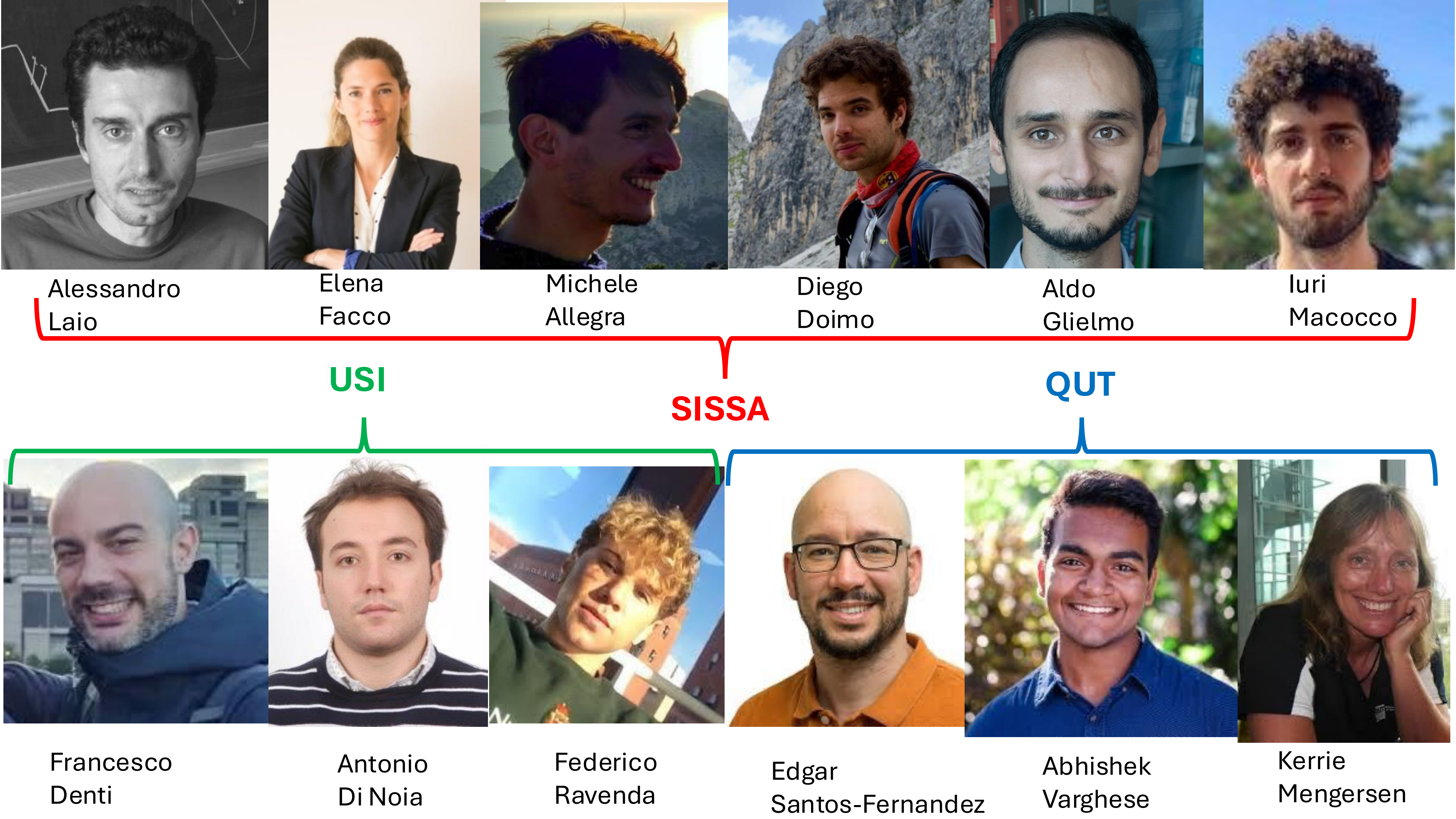
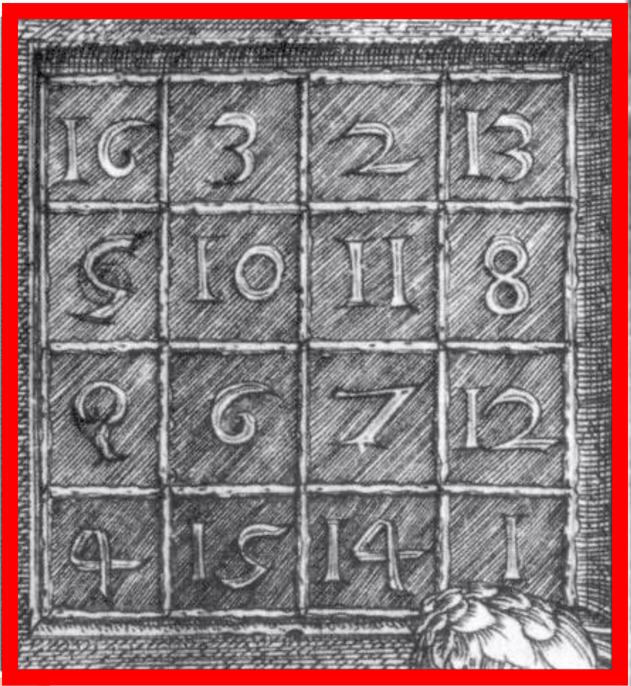
Intrinsic Dimension for continuous data

Antonietta Mira
Università della Svizzera italiana, Lugano



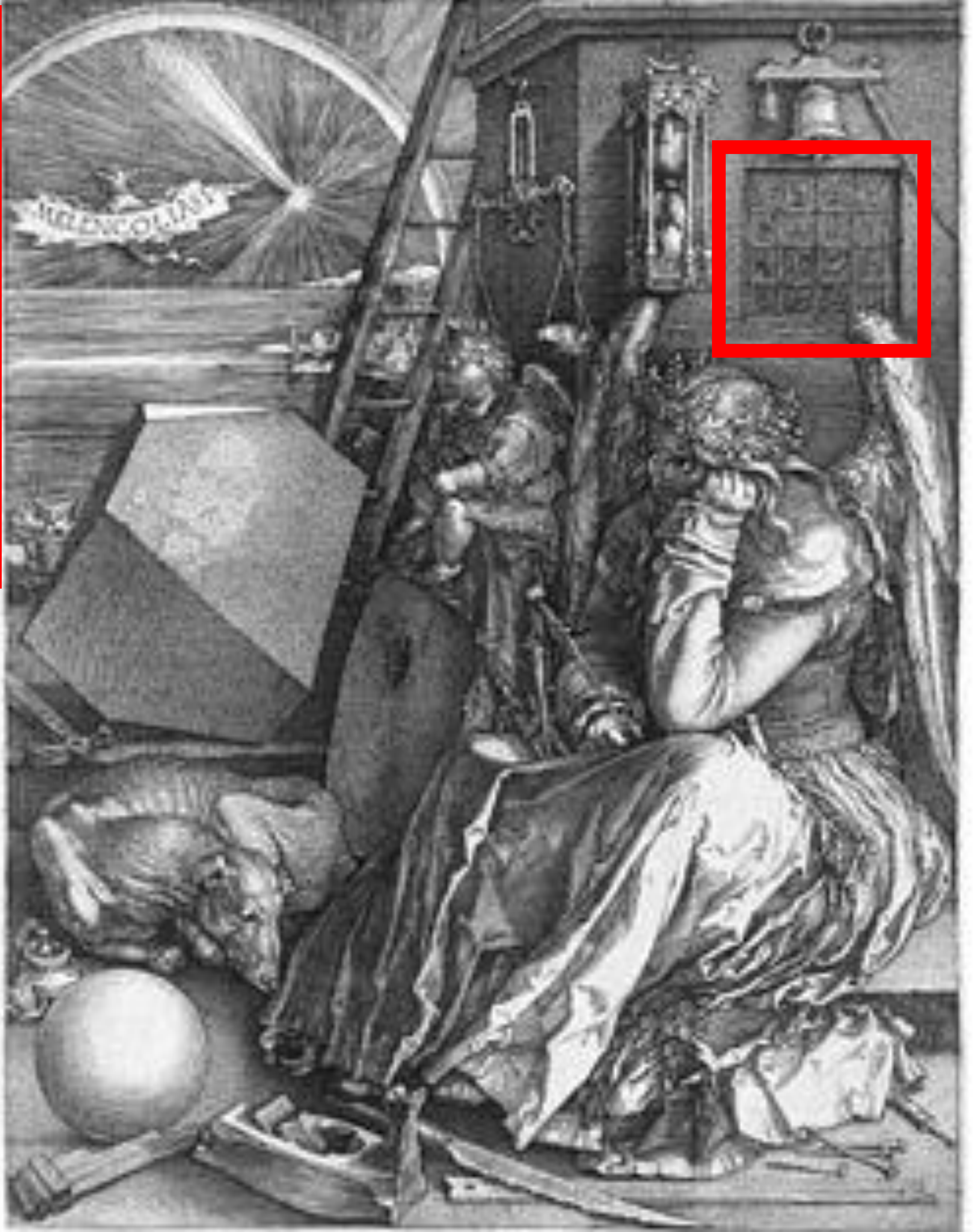
Alessandro Laio
Elena Facco
Michele Allegra
Diego Doimo
Aldo Glielmo
Iuri Macocco
USI
SISSA
QUT
Francesco Denti
Antonio Di Noia
Federico Ravenda
Edgar Santos-Fernandez
Abhishek Varghese
Kerrie Mengersen

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1



A. Durer, 1514

synthesis of Earth and Heaven, search for the intrinsic dimension of all things



- Interdisciplinary: statistics + physics + domain experts
- 2019 - ongoing:

  - 5 (SISSA + USI)
  - 2 (QUT + USI)
  - 2 (USI)

Quote / Quiz

”The non-mathematician, is seized by a mysterious shuddering when (s)he hears of ‘four-dimensional’ things, (s)he is seized by a feeling, which is very similar to the thoughts awakened by the occult.”

Albert Einstein, 1920

Classical mechanics:
space and time = separate entities

Special relativity:
space and time = interwoven into a four dimensional

Unified construct known as ”space-time”

From physics to statistics / DS / ML

Dimension expansion

- data: EM, ABC, Knockoffs, LLM …
- parameter: hybrid MCMC, Slice sampler, mixture models

Dimension reduction

- data:
  - Kernel PCA, Canonical Correlation Analysis, Clustering,
  - VAE, t-SNE, Isomap, manifold learning, scketching,
  - random projections, projection pursuit
- parameter:
  - Variable / feature selection, Factor analysis

Focus on dimension reduction

Open questions

A subset of the variables or non-linear combinations thereof
are often sufficient to describe a real-world data set

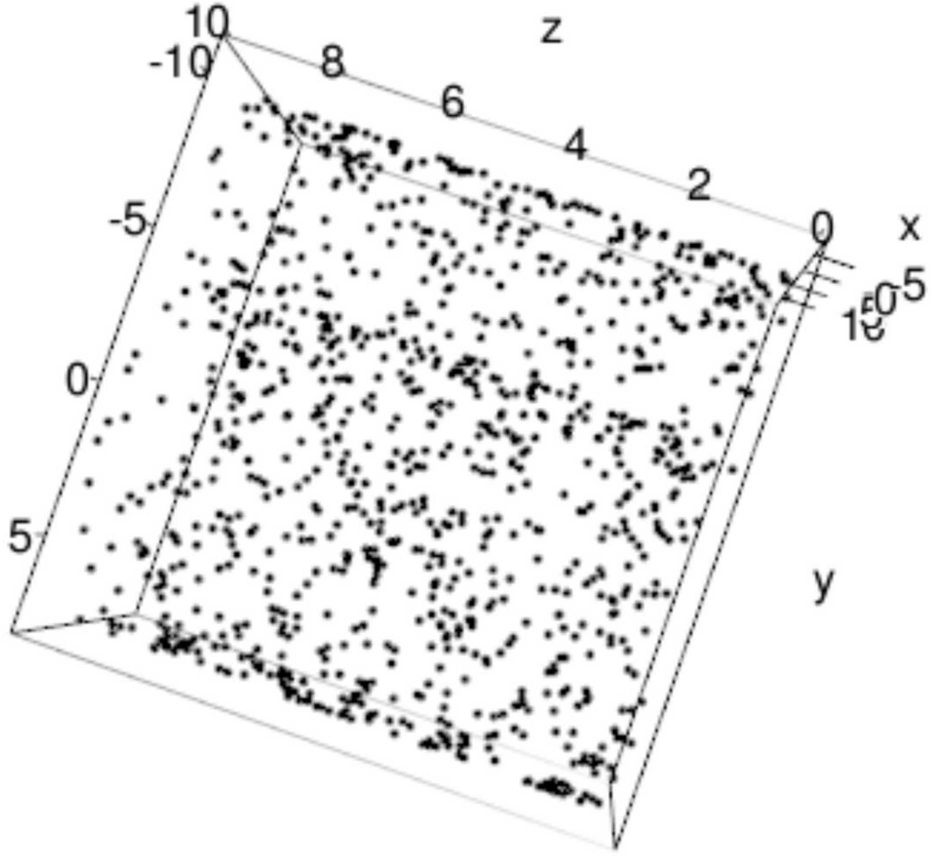
How many (and which) variables are needed to summarize a data set
without significant information loss?

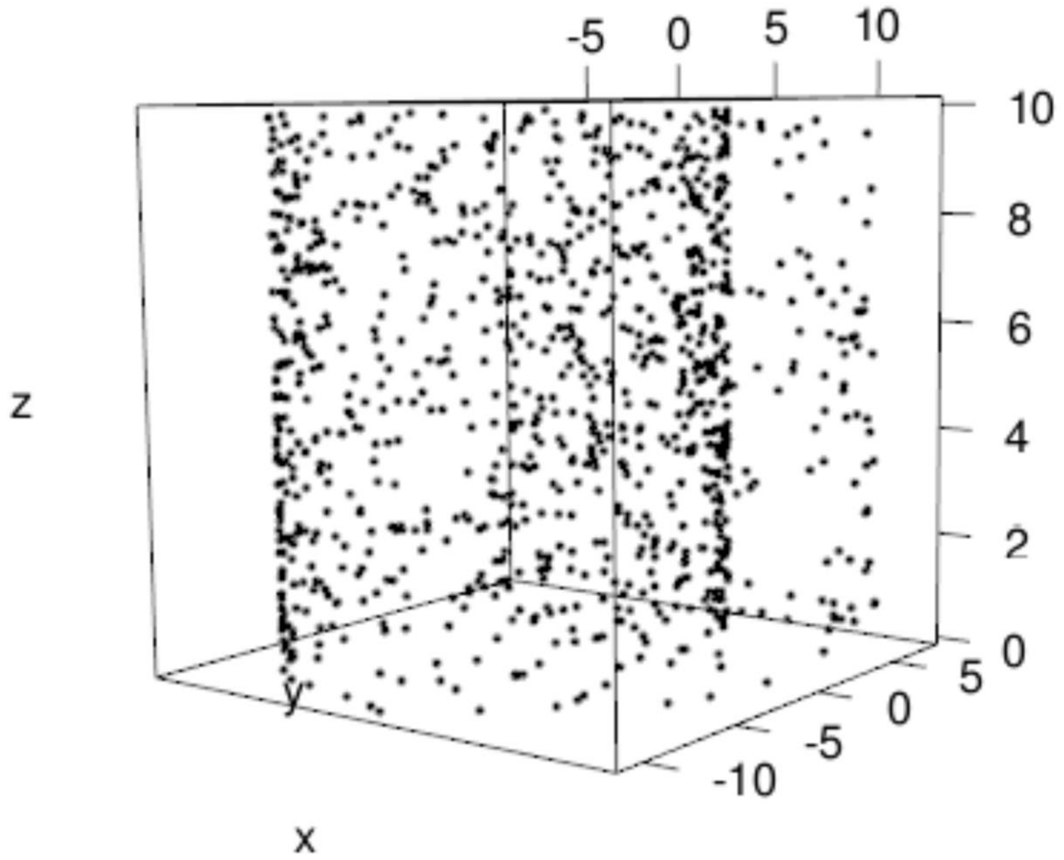
What is the appropriate scale at which one should analyze data?

Two questions, often considered unrelated, but strongly entangled,
can be addressed within a unified ”scale-dimension” framework

We introduce an approach in which the optimal number of variables and the optimal scale
are determined self-consistently, bypassing the scale at which data are affected by noise

A matter of perspective





Can you guess the dimension of the support of the data generating process, d?

Motivation: dimensionality reduction

In the Swissroll example $D = 3$ and $d = 2$

d, is called the intrinsic dimension (ID) of the data

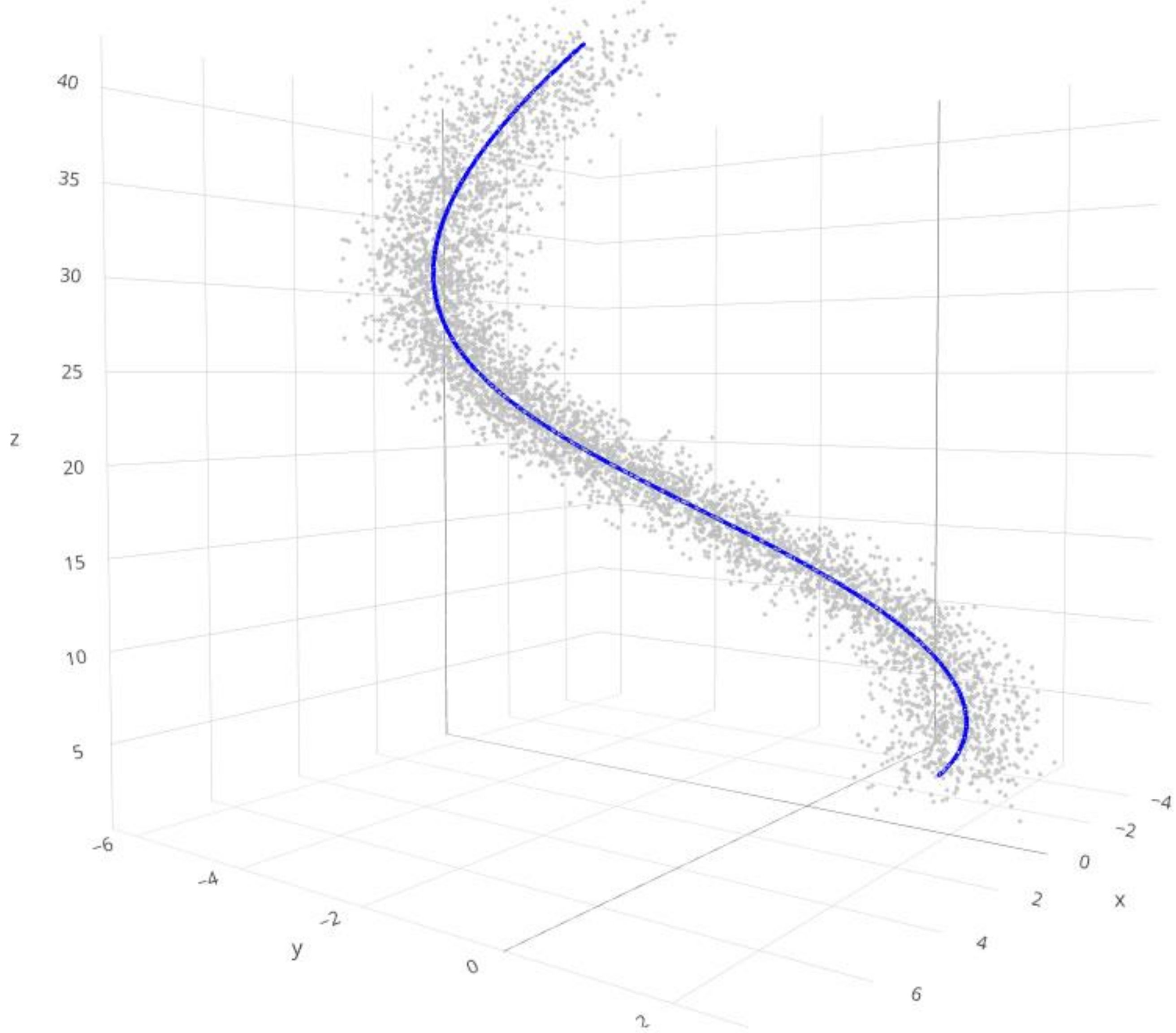
ID = needed for manifold learning (D. Dunson et al.)

ID = regression with covariates on a manifold (J. Rousseau et al.)

ID = key concept in unsupervised learning and feature selection

ID = lower bound to number of variables needed to describe a system

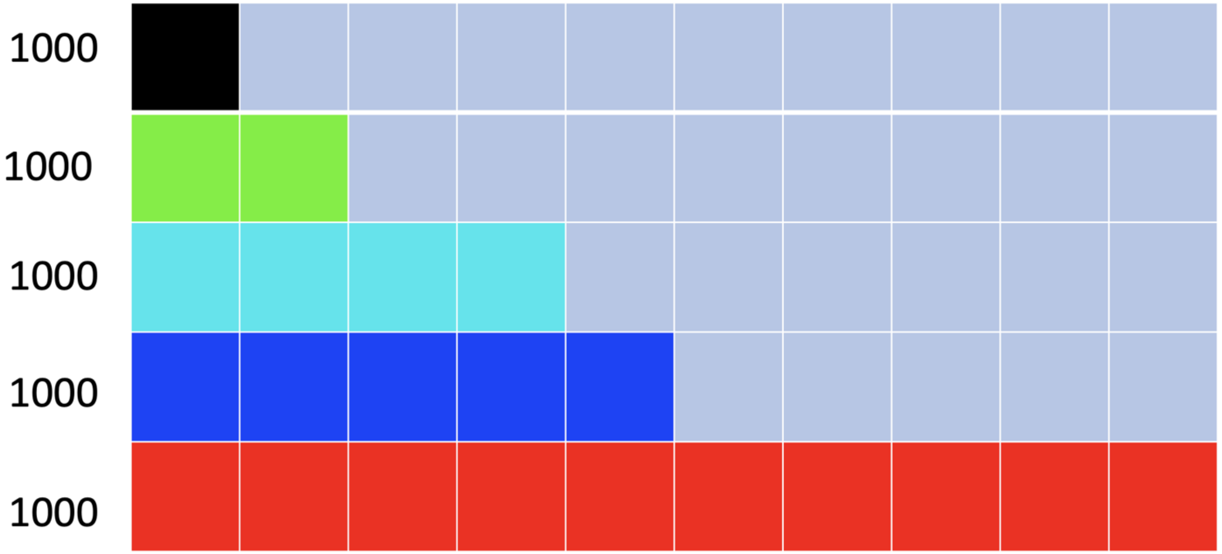
PROBLEM 1: Noise makes the identification of the ID har

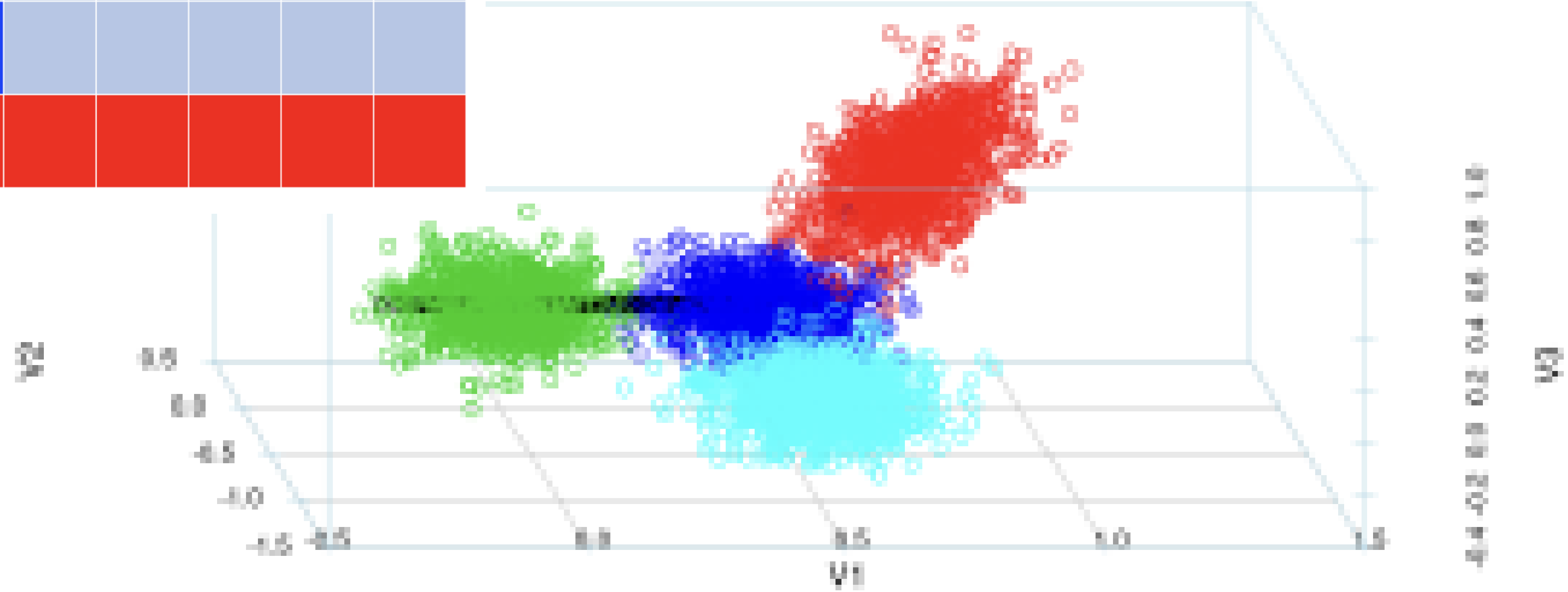


PROBLEM 2: Sometimes the ID can vary within the same dataset

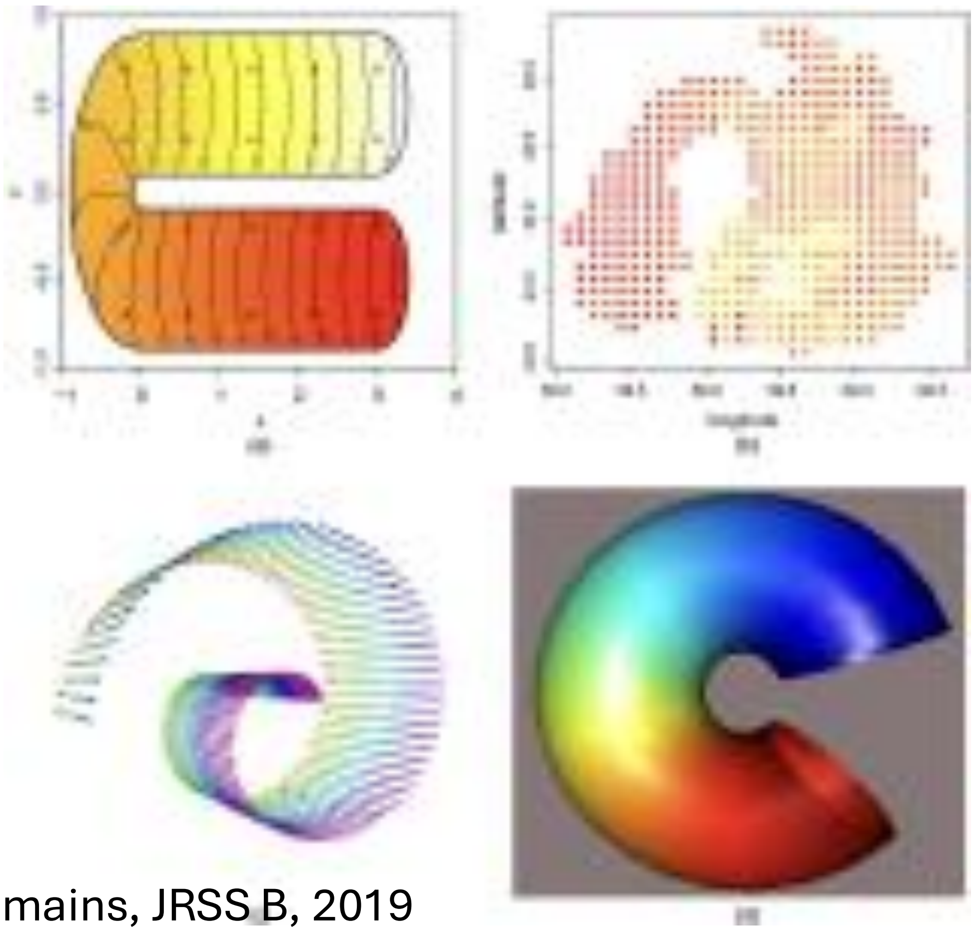
5 x1000 observations generated from 5 Gaussians of dimensions 1, 2, 4, 5 and 9, partially overlapping: 3 dim projection

Data matrix = 5000 x 9





PROBLEM 3: data can be discrete or continuous



D. Dunson et al. GP on contrained domains, JRSS B, 2019

A - a test function increases smoothly within a U-shaped boundary B - remote sensed chlorophyll data in the Aral sea

C - a spiralling band in a three-dimensional Euclidean space + data D - Bitten torus + data on the surface

Regions with the same ID host points differing in core properties:

- folded vs unfolded state in protein configurations
- active vs non-active regions in brain imaging data
- patients vs controls in gene expression data
- firms with different financial risk in balance sheets data
- winning vs losing teams in basketball data (AoAS)
- country specific NPI in pandemic evolution (Sci. Rep.)
- ID of an undirected unweighted network (Sci. Rep.)
- Identified vs unidentified models via ID in MCMC
- ID of layers in a CNN and transformes in LLM

A simple topological feature uncovers a rich data structure

What is the Intrinsic Dimension (ID)?

Dimension of the support of the data generating distribution

Different estimation approaches:

projective, topological, Nearest-Neighbours (NN) ...

Lack of statistically sound estimation procedures

Aim: directly target the ID as a parameter to estimate

Fundamental issue: scale dependence

Goal: statistical guarantees

- Allow for uncertainty quantification
- Rely on exact or asymptotic distributional results

NN based ID estimators: 2 classes

For each point i in the data set

- NN order class
  fix two NN orders: $n_2 > n_1$
  find distance from i to the two NN: $r_{i,n_2} > r_{i,n_1}$
  statistics: $\mu_i = r_{i,n_2}/r_{i,n_1}$

- NN distance class
  fix two distances: $r_2 > r_1$
  count points within that distance from i: $n_{i,r_2} > n_{i,r_1}$
  statistics: $n_{i,r_1}|n_{i,r_2}$

We find the distribution of the two statistics = function (d)

CLASS I: ID Estimators based on fixing NN order:

- TWO-NN = ratios of distances 2nd to 1st NN
- GRIDE = Generalized Ratios ID Estimator
- Hidalgo = Finite Mixture of TWO-NN ratios
- BNP Hidalgo = Infinite Mixture of TWO-NN ratios

CLASS II: ID Estimators based on fixing NN distance:

- I3D = ID for Discrete Data
- BIDE = Binomial ID Estimator
- ABIDE = Adaptive BIDE
- BABIDE = Bayesian ABIDE

PROBLEM SOLVING

TWO-NN requires a weaker local homogeneity assumption

GRIDE is more robust to noise in the data

Hidalgo and BNP Hidalgo robust if more than one ID exists

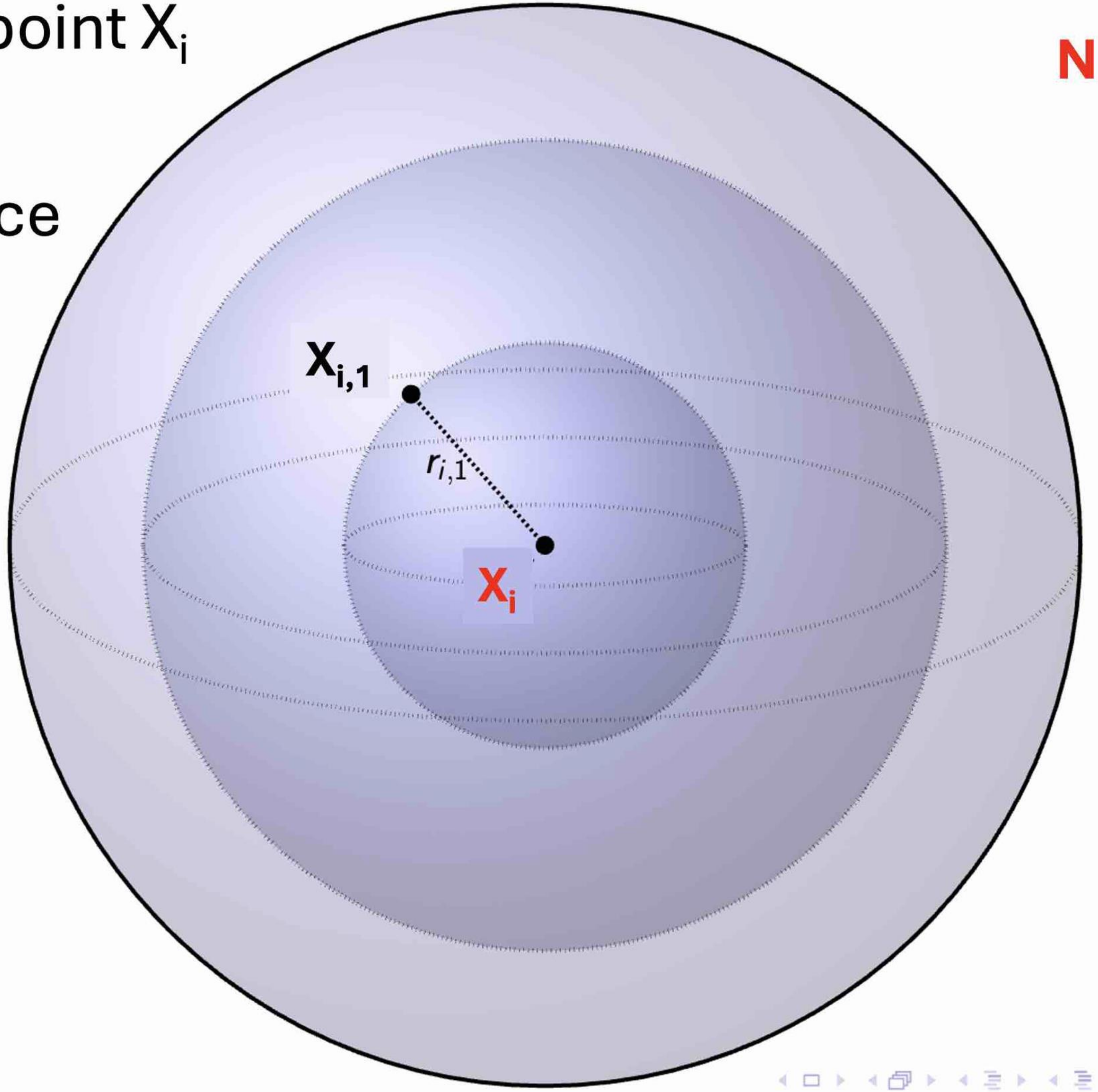
Adaptive ID estimators allow to escape the noise

Adaptive I3D for discrete data

$X_{i,1}$ = 1-st NN to point X_i

$r_{i,1}$ is their distance

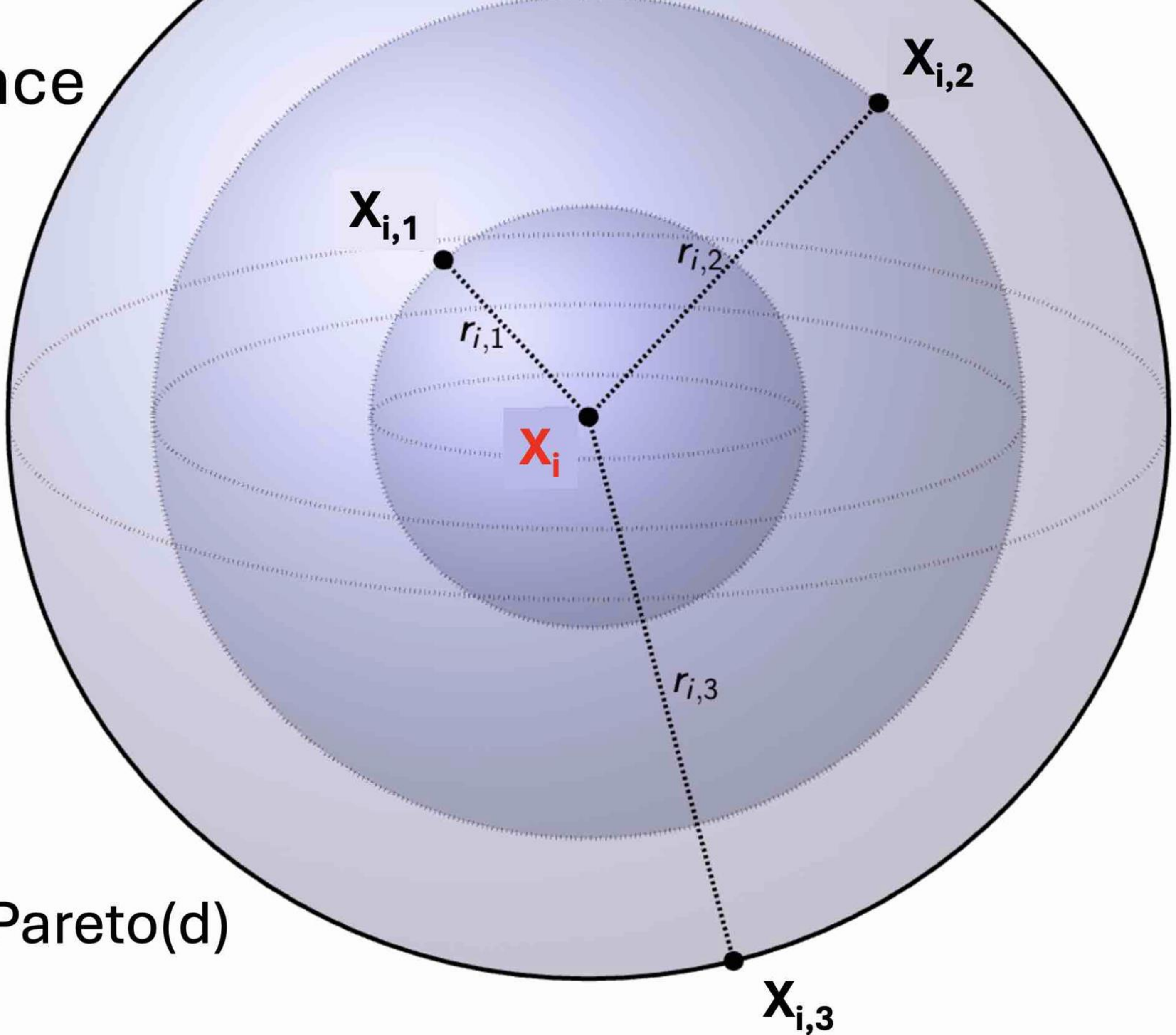
NN order class



$X_{i,j}$ = j-th NN to point X_i

NN order class

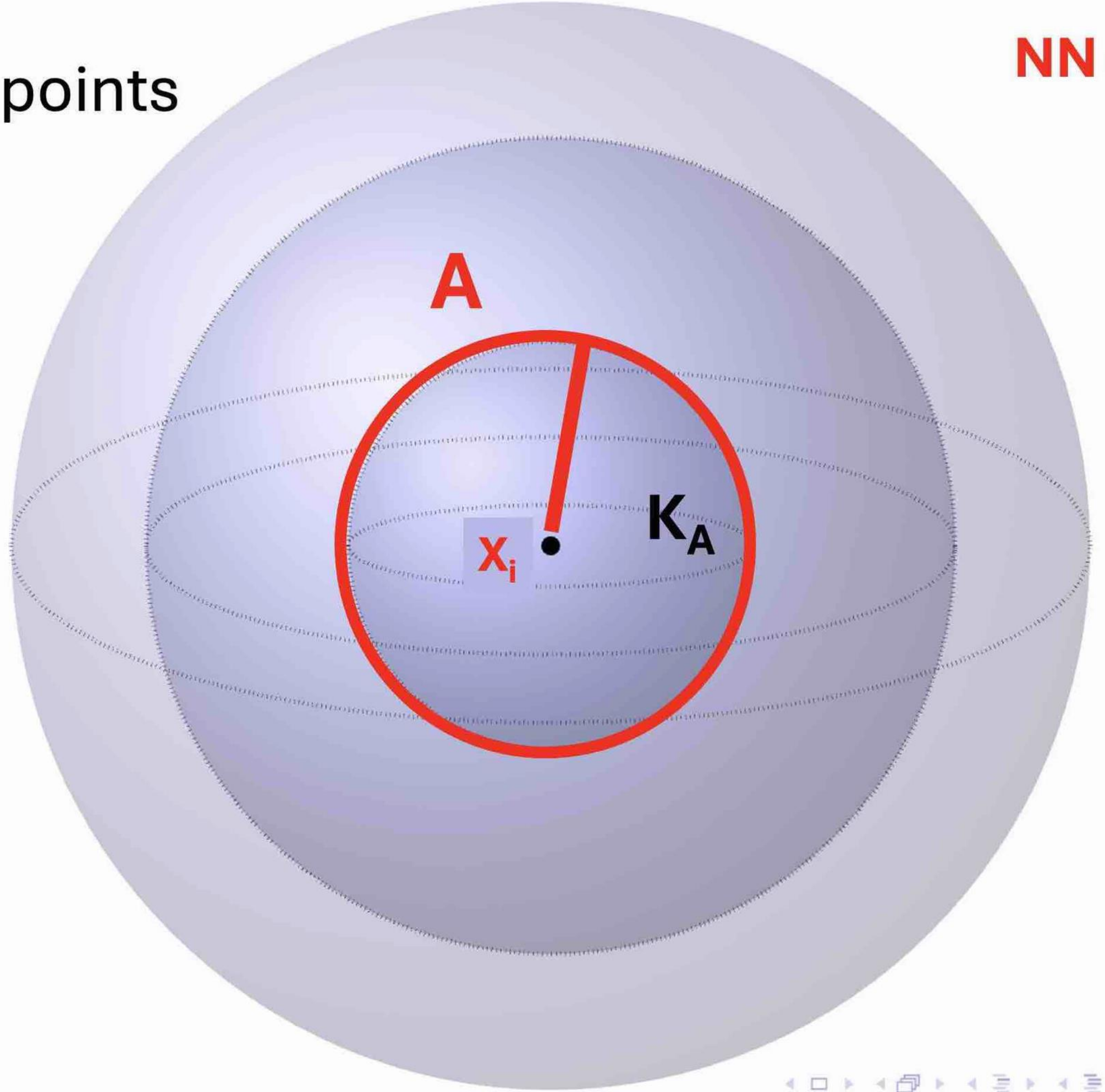
$r_{i,j}$ is their distance



$\mu_i = r_{i,2} / r_{i,1} \sim \text{Pareto}(d)$

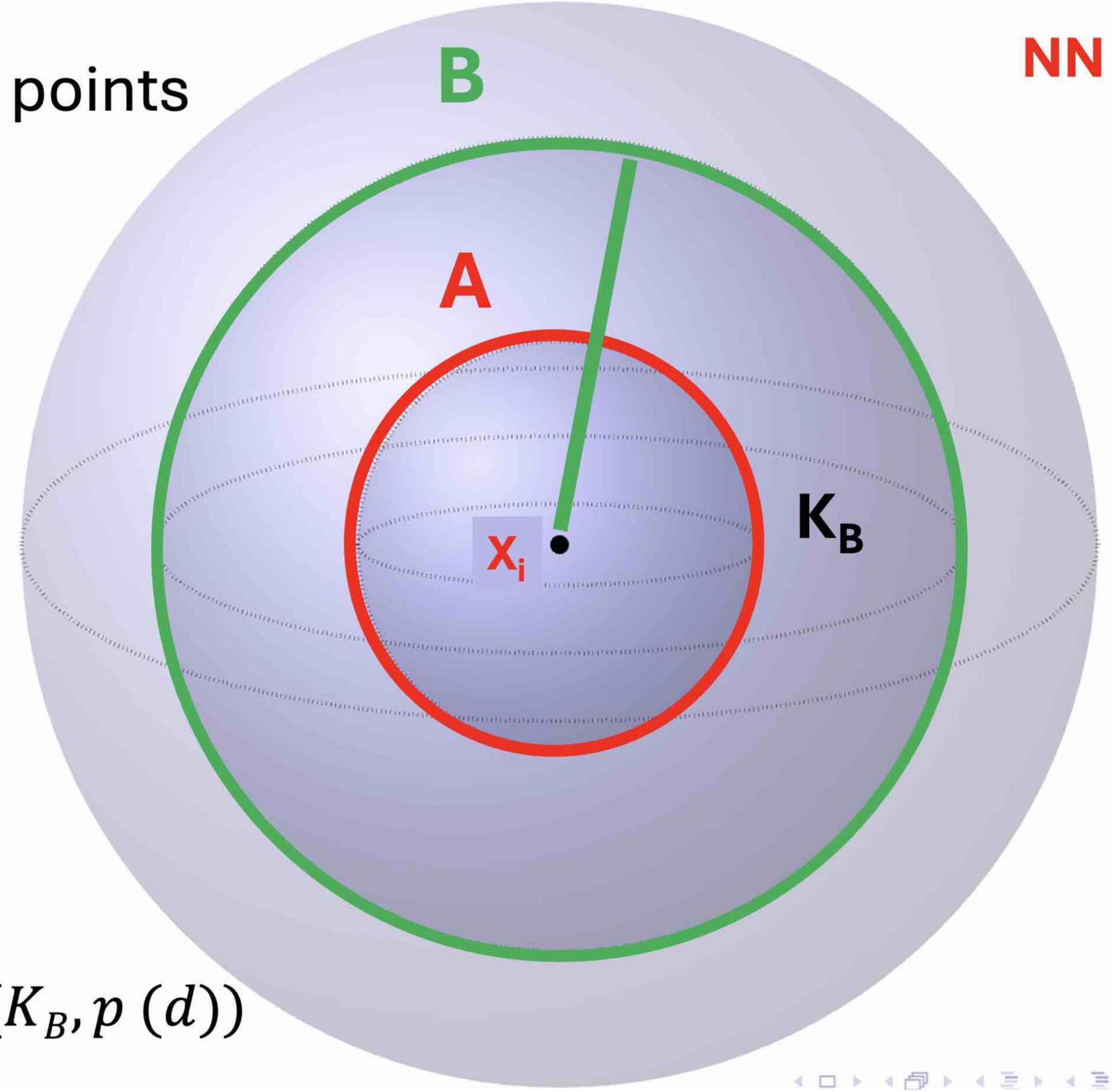
K_A = number of points in ball A

NN distance class



NN distance class

K_B = number of points in ball B



$K_A \mid K_B \sim Binom\,(K_B, p\,(d))$

- Let n_1, n_2 be two positive integers with $n_2 > n_1$ and define the ratio $\mu_{i,n_1,n_2} = r_{i,n_2}/r_{i,n_1}$ which is a.s. well defined for continuous data

Theorem

μ_{i,n_1,n_2} *has density*

$$f\mu_{i,n_1,n_2}(\mu) = \frac{d(\mu^d - 1)^{n_2 - n_1 - 1}}{\mu^{d(n_2-1)+1} B(n_2 - n_1, n_1)}, \quad \mu > 1 \tag{1}$$

where $B(\cdot,\cdot)$ is the Beta function and d is the id.

The proof (F. Denti) follows from the PPP local homegeneity assumption that implies that the volumes of the shell $v_{i,j}$ between the i-th and the j-th NN have an exponential distribution with parameter ρ_i

- Take $n_1 = 1$ and $n_2 = 2$, the density in (1) reduces to a $Pareto(1, d)$.
- TWO-NN MLE estimator is given by

$$\hat{d} = \frac{n-1}{\sum_{i=1}^{n} \log(\mu_i)}, \tag{2}$$

with $1-\alpha$ confidence interval $\tau_{1-\alpha} = \left[\frac{\hat{d}}{q^{1-\alpha/2}_{IG_{n,n-1}}}, \frac{\hat{d}}{q^{\alpha/2}_{IG_{n,n-1}}} \right]$.

- TWO-NN Bayesian estimator: prior $d \sim Gamma(a, b)$

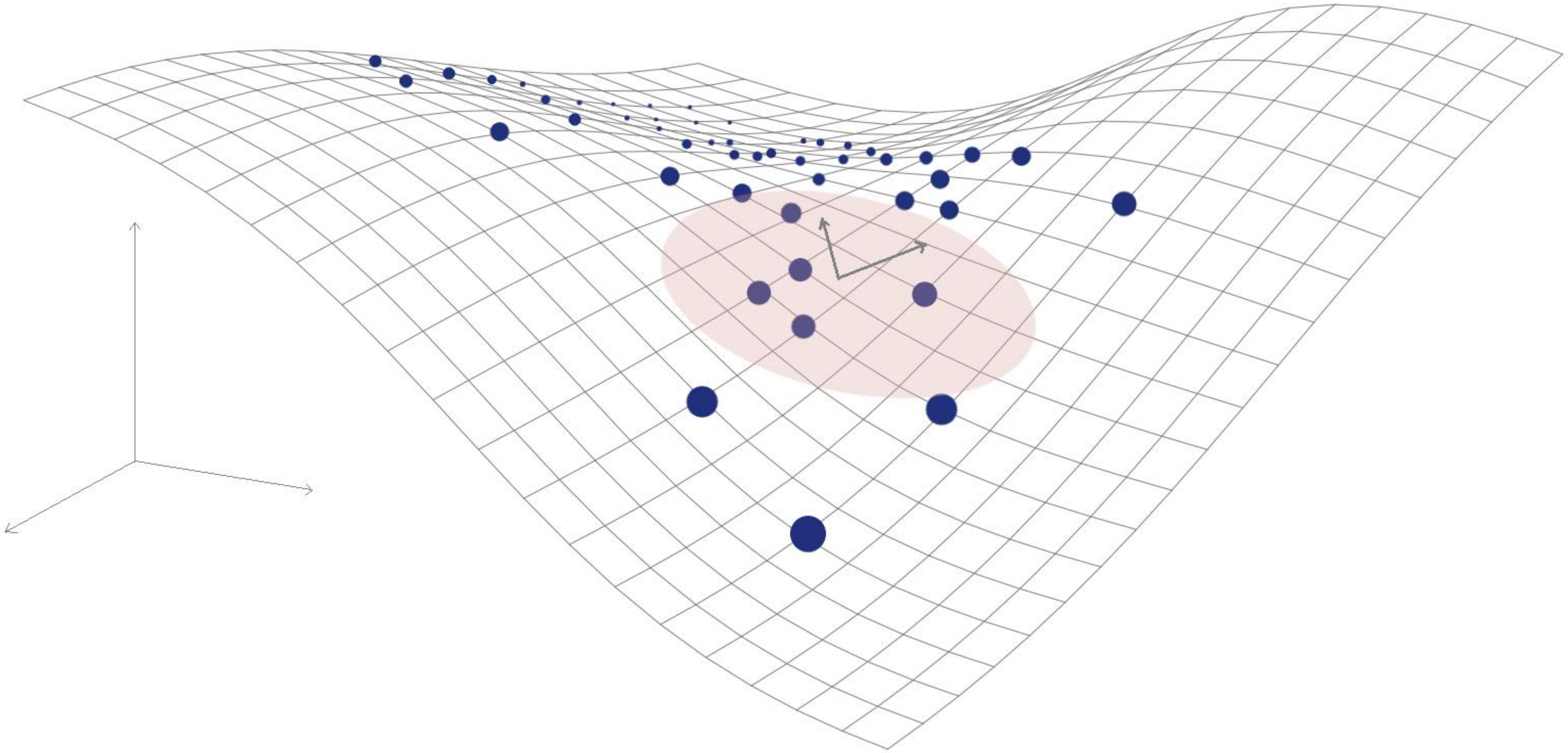
$$d | \mu_1, \ldots, \mu_n \sim Gamma\Big(a + n, b + \sum_{i=1}^{n} \log(\mu_i)\Big), \tag{3}$$

whose mean is asymptotically equivalent to (2)

- GRIDE: $n_1 = n_2/2$

D = 3 = embedding dimension

d = 2 = intrinsic dimension



	Body temp.	Heart rate	Blood pressure
Person 1			
Person 2			
Person 3			
Person 4			
Person 5			

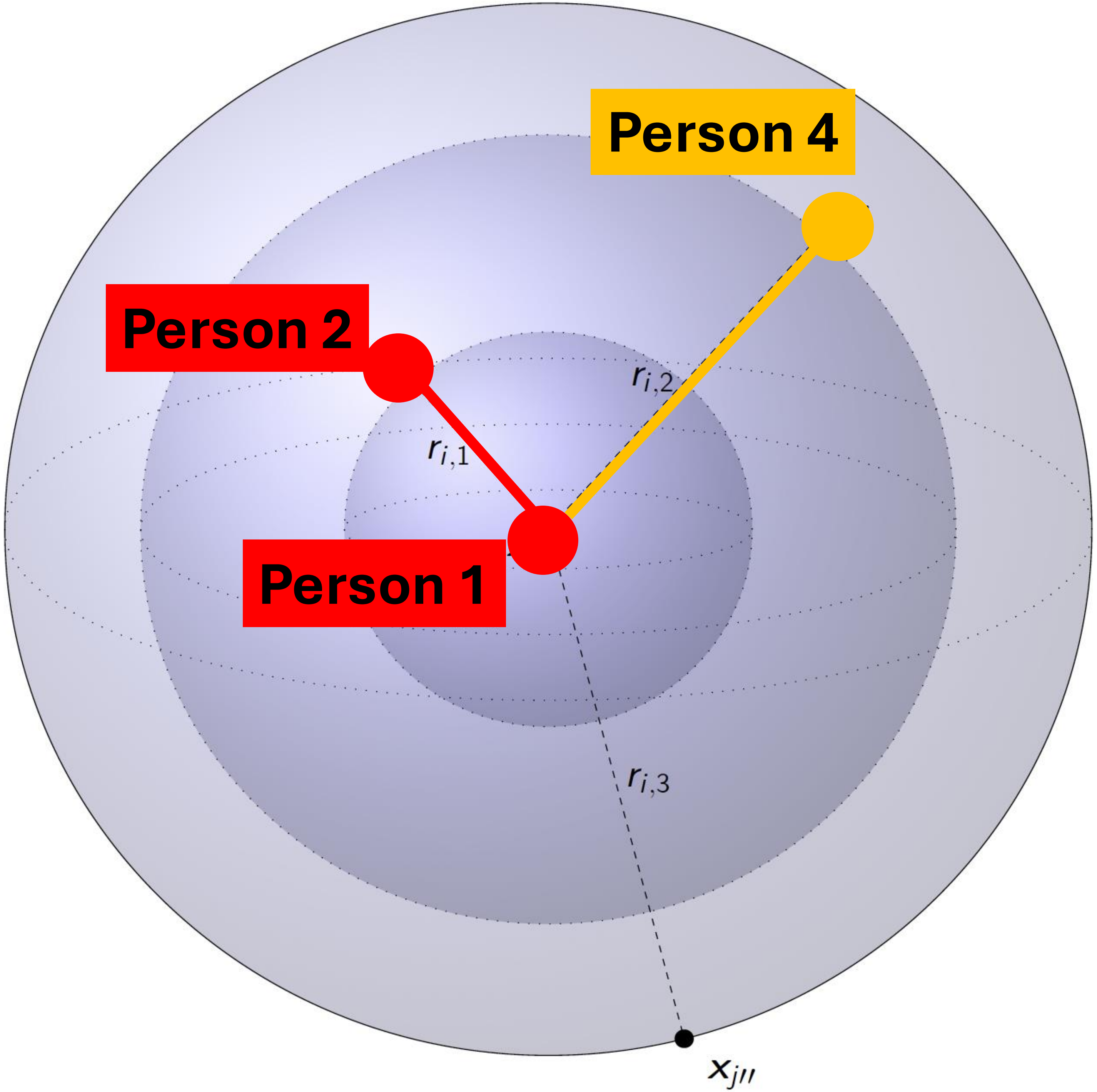
$D = 3$

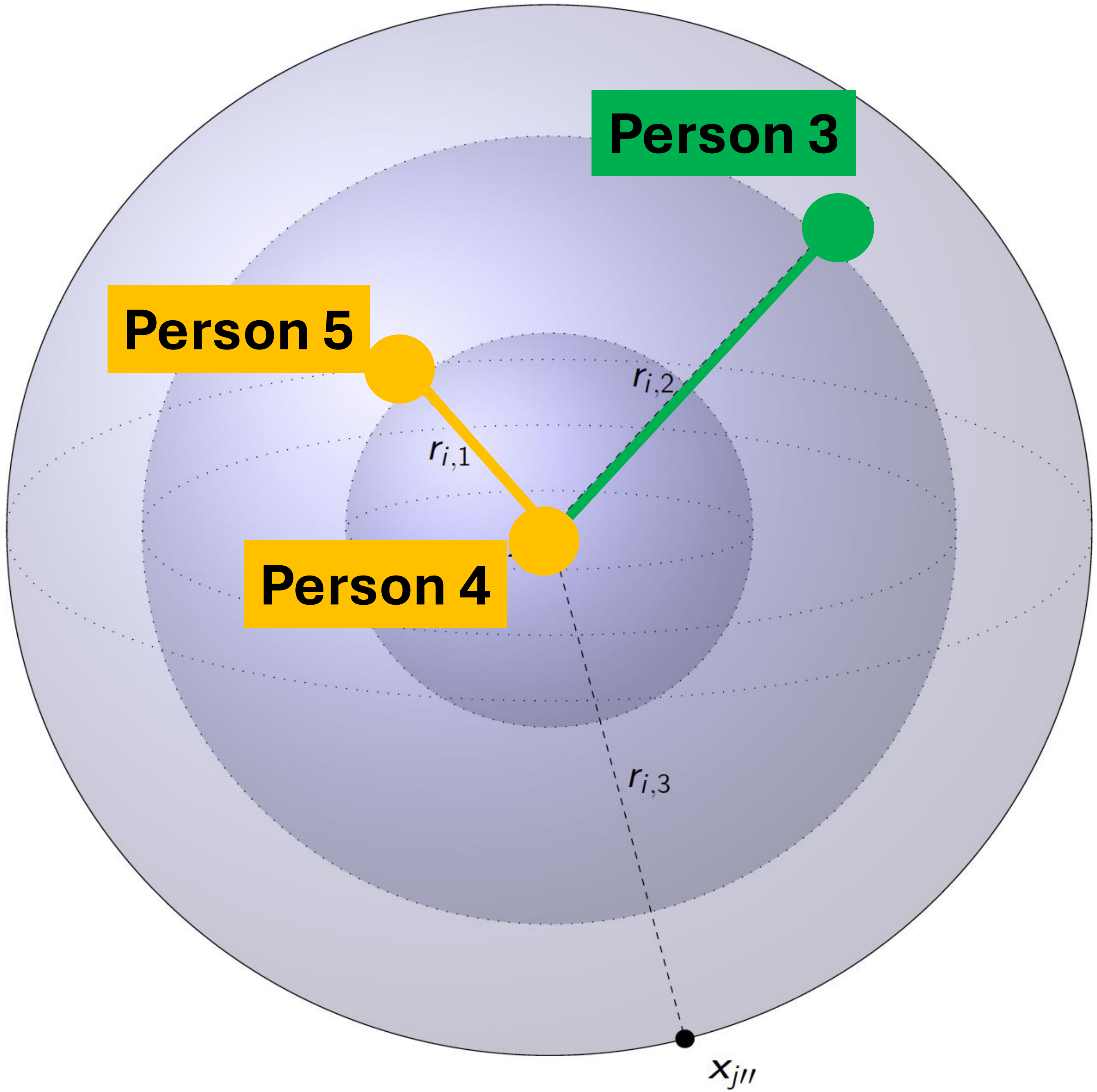
$d1 = 0$

$d2 = 2$

$d3 = 3$

$N = 5$


Person 4
Person 2
$r_{i,2}$
$r_{i,1}$
Person 1
$r_{i,3}$
$x_{j''}$


Person 3
Person 5
Person 4
$r_{i,1}$
$r_{i,2}$
$r_{i,3}$
$x_{j''}$

How to deal with heterogeneous ID case? HIDALGO!

Heterogeneous ID algorithm - Hidalgo model
allows for the possibility that the ID may not be uniform in the dataset.

Under some assumptions the distribution of $\mu_i = r_{i2}/r_{i1}$ is a mixture of Pareto distributions

$$f(\mu_i) = \sum_{k=1}^{K} p_k d_k \mu_i^{-d_k - 1}$$

Likelihood and Estimation

The likelihood of the data is

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k d_k \mu_i^{-d_k-1}$$

where $\boldsymbol{\mu} = (\mu_1 \dots \mu_N)$
Then we can again estimate

$$\mathbf{d} = (d_1 \dots d_K), \quad \mathbf{p} = (p_1 \dots p_K)$$

Likelihood and Estimation

The likelihood of the data is

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k d_k \mu_i^{-d_k-1}$$

where $\boldsymbol{\mu} = (\mu_1 \ldots \mu_N)$
Then we can again estimate

$$\mathbf{d} = (d_1 \ldots d_K), \quad \mathbf{p} = (p_1 \ldots p_K)$$

Fix $P_{prior}(\mathbf{d},\mathbf{p})$ and compute the posterior means
$P_{post}(\mathbf{d},\mathbf{p}) \propto \mathcal{L}(\mu|\mathbf{d},\mathbf{p}) P_{prior}(\mathbf{d},\mathbf{p})$

Priors

Independent priors on **d** and **p**

Prior on **d** : $d_k \sim Gamma(a_0, b_0), \quad k = 1, \ldots, K$

Prior on **p** $\sim Dir(\alpha_1, \ldots, \alpha_K)$

Prior on $\mathbf{Z}|\mathbf{p}$ $\sim$discrete distribution on $(1, \ldots K)$ w.p. **p**

BNP - HIDALGO

- To adopt a full Bayesian approach, we need to address the uncertainty on the number of mixture components K
- Instead of making K stochastic, we adopt a Bayesian nonparametric approach, letting $K \to \infty$

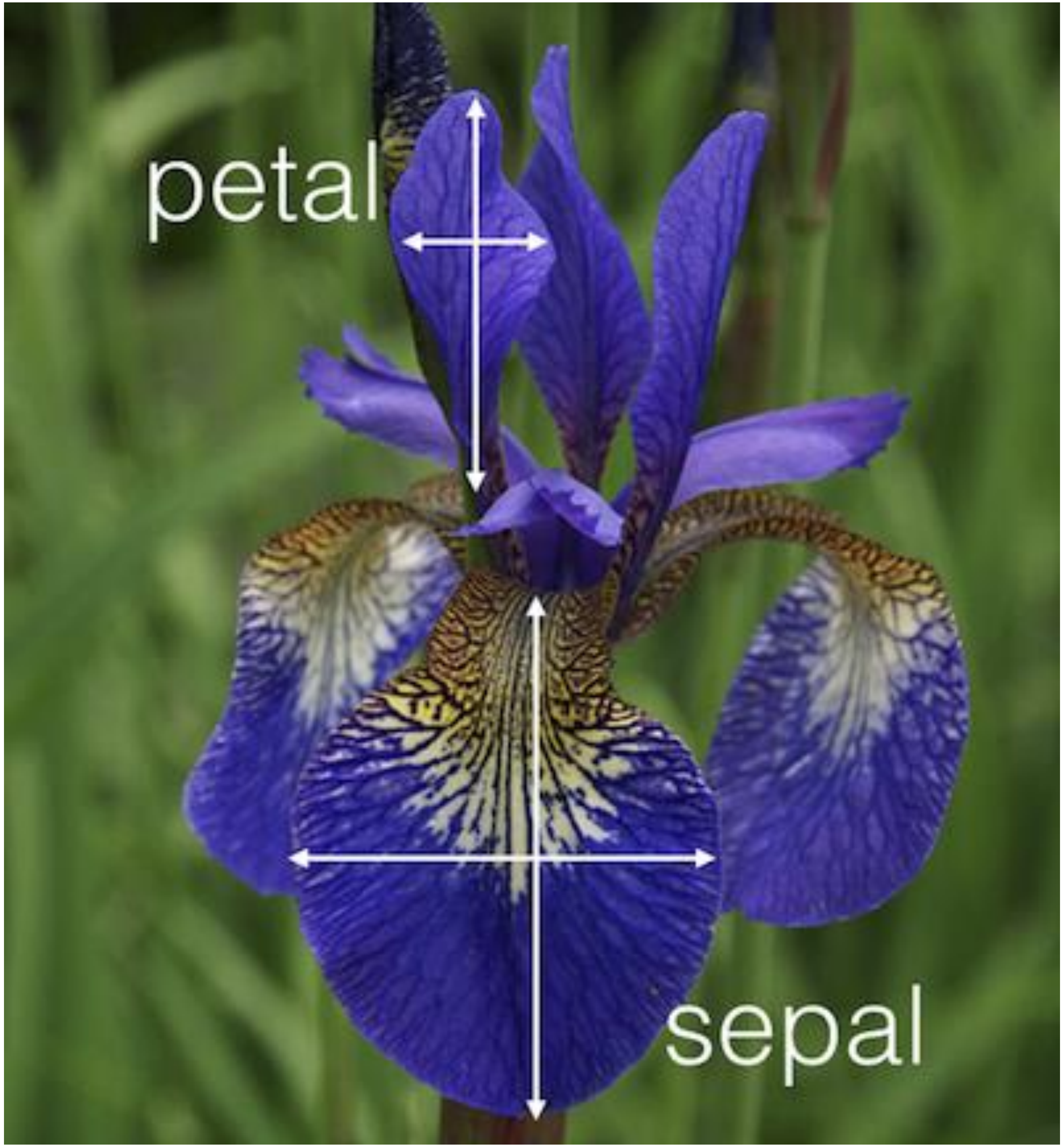
We now model μ_i as a infinite mixture of Pareto distributions:

$$\sum_{i=1}^{+\infty} p_i \cdot \mathcal{P}(\mu_i | d_i)$$

We adopt a Dirichlet process prior for the parameters that model the ID

Three types of Iris flowers

N = 50



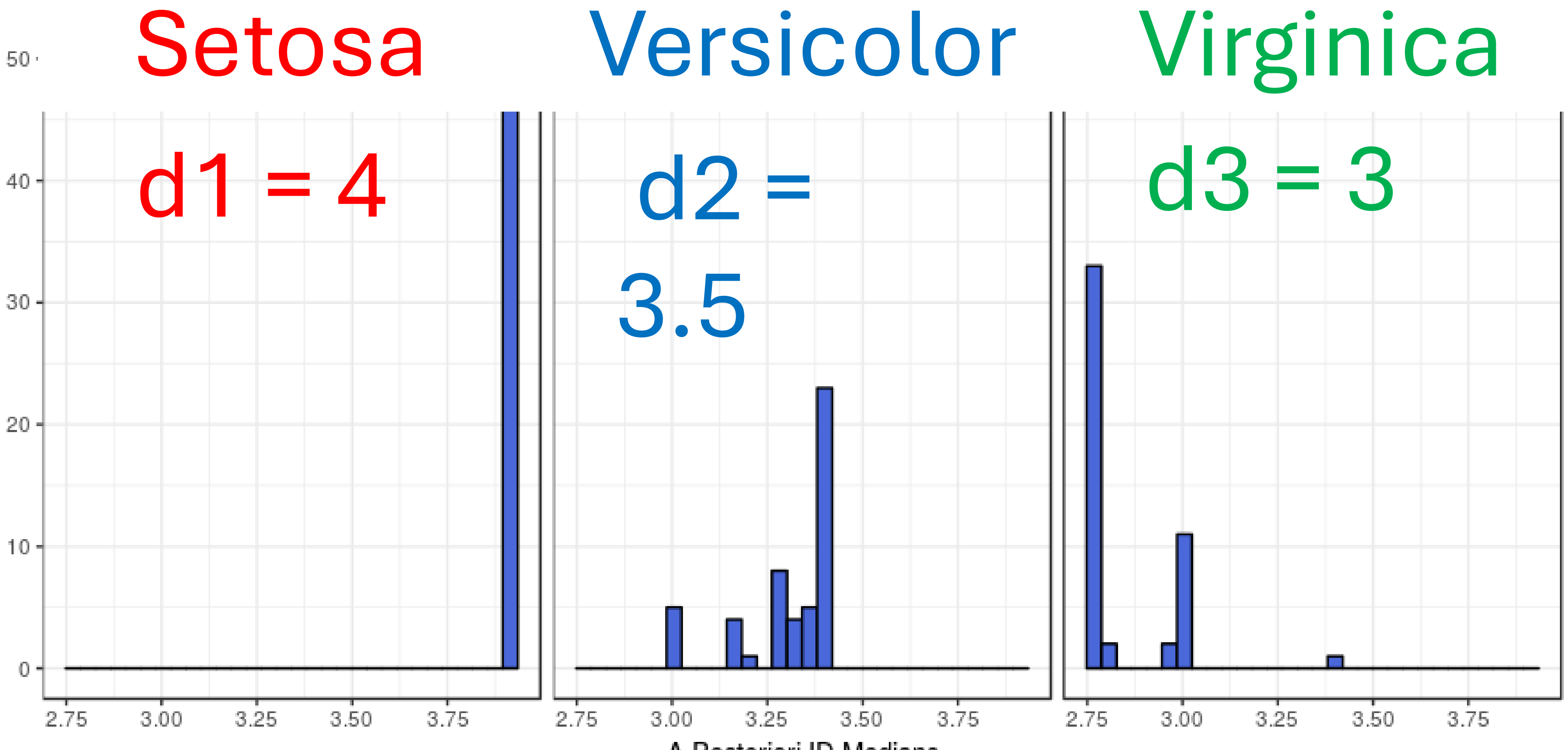
D = 4

	Petal length	Petal width	Sepal length	Sepal width
Flower 1				
Flower 2				
Flower 3				
Flower 4 . . . Flower 50				

D = 4

N = 50

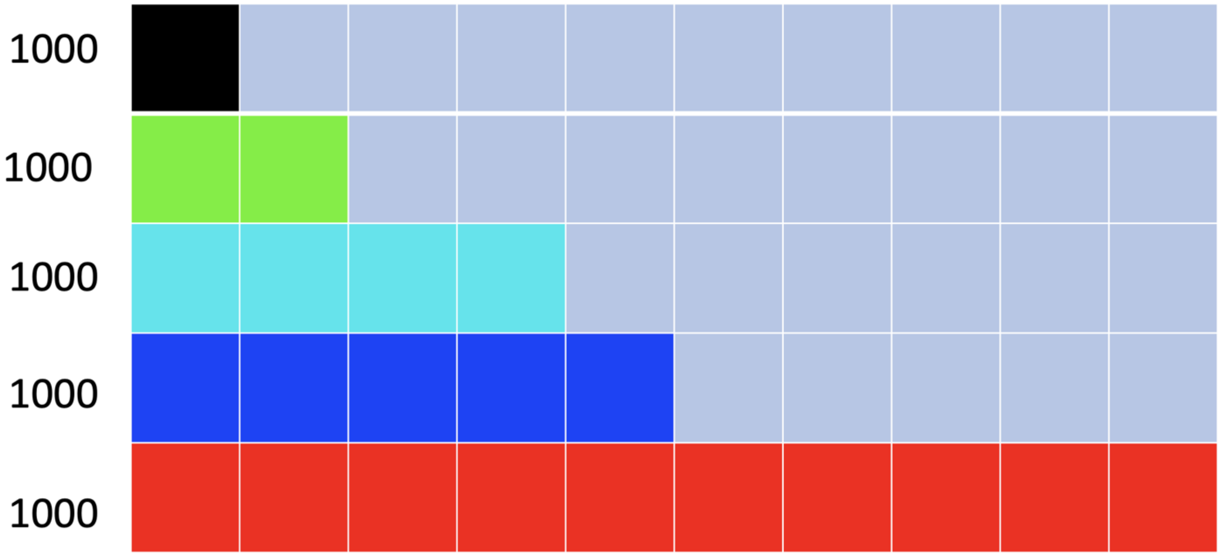
3 clusters
almost coincident
with the flower species

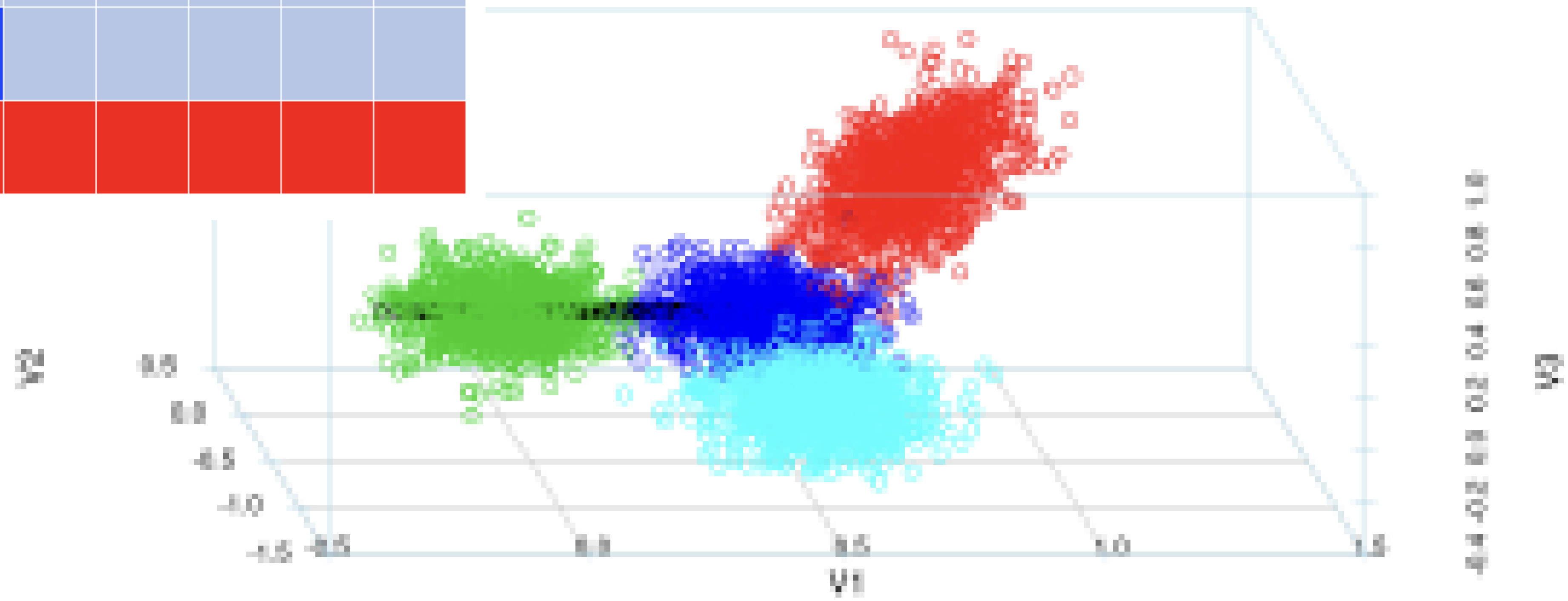


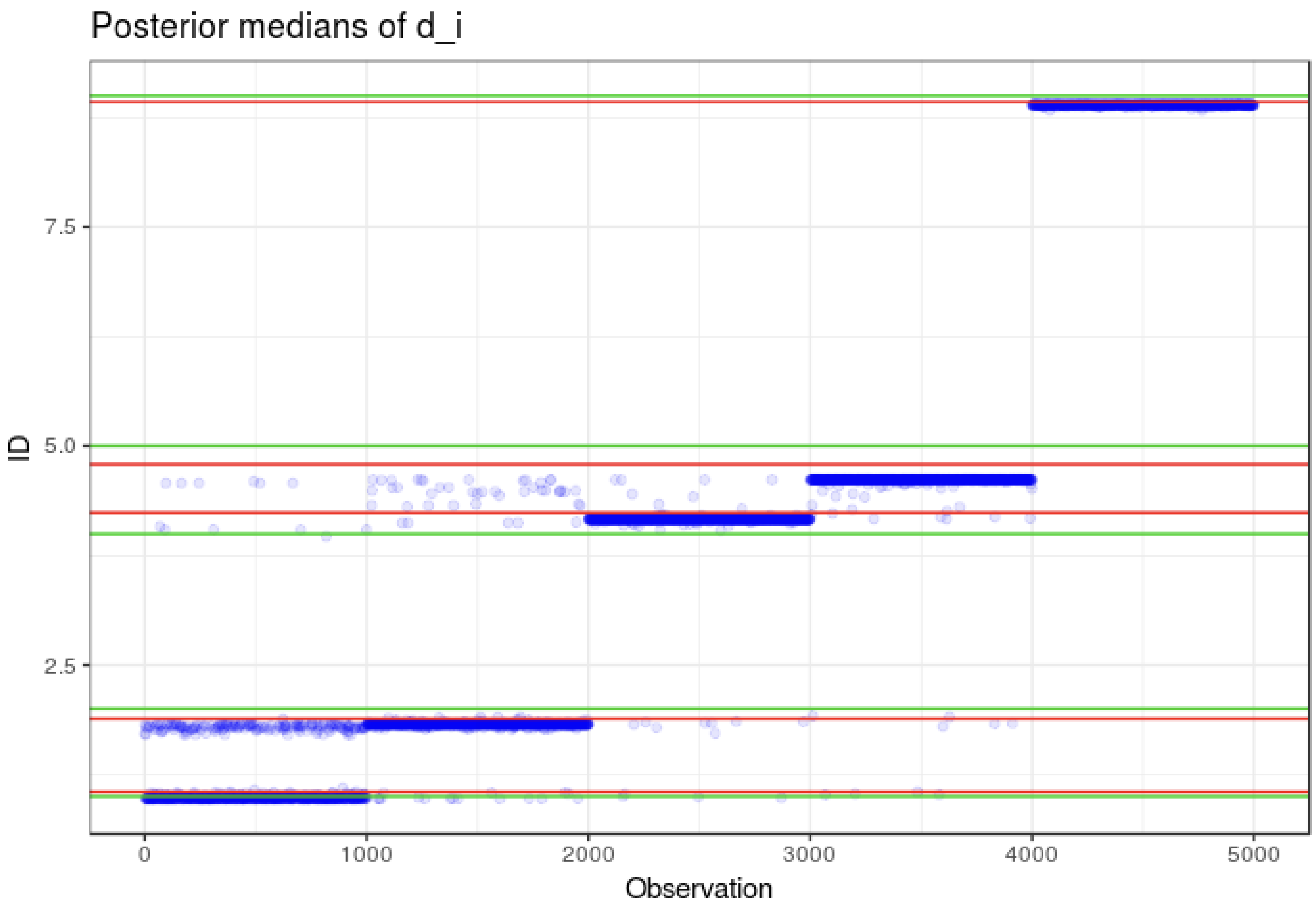
Simulated data

5 x1000 observations generated from 5 Gaussians of dimensions 1, 2, 4, 5 and 9, partially overlapping:
3 dim projection

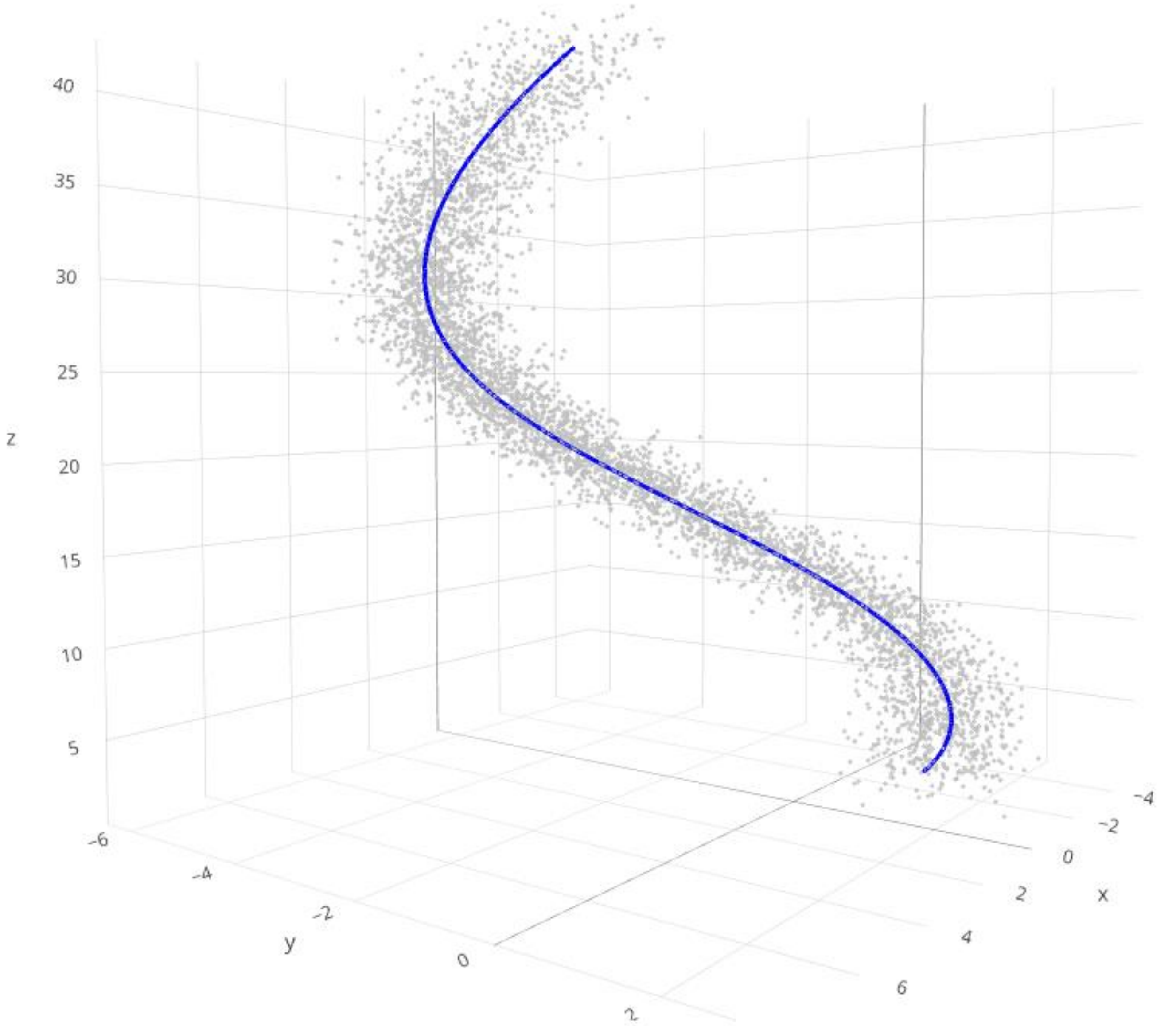
Data matrix = 5000 x 9

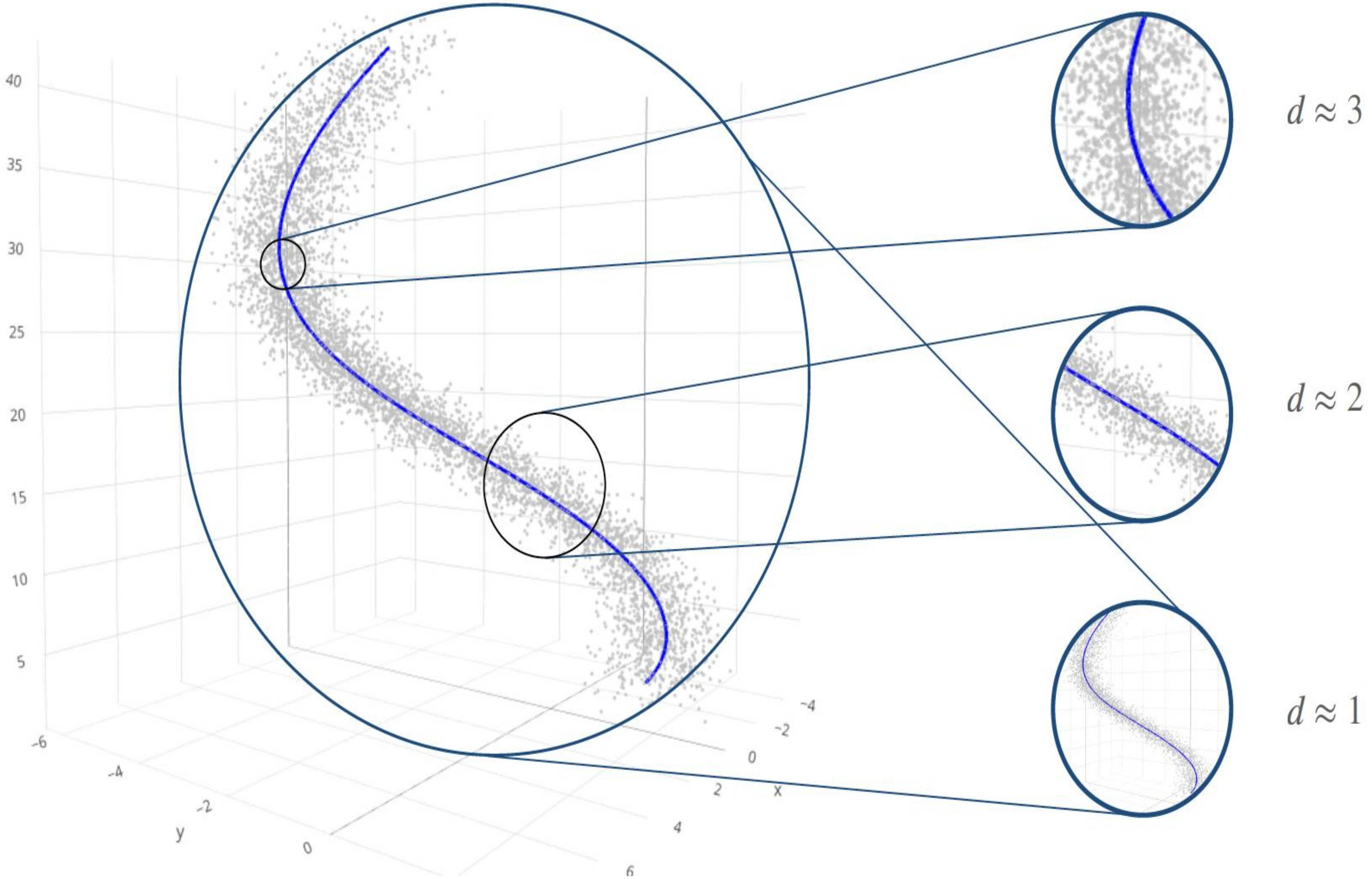





Posterior medians of d_i
ID
7.5
5.0
2.5
0
1000
2000
3000
4000
5000
Observation

NOISE

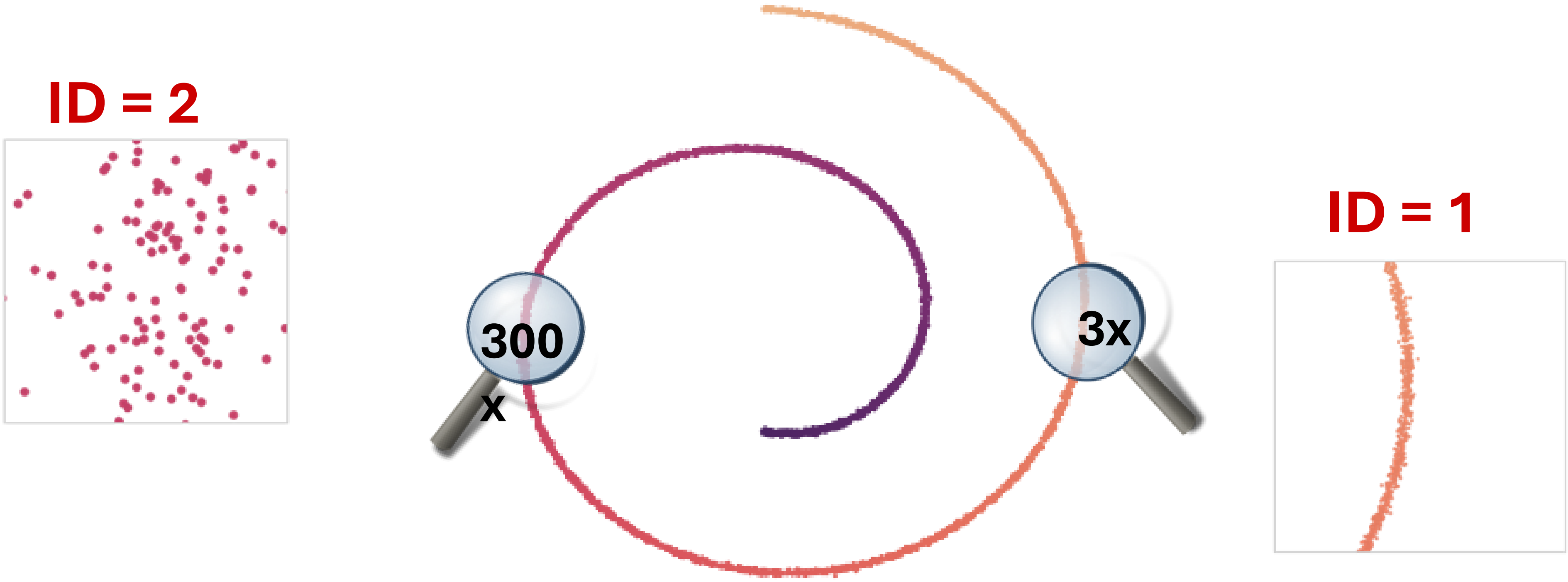




d ≈ 3
d ≈ 2
d ≈ 1
40
35
30
25
20
15
10
5
-6
-4
-2
0
y
-4
-2
0
2
4
6
x

Intrinsic dimension estimation on noisy datasets

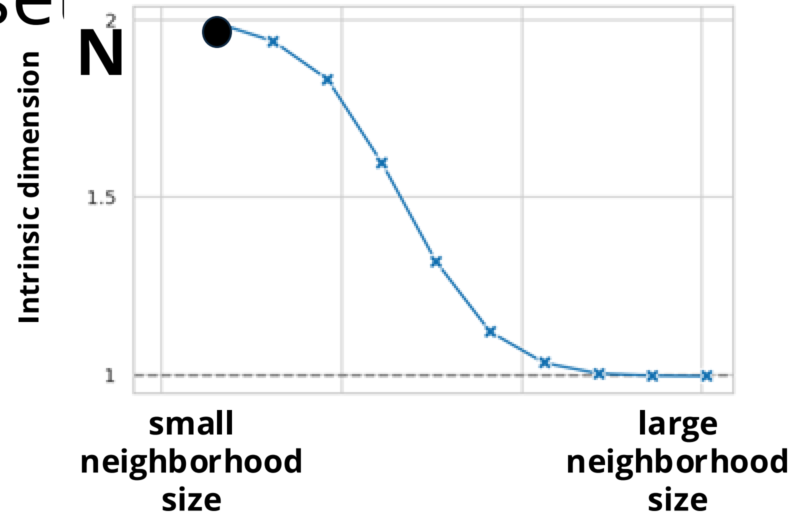
- The estimated intrinsic dimension is a **scale dependent** quantity



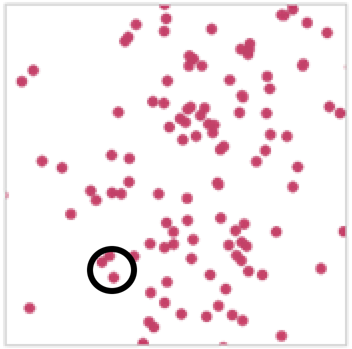
Intrinsic dimension estimation on noisy datasets

- The intrinsic dimension is a **scale dependent** quantity

TwoN N

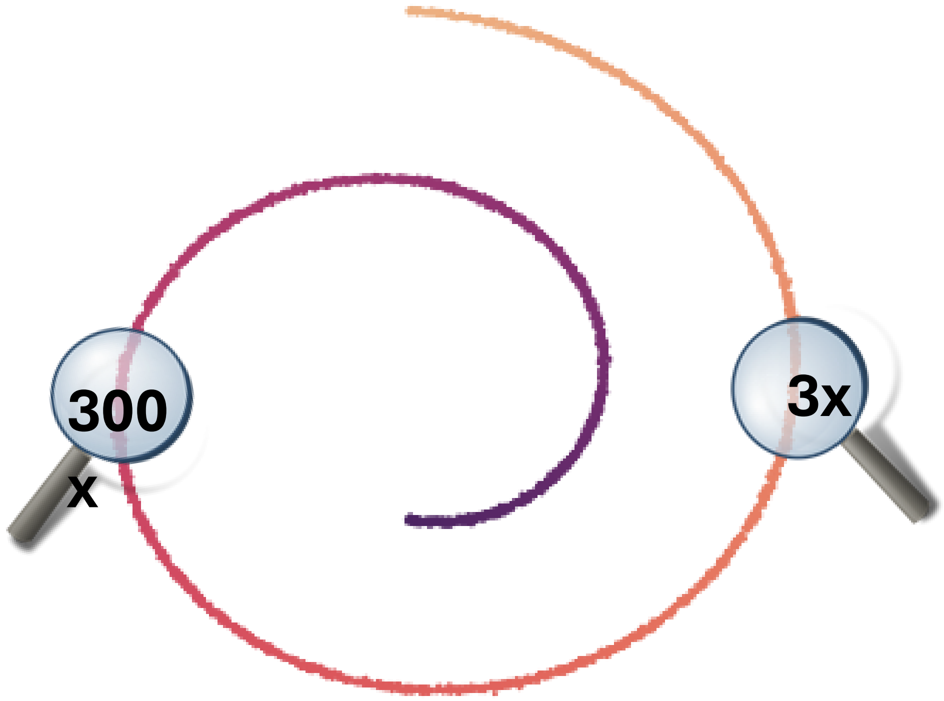


ID = 2

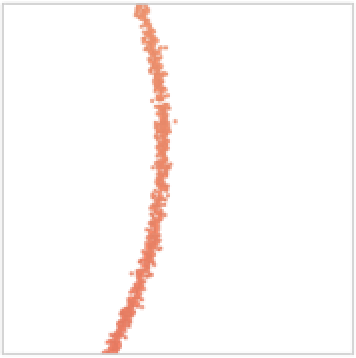


k = 2

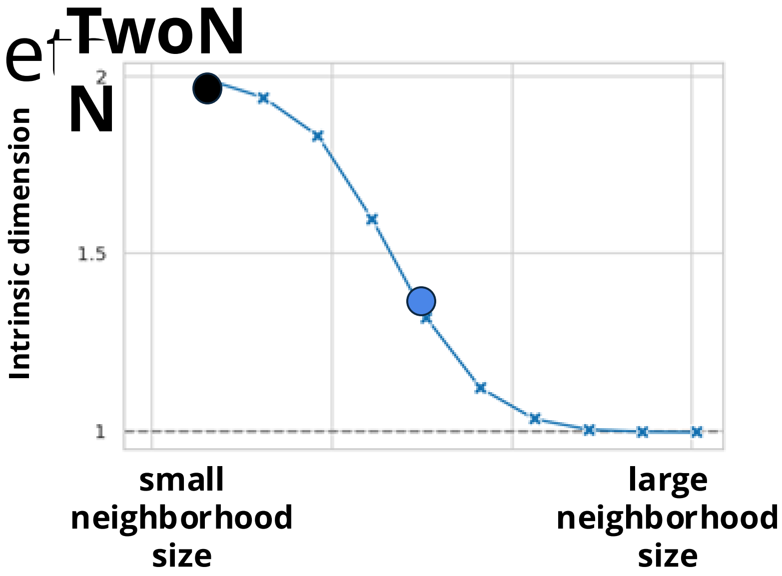
TwoNN



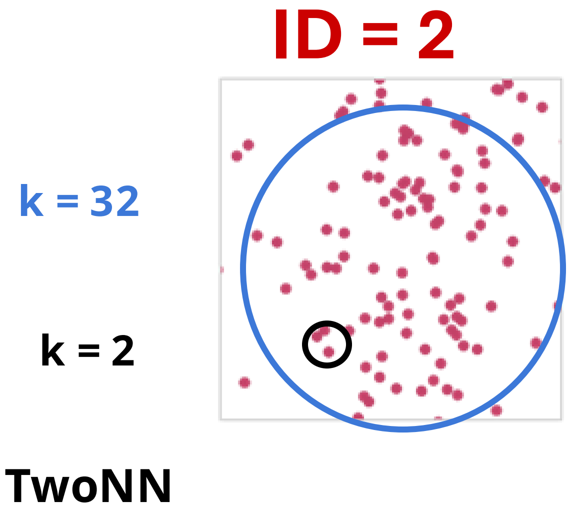
ID = 1

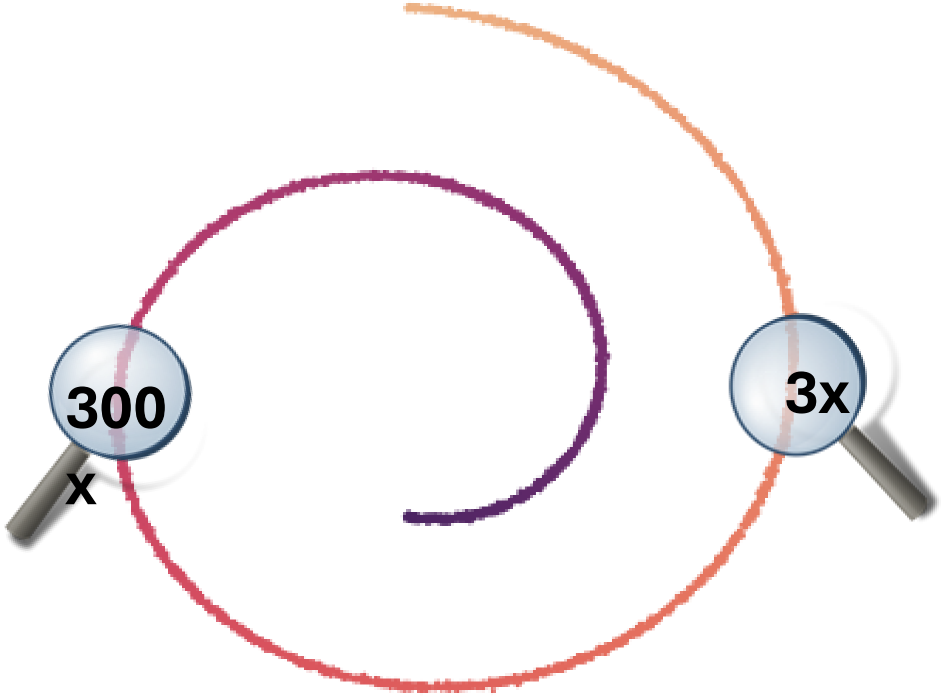


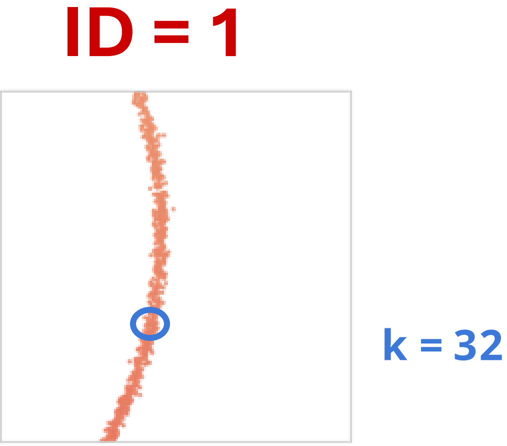
Intrinsic dimension estimation on noisy datasets



- The intrinsic dimension is a **scale dependent** quantity

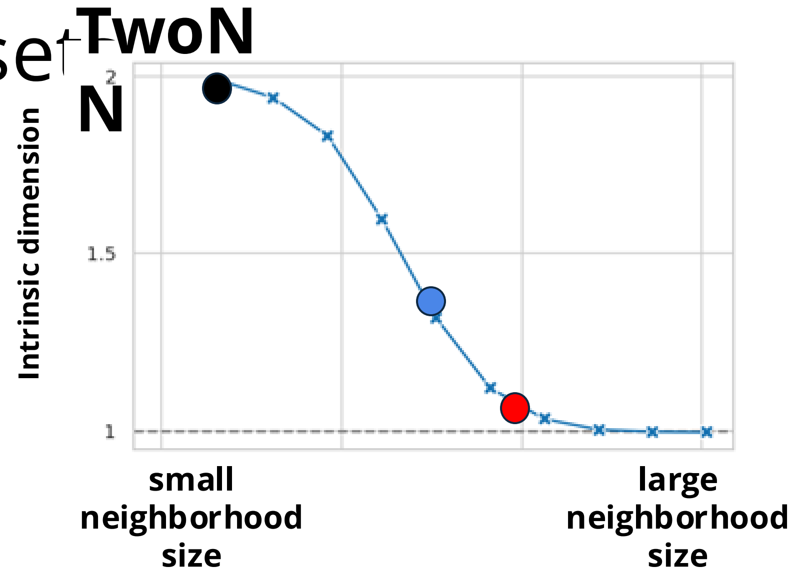




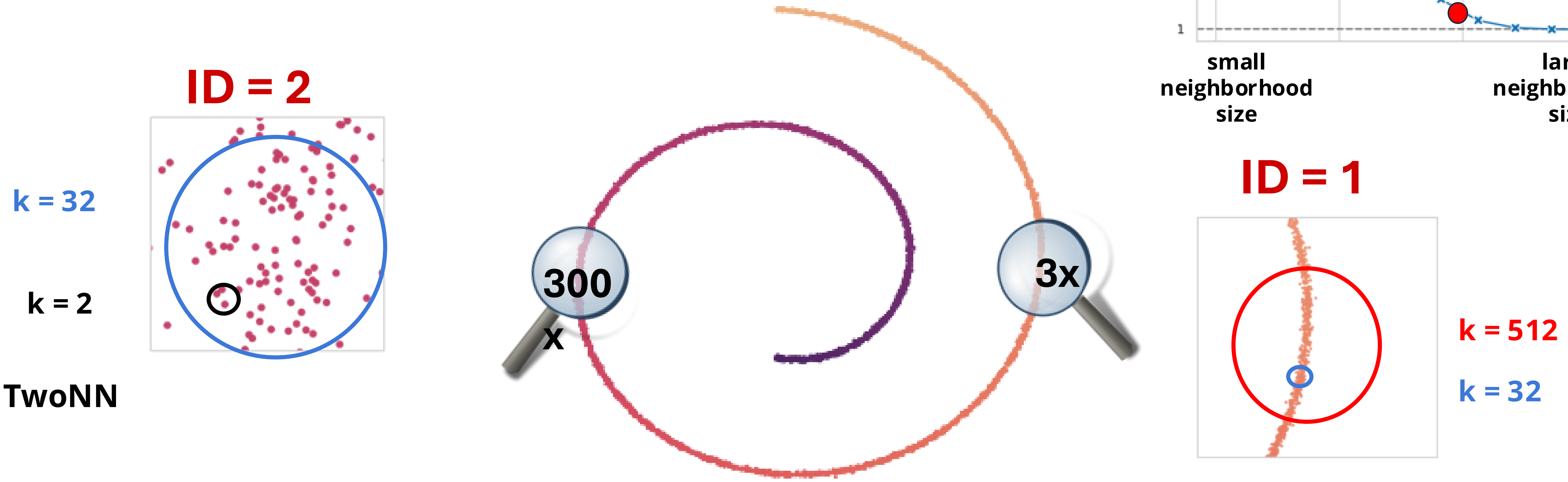


- The scale dependence of the intrinsic dimension can be probed with an **higher order ratio approach**

Intrinsic dimension estimation on noisy datasets

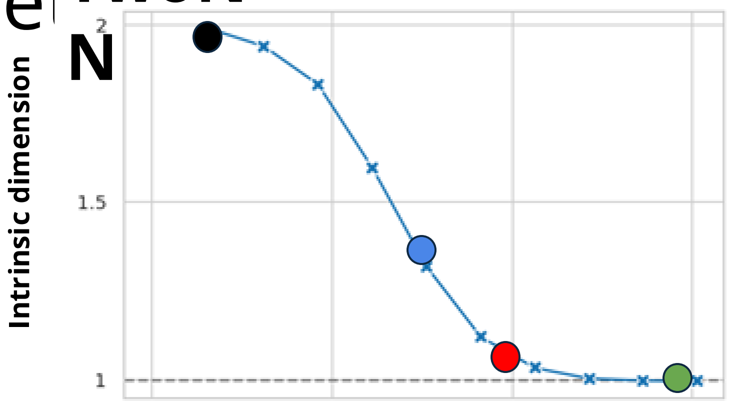


- The intrinsic dimension is a **scale dependent** quantity

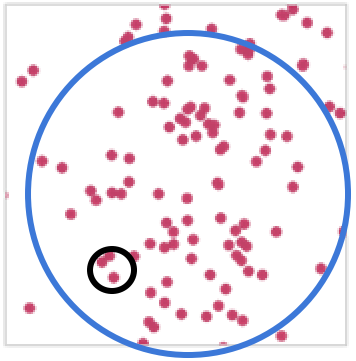


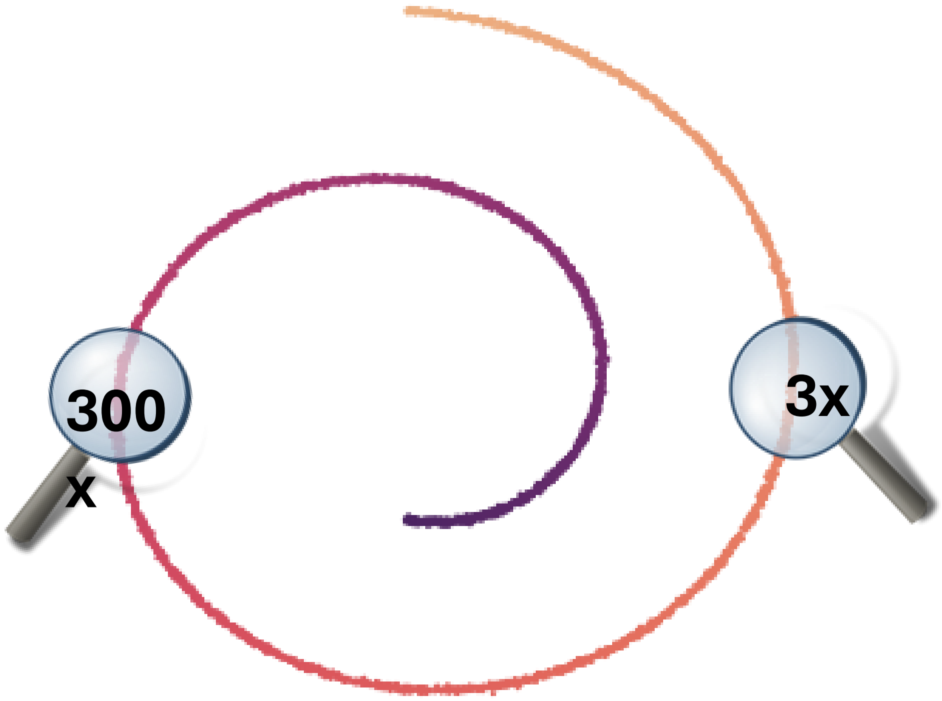
- The scale dependence of the intrinsic dimension can be probe with an **higher order ratio approach**

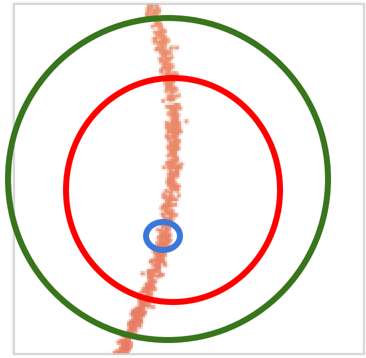
Intrinsic dimension estimation on noisy datasets



- The estimated intrinsic dimension is a **scale dependent** quantity

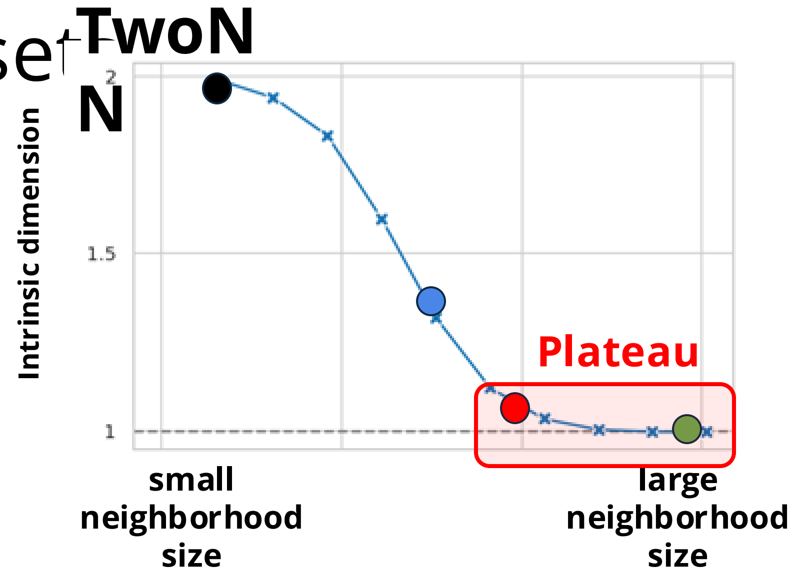




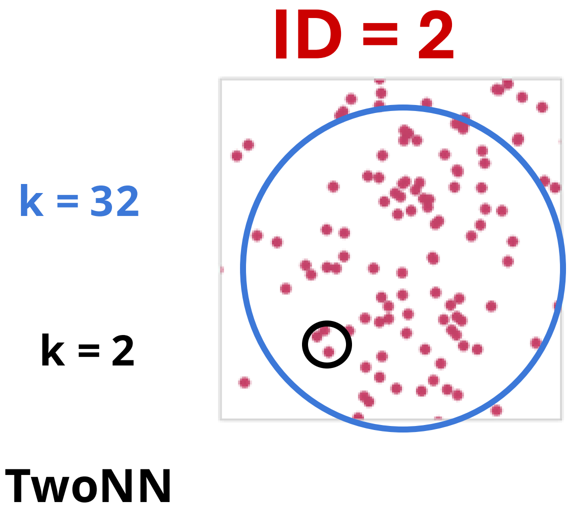


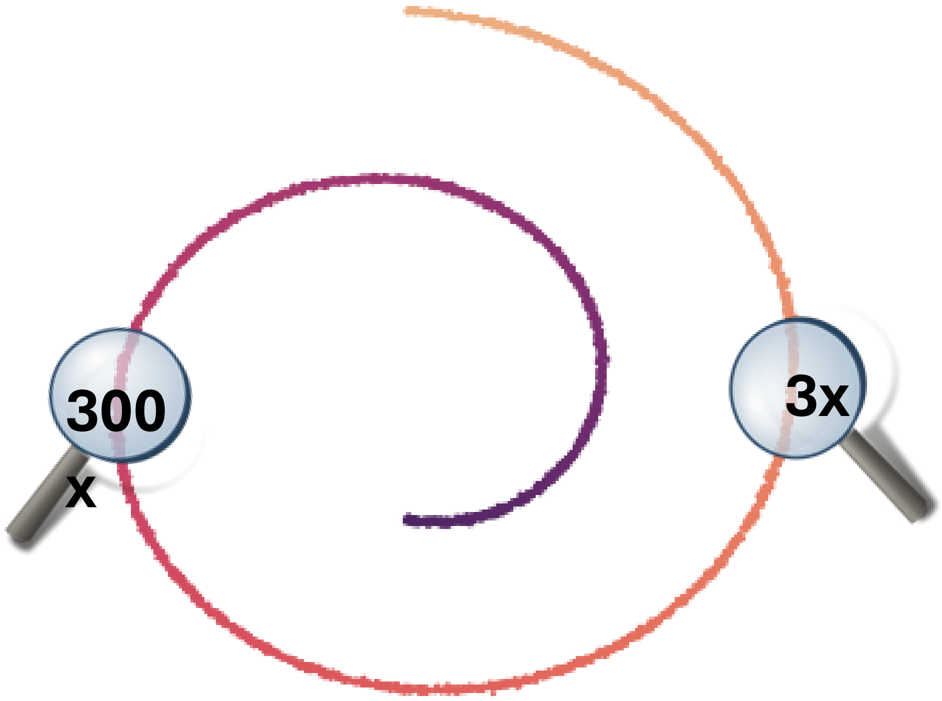
- The scale dependence of the intrinsic dimension can be probe with an **higher order ratio approach**

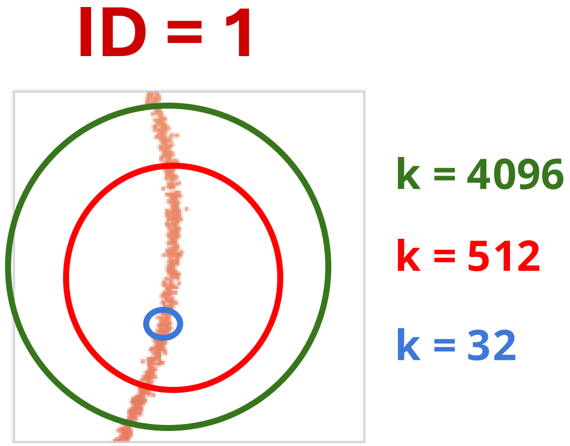
Intrinsic dimension estimation on noisy datasets



- The intrinsic dimension is a **scale dependent** quantity







- The scale dependence of the intrinsic dimension can be probe with an **higher order ratio approach**

Generalized ratios ID estimator (GRIDE)

Homogeneous Poisson process assumption:

1. The **number of points** $k(A_1)$, $k(A_2)$ falling in two non overlapping regions A1, A2 are **independent random variables**

1. The **number of points** k in a region of volume V is a **Poisson random variable**:

$$P(k, V) = \frac{(\rho V)^k}{k!} e^{-\rho V}$$

Generalized ratios ID estimator (GRIDE)

k^{th}, $2k^{th}$ neighbors

$$\mu_{k_i} = \frac{r_{2k_i}}{r_{k_i}} \Rightarrow p(\mu_{k_i}|d) = \frac{d(\mu_{k_i}^d - 1)^{k-1}}{\mu_{k_i}^{(2k-1)d+1}\mathrm{Beta}(k,k)}$$

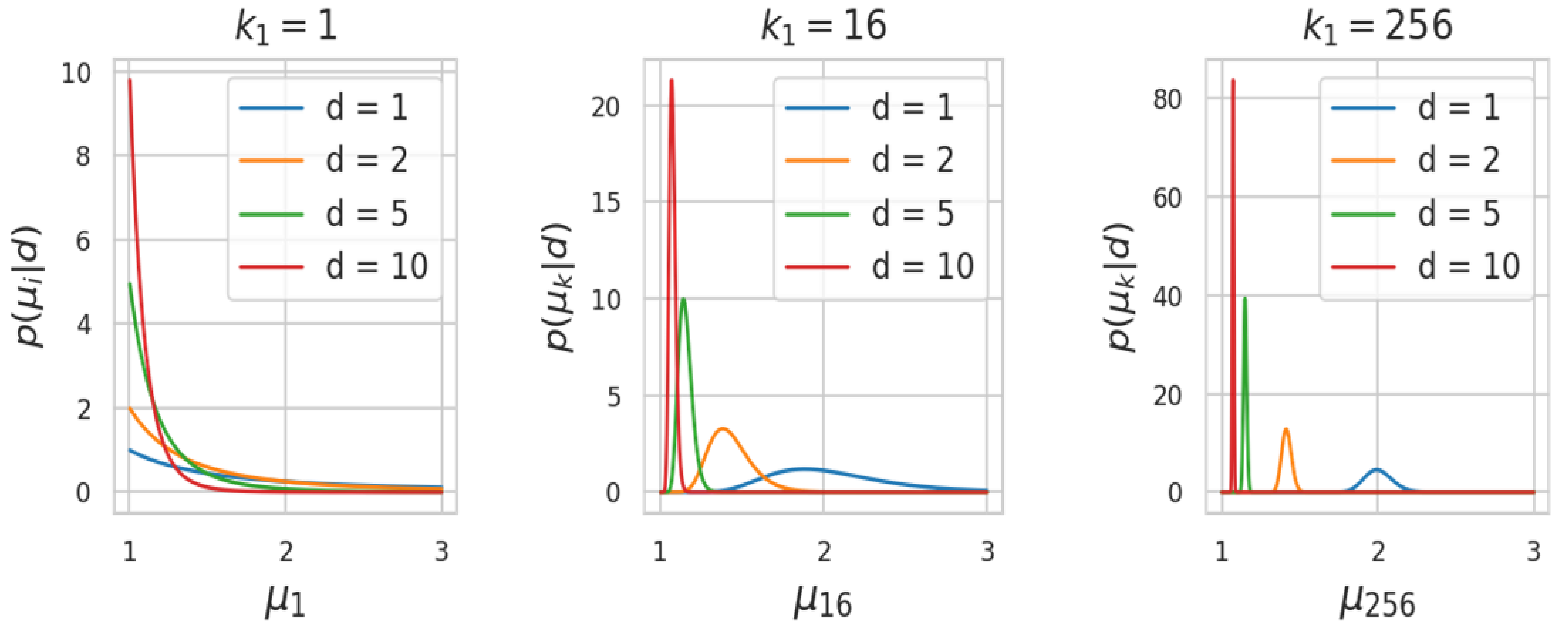
Data point likelihood

$$p(\boldsymbol{\mu}|d) = \prod_{i}^{N} p(\mu_i|d)$$

Data set likelihood

Generalized ratios ID estimator (GRIDE)

TwoNN **Likelihood variance decreases increasing k**

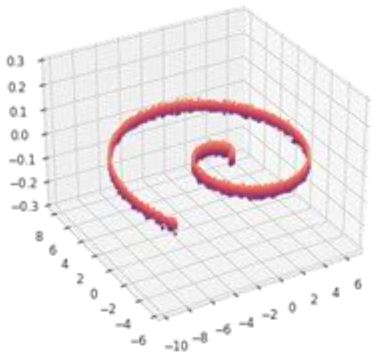


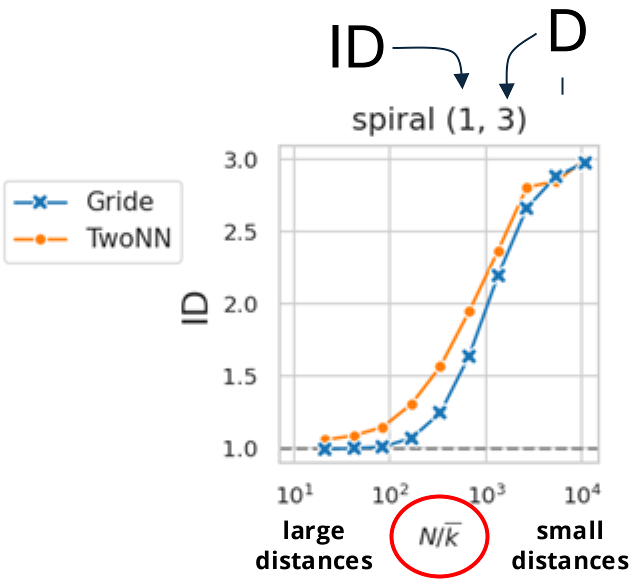
Scale analysis of the ID on synthetic datasets

n° data points = N = 16000

$$\sigma = \frac{0.01}{\sqrt{D}}$$

spiral 1d



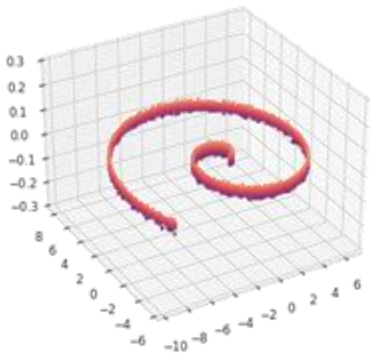


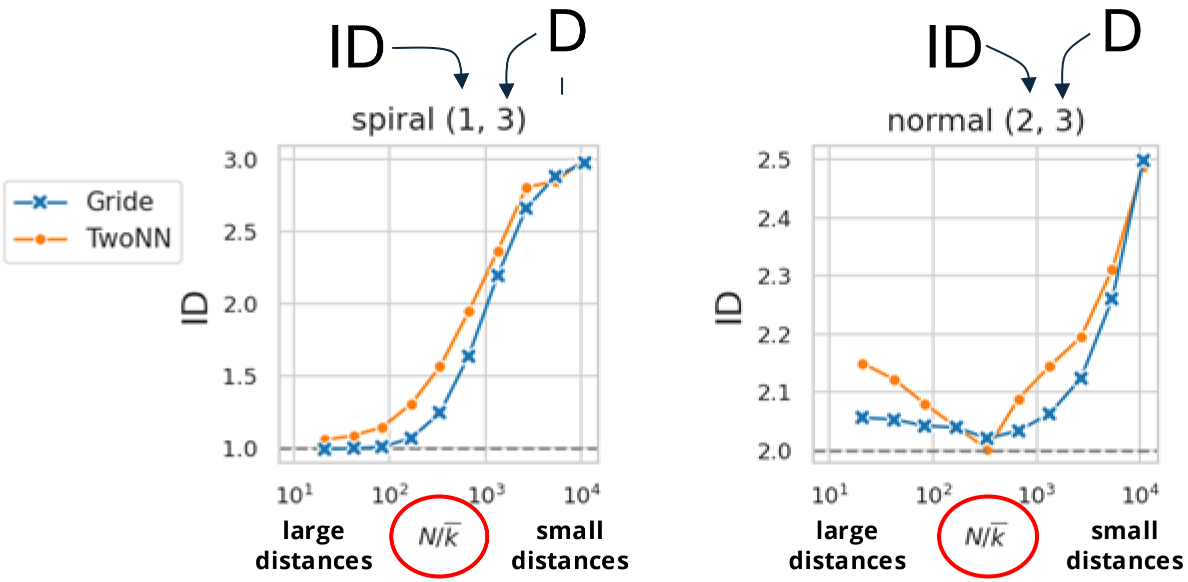
Scale analysis of the ID on synthetic datasets

n° data points = N = 16000

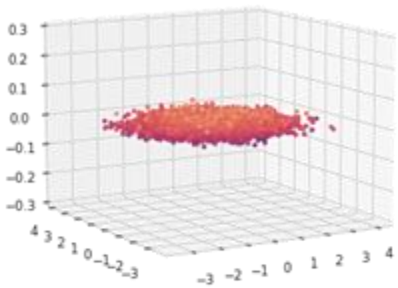
$$\sigma = \frac{0.01}{\sqrt{D}}$$

spiral 1d





gaussian 2d

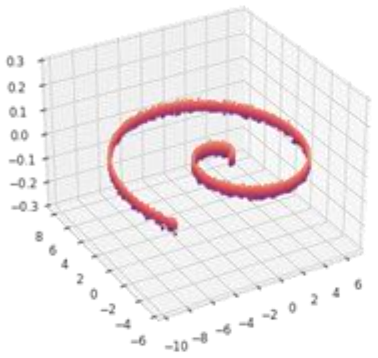


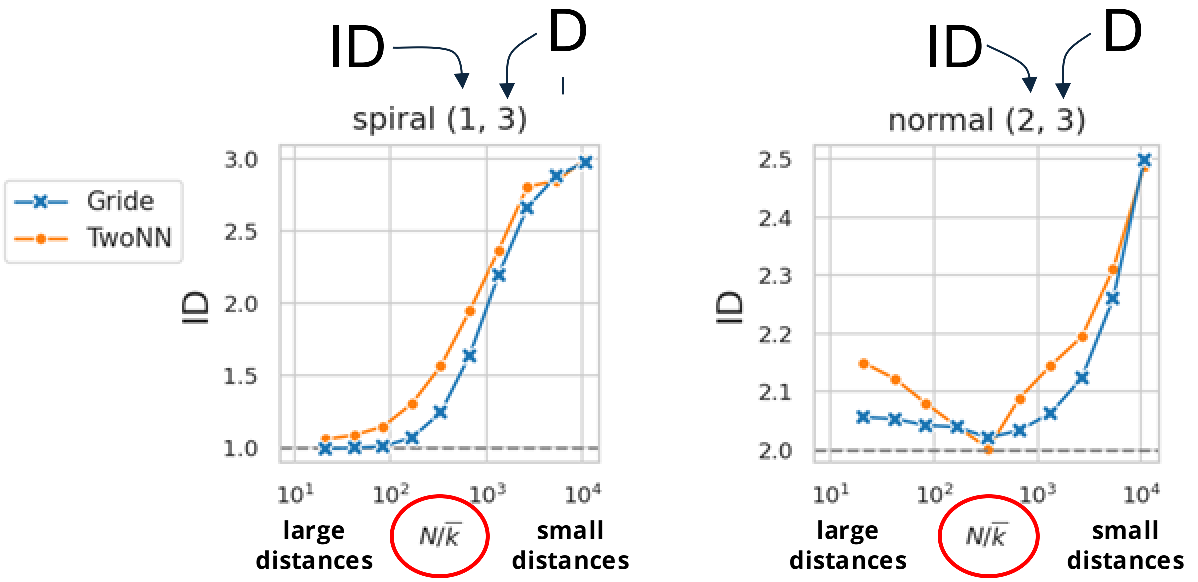
Scale analysis of the ID on synthetic datasets

n° data points = 16000

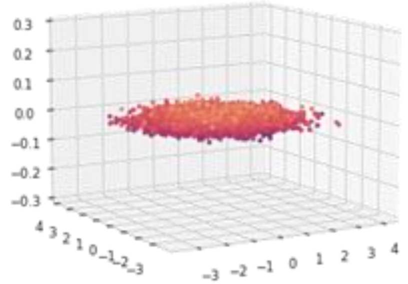
$$\sigma = \frac{0.01}{\sqrt{D}}$$

spiral 1d

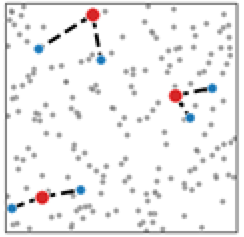




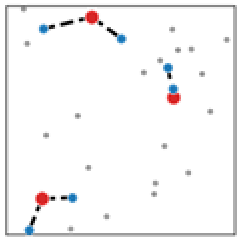
gaussian 2d



Gride



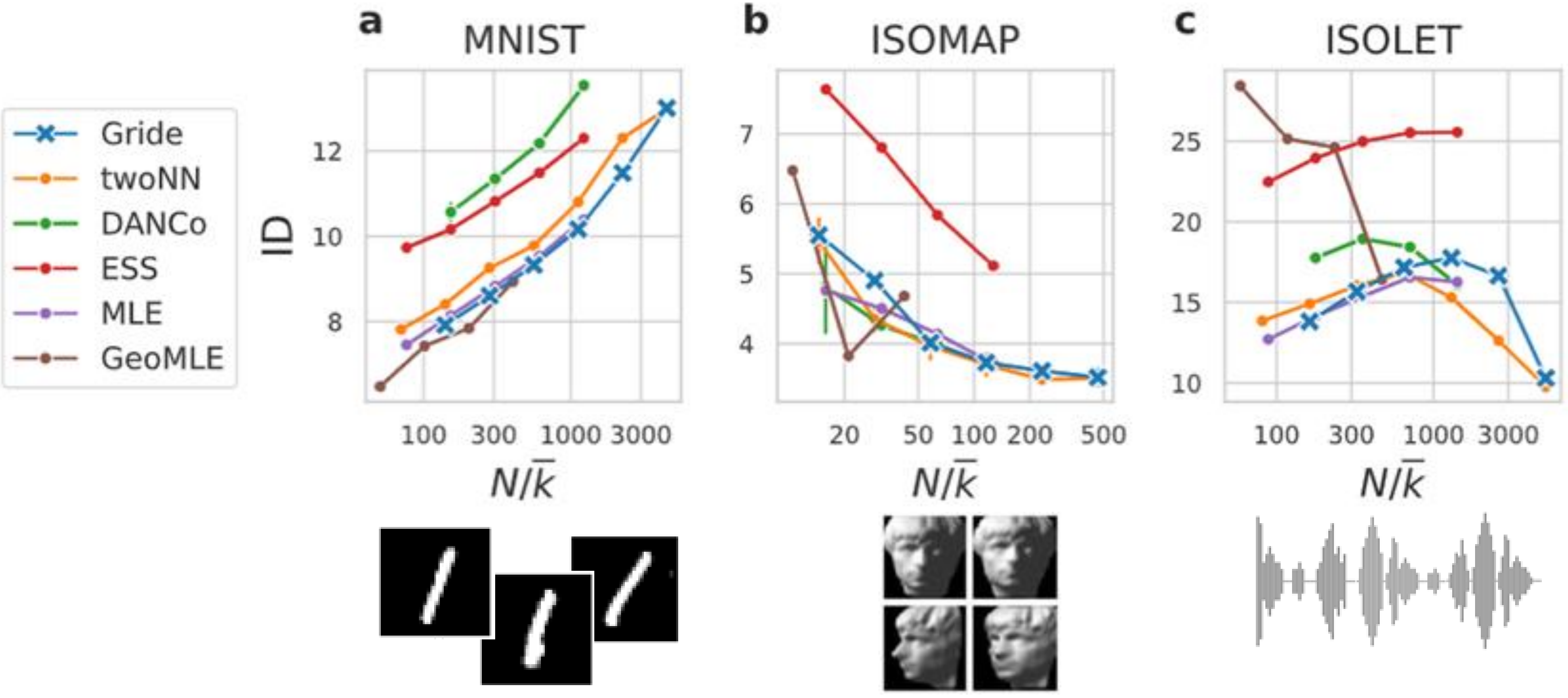
Decimation



We define the scale with N/k

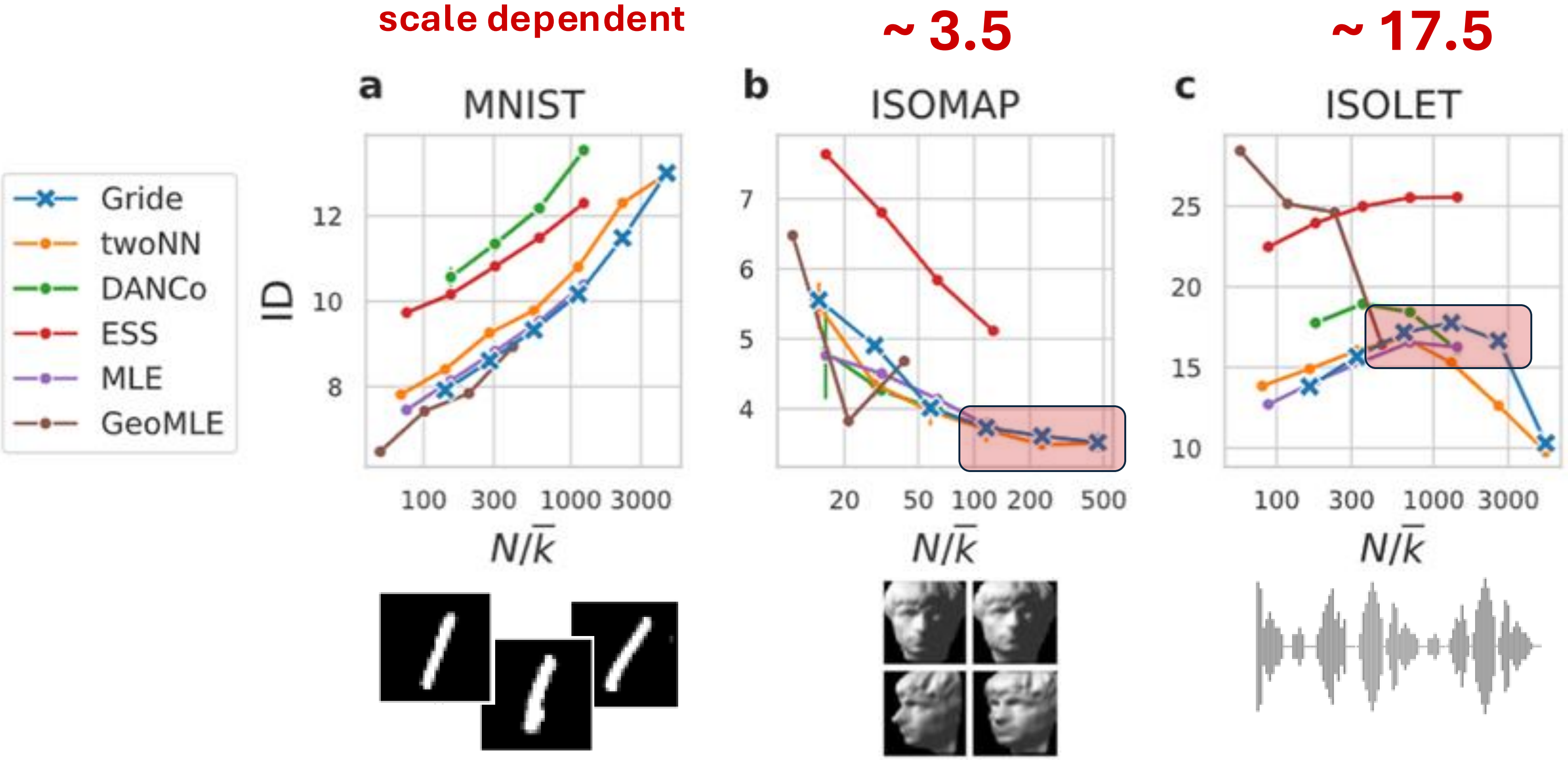
➔ It works better when the data are high dimensional

Scale analysis of the ID on real data sets: unknown intrinsic dimension

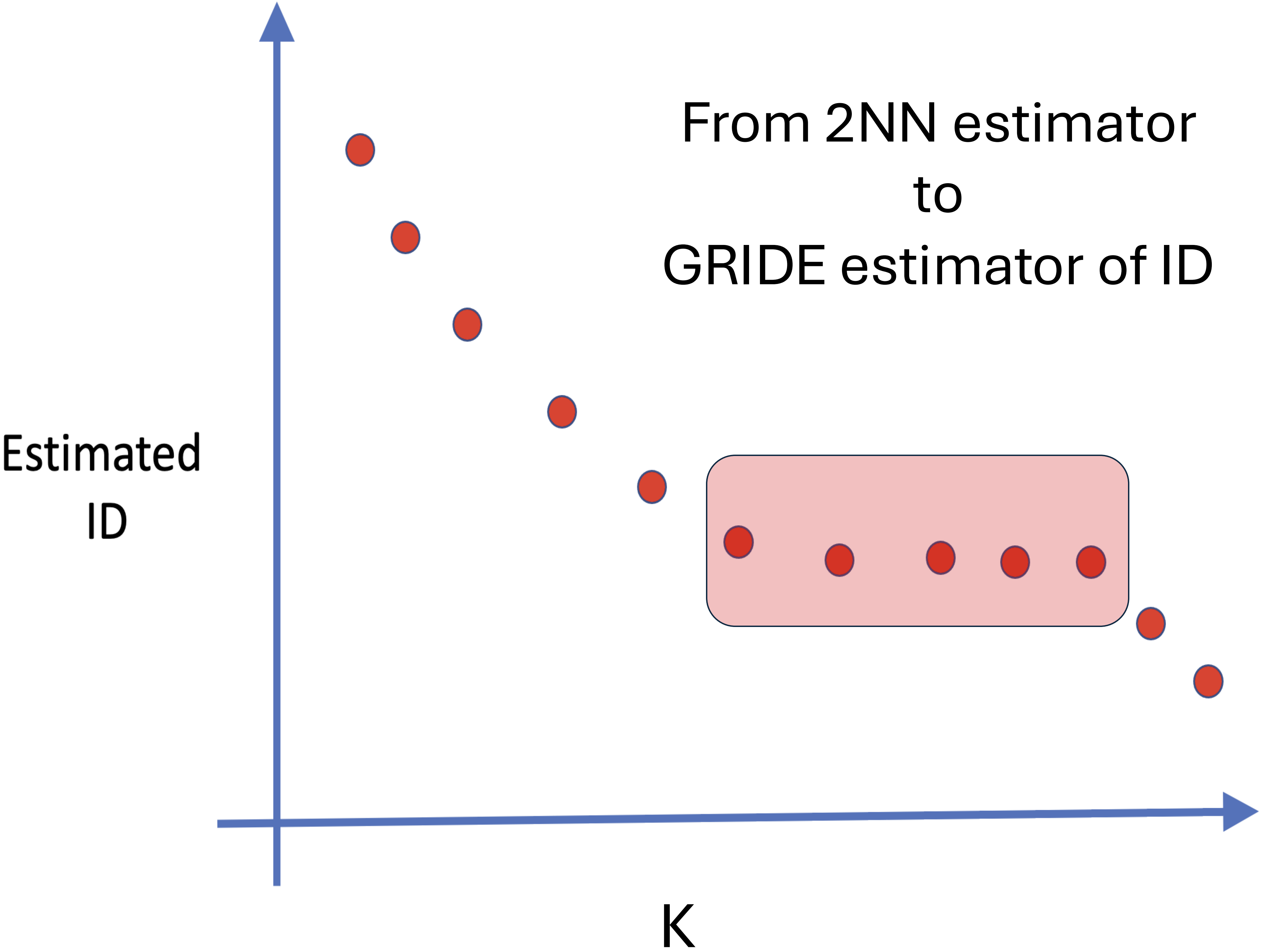


	MNIST	ISOMAP	ISOLET
number of data (N) – number of features (P)	N_{tot} = 6742 P = 784	N_{tot} = 698 P = 4096	N_{tot} = 7797 P = 617
Consensus around plausible ID ranges:	**9-14**	**3**	**17-22**

Scale analysis of the ID on real data sets: unknown intrinsic dimension



number of data (N) – number of features D	**N_{tot} = 6742, D = 784**	**N_{tot} = 698, D = 4096**	**N_{tot} = 7797, D = 617**
Consensus around plausible ID ranges:	**9-14**	**3**	**17-22**


From 2NN estimator
to
GRIDE estimator of ID
Estimated
ID
K

Computational efficiency of GRIDE

Dataset CIFAR10: 32 x 32 color images

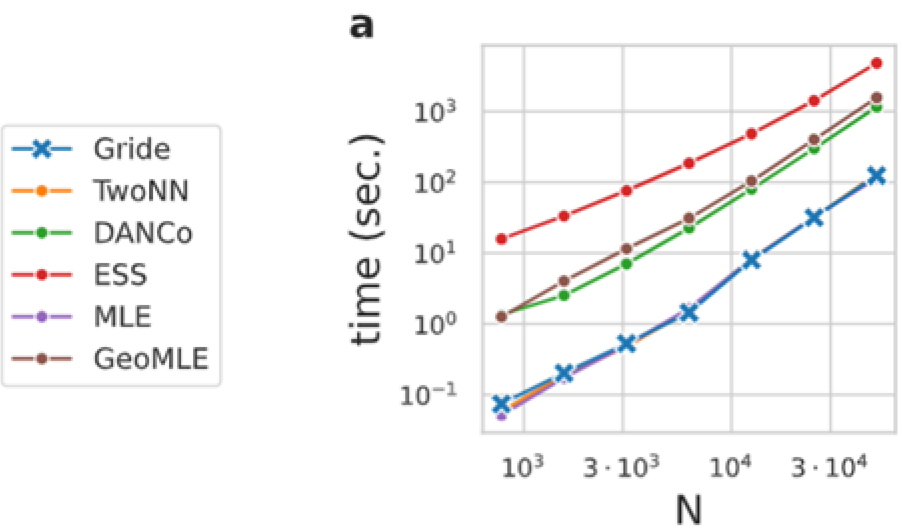
ID ~ 30

Computational efficiency of GRIDE

Dataset CIFAR10: 32 x 32 color images

ID ~ 30

D = 3072 = 32x32x3

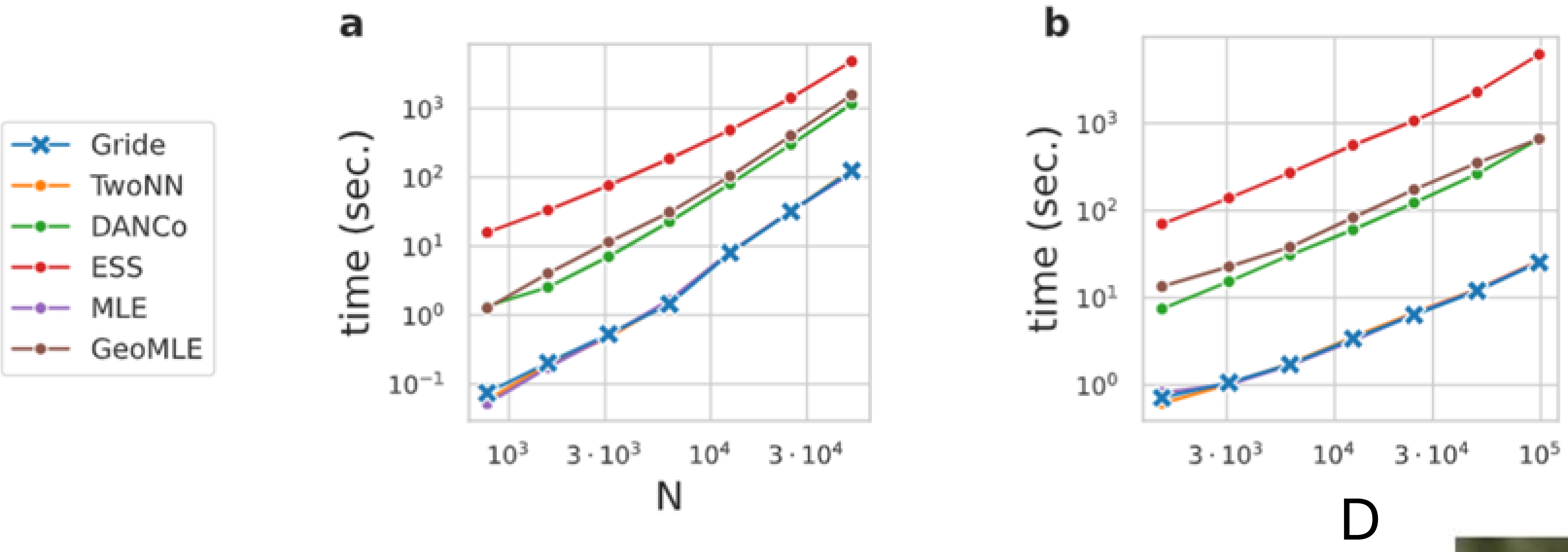


Computational efficiency of GRIDE

Dataset CIFAR10: 32 x 32 color images

ID ~ 30

D = 3072 = 32x32x3 **N = 5000**



A simple topological feature uncovers a rich data structure

Folded vs unfolded state in protein configurations

Active vs non-active regions in brain imaging data

Patients vs controls in gene expression data

Firms with different financial risk in balance sheets data

Other applications:

Winning vs losing teams in basketball data

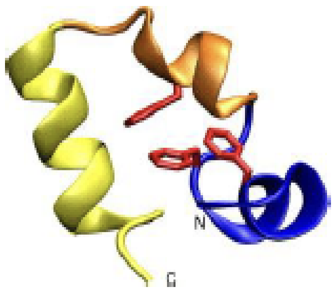
Country specific NPI in Covid-19 pandemic evolution

Identified vs unidentified models in MCMC simulation

Layers in a Deep Neural Network

Communities in Network data

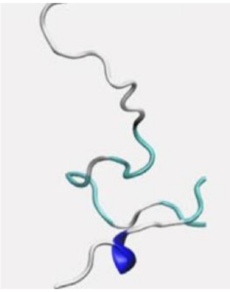
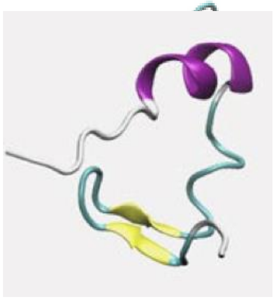
Molecular dynamics

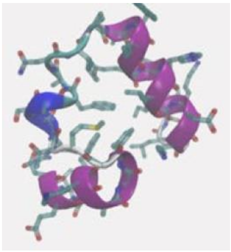


- consider a MD of unfolding/refolding villing headpiece
- for each of the N ~ 32000 configurations, D=32 dihedral angles.

We find four manifolds

- d=12 d=13 d=13 d=23
- Q=0.53 Q=0.58 Q=0.64 Q=0.89 Fraction of native contacts



The folded state is recognized from its higher ID!

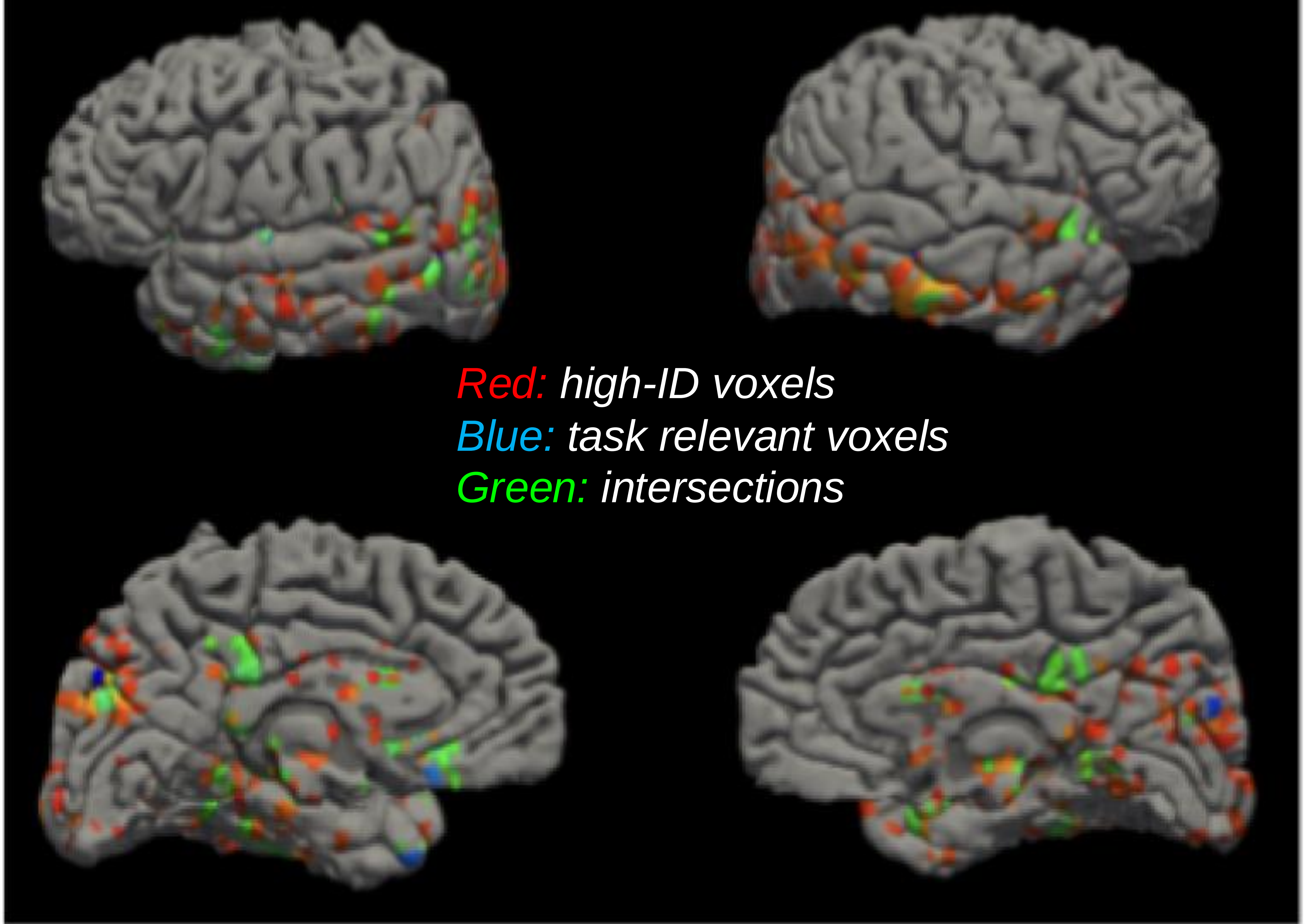
fMRI time series of BOLD signal

N ~ 30'000 voxels with
D = 202 scans

We find two manifolds d1 = 16 and
d2 = 32

Task-relevant voxels are in the manifold with higher ID

Low-dimensional manifold mostly includes “noise” voxels

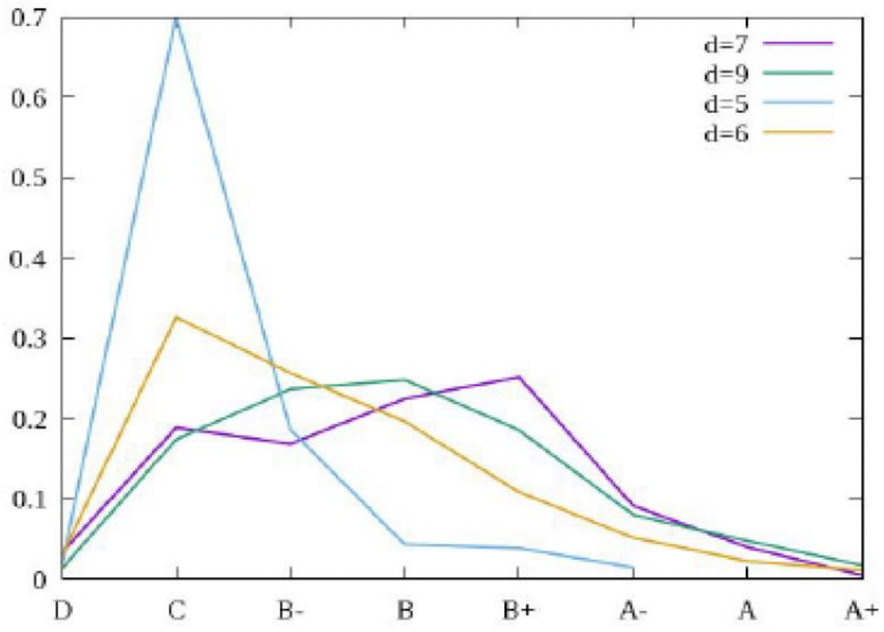

Red: high-ID voxels
Blue: task relevant voxels
Green: intersections

Firms from Compustat

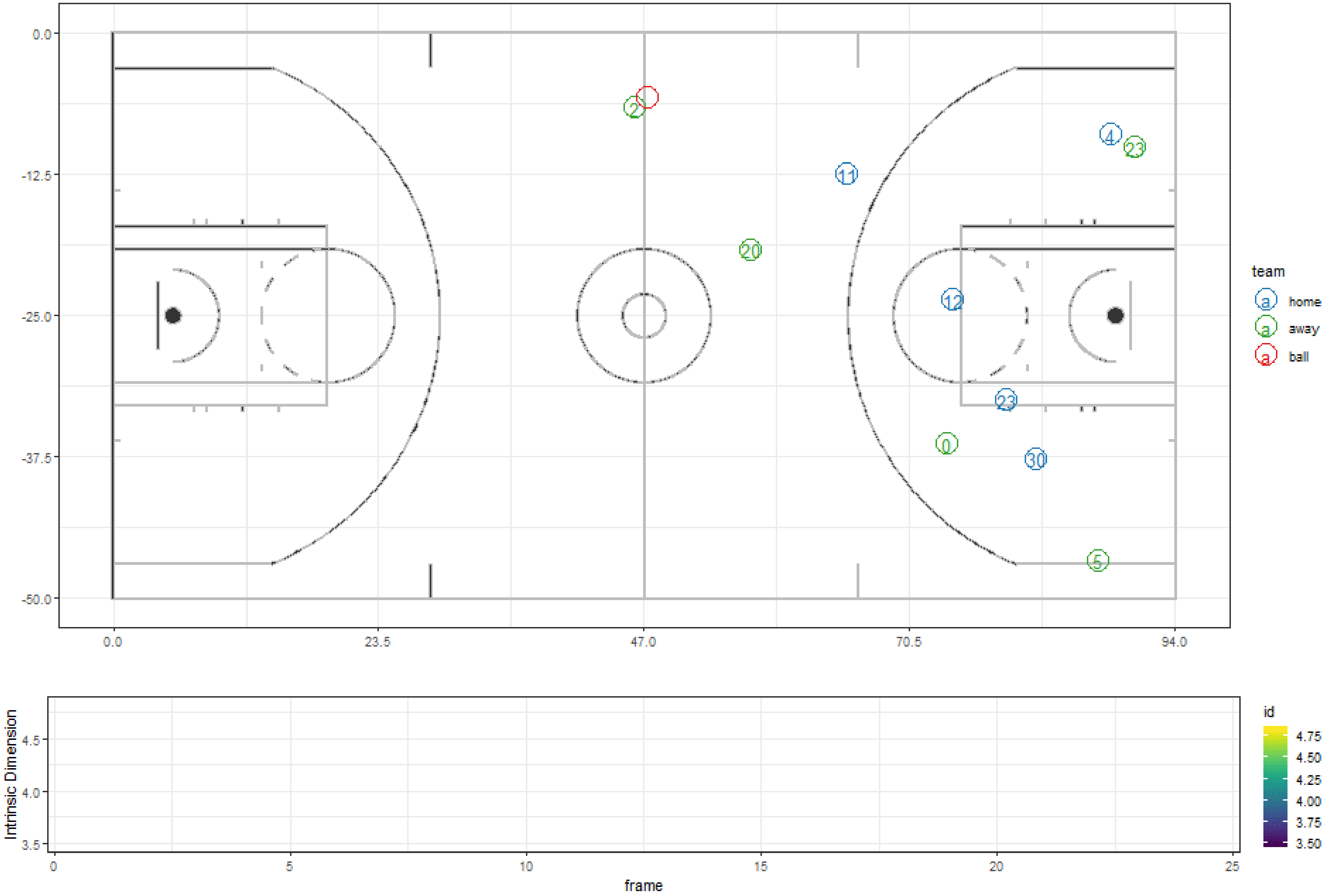
- consider ~8000 firms in the Compustat Database
- for each of the firms, D=31 balance sheet variables

We find four manifolds: d=5, d=6, d=7, d=9

We compute S&P ratings for the different manifolds



Lower dimension tends to have lower ratings!



Curtesy of Edgar Santos-Fernando

Gene expression data

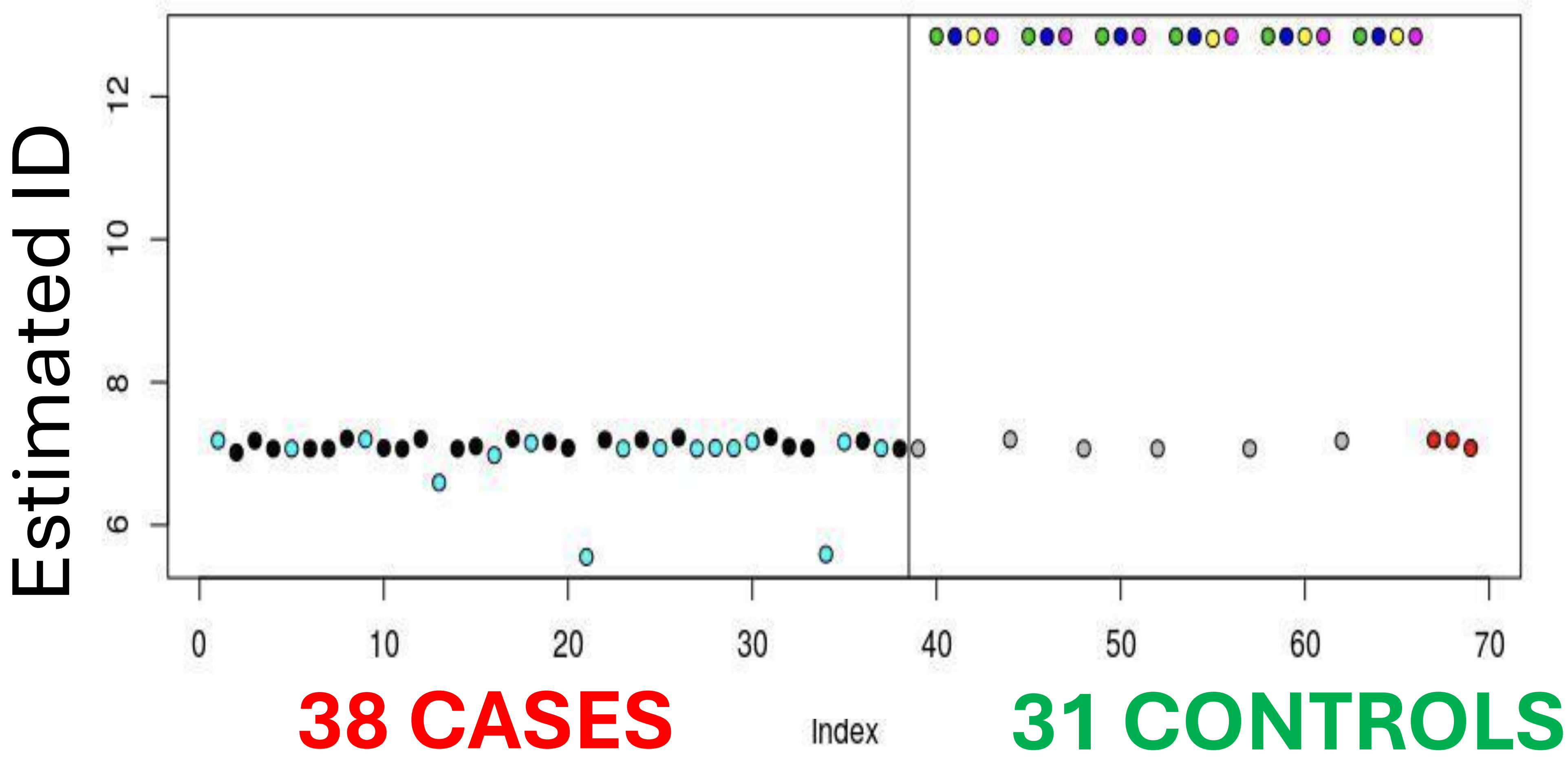
Joint work with Luciano Cascione, Institute of Oncology Research, USI

$D \approx$ 16.900 genes expressions
N = 69 tissue samples

The first 38 samples are CASES affected by Diffuse large B-cell lymphoma - DLBCL

The last 31 constitute the CONTROL group

69 tissues colored by cell typology
The vertical line divides case / control groups
The y-axis: posteriori medians of the ID for each tissue sample



Binomial Intrinsic Dimension Estimator (BIDE) ... and its Bayesian counterpart

BIDE (MLE):

$$\widehat{d} = \frac{\log\left(\frac{1}{n}\sum_{i=1}^{n} k_{A,i} / \frac{1}{n}\sum_{i=1}^{n} k_{B,i}\right)}{\log(\tau)}$$

BBIDE Bayesian estimator:

- Prior: $p = \tau^d \sim \mathrm{Beta}(\alpha_0, \beta_0)$
- Posterior:

$$p | k_{B,1}, \ldots, k_{B,n} \sim \mathrm{Beta}(\alpha, \beta)$$

  where $\alpha = \alpha_0 + \sum_{i=1}^{n} k_{A,i}$ and $\beta = \beta_0 + \sum_{i=1}^{n} (k_{B,i} - k_{A,i})$.
- Posterior expectation and variance:

$$\mathsf{E}[d] = \frac{\psi_0(\alpha) - \psi_0(\alpha+\beta)}{\log(\tau)}, \quad \mathsf{Var}[d] = \frac{\psi_1(\alpha) - \psi_1(\alpha+\beta)}{\log(\tau)^2}$$

Adaptive BIDE (ABIDE) and its Bayesian twin

Algorithm Adaptive-BIDE

$d_{\text{current}} \leftarrow$ **ID from Two-NN**
$d_{\text{next}} \leftarrow 0$
for $it <$ max_iter **do**
  $\tau = 0.2032^{1/d_{\text{current}}}$
  for $i < n$ **do**
    compute k_i^* (using d_{current}) and set $k_{B,i}^* = k_i^*$
    $t_{B,i}(k_i^*) = r_{ik_i^*}$
    $t_{A,i}(k_i^*) = \tau\, t_{B,i}(k_i^*)$
    $k_{A,i}^* = \sum_{j=1}^{n} \mathbf{1}\{t_{A,i}(k_i^*) - r_{i,j} > 0\}$
  end for
  $d_{\text{next}} = \frac{\log\left(\frac{1}{n}\sum_{i=1}^{n} k_{A,i}^* / \frac{1}{n}\sum_{i=1}^{n} k_{B,i}^*\right)}{\log(\tau)}$ *or* $d_{\text{next}} = \frac{\psi_0(\alpha^*) - \psi_0(\alpha^* + \beta^*)}{\log(\tau)}$,
  if then $|d_{\text{current}} - d_{\text{next}}| < \delta$ **break**
  end if
  $d_{\text{current}} = d_{\text{next}}$
end for
$d^* = d_{\text{next}}$
for $i <$ n **do**
  compute k_i^* (using d^*)
end for
return $d^*,\ k_i^*$

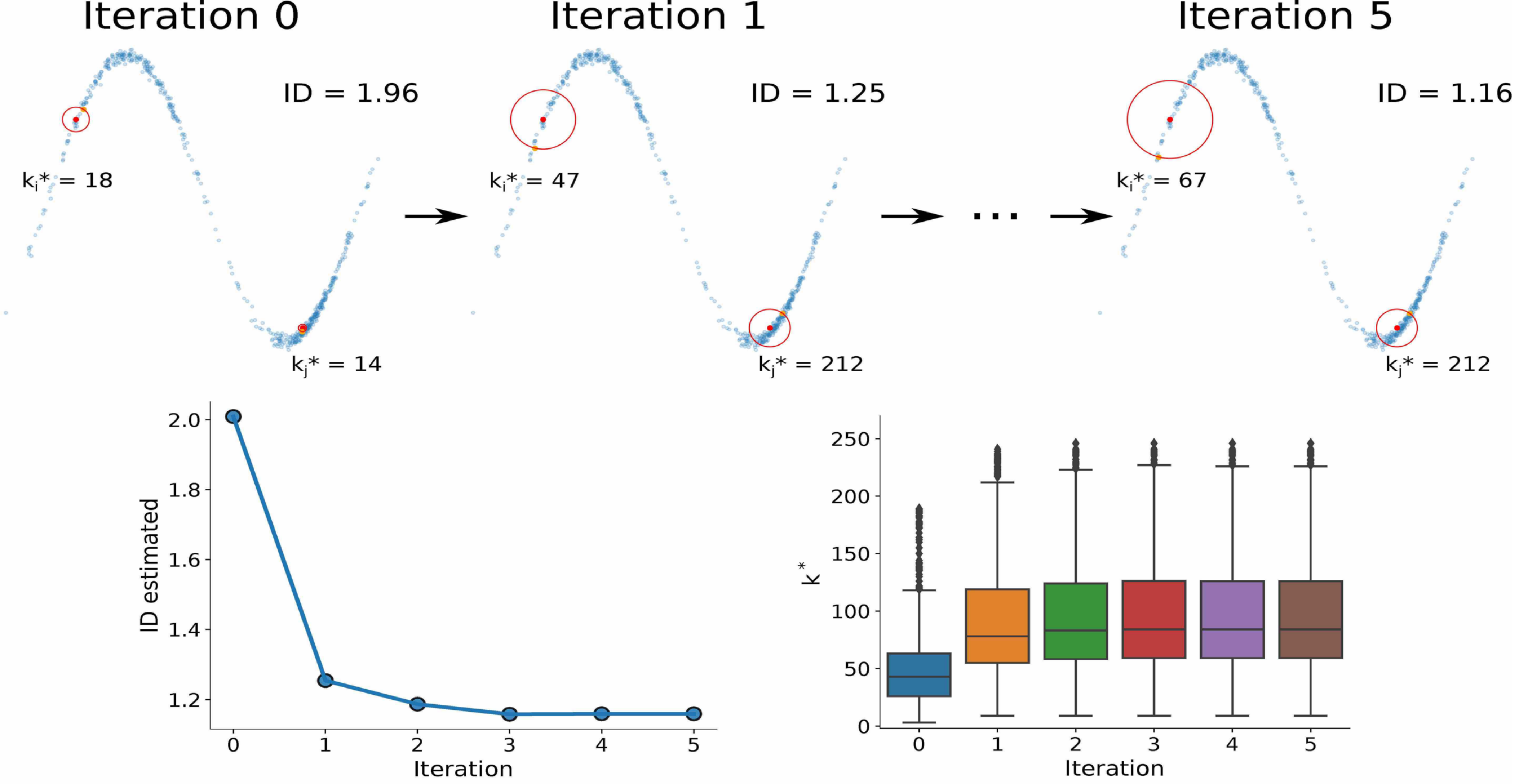


Figure: Consider $\mathcal{N}(\frac{\pi}{2}, 1)$ and $\mathcal{N}(\frac{5}{3}\pi, 0.5^2)$, sample $X_1, \ldots, X_{1000}$ points, 500 from each one of the two. Then map the points on a curved manifold by adding a second coordinate given by $Y_i = \sin(X_i)$.

Locally Linear Embedding - LLE

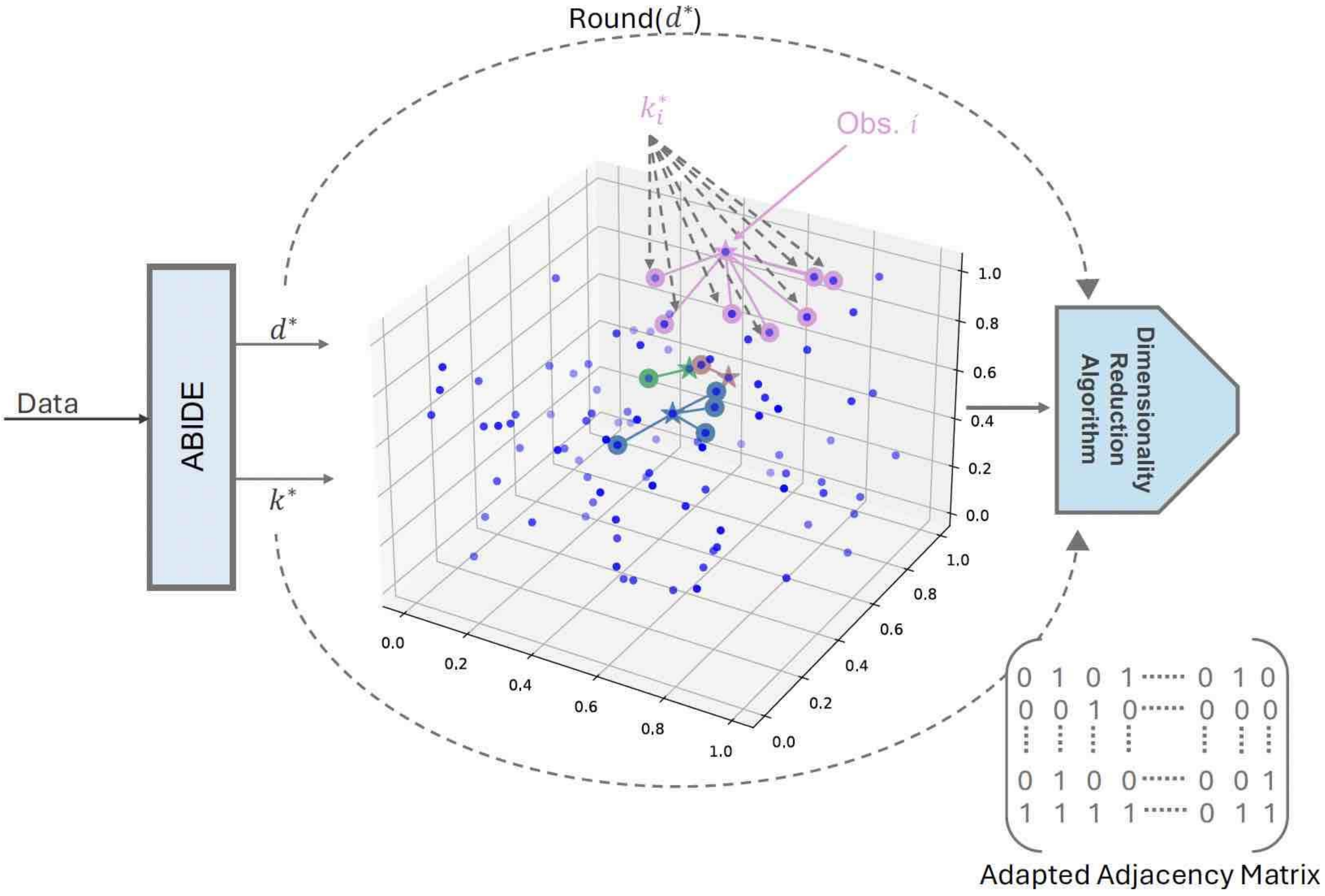
LLE: unsupervised learning algorithm that computes low dimensional, neighborhood preserving embeddings

TS Roweis and LK Saul, Science, 2000, $\sim$ 20.000 citations

LLE: eigenvector method for nonlinear dimensionality reduction

Given a dataset $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^D$, LLE finds low dimensional embedding vectors $\{\mathbf{y}_i\}_{i=1}^n \subset \mathbb{R}^d$

Given inputs: k and d
1. Compute the k neighbors of each data point x_i
2. Compute the weights that best reconstruct each data point from its neighbors, minimizing the reconstruction error
3. Compute the vectors y_i best reconstructed by the weights by minimizing a cost function



Round(d^*)
Data
ABIDE
d^*
k^*
k_i^*
Obs. i
Dimensionality Reduction Algorithm
0 1 0 1 ⋯ 0 1 0
0 0 1 0 ⋯ 0 0 0
0 1 0 0 ⋯ 0 0 1
1 1 1 1 ⋯ 0 1 1
Adapted Adjacency Matrix

Diffuse Large B-Cell Lymphoma - DLBCL

$D \approx 16.900$ genes expressions recorded on $N = 69$ tissues

38 patients with DLBCL with different variants of lymphoma

- Activated B-Cell-like - ABC
- Germinal Center B-Cell-like - GCB

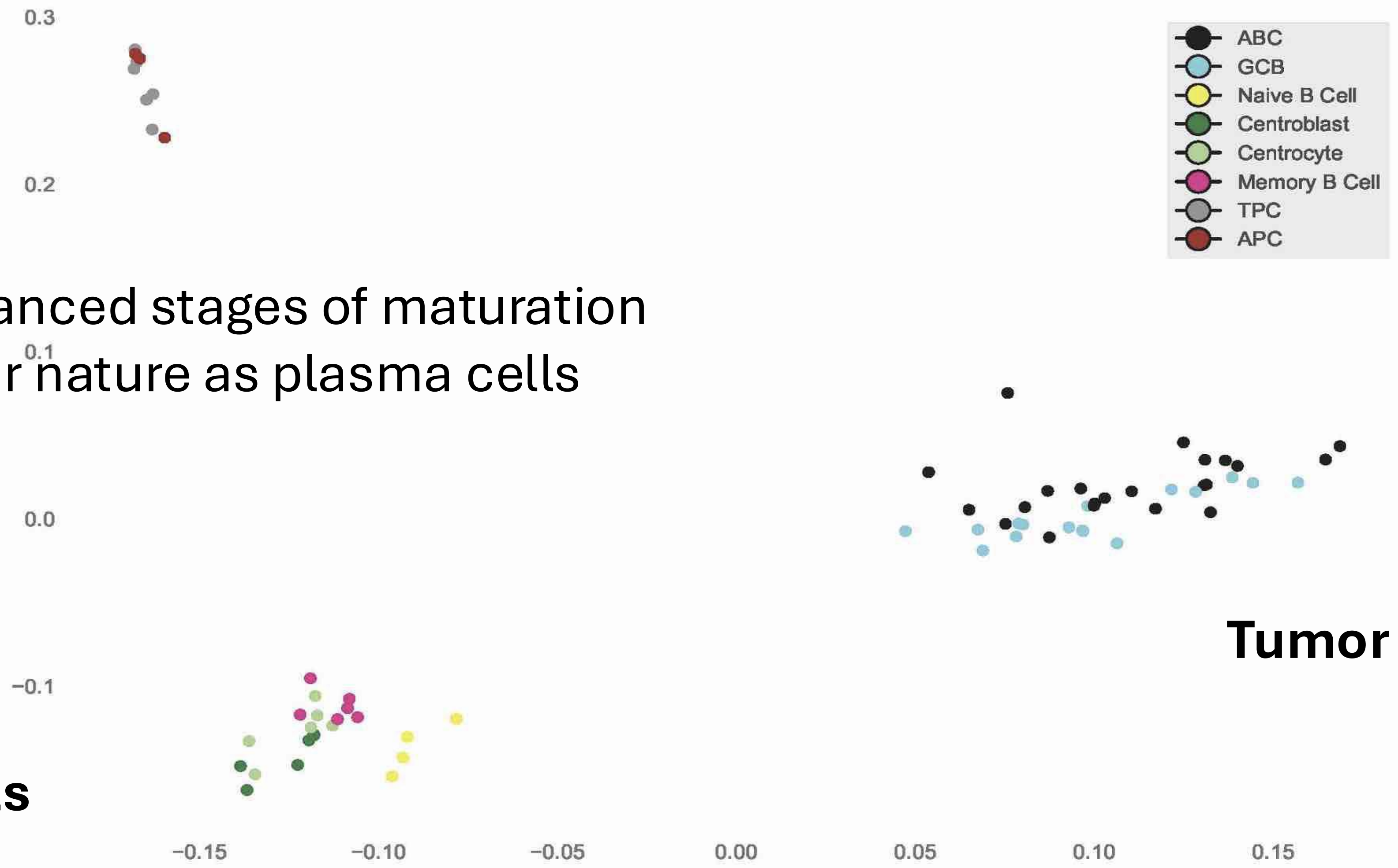
31 healthy donors of B-cell samples at various stages of maturation

- Naive B Cell - NB
- CentroBlast - CB
- CentroCyte - CC
- Memory B Cell - MEM
- Transitional Plasma Cells - TCP
- Advanced-stage Plasma Cells in Bone Marrow - APC / BMPC

TPC and APC, being plasma cells, represent the most advanced stage of B-cell maturation

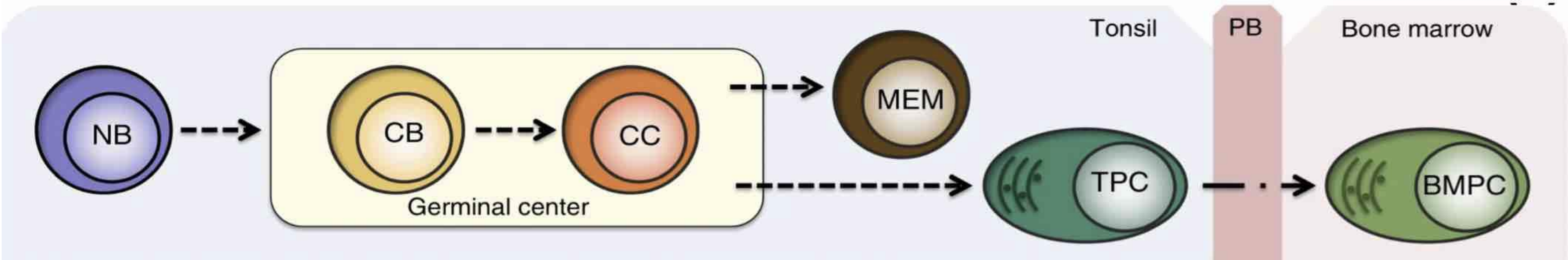
TPC and APC cells

are in the most advanced stages of maturation consistent with their nature as plasma cells



Tumor cells

Healthy cells



ID of Covid-19

Joint with

A. Varghese, E. Santos-Fernandez, F. Denti, Kerrie Mengersen

	CSI 1.3.20 – 29.5.21	New Cases pmp 1.3.20 – 29.5.21	New Deaths pmp 1.3.20 – 29.5.21
Country 1			
Country 2			
Country 3			
Country 4 , , ,			
Country 115			

D =

454x3 =

1362

N = 115

excluded: > than 20% missing

< 1 million population

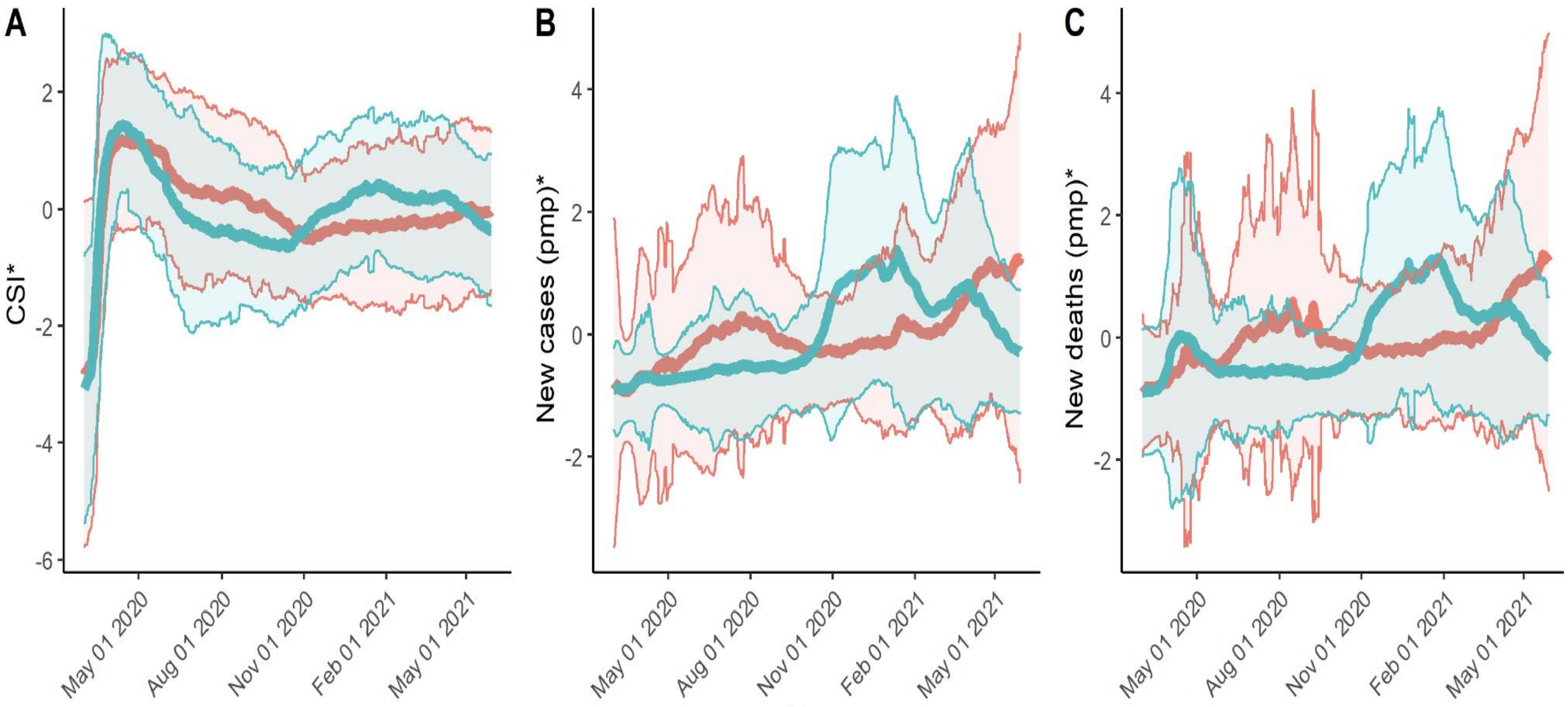
MAIN CONCLUSION

high-income countries are more likely to lie on low-dimensional manifolds,

likely arising from aging populations and comorbidities,

causing increased per capita mortality from COVID-19

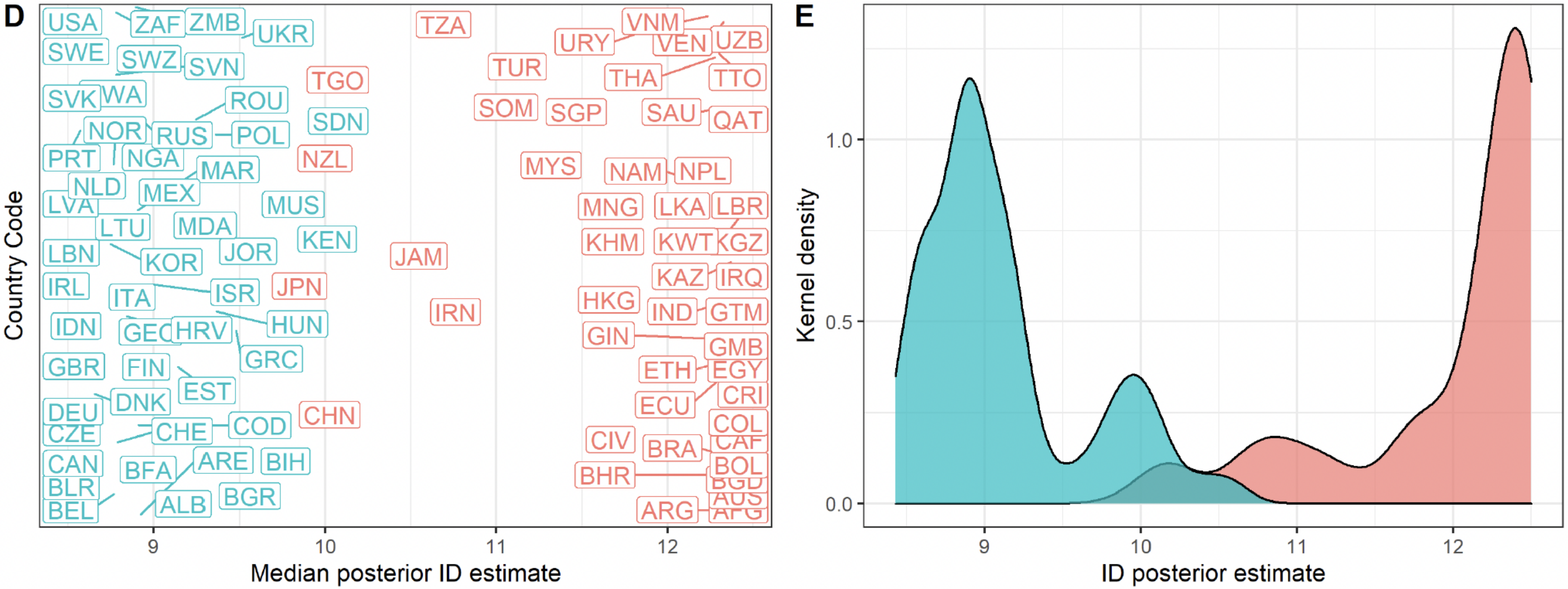
1st Mar 2020 to 29th May 2021

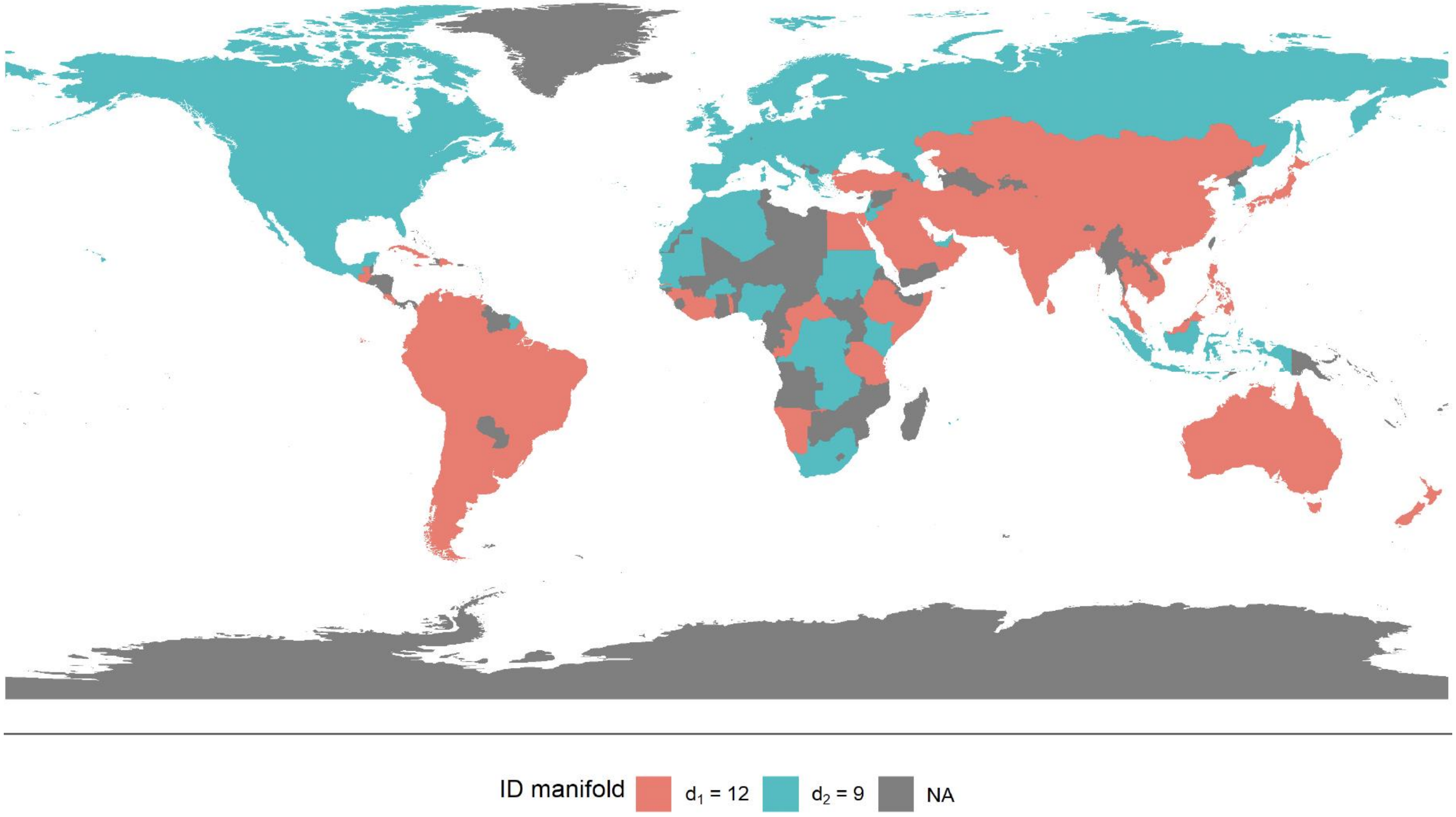


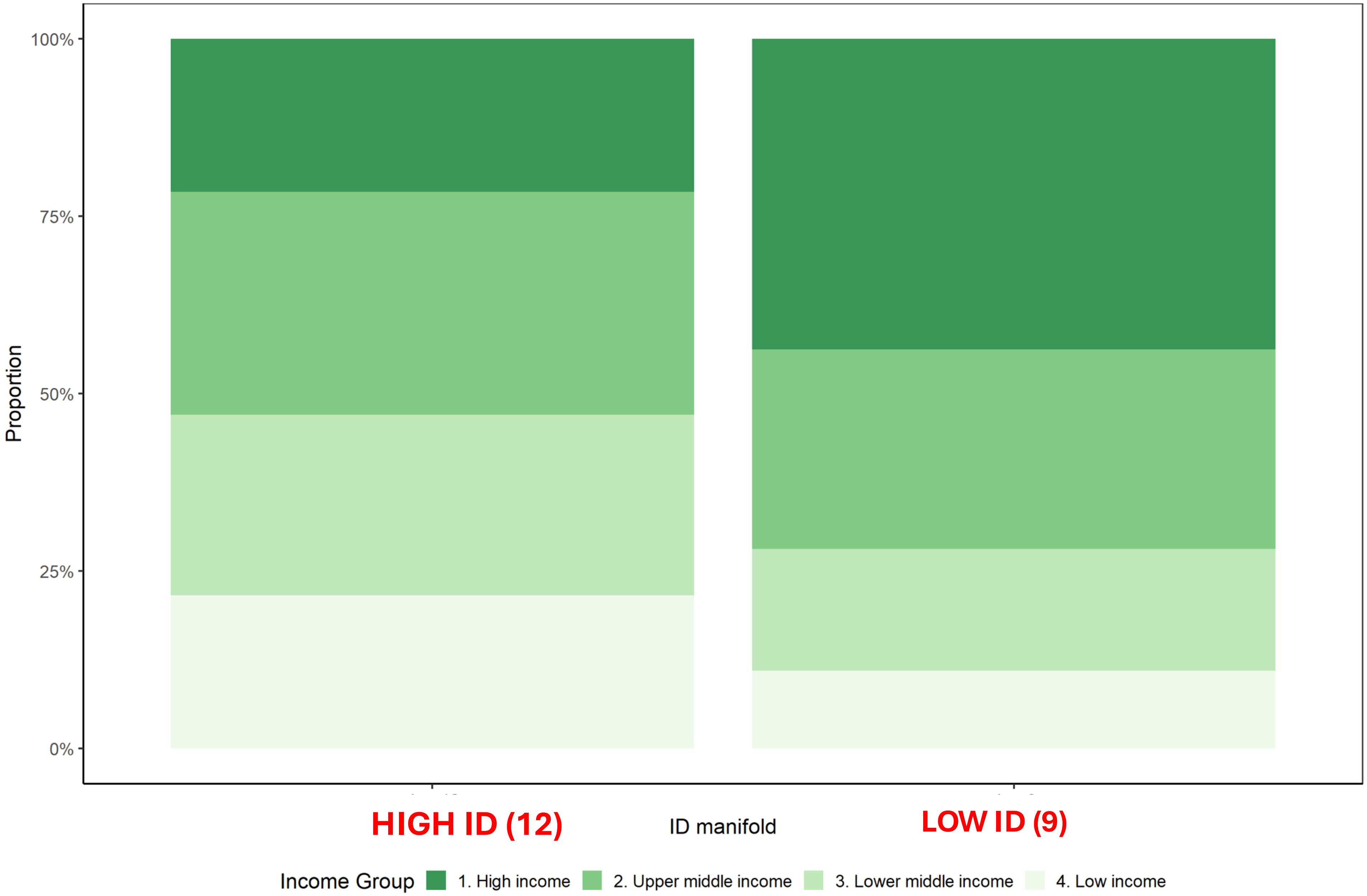
CSI

new cases pmp

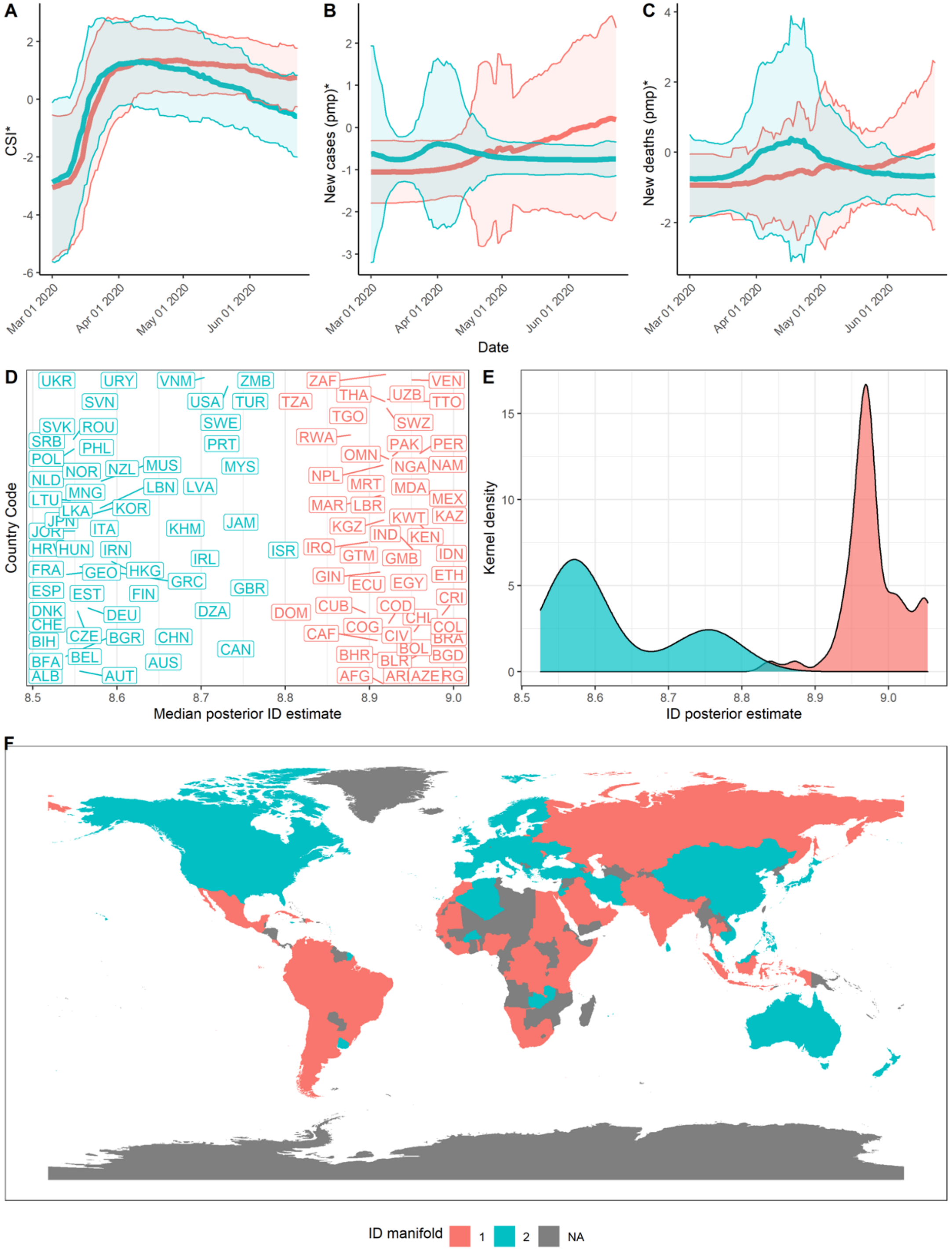
new deaths pmp

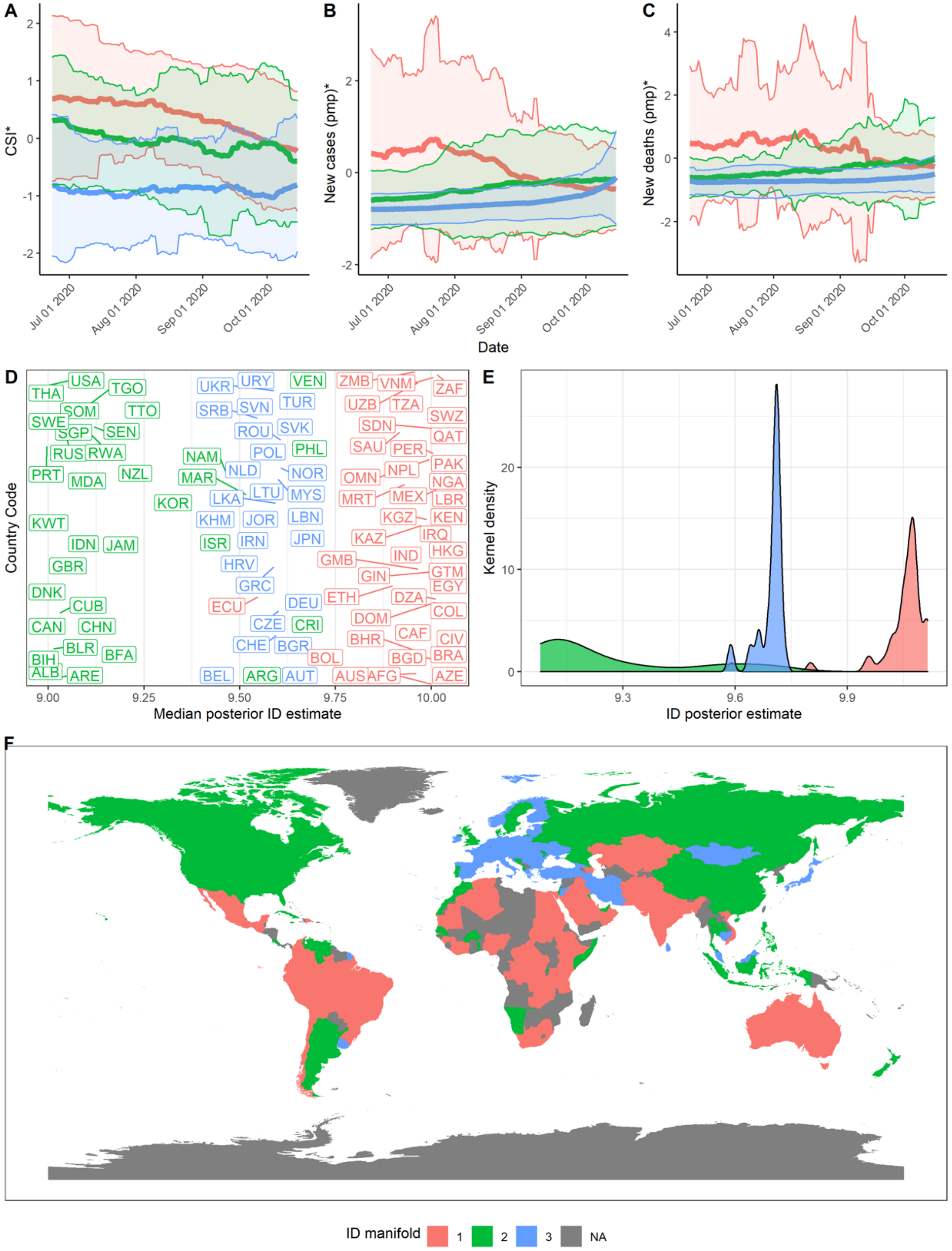

D
E
Country Code
Median posterior ID estimate
9
10
11
12
USA
ZAF
ZMB
UKR
SWE
SWZ
SVN
SVK
WA
ROU
NOR
RUS
POL
PRT
NGA
MAR
NLD
MEX
LVA
LTU
MDA
MUS
LBN
KOR
JOR
KEN
IRL
ITA
ISR
IDN
GEO
HRV
HUN
GRC
GBR
FIN
EST
DNK
DEU
CZE
CHE
COD
CAN
BFA
ARE
BIH
BLR
BEL
ALB
BGR
TZA
TGO
SDN
NZL
JAM
JPN
IRN
CHN
TUR
SOM
SGP
MYS
URY
VNM
VEN
UZB
THA
TTO
SAU
QAT
NAM
NPL
MNG
LKA
LBR
KHM
KWT
KGZ
KAZ
IRQ
HKG
IND
GTM
GIN
GMB
ETH
EGY
ECU
CRI
COL
CIV
CAF
BRA
BOL
BHR
BGD
AUS
ARG
AFG
Kernel density
ID posterior estimate
0.0
0.5
1.0


ID manifold
$d_1 = 12$
$d_2 = 9$
NA



100%
75%
50%
25%
0%
Proportion
HIGH ID (12)
ID manifold
LOW ID (9)
Income Group
1. High income
2. Upper middle income
3. Lower middle income
4. Low income


A
B
C
CSI*
New cases (pmp)*
New deaths (pmp)*
Mar 01 2020
Apr 01 2020
May 01 2020
Jun 01 2020
Date
D
E
Country Code
UKR URY VNM ZMB USA TUR SVN SWE SVK ROU SRB PHL PRT POL MUS NZL NOR NLD MYS LBN LVA MNG LTU KOR LKA JPN JOR ITA KHM JAM HRV HUN IRN ISR IRL FRA GEO HKG GRC GBR ESP EST FIN DNK DEU DZA CHE CZE BGR CHN BIH CAN BEL AUS BFA ALB AUT
ZAF VEN THA UZB TZA TTO TGO SWZ RWA PAK PER OMN NGA NAM NPL MRT MDA MEX MAR LBR KWT KAZ KGZ IND KEN IRQ GTM GMB IDN GIN ECU EGY ETH CUB COD CRI DOM CHL COL COG CAF CIV BRA BOL BHR BLR BGD AFG ARG AZE
Median posterior ID estimate
Kernel density
ID posterior estimate
F
ID manifold
1
2
NA


A
B
C
CSI*
New cases (pmp)*
New deaths (pmp)*
Jul 01 2020
Aug 01 2020
Sep 01 2020
Oct 01 2020
Date
D
Country Code
USA THA TGO SOM TTO SWE SGP SEN RUS RWA PRT MDA NZL KOR KWT IDN JAM GBR DNK CUB CAN CHN BLR BFA BIH ALB ARE ARG VEN PHL NAM MAR
UKR URY SRB SVN TUR ROU SVK POL NLD NOR LKA LTU MYS KHM JOR LBN ISR IRN JPN HRV GRC DEU CZE CRI CHE BGR BEL AUT
ZMB VNM ZAF UZB TZA SWZ SDN QAT SAU PER PAK OMN NPL NGA MRT MEX LBR KGZ KEN KAZ IRQ IND HKG GMB GTM GIN EGY ECU ETH DZA COL DOM CAF BHR CIV BGD BRA BOL AUS AFG AZE
9.00
9.25
9.50
9.75
10.00
Median posterior ID estimate
E
Kernel density
0
10
20
9.3
9.6
9.9
ID posterior estimate
F
ID manifold
1
2
3
NA

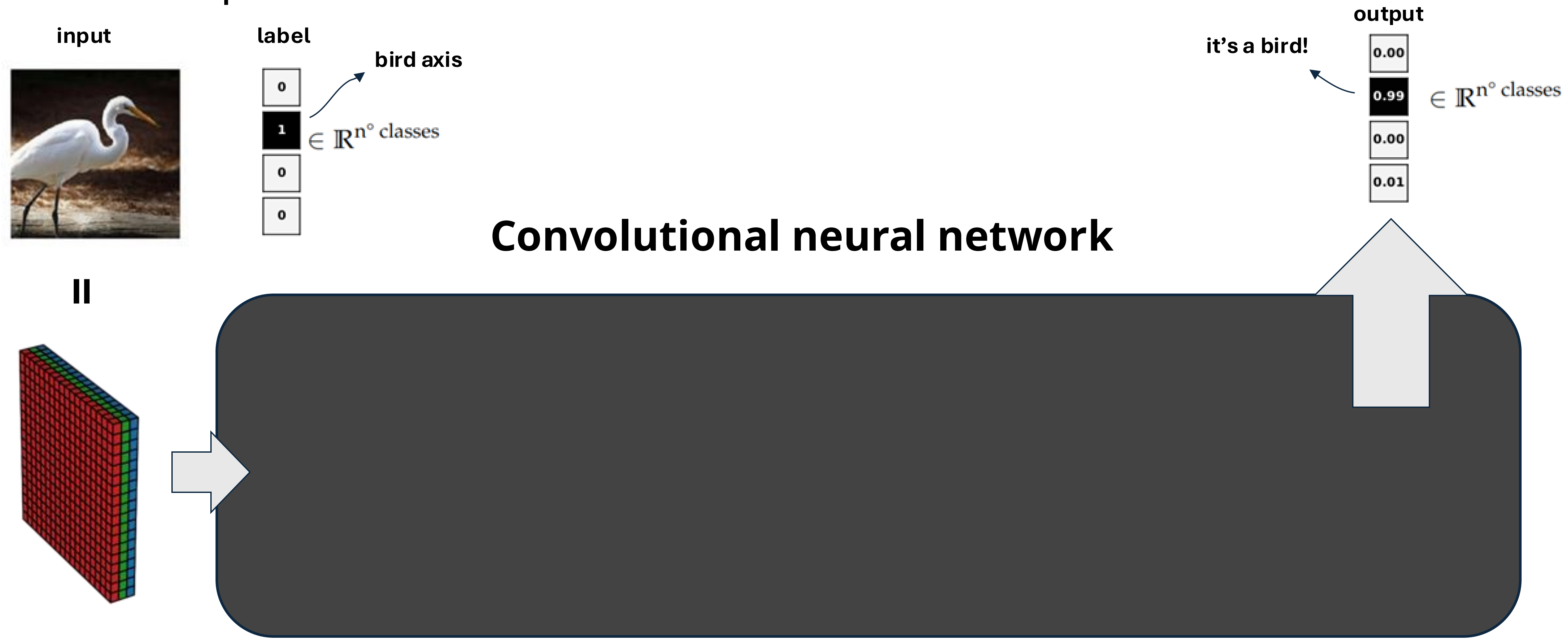
Ansuini, Laio, Macke, Zoccolan, Advances in NIPS , 2019

- Study the ID of data representations in CNN
- In a trained CNN, the ID is orders of magnitude smaller than the number of units in each layer
- Across layers, the ID first increases then decreases
- The ID of the last hidden layer predicts classification accuracy on the test set in CNN to classify images
- Not true for untrained networks
  Not true for networks trained on randomized labels

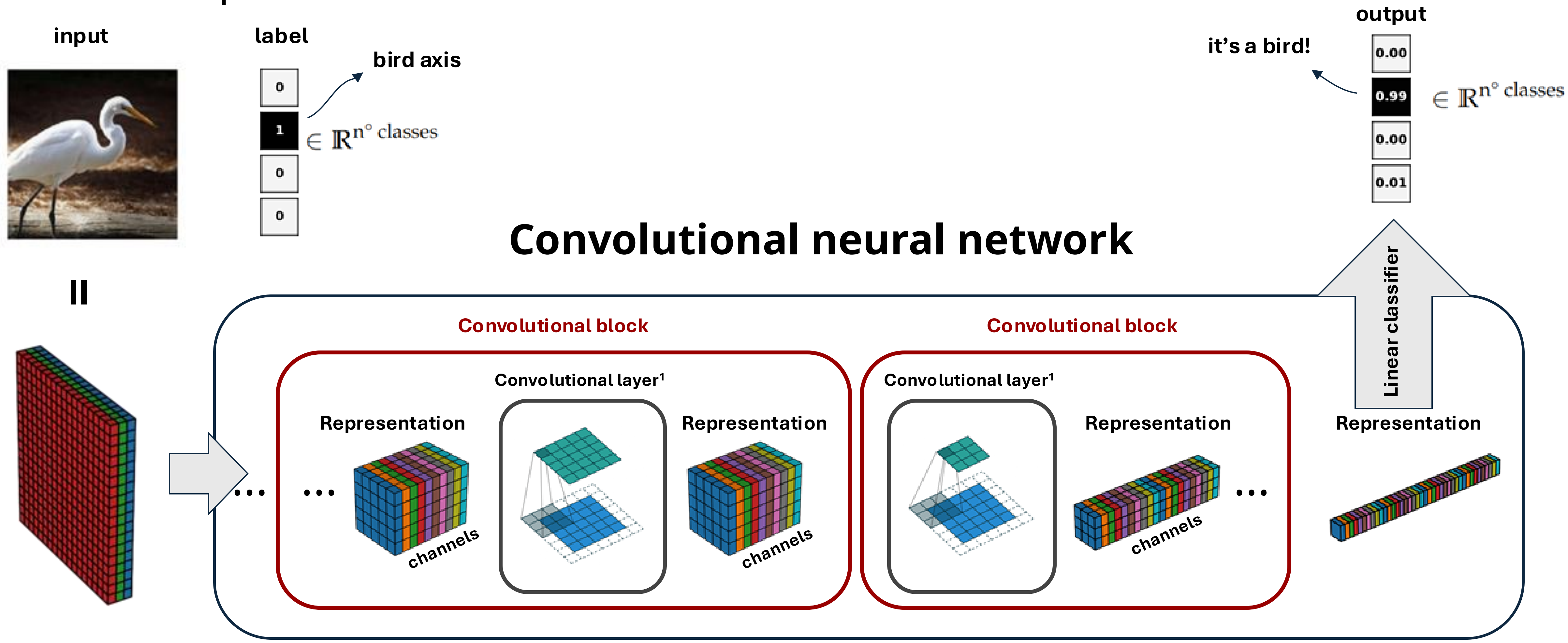
Conclusion: NN that can generalize are those that transform the data into low-dimensional, but not necessarily flat manifolds

ID exploitation for selecting CNN architectures and training procedures

Hidden representations of convolutional networks

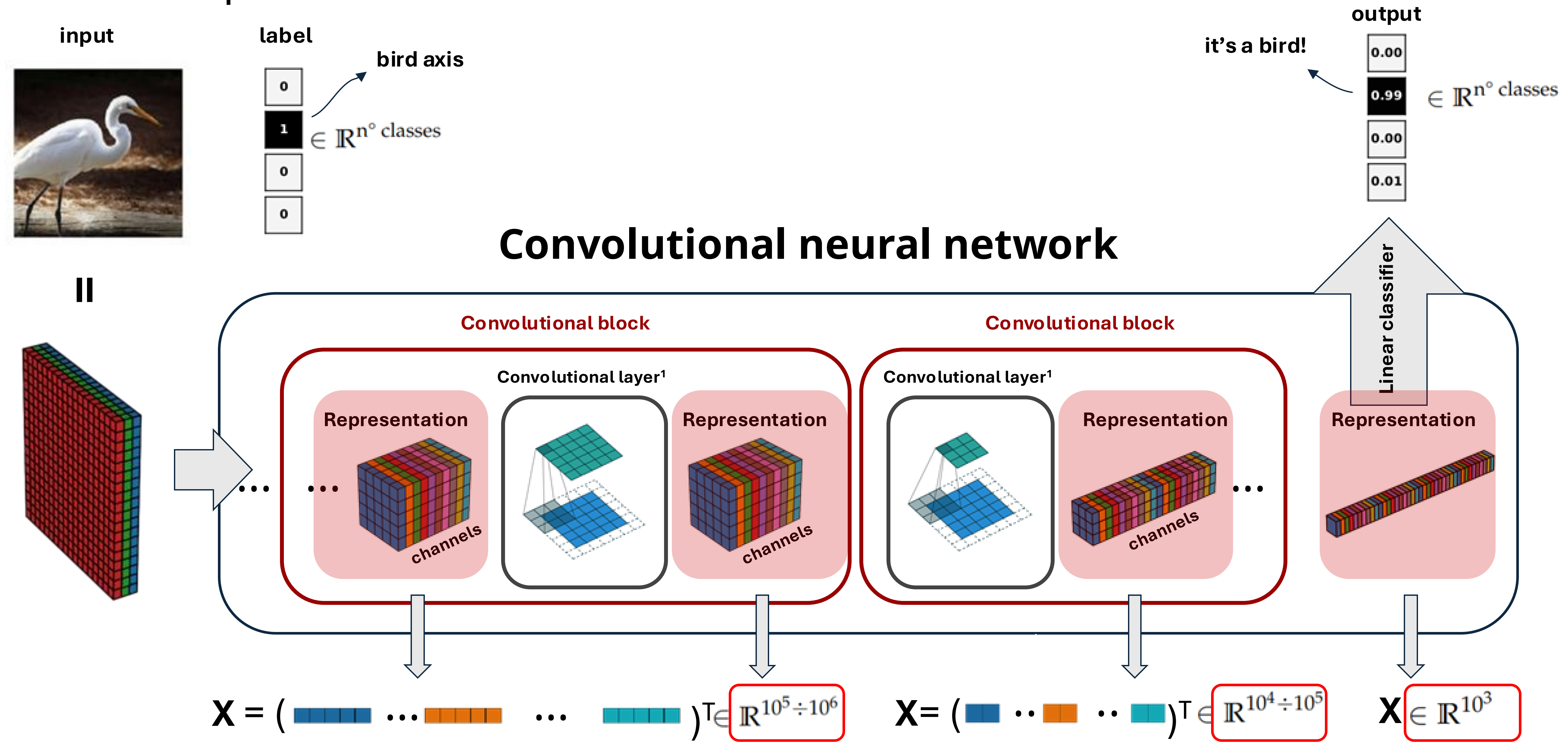


Hidden representations of convolutional networks



1 https://github.com/vdumoulin/conv_arithmetic

Hidden representations of convolutional networks



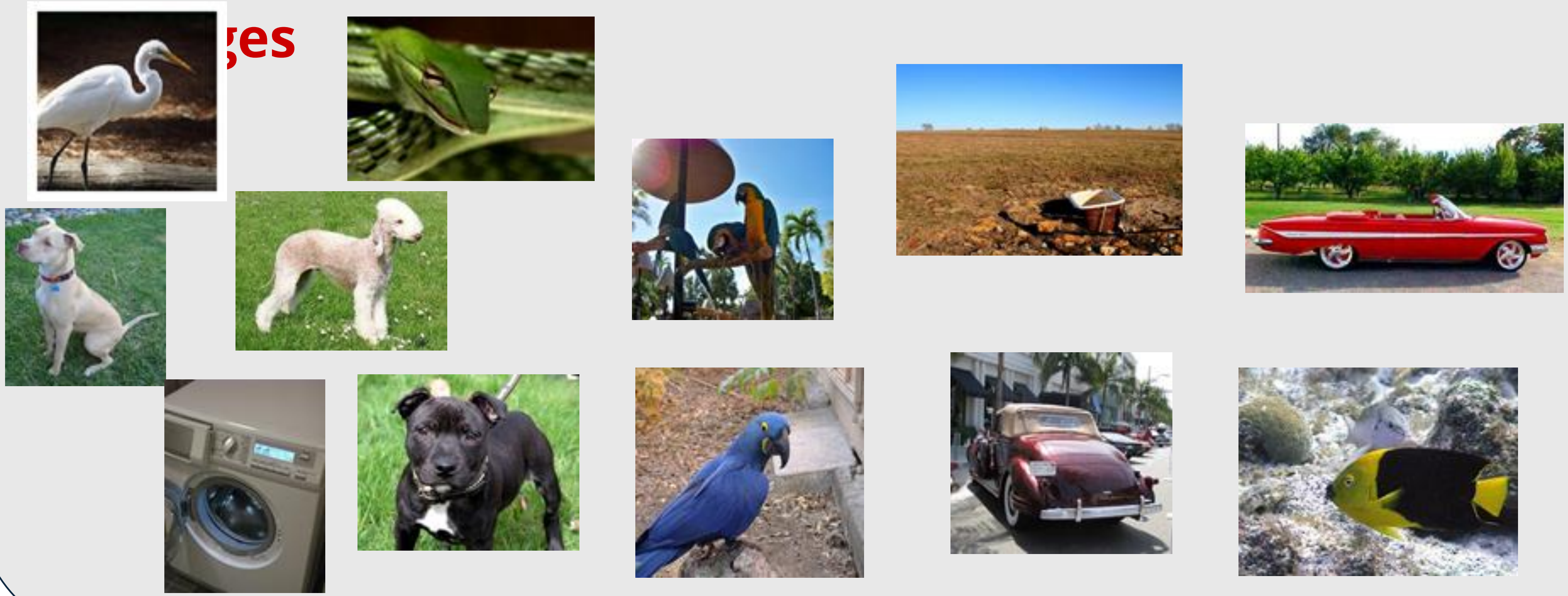
1 https://github.com/vdumoulin/conv_arithmetic

ID of hidden representations of ResNet152

Dataset ImageNet: 1.2 million images, 1000 classes

We consider a subset of **300 classes with 300 images per class** for a total of **90 000**

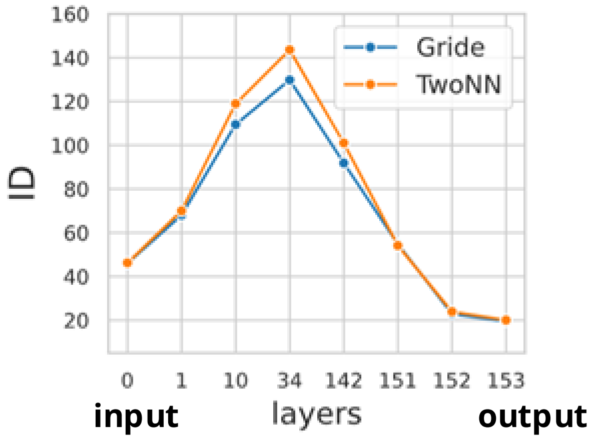
ges



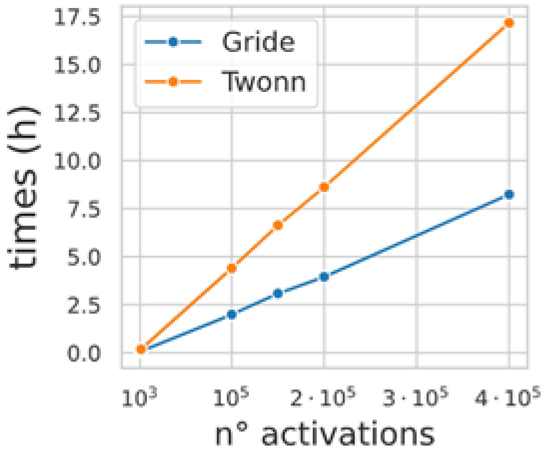
ID of hidden representations of ResNet152

Ansuini et al., NeurIPS, 2019

Trained network



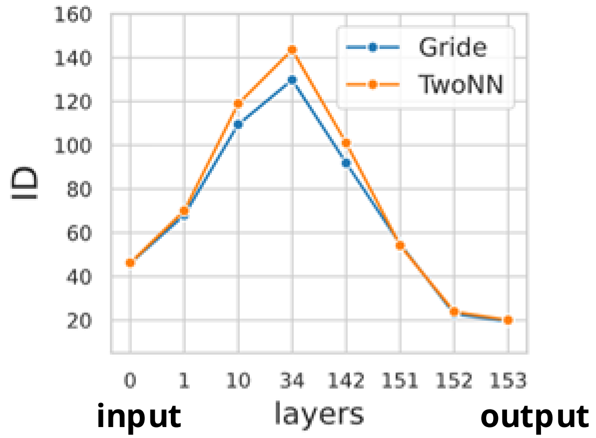
2x speed-up wrt TwoNN



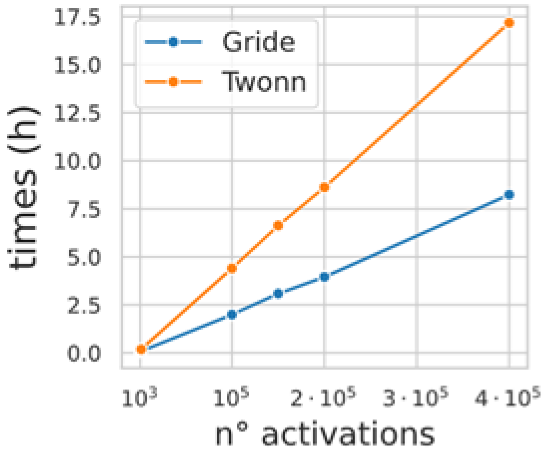
ID of hidden representations of ResNet152

Ansuini et al., NeurIPS, 2019

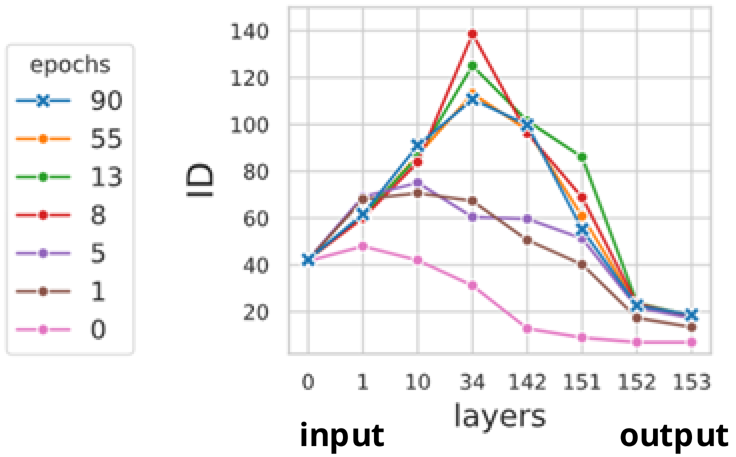
Trained network



2x speed-up wrt TwoNN



Evolution of the ID during training

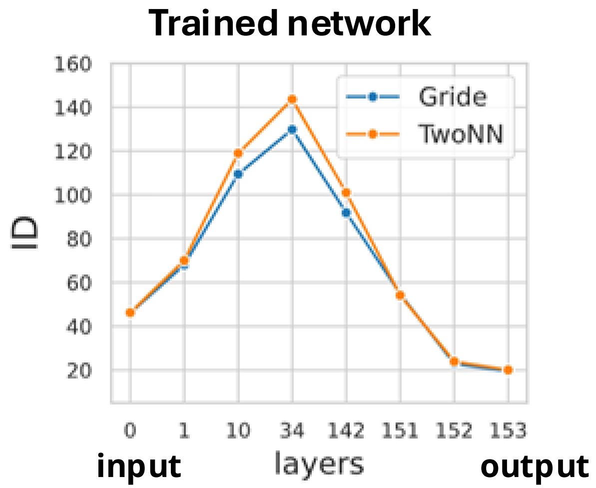


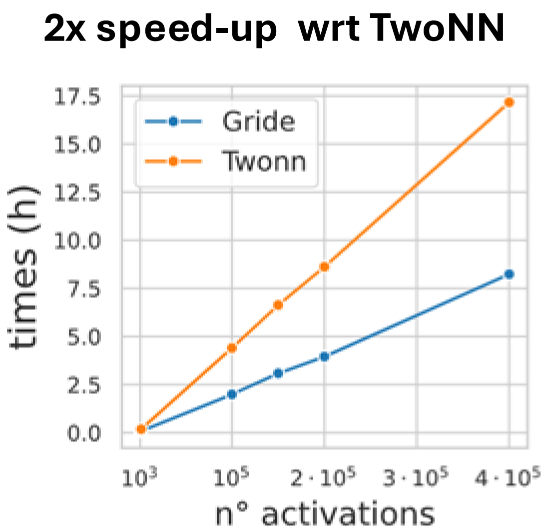
90 epochs

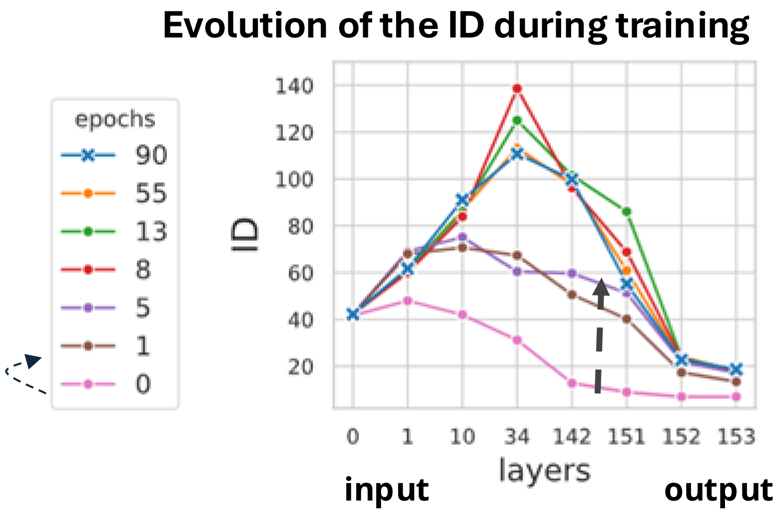
Greatest change of the ID in the first 10 epochs

ID of hidden representations of ResNet152

Ansuini et al., NeurIPS, 2019





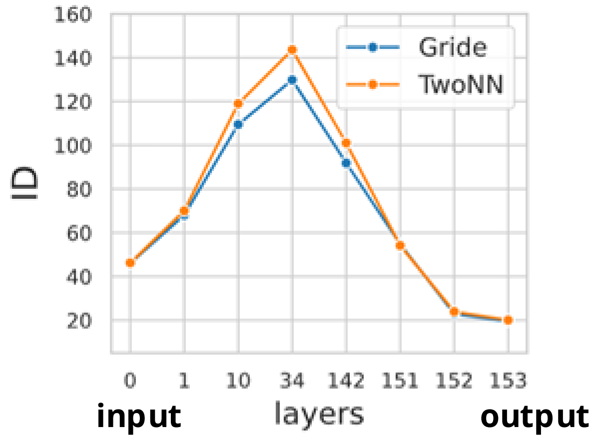


90 epochs
Greatest change of the ID in the first 10 epochs

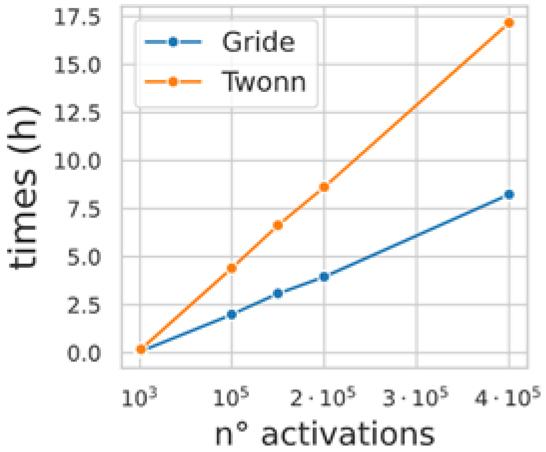
ID of hidden representations of ResNet152

Ansuini et al.,
NeurIPS, 2019

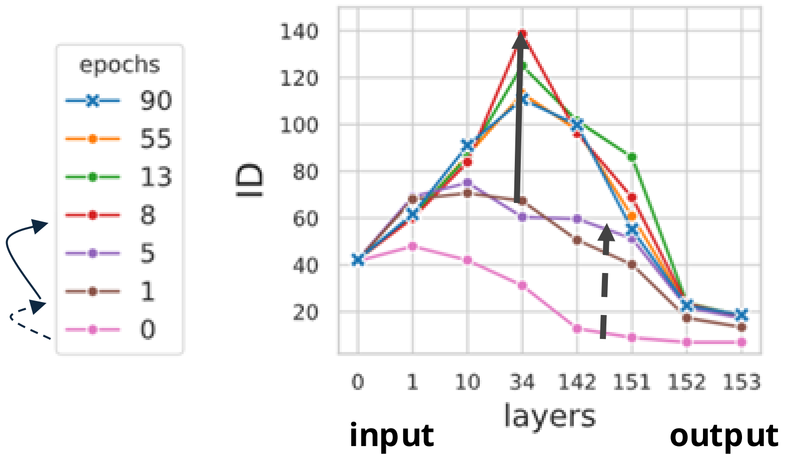
Trained network



2x speed-up wrt TwoNN



Evolution of the ID during training

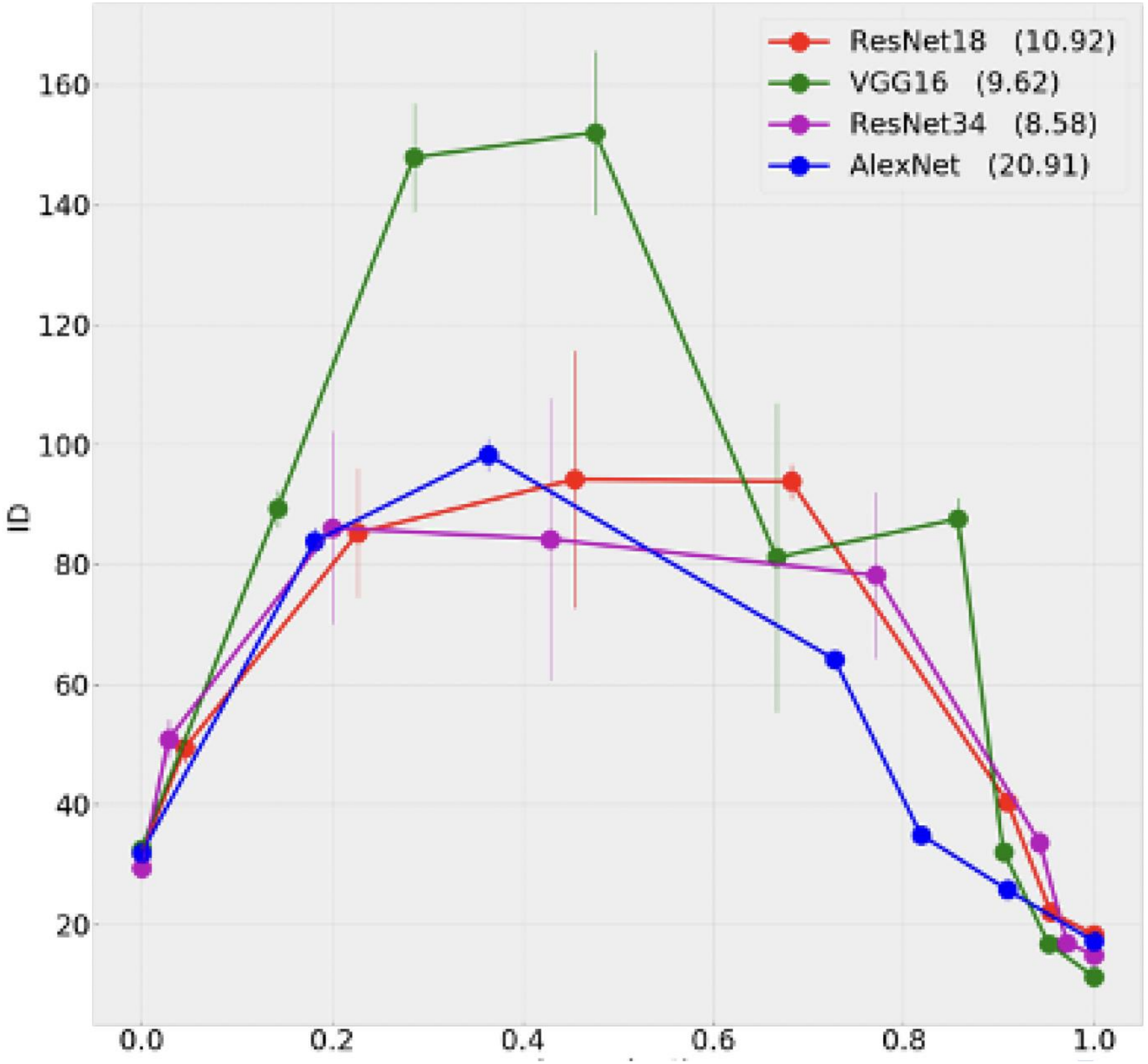


90 epochs
Greatest change of the ID
in the first 10 epochs

CNN architectures: AlexNet, Vgg, Resnet pretrained on ImageNet

50 samples of $\sim$ 2000 images

ID computed as function of the NN layer depth



R package by F. Denti
Cattolica University, Italy

Python suite by SISSA,
Trieste, Italy

References

Facco, d'Errico, Rodriguez, Laio;
Estimating the intrinsic dimension of datasets by a minimal neighborhood information.
Scientific Reports 2017

Allegra, Facco, Denti, Laio, Mira
Data segmentation based on the local ID,
Scientific Reports 2020

Denti, Doimo, Laio, Mira,
Distributional Results for Model-Based ID Estimators,
Scientific Reports 2022

Vargehese, Santos-Fernandez, Denti, Mira, Mengersen,
A global perspective on the intrinsic dimensionality of COVID-19 data
Scientific Reports 2023

Denti,
intRinsic: an R package for model-based estimation of the ID of a dataset,
J Stat Software 2023