“A single cell approach to interrogating network rewiring in EMT”

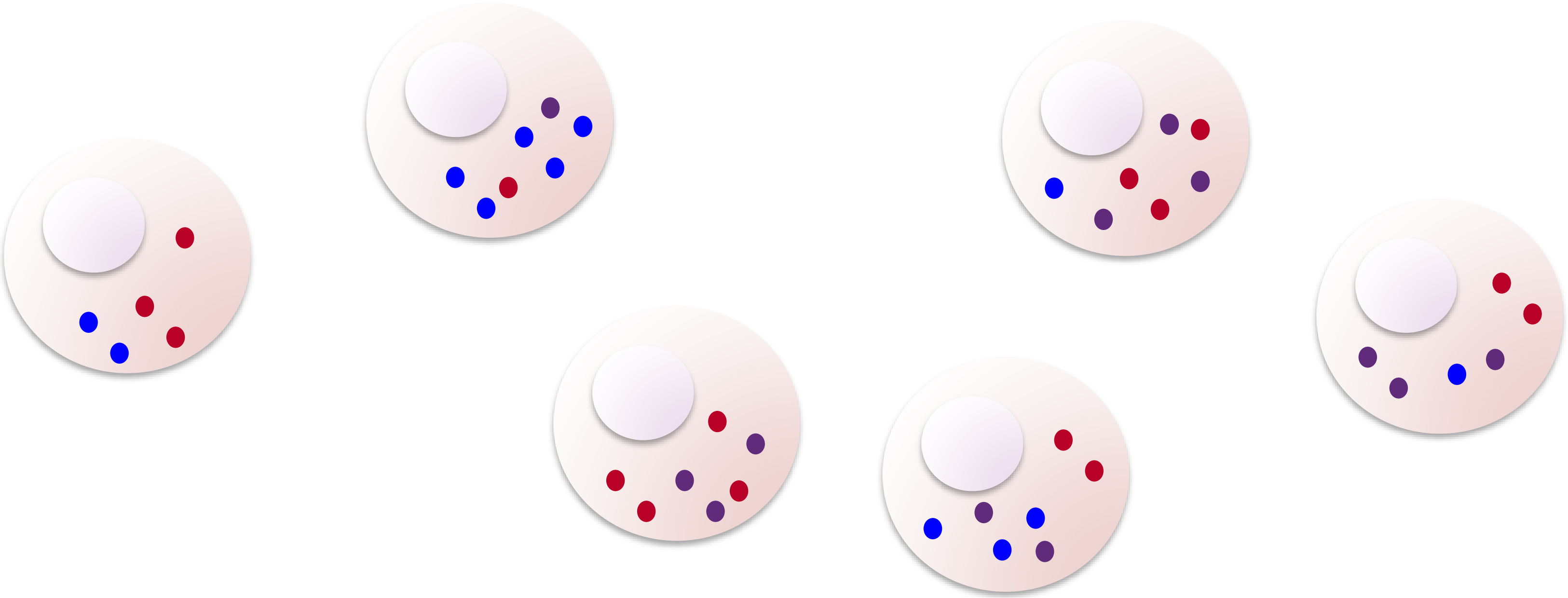
Dana Pe’er

Department of Biological Science

Department of Systems Biology

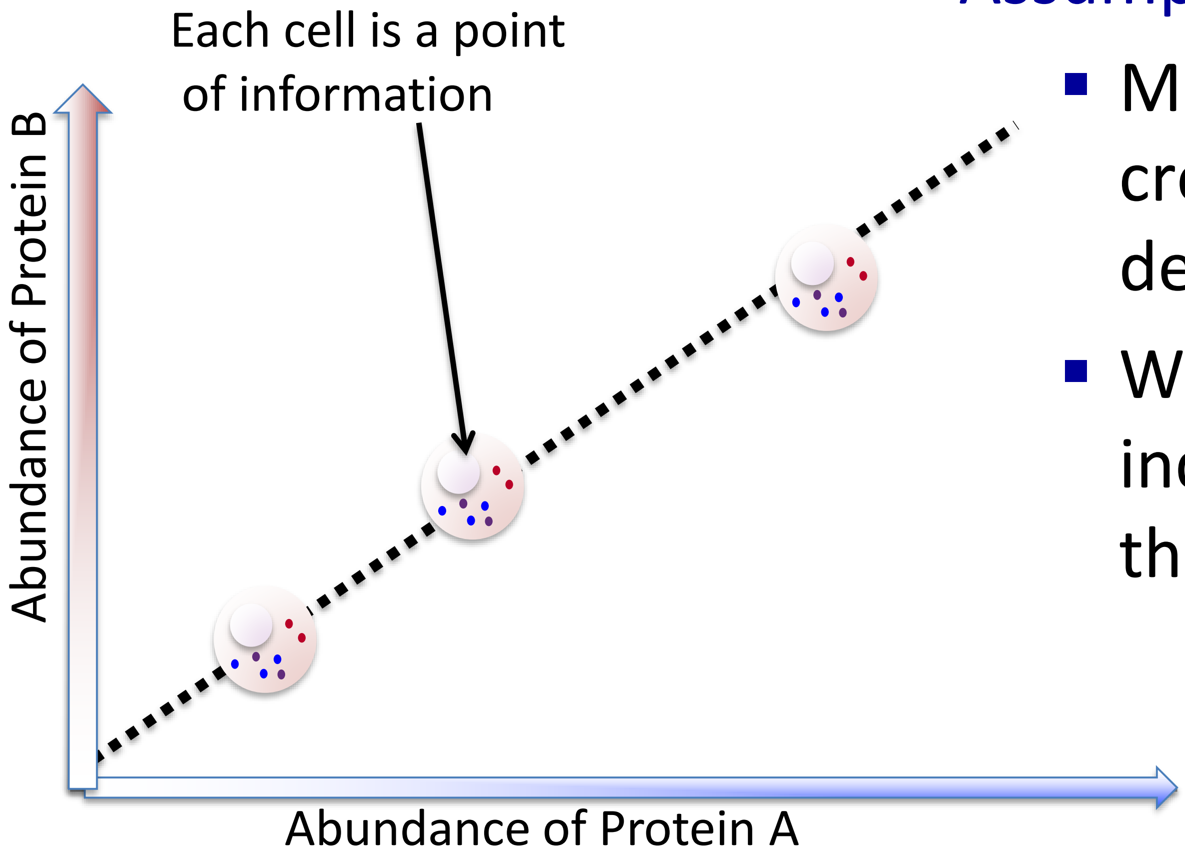
Columbia University

Learning Networks from Single Cells



- Idea: Use natural stochastic variation within a cell population and treat measurements of each individual cell as a sample for learning

Data-Driven Learning



Assumptions:

- Molecular influences create statistical dependencies
- We treat each cell as an independent sample of these dependencies.

How does protein A influence protein B?

Can we use single cells to learn signaling networks?

Karen Sachs

Omar Perez

Doug Lauffenburger

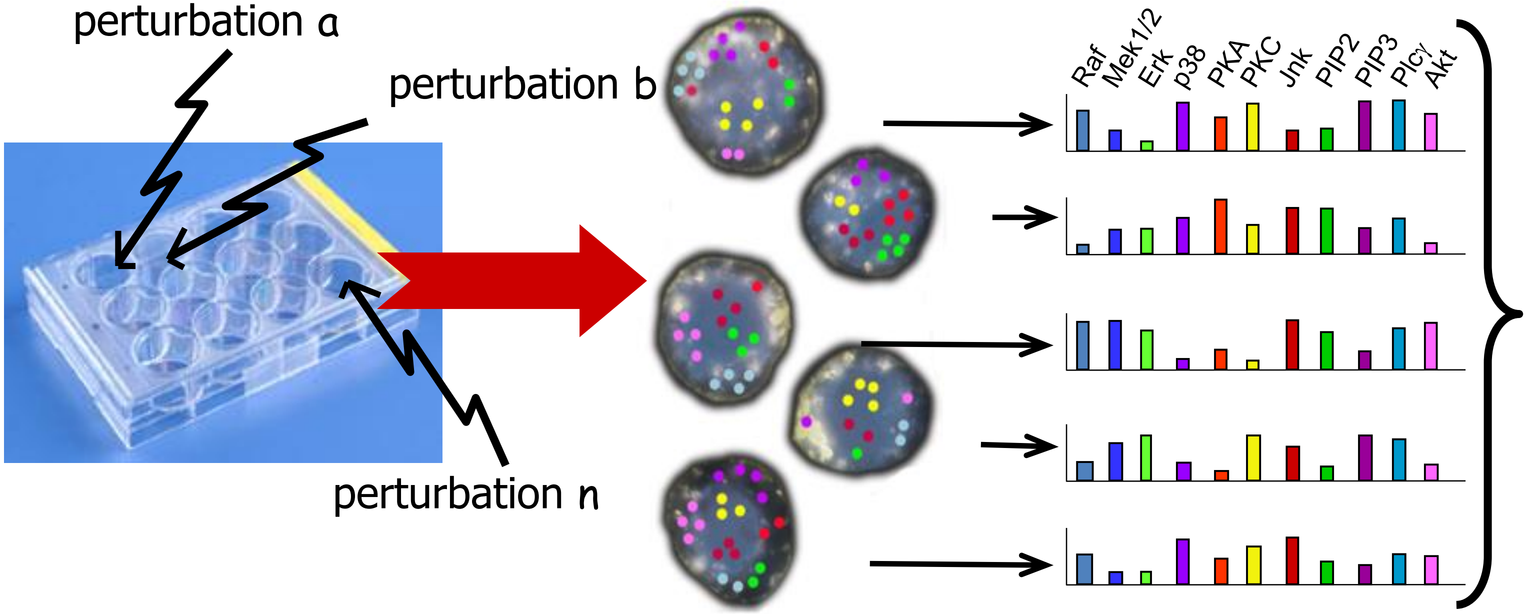
Garry Nolan

Sachs, Perez*, Pe'er* et.al. Science 2005*

Primary Human T-Lymphocyte Data

Conditions (96 well format)

12 Color Flow Cytometry



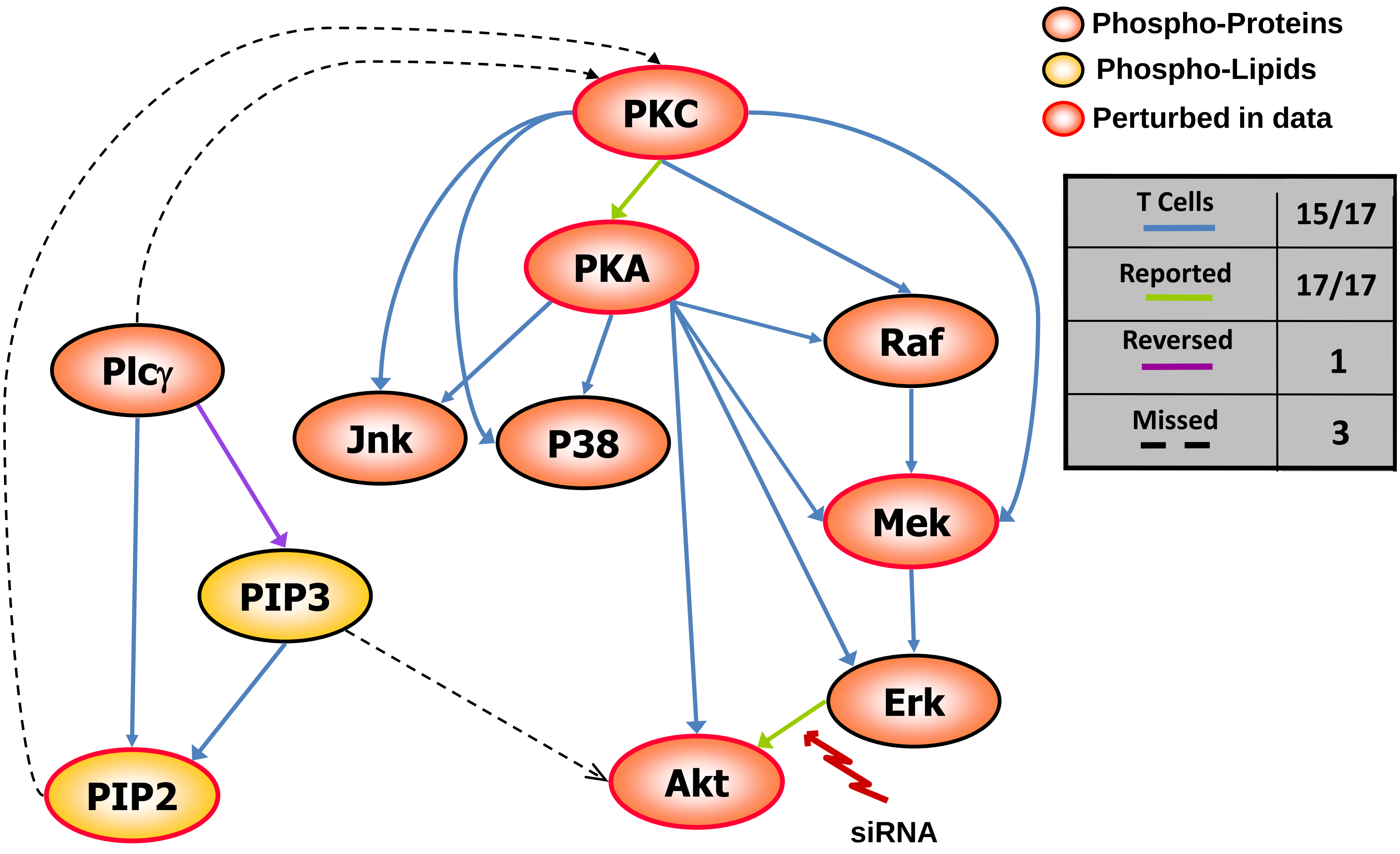
Datasets of cells

- *condition 'a'*
- *condition 'b'*
- *condition… 'n'*

Assumptions:

- Treat perturbation as an "ideal intervention" (Cooper, G. and C. Yoo (1999).

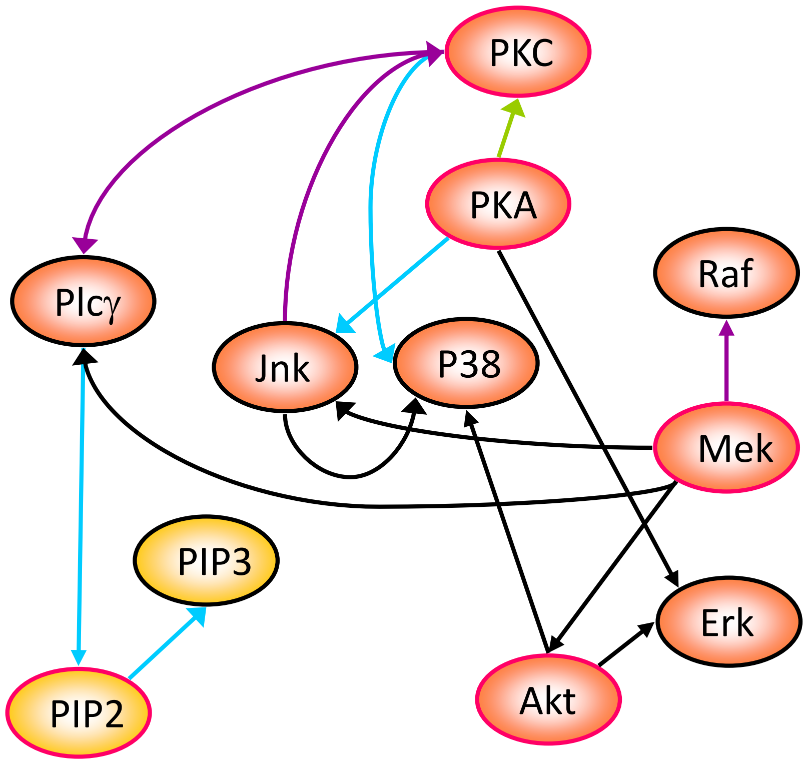
Inferred T cell signaling map



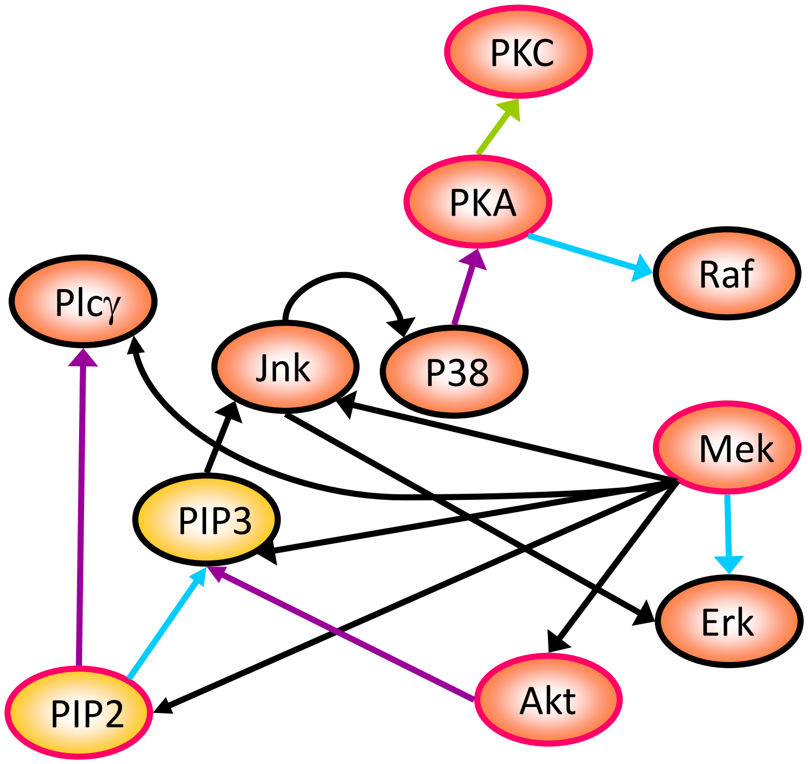
T Cells ━━	15/17
Reported ━━	17/17
Reversed ━━	1
Missed ▬ ▬	3

[Sachs *et al, Science* 2005]

What did we need to succeed?

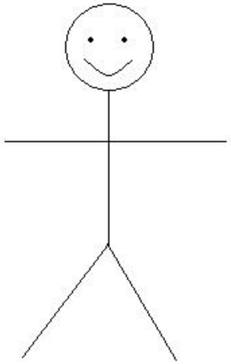


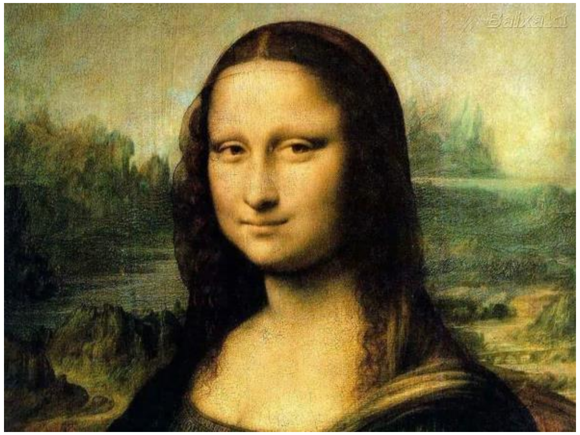
420 instead of 6000 samples



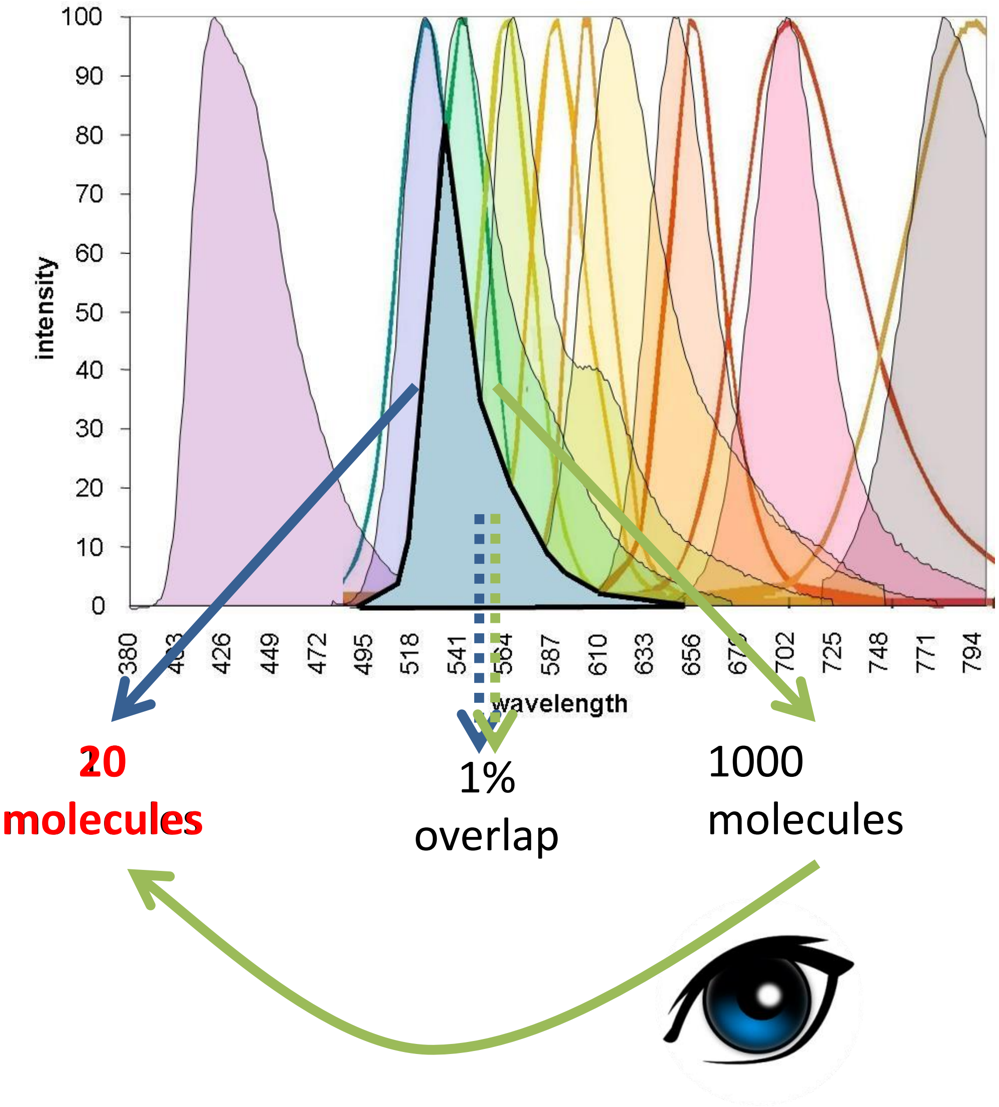
420 averaged samples

Large number of samples and single cell resolution are needed for success





Spectral overlap in flow cytometry



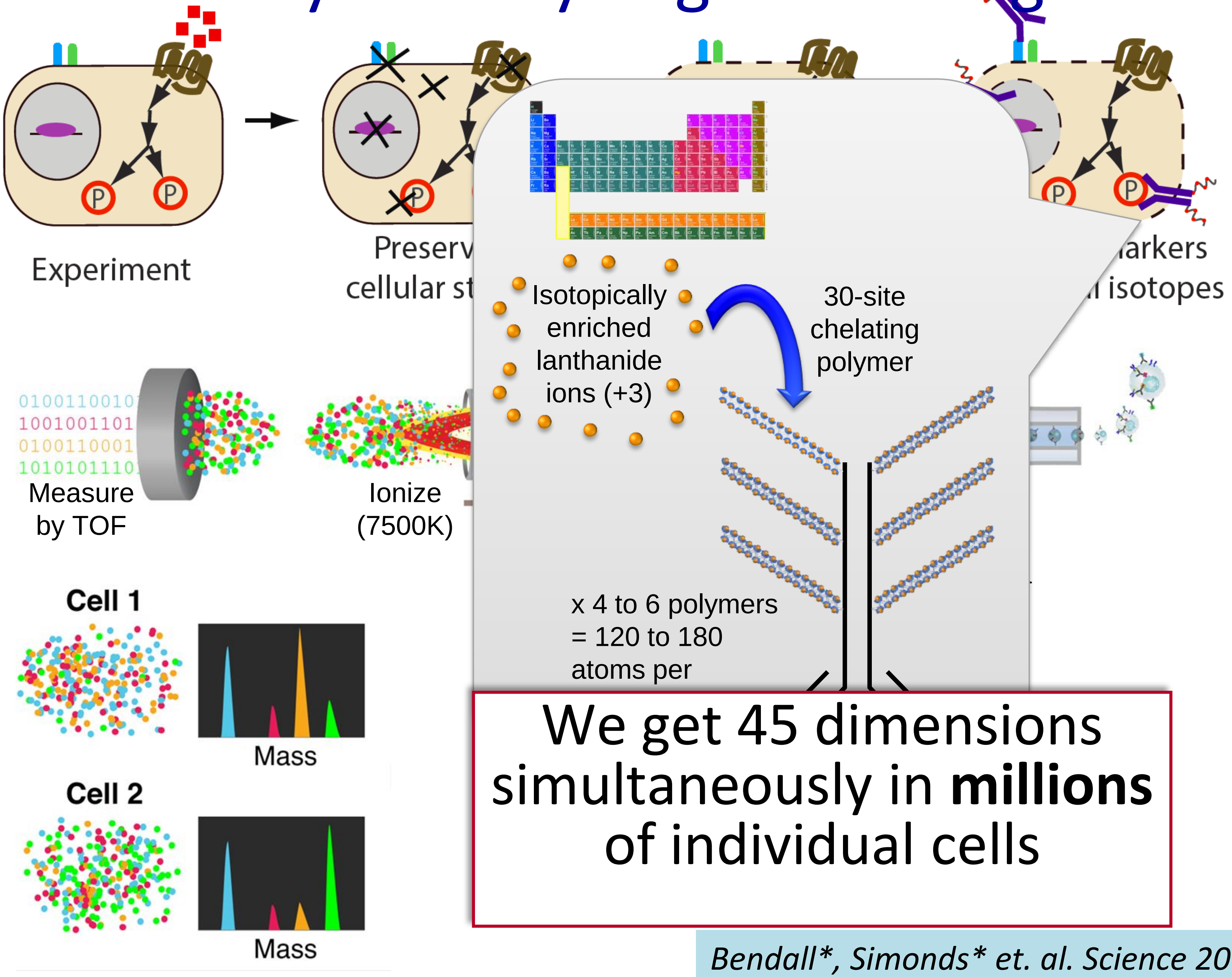
20 molecules

1% overlap

1000 molecules

http://www.dvssciences.com/technical.html

Mass cytometry: a game changer

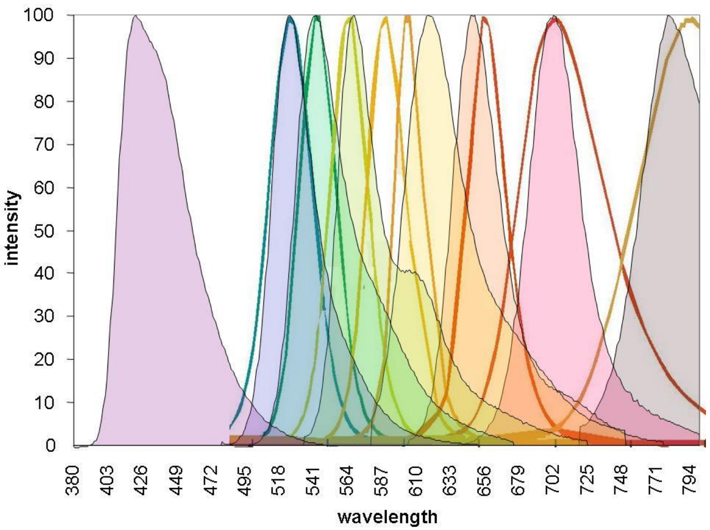


We get 45 dimensions simultaneously in **millions** of individual cells

Bendall, Simonds* et. al. Science 2011*

Mass cytometry

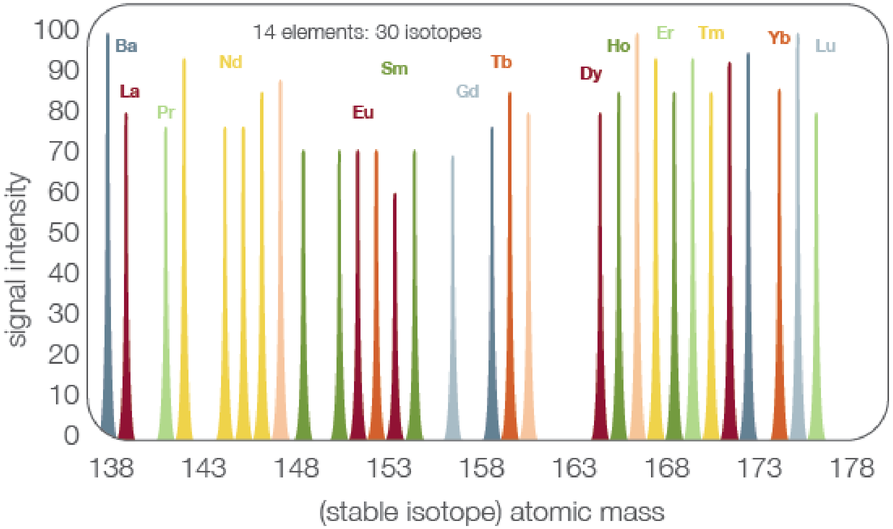
45 dimensions and counting



Decreased spectral overlap

➔

Increased dimensionality



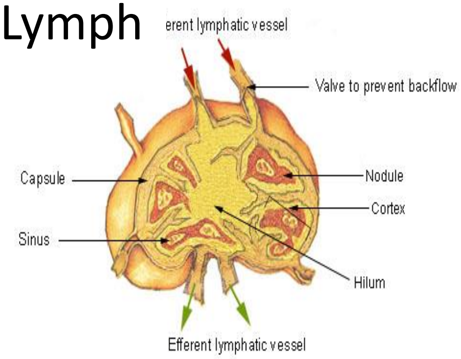
How does signal processing differ between subtypes?

Smita Krishnaswamy
Matthew H. Spitzer
Michael Mingueneau
Sean C Bendall
Oren Litvin,
Erica Stone
Garry Nolan

Krishnaswamy et.al. Science 2014

Signaling Through T-cell Maturation

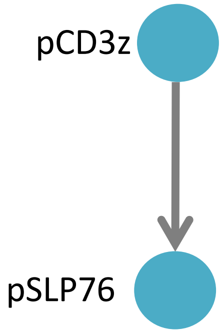
Lymph



Naïve (CD44-) → Effector/Memory (CD44+)

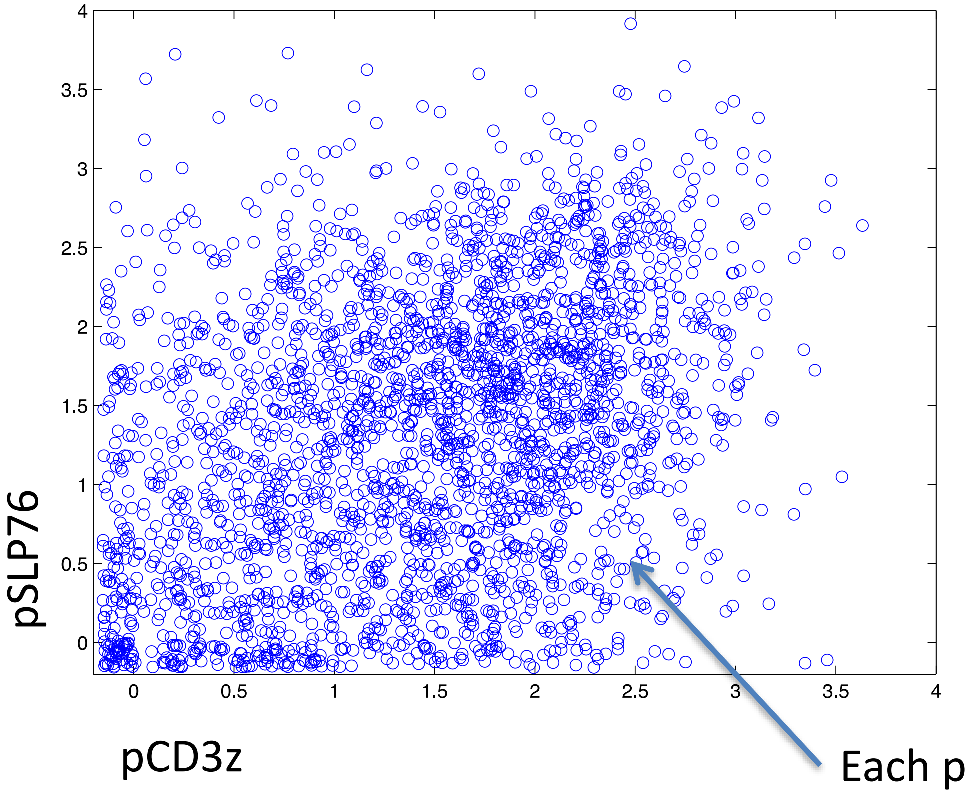
- Naïve and effector memory CD4+ T-cells have similar signaling network, yet these respond differently
- Our surface panel has enough markers to resolve key T-cell subsets together with their signaling
- They have been stimulated and processed in the same tube allowing for direct comparison

Real Mass Cytometry Data



pCD3z

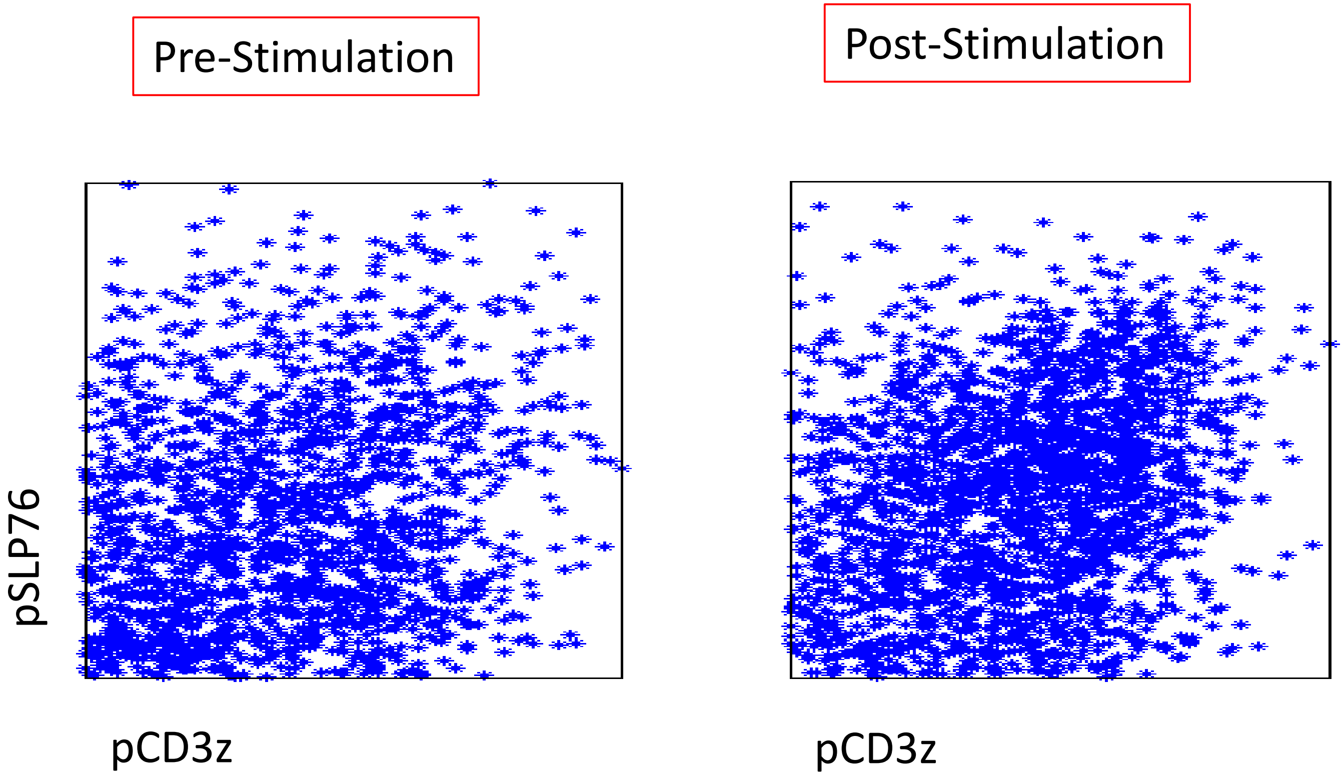
pSLP76



Each point is a cell

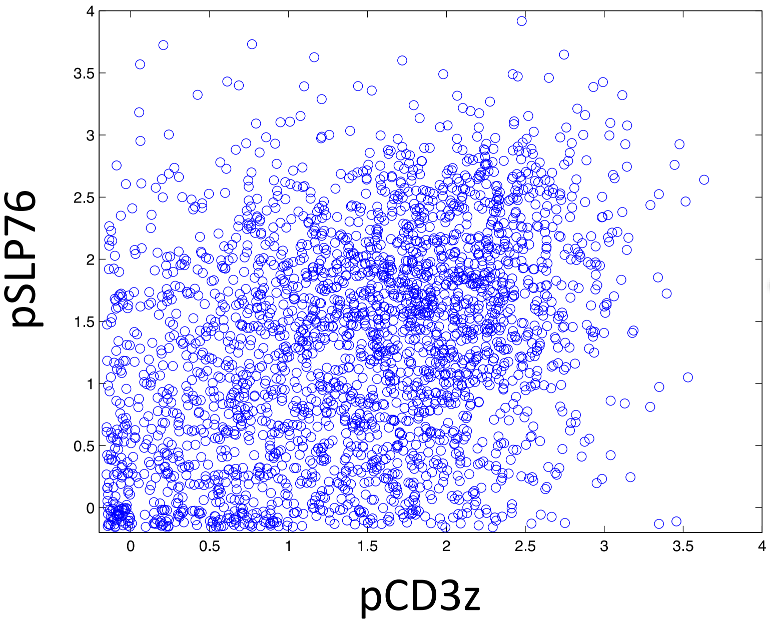
Units of measurement: log-scale transformed molecule counts

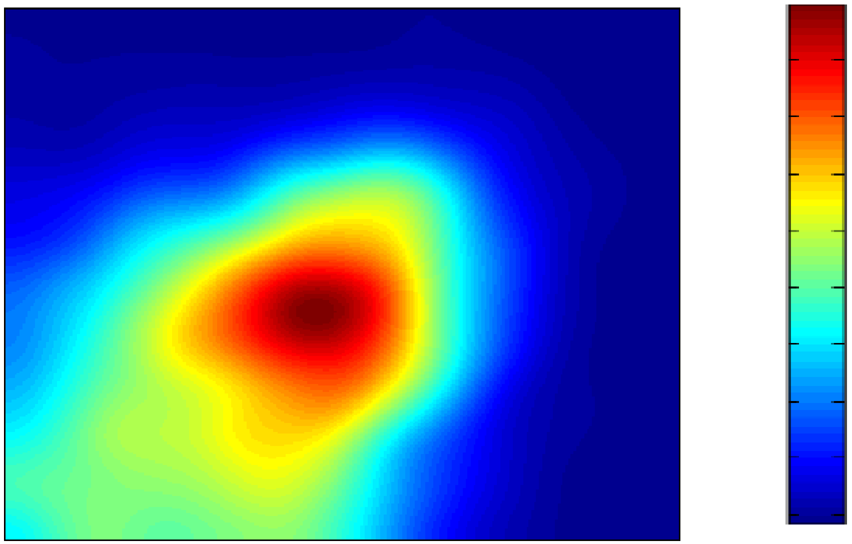
Scatterplots Reveal Only Range



Cannot discern effect of stimulation

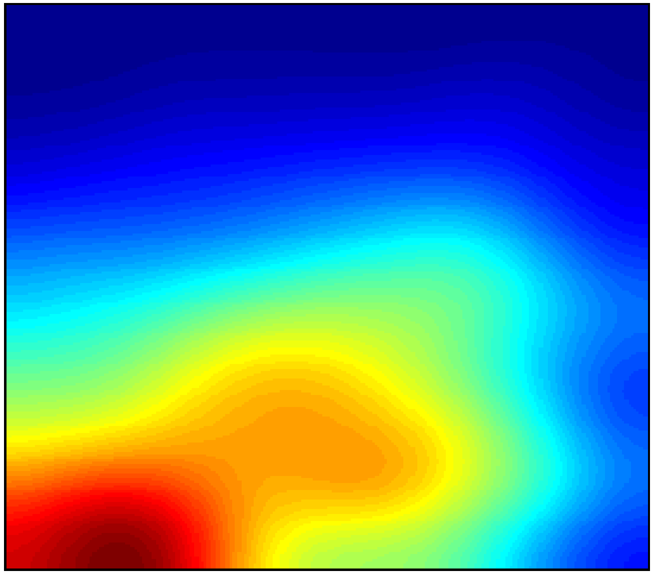
Kernel Density Estimation (KDE) learns underlying probability distribution



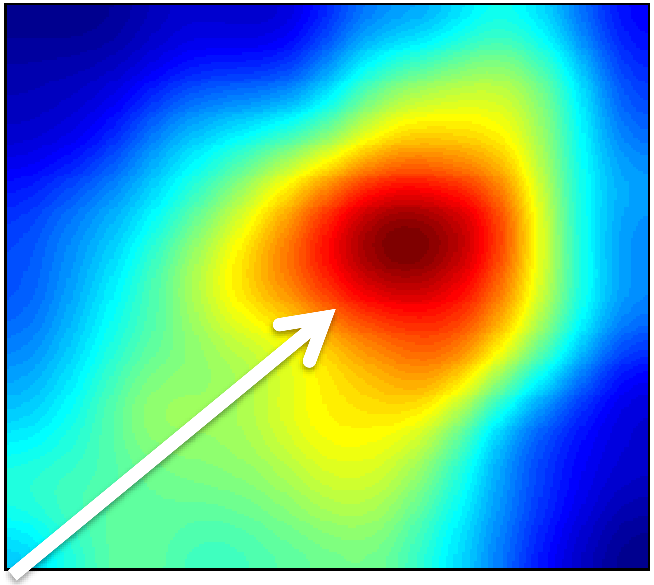


KDE obscures X-Y relationship

Pre-Stimulation

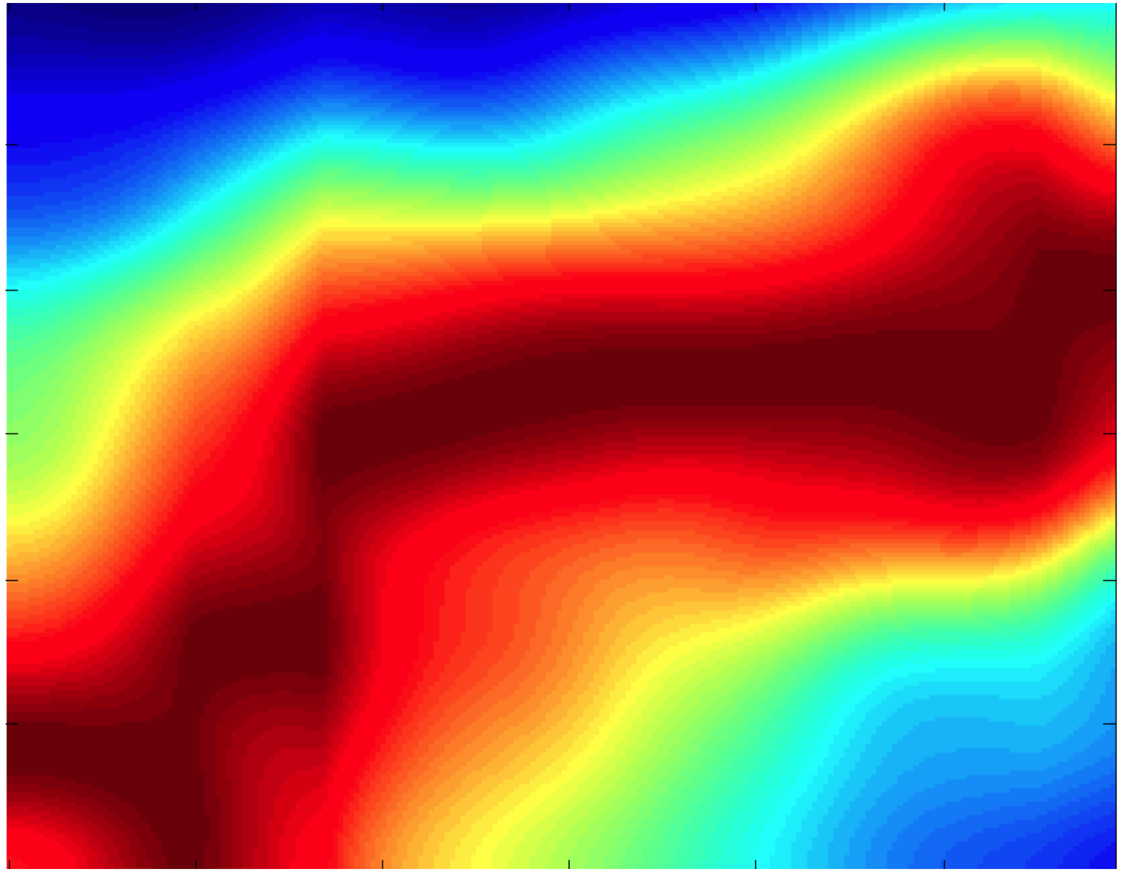


Post-Stimulation



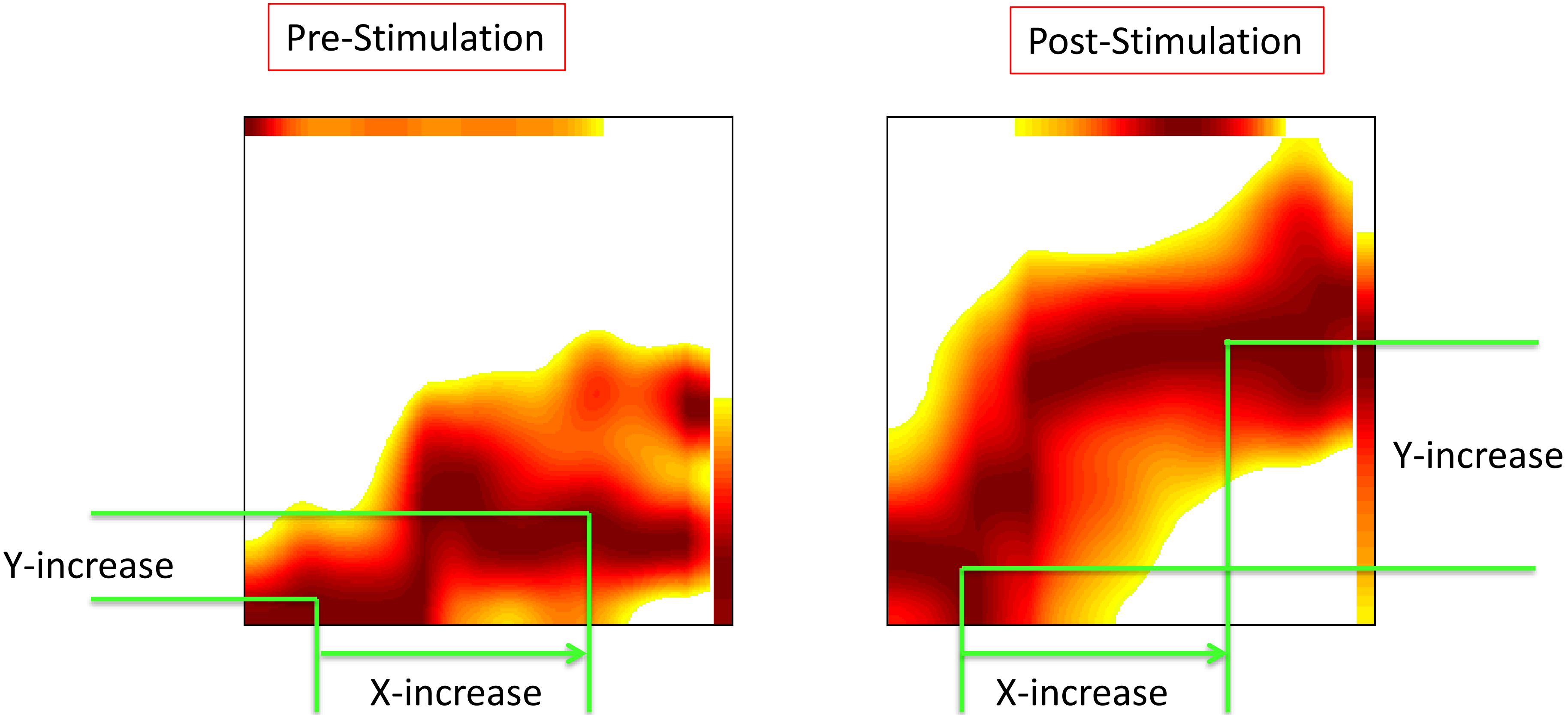
- Molecules shift together
- Coarse functional relationship

Conditioning unveils X-Y Relationship



- Captures behavior across full dynamic range
- Captures behavior of small populations of responding cells

Change in Signal Transfer Relationship

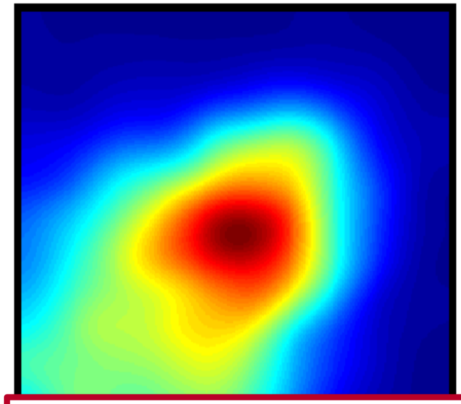


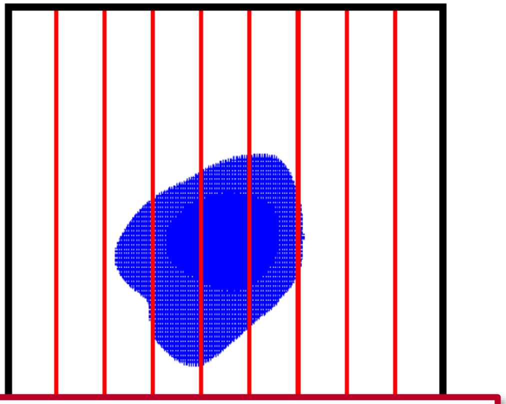
Pre-Stimulation

Post-Stimulation

This is beyond “increasing pCD3z levels”

How do we quantify information transmitted by an edge?

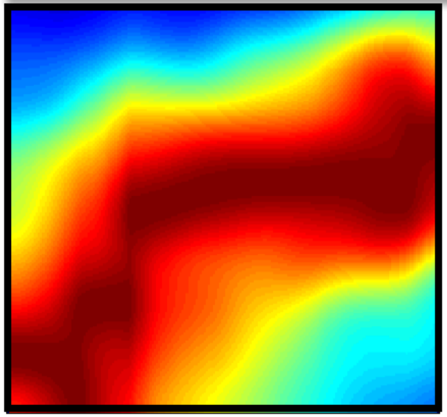


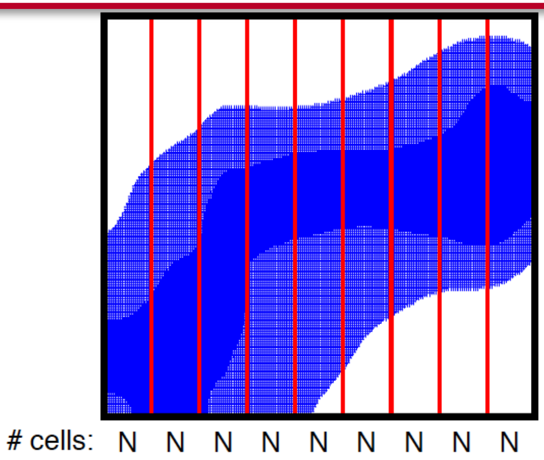


The high local joint density biases mutual information assessment

The key is we want to model P(Y|X)

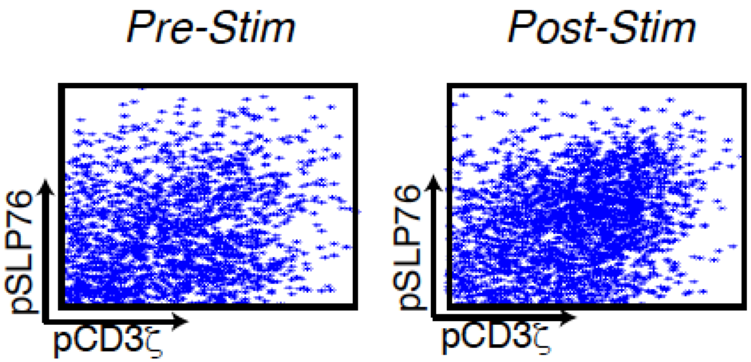
Rather than P(X,Y)





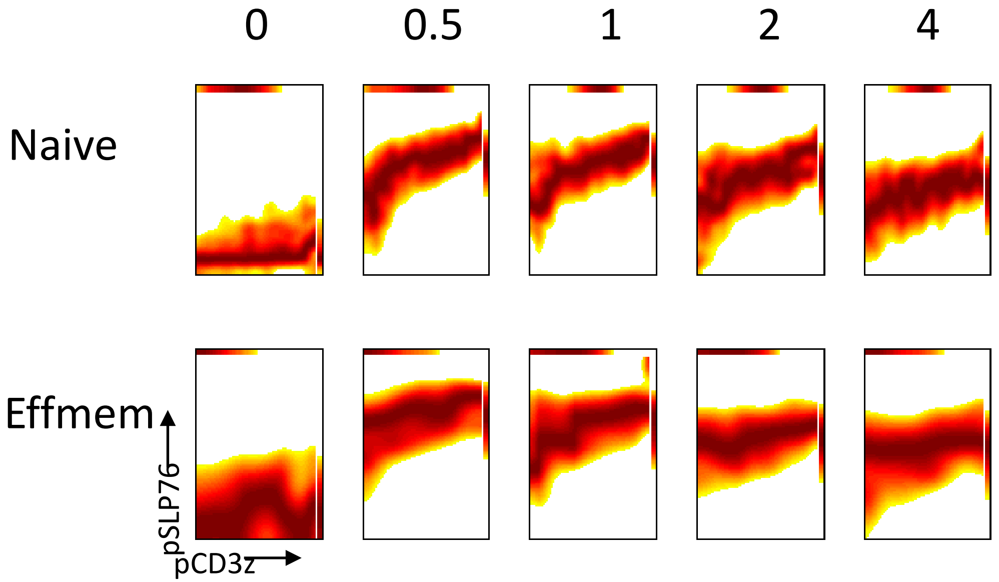
DREMI resamples Y from conditional density in each X-slice to reveal relationship between X and Y

DREMI captures “edge strength”

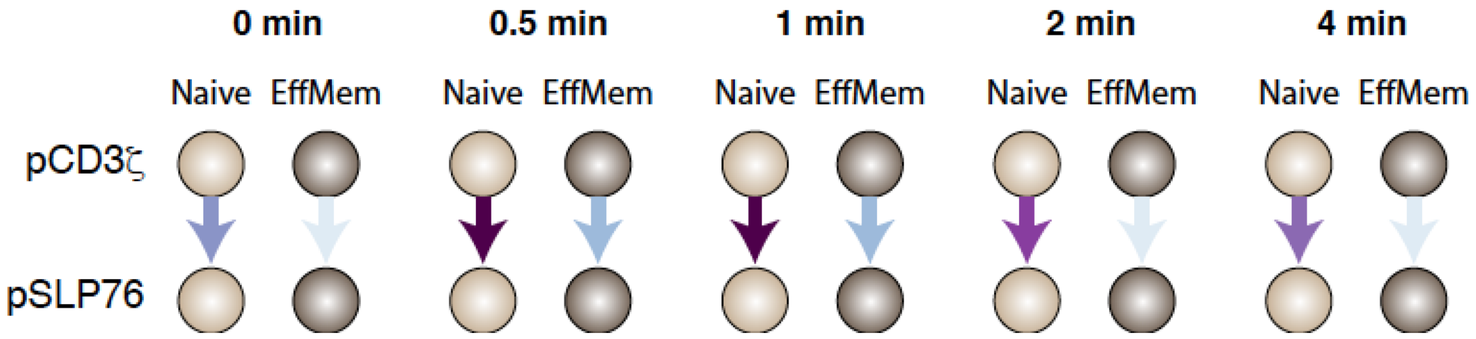


	Pre	*Post*
R^2	0.32	0.35
MI	0.07	0.06

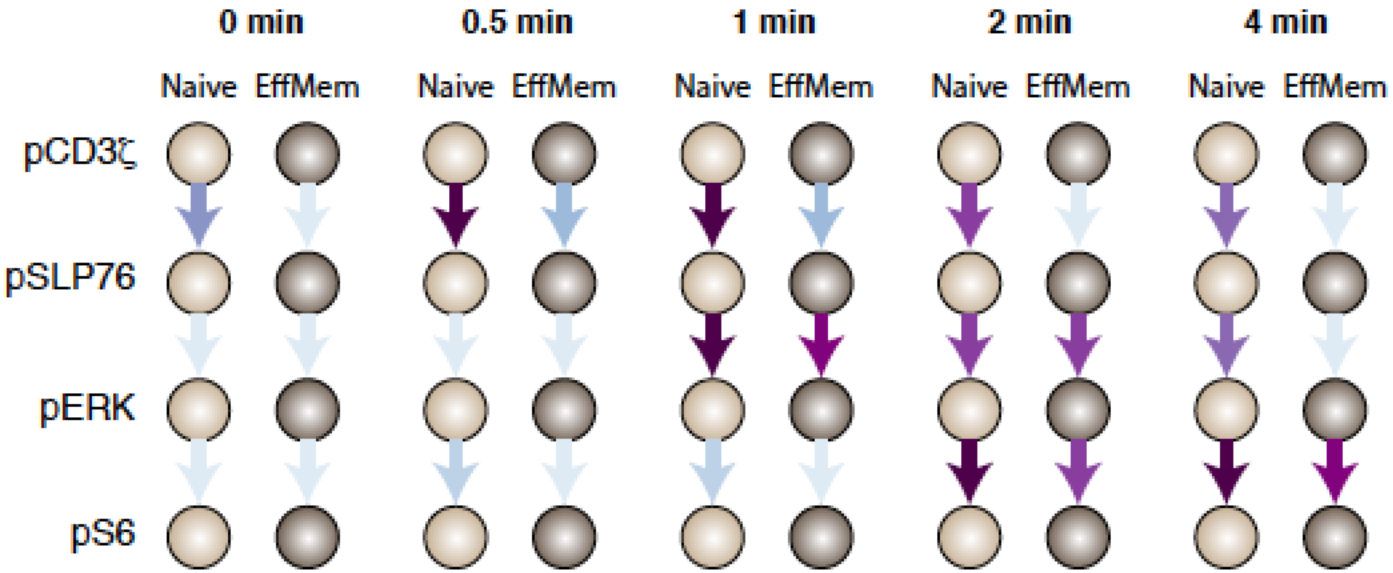
Comparing Naïve to Effector memory T-cells



- pSLP76 responds more strongly in effmem T-cells
- The “edge” transmits pCD3z levels more faithfully in naïve T-cells

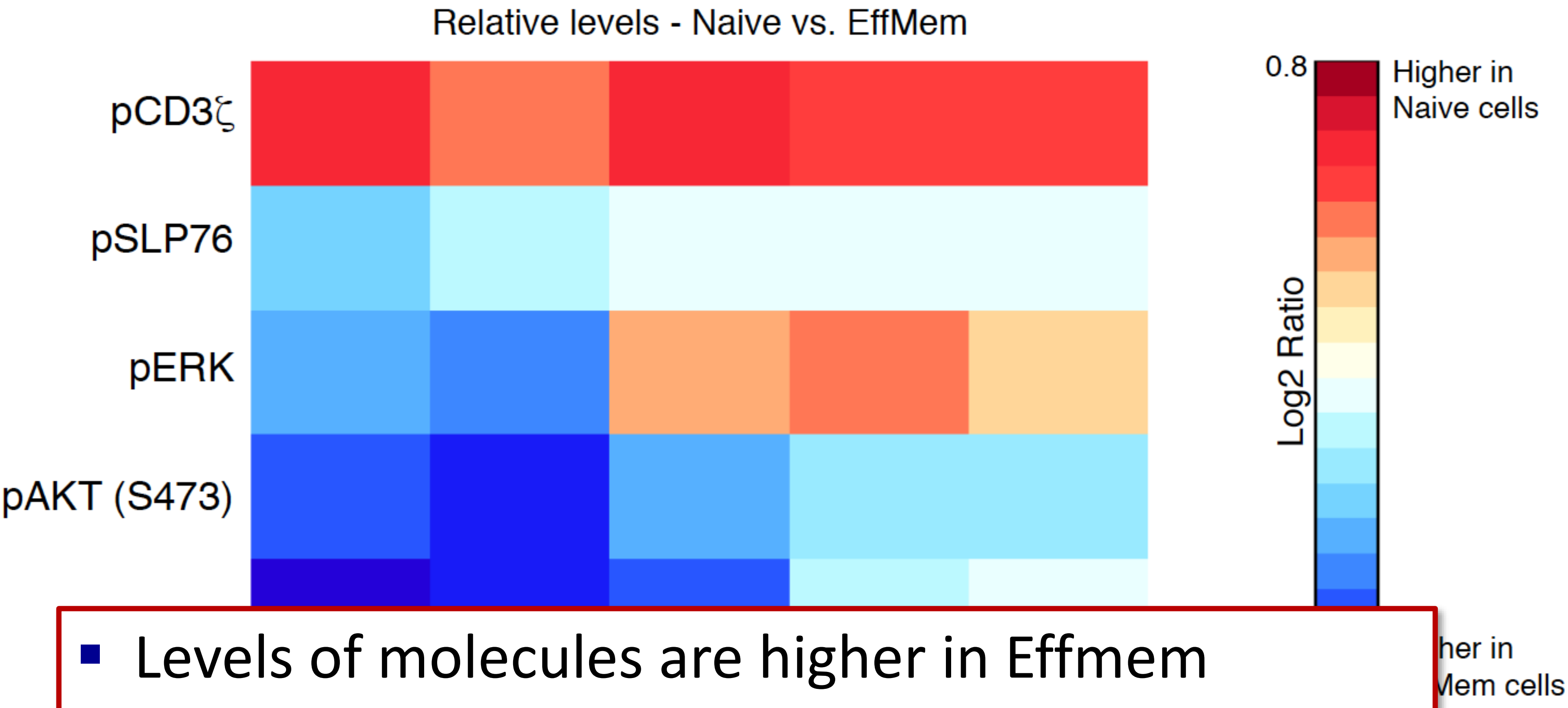


Comparing Naïve to Effector memory T-cells



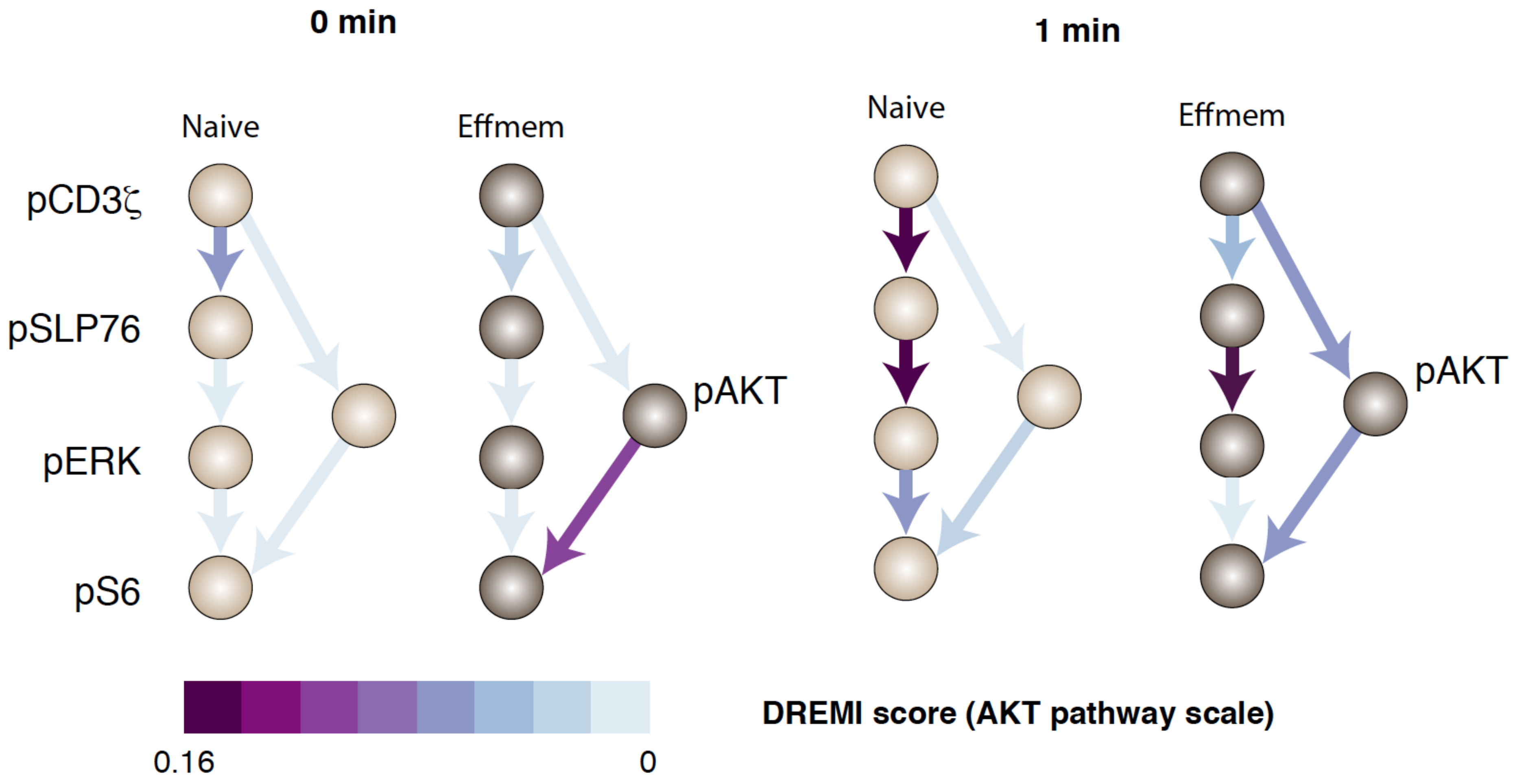
- Increased transmission of input in naïve T-cells propagates down
- For a longer duration

Protein Activation: a Different View



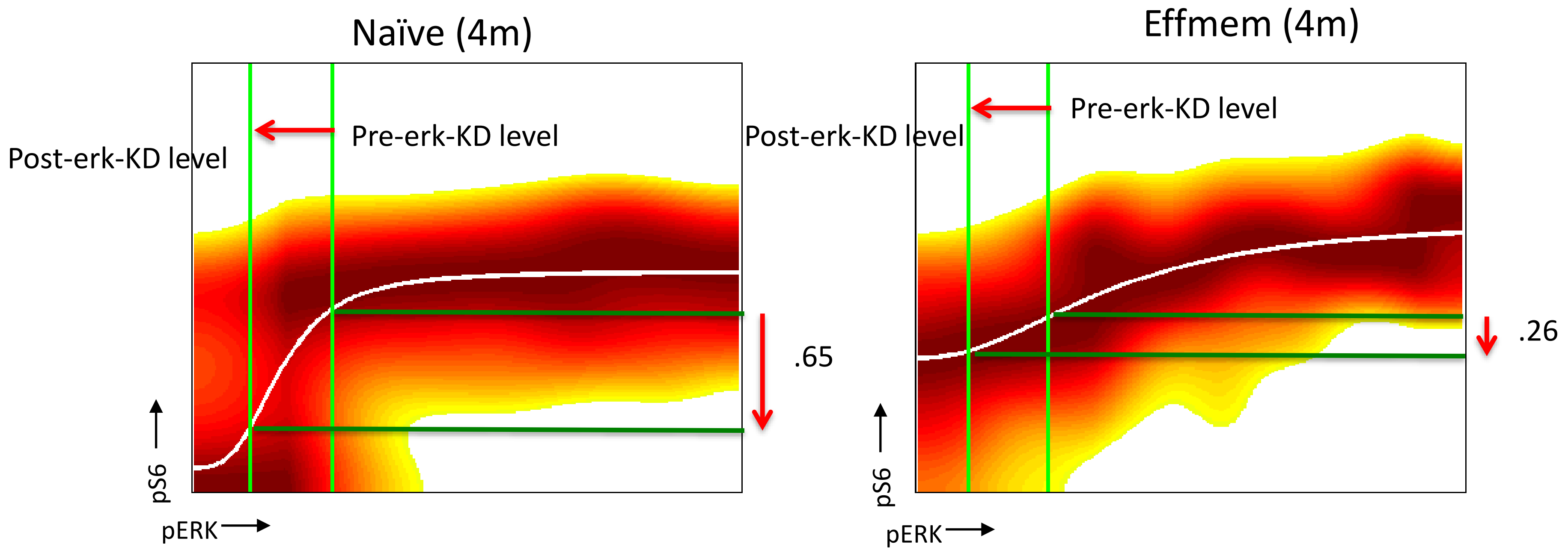
- Levels of molecules are higher in Effmem
- Effmem cells need less antigen to trigger
- Naïve cell responses are more tailored to input

DREMI Reveals Alternative Pathway



Effmem cells have alternate input via AKT pathway

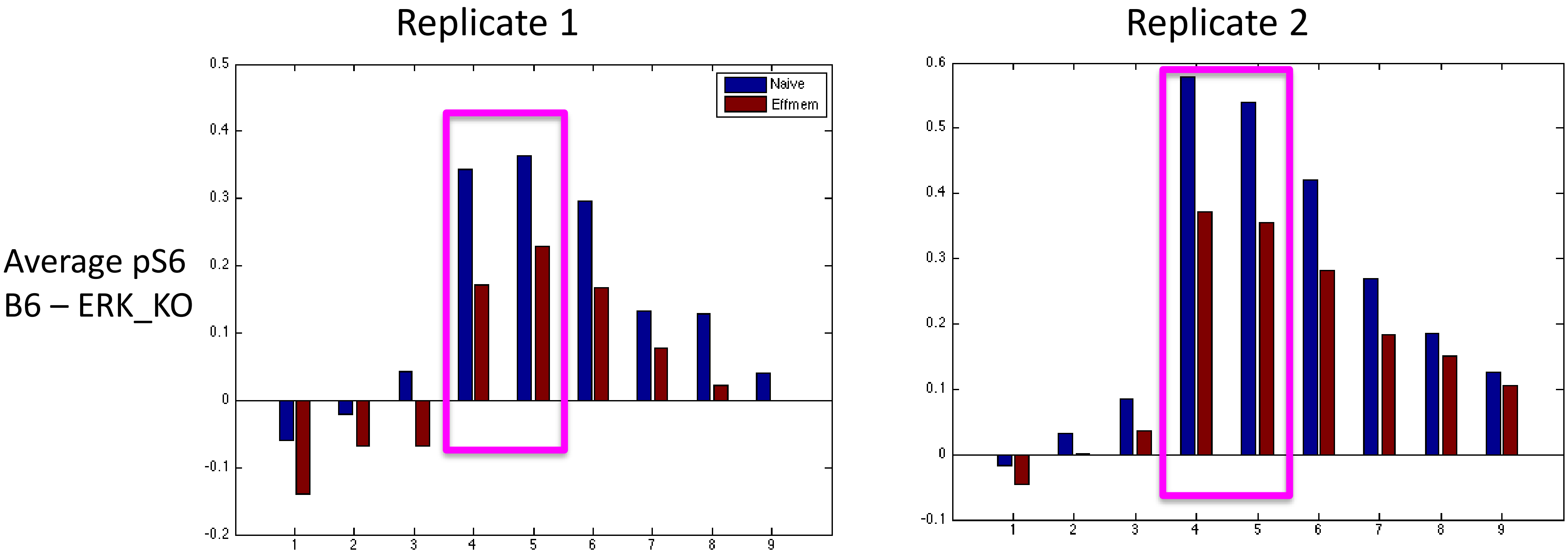
Predicting differences in “edge” strength



Predictions for ERK KO mouse

- Erk_KO should impact pS6 more in Naïve cells
- Difference should accentuate at the 3 minutes after stimulus

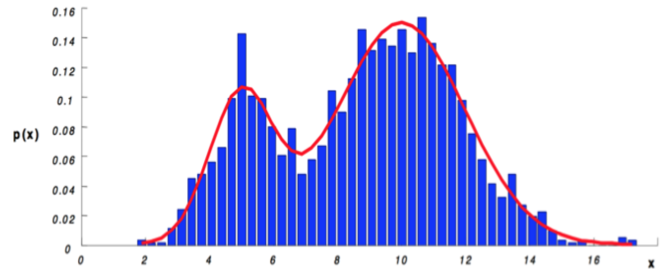
Validation of edge strength prediction



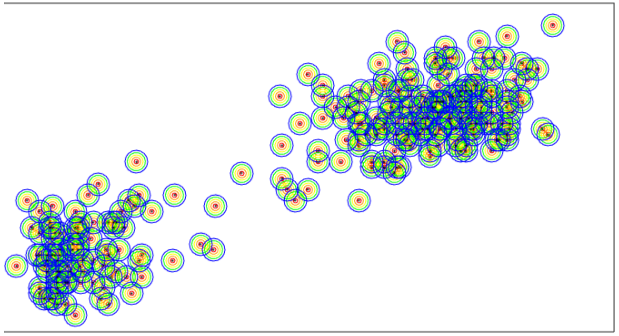
- We validated that the influence of pERK on pS6 is stronger in Naïve T-cells.
- Similar validation for differences between CD4 and CD8

The devil is in the details

- KDE's interpolate over areas where there are no samples, so they correct for gaps to some extent.
- Histogram approach, fast, but sensitive to bandwidth



- Kernel approach, slow and tedious need to integrate all kernels at every point of evaluation, most heuristics sensitive to noise



Hybrid Method for Density Estimation

- We take a hybrid method for density estimation.

- Use the speed of histogram and the smoothness of Kernels:
    - 1. Build a histogram of the initial data
    - 2. Obtain a good estimate of the bandwidth
    - 3. Smooth the histogram using the bandwidth.

- Goal:

$$\hat{f}_h(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{h^2(x-x_i)^2}{2}}$$

Botev *et.al.,* Annals of Statistic, 2010

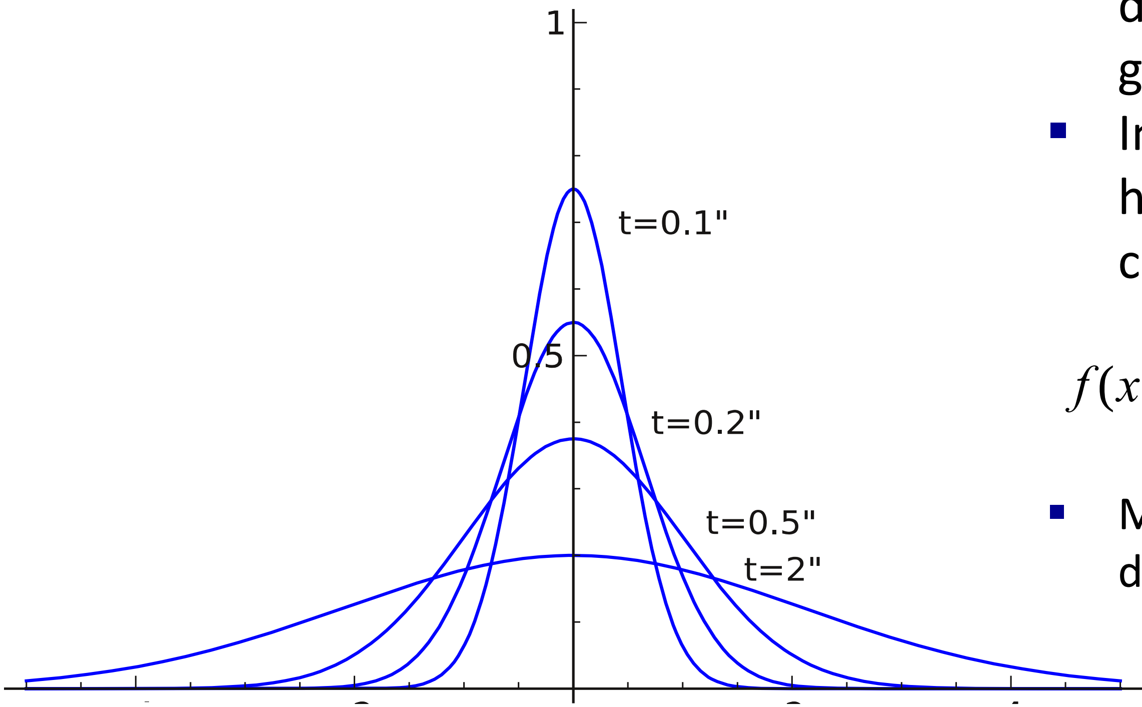
Connection to heat equation

- Heat Equation: $\frac{\partial f}{\partial t}=\frac{1}{2}\frac{\partial^2 f}{\partial x^2}$, with initial condition: $f\left(x,0\right)=D$

- It governs the distribution of temperature in a region over time.

A Gaussian kernel, $\hat{f}_h(x)=\frac{1}{nh\sqrt{2p}}\overset{n}{\underset{i=1}{\text{å}}}\,e^{-\frac{h^2(x-x_i)^2}{2}}$ **(which is what we want) is the unique solution to the above equation!**

“Spreading of Heat” over time akin to Smoothing Data



- At t = 0, the initial condition is a delta peak at 0. For any t>0, we get a Gaussian.
- In finite domain, the solution to heat equation is a Fourier series in cosine

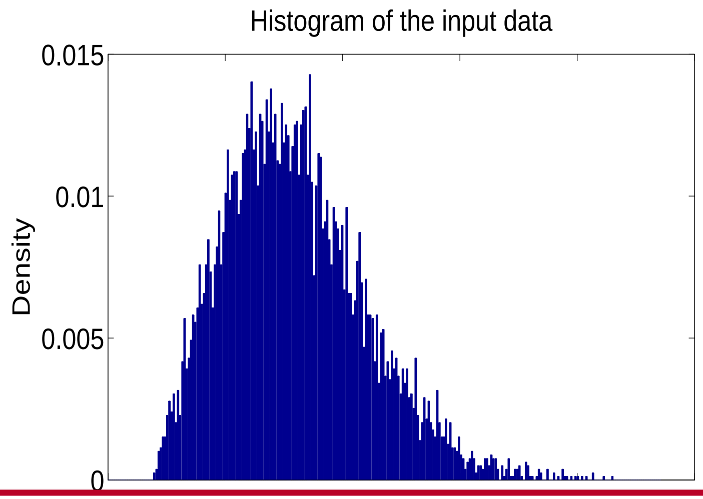
$$f(x) = \sum_{m=0}^{\infty} a_m \cos(m\pi x)\exp\left(-\frac{m^2\pi^2 t}{2}\right)$$

- Motivates us to work in frequency domain.

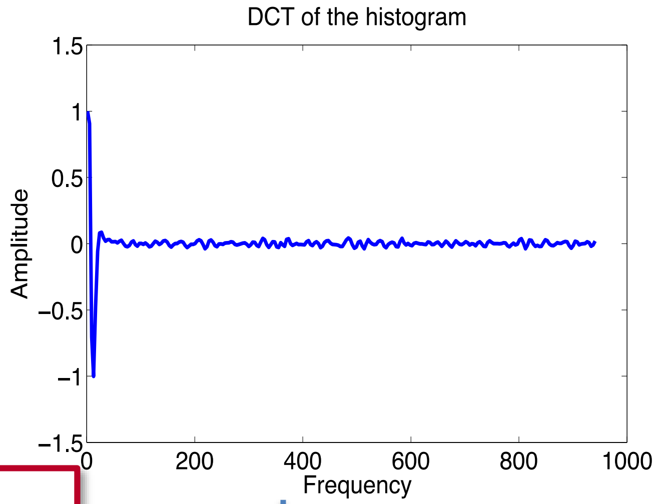
  => Solution = Discrete Cosine Transforms

- Facilitates rapid computation

Computing in frequency domain

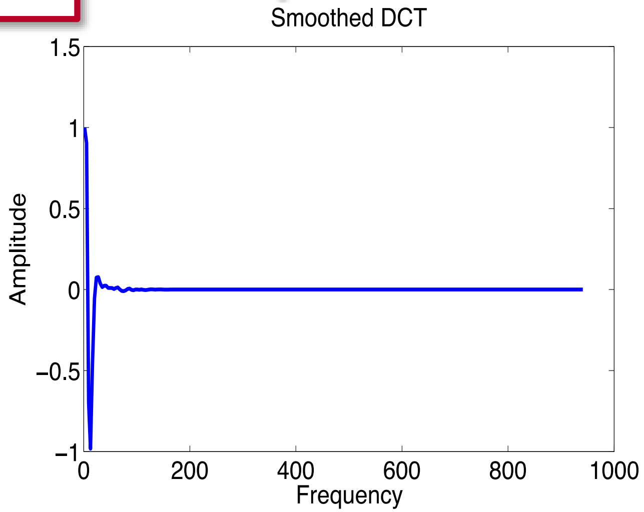


DC
T

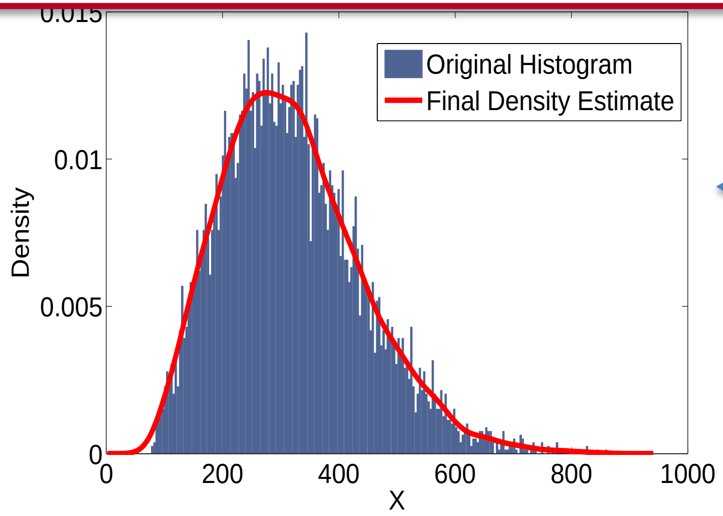


Smooth
DCT

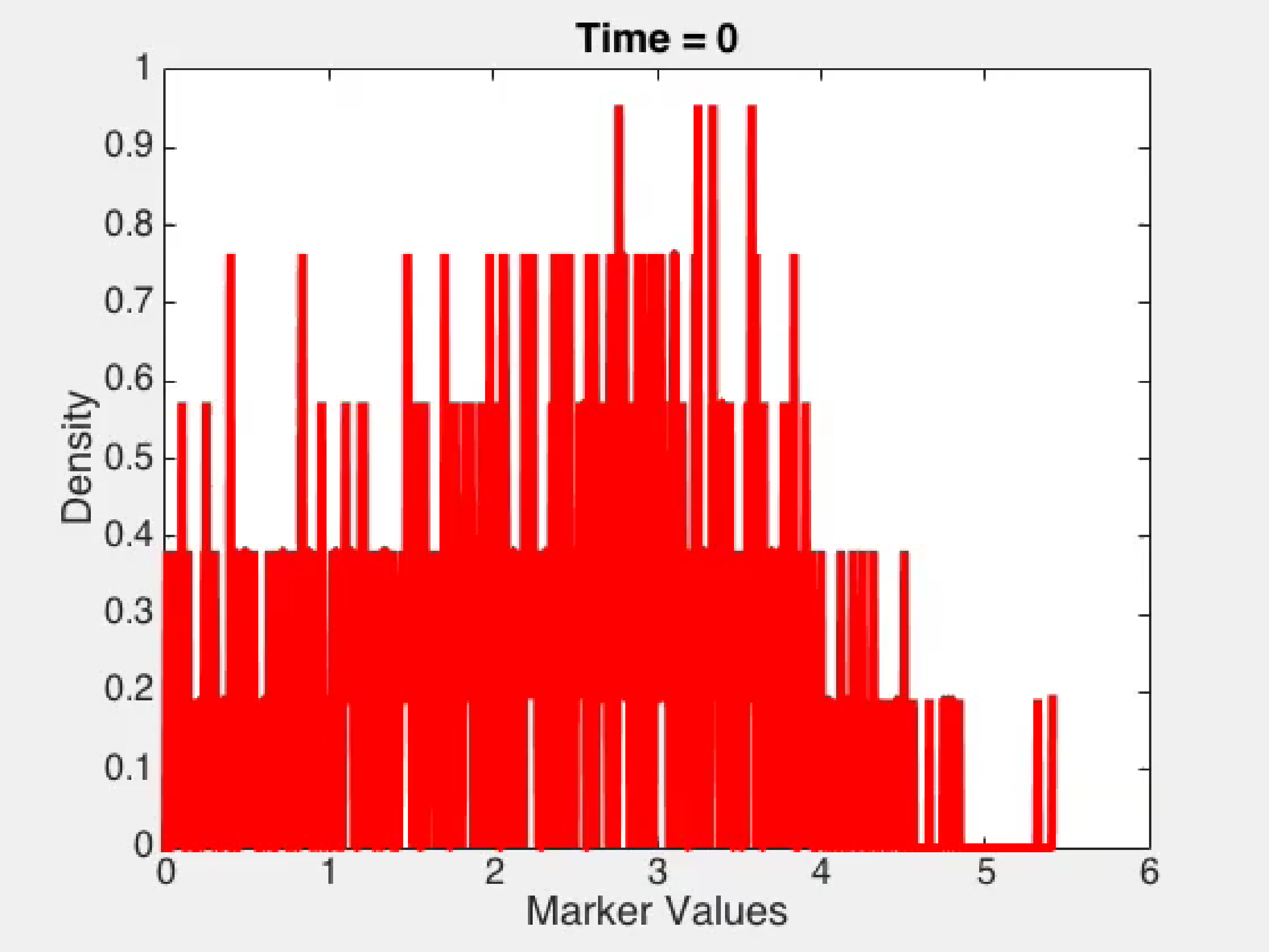
This is equivalent to solving heat diffusion in a bound space



Invert
Smooth
DCT



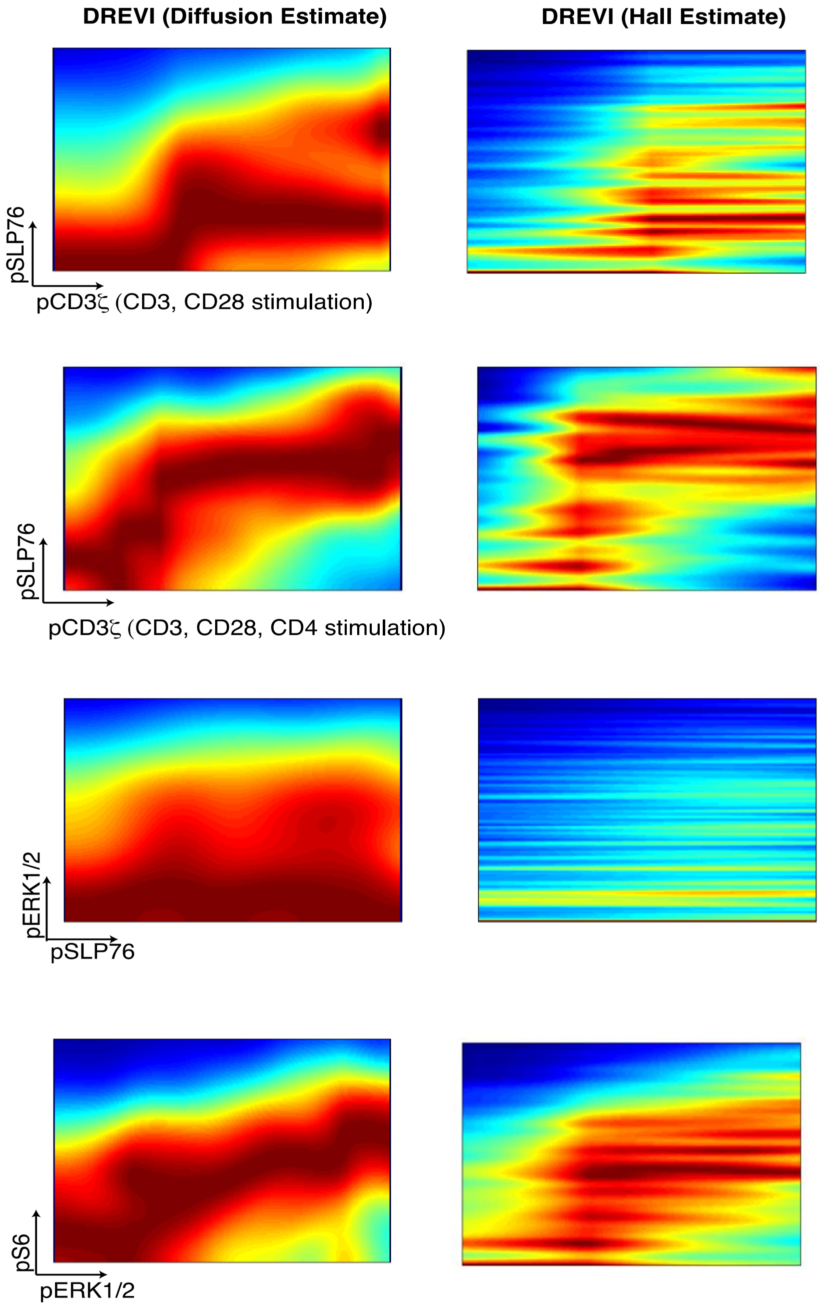
Smoothing in action: increasing the diffusion



Diffusion KDE

Diffusion-based KDE estimate is faster and smoother

Botev, et al., Annals of Stats, 2011



DREVI (Diffusion Estimate)

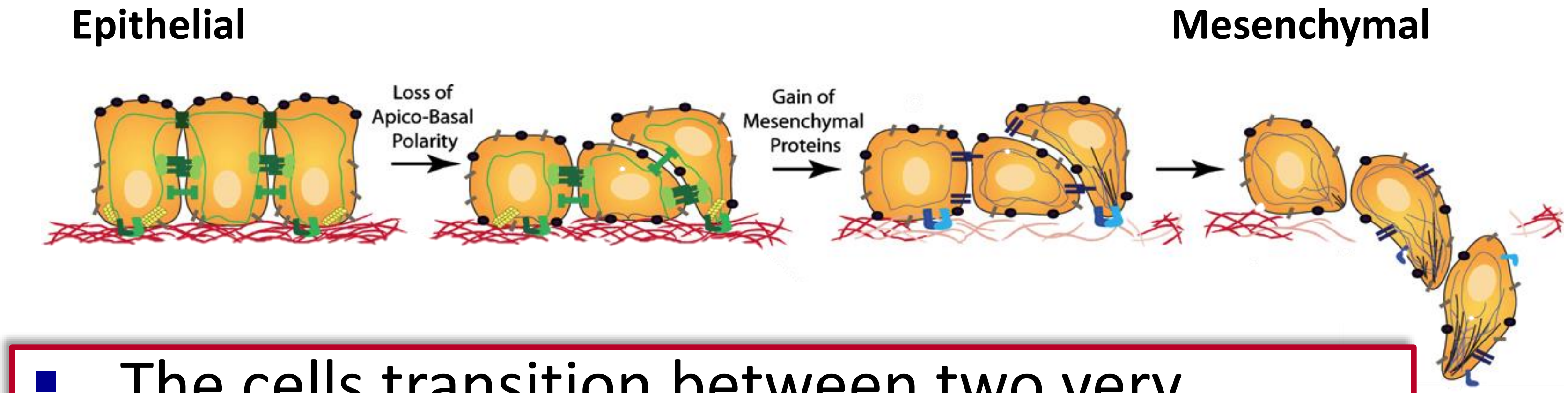
DREVI (Hall Estimate)

	DIffusion Estimate	Hall Estimate
DREMI	0.20	0.21
Runtime	0.26s	178.70s
DREMI	0.27	0.26
Runtime	0.25s	211.08s
DREMI	0.08	0.00
Runtime	0.30s	1187s
DREMI	0.20	0.27
Runtime	0.31s	899s

Reconfiguring Signaling Edges Driving EMT

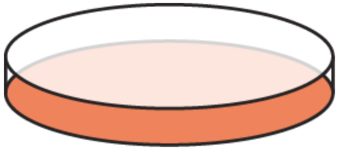
Smita Krishnaswamy
Roshan Sharma
Nevana Zivanovic
Bernd Bodenmiller

Epithelial-mesenchymal transition (EMT)



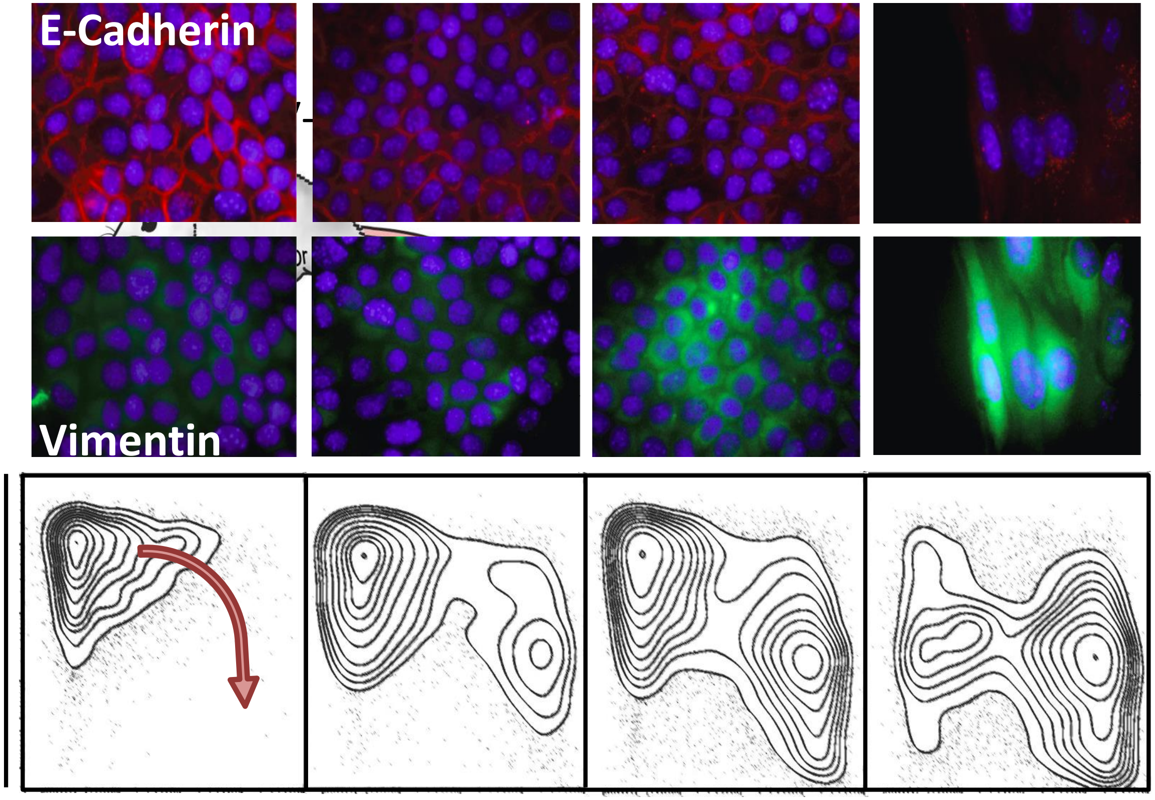
- The cells transition between two very different states.
- Can we understand the changes in signaling and phenotype underlying this transition?

Induce EMT by treating a breast cancer cell line with TGFB



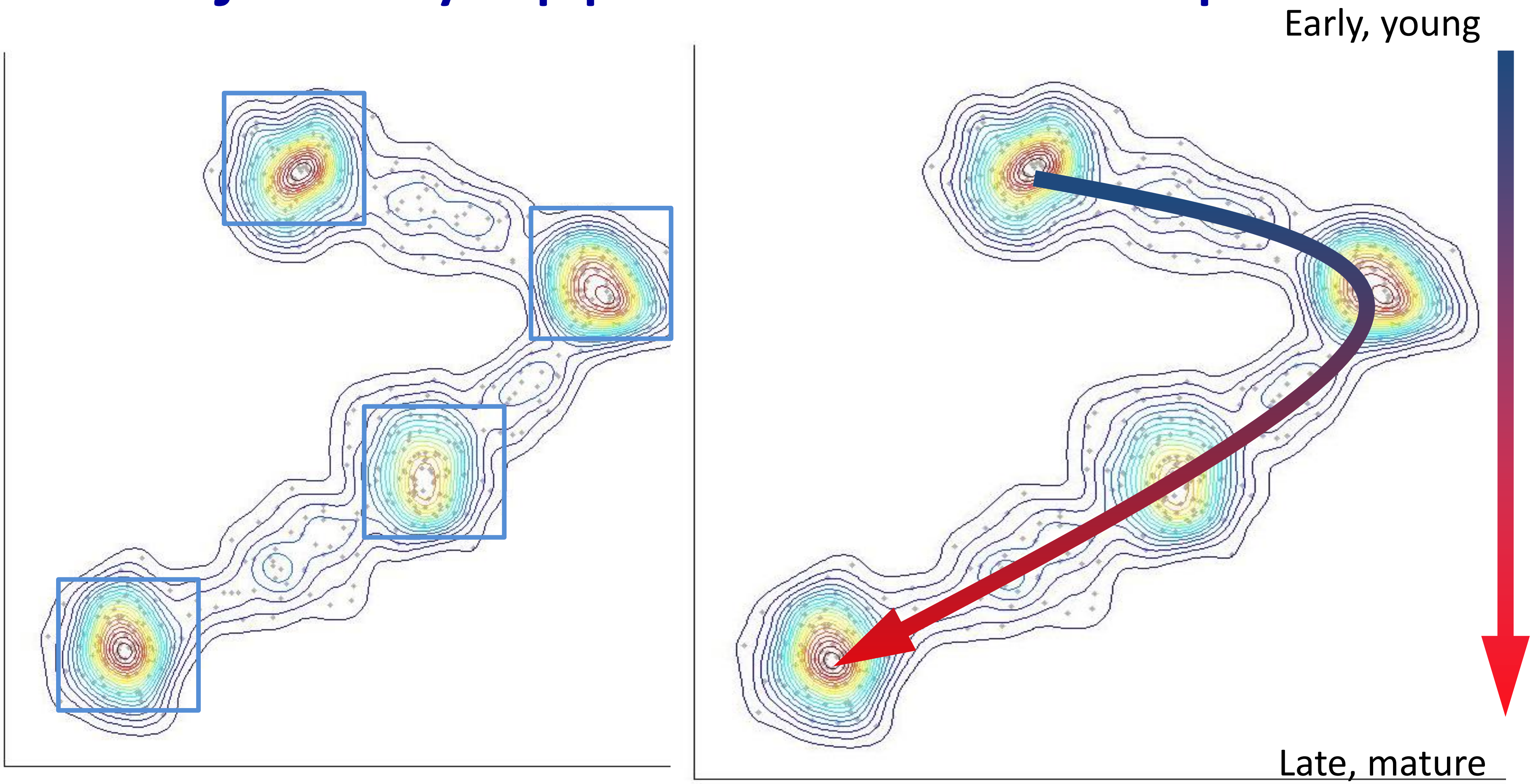
EMT: State Change in Cells

- Cellular heterogeneity: both epithelial and mesenchymal cells coexist during transition.



Both epithelial and mesenchymal cells at day 3

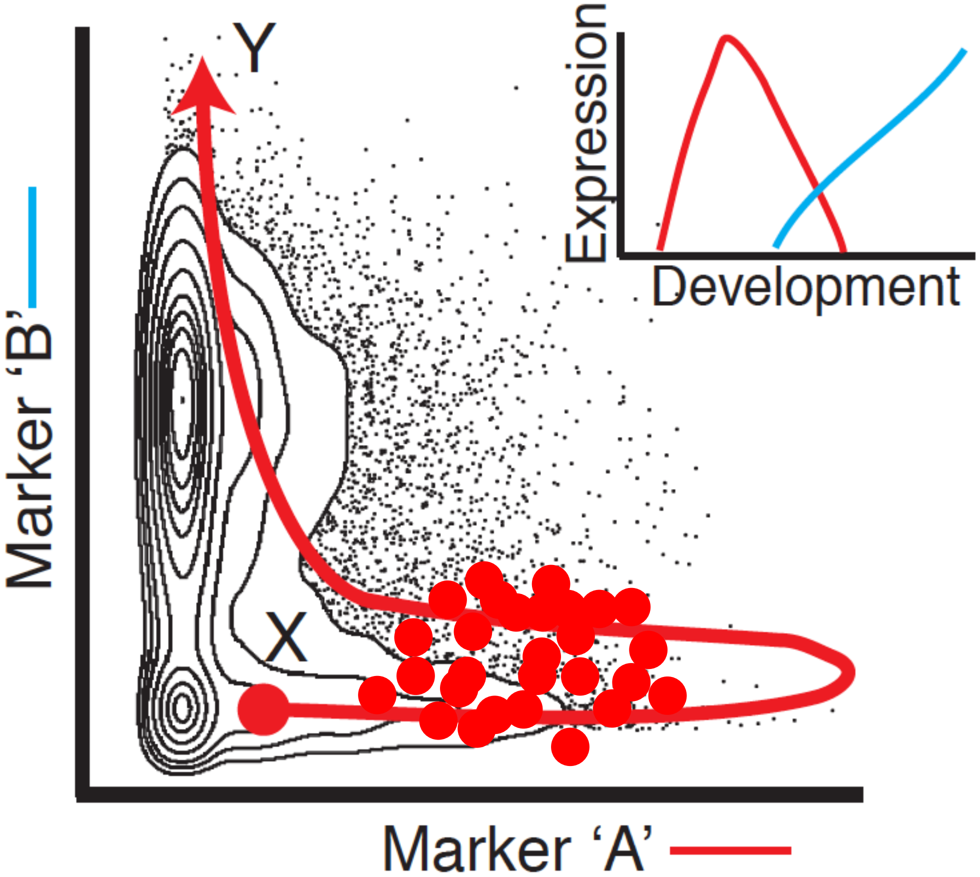
A trajectory approach to development



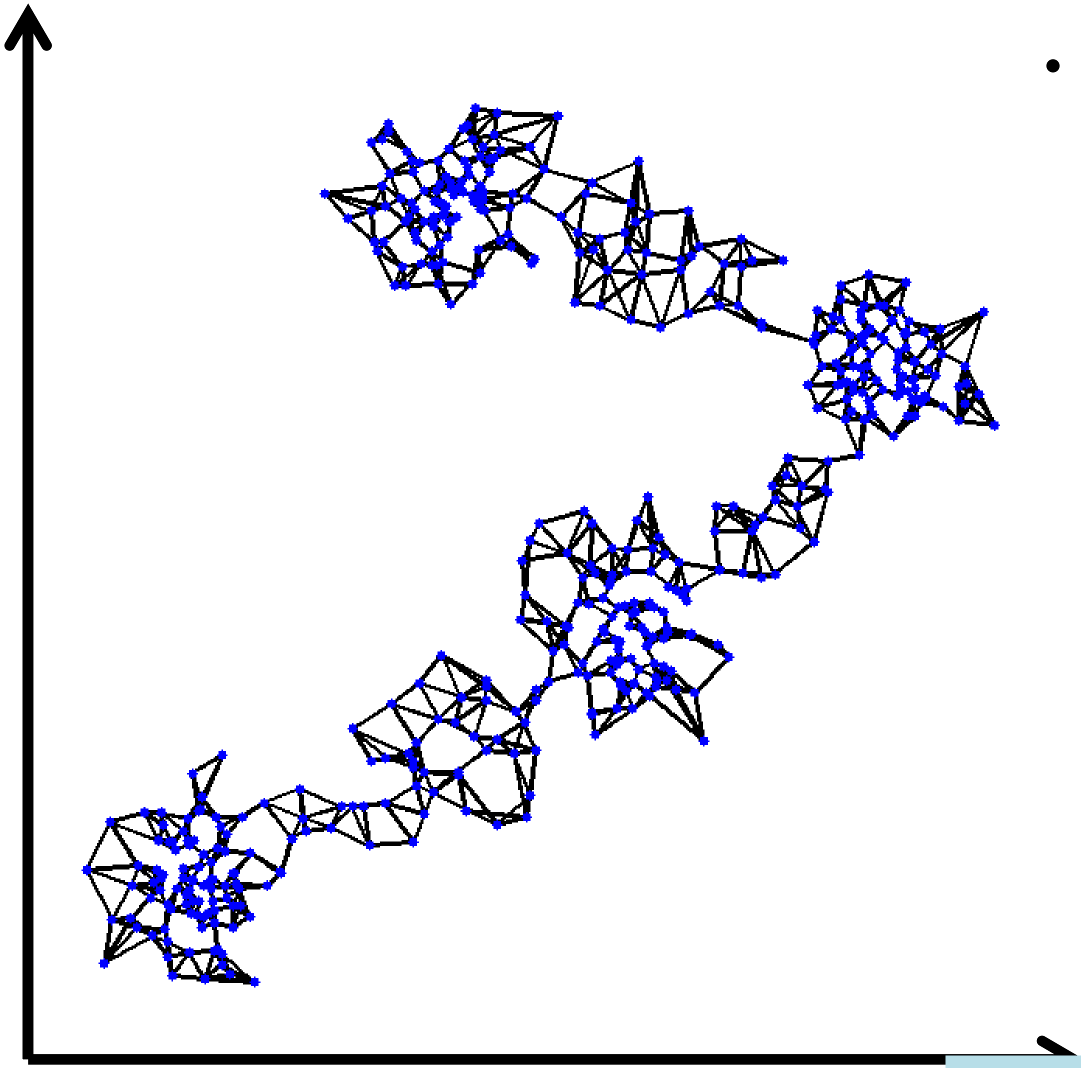
- Single cell studies are finding that sometimes development is a continuous progression
- Strong signal in the data, simple methods get rough approximation, but hard to get accurate progression.

The Challenge: Non-Linearity

- Development is highly non-linear in n-D space
- Euclidian distance is a poor measure for chronological distance



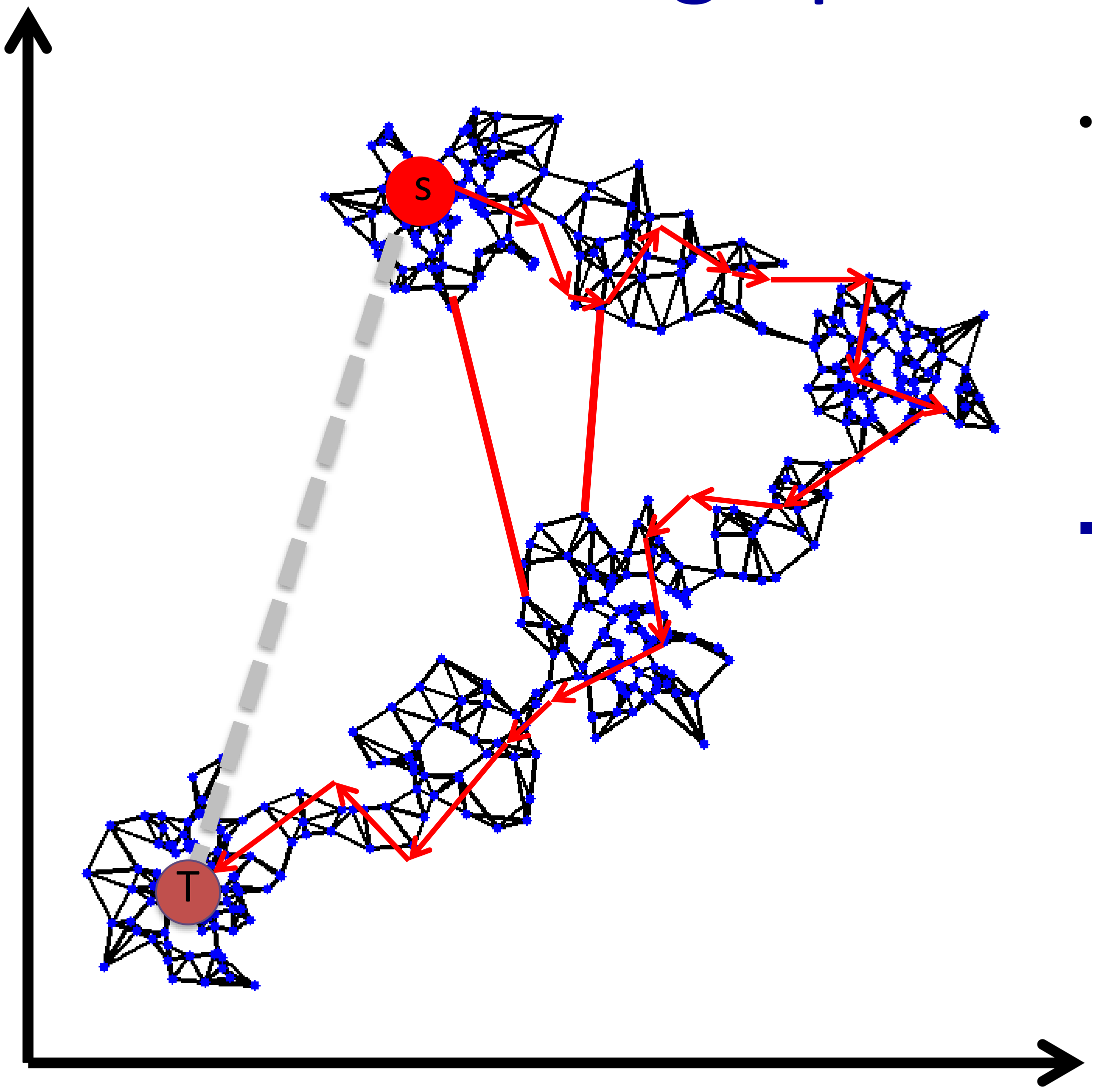
Wanderlust Approach



- Convert data to a k nearest neighbors graph
    - Each cell is a node
    - Each cell only “sees” its local neighborhood

Bendall, Davis*, Amir* et.al. Cell 2014*

Derive Trajectory using “graph walk”



- What is the position of a cell along the trajectory?
    - Start from an early cell
    - Define distance by walking along graph

- But, very noisy data, many additional tricks needed.

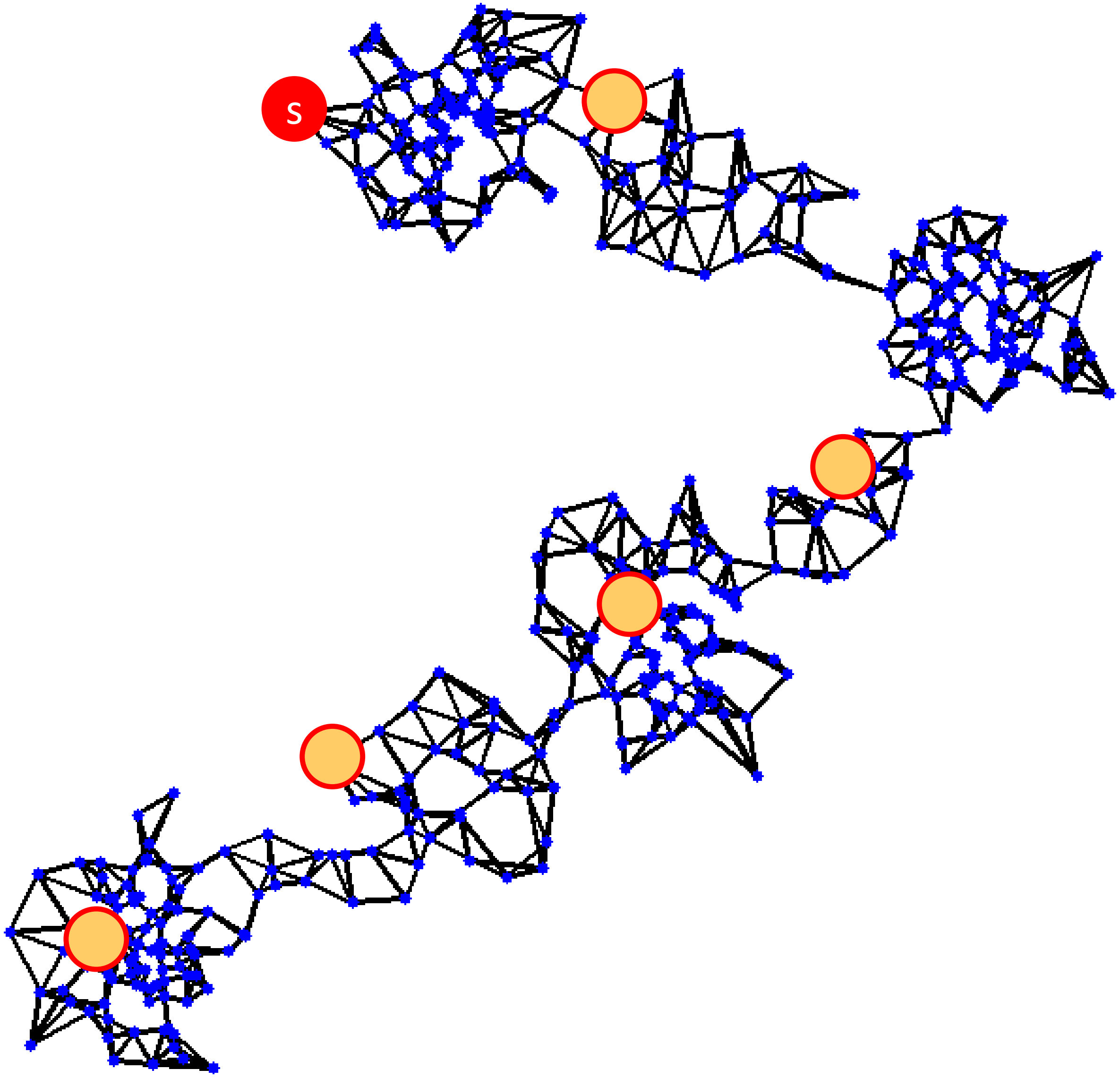
Wanderlust

A graph based trajectory detection algorithm. Wanderlust is scalable, robust and resistant to noise

We use randomness to overcome noise!

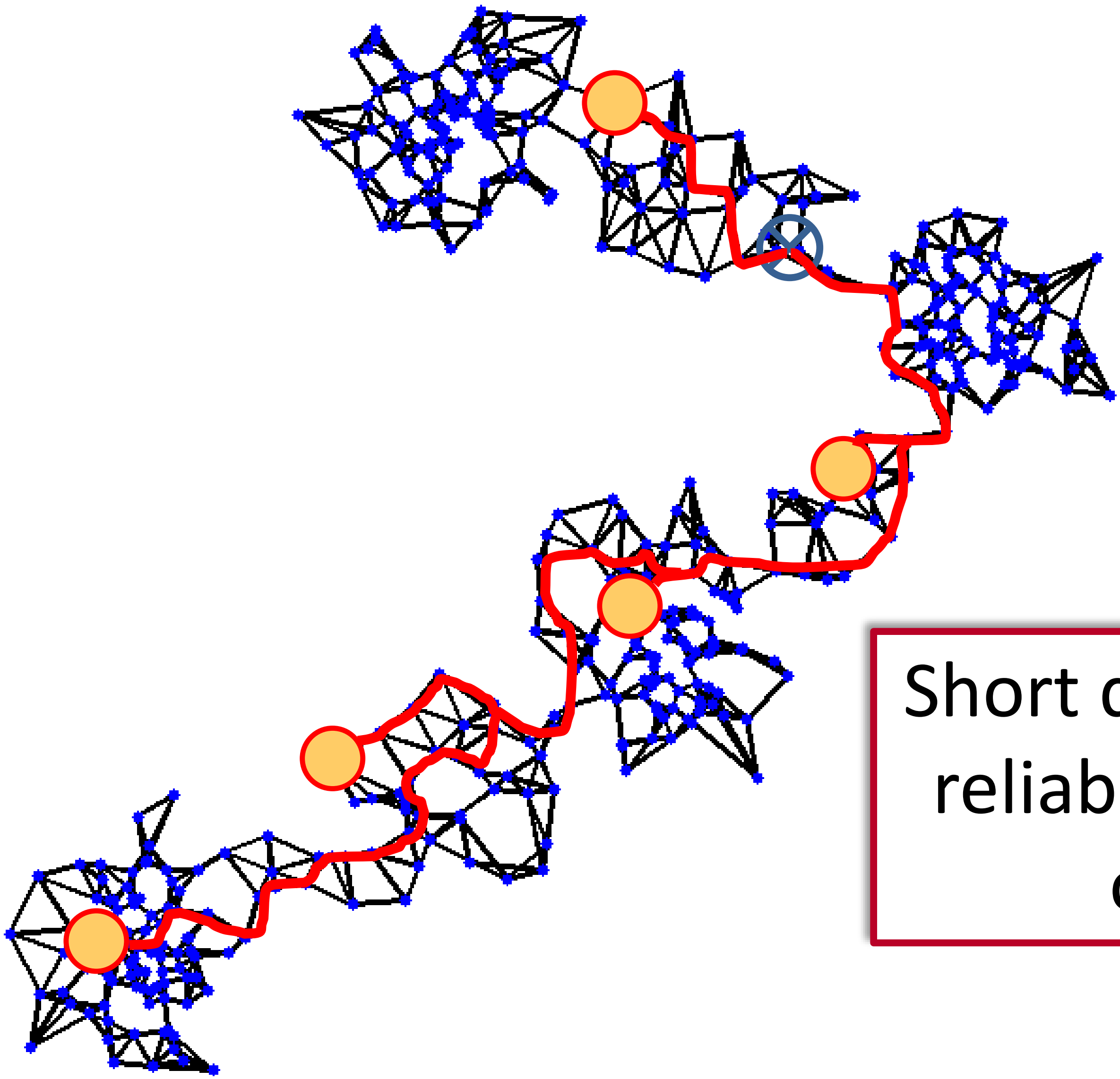
1. Convert data into a set of klNN graphs
2. In each graph, iteratively refine a trajectory using a set of random waypoints
3. The solution trajectory is the average over all graph trajectories

Refine distances using waypoints



Choose M random waypoints, $l_1 \ldots l_M$

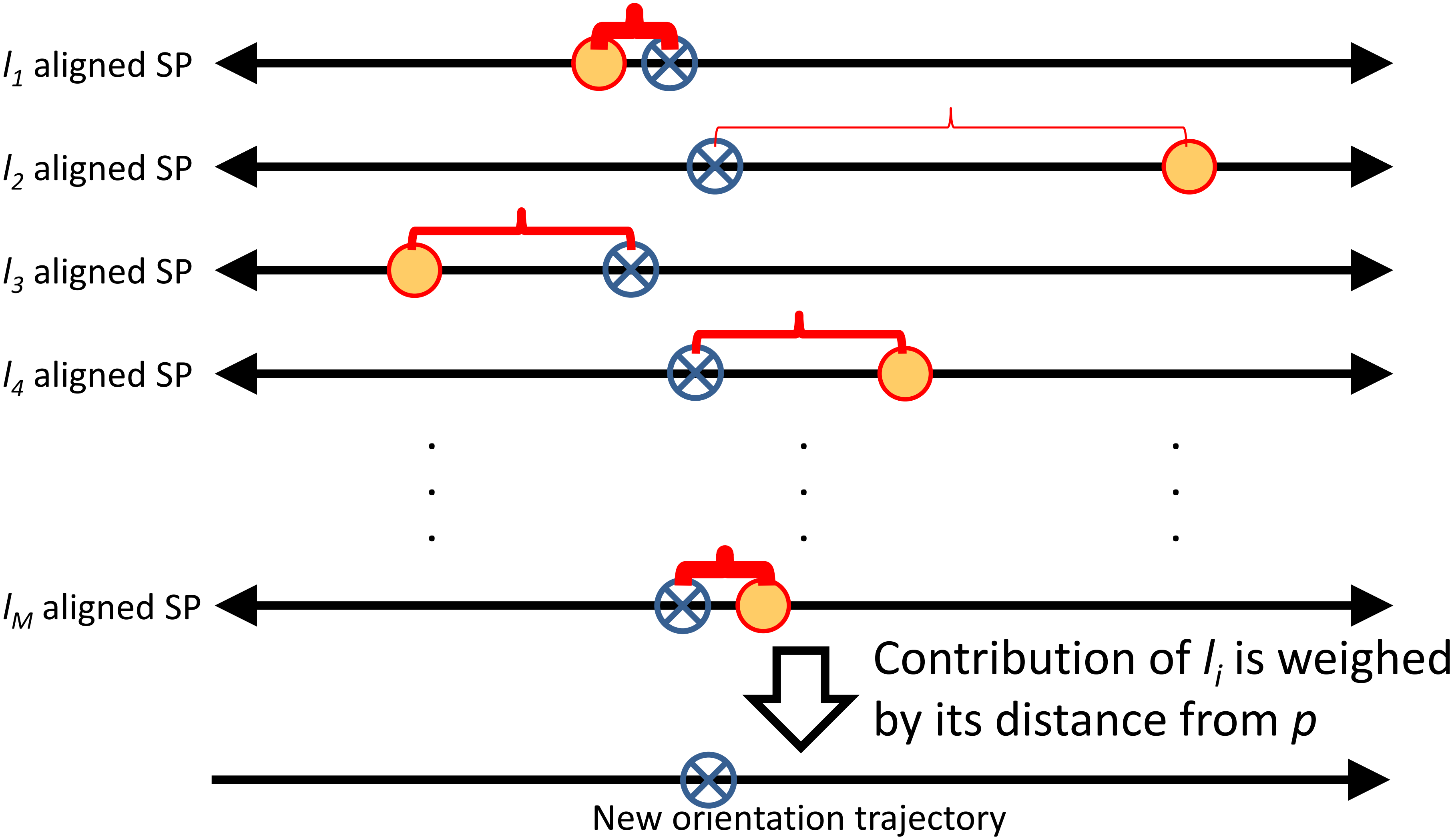
Refine distances using waypoints



Next, find the shortest path from each waypoint l_i to *n*

Short distances are more reliable and help refine order locally

Refine distances using waypoints



l_1 aligned SP

l_2 aligned SP

l_3 aligned SP

l_4 aligned SP

l_M aligned SP

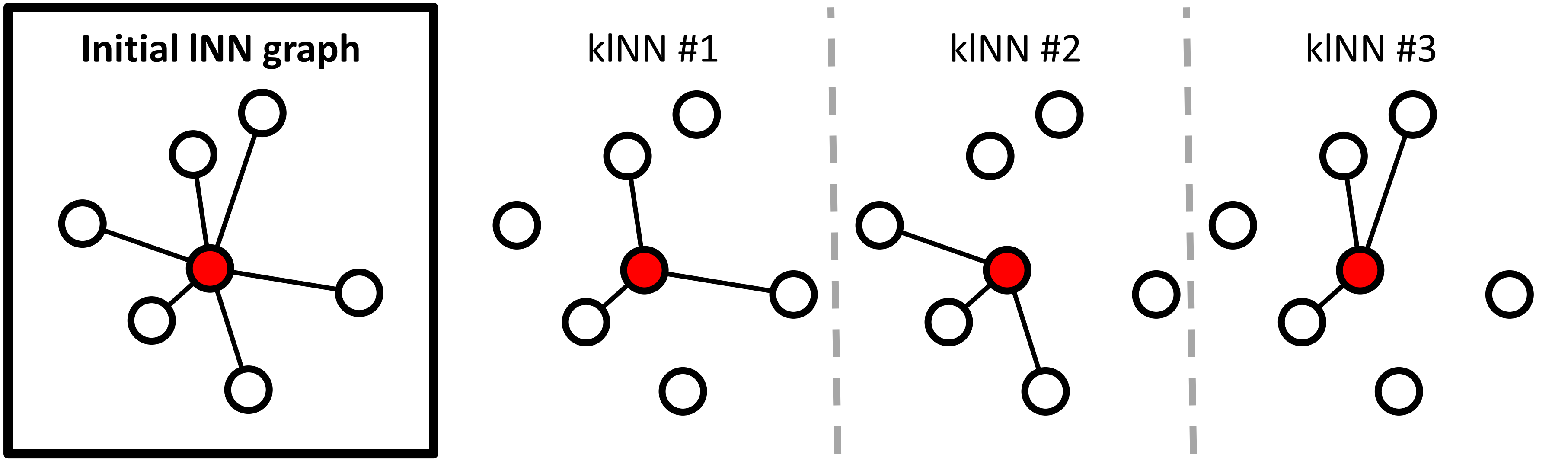
Contribution of *l_i* is weighed by its distance from *p*

New orientation trajectory

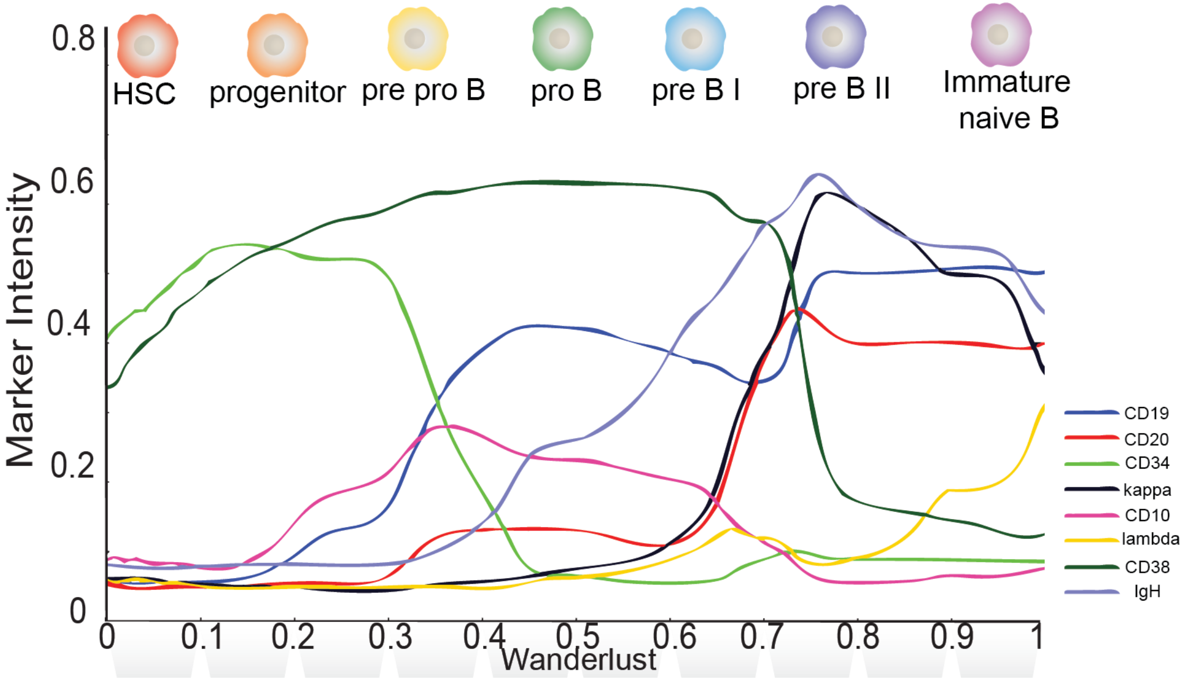
klNN graph

- klNN: k-out-of-l nearest neighbors
- Generate l nearest neighbors graph

Each shortcut appears in only a small number of klNN-graphs



Wanderlust Trajectory



- Wanderlust infers path from Hematopoietic Stem Cells to immature B cells from a single sample of human bone marrow.
- Matches prior knowledge, robust and reproducible across 7 individuals.
- Identified and validated 3 novel rare progenitor states (0.007% of cells)

Acknowledgements

Smita Krishnaswamy

Roshan Sharma

David van Dijk

Ambrose Carr

Linas Mazutis

Josh Nainys

Oren Litvin

El-ad David Amir

Michelle Tadmor

Jacob Levine

Manu Setty

Bodenmiller Lab (U Zurich)

Bernd Bodenmiller

Nevana Zivanovic

Nolan Lab (Stanford)

Garry Nolan

Sean Bendall

Matt Spitzer

Kara Davis

Erin Simons

Tiffany Chen

S↑2C™

NIH DIRECTOR'S PIONEER AWARD

NATIONAL CANCER INSTITUTE®

COLUMBIA SYSTEMS BIOLOGY

the David & Lucile Packard FOUNDATION

NSF

NIH DIRECTOR'S NEW INNOVATOR AWARD