An Introduction to Synthetic Clinical Trial Data

October 4th 2019

Replica Analytics

Agenda

Time (all EST)	Person Presenting	Topic
1100 – 1105	Richard Hoptroff	Introduction
1105 – 1135	Lucy Mosquera	Data synthesis – theory and practice
1135 - 1142	Rebecca Li	Case study with the Vivli-Microsoft datathon: • An overview of the datathon objectives • How synthetic data helped with the competition
1142 - 1149	Ben Szekely	Case study with Cambridge Semantics: • An overview of the graph database and its application to clinical trial data harmonization • How synthetic data helped to expedite the technology evaluation project
1149 - 1200	Richard Hoptroff	Q&A using chat

Replica Analytics

For more information:

info@replica-analytics.com

Replica Analytics

Appendices

- Presenter biographies

Richard Hoptroff

Richard Hoptroff is a long term technology inventor, investor and entrepreneur. Awarded a Ph.D. in Physics by King's College London for his work in optical computing and artificial intelligence. In 1992, he co-founded Right Information Systems, a neural net forecasting software company that was later sold to Cognos Inc (part of IBM).

He then worked as a postdoc at the Research Laboratory for Archaeology and the History of Art at Oxford University and in 2001, created Flexipanel Ltd, a company supplying Bluetooth modules to the electronics industry.

In 2010, he founded Hoptroff London, with the aim to develop smart, hyper-accurate watch movements and create a new watch brand. In 2013 Richard Hoptroff established a new commercial category when he brought to market the first commercial atomic timepiece and an atomic wristwatch.

Richard Hoptroff then leveraged his expertise in timing technology and software to develop a hyper-accurate synchronized timestamping solution for the financial services sector, based on a unique combination of grandmaster atomic clock engineering and proprietary software.

Lucy Mosquera

Lucy Mosquera has a background in biology and mathematics, having done her studies at Queen's University in Kingston and the University of British Columbia. In the past she has provided data management support to clinical trials and observational studies at Kingston General Hospital. She also worked on clinical trial data sharing methods based on homomorphic encryption and secret sharing protocols with various companies.

At Replica Analytics, Lucy is responsible for integrating her subject area expertise in health data into innovative methods for synthetic data generation and the assessment of that data, as well as overseeing our analytics program.

Replica Analytics

Rebecca Li

Rebecca Li, PhD, is the Executive Director of Vivli and on faculty at the Center for Bioethics at the Harvard Medical School. Previous to her current role she was the Executive Director of the MRCT Center of Brigham and Women's Hospital and Harvard for over 5 years and remains a Senior Advisor at the Center. The MRCT Center is a neutral convening organization that works to define actionable policy solutions for the clinical trial enterprise. She has over 20 years of experience spanning the entire drug development process with experience in Biotech, Pharma and CRO environments. She completed a Fellowship in 2013 in the Division of Medical Ethics at Harvard Medical School. Dr. Li also served as the VP of Clinical Research at the New England Research Institutes for 6 years. She was also previously employed at Wyeth Research as the Associate Director in Translational Clinical Research. She earned her PhD in Chemical and Biomolecular Engineering from Johns Hopkins University.

Replica Analytics

Ben Szekely

Ben is Senior Vice President, Head of Field Operations at Cambridge Semantics Inc. As a Founding Engineer of Cambridge Semantics, Ben has impacted all sides of the business from developing the core of the Anzo Platform to leading early business development and customer engagements.

Ben currently leads a rapidly growing team of software engineers and data gurus to identify and deliver high value Anzo Smart Data solutions to Cambridge Semantics' customers and partners across Pharma, Financial Services and Government.

Before joining the founding team at Cambridge Semantics, Ben worked as an Advisory Software Engineer at IBM with CSI founder and CTO Sean Martin on early research projects in Semantic Technology.

He has BA in Math and Computer Science from Cornell University and an SM in Computer Science from Harvard University.

Replica Analytics

Introduction to Synthetic Data

Lucy Mosquera

October 4, 2019

Replica Analytics

What is Synthetic Data ?

- Data that is generated from real data, but is not real data. It has the same statistical properties as real data.
- Models (statistical machine learning and deep learning) are built from the real data and sample from these models to create synthetic data.
- Because it is not real data, it will not have the same privacy risks as real data. We can explicitly test that assumption.

Replica Analytics

Deep Fakes

Replica Analytics

Types of Synthetic Data

- Data synthesis can be performed on different types of data:
  - Structured data
  - Images
  - Video
  - Audio
  - Text
- Our focus is on structured data

Replica Analytics

Basic Example - Method

This is a simple example to illustrate how synthesis is performed – to make the process concrete. The method used in the example is not something to apply in practice.

- State-wide hospital discharge dataset
- Three variables of interest to illustrate the concepts:
  - Age at time of visit (in years)
  - Days since last visit
  - Length of stay
- Removed all births from the dataset (n=189,047 discharges)

Replica Analytics

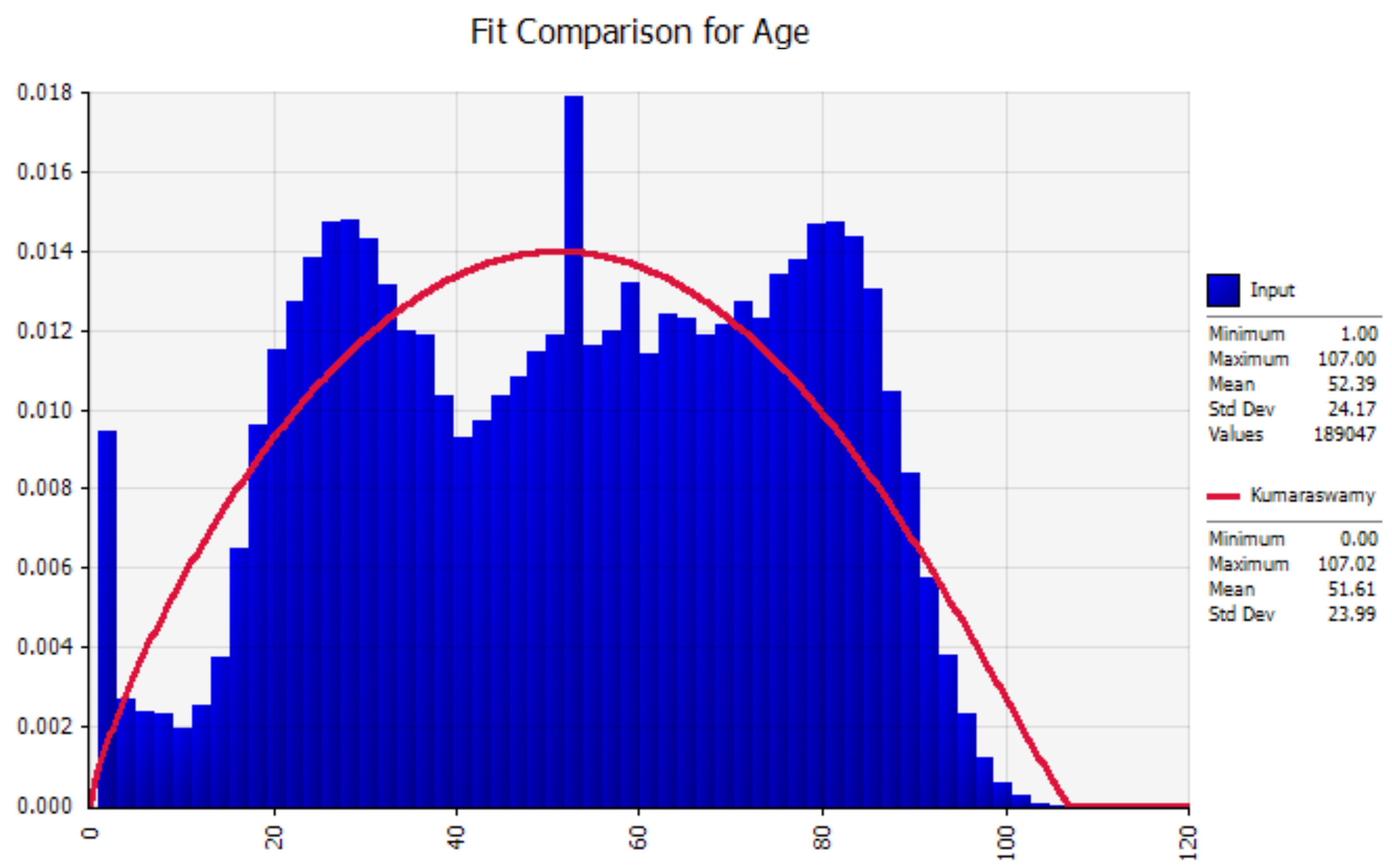
Basic Example - Method

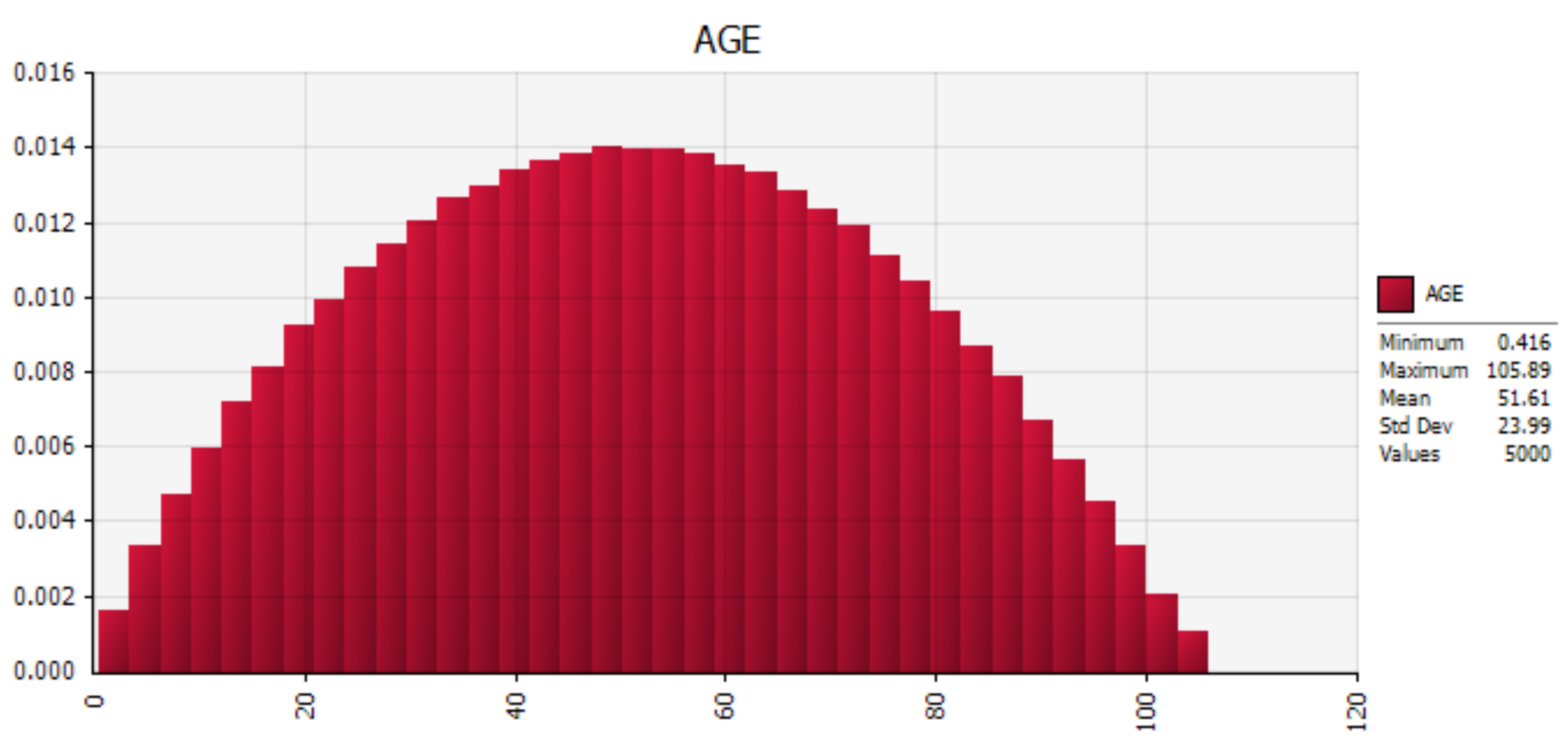
- Synthetic data is generated by:
  - Sampling from the fitted distributions
  - Inducing the non-parametric correlations among the values during sampling
  - We use automated distribution fitting using the AIC criterion to determine the best fit for each variable
  - We compute the empirical non-parametric correlation

The result is a synthetic dataset that has the same distributions as the original data and has the same bivariate correlations as the original data

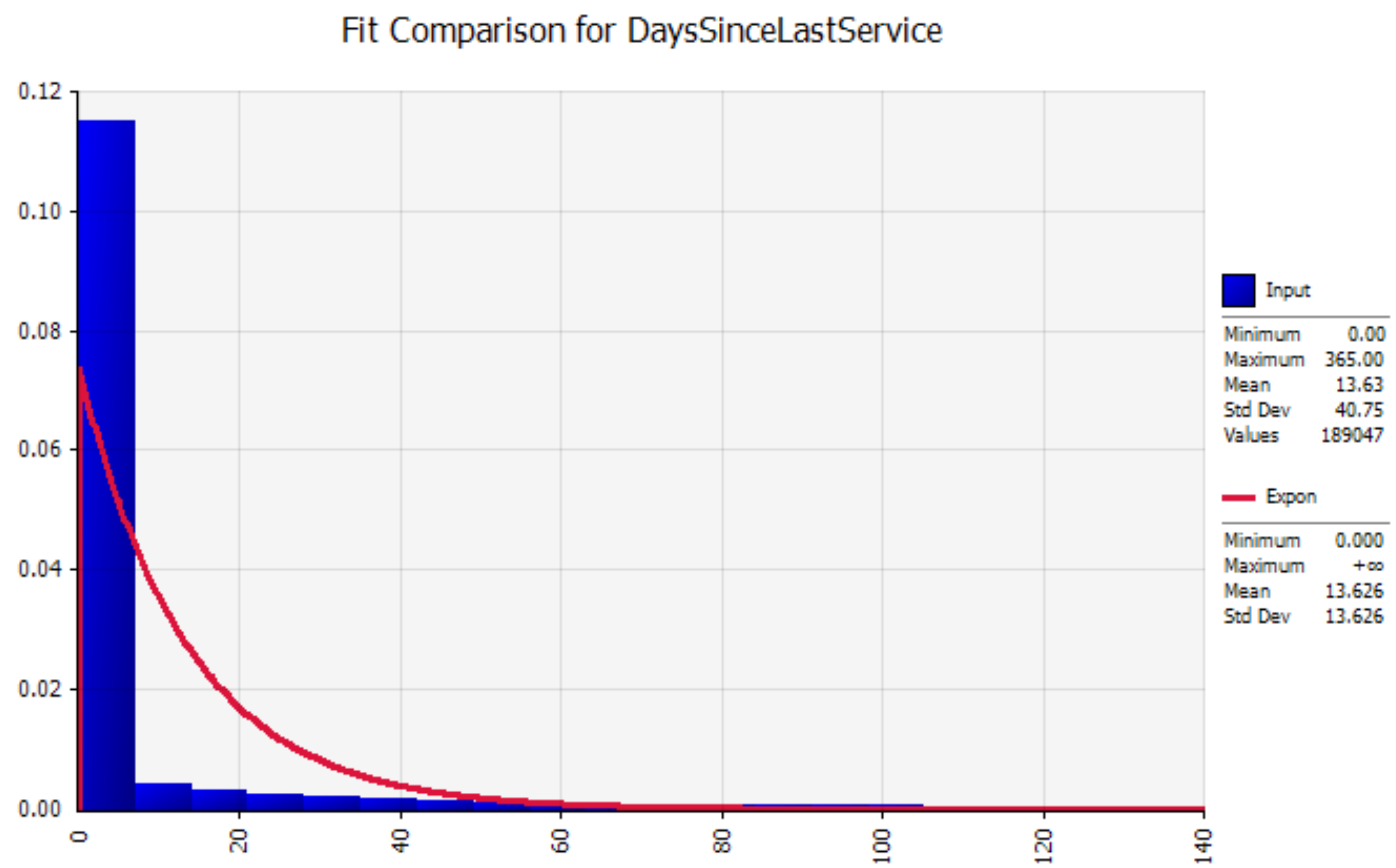
Replica Analytics

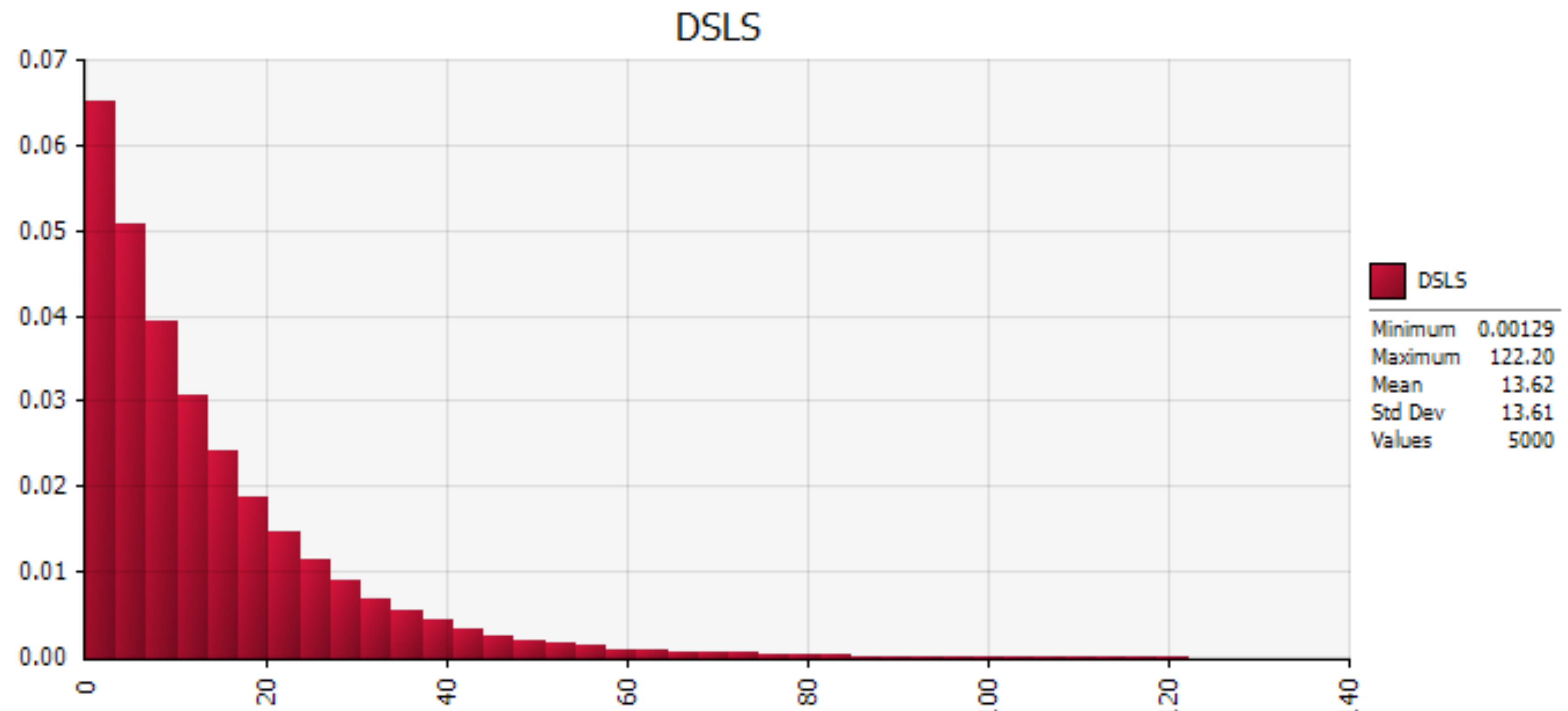
Marginal Distributions





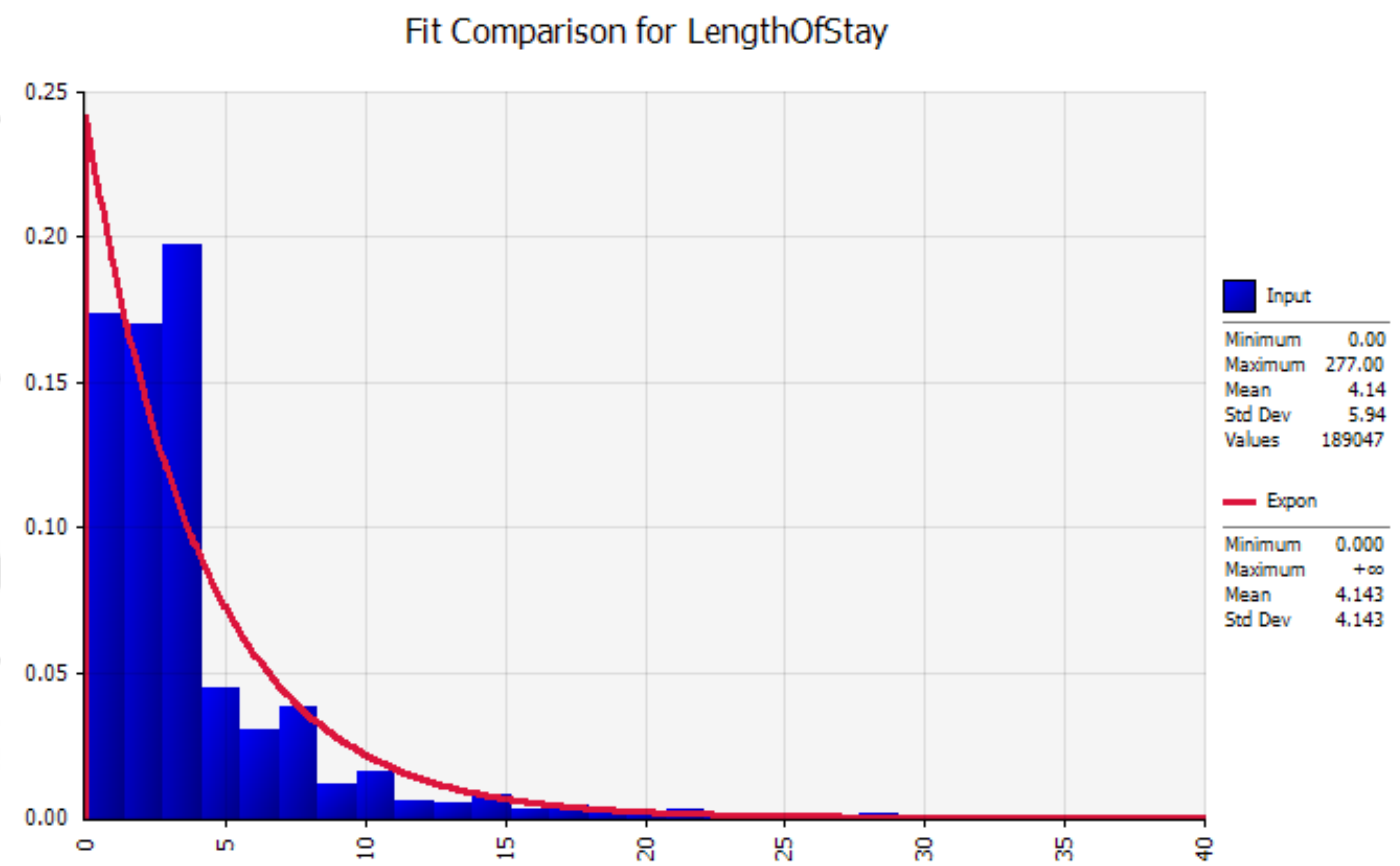
Marginal Distributions

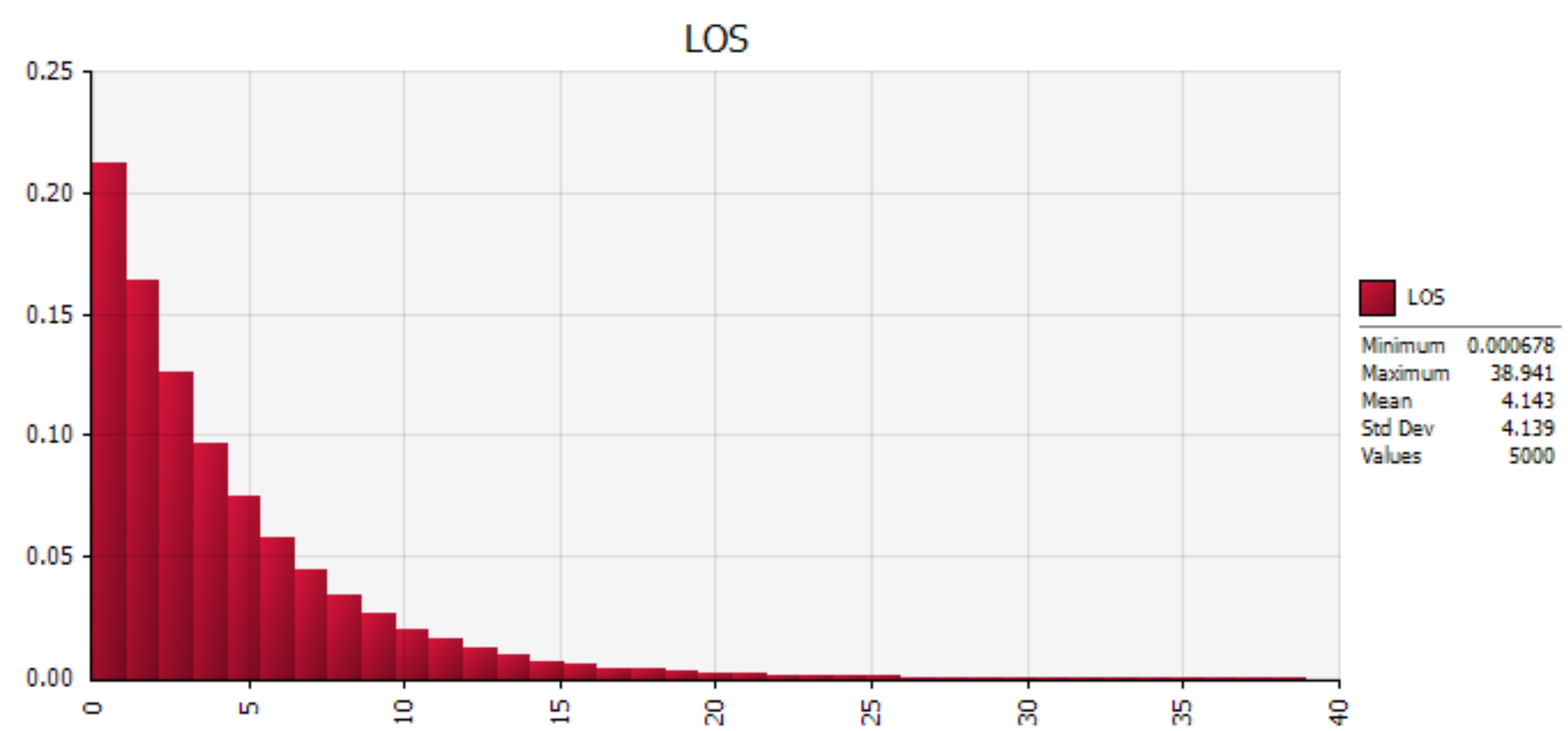




Replica Analytics

Marginal Distributions



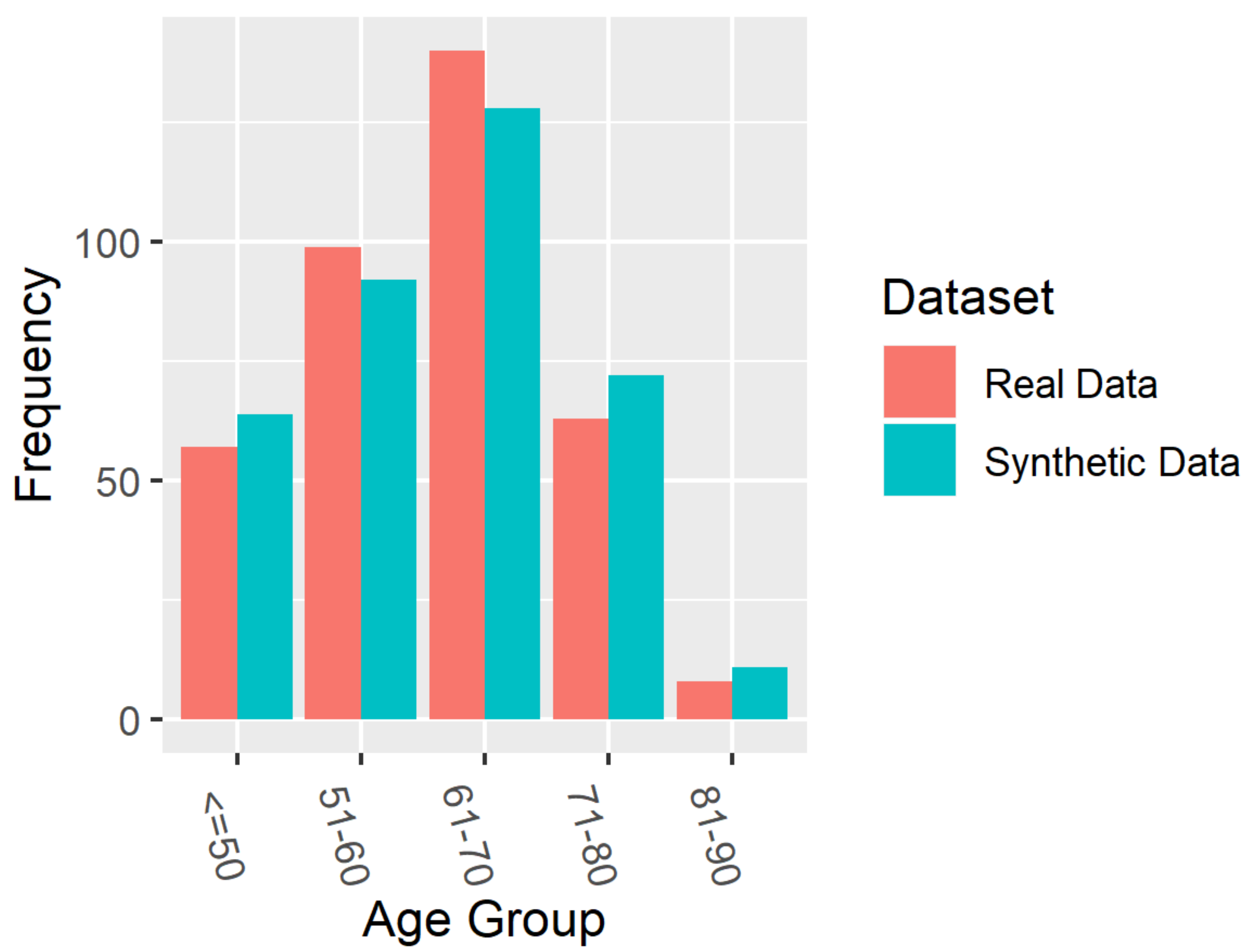


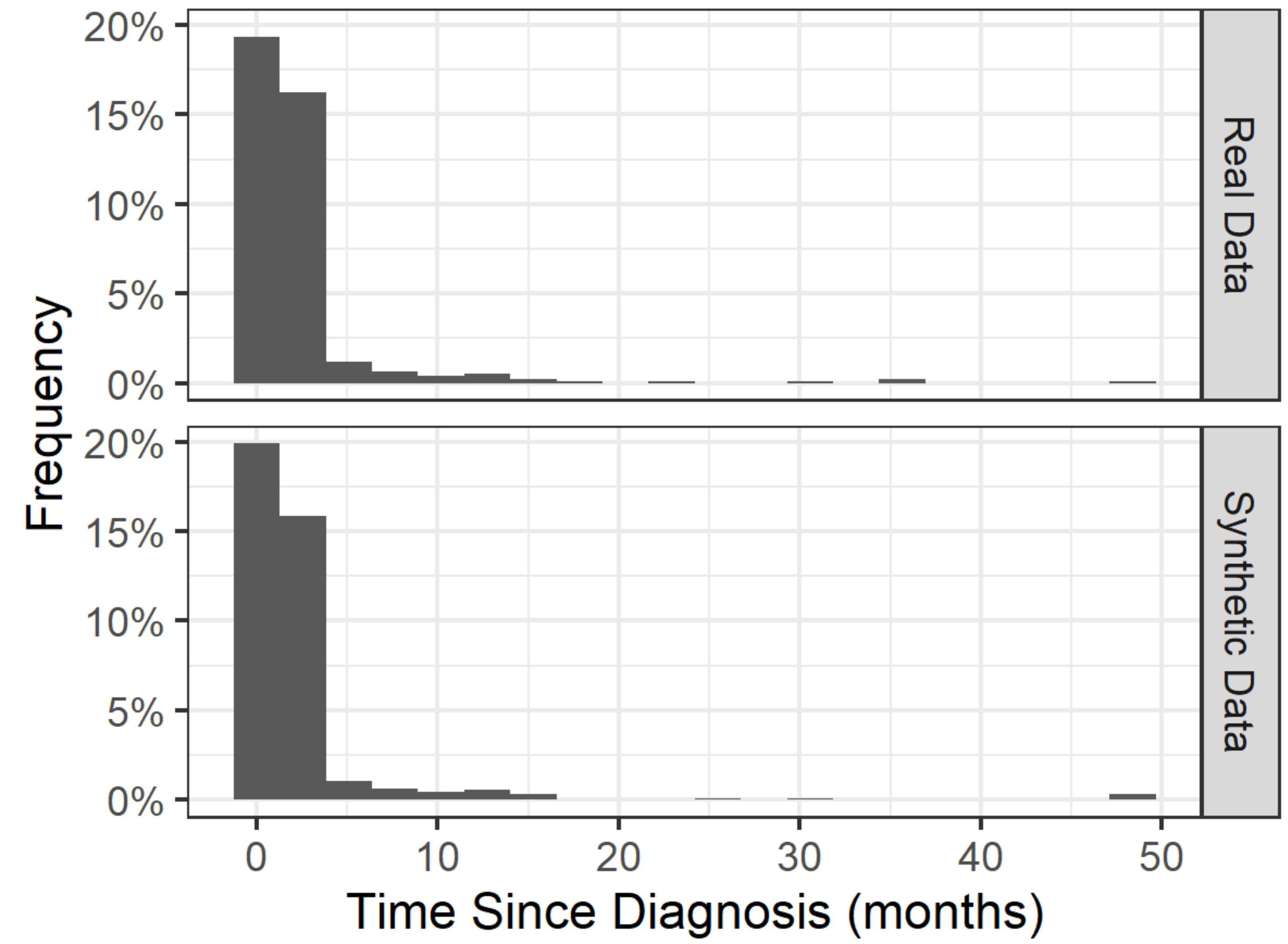
Replica Analytics

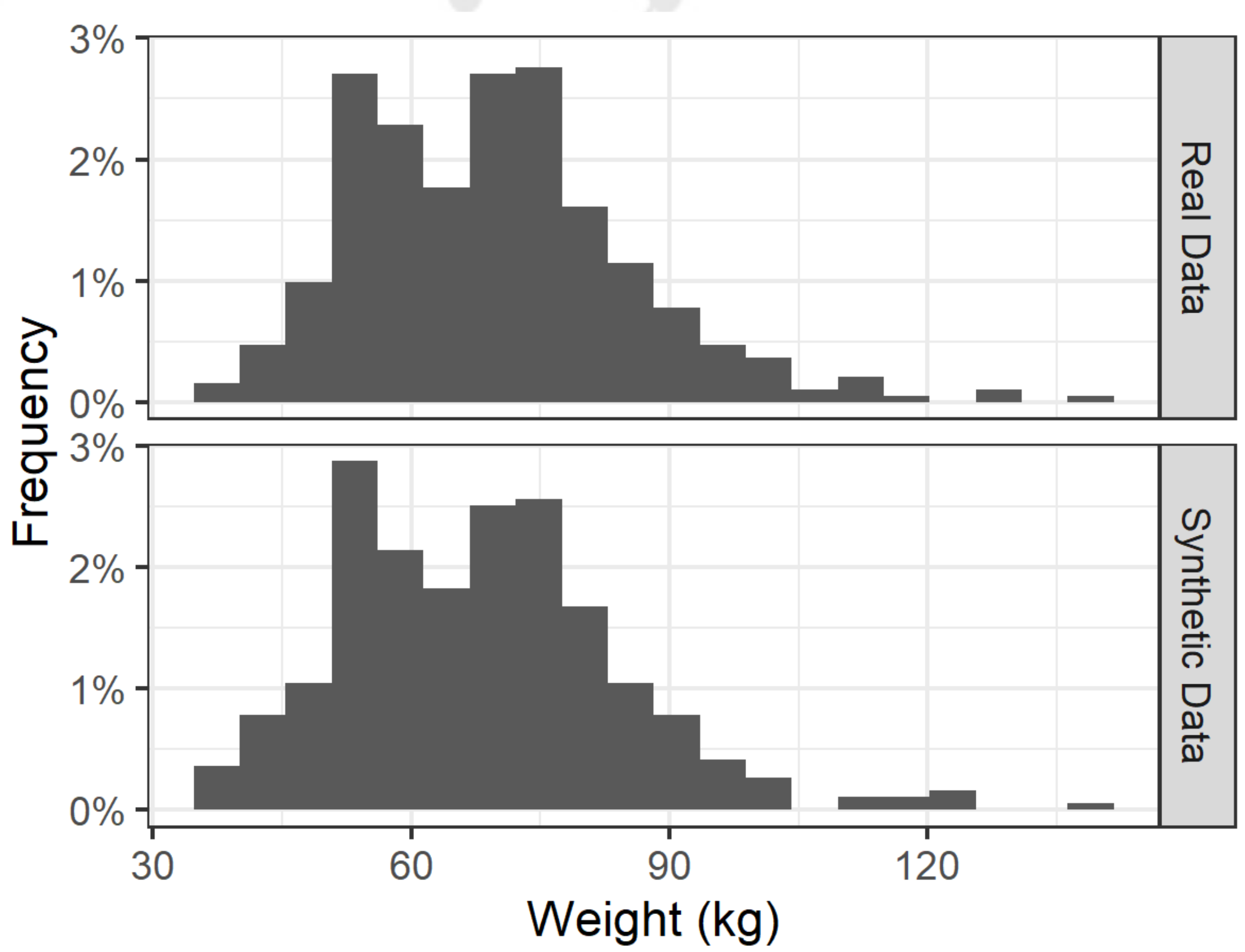
Correlations

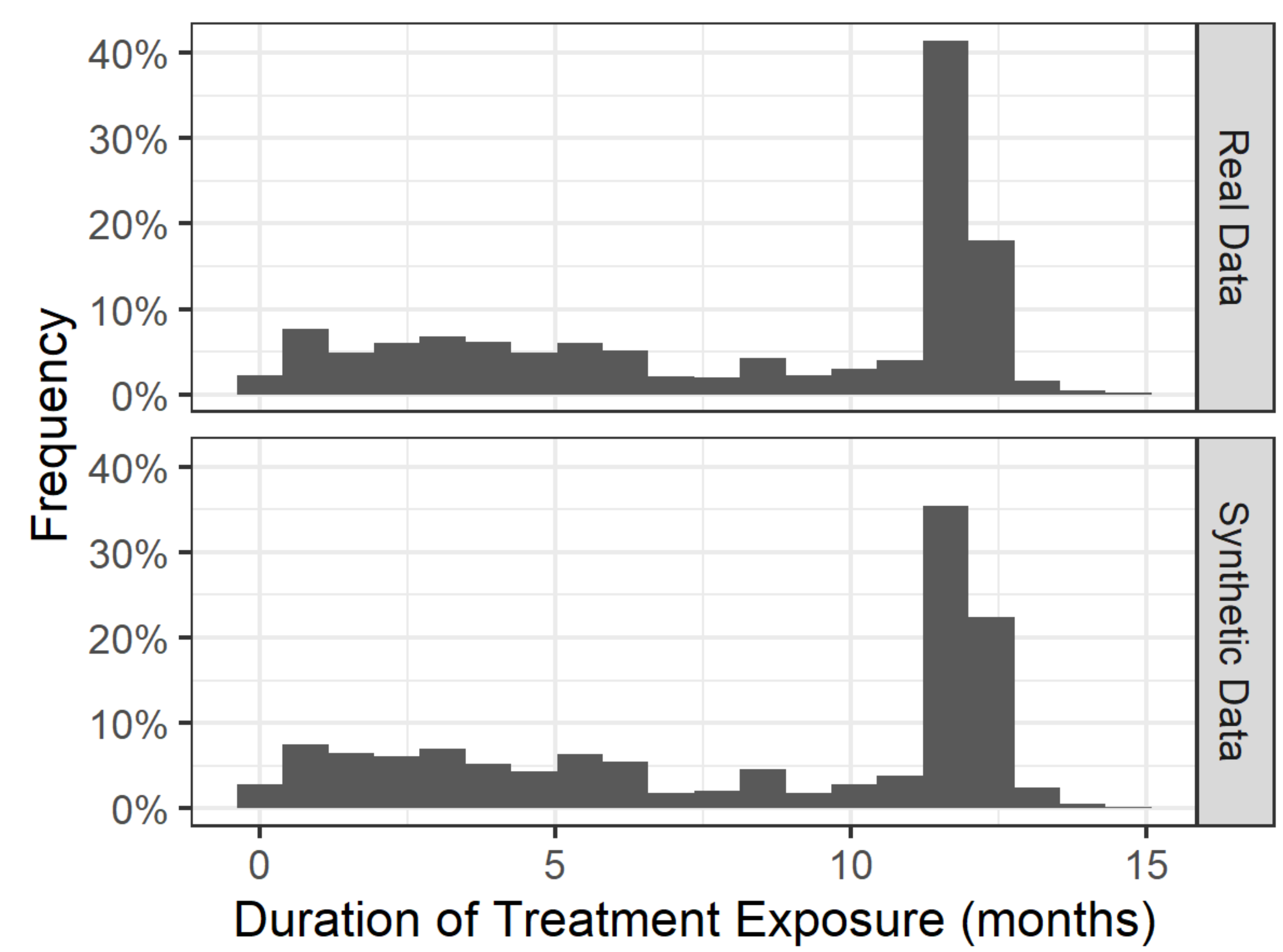
	AGE	DSLS	LOS
AGE	1	0.1617 (0.165)	0.1968 (0.223)
DSLS	0.1617 (0.165)	1	0.1424 (0.168)
LOS	0.1968 (0.223)	0.1424 (0.168)	1

Similarity to Real CT Data









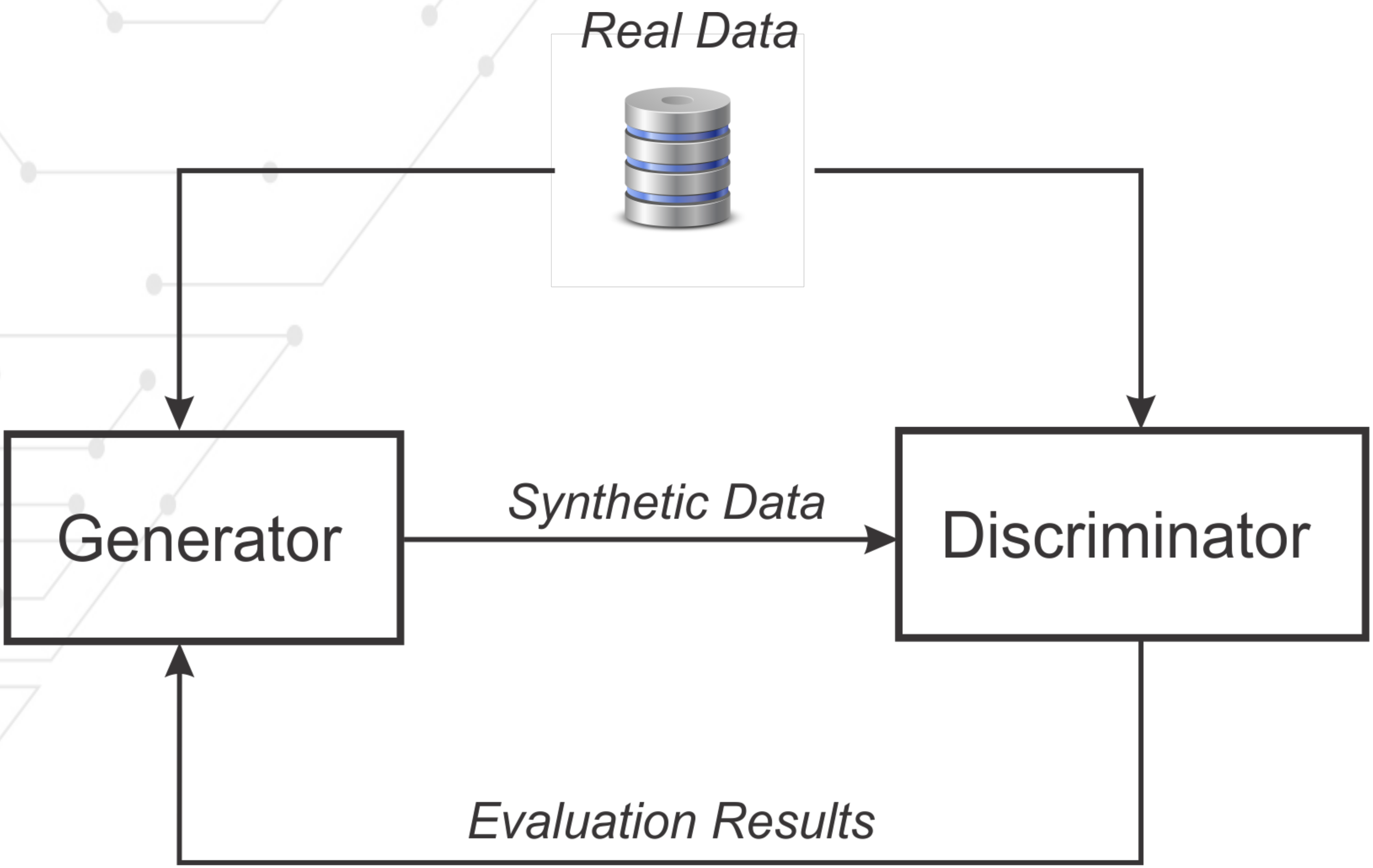
Replica Analytics

Why Synthetic Data ?

- Getting access to good quality realistic data for AI and machine learning projects is difficult (time consuming and costly, and in some cases not possible) – data protection regulations and cross-border data transfer concerns are a major factor causing this
- This slows down the ability to build and test models, and to evaluate models and analytics technologies

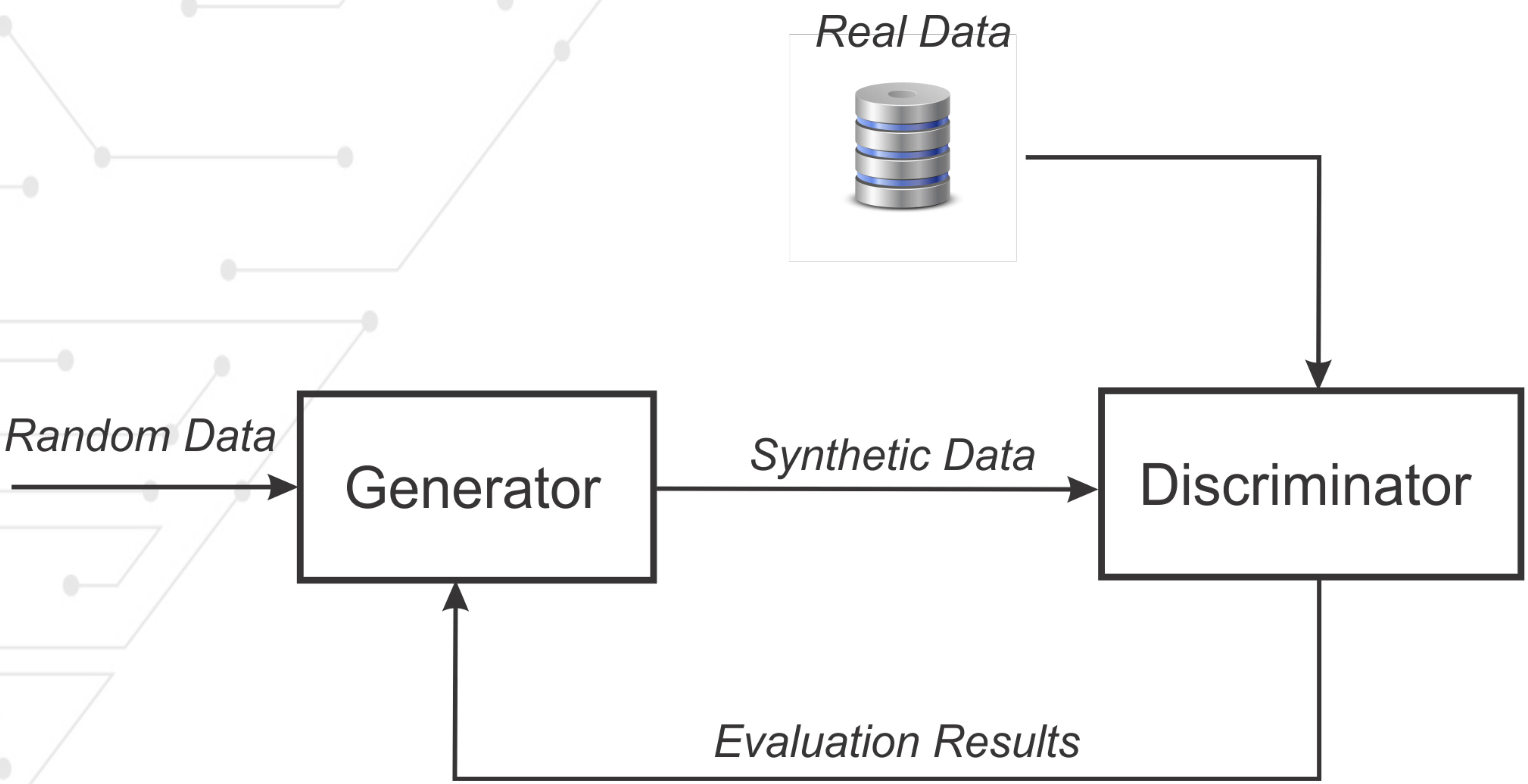
Synthetic data is a cost effective and scalable way to solve this problem

Generation Methods - A



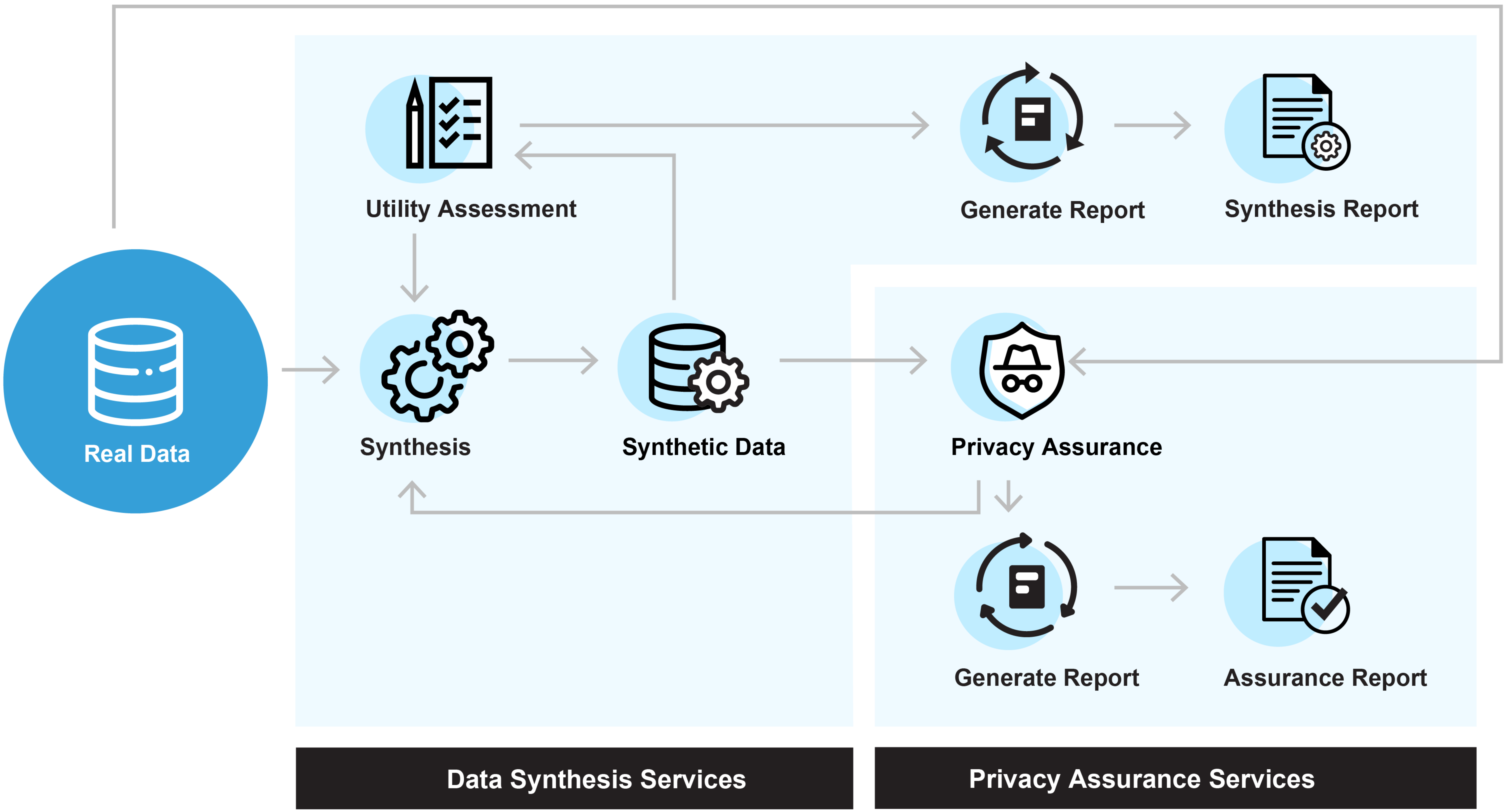
Replica Analytics

Generation Methods - B



Replica Analytics

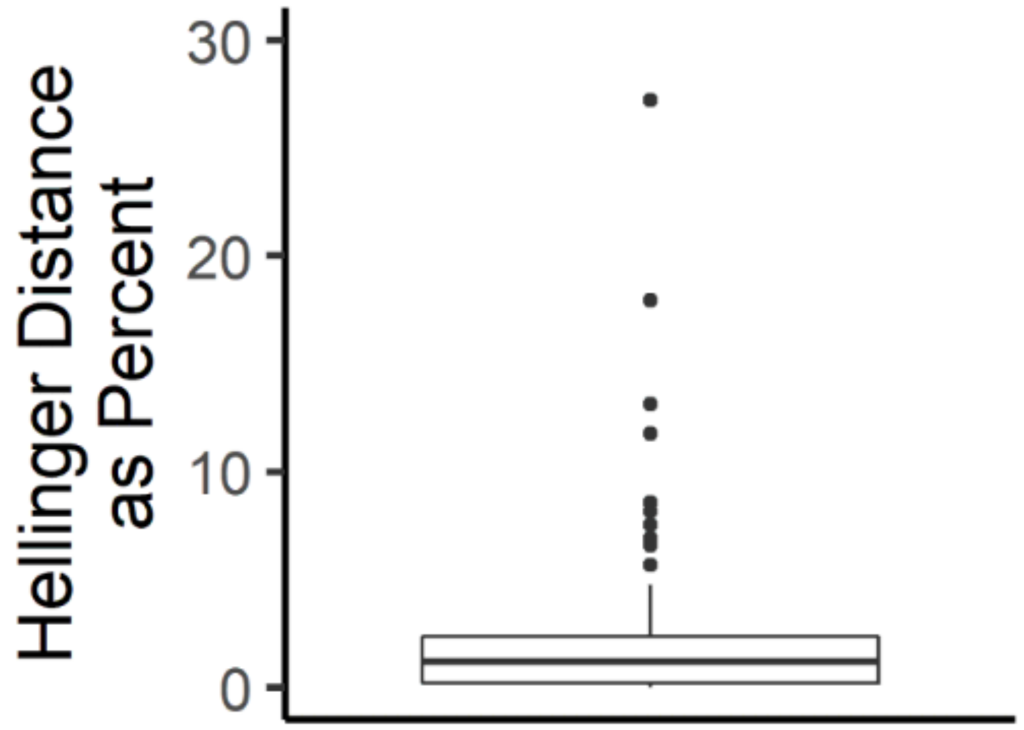
General Workflow



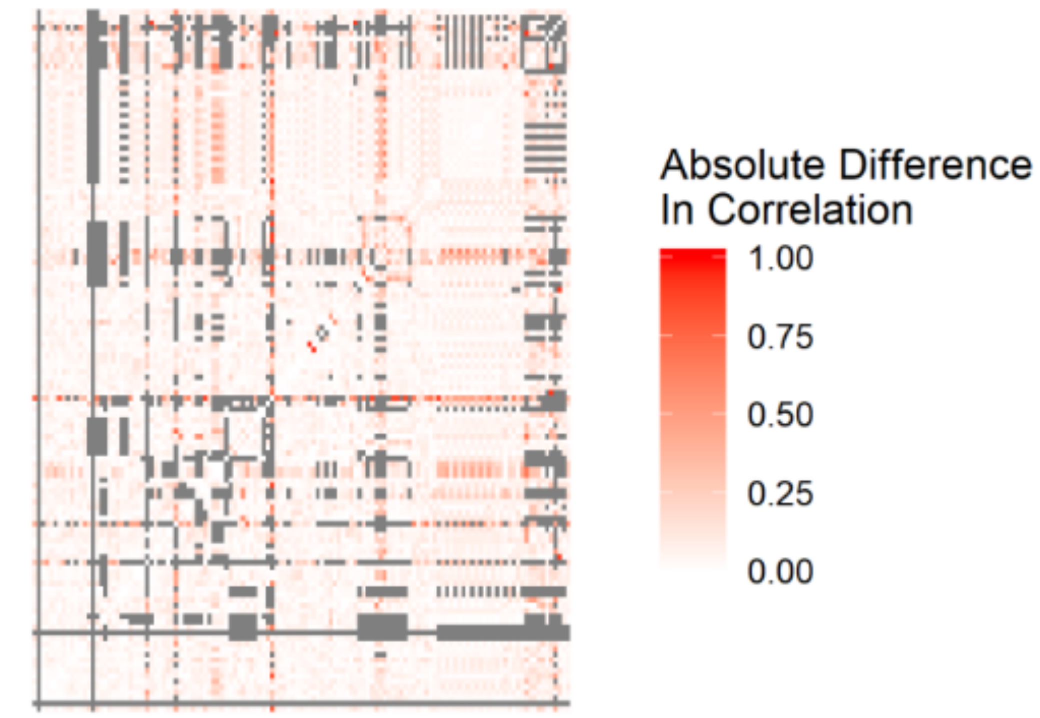
Replica Analytics

How Good Is the Data ?

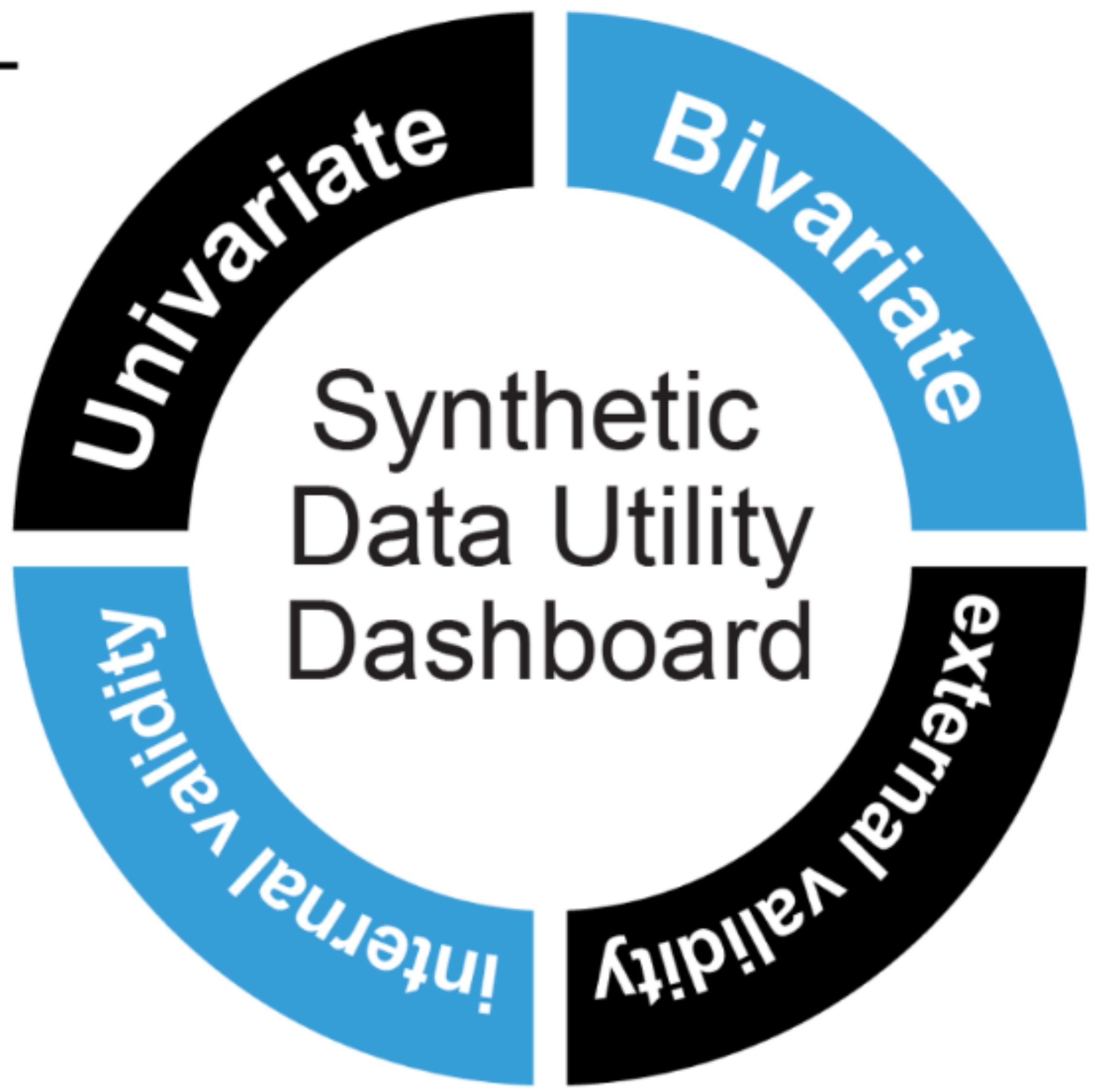
- Utility tests are performed on the data generated
- They compare analysis results from the real vs synthetic data (i.e., the absolute difference between the distributions or parameters for the real vs synthetic data)
- There are different ways to do this:
  - All univariate, bivariate, and multivariate models are compared (all models test)
  - Replicate on synthetic data the published analyses from real data



clinical trial dataset



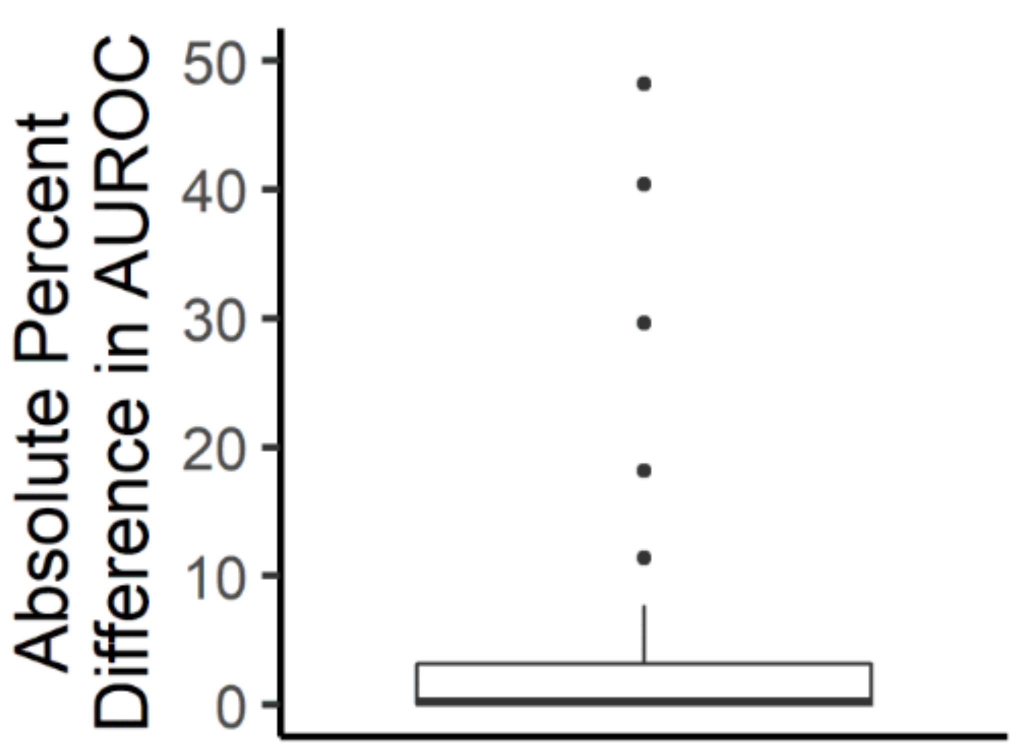
The difference in the univariate distributions between the real and the synthetic data

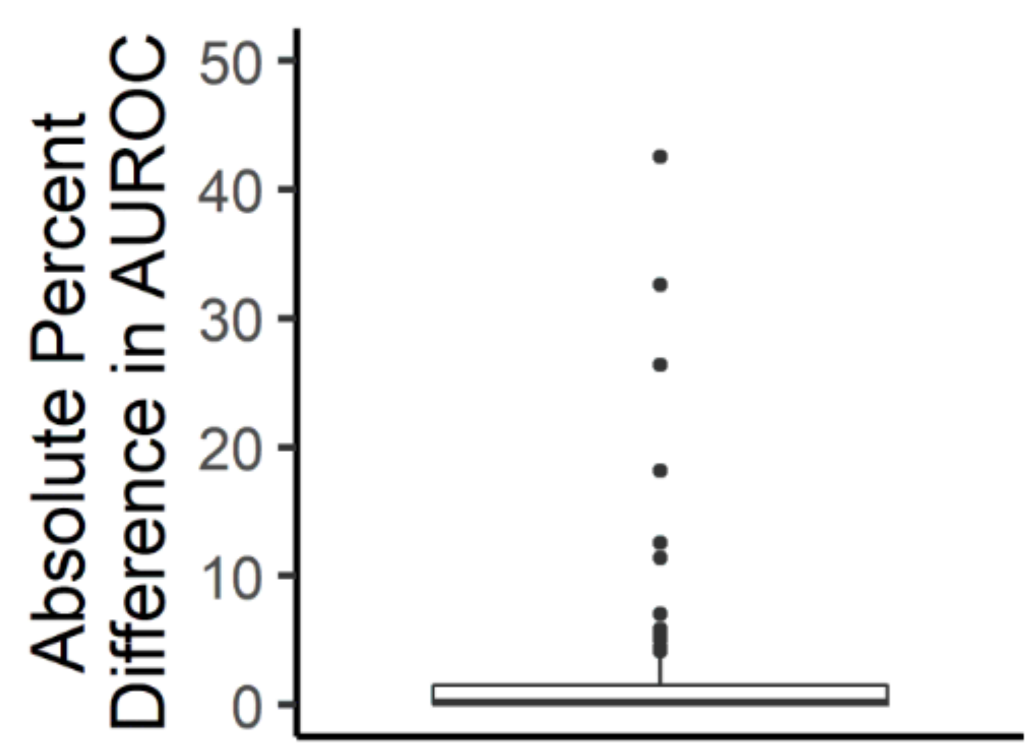


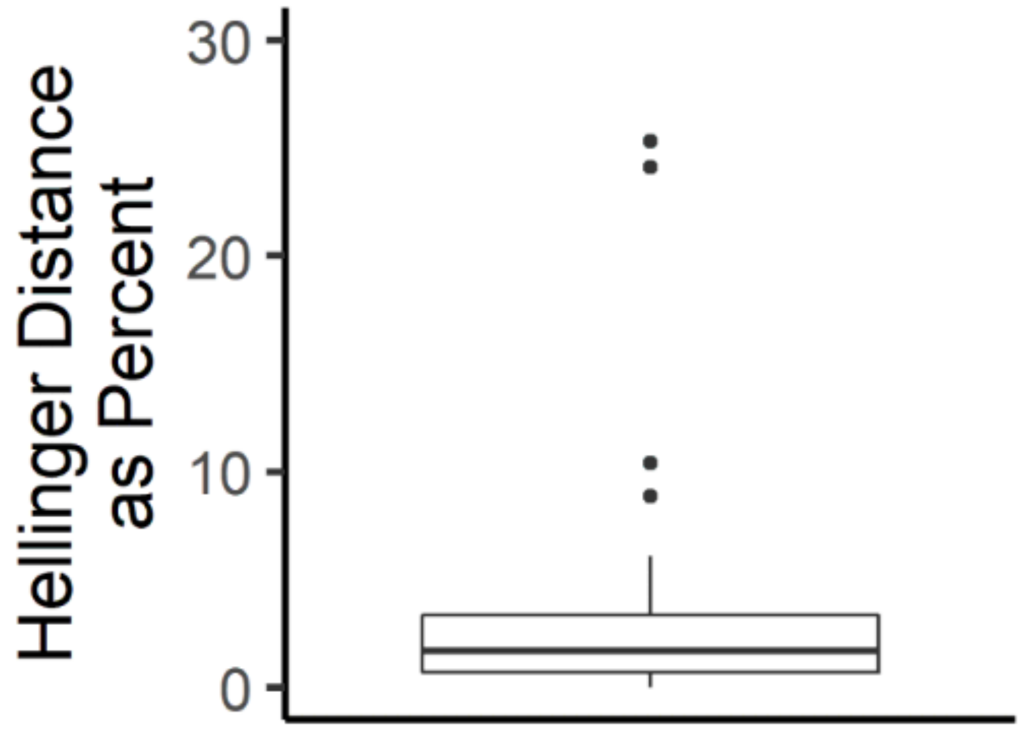
The difference in the correlations between the real and the synthetic data

The difference in all multivariate models' accuracy between real and synthetic data **tested on an internal validity scenario**

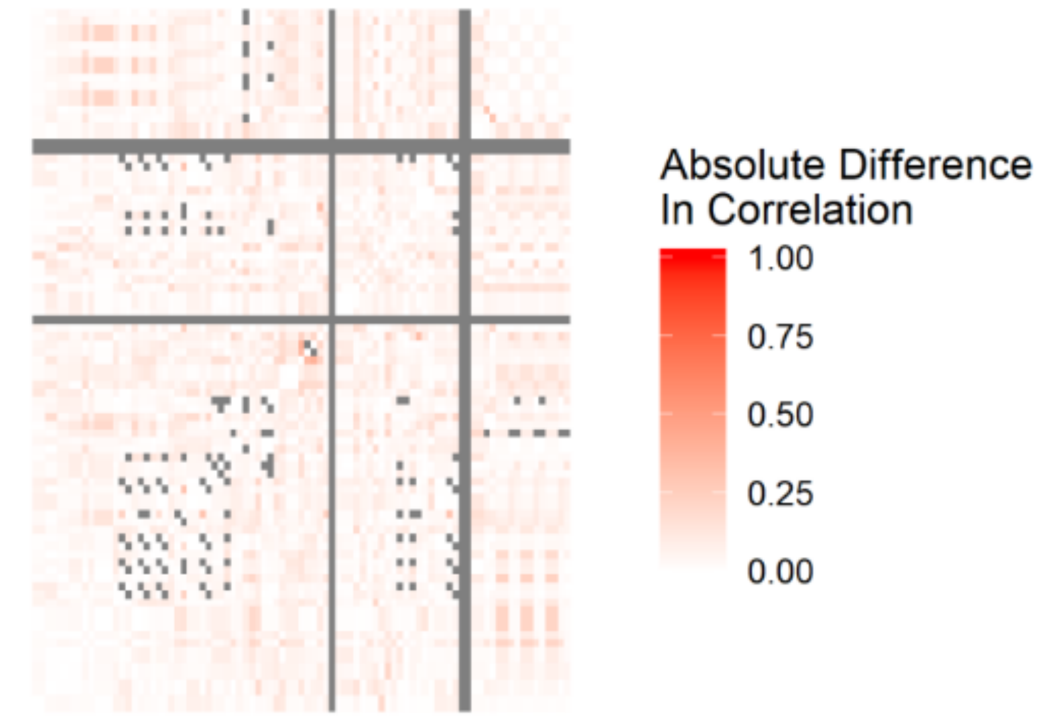
The difference in all multivariate models' accuracy between real and synthetic **data tested on an external validity scenario**



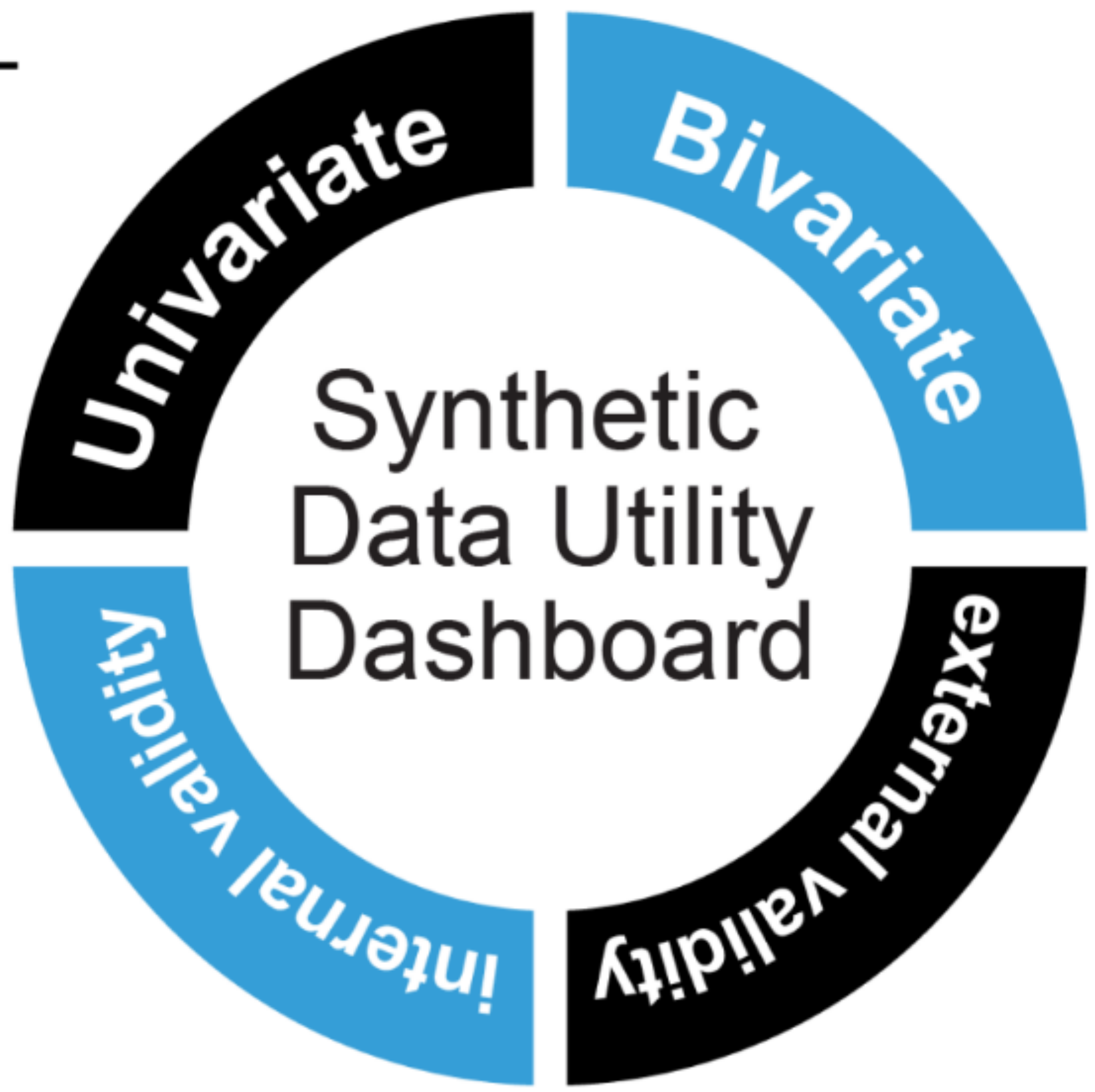




clinical trial dataset



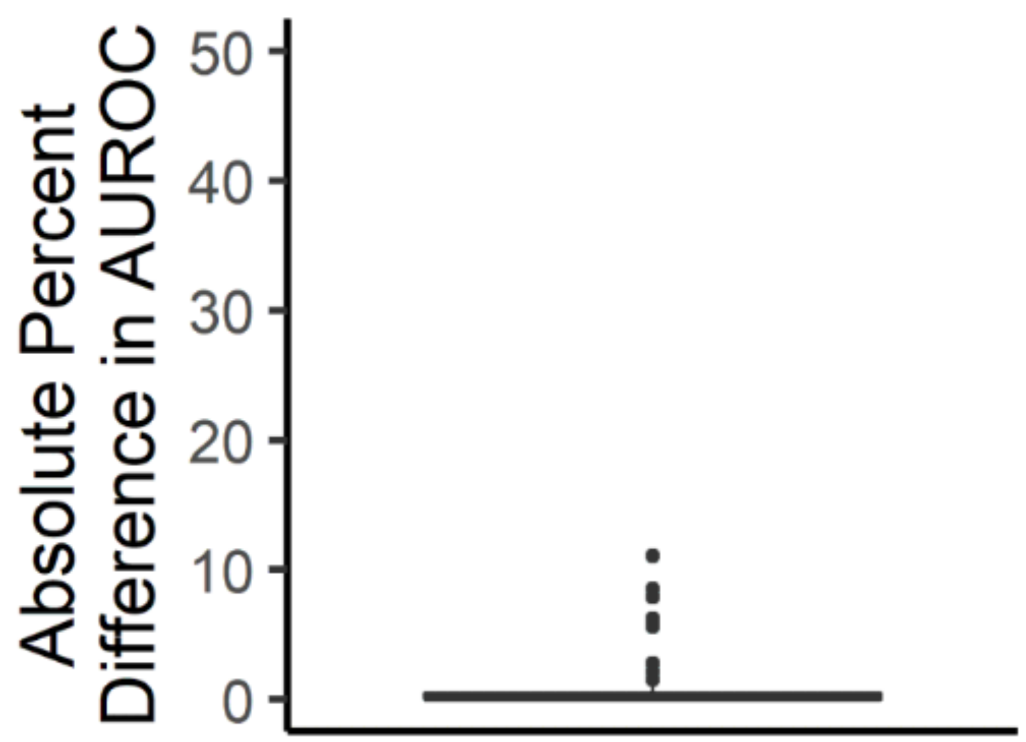
The difference in the univariate distributions between the real and the synthetic data

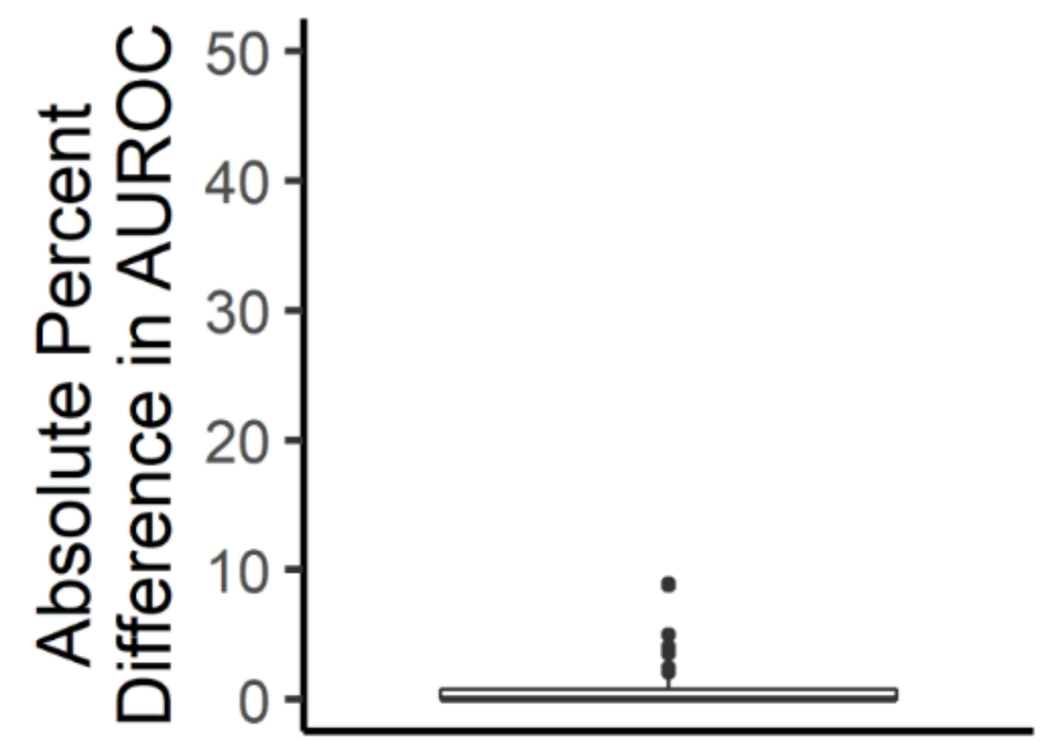


The difference in the correlations between the real and the synthetic data

The difference in all multivariate models' accuracy between real and synthetic data **tested on an internal validity scenario**

The difference in all multivariate models' accuracy between real and synthetic **data tested on an external validity scenario**





Privacy Assurance

- Replica Analytics has a unique privacy assurance framework to quantitatively assess the risk of meaningful identity disclosure:
  1. The likelihood that an individual in the synthetic data can be matched with a real person
  2. If such a match is possible, will the adversary learn something new from such a match (because the data is fully synthetic, even if there is a match the information may be sufficiently different that nothing meaningful would be learned)
- Both tests must pass for a meaningful identity disclosure to have occurred

Replica Analytics

Synthesis CoE

- A data synthesis Center of Excellence (CoE) is an internal resource within the organization to:
  - Generate synthetic data for client projects and for internal analytics and software testing efforts
  - Consult on data synthesis with groups within the business
  - Provide external facing experts to clients, regulators, and the media about data synthesis methods and their application within the enterprise
  - Promote best practices on the use of synthetic data within business units and with clients (an education role)

Replica Analytics

Setting up a CoE

- A Synthesis CoE can be set up by:
  - Providing technology for data synthesis
  - Training the CoE team on data synthesis methods
  - Supports the development of SOPs and policies for the use of synthetic data within the enterprise
  - Advises on the implementation of governance mechanisms supporting the use of synthetic data internally and externally

Replica Analytics

Why Synthetic Data ?

- The data utility is improving rapidly and models built with synthetic data give similar results as the original data
- Data synthesis can be largely automated and scaled – there is little manual effort needed in the generation process
- The number of use cases where this is a good solution is increasing

Replica Analytics

Case Studies

Rebecca Li,
Vivli

Ben Szekely
Cambridge Semantics

Replica Analytics

Vivli

CENTER FOR GLOBAL CLINICAL RESEARCH DATA

Vivli Clinical Research Data Sharing: Share. Discover. Innovate.

Rebecca Li

October 4th, 2019

Vivli is a Global Data Platform – Agnostic to Disease, Funder or Data Contributor

Irritable Bowel Syndrome
Bacterial Peritonitis Glaucoma Endometriosis
Kidney cancer Non Hodgkins Lymphoma Epilepsy HIV
Breast cancer Cystic Fibrosis Diabetes Mellitus Insomnia
Coronary Artery Bypass Surgery Schizophrenia Bariatric Obesity
Atrial Fibrillation Fibromyalgia Cancer Traumatic Brain injury Trauma
Huntington's Disease Dabigatran
Influenza Crohn's Diabetes Hepatitis CHepatitis Autism
Atorvastatin
Hidradenitis Disease Hypertension Myocardial Arthritis
Psoriasis Statin Endometriosis Depression Interleukin-6 Zoloft
Tysabri Tuberculosis Heart-Failure
Bipolar disorder Cannabinoids Asthma Lung cancer Lymphoma
Multiple Sclerosis Sickle Cell disease Atopic Dermatitis
Tumor burden Vitamin D Total Joint Replacement Cancer
Vedolizumab Pulmonary Arterial Hypertension Infarction
Hemophilia Sleep Apnea Edoxaban Type 1 Diabetes Mellitus
HPV Humira Colorectal Cancer Osteoarthritis
Lymphoma Stroke Ulcerative Colitis Vitiligo

Vivli

Vivli Members

- abbvie
- Boehringer Ingelheim
- Biogen
- BioLINCC
  Biologic Specimen and Data Repository Information Coordinating Center
- Celgene
- CRITICAL PATH INSTITUTE
- Daiichi-Sankyo
- DORIS DUKE CHARITABLE FOUNDATION
- Duke UNIVERSITY
- gsk do more feel better live longer
- HARVARD UNIVERSITY
- THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST
- IMMPORT
  BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY
- JOHNS HOPKINS UNIVERSITY
- Johnson & Johnson
- Lilly
- Pfizer
- Project Data Sphere
- Takeda
- UCSF
  University of California San Francisco
- Vivli

How Vivli works

SEARCH

Search Vivli platform for information about available studies.

REQUEST

Request IPD Data sets.

Each Data Request will be **reviewed** according to contributors' publicly stated requirements.

ACCESS

Data from approved requests can be **accessed** in Vivli's secure research environment or **downloaded** with permission.

ANALYZE

Use robust **analytical tools** to combine and analyze multiple data sets.

DISSEMINATE

Completed **research results** will be assigned a DOI.

Researchers may use the Vivli platform to meet their **publication** requirements.

Vivli

Vivli-Microsoft Datathon Scientific Objective

- **Background** - More than 60 individuals formed 11 teams and participated in the first Vivli Microsoft Data Challenge. Participants were from universities, hospitals, pharmaceutical, biotech and software companies.

- **Objective** – To find innovative solutions for how to safeguard participant privacy and minimize privacy loss while maintaining the scientific analytic value of the data for rare disease data sets that are more highly identifiable.

Microsoft

Replica Analytics

Vivli

At a Glance:

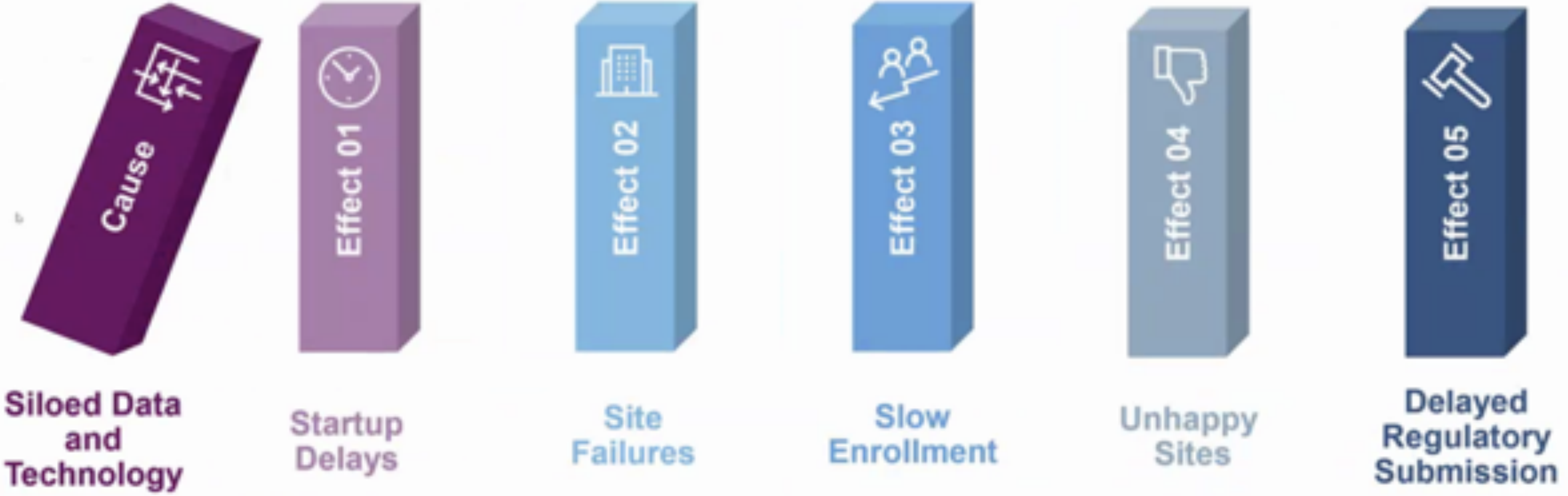
We apply semantics and graph to a data fabric – so anyone can find, understand, blend, and use enterprise data.

- Based in Boston
- 100+ employees
- Origins in IBM and Netezza
- Anzo 4.0 GA 2017
- Added enterprise-scale OLAP graph database engine in 2015

CAMBRIDGE SEMANTICS

Clinical Development Challenges

Domino Effect Complexity



MILLIONS $$ PER DAY

of lost sales

NEWSROOM

Press Release

Gartner Identifies Top 10 Data and Analytics Technology Trends for 2019

Augmented Analytics and Artificial Intelligence in the Spotlight

Trend No. 1: Augmented Analytics
Trend No. 2: Augmented Data Management
Trend No. 3: Continuous Intelligence
Trend No. 4: Explainable AI

Trend No. 5: Graph

Trend No. 6: Data Fabric

Trend No. 7: NLP/ Conversational Analytics
Trend No. 8: Commercial AI and ML
Trend No. 9: Blockchain
Trend No. 10: Persistent Memory Servers

….Graph processing to continuously **accelerate data preparation** and enable more complex and adaptive **data science.**

… to efficiently model, explore and query data with **complex interrelationships across data silos**

….. **the need to ask complex questions across complex data,** which is not always practical or even possible at scale using SQL queries.

…..Data fabric enables **frictionless access and sharing of data** in a **distributed data** environment.

……It enables a single and consistent data management framework, which allows **seamless data access and processing** by design across otherwise siloed storage.

Anzo Difference:

Graph Data Models & Semantics

Simplifies access to complex and blended data to address unanticipated questions

Quickly profiles, connects and harmonizes data from multiple sources, including unstructured textual sources

Presents tailored views and experiences to different personas with conceptual models that use business terms

Flexibly accommodates new data sources and use cases on the fly, with minimal impact

Scales horizontally to accommodate enterprise data fabric scale

A modern data discovery and integration platform for your enterprise data fabric.

Anzo lets business users find, connect, and blend enterprise data into analytic ready datasets.

Map and Explore Enterprise Data

Build Blended Analytic-Ready Datasets

Apply Enterprise-Ready Data Management

Anzo Proof of Concept with Replica Analytics

Our Plan for the PoC

- Blend data into a common model at scale
- Find insights from data across studies, domains, and subjects

Data

- 2 Synthetic Study Datasets
- 12 CSV files

Graph Approaches Applied

- Map study datasets and their metadata to a common graph model
- Use the graph to automate the mappings to the model
- Ask questions on a single graph that contains cleaned, conformed data from across datasets

Summary of PoC Achievements To-Date

Onboard and Model

- Mapped studies to SDTM v3.2 standard
- Conformed 8 domains from 12 CSV files

Blend

- Applied graph model to connect entities via subject relationships
- Connected studies via common entities (e.g. Disease)

Access

- Answered exploratory questions across multiple studies
- Demonstrated integration with Spotfire

SDTM 3.2 Model for the PoC

Centralized key concepts and related properties

- 2 studies
- 8 domains
- 75 properties

Conformed data accessible in Anzo

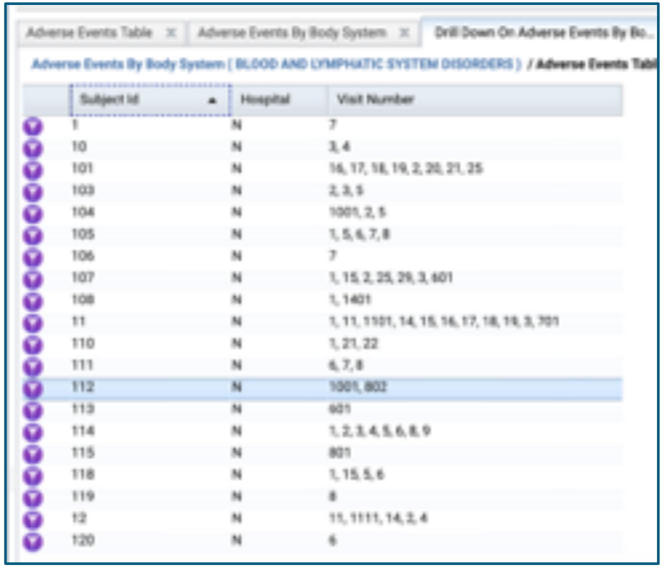
- 9 Classes
- 14 Data Layers
- 5 Dashboards
- 10+ visualizations

Search Across Studies in Dashboards

- Search and analyze related data across studies from a unified dashboard
  - Adverse events by comorbidities
  - Treatments by demographics
  - Responses by labs
- Easily include additional studies

Drill Down to Understand Root Causes

- Identify root causes
- Drill-down to examine trends and outliers



Adverse Events Table | Adverse Events By Body System | Drill Down On Adverse Events By Bo...

Adverse Events By Body System (BLOOD AND LYMPHATIC SYSTEM DISORDERS) / Adverse Events Tabl

Subject Id	Hospital	Visit Number
1	N	7
10	N	3, 4
101	N	16, 17, 18, 19, 2, 20, 21, 25
103	N	2, 3, 5
104	N	1001, 2, 5
105	N	1, 5, 6, 7, 8
106	N	7
107	N	1, 15, 2, 25, 29, 3, 601
108	N	1, 1401
11	N	1, 11, 1101, 14, 15, 16, 17, 18, 19, 3, 701
110	N	1, 21, 22
111	N	6, 7, 8
112	N	1001, 802
113	N	601
114	N	1, 2, 3, 4, 5, 6, 8, 9
115	N	801
118	N	1, 15, 5, 6
119	N	8
12	N	11, 1111, 14, 2, 4
120	N	6

Access Data in Anzo

OData

Query using OData standard based on HTTP/REST

Standard with supporting libraries and integrations with many BI and analytics tools

Java JDBC

JDBC is a standard protocol for connecting Java Applications to relational databases.

Anzo also supports ODBC.

SPARQL

Query the graph through an HTTP endpoint with a SPARQL query

```
SELECT ?mutation ?id
WHERE {
  ?mutation tcga:ssm_id ?id .
}
```

SPARQL is a query language that is familiar to those with SQL experience

Highly flexible for extracting datasets from the complete graph

Hi-Res Analytics

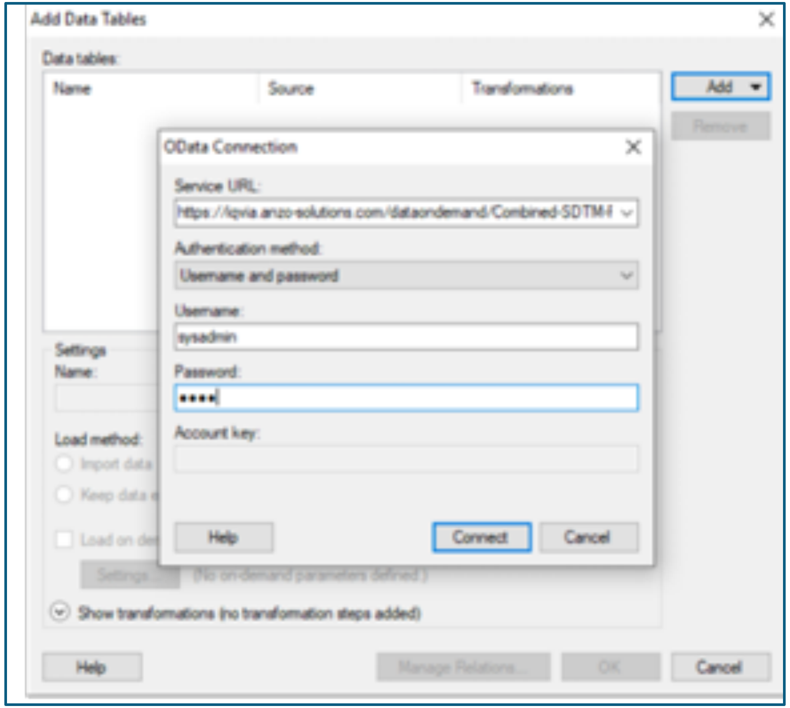
Anzo's exploratory analytics tool for querying and visualizing data

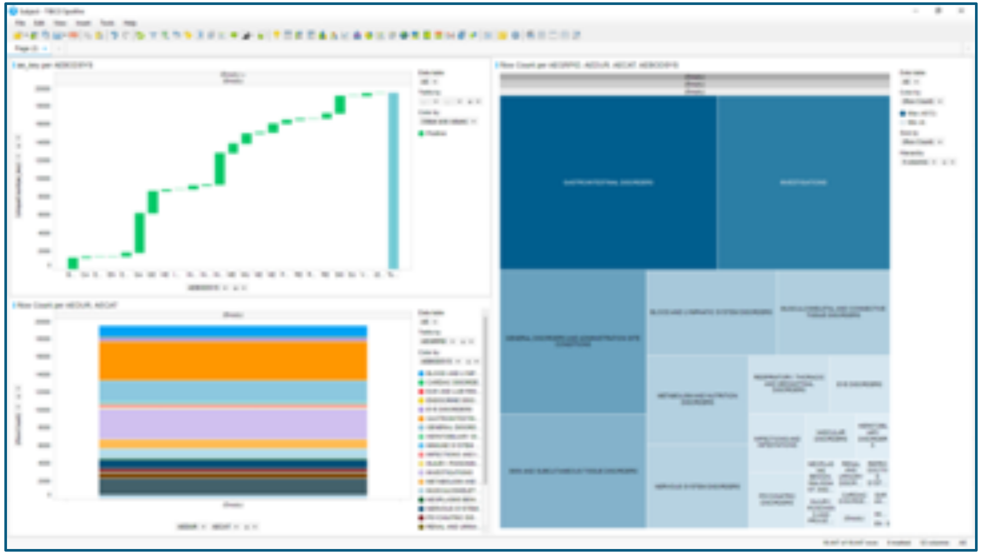
Great for exploration, discovering relationships in data, and drilling-down into hierarchical data.

Easy to create self-service data products for exports with the other tools mentioned here.

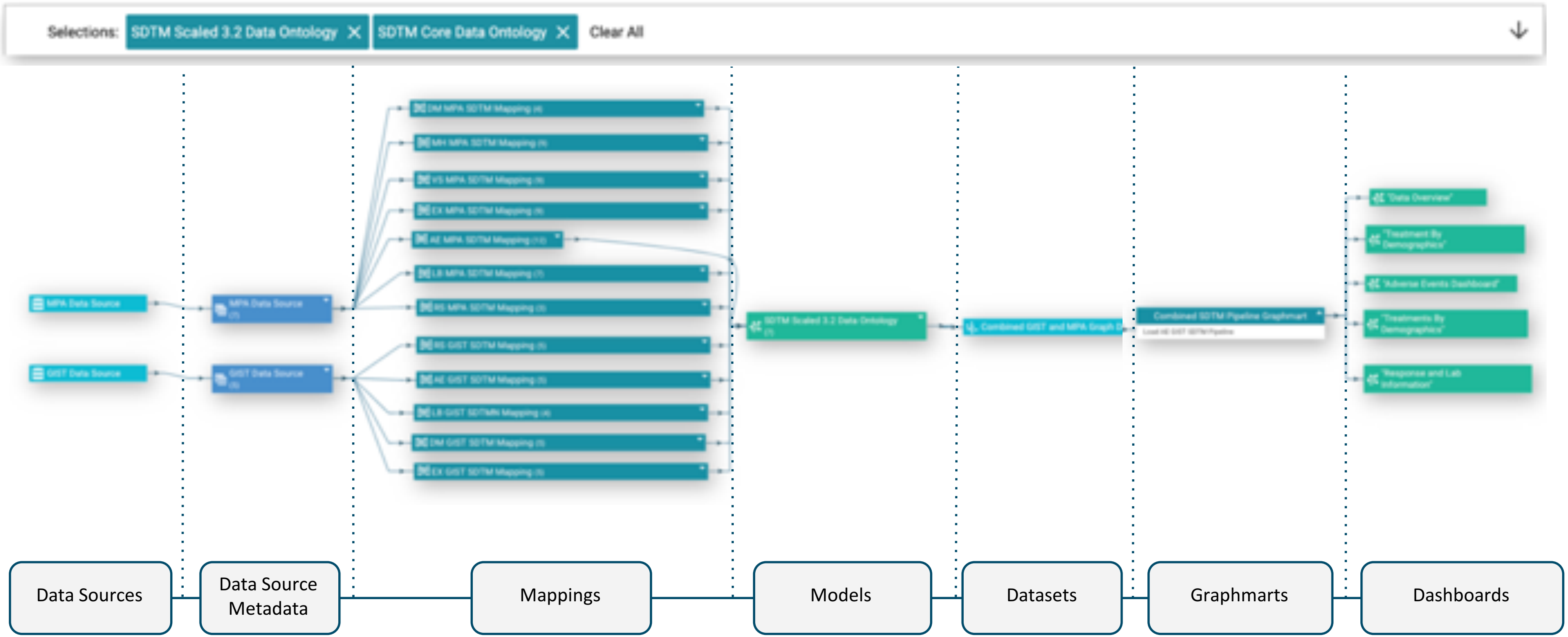
Access Data in Spotfire

All data is accessible in Spotfire via an OData endpoint





End-to-end Data Provenance




QUESTIONS
cc: an untrained eye - https://www.flickr.com/photos/

You will receive

- The materials from this webinar
- Over the next couple of weeks we will send you the reports that were mentioned in this webinar as well:
  - *Accelerating AI Using Synthetic Data* report
  - Technical report on data utility evaluation
  - Notice about the book on synthetic data generation when it comes out in 2020
- An invitation to the synthetic data privacy assurance webinar later this year

Replica Analytics

Next Steps

- If you want to learn more about synthetic data please contact us: info@replica-analytics.com

- You will be asked to opt-in to receive more information from us (technical reports and whitepapers, future webinars on AI and synthetic data, live events that we organize, and newsletters) that may be of interest to you. It is important that you opt-in to continue to receive these communications.

Replica Analytics

info@replica-analytics.com

Lucy Mosquera: lmosquera@replica-analytics.com
Rebecca Li: rli@vivli.org
Ben Szekely: ben@cambridgesemantics.com

Replica Analytics