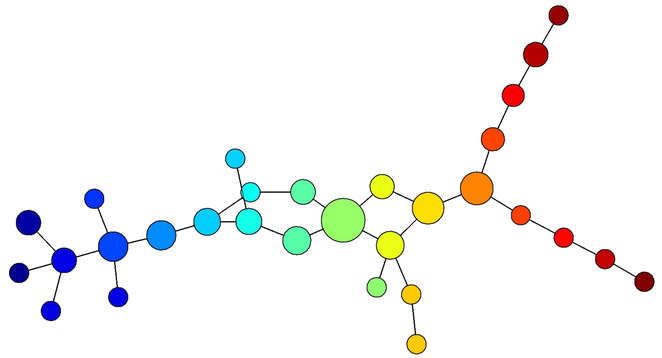


Topological Data Analysis

-Methods and Examples-

Sunghyon Kyeong

Severance Biomedical Science Institute,

Yonsei University College of Medicine

Contents

- Brief overview of topological data analysis
- About **Ayasdi**

  Startup company providing solutions for data analytics
- Topological data analysis - Methods
- Applications to medical science data
- Applications to social science data

Brief Overview of
Topological Data Analysis

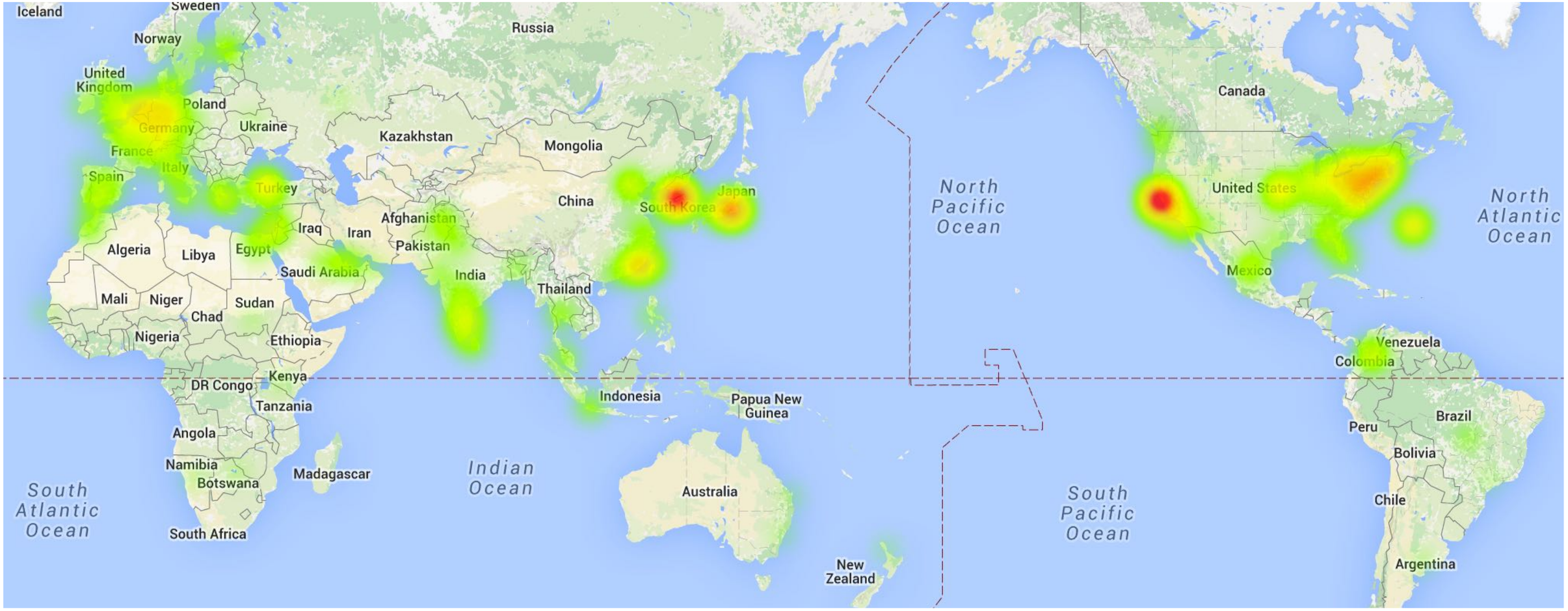
Machine Learning

- Supervised Machine Learning
  → Classification of new input data
  (LDA, bayesian, support vector machine, neural network, and so on)

- Unsupervised Machine Learning
  → Clustering of given dataset / Community detection
  (k-means clustering, modularity optimisation, ICA, PCA, and so on)

- Topological Data Analysis
  → partial clustering with allowing overlaps among clusters

World Interests for TDA



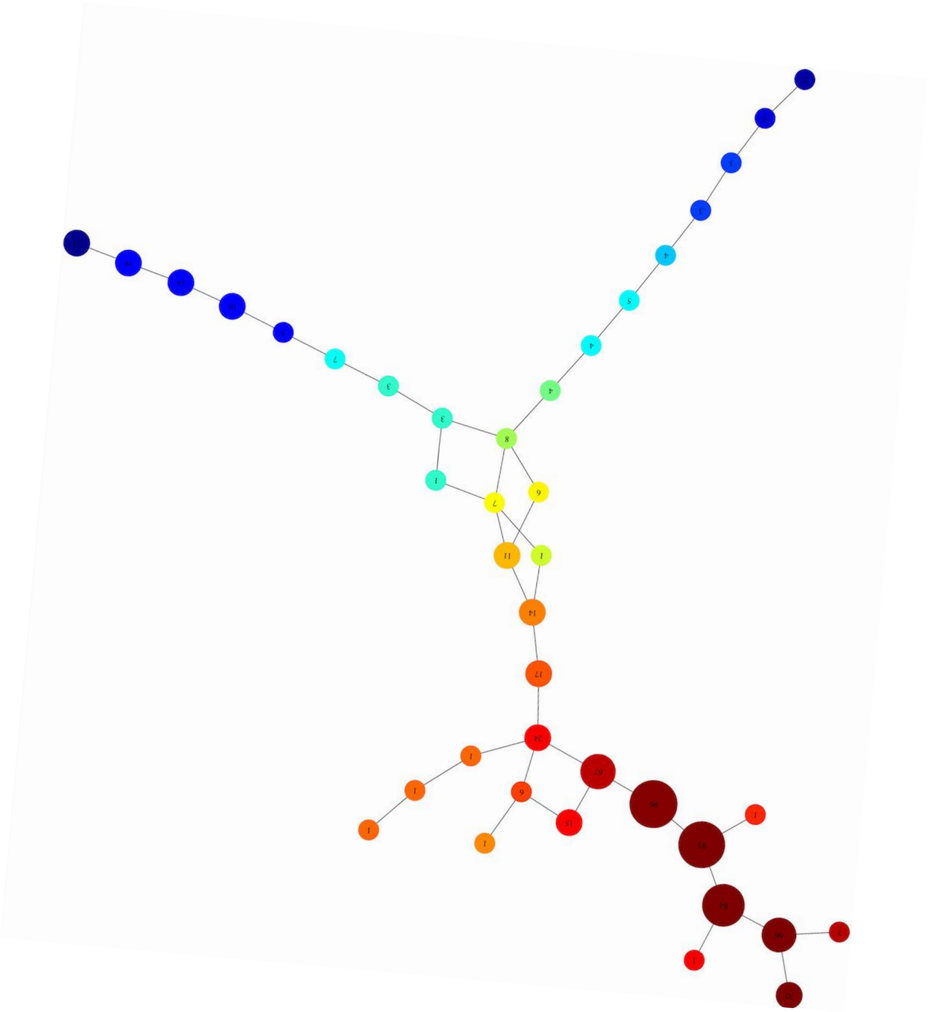
Heat map for viewers of my TDA slide at sliceshare (for 2500 viewers during 2015.2.14. - 2014.11.31.)

Data has **Shape**

An example

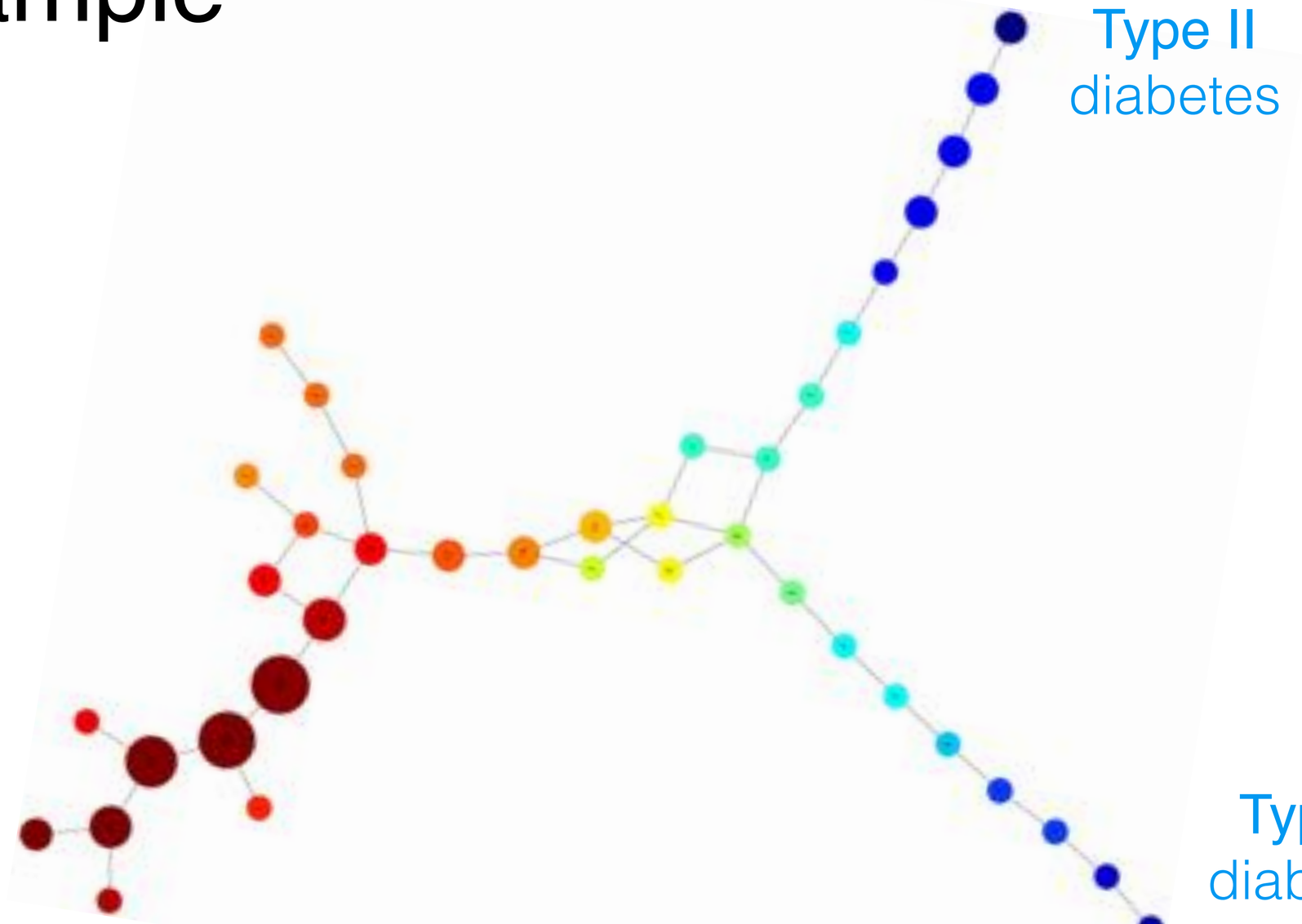
Raw Data
(diabetes related data)

id	weight	fpg	ga	ina	sspg
1	0.81	80	356	124	55
2	0.95	97	289	117	76
3	0.94	105	319	143	105
4	1.04	90	356	199	108
5	1	90	323	240	143
6	0.76	86	381	157	165
7	0.91	100	350	221	119
8	1.1	85	301	186	105
9	0.99	97	379	142	98
10	0.78	97	296	131	94
11	0.9	91	353	221	53
12	0.73	87	306	178	66
13	0.96	78	290	136	142
14	0.84	90	371	200	93
15	0.74	86	312	208	68
16	0.98	80	393	202	102
17	1.1	90	364	152	76
18	0.85	99	359	185	37
19	0.83	85	296	116	60
20	0.93	90	345	123	50



Shape has **Meaning**

An example



Type II
diabetes

Normal

Type I
diabetes

Meaning drives **“Values”**

When to use TDA?

- To study complex **high-dimensional data**
  : feature selections are not required in TDA
- **Extracting shapes** (patterns) of data
- Insights qualitative information is needed.
- Summaries are more valuable than individual parameter choices.

Algebraic

Topology

$Betti_0$: 10
$Betti_1$: 5

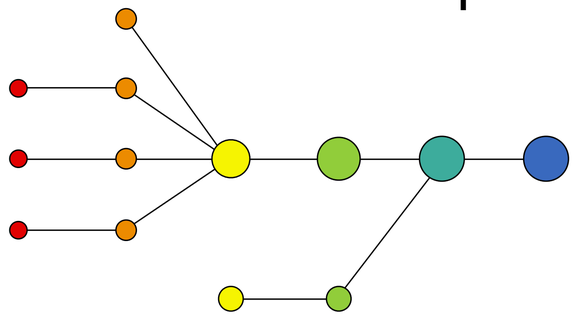
$Betti_0$: 8
$Betti_1$: 5

mathematically
defined “holes” in data

$Betti_0$: clusters
$Betti_1$: holes
$Betti_2$: voids

Geometric

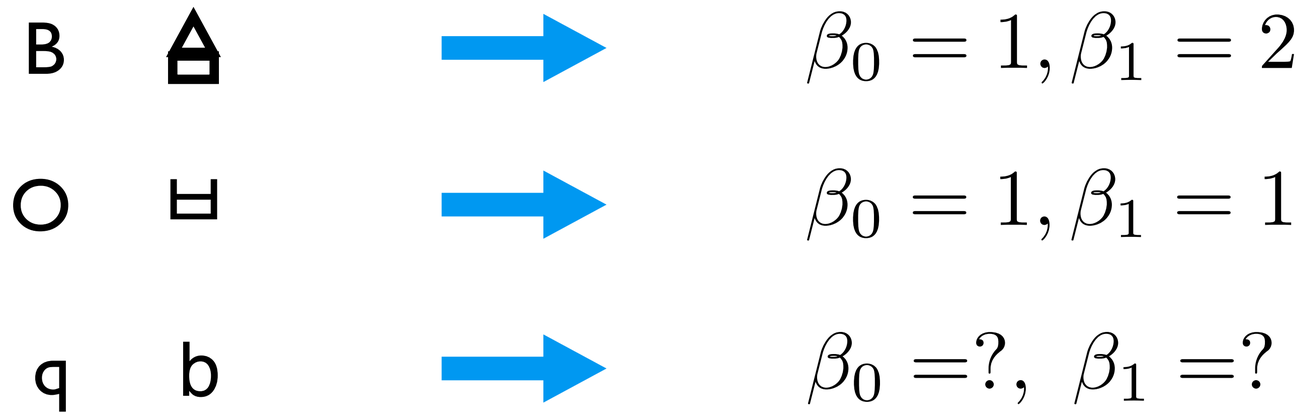
Topology



Algebraic Topology

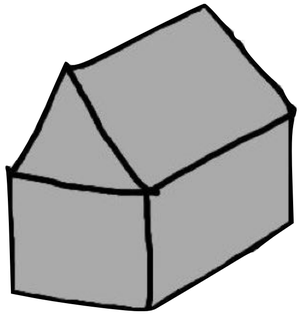
Quantitive Information

Persistent homology is a spatial type of homology that is useful for data analysis. **Betti numbers**, which come from computing **homology**, reflect the topological properties of an object.

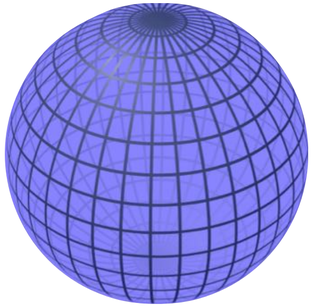


β_0 : the connected components β_1 : the number of holes

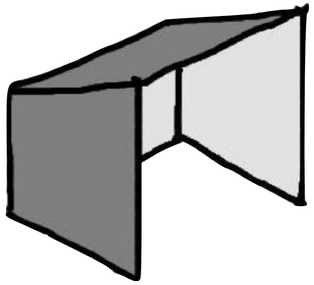
Homology



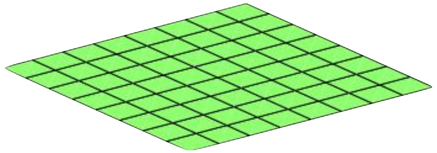
homeomorphic to



, $Betti_2 = 1$



homeomorphic to



, $Betti_2 = 0$

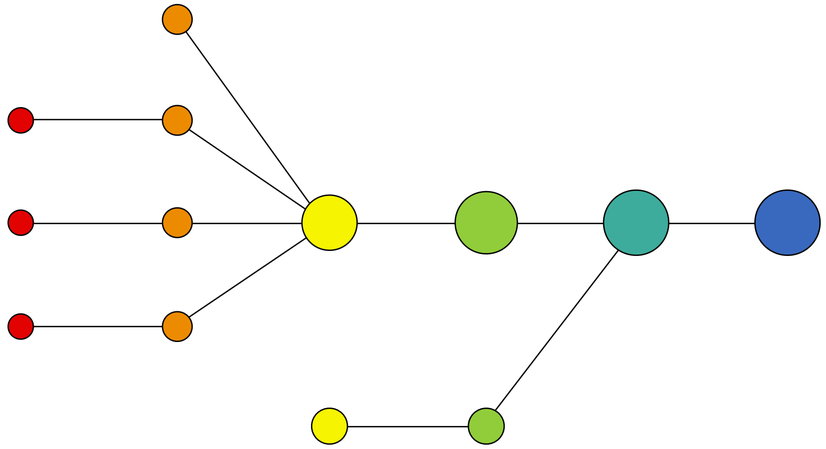
Ref) Xiaojin Zhu, IJCAI 2013 presentation slide

Geometric Topology

Extracting Shapes of Data

points cloud data

?



topology

Ref) Figures are obtained from Y.P. Lum et al (2013) **Scientific Reports** | 3: 1236

Topological Data Analysis using *Mapper*

Two input functions

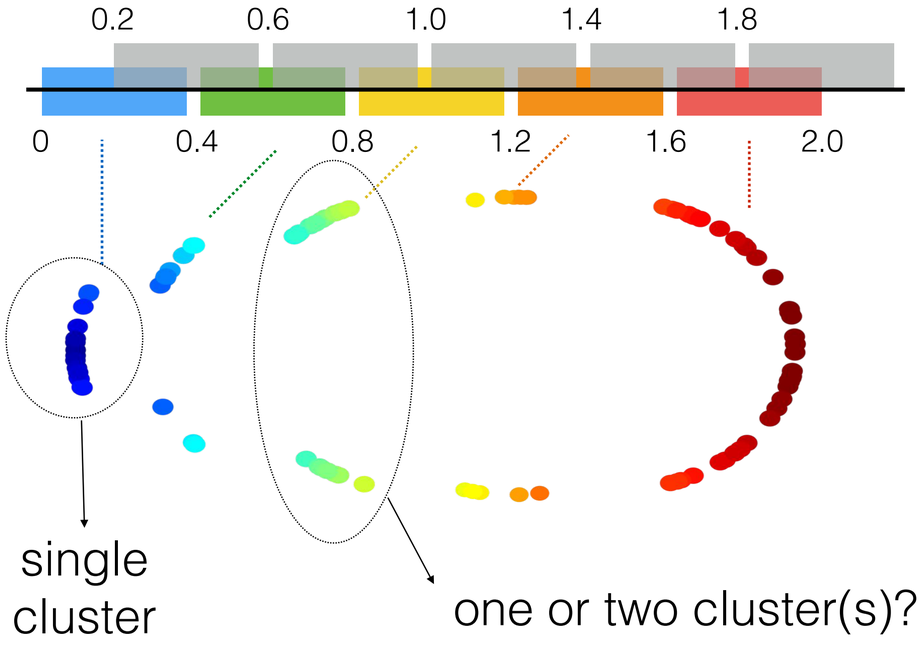
- **filter** is to collapse high-dimensional data set into a single point
- **distance** as a measure of distance between data points

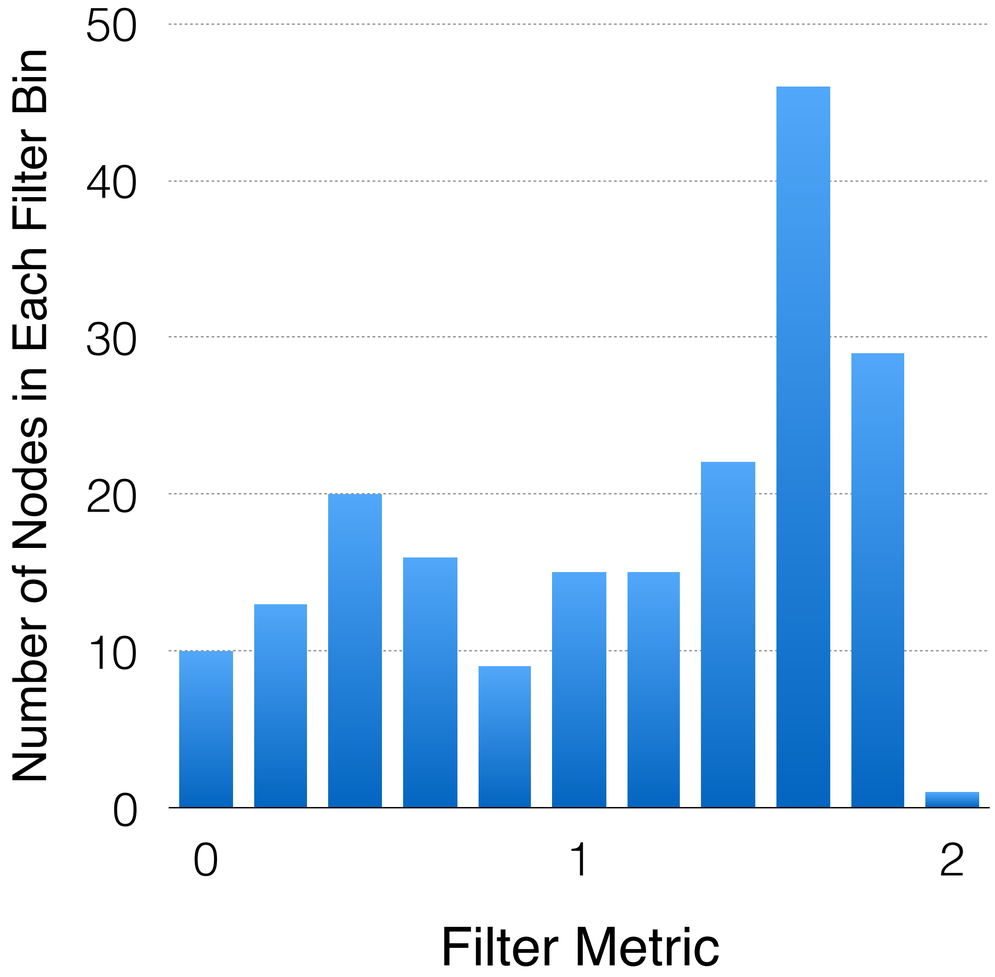
Resolution Parameters

- *Intervals, overlap, magic fudge*

Filter : Divide point clouds into each filter bin

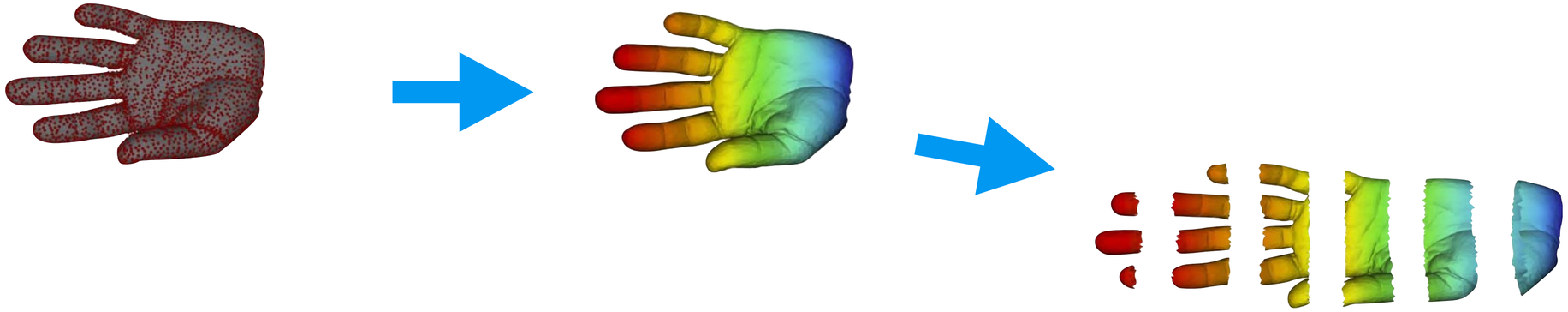
Filter Range : 0.0~2.0
Intervals : 5
Overlap : 50%
Interval Length : 0.4





Filter Function

- Filter function is not necessarily linear projections on a data matrix.
- People often uses functions that depend only on the distance function itself, such as a measure of centrality.
- Some filter functions may not produce any interesting shapes.



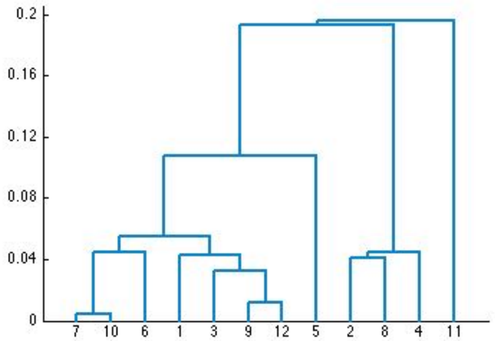
Ref) Figures are obtained from Y.P. Lum et al (2013) **Scientific Reports** | 3: 1236

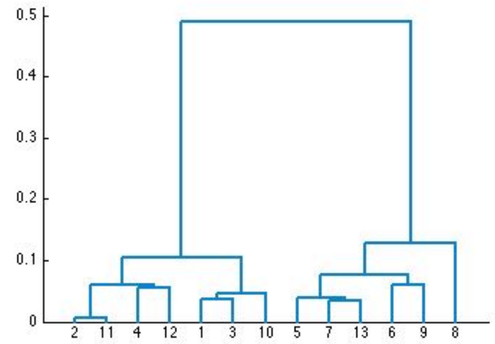
Distance Function

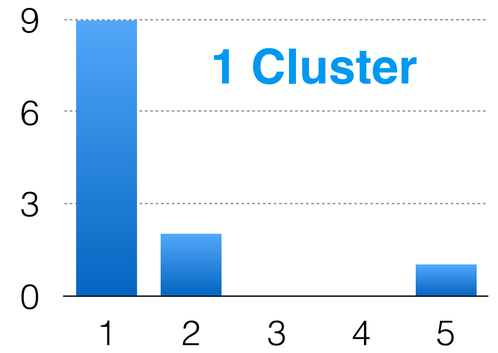
- distance between all pairs of data points.
- both euclidean or geodesic distances could be used.

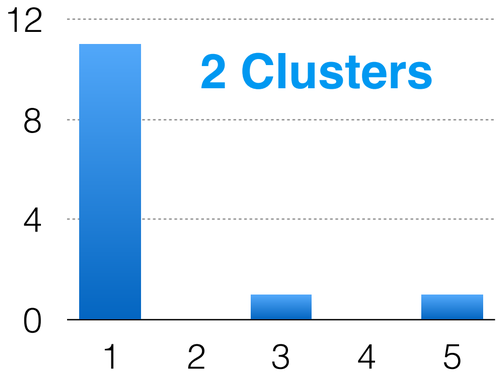
Distance & Clustering

- Single linkage dengrogram is used for clustering point clouds based on distance between two nodes.







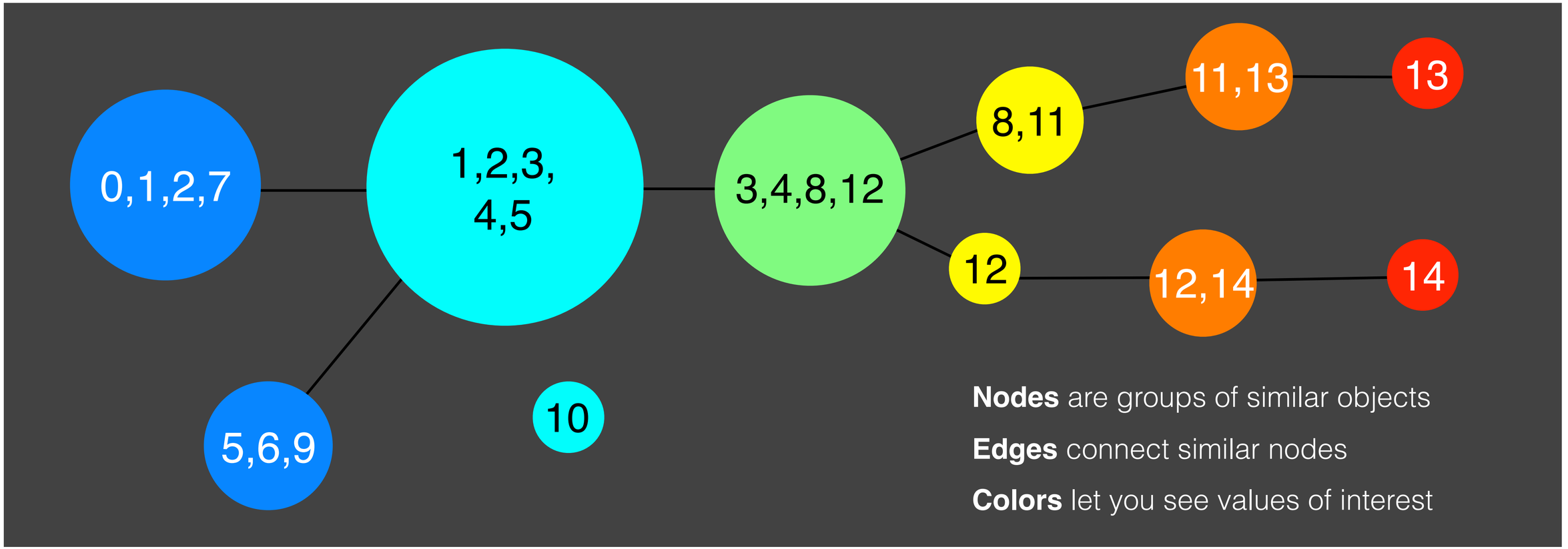


Magic Fudge is the number of bins in the distribution of the distance obtained from single linkage dendrogram.

No. of clusters are estimated from the number of continuous bins having zero elements.

Nodes, Edges, Colors

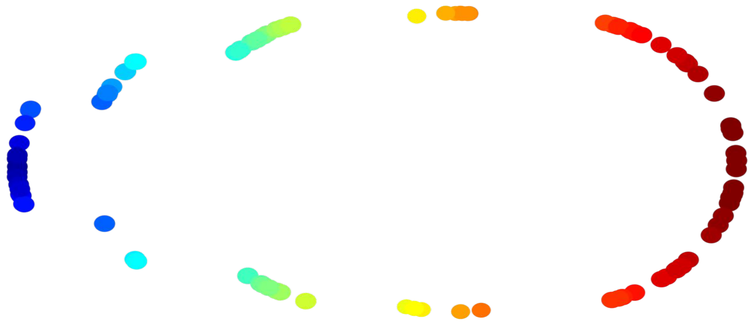
indices for points cloud: 0,1,2,3,4,5,6,7,8,9,10,11,12,13,14



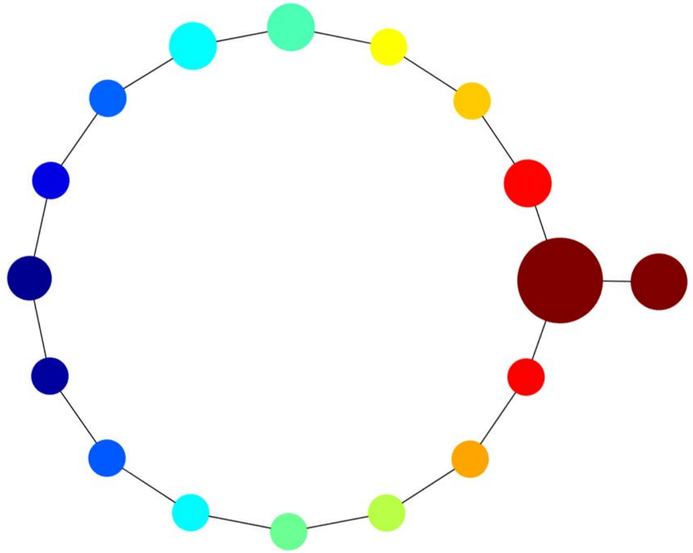
Topology extraction

Filter Range : 0 ~ 2
Intervals : 5
Overlap : 50%
Magic Fudge : 5

Filter



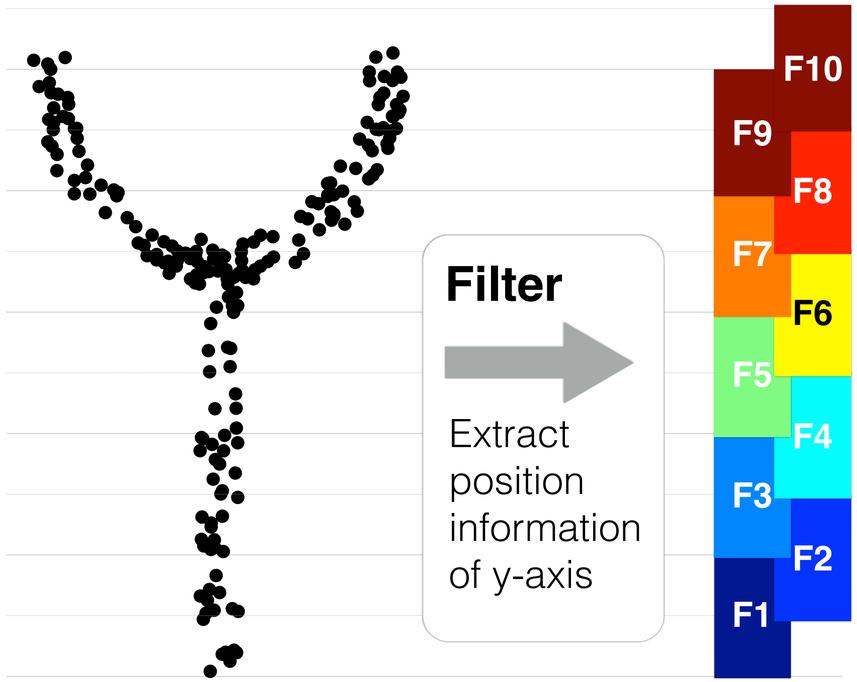
Partial Clustering



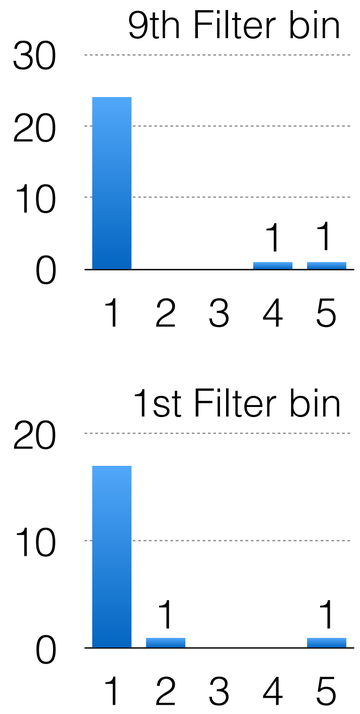
The size of node represents the number of point clouds in each node.

Topology of Y-shape points cloud

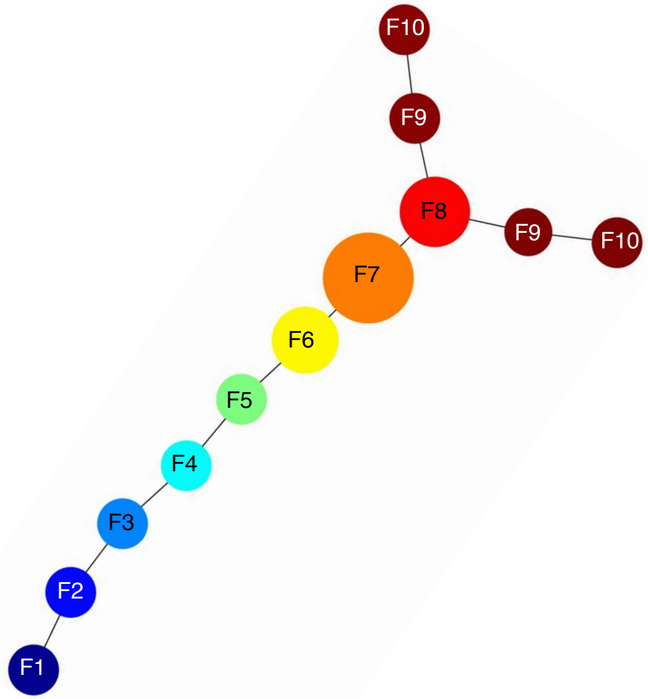
A)



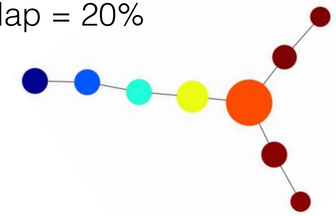
B)



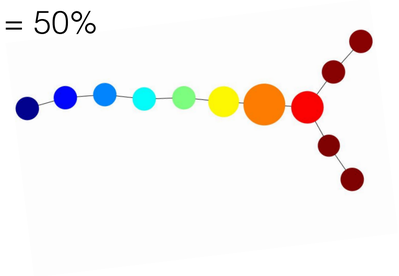
C)



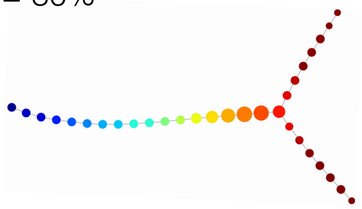
interval = 5
overlap = 20%



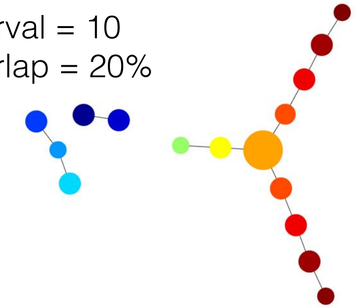
interval = 5
overlap = 50%



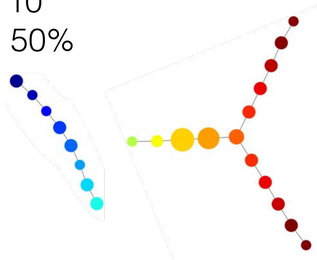
interval = 5
overlap = 80%



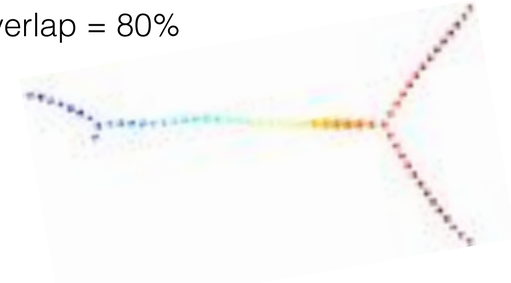
interval = 10
overlap = 20%



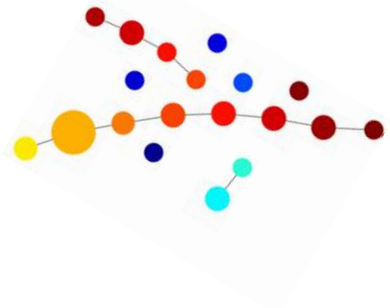
interval = 10
overlap = 50%



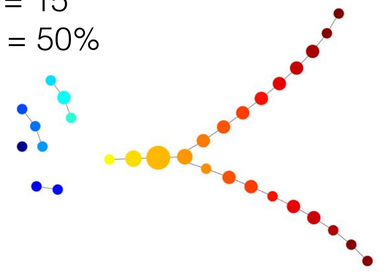
interval = 10
overlap = 80%



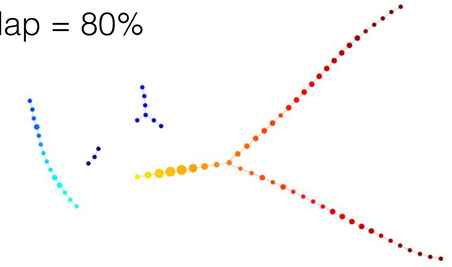
interval = 15
overlap = 20%



interval = 15
overlap = 50%



interval = 15
overlap = 80%



Mathematical details can be found at

- Gurjeet Singh et al. (2007), Topological methods for the analysis of high dimensional data sets and 3D object recognition.
- Gunnar Carlsson (2009), Topology and data, Bull. Amer. Math. Soc. 46. 255-308.

Gurjeet Singh and Gunnar Carlsson are co-founders of Ayasdi

AYASDI

TECHNOLOGY | PRODUCTS | APPROACH | INDUSTRIES | CUSTOMERS | COMPANY

Ayasdi Core

Discover The Critical Intelligence In Your Data

US Patent by Ayasdi

(12) **United States Patent**
Carlsson et al.

(10) **Patent No.: US 8,972,899 B2**
(45) **Date of Patent: Mar. 3, 2015**

(54) **SYSTEMS AND METHODS FOR VISUALIZATION OF DATA ANALYSIS**

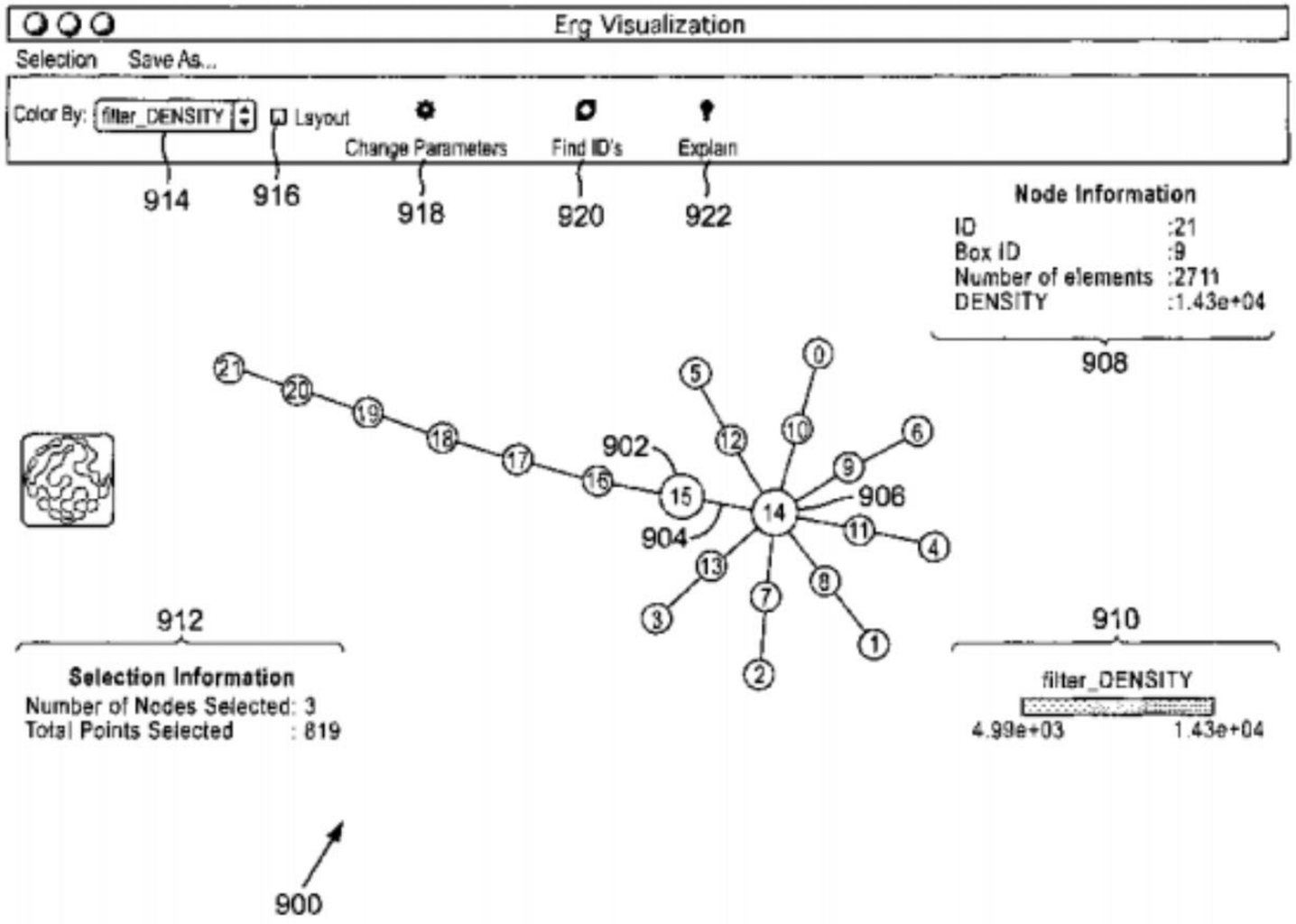
(75) Inventors: **Gunnar Carlsson**, Stanford, CA (US); **Harlan Sexton**, Palo Alto, CA (US); **Gurjeet Singh**, Menlo Park, CA (US)

(73) Assignee: **Ayasdi, Inc.**, Menlo Park, CA (US)

(57) **ABSTRACT**

Exemplary systems and methods for visualization of data analysis are provided. In various embodiments, a method comprises accessing a database, analyzing the database to identify clusters of data, generating an interactive visualization comprising a plurality of nodes and a plurality of edges wherein a first node of the plurality of nodes represents a cluster and an edge of the plurality of edges represents an intersection of nodes of the plurality of nodes, selecting and dragging the first node in response to a user action, and reorienting the interactive visualization in response to the user action of selecting and dragging the first node.

27 Claims, 11 Drawing Sheets


Erg Visualization
Selection
Save As...
Color By:
filter_DENSITY
Layout
Change Parameters
Find ID's
Explain
914
916
918
920
922
Node Information
ID :21
Box ID :9
Number of elements :2711
DENSITY :1.43e+04
908
902
904
906
912
Selection Information
Number of Nodes Selected: 3
Total Points Selected : 819
910
filter_DENSITY
4.99e+03
1.43e+04
900

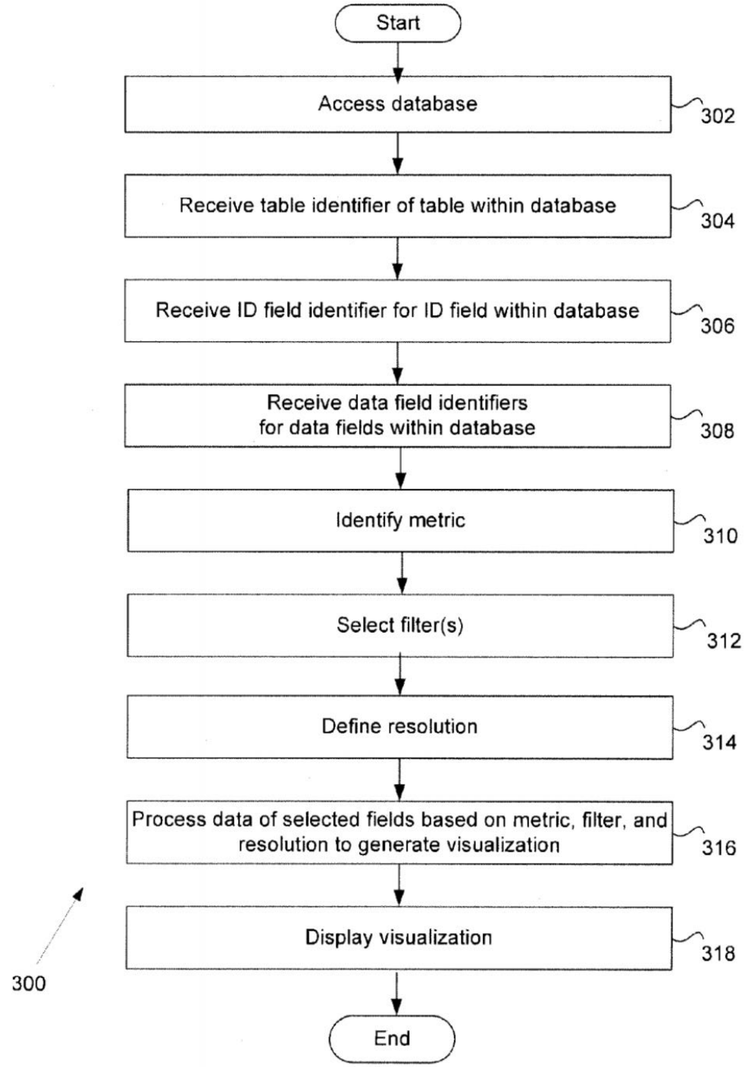

Start
Access database
302
Receive table identifier of table within database
304
Receive ID field identifier for ID field within database
306
Receive data field identifiers for data fields within database
308
Identify metric
310
Select filter(s)
312
Define resolution
314
Process data of selected fields based on metric, filter, and resolution to generate visualization
316
Display visualization
318
End
300

FIG. 3

Applications to Neurobiological/Clinical Data

International Neuroimaging Data-sharing Initiative (INDI)

- URL: http://fcon_1000.projects.nitrc.org/index.html
- Healthy control, ADHD, ASD data sets are available.
- resting state fMRI, diffusion tensor imaging,
- phenotype information such as intelligence scale and ADHD symptom severity are available.

Dataset : ADHD Symptoms & IQ

Data set:

	A	H	I	J	K	L	M
1	ScanDir ID	ADHD Index	Inattentive	Hyper/Impulsive	VIQ	PIQ	FSIQ-IV
2	0010001	90	90	80	106	91	99
3	0010002	66	65	62	65	89	75
4	0010003	42	42	43	107	93	100
5	0010005	63	59	70	98	118	108

$d(1,2)$

(or Euclidean)

L_2-Distance (or Euclidean distance) for all pairwise subjects:

$$d(1,2) = \sqrt{(90-66)^2 + (90-65)^2 + (80-62)^2 + (106-65)^2 + (91-89)^2 + (99-75)^2}$$

$$d(1,3) = \sqrt{(90-42)^2 + (90-42)^2 + (80-43)^2 + (106-107)^2 + (91-93)^2 + (99-100)^2}$$

Distance Matrix:

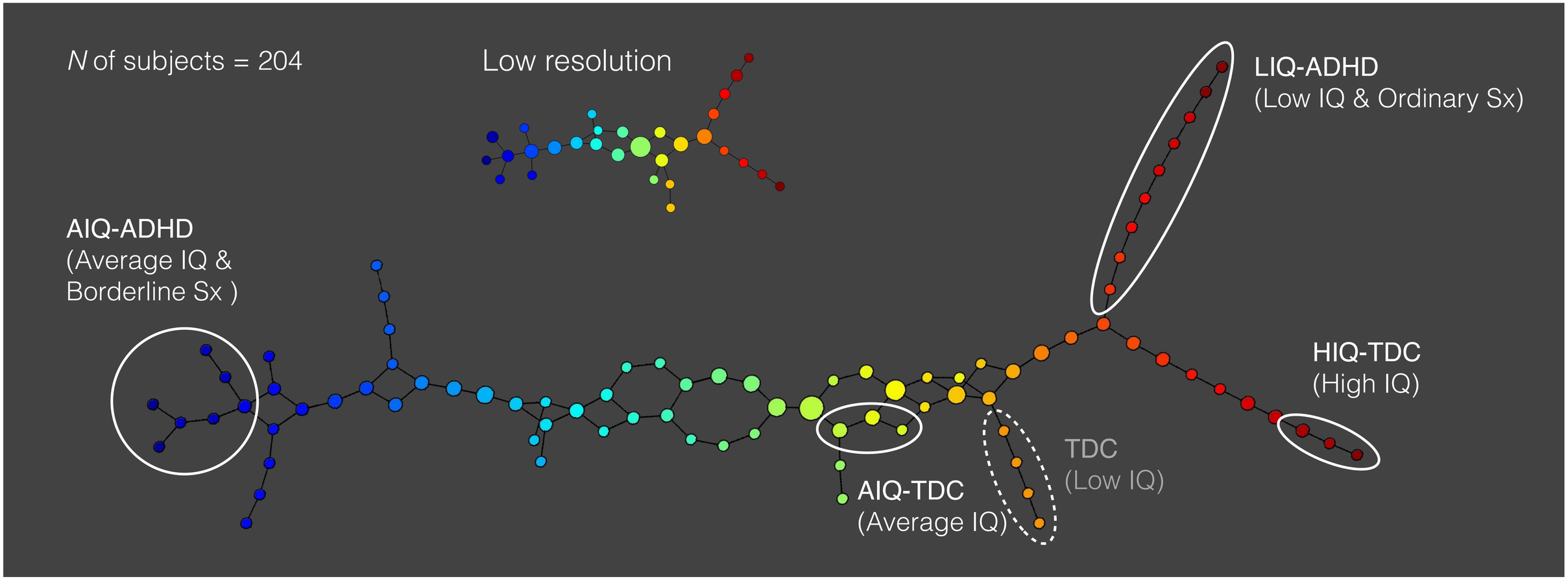
	1	2	3	4	5
1	0.0	61.5	77.3	51.6	77.0
2	61.5	0.0	62.2	55.9	69.0
3	77.3	62.2	0.0	47.2	11.9
4	51.6	55.9	47.2	0.0	52.4
5	77.0	69.0	11.9	52.4	0.0

Filter Function: L-infinity eccentricity

$$f(x) = \max_{y \in x} d(x,y)$$

$$f(x) = [77.3, 69.0, 77.3, 55.9, 77.0]$$

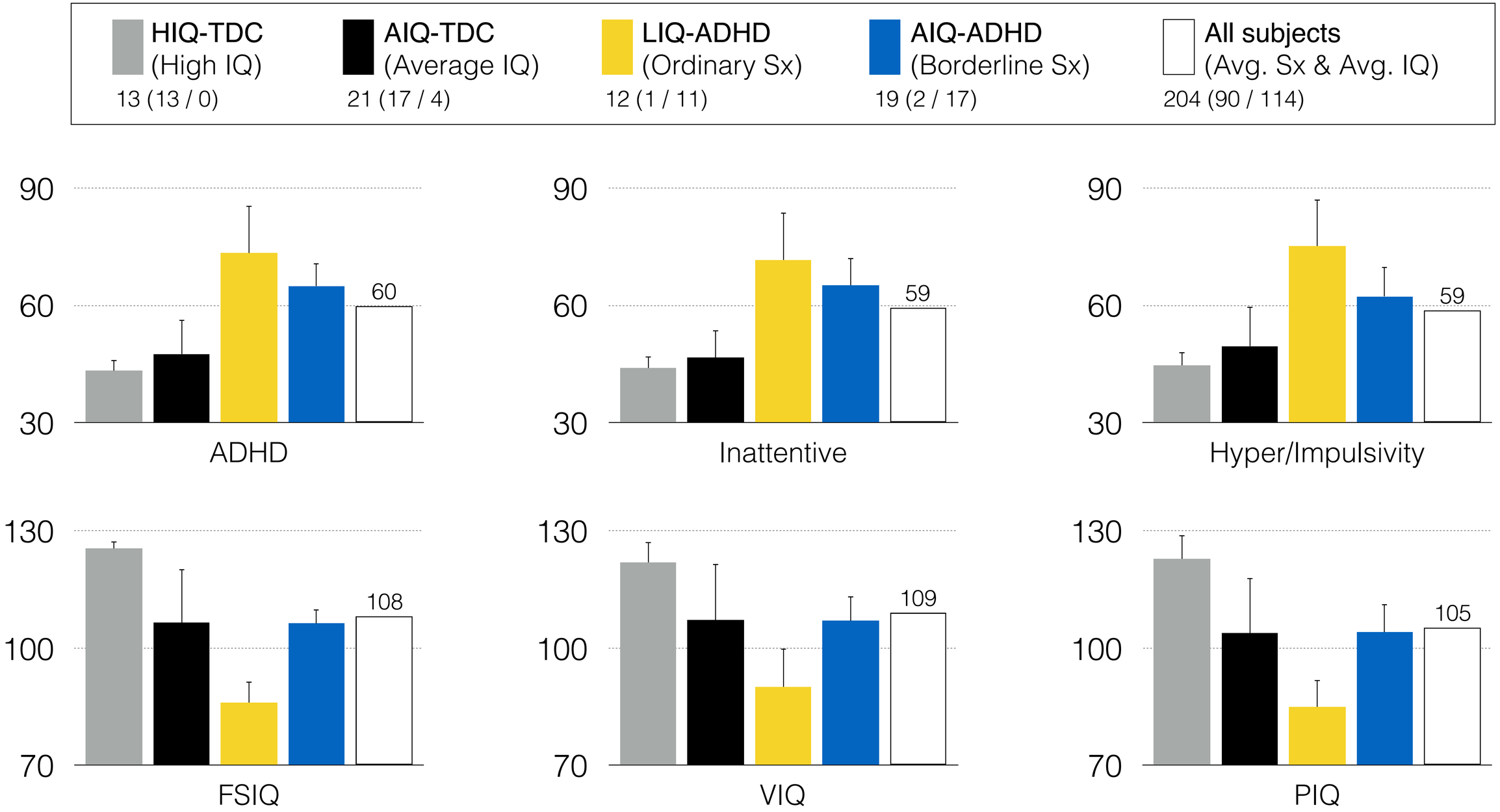
Phenotypic subgroups of ADHD



Distance function: L_2-distance

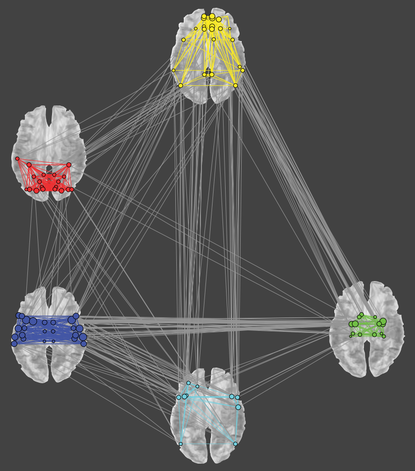
Filter function: L-infinity eccentricity

Low — High

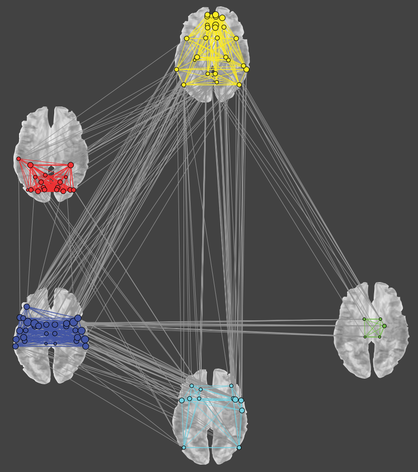

HIQ-TDC
(High IQ)
13 (13 / 0)
AIQ-TDC
(Average IQ)
21 (17 / 4)
LIQ-ADHD
(Ordinary Sx)
12 (1 / 11)
AIQ-ADHD
(Borderline Sx)
19 (2 / 17)
All subjects
(Avg. Sx & Avg. IQ)
204 (90 / 114)
90
60
30
60
ADHD
90
60
30
59
Inattentive
90
60
30
59
Hyper/Impulsivity
130
100
70
108
FSIQ
130
100
70
109
VIQ
130
100
70
105
PIQ

Functional Modular Architectures

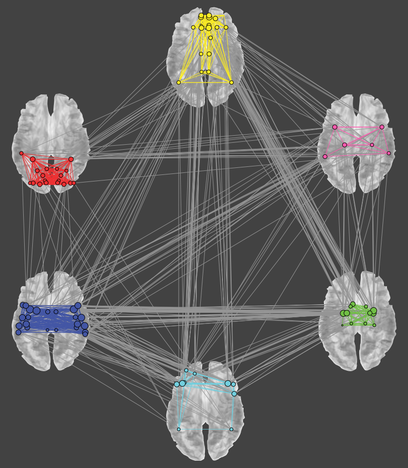
HIQ-TDC
(TDC with High IQ)



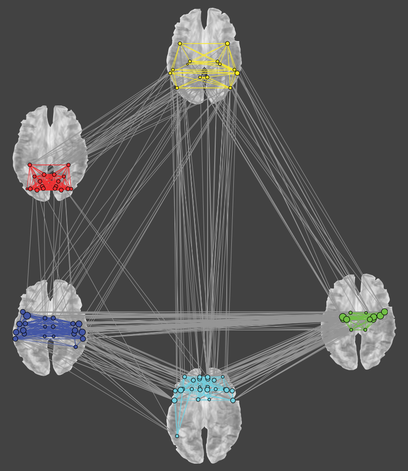
AIQ-TDC
(TDC with Average IQ)

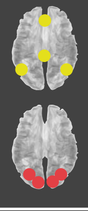


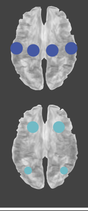
LIQ-ADHD
(ADHD with Ordinary Sx.)

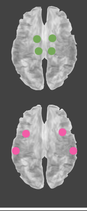


AIQ-ADHD
(ADHD with Borderline Sx.)




DMN module


Central module


BG/THL module

Occipital module

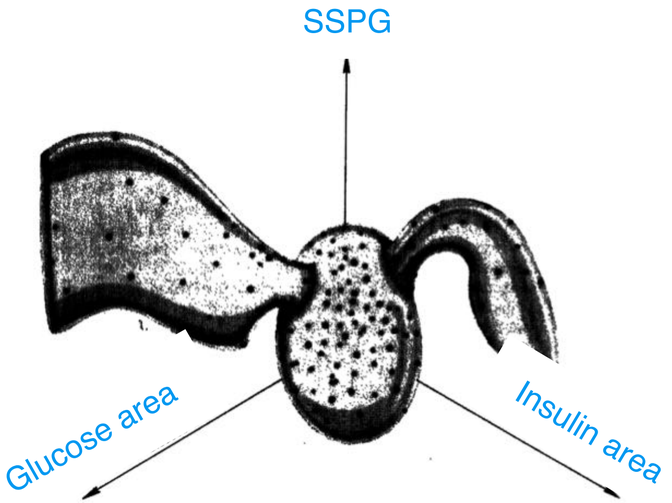
Parietofrontal module

Temporal/Limbic module

※ Only positive weights of the functional connectivity are used for module analysis and visualised.

Applications to
Medical science data

Diabetes Subtypes



based on six quantities:

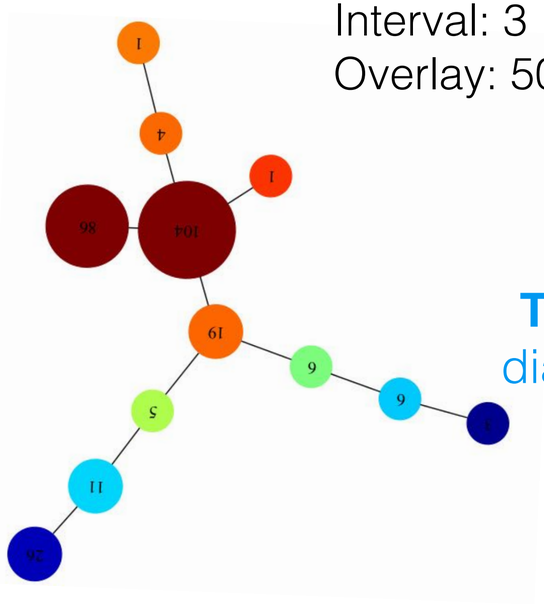
- age
- relative weight
- fasting plasma glucose
- area under the plasma glucose curve for the three hour glucose tolerance test (OGTT)
- area under the plasma insulin curve for the OGTT
- steady state plasma glucose (SSPG) response

Ref) Eurographics Symposium on Point-Based Graphics, Singh G *et al.* (2007)

Subtypes of diabetes

Low-resolution

Interval: 3
Overlay: 50%

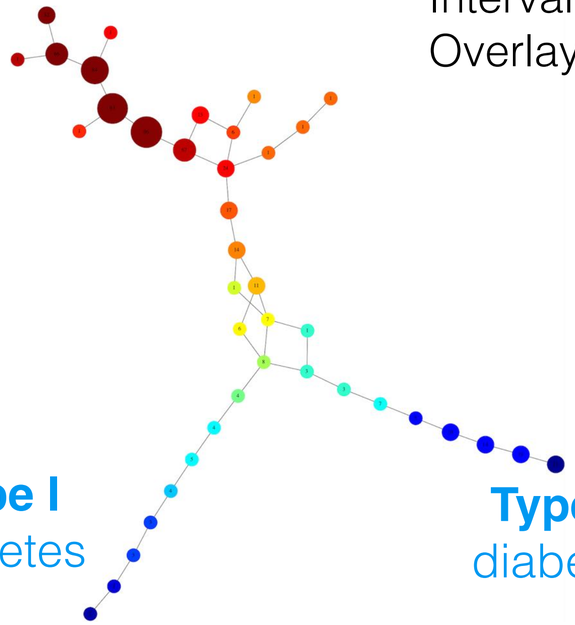


Type II diabetes

Type I diabetes

High-resolution

Interval: 4
Overlay: 80%



Type I diabetes

Type II diabetes

Type I: adult onset **TypeII:** juvenile onset

Distance function: L2-distance

Filter function: density kernel with e=130,000

$$f_\varepsilon(x) = C_\varepsilon \sum_y \exp\left(\frac{-d(x,y)^2}{\varepsilon}\right)$$

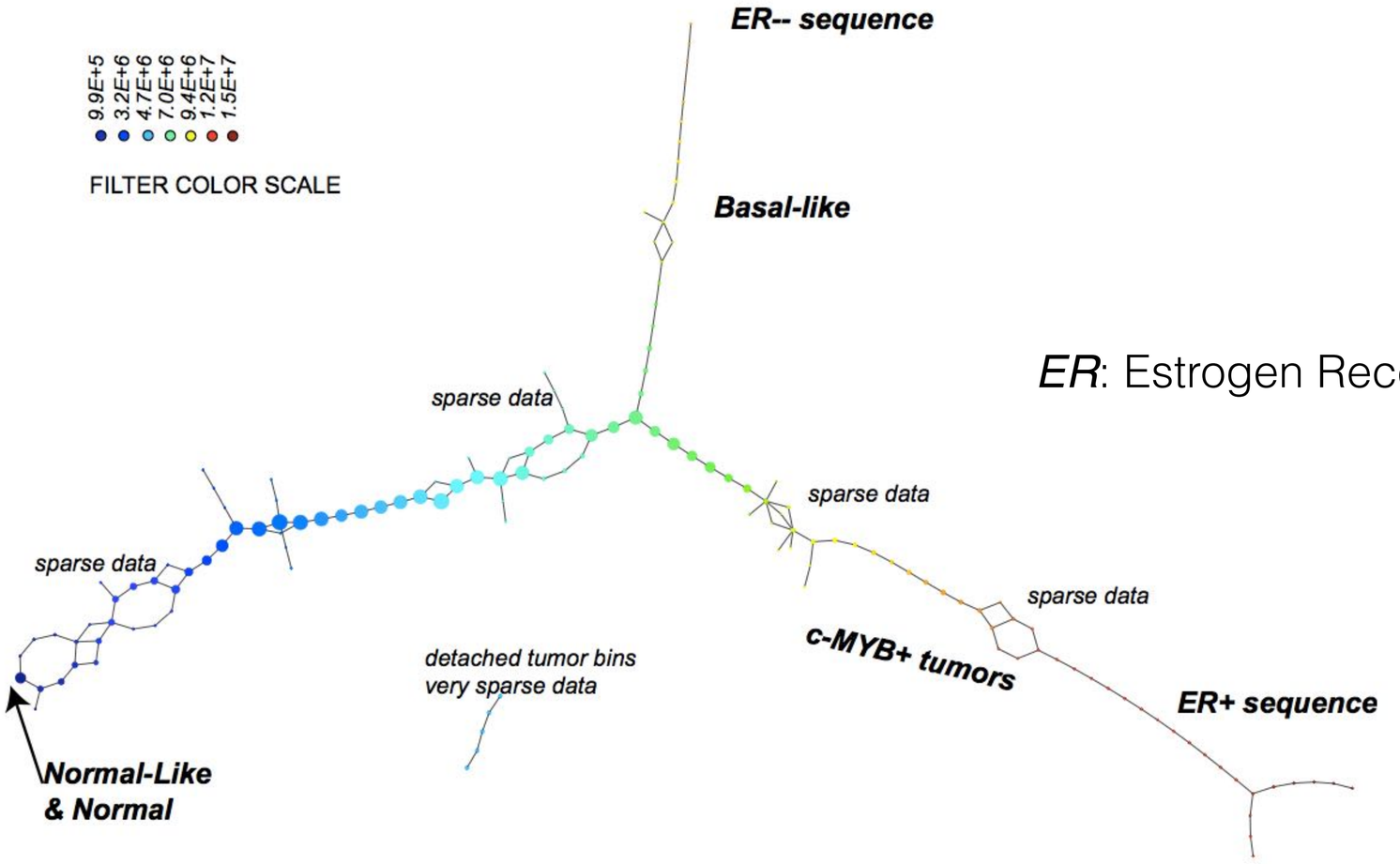
Application to **Biological Data**,

Subtypes of Breast Cancer

using breast cancer microarray gene expression data set
- disease specific genomic analysis (DSGA) transformed data

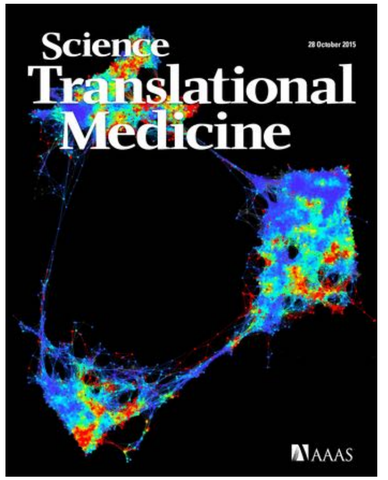
PNAS, Monica **Nicolau** *et al.* (2010)

Breast Cancer Subtype



ER: Estrogen Receptor

PNAS, Monica **Nicolau** *et al.* (2010)



RESEARCH ARTICLE

PRECISION MEDICINE

Identification of type 2 diabetes subgroups through topological analysis of patient similarity

Li Li,[1] Wei-Yi Cheng,[1] Benjamin S. Glicksberg,[1] Omri Gottesman,[2] Ronald Tamler,[3] Rong Chen,[1] Erwin P. Bottinger,[2] Joel T. Dudley[1,4]*

Diabetes Mellitus (DM)

- Commonly referred to as diabetes, a group of metabolic disease.
- If left untreated, diabetes can cause many complications:
  cardiovascular disease, stroke, chronic kidney failure, foot ulcers, and damage to eyes.
- Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced.
- **Type 1 DM**: results from the pancreas's failure to produce enough insulin.
- **Type 2 DM**: begins with insulin resistance, a condition in which cells fail to respond to insulin properly. heavy weight and not enough exercise are causes of T2D.
- Gestational diabetes, is the third main form and occurs when pregnant women without a previous history of diabetes develop a high blood-sugar level.

Why subgroups of type 2 DM?

- Risk factors of Type 2 DM are:
  obesity, family history of diabetes, physical inactivity, ethnicity, and advanced age.
- Type 2 DM is heterogenous complex disease affecting
  more than 29 million in American (9.3%). 2 million in Korea.
- Increasing needs for early prevention and clinical management of Type 2 DM.

Methods

- High dimensional EMRs and genotype data from 11,210 individuals from Mount Sinai Medical Center (MSMC)'s outpatient population (46% Hispanic, 32% African american, 20% European white, and 2% others).

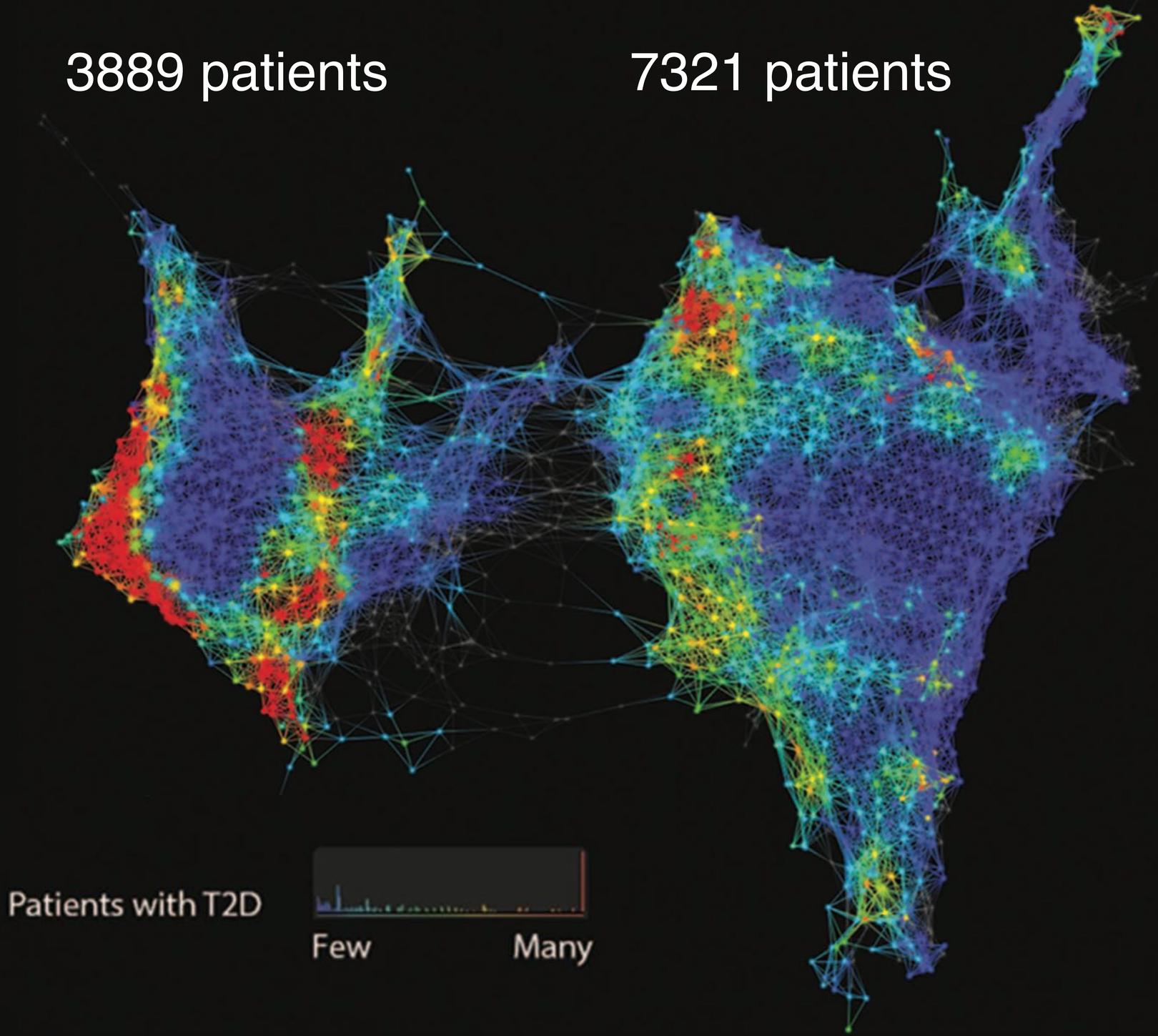
- Type 2 DM and non-Type 2 DM were defined by an electronic phenotyping algorithm (eMERGE network) based on ICM-9-CM diagnosis codes, laboratory test, prescribed medications (RxNorm), physician notes (natural language processing), and family history.

- Form of preprocessed data matrix is ***n* patients by *P* medical variables**.
  Medical variables included 505 clinical variable, 7097 unique ICM-9-CM codes (1 to 218 per patients). On average, 64 clinical variables per patients. To avoid overfitting, select the variable with at least 50% of patients who had the variables, resulting in 73 (of 505) variables to perform the analysis.

TDA pipeline

- Distance metric: cosine distance metric was used to assess the similarity of the data points.

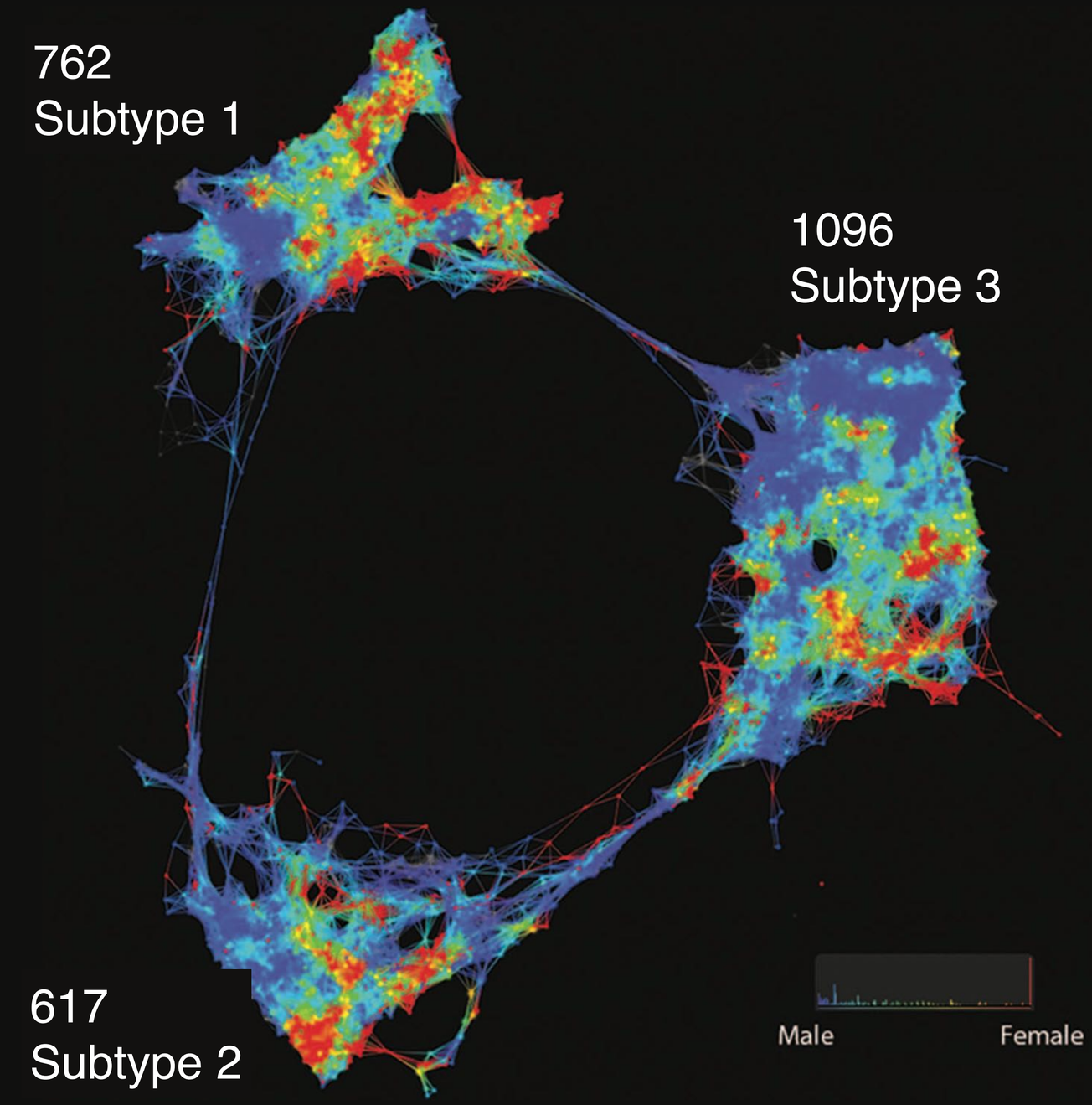
- Filter metric: L-infinity centrality and principal metric singular value decomposition (SVD1).
  L-infinity centrality is defined for each data point y to be the maximum distance from y to any other data point in the data set. Large values of this function correspond to points that are far from the centre of data set.


3889 patients
7321 patients
Patients with T2D
Few
Many

TDA with only 2551 Type II DM

Gender is not an organising factor in the topology

Reproducibility test (2/3 training, 1/3 test data set, 10 times) revealed that the overall accuracy was 96% for a subtype classification.



Applications to
Social Science Data

- Classification of (basket ball) player types
- Partial clustering of personality using temperamental traits
- Relationship among welfare, civil construction, and suicide rate

Extracting insights from the shape of complex data using topology

P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

SCIENTIFIC REPORTS

Classification of player types

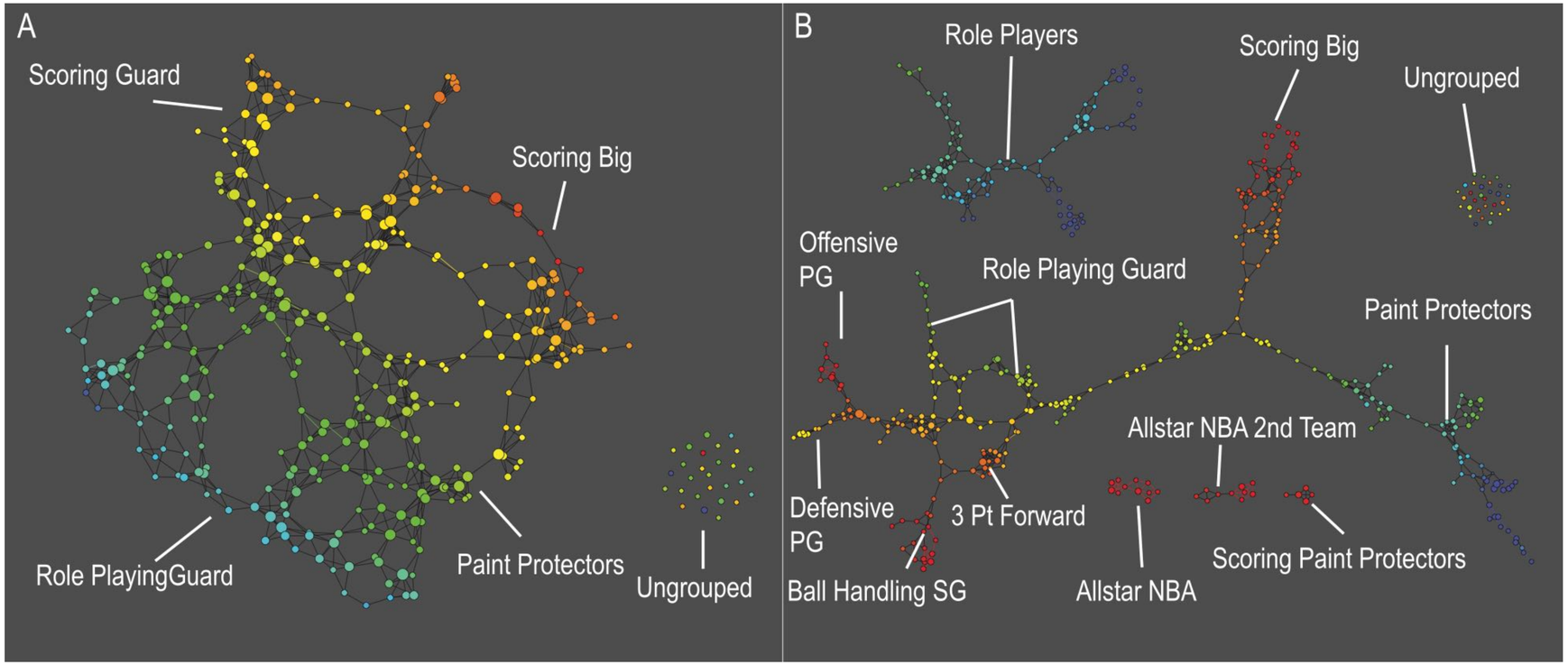
based on their in-game performance such as:

- rates (per minute played) of rebounds, assists, turnovers, steals, blocked shots, personal fouls, and points scored (7 performance measures)

Map of Players

low resolution map at 20 intervals

high resolution map at 30 intervals



Traditionally, basket ball players are categorised into guards, forwards, and center.

Points Per Game

Distance function:

Variance normalised L_2-distance

Filter function:

Principal and secondary SVD values

TDA versus K-means clustering

Grouping subjects into personality types

Novelty seeking

Reward dependance

Harm avoidance

Persistence

Name	NS	HA	RD	P	SD	C	ST
subj001	28	57	44	30	30	42	27
subj002	52	32	45	54	55	48	46
subj003	37	33	47	36	46	59	25
subj004	36	26	43	53	54	40	20
subj005	52	45	35	59	46	64	31
subj006	49	39	47	48	38	44	18
subj007	39	46	42	34	34	50	22
subj008	38	41	52	51	49	60	10
subj009	26	48	40	37	39	48	15
subj010	38	25	47	58	56	55	22
subj012	30	67	17	13	19	38	13
subj013	43	26	58	50	53	60	30
subj014	41	34	59	62	60	74	48
subj015	38	51	36	30	19	56	9
subj016	30	47	39	43	41	55	14
subj017	32	65	58	49	33	72	40
subj018	57	23	53	52	58	72	29

Temperament and character inventory (TCI) scores from 40 normal subjects

Subject Grouping

The goal of ***k*-means clustering** is to minimise the within-cluster sum of squares.

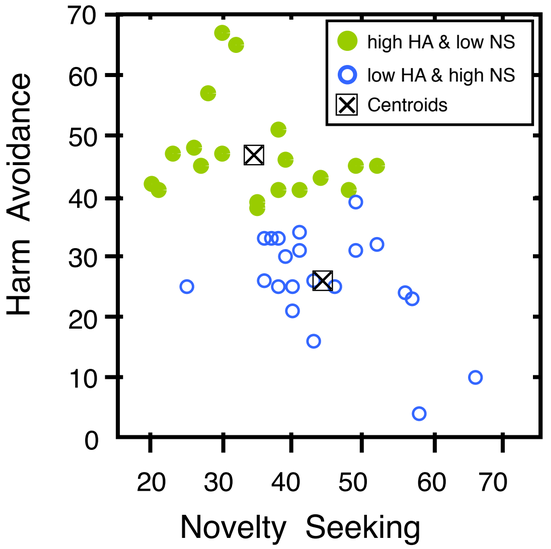
$$V = \sum_{i=1}^{2} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \quad S = \{S_1, S_2\}$$

μ_i: the mean of points in S_i

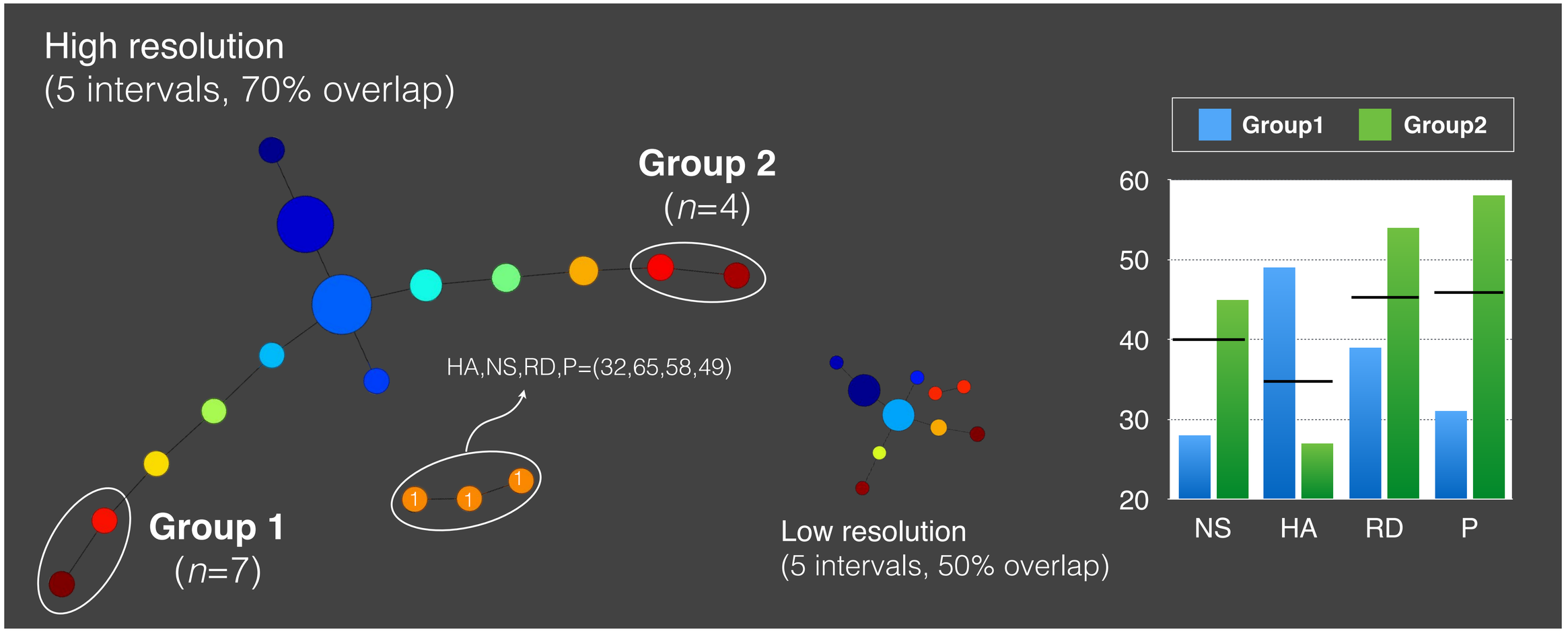
A. Input Dataset

Name	NS	HA	RD	P	SD	C	ST
subj001	28	57	44	30	30	42	27
subj002	52	32	45	54	55	48	46
subj003	37	33	47	36	46	59	25
subj004	36	26	43	53	54	40	20
subj005	52	45	35	59	46	64	31
subj006	49	39	47	48	38	44	18
subj007	39	46	42	34	34	50	22
subj008	38	41	52	51	49	60	10
subj009	26	48	40	37	39	48	15
subj010	38	25	47	58	56	55	22
subj012	30	67	17	13	19	38	13
subj013	43	26	58	50	53	60	30
subj014	41	34	59	62	60	74	48
subj015	38	51	36	30	19	56	9
subj016	30	47	39	43	41	55	14
subj017	32	65	58	49	33	72	40
subj018	57	23	53	52	58	72	29

B. *k*-means Clustering



TDA to extract personality groups



Distance function: L_2-distance

Filter function: L-infinity eccentricity

Low High

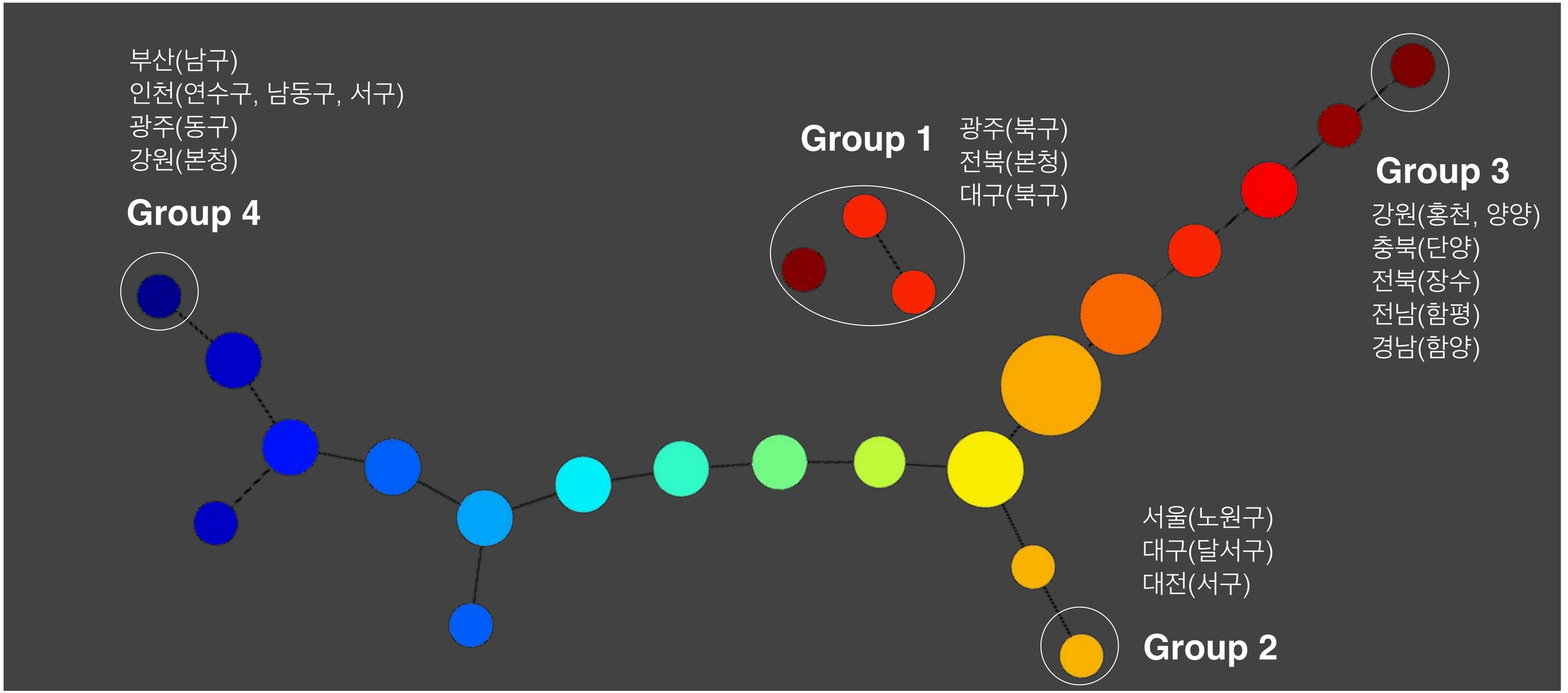
Welfare / civil engineering / suicide ratio

	A	B	C	D	E	F	G	H	I	J
1	순번	광역단체	기초단체	2009년 토건예산	2012년 토건예산	토건예산증감(%p)	2009년 복지예산	2012년 복지예산	복지예산증감(%p)	2012년 10만명 당 자살자수(연령표준화)
2	1	서울	본청	21.41%	12.71%	-8.70%	22.1%	26.6%	4.5%	21.2
3	2	서울	종로구	36.41%	19.97%	-16.44%	18.6%	26.3%	7.7%	14.2
4	3	서울	중구	33.54%	18.50%	-15.04%	23.6%	28.6%	5.0%	25.3
5	4	서울	용산구	34.44%	15.04%	-19.40%	22.1%	34.9%	12.8%	23
6	5	서울	성동구	27.87%	19.95%	-7.92%	23.6%	33.0%	9.4%	23.1
7	6	서울	광진구	26.49%	11.83%	-14.66%	31.6%	37.3%	5.7%	16
8	7	서울	동대문구	33.97%	13.55%	-20.42%	29.1%	38.4%	9.3%	24.9
9	8	서울	중랑구	23.90%	8.08%	-15.82%	39.7%	46.9%	7.2%	23.4
10	9	서울	성북구	23.76%	9.78%	-13.98%	33.7%	43.5%	9.8%	19.3
11	10	서울	강북구	29.49%	10.65%	-18.84%	40.7%	47.7%	7.0%	23
12	11	서울	도봉구	23.71%	11.63%	-12.08%	38.2%	44.1%	5.9%	22.3
13	12	서울	노원구	24.97%	8.00%	-16.97%	43.5%	53.9%	10.4%	23
14	13	서울	은평구	35.77%	9.43%	-26.34%	37.4%	49.6%	12.2%	22.4
15	14	서울	서대문구	33.47%	13.08%	-20.39%	30.7%	36.5%	5.8%	23.7
16	15	서울	마포구	26.41%	10.26%	-16.15%	35.4%	41.7%	6.3%	21.9
17	16	서울	양천구	22.90%	14.25%	-8.65%	32.8%	43.6%	10.8%	21.1
18	17	서울	강서구	22.28%	15.20%	-7.08%	44.0%	49.5%	5.5%	24.8
19	18	서울	구로구	22.15%	17.28%	-4.87%	34.6%	44.1%	9.5%	23.2
20	19	서울	금천구	29.19%	14.56%	-14.63%	35.4%	46.5%	11.1%	25.4
21	20	서울	영등포구	28.00%	14.16%	-13.84%	31.9%	35.7%	3.8%	19.7
22	21	서울	동작구	26.71%	13.62%	-13.09%	35.8%	42.7%	6.9%	20.1
23	22	서울	관악구	24.85%	7.90%	-16.95%	38.9%	46.0%	7.1%	21.7
24	23	서울	서초구	47.88%	27.72%	-20.16%	22.1%	27.2%	5.1%	13.3

Data Download: http://newstapa.com/news/201411935

Nation’s public data analysis

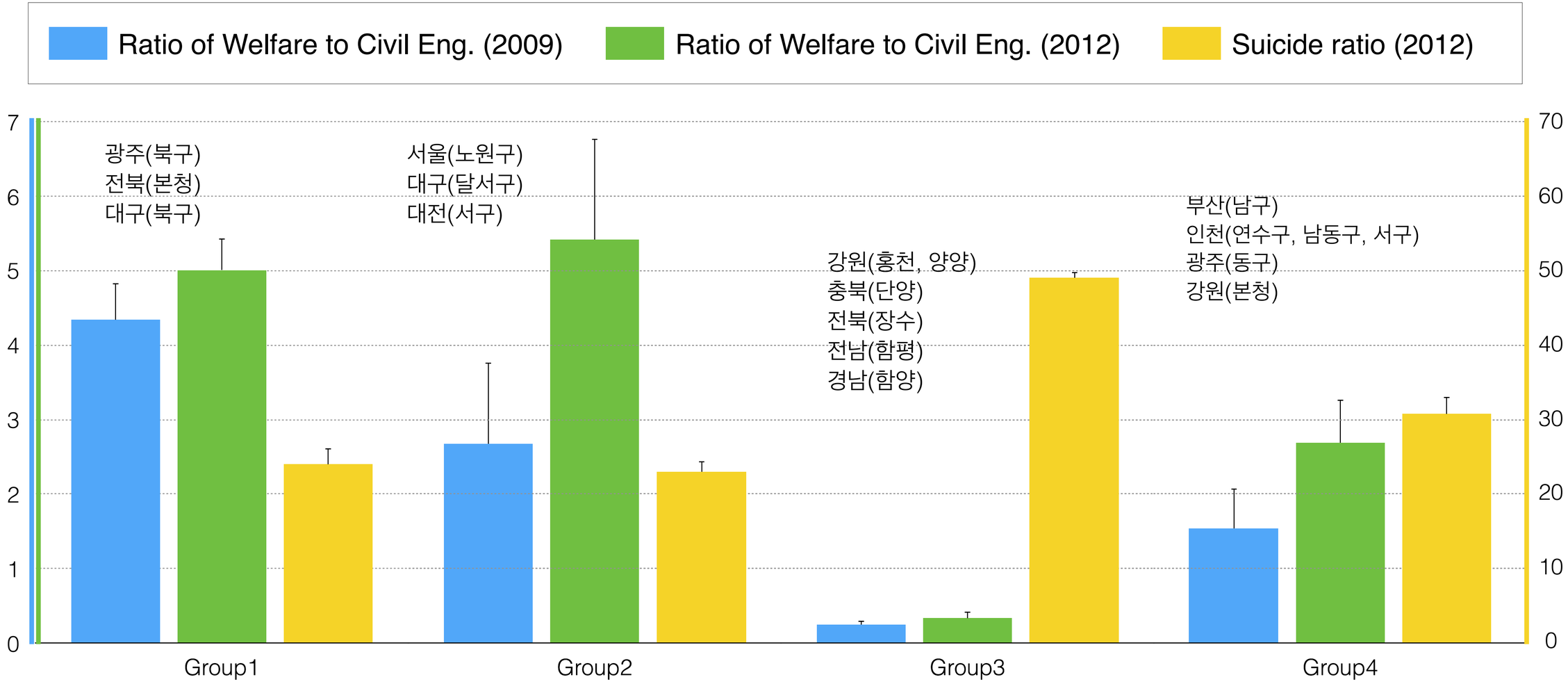
Input data: ratio of welfare to civil engineering (2009), ratio of welfare to civil engineering (2012), Suicide rate (2012)



Distance function: L_2-distance

Filter function: L-infinity eccentricity

Welfare I Civil Eng. I Suicide ratio



Blog posting: http://skyeong.tistory.com/136/

Topological Data Analysis,
new weapon for
discovering new **insight** from data.

Conclusion

- TDA can **be applied to various dataset** and has a coordinate free characteristic.
- Useful for analysing non-linear dataset.
- Selection and optimisation of distance and filter metric are important issue.
- TDA will be a powerful weapon for those who want to find a new insight from data
- It’s possible to make a supervised machine learning system using TDA.

References

1. Gurgeek Singh *et al.*, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, Eurographics Symposium on Point-Based Graphics, 2007.
2. Gunnar Carlsson, Topology and Data, Bull. Amer. Math. Soc. **46**(2):255-308, 2009.
3. Monica Nicolau *et al.*, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, PNAS **108**(17):7265-7270, 2011.
4. P. Y. Lum *et al.*, Extracting insights from the shapes of complex data using topology, Nature Scientific Reports **3**:1236, 2013.
5. Li Li *et al.*, Identification of type 2 diabetes subgroups through topological analysis of patients similarity, Science Translational Medicine **7**(311):311ra174, 2015.
6. Jessica L. Nielson *et al.*, Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury, Nature communications **6**:8581, 2015.
7. AYASDI, a commercial software for TDA, http://www.ayasdi.com/