Agency for
Science, Technology
and Research
SINGAPORE

BII SCIENTIFIC CONFERENCE

Ming Zhen TAN

Asst PI, BII

Biomed DAR

29 March 2022

BioMed DAR: Structure and Operatives

<u>**Mission statement :**</u>

To operationalize a SSSO (Standard Systems Support Office) for clinical research data management to support strategic (past, current, future) A*STAR BMRC programmes and beyond to health clusters

@BMRC (accountability)

@BII (responsibilities)

Data Governance

Dr. Max Fun (Project manager)

<u>IAF programs</u>
- ATTRaCT†
- CaliBre†
- WCIT†
- GUSTO†

DSM Ops

Dr. Tan Ming Zhen (Assistant PI)

supports

S/W DevOps

Tay Weihong (s/w development manager)

† currently supported

Governance
- *oversees data access (DAC)*
- *ensures compliance (security+regulations)*

Repository
- *DM ops (data life cycle)*
- *Synthetic Data Lab (data modeling)*

Architecture
- *S/W engineering*
(Tech refresh, app dev)
- *Pen testing*

DAC : Data Access Committee
DSM : Data Science Management
ML : machine learning
NLP : Natural Language Processing
S/W DevOps: Software Development & Operations

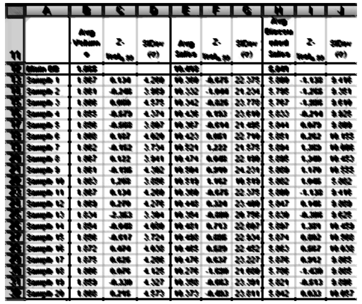
Edit	Accountability	Responsibility
Introduction	In ethics and governance, accountability is answerability, blameworthiness, liability, and the expectation of account-giving.	Responsibility may refer to: being in charge, being the owner of a task or event.
Explanation owed	Yes	Not necessarily

Synthetic Data Lab

Extant Clinical Data Sharing and Protection Framework

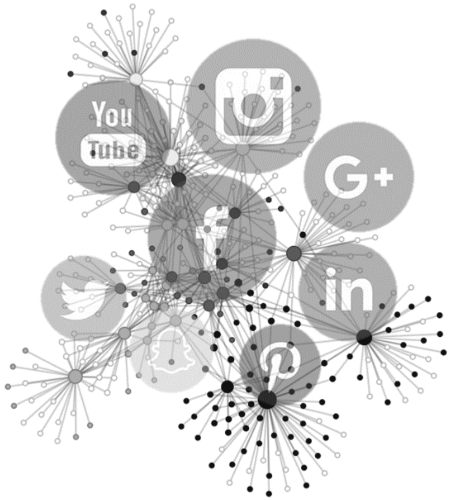
Sources of Auxiliary Information

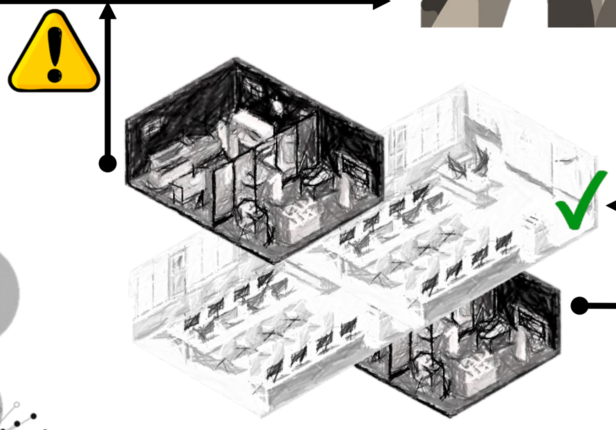
A combination of auxiliary knowledge sourced from social networks and other sources augments the tools at the adversary’s disposal. Recovery of pertinent, sensitive information of the data owner from sanitized / deidentified data is already a distinct possibility.

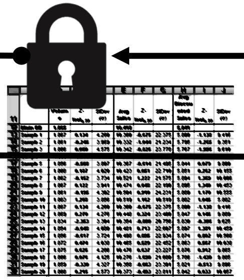


Accumulation of Data

Methodical accumulation of a huge trove of clinical data heralds the advent of data-centric scientific research in the biosciences.





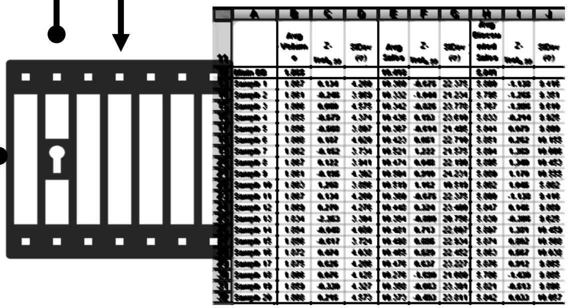


Cryptographic Protections

Encrypted/Sanitised data cannot be directly reinstated to their original form and can be released for wider collaborative efforts. Improving sanitisation requirements and cryptographic methods keep prying eyes at bay.

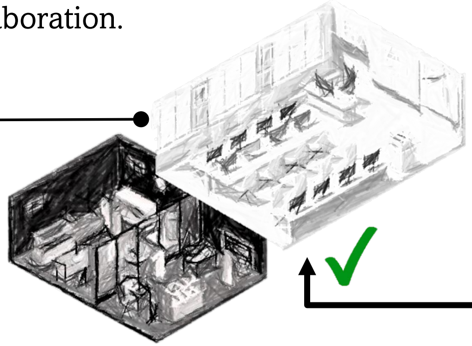
Wider Collaborative Efforts

Sanitised/Deidentified databases released to the wider community for industrial and academic research. Loose controls over data usage and movement facilitates collaboration.



Authorised Entities

Authorised Individuals and Entities use data responsibly. Data movements monitored and tracked.



Security Frameworks

Security Frameworks and agreements limit data access to authorised individuals and entities. Prohibitive data sharing and usage policies frustrates attempts at data leakage.

Last Updated: 21 March 2022 MZ

Data Sharing and Protection Framework with SYNTHETIC DATA

CREATING GROWTH, ENHANCING LIVES

Wider Collaborative Efforts

Collaboration with the wider community can be done at minimal risk of privacy loss.

Reduced Red Tape

Reduced efforts to adhere to personal data protection requirements because data did not come from any person.

No Connection to Data Source

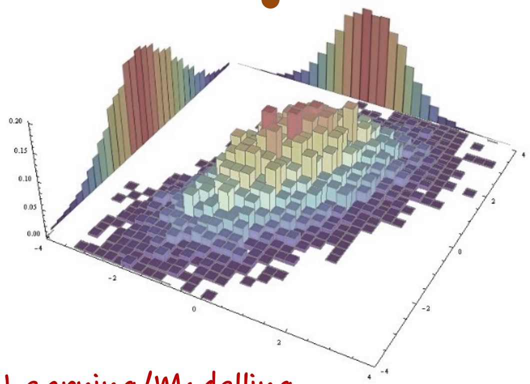
Generated synthetic data has no conceivable personal links to the individual from which the raw data was sampled. Auxiliary information contributes little towards privacy infringement.

Synthetic Data

Synthetic data sampled from the learned probability distributions. Process is inexpensive as compared to traditional clinical data collection and a new set can be generated for different users or uses.

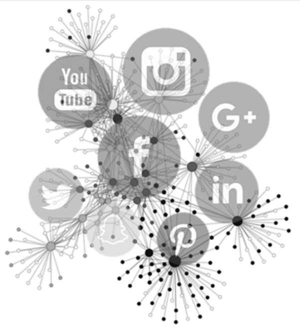
Major Utility

Synthetic data shares the same probability distributions with the raw data and can be used as a substitute for algorithm development, etc.



Sources of Auxiliary Information

A combination of auxiliary knowledge sourced from social networks and other sources augments the tools at the adversary's disposal. Recovery of pertinent, sensitive information of the data owner from sanitized / deidentified data is already a distinct possibility.



Accumulation of Data

Methodical accumulation of a huge trove of clinical data heralds the advent of data-centric scientific research in the biosciences.

Learning/Modelling

The underlying data distributions (univariate & joint) are learned/modelled from the raw data. This forms the basis from which new data could be synthesised.

Cryptographic Protections

Encrypted/Sanitised data cannot be directly reinstated to their original form and can be released for wider collaborative efforts. Improving sanitisation requirements and cryptographic methods keep prying eyes at bay.

Wider Collaborative Efforts

Sanitised/Deidentified databases released to the wider community for industrial and academic research. Loose controls over data usage and movement facilitates collaboration.

Major Cost Savings

Synthetic data does not require encryption-at-rest protocols, secure database warehousing, auditing, cryptographic cleansing, approvals, collaboration agreements, etc.

Authorised Entities

Authorised Individuals and Entities use data responsibly. Data movements monitored and tracked.

Security Frameworks

Security Frameworks and agreements limit data access to authorised individuals and entities. Prohibitive data sharing and usage policies frustrates attempts at data leakage.

Last Updated: 21 March 2022 MZ

CREATING GROWTH, ENHANCING LIVES

Possible Applications / Collaborations for Synthetic Data Lab

- **Types of Synthetic Data**
  - Tabular Clinical Data
  - Medical Images (CT scans / MRI)
  - Time Series (fMRI signals, ECG signals)
  - Medical Records
- **Applications of Synthetic Data**
  - Training Datasets to accelerate development of algorithms
  - Independent Test Datasets to validate algorithms
  - Mock data generation for systems development
  - Mock data generation for improved statistical power
  - Combination of multiple data sources to create joint probability distribution

Synthetic Data Lab Real Working Example

REAL TABULAR DATA EXAMPLE: ATTRACT DATA SET

Number of Variables

315

Number of Data Points

3964

Longitudinal Data Collection (#Variables)

Baseline: 85
Baseline/6 months: 45
Baseline/6 weeks/6 months: 88
Baseline/6W/6M/1Y/2Y/3Y: 94
Reference: 3

Longitudinal Data Collection (#Data points)

Baseline: 912
6 weeks: 428
6 months: 420
1Y: 816
2Y: 787
3Y: 601

Type of Variables

Date: 41
Mixture (numeric): 24
Nominal: 107
Numeric: 80
Ordinal: 10
Percentage: 1
Range: 1
String: 5
Text: 44
Unknown: 2

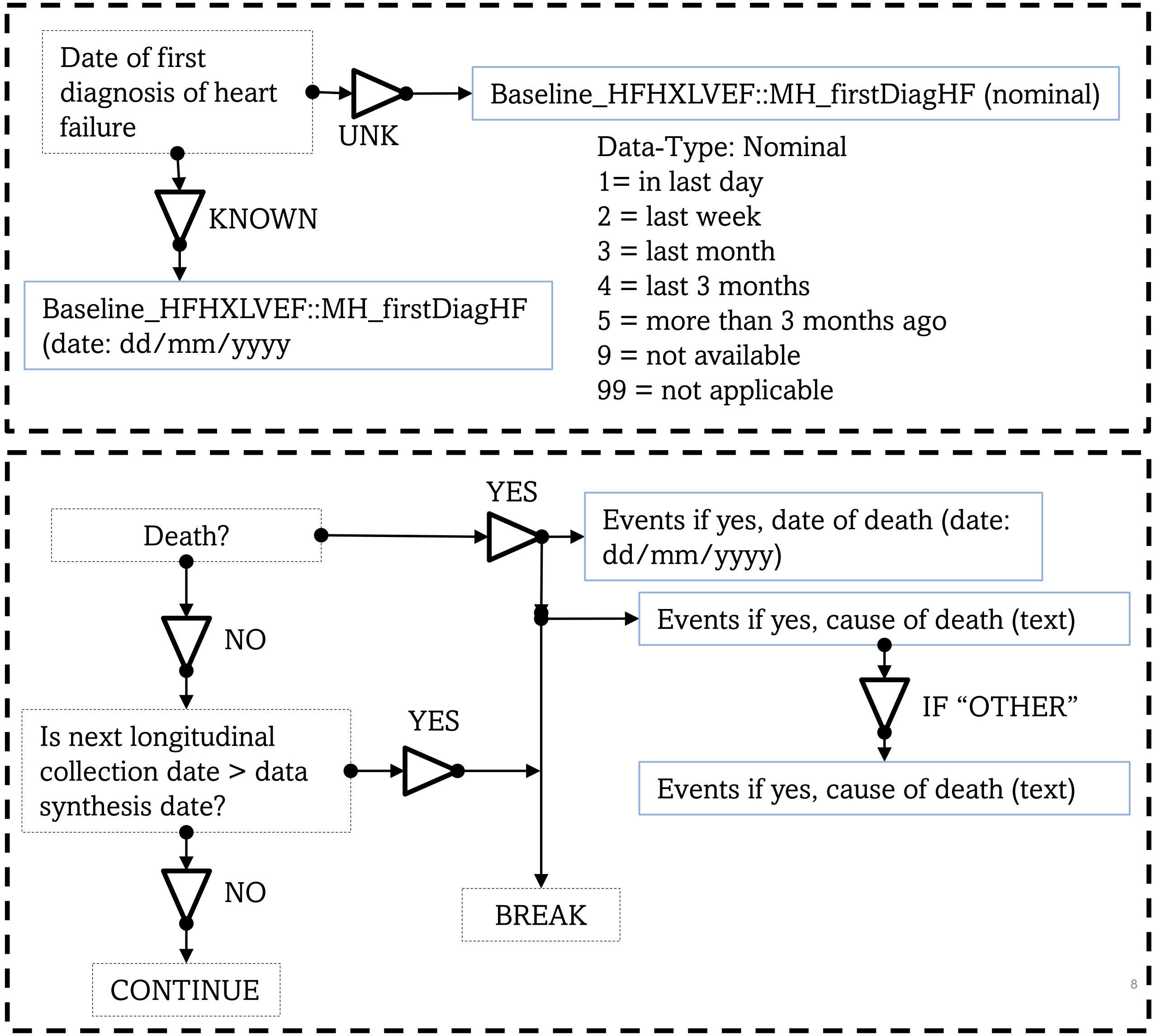
Missing Datapoints

Yes

Date Dependencies

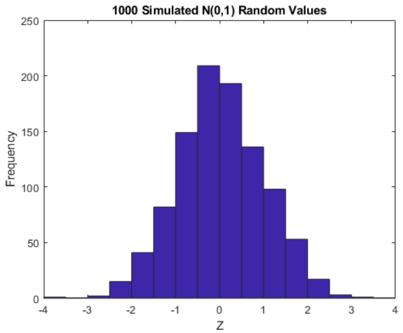
Yes

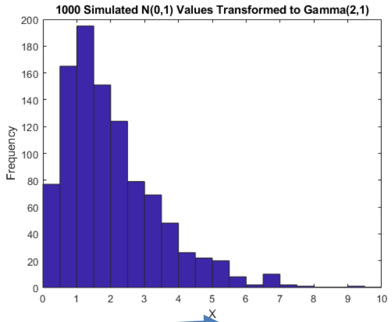
Tree Dependencies



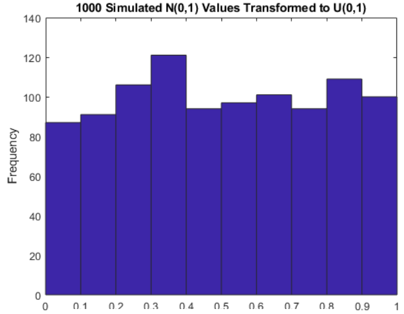
Marginal Probabilities

The various multivariate probability distributions are a generalisation of univariate probability distributions to one or more variables.





The probability integral transform of a random variable is a random variable that is uniform on [0,1]



$$F_i(x) = \Pr[X_i \leq x]$$

Applying the inverse probability integral transform of a distribution F to a random variable gives a random variable whose distribution is F.

$$(X_1, X_2, \ldots, X_d)$$

$$(U_1, U_2, \ldots, U_d) = (F_1(X_1), F_2(X_2), \ldots, F_d(X_d))$$

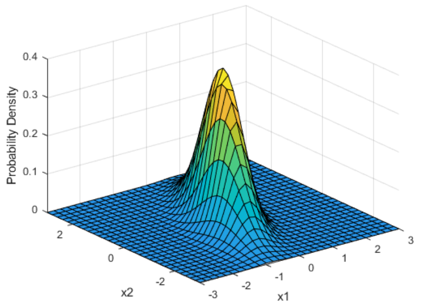
$$(X_1, X_2, \ldots, X_d) = \left(F_1^{-1}(U_1), F_2^{-1}(U_2), \ldots, F_d^{-1}(U_d)\right).$$

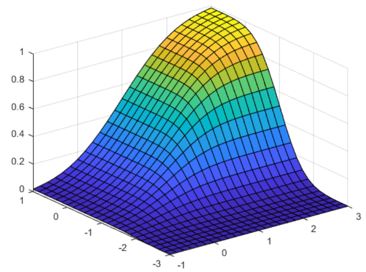
Joint Probabilities

The various multivariate probability distributions are a generalisation of univariate probability distributions to one or more variables.

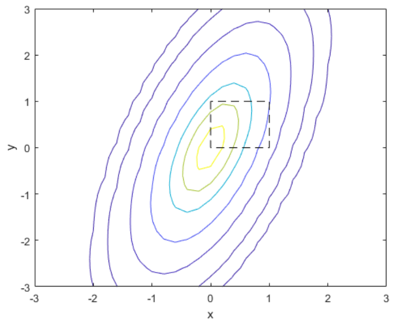
Closed form of PDF of d-dim multivariate normal dist.

$$y = f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} \exp\left(-\frac{1}{2}(x-\mu)\,\Sigma^{-1}(x-\mu)'\right)$$





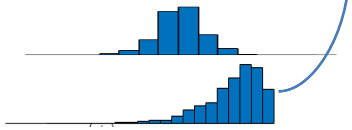
If we have the closed form of a multivariate PDF/CDF, we can generate sets of values from the multivariate PDF.



But we don't have closed forms for varying marginal distributions with varying correlations.

Copula

$$(X_1, X_2, \ldots, X_d)$$

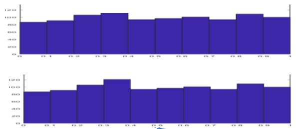


Compute data correlations from training data

Use correlations in some known form of multivariate probability distribution for uniform RV on [0,1].

Sample from the multivariate probability distribution of random variables.

$$(U_1, U_2, \ldots, U_d)$$



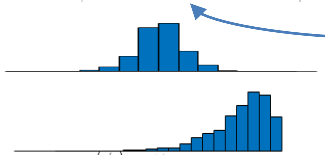
COPULA: dependency structure between marginals.

$$\Pr[U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d].$$

A CDF based on uniform random variables on [0,1]

Synthetic

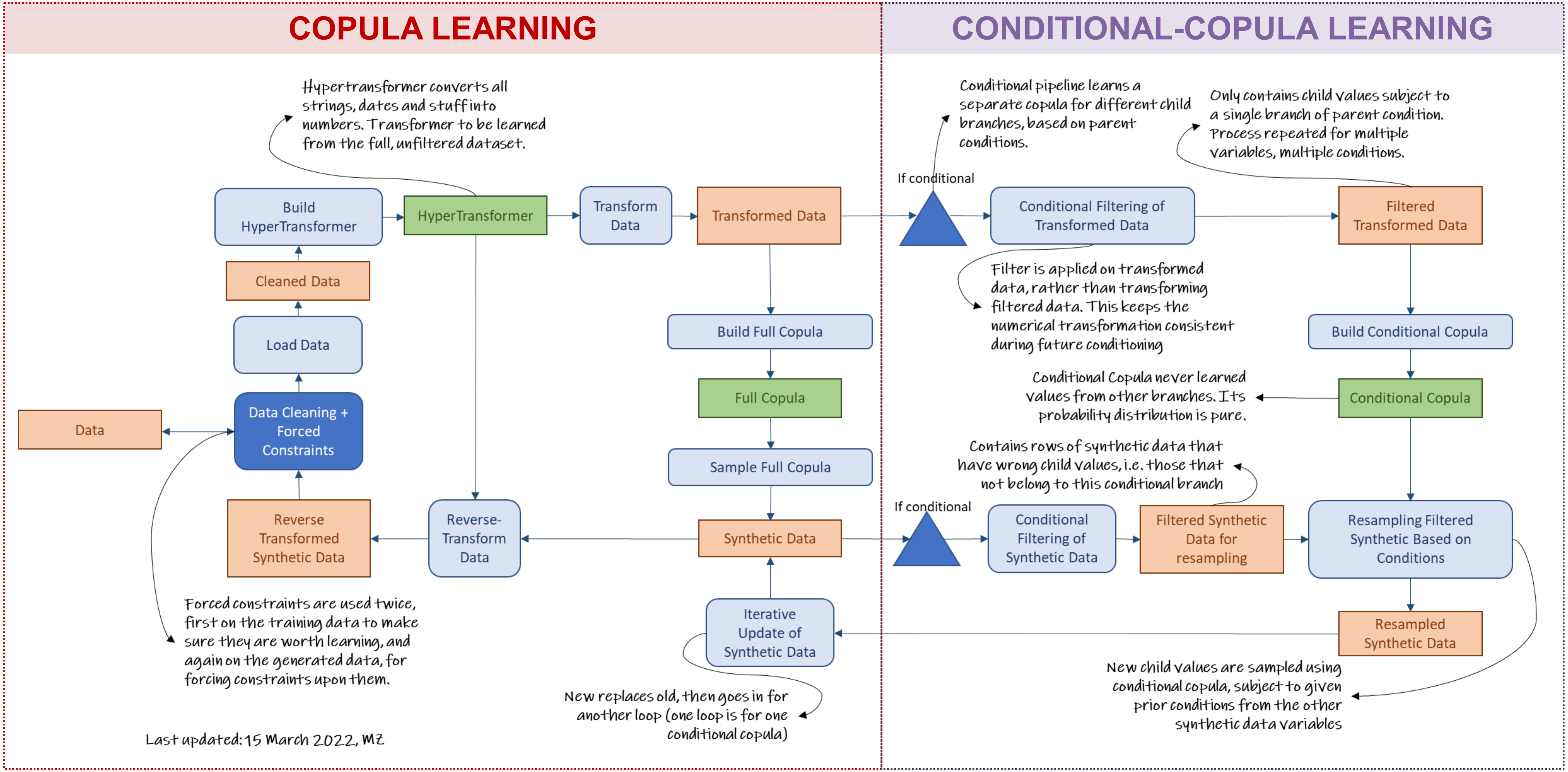
$$(X_1, X_2, \ldots, X_d)$$



Do inverse transform on sampled variables, based on what I think their marginal distribution is.

$$\left(F_1^{-1}(U_1), F_2^{-1}(U_2), \ldots, F_d^{-1}(U_d)\right)$$

MODELLING JOINT DEPENDENCIES



This pipeline uses packages from the Synthetic Data Vault project. https://sdv.dev

CREATING GROWTH, ENHANCING LIVES

QUICK GLANCE

Experiment

Multivariate copula learning:

- 912 baseline datapoints;
- 315 variables;

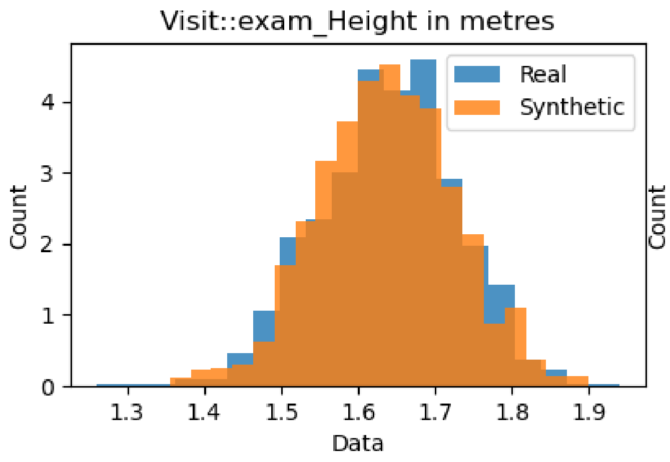
Qualitative Look at Histograms

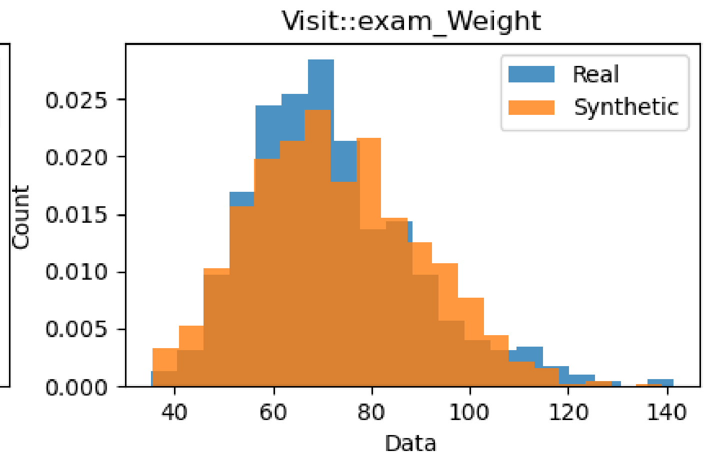
Qualitative inspection of histogram plots of individual continuous variables. Multivariate copula can model univariate continuous distributions to reasonable accuracy.

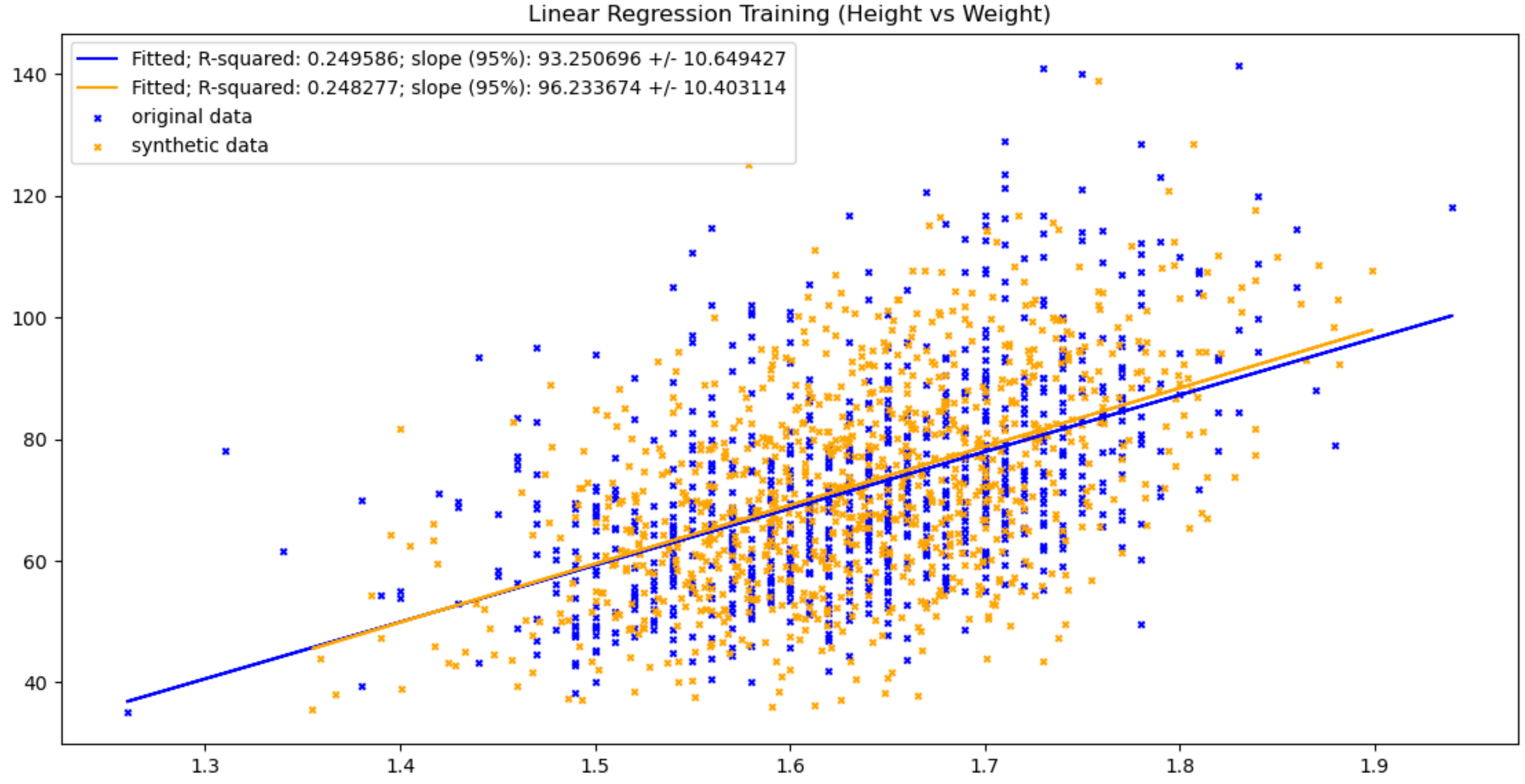
Qualitative Look at Regressions (Linear)

Simple linear regression using height and weight variables from both original and synthetic data. Figure shows superimposed scatterplots and regression trendlines from both the original and synthetic datasets. There is visible high overlap of the two scatterplots, though slope and intercept confidence levels are slightly different.

Original slope (95%): 93.250696 +/- 10.649427
Original slope (95%): 96.233674 +/- 10.403114
Original intercept (95%): -80.618723 +/- 17.465537
Original intercept (95%): -84.849367 +/- 17.052326







CREATING GROWTH, ENHANCING LIVES

Agency for Science, Technology and Research

SINGAPORE

THANK YOU

www.a-star.edu.sg