Statistical Methods for High-dimensional Biology

Model-based data analysis: Survey of biologically-inspired models

Yongjin Park, UBC Path&Lab, STAT, BC Cancer

Learning objectives

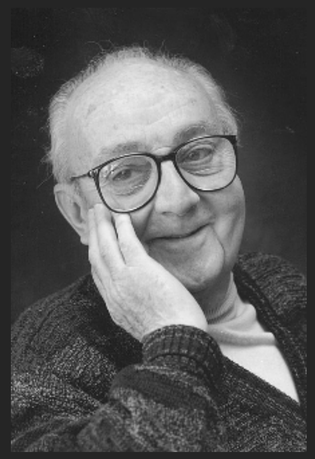
- How does modelling play a role in statistical learning in computational biology problems?
- Walk through several examples of model-based learning:
  - Q. What was the underlying statistical model?
  - Q. How did they do statistical inference?
  - Discussions

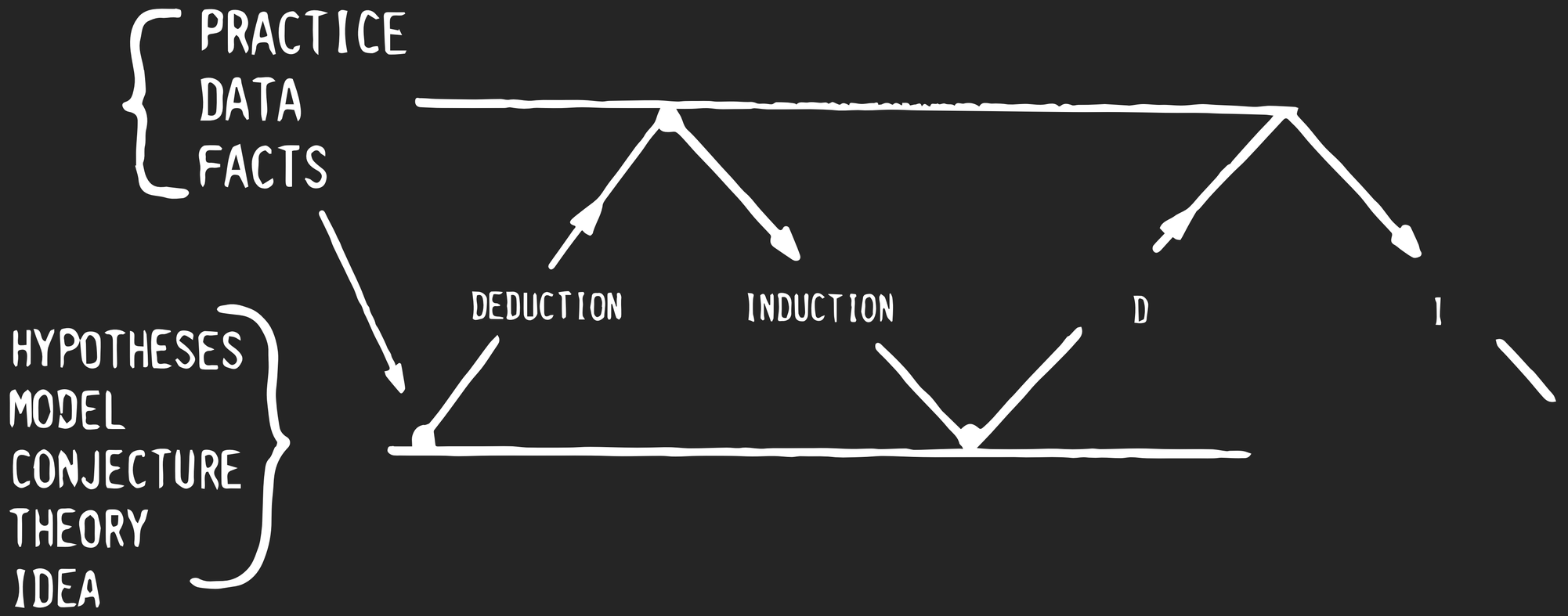
Today’s lecture: Model-based Data Analysis

- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

All models are wrong.

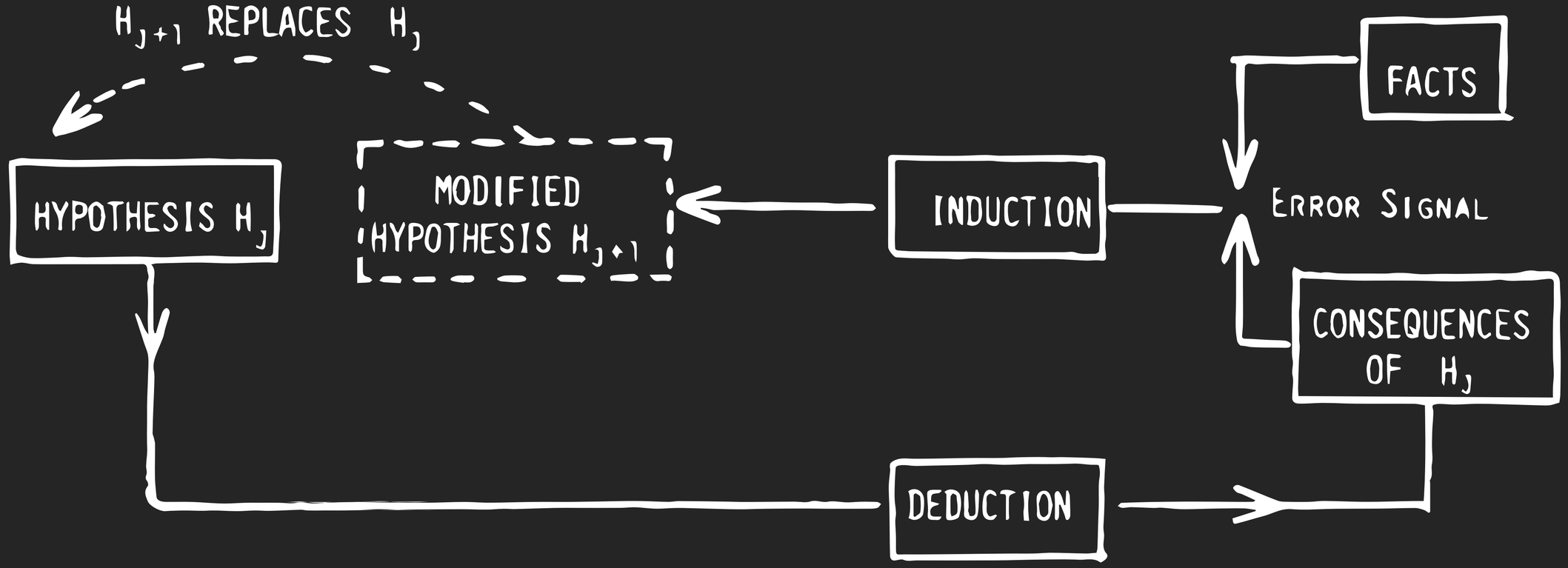


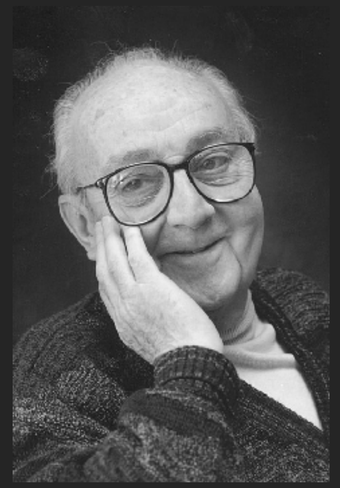


https://en.wikipedia.org/wiki/All_models_are_wrong

Box, JASA (1976)

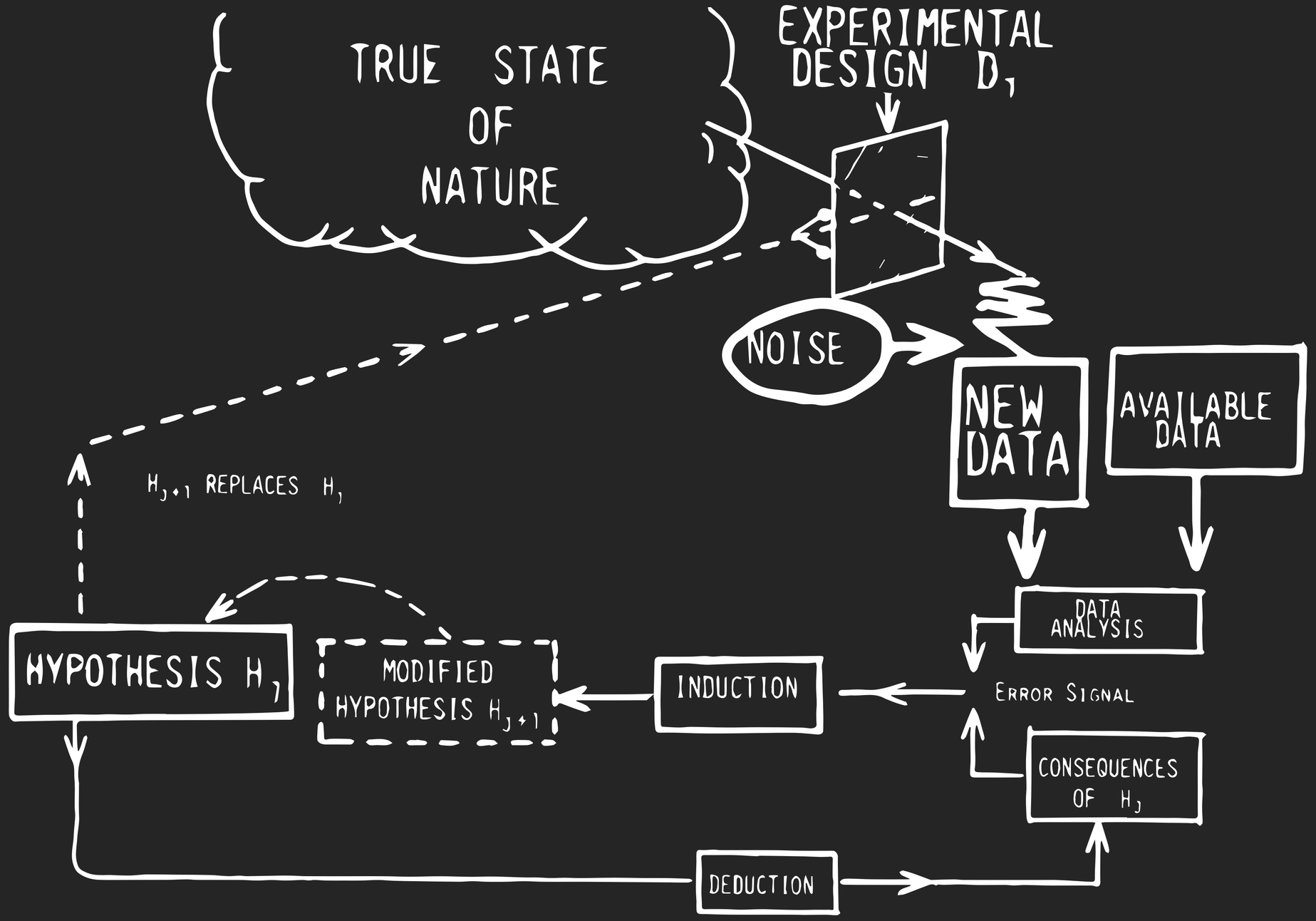
All models are wrong.





Box, JASA (1976)

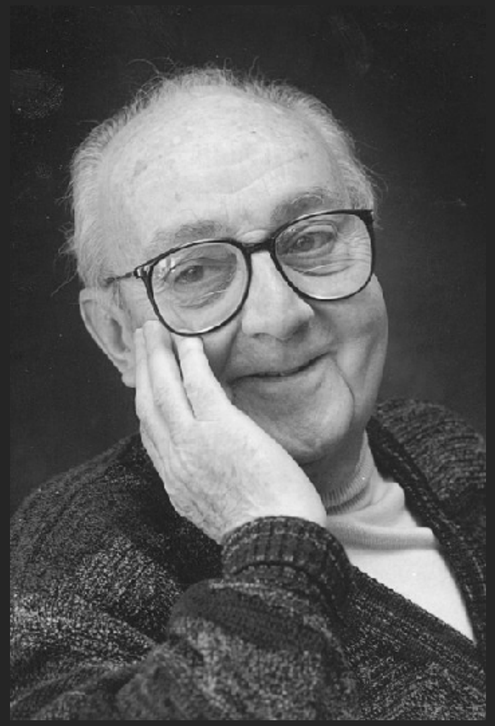
Data Analysis and Data Getting in the Process of Scientific Investigation



Box, JASA (1976)

Statistician-scientist

- Flexibility: "courage to seek out ... errors" "must not fall in love with model"
- Parsimony: "overelboration and overparameterization is often the mark of mediocrity"
- Worrying selectively: "in appropriate to be concerned about mice when there are tigers abroad"



Box, JASA (1976)

All models are wrong.

Box, JASA (1976)

Modelling is a process of abtraction

1. Set your goal with "flexibility" and "worrying selectively."
2. What data sets are (or will be) available?
   - What are the variables and sample size?
   - What level of detail do we want/need?
3. Focus on your goal (if needed, revise it)
4. Are there known problems and algorithms?
5. Is it worth a new algorithm? Or should we revise the problem?

Why should we care about modelling in biological data analysis?

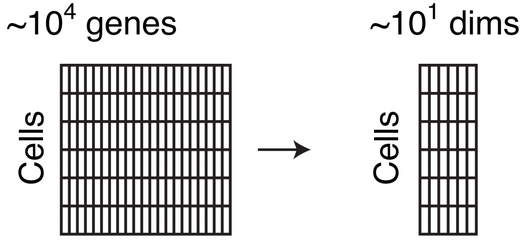
Today’s lecture: Model-based Data Analysis

- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

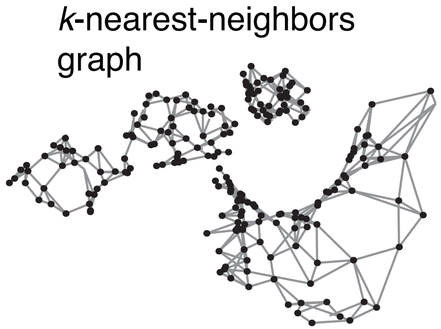
Learning dynamics from single-cell data

Reduction to a medium-dimensional space



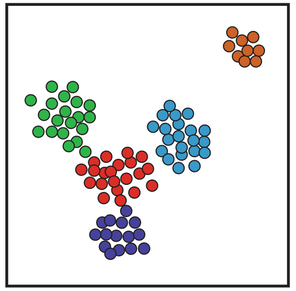
Find most informative set of reduced latent axes (10–50), use it to assess cell–cell similarity

Manifold representation



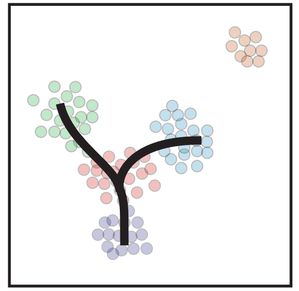
Capture complex, curved arrange-ments of cells in the expression space

Clustering and differential expression



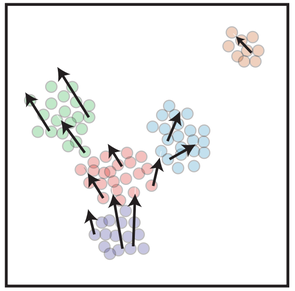
Identify discrete subpopulations of cells, and genes distin-guishing them

Trajectories



Capture contin-uous variation of cell state with trees or curves

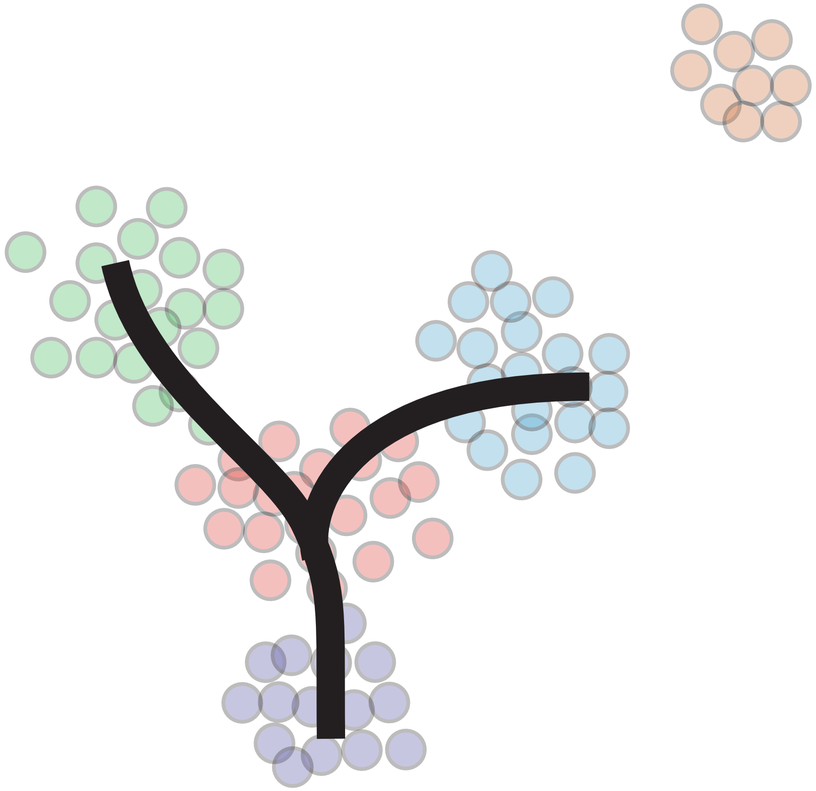
Velocity estimation



Predict state of the cells in the near future

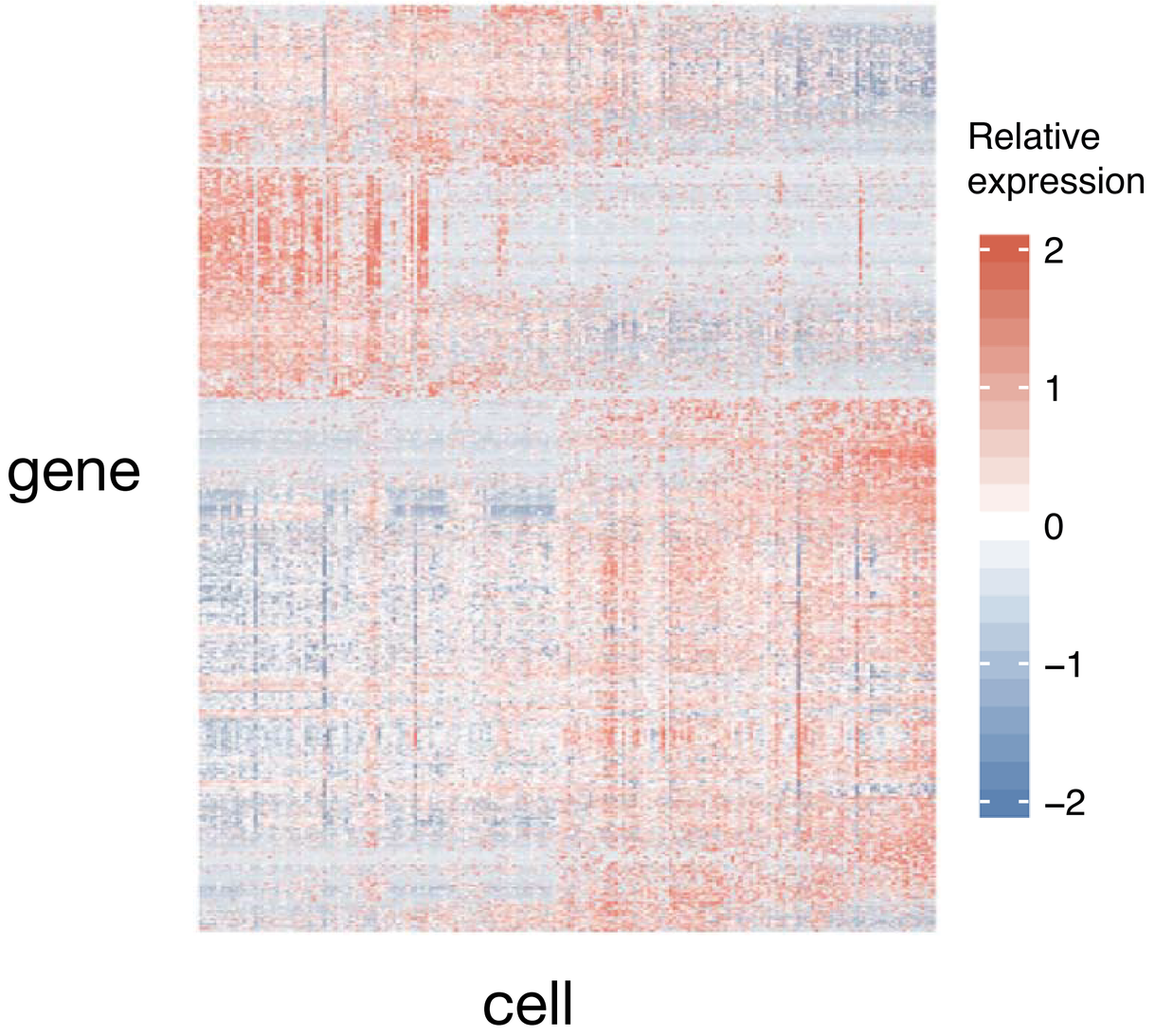
Karchenko (2021)

What is the goal of learning cell "trajectory?"



1. What do we mean by "trajectory?"
    - Why is it important or at least interesting?
2. What will be the goal?

Trajectory inference in bulk RNA-seq data

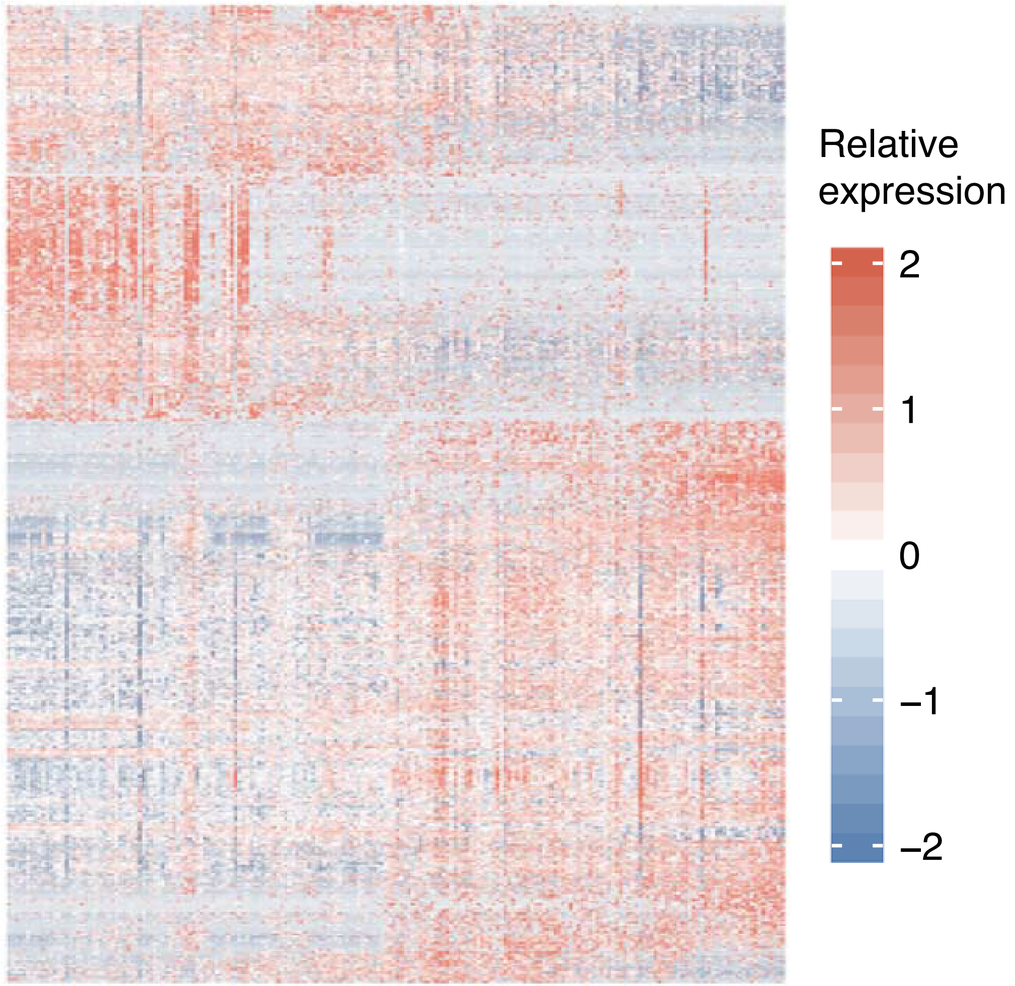


- What are the data sets?
- What are the variables?
- Which side is more important and interpretable?
- What level of detail do we hope to achieve?

Trapnell .. Rinn, Nature Method (2014)

What level of detail do we care about?

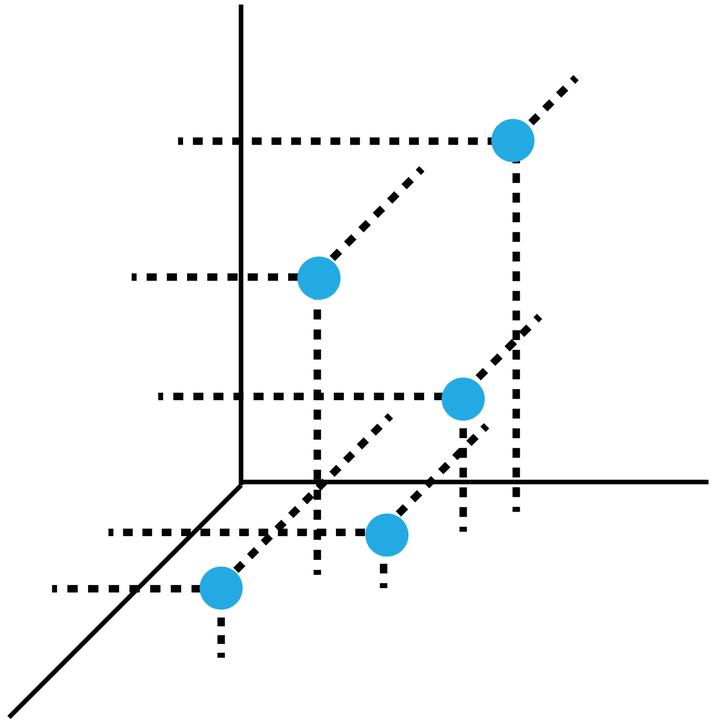
gene



cell

vs.

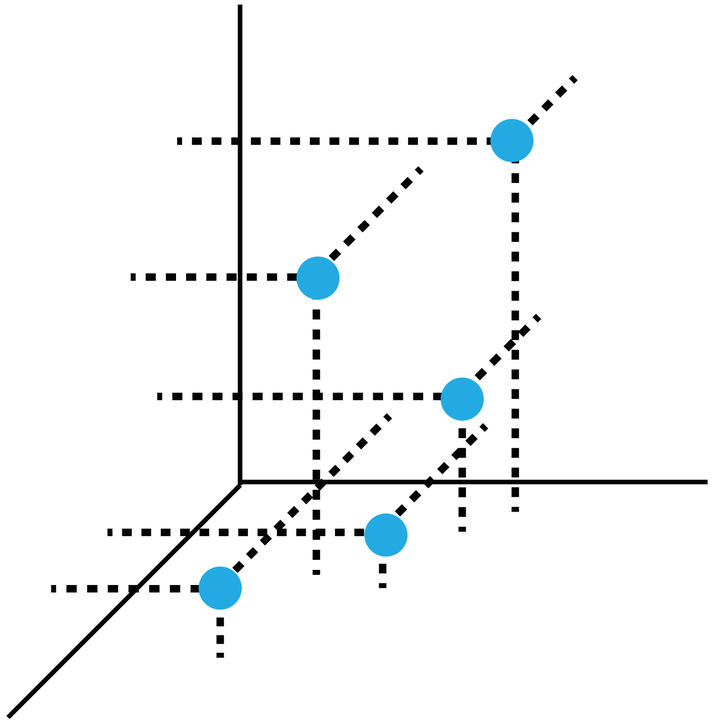
Cells represented as
points in expression space



Trapnell .. Rinn, Nature Method (2014)

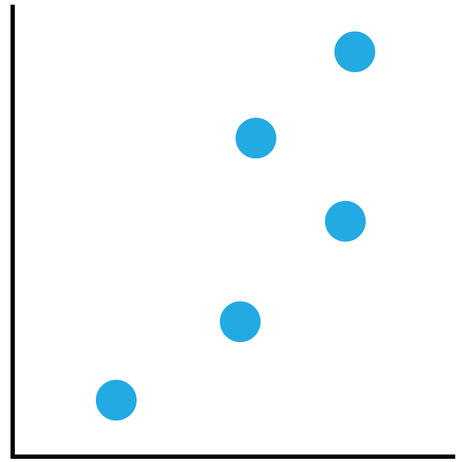
Trajectory: What level of detail?

Cells represented as
points in expression space



Each dot = a vector of gene expressions

Reduce dimensionality

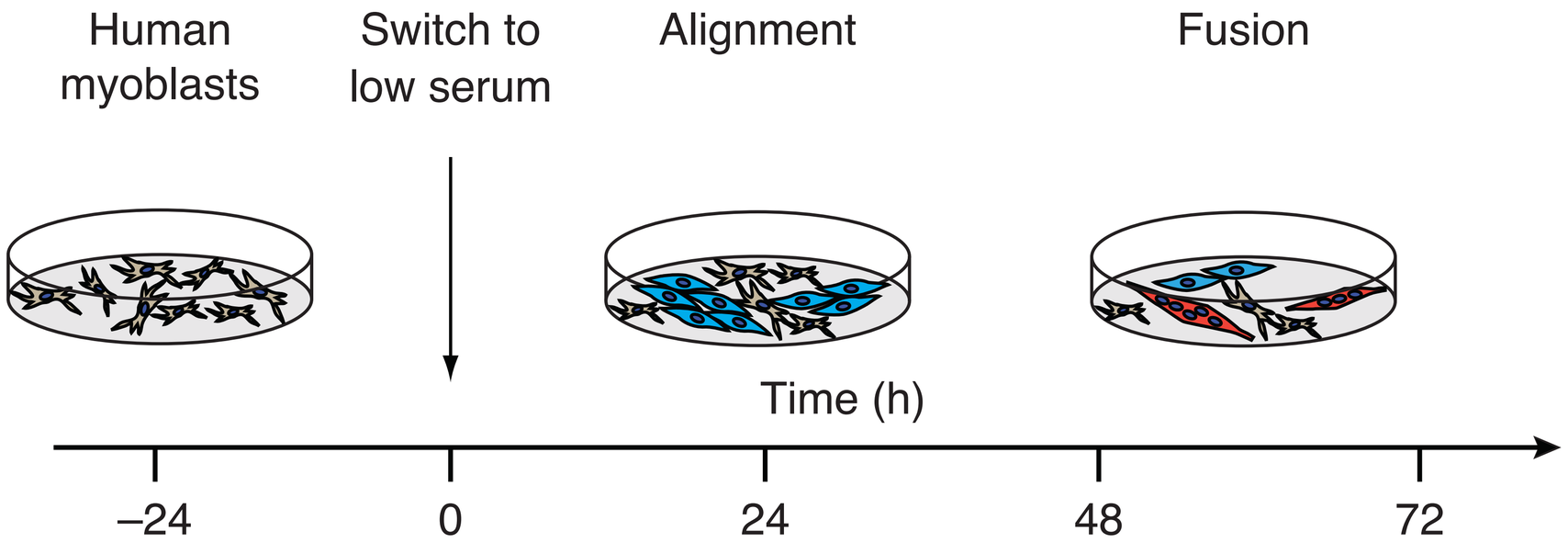


Each dot = a vector of PCs

Give up on gene level details

Trapnell .. Rinn, Nature Method (2014)

The goal of trajectory analysis is to assign "pseudo" time to each cell.

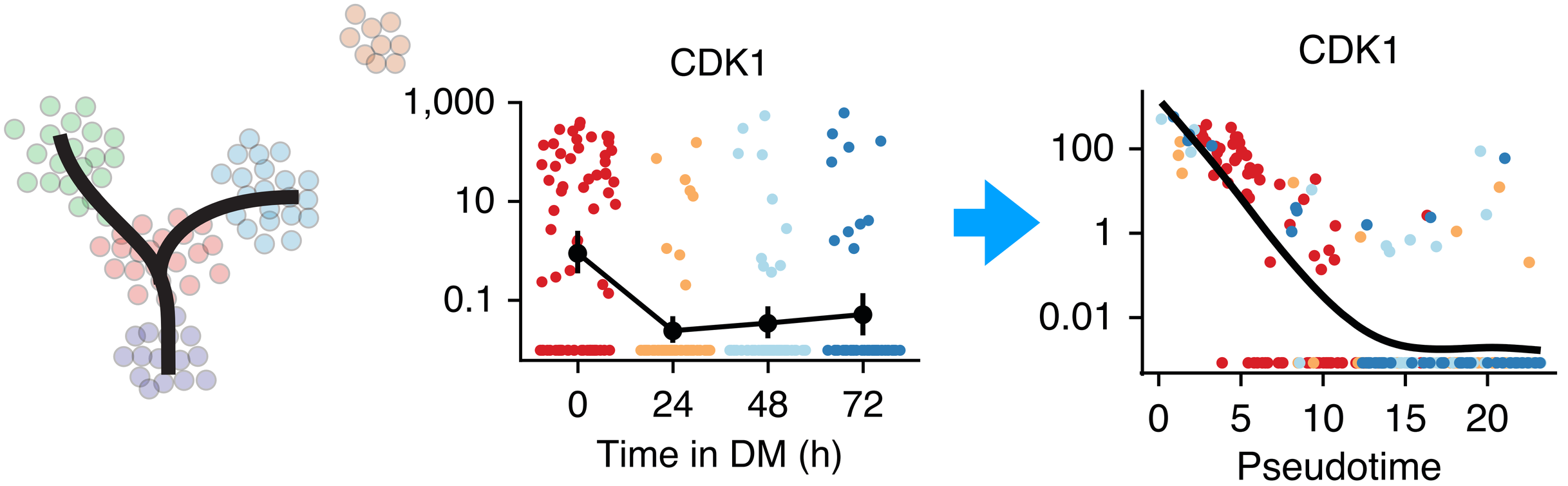


Cells captured	70	74	77	49
Reads/cell (millions)	4.8	5.2	4.5	4.3

Experimental time is not "precise" enough at a cell level

Trapnell .. Rinn, Nature Method (2014)

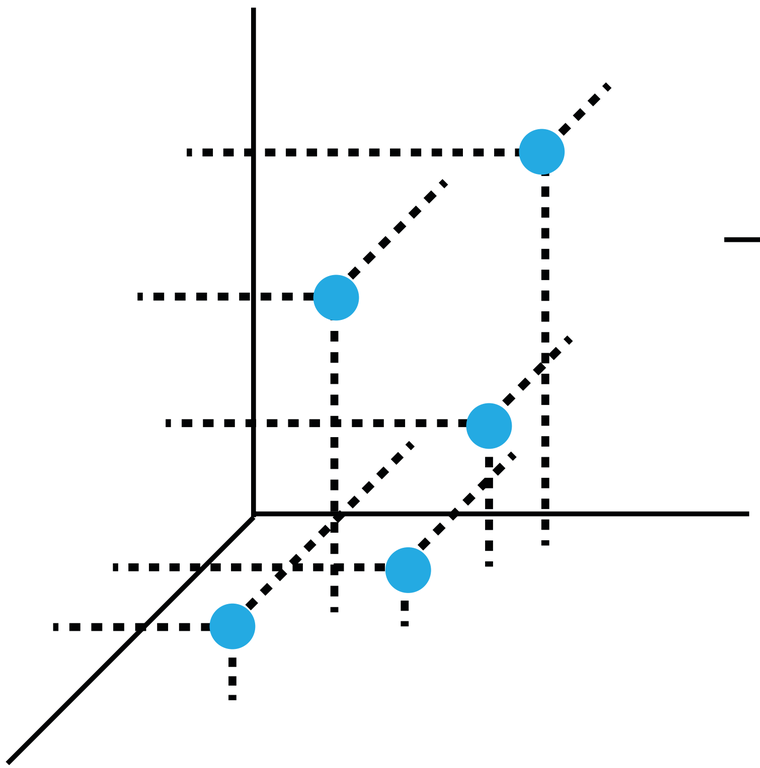
The goal of trajectory analysis is to assign "pseudo" time to each cell.



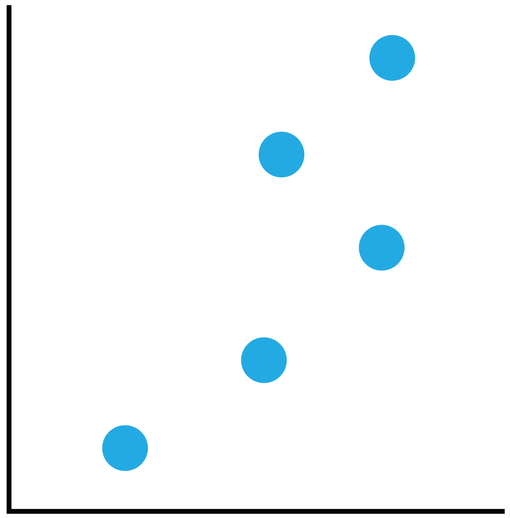
Trapnell .. Rinn, Nature Method (2014)

How can we resolve pseudotime?

Cells represented as
points in expression space

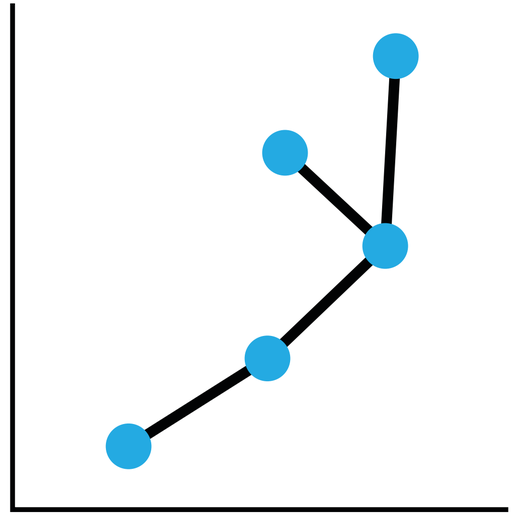


Reduce dimensionality



Each dot = a vector of PCs

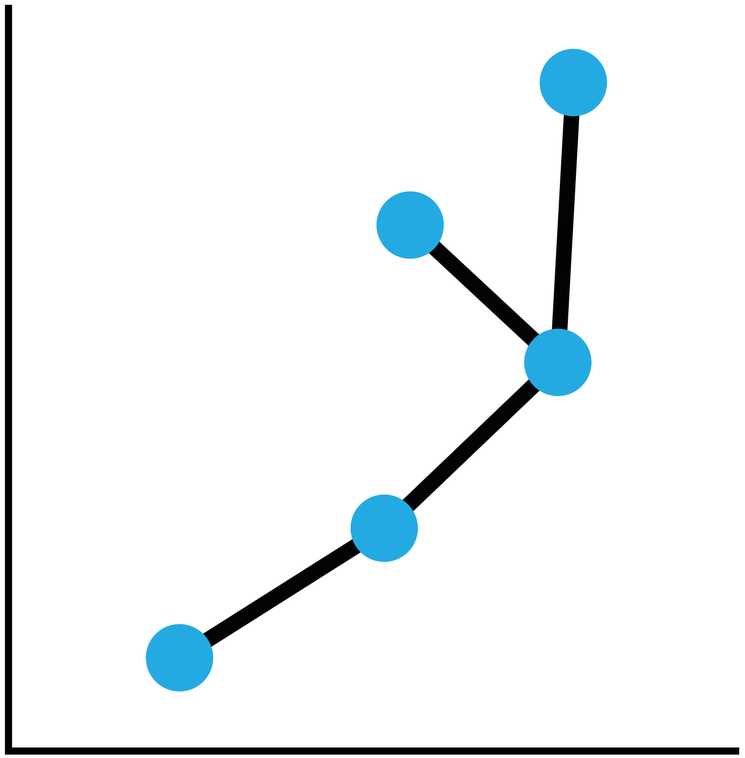
Build MST on cells

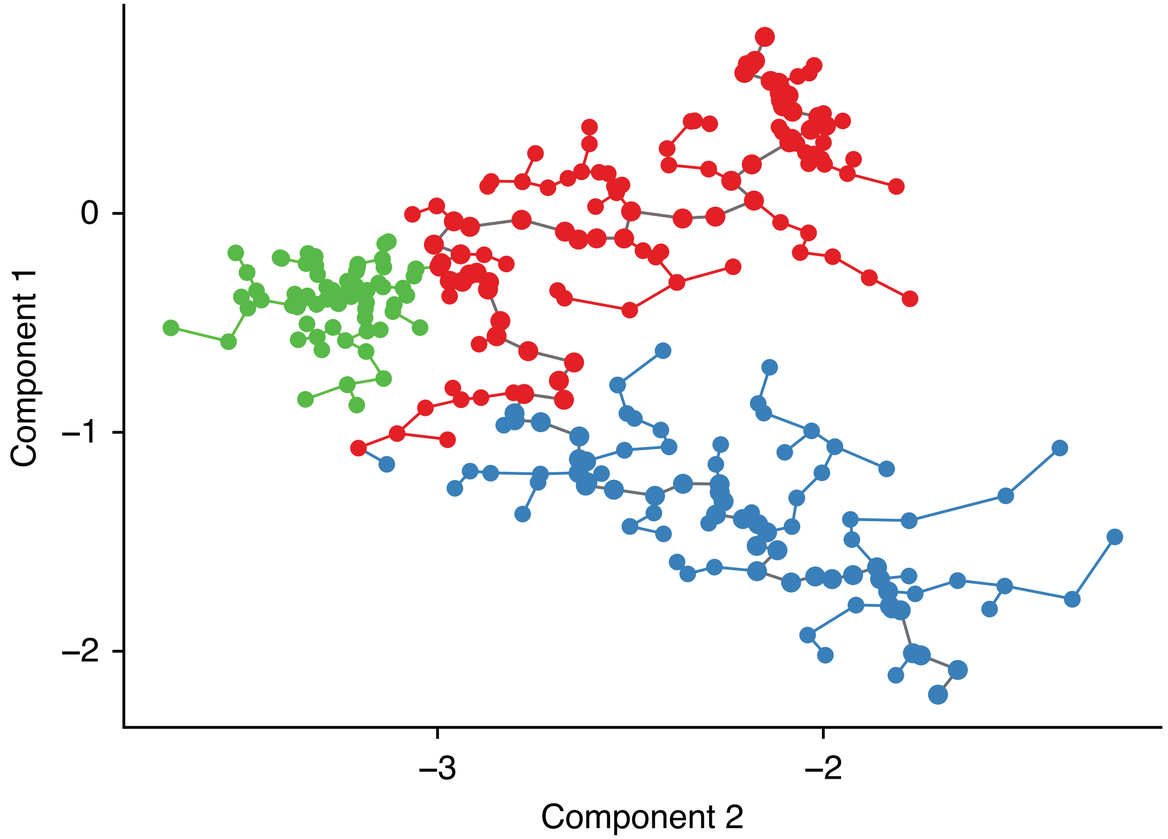


Cell-cell graph

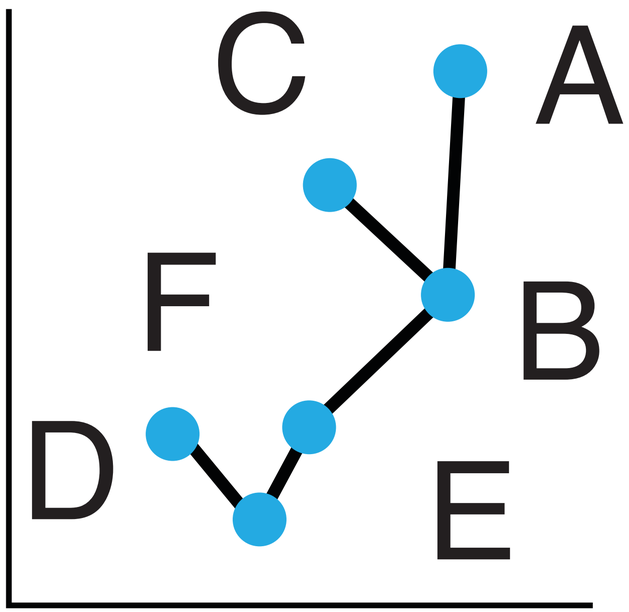
Trapnell .. Rinn, Nature Method (2014)

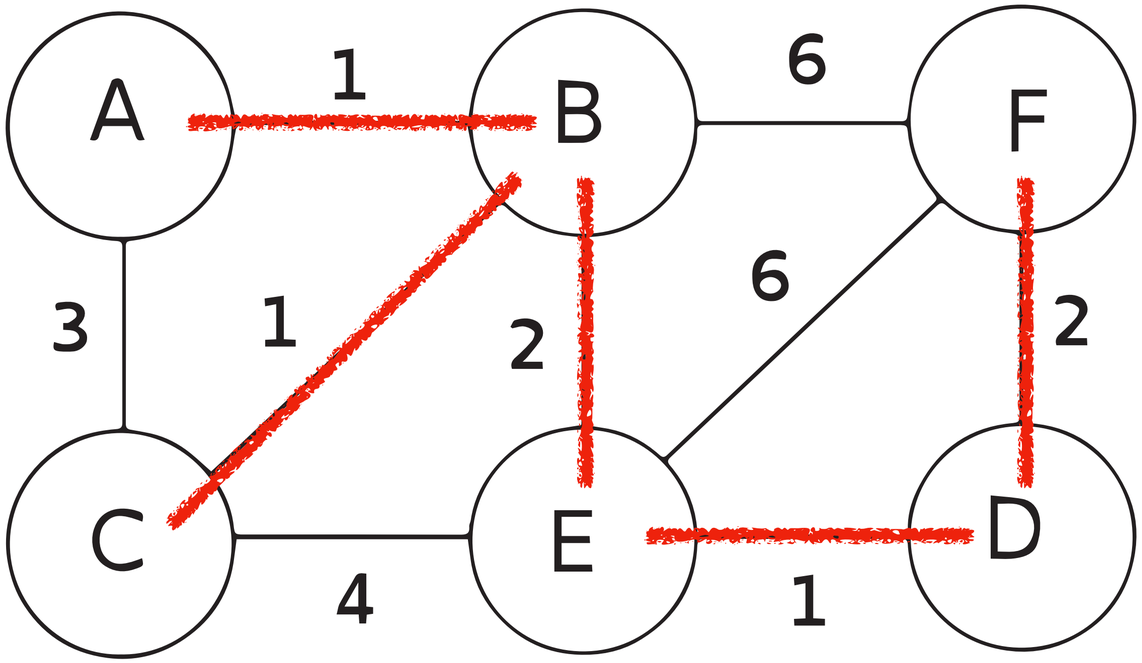
Cast our problem into a well-known problem





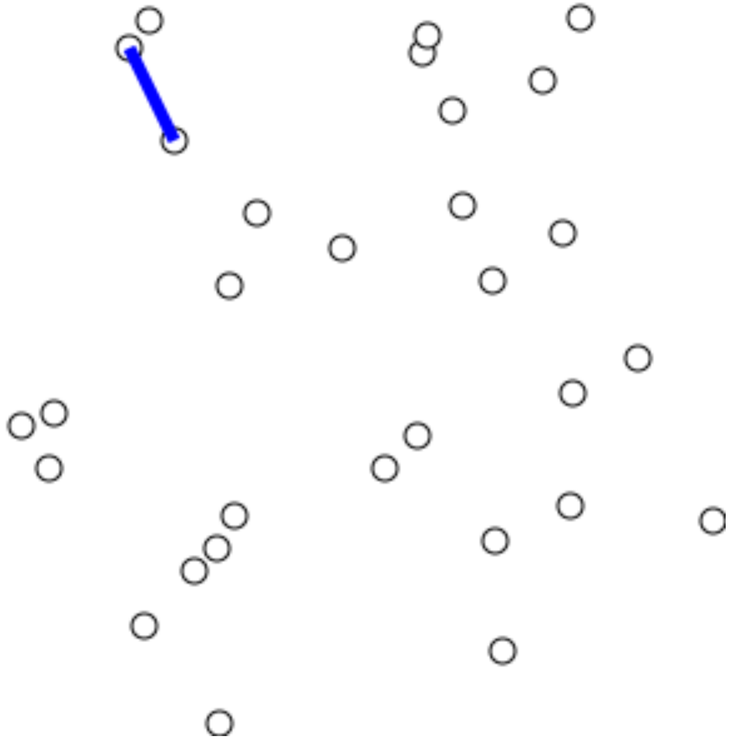
A minimum spanning tree problem





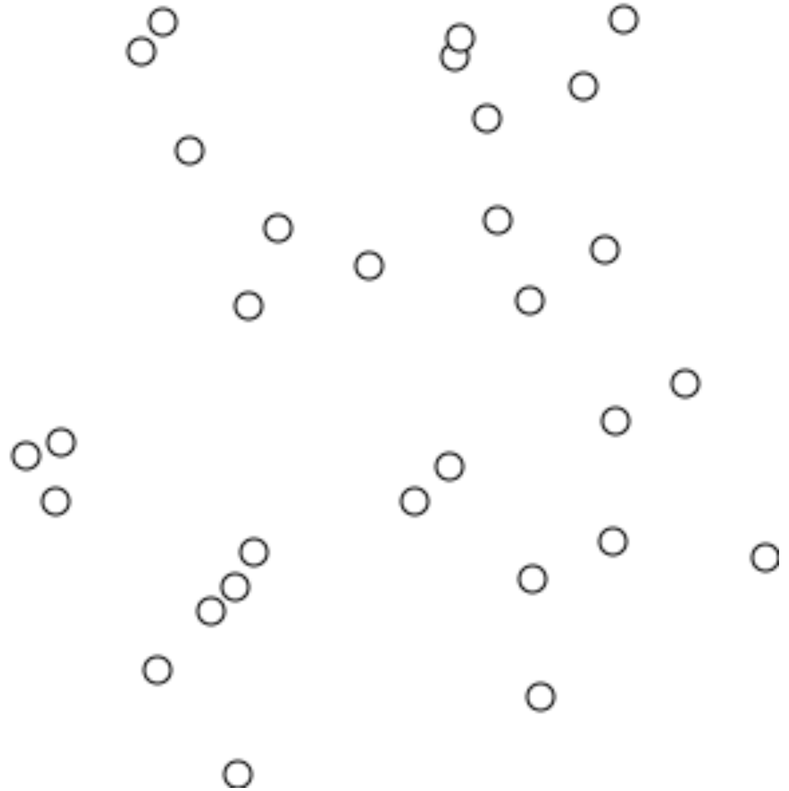
A minimum spanning tree problem

Prim's algorithm



https://en.wikipedia.org/wiki/Prim%27s_algorithm

Kruskal's algorithm



https://en.wikipedia.org/wiki/Kruskal%27s_algorithm

R igraph manual pages

Use this if you are using igraph from R

Side panel

mst {igraph} R Documentation

Minimum spanning tree

Description

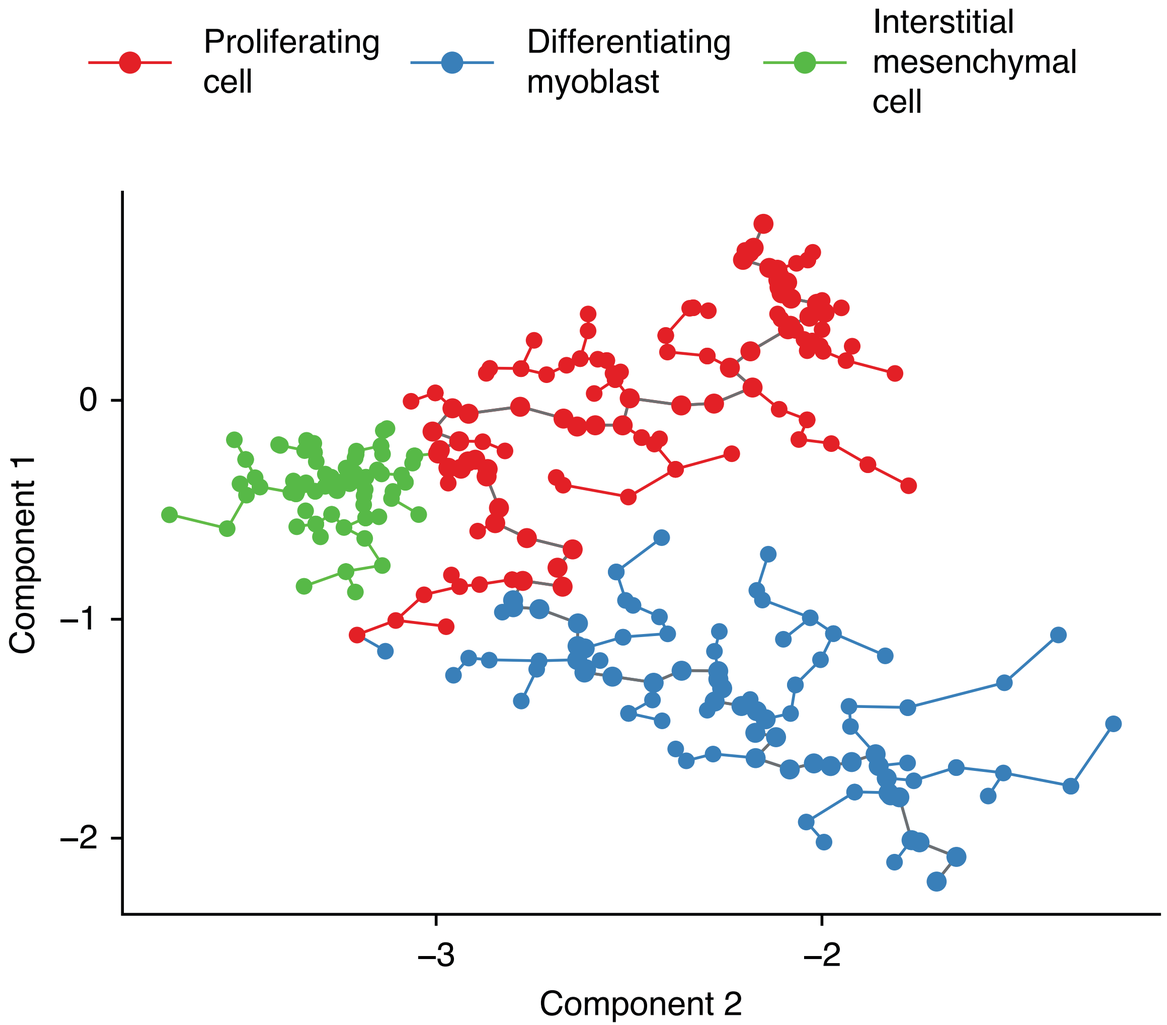
A subgraph of a connected graph is a *minimum spanning tree* if it is tree, and the sum of its edge weights are the minimal among all tree subgraphs of the graph. A minimum spanning forest of a graph is the graph consisting of the minimum spanning trees of its components.

Usage

```
mst(graph, weights = NULL, algorithm = NULL, ...)
```

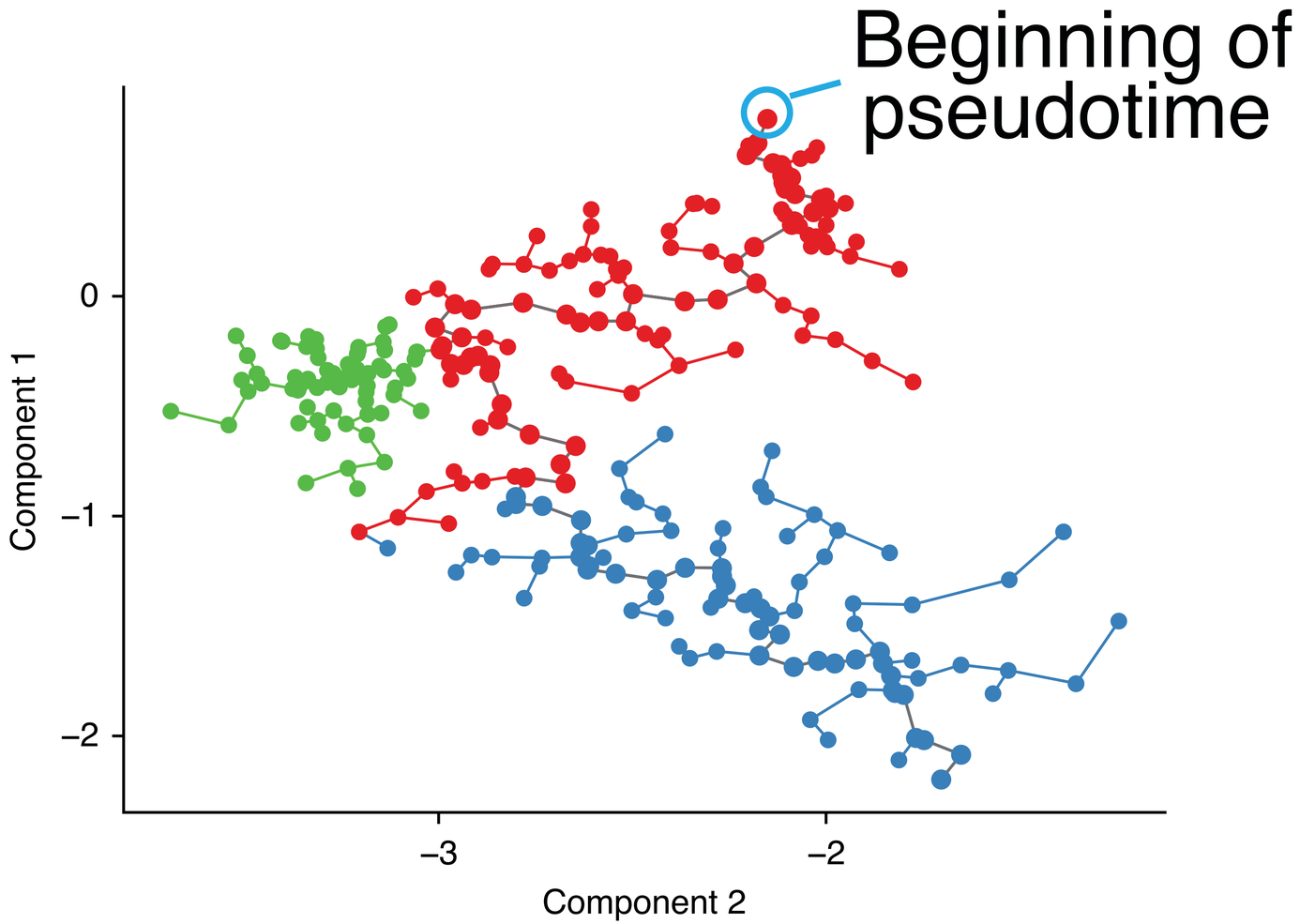
https://igraph.org/r/doc/mst.html

MST captures the backbone of time ordering



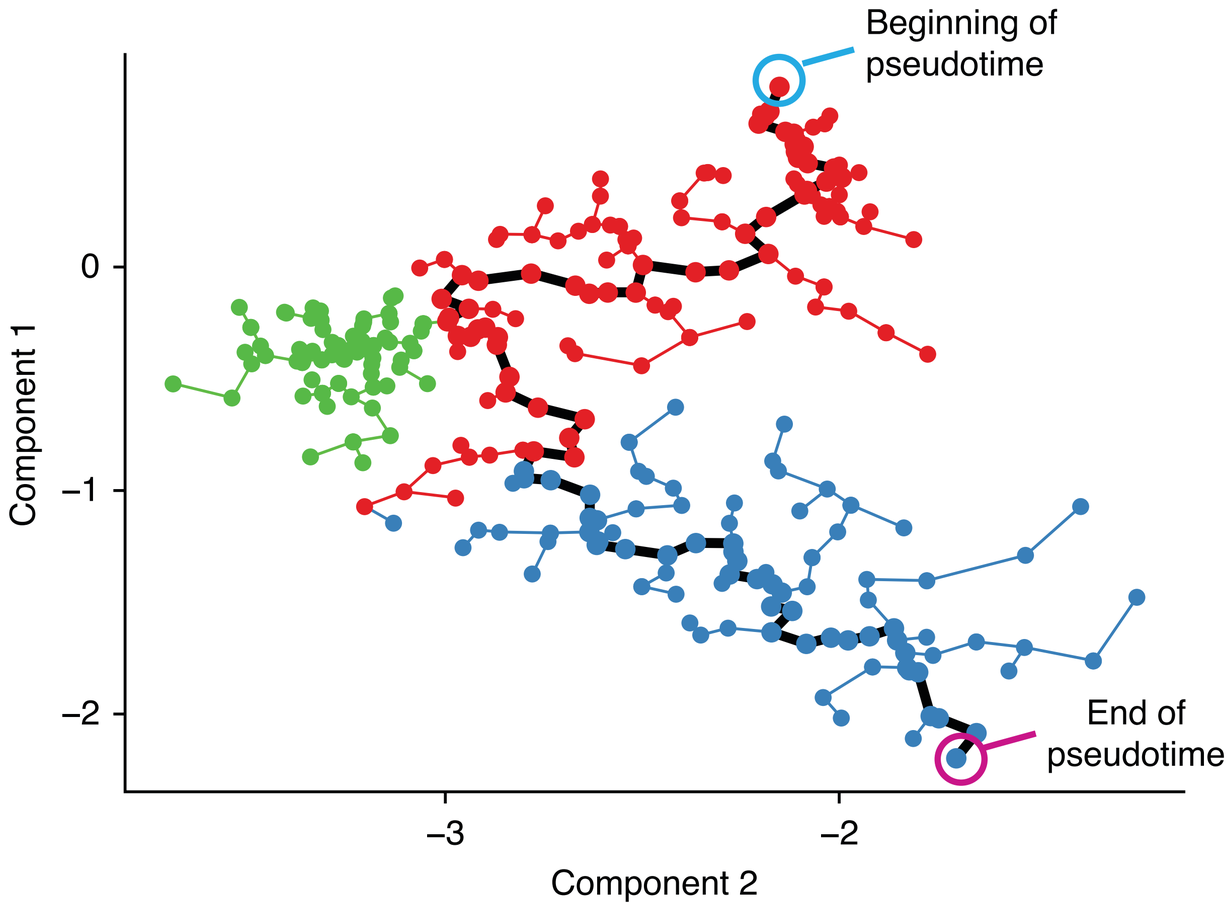
Trapnell .. Rinn, Nature Method (2014)

If we knew the starting point in MST



Trapnell .. Rinn, Nature Method (2014)

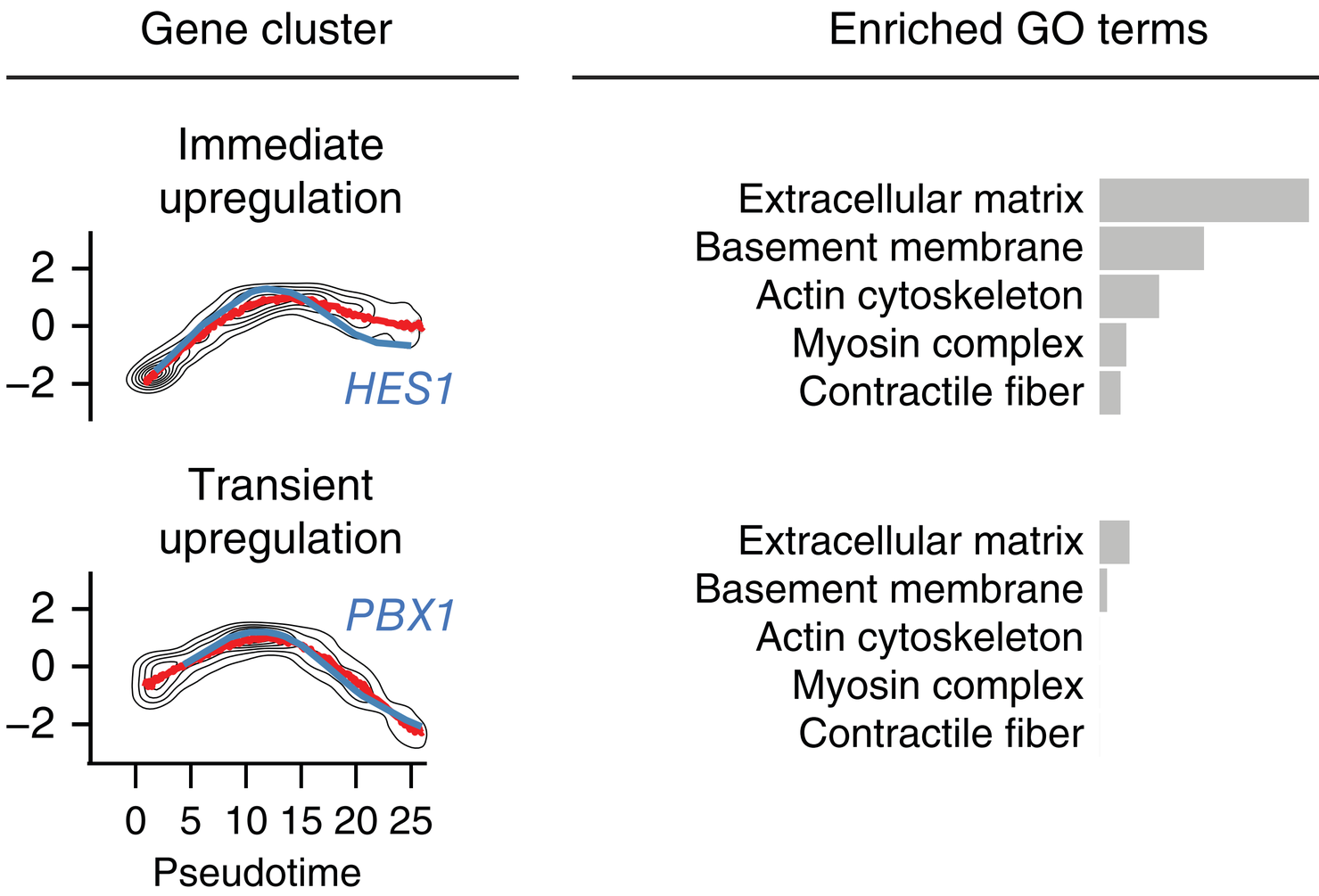
Assign pseudo time to cells along the longest path from the beginning to the end



1. How is it possible to assign pseudotime along a path from the root (beginning)?
2. How many parental nodes for each tree node?

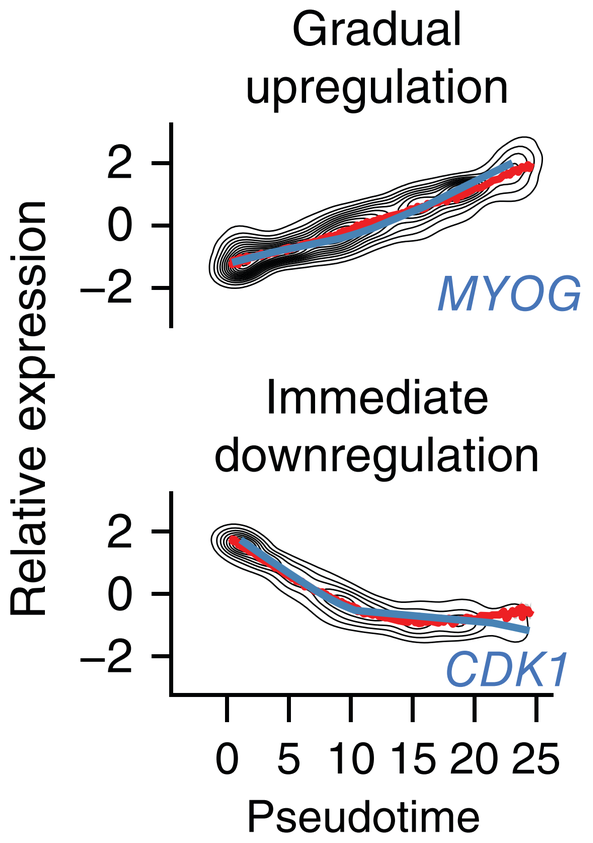
Trapnell .. Rinn, Nature Method (2014)

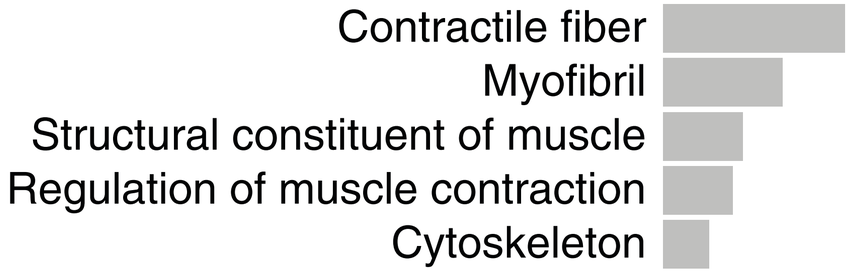
Dynamic gene programs identified by trajectory analysis and the following clustering

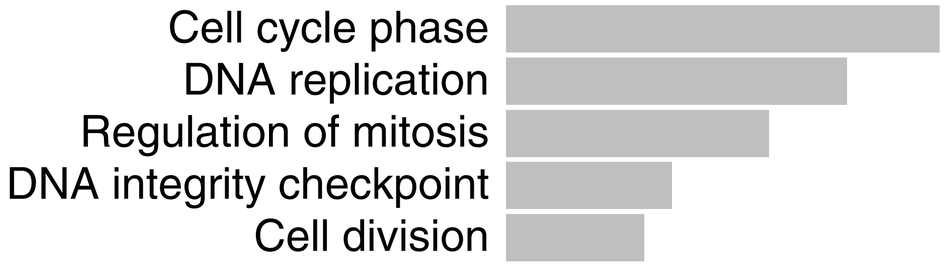


Trapnell .. Rinn, Nature Method (2014)

Dynamic gene programs identified by trajectory analysis and the following clustering







Trapnell .. Rinn, Nature Method (2014)

Gain more robustness in MST estimation

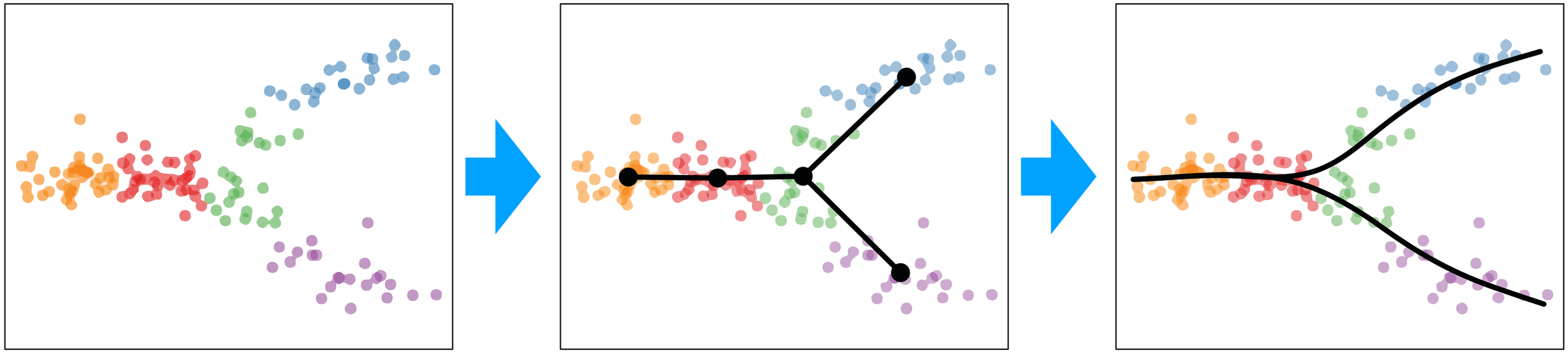
Street *et al. BMC Genomics* (2018) 19:477
https://doi.org/10.1186/s12864-018-4772-0

BMC Genomics

METHODOLOGY ARTICLE Open Access

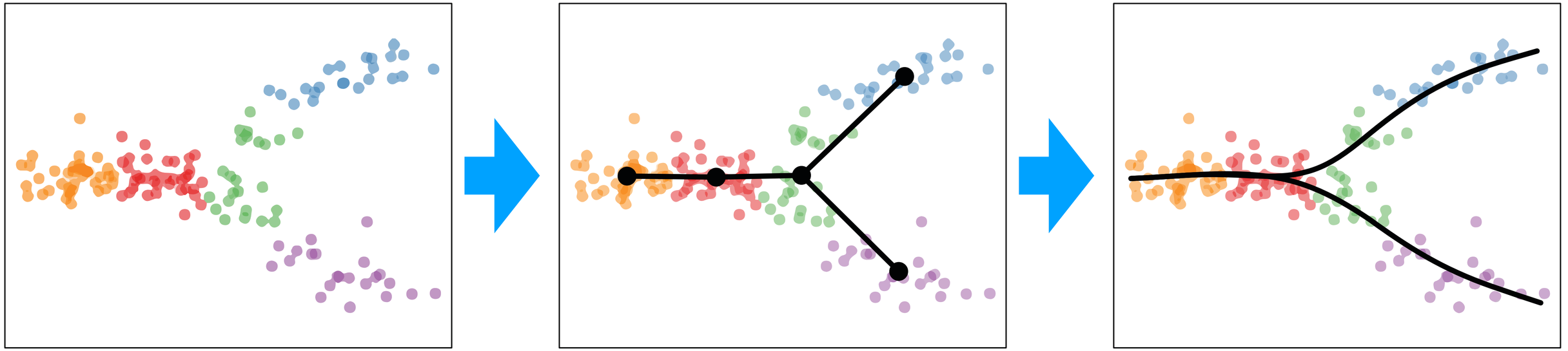
Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics

Kelly Street[1,8], Davide Risso[2], Russell B. Fletcher[3], Diya Das[3,9], John Ngai[3,6,7], Nir Yosef[4,8], Elizabeth Purdom[5,8] and Sandrine Dudoit[1,5,8,9*]



Street .. Dudoit, BMC Genomics (2018)

Single-cell-cluster level MST

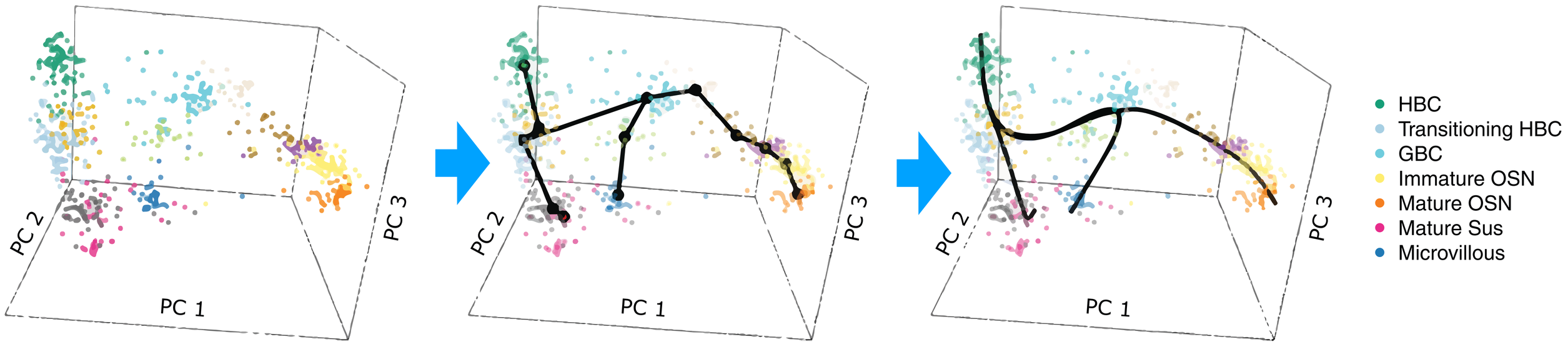


Clustering

Black dot: the centre of each cluster

Add the centre dots by Minimum Spanning Tree

What is the key assumption?



Clustering (colours)

MST of the centre points

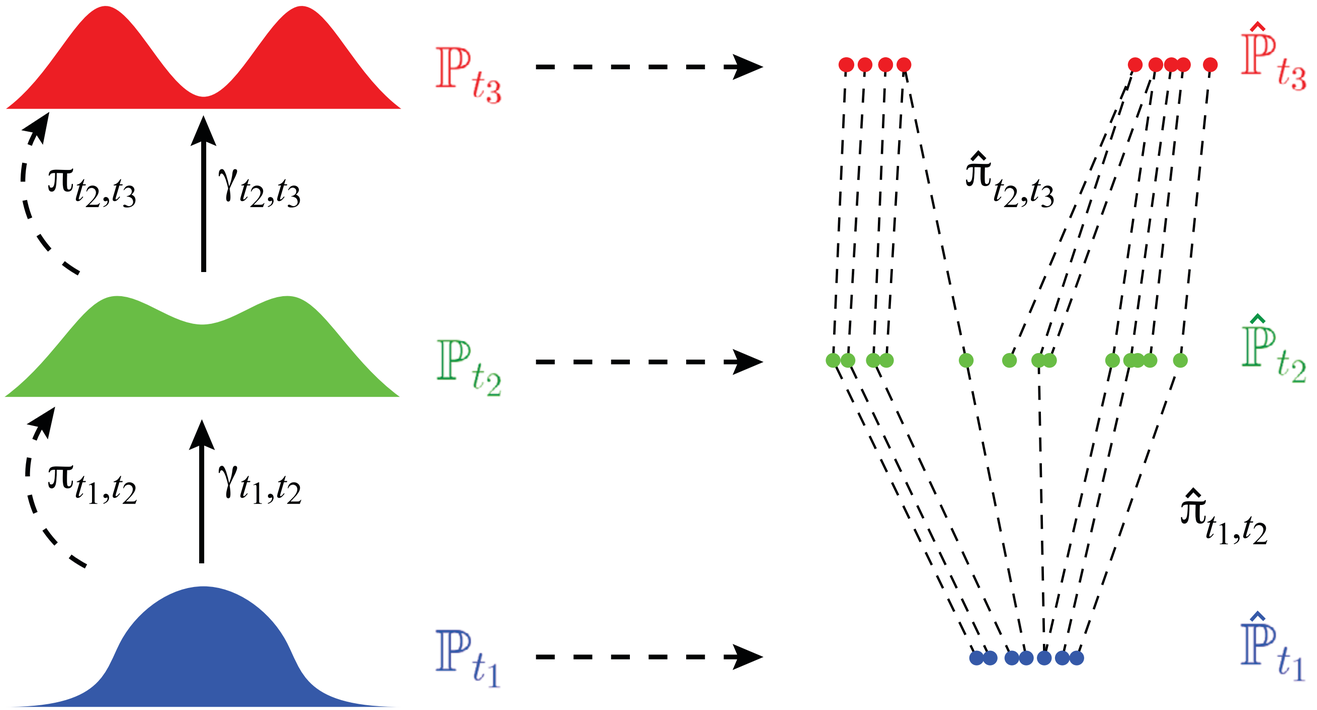
Non-parametric smoothing

Today’s lecture: Model-based Data Analysis

- **Model-based scientific investigation**

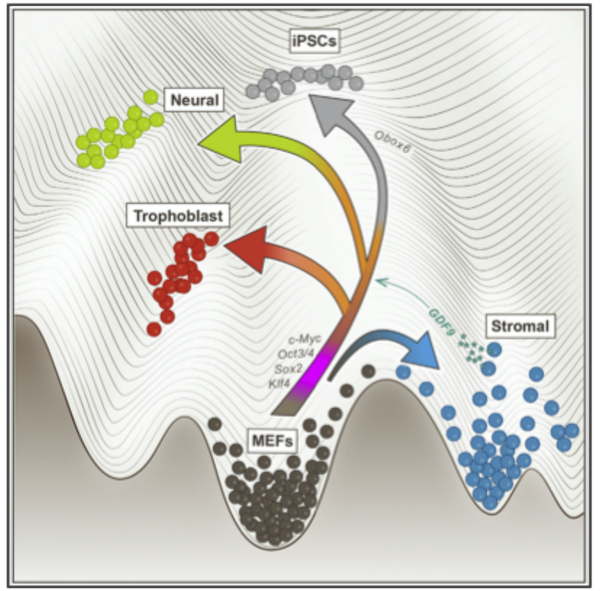
- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

Optimal Transport to interrogate developmental trajectories from time-series scRNA-seq data



Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

Graphical Abstract



Authors

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, ..., Rudolf Jaenisch, Aviv Regev, Eric S. Lander

Correspondence

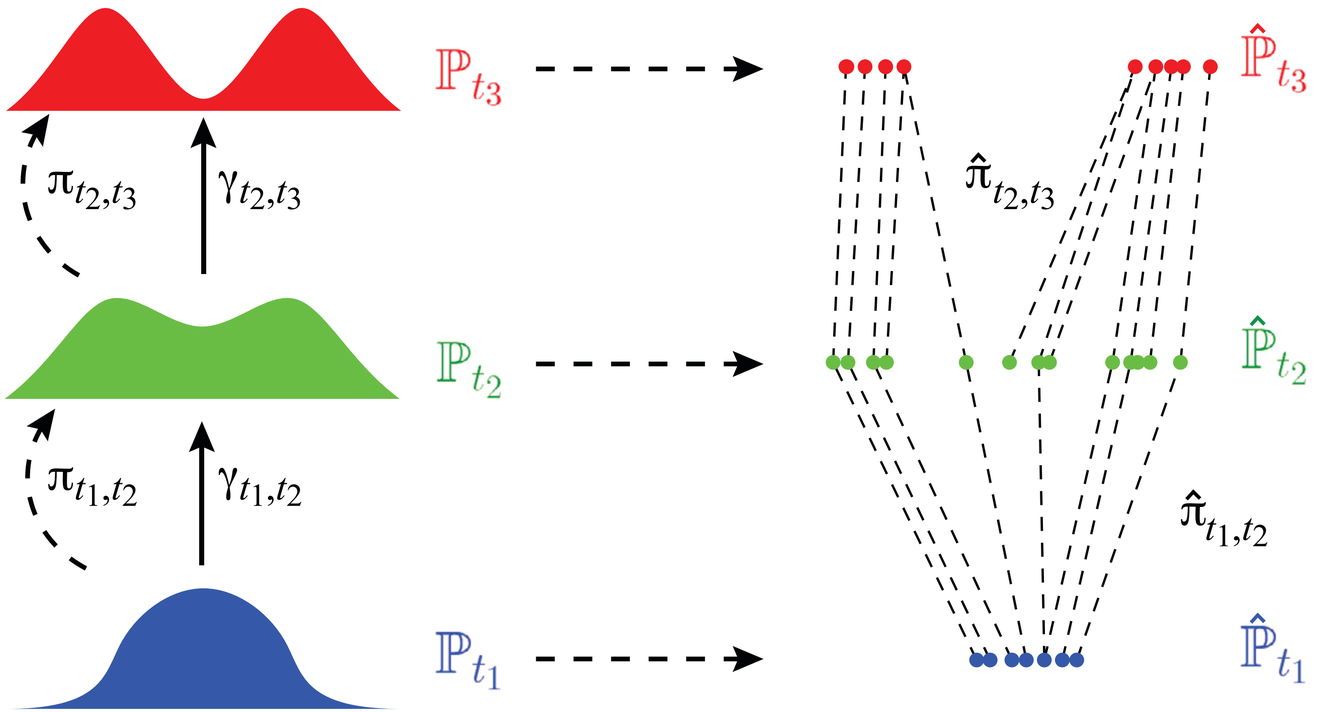
jianshu@broadinstitute.org (J.S.), aregev@broadinstitute.org (A.R.), lander@broadinstitute.org (E.S.L.)

In Brief

Application of a new analytical approach to examine developmental trajectories of single cells offers insight into how paracrine interactions shape reprogramming.

Schiebinger *et al.* Cell (2019)

Optimal Transport to couple cells between adjacent time points



Input (Coupling t1 and t2):

- Cost function (distance matrix) cells in t1 (x) cells in t2 (y)

  C(x,y)

- Marginal (empirical) distributions over G genes:

  Q(x) and P(y) on $\mathbb{R}^G$

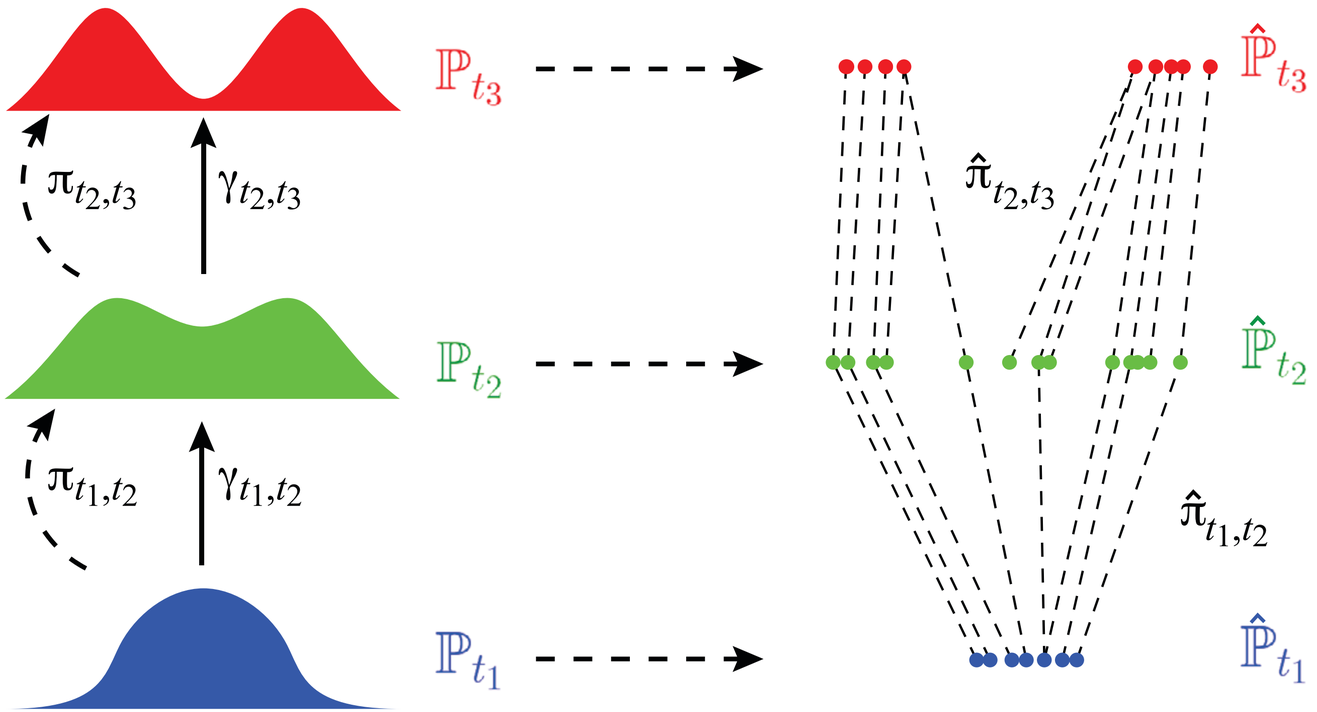
Goal:

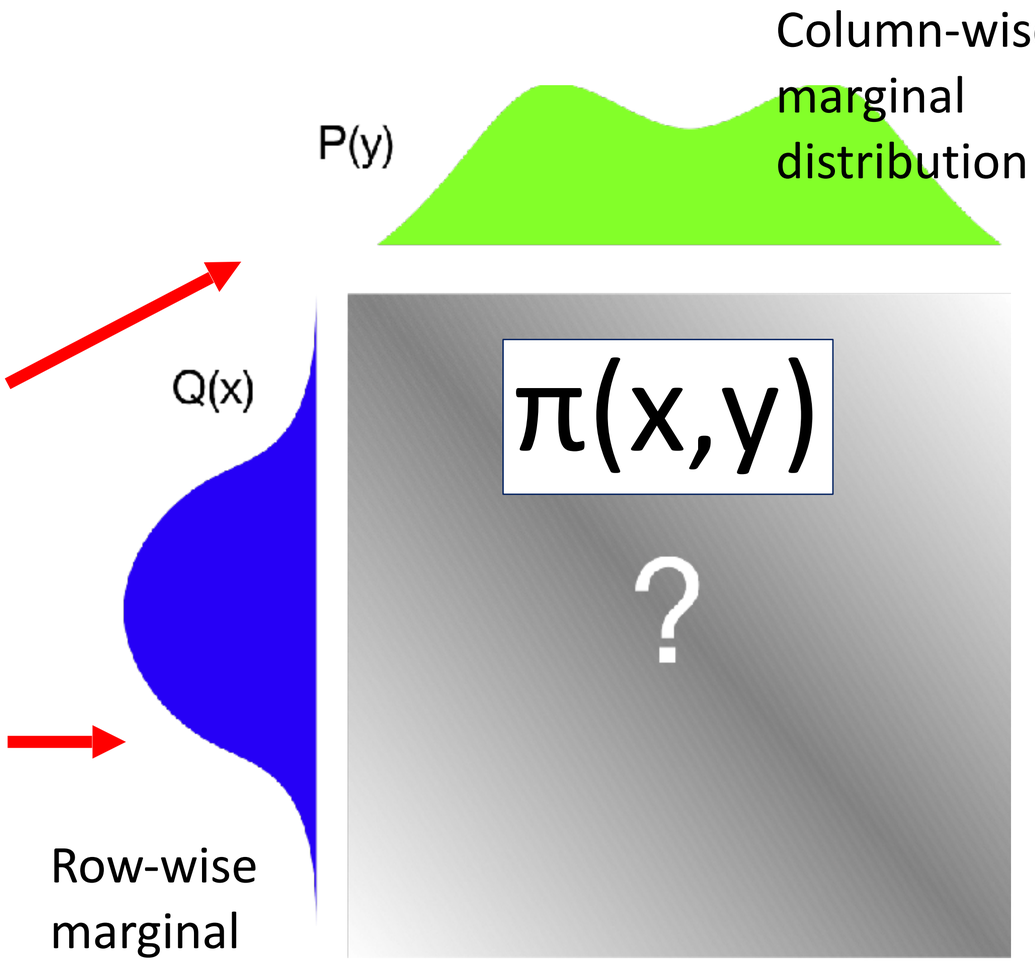
- Joint probability

  π(x,y)

Schiebinger *et al.* Cell (2019)

OT: Two marginal probs → Joint probability





Assume: biology is optimal

Schiebinger *et al.* Cell (2019)

Solving OT by constrained optimization

Objective function

$$\min_{\pi} \sum_{x} \sum_{y} C(x,y)\pi(x,y) - \epsilon H(\pi)$$

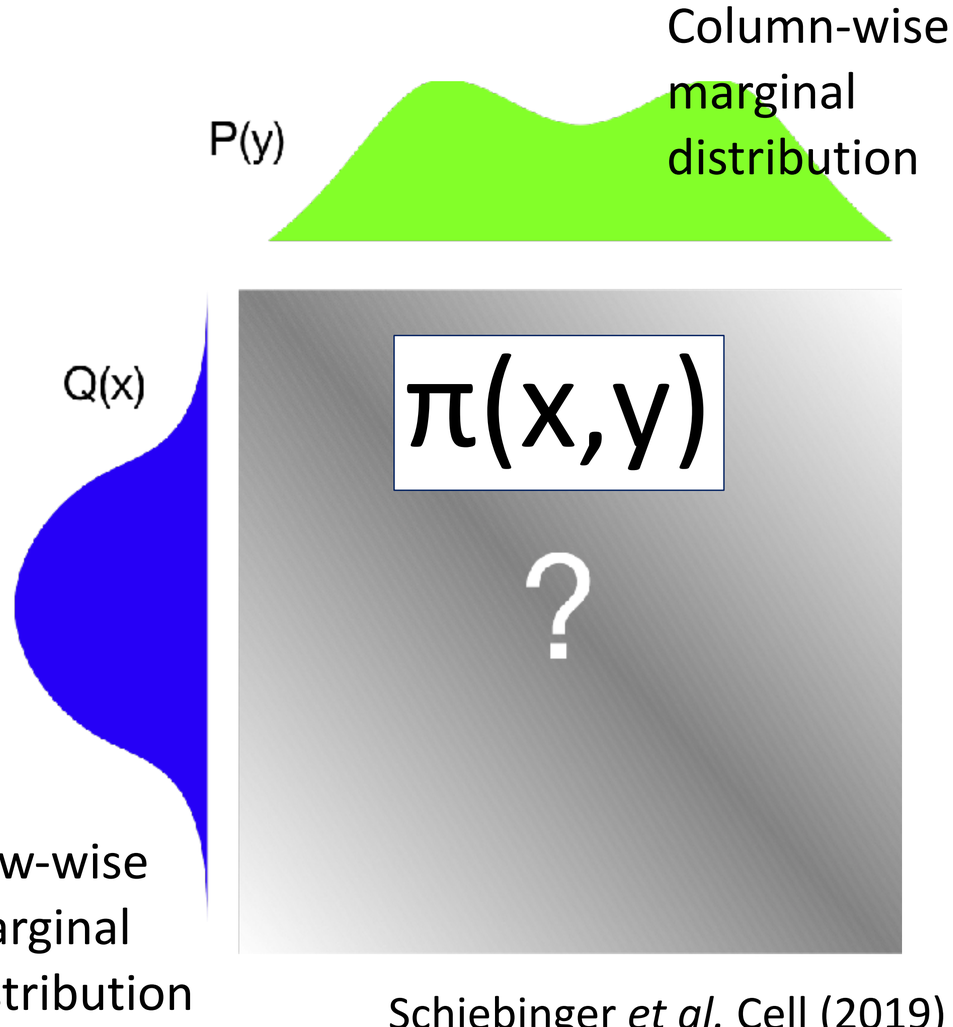
Entropy for smoothness & a unique soln.

Constraints

$$D_{KL}\left(\sum_{x} \pi(x,y) \,||\, P(y) \right) < s_1$$

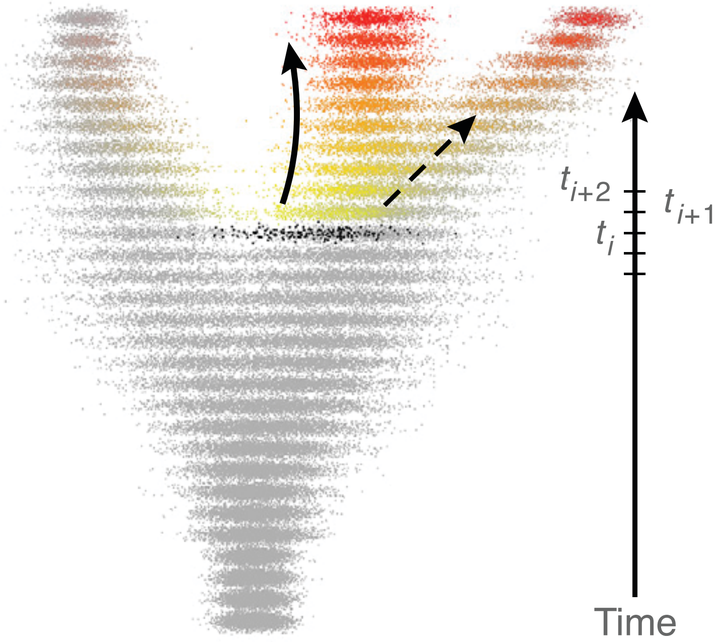
$$D_{KL}\left(\sum_{y} \pi(x,y) \,||\, Q(x) \right) < s_2$$

Row and col-wise marginal distrib. should match

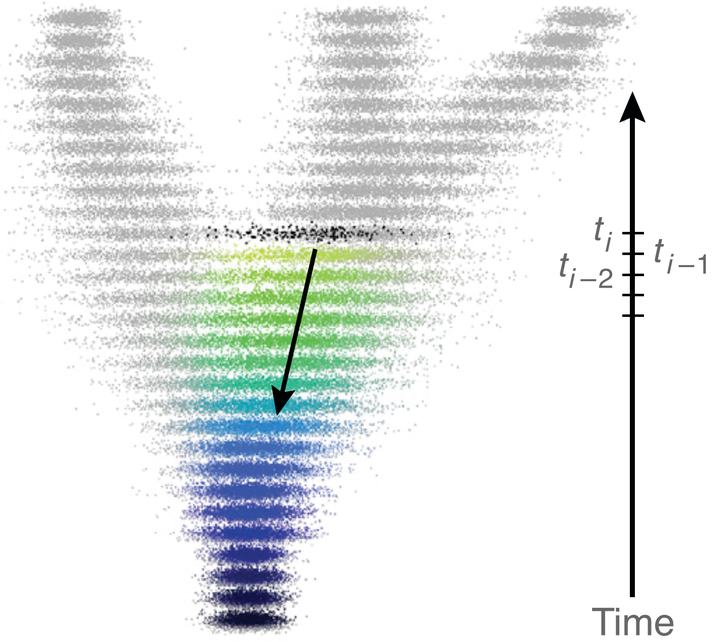


Schiebinger *et al.* Cell (2019)

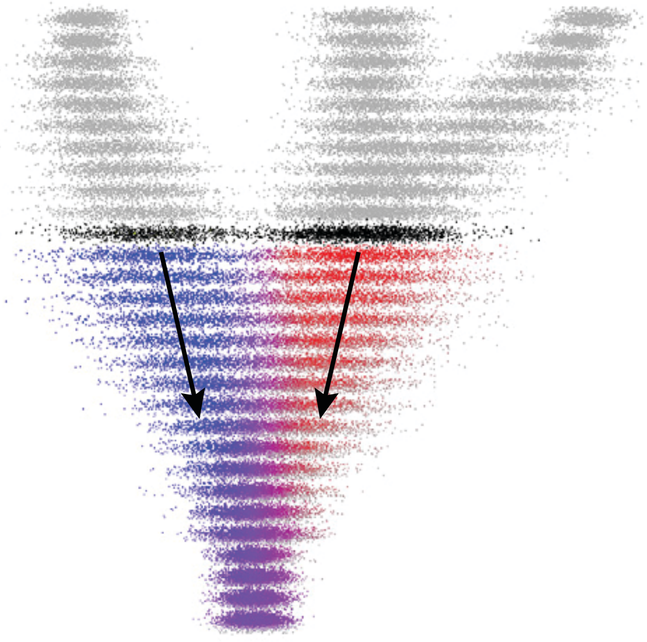
OT estimate trajectories between time points



Descendants



Ancestors

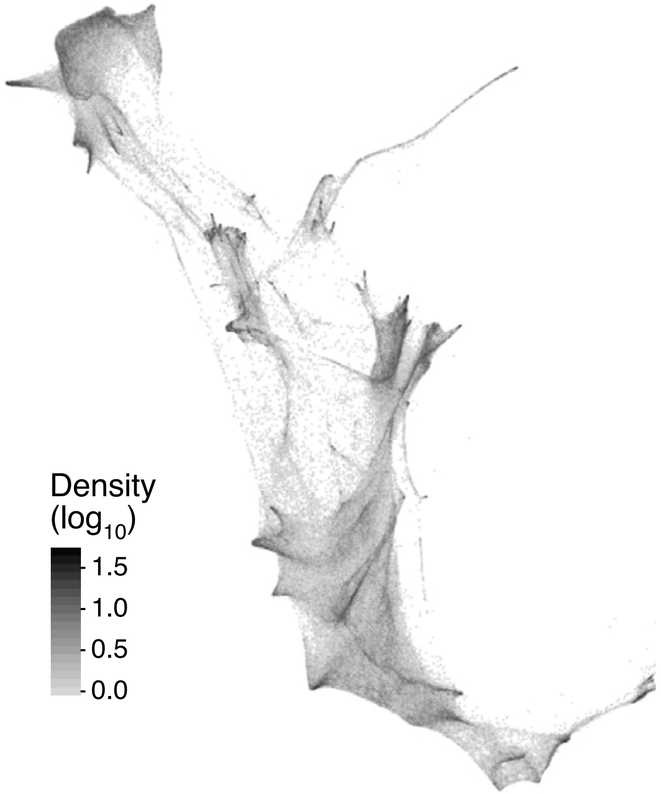


Shared ancestry

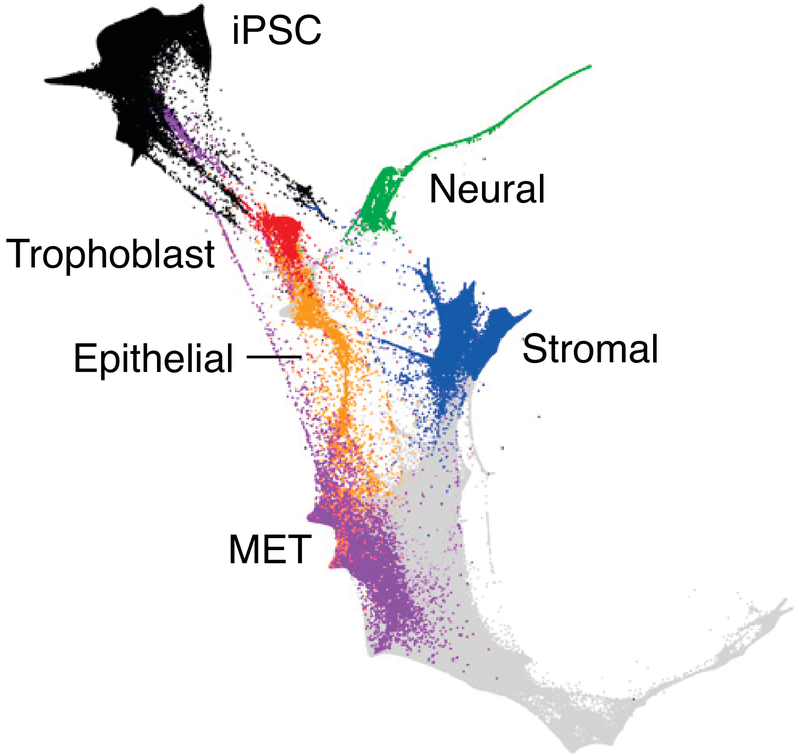
Schiebinger *et al.* Cell (2019)

OT to reconstruct a developmental trajectory

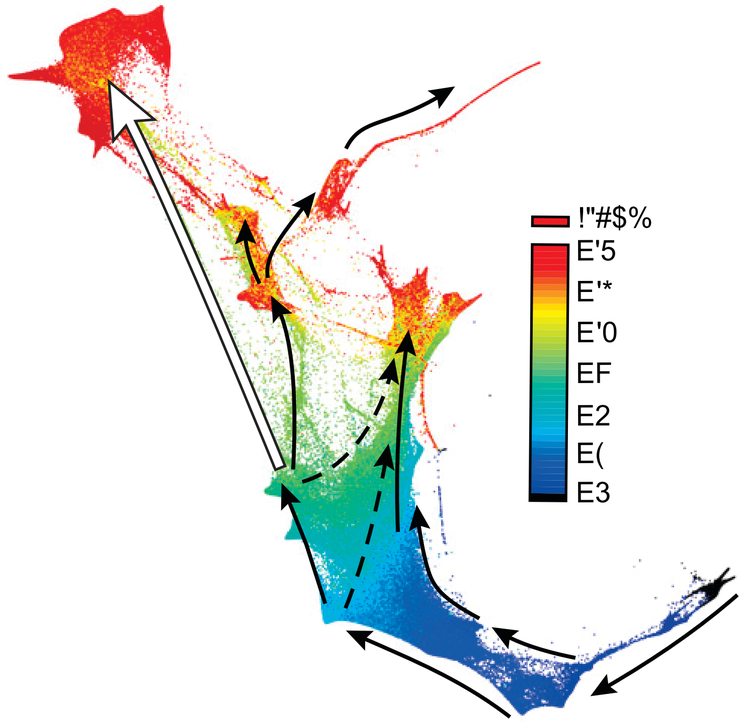
Observed marginal density



Cell type annotation



Trajectory inference



J

Schiebinger *et al.* Cell (2019)

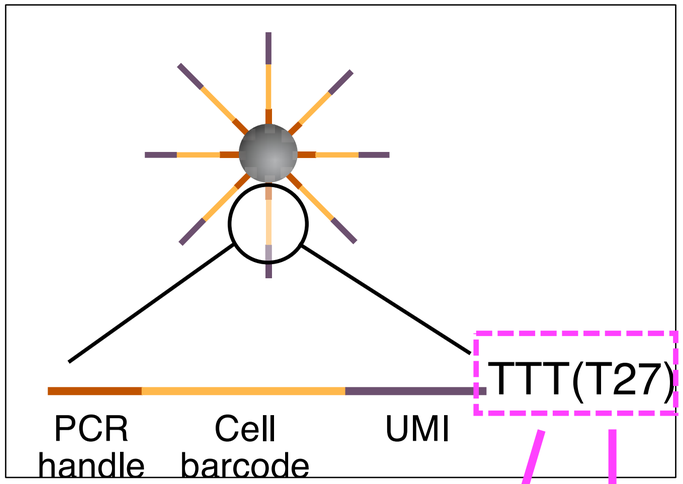
Today’s lecture: Model-based Data Analysis

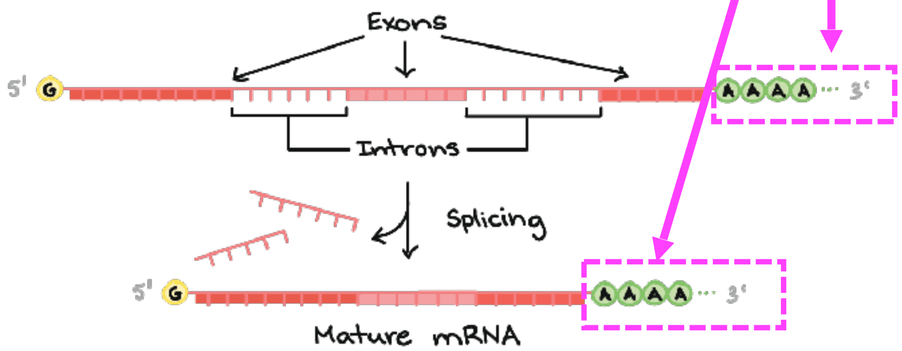
- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

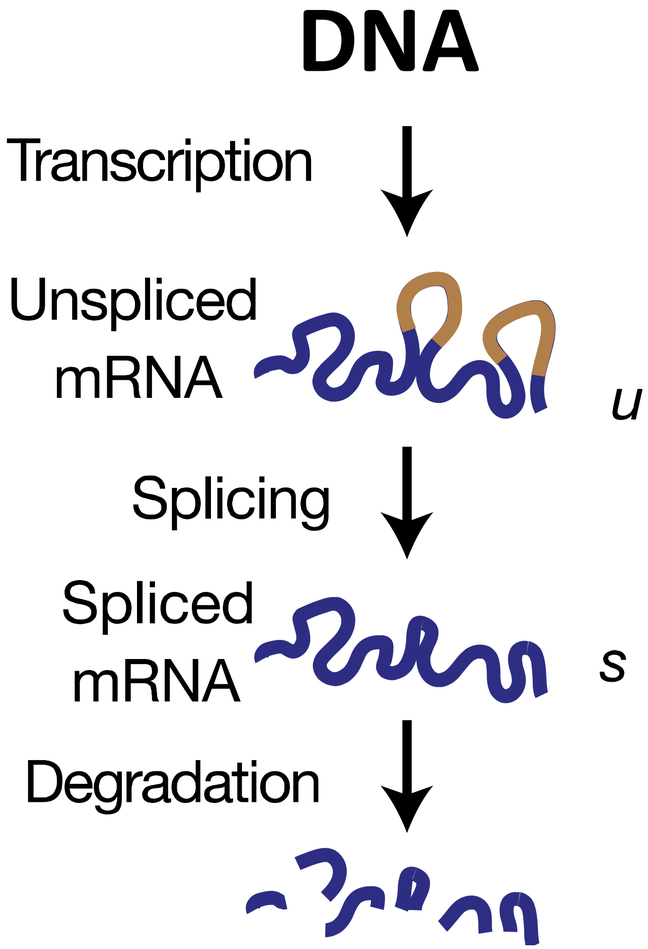
Capturing the dynamics of gene expression regulation: pre mRNA → mature mRNA

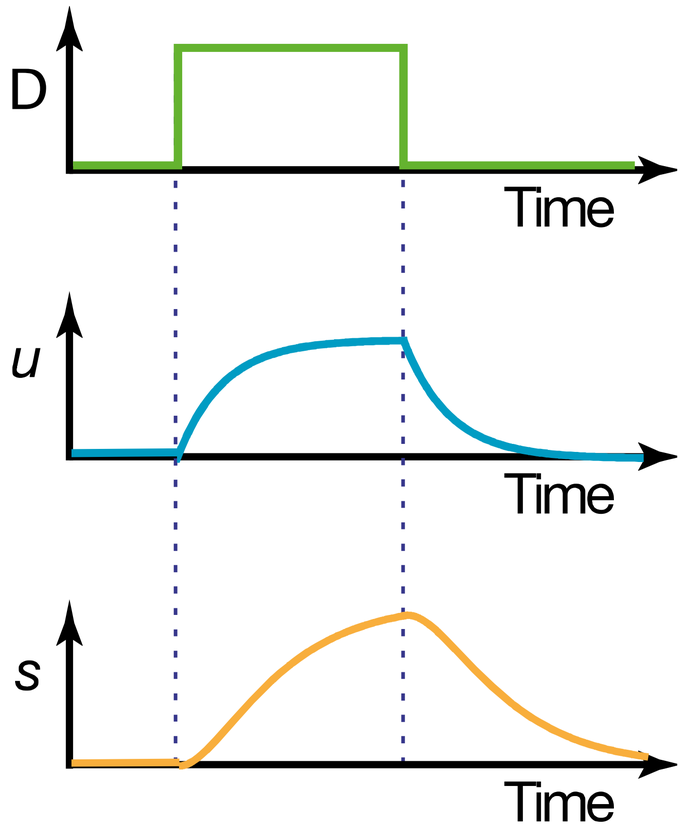
poly-A/T for both U & S





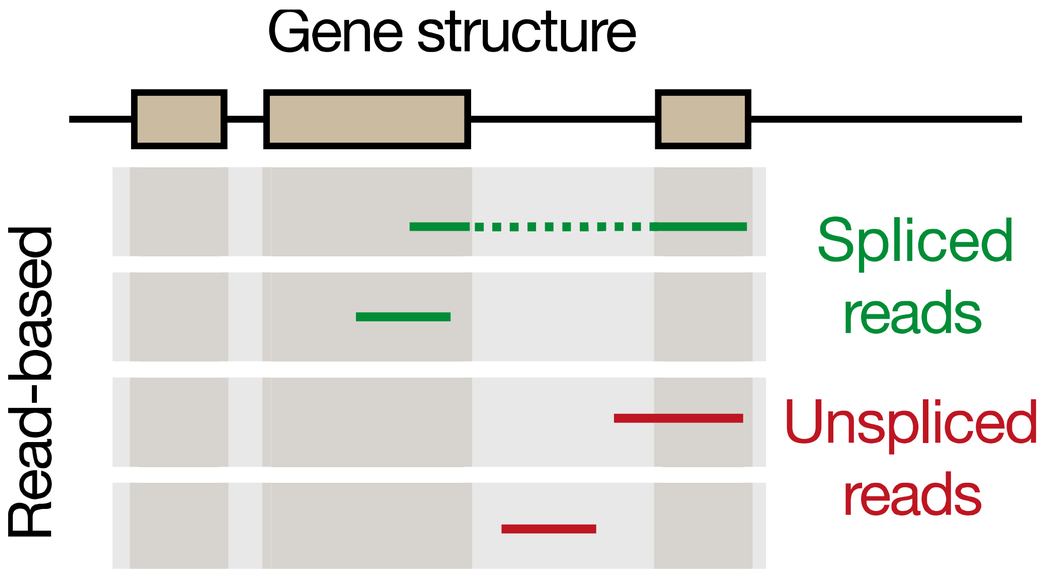
La Manno *et al*. Nature (2018)





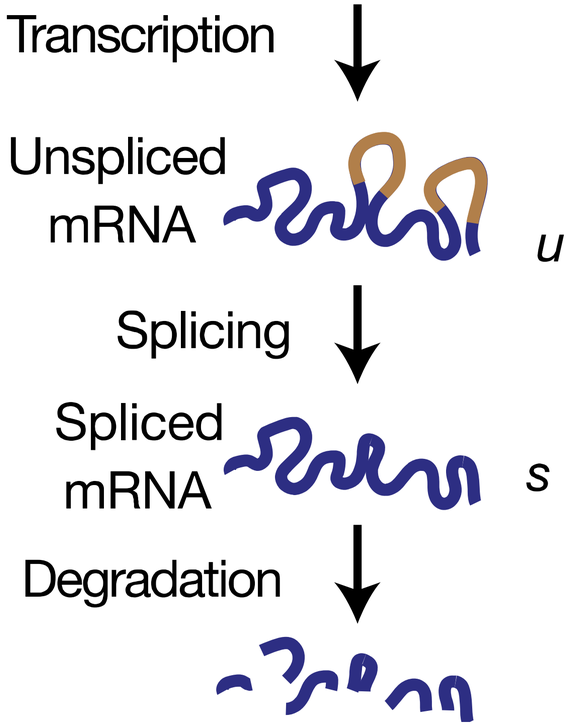
Goal: mRNA dynamics (ODEs)

Estimate the state of each gene in each cell by counting the number of spliced/unspliced



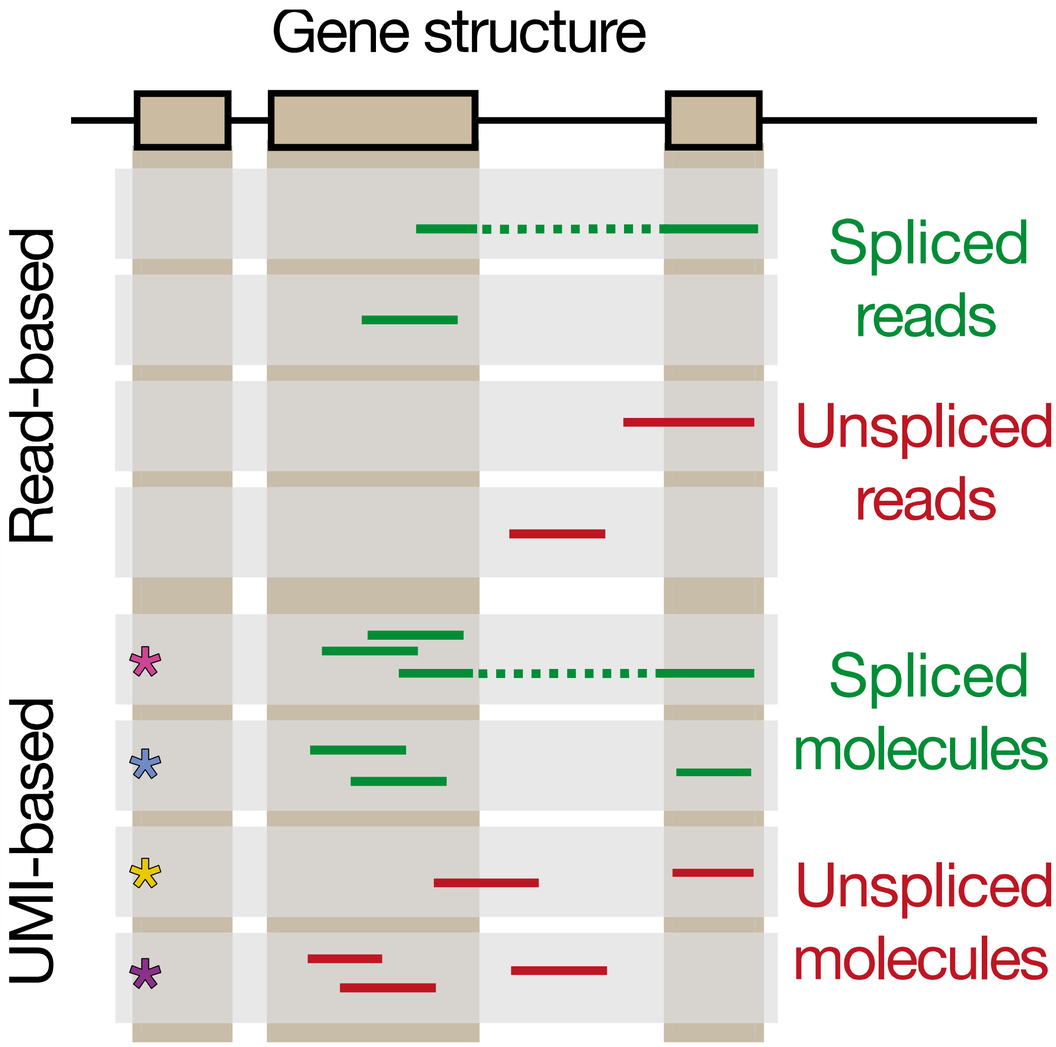
Unspliced: short reads landed on introns (gap) and the boundary between exon and intron.

Spliced: short reads, not including any intronic regions

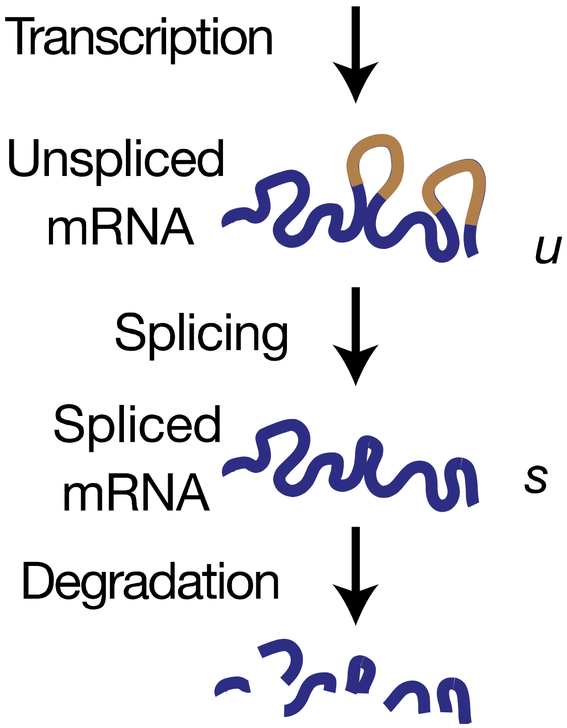


La Manno *et al*. Nature (2018)

Estimate the state of each gene in each cell by counting the number of spliced/unspliced

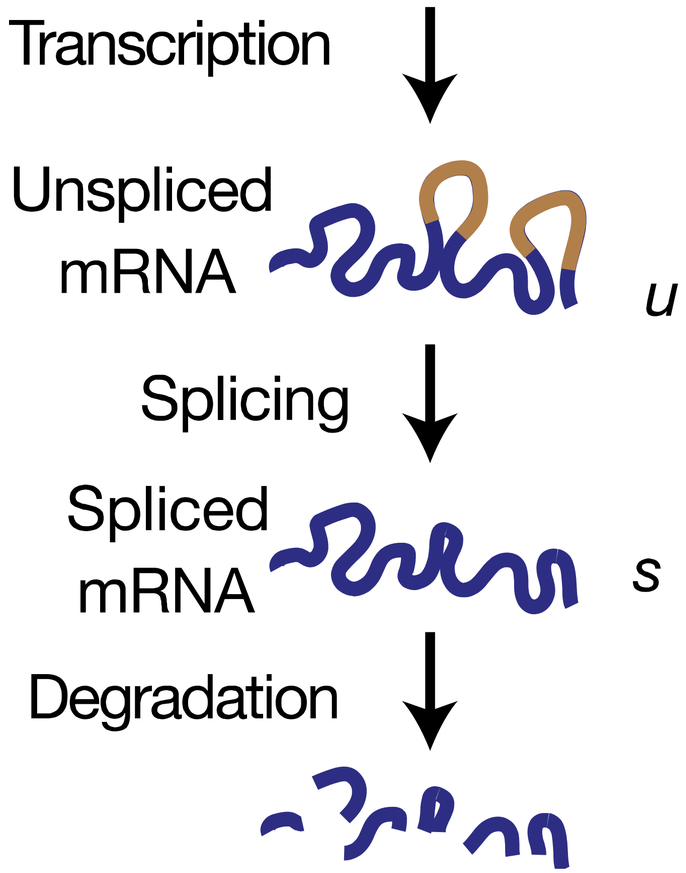


With UMI, we count the number of unique molecules (we could have multiple reads per UMI)



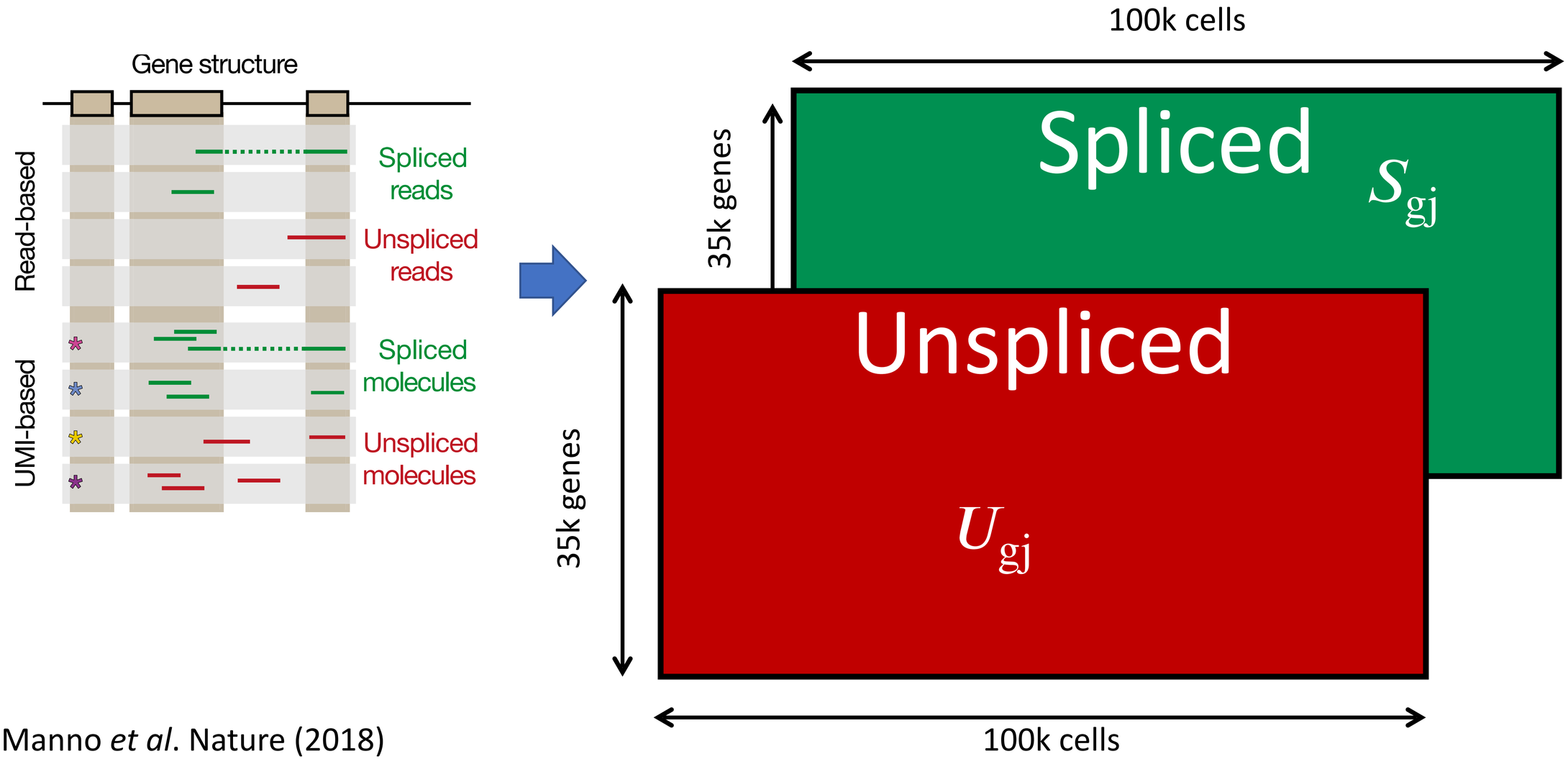
La Manno *et al*. Nature (2018)

Goal: How can we estimate RNA velocity?



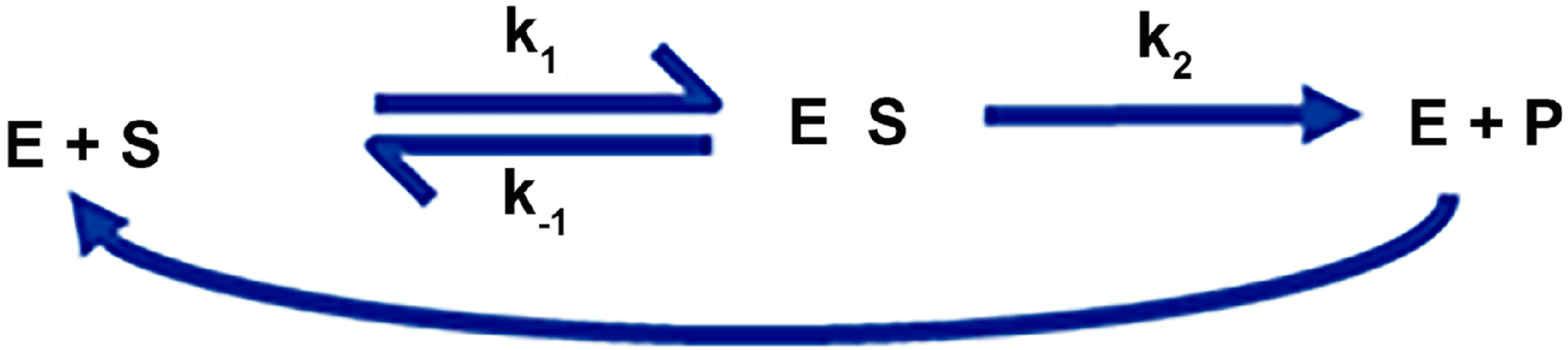
1. Define RNA velocity
   - transcription speed
2. What will be the goal?
   1. Directional information
   2. Measure reaction rate parameters

What are the data for RNA velocity problem?



La Manno *et al*. Nature (2018)

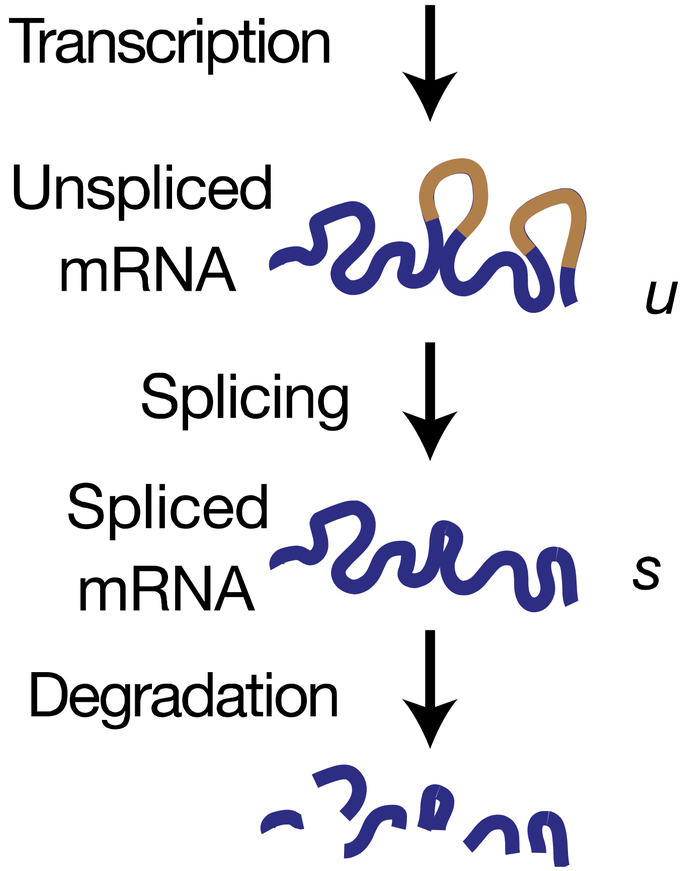
Have we seen a similar problem before?



Mass action law defines how a chemical reaction takes place.

History: Voit, Martens, Omholt (2015)

Modelling splicing dynamics by mass action law



$$\frac{dU}{dt} = \alpha - \beta U(t)$$

splicing rate

transcription initiation rate

$$\frac{dS}{dt} = \beta U(t) - \gamma S(t)$$

splicing rate

mRNA degradation

Phase diagram of RNA velocity

$$\frac{dU}{dt} = \alpha - \beta U(t)$$

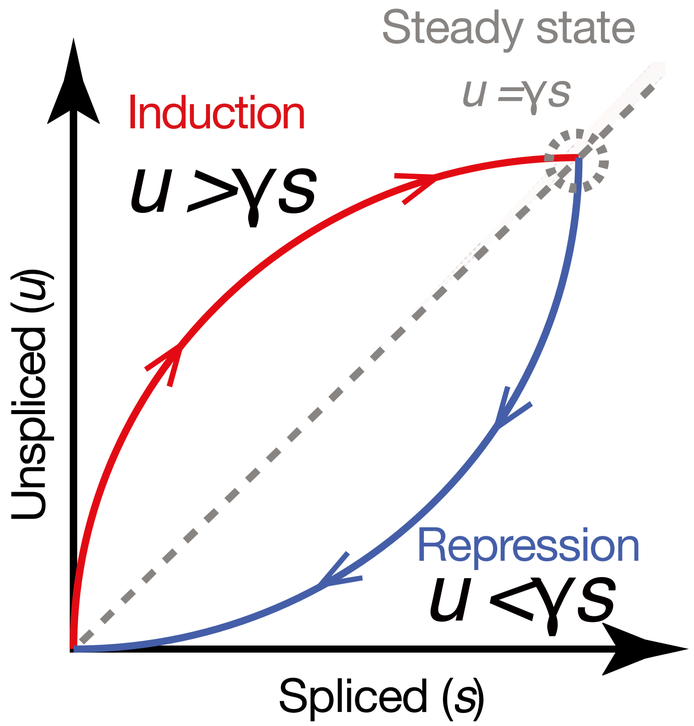
splicing rate

transcription initiation rate

$$\frac{dS}{dt} = \beta U(t) - \gamma S(t)$$

splicing rate

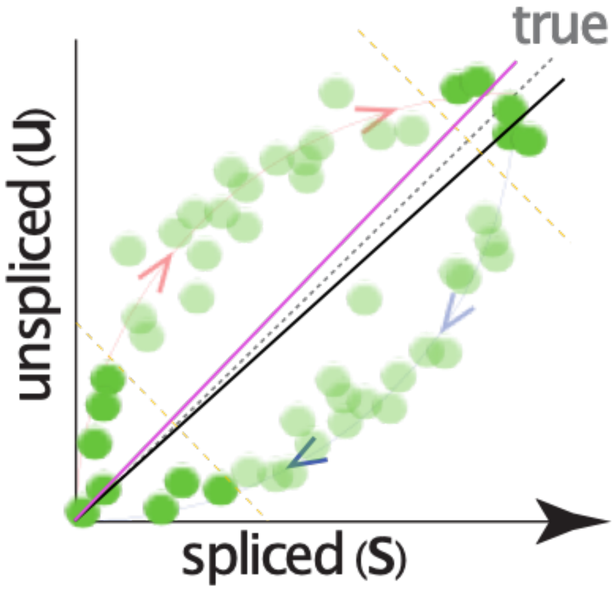
mRNA degradation



La Manno *et al*. Nature (2018)

How can we estimate the model parameters for each gene?

If scRNA-seq profiled all the steady state for this gene (each point = cell)



A quick solution can be derived by least-square estimates

$$u_{gj} \sim s_{gj}\,\gamma_g$$

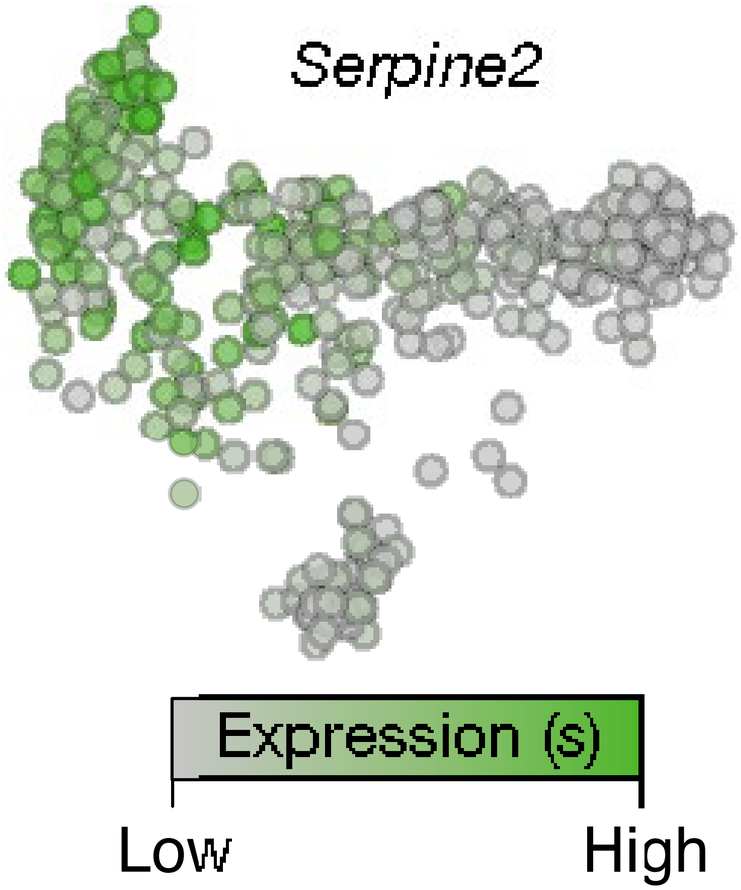
$$\nabla x_{gj} \leftarrow u_{gj} - \hat{\gamma}_g s_{gj}$$

$$x_{gj}(\Delta t) \leftarrow x_{gj} + \nabla x_{gj}\Delta t$$

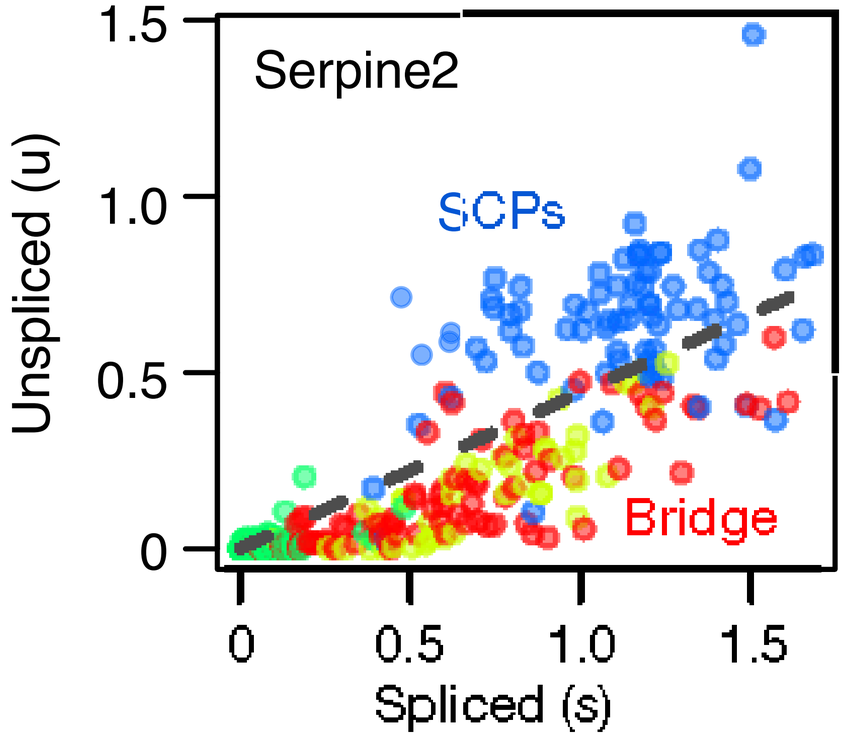
La Manno *et al*. Nature (2018)

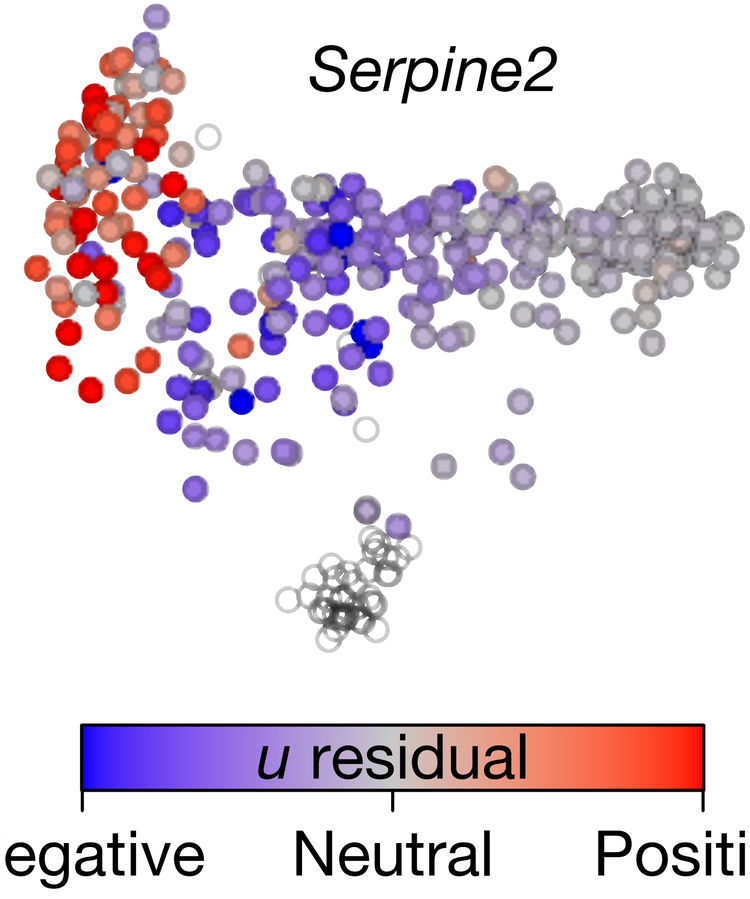
La Manno *et al*. Nature (2018)

Assign cells within the phase diagram



$U_j \sim S_j \gamma$

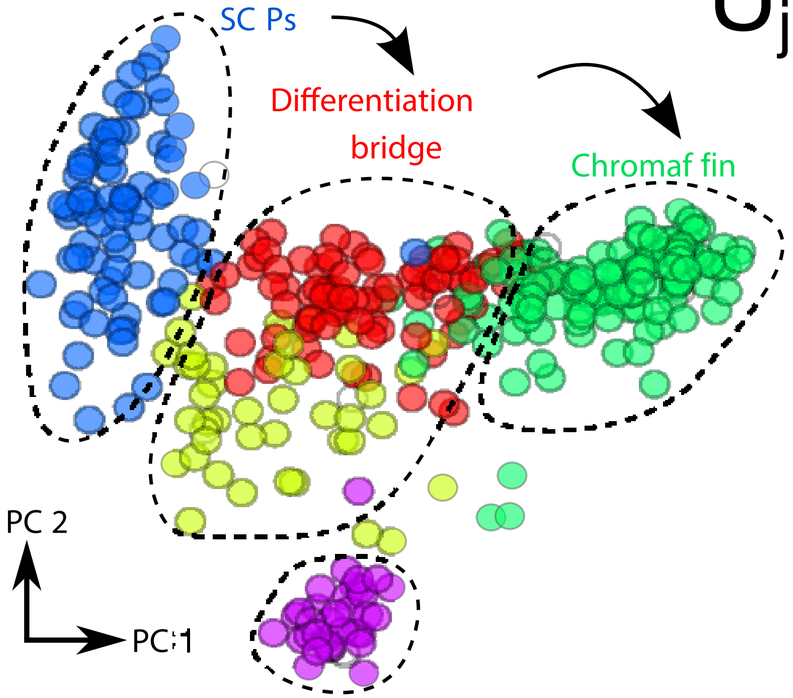




Gene-level velocity → Vector field

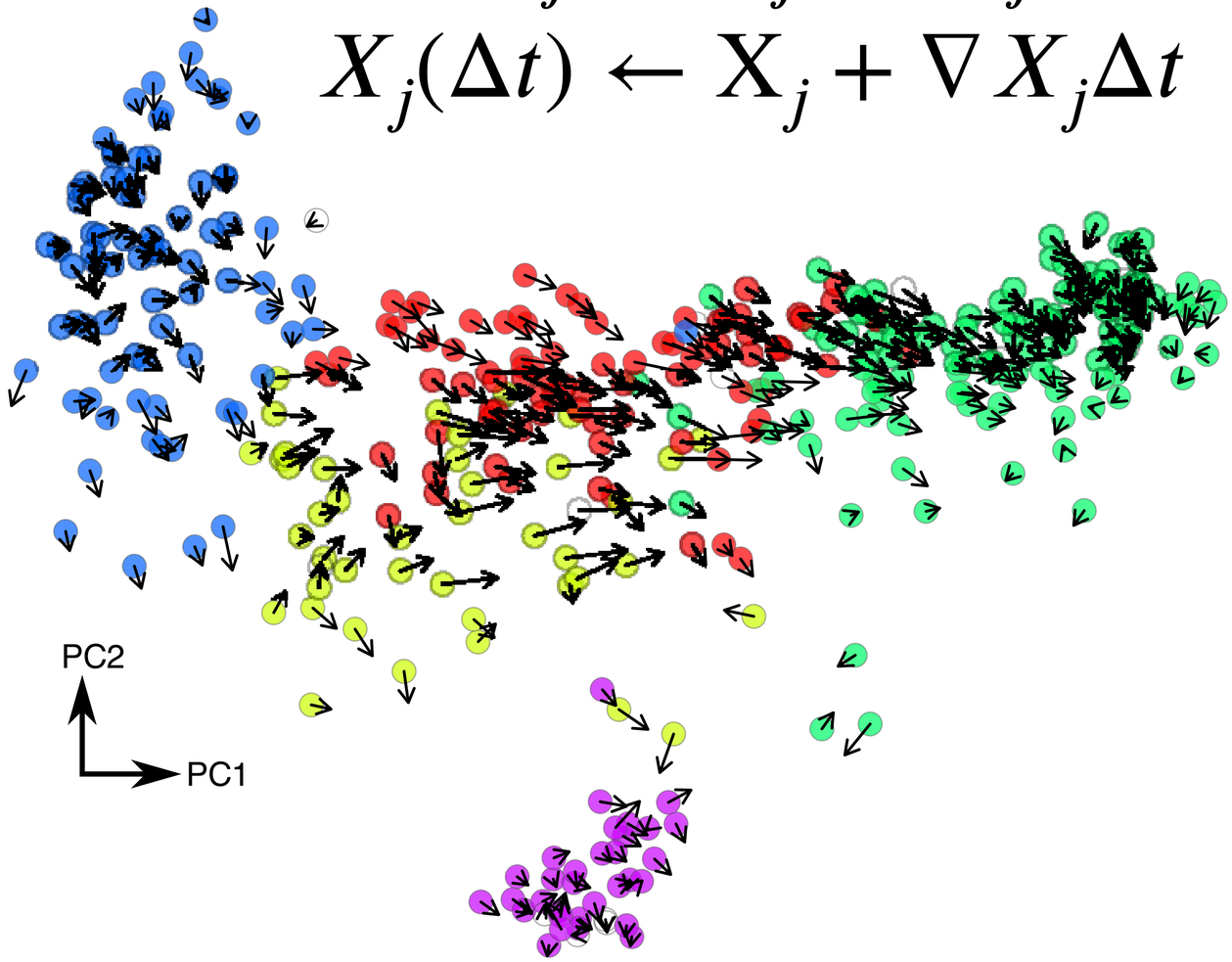
Estimate γ for the aggregate U and S

$$U_j \sim S_j \gamma$$



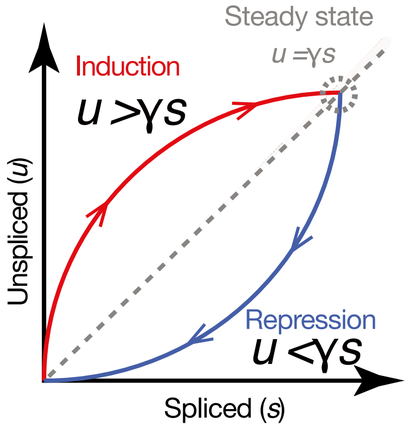
$$\nabla X_j \leftarrow U_j - \hat{\gamma} S_j$$

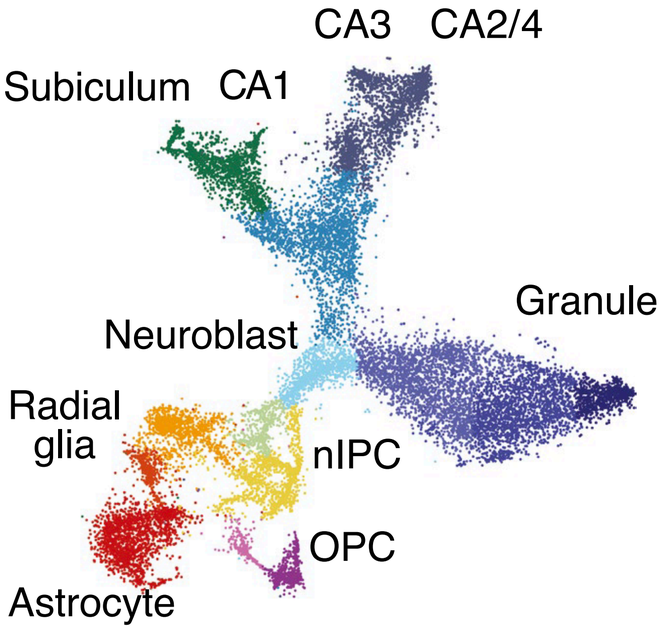
$$X_j(\Delta t) \leftarrow X_j + \nabla X_j \Delta t$$

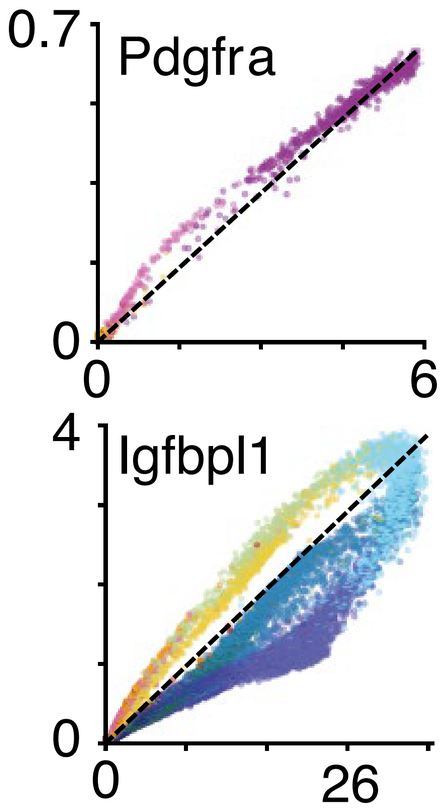


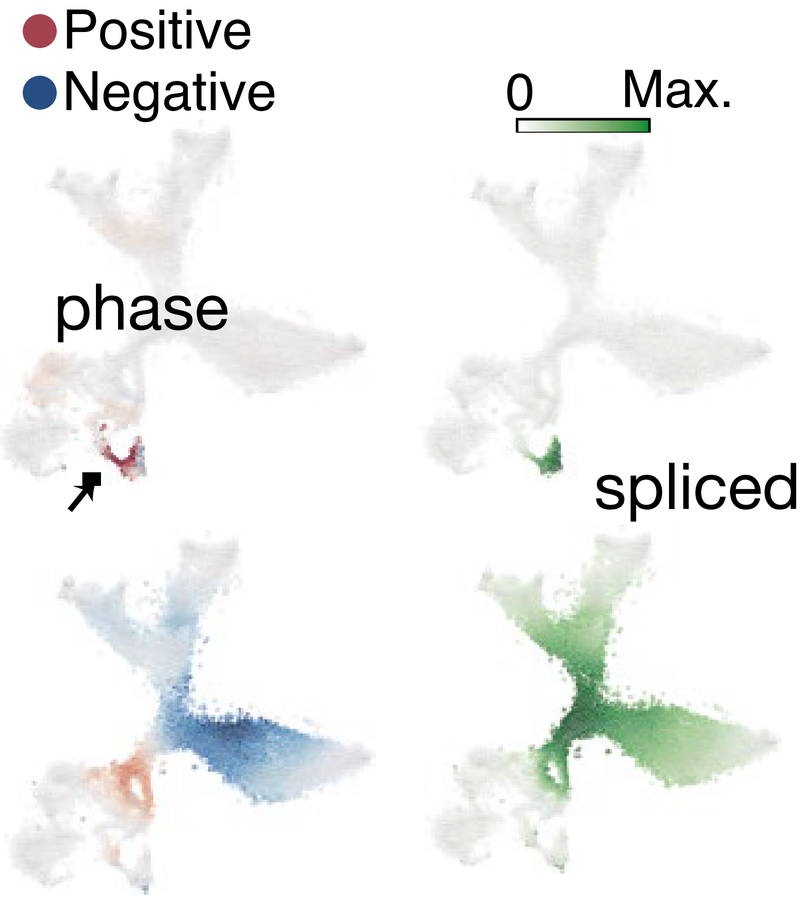
La Manno *et al*. Nature (2018)

Differentiation trajectory ~ RNA velocity









La Manno *et al*. Nature (2018)

However, we can estimate the model parameters in a better way…

transcription initiation rate

splicing rate

$$\frac{dU}{dt} = \alpha - \beta U(t)$$

$$\frac{dS}{dt} = \beta U(t) - \gamma S(t)$$

splicing rate

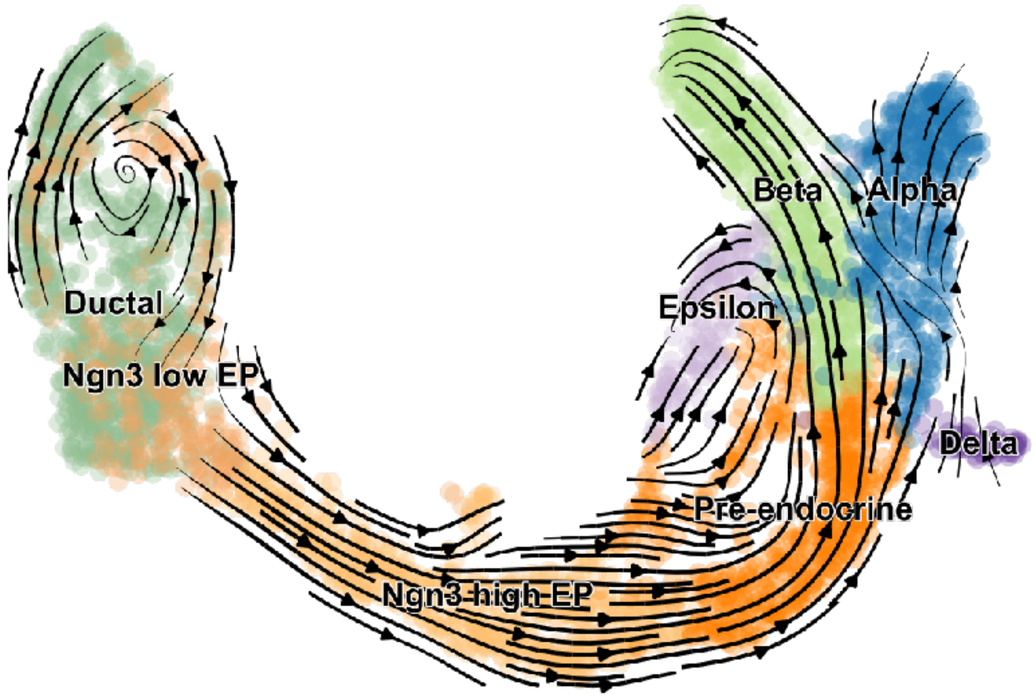
mRNA degradation

Analytical solution:

$$U(t) = U_0 e^{-\beta t} + (\alpha/\beta)(1 - e^{-\beta t})$$

$$S(t) = S_0 e^{-\gamma t} + \frac{\alpha - \beta U_0}{\gamma - \beta}(e^{-\gamma t} - e^{-\beta t})$$

scVelo - RNA velocity generalized through dynamical modeling



Bergen *et al.* 2020

$$U(t) = U_0 e^{-\beta t} + (\alpha/\beta)(1 - e^{-\beta t})$$

$$S(t) = S_0 e^{-\gamma t} + \frac{\alpha - \beta U_0}{\gamma - \beta}(e^{-\gamma t} - e^{-\beta t})$$

EM-algorithm

- E-step: Estimate latent time for each data point
- M-step: Optimize the ODE parameters

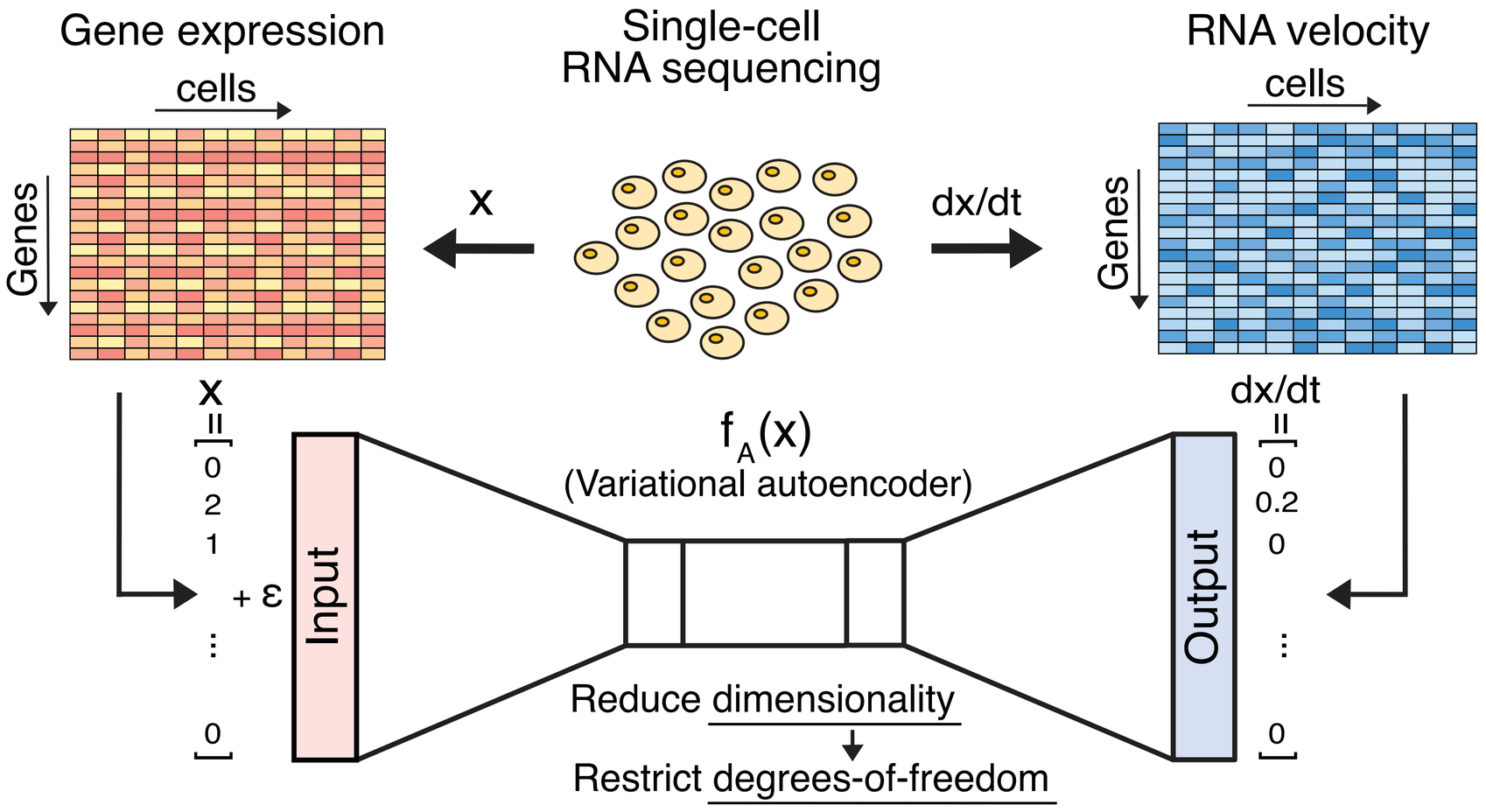
https://scvelo.readthedocs.io/en/stable/

Today’s lecture: Model-based Data Analysis

- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

DeepVelo: Directly learn f: X → dX/dt



Chen *et al.* Science Adv. (2022)

DeepVelo: supervised learning of velocity estimates

MATERIALS AND METHODS

Data collection and preprocessing

For preprocessing and computing RNA velocities, we followed the procedure recommended by *scVelo* (*31*). We selected the top

(...)

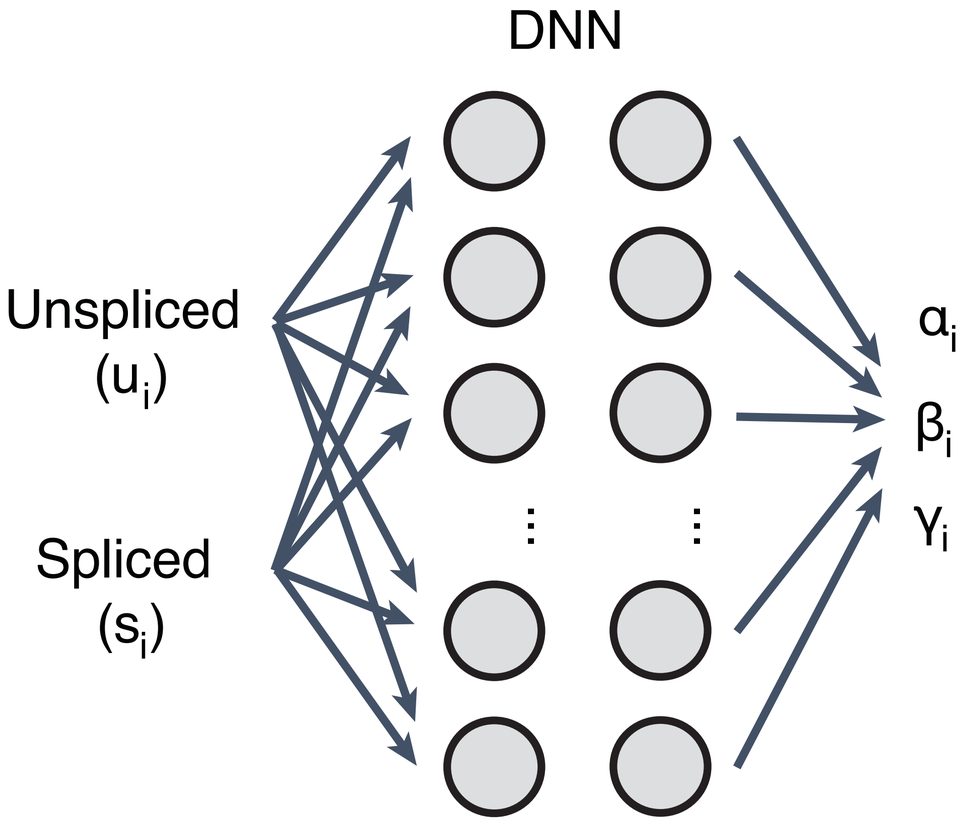
(using the *scv*. *pp*. *moments* function in *scVelo*). After recovering the dynamics using the moments, RNA velocity was computed with the generalized dynamical model from the raw normalized reads (using

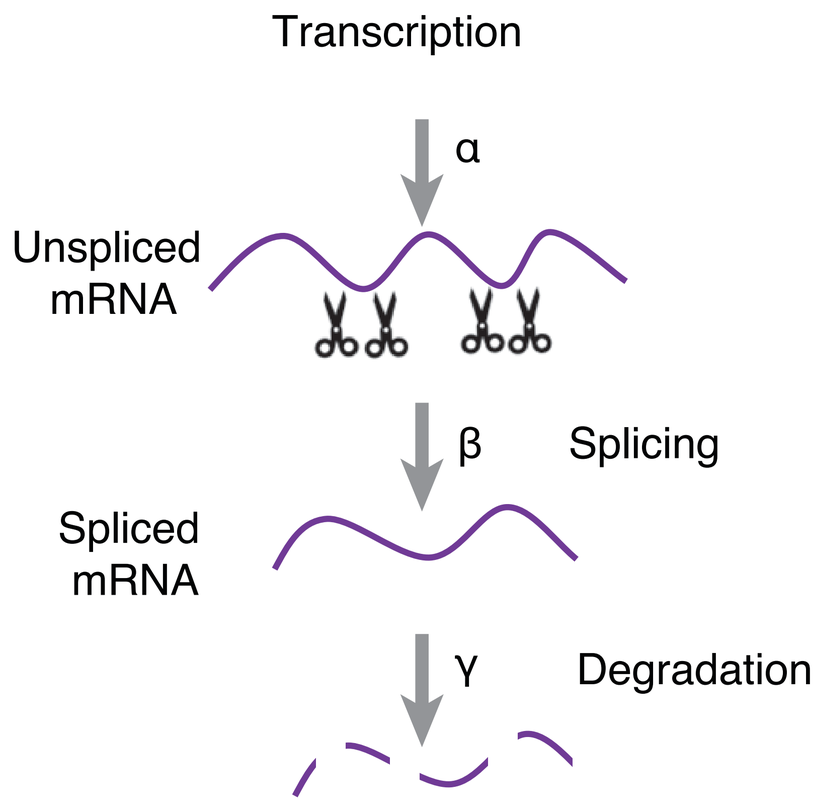
(...)

Only the velocity genes were used as features for the neural ODE. The numbers of cells and available velocity genes for each dataset are shown in Table 1.

Chen *et al.* Science Adv. (2022)

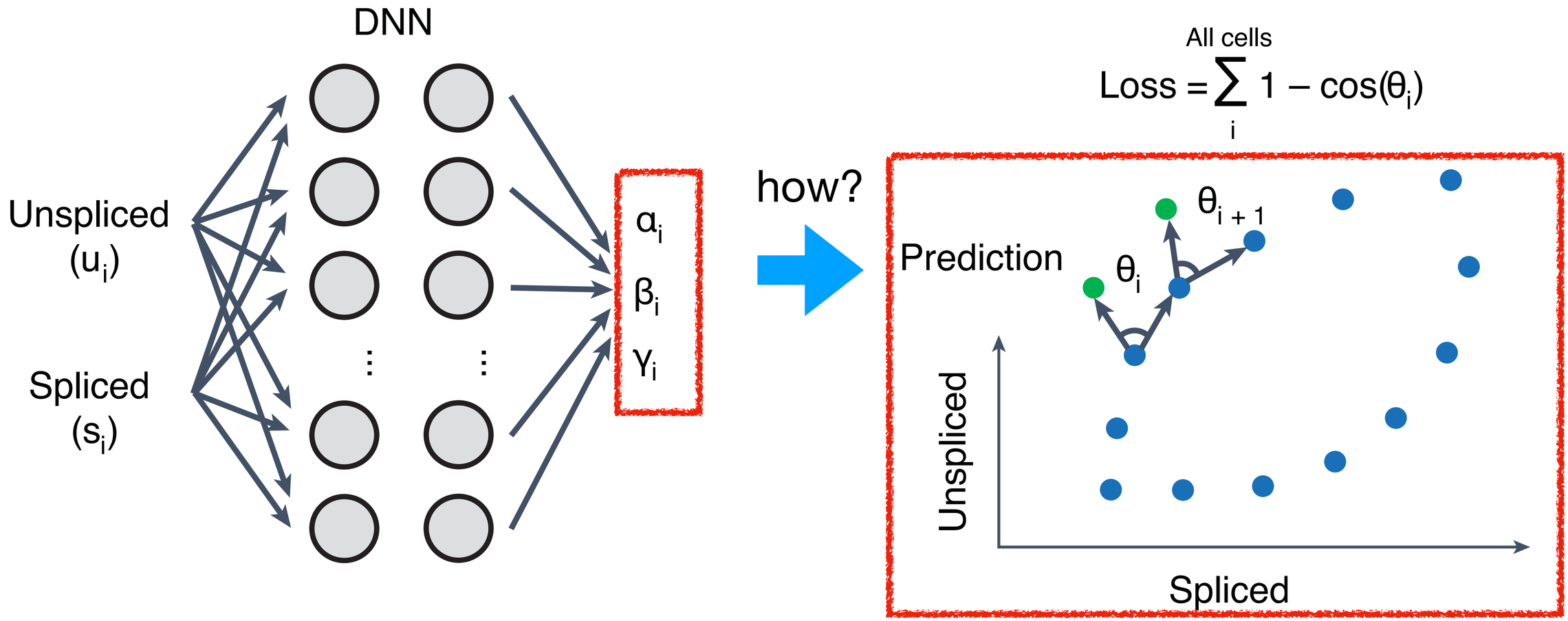
cellDancer: supervised learning of the rate parameters





Li *et al.* Nature Biotech. (2023)

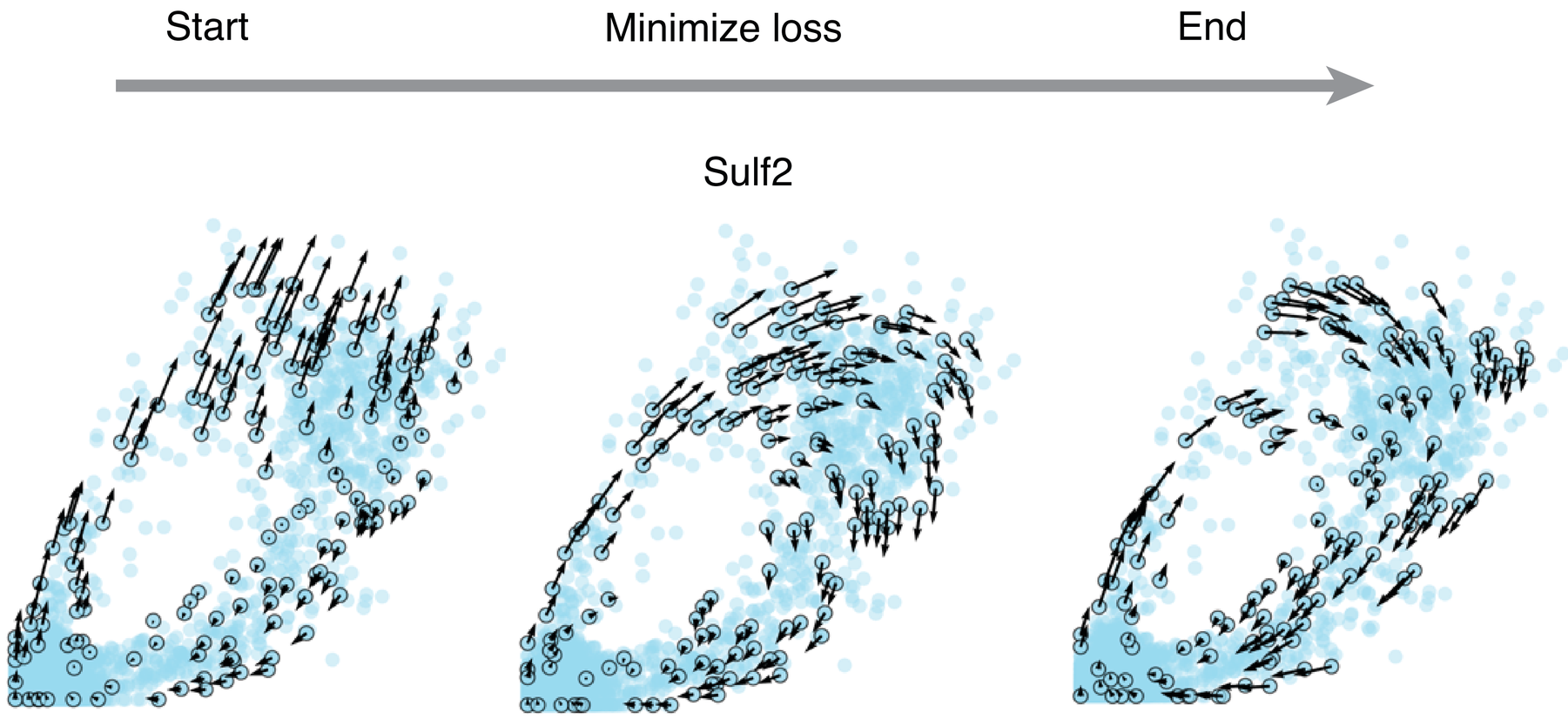
cellDancer: supervised learning of the rate parameters



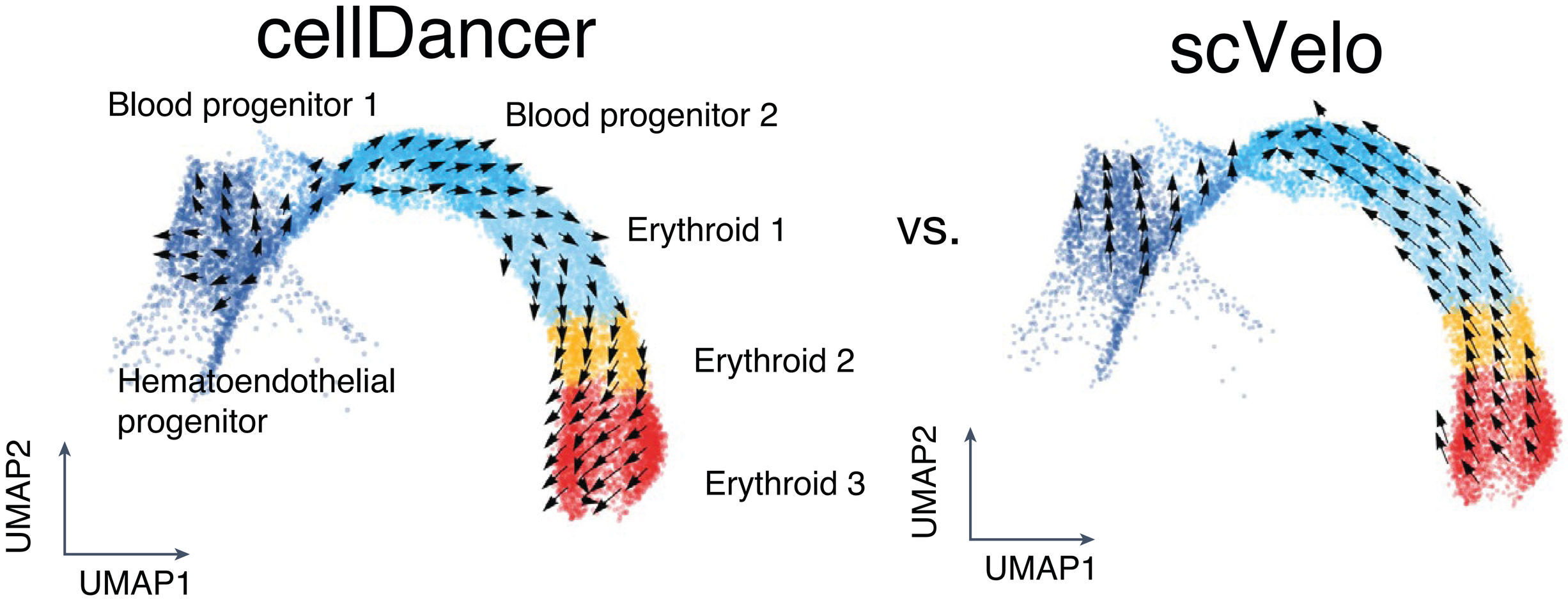
prediction: Deep Neural Network → simulate 1 step

observed outcome: nearest neighbour cells

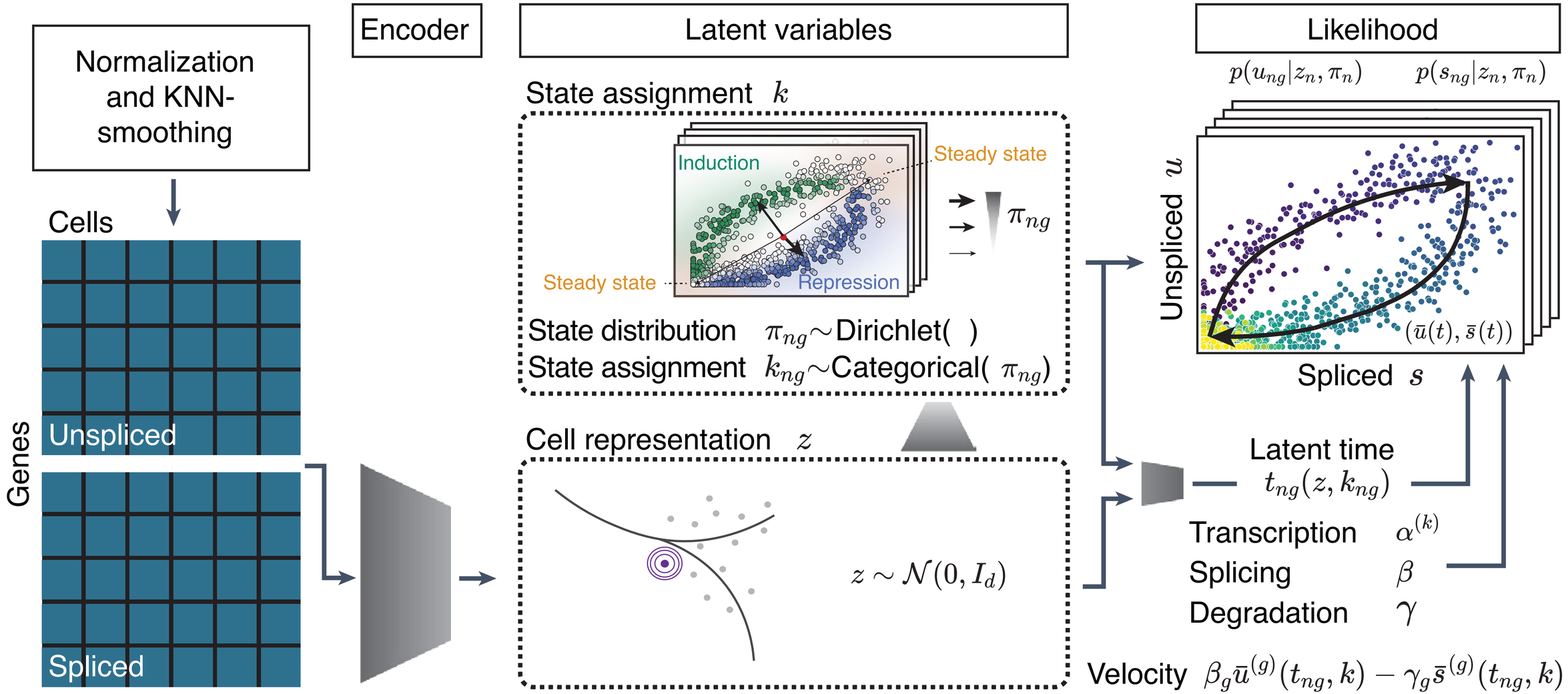
cellDancer adaptively optimizes ODE models



cellDancer fine-tunes scVelo results



veloVI: deep VAE directly models the data likelihood



Gayoso .. Theis, Yosef, *Nature Methods* (2023)

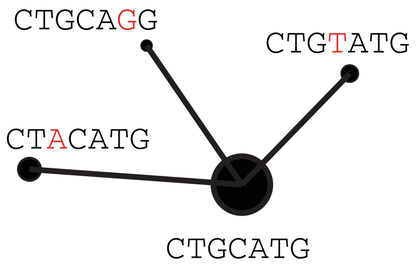
Today’s lecture: Model-based Data Analysis

- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

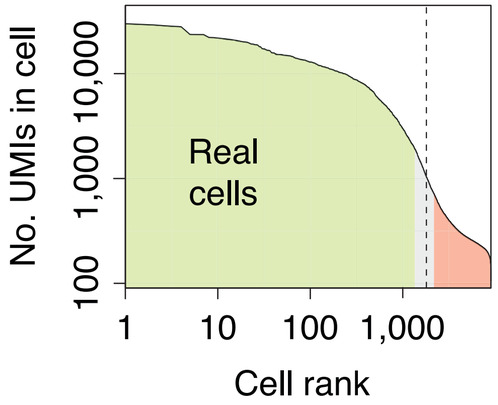
A general workflow of single-cell data anlsysis

Alignment and molecular counting



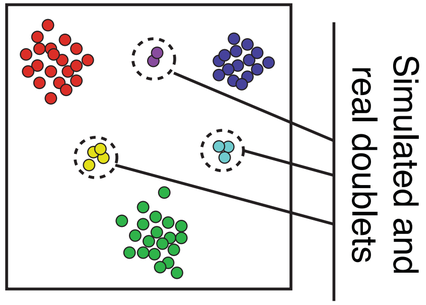
Estimate transcript abundance, correct barcode sequencing errors

Cell filtering and quality control



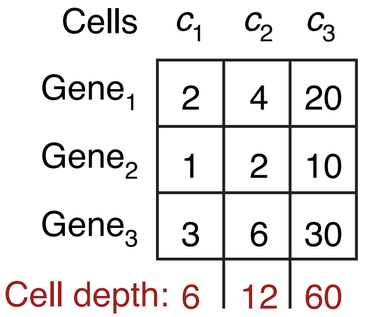
Distinguish empty droplets/barcodes, dying cells, outliers

Doublet scoring



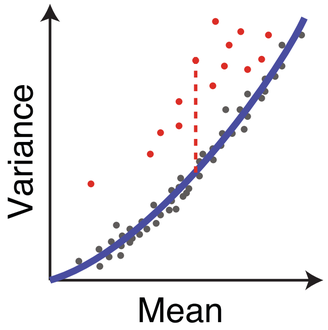
Identify potential doublets resulting from co-encapsu-lation or barcode collisions

Cell size estimation



Estimate effective cell sampling depth

Gene variance analysis

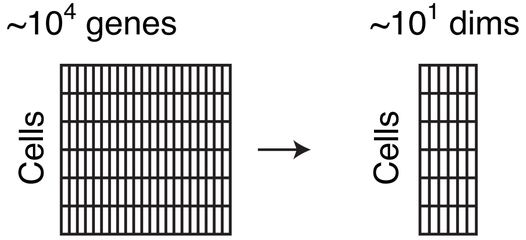


Identify overdis-persed genes, rescale gene variance for down-stream analysis

Karchenko (2021)

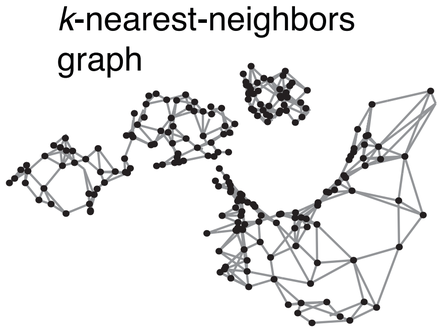
Learning dynamics from single-cell data

Reduction to a medium-dimensional space



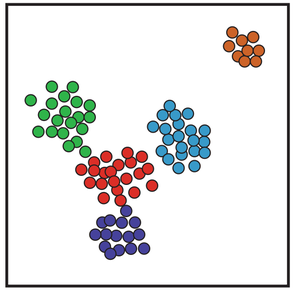
Find most informative set of reduced latent axes (10–50), use it to assess cell–cell similarity

Manifold representation



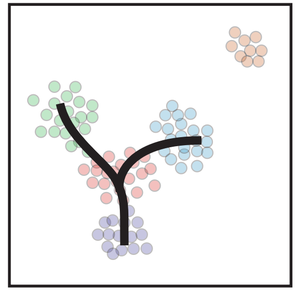
Capture complex, curved arrange-ments of cells in the expression space

Clustering and differential expression



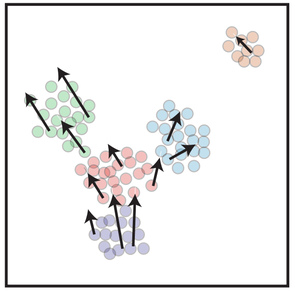
Identify discrete subpopulations of cells, and genes distin-guishing them

Trajectories



Capture contin-uous variation of cell state with trees or curves

Velocity estimation



Predict state of the cells in the near future

Karchenko (2021)

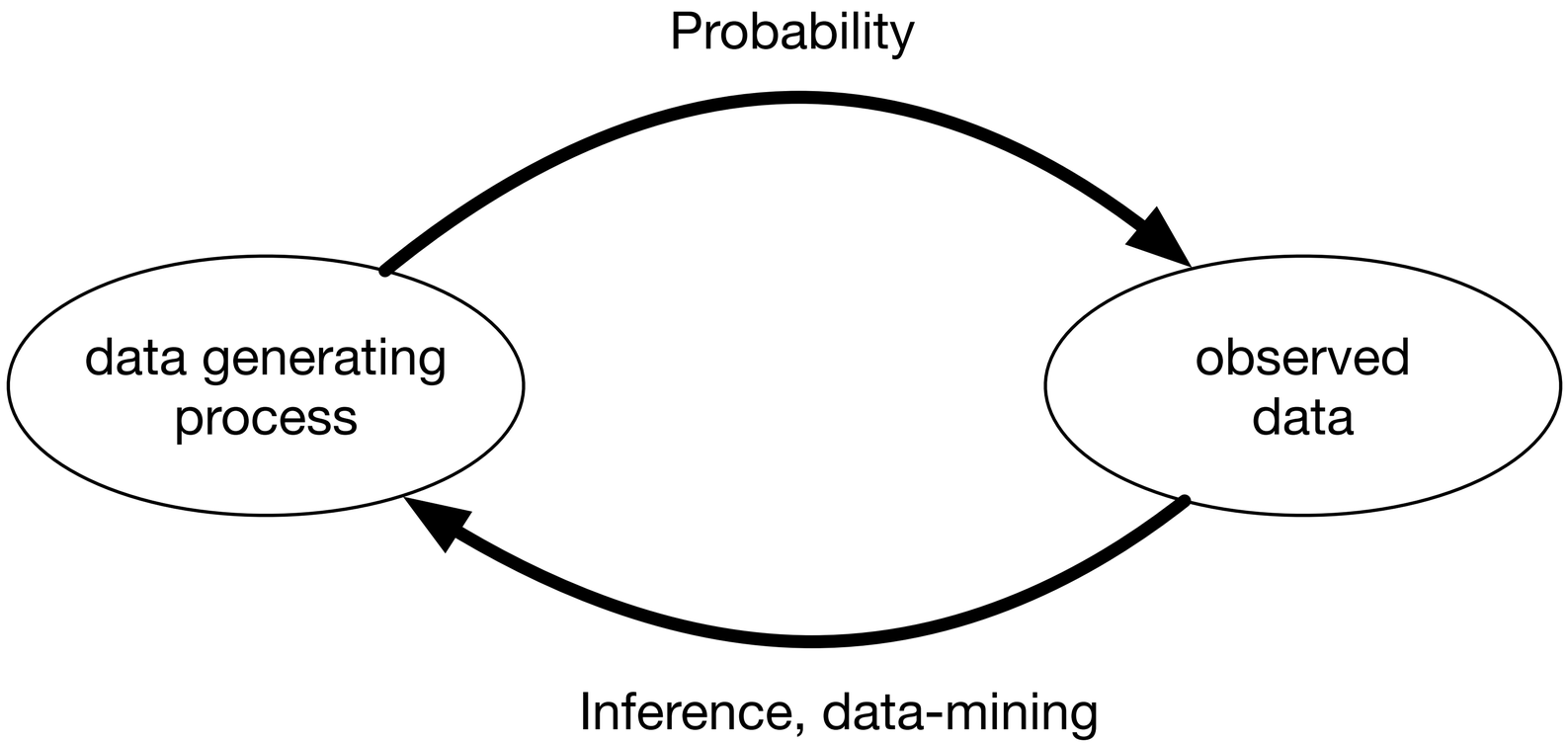
Model-based clustering (if you were a Bayesian statistician...)

Simulation/data generation

Given a data generating process, what are the properties of the outcomes?

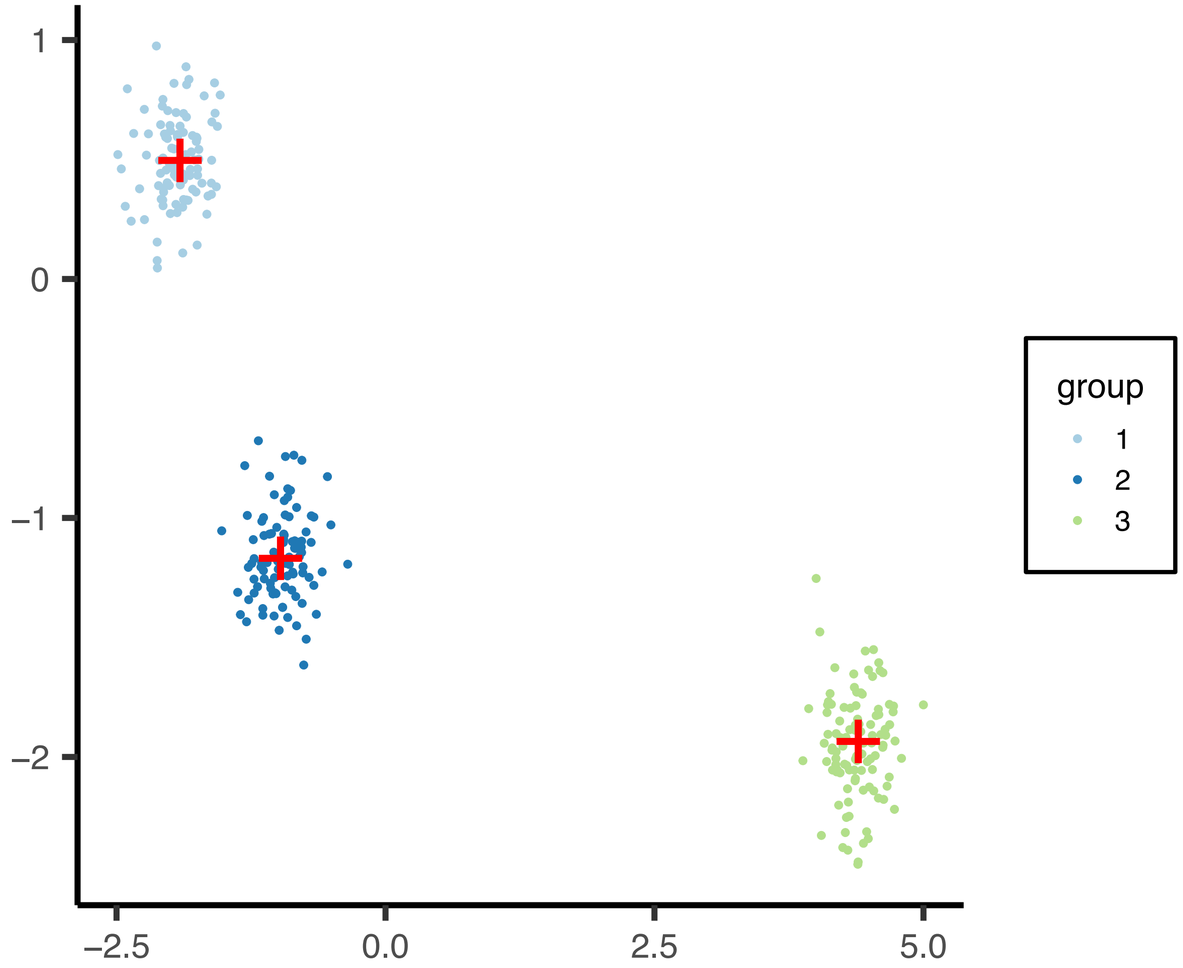
Model inference

Given the outcomes, what can we say about the process that generated the data?



Wassermann, *All of Statistics* (2004)

What do we want to know from data?

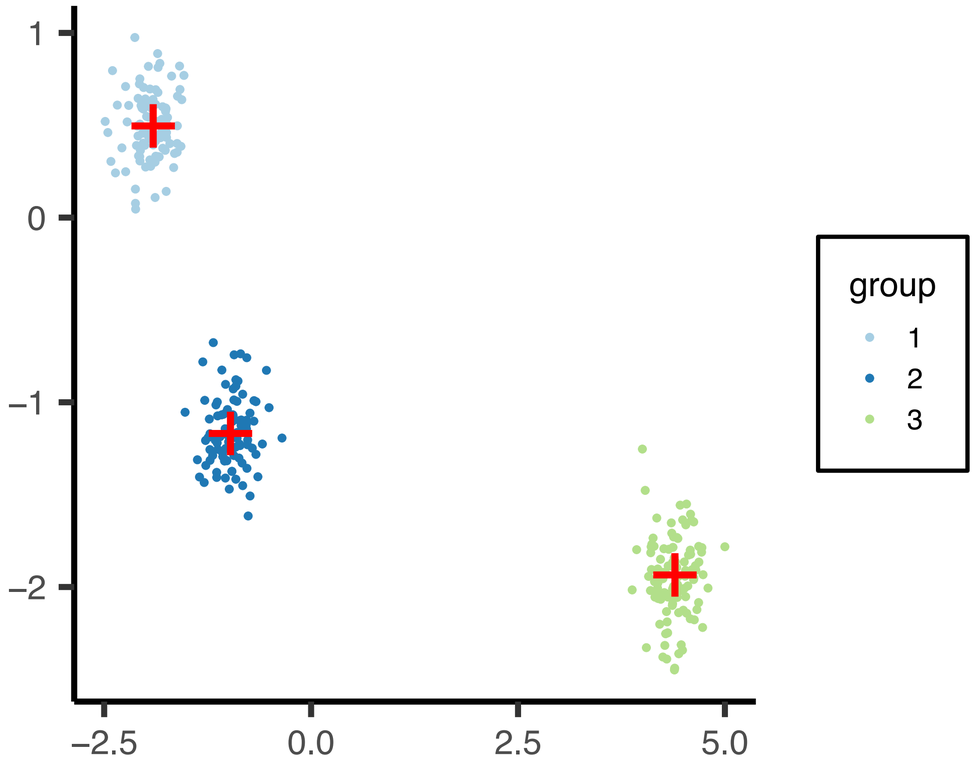


Two goals: recover (1) group membership and (2) the centroids (red marks)

A chicken-and-egg problem: guessing latent membership vs. parametric inference

- If we knew the membership of all the points, we can simply estimate the centre (e.g., taking sample mean within each cluster)
- If we knew the centre coordinates, we would be able to assign points to most probable groups easily based on distance from the centre points.
- Statistical answer: Solve the underlying inference problem.
- To a parameter estimator, the membership assignments are *hidden* (latent).

Let's think about the data-generating process to "reverse" it

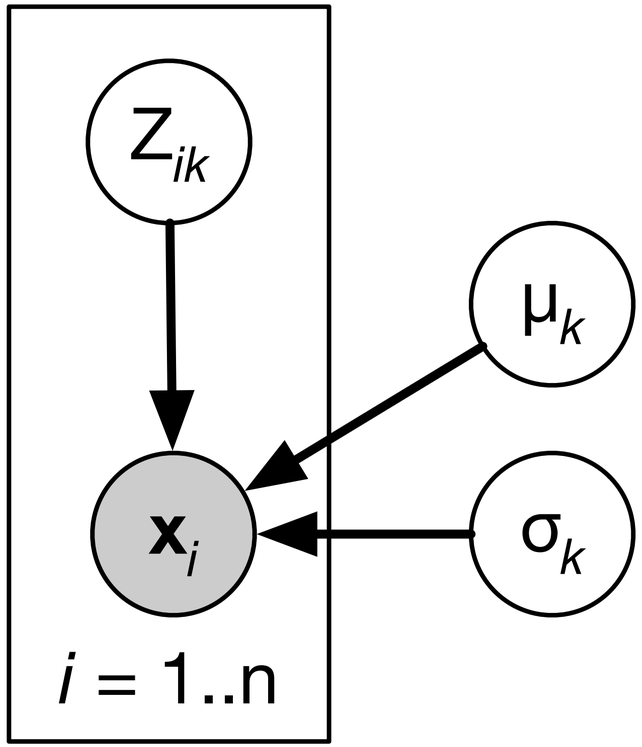


Latent Variable Model

What could have been done to generate observe data?

- Latent membership: Z_{ik}
- Model parameters: μ_k and σ_k

Gaussian Mixture Model (k-means)



How do we infer Z and μ, σ?

GMM data generating process

1. Initialize μ_k (the centre of each group) and σ_k (the spread within each group)
2. Randomly assign group membership, $Z_{ik} = 1$ iff a point i belongs to a group k.
3. Generate: $\mathbf{x}_i | Z_{ik} = 1, \mu_k \sim \mathcal{N}(\mu_k, \sigma^2 I)$

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

Harvard University and Educational Testing Service

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. SILVEY in the Chair]

SUMMARY

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

Keywords: MAXIMUM LIKELIHOOD; INCOMPLETE DATA; EM ALGORITHM; POSTERIOR MODE

Expectation Maximization for GMM MLE

$$J \equiv \log \prod_{i=1}^{n} p(\mathbf{x}_i|\mu, \sigma)$$

Expectation Maximization for GMM MLE

$$
\begin{aligned}
J &\equiv \log \prod_{i=1}^{n} p(\mathbf{x}_i|\mu, \sigma) \\
&= \log \prod_{i=1}^{n} \sum_{Z} p(\mathbf{x}_i|Z, \mu, \sigma) p(Z)
\end{aligned}
$$

* It might be difficult to enumerate all the Z's... so let's introduce some other distributions that will help our "guessing" work, which we call it $q(Z)$

Expectation Maximization algorithm = expected MLE

The goal:

$$\log p(X|\mu,\sigma) = \log \sum_Z p(X,Z|\mu,\sigma) \tag{1}$$

$$\geq_{\text{Jensen}} \sum_Z \underbrace{p(Z|X,\mu,\sigma)}_{\text{E-step}} \underbrace{\log p(X|Z,\mu,\sigma)}_{\text{M-step}} \tag{2}$$

$$= \underbrace{\mathbb{E}_{p(Z|X,\mu,\sigma)}}_{\text{E-step}} \left[\underbrace{\log p(X|Z,\mu,\sigma)}_{\text{M-step}} \right] \tag{3}$$

(Z is discrete, e.g., a membership indicator)

Solution: **Maximize** the lower bound by taking **the expectation** over the posterior probability.

EM algorithm of GMM: E-step

Log-likelihood under some group (μ_k and σ_k):

$$\log p(\mathbf{x}_i|\mu_k, \sigma_k) = \log \mathcal{N}(\mathbf{x}_i|\mu_k, \sigma_k)$$

How to estimate the posterior?

$$p(Z_{ik}|\mathbf{x}_i, \mu, \sigma) = \frac{\exp\{\log p(\mathbf{x}_i|\mu_k, \sigma_k)\}}{\sum_{k'} \exp\{\log p(\mathbf{x}_i|\mu_{k'}, \sigma_{k'})\}}$$

Remark: We can stochastically sample $Z_{ik} = 1$ with the posterior probability.

EM algorithm of GMM: M-step

Maximization step to optimize model parameters

Let this expected lower-bound (ELBO)

$$\mathcal{L}(\mathbf{x}_i; \{\mu_k\}, \{\sigma_k\}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log p(\mathbf{x}_i | Z_{ik}, \mu_k, \sigma_k)$$

Given Z, what are the unknown? We can take gradient steps (e.g., `torch`)

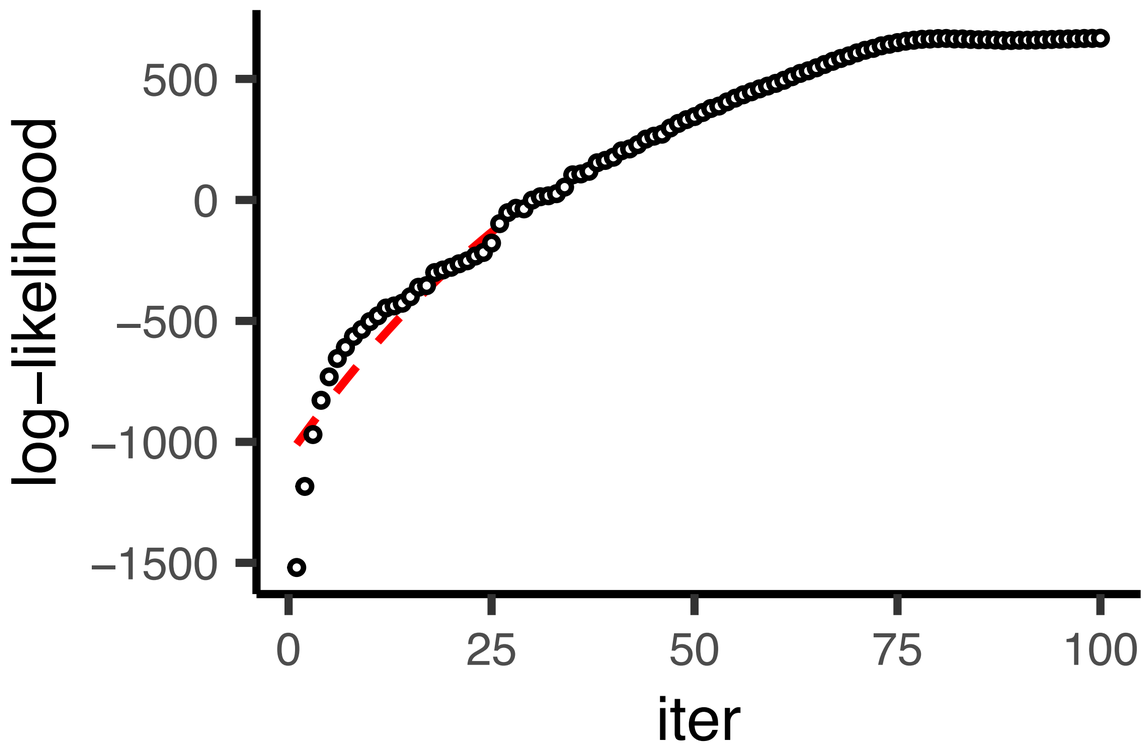
$$\mu_k^{(t)} \leftarrow \mu_k^{(t-1)} + \rho \nabla_{\mu_k} \sum_i \mathcal{L}(\mathbf{x}_i)$$

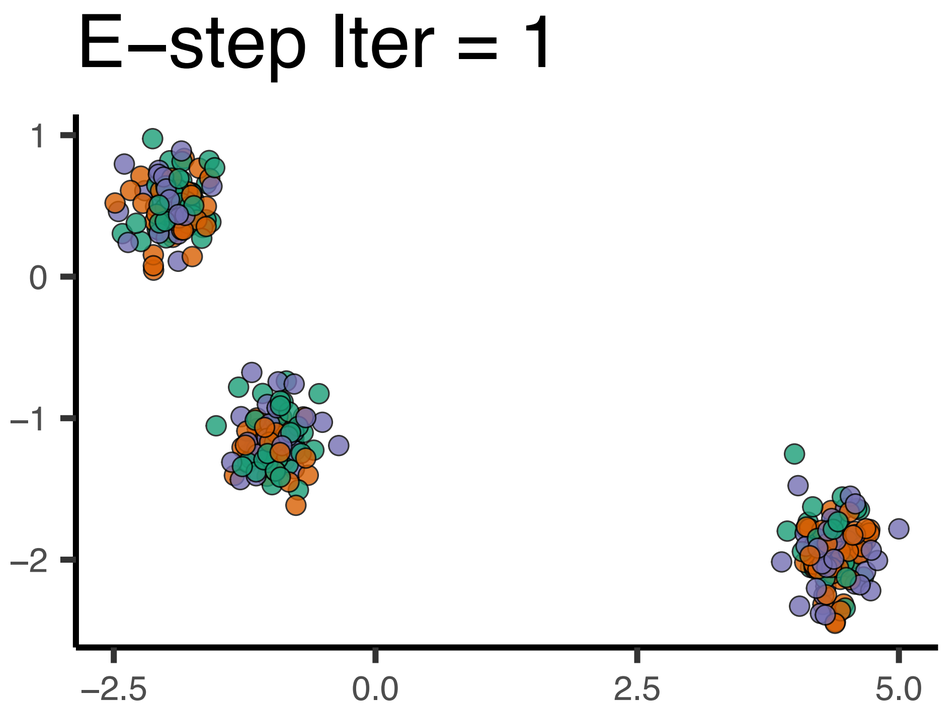
$$\sigma_k^{(t)} \leftarrow \sigma_k^{(t-1)} + \rho \nabla_{\sigma_k} \sum_i \mathcal{L}(\mathbf{x}_i)$$

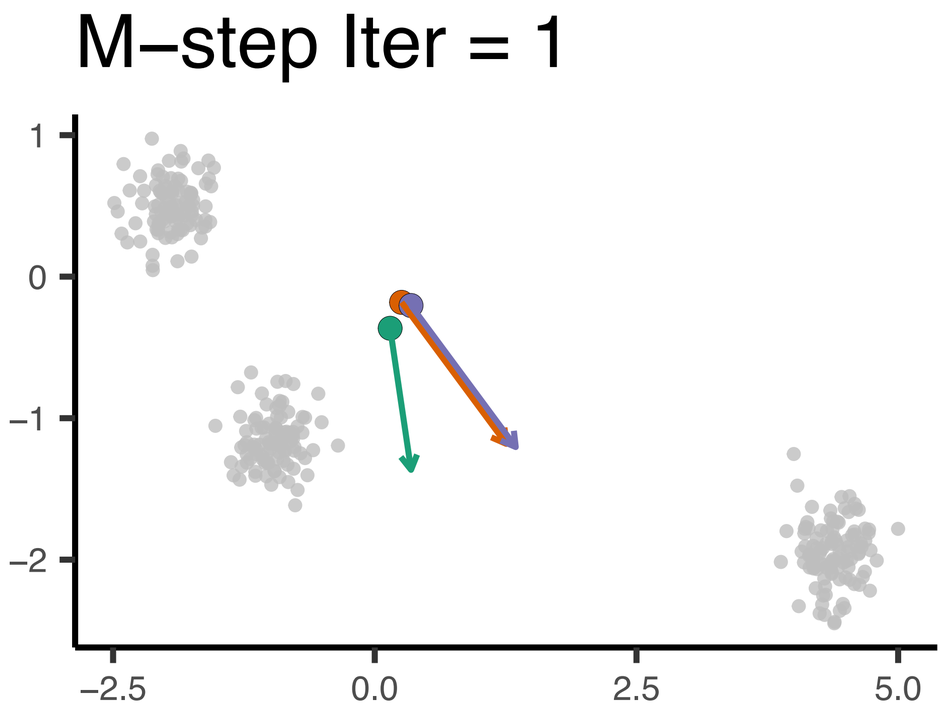
Remark: We have an analytical solution for μ and σ in this example.

Expectation Maximization

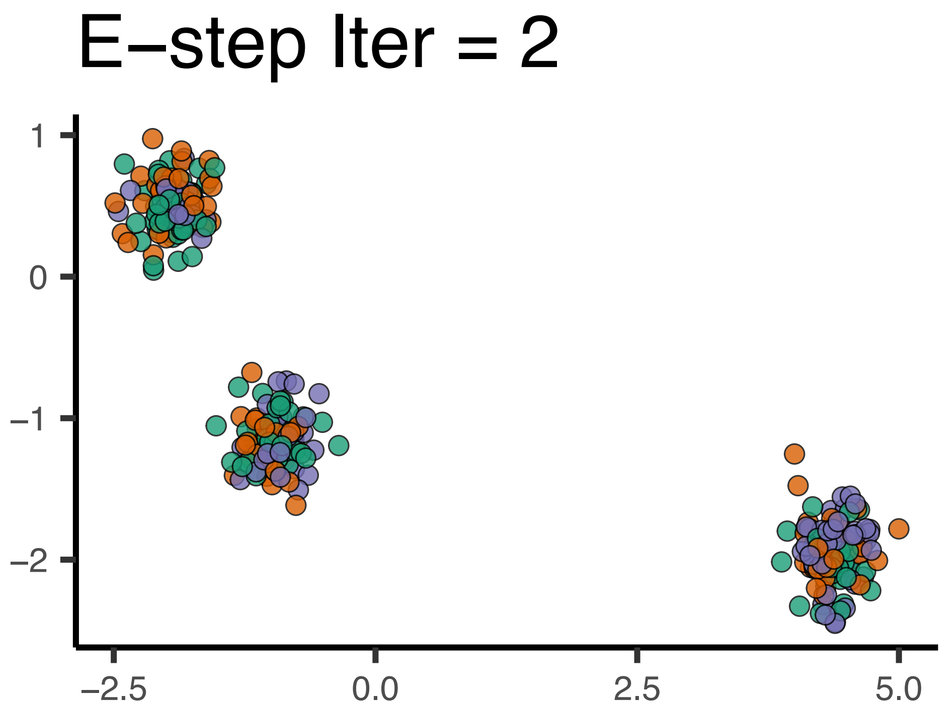
```
for(tt in 1:100){
    rand.idx <- take.estep()
    z <- nnf_one_hot(rand.idx)
    llik <- take.mstep(z)
}
```

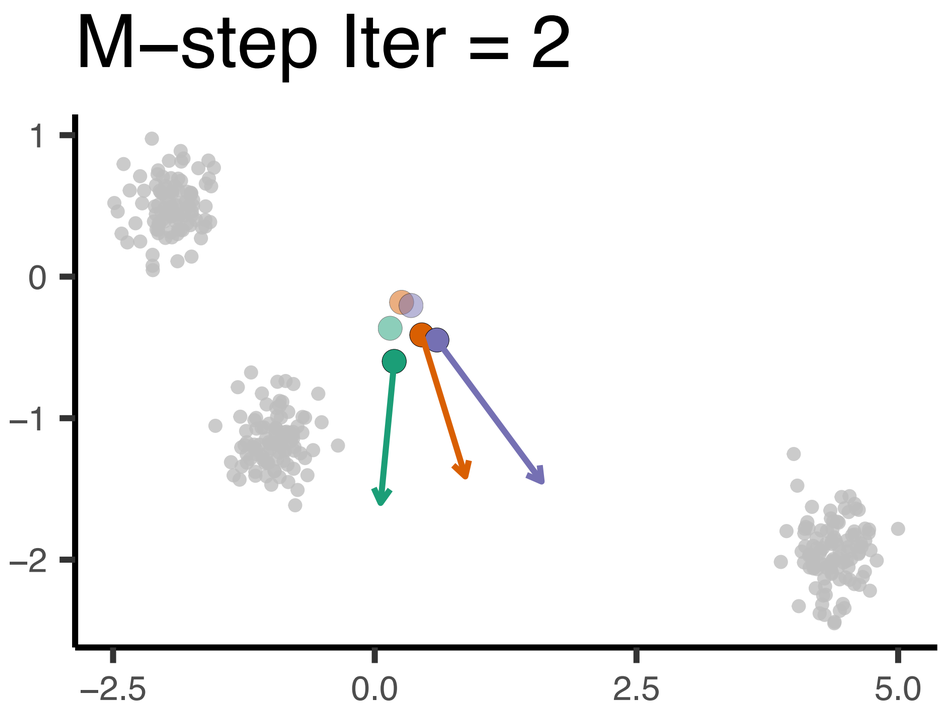




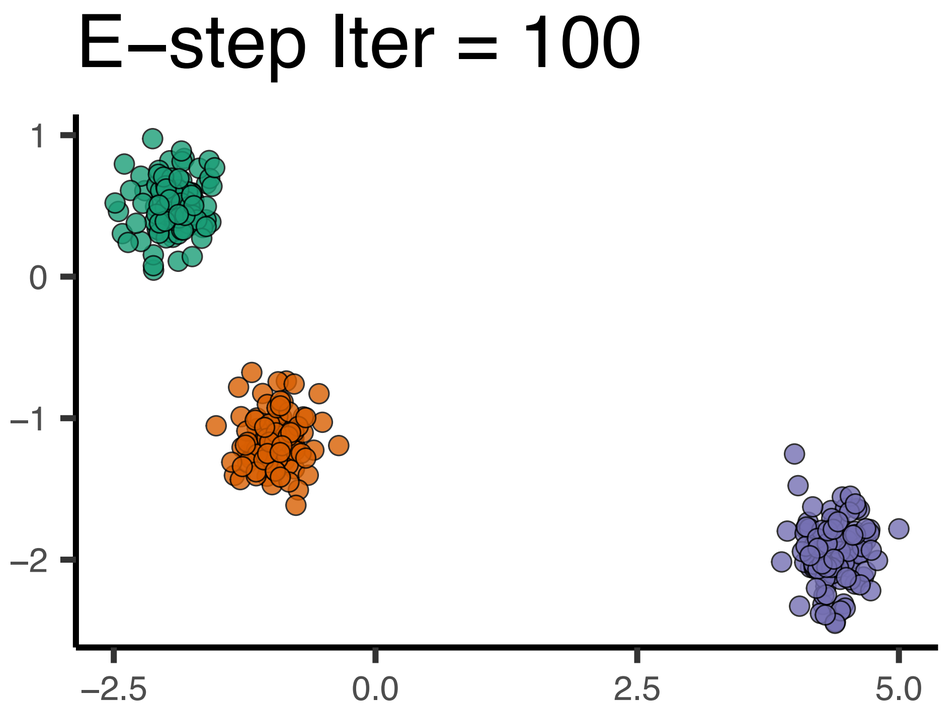


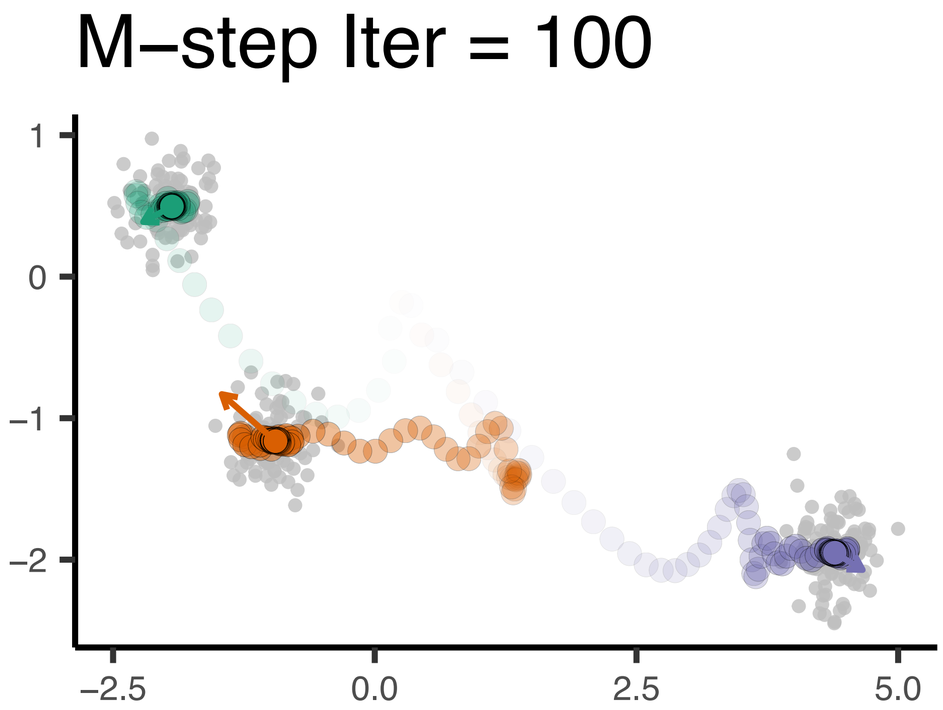
- Arrows: stochastic gradient $\nabla \mu$
- Colour: latent membership





- Arrows: stochastic gradient $\nabla \mu$
- Colour: latent membership





- Arrows: stochastic gradient $\nabla \mu$
- Colour: latent membership

Today’s lecture: Model-based Data Analysis

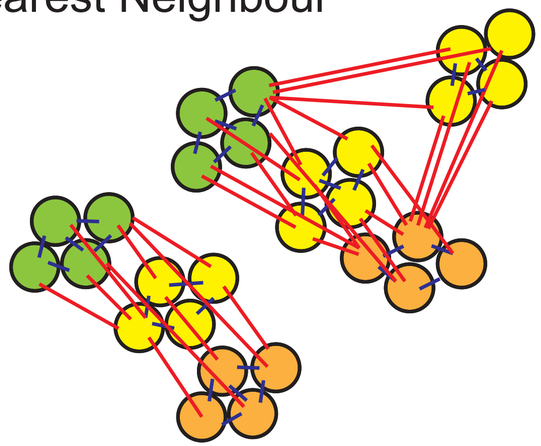
- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

Graph-based clustering of cells (cell-cell interaction network)

Where is the graph?

Batch Balanced
K-Nearest Neighbour



Most existing scRNA-seq toolboxes aim to produce best possible k-NN graph

Polanski *et al*. Bioinformatics (2019)

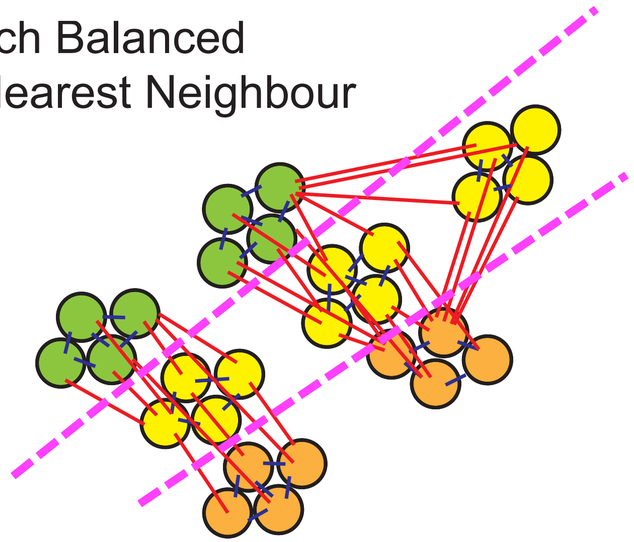
And why graph-based?

1. High-dimensionality of data
2. Unknown or unfriendly (non-Gaussian) distribution of the data vector, whereas an intercellular distance is easy to think of.
3. Sparse N x N (lots of zeros) adjacency matrix vs. D x N matrix (deeper sequencing → dense matrix)
4. Just legacy, or pretty picture

Graph-based clustering of cells

Where is the graph?

Batch Balanced
K-Nearest Neighbour



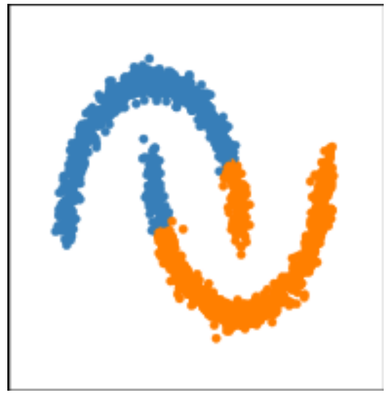
Most existing scRNA-seq toolboxes aim to produce best possible k-NN graph

Polanski *et al*. Bioinformatics (2019)

And why graph-based?

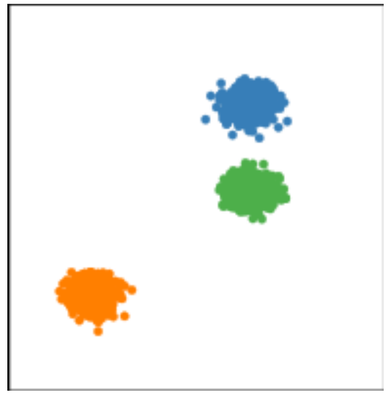
k-means

spectral clustering based on k-NN graph

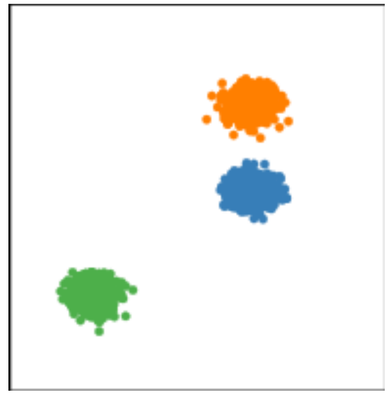


vs.

We may not know the distrib. of scRNA-seq



vs.



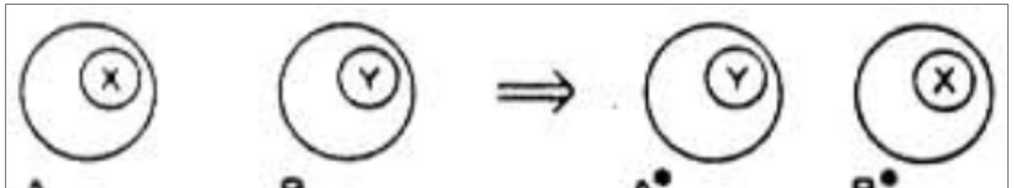
It's hard to assume that single-cell data can be faithfully modeled by a multivar. linear Gaussian model

https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Digression

Graph clustering/partitioning is a classic discrete optimization problem

Two-cut problems (vertex set A and B)



Kernighan & Lin (1970)

$A^{\bullet} = A - X + Y$

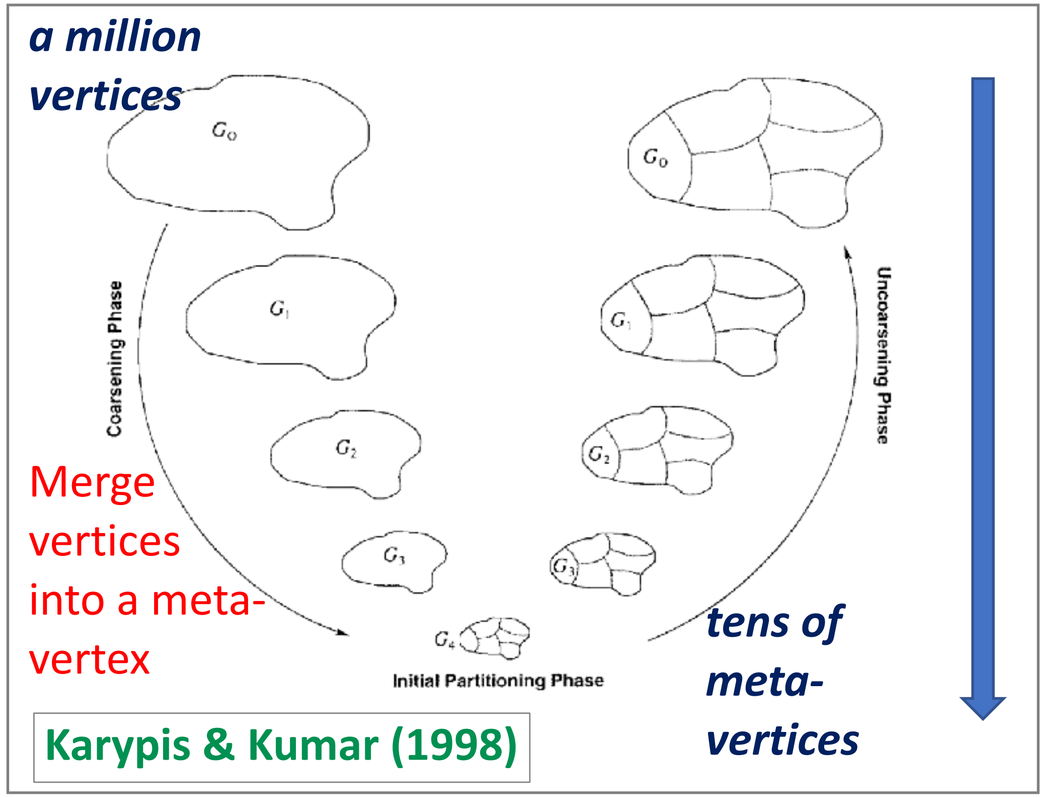
$B^{\bullet} = B - Y + X$

Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming

0.87856-approximation to opt

Goemans & Williamson (1995)

k-metis algorithm (heavily used in SciPy)



Graph contraction to simplify for high-level greedy move

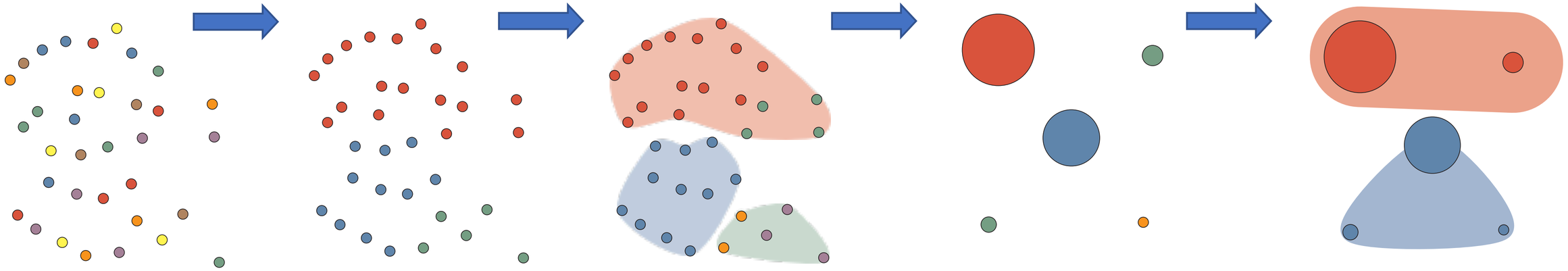
Leiden algorithm to resolve densely-connected community of nodes (cells)

Initial move (greedy update)

Initial partition

Create meta-vertices

Refine vertex assignment (locally)



From Louvain to Leiden: guaranteeing well-connected communities

V.A. Traag,* L. Waltman, and N.J. van Eck
Centre for Science and Technology Studies, Leiden University, the Netherlands
(Dated: October 22, 2018)

Louvain, Belgium

Leiden, Netherland

Digression

Stochastic block structure to model the likelihood of a graph {n,m,p}

Mathematics

cluster/block k=1

Erdos

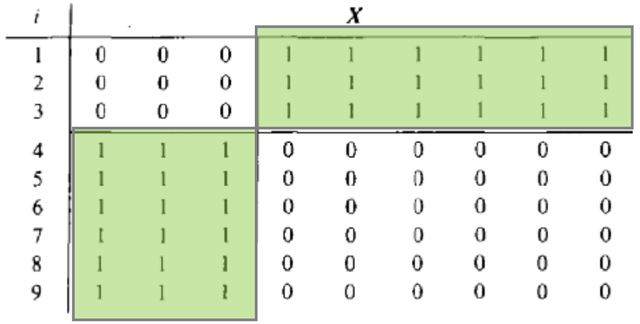
Renyi

$$p^m(1-p)^{\binom{n}{2}-m}$$

Erdos &Renyi (1959, 1960)

Social Science

cluster/block k > 1

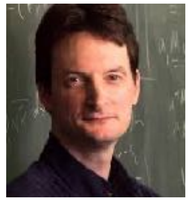


i				X					
1	0	0	0	1	1	1	1	1	1
2	0	0	0	1	1	1	1	1	1
3	0	0	0	1	1	1	1	1	1
4	1	1	1	0	0	0	0	0	0
5	1	1	1	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0	0
7	1	1	1	0	0	0	0	0	0
8	1	1	1	0	0	0	0	0	0
9	1	1	1	0	0	0	0	0	0

$$L = \prod_{r,s} \pi_{rs}^{M(r,s)} \left(1-\pi_{rs}\right)^{n_r n_s - M(r,s)}$$

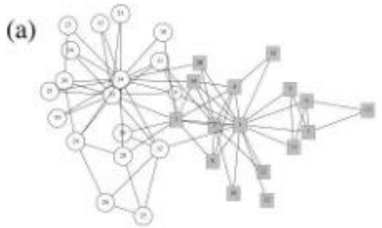
Holland, Laskey, Leinhardt (1983)

Network science



Newman

Girvan



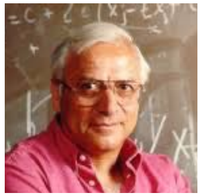
Girvan & Newman (2002)

Statistics/ML

Mixed membership (2008)

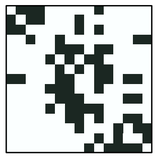
Airoldi

Blei



Fienberg

Xing

P(A  **|Θ**  **)**

Today’s lecture: Model-based Data Analysis

- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning

A cell type identification problem

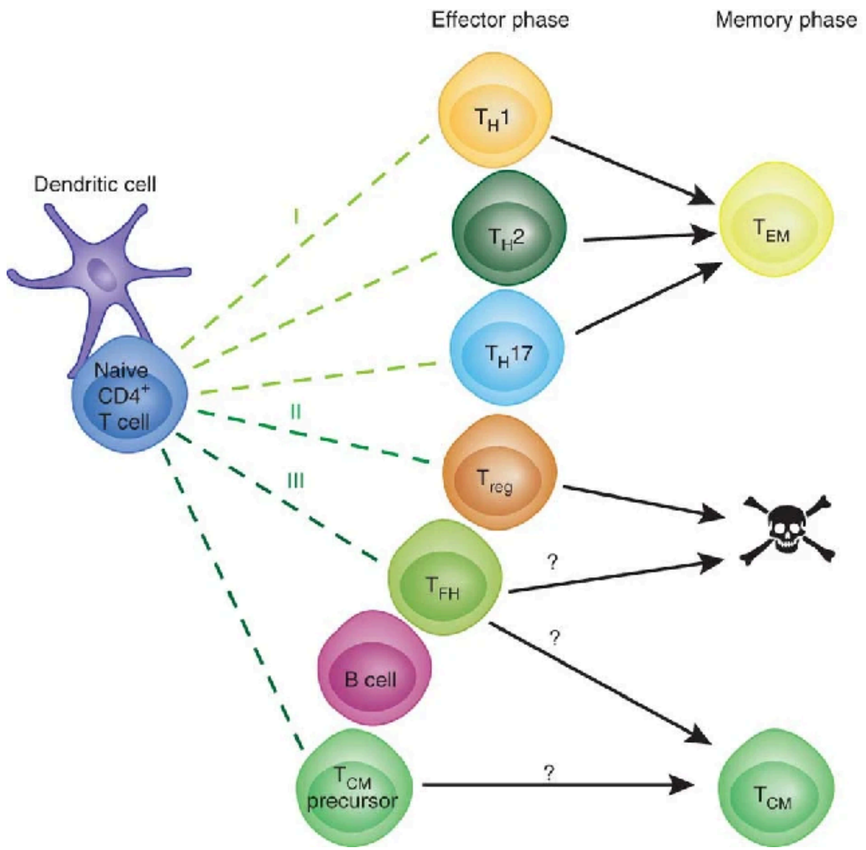
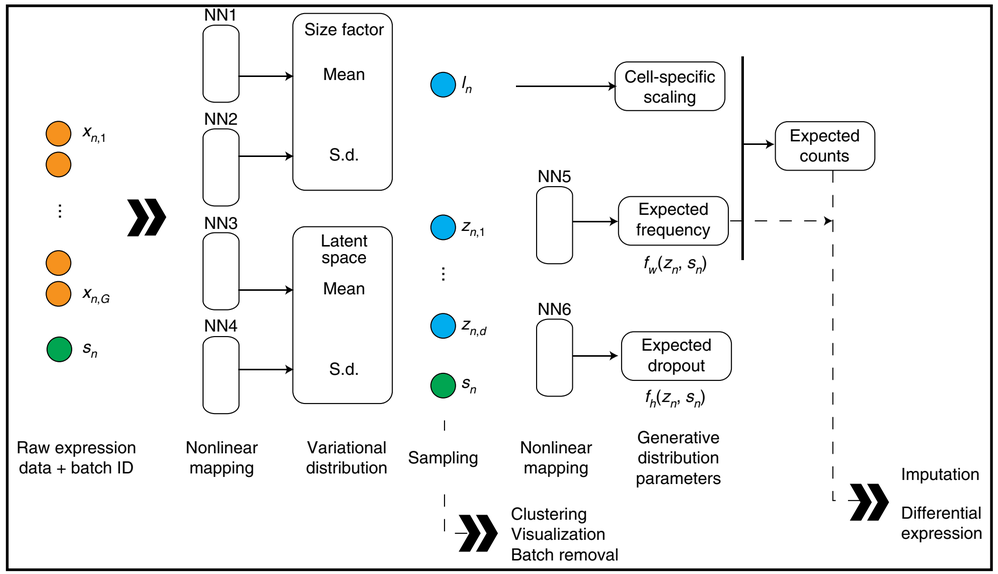


Image: Pepper & Jenkins (2011)

My first idea: Deep Autoencoder model



Lopez *et al.* (2019)

Find latent “representation” of known and novel cell types

Different opinions on a cell type identification problem

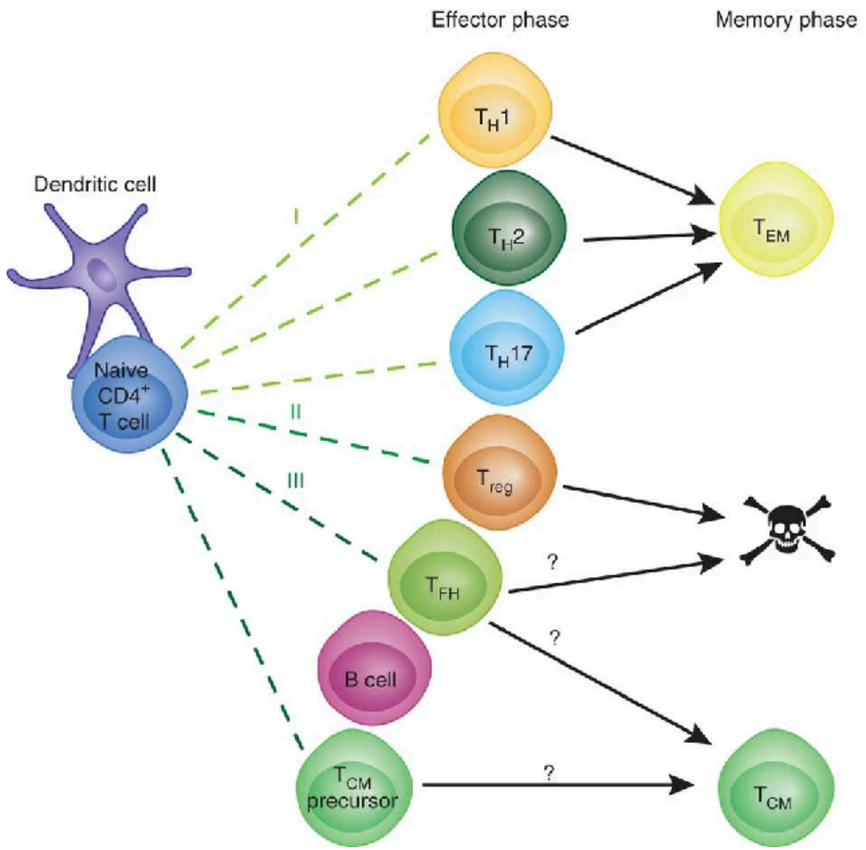


Image: Pepper & Jenkins (2011)

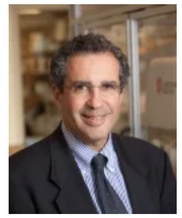
What does this latent variable mean?

How do you know this factor represents memory Treg?

VAE model is difficult to train and produces inferior results than PCA


Tomo Sumida


Matt Lincoln


David Hafler


Liang He

Cell type identification = cell sorting

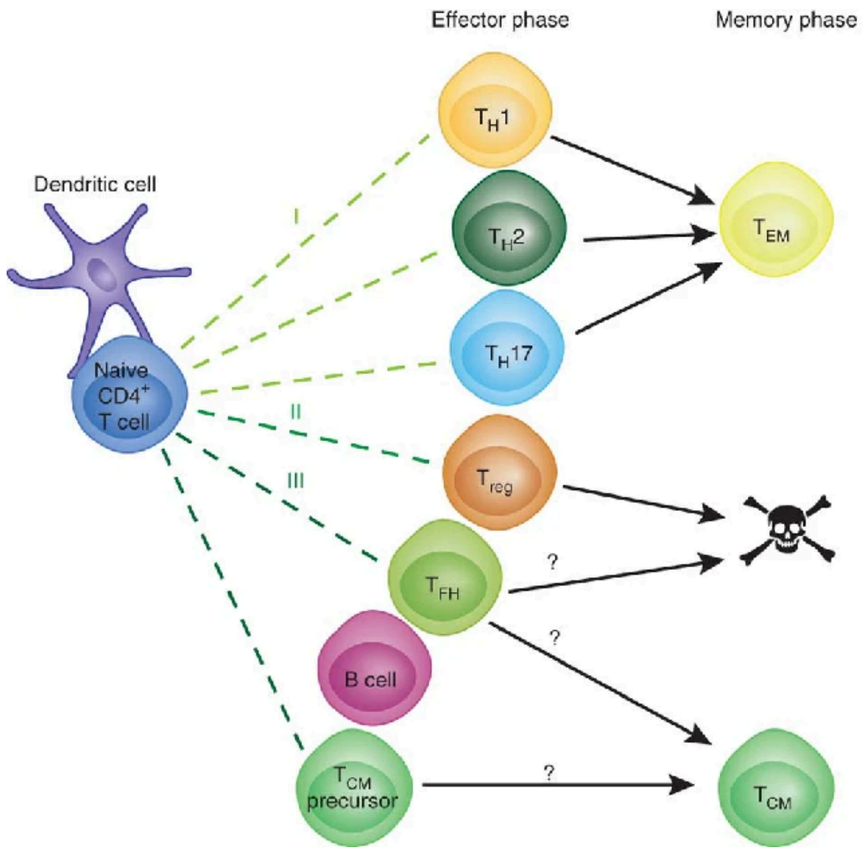
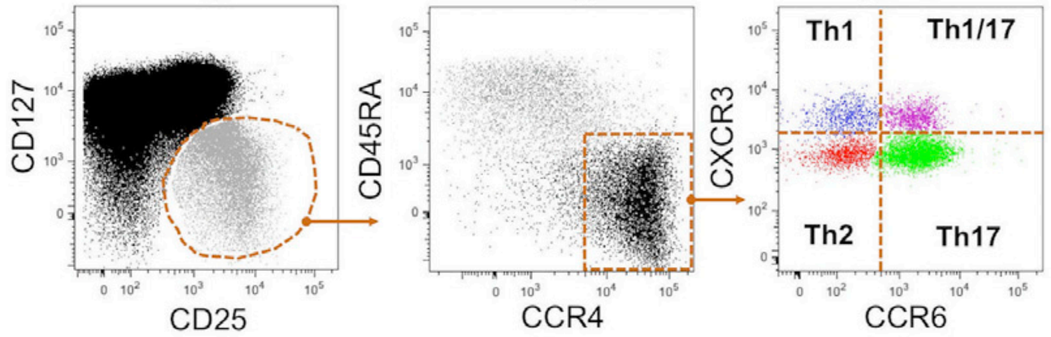


Image: Pepper & Jenkins (2011)

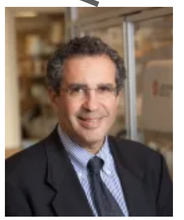
Image: Halim *et al*. (2017)



Immunologists @ Yale

Tomo Sumida

Matt Lincoln

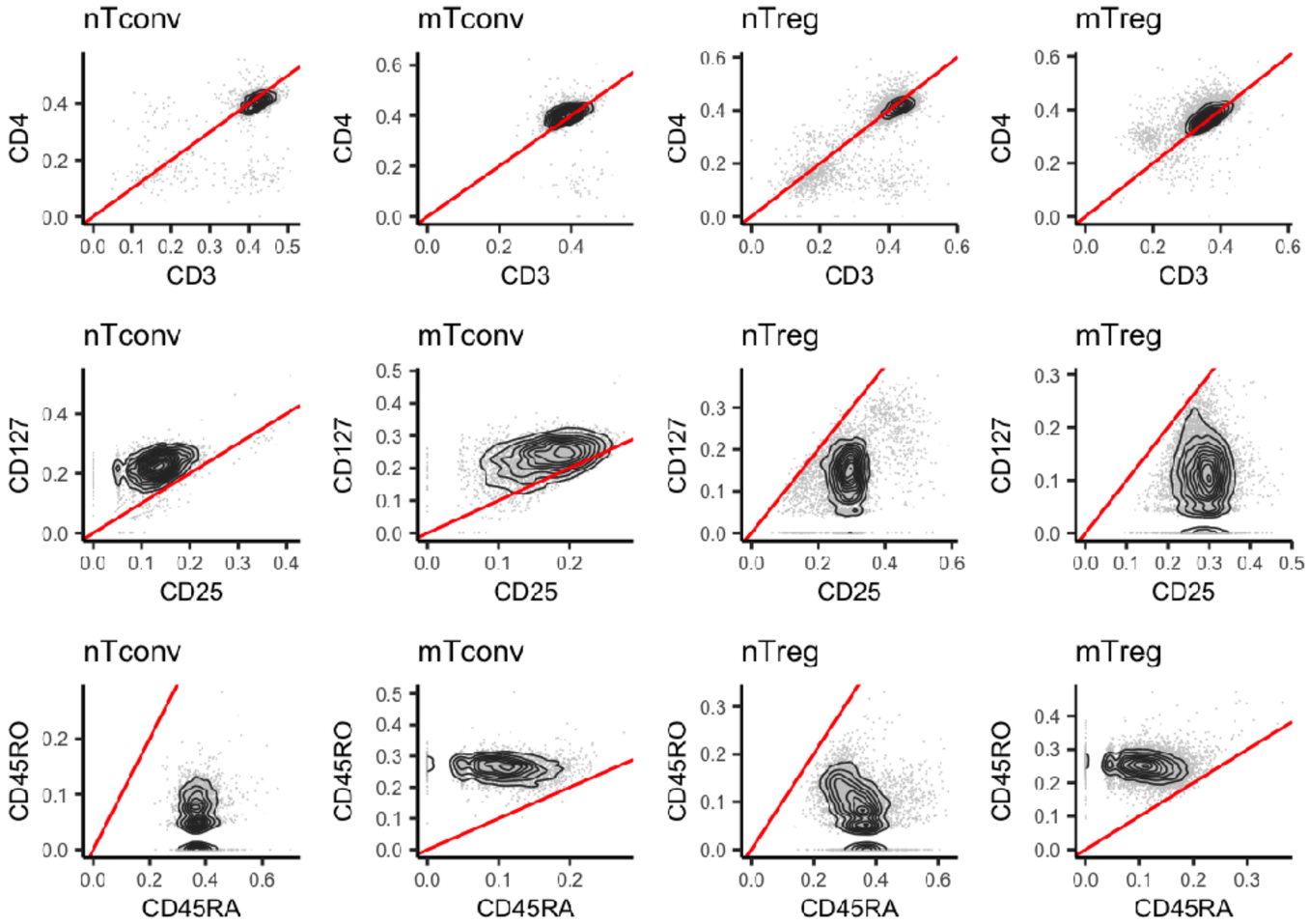


David Hafler

Epidemiologist @ Duke

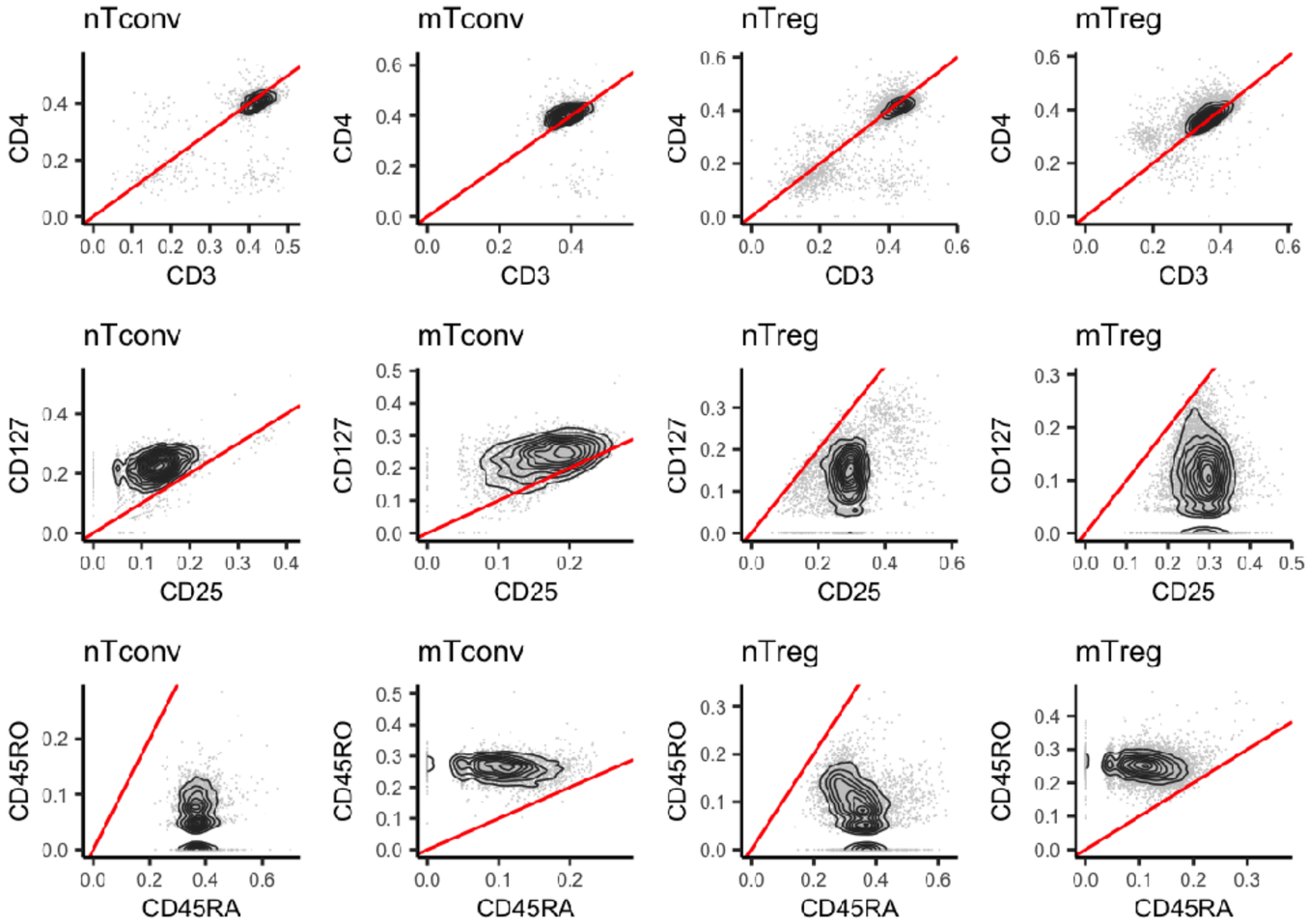
Liang He

Cell type identification = cell sorting



- Identify informative cell surface markers (prior knowledge) or marker genes (feature selection).
- Draw a decision boundary between different cell types (e.g., CD127+/CD25- vs. CD127+/CD25+)

Cell type identification = cell sorting



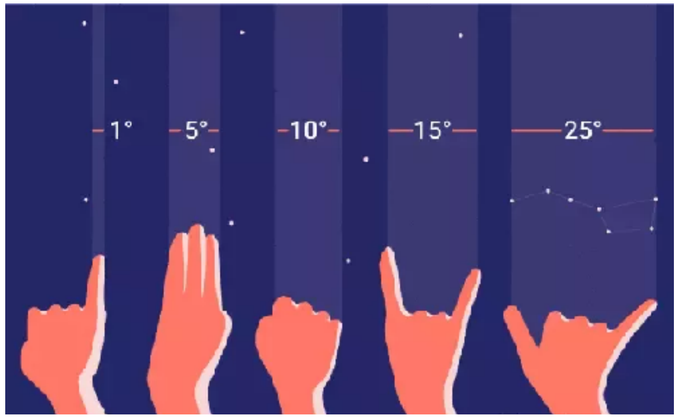
CD3+/CD4+ Q/C

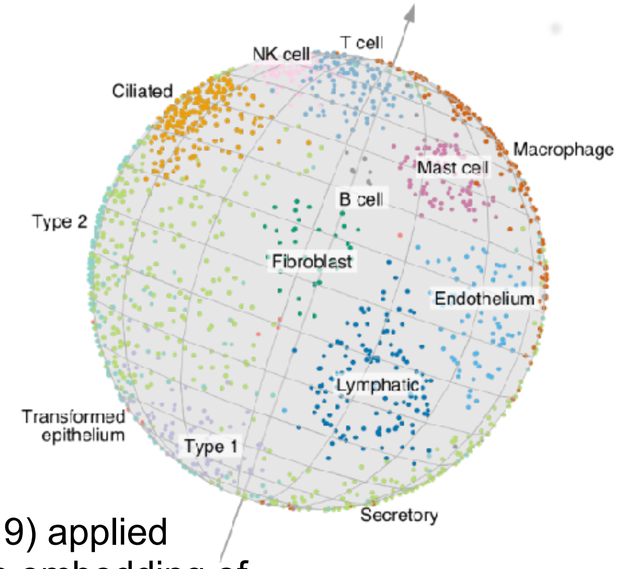
Regulatory T-cells vs. conventional (effector) T-cells

Memory vs. naive T-cells

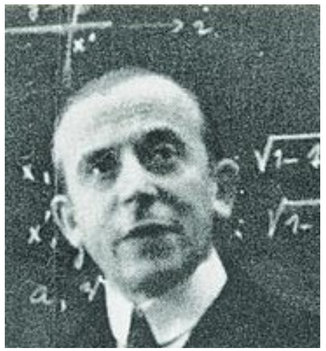
Formulation in Math: Angular distance and von Mises-Fisher

Angular distance between the stars observed on Earth





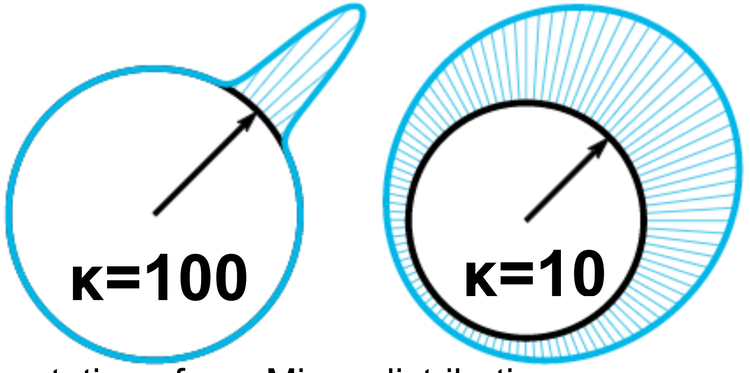
Ding and Regev (2019) applied Spherical VAE to learn embedding of scRNA-seq (robust against unknown batch effects & outliers)



Von Mises distribution (p=2)

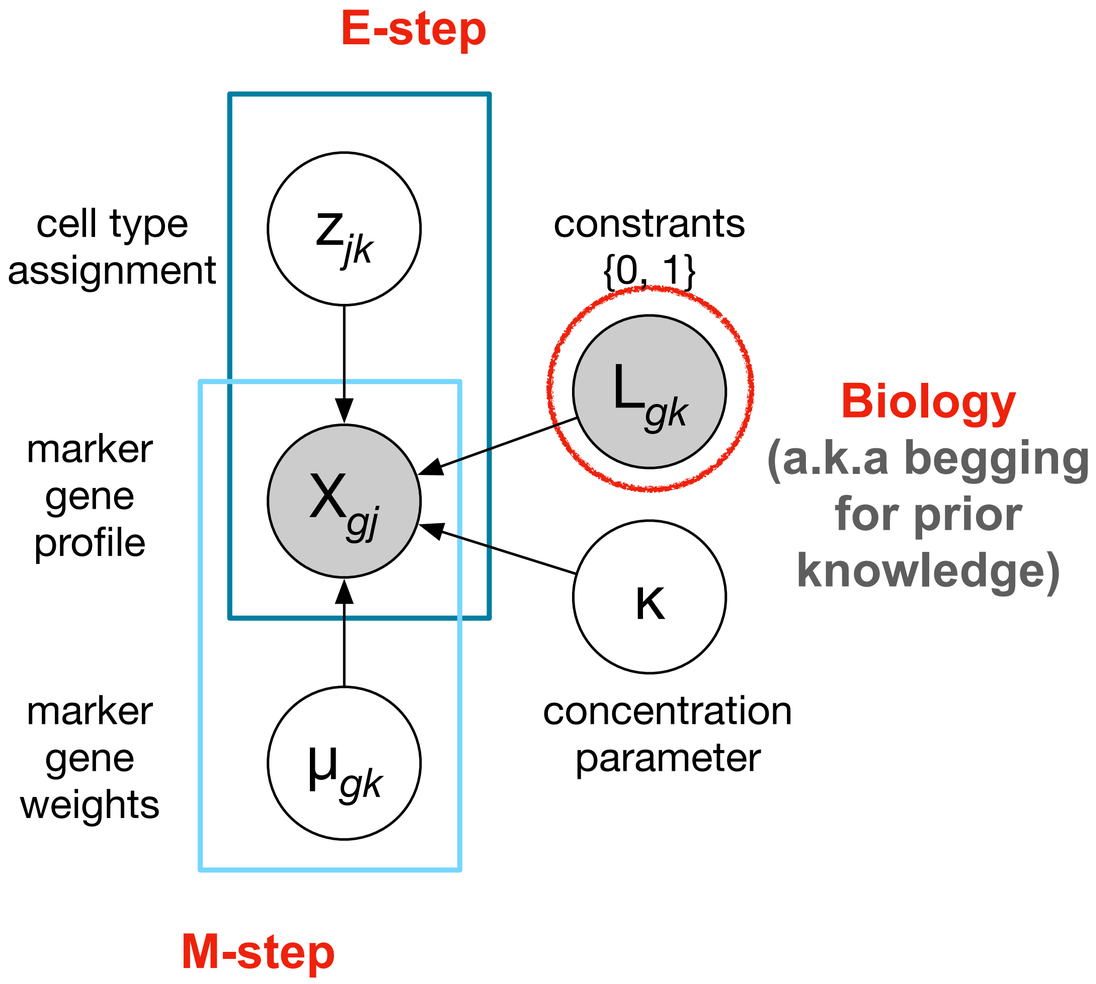
Fisher extends p > 2

$$\exp\{\kappa \mathbf{x}_j^\top \boldsymbol{\mu}_k\} C(\kappa)$$

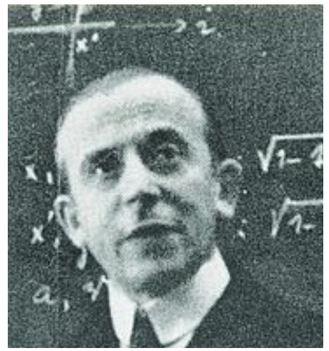


2D representation of von Mises distribution.
(left) high precision/low variance (right) low precision
Kappa = concentration parameter
C(κ) = normalizer,
μκ = cluster mean for that gaussian

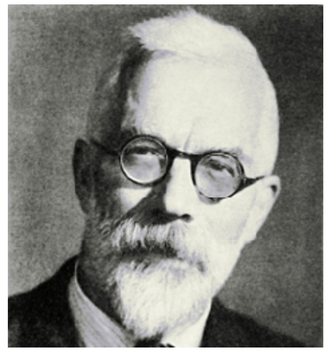
Fit a mixture of von Mises-Fisher distributions



Biology (a.k.a begging for prior knowledge)

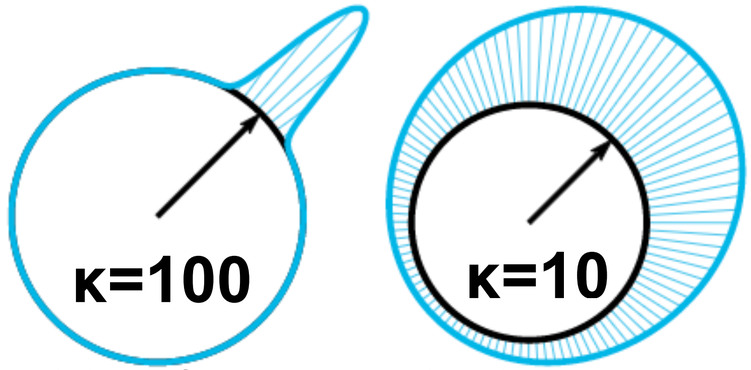


Von Mises distribution (p=2)



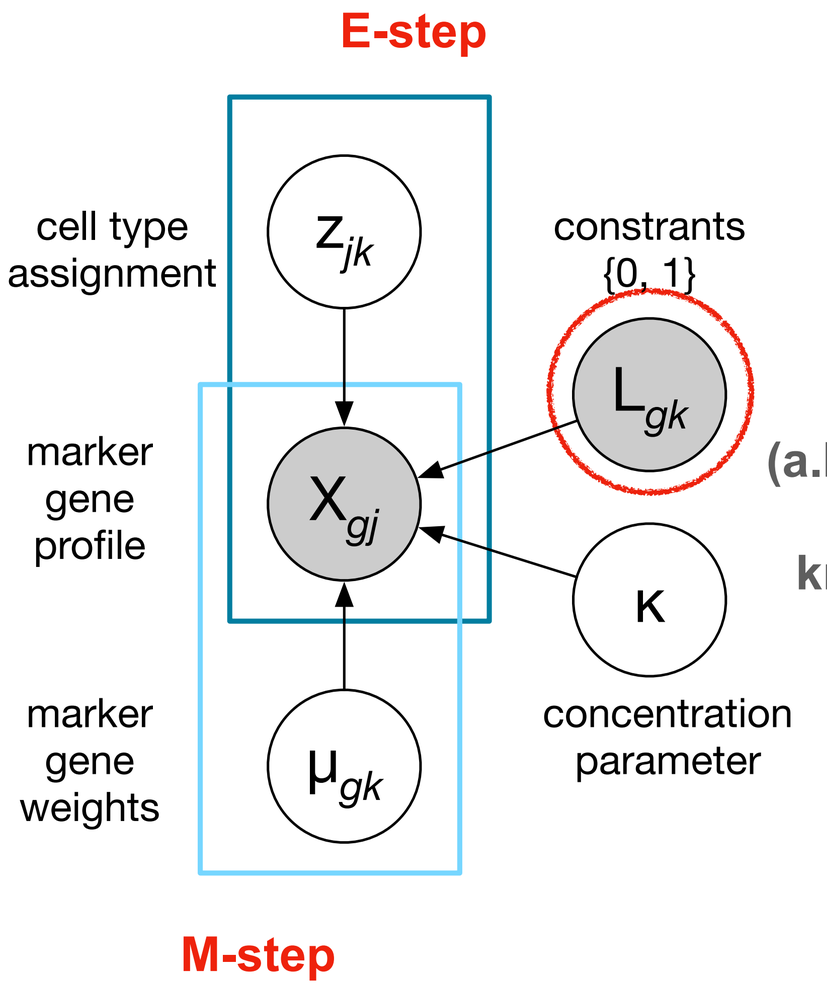
Fisher extends p > 2

$$\exp\{\kappa \mathbf{x}_j^\top \boldsymbol{\mu}_k\} C(\kappa)$$



2D representation of von Mises distribution.
(left) high precision/low variance (right) low precision
Kappa = concentration parameter
C(κ) = normalizer,
μκ = cluster mean for that gaussian

Fit a mixture of von Mises-Fisher distributions



Biology (a.k.a begging for prior knowledge)


Tomo Sumida

Matt Lincoln

For Multiple Sclerosis & T-cell biology

For brain data analysis

NIMH REPOSITORY & GENOMICS RESOURCE

https://www.nimhgenetics.org/

Today’s lecture: Model-based Data Analysis

- **Model-based scientific investigation**

- **Example: Dynamics and trajectory inference**
  - Minimum spanning tree
  - RNA velocity and ordinary differential equation
  - Variational autoencoder models
- **Example: Cell type annotation**
  - Clustering by expectation maximization
  - Graph-based clustering
  - Supervised learning