UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

Data structure in machine learning: estimators and models

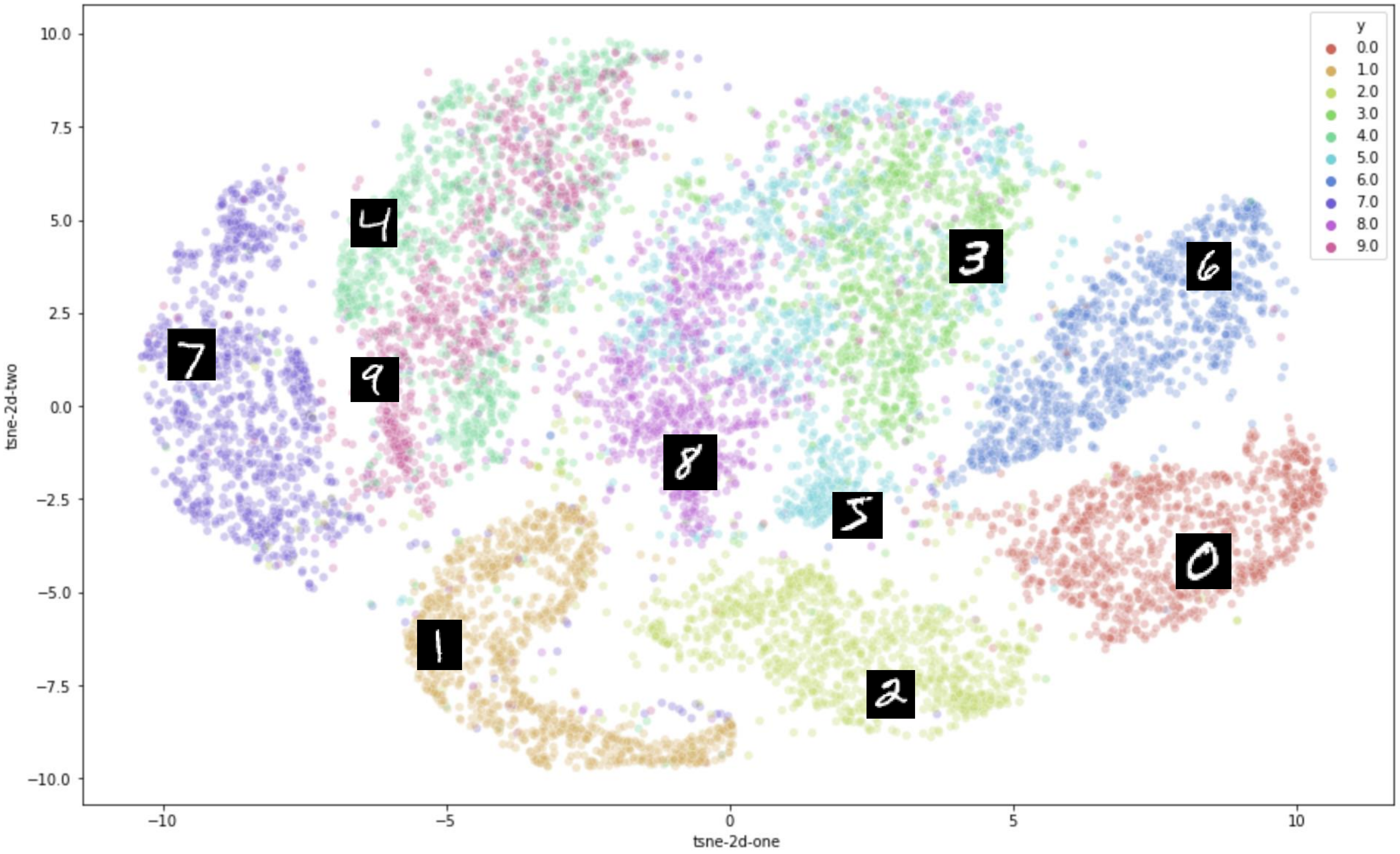
Vittorio Erba[1,2]

Marco Cosentino Lagomarsino[1], Marco Gherardi[1,2], Mauro Pastore[1,2], Pietro Rotondo[2]

EPFL, 15 December 2020

1. Università degli Studi di Milano
2. INFN, Sezione di Milano & FELLINI fellowship

Real data is geometrically structured



2d t-SNE projection of MNIST dataset

Image taken from: https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b

Understanding data structure requires many tools

1) UNDERSTAND STRUCTURE IN REAL DATASETS (Manifold learning)
Clustering, Intrinsic dimension estimation, Dimensionality reduction, ...

TODAY

PART 1:
INTRINSIC DIMENSION ESTIMATION

2) HOW TO EMBED REAL DATASETS IN MATHEMATICALLY TRACTABLE SPACES
Word embeddings, One-hot encodings, … +
Euclidean VS Non-Euclidean metrics (Wasserstein, p-norms, ...)

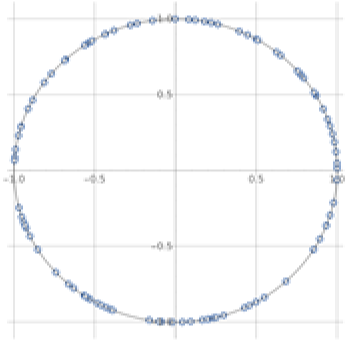
3) THEORETICAL MODELS
Perceptual manifolds, Teacher-student, Hidden manifold model

PART 2:
LINEAR CLASSIFICATION OF GEOMETRICALLY STRUCTURE DATA

Part 1: Intrinsic Dimension Estimation (IDE)

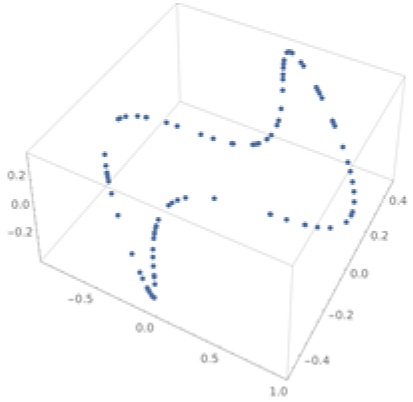
1) Define the problem
2) Overview of algorithms and issues
3) Glimpse of our novel estimator

IDE: retrieve the dimension d of manifold from a discrete, random sample of N points



Low d-dim manifold
+
Random sampling

Embedding



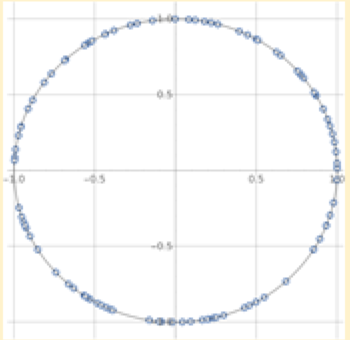
High D-dim representation

IDE algorithm

Intrinsic dimension d

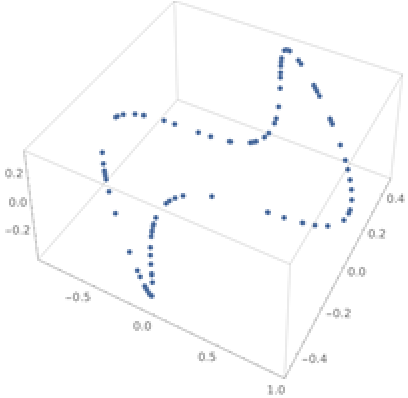
IDE: retrieve the dimension d of manifold from a discrete, random sample of N points

HIDDEN INFORMATION



Low d-dim manifold
+
Random sampling

Embedding



High D-dim representation

IDE algorithm

Intrinsic dimension d

MANIFOLD HYPOTHESYS

Real datasets are random samples of smooth manifolds

MAY NOT BE TRUE + EMBEDDING DEPENDENT

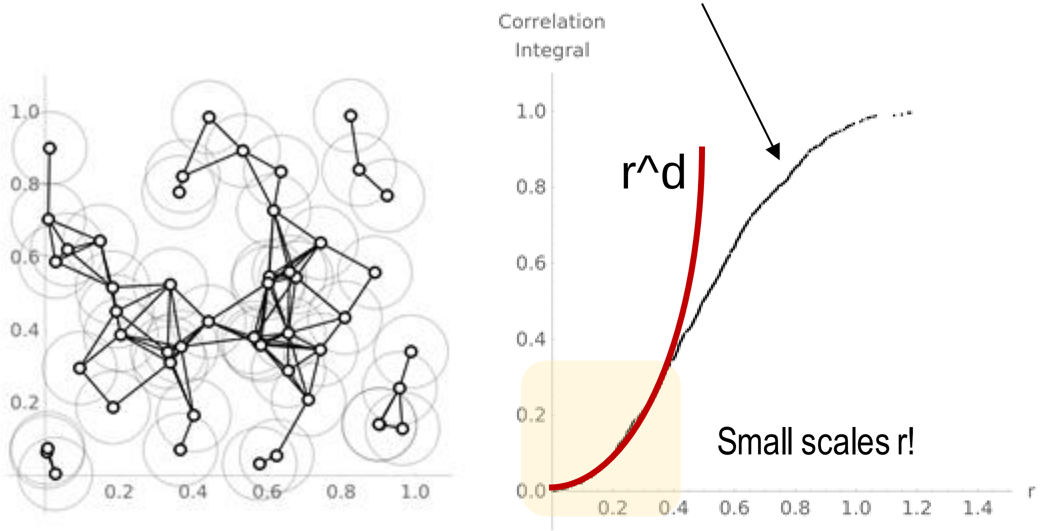
INTRINSIC DIMENSION

Minimal number of degrees of freedom that encode all information of the dataset

There are two classes of IDE algorithms

GEOMETRIC ESTIMATORS (Corr Dim)

$$\rho(r) = \frac{2}{N(N-1)} \sum_{1\leq i<j\leq N} \theta(r - ||\vec{x}_i - \vec{x}_j||)$$



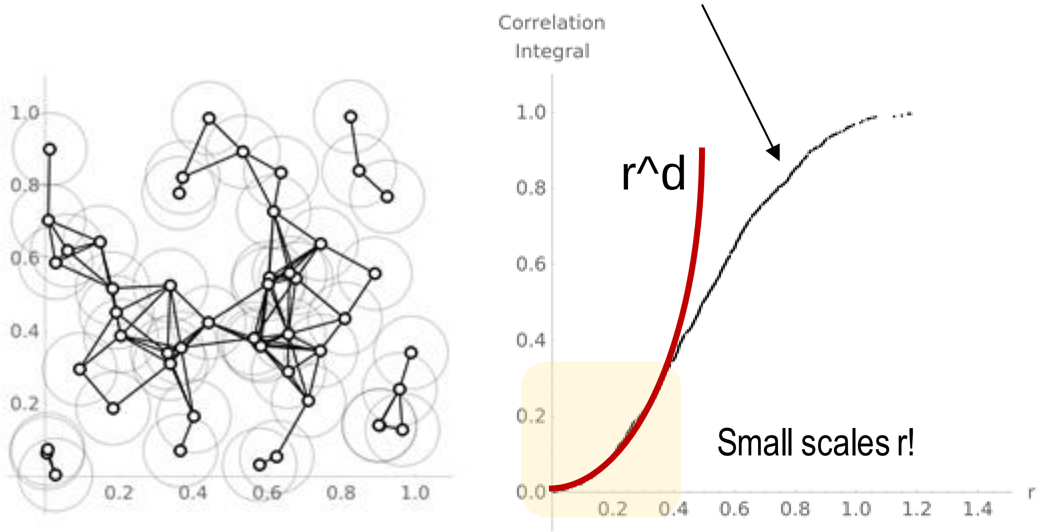
ASSUME LOCAL LINEARITY +
MEASURE LOCAL DENSITY AT SMALL SCALE r
=
FIT AGAINST r^d

Grassberger, P., & Procaccia, I. (2004). Measuring the strangeness of strange attractors. In *The Theory of Chaotic Attractors* (pp. 170-189). Springer, New York, NY.

There are two classes of IDE algorithms

GEOMETRIC ESTIMATORS (Corr Dim)

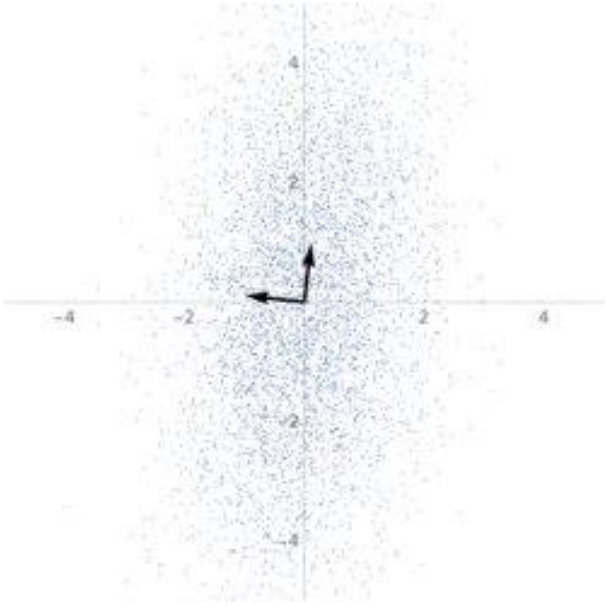
$$\rho(r) = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \theta(r - ||\vec{x}_i - \vec{x}_j||)$$



ASSUME LOCAL LINEARITY +
MEASURE LOCAL DENSITY AT SMALL SCALE r
=
FIT AGAINST r^d

Grassberger, P., & Procaccia, I. (2004). Measuring the strangeness of strange attractors. In *The Theory of Chaotic Attractors* (pp. 170-189). Springer, New York, NY.

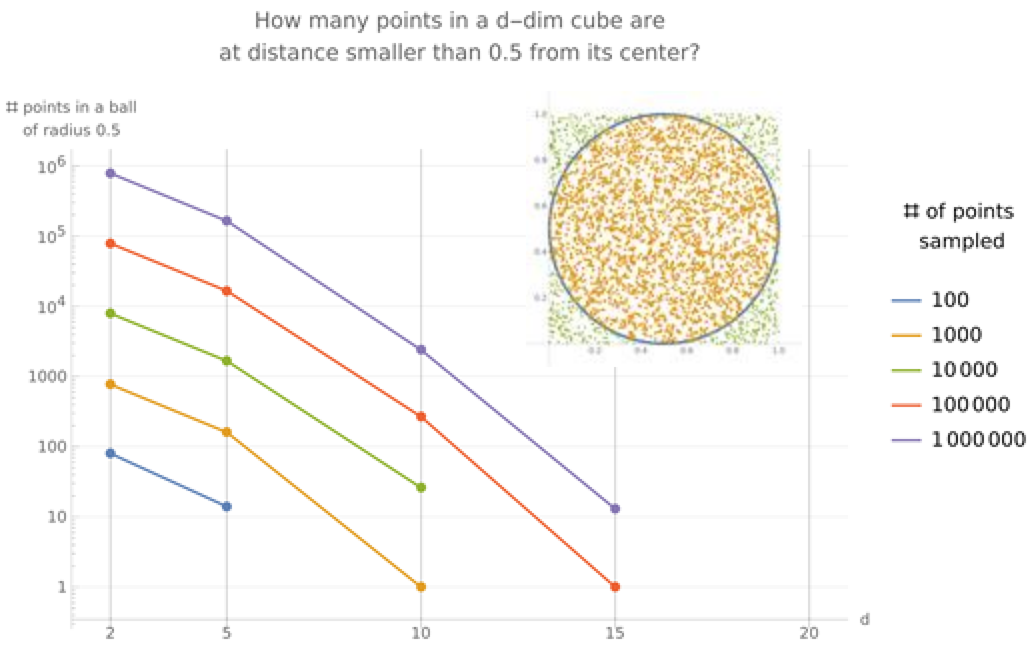
PROJECTIVE ESTIMATORS (PCA)



ASSUME GLOBAL LINEARITY
=
USE LINEAR ALGEBRA TO DISTINGUISH BETWEEN INTRINSIC AND SPURIOUS DIMENSIONS

Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901): 559-572.

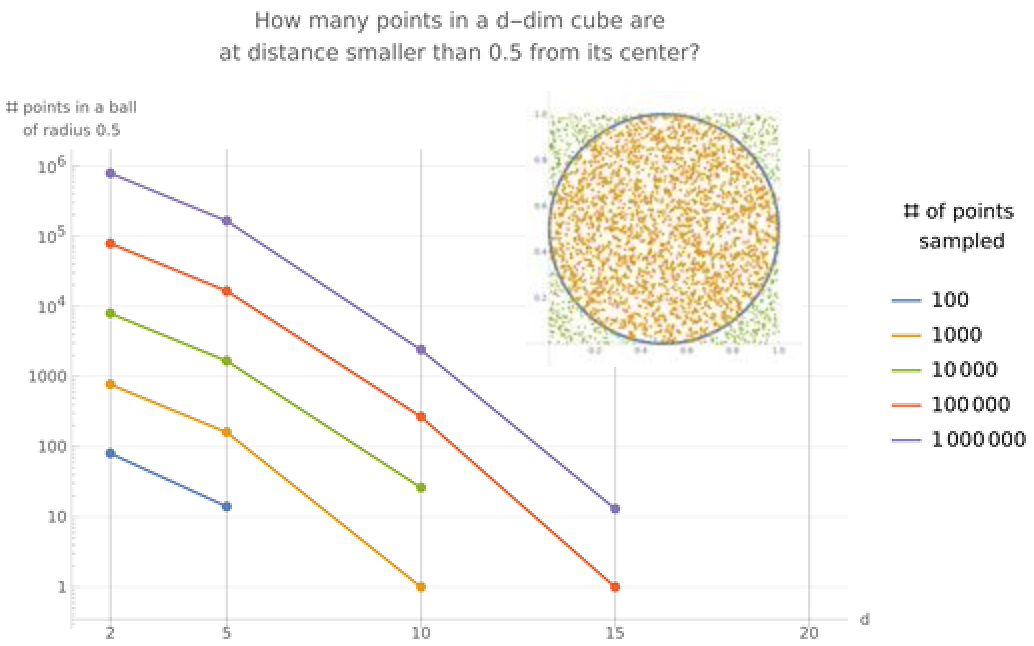
High ID + Curvature are the main enemies of IDE algorithms



EXPONENTIAL UNDERSAMPLING IN HIGH DIMENSION (d > 6) = LOCAL LINEARITY DIFFICULT TO PROBE

Eckmann, J-P., and David Ruelle. "Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems." *Physica D: Nonlinear Phenomena* 56.2-3 (1992): 185-187.

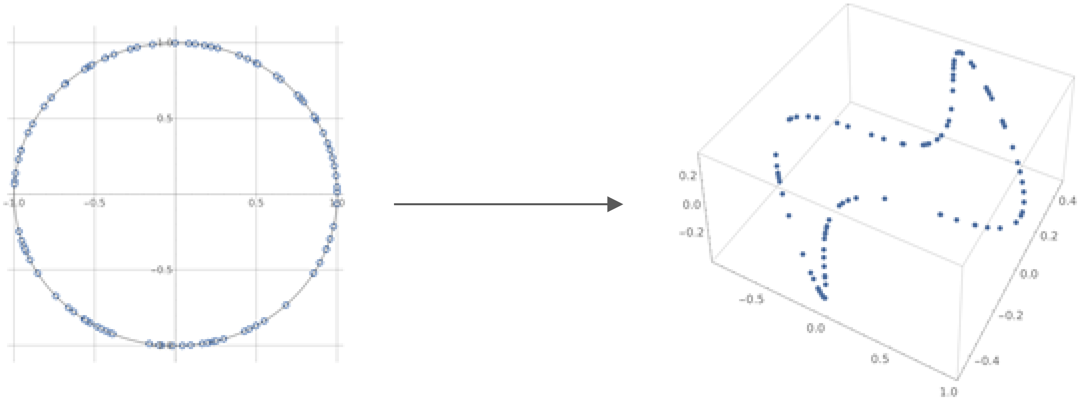
High ID + Curvature are the main enemies of IDE algorithms



EXPONENTIAL UNDERSAMPLING IN HIGH DIMENSION (d > 6) = LOCAL LINEARITY DIFFICULT TO PROBE

Eckmann, J-P., and David Ruelle. "Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems." *Physica D: Nonlinear Phenomena* 56.2-3 (1992): 185-187.

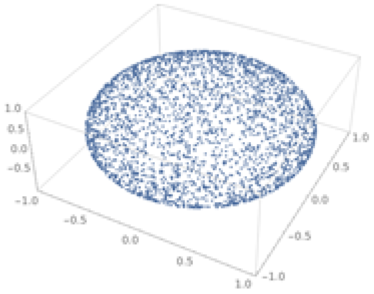
CURVATURE => PROLIFERATION OF POSSIBLE GEOMETRIES + NO GLOBAL LINEARITY



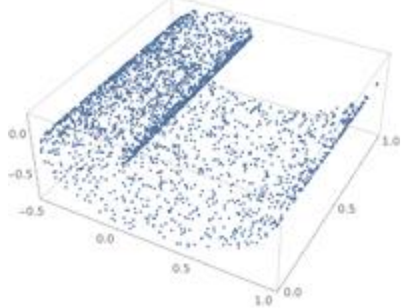
Low d-dim manifold
+
Random sampling

High D-dim representation

INTRINSIC CURVATURE



EXTRINSIC CURVATURE

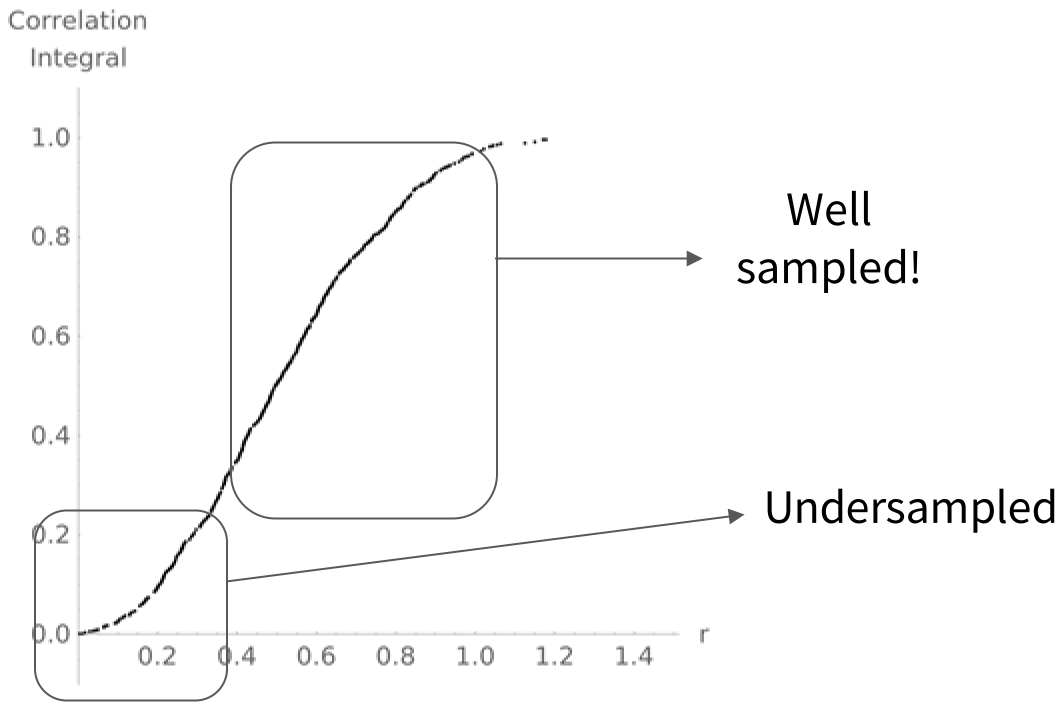


A new estimator: Full Correlation Integral (FCI)

FCI leverages not-so-small r regime to avoid undersampling

Linear manifolds
Isotropic sampling measure
Linear embeddings

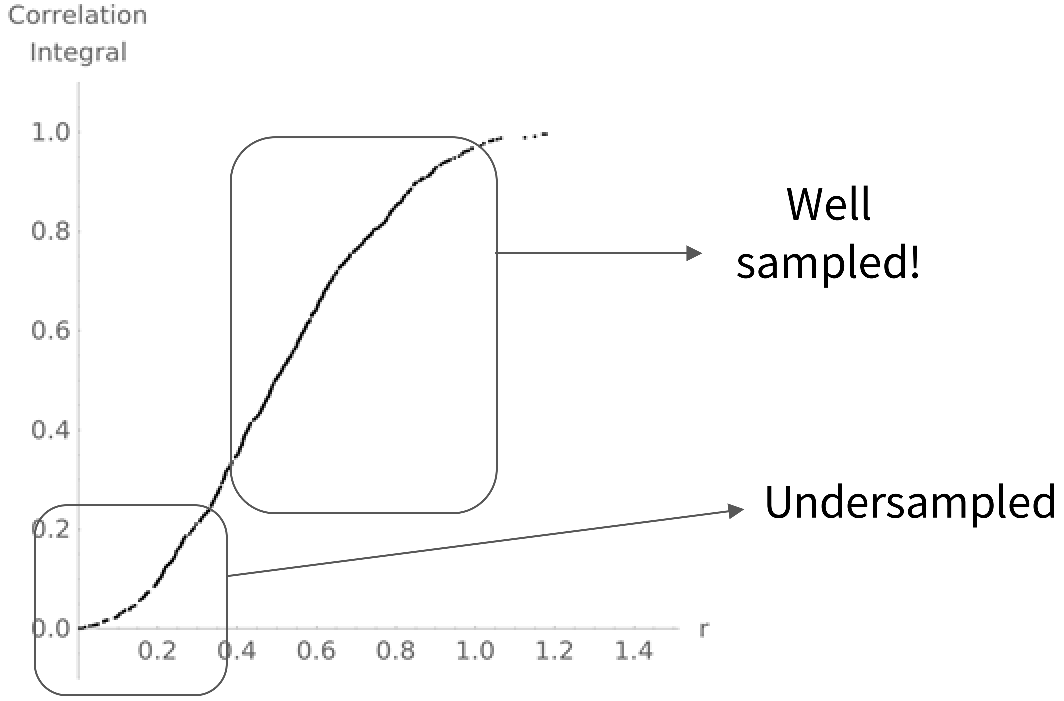
$$\rho(r;d) = \frac{1}{2} + \frac{\Omega_{d-1}}{\Omega_d}(r^2-2)\,{}_2F_1\left(\begin{matrix}\frac{1}{2}, 1-\frac{d}{2} \\ \frac{3}{2}\end{matrix}\middle|(r^2-2)^2\right)$$



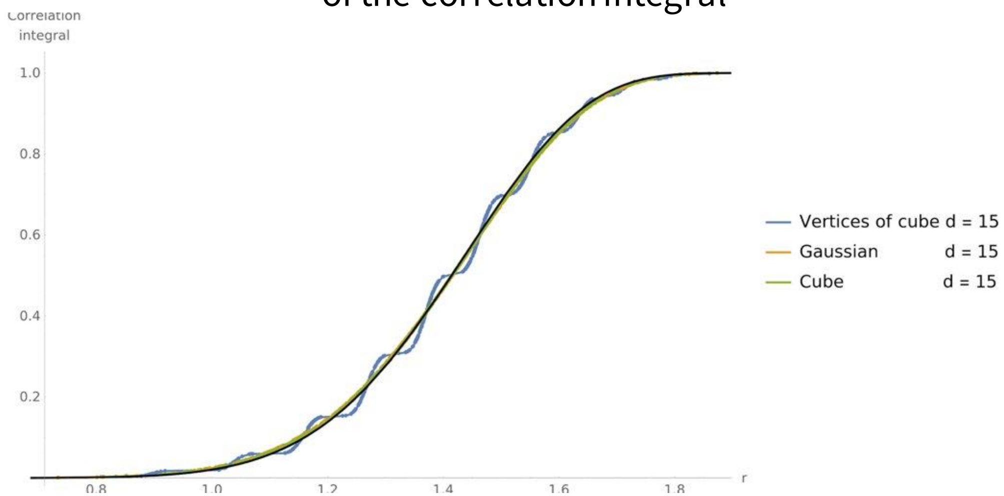
FCI leverages not-so-small r regime to avoid undersampling

Linear manifolds
Isotropic sampling measure
Linear embeddings

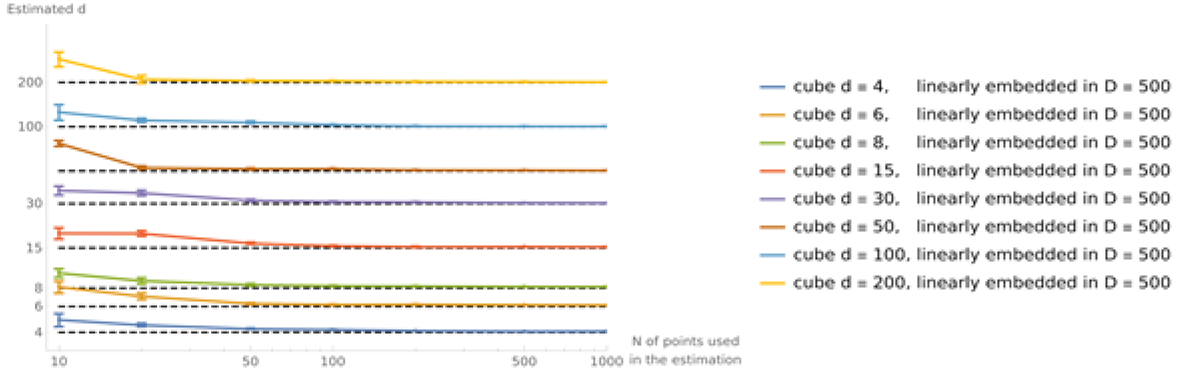
$$\rho(r;d) = \frac{1}{2} + \frac{\Omega_{d-1}}{\Omega_d}(r^2-2)\,{}_2F_1\left(\begin{matrix}\frac{1}{2}, 1-\frac{d}{2} \\ \frac{3}{2}\end{matrix}\middle|(r^2-2)^2\right)$$



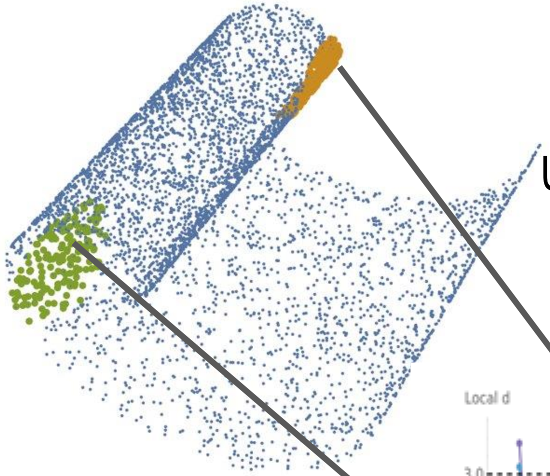
Manifold dependent features do not alter the overall shape of the correlation integral



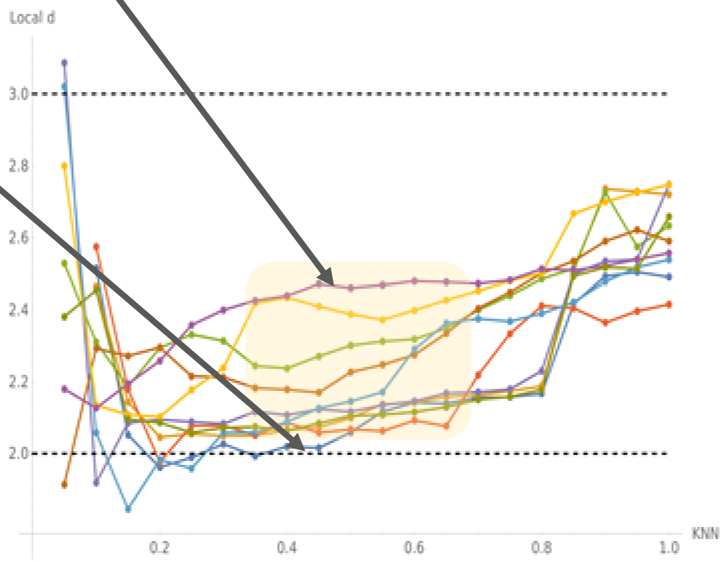
Able to estimate in the extreme undersampled regime N < d
(Geometric: exp d | Projective: d log d)



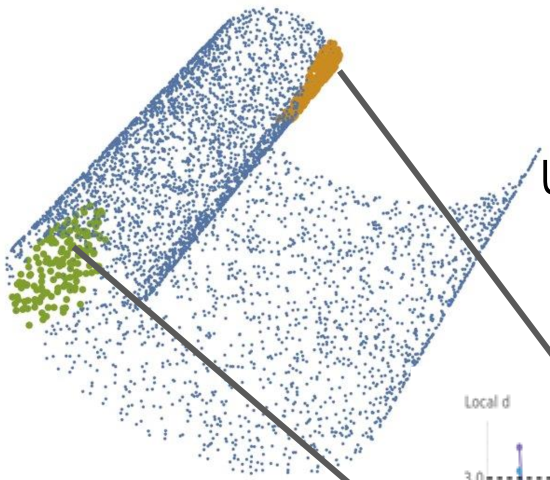
FCI can be easily multiscaled to extend to curved manifolds



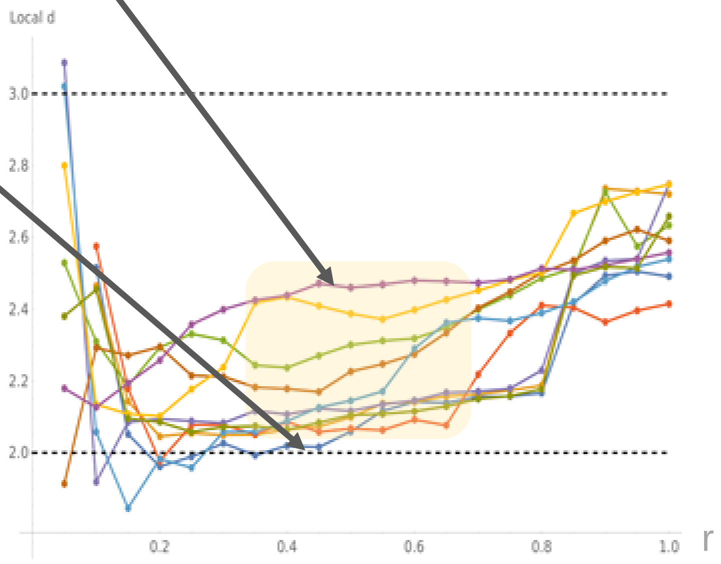
Use the “**most persistent" + lowest minimum** as the estimator of the Intrinsic Dimension



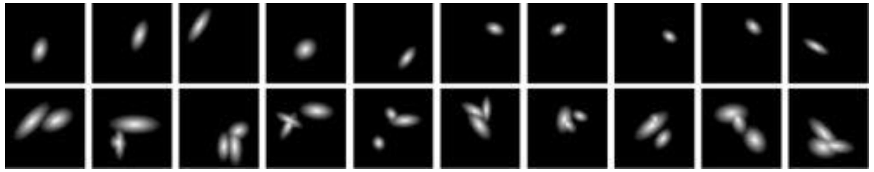
FCI can be easily multiscaled to extend to curved manifolds

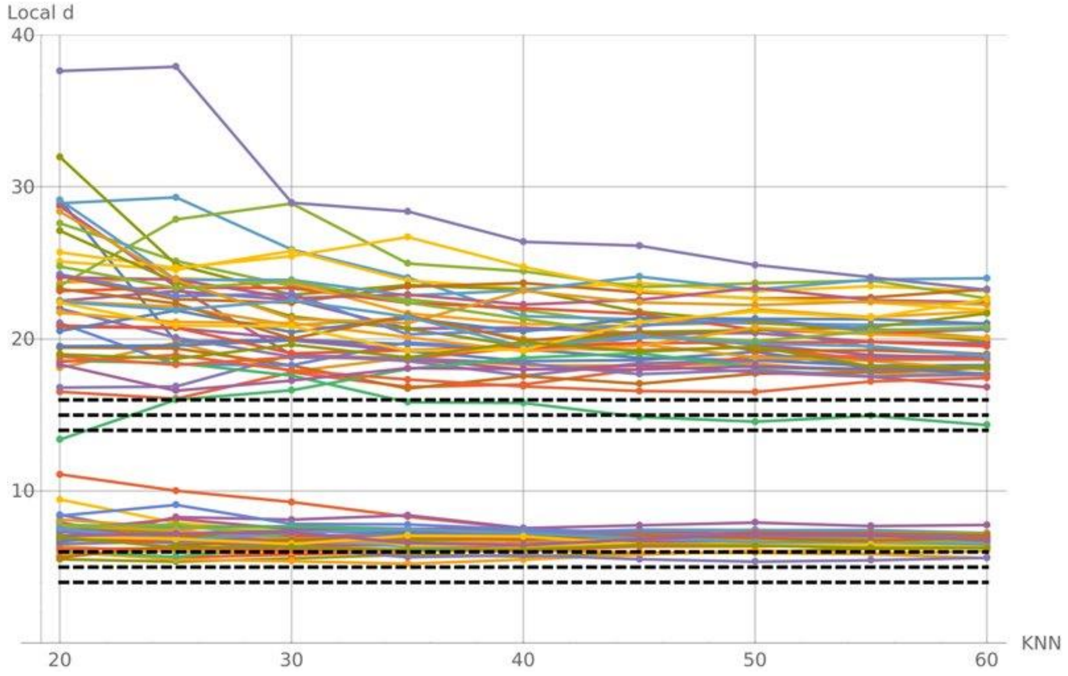


Use the “**most persistent" + lowest minimum** as the estimator of the Intrinsic Dimension



5 degrees of freedom per blob:
translation x, translation y, eccentricity, scale, tilt





Part 1: Intrinsic Dimension Estimation (IDE)

1) Accurate IDE is tricky
2) FCI + multiscalability => first step towards more robust IDE
3) A lot of work to do! Rationalize the multiscale approach into a statistically robust estimator

Erba, V., Gherardi, M., & Rotondo, P. (2019). Intrinsic dimension estimation for locally undersampled data. *Scientific reports*, *9*(1), 1-9

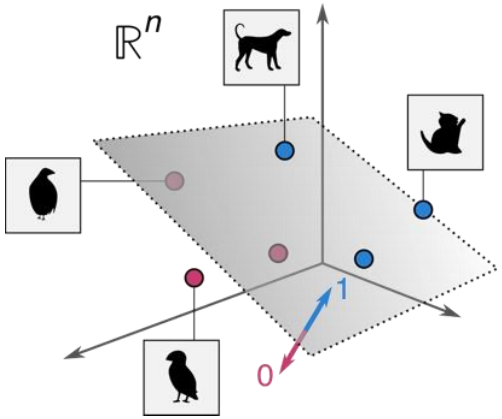
https://github.com/vittorioerba/pyFCI

Part 2: Linear classification of geometrically structured data

1) Define the polytope model
2) Review of expressivity of linear classifiers for unstructured data
3) Expressivity of linear classifiers for structured data
4) A data-driven phase transition

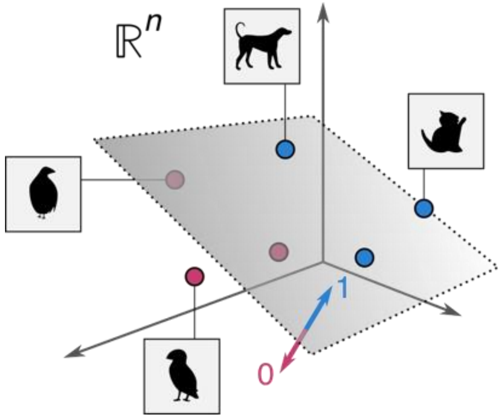
Geometric correlations => Classification of manifolds

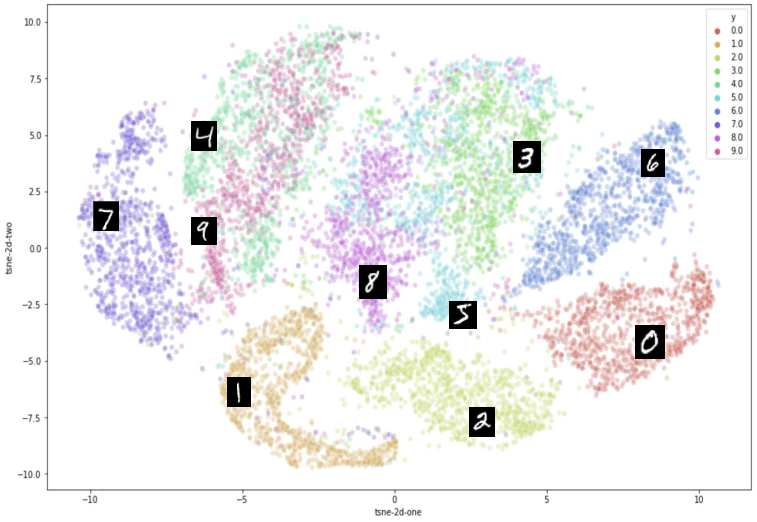
Before:
Linear classification of
UNSTRUCTURED data

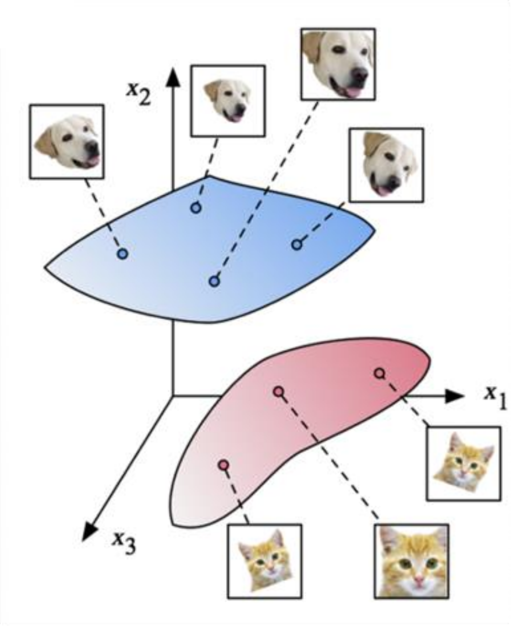


Geometric correlations => Classification of manifolds

Before:
Linear classification of
UNSTRUCTURED data







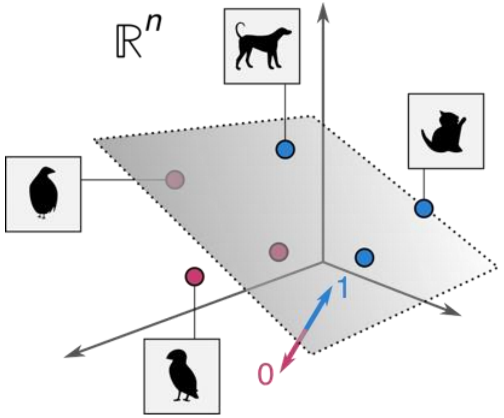
Chung, S., Lee, D. D., & Sompolinsky, H. (2016). Linear readout of object manifolds. *Physical Review E*, *93*(6), 060301.

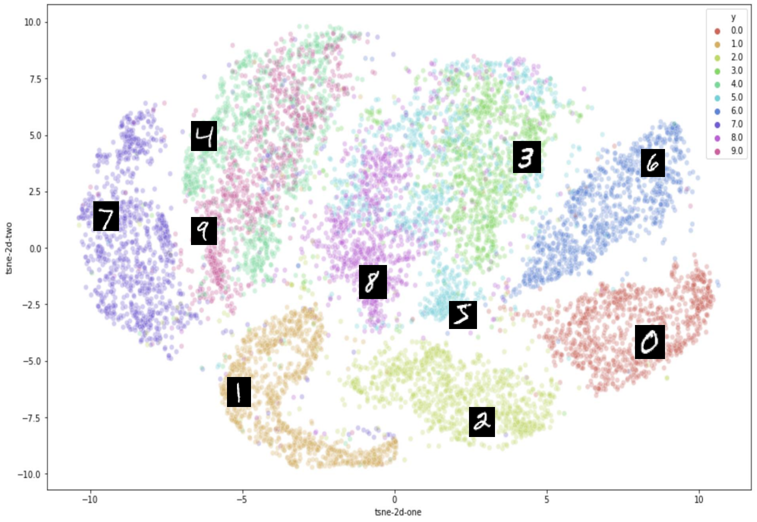
Chung, S., Lee, D., & Sompolinsky, H. (2018). Classification and Geometry of General Perceptual Manifolds *Phys. Rev. X, 8*, 031003.

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature communications*, *11*(1), 1-13.

Geometric correlations => Classification of manifolds

Before:
Linear classification of
UNSTRUCTURED data

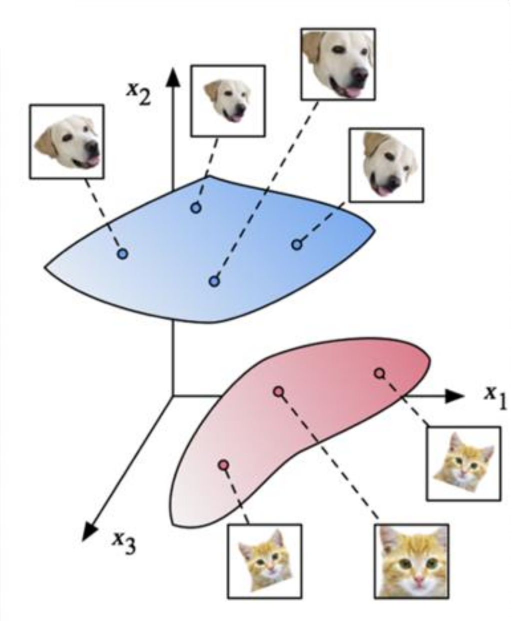


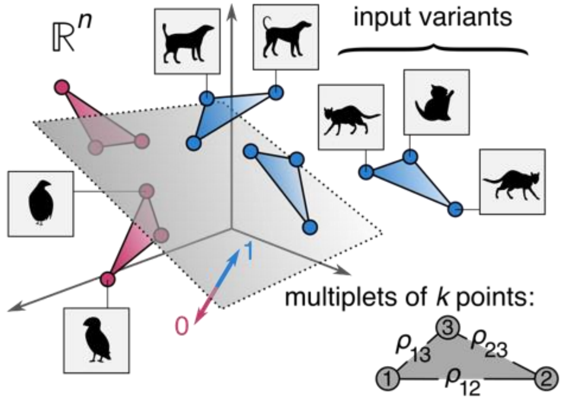


1) Expand points into manifolds
2) Restrict the learning architectures to those that classify coherently points in the same manifolds

After:
Linear classification of
STRUCTURED data

Classification of p convex polytopes with k vertices with prescribed overlaps





$$\xi_a^\mu \in S^n$$

$$\rho_{ab} = \sum_{i=1}^{n} (\xi_a^\mu)_i (\xi_b^\mu)_i$$

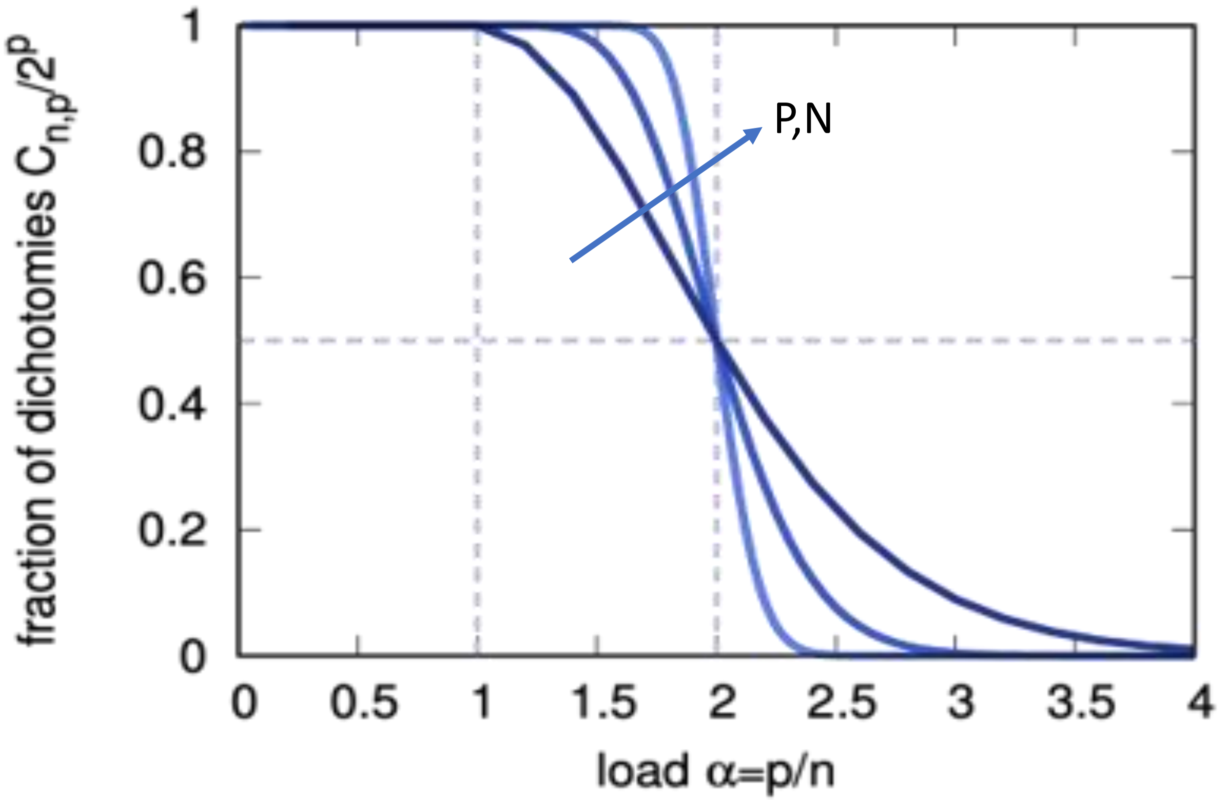
Chung, S., Lee, D. D., & Sompolinsky, H. (2016). Linear readout of object manifolds. *Physical Review E*, *93*(6), 060301.

Chung, S., Lee, D., & Sompolinsky, H. (2018). Classification and Geometry of General Perceptual Manifolds *Phys. Rev. X*, *8*, 031003.

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature communications*, *11*(1), 1-13.

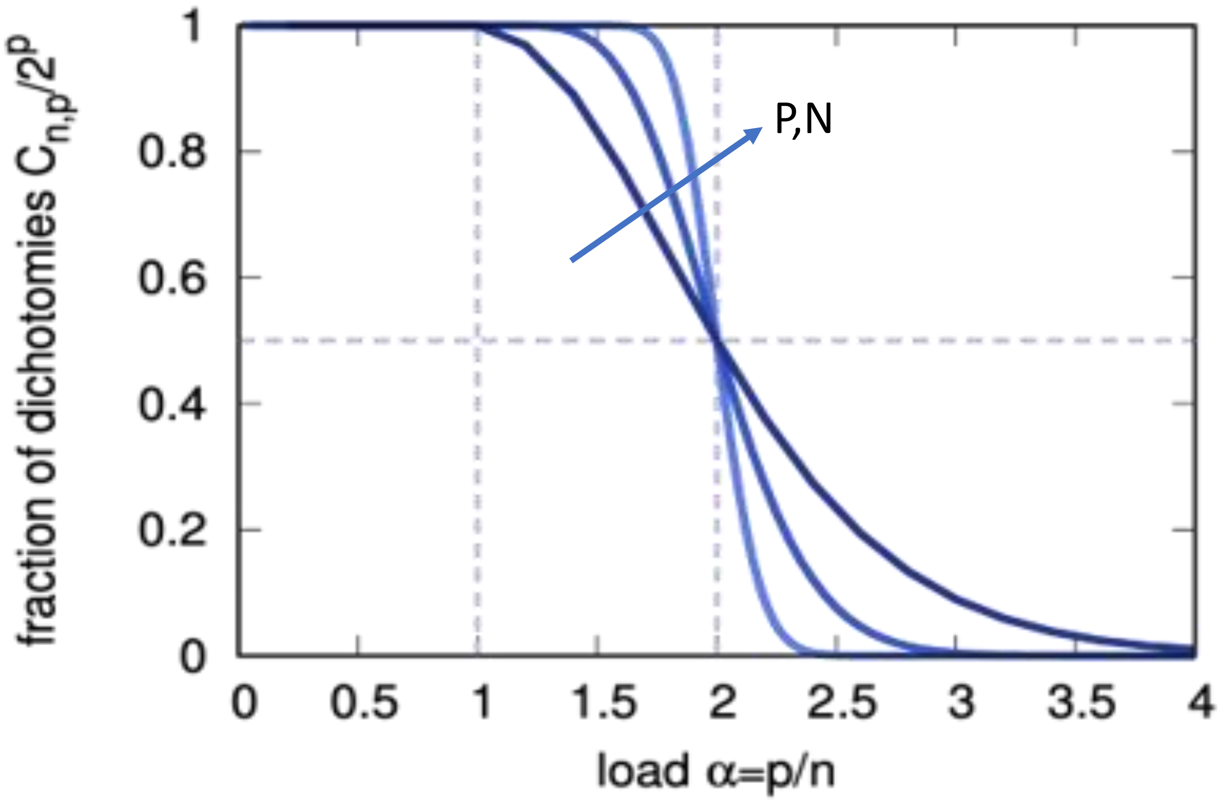
Understand how linear classification properties are changed by data structure

The expressivity of linear classifiers can be computed using two complementary frameworks



P = # of training samples
N = # of dimensions
C(N,P) = # of dichotomies realizable by a linear classifiers

The expressivity of linear classifiers can be computed using two complementary frameworks



P = # of training samples
N = # of dimensions
C(N,P) = # of dichotomies realizable by a linear classifiers

Combinatorial (Cover)

$$C_{n,p} = 2\sum_{k=0}^{n-1}\binom{p-1}{k}$$

$$\alpha_c = 2$$

- Exact results for full C(N,P) curves

- Additional insights:
  C(N,P) independent on the position of the points
  Finite size corrections

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3), 326-334.

Statistical Physics (Gardner)

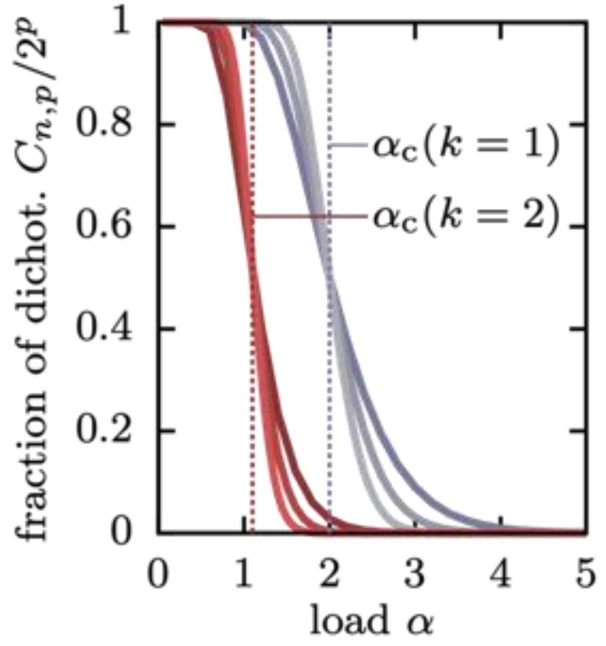
$$\int D^n W \prod_{\mu=1}^{p} \theta\left(\sigma^\mu \sum_{i=1}^{n} W_i \xi_i^\mu\right)$$

- Exact results for storage capacity

- Position of the points + labels = quenched disorder

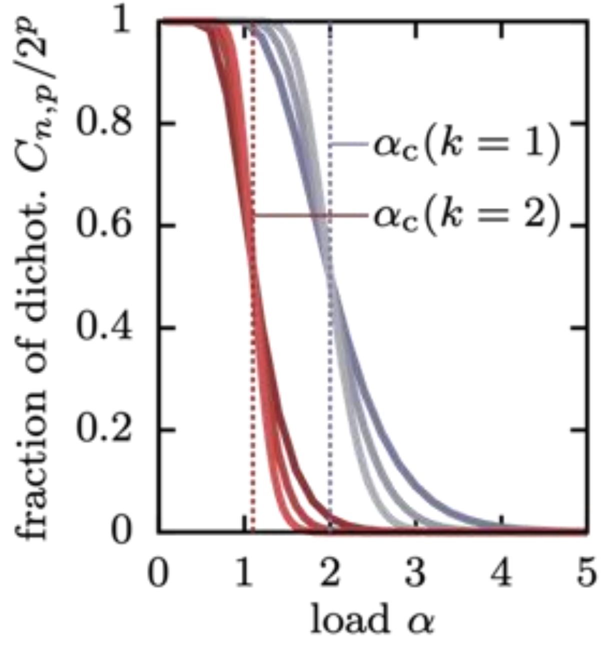
Gardner, E. (1988). The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, *21*(1), 257.

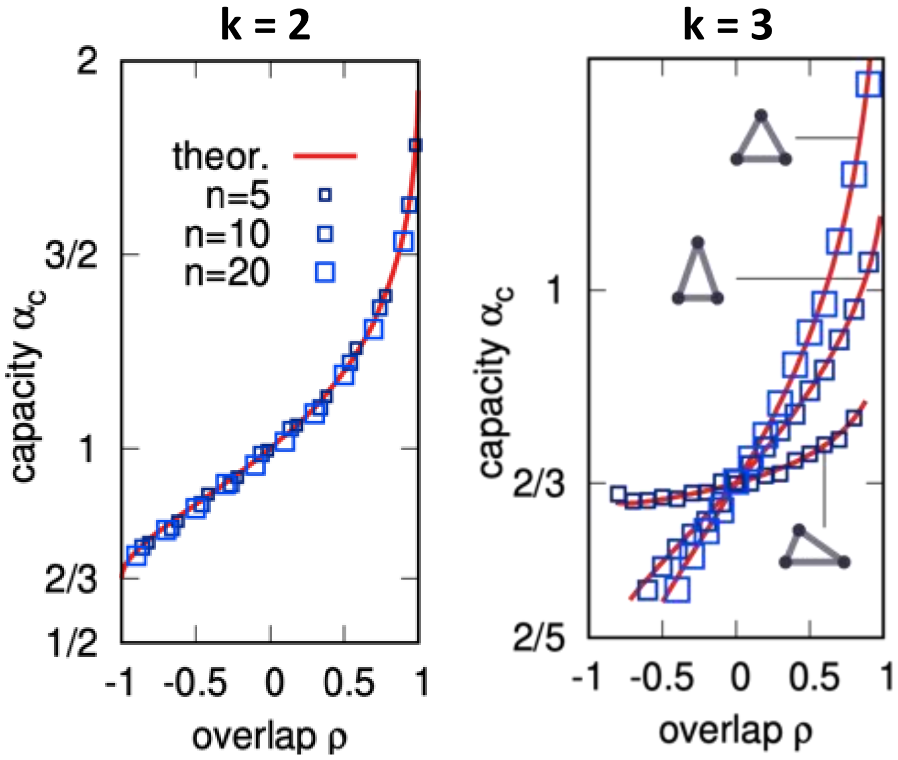
Both frameworks extend to geometrically structured data

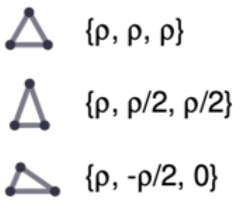


P = # of training samples
N = # of dimensions
K = # of vertices of polytopes
C(N,P,K) = # of dichotomies of polytopes realizable by a linear classifiers

Both frameworks extend to geometrically structured data

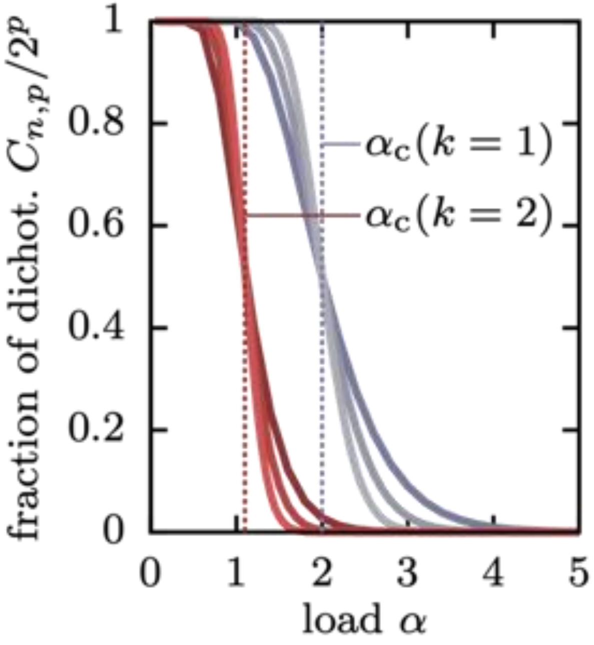






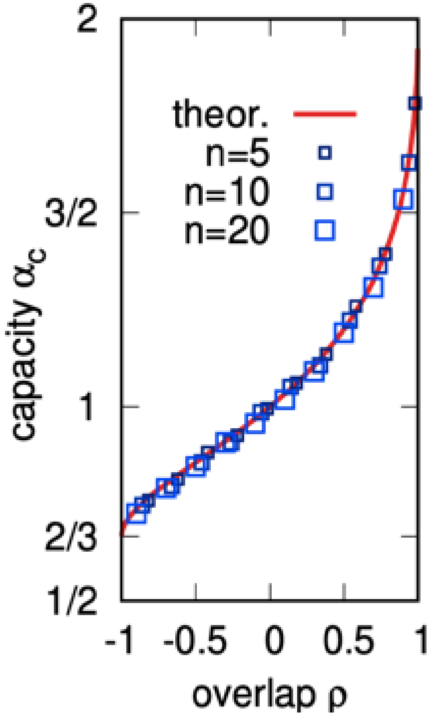
P = # of training samples
N = # of dimensions
K = # of vertices of polytopes
C(N,P,K) = # of dichotomies of polytopes realizable by a linear classifiers

Both frameworks extend to geometrically structured data

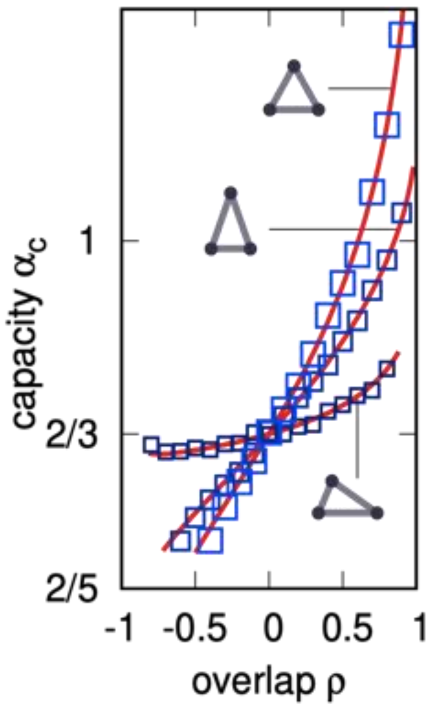


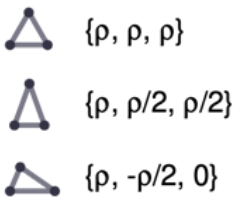
P = # of training samples
N = # of dimensions
K = # of vertices of polytopes
C(N,P,K) = # of dichotomies of polytopes realizable by a linear classifiers

k = 2



k = 3





Combinatorial

$$C^{(2)}_{n,p+1} = \Psi_2(\rho) C^{(2)}_{n,p} + C^{(2)}_{n-1,p} + [1-\Psi_2(\rho)] C^{(2)}_{n-2,p}$$

Rotondo, P., Lagomarsino, M. C., & Gherardi, M. (2020). Counting the learnable functions of geometrically structured data. *Physical Review Research*, *2*(2), 023169.

Statistical Physics

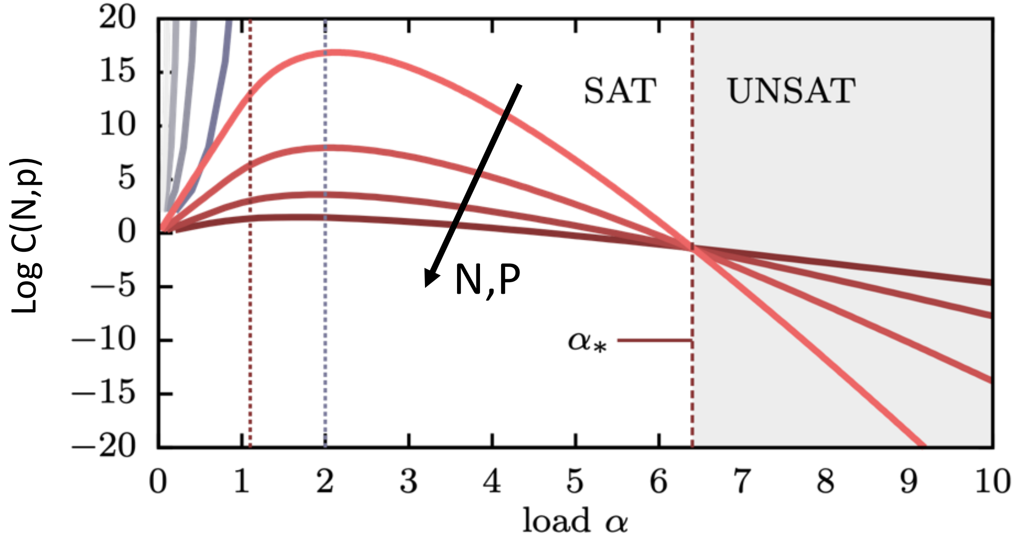
$$\int D^n W \prod_{\mu,a=1}^{p,k} \theta\left[\sigma^\mu \sum_{i=1}^{n} W_i (\xi^\mu_a)_i\right]$$

$$\alpha_c = \frac{2}{3 - 2\Psi_2(\rho)}$$

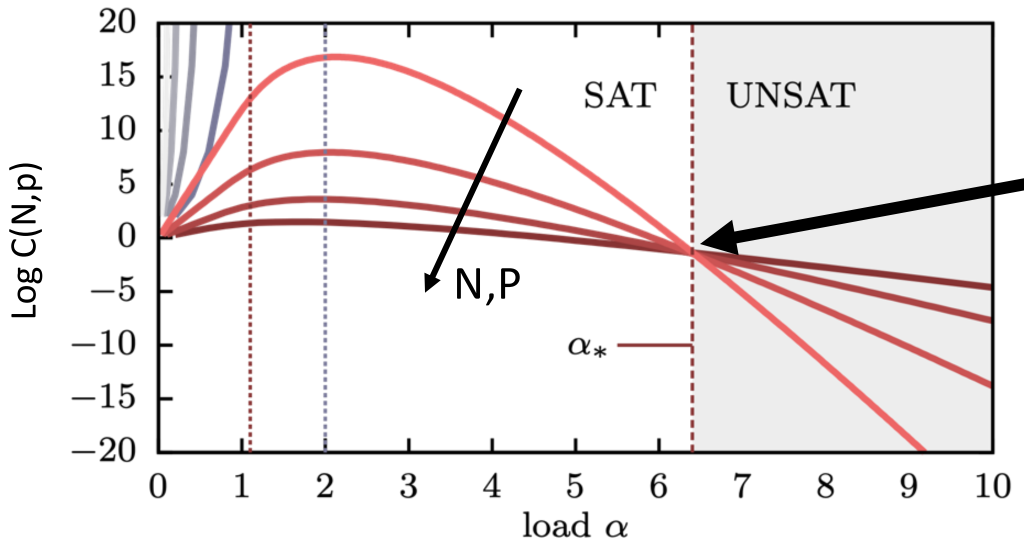
Borra, F., Lagomarsino, M. C., Rotondo, P., & Gherardi, M. (2019). Generalization from correlated sets of patterns in the perceptron. *Journal of Physics A: Mathematical and Theoretical*, *52*(38), 384004.

Chung, S., Lee, D., & Sompolinsky, H. (2018). Classification and Geometry of General Perceptual Manifolds *Phys. Rev. X, 8*, 031003.

One more thing: a novel phase transition



One more thing: a novel phase transition



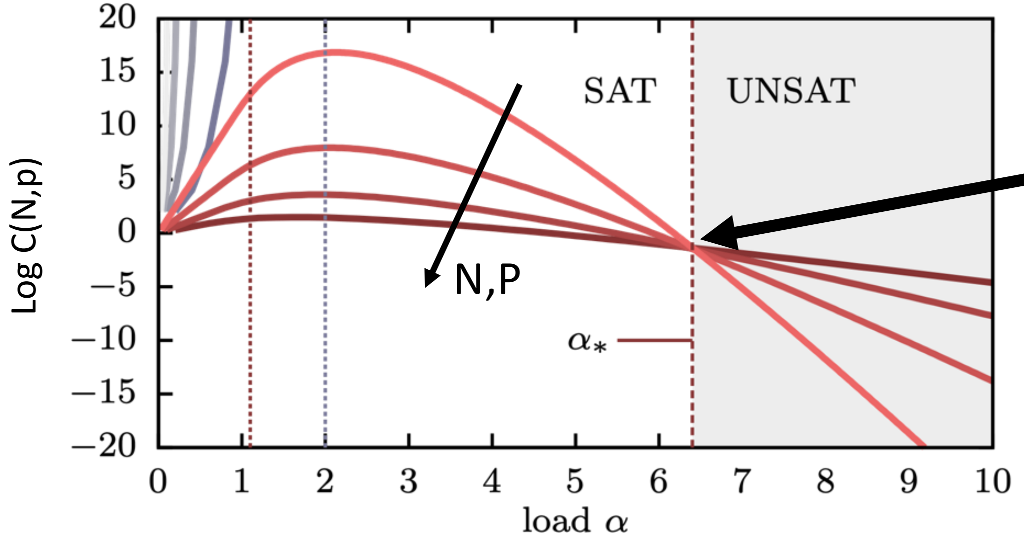
Combinatorial

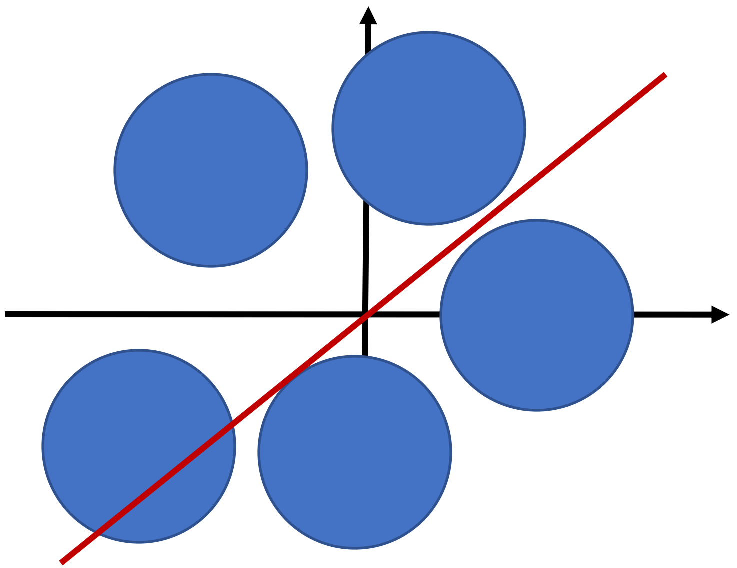
$$(\alpha_* + 1)\log(\alpha_* + 1) - \alpha_* \log \alpha_* + \\ + (\alpha_* - 1)\log \theta_k(0) + \log \theta_k(1) = 0$$

Mauro Pastore, Pietro Rotondo, Vittorio Erba, & Marco Gherardi (2020). Statistical learning theory of structured data *Physical Review E, 102*(3).

Rotondo, P., Pastore, M., & Gherardi, M. (2020). Beyond the Storage Capacity: Data-Driven Satisfiability Transition *Phys. Rev. Lett., 125*, 120601.

One more thing: a novel phase transition





Combinatorial

$$(\alpha_* + 1)\log(\alpha_* + 1) - \alpha_* \log \alpha_* + \\ + (\alpha_* - 1)\log \theta_k(0) + \log \theta_k(1) = 0$$

Mauro Pastore, Pietro Rotondo, Vittorio Erba, & Marco Gherardi (2020). Statistical learning theory of structured data*Physical Review E, 102*(3).

Rotondo, P., Pastore, M., & Gherardi, M. (2020). Beyond the Storage Capacity: Data-Driven Satisfiability Transition*Phys. Rev. Lett., 125*, 120601.

Is a set of manifolds classifiable at all?

$$\int D^p\sigma \int D^n W \prod_{\mu,a=1}^{p,k} \theta\left[\sigma^\mu \sum_{i=1}^{n} W_i(\xi_a^\mu)_i\right]$$

Transition common to all models of extended geometry

Part 2: Linear classification of structured data (simplex model)

1) Polytope model is analytically tractable
2) Classification of extended datasets induces new phase transition
3) Check other kinds of data structure models?

https://journals.aps.org/prresearch/abstract/10.1103/PhysRevResearch.2.023169

https://journals.aps.org/pre/abstract/10.1103/PhysRevE.102.032119

https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.125.120601

Thank you!