ALGORITHMS FOR BIG DATA PROJECT

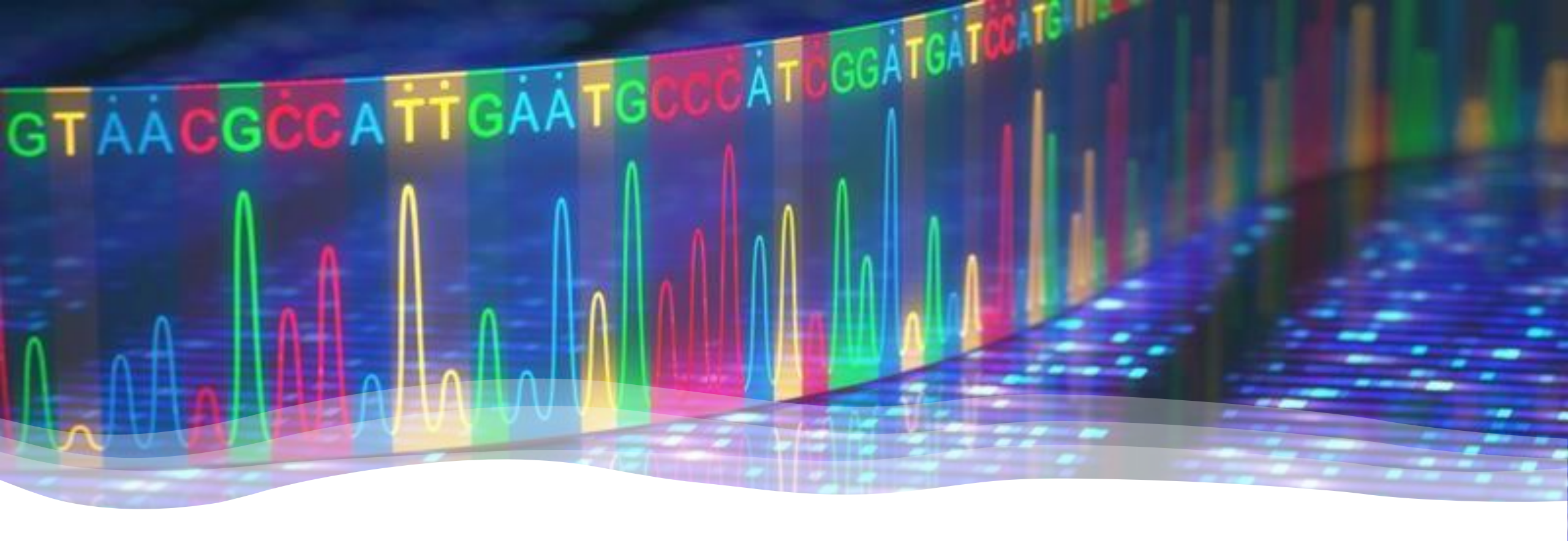
Topic: "Estimating the intrinsic dimension of datasets by a minimal neighborhood information"

Plan

- ➔ Definition of Intrinsic Dimension (ID)
- ➔ Importance of ID
- ➔ History about ID algorithm
- ➔ Presentation of our algorithm
- ➔ Runtime and storage analysis of our algorithm: ***2NN ID estimation***
- ➔ Evaluation of our algorithm with known ID dataset
- ➔ Some example

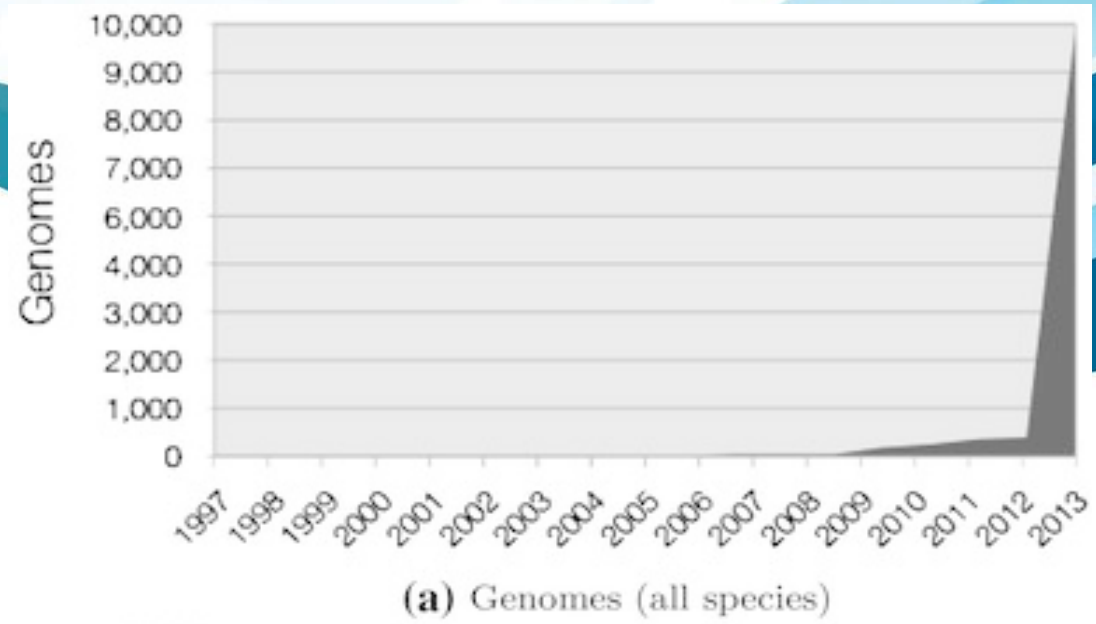
WHAT IS THE INTRINSIC DIMENSION AND WHY IS IT IMPORTANT?

HIGH DIMENSIONAL DATASETS AND THEIR PROBLEMS

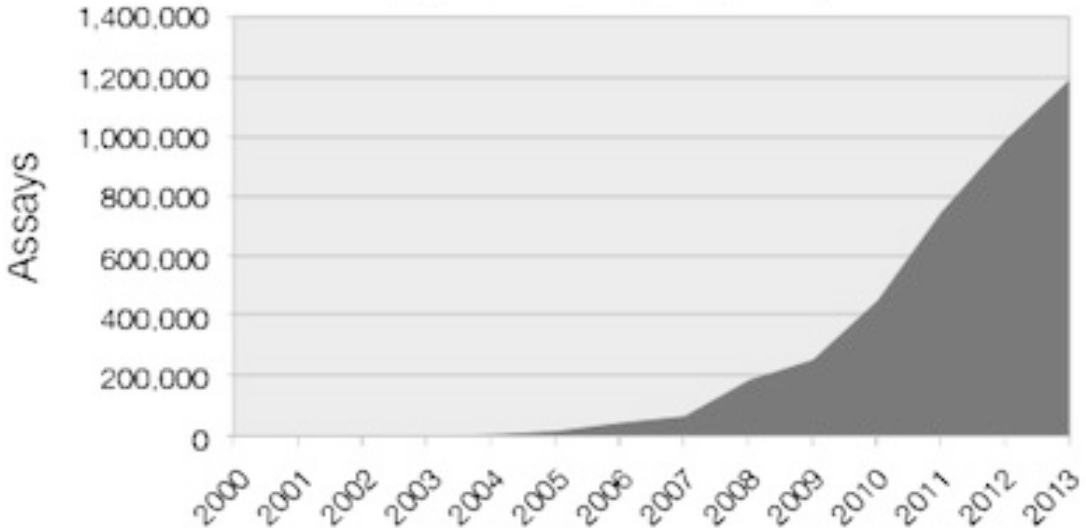


BIOINFORMATICS

NUMBER OF DATA INCREASE ALONG THE YEARS



(a) Genomes (all species)



(b) Gene expression data

Immage refference:
(PDF) Big Data Analytics in Bioinformatics: Architectures, Techniques, Tools and Issues (researchgate.net)

TEXT MINING

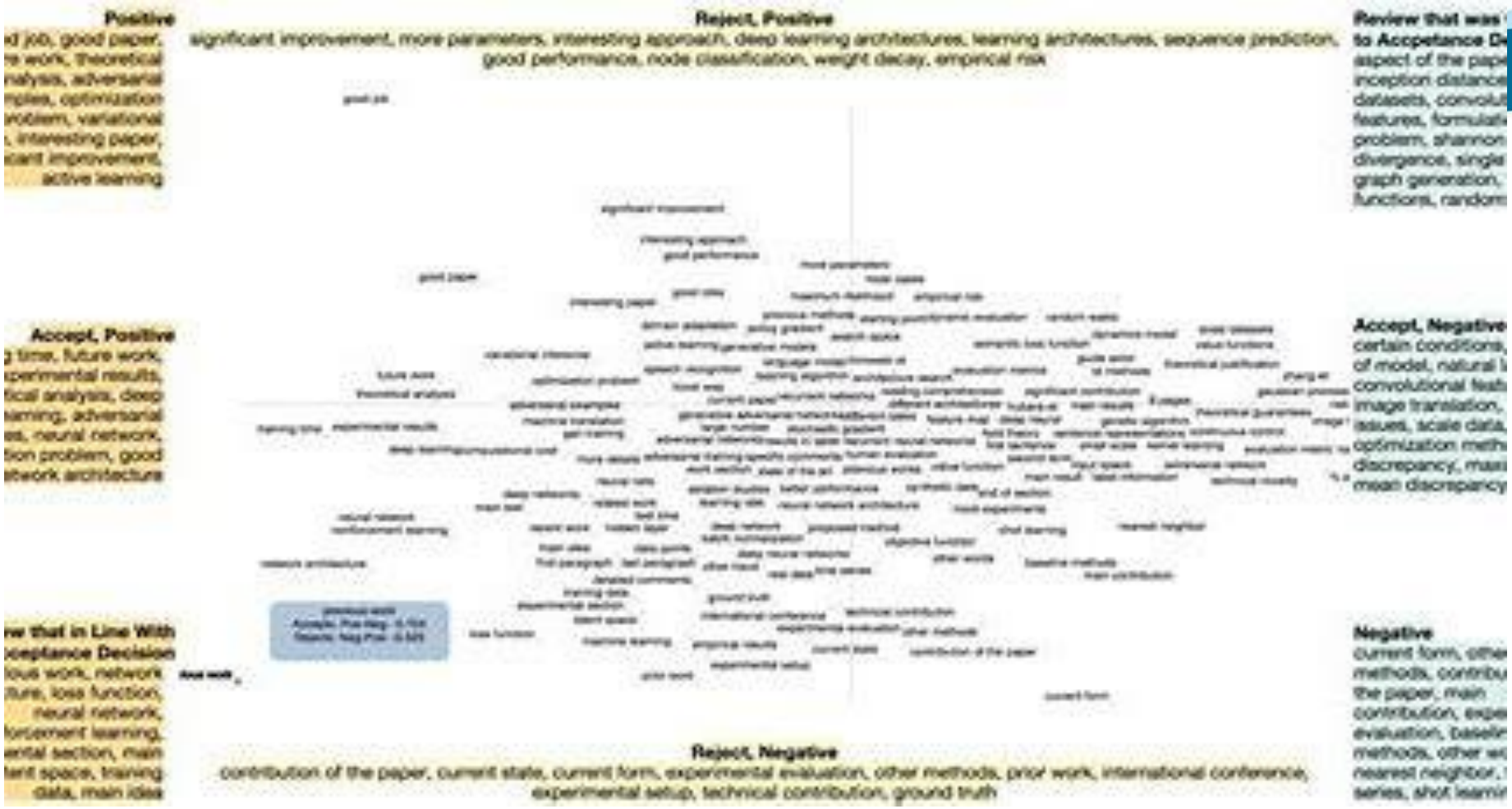


Image source: [P] How to write a persuasive ICLR review: text mining the OpenReview dataset : MachineLearning (reddit.com)

TEXT MINING DATASET

	proverbi	candi	abliti	milk	edit	book	uncompl	ferrer	intro	kid	...	dad	why-did-the-illustrator-make-that-choic	pavilion	open-mind	doctrin	intimid	self-absorb	misunderstood	collid	brule
0	0.0	0.0	0.0	0.0	0.0	0.072099	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.265174	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.282634	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.132491	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.213154	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0.0	0.0	0.0	0.0	0.0	0.457021	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
996	0.0	0.0	0.0	0.0	0.0	0.147632	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
997	0.0	0.0	0.0	0.0	0.0	0.060985	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
998	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
999	0.0	0.0	0.0	0.0	0.0	0.679095	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1000 rows × 11529 columns

IMAGE ANALYSIS AND PROCESSING

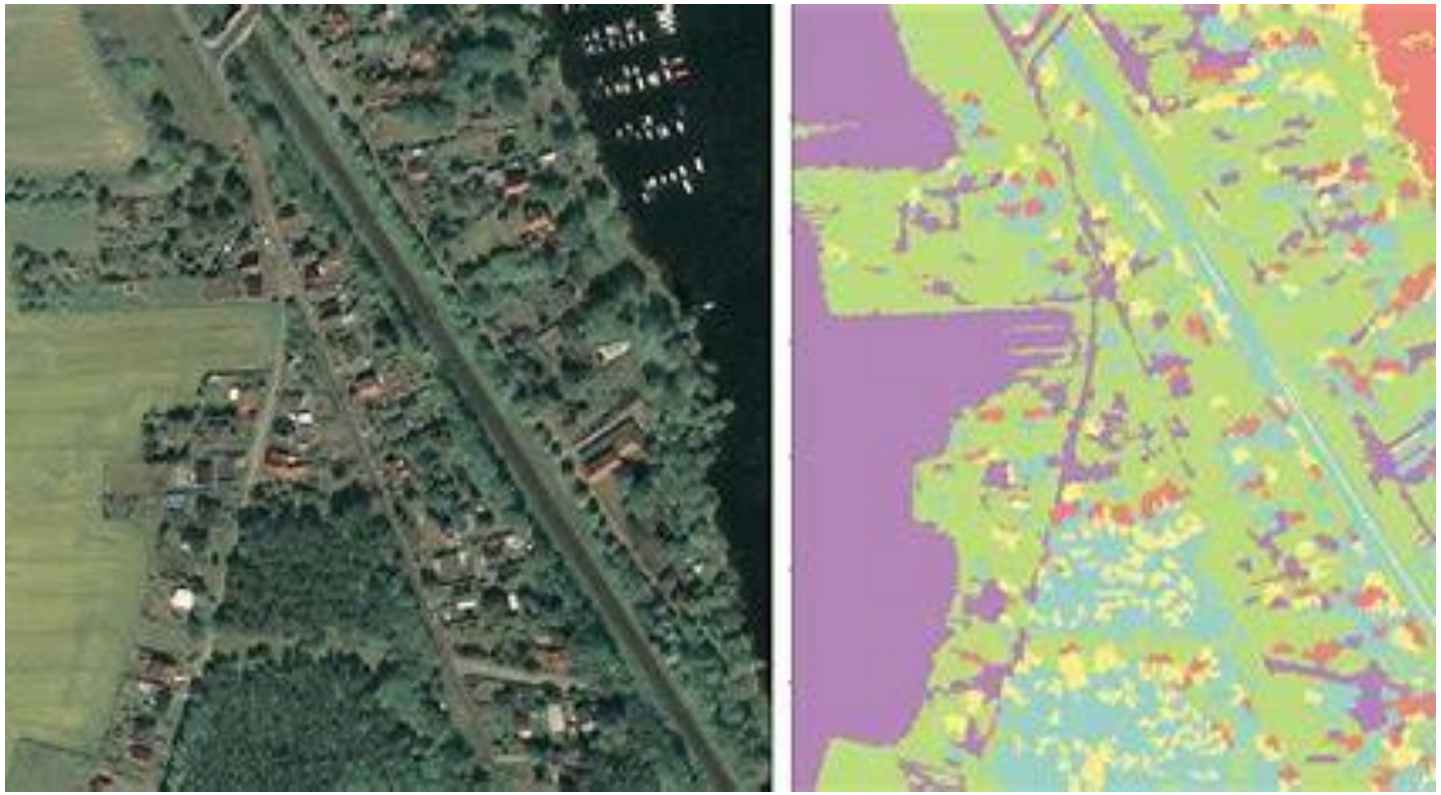


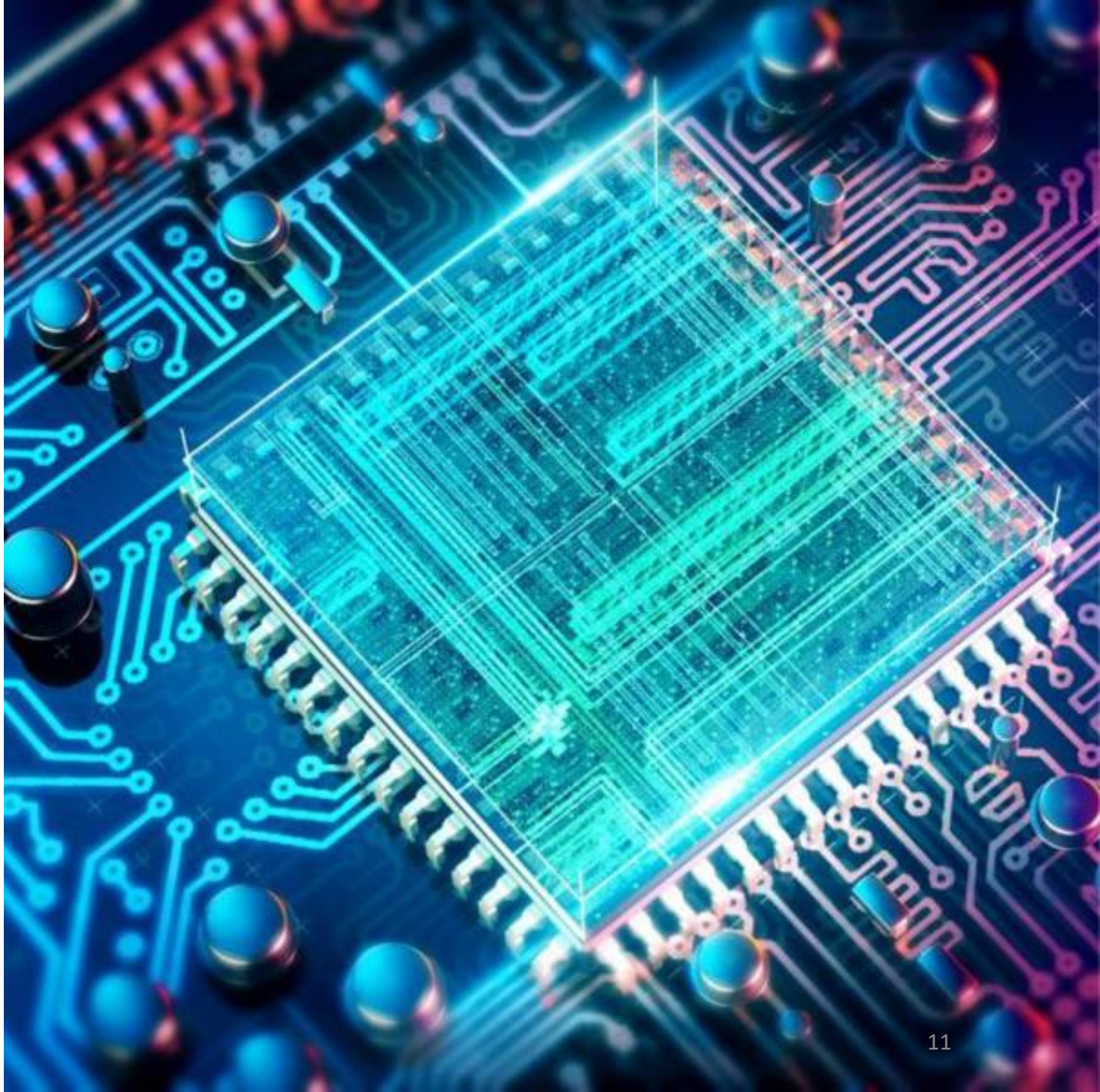
Image source: Satellite Image Processing Made Simple · UP42

WHAT DOES IT HAPPEN IF «n» IS LARGER THAN «m» ?

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

1) COMPUTATIONAL COST

SCALE-IN / SCALE-OUT



2) HIGH RISK OF OVERFITTING

IMAGE SOURCE: Overfitting, bias-variance and learning curves - rmartinshort (jimdofree.com)

3) LEAST SQUARE LINEAR REGRESSION FORMULA

$$\hat{B} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

Increasing the number of dimensions $\mapsto$ $det\left(X^T \cdot X\right) \to 0$

4) COURSE OF DIMENSIONALITY

Consider a point X_q in an n dimensional space

Consider a hypercube of edge $d < 1$ and thus volume $V < 1$

$$d^n = V \rightarrow V^{\frac{1}{n}} = d$$

Define K as the number of neighbors of X_q inside in the hypercube. For a fixed N (number of points in the space) and $V < 1$:

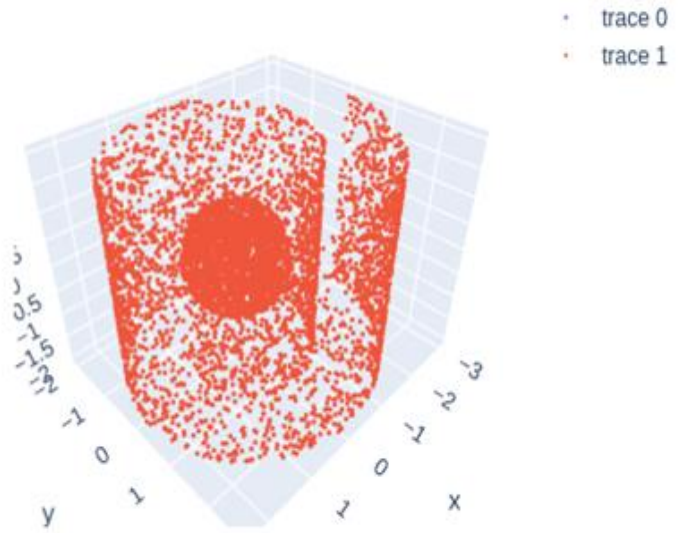
As $n \longrightarrow \infty$ then $K \longrightarrow 0$:

$$K = N \cdot V = N d^n$$

INTRINSIC DIMENSION (ID)

HISTORY OF ID ESTIMATION

- Manifolds: **a generalization and abstraction of the notion of a curved surface** **(from Britannica)**
- Notion of fractal
- **the minimal and accurate representation of dataset.**
- Generally 2 types of methods:
  - the **projection method** type:
    - Project the data into a new space with a lower dimension
    - eigenvalue methods
    - Example: PCA,
  - the **geometric method** type:
    - it is based on the nearest neighbor (NN) distance or fractal dimensions
    - Example: NN, MLE



History of ID Estimation
- some Algorithms

- Johnson Lidenstrauss projection method
- Gaussian Random projection
- MLE algorithm
- PCA algorithm

History of ID Estimation - some algorithms: Johnson Lidenstrauss

- low-distortion embeddings of points from high-dimensional into low-dimensional Euclidean space
- Dimension k
- failure probability of δ

$$k \geq \frac{4+2\delta}{\epsilon^2/2+\epsilon^3/3} \ln n$$

Let $X_1, \ldots, X_d$ be d independent Gaussian $N(0, 1)$ random variables, and let $Y = \frac{1}{\|X\|}(X_1, \ldots, X_d)$. It is easy to see that Y is a point chosen uniformly at random from the surface of the d-dimensional sphere S^{d-1}. Let the vector $Z \in \mathbf{R}^k$ be the projection of Y onto its first k coordinates, and let $L = \|Z\|^2$. Clearly the expected squared length of Z

History of ID Estimation - some Algorithms: Gaussian Random projection

- Matrix dimension as parameter
- Gaussian random method projects the original input space on a randomly generated matrix to reduce dimensions
- No relationship with the dataset different values
- The distribution

$$N\left(0, \frac{1}{n_{\text{components}}}\right)$$

History of ID Estimation - some Algorithms: MLE

- Maximum Likelihood Estimator
- ID: m
- Book link: https://papers.nips.cc/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf

MLE of intrinsic dimension

The maximum likelihood estimator in general belongs to the geometric methods.

The key assumption for deriving the maximum likelihood estimator of intrinsic dimension is that: For every point $y \in \mathbb{R}^m$, there exists a small sphere $S_y(R)$ of radius R around y, such that the density $f(y) \approx const$ within the sphere (i.e. the data points are approximately uniform in each small local region).

For the fixed point $x = \psi(y)$, consider the spatial point process (for more math details about spatial point process, see [2]) $\{N(t,x), 0 \leq t \leq R\}$,

$$N(t,x) = \sum_{i=1}^{n} I(\mathbf{x}_i \in S_x(t))$$

which counts observations within distance t from point x.

With the key assumption and recall the fact that we assume the mapping ψ is local isometric, we can approximate $N(t,x)$ by a poisson process. For the fixed x, ignore the dependence of $N(t,x)$ on x for now, the rate $\lambda(t)$ of $N(t)$:

$$\lambda(t) = f(y)V(m)mt^{m-1}$$

Source: wiki.math.uwaterloo.ca

where $T_j(x)$ is the Euclidean distance from the fixed point x to its j-th nearest neighbor in the sample.

In practice, it can be also expressed in term of the number of neighbors k rather than the radius of the sphere R:

$$\widehat{m}_k(x) = \left[\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(x)}{T_j(x)}\right]^{-1}$$

History of ID Estimation - some Algorithms: PCA

- Use of eigenvalues
- The intrinsic dimension is determined by the number of eigenvalues that are greater than a given threshold.
- PCA can be used to estimate the dimensionality of a globally linear subspace
- PCA aims to find the directions of maximum variance in high-dimensional data and projects them onto a new subspace with equal or fewer dimensions than the original one.

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_d], \quad \boldsymbol{x} \in \mathbb{R}^d$$

$$\downarrow \boldsymbol{x}\boldsymbol{W}, \quad \boldsymbol{W} \in \mathbb{R}^{d \times k}$$

$$\boldsymbol{z} = [z_1, z_2, \ldots, z_k], \quad \boldsymbol{z} \in \mathbb{R}^k$$

PROPOSED APPROACH:

TWO-NN ID ESTIMATOR

PROBLEMS RELATED TO ID ESTIMATION

- EFFECT OF CURVATURE
- DENSITY VARIATION

ADVANTAGES OF THE TWO-NN APPROACH

- MINIMAL INFORMATION USED IN THE ID ESTIMATION RESULTING IN:

  - REDUCED EFFECT OF CURVATURE

  - REDUCED DENSITY VARIATION

MATHEMATICAL ANALYSIS

Definitions:

- d is the intrinsic dimension
- ℓ labels data points in the dataset
- r_j is the distance between ℓ and its j-th nearest neighbor ($r_0 = 0$)
- ω_d is the volume of a d-sphere with unit radius
- $\Delta v_\ell \equiv \omega_d \left(r_\ell^d - r_{\ell-1}^d \right)$ is the volume of the hyperspherical shell enclosed between two successive neighbors ℓ and $\ell - 1$

A. Hypothesis: The density is approximately constant on the lengthscale defined by the typical distance to the second neighbor

B. Thesis: The cumulative distribution of the ratio of the second distance to the first one is assessable and it is a function of $\hat{d}$, and so it is possible to compute $\hat{d}$:

$$\hat{d} = \frac{-log(1-F(\mu))}{log(\mu)}$$

Considering the set of Δv_ℓ for $\ell = 1, \ldots, K$, it can be proved that, if the density around point ℓ is constant, ρ, then all Δv_ℓ are independently drawn from an exponential distribution with expected value $1/\rho$:

$$P\Big(\Delta v_\ell \in \big[v, v + dv\big]\Big) = \rho\, \mathrm{e}^{-\rho v}\, dv$$

Let $R = \frac{\Delta v_i}{\Delta v_j}$, from the previous consideration it is possible to compute exactly the probability distribution of R:

$$P\Big(R \in \big[\bar{R}, \bar{R}+d\bar{R}\big]\Big) = \int_{\bar{R} \leq \frac{v_i}{v_j} \leq \bar{R}+d\bar{R}} dv_i\, dv_j\, \rho\, \mathrm{e}^{-\rho v_i}\ \rho\, \mathrm{e}^{-\rho v_j} = \frac{d\bar{R}}{\big(1+\bar{R}\big)^2}$$

thus

$$g(R) = \frac{1}{\big(1+R\big)^2}$$

The distribution does not depend explicitly on the dimension d, but, defining $\mu = \frac{r_2}{r_1} \in [1, \infty]$, μ and R are related by:

$$R = \mu^d - 1$$

and so we can exactly compute the probability distribution and the cumulative distribution of μ.
Probability distribution:

$$f(\mu) = \frac{d}{\mu^{d+1}} \, 1_{\mu \in [1,\infty]}$$

Cumulative distribution

$$F(\mu) = \int_1^{\infty} f(\mu)d\mu = (1 - \mu^{-d})1_{\mu[1,+\infty]}$$

It can be seen that $F(\mu)$ and $f(\mu)$ depend on d but not on the local density.

From the cumulative distribution formula we can derive $\hat{d}$:

$$F(\mu) = (1 - \mu^{-d}) \rightarrow \mu^{-d} = 1 - F(\mu) \rightarrow -d = log_{\mu}(1 - F(\mu))) \rightarrow d = \frac{-log(1-F(\mu))}{log(\mu)}$$

IMPLEMENTATION

PSEUDO-CODE

1. Compute the pairwise distances for each point in the dataset $i = 1, \ldots, N$.
2. For each point i find the two shortest distances r_1 and r_2.
3. For each point i compute $\mu_i = \frac{r_2}{r_1}$.
4. Compute the empirical cumulate $F^{emp}(\mu)$ by sorting the values of μ in an ascending order through a permutation σ, then define $F^{emp}(\mu_{\sigma(i)}) \doteq \frac{i}{N}$.
5. Fit the points of the plane given by coordinates $\{(log(\mu_i), -log(1 - F^{emp}(\mu_i)))|i = 1, \ldots, N\}$ with a straight line passing through the origin.

Implementation procedure

A. Compute the pairwise distances dataframe.

A. Iterate the pairwise distances dataframe to find r1 and r2 for each point and return a list of mu values.

A. Creation of a function that given a list of mu values and a specific mu return the empirical cumulative probability.

A. Fit the points on a plane using the coordinates and do a linear regression with intercept=0.

A. return the slope of the line computed by the linear regression.

RUNTIME ANALYSIS

PSEUDO-CODE

1. Compute the pairwise distances for each point in the dataset $i = 1, \ldots, N$.
2. For each point i find the two shortest distances r_1 and r_2.
3. For each point i compute $\mu_i = \frac{r_2}{r_1}$.
4. Compute the empirical cumulate $F^{emp}(\mu)$ by sorting the values of μ in an ascending order through a permutation σ, then define $F^{emp}(\mu_{\sigma(i)}) \doteq \frac{i}{N}$.
5. Fit the points of the plane given by coordinates $\{(log(\mu_i), -log(1 - F^{emp}(\mu_i)))|i = 1, \ldots, N\}$ with a straight line passing through the origin.

IMAGE REFERENCE: https://www.nature.com/articles/s41598-017-11873-y.pdf

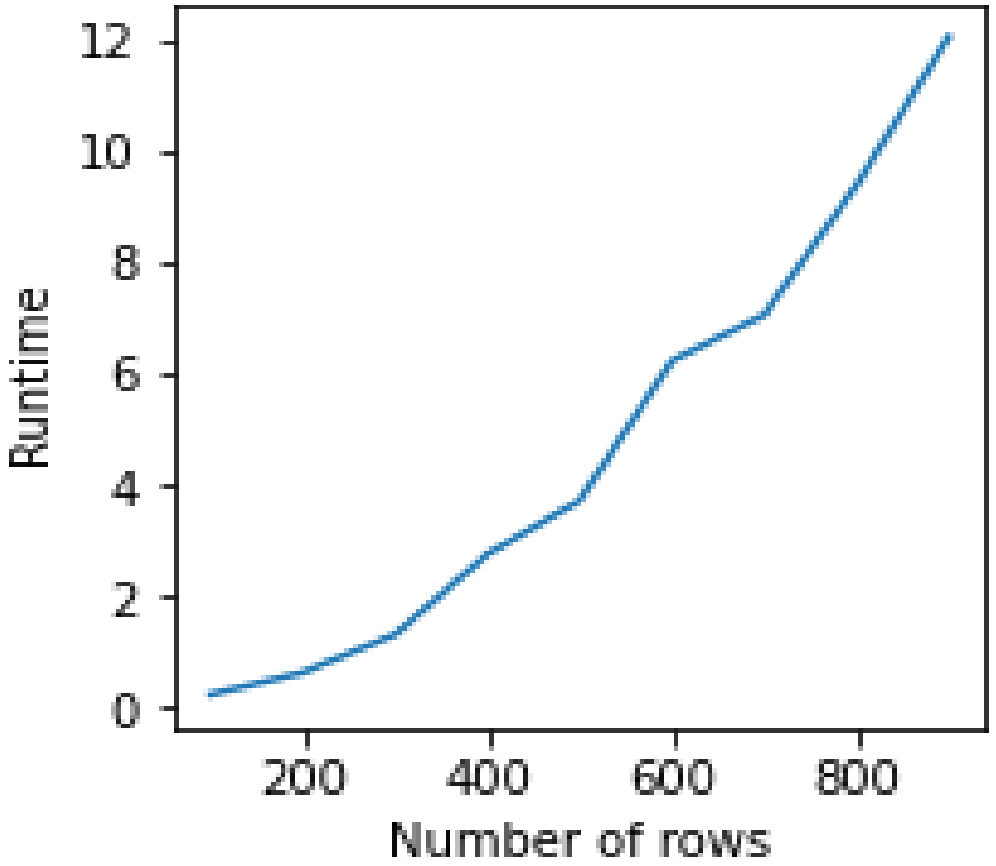
- 1) $O(n^2)$
- 2) $O(n^2)$
- 3) $O(n)$
- 4) $O(n)$
- 5) $O(2n + 4n + 8 + 2n + 4) + O(n)$

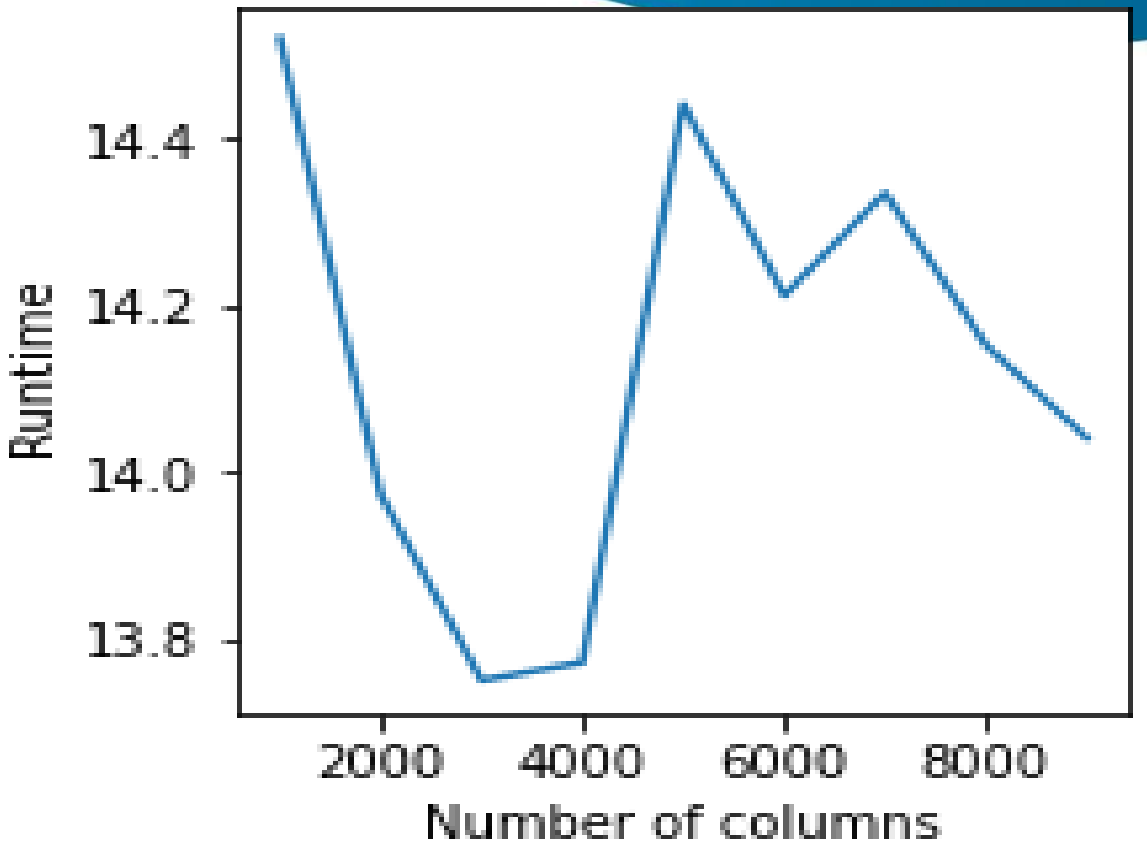
$\Rightarrow$ $O(n^2)$

RUNTIME FOR LINEAR REGRESSION HAVING X (n x m) and Y (n x 1)

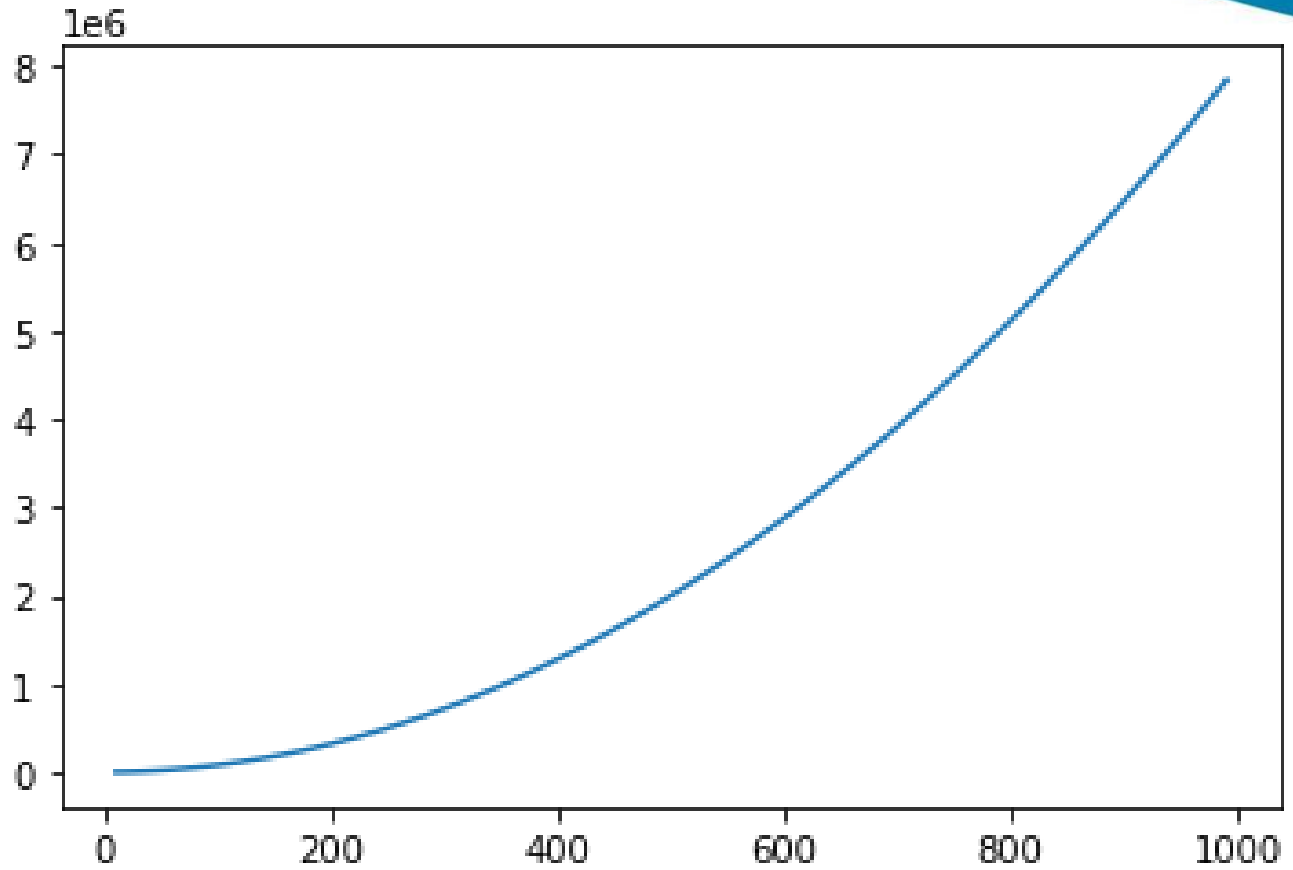
$$O(n \cdot m + n \cdot m^2 + m^3 + n \cdot m + m^2)$$

EMPIRICAL RUNTIME





STORAGE ANALYSIS



Evaluation of the different algorithm with datasets having known ID

The dataset

- Origin: **scikit-dimension package**

Shape of our truth table (24, 4)

	dataset_type	Intrinsic Dimension	Number of variables	Description
0	M1_Sphere	10	11	10D sphere linearly embedded
1	M2_Affine_3to5	3	5	Affine space
2	M3_Nonlinear_4to6	4	6	Concentrated figure, mistakable with a 3D one
3	M4_Nonlinear	4	8	Nonlinear manifold
4	M5a_Helix1d	1	3	1D helix
5	M5b_Helix2d	2	3	2D helix
6	M6_Nonlinear	6	36	Nonlinear manifold
7	M7_Roll	2	3	Swiss Roll
8	M8_Nonlinear	12	72	Nonlinear (highly curved) manifold
9	M9_Affine	20	20	Affine space
10	M10a_Cubic	10	11	10D hypercube
11	M10b_Cubic	17	18	17D hypercube
12	M10c_Cubic	24	25	24D hypercube
13	M10d_Cubic	70	71	70D hypercube
14	M11_Moebius	2	3	Möebius band 10-times twisted
15	M12_Norm	20	20	Isotropic multivariate Gaussian
16	M13a_Scurve	2	3	2D S-curve
17	M13b_Spiral	1	13	1D helix curve
18	Mbeta	10	40	Manifold generated with a smooth nonuniform pd...
19	Mn1_Nonlinear	18	72	Nonlinearly embedded manifold of high ID (see ...

Evaluation Process

- Use the default dataset
- For each manifold apply the presented algorithms to estimate the ID:
  - Johnson Lidenstrauss projection method
  - Gaussian Random projection
  - MLE algorithm
  - PCA algorithm
  - 2NN ID estimator
- Determine the square distance error: (real_ID - calculate_ID)^2
- For each manifold select the method having the lowest error

Evaluation of the ID

	dataset_type	Intrinsic Dimension	pca_id	pca_time	pca_error
0	M1_Sphere	10	9.9604	1.506254	1.568160e-03
1	M2_Affine_3to5	3	3.0000	0.816053	0.000000e+00
2	M3_Nonlinear_4to6	4	4.6932	0.949426	4.805262e-01
3	M4_Nonlinear	4	5.1956	1.159452	1.429459e+00
4	M5a_Helix1d	1	1.0000	0.761038	0.000000e+00
5	M5b_Helix2d	2	2.9464	1.137240	8.956730e-01
6	M6_Nonlinear	6	8.9308	3.432740	8.589589e+00
7	M7_Roll	2	2.0000	1.623672	0.000000e+00
8	M8_Nonlinear	12	15.8224	4.701817	1.461074e+01

	dataset_type	Intrinsic Dimension	2nn_id	2nn_time	2nn_error
0	M1_Sphere	10	9.405339	77.774151	0.353621
1	M2_Affine_3to5	3	2.959170	88.092001	0.001667
2	M3_Nonlinear_4to6	4	3.791889	84.511519	0.043310
3	M4_Nonlinear	4	3.689355	117.228837	0.096500
4	M5a_Helix1d	1	1.009278	89.975596	0.000086
5	M5b_Helix2d	2	1.982765	78.589640	0.000297
6	M6_Nonlinear	6	5.719325	79.124561	0.078779
7	M7_Roll	2	1.952463	77.934621	0.002260
8	M8_Nonlinear	12	13.472597	78.058962	2.168543

Error Analysis - Best methods

	dataset_type	Intrinsic Dimension	pca_id	pca_time	pca_error	mle_id	mle_time	mle_error	2nn_id	2nn_time	2nn_error	jl_id	jl_time	jl_error	best_id
0	M1_Sphere	10	9.9604	1.506254	0.001568	8.947787	0.219506	1.107152	9.405339	77.774151	0.353621	57	4.559997e-07	2209	skdim_pca
1	M2_Affine_3to5	3	3.0000	0.816053	0.000000	2.855454	0.057479	0.020894	2.959170	88.092001	0.001667	38	5.059992e-07	1225	skdim_pca
2	M3_Nonlinear_4to6	4	4.6932	0.949426	0.480526	3.737550	0.080906	0.068880	3.791889	84.511519	0.043310	43	4.699996e-07	1521	2nn
3	M4_Nonlinear	4	5.1956	1.159452	1.429459	3.935440	0.110207	0.004168	3.689355	117.228837	0.096500	49	2.789993e-07	2025	2nn
4	M5a_Helix1d	1	1.0000	0.761038	0.000000	1.005041	0.045372	0.000025	1.009278	89.975596	0.000086	26	2.669995e-07	625	skdim_pca

```
1 print('total value = ',new_df['best_id'].size)
2 new_df['best_id'].value_counts()
```

```
total value =  24
skdim_pca    13
2nn          11
Name: best_id, dtype: int64
```

```
[ ]  1 print('% of each method')
     2 new_df['best_id'].value_counts()/new_df['best_id'].size * 100
```

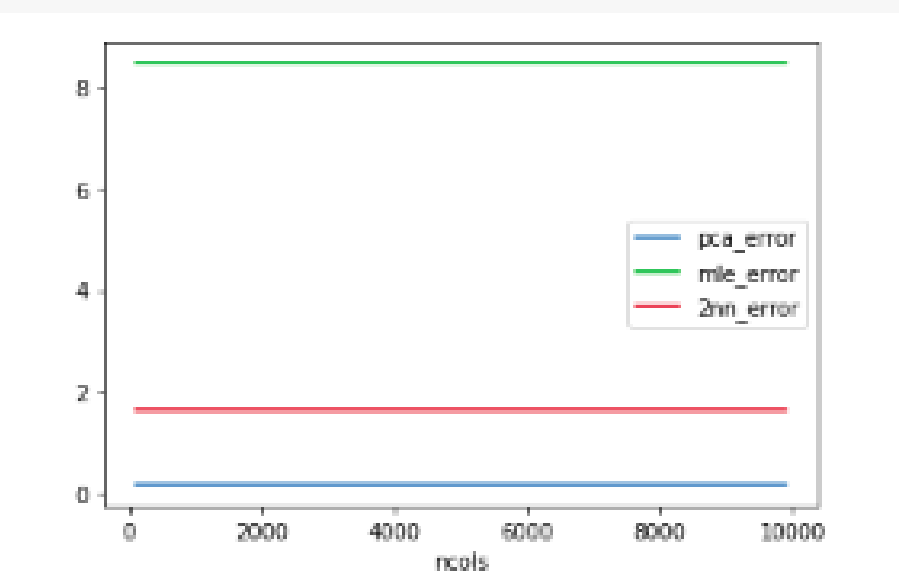
```
% of each method
skdim_pca    54.166667
2nn          45.833333
Name: best_id, dtype: float64
```

Evaluation:
Varying columns vs error

Process:

- Fixed number of rows in the dataset (case Sphere)
- Varying the column number from 100 to 10000

Comment: The number of columns or dataset features do not impact on the value of the error.

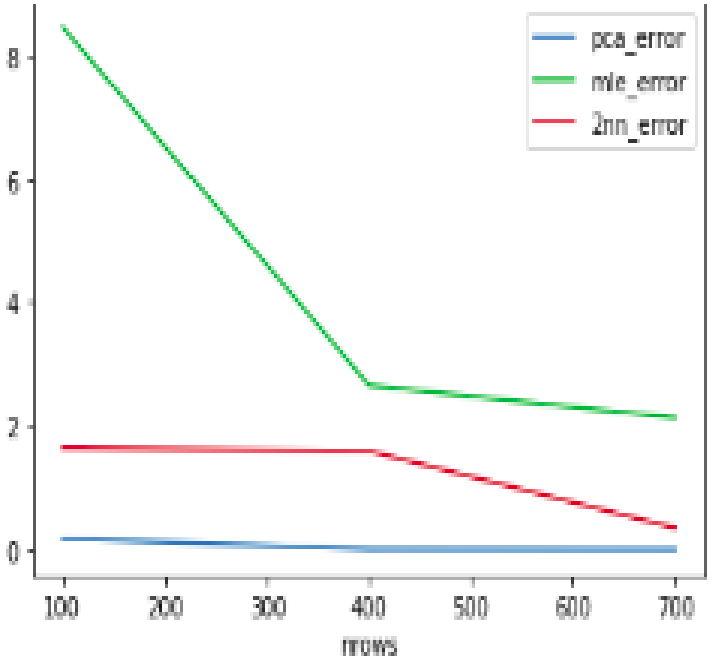


Evaluation:
Varying rows vs error

Process:

- Fixed number of columns in the dataset (case Sphere)
- Varying the number of rows from 100 to 1000

Comment: The number of rows has an impact on the error.



TEXT MINING EXAMPLE

“Represent a group of 1000 texts as an high dimensional dataset and then assess the intrinsic dimension”

GUIDELINE OF THE EXAMPLE:

- Preprocess a set of texts about positive book reviews.
  - TOKENTIZATION
  - REGULARIZATION (Deletion of punctuation)
  - DELETION OF FREQUENT ENGLISH WORDS
  - STEMMING
- Determine the dimensions (python set of all possible preprocessed words)
- Creation of a dataset where each line is a text and each column is a dimension (the value of the text components is the TF-IDF)
- Application of ID estimator to the dataset

EXAMPLE OF PREPROCESSING AND TF-IDF

Preprocessing:

```
preprocess_text("This is how the algorithm works")
```

```
['algorithm', 'work']
```

TF-IDF:

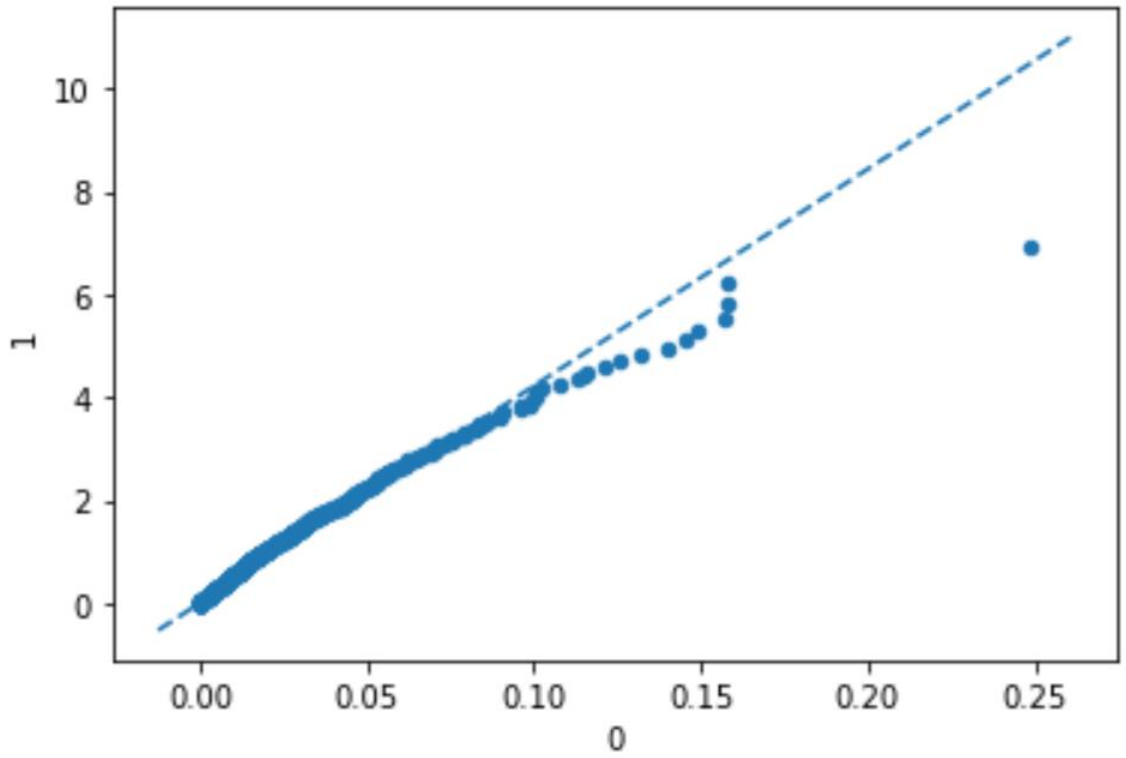
```
[('algorithm', 0.9167635831535623), ('work', 0.39943025999971693)]
```

ORIGINAL DATASET

	proverbi	candi	abliti	milk	edit	book	uncompl	ferrer	intro	kid	...	dad	why-did-the-illustrator-make-that-choic	pavilion	open-mind	doctrin	intimid	self-absorb	misunderstood	collid	brule
0	0.0	0.0	0.0	0.0	0.0	0.072099	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.265174	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.282634	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.132491	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.213154	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0.0	0.0	0.0	0.0	0.0	0.457021	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
996	0.0	0.0	0.0	0.0	0.0	0.147632	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
997	0.0	0.0	0.0	0.0	0.0	0.060985	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
998	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
999	0.0	0.0	0.0	0.0	0.0	0.679095	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1000 rows × 11529 columns

APPLICATION OF ID ESTIMATOR



THE ID RESULT IS ALMOST 42

REDUCED DATASET

	book	read	stori	like	love	great	charact	recommend	time	good	...	thing	understand	look	highli	easi	inform	review	feel	live
0	0.072099	0.071014	0.075514	0.000000	0.000000	0.000000	0.000000	0.082228	0.000000	0.0	...	0.125524	0.000000	0.0	0.000000	0.000000	0.070086	0.00000	0.0	0.064728
1	0.265174	0.130592	0.069433	0.000000	0.000000	0.058104	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
2	0.282634	0.139191	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
3	0.132491	0.000000	0.000000	0.098807	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
4	0.213154	0.069982	0.223249	0.052987	0.000000	0.062274	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.08483	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0.457021	0.225073	0.000000	0.000000	0.446518	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
996	0.147632	0.000000	0.154624	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.128513	0.149657	0.0	0.000000	0.000000	0.143511	0.00000	0.0	0.132540
997	0.060985	0.060067	0.000000	0.090960	0.059583	0.053451	0.067924	0.000000	0.086641	0.0	...	0.053087	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
998	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.095638	0.238270	0.0	...	0.072997	0.085007	0.0	0.000000	0.191918	0.000000	0.10012	0.0	0.000000
999	0.679095	0.167220	0.000000	0.000000	0.000000	0.148801	0.000000	0.193624	0.000000	0.0	...	0.000000	0.000000	0.0	0.201815	0.000000	0.000000	0.00000	0.0	0.000000

1000 rows × 42 columns

NEW DATASET DIMENSIONS

['book', 'read', 'stori', 'like', 'love', 'great', 'charact',
'recommend', 'time', 'good', 'work', 'author', 'realli', 'life',
'mani', 'make', 'novel', 'enjoy', 'reader', 'way', 'best', 'use',
'peopl', 'written', 'want', 'write', 'world', 'page', 'know', 'think',
'year', 'help', 'thing', 'understand', 'look', 'highli', 'easi',
'inform', 'review', 'feel', 'live', 'histori']

Conclusion

IMPLEMENTATION SUMMARY

1) What is the intrinsic dimension (ID) of a dataset and why is important to assess it.
2) Existing approaches
3) Proposed approach TWO-NN
4) Implementation (N.B. part 4.5. Estimation of limit ID of a fixed distribution takes about 2h 30min to run due to high number of iterations and the large number of rows (10 000). It is not necessary to run it for the rest of the project, the result will be available on the next slide.)
5) Application to a text mining example
6) EVALUATION AND ASSESSMENT OF THE ID ESTIMATE

Result of execution of section 4.5. Estimation of limit ID of a fixed distribution

As the number of points tends to infinity the estimation of the cumulative distribution of μ gets more and more accurate
($F^{emp}(\mu) \to F(\mu)$)
as well as the estimation of the intrinsic dimension. As a result the estimation of ID tends to the real value of ID.