Topological Analysis of Modern Data

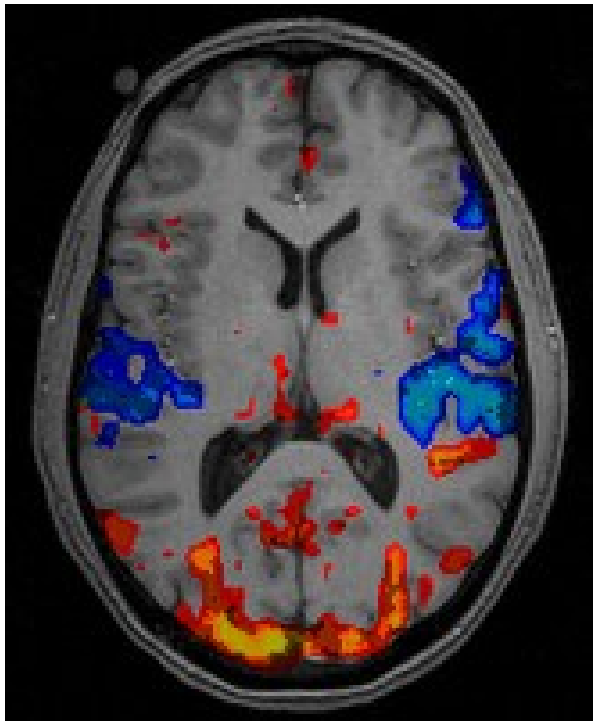
Chao Chen

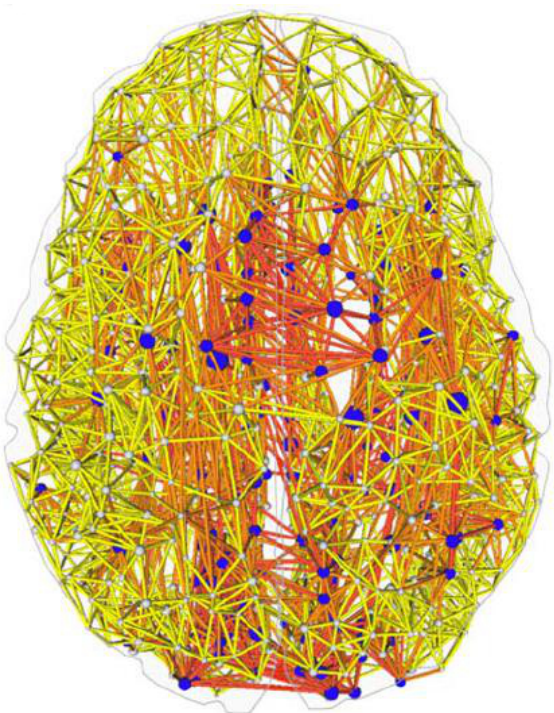
Rutgers University

April 2015

Modern: Huge Volume, High Dimension, Complex

- Neuroscience ($300 million [White House])
- The largest collection of functional Magnetic Resonance Imaging (fMRI) datasets, 3M scans, 135K dimension



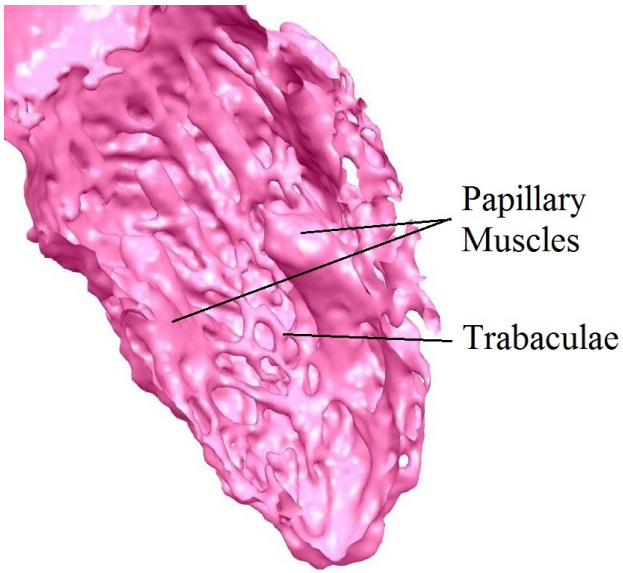


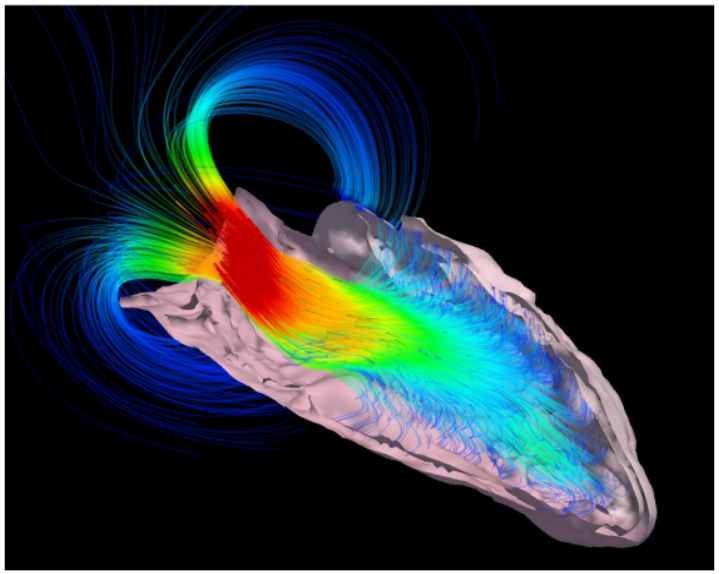
Pic from Sporns

Collaboration: Han Liu (Princeton). Funding opportunities: NIH-NIA/NIMH

Modern: Huge Volume, High Dimension, Complex

- High resolution cardiac data, > 10 GB per data/patient
  Collaboration: Leon Axel (NYU Medical School).
  Funding opportunities: NIH-NHLBI





- Climate data, Social science, Genomic data, etc.

Modern Data: New Opportunities, New Challenges

- Learning: e.g. better diagnosis/prediagnosis
- Exploration: deeper understanding through data
  - Enable domain experts to explore the data
  - Communicate learning results effectively

Modern Data: New Opportunities, New Challenges

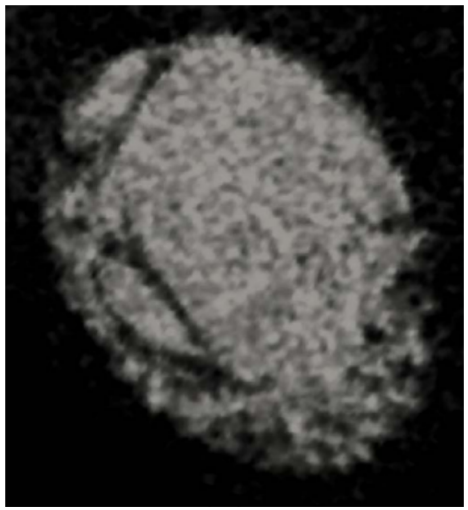
- Learning: e.g. better diagnosis/prediagnosis
- Exploration: deeper understanding through data
  - Enable domain experts to explore the data
  - Communicate learning results effectively

Structure is the key!

Why Topology Data Analysis?

Topology Data Analysis: a Robust Way to Extract Structures of Data

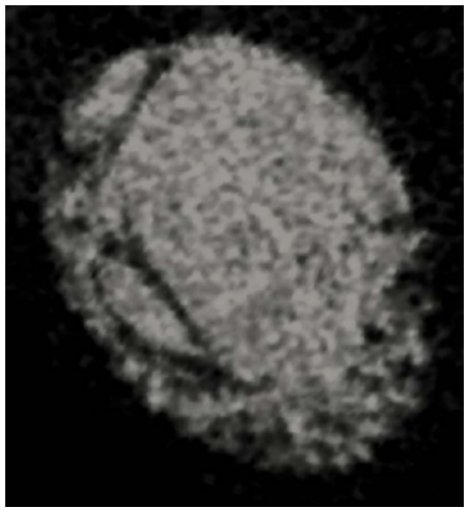
- cardiac data (**Demo** Visible Heart)

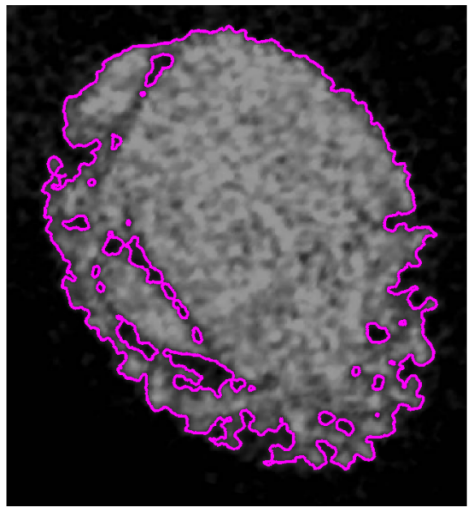


Why Topology Data Analysis?

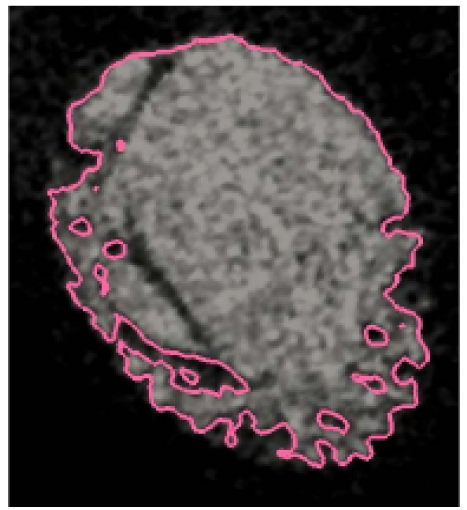
Topology Data Analysis: a Robust Way to Extract Structures of Data

- cardiac data (**Demo** Visible Heart)





Thresholding

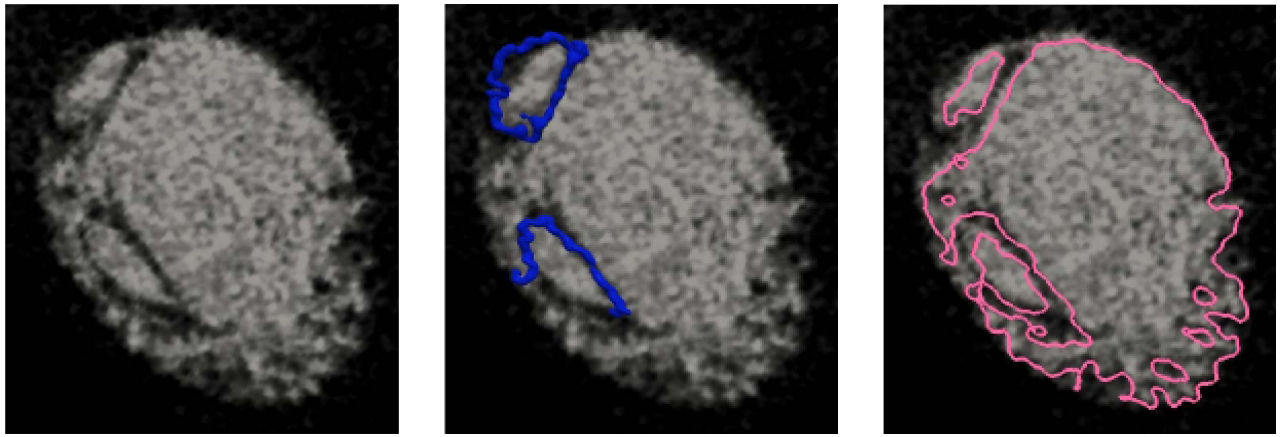


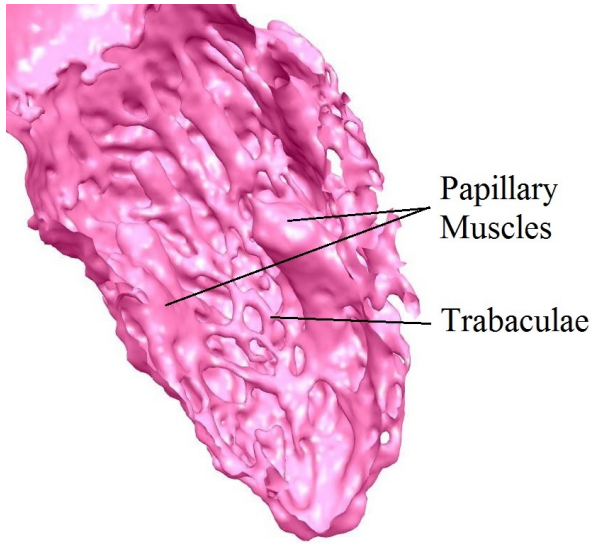
Advanced

- Thresholding: local evidence
- Advanced: pairwise local evidence

Why Topology Data Analysis?

Recovering missing trabeculae:

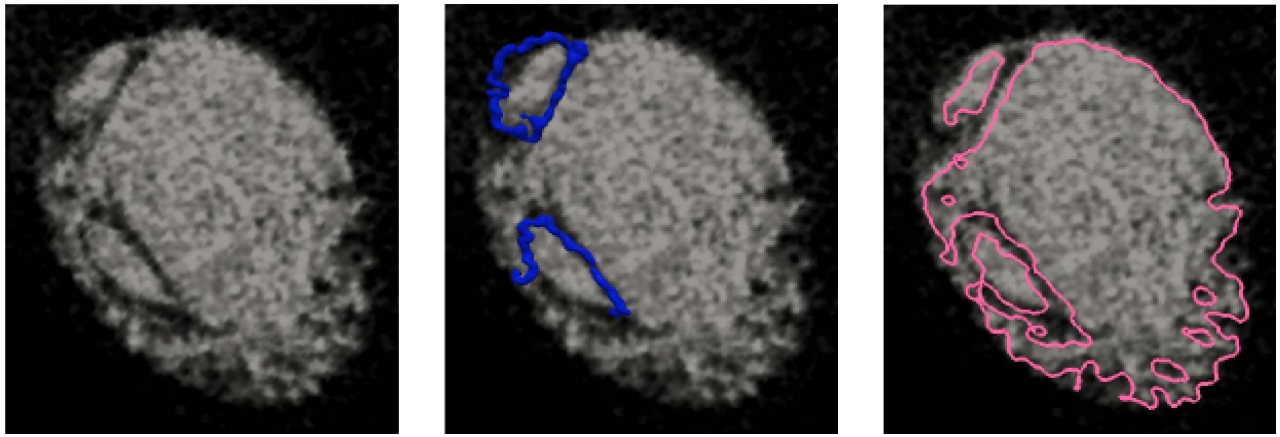


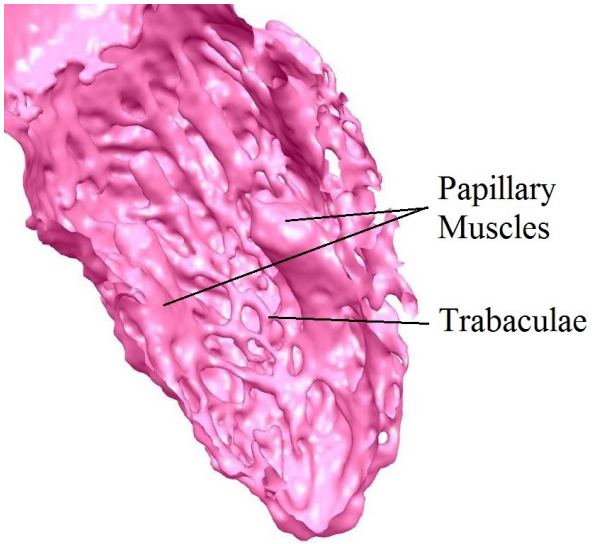


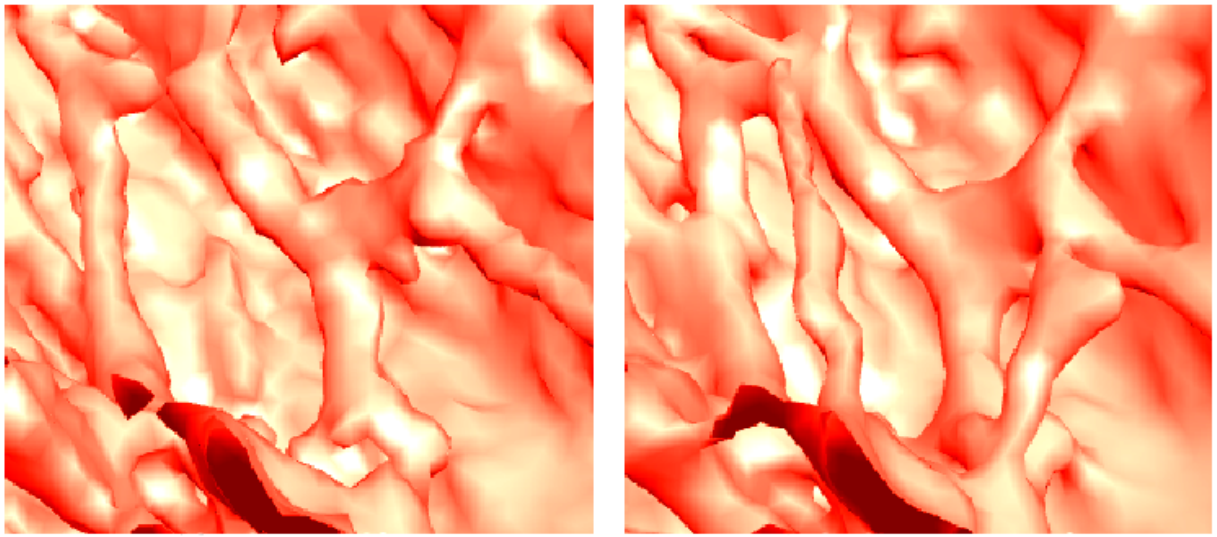
[Gao, **Chen** , *et al.* IPMI 2013]

Why Topology Data Analysis?

Recovering missing trabeculae:



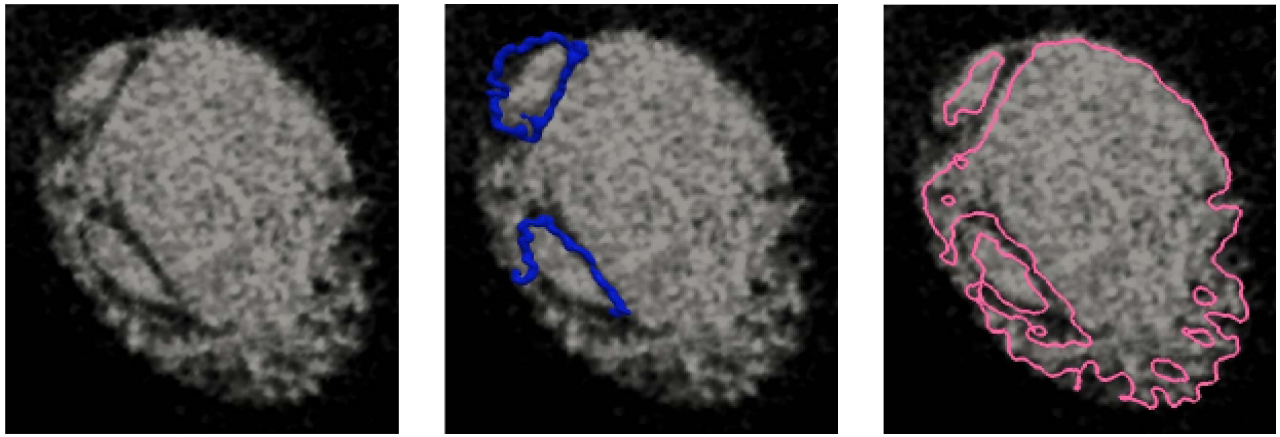


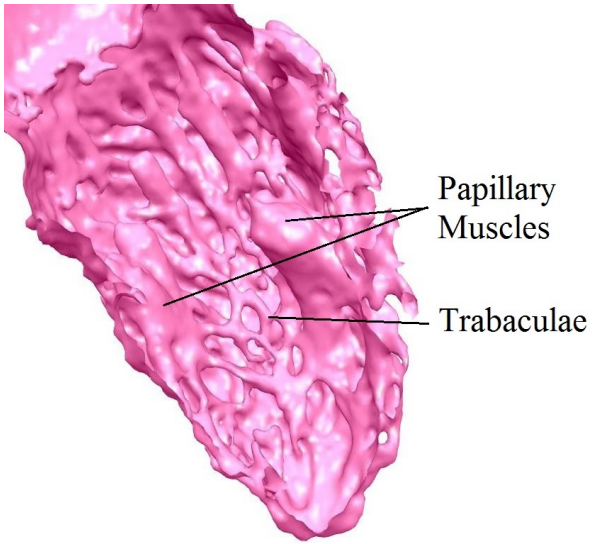


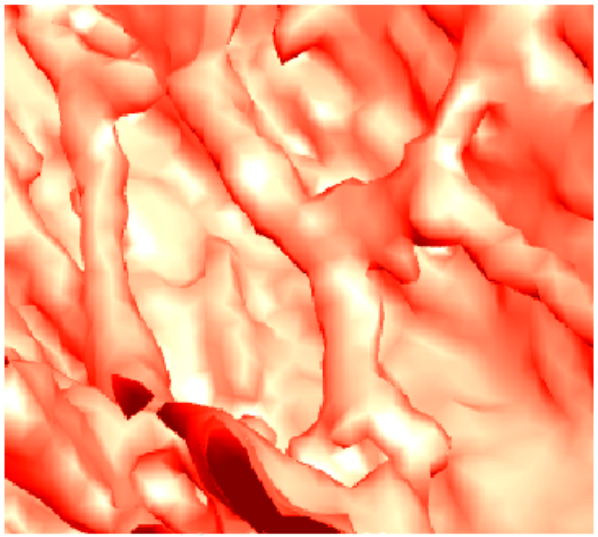
[Gao, **Chen** , *et al.* IPMI 2013]

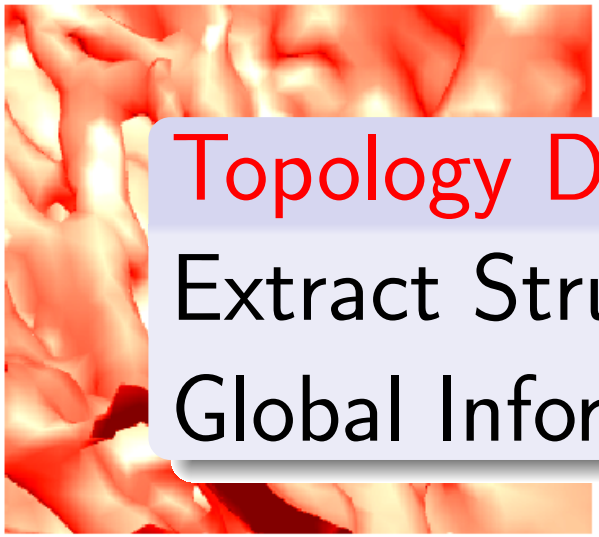
Why Topology Data Analysis?

Recovering missing trabeculae:









Topology Data Analysis

Extract Structures Using Global Information.

[Gao, **Chen** , *et al.* IPMI 2013]

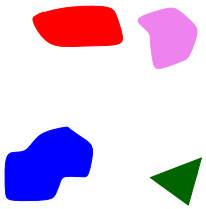
Outline

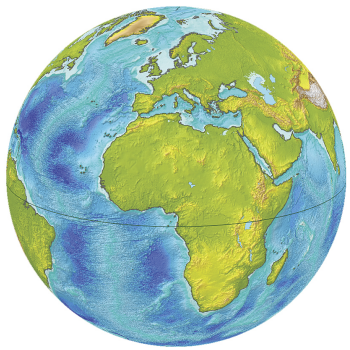
Topology 101

Topology: the science of structures.

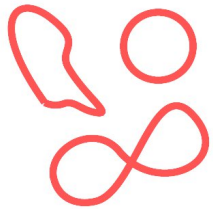
- The most global structural information.
- Forgetting local deformation.

0 dim: components

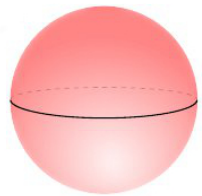




1 dim: loops

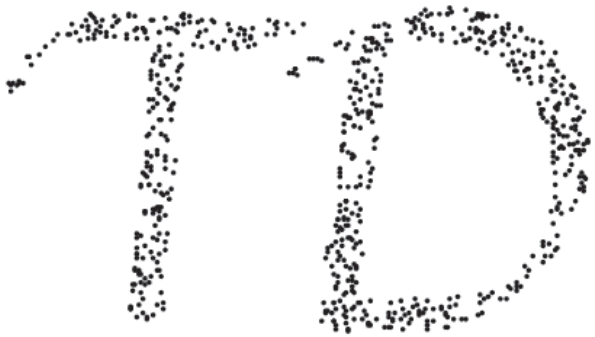


2 dim: voids



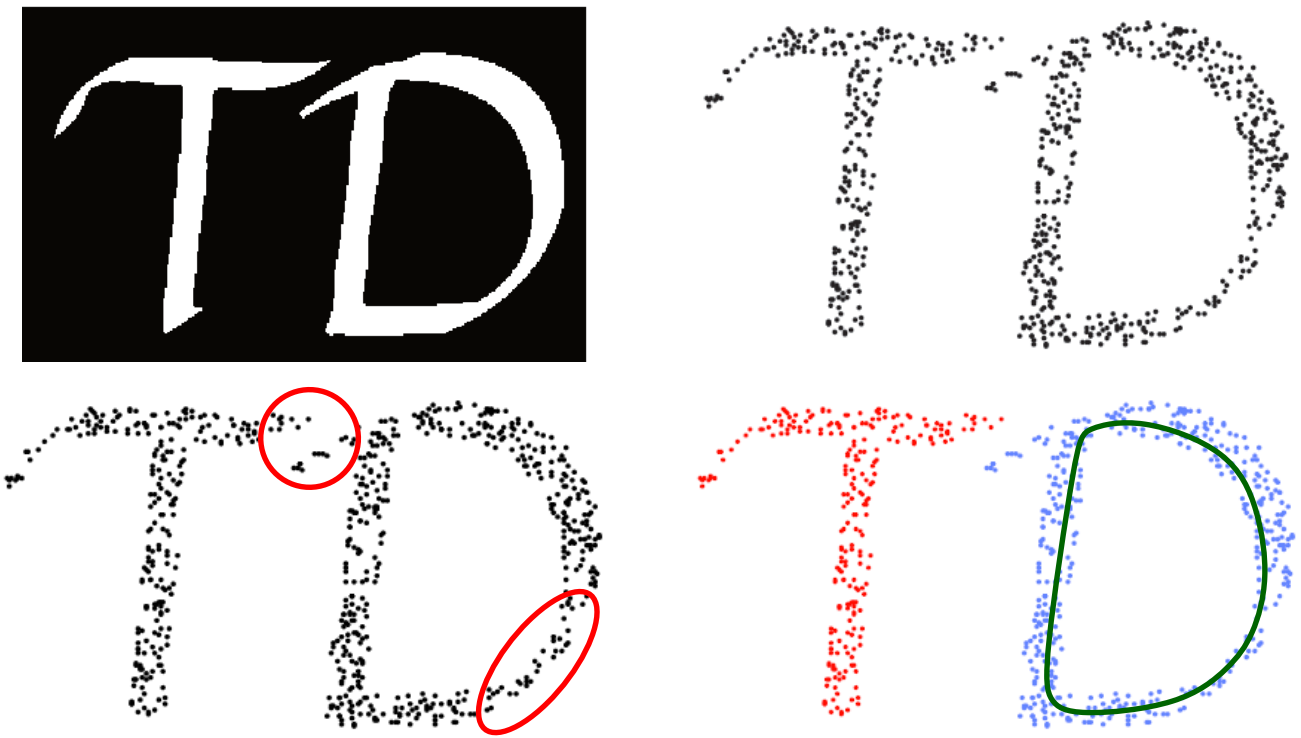
Topological Structures of Data

- For a dataset, what are the components and loops of the data?
- TDA: detect these structures in a robust way.



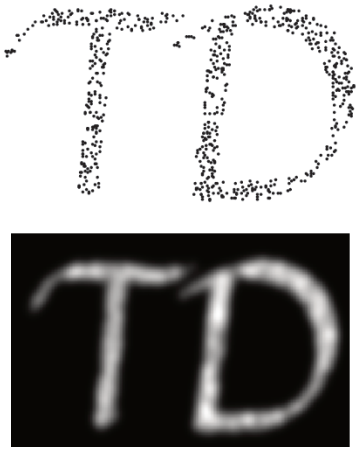
Topological Structures of Data

- For a dataset, what are the components and loops of the data?
- TDA: detect these structures in a robust way.

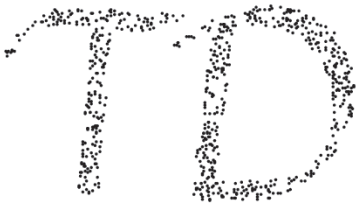


Persistent Homology

- Input: a density function, f
- Output:
  topo. structures & their persistence

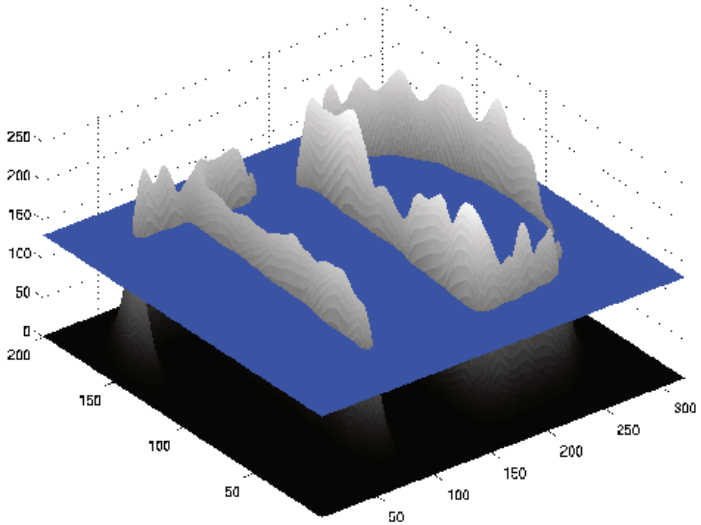


Persistent Homology



- Input: a density function, f
- Output:
  topo. structures & their persistence

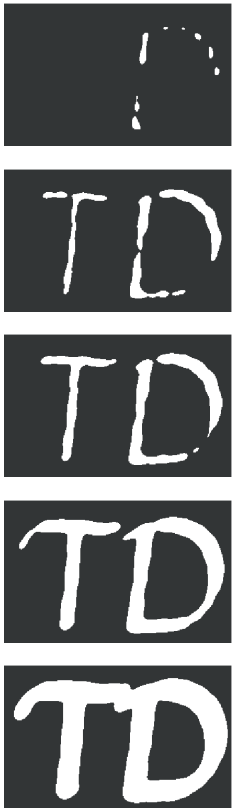
- Def: given threshold t, the superlevel set $f^{-1}[t, +\infty) := \{x | f(x) \geq t\}$

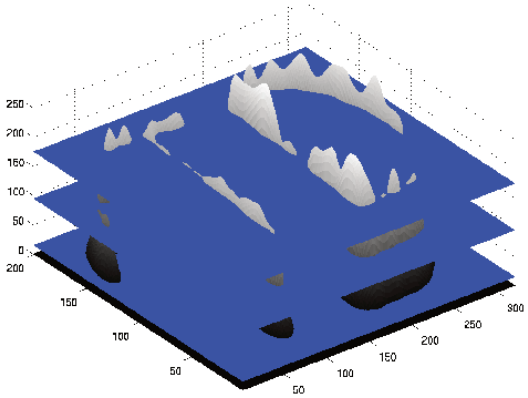


Persistent Homology (continued)

- the true structures are hidden in superlevel sets
- consider the whole stack of superlevel sets
- identify structures that often appear (high persistence)
- Output: persistence diagram – dots representing all structures

Demo

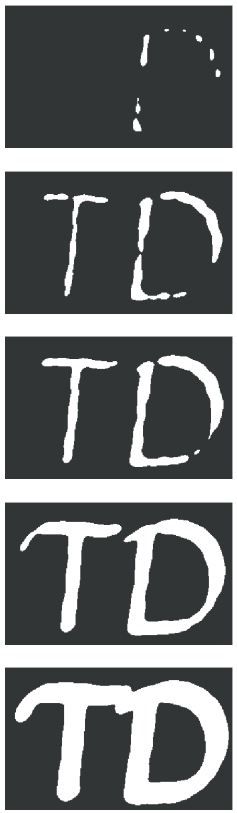


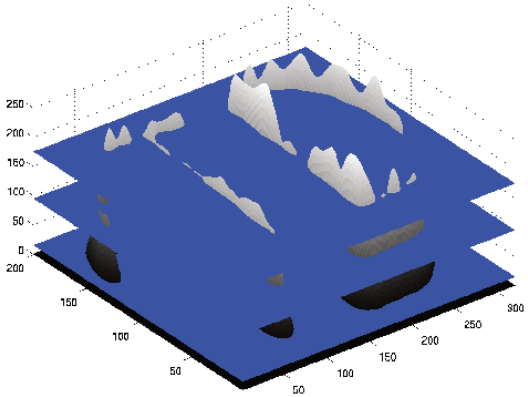


Persistent Homology (continued)

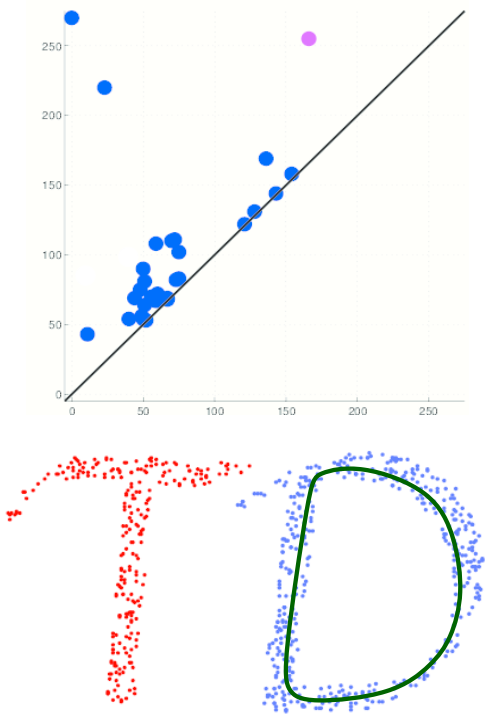
- the true structures are hidden in superlevel sets
- consider the whole stack of superlevel sets
- identify structures that often appear (high persistence)
- Output: persistence diagram – dots representing all structures

Demo





Diagram

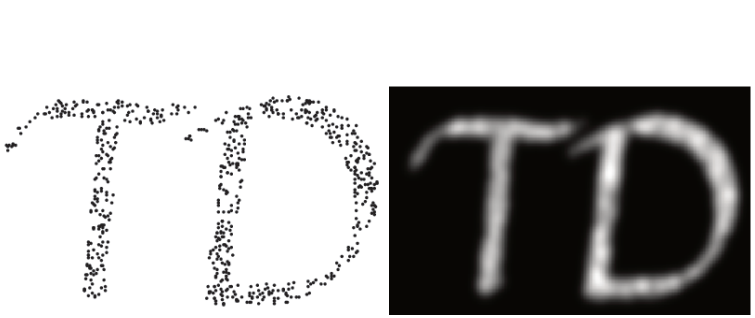


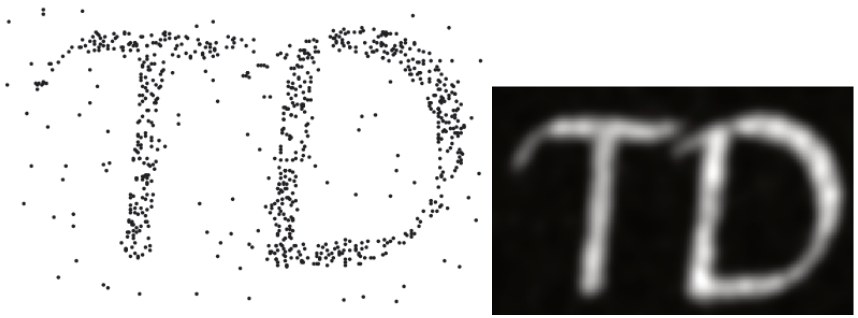
Theoretical Guarantee: Stability

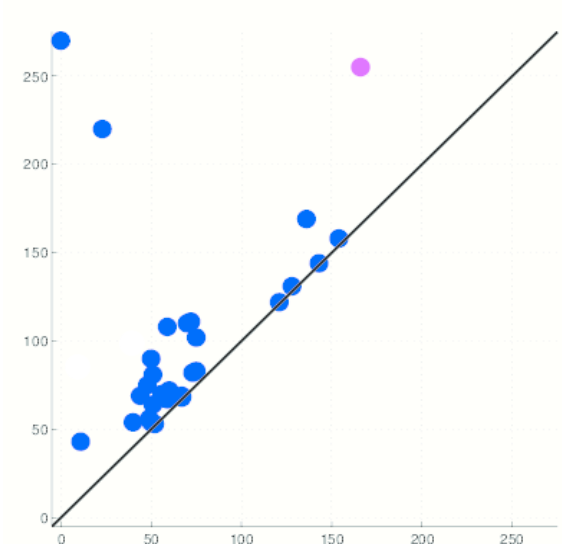
The Stability Theorem [Cohen-Steiner *et al.* DCG 2007]

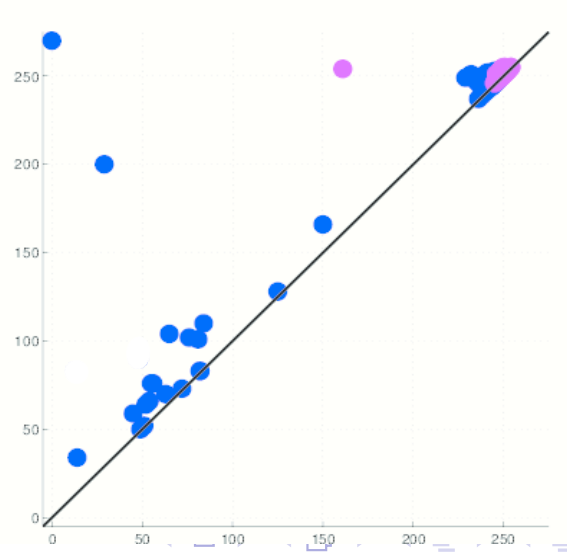
$$\|diagram(f) - diagram(g)\| \leq \|f - g\|_\infty$$

need new pic



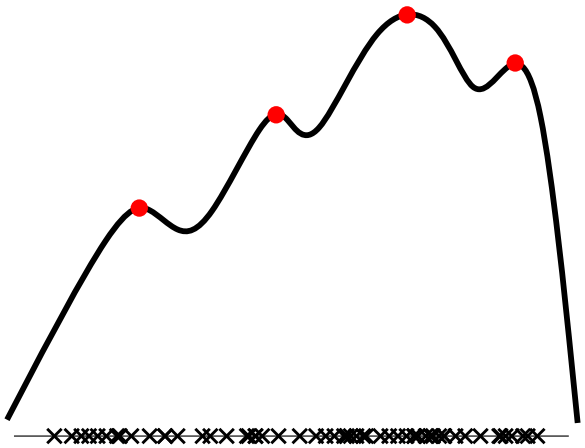


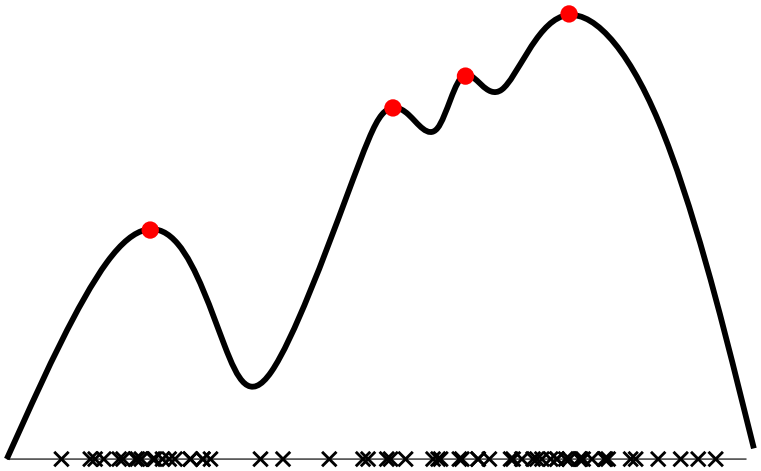




Connection to Known Structures

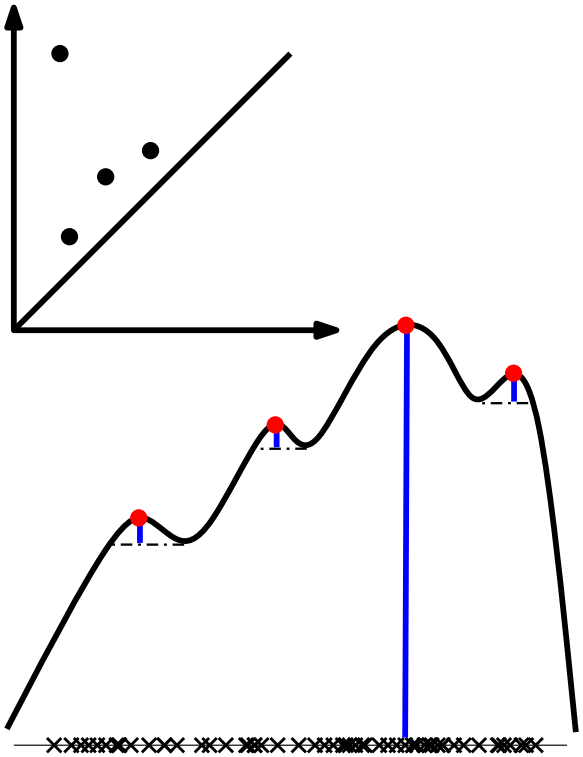
- Clusters and more

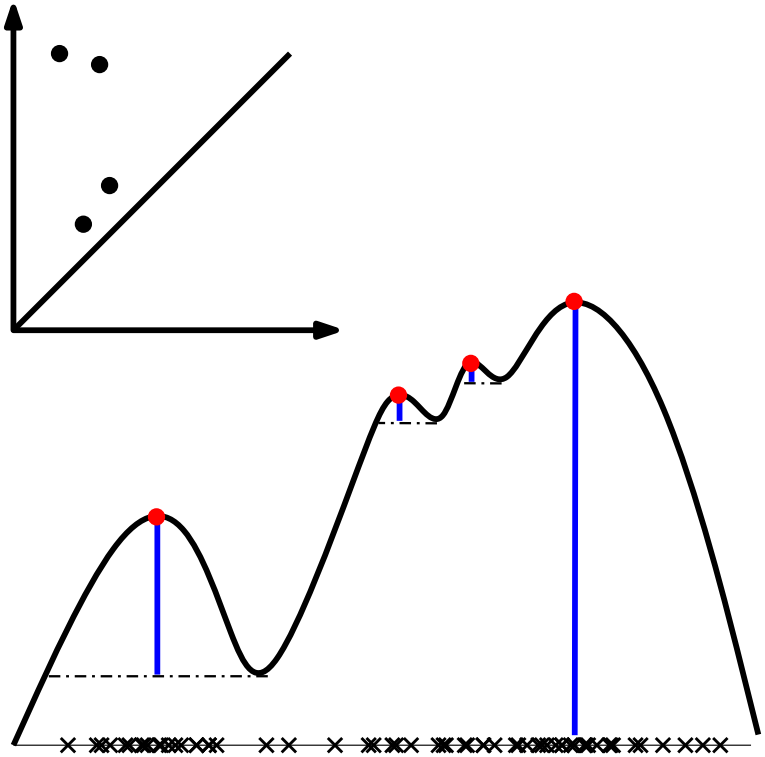




Connection to Known Structures

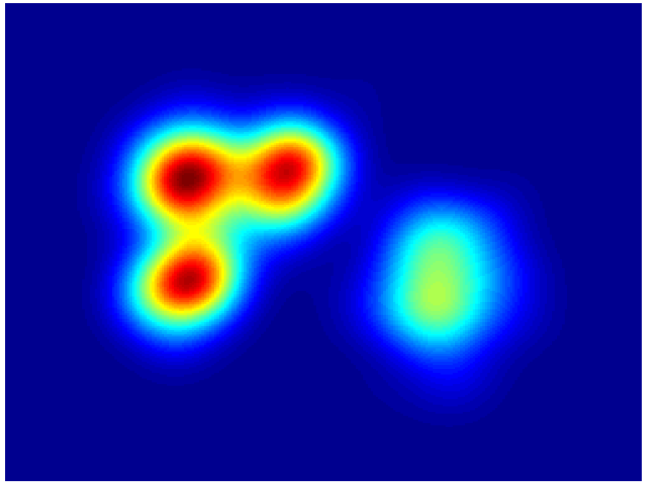
- Clusters and more





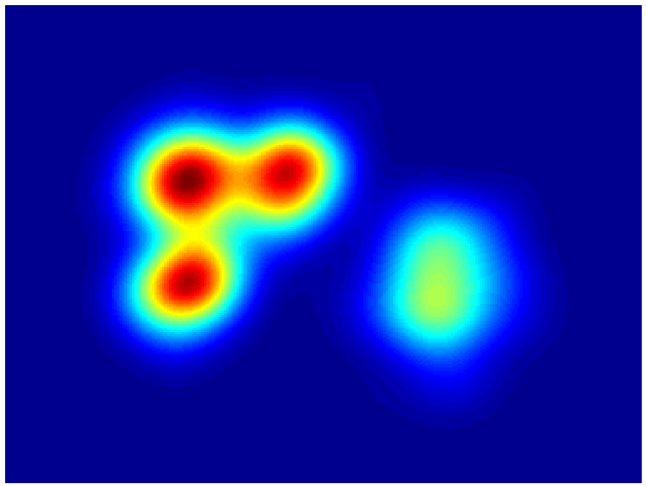
Connection to Known Structures

- Zero-dimensional structures: clusters

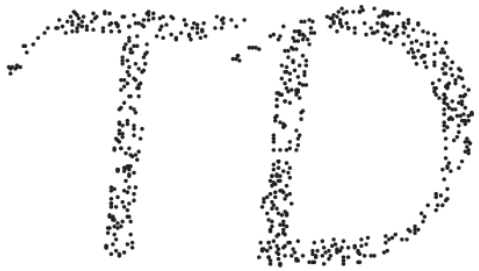


Connection to Known Structures

- Zero-dimensional structures: clusters



- High-dimensional structures

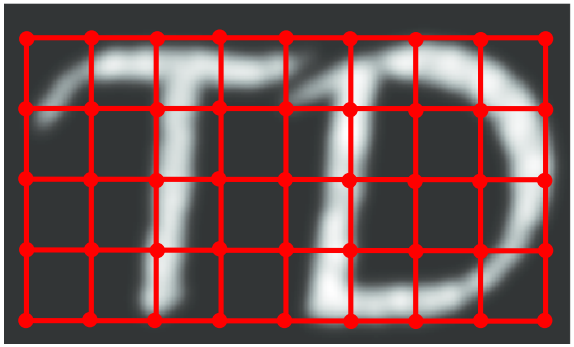


- Data living in D dimension:
  topological structures of dimension $0, \cdots, D-1$.

The Classic Algorithm

[Edelsbrunner *et al.* DCG 2002]
(based on algebraic topology)

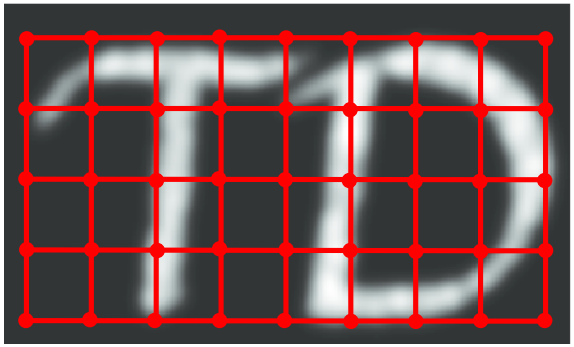
- Density estimation
- Discretize the domain
- Assign function values to elements

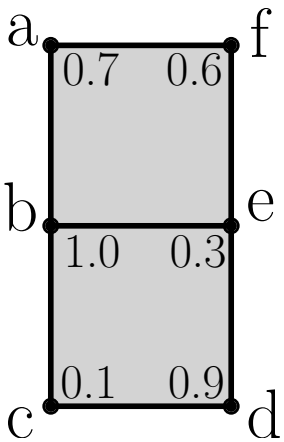


The Classic Algorithm

[Edelsbrunner *et al.* DCG 2002]
(based on algebraic topology)

- Density estimation
- Discretize the domain
- Assign function values to elements
- Sorted boundary matrices





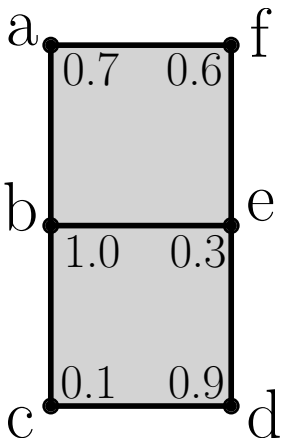
function values: 1.5, 0.0, 2.5, 1.3, 2.0, 1.7

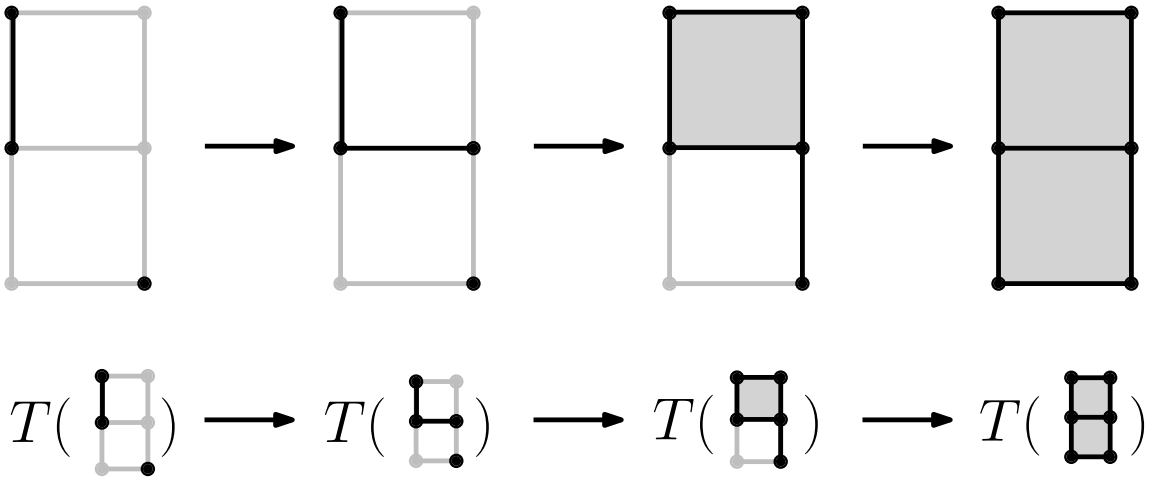
incidence matrix

	ab	af	be	de	ef	bc	cd
b	1		1			1	
d				1			1
a	1	1					
f		1			1		
e			1	1	1		
c						1	1

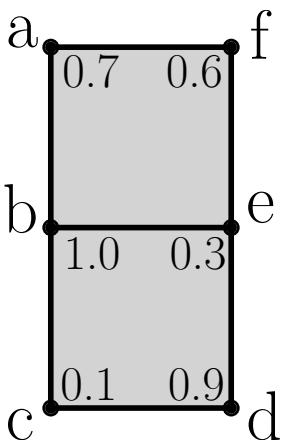
	$abef$	$bcde$
ab	1	
af	1	
be	1	1
de		1
ef	1	
bc		1
cd		1

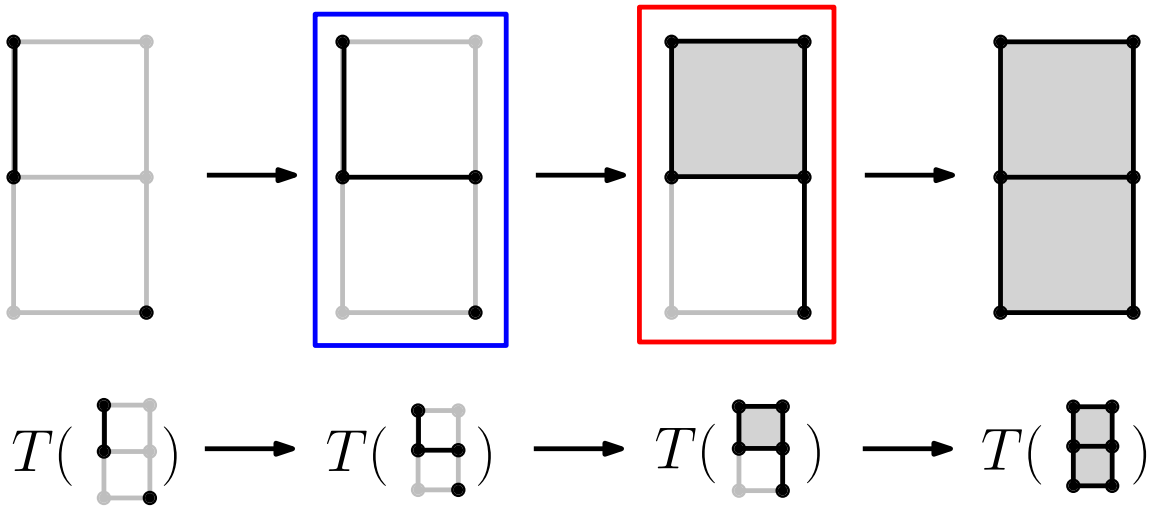
Computation: the Classic Algorithm





Computation: the Classic Algorithm





$$\begin{array}{c|ccccccc} & ab & af & be & de & ef & bc & cd \\ \hline b & 1 & & 1 & & & 1 & \\ d & & & & 1 & & & 1 \\ a & 1 & 1 & & & & & \\ f & & 1 & & & 1 & & \\ e & & & 1 & 1 & 1 & & \\ c & & & & & & 1 & 1 \end{array}$$

$$\begin{array}{c|ccccccc} & ab & af & be & de & ef & bc & cd \\ \hline b & 1 & & 1 & & & 1 & \\ d & & & & 1 & & & 1 \\ a & 1 & 1 & & & & & \\ f & & 1 & & & 1 & & \\ e & & & 1 & 1 & 1 & & \\ c & & & & & & 1 & 1 \end{array}$$

Computation: the Classic Algorithm

- Gaussian elimination

	ab	af	be	de	ef	bc	cd
b	1		1	1		1	
d				1			
a	1	1					
f		1					
e			1				
c						1	

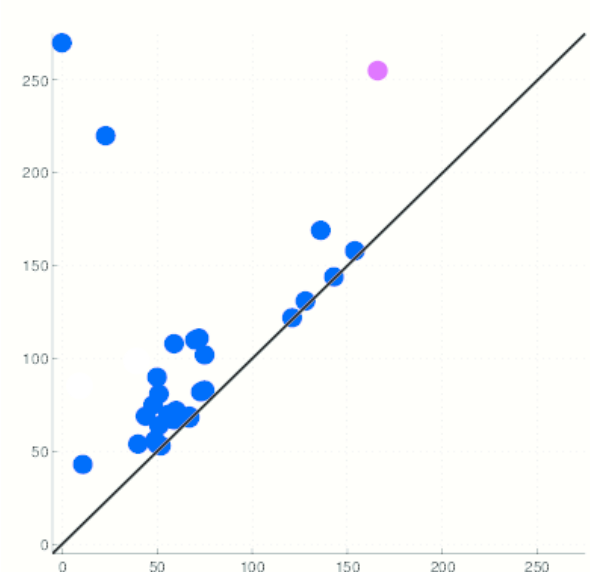
	$abef$	$bcde$
ab	1	
af	1	
be	1	1
de		1
ef	1	
bc		1
cd		1

Computation: the Classic Algorithm

- Gaussian elimination

	ab	af	be	de	ef	bc	cd
b	1		1	1		1	
d				1			
a	1	1					
f		1					
e			1				
c						1	

	$abef$	$bcde$
ab	1	
af	1	
be	1	1
de		1
ef	1	
bc		1
cd		1



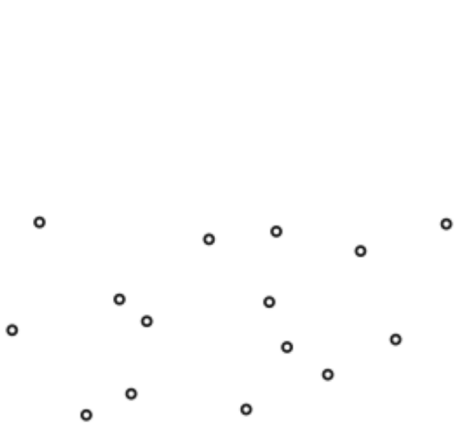
- Pivoting entries $\rightarrow$ the persistence diagram.
- $O(n^3)$, n size of the discretization

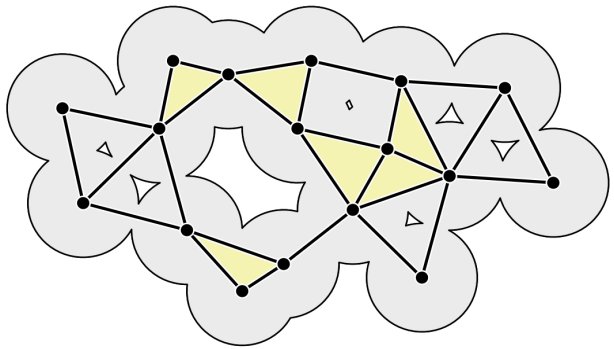
For Real World Data:

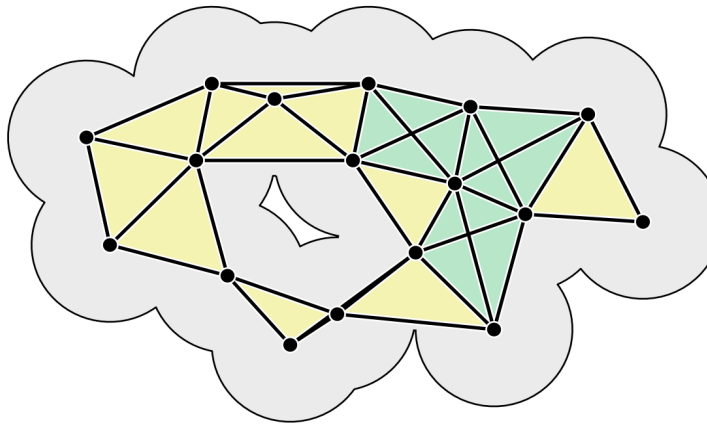
- Cubic is still expensive; Density estimation; Discretization size

Alternative for High Dimension

- Distance from the data
- Discretization: Čech Complex (Vietoris-Rips Complex)
  - Use the data as the set of vertices







Pic from Erickson

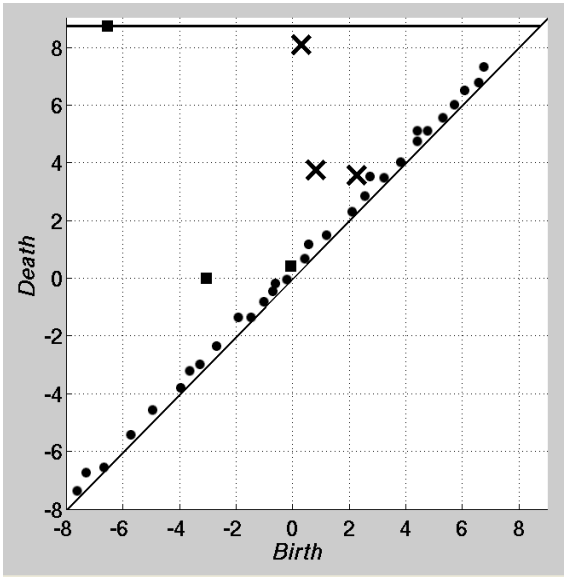
- Issues:
  - Extend to probability distribution?
  - Complex size is still exponential to the dimension of the space (ambient or intrinsic) [Sheehy SoCG 2012]

Outline

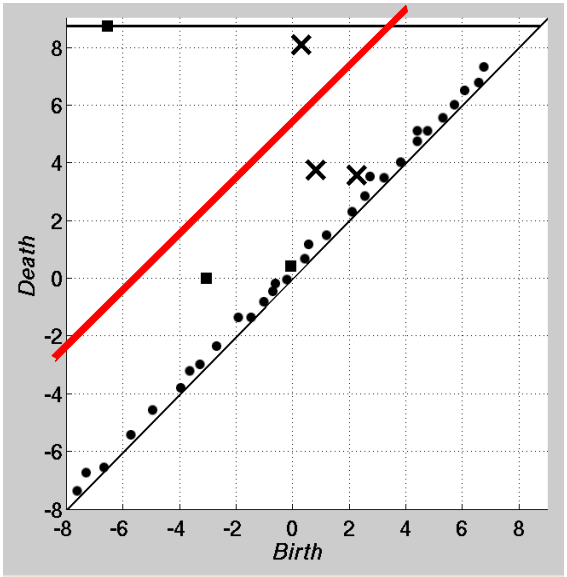
Outline

My Contributions: Output-Sensitive

- The first output-sensitive algorithm.
- Only structures with high persistence ($\geq \theta$)
- $O(P_\theta n^2)$, $P_\theta :=$ number of important structures
- The fastest.



Diagram



Output Sensitive

[**Chen** and Kerber SOCG 2011(Top 5); CGTA 2013]

My Contributions: Output-Sensitive (continued)

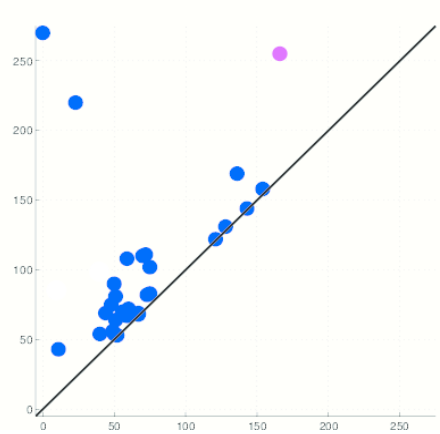
Original Matrix

	ab	af	be	de	ef	bc	cd
b	1		1			1	
d				1			1
a	1	1					
f		1			1		
e			1	1	1		
c						1	1

Reduced Matrix

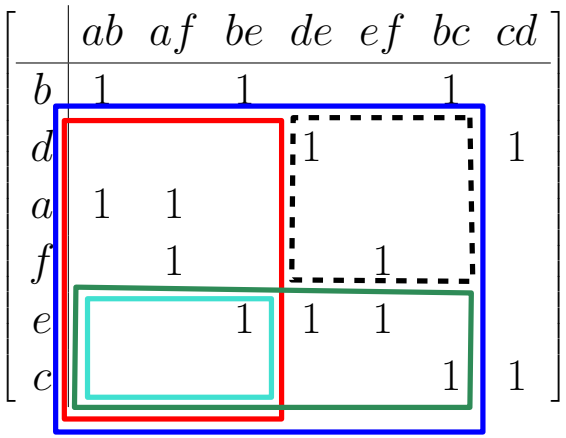
	ab	af	be	de	ef	bc	cd
b	1		1	1		1	
d				1			
a	1	1					
f		1					
e			1				
c						1	

Diagram

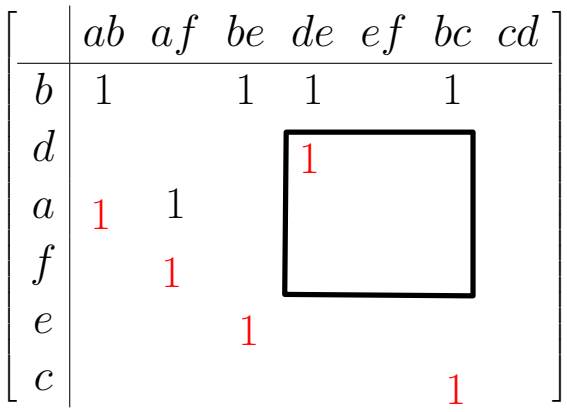


My Contributions: Output-Sensitive (continued)

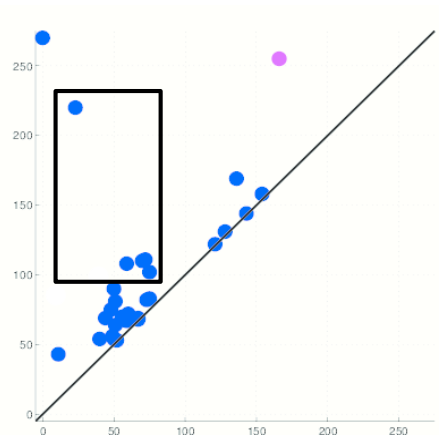
Original Matrix



Reduced Matrix



Diagram

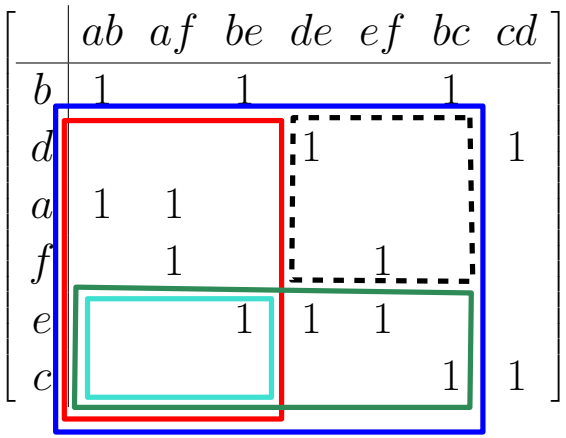


Lemma

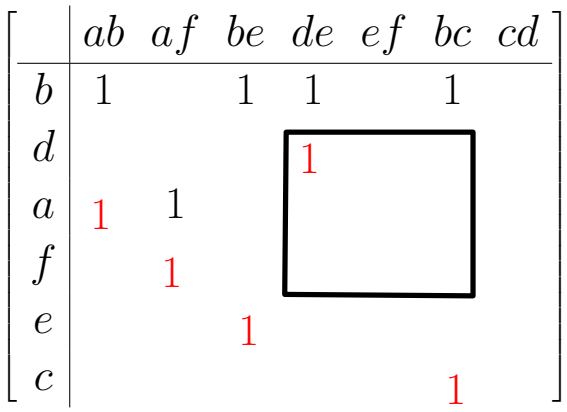
Number of dots in any box can be counted by computing ranks of submatrices of the original matrix.

My Contributions: Output-Sensitive (continued)

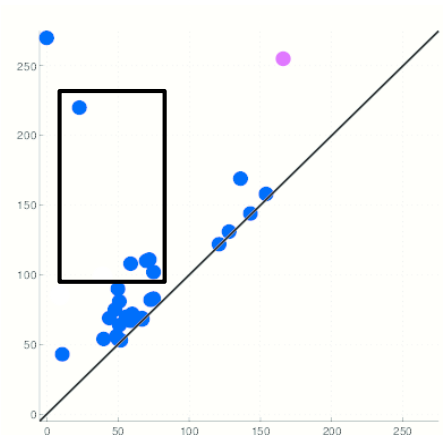
Original Matrix



Reduced Matrix

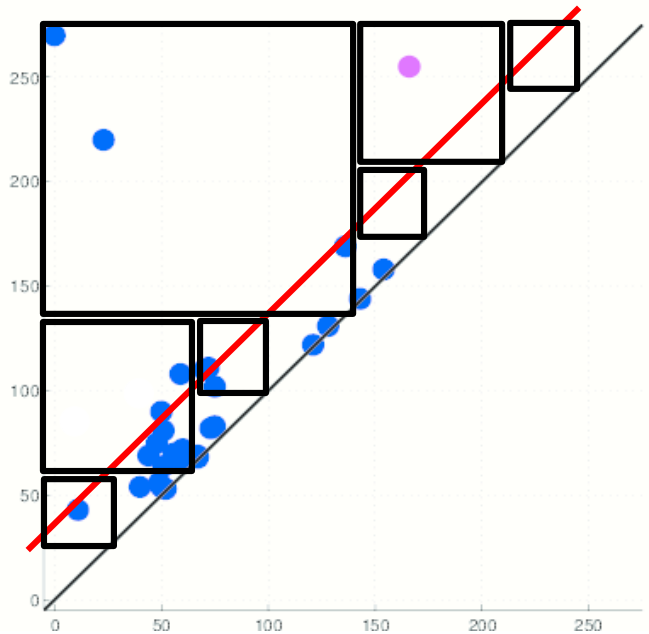


Diagram



Lemma

Number of dots in any box can be counted by computing ranks of submatrices of the original matrix.

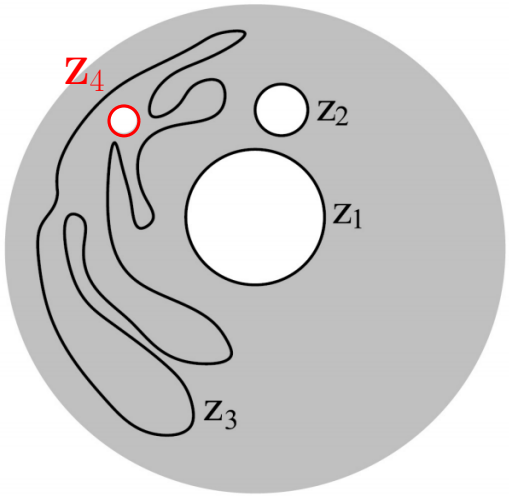


- Finite field rank (sparse matrix)
- Divide and conquer
- Master theorem

My Contributions: Topology Representations and Memory

- Representation of topological structures $\{Z_1, Z_2, Z_3\}$ vs. $\{Z_1, Z_2, Z_4\}$
  - NP-hard in the most generic form
  - Various algorithms under different assumptions

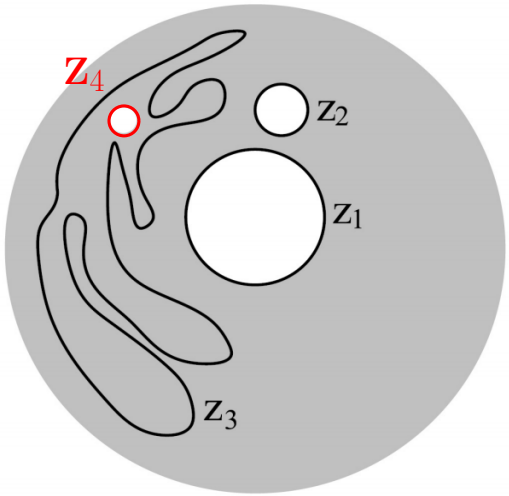
[**Chen** and Freedman SODA'10; CGTA'10; DCG'11]



My Contributions: Topology Representations and Memory

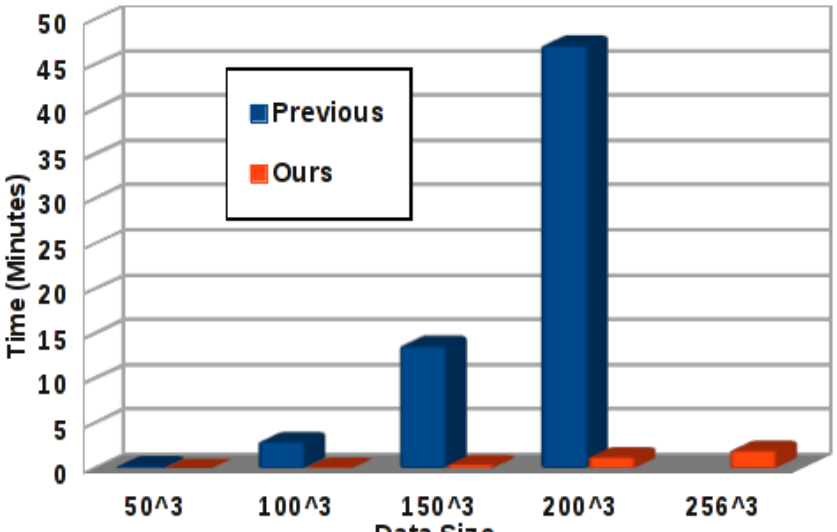
- Representation of topological structures $\{Z_1, Z_2, Z_3\}$ vs. $\{Z_1, Z_2, Z_4\}$
  - NP-hard in the most generic form
  - Various algorithms under different assumptions

[**Chen** and Freedman SODA'10; CGTA'10; DCG'11]

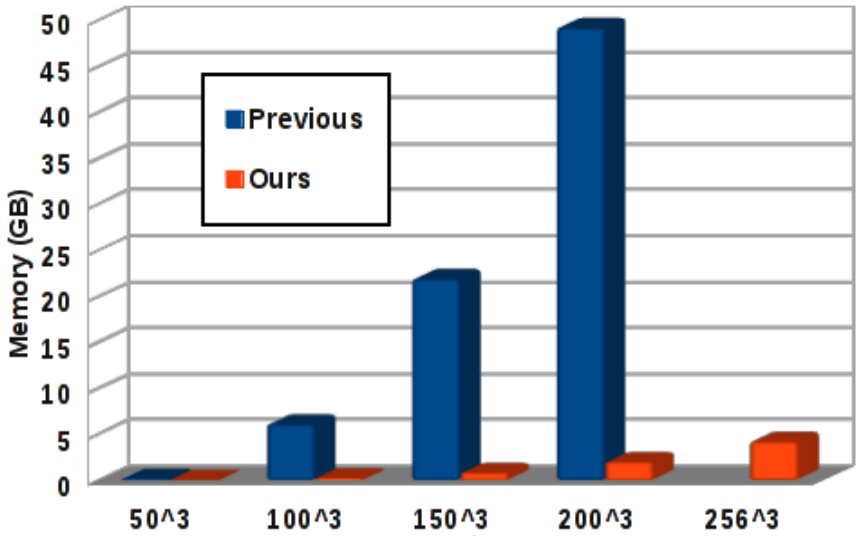


- Better speed and memory efficiency [**Chen** and Kerber, EuroCG 2011] [Wagner, **Chen** and Vucini, TopoInVis 2011 (best paper runner-up)]

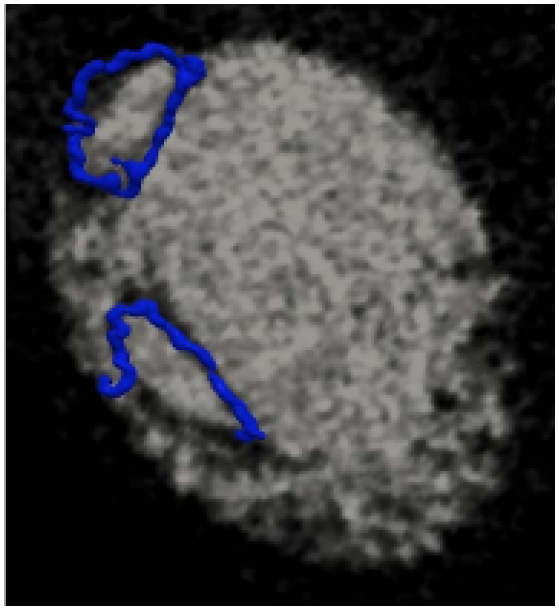
Time

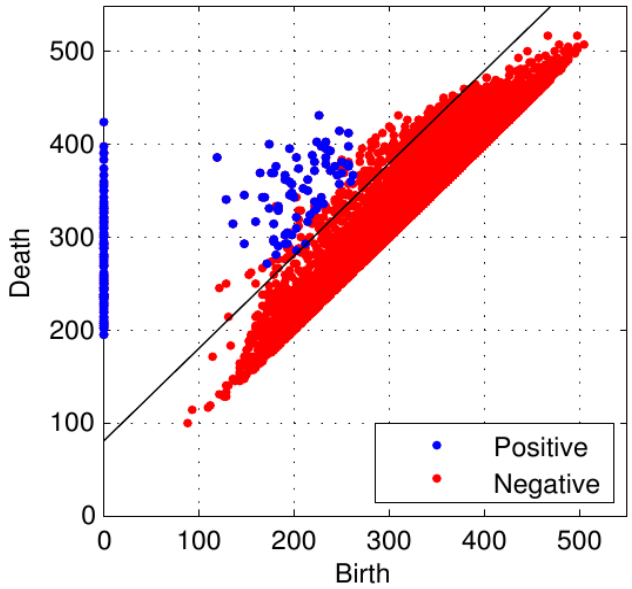


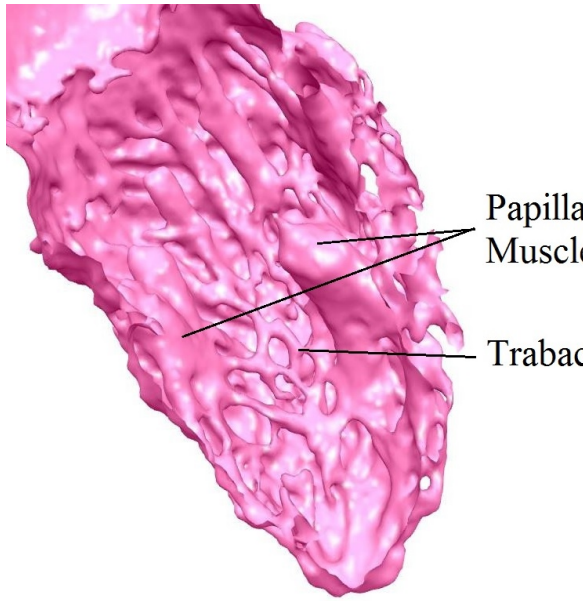
Memory



Cardiac Data Revisited

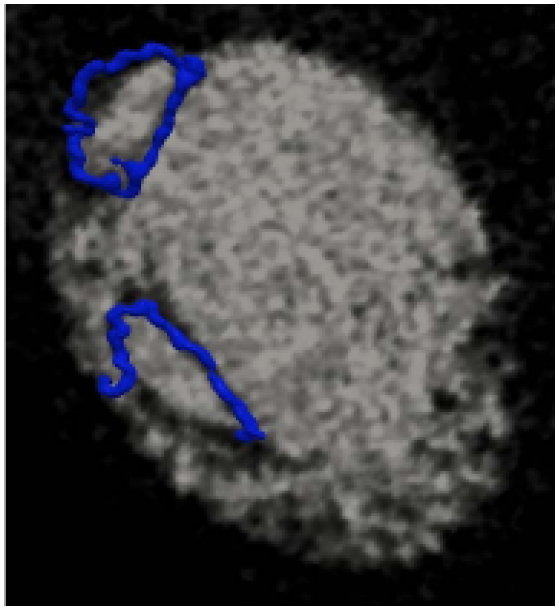


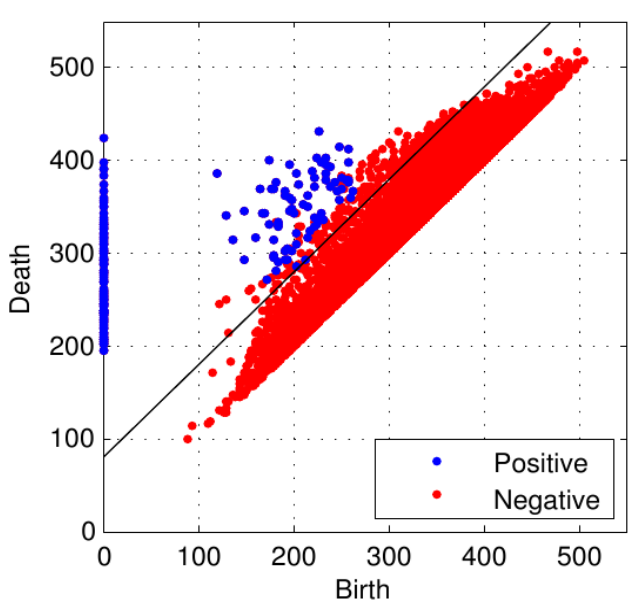


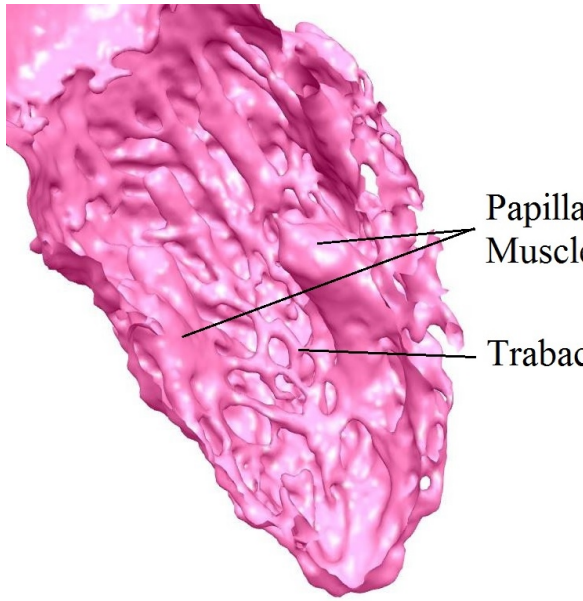


[Gao, **Chen** , *et al.* IPMI 2013]

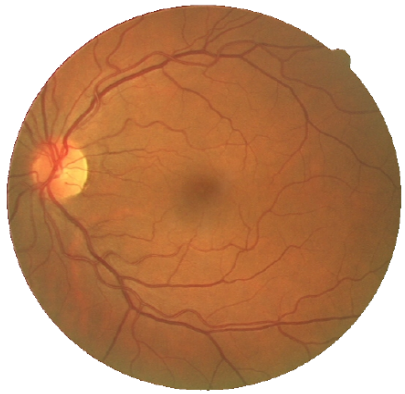
Cardiac Data Revisited

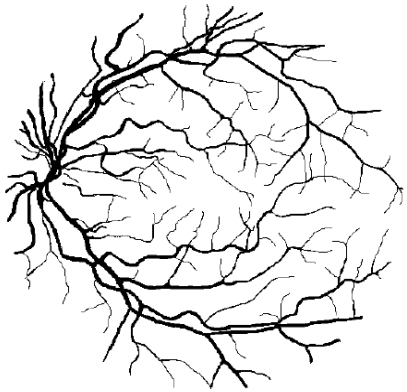






[Gao, **Chen** , *et al.* IPMI 2013]

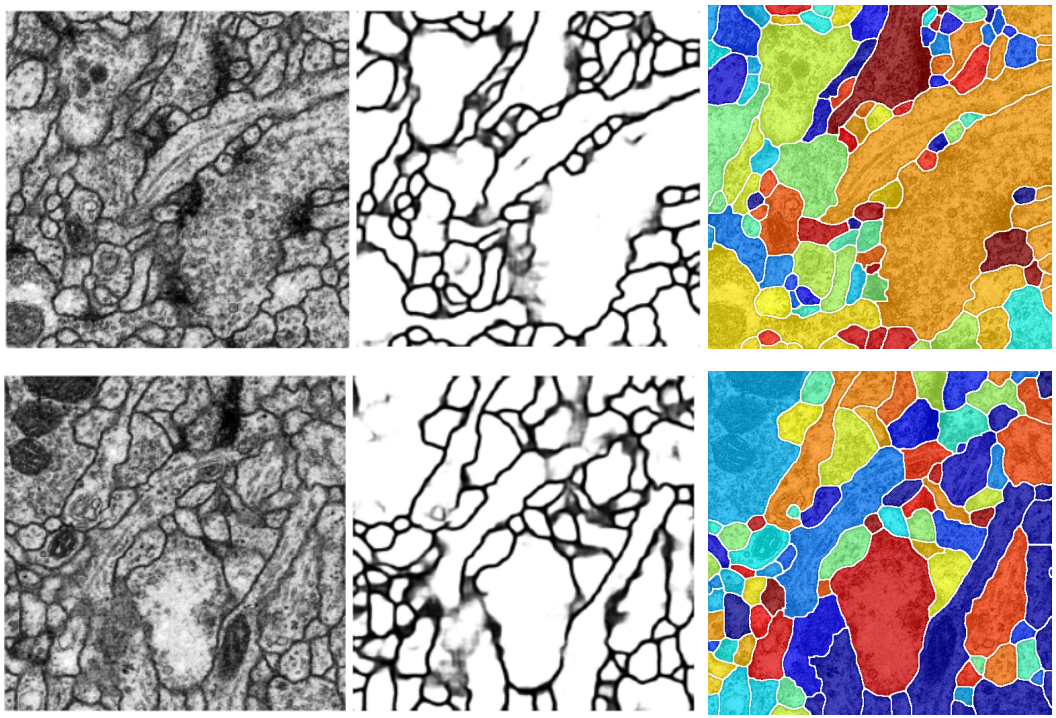




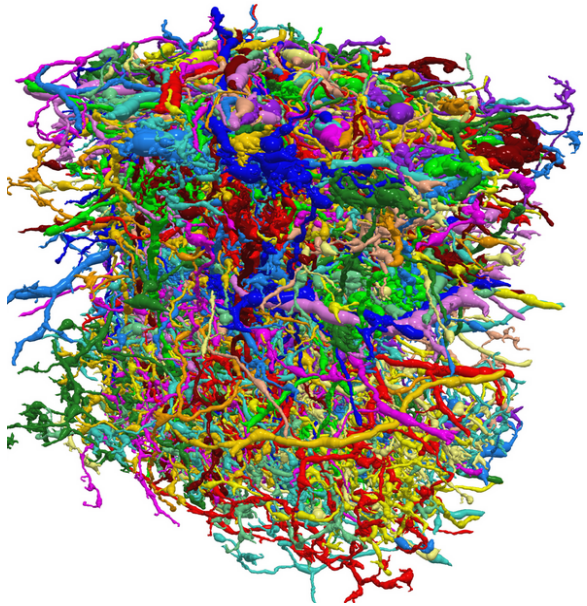
[**Chen** , *et al.* CVPR 2011]

More Examples

- Electron Microscopy Images of Fly/Mouse Brains
  - Input: 2D or 3D EM images; boundary likelihood map
  - Output: partitioning of the image
  - Topology: help making accurate decision at uncertain/blurred region



EM Images Likelihood Results



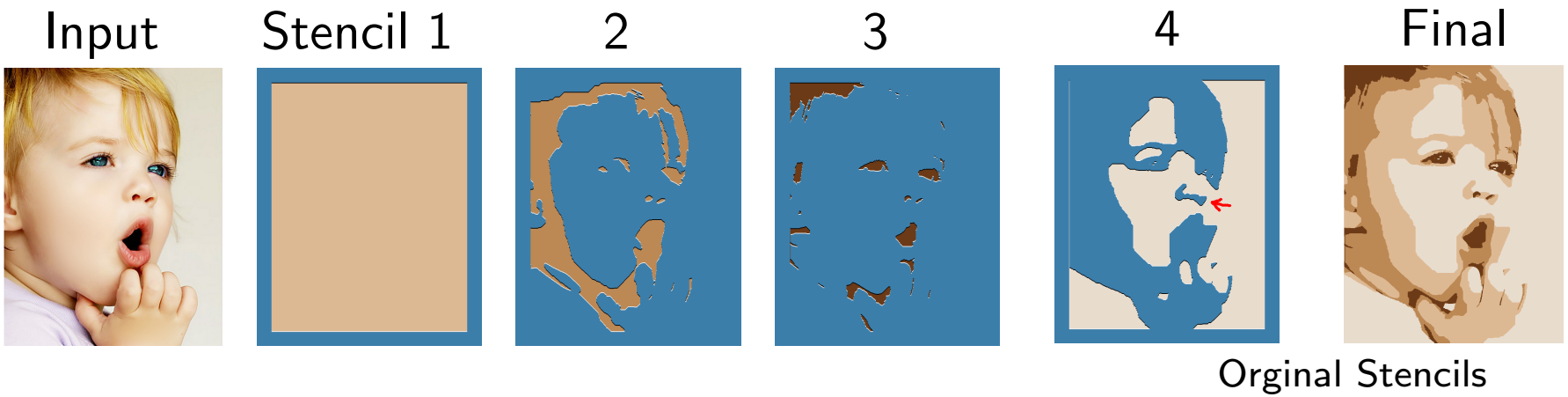
[Uzunbaş, **Chen** and Metaxas, MICCAI 2014]

Online Challenge, International Symposium on Biomedical Imaging (ISBI)

Leading Groups

Group name	Rand Error
optree-idsia	0.022777620
** human values **	0.026546995
SCI	0.031457892
IDSIA-SCI	0.040982878
IDSIA	0.050399038
MLL-ETH	0.063919883

Application: Multi-Layer Stencil Creation

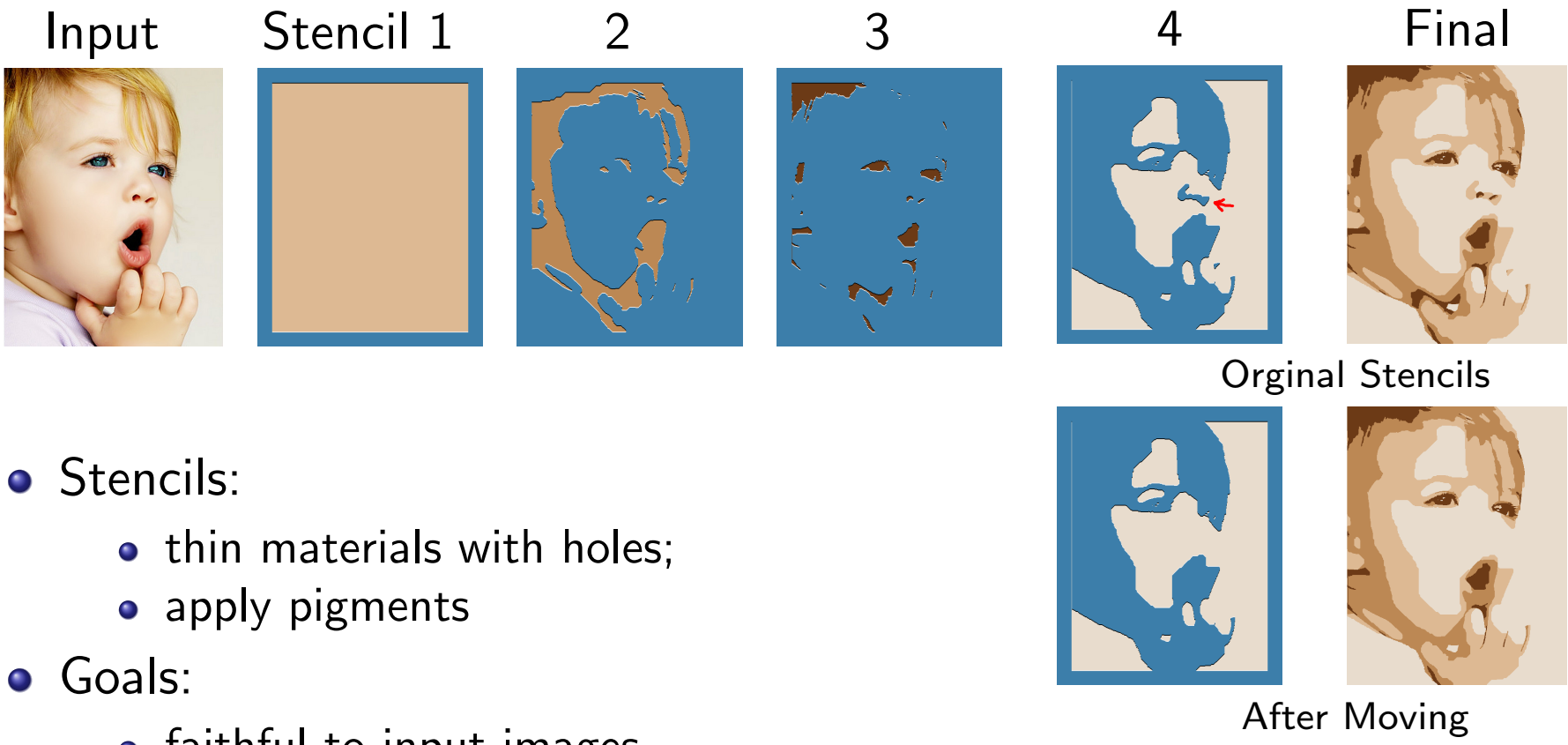


Orginal Stencils

- Stencils:
  - thin materials with holes;
  - apply pigments
- Goals:
  - faithful to input images,
  - simple boundary,
  - connected

[Jain, **Chen** , *et al.* , Computer & Graphics, 2015]

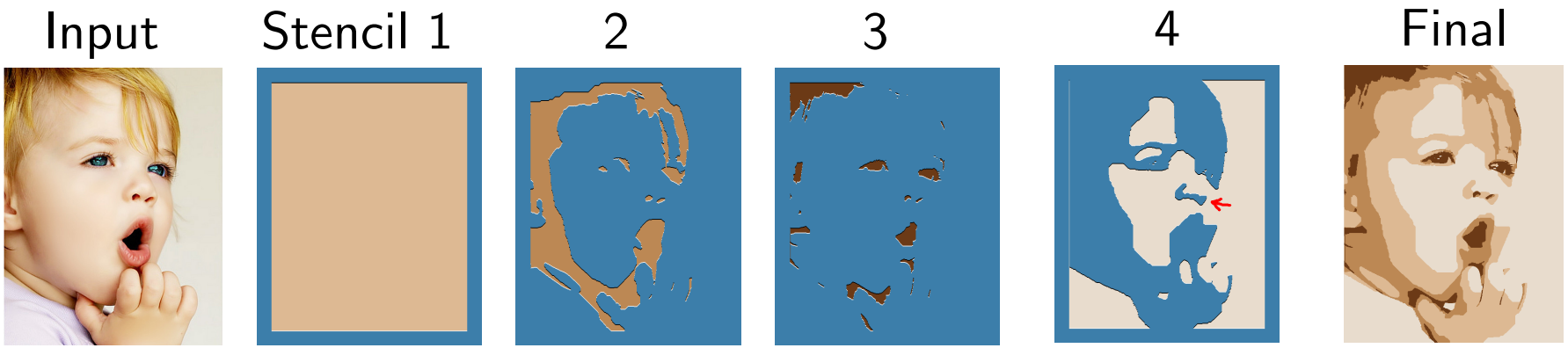
Application: Multi-Layer Stencil Creation



- Stencils:
  - thin materials with holes;
  - apply pigments
- Goals:
  - faithful to input images,
  - simple boundary,
  - connected

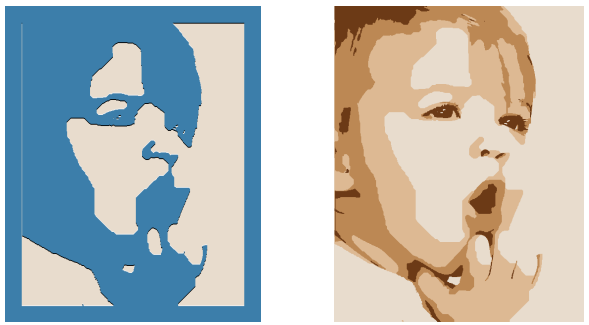
[Jain, **Chen** , *et al.* , Computer & Graphics, 2015]

Application: Multi-Layer Stencil Creation



Orginal Stencils

After Moving



With Connectivity Constraints

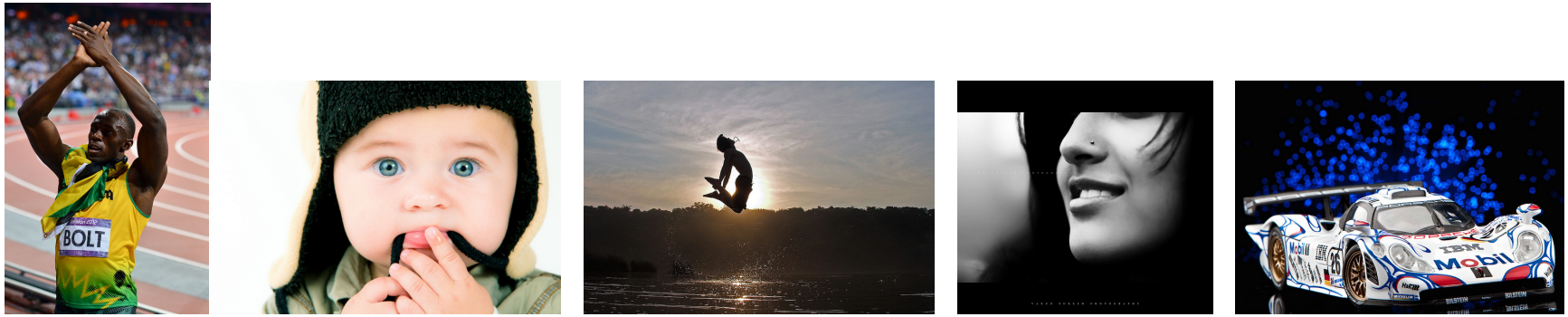
- Stencils:
  - thin materials with holes;
  - apply pigments
- Goals:
  - faithful to input images,
  - simple boundary,
  - connected

[Jain, **Chen** , *et al.* , Computer & Graphics, 2015]

Multi-Layer Stencil Creation

With vs. without connectivity constraints

Input:



Without:

With:

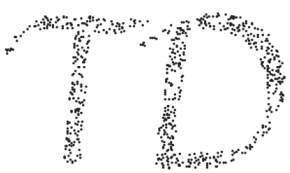
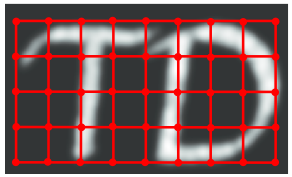
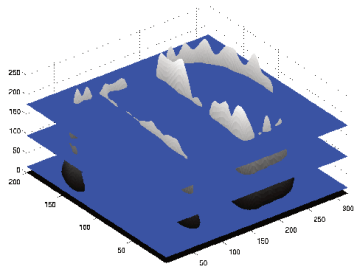
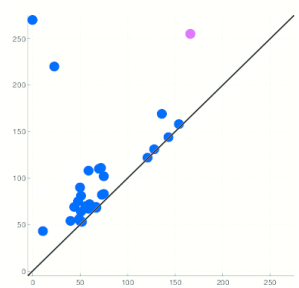
Application: Multi-Layer Stencil Creation

Canvas/wall result:

- Website, interactive

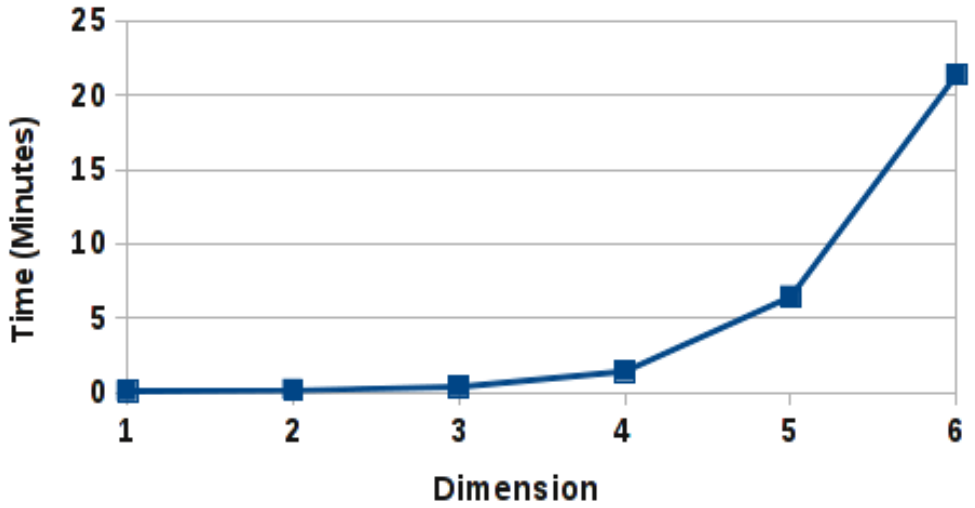
Outline

Curse of dimensionality

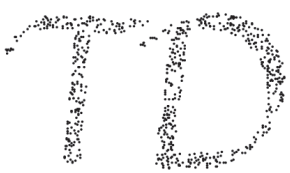
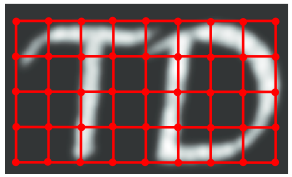
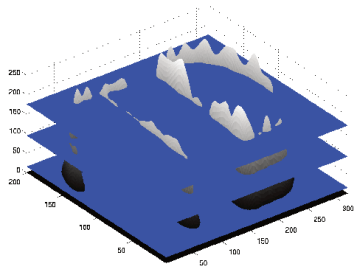
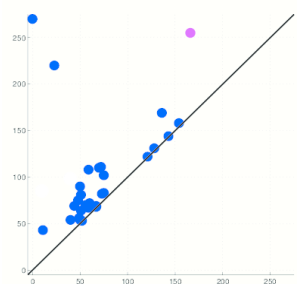





Previous methods:

- kernel density estimation; discretize the domain



Curse of dimensionality

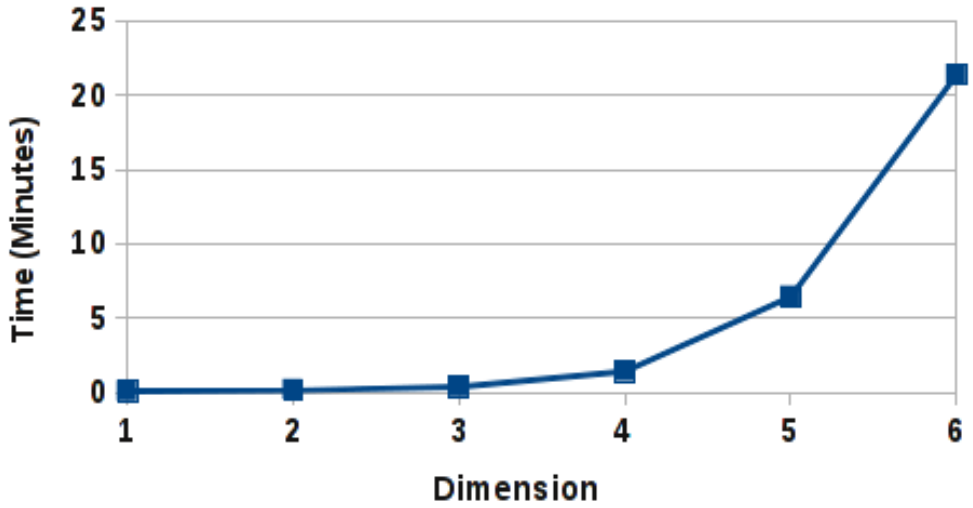
Previous methods:

- kernel density estimation; discretize the domain

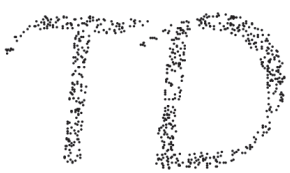
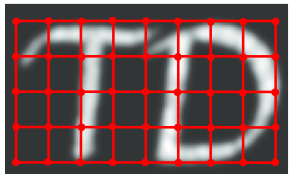
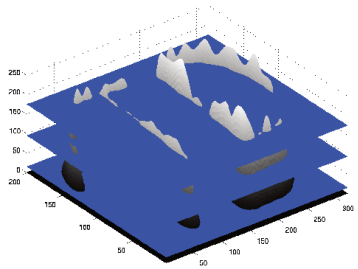
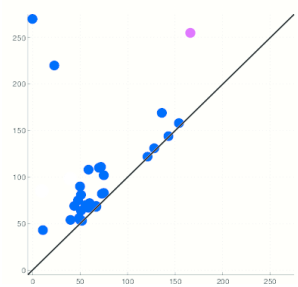
Curse of dimensionality:

model/algorithm fails as the dimension D increases

- Issue 1: necessary sample size grows exponential to D



Curse of dimensionality

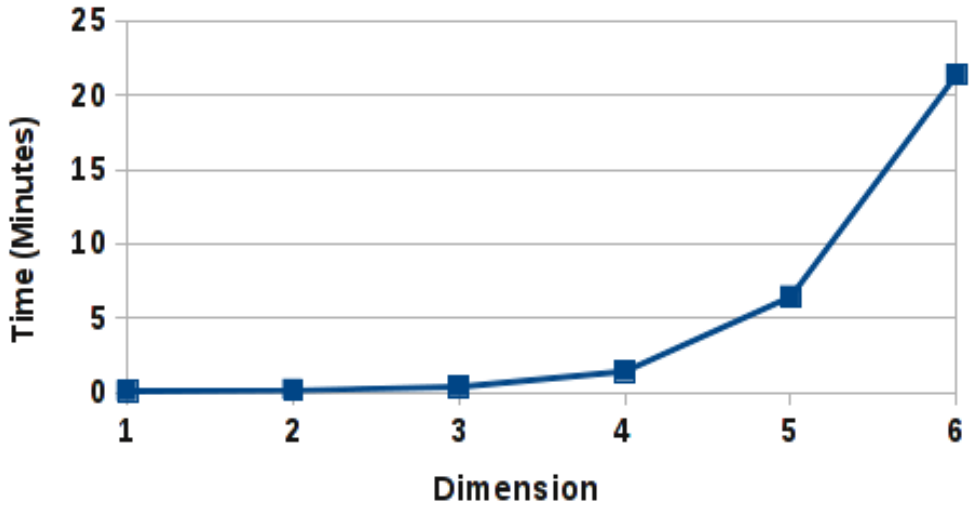
Previous methods:

- kernel density estimation; discretize the domain

Curse of dimensionality:

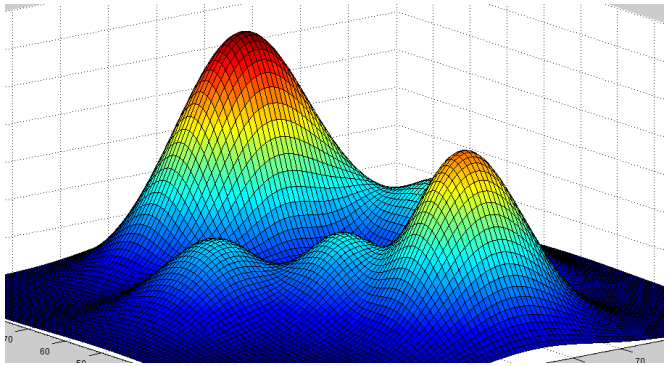
model/algorithm fails as the dimension D increases

- Issue 1: necessary sample size grows exponential to D
- Issue 2: discretization has exponential size $n = \exp(D)$, $O(n^3)$ or $O(n^2)$

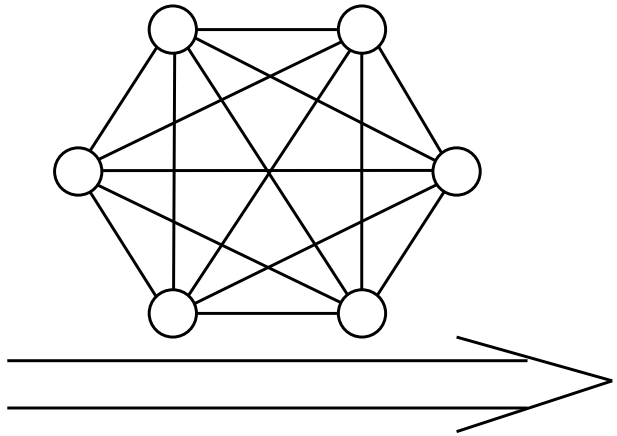


Stepping Toward Big Data

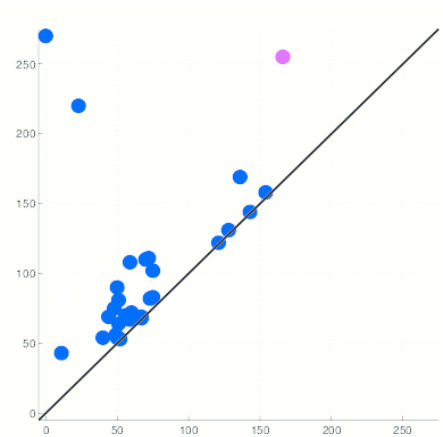
- Solution: combine with state-of-the-art statistical tools
  - Probabilistic Graphical Models
  - Develop algorithms on a size D graph
  - $O(\exp(D)) \to O(D^2)$
  - Only need $\log D$ samples (vs. $\exp D$)



D-dimensional domain



via a size D graph

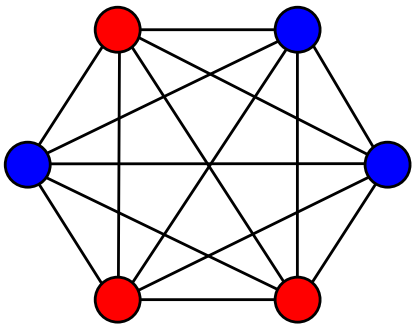


persistence diagram

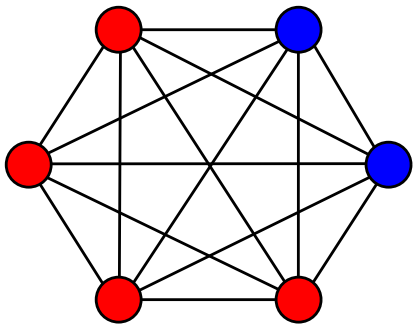
A One-Page Introduction to Graphical Models

Represent a D-dimensional distribution using a weighted graph

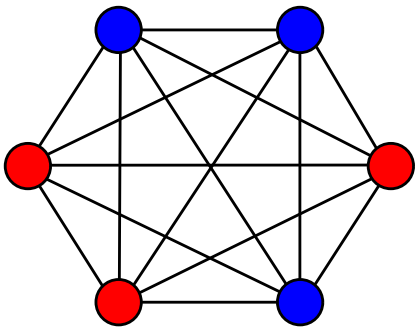
- D variables $\leftrightarrow$ D nodes of the graph, $\{1, \cdots, L\}^D = \{red, blue\}^6$
- Edge weights describe the (in)dependence between different variables



$\mathbf{Prob = 0.05}$



$\mathbf{Prob = 0.02}$

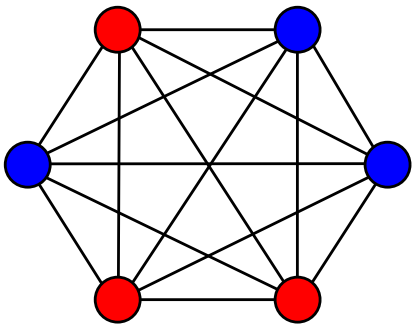


$\mathbf{Prob = 0.00}$

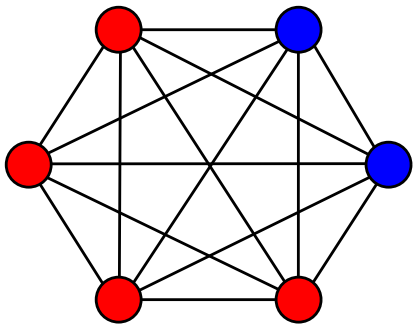
A One-Page Introduction to Graphical Models

Represent a D-dimensional distribution using a weighted graph

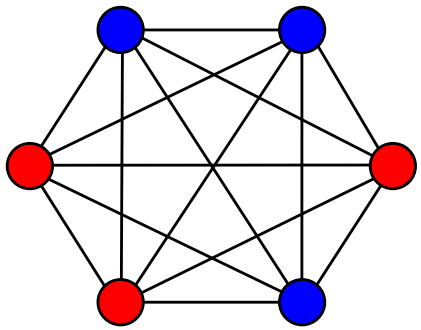
- D variables $\leftrightarrow$ D nodes of the graph, $\{1, \cdots, L\}^D = \{red, blue\}^6$
- Edge weights describe the (in)dependence between different variables



$\mathbf{Prob = 0.05}$



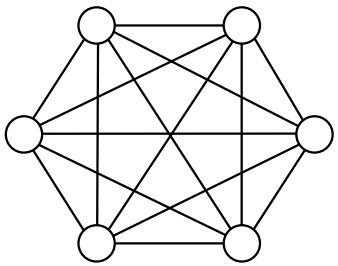
$\mathbf{Prob = 0.02}$



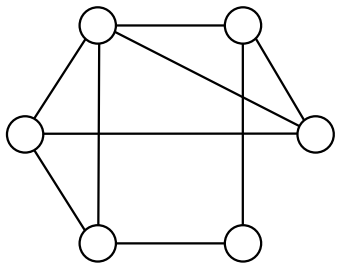
$\mathbf{Prob = 0.00}$

Estimation

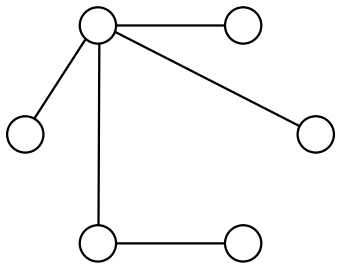
- Graph structure + weights
- Simplified graph: less flexibility; easier to develop algorithms



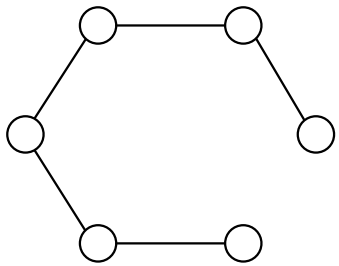
Complete



Sparse

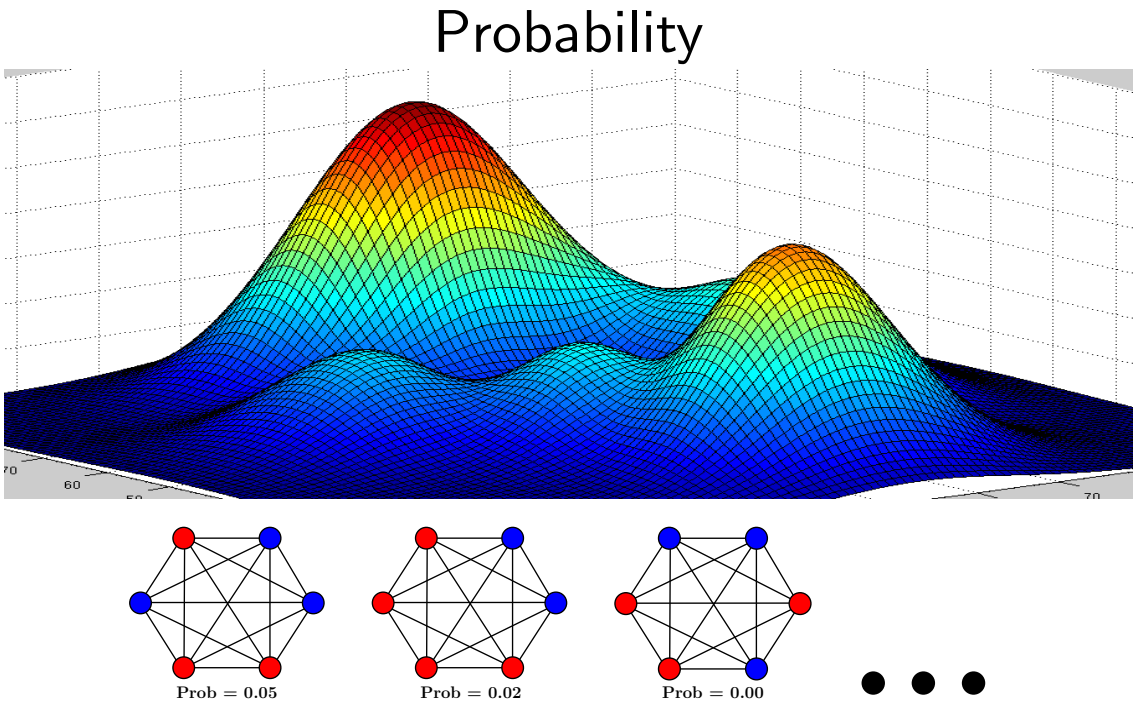


Tree

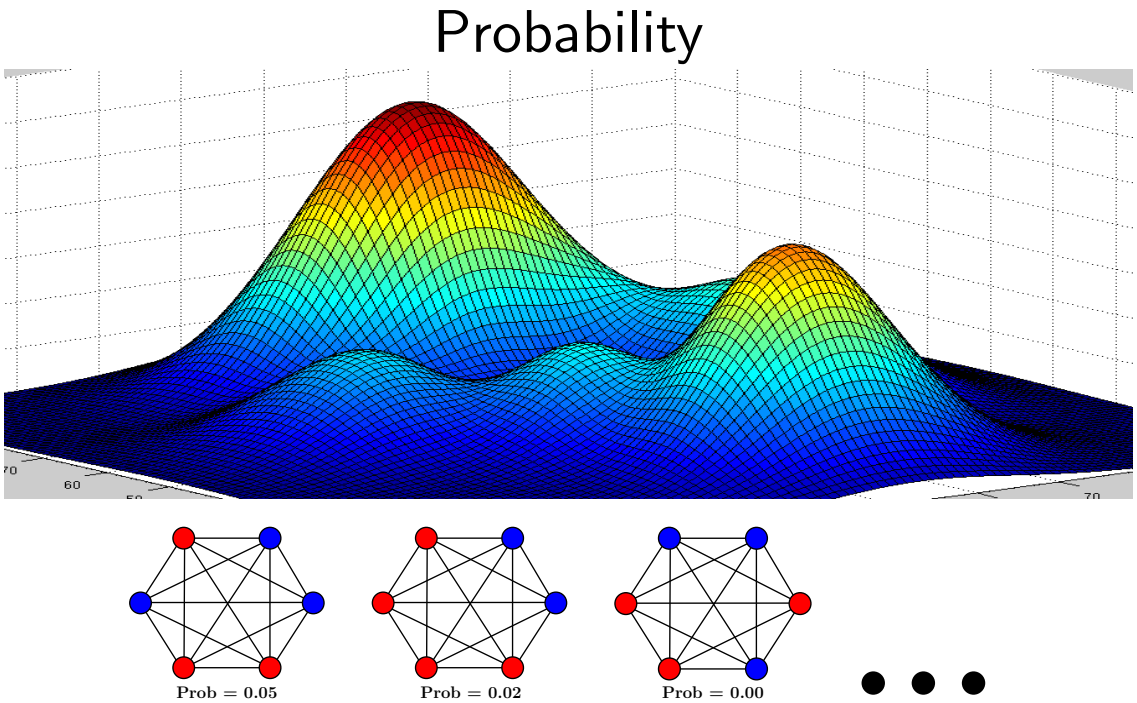


Chain

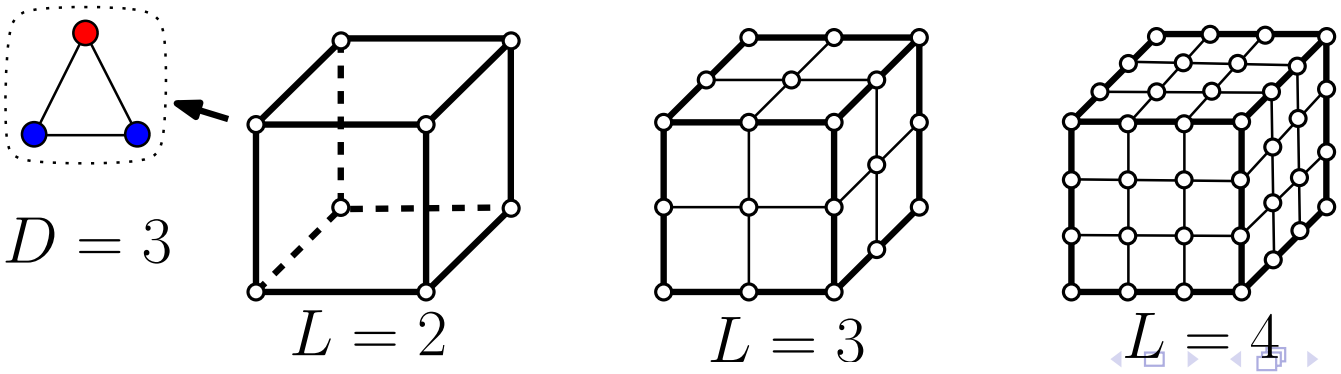
Computing Topological Structures



Computing Topological Structures

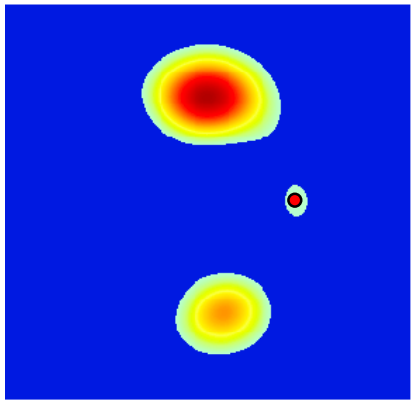


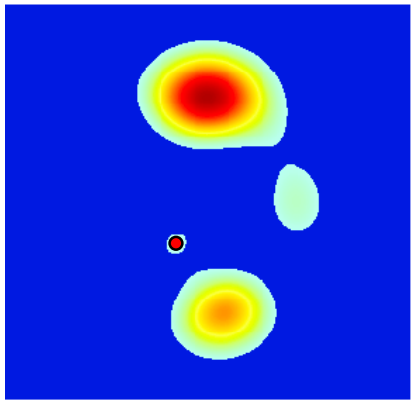
- Discretized Domain (lattice with L^D elements):



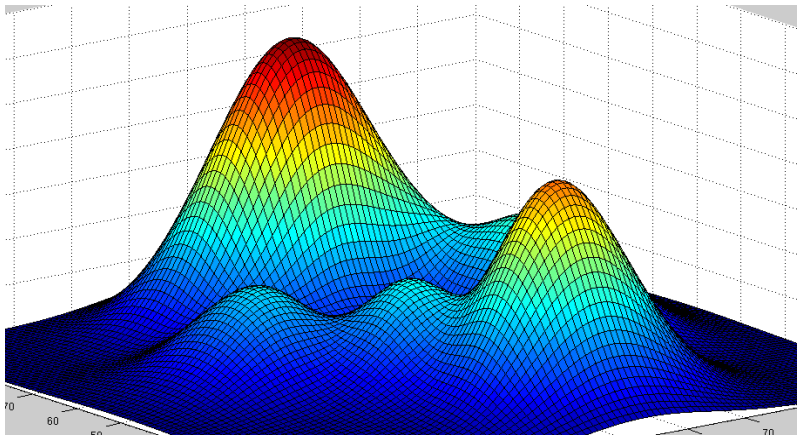
Critical points

Maxima (modes)



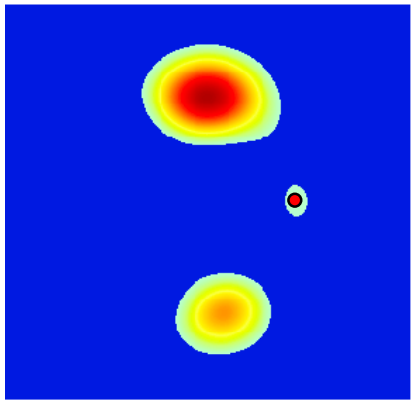


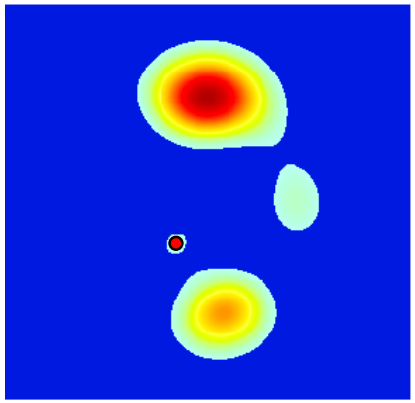
Saddle Points



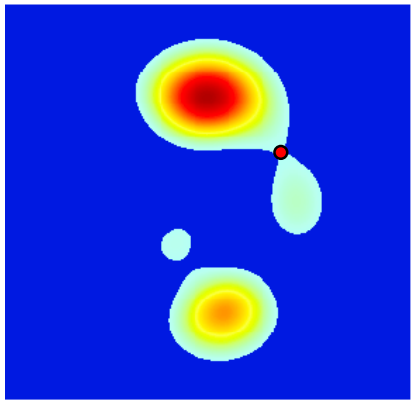
Critical points

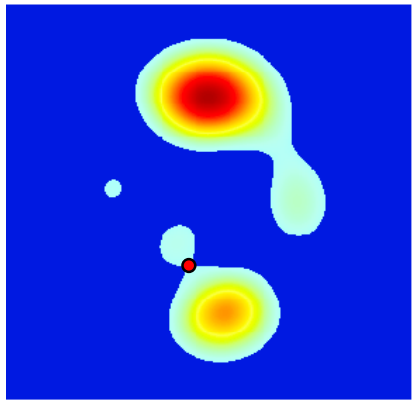
Maxima (modes)

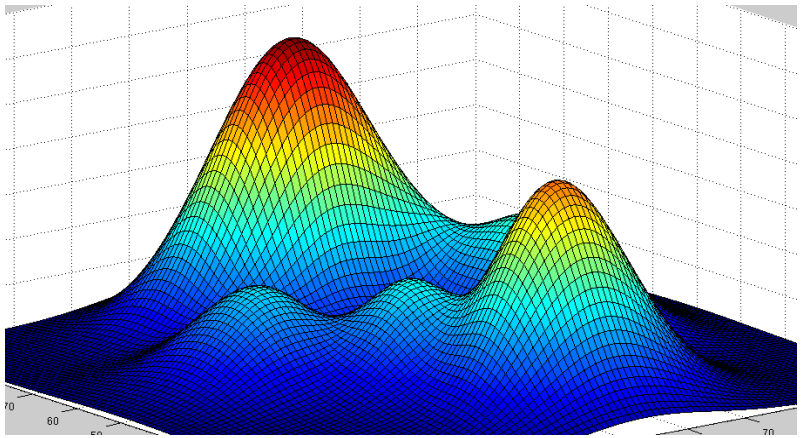


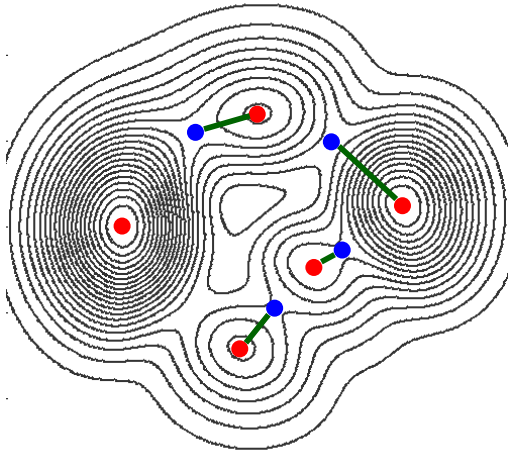


Saddle Points









- Key: locate and pair critical points,
  function values $\rightarrow$ coordinates of dots in the diagram

Algorithm

[**Chen** *et al.* AISTATS'13 (Oral); NIPS 2014 (Spotlight); Submitted to JMLR]

- Input: data samples
- Step 1: estimate a graphical model [Liu et al. JMLR'11]
- Step 2: compute modes
- Step 3: compute saddles
- Step 4: infer topological structures and their persistence

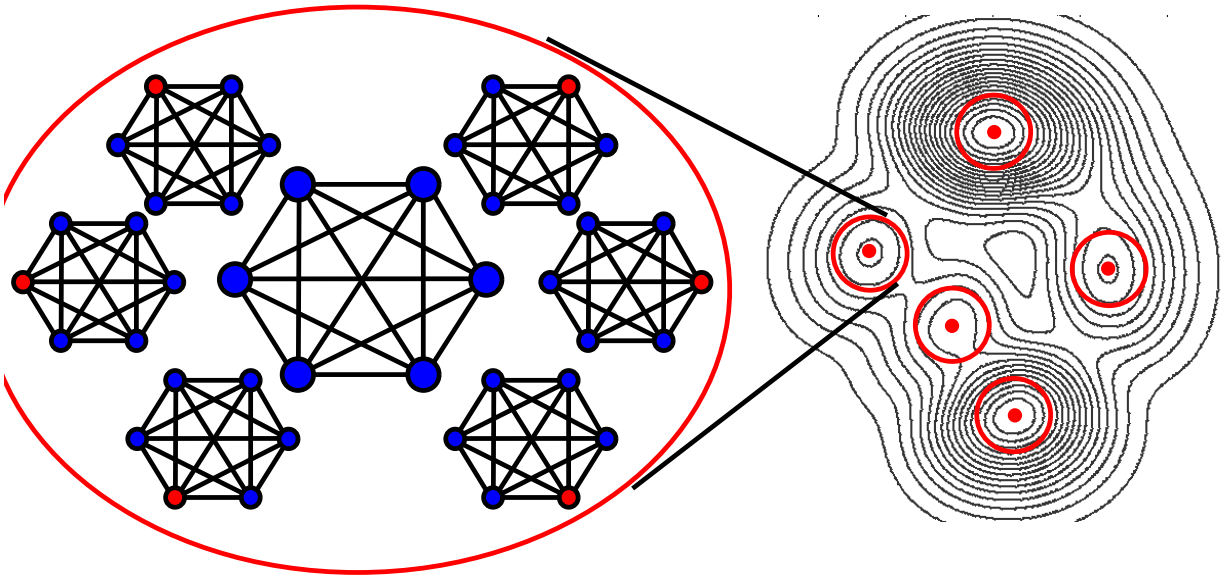
Algorithm

[**Chen** *et al.* AISTATS'13 (Oral); NIPS 2014 (Spotlight); Submitted to JMLR]

- Input: data samples
- Step 1: estimate a graphical model [Liu et al. JMLR'11] (skip)
- Step 2: compute modes
- Step 3: compute saddles (skip)
- Step 4: infer topological structures and their persistence (skip)

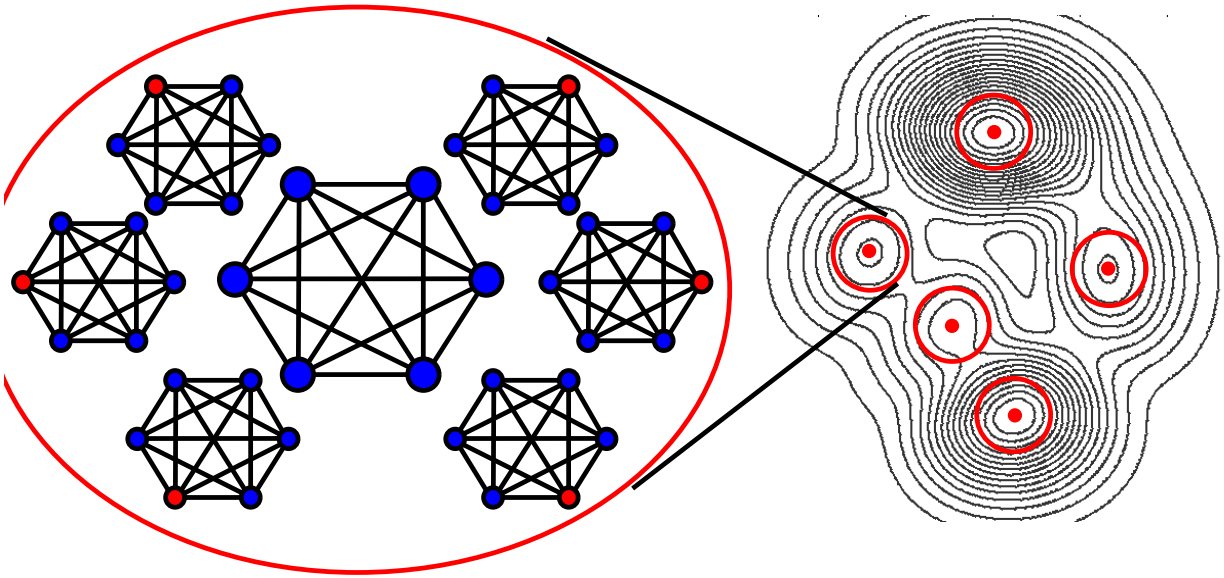
Modes in Discrete Domain

- Domain: discrete ($\{1, \cdots, L\}^D$)
- Neighborhood: $\mathcal{N}_\delta(x)$, Hamming ball centered at x radius δ
- Mode: bigger probability than all its neighbors



Modes in Discrete Domain

- Domain: discrete ($\{1, \cdots, L\}^D$)
- Neighborhood: $\mathcal{N}_\delta(x)$, Hamming ball centered at x radius δ
- Mode: bigger probability than all its neighbors



Compute top M modes for a given δ

- Previously: compute the global maximum (top 1 mode)
  Trees: polynomial, general graphs: NP-hard
- Our setting: use tree graphs

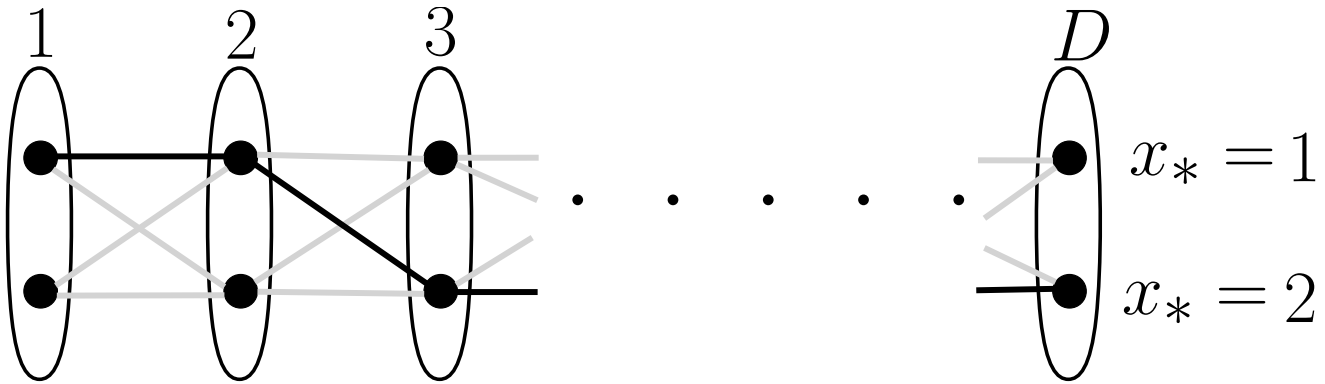
Compute top M modes for chains or trees

$$\mathcal{V} = \{1, \ldots, D\}, \quad \mathcal{E} = \{(i, i+1)\}$$

Compute top M modes for chains or trees

$$\mathcal{V} = \{1, \ldots, D\}, \quad \mathcal{E} = \{(i, i+1)\}$$

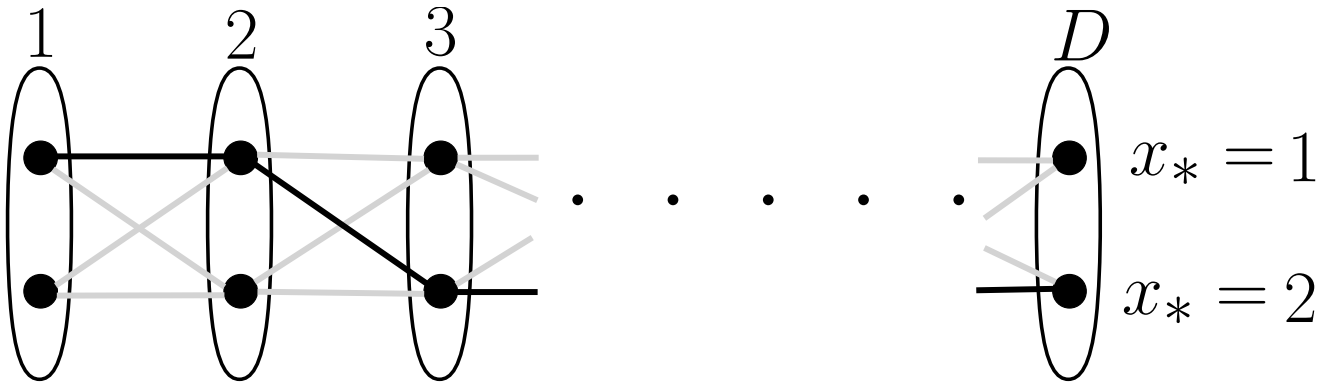
- labeling $x = (x_1, \ldots, x_D), x_i \in \mathcal{L}$



Compute top M modes for chains or trees

$$\mathcal{V} = \{1, \ldots, D\}, \quad \mathcal{E} = \{(i, i+1)\}$$

- labeling $x = (x_1, \ldots, x_D), x_i \in \mathcal{L}$



$$E(x) = \sum_{(i,i+1)\in\mathcal{E}} \phi_{i,i+1}(x_i, x_{i+1}) \qquad P(x) = \frac{1}{Z} \exp\left(-E(x)\right)$$

Algorithm: Chains

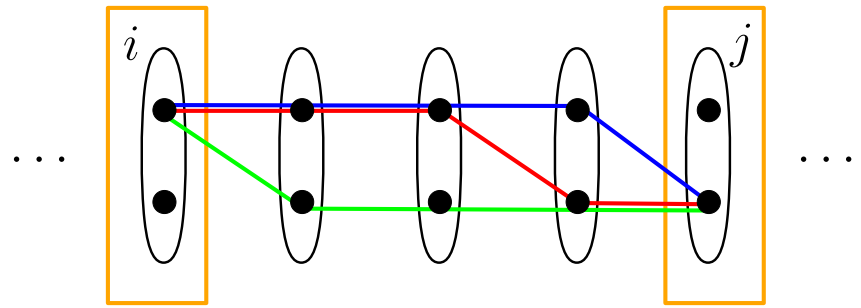
Idea: divide and conquer

- The whole chain $[1, D] \rightarrow$ subchains $[i, j]$

Algorithm: Chains

Idea: divide and conquer

- The whole chain $[1, D] \rightarrow$ subchains $[i, j]$
- A partial labeling $x_{i:j}$

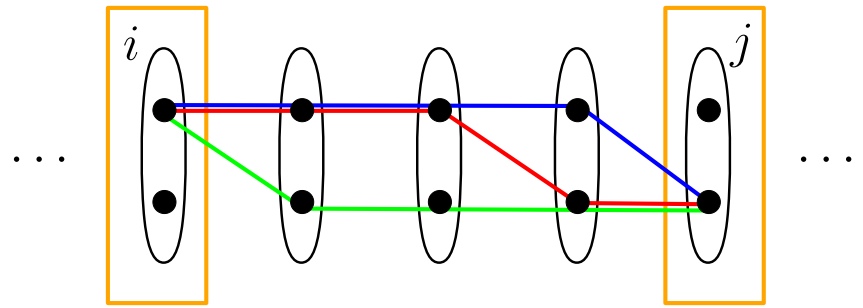


- $x_{i:j}$ is a local mode iff for any $y_{i:j}$ s.t. $y_i = x_i$, $y_j = x_j$ $E(x_{i:j}) < E(y_{i:j})$

Algorithm: Chains

Idea: divide and conquer

- The whole chain $[1, D] \rightarrow$ subchains $[i, j]$
- A partial labeling $x_{i:j}$



- $x_{i:j}$ is a local mode iff for any $y_{i:j}$ s.t. $y_i = x_i$, $y_j = x_j$ $E(x_{i:j}) < E(y_{i:j})$

Lemma

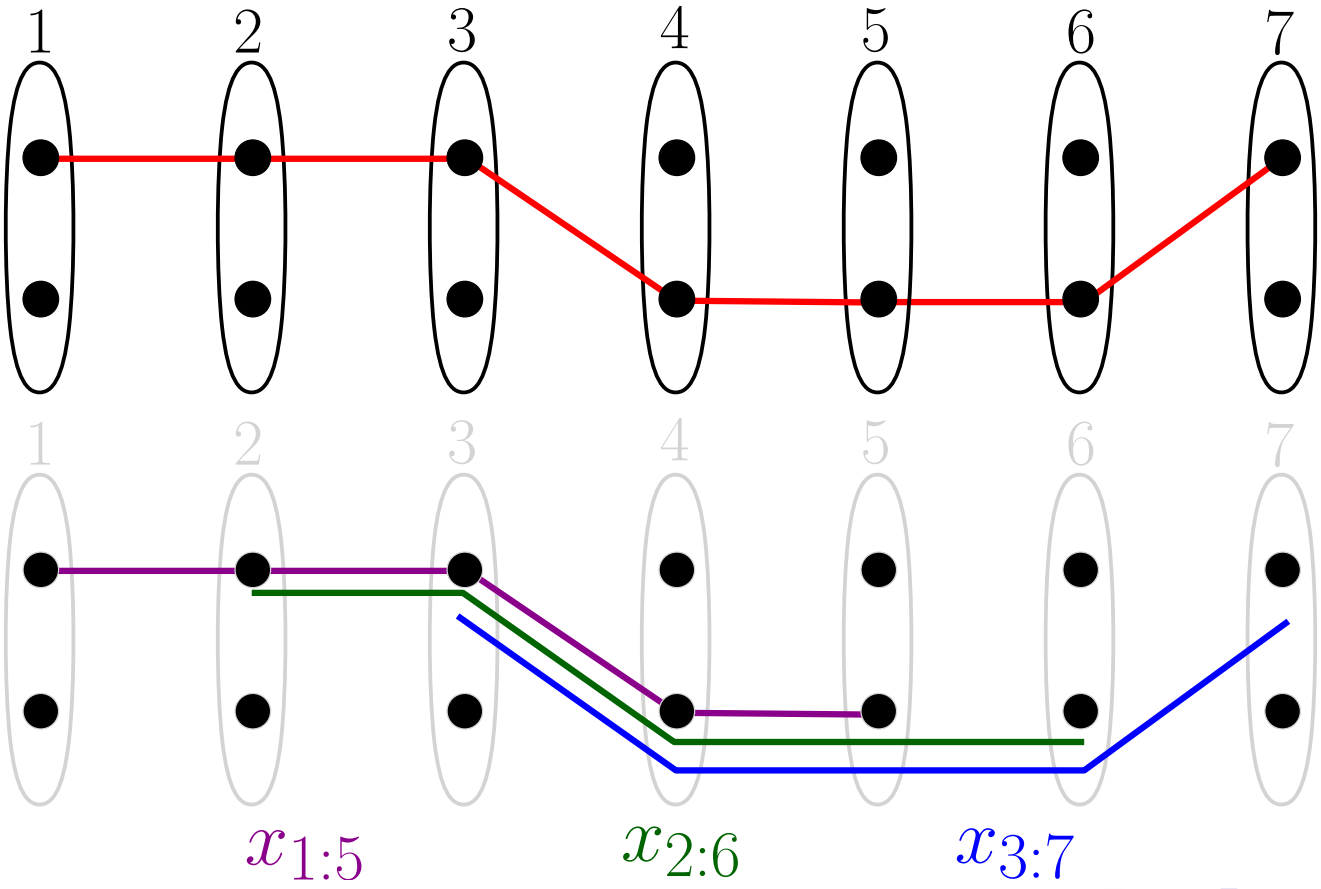
any $[i, j]$ has L^2 local modes, computable in polynomial time

Algorithm: Chains

Theorem (local-global)

$\forall \delta$, x is a mode iff it is a local mode in every length $(\delta + 2)$ subchain.

An example: $D = 7$, $\delta = 3$

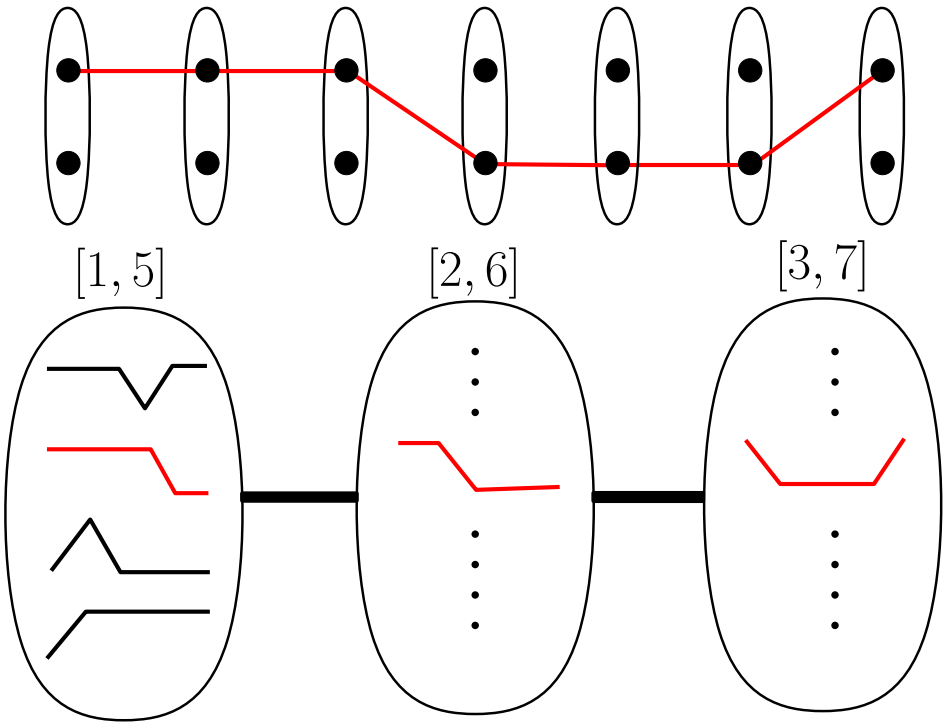


Algorithm: Chains

- combinations of local modes = global modes
  efficiently search through this space?

Algorithm: Chains

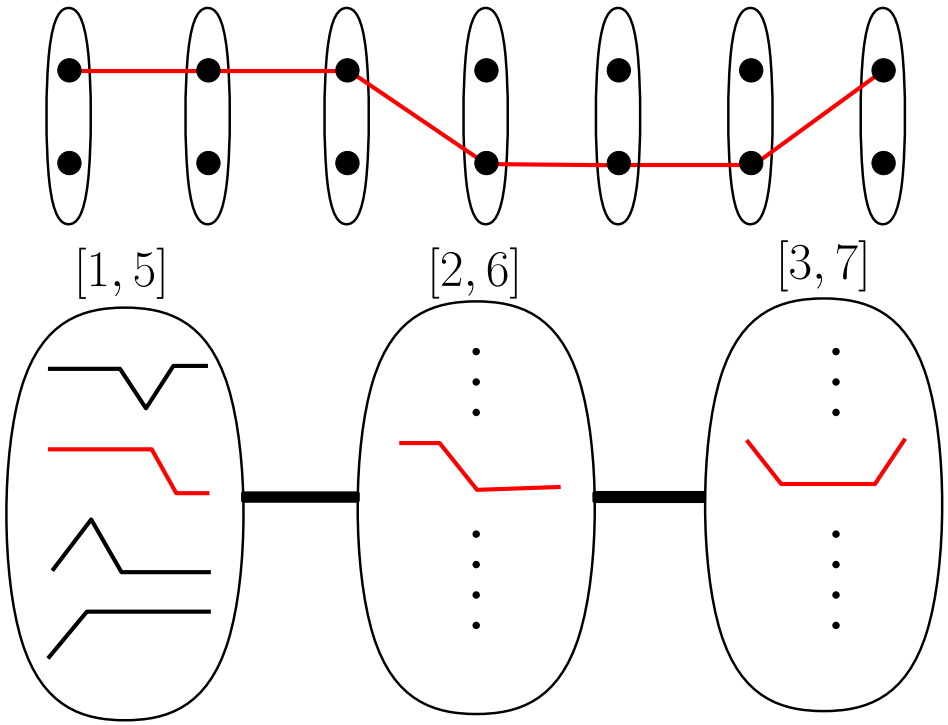
- combinations of local modes = global modes
  efficiently search through this space?
- Construct a new (junction) chain



- supernodes $[i, j]$
- labels = {local modes of $[i, j]$}
- preserve the energy

Algorithm: Chains

- combinations of local modes = global modes
  efficiently search through this space?
- Construct a new (junction) chain



- supernodes $[i, j]$
- labels = {local modes of $[i, j]$}
- preserve the energy

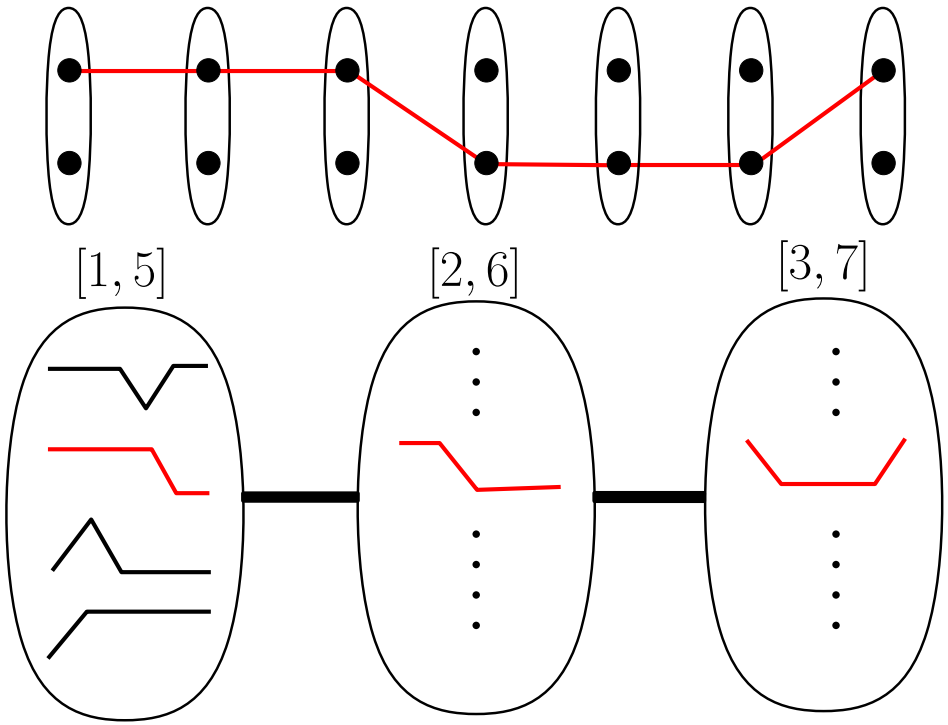
Fact

New chain labeling space
= space of all modes

- compute the top M labeling of the new chain [Nilsson'98]

Algorithm: Chains

- combinations of local modes = global modes
  efficiently search through this space?
- Construct a new (junction) chain



- supernodes $[i, j]$
- labels = {local modes of $[i, j]$}
- preserve the energy

Fact

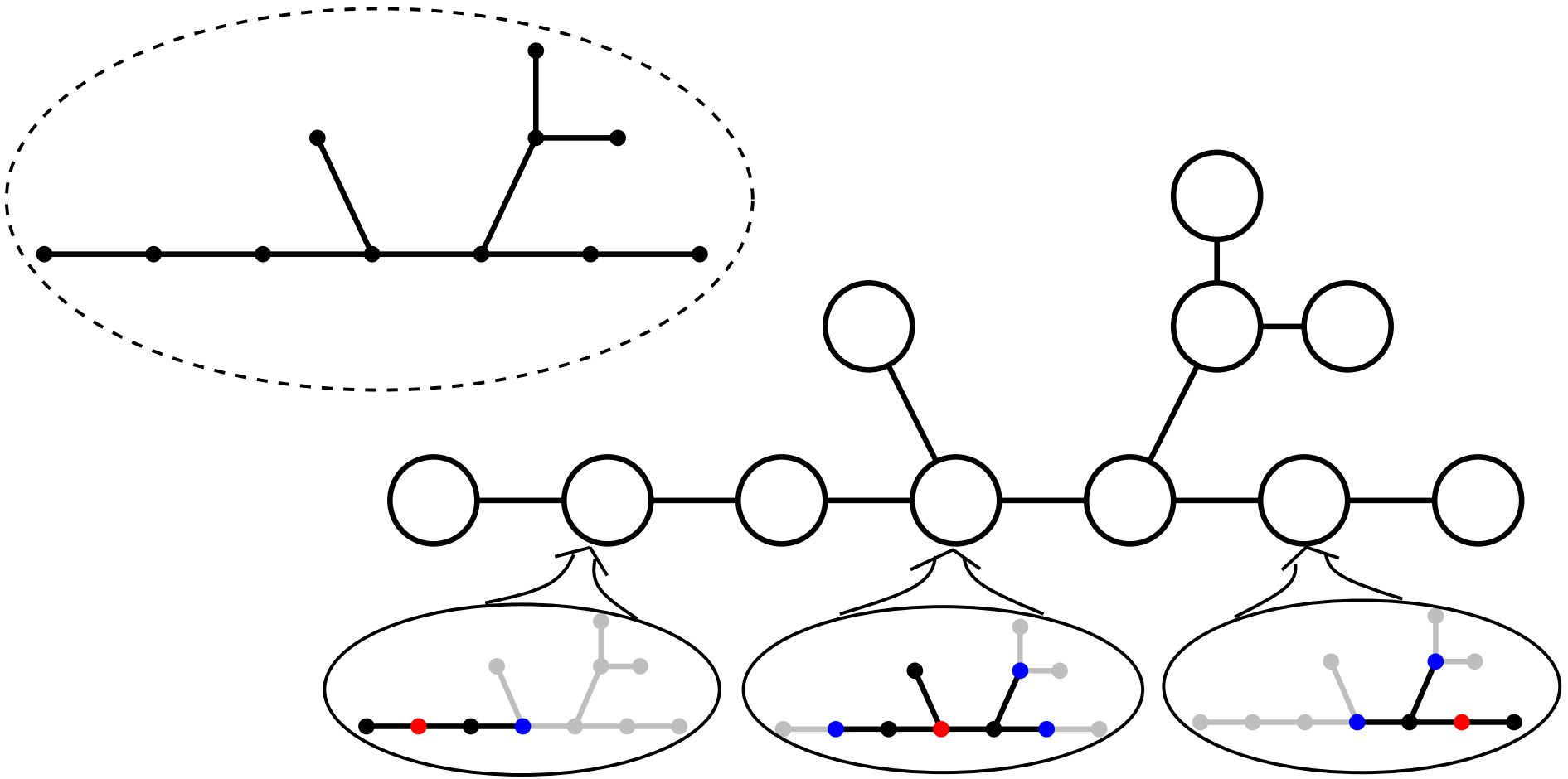
New chain labeling space
= space of all modes

- compute the top M labeling of the new chain [Nilsson'98]

- Complexity $O(DL^2(L\delta + M) + DM\log(DM))$

Trees

- Chains $\rightarrow$ trees
- subchains of length $\delta + 2 \rightarrow$ geodesic balls of radius $r = \lfloor \delta/2 \rfloor + 1$
  - Geod dist: # of edges



Computing Top M Modes

- Complexity:

$$O\left(D^2 dL\delta^2(L+\delta)(L^d+\lambda^d)+D\lambda^2+MD\lambda+MD\log(MD)\right)$$

  - D # of dimensions, L number of discrete values
    d tree degree, δ nbd radius, M number of modes to compute
    λ max # of used local modes for any ball

Computing Top M Modes

- Complexity:

$$O\left(D^2 dL\delta^2(L+\delta)(L^d+\lambda^d)+D\lambda^2+MD\lambda+MD\log(MD)\right)$$

  - D # of dimensions, L number of discrete values
    d tree degree, δ nbd radius, M number of modes to compute
    λ max # of used local modes for any ball

Theoretical Guarantee

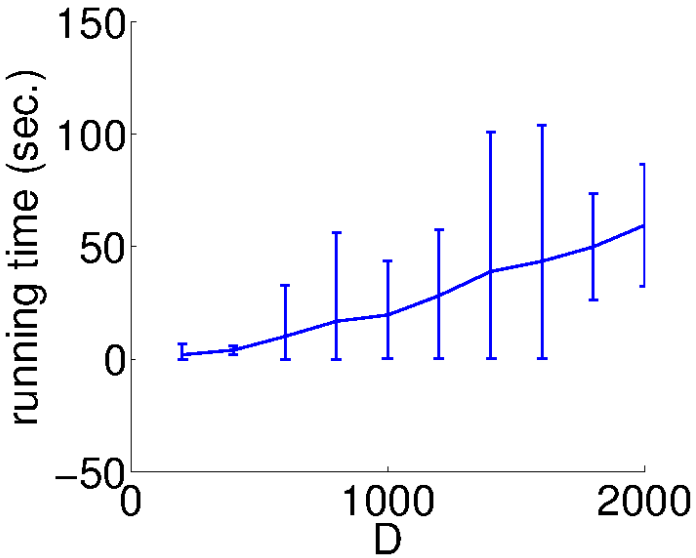
$$P(\widehat{\mathcal{M}^\delta}=\mathcal{M}^\delta)\geq 1-c_0\exp(-c_1 S),\ \text{where } S \text{ number of samples}$$

- $S\to\infty$, accuracy $\to 100\%$
- Assume the distribution can be well modeled with trees
- Assume a probability gap between the top M modes and others
  - When relaxed: top M modes can change,
    but topology is stable (stability theorem)

Scalability and Accuracy

Scalability

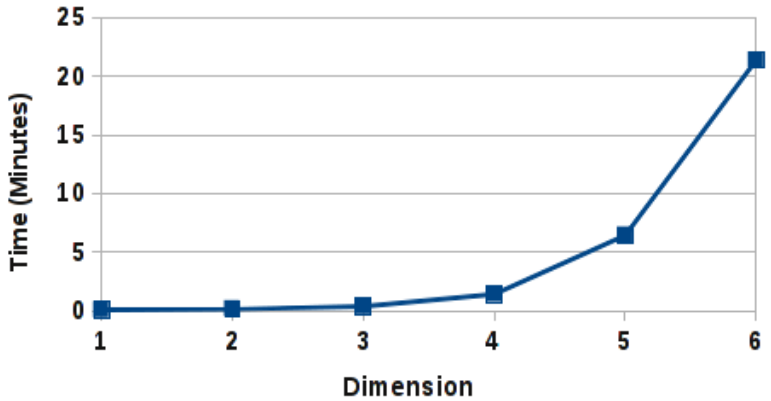
- Overall: $O(D^2)$, D # of dimensions



Real World Data

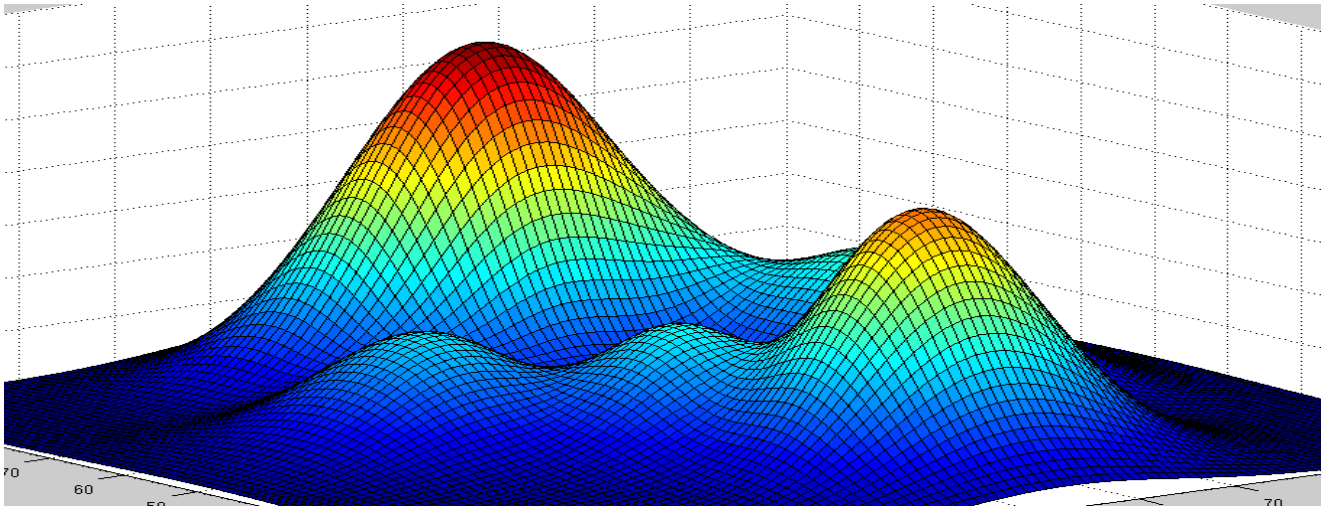
D	Time (Sec.)
434	6.42
548	4.27
1072	18.86

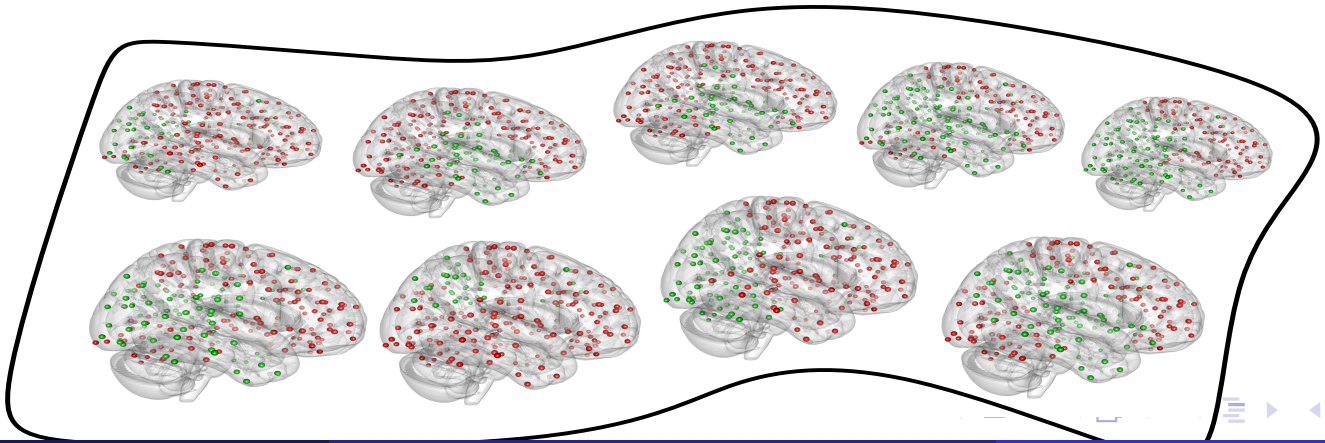
- Previous TDA methods: explode quickly as D increases, could not finish when $D \geq 10$



Application 1: Big Data Analytics

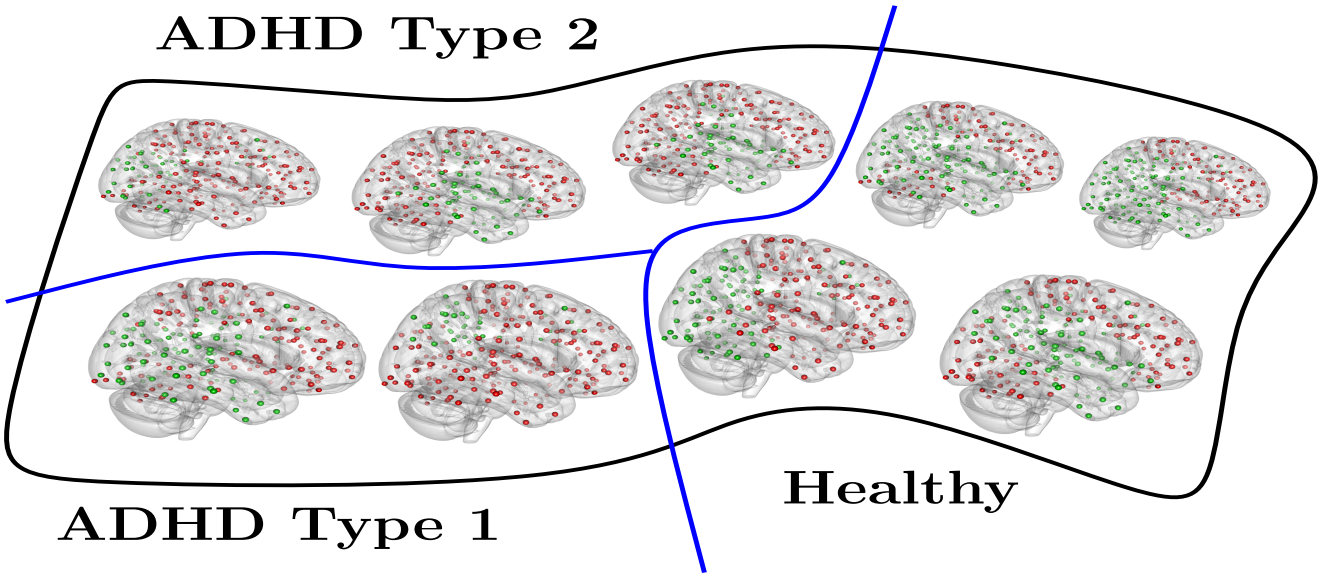
- Structures of brain fMRI data:
  - Each scan is a high-dimensional data
  - Deeper insight
  - Mid-level features to improve predictions





Validation via Clustering

- Partition the dataset using topological structures extracted
- Validate by comparing to human diagnosis
- ADHD-200 dataset (Attention Deficit Hyperactivity Disorder)

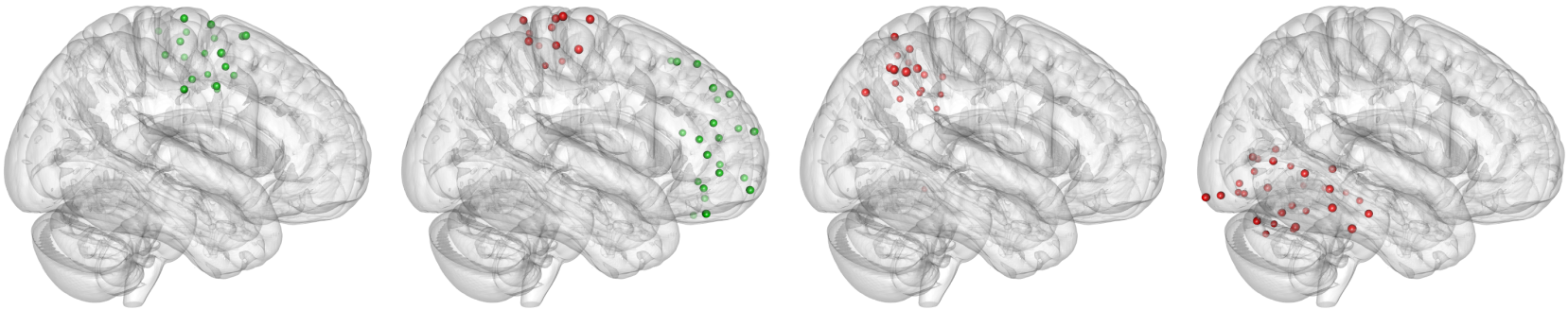


Results

Error Metrics	Modes	k-means
Random Index	0.71	0.57

Mid-level Substructures/Features

- Identifying subregions of brain



Results:

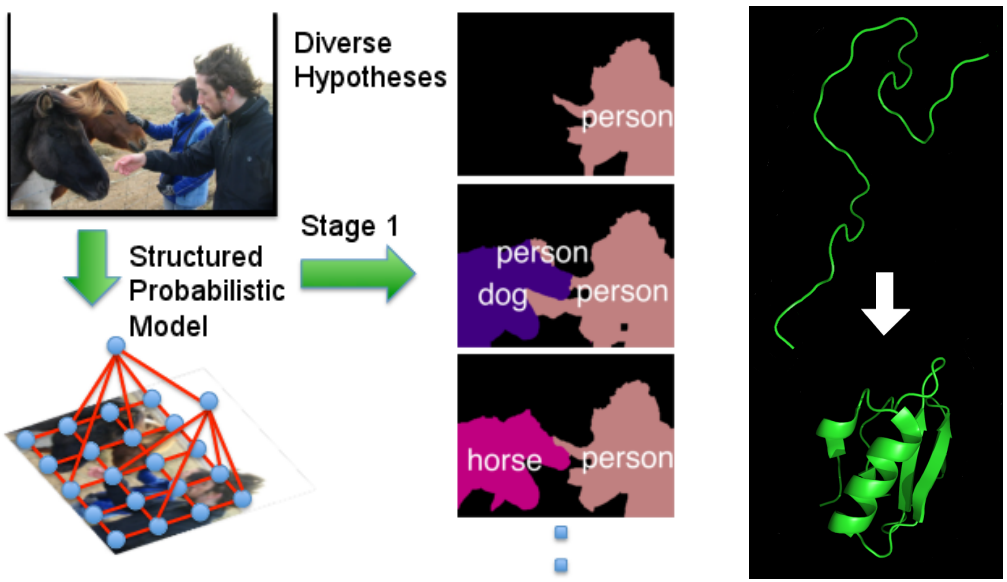
- consistent with known partition of human brain
- Promising learning results:
  - improve performance of previous methods by $\geq 10\%$

[**Chen** *et al.* MLINI 2014]

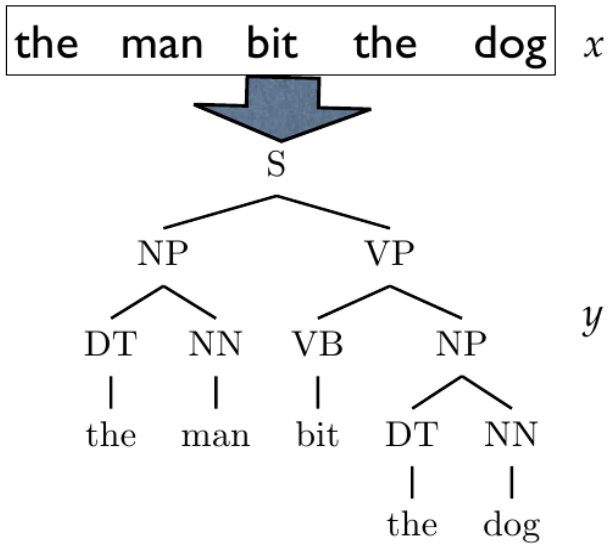
Application 2: Prediction

- Structured prediction:
  finding the highest probability labeling of a given graphical model
- Issue: single prediction not accurate, model limitation

Alternative: multiple predictions (high probability, diverse)
[Lampert NIPS'11, Guzman-Rivera et al. NIPS'12, Prasad *et al.* NIPS'14]



Picture from Batra



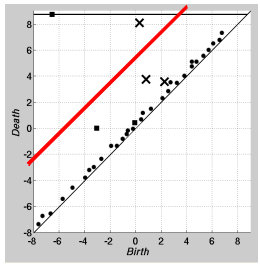
Picture from Huang

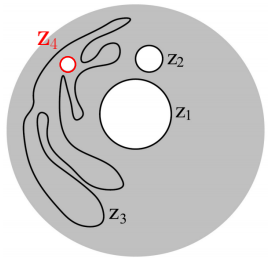
- Our method outperforms state-of-the-art
- Other applications: protein folding, natural language processing, etc.

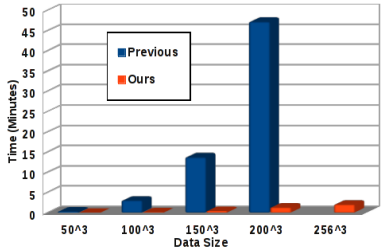
[**Chen** *et al.* AISTATS'13(Oral); NIPS'14 (Spotlight)]

Summary

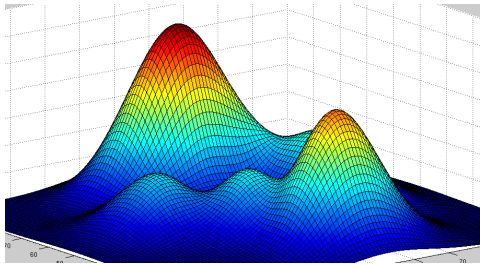
- Topology data analysis: robustly extract global structures
- Improve the method for real world applications

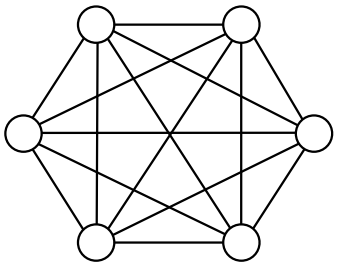


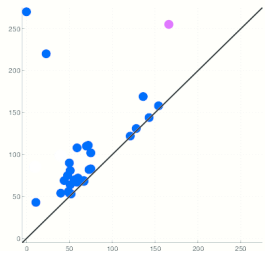




- New approach: using graphical models ($O(D^2)$)

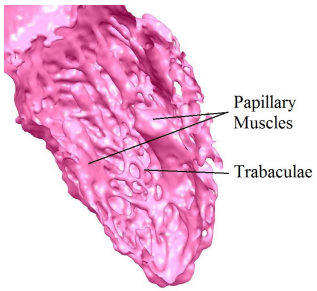


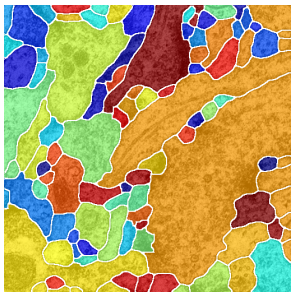




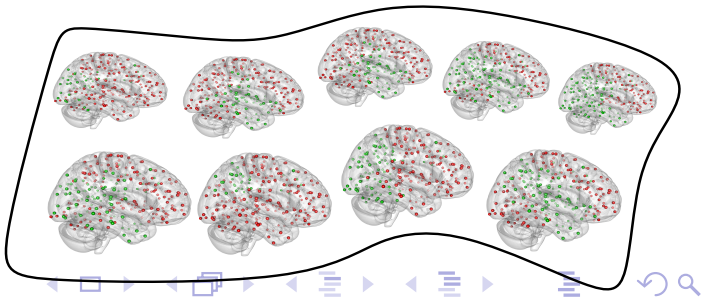
- Applications

low-dim



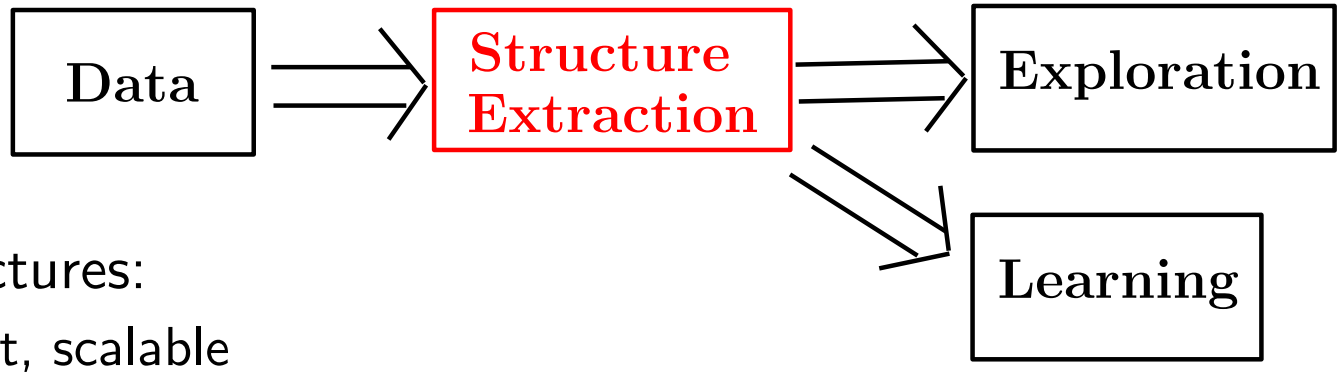


High-dim

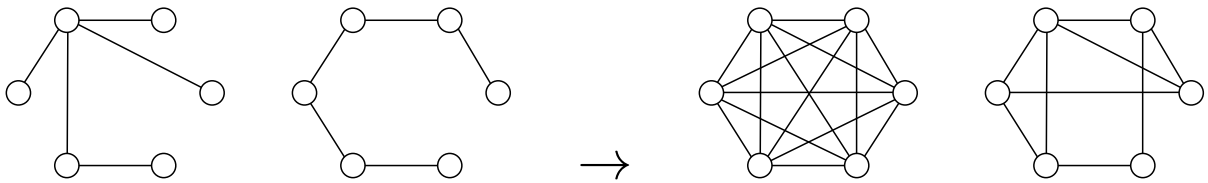


Outline

Topological Summary for Modern Data



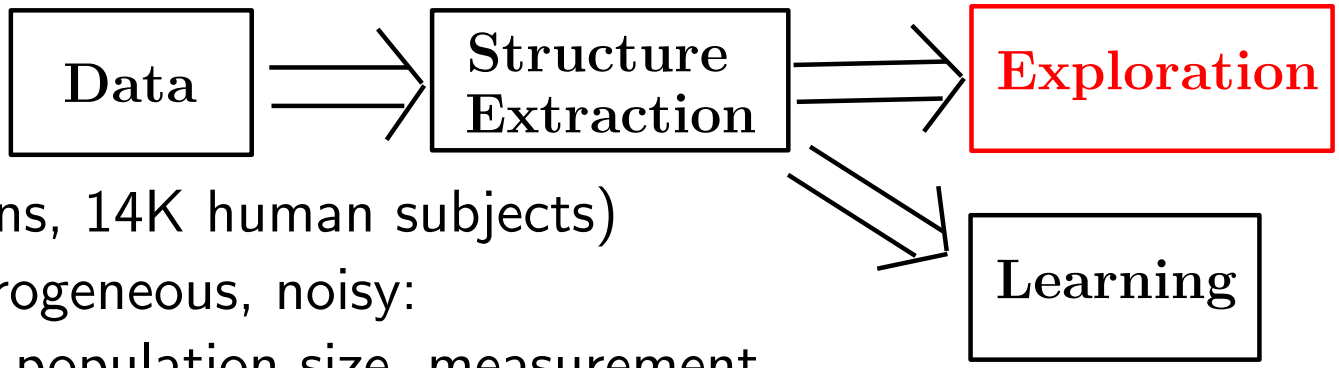
- Topological structures:
  - Global, robust, scalable
  - No assumption on geometry/dimension/smoothness
- Milestones:
  - High-dimensional structures
  - More flexible models



  - Theoretical guarantees
  - Tracking structures through time

Funding Opportunities: NSF-IIS/CCF/DMS

Application: Exploration



fMRI (3.3M data/scans, 14K human subjects)

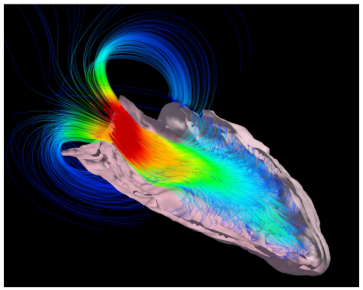
- Challenges: heterogeneous, noisy:
  age, disease type,population size, measurement
- Use topological structures to
  - Characterize the whole collection at different granularities
    structures of each dataset, datasets for each disease, everything
  - Better prediction
- Multimodel:
  - Neuroimaging genomics: combining with genomic data
    Collaboration with Prof. Li Shen (Indiana U. Medical School)
  - Topo. structures at different feature spaces and the joint feature space

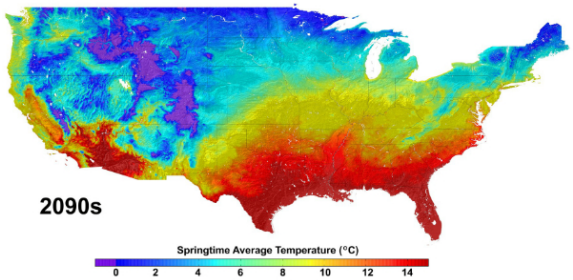
Opportunities: NIH-NIA/NIMH.

Collaboration with Prof. Han Liu (Princeton)

Application: Complex Data

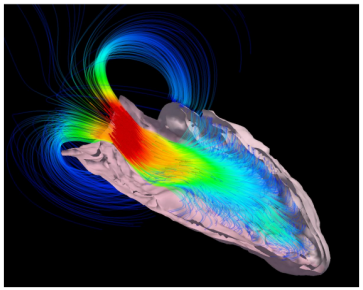
- Spatiotemporal Complex Data
  - Often involves physical phenomena
  - Examples: cardiac flow, climate, etc.

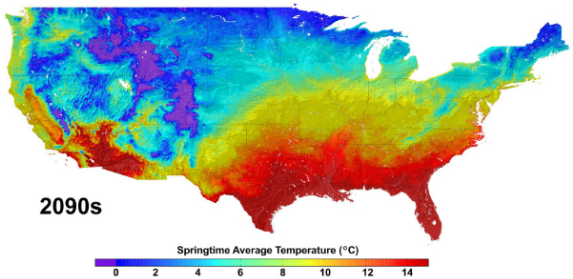


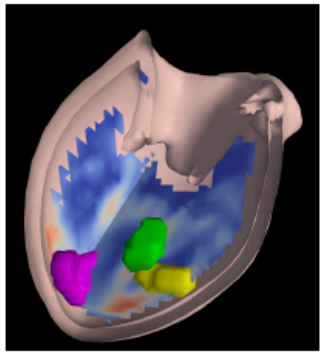


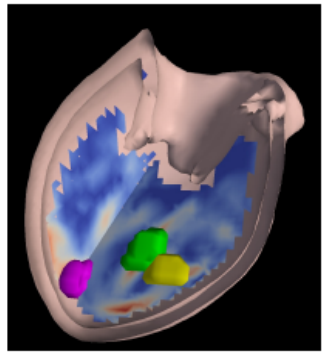
Application: Complex Data

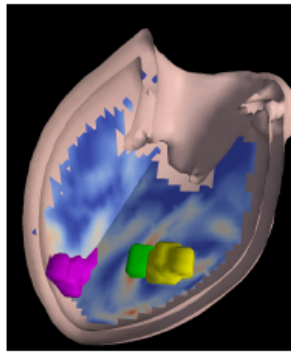
- Spatiotemporal Complex Data
  - Often involves physical phenomena
  - Examples: cardiac flow, climate, etc.











[Kulp, **Chen** *et al.* ISBI 2015]

- Topological Structures
  - As mid-level features
    Proposal submitted to NSF-IIS, Senior Personnel.
  - Long term: Bridging the gap between physics-driven models and data-driven models.
    Especially important for domain experts.

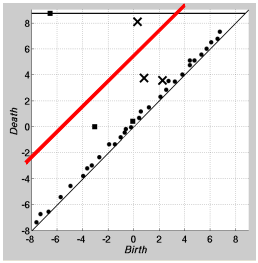
Opportunities: NIH-NHLBI.

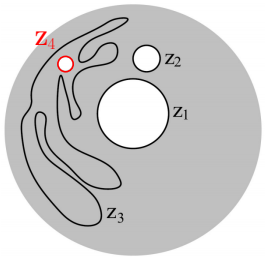
Collaboration with Dr. Leon Axel (NYU Medical School), Climate Central

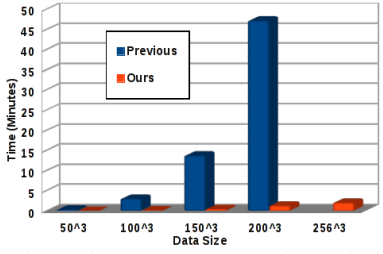
More Applications

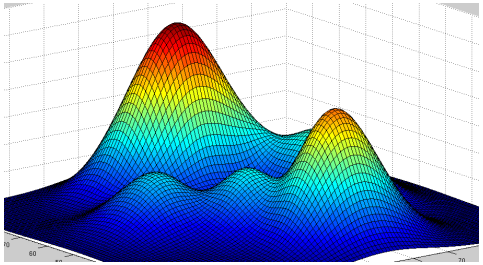
- Complex graph data
  - Social network, gene regulatory network, brain network, etc.
  - Topological structures: connected components, loops
    [Quadrianto, **Chen** and Lampert ICML 2012]
  - Help understand the (MANY) existing community detection methods
- Computer vision
  - Topology: fundamental information of objects
  - Applications: object recognition, tracking, robotic vision, etc.
    [**Chen** *et al.* CVPR'11, ICCV'11]
- System: parallel computing, out-of-core, high performance computing
- Theory: topological structures vs. dimension reduction,
  different model regularizers, deep learning, etc.

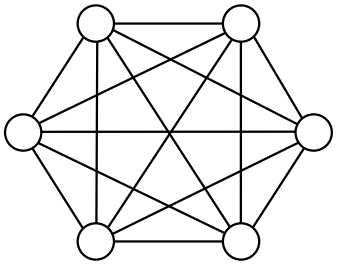
The End

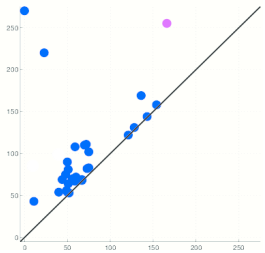


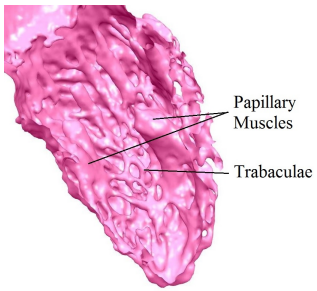


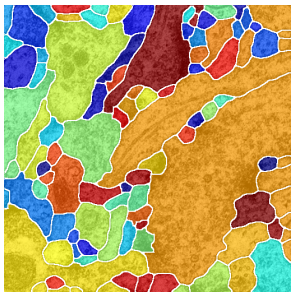


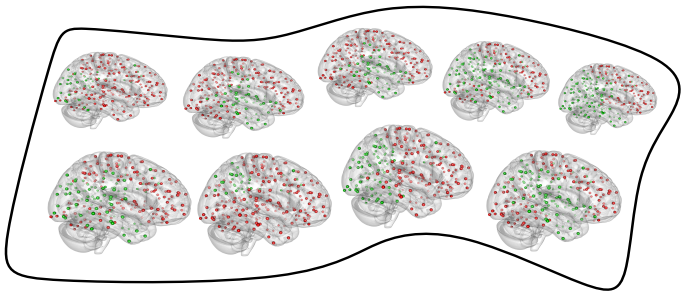












Thank you! Questions?

The End

Topology: The New 'Shape' of Data!

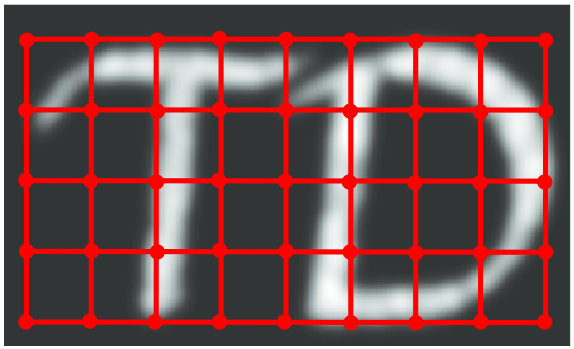
Thank you! Questions?

Bottleneck Distance Between Persistence Diagrams

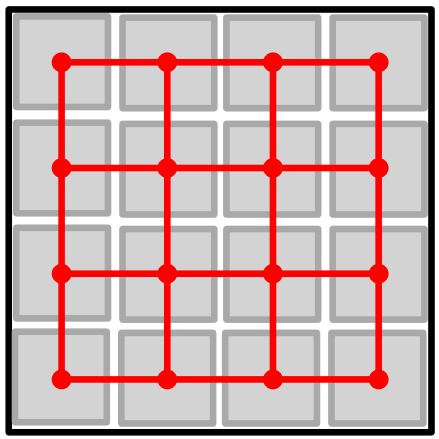
Discretization

[Edelsbrunner *et al.* DCG 2002]
(based on algebraic topology)

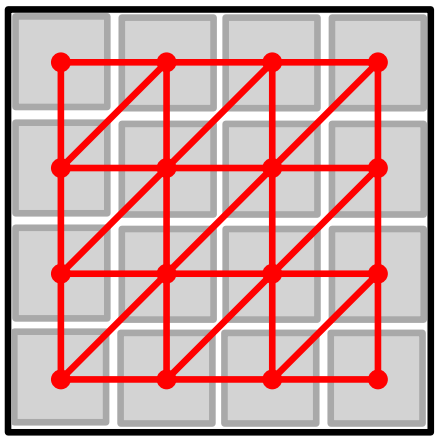
- Density estimation
- Discretize the domain



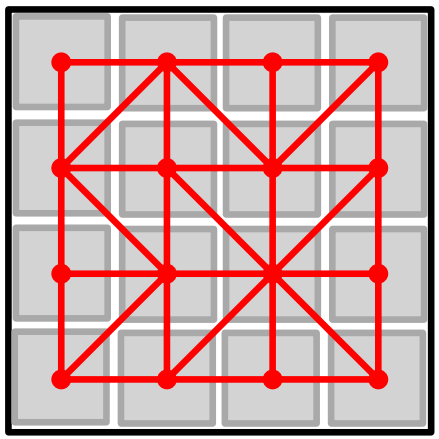
Different discretizations lead to different topologies



digital topology



Freudenthal



adaptive topology

[Kong and Rosenfeld, 1989], [Edelsbrunner and Kerber, 2012],
[Edelsbrunner and Symonova, 2012]

Computation of Saddles ($O(D^2)$)