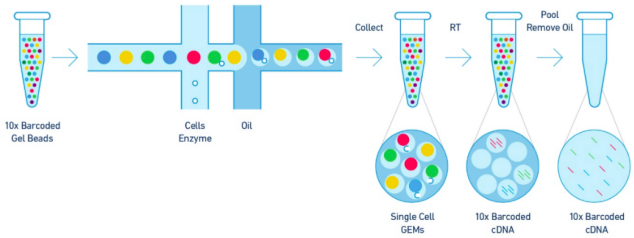
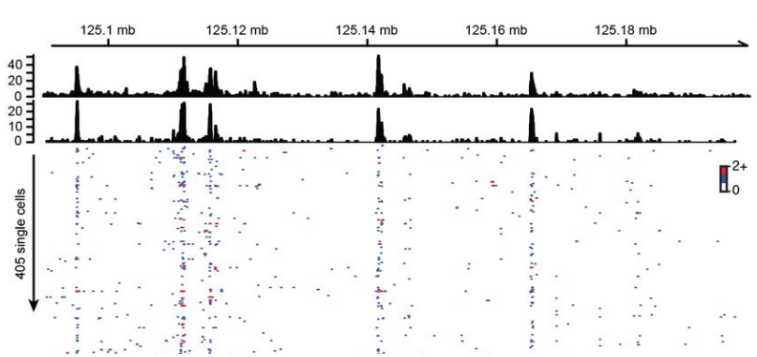
Challenges and Opportunities in Single Cell Data

Machine Learning for Single Cell Analysis

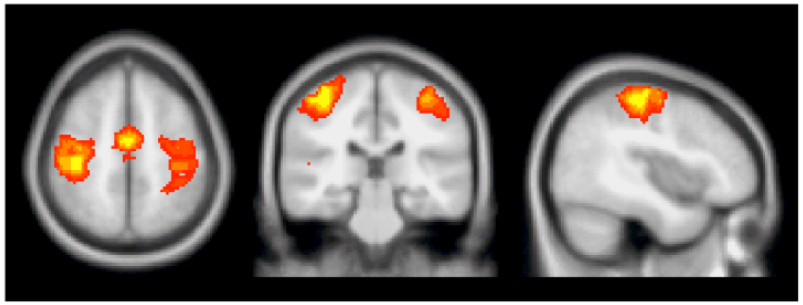
Big Biomedical Data



ScRNA-seq



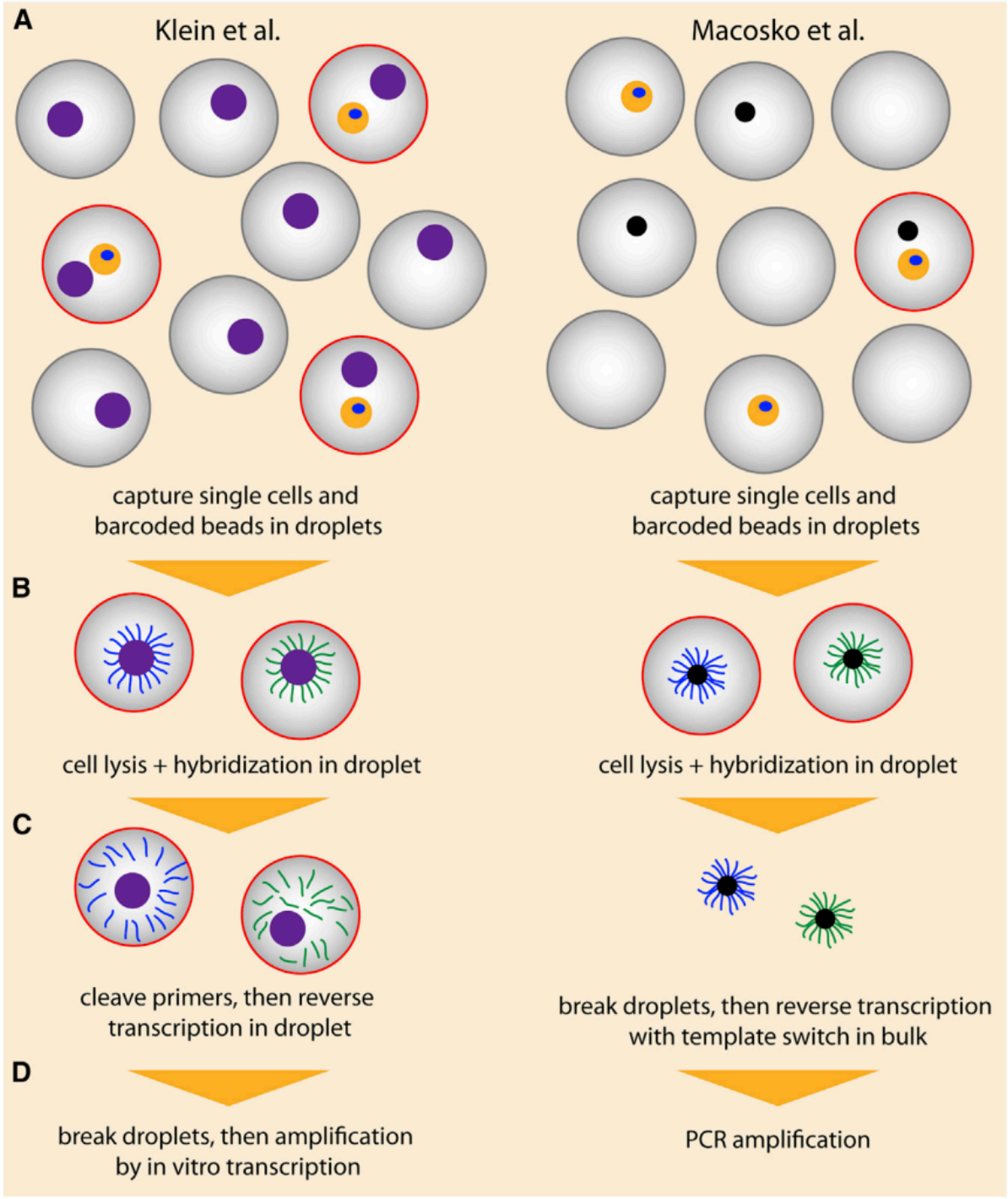
ScATAC-seq



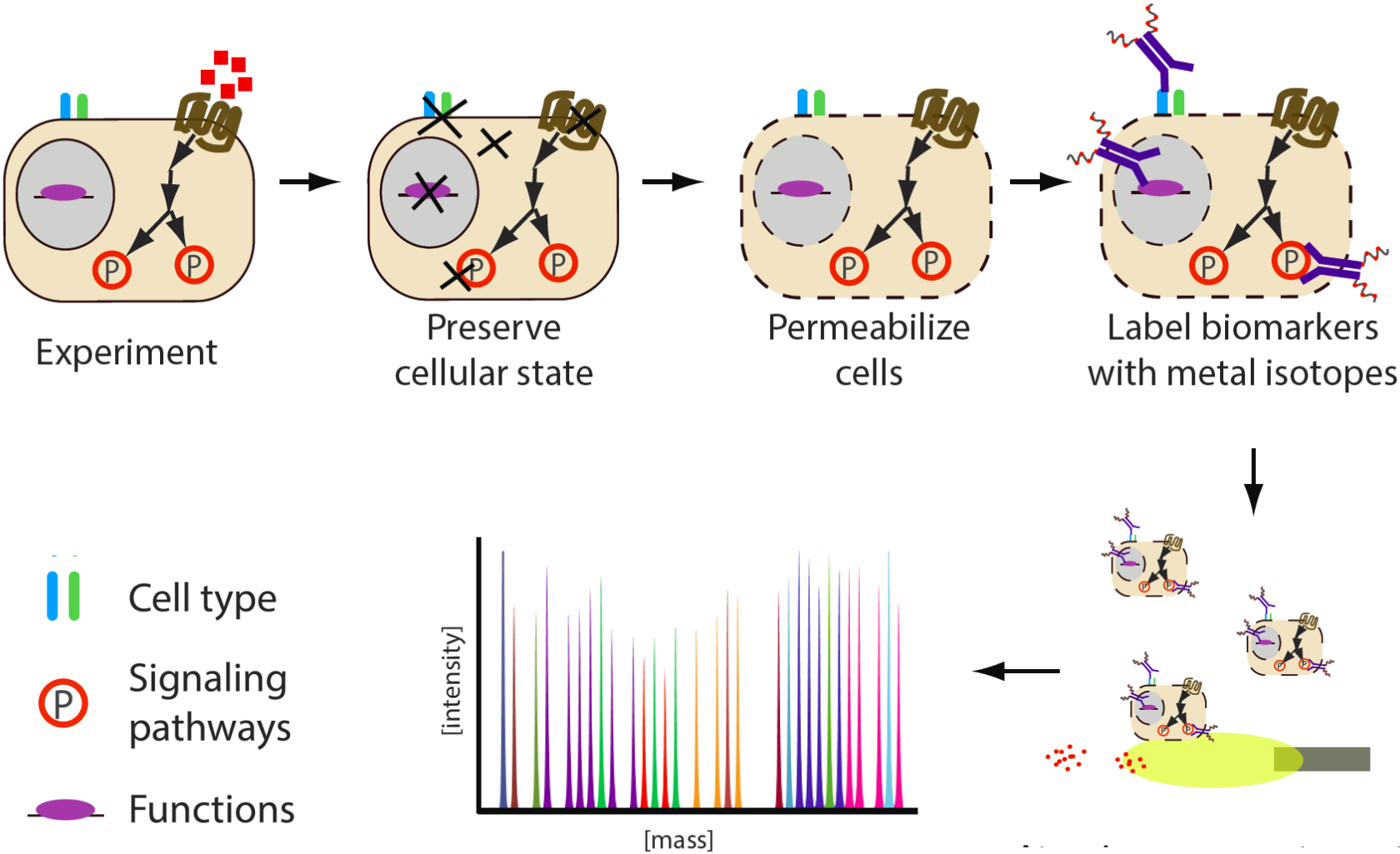
FMRI

Patient Data

Droplet-based Technologies



Single-Cell Proteomics: Mass Cytometry



Targeted & non-redundant

MASS CYTOMETRY

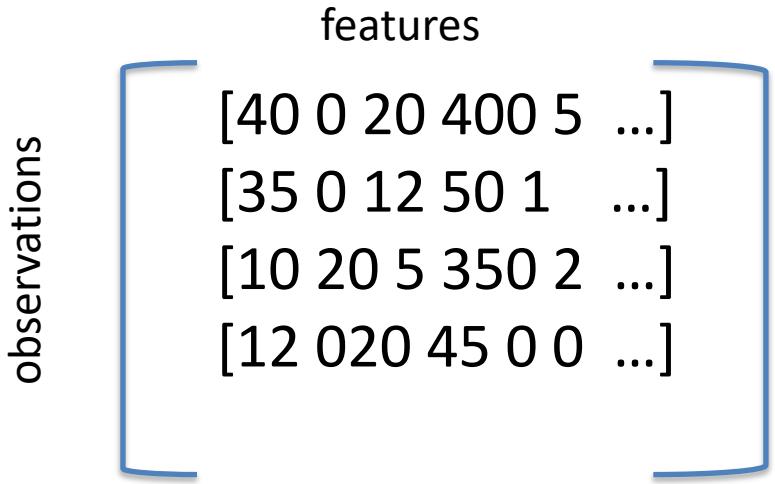
Perez O.D. *et al.* Nat. Biotech. 2002. Bandura D. *et al*. Anal. Chem. 2009.
Bendall S.C. *et al.* Science 2011. Bodenmiller B. *et al.* Nat. Biotechnol. 2012.

Single Cell Data

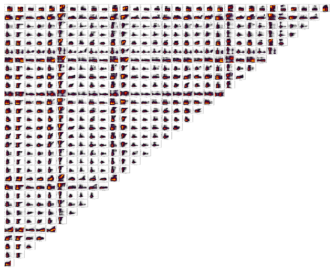
Cell 1 = [40 0 20 18 5 0 …]

Each cell is a vector of measurements

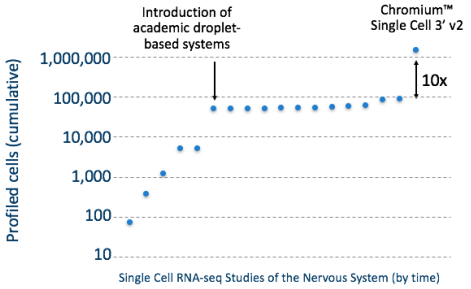
The whole data is a matrix with many observations (cells) and features (proteins, genes)



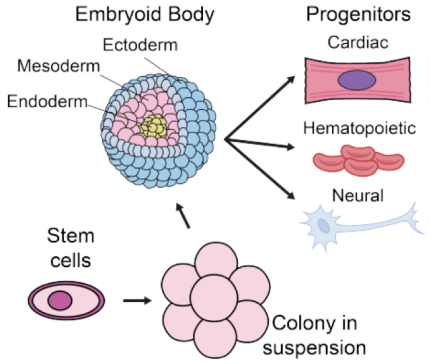
High Dimensional



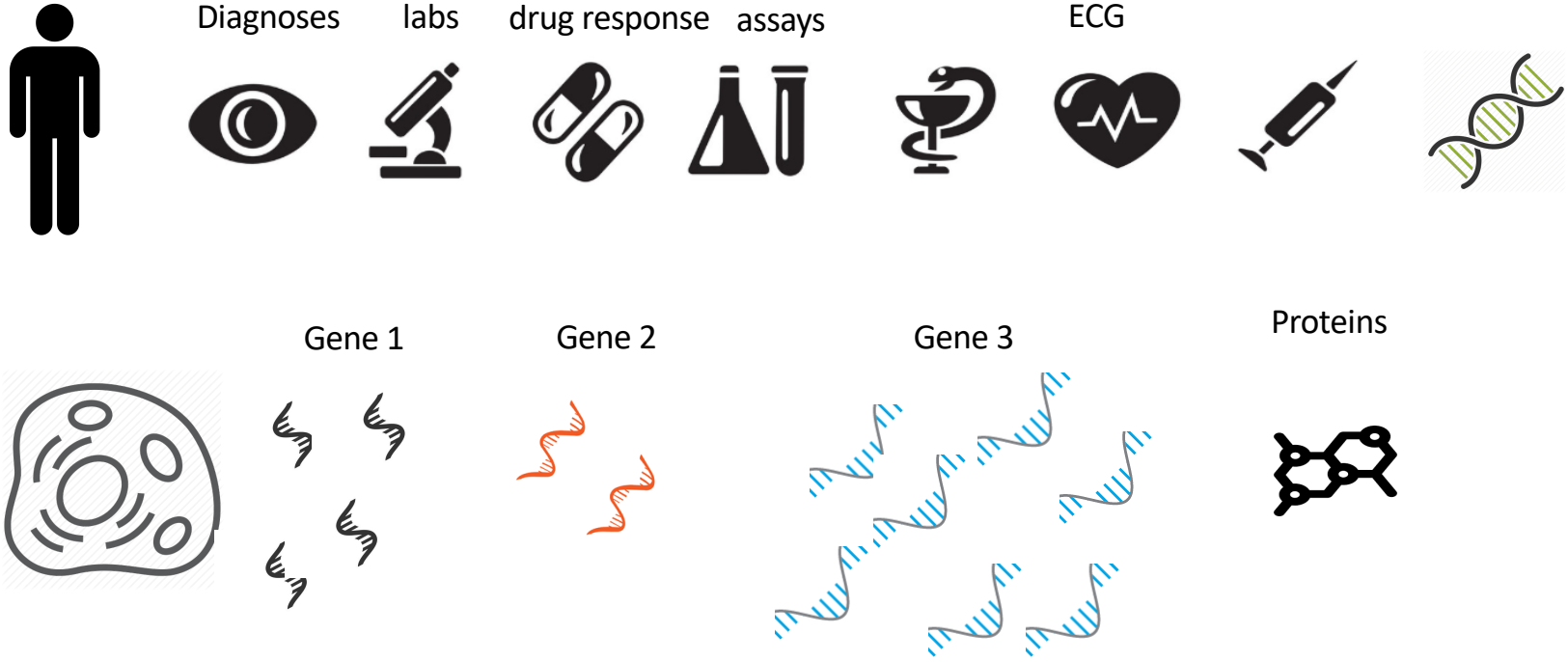
High Throughput



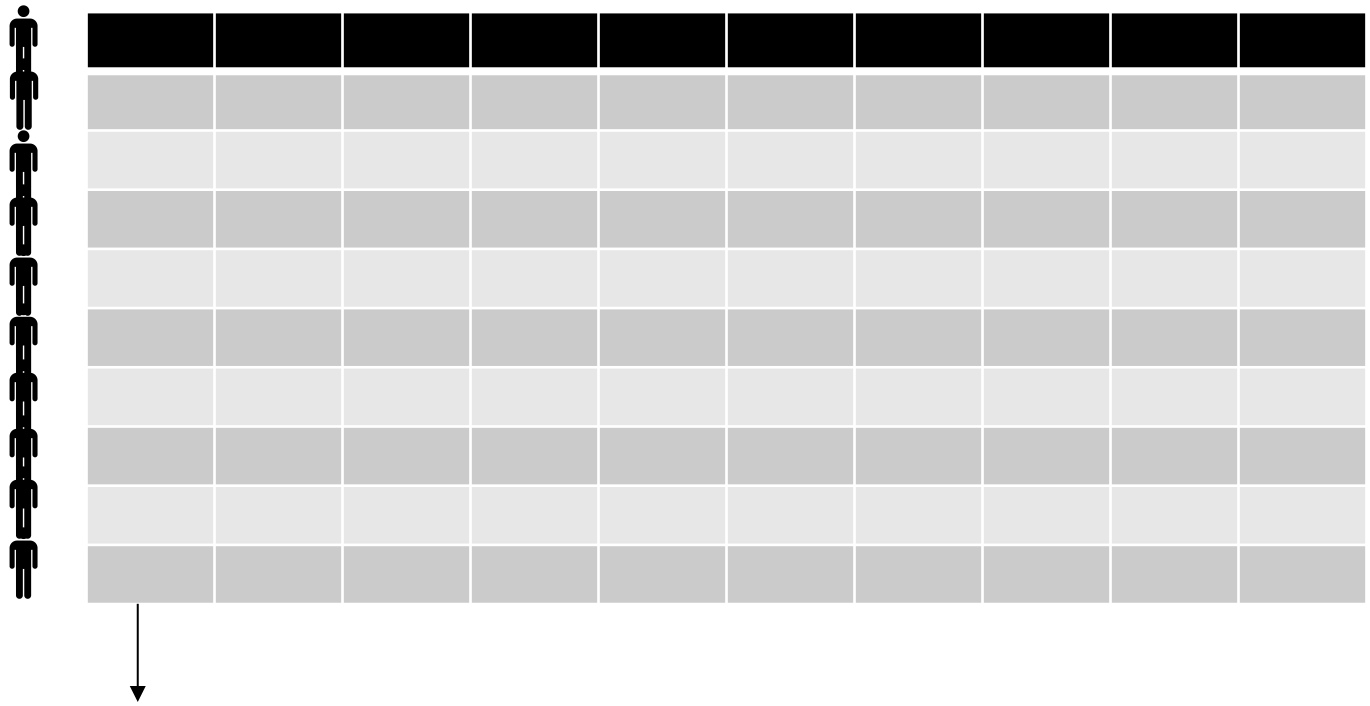
Heterogeneous



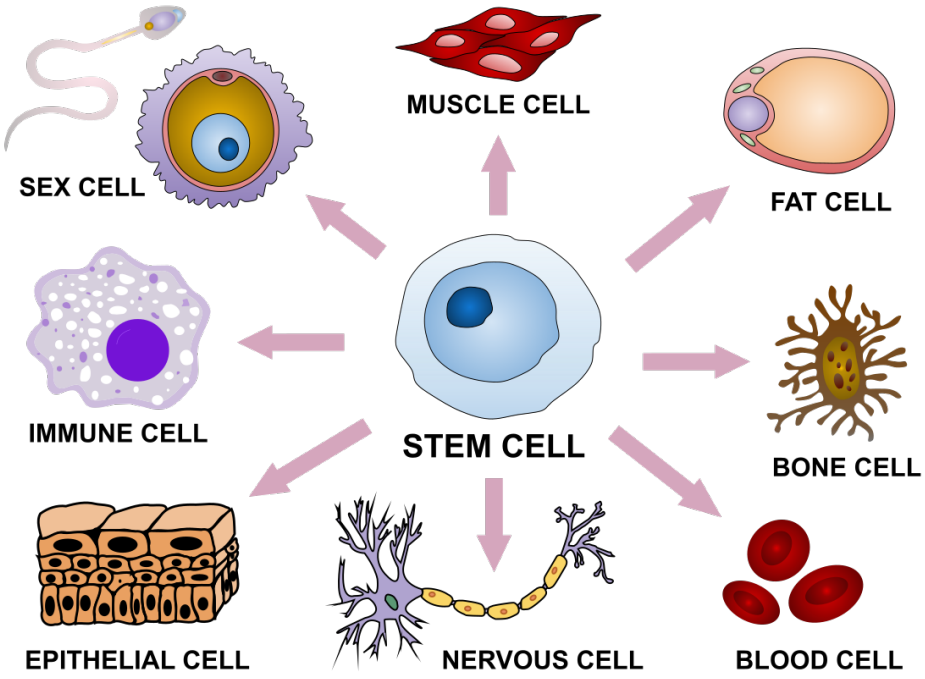
Many dimensions = many measurements



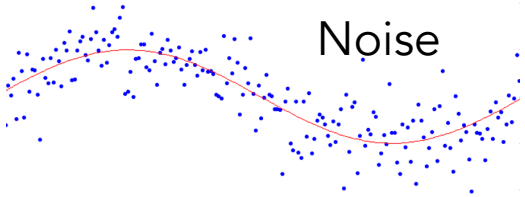
High Throughput = Many observations



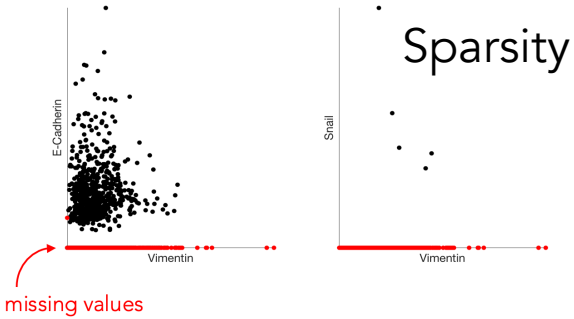
Heterogeneous Observations



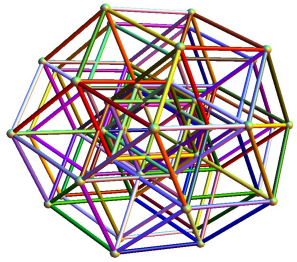
Challenges



Noise



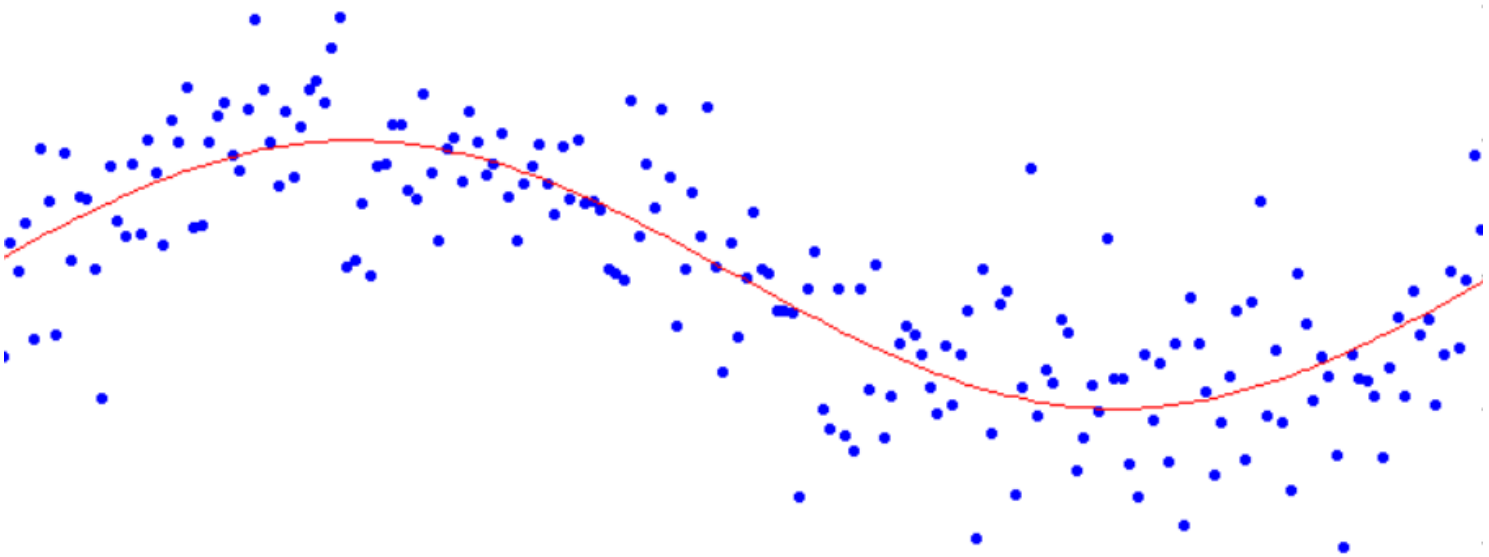
Sparsity



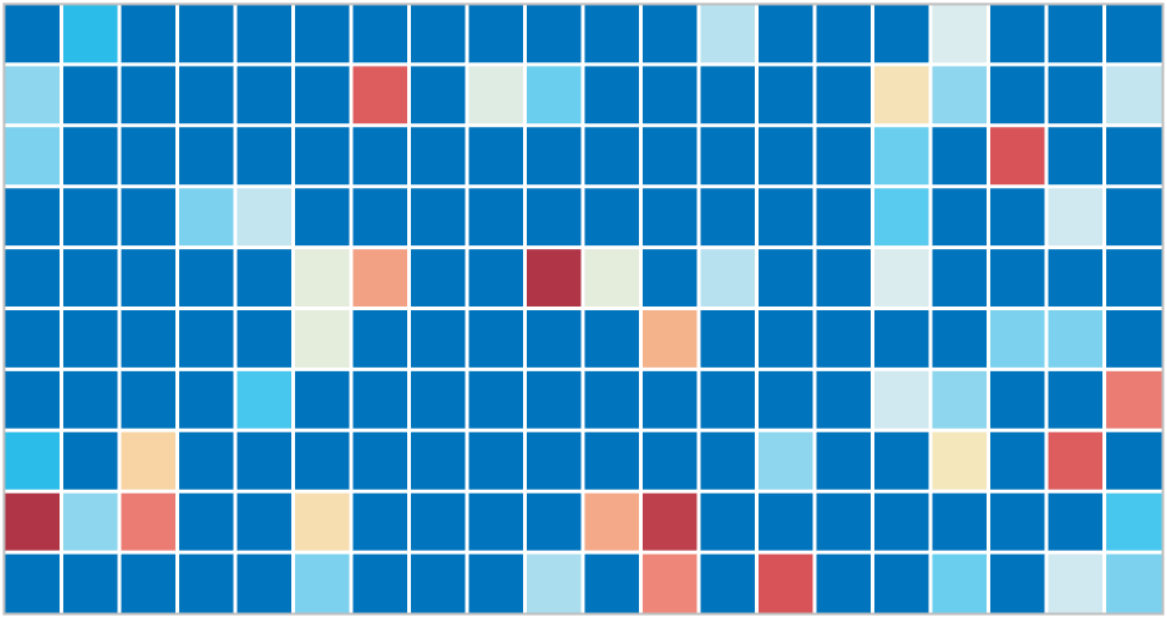
Dimensionality

Scale

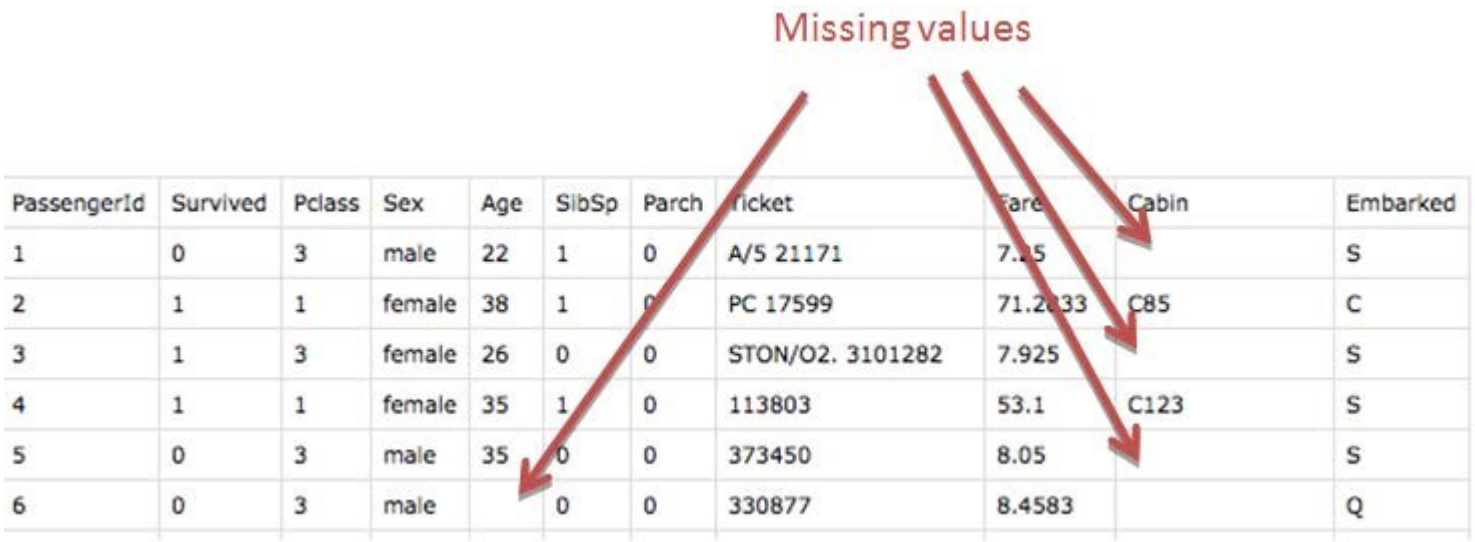
Noise



Dropout

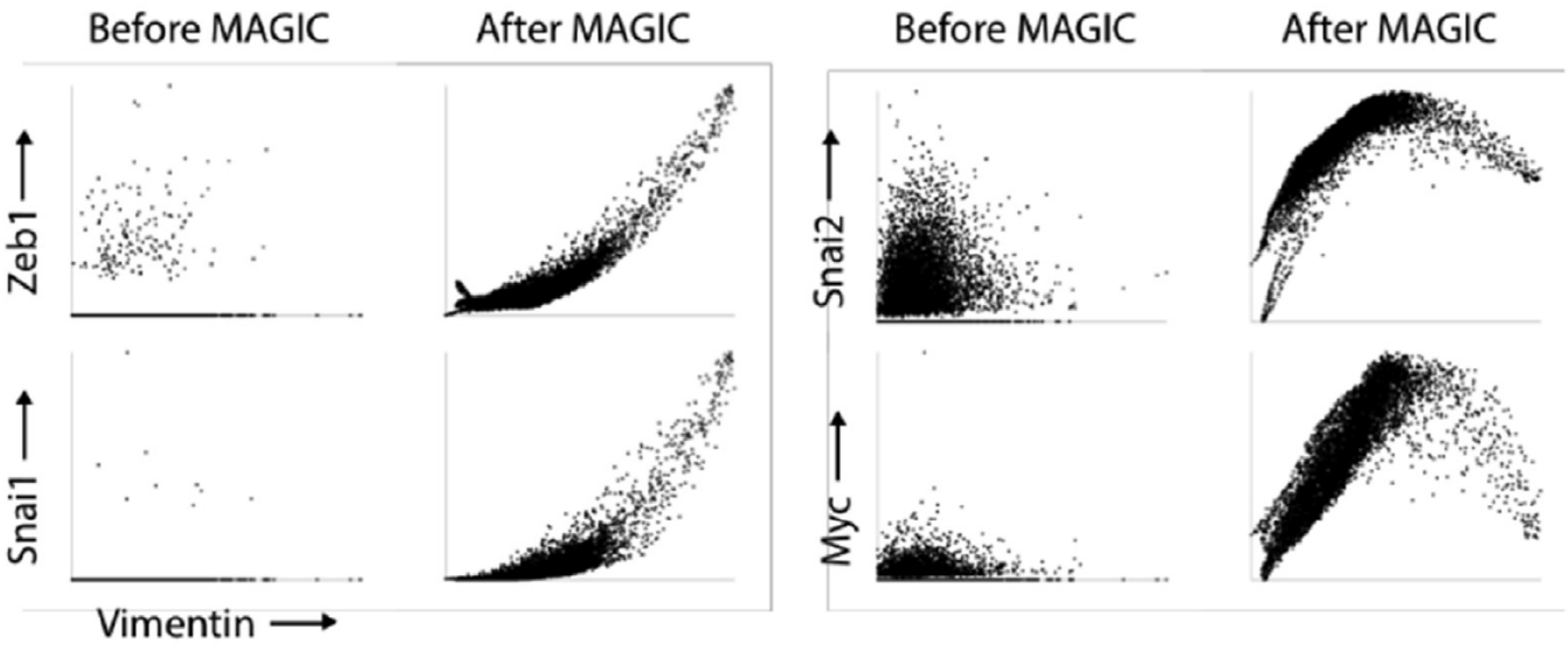


Dropout vs Missing Data

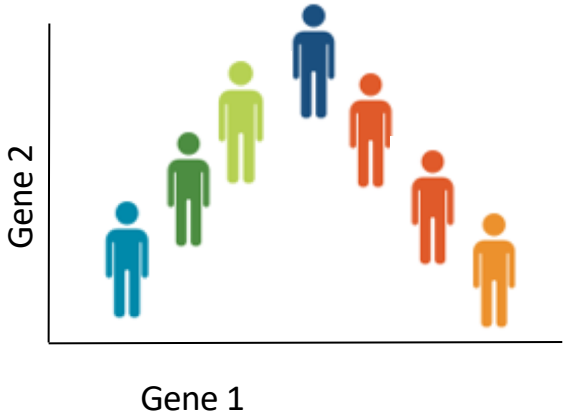


PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Denoising Data

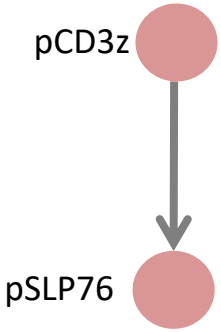


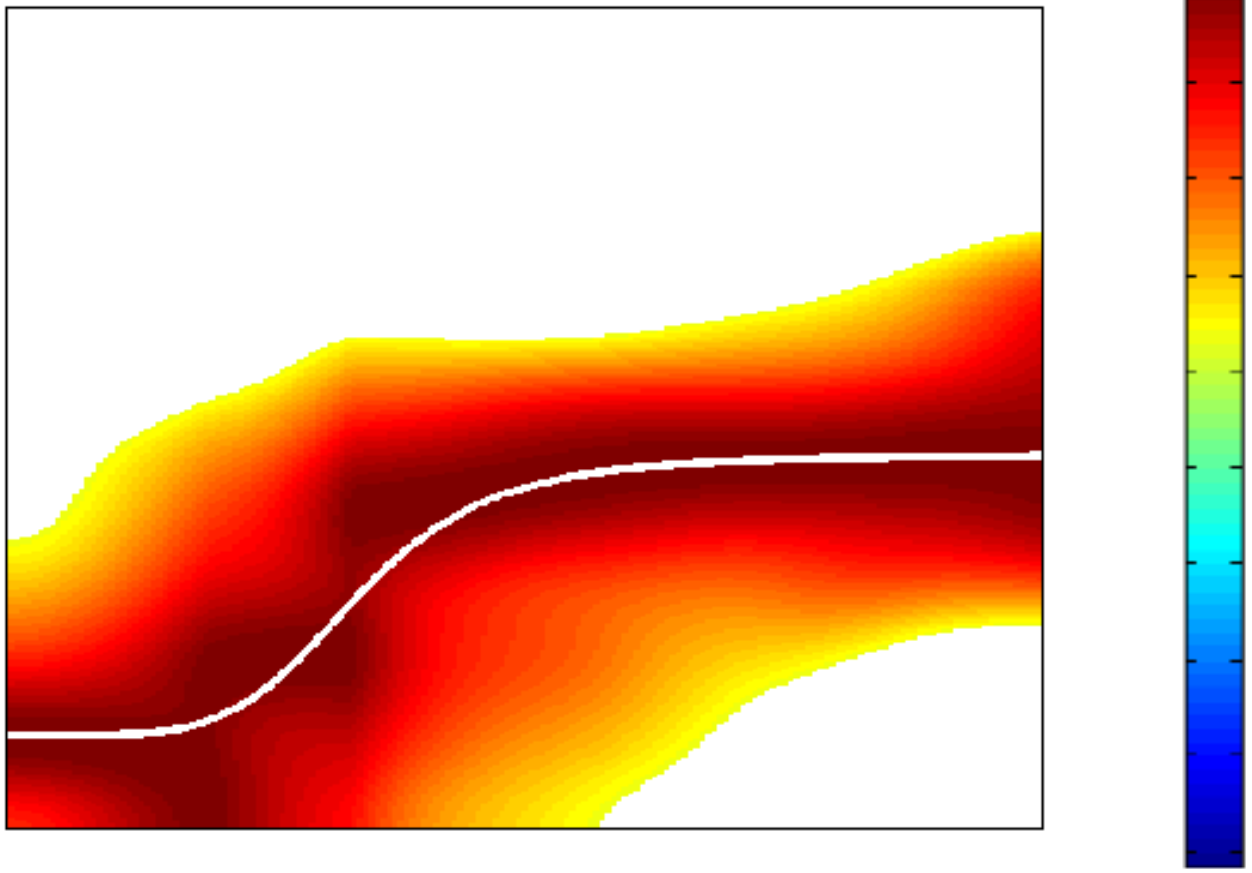
Gene-gene Relationships



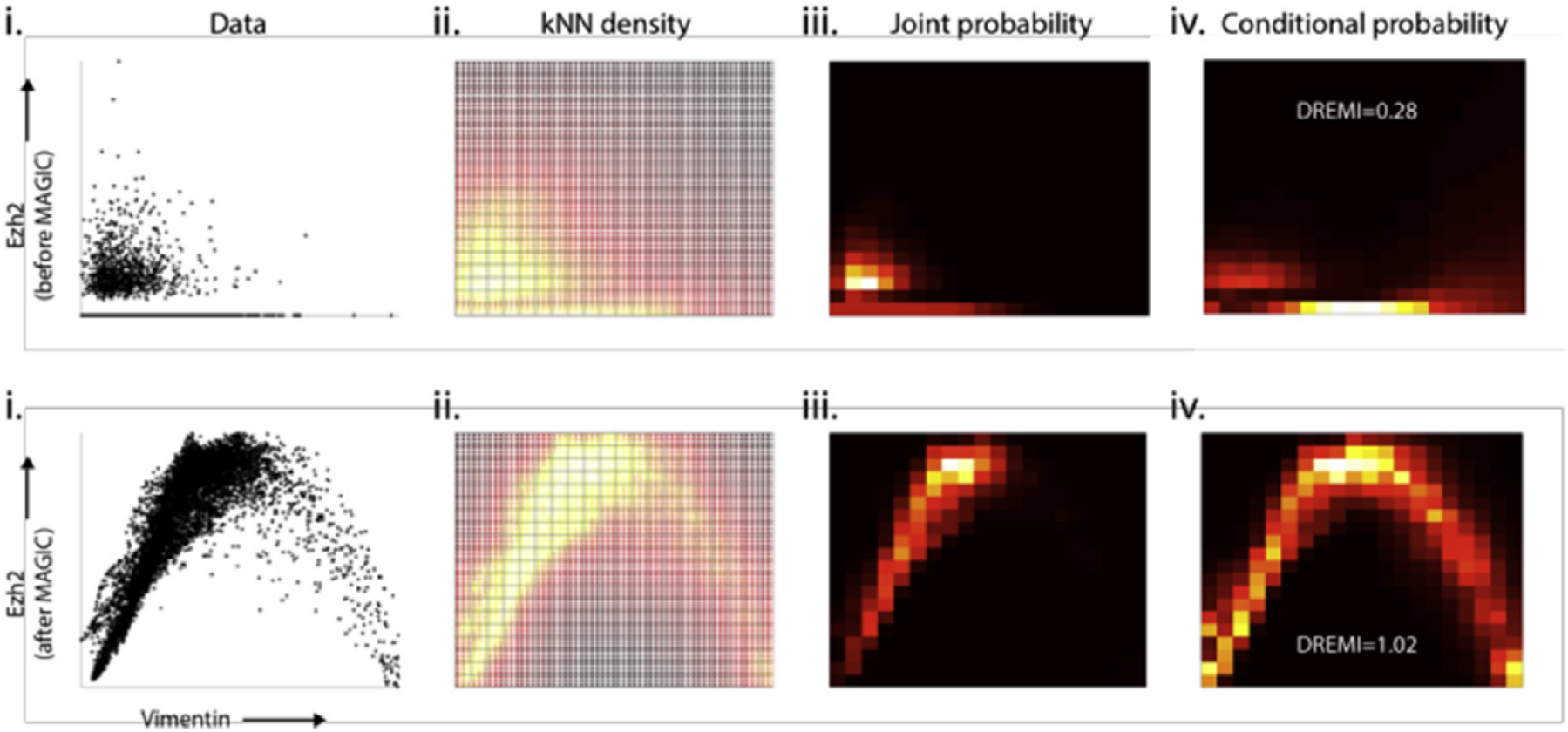
Relationship between features?

Non-linear Relationships

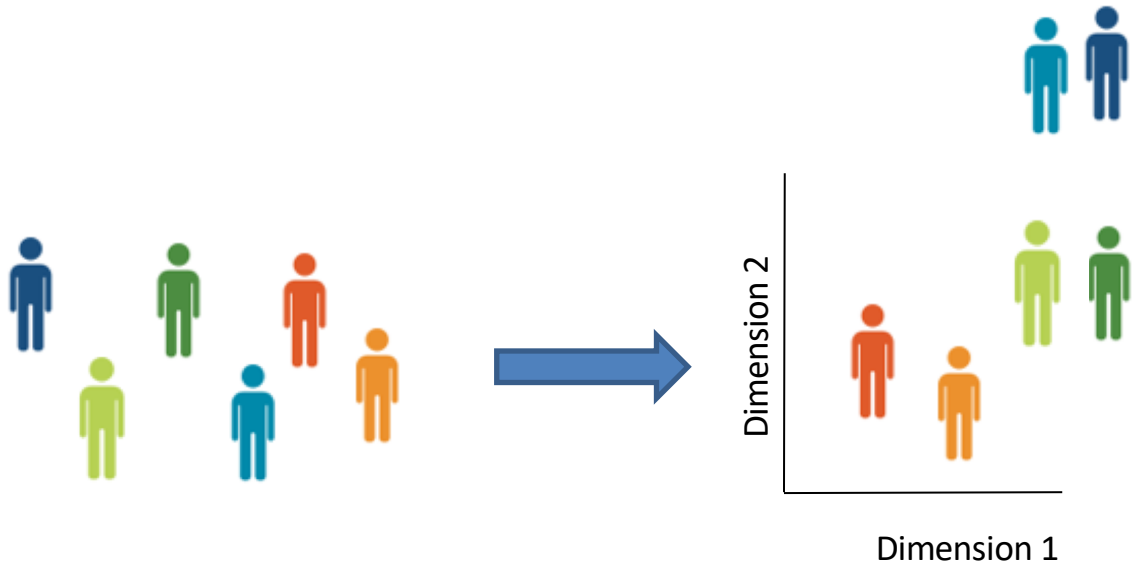




Mutual Information



Embedding reveals structure



Use high dimensional features and high throughput to understand shape of data

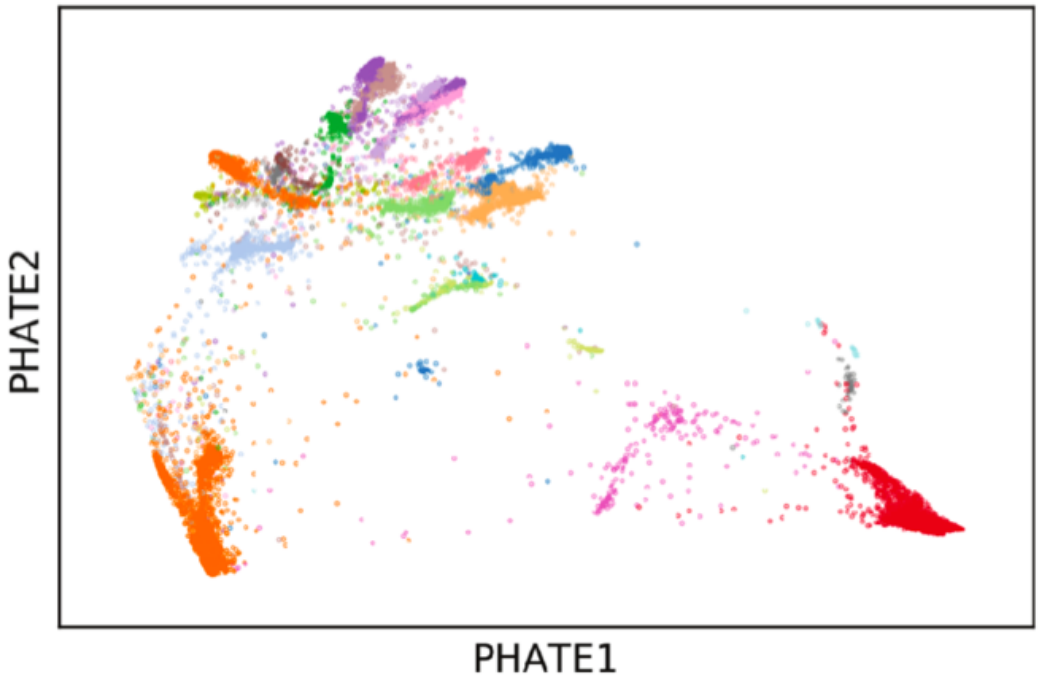
Cluster structure

Excellent Responders

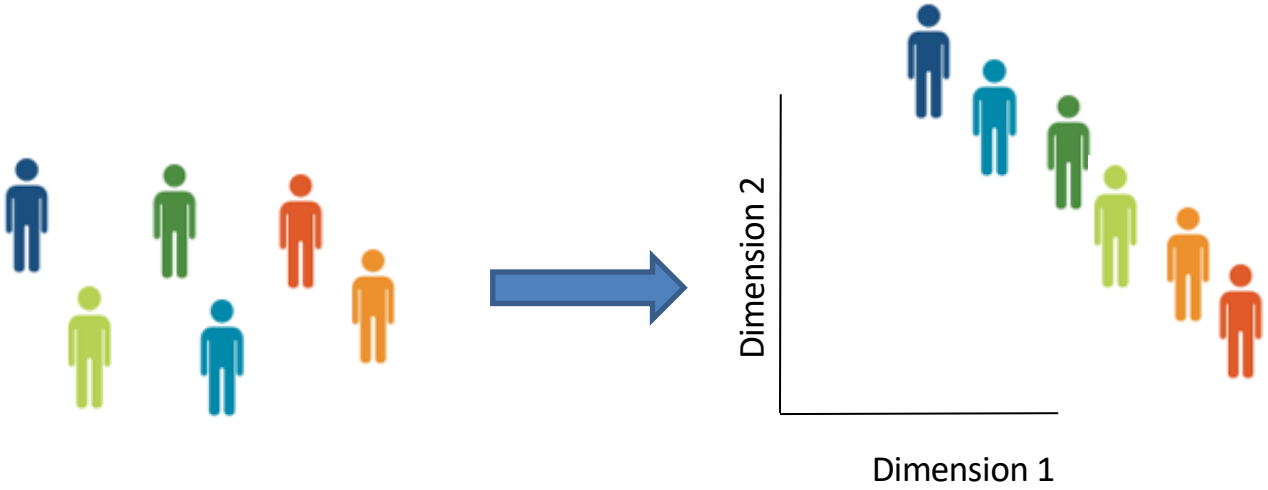
Fair Responders

Poor responders

Retinal Bipolar Cells



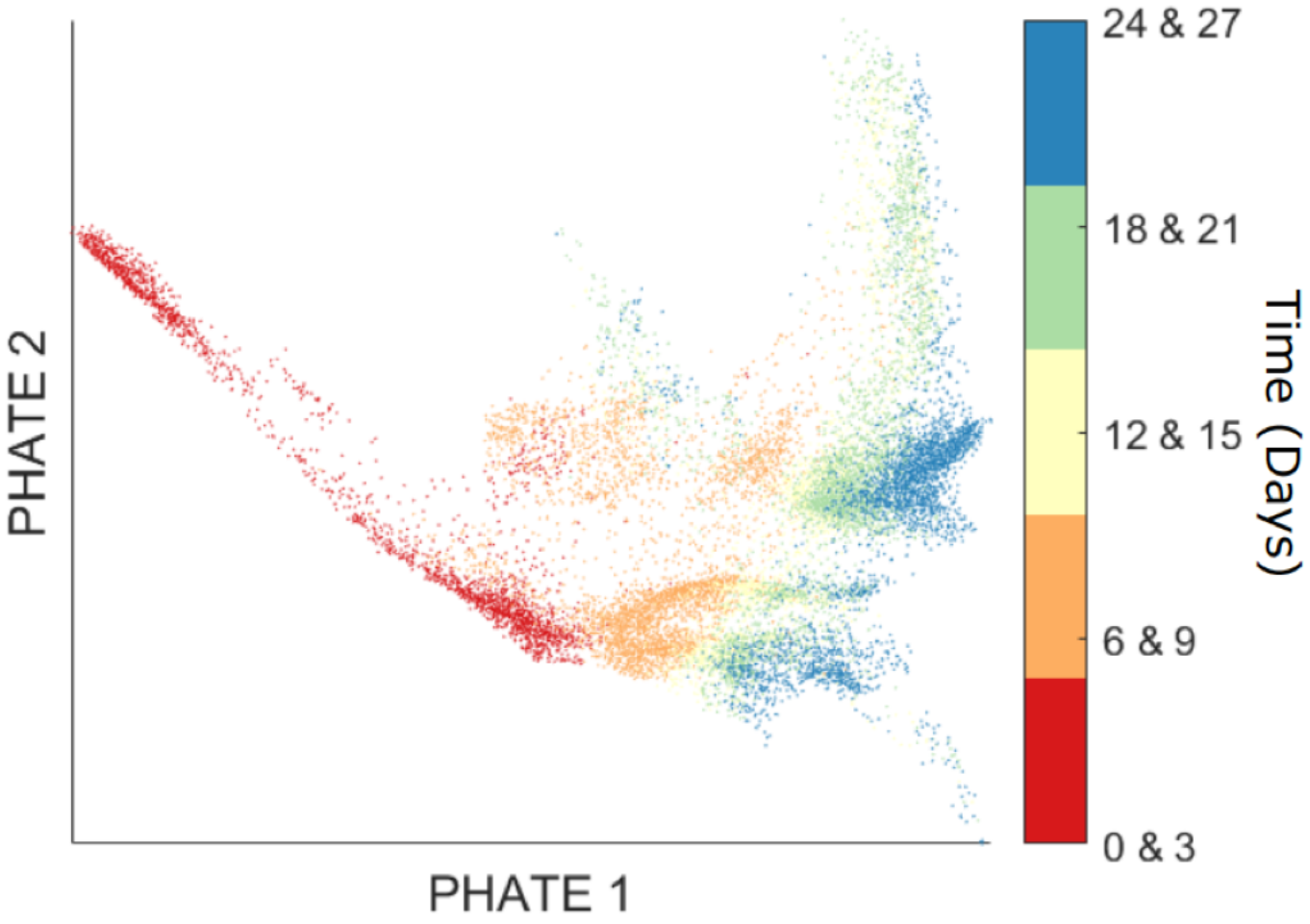
Progression continuum



Use high dimensional features and high throughput to understand shape of data

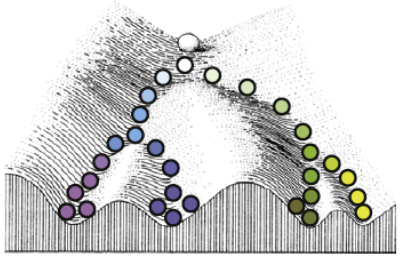
Pseudotime

Progressions

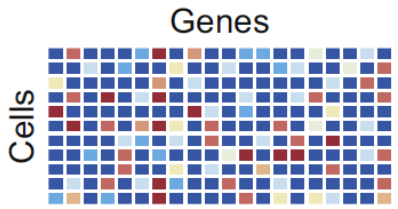


Manifold Learning

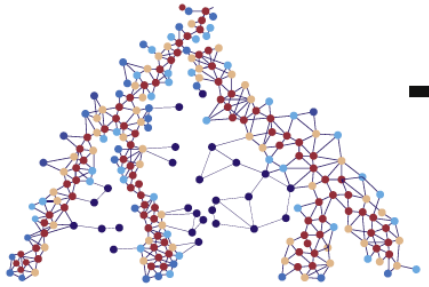
Cells are sampled from an underlying manifold



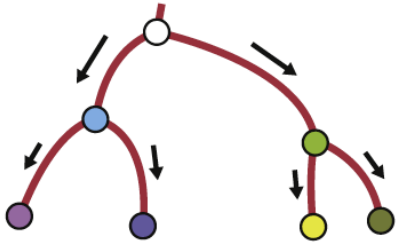
Each cell is represented by a vector of gene expression



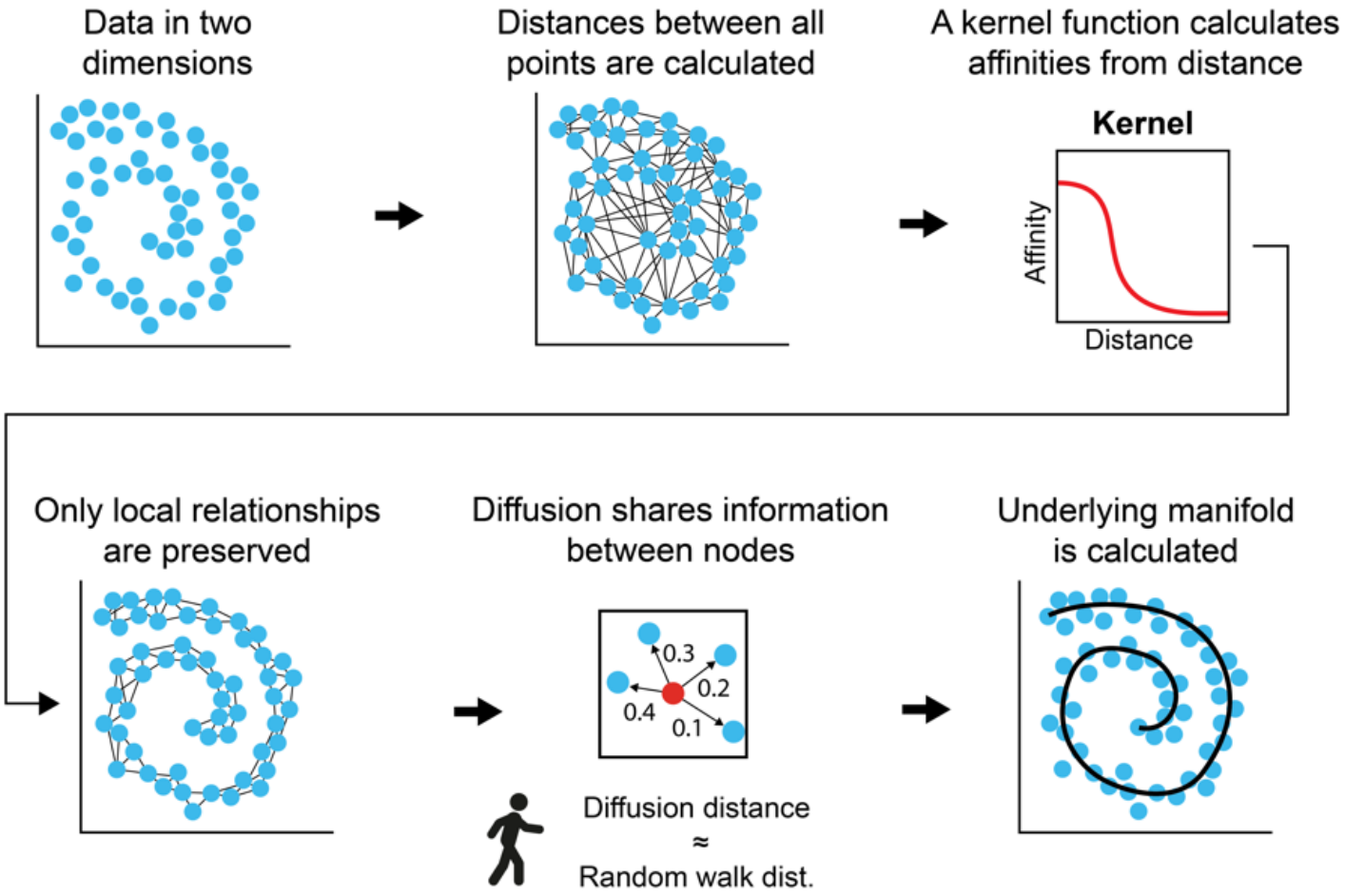
Neighborhood structure of the observations is identfied



The latent manifold is learned from the data

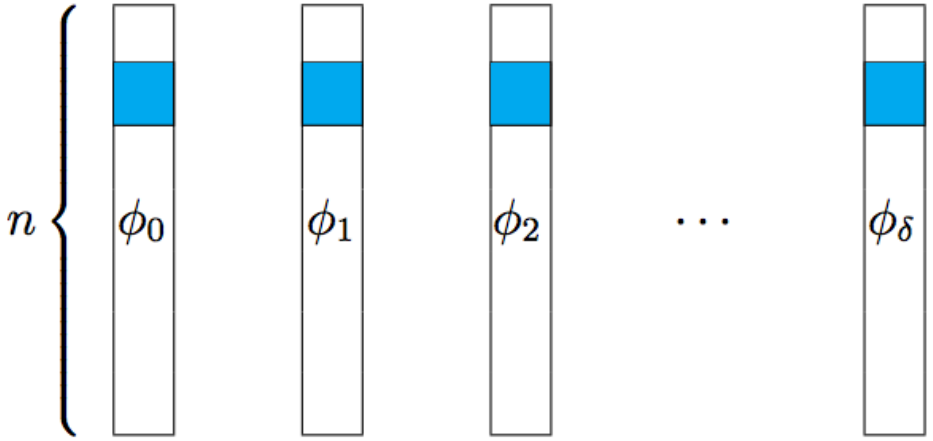


Understanding the shape of data



Diffusion Maps

$$1 = \boxed{\lambda_0} \geq \boxed{\lambda_1} \geq \boxed{\lambda_2} \geq \cdots \geq \boxed{\lambda_\delta} > 0$$

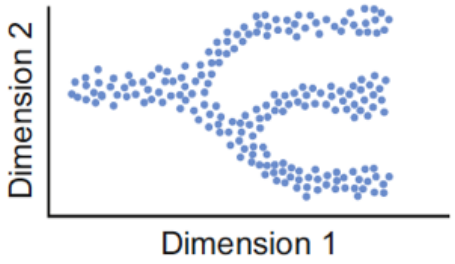


$$x \mapsto \Phi(x) \triangleq [\lambda_0\phi_0(x)\,,\, \lambda_1\phi_1(x)\,,\, \lambda_2\phi_2(x)\,,\, \ldots\,,\, \lambda_\delta\phi_\delta(x)]^T$$

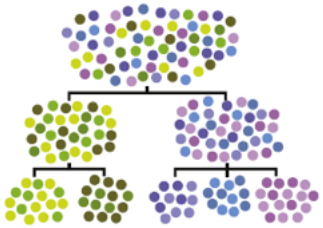
[Coifman, Lafon, ACHA 2006]

Analysis Tasks

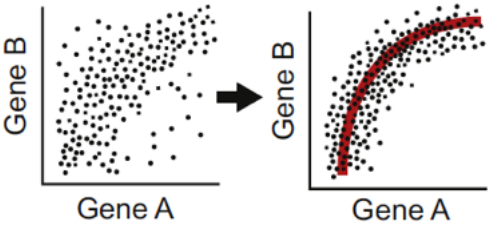
Vizualization



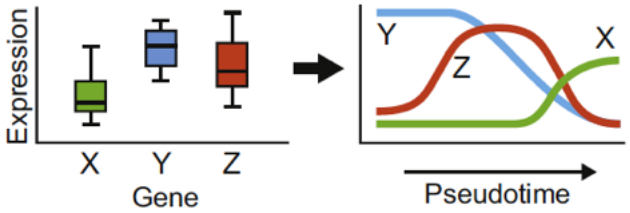
Clustering



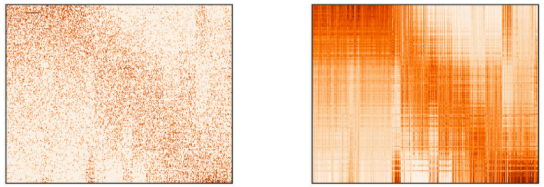
Denoising



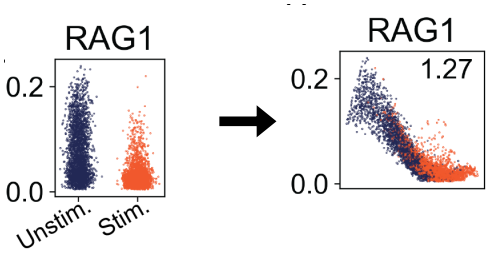
Pseudotime analysis



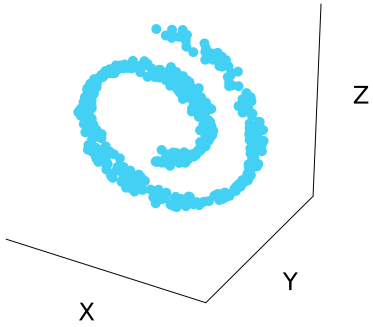
Imputation - MAGIC



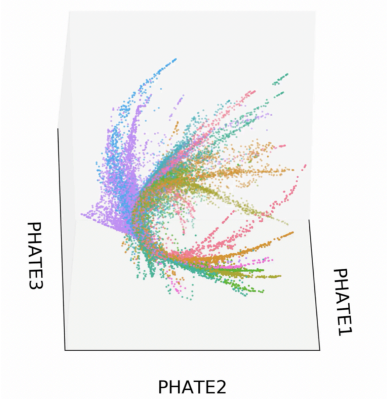
Experimental analysis - MELD



Manifold Learning



Visualization - PHATE



Scalable analytics - SAUCIE