

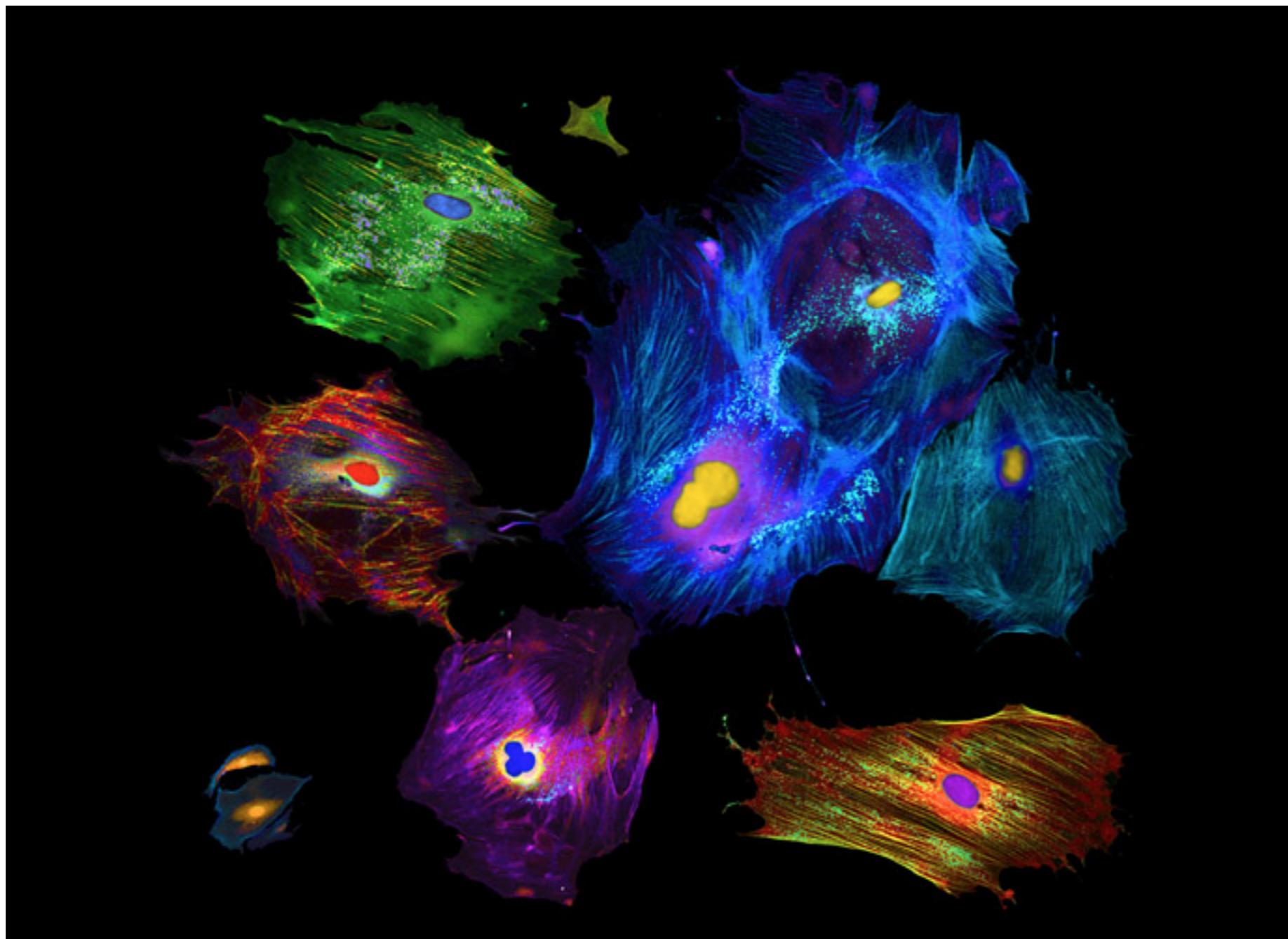
Analyzing single-cell data

Noah Apthorpe, Josh Fass, Maithra Raghu

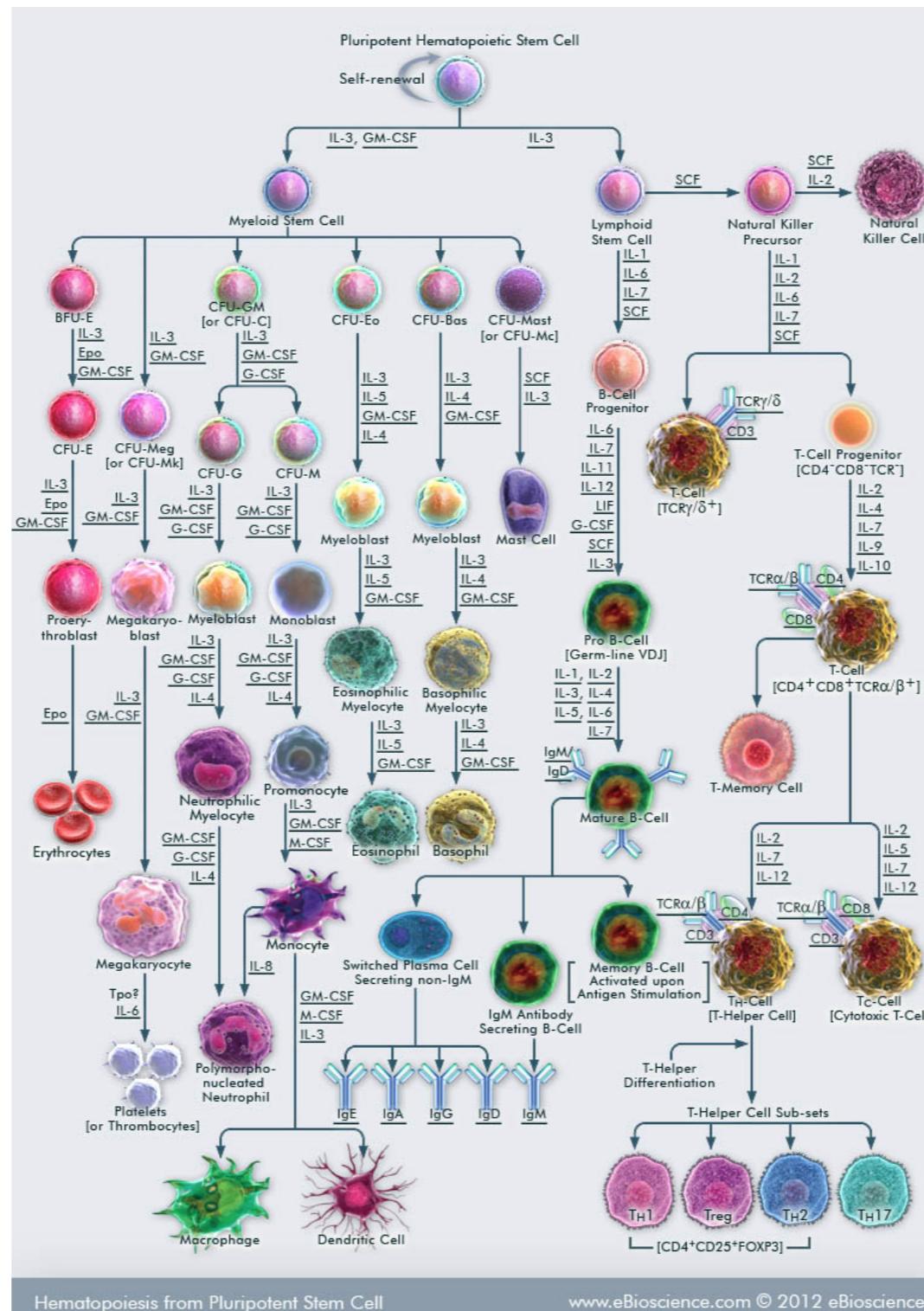
Outline

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

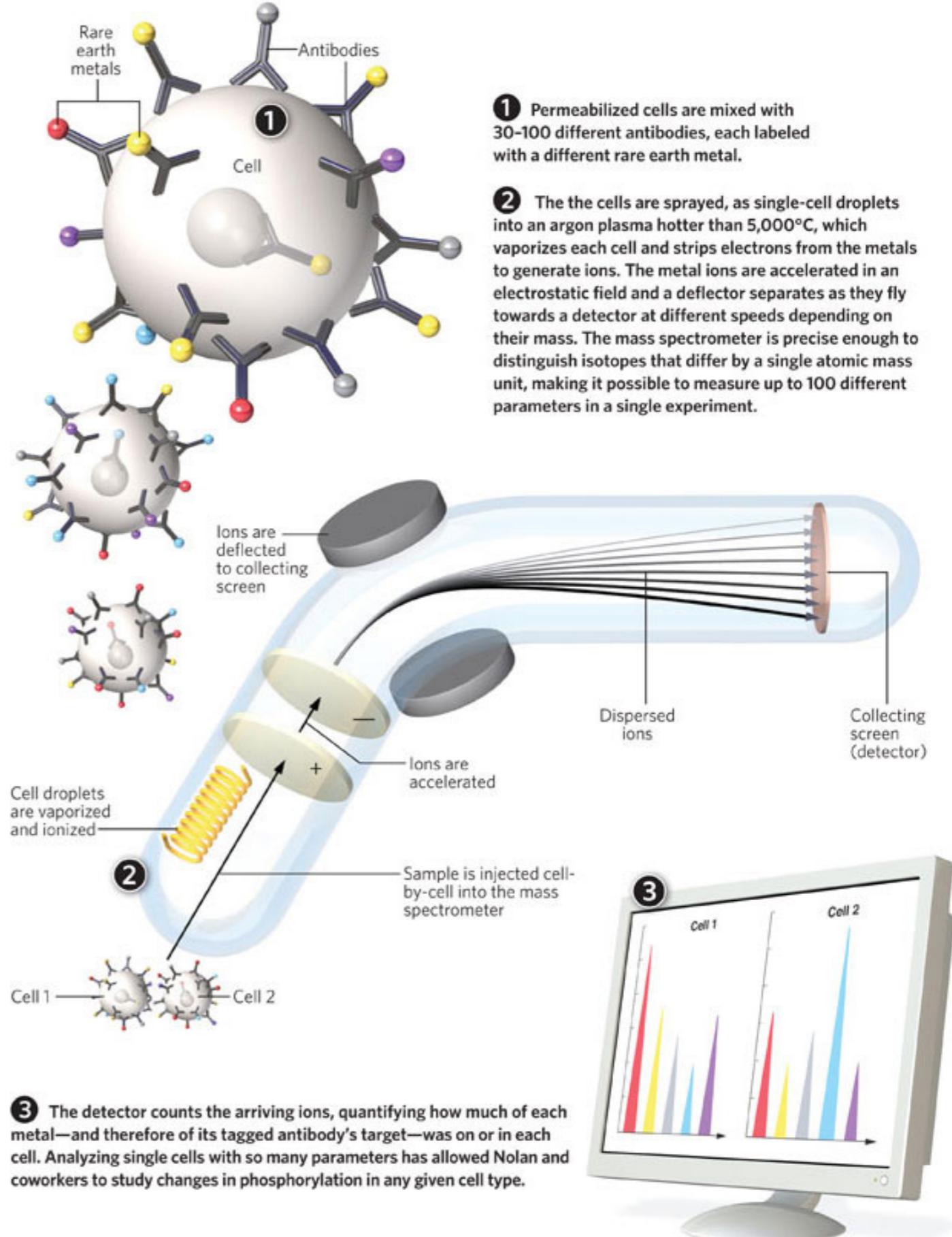
Every cell in the human body is different...



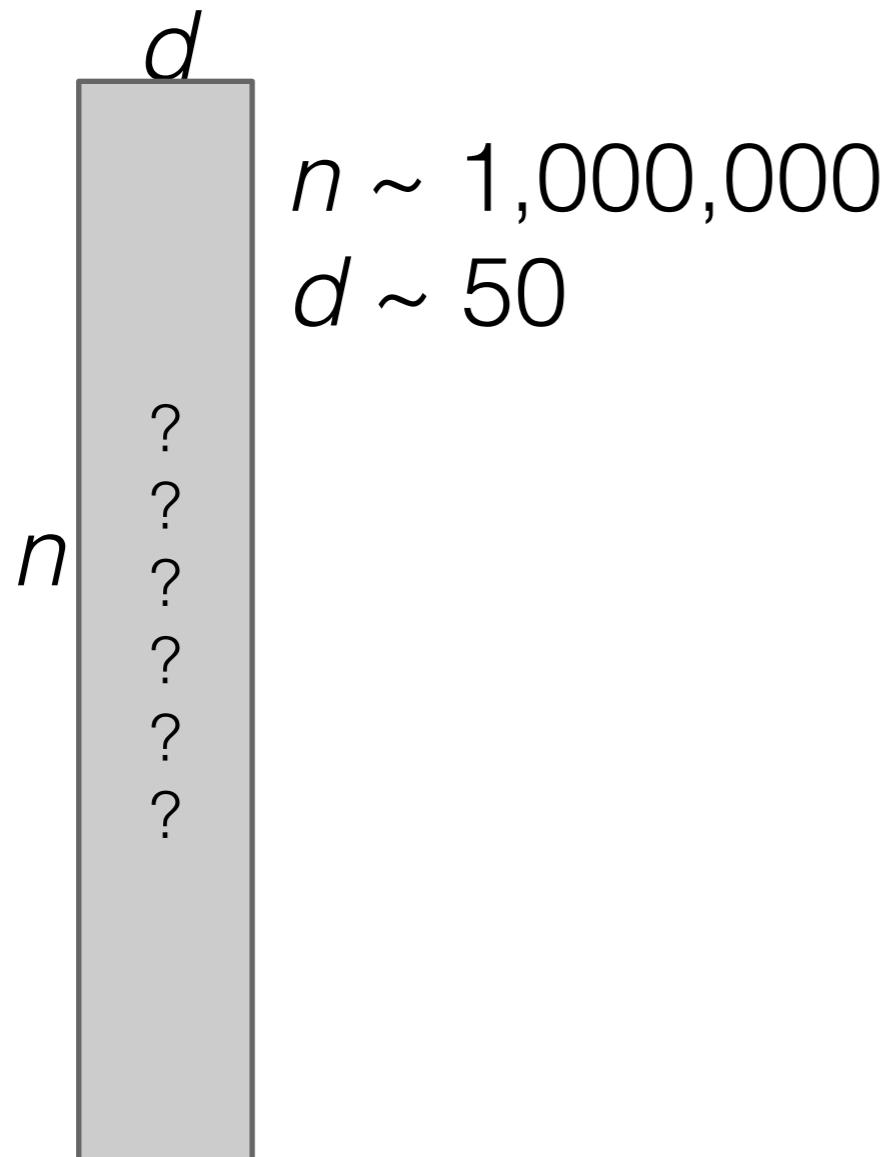
... due to repeated “specialization”



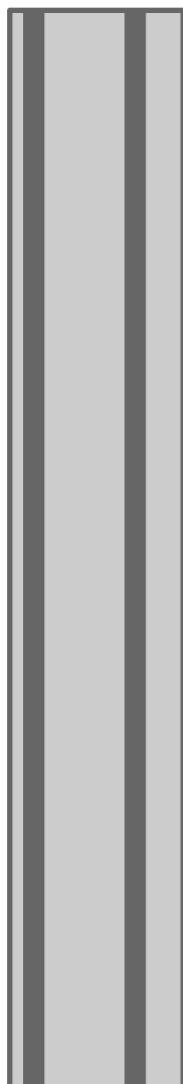
We can now
make high-
dimensional
single-cell
measurements...



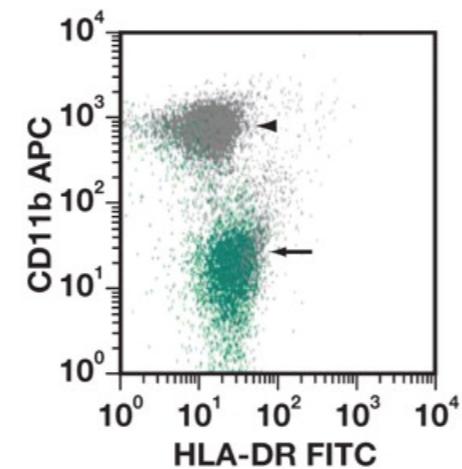
... resulting in large
single-cell datasets



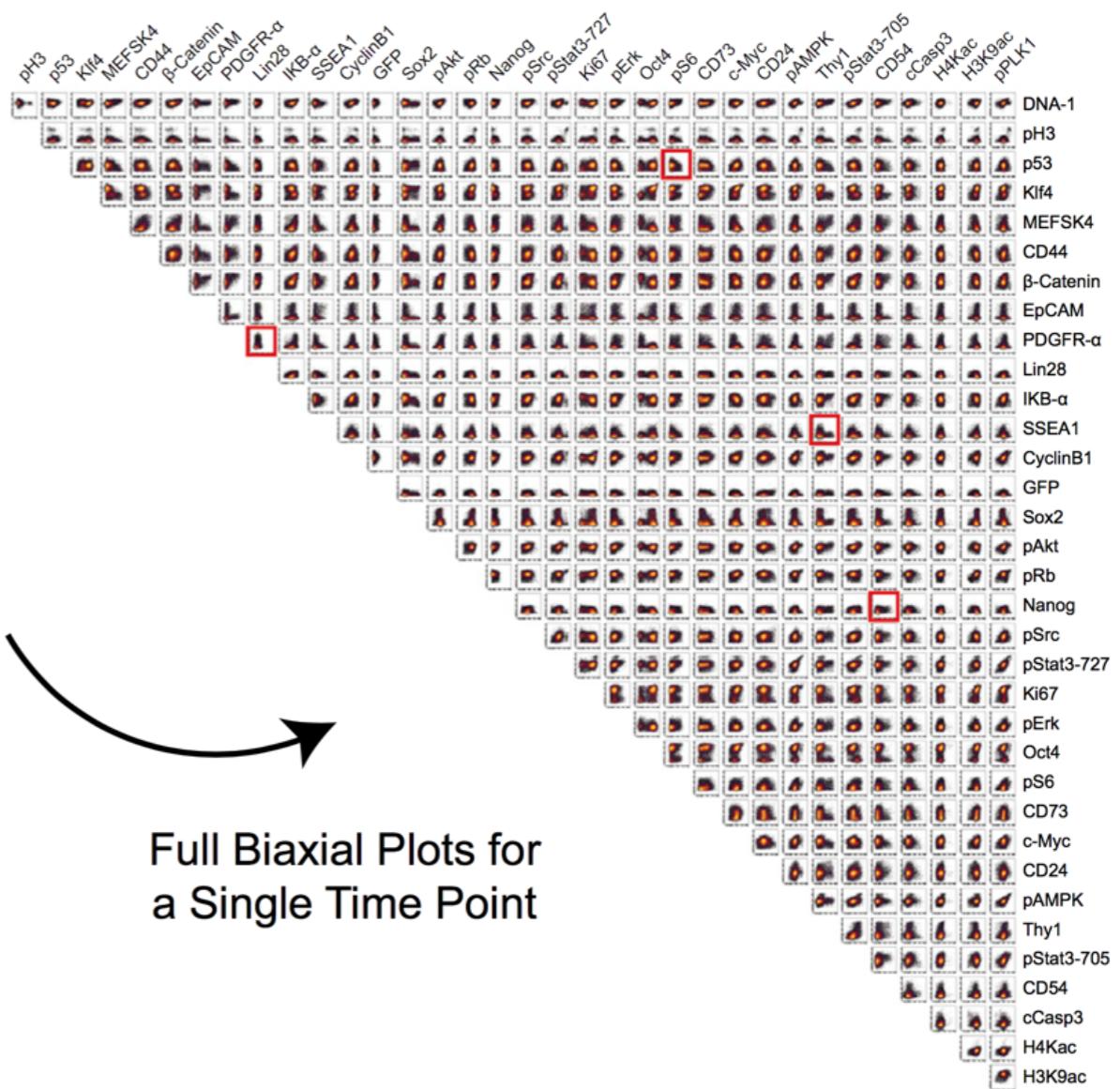
Traditionally analyzed by manual bivariate “gating”



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



Make
scatterplots of
2 columns at a
time



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

**nature
biotechnology**

ARTICLES

viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

El-ad David Amir¹, Kara L Davis^{2,3}, Michelle D Tadmor^{1,3}, Erin F Simonds^{2,3}, Jacob H Levine^{1,3}, Sean C Bendall^{1,3}, Daniel K Shenfeld^{1,3}, Smita Krishnaswamy¹, Garry P Nolan^{2,4} & Dana Pe'er^{1,4}

New high-dimensional, single-cell technologies offer unprecedented resolution in the analysis of heterogeneous tissues. However, because these technologies can measure dozens of parameters simultaneously in individual cells, data interpretation can be challenging. Here we present viSNE, a tool that allows one to map high-dimensional cytometry data onto two dimensions, yet conserves the high-dimensional structure of the data. viSNE plots individual cells in a visual similar to a scatter plot, while using all pairwise distances in high dimension to determine each cell's location in the plot. We integrated mass cytometry with viSNE to map healthy and cancerous bone marrow samples. Healthy bone marrow automatically maps into a consistent shape, whereas leukemia samples map into malformed shapes that are distinct from healthy bone marrow and from each other. We also use viSNE and mass cytometry to compare leukemia diagnosis and relapse samples, and to identify a rare leukemia population reminiscent of minimal residual disease. viSNE can be applied to any multi-dimensional single-cell technology.

Emerging single-cell technologies have revealed an extensive degree of heterogeneity between and within tissues¹. Analysis of single-cell data has shed light on many different cellular processes^{2–7} and recent technological advances have enabled the study of a large number of parameters in single cells at unparalleled resolution. For example, mass cytometry⁸ can measure up to 45 parameters simultaneously in tens of thousands of individual cells. High-resolution microscopy^{9,10} and single cell RNA quantification^{11–13} allow analysis of 100 parameters in dozens and soon hundreds of individual cells. These innovations promise to transform the way we think about development, differentiation, and disease^{1,14,15}.

However, it is difficult to visualize such high numbers of dimensions in a meaningful manner. Single cell data are often examined in two dimensions at a time in the form of a scatter plot¹⁶. Yet, as the number of parameters increases, the number of pairs becomes overwhelming. A typical mass cytometry data set allows several hundred pairwise combinations. In addition, a pairwise viewpoint could miss biologically meaningful multivariate relationships that cannot be discerned in two dimensions. Several computational tools, such as SPADE¹⁶, have been developed to address these problems^{16–18}. However, these approaches typically cluster cells and examine the association of each cluster with specific parameters or combinations of the data. Principal component analysis (PCA), another computational tool, has been applied to mass cytometry data sets¹⁹ and can be used to project data into two dimensions while maintaining single cell resolution. However, PCA is a linear transformation that cannot fully capture the nonlinear relationships that are a hallmark of many

RESULTS
Preserving high-dimensional relationships in single-cell data
In viSNE, each cell is represented as a point in high-dimensional space. Each dimension is a parameter (that is, the expression level of one protein). An optimization algorithm searches for a projection of the points from the high-dimensional space into two or three dimensions such that pairwise distances between the points are best conserved between the high- and low-dimensional space (Online Methods). The resulting low-dimensional projection, which we call the viSNE map,

© 2013 Nature America, Inc. All rights reserved.
New high-dimensional, single-cell technologies offer unprecedented resolution in the analysis of heterogeneous tissues. However, because these technologies can measure dozens of parameters simultaneously in individual cells, data interpretation can be challenging. Here we present viSNE, a tool that allows one to map high-dimensional cytometry data onto two dimensions, yet conserves the high-dimensional structure of the data. viSNE plots individual cells in a visual similar to a scatter plot, while using all pairwise distances in high dimension to determine each cell's location in the plot. We integrated mass cytometry with viSNE to map healthy and cancerous bone marrow samples. Healthy bone marrow automatically maps into a consistent shape, whereas leukemia samples map into malformed shapes that are distinct from healthy bone marrow and from each other. We also use viSNE and mass cytometry to compare leukemia diagnosis and relapse samples, and to identify a rare leukemia population reminiscent of minimal residual disease. viSNE can be applied to any multi-dimensional single-cell technology.

1Department of Biological Sciences, Columbia University, New York, New York, USA. 2Barker Laboratory – Stem Cell Biology Department of Biological Sciences, Columbia University, New York, New York, USA. 3These authors contributed equally to this work. 4These authors jointly directed this work. Correspondence should be addressed to D.P. (gnolan@biology.columbia.edu).

Received 18 December 2012; accepted 22 April 2013; published online 19 May 2013; doi:10.1038/nbt.2594

Cell Stem Cell
Resource

CellPress

A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry

Eli R. Zunder,^{1,2} Ernesto Lujan,^{1,2} Yury Goltsveit,¹ Marius Wernig,² and Gary P. Nolan,^{1,2*}

¹Department of Microbiology and Immunology, Baxter Laboratory for Stem Cell Biology
²Department of Pathology, Institute for Stem Cell Biology and Regenerative Medicine
³Department of Genetics
Stanford University School of Medicine, Stanford, CA 94305, USA
*Co-first author
Correspondence: gary.nolan@stanford.edu
http://dx.doi.org/10.1016/j.stem.2013.01.015

SUMMARY
To analyze cellular reprogramming at the single-cell level, mass cytometry was used to simultaneously measure markers of pluripotency, differentiation, cell-cycling status, and cellular signaling throughout the reprogramming process. Time-resolved progression analysis of the resulting data sets was used to construct a continuous molecular roadmap for three independent reprogramming systems. Although these systems varied substantially in Oct4, Sox2, Klf4, and c-Myc stoichiometry, they presented a common set of reprogramming landmarks. Early in the reprogramming process, Oct4^{high} Klf4^{high} cells transitioned to a CD73^{high} CD104^{high} CD54^{low} partially reprogrammed state. Kit67^{low} cells from this intermediate population reverted to a MEF-like phenotype, but Kit67^{high} cells advanced through the M-E-T and then bifurcated into two distinct populations: an ESC-like Nanog^{high} Sox2^{high} CD43^{high} population and a mesendoderm-like Nanog^{low} Sox2^{low} Lin28^{high} CD24^{high} PDGF β -R^{high} population. The methods developed here for time-resolved, single-cell progression analysis may be used for the study of additional complex and dynamic systems, such as cancer progression and embryonic development.

INTRODUCTION
Reprogramming somatic cells to a pluripotent state by forced expression of transcription factors is a dynamic process. How a somatic cell successfully undergoes this transition is poorly understood because low efficiencies, long latency times, and asynchronous progression impede molecular analysis (Hanna et al., 2009; Wernig et al., 2008). Characterization of bulk populations over time has given insight into how entire reprogramming populations progress (Li et al., 2010; Mikkelsen et al., 2008; Samavarchi-Tehrani et al., 2010; Soufi et al., 2012), but as most cells undergoing this process fail to reprogram, bulk analyses of

© 2011 Nature America, Inc. All rights reserved.
The ability to analyze multiple single-cell parameters is critical for understanding cellular heterogeneity. Despite recent advances in measurement technology, methods for analyzing multiple single-cell data are often subjective, labor intensive and require prior knowledge of the biological system. To objectively uncover cellular heterogeneity from single-cell measurements, we present a versatile computational approach, spanning-tree progression analysis of density-normalized events (SPADE). We applied SPADE to flow cytometry data of mouse bone marrow and to mass cytometry data of human bone marrow, in both cases, SPADE organized cells in a hierarchy of related phenotypes that partially recapitulated well-described patterns of hematopoiesis. We demonstrate that SPADE is robust to measurement noise and to the choice of cellular markers. SPADE facilitates the analysis of cellular heterogeneity, the identification of cell types and comparison of functional markers in response to perturbations. Some algorithms, such as a recent approach for automated gating termed SansSPECTRAL, have begun to include mechanisms for rare cell type identification.
Traditional cytometry data analysis methods also often cannot effectively accommodate and visualize the increasing numbers of measurements per single cell. For instance, to fully visualize an n-dimensional flow data set, $n(n-1)/2$ biaxial plots are needed, where each biaxial plot displays the correlation of only two measurements at a time. It is difficult to identify the correlations in high-dimensional data ($n \geq 3$) from a series of biaxial plots. One recent approach that partly addresses the scalability issue is the probability state model, implemented in the Genstat software package. That approach rearranges cells into a nonbranching linear order, according to an investigator's knowledge or expectation of how known markers fluctuate along a progression underlying the measured cell population (0 or more single cell parameters).
Despite the technological advances in acquiring an increasing number of parameters per single cell, methods for analyzing multidimensional single-cell data remain inadequate. Traditional methods are often subjective, labor intensive and require expert knowledge of the underlying cellular phenotypes. One common but cumbersome step is the selection of subsets of cells in a process

*Correspondence should be addressed to N.G. (gnolan@stanford.edu).
Received 15 January, accepted 31 August, published online 2 October 2011; doi:10.1038/nbt.2591

ANALYSIS

Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE

Peng Qiu^{1,2}, Erin F Simonds³, Sean C Bendall³, Kenneth D Gibbs Jr³, Robert V Bruggner³, Michael D Linderman⁴, Karen Sachs³, Garry P Nolan³ & Sylvia K Levitt¹

**The ability to analyze multiple single-cell parameters is critical for understanding cellular heterogeneity. Despite recent advances in measurement technology, methods for analyzing multiple single-cell data are often subjective, labor intensive and require prior knowledge of the biological system. To objectively uncover cellular heterogeneity from single-cell measurements, we present a versatile computational approach, spanning-tree progression analysis of density-normalized events (SPADE). We applied SPADE to flow cytometry data of mouse bone marrow and to mass cytometry data of human bone marrow, in both cases, SPADE organized cells in a hierarchy of related phenotypes that partially recapitulated well-described patterns of hematopoiesis. We demonstrate that SPADE is robust to measurement noise and to the choice of cellular markers. SPADE facilitates the analysis of cellular heterogeneity, the identification of cell types and comparison of functional markers in response to perturbations. Some algorithms, such as a recent approach for automated gating termed SansSPECTRAL, have begun to include mechanisms for rare cell type identification.
Traditional cytometry data analysis methods also often cannot effectively accommodate and visualize the increasing numbers of measurements per single cell. For instance, to fully visualize an n-dimensional flow data set, $n(n-1)/2$ biaxial plots are needed, where each biaxial plot displays the correlation of only two measurements at a time. It is difficult to identify the correlations in high-dimensional data ($n \geq 3$) from a series of biaxial plots. One recent approach that partly addresses the scalability issue is the probability state model, implemented in the Genstat software package. That approach rearranges cells into a nonbranching linear order, according to an investigator's knowledge or expectation of how known markers fluctuate along a progression underlying the measured cell population (0 or more single cell parameters).
Despite the technological advances in acquiring an increasing number of parameters per single cell, methods for analyzing multidimensional single-cell data remain inadequate. Traditional methods are often subjective, labor intensive and require expert knowledge of the underlying cellular phenotypes. One common but cumbersome step is the selection of subsets of cells in a process**

1Department of Haematology, Imperial College London, London, UK.
2Department of Bioinformatics and Computational Biology, University of Texas, M.D. Anderson Cancer Center, Houston, Texas, USA.
3Departments of Microbiology and Immunology, Stanford University, Stanford, California, USA.
4Computer Systems Laboratory, Stanford University, Stanford, California, USA.
Correspondence should be addressed to N.G. (gnolan@stanford.edu).
Received 15 January, accepted 31 August, published online 2 October 2011; doi:10.1038/nbt.2591

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

(i) Cytometry data

Density-dependent
down-sampling

(ii) Down-sampled data

Agglomerative
clustering

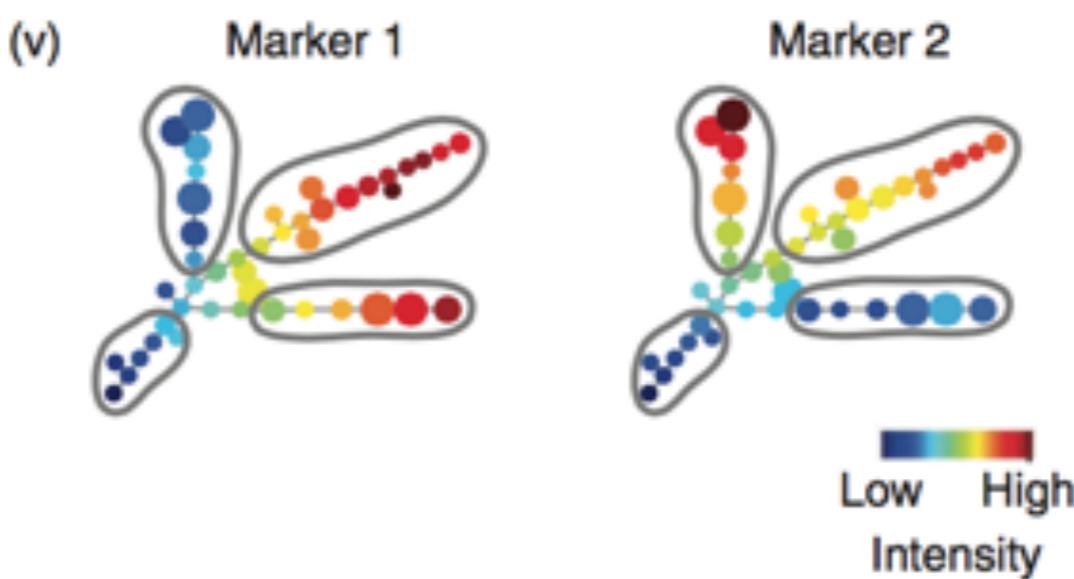
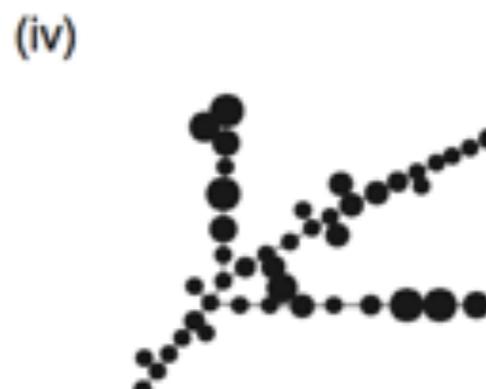
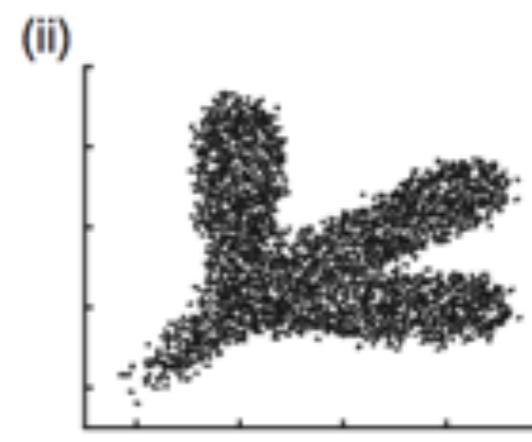
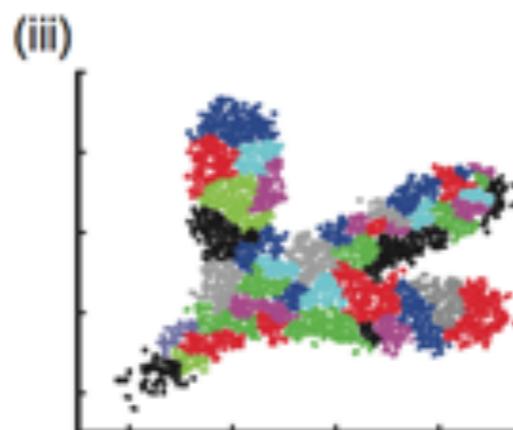
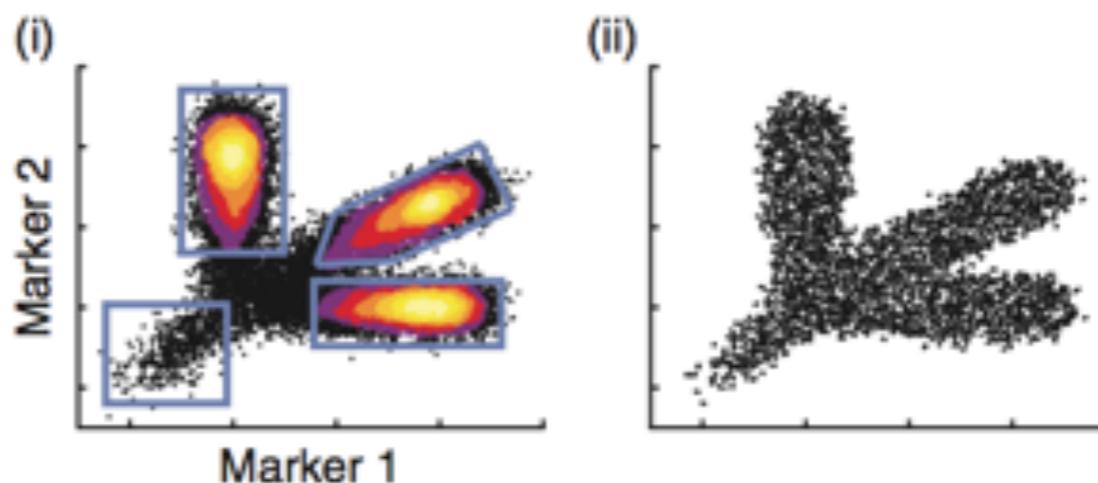
(iii) Clustering result

Minimum spanning
tree construction

(iv) SPADE tree

Up-sampling

(v) Colored tree showing
cellular heterogeneity



SPADE: Overview

Design goals

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

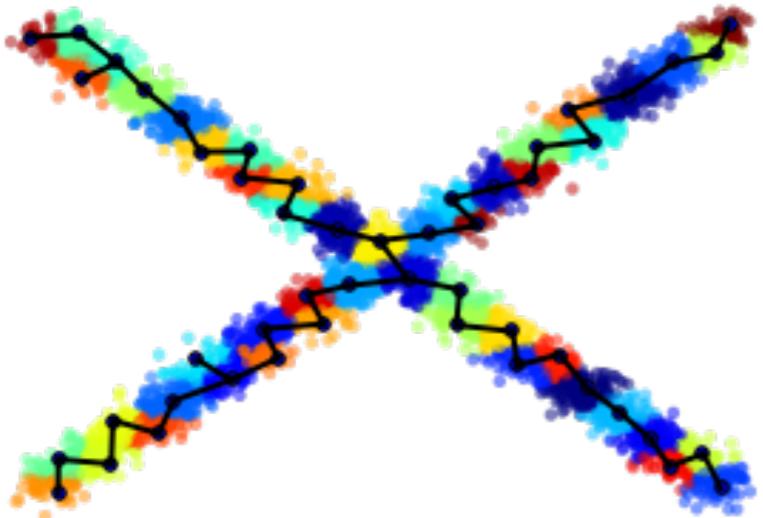
Limitations

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

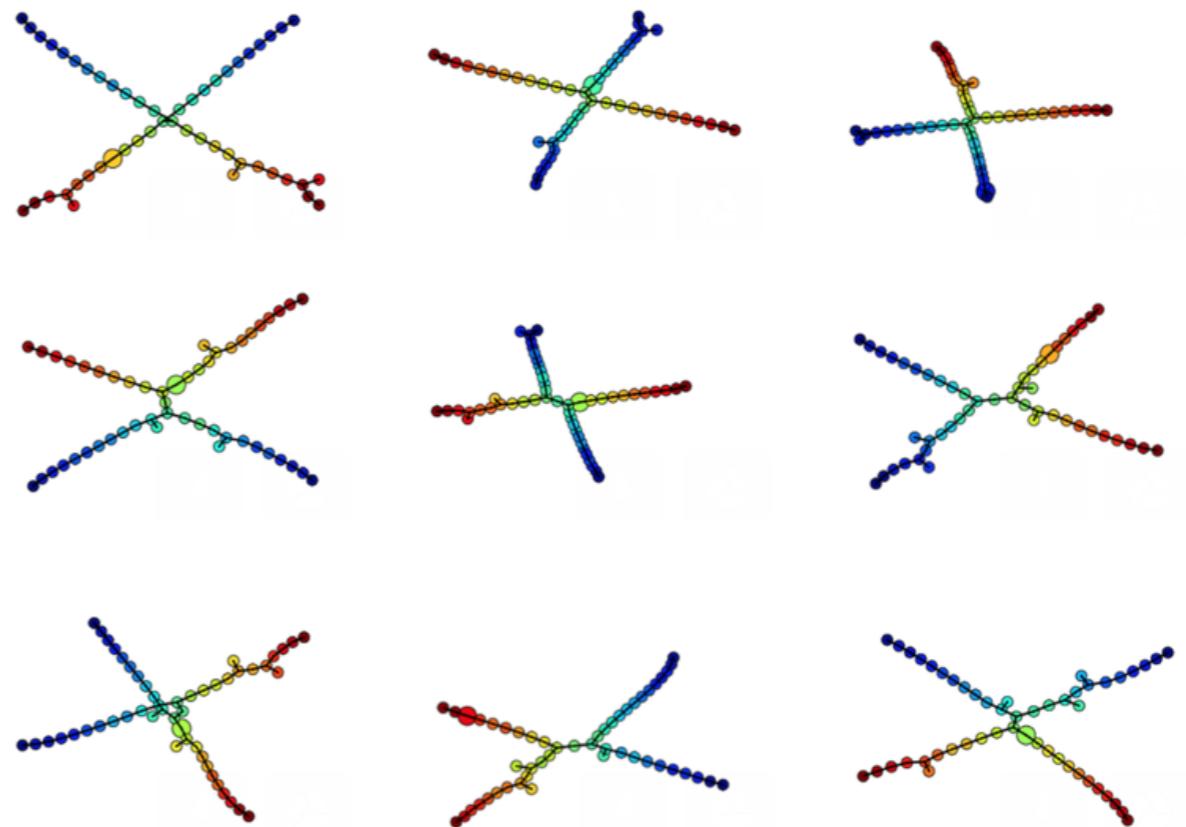
Our work

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

SPADE: Consistency experiments

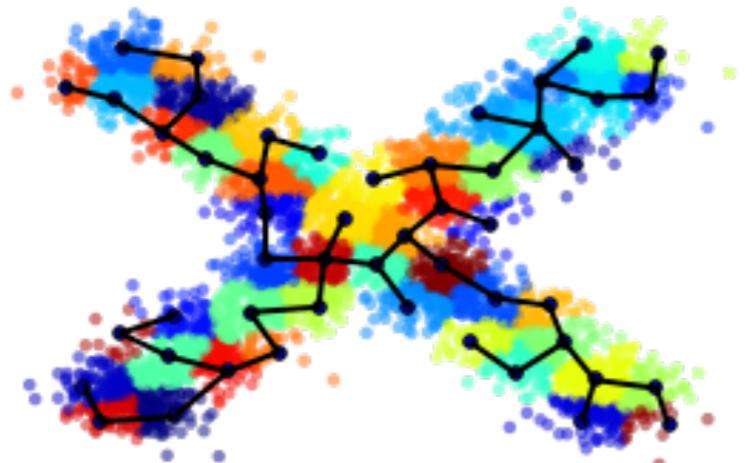


Synthetic data: long
“branches”

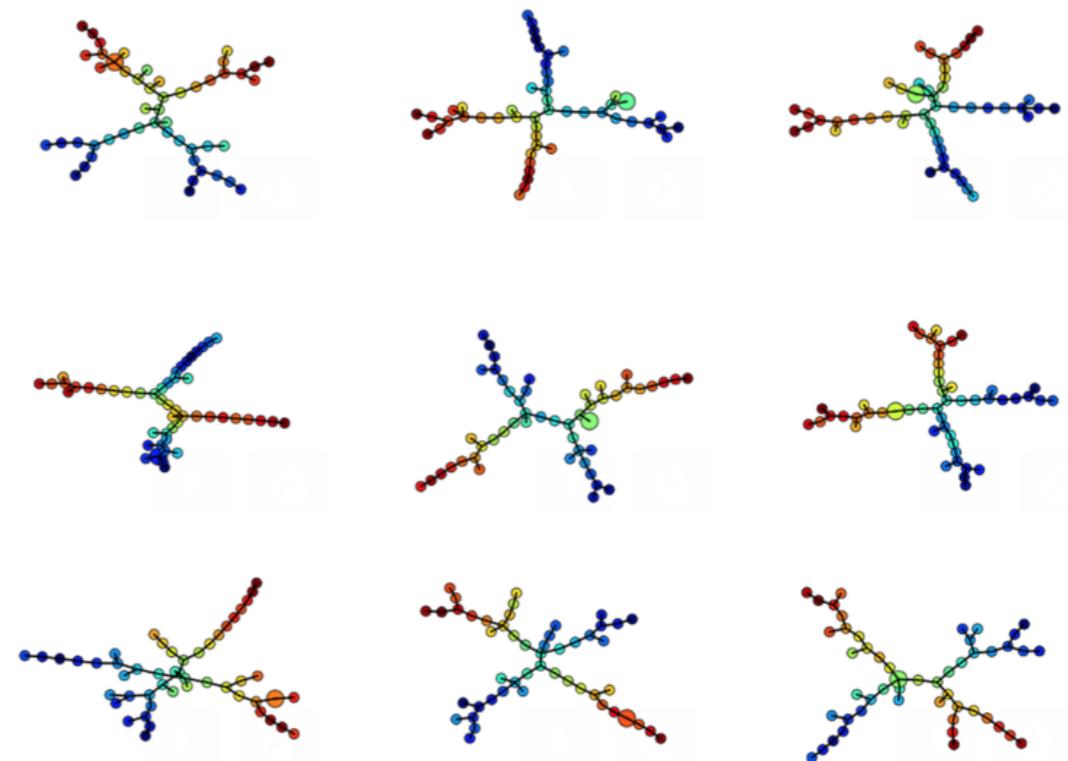


Outputs of repeated
trials

SPADE: Consistency experiments

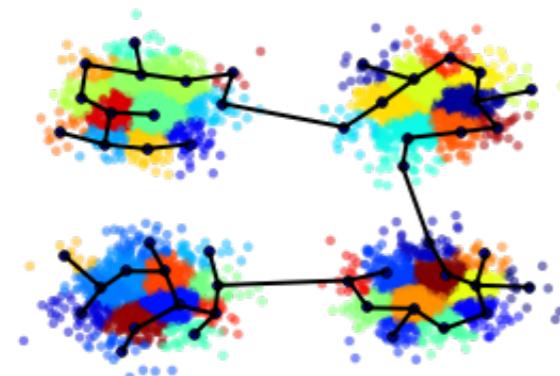


Synthetic data:
shorter “branches”

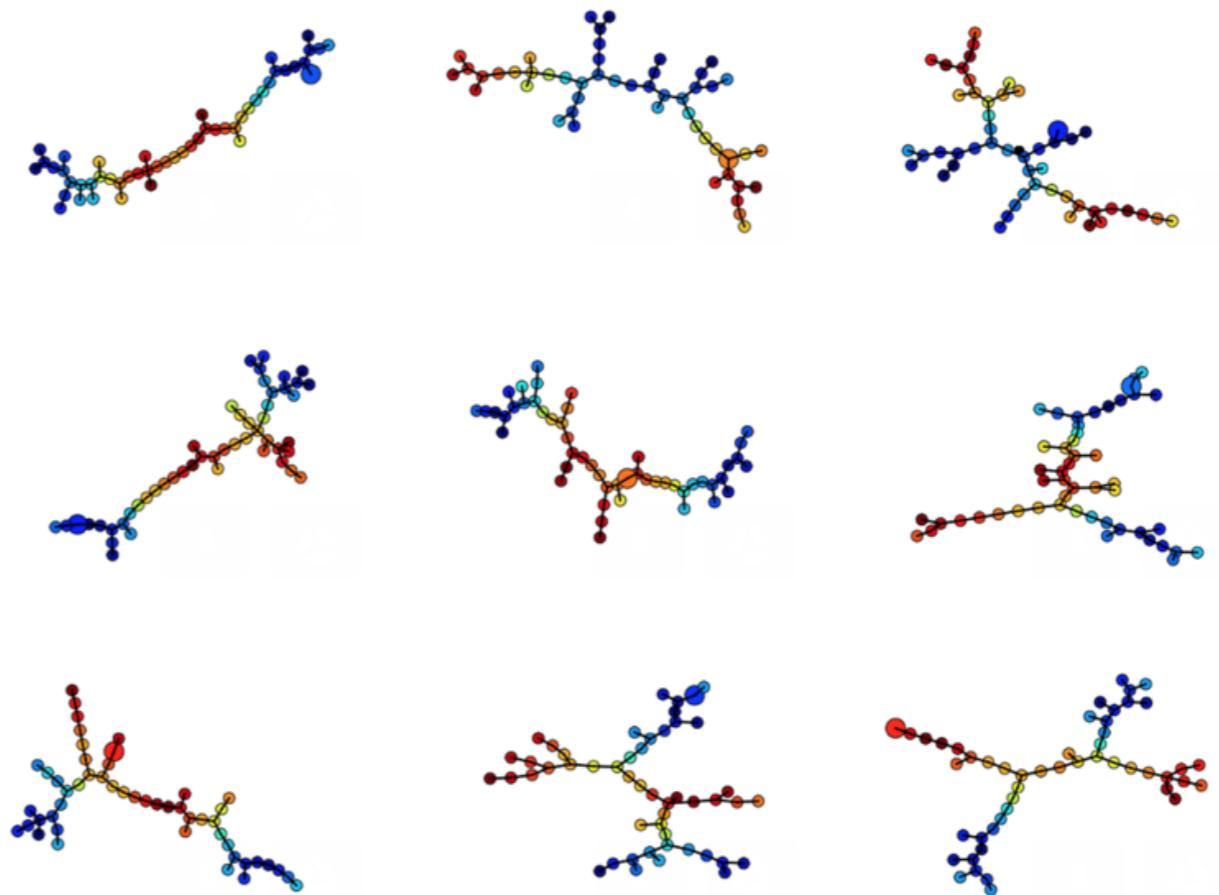


Outputs of repeated
trials

SPADE: Consistency experiments



Synthetic data:
mixture of Gaussians

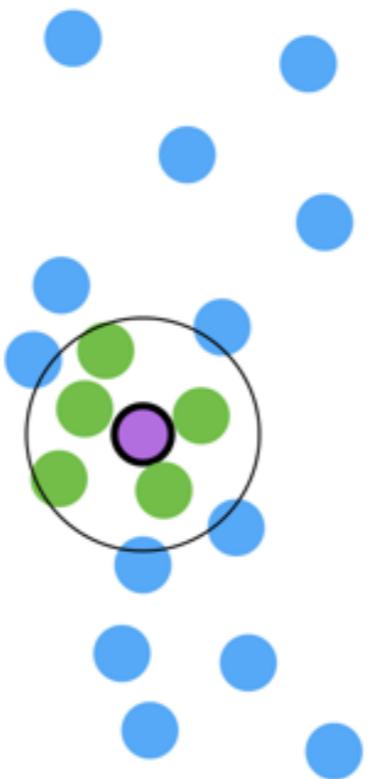


Outputs of repeated
trials

Our work

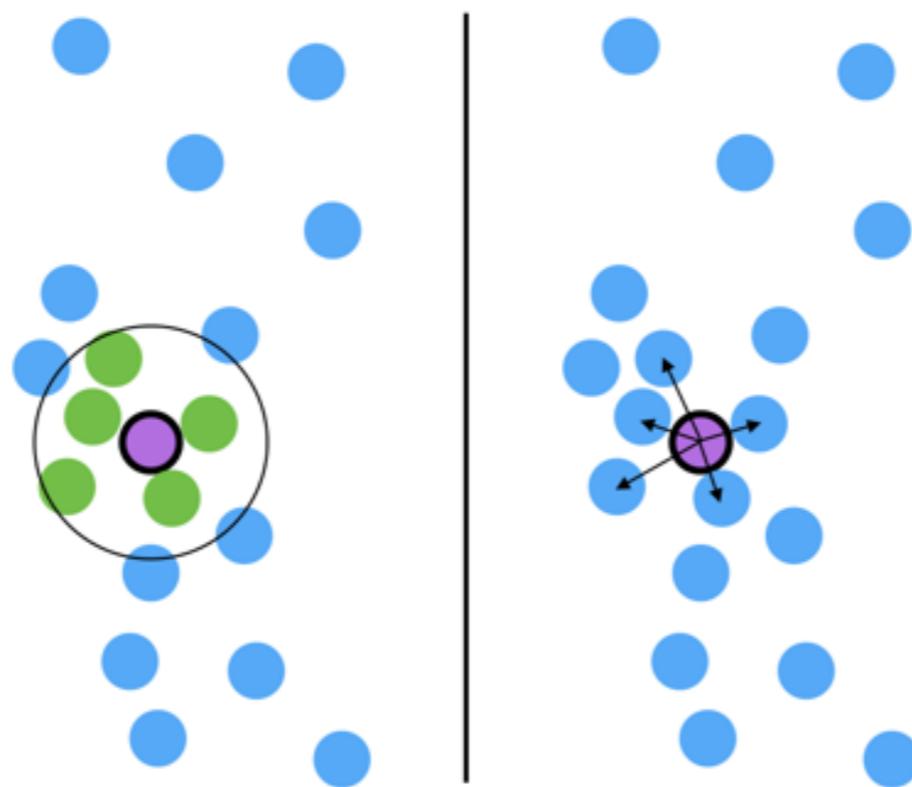
- SPADE: consistency experiments
- **SPADE: density estimation**
- Approximate “tree-preserving embedding”
- Is FLOW-MAP misleading?

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

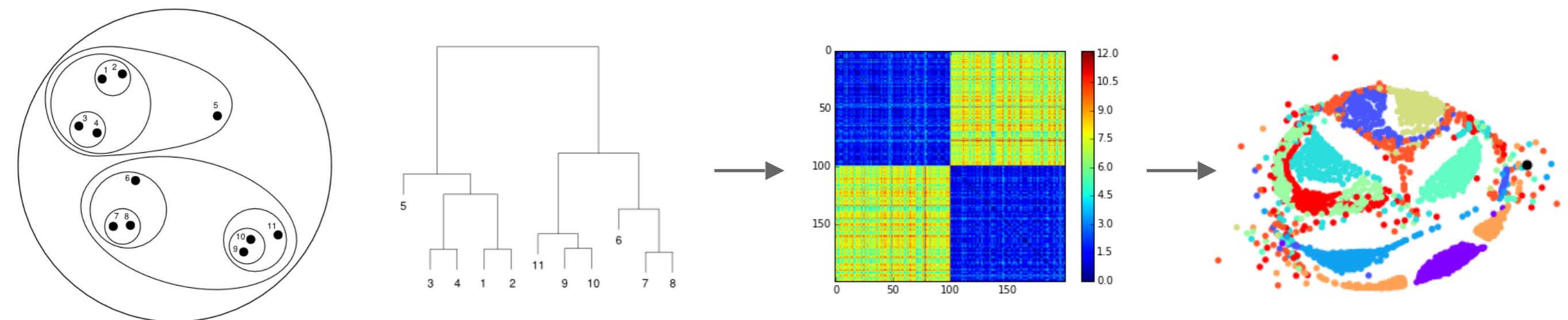
- SPADE: consistency experiments
- SPADE: density estimation
- **Approximate “tree-preserving embedding”**
- Is FLOW-MAP misleading?

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

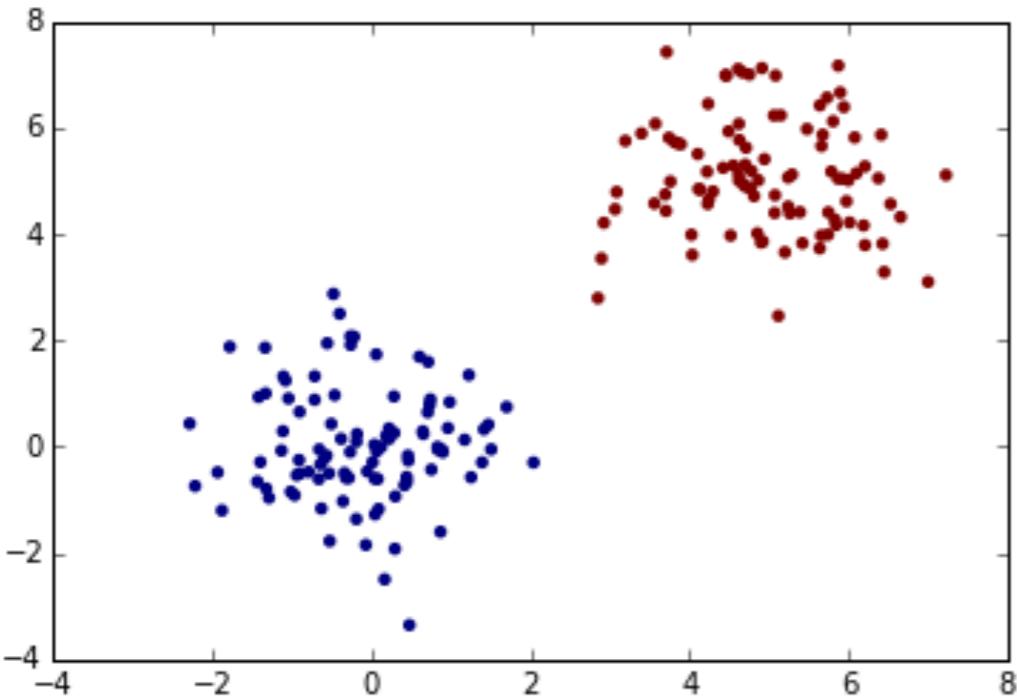


Hierarchical agglomerative clustering

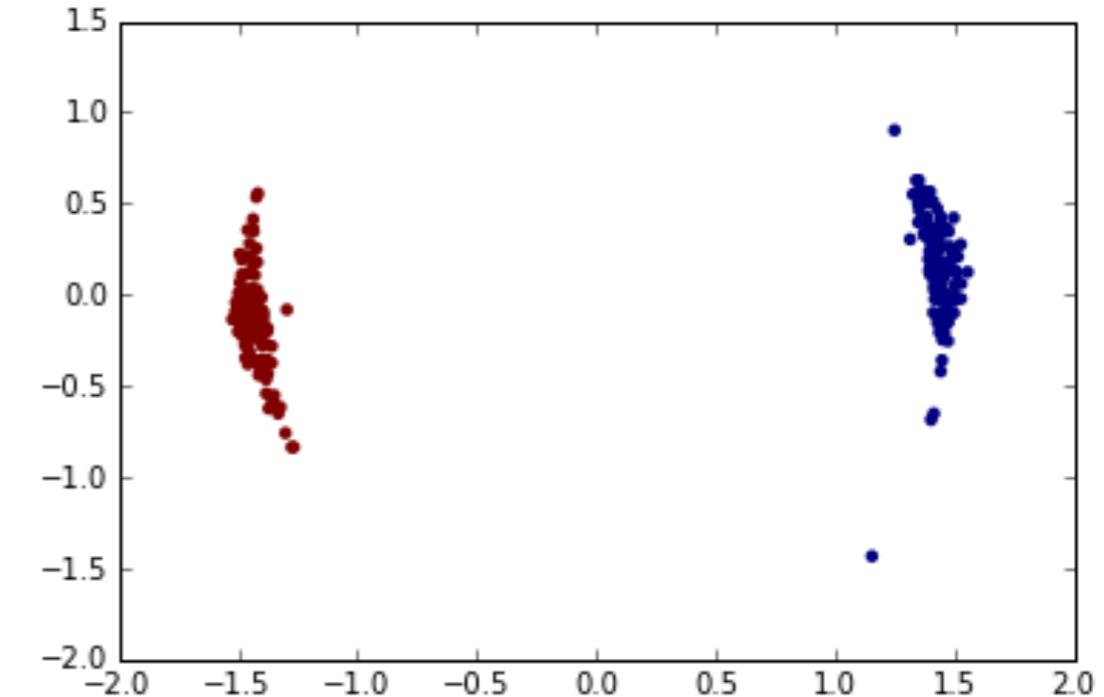
Cophenetic distance matrix

2D embedding

Approximate TPE: toy problem

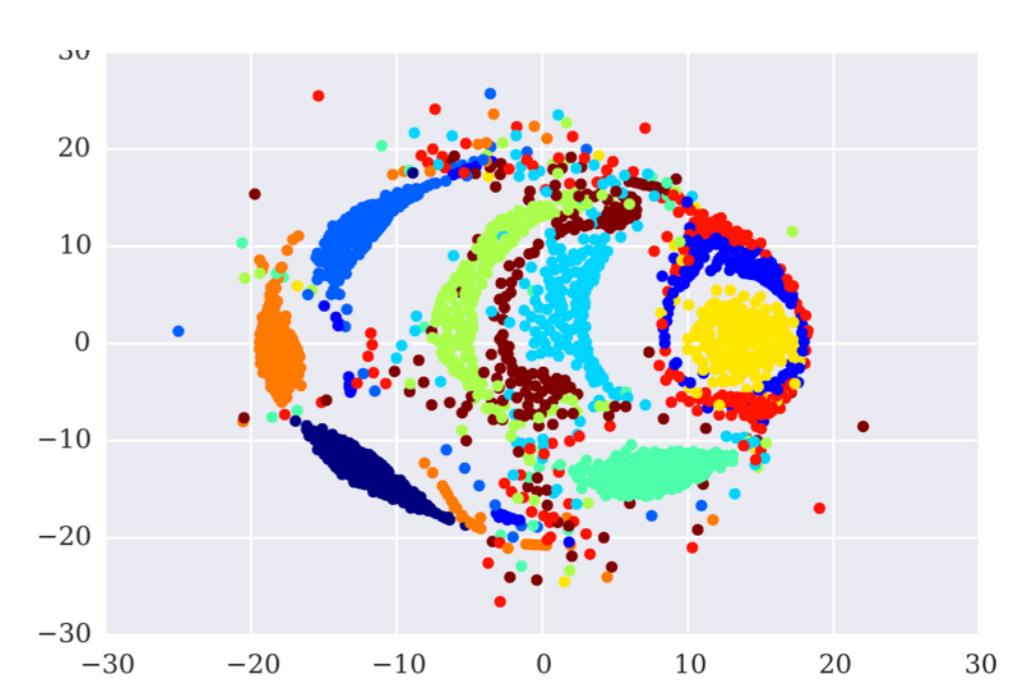
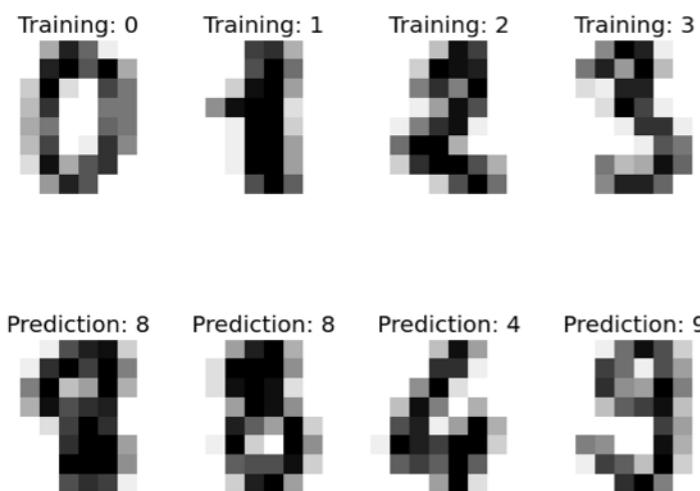


Synthetic data: 2D
mixture of Gaussians



Approximate TPE
solution

Approximate TPE: benchmark data

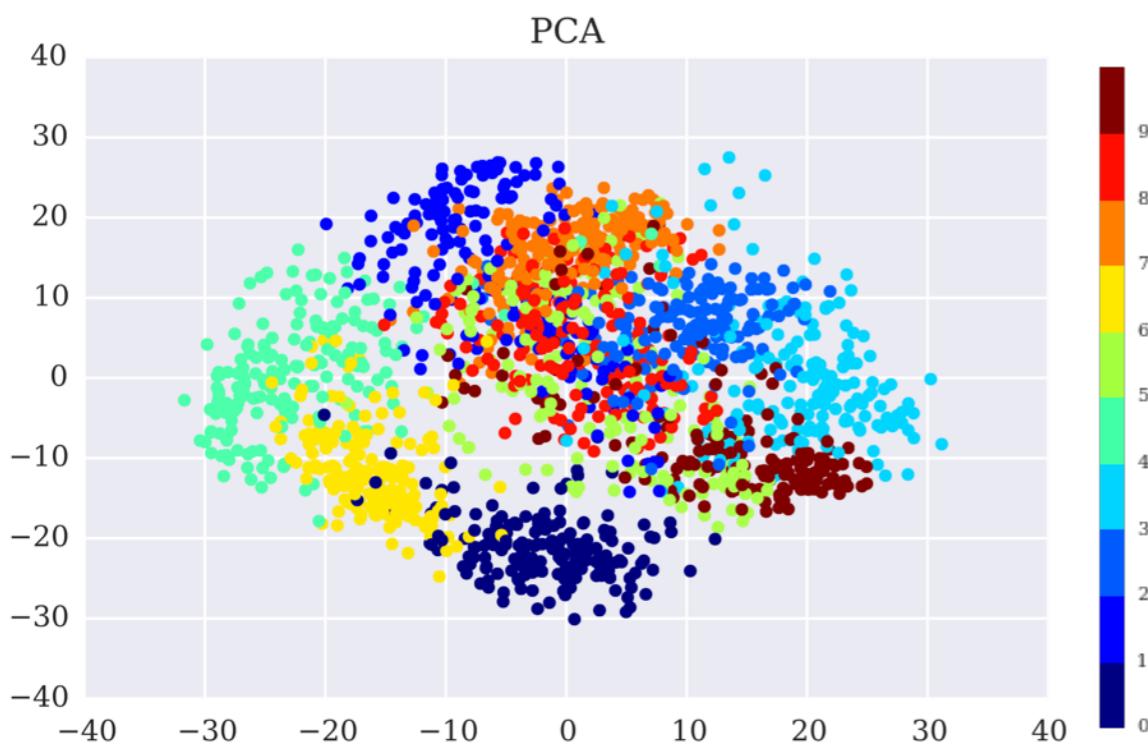


Example dataset:
handwritten digits (from
sklearn.datasets,
 $n=1797$, $d=64$)

Approximate TPE
solution

Comparing embedding quality

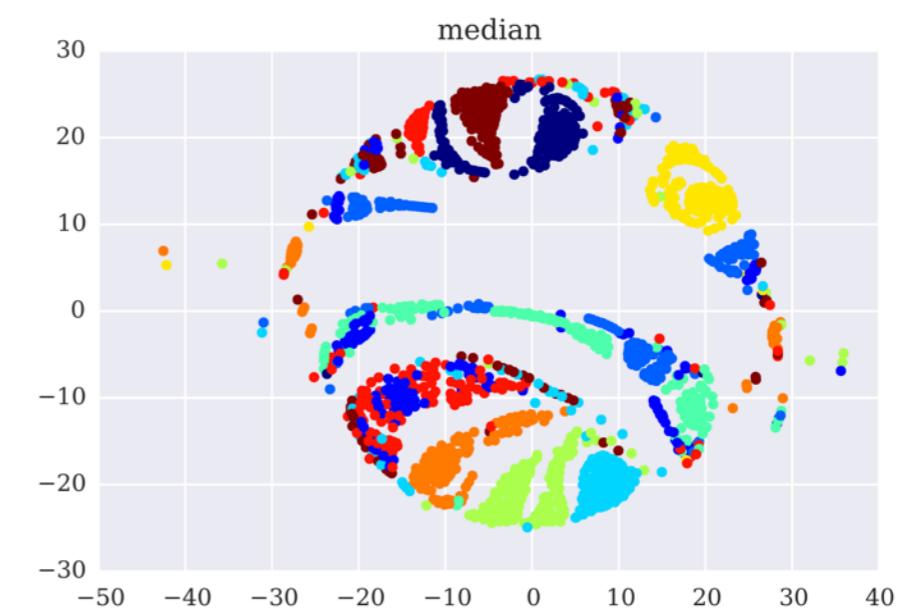
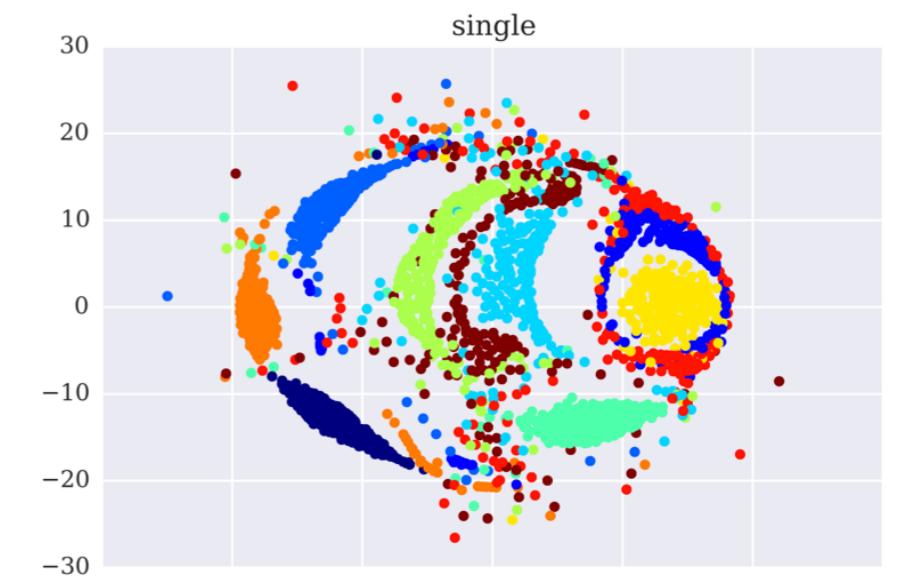
Baseline (PCA)



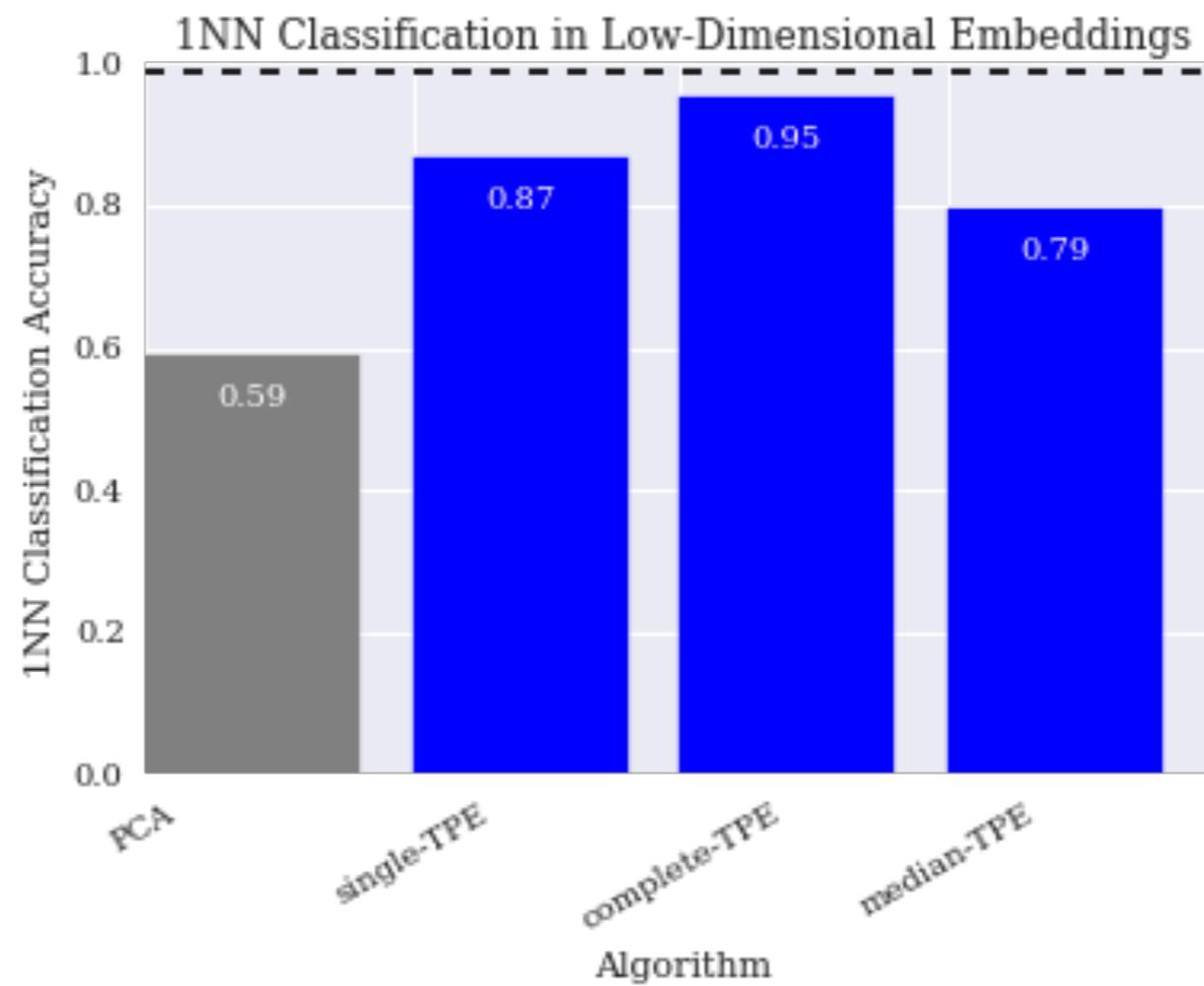
Approx. TPE with

- single-linkage
- median-linkage

VS.



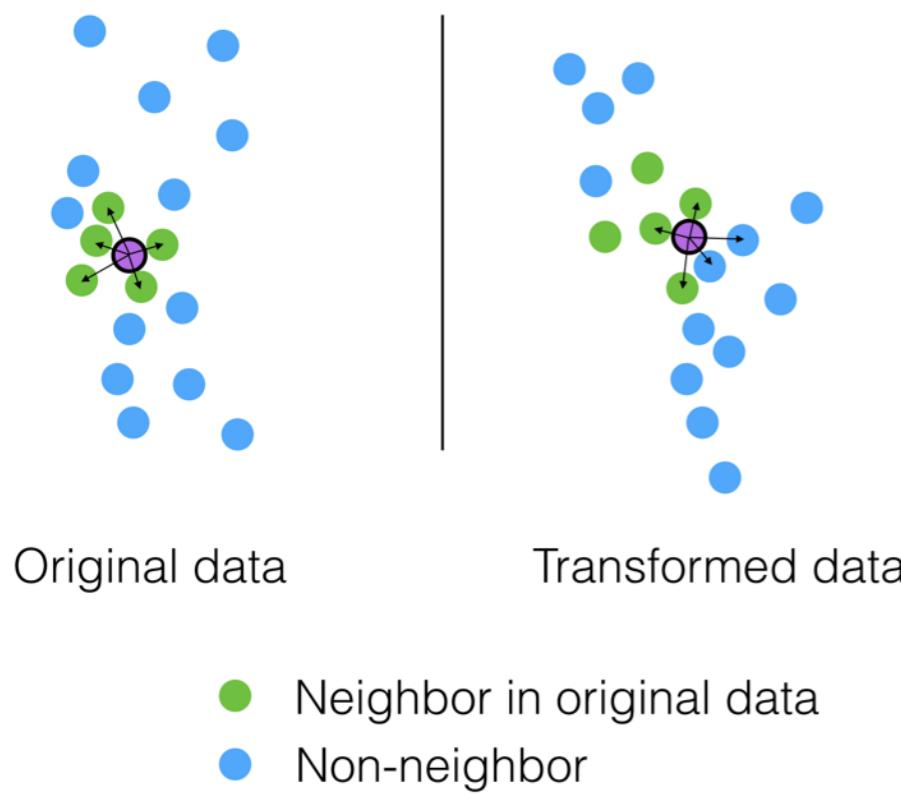
Measuring embedding fidelity: 1NN classification accuracy



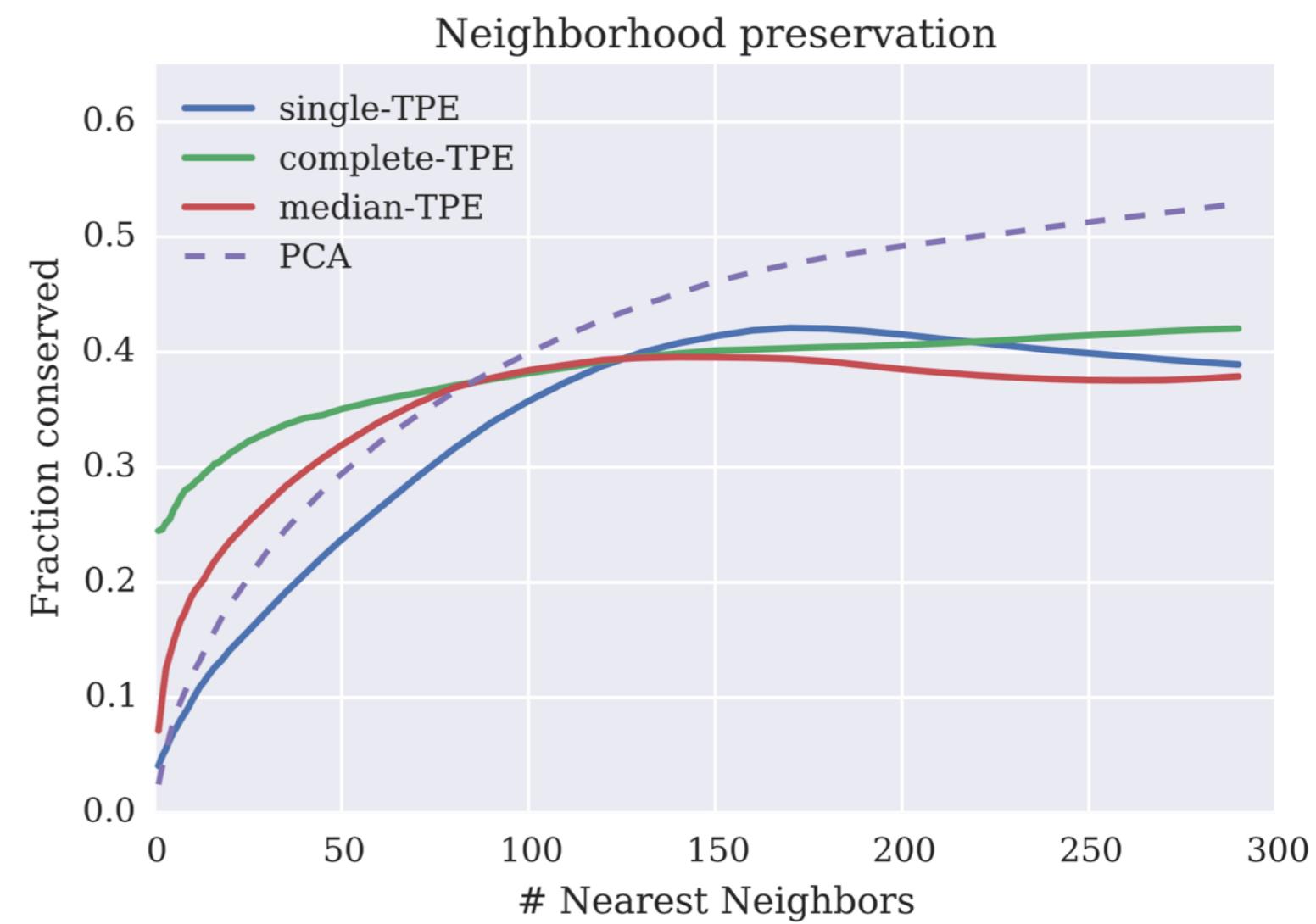
Measuring embedding fidelity: neighborhood preservation

Neighborhood preservation:

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



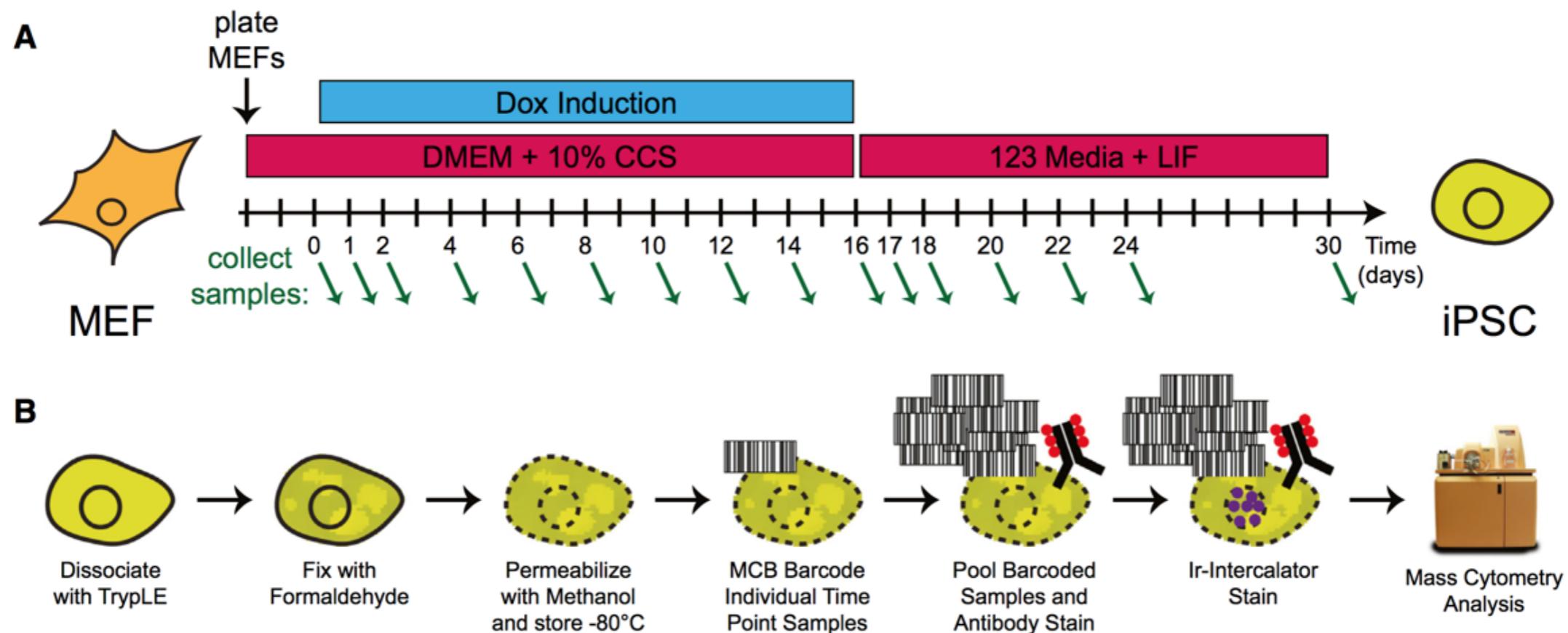
Results:



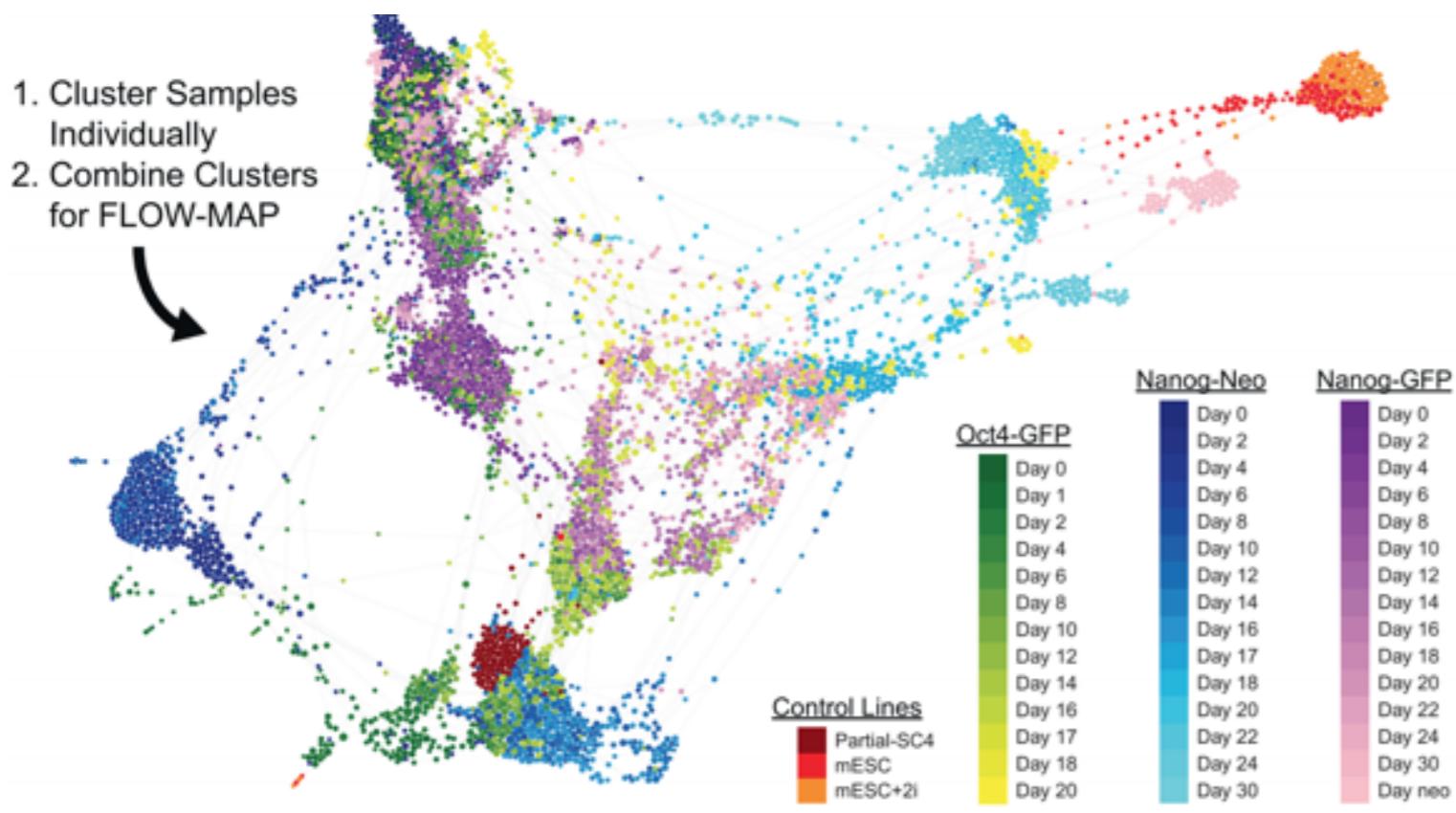
Our work

- 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”



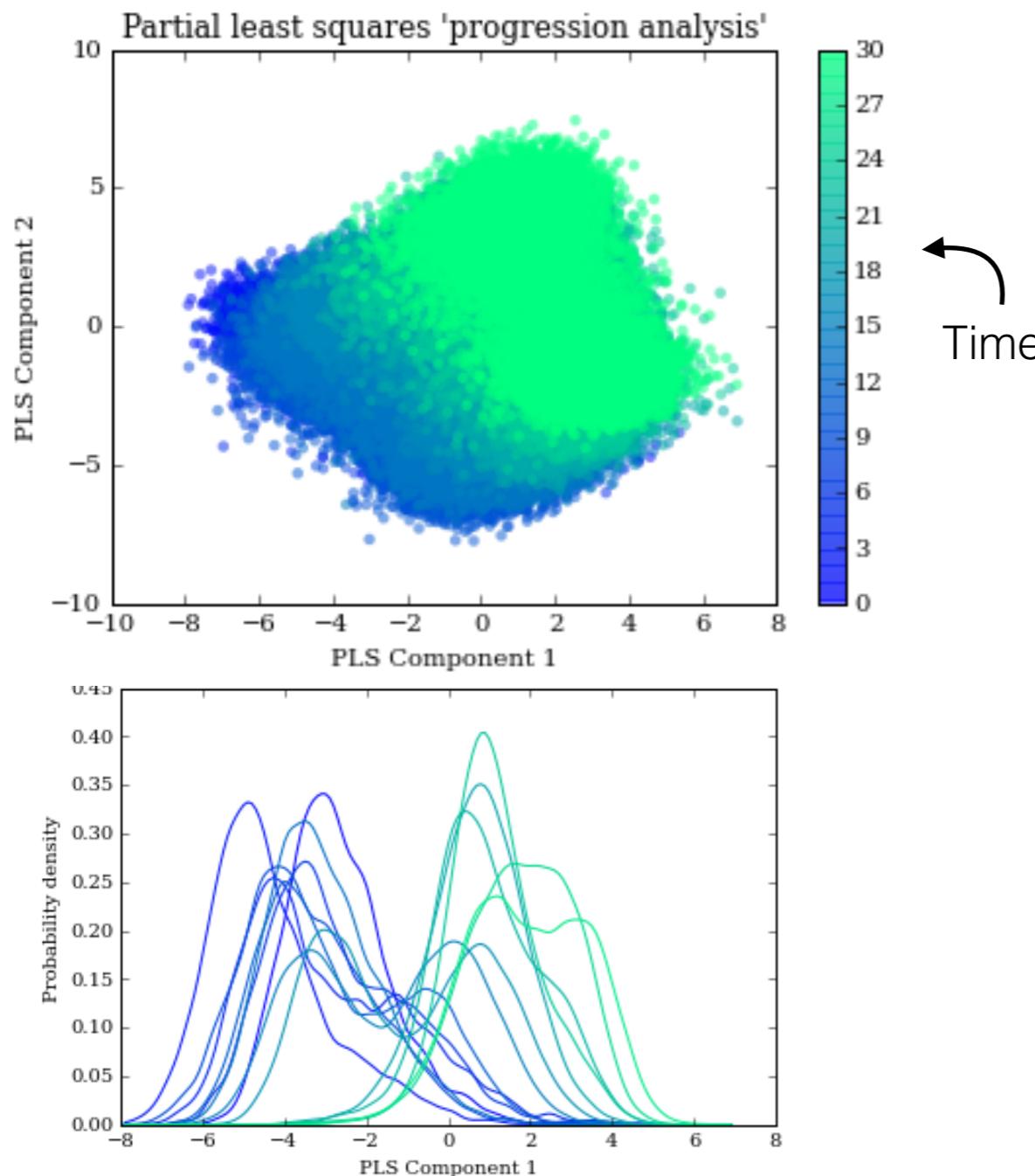
- Visualization algorithm based on force-directed graph layout
- Very different conclusions than linear methods

Partial Least Squares “progression analysis”

Goal: induce a linear transformation that optimally captures progression over time

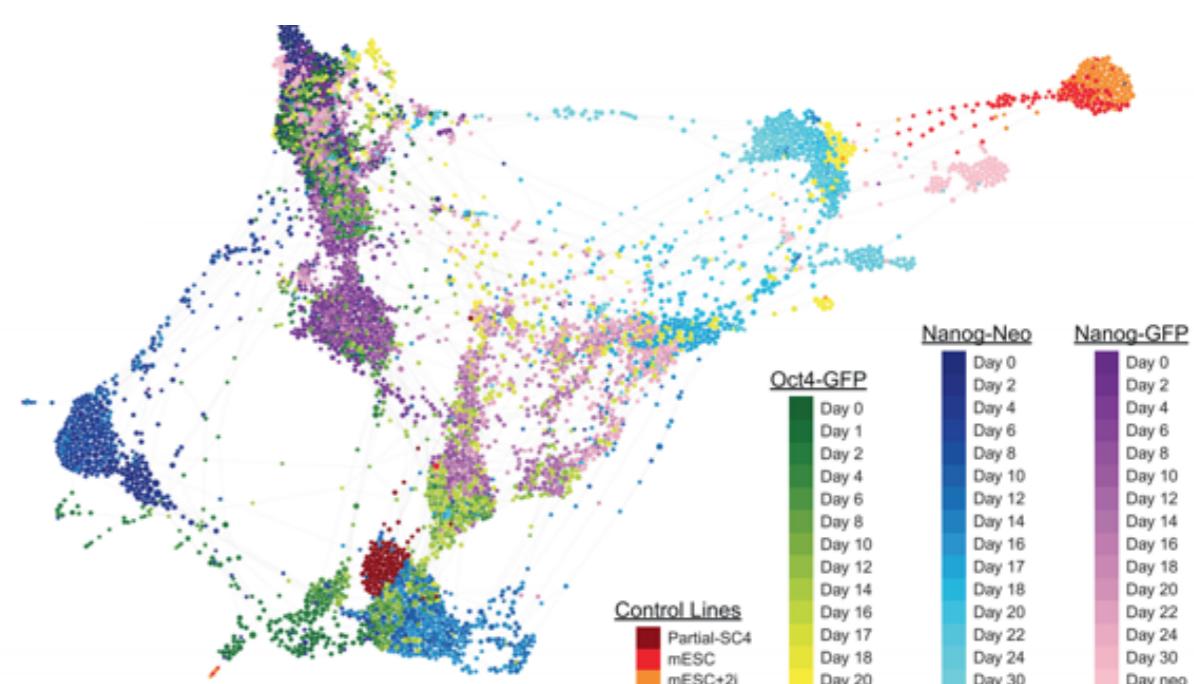
Tactic: apply PLS, predict time-stamp from features

Results: qualitatively different from FLOW-MAP



FLOW-MAP results:

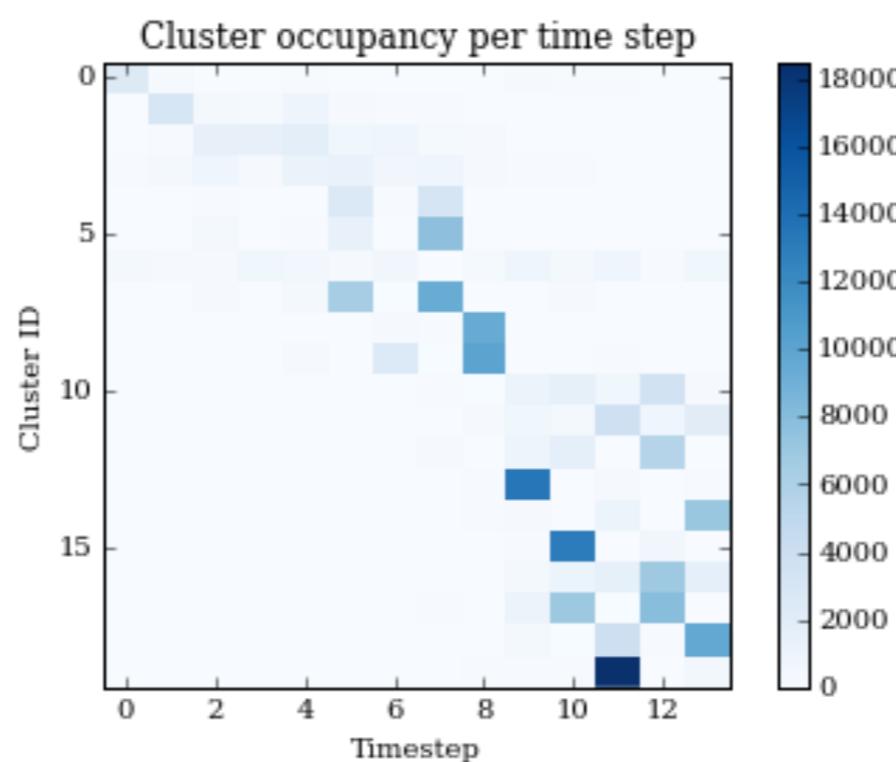
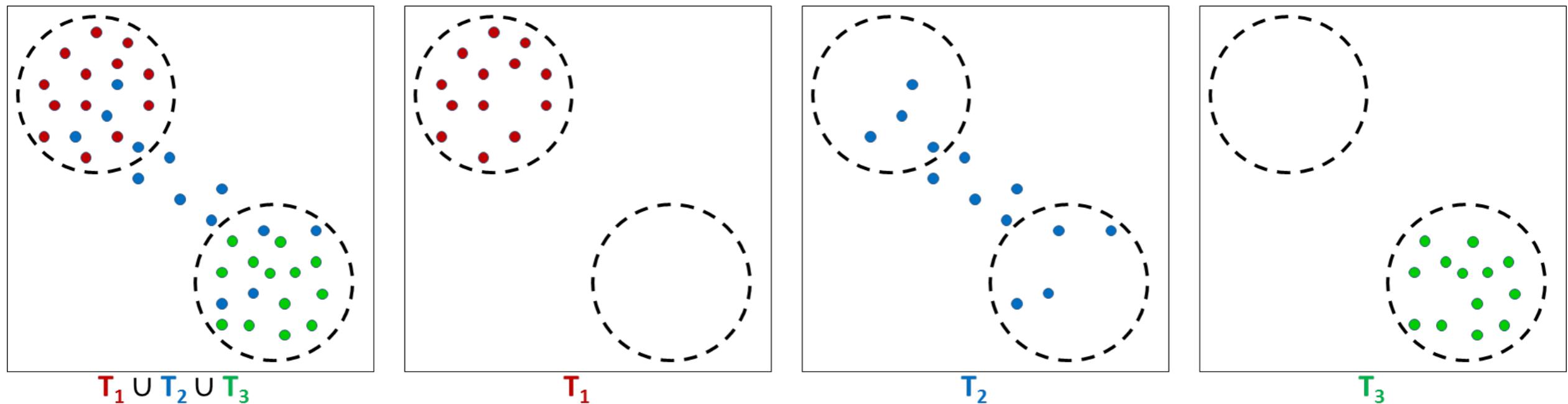
vs.



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