



INTEGO
GROUP

EXPLORING CATEGORICAL DATA USING UNSUPERVISED GRAPH-BASED MACHINE LEARNING



US Connect 2021



Kostiantyn Drach, PhD

Researcher, Aix-Marseille University, Marseille, France / Intego Group, Kharkiv, Ukraine

Kostiantyn graduated from Kharkiv National University with a Master's degree in Mathematics. He obtained his Ph.D. in Mathematics with a focus on geometry and topology from the Institute for Low Temperature Physics in Kharkiv, Ukraine. Over the last several years, he has participated in various research projects in pure and applied mathematics. Kostiantyn is a core external member of an R&D team of data scientists studying clinical data using methods of applied topology.

Iryna Kotenko

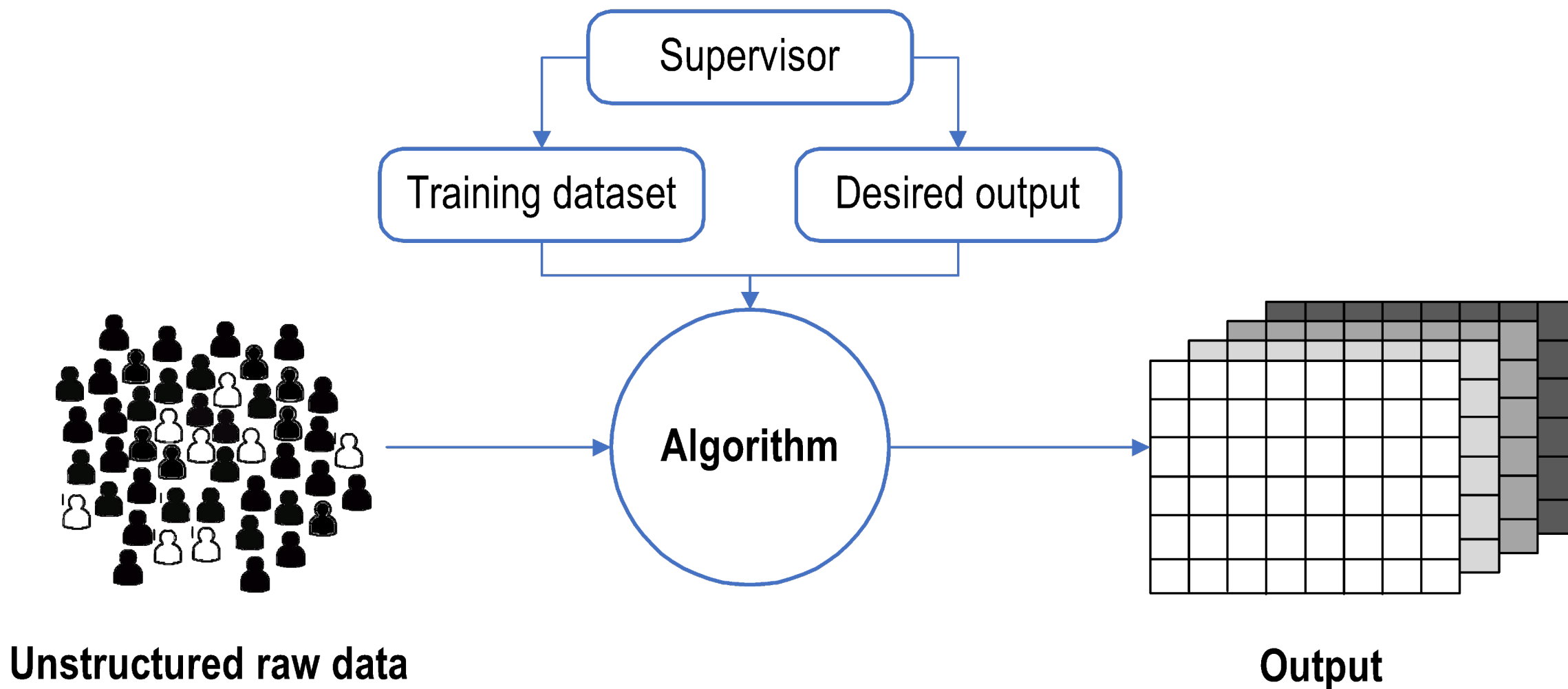
Site Lead, Intego Group, Kharkiv, Ukraine

Iryna graduated with a Master's in Applied Mathematics from Kharkiv Polytechnic University. Her fourteen-year career as a statistical programmer spans numerous CROs as a clinical SAS programmer and data analyst. Over the last eight years, she has held the position of Site Lead for Clinical Programming Services, where she leads a group of over 150 statistical programmers in Ukraine. She is also heavily involved in a partnership program with Kharkiv National University.

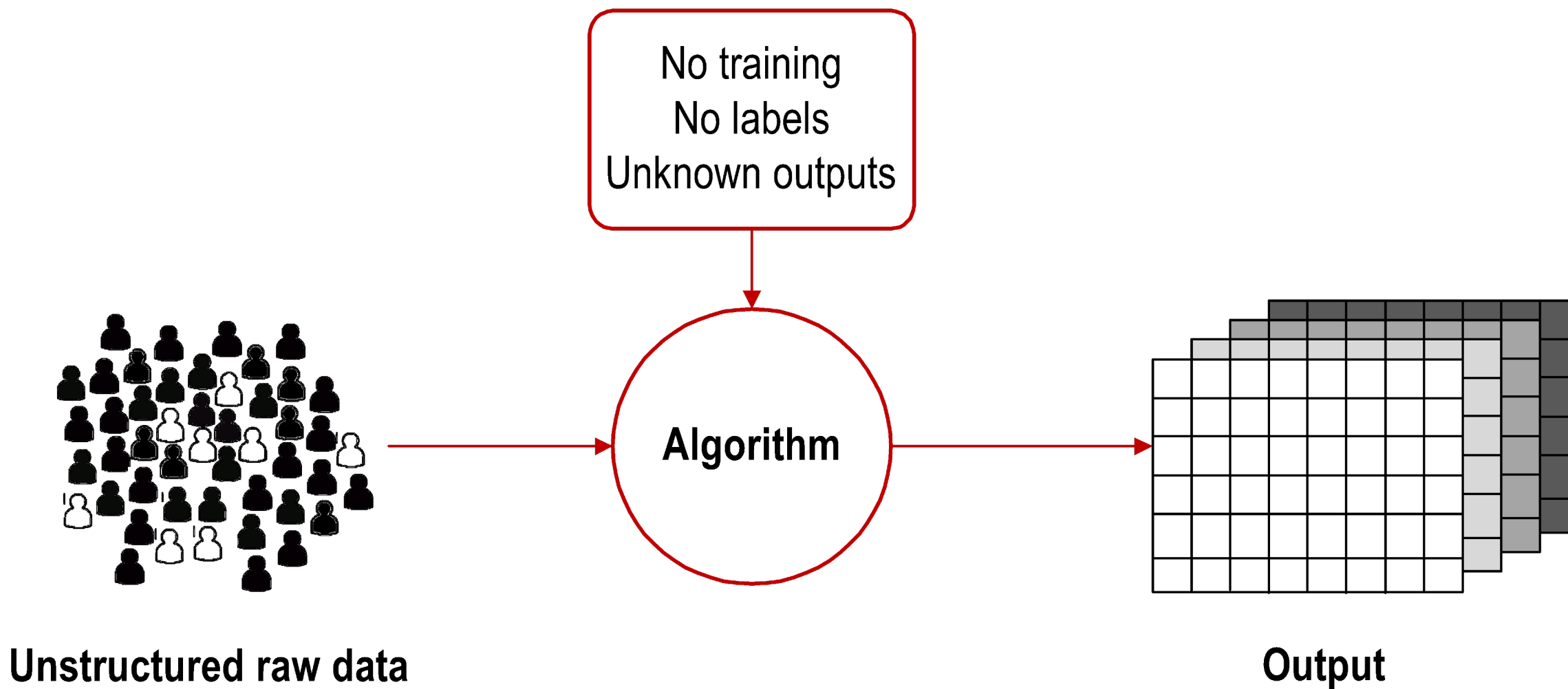
Today we'll cover:

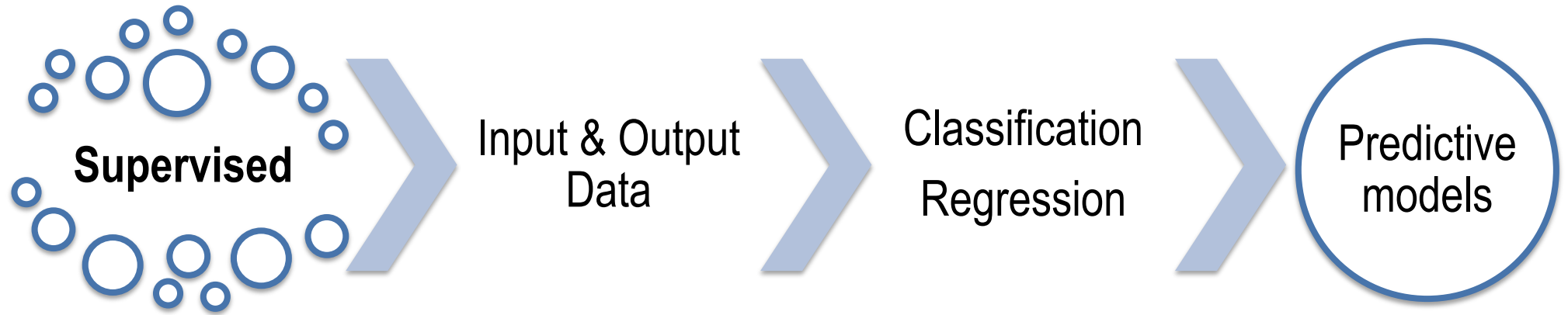
- I. Unsupervised vs. supervised machine learning
- II. Introduction to Topological Data Analysis
- III. Analysis of categorical data in clinical trials: binary logistic regression
- IV. Experiment with a large dataset of categorical data: setup, modeling, comparison of methods

Supervised Machine Learning



Unsupervised Machine Learning

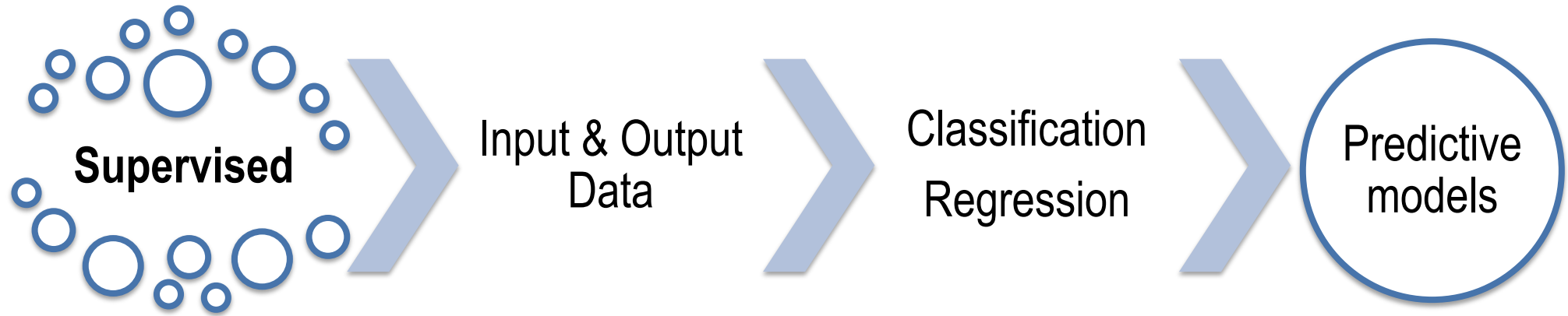




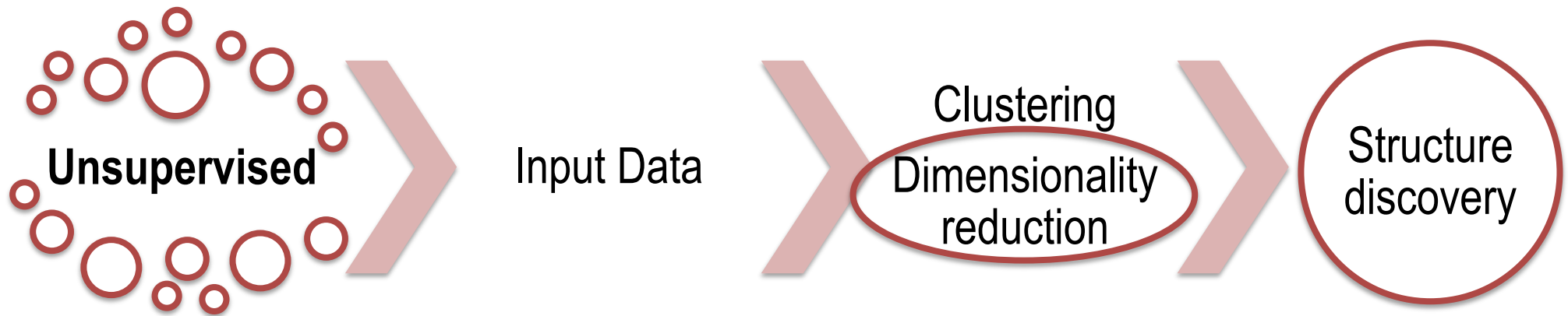
In supervised learning historical input and output data is required to develop predictive models by making iterative predictions



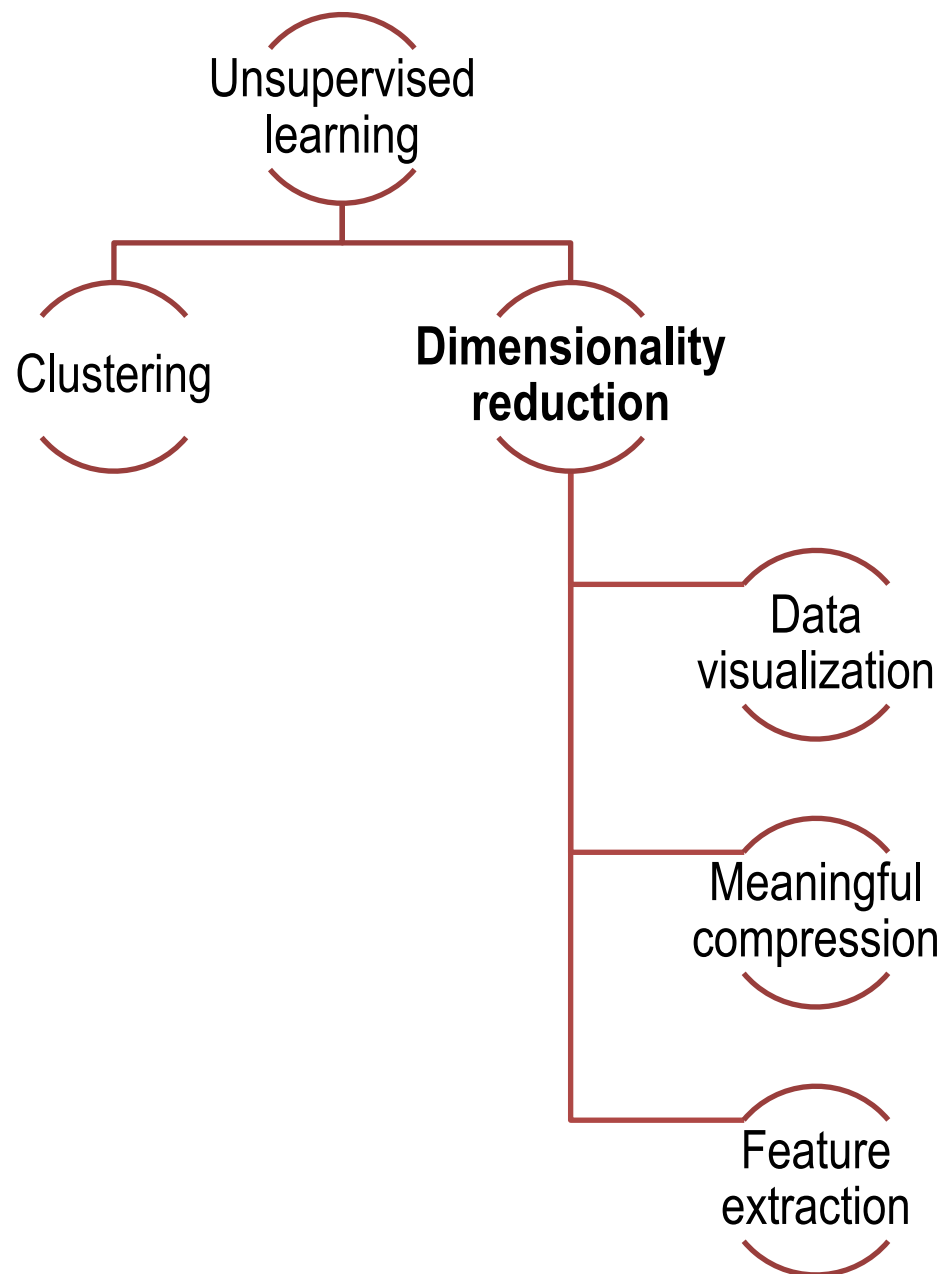
In unsupervised learning only input data is used to discover hidden structure of the dataset and generate new hypothesis



In supervised learning historical input and output data is required to develop predictive models by making iterative predictions

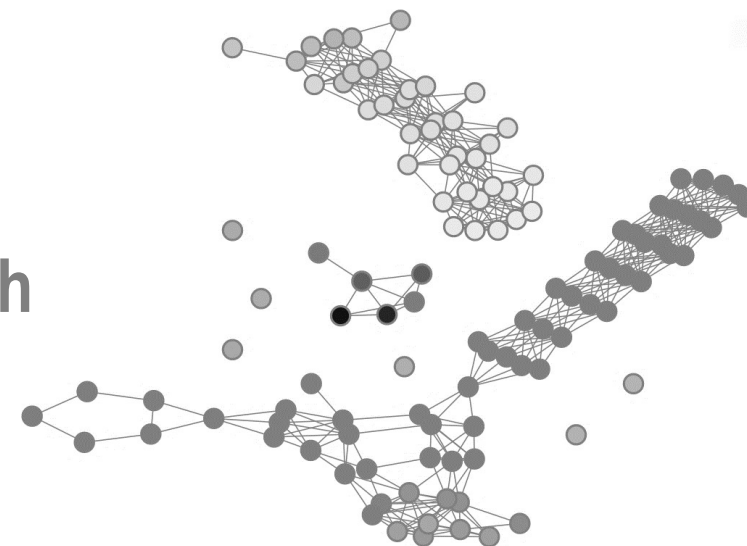


In unsupervised learning only input data is used to discover hidden structure of the dataset and generate new hypothesis

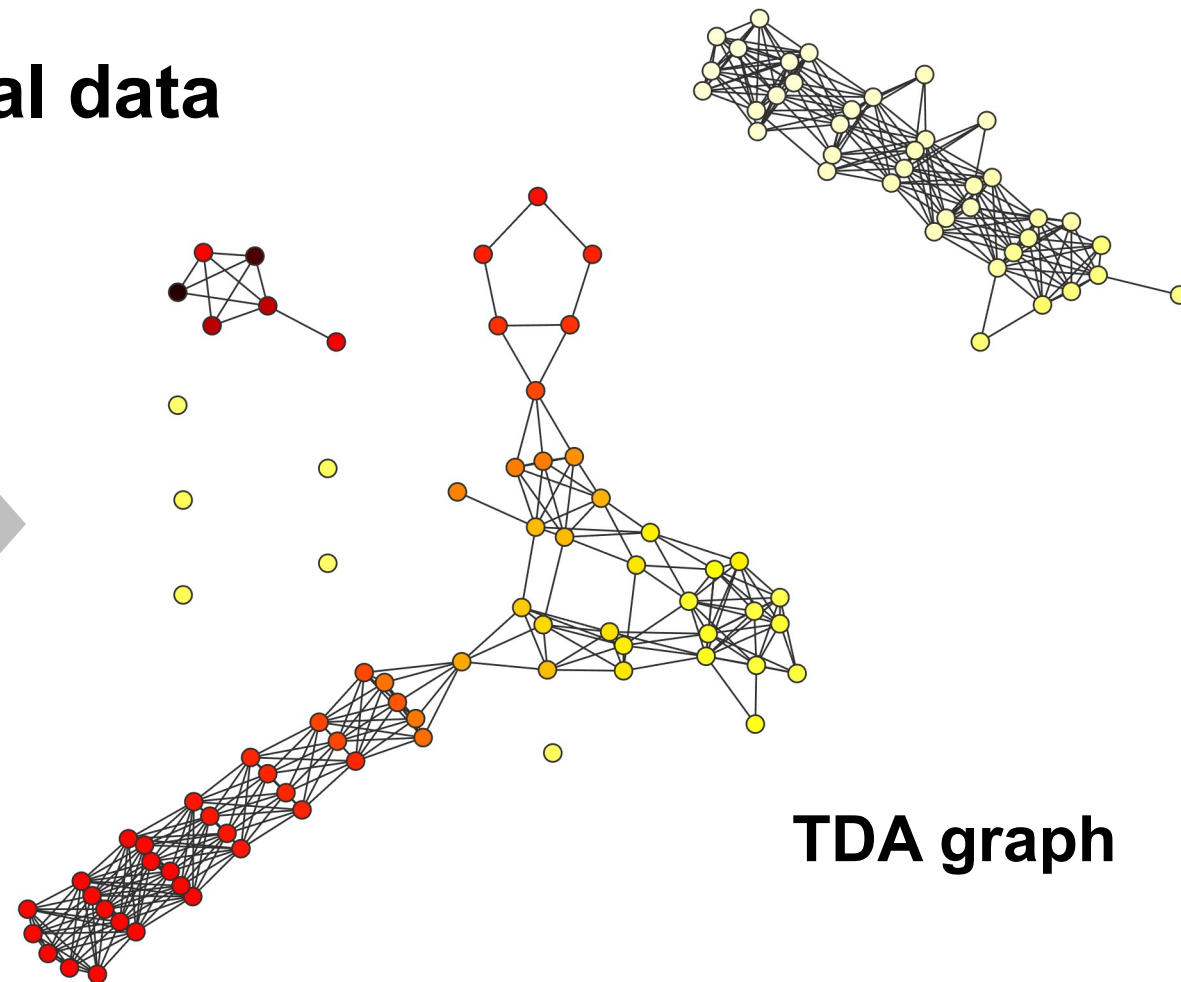


Topological Data Analysis (TDA) – modern approach for data analysis based on geometric properties of the dataset

TDA graph



Topological Data Analysis for clinical data



TDA graph

Unstructured group of
clinical trials (**subjects**)

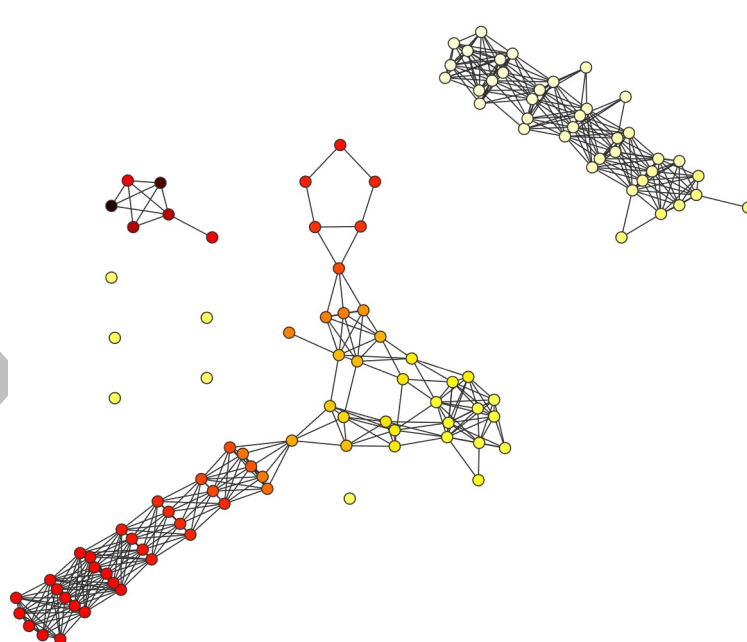
Each **node** represents one subject.

Two nodes are connected if they are **similar** in
terms of predefined **features**

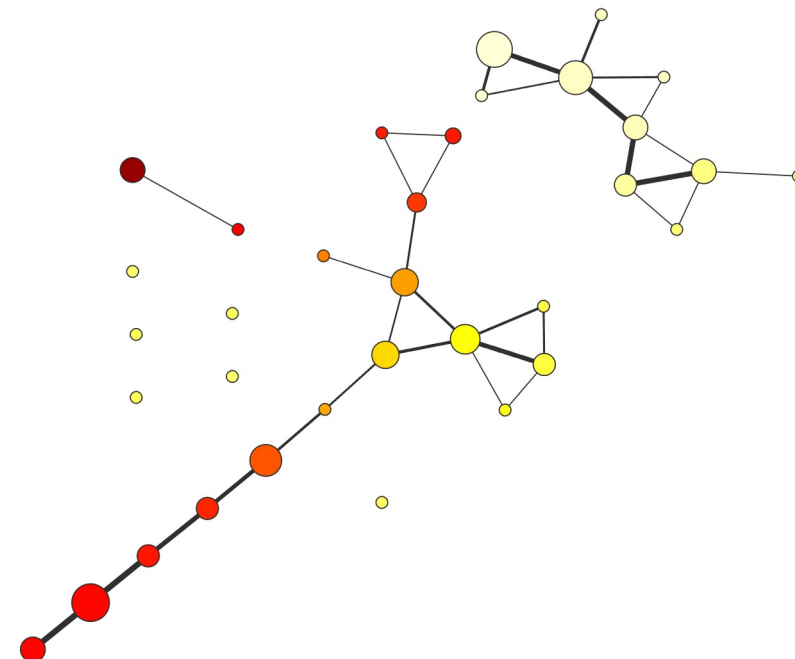
Compressing TDA graph



Data



Single-node
TDA graph



In a **compressed** TDA graph, each node represents a **group of subjects** with similar features.

TDA algorithm can deal with numerical and categorical outcomes

Interrelated biomarkers evaluations

E.g. patients' vital signs or basic metabolic panel's results on a specific day of study

Series of repeated measurements

E.g. weekly hemoglobin levels during chemotherapy in oncological patients

Questionnaire data

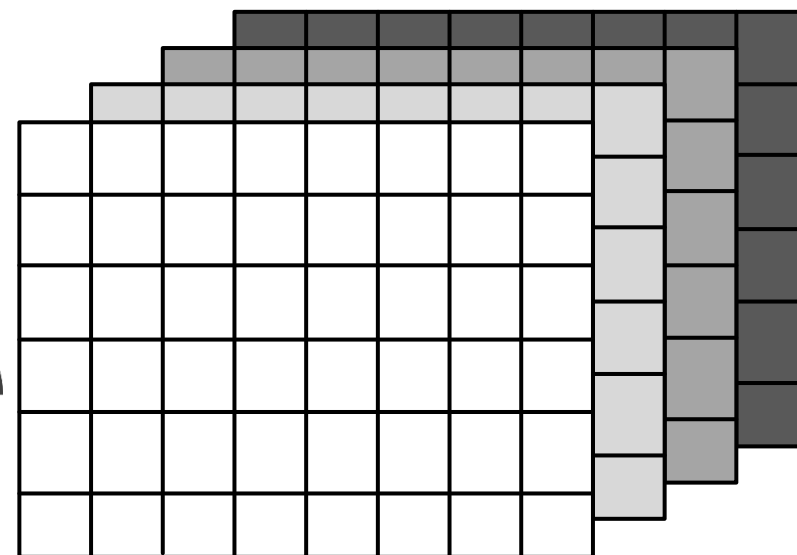
Binary or ordinal responses to the items of a questionnaire, aggregate scores



Outcomes



Patients



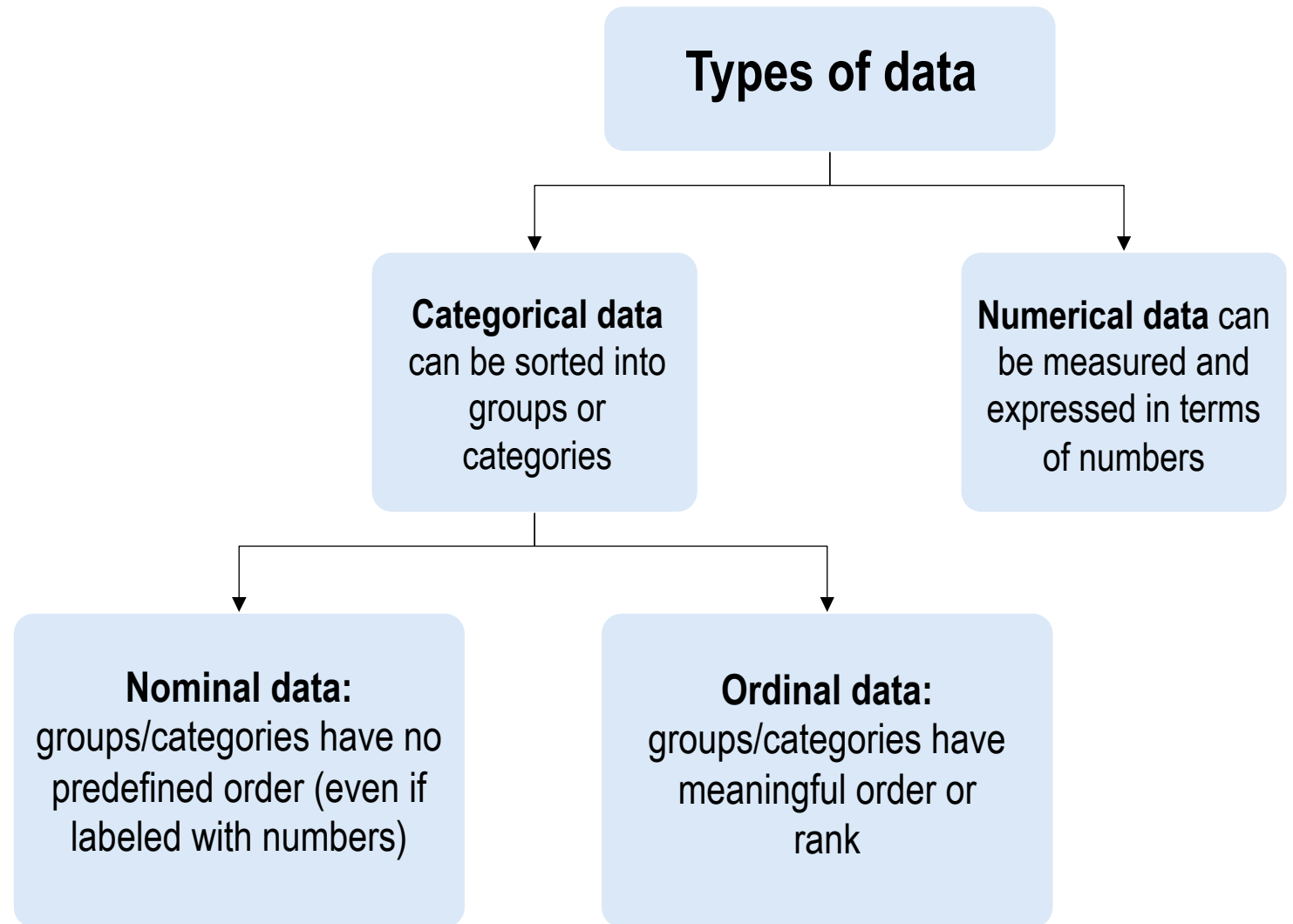
Standard statistical approach vs. TDA



Standard statistical approach	Outcome variables are studied separately	Requires certain assumptions or a data model	Hypotheses testing for pre-specified subgroups
Topological Data Analysis	Allows analysis of multiple interrelated outcomes	Assumption-free and model-independent	Discovery of subgroups of subjects with similar outcomes

Types of data

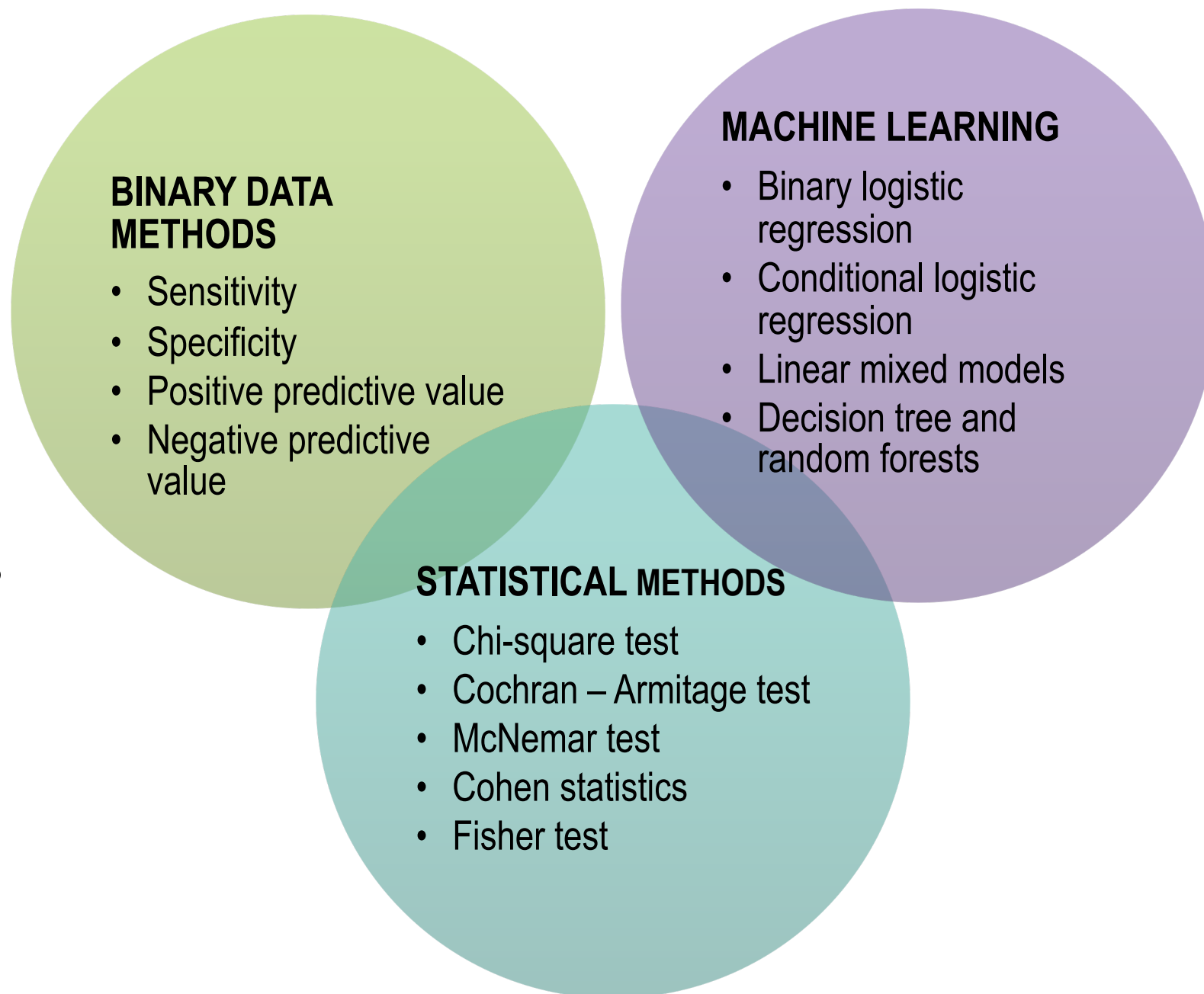
- Type of data is an essential aspect for statistical analysis
- Must be understood in order to apply correct statistical methods
- Main types:
 - *numerical data* (**quantitative**)
 - *categorical data* (**qualitative**)



Key challenges with categorical data:

- Could potentially have too many levels that pull down the performance of the model
- Rarely occurred values may bias the results
- Frequently occurred values fail to make a positive impact on the model
- Most algorithms (ex. regression) require data pre-processing and can't fit raw data into the model

Examples of methods used for analysis of categorical data in clinical trials



Goals of the experiment:

- Explore a large publicly available dataset (health survey, about 27,000 respondents) with categorical data (questionnaires)
- Perform logistic regression analysis to assess the impact of various predictors on the selected outcomes
- Apply unsupervised learning (TDA-algorithm):
 - to confirm logistic regression results
 - to reveal hidden patterns undetected by the regression analysis
- Compare results and discuss advantages and disadvantages of both approaches

Dataset for the experiment

2019 National Health Interview Survey by CDC

<https://www.cdc.gov/nchs/nhis/2019nhis.htm>

Data specification

- Insurance coverage data
- Demographic and social data
- Selected outcomes:
 - Whether the respondents are worried that they will be able to pay the medical bills or not.
 - Whether the respondents have any problems with paying for health care or prescription medication.

Methods of analysis

- Binary logistic regression
- Topological Data Analysis (TDA)
- Statistical analysis

National Health Interview Survey

About NCHS

The National Center for Health Statistics (NCHS) is the nation's principal health statistics agency, providing data to identify and address health issues. NCHS compiles statistical information to help guide public health and health policy decisions.

Collaborating with other public and private health partners, NCHS uses a variety of data collection mechanisms to obtain accurate information from multiple sources. This process provides a broad perspective to help understand the population's health, influences on health, and health outcomes.

National Health Interview Survey

The National Health Interview Survey (NHIS) provides information on the health of the U.S. civilian noninstitutionalized population through confidential interviews conducted in households. NHIS is one of the nation's largest in-person household health surveys. It provides data for analyzing health trends and tracking progress toward achieving national health objectives.

These data, continuously collected throughout the year, are also used for epidemiological and policy analysis, such as characterizing those with various health conditions, determining barriers to accessing and using appropriate health care, and evaluating federal health programs.

Major health topics addressed

The NHIS questionnaire was redesigned in 2019 to increase relevance, enhance data quality, and minimize respondent burden. Topics covered in 2020 include:

- Physical and mental health status
- Chronic conditions, including asthma and diabetes
- Measures of functioning and disability
- Access to and use of health care services
- Health insurance coverage and type of coverage
- Chronic pain, pain management, and prescription opioid use
- Demographics and social determinants of health

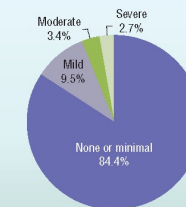
Examples of NHIS data

Mental Health

Data from 2019 on symptoms of anxiety and depression in adults show:

- The percentage of adults aged 18 and over who experienced mild, moderate, or severe symptoms of anxiety in the past 2 weeks was 9.5%, 3.4%, and 2.7%, respectively.
- The percentage of adults aged 18 and over who experienced any severity of symptoms of anxiety in the past 2 weeks decreased with age, were higher among women (19.0%) than men (11.9%), and were higher among non-Hispanic white adults (16.5%) compared with non-Hispanic black (14.6%), Hispanic (14.5%), and non-Hispanic Asian (8.5%) adults.

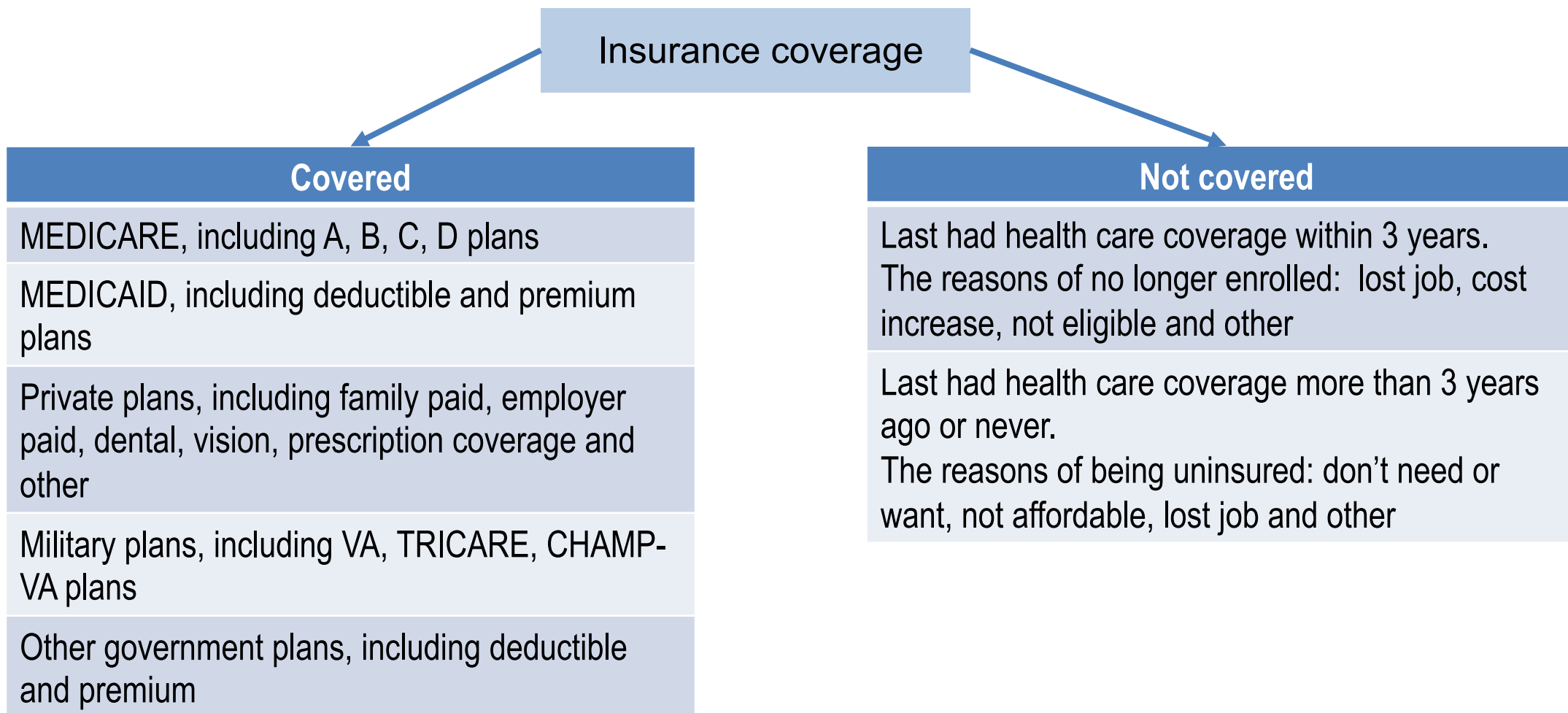
Percent distribution of severity of anxiety symptoms in the past 2 weeks among adults aged 18 and over: United States, 2019



NOTES: Severity of anxiety symptoms was based on the seven-item Generalized Anxiety Disorder scale (GAD-7) and summarized into none or minimal (values 0-4), mild (values 5-9), moderate (values 10-14), and severe (values 15-21). Estimates are based on household interviews of a sample of the civilian noninstitutionalized population. Access data table for this figure at: <https://www.cdc.gov/nchs/data/behavioral/d375-tables-508.pdf>

SOURCE: National Center for Health Statistics, National Health Interview Survey, 2019.

Structure of the dataset



Categorical variables were converted into binary variables. Number of predictors: **52**

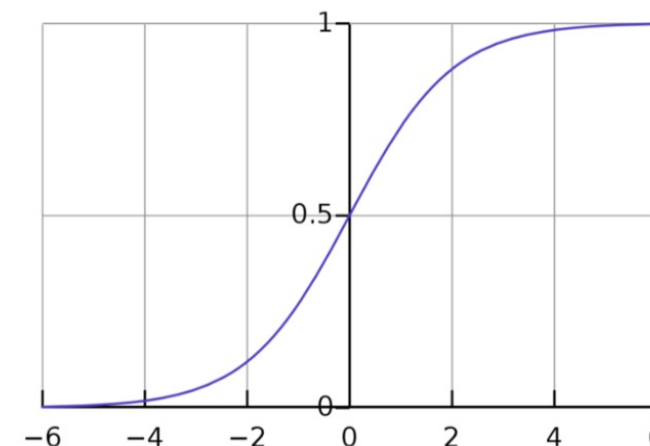
Binary logistic regression

The model is a *log-linear binary classifier*, with classes 0 and 1.
In the model,

$$P(\mathbf{y} = 1) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + w_2 x_2 \dots + w_n x_n)}}$$

is the probability that the binary *outcome* $\mathbf{y}$ equals 1.

Each *weight coefficient* w_k shows how much the *predictor* x_k contributes to the outcome $\mathbf{y}$.



The model in our experiment:

Assesses the impact of insurance coverage predictors on the following selected outcomes:

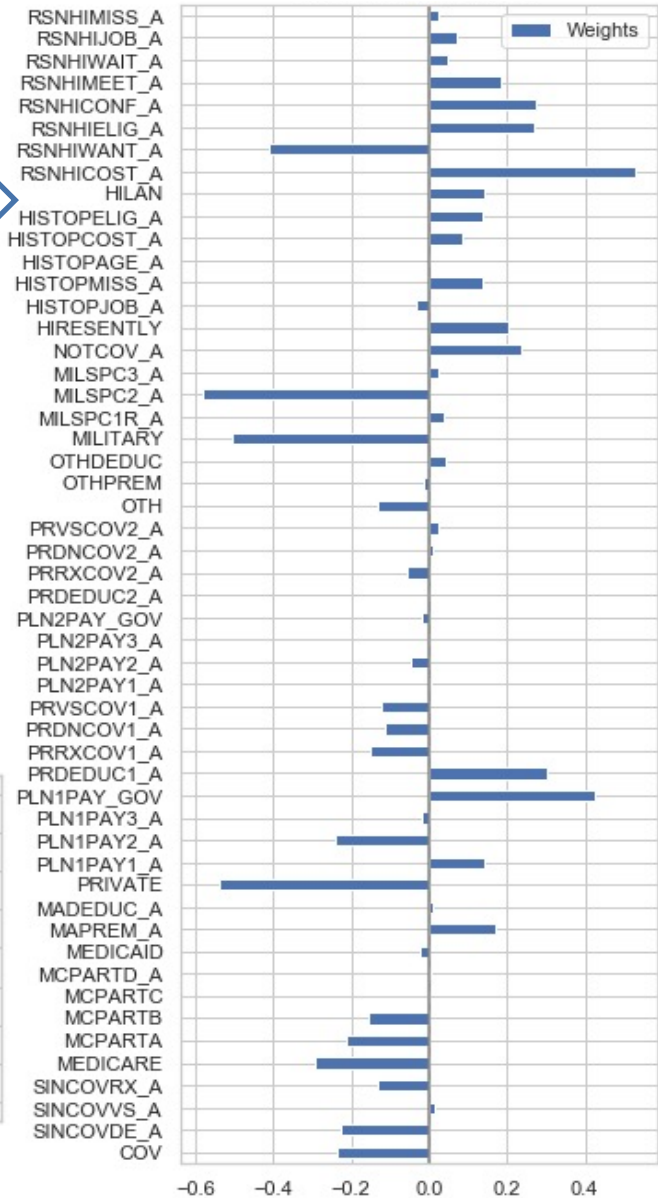
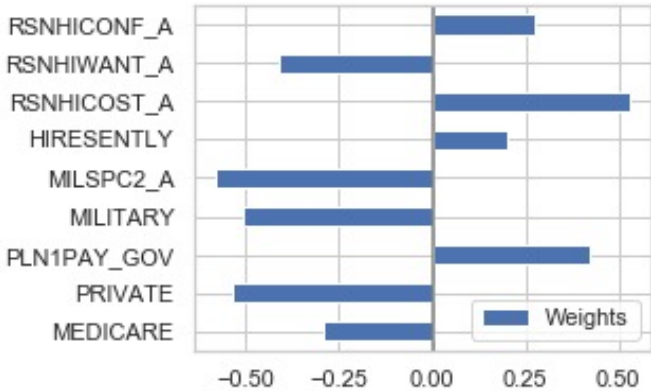
- **worry about paying** medical bills if get sick or have accident
- (likelihood of having) **problems with paying** for health care or prescription medication

Results of logistic regression

Worry about
paying medical
bills if get sick
or have
accident

All
weights

Significant
weights



RSNHICONF_A: no coverage, signing up for coverage too difficult or confusing
RSNHICOST_A: no coverage, coverage is not affordable
HIRECENTLY: no coverage, last had coverage within 3 years
PLN1PAY_GOV: private plan paid by Medicare, Medicaid or other government program



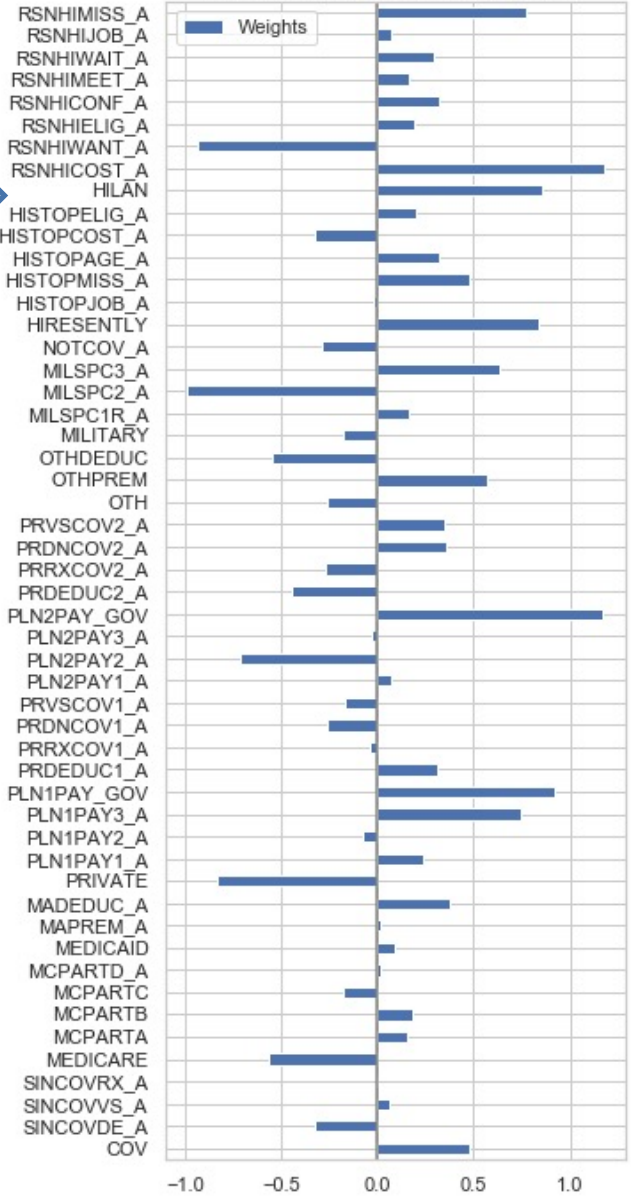
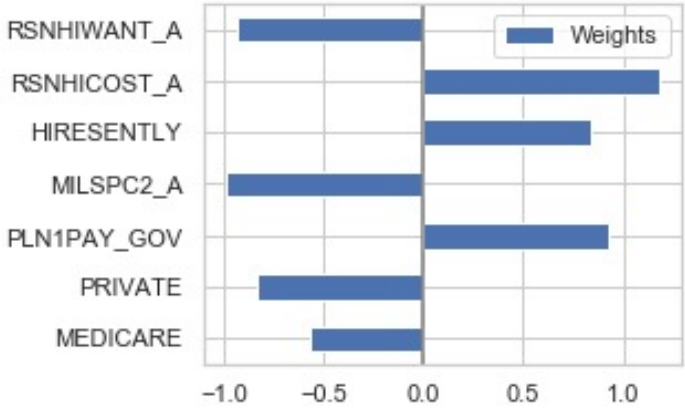
RSNHIWANT_A: do not need or want coverage
MILSPC2_A: military TRICARE health coverage
MILITARY: military health care coverage
PRIVATE: private health insurance
MEDICARE: Medicare insurance

Results of logistic regression

Problems with
paying for
health care or
prescription
medication

All
weights

Significant
weights



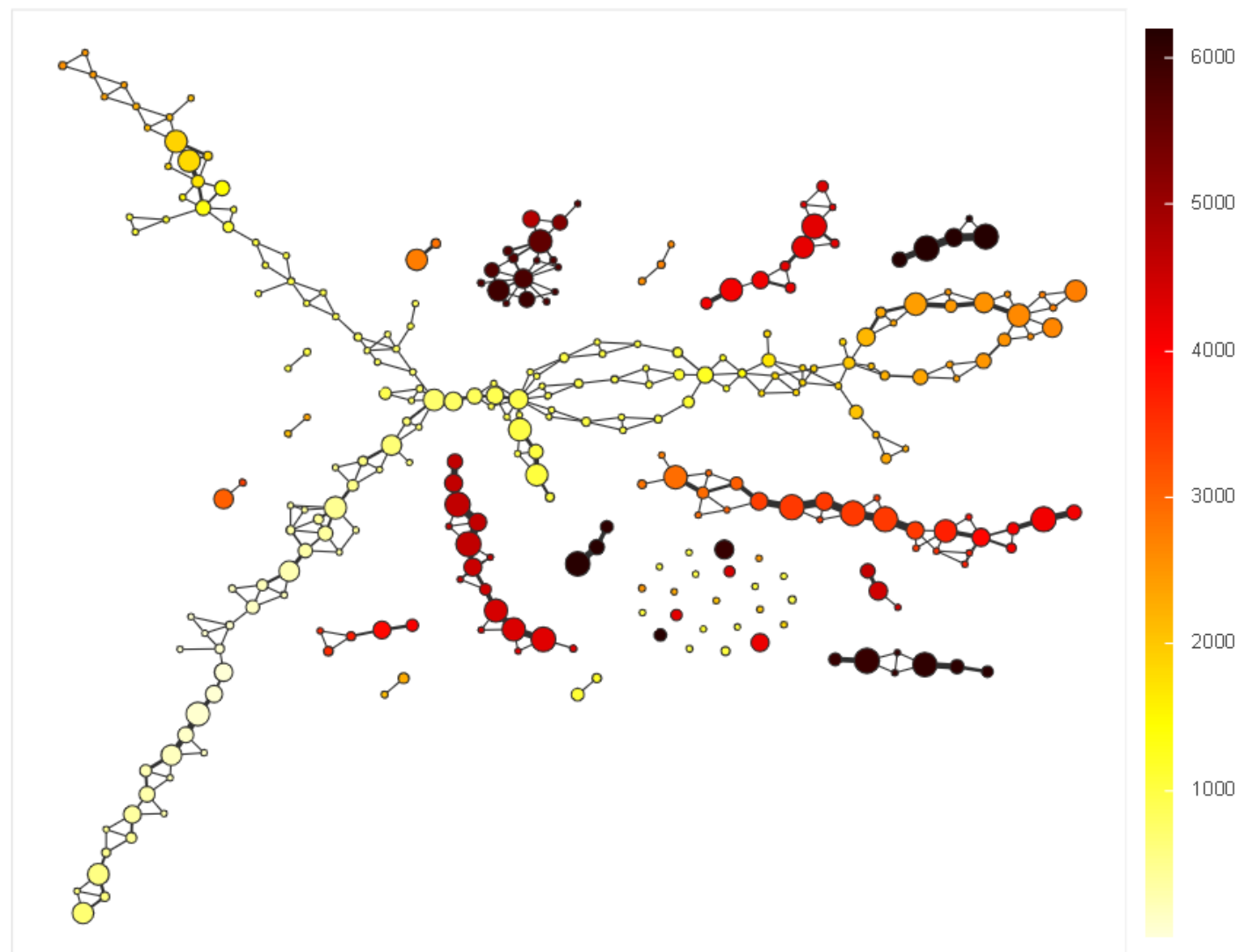
RSNHICOST_A: no coverage, coverage is not affordable
HIRECENTLY: no coverage, last had coverage within 3 years
PLN1PAY_GOV: private plan paid by Medicare, Medicaid or other government program



RSNHIWANT_A: do not need or want coverage
MILSPC2_A: military TRICARE health coverage
PRIVATE: private health insurance
MEDICARE: Medicare insurance

TDA graph (compressed)

- Nodes represent the groups of respondents with *similar insurance plans*
- The similarity of respondents is measured by *weighted Jaccard metric*
- 292 nodes** represent **26,959 respondents**
- Color of nodes reflects the *density* in the Jaccard metric of respondents within the group



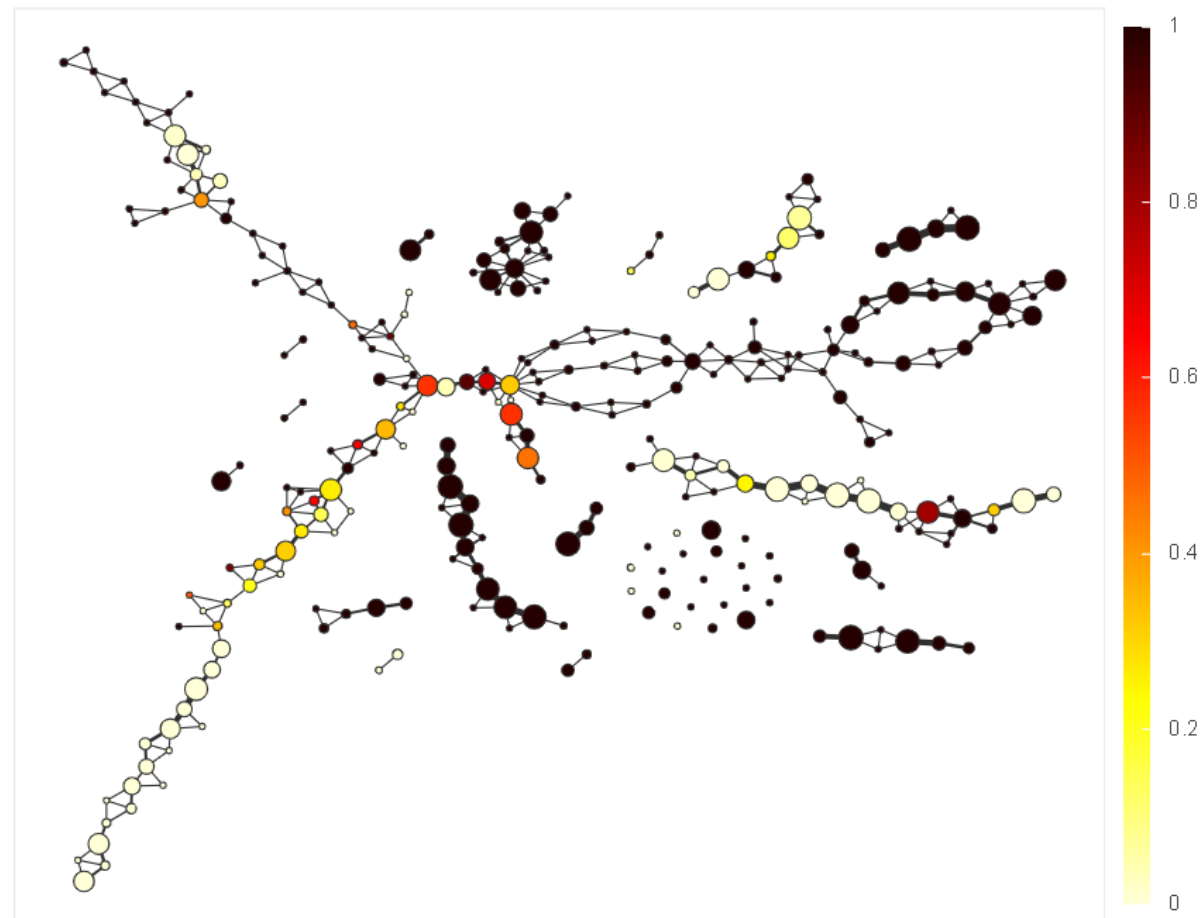
Color: density projection

Insurance coverage groups

Color: coverage status as used in Health United States



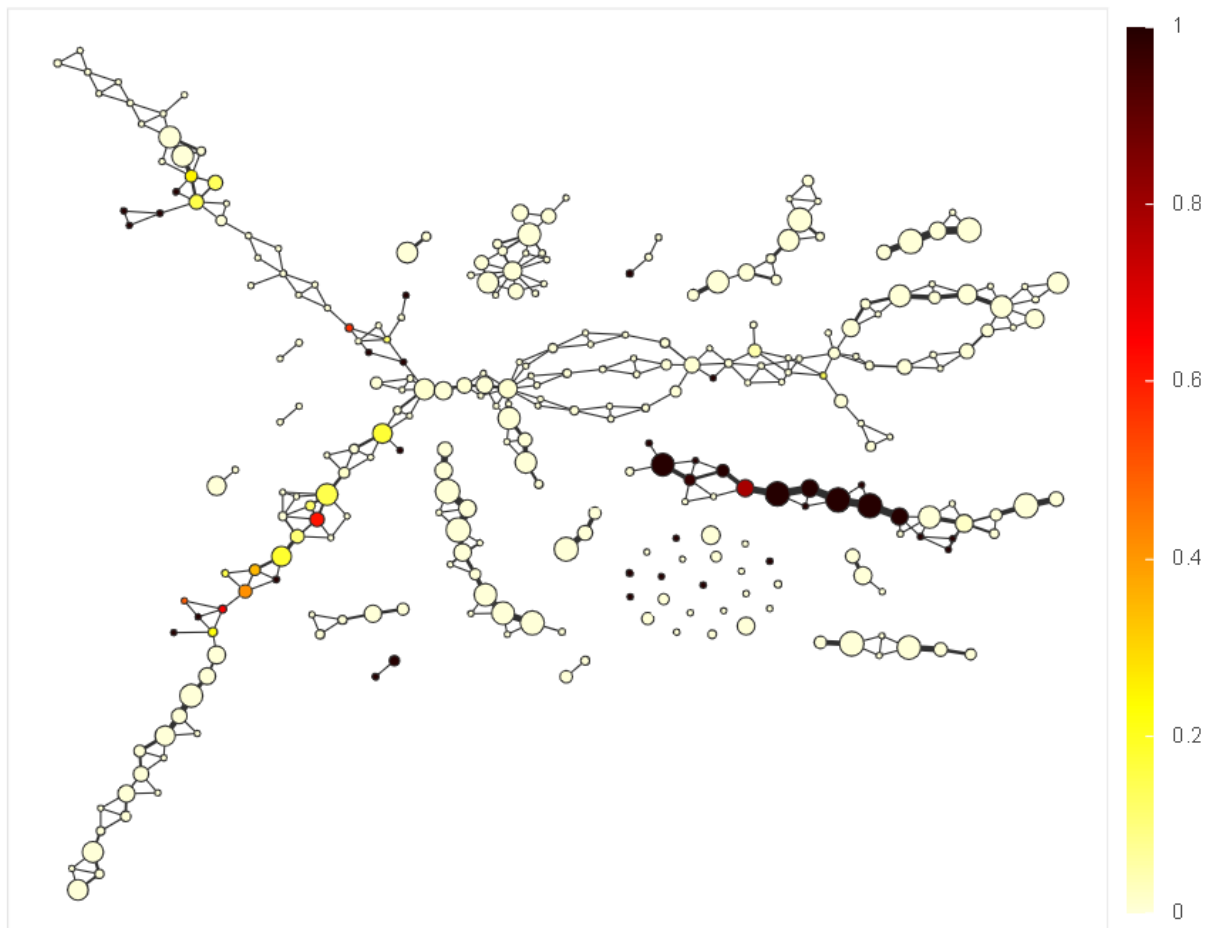
Color: private health insurance



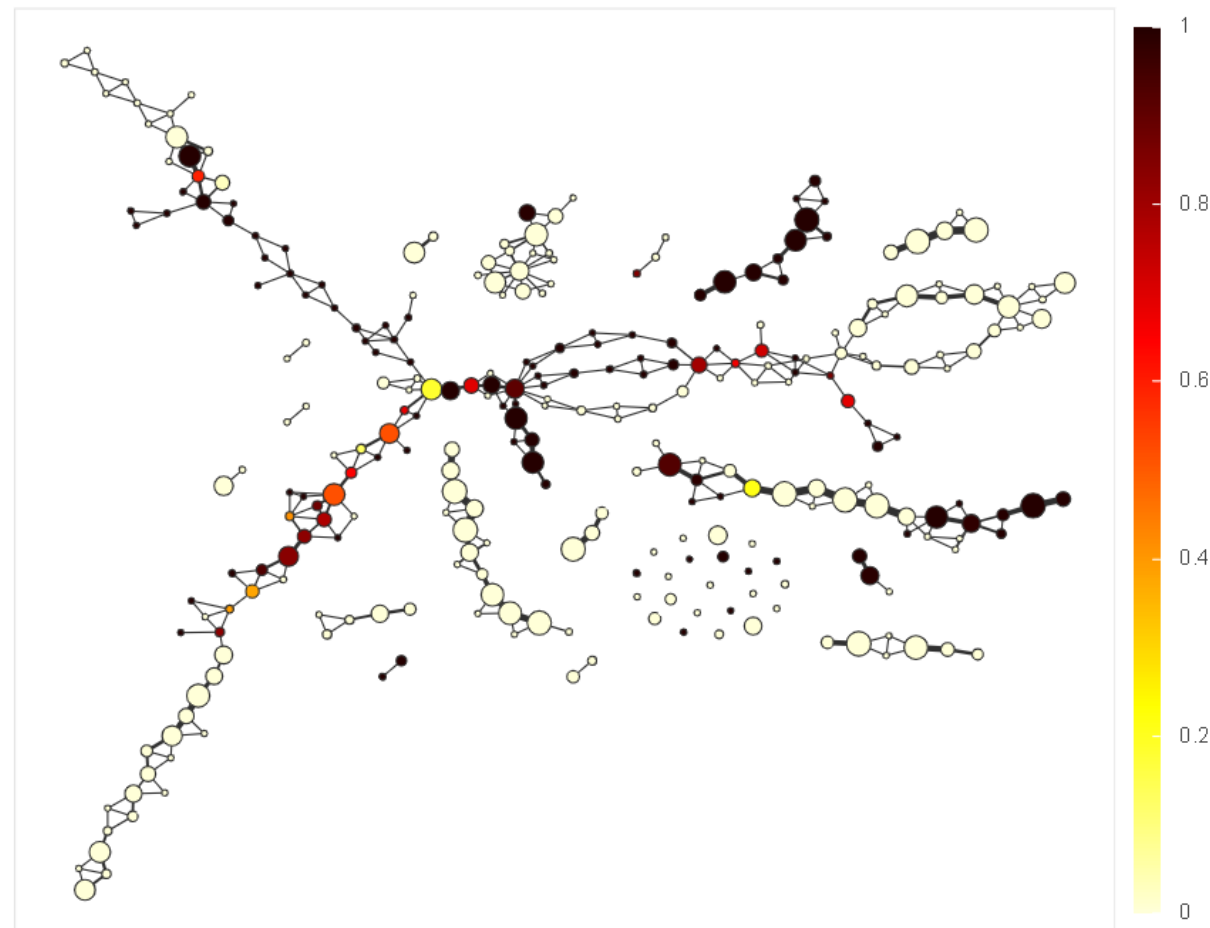


Insurance coverage groups

Color: Medicaid



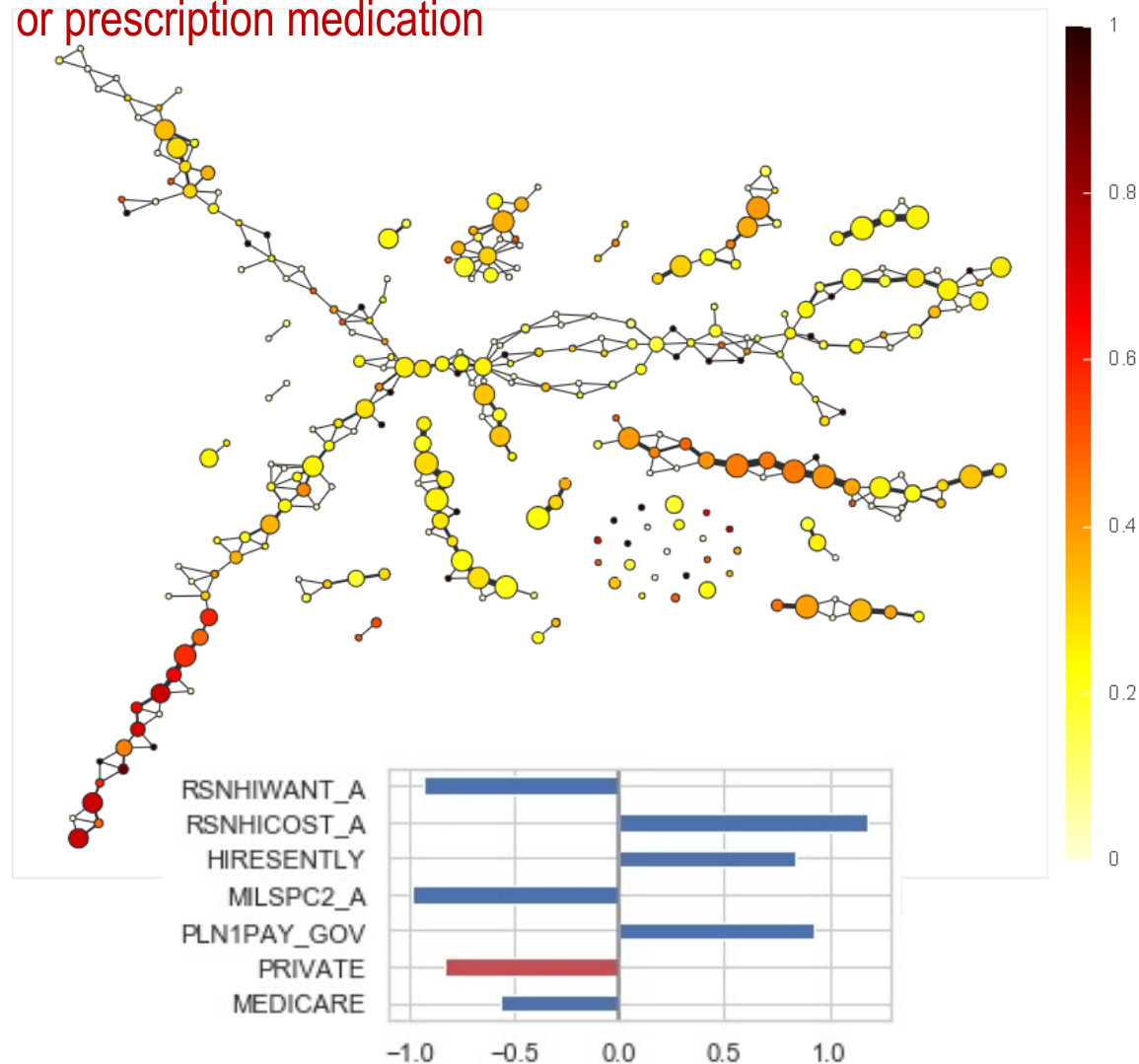
Color: Medicare



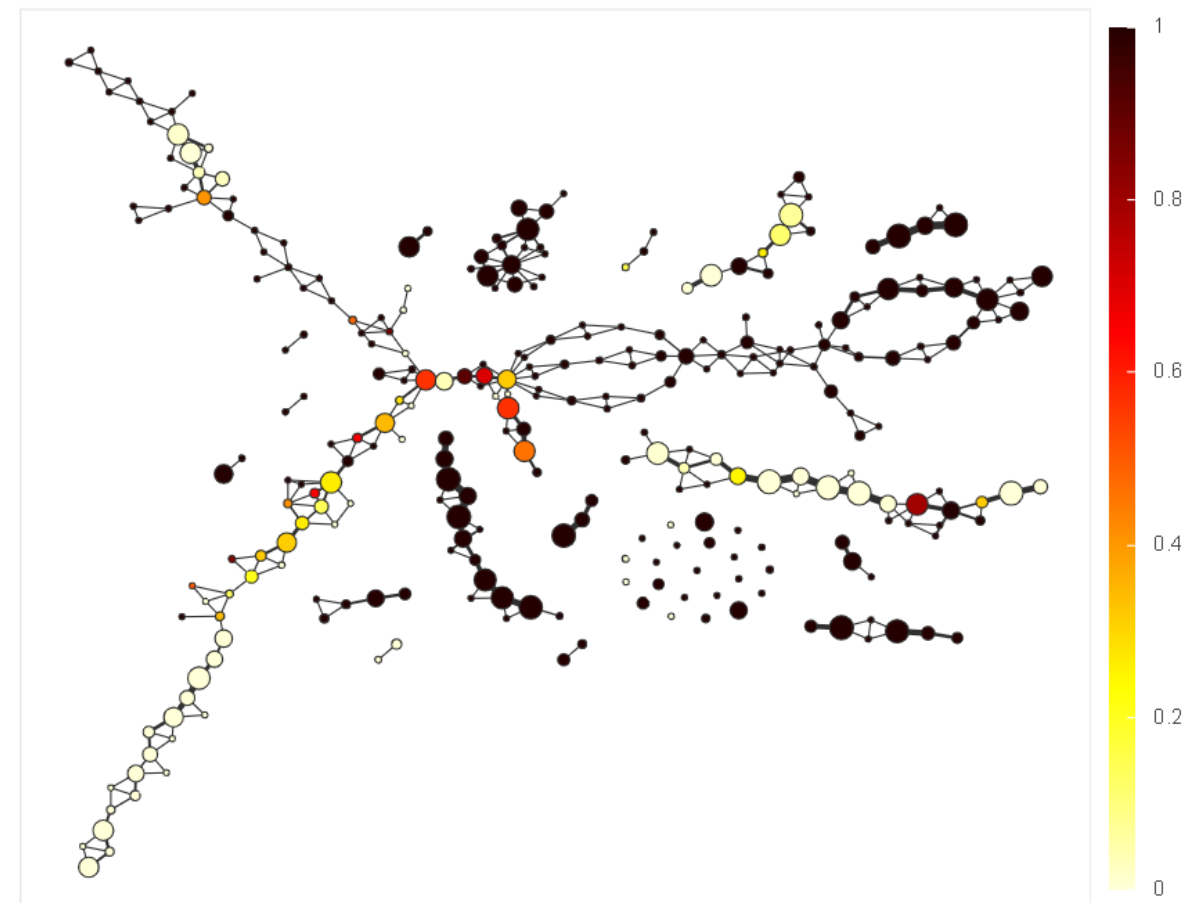


Problems with paying: logistic regression vs. TDA

Color: problems with paying for health care
or prescription medication



Color: private health insurance



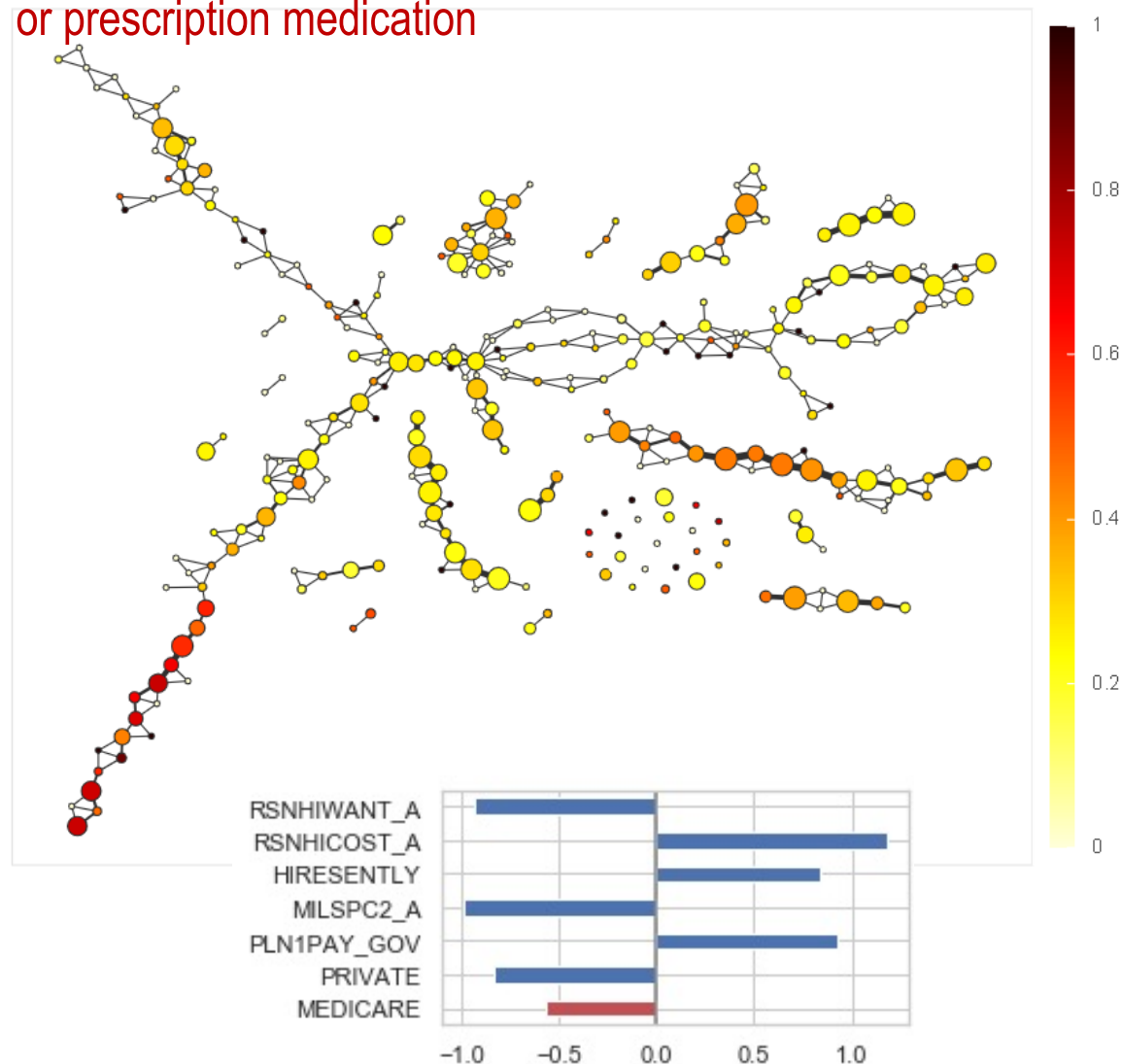
Private:
Odds ratio : 0.4360
Percent of respondents: 62.05%



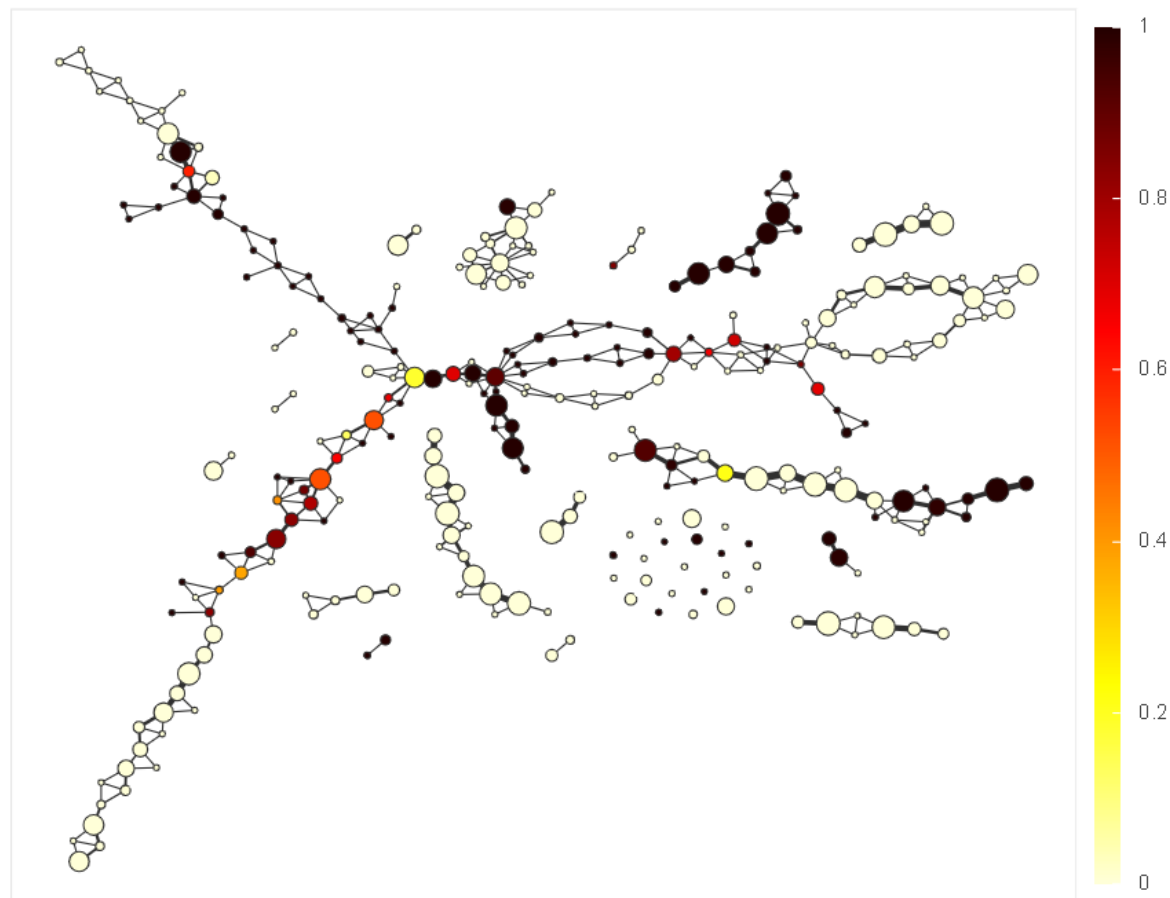


Problems with paying: logistic regression vs. TDA

Color: problems with paying for health care
or prescription medication



Color: Medicare



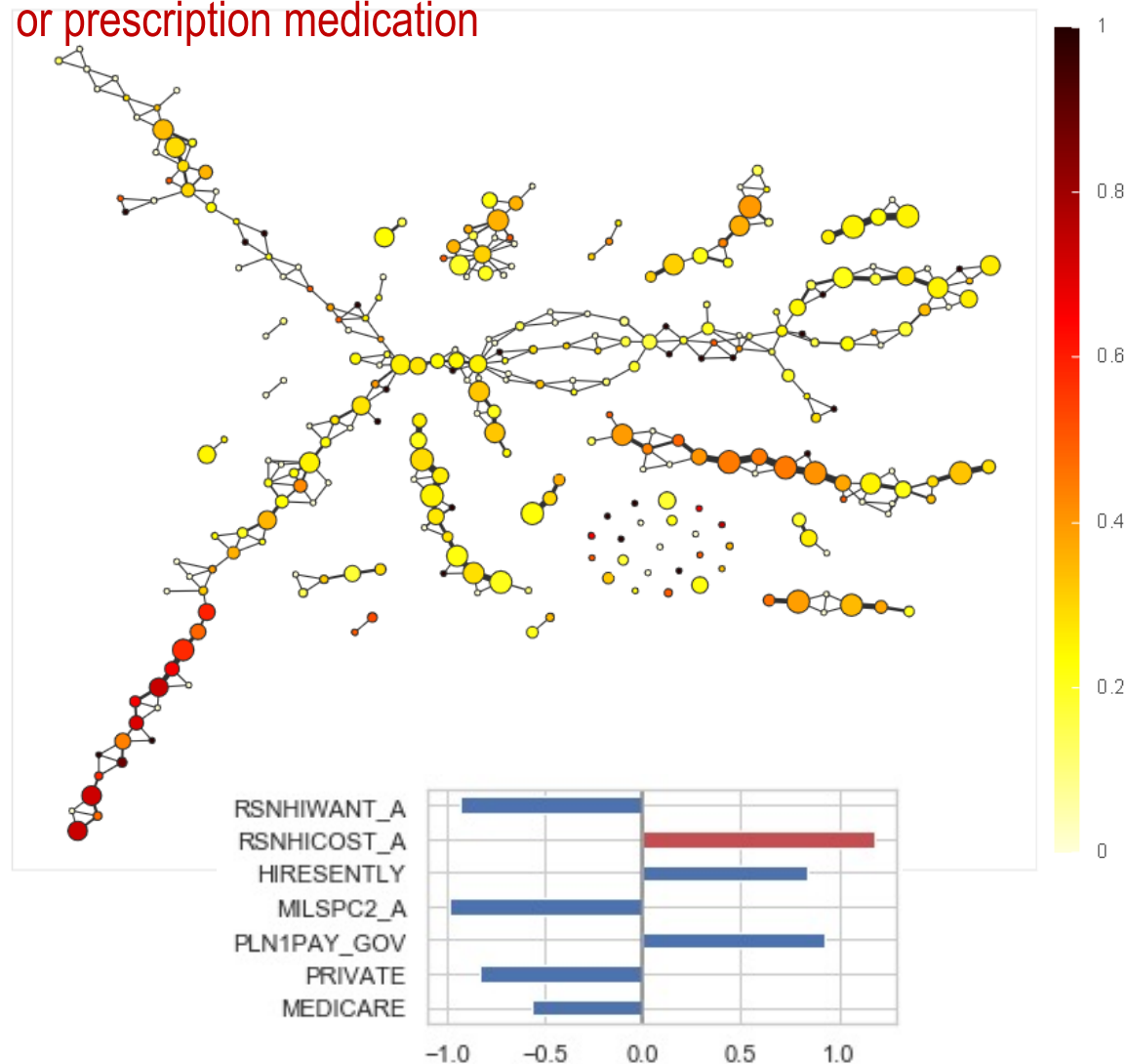
Medicare:
Odds ratio : 0.5666
Percent of respondents: 30.79%



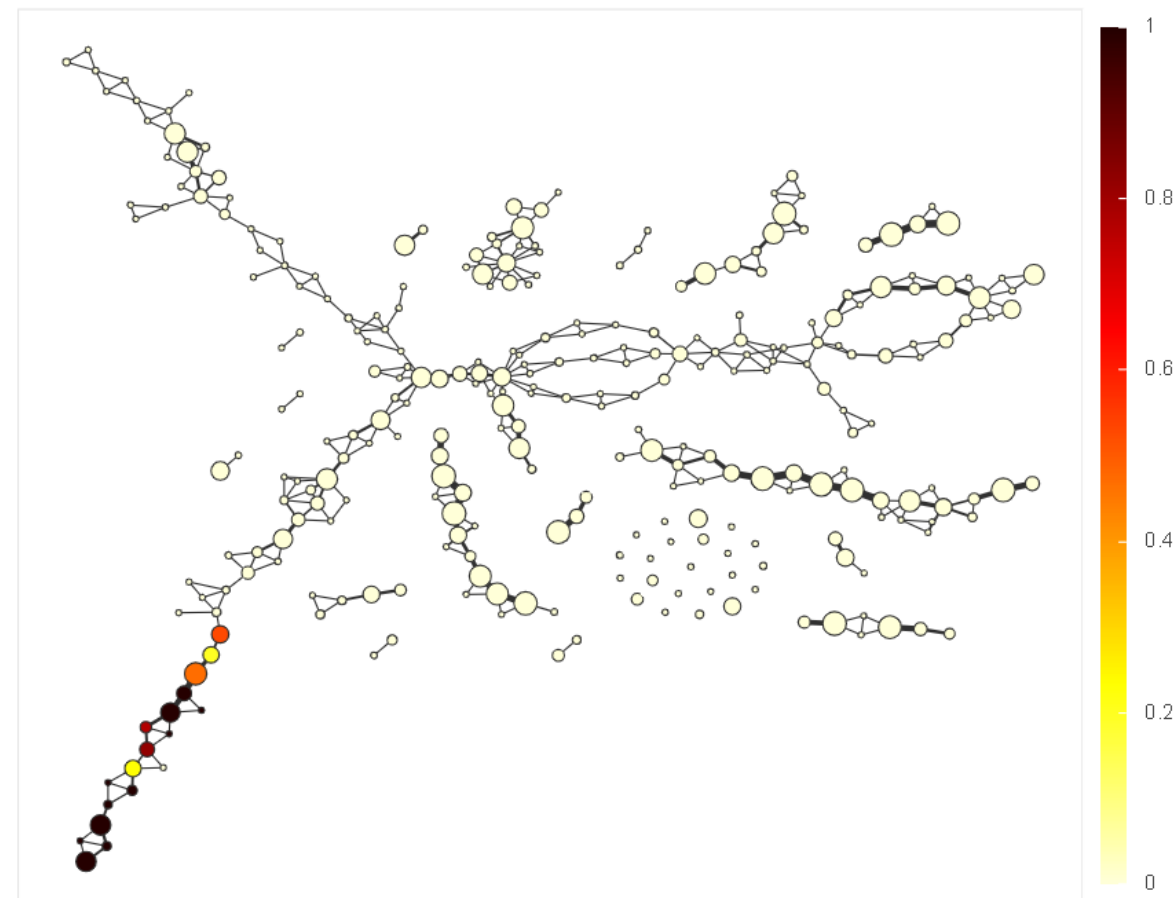


Problems with paying: logistic regression vs. TDA

Color: problems with paying for health care
or prescription medication



Color: coverage is not affordable (RSNHICOST_A)

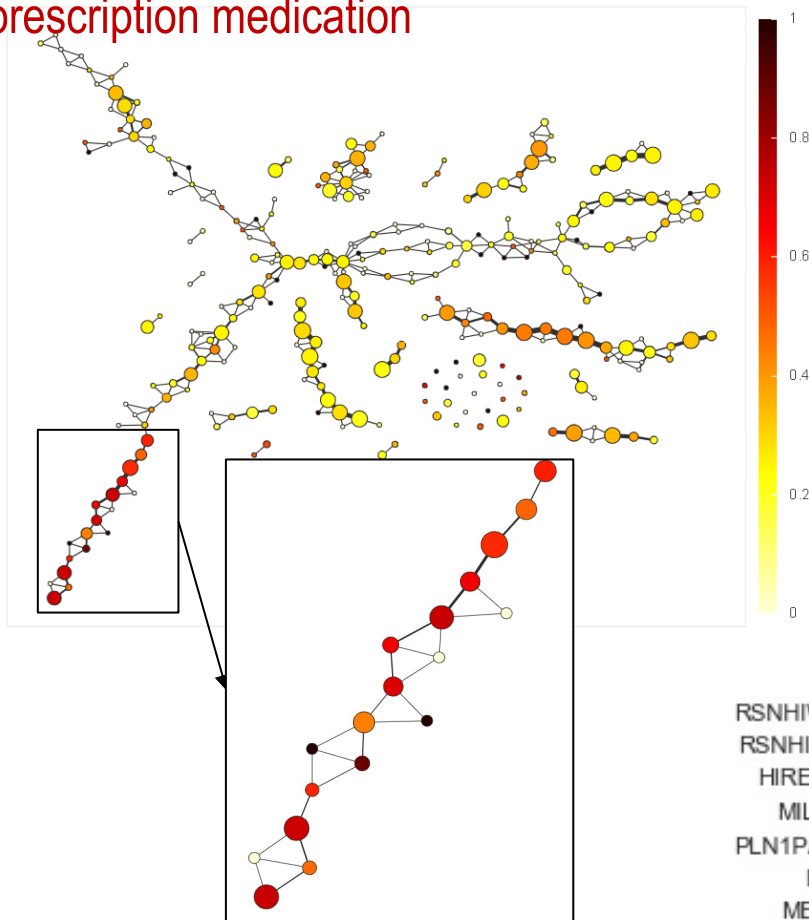


RSNHICOST_A :
Odds ratio : 3.2631
Percent of respondents: 6.80%

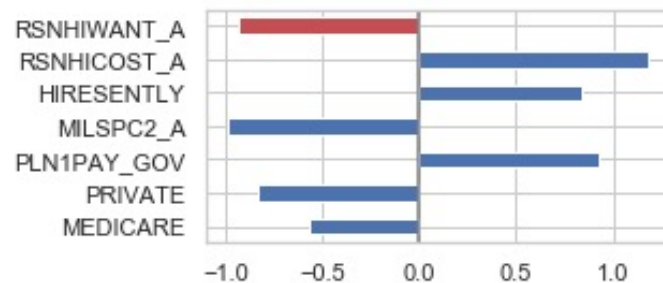
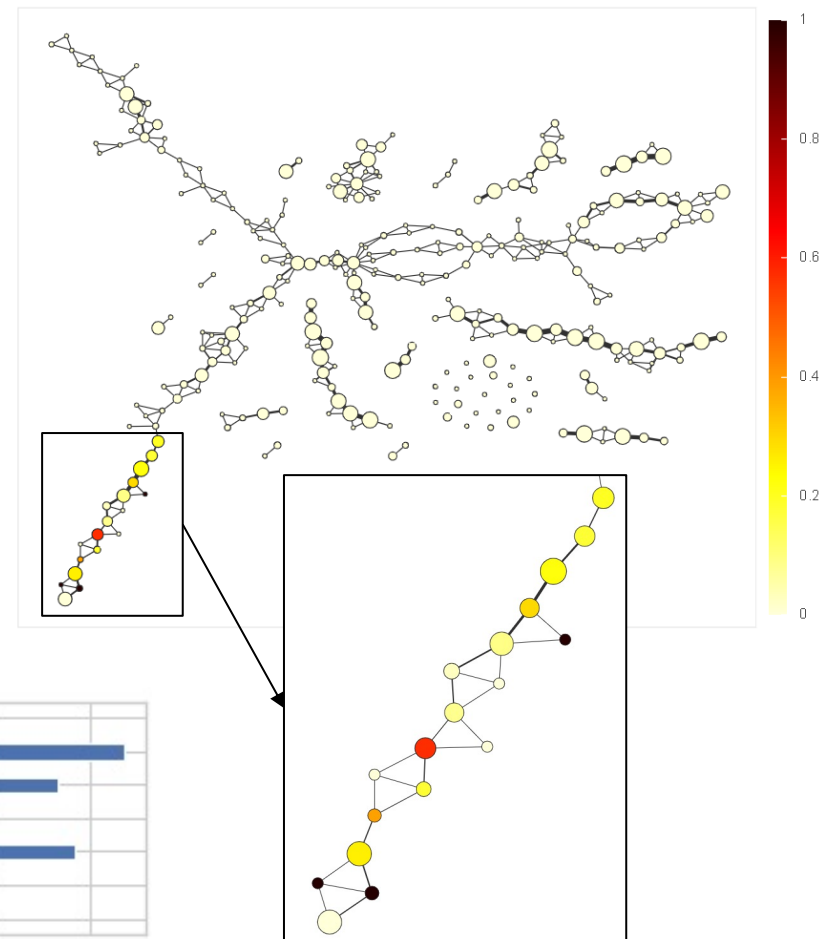


Problems with paying: logistic regression vs. TDA

Color: problems with paying for health care or prescription medication



Color: do not need or want coverage (RSNHIWANT_A)



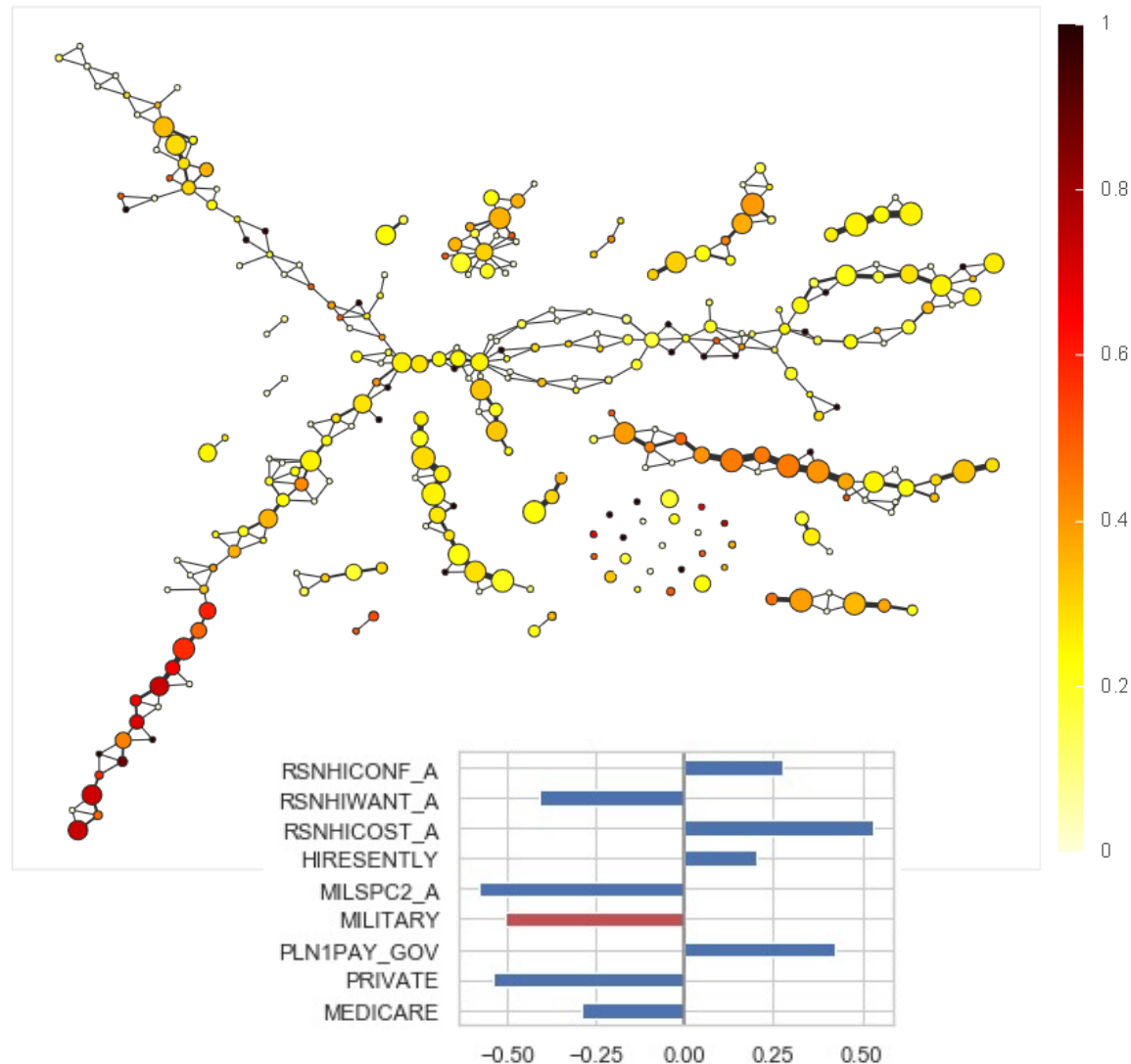
RSNHIWANT_A :
Odds ratio : 0.3916
Percent of respondents: 1.97%



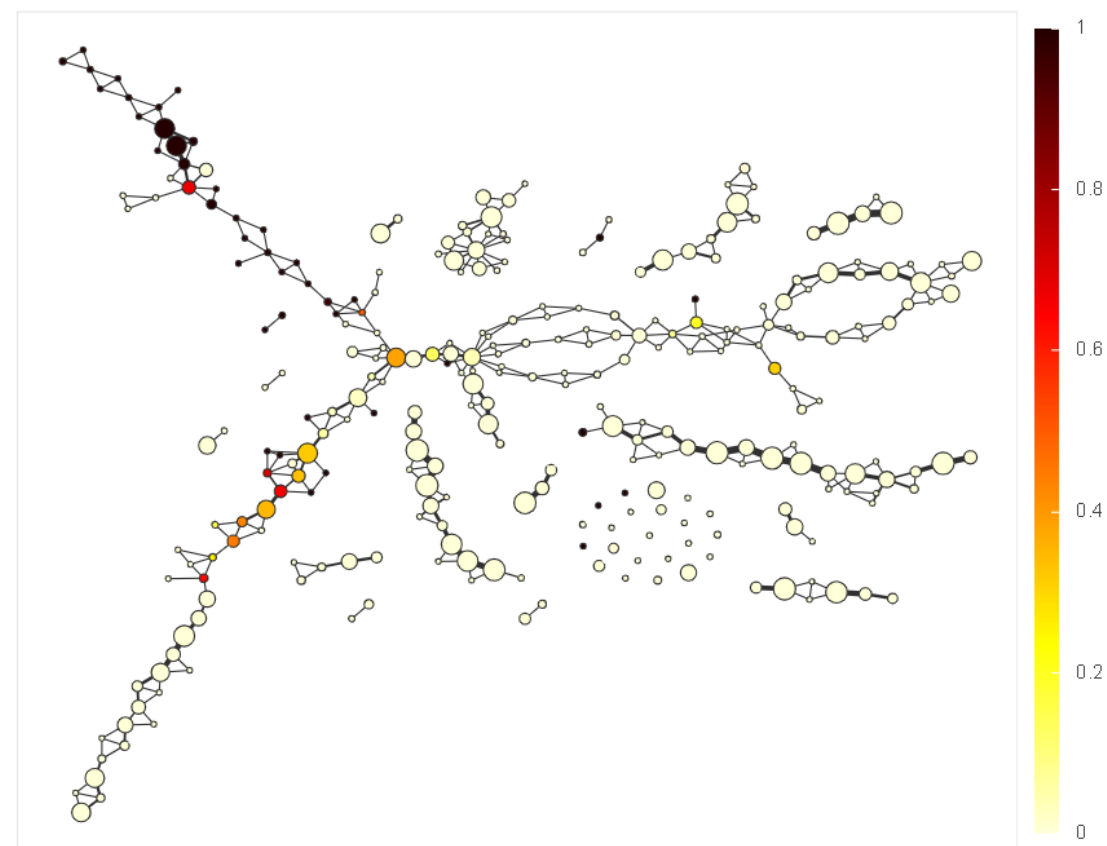


Worry about paying: logistic regression vs. TDA

Color: worry about paying medical bills



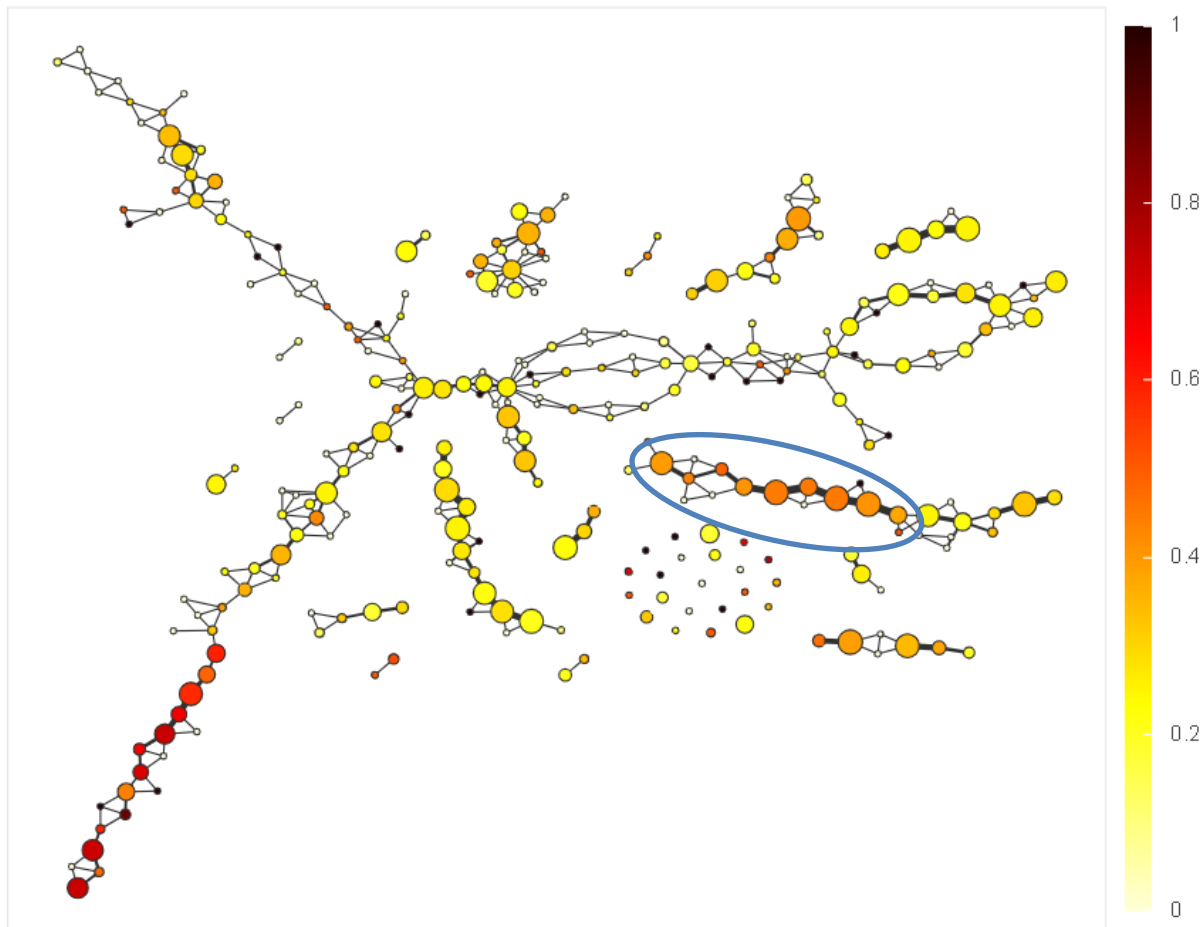
Color: military health care coverage



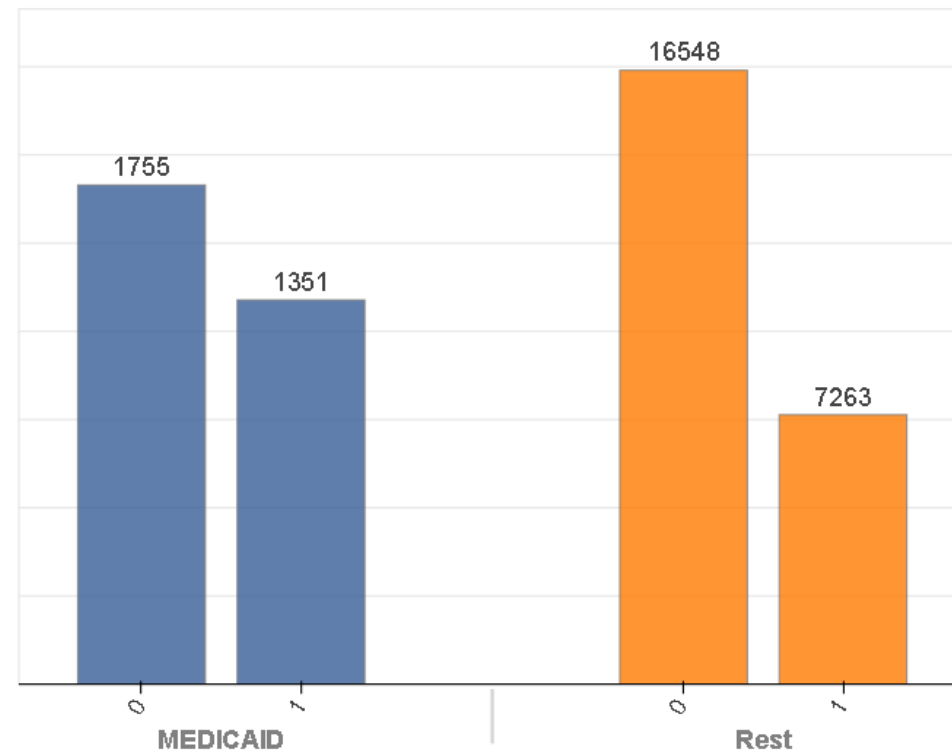
Military :
Odds ratio : 0.2581
Percent of respondents: 6.99%

TDA: additional predictors

Color: problems with paying



Fisher test: $p < 0.01$

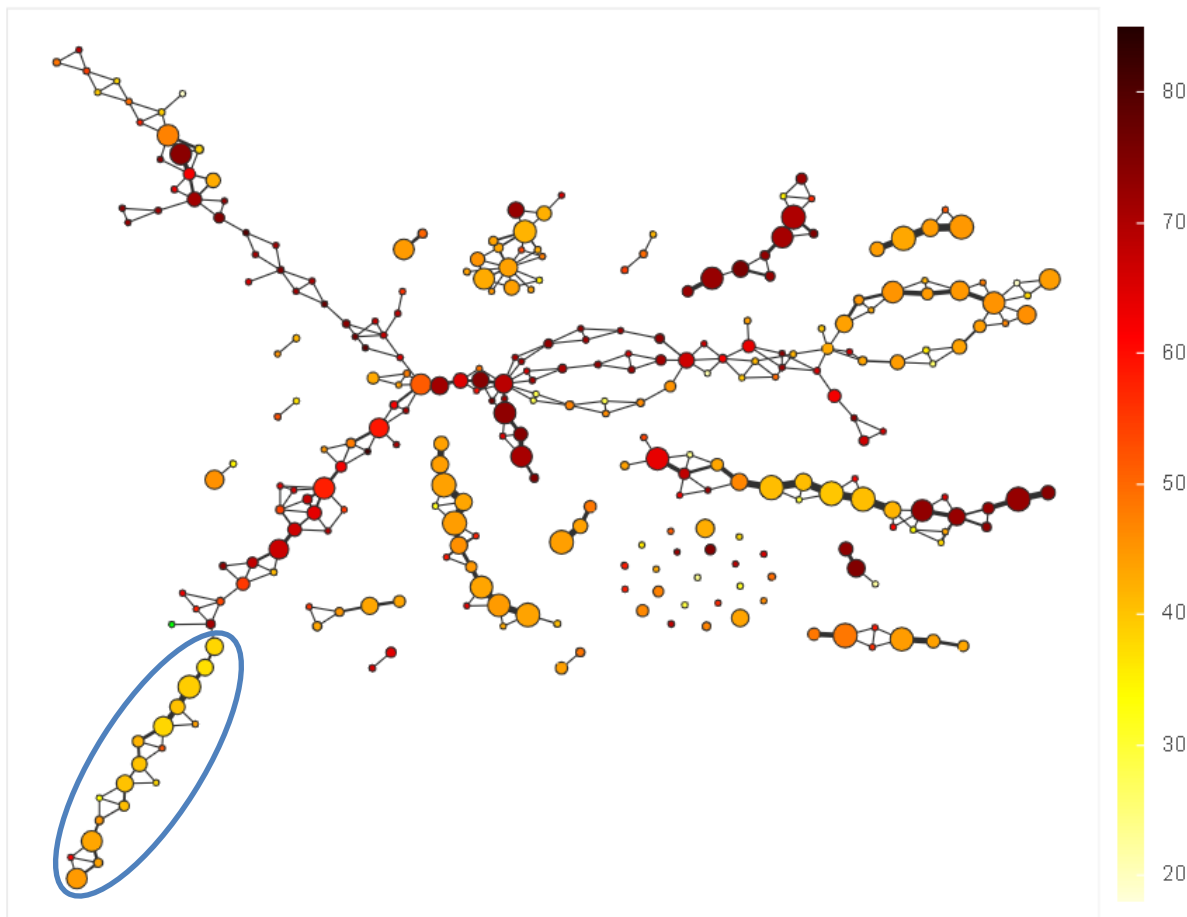


Selection: MEDICAID – 3118 respondents

Rest: 23841 respondents

TDA: insurance coverage and social predictors

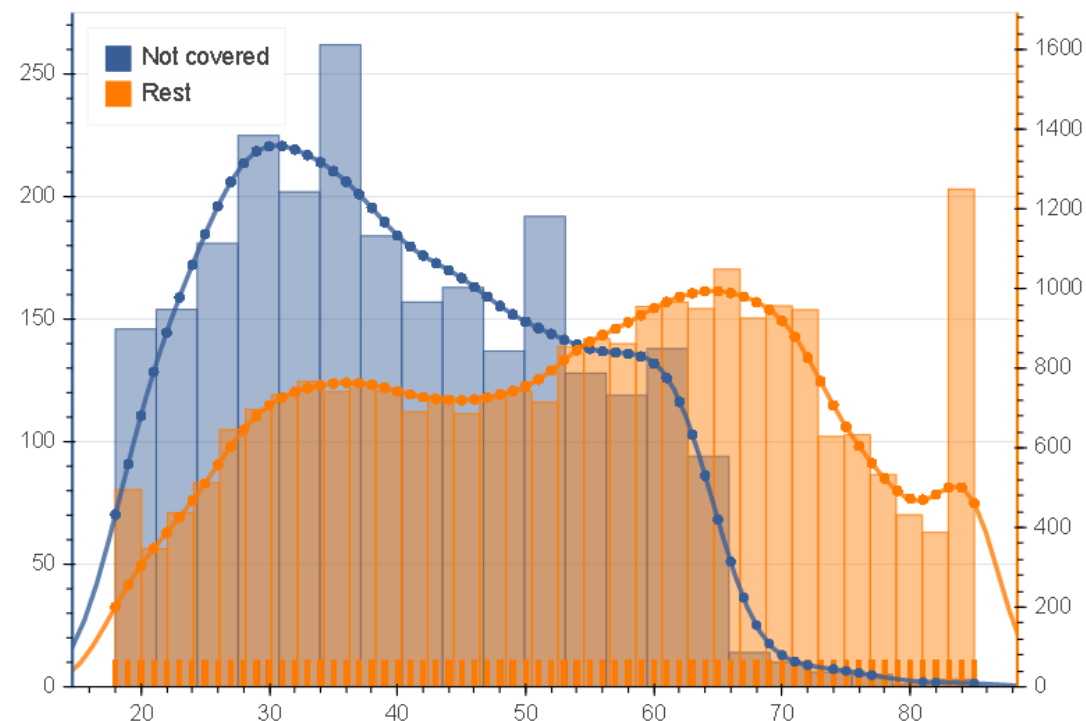
Color: age of respondents



Selection: Not covered – 2525 respondents

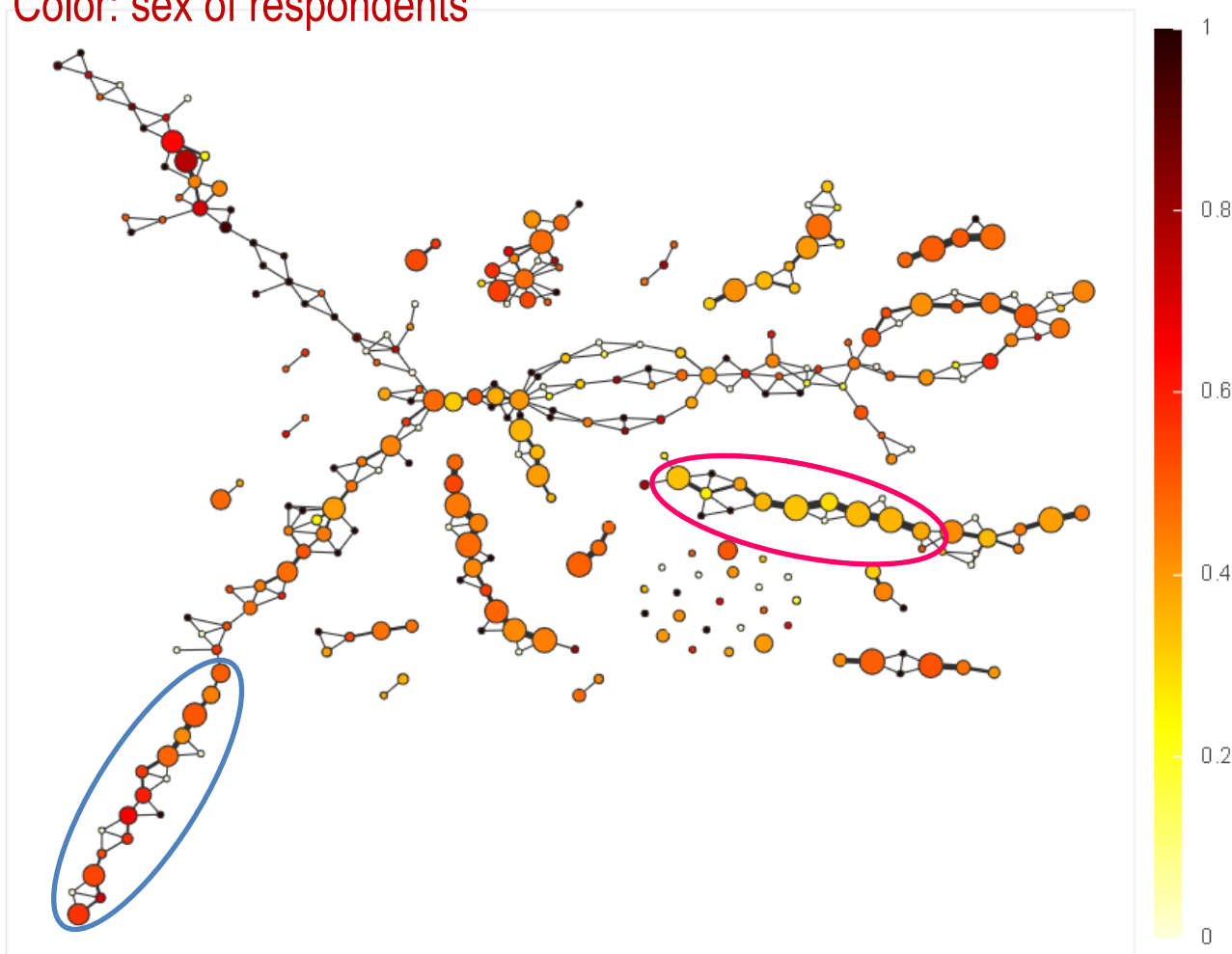
Rest: 24434 respondents

Mann-Whitney rank test: $p < 0.01$



TDA: insurance coverage and social predictors

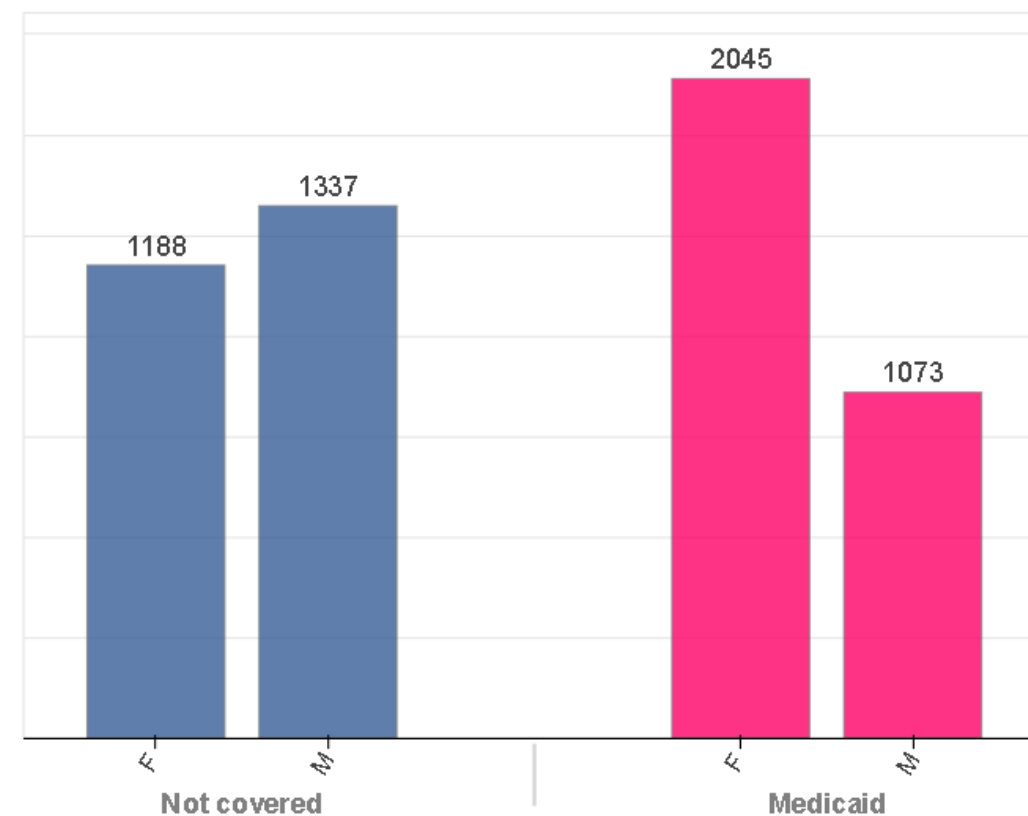
Color: sex of respondents



Selection A: Not covered – 2525 respondents

Selection B: MEDICAID – 3118 respondents

Chi-square test: $p < 0.01$



Experiment takeaways

- We built two models:
 - Supervised machine learning model using binary logistic regression
 - Unsupervised machine learning model using Topological Data Analysis (TDA)
- Results of logistic regression were confirmed by TDA
- TDA revealed additional significant predictors for selected outcomes
- Supervised and unsupervised machine learning methods can complement each other in exploring large datasets with categorical data



THANK YOU

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

