Geometry of large data point clouds in high dimensions: examples from biology

Andrei Zinovyev

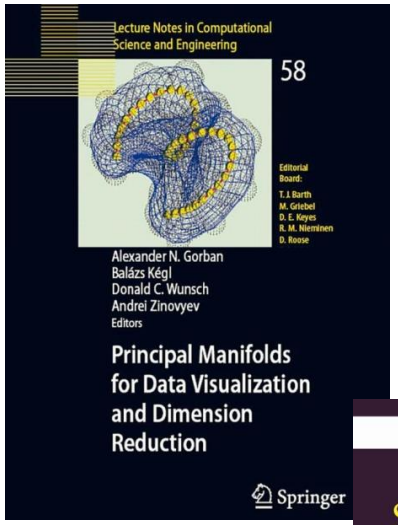
Institut Curie - INSERM U900 - PSL Research University / Mines ParisTech

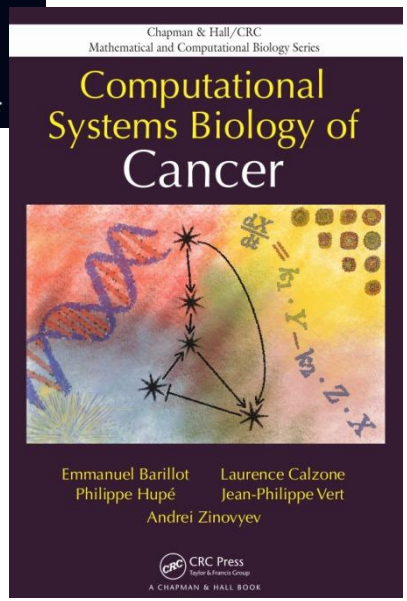
Computational Systems Biology of Cancer

About myself

https://auranic.github.io/

- Master degree in Theoretical physics (Cosmology)
- PhD in Machine Learning (principal manifolds, elastic maps)
- HDR in biology





- Post-doc at Institut des Hautes Etudes Scientifiques (IHES) (chemical kinetics, invariant manifolds, math biology)
- Since 2005: Scientific coordinator of Computational Systems Biology of Cancer group at Institut Curie
- Learning in high dimensions for cancer biology (dimensionality reduction, classification, etc.)

institutCurie

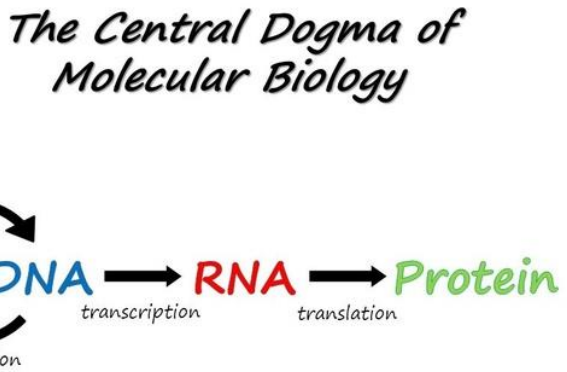
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

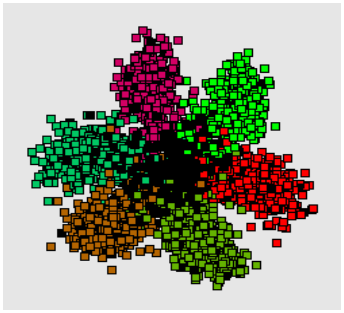
Modern molecular biotechnology is one of the main providers of large-scale real-life datasets (and related questions)

- OMICS data: systematic global measurements of a biological sample
  - Gen**omics** (all DNA in a sample), also epigen**omics** (all states of DNA, methylation, histones, 3D conformation)
  - Transcript**omics** (all RNA in a sample)
  - Prote**omics** (all proteins in a sample)
- Main technology : sequencing, mass-spectrometry
- *OMICs profile* of a biological sample – typically from 10^3 to 10^6 features (can be much more), *p>>n* problematics

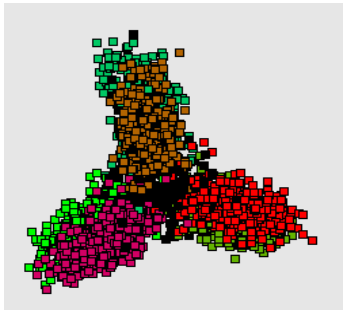


7 cluster structure of bacterial genomes

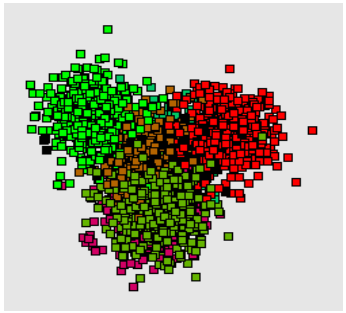
(Gorban A., Popova T., Zinovyev A. *Physica A,* 2005)



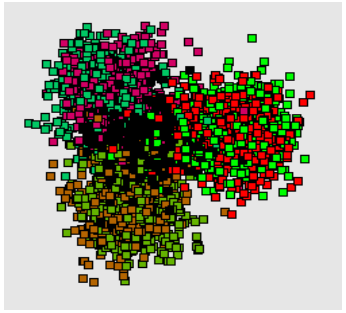
Streptomyces coelicolor



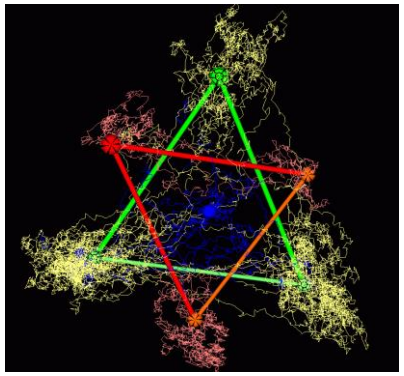
Fusobacterium nucleatum



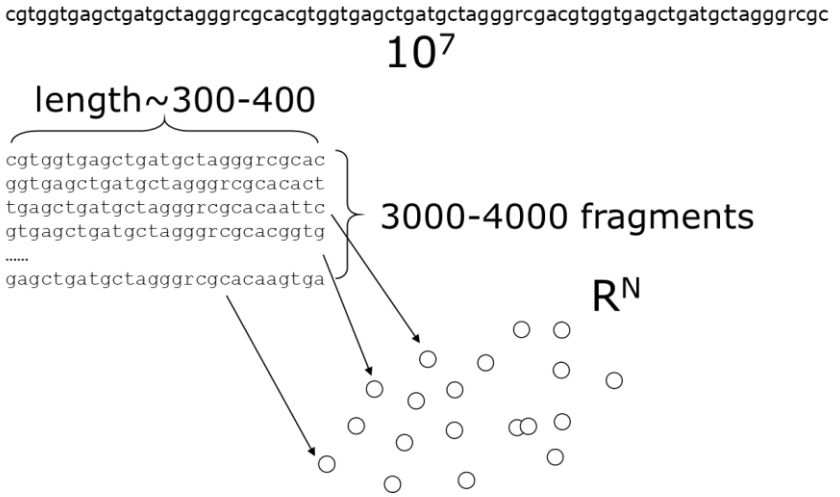
Bacillus halodurans

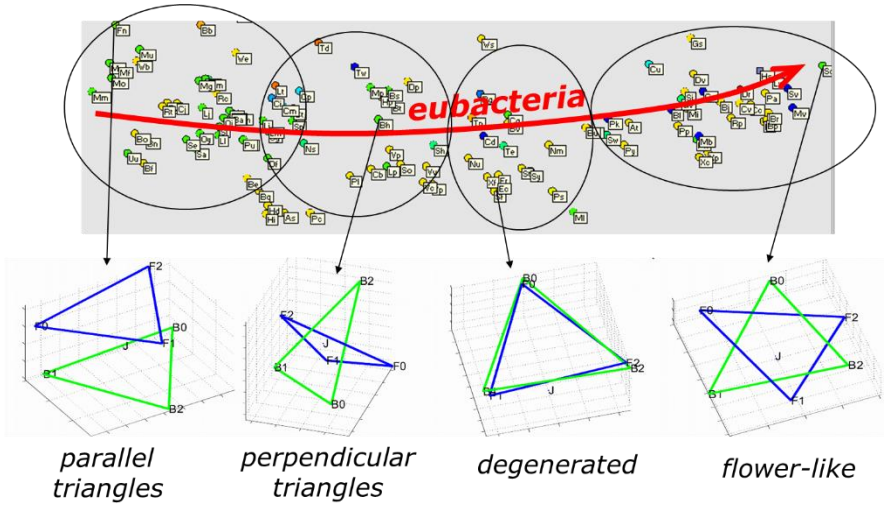


Erchericha coli



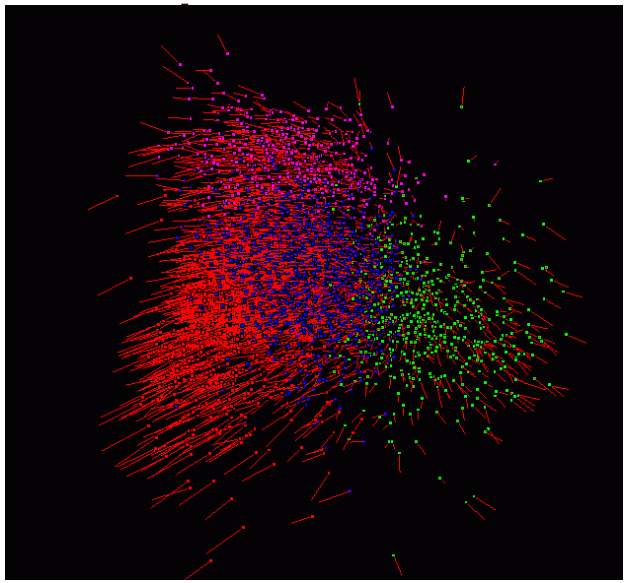
Visualization of random walk along the genome





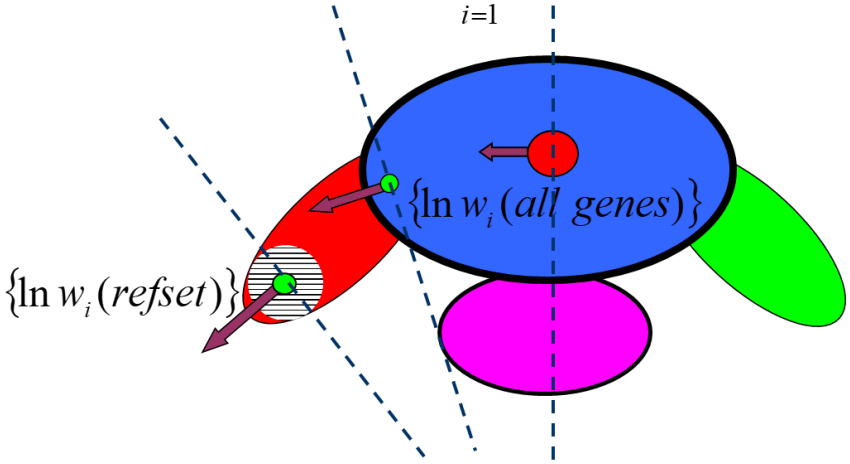
Stereotypical structure of codon frequency distribution in fast-growing bacteria and eukaryotes

(Carbone, Zinovyev, Képès, Bioinformatics, 2003;
Carbone, Képès, Zinovyev, Mol Biol Evol, 2005)



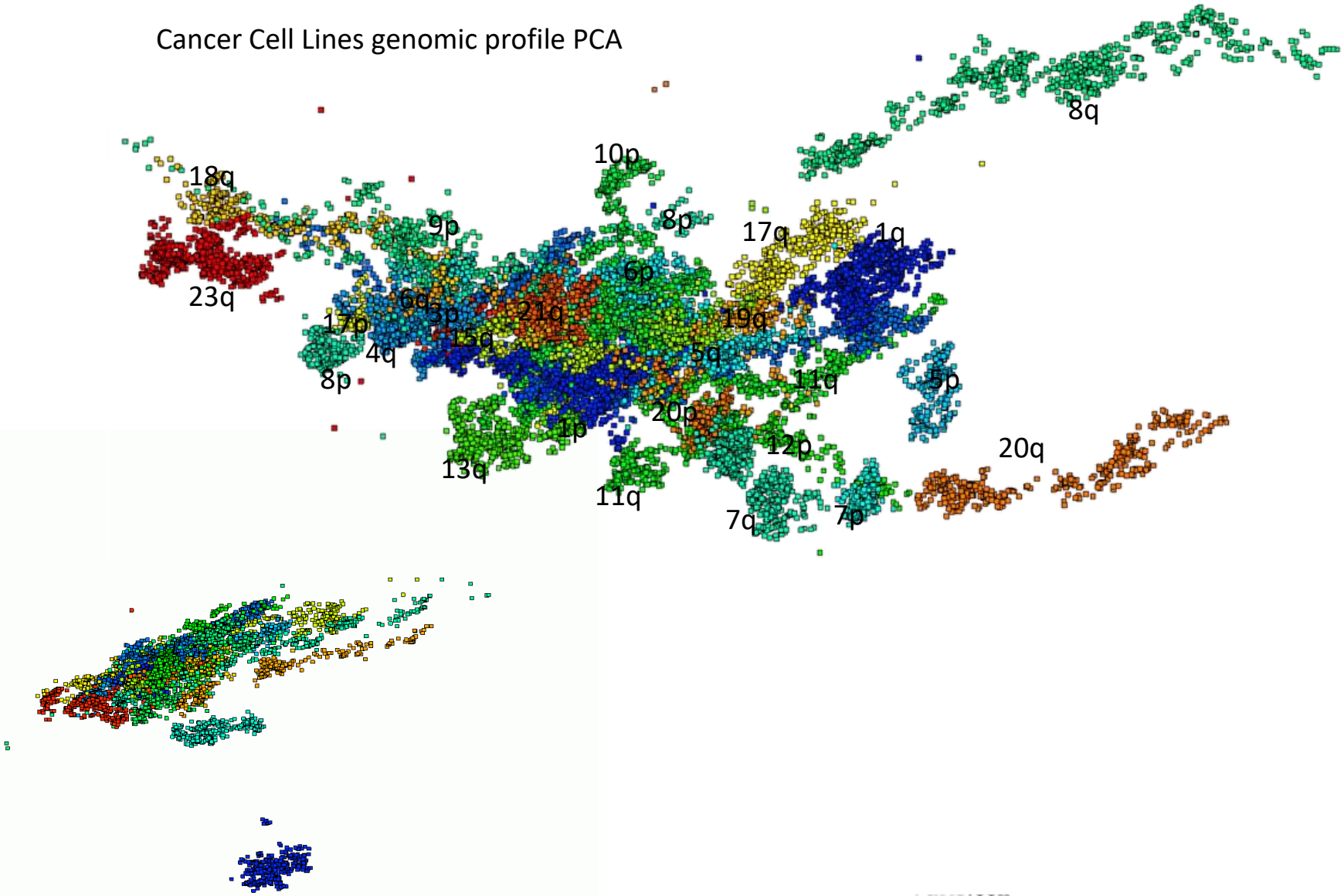
- **Genes of class I (most of)**
- **Genes of class II (higly expressed)**
- **Genes of class III (unusual)**
- **Genes of class IV (hydrophobic proteins)**

$$\ln CAI(gene) = \sum_{i=1}^{64} g_i \ln w_i = \langle g_i, \ln w_i \rangle$$

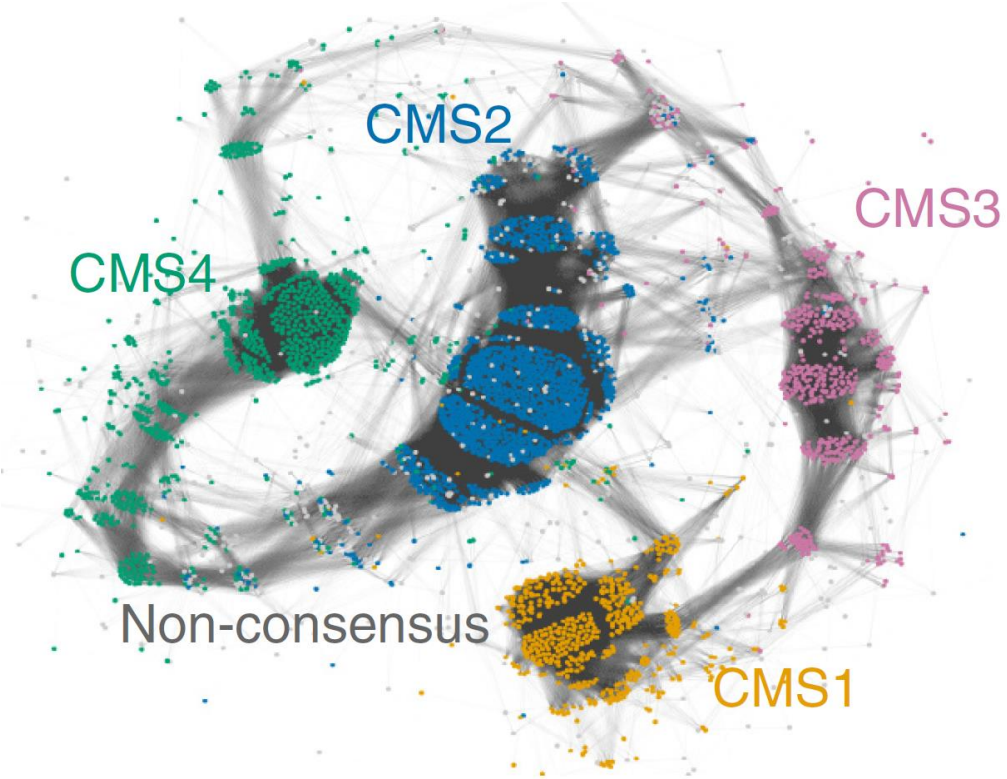


Geometry of the cancer genome copy number profiles

160 breast and ovary cancer cell lines, principal component analysis of SNPs

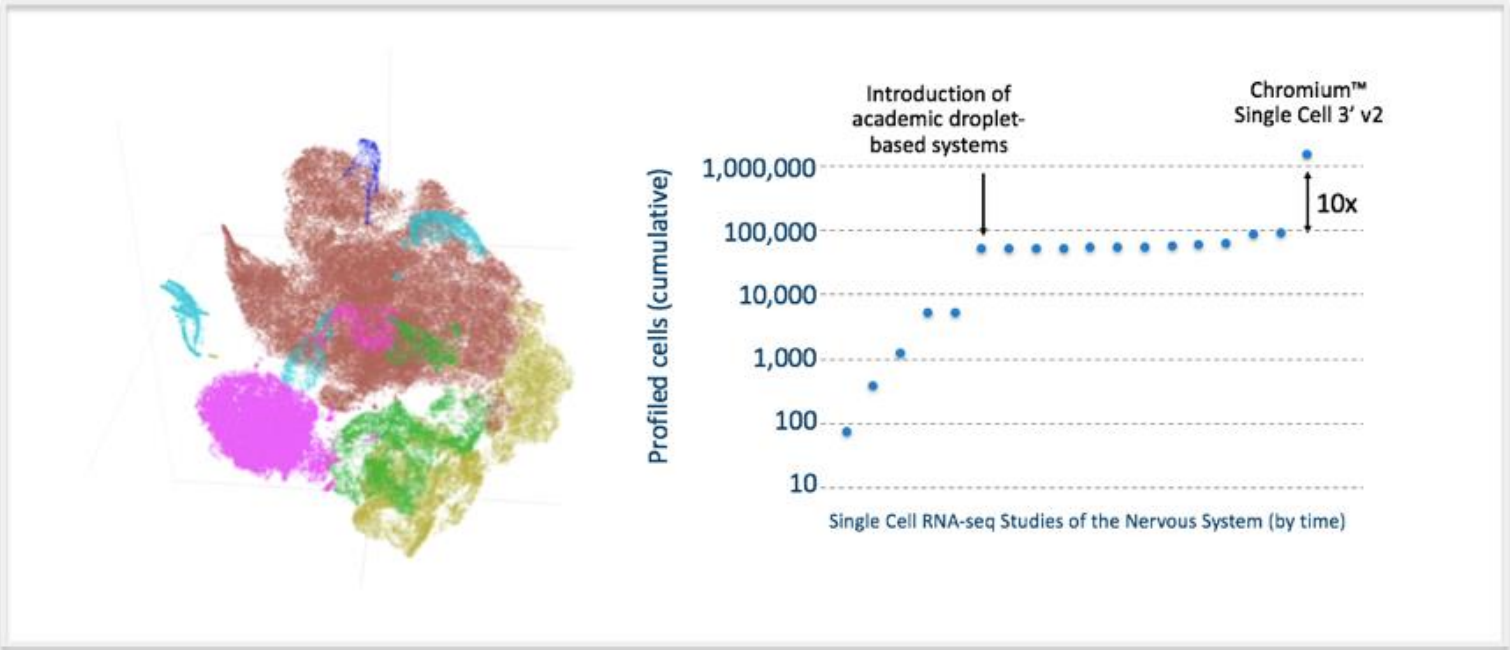


Molecular cancer subtypes



4000 cases of colorectal cancer transcriptomes
(from Guinney et al, Nat Med, 2015)

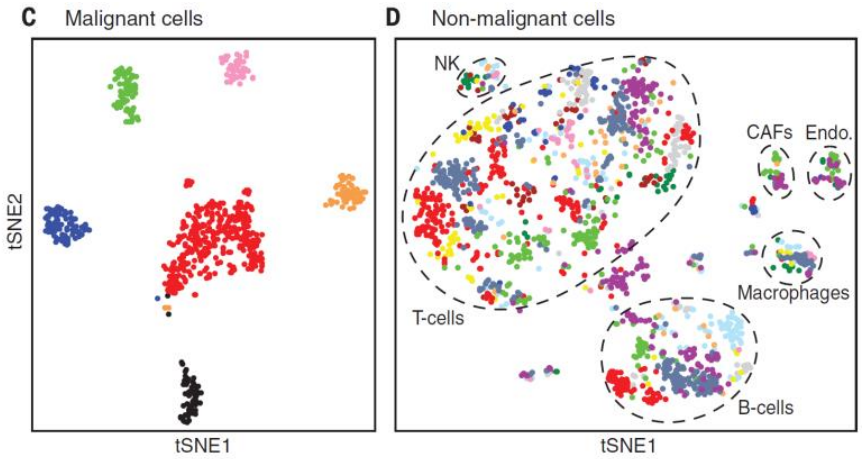
Single cell RNASeq data



- Measurements are not limited by availability of samples
- Each biological sample can be represented as a cloud of points in multidimensional space
- Importance of data exploratory/geometrical methods

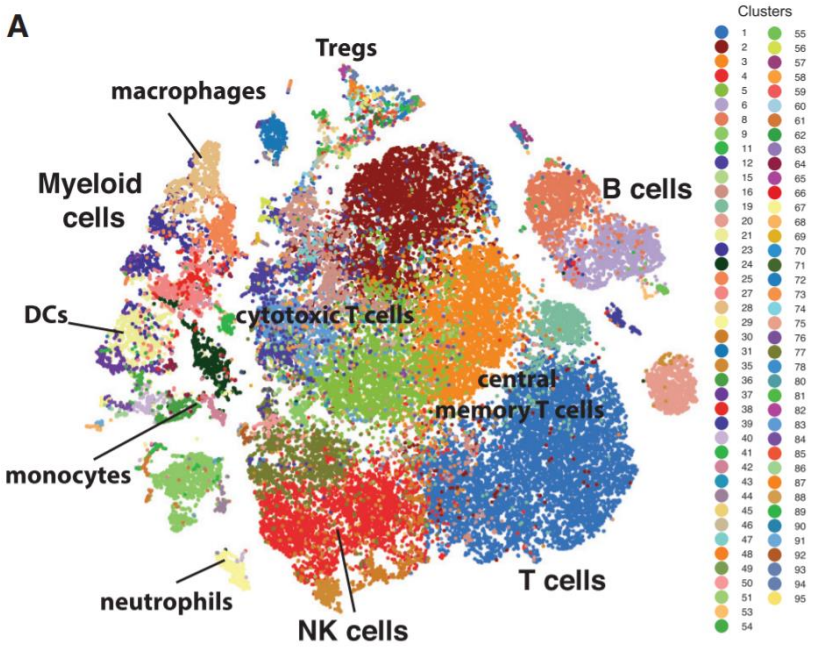
Examples from cancer biology

8000 cells



Tirosh et al, 2015, Science

45000 cells



Aiziz et al, 2018, Cell

institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Single cell data cartography of Planarian
(Plass et al, 2018)

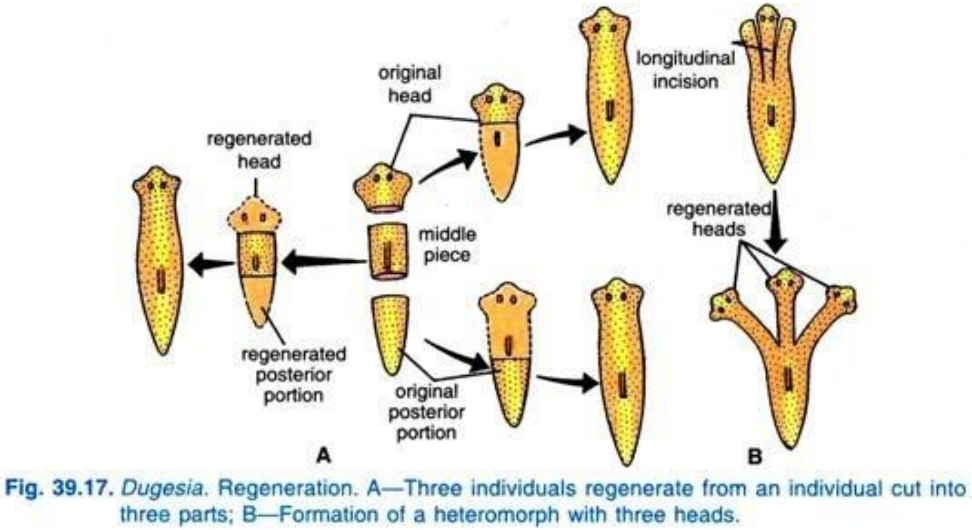
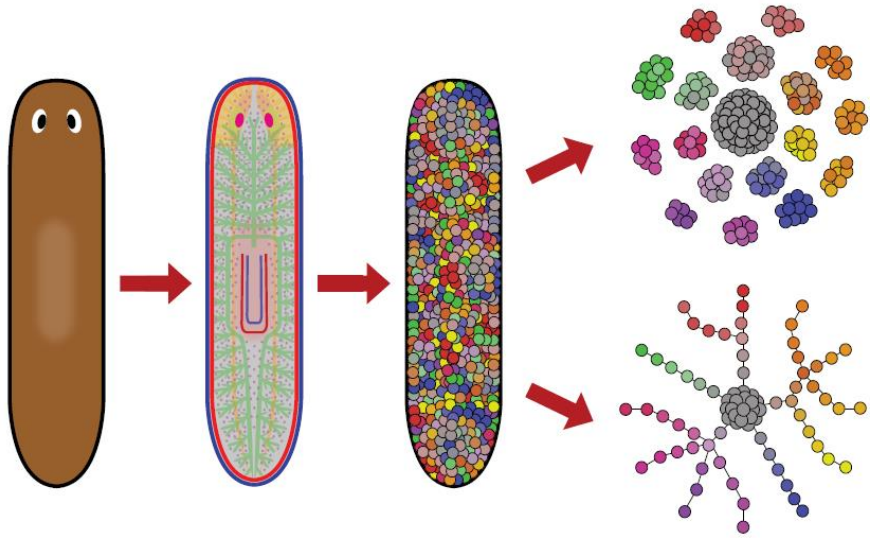
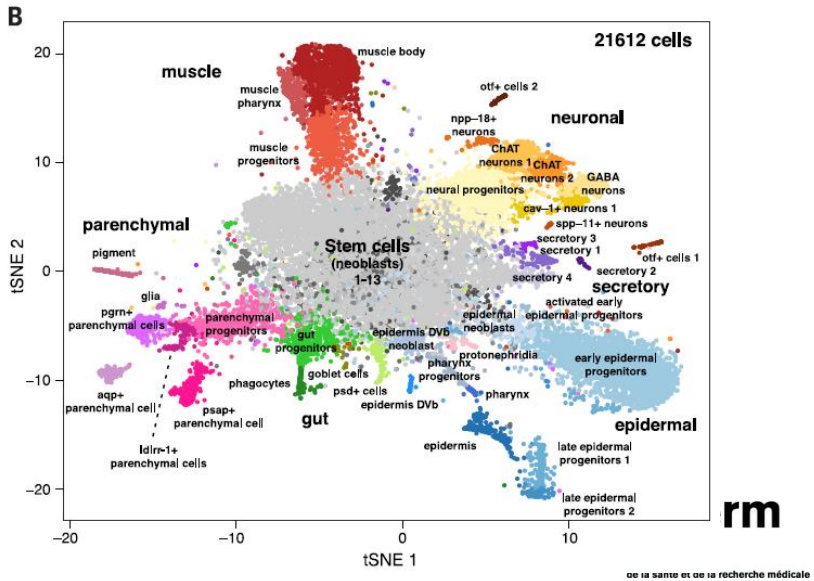
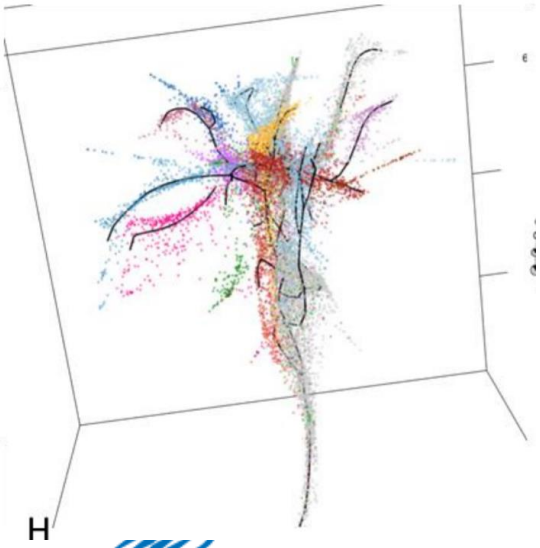


Fig. 39.17. *Dugesia*. Regeneration. A—Three individuals regenerate from an individual cut into three parts; B—Formation of a heteromorph with three heads.







institu rm

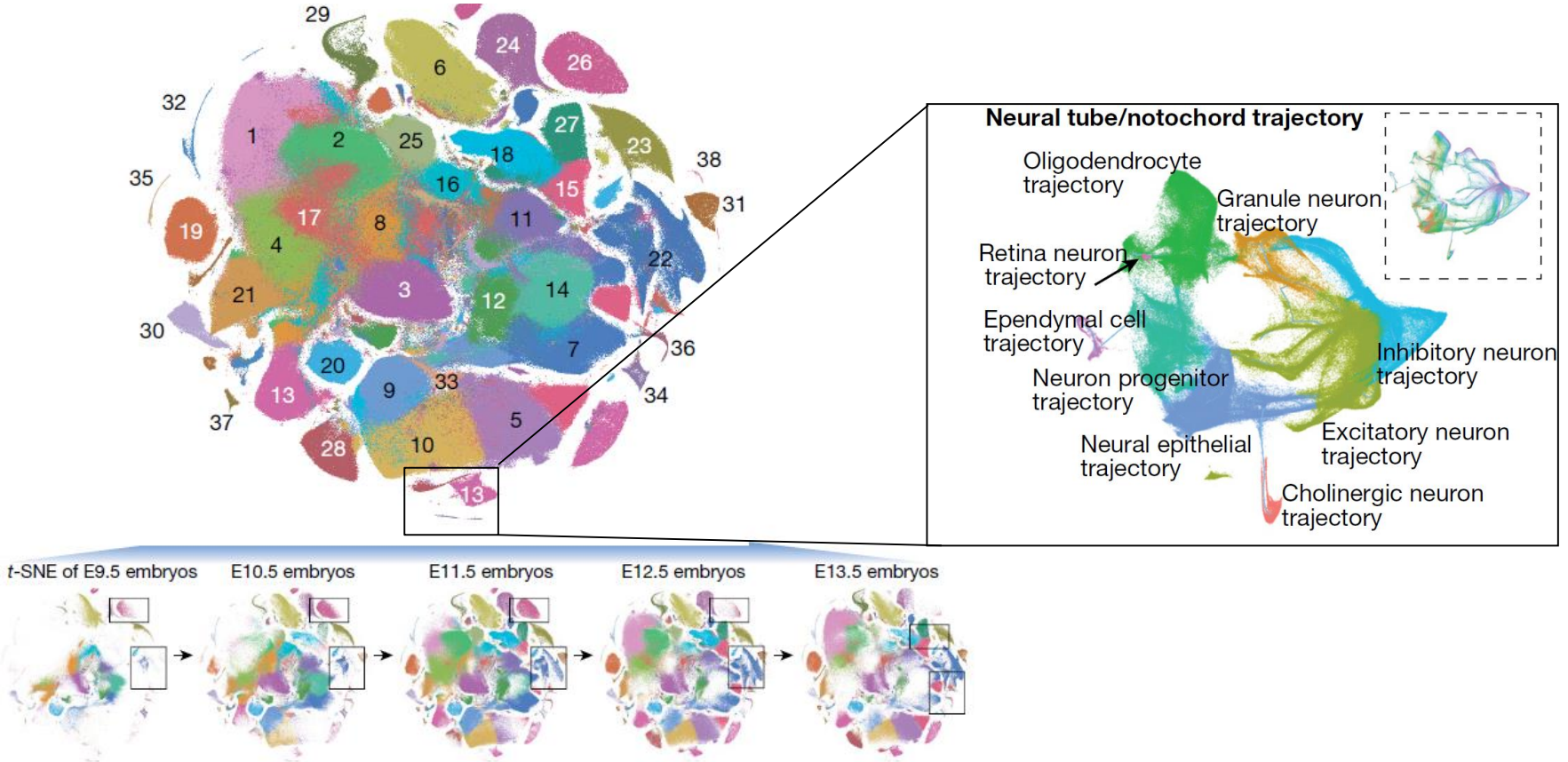
de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Mouse organogenesis at single cell level (~2 millions of cells, Cao et al, 2019, Nature)

tSNE, Louvain clustering of kNN graph



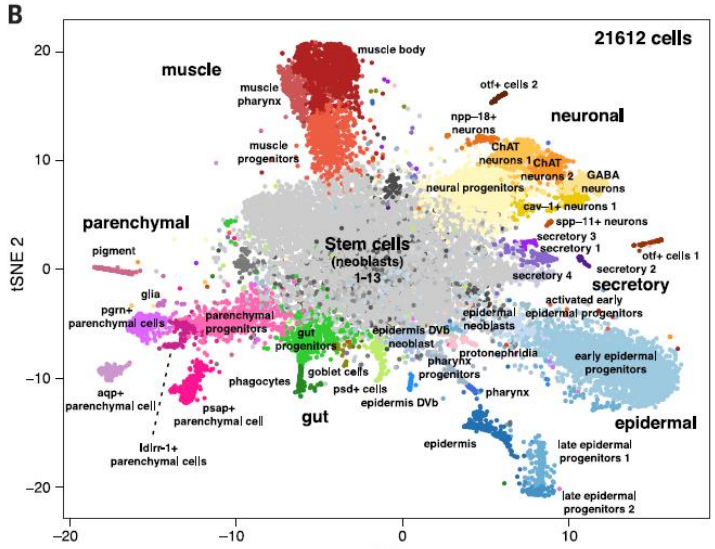
institutCurie

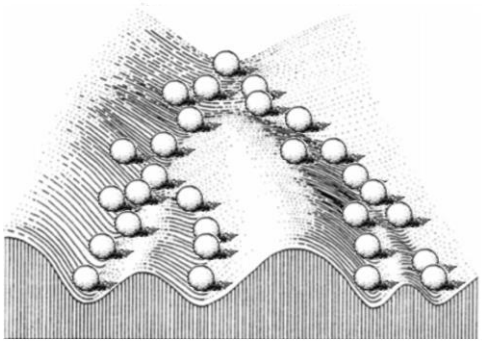
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

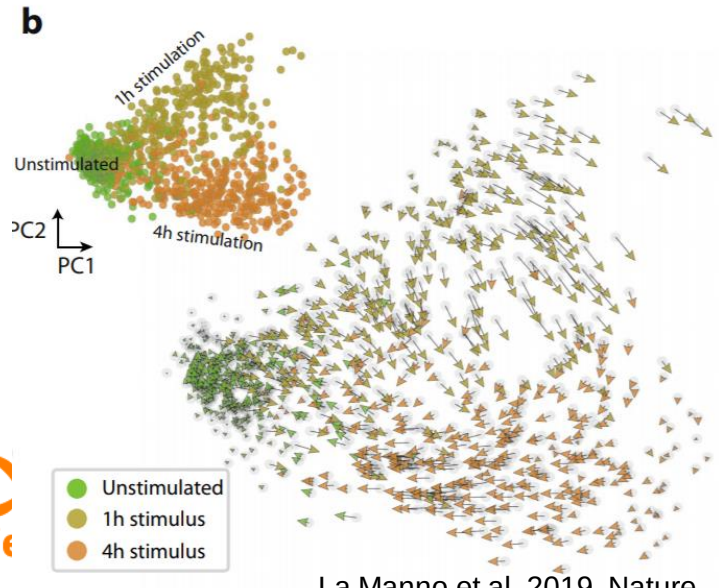
MINES ParisTech

PSL RESEARCH UNIVERSITY

Understanding cell fate decisions







1) visualize landscape
2) Introduce intrinsic coordinates of the landscape

La Manno et al, 2019, Nature

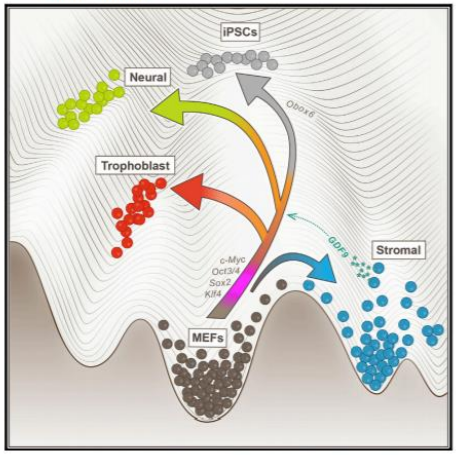
Using mathematics to understand the single cell trajectories

Resource

Cell

Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

Graphical Abstract



Authors

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, ..., Rudolf Jaenisch, Aviv Regev, Eric S. Lander

Correspondence

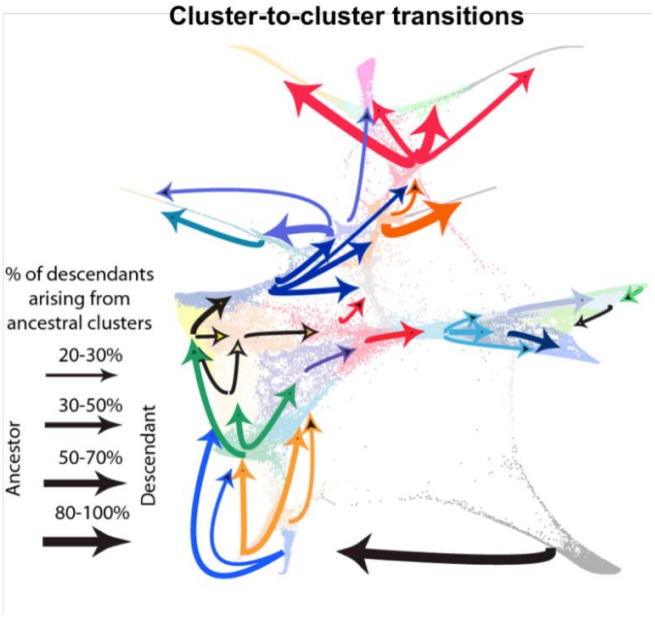
jianshu@broadinstitute.org (J.S.), aregev@broadinstitute.org (A.R.), lander@broadinstitute.org (E.S.L.)

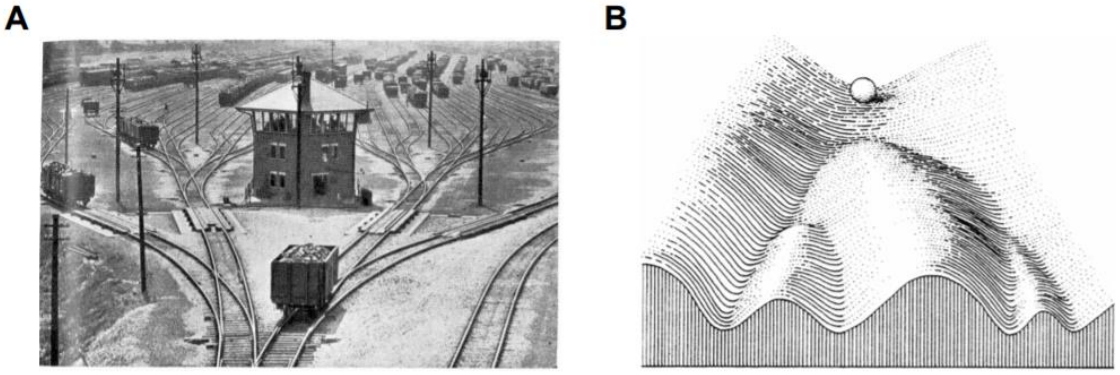
In Brief

Application of a new analytical approach to examine developmental trajectories of single cells offers insight into how paracrine interactions shape reprogramming.

Highlights

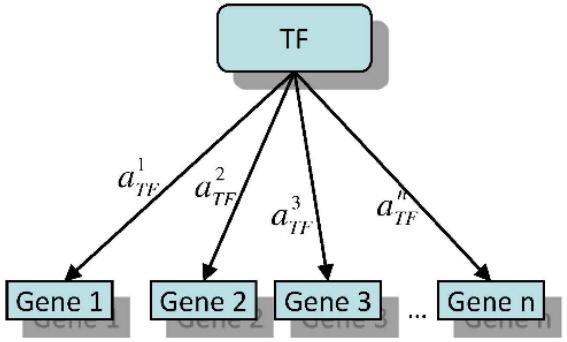
- Optimal transport analysis recovers trajectories from 315,000 scRNA-seq profiles



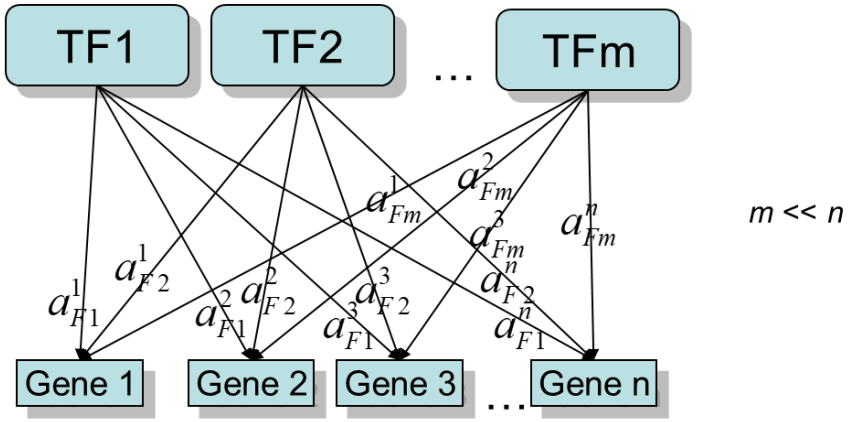


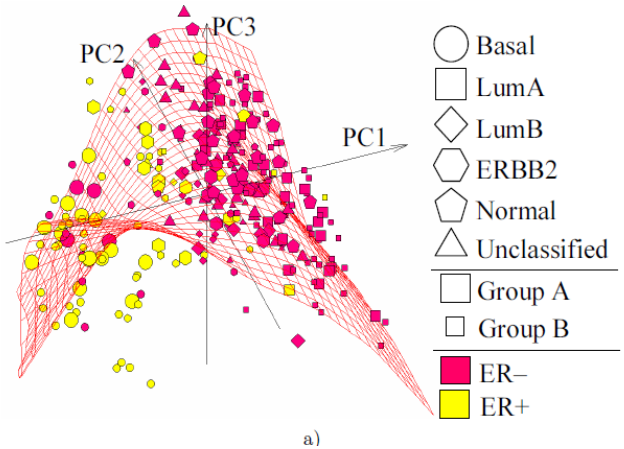
Unsupervised analysis of omics data

One factor, linear response

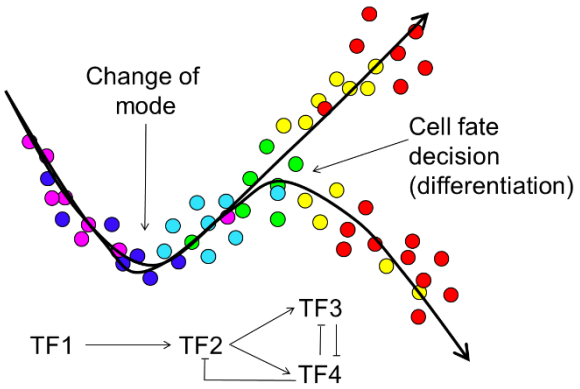


Many factors, linear response



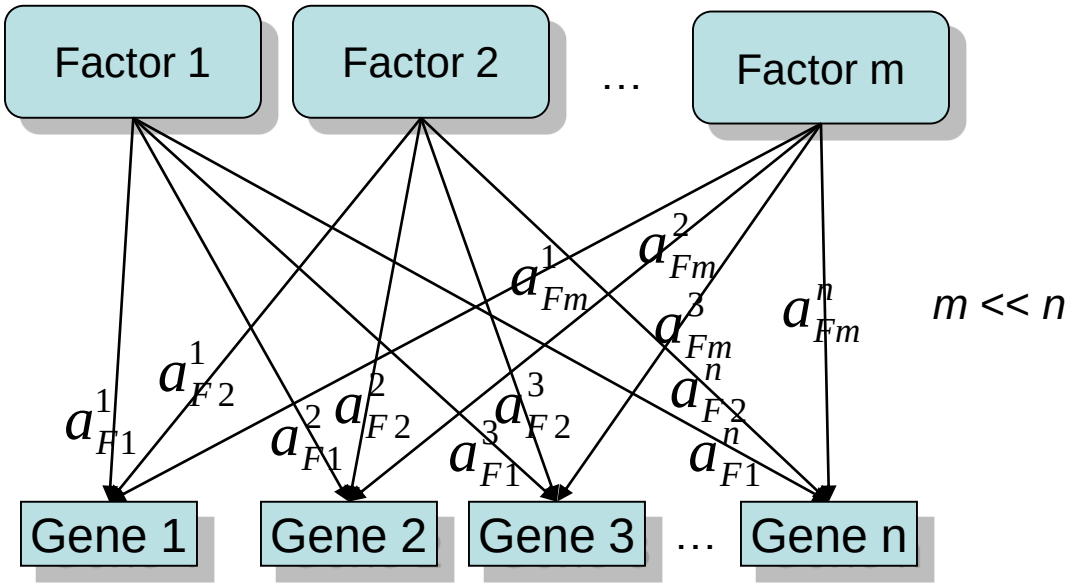


Non-linear manifold learning

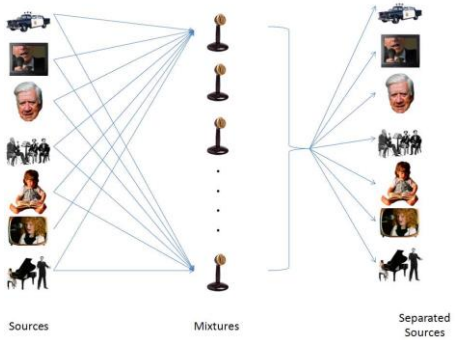


Non-linear branching data approximators (e.g., *principal trees*)

Mixture of independent sources as the simplest representation of regulation



$m << n$



$$Expression(gene\ g, sample\ s) \approx \sum_{i=1}^{m} a^g_{Fi} \cdot Activity(Fi, sample\ s)$$

Principal component analysis (PCA):

«Orthogonality constraint» (just mean-square approximation)

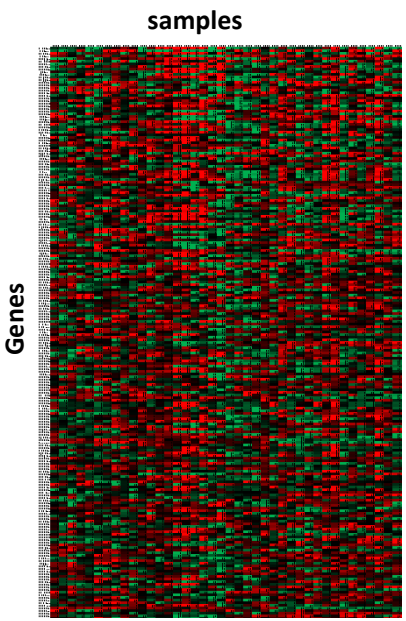
Non-negative matrix factorization (NMF):

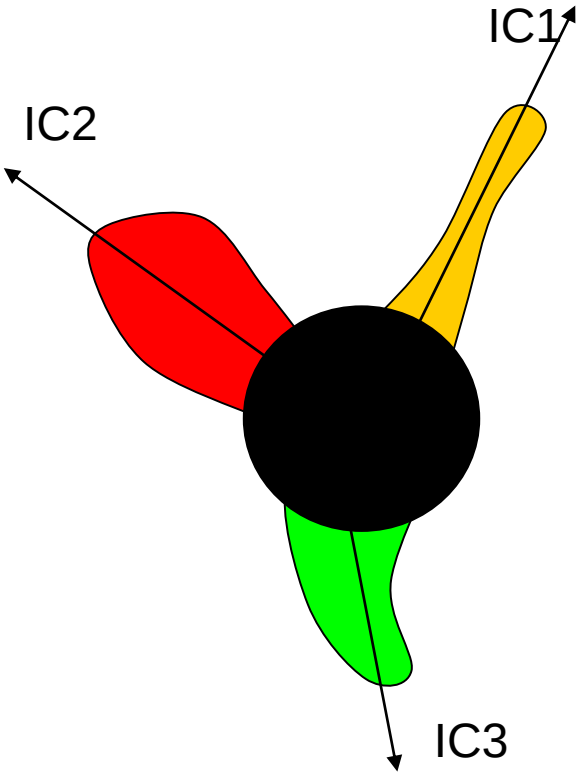
a_{ij} and Activities should be non-negative. Sparsity effect.

Independent Component Analysis (ICA):

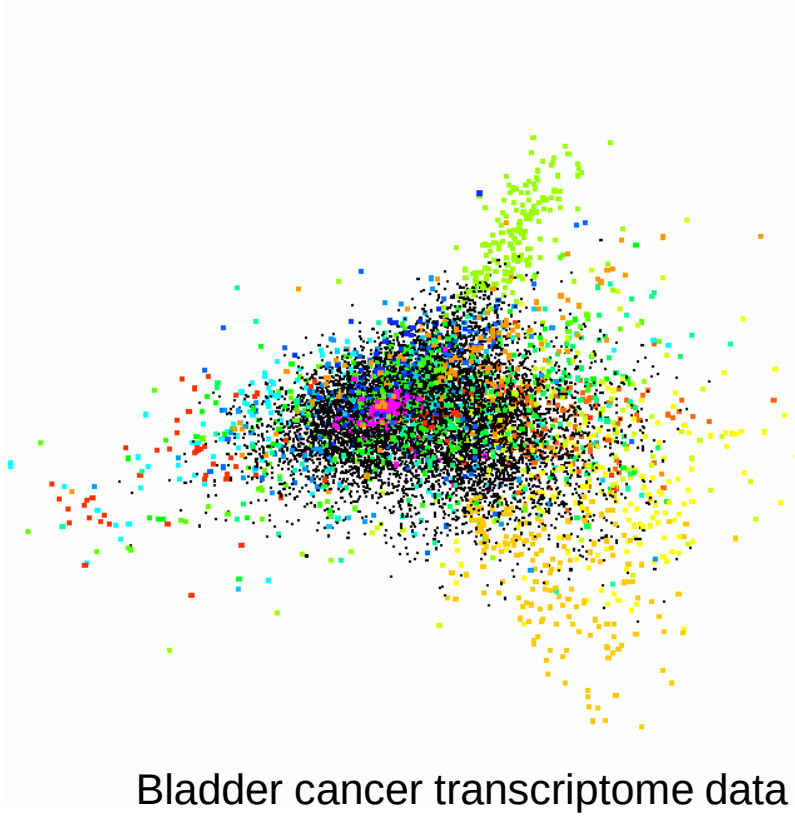
Assumption of statistical independence of Factor activities

Independent Component Analysis in the gene space





IC is a vector (direction) in the gene space



Bladder cancer transcriptome data (Stransky, 2007), projection from 81-dimensional space

institutCurie

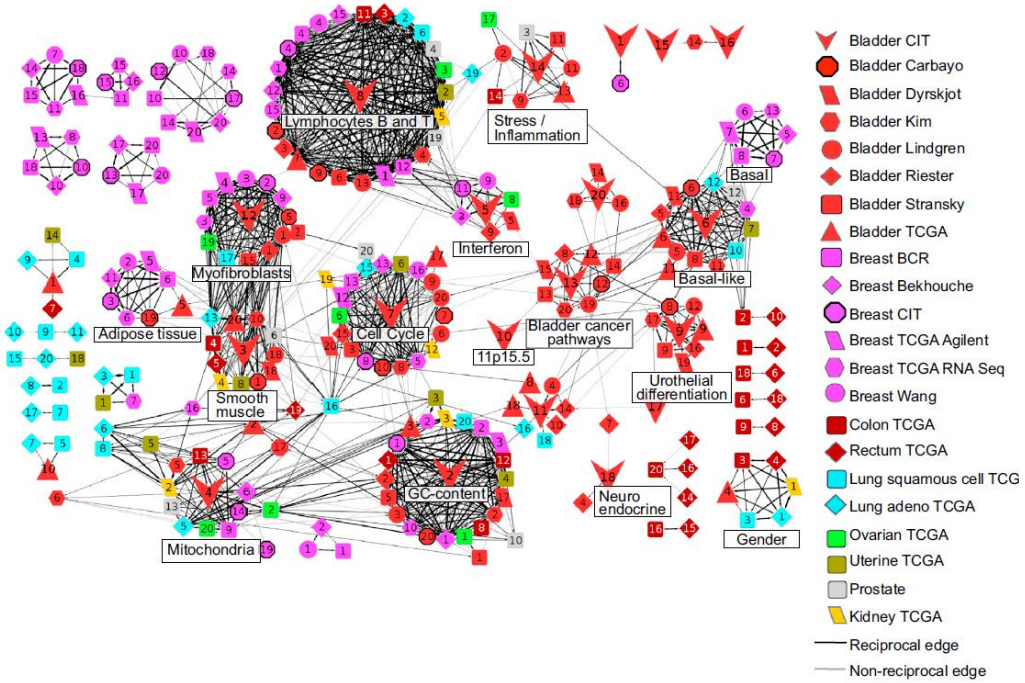
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

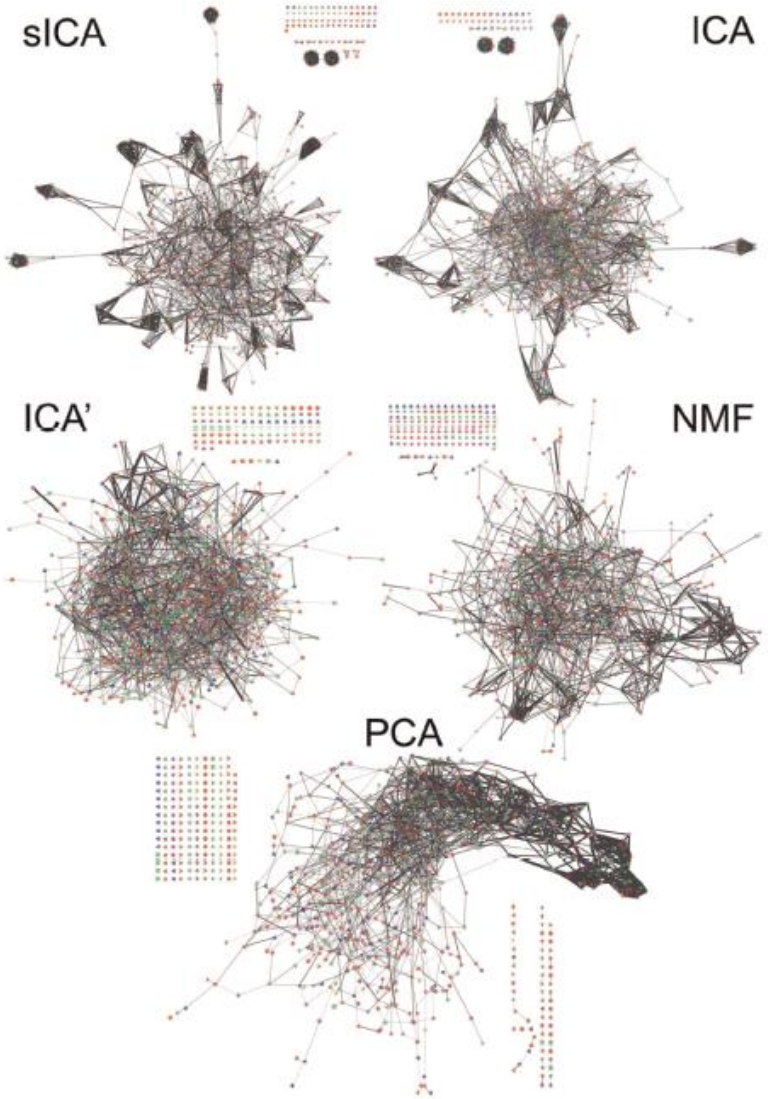
MINES ParisTech

PSL RESEARCH UNIVERSITY

Stabilized ICA in the gene space generalizes well for transcriptomic data

Biton et al, 2014; Kairov et al, 2017; Cantini et al, 2019



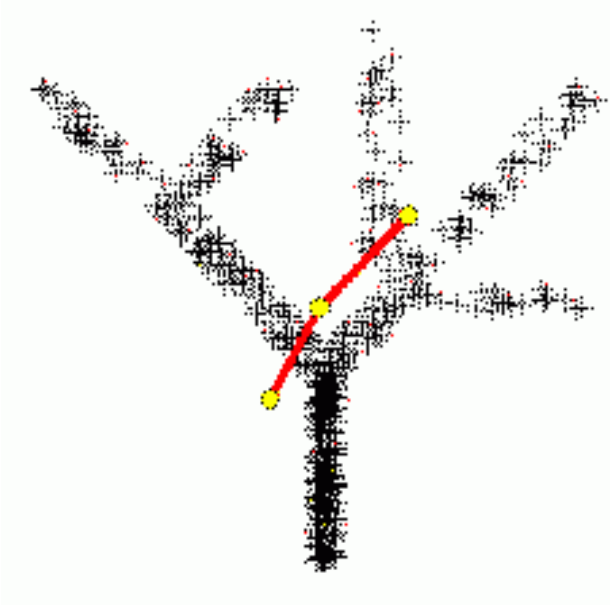


institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

Elastic principal graphs (ElPiGraph)

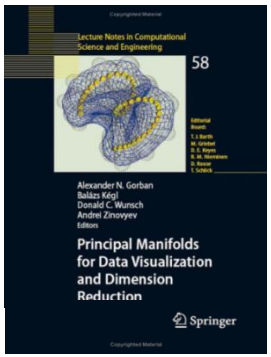
(Gorban&Zinovyev,2007; Zinovyev&Mirkes, 2013; Gorban&Zinovyev, 2010)
book Gorban, Kegl, Wunch, Zinovyev, LNSC, 2008)



- History:
  - principal curves were introduced by Trevor Hastie in 1989: "**Principal curves** are smooth one-dimensional curves that pass through the middle of a p-dimensional data set, providing a nonlinear summary of the data. They are nonparametric, and their shape is suggested by the data"
  - principal graphs were introduced by Kegl and Kryzak in 2002 for finding skeleton **graph** of handwritten characters (not a general approach)
  - **elastic principal graphs based on graph grammars were introduced in 2007 (**Gorban and Zinovyev, *Applied Mathematics Letters, 2007***)**
  - currently principal graphs are used in the analysis of single cell data as a part of MONOCLE 2 (reverse graph embedding method). Still based on kNN graph and require drastic dimension reduction.
  - Implementations in **MATLAB, R, Python, Scala, Java**

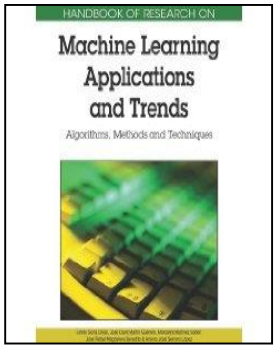
https://github.com/sysbio-curie/ElPiGraph.R
https://github.com/sysbio-curie/ElPiGraph.M
https://github.com/sysbio-curie/ElPiGraph.P

2000



2008

Principal graphs and manifolds
chapter

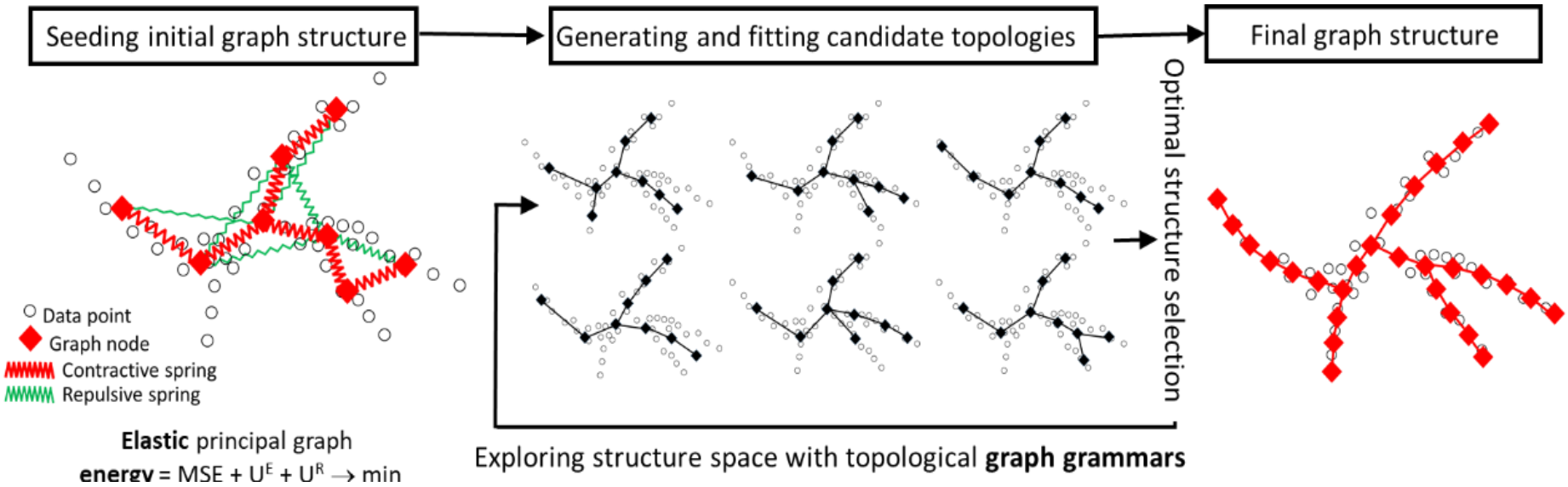


2010

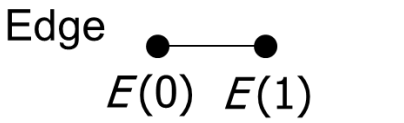
Elastic principal graphs (ElPiGraph)

(Gorban&Zinovyev,2007; Zinovyev&Mirkes, 2013;
Gorban&Zinovyev, 2010; Albergante et al, 2018; Chen et al, 2019)

book Gorban, Kegl, Wunch, Zinovyev, LNSC, 2008)

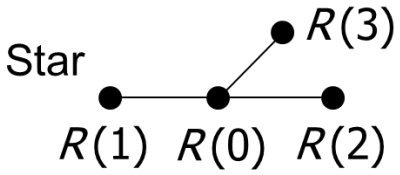


Penalty on **total length**:



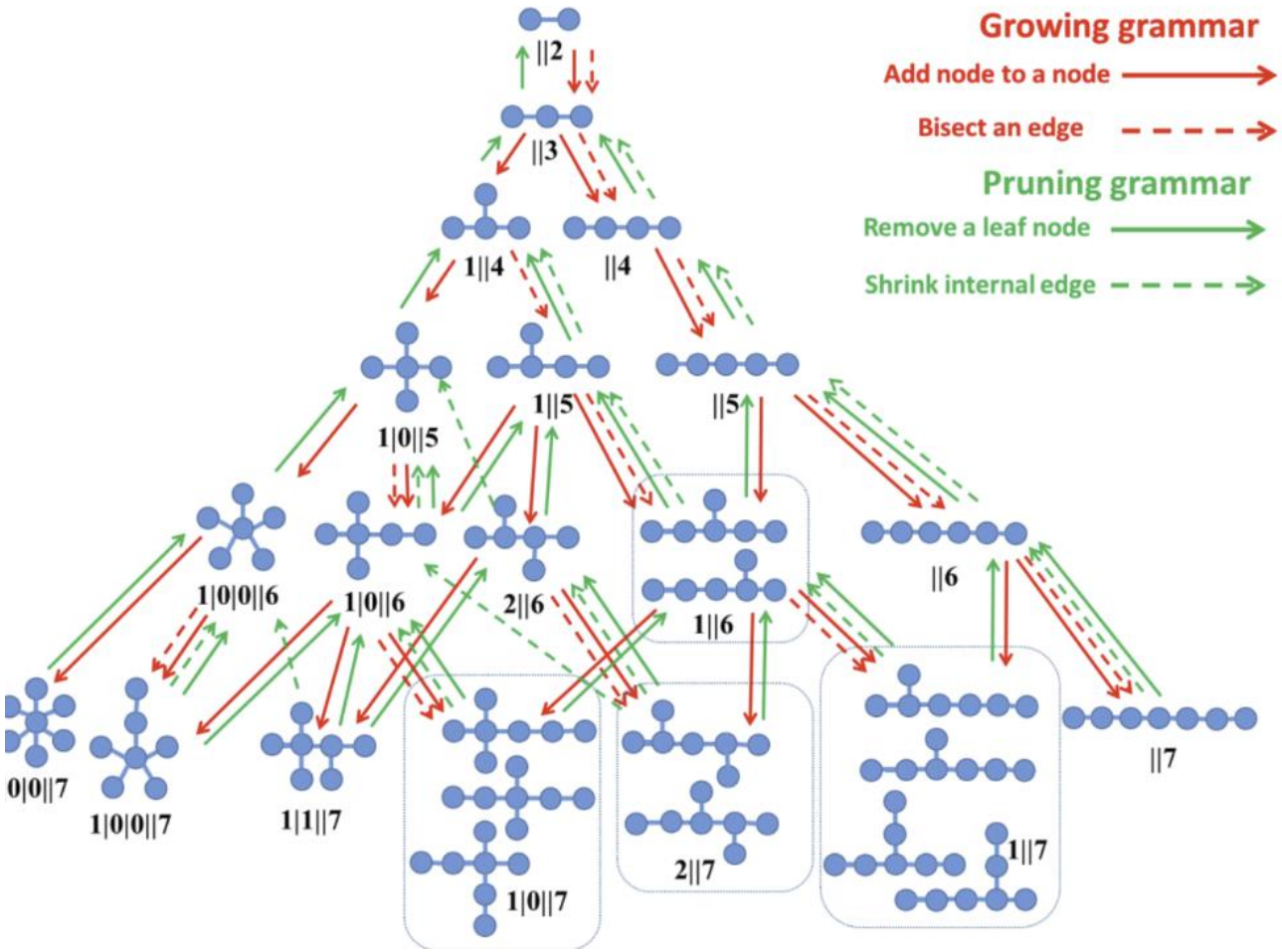
$$U^{(E)} = \sum_{i=1}^{s} \lambda_i \left\| E^{(i)}(1) - E^{(i)}(0) \right\|^2$$

Penalty on deviation from **harmonicity**:



$$U^{(R)} = \sum_{i=1}^{r} \mu_i \left\| R^{(i)}(0) - \frac{1}{k} \sum_{j=1..k} R^{(i)}(j) \right\|^2$$

Topological grammars



institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

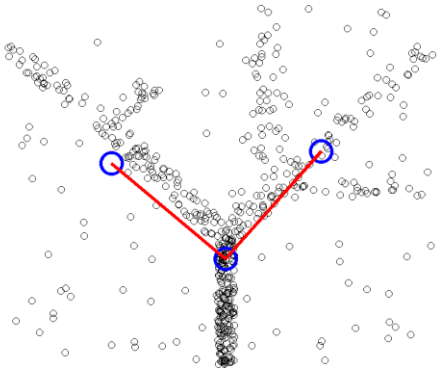
PSL RESEARCH UNIVERSITY

Robust principal graphs

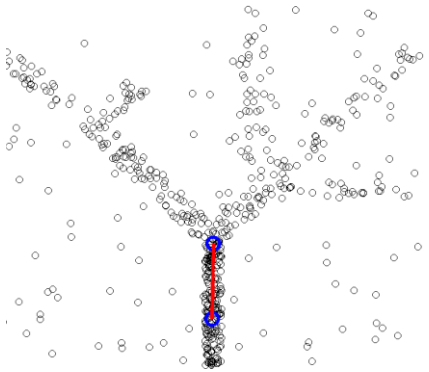
A.N. Gorban, E.M. Mirkes, A. Zinovyev.

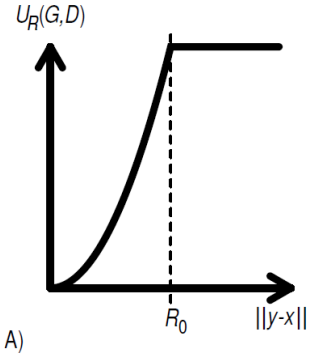
Robust principal graphs for data approximation. 2018.

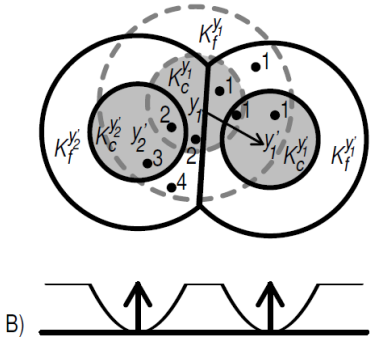
“Global” non-robust version



“Local” robust version
sees only close data points

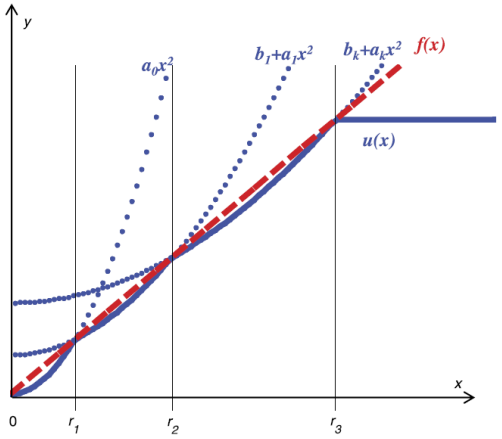






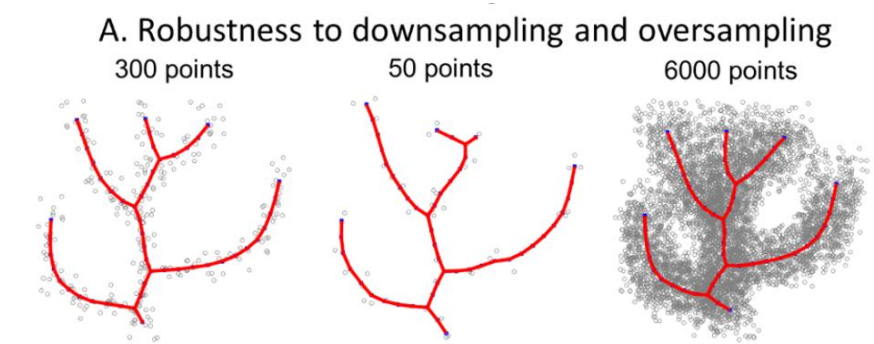
PQSQ (Piece-wise Sub-Quadratic) approach to robustness

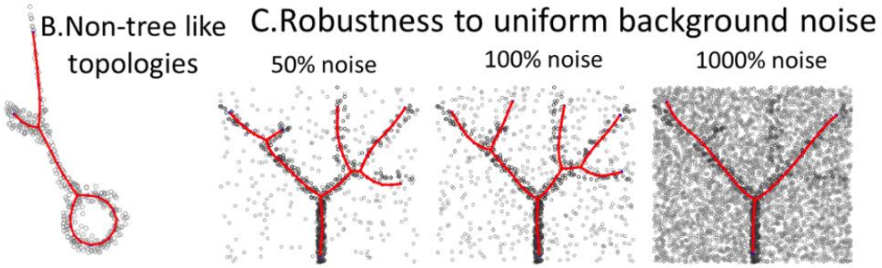
Gorban, Mirkes, Zinovyev, Neural Networks, 2016

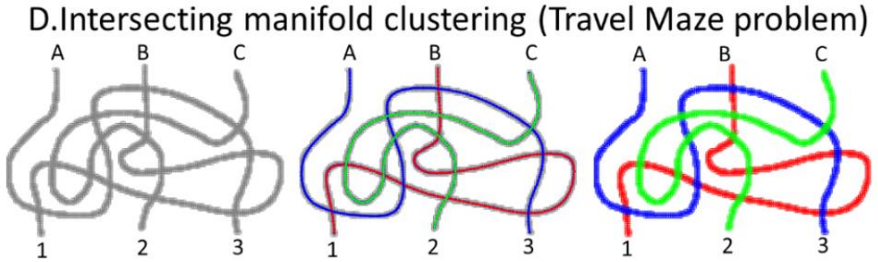


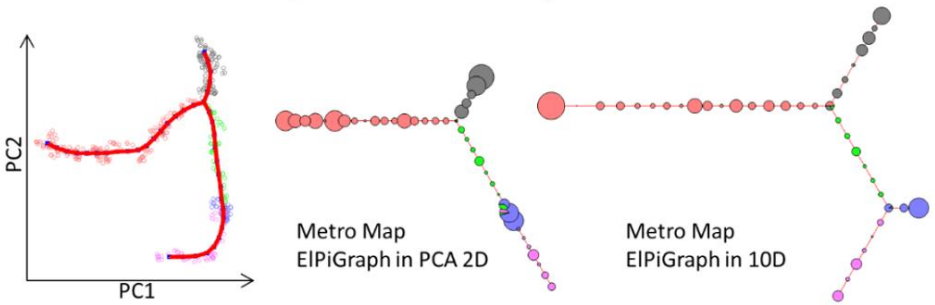
Advanced features of ElPiGraph

(Albergante et al, 2018)









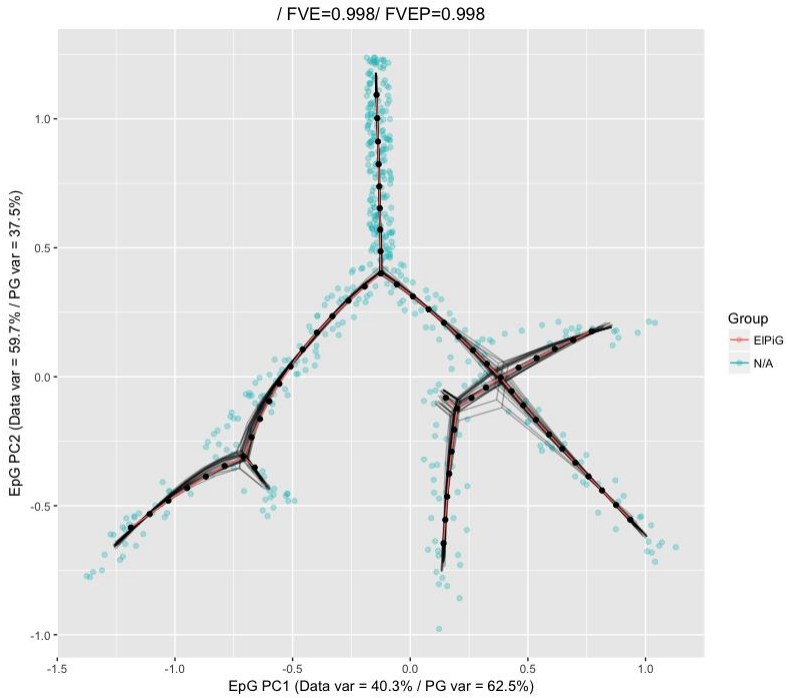
institutCurie

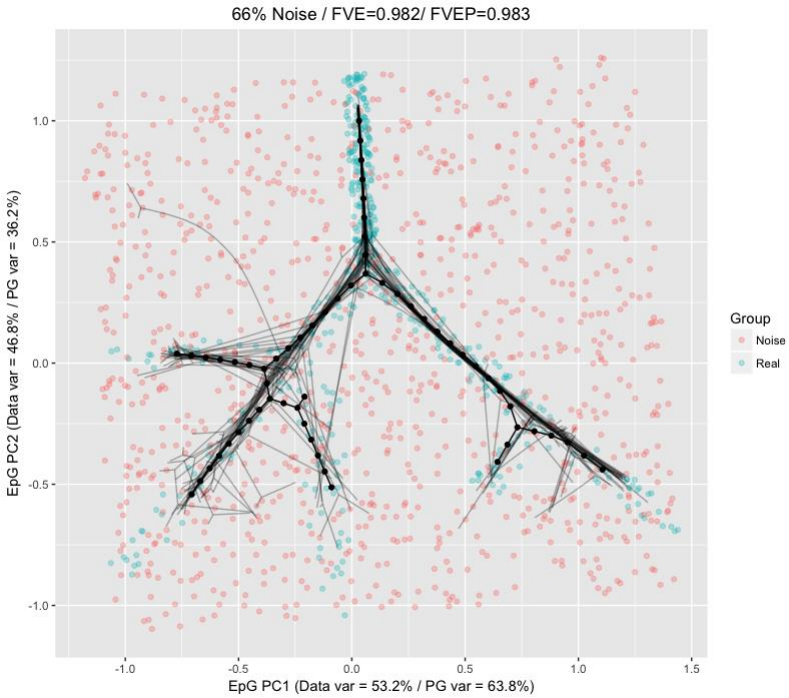
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Principal graphs ensembles and consensus principal graph





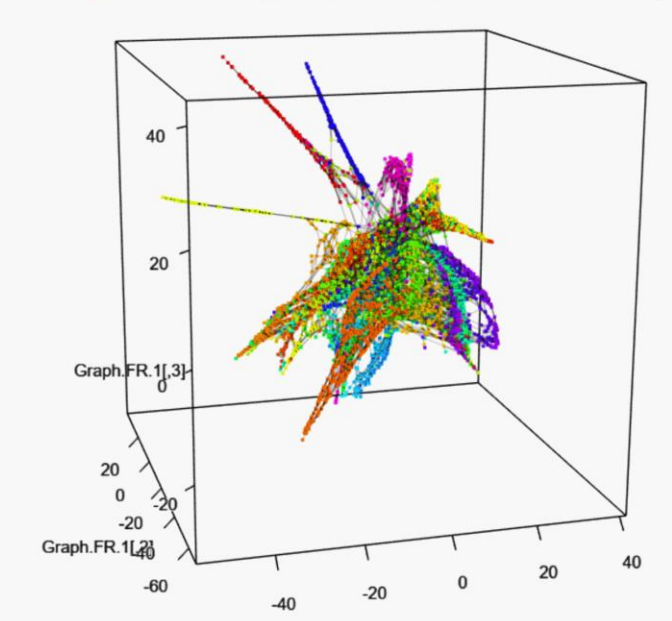
institutCurie

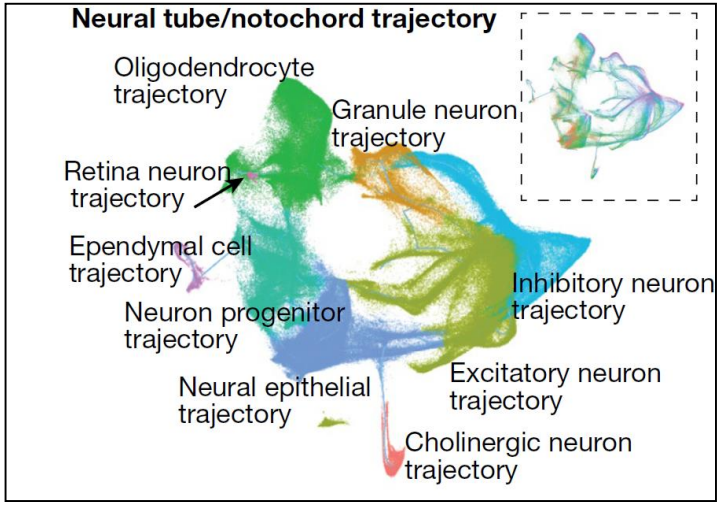
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

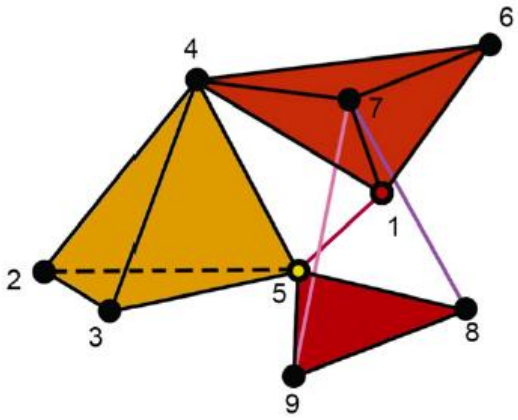
Local intrinsic dimensionality

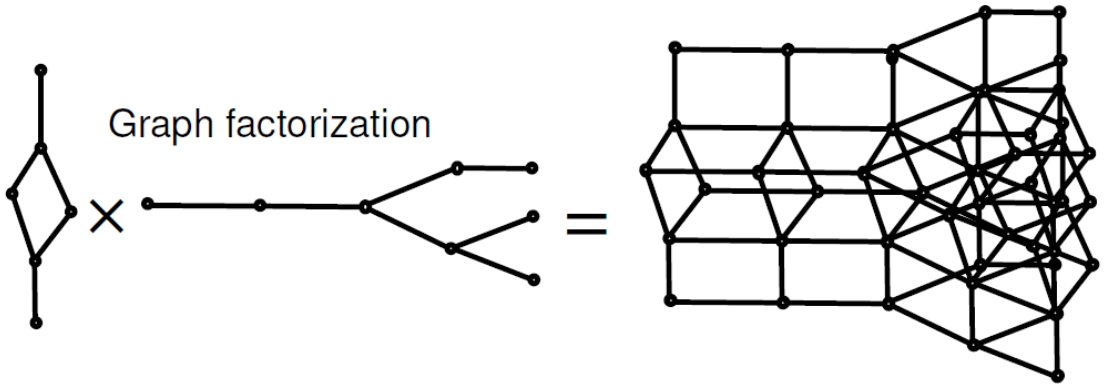




Simplicial complexes, Principal cubic complexes?

(Gorban, Sumner, Zinovyev, AML, 2007;
Gorban&Zinovyev, Handbook of ML; 2009)



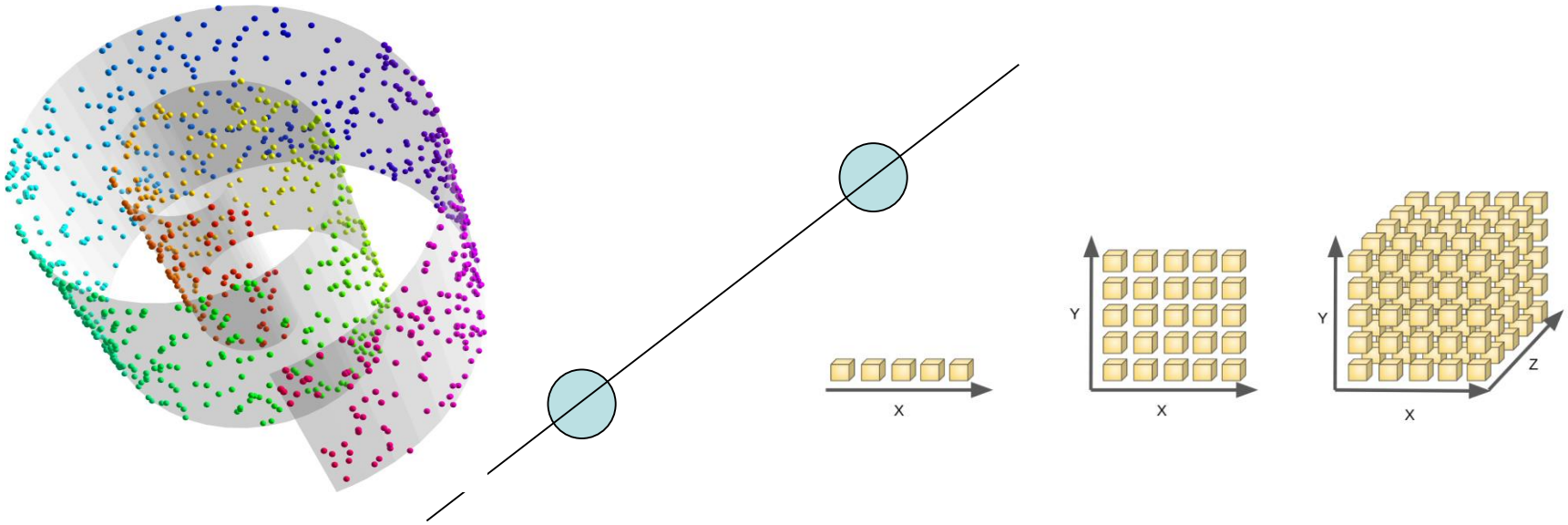


“Curse of dimensionality”

Origin: Bellman, R.E. (1957). Dynamic programming. Princeton University Press, Princeton, NJ.

~~When number of features >> number of objects~~

When the *intrinsic dimension of the data* > log2(number of objects)

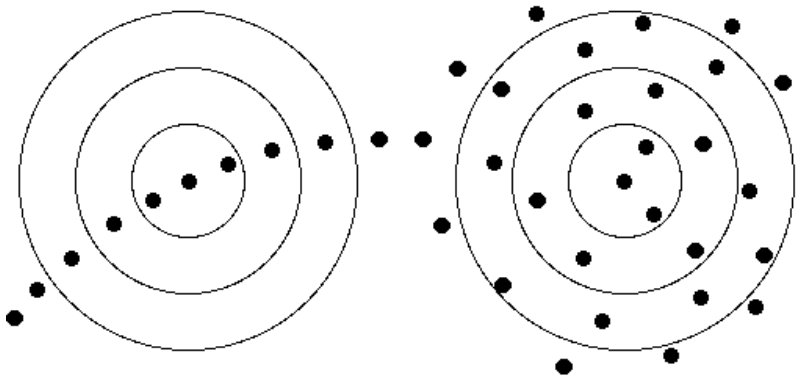


- Vastness of high-dimensional spaces, $2^{100} = 10^{30}$
- Machine learning, based on the notion of point neighbourhood, fails
- Model non-uniqueness increases

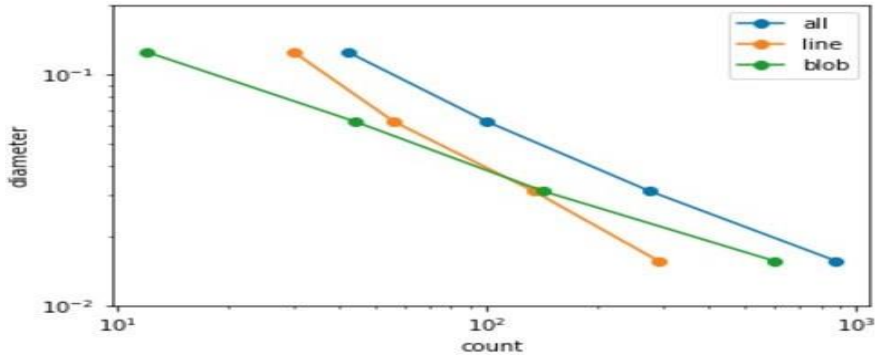
“Standard” measures of intrinsic dimensionality

The correlation dimension

- Count the number of points at a distance less than a radius **r**



Does not work well for high-dimensions (because of the curse and non-uniformity)

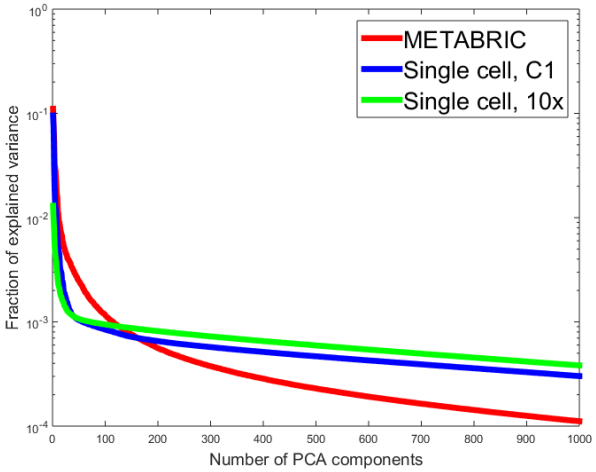


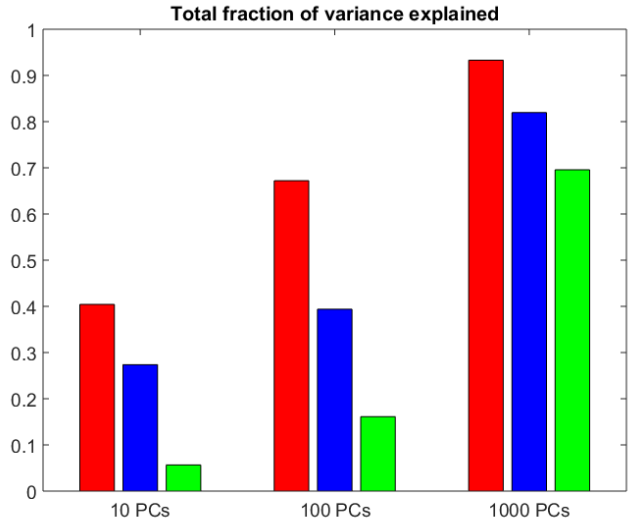
$$d = -\frac{\log n_2 - \log n_1}{\log r_2 - \log r_1}$$

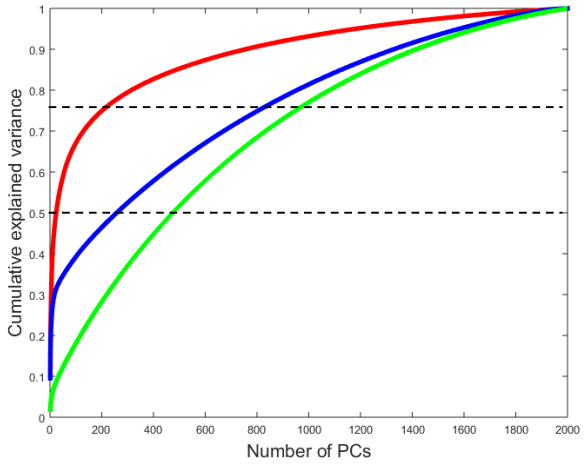
Steeper decline = lower dimension

Do we deal with curse of dimensionality in genomics data?

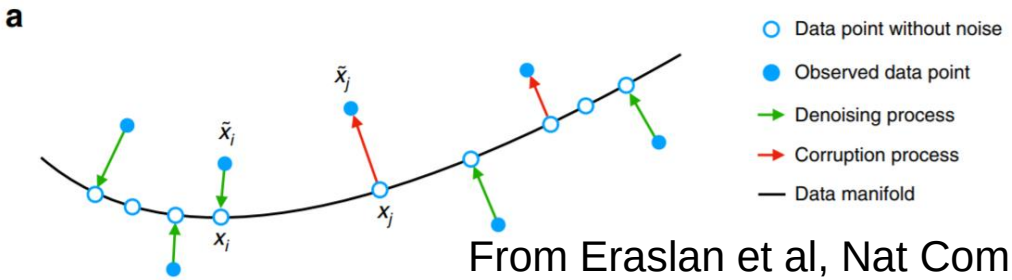
Three datasets of ~2000 samples







Single cell data:



From Eraslan et al, Nat Comm, 2019

institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

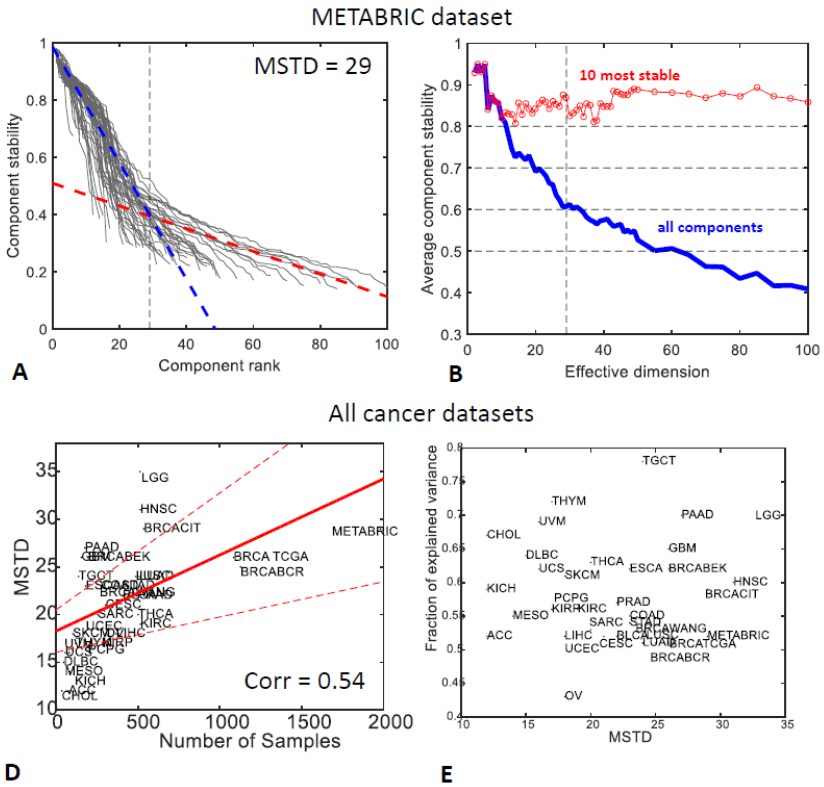
Do we deal with curse of dimensionality in genomics data?

- May be not, may be yes:

$2^{30} \sim 10^{8}$

$2^{20} \sim 10^{6}$

$2^{10} \sim 10^{3}$



From Kairov et al, BMC Genomics, 2017

Blessing of dimensionality?

« Reference for variable selection

Why poll numbers keep hopping around by Philip Meyer »

The blessing of dimensionality

Posted by Andrew on 27 October 2004, 1:00 pm

The phrase “curse of dimensionality” has many meanings (with 18800 references, it loses to “bayesian statistics” in a googlefight, but by less than a factor of 3). In numerical analysis it refers to the difficulty of performing high-dimensional numerical integrals.

But I am bothered when people apply the phrase “curse of dimensionality” to statistical inference.

In statistics, “curse of dimensionality” is often used to refer to the difficulty of fitting a model when many possible predictors are available. But this expression bothers me, because more predictors is more data, and it should not be a “curse” to have more data. Maybe in practice it’s a curse to have more data (just as, in practice, giving people too much good food can make them fat), but “curse” seems a little strong.

With multilevel modeling, there is no curse of dimensionality. When many measurements are taken on each observation, these measurements can themselves be grouped. Having more measurements in a group gives us more data to estimate group-level parameters (such as the standard deviation of the group effects and also coefficients for group-level predictors, if available).

Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification

4 Author(s) Dong Chen ; Xudong Cao ; Fang Wen ; Jian Sun **View All Authors**

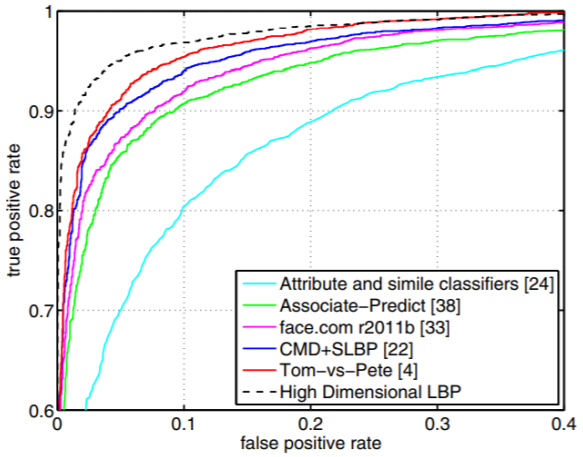
241 Paper Citations	1 Patent Citation	1764 Full Text Views

Abstract

Document Sections

Abstract:

Making a high-dimensional (e.g., 100K-dim) feature for face recognition seems not a good idea because it will bring difficulties on consequent training, computation, and storage. This prevents further exploration



institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Counterintuitive properties of high-dimensional data spaces

- Any two random vectors are (almost) orthogonal
- Any basis of random n vectors is (almost) orthogonal
- (Almost) any point is linearly separable from all other points

- Can it be “blessing”?

Blessing of dimensionality: measure concentration phenomena

PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Review

Article submitted to journal

Blessing of dimensionality: mathematical foundations of the statistical physics of data

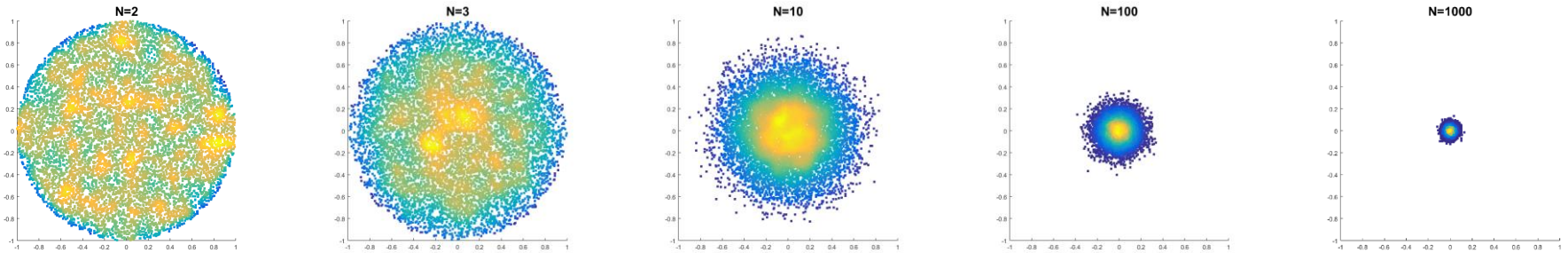
A.N. Gorban[1], I.Y. Tyukin[2]

[1]Department of Mathematics University of Leicester, Leicester LE1 7RH, UK

[2] Department of Mathematics University of Leicester, Leicester LE1 7RH, UK and Department of Automation and Control Processes, Saint-Petersburg State

Counterintuitive properties of high-dimensional distributions

Uniformly sampled ball in R^N *observed in* R^2



Does not matter what distribution, it will look normal in any 2D or 3D projection (law of big numbers)

institutCurie

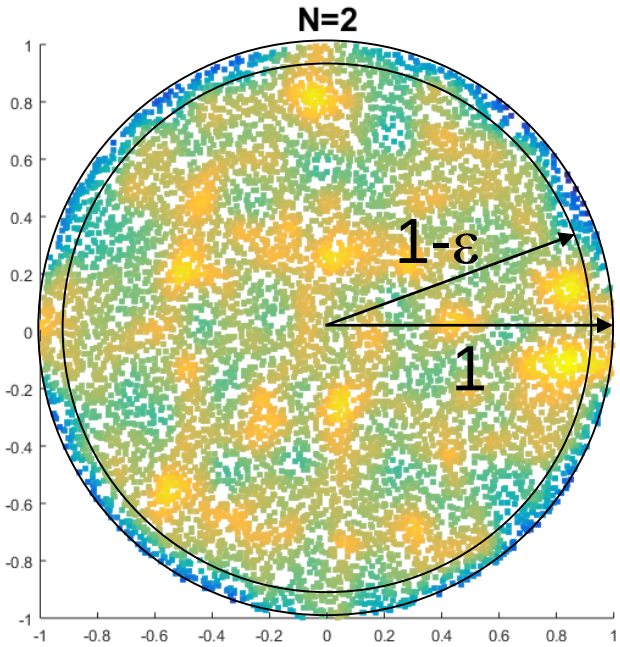
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Counterintuitive properties of high-dimensional distributions

- Concentration of the volume of a ball near its surface



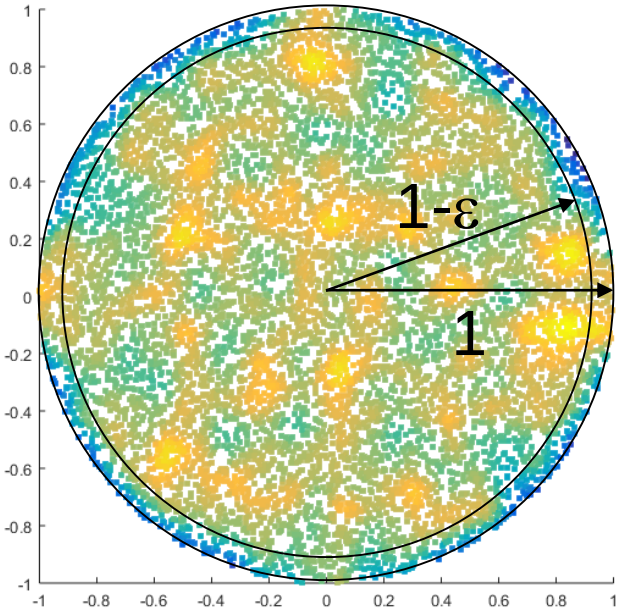
Fraction of volume in vicinity of a surface =

$$f = 1 - (1-\varepsilon)^n \approx 1 - \exp(-\varepsilon n)$$

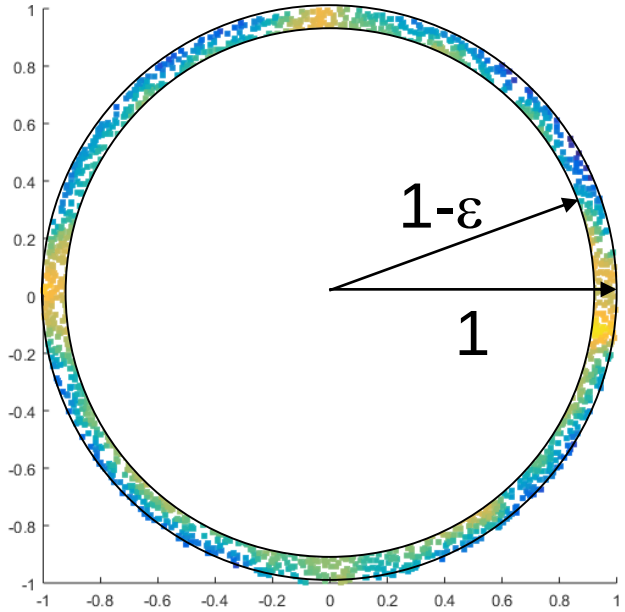
$\varepsilon = 0.01$, n = 2, f = 0.0199
$\varepsilon = 0.01$, n = 3, f = 0.0297
$\varepsilon = 0.01$, n = 10, f = 0.0956
$\varepsilon = 0.01$, n = 100, f = 0.6340
$\varepsilon = 0.01$, n = 500, f = 0.9934

- Concentration of the volume of a ball near its surface

N=2

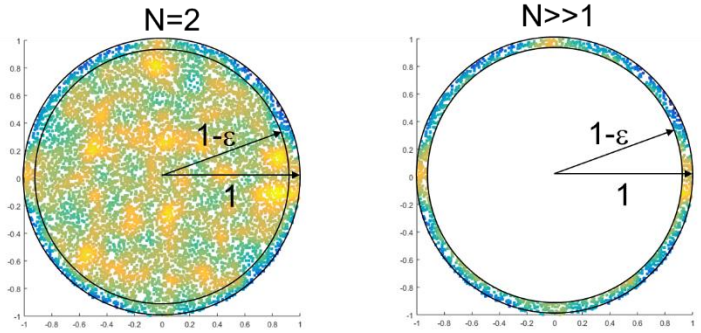


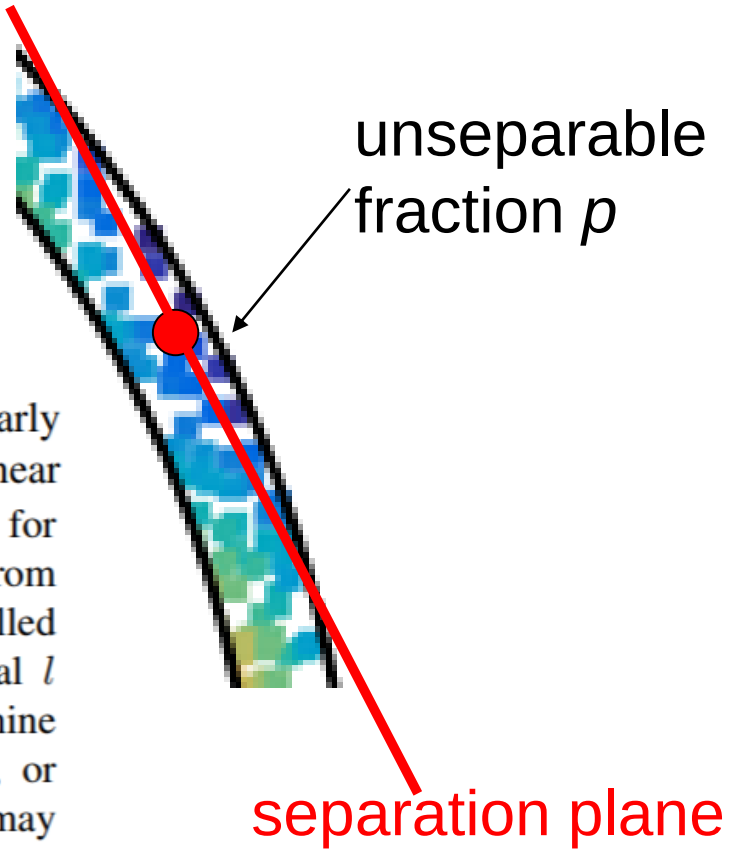
N>>1



This is a “mental image” not projection!

How to quantify separability?





We remind the reader that a point $\mathbf{x} \in R^n$ is linearly separable from a finite set $Y \subset R^n$ if there exists a linear functional l such that $l(\mathbf{x}) > l(\mathbf{y})$ for all $y \in Y$. If for any point $\mathbf{x}$ there exists a linear functional separating it from all other data points, then such a data point cloud is called *linearly separable* or 1-convex. The separating functional l may be computed using the linear Support Vector Machine (SVM) algorithms, the Rosenblatt perceptron algorithm, or other comparable methods. However, these computations may be rather costly for large-scale estimates. Hence, it has been suggested to use the simplest non-iterative estimate of the linear functional by Fisher's linear discriminant which is computationally inexpensive, after a well-established standardised pre-processing described below [7].

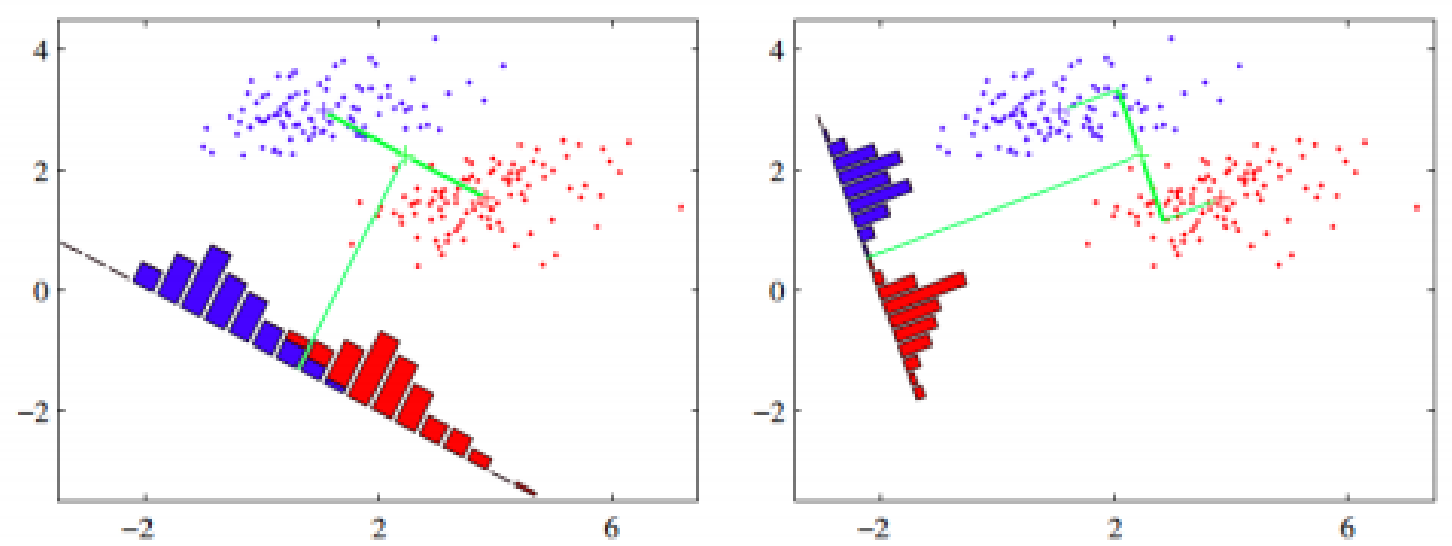
ins

MINES ParisTech

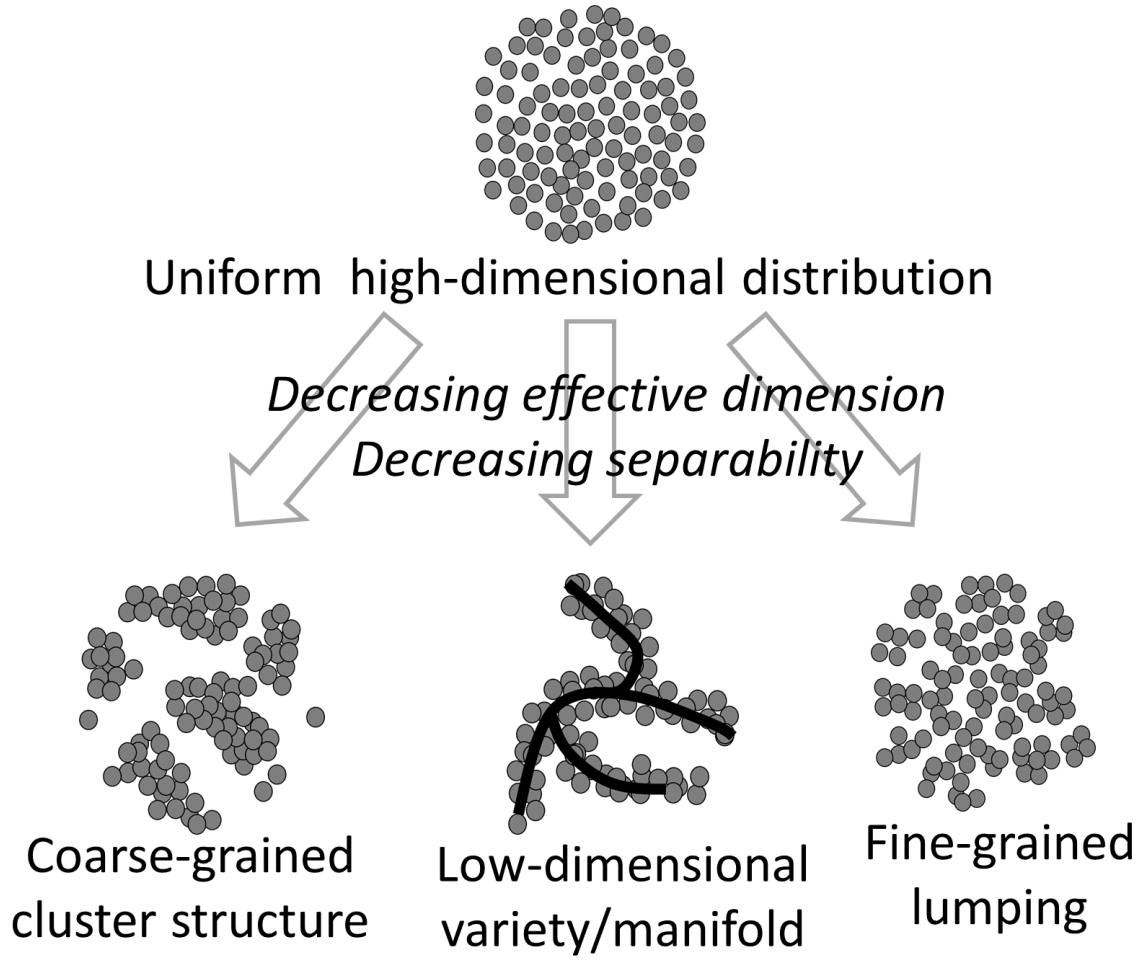
PSL RESEARCH UNIVERSITY

Fisher discriminant analysis

Computed in explicit formula, without iterations!



Intrinsic or effective data dimensionality from the point of view of separability



From Albergante et al, IJCNN proceedings, 2019, arxiv:1901.06328

How to quantify dimensionality from separability?

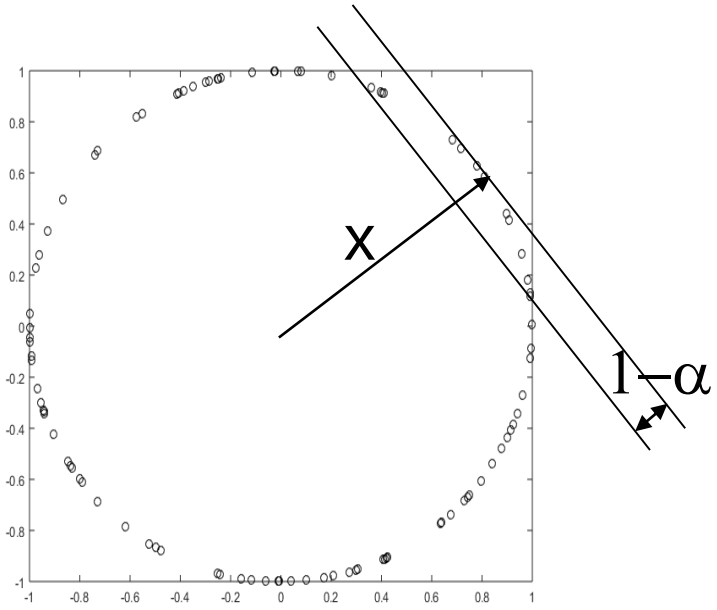
Albergante et al, IJCNN proceedings, 2019, arxiv:1901.06328

1. We first make our dataset comparable to a multidimensional sphere S^{n-1} in R^n (centering, whitening + scaling data vectors)

2. For each point x we compute how many other points $y \neq x$ have $(x,y)>\alpha$, this gives $p(\alpha)$

3. For uniform distribution on a sphere

$$p_\alpha = \bar{p}_\alpha = \frac{(1-\alpha^2)^{\frac{n-1}{2}}}{\alpha\sqrt{2\pi n}}$$

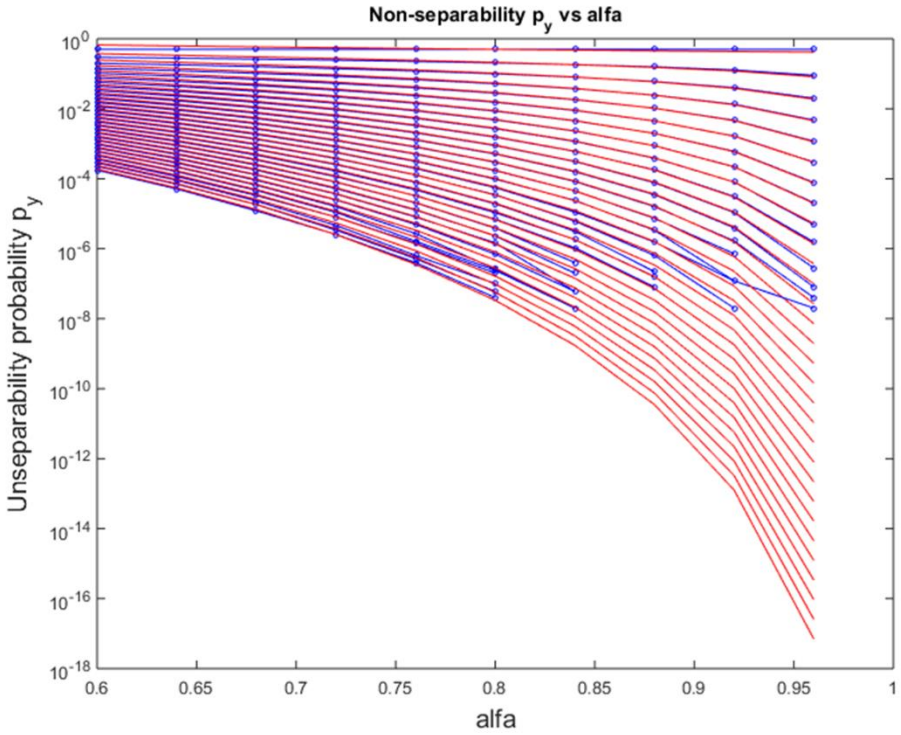


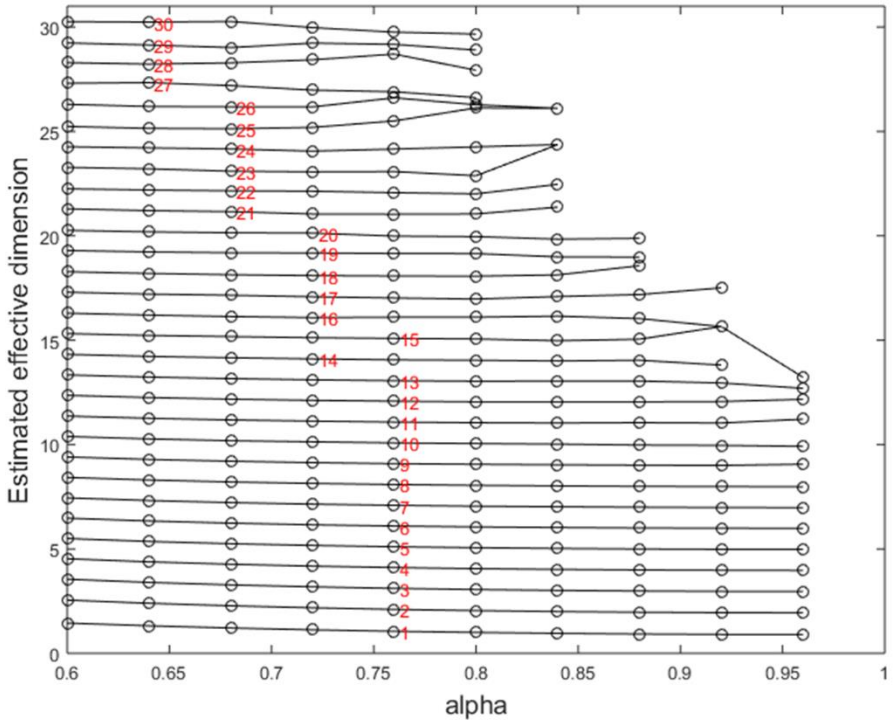
4. From here we can find

$$n_\alpha = \frac{W\left(\frac{-\ln(1-\alpha^2)}{2\pi\bar{p}_\alpha^2\alpha^2(1-\alpha^2)}\right)}{-\ln(1-\alpha^2)}$$

where W is Lambert function (solution of equation $y = xe^x$)

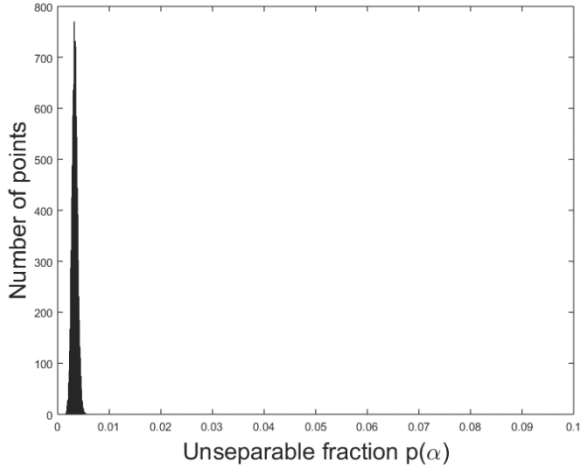
How it works for uniformly sampled spheres

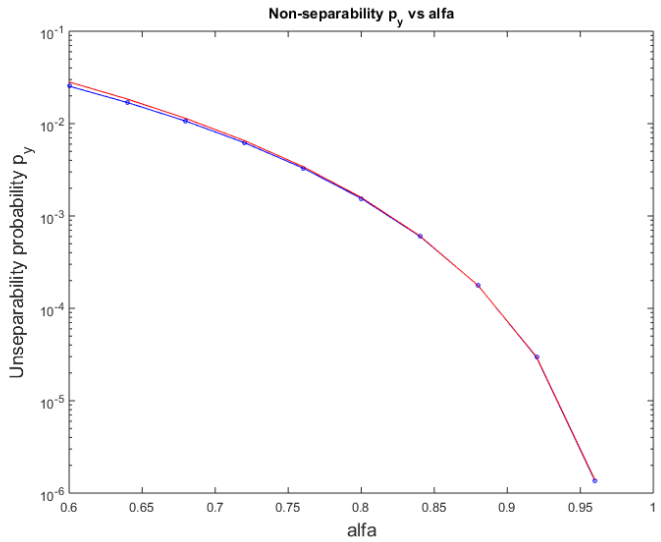




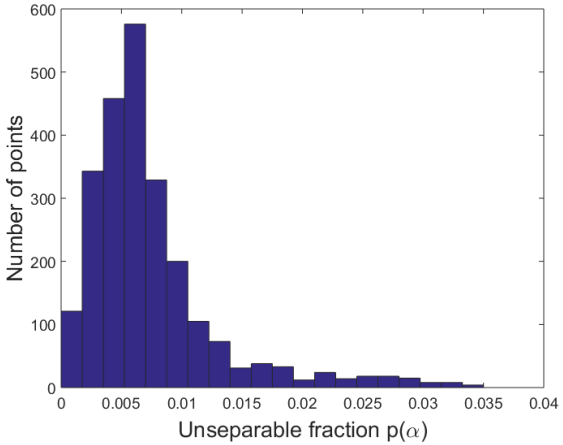
Unseparable fraction distribution

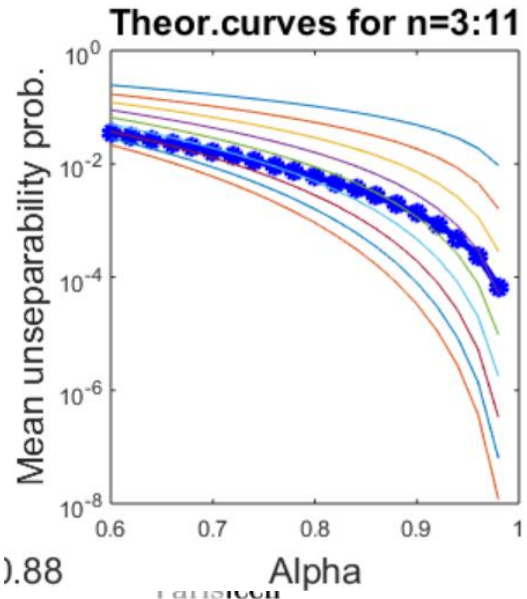
Uniform data distribution (n=30)





Some real single cell data distribution (Formal n=20000, PCA-estimated n=28)





- Synthetic manifolds (M. Hein and J.-Y. Audibert, 2005)

Dataset	Underlying manifold name	Description	d	D
Synthetic	$\mathcal{M}_1^H$	d-dimensional sphere linearly embedded	D – 1	*User-defined*
	$\mathcal{M}_2^H$	Affine space	3	5
	$\mathcal{M}_3^H$	Concentrated figure, mistakable with a 3-dimensional one	4	6
	$\mathcal{M}_4^H$	Nonlinear manifold	4	8
	$\mathcal{M}_5^H$	2-dimensional helix	2	3
	$\mathcal{M}_6^H$	Nonlinear manifold	6	36
	$\mathcal{M}_7^H$	Swiss-Roll	2	3
	$\mathcal{M}_8^H$	Nonlinear (highly curved) manifold	12	72
	$\mathcal{M}_9^H$	Affine space	D	*User-defined*
	$\mathcal{M}_{10}^H$	d-dimensional hypercube	D – 1	*User-defined*
	$\mathcal{M}_{11}^H$	Möebius band 10-times twisted	2	3
	$\mathcal{M}_{12}^H$	Isotropic multivariate Gaussian	D	*User-defined*
	$\mathcal{M}_{13}^H$	1-dimensional helix curve	1	*User-defined*

The estimation worked surprisingly well on a benchmark when noise was added

TABLE I

Predicted ID for synthetic datasets evaluated globally, with added multidimensional isotropic Gaussian noise (standard deviation $\sigma = .05$), and the ISOMAP Faces dataset. Cardinality: Number of points of the dataset, N: embedding dimension, n: intrinsic dimension. *FisherS*: Fisher Separability (The number in parentheses indicates the number of components retained by PCA preprocessing for the separability-based method), *CD*: Correlation Dimension [23], *GMSTL*: Geodesic Minimum Spanning Tree Length [19], *DANCo*: Dimensionality from Angle and Norm Concentration, *LBMLE*: Levina-Bickel Maximum Likelihood Estimation [29], *ESS*: Expected Simplex Skewness, *FanPCA*: PCA based on [30], *TwoNN*: Two Nearest Neighbors [28]

	Cardinality	N	n	FisherS	CD	GMSTL	DANCo	LBMLE	ESS	FanPCA	TwoNN
$\mathbf{M_{13}}$	2500	13	1	1.67 (3)	1.64	3.73	4	3.74	3.16	2	5.50
$\mathbf{M_{5}}$	2500	3	2	2.57 (3)	2.14	2.47	3	2.66	2.74	1	2.73
$\mathbf{M_{7}}$	2500	3	2	2.94 (3)	2	2.24	2	2.39	2.93	2	2.67
$\mathbf{M_{11}}$	2500	3	2	1.96 (2)	2.33	2.21	2	2.49	2.34	1	2.69
Faces	698	4096	3	3.12 (28)	0.78	1.64	4	4.31	7.49	8	3.49
$\mathbf{M_{2}}$	2500	5	3	2.66 (3)	3.60	4.61	4	4.42	2.66	2	4.69
$\mathbf{M_{3}}$	2500	6	4	2.87 (4)	3.16	3.36	4	4.40	3.11	2	4.36
$\mathbf{M_{4}}$	2500	8	4	5.78 (8)	3.90	4.33	4	4.38	7.79	5	3.96
$\mathbf{M_{6}}$	2500	36	6	8.50 (12)	5.99	6.62	7	7.05	11.98	9	6.27
$\mathbf{M_{1}}$	2500	11	10	11.03 (11)	8.96	9.02	11	9.88	10.81	7	9.43
$\mathbf{M_{10a}}$	2500	11	10	9.46 (10)	7.86	9.50	10	8.90	10.31	7	8.57
$\mathbf{M_{8}}$	2500	72	12	17.41 (24)	10.97	13.04	17	14.74	24.11	18	13.15
$\mathbf{M_{10b}}$	2500	18	17	15.94 (17)	11.88	13.15	16	13.89	17.35	13	13.59
$\mathbf{M_{12}}$	2500	20	20	19.83 (20)	10.62	16.05	20	17.07	19.90	11	16.94
$\mathbf{M_{9}}$	2500	20	20	19.07 (20)	13.51	14.26	19	15.73	20.26	11	15.68
$\mathbf{M_{10c}}$	2500	25	24	22.62 (24)	15.15	21.94	23	18.24	24.42	17	17.36
$\mathbf{M_{10d}}$	2500	71	70	68.74 (70)	29.89	36.62	71	38.92	71.95	43	39.18
Mean%error				28.82	32.45	36.35	43.04	43.83	66.78	67.56	74.91

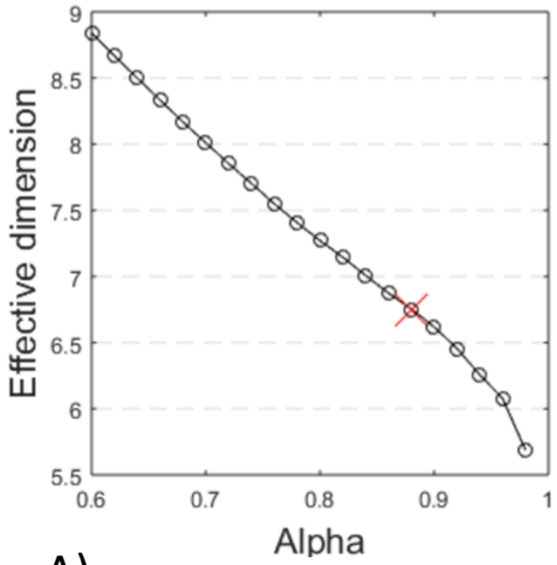
institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

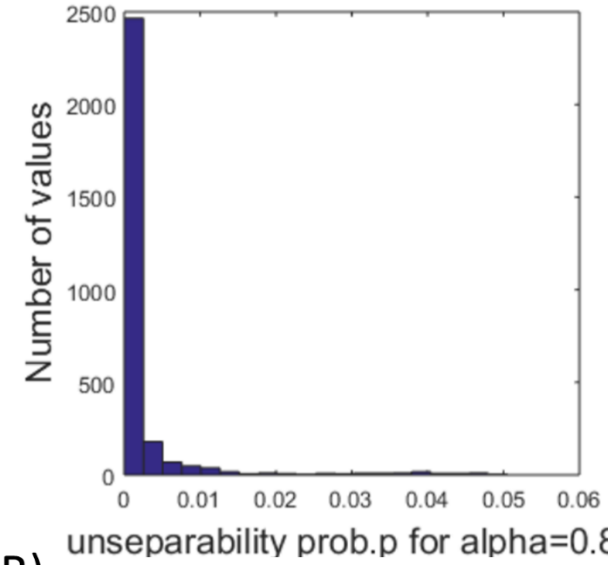
MINES ParisTech

PSL RESEARCH UNIVERSITY

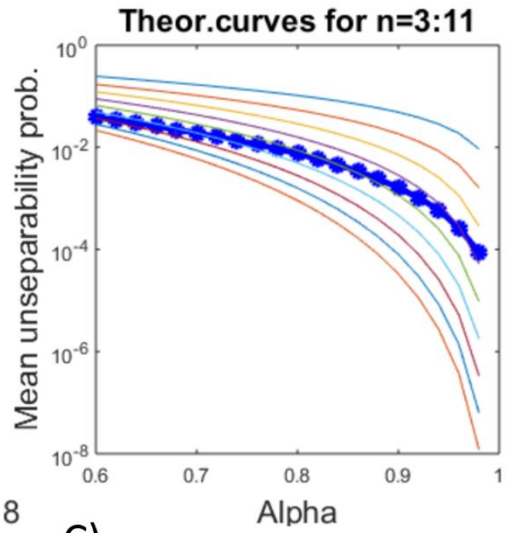
Mutation data (BRCA), microcluster structure



A)

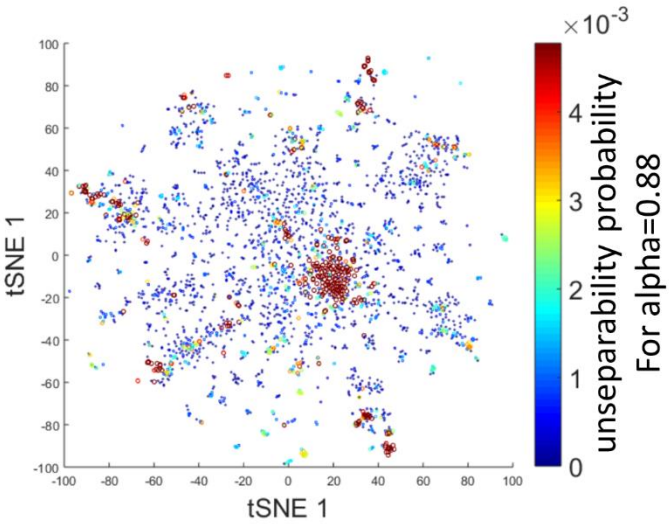


B)



8

C)

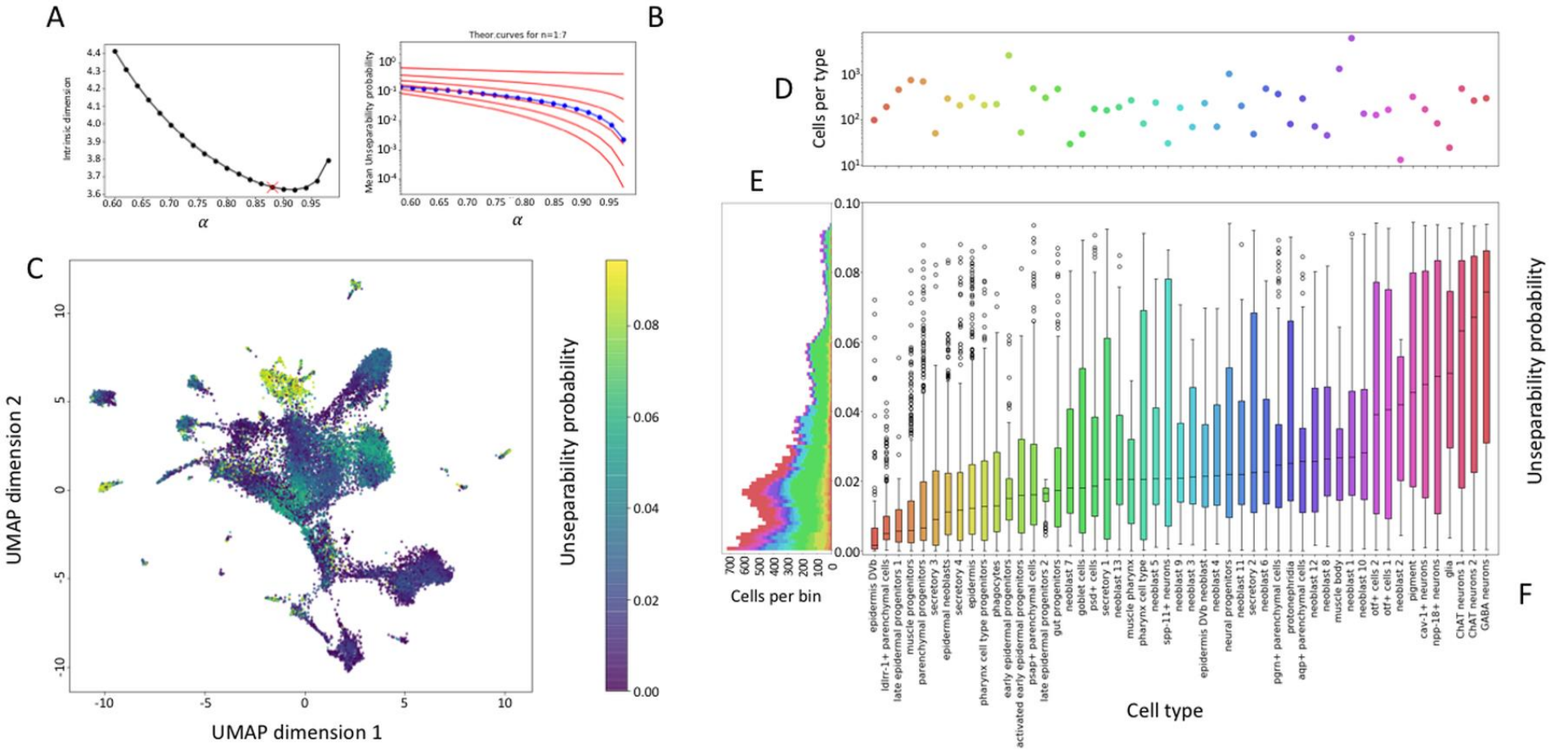


D)

institutCurie

PSL RESEARCH UNIVERSITY

Single cell RNASeq data (planarian transcriptome)



institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

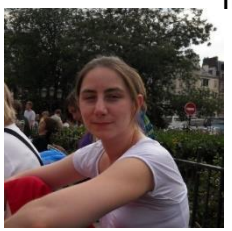
PSL RESEARCH UNIVERSITY

Acknowledgements

Systems biology group at Institut Curie

Emmanuel Barillot
Institut Curie

Nicolas Sompairac
Institut Curie



Jane Mervelede
Institut Curie

Independent Component Analysis methodology

Urszula Czerwinska

Laura Cantini
ENS

Francois Radvanyi
Institut Curie

Peter Nazarov
Luxembourg Institute of Health

Ulykbek Kairov
Nazarbayev University
Kazakhstan

Blessing of dimensionality team

Luca Albergante
Sensyne Health

Alexander Gorban
University of Leicester
UK

Eugene Mirkes
University of Leicester
UK

iPaediatricCure EU Horizon-2020
IMMUCAN EU IMI
Scalable Artificial Intelligence Networks for Data Analysis in Growing Dimensions, Russian Ministry of Science