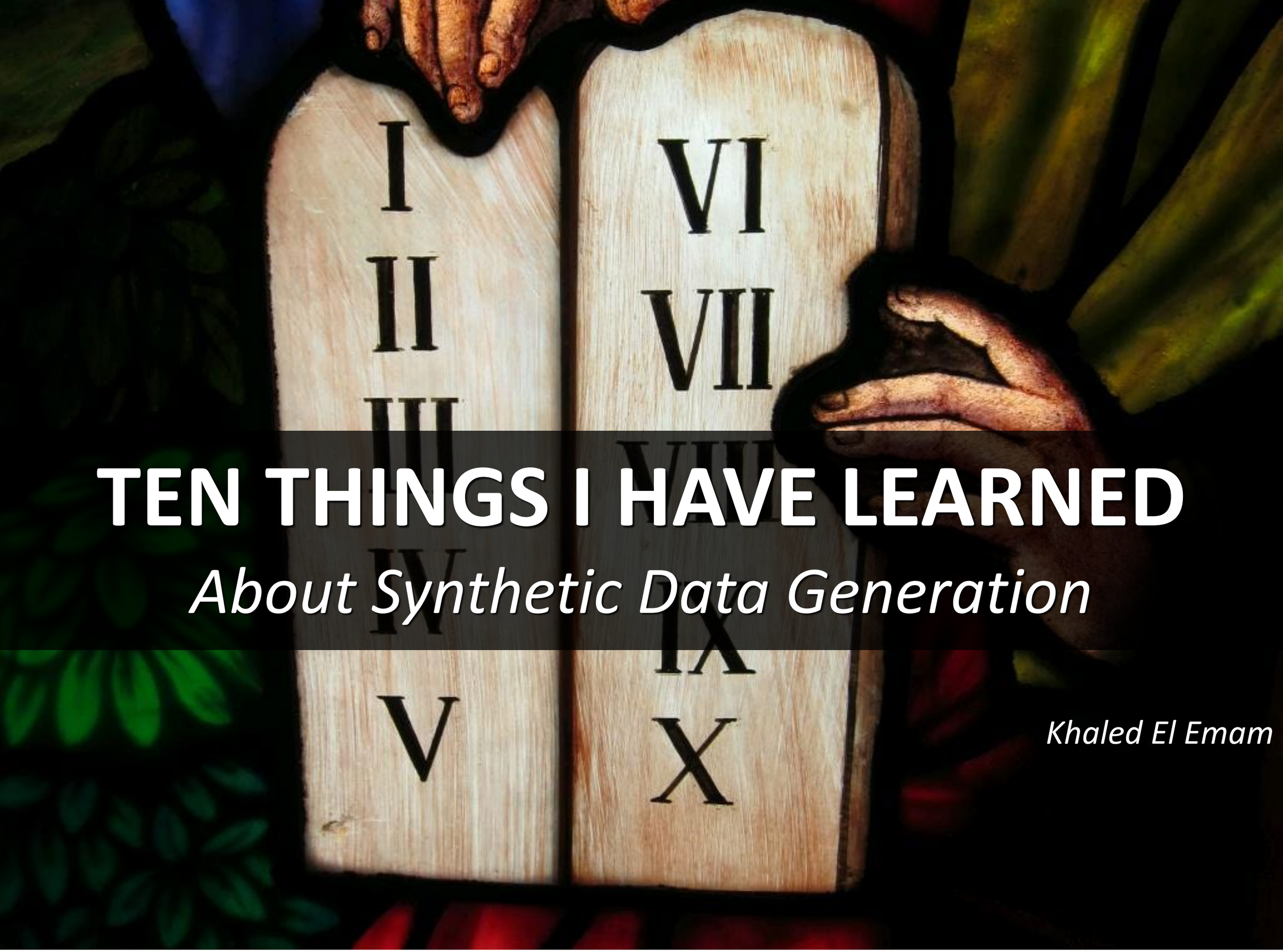


TEN THINGS I HAVE LEARNED

About Synthetic Data Generation

Khaled El Emam

Scope of Presentation

- Health data
- Tabular data
- Structured data
- Out-of-the-box generative models
- Real world perspective (i.e., implementation focus)

Original Reports | Artificial Intelligence

Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets

Samer El Kababji, PhD, MEng, MSc[1]; Nicholas Mitsakakis, PhD, MSc[1]; Xi Fang, MSc[2]; Ana-Alicia Beltran-Bless, MD[3,4]; Greg Pond, PhD[5]; Lisa Vandermeer, MSc[3]; Dhenuka Radhakrishnan, MD, MSc[1,6]; Lucy Mosquera, MSc[1,2]; Alexander Paterson, MD[7]; Lois Shepherd, MD[8]; Bingshu Chen, PhD[8]; William E. Barlow, PhD[9]; Julie Gralow, MD[10]; Marie-France Savard, MD[3,4]; Mark Clemons, MD, MB[3,4]; and Khaled El Emam, PhD, BEng[1,2,11]

DOI https://doi.org/10.1200/CCI.23.00116

ABSTRACT

PURPOSE There is strong interest from patients, researchers, the pharmaceutical industry, medical journal editors, funders of research, and regulators in sharing clinical trial data for secondary analysis. However, data access remains a challenge because of concerns about patient privacy. It has been argued that synthetic data generation (SDG) is an effective way to address these privacy concerns. There is a dearth of evidence supporting this on oncology clinical trial data sets, and on the utility of privacy-preserving synthetic data. The objective of the proposed study is to validate the utility and privacy risks of synthetic clinical trial data sets across multiple SDG techniques.

METHODS We synthesized data sets from eight breast cancer clinical trial data sets using

ACCOMPANYING CONTENT

Data Supplement

Accepted September 19, 2023
Published November 27, 2023

JCO Clin Cancer Inform
7:e2300116

S. El Kabaji, N. Mitsakakis, X. Fang, A.Beltran-Bless, G. Pond, L. Vandermeer, D. Radhakrishnan, L. Mosquera, A. Paterson, L. Shepherd, B. Chen, W. Barlow, J. Gralow, M-F Savard, M. Clemons, K. El Emam: “Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Datasets," Journal of Clinical Oncology: Clinical Cancer Informatics, 2023.

Common Definitions of Utility

Fidelity

Generic utility

Show how similar synthetic data is to the real data it was generated from without referencing a specific analysis

Replicability

Workload aware utility

Illustrate how well synthetic data can be used as a drop-in replacement or proxy for real data for a specific analysis

Expert discrimination

A clinician would manually examine multiple records and classify each one as real or synthetic

Fidelity

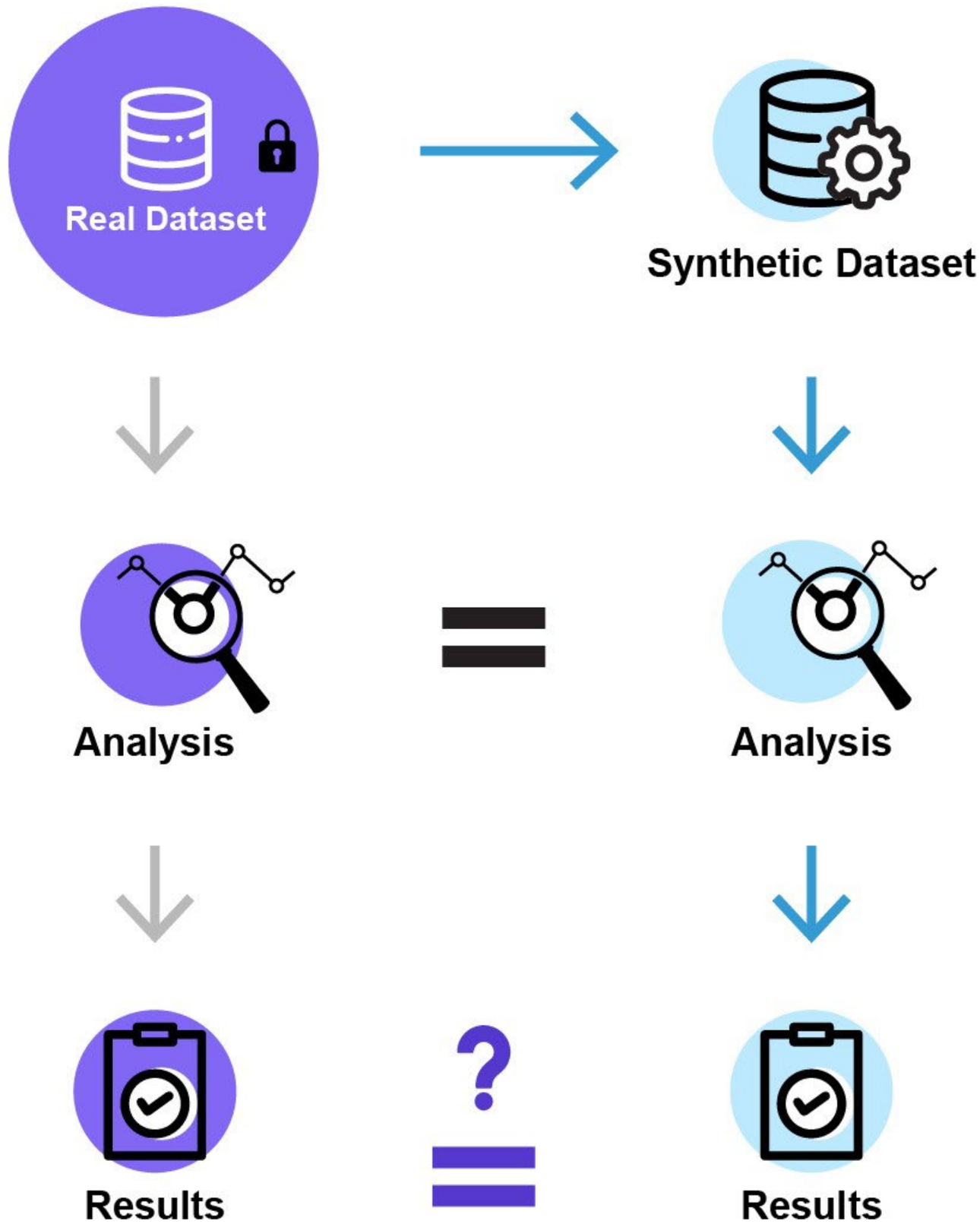
Generic utility

Show how similar synthetic data is to the real data it was generated from without referencing a specific analysis

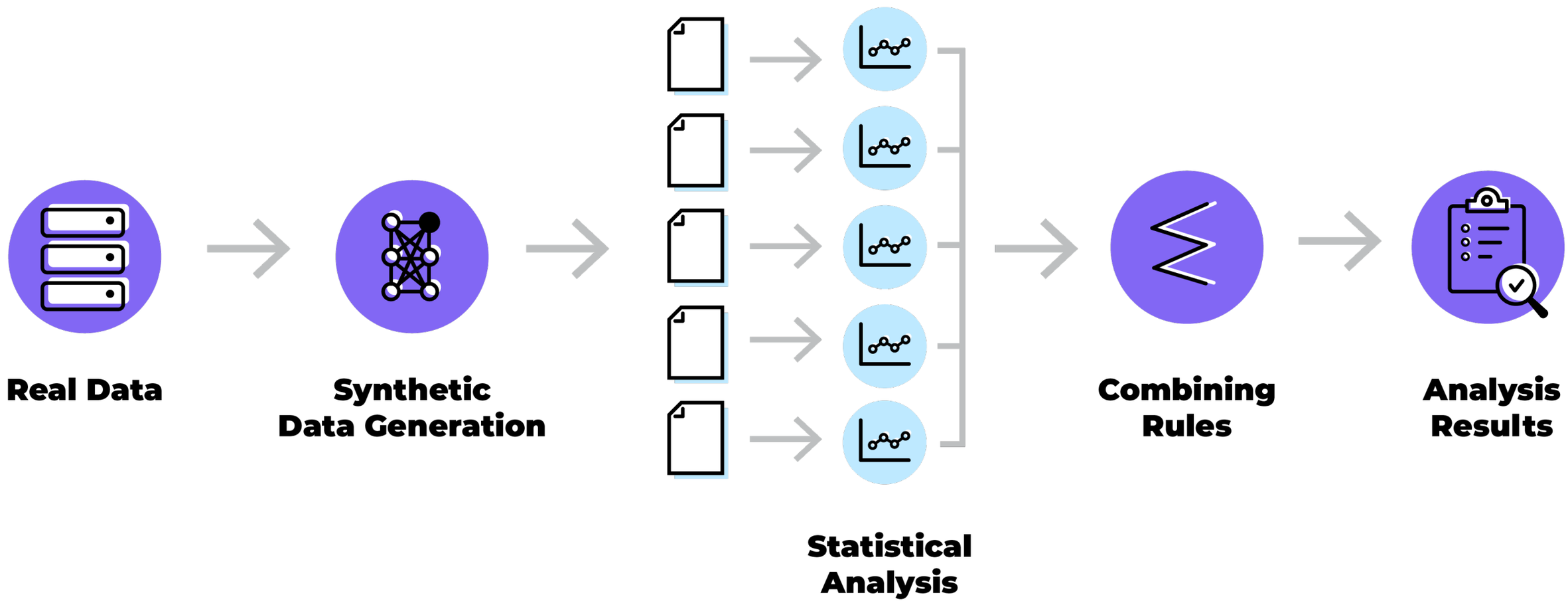
Replicability

Workload aware utility

Illustrate how well synthetic data can be used as a drop-in replacement or proxy for real data for a specific analysis


Real Dataset
Synthetic Dataset
Analysis
Analysis
Results
Results

Model Averaging



Data Set	Sample Size	SEQ			GAN			VAE		
		Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap
REaCT-HER2+	48	1	1	0.77	1	1	0.88	1	1	0.94
REaCT-G/G2	401	1	1	0.91	[a]	[a]	[a]	1	1	0.67
REaCT-ILIAD	218	1	1	0.99	1	1	0.85	1	0	0.74
REaCT-ZOL	211	1	[b]	0.98	1	[b]	0.88	0	[b]	0.61
REaCT-BTA	230	1	1	0.85	1	0	0.68	1	0	0.72
CCTG MA27	7,576	1	1	0.90	1	1	0.62	1	1	0.82
SWOG 0307	6,097	1	1	0.93	1	0	0.50	1	1	0.95
NSABP B34	3,323	1	1	0.93	1	1	0.83	1	1	0.61

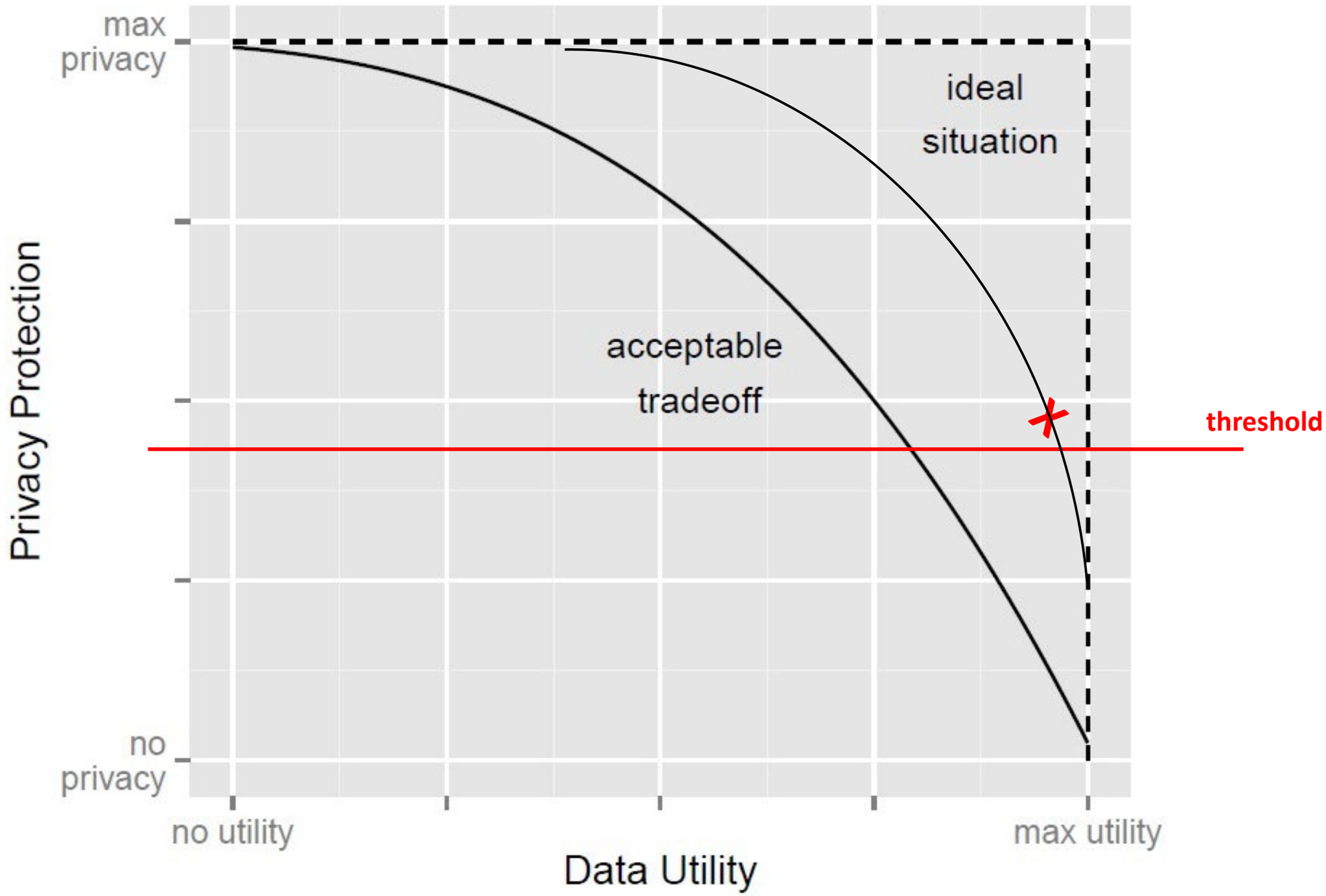
Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

[a]Training the generative model failed.

[b]The analysis is descriptive and hence decision agreement does not apply.

Some Observations

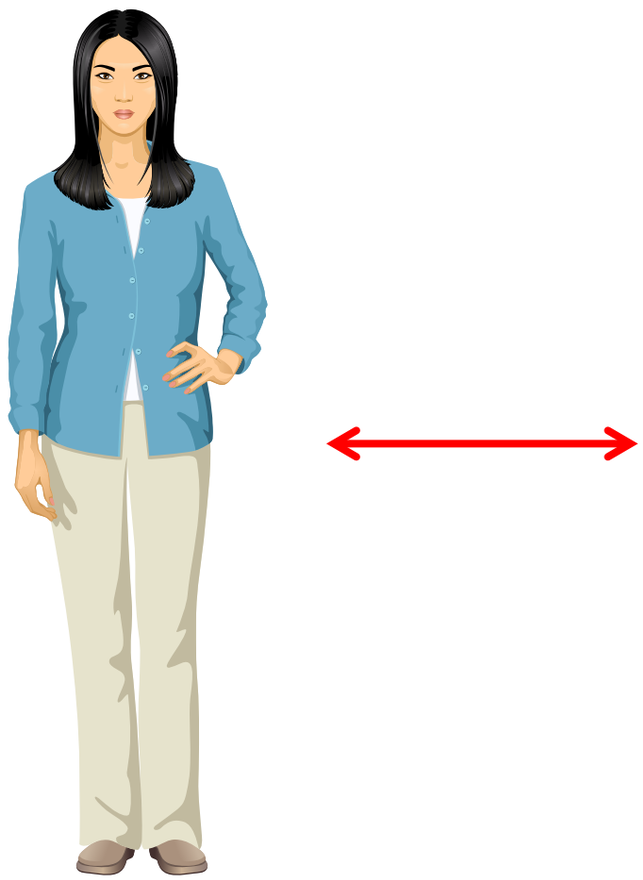
- Not all generative models are created equal
- Model averaging is important
- Data pre-processing is important - it is not all about the training


max
privacy
no
privacy
Privacy Protection
ideal
situation
acceptable
tradeoff
threshold
no utility
max utility
Data Utility

Similarity Metrics

- Ignore re-identification risk of original dataset
- Ignore information gain
- Ignore adversary knowledge

Attribution Disclosure



Quasi-identifiers

New Information

Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	**1989**	**65862-403**
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

Attribution Disclosure

- Contingent on re-identification risk of real dataset
- Considers similarity on quasi-identifiers
- Accounts for information gain (outliers have more information gain than the average)

Attribution Disclosure

	SEQ		GAN		VAE	
Data Set	Risk Value	Risk	Risk Value	Risk	Risk Value	Risk
REaCT-HER2+	2.56E-04	LO	2.35E-04	LO	2.35E-04	LO
REaCT-G/G2	1.10E-04	LO	1.10E-04	LO	1.10E-04	LO
REaCT-ILIAD	2.90E-05	LO	2.90E-05	LO	2.90E-05	LO
REaCT-ZOL	1.58E-03	LO	1.41E-03	LO	1.10E-03	LO
REaCT-BTA	6.48E-04	LO	6.43E-04	LO	6.43E-04	LO
CCTG MA27	1.37E-03	LO	1.37E-03	LO	1.38E-03	LO
SWOG 0307	2.09E-03	LO	2.17E-03	LO	2.02E-03	LO
NSABP B34	2.25E-02	LO	2.02E-02	LO	1.83E-02	LO

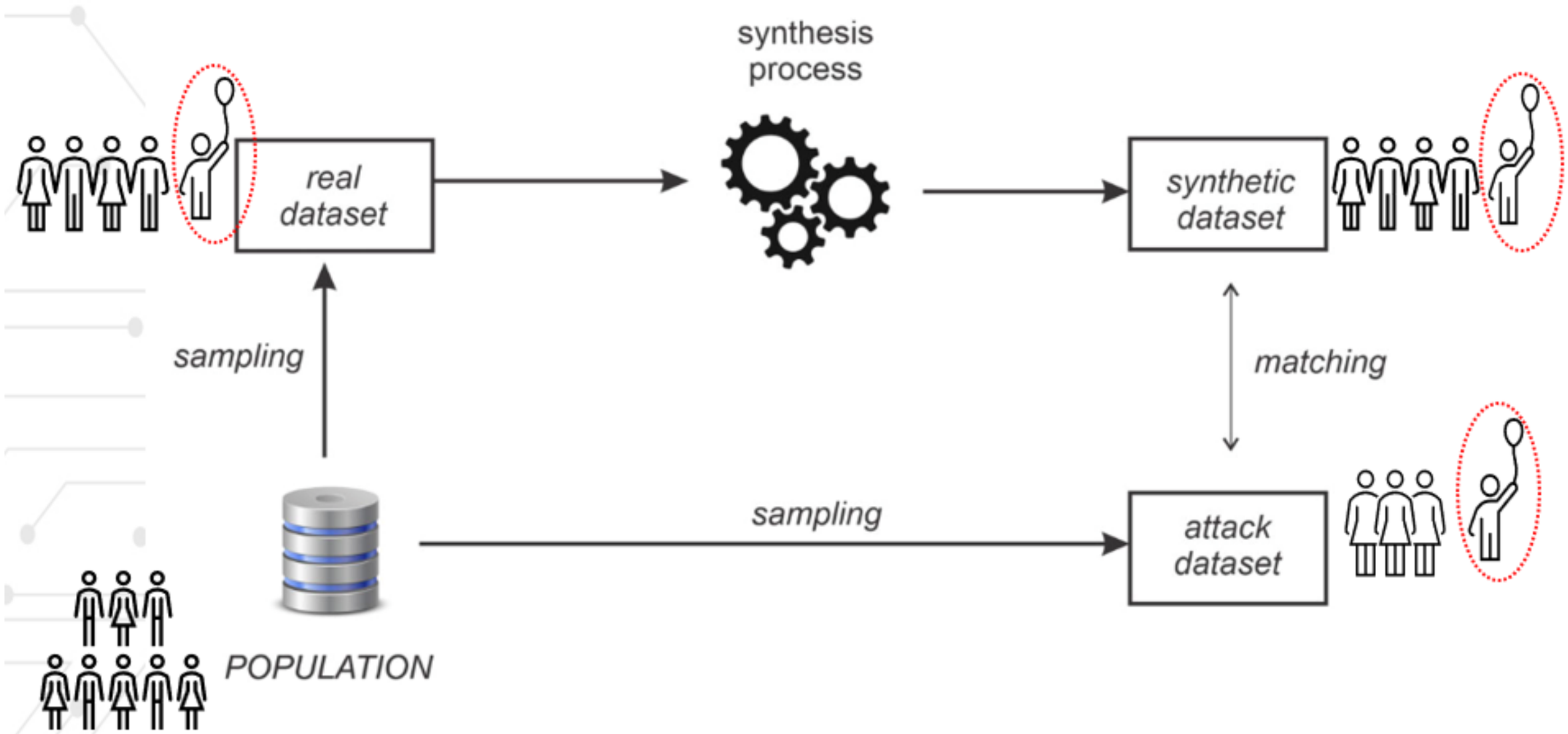
Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; LO, low risk; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

Commonly used threshold of 0.09 for disclosure risk

Attribute Disclosure

- Defined as making inferences from models - if an analyst is able to train an accurate prognostic model from the data then that is an attribute disclosure
- That is the essence of data analysis
- Sensitivity of inferences should be dealt with through an ethics review rather

Membership Disclosure



Membership Disclosure

- Some attacks assume a large reference dataset is available
- Should we focus only on quasi-identifiers ?
- Sampling fraction of the real data is an important factor

Membership Disclosure

		SEQ		GAN		VAE	
Data Set	n/N (sampling fraction)	F_rel	Risk	F_rel	Risk	F_rel	Risk
REaCT-HER2+	0.021	0.15	LO	0.07	LO	0.09	LO
REaCT-G/G2	0.062	0.06	LO	0.06	LO	0.06	LO
REaCT-ILIAD	0.004	0.02	LO	0.02	LO	0.02	LO
REaCT-ZOL	0.023	0.02	LO	0.02	LO	0.02	LO
REaCT-BTA	0.207	0.13	LO	0.18	LO	0.18	LO
CCTG MA27	0.573	0.31	HI	0.32	HI	0.34	HI
SWOG 0307	0.147	0.13	LO	0.13	LO	0.13	LO
NSABP B34	0.158	−0.02	LO	−0.15	LO	−0.19	LO

NOTE. The threshold for the sampling fraction is 0.33, and 0.2 for the relative F1 score (F_rel).
Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; HI, high risk; LO, low risk; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

Commonly used threshold of 0.2 for membership disclosure

Regulation

- Is it possible to regulate at the same pace as technology development ?
- Do regulators have the full expertise to cover deep technical topics ?
- Should regulation refer to current best practices ?

Zero Risk

Standards

- Generic frameworks are not very useful and may actually increase uncertainty
- We have well defined utility and privacy metrics that can be used to benchmark
- More is not better

SOCIAL LICENSE

Data Augmentation

- Augmentation for machine learning models
- Simulating patients to deal with accrual problems or attrition
- Simulating patients by design to enable smaller data collection
- Simulating under-represented patients

Scalability

- Training generative models on large datasets is a challenge (many observations and many variables) - the compute requirements can be cost and time prohibitive
- Inference / synthesis has to be very efficient to enable data generation on demand

TEN THINGS

1. Replicability is the most important utility measure
2. There are metrics to measure replicability
3. Final analysts from synthetic data must be averaged across models
4. There is variation across generative model performance
5. Sequential synthesis produces competitive results on tabular data
6. Stop using similarity metrics for privacy assessment
7. Use ethics reviews to manage attribute disclosure
8. Simulation of patients is the next major application
9. There is a need for operational standards and guidelines supported by regulators
10. Scalability of generative model training is a becoming a practical challenge