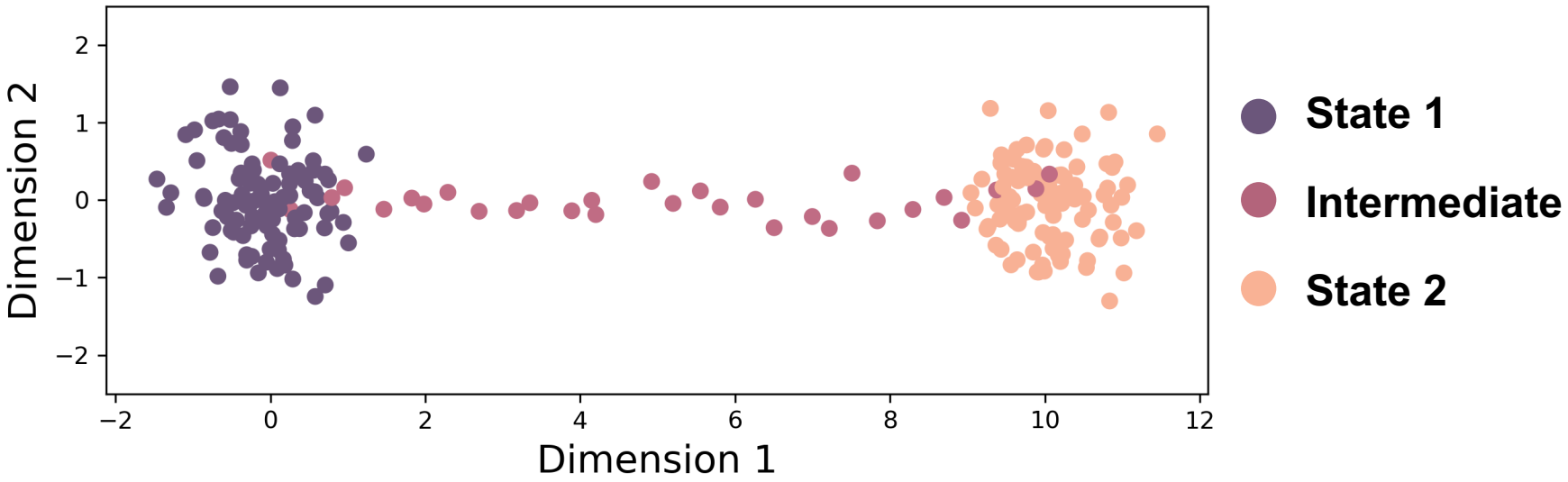
Learning trajectories from scRNA-seq data

Machine Learning for Single Cell Analysis

Goals for this unit

- Understanding pseudotime
- Comparison of diffusion and
- Calculating distances between points
- Introduction to the Numpy / Scipy / Pandas / Sklearn ecosystem

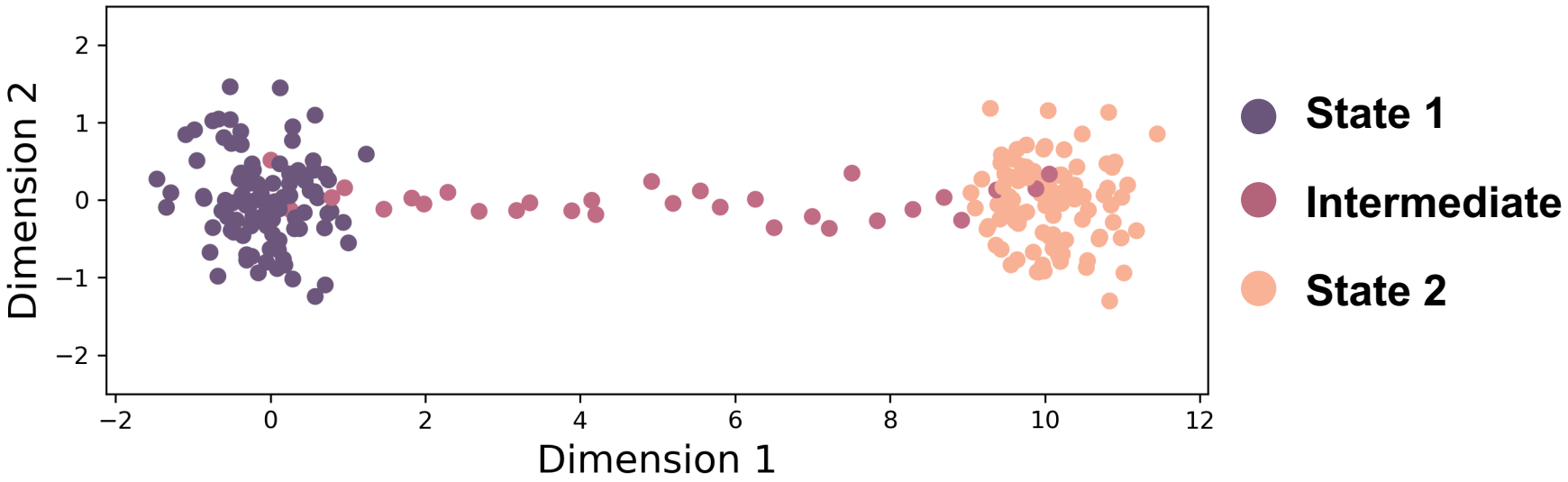
A simple developmental system



Dimension 1: “Genes that change from State 1 to State 2”

Dimension 2: “Noise genes”

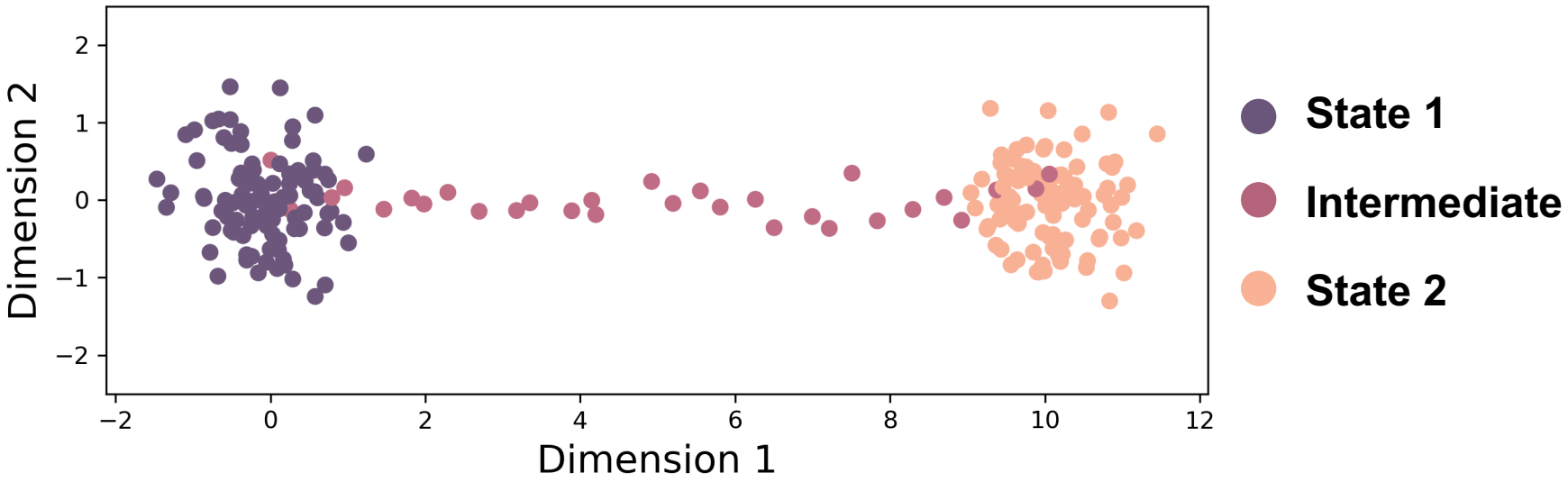
A simple developmental system

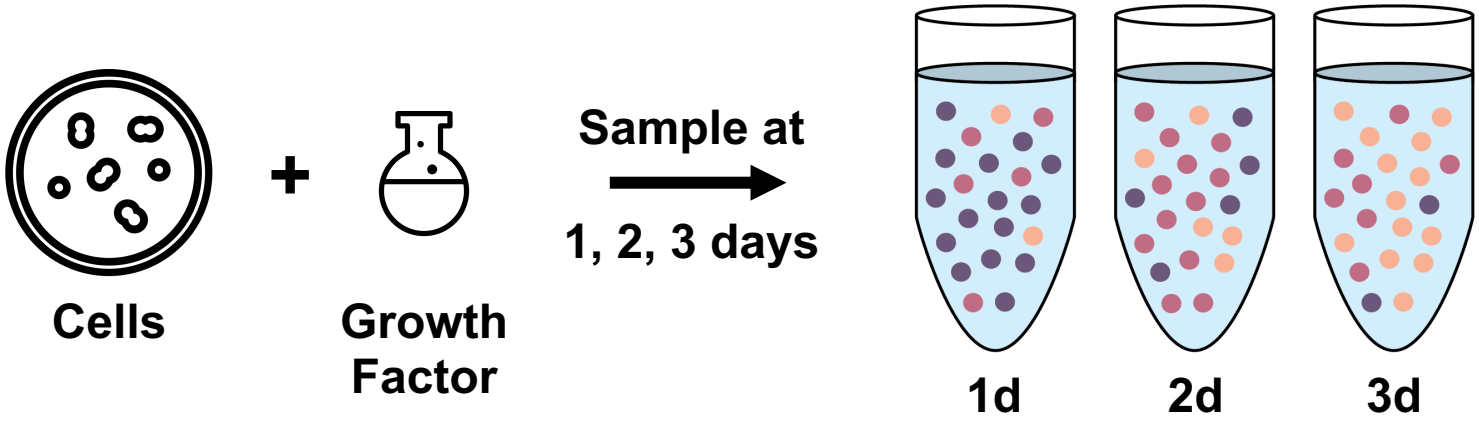


Things we might like to know:

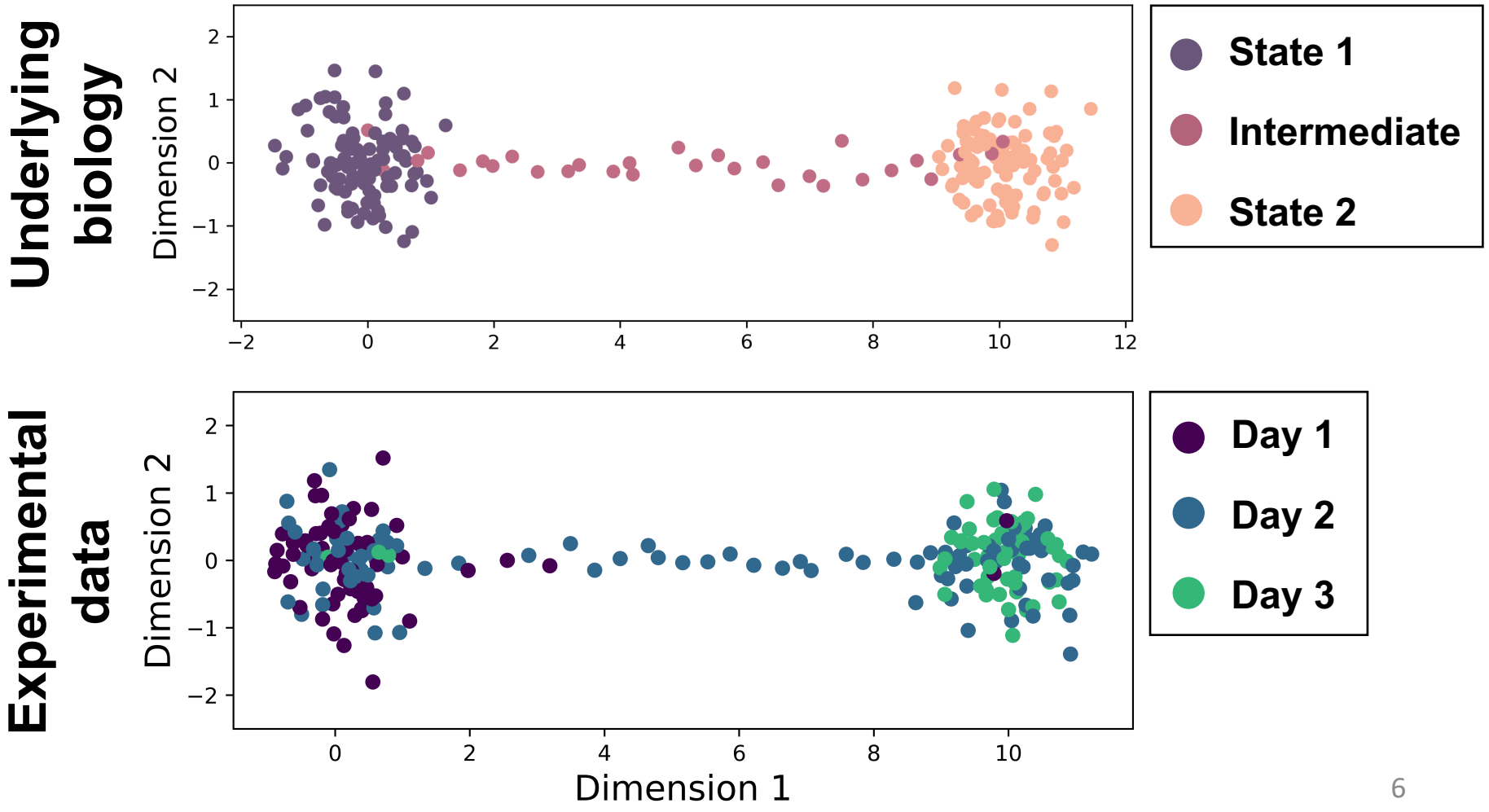
- What genes are expressed in each population?
- What genes change the most from State 1 -> State 2?
- What is the ordering of gene expression?
- How do gene-gene relationships change across states?

A simple developmental system





We must infer underlying biology from experimental observations

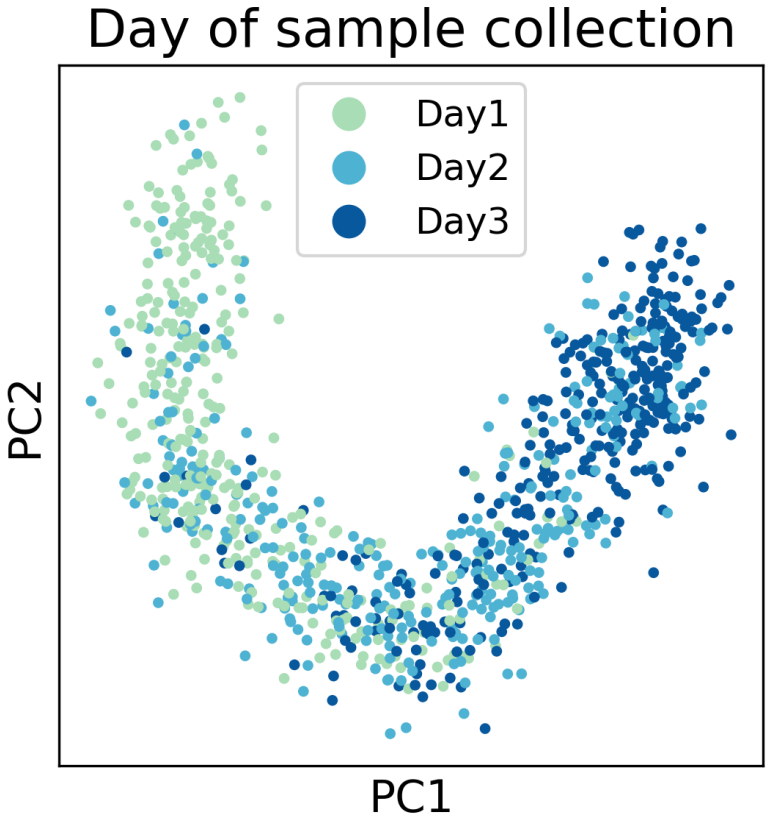


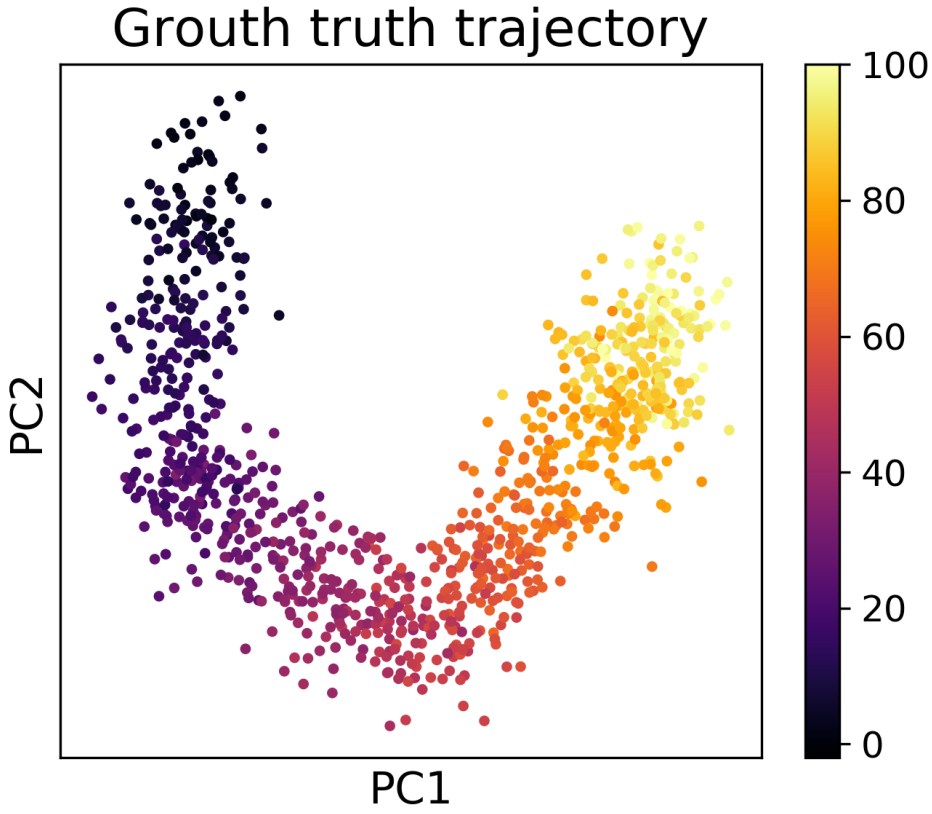
The goal of trajectory inference (TI) is to infer a temporal ordering of cells

Summary of 45 TI methods

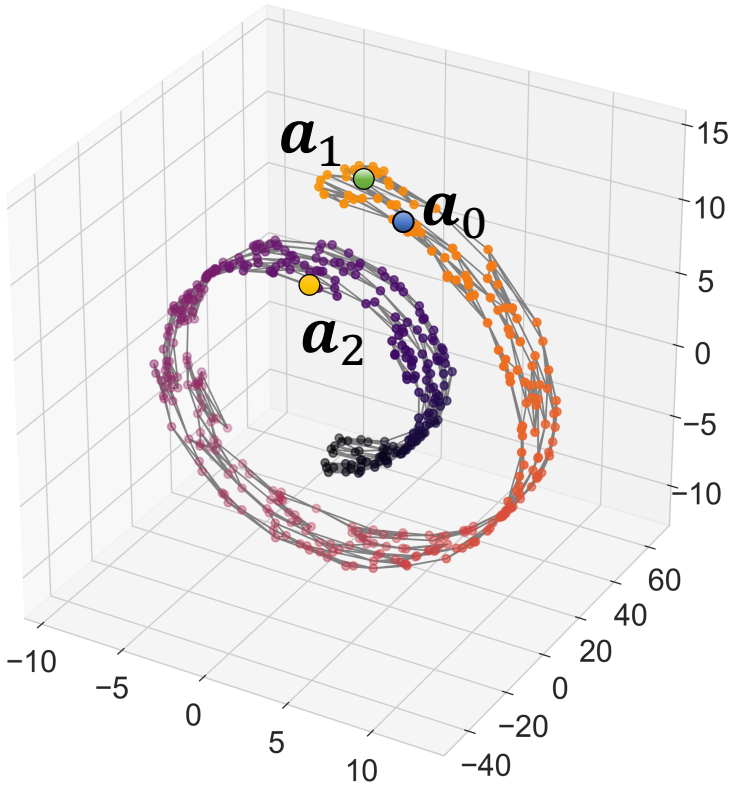
dynverse.org

Learning pseudotime via graph walks

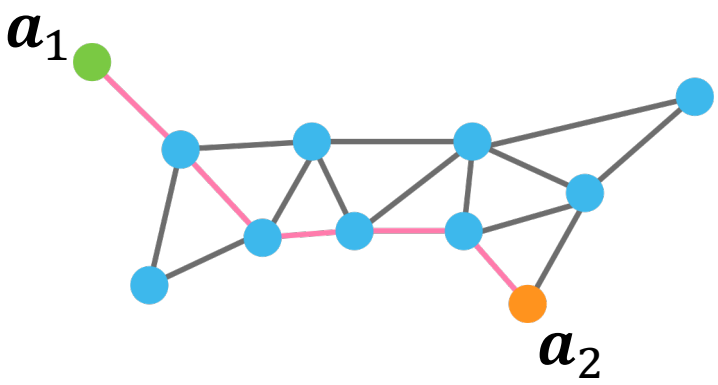




Graph walks approximate manifold distances

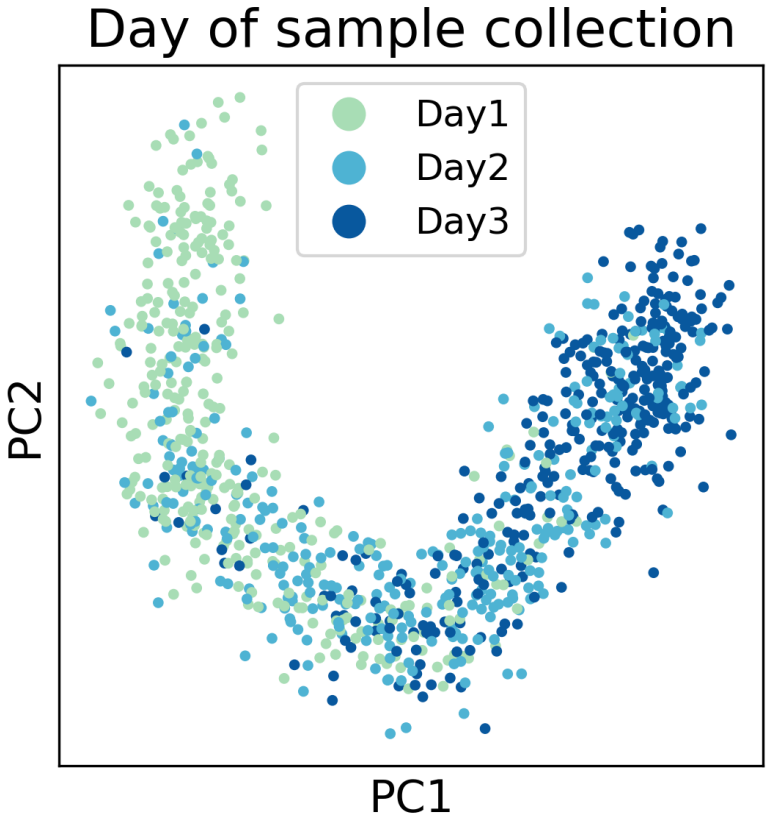


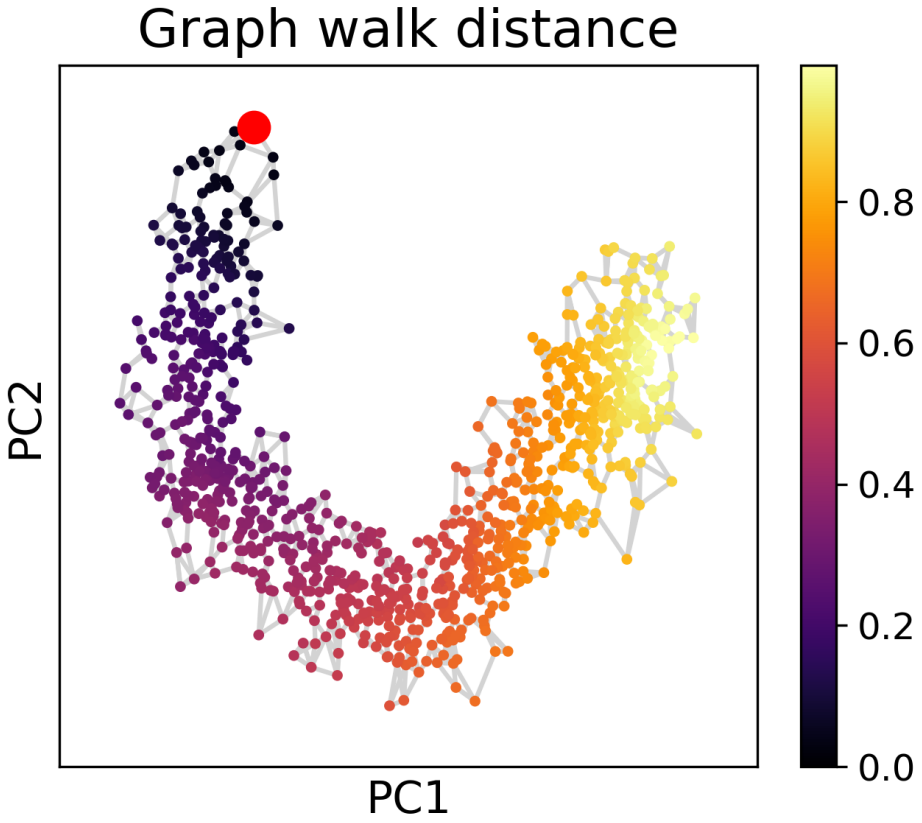
Shortest path between points



$$d_{geodesic}(a_1, a_2) = \text{shortestpath}(a_1, a_2)$$
$$= 5$$

Learning pseudotime via graph walks

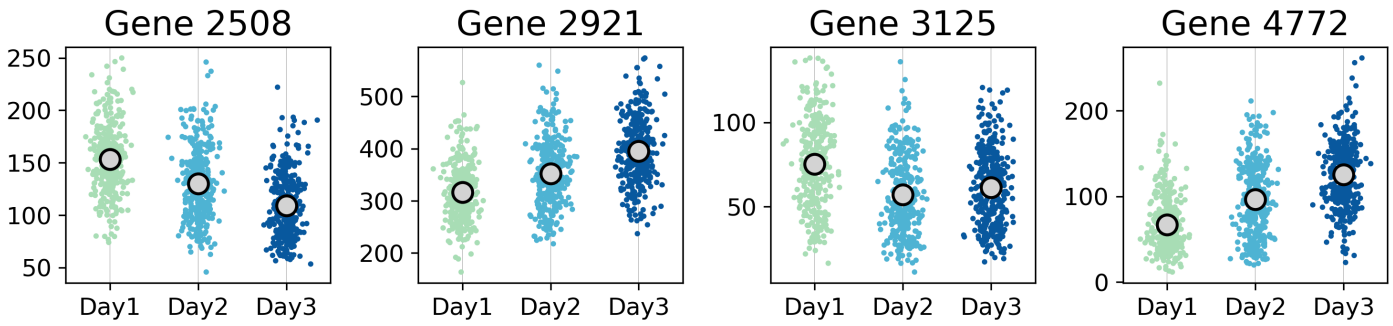




Pseudotime facilitates analysis of genes driving cellular progression

Ordering

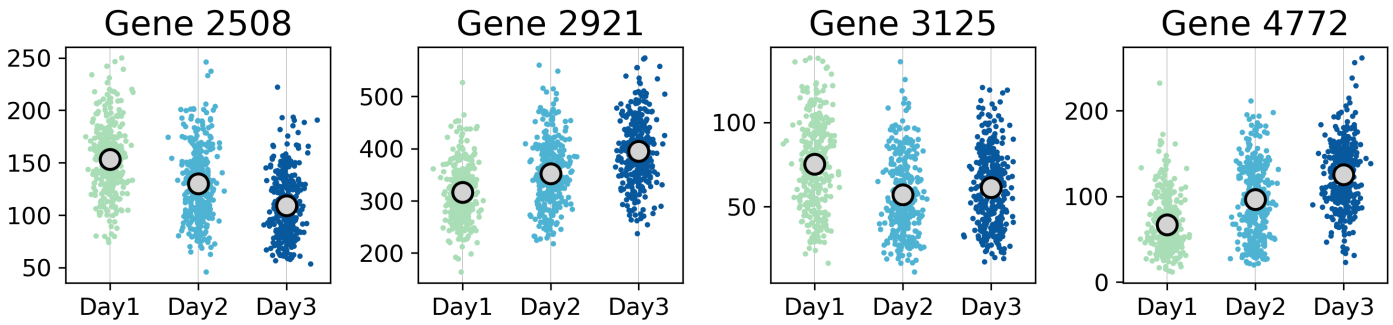
Sample



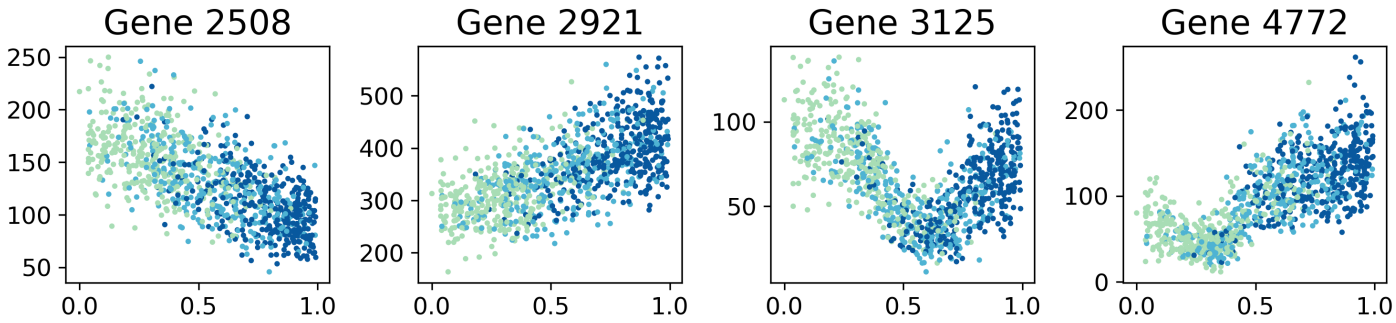
Pseudotime facilitates analysis of genes driving cellular progression

Ordering

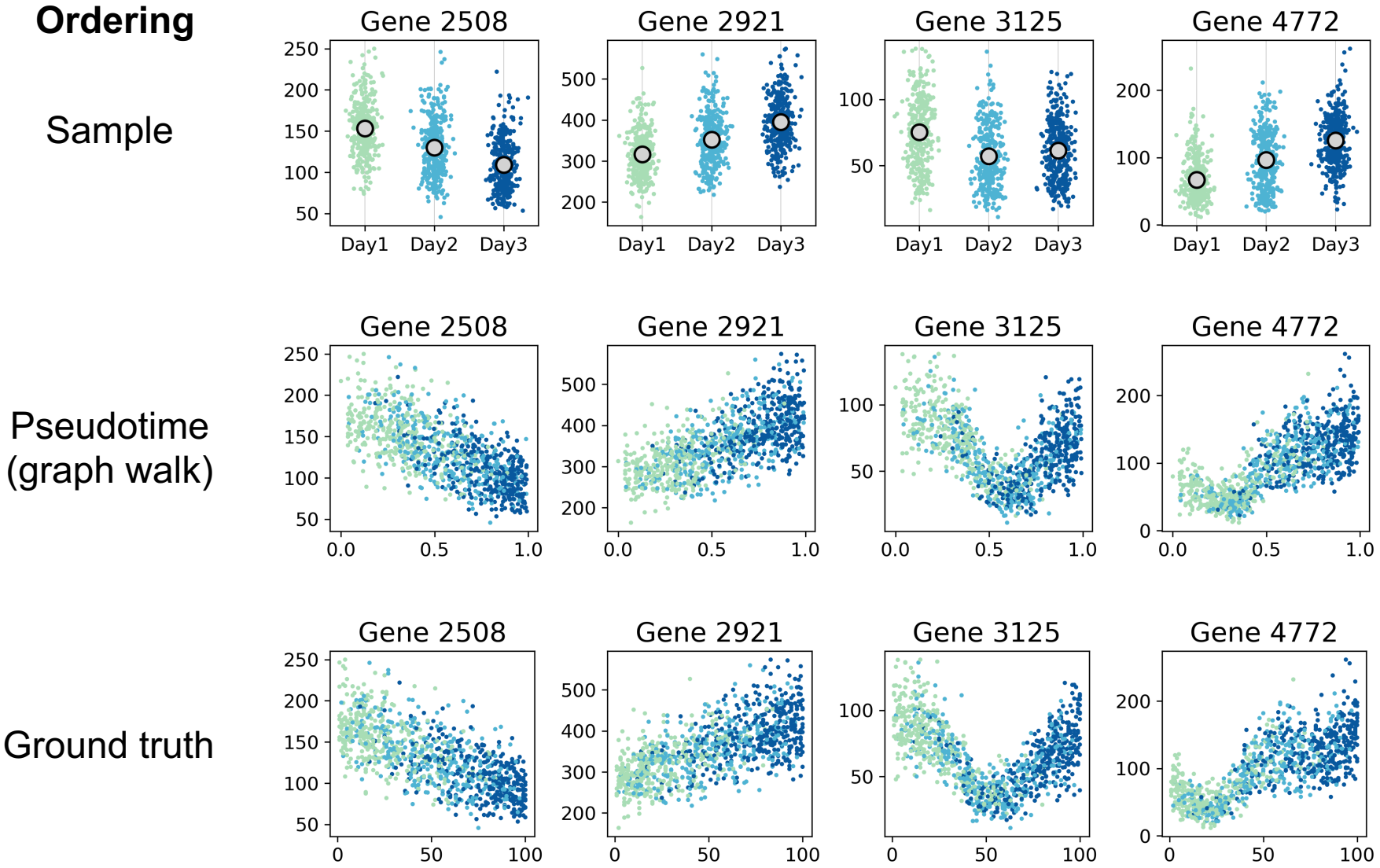
Sample



Pseudotime (graph walk)



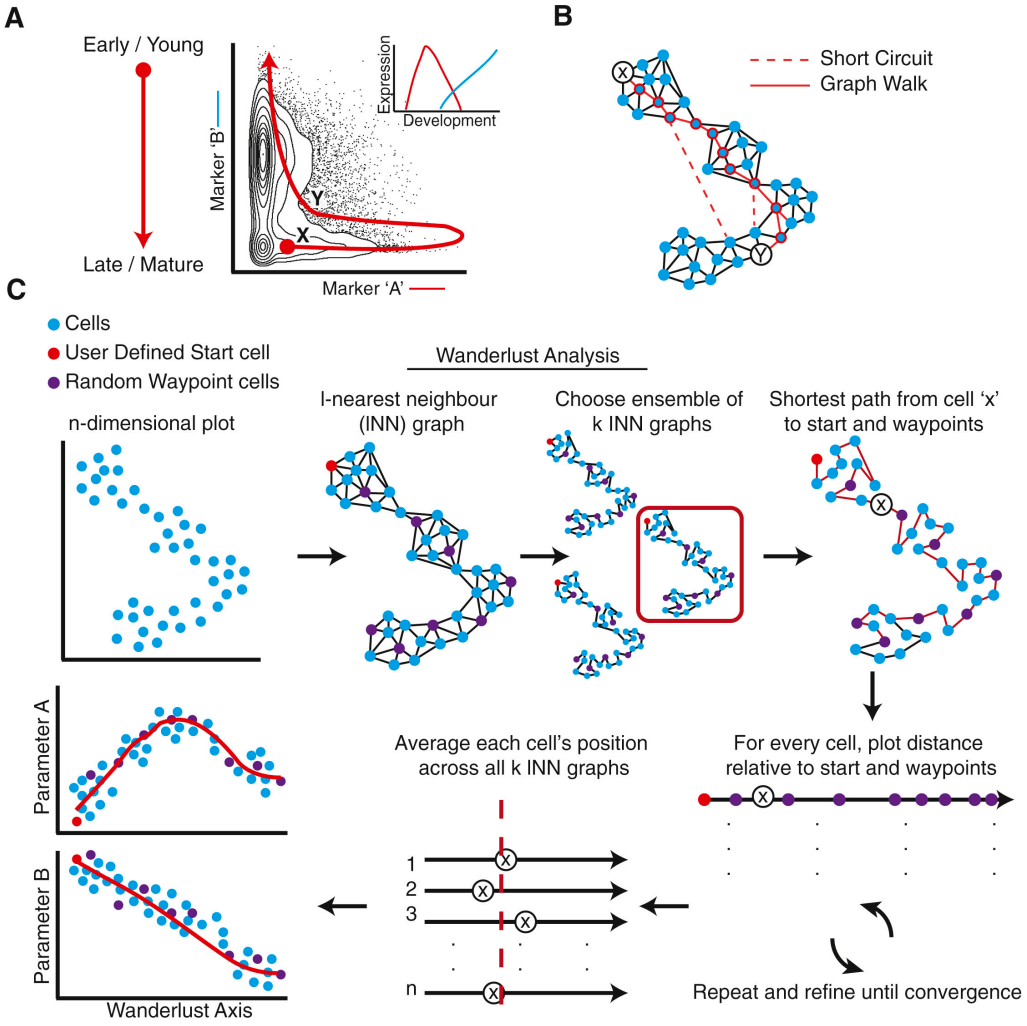
Pseudotime facilitates analysis of genes driving cellular progression



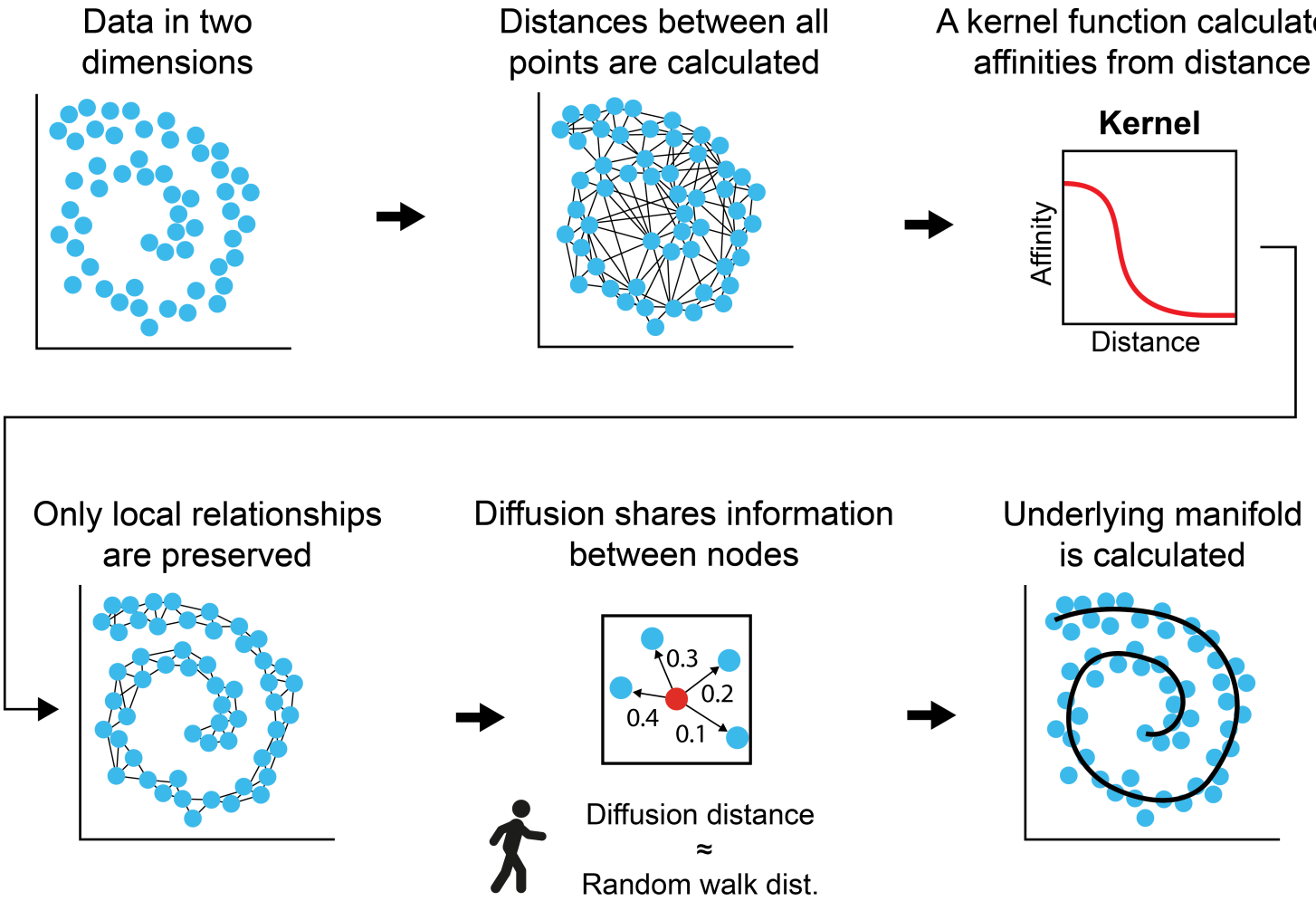
Wanderlust – Bendal et al. (2014)

Pseudocode

1. User selects a ”start cell”
2. Randomly select *n* “waypoints”
3. Learn a set of k-NN graphs
4. For each graph:
    1. For each cell:
        1. Calculate walk distance to the start cell
        2. Calculate walk distance to waypoints
    2. Calculate each cell’s relative position to start and endpoints
5. Average across all k-NN graphs



Diffusion pseudotime – from graph distance to random walks



Diffusion pseudotime – from graph distance to random walks

Euclidean distance between two vectors

Sum over values of *t*

$$\mathrm{dpt}(x,y) = ||\, \boldsymbol{M}(x,.) - \boldsymbol{M}(y,.)\,||,\ \boldsymbol{M} = \sum_{t=1}^{\infty} \tilde{\boldsymbol{T}}^{t}.$$

$$\boldsymbol{M} = \sum_{t=1}^{\infty} \tilde{\boldsymbol{T}}^{t} = (I - \tilde{\boldsymbol{T}})^{-1} - I \text{ where } \tilde{\boldsymbol{T}} = \boldsymbol{T} - \psi_0 \psi_0^{T},$$

ψ is used for eigenvectors
ψ_0 represents starting probabilities for T

Provides a calculatable solution to the infinite series

T is the random walk matrix

I is the identity matrix with 1's on the diagonal

$$\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

Diffusion pseudotime – from graph distance to random walks

Euclidean distance between two vectors

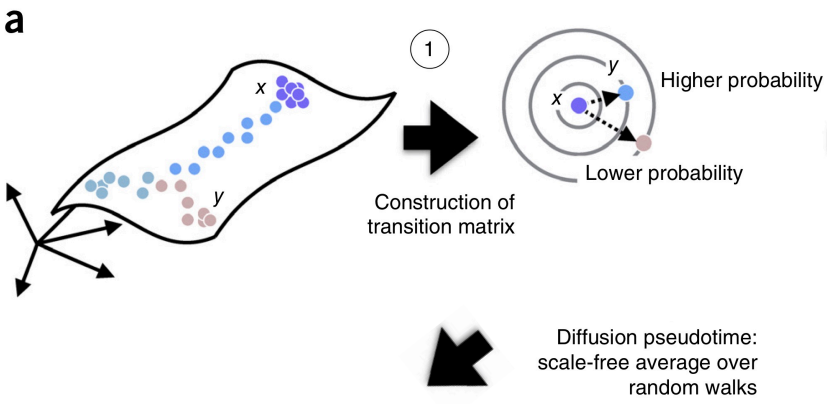
Sum over values of t

$$\mathrm{dpt}(x,y) = ||\, \boldsymbol{M}(x,.) - \boldsymbol{M}(y,.)\,||,\ \boldsymbol{M} = \sum_{t=1}^{\infty} \tilde{\boldsymbol{T}}^t.$$

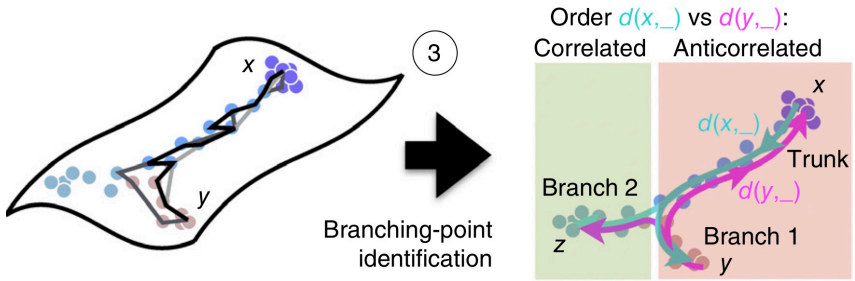
$$\boldsymbol{M} = \sum_{t=1}^{\infty} \tilde{\boldsymbol{T}}^t = (I - \tilde{\boldsymbol{T}})^{-1} - I \text{ where } \tilde{\boldsymbol{T}} = \boldsymbol{T} - \psi_0 \psi_0^T,$$

Provides a calculatable solution to the infinite series

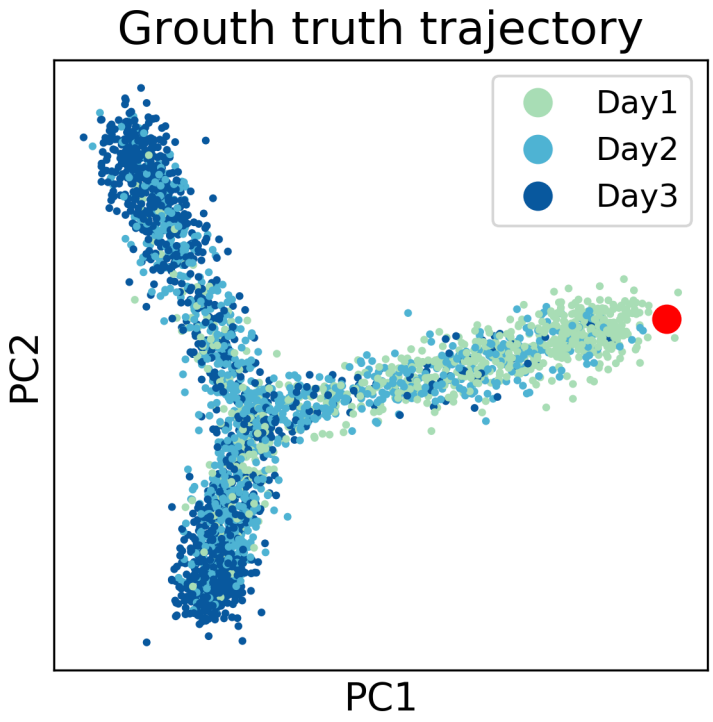
Transition probabilities give data geometry

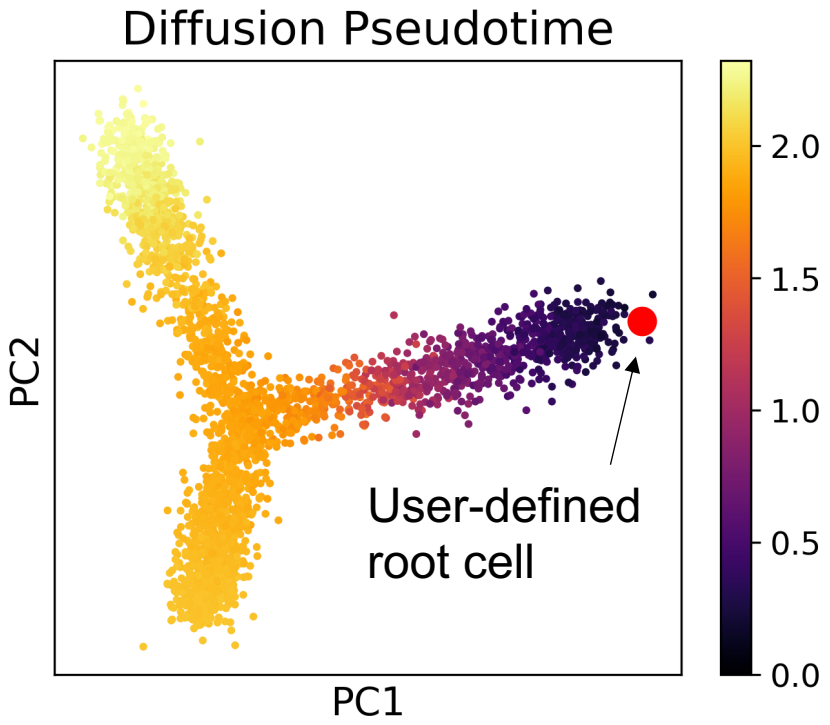


Correlation reveals branches

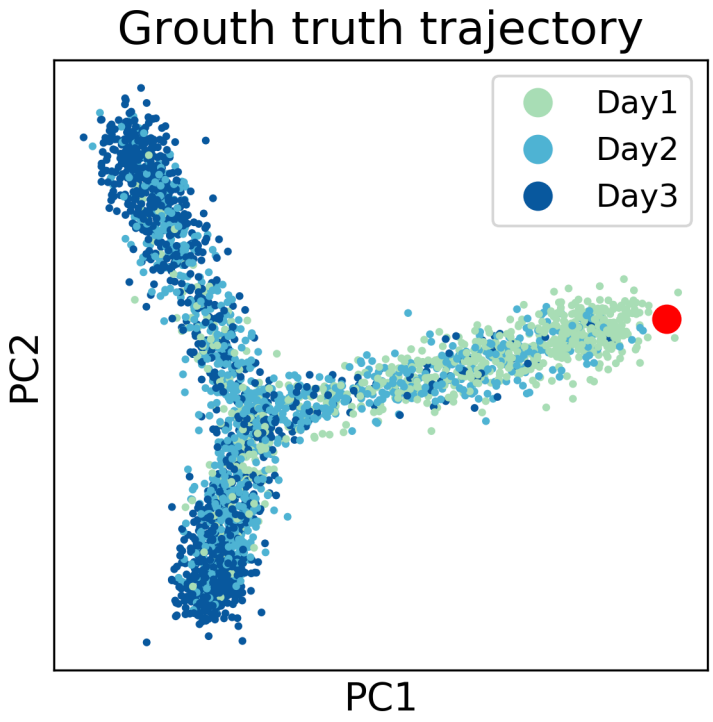


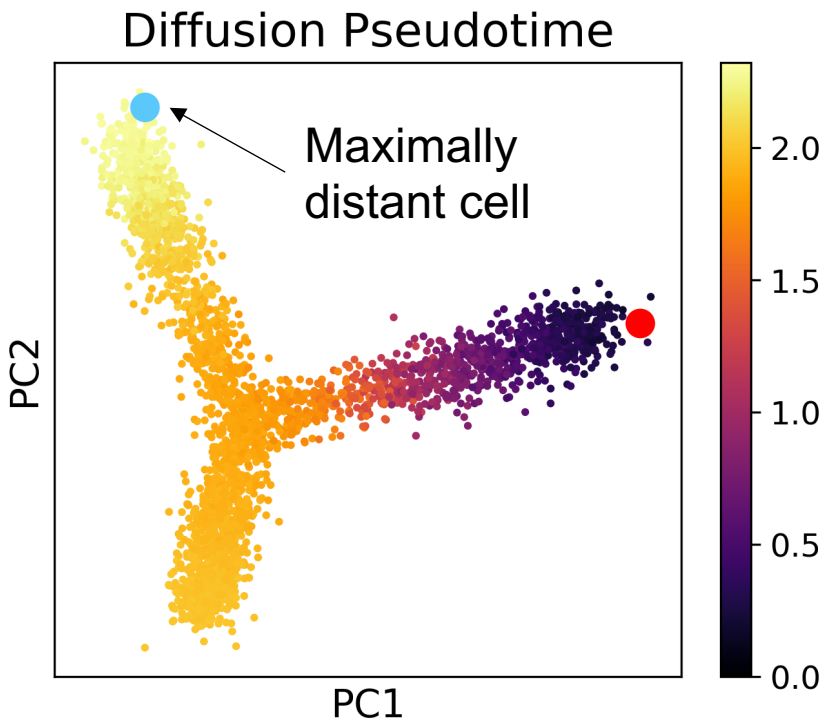
Using DPT to detect branches



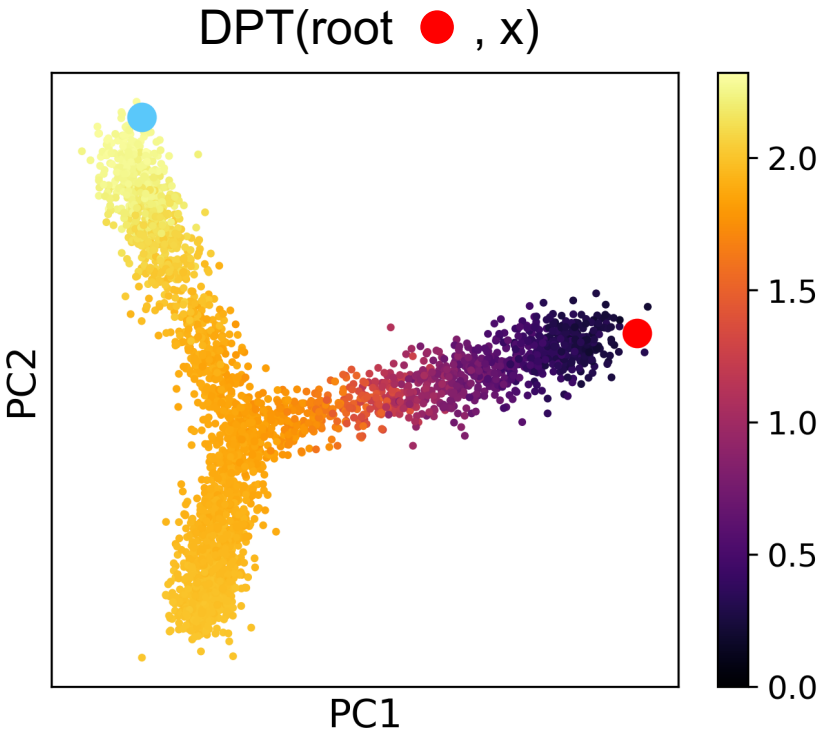


Using DPT to detect branches

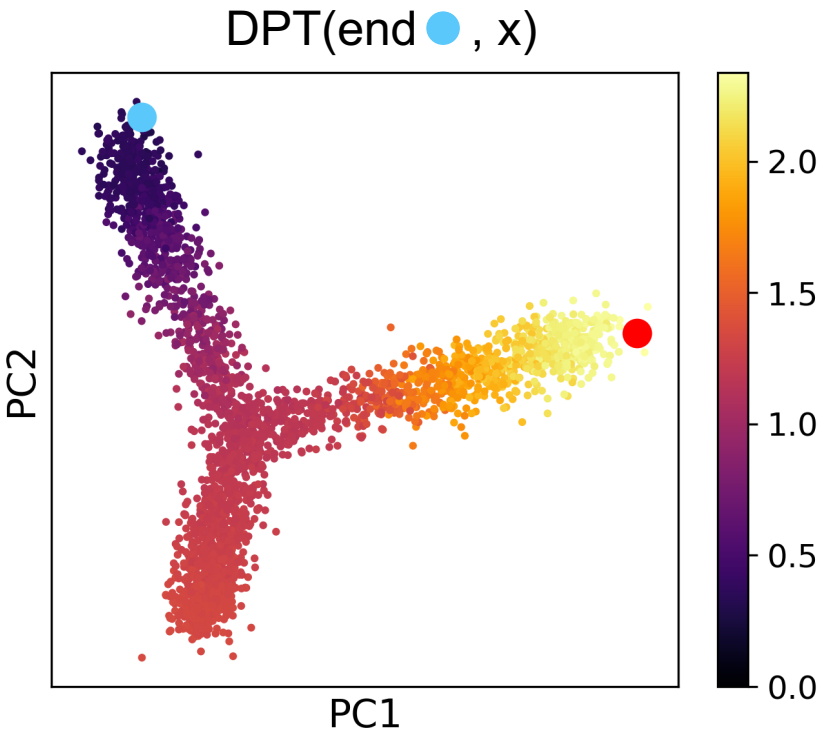




Using DPT to detect branches



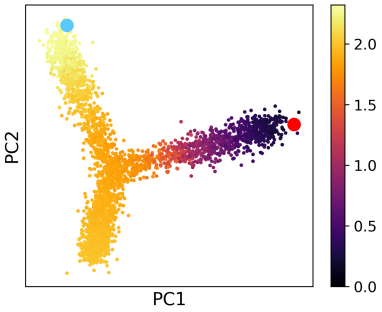
“Forward pseudotime”



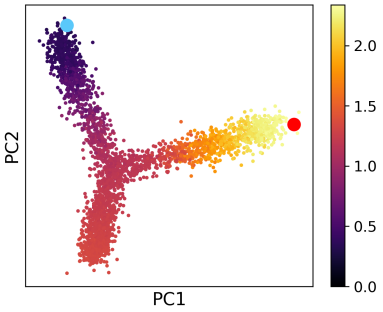
“Backward pseudotime”

Using DPT to detect branches

DPT(root ● , x)

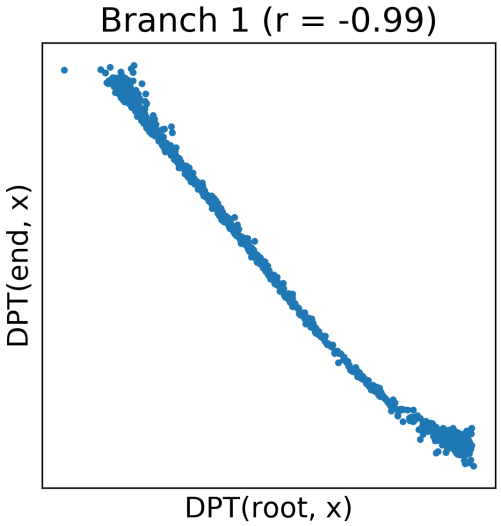


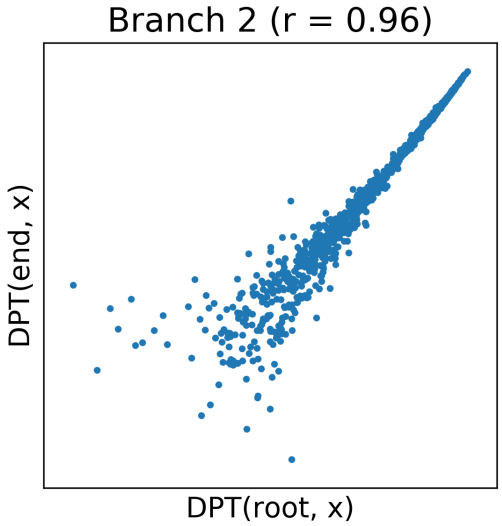
DPT(end ● , x)

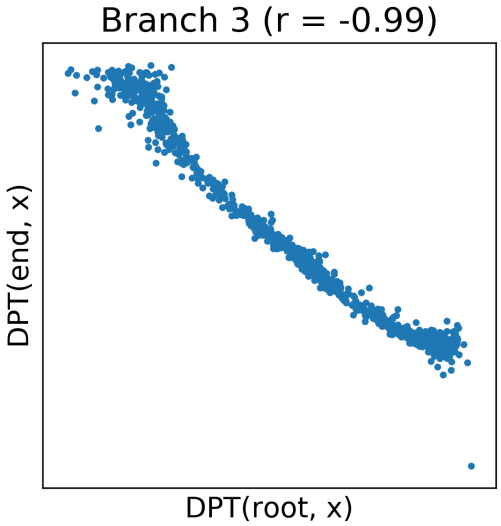


Cells on the same “branch” as the root and end cell have negative DPT correlation.

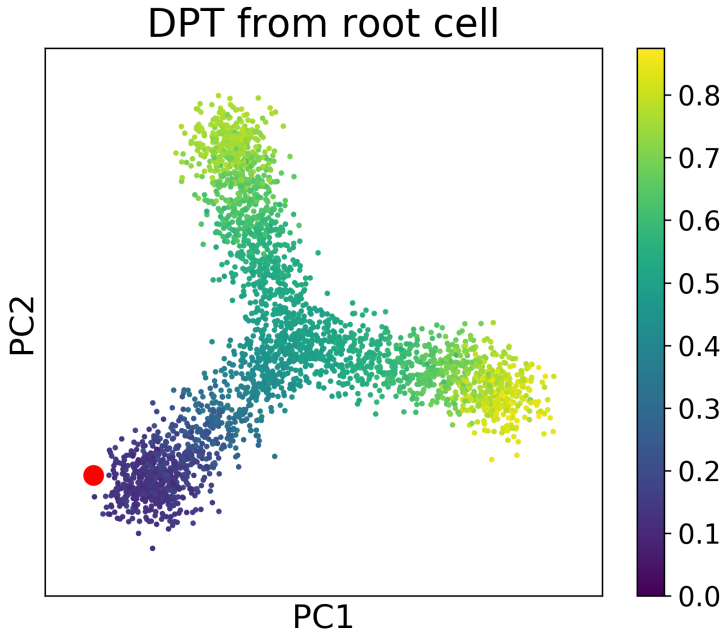
Cells on different branches have positive DPT correlation

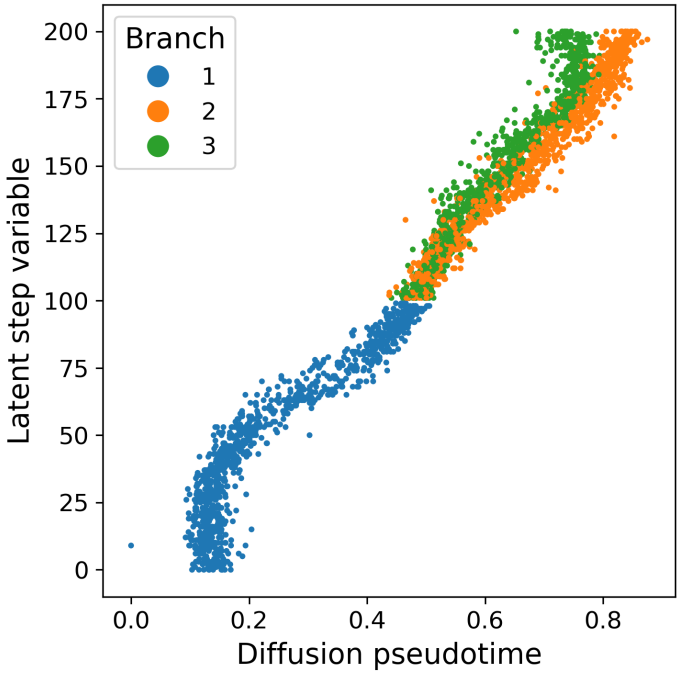






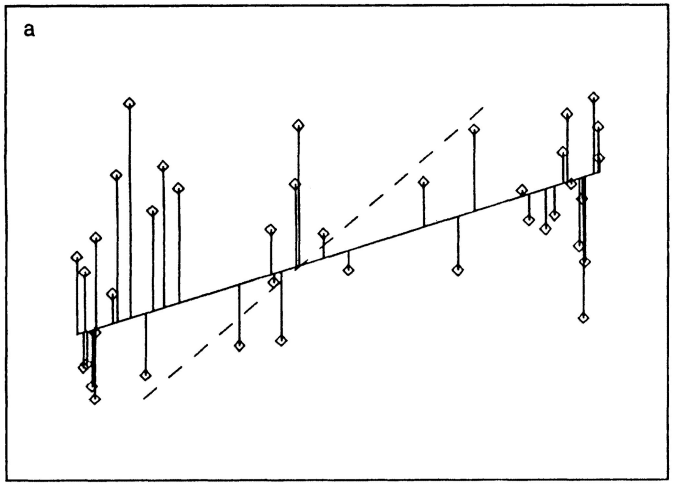
Exercise – Run diffusion pseudotime from scratch



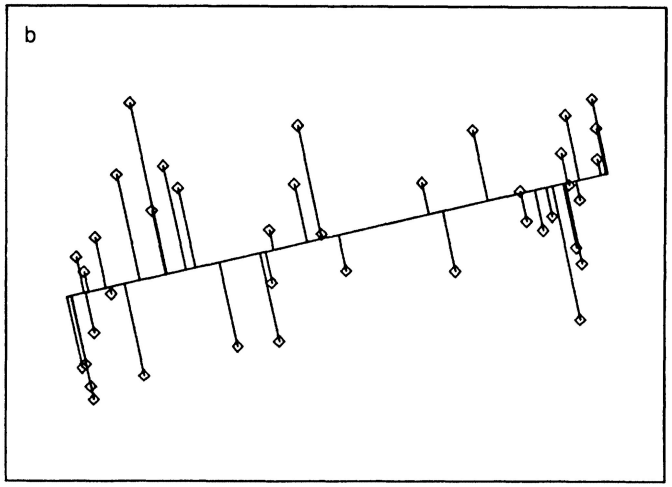


Principal curves – non-linear PCA

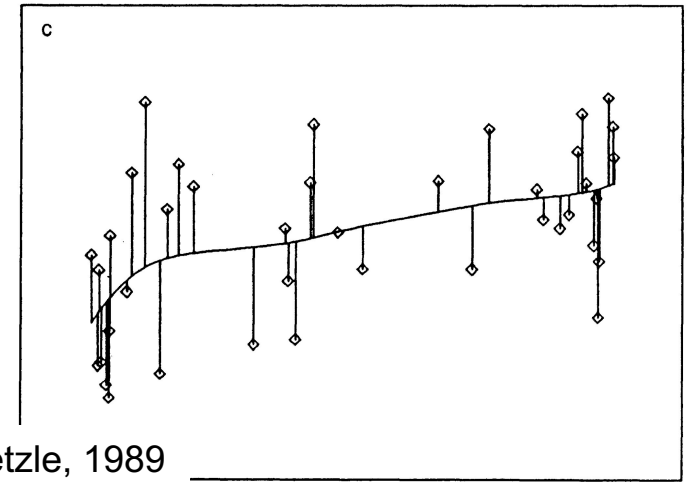
Linear Regression



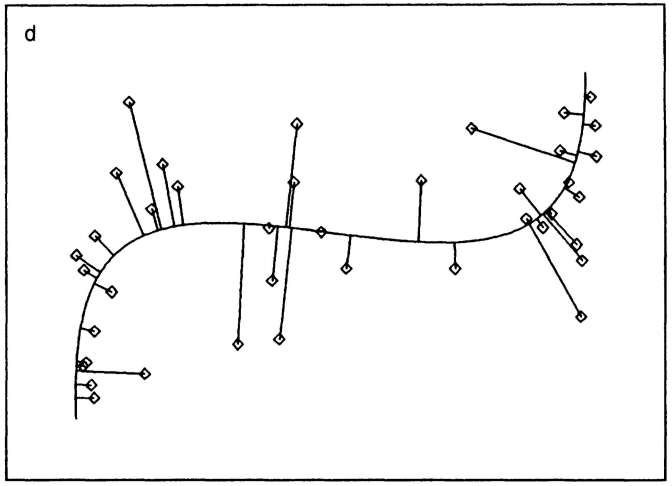
PCA



Non-linear Regression

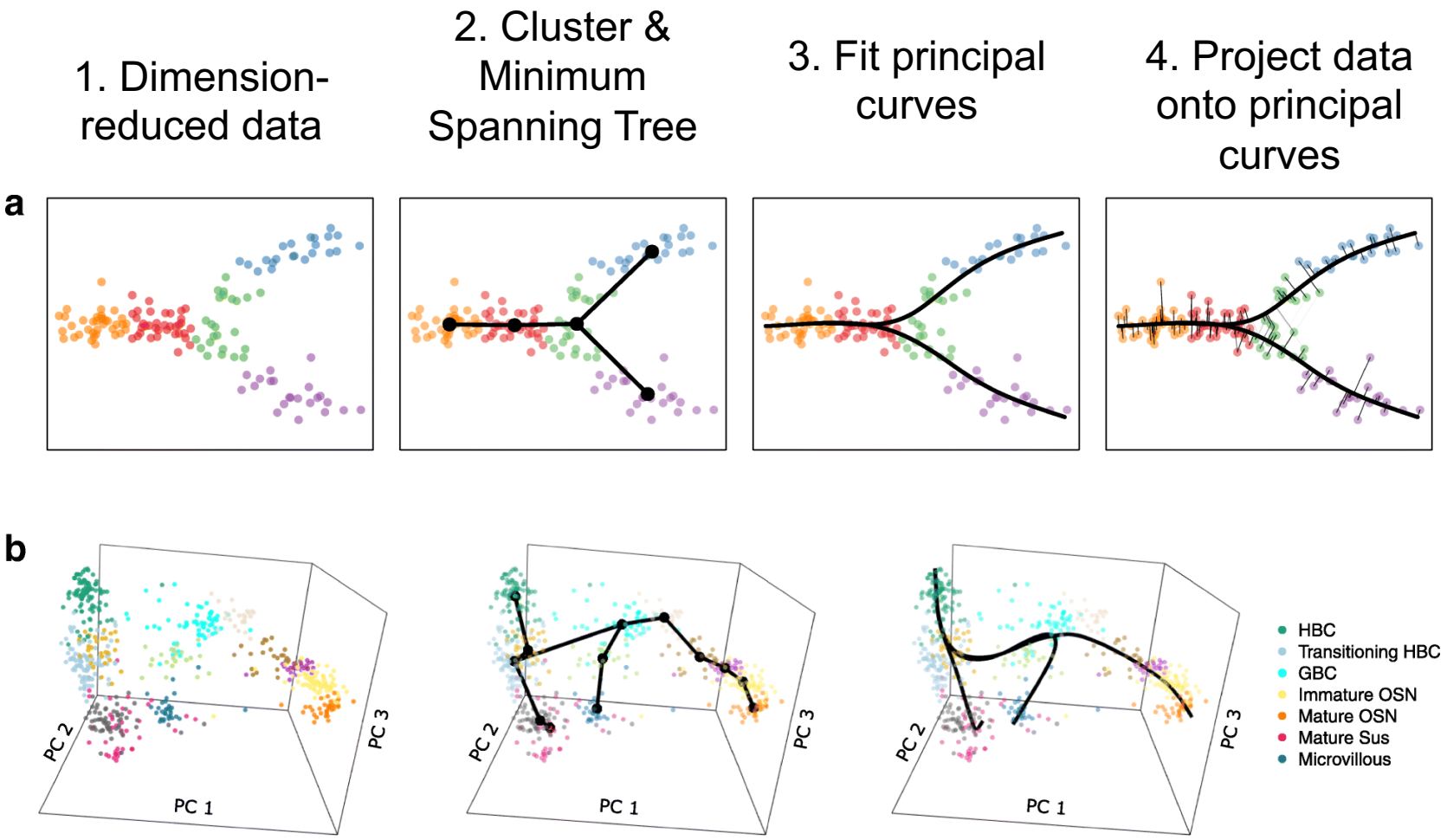


Principal Curves



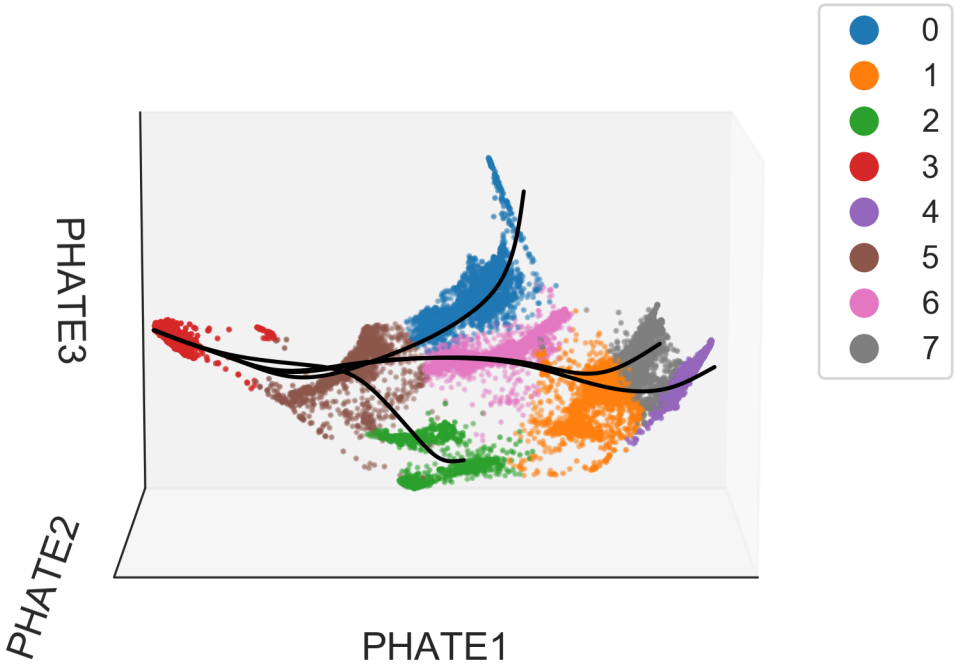
Hastie and Stuetzle, 1989

Slingshot - Street et al. 2018

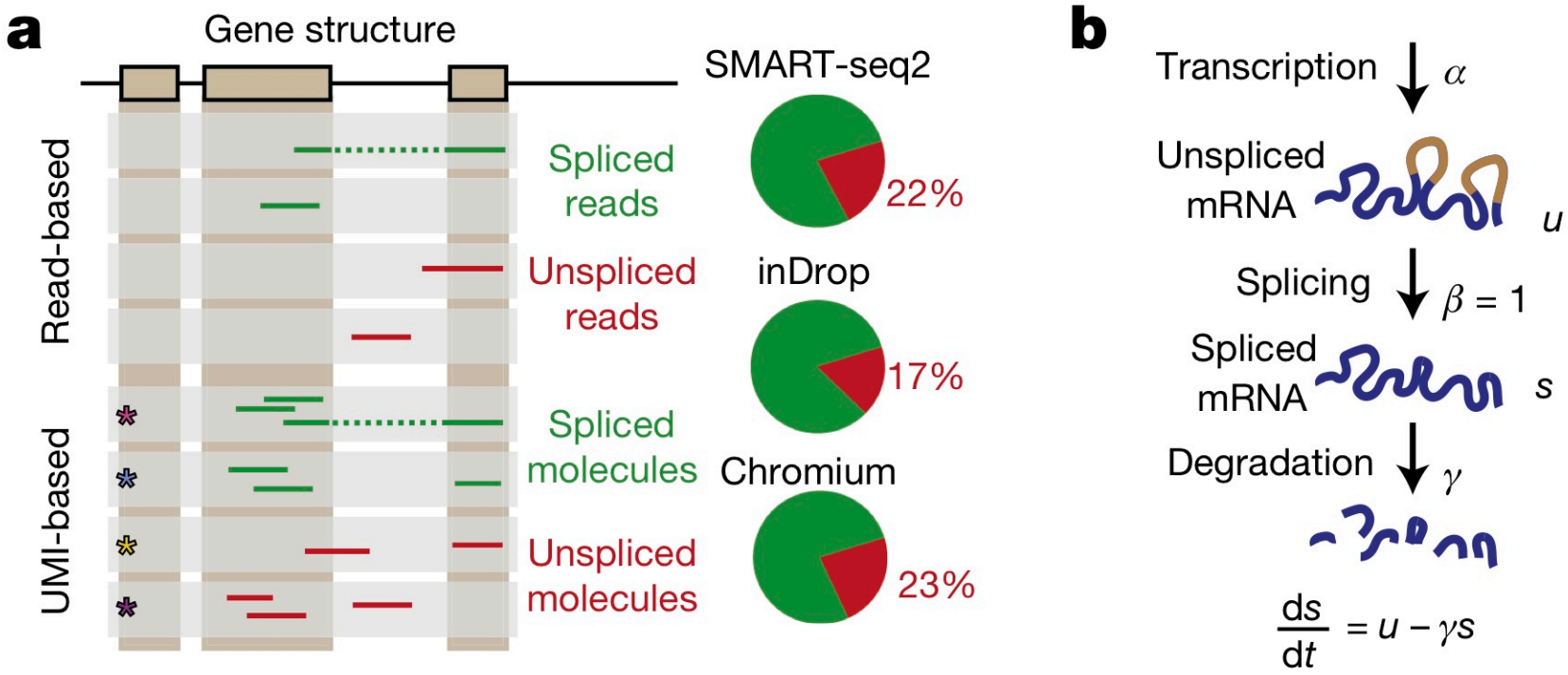


Exercise: Compare Slingshot and Diffusion Pseudotime on MEFs

Exercise: Apply Slingshot to EB data

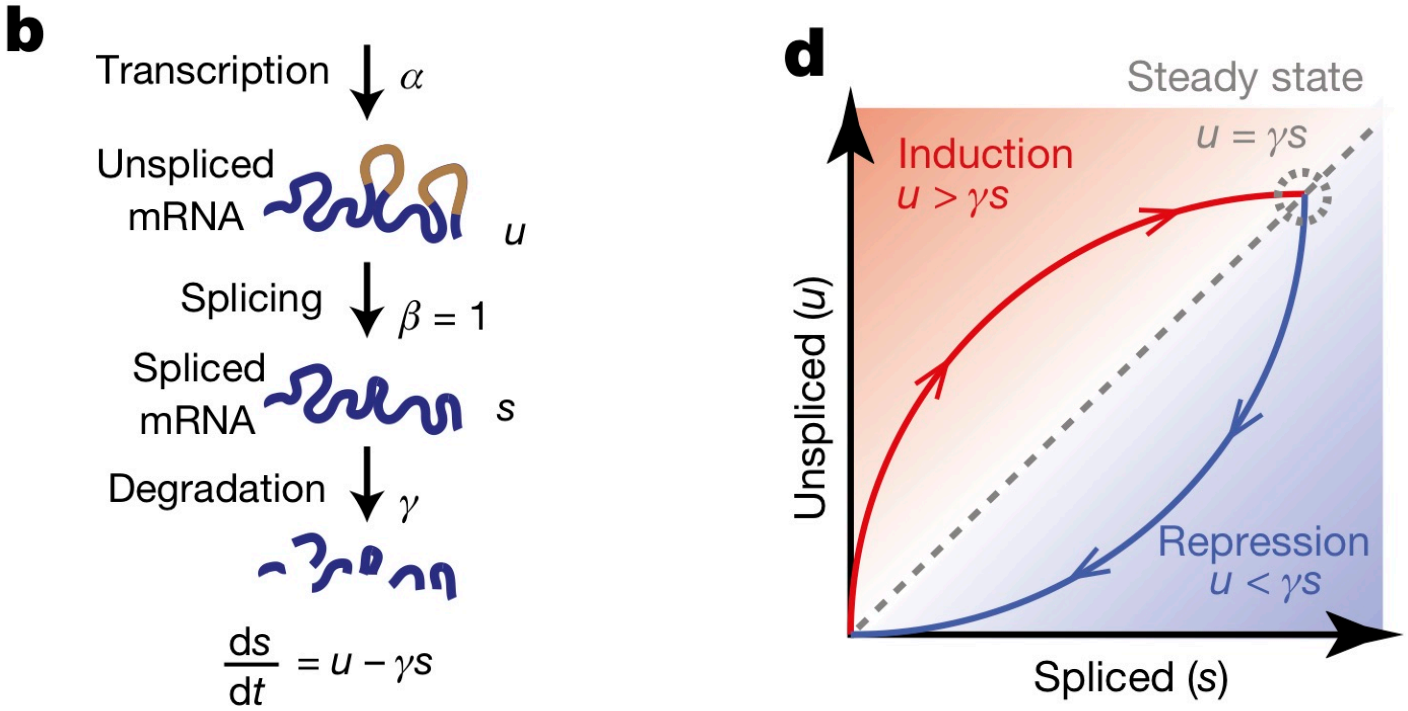


RNA Velocity: using gene splicing genes to infer future cell state



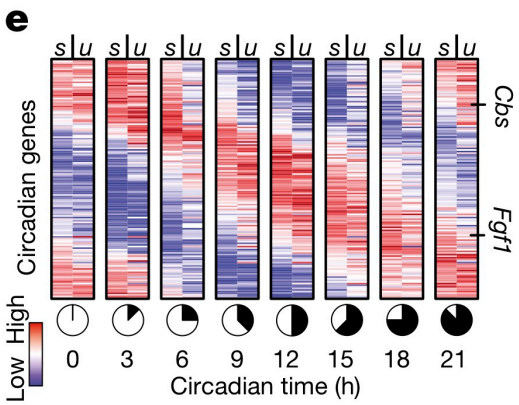
Le Manno, et al. Nature. 2018.

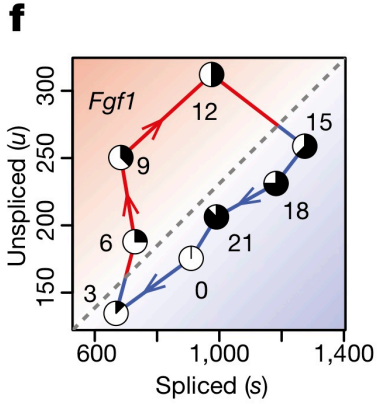
RNA Velocity: using gene splicing genes to infer future cell state

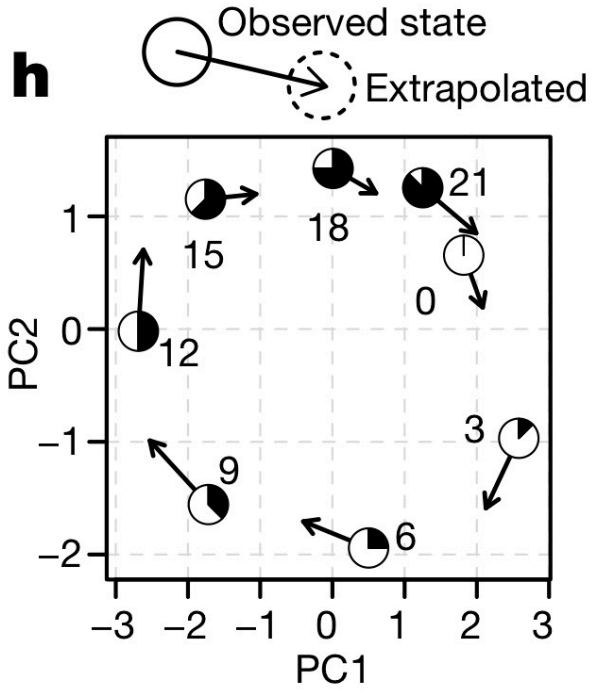


Le Manno, et al. Nature. 2018.

RNA velocity correctly infers circadian rhythm in mouse liver

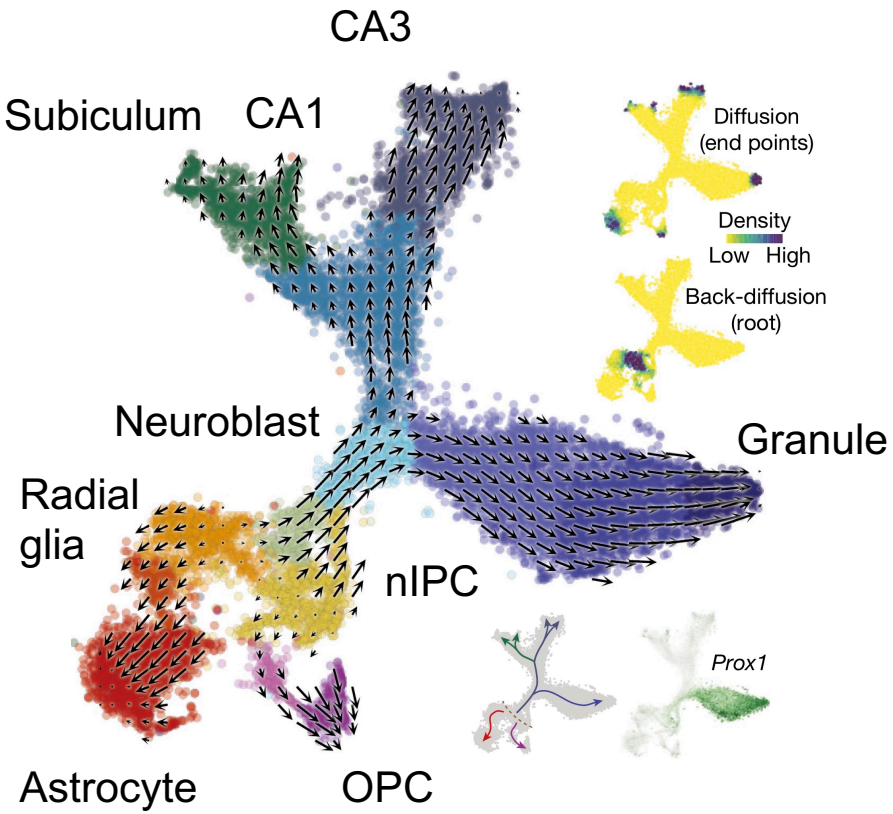




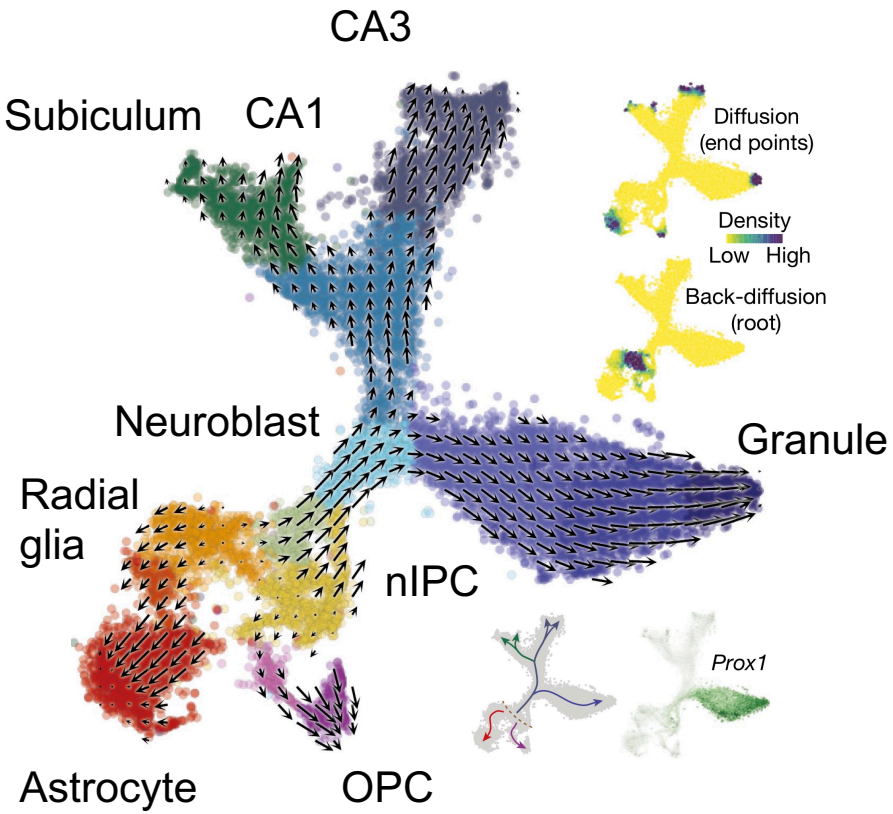


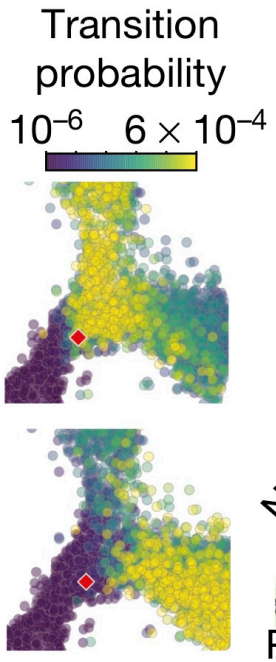
mRNA was measured in bulk over 24 hours in the mouse liver

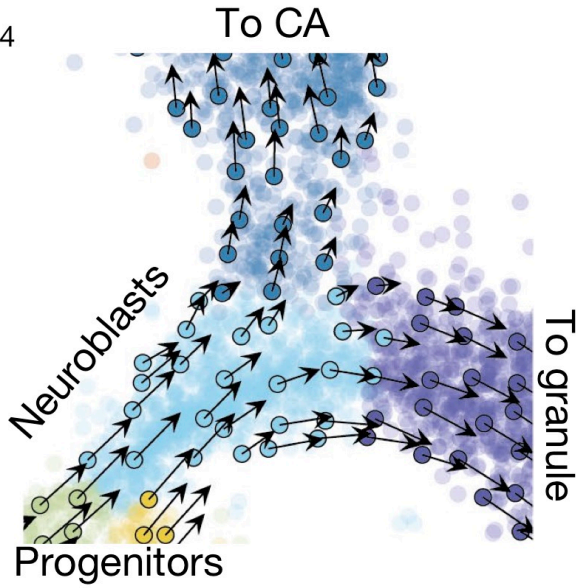
Neural fate decisions in the mouse hippocampus



Neural fate decisions in the mouse hippocampus







Exercise: RNA velocity on EB data