

Machine learning in single cell genomics

Fabian J. Theis

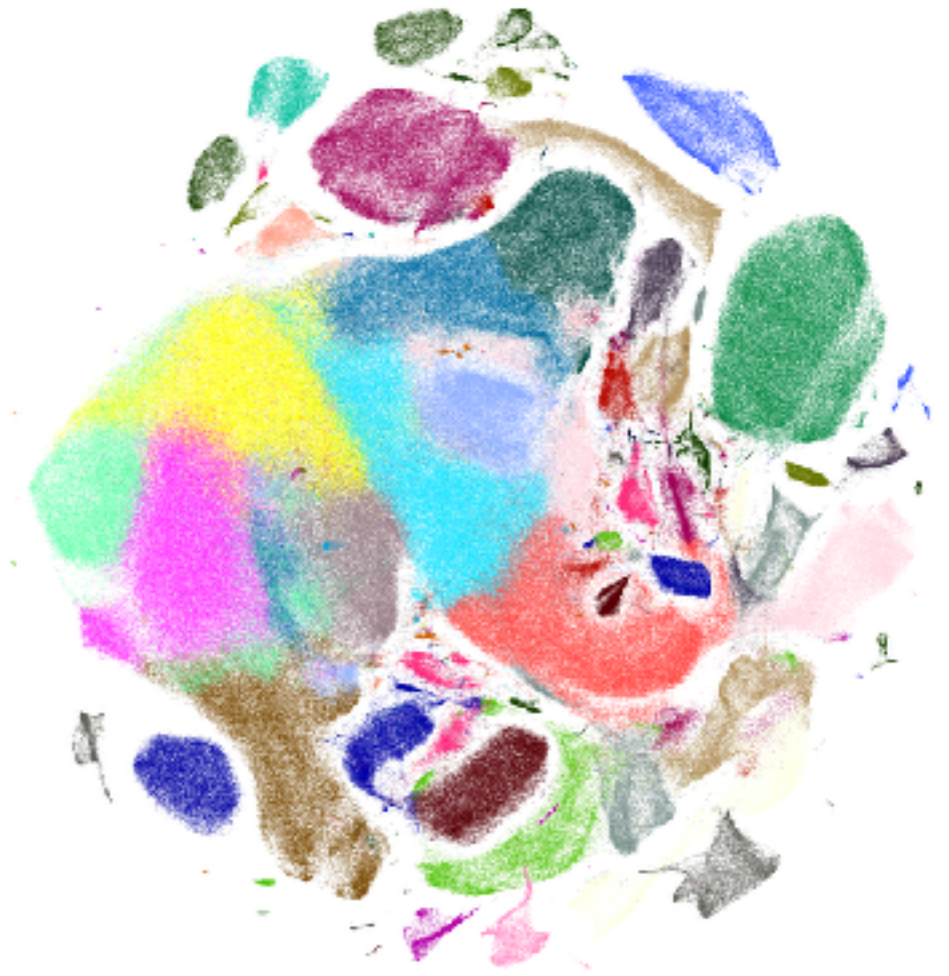
Institute of Computational Biology, Helmholtz Center Munich &
Department of Mathematics, TU Munich



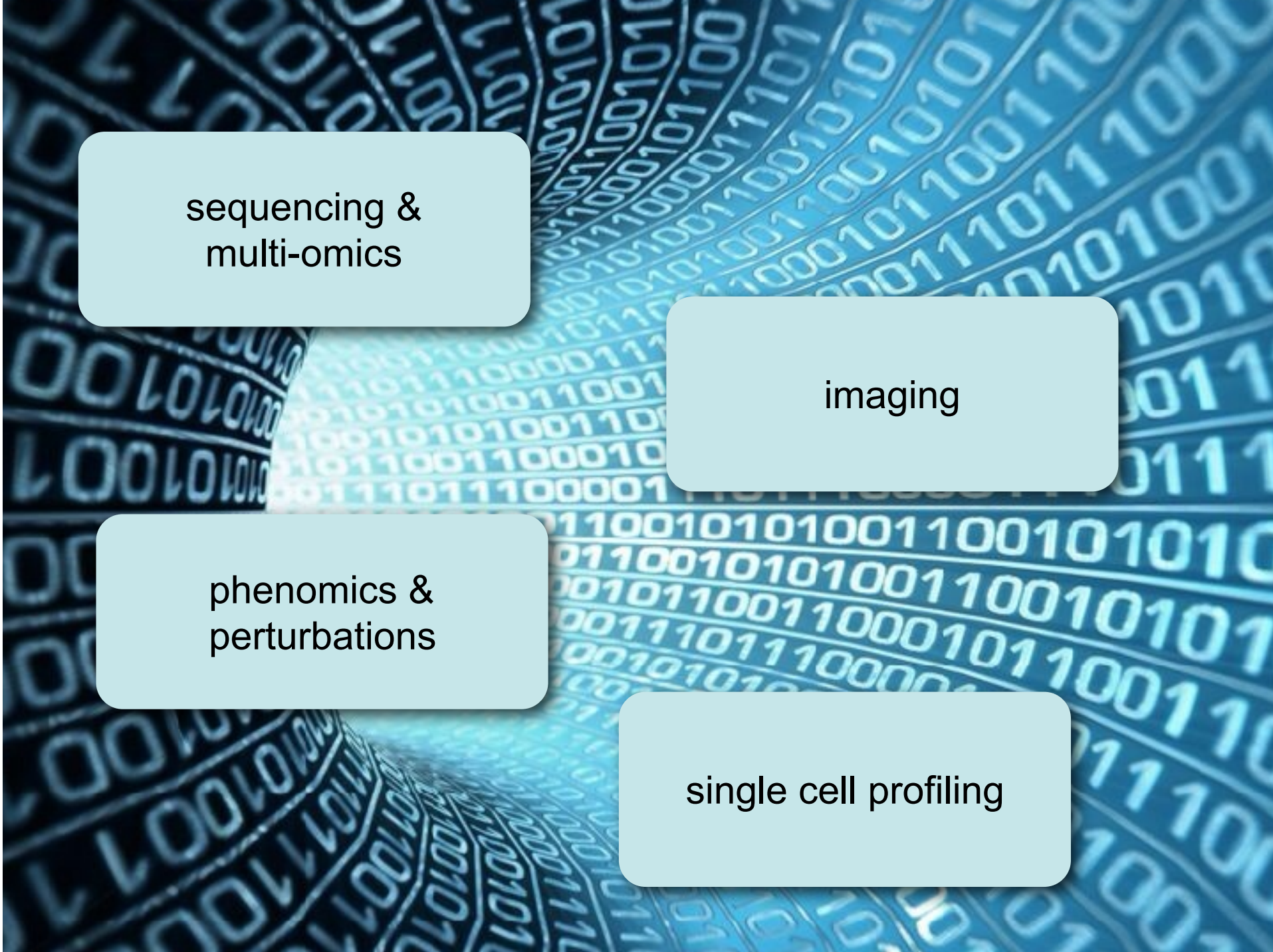
www.comp.bio



[@fabian_theis](https://twitter.com/fabian_theis)



big data in biology & biomedicine



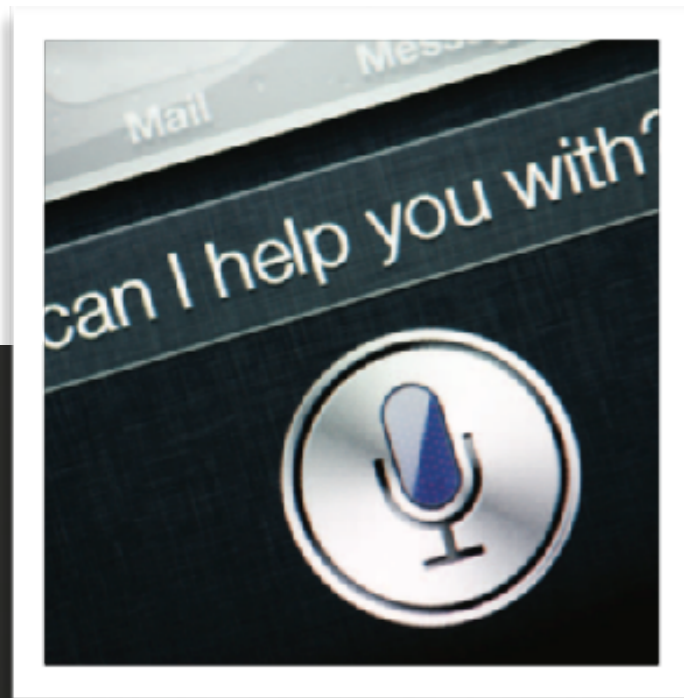
sequencing &
multi-omics

imaging

phenomics &
perturbations

single cell profiling

Big data analytics? → machine learning

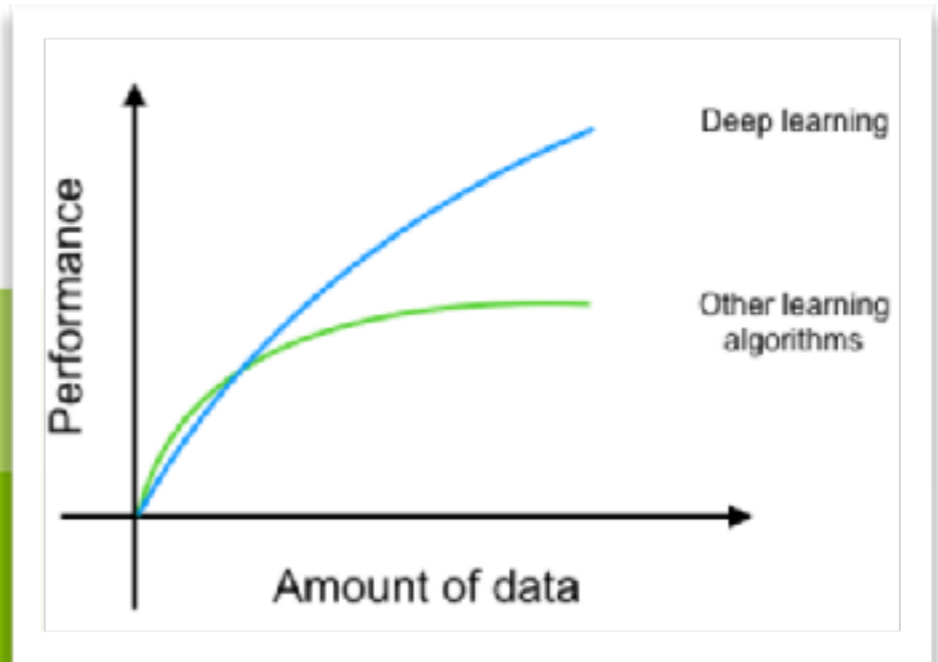
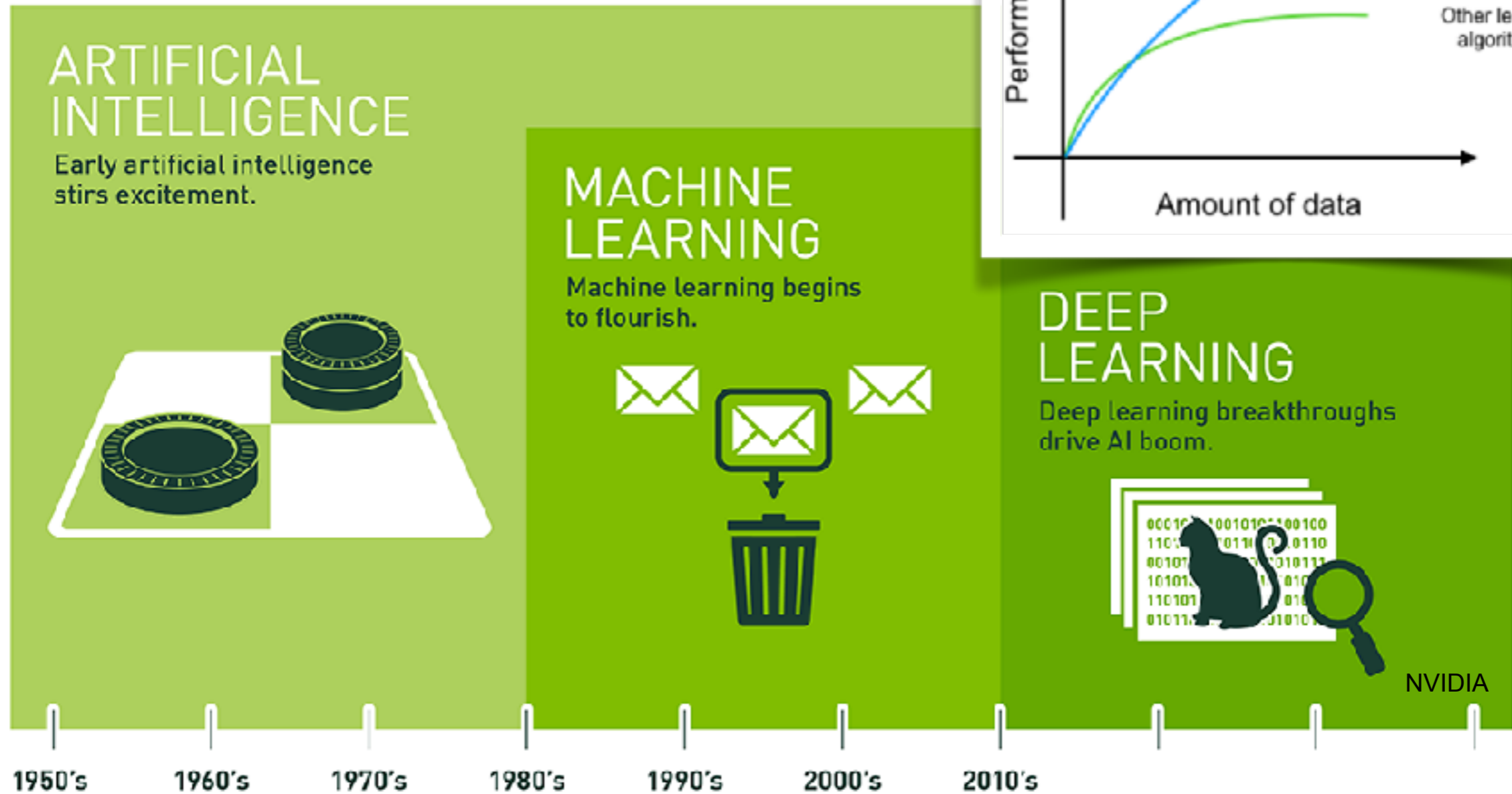


RISE OF MACHINE



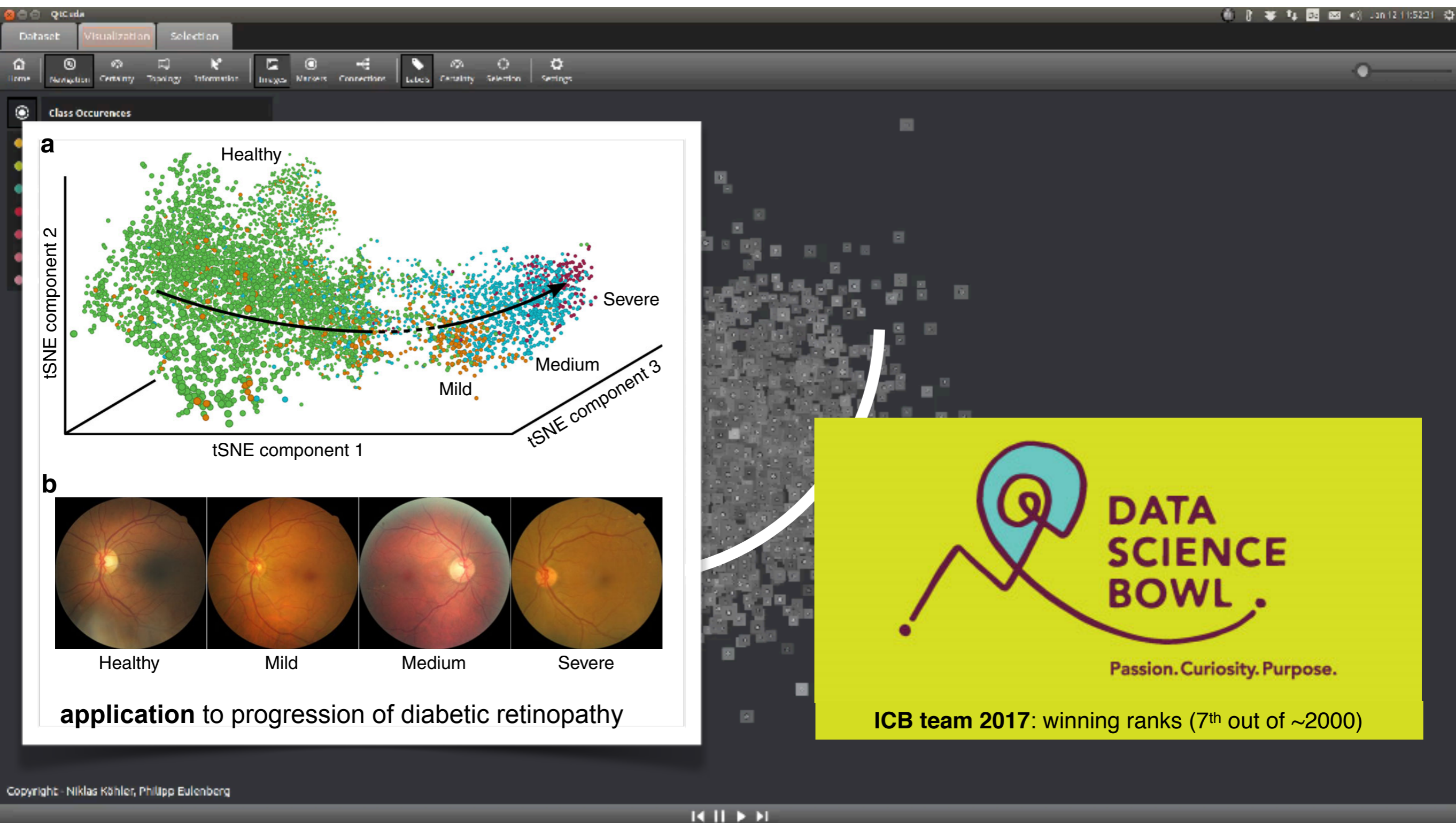
By Jelena Stajic, Richard Stone, Gilbert Chin, and Brad Wible

AI, machine & deep learning



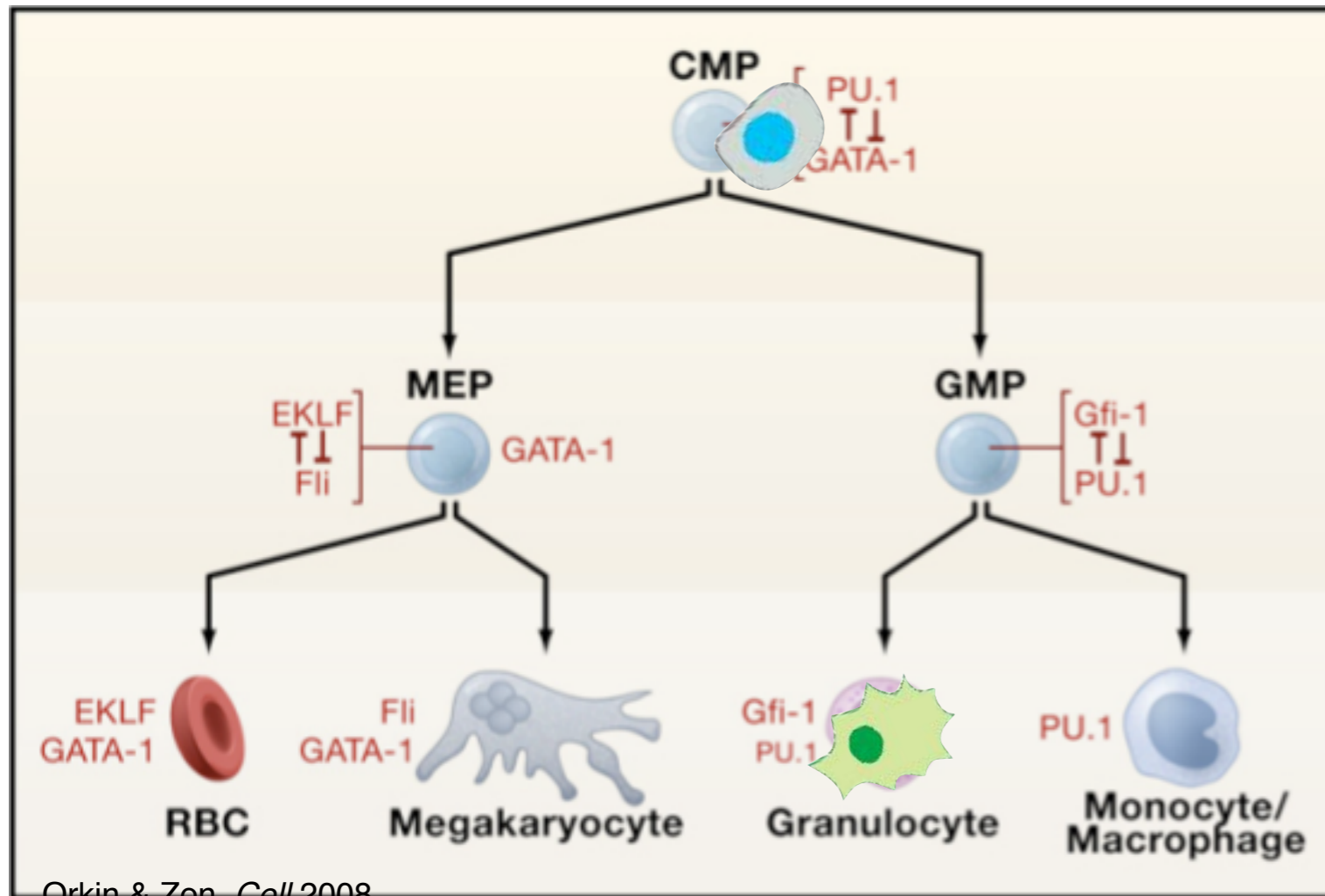
NVIDIA

example: predicting cell cycle from morphometry

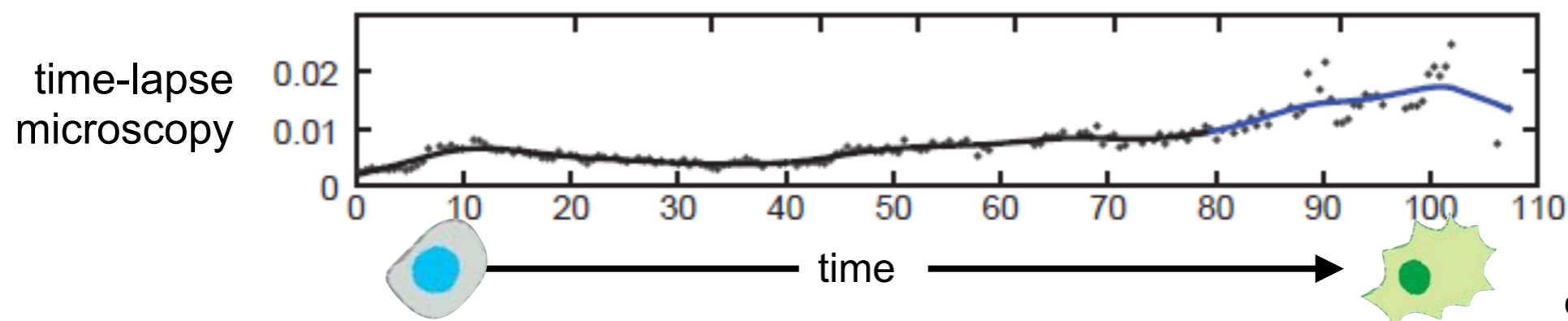
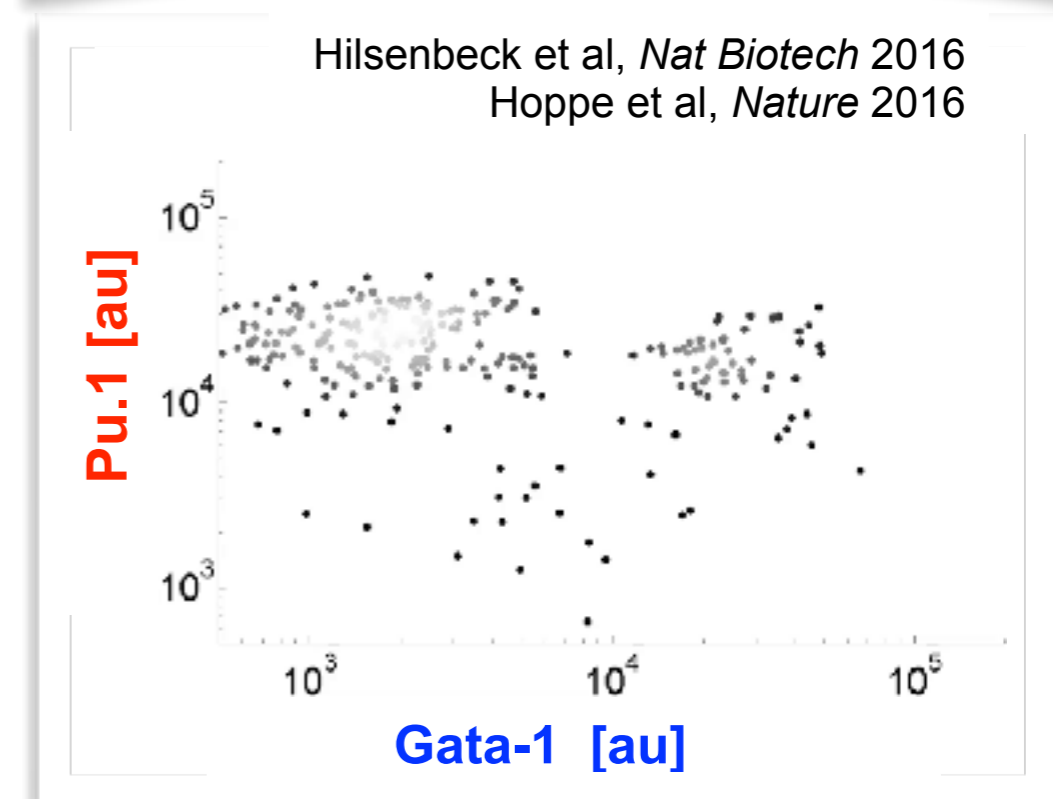
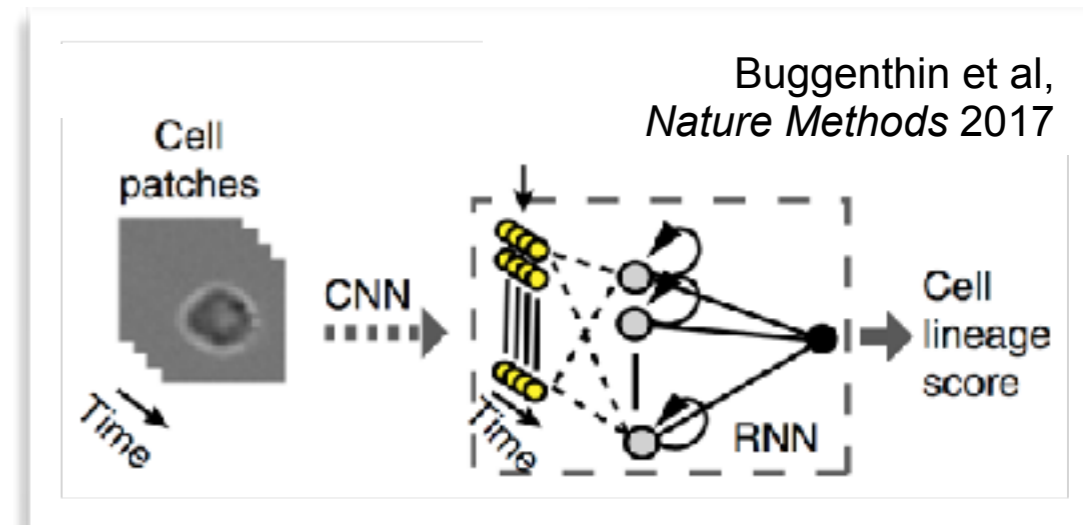


→ mean learning accuracy 98.4% (10-fold CV)

my aim: model single-cell decisions



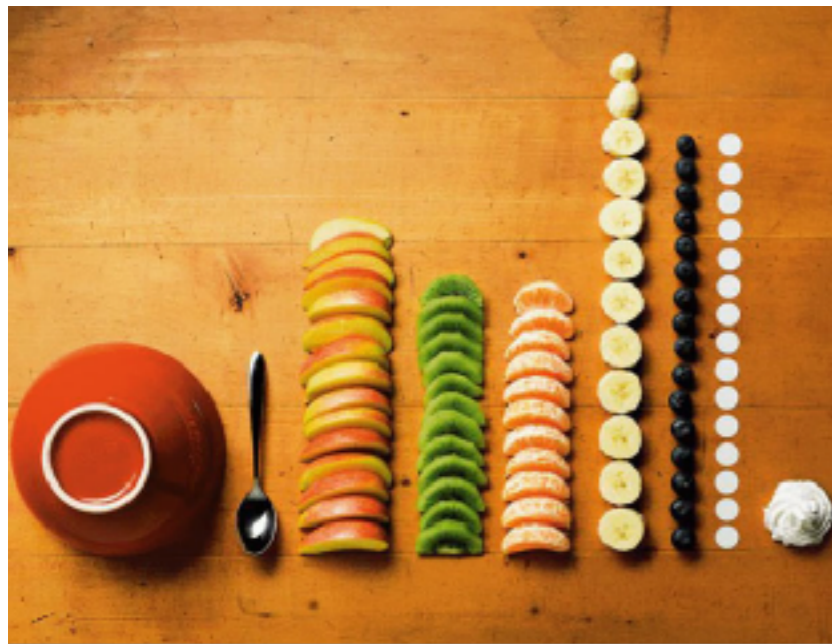
Orkin & Zon, *Cell* 2008



unbiased description of cellular state by transcriptomics

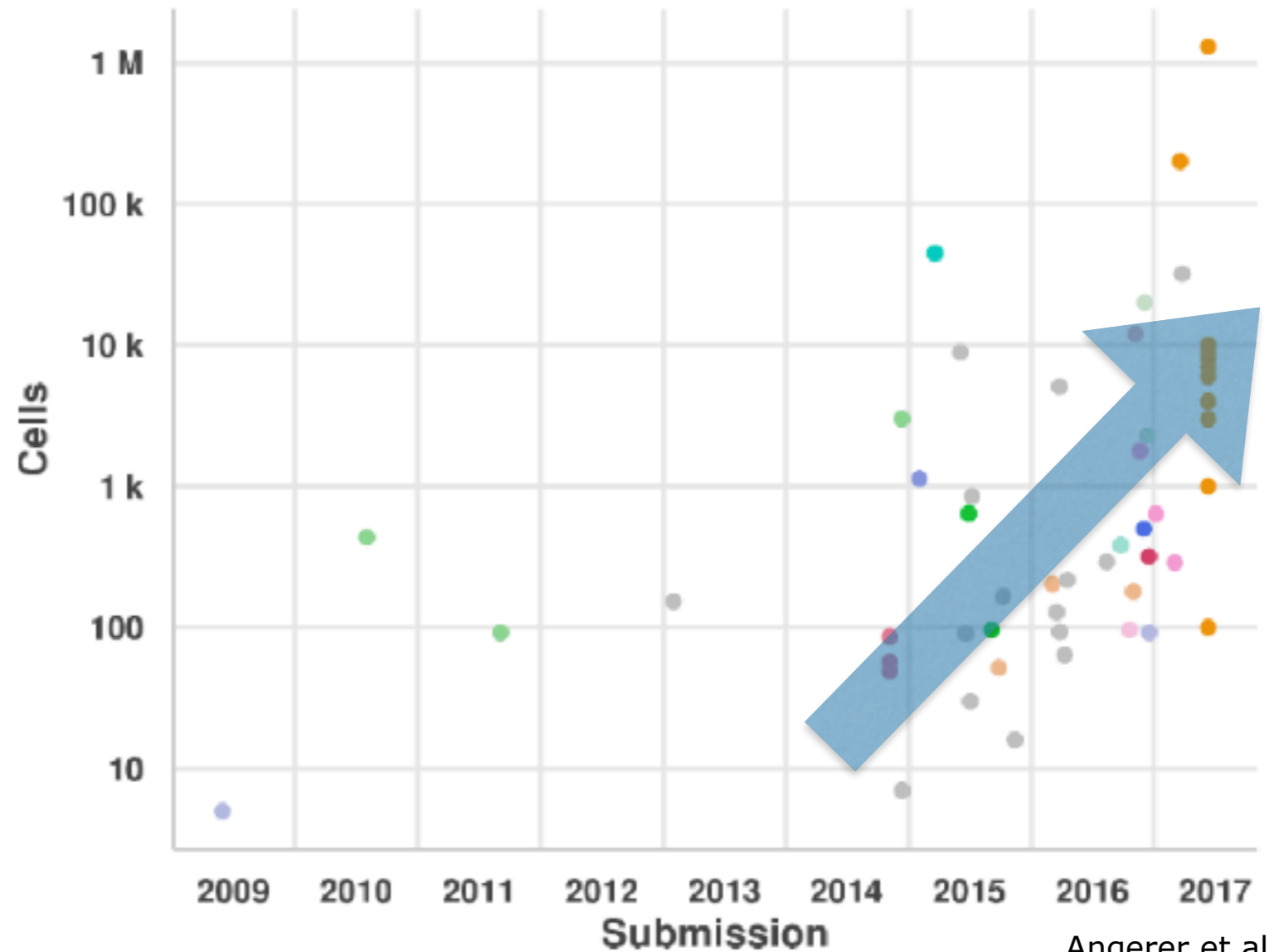


bulk genomics



single-cell genomics

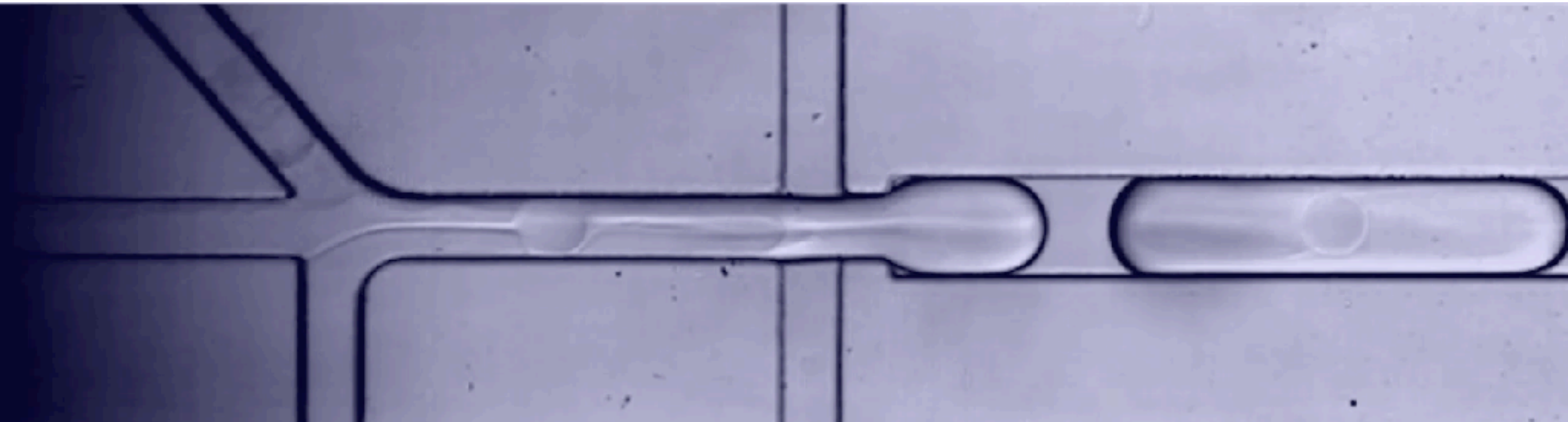
single-cell genomics is becoming big data



Angerer et al,
Curr Op Sys Bio 2017

single-cell transcriptomics

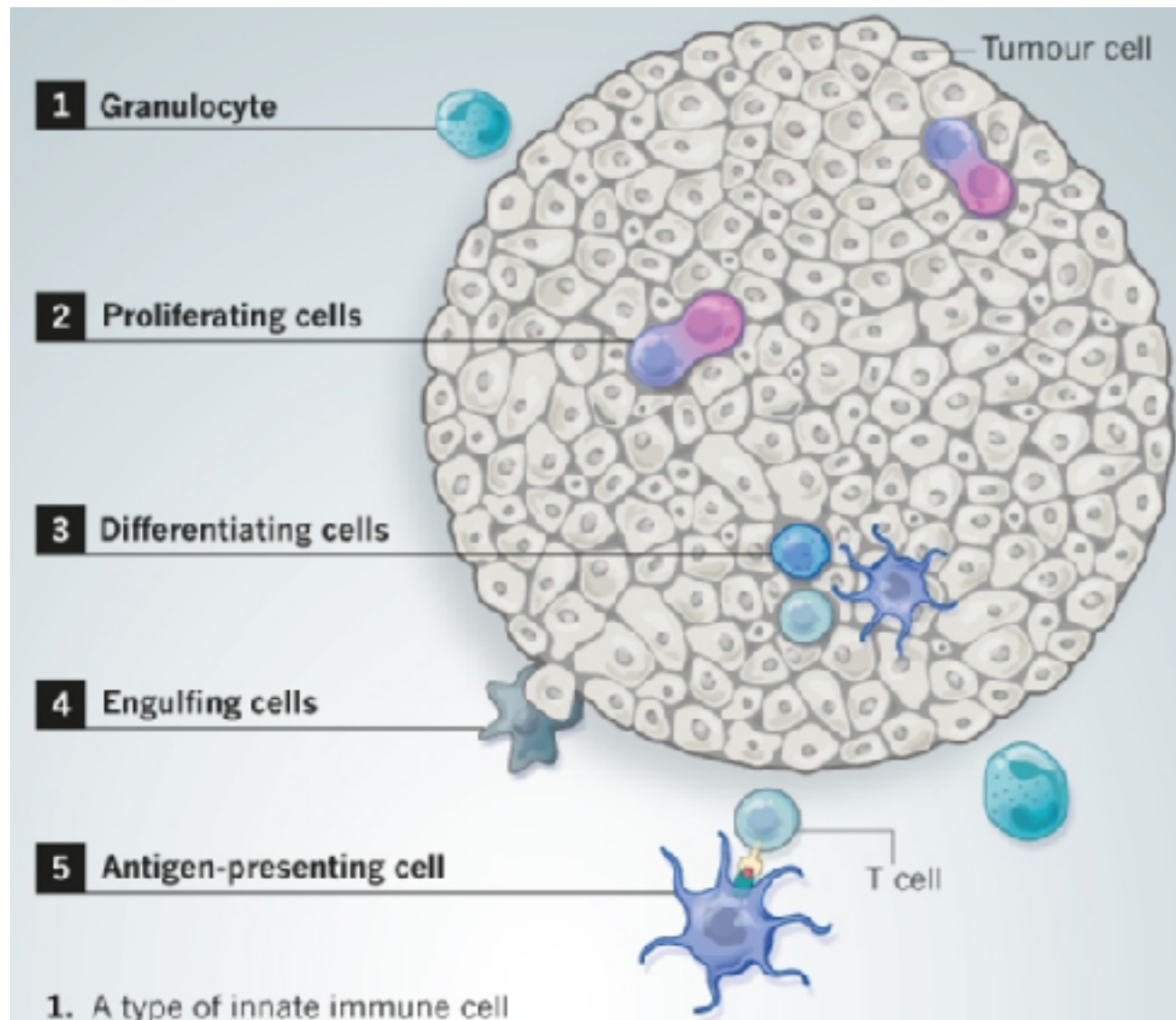
NATURE METHODS 2014
METHOD OF THE YEAR



using droplet microfluidics, isolate single cells and quantify their transcripts

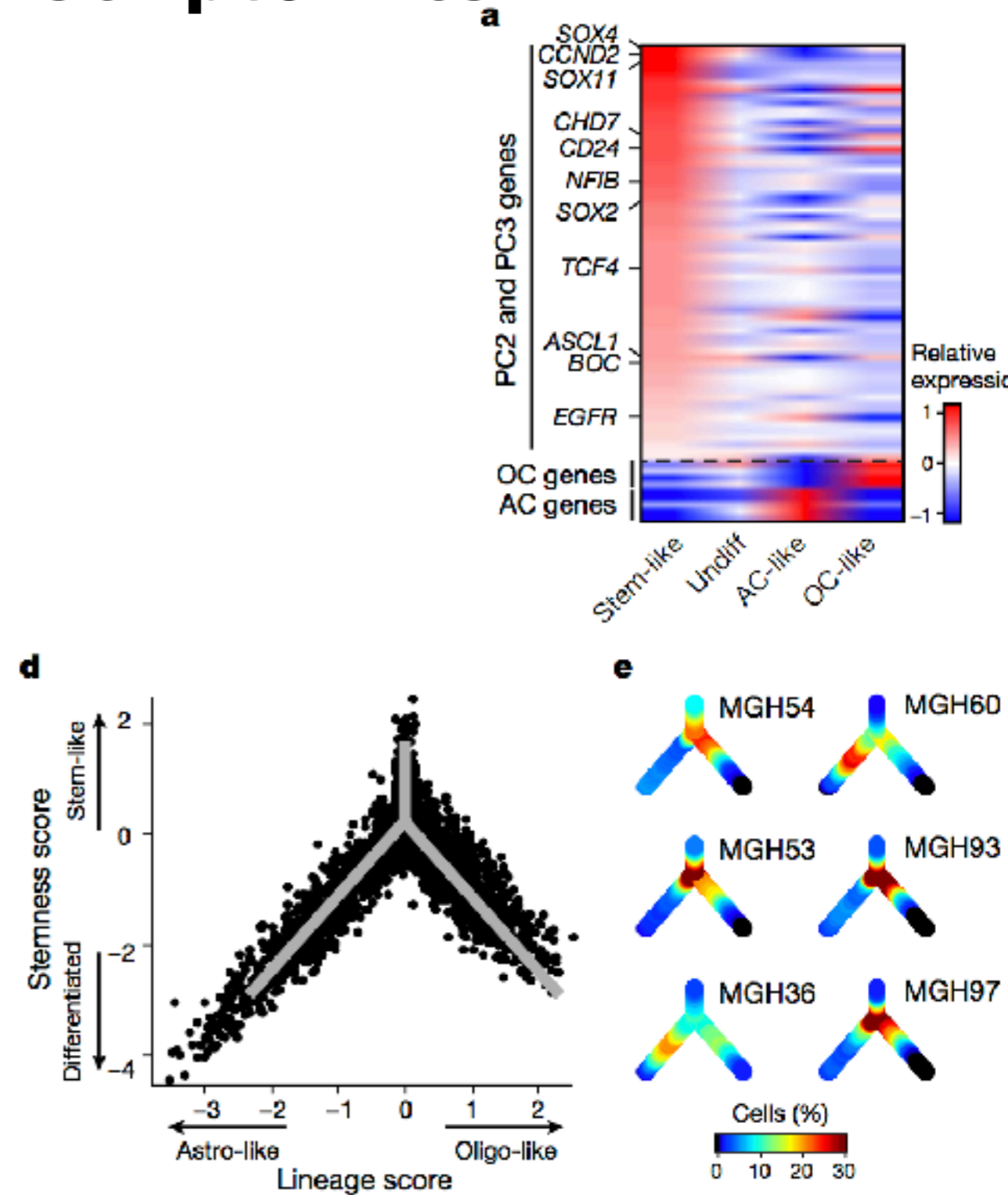
Klein et al., Cell, 2015

Promises of single cell transcriptomics



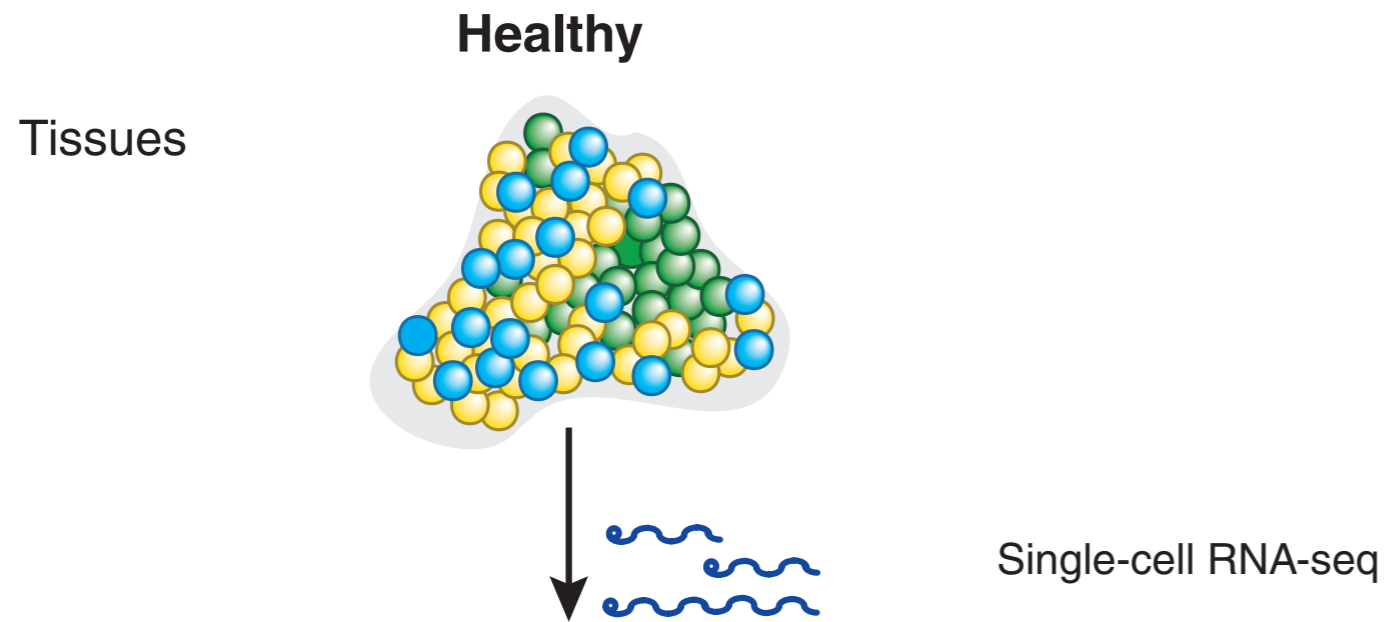
a genetic microscope

Giladi & Amit, Nature 547, 2017

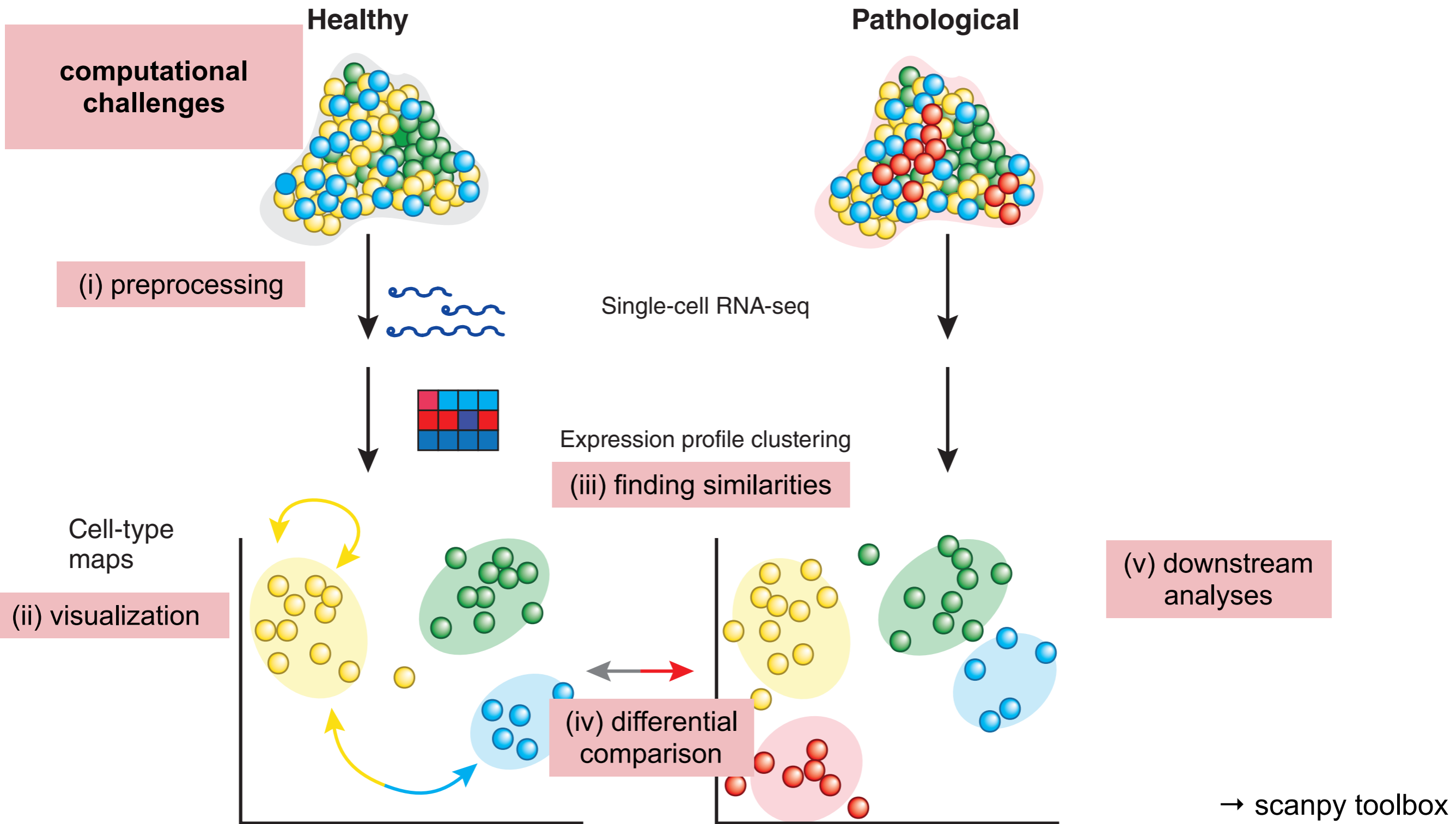


scRNAseq lineages in oligodendrogloma
Tirosh, Venteicher ... Regev, Suva, Nature 2016

single-cell transcriptome analysis



single-cell transcriptome analysis

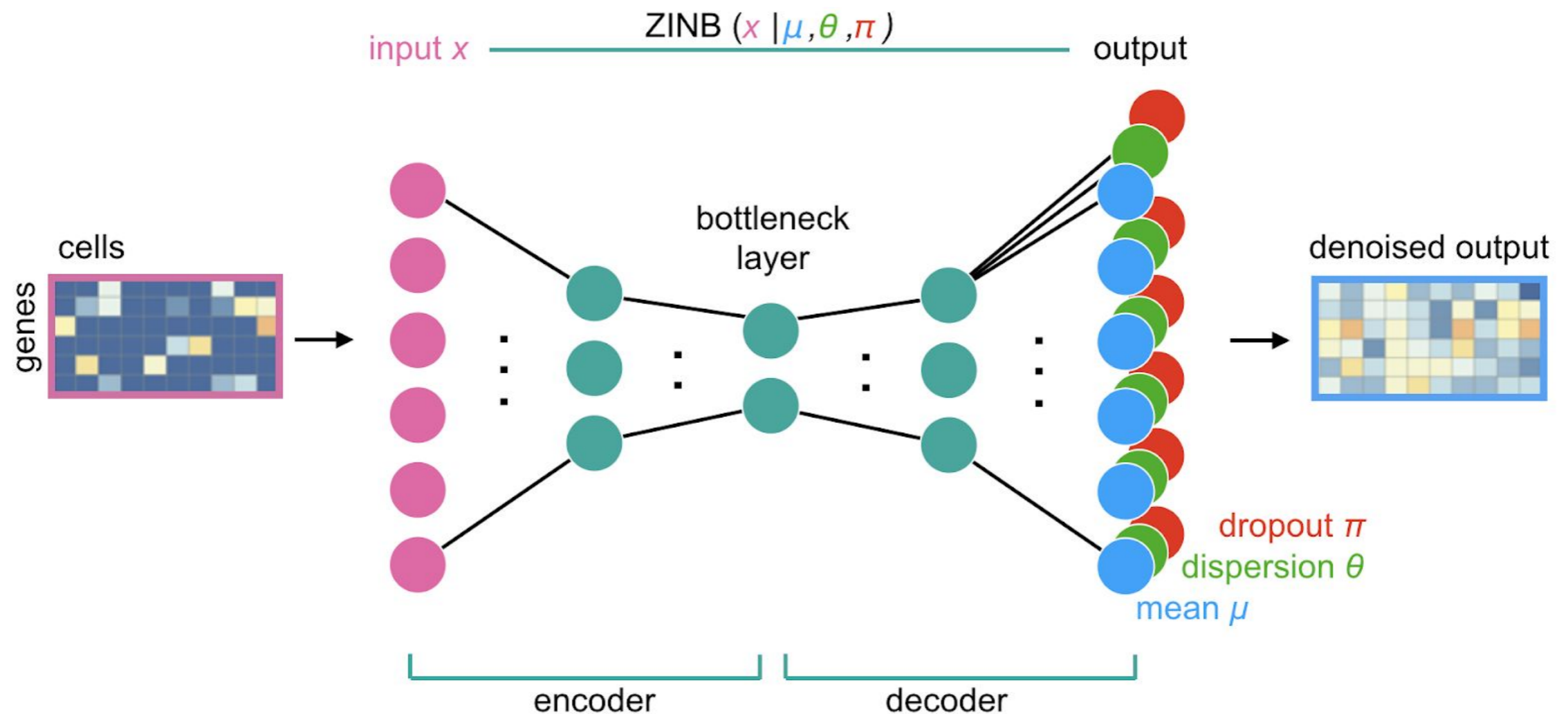


→ scanpy toolbox
github.com/theislab/scanpy

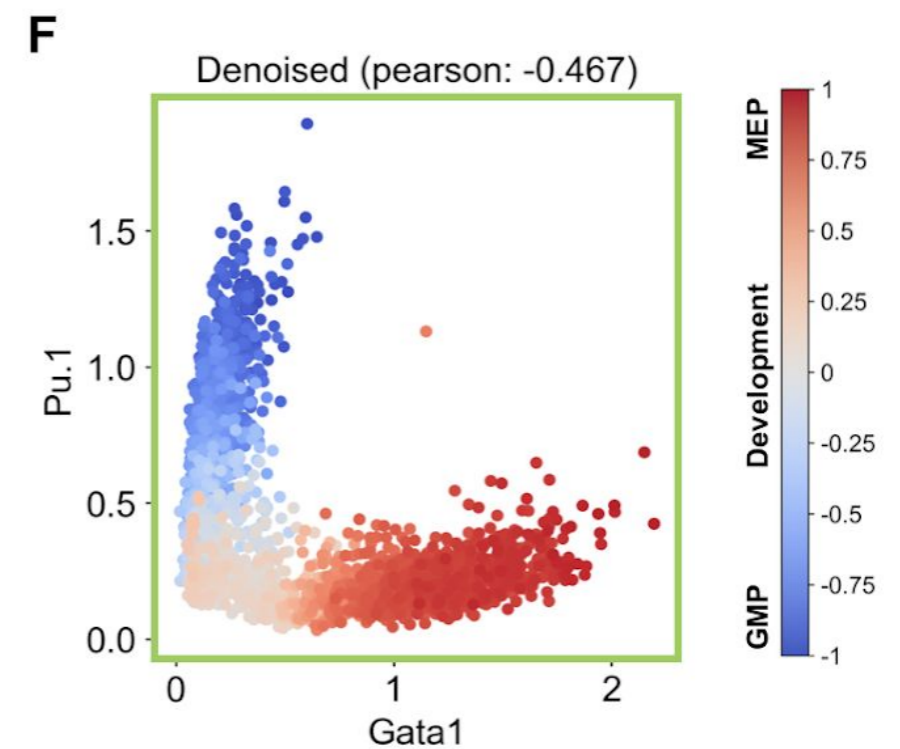
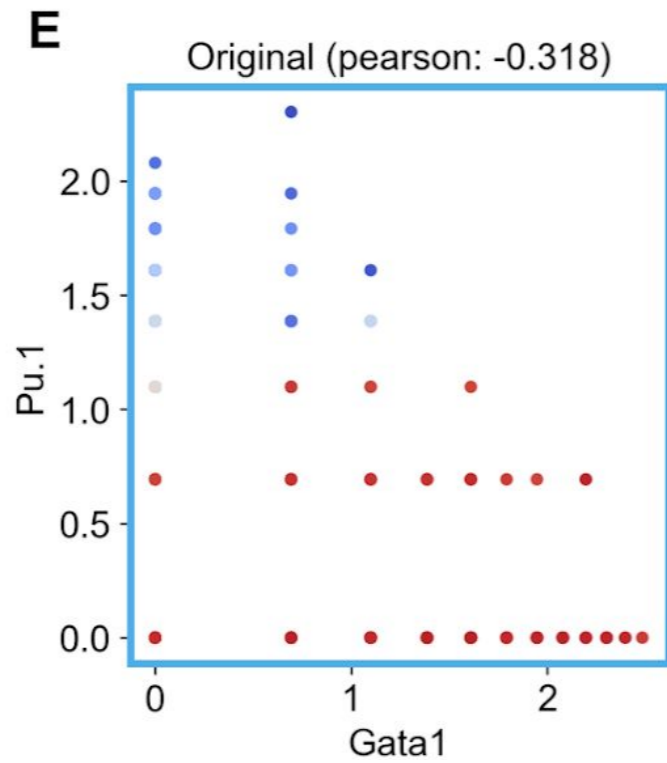
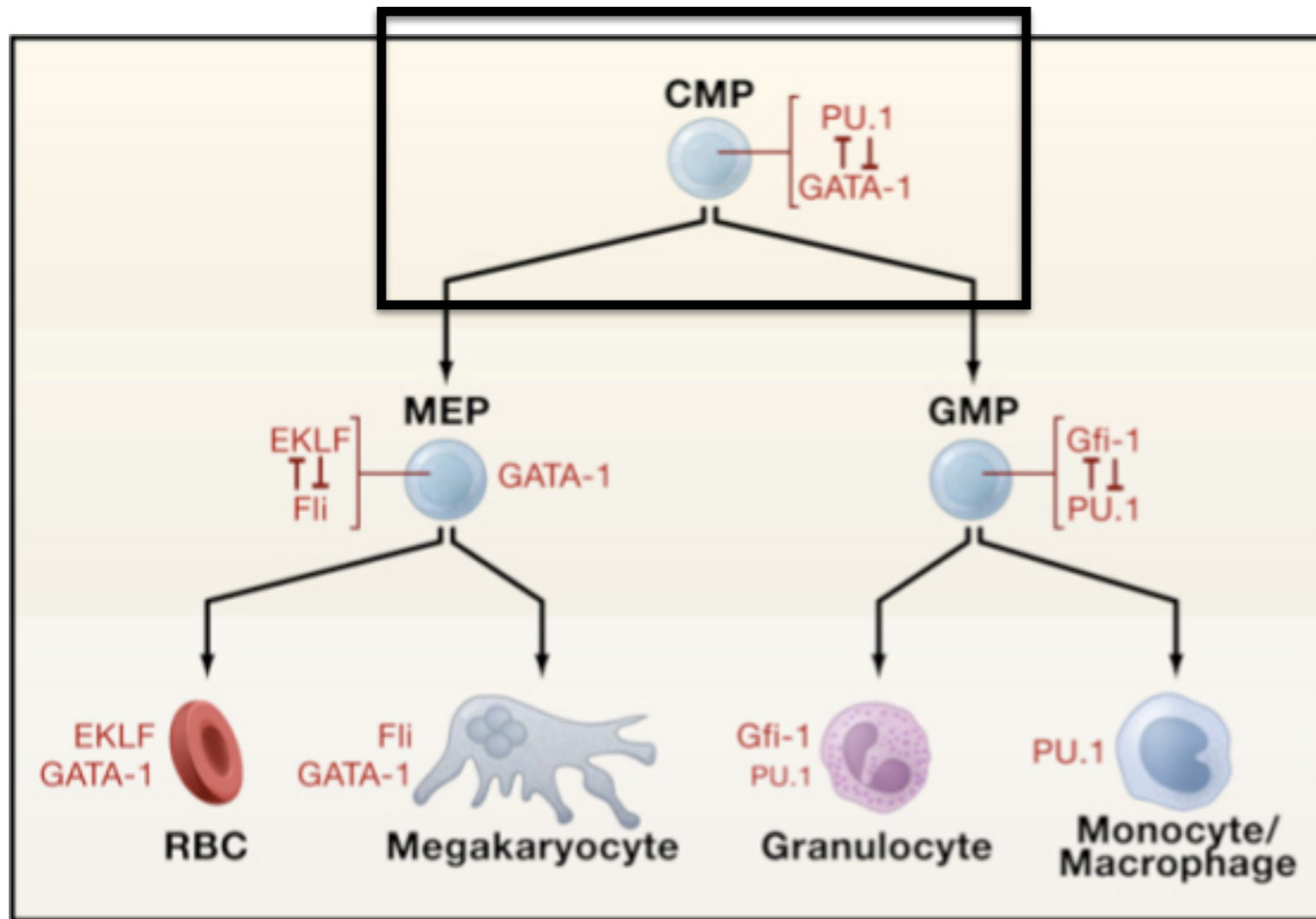
preprocessing: scRNAseq denoising using a deep count autoencoder

idea: replace MSE cost function by adapted ZINB loss

implementation: <https://github.com/theislab/dca>



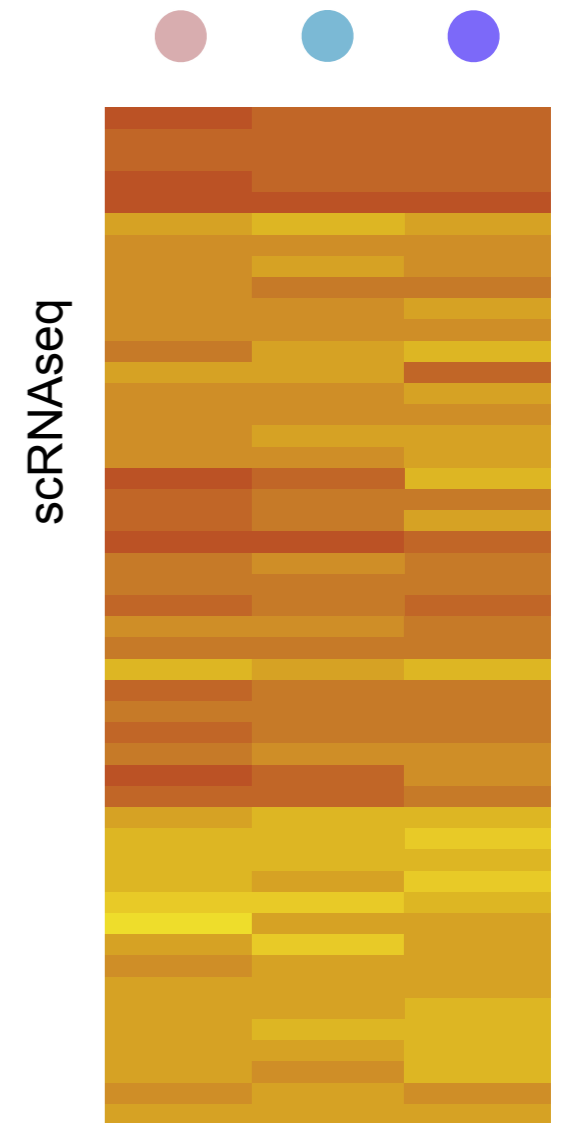
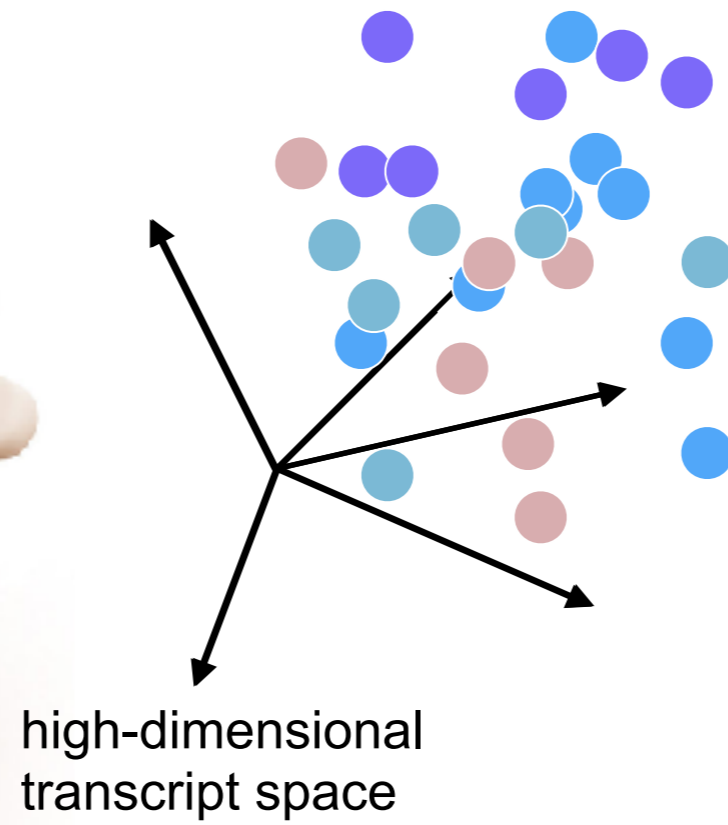
DCA increases correlation structure of key regulatory genes



visualizing high-dimensional single cell RNA-seq

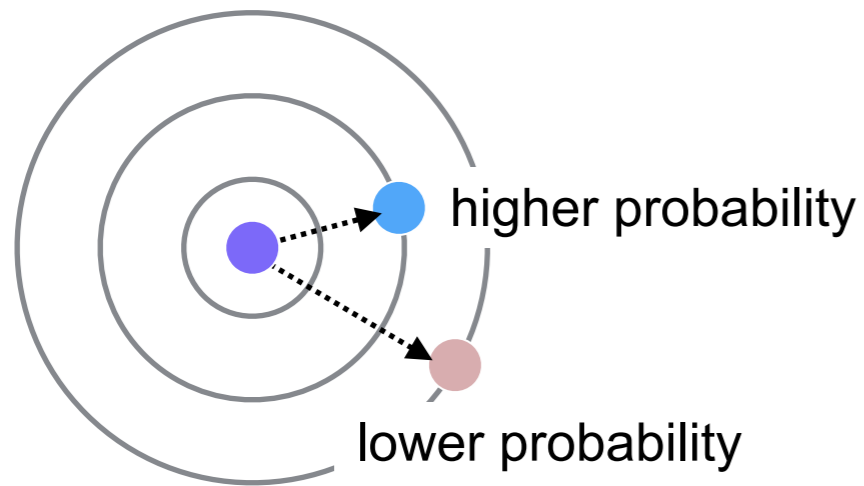
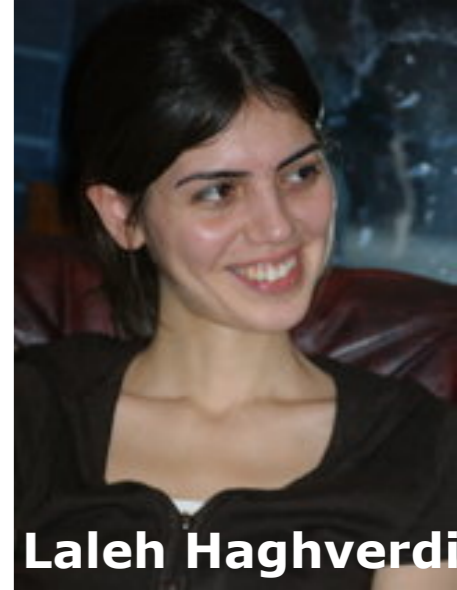


visualizing high-dimensional single cell RNA-seq

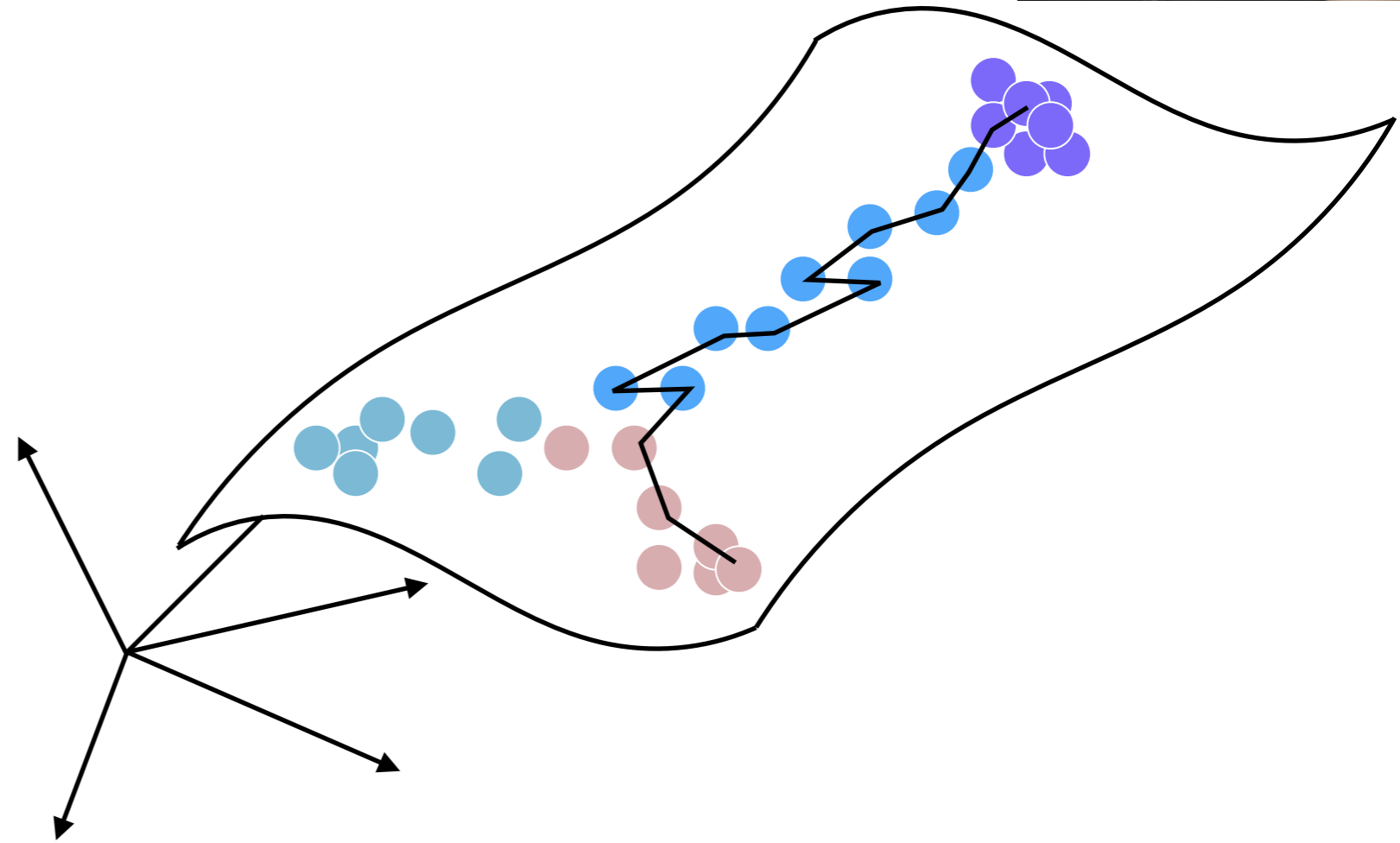


Single-cell diffusion maps

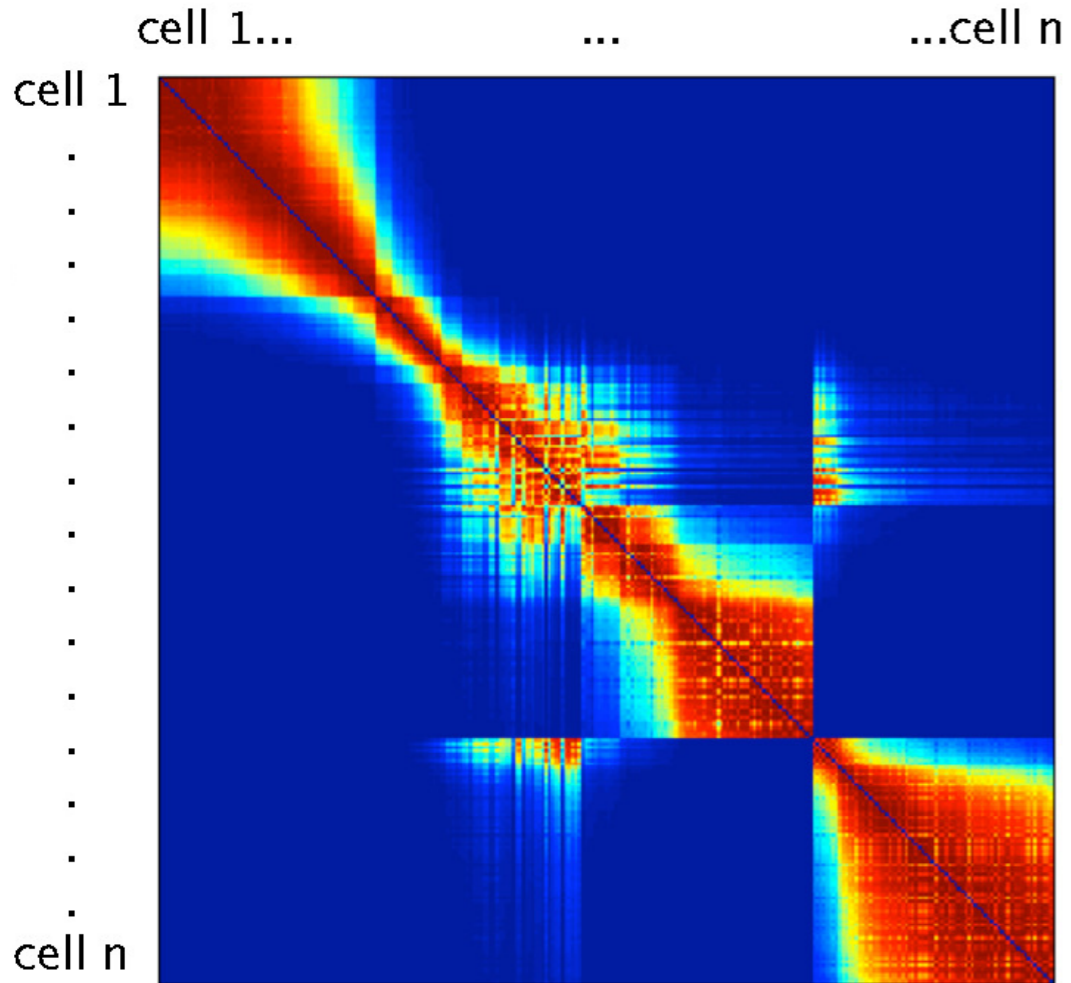
approach: visualize cellular dynamics by analyzing random walks between close-by cells



local diffusion of each cell x

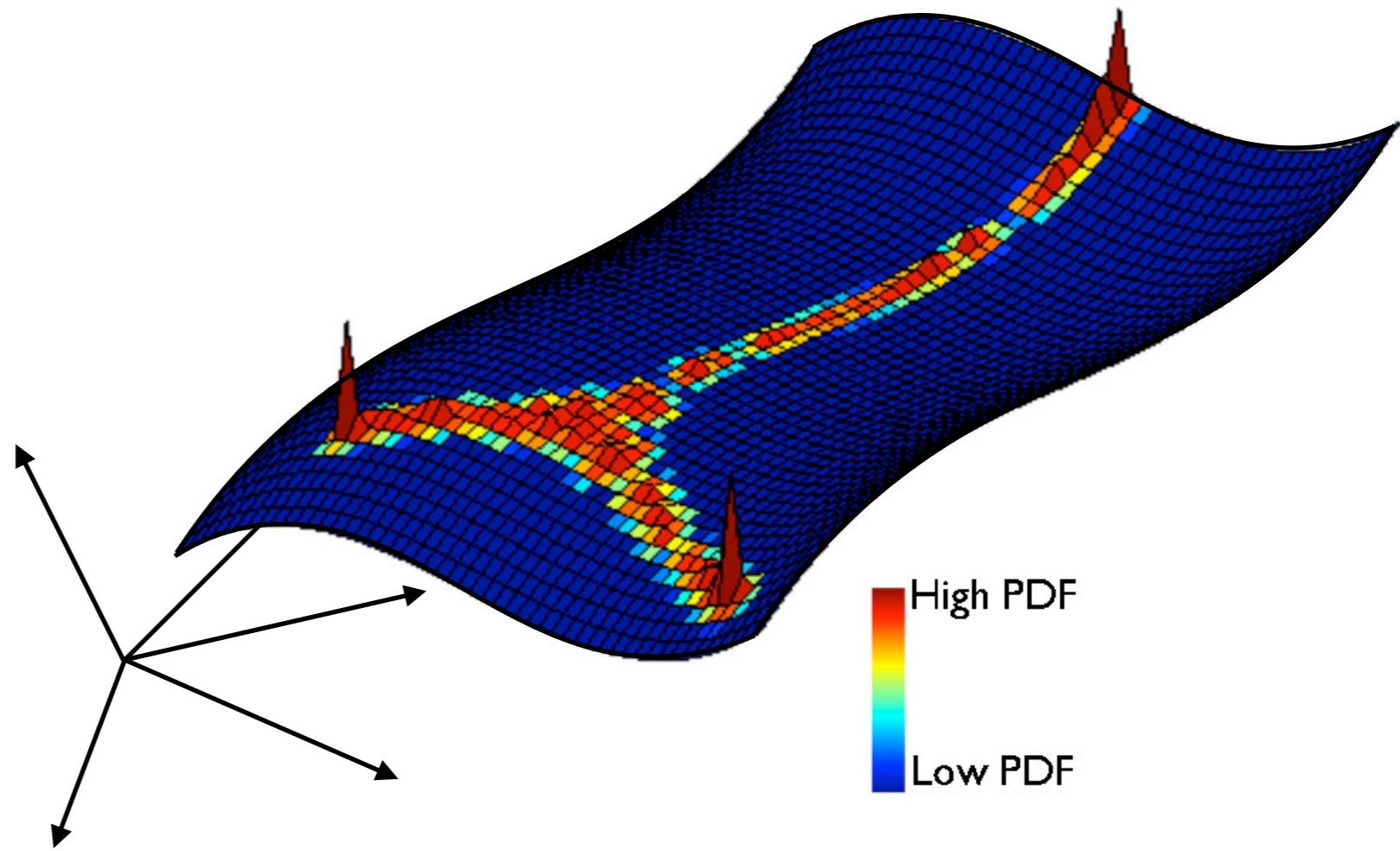


Single-cell diffusion maps



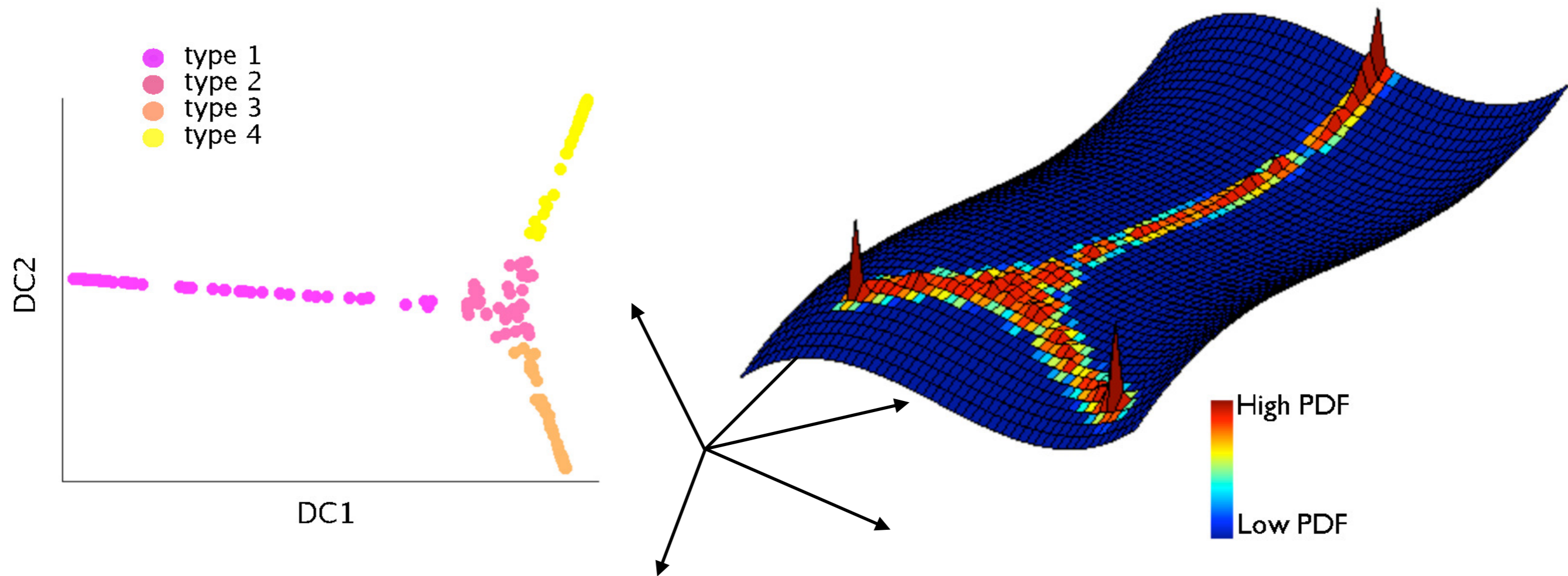
Markovian transition matrix T

$$T_{xy} \propto \exp\left(-\frac{1}{2\sigma^2}\|x - y\|\right)$$



diffusion paths form on data manifold because of superposition of Gaussians

Single-cell diffusion maps

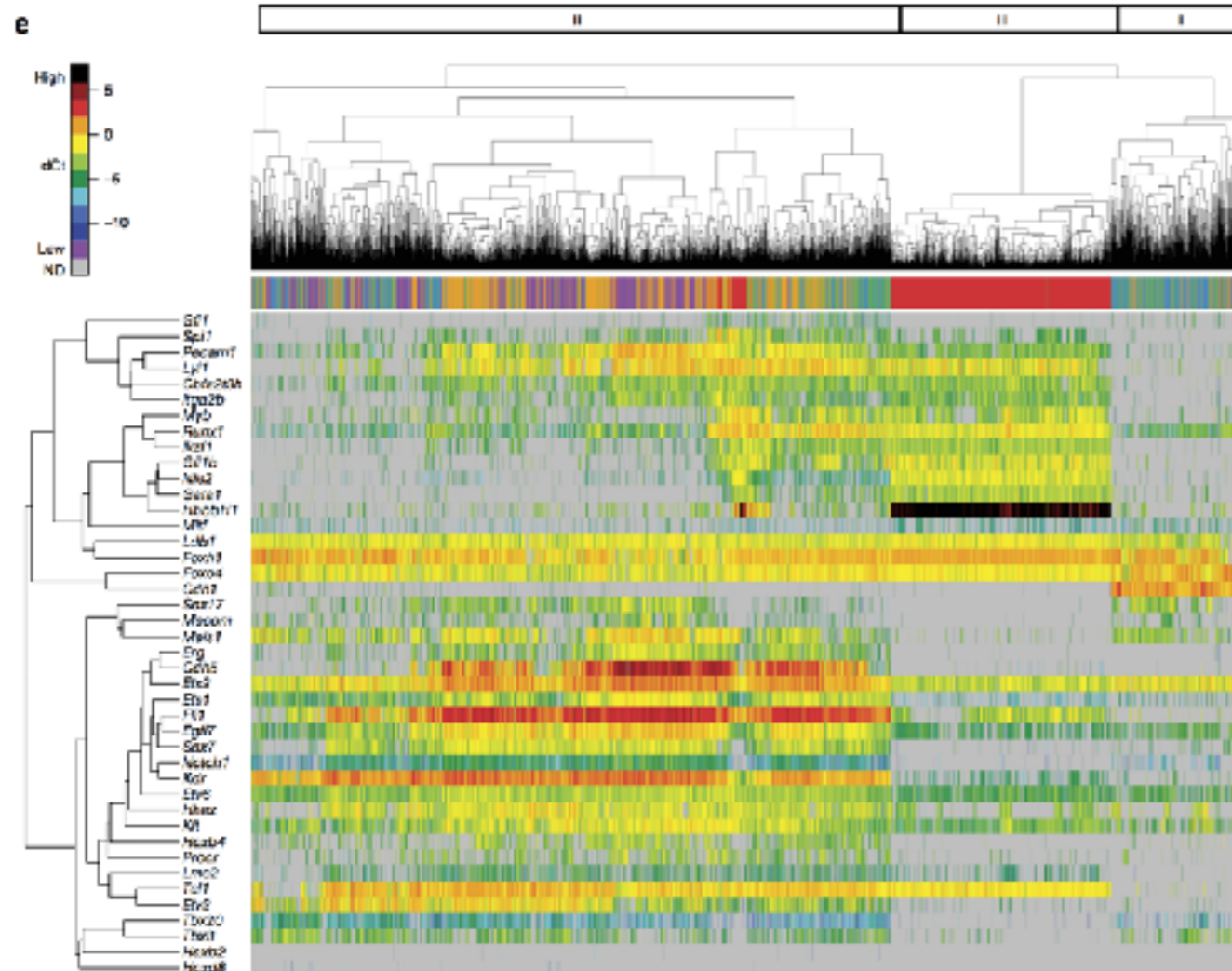
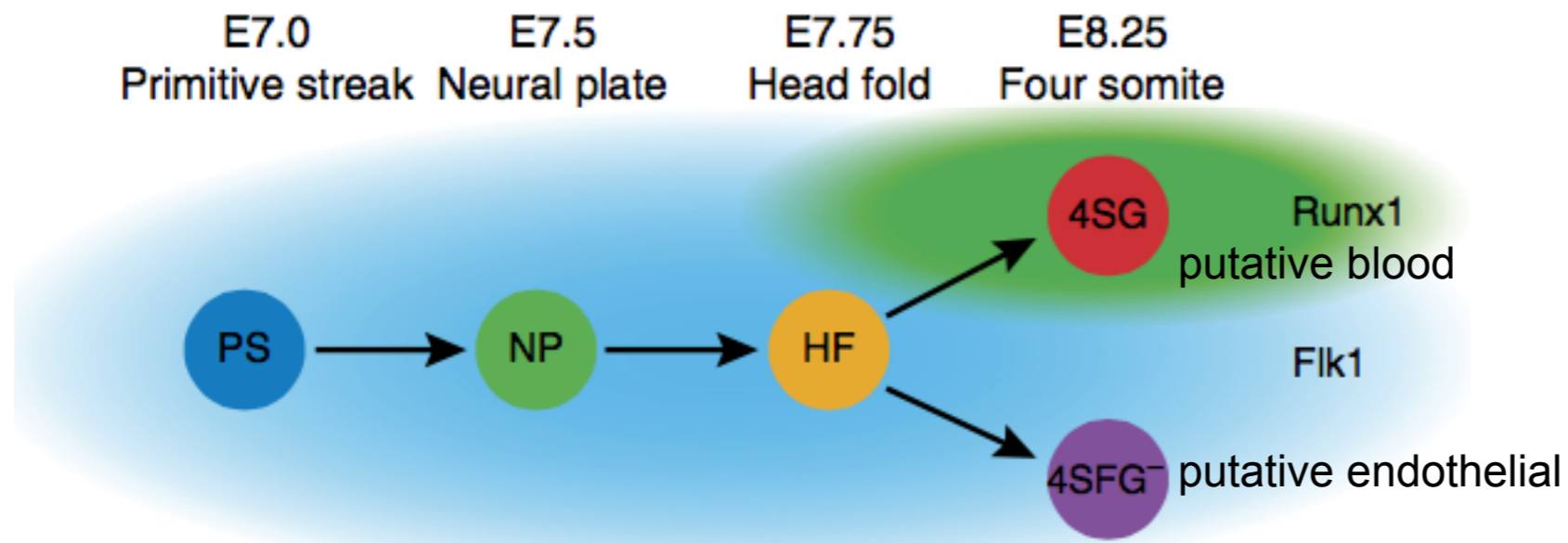


data projection onto the first two eigenvectors of T , diffusion component DC1 and DC2

example: visualizing early blood development



collab Göttgens lab
MRC, Uni Cambridge

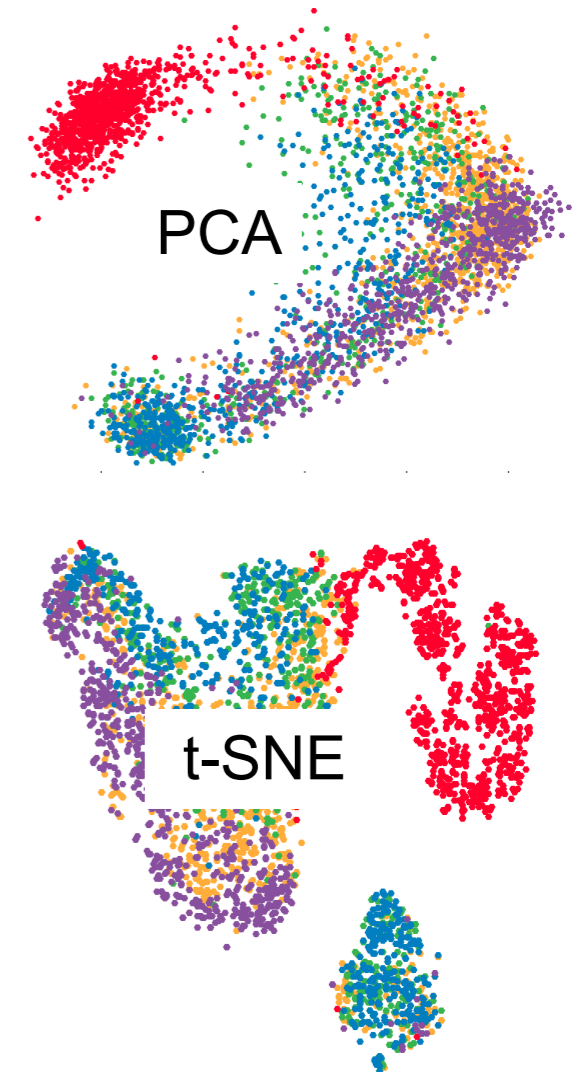
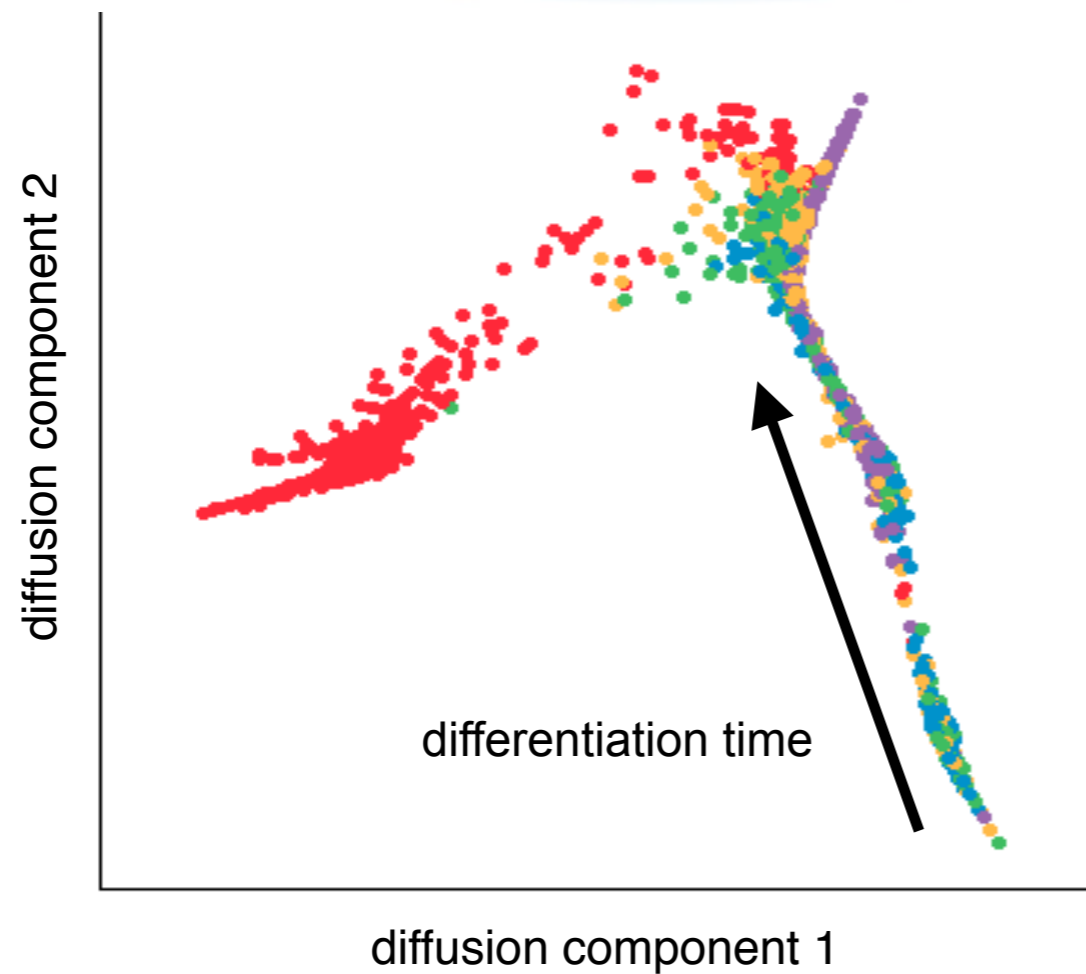
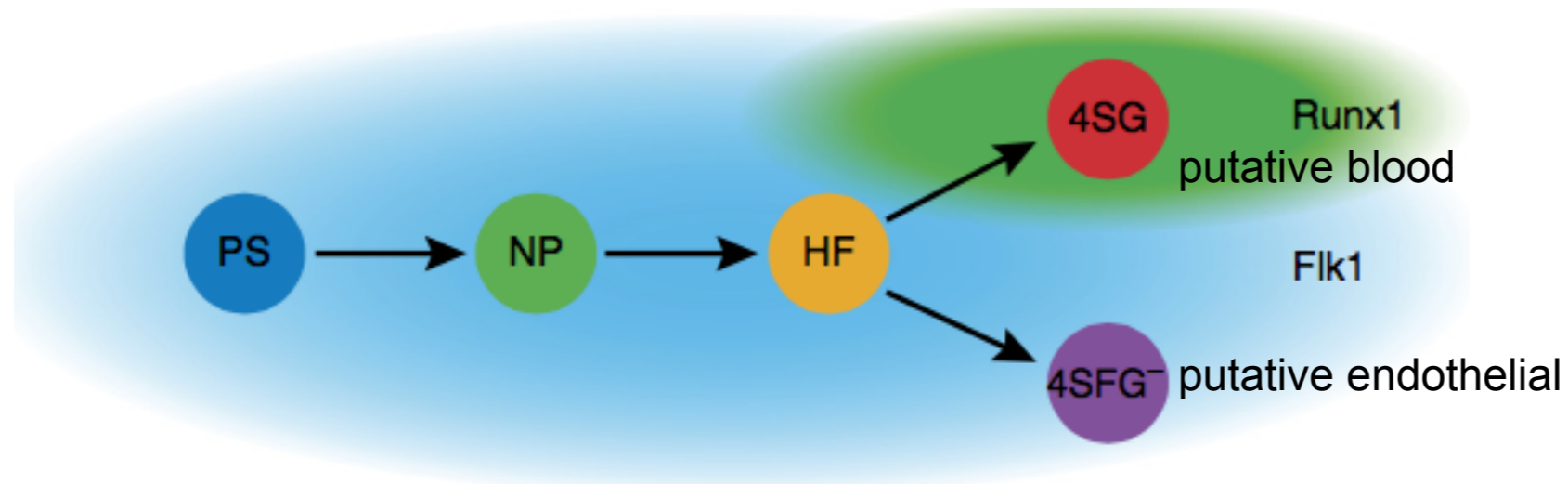


hematopoietic development in the mouse embryo, single-cell qPCR of 3,934 cells

example: visualizing early blood development



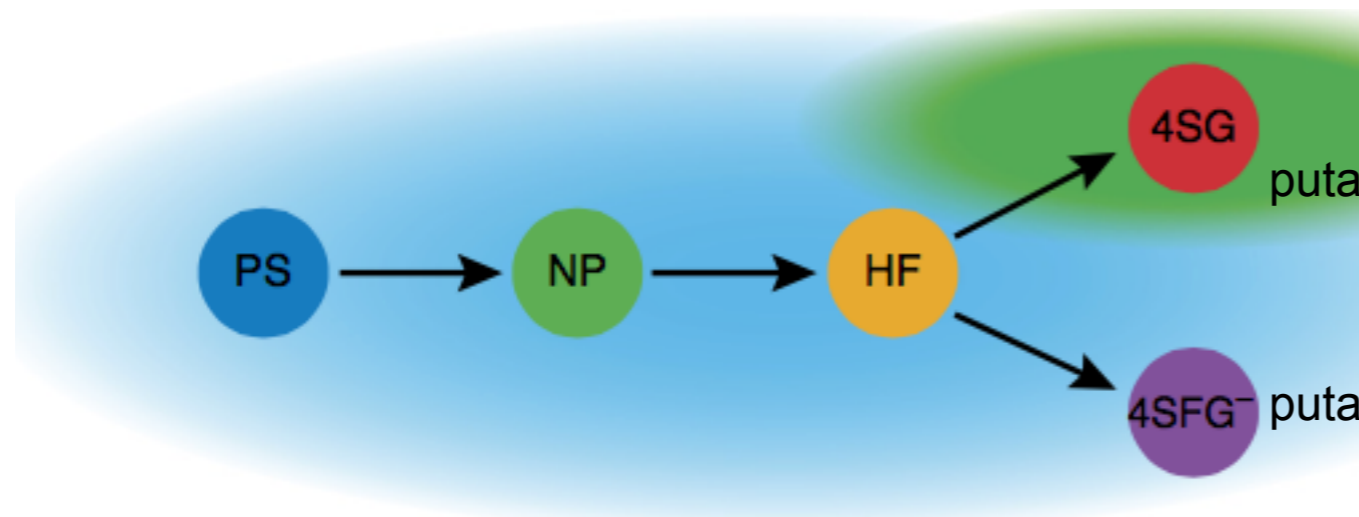
collab Göttgens lab
MRC, Uni Cambridge



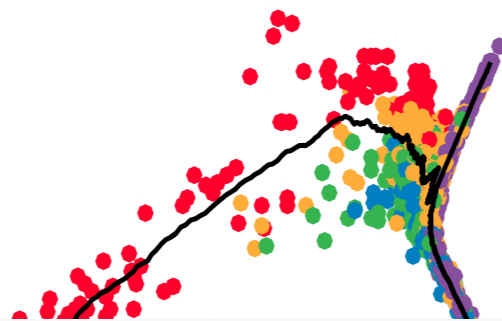
example: visualizing early blood development



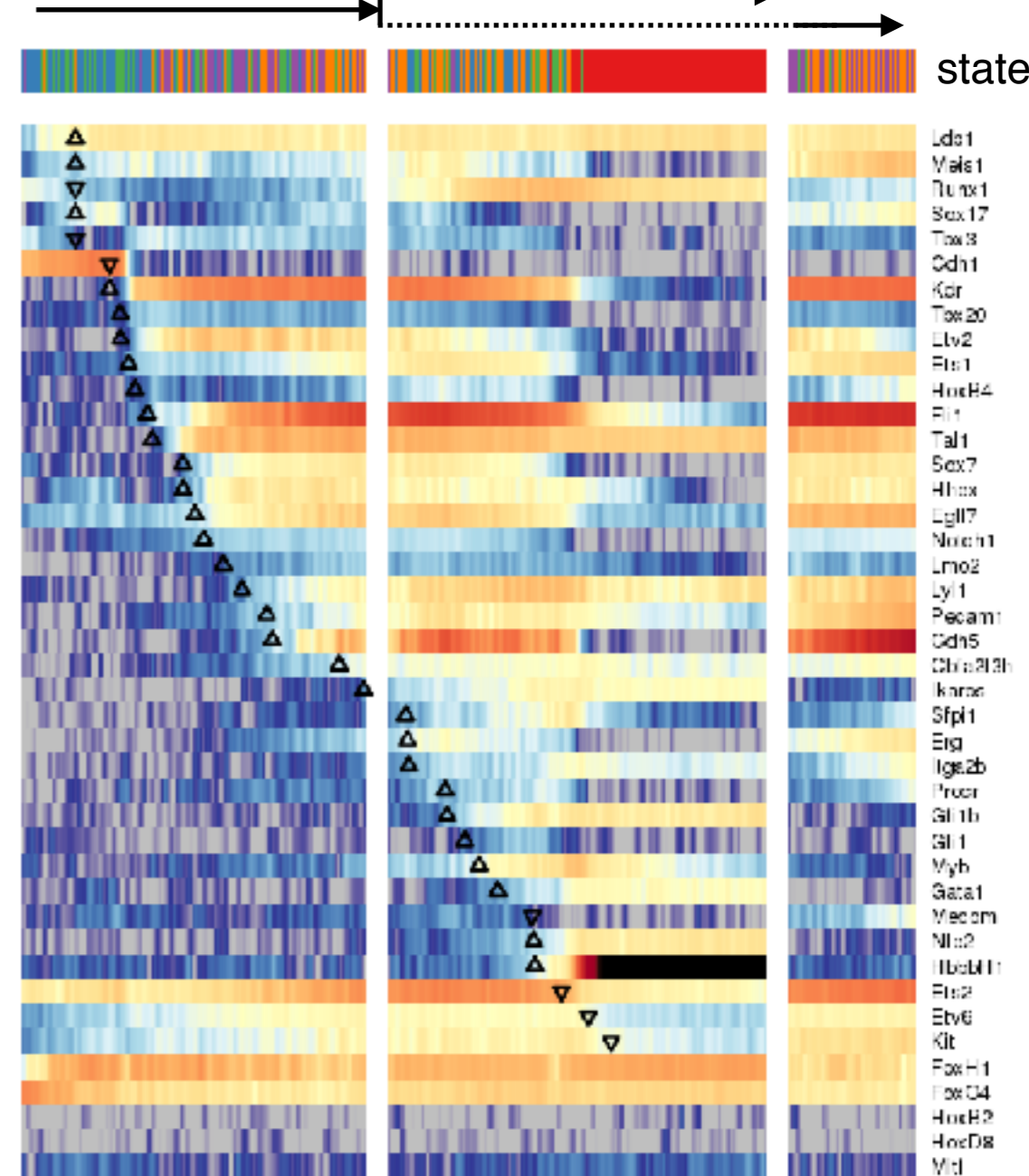
collab Göttgens lab



component 2



diffusion pseudotime

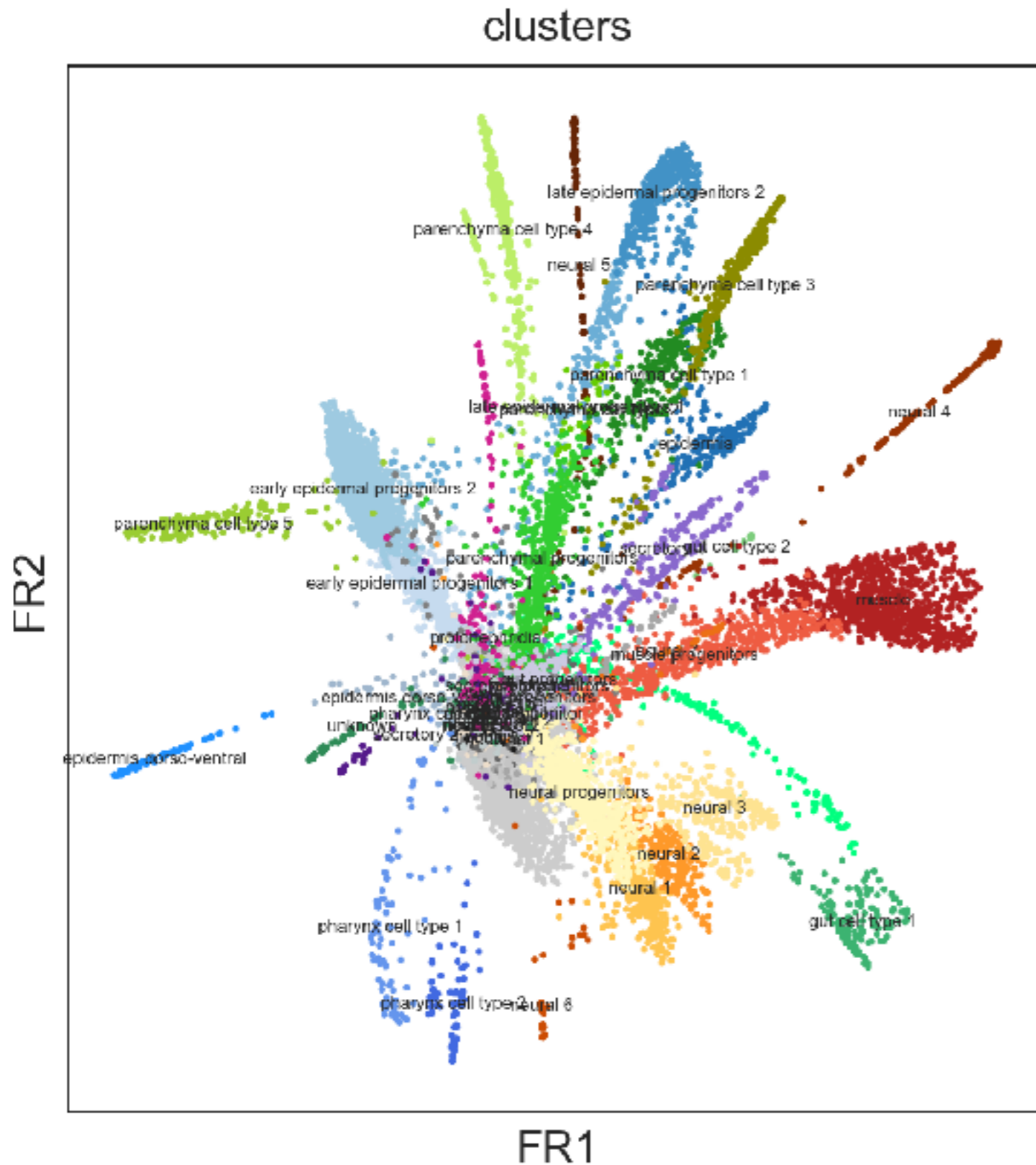


diffusion map and DPT implementation:
destiny R-package

```
library(destiny)
dm <- DiffusionMap(data, ...)
plot(dm, col.by = 'variable')
```

www.helmholtz-muenchen.de/icb/destiny

whole organism lineage tree



Which cell types / clusters exist, which are connected?

Which paths do cells take, where do branchings occur?

Trace gene dynamics / changes along paths?

Discrete & continuous topologies

goal: unify this!

discrete topology

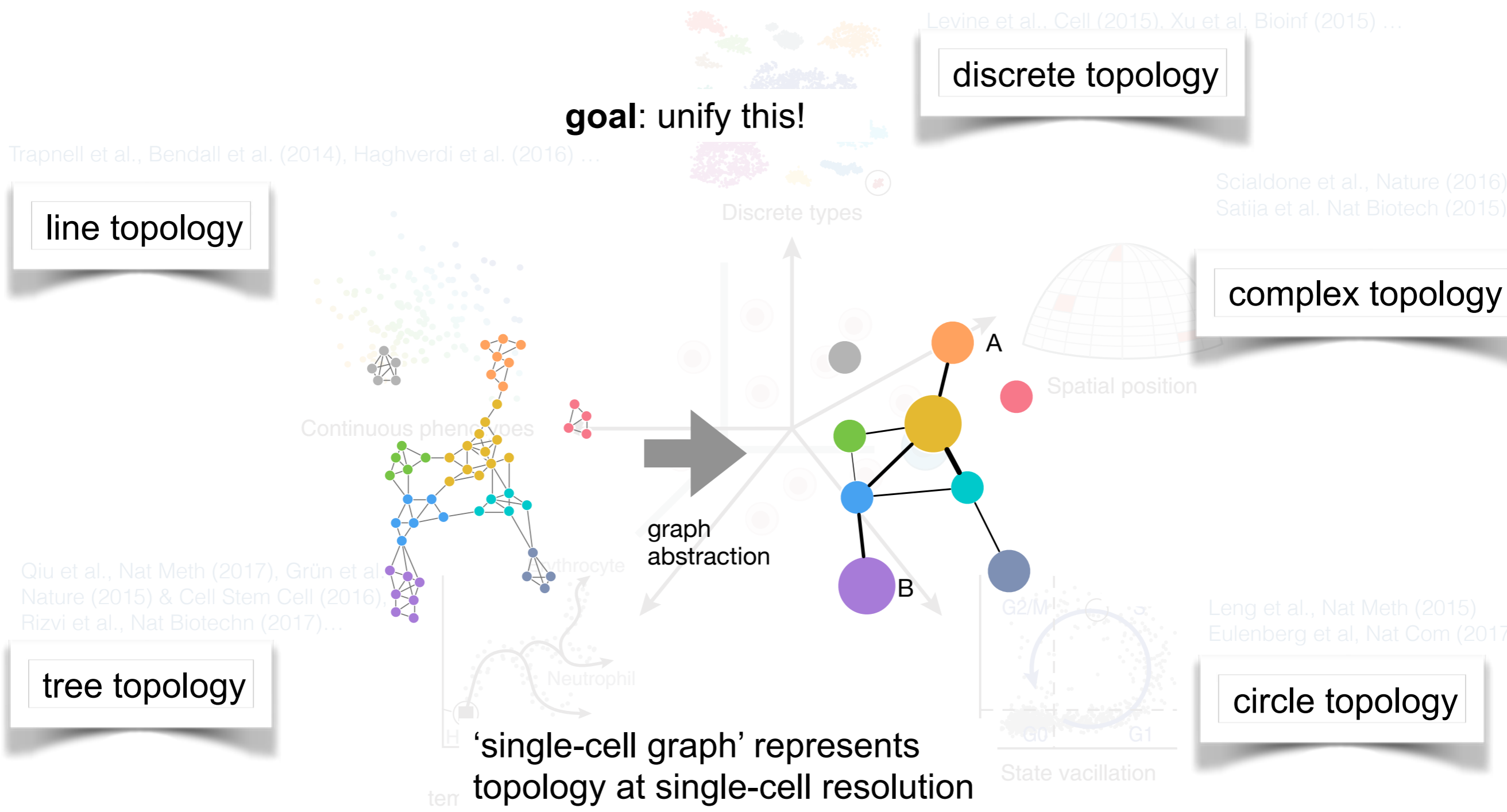
line topology

complex topology

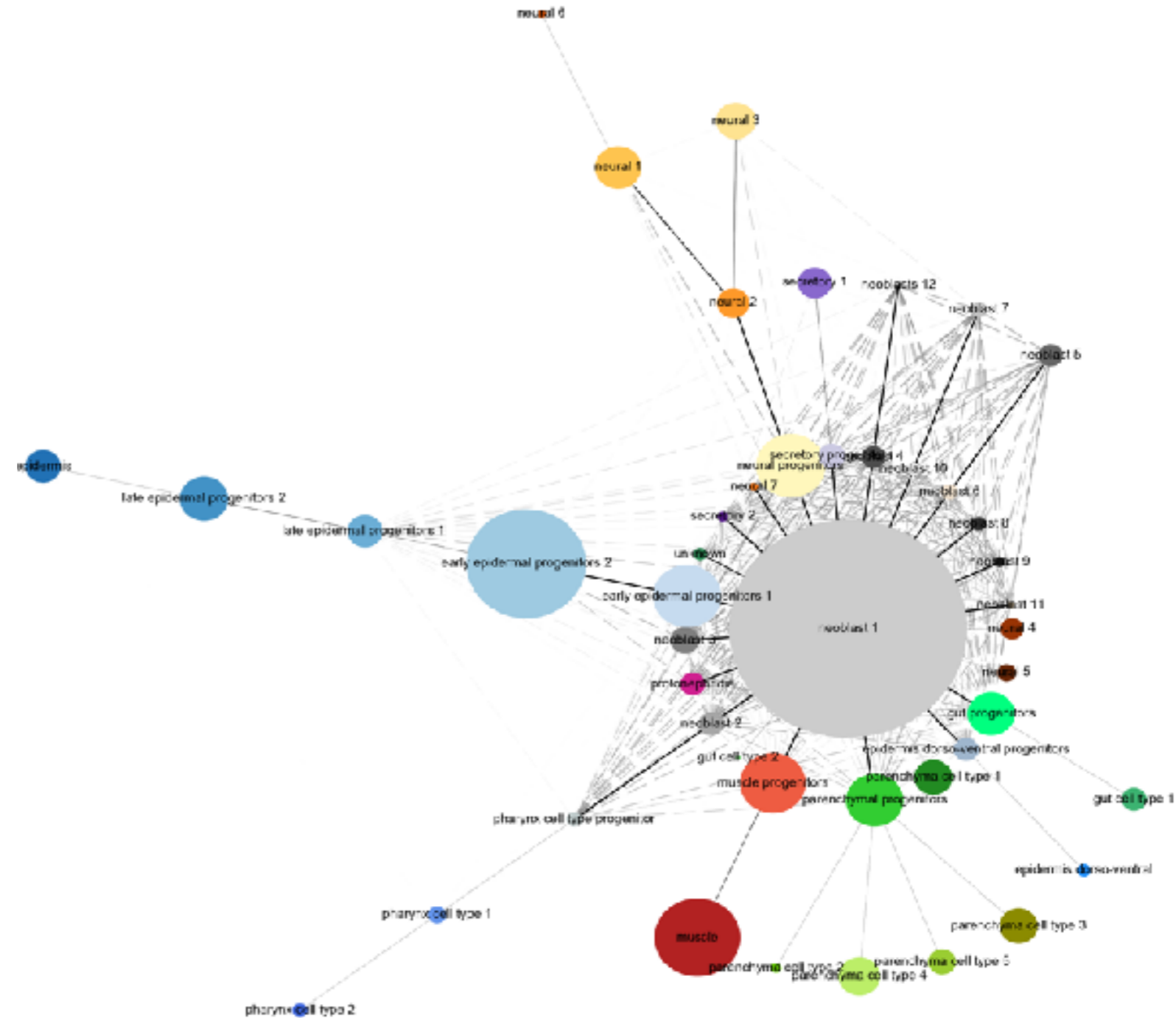
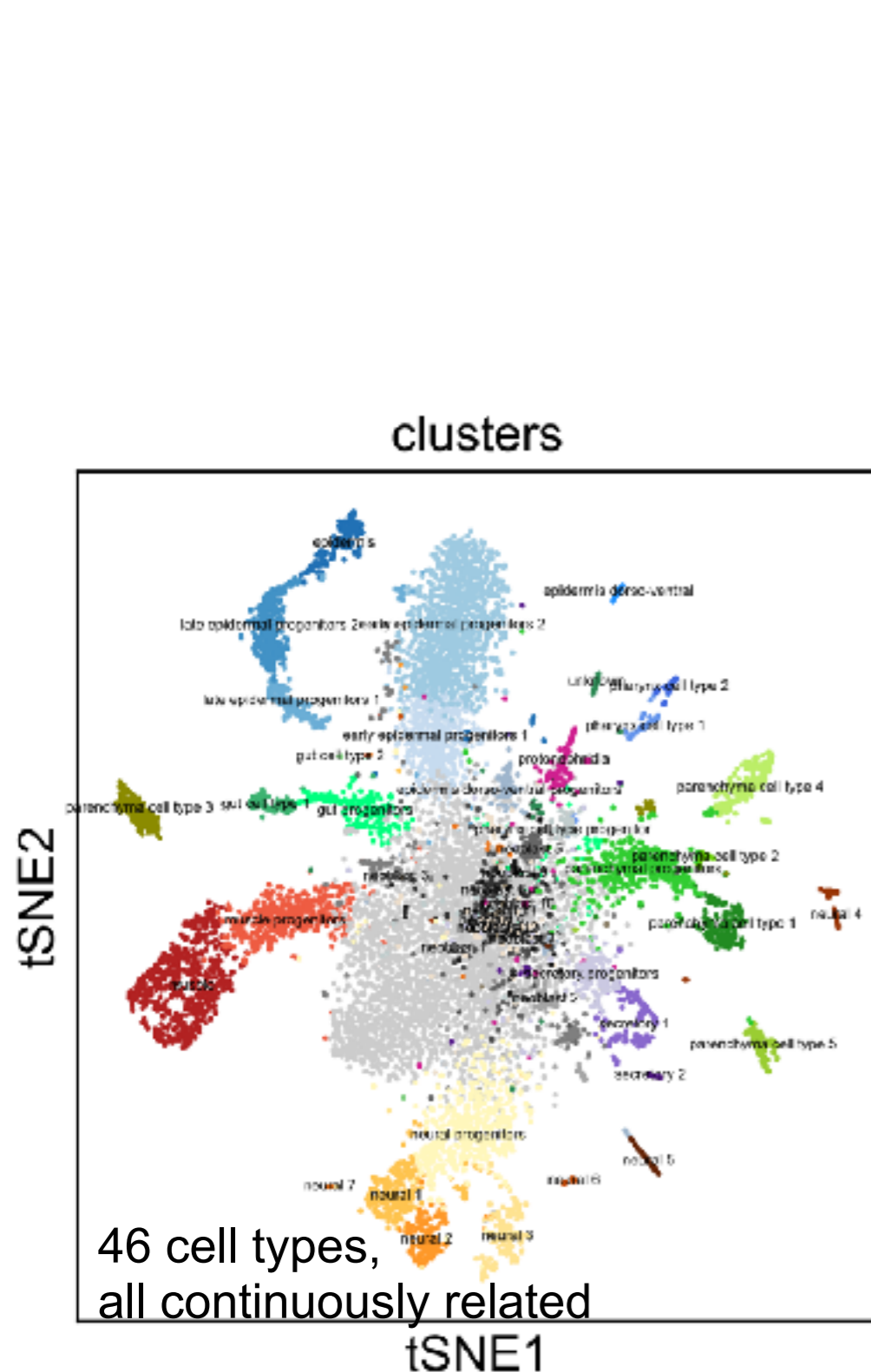
tree topology

circle topology

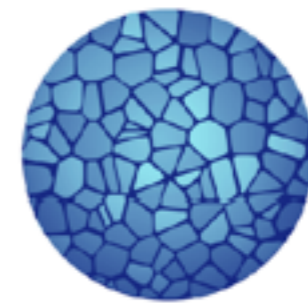
'single-cell graph' represents topology at single-cell resolution



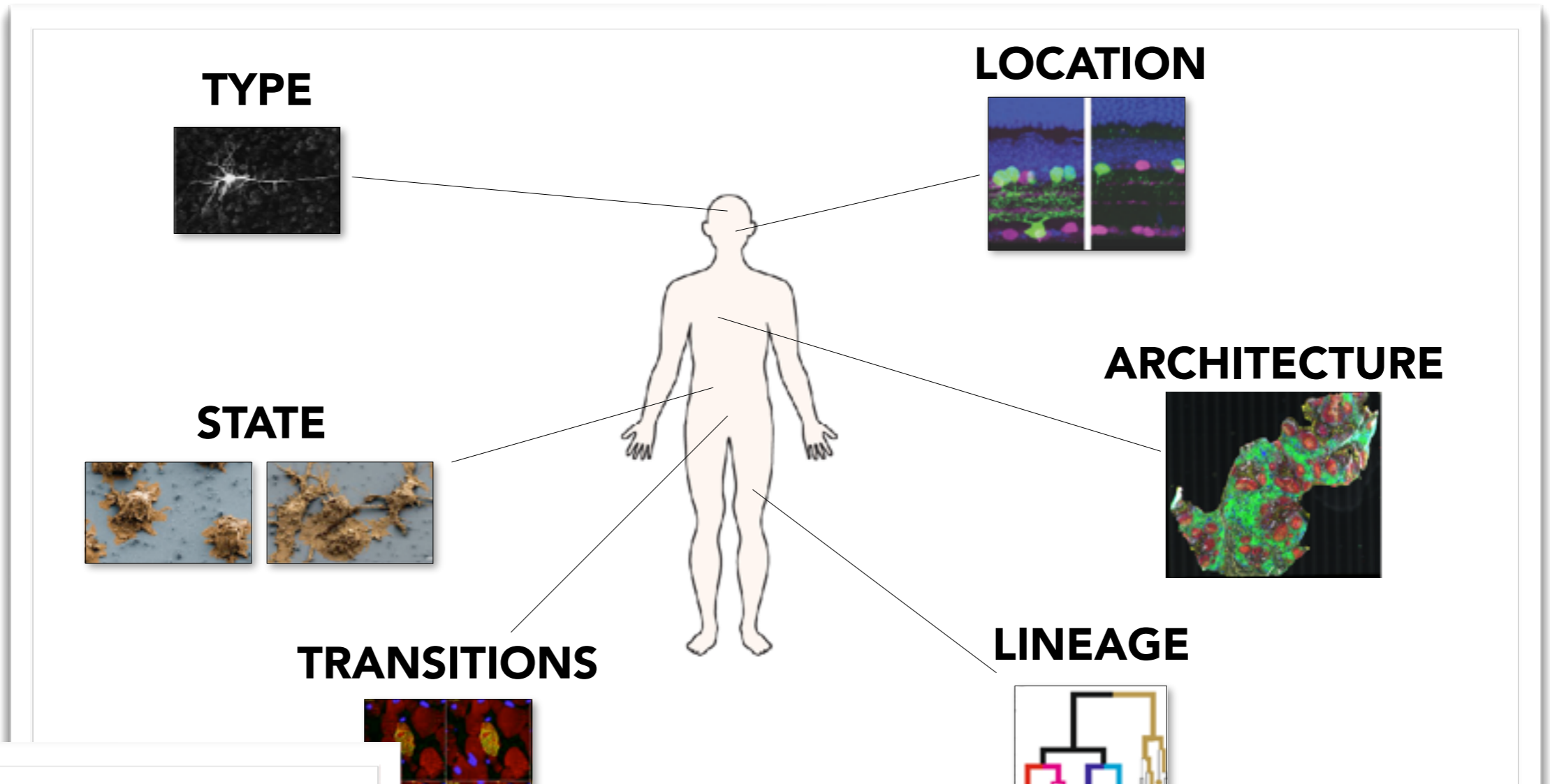
Inferring the lineage tree of planaria



outlook



**HUMAN
CELL
ATLAS**



www.single-cell.de



**CHAN
ZUCKERBERG
INITIATIVE**

Big data skill gap & education

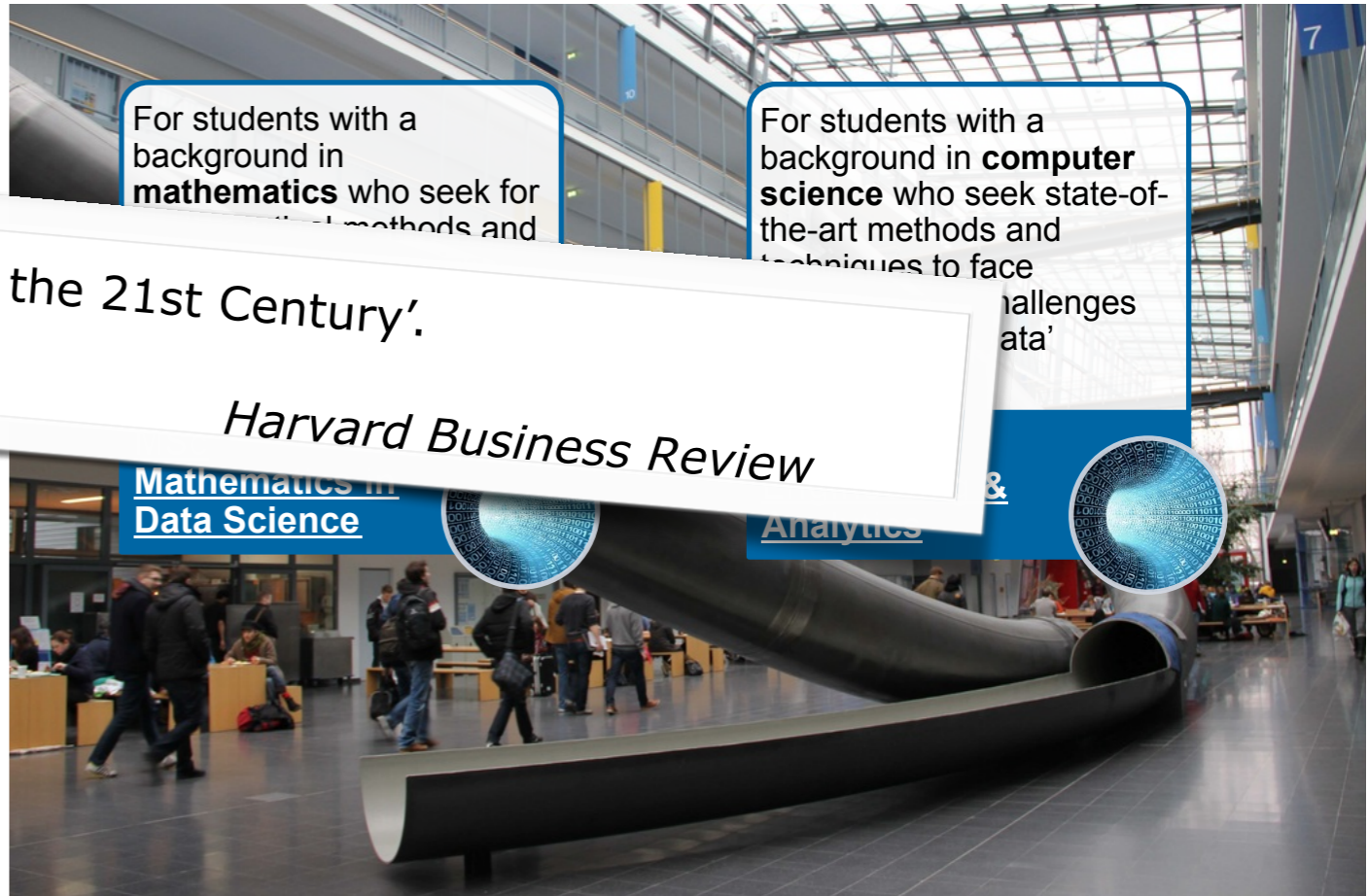
FACULTY OF MATHEMATICS, COMPUTER SCIENCE AND STATISTICS

ELITE MASTER PROGRAM DATA SCIENCE



Munich School for
Data Science
@ Helmholtz, TUM & LMU

Data scientist ... 'sexiest job of the 21st Century'.
Harvard Business Review



For students with a background in **mathematics** who seek for advanced methods and

For students with a background in **computer science** who seek state-of-the-art methods and techniques to face challenges 'data'

Mathematics in Data Science

Analytics



Conclusion

summary

- » *preprocessing*: deep count autoencoder denoising
- » *diffusion pseudotime*: understand temporal structure of differentiation processes
- » *graph abstraction*: robust multi-branch analysis
- » applications to hematopoiesis and epithelial gut

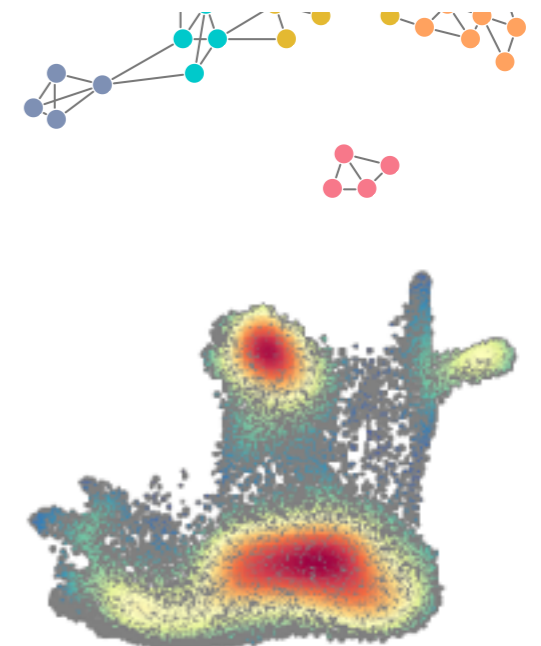
outlook

- » challenges of *large-scale scRNAseq*
- » *human cell atlas* as single-cell resolved background map for complex diseases
- » data scientist education

Mike Inouye @minouye271



The @VicGovAu runs some great anti-gambling advertising pic.twitter.com/xZsnUqkKnf



Institute of Computational Biology



Bundesministerium
für Bildung
und Forschung

Deutsche
Forschungsgemeinschaft
DFG



www.comp.bio

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

HELMHOLTZ SPITZENFORSCHUNG FÜR
GROSSE HERAUSFORDERUNGEN



 @fabian_theis