



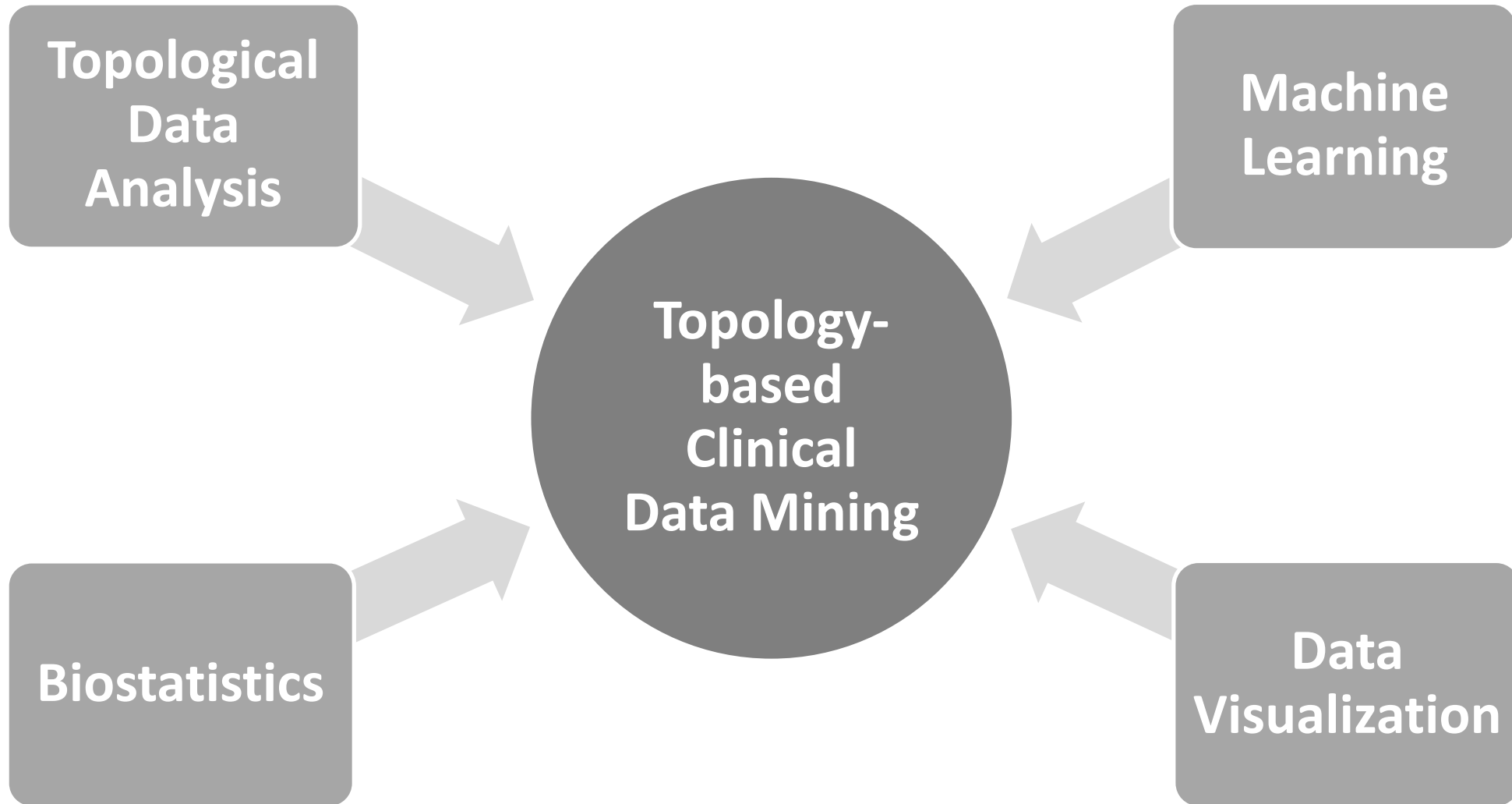
INTEGO
GROUP

HANDLING MISSING DATA IN CLINICAL TRIALS USING TOPOLOGICAL DATA ANALYSIS



Frankfurt 2018





Application of data mining techniques that involves topological data analysis, biostatistics, machine learning and visualization for extraction, analysis and interpretation of available datasets obtained during clinical trials



PART I.

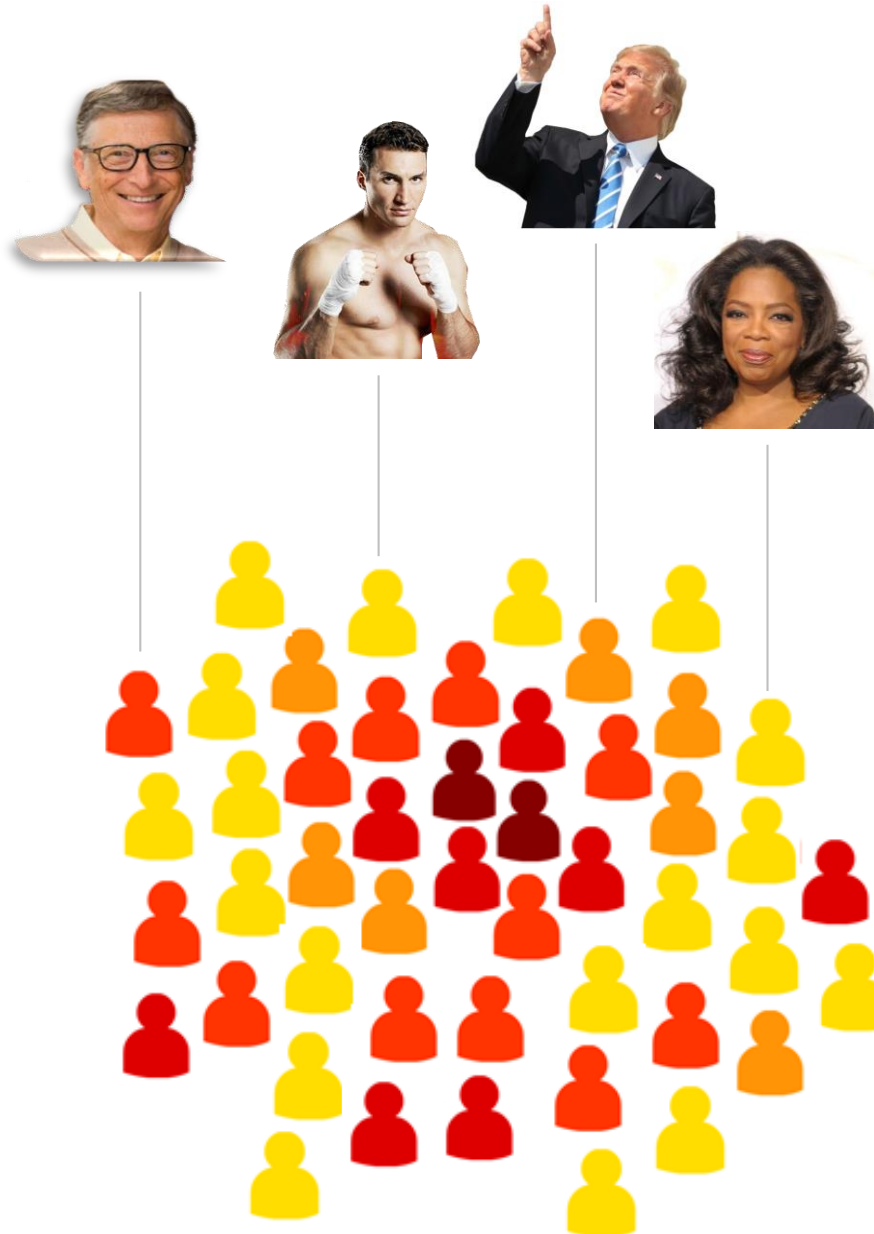
MINING DATA USING TOPOLOGY



REAL-LIFE EXAMPLE: How, as a parent, can I help my kids become successful?

- 1) Take a sample of successful people
- 2) Define what success is
- 3) Integrate possible reasons that might lead to success
- 4) Build a model using topological data analysis
- 5) Mine data to discover hidden and not-obvious relationships

100 successful people



100 successful people

25 outcomes of success



Wealth earned



Social media



Health Index



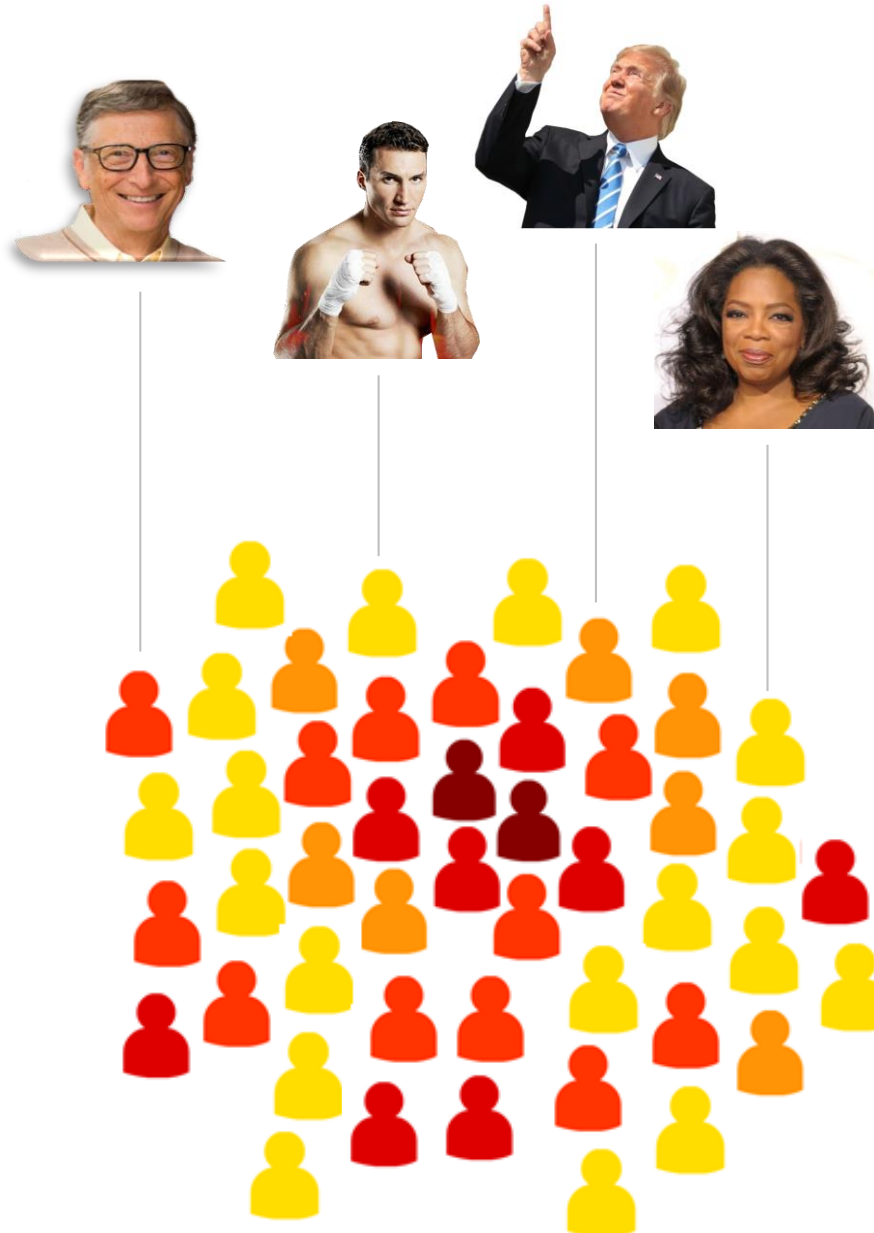
GPA



Money for charity



Happiness index



100 successful people

25 outcomes of success



Wealth earned



Social media



Health Index



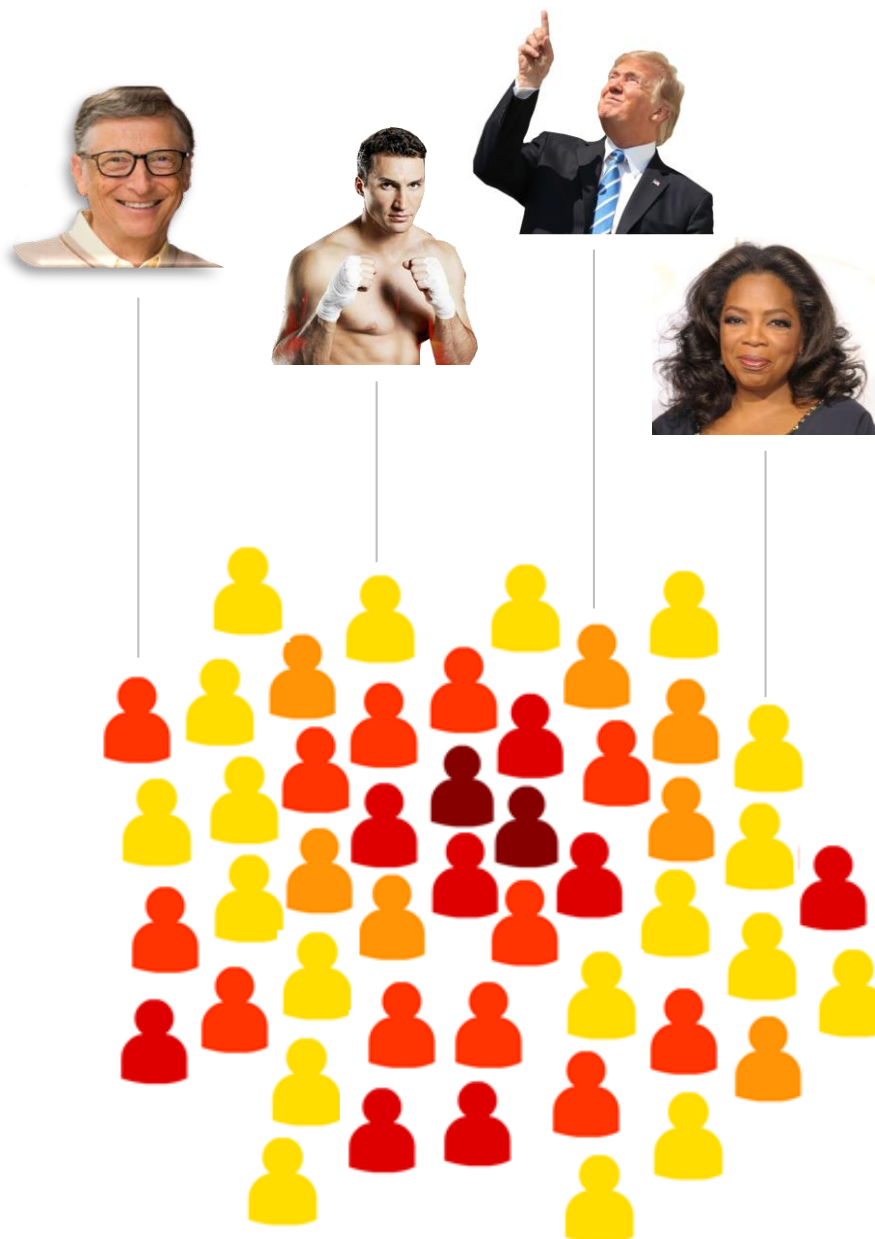
GPA



Money for charity



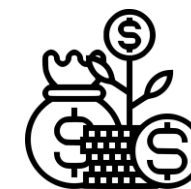
Happiness index



50 possible predictors



School ranking



Family wealth



IQ score



Hours of reading



Languages spoken

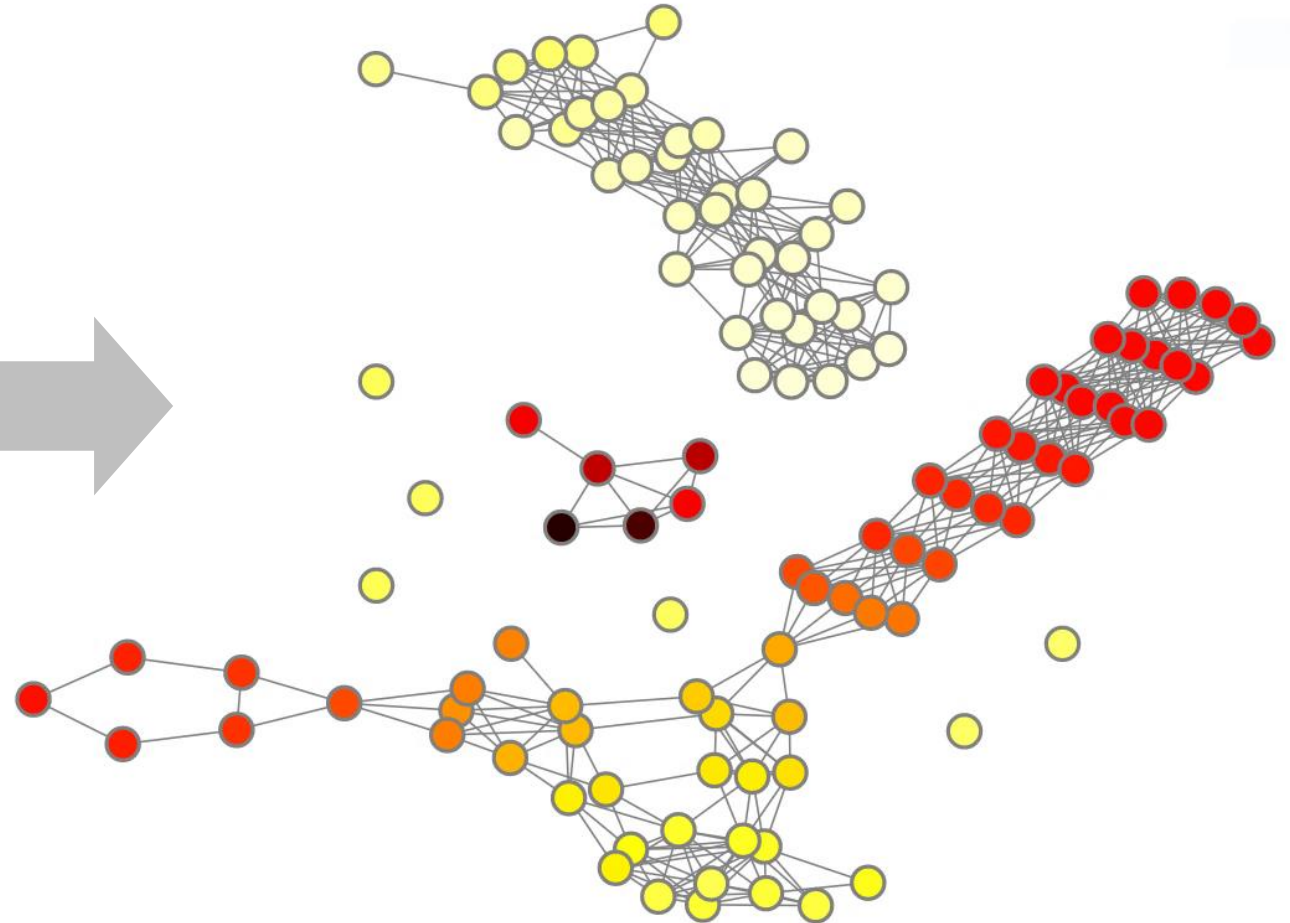


Number of sports



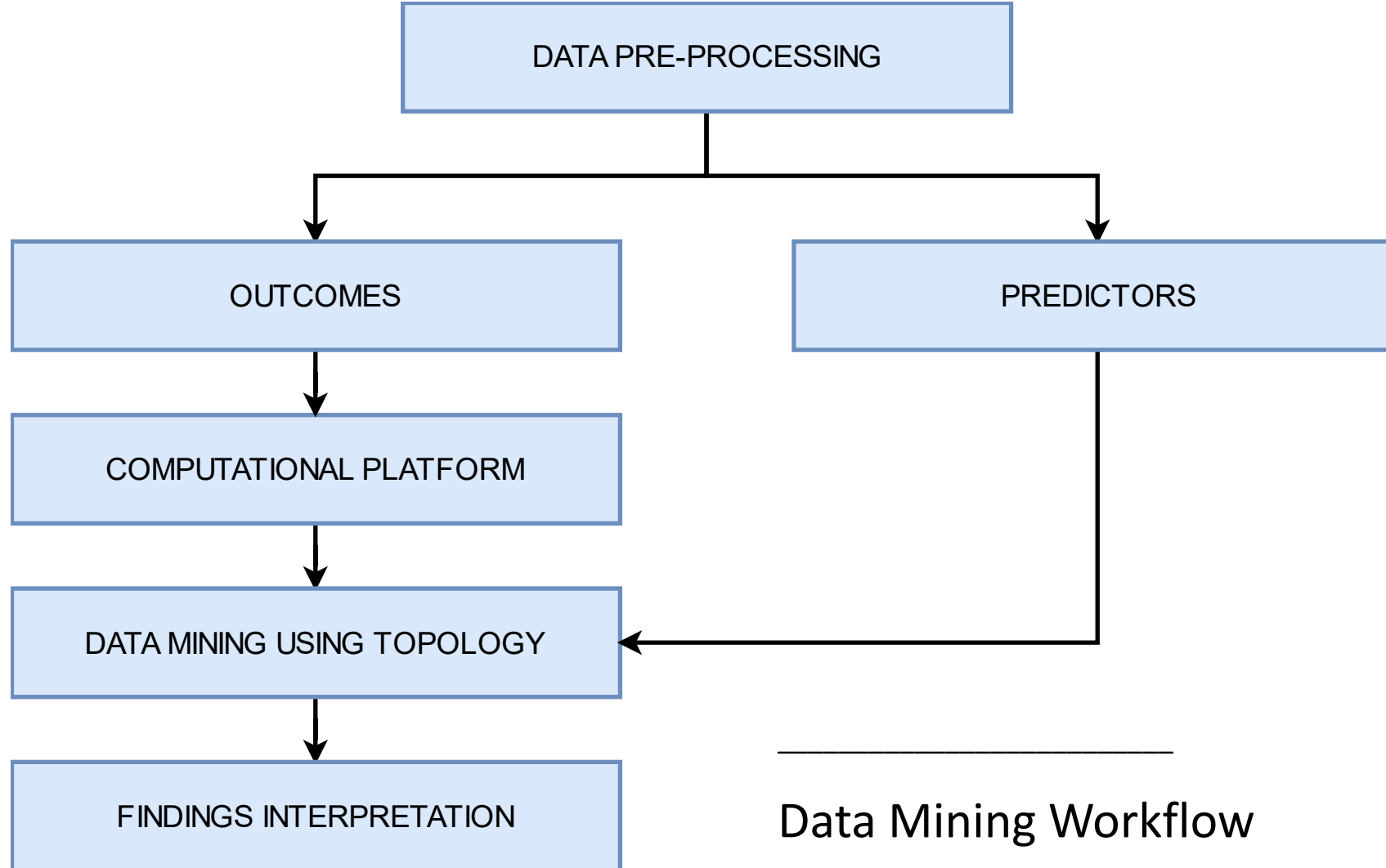
**1) Each node
represents one person**

**TCDM
algorithm**



**2) Two nodes are connected
if they are similar in terms of outcomes**

live demo



Data Mining Workflow

Workflow involves data preprocessing, automated generation of topological data maps, visual inspection of interesting features, and statistical analysis of emerging patterns



PART II.

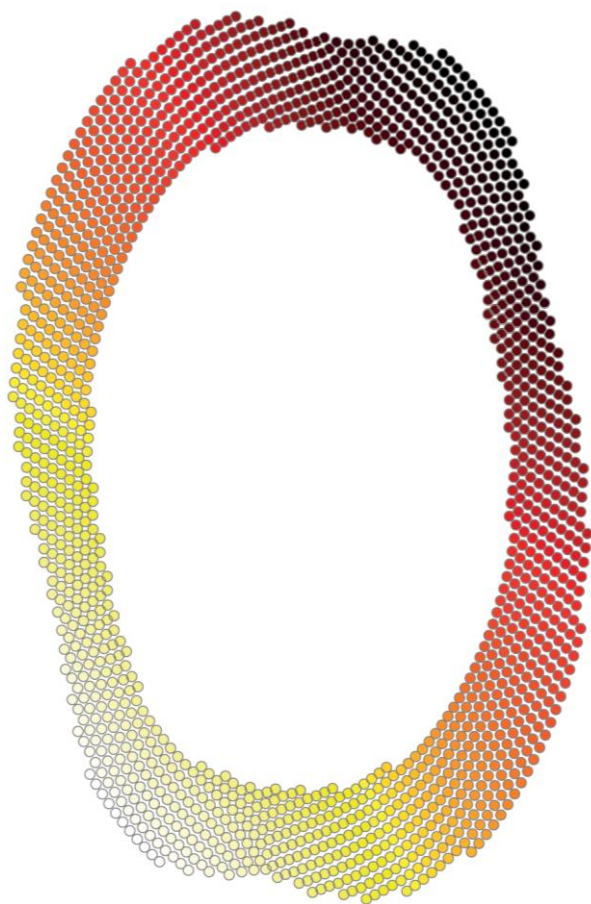
**HANDLING MISSING DATA USING
DATA GEOMETRY**



KEY STATEMENTS ABOUT TOPOLOGICAL DATA ANALYSIS (TDA):

- 1) TDA zooms in on robust patterns in data that are not affected by random noise, and therefore topological data maps extracted from a clinical dataset should not experience significant transformation if some data were missing at random
- 2) This focus on topological properties helps identify data geometry of a clinical dataset and determine whether missing values filled in by imputation methods are similar to the data geometry of the original dataset

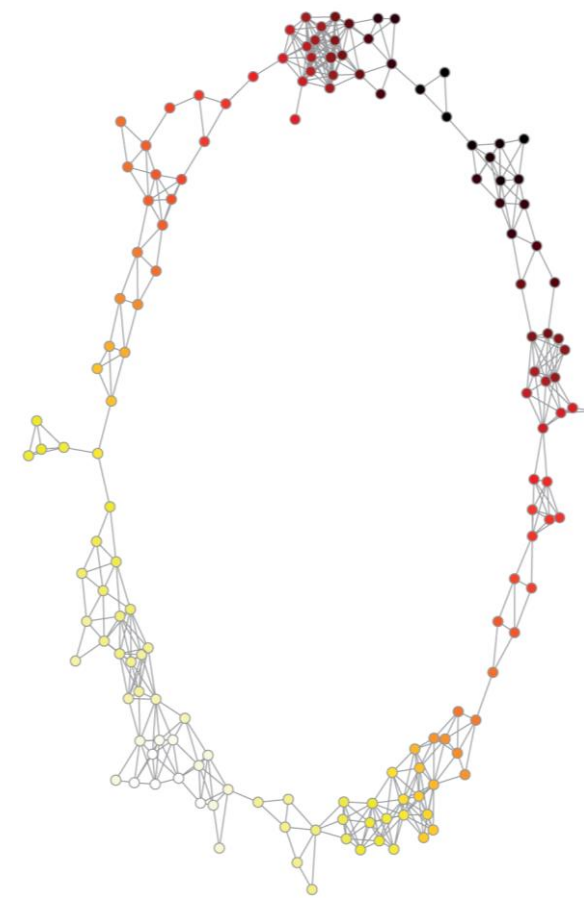
ROBUST GEOMETRIC PROPERTIES OF DATASET



No missing data



50% missingness



90% missingness

The Panel on Handling Missing Data in Clinical Trials under authority of FDA

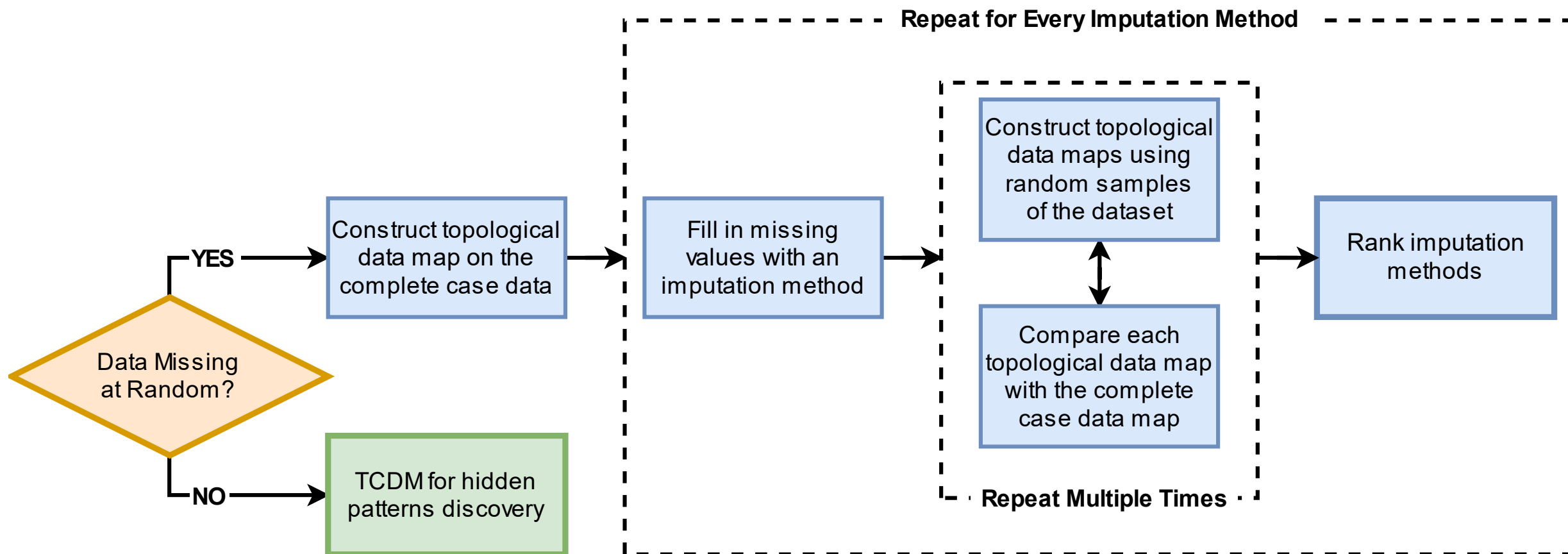
RECOMMENDATIONS:

- 1) Use of advanced imputations methods is highly encouraged
- 2) Development of robust methodologies and analytic tools is very important
- 3) Don't use single imputation methods such as LOCF and BOCF

REALITY:

- 1) Advanced methods are complicated and time consuming
- 2) Not many popular analytic tools are available (especially visualization)
- 3) LOCF is the most commonly used imputation method

WORKFLOW FOR EVALUATION OF DATA IMPUTATION METHODS



CASE STUDY: Treatment Effectiveness Assessment Questionnaire (TEAHL)

The TEAHL questionnaire asks the patient to express the extent of changes for the better from his/her involvement in the clinical study. For each area, the patient is asked to think about how things have become better and to circle the results on a 10-point scale.

TEAHL dataset:

- 302 patients with 5 observations for each patient
- 248 patients had no missing values (complete case)
- 104 out of 1510 values in total were missing

Clinical study that was conducted with the support of the NIDA Clinical Trials Network.
Study data are available at <https://datashare.nida.nih.gov/study/nida-ctn-0048>

THREE DATA IMPUTATION METHODS WERE COMPARED

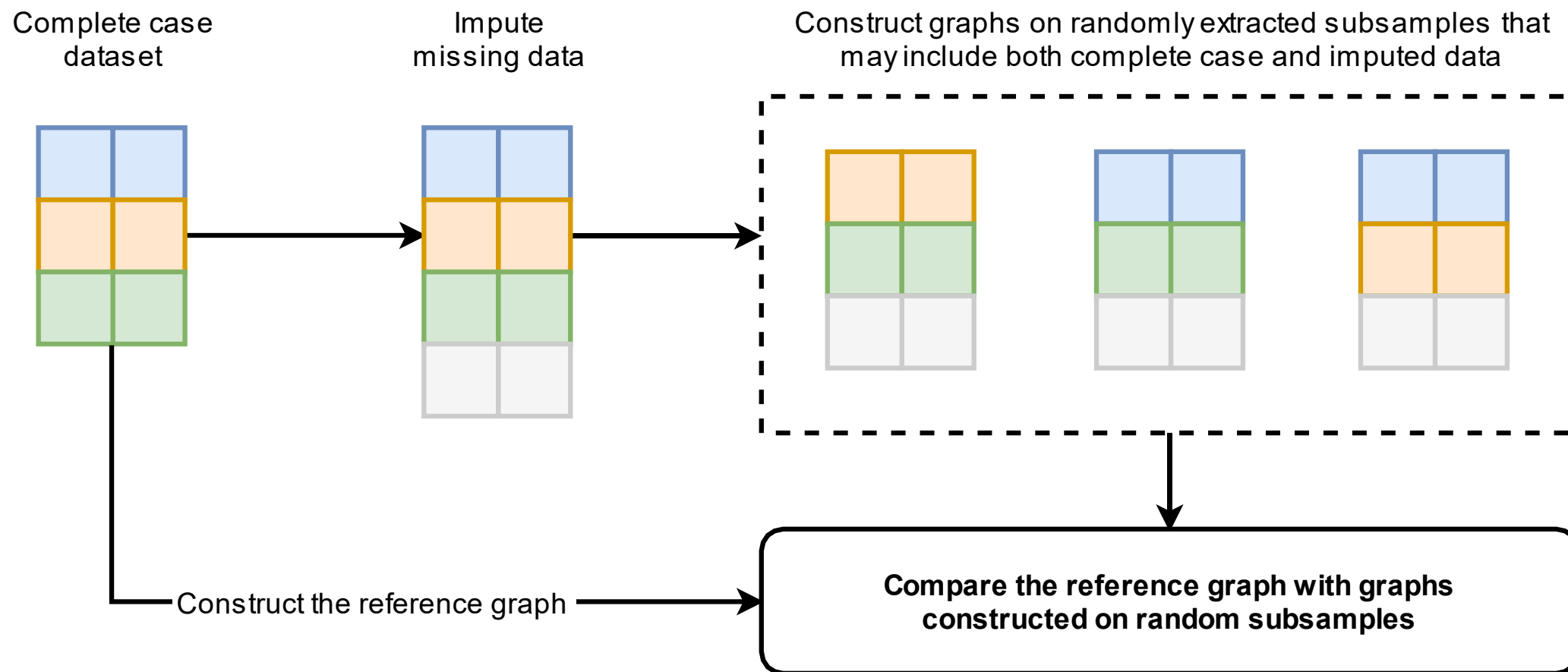
- 1) MICE
Multivariate Imputation by Chained Equations
- 2) k-NN
K-Nearest Neighbor Imputation Algorithms
- 3) LOCF
Last Observation Carried Forward

More information is available:

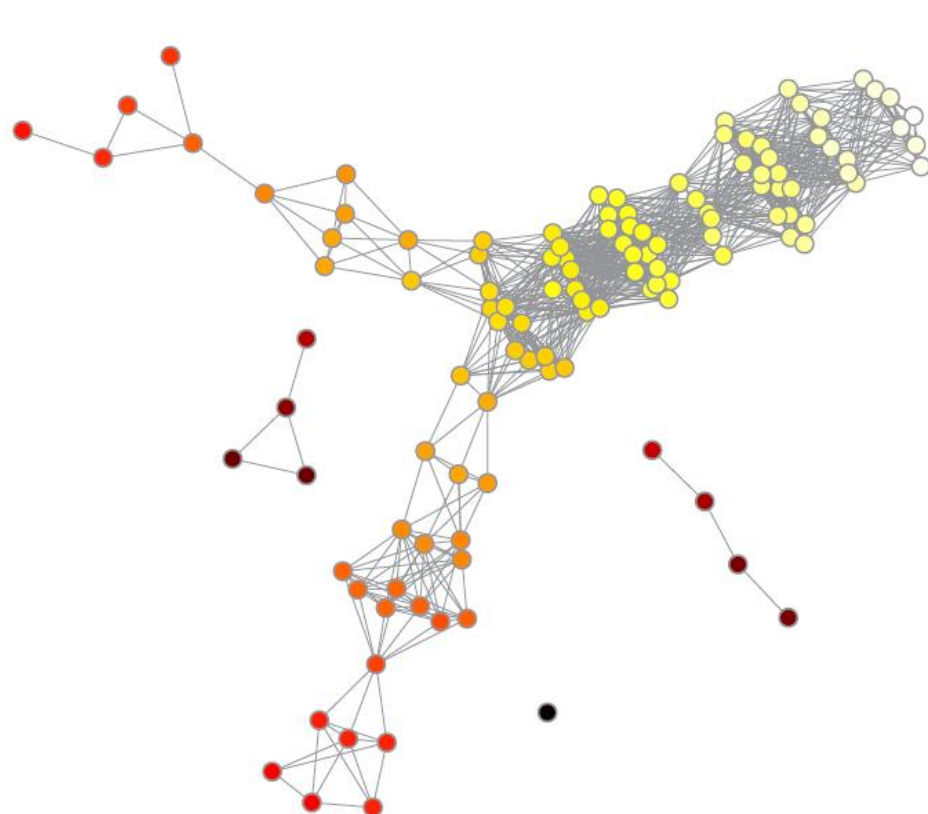
- MICE: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/>
- k-NN: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4959387/>
- LOCF: <https://www.ncbi.nlm.nih.gov/pubmed/19947800>



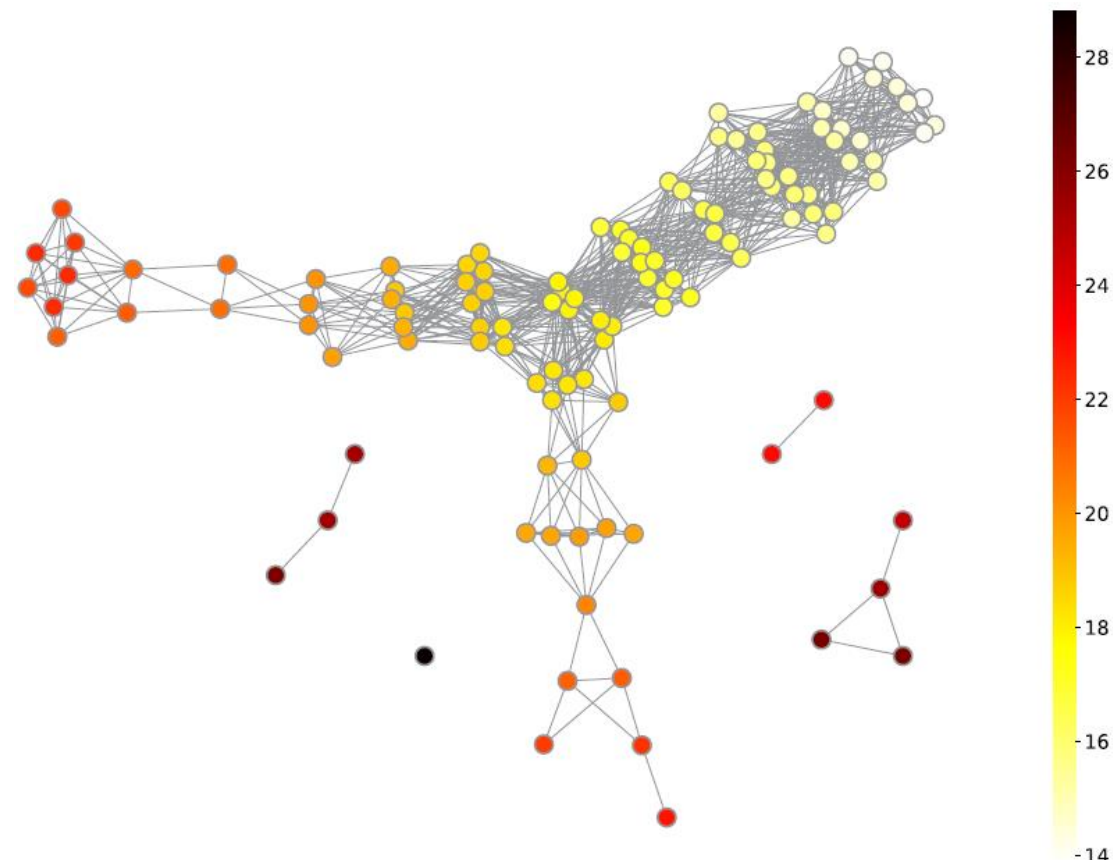
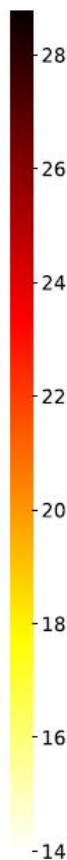
GENERATION OF SUBSAMPLES WITH IMPUTED VALUES



COMPLETE CASE VERSUS MICE METHOD

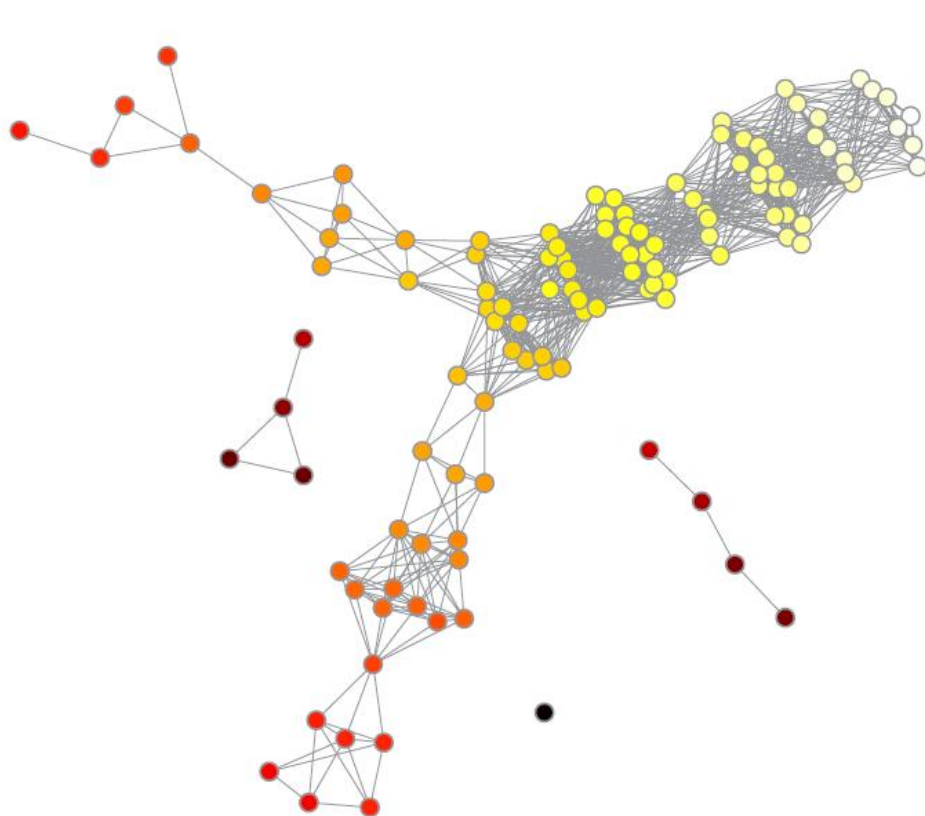


Reference graph (complete case data)

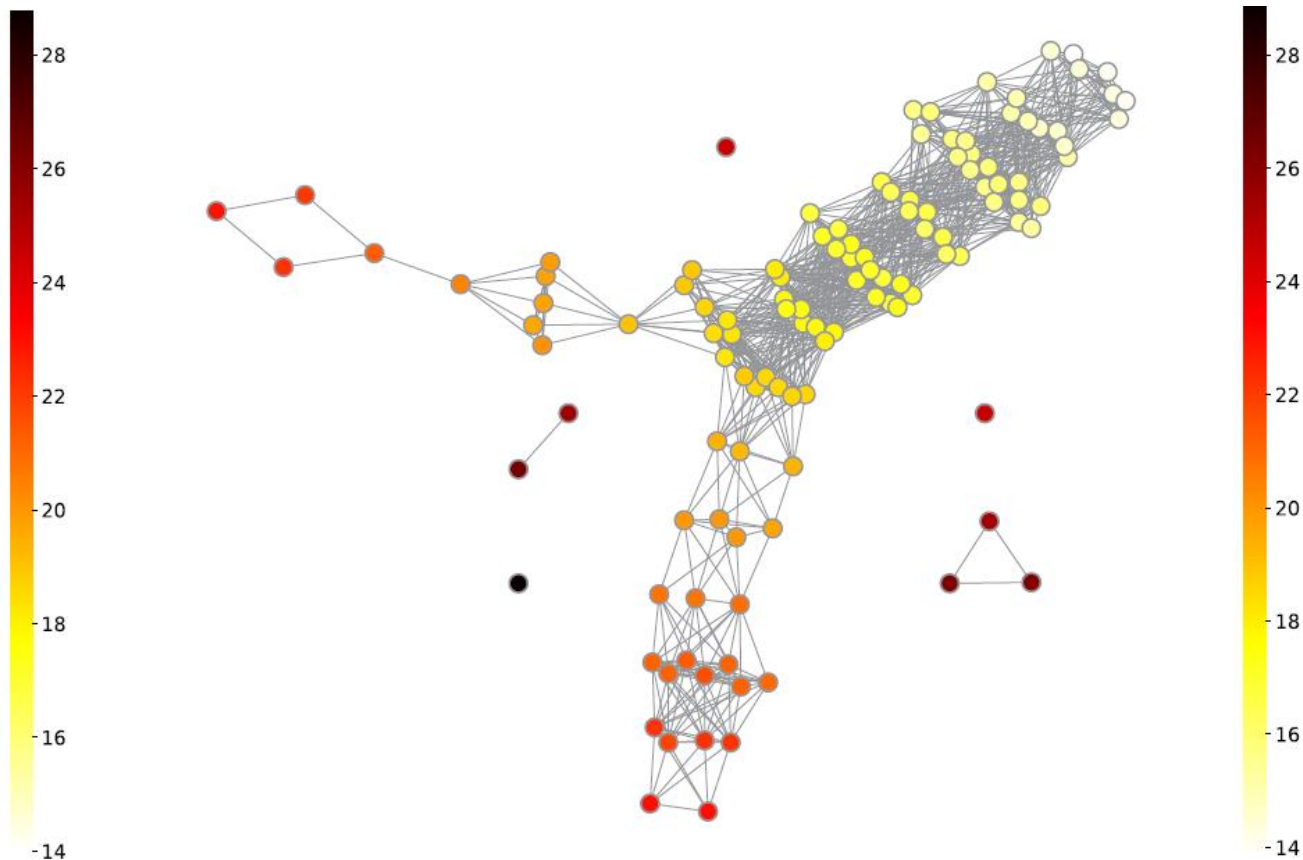


MICE imputation method

COMPLETE CASE VERSUS k -NN METHOD

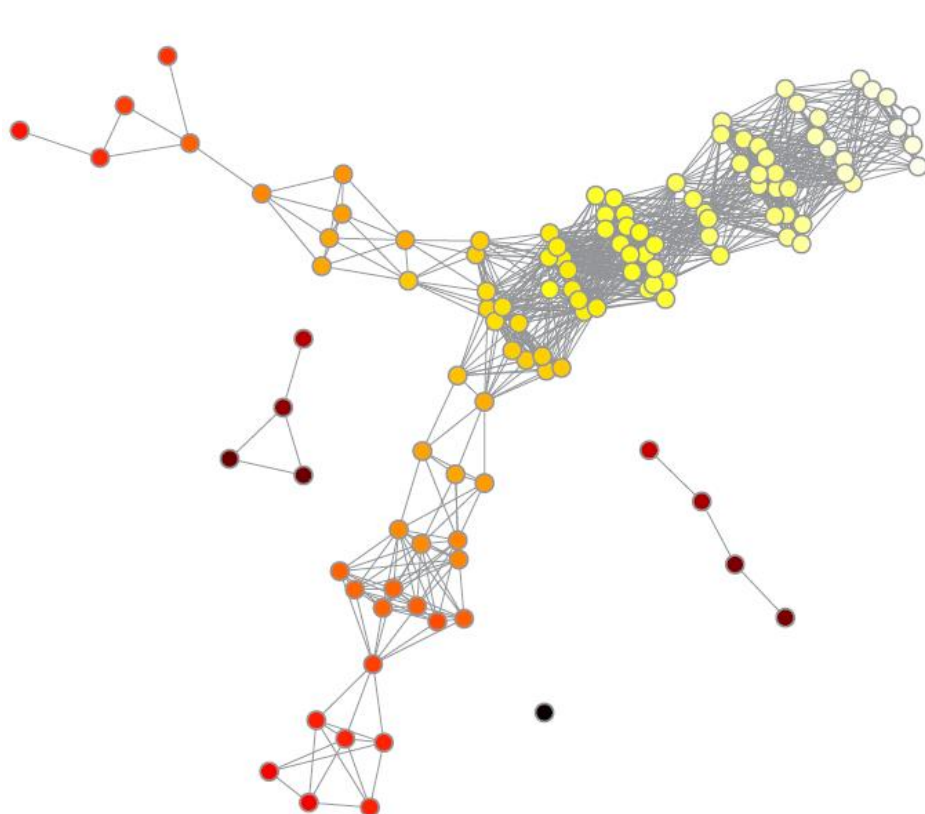


Reference graph (complete case data)

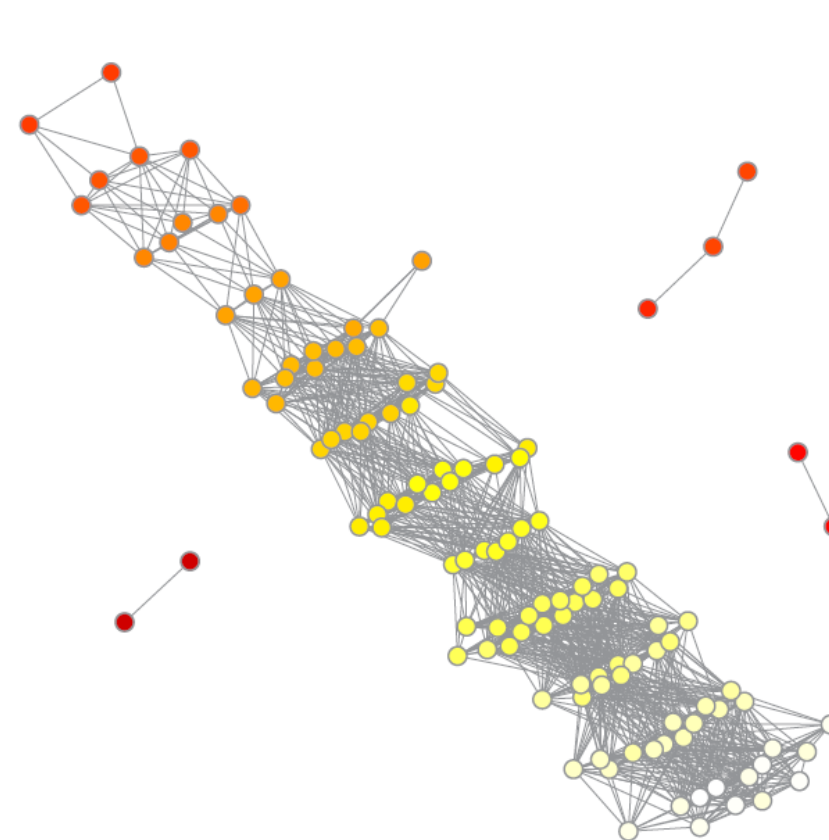
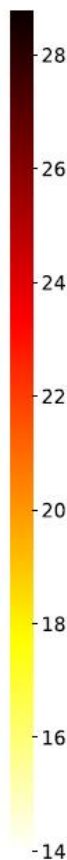


k -NN imputation method ($k = 5$)

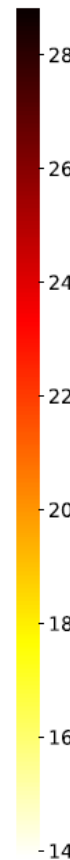
COMPLETE CASE VERSUS LOCF METHOD



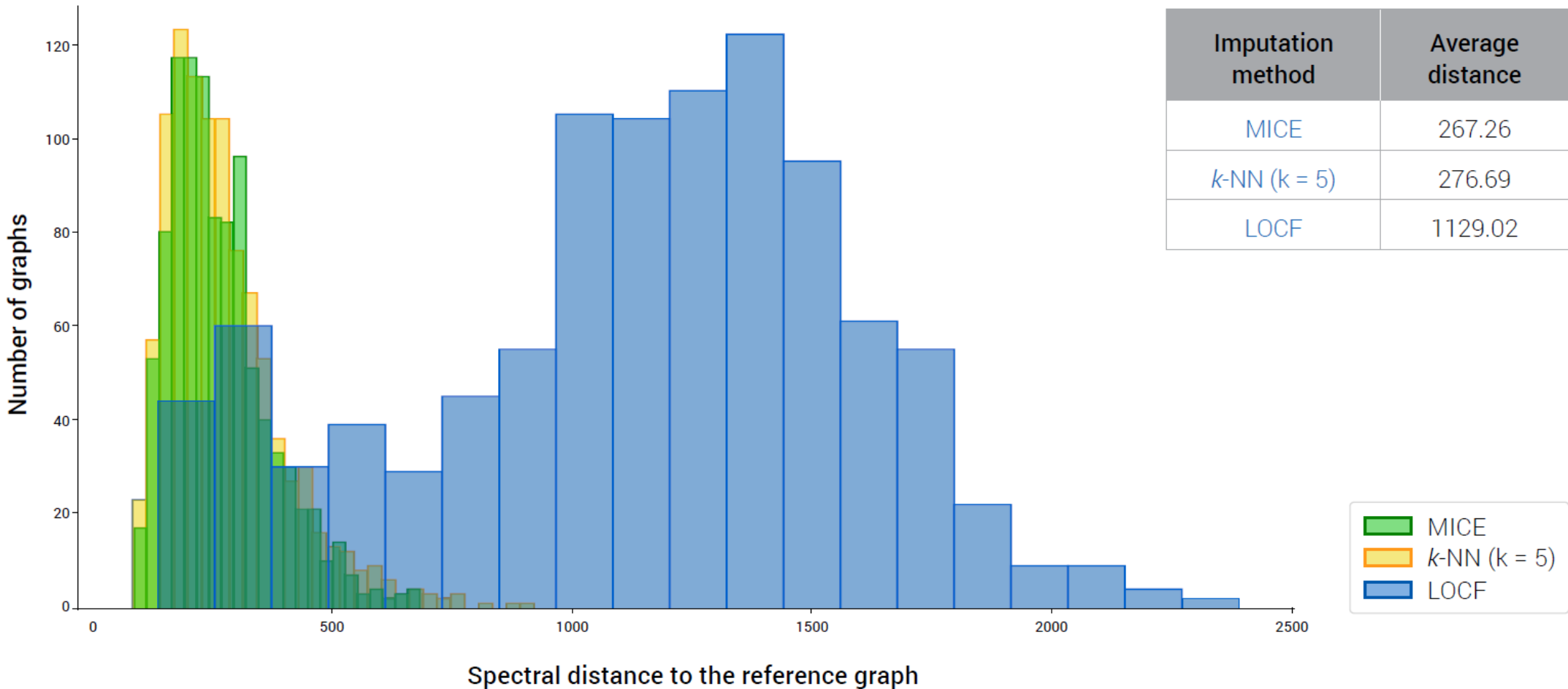
Reference graph (complete case data)



LOCF imputation method



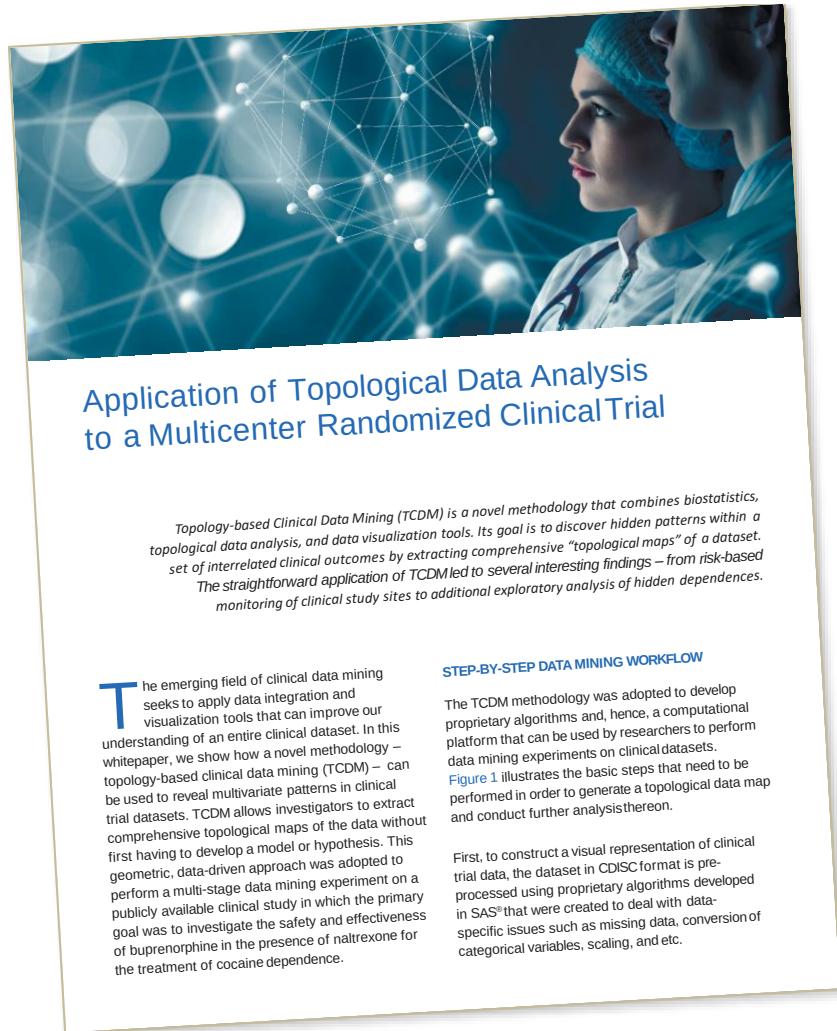
DISTRIBUTION OF DISTANCES TO THE REFERENCE GRAPH



PRESENTATION TAKEAWAYS:

- TDA allows visualize clinical dataset by extracting topological data maps
- Robust data maps can help select imputation method based on geometry of the dataset
- We can automatically evaluate the effectiveness of each imputation method for a specific dataset

WHITEPAPERS AVAILABLE



THANK YOU

*Intego Group, LLC
555 Winderley Place, Ste. 129,
Maitland, FL 32751
Phone: +1 (407) 512-1006
www.intego-group.com/tcdm*

