The Shape of Data: Machine Learning and Topology

Minnesota Developers Conference 2018

Kaisa Taipale

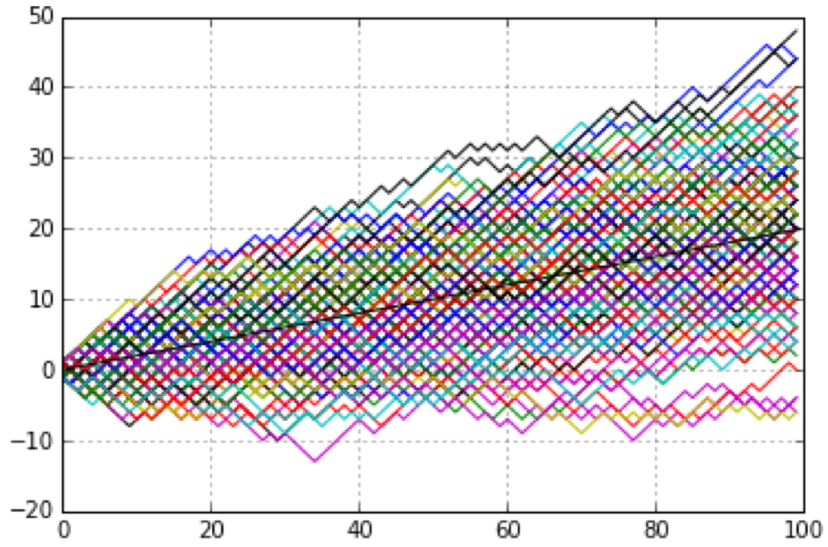
Disclosures

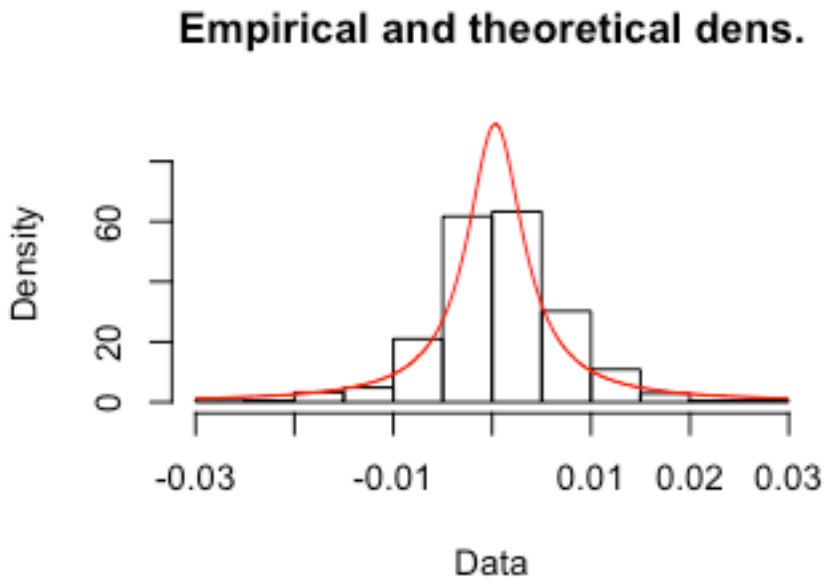
- Grew up in pure math
- I hack through code like someone lost in the Amazon with only a machete (it’s not pretty, but it gets me someplace!)
- Reproducibility matters to me because in working with students, non-reproducible code causes me personal pain
- Finance is interesting to me as a dynamical system and as an exploration of sociology

The plan:

- First, a high-level look at the context of applications of topology
- Next, geek out on pure math
- Last, put it into practice: packages and tools

Math finance, statistics…

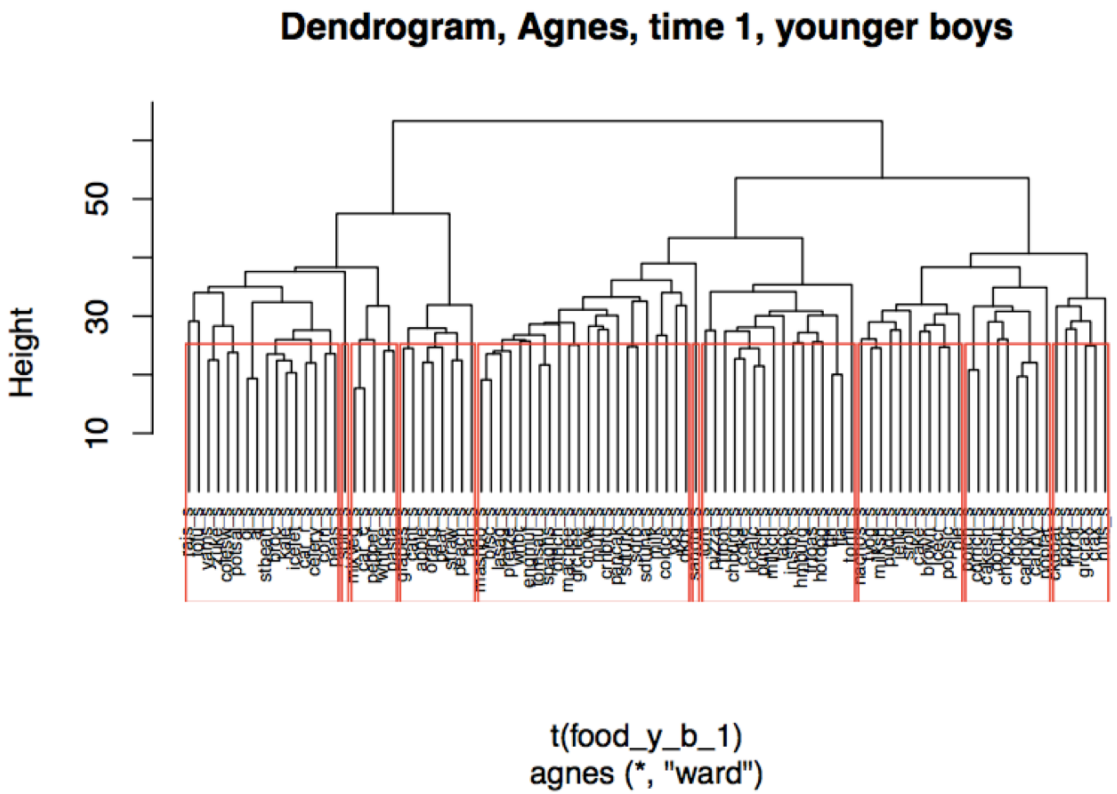


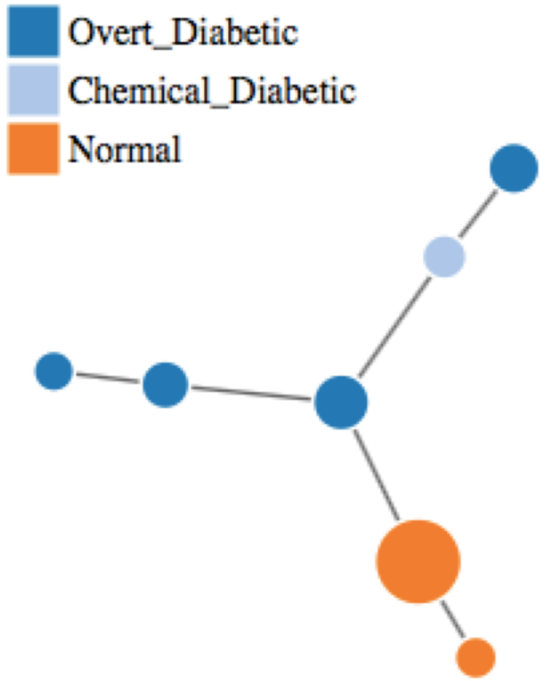


Financial math: Black-Scholes, binomial trees, time series, stochastic modeling

Statistics: linear regression, p-values, distributions

….machine learning… topology?

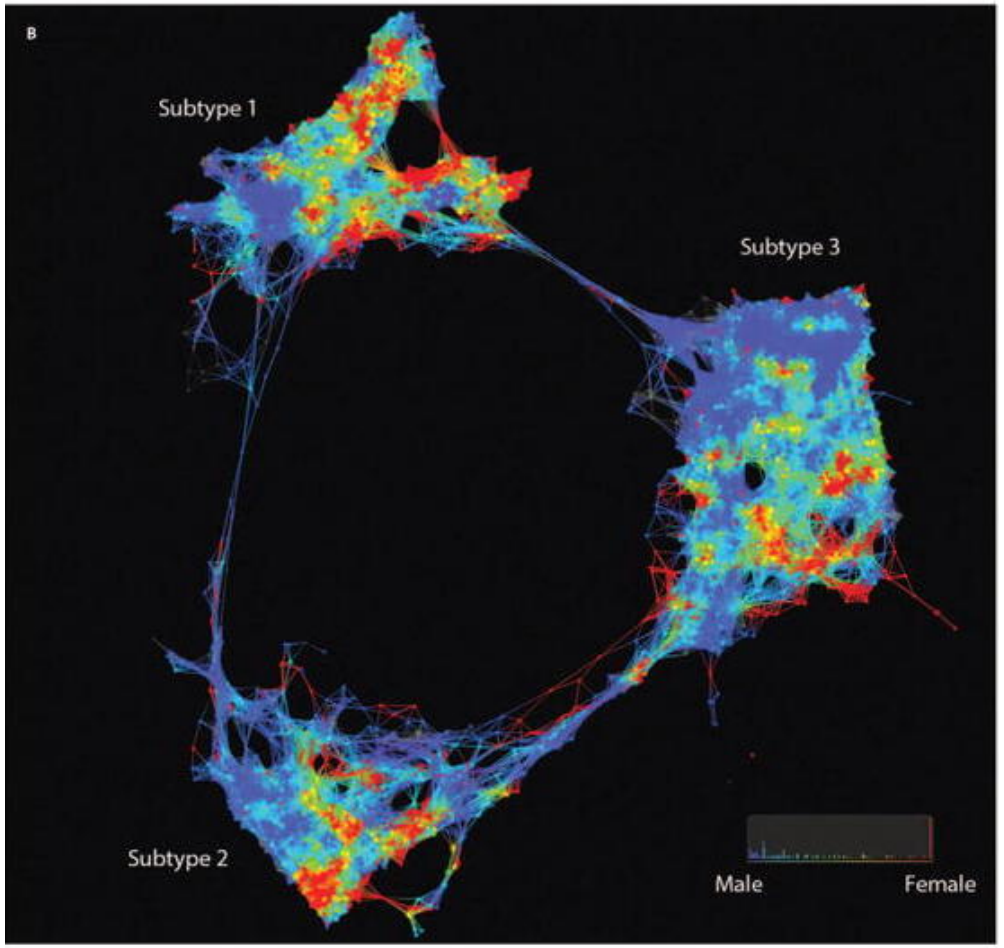




Machine learning: Neural networks, clustering, manifold learning

Topology?! Shape of data, for feature discovery and interpolation between clusters and manifold learning

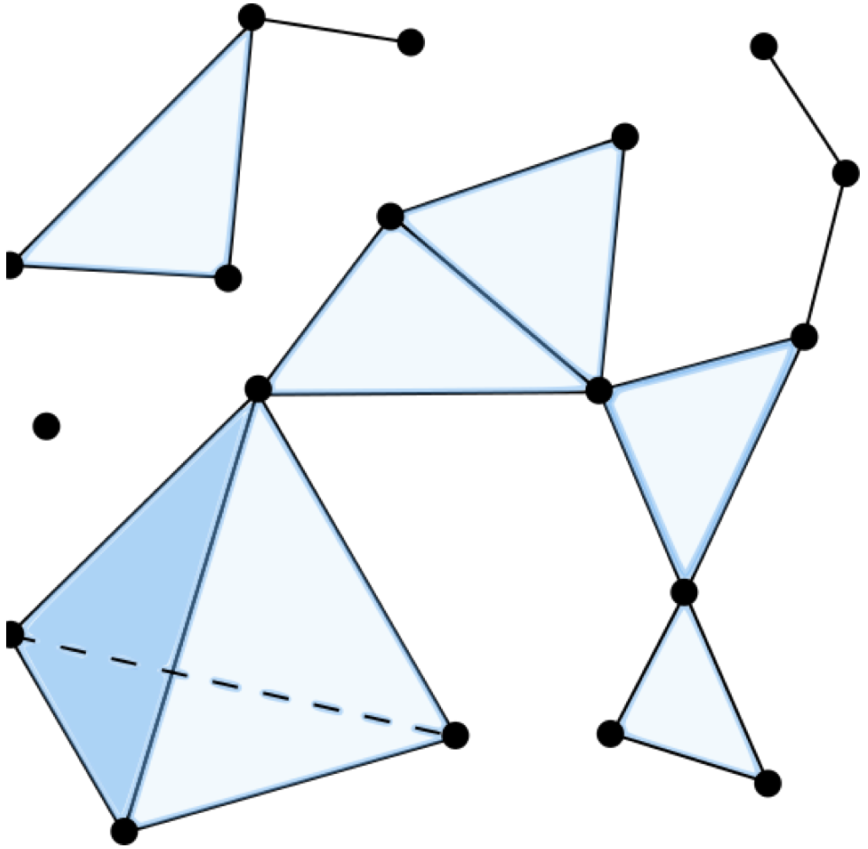
Finding use in diabetes research…



Identification of type 2 diabetes subgroups through topological analysis of patient similarity
Li Li,[1] Wei-Yi Cheng,[1] Benjamin S. Glicksberg,[1] Omri Gottesman,[2] Ronald Tamler,[3] Rong Chen,[1] Erwin P. Bottinger,[2] and Joel T. Dudley[1,4,*]

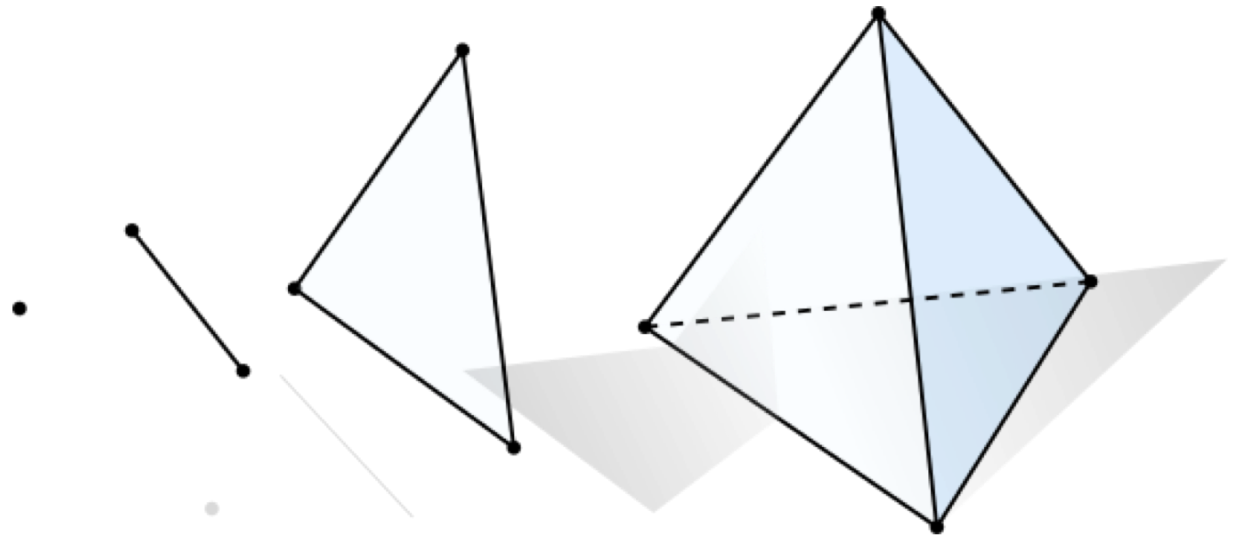
What’s topology?

Simplicial complexes



By cflm (talk) - Own work. Derived from File:Simplicial complex example.png by Trevorgoodchild(en.wp), released under PD-self. Coloured using Inkscape., Public Domain, https://commons.wikimedia.org/w/index.php?curid=7937755

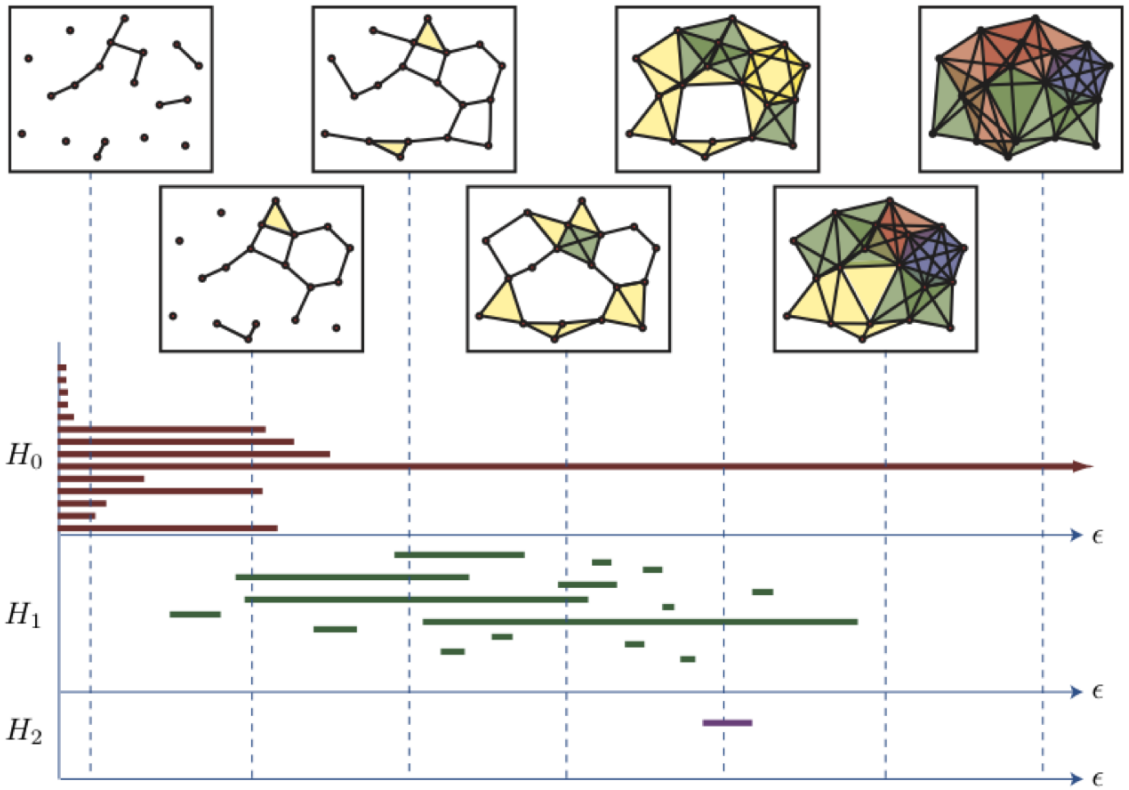
What are simplices?



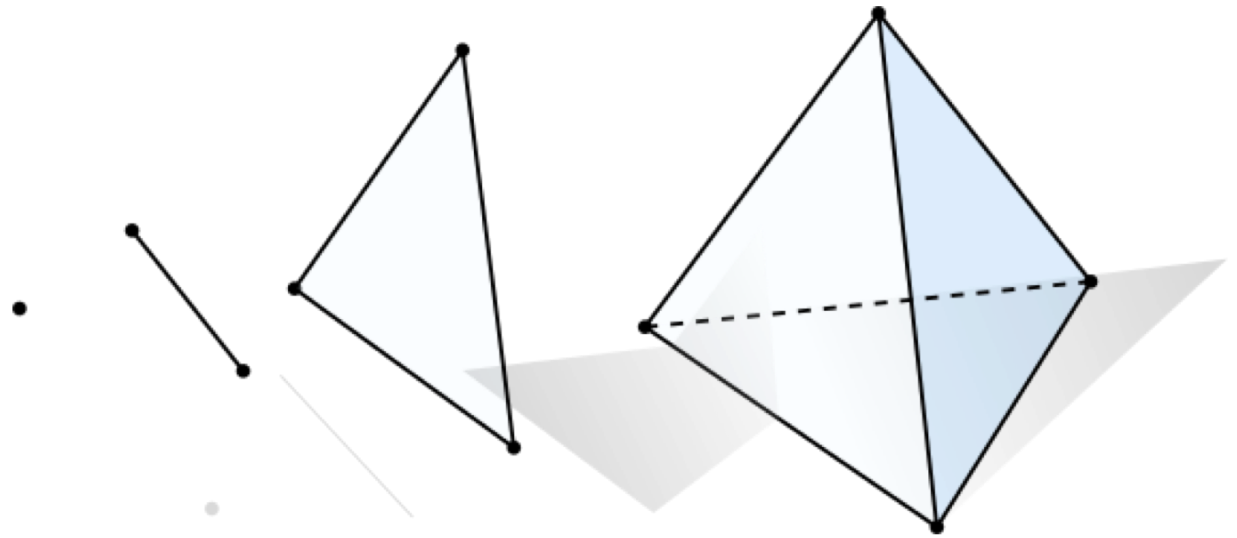
Persistent homology

Persistent homology: what topological features *persist* as we vary the cutoff parameter?

(Image from BARCODES: THE PERSISTENT TOPOLOGY OF DATA by Robert Ghrist)



Betti numbers: just count simplices



Big topology ideas

I'll talk about four major ideas from the TDA toolbox and then apply them to finance examples.

- Super-level sets
- Persistent homology
- The Mapper algorithm for visualization
- Betti numbers

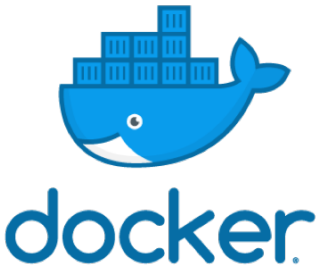
Big topology ideas

I'll talk about four major ideas from the TDA toolbox and then apply them to finance examples.

- Super-level sets (look at graph from $\delta > c$)
- Persistent homology (look at shapes that persist as δ varies)
- The Mapper algorithm for visualization (slice and cluster, then build a simpler graph)
- Betti numbers (counting number of simplices in each dimension)

Tools

- Python
  - Kepler-mapper, moguTDA, …
- R
  - TDA, TDAmapper, Igraph, NetworkD3…
- Docker, Git, etc



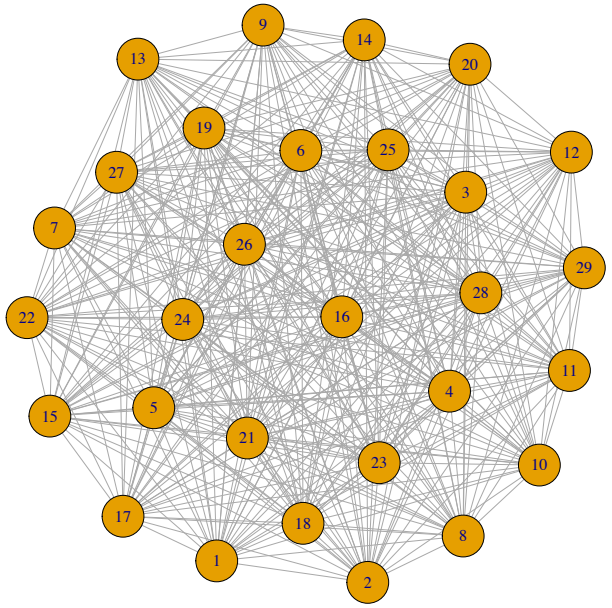
- Gigantum?
- Julia: Eirene (fast!)

How do we use topology on data?

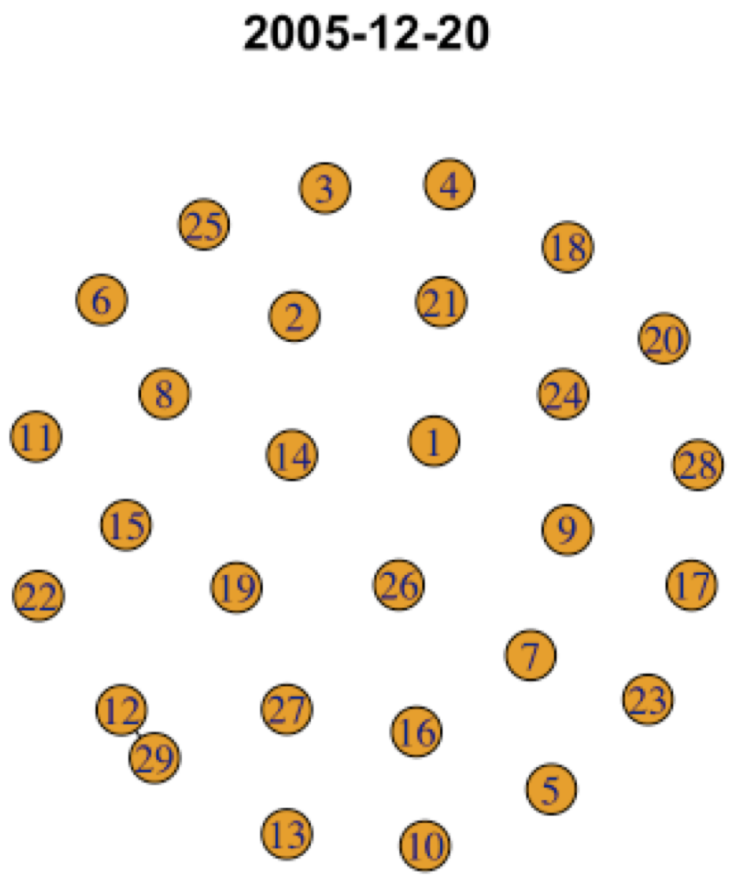
- Data generally comes as a point cloud (a set of points), ideally in a csv file.
- Load the data
- Specify a metric (a way to quantify “nearness”) – this is a choice
- Build simplices on the data set
- Analyze the topology of the simplicial complexes

Dow Jones stocks: correlation network in R

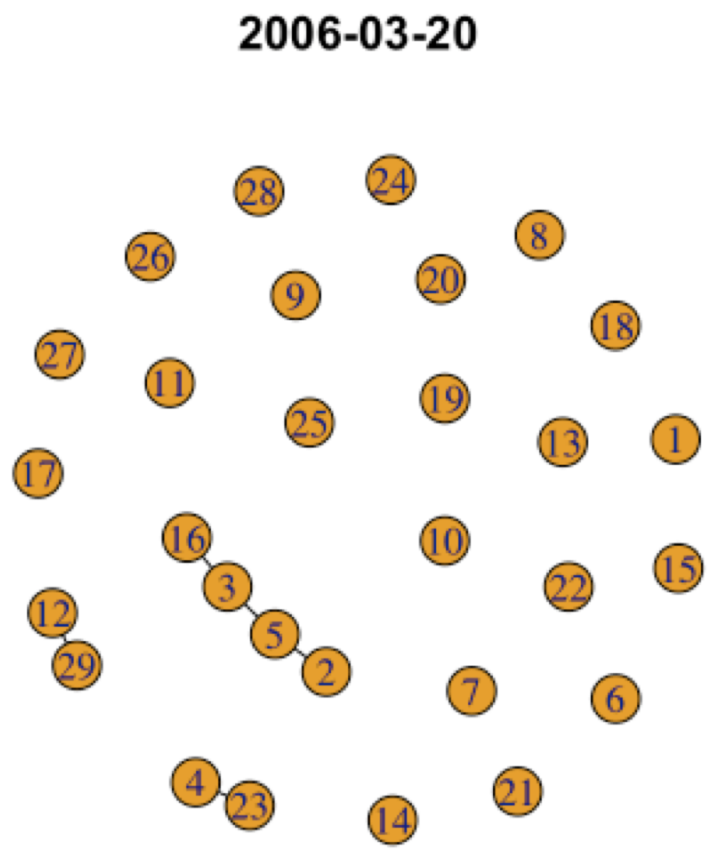
Transform correlation and build correlation networks. First, all edges :



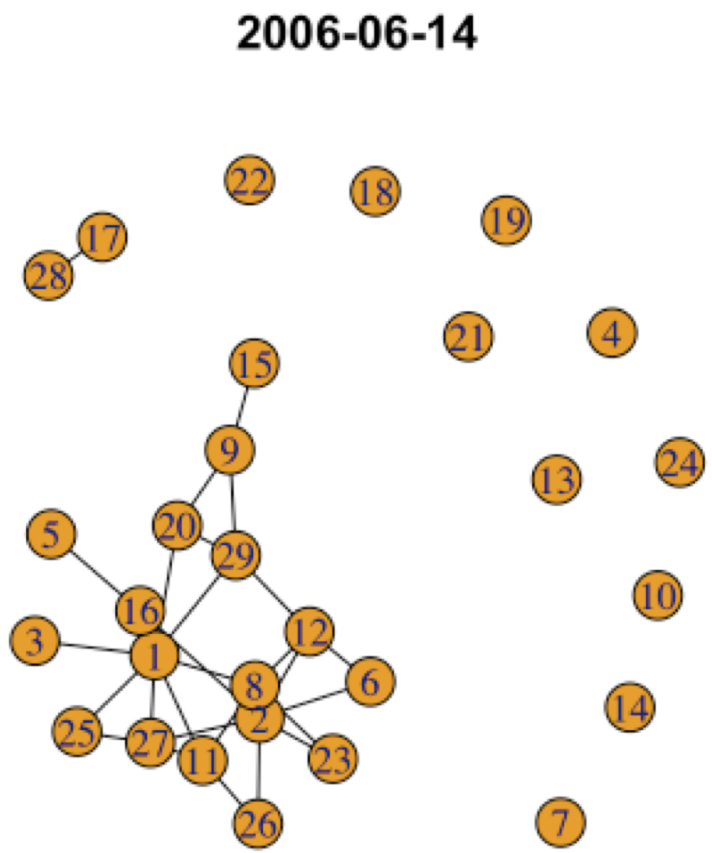
Dow Jones stocks: superlevel sets in R



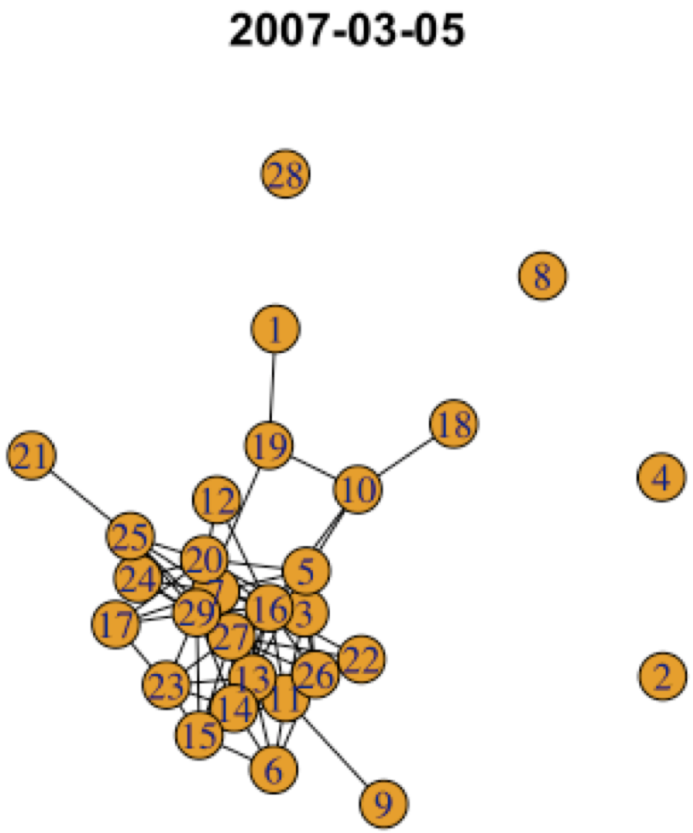
Dow Jones stocks: superlevel sets in R



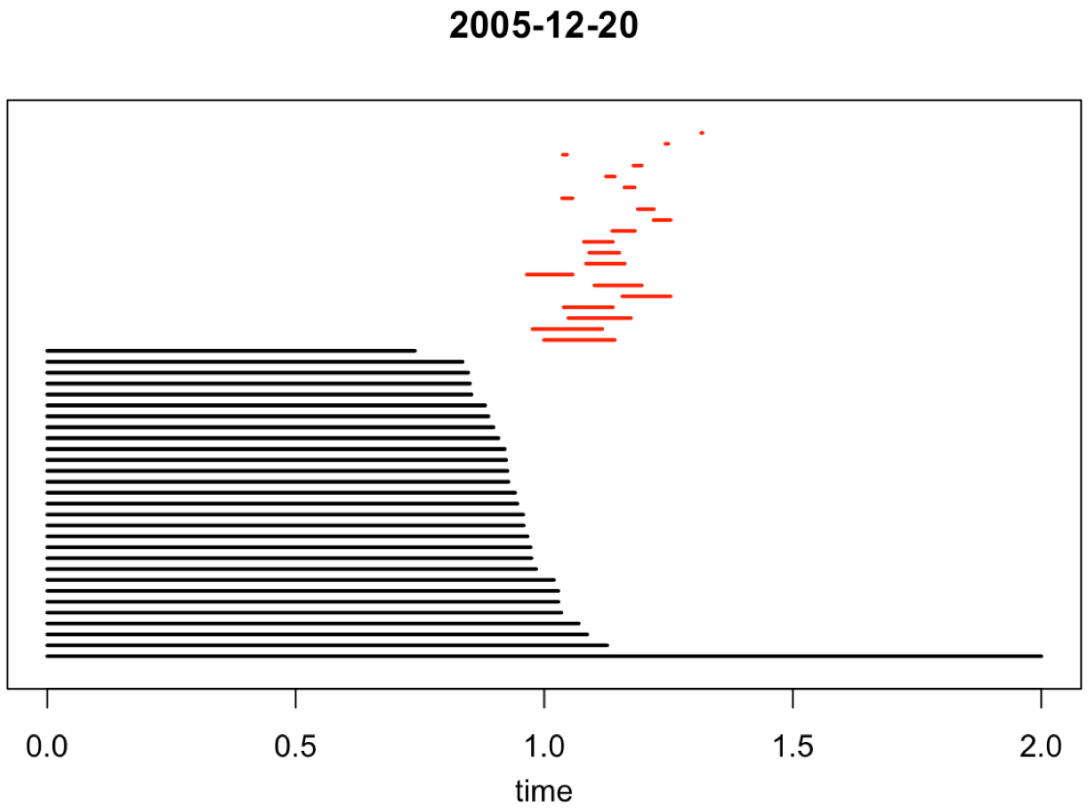
Dow Jones stocks: superlevel sets in R



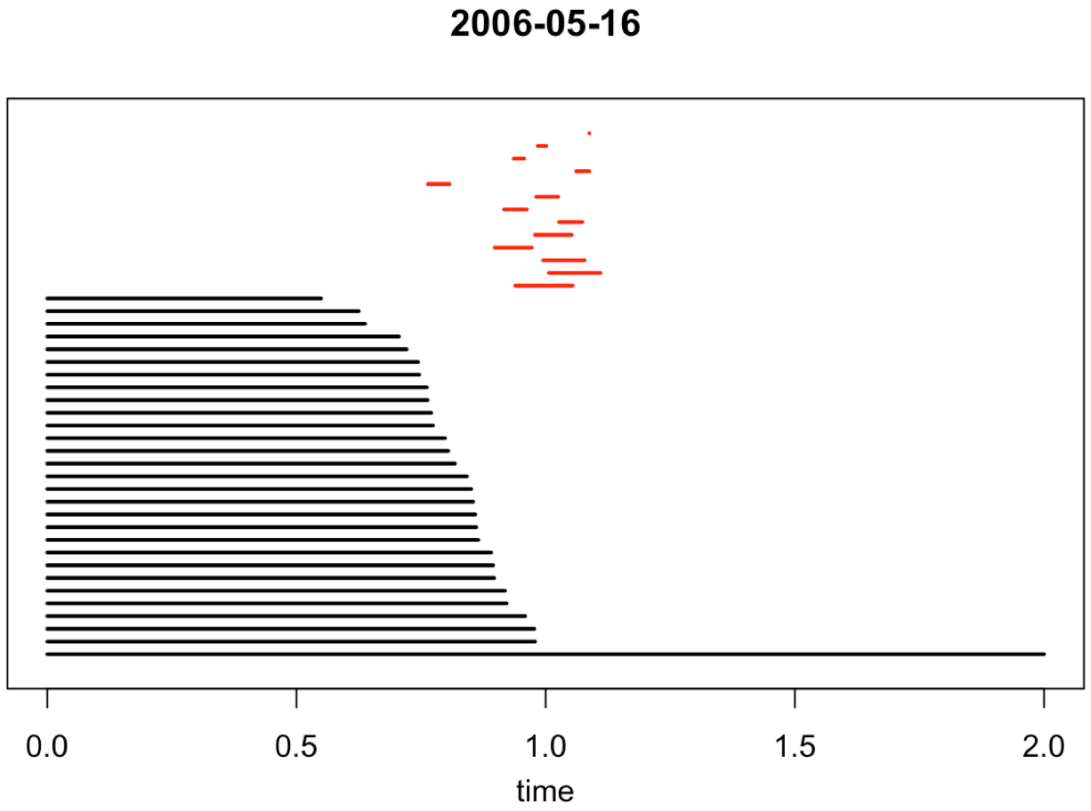
Dow Jones stocks: superlevel sets in R



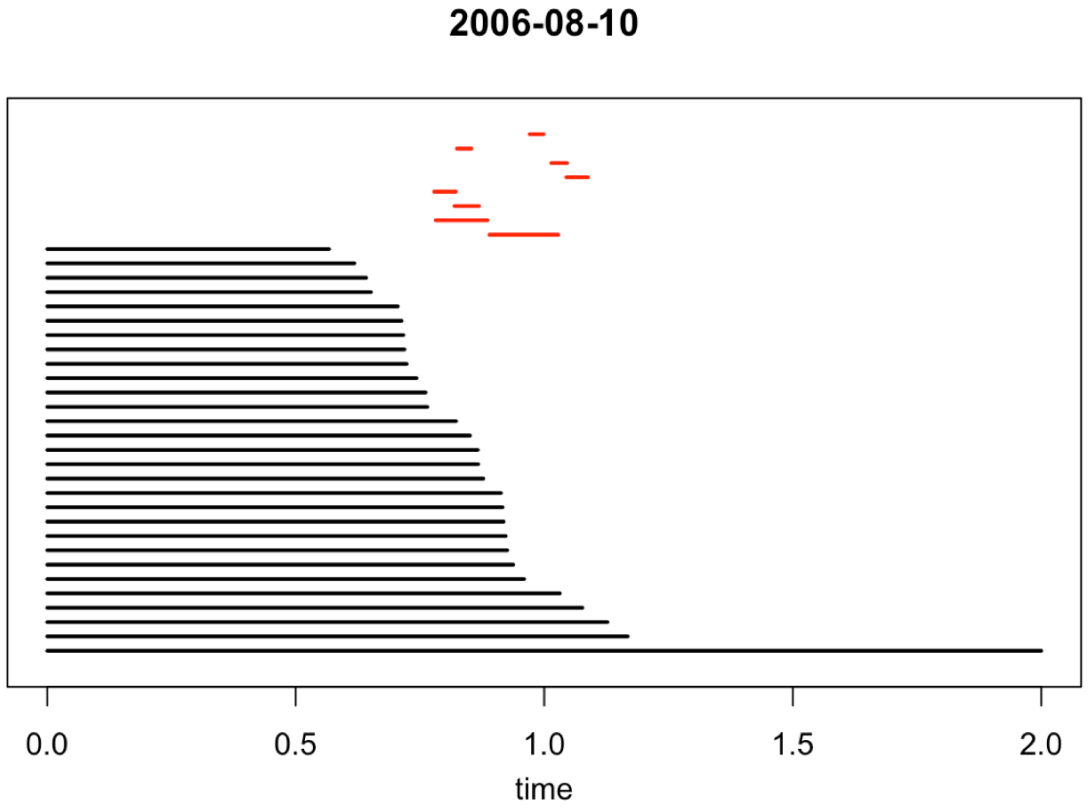
Dow Jones stocks: persistent homology



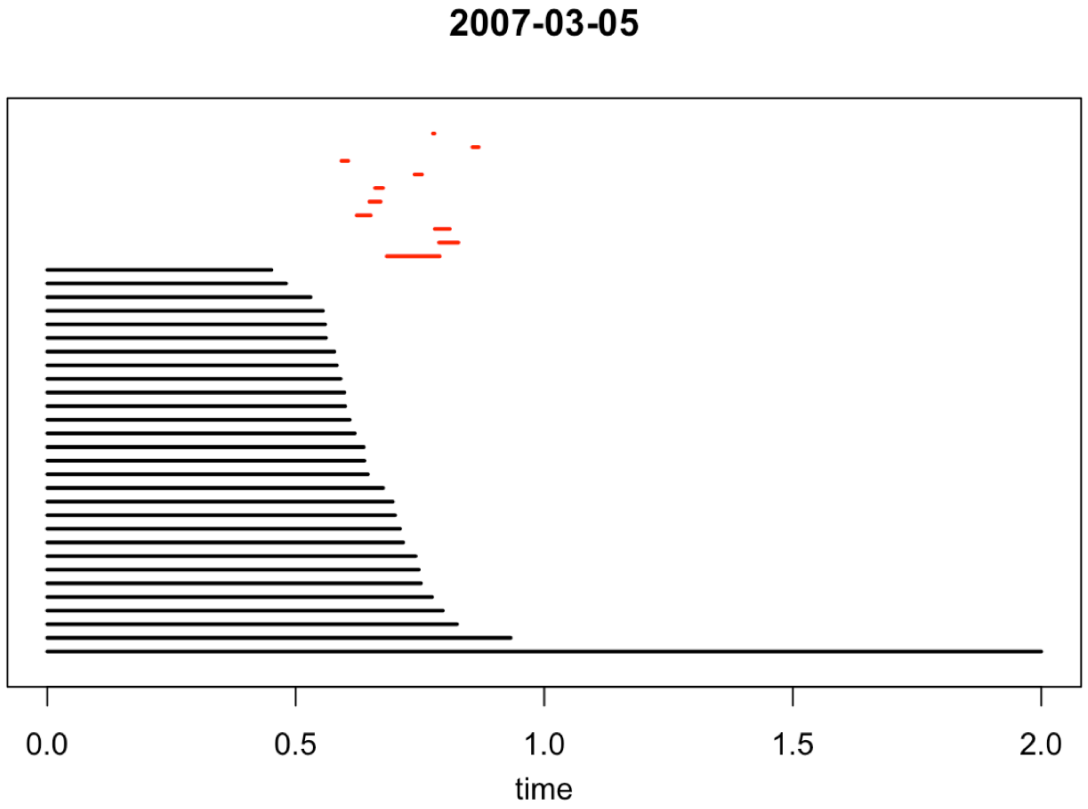
Dow Jones stocks: persistent homology



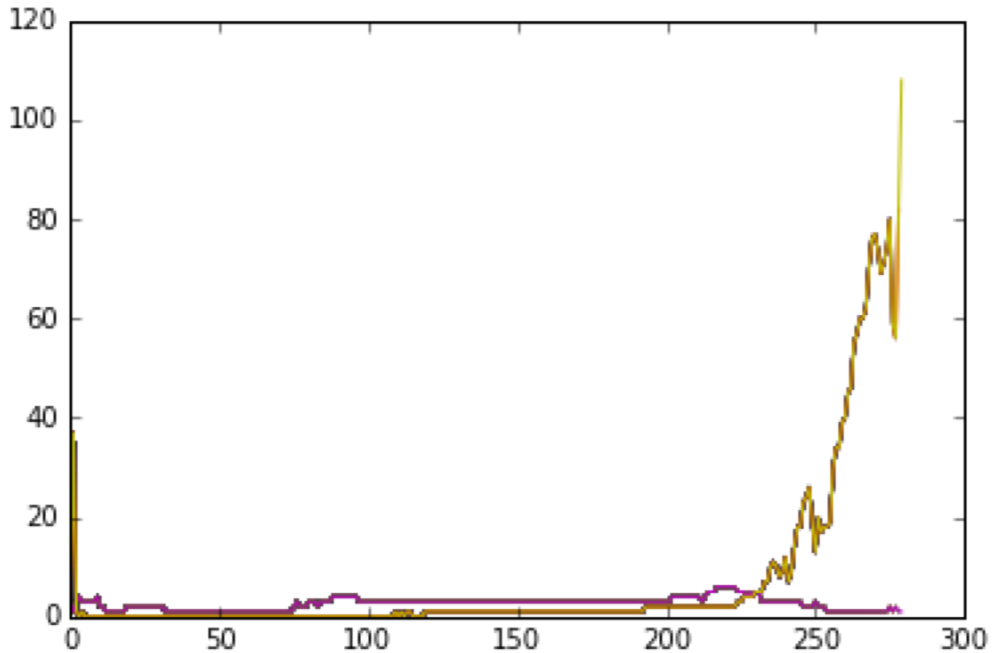
Dow Jones stocks: persistent homology

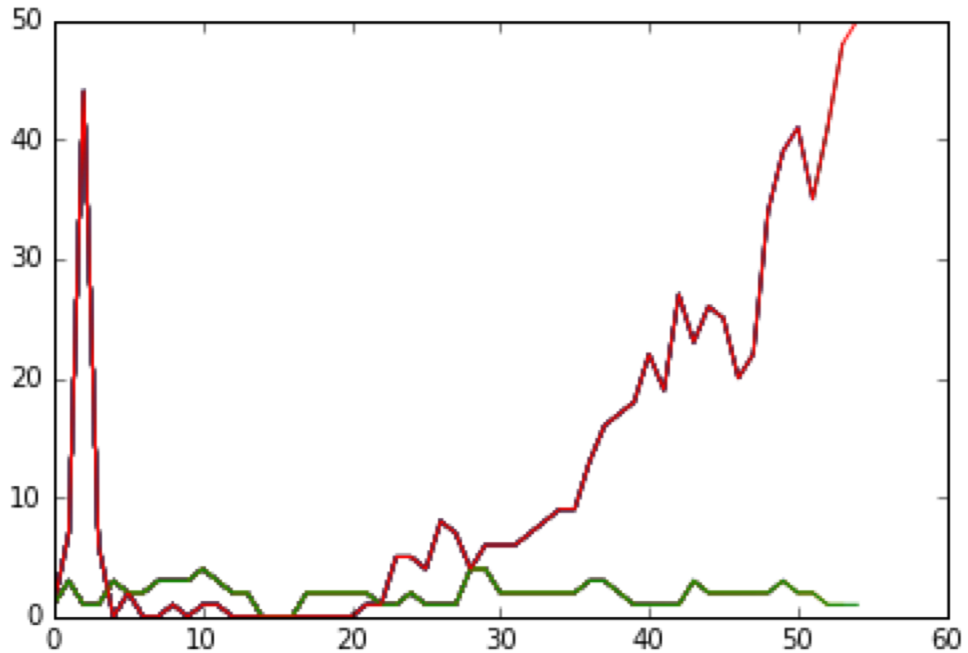


Dow Jones stocks: persistent homology

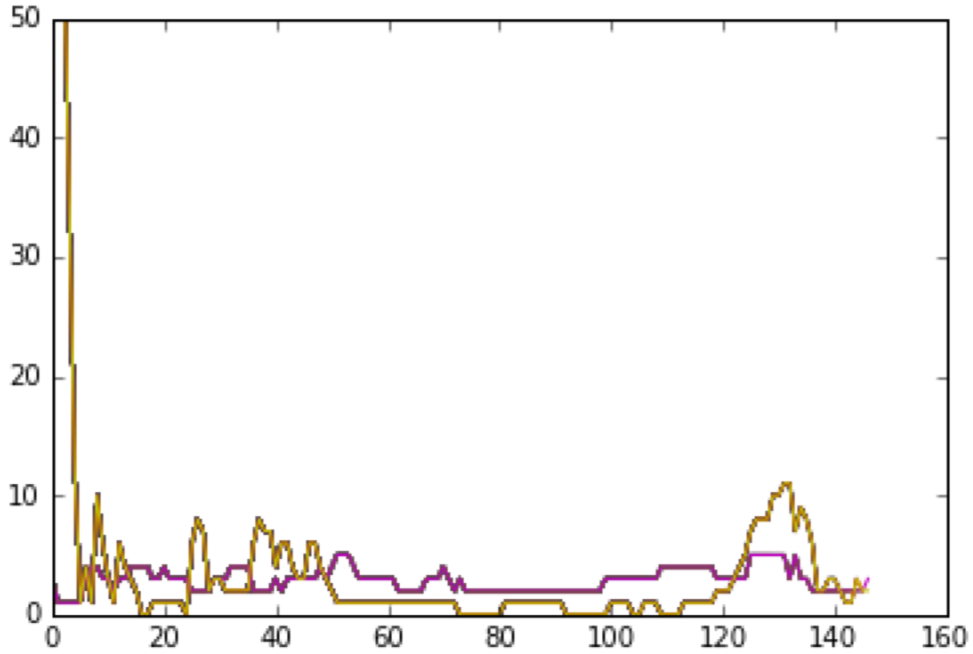


Dow Jones: Betti numbers (Python)





Dow Jones Betti numbers today (Python)



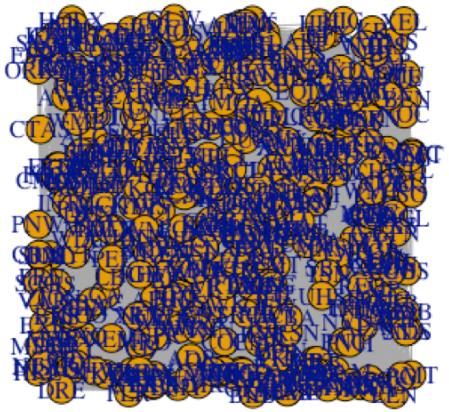
The bump is in January 2018.

S&P 500: data issues

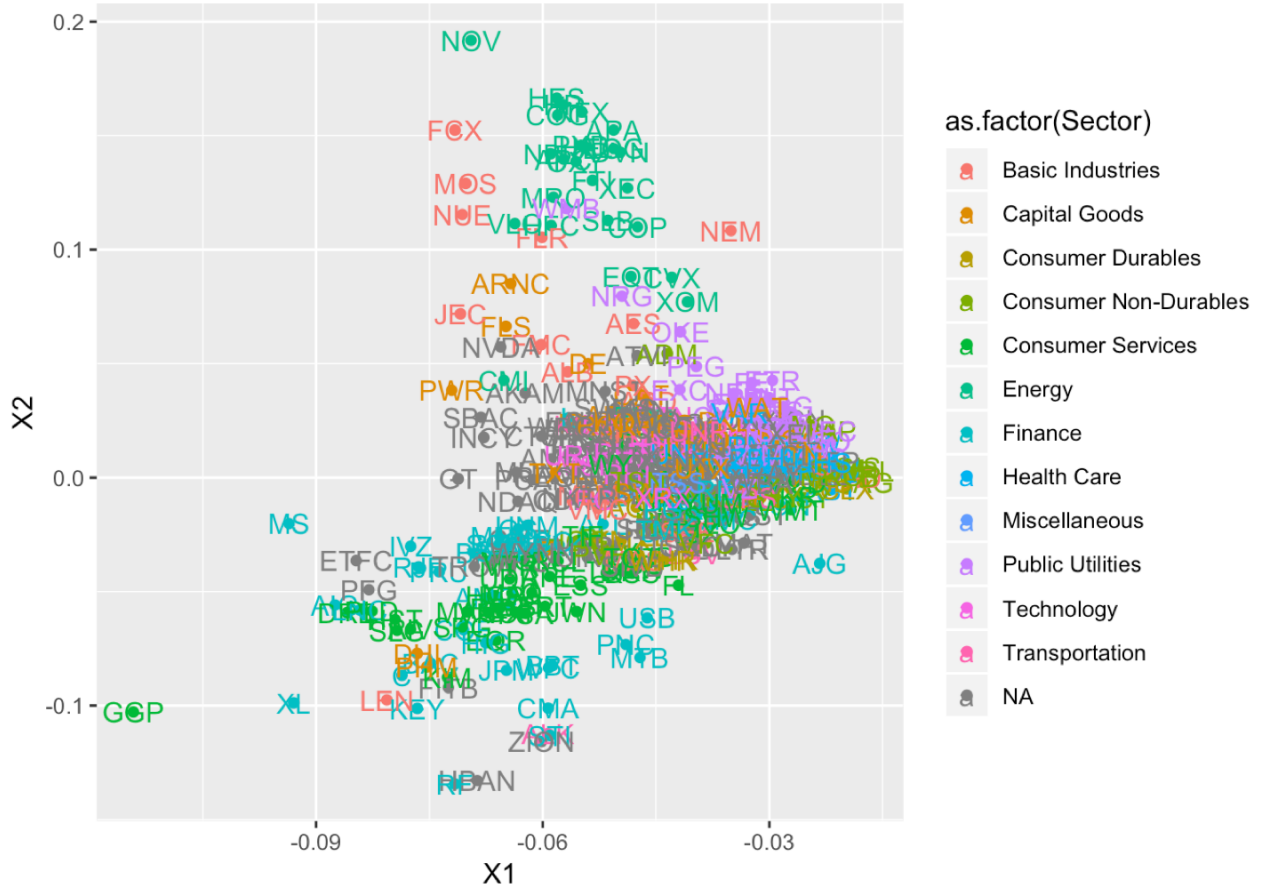
- Stocks enter and leave the S&P 500 based on company characteristics
- By definition, the S&P represents only a particular view of the equities market!
- Problems with data continuity: mergers, bankruptcies, etc.

S&P 500: superlevel sets

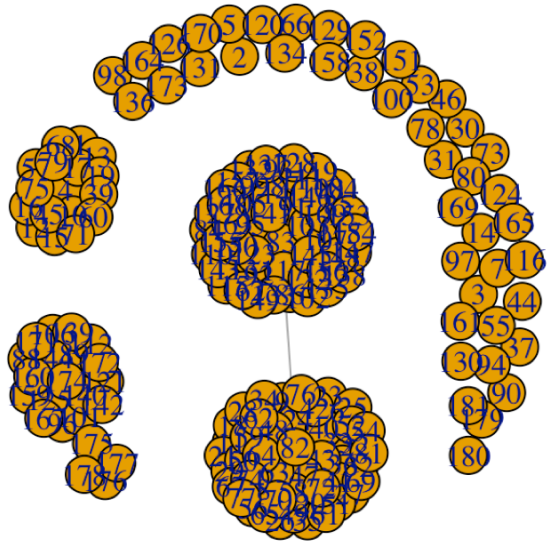
Just a mess!!! Very hard to interpret.



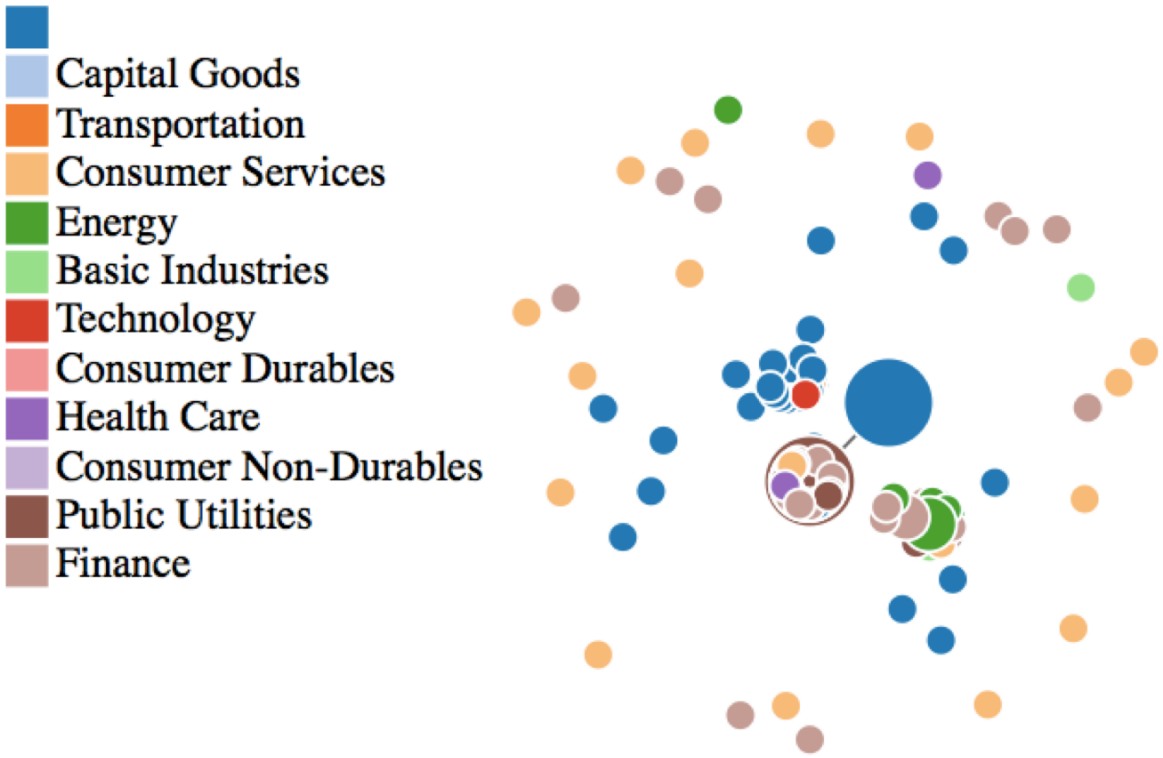
S&P 500: singular value decomposition



S&P 500: Mapper (iGraph in R)



S&P 500: Mapper (networkD3 in R)



Thank you!

- *Thank you, Minnesota Developers Conference audience and organizers!*

- Thank you to the MCFAM summer seminar participants: Jialing Cai, Yunpeng Liu, Ayman Ahmed, John Burbidge, Fiona Jiang, Ziqi Dong, John Nguyen, Zhongwu Wang, Ayush Bansal, Ameya Phadke, Ziran Xu, Heng, Qinzheng Xu, Yin Xu, Doreen Vescelius, Yifan Xu, Bo Zhu, Jianfeng Liu….

References

News/pop science:

- https://www.technologyreview.com/s/602234/how-the-mathematics-of-algebraic-topology-is-revolutionizing-brain-science/
- https://www.wired.com/story/the-mind-boggling-math-that-maybe-mapped-the-brain-in-11-dimensions/

Academic:

- A talk with some different applications – mode discovery, image analysis, http://www.sci.utah.edu/~beiwang/acmbcbworkshop2016/slides/ChaoChen.pdf
- Robert Ghrist has written great mathy notes! https://www.math.upenn.edu/~ghrist/
- Kathryn Hess is a mathematician working on neuroscience problems – mentioned in Wired article above – see technical talk at https://www.youtube.com/watch?v=vD27zKxoio0&index=6&list=PL4kY-dS_mSmJ4DU2OmOUWB8QIN5nG0CMv
- Marian Gidea, some of the first public applications to finance: https://arxiv.org/abs/1701.06081
- Some of the founders of the field are Gunnar Carlsson, Gurjeet Singh, Afra Zomorodian – look for papers with their names.

Follow me if you want to see upcoming applications to finance!

- http://www.kaistataipale.net/blog or http://www-users.math.umn.edu/~taipale/