Exploring Topological Data Analysis: In the perspective of Data Science

INVITED LECTURE:
Date: 19 March 2021
Time: 1.30-3.30PM (Indonesian Time)
2.30PM-4.30PM (Malaysian Time)

Gobithaasan Rudrusamy
Faculty of Ocean Engineering Technology & Informatics,
University Malaysia Terengganu.

email: gr@umt.edu.my / gobithaasan@gmail.com
https://sites.google.com/site/gobithaasan/

Sample codes: https://github.com/gob1thaasan/UndipTDA2021/

2

Content:

From theorems to applications

1. Motivation: Lifecycle of Data Science and job prospects
2. Introduction to Dataset & Machine Learning
3. Exploratory Data Analysis: Topological Data Analysis
4. Persistent Homology.
   I. Invariants: Euler Characteristics & Betti Numbers
   II. Homology computation & Persistence Diagrams
   III. PH Applications
5. Mapper:
   V. Mapper Computation & Graphs
   VI. Mapper Applications
6. Acknowledgement

© 2021 R.U. Gobithaasan All Rights Reserved.
Cite as: R.U.Gobithaasan (19 March, 2021). Exploring Topological Data Analysis. Invited Lecture for Master of Mathematics Study Program, Faculty of Science & Mathematics, Diponegoro University. (Online Presentation)

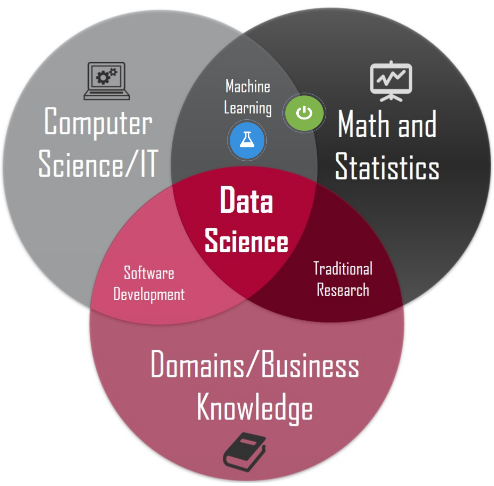
3

Motivation: Data Science

4

Data Science

A picture is worth a thousand words



Complex Data → Useful knowledge

Network Analysis

group structure

Topological Data Analysis

Persistent Homology

Mapper

5

Indonesia as Regional Data Center

https://news.microsoft.com/apac/2021/02/25/microsoft-to-establish-first-datacenter-region-in-indonesia-as-part-of-berdayakan-digital-economy-indonesia-initiative/

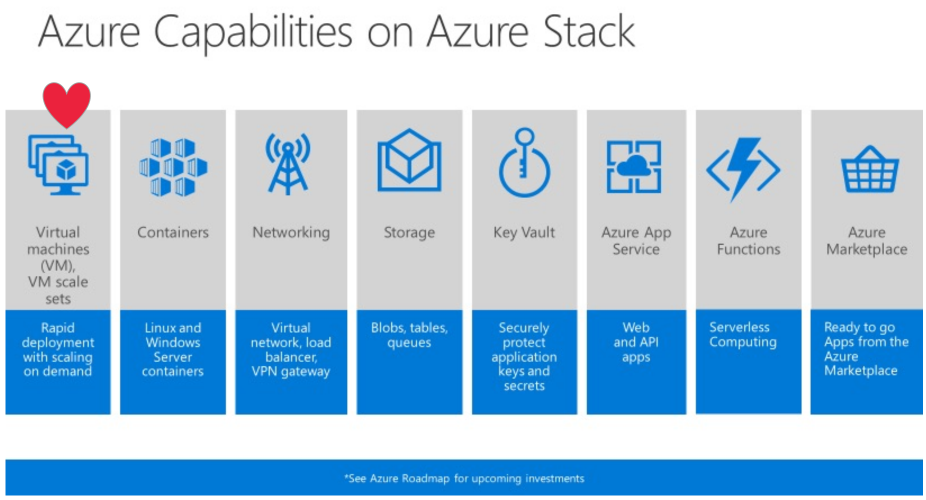
Microsoft increases its investment in Indonesia's digital economy, with plans to establish a new datacenter region and a commitment to skill over 24 million Indonesians by end of 2021 through its decades' long skilling investment

Jakarta, Indonesia, 25 February 2021 – Microsoft today announced its *Berdayakan Ekonomi Digital Indonesia* initiative, which marks a significant commitment to advancing growth and digital transformation for Indonesia, its vibrant developer and startup ecosystem, enterprises, and the public sector. As part of the plan, Microsoft will establish its first datacenter region in Indonesia to deliver trusted cloud services locally, with world-class data security, privacy, and the ability to store data in country. Microsoft also announced plans to skill an additional 3 million Indonesians to achieve its goa of empowering over 24 million Indonesians by the end of 2021, through its long-established skills programs designed to create inclusive economic opportunities in the digital era.

6

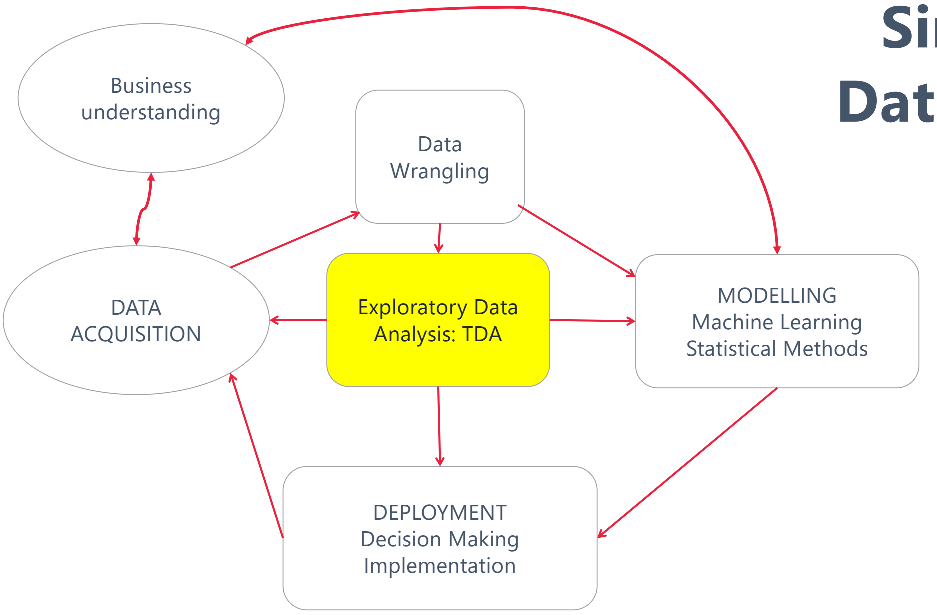
Data Center equipped with Microsoft Azure

https://news.microsoft.com/apac/2021/02/25/microsoft-to-establish-first-datacenter-region-in-indonesia-as-part-of-berdayakan-digital-economy-indonesia-initiative/



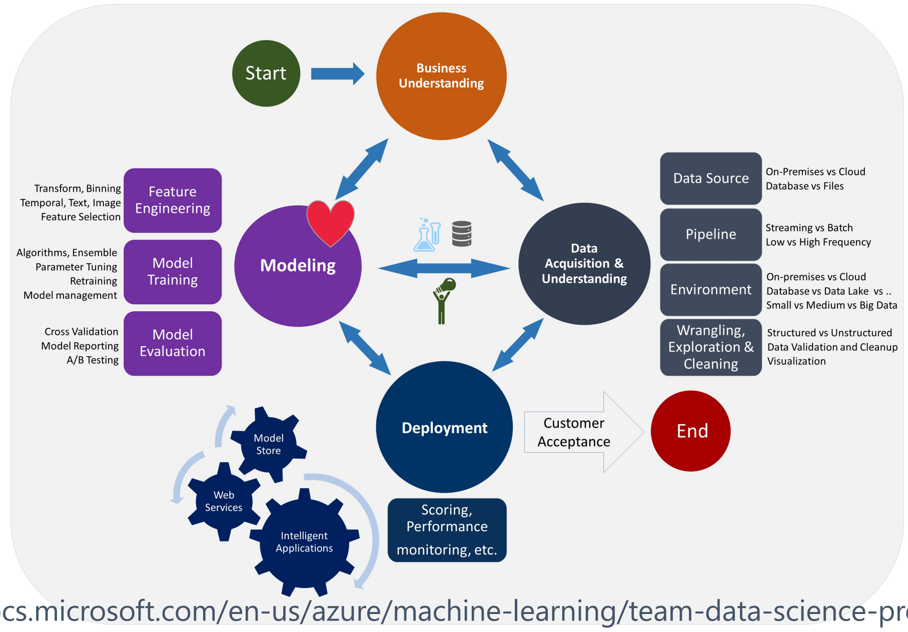
7

Simplified Data lifecycle



8

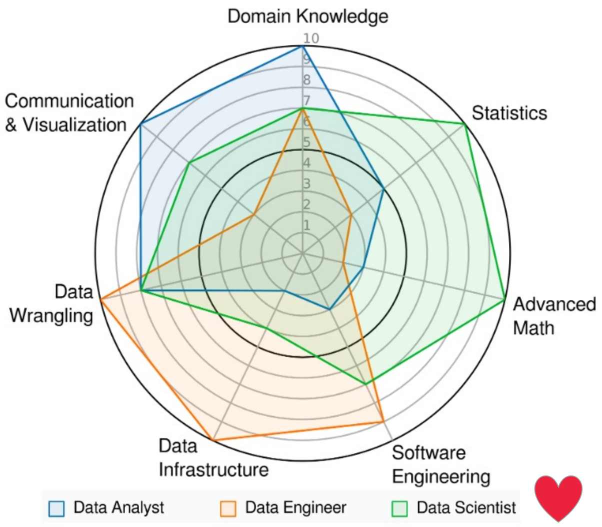
Data Science Lifecycle



https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

9

Spider Chart: Job prospects in Data Science



Project roles:

- Data scientist
- Data Analyst
- Data engineer
- Application developer
- Project lead
- Solution architect
- Project manager

https://towardsdatascience.com/data-science-and-the-data-scientist-db200aac4ea0

10

Introduction:
Dataset & Machine Learning

11

Datasets!

12

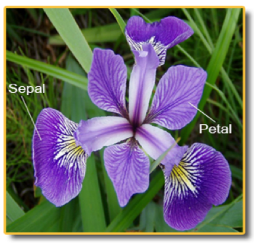
Example: Iris Dataset (labelled)

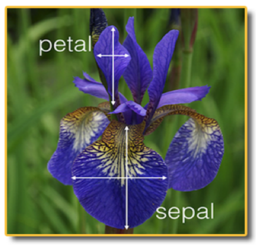
Points $p_i = (\underbrace{Sepal\ Length, Sepal\ Width, Petal\ Length, Petal\ Width}, Species)$

$\boldsymbol{p_i} \epsilon \mathbb{R}^4$: variables x_j: *can be used as a predictor variable called feature*

Iris Versicolor Iris Setosa Iris Virginica

{7.6, 3.0, 6.6, 2.1}->"virginica";
{4.8, 3.4,1.9, 0.2}->"setosa";
{6.4, 2.9, 4.3,1.3}->"versicolor";





Label species $= \{versicolor, sentosa, virginica\}$

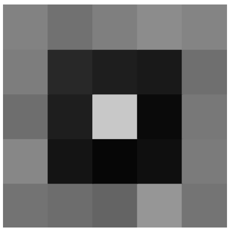
Point cloud $X = \{p_1, p_2, \ldots, p_{50}\}$

Classification problem: Find Species?

13

Point cloud from an image

Higher Dimension: MNIST database of handwritten digits



$$G = \begin{pmatrix} 130 & 113 & 127 & 140 & 132 \\ 125 & 40 & 30 & 25 & 111 \\ 110 & 30 & 200 & 10 & 120 \\ 135 & 20 & 6 & 15 & 123 \\ 115 & 109 & 100 & 150 & 116 \end{pmatrix} \left.\vphantom{\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix}}\right\} \text{Each pixel is a feature}$$

Each picture as a point: (linearized vector)

L: Picture $\mapsto$ Vector

A picture: $p_i = (G_{11}, G_{12}, \dots, G_{15}, G_{21}, \dots, G_{55})$

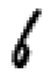
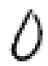
$= (x_1, x_2, \dots, x_{25})$ $\quad p_1 \epsilon \mathbb{R}^{25}$

Point cloud as a collection of pictures: $X = \{p_1, p_2, \dots, p_n\}$ $\quad p_i \epsilon \mathbb{R}^n$

14

MNIST database of handwritten digits (labeled)

Higher Dimension:

{  → 8,  → 6,  → 0,  → 9}

One scanned picture: $p_i = (G_{11}, G_{12}, \dots, G_{15}, G_{21}, \dots, G_{28,28})$

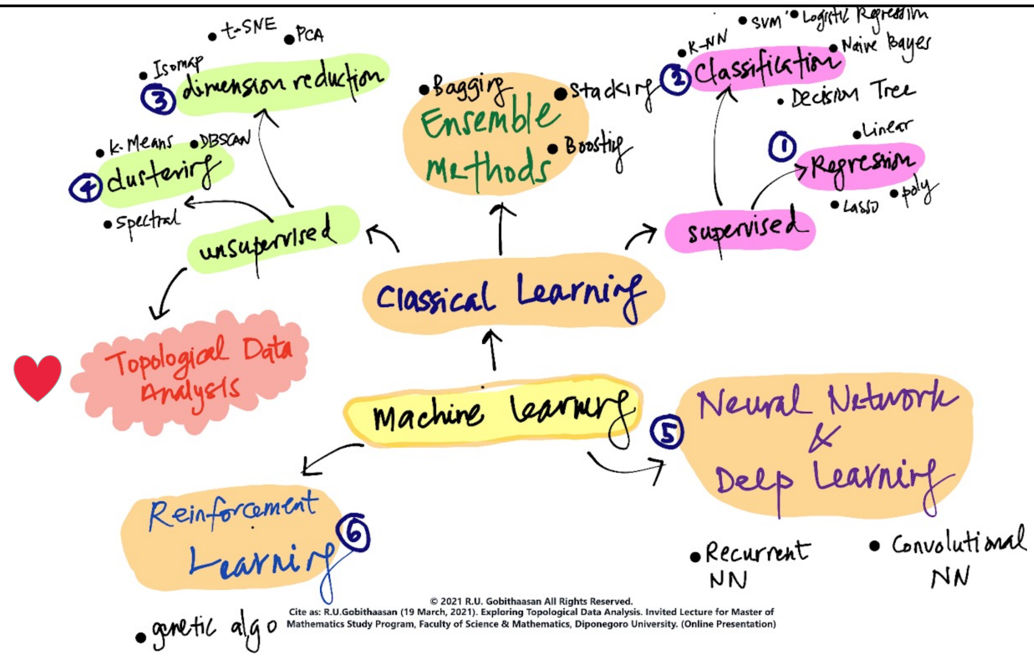
$= (x_1, x_2, \dots, x_{784})$ $\quad p_i \epsilon \mathbb{R}^{784}$

Training dataset (60,000 pictures): Dimension: $60{,}000 \times 784$ $\quad p_i \epsilon \mathbb{R}^{784}$

Test dataset (10,000 pictures): Dimension: $10{,}000 \times 784$ $\quad p_j \epsilon \mathbb{R}^{784}$

Point cloud as a collection of MINIST dataset: $X = (p_1, p_2, \dots, p_{70,000})$ $\quad p_i \epsilon \mathbb{R}^{784}$

15



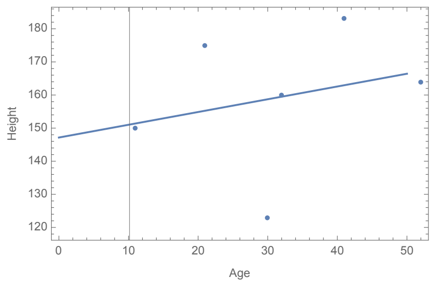
16

1, Classical Regression

Point Cloud

age	height	gender
32	160	female
41	183	female
30	123	female
21	175	male
11	150	male
52	164	female

removed



	DF	SS	MS	F-Statistic	P-Value
x	1	154.838	154.838	0.297766	0.614291
Error	4	2080.	519.999		
Total	5	2234.83			

17

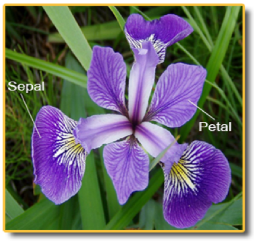
Classification problem: Iris Dataset

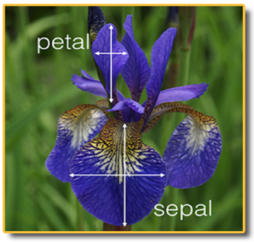
Points $p_i = (\underbrace{Sepal\ Length, Sepal\ Width, Petal\ Length, Petal\ Width}_{}, Species)$

$p_i \epsilon \mathbb{R}^4$: variables x_j: *can be used as a predictor variable called feature*

Iris Versicolor Iris Setosa Iris Virginica

{7.6, 3.0, 6.6, 2.1}->"virginica";
{4.8, 3.4,1.9, 0.2}->"setosa";
{6.4, 2.9, 4.3,1.3}->"versicolor";





Label species $= \{versicolor, sentosa, virginica\}$

Point cloud $X = \{p_1, p_2, \ldots, p_{50}\}$

Classification problem: Find Species?

$given = (S.Length, S.Width, P.Length, P.Width)$

18

2, Classification examples

Point Cloud

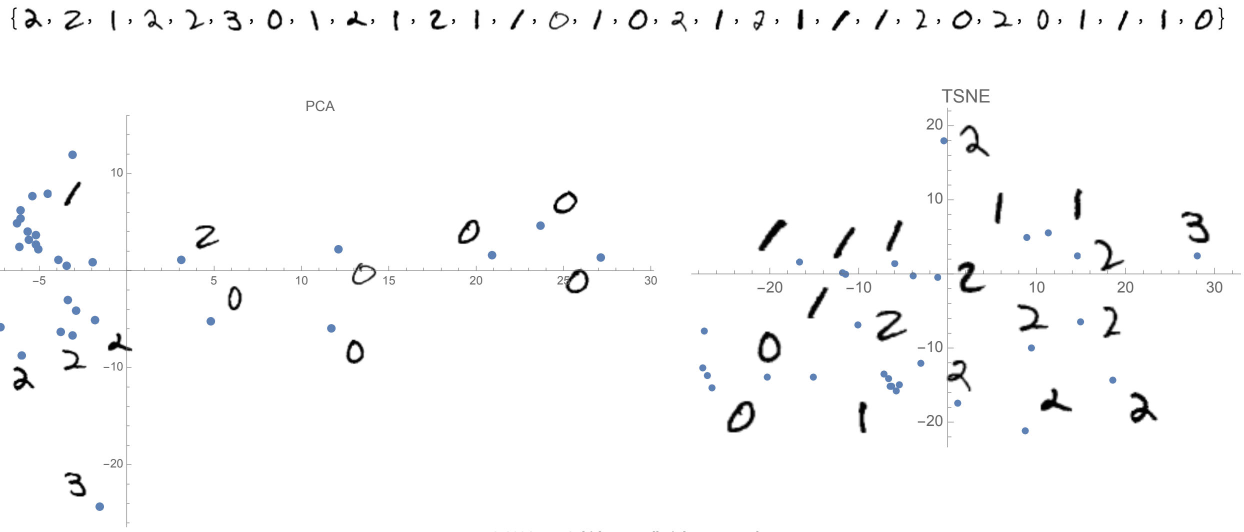
age	height	gender
32	160	female
41	183	female
30	123	female
21	175	male
11	150	male
52	164	female

```
LogisticRegression[<|"age" → 12, "height" → 120|>, "Probabilities"]
<|female → 0.523778, male → 0.476222|>

RandomForest[<|"age" → 12, "height" → 120|>, "Probabilities"]
<|female → 0.324388, male → 0.675612|>

NeuralNet[<|"age" → 12, "height" → 120|>, "Probabilities"]
<|female → 0.000185764, male → 0.999814|>
```

- Spam filtering: logistic regression
- Sentiment Analysis: Naïve Bayes
- Fraud Detection

19

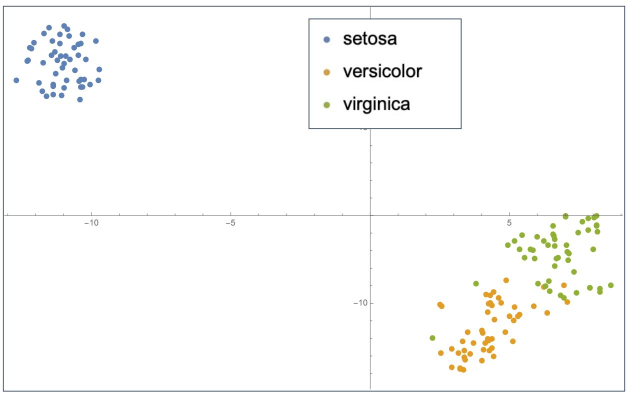
3,Dimension reduction:
Handwriting (assume we lost its labels)



20

3,Dimension reduction:
Iris Dataset (assume we lost its labels)

$X = (p_1, p_2, \dots, p_{50}), p_i \epsilon \mathbb{R}^4$



t-SNE: Stochastic Neighbor Embedding
- low dimensional neighborhood similar to its original neighborhood (preserve local structure).

Applications
- Recommender systems
- Visualization
- Fake image analysis
- Risk management

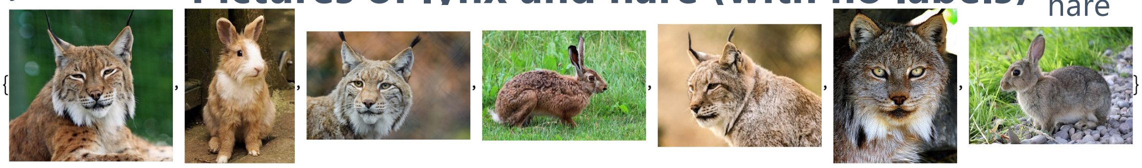
21

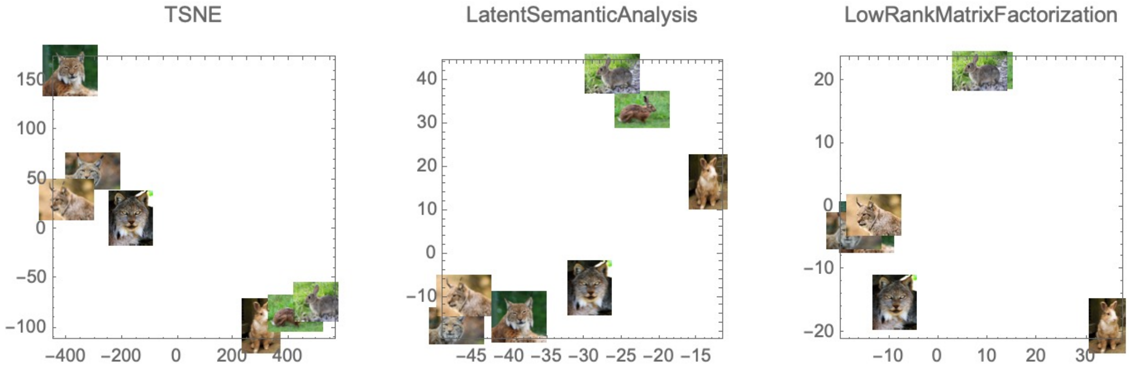
3,Dimension reduction:

Pictures of lynx and hare (with no labels)

lynx

hare





22

Point Cloud

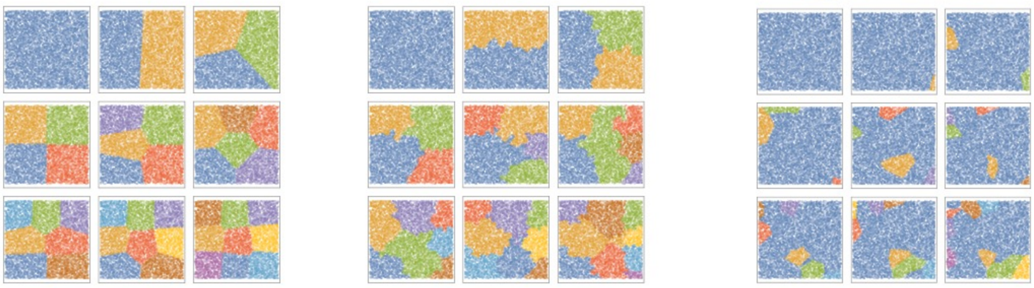
−0.204096	−0.709134
−0.650626	−0.165738
−0.293959	0.350446
0.0888485	0.00588744
−0.681741	−0.110084
0.755913	0.444917
0.176241	−0.429377
0.214507	−0.532368
0.117836	−0.721169
−0.772519	0.724612
−0.799325	0.467894
0.500291	−0.680519
−0.782487	0.10502
−0.646488	0.67687
0.402242	−0.027681
−0.931735	0.370062
0.334873	−0.231835
−0.568243	0.91646
−0.692876	0.145792
−0.638953	−0.0526831

rows 1–20 of 10000

4, Clustering

10,000 Random Real Points between {-1,1}

KMeans Spectral Agglomerate

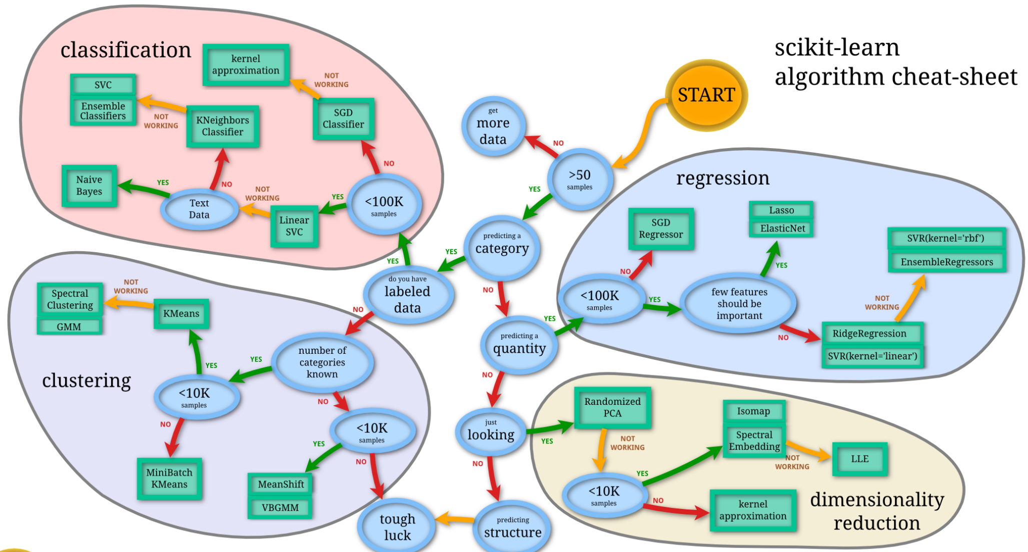


- Market segmentation (types of customers)
- Image compression
- Label new data
- Detect anomalies

23

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

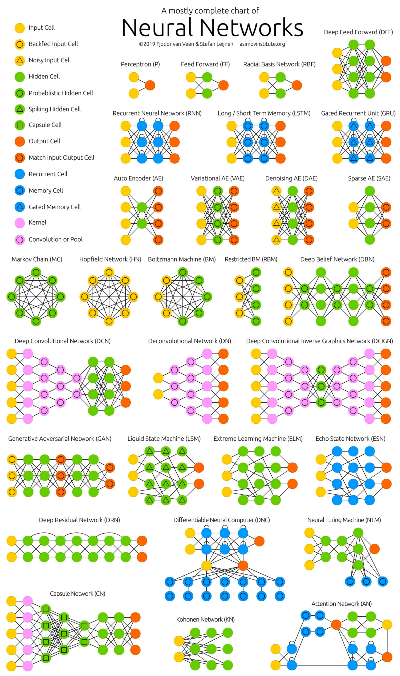
scikit-learn
algorithm cheat-sheet



Back

Different dataset suits different algorithm

24



The Neural Network Zoo

https://www.asimovinstitute.org/neural-network-zoo/

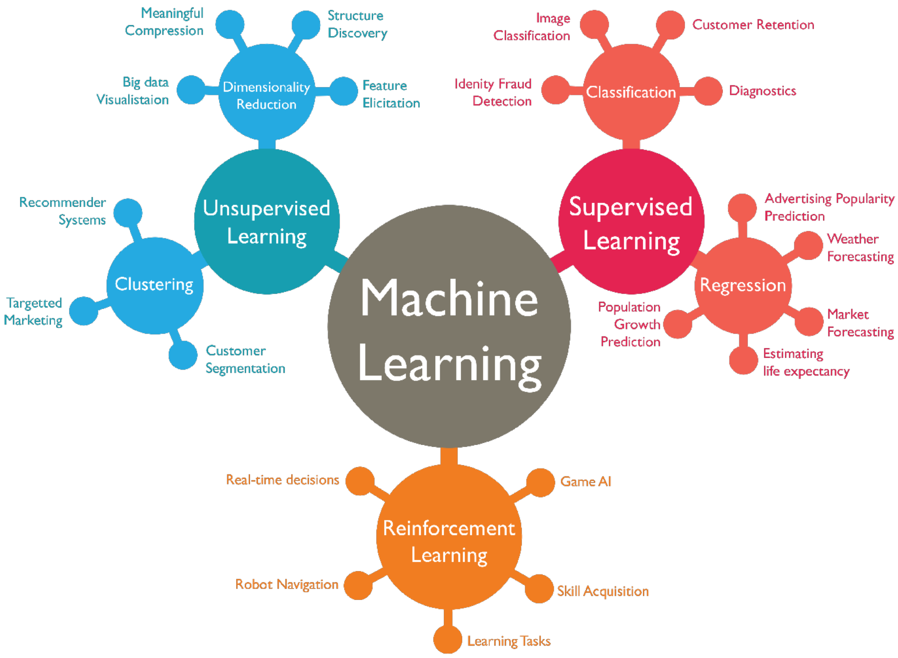
Python library for Neural Network/Deep Learning:

TensorFlow PyTorch

25

Application & Productions

https://vitalflux.com/wp-content/uploads/2020/12/machine_learning_mind_map_3.png



26

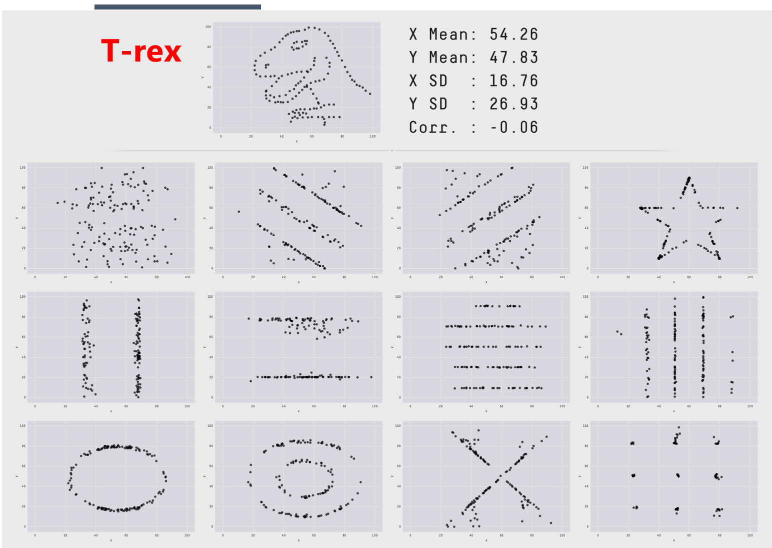
Topological Data Analysis TDA

27

The **failure** of Descriptive Stat: Datasaurus Dozen

https://www.autodesk.com/research/publications/same-stats-different-graphs

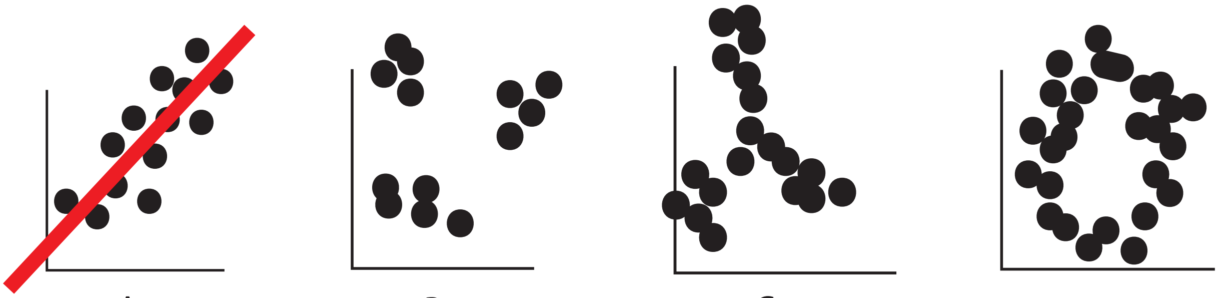
- How to differentiate when they are quantitatively similar?
  - qualitative approach
  - Investigate the **shape** **visually**
- How to visually see high dimensional dataset?
  - $p_i \epsilon \mathbb{R}^n$
  - **Dimension reduction:**
  We lose some info while reducing dimension…



28

What method to use?

- Supervised (regression / classification)
- Unsupervised (clustering / dimension reduction)



A B C D

29

DATA HAS SHAPES!

Topological Data Analysis: Shape has a meaning, meaning has a value

Advantage of TDA:

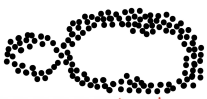
1. **Investigation at different scale:** TDA sensitive to both large and small scale patterns.
2. **Resistance to noise and missing data.** TDA retains significant features of the data.
3. **Invariance under deformation.** The stretch, resize, or orientation of data does not fundamentally change the 'underlying shape' of the data.

As a tool for:

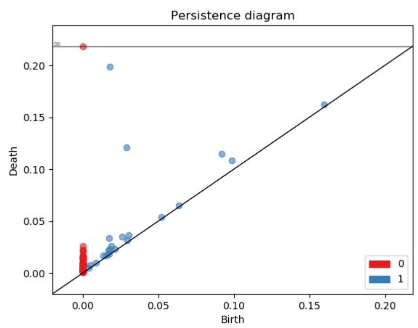
A data exploration tool. Topological exploration of data: qualitative approach.
Point cloud as a signature: TDA can identify global topology and local geometry

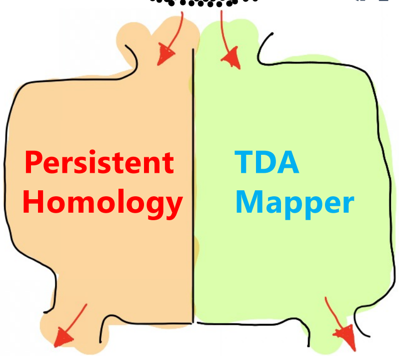
30

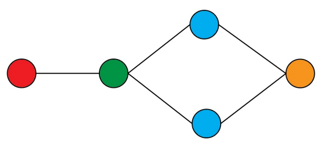
TDA: Two Methodologies



Point cloud
$X = (p_1, p_2, \dots, p_m), p_i \epsilon \mathbb{R}^n$





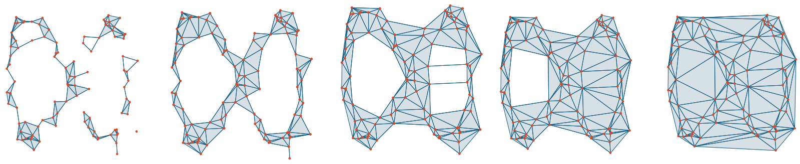


Persistence Diagram:
Topological structures: n-dimension holes

A graph G(V,E):
Topological structures :loops & flares

31

Topological Invariants & Persistent Homology



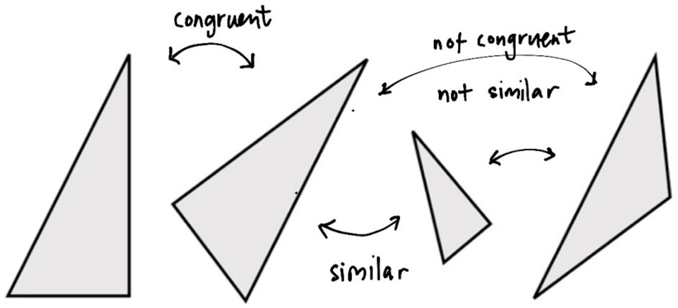
32

Topology: Study of *shape* in $\mathbb{R}^n$

- A fundamental problem of topology is that determining whether given two spaces are **homeomorphic**. (Munkres, elements of algebraic topology)
  - A suitable shape descriptor is known as **invariant.**
- Def. A *(topological)* **invariant** is a map that assigns the same object to spaces of the same topological type.
- **Euclidean geometry** studies invariants under rigid motion in $\mathbb{R}^n$, e.g., moving a cube in space does not change its geometry.
- **Topology** studies invariants under continuous, and continuously invertible, transformations; e.g., mold and stretch a play-doh ball into a filled cube, but not into a donut shape.

33

Invariants in Euclidean Geometry

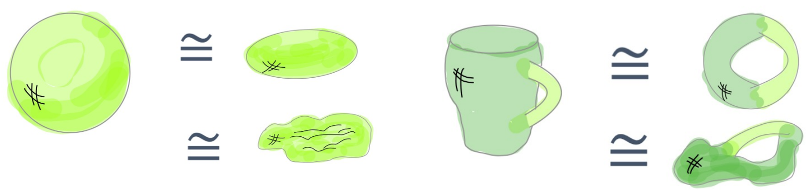


- **Congruence: similar in shape and size.**
- **Unchanged property is called invariants.**
- **Rigid motions does not change the shape or size.**

34

Topology: Study of *shape* in $\mathbb{R}^n$

➢ **Topology studies the connectivity.**
➢ **Homeomorphism: Shapes can be deformed into each other without cutting or gluing.**



➢ **Homeomorphic:**
- Continuous bijective map
- Continuous inverse map

➢ Length and area are not preserved.
 - invariant under deformation

35

Invariant: Betti number!

The *k*th Betti β_k (*X*) number refers to the number of *k*-dimensional holes in a topological space.

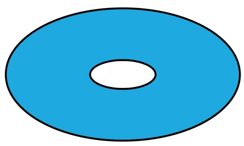
- β_0: # connected components.
- β_1: # one-dimensional or "circular" holes.
- β_2: # two-dimensional "voids".
- β_k: # kth-dimensional 'hole'.

the maximum number of cuts (must cut through) that must be made before separating a topological surface into two pieces or 0-cycles, 1-cycles, etc.

36

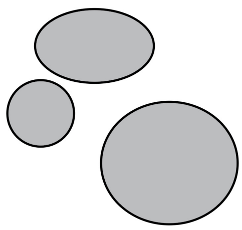
Invariant: Betti number!

β_o (line)=1
β_1 (line)=0



β_o (cd)=**1**
β_1 (cd)=**1**

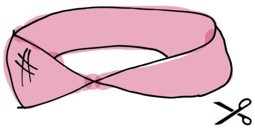
β_o (circle)=1
β_1 (circle)=1



β_o (3 discs)=**3**
β_1 (3 discs) =**0**

37

Invariant: Betti number!



β_o (mobius strip)=1

β_1 (mobius strip)=1

β_o (Torus)=1

β_1 (Torus)=2

β_2 (Torus)=1

β_o (sphere)=1

β_1 (sphere)= 0

β_2 (sphere)= 1

38

Algo: Persistent Homology

1. Simplicial Complexes

- Vary scale distance, ϵ to get various filtrations level
- Various types of filtrations:
  - α: alpha complex
  - Č: Čech complex
  - VR: Vietoris–Rips complex

2. Boundary Matrix Reduction

- Operation on mathematical structure of simplicial complex.

The *nth homology group* H_n

$$H_n = \ker \partial_n / \operatorname{Im} \partial_{n+1}.$$

- $\boldsymbol{\beta_n = rank\ (H_n)}$

3. Persistent Homology

Birth/death: track homological features across the entire filtration:

- Barcodes
- Persistence Diagram (PD)

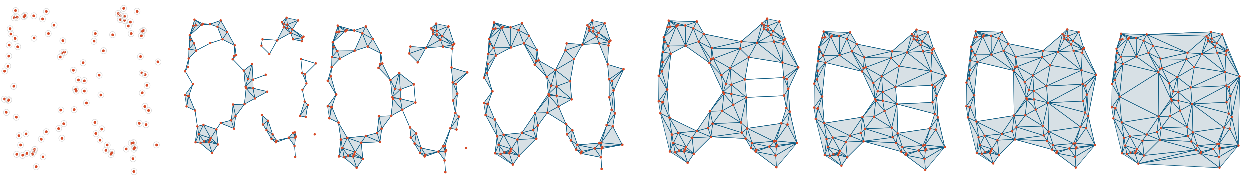
4. Persistence Summaries

- Various kinds of PD representation as features.
- Statistics of PDs: Averages, standard deviations, distributions, etc.

39

Multiresolution and various kind of filtrations

FILTRATION DEMO

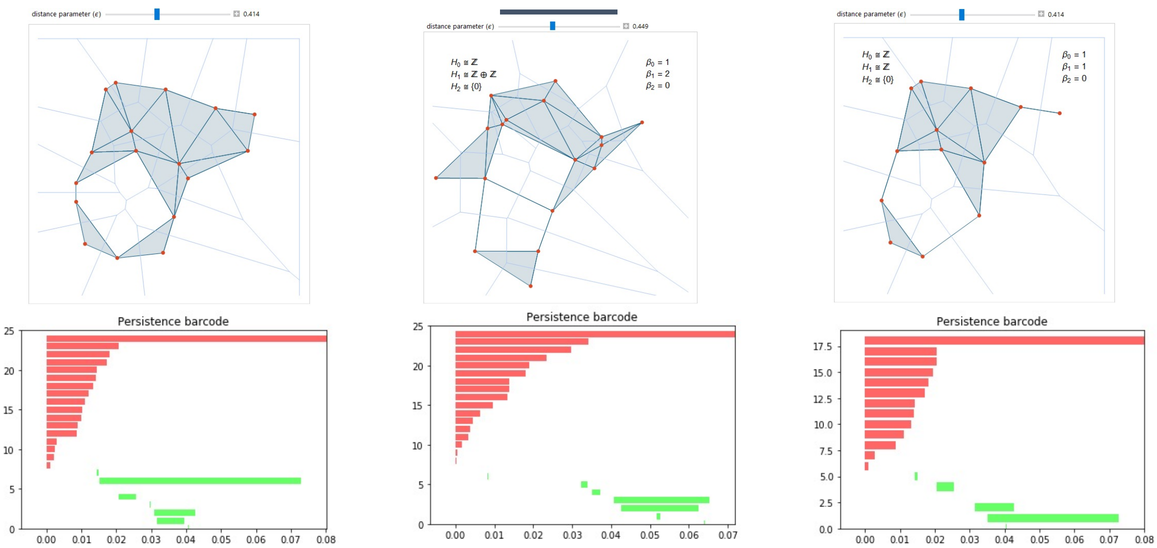


Can I identify the nth dimension hole which persist over various resolution?

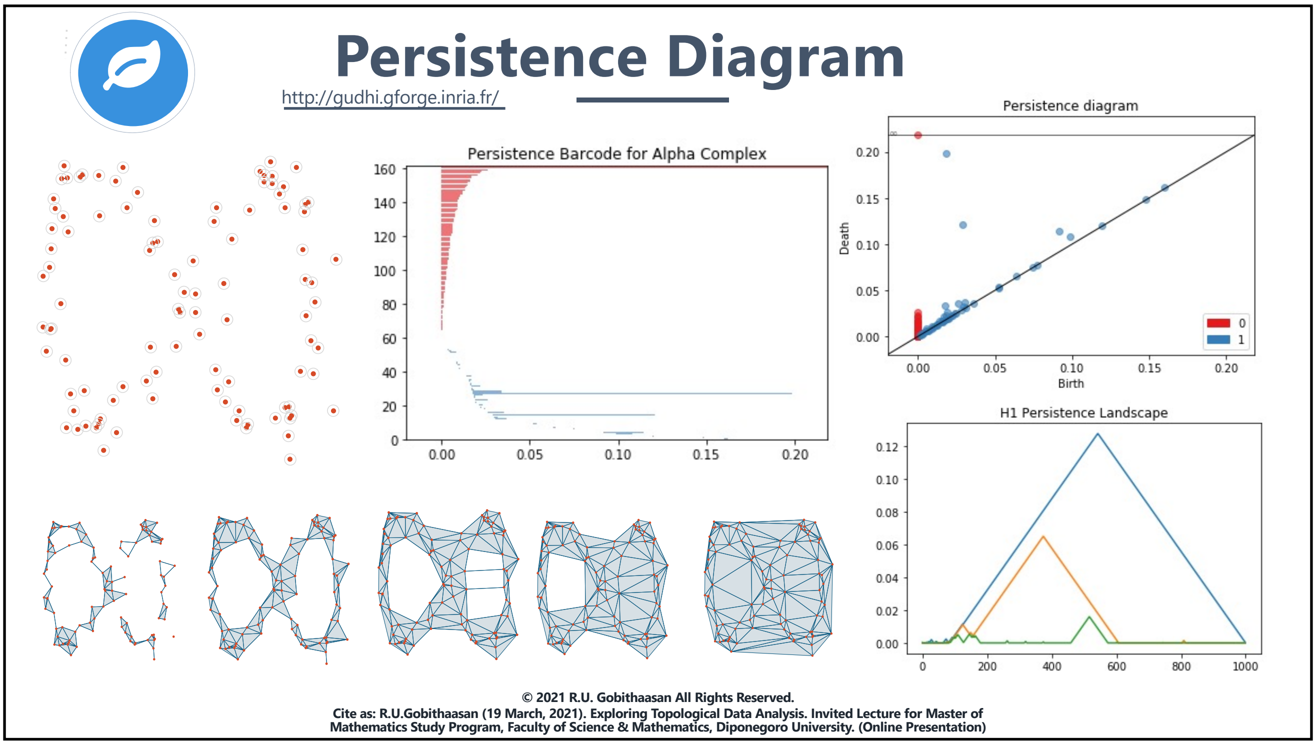
40

Persistence Diagram

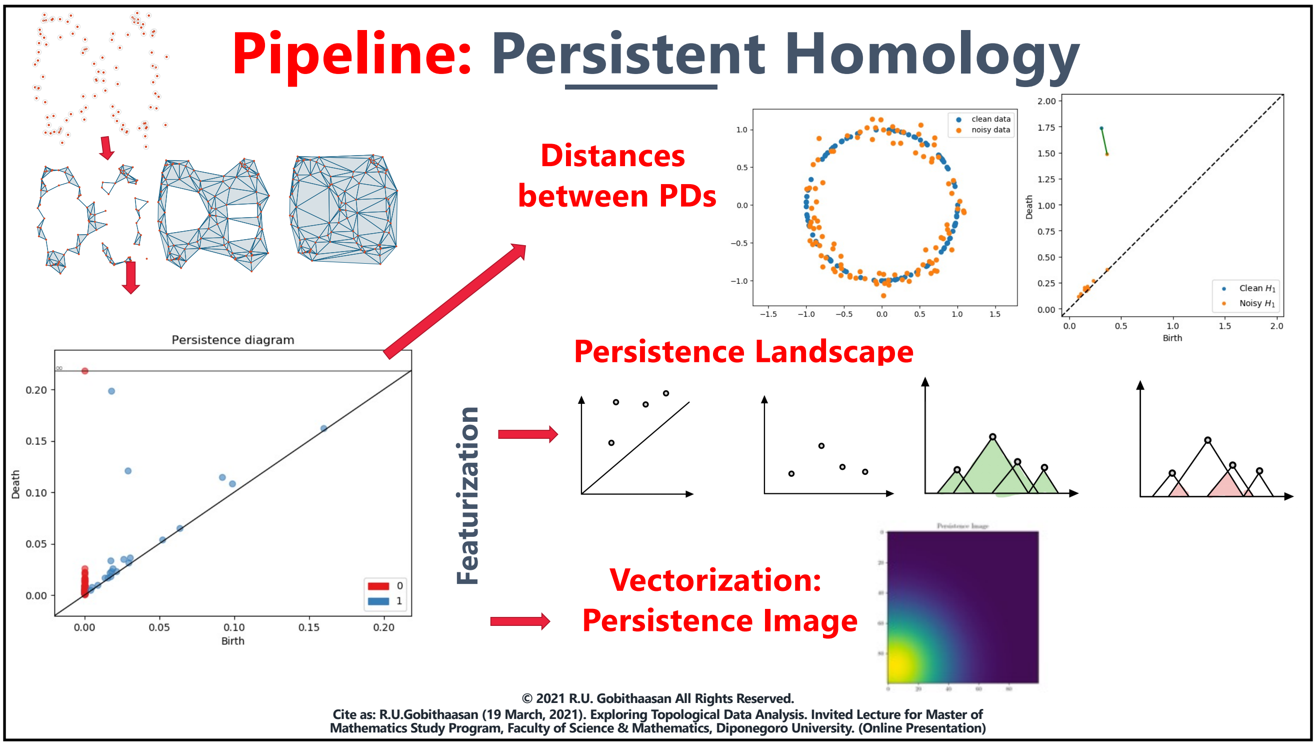
Track homological features across the entire filtration:



41


Persistence Diagram
http://gudhi.gforge.inria.fr/
Persistence Barcode for Alpha Complex
Persistence diagram
Death
Birth
H1 Persistence Landscape
© 2021 R.U. Gobithaasan All Rights Reserved.
Cite as: R.U.Gobithaasan (19 March, 2021). Exploring Topological Data Analysis. Invited Lecture for Master of Mathematics Study Program, Faculty of Science & Mathematics, Diponegoro University. (Online Presentation)

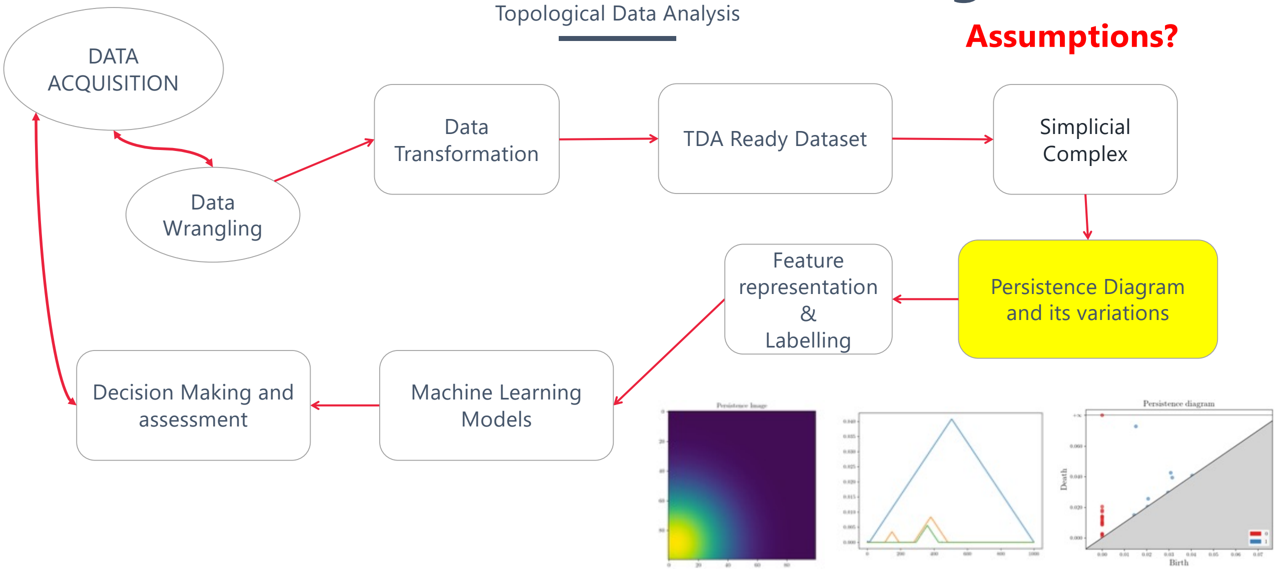
42


Pipeline: Persistent Homology
Distances between PDs
clean data
noisy data
Death
Birth
Clean H_1
Noisy H_1
Persistence diagram
Featurization
Persistence Landscape
Vectorization: Persistence Image
© 2021 R.U. Gobithaasan All Rights Reserved.
Cite as: R.U.Gobithaasan (19 March, 2021). Exploring Topological Data Analysis. Invited Lecture for Master of Mathematics Study Program, Faculty of Science & Mathematics, Diponegoro University. (Online Presentation)

43

TDA for Machine Learning

Topological Data Analysis



44

PH: Success stories

- Jose Perea (MSU), Elizabeth Munch (MSU), Chris Tralie (Ursinus College)
  - time series dataset (dynamical systems, audio and video dataset)
- Robert Ghrist's Gr. (math. UPenn)
  -network flow, sensor network
- Yasuaki Hiraoka (Riken, Kyoto)
  -materials science
- Michelle Feng & Mason A. Porter (UCLA)
  -political science: voting (geospatial Data)

45

Software: Persistent Homology

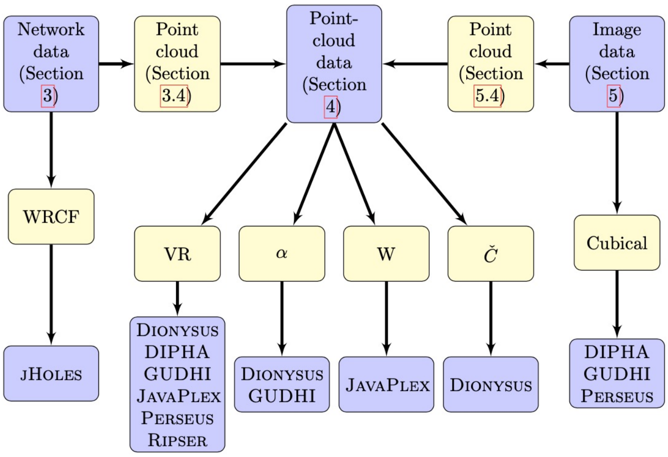


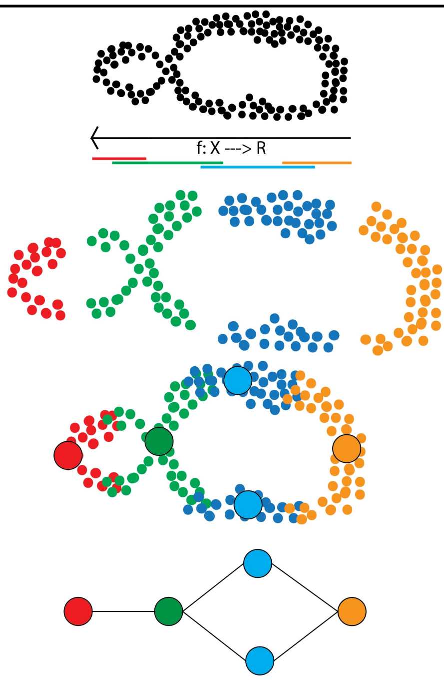
Figure 1: Schematic for how to navigate the tutorial.

Types of complexes
- α: alpha complex
- Č: Čech complex
- VR: Vietoris–Rips complex
- W: weak witness complex
- WRCF: weight rank clique filtration

https://github.com/n-otter/PH-roadmap

46

TDA Mapper



47

TDA Mapper

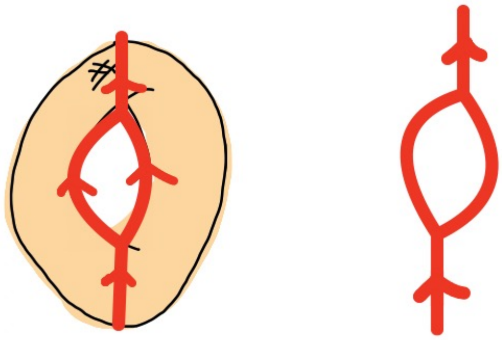
- Compressed representation approximating Reeb graph

Extracting insights from the shape of complex data using topology

P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson & G. Carlsson

Affiliations | Contributions | Corresponding authors

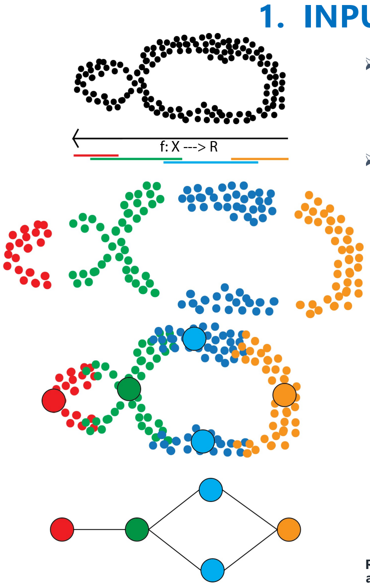
Scientific Reports 3, Article number: 1236 | doi:10.1038/srep01236
Received 13 September 2012 | Accepted 06 December 2012 | Published 07 February 2013



Reeb graph from height function

48

Mapper: Discrete approximation of Reeb Graph



1. INPUT (dissimilarity measure)

- Choose a filter function $f: X \longrightarrow \mathbb{R}$
  - Density measure, score of measure difference, eccentricity measure.
- a cover U of $f(X)$
  - Resolution r (diameter of intervals)
  - Gain g (% of interval overlaps):

2. Decompose into clusters

- From the covering $U = \{U_\alpha\}_{\alpha \in A}$ construct the clusters: $X_\alpha = f^{-1}(U_\alpha) \subseteq X$
- k-NN, Rips, various Clustering methods

3. Compute The Nerve

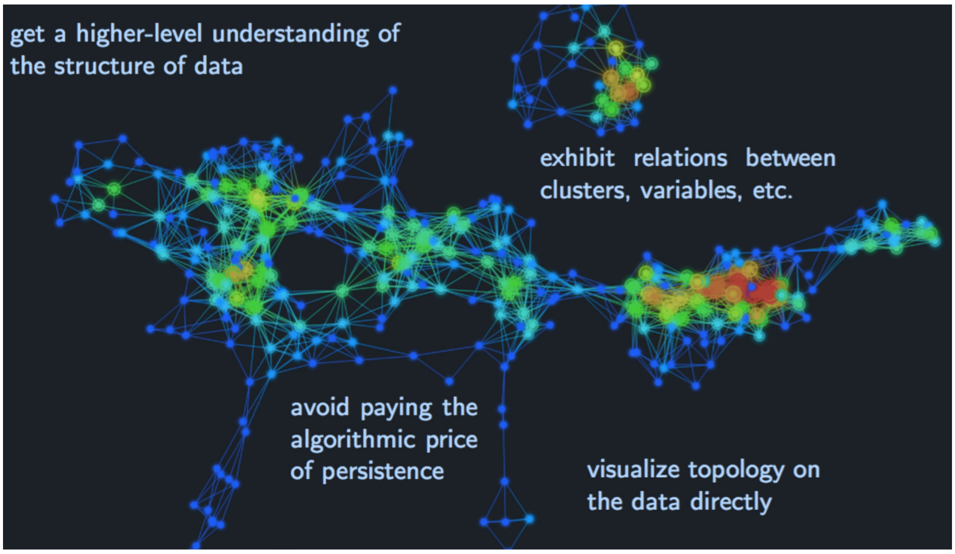
- A vertex for each cluster.
- An edge iff the clusters have a common point

4. OUTPUT: Simplicial Complex

- Graph G(V,E) with connected vertices of nonempty intersection

49

Mapper: What to investigate from the graph?



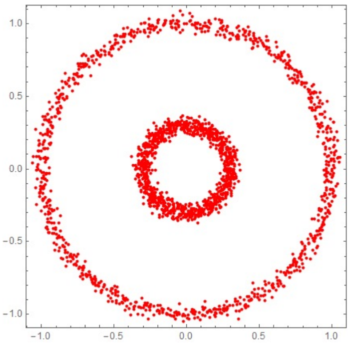
http://www.enseignement.polytechnique.fr/informatique/INF556/index.html

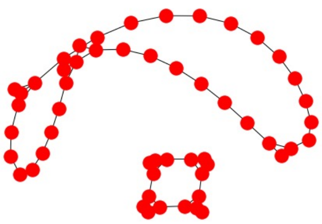
50

TDA Mapper

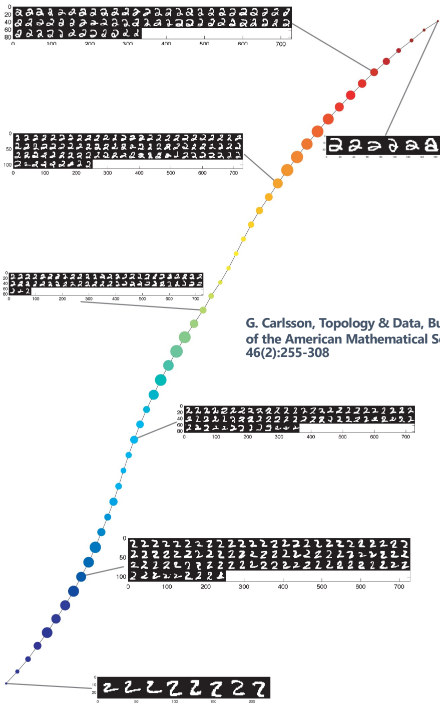
- Coordinate freeness
- Invariance to deformation
- Compressed representation

make_circles(n_samples=2000, noise=0.03, factor=0.3)





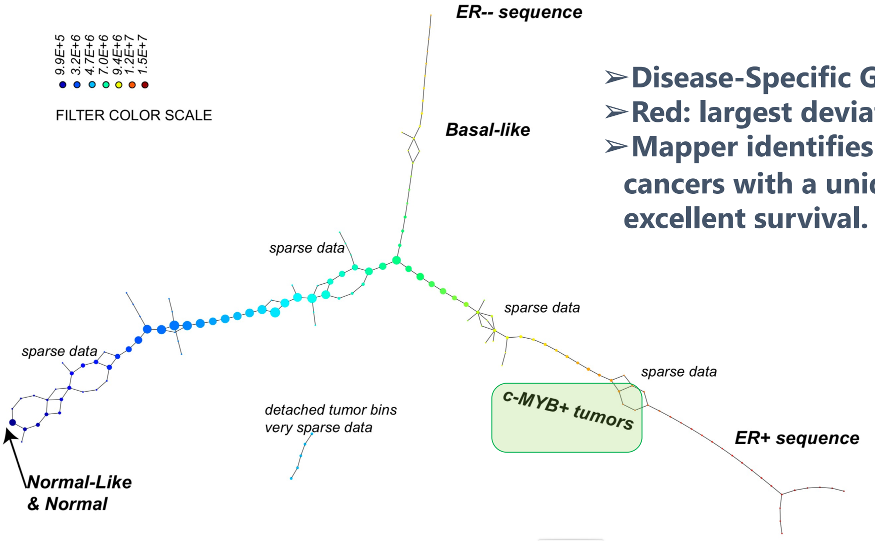
https://kepler-mapper.scikit-tda.org/



G. Carlsson, Topology & Data, Bulletin of the American Mathematical Society 46(2):255-308

51

Mapper: Success Story

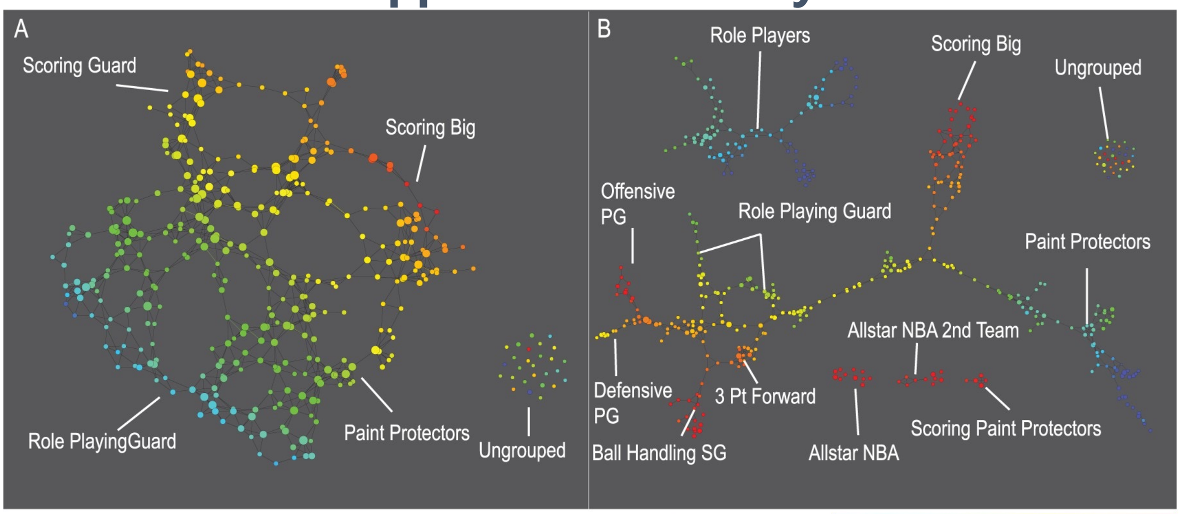


- ➢Disease-Specific Genomic Analysis (DSGA)
- ➢Red: largest deviation from tumor cells.
- ➢Mapper identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.

Nicolau, M., Levine, A. J., & Carlsson, G. (2011). Proceedings of the National Academy of Sciences, 108(17), 7265–7270. doi:10.1073/pnas.1102826108

52

Mapper: Success Story



Extracting insights from the shape of complex data using topology
P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson & G. Carlsson ,Scientific Reports volume 3, Article number: 1236 (2013)

53

Mapper: Success stories

Symphony
AYASDI

- Financial Crimes,
- Disease Recovery,
- Genomics,
- Spinal Cord Injury/Traumatic Brain Injury,
- Asthma, Diabetes, Oncology, Proteomics,
- Predicting Earthquakes,
- Macroeconomic loops,
- Remote sensing

https://www.ayasdi.com/knowledge-center/?type=publication

54

Software: Mapper

- Python Mapper

http://danifold.net/mapper/index.html

- Skicit-tda: Keppler Mapper

Scikit-TDA
Topological Data Analysis for the Python ecosystem.
http://scikit-tda.org
https://kepler-mapper.scikit-tda.org/en/latest/

- Gioto-tda

Giotto-tda

https://giotto-ai.github.io/gtda-docs/0.4.0/notebooks/mapper_quickstart.html

55

Acknowledgements

'bumping into' the right person at the right time.

- This research was supported by Ministry of Education (MOE) through Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UMT/02/2).
- Paweł Dłotko (Institute of Mathematics, Polish Academy of Sciences)
- Vidit Nanda (Mathematical Institute, University of Oxford)
- Martin Crossley (Dept of Mathematics, Swansea University)

56

Thank you...

- Prof. Dr. Widowati, S.Si, MS.i
- Dr. Heru Tjahjana, M.Si
- Ratna Herdiana, M.Sc., Ph.D.
- Audience for your interest!

Demo codes using Python is available at:

https://github.com/gob1thaasan/UndipTDA2021/

email: gr@umt.edu.my / gobithaasan@gmail.com

57