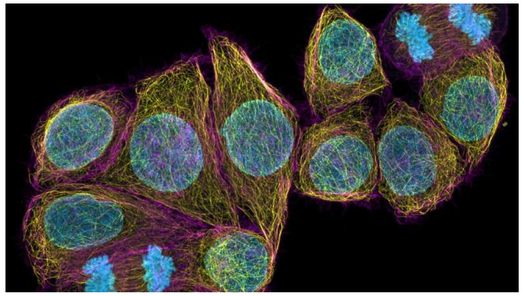
Me, 31 years ago

Me today









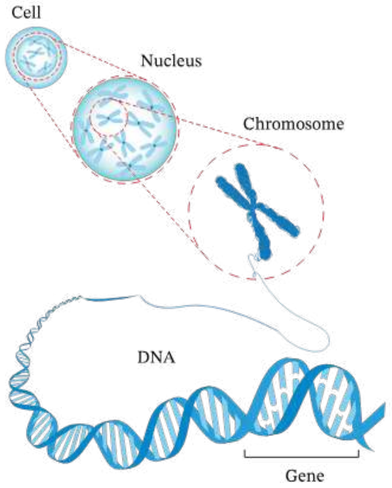
…et al.

We are made up of ~10^{14} cells, coming from approximately 200 distinct celltypes

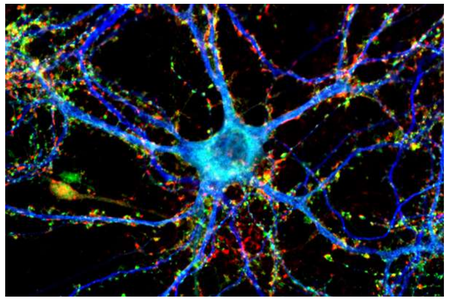
But we start life as only 1 cell!

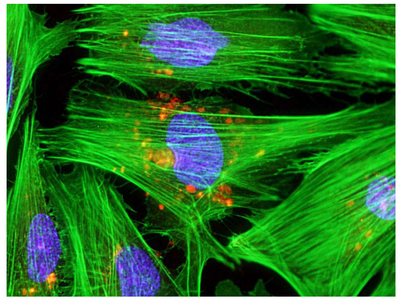
What are you made of?

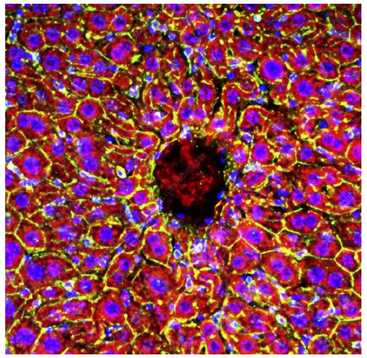
Me, 31 years ago

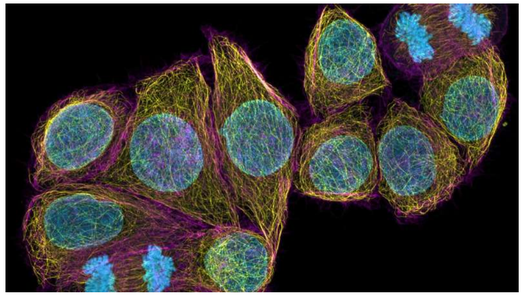


Me today









…et al.

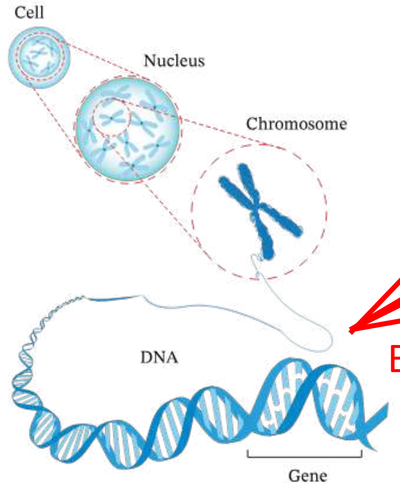
All the information I need to make every single celltype is in the DNA I have when I am 1 cell old

We are made up of ~10^{14} cells, coming from approximately 200 distinct celltypes

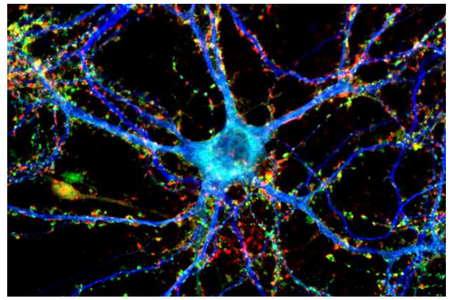
But we start life as only 1 cell!

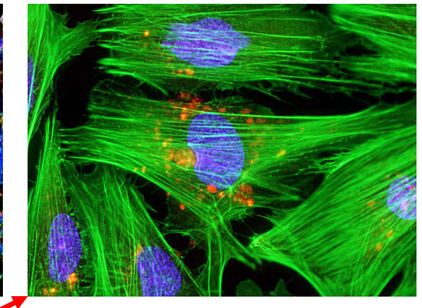
What are you made of?

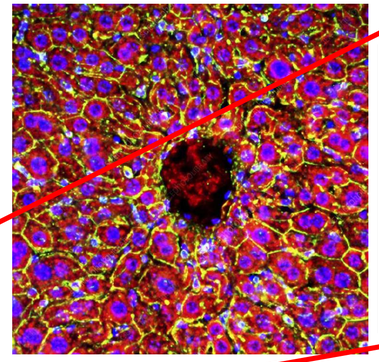
Me, 31 years ago

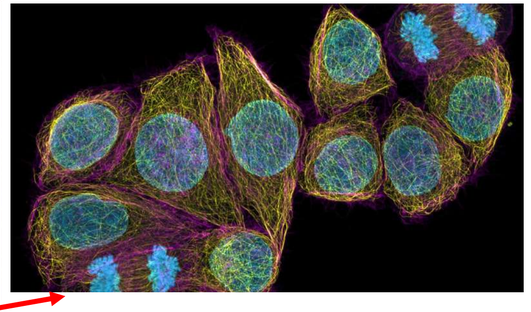


Me today









...et al.

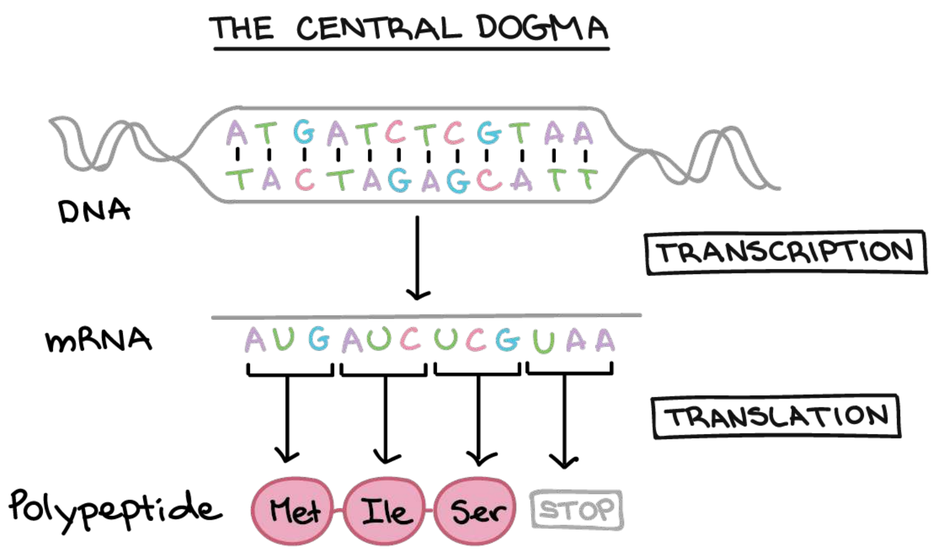
But how do we get from this

To **all** these?

We are made up of ~10^{14} cells, coming from approximately 200 distinct celltypes

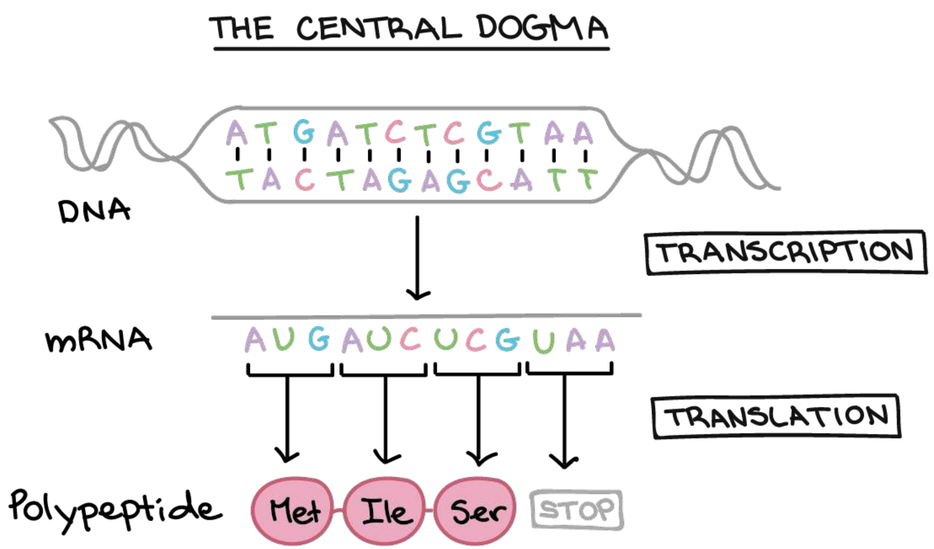
But we start life as only 1 cell!

The central dogma

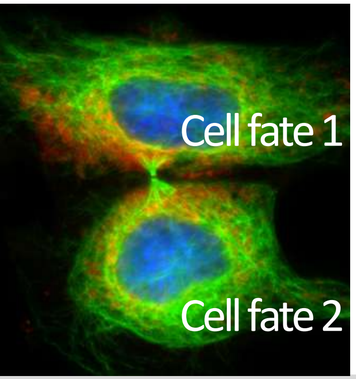


DNA begets RNA
RNA begets Protein

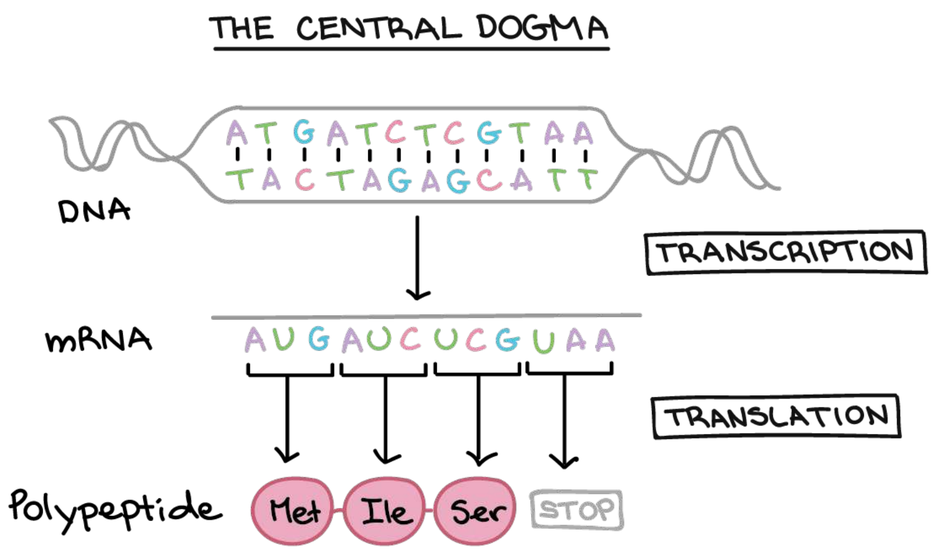
The central dogma



DNA begets RNA
RNA begets Protein

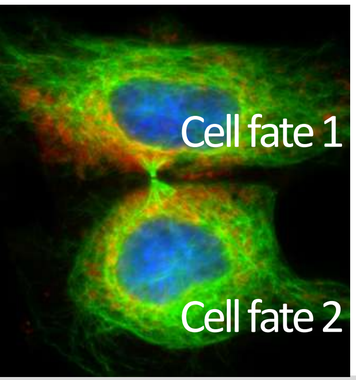


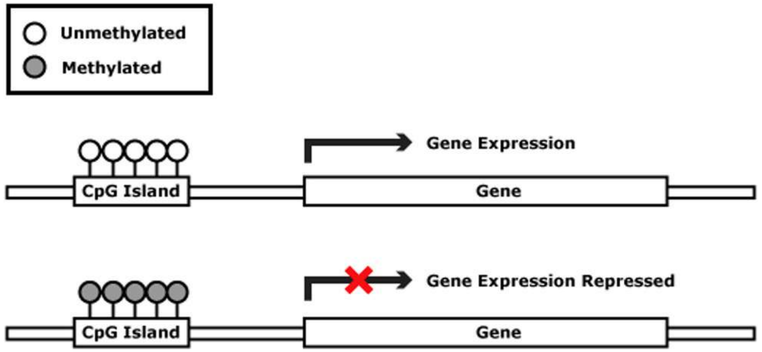
The central dogma



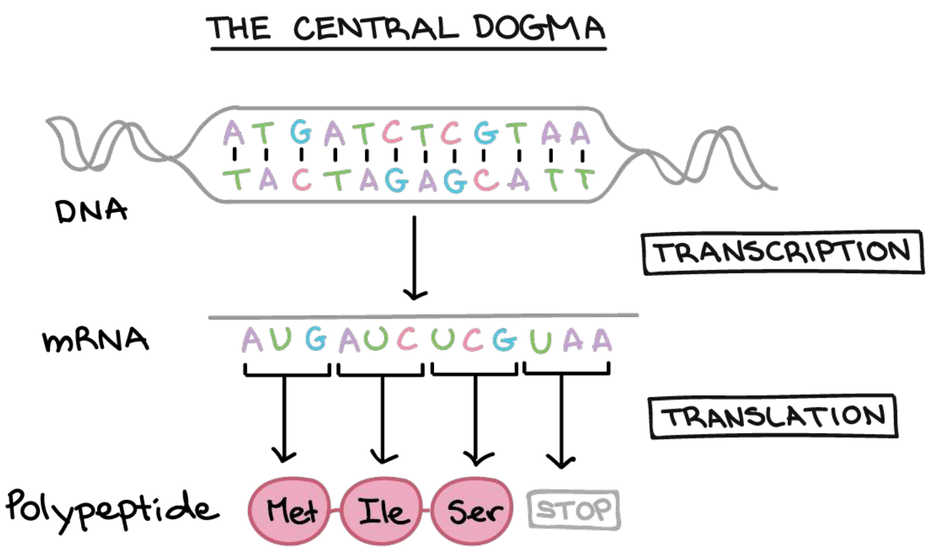
DNA begets RNA
RNA begets Protein

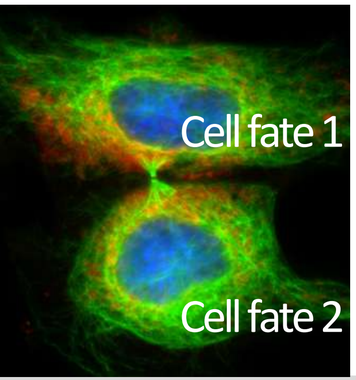
Epigenetic marks give rise to different genes expressed

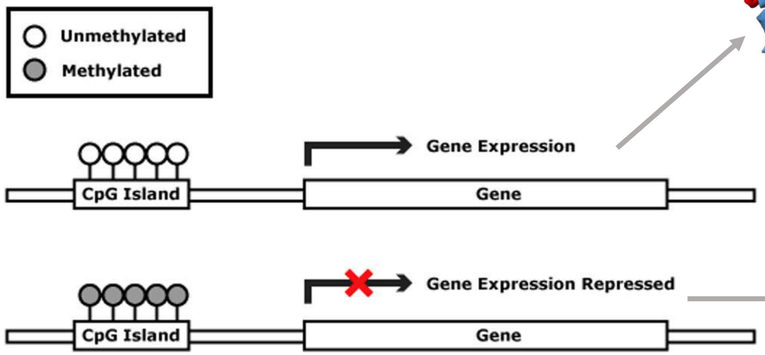


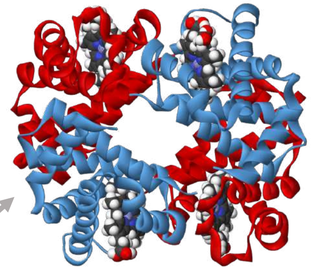


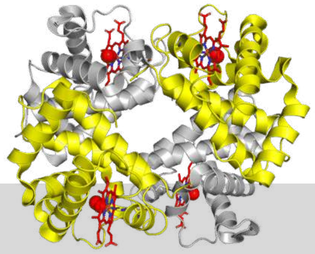
The central dogma







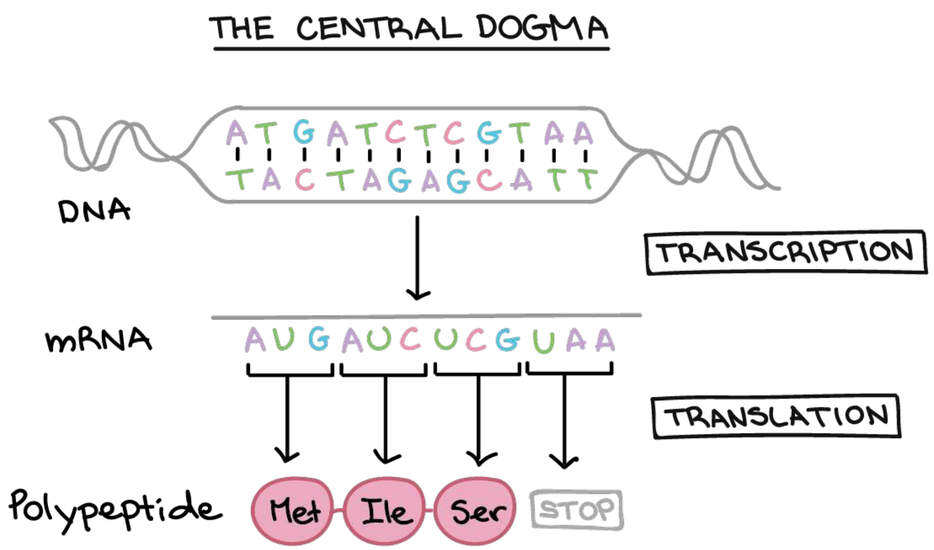


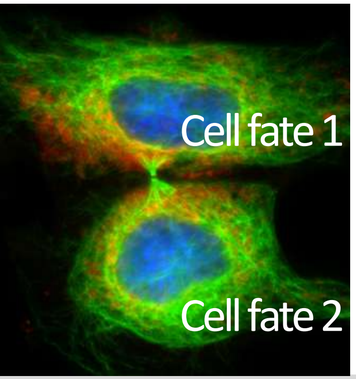


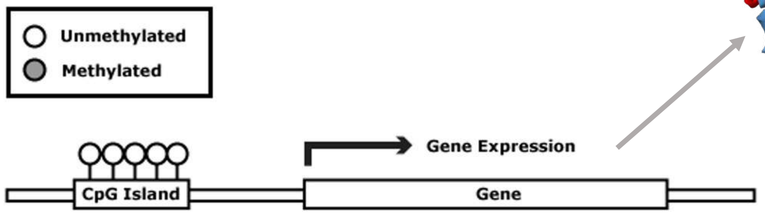
DNA begets RNA
RNA begets Protein

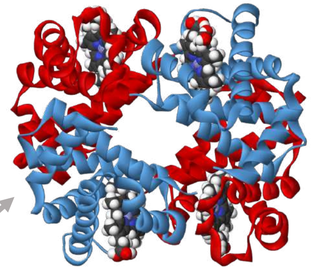
Epigenetic marks give rise to different genes expressed, and different proteins

The central dogma

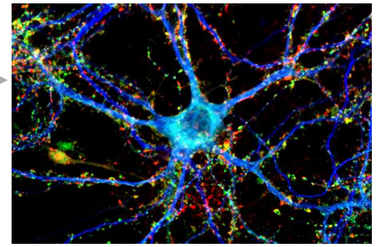


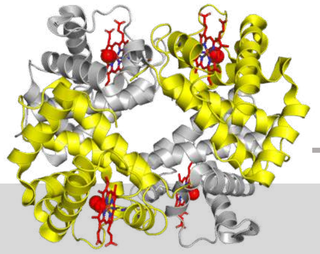




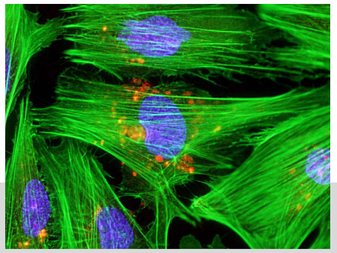


Celltype 1





Celltype 2

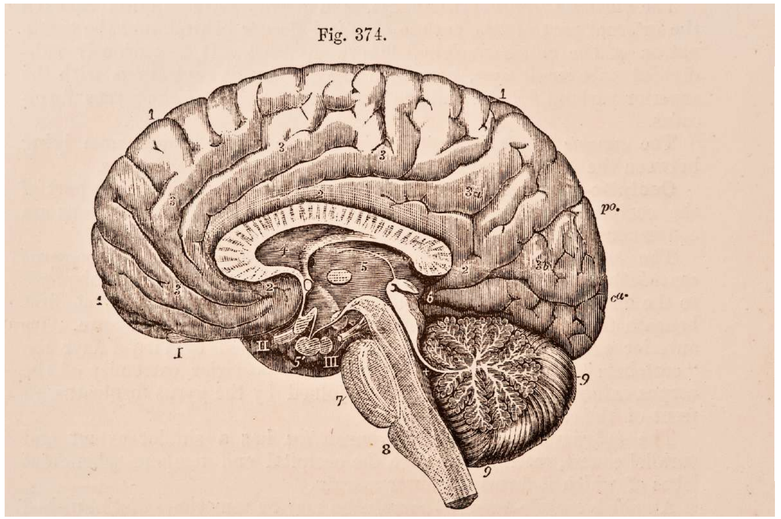


DNA begets RNA
RNA begets Protein

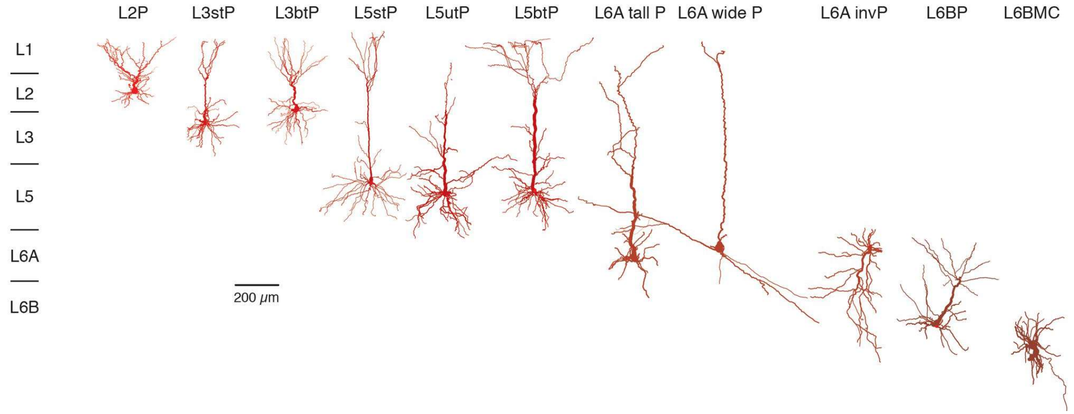
Epigenetic marks give rise to different genes expressed, and different proteins

This is how we get different celltypes!

Why single cells matter…

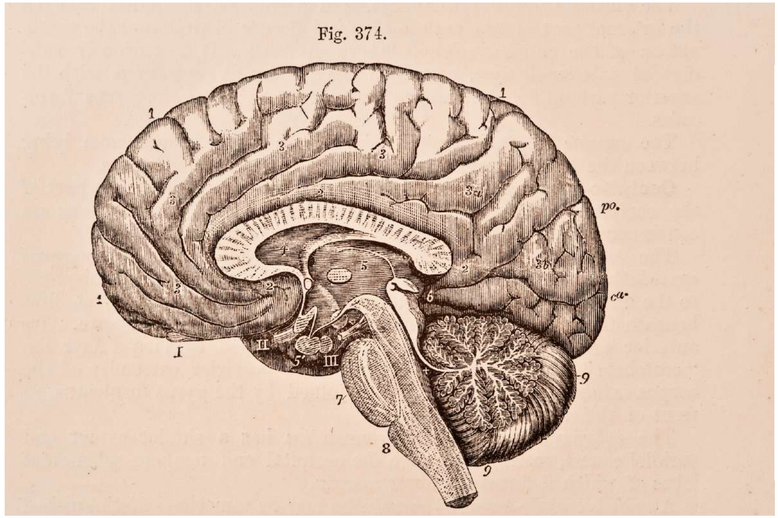


Different celltypes serve different functions!

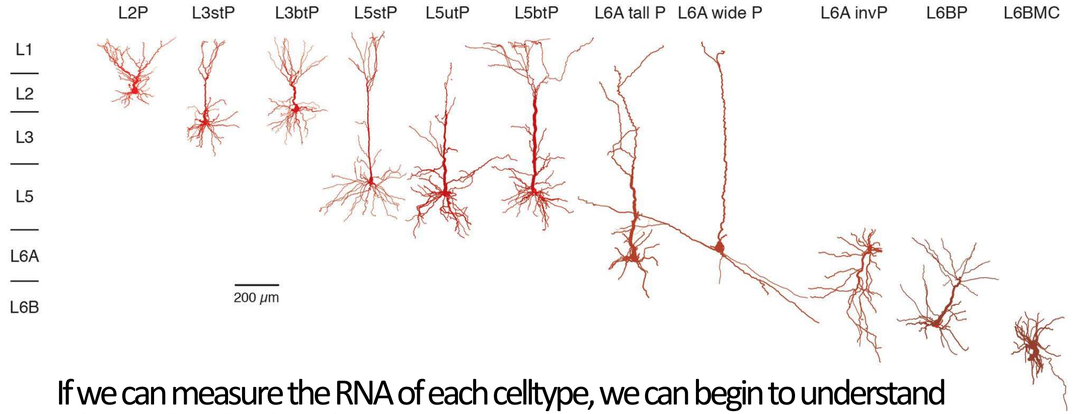


We know surprisingly little about the function of certain cells

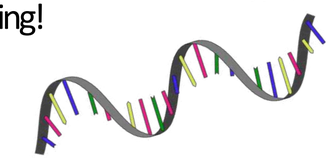
Why single cells matter…



Different celltypes serve different functions!

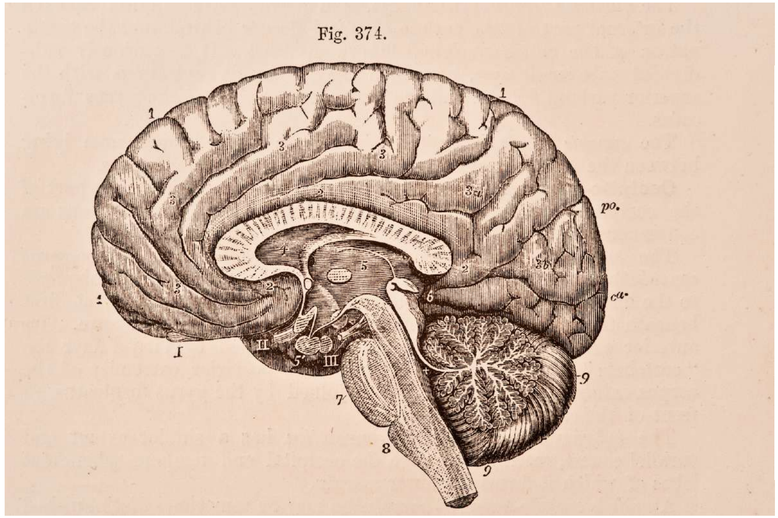


If we can measure the RNA of each celltype, we can begin to understand what each cell is doing!

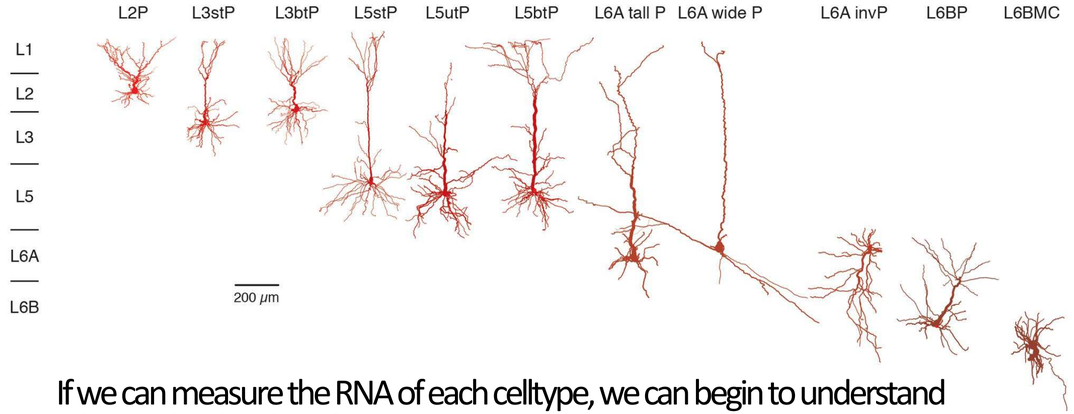


We know surprisingly little about the function of certain cells

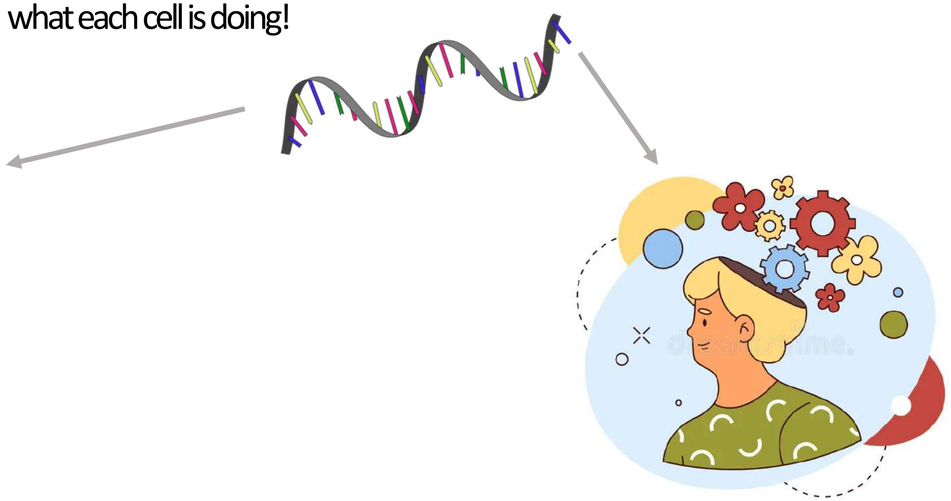
Why single cells matter…



Different celltypes serve different functions!

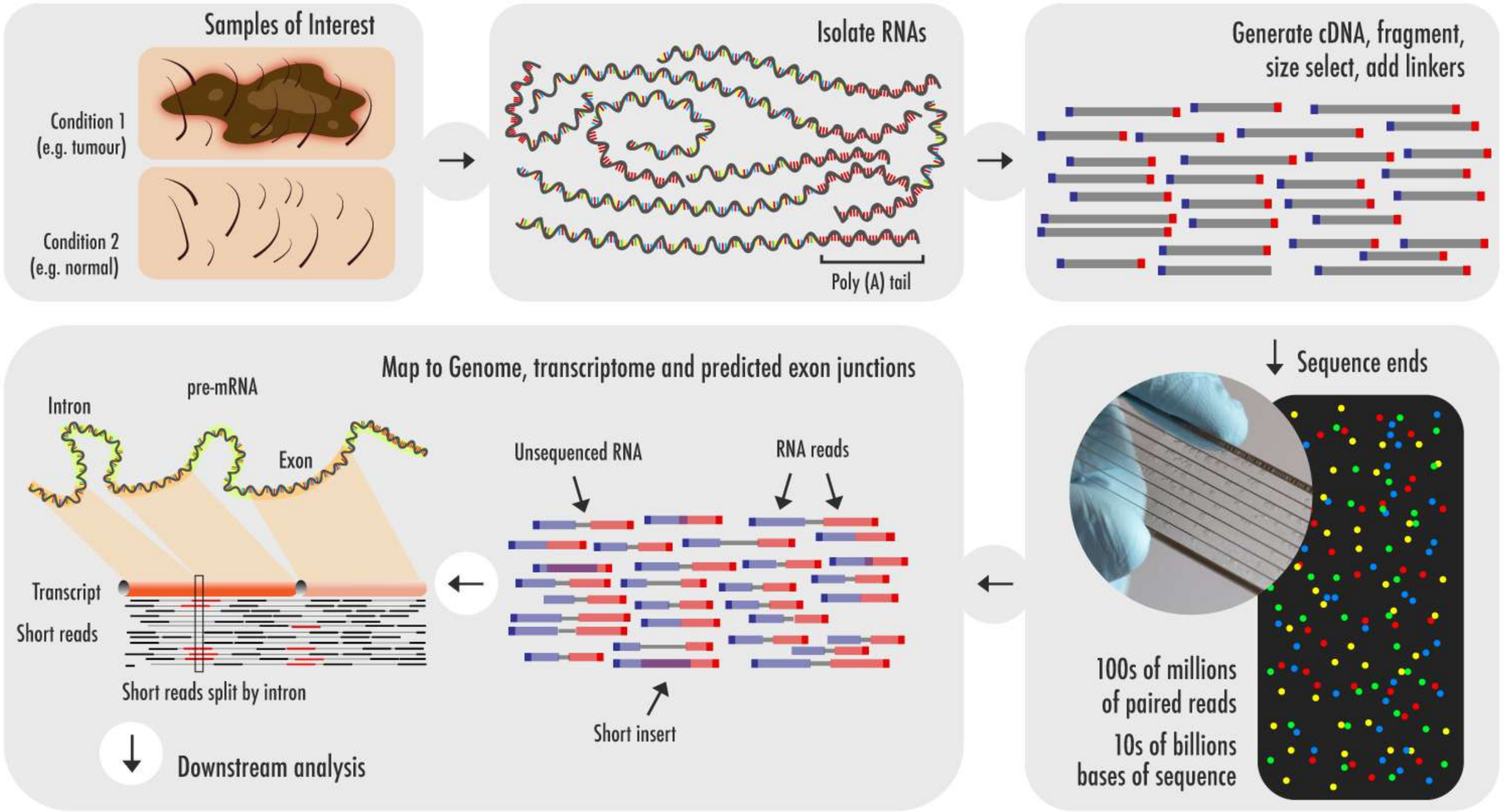


If we can measure the RNA of each celltype, we can begin to understand what each cell is doing!



We know surprisingly little about the function of certain cells

RNA sequencing

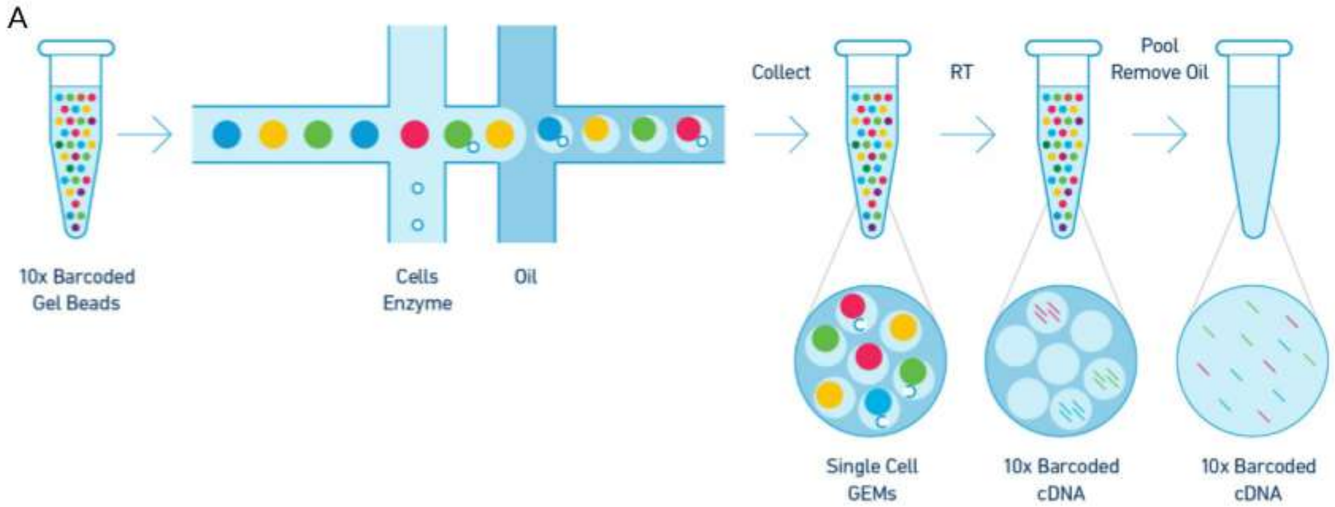


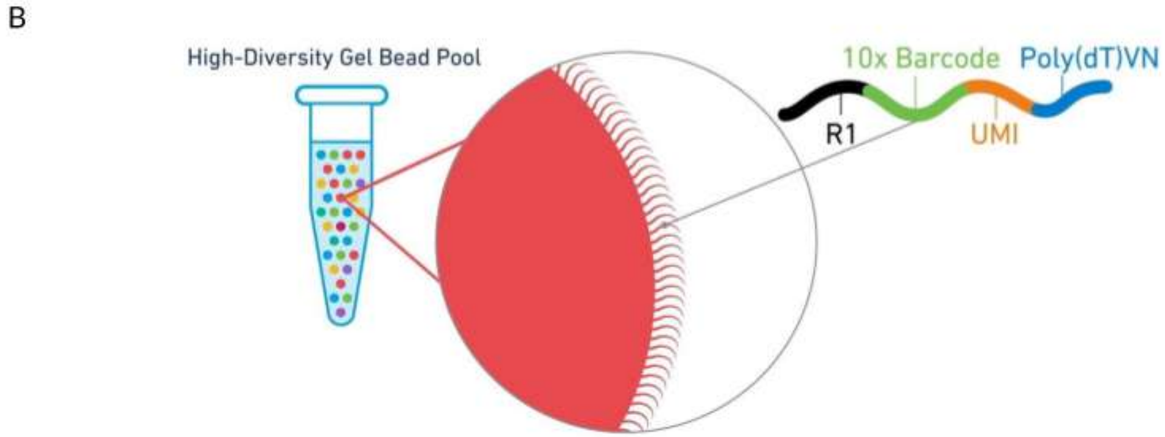
For the best explanation of how RNAseq is done, check out Illumina's video:

https://www.youtube.com/watch?v=fCd6B5HRaZ8

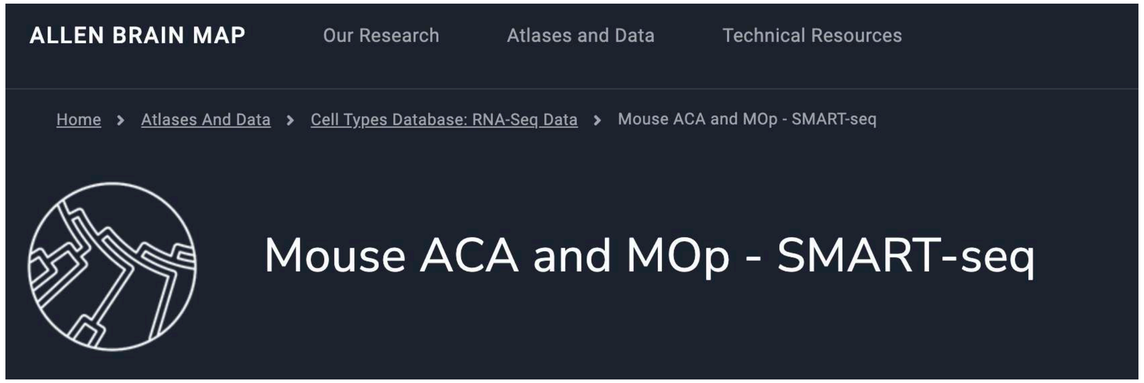
https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

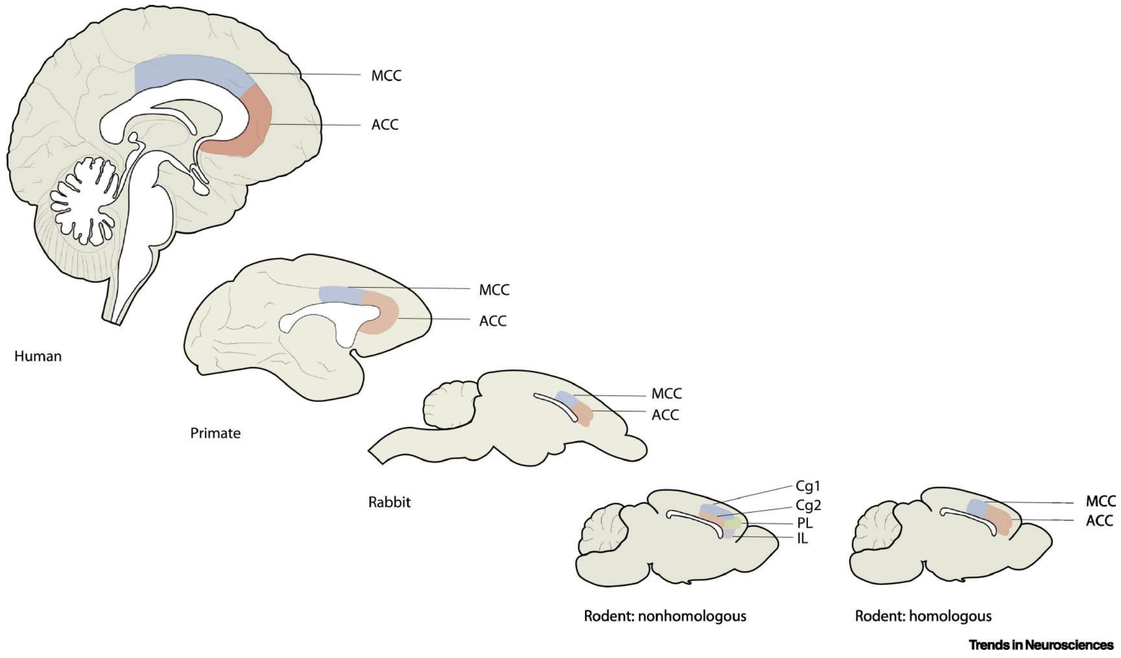
Single cell RNA sequencing





The data





The ACC is involved in decision making, error monitoring, goal directed behavior, emotional regulation

Start by loading in the data!

https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-aca-and-mop-smart-seq

The data

These are our genes

index	0610005C13Rik	0610006L08Rik	0610007P14Rik	0610009B22Rik	0610009E02Rik
SM-DD44B_S81_E1-50	0	0	0	912	0
SM-DD44B_S82_E1-50	0	0	0	0	0
SM-DD44B_S83_E1-50	0	0	0	1507	0
SM-DD44B_S84_E1-50	0	0	344	0	0
SM-DD44B_S85_E1-50	0	0	143	103	140
SM-DD44B_S86_E1-50	0	0	345	195	0
SM-DD44B_S87_E1-50	0	0	551	12	0
SM-DD44B_S88_E1-50	0	0	3	913	0
SM-DD44B_S89_E1-50	0	0	0	15	137
SM-DD44B_S90_E1-50	0	0	0	1	0
SM-DD44B_S91_E1-50	0	0	309	0	0
SM-DD44B_S92_E1-50	0	0	313	446	0
SM-DD44B_S93_E1-50	0	0	347	0	0
SM-DD44B_S94_E1-50	0	0	75	191	0
SM-DD44B_S95_E1-50	0	0	0	1	0

These are our cells

Each entry tells us how many RNA fragments mapped to that gene, for each cell!

The ACC is involved in decision making, error monitoring, goal directed behavior, emotional regulation

The data consists of:
5,028 cells
31,995 genes

https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-aca-and-mop-smart-seq

The data

Our goal: can we use this data to identify different celltypes in the brain?

index	0610005C13Rik	0610006L08Rik	0610007P14Rik	0610009B22Rik	0610009E02Rik
SM-DD44B_S81_E1-50	0	0	0	912	0
SM-DD44B_S82_E1-50	0	0	0	0	0
SM-DD44B_S83_E1-50	0	0	0	1507	0
SM-DD44B_S84_E1-50	0	0	344	0	0
SM-DD44B_S85_E1-50	0	0	143	103	140
SM-DD44B_S86_E1-50	0	0	345	195	0
SM-DD44B_S87_E1-50	0	0	551	12	0
SM-DD44B_S88_E1-50	0	0	3	913	0
SM-DD44B_S89_E1-50	0	0	0	15	137
SM-DD44B_S90_E1-50	0	0	0	1	0
SM-DD44B_S91_E1-50	0	0	309	0	0
SM-DD44B_S92_E1-50	0	0	313	446	0
SM-DD44B_S93_E1-50	0	0	347	0	0
SM-DD44B_S94_E1-50	0	0	75	191	0
SM-DD44B_S95_E1-50	0	0	0	1	0

The ACC is involved in decision making, error monitoring, goal directed behavior, emotional regulation

The data consists of:
5,028 cells
31,995 genes

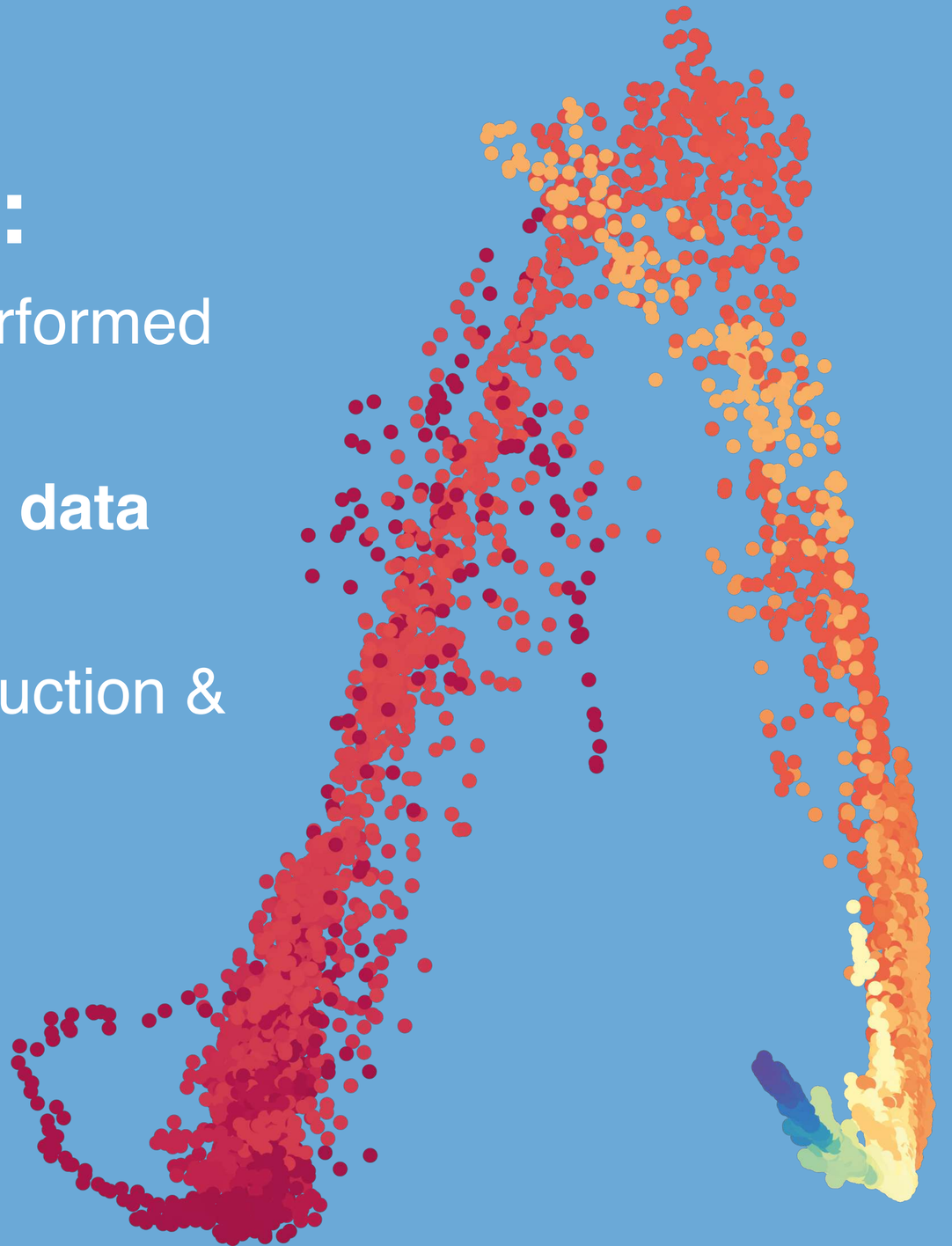
https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-aca-and-mop-smart-seq

Today’s learning objectives:

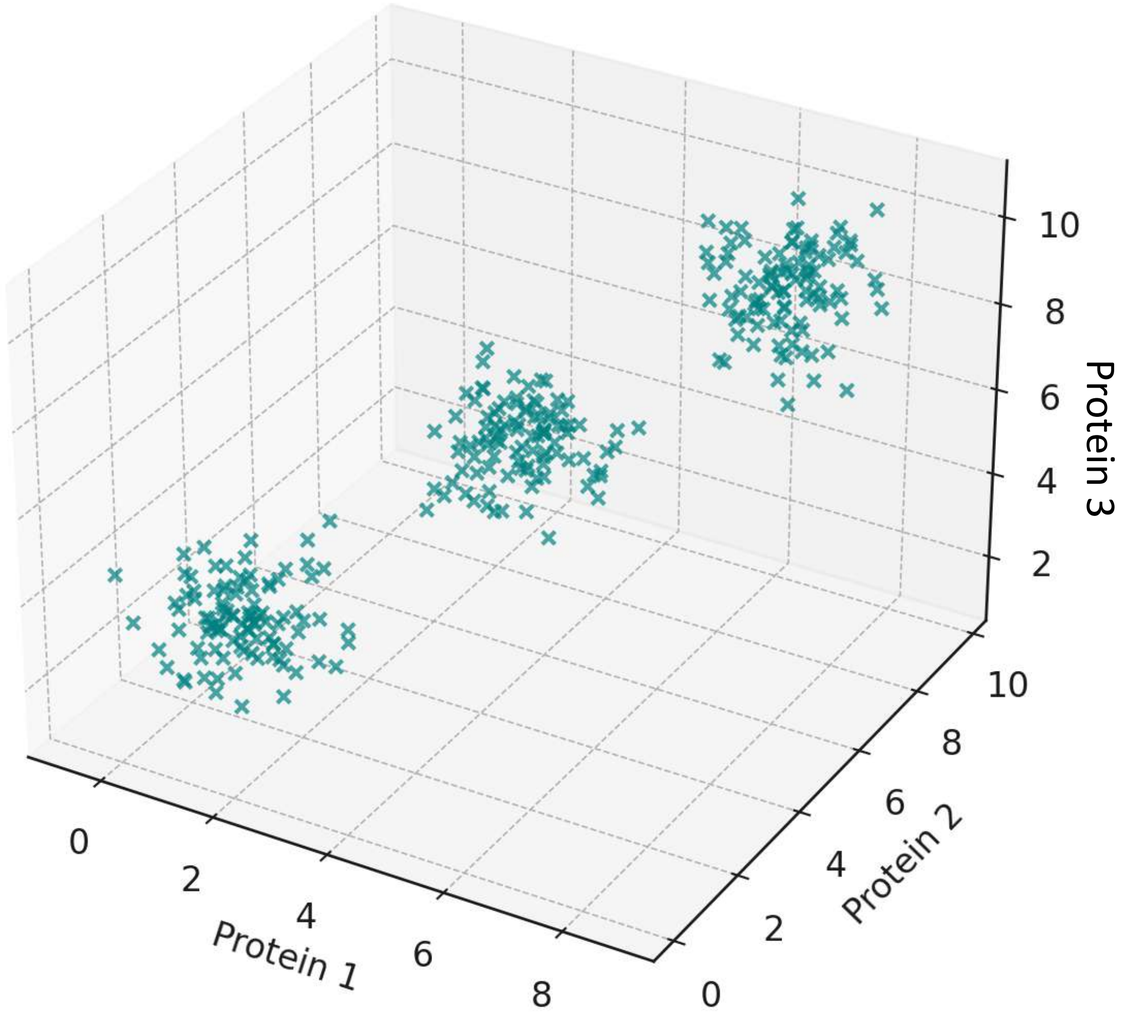
- Understand how single cell RNA seq is performed

- **Understand how to process scRNA seq data**

- Understand how to use dimensionality reduction & clustering to identify celltypes



The curse of dimensionality

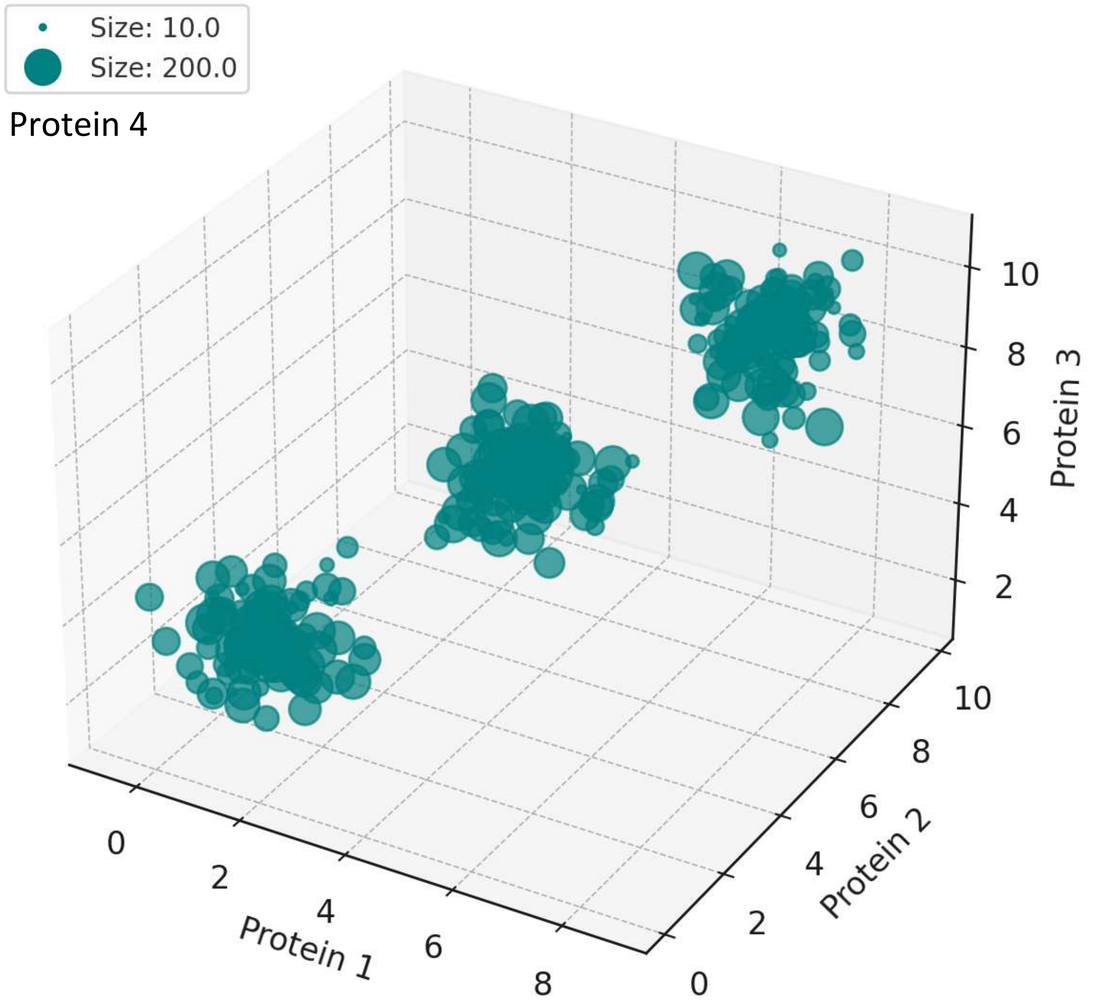


This is our feature space

Each position in this space has a coordinate
e.g. 3d space has 3 coordinates x, y and z

But instead of corresponding to physical location
x, y and z correspond to P1, P2 and P3

The curse of dimensionality



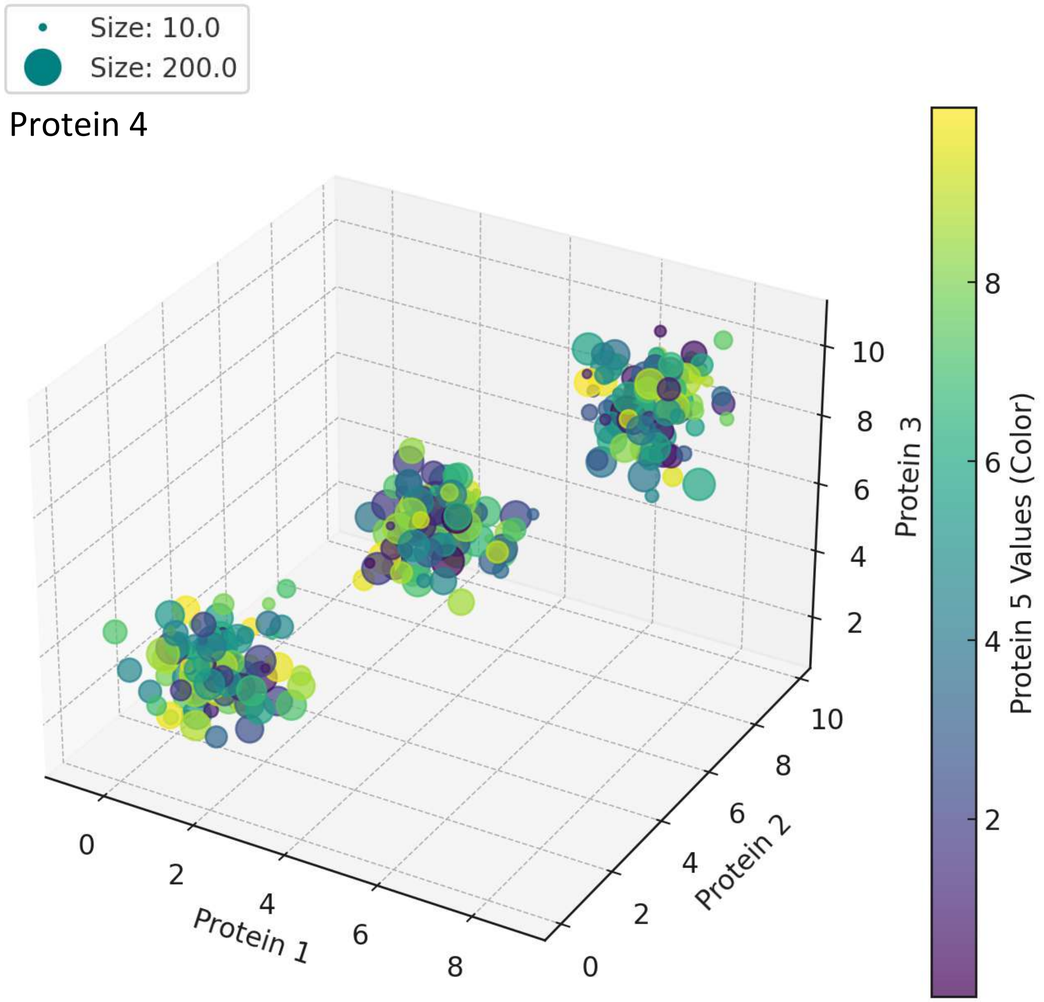
This is our feature space

Each position in this space has a coordinate
e.g. 4d space has 4 coordinates x, y, z, and size

But instead of corresponding to physical location
x, y, z and size correspond to P1, P2, P3 P4

What about protein 5?

The curse of dimensionality

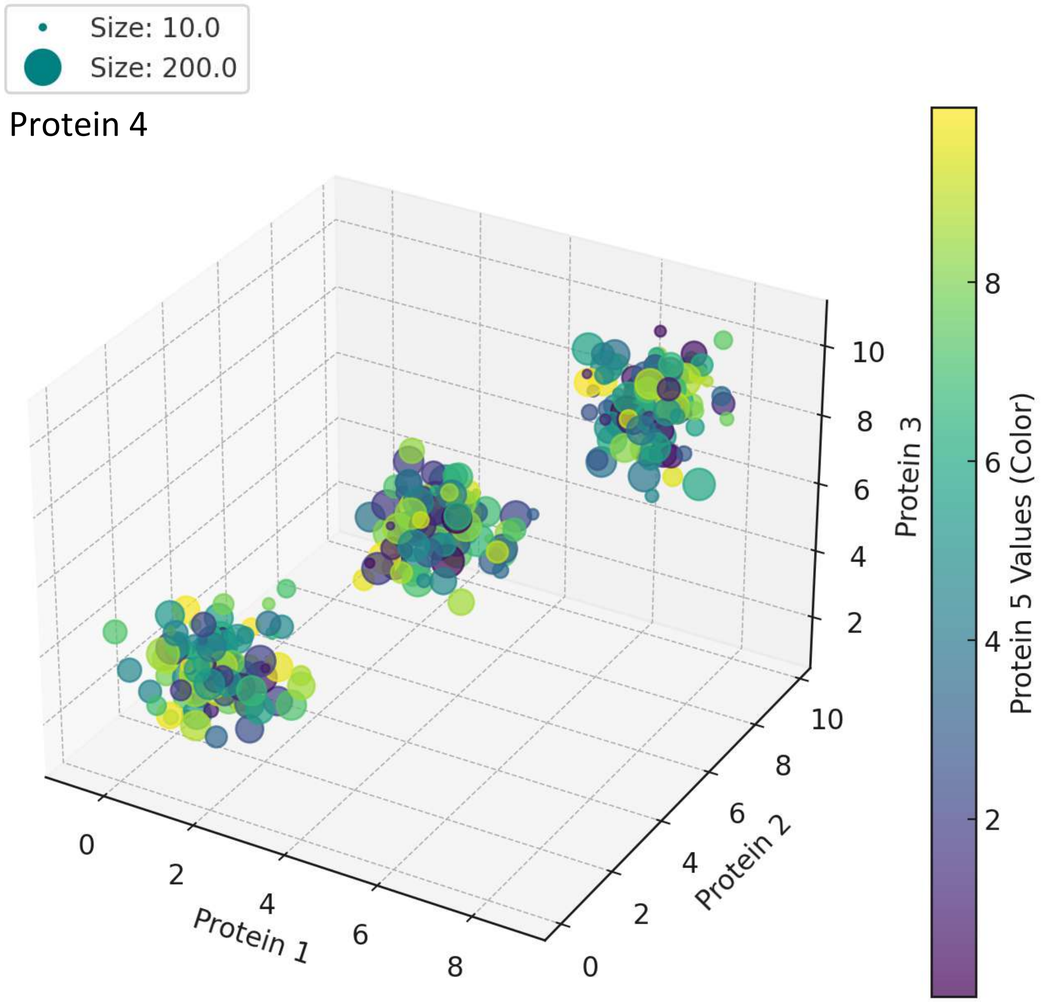


This is our feature space

Each position in this space has a coordinate
e.g. 5d space has 5 coordinates x, y, z, and size and color

But instead of corresponding to physical location
x, y, z, size and color correspond to P1, P2, P3, P4, P5

The curse of dimensionality



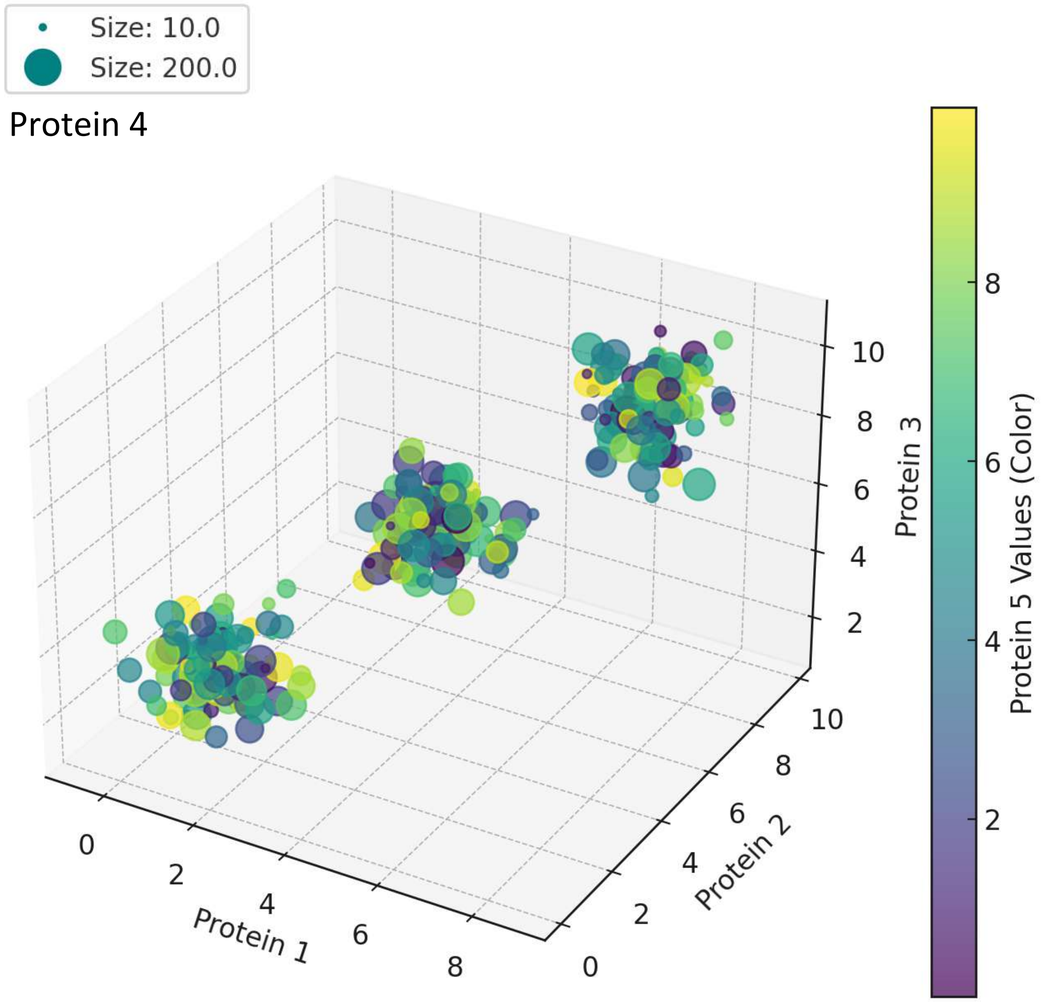
This is our feature space

Each position in this space has a coordinate
e.g. 5d space has 5 coordinates x, y, z, and size and color

But instead of corresponding to physical location
x, y, z, size and color correspond to P1, P2, P3, P4, P5

Around ~4-5 dimensions, things get very hard to visualize!
- visualising patterns in our data is central to identifying relationships, errors and refining hypotheses

The curse of dimensionality



This is our feature space

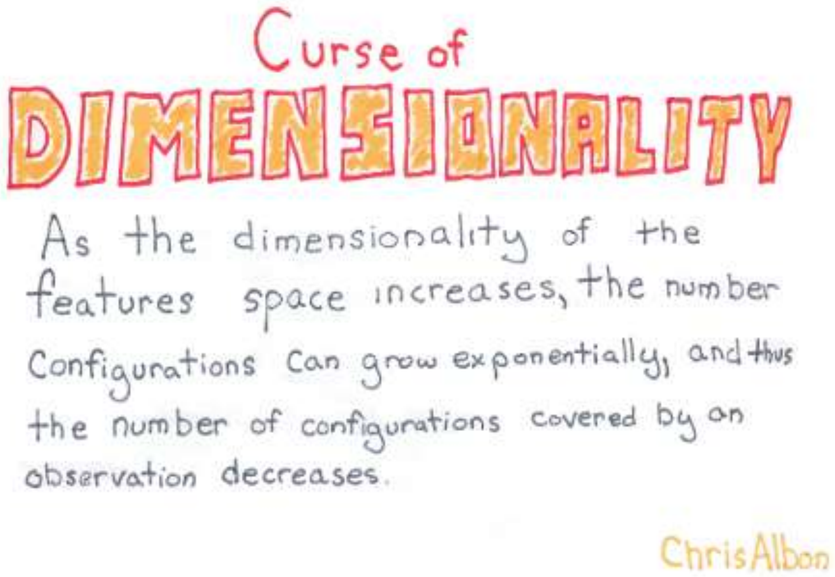
Each position in this space has a coordinate
e.g. 5d space has 5 coordinates x, y, z, and size and color

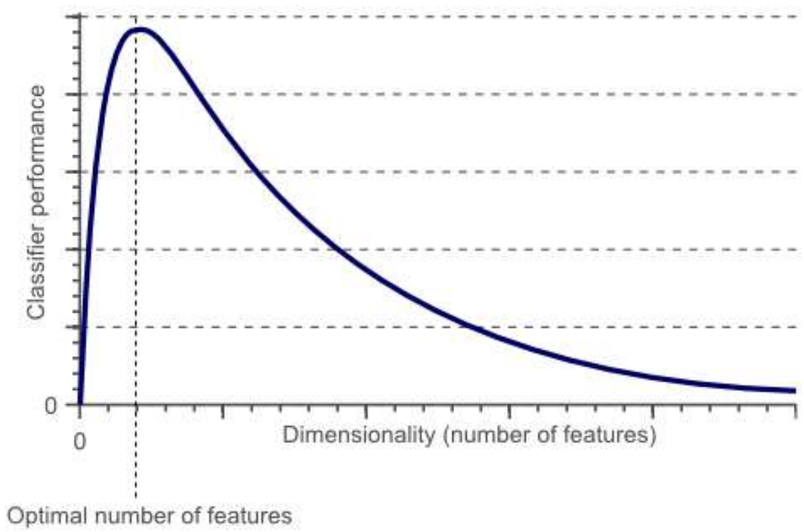
But instead of corresponding to physical location
x, y, z, size and color correspond to P1, P2, P3, P4, P5

Around ~4-5 dimensions, things get very hard to visualize!
- visualising patterns in our data is central to identifying relationships, errors and refining hypotheses

We still have >29,000 genes to visualise

The curse of dimensionality





Hughes phenomenon: as the number of features (genes) increases, a model's performance increases until an optimal number is reached…adding further features degrades the performance

Not only is high dimensional data hard to visualize, it is hard to learn from!

We want to reduce the number of features where possible

Normalising our data to estimate Counts per Million (CPM)

- Normalising makes sure our data is in a comparable scale

- The reaction inside each cell can take place with different efficiencies, the overall amplification of RNA can be different across cells

- Directly comparing genes across cells might lead us to falsely conclude there are differences in gene levels when really it is just due to differenes in the amplification reaction

- Therefore, we divide each gene by the sum of all counts for a cell - this ensures we are looking at relative gene expression, given the total amount of RNA in the cell

- We then multiple by a million to make the number in an easier to visualise range

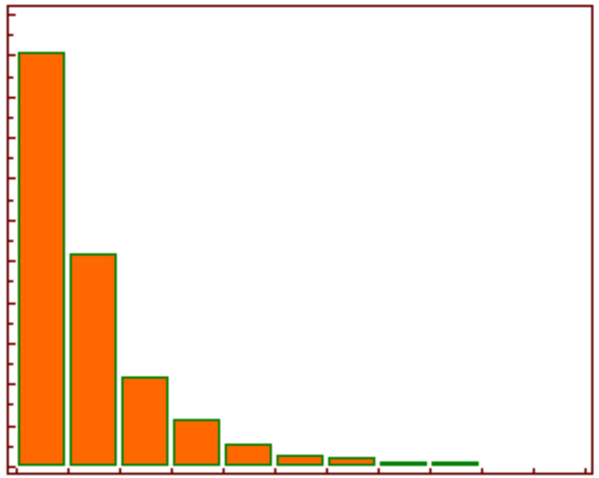
Log transforming our data

Genetic data is strongly skewed

Taking the log of the data ensures that we deal with powers that give rise to our CPM values, instead of the CPM values alone

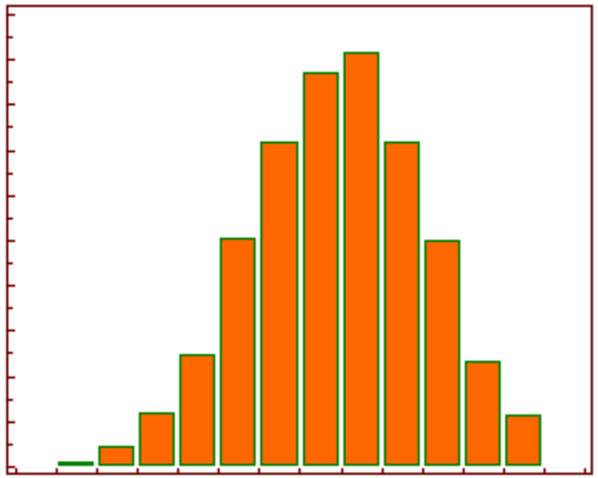
Because CPM values at smaller scales are encoded by further apart powers, compared with CPM values at larger scales, taking the log ensures that smaller values are more spread out!

Before log transform



CPM

After log transform



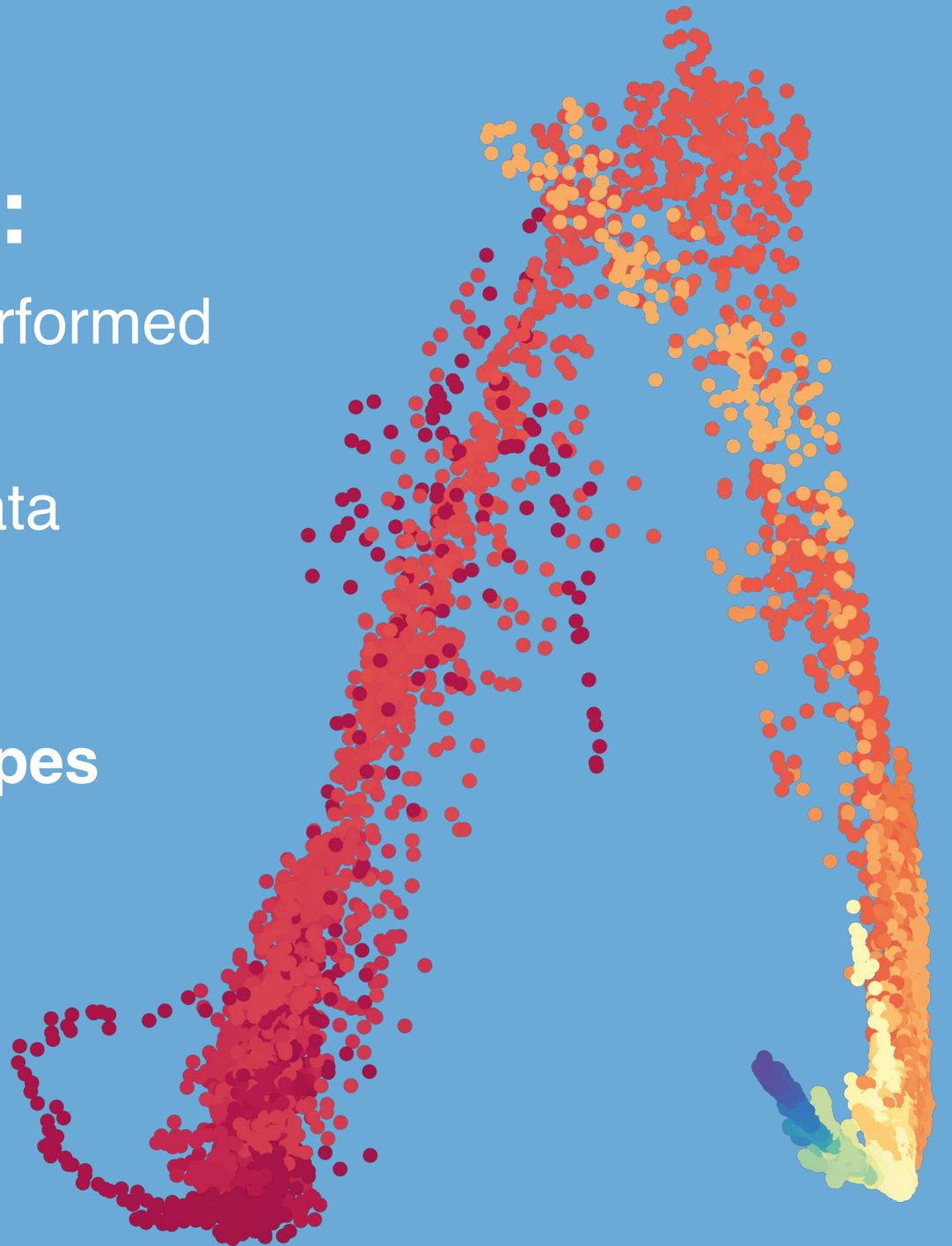
Log(CPM)

Today’s learning objectives:

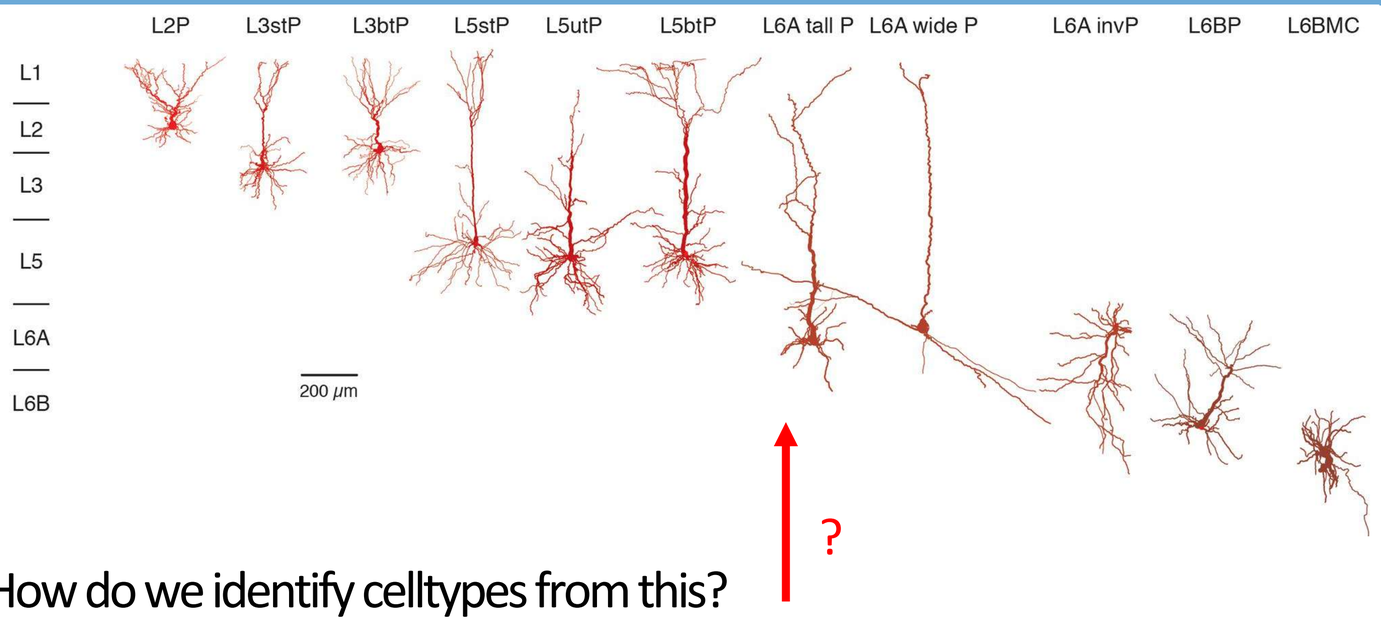
- Understand how single cell RNA seq is performed

- Understand how to process scRNA seq data

- **Understand how to use dimensionality reduction & clustering to identify celltypes**



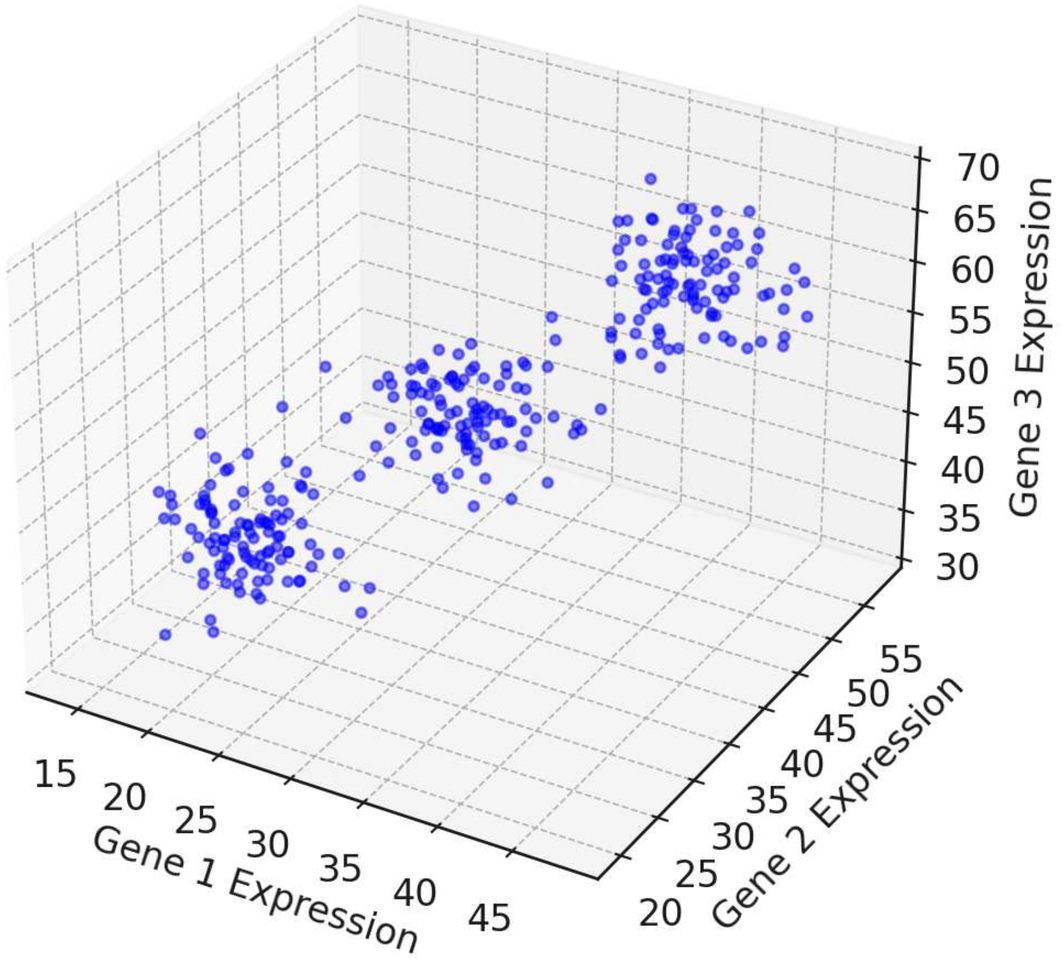
Identifying celltypes



How do we identify celltypes from this?

	0610007P14Rik	0610009B22Rik	0610009E02Rik	0610009L18Rik	0610009O20Rik	0610010F05Rik	0610010K14Rik	06
SM-DD44B_S81_E1-50	0.000000	7.822666	0.000000	0.000000	0.000000	8.033618	7.893425	
SM-DD44B_S82_E1-50	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
SM-DD44B_S83_E1-50	0.000000	8.324744	0.000000	0.000000	0.000000	0.000000	0.000000	
SM-DD44B_S84_E1-50	6.768148	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
SM-DD44B_S85_E1-50	5.891966	5.644897	11.071925	0.000000	0.000000	6.827164	5.497695	
...	...	...	...	...	...	...	...	
SM-GE671_S236_E1-50	5.043701	4.679514	0.000000	4.906608	6.125131	4.287863	4.808635	
SM-GE671_S237_E1-50	4.236753	5.441784	0.000000	4.504835	0.000000	4.660831	5.127973	

Identifying celltypes: Clustering



Imagine we had 3 genes only

We can look at this data and discern 3 different celltypes

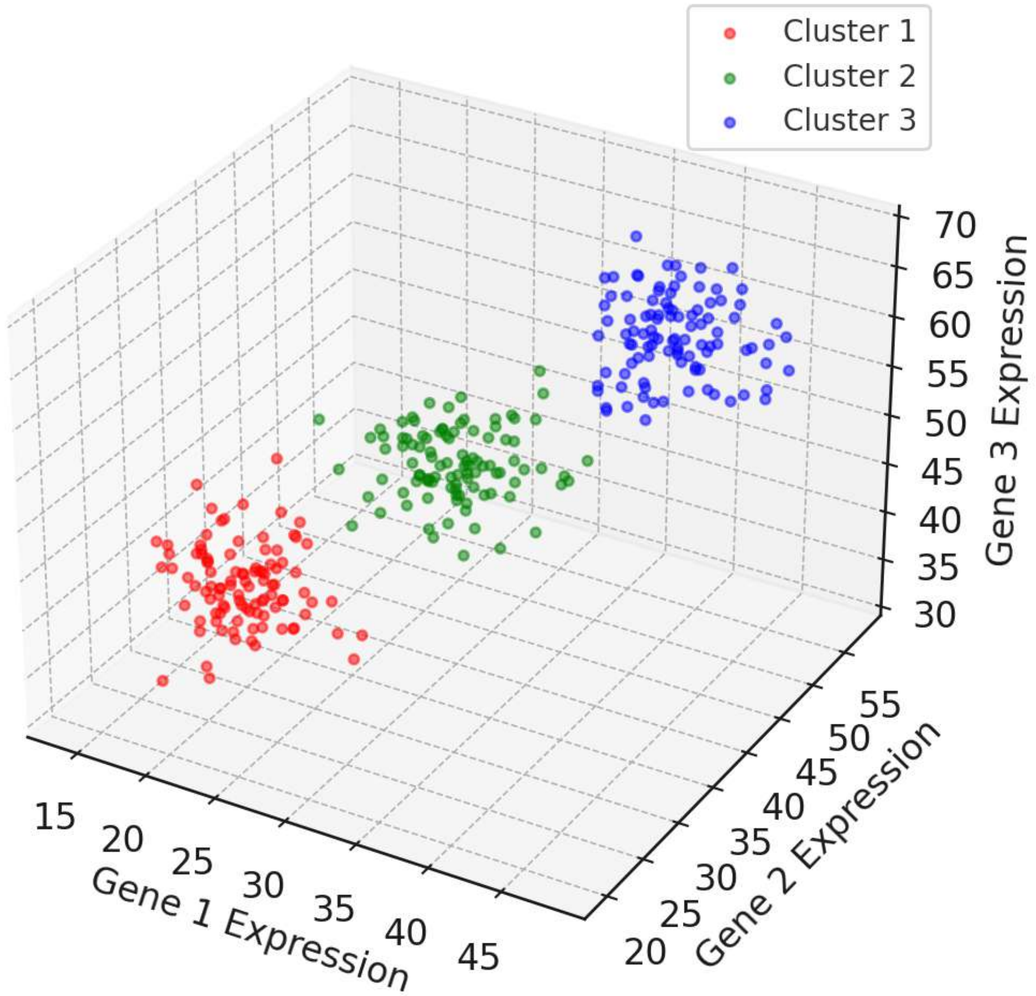
Why?

Each member of a group has similar gene expression to other members of that group

&

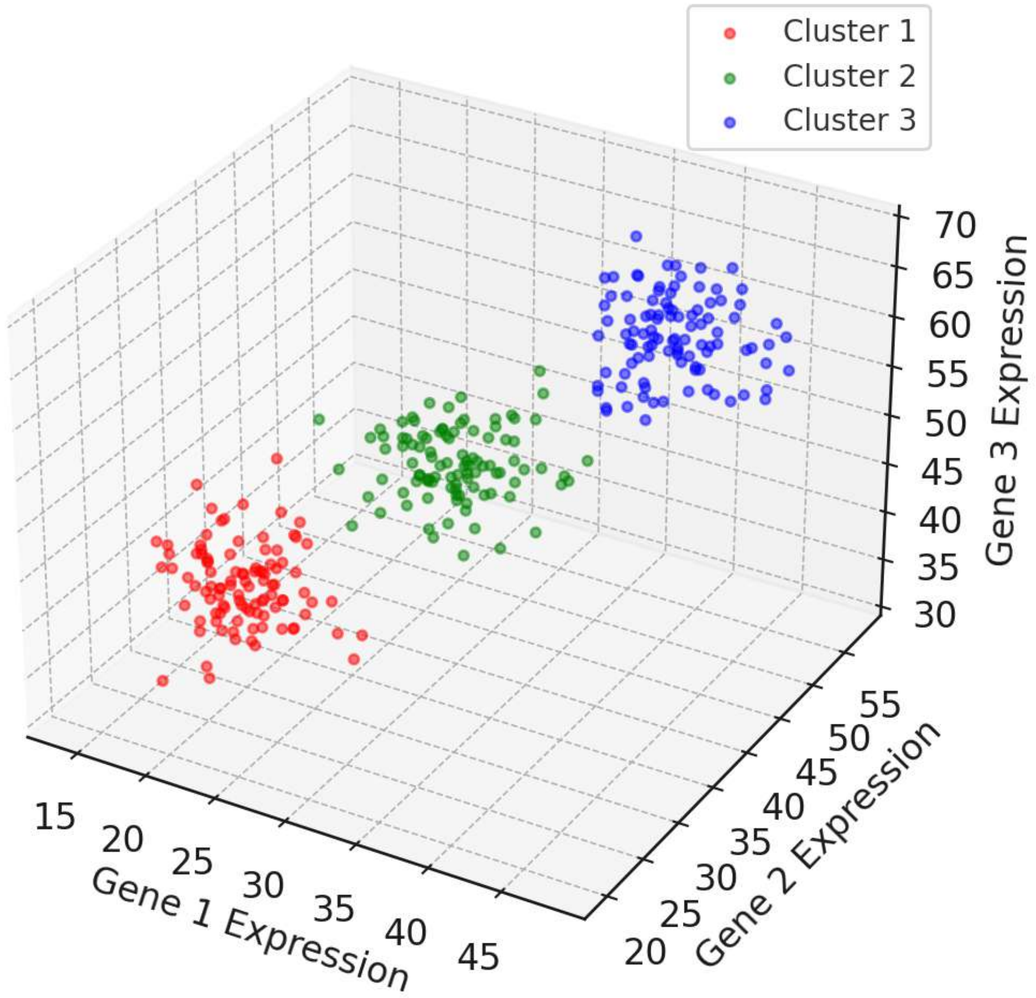
Each member of a group has different gene expression to other groups!

Identifying celltypes: Clustering



In theory we could just draw lines separating the groups and define our clusters

Identifying celltypes: Clustering



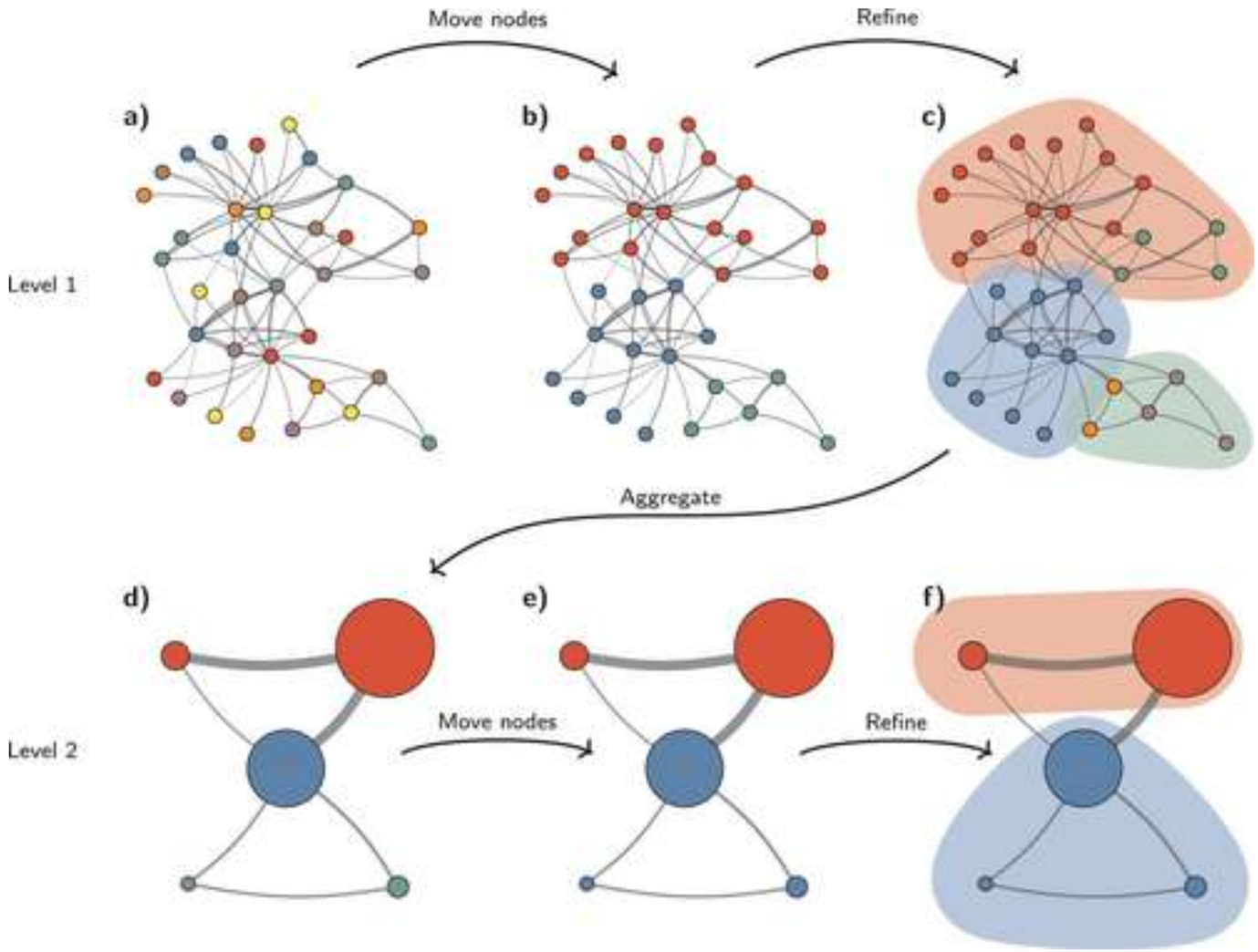
In theory we could just draw lines separating the groups and define our clusters

but remember we have 18,117 genes!

We can't just draw groups by eye

We need a way of identifying groups of cells I high dimensional spaces!

Identifying celltypes: Clustering



Leiden clustering – an algorithm that tries to find communities of datapoints that are highly similar to eachother

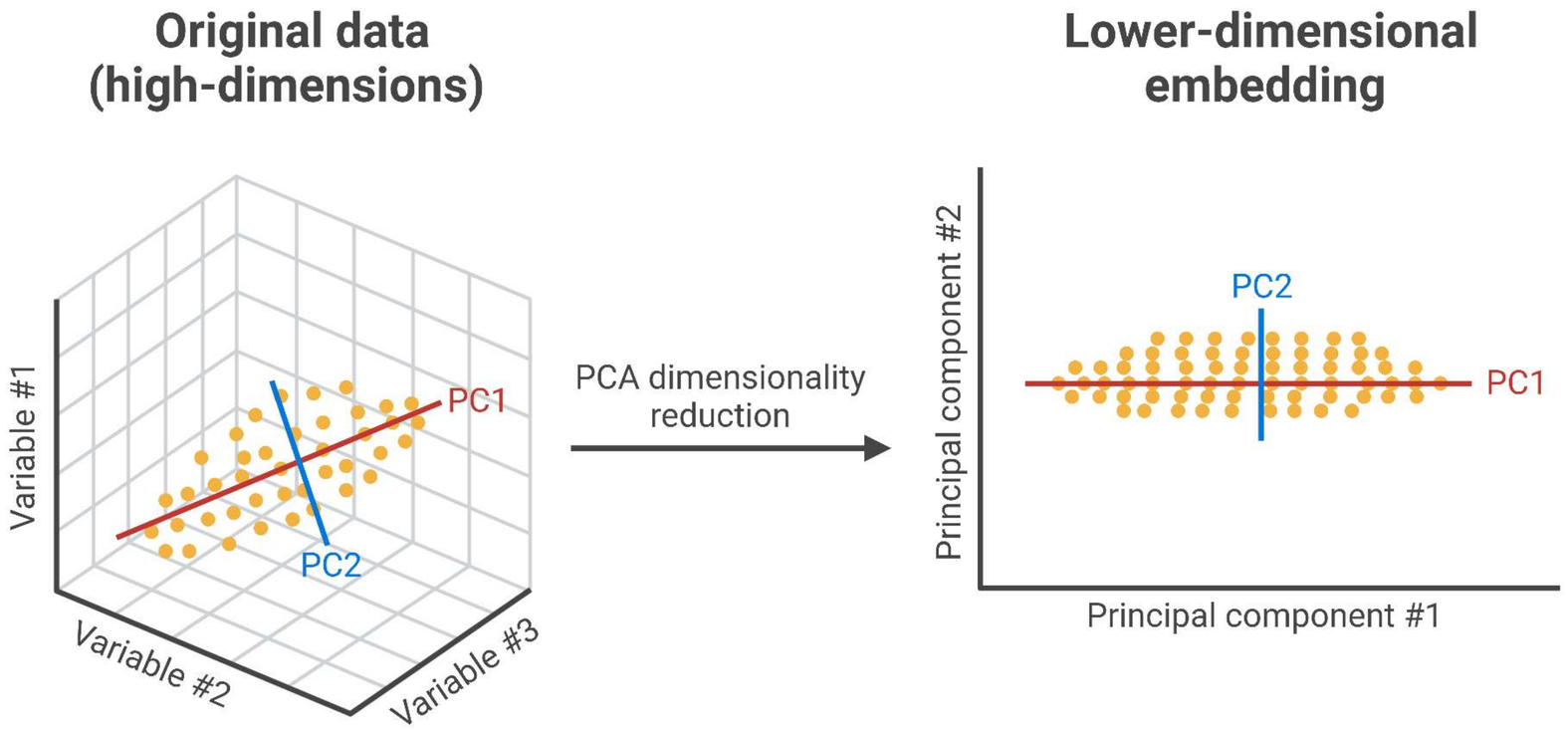
It does this by maximizing the modularity within a cluster

Modularity – quantifies how similar datapoints are (that belong to a cluster) compared to a random graph

<u>Clustering works well in high dimensional spaces!</u>

Identifying celltypes: Dimensionality reduction with PCA

$$Cv_i = \lambda_i v_i$$



C = Covariance matrix of our data

v = Eigenvectors or PCs

lambda = eigenvalues or variance captured by each PC

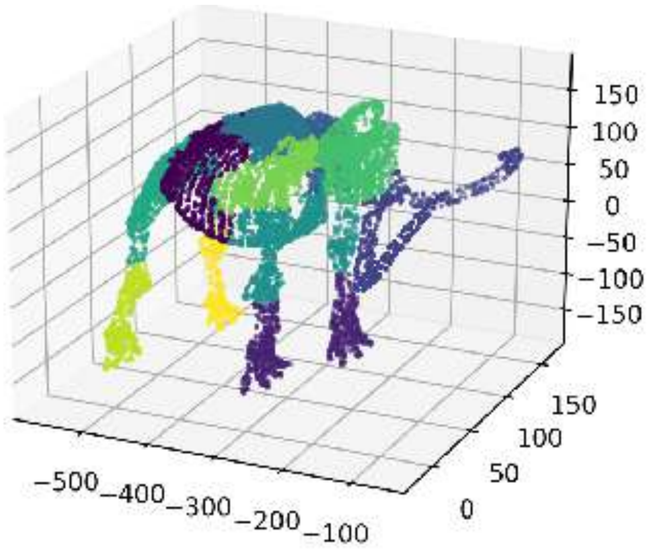
It can be shown that the eigenvectors of our covariance data matrix are:

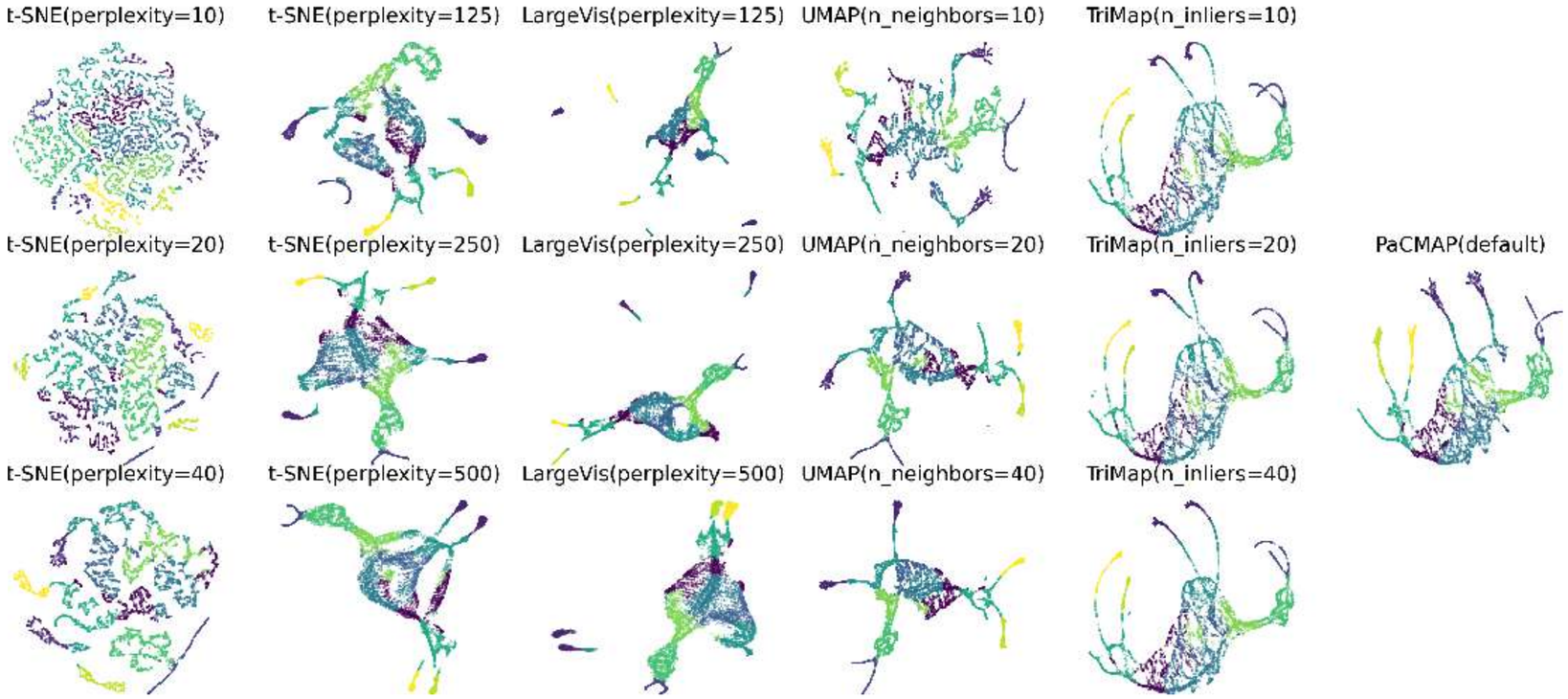
1. Capture the principal axes of variation in our data
2. While also being orthogonal
3. The eigenvalues tell us how much variance each PC captures, so we can rank them in order!

This allows us to choose n PCs

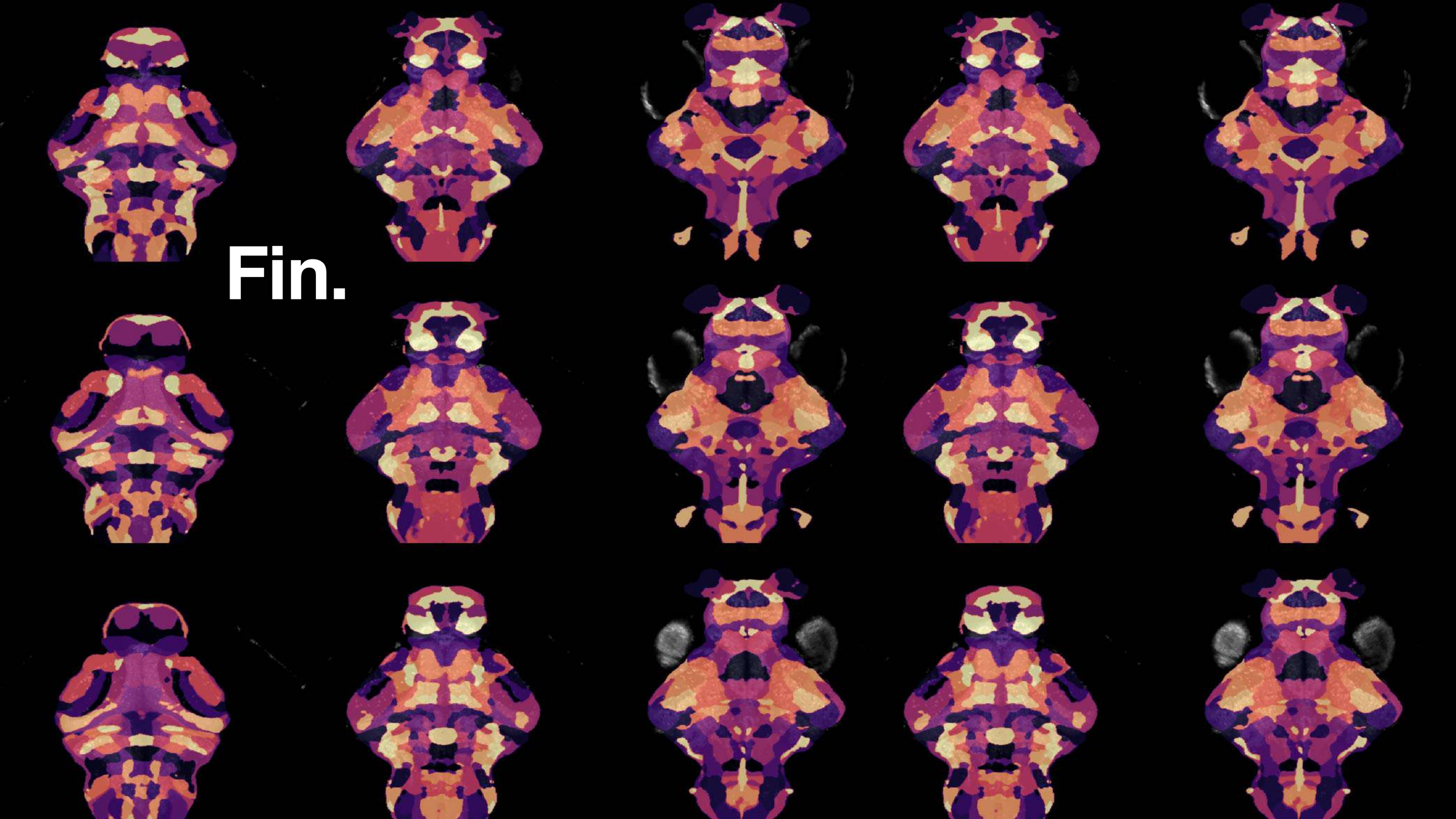
Identifying celltypes: Visualising our clusters with UMAP

TSNE/UMAP:
Tries to place similar points in high dimensional space closer to similar points in low dimensional space





Good at preserving local but not global structure!


Fin.