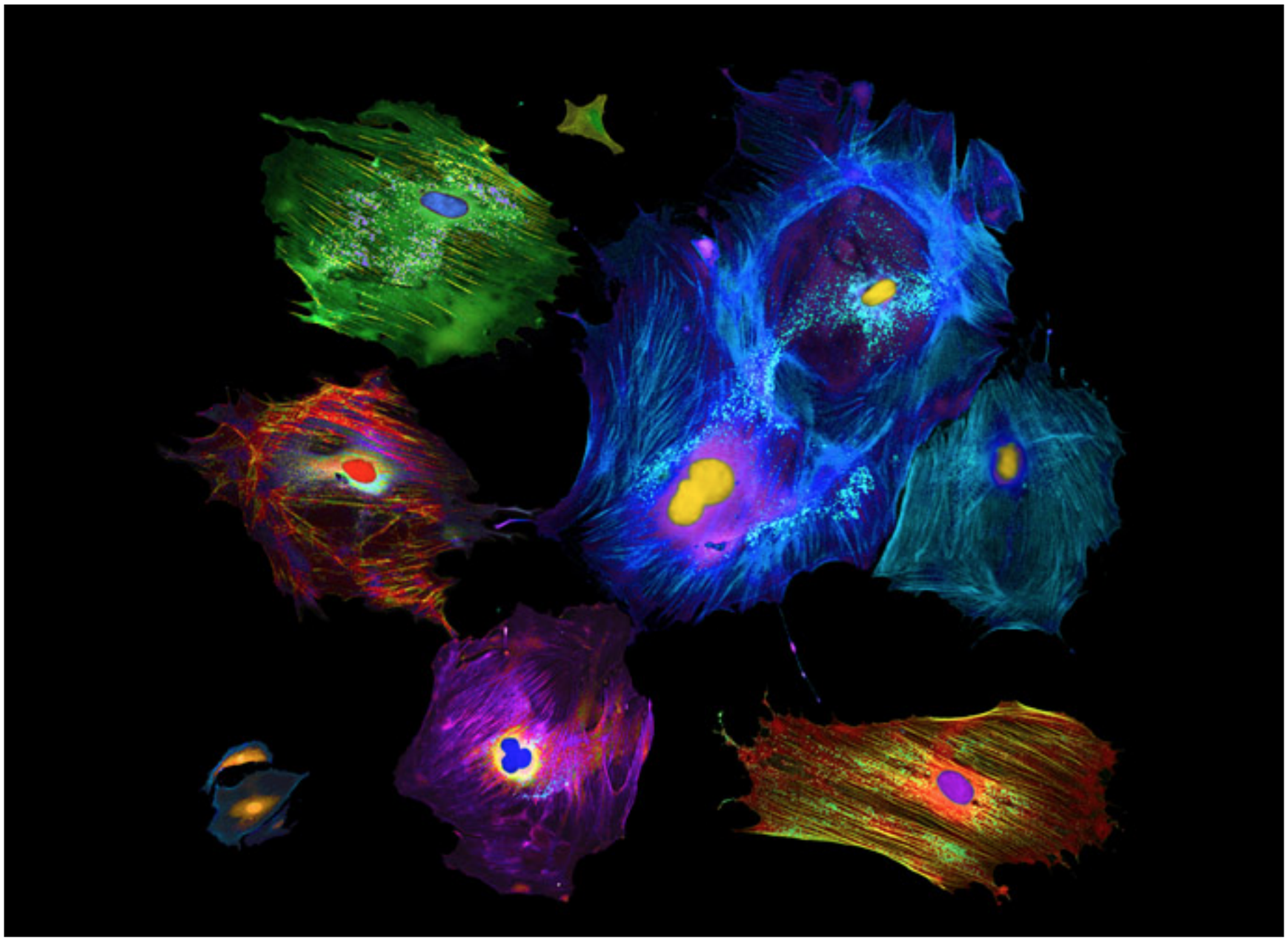
Analyzing single-cell data

Noah Apthorpe, Josh Fass, Maithra Raghu

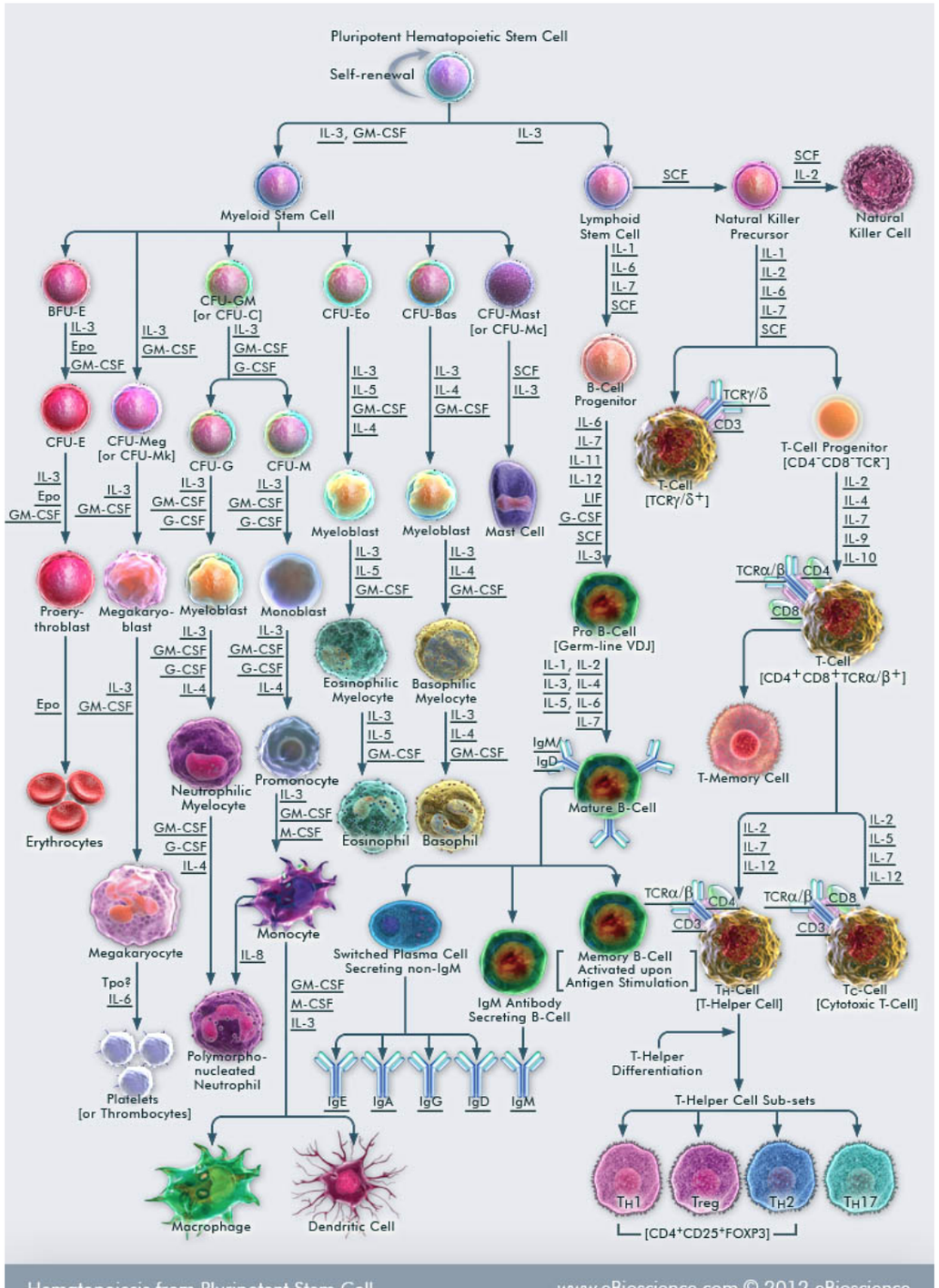
Outline

- Biological motivations
- Prior work
- Current work
  - Experiments with SPADE
  - Approximate “tree-preserving embedding”
  - Partial Least Squares-based analysis

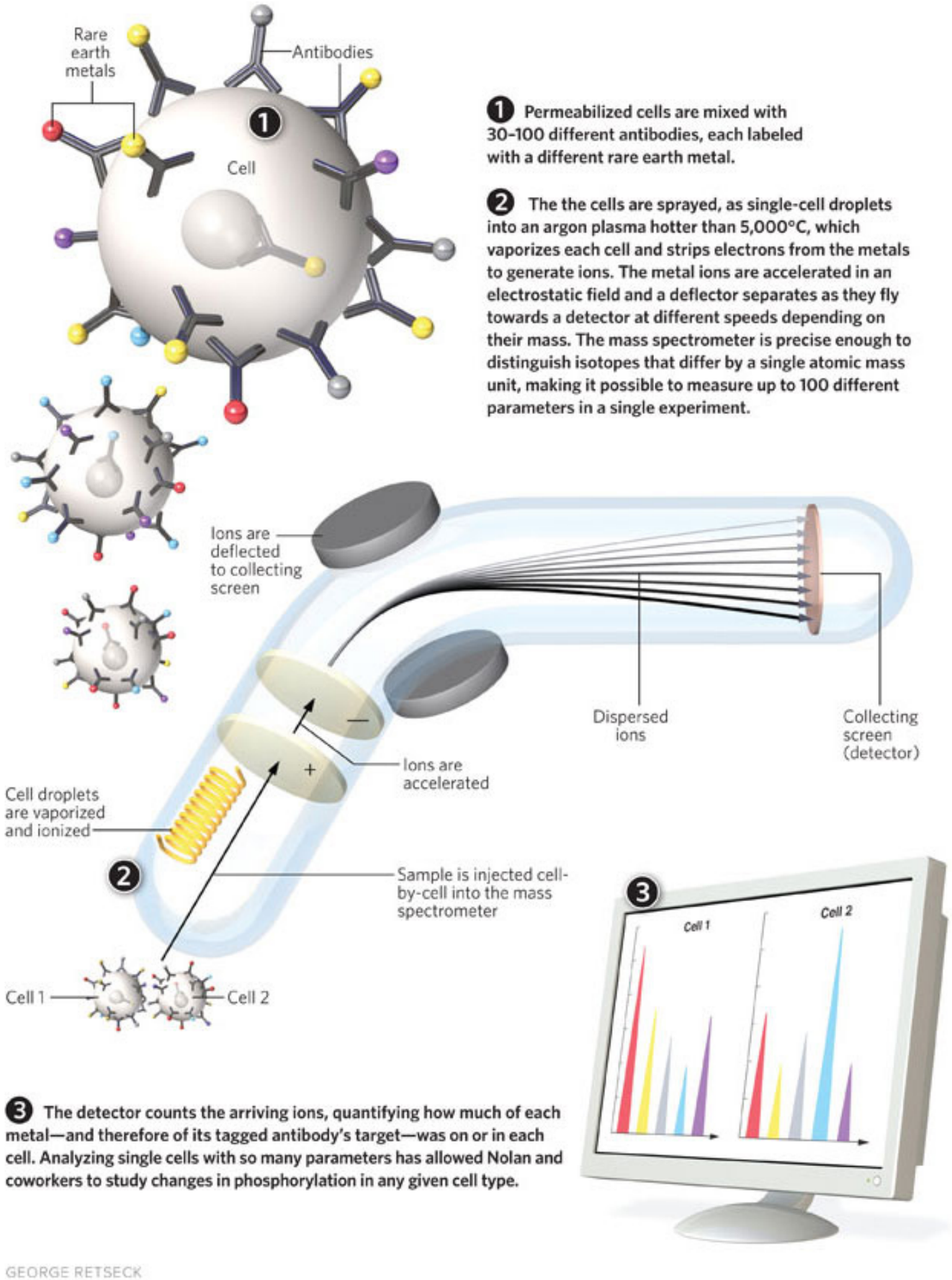
Every cell in the human body is different...



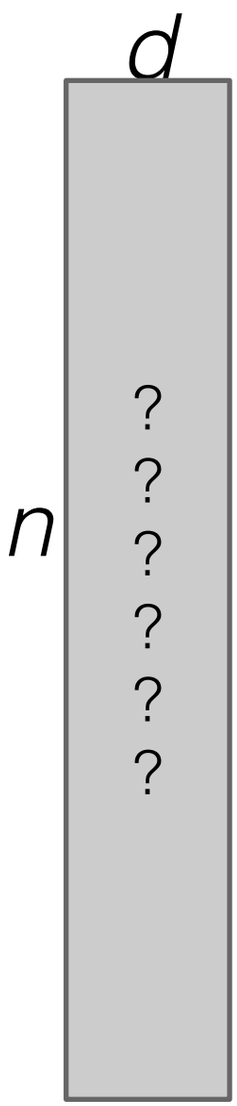
… due to repeated “specialization”



We can now make high-dimensional single-cell measurements...

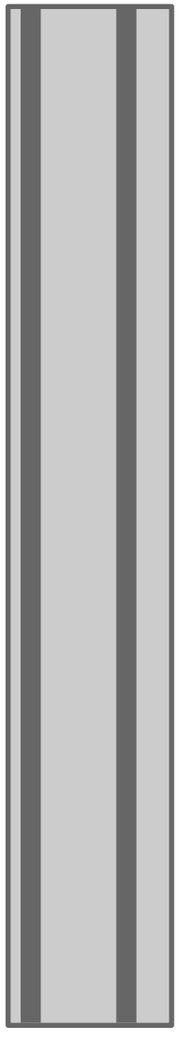


… resulting in large single-cell datasets



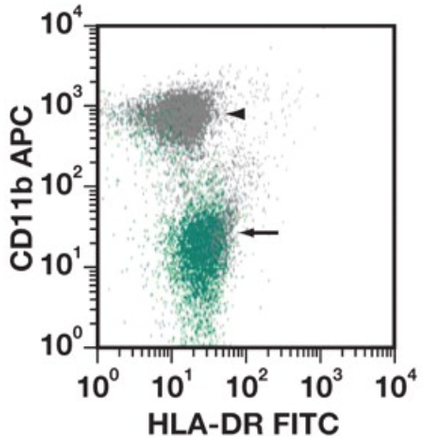
$n \sim 1{,}000{,}000$
$d \sim 50$

Traditionally analyzed by manual bivariate “gating”

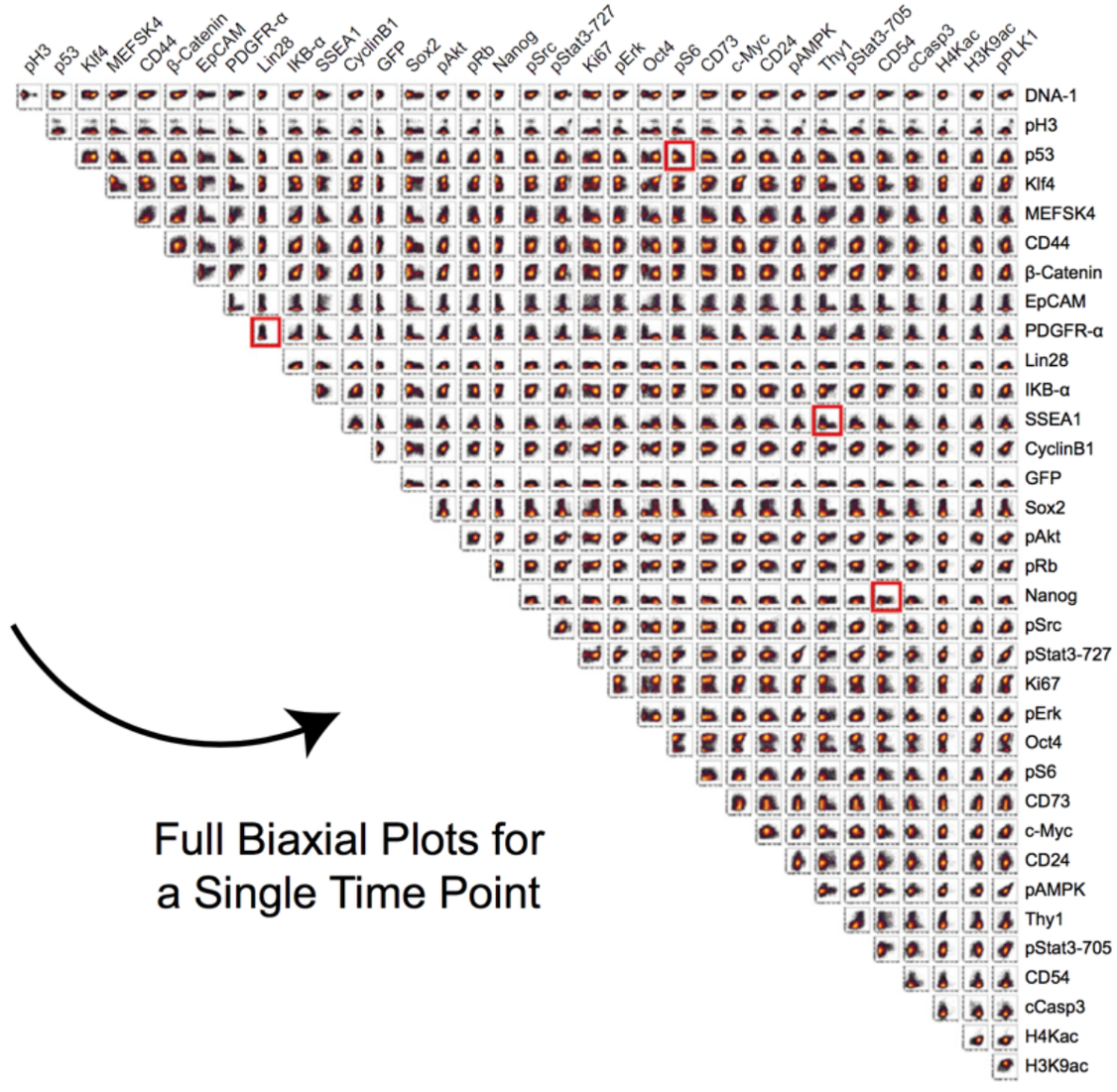


n ~ 1,000,000

d ~ 50



Make scatterplots of 2 columns at a time

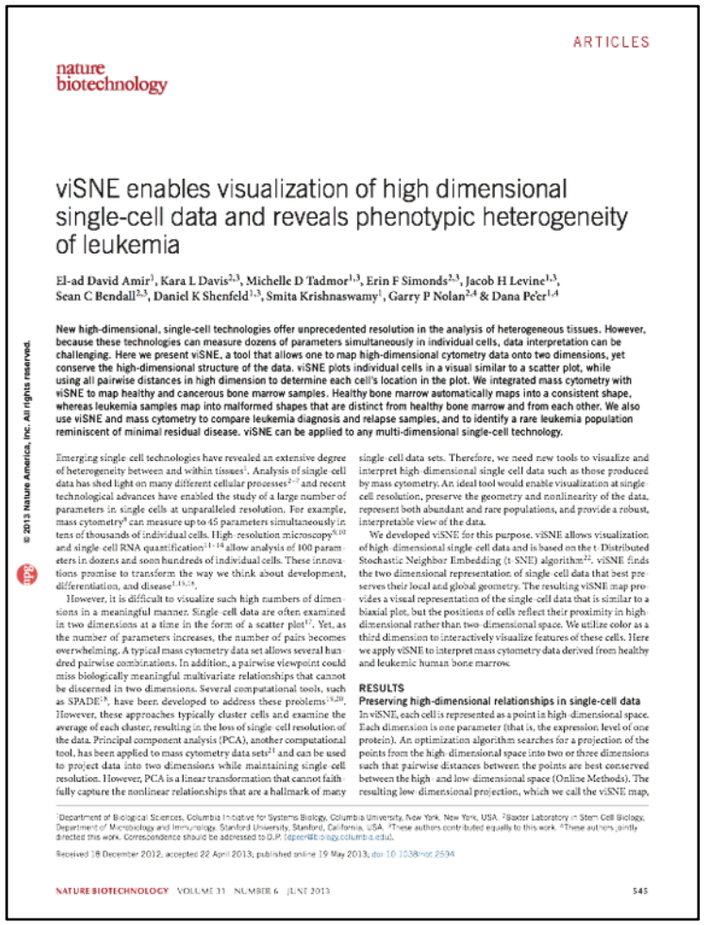


Full Biaxial Plots for a Single Time Point

Machine learning objectives

- High-resolution data available
- Big goals:
  - Visualize in 2D (**dimensionality reduction**)
  - Identify and characterize subtypes (**clustering**)
  - Understand structure of cellular variability (**representation learning**)
  - Understand stem cell reprogramming (**regression?**)

Recent high-profile work: domain-specific ML algorithms



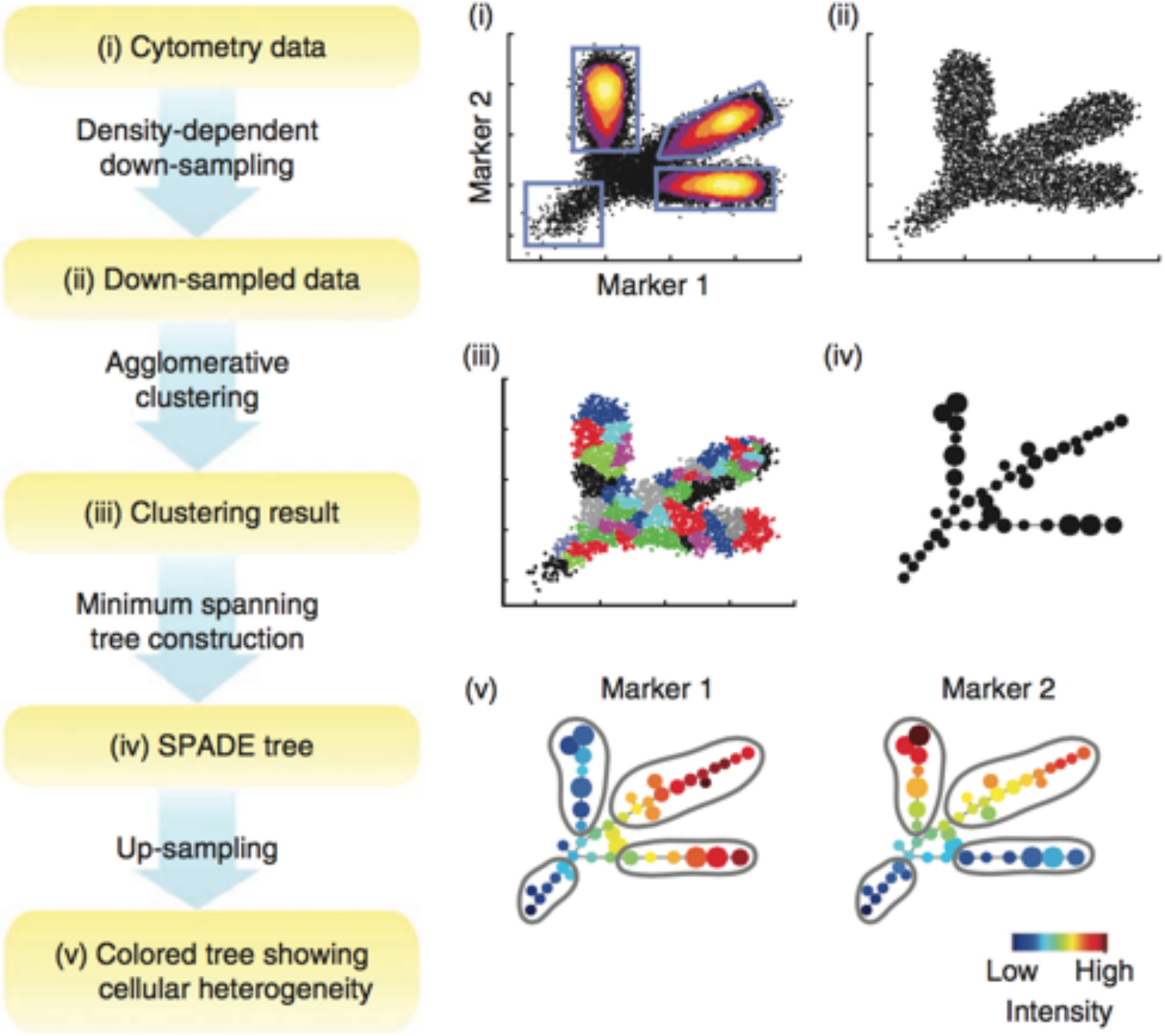
Dimensionality reduction

Structured output prediction

Limitations of these domain-specific ML approaches

- Complicated pipelines with many user-defined parameters
- Incompletely specified preprocessing
- Limited probabilistic motivation
- **Do they distort or exaggerate structure in data?**

Example: SPADE



SPADE: Overview

Design goals

- 2D visualization
- Illustrate clusters
- Recover likely tree of cellular differentiation events

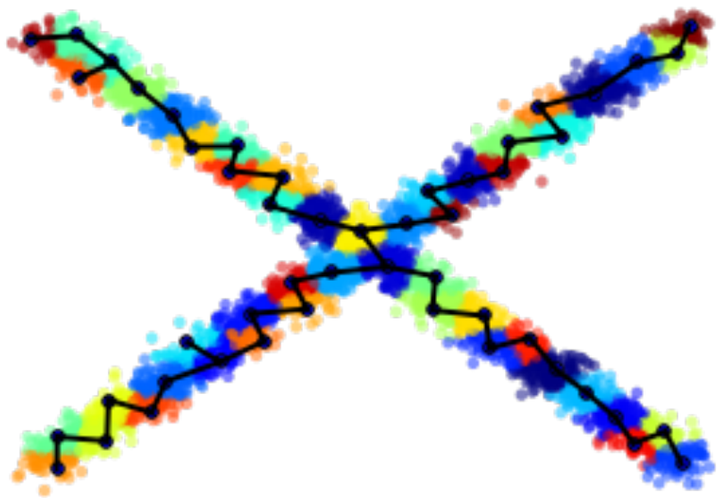
Limitations

- Always returns one tree
- Returns inconsistent trees when run repeatedly

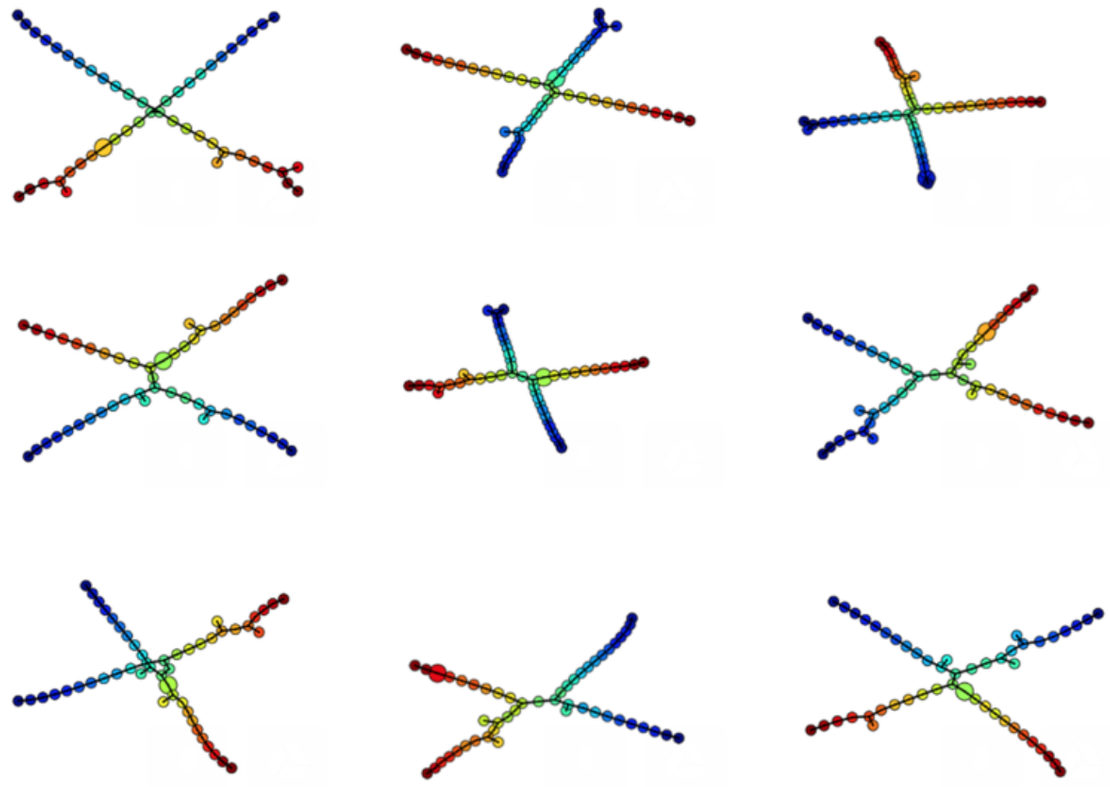
Our work

- **SPADE: consistency experiments**
- SPADE: density estimation
- Approximate “tree-preserving embedding”
- Is FLOW-MAP misleading?

SPADE: Consistency experiments

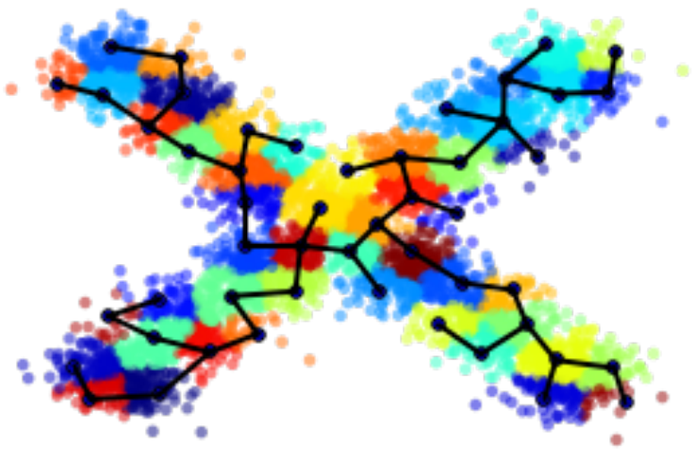


Synthetic data: long “branches”

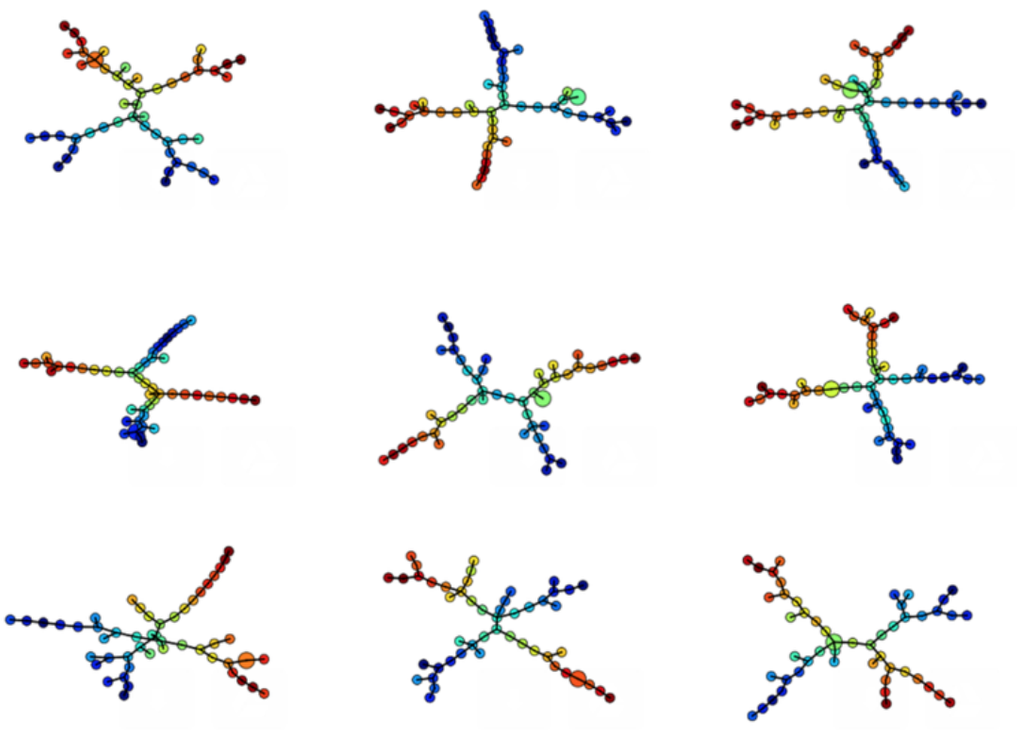


Outputs of repeated trials

SPADE: Consistency experiments

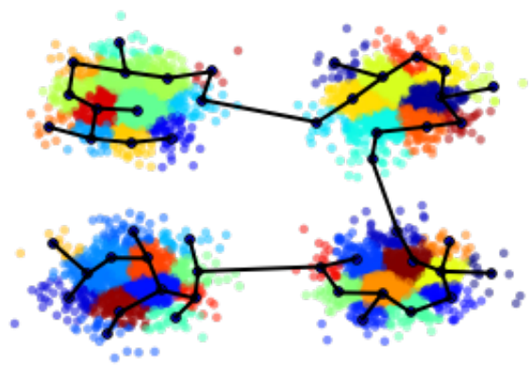


Synthetic data: shorter “branches”

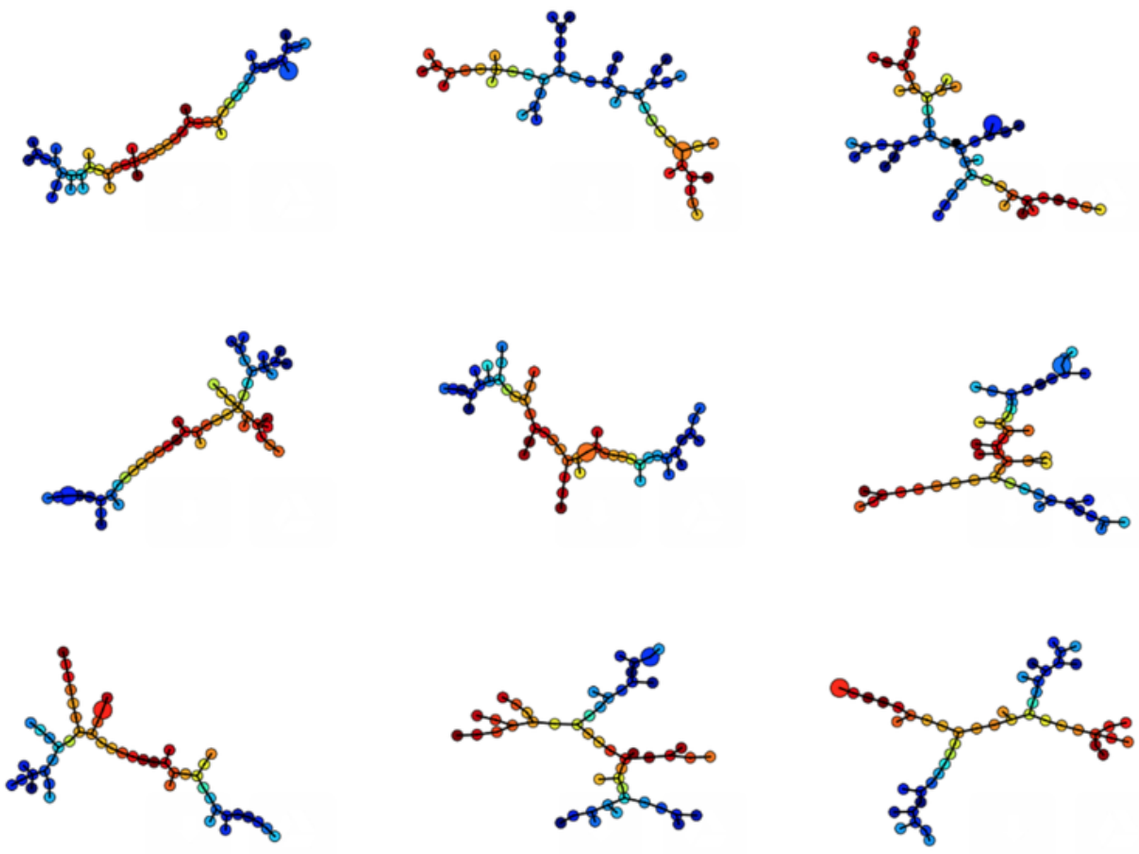


Outputs of repeated trials

SPADE: Consistency experiments



Synthetic data:
mixture of Gaussians

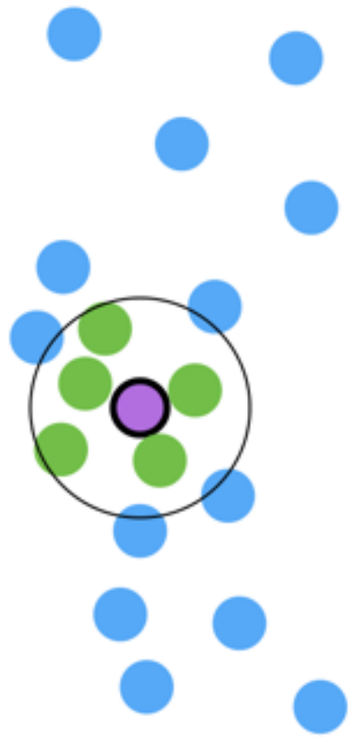


Outputs of repeated trials

Our work

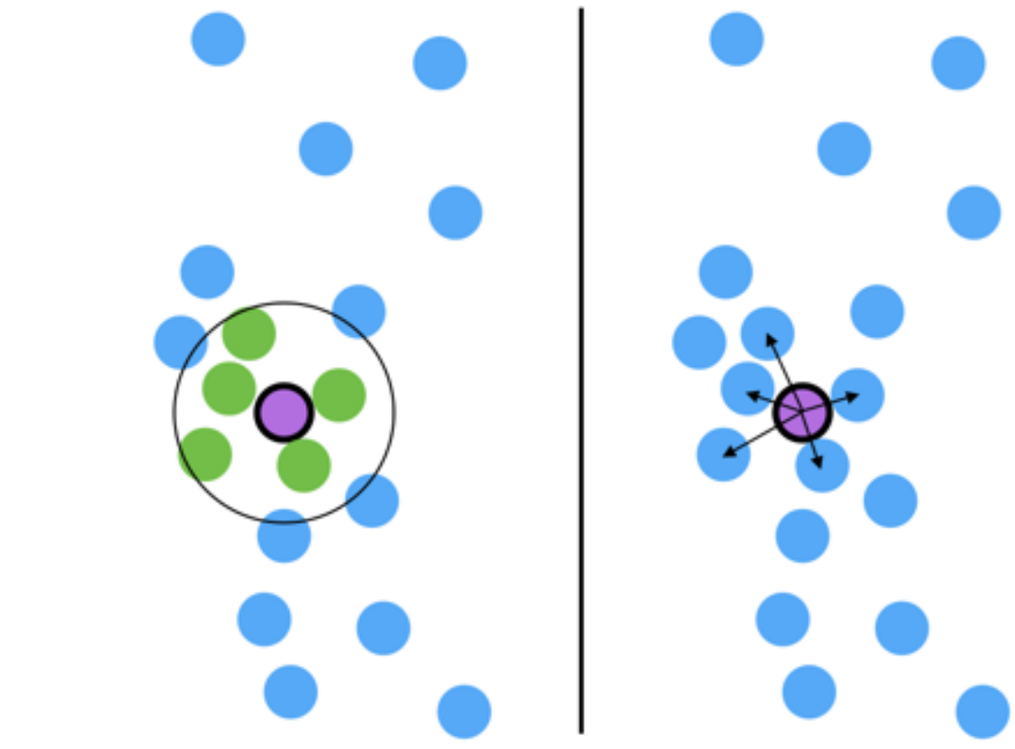
- SPADE: consistency experiments
- **SPADE: density estimation**
- Approximate “tree-preserving embedding”
- Is FLOW-MAP misleading?

SPADE: Non-standard density estimator, possibly inaccurate when *d*>2



SPADE approach:
count neighbors
within radius *r*

SPADE: More accurate density-estimators possible?



SPADE approach:
count neighbors
within radius *r*

Alternative: Measure
average distance to *k*
nearest neighbors

Preliminary results:
more stable when *d*>2

SPADE: Conclusions

- Complex algorithms have many possible failure points, and may not be rigorously tested
- Open need for this particular kind of structure-learning

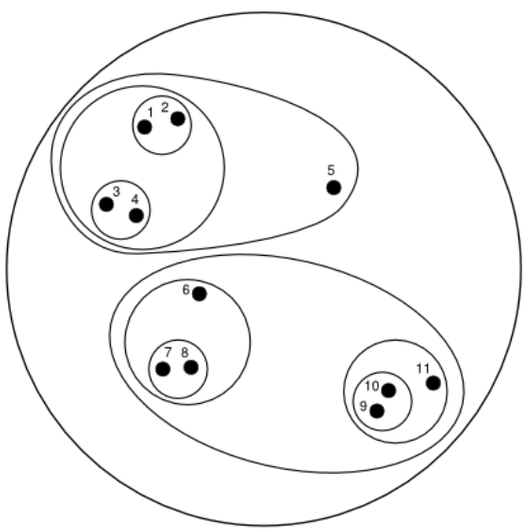
Our work

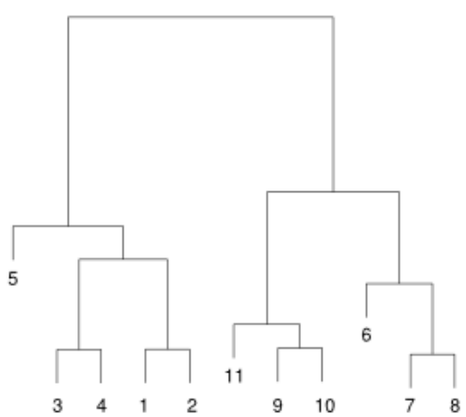
- SPADE: consistency experiments
- SPADE: density estimation
- **Approximate “tree-preserving embedding”**
- Is FLOW-MAP misleading?

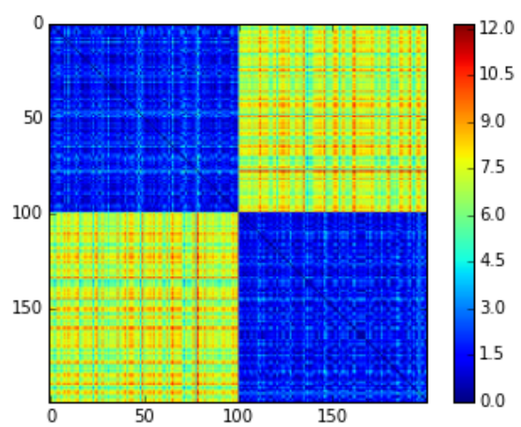
Approximate “tree-preserving embedding”

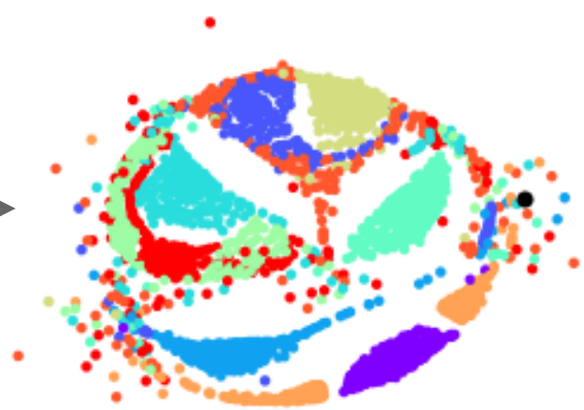
Goal: 2D visualization that preserves cluster structure
Tactic: approximately preserve hierarchical cluster tree

- Compute hierarchical clustering
- Extract cophenetic distance matrix
- Apply Multidimensional Scaling to distance matrix







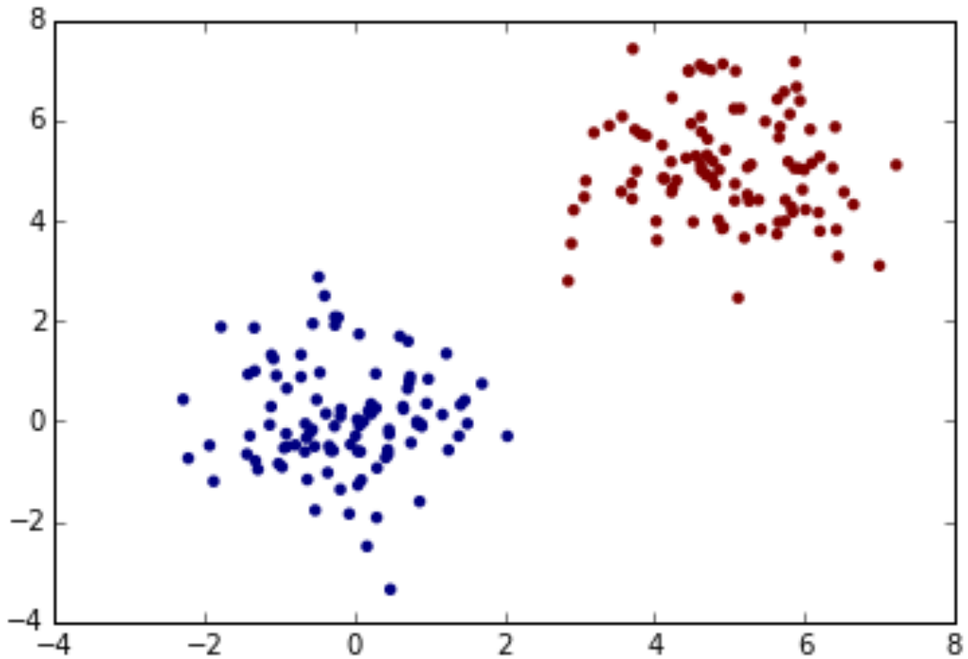


Hierarchical agglomerative clustering

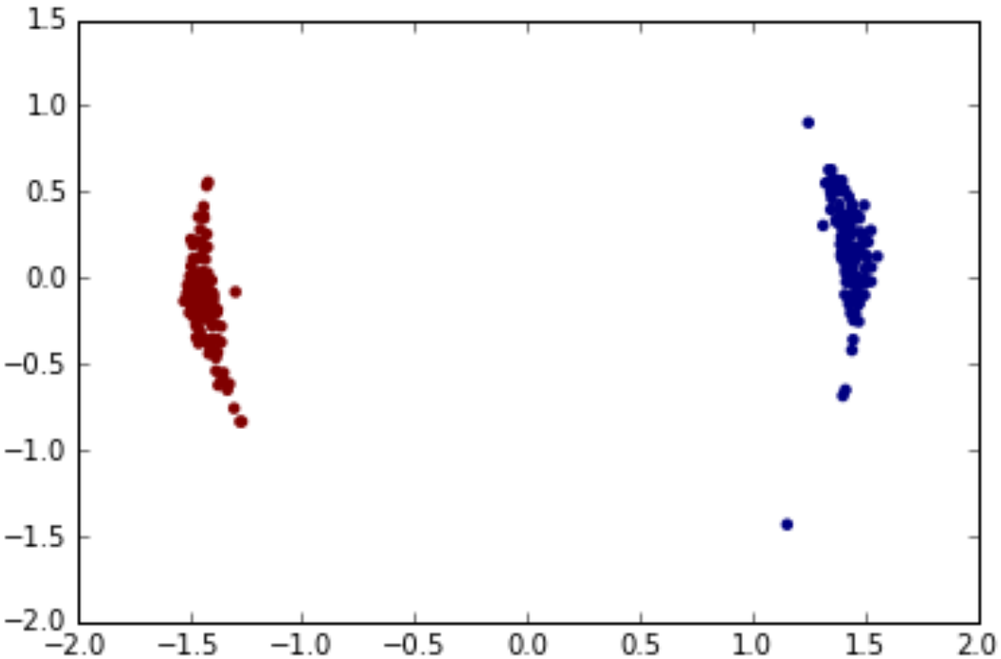
Cophenetic distance matrix

2D embedding

Approximate TPE: toy problem

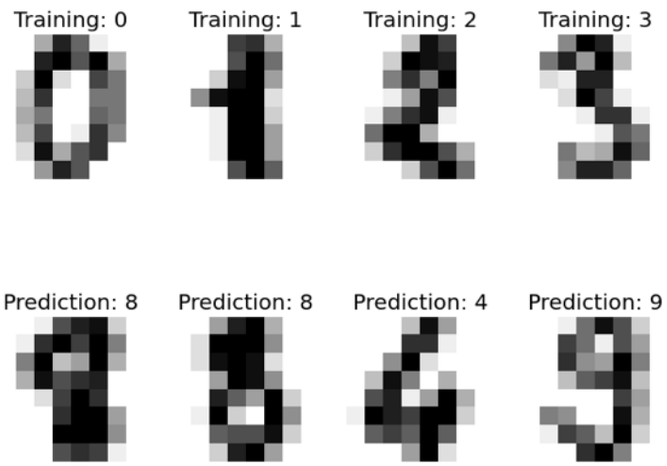


Synthetic data: 2D mixture of Gaussians

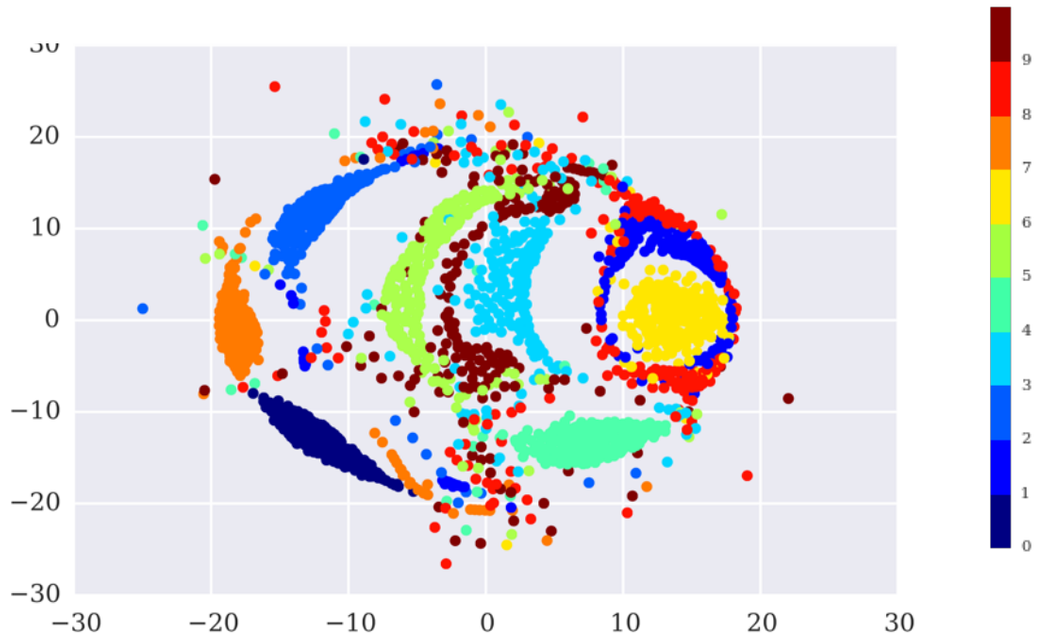


Approximate TPE solution

Approximate TPE: benchmark data



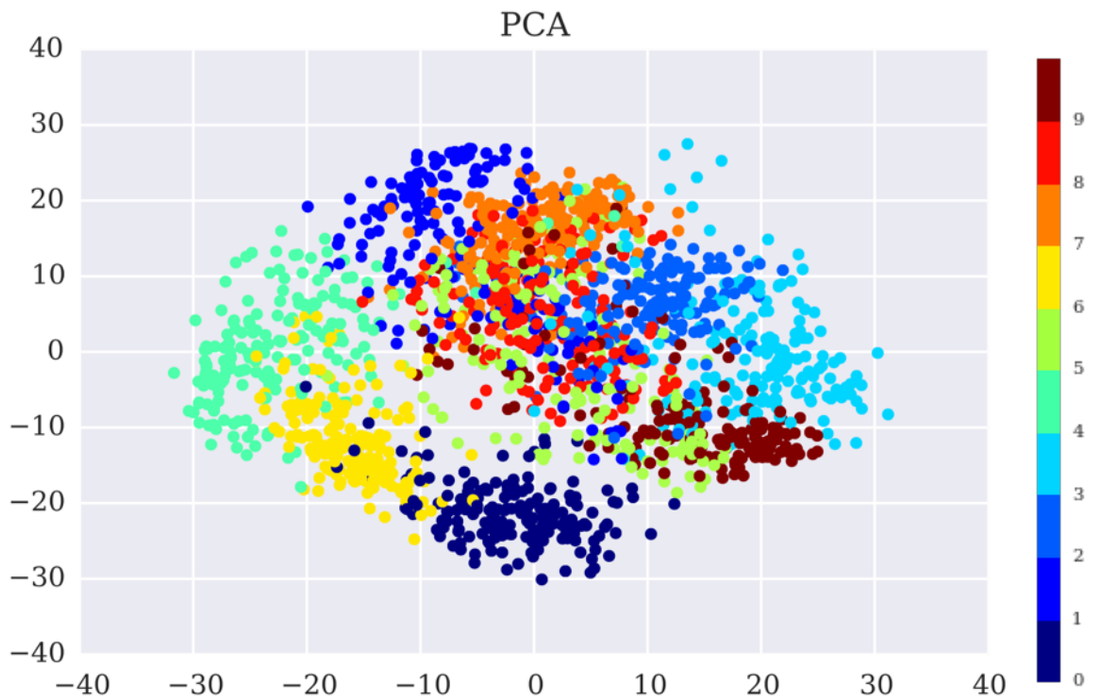
Example dataset: handwritten digits (from `sklearn.datasets`, *n*=1797, *d*=64)



Approximate TPE solution

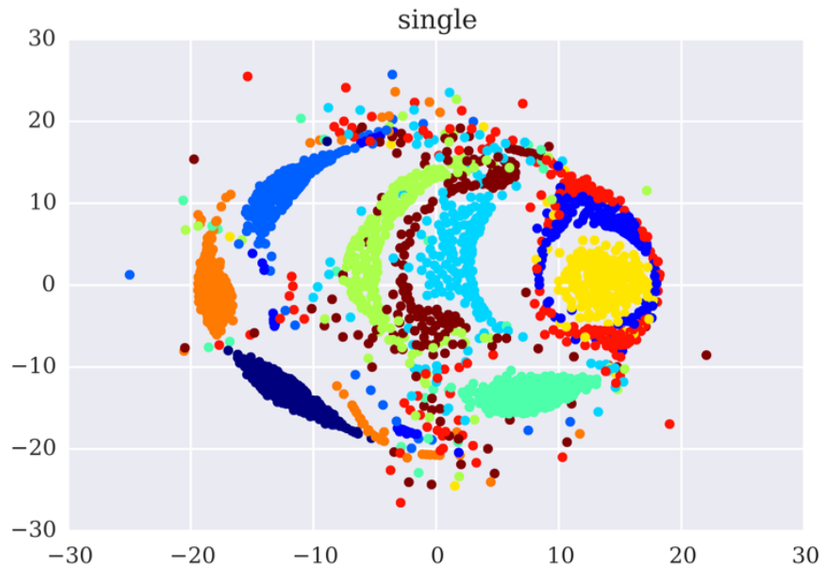
Comparing embedding quality

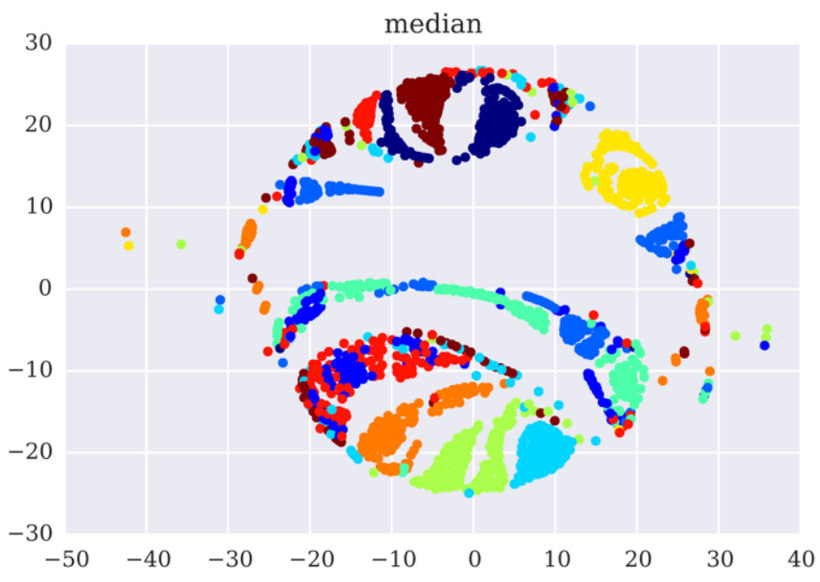
Baseline (PCA)



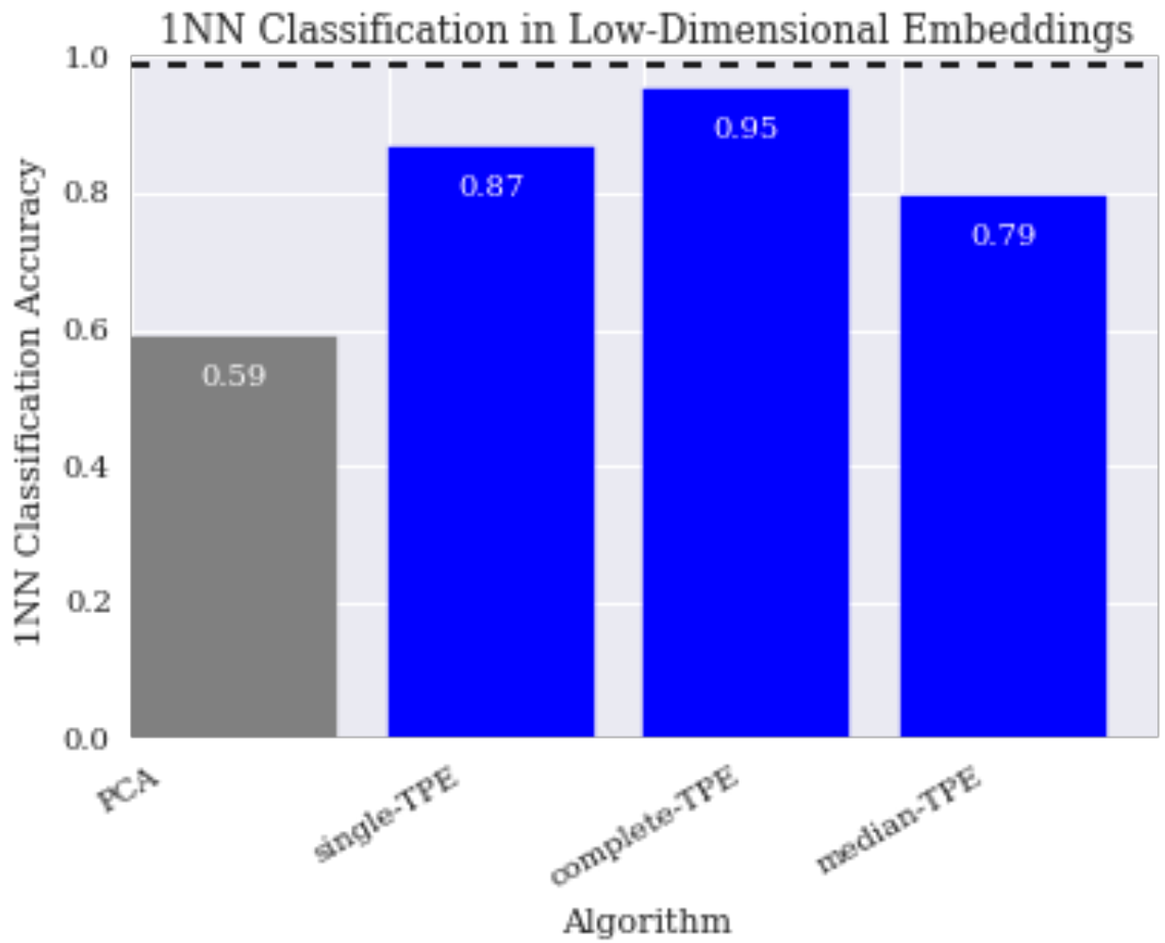
vs.

Approx. TPE with
- single-linkage
- median- linkage



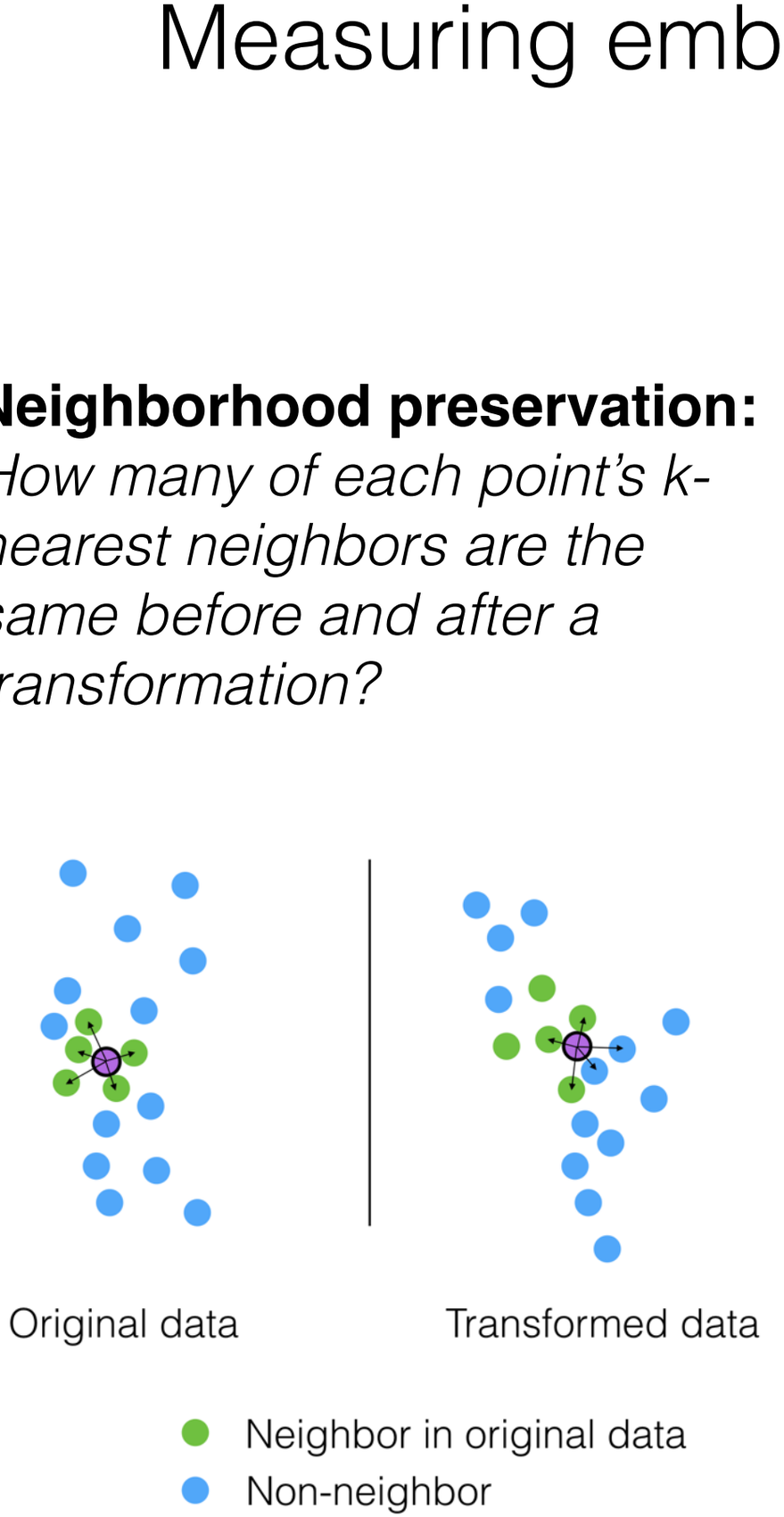


Measuring embedding fidelity: 1NN classification accuracy

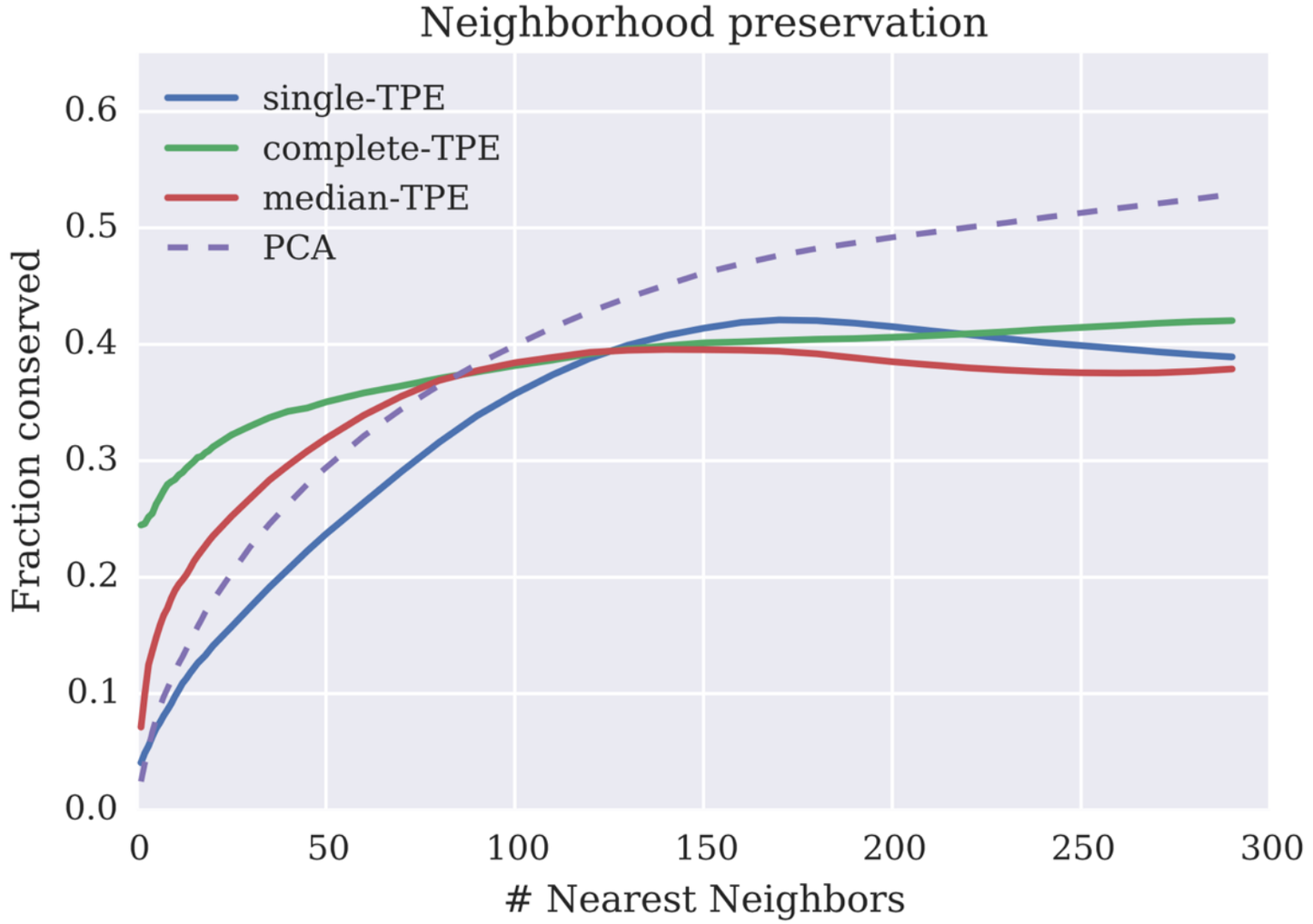


Measuring embedding fidelity: neighborhood preservation

Neighborhood preservation:
How many of each point's k-nearest neighbors are the same before and after a transformation?



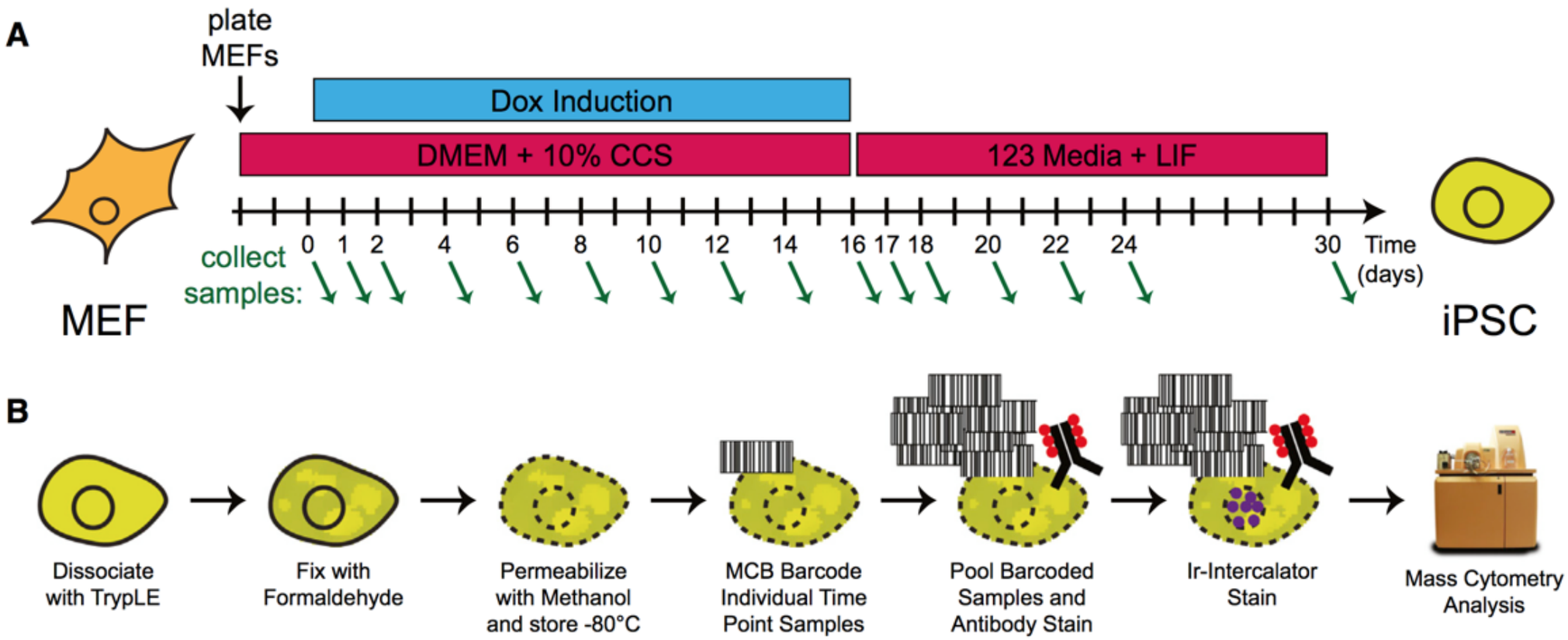
Results:



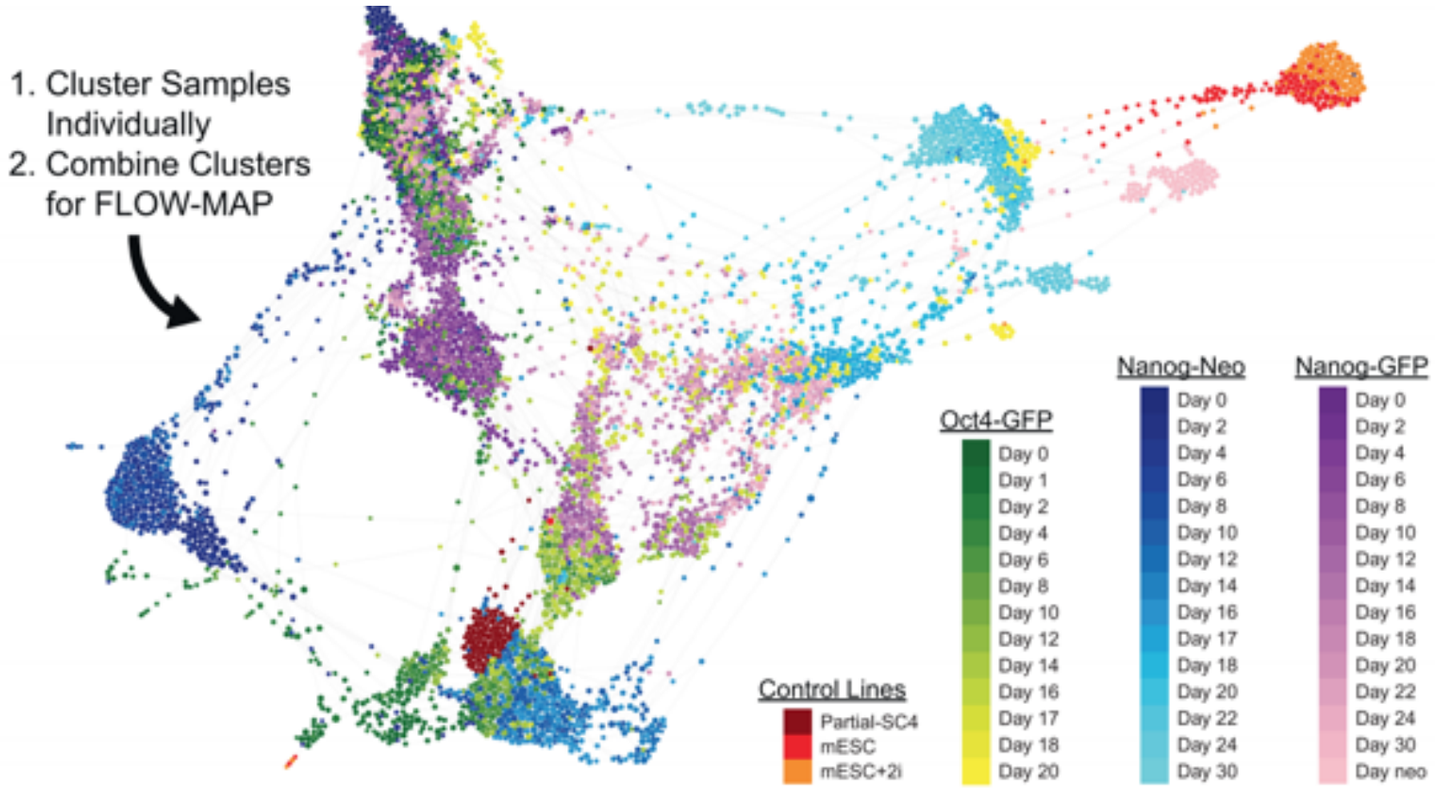
Our work

- SPADE: consistency experiments
- SPADE: density estimation
- Approximate “tree-preserving embedding”
- **Is FLOW-MAP misleading?**

New dataset: population snapshots during stem cell reprogramming



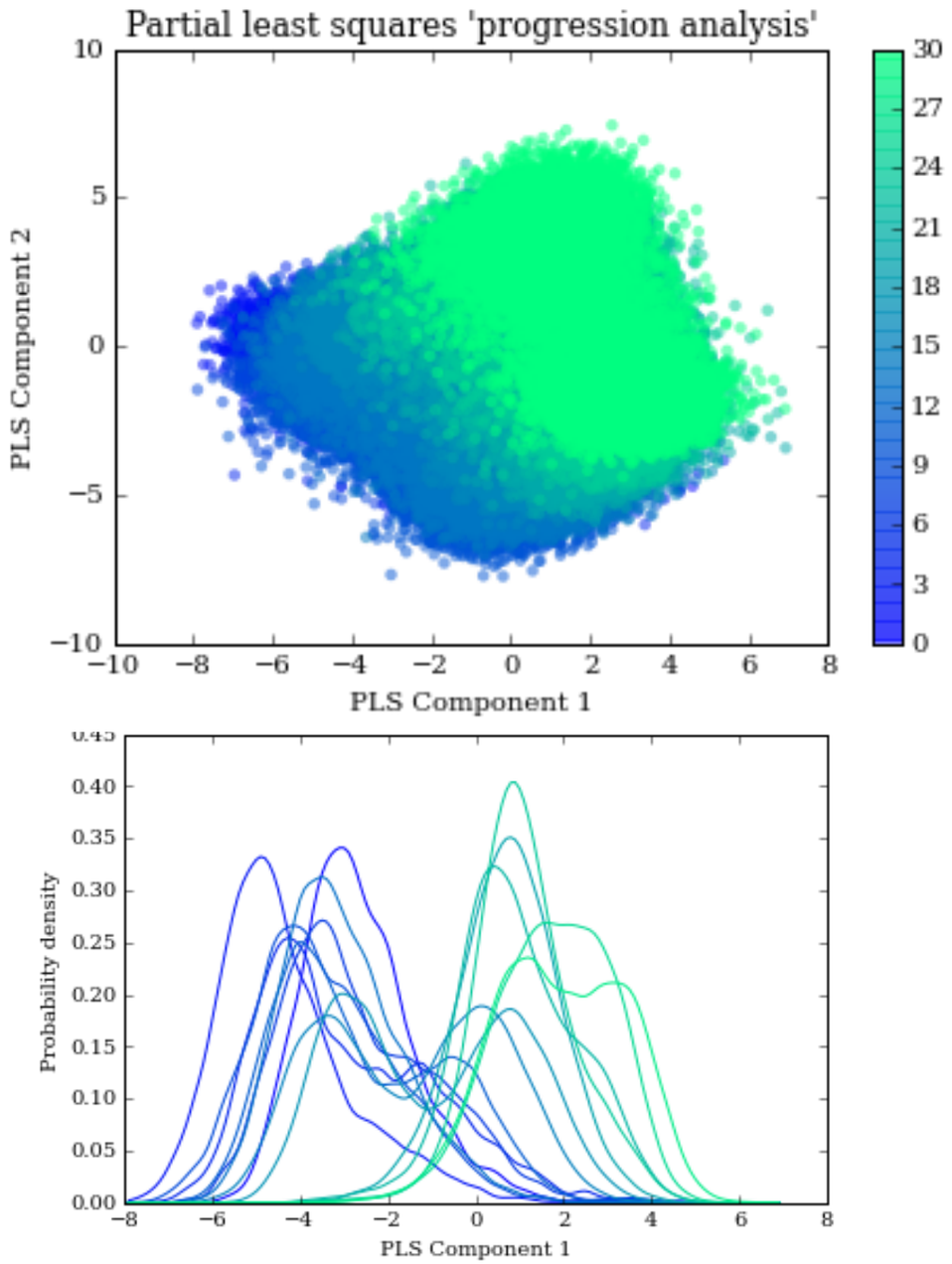
Prompted development of “FLOW-MAP”



- Visualization algorithm based on force-directed graph layout
- Very different conclusions than linear methods

Partial Least Squares “progression analysis”

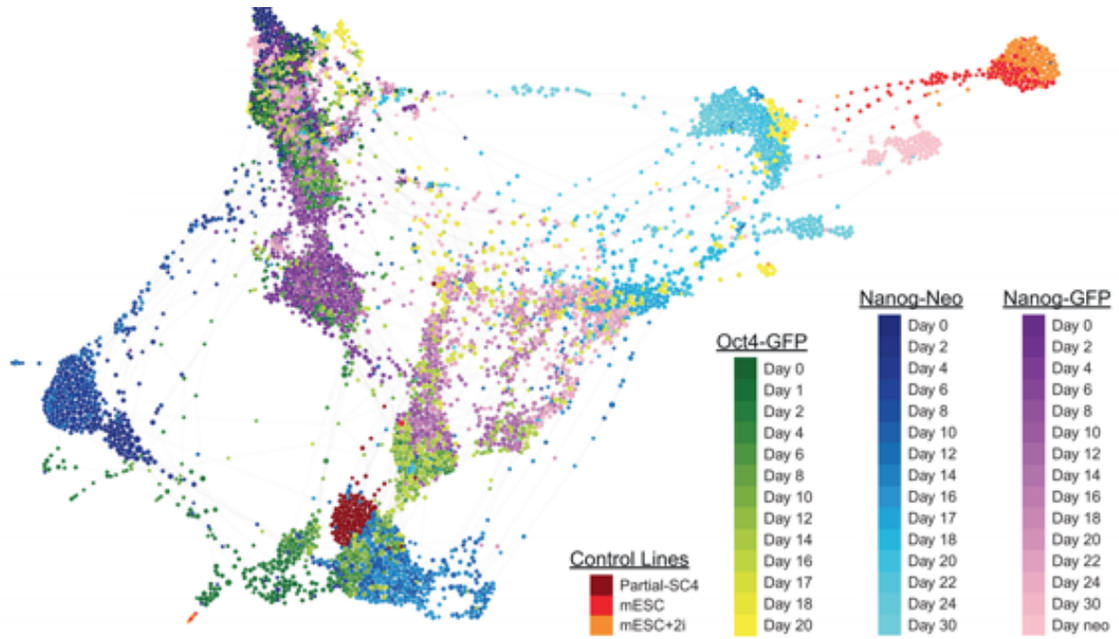
Goal: induce a linear transformation that optimally captures progression over time
Tactic: apply PLS, predict time-stamp from features
Results: qualitatively different from FLOW-MAP



Time

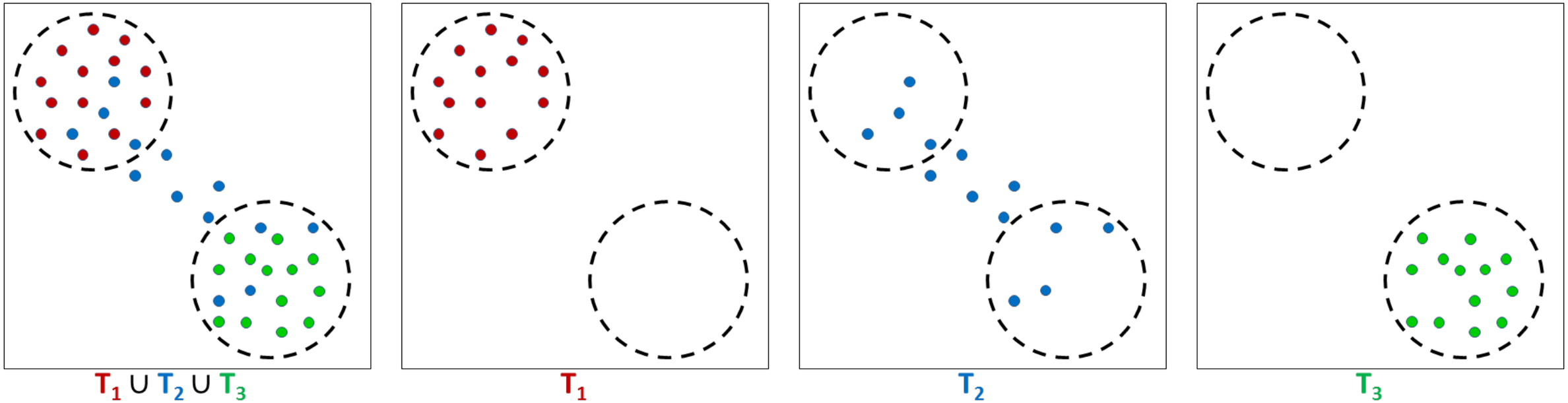
vs.

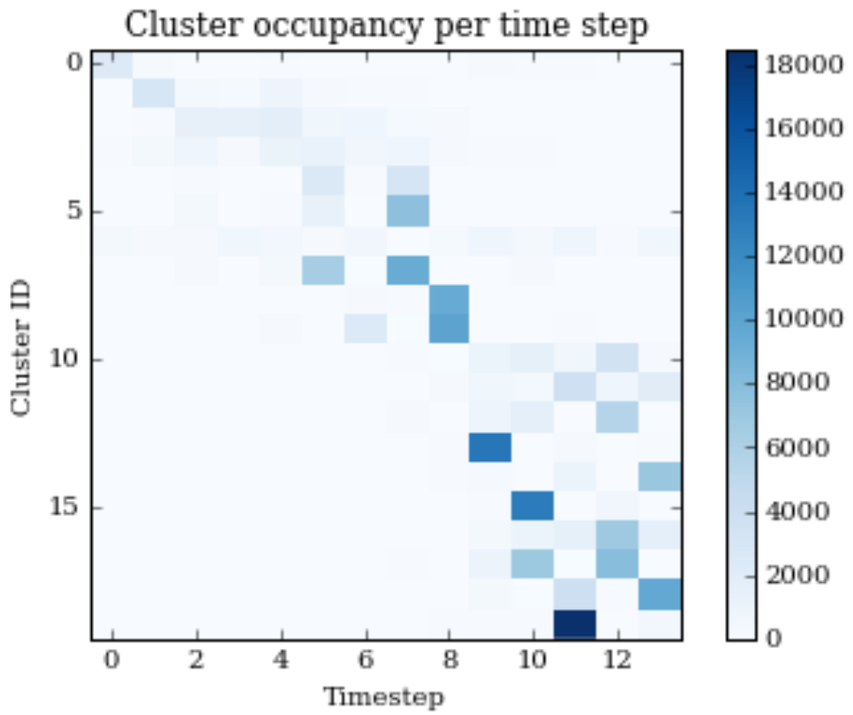
FLOW-MAP results:



Cluster-discretized “progression analysis”

Goal: quantify temporal change in population structure
Tactic: compute a clustering of the full dataset, observe how cluster occupancy changes over time





Conclusions and next steps

- Lots of cool new datasets from biology
    - plenty of room for (and interest in) better analysis methods
- Difficult to measure whether a visualization accurately depicts the data
- Significance: can influence how biologists / clinicians interpret their experiments