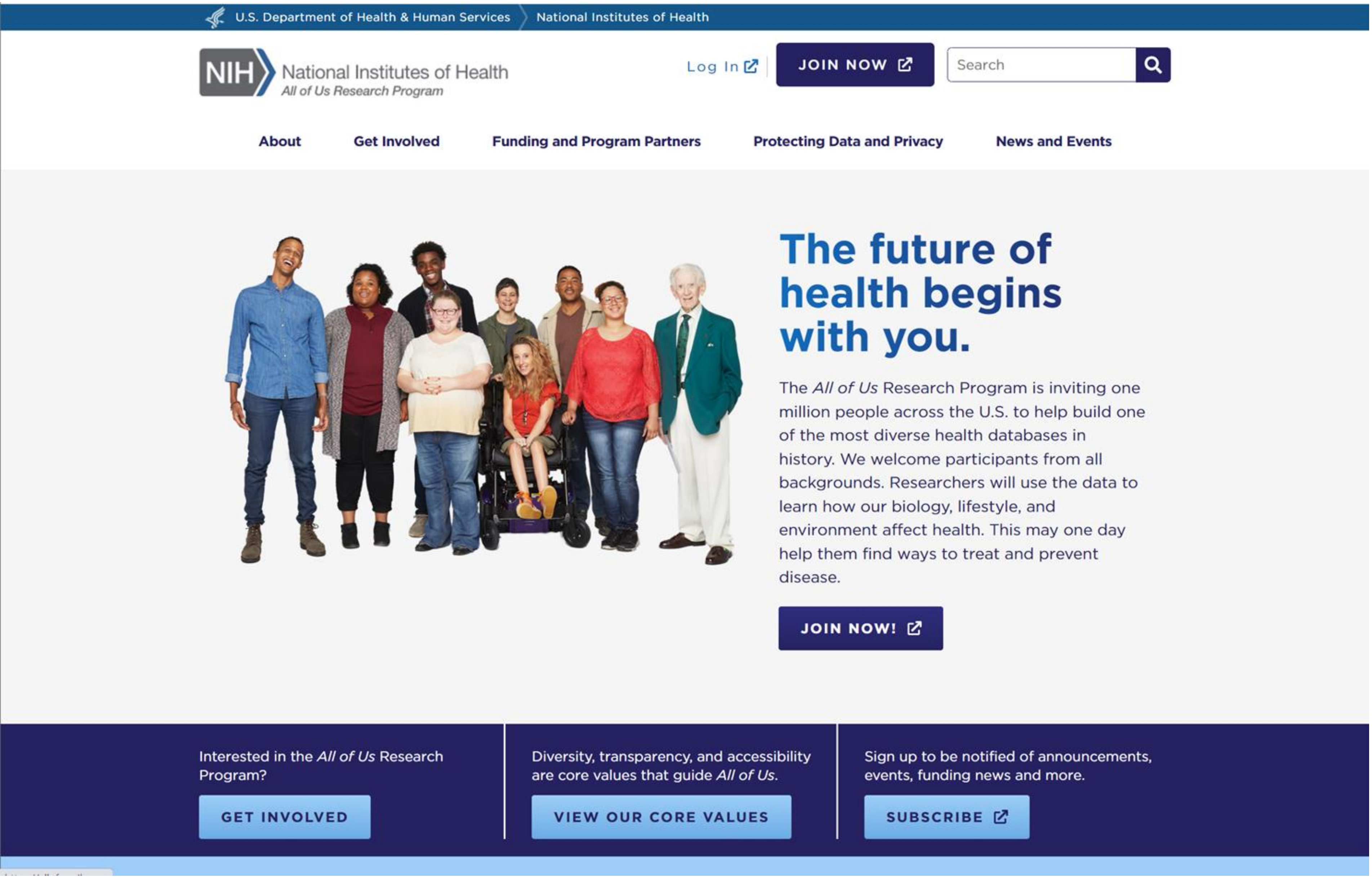
Achieving Better Privacy and Better Health Through Machine Learning

Bradley Malin, Ph.D.
Vice Chair for Research Affairs, Department of Biomedical Informatics
Director, Health Data Science Center
Vanderbilt University Medical Center

Accenture Professor of Biomedical Informatics, Biostatistics, and Computer Science
Vanderbilt University

9 October 2023


U.S. Department of Health & Human Services
National Institutes of Health
NIH National Institutes of Health
All of Us Research Program
Log In
JOIN NOW
Search
About
Get Involved
Funding and Program Partners
Protecting Data and Privacy
News and Events

The future of health begins with you.

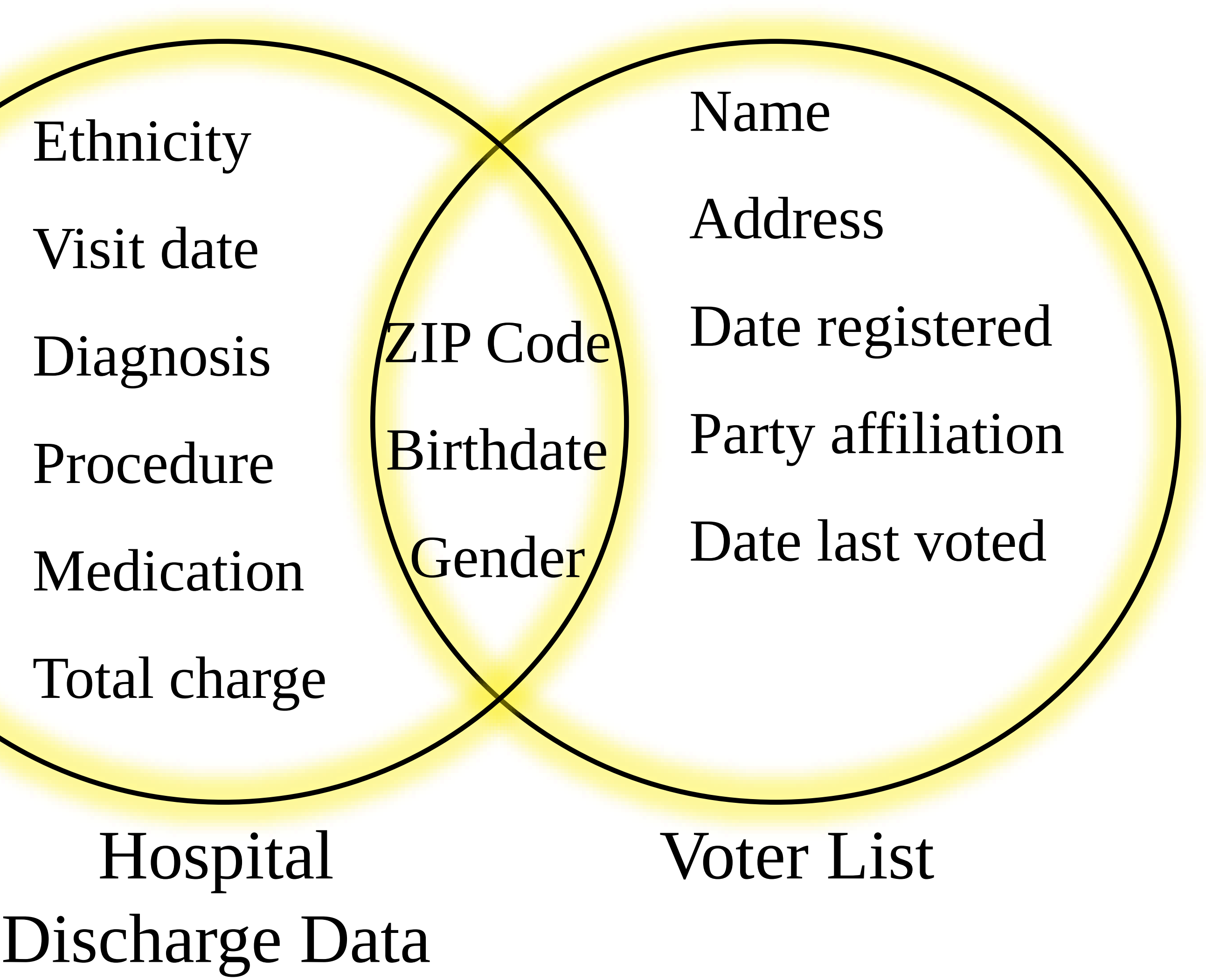
The *All of Us* Research Program is inviting one million people across the U.S. to help build one of the most diverse health databases in history. We welcome participants from all backgrounds. Researchers will use the data to learn how our biology, lifestyle, and environment affect health. This may one day help them find ways to treat and prevent disease.

Interested in the *All of Us* Research Program?

Diversity, transparency, and accessibility are core values that guide *All of Us*.

Sign up to be notified of announcements, events, funding news and more.

What We Fear



Ethnicity

Visit date

Diagnosis

Procedure

Medication

Total charge

ZIP Code

Birthdate

Gender

Name

Address

Date registered

Party affiliation

Date last voted

Hospital Discharge Data

Voter List

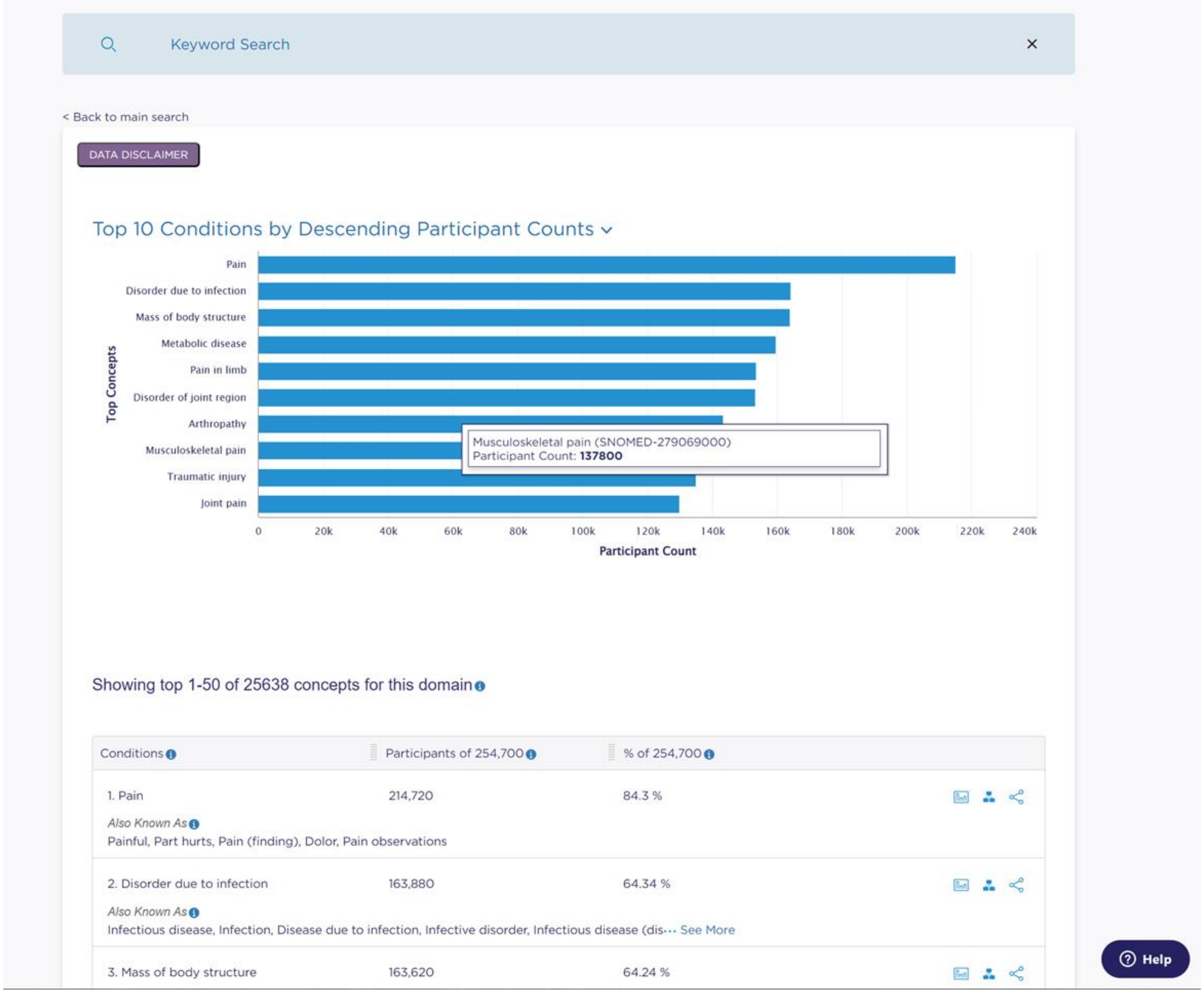
Sweeney L. Journal of Law, Medicine, and Ethics. 1997

Tiered Levels of Access

- Public
  - Can be accessed without logging in
  - Summary statistics only

Public

https://databrowser.researchallofus.org/ehr/conditions



Top 10 Conditions by Descending Participant Counts

Musculoskeletal pain (SNOMED-279069000)
Participant Count: **137800**

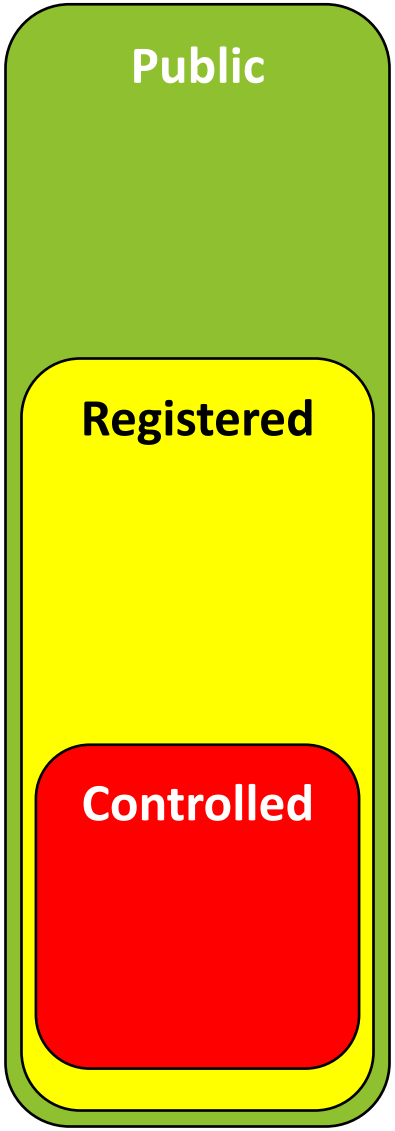
Showing top 1-50 of 25638 concepts for this domain

Conditions	Participants of 254,700	% of 254,700
1. Pain *Also Known As* Painful, Part hurts, Pain (finding), Dolor, Pain observations	214,720	84.3 %
2. Disorder due to infection *Also Known As* Infectious disease, Infection, Disease due to infection, Infective disorder, Infectious disease (dis··· See More	163,880	64.34 %
3. Mass of body structure	163,620	64.24 %

Tiered Levels of Access

- Public
  - Can be accessed without logging in
  - Summary statistics only

- Sandbox Environments
  - Registered
    - Available to anyone within a trusted organization who proves their identity and enters into use agreement … expanding to citizen scientists
    - Individual-level data with very low risk of re-identification
  - Controlled
    - Available to trusted investigators
    - Genomic data

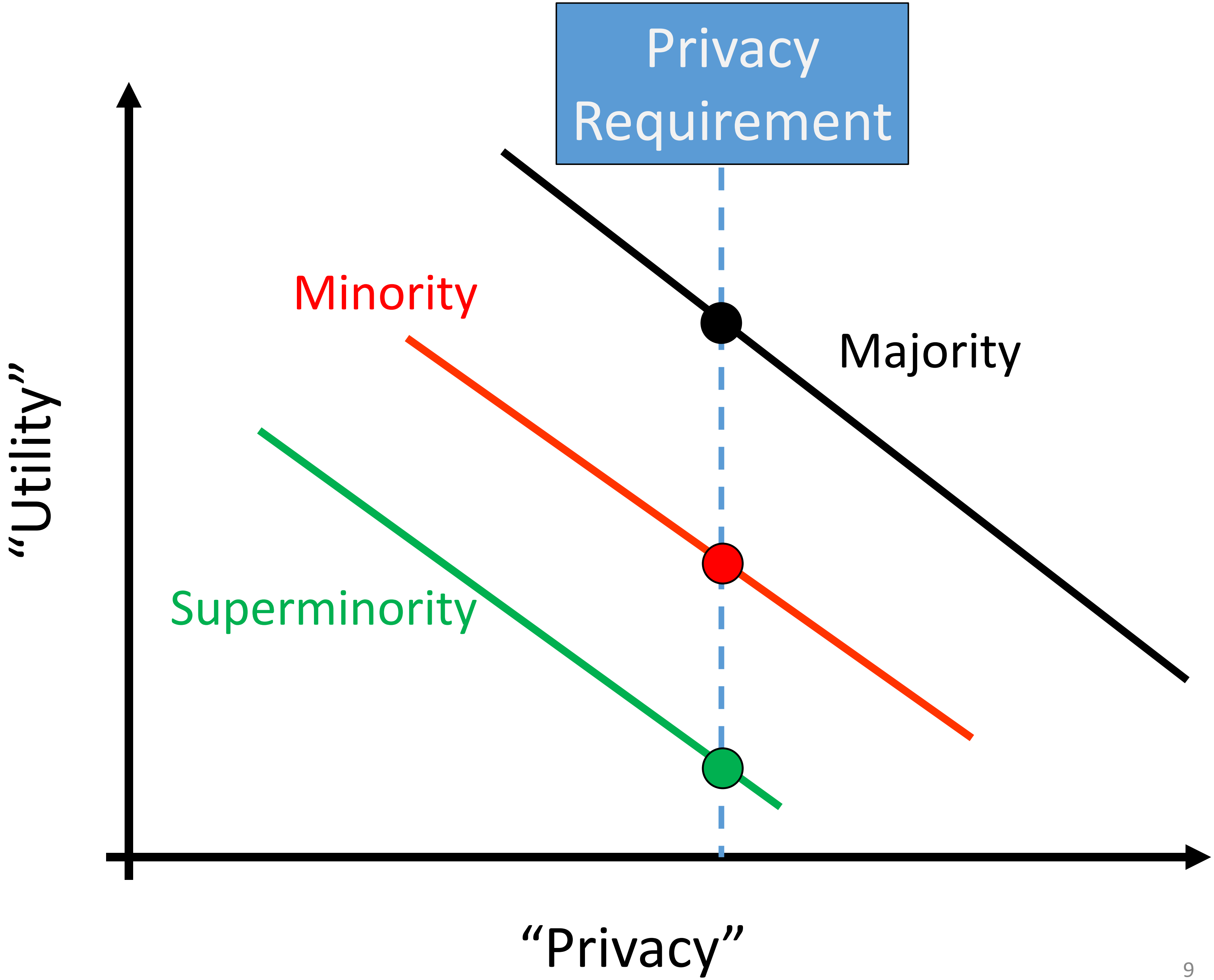


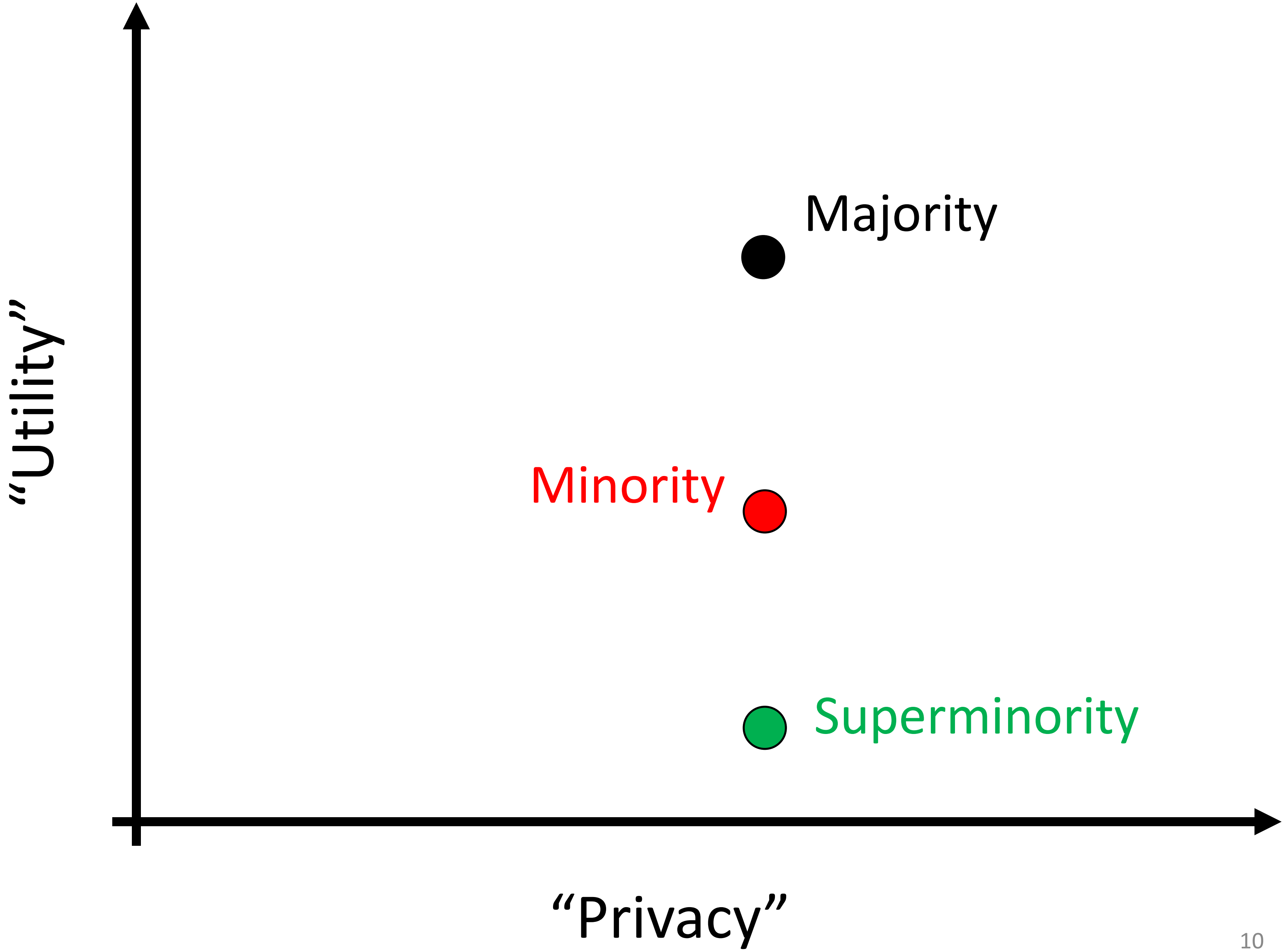
Demographic Fields	**Registered Tier**
Date of Birth	Random Shift (one year)
Dates of Events	Random Shift (one year)
Date of Death	Random Shift (one year)
Geolocation	Generalize (US State)
Marital status	No change
Own or rent	No change
Race/Ethnicity	Generalize
Sex/gender	Generalize
Sexuality	Generalize
Education	Generalize
Employment status	Generalize
Born in US or not	No change
Annual household income	No change
Death cause	Suppress
Living situation	Suppress
Active duty military	Suppress
Genomic data	Suppress
Motor vehicle accidents	Suppress
Free Text	Suppress

De-identification Is Not Ideal

- The smaller the population, the more likely facts (e.g., geographic area, race, sexual orientation) are amended in some way

- This can have major implications on bias and generalizability


Privacy Requirement
Minority
Majority
Superminority
"Utility"
"Privacy"

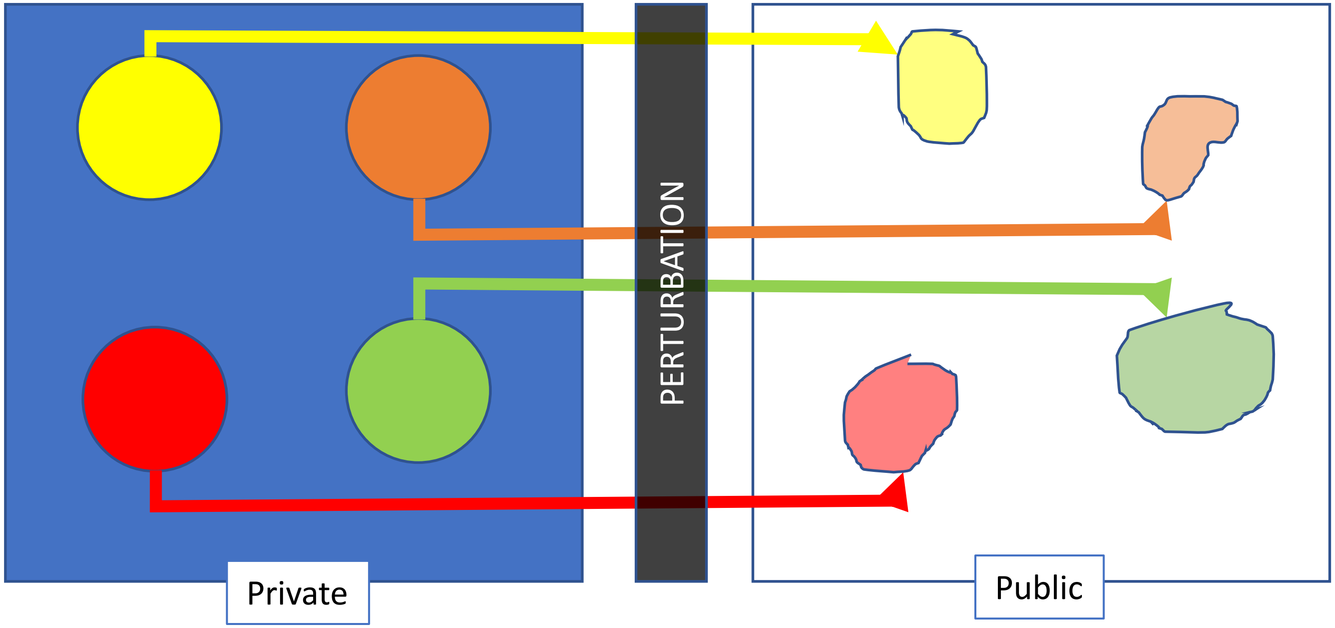

“Utility”
Majority
Minority
Superminority
“Privacy”

Consider An Alternative

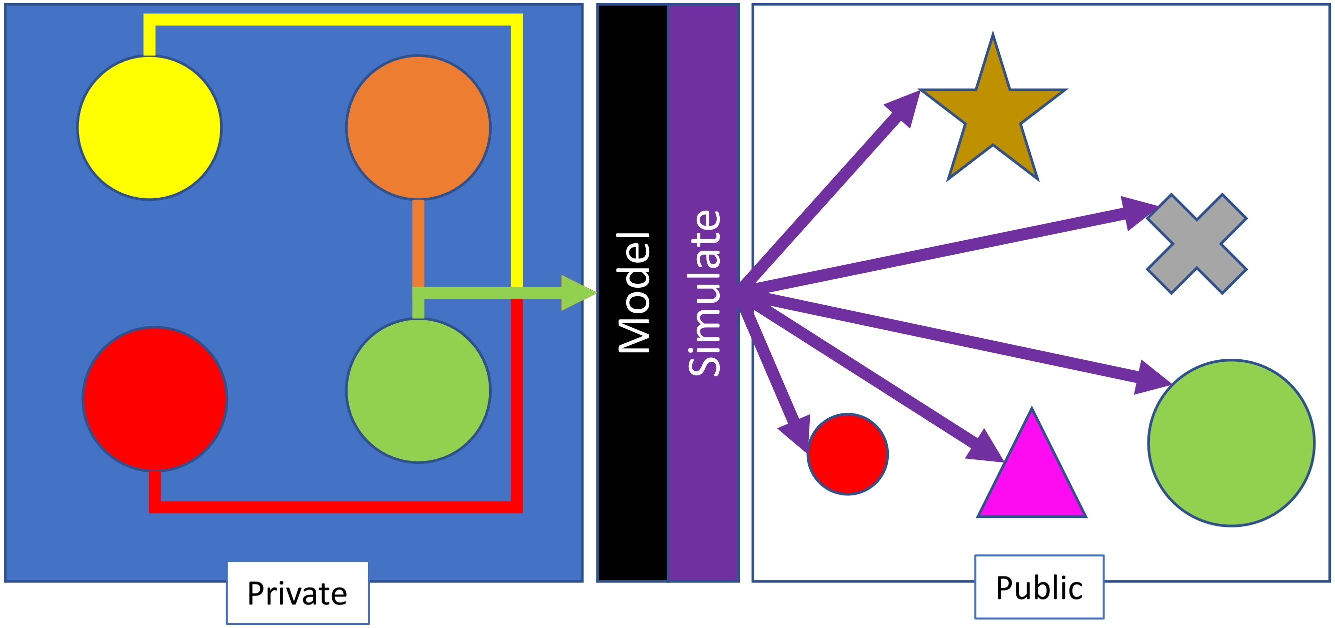
- Algorithmic bias often happens when there’s insufficient data on one population

- Can we “make” records for them?

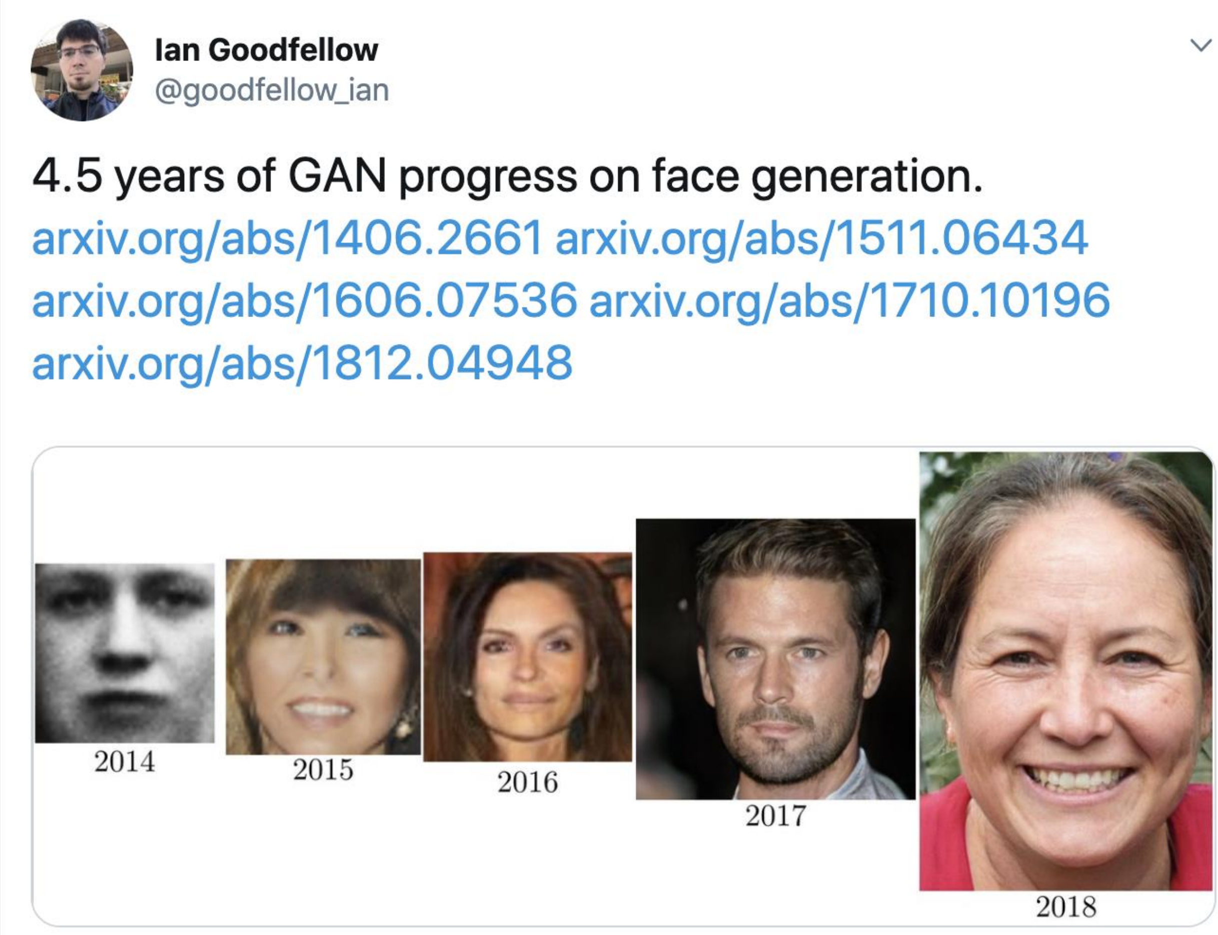
Generating Synthetic Data: Perturbation



Generating Synthetic Data: Simulation



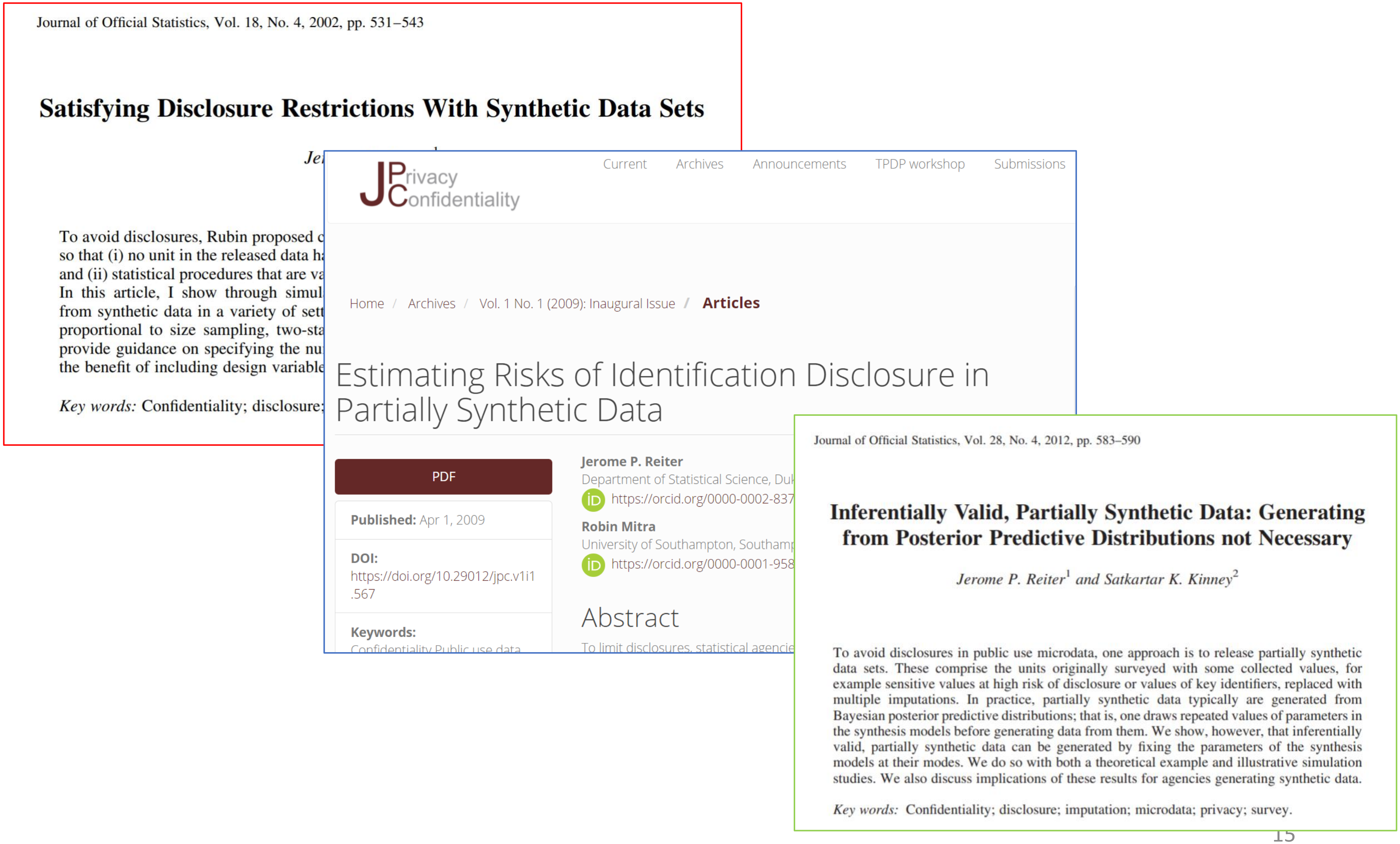
Synthetic Data is Not New



Ian Goodfellow
@goodfellow_ian

4.5 years of GAN progress on face generation.
arxiv.org/abs/1406.2661 arxiv.org/abs/1511.06434
arxiv.org/abs/1606.07536 arxiv.org/abs/1710.10196
arxiv.org/abs/1812.04948

And It's Older than You Think



Journal of Official Statistics, Vol. 18, No. 4, 2002, pp. 531–543

Satisfying Disclosure Restrictions With Synthetic Data Sets

Je

To avoid disclosures, Rubin proposed
so that (i) no unit in the released data h
and (ii) statistical procedures that are va
In this article, I show through simul
from synthetic data in a variety of sett
proportional to size sampling, two-sta
provide guidance on specifying the nu
the benefit of including design variable

Key words: Confidentiality; disclosure;

Home / Archives / Vol. 1 No. 1 (2009): Inaugural Issue / **Articles**

Estimating Risks of Identification Disclosure in Partially Synthetic Data

Published: Apr 1, 2009

DOI: https://doi.org/10.29012/jpc.v1i1.567

Keywords: Confidentiality Public use data

Jerome P. Reiter
Department of Statistical Science, Du
https://orcid.org/0000-0002-837

Robin Mitra
University of Southampton, Southam
https://orcid.org/0000-0001-958

Abstract

To limit disclosures, statistical agencie

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 583–590

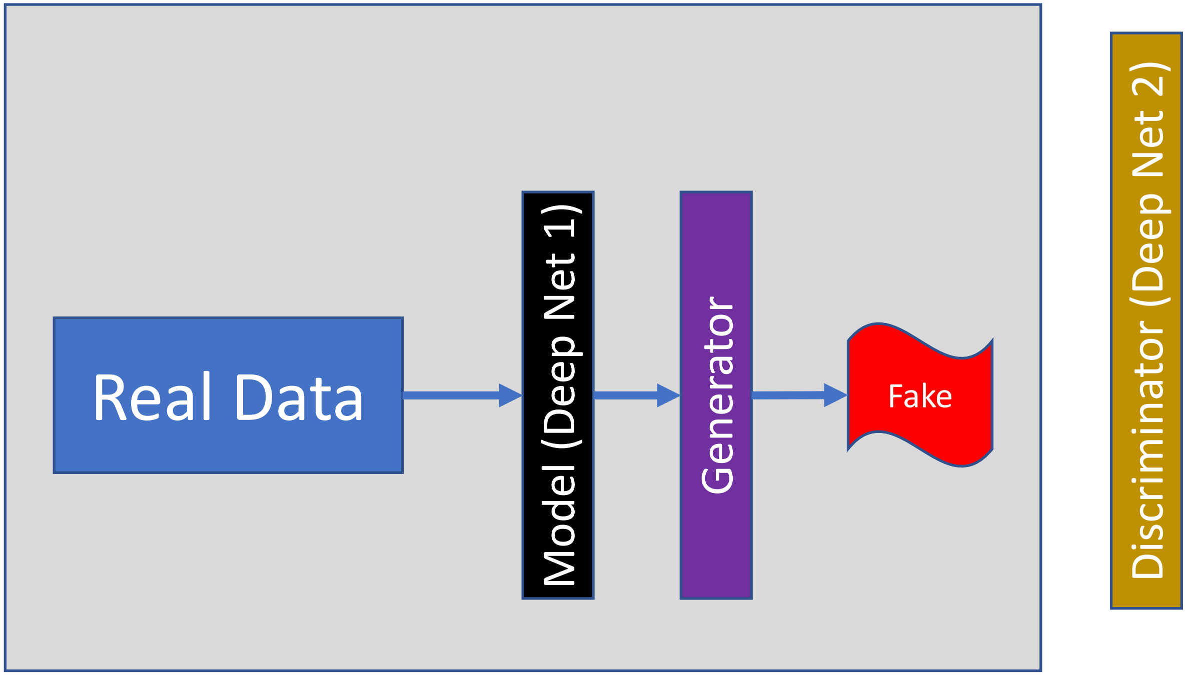
Inferentially Valid, Partially Synthetic Data: Generating from Posterior Predictive Distributions not Necessary

Jerome P. Reiter[1] *and Satkartar K. Kinney*[2]

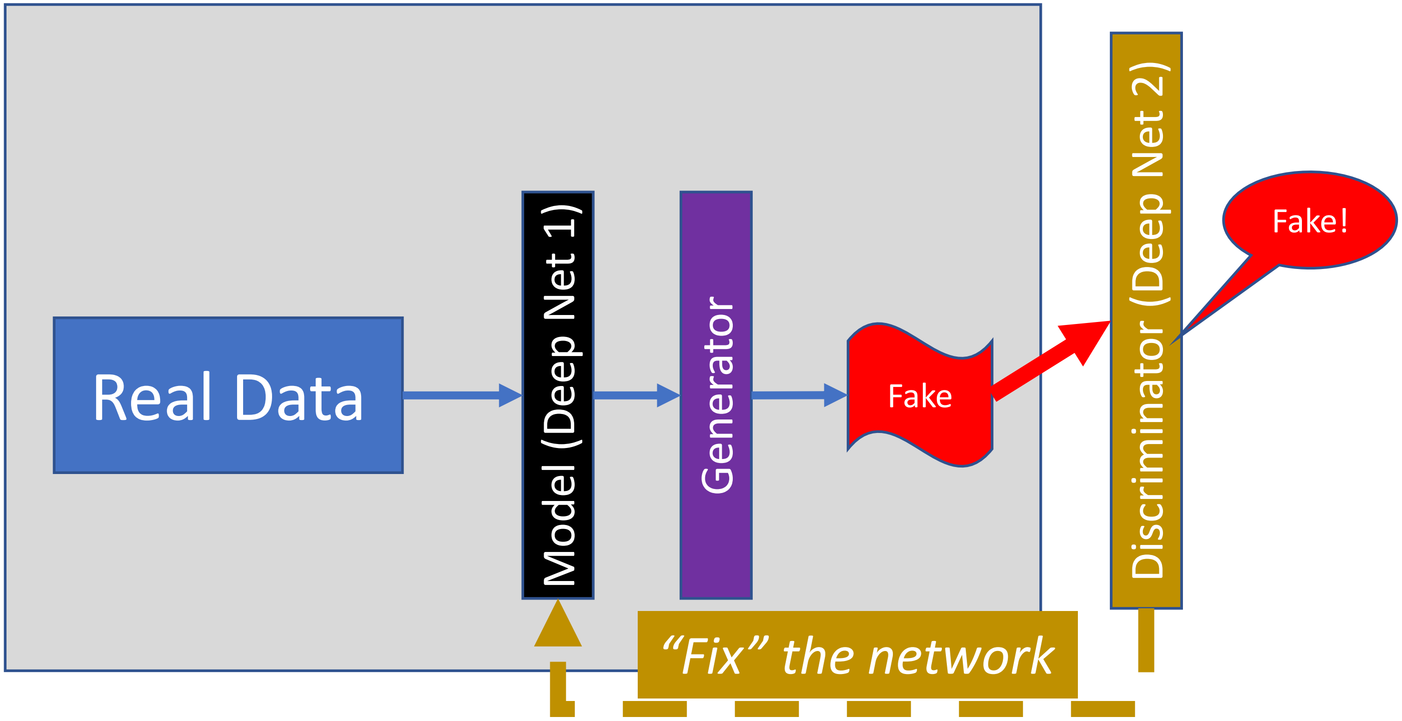
To avoid disclosures in public use microdata, one approach is to release partially synthetic data sets. These comprise the units originally surveyed with some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. In practice, partially synthetic data typically are generated from Bayesian posterior predictive distributions; that is, one draws repeated values of parameters in the synthesis models before generating data from them. We show, however, that inferentially valid, partially synthetic data can be generated by fixing the parameters of the synthesis models at their modes. We do so with both a theoretical example and illustrative simulation studies. We also discuss implications of these results for agencies generating synthetic data.

Key words: Confidentiality; disclosure; imputation; microdata; privacy; survey.

Generative Adversarial Networks (GANs)



Playing the GAN Game



Real Data

Model (Deep Net 1)

Generator

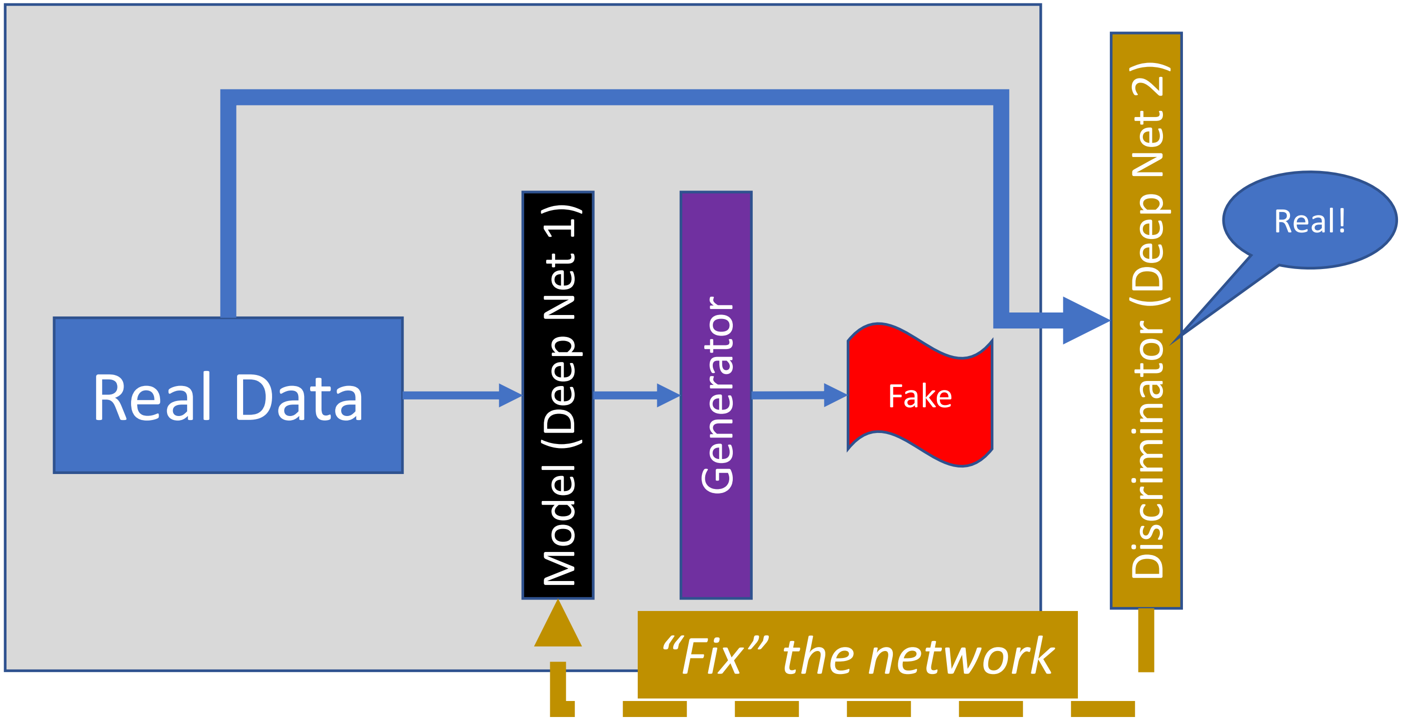
Fake

Discriminator (Deep Net 2)

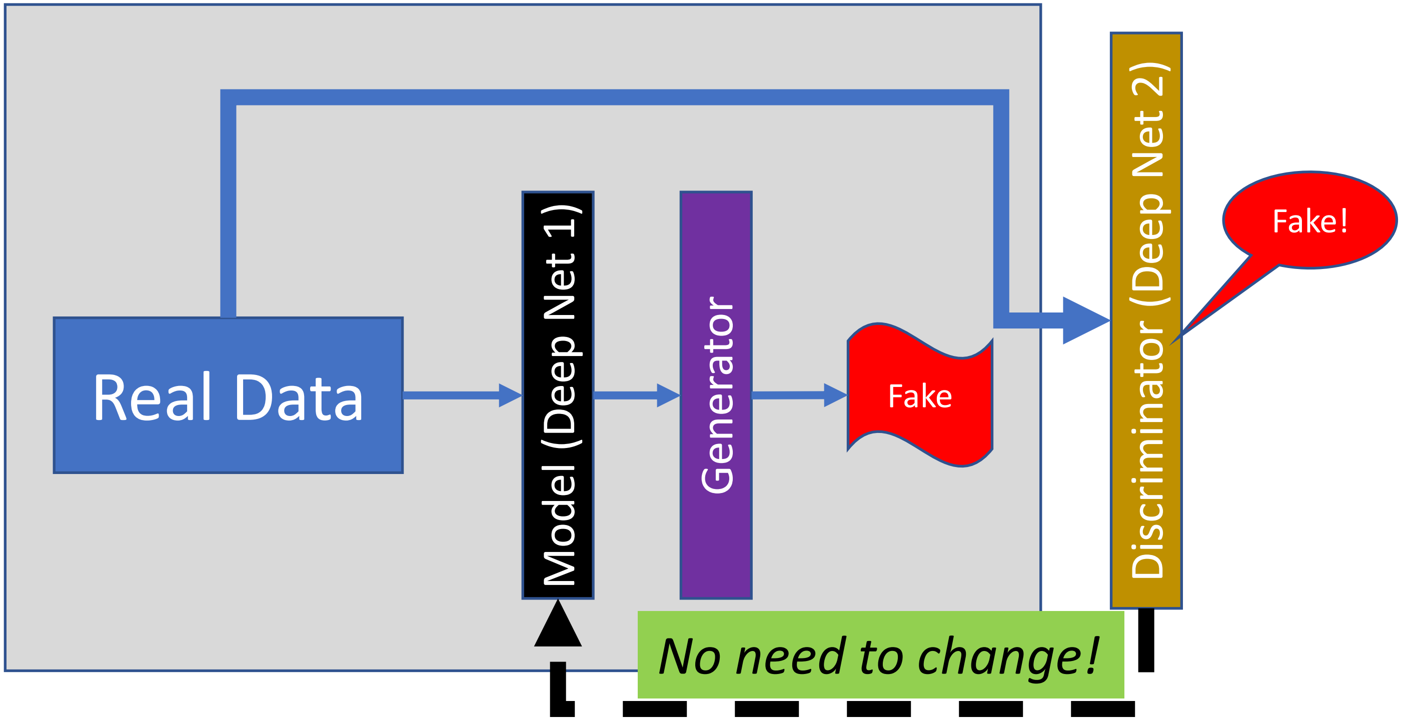
Fake!

“Fix” the network

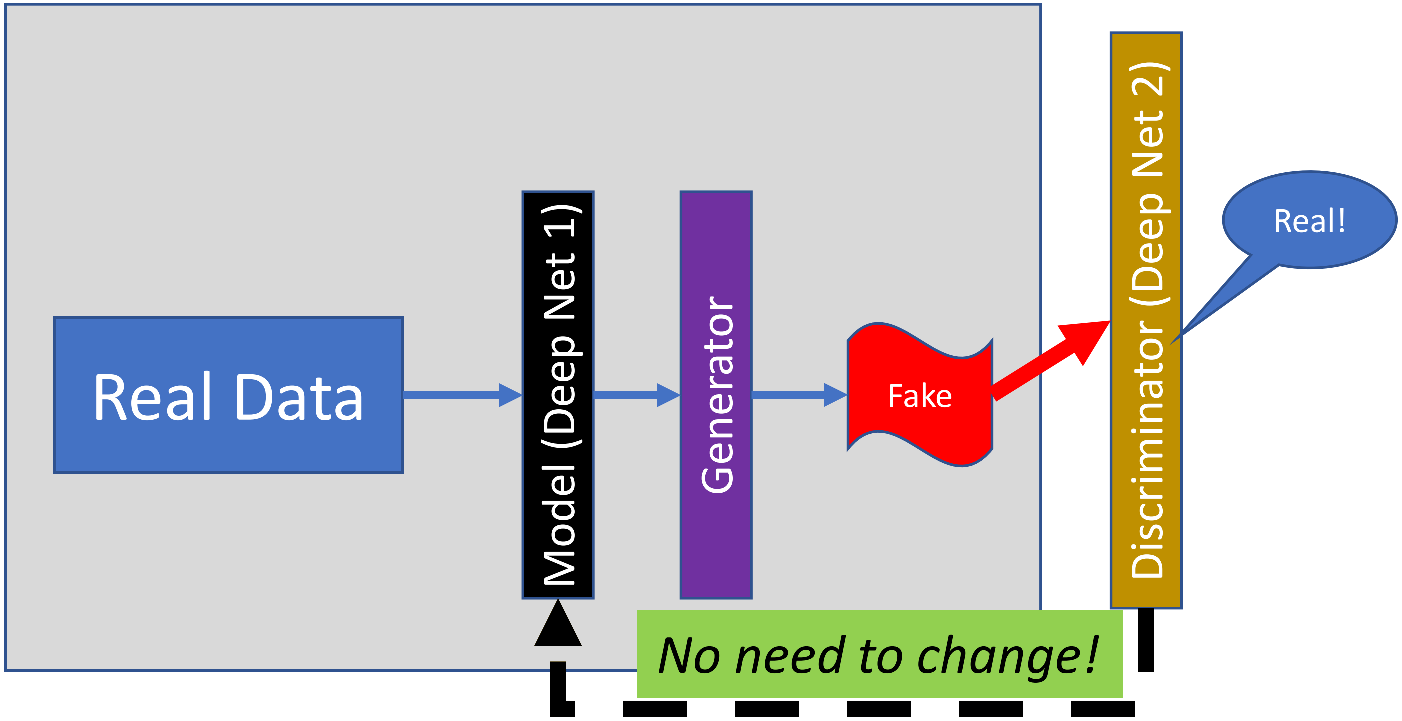
Playing the GAN Game



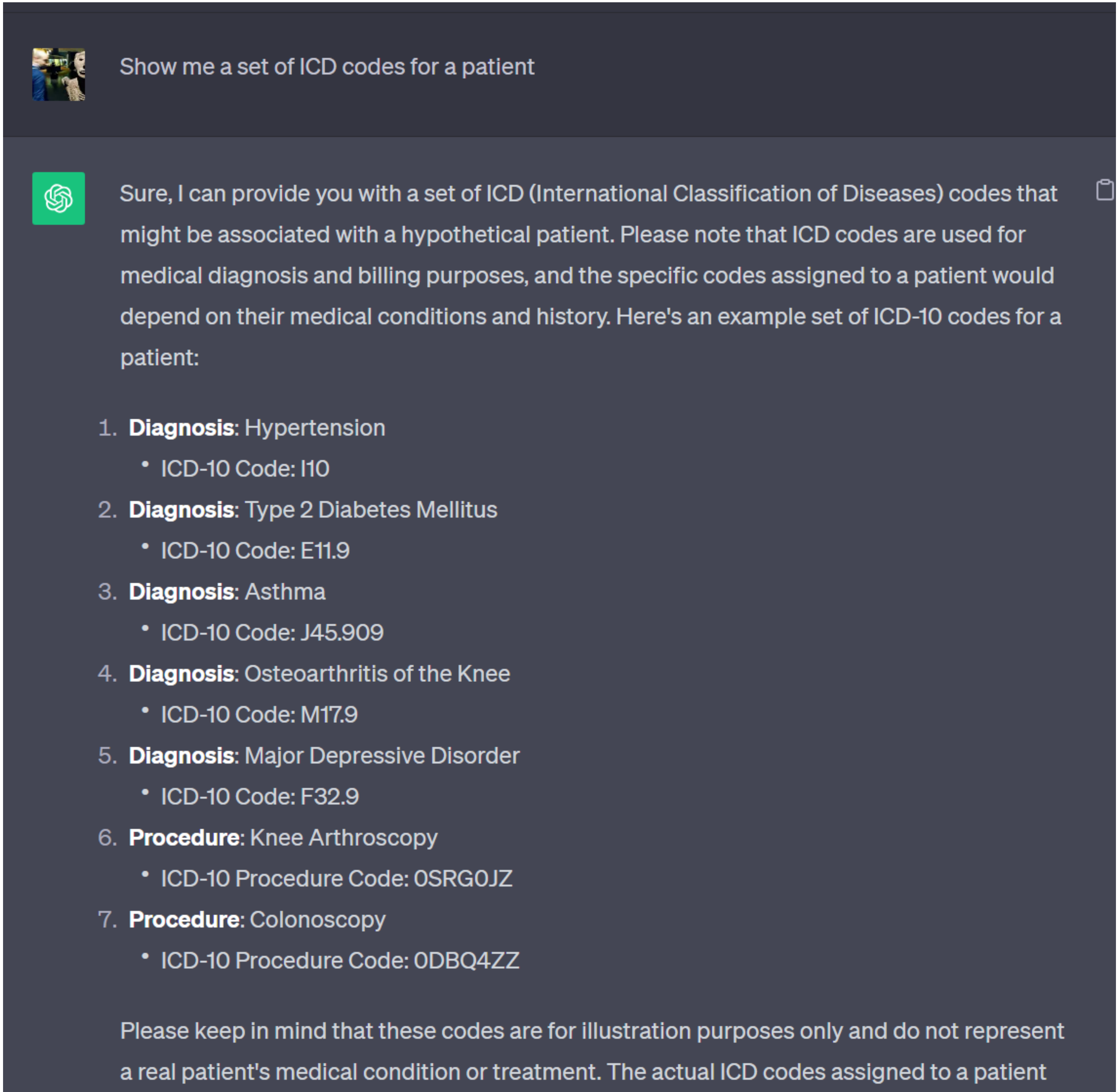
Playing the GAN Game



Playing the GAN Game



No need to change!



Show me a set of ICD codes for a patient

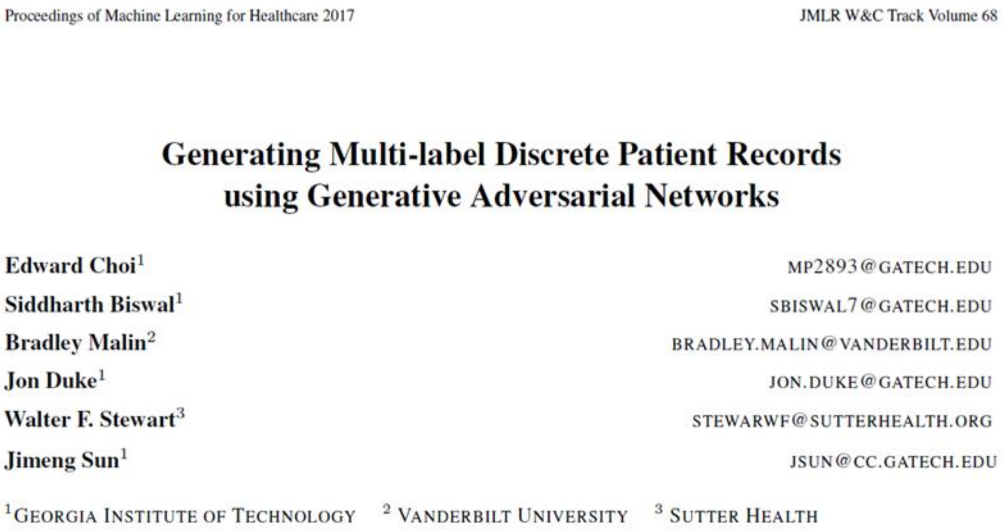
Sure, I can provide you with a set of ICD (International Classification of Diseases) codes that might be associated with a hypothetical patient. Please note that ICD codes are used for medical diagnosis and billing purposes, and the specific codes assigned to a patient would depend on their medical conditions and history. Here's an example set of ICD-10 codes for a patient:

1. **Diagnosis**: Hypertension
   - ICD-10 Code: I10
2. **Diagnosis**: Type 2 Diabetes Mellitus
   - ICD-10 Code: E11.9
3. **Diagnosis**: Asthma
   - ICD-10 Code: J45.909
4. **Diagnosis**: Osteoarthritis of the Knee
   - ICD-10 Code: M17.9
5. **Diagnosis**: Major Depressive Disorder
   - ICD-10 Code: F32.9
6. **Procedure**: Knee Arthroscopy
   - ICD-10 Procedure Code: 0SRG0JZ
7. **Procedure**: Colonoscopy
   - ICD-10 Procedure Code: 0DBQ4ZZ

Please keep in mind that these codes are for illustration purposes only and do not represent a real patient's medical condition or treatment. The actual ICD codes assigned to a patient

Medical Record Data Simulation

(Choi MLHC 2017)



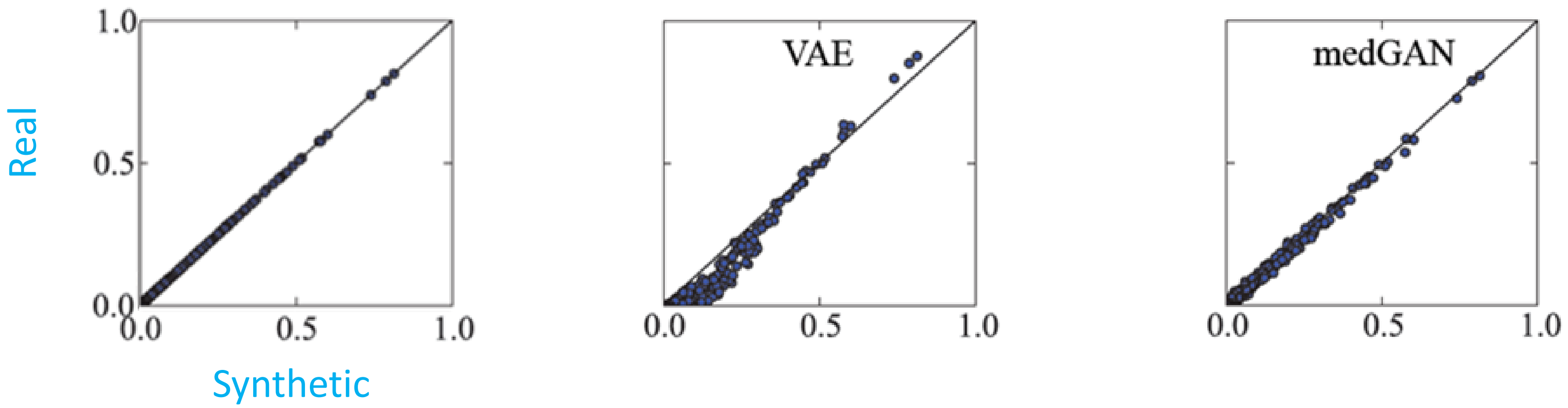
Proceedings of Machine Learning for Healthcare 2017 JMLR W&C Track Volume 68

Generating Multi-label Discrete Patient Records using Generative Adversarial Networks

Edward Choi[1] MP2893@GATECH.EDU
Siddharth Biswal[1] SBISWAL7@GATECH.EDU
Bradley Malin[2] BRADLEY.MALIN@VANDERBILT.EDU
Jon Duke[1] JON.DUKE@GATECH.EDU
Walter F. Stewart[3] STEWARWF@SUTTERHEALTH.ORG
Jimeng Sun[1] JSUN@CC.GATECH.EDU

[1] GEORGIA INSTITUTE OF TECHNOLOGY [2] VANDERBILT UNIVERSITY [3] SUTTER HEALTH

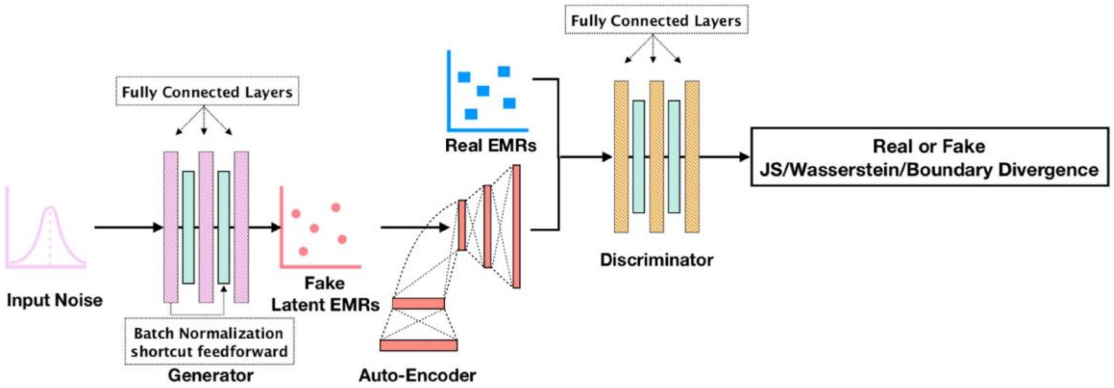
- Sutter Health & MIMIC
- Demographics, Diagnoses, Procedures, & Meds
- Prediction of presence / absence clinical concept



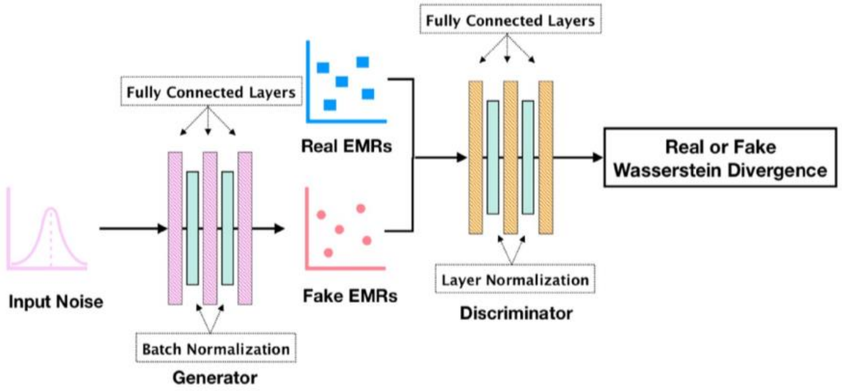
Evolution

- Better training and evaluation methods

(latent dimensions) (Zhang et al JAMIA 2020)



(a) medGAN, medWGAN and medBGAN



(b) EMR-WGAN

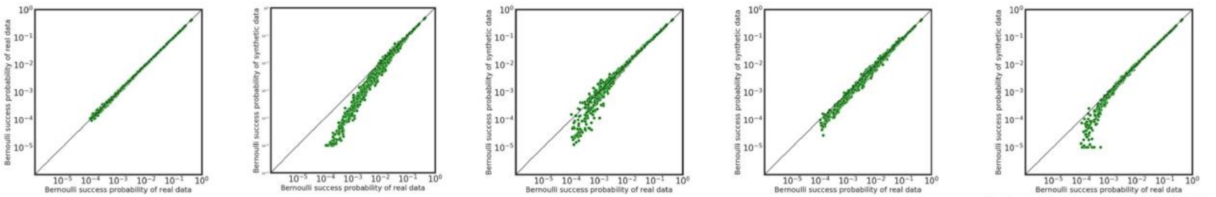
Zhang, Yan, Mesa, Sun, & Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. JAMIA. 2020; 27: 99-108.
Yan, Zhang, Nyemba, & Malin. Generating electronic health records with multiple data types and constraints. Proc AMIA Symp. 2020: 1335-1344.
Zhang, Yan, Lasko, Sun, & Malin. SynTEG: A framework for temporal structured electronic health data simulation. JAMIA. 2021; 28: 596-604.
Zhang, Yan, & Malin. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. JAMIA 2022; 29: 1890-1898.

Evolution

- Better training and evaluation methods

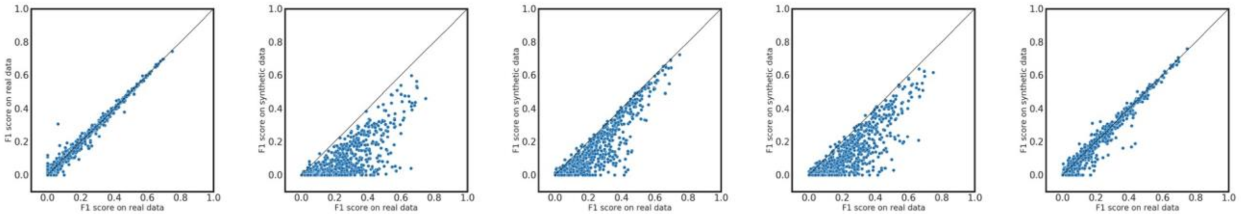
(latent dimensions) (Zhang et al JAMIA 2020)

First Order Statistics

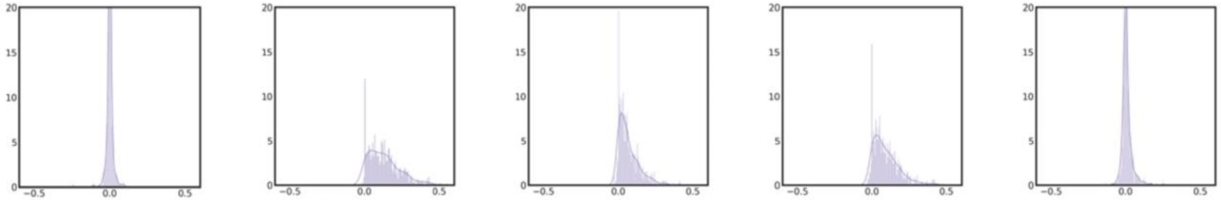


(a) Real (b) medGAN (c) medWGAN (d) medBGAN (e) EMR-WGAN.

Predictive Performance



(a) Real. (c) medGAN. (e) medWGAN. (g) medBGAN. (i) EMR-WGAN.



(b) Distribution of dot-to-diagonal distance of (a). (d) Distribution of dot-to-diagonal distance of (c). (f) Distribution of dot-to-diagonal distance of (e). (h) Distribution of dot-to-diagonal distance of (g). (j) Distribution of dot-to-diagonal distance of (i).

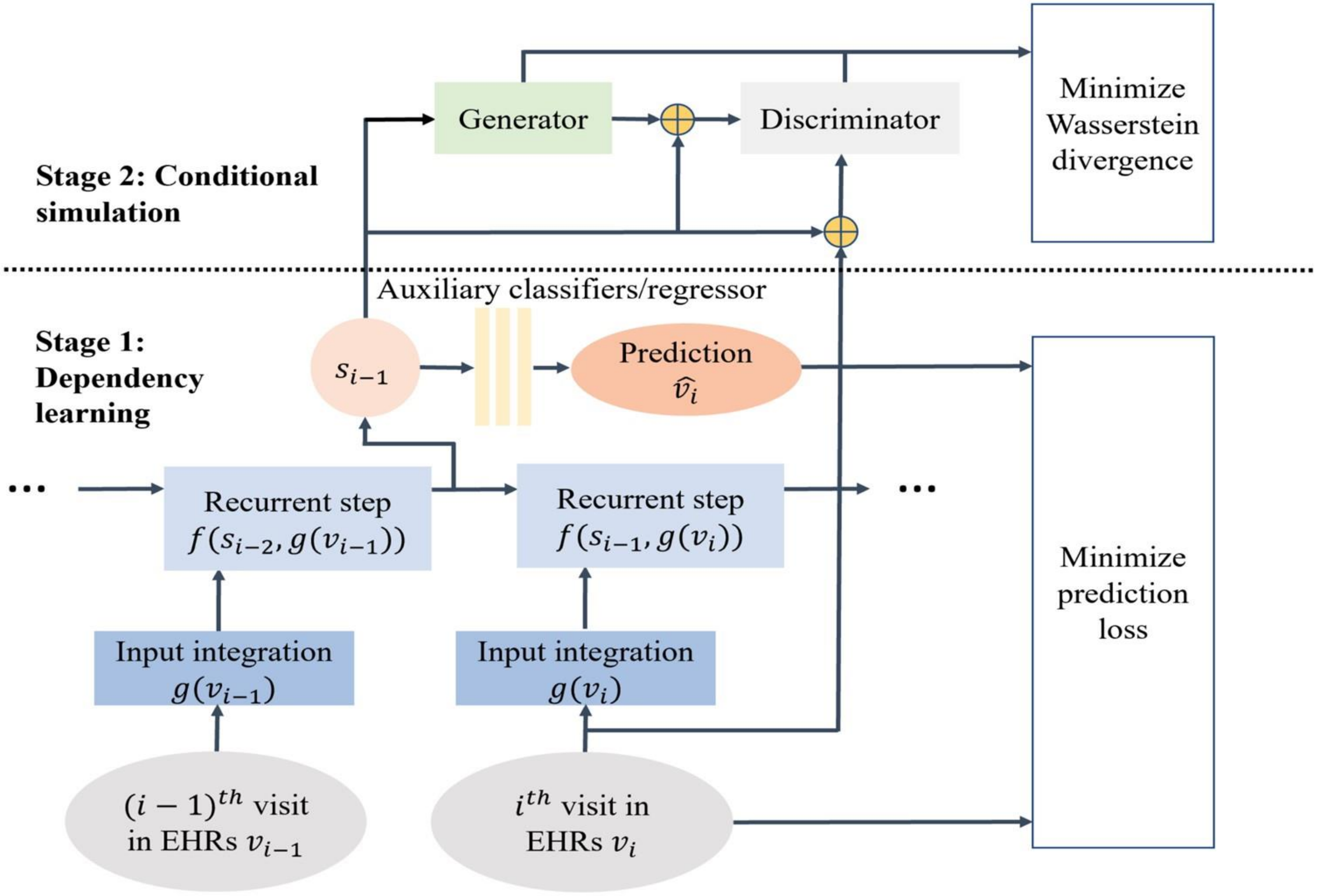
Zhang, Yan, Mesa, Sun, & Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. JAMIA. 2020; 27: 99-108.
Yan, Zhang, Nyemba, & Malin. Generating electronic health records with multiple data types and constraints. Proc AMIA Symp. 2020: 1335-1344.
Zhang, Yan, Lasko, Sun, & Malin. SynTEG: A framework for temporal structured electronic health data simulation. JAMIA. 2021; 28: 596-604.
Zhang, Yan, & Malin. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. JAMIA 2022; 29: 1890-1898.

Evolution

- Better training and evaluation methods (latent dimensions) (Zhang et al JAMIA 2020)
- Enabling constraints (e.g., preventing women from having prostate cancer) (Yan et al AMIA 2020)
- Move from static to longitudinal data (Zhang et al JAMIA 2021; JAMIA 2022)

Zhang, Yan, Mesa, Sun, & Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. JAMIA. 2020; 27: 99-108.
Yan, Zhang, Nyemba, & Malin. Generating electronic health records with multiple data types and constraints. Proc AMIA Symp. 2020: 1335-1344.
Zhang, Yan, Lasko, Sun, & Malin. SynTEG: A framework for temporal structured electronic health data simulation. JAMIA. 2021; 28: 596-604.
Zhang, Yan, & Malin. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. JAMIA 2022; 29: 1890-1898.

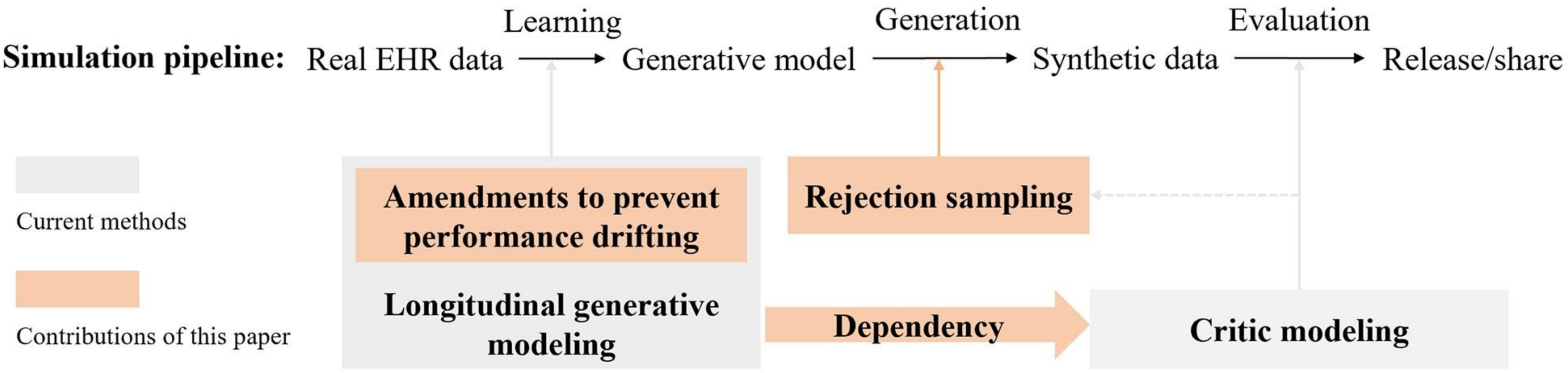
Evolution (Zhang 2021)


Generator
Discriminator
Minimize Wasserstein divergence
Stage 2: Conditional simulation
Auxiliary classifiers/regressor
Stage 1: Dependency learning
s_{i-1}
Prediction $\hat{v}_i$
Recurrent step $f(s_{i-2}, g(v_{i-1}))$
Recurrent step $f(s_{i-1}, g(v_i))$
…
…
Minimize prediction loss
Input integration $g(v_{i-1})$
Input integration $g(v_i)$
$(i-1)^{th}$ visit in EHRs v_{i-1}
i^{th} visit in EHRs v_i

Zhang, Yan, Mesa, Sun, & Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. JAMIA. 2020; 27: 99-108.
Yan, Zhang, Nyemba, & Malin. Generating electronic health records with multiple data types and constraints. Proc AMIA Symp. 2020: 1335-1344.
Zhang, Yan, Lasko, Sun, & Malin. SynTEG: A framework for temporal structured electronic health data simulation. JAMIA. 2021; 28: 596-604.
Zhang, Yan, & Malin. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. JAMIA 2022; 29: 1890-1898.

Evolution (Zhang 2022)

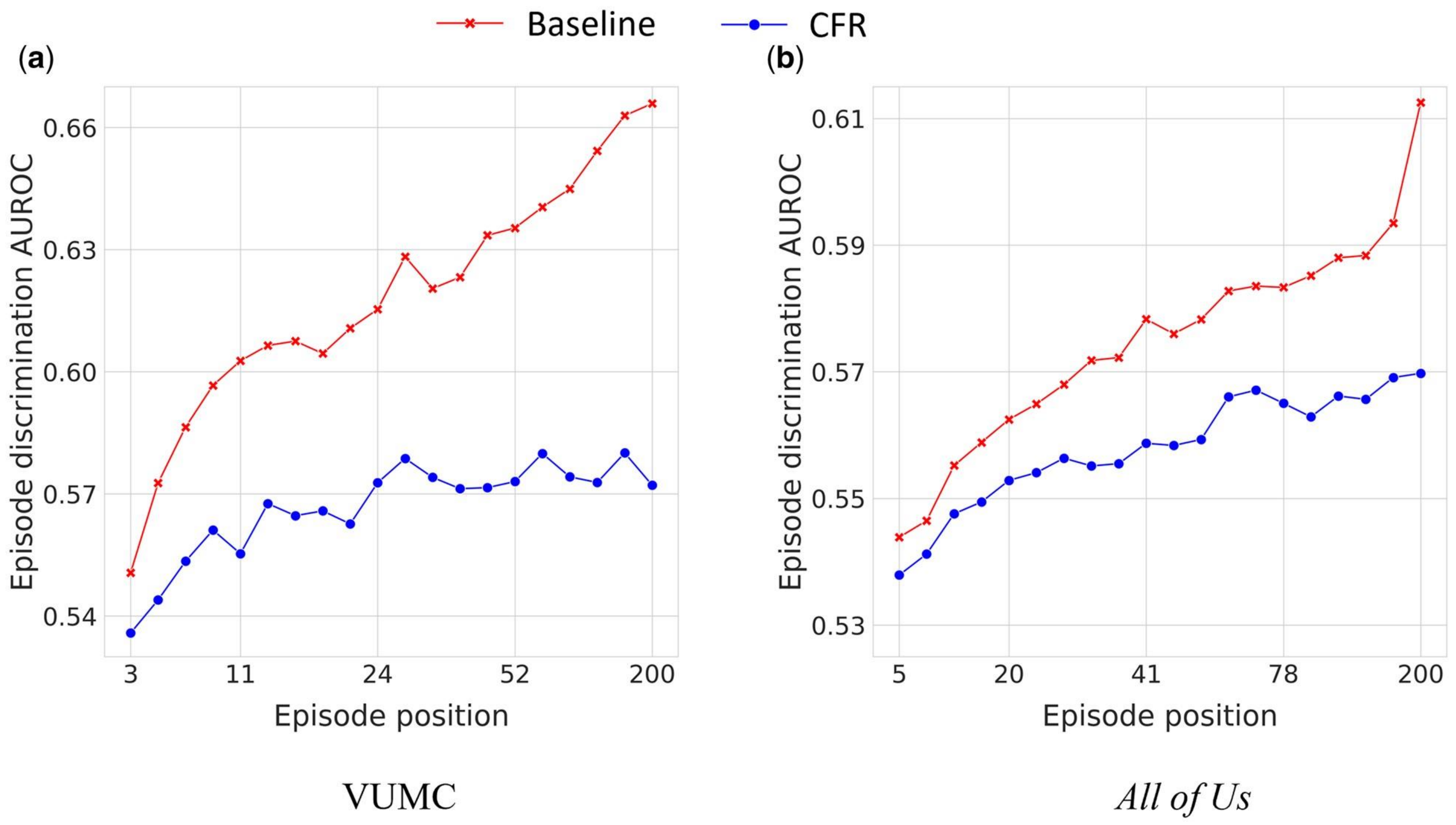
Condition Fuzzing & Regularization (CFR)



Learning
Generation
Evaluation
Simulation pipeline: Real EHR data → Generative model → Synthetic data → Release/share
Current methods
Contributions of this paper
Amendments to prevent performance drifting
Longitudinal generative modeling
Rejection sampling
Dependency
Critic modeling

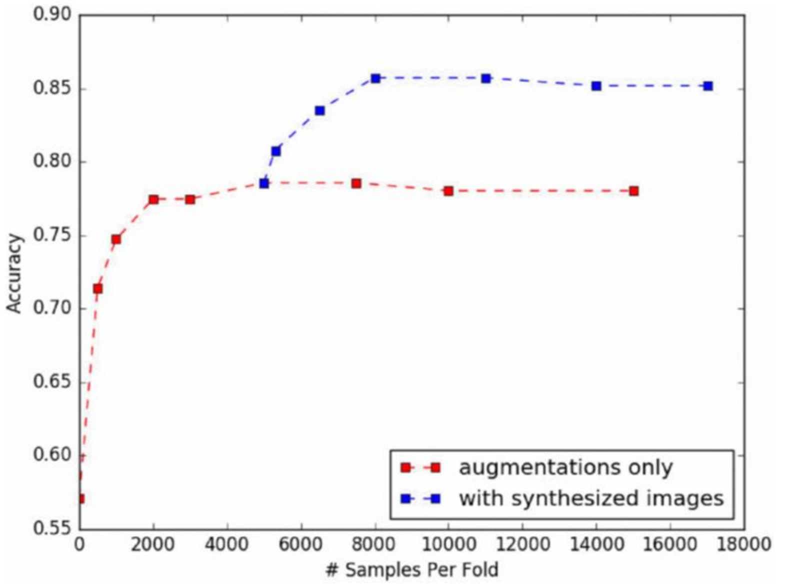
Zhang, Yan, Mesa, Sun, & Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. JAMIA. 2020; 27: 99-108.
Yan, Zhang, Nyemba, & Malin. Generating electronic health records with multiple data types and constraints. Proc AMIA Symp. 2020: 1335-1344.
Zhang, Yan, Lasko, Sun, & Malin. SynTEG: A framework for temporal structured electronic health data simulation. JAMIA. 2021; 28: 596-604.
Zhang, Yan, & Malin. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. JAMIA 2022; 29: 1890-1898.

Stop the Drift! (Zhang 2022)

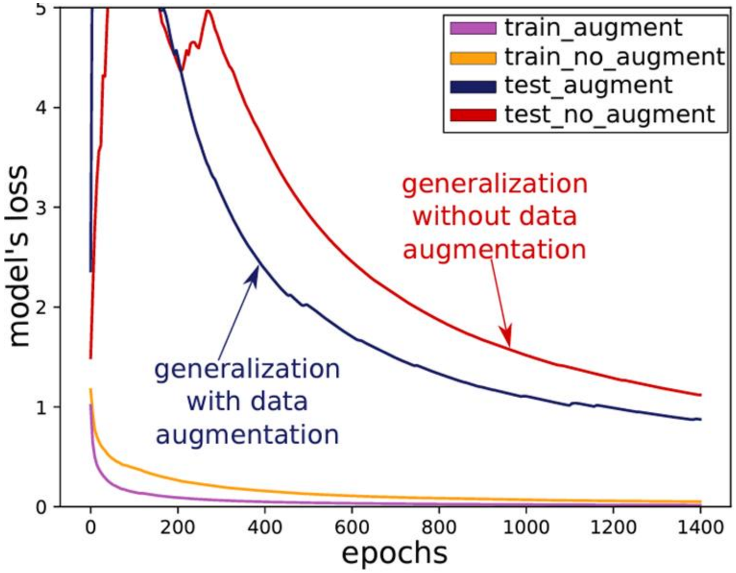


Zhang, Yan, Mesa, Sun, & Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. JAMIA. 2020; 27: 99-108.
Yan, Zhang, Nyemba, & Malin. Generating electronic health records with multiple data types and constraints. Proc AMIA Symp. 2020: 1335-1344.
Zhang, Yan, Lasko, Sun, & Malin. SynTEG: A framework for temporal structured electronic health data simulation. JAMIA. 2021; 28: 596-604.
Zhang, Yan, & Malin. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. JAMIA 2022; 29: 1890-1898.

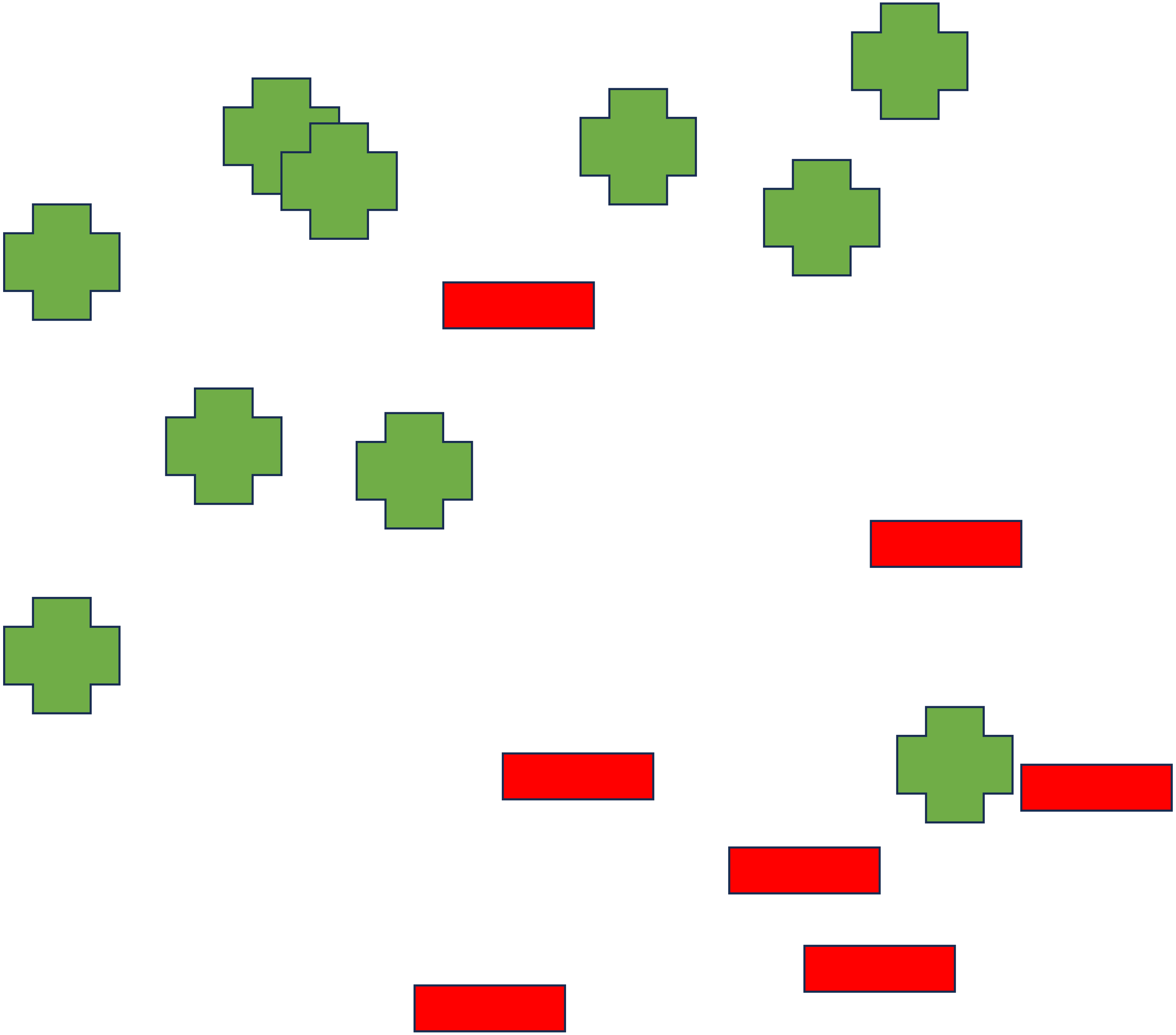
Evidence for Improved Classifiers

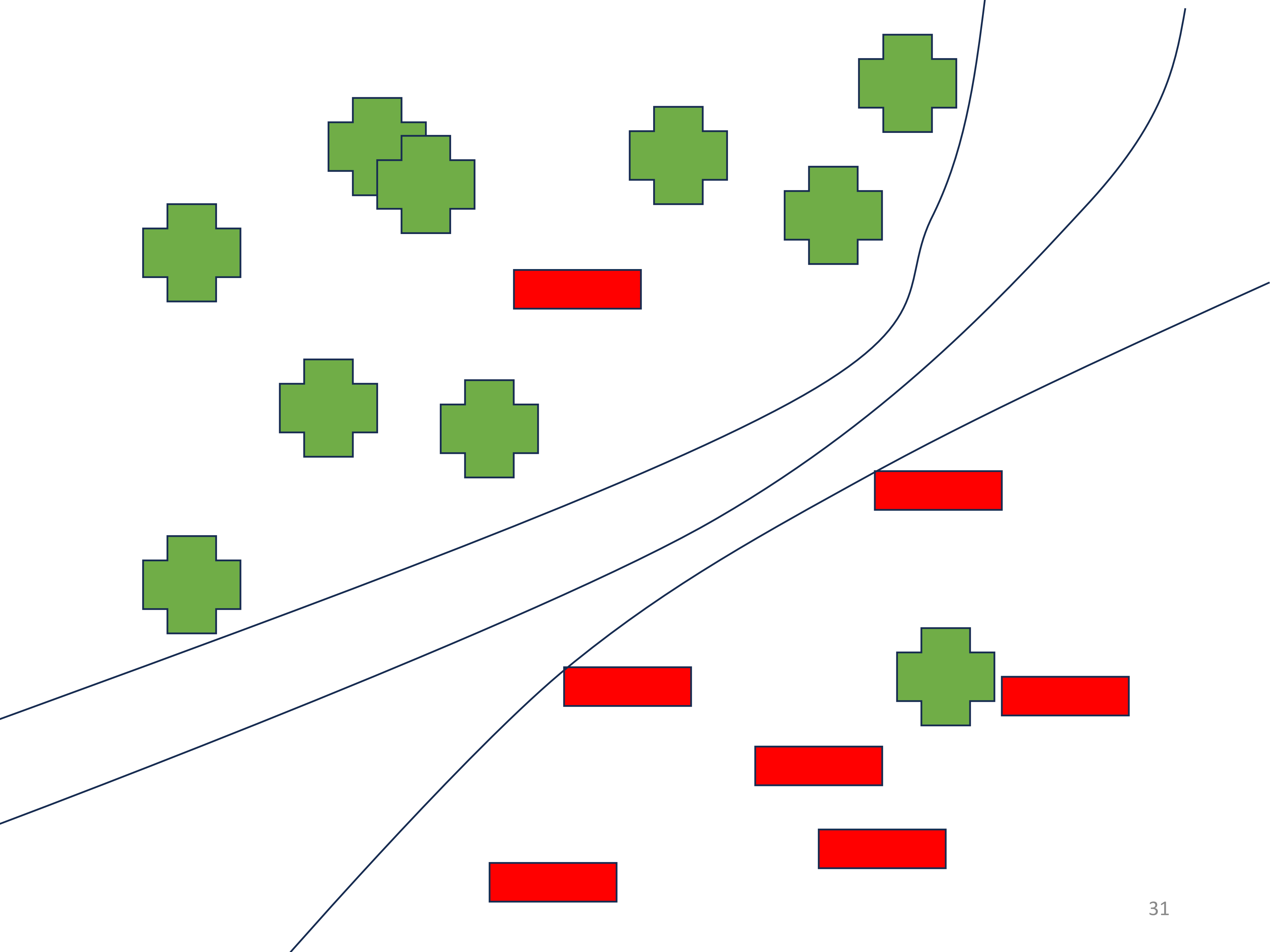


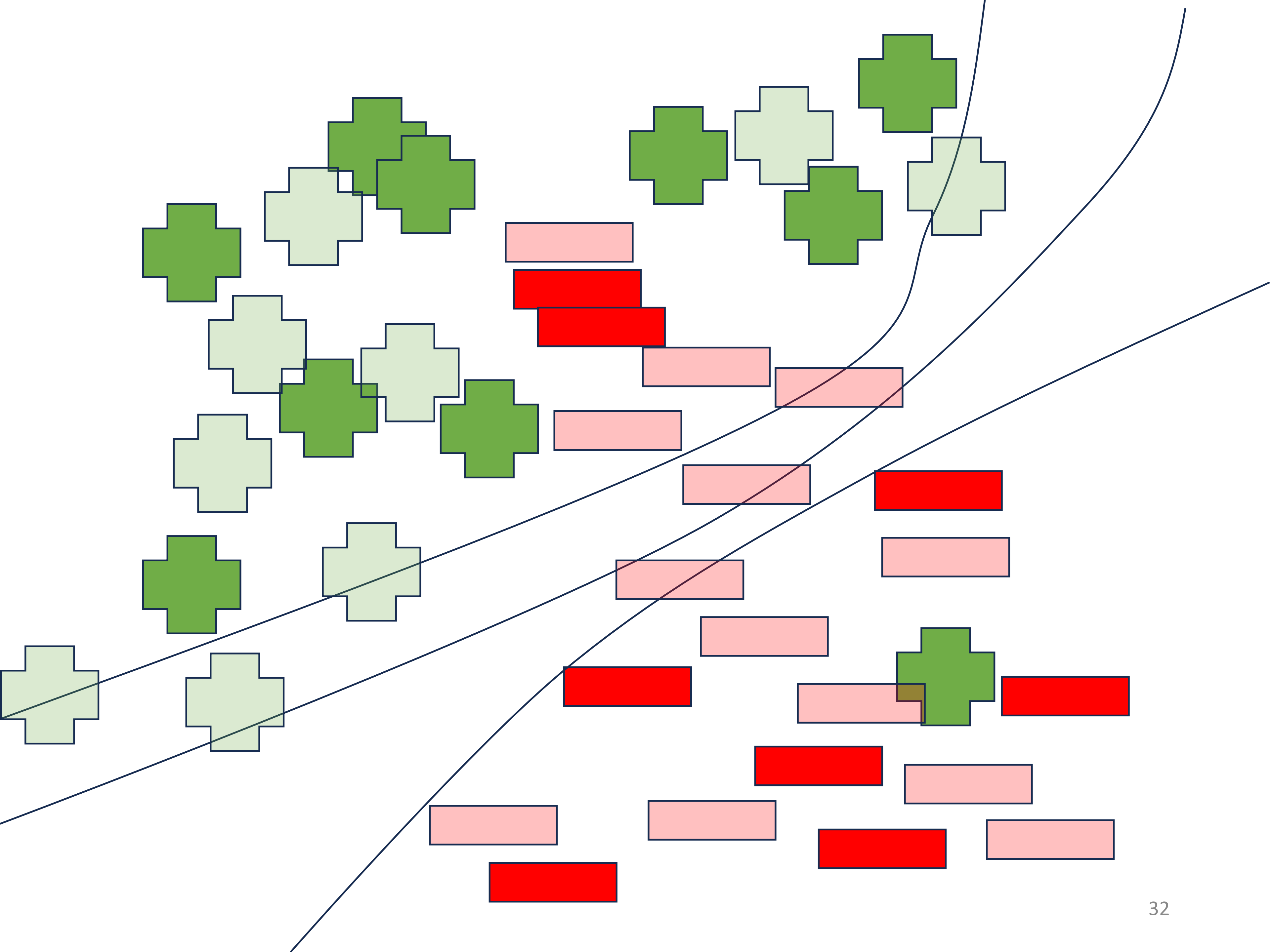
Frid-Adar M, et al. Synthetic Data Augmentation Using GAN for Improved Liver Lesion Classification. 15th International Symposium on Biomedical Imaging. **2018**.

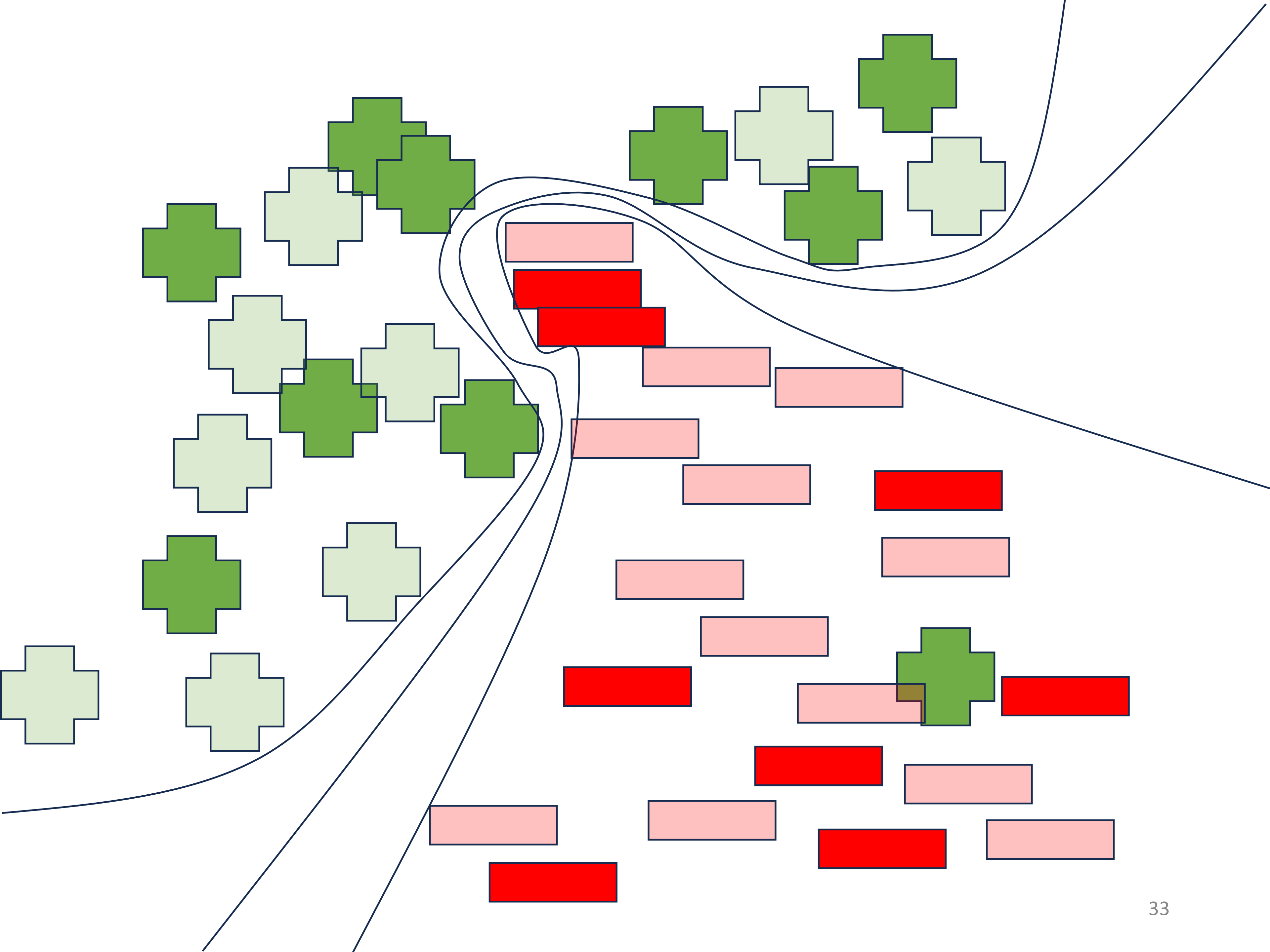


Fawaz, Forestier, Weber, Idoumghar, and Muller. Data augmentation using synthetic data for time series classification with deep residual networks. ECML/PKDD Workshop. **2018**.









Did Generative AI Just Save the World?

- Not quite

- Hallucinations are real…

  … research is advancing on how to prevent it, but still a ways to go

- Need more to validate the clinical face validity…

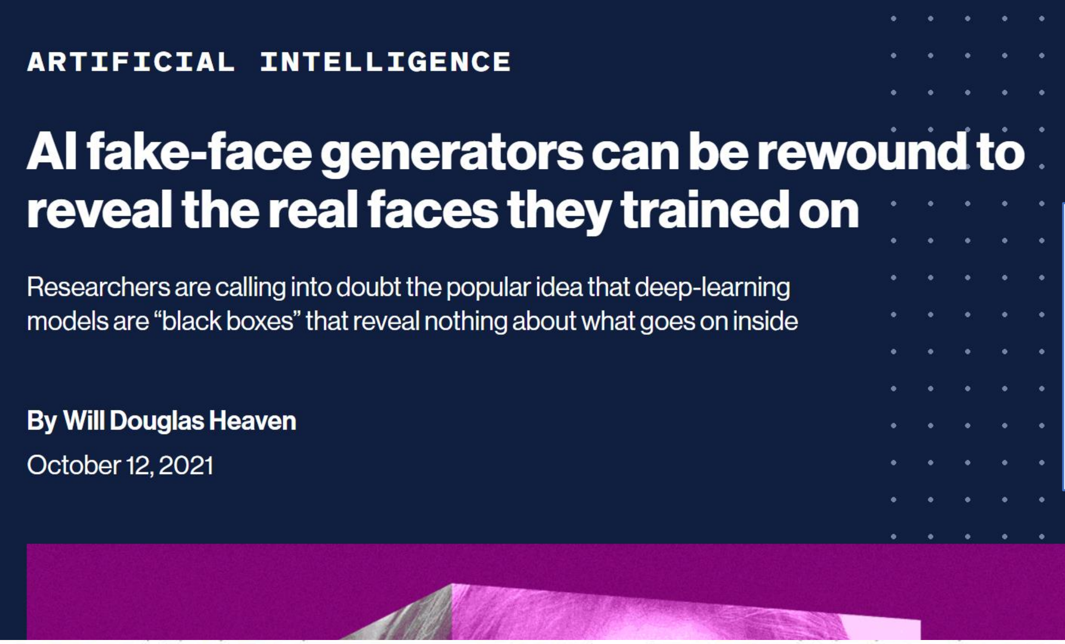
  … easier said than done

- If there isn’t a sufficient amount of data to begin with…

  … generating synthetic data is just generating noise

- There is a limit to learning…

  … as the number of variables grows, the chance that the data is “useful”



ARTIFICIAL INTELLIGENCE

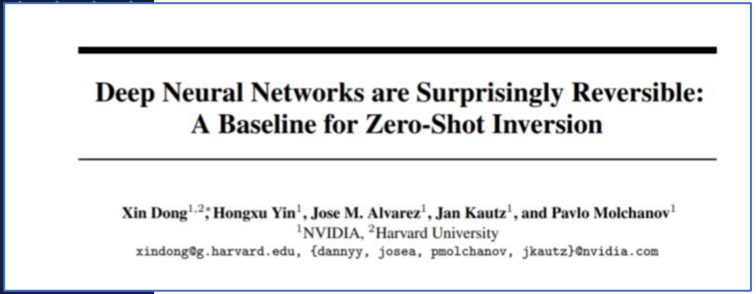
AI fake-face generators can be rewound to reveal the real faces they trained on

Researchers are calling into doubt the popular idea that deep-learning models are “black boxes” that reveal nothing about what goes on inside

By Will Douglas Heaven

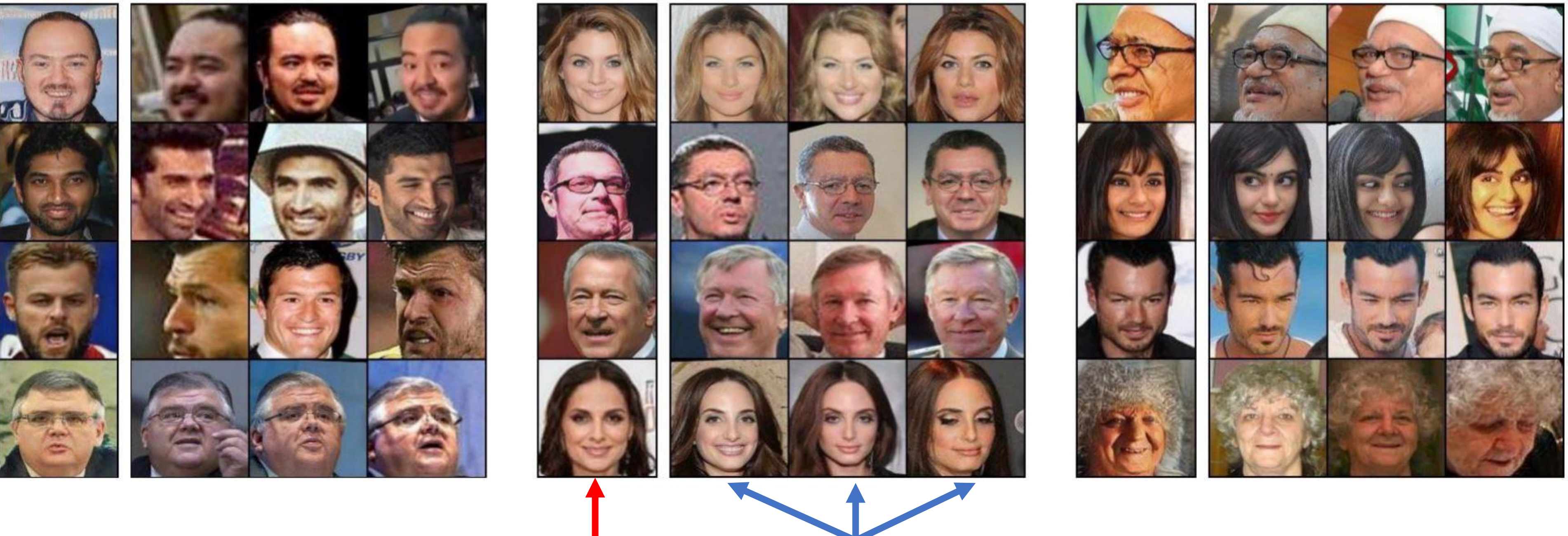
October 12, 2021

https://arxiv.org/abs/2107.06304



Deep Neural Networks are Surprisingly Reversible: A Baseline for Zero-Shot Inversion

Xin Dong[1,2]*, Hongxu Yin[1], Jose M. Alvarez[1], Jan Kautz[1], and Pavlo Molchanov[1]
[1]NVIDIA, [2]Harvard University
`xindong@g.harvard.edu, {dannyy, josea, pmolchanov, jkautz}@nvidia.com`



FAKE

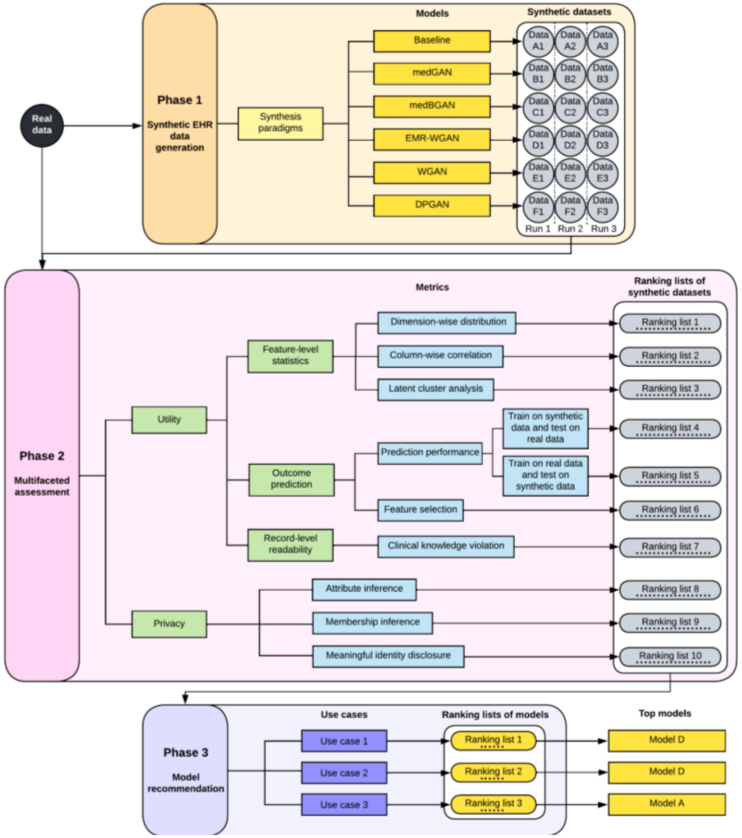
Real

Concerns for Health Data

- Mimicry
  - Insufficient training data can lead to “mimicking” of original records
- Membership Inference
  - User can test if features of someone they know appear to be in the training data
  - Requires knowing the features in question
- Attribute Inference
  - User can predict features (they don’t know) about someone based on features they do know
- Combining Membership and Attribute is where disclosure occurs

Benchmarking

- Software to evaluating synthetic EHR data privacy and utility
  - https://github.com/yy6linda/synthetic-ehr-benchmarking
- Companion report out in Nature Communications*



*Yan, Yan, Wan, Zhang, Omberg, Guinney, Mooney, & Malin. A Multifaceted benchmarking of synthetic electronic health record generation models. Nature Communications. 2022.

Questions?

brad.malin@gmail.com

Center for Genetic Privacy and Identity in Community Settings
http://www.vumc.org/getprecise/

Health Data Science Center
http://www.vumc.org/heads/