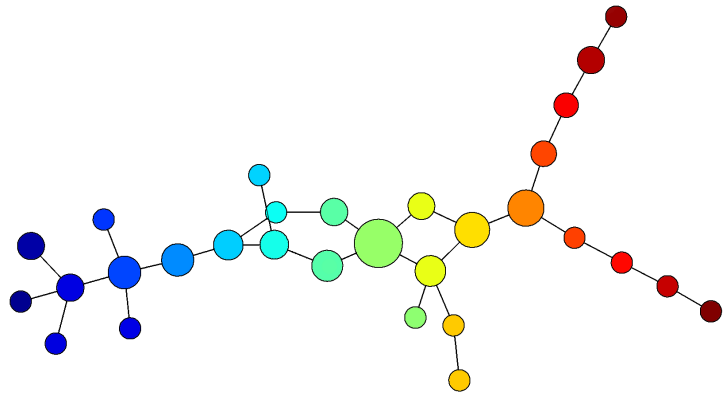


Topological Data Analysis

-Methods and Examples-

Sunghyon Kyeong

Institute of Behavioral Science in Medicine,

Yonsei University College of Medicine

Machine Learning

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

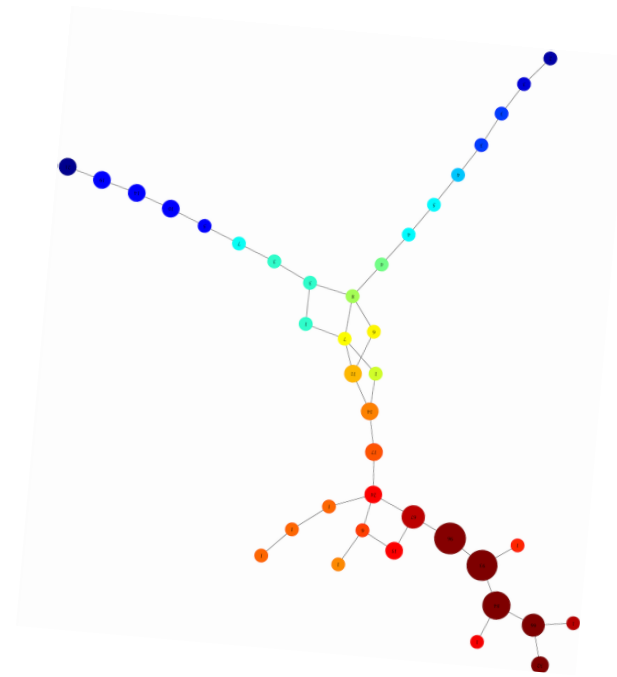
- Unsupervised Machine Learning
  → Clustering of given dataset / Community detection
  (k-means clustering, modularity optimisation, **TDA**, ICA, PCA, and so on)

Data has Shape

An example

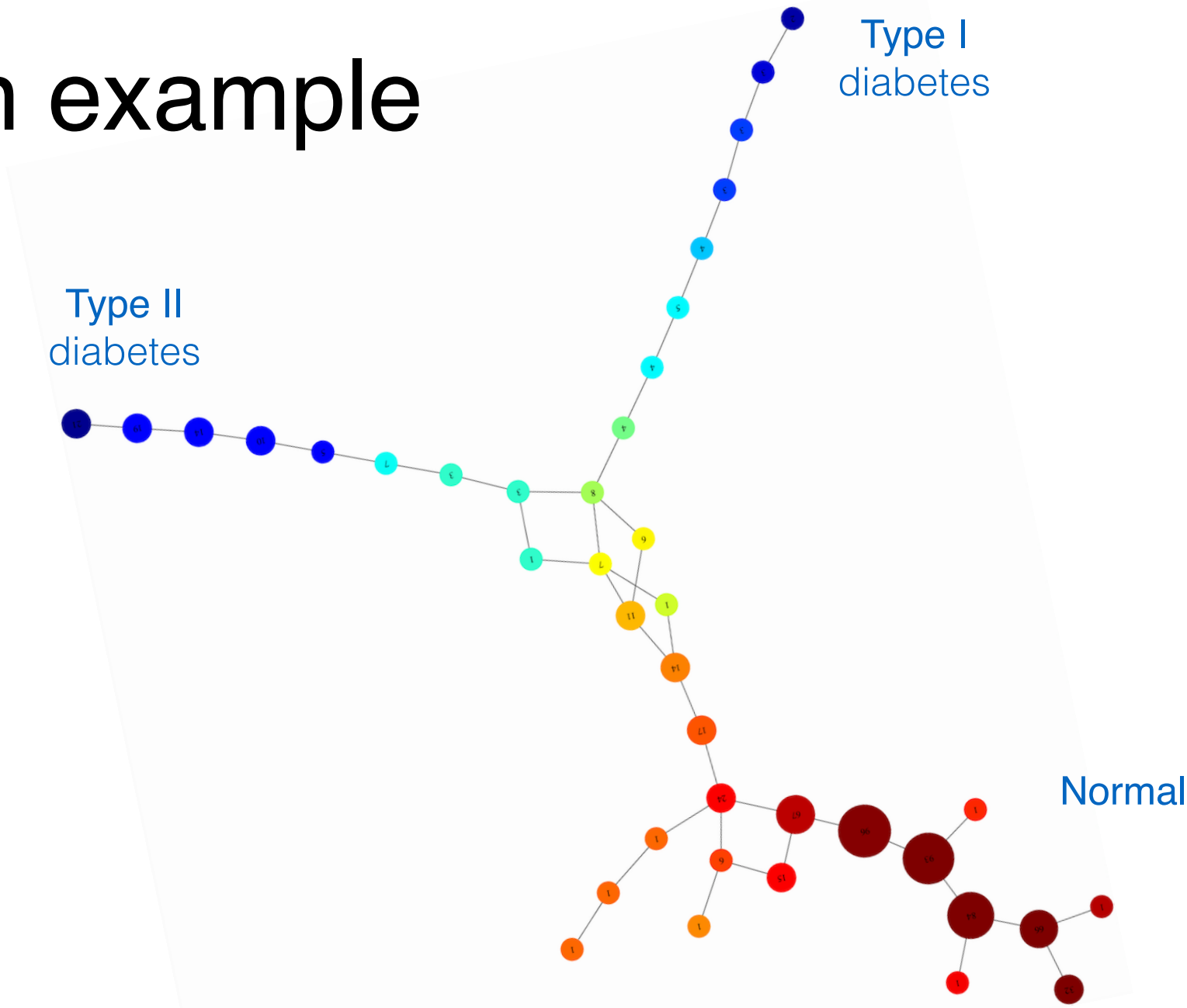
Raw Data
(diabetes related data)

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



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

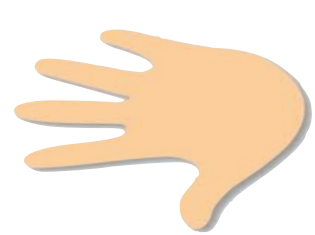
$Betti_0$: 8
$Betti_1$: 5

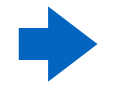
mathematically
defined “holes” in data

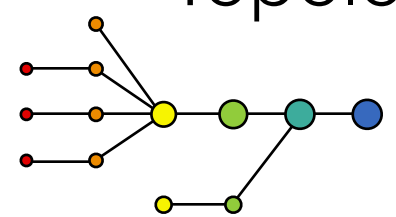
$Betti_0$: clusters
$Betti_1$: holes
$Betti_2$: voids

Geometric

Topology



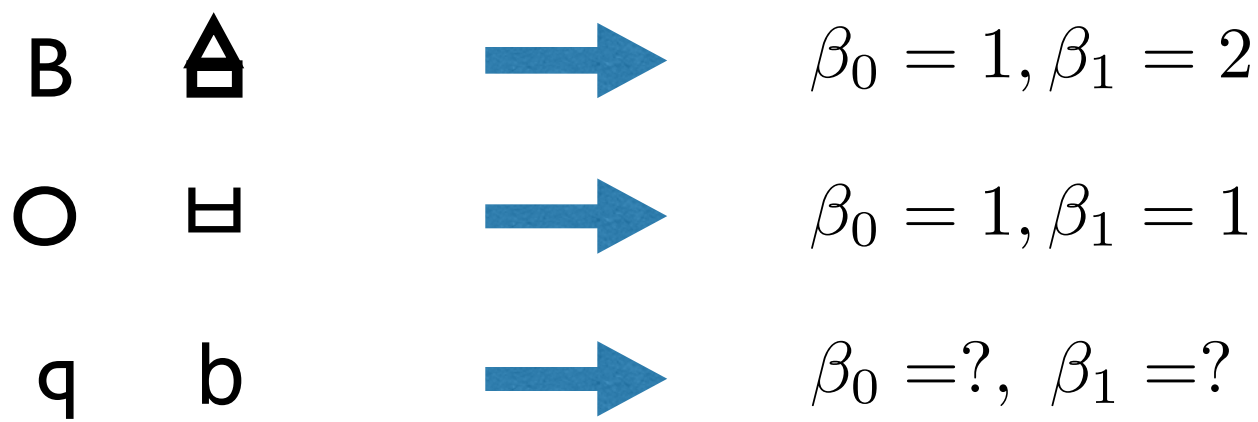




Algebraic Topology

Quantitive Information

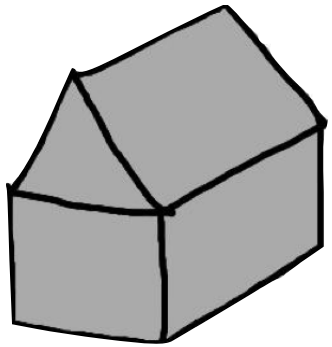
Persistent homology is a spatial type of homology that is useful for data analysis. **Betti numbers**, which come from computing **homology**, reflect the topological properties of an object.



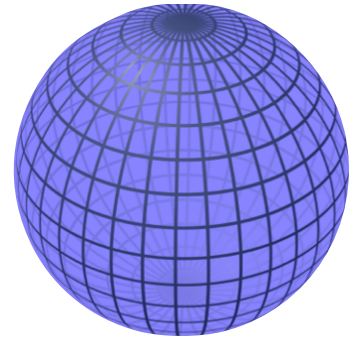
β_0 : the connected components

β_1 : the number of holes

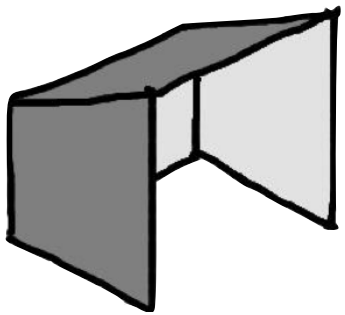
Homology



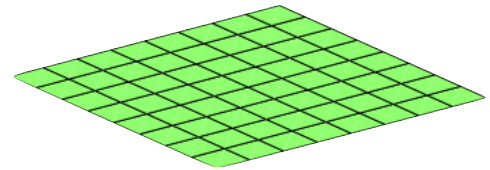
homeomorphic to



, $Betti_2 = 1$



homeomorphic to



, $Betti_2 = 0$

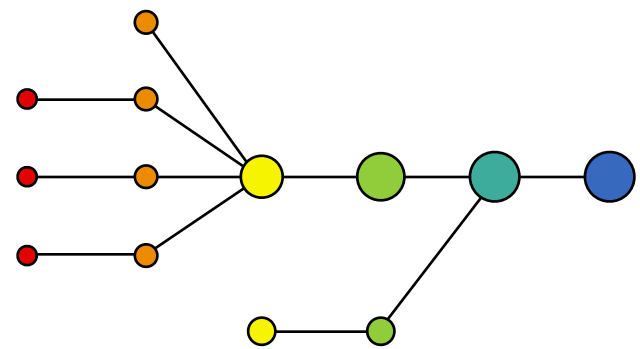
Ref) Xiaojin Zhu, IJCAI 2013 presentation slide

Geometric Topology

Extracting Shapes of Data

points cloud data

?



topology

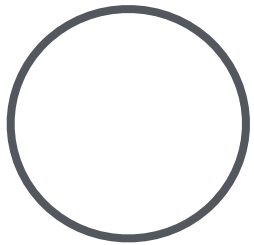
Ref) Figures are obtained from Y.P. Lum et al (2013) **Scientific Reports** | 3: 1236

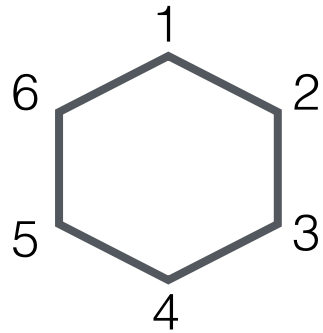
Properties of **TDA**

- Coordinate invariance

- Deformation invariance

- Compressed representations





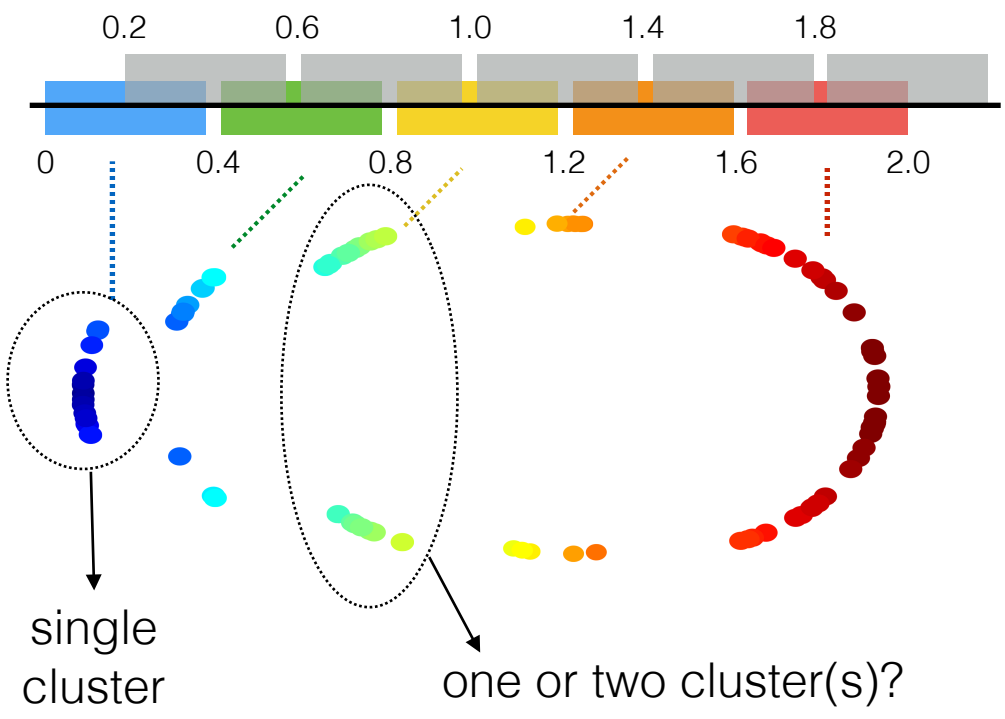
Topological Data Analysis using *Mapper*

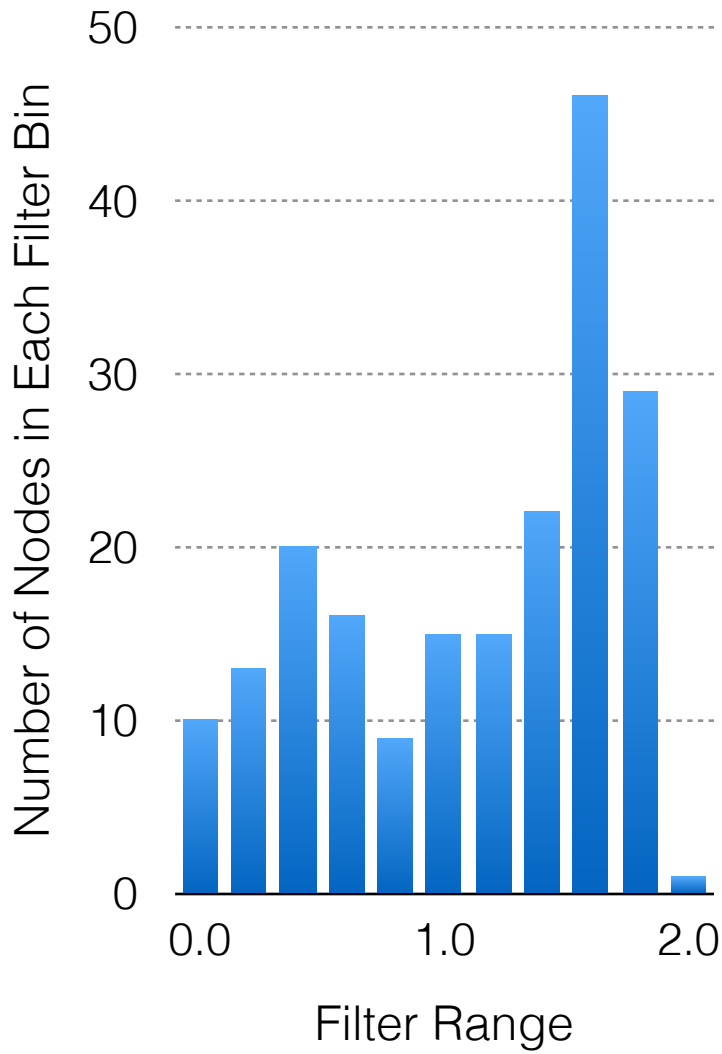
- Two input functions
  :**filter** as a measure of length of disease component
  :**distance** as a measure of distance between data points

- Resolution Parameters
  : *Intervals, overlap, magic fudge*

Filter : Divide point clouds into each filter bin

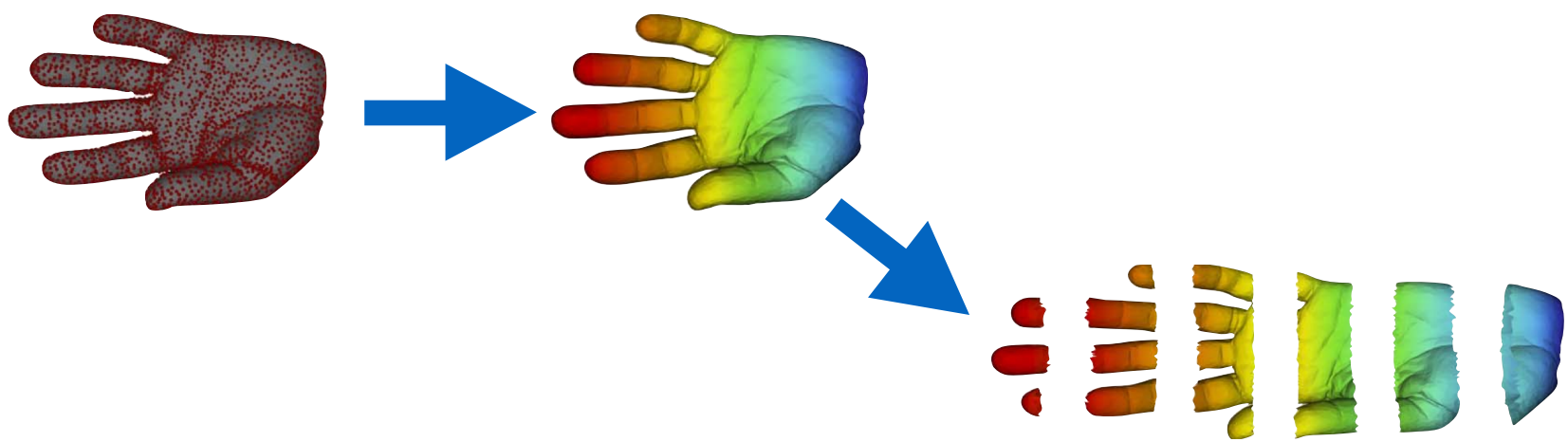
Filter Range : 0.0~2.0
Intervals : 5
Overlap : 50%
Interval Length : 0.4





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.

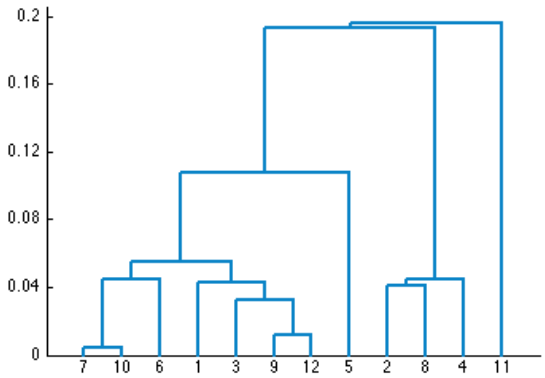


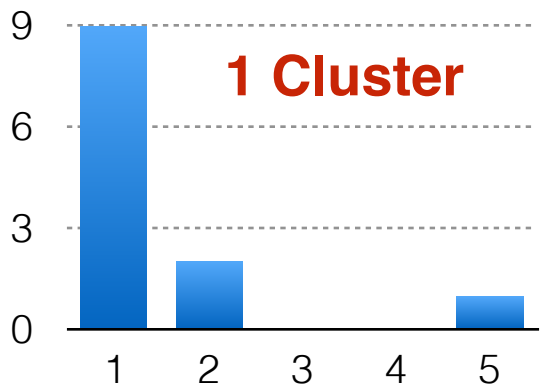
Distance Function

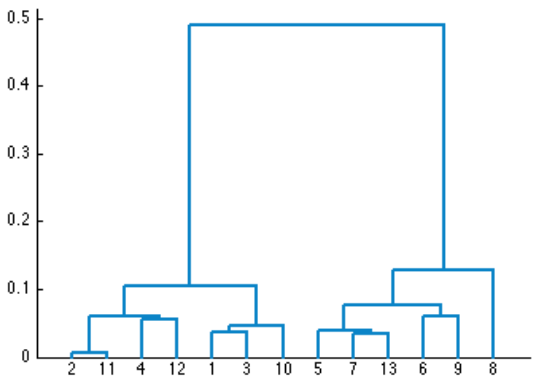
- distance between all pairs of data points.
- both euclidean or geodesic distances could be used.

Distance & Clustering

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



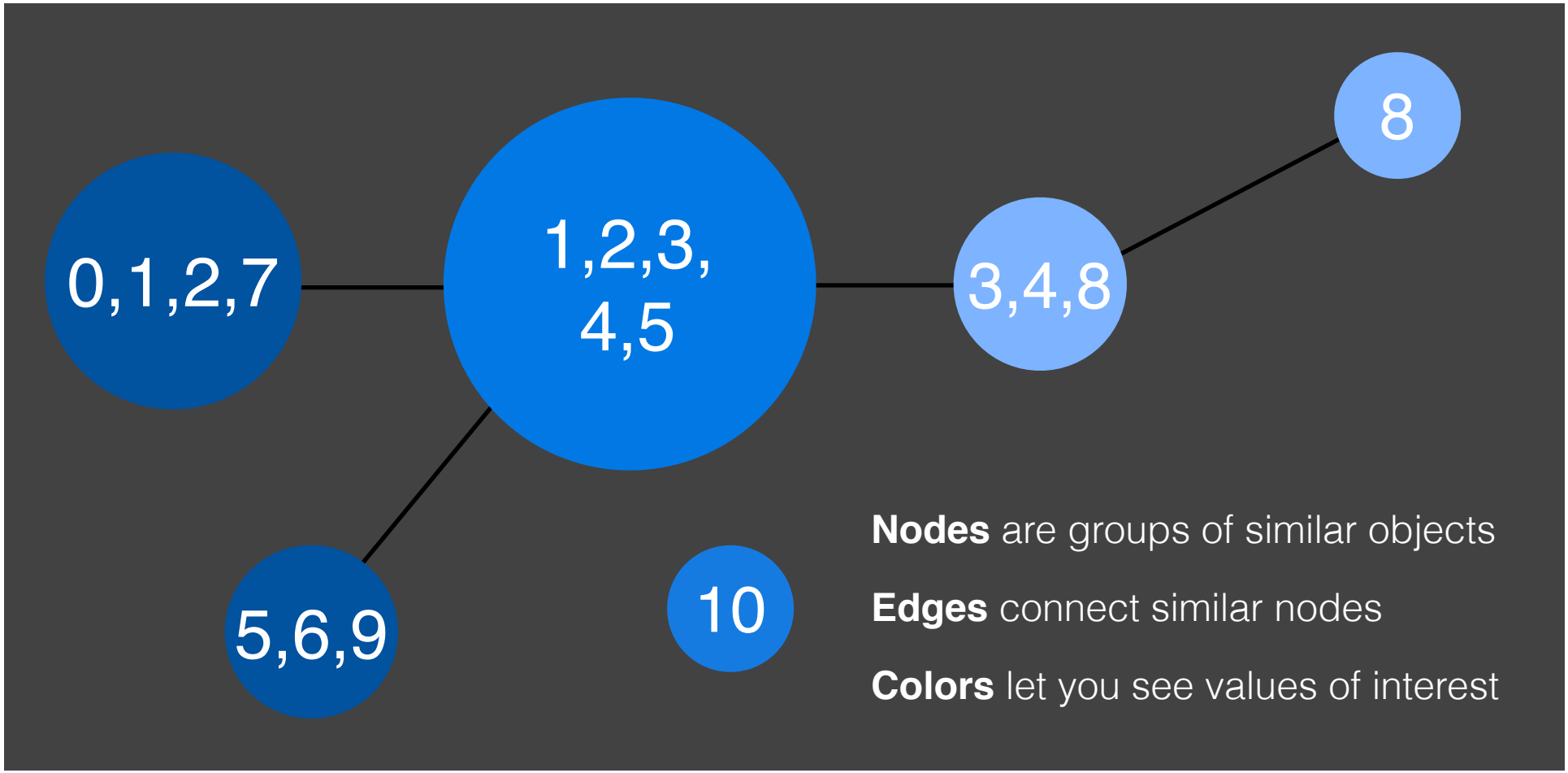




Magic Fudge is the number of bins in the distribution of the distance obtained from single linkage dendrogram.

N of clusters is estimated from the number of continuous bins having zero elements.

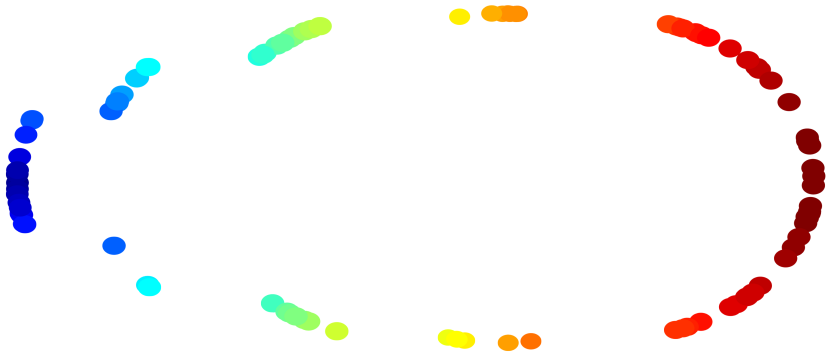
Nodes, Edges, Colors



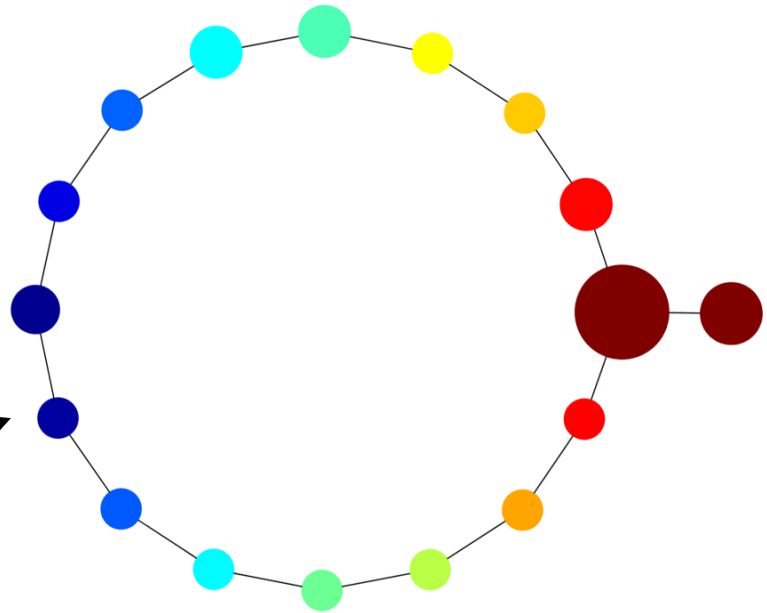
Topology extraction

Filter Range : 0.0~2.0
Intervals : 5
Overlap : 50%
Magic Fudge : 5

Filter



Clustering



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

Applying to Clinical Data

International Neuroimaging Data-sharing Initiative (INDI)
: 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:

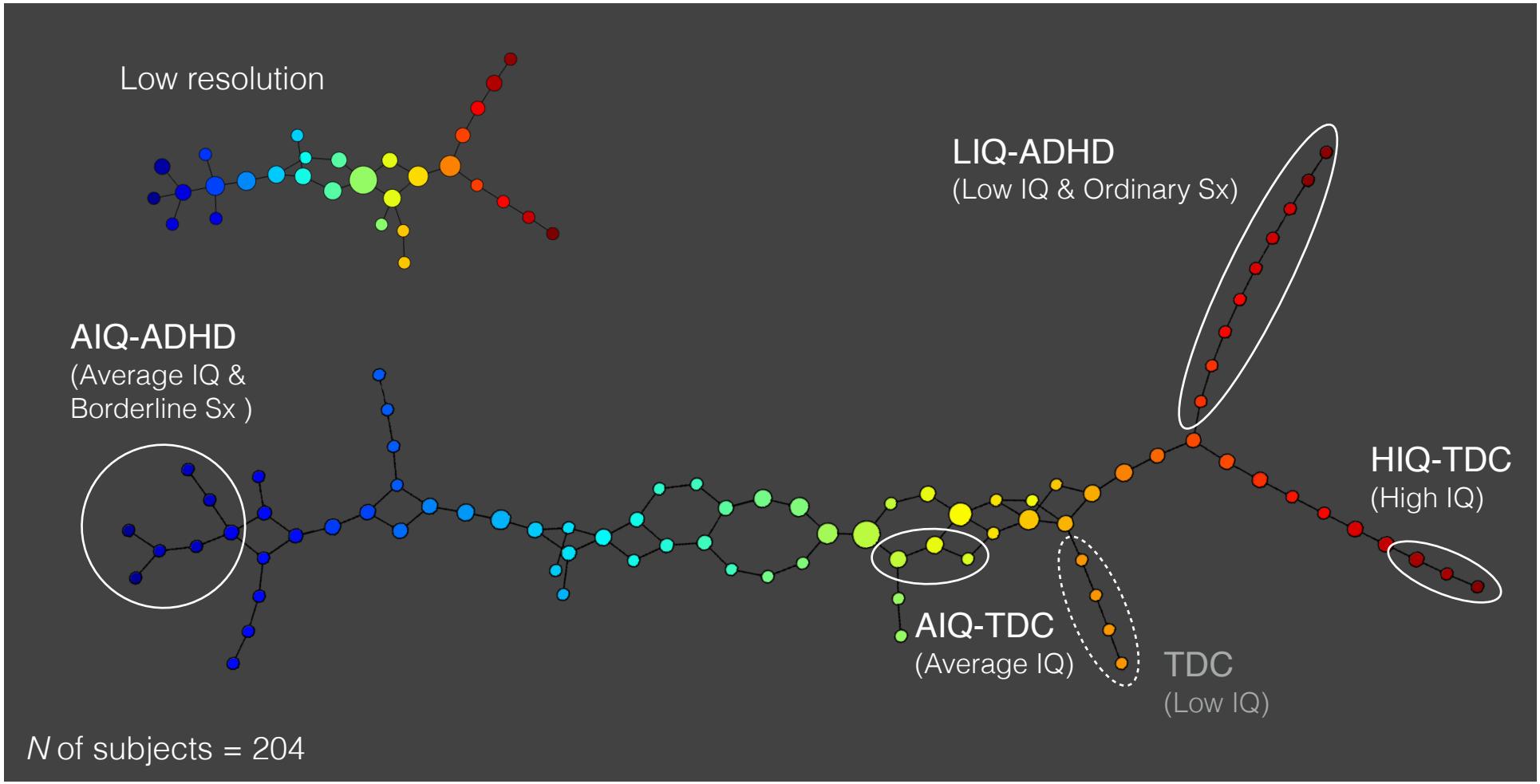
	1	2	3	4	5
1	0.0	61.5	77.3	51.6	77.0
2	61.5	0.0	62.2	55.9	69.0
3	77.3	62.2	0.0	47.2	11.9
4	51.6	55.9	47.2	0.0	52.4
5	77.0	69.0	11.9	52.4	0.0

Filter Function: L-infinity eccentricity

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

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

Clinical data analysis



Distance function: L_2-distance **Filter function:** L-infinity eccentricity

Low High

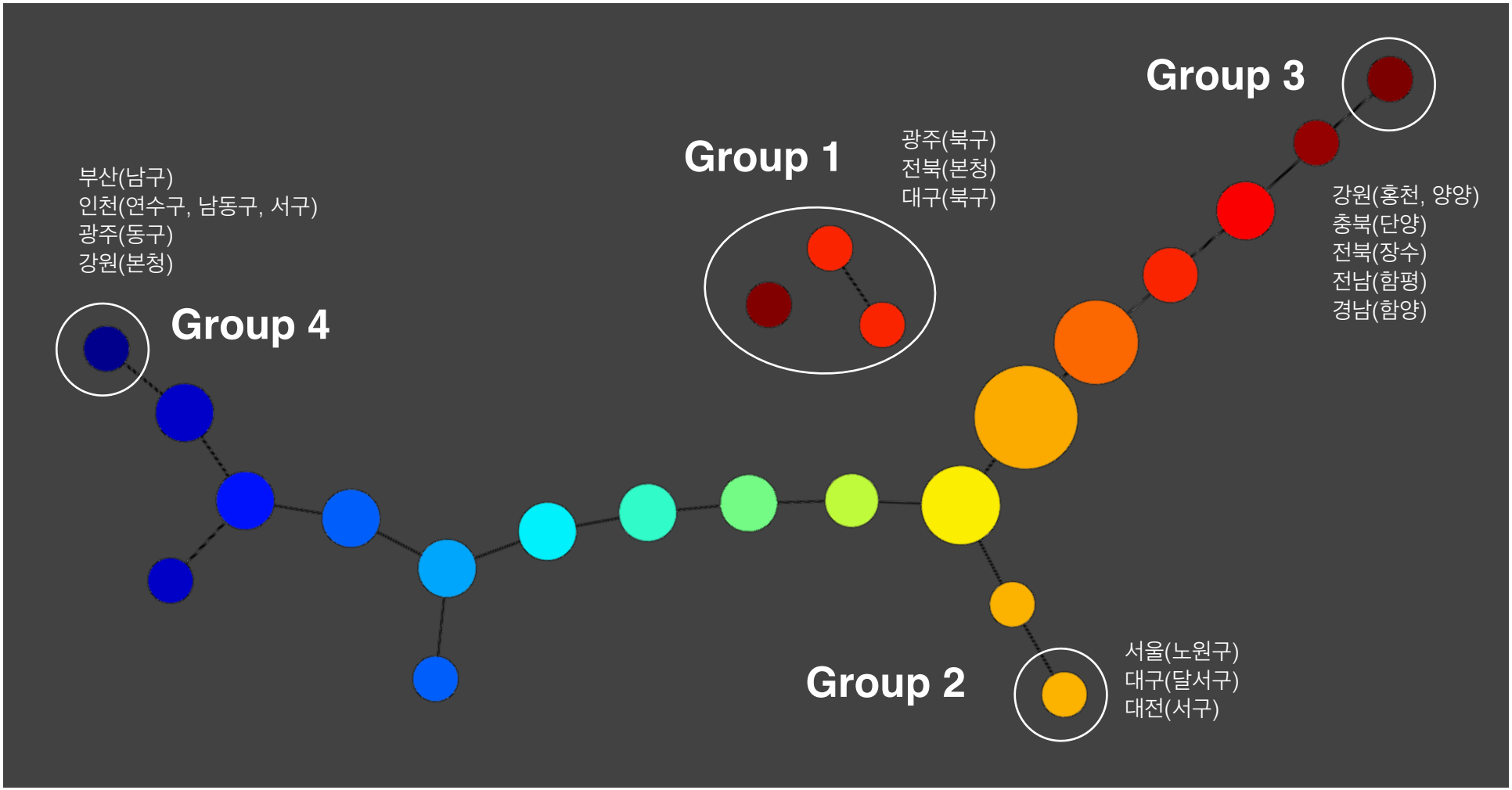
복지/토건/자살자수 데이터

	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

공공데이터 분석

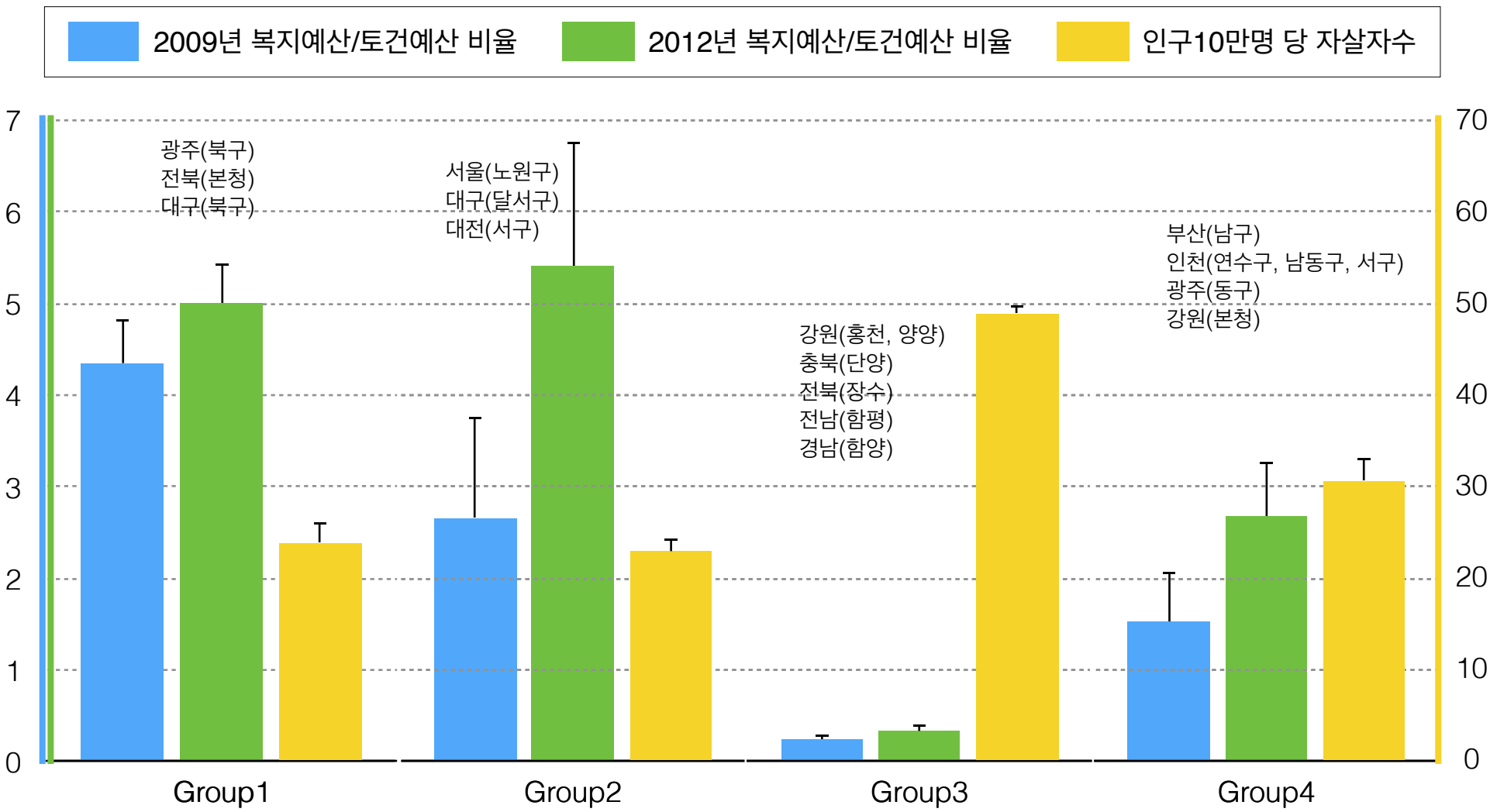
Input data: 2009년 복지예산/토건예산 비율, 2012년 복지예산/토건예산 비율, 2012년 10만명당 자살자수 (연령표준화)



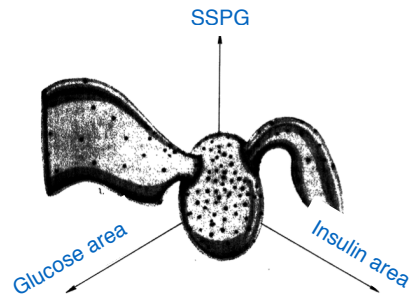
Distance function: L_2-distance **Filter function:** L-infinity eccentricity

Low High

복지예산 I 토건예산 I 자살자수



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



Diabetes Subtypes

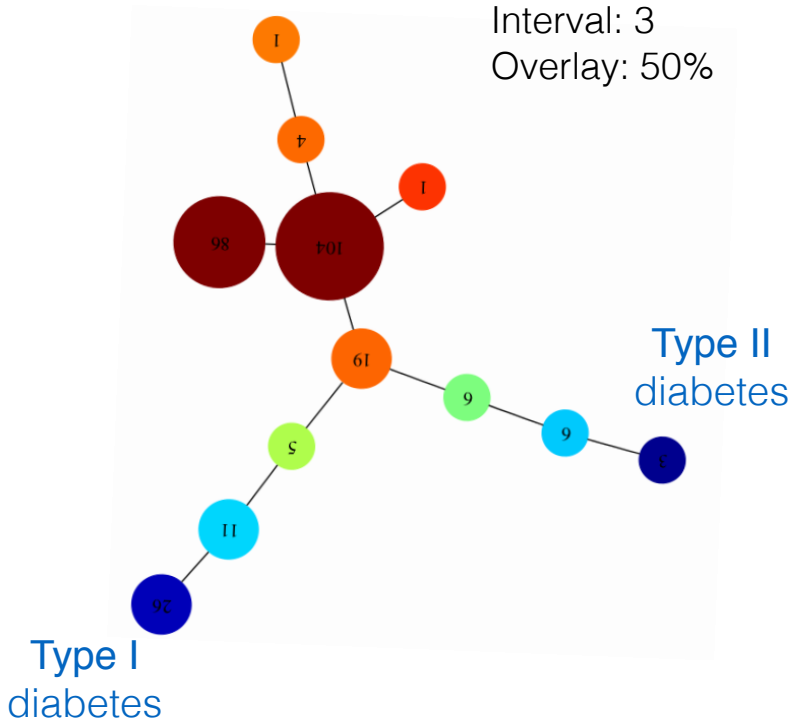
based on six quantities:

- 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

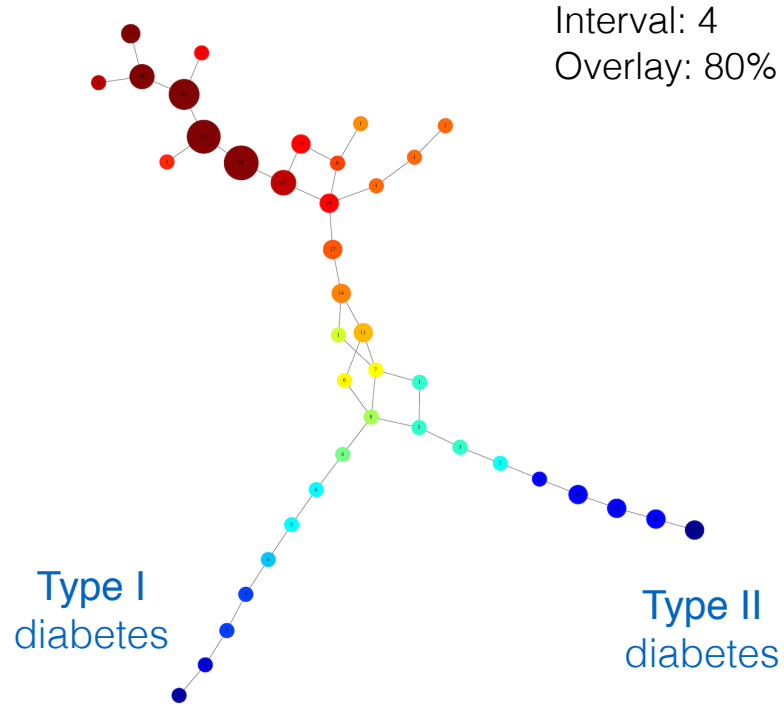
Eurographics Symposium on Point-Based Graphics, Singh G *et al.* (2013)

Two types of diabetes

Low-resolution



High-resolution



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**,

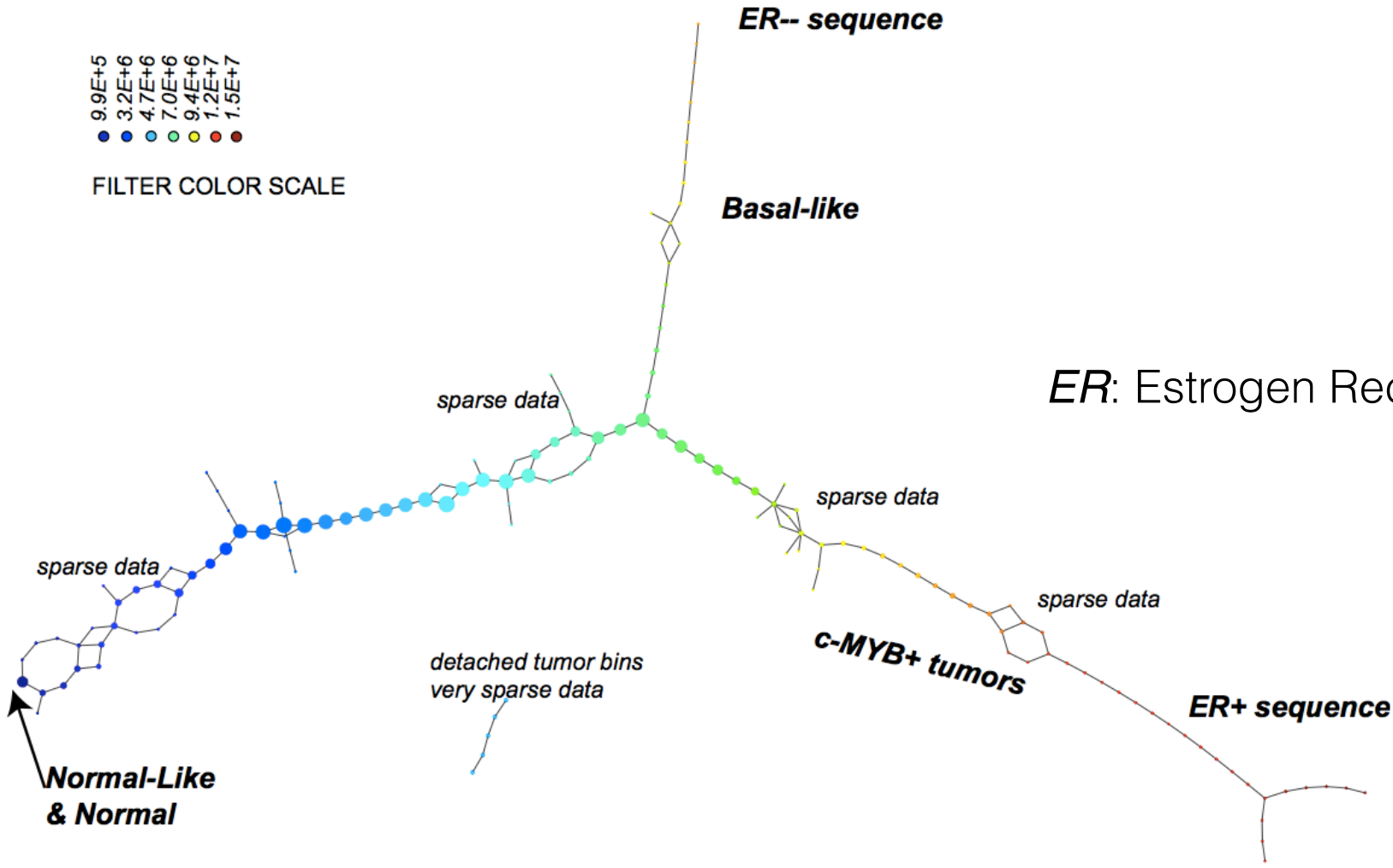
Subgroups of Breast Cancer

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)

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:

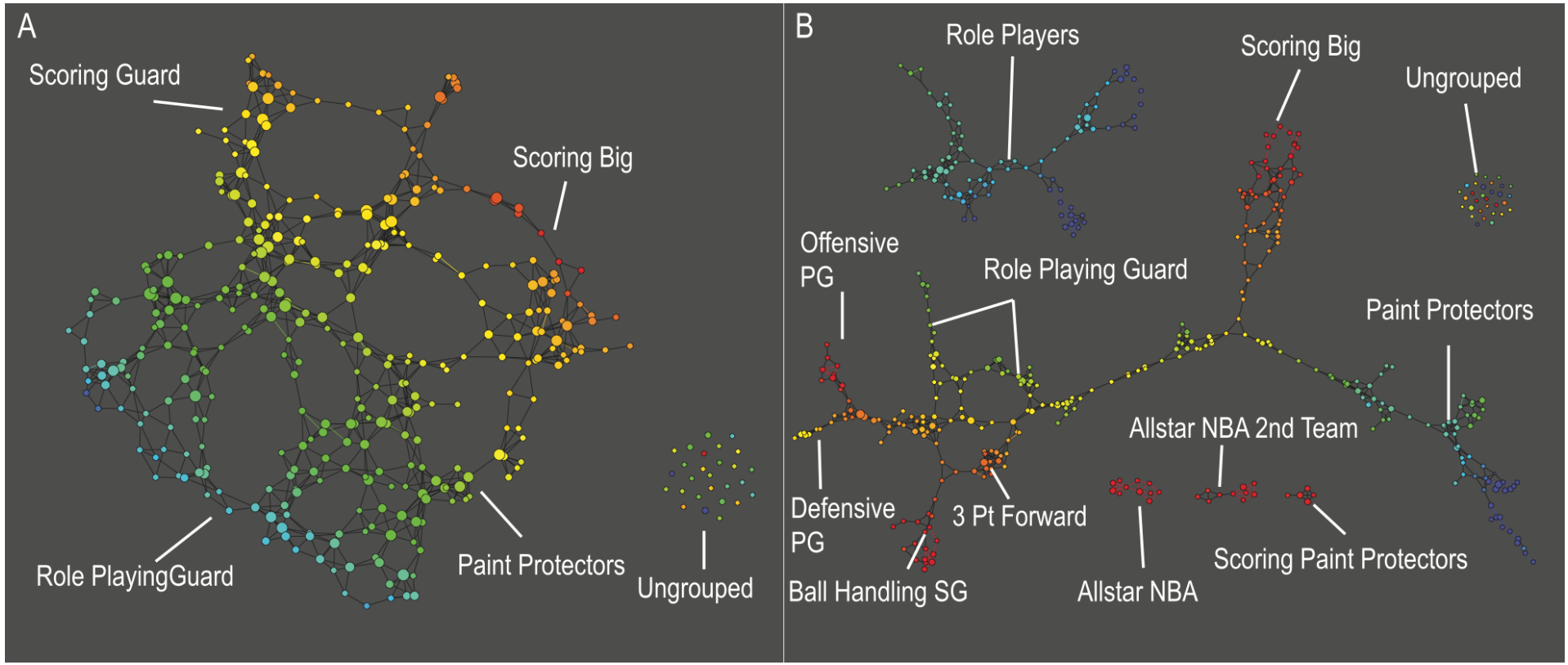
- rates (per minute played) of rebounds, assists, turnovers, steals, blocked shots, personal fouls, and points scored (7 performance measures)

Scientific **Reports**, P.Y. Lum *et al.* (2013)

Map of Players

low resolution map at 20 intervals

high resolution map at 30 intervals



Distance function:

Variance normalised L_2-distance

Filter function:

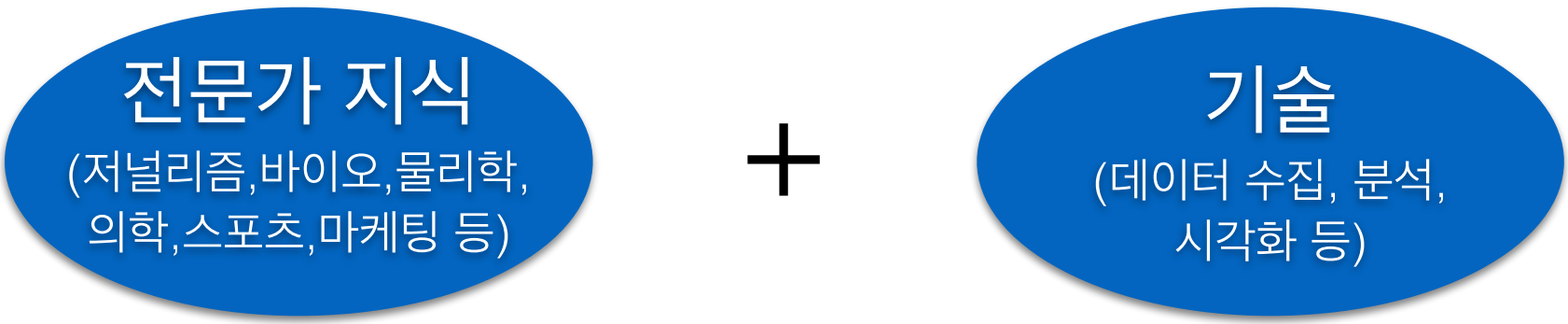
Principal and secondary SVD values

Points Per Game

Low — High

Topological Data Analysis,

as a new weapon for
data scientists.



Conclusion

- TDA는 다양한 형태의 데이터에 적용될 수 있으며, coordinate free 특징이 있다.
- Distance and filter function은 각각의 데이터에 맞게 디자인하고 최적화 해야 한다.
- 데이터에 대한 새로운 insight를 찾고자하는 데이터 과학자에게 TDA는 강력한 무기가 될 것이다.

References

1. Gunnar Carlsson, Topology and Data, Bull. Amer. Math. Soc. volume 46 (2), pages 255-308, 2009.

2. Monica Nicolau *et al.*, Disease-specific genomic analysis: identifying the signature of pathologic biology, Bioinformatics, volume 23 (8), pages 957-965, 2007.

3. Monica Nicolau *et al.*, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, PNAS early edition, 2011.

4. 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.

5. AYASDI, a commercial software for TDA, http://www.ayasdi.com/

6. ***Mapper*** (matlab & python) download:
   http://comptop.stanford.edu/programs/ % matlab version
   http://math.stanford.edu/~muellner/mapper/ % python version

7. PHOM: Persistent Homology in *R*: (for algebraic topology)
   http://cran.r-project.org/web/packages/phom/