SAND2024-04063C

Sandia National Laboratories

*Exceptional service in the national interest*

# STATISTICAL INFERENCE WITH TOPOLOGICAL DATA ANALYSIS

## *With Applications to Precision Medicine*

Esha Datta

*S. Scott Collis Data Science Fellow*
*Sandia National Laboratories*

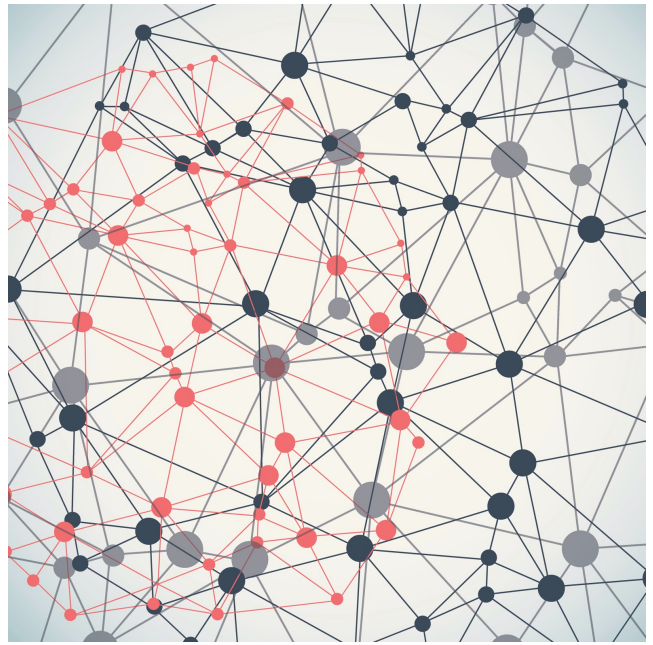
April 3rd, ACTS Meeting, Las Vegas

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

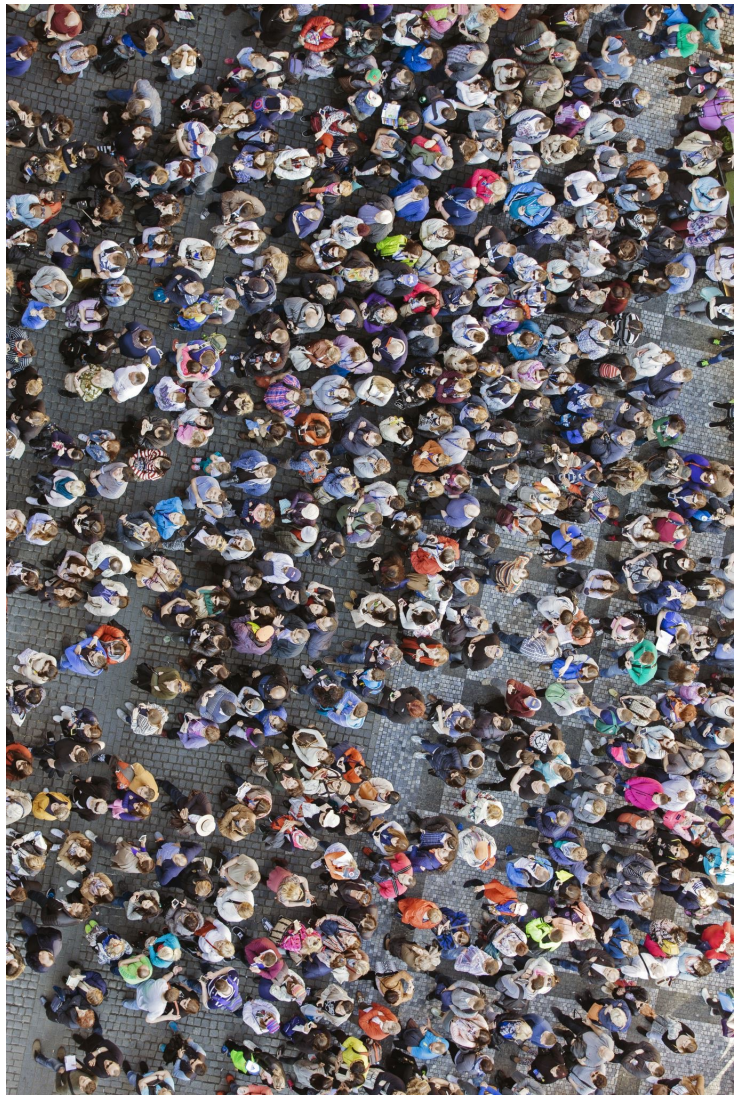
# BIOMEDICAL DATA IS BIG (AND GETTING BIGGER)

- Three V's of Big Data: **volume, velocity, and variety**
- Especially true for biomedical data:
- **Omics revolution** provides novel perspectives on understanding disease and patients
- **Problem:** the analysis of biomedical data is a bottleneck

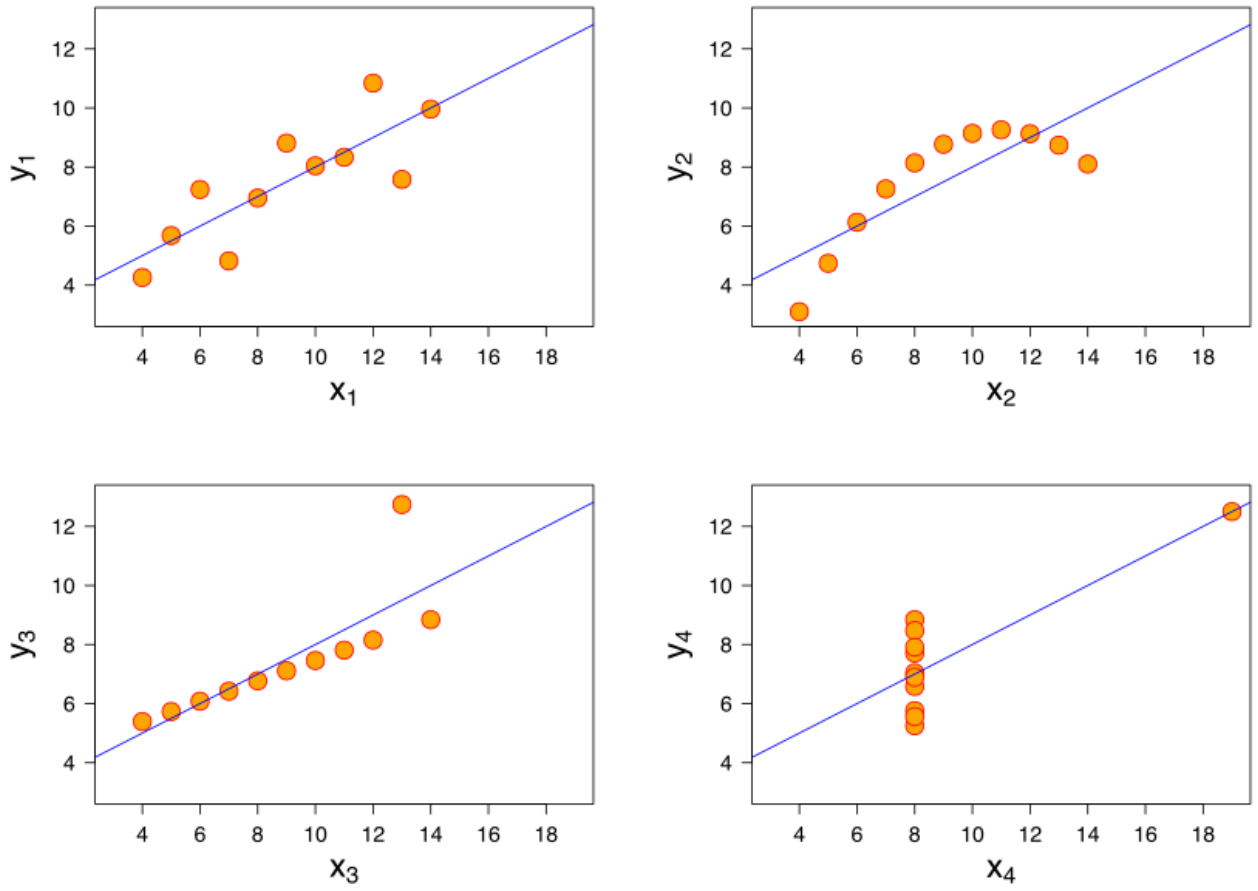


# PRECISION MEDICINE NEEDS CLUSTERING

- **Patient stratification**: ”the division of a patient population into distinct subgroups based on the presence or absence of particular disease characteristics” (Abdelnour et al 2022)
- Enables targeted treatments and more accurate prognosis for patients
- Agnostic, robust clustering is necessary
  - Number of subtypes may not be known for a given condition
  - The clusters must be validated, not *ad hoc*
  - The clusters must be interpretable in the given domain



# THE SHAPE OF DATA MATTERS



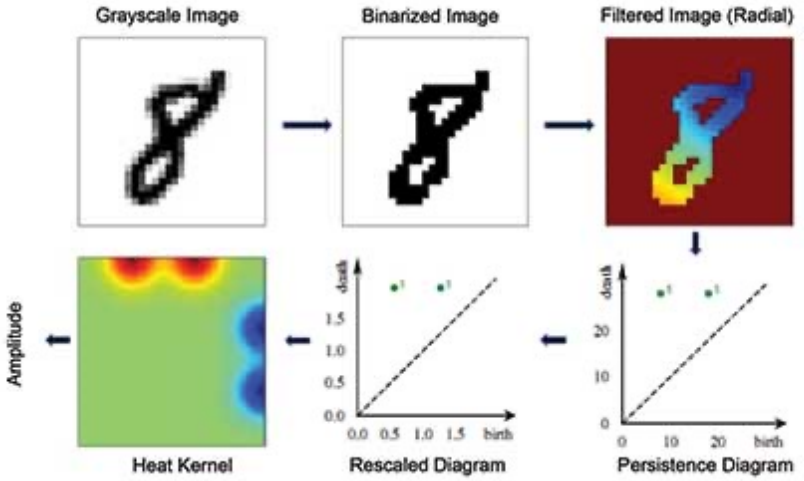

Anscombe's quartet: 4 distinct datasets with same x-, y- mean, standard deviation, and correlation

# THE ROLE OF TOPOLOGY

- A field of mathematics that considers shape, space, and relationships between geometric objects
- Rigorously defines notions of equivalence across distinct structures
- What properties are **invariant,** i.e., preserved under continuous deformation?
- Topological data analysis (TDA) is a broad set of methods for the analysis, visualization, and exploration of data using topological techniques

# TOPOLOGICAL DATA ANALYSIS

- TDA refers to a broad set of methods for the analysis, visualization, and exploration of data using topological techniques
- Major algorithms:
  - Persistent homology, used for clustering and shape identification
  - Mapper, used for visualization
  - U-Map, used for manifold learning and variations on t-SNE
- Advantages of using topology:
  - Coordinate-free due to reliance on metric
  - Invariance gives stability relative to noise in data

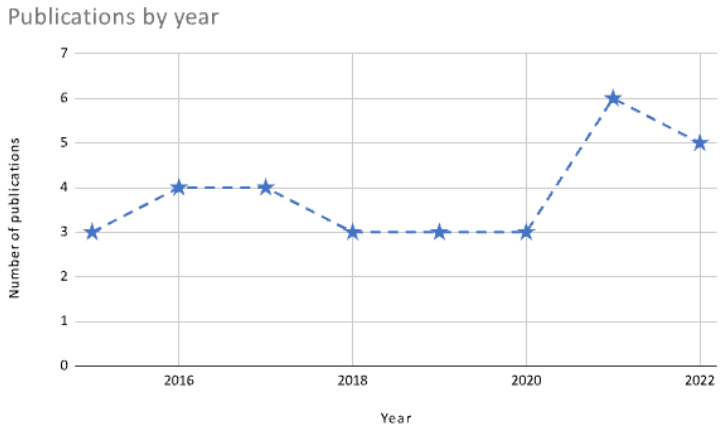


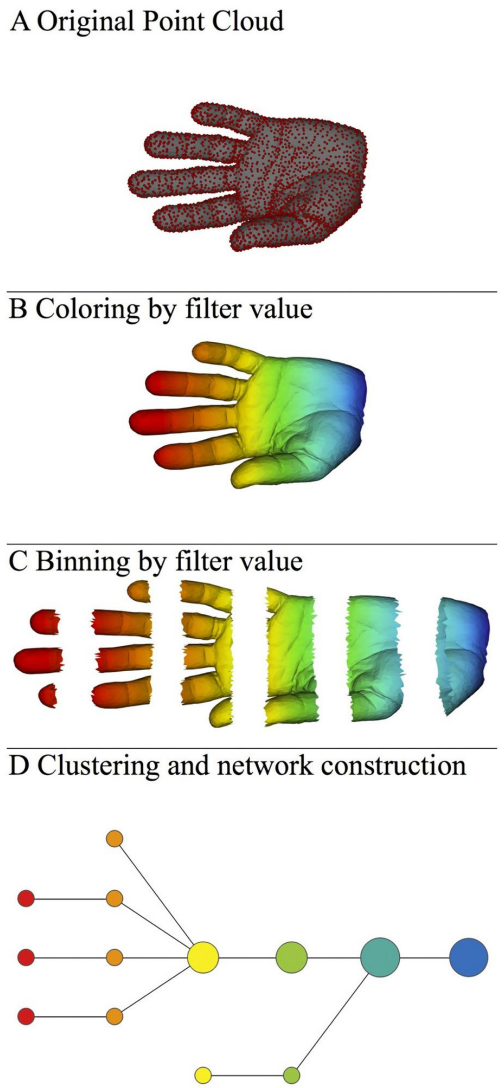

Garin et al 2019

# MAPPER AND MAPPERPLUS

# MAPPER: A TOPOLOGICAL TOOL

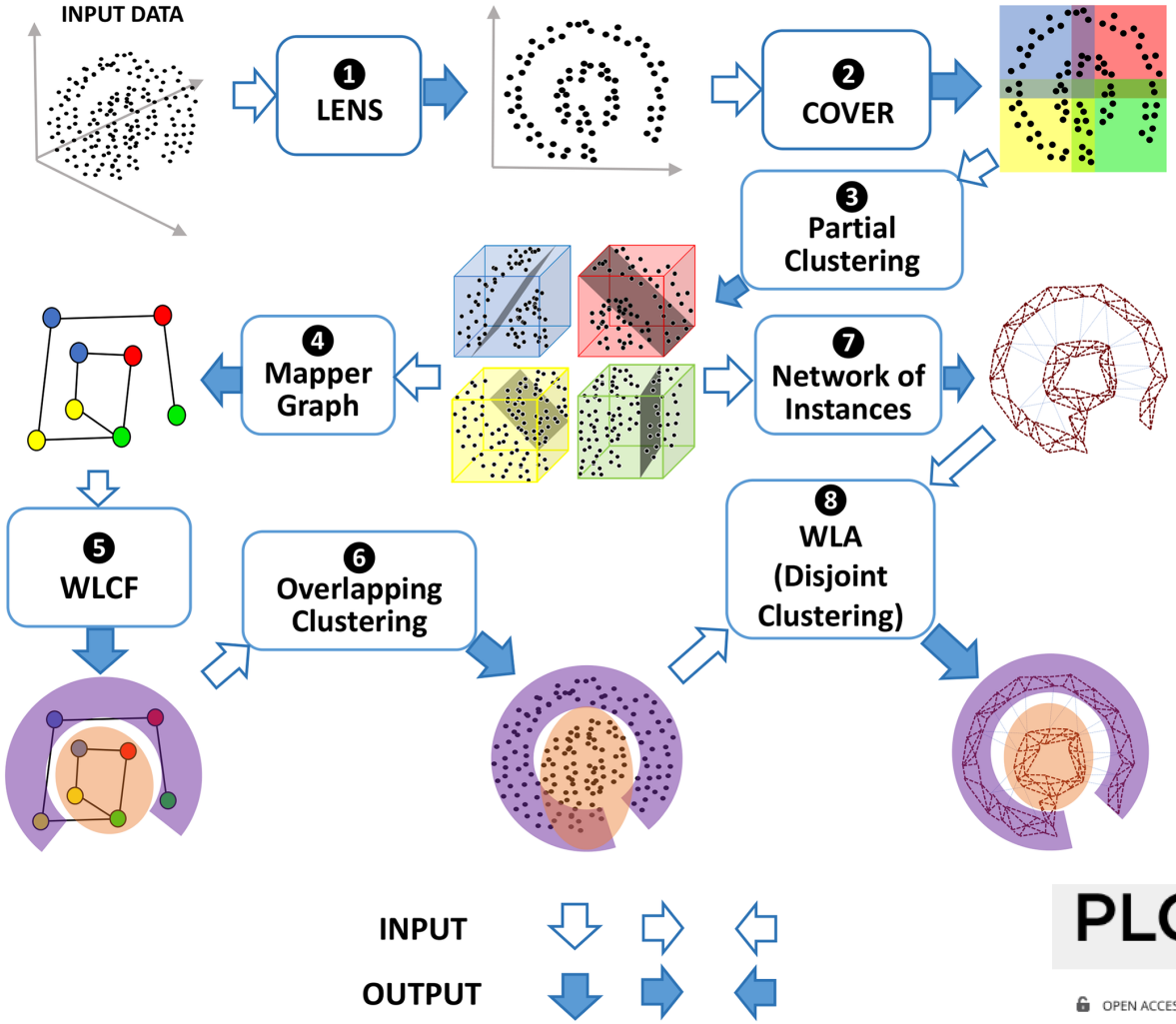
- Leverages the geometric properties of a dataset to generate a graph
- Projects high-dimensional data to a lower dimension
- Clusters data at different lens values to preserve the local connectivity
- Originally introduced for **visualization**, but is applied to clustering (with major caveats)
- Used in patient stratification with increasing frequency




Singh, Memoli, Carlsson 2007

# MAPPERPLUS: GENERATING MEANINGFUL CLUSTERS




PLOS DIGITAL HEALTH

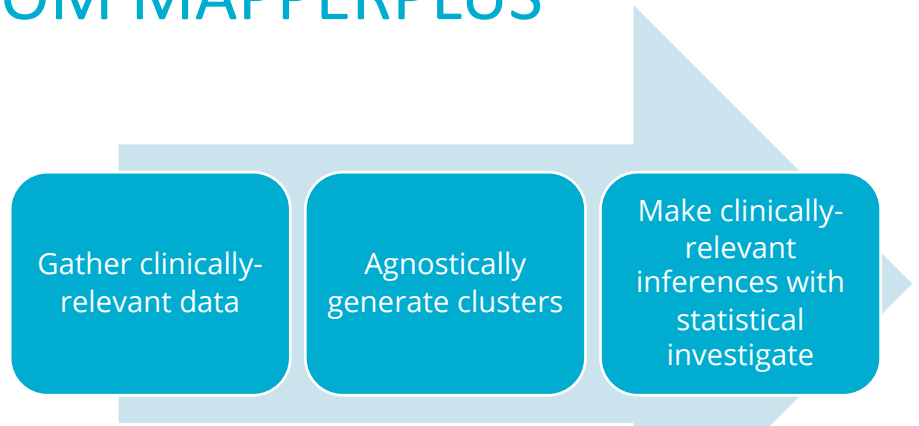
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

## MapperPlus: Agnostic clustering of high-dimension data for precision medicine

Esha Datta, Aditya Ballal, Javier E. López, Leighton T. Izu

# STATISTICAL INFERENCE FROM MAPPERPLUS




| | All (n = 187) | N1 (n = 83) | N2 (n = 82) | N3 (n = 22) | p-value |
|---|---|---|---|---|---|
| **Pre-transplant Features** | | | | | |
| Donor age | 33.47 ± 8.27 | 32.82 ± 9.27 | 31.61 ± 8.55 | 31.70 ± 6.83 | 0.22 |
| Recipient age | 9.93 ± 5.31 | 5.61 ± 2.71 | 14.19 ± 3.79 | 10.35 ± 4.54 | **<0.001** |
| CD34+ cell doses per kilogram (10^6/kg) | 11.89 ± 9.91 | 15.03 ± 10.44 | 8.53 ± 7.53 | 12.56 ± 11.94 | **<0.001** |
| CD3+ cell doses per kilogram (10^6/kg) | 4.74 ± 3.86 | 6.11 ± 4.50 | 3.15 ± 2.75 | 4.43 ± 2.87 | **<0.001** |
| CD3+ cell to CD34+ cell ratio | 5.39 ± 9.60 | 5.16 ± 11.17 | 5.61 ± 8.80 | 4.16 ± 3.69 | 0.81 |
| Recipient stem cell body mass (kg) | 35.81 ± 19.65 | 20.83 ± 9.14 | 50.27 ± 18.01 | 35.08 ± 15.02 | **<0.001** |
| Donor age below 35 | 104 (55.61%) | 48 (57.83%) | 40 (48.78%) | 16 (72.73%) | 0.115 |
| Presence of cytomegalovirus infection (CMV) in donor prior to transplantation | 72 (38.50%) | 33 (39.76%) | 30 (36.59%) | 9 (69.23%) | 0.889 |
| Recipient age below 10 | 99 (52.94%) | 83 (100%) | 7 (8.54%) | 9 (40.91%) | **0.001** |
| Recipient gender | 112 males (65.24%) | 46 (55.42%) | 52 (63.41%) | 14 (63.64%) | 0.537 |
| Presence of CMV in recipient prior to transplantation | 100 (53.48%) | 42 (50.60%) | 44 (53.66%) | 14 (63.64%) | 0.552 |
| Disease categorized as malignant | 155 (82.89%) | 62 (74.70%) | 75 (91.46%) | 18 (81.82%) | **0.017** |
| Compatibility of donor and recipient according to gender (incidence of female to male) | 32 (17.11%) | 11 (13.25%) | 16 (19.51%) | 5 (22.73%) | 0.429 |
| Compatibility of donor and recipient according to blood group (incidence of matched) | 52 (27.81%) | 21 (25.30%) | 26 (31.71%) | 6 (27.27%) | 0.655 |
| Incidence of HLA match | 159 (85.03%) | 82 (98.80%) | 62 (75.61%) | 15 (68.18%) | **<0.001** |
| Risk Group (high, low) | 69 high risk (36.90%) | 26 (31.33%) | 34 (41.47%) | 9 (40.91%) | 0.369 |
| Stem cells source (peripheral, bone marrow) | 145 peripheral (77.54%) | 67 (80.72%) | 60 (73.17%) | 18 (81.82%) | 0.446 |
| **Outcomes** | | | | | |
| 1-year Survival Time (Days) | 286 ± 132.39 | 312.11 ± 116.25 | 259.02 ± 143.87 | 291.72 ± 130.60 | **0.035** |
| 1-year Survival Rate | 64.71% | 74.39% | 54.88% | 68.18% | **0.041** |
| Study Duration Survival Rate | 54.54% | 61.45% | 43.90% | 68.18% | **0.03** |

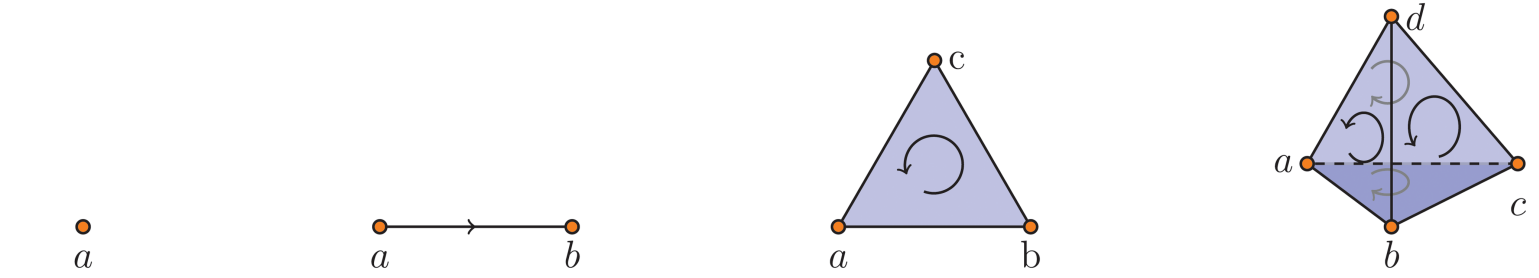
https://doi.org/10.1371/journal.pdig.0000307.t004

- Statistics enable **robust** inferences about our dataset
- **Validate observed relationships** in high-consequence areas, such as medicine
- Exploratory data analysis **must** be paired with **reproducible, principled inference**

# PERSISTENT HOMOLOGY

# HOMOLOGY: MEASURING SIGNIFICANT STRUCTURE

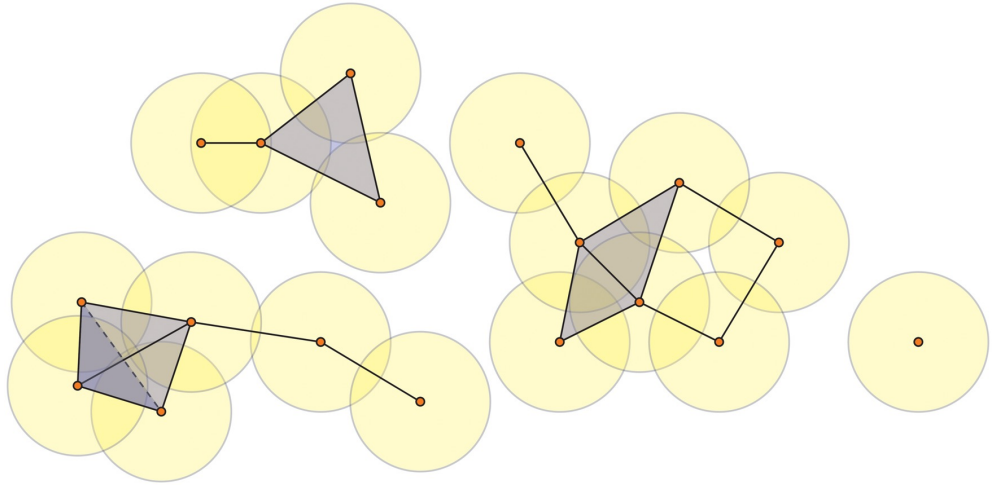
- Homology measures features like connected components, topological circles, trapped volumes, etc.
- A finite set of data points is a (noisy) sampling of an underlying topological space
- We can measure the homology of the data by creating connections between proximate observations
- The *persistence* is computed by varying the scale at which connections are made and seeing what features continue to exist




Topaz et al 2015

# BUILDING STRUCTURE ON DATA POINTS

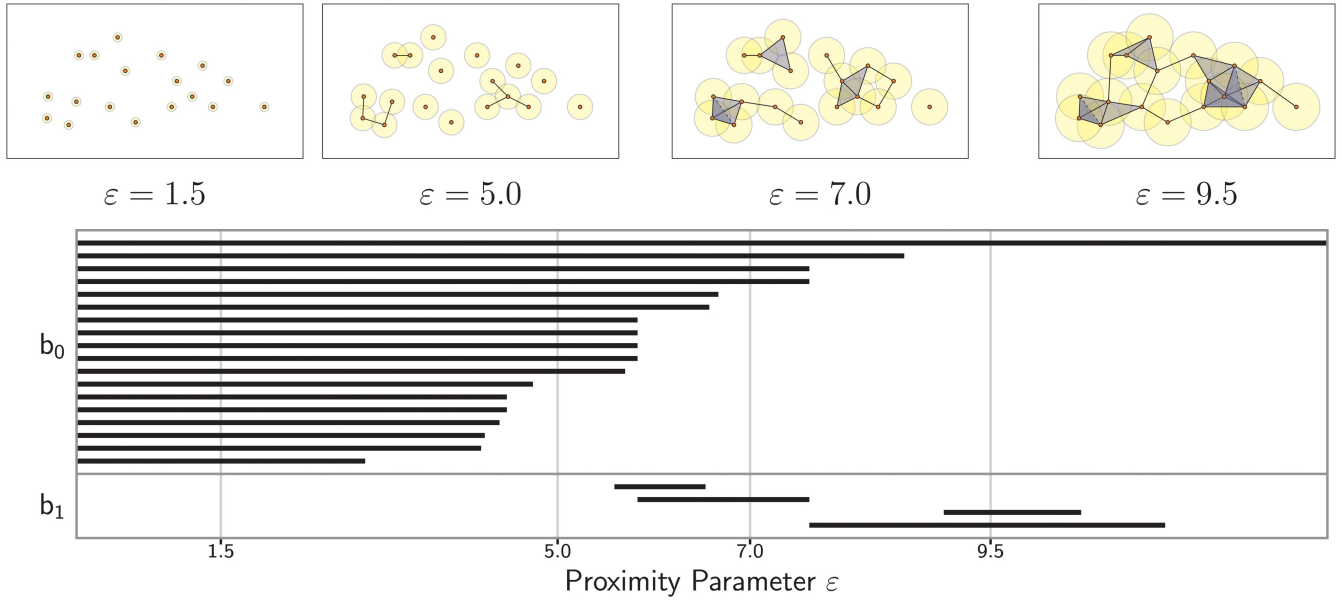
- For a given distance ε, we connect points within that distance to form a simplicial complex
- Vietoris-Rips complex on 18 points:
  - 18 points are 0-simplices
  - 2 0-simplices form a 1-simplex (edge) if their ε/2-neighborhoods intersect
  - 3 vertices form a 2-simplex (triangle) if they are pairwise connected by edges



Topaz et al 2015

# WHAT STRUCTURE PERSISTS?

- We vary the distance to observe the different structures that arise
- $b_0$ encodes connected components, $b_1$ encodes topological holes, etc.
- We consider features that persist for a long period to be more significant, while short-lived ones are considered noise
- The length of all the bars is the “lifetime” of a dimension




Topaz et al 2015

# DOES STRUCTURE MATTER?

- We seek **domain-relevant structure**, not just any structure
- In high-consequence settings, we want some **confidence that detected structures are important**
- Statistical inference with persistent homology remains a challenging problem
- Widely applied to phylogenetic trees, bacterial evolution, population genetics, cancer genomics, and single cell expression data
- Can build **confidence intervals** to measure the significance of structure
- **Cannot update beliefs in a Bayesian manner**