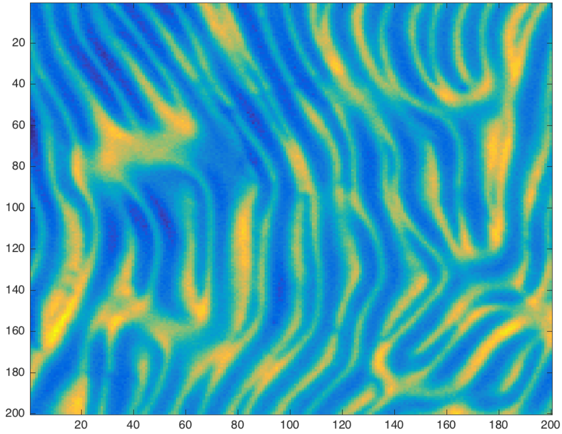
Nonlinear Data Analysis

Lessons and Challenges

Tyrus Berry
George Mason University

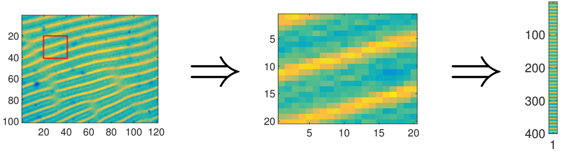
Nov. 29, 2017

MOTIVATING EXAMPLE: NEMATIC LIQUID CRYSTAL



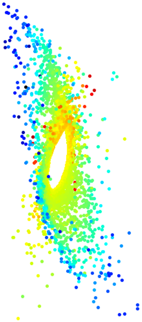
Video provided by Rob Cressman, Krasnow Institute, GMU

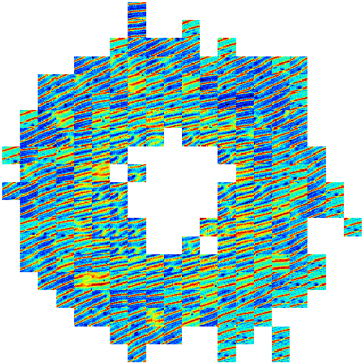
FINDING HIDDEN STRUCTURE IN DATA



$\in \mathbb{R}^{400}$

The sub-image geometry:





OUTLINE

Lessons:

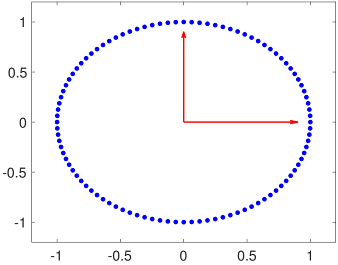
- **Dimensionality:** Intrinsic vs. Extrinsic
- **Non-uniformity:** Respect the density
- **Meta-structure:** Images and times series

Challenges:

- Curse-of-dimensionality (intrinsic)
- Extrapolation

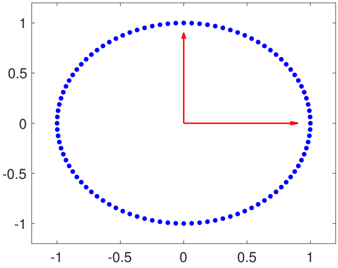
Intrinsic vs. Extrinsic Dimension

100 points on a Circle



θ	x	y
0.0628	0.9980	0.0628
0.1257	0.9921	0.1253
0.1885	0.9823	0.1874
0.2513	0.9686	0.2487
0.3142	0.9511	0.3090
0.3770	0.9298	0.3681
0.4398	0.9048	0.4258
0.5027	0.8763	0.4818
$\vdots$	$\vdots$	
6.0319	0.9686	-0.2487
6.0947	0.9823	-0.1874
6.1575	0.9921	-0.1253
6.2204	0.9980	-0.0628
6.2832	1.0000	-0.0000

INTRINSIC VS. EXTRINSIC DIMENSION



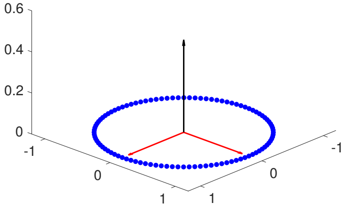
- Intrinsic Dimension $= 1$

$$\theta_i = 2\pi \frac{i}{100}$$

- Extrinsic Dimension $= 2$

$$(x_i, y_i) = (\cos(\theta_i), \sin(\theta_i))$$

INTRINSIC VS. EXTRINSIC DIMENSION



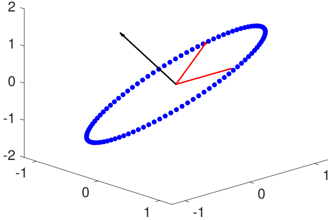
- Intrinsic Dimension $= 1$

$$\theta_i = 2\pi \frac{i}{100}$$

- Extrinsic Dimension $= 3$

$$(x_i, y_i, z_i) = (\cos(\theta_i), \sin(\theta_i), 0)$$

INTRINSIC VS. EXTRINSIC DIMENSION



- Intrinsic Dimension $= 1$

$$\theta_i = 2\pi \frac{i}{100}$$

- Extrinsic Dimension $= 3$

$$x_i = \cos(\theta_i)$$
$$y_i = \sin(\theta_i)$$
$$z_i = x_i + y_i$$

INTRINSIC VS. EXTRINSIC DIMENSION

- Intrinsic Dimension $= 1$

$$\theta_i = 2\pi \frac{i}{100}$$

- Extrinsic Dimension $= 2 + N$

$$\begin{aligned} x_i &= \cos(\theta_i) \\ y_i &= \sin(\theta_i) \\ z_i^1 &= a_1 x_i + b_1 y_i \\ &\vdots \\ z_i^N &= a_N x_i + b_N y_i \end{aligned} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a_1 & a_2 \\ \vdots & \vdots \\ a_N & b_N \end{bmatrix} \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} = A \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix}$$

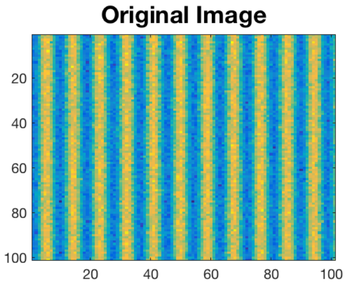
A is a $(2 + N) \times 2$ matrix

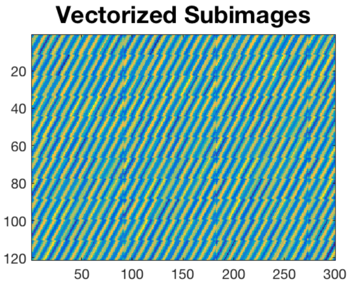
PRINCIPAL COMPONENT ANALYSIS (PCA)

- Matrix times *intrinsic* data $\Rightarrow$ limitless redundancy
- These *linear* redundancies are easy to remove
- PCA finds X given $Y = AX$
- Does this really happen?

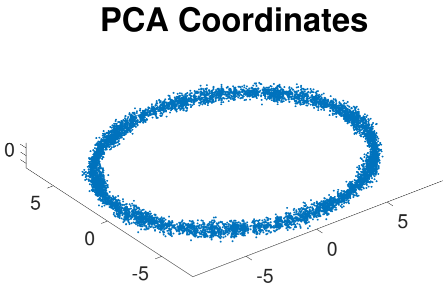
DOES THIS REALLY HAPPEN?

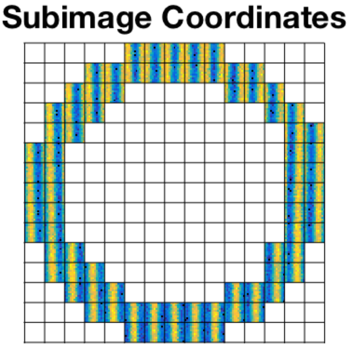
Consider 11×11 subimages from a pattern:



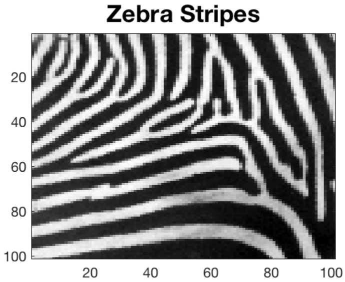


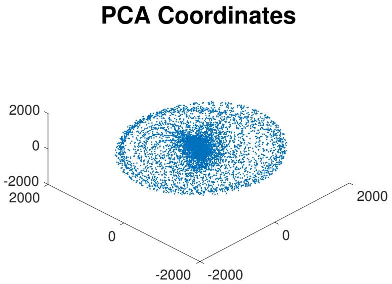
DOES THIS REALLY HAPPEN?

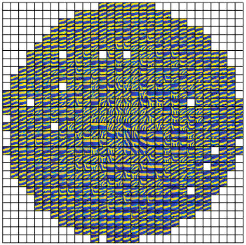




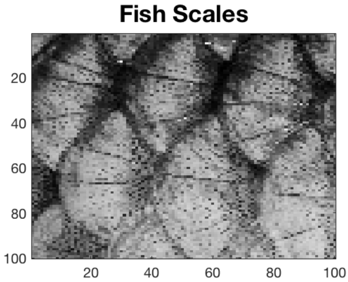
DOES THIS REALLY HAPPEN?

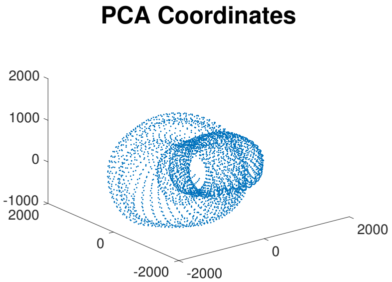


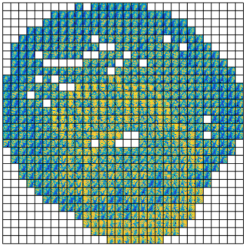




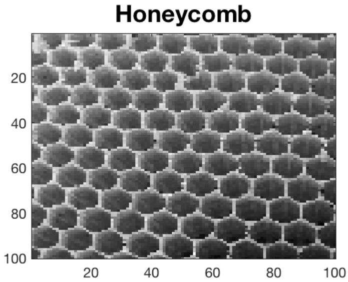
DOES THIS REALLY HAPPEN?

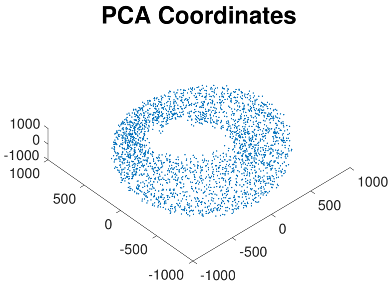


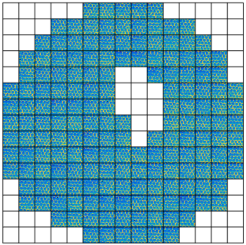




DOES THIS REALLY HAPPEN?





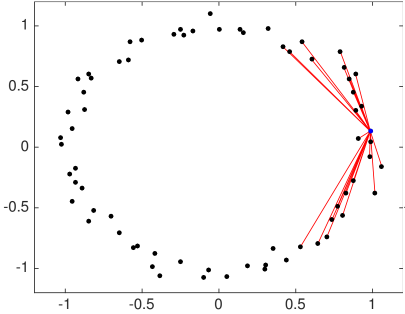


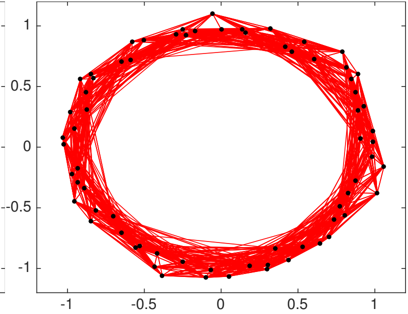
Principal Component Analysis (PCA)

- Matrix times *intrinsic* data $\Rightarrow$ limitless redundancy
- These *linear* redundancies are easy to remove
- PCA finds X given $Y = AX$
- What about nonlinear redundancies?

NONLINEAR ⇒ GRAPH

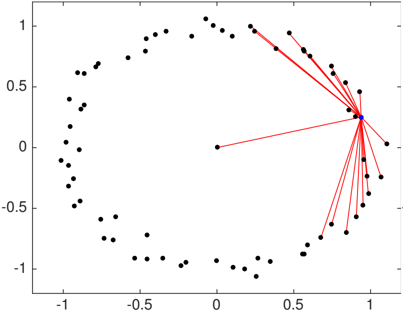
- Represent the nonlinear curved structure with a graph
- Locally linear ⇒ Connect nearby points

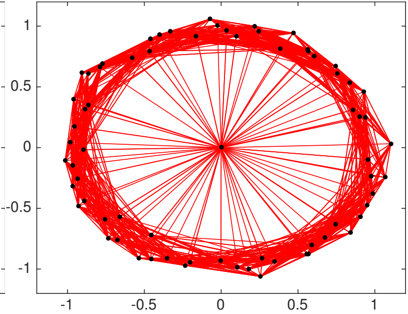




NONLINEAR ⇒ GRAPH

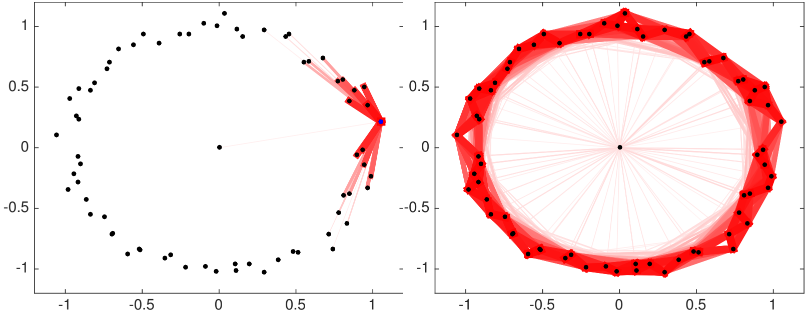
- **Problem:** Noise and outliers can lead to bridging





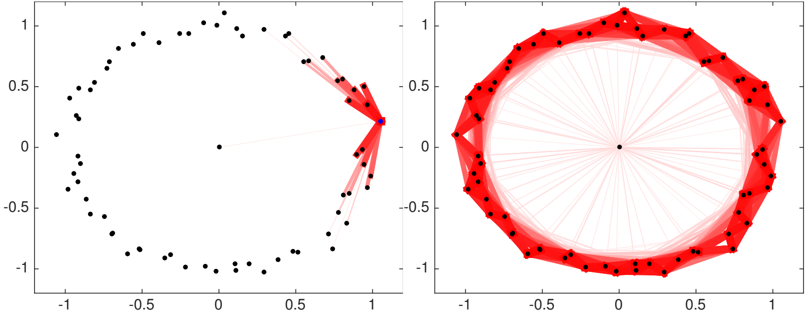
NONLINEAR ⇒ GRAPH

- To prevent bridging, edges weighted: $K_\delta(x,y) = e^{-\frac{||x-y||^2}{4\delta^2}}$
- **Theorem:** Graph encodes all nonlinear information



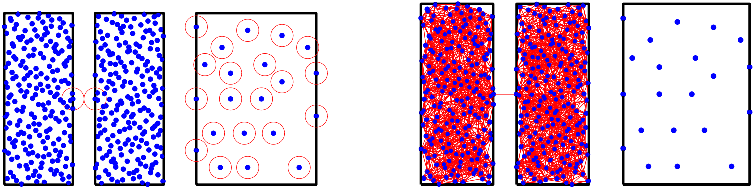
Nonlinear ⇒ Graph

- Equivalently: Restrict to closer points
- Does this always work?



Nonuniform Density: Fixed Balls

Black outlines indicate true clusters:



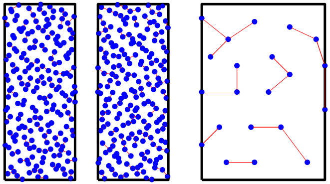
(a) (b)

(a) Dense regions bridged before connecting sparse region

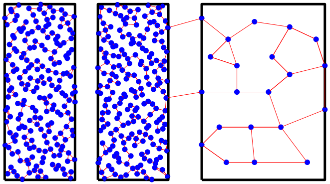
(b) Graph connecting all points with distance less than ϵ

$$||x - y|| < \epsilon$$

NONUNIFORM DENSITY: NEAREST NEIGHBORS (NN)



(c)

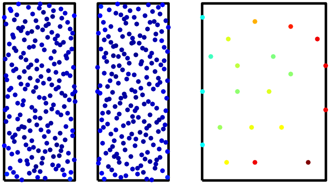


(d)

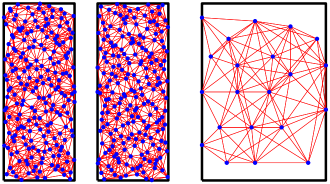
(c) Connect each point to its nearest neighbor (NN)

(d) Connect each point to its two nearest neighbors (2NN)

NONUNIFORM DENSITY: CKNN



(e)



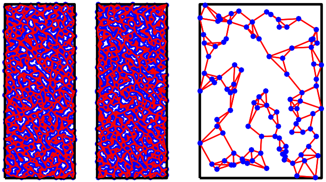
(f)

(e) Distance to 10-th nearest neighbor

(f) **Continuous k-Nearest Neighbors (CkNN)**

$$\frac{||x - y||}{\sqrt{||x - \mathrm{kNN}(x)|| \cdot ||y - \mathrm{kNN}(y)||}} < \delta$$

NONUNIFORM DENSITY: MORE DATA?

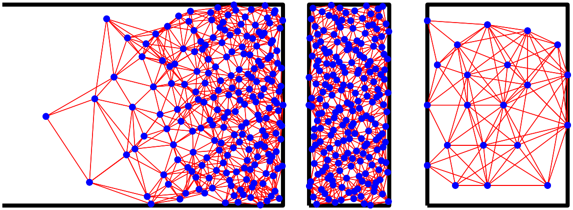


(g)

(g) Five times more data, 4 nearest neighbors works

Does nearest neighbors always work given sufficient data?

NONUNIFORM DENSITY: CONCLUSION



(h)

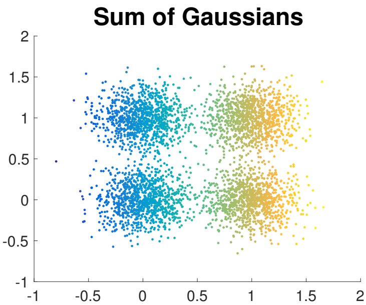
(h) Real data has sparse tails: More data = bigger gaps!

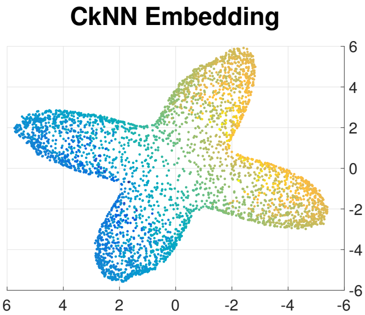
Theorem: NN fails even with infinite data. CkNN succeeds.

HOW CKNN 'SEES' DATA

CkNN defines a symmetric measure of dissimilarity:

$$d_{\mathrm{CkNN}}(x,y) = \frac{||x-y||}{\sqrt{||x-\mathrm{kNN}(x)|| \cdot ||y-\mathrm{kNN}(y)||}}$$

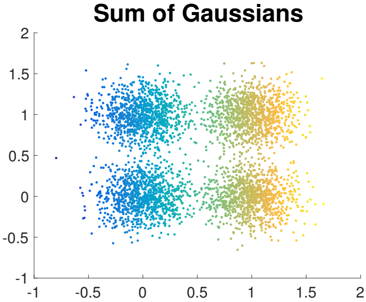


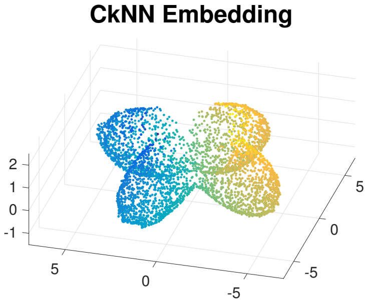


HOW CKNN 'SEES' DATA

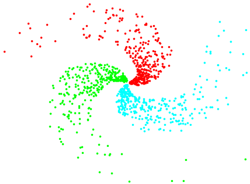
CkNN defines a symmetric measure of dissimilarity:

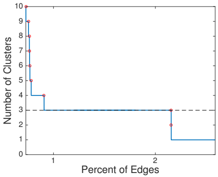
$$d_{\mathrm{CkNN}}(x,y) = \frac{||x-y||}{\sqrt{||x-\mathrm{kNN}(x)|| \cdot ||y-\mathrm{kNN}(y)||}}$$

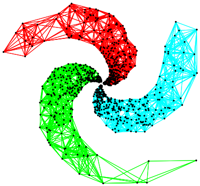


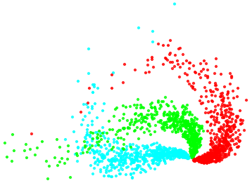


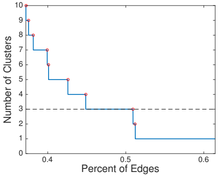
IMPROVED CLUSTERING USING CKNN











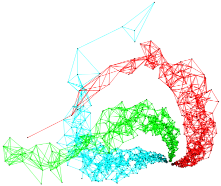
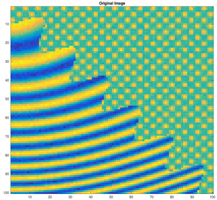
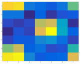


IMAGE SEGMENTATION

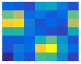
Original Image: Break into subimages



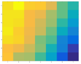
(a)



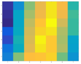
(b)



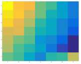
(c)



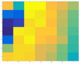
(d)



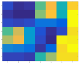
(e)



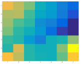
(f)



(g)



(h)



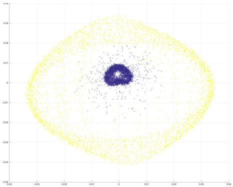
(i)

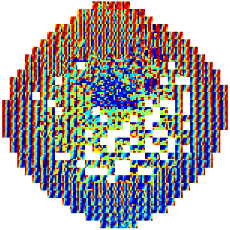
Images produced by Marilyn Vazquez.

IMAGE SEGMENTATION

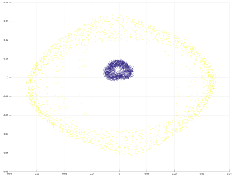
Clustering shown projected to two principal components

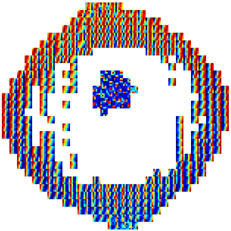
all points





with low density points removed

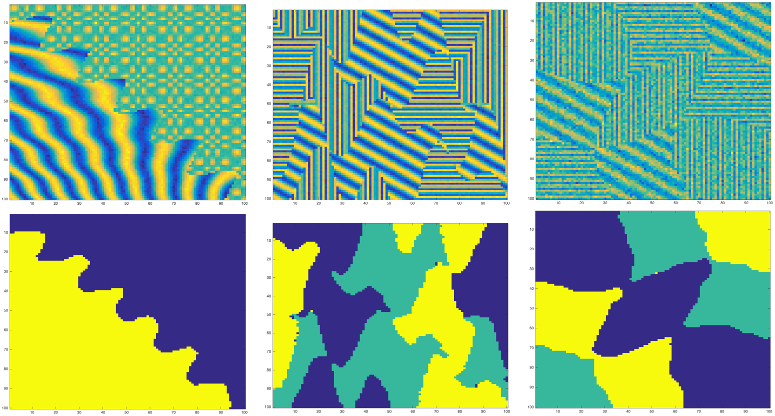




Images produced by Marilyn Vazquez.

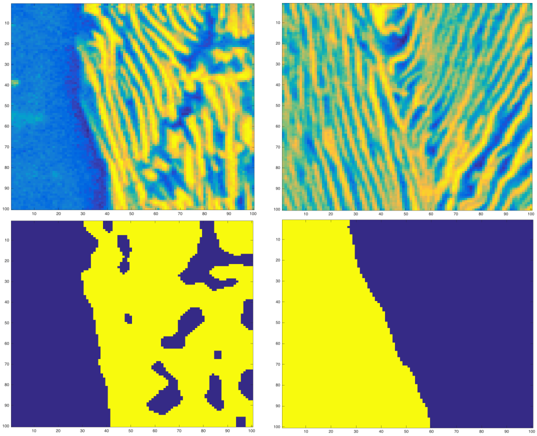
Image segmentation

Results - synthetic images



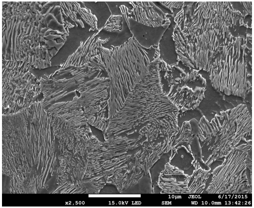
Images produced by Marilyn Vazquez.

IMAGE SEGMENTATION: REAL IMAGES

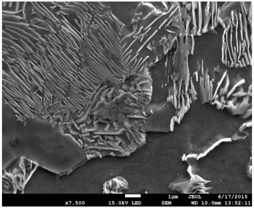


Images produced by Marilyn Vazquez.

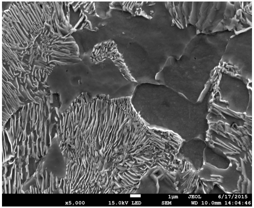
IMAGE SEGMENTATION: REAL IMAGES



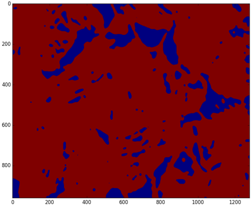
(g)



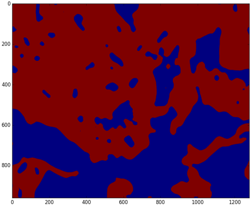
(h)



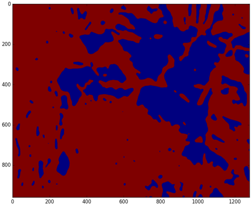
(i)



(j)



(k)

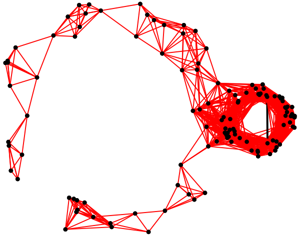


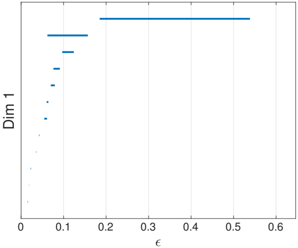
(l)

Original images by Mark R. Stoudt and Steve P. Mates. Analysis by Marilyn Vazquez.

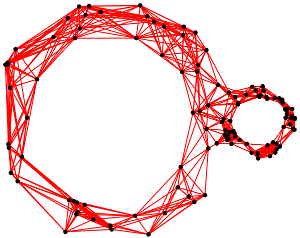
Persistent vs. Consistent homology

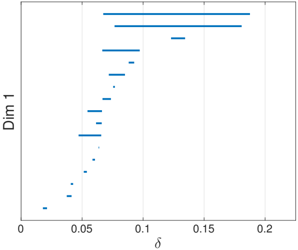
ε-ball





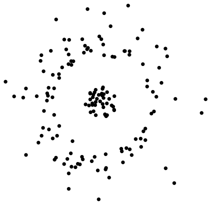
CkNN

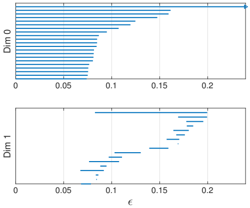


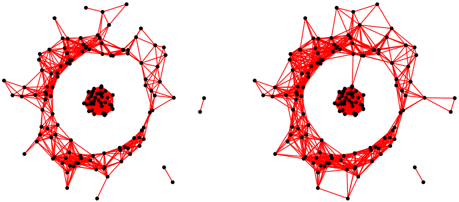


Persistent vs. Consistent homology

A noncompact example, with ε-balls



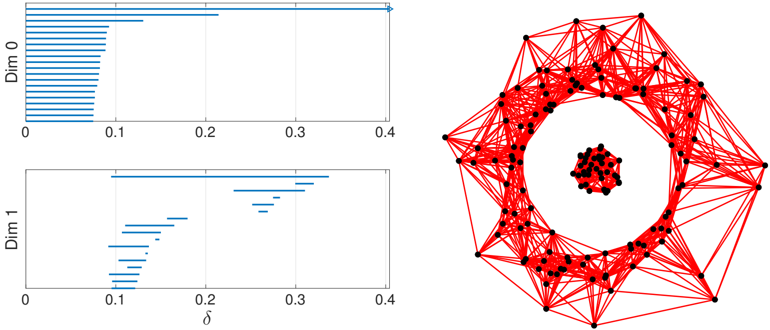




Adding data cannot help.

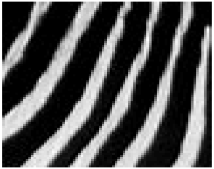
PERSISTENT VS. CONSISTENT HOMOLOGY

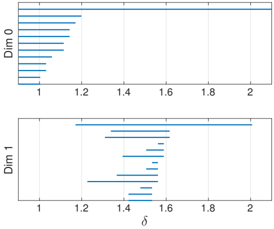
Noncompact example, with CkNN

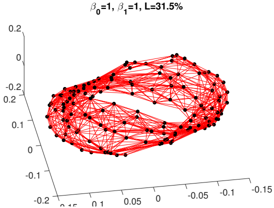


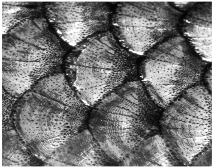
Identifying patterns

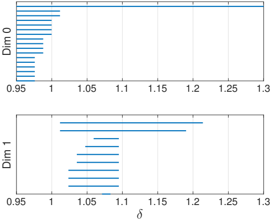
Compute homology of point cloud of $p \times p$ subimages

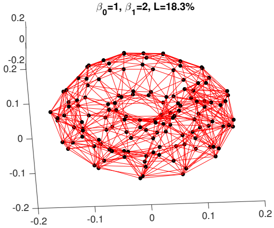






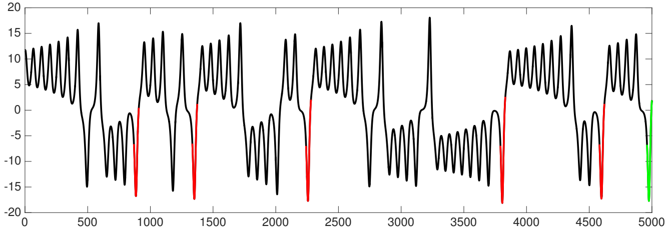




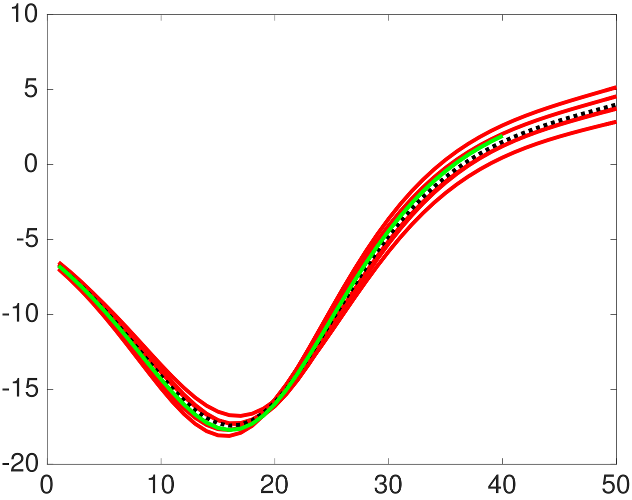


SPATIOTEMPORAL DATA

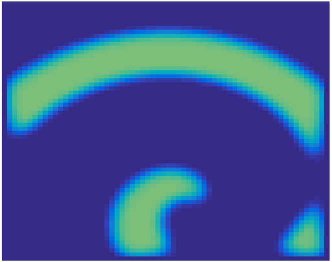
- Spatial ⇒ Short spatial windows (subimages)
- Temporal ⇒ Short time windows (delay embedding)



TAKENS RECONSTRUCTION



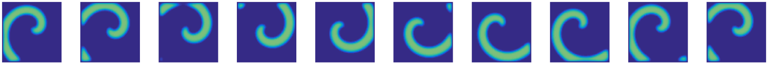
SPIRAL WAVES



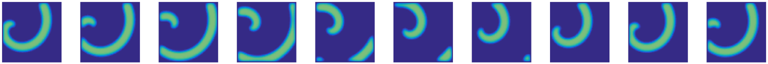
$$
\begin{aligned}
u_t &= \Delta u + \frac{1}{\rho} u(1-u)\left(u - \frac{v+b}{a}\right) \\
v_t &= u - v
\end{aligned}
$$

(D. Barkley 1991)

SPIRAL WAVES

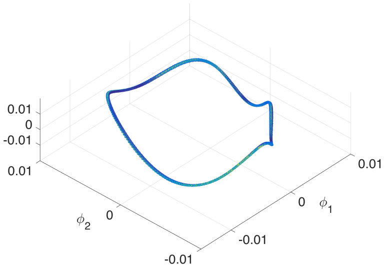


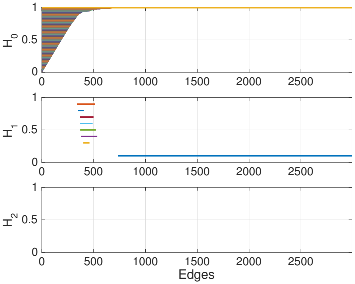
... and later ...



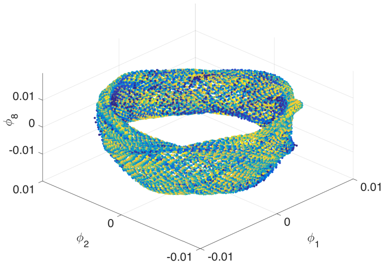
Depending on *a* and *b*, spirals may or may not meander.

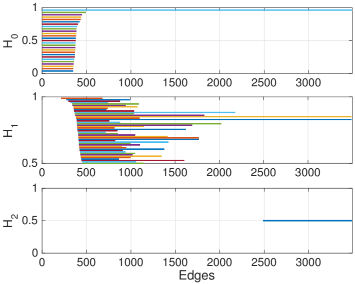
Spiral waves: Non-meandering



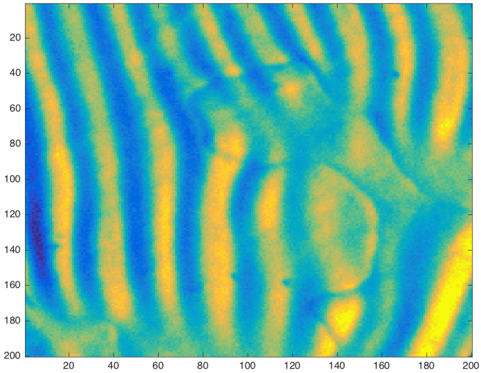


Spiral waves: Meandering





LIQUID CRYSTAL EXPERIMENT

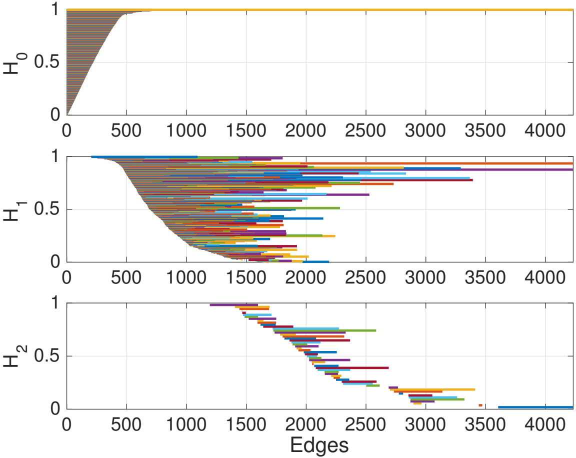


Electroconvection in liquid crystal produces spatiotemporal patterns.

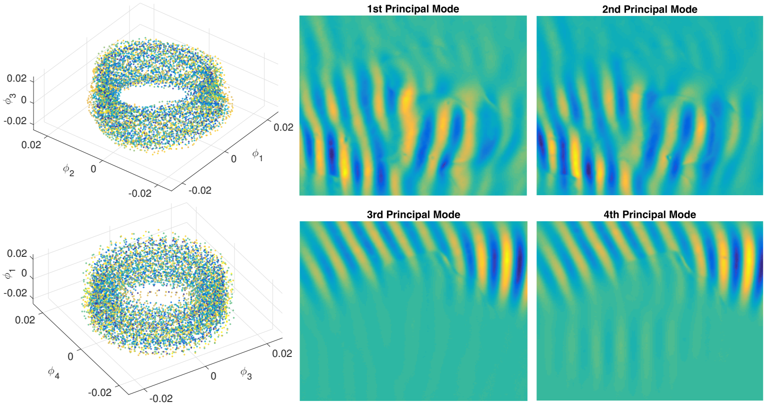
Sample is 0.1×0.1 mm and 25 μm thick.

Driven at 22 V.

LIQUID CRYSTAL



LIQUID CRYSTAL

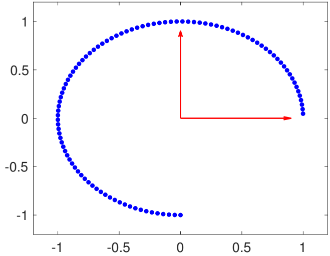


Curse-of-(intrinsic)-Dimensionality

- **Try** to cut into independent components
- Otherwise math/stat says it is impossible
- Need more/better assumptions and/or questions
- **Better assumptions:** Smoothness
- **Better questions:** Feature of interest (supervised)

EXTRAPOLATION

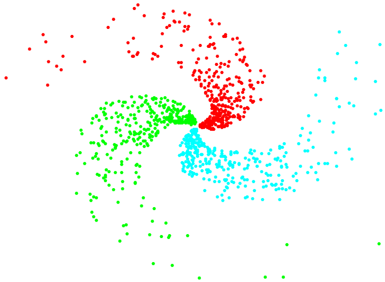
- Given only part of a structure recover the whole



- Need to exploit symmetry

EXTRAPOLATION

- Given only part of a structure recover the whole



?

- Need to exploit symmetry