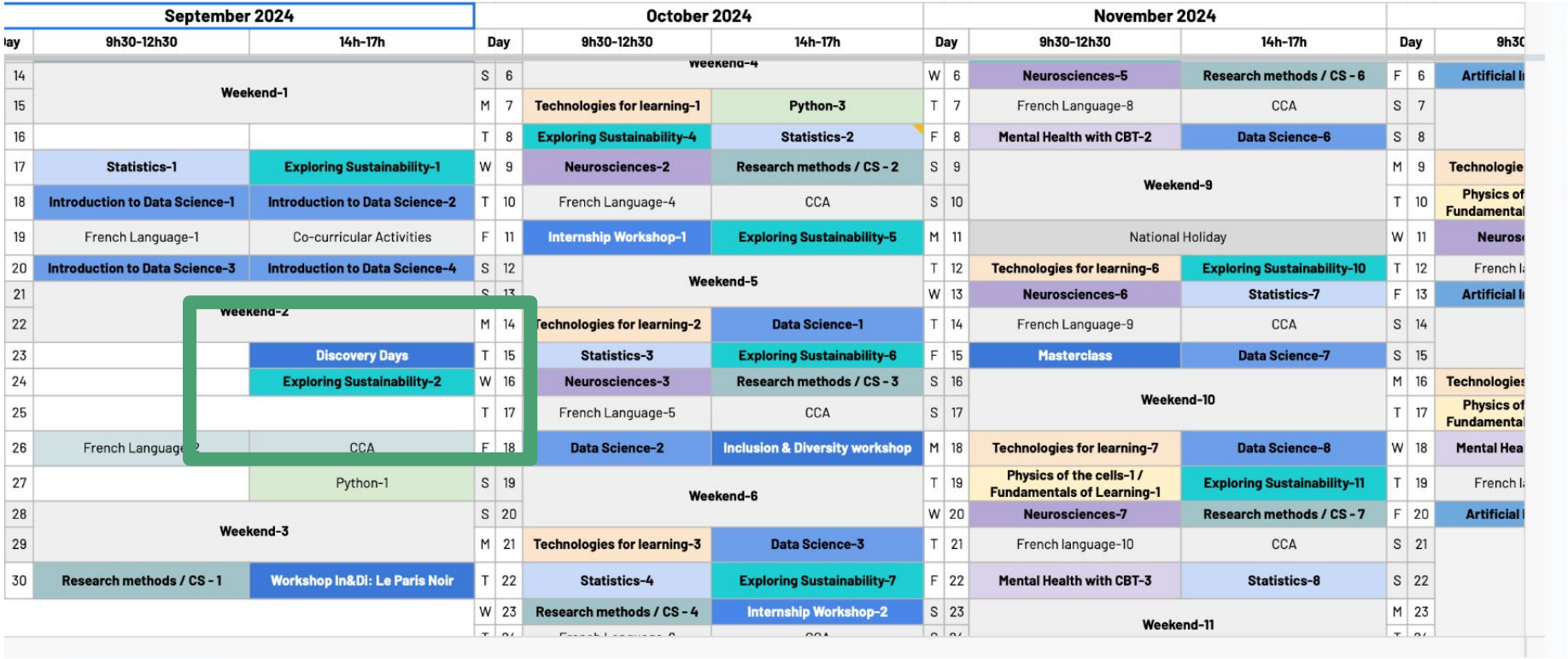
Foundations of AI

LIUBOV TUPIKINA
LPI, 2024-2025

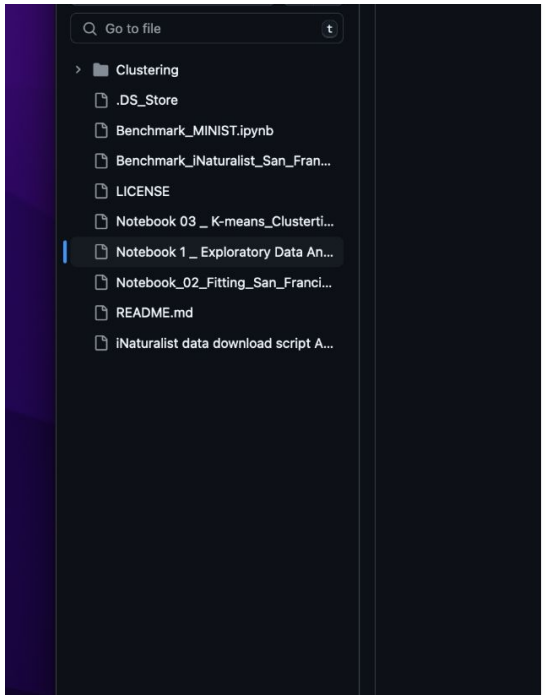
Outline and connection to other courses



September 2024				October 2024				November 2024					
ay	9h30-12h30	14h-17h	Day		9h30-12h30	14h-17h	Day		9h30-12h30	14h-17h	Day		9h30
14	Weekend-1		S	6	Weekend-4		W	6	Neurosciences-5	Research methods / CS - 6	F	6	Artificial I
15			M	7	Technologies for learning-1	Python-3	T	7	French Language-8	CCA	S	7	
16			T	8	Exploring Sustainability-4	Statistics-2	F	8	Mental Health with CBT-2	Data Science-6	S	8	
17	Statistics-1	Exploring Sustainability-1	W	9	Neurosciences-2	Research methods / CS - 2	S	9	Weekend-9		M	9	Technologie
18	Introduction to Data Science-1	Introduction to Data Science-2	T	10	French Language-4	CCA	S	10			T	10	Physics of Fundamental
19	French Language-1	Co-curricular Activities	F	11	Internship Workshop-1	Exploring Sustainability-5	M	11	National Holiday		W	11	Neuros
20	Introduction to Data Science-3	Introduction to Data Science-4	S	12	Weekend-5		T	12	Technologies for learning-6	Exploring Sustainability-10	T	12	French l
21	Weekend-2		S	13			W	13	Neurosciences-6	Statistics-7	F	13	Artificial I
22			M	14	Technologies for learning-2	Data Science-1	T	14	French Language-9	CCA	S	14	
23		Discovery Days	T	15	Statistics-3	Exploring Sustainability-6	F	15	Masterclass	Data Science-7	S	15	
24		Exploring Sustainability-2	W	16	Neurosciences-3	Research methods / CS - 3	S	16	Weekend-10		M	16	Technologies
25			T	17	French Language-5	CCA	S	17			T	17	Physics of Fundamental
26	French Language-2	CCA	F	18	Data Science-2	Inclusion & Diversity workshop	M	18	Technologies for learning-7	Data Science-8	W	18	Mental Hea
27		Python-1	S	19	Weekend-6		T	19	Physics of the cells-1 / Fundamentals of Learning-1	Exploring Sustainability-11	T	19	French l
28	Weekend-3		S	20			W	20	Neurosciences-7	Research methods / CS - 7	F	20	Artificial I
29			M	21	Technologies for learning-3	Data Science-3	T	21	French language-10	CCA	S	21	
30	Research methods / CS - 1	Workshop In&Di: Le Paris Noir	T	22	Statistics-4	Exploring Sustainability-7	F	22	Mental Health with CBT-3	Statistics-8	S	22	
			W	23	Research methods / CS - 4	Internship Workshop-2	S	23	Weekend-11		M	23	

Example of sustainability related projects

Github [notebook]




```
Loading:  Francisco_Bay 2021
Loading:  Francisco_Bay 2022
Loading:  Francisco_Bay 2023
```

Out[4]:

	id	observed_on_string	observed_on	time_observed_at	created_time_zone	created_at	
0	20069	1:15 pm.	2016-07-14	2016-07-14T13:15:00-07:00	America/ Los_Angeles	2011-06-03T14:51:45-07:00	2020-0
1	20070	1:00 pm.	2016-03-25	2016-03-25T13:00:00-07:00	America/ Los_Angeles	2011-06-03T14:53:13-07:00	2020-0
2	68373	6:30	2016-02-12	2016-02-12T06:30:00-08:00	America/ Los_Angeles	2012-04-20T20:36:48-07:00	2020-(
3	158736	2:19	2016-10-14	2016-10-14T14:19:00-07:00	America/ Los_Angeles	2012-12-06T20:23:52-08:00	2016-1
4	538018	2016-04-10 2:20:00 PM PDT	2016-04-10	2016-04-10T14:20:00-07:00	America/ Los_Angeles	2014-02-20T15:40:40-08:00	2016-0

5 rows × 38 columns

Dataset comprises of 4214727 observations and 38 characteristics.

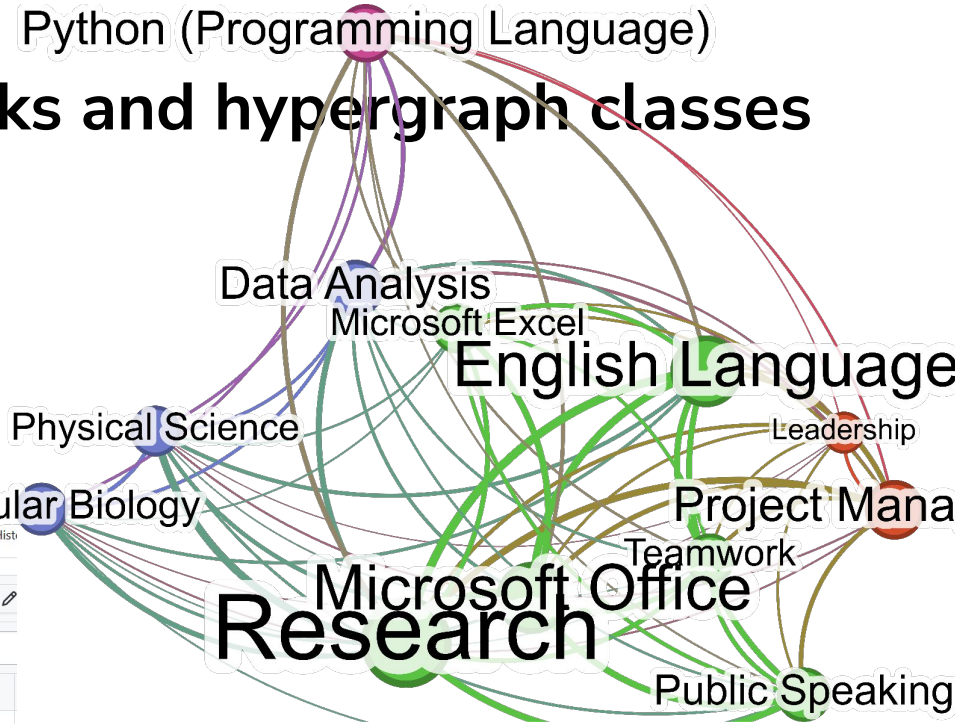
In [10]:
```
dfall.shape
```

Out[10]:
```
(4214727, 38)
```

Some reminder from networks and hypergraph classes

Process data: what is important to do with data?

Github of the course



ziqingcheryl organized_version 688a7a3 · 9 months ago

Preview | Code | Blame | 8 lines (8 loc) · 56.3 KB | Raw

Search this file

		Name	Occupation	City	State	Country
2	1	7f87716741aeae4d227491b5a7d87d4e	Stagiaire at INSTITUT GUSTAVE ROUSSY			France
3	2	c10674167315089247ea5fa8c98256f6	Initiatrice de projet at Tous Tes Possibles	Paris	Île-de-France	France
4	3	bf83d3cbf7ebfeeecdedd9519df56af8	PhD Student at Medical University of Vienna	Österreich		Austria
5	4	94369f5f833008a9b93d6e9ae7a6a533	Chargée de communication grand public et jeunes at ADEME	Paris	Île-de-France	France
6	5	67af1831de65104374b77b9a597f4671	Stagiaire UX/ Product Owner at Tyllt	Talence	Nouvelle-Aquitaine	France
7	6	95dcb67fb5d78d0c099249d553343451	Research Associate at King's College London			United Kingdom

Orientation every class

Outline of the course

Syllabus and Agenda:

18th September:

- **Morning**: Elements of statistics for data analysis: building intuition with a dataset)
- **Afternoon**: Introduction to data science, network science

20th September

- **Morning**: Foundations on AI for data science: from theory to practice on data fitting, embedding, modeling
- **Afternoon**: Spatial data analysis, Data and Network visualization

Outline of the day

1. Introduction to algorithms and foundations of AI
2. Practical part: notebooks

Some sources and additional materials (Darmouth University)

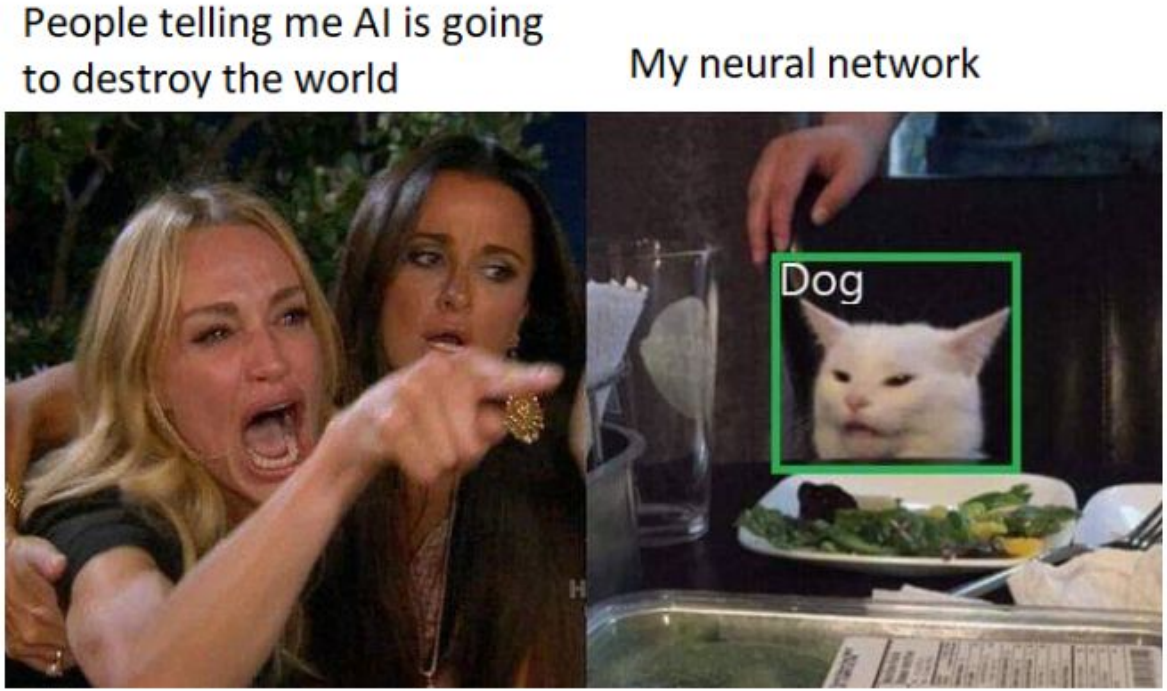
Some external resources:
Neuroscience Pasteur Institute, Marie Curie (Roberto Toro)
Computer Science (Telecom Paris, Saclay)
Network seminar Thursdays (2 weeks)
Data Science breakfast (telegram channel, Viacheslav can add:)

Artificial intelligence: Ideas & their evolution

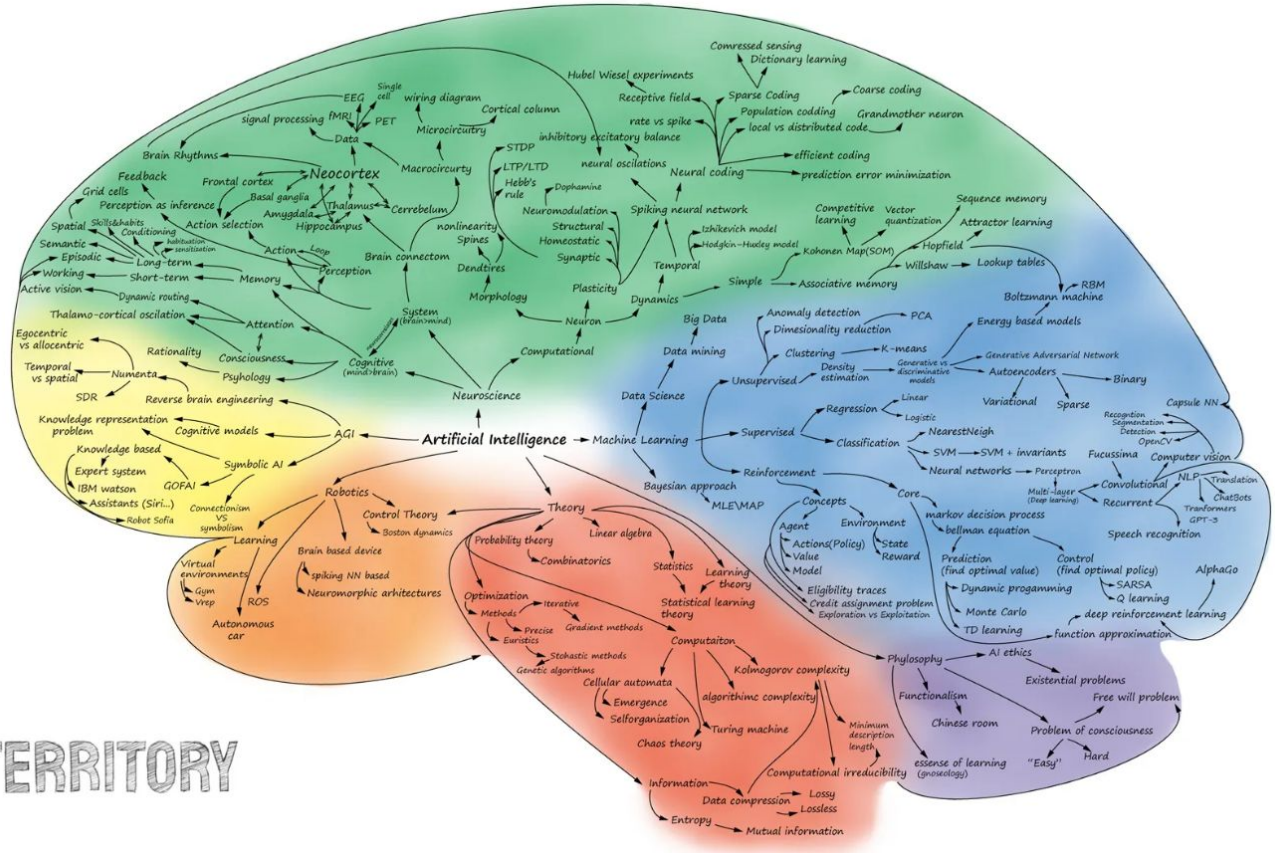
Lecture 1 of “Mathematics and AI”

AI introduction
Mathematical foundations

AI introduction
Mathematical foundations

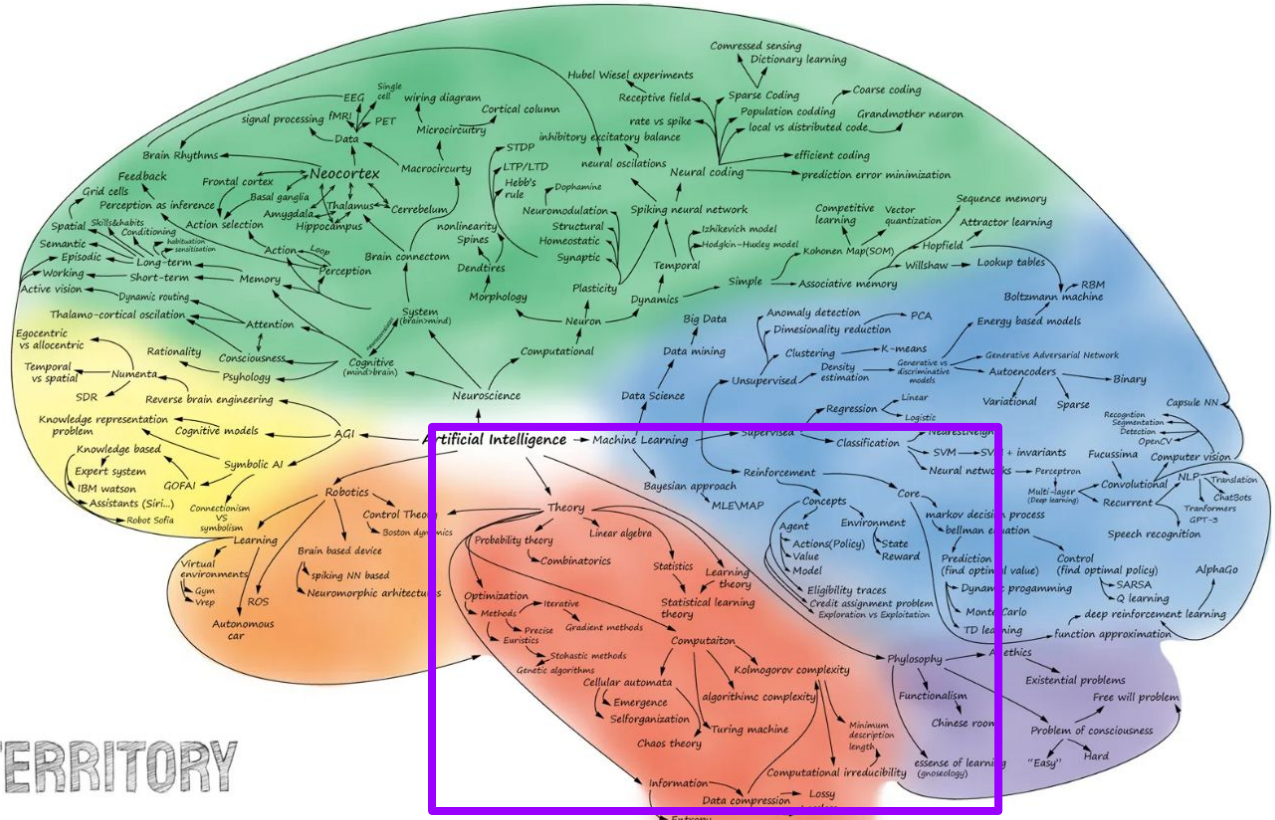


AI introduction
Mathematical foundations



AI TERRITORY

AI introduction
Mathematical foundations



AI TERRITORY

Approaches to AI systems

Approaches to AI systems

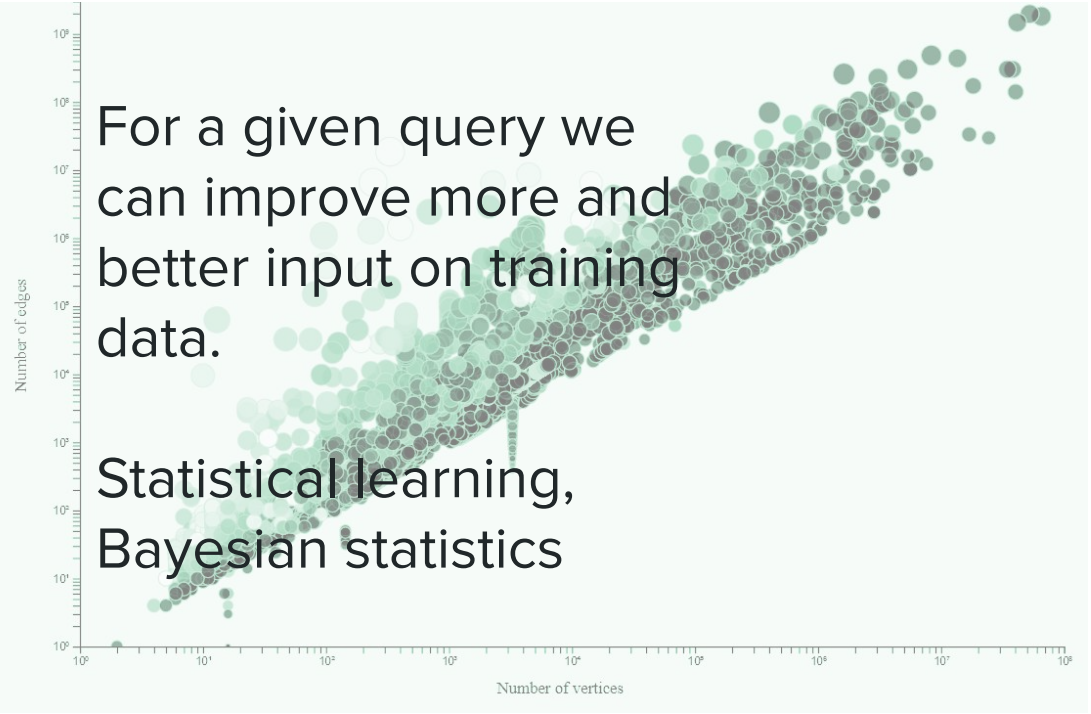
Probabilitstic, neuro-inspired approaches vs.

For a given query we can improve more and better input on training data.

Statistical learning, Bayesian statistics

Approaches to AI systems

Probabilitstic approach vs.

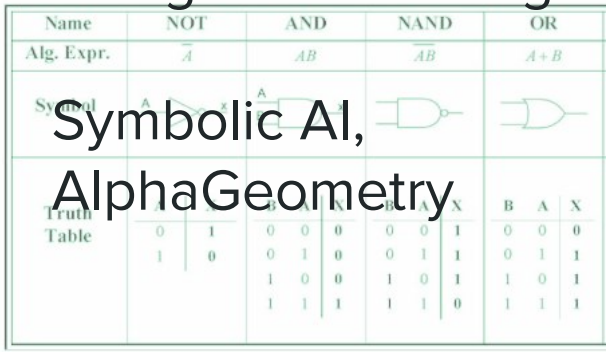


For a given query we can improve more and better input on training data.

Statistical learning, Bayesian statistics

Logical approach

If computer to give right set of axioms, rule to reason, it can learn basics of logical reasoning.

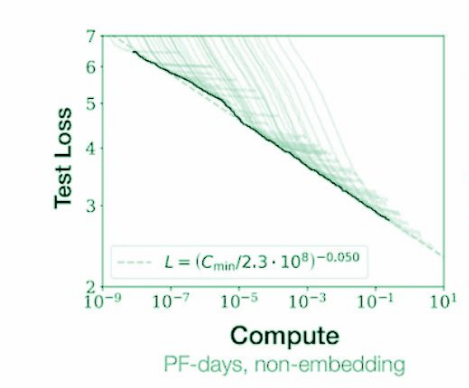


Symbolic AI, AlphaGeometry

Limits of approaches to AI systems

Probabilitstic approach

Model's quality evolves with increases in its size, training data volume, and compute

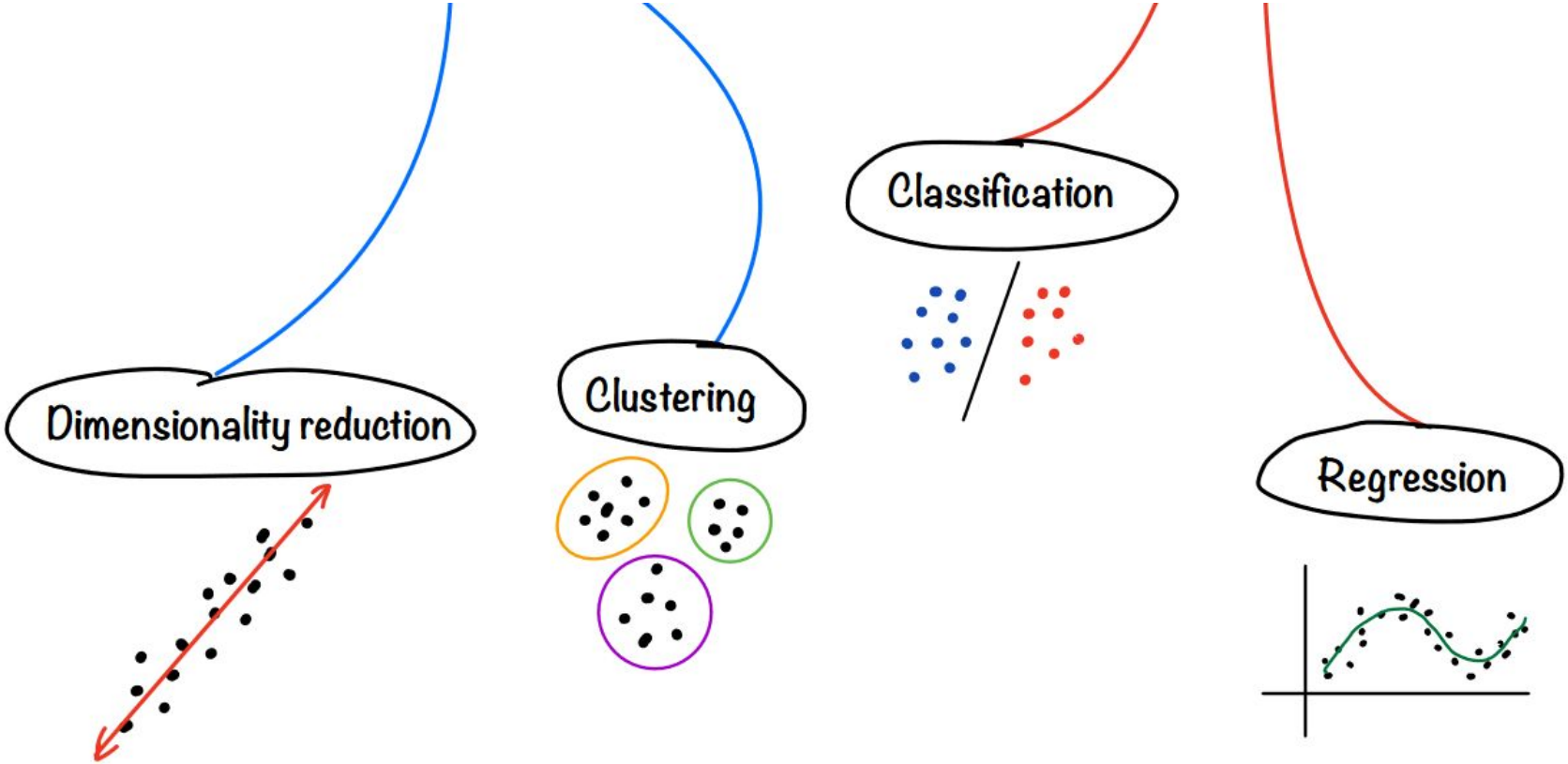


Logical approach

Goedel's theorem on limits of prove-ability of logical statements

Name	NOT	AND	NAND	OR
Alg. Expr.	$\overline{A}$	AB	$\overline{AB}$	$A+B$
Symbol				
Truth Table	A X 0 1 1 0	B A X 0 0 0 0 1 0 1 0 0 1 1 1	B A X 0 0 1 0 1 1 1 0 1 1 1 0	B A X 0 0 0 0 1 1 1 0 1 1 1 1

Main methods in machine learning



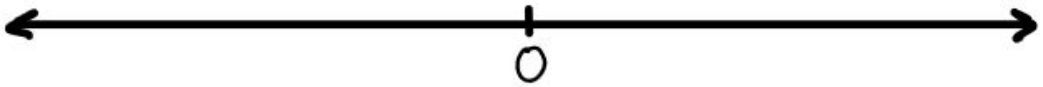
Embedding methods

What to do when you have really big data?

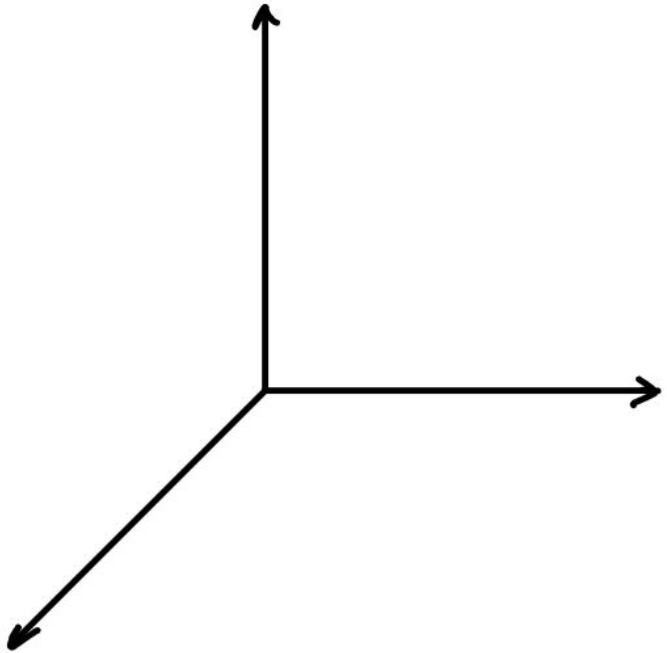
Embedding methods idea

	species_short	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

Concept of dimension in data

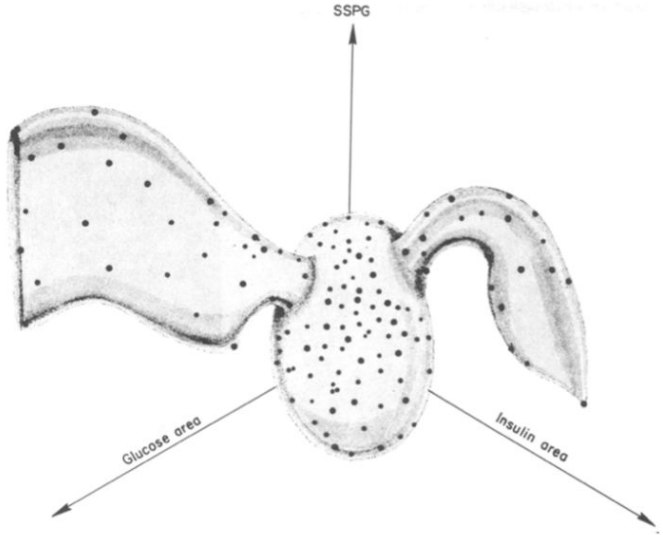


Concept of dimension in data



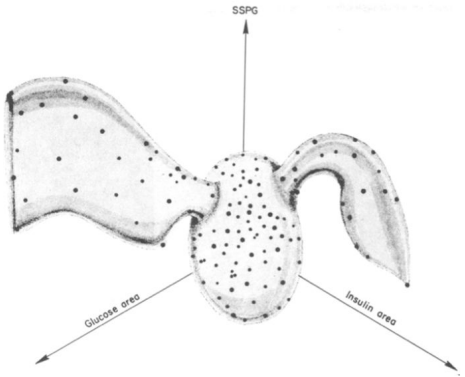
Concept of dimension in data

Example: Diabetes study
145 points in 5-dimensional space



An attempt to define the nature of chemical diabetes using a multidimensional analysis by G. M. Reaven and R. G. Miller, 1979

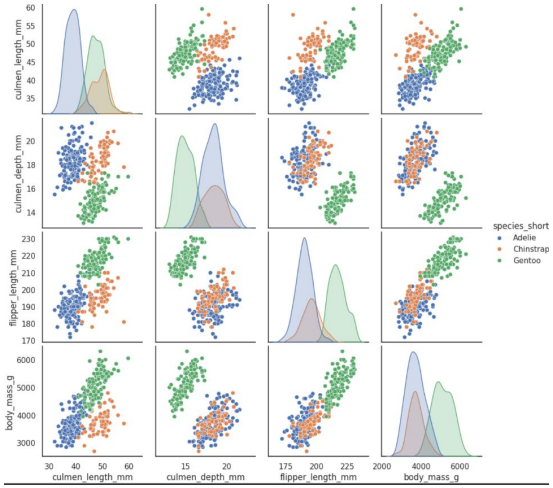
Concept of dimension in data



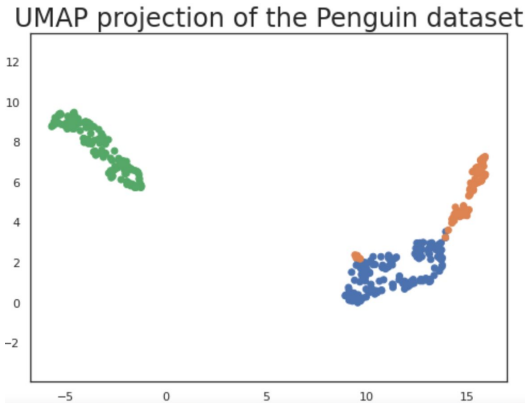
N-dimensional data
of all features

2-dimensional representation of
Important features

Main idea of the dimensionality reduction methods



N-dimensional data
of all features (not all are easy to display and analyze common patterns)



2-dimensional representation of Important features

Embedding methods idea example

Your data is very large data to work with with features

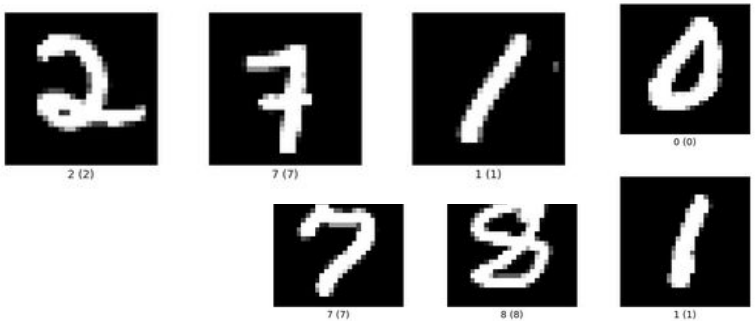
species_short	island	culmen_length_mm	culmen_depth_mm
Adelie	Torgersen	39.1	18.7
Adelie	Torgersen	39.5	17.4
Adelie	Torgersen	40.3	18.0

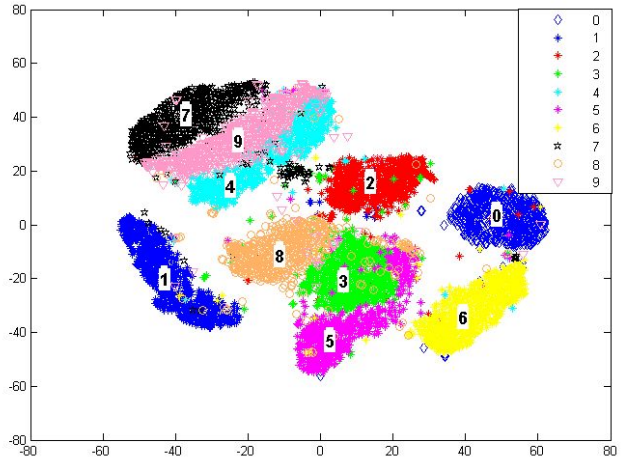
Or with some specific properties

Embedding methods idea

You want your complex data to be presented as a whole

species_short	island	culmen_length_mm	culmen_depth_mm
Adelie	Torgersen	39.1	18.7
Adelie	Torgersen	39.5	17.4
Adelie	Torgersen	40.3	18.0

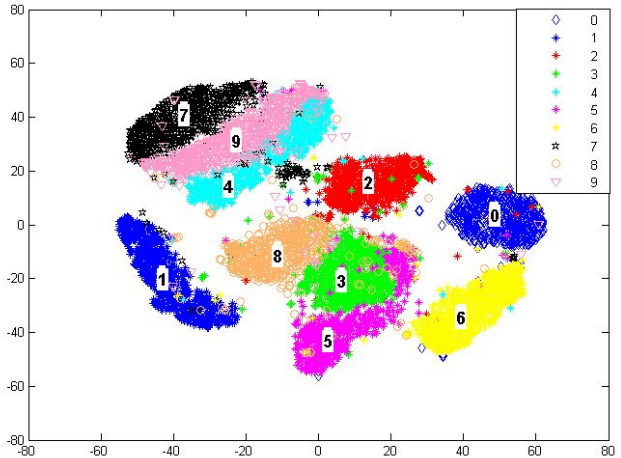




Embedding methods idea

Main idea of the algorithm:
To reduce dimensions (number of features) of your data using linear algebra transformations (PCA), distributions (tSNE), simplicial complex (UMAP)

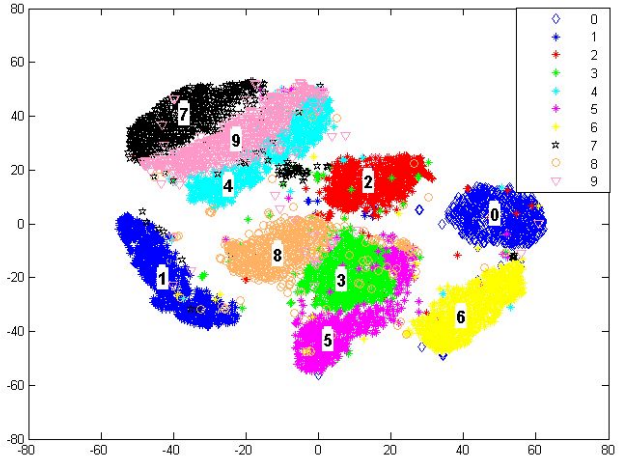
species_short	island	culmen_length_mm	culmen_depth_mm
Adelie	Torgersen	39.1	18.7
Adelie	Torgersen	39.5	17.4
Adelie	Torgersen	40.3	18.0



Embedding methods idea

[Notebook]

species_short	island	culmen_length_mm	culmen_depth_mm
Adelie	Torgersen	39.1	18.7
Adelie	Torgersen	39.5	17.4
Adelie	Torgersen	40.3	18.0

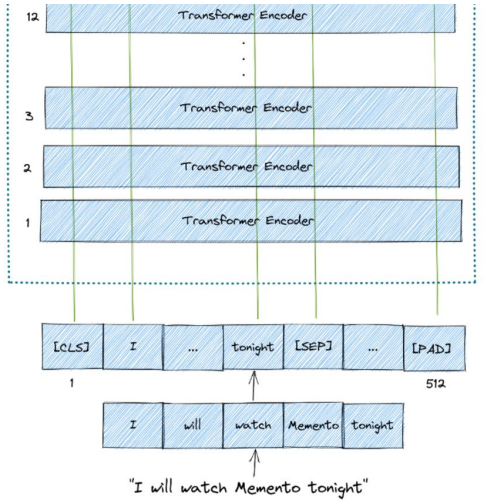


Embedding methods idea

What to do when you have some features of your data, which are not in numerical format?

Tokenisation process for your data may help BERT tokenisation

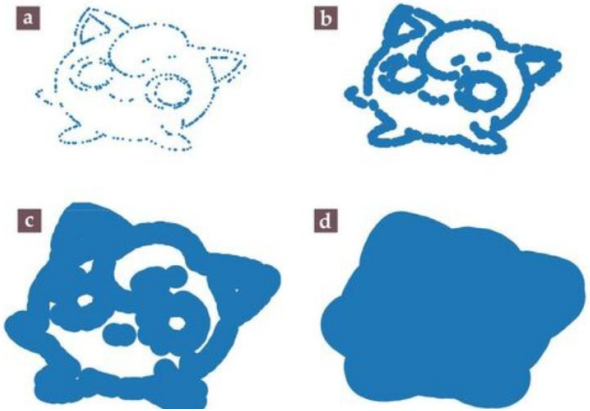
species_short	island	culmen_length_mm	culmen_depth_mm
Adelie	Torgersen	39.1	18.7
Adelie	Torgersen	39.5	17.4
Adelie	Torgersen	40.3	18.0



Geometry of data

To identify “holes” in a data set and thereby describe its topological “shape,” we need to assign a topological structure to the data and compute topological invariants. Homology groups are good invariants because there are efficient algorithms to compute them.

The most widely used tool in TDA is persistent homology (P



Key idea: two data objects are close to each other when on
obtained from the other by perturbing the other dataset
https://physicstoday.scitation.org/doi/10.1063/PT.3.5157

Types of embedding methods:

linear methods:
PCA

non-linear methods
tSNE, diffusion maps
UMAP algorithms
Node2Vec

Embedding methods

linear methods:
PCA
(fast)

non-linear methods
tSNE, diffusion maps
UMAP algorithms
Node2Vec (slower)

1. Big data visualisation (always plot your data), but what if your data is multidimensional?
Example of simple dimension reduction method: 3D data of Earth -> 2D data of coordinates

2. Idea on visualisation when we have data with N-dimensions, N>3:
Minimize sum of squares of distances on the line.
We can use two main ideas for embedding (or dimension reduction):
A. use linear methods from linear algebra: eigenvalues decomposition.
PCA idea
S = U A U ^-1
A - matrix from eigenvalues, U are eigenvectors, which are transforming the space.
KNN idea

B. use non-linear methods:
tSNE, stochastic embedding which optimise loss-functions

C. (Chakresh part)... Umap and other techniques

3. Hands-on: letting them to visualise their dataset or our dataset (for their choice)
Comparison of embeddings:
We compare embeddings using rainbow plot, for which we estimate the value of the loss function for the triplets series [Wang et al.]
For each triplet of nodes (i,j,k) the loss function is estimated as
$Loss(i,j,k) = f(d_{ij}, d_{ik})$
Comparing these rainbow plots for various embeddings we can find the most suited embedding for each dataset type.

Embedding types:

sklearn.decomposition.PCA - Principal component analysis that is a linear dimensionality reduction method

sklearn.decomposition.KernelPCA - Non-linear dimensionality reduction using kernels and PCA

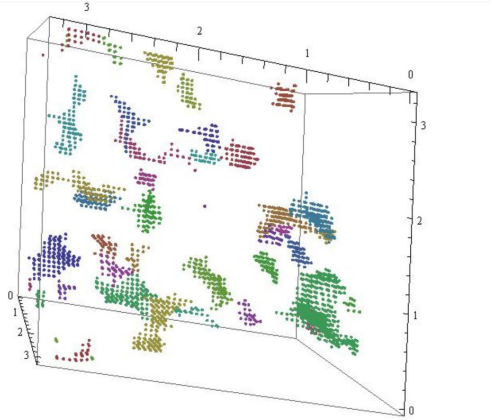
MDS Manifold learning using multidimensional scaling

Isomap Manifold learning based on Isometric Mapping

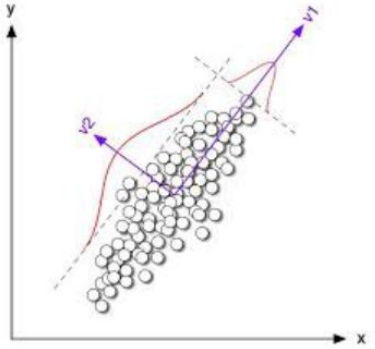
LocallyLinearEmbedding Manifold learning using Locally Linear Embedding

SpectralEmbedding Spectral embedding for non-linear dimensionality

Why embedding?

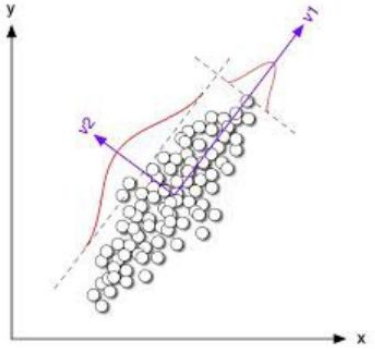


1. reduce the number of variables/dimensions:
   e.g. height, size, number of days etc. in multidimensional setting
2. ensure your variables are independent of one another
   not all variables are independent and how to interpret the independence...
3. visualise high-dimensional dataset, learn new hidden patterns



Simple embedding: PCA

Simple embedding: PCA



Idea PCA:

fitting a *p*-dimensional ellipsoid to the data.

We find the direction of v1 variable where the variance is maximum.
We rotate the x variable axis on the plane:
v1 has maximum variance and v2 have minimum variance so v1 has more information about the dataset
(see notebook with MNIST)

MNIST dataset consist of 60k handwritten images of numbers from 0-9 and is commonly used for training various image processing systems. Data can be downloaded from here.

Simple embedding: KNN

Idea KNN:
find k-nearest neighbors around each point
k is predefined constant, new data point is classified by assigning the label which is most frequent among the *k* training samples nearest to that new data point.

Issues with the method:

Simple embedding: KNN

Idea KNN:
find k-nearest neighbors around each point
k is predefined constant, new data point is classified by assigning the label which is most frequent among the *k* training samples nearest to that new data point.

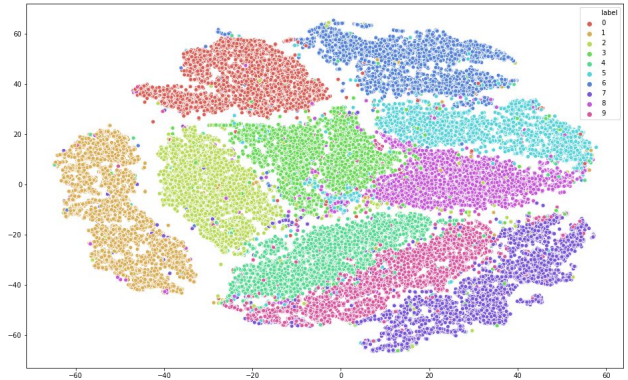
Issues with the method:
Supervised method (k is specified).
Presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance, metric dependence

Simple classification for data: KNN algorithm

See notebook classroom

Issues with embeddings

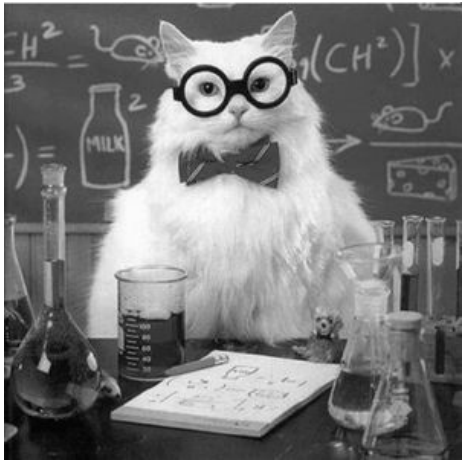
The curse of dimensionality:
issues when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience: **crowded problem.**
Preserve the local structure of data while keeping global one

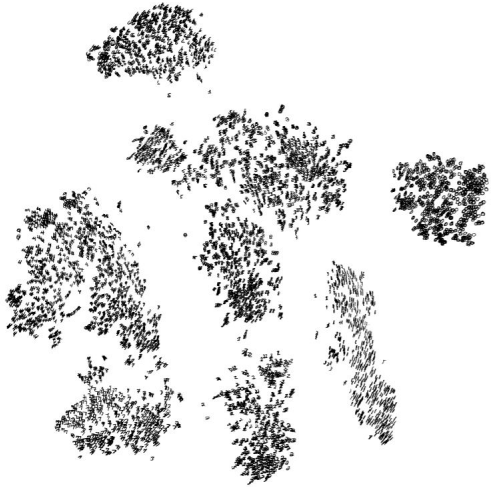


Non-linear embedding tSNE: stochastic embedding

The main idea of the stochastic embedding:
Measure **similarity** between points in high-dimensions.

Brilliant idea:
Cost function - minimisation of distances between distributions of points P(initial) and Q (projected) spaces.



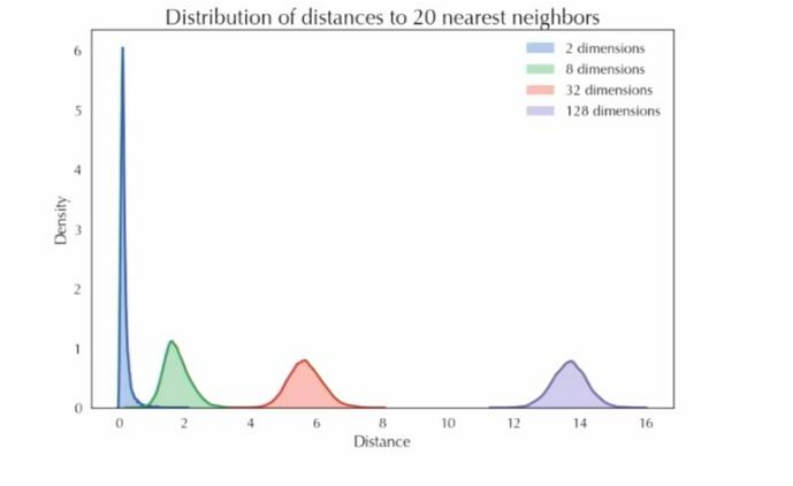


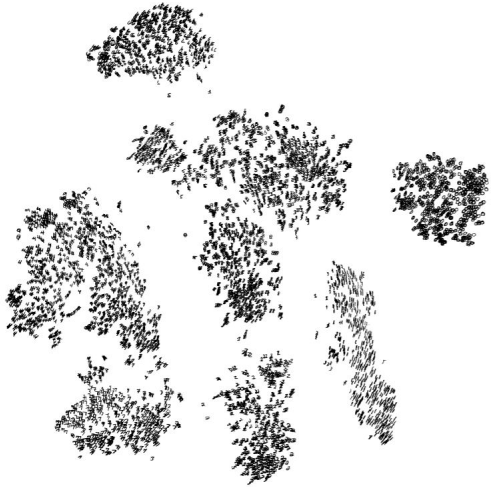
Non-linear embedding tSNE: stochastic embedding

The main idea of the stochastic embedding:
Measure **similarity** between points in high-dimensions.

Brilliant idea:
Cost function - minimisation of distances between distributions of points P(initial) and Q (projected) spaces.

BUT: cost of calculating distances distribution between all points is high...





Non-linear embedding tSNE: stochastic embedding

The main idea of the stochastic embedding:
Measure **similarity** between points in high-dimensions.

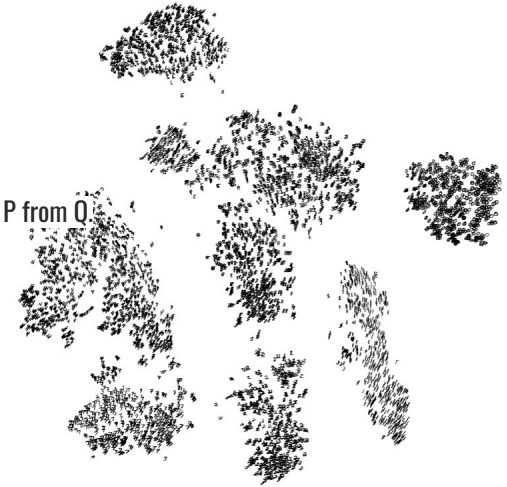
We pick up points to be close to each other **with some probability (!)** and not with certainty (as for PCA), which are close to each other and not the points,
we denote distance between them by by p_ij. Then we follow:

step 1, we compute the similarity between two data points using a conditional probability p. For example, the conditional probability of j given i represents that *x_j* would be picked by *x_i* as its neighbor assuming neighbors are picked in proportion to their probability density under a **Gaussian** distribution centered at *x_i* .

step 2, we let *y_i* and *y_j* to be the low dimensional counterparts of *x_i* and *x_j*.
Then we consider q to be a similar conditional probability for *y_j* being picked by *y_i*
We employ a **student t-distribution** in the low dimension map.

Cost function - minimisation of distances between distributions of points P(initial) and Q (projected)
The locations of the low dimensional data points are determined by minimizing the **Kullback–Leibler divergence** of probability distribution P from Q

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

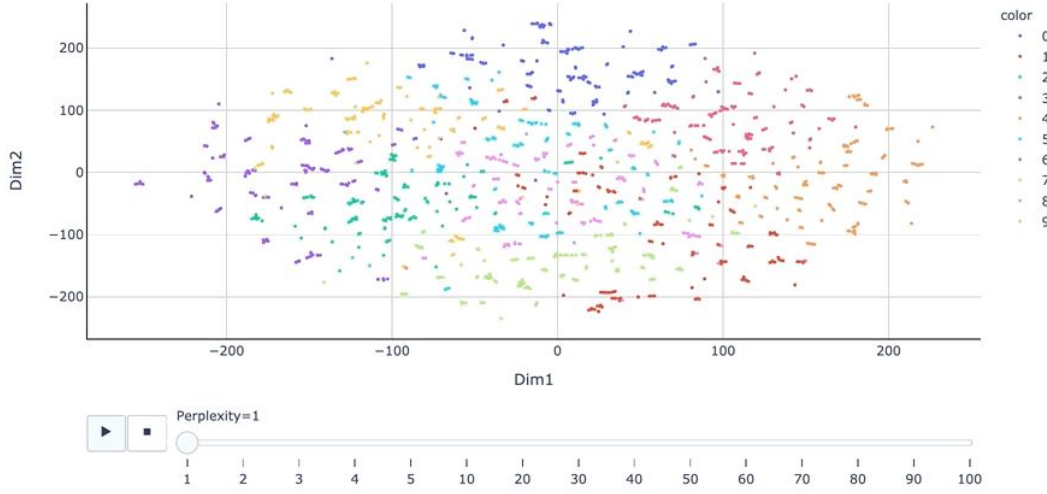


Embedding: tSNE

Methods parameters:
Gaussian variance for estimating distribution between points
Perplexity level - hyper-parameter analogy with nearest neighbors - global parameter.
Number of simulations (iterations)

Initial implementation
https://lvdmaaten.github.io/tsne/

Paper http://www.jmlr.org/papers/volume9/vandermaater



How to use tSNE effectively?

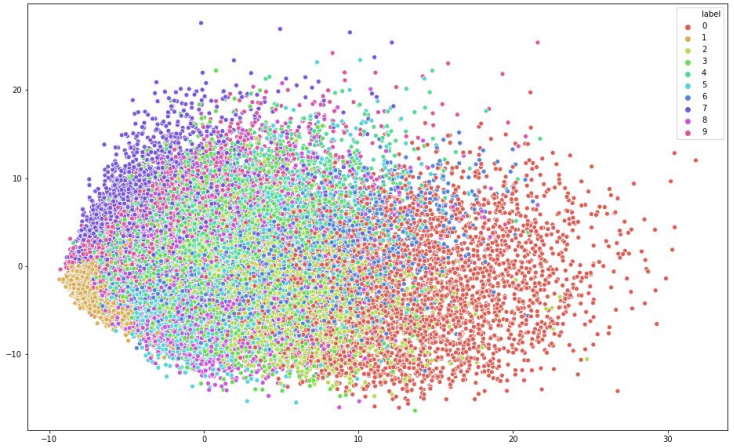
Parameters choice

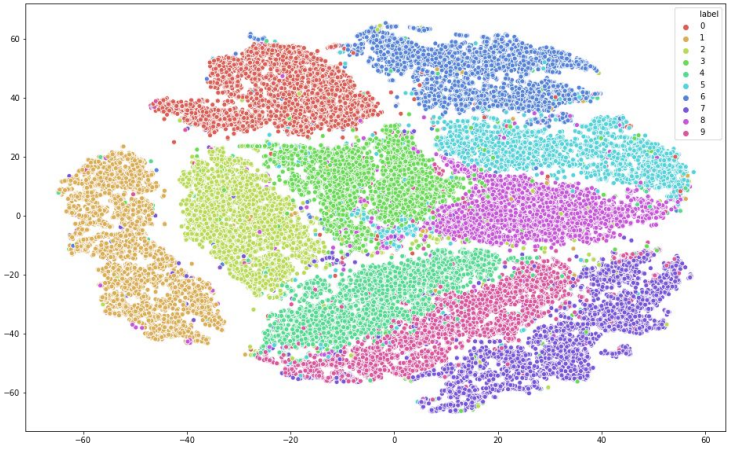
https://distill.pub/2016/misread-tsne/

How to use tSNE effectively?

Parameters choice

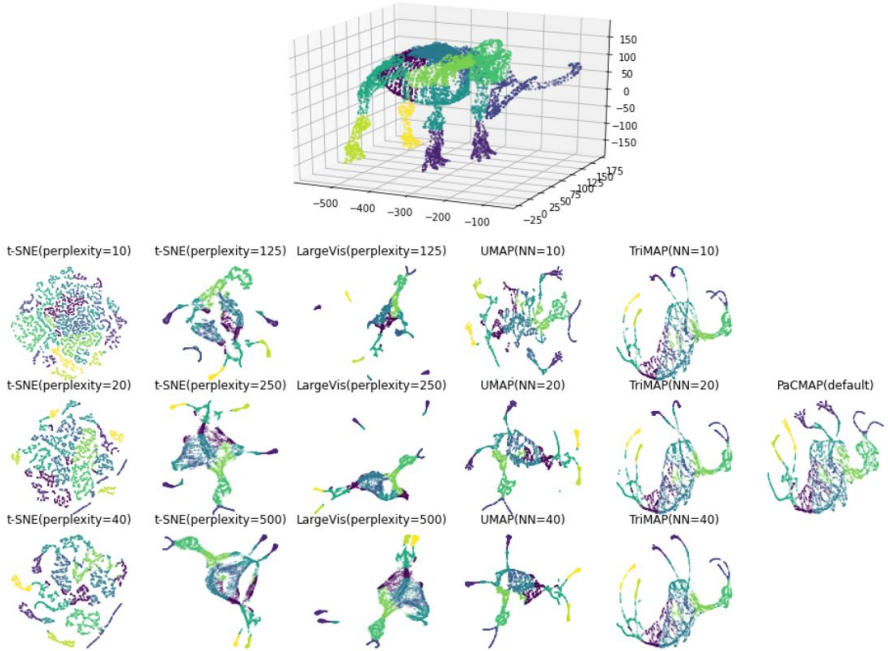
Embeddings comparison
How to compare?





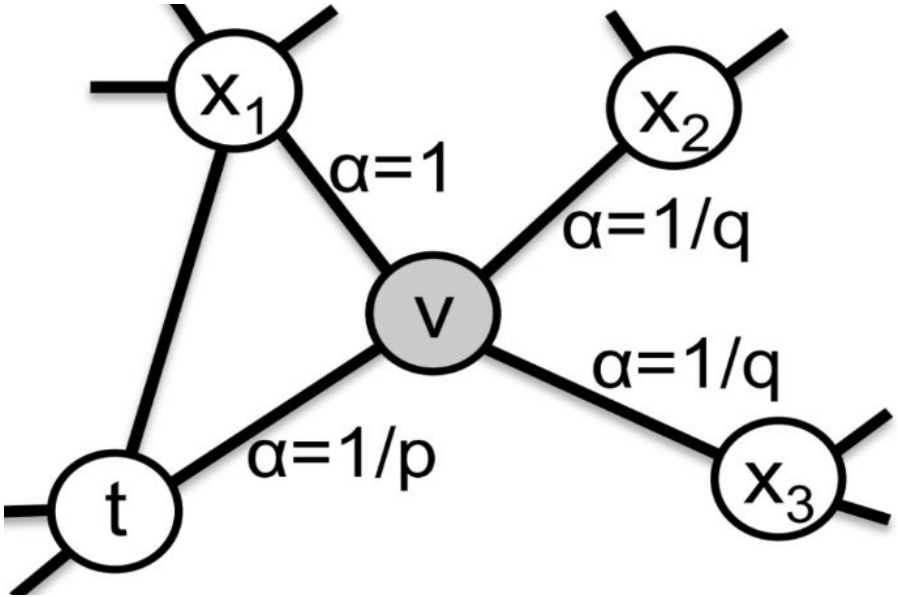
Embeddings comparison:

Choose the dataset we all know well



Node2Vec

Network embeddings
See notebook



Clustering and patterns recognition

https://scikit-learn.org/stable/modules/clustering.html