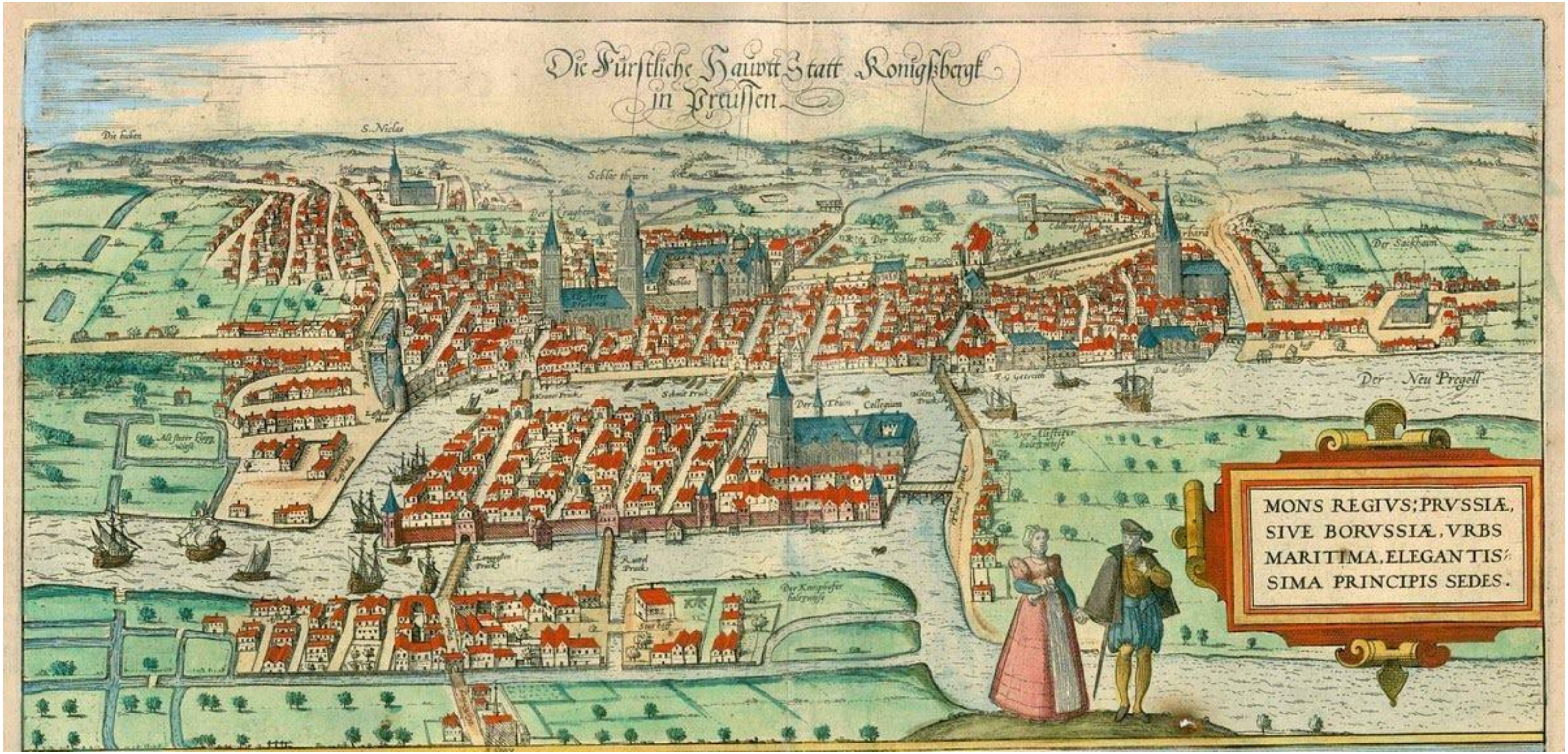
The Underlying Topology of Data

Jose Perea

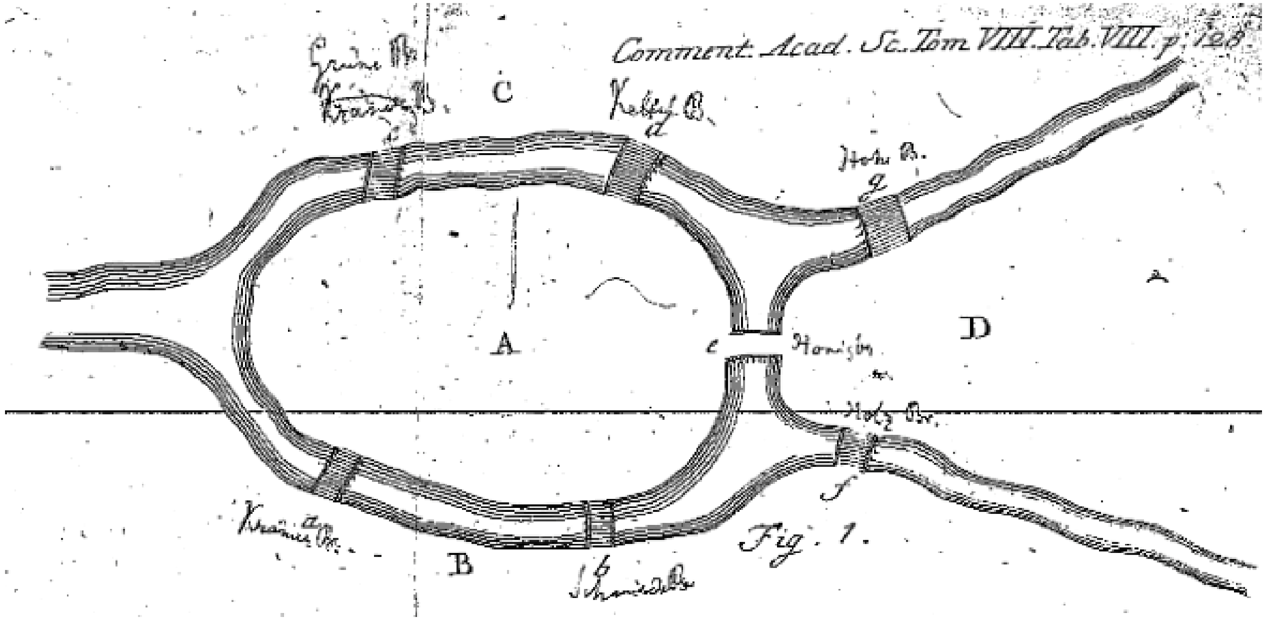
Mathematics

Computer Sciences



Königsberg, 1700s

“This question is so banal, but seemed to me worthy of attention in that [neither] geometry, nor algebra, nor even the art of counting was sufficient to solve it”

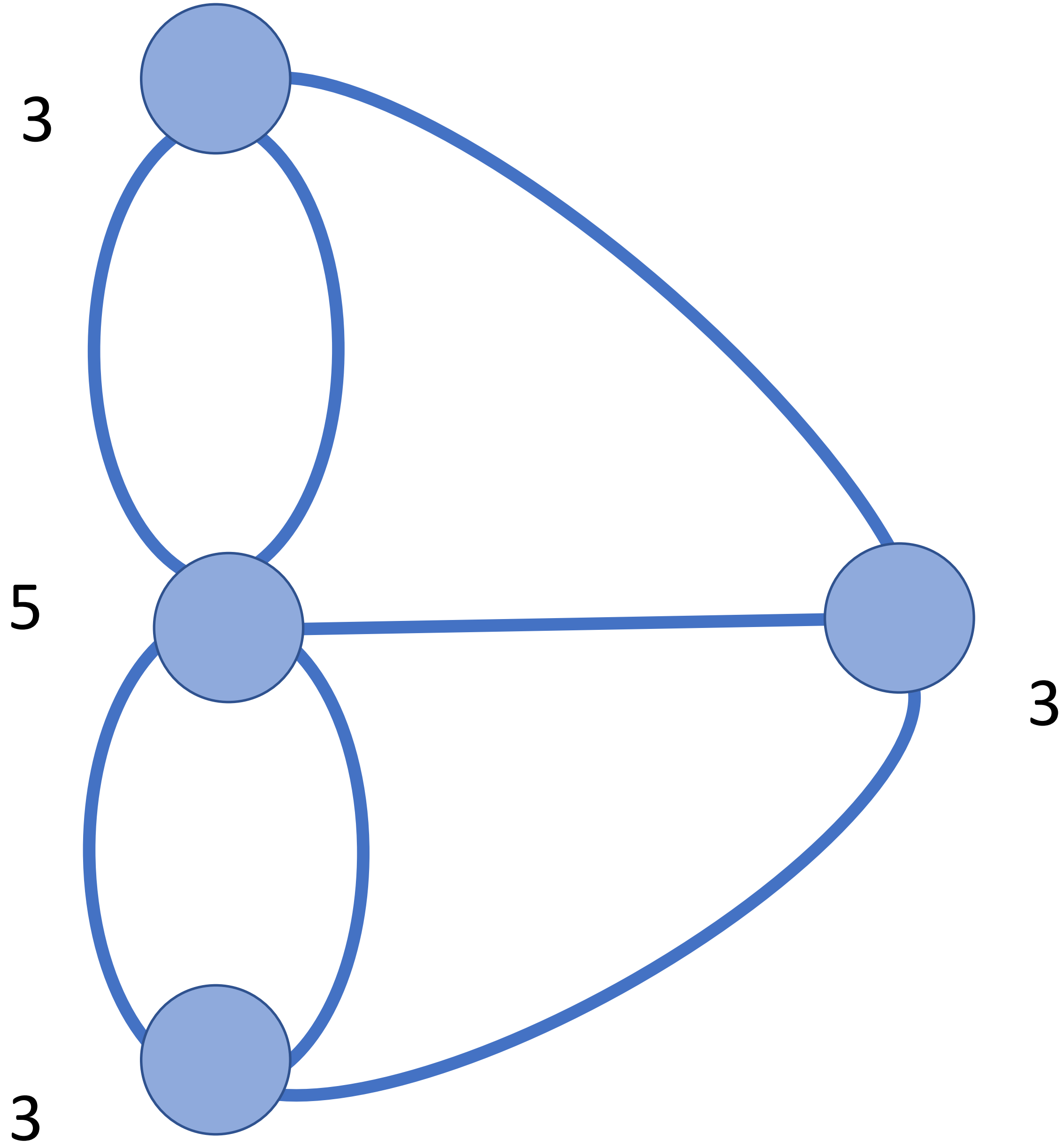


Leonhard Euler, 1707 - 1783

Hopkins, Brian, and Robin Wilson. "The Truth about Königsberg." *College Mathematics Journal* (2004), 35, 198-207.

degree

$\mathcal{G}$



Leonhard Euler, 1707 - 1783

Leonhard Euler, 1707 - 1783

Theorem

$\mathcal{G}$

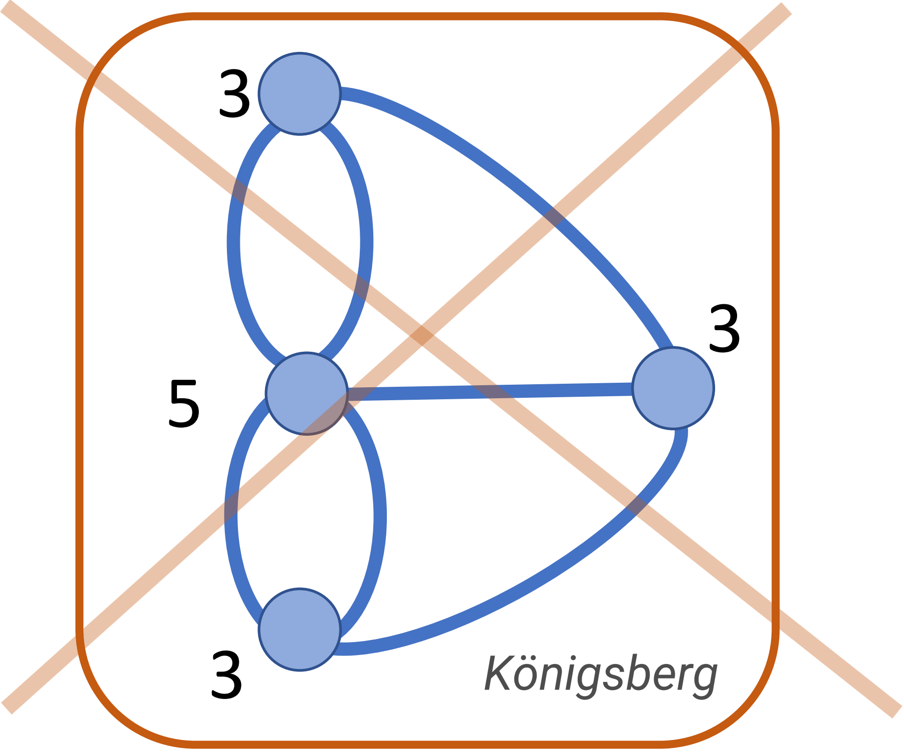
Has an Eulerian path

THEN

\# of nodes with odd degree

=

0 or 2



Euler, L. *Solutio problematis ad geometriam situs pertinentis*. Commentarii academiae scientiarum Petropolitanae, 1741

In Topology

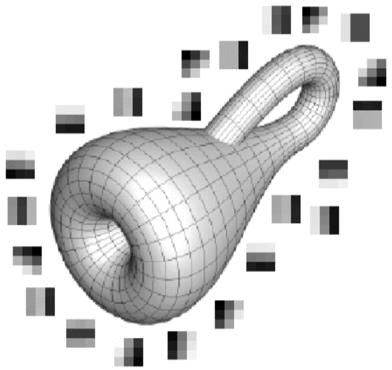
Objects are equal up to **continuous** deformations:

=

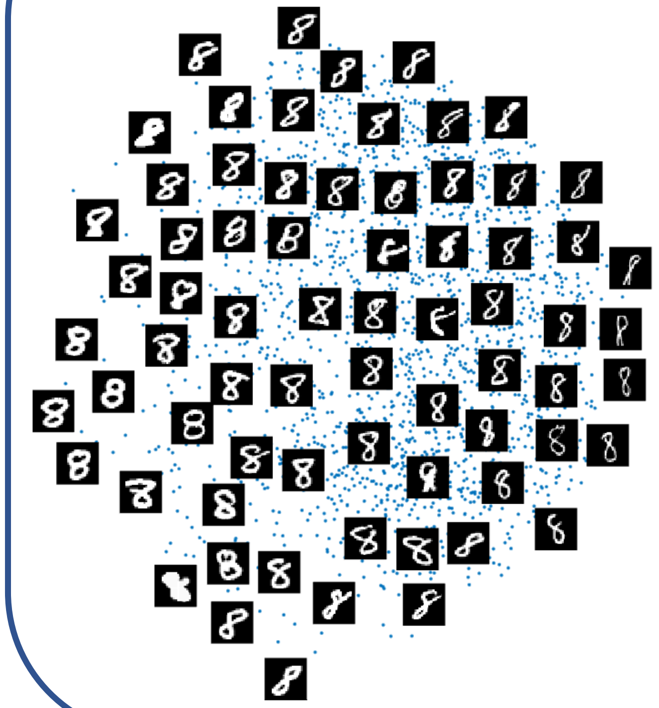
=

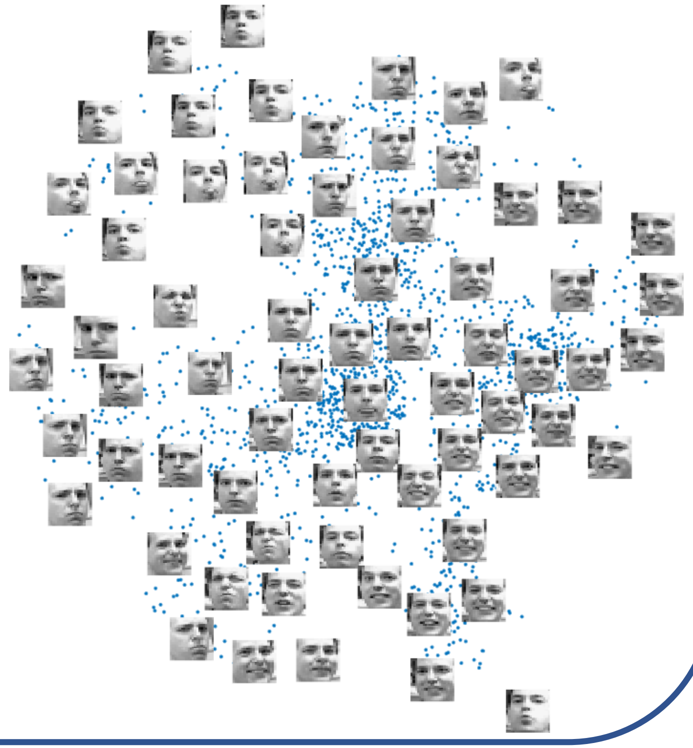
Topological Data Analysis

Computer Vision

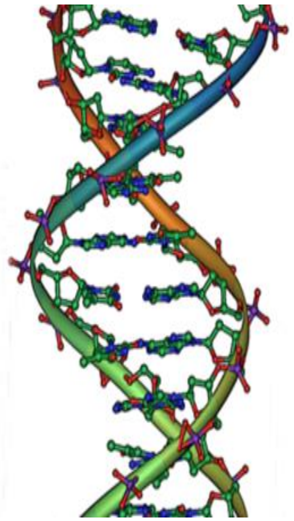


Dimensionality Reduction

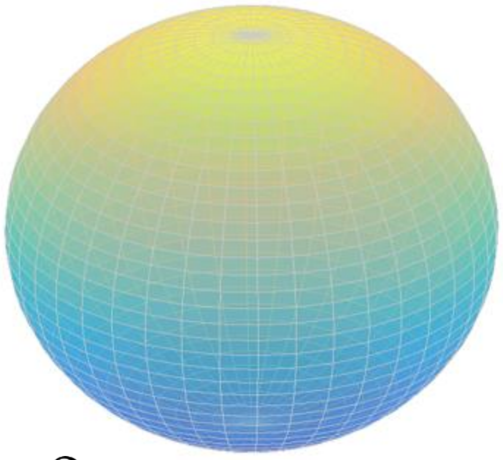




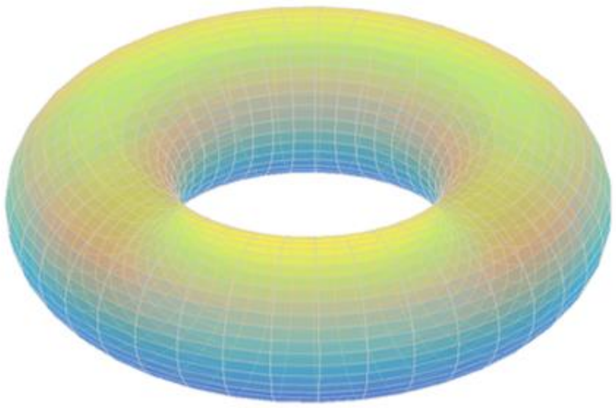
Computational Biology



Betti Numbers: $\beta_n(K) \sim \#$ number of n-dim holes



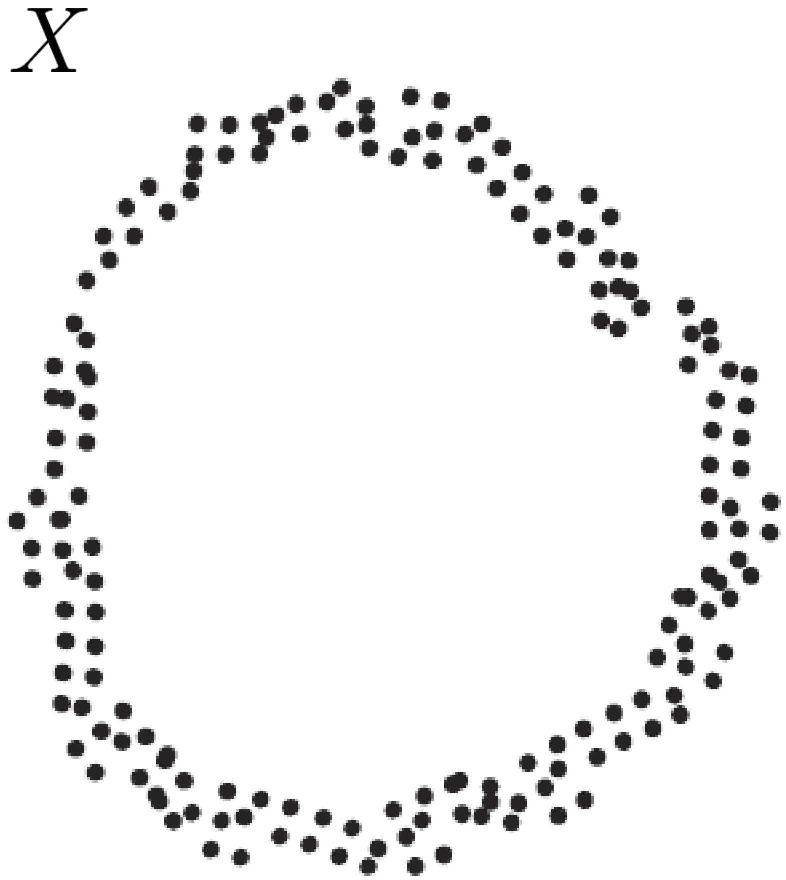
S^2



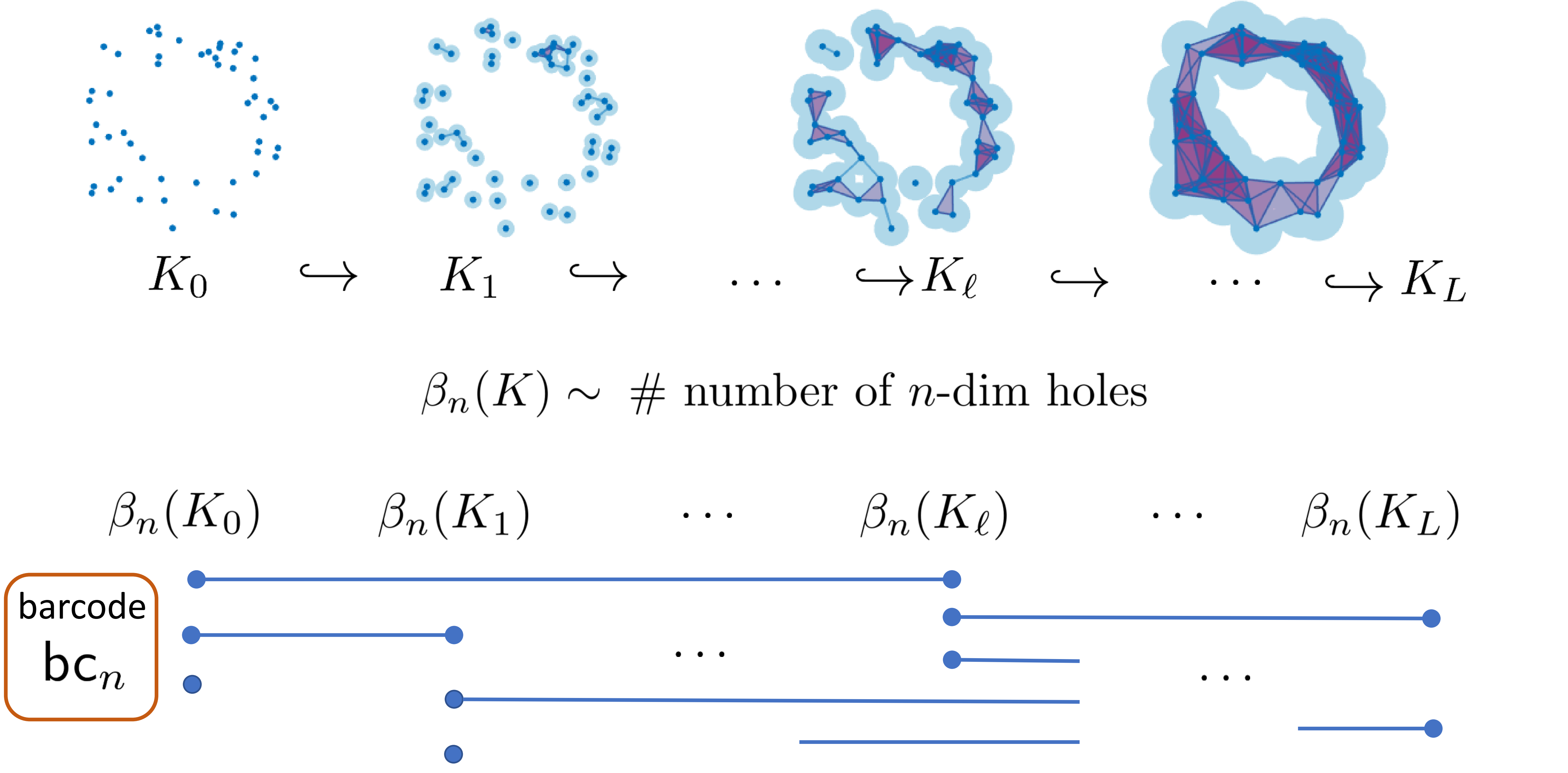
T

$\beta_0(S^2) = 1$	Components	$\beta_0(T) = 1$
$\beta_1(S^2) = 0$	Holes	$\beta_1(T) = 2$
$\beta_2(S^2) = 1$	Voids	$\beta_2(T) = 1$

$$\beta_0(X) = \#(X)$$



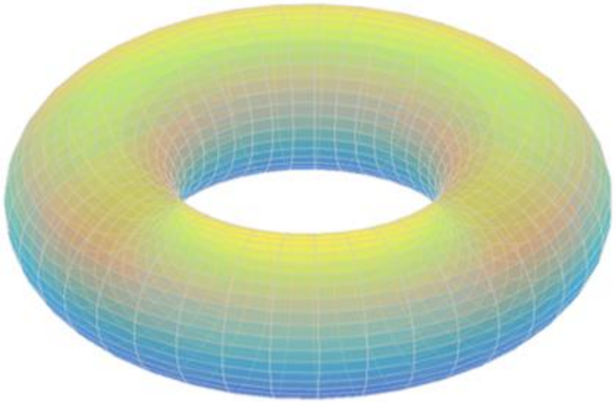
$$\beta_1(X) = 0$$



$K_0 \hookrightarrow K_1 \hookrightarrow \cdots \hookrightarrow K_\ell \hookrightarrow \cdots \hookrightarrow K_L$
$\beta_n(K) \sim$ # number of n-dim holes
$\beta_n(K_0)$ $\beta_n(K_1)$ $\cdots$ $\beta_n(K_\ell)$ $\cdots$ $\beta_n(K_L)$
barcode
bc_n

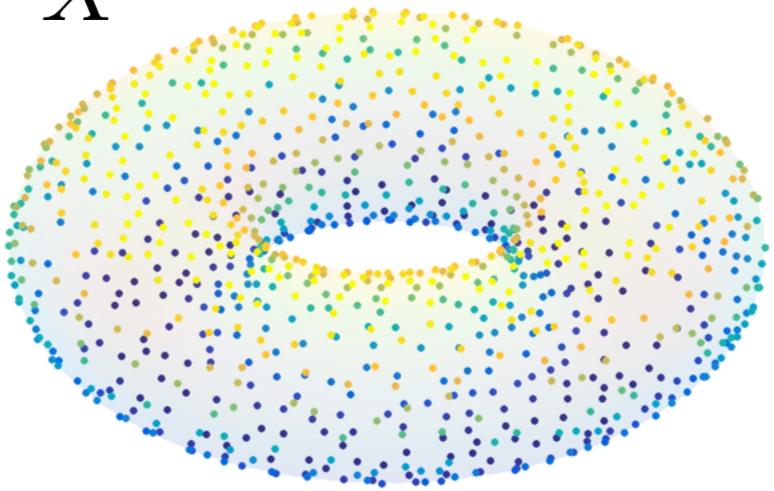
The Persistent Homology of Data:

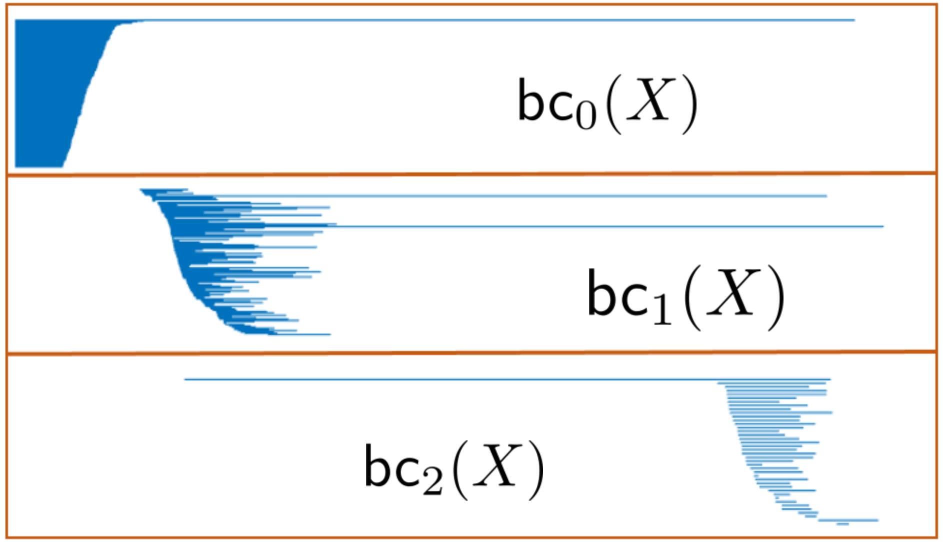
Betti Numbers



$$\beta_0(T) = 1$$
$$\beta_1(T) = 2$$
$$\beta_2(T) = 1$$

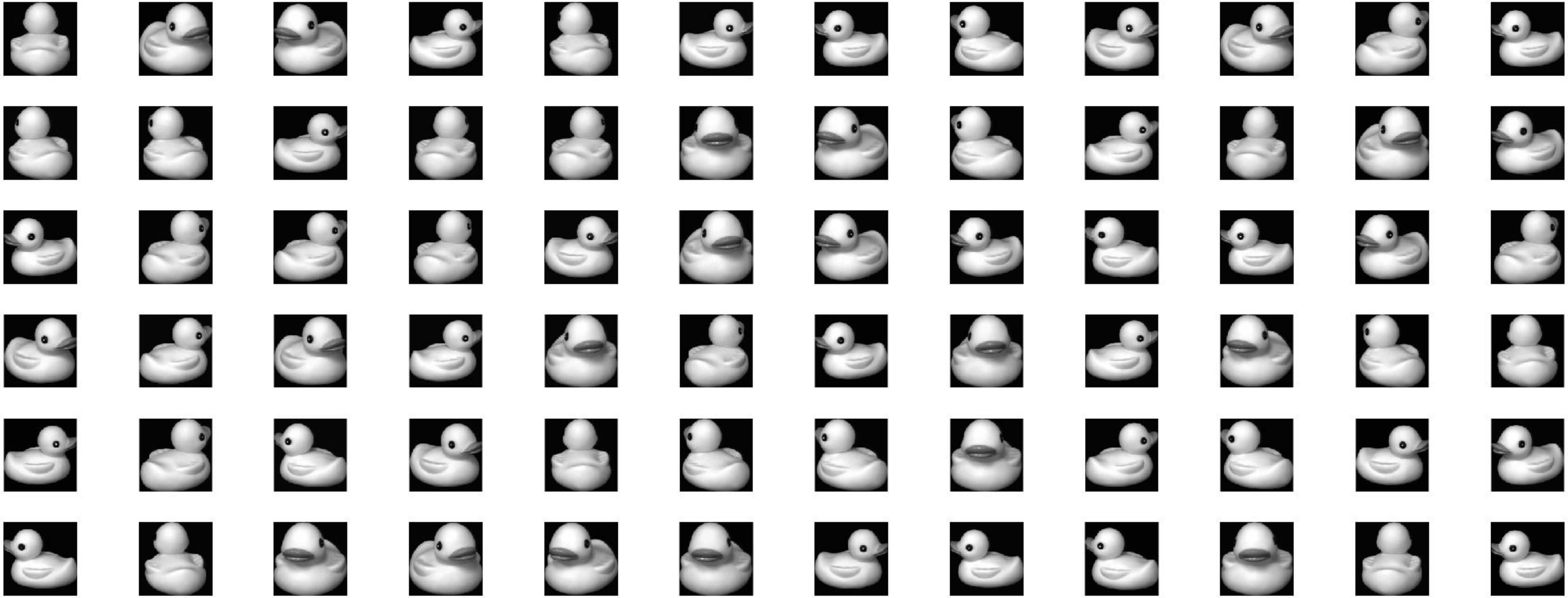
X



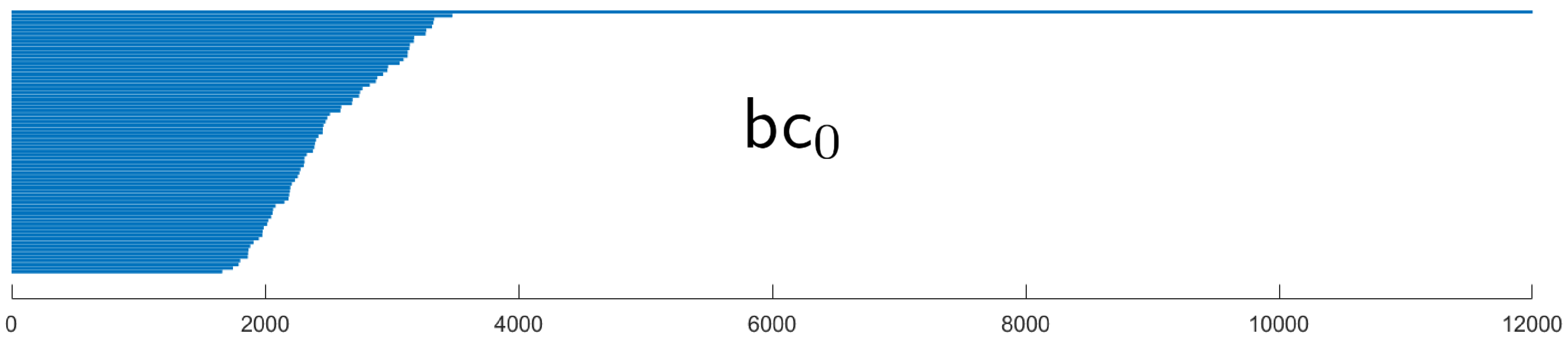


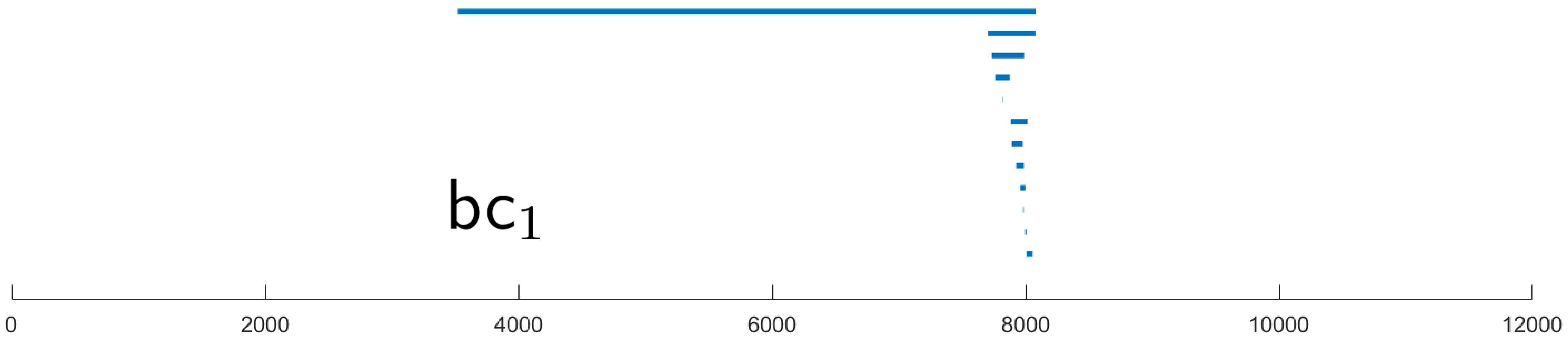
Barcodes

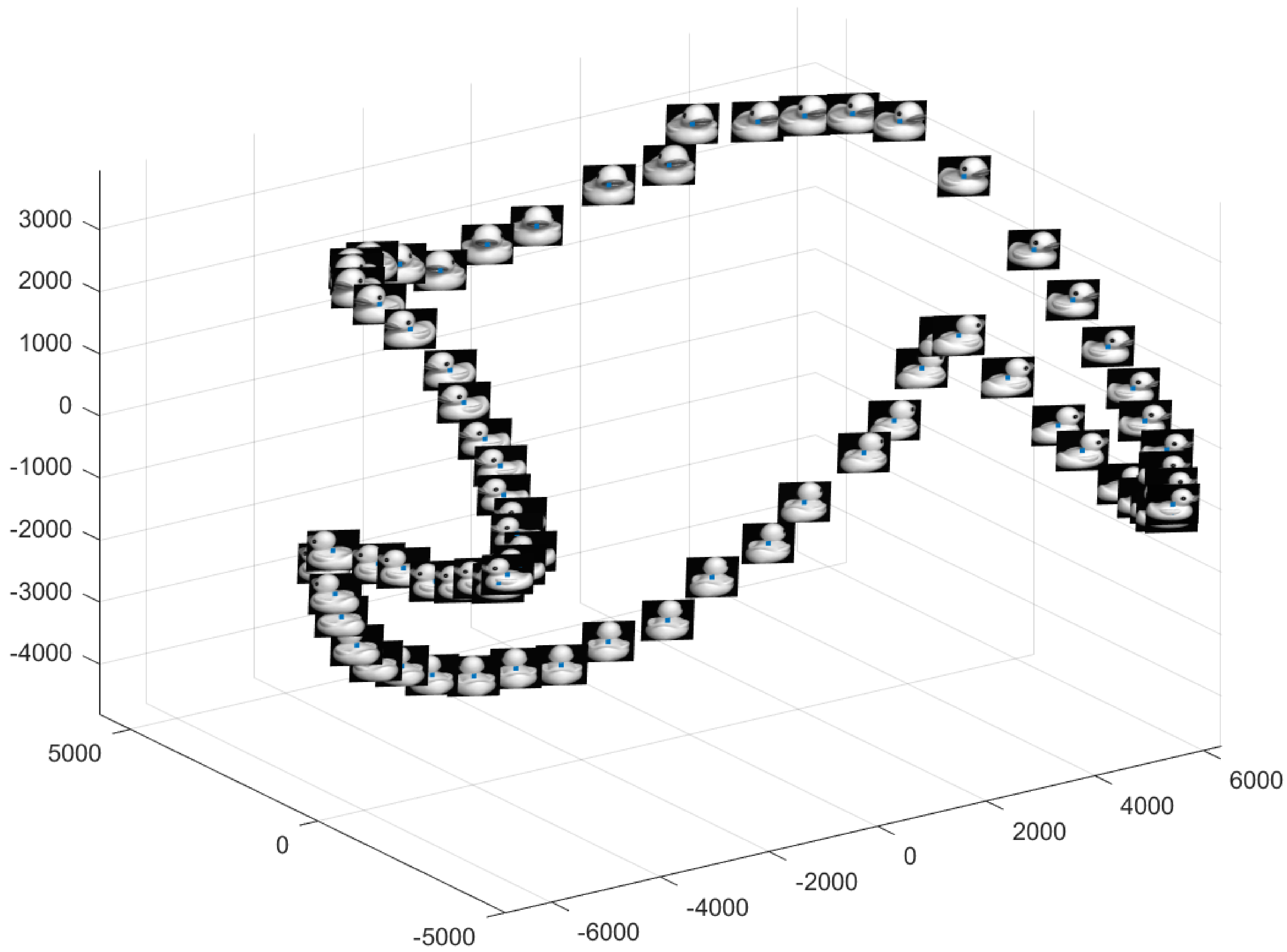
Data



The Persistent Homology of Data:



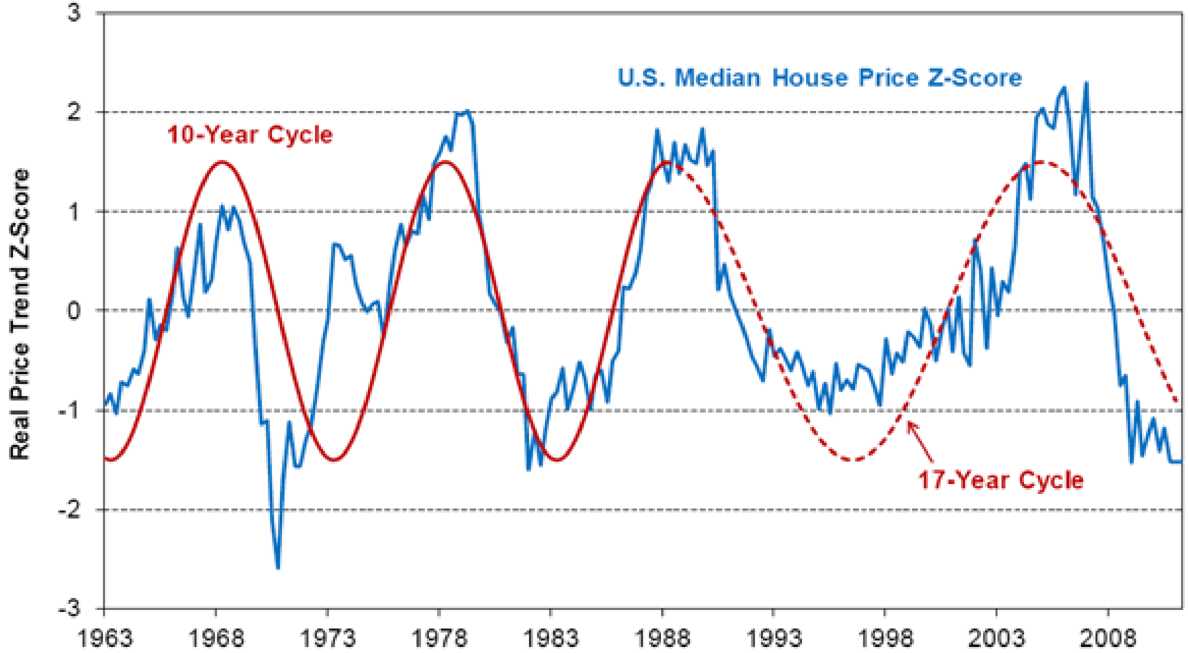



3000
2000
1000
0
-1000
-2000
-3000
-4000
5000
0
-5000
-6000
-4000
-2000
0
2000
4000
6000

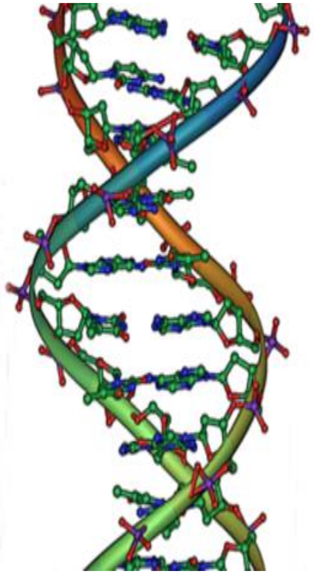
Detecting Recurrence in Time Series Data

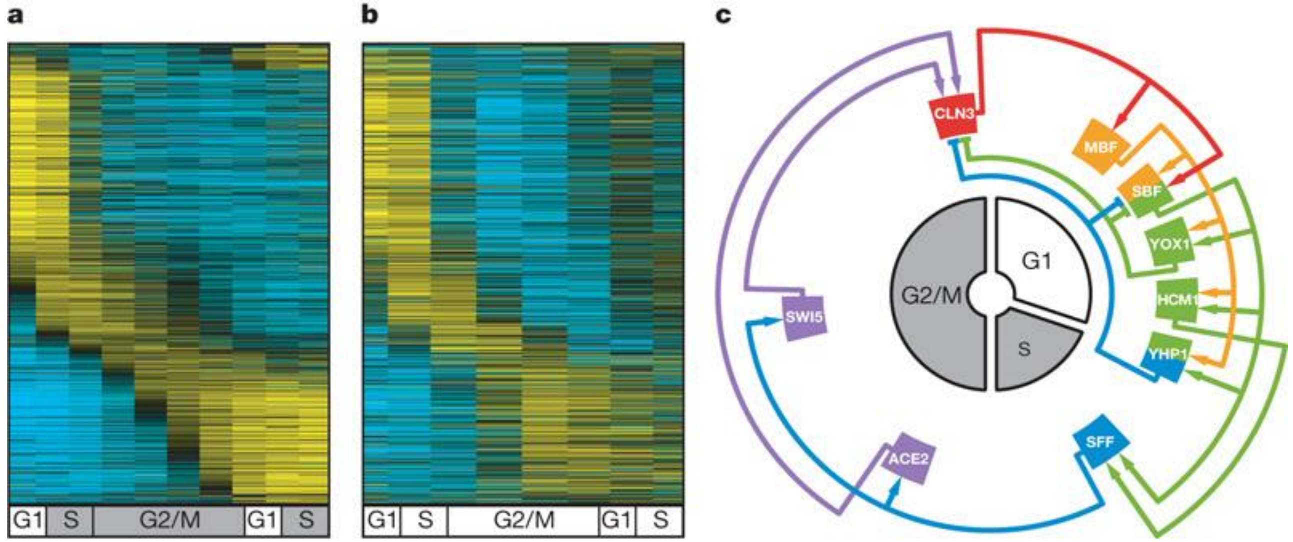
Exhibit 1

U.S. Housing Follows a More or Less Regular Cycle

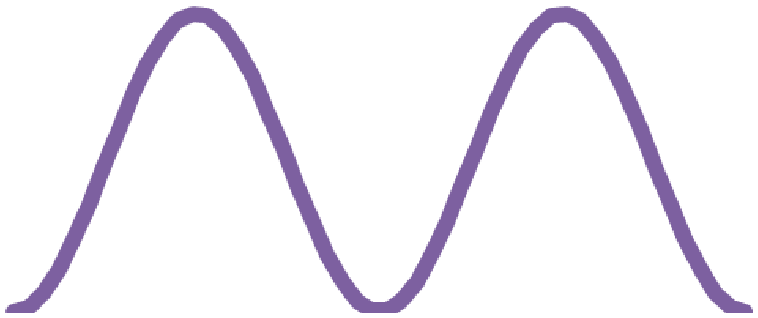


Source: Bureau of the Census, GMO   As of 6/30/11

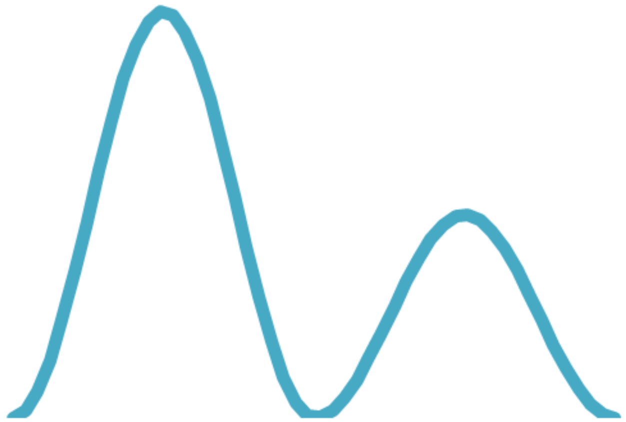




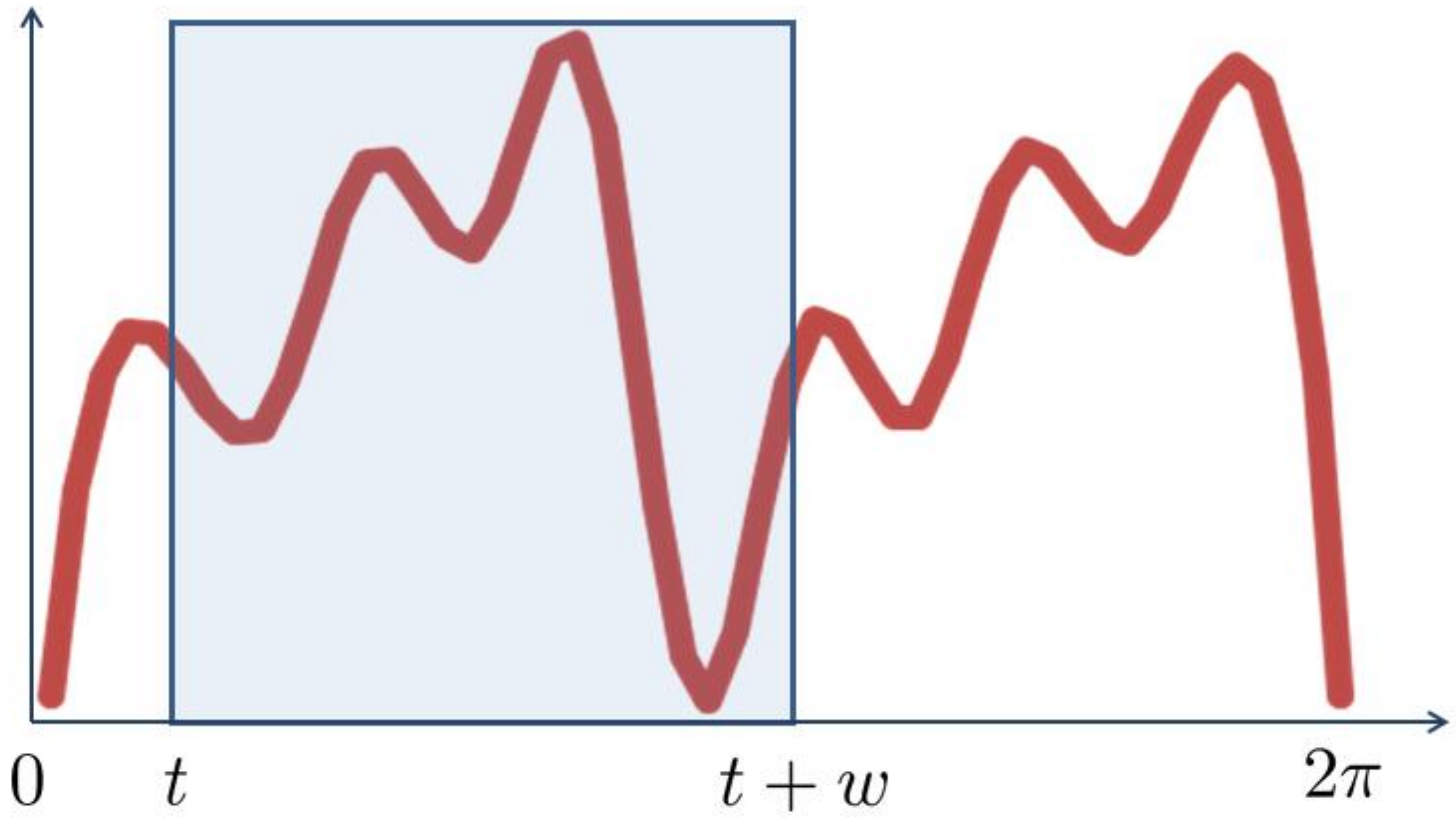
Global control of cell-cycle transcription by coupled SDK and network oscillators, D. Orlando et. al., Nature, 2008



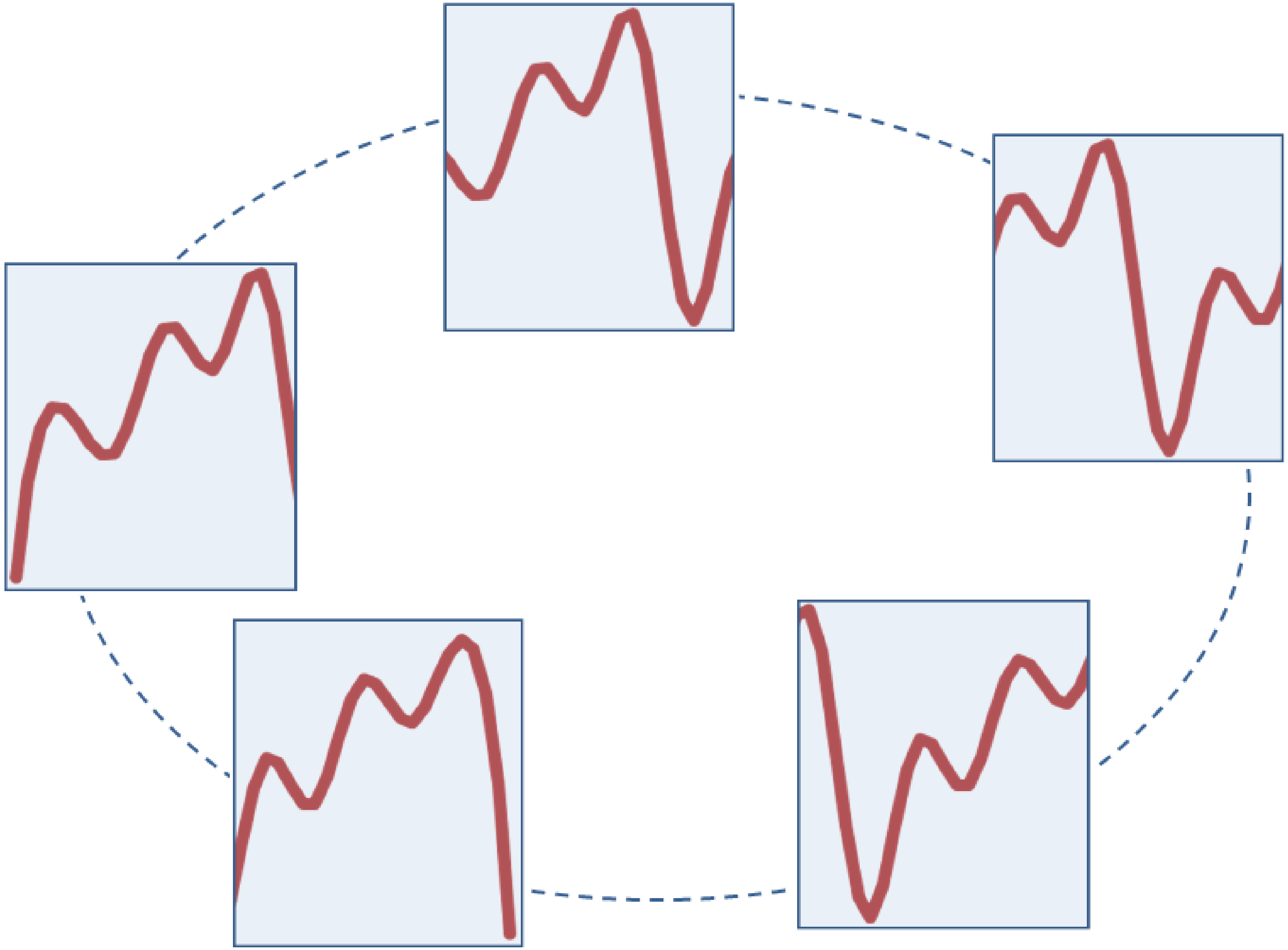

What is recurrence, and
how do we quantify it?




Sliding Windows

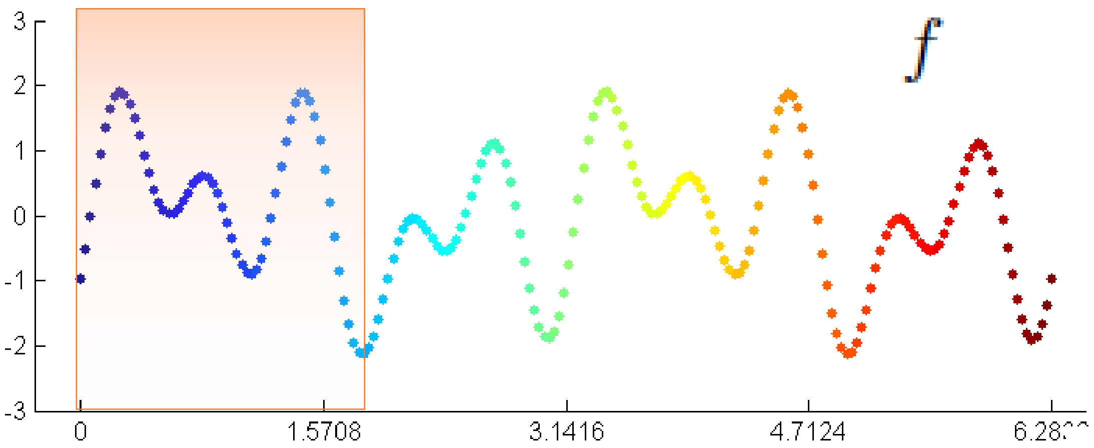


Sliding Windows



Sliding window embedding

window

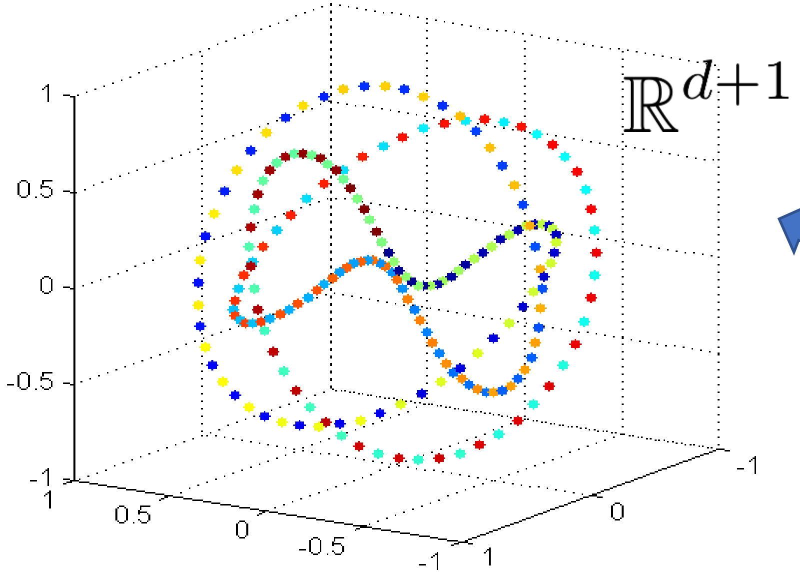


t $\quad$ $t + d\tau$

τ → Step/delay

$d + 1$ → Dimension

$$SW_{d,\tau} f(t) = \begin{bmatrix} f(t) \\ f(t+\tau) \\ \vdots \\ f(t+d\tau) \end{bmatrix}$$



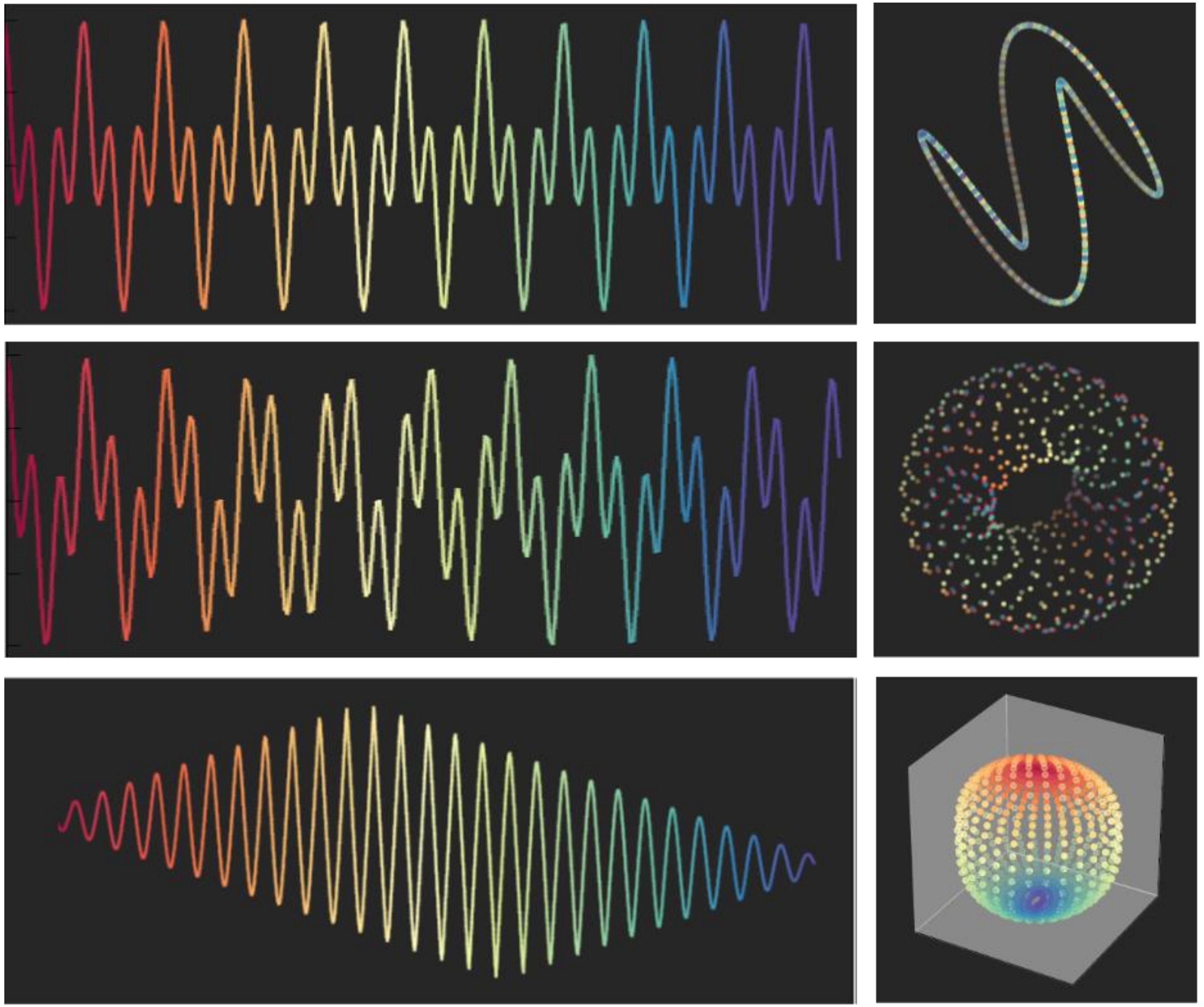
Sliding window point-cloud

$$\mathbb{SW}_{d,\tau} f$$
$$||$$
$$SW_{d,\tau} f(I)$$

$$I \subset \mathbb{R}$$

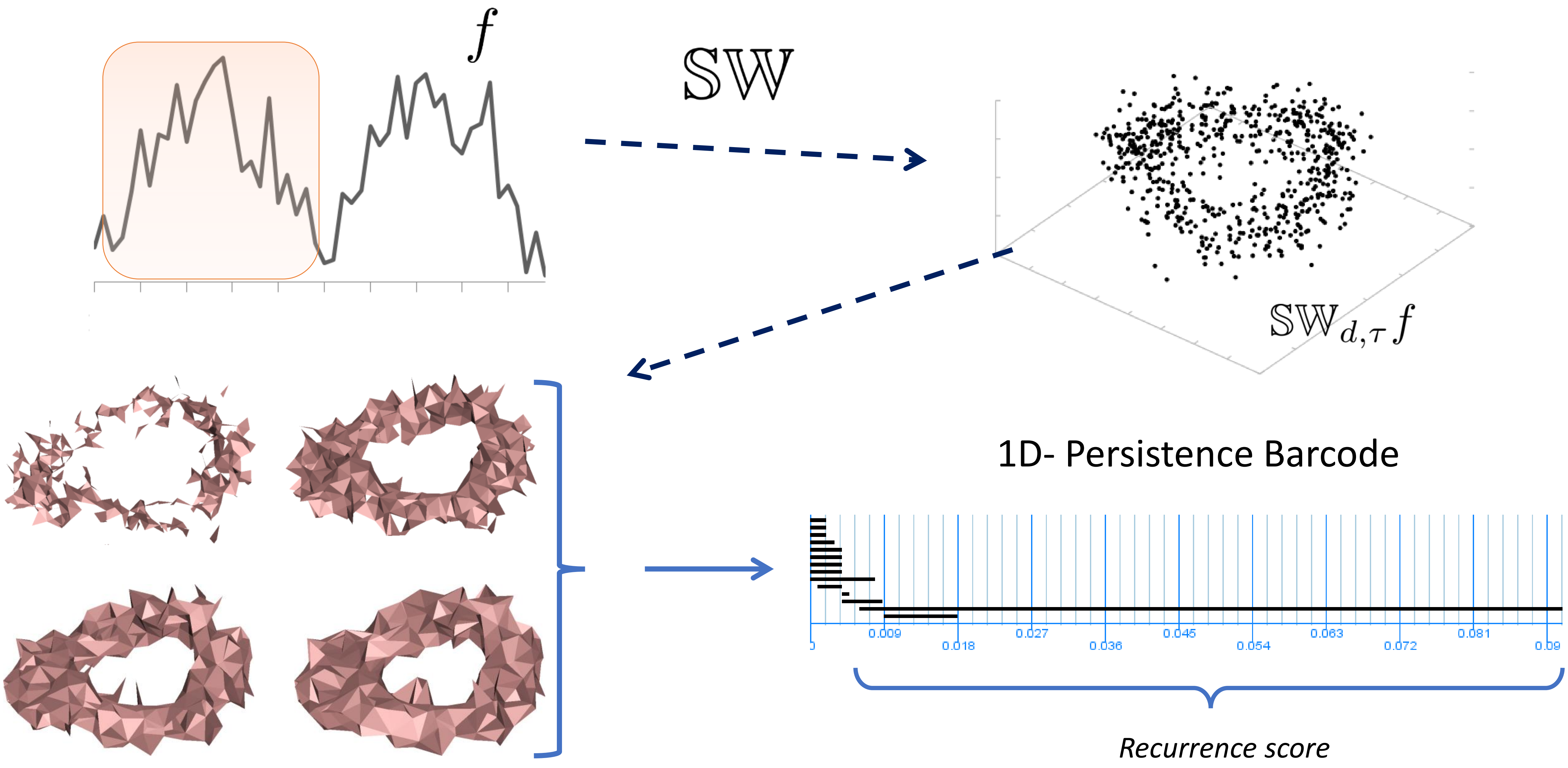
Sliding Windows and Persistence: An application of topology to signal analysis, J. Perea and J. Harer, **FOCM** 2015

f



$\mathbb{SW}_{d,\tau} f$

SW1PerS: Sliding Windows and 1-Persistence Scoring



Sliding Windows and Persistence: An application of topology to signal analysis, J. Perea and J. Harer, **FOCM** 2015

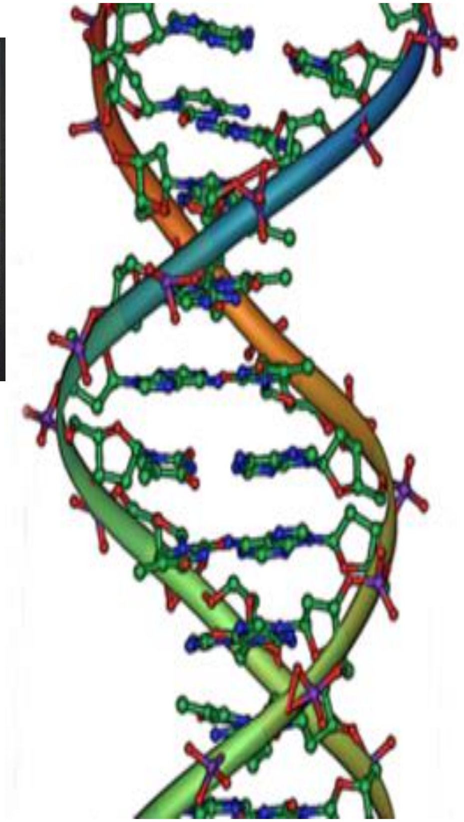
Found Comput Math (2015) 15:799–838
DOI 10.1007/s10208-014-9206-z

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis

Jose A. Perea · John Harer

Biological Clocks



Perea et al. BMC Bioinformatics (2015) 16:257
DOI 10.1186/s12859-015-0645-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data

Jose A. Perea[1,2*], Anastasia Deckard[3], Steve B. Haase[4,5] and John Harer[1,4,6]

hughes2009-liver_res - Excel

Norm Plot	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	11895.2	12898	11911.3	10194.2	11634.3	9322.2	11362.2	9880.3	11402.3	9936.2	8346.8	9777.8	10107.9	6893.2	7986.2	8590.2	8712	9665.2	9078.2
	19494.2	17352.1	18554.1	18380.1	19133.7	17574.4	16778.7	14786.6	17291.2	17252.2	13845.6	16773.6	18236.5	17100.1	17017.1	18457.5	16770.5	18643.4	21593.4
	25261.2	24843.7	23924.7	27841	24070.2	26161.9	23777.7	27511.9	21394.7	22450.8	24288.4	26018.1	26731.8	22372.5	22325.2	22496	23327.2	24087.7	24628
	3117.3	3732.7	3964.2	3111.2	3379	2959.5	2460.3	3250.3	2853	2272.1	3588.5	2691.6	2488.9	1873.1	1733.9	1710.5	2064.8	1919.4	2168.1
	5019.7	4949.6	4661.2	4852.6	4287.6	3668.6	4460.2	4930.9	3365.7	3093.2	3951.8	5614.6	6065.7	3268.9	4349.1	3656.8	5440.7	5067.7	5025.1
	2291.9	1924.7	1974.4	2114.4	2601.3	2184	2781.6	2603.2	2565.3	2421.2	2609.9	2189	2186.3	2167	2106	1908.3	2348.1	2218.4	2171.6
	31688.7	32265.8	34335.7	32932.3	34843.3	28885.1	26908.4	28641.3	26995.8	29937.8	32894.9	32400.3	33468.2	26061.7	25781	24484.8	29700.4	31629	32161
	32270.2	32399.8	30527.7	33447	28200.7	26104.6	24098.6	27367.5	17303.9	17192.4	18305.2	29987.3	28857.8	18378.8	22644.6	22858.7	26571.7	27955.2	30859
	12132.3	12016.1	12976.3	11842.6	11746.3	12261.7	11203.4	9859.5	11791.3	13934.8	13435.6	12962.6	14130.5	15687.7	15616.2	14913.9	16606.3	16124.5	16905
	27906.8	24578	28547.9	25954.4	27405	28009.8	26748.6	26927.5	23337.6	25312.5	27580.3	26473.5	28164	25583.9	27012.6	25628.8	28332.6	29554	27519.3
	4213.7	3803.1	4142	3671.7	4110.5	3930.1	4336.3	3128.7	4498.3	4249.8	4238.2	3936.1	3804.3	4242.1	4079.7	3252.3	4084.6	4041	4102.6

hughes-liver-v1_swft3_rescomb_d

SW1PerS: Sliding Windows and 1-Persistence Scoring, J. Perea et. al., 2015

Probe	Symbol	SW_rank	DL_rank	LS_rank	JTK_rank	Max-Min	Norm Plot	18	19	20	21	22
1450869_at	Fgf1	1	1451.5	39	116	10916.2		15201.2	13204.8	15041.3	14251	11082.3
1416958_at	Nr1d2	2	26.5	10	25.5	62708.9		13006.6	12287.7	10224	9298	7393.1
1417190_at	Nampt	3	242.5	48	5	17275.4		10043.8	9446.7	9994.4	7548.1	6338.7
1450714_at	Azin1	4	4053.5	134	121	8391.4		13494.8	14060.1	13963.6	12193.6	11334.7
1436590_at		5	98	86.5	144.5	49573.2		47739.7	38509.4	38855.4	30598.5	35784.8
1420722_at	Elovl3	6.5	26.5	1	1	149978.3		76307.7	93712.8	98998.6	121038.2	127449
1437250_at	Mreg	6.5	1534.5	38	34.5	30373.3		29040.6	29839.6	31687.1	39536	35259.5

SW1PerS: Sliding Windows and 1-Persistence Scoring, J. Perea et. al., 2015

hughes2009-liver_res - Excel

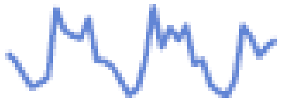
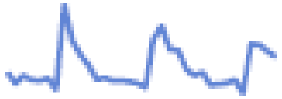
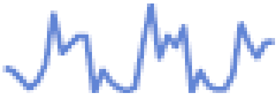
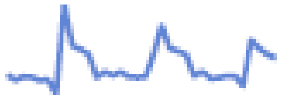
A2 | 1450869_at

1	Probe	Symbol	SW_rank	DL_rank	LS_rank	JTK_rank	Max-Min	Norm Plot	18	19	20	21	22
45096	1459877_x_at		44920	39869.5	40494	37728.5	2233.2		4848.7	4365.1	4505.4	4685.3	4858
45097	1459917_at	Ggnbp2	44920	24121	30730.5	37728.5	1659.9		1132.3	1077.1	1122.8	1250	1164
45098	1459948_at		44920	6142	21561	37728.5	7536.4		3512.8	3705.9	4896	2755.5	2715
45099	1459957_at		44920	27801.5	26601.5	21692.5	172.4		144.6	149.9	111.6	127.9	122.7
45100	1460126_at		44920	27184	40494	37728.5	1024		1065.1	734.9	986.6	810	883.5
45101	1460610_at	Ost4	44920	39204.5	31251	37728.5	55.7		95	94	95	95.3	90.4
45102	FX-MURINE_b1_	NA	44920	29424	21164.5	29740	12663.2		12215.2	10674.9	10611.6	11054.7	13470.4

hughes-liver-v1_swft3_rescomb_d

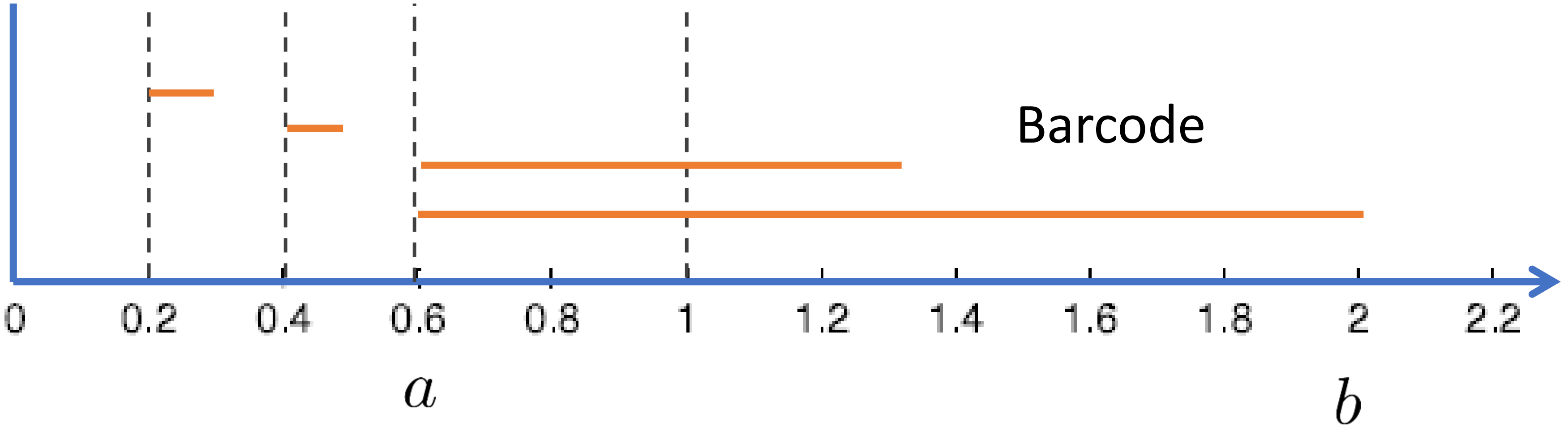
SW1PerS: Sliding Windows and 1-Persistence Scoring, J. Perea et. al., 2015

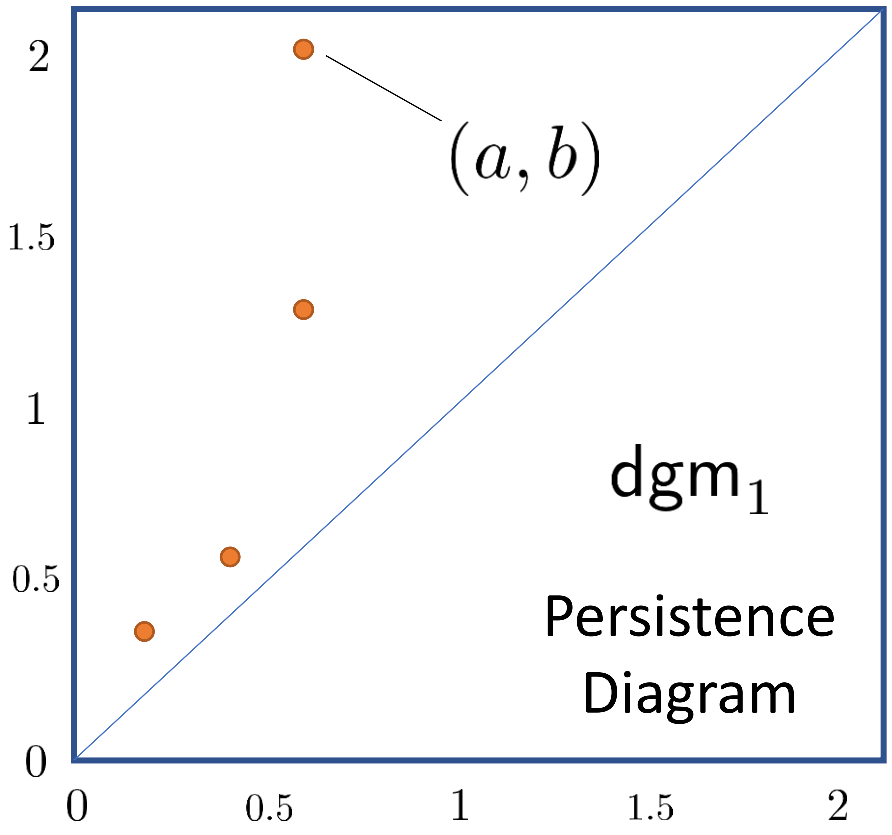
Yeast Metabolic Cycle Data

Gene	SW	DL	LS	JTK	Amp	Plot
ECM33	137	1552	1194.5	1492	35.86	
CDC9	291	1494	1993.5	2714.5	2.81	
SAM1,2	628	1133	1723	3289.5	60.82	
MSH6	715	3569	2381	3341.5	5.06	

Rankings of genes in the top 10% (out of 9,330) according to SW, and not in the top 10% for any other algorithm

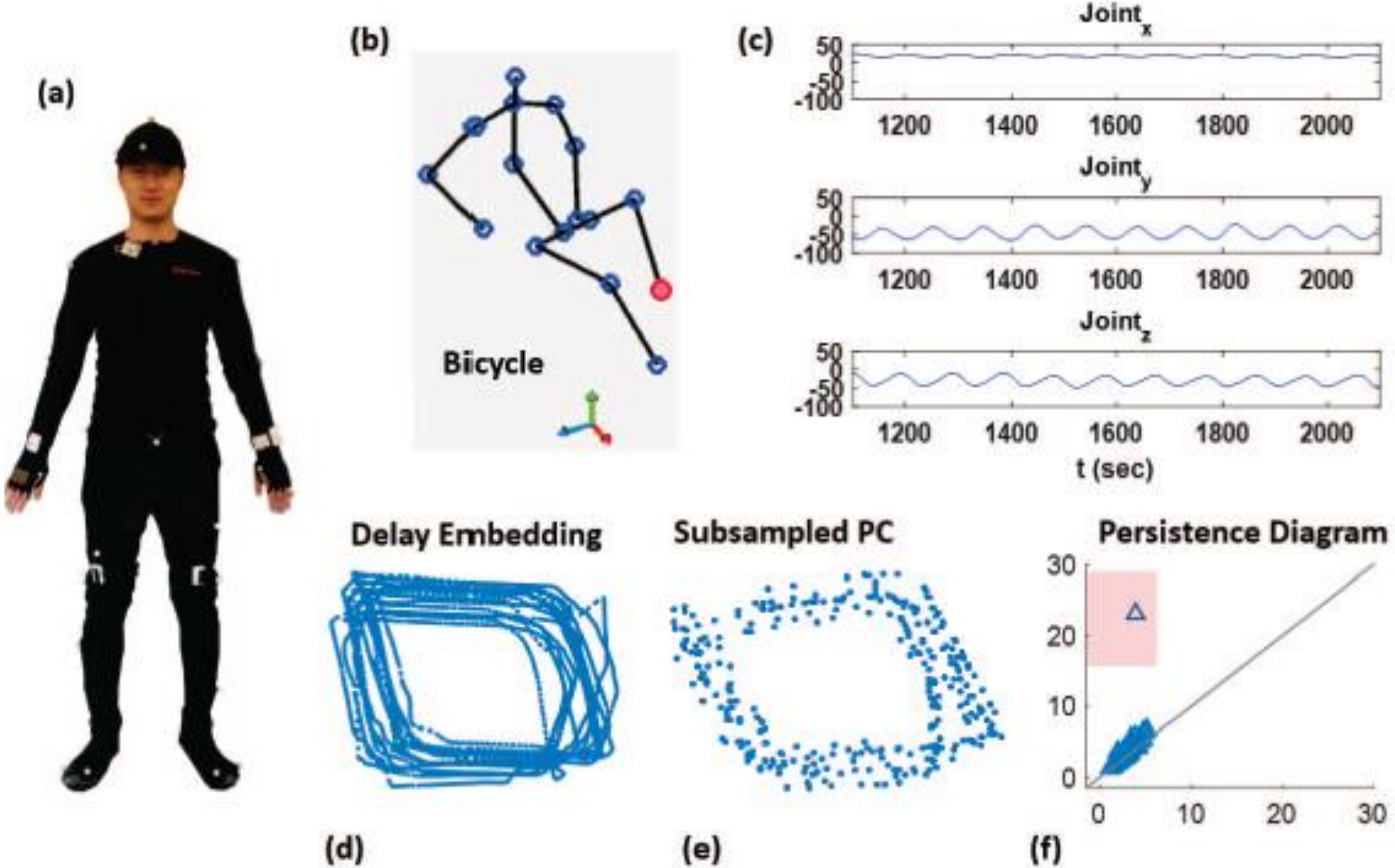
SW1PerS: Sliding Windows and 1-Persistence Scoring, J. Perea et. al., 2015


Barcode
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
a
b

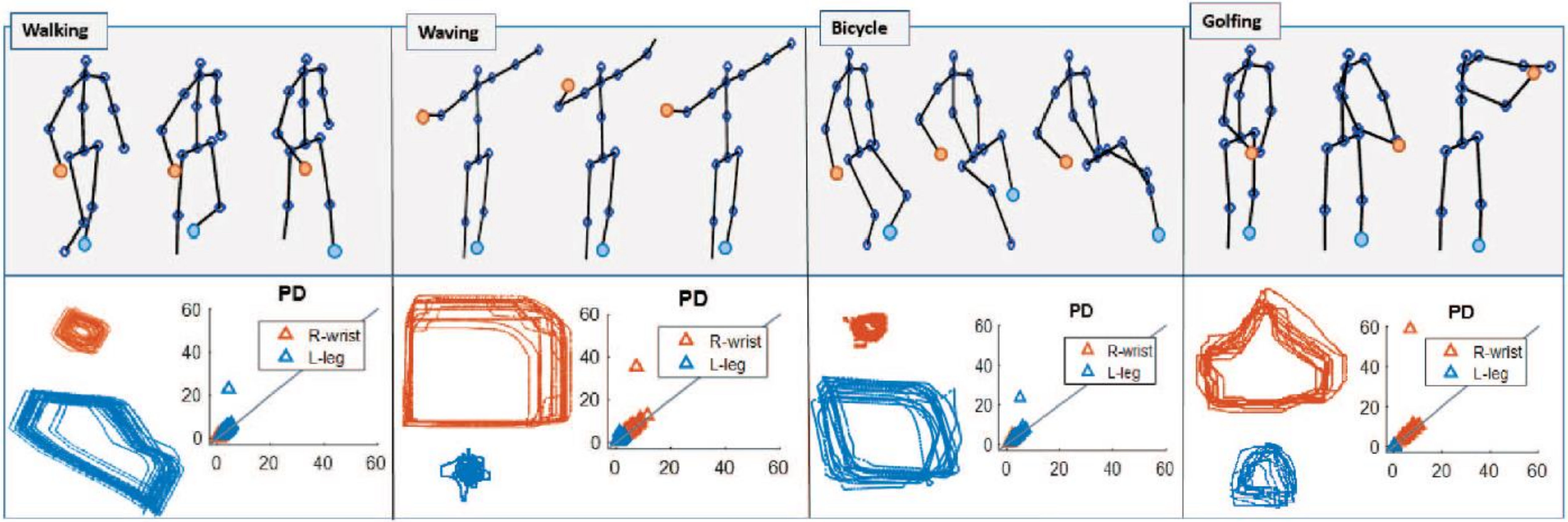

(a, b)
dgm_1
Persistence Diagram
0
0.5
1
1.5
2

ACTION CLASSIFICATION FROM MOTION CAPTURE DATA USING TOPOLOGICAL DATA ANALYSIS

Alireza Dirafzoon, Namita Lokare and Edgar Lobaton



(a)
(b)
Biicycle
(c)
Joint$_x$
Joint$_y$
Joint$_z$
50
0
-50
-100
1200
1400
1600
1800
2000
t (sec)
Delay Embedding
Subsampled PC
Persistence Diagram
30
20
10
0
(d)
(e)
(f)



Walking
Waving
Bicycle
Golfing
PD
R-wrist
L-leg
0
20
40
60

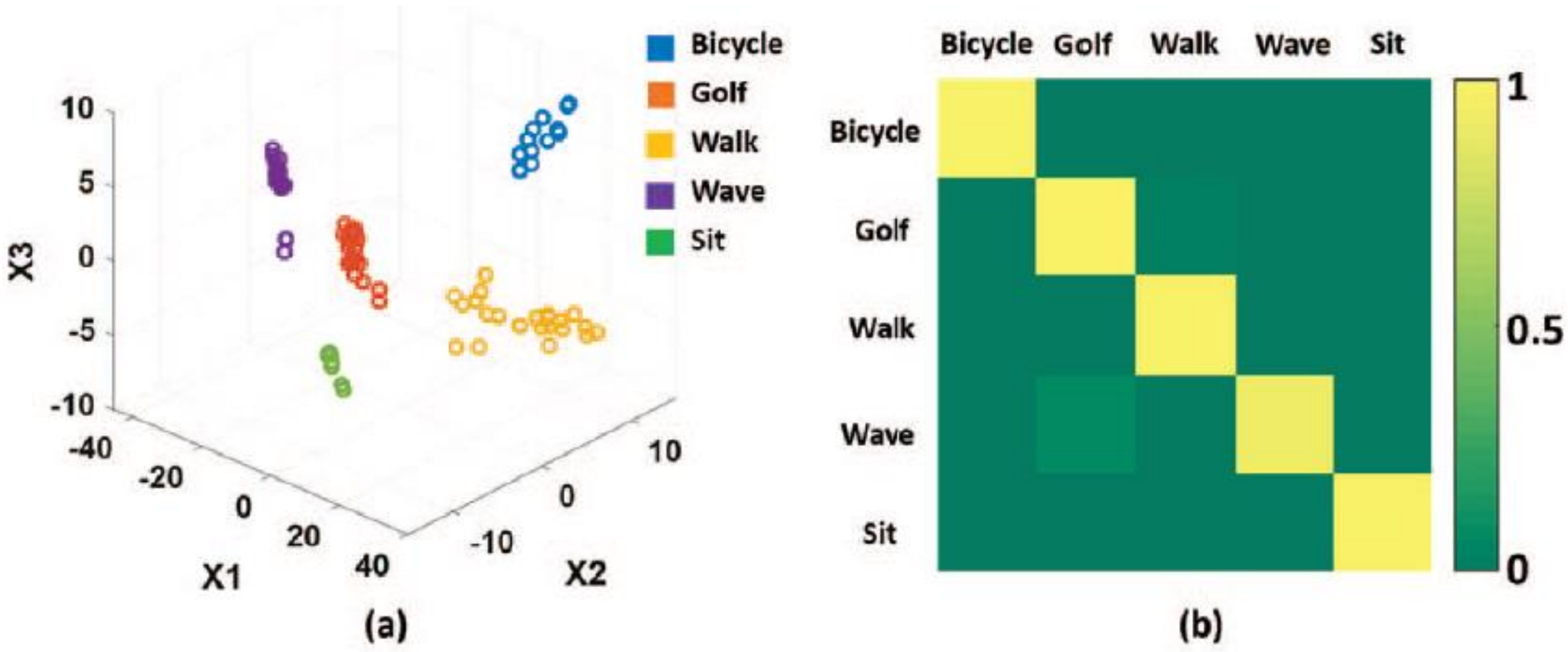


Fig. 4. (a) Separation of classes from the training set, (b) Confusion matrix over the predicted and true classes

Table 1. Class accuracy results for the activities

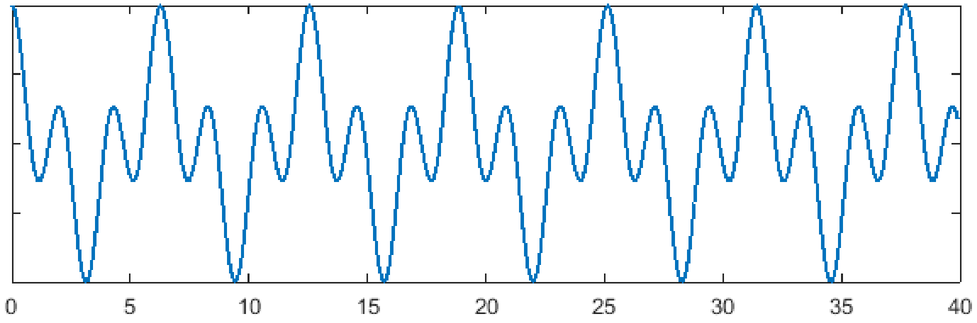
Action	Bicycle	Golf	Walk	Wave	Sit
Accuracy	1.00	0.9787	0.9929	0.9858	1.00

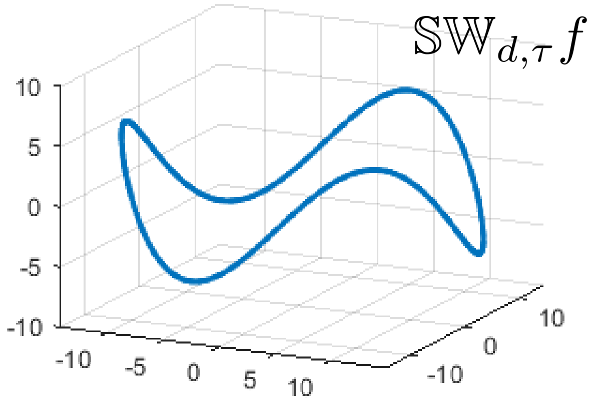
SLIDING WINDOW PERSISTENCE OF QUASIPERIODIC FUNCTIONS

HITESH GAKHAR AND JOSE A. PEREA

arXiv:2103.04540

$$f(t) = \cos(t) + \cos(3t)$$

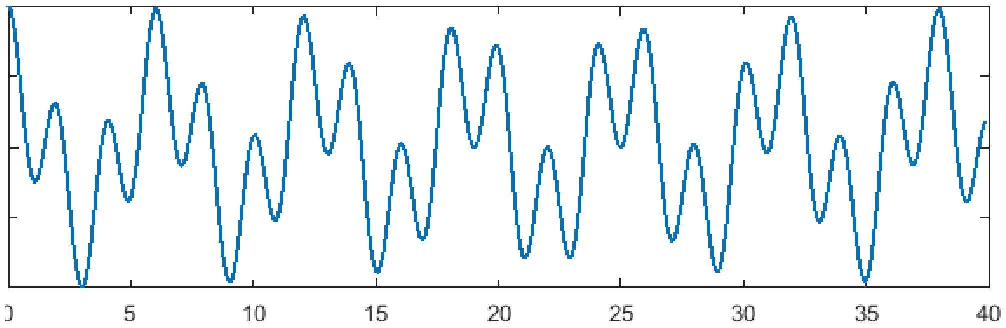


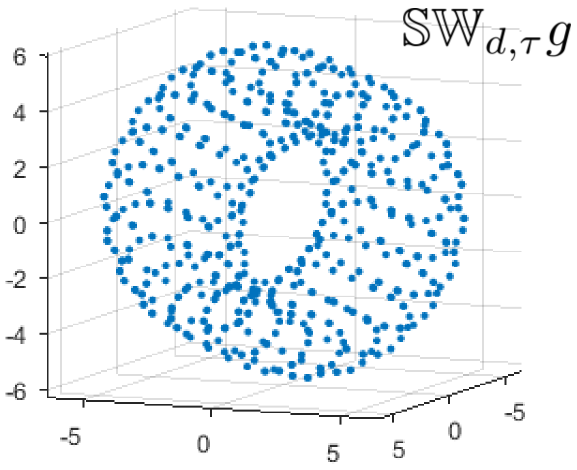


Commensurate

$$\frac{1}{3} \in \mathbb{Q}$$

$$g(t) = \cos(t) + \cos(\pi t)$$



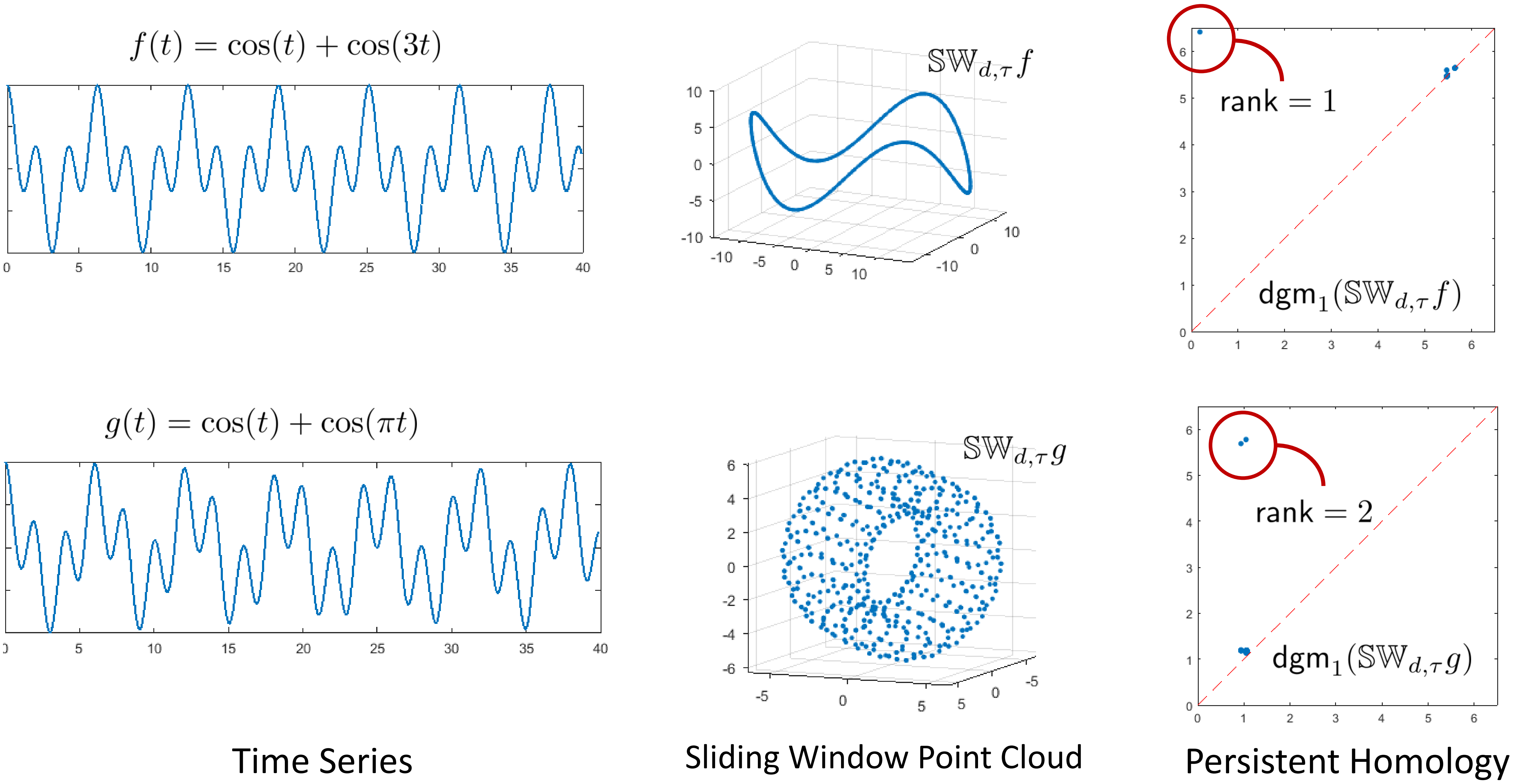


Non-Commensurate

$$\frac{1}{\pi} \notin \mathbb{Q}$$

Time Series

Sliding Window Point Cloud



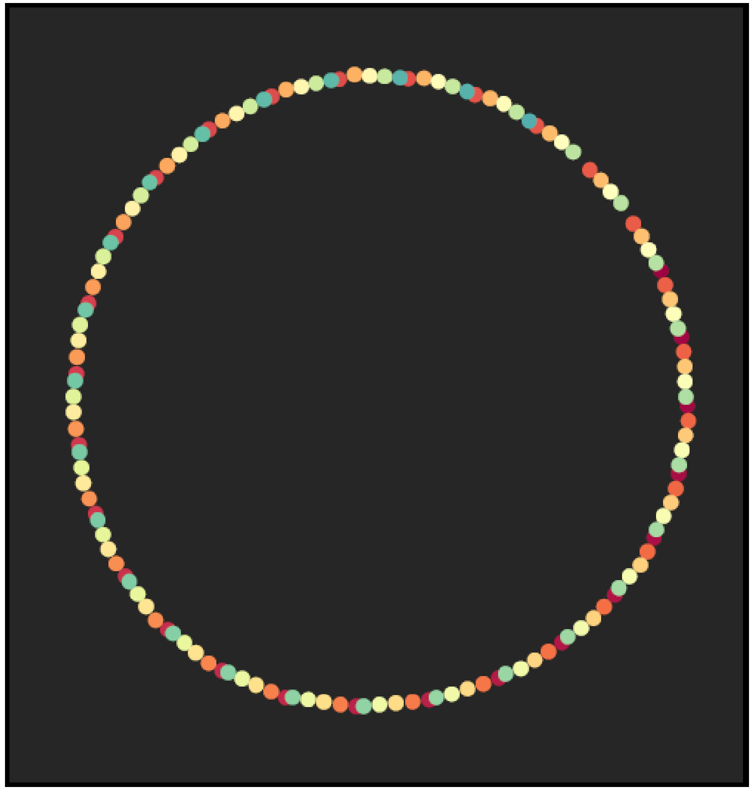
$f(t) = \cos(t) + \cos(3t)$
$\mathbb{SW}_{d,\tau} f$
rank = 1
$\mathrm{dgm}_1(\mathbb{SW}_{d,\tau} f)$
$g(t) = \cos(t) + \cos(\pi t)$
$\mathbb{SW}_{d,\tau} g$
rank = 2
$\mathrm{dgm}_1(\mathbb{SW}_{d,\tau} g)$
Time Series
Sliding Window Point Cloud
Persistent Homology

SIAM J. IMAGING SCIENCES
Vol. 11, No. 2, pp. 1049–1077

(Quasi)Periodicity Quantification in Video Data, Using Topology*

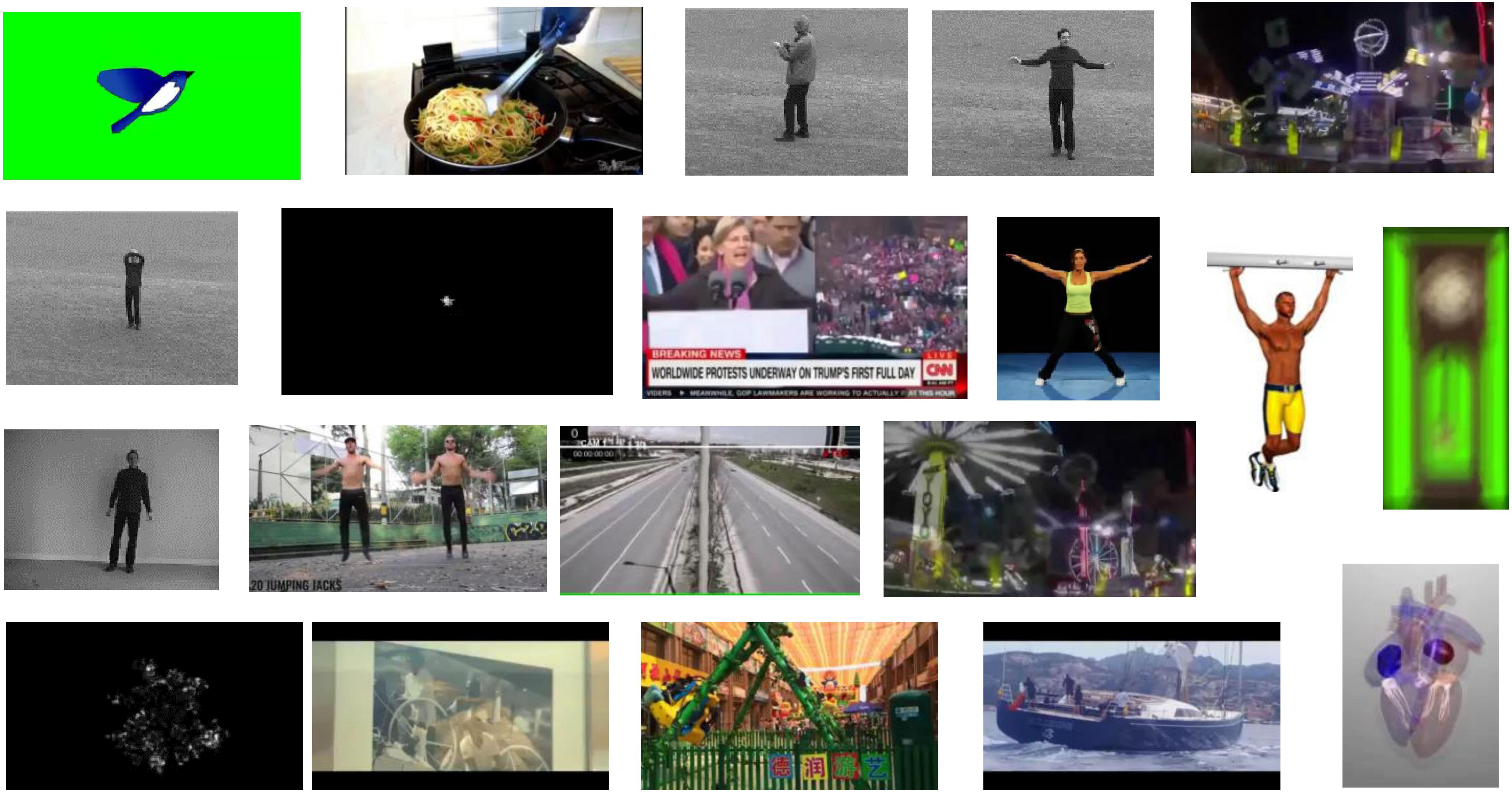
Christopher J. Tralie† and Jose A. Perea‡

Recurrence in video data (SW1PerS-video)

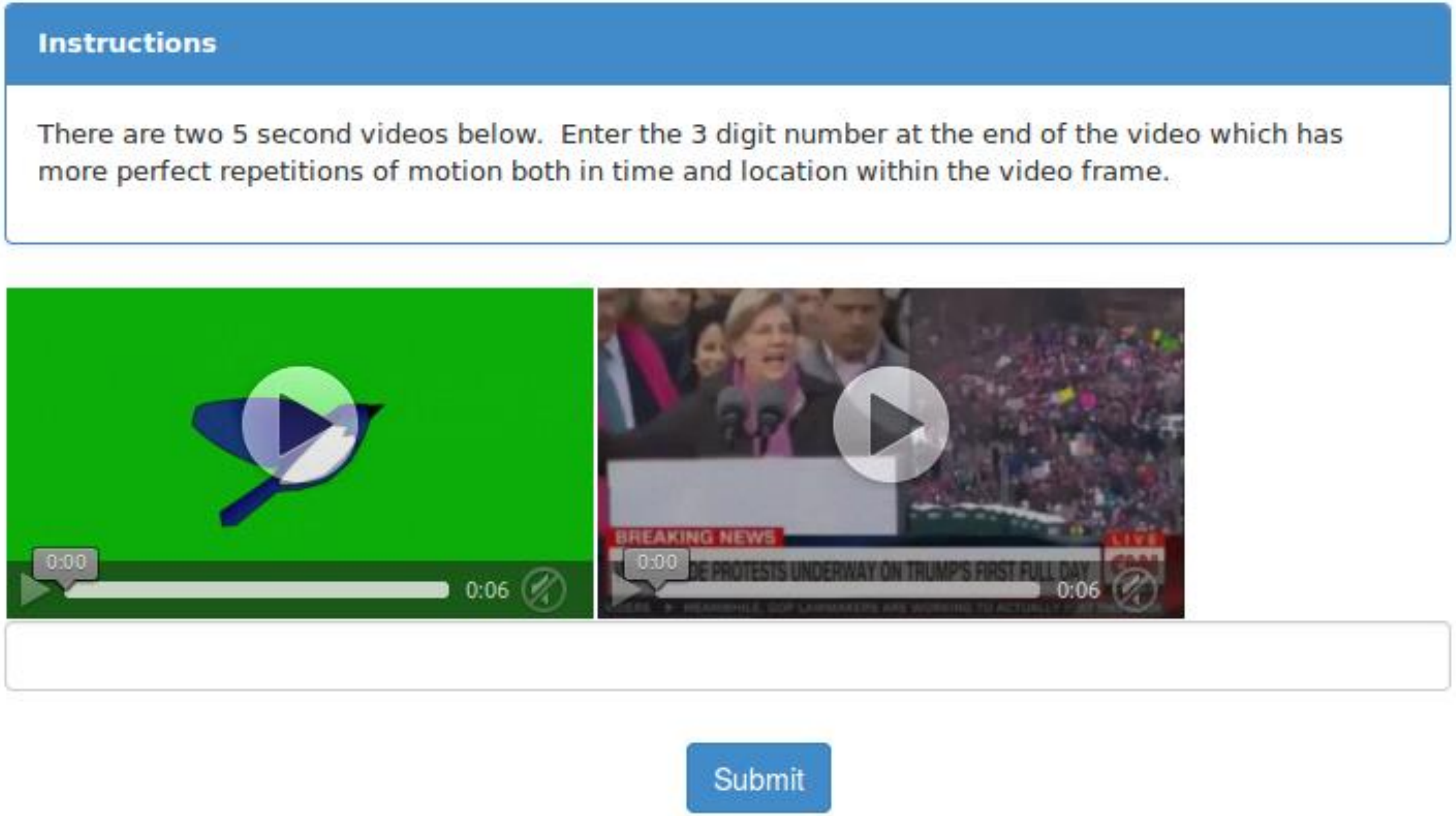


Sliding window embedding

C. Tralie and J. Perea, *(Quasi)Periodicity Quantification in Video Data, Using Topology,* 2018.


BREAKING NEWS
WORLDWIDE PROTESTS UNDERWAY ON TRUMP'S FIRST FULL DAY
CNN
LIVE
20 JUMPING JACKS
德润游艺

Experiment: amazon Mechanical Turk



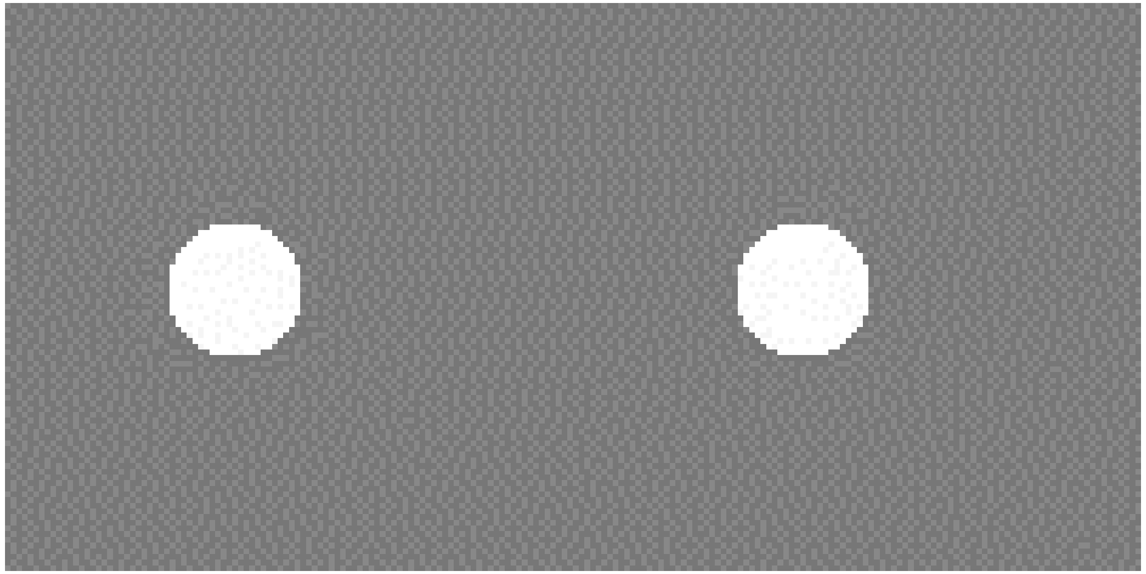
C. Tralie and J. Perea, *(Quasi)Periodicity Quantification in Video Data, Using Topology,* 2018.

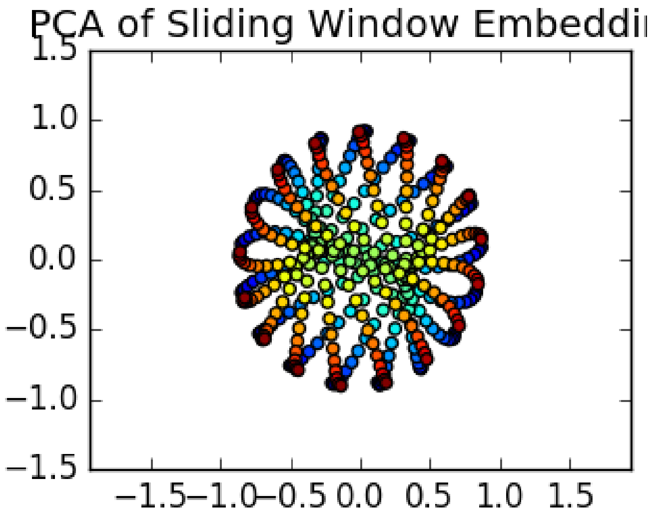
Results: Humans (amazon turk) vs Computers

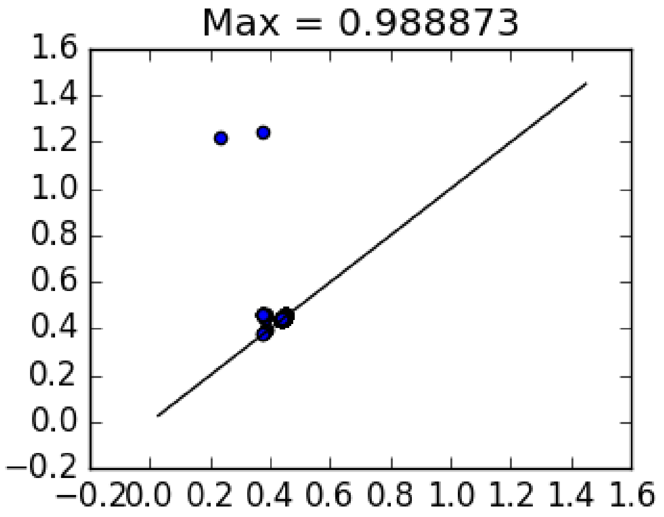
Correlation of rakings
(from most periodic to least periodic)
across 20 videos

Kendall's Tau	SW	CutlerDavis Freq	CutlerDavis Lattice	Humans
SW	1	-0.315	0.221	**0.663**
CutlerDavis Freq		1	-0.0842	-0.294
CutlerDavis Lattice			1	0.347
Humans				1

Recurrence in video data (SW1PerS-video)



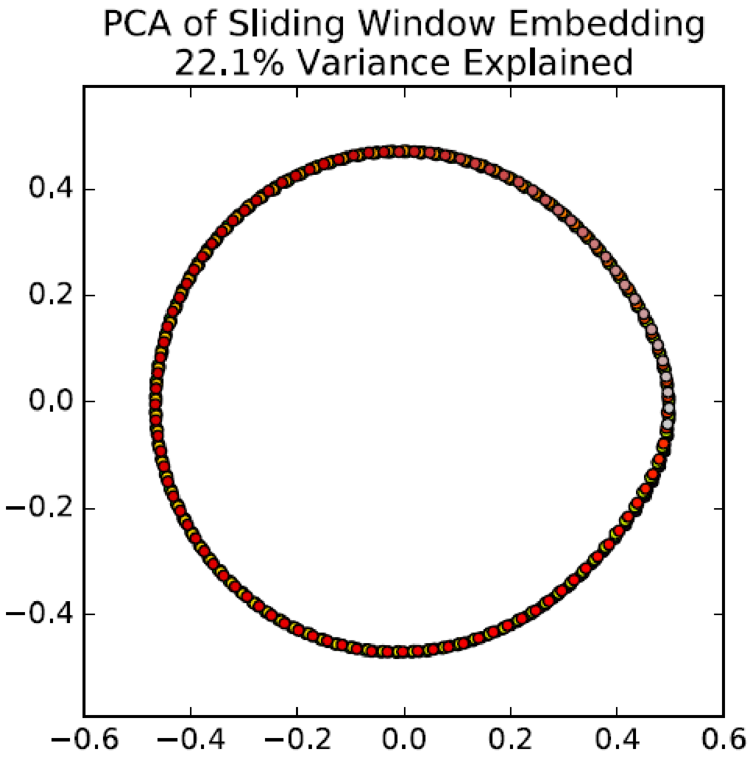


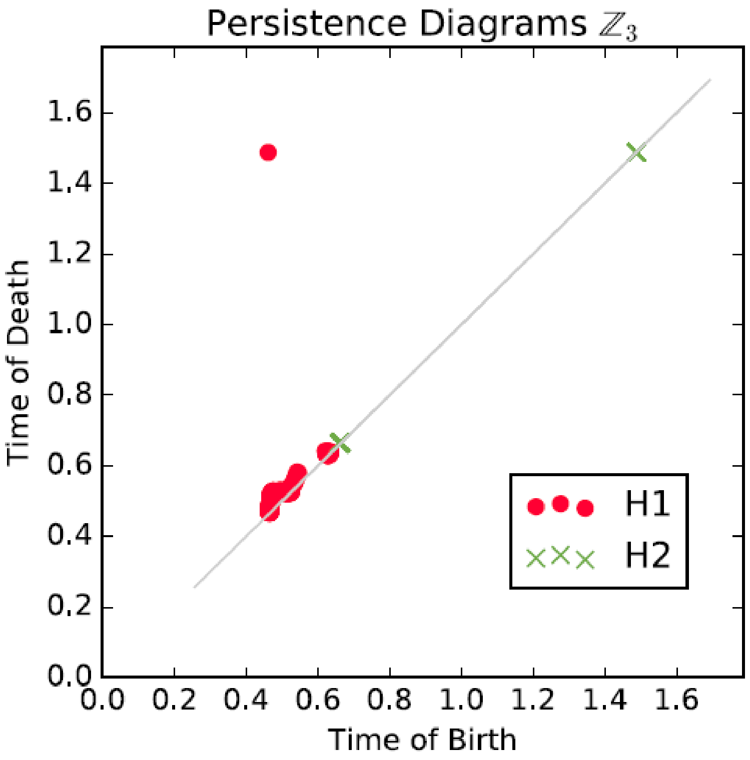


C. Tralie and J. Perea, *(Quasi)Periodicity Quantification in Video Data, Using Topology,* 2018.

Recurrence in video data (SW1PerS-video)

normal

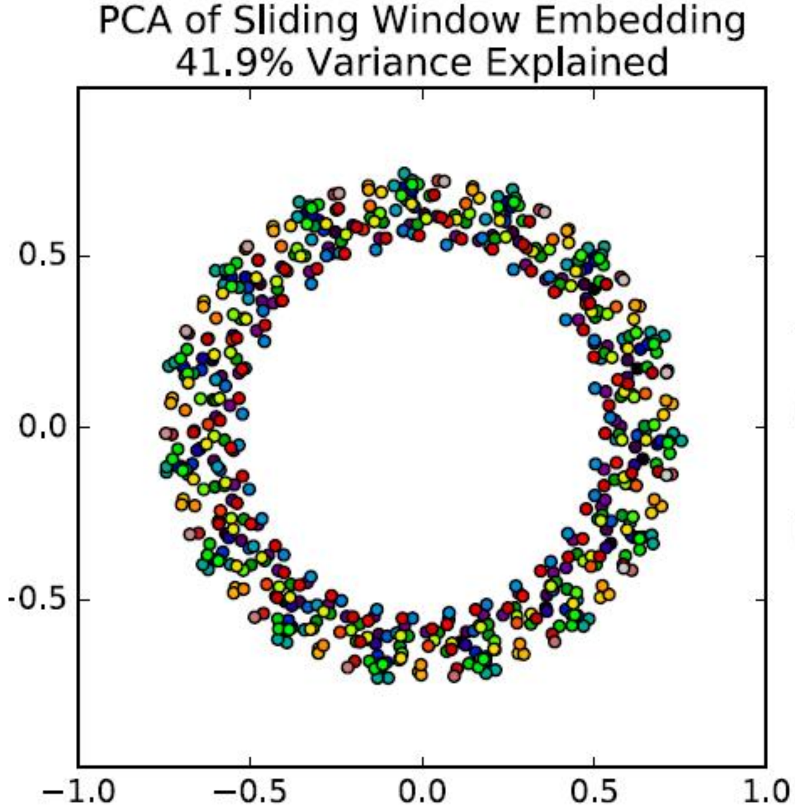


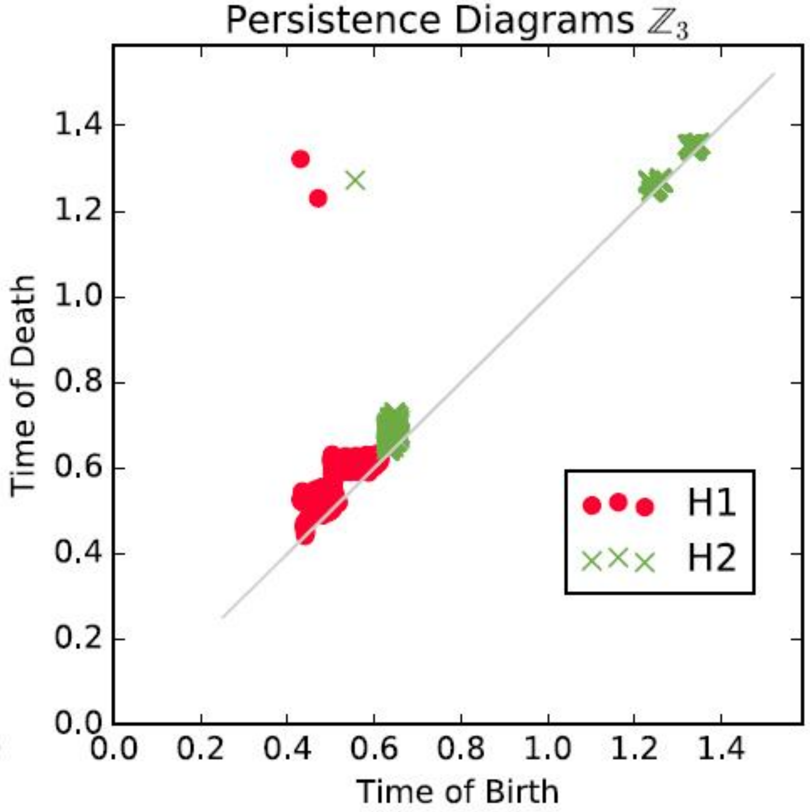


C. Tralie and J. Perea, *(Quasi)Periodicity Quantification in Video Data, Using Topology,* 2018.

Recurrence in video data (SW1PerS-video)

Clinical asymmetry



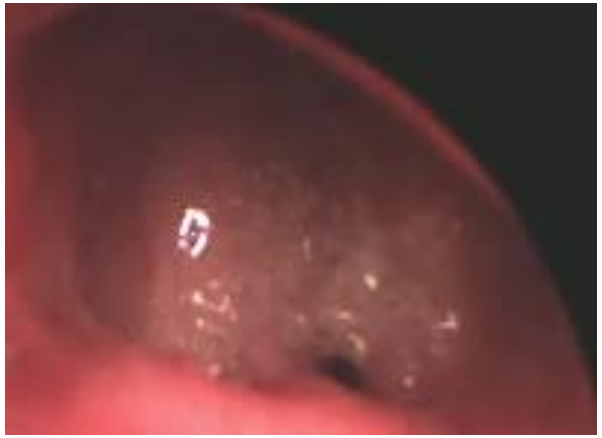


C. Tralie and J. Perea, *(Quasi)Periodicity Quantification in Video Data, Using Topology,* 2018.

Laryngeal video-endoscopy

normal

Clinical asymmetry



Mucus irregular

AP-Biphonation

C. Tralie and J. Perea, *(Quasi)Periodicity Quantification in Video Data, Using Topology*, 2018.

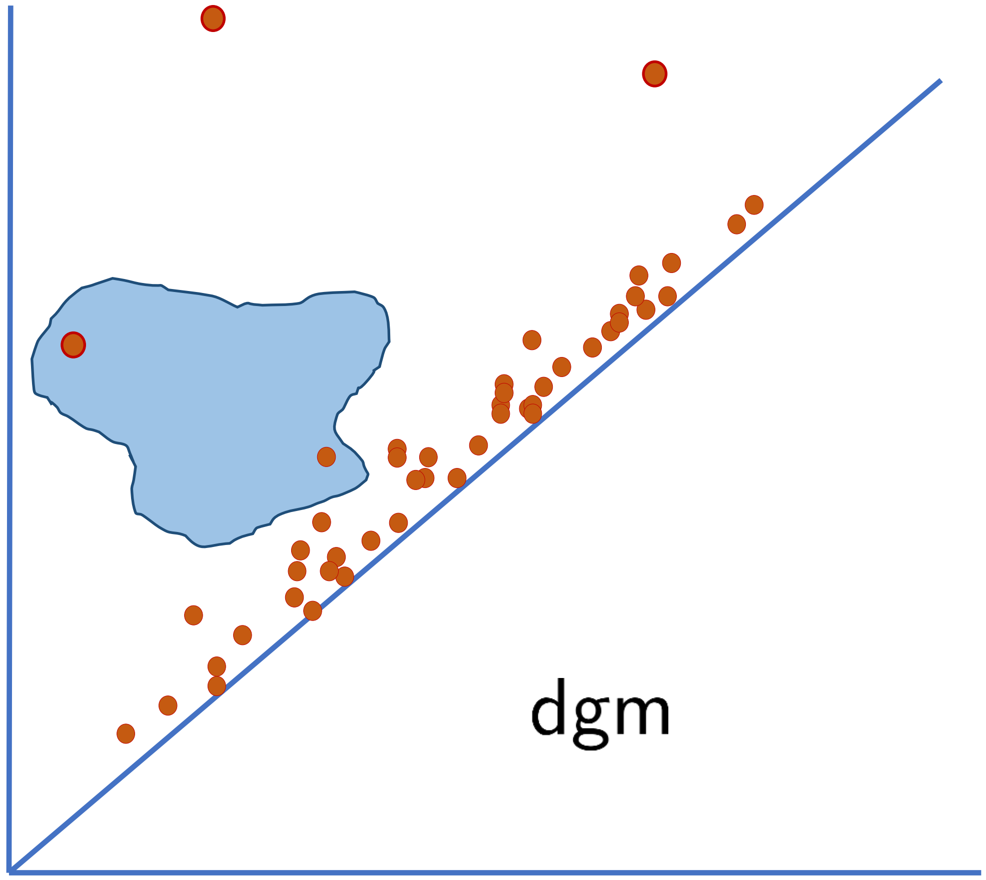
$$\mathsf{dgm} \mapsto predict(\mathsf{dgm}) \in \mathbb{R}$$

Approximating Continuous Functions on Persistence Diagrams Using Template Functions

Jose A. Perea · Elizabeth Munch · Firas A. Khasawneh

To appear in **FoCM** 2022
https://arxiv.org/abs/1902.07190

$f \in C_c(\mathbb{W}, \mathbb{R})$ Continuous and compactly supported



$$\mathbb{W} = \{(x, y) \in \mathbb{R}^2 \;:\; 0 \leq x < y\}$$

Remark:

$$\text{dgm} \mapsto \sum_{\mathbf{x} \in \text{dgm}} f(\mathbf{x})$$

Is continuous.

J. A. Perea, E. Munch and F. Khasawneh, *Approximating Continuous Functions on Persistence Diagrams*, **FoCM** *2022*

Theorem

Let $\mathcal{C} \subset \mathscr{D}$ be compact and let $F : \mathcal{C} \longrightarrow \mathbb{R}$ be continuous.

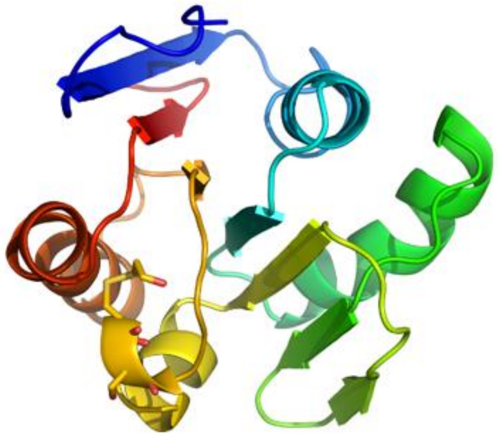
Then, given $\epsilon > 0$, there exist functions $f_1, \ldots, f_n \in C_c(\mathbb{W}, \mathbb{R})$

and a polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ so that

$$\left| p \left(\sum_{\mathbf{x} \in \mathsf{dgm}} f_1(\mathbf{x}), \ldots, \sum_{\mathbf{x} \in \mathsf{dgm}} f_n(\mathbf{x}) \right) - F(\mathsf{dgm}) \right| < \epsilon$$

for every $\mathsf{dgm} \in \mathcal{C}$.

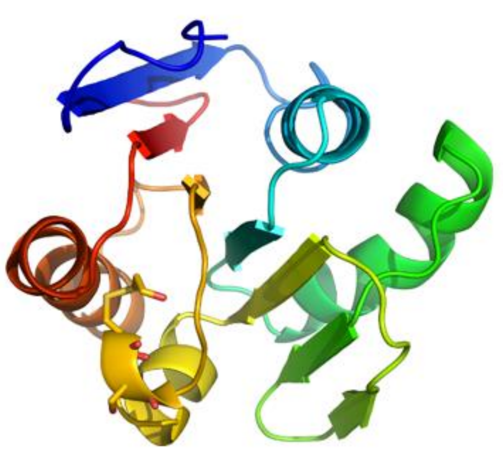
J. A. Perea, E. Munch and F. Khasawneh, *Approximating Continuous Functions on Persistence Diagrams,* **FoCM** *2022*

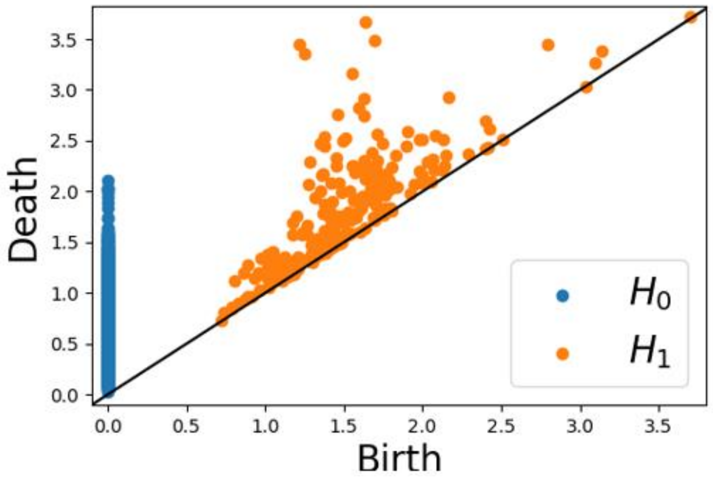


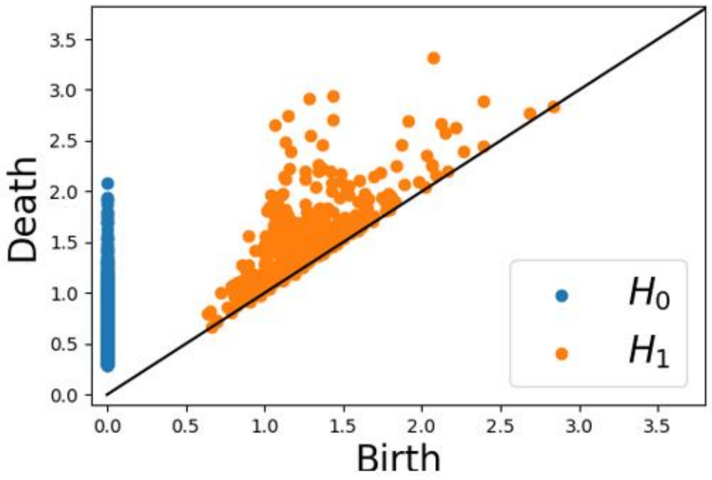
Protein Classification Benchmark Collection (PCB00019) – SCOP40mini

1,357 proteins	# atoms ~ 1K	55 classification tasks

A protein classification benchmark collection for machine learning, P. Sonego et. al., Nucleic acids research, 2007.







		Train	Test
CDER	Polynomial	0.90 ± 0.07	0.98 ± 0.02
	RBF	0.91 ± 0.06	0.97 ± 0.02
	Sigmoid	0.90 ± 0.07	0.98 ± 0.02
Topological features in [4]		-	$0.82 \pm$ —-

L. Polanco and J. A. Perea, *Adaptive template systems: Data-driven feature selection for learning with persistence*, **ICMLA** 2019

Thanks!

http://www.joperea.com