SESSION 5: APPLICATIONS IN COMPLEX HEALTHCARE SETTINGS

uOttawa

Replica Analytics
AN AETION COMPANY

GENERATING SYNTHETIC DATA FOR THE NHS

Presented by:

Jonathan Pearson,
Lead Data Scientist,
NHS England

Generating Synthetic Data for the NHS

Synthetic Data Summit -

30th November 2023

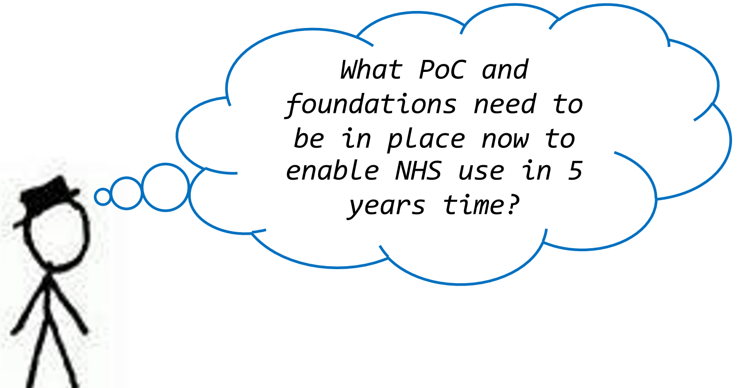
Dr. Jonny Pearson
Lead Data Scientist
Digital Analytics and Research Team
NHS England
jonathanpearson@nhs.net

Who We Are and Where We Fit

Academia (and industry research)

Cutting edge models and thinking around getting value from large and complex data sources.

- Work often siloed and caught in long term projects
- Application often focussed on edge cases and ideal circumstances
- Sometime lacking real world or domain specific application



DART Innovation

Work with academia and NHS Ops to develop **both** push and pull (in different time scales)

- Short tangible outputs that clearly build towards wider context
- Take risks with expectation of high rate of failure
- Include development cycle and tech transition plan

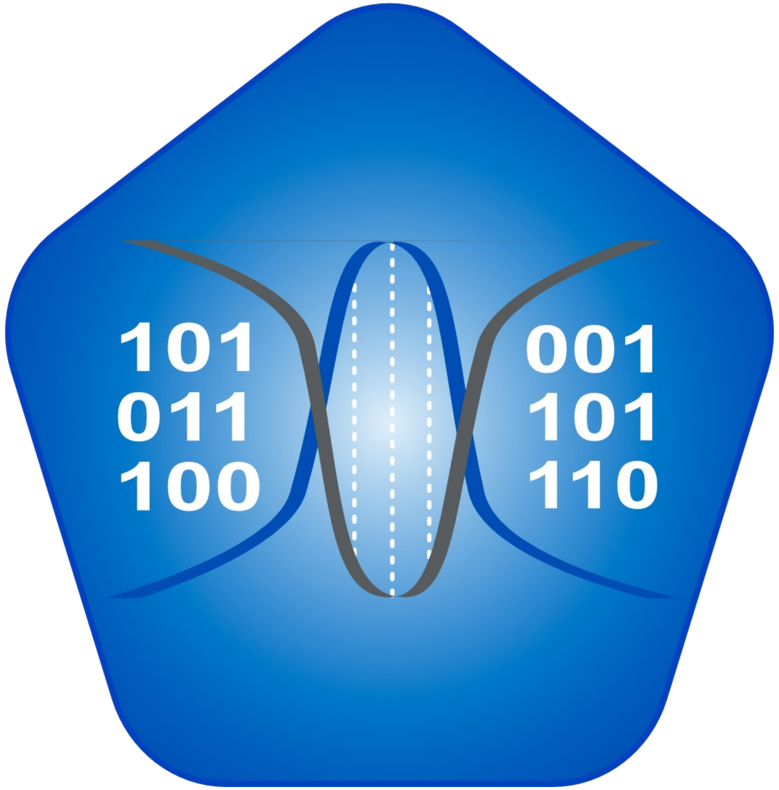
NHS Operations and Decision Making

Need often driven by short-term priorities reducing desire for R&D.

- Evidence-based decision making from robust data insights
- Live modelling and visualisation of data to support daily operations
- Linking data across a complex landscape

What’s Coming

- Fidelity - Simple is Often Better
- The Generation Landscape
- Our Approach
- Evaluating the Data
- What to care about (for healthcare)

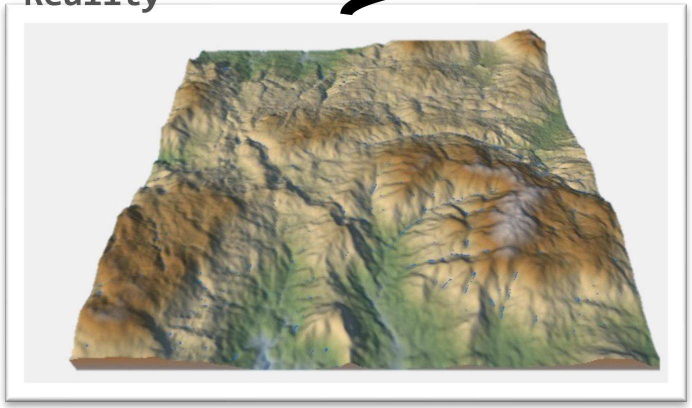


Synthetic Data

Setting the Scene

Synthetic data generation is not a silver bullet and often not an easy alternative, but it does have huge potential for datasets that are too low quality to use, too sensitive to share or, just doesn’t exist.

Reality



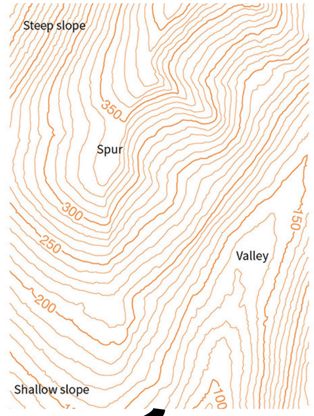
Data Quality and Bias Issues

Raw Data

lame	Waypoint	Town	County	Grid Ref	Current Use	Historic Use	Type	Co.
[Great Gidding Windmill Apex]	TP9905	OUNDLE	Cambridgeshire	TL122831	None	3rd order	Intersected Station	
Site RAF Boiler House Tower Heald Green Centre	TP13396	WILMSLOW	Greater Manchester	SJ852855	none	4th order	Intersected Station	
Site Tower	TP14709	HEYWOOD	Greater Manchester	SD856091	none	4th order	Bolt	
202 F D Building	TP17796	WREXHAM	Clwyd	SJ386495	none	4th order	Rivet	
208 Length Hut	TP14096	WIDNES	Merseyside	SJ475862	none	4th order	Rivet	
RAF Boiler House Tower [illegible] Centre	TP13395	WILMSLOW	Cheshire	SJ849838	none	4th order	Intersected Station	

Model and Transformations

Representation



Fidelity, Fairness, Privacy Metrics

Generated Data

[illegible] Seaside	[illegible]	EASTBOURNE	East Sussex	[illegible]	none	3rd order	Bolt	
53 Colebrook Road	TP24951	SOUTHBOROUGH	Kent	TQ591415	none	3rd order	Bolt	
A Chroic	TP12933	TOBERMORY	Strathclyde Region	NM229627	none	4th order	Bolt	
A Chruach	TP8120	FORT WILLIAM	Highland Region	NM771977	None	3rd order	Bolt	
A E C Works	TP11380	ELLESMERE PORT	Cheshire	SJ416766	none	4th order	Rivet	
A H Palmers	TP12059	GLOUCESTER	Gloucestershire	SO818155	none	4th order	Bolt	
A I D Laboratories	TP23764	HAREFIELD	Greater London	TQ051903	none	4th order	Rivet	
[illegible] Tower	TP24850	CRAWLEY	West Sussex	TQ274385	none	4th order	Rivet	

Use-Case and Amendments

Fidelity - Simple is Often Better

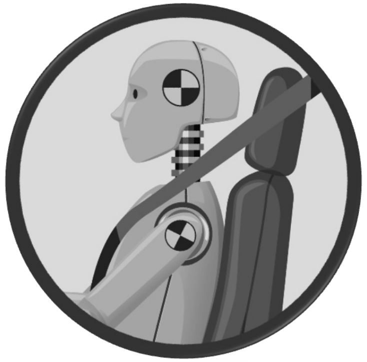
NHS

***Range of Fidelity** (how similar the generated data is to the ground truth)*

Structural → Valid → Plausible → Multivariate Plausible → Multivariate Detailed → Replica

Source: [Office for National Statistics]

Range of Use-cases

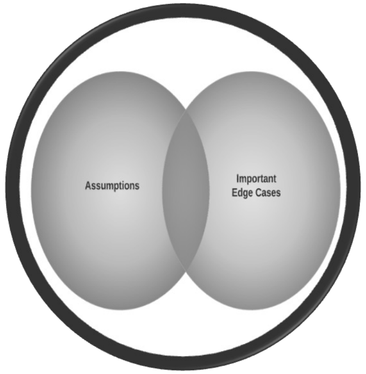


End-to-end software testing
e.g. Interoperability of architectures and systems to pass health data in FHIR formats

Tool Demonstration
e.g. New geospatial tool for showing impact of service planning on travel distance

Faster Innovation
e.g. Internal or external development of patient safety report classifier

Novel Linkage
e.g. generation of patient cancer pathways



Evaluation of Solutions
e.g. test clinical risk score prediction on rare patients.

Addressing Bias and Quality
e.g. creating a de-biased data set to highlight the impact that bias is having on the real data

Fidelity - Simple is Often Better

Range of Fidelity *(how similar the generated data is to the ground truth)*

Structural → Valid → Plausible → Multivariate Plausible → Multivariate Detailed → Replica

Source: [Office for National Statistics]

Range of Use-cases

End-to-end software testing

e.g. Interoperability of architectures and systems to pass health data in FHIR formats

Tool Demonstration

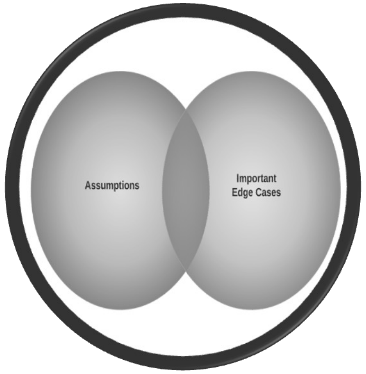
e.g. New geospatial tool for showing impact of service planning on travel distance

Faster Innovation

e.g. Internal or external development of patient safety report classifier

Novel Linkage

e.g. generation of patient cancer pathways



Evaluation of Solutions

e.g. test clinical risk score prediction on rare patients.

Addressing Bias and Quality

e.g. creating a de-biased data set to highlight the impact that bias is having on the real data

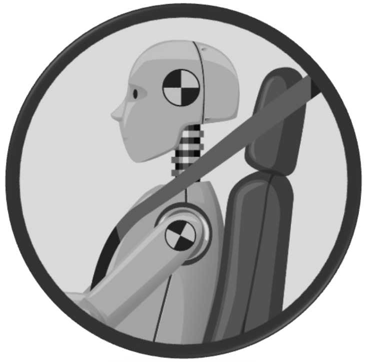
Fidelity - Simple is Often Better

***Range of Fidelity** (how similar the generated data is to the ground truth)*

Structural → Valid → Plausible → Multivariate Plausible → Multivariate Detailed → Replica

Source: [Office for National Statistics]

Range of Use-cases

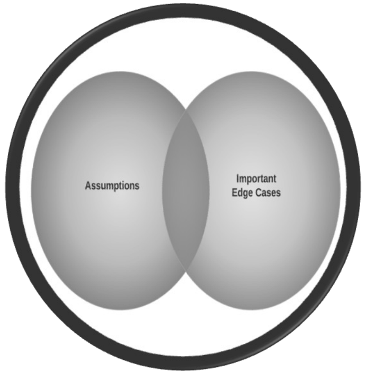


End-to-end software testing
e.g. Interoperability of architectures and systems to pass health data in FHIR formats

Tool Demonstration
e.g. New geospatial tool for showing impact of service planning on travel distance

Faster Innovation
e.g. Internal or external development of patient safety report classifier

Novel Linkage
e.g. generation of patient cancer pathways



Evaluation of Solutions
e.g. test clinical risk score prediction on rare patients.

Addressing Bias and Quality
e.g. creating a de-biased data set to highlight the impact that bias is having on the real data

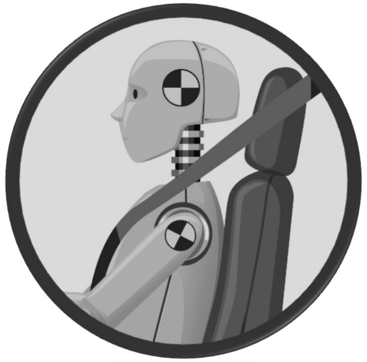
Fidelity - Simple is Often Better

Range of Fidelity *(how similar the generated data is to the ground truth)*

Structural → Valid → Plausible → Multivariate Plausible → Multivariate Detailed → Replica

Source: [Office for National Statistics]

Range of Use-cases



End-to-end software testing

e.g. Interoperability of architectures and systems to pass health data in FHIR formats

Tool Demonstration

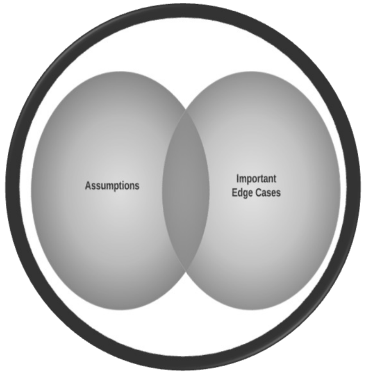
e.g. New geospatial tool for showing impact of service planning on travel distance

Faster Innovation

e.g. Internal or external development of patient safety report classifier

Novel Linkage

e.g. generation of patient cancer pathways



Evaluation of Solutions

e.g. test clinical risk score prediction on rare patients.

Addressing Bias and Quality

e.g. creating a de-biased data set to highlight the impact that bias is having on the real data

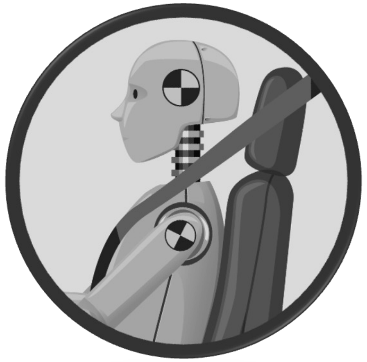
Fidelity - Simple is Often Better

Range of Fidelity *(how similar the generated data is to the ground truth)*

Structural → Valid → Plausible → Multivariate Plausible → Multivariate Detailed → Replica

Source: [Office for National Statistics]

Range of Use-cases



End-to-end software testing

e.g. Interoperability of architectures and systems to pass health data in FHIR formats

Tool Demonstration

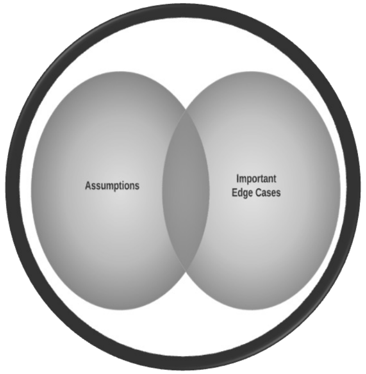
e.g. New geospatial tool for showing impact of service planning on travel distance

Faster Innovation

e.g. Internal or external development of patient safety report classifier

Novel Linkage

e.g. generation of patient cancer pathways



Evaluation of Solutions

e.g. test clinical risk score prediction on rare patients.

Addressing Bias and Quality

e.g. creating a de-biased data set to highlight the impact that bias is having on the real data

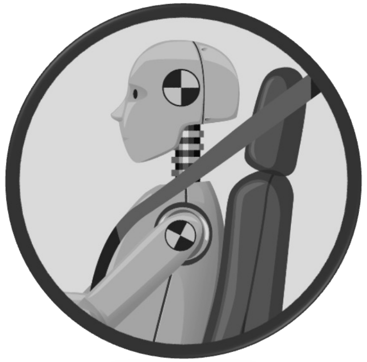
Fidelity - Simple is Often Better

***Range of Fidelity** (how similar the generated data is to the ground truth)*

Structural → Valid → Plausible → Multivariate Plausible → Multivariate Detailed → Replica

Source: [Office for National Statistics]

Range of Use-cases



End-to-end software testing

e.g. Interoperability of architectures and systems to pass health data in FHIR formats

Tool Demonstration

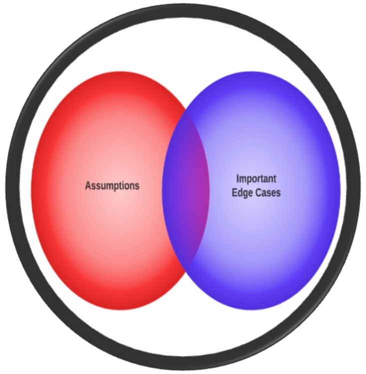
e.g. New geospatial tool for showing impact of service planning on travel distance

Faster Innovation

e.g. Internal or external development of patient safety report classifier

Novel Linkage

e.g. generation of patient cancer pathways



Evaluation of Solutions

e.g. test clinical risk score prediction on rare patients.

Addressing Bias and Quality

e.g. creating a de-biased data set to highlight the impact that bias is having on the real data

Fidelity - Simple is Often Better

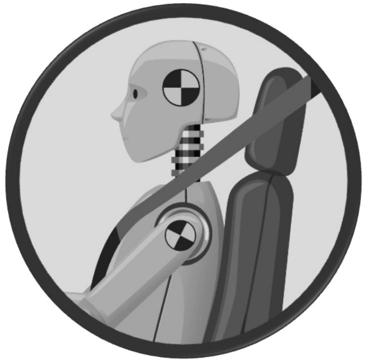
NHS

Range of Fidelity *(how similar the generated data is to the ground truth)*

Structural → Valid → Plausible → Multivariate Plausible → Multivariate Detailed → Replica

Source: [Office for National Statistics]

Range of Use-cases



End-to-end software testing

e.g. Interoperability of architectures and systems to pass health data in FHIR formats

Tool Demonstration

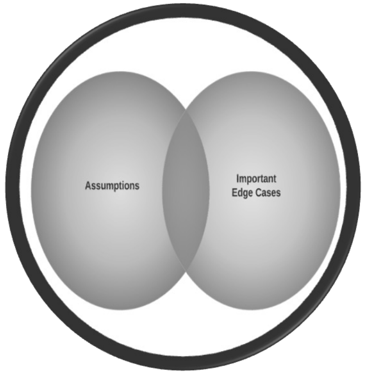
e.g. New geospatial tool for showing impact of service planning on travel distance

Faster Innovation

e.g. Internal or external development of patient safety report classifier

Novel Linkage

e.g. generation of patient cancer pathways



Evaluation of Solutions

e.g. test clinical risk score prediction on rare patients.

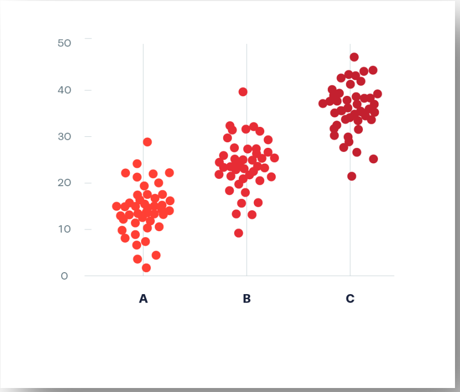
Addressing Bias and Quality

e.g. creating a de-biased data set to highlight the impact that bias is having on the real data

The Generation Landscape – By Technique

NHS

Adding Noise / Data Erosion



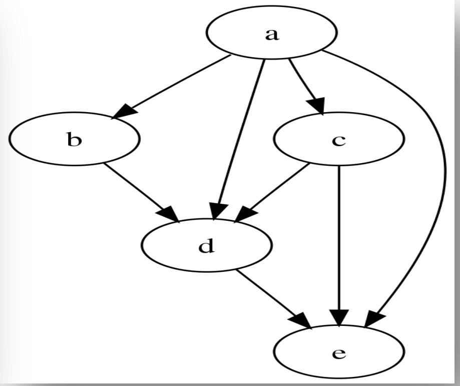
Adding Noise/Jitter

Suppression and relocation

Generalisation

[A Review of Anonymization for Healthcare Data]

Statistical/Probabilistic models

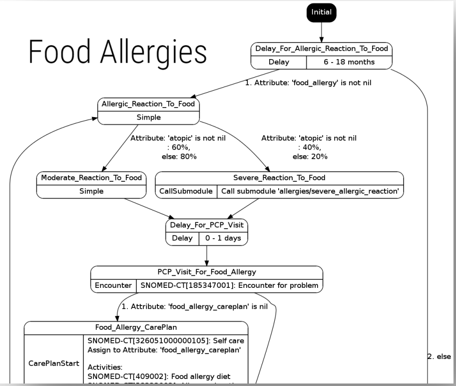


Sampling from independent marginals

Sampling from joint probabilities

[SynthPop]
[Faker]
[Simlacrum]
[CPRD Syntehtic Data Generation]

Simulations



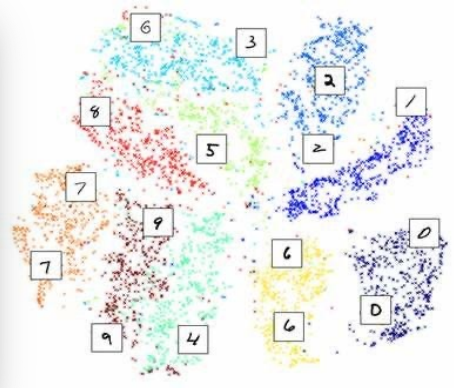
Digital Twins

Clinical Practice guidelines (CPGs)

Agent based simulations

[Synthea], [simhospital]

Perturbations of the manifold

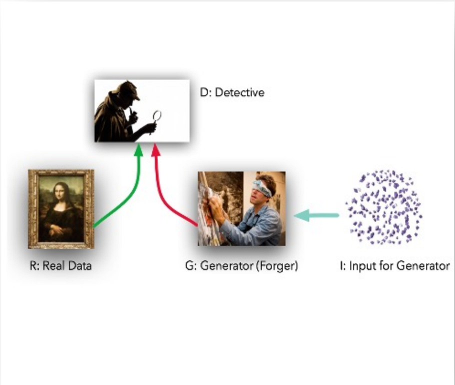


Synthetic Minority Over-Sampling Technique ([SMOTE])

Variational Autoencoders ([VAE])

[TVAE]
[Synthetic Patient Generation]

Iterative Comparisons

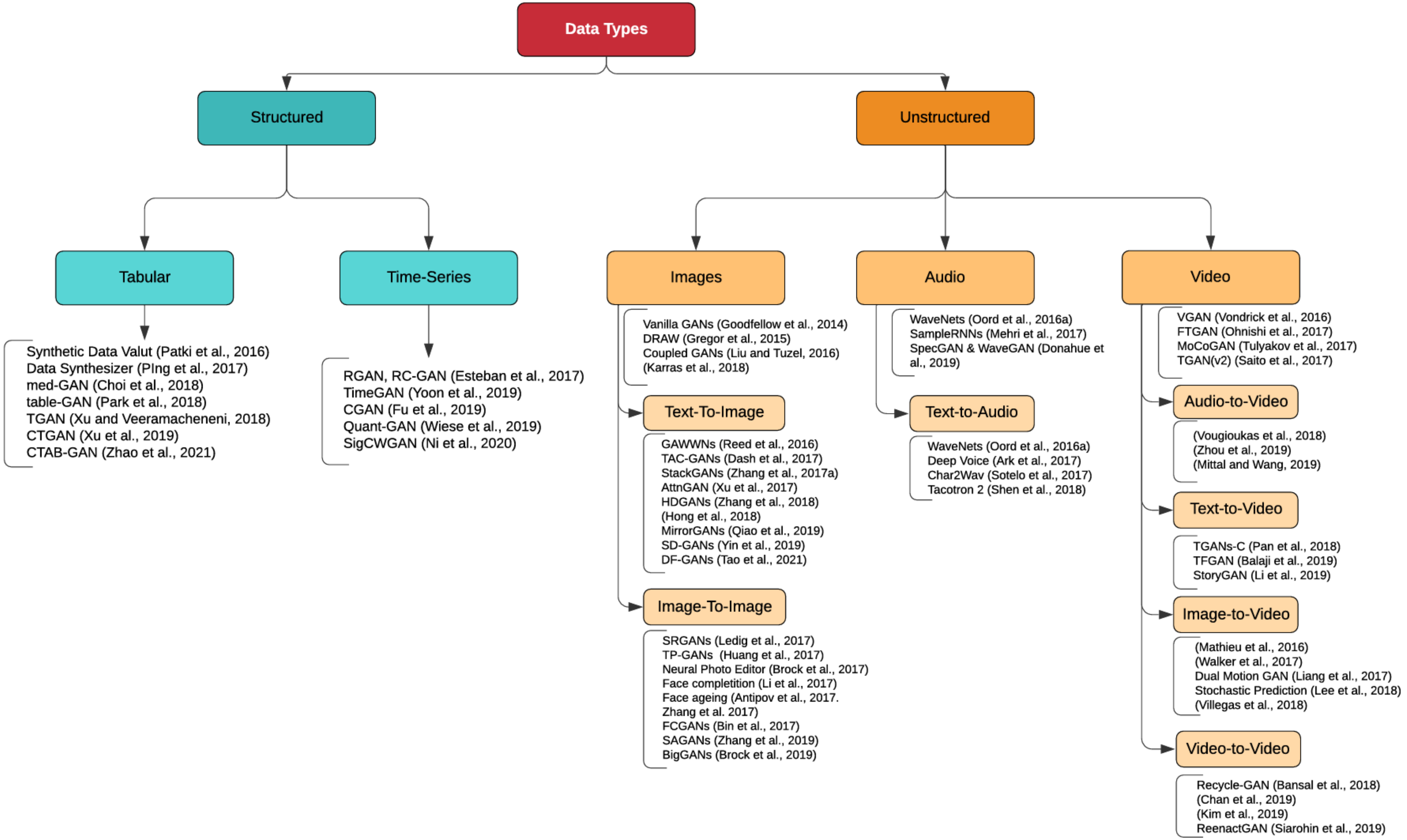


Generative Adversarial Networks ([GAN])

GPT-3-based architecture

[CT-GAN]
PATE-GAN, ADS-GAN, DECAF
[Van Der Schaar Lab]
[SynGatorTron]

The Generation Landscape – By Modality

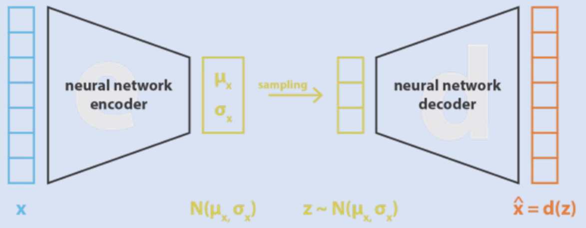


Source: [“Synthetic Data – What, Why and How?” - the Alan Turing Institute and The Royal Society.]

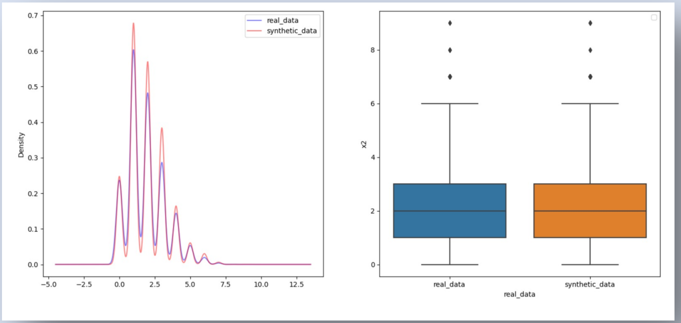
Our Approach - SynthVAE

Dom Danks and David Brind joined our team as a PhD Data Science Interns developing a variational autoencoder with differential privacy (SynthVAE).

Fidelity



$$\text{loss} = \|x - \hat{x}\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)]$$



The NHS AI Lab Skunkworks team published a case study of their use of our original tool in a user-friendly end-to-end process using QuantumBlack's Kedro.

Privacy

The privacy is the most difficult element to quantise and demonstrate confidence in.

We investigated adding differential privacy through it's impact on privacy metrics from SDMetrics.

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

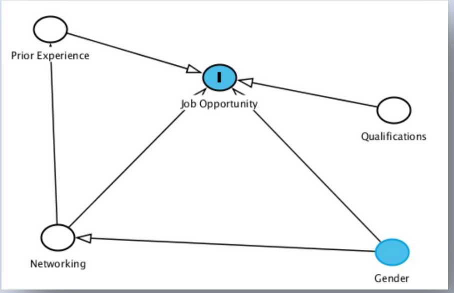
Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Figure 2: The DP-SGD algorithm. Credit: Abadi et al. (2016).

Fairness



Fairness can be considered before, during or after model application.

We have explored incorporating manually adjusting a learnt DAG representation through the ability to meet different fairness metrics

	Metric Type		
Name	Group	Subgroup	Individual
Demographic parity	✓		
Conditional statistical parity	✓		
Equalized odds	✓		
Equal opportunity	✓		
Treatment equality	✓		
Test fairness	✓		
Subgroup fairness		✓	
Fairness through unawareness			✓
Fairness through awareness			✓
Counterfactual fairness			✓

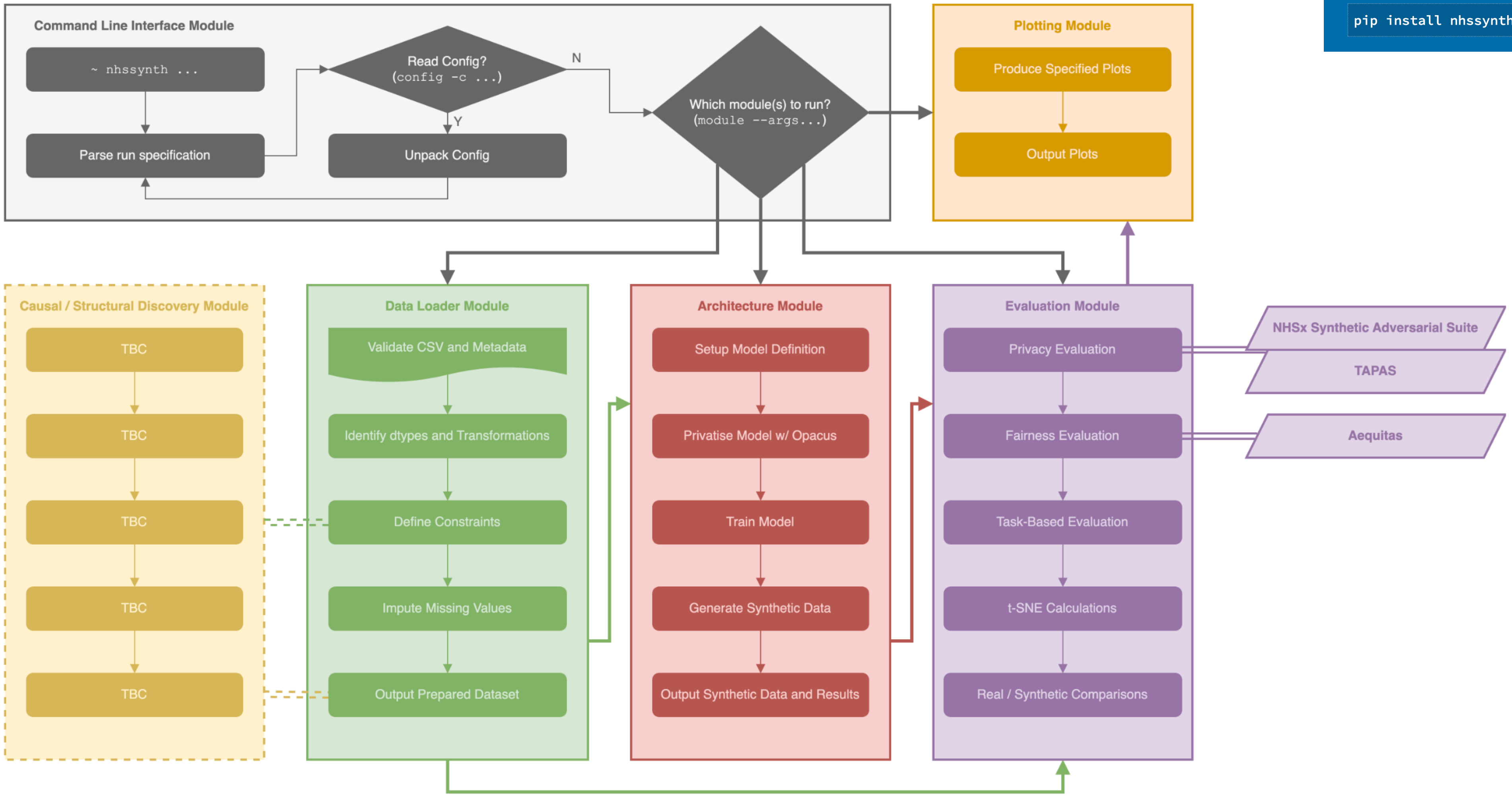
Table 1: Table 1 from review by Mehrabi et al. [9] showing the breakdown of fairness metrics and segregating them into similar groups

Our Approach – NHSSynth

In 2023, Harry Wilde has taken on the code with the task of bring everything together into a public facing python package called NHSSynth.

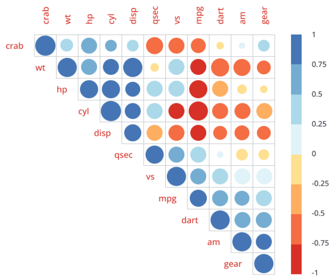
nhssynth 0.10.0

`pip install nhssynth`



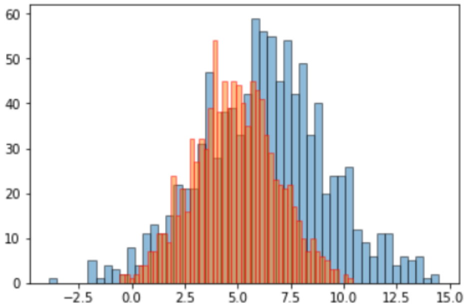
Evaluating the Data - Quality

Profile Comparisons
Are variable relationships maintained?



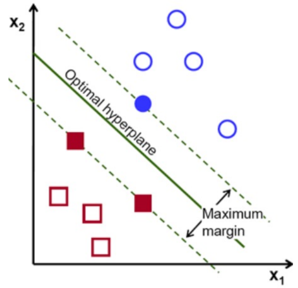
e.g. Pearson’s / **Similarity Metrics**

Distribution Comparisons
Are variable profiles consistent?



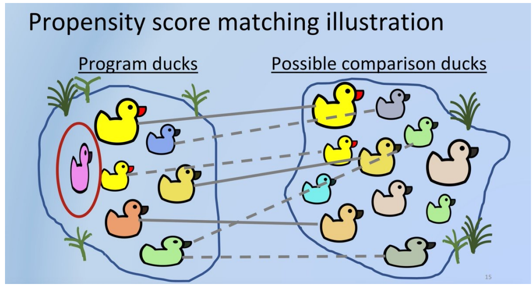
e.g. KS test, Chi-Squared

Detection Metrics
Can a classifier differentiate between real and synthetic data?



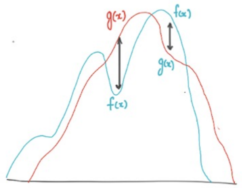
e.g. **logistic**, SVM

Variance Metrics
Is the range of variation of data points consistent?



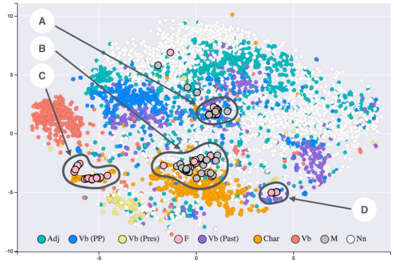
Voas-Williamson or propensity scores

Aggregate difference Metrics
How much work is required to align data points to expectations?



e.g. **KL divergence, Gower distance** or Wasserstein metric

Off-Manifold and Latent Space Checks
Are there unexpected features in the latent space?



e.g. PCA, t-SNE

End use case comparisons and task performance

Evaluating the Data - Privacy

"Synthetic Data – Anonymisation Groundhog Day" by Stadler, Oprisanu, and Troncoso proposed using shadow modelling to apply generalised membership inference and attribute inference attacks on any synthetic data model.

Tapas: a Toolbox for Adversarial Privacy Auditing of Synthetic Data

Project is under active development. A Python library for evaluating the privacy of synthetic data from an adversarial perspective.

tapas
latest
Search docs

- tapas
- Installation
- Quick Start
- Data format
- Modelling Threats
- Library of Attacks
  - Summary
  - Trainable-threshold attacks
  - Closest-distance Attacks
  - Shadow Modelling Attacks
  - Inference-on-Synthetic Attacks
  - References
- Implementing Attacks
- Evaluation

/ Library of Attacks — Edit on GitHub

Library of Attacks

Adversarial approaches for privacy require auditors to run a large range of diverse attacks, in order to test for as many potential vulnerabilities as possible. For this reason, `TAPAS` implements a large range of attacks. While some of these attacks are technically involved, many are straightforward and are intended mostly as safety checks. We here present the different attacks implemented in `TAPAS`, grouped by theme. For each attack, we specify the additional parameters it requires, and the attack models (see *Modelling Threats <modelling-threats.rst>*) it applies to. `TAPAS` attacks inherit from the `tapas.attacks.Attack` abstract class (see *Implementing Attacks <implementing-attacks.rst>* for details). Note that the constructor of *all* the attacks described below allows for an optional `label` parameters, which we exclude from the descriptions for the sake of concision.

Notations: we denote by $D^{(r)}$ the real, private dataset, and $D^{(s)}$ the synthetic dataset obtained with the generation method $\mathcal{G}$. For targeted attacks, the attacker aims to learn information about a record x, either membership ($x \in D^{(r)}$) or the value of a sensitive attribute s (v s.t. $x|v \in D^{(r)}$).

Summary

Class	Threat Model	Parameters	Decision uses
`ClosestDistanceMIA`	MIA	`distance`, `criterion`	Distance of close
`ClosestDistanceAIA`	AIA	`distance`, `criterion`	For each value, d
`LocalNeighbourhoodAttack`	MIA/AIA	`distance`, `radius`, `criterion`	Sphere of given r
`ShadowModellingAttack`	MIA/AIA	`SetClassifier`	Train a set classif
`GroundhogAttack`	MIA/AIA	`features`, `classifier`	Shadow modellin
`ProbabilityEstimationAttack`	MIA	`estimator`, `criterion`	Density estimato
`SyntheticPredictorAttack`	AIA	`estimator`, `criterion`	Classifier fit on sy

License MIT | Python 3.8.5

NHSX Synthetic Adversarial Suite

NHSX Synthetic Adversarial Suite was selected as a project in November 2021.

Intended Use

This suite of tools is designed to cover attack scenarios where an attacker may have different levels of access to a synthetic data model or synthetic dataset and wishes to determine if any original data can be asertained.

This suite of tools is a python package designated `synthetic_adversarial_suite`.

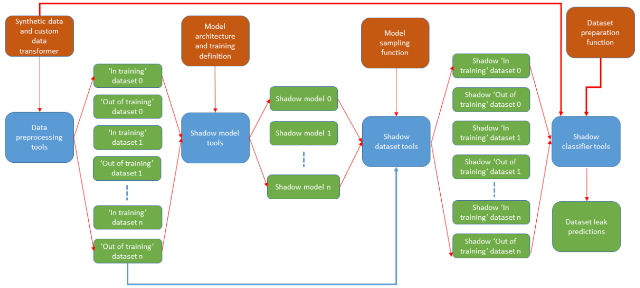
README.md

Data Pipeline

The data flow for each scenario is shown in the diagrams below. The elements are catagorised as follows:

- Orange - Attacker provided
- Blue - Pipeline element
- Green - Generated artefact

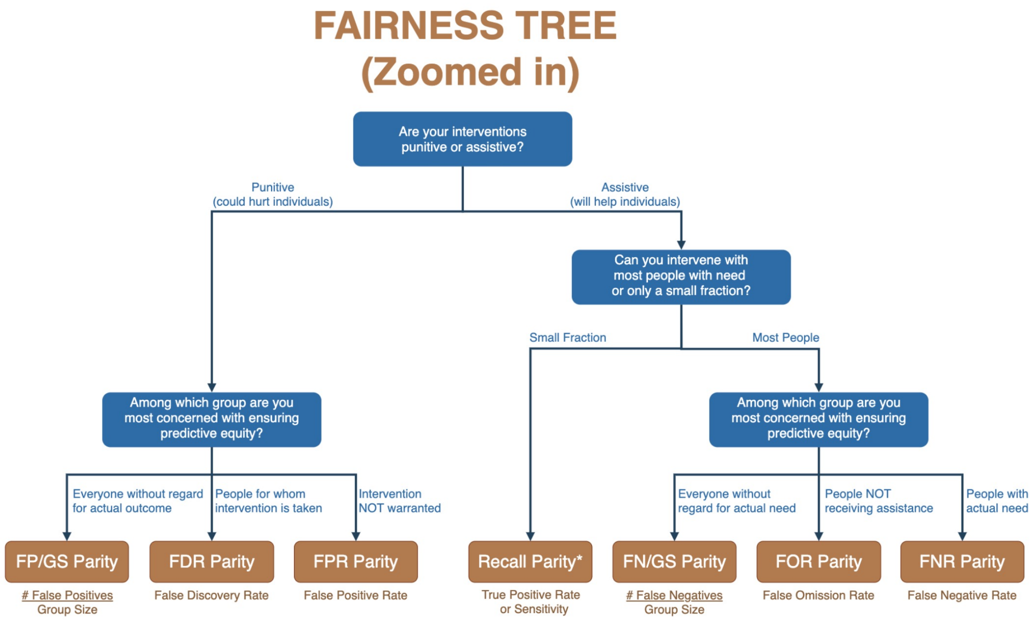
Scenario One

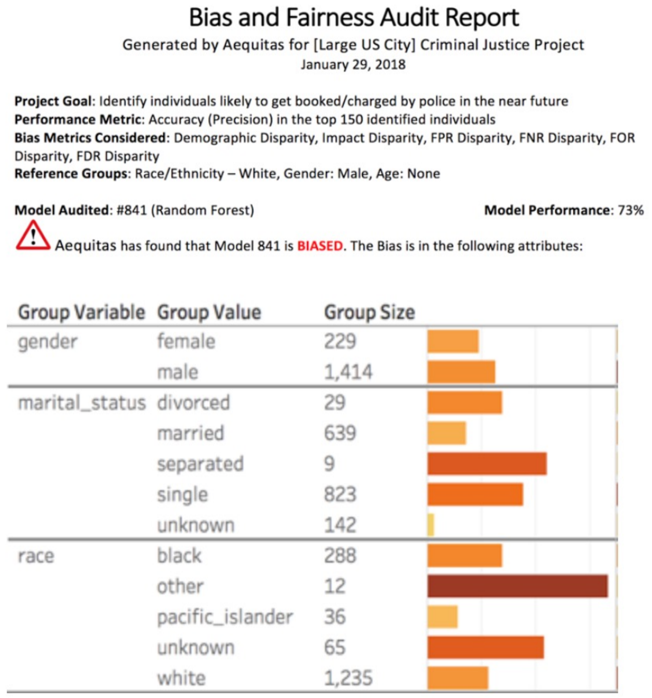


Evaluating the Data - Fairness

It is desirable to be able to “de-bias” originally unfair training data or at least understand if the synthetic process has changed bias in the data.

Aequitas: an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive risk-assessment tools. Developed by the Centre for Data Science and Public Policy at University of Chicago.





Bias and Fairness Audit Report

Generated by Aequitas for [Large US City] Criminal Justice Project

January 29, 2018

Project Goal: Identify individuals likely to get booked/charged by police in the near future
Performance Metric: Accuracy (Precision) in the top 150 identified individuals
Bias Metrics Considered: Demographic Disparity, Impact Disparity, FPR Disparity, FNR Disparity, FOR Disparity, FDR Disparity
Reference Groups: Race/Ethnicity – White, Gender: Male, Age: None

Model Audited: #841 (Random Forest) **Model Performance**: 73%

Aequitas has found that Model 841 is BIASED. The Bias is in the following attributes:

Group Variable	Group Value	Group Size
gender	female	229
	male	1,414
marital_status	divorced	29
	married	639
	separated	9
	single	823
	unknown	142
race	black	288
	other	12
	pacific_islander	36
	unknown	65
	white	1,235

Success / Failures

Success

+ Data generation using Bayesian Network, VAE, GAN & Transformer approaches

+ Development of end-to-end framework allowing experiments beyond single generation

+ IG Buy-In for data erosion approach

+ Partial success on data transformers for categorical, date and numerical variables

Further examples required in this area

Failures (yet to succeed)

- Clarity around privacy threshold for different data types and sensitivities (characterising metrics as confidence levels)

- Clear privacy metrics

- Longitudinal Data generation

- Multi-table data generation

- Multimodal Data generation

- Implementing causal modelling into the generation algorithm

Further research required in this area

Main Take-Away

NHS

The balance for utility/privacy is key for the actual generation of synthetic data but two components which are fundamental to a successful project are:

Explainability

A project which is clear in how it has handled the data, where the data is limited and what level of risk is present is more useful that high quality or absolute privacy.

Adoption

Putting user need and route to deployment above technical opportunity

Final Point:

For healthcare a golden thread needs to be discussed throughout the work that allows the end decision to be set in context of the level or quality/risk created by the synthetic generation