Open Source Software for TDA

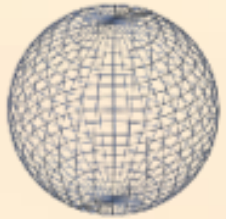
ACM-BCB Workshop on TDA
October 2, 2016

by Svetlana Lockwood

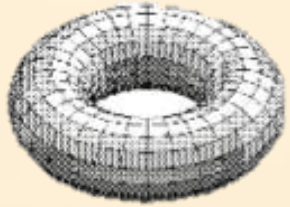
Topological Data Analysis

1. Persistence-Way

- Topological analysis using persistent homology
- Finds topological invariants in data (# of connected components, enclosed voids, etc.)



$\beta_0 = 1$
$\beta_1 = 0$
$\beta_2 = 1$

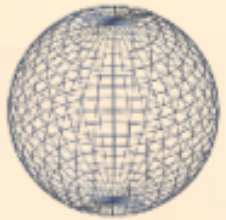


$\beta_0 = 1$
$\beta_1 = 2$
$\beta_2 = 1$

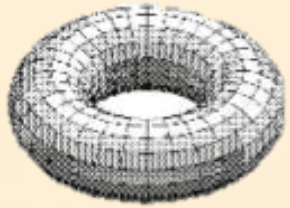
Topological Data Analysis

1. Persistence-Way

- Topological analysis using persistent homology
- Finds topological invariants in data (# of connected components, enclosed voids, etc.)



$\beta_0 = 1$
$\beta_1 = 0$
$\beta_2 = 1$



$\beta_0 = 1$
$\beta_1 = 2$
$\beta_2 = 1$

2. Mapper-Way

- Apply a filter function to project data onto a lower dimensional space
- Performs partial clustering in the level sets

TDA: the Persistence-Way (# 1)

- A number of free software has appeared recently
- R package – “TDA”
- A number of benefits:
  - Familiar R environment
  - Implements 2 types of representation (barcodes & birth-death)
  - R interface to efficient C++ libraries of **GUDHI**, **Dionysus** and **PHAT**

TDA: the Persistence-Way (# 1)

- TDA package for R is developed by
  - Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria, Vincent Rouvreau
- Some of examples from:
  - Fasy, Brittany Terese, Jisu Kim, Fabrizio Lecci, and Clément Maria. "Introduction to the R package TDA." arXiv preprint arXiv:1411.1830 (2014).
  - Kim, Jisu. "Tutorial on the R package TDA."

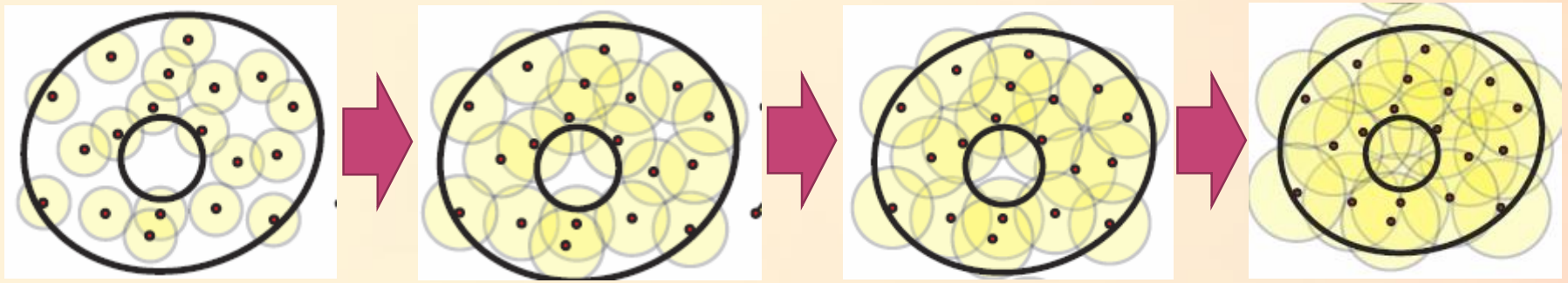
TDA: the Persistence-Way (# 1)

- Goal: to discover underlying shape of data

TDA: the Persistence-Way (# 1)

- Goal: to discover underlying shape of data

Data

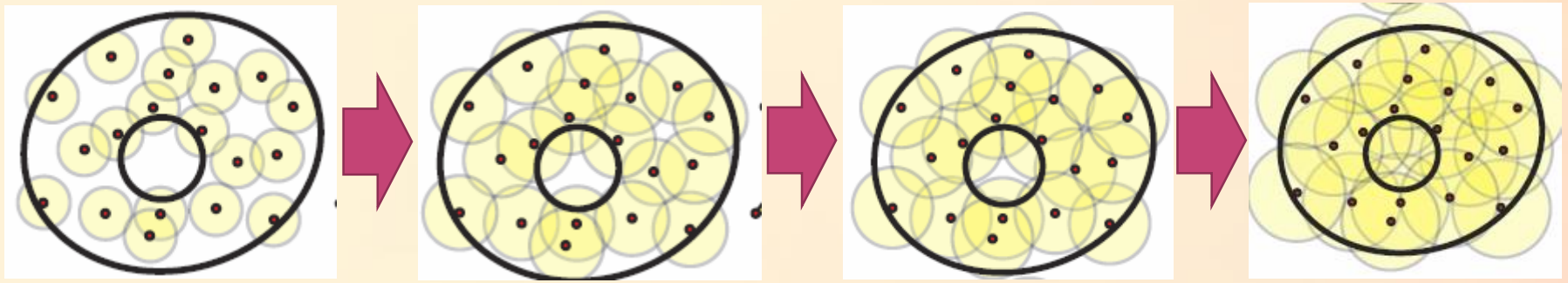


Ghrist, R., 2008. Barcodes: the persistent topology of data.

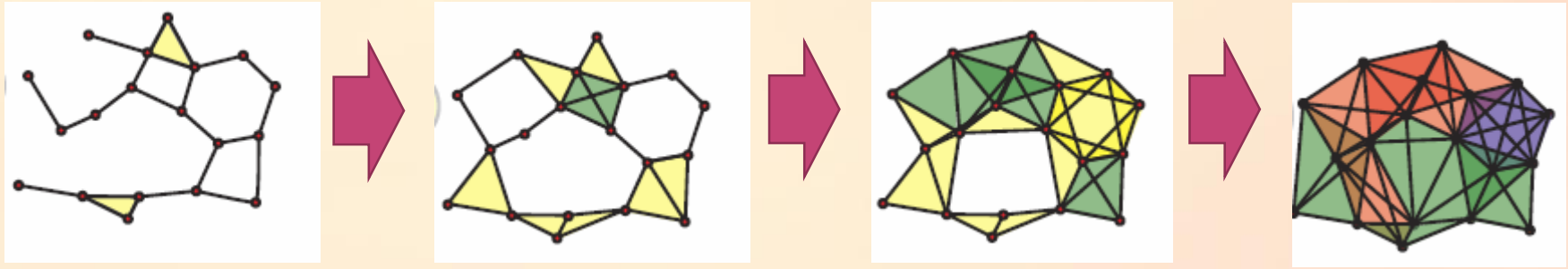
TDA: the Persistence-Way (# 1)

- Goal: to discover underlying shape of data

Data



Topological Features

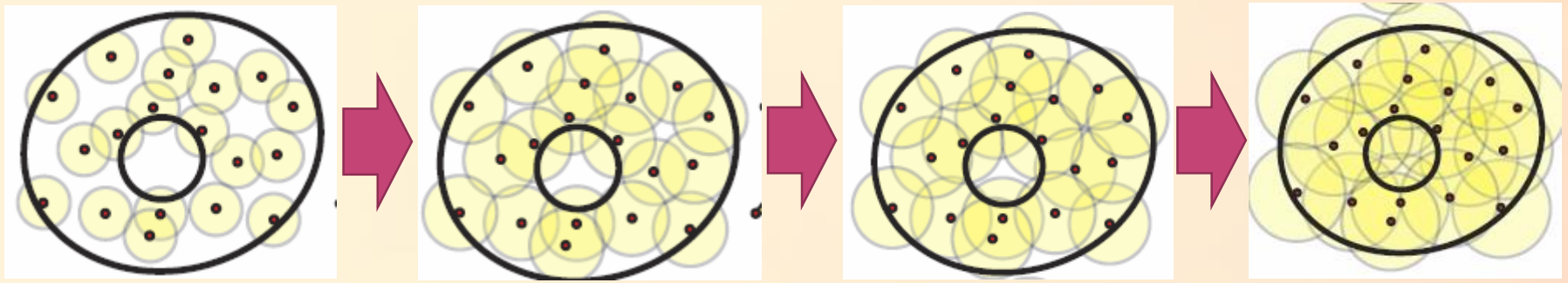


Ghrist, R., 2008. Barcodes: the persistent topology of data.

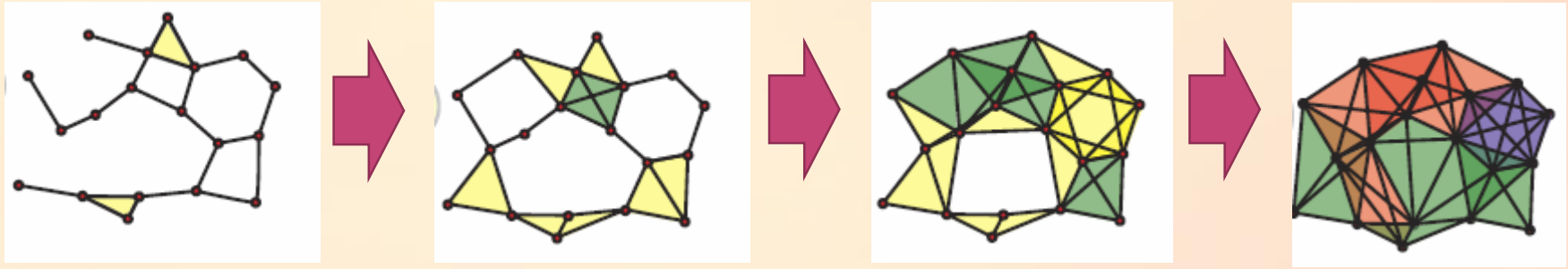
TDA: the Persistence-Way (# 1)

- Goal: to discover underlying shape of data
- (switch to R)

Data



Topological Features

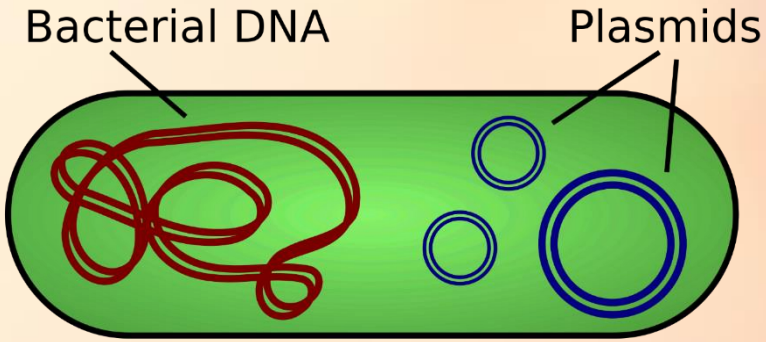


Ghrist, R., 2008. Barcodes: the persistent topology of data.

Plasmids Data

- Plasmids are mobile elements
- Exchange genetic material
- 831 plasmids (see table)
- Original data: 831 plasmids by **81898** features
- Computed pairwise genetic distance → 831 x 831 matrix
- Want to see if there is any "interesting" structure

(switch to R)



Subgroup	Count
1. Alpha	159
2. Beta	85
3. Gamma	519
4. Delta/epsilon	68
Total plasmids	**831**

Pictures adapted from http://www.scienceprofonline.com

Plasmids Data

	[,1]	[,2]
[1,]	351	471
[2,]	292	471
[3,]	351	570
[4,]	292	570

351

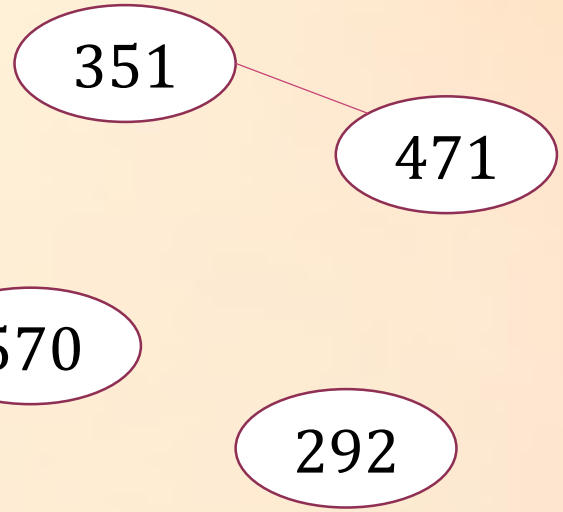
471

570

292

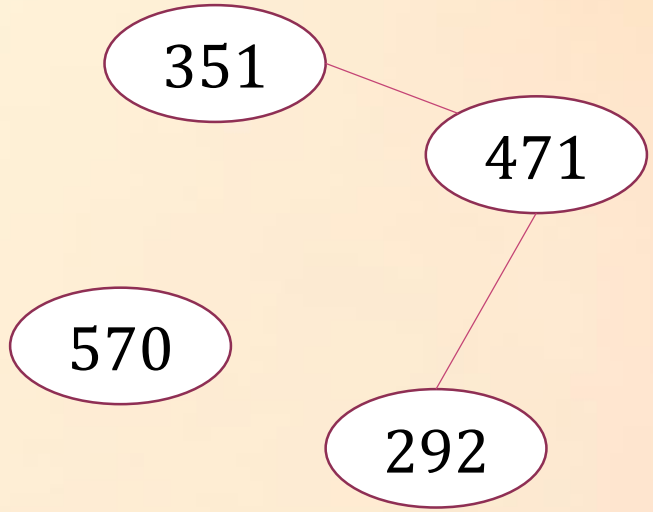
Plasmids Data

	[,1]	[,2]
[1,]	351	471
[2,]	292	471
[3,]	351	570
[4,]	292	570



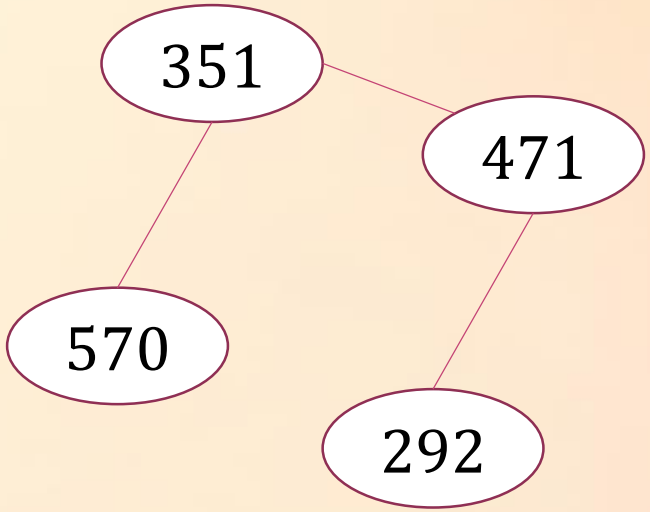
Plasmids Data

	[,1]	[,2]
[1,]	351	471
[2,]	292	471
[3,]	351	570
[4,]	292	570



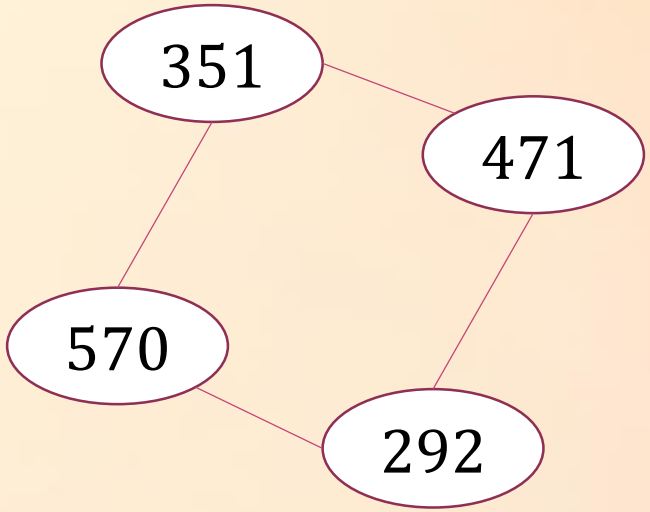
Plasmids Data

	[,1]	[,2]
[1,]	351	471
[2,]	292	471
[3,]	351	570
[4,]	292	570



Plasmids Data

	[,1]	[,2]
[1,]	351	471
[2,]	292	471
[3,]	351	570
[4,]	292	570



Other Software For Persistent Homology

Other open source software is available for computing persistent homology

Interface to Matlab/Octave

Software	Installation	Complex	Boundary matrix	Barcodes	Visualization	Data Set Size	Ease of Use
JavaPlex	✓	✓	✓	✓	✓	small	easy
Perseus	✓	✓	✓	✓	✓	small	easy
Dionysus	--	✓	✓	✓	--	medium	medium
DIPHA	--	✓	✓	✓	✓	large	hard
GUDHI	--	✓	✓	✓	--	large	hard

arxiv 2015, N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington

TDA: the Mapper-Way (# 2)

TDA: the Mapper-Way (# 2)

- Apply a filter function to project data onto a lower dimensional space

TDA: the Mapper-Way (# 2)

- Apply a filter function to project data onto a lower dimensional space
- Performs partial clustering in the level sets using standard clustering algorithms to subsets of the original data

TDA: the Mapper-Way (# 2)

- Apply a filter function to project data onto a lower dimensional space
- Performs partial clustering in the level sets using standard clustering algorithms to subsets of the original data
- Goal: to understand the interaction of the partial clusters formed in this way with each other

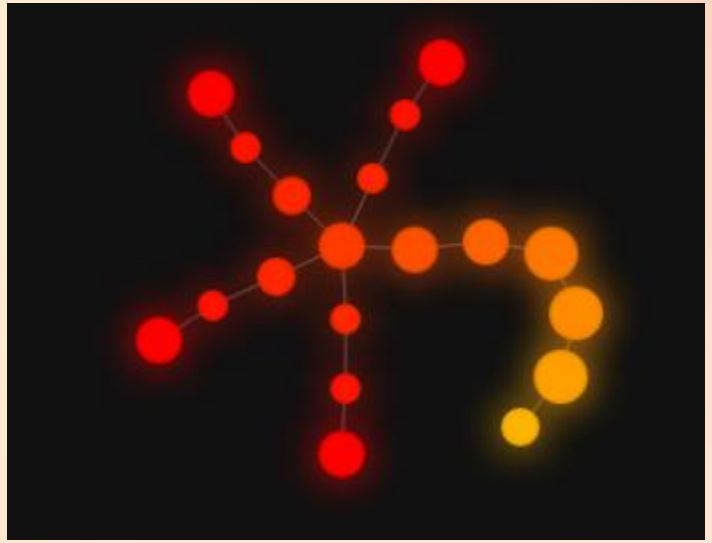
TDA: the Mapper-Way (# 2)

- Apply a filter function to project data onto a lower dimensional space
- Performs partial clustering in the level sets using standard clustering algorithms to subsets of the original data
- Goal: to understand the interaction of the partial clusters formed in this way with each other
- A few open source software exists
  - However all have some limitations

TDA: the Mapper-Way (# 2)

- I'll present Python-based version developed by MLWave & examples from https://github.com/MLWave/kepler-mapper

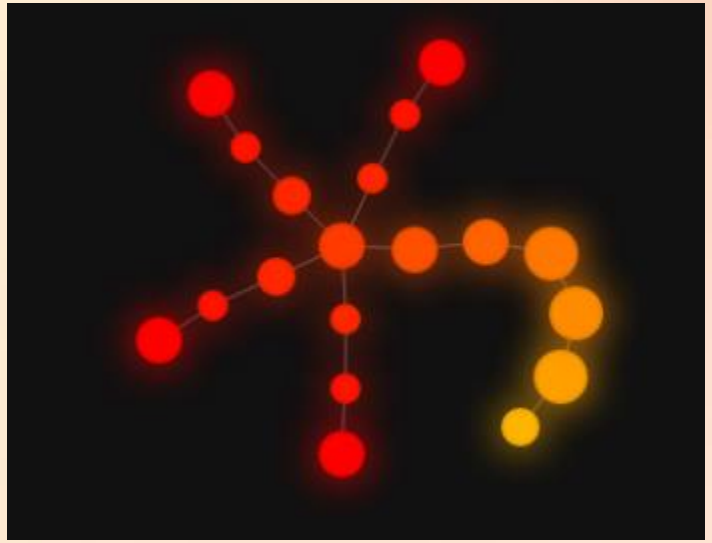
MLWave



TDA: the Mapper-Way (# 2)

- I'll present Python-based version developed by MLWave & examples from https://github.com/MLWave/kepler-mapper
- Pros:
  - Simple programming interface
  - Makes use of existing python ML libraries
  - Nice visualizations
- Cons:
  - Limited coloring
  - Not completely automated

MLWave



Python Mappers: Prerequisites

- I highly recommend installing Anaconda
  - Saves a lot of troubles
  - Comes with SciPy, NumPy, scikit-learn
  - Includes Python IDE and package manager (pip)
- Copy **km.py** from MLWave into Anaconda Lib folder

Intro Mapper Example: MNIST digits

Intro example from MLWave

- The MNIST database of handwritten digits
- Thousands of digits

Intro Mapper Example: MNIST digits

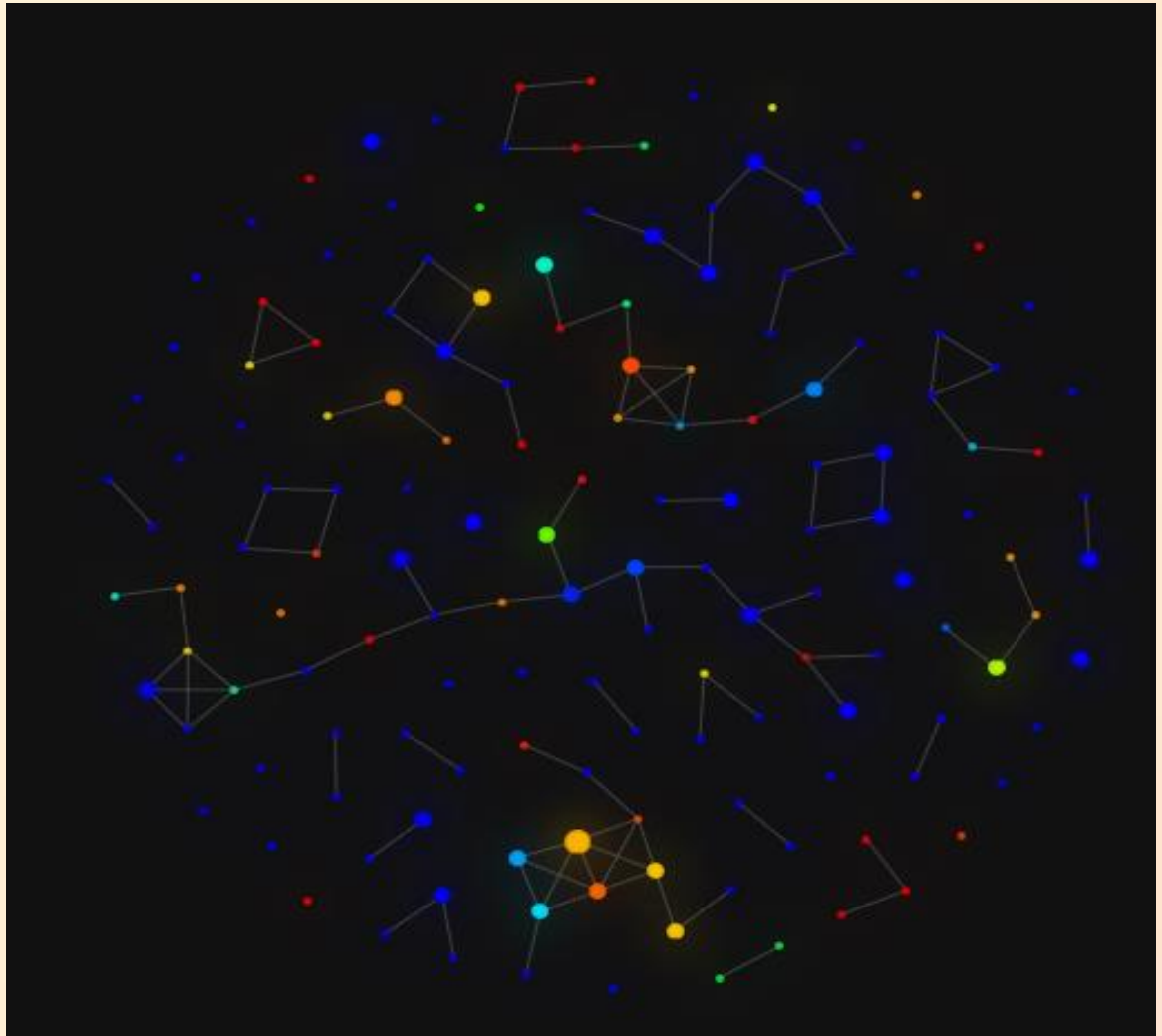
Intro example from MLWave

- The MNIST database of handwritten digits
- Thousands of digits
- Each digit is represented by 8x8 pixel image
- Goal: cluster handwritten digits according to their value

(switch to python)

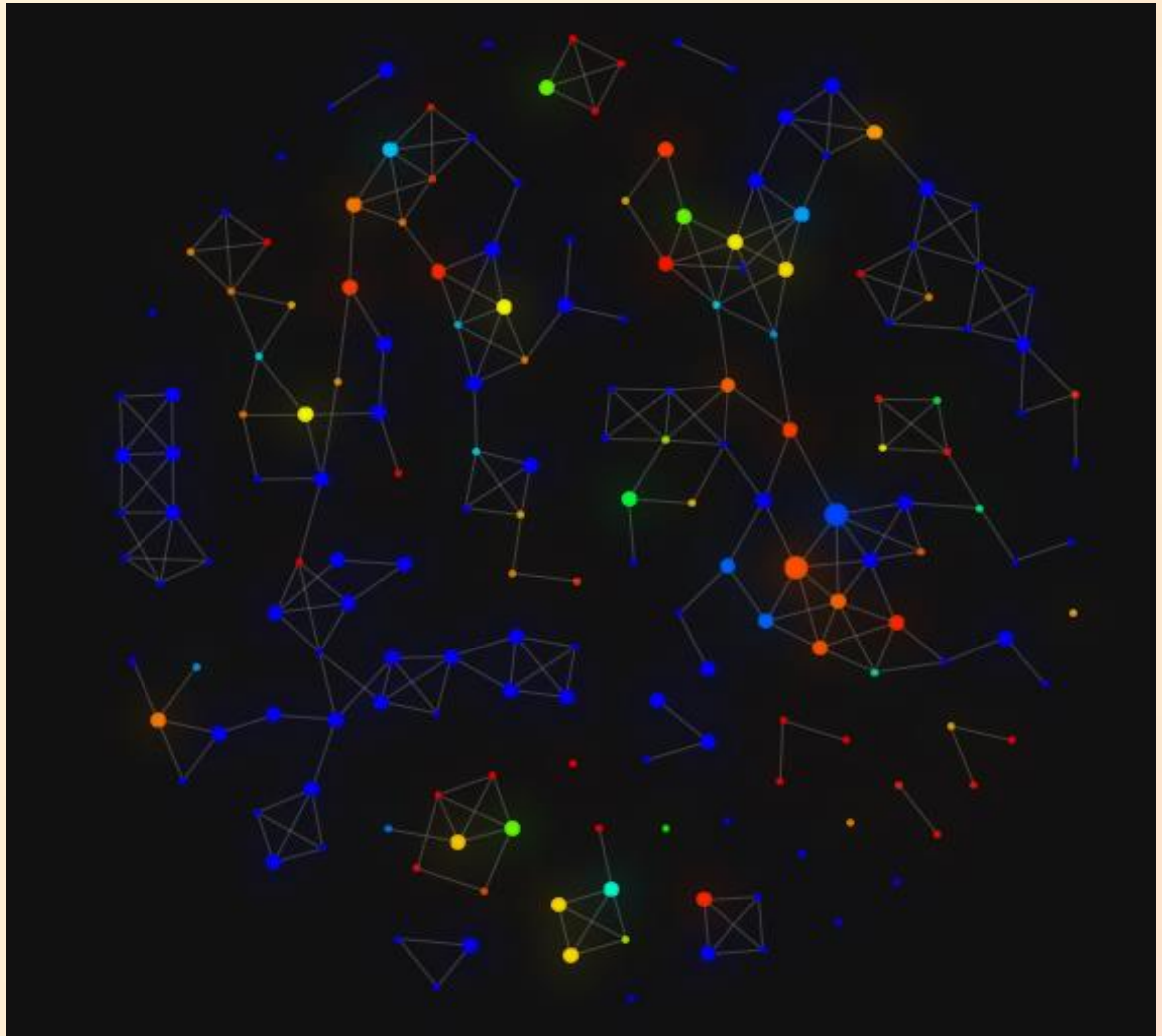
Plasmids Network

Overlap – 10%



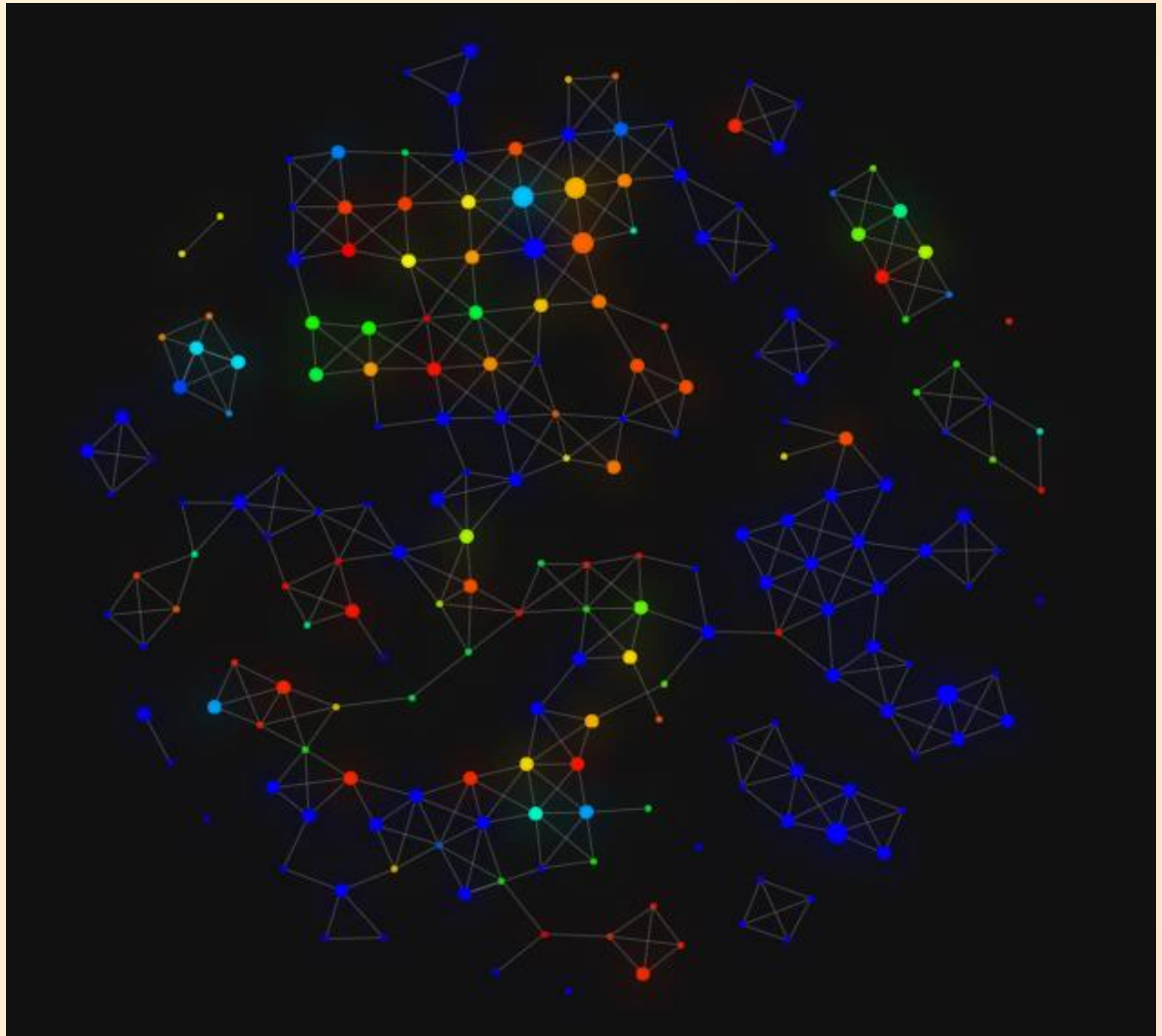
Plasmids Network

Overlap – 30%



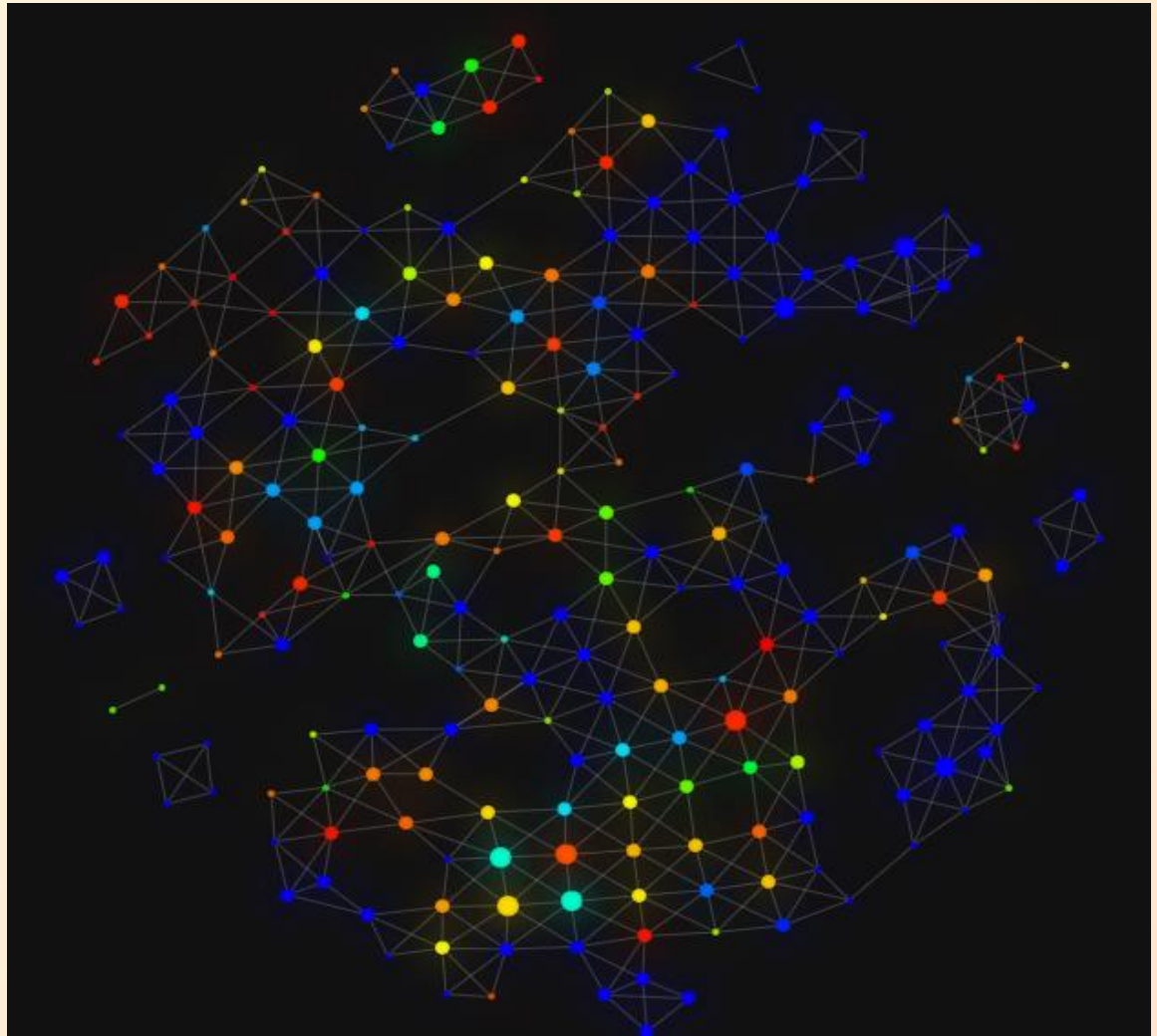
Plasmids Network

Overlap – 50%



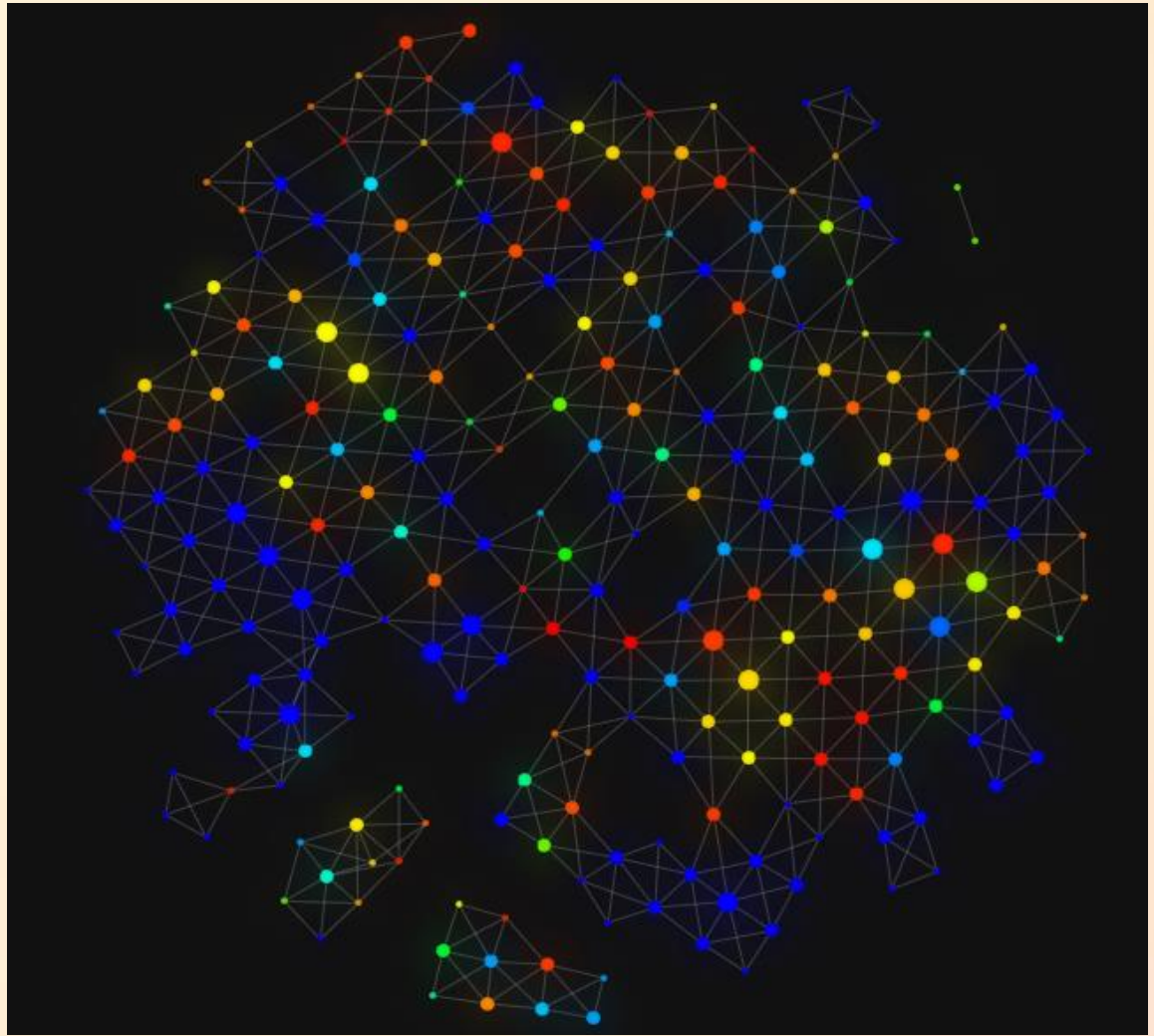
Plasmids Network

Overlap – 70%



Plasmids Network

Overlap – 90%



Other Mapper Software

- Mapper by Daniel Müllner
- Installation and the list of dependencies
  - http://danifold.net/mapper/installation/
- Website also contains Mapper documentation
- Nice GUI (show)
- More complex

Other Mapper Software

- R package “TDAmapper”
- A walkthrough and a tutorial by *Frederic Chazal and Bertrand Michel* at
  - http://www.lsta.upmc.fr/michelb/Enseignements/TDA/Mapper_solutions.html
- Familiar R environment
- Visualizations are somewhat limited (show)

References

1. Fasy, Brittany Terese, Jisu Kim, Fabrizio Lecci, and Clément Maria. "Introduction to the R package TDA." *arXiv preprint arXiv:1411.1830* (2014).

2. Kim, Jisu. "Tutorial on the R package TDA.“

3. Daniel Muller’s Mapper http://danifold.net/mapper/installation/

4. TDAmapper in R http://www.lsta.upmc.fr/michelb/Enseignements/TDA/Mapper_solutions.html

5. Python Mapper by MLWave https://github.com/MLWave/kepler-mapper

6. Ghrist, R., 2008. Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society, 45(1), pp.61-75.


Thank You!
Questions?