
AIDC
INTEL AI DEVCON 2018

TOPOLOGICAL DATA ANALYSIS WITH PERSISTENT HOMOLOGY

Bryn Keller

24-May-2018

ABSTRACTION AND GENERALIZATION

... a necessary transformational development that we can expect in the field of machine learning is a move away from models that perform purely *pattern recognition* and can only achieve *local generalization*, towards models capable of *abstraction* and *reasoning*, that can achieve *extreme generalization.*

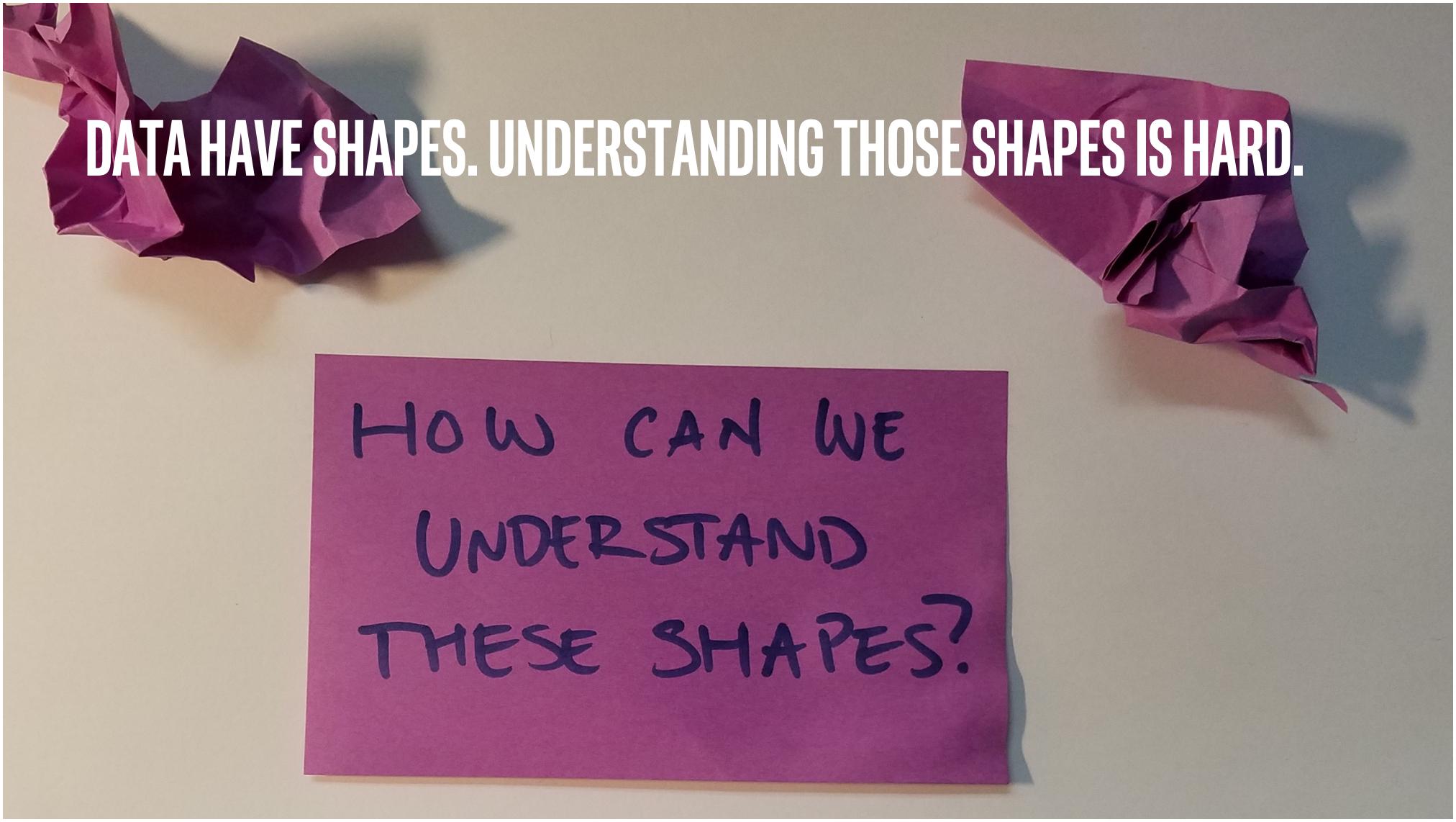
https://blog.keras.io/the-future-of-deep-learning.html Francis Chollet

ABSTRACTION AND GENERALIZATION

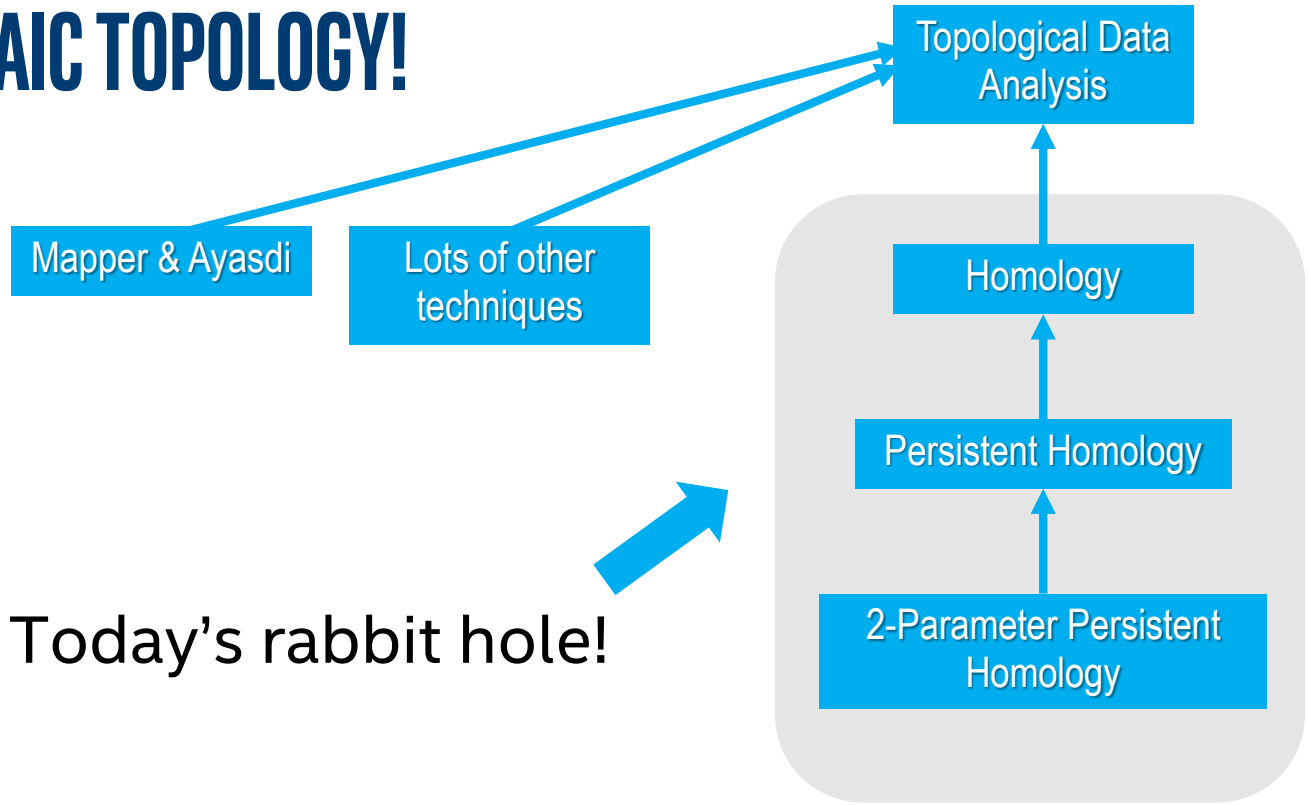
... a necessary transformational development that we can expect in the field of machine learning is a move away from models that perform purely *pattern recognition* and can only achieve *local generalization*, towards models capable of *abstraction* and *reasoning*, that can achieve *extreme generalization*.

https://blog.keras.io/the-future-of-deep-learning.html

Francis Chollet

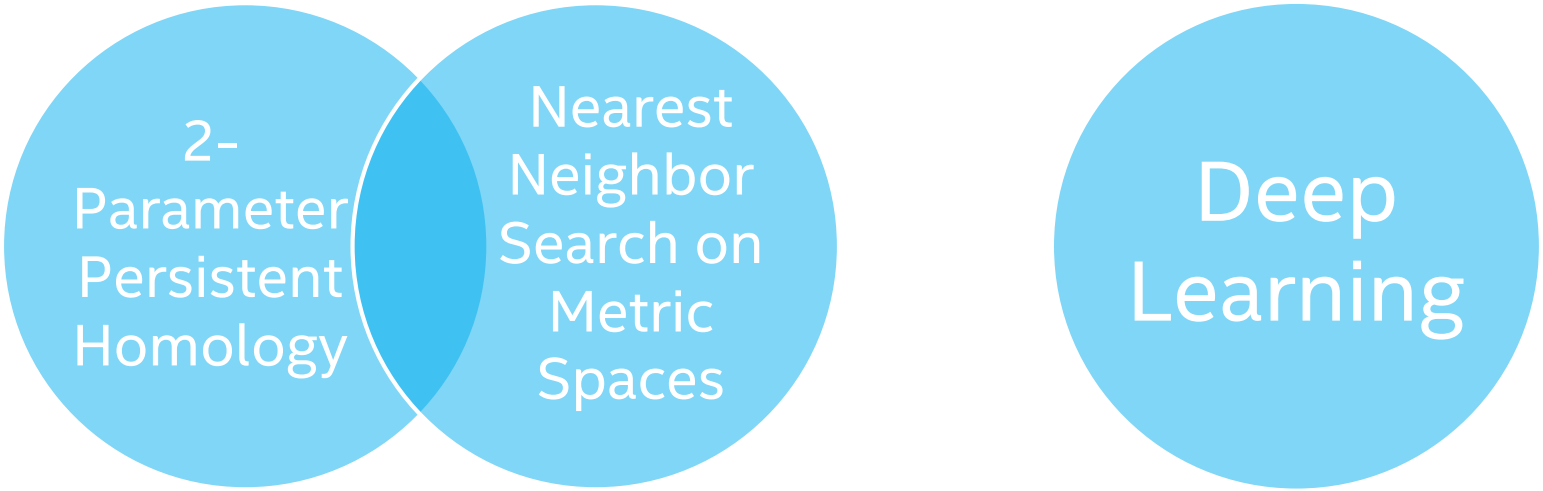

DATA HAVE SHAPES. UNDERSTANDING THOSE SHAPES IS HARD.
HOW CAN WE UNDERSTAND THESE SHAPES?

ALGEBRAIC TOPOLOGY!



Today's rabbit hole!

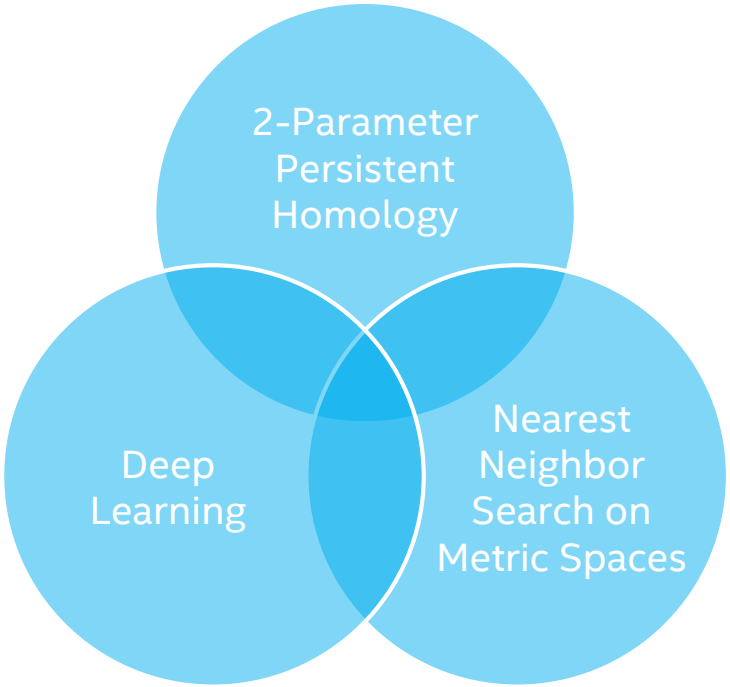
MY RECENT WORK



Search for new medicines

Humpback whale health assessment

AN IDEA

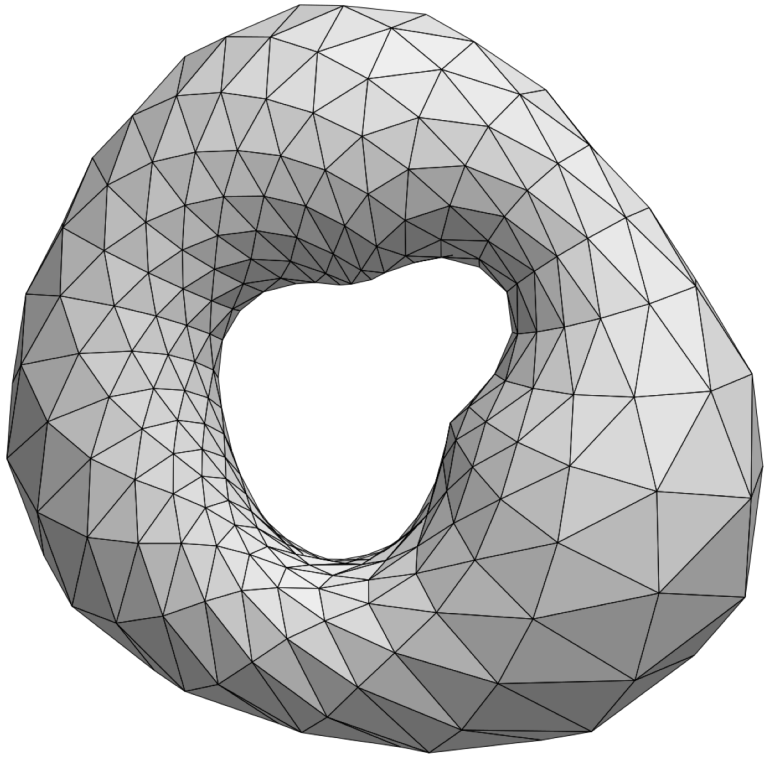


Even better search for new medicines?

HOMOLOGY

Power tool of modern mathematics

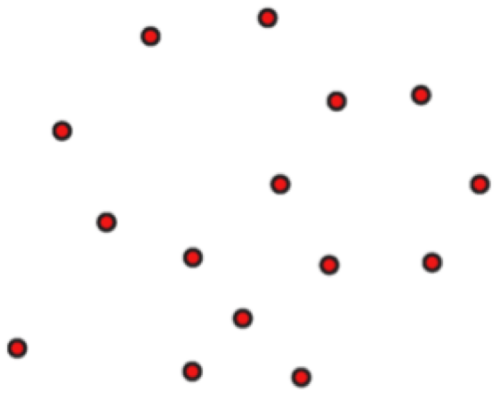
ORIGINAL SHAPE



Fashionslide at English Wikipedia

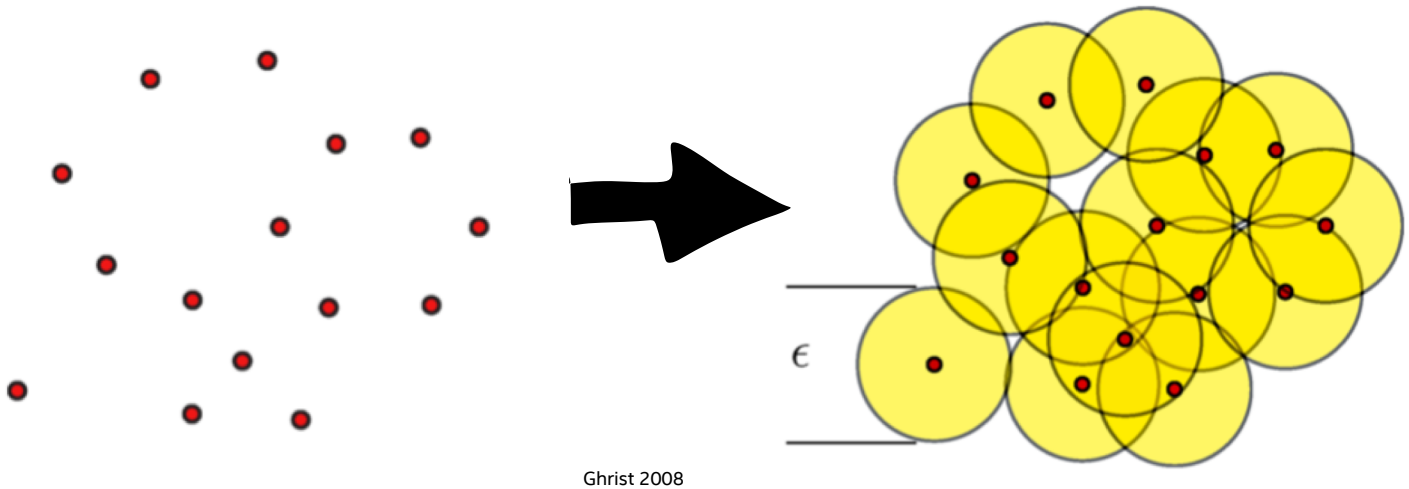
AIDC
INTEL AI DEVCON 2018

NOISY SAMPLES FROM THAT SHAPE



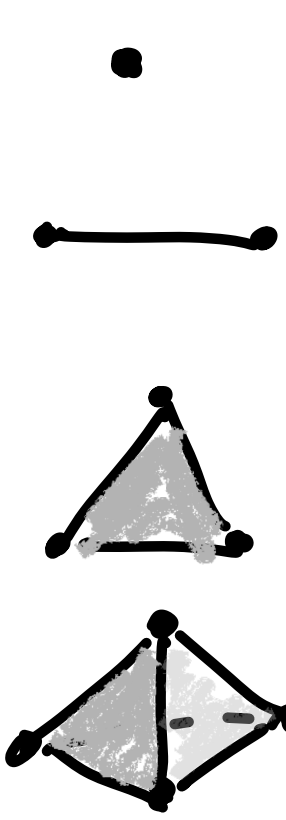
Ghrist 2008

PICK A DISTANCE ϵ, MAKE ϵ-BALLS AROUND THE POINTS



Ghrist 2008

SIMPLICES (GENERALIZED VERTICES & EDGES)



0 - SIMPLEX

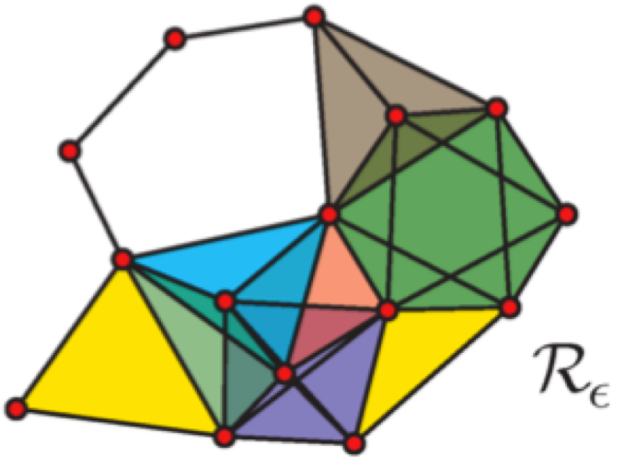
1 - SIMPLEX

2 - SIMPLEX

3 - SIMPLEX

AND SO ON FOR n - SIMPLICES

CAN WE RECOVER A TOPOLOGICALLY EQUIVALENT SHAPE?



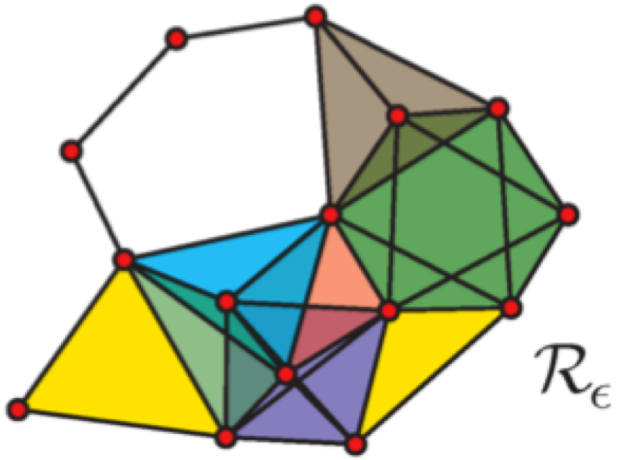
Ghrist 2008

A simplicial complex!

BETTI NUMBERS: A CODE FOR COMPARING SHAPES

- β_0 counts **connected components** - zero dimensional "holes": solid objects
- β_1 counts **tunnels**, or holes you can poke a finger through
- β_2 counts **voids** or bubbles
- β_n counts n-dimensional holes generally

A SIMPLE DESCRIPTION



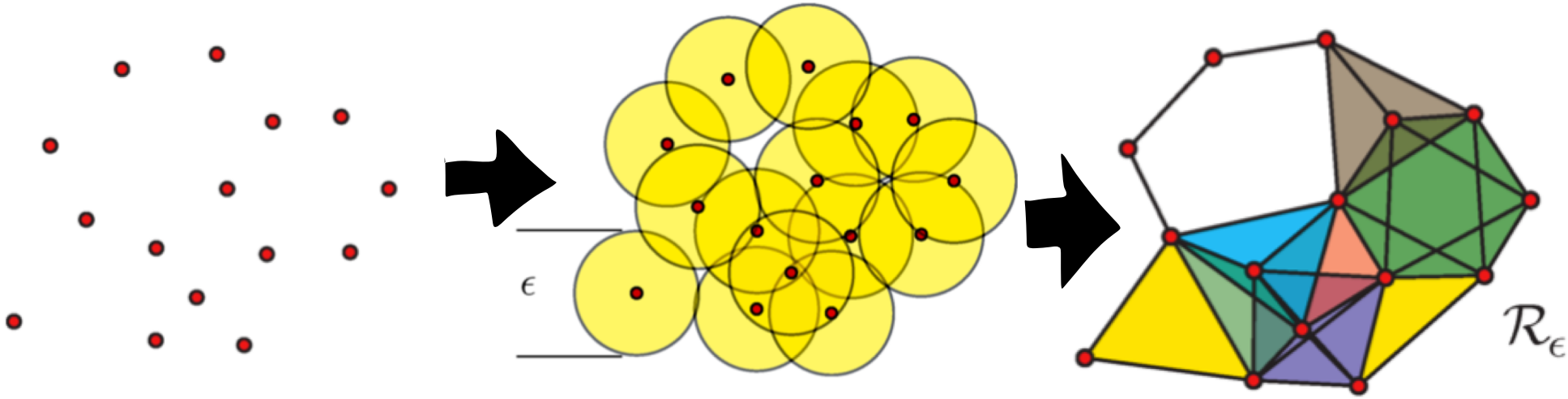
Ghrist 2008

β_0 (connected components): 1

β_1 (tunnels): 1

β_2 (voids): 0

HOMOLOGY: MAKE A COMPLEX OUT OF DATA, COUNT THE HOLES

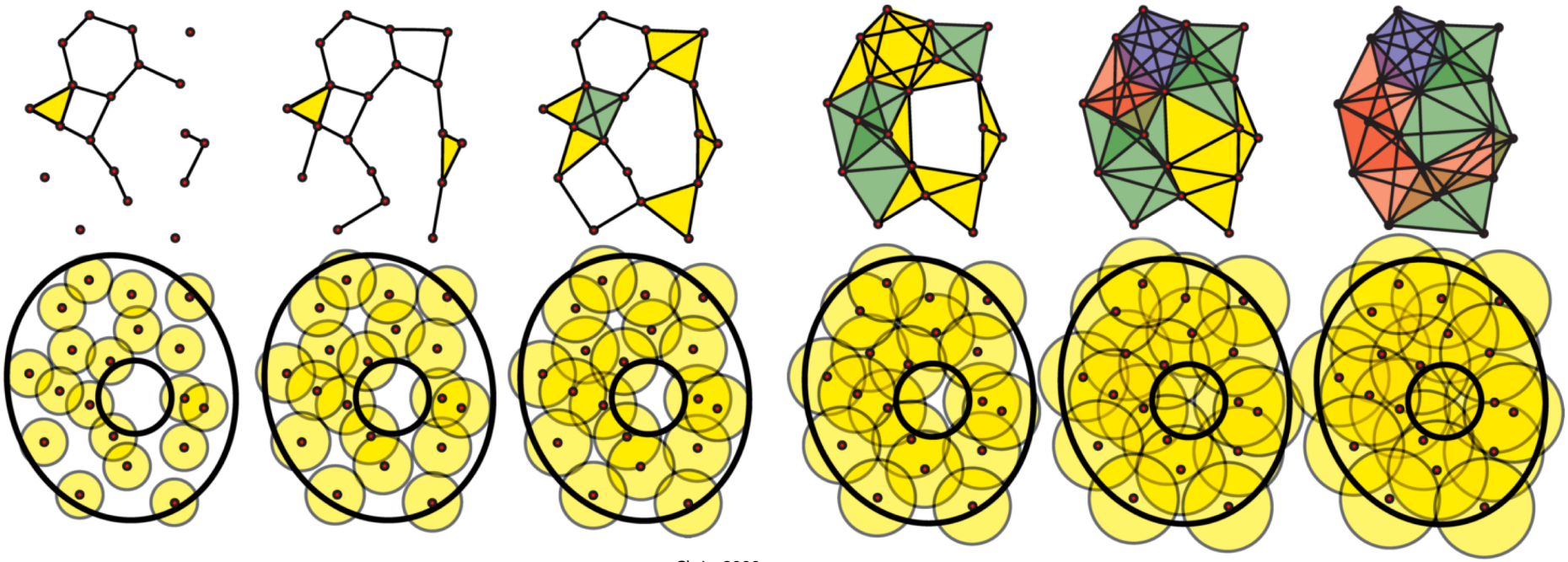


Ghrist 2008

PERSISTENT HOMOLOGY

It’s all a question of scale

WHICH ϵ SHOULD WE PICK?



Ghrist 2008

INCREASING ϵ

VISUALIZE VR-COMPLEX BUILDING

Vietoris-Rips Complex

Radius: 10

Vertices: 35

Edges: 0

Triangles: 0

Tetrahedrons: 0

Homology group

β_0 (connected components): 34

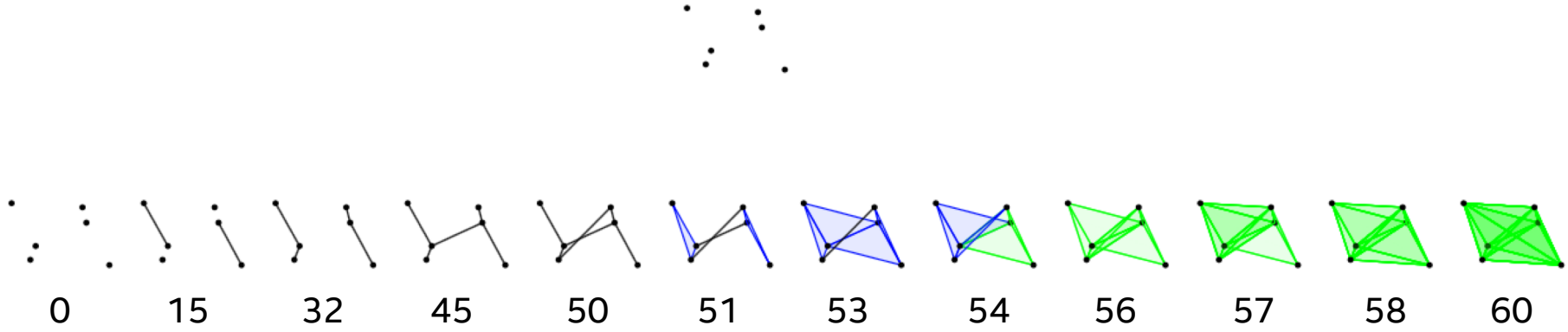
β_1 (holes): 0

Shadow

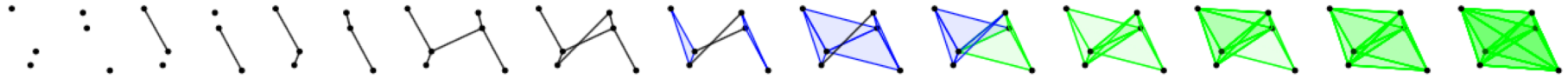
BIG IDEA OF PERSISTENT HOMOLOGY

- Don't try to *pick the right value* for ϵ, try them all!
  - as the parameter changes, holes will come and go
  - long lasting holes are the ones we should care about
- So we build a *collection* of simplicial complexes
  - one complex for each value of ϵ where we get different results
  - this collection is called a ***filtration***

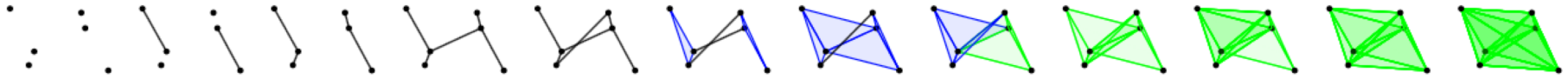
FILTRATION



CONNECTED COMPONENTS

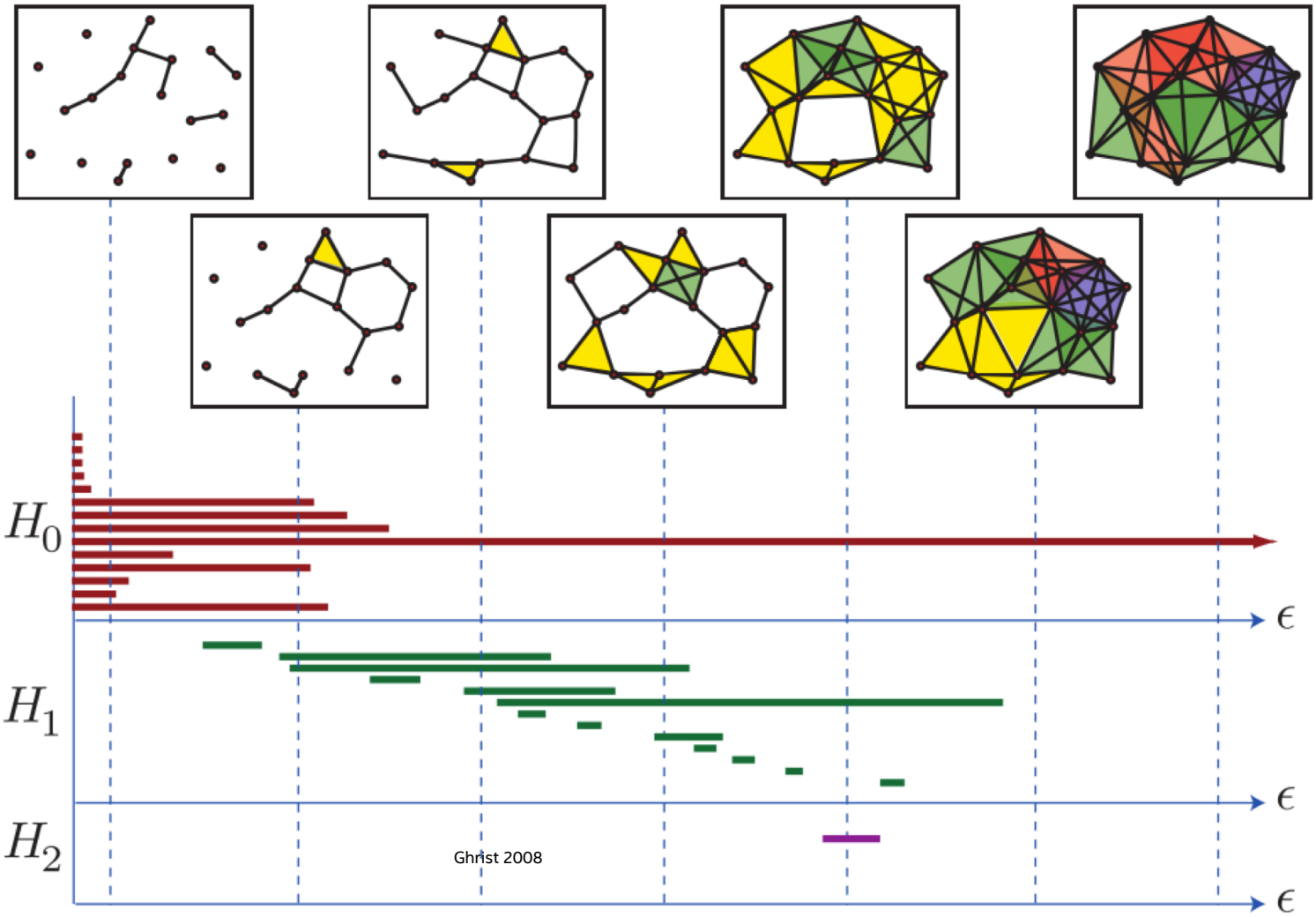


CONNECTED COMPONENTS & TUNNELS

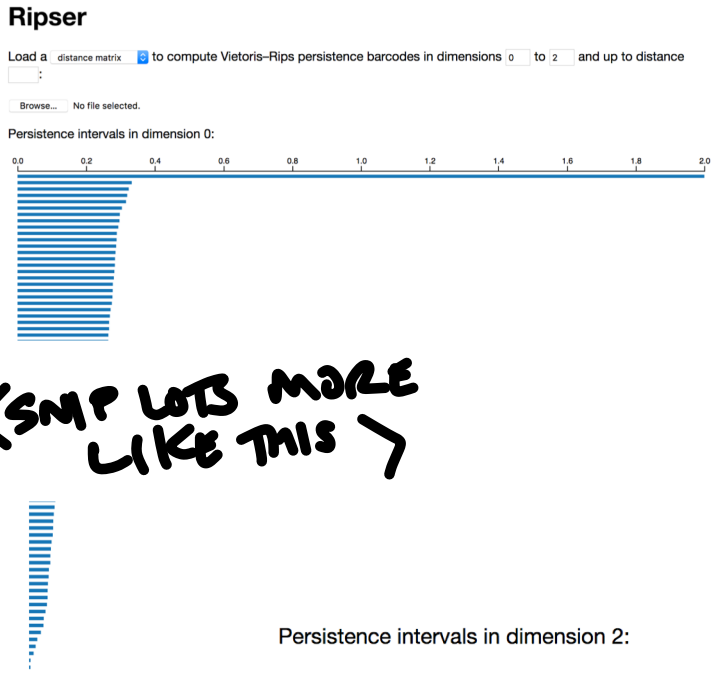


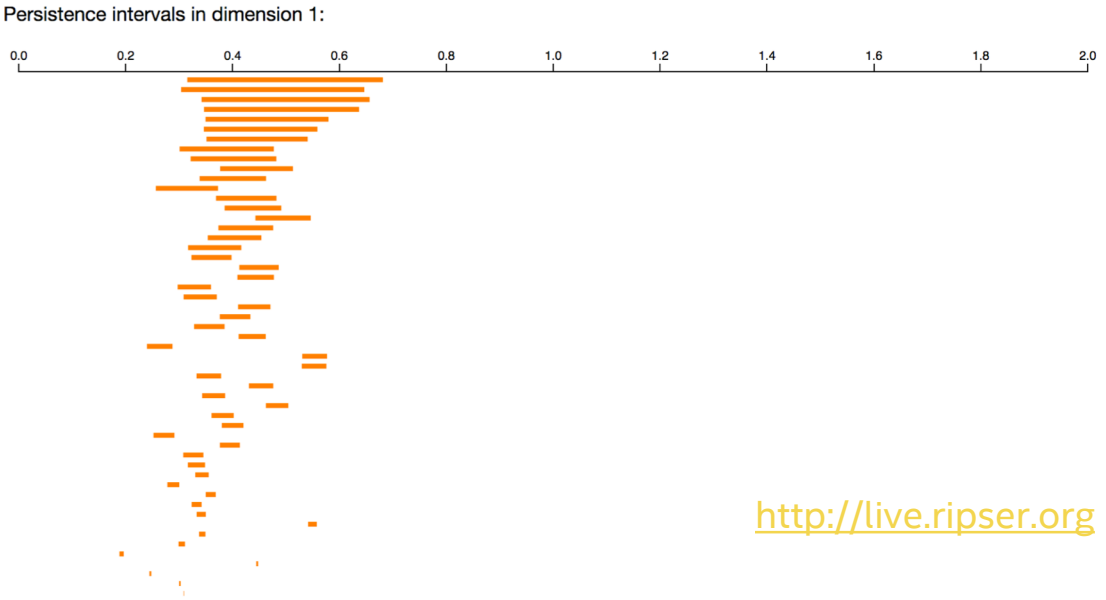
0 15 32 45 50 51 53 54 56 57 58 60

BARCODES

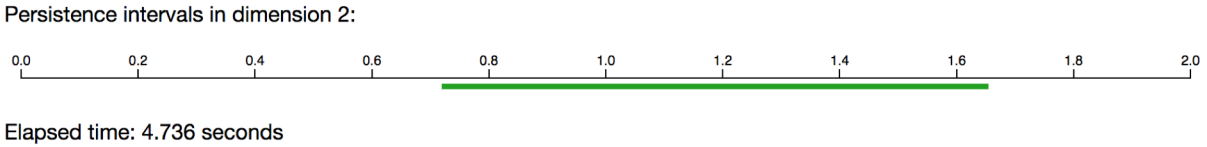


WHAT SHAPE IS THIS DATASET?





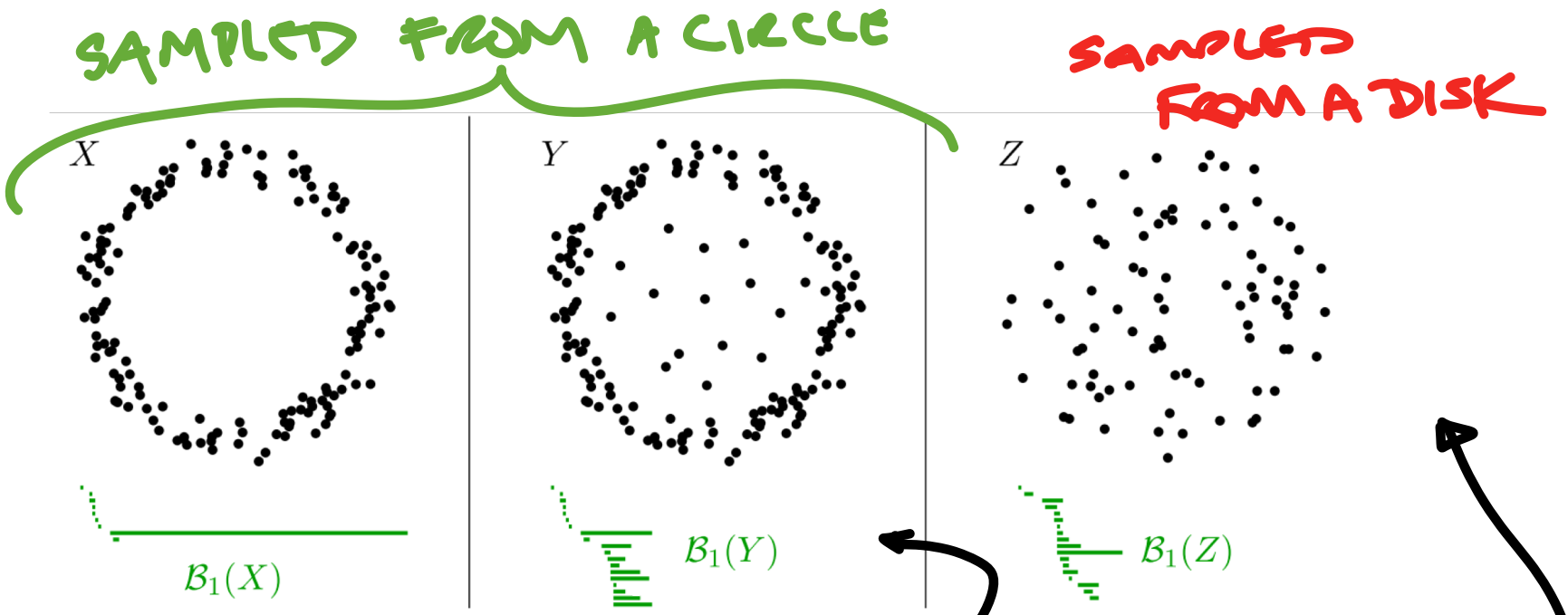
http://live.ripser.org



2-PARAMETER PERSISTENT HOMOLOGY

More is better.

THE TROUBLE WITH OUTLIERS

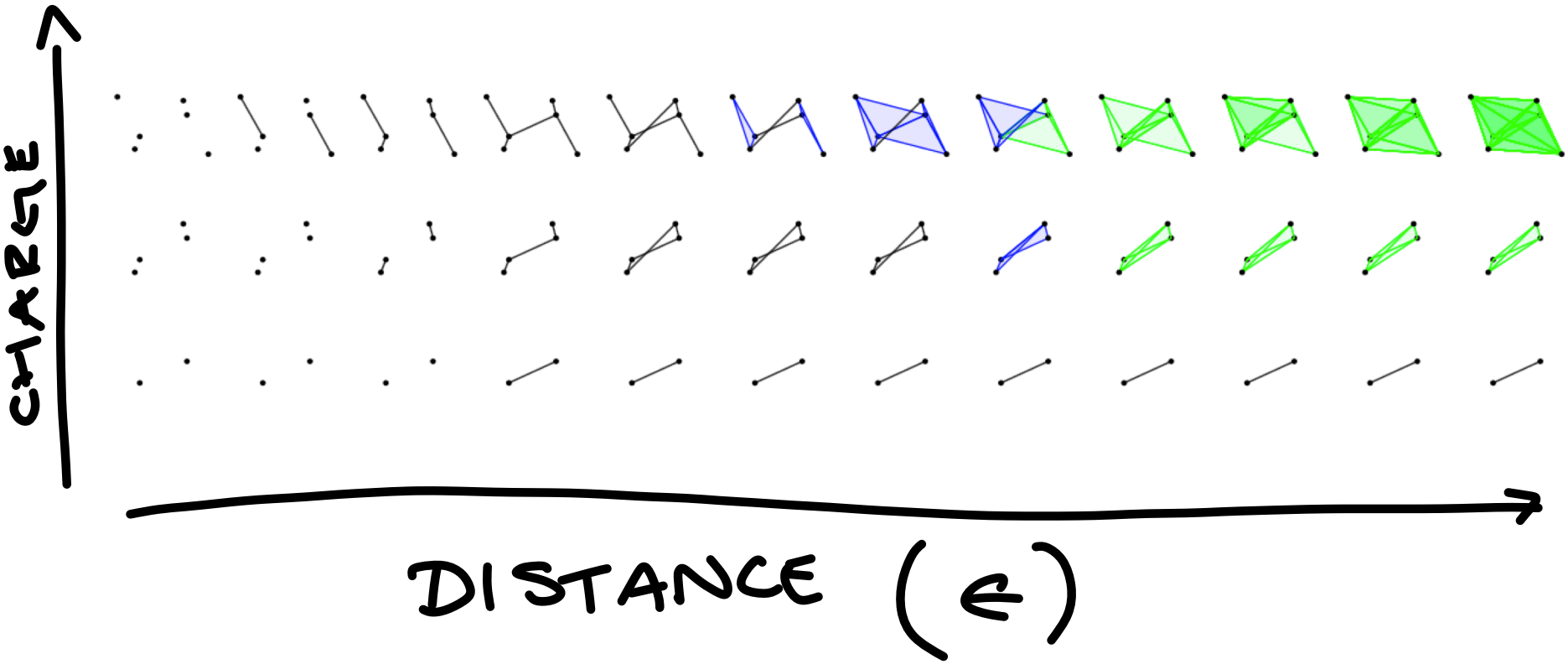


Lesnick & Wright 2015

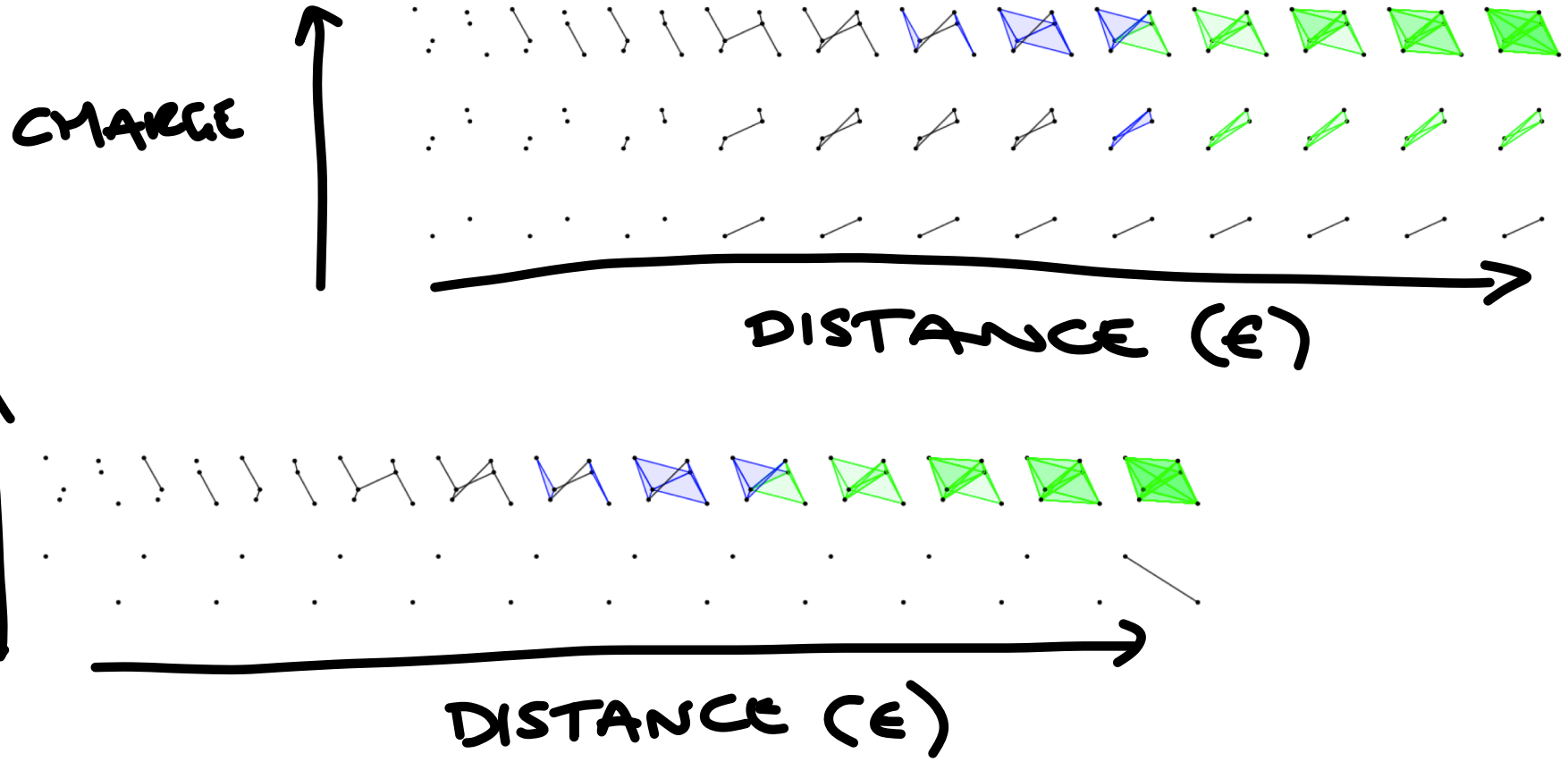
SOLUTION: 2-PARAMETER PERSISTENT HOMOLOGY

- Sweep 2 parameters, not just one, build ***bifiltrations*** instead of filtrations
- Second parameter depends on data & problem
  - Can be structural, e. g.:
    - "how many points are in close proximity of this point?"
    - "how close is this point to the center of the graph?"
  - Can be domain-specific, like atomic mass in a molecule

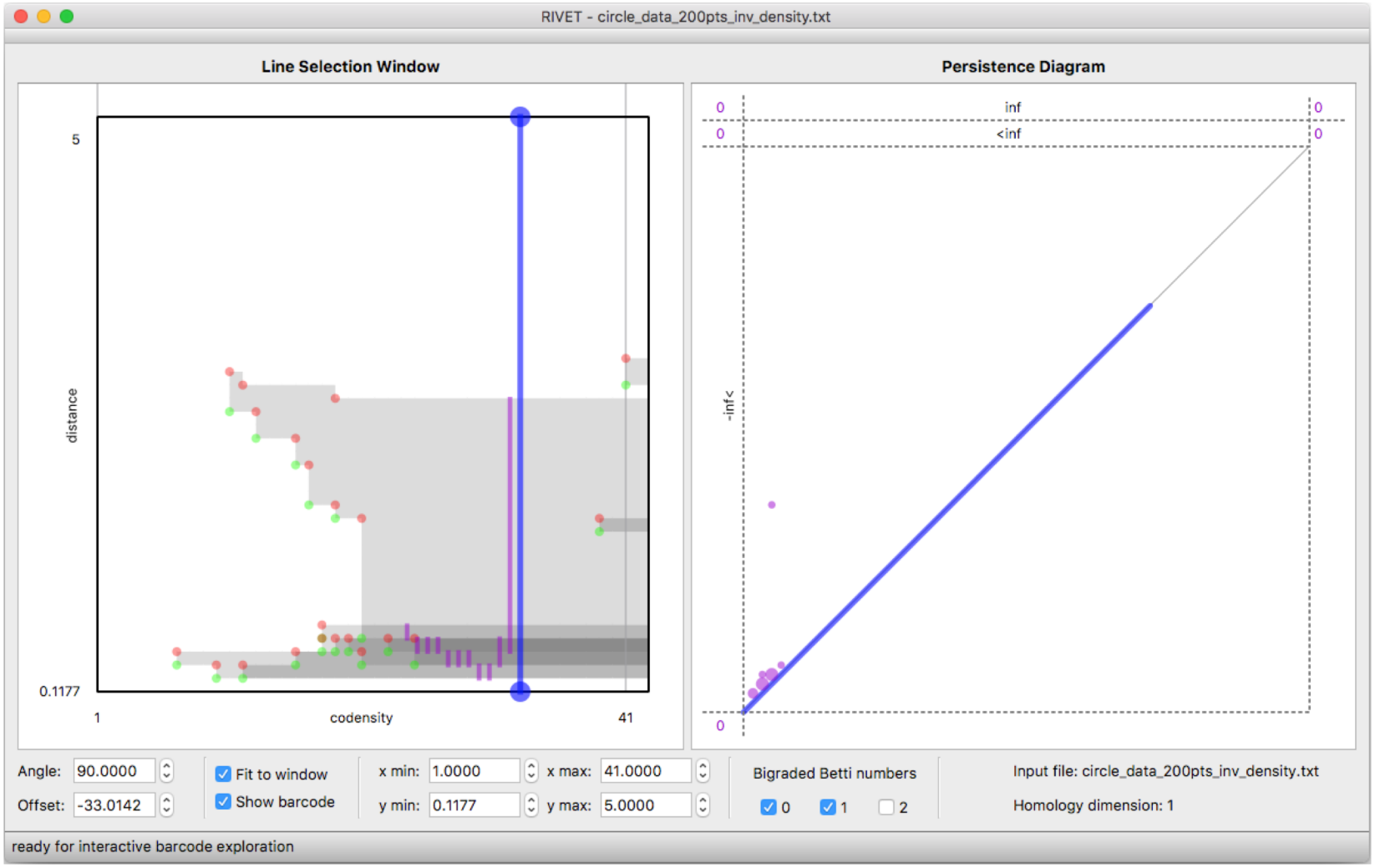
BIFILTRATIONS



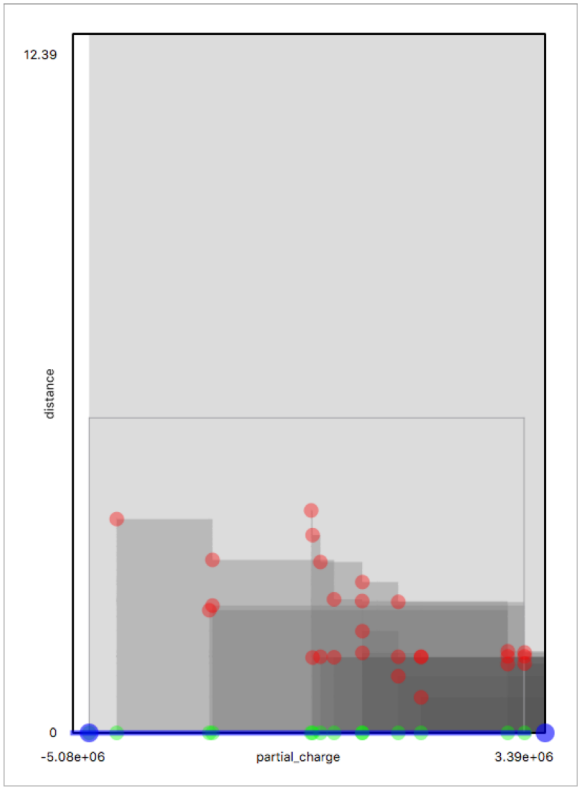
DIFFERENT BIFILTRATIONS CAN SHOW DIFFERENT STRUCTURE



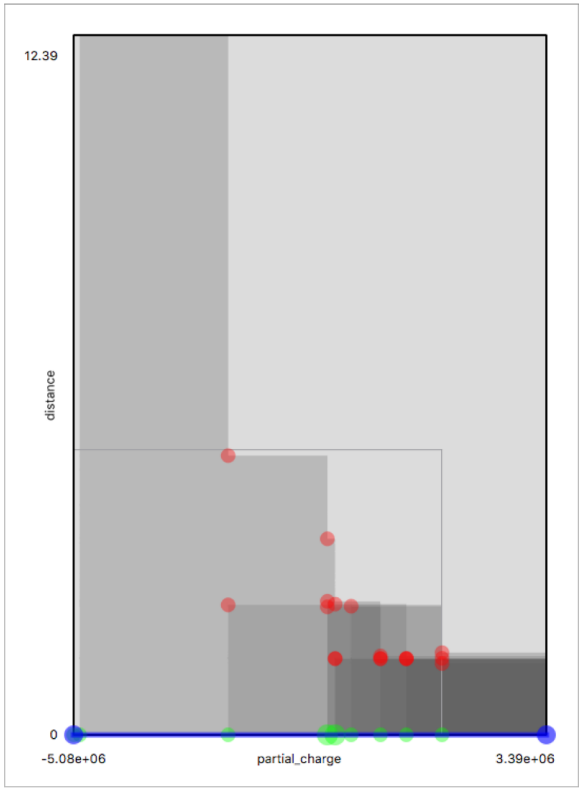
RIVET


RIVET - circle_data_200pts_inv_density.txt
Line Selection Window
Persistence Diagram
5
distance
0.1177
1
codensity
41
0
inf
0
0
<inf
0
-inf<
0
Angle: 90.0000
Offset: -33.0142
Fit to window
Show barcode
x min: 1.0000
x max: 41.0000
y min: 0.1177
y max: 5.0000
Bigraded Betti numbers
0
1
2
Input file: circle_data_200pts_inv_density.txt
Homology dimension: 1
ready for interactive barcode exploration

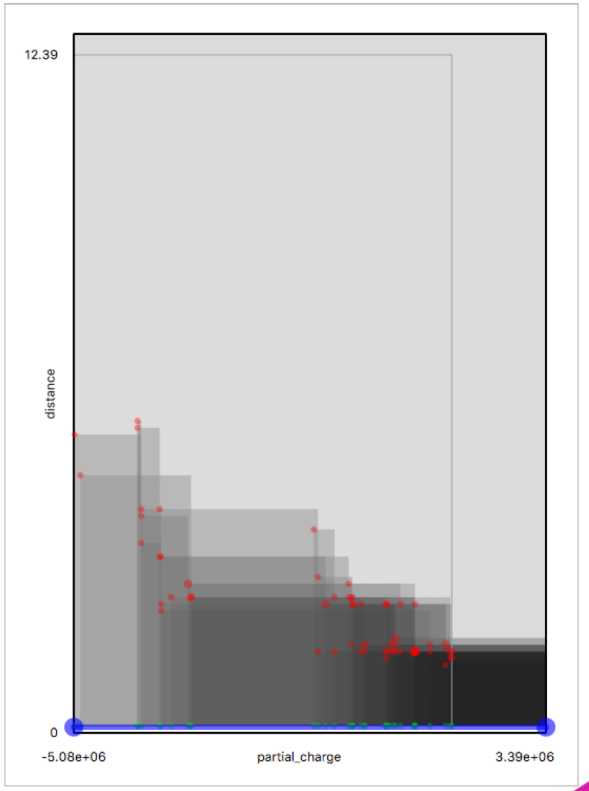
GRADED DIMENSION VISUALIZATION WITH RIVET



Aspirin

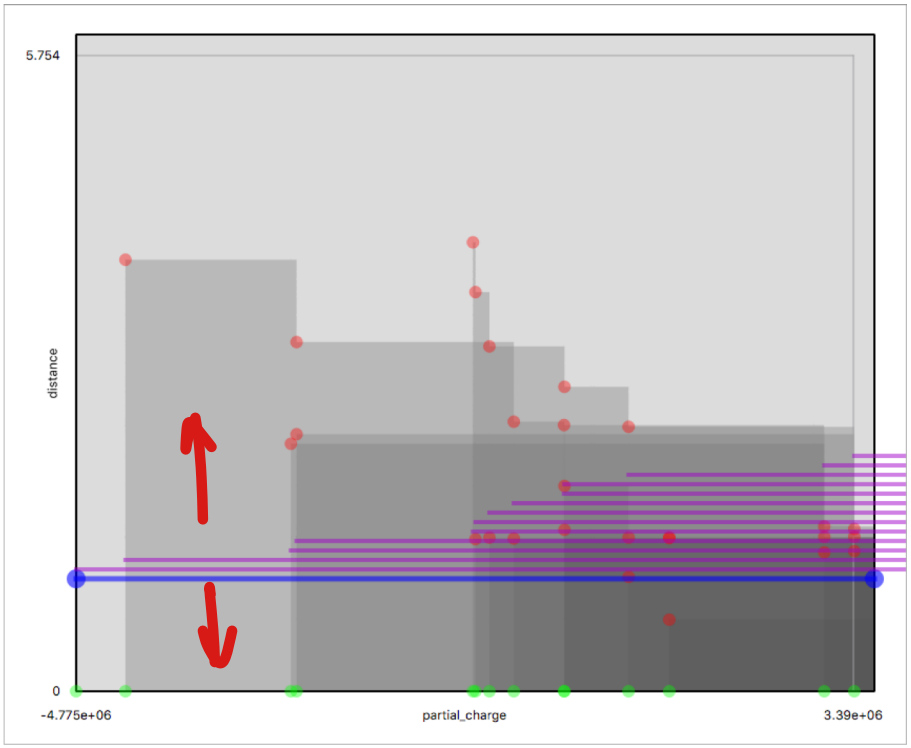


Tylenol

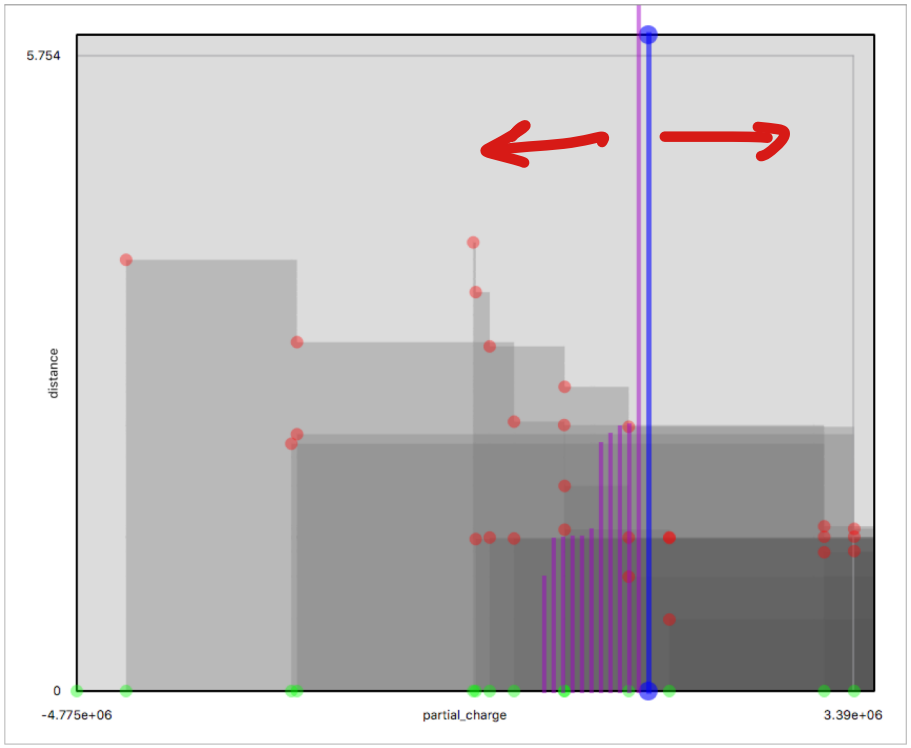


Doxyrubicin

2 VIEWS OF ASPIRIN

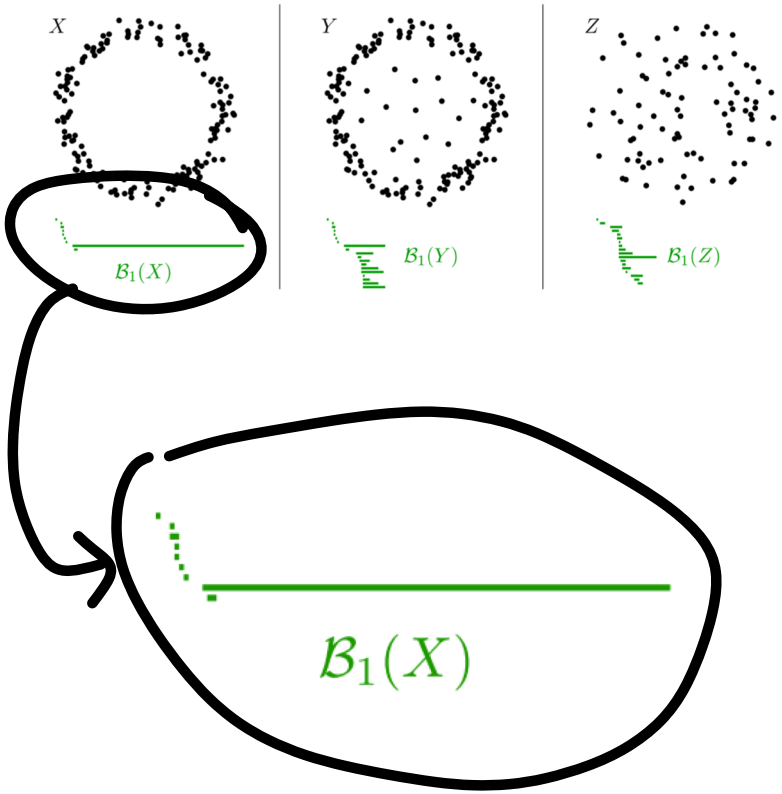


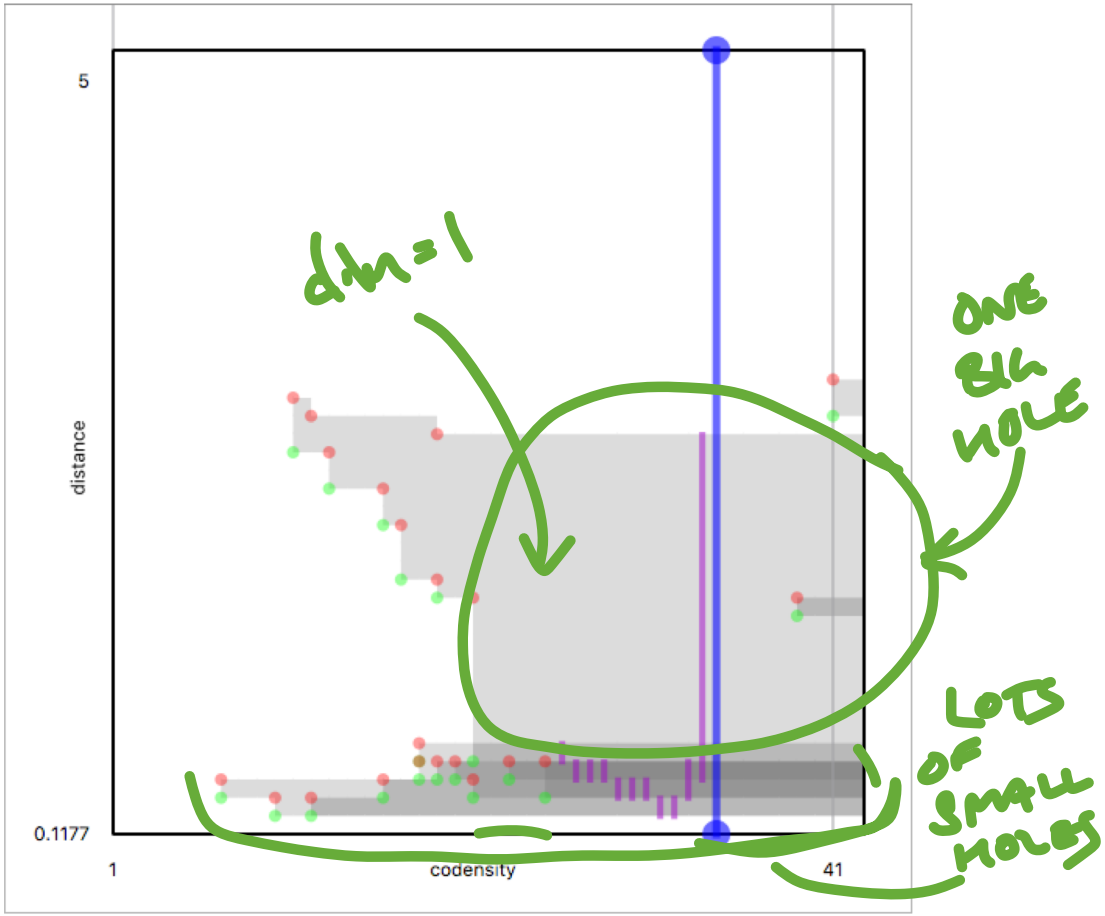
How barcodes vary with ϵ distance



How barcodes vary with partial charge

THE NOISY CIRCLE IN 2-PARAM PERSISTENCE





THE NOISY CIRCLE IN 2-PARAM PERSISTENCE

$\mathcal{B}_1(X)$

ALMOST SAME!

WE'VE FOUND THE CIRCLE IN THE NOISE.

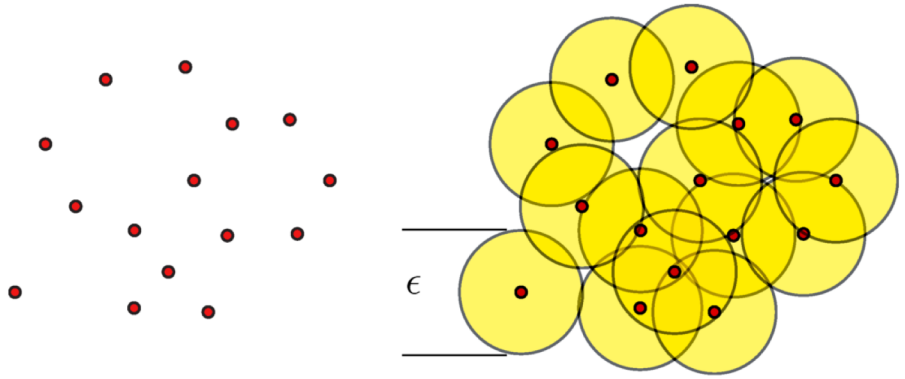
LEARNING & PERSISTENT HOMOLOGY

Better together.

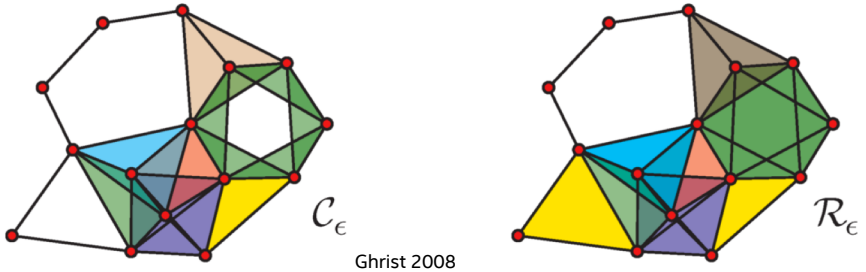
MANY KINDS OF LEARNING

- Learn the parameters for persistent homology itself
- Topological features as input for GLM
- Topological features as input for DL
- Analyze data sets with TDA to drive DL architecture
- And more…

WHICH COMPLEX BUILDING ALGORITHM MATTERS



Čech complex for distance ϵ

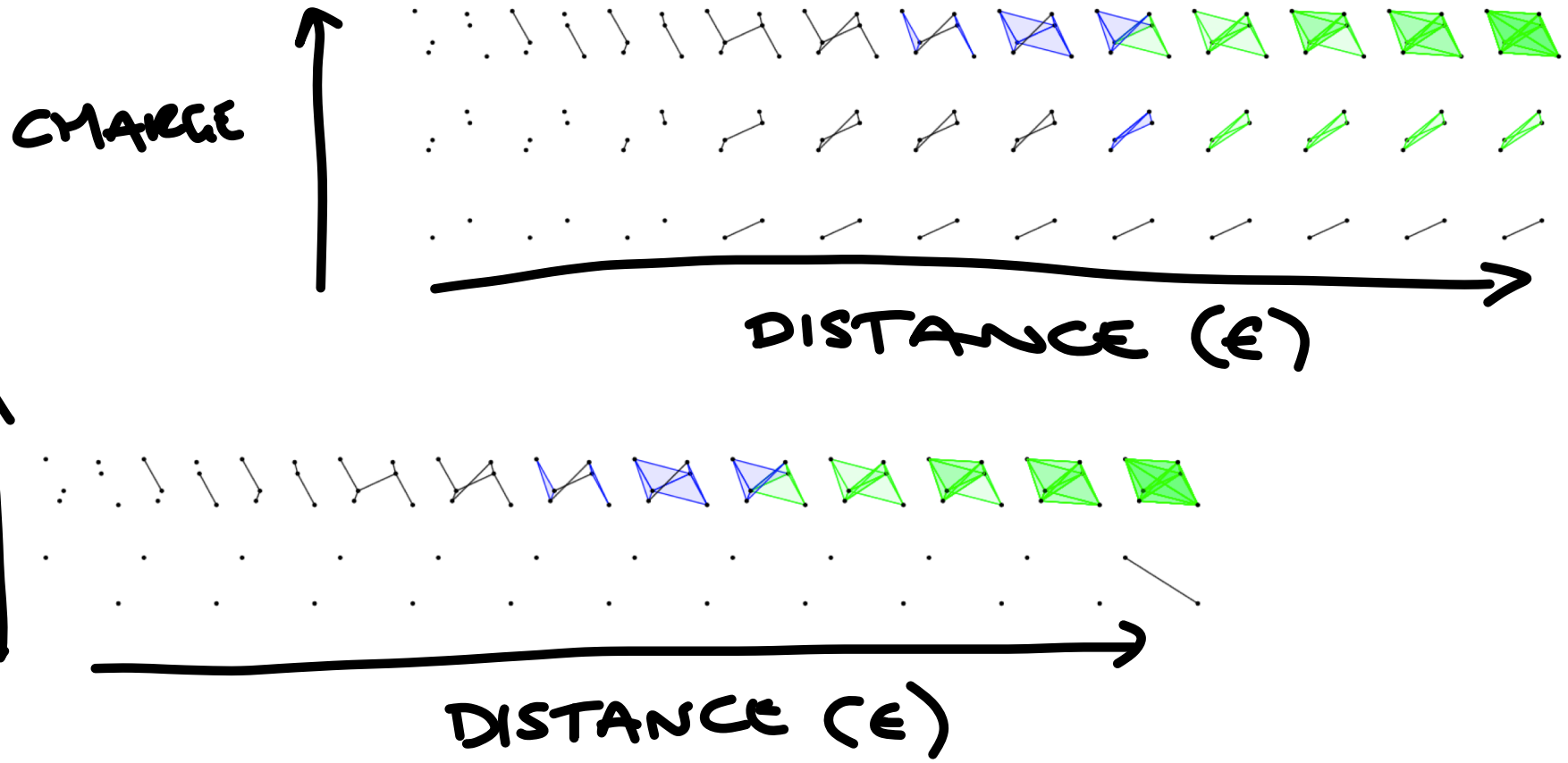


Ghrist 2008

Vietoris-Rips Complex for distance ϵ

Same data, different shapes!

CHOICE OF SWEEP PARAMETERS MATTERS



TDA + GLM LEARNING

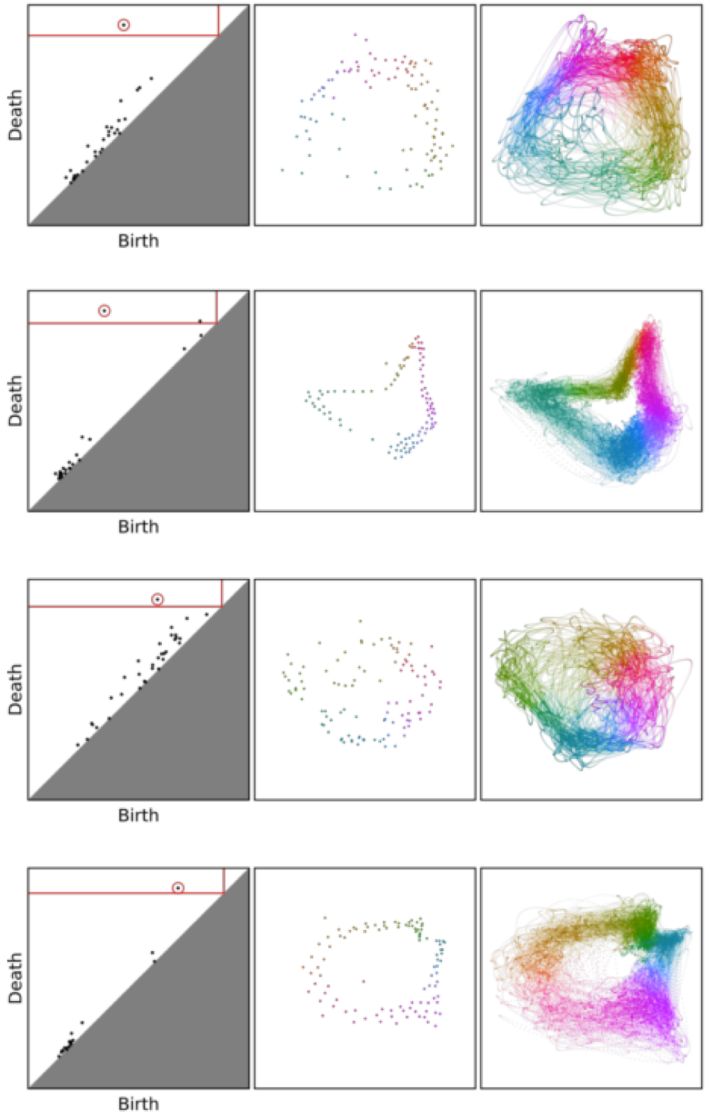
PCA

GLM

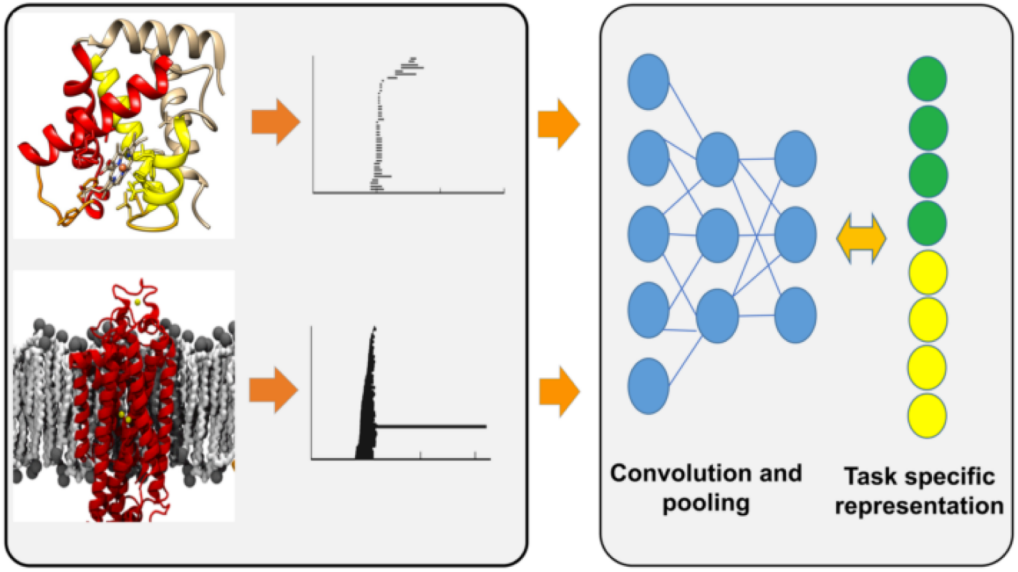
Persistent (co)homology

Circular parameterization

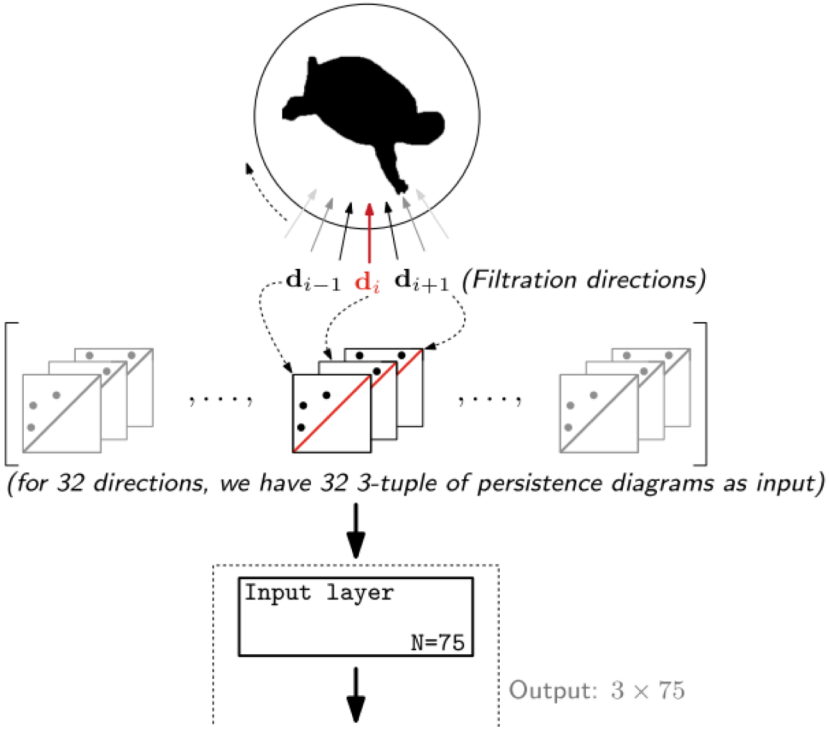
Rybakken, et al. 2017



TDA (PERSISTENT HOMOLOGY) INPUTS FOR DEEP LEARNING

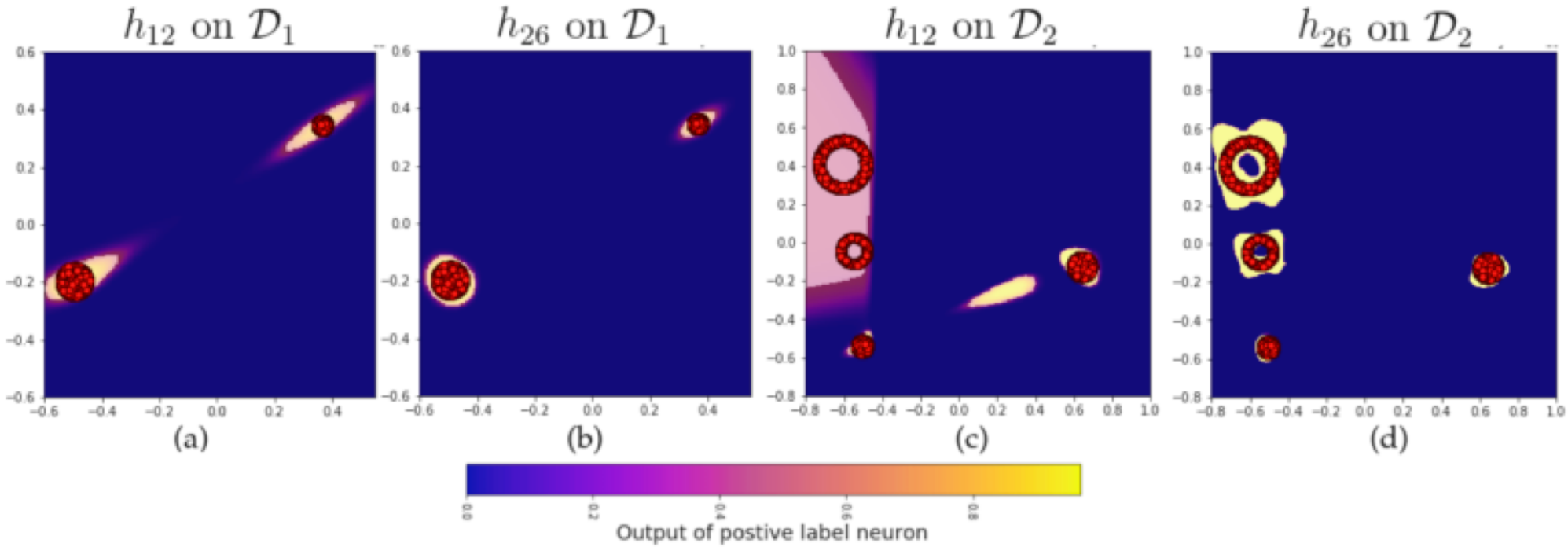


Cang & Wei 2017



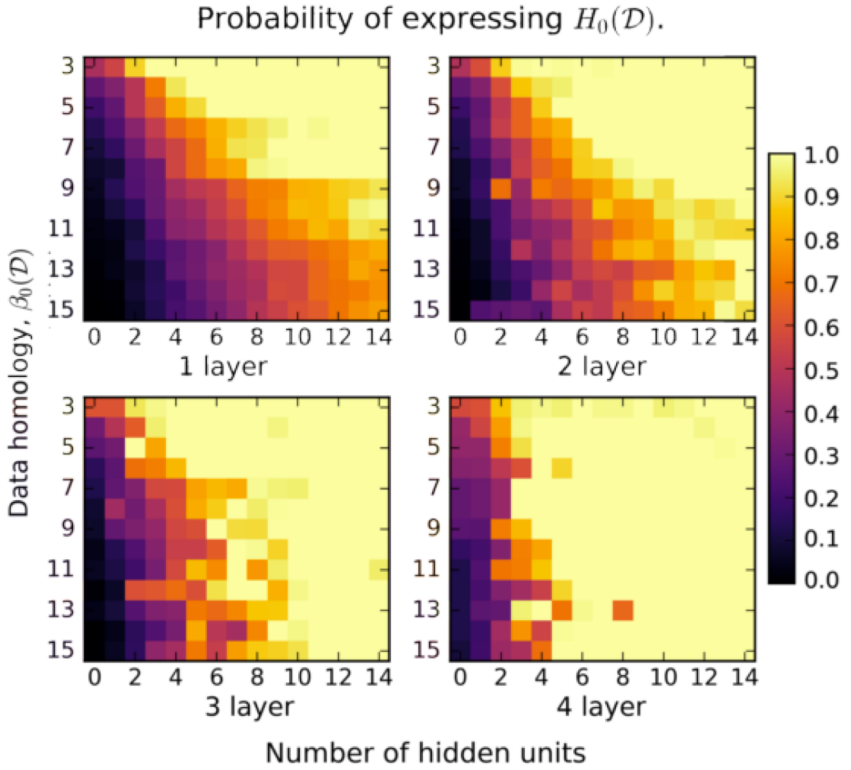
Hofer et al. 2017

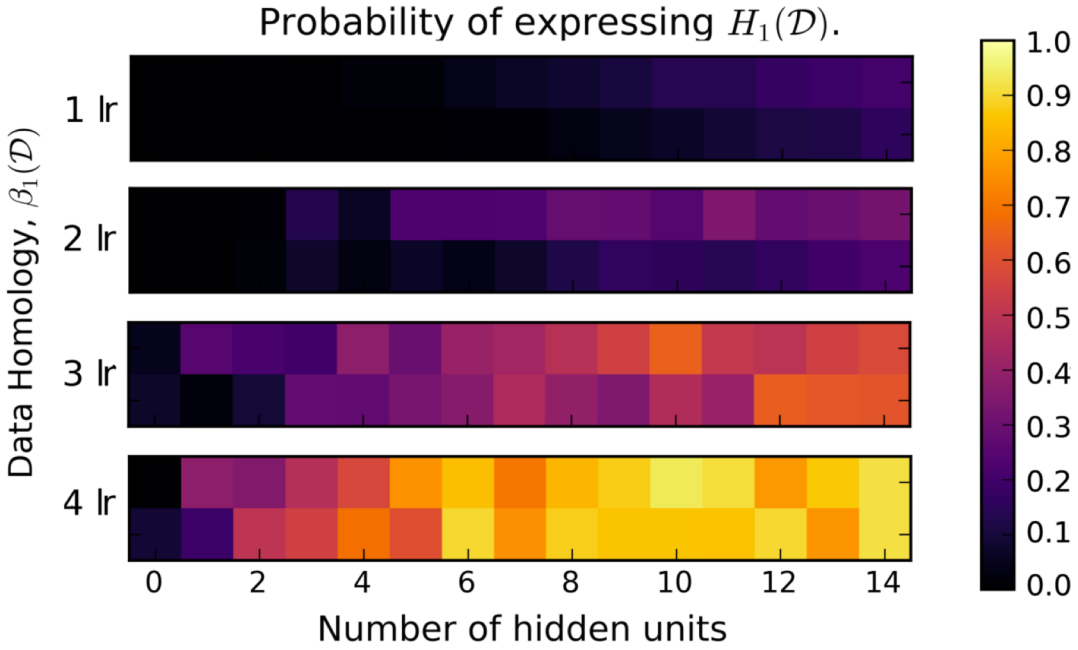
DATA-FIRST ARCHITECTURE SELECTION



Guss & Salakhutdinov 2018: On Characterizing the Capacity of Neural Networks using Algebraic Topology

DATA-FIRST ARCHITECTURE SELECTION

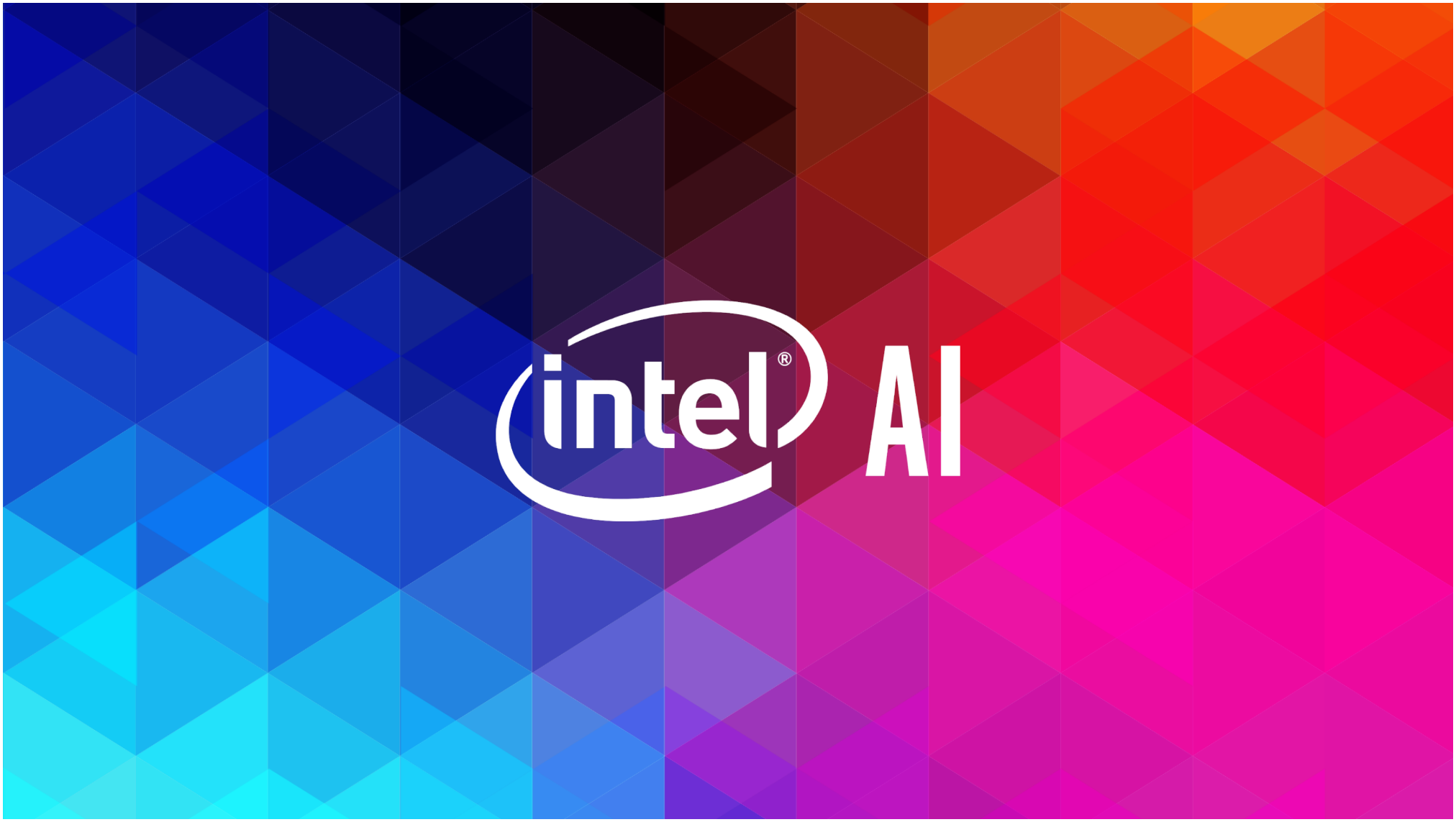




Guss & Salakhutdinov 2018

PERSISTENT HOMOLOGY

- Powerful technique for understanding data
- Strong theoretical foundations
- Rapidly growing interest among researchers
- A new frontier for enhancing AI


intel
AI

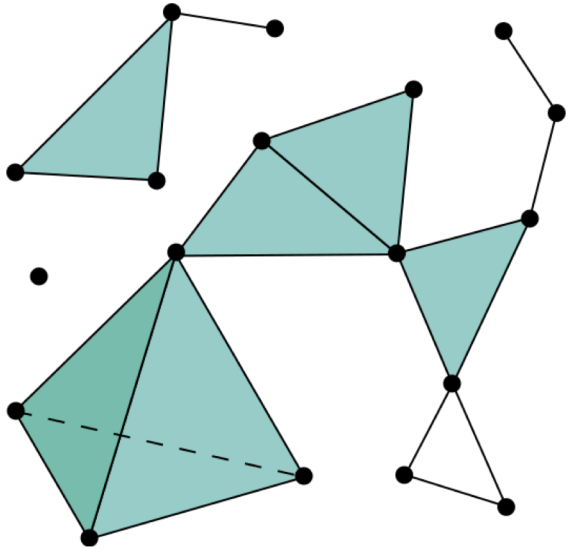
RIVET (RANK INVARIANT VISUALIZATION AND EXPLORATION TOOL)

- Tool for calculating 2-parameter persistence from data and visualizing it
- Invented by Mike Lesnick & Matthew Wright in about 2013, with help for the last couple of years from me and others
  - Paper: Lesnick, M., & Wright, M. (2015). Interactive Visualization of 2-D Persistence Modules. https://arxiv.org/abs/1512.00180
- Get it at http://rivet.online

RIPSER

- By Ulrich Bauer, paper not yet published but software already freely available at http://ripser.org
- Fast C++ tool for persistent homology, including a web version you can try out at http://live.ripser.org
- Feed it your distance matrix (or some other formats) and it will give you barcodes

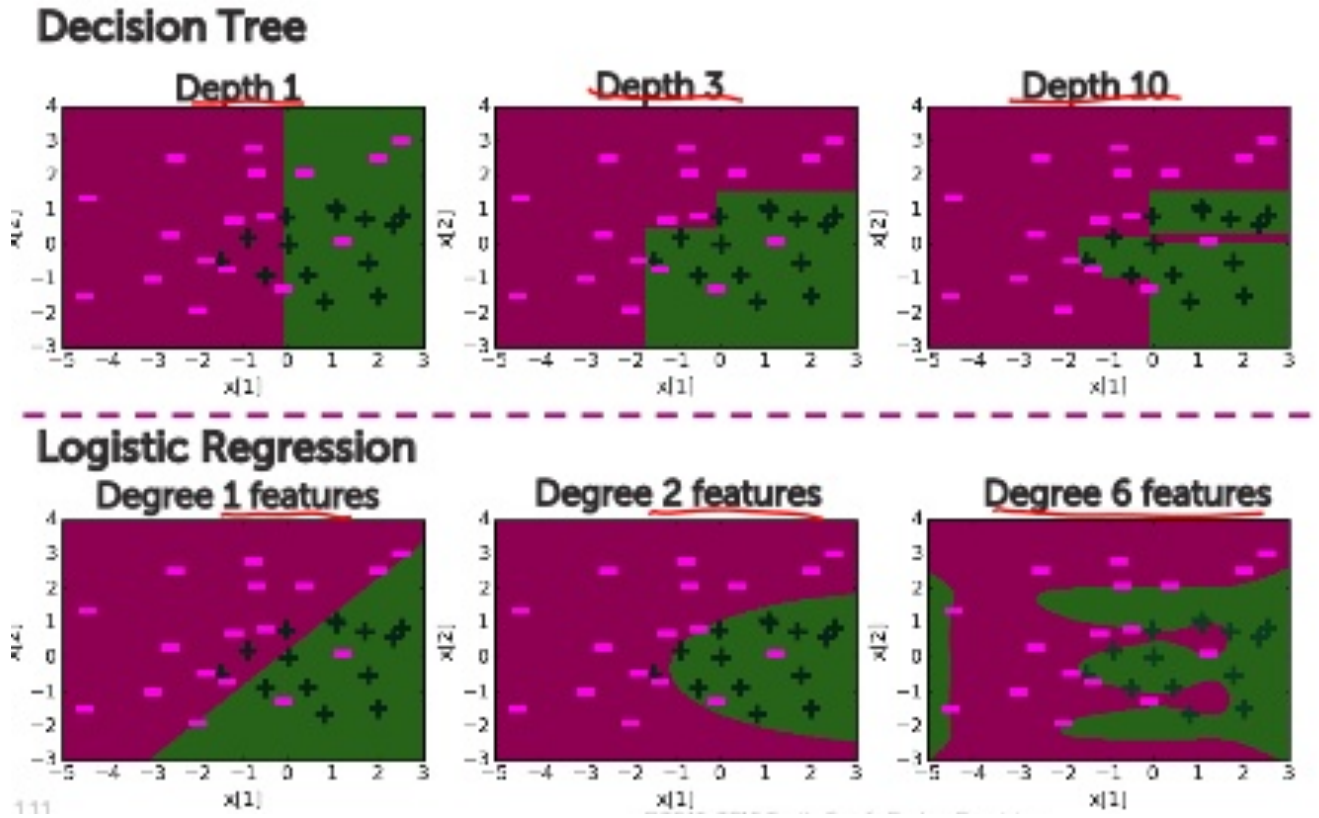
SIMPLICIAL COMPLEX



Wikipedia

AIDC
INTEL AI DEVCON 2018

DECISION BOUNDARIES



Fox & Guestrin 2015-2016