Tutorial on

Synthetic Healthcare Data Generation and Assessment: Challenges, Methods, and Impact on Machine Learning

Mihaela van der Schaar
Ahmed M. Alaa

University of Cambridge
The Alan Turing Institute
University of California, Los Angeles

vanderschaar-lab.com

mv472@cam.ac.uk, ahmedmalaa@ucla.edu

ICML
International Conference
On Machine Learning

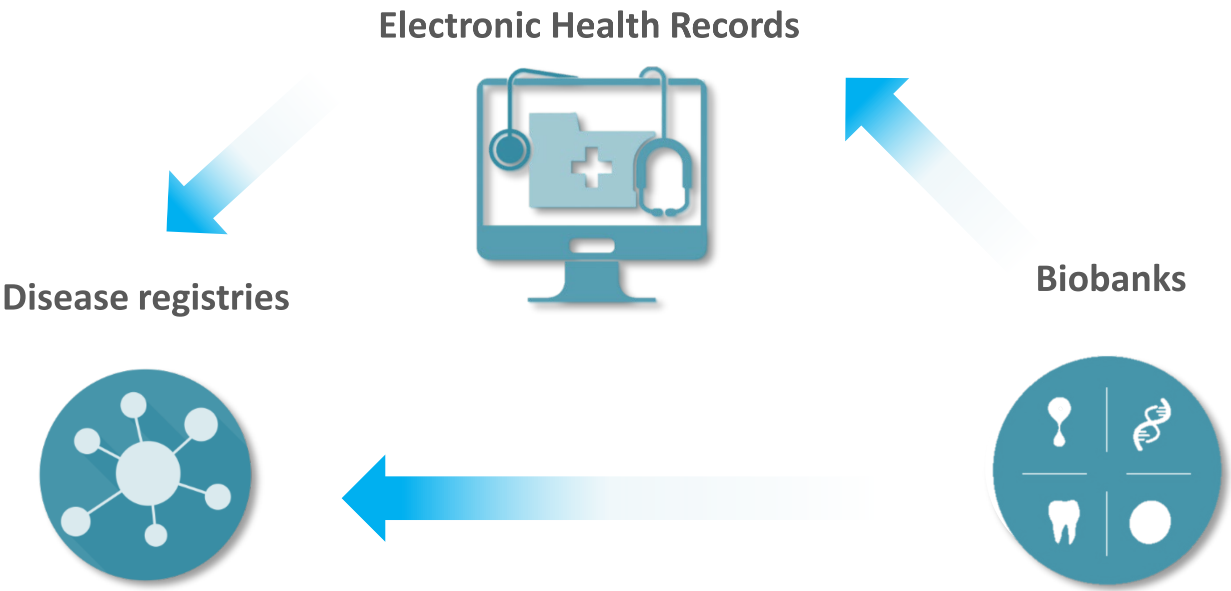
van_der_Schaar
\ LAB

UNIVERSITY OF
CAMBRIDGE

Healthcare data: an essential resource

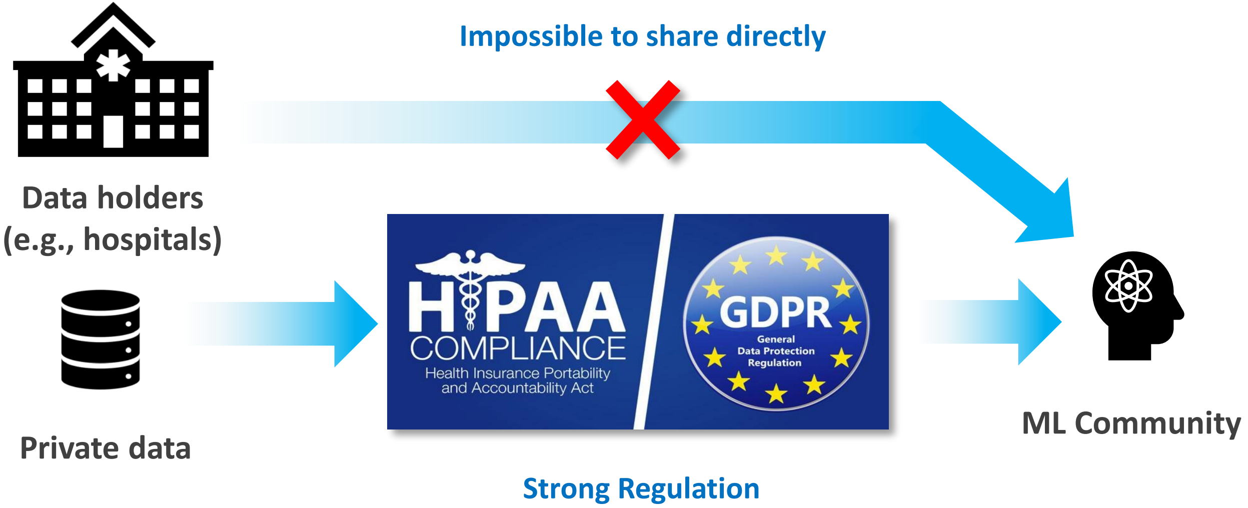
- **Availability of healthcare data resources:**

can catalyze a complete transformation in healthcare using ML – improve clinical practice & clinical workflows, empower patients, enhance drug development, enable medical knowledge discovery etc.



Healthcare data: not easy to access

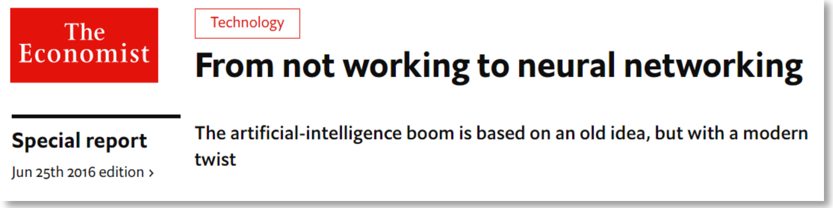
- Strict regulations for data access
- …the result of perfectly valid concerns regarding privacy



- **Lack of high-quality healthcare data: impedes ML research in healthcare!**

What history teaches us...

- Open access data sets = significant progress!



The Economist

Technology

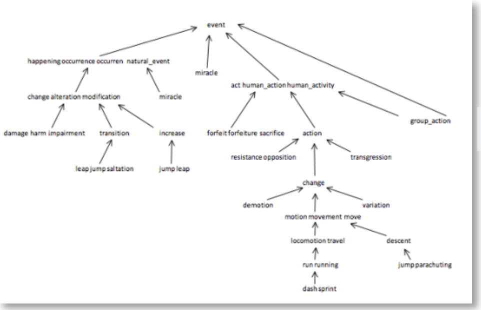
From not working to neural networking

Special report

Jun 25th 2016 edition ›

The artificial-intelligence boom is based on an old idea, but with a modern twist

[C. Fellbaum et al, 1980s]



WordNet

Progress in NLP

[Fei-Fei Li et al, 2006]

ImageNet

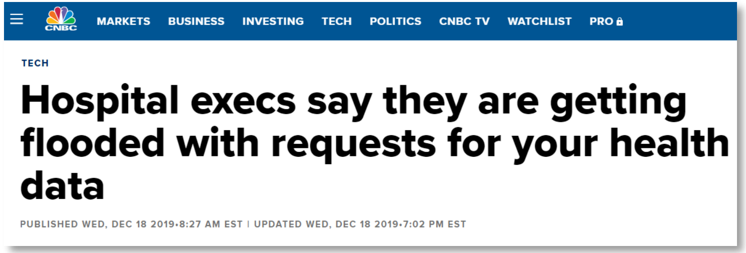
Progress in imaging, vision

The story is more complex for healthcare data...

- Unlike ImageNet, clinical data is about *people and their health*: ethical considerations!
- One dataset is not enough: medicine is complex; many diseases, different types of data (images, genetic, tabular, longitudinal, time-series, …), diverse populations, different uses, patient characteristics and treatments change over time…etc.

- One of the few initiatives: MIMIC dataset from Beth Israel Deaconess Medical center…

- Companies/organizations are trying to lock up access to data to productize their models!



TECH

Hospital execs say they are getting flooded with requests for your health data

PUBLISHED WED, DEC 18 2019·8:27 AM EST | UPDATED WED, DEC 18 2019·7:02 PM EST

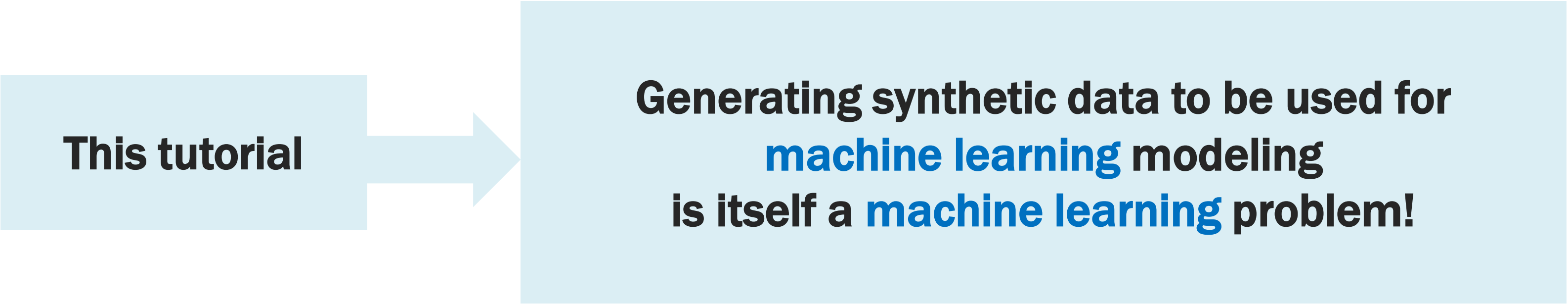
[1] A. E. Johnson et al, Scientific data, 2016

De-identified data vs Synthetic data

- De-identified/anonymized data: real data with all personal identifiers removed/data fields scrambled

- Synthetic data: data created from scratch, cannot be synced back to any individual (if modeled properly)

Requires ML/statistical modelling!

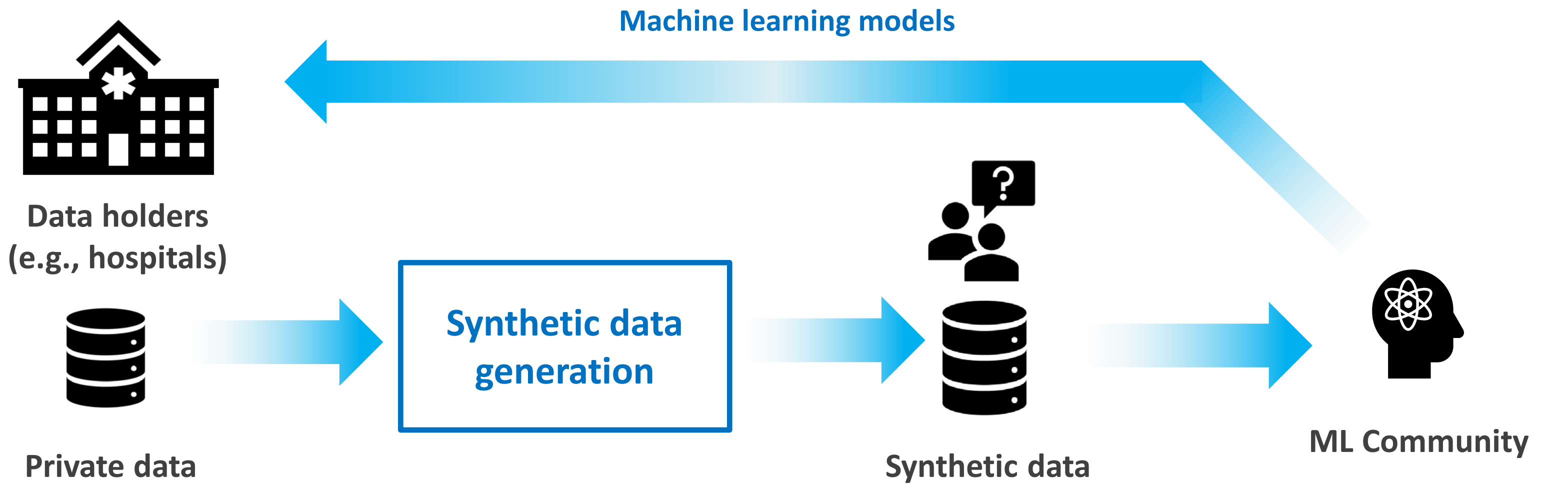


This tutorial

Generating synthetic data to be used for machine learning modeling is itself a machine learning problem!

How can synthetic data help?

- Share a synthetic (proximal) version of the data that resembles real data but contains no real samples for any specific individual…



Use cases for synthetic data

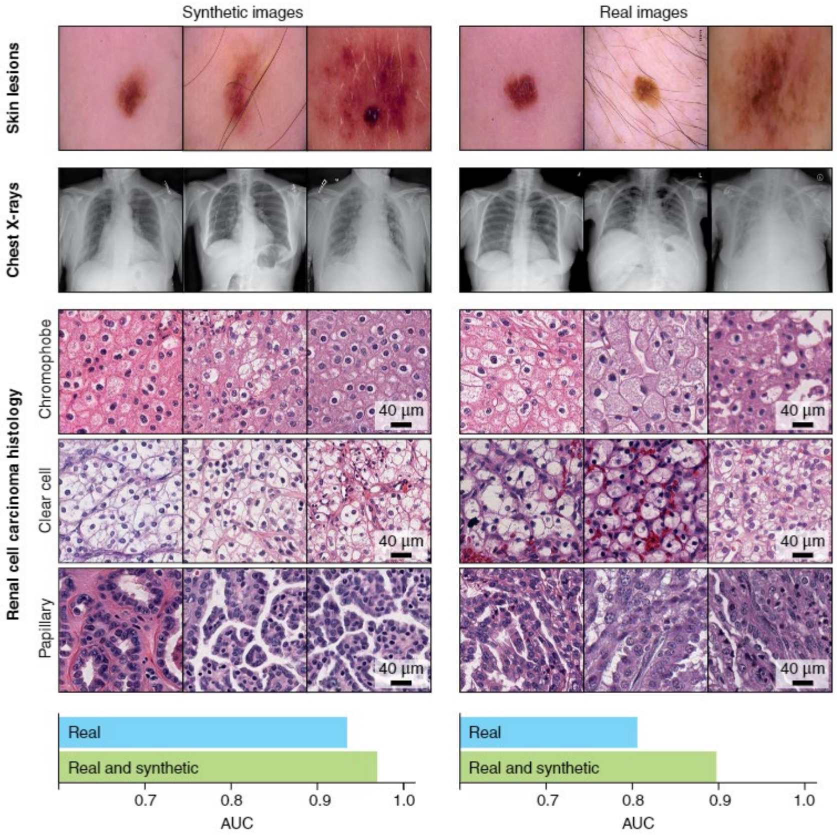
- Developing analytics
- Facilitating reproducibility of clinical studies and analyses
- Augmenting small-sample data sets:
  - Data for rare diseases
  - Data for underrepresented patient subgroups (to guard against model bias)
- Increasing robustness and adaptability of ML models (transferring across hospitals)
- Simulating forward-looking data

Synthetic clinical data in action: biomedical imaging

- Synthesizing skin lesion, chest x-rays and histology images using GANs[3]

- A classifier trained on real + synthetic data improved predictions of a rare subtype of renal cell carcinomas

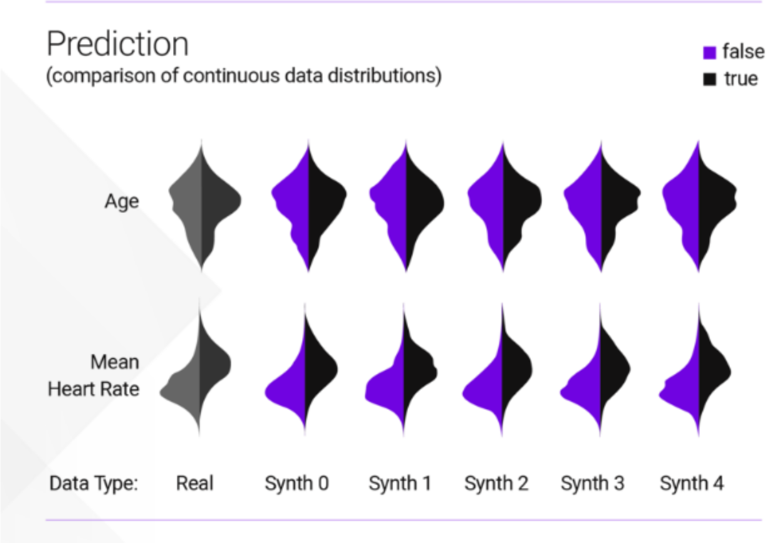
- Even with access to real data, synthetic data can still be used to augment training data.

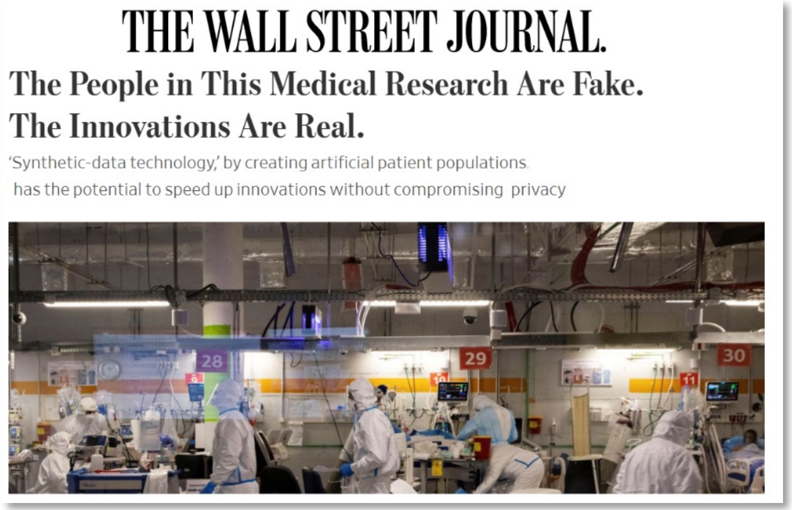


[3] R. J. Chen et al, Nature Biomedical Engineering, 2021

Synthetic clinical data in action: COVID-19

- **MDClone:** a synthetic data generation platform used by Sheba Medical Center, Maccabi Institute to accelerate time-to-insight for COVID-19





THE WALL STREET JOURNAL.

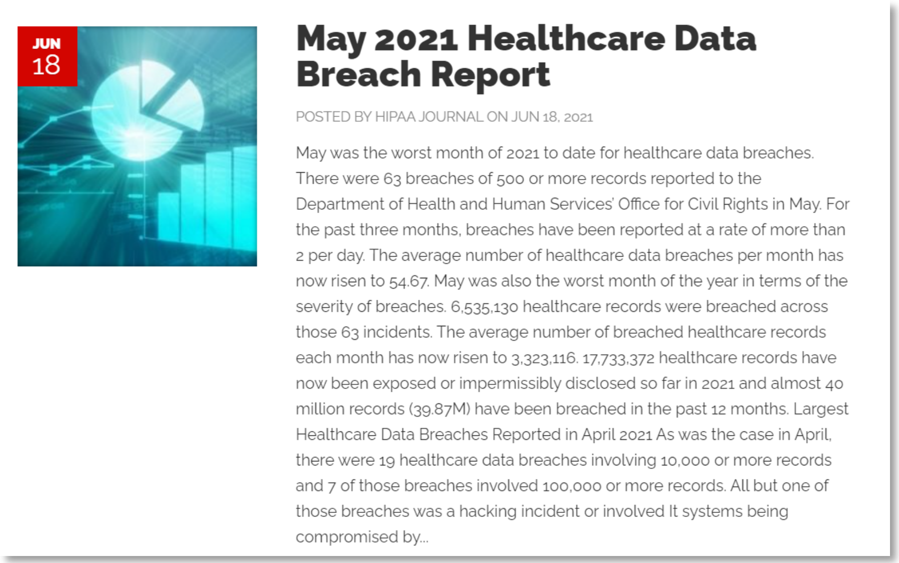
The People in This Medical Research Are Fake. The Innovations Are Real.

'Synthetic-data technology,' by creating artificial patient populations has the potential to speed up innovations without compromising privacy

[4] MDClone: https://www.mdclone.com/

Privacy concerns are central to clinical data sharing

- **Irresponsible sharing of clinical data = legal consequences for researchers, privacy breaches for patients**

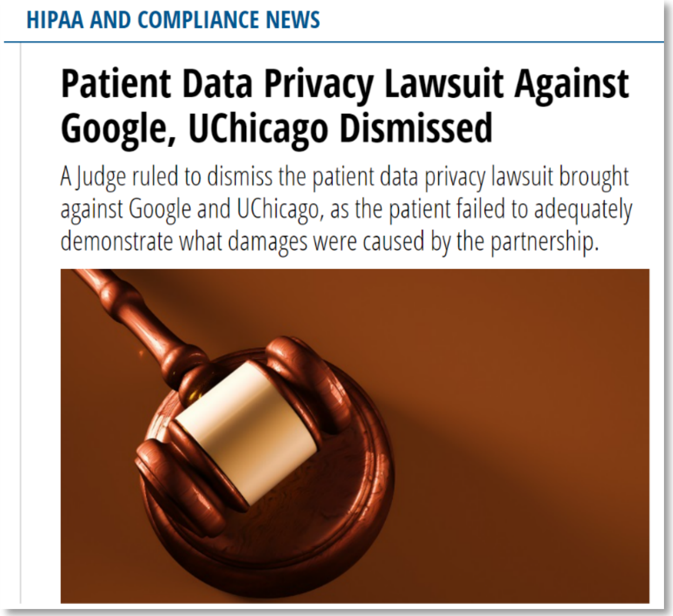


JUN 18

May 2021 Healthcare Data Breach Report

POSTED BY HIPAA JOURNAL ON JUN 18, 2021

May was the worst month of 2021 to date for healthcare data breaches. There were 63 breaches of 500 or more records reported to the Department of Health and Human Services' Office for Civil Rights in May. For the past three months, breaches have been reported at a rate of more than 2 per day. The average number of healthcare data breaches per month has now risen to 54.67. May was also the worst month of the year in terms of the severity of breaches. 6,535,130 healthcare records were breached across those 63 incidents. The average number of breached healthcare records each month has now risen to 3,323,116. 17,733,372 healthcare records have now been exposed or impermissibly disclosed so far in 2021 and almost 40 million records (39.87M) have been breached in the past 12 months. Largest Healthcare Data Breaches Reported in April 2021 As was the case in April, there were 19 healthcare data breaches involving 10,000 or more records and 7 of those breaches involved 100,000 or more records. All but one of those breaches was a hacking incident or involved It systems being compromised by...



HIPAA AND COMPLIANCE NEWS

Patient Data Privacy Lawsuit Against Google, UChicago Dismissed

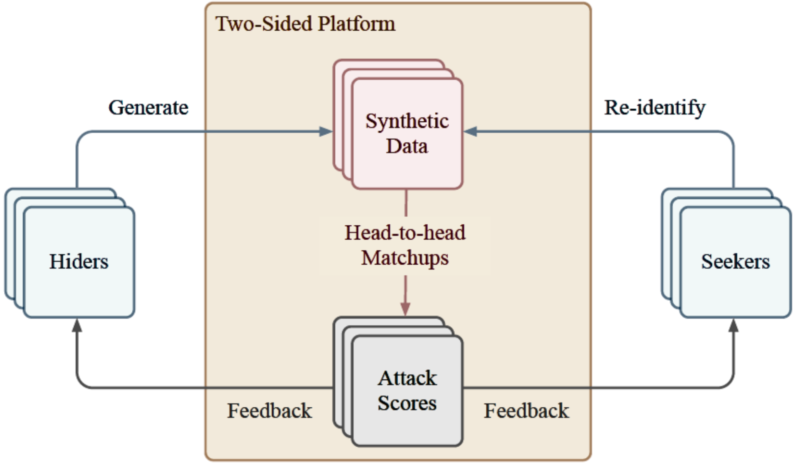
A Judge ruled to dismiss the patient data privacy lawsuit brought against Google and UChicago, as the patient failed to adequately demonstrate what damages were caused by the partnership.

[2] HIPAA Journal: https://www.hipaajournal.com/

Synthetic clinical data in action: Hide-and-seek competition

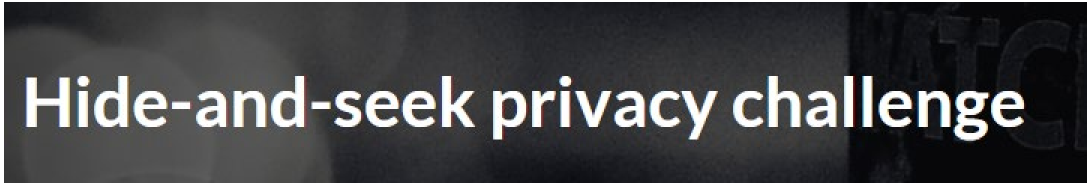
Hide-and-seek privacy challenge

https://www.vanderschaar-lab.com/privacy-challenge/



[5] J. Jordon and M. van der Schaar, 2020

Synthetic clinical data in action: Hide-and-seek competition



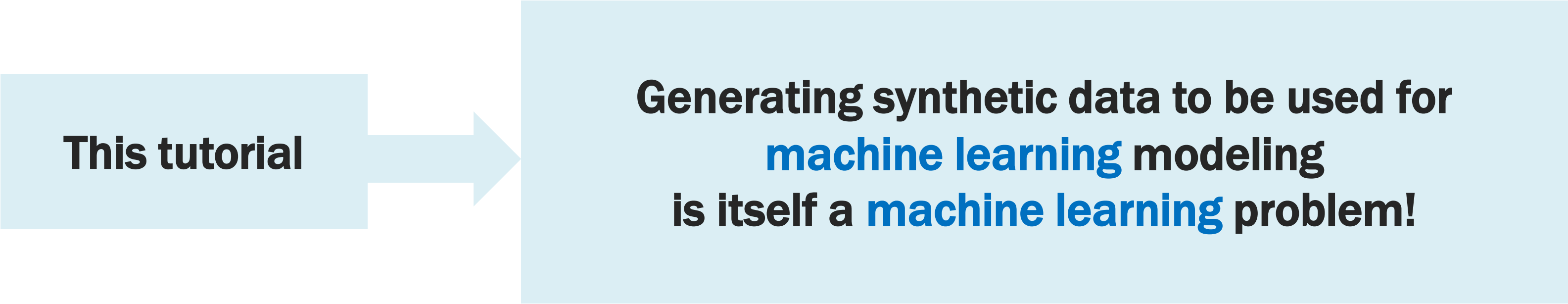
Hide-and-seek privacy challenge

https://www.vanderschaar-lab.com/privacy-challenge/

- **Organized by van der Schaar-lab, with support from Amsterdam UMC and Microsoft Research Cambridge**
  - Competitors are either “hiders” aiming to design a model for generating synthetic ICU data or “seekers” aiming to re-identify patients in the original data set.
  - Key lesson learned: defining the metrics for evaluating synthetic data and privacy preservation is a tricky problem!

Inspiration Exchange - synthetic data evaluation: www.youtube.com/watch?v=2FoazjRtTUI

Desiderata for synthetic data generation

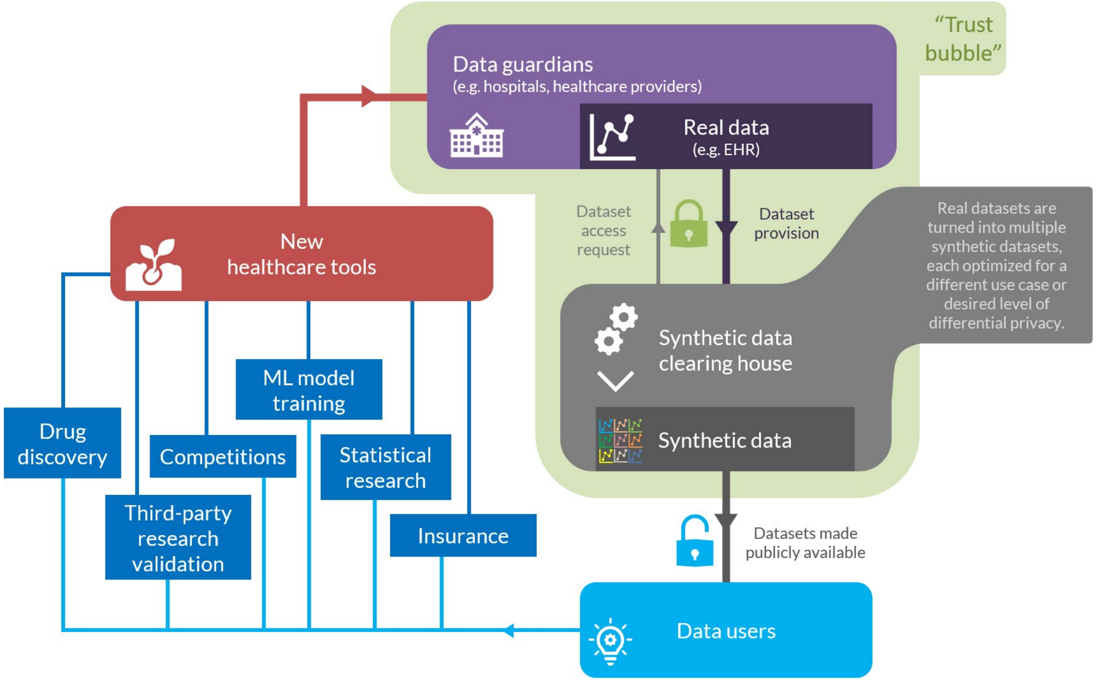


This tutorial

Generating synthetic data to be used for machine learning modeling is itself a machine learning problem!

- **Two key questions:**
  - How to build a machine learning model to generate synthetic data?
  - How to evaluate the quality of a synthetic dataset?

A clearinghouse for healthcare data resources

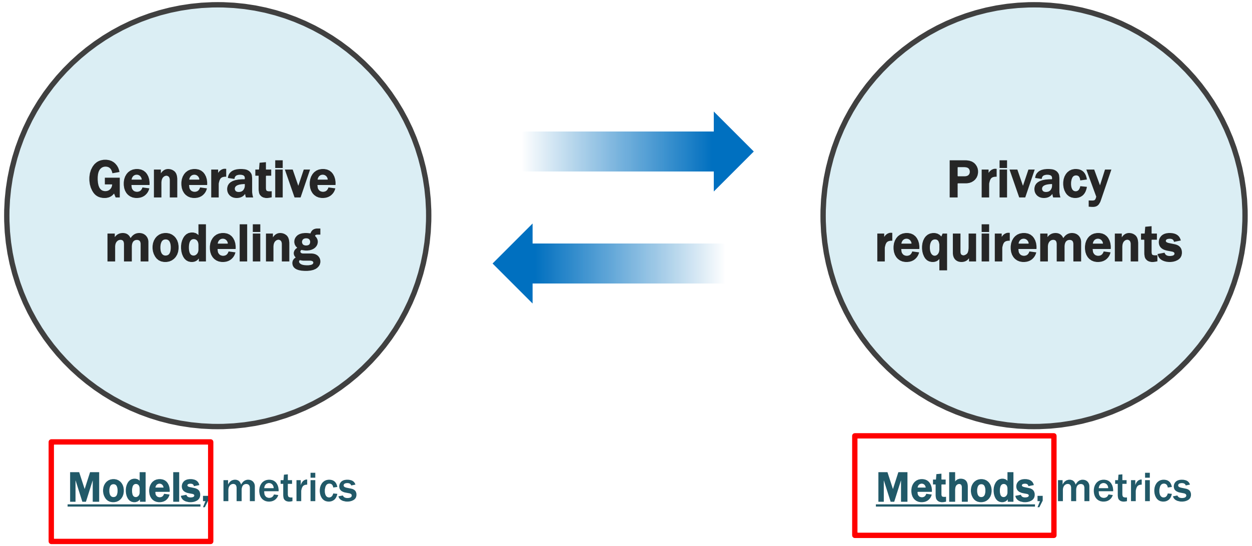

"Trust bubble"
Data guardians
(e.g. hospitals, healthcare providers)
Real data
(e.g. EHR)
Dataset access request
Dataset provision
Real datasets are turned into multiple synthetic datasets, each optimized for a different use case or desired level of differential privacy.
Synthetic data clearing house
Synthetic data
Datasets made publicly available
Data users
New healthcare tools
Drug discovery
Third-party research validation
Competitions
ML model training
Statistical research
Insurance

www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/

Synthetic generation of healthcare data: A brief technical background

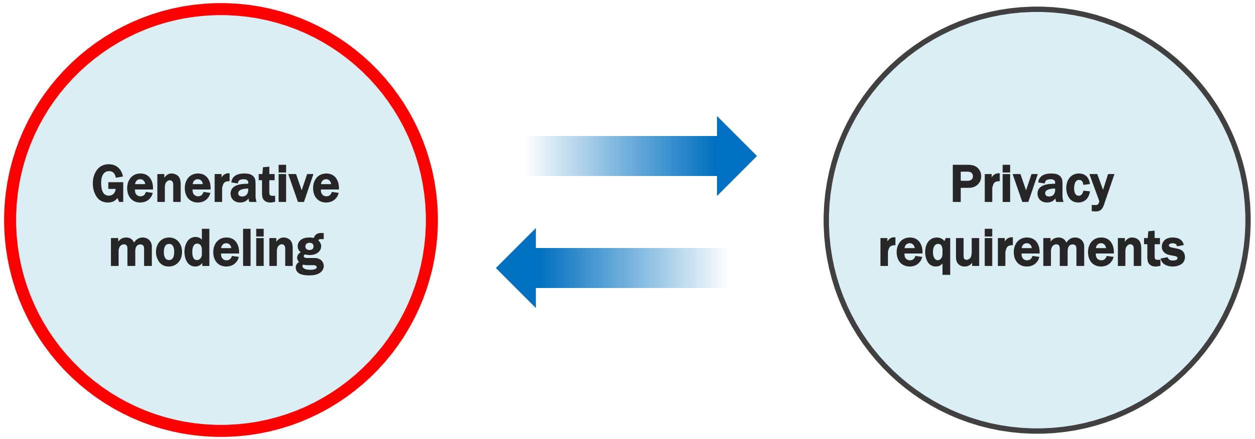
Desiderata for synthetic data generation

- We want synthetic data to enable learning statistical patterns without compromising the privacy of individual patients in the dataset



Desiderata for synthetic data generation

- We want synthetic data to enable learning statistical patterns without compromising the privacy of individual patients in the dataset



Models, metrics

Methods, metrics

Generative modeling

A model that simulates data

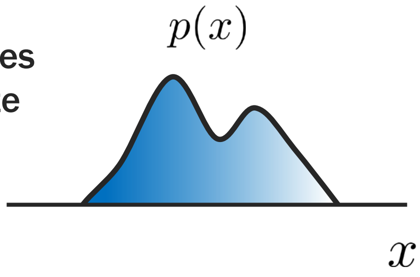
Estimates a (high-dimensional) probability density

Unsupervised

- Generative modeling

Sample feature instances from a density estimate

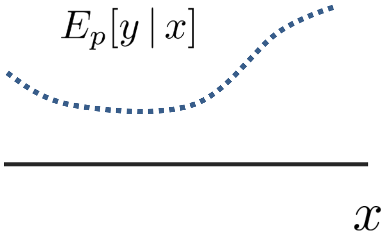
$$x \sim \hat{p}(x)$$



- Discriminative modeling

Given a feature, predict a target output

$$\hat{y} = E_{\hat{p}}[y \,|\, x]$$

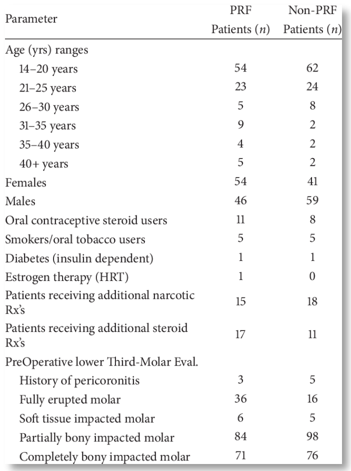


[6] Shakir Mohamed, UAI 2017

Generative modeling for clinical data

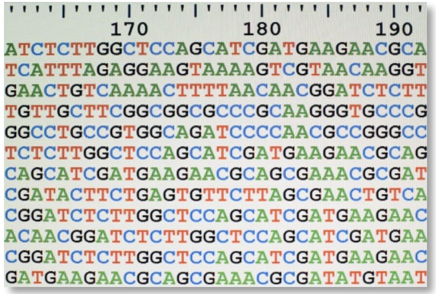
- **What makes generation of synthetic clinical data different?**
  - Complex and diverse data structures
  - Domain knowledge

Clinical variables



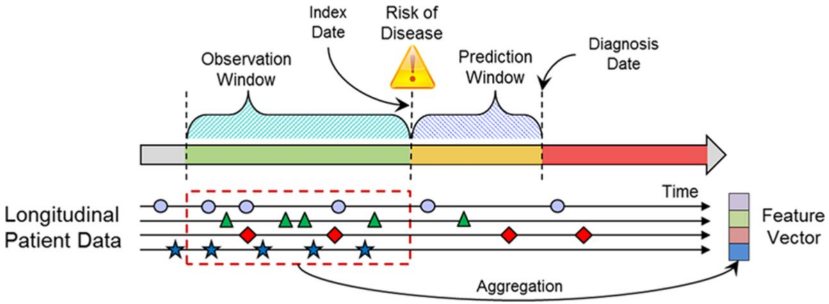
Parameter	PRF Patients (n)	Non-PRF Patients (n)
Age (yrs) ranges		
14–20 years	54	62
21–25 years	23	24
26–30 years	5	8
31–35 years	9	2
35–40 years	4	2
40+ years	5	2
Females	54	41
Males	46	59
Oral contraceptive steroid users	11	8
Smokers/oral tobacco users	5	5
Diabetes (insulin dependent)	1	1
Estrogen therapy (HRT)	1	0
Patients receiving additional narcotic Rx's	15	18
Patients receiving additional steroid Rx's	17	11
PreOperative lower Third-Molar Eval.		
History of pericoronitis	3	5
Fully erupted molar	36	16
Soft tissue impacted molar	6	5
Partially bony impacted molar	84	98
Completely bony impacted molar	71	76

Genetic data



```
        170             180             190
ATCTCTTGGCTCCAGCATCGATGAAGAACGCA
TCATTTAGAGGAAGTAAAAGTCGTAACAAGGT
GAACTGTCAAAACTTTTAACAACGGATCTCTT
TGTTGCTTCGGCGGCGCCCGCAAGGGTGCCCG
GGCCTGCCGTGGCAGATCCCCAACGCCGGGCC
TCTCTTGGCTCCAGCATCGATGAAGAACGCAG
CAGCATCGATGAAGAACGCAGCGAAACGCGAT
CGATACTTCTGAGTGTTCTTAGCGAACTGTCA
CGGATCTCTTGGCTCCAGCATCGATGAAGAAC
ACAACGGATCTCTTGGCTCCAGCATCGATGAA
CGGATCTCTTGGCTCCAGCATCGATGAAGAAC
GATGAAGAACGCAGCGAAACGCGATATGTAAT
```

Electronic health record data



The generative modeling problem

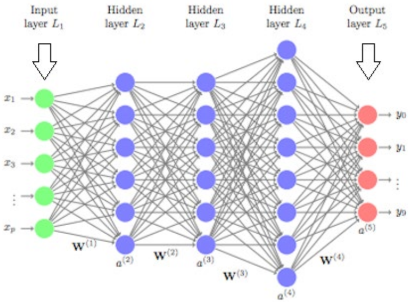
- **We have a data set of real patient data:** $\{X_1, \ldots, X_N\}$
- **A generative model is effectively an estimate** $\hat{P}(X)$ **of the real distribution** $P(X)$
- **A synthetic data is a sample from the generative model:** $\hat{X} \sim \hat{P}(X)$
- **We want the generative model to be:**
  - able to handle high-dimensional data
  - easy to train
  - easy to sample from

Two modeling approaches

Deep learning

- Rich non-linear models
- Scalable learning, end-to-end training
- Data-driven

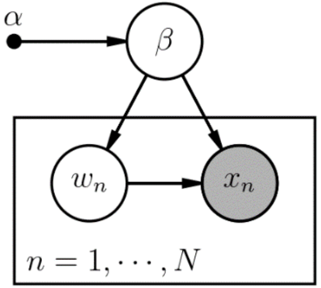
Used for learning complex function with the goal of making predictions



Probabilistic models

- Simpler, often linear models
- Principled inference, but interactable
- Encodes prior assumptions

Used for modeling data distributions

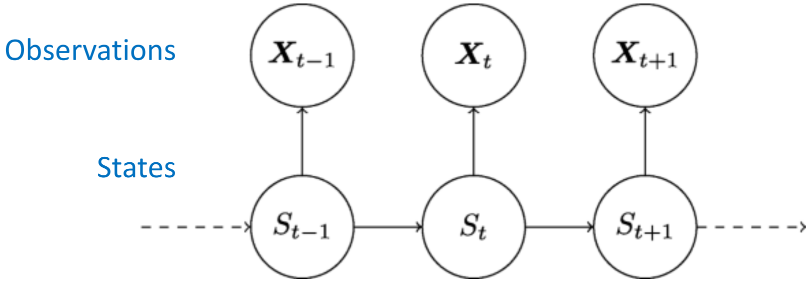


Probabilistic modeling

- **Use our intuition about the data to factorize the joint distribution of all variables**
- **Learn the model using maximum likelihood estimation**
- **Example: Hidden Markov models**
  - Data type: time-series, sequential
  - Intuition: each new observation only depends on the previous observation
  - Data distribution factorizes to:

$$P(X_1, \ldots, X_T) = \prod_{t=1}^{T} P(X_t \,|S_t) \cdot P(S_t \,|\, S_{t-1})$$

Patient health state = latent variable

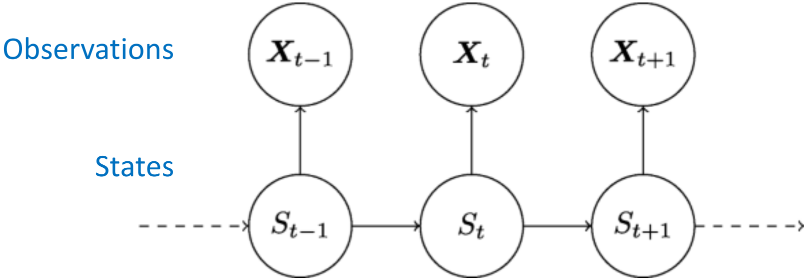


Probabilistic modeling: Learning

- Analytical derivation of the log-likelihood = depends on the model
- HMMs: no tractable algorithm for maximum likelihood estimation
- Expectation-Maximization (Baum-Welch) algorithm (approximate inference)
- HMMs are generative: we can sample synthetic time-series from them

$$S_t \sim P(S_t \mid S_{t-1})$$
$$X_t \sim P(X_t \mid S_t)$$

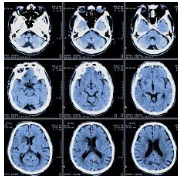
Patient health state = latent variable

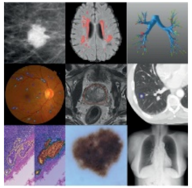


Why not just use probabilistic models?

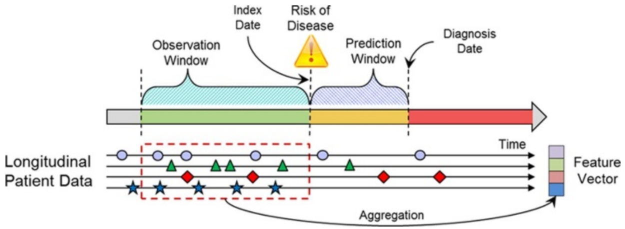
- **Strong assumption, limited representation complexity, inference routines need to be designed for each modeling scenario…**
  - Can we find a fixed factorization of joint pixel distribution for biomedical images?

$$P(X_1, \ldots, X_T)$$





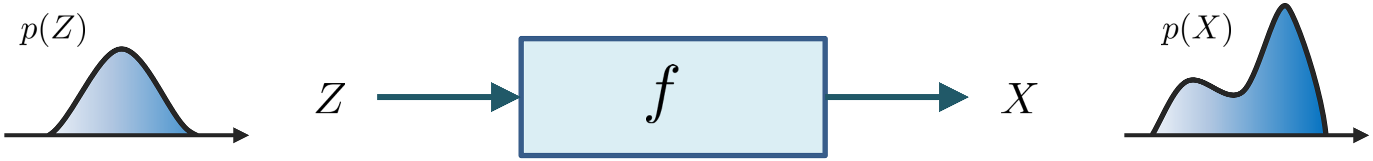
  - How can we model unstructured, irregularly sampled EHR data?



Categorization of modern (deep) generative models

- **Data-driven, SGD-based learning, complex representations**

- **Key idea: map a noise variable Z to a random variable X through a neural net f**

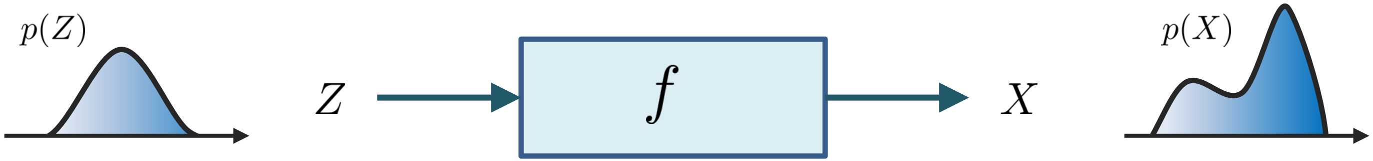


  - Different deep generative models specify a different loss function to learn f

  - Unlike standard NNs, these are models that we can **sample** data from

Categorization of modern (deep) generative models

- Data-driven, SGD-based learning, complex representations
- Key idea: map a noise variable Z to a random variable X through a neural net f



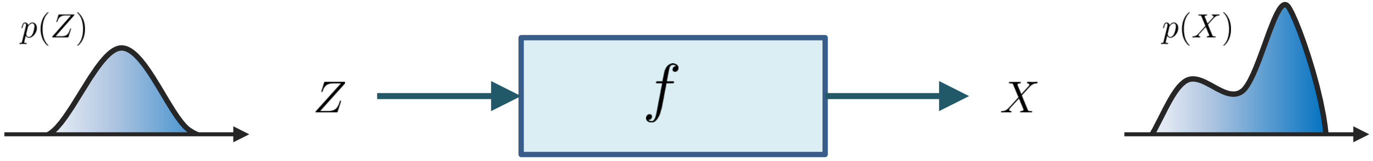
Implicit-likelihood models

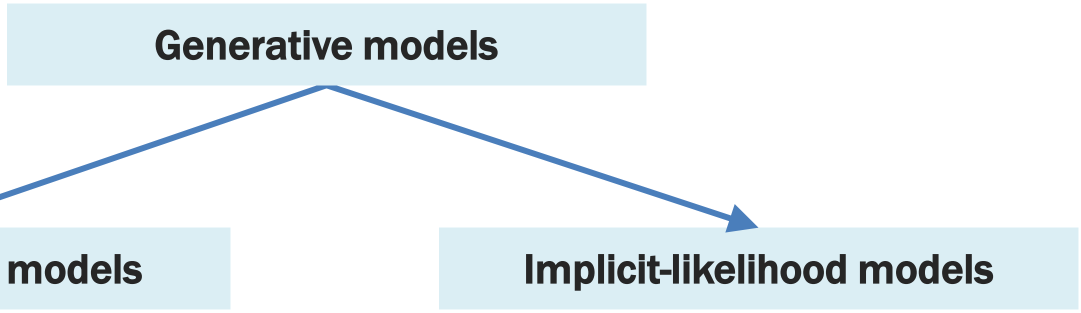
- A model that learns how to sample synthetic data without explicitly defining the likelihood function
- Learns by comparison

Explicit-likelihood models

- A model that defines the model likelihood and trains via maximum likelihood estimation
- Fits a distribution

Taxonomy of deep generative models





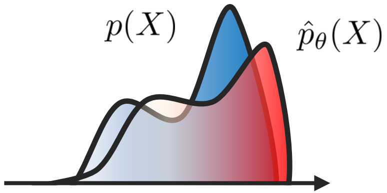
Generative models

Explicit-likelihood models

Implicit-likelihood models

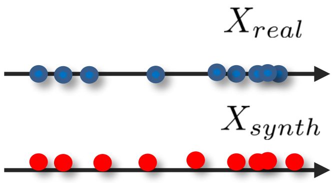
- Maximize likelihood

$$\hat{\theta} = \arg\max_{\theta} \log L(X_{real}; \theta)$$

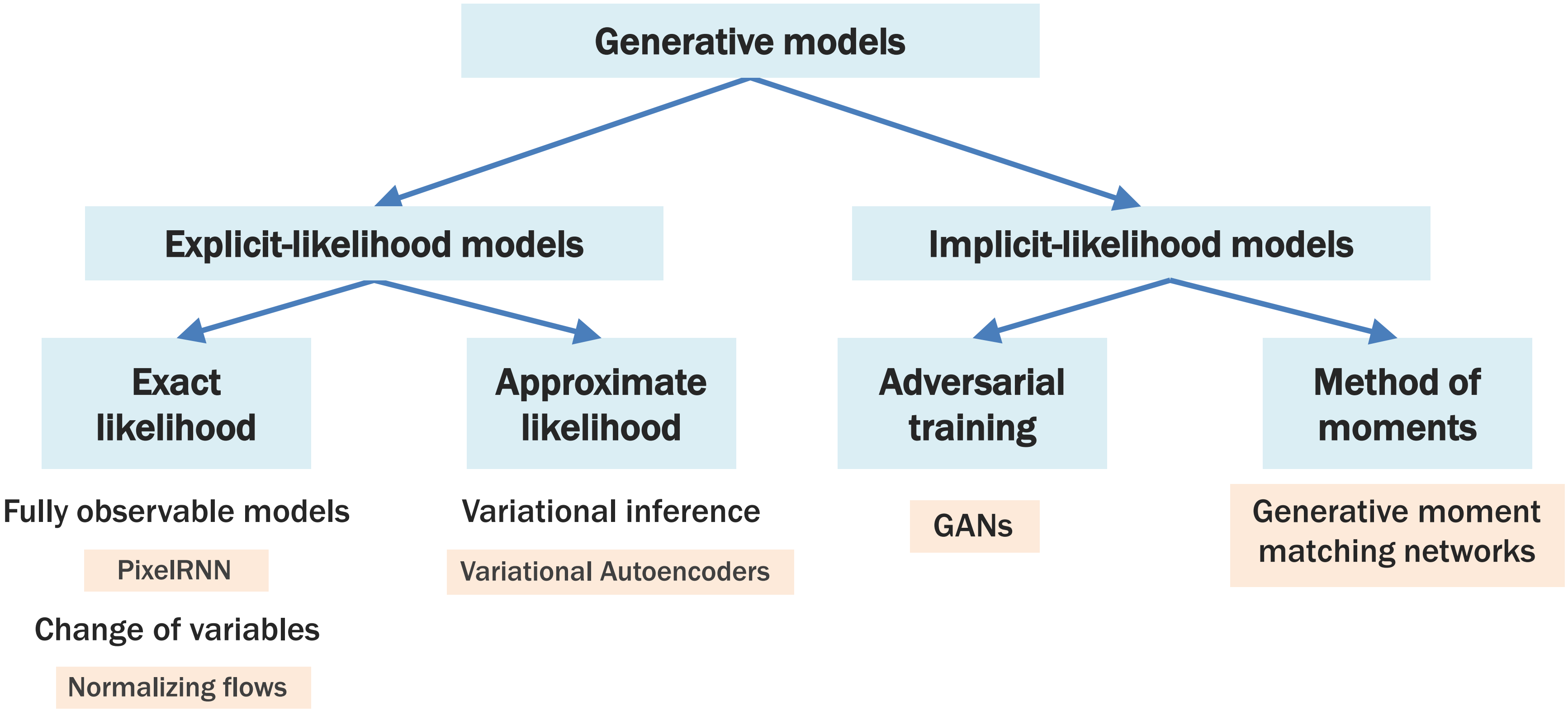


- Two-sample tests, sample comparison

$$L(X_{real}, X_{synth})$$



Taxonomy of deep generative models



Generative models

Explicit-likelihood models

Implicit-likelihood models

Exact likelihood

Approximate likelihood

Adversarial training

Method of moments

Fully observable models

PixelRNN

Change of variables

Normalizing flows

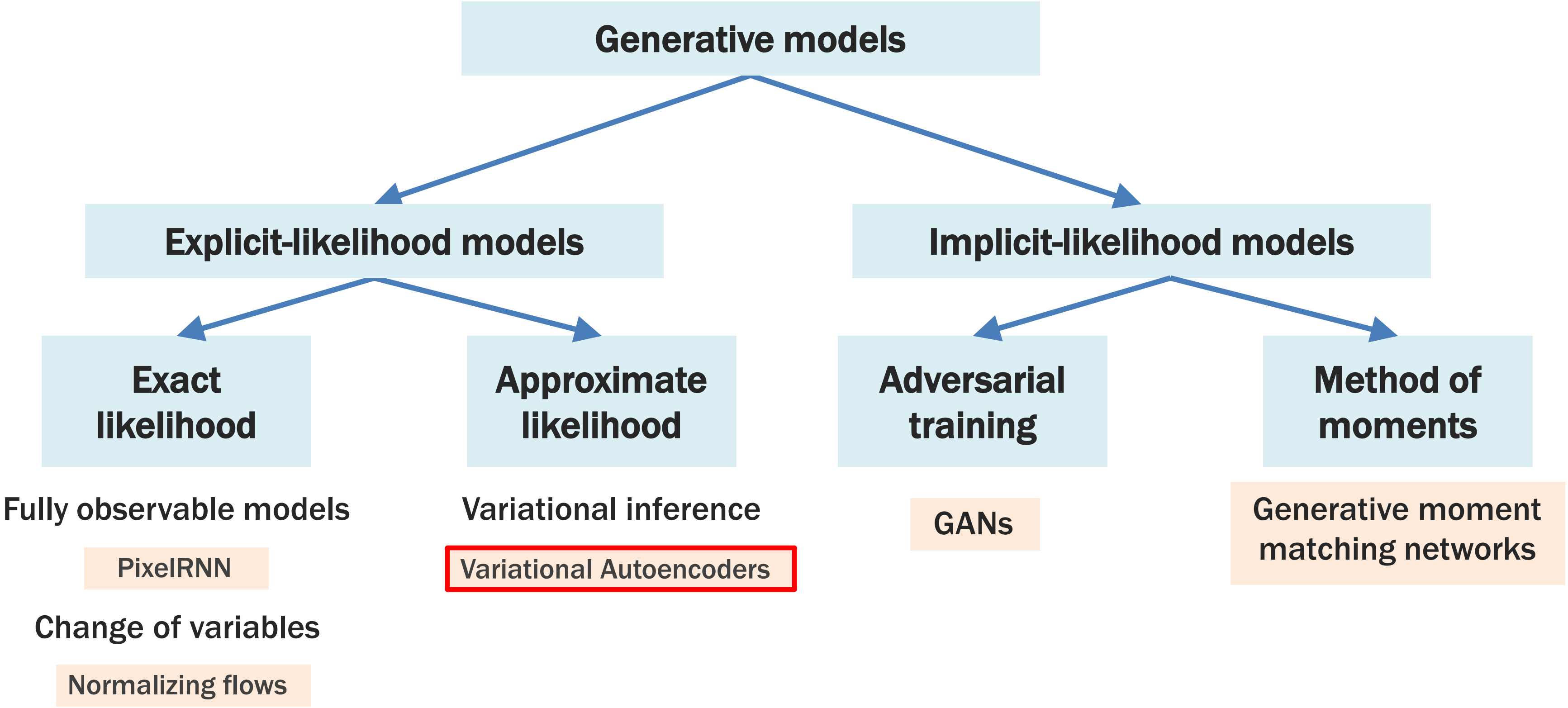
Variational inference

Variational Autoencoders

GANs

Generative moment matching networks

Taxonomy of deep generative models



Generative models

Explicit-likelihood models

Implicit-likelihood models

Exact likelihood

Approximate likelihood

Adversarial training

Method of moments

Fully observable models

PixelRNN

Change of variables

Normalizing flows

Variational inference

Variational Autoencoders

GANs

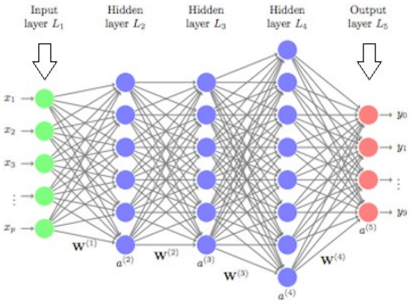
Generative moment matching networks

Explicit density models (I): Variational Autoencoders (VAEs)

- **VAEs:** probabilistic spinoff of Autoencoders, a latent variable model

Autoencoder

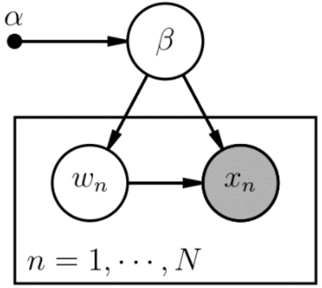
Deep learning



+

Variational inference

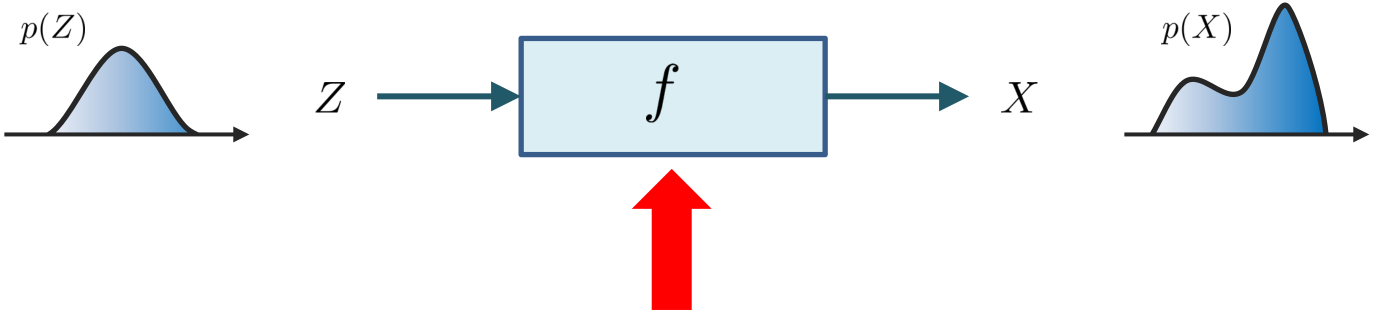
Probabilistic models



[7] Kingma and Welling, 2014, Rezende et al., 2014

Explicit density models (I): Variational Autoencoders (VAEs)

- **VAEs:** probabilistic spinoff of Autoencoders, a latent variable model



An autoencoder as a latent variable model

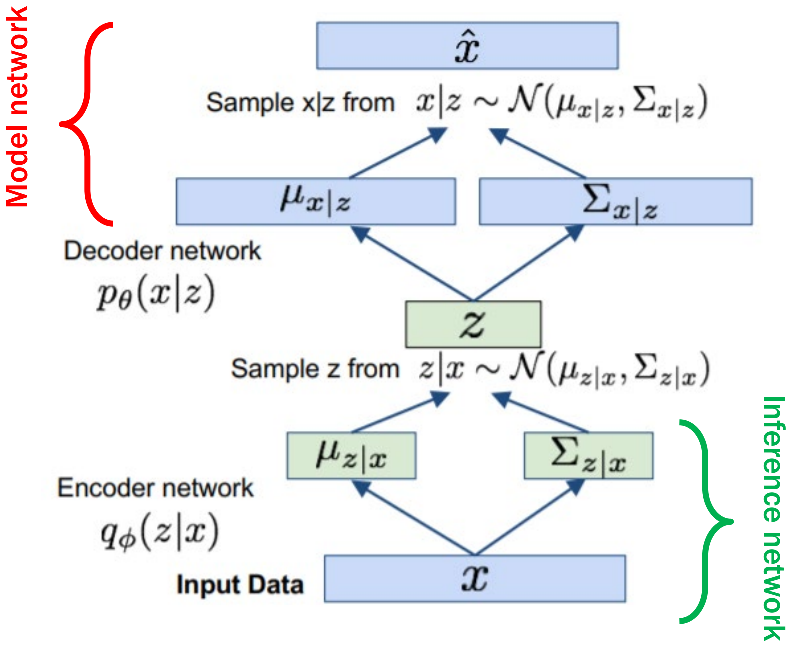
[7] Kingma and Welling, 2014, Rezende et al., 2014

Variational Autoencoders

- Latent variable model → marginalizing over latent variables → likelihood is intractable

- Optimize Evidence lower bound (ELBO) on model likelihood:

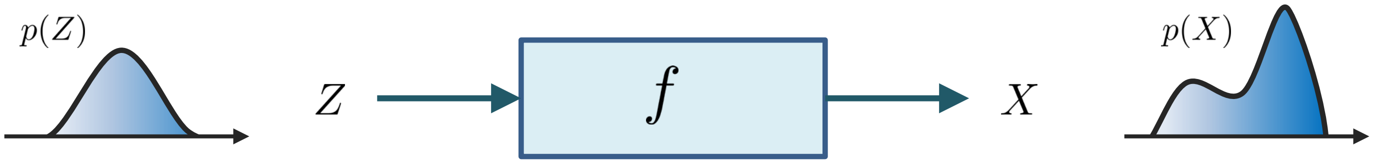
$$\underbrace{\overbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right]}^{\text{Model network}} - \overbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}^{\text{Inference network}}}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$



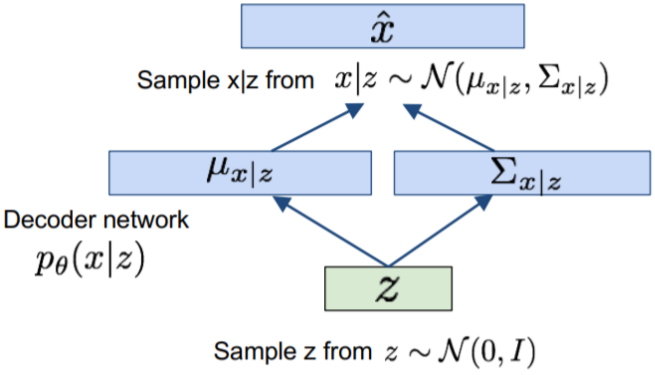
[8] Figure courtesy of: Dhruv Batra

Sampling from VAEs

- **Recall key idea:** map a noise variable Z to a random variable X through f



- Use decoder network and sample from prior…

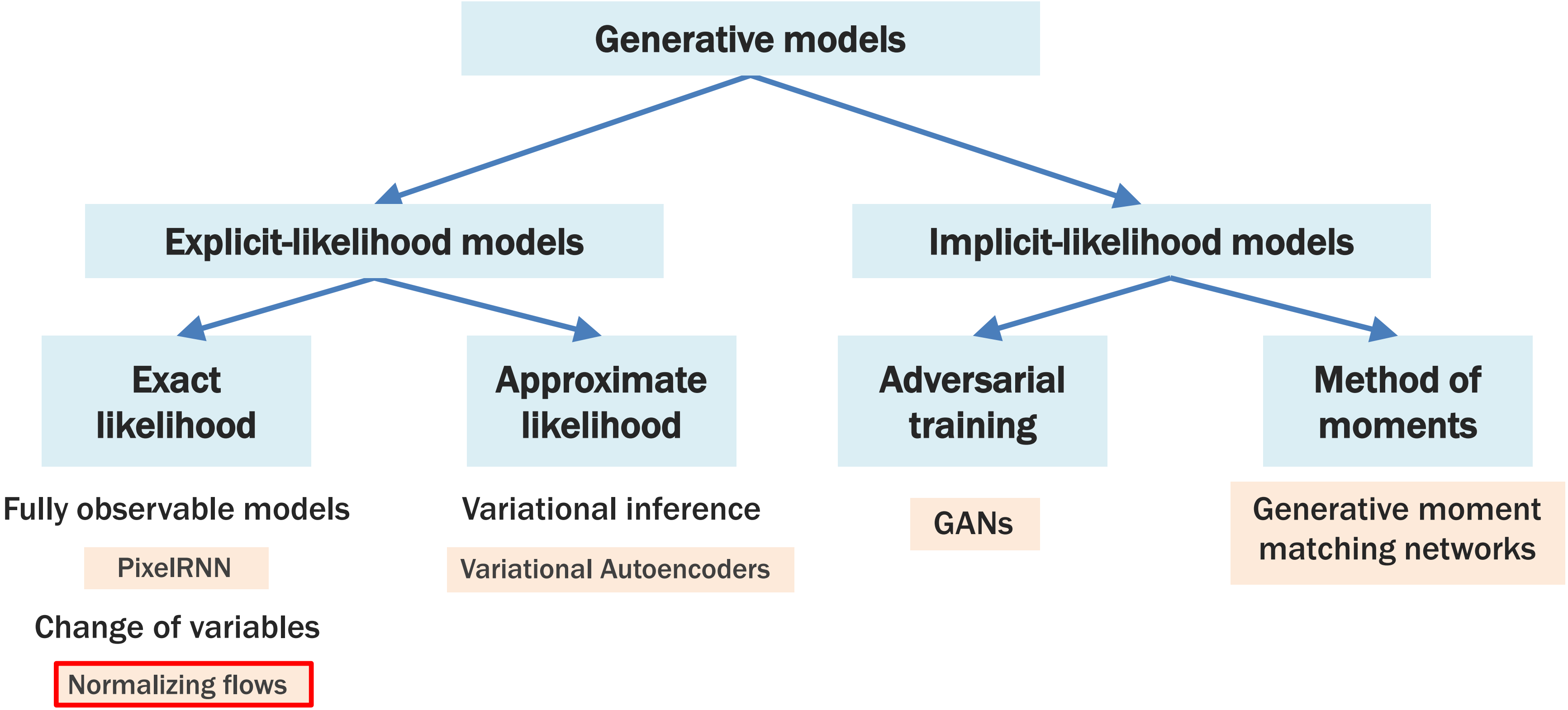


Sample x|z from $x|z \sim \mathcal{N}(\mu_{x|z}, \Sigma_{x|z})$

Decoder network $p_\theta(x|z)$

Sample z from $z \sim \mathcal{N}(0, I)$

Taxonomy of deep generative models



Generative models

Explicit-likelihood models

Implicit-likelihood models

Exact likelihood

Approximate likelihood

Adversarial training

Method of moments

Fully observable models

PixelRNN

Change of variables

Normalizing flows

Variational inference

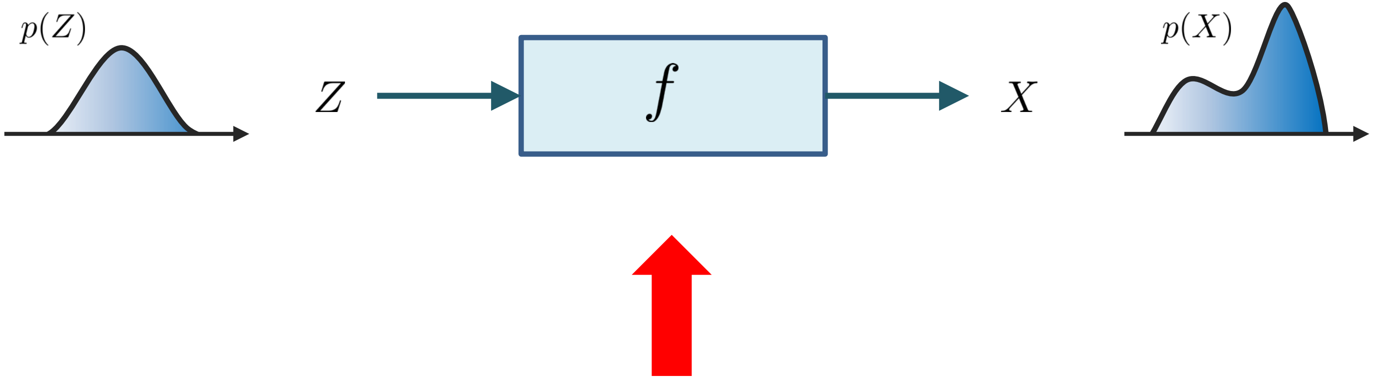
Variational Autoencoders

GANs

Generative moment matching networks

Explicit density models (II): Normalizing flows

- **Recall key idea:** map a noise variable Z to a random variable X through f

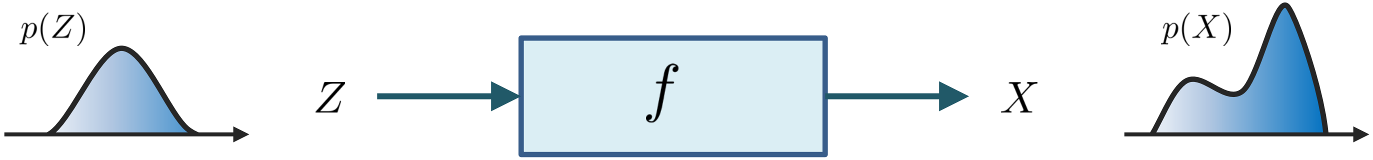


Use an invertible transform + change of variables formula

[9] D. Rezende and S. Mohamed, 2015

Normalizing Flows

- **Change of variables formula:**



- **Distribution of transformed variable:** $p_X(x) = p_Z(f^{-1}(x)) \left| \det \left(\frac{\partial f^{-1}(x)}{\partial x} \right) \right|$

- **For high-dimensional random variables:** $p_X(x) = p_Z(f^{-1}(x)) \underbrace{\left| \det \boldsymbol{J}[f] \right|}_{\text{Jacobean matrix}}$

Normalizing Flows

- **For high-dimensional random variables:** $p_X(x) = p_Z(f^{-1}(x))\,\overbrace{|\det \boldsymbol{J}[f]|}^{\text{Jacobean matrix}}$

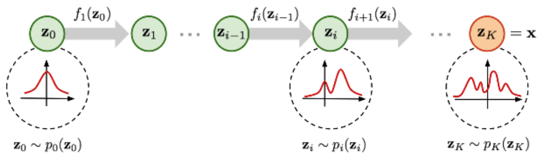
- **Flow composition:** $f^{(1)} \circ f^{(2)} \circ \ldots \circ f^{(M)}$

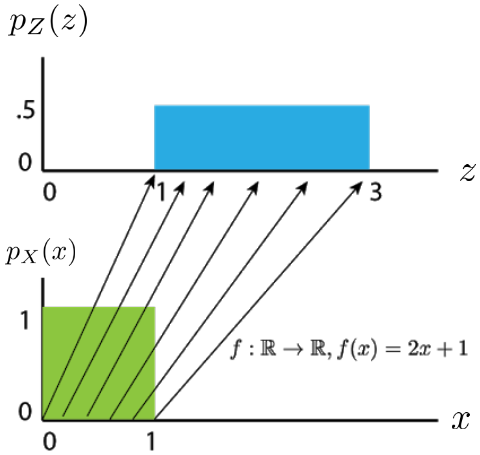
- **Need to <u>constrain</u> the transformation so that:**
  - The Jacobean determinant is easy to compute

    Needed for efficient training

  - The forward and inverse map are easy to compute

    Needed for efficient sampling

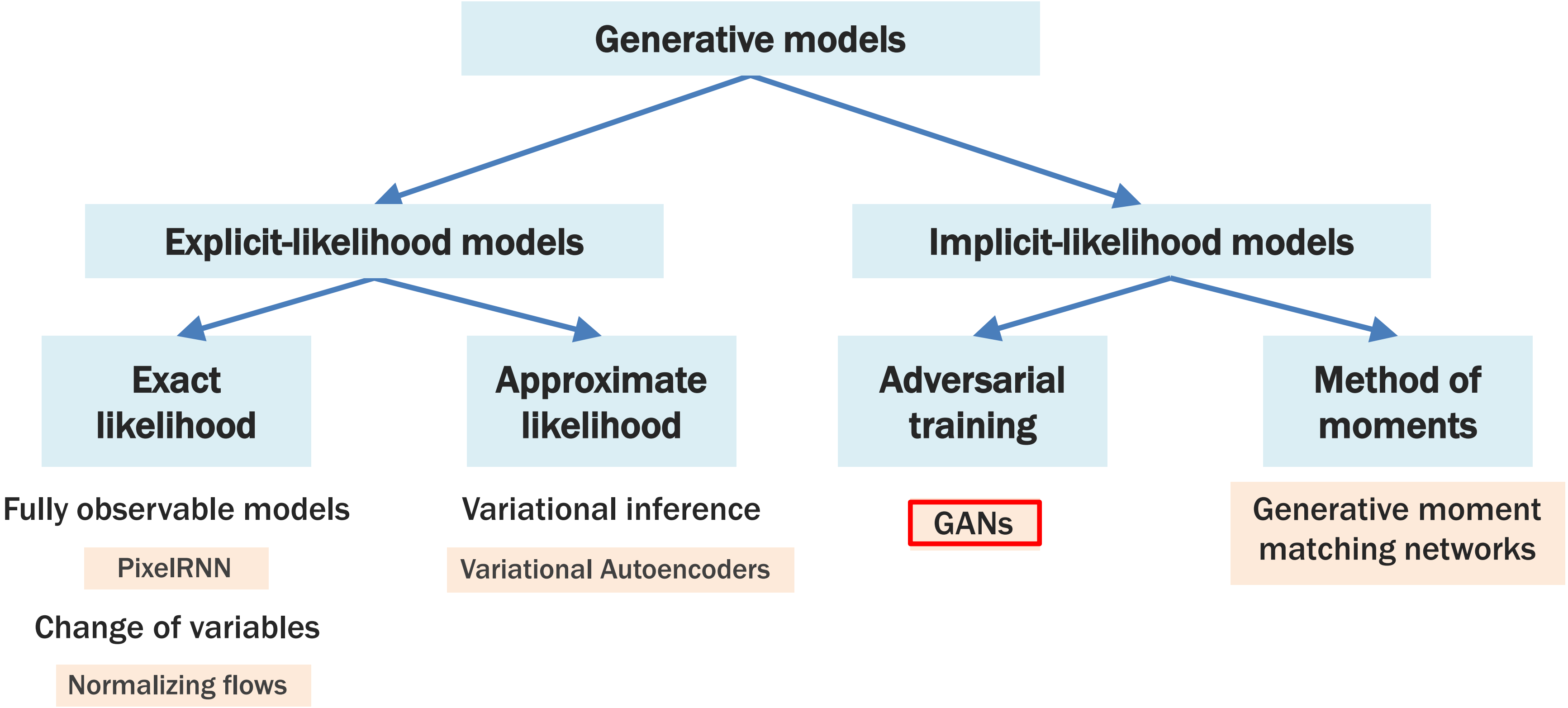




Key advantage: Exact likelihood for training and evaluation!

[10] Figure courtesy of: Eric Jang, 2019

Taxonomy of deep generative models



Generative models

Explicit-likelihood models

Implicit-likelihood models

Exact likelihood

Approximate likelihood

Adversarial training

Method of moments

Fully observable models

PixelRNN

Change of variables

Normalizing flows

Variational inference

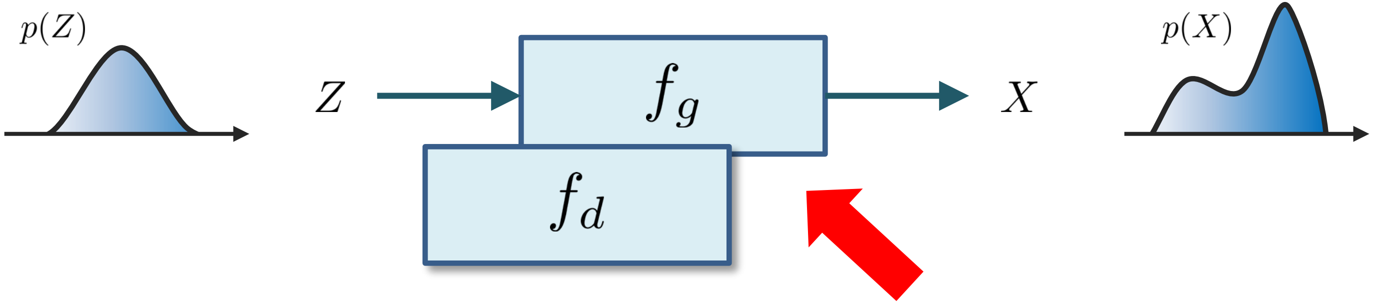
Variational Autoencoders

GANs

Generative moment matching networks

Implicit density models: Generative Adversarial Nets

- **Recall key idea:** **map a noise variable Z to a random variable X through f**



One player in a 2-player minimax game

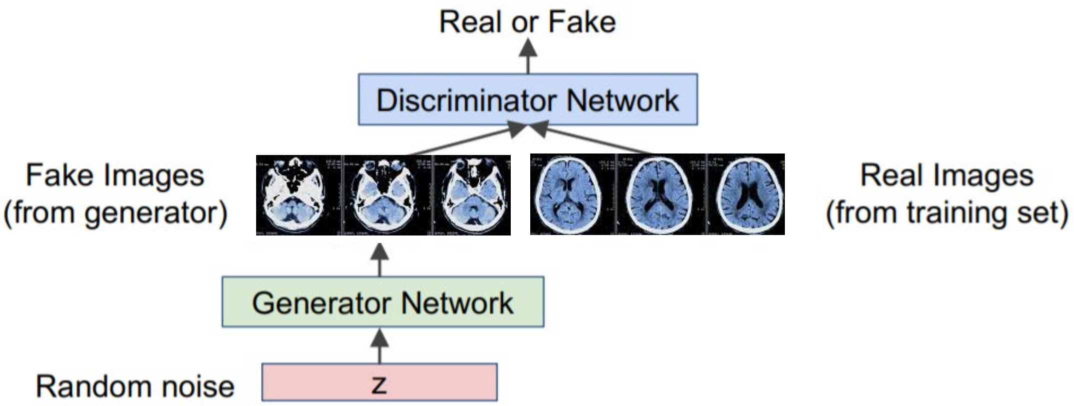
  - The transformation is optimized with the aid of a second player (discriminator network).

[11] I. J. Goodfellow et al, 2014

Generative Adversarial Nets

- **Minimax loss function**

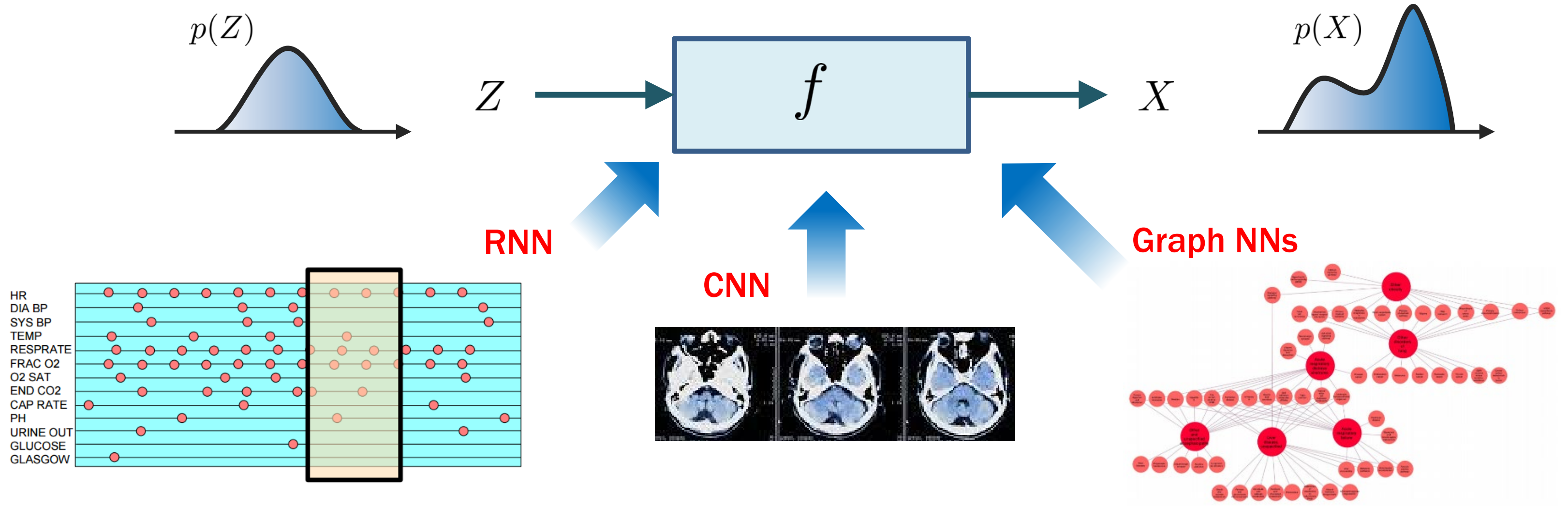
$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$



Adapting deep generative models to your healthcare data

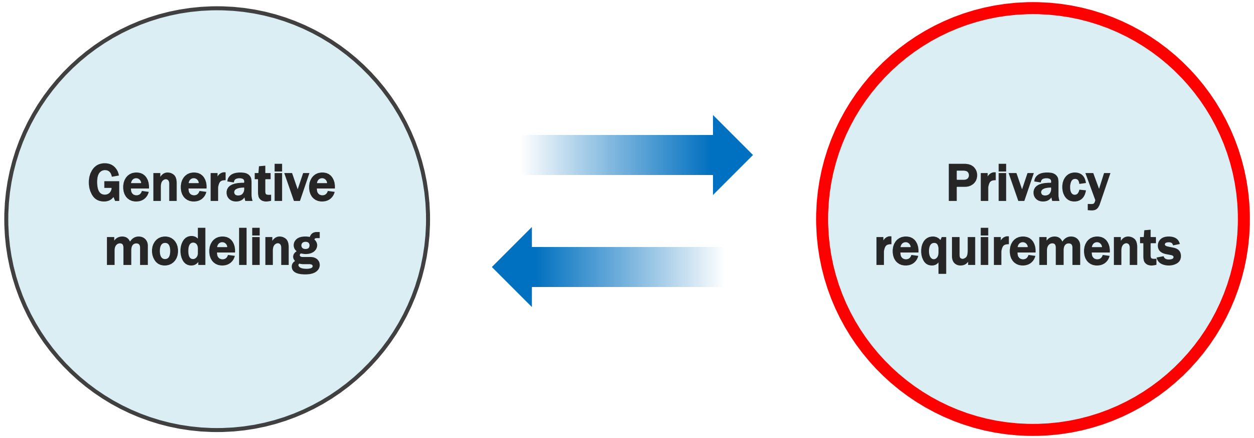
- In most cases, an off-the-shelf generative model will not be sufficient

- We need to make design choices…



Desiderata for synthetic data generation

- We want synthetic data to enable learning statistical patterns without compromising the privacy of individual patients in the dataset



Generative modeling

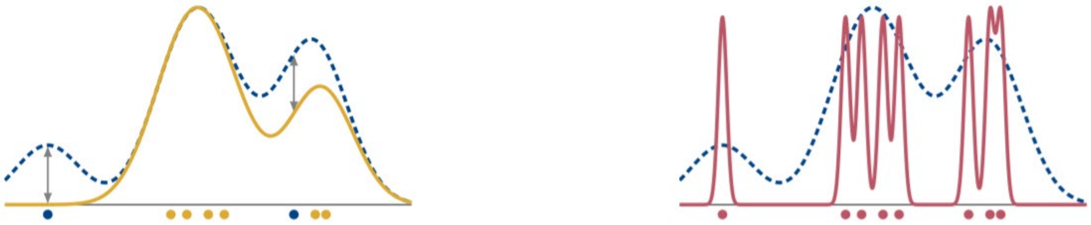
Privacy requirements

Models, metrics

Methods, metrics

Memorization and overfitting in generative models

- Synthetic data violates privacy if it copies real data
- Data copying happens when the generative model memorizes training data
- Is memorization the same thing as overfitting?



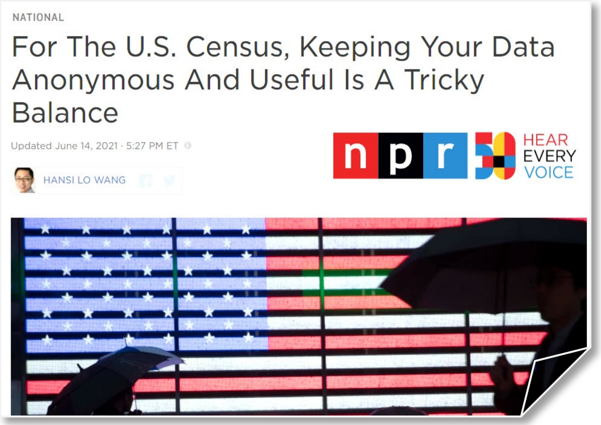
(a) Memorization

(b) Overfitting

[12] Figure is courtesy of: G. J. J. van den Burg and C. K.I. Williams, 2021

Statistical inference and privacy

- **More generally: we want synthetic data to enable learning statistical patterns without compromising the privacy of individual members of the dataset**
  - Common mistake: conflating inference with privacy violation

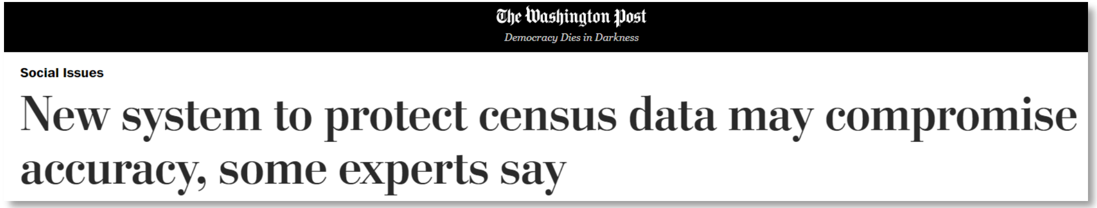


NATIONAL

For The U.S. Census, Keeping Your Data Anonymous And Useful Is A Tricky Balance

Updated June 14, 2021 · 5:27 PM ET

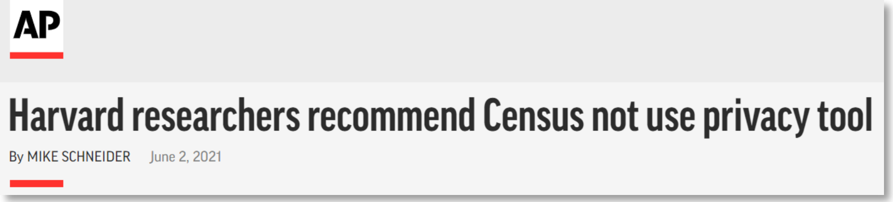
HANSI LO WANG



The Washington Post
Democracy Dies in Darkness

Social Issues

New system to protect census data may compromise accuracy, some experts say



AP

Harvard researchers recommend Census not use privacy tool

By MIKE SCHNEIDER June 2, 2021

[13] C. T. Kenny et al, 2021

Statistical inference is not a privacy violation

- **We want synthetic data to enable learning statistical patterns without compromising the privacy of individual members of the dataset**
  - Privacy in data sharing is a difficult concept to formalize…

Analysis on real data

Smoker → Cancer

Analysis on synthetic data

Mr. X is a Smoker → Mr. X has an elevated risk of cancer

[14] https://differentialprivacy.org/inference-is-not-a-privacy-violation/

Differential privacy

- **Definition**
  - A randomized mechanism $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$ is Ɛ-differentially private if for all neighbouring inputs and for all sets of outputs $E \subseteq \mathcal{Y}$ we have

$$\mathbb{P}(\mathcal{M}(x) \in E) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(x') \in E)$$

  - Intuition: the result of an analysis conducted using a data set would not change if any individual (synthetic) patient data point was removed from the data set.

[15] Dwork et al., 2006

Incorporating differential privacy into generative models

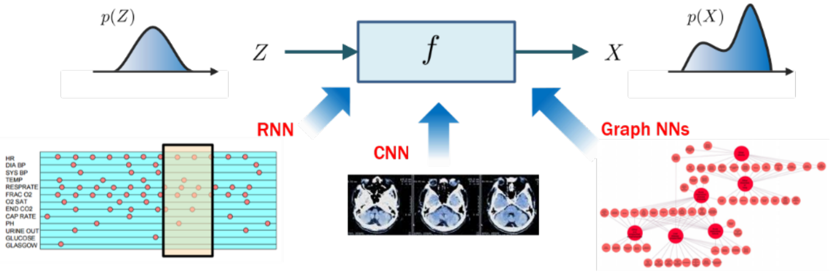
- **Differential privacy is incorporated by adding noise**
- **Three different approaches:**
  - Add noise to the model output
  - Add noise to the model parameters
  - Add noise to the model gradients during training

Recipe for generating synthetic healthcare data

- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.

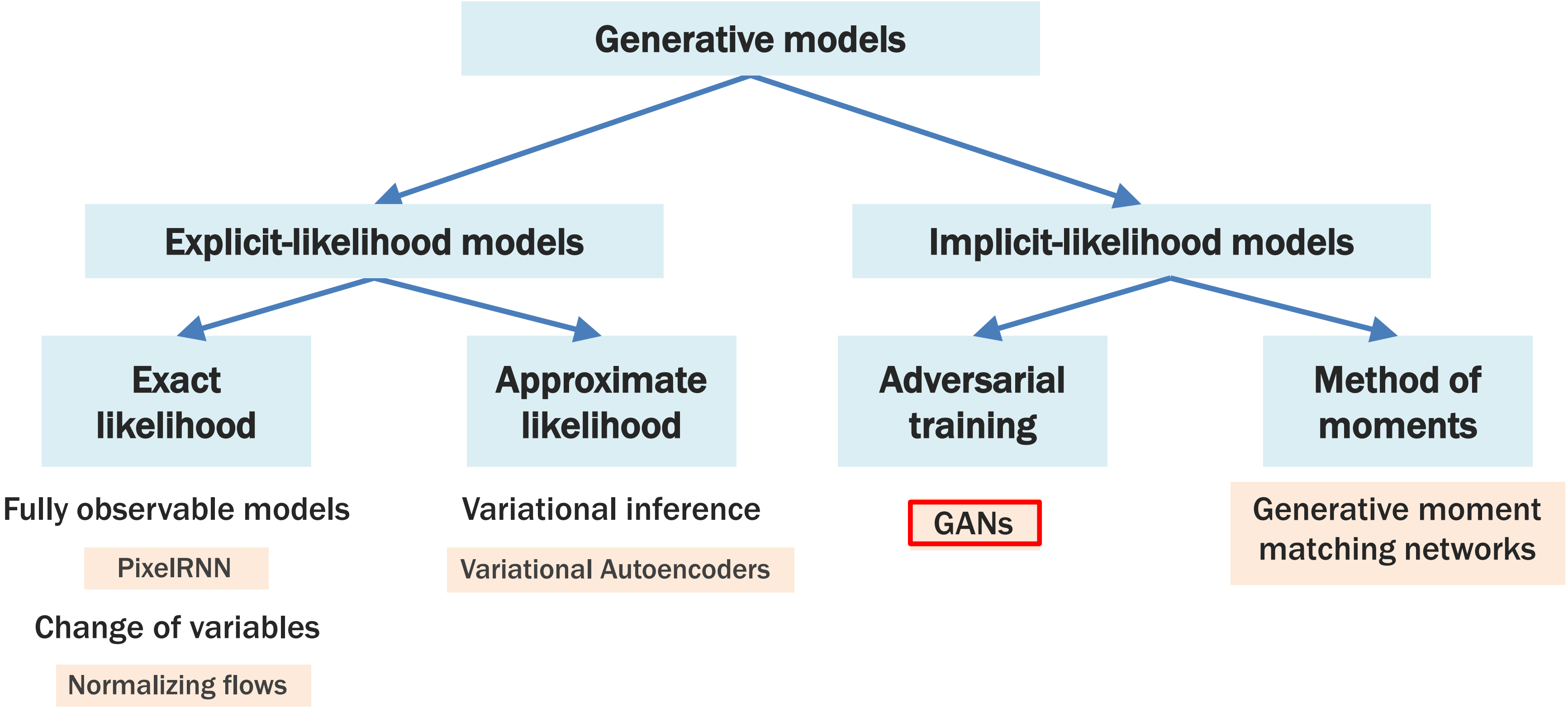
- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., NNs, RNN, CNN, transfer representations etc.

- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**



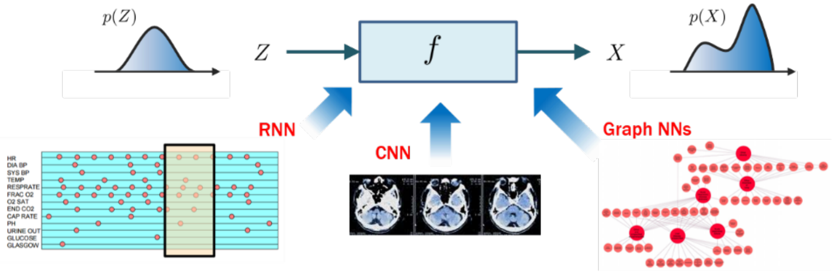
Generative models for synthetic tabular data

Taxonomy of deep generative models: Focus on GANs



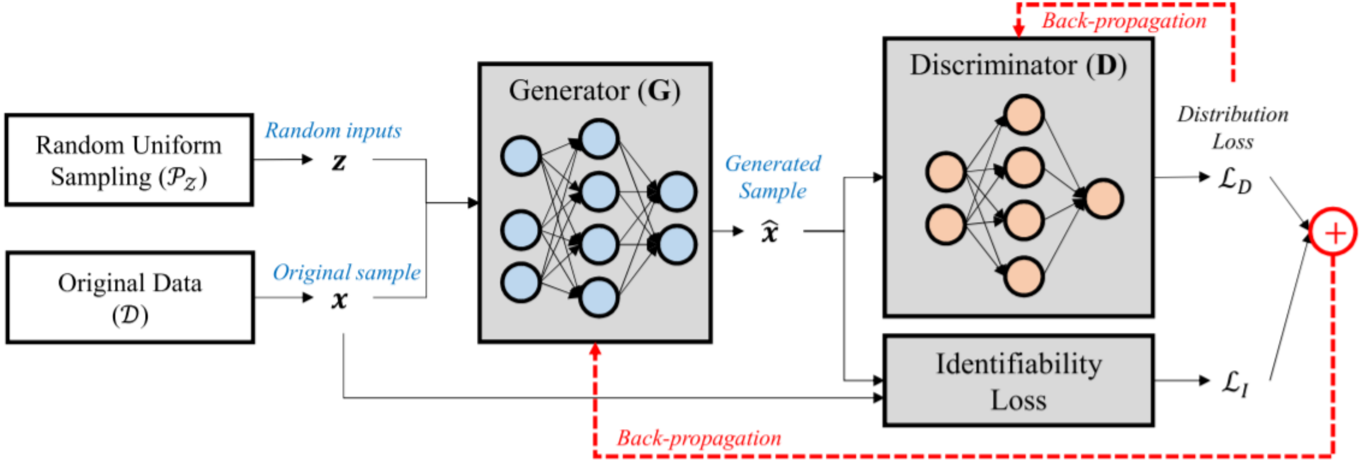
Recipe for generating synthetic healthcare data

- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.
- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., NNs, RNN, CNN, transfer representations etc.
- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**



ADS-GAN: Synthetic data fulfilling GDPR privacy requirements

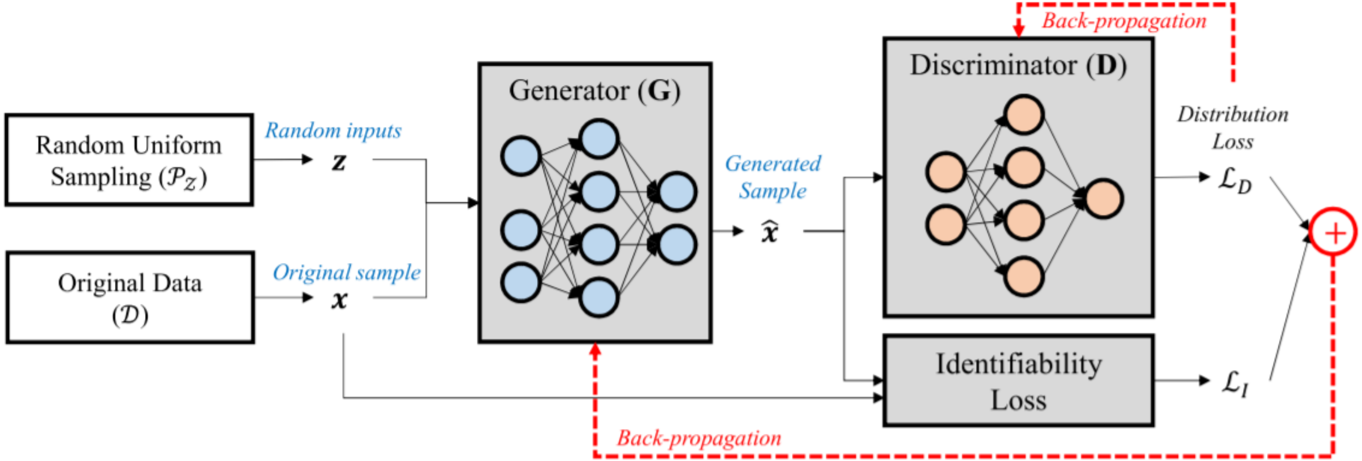
- Generating GDPR-compliant anonymized clinical data using GANs.
- Can be applied to any clinical datasets with tabular data (e.g. transplantation)



[16] J. Yoon, L. Drumright and M. van der Schaar, archive 2018, IEEE JHBI 2020.

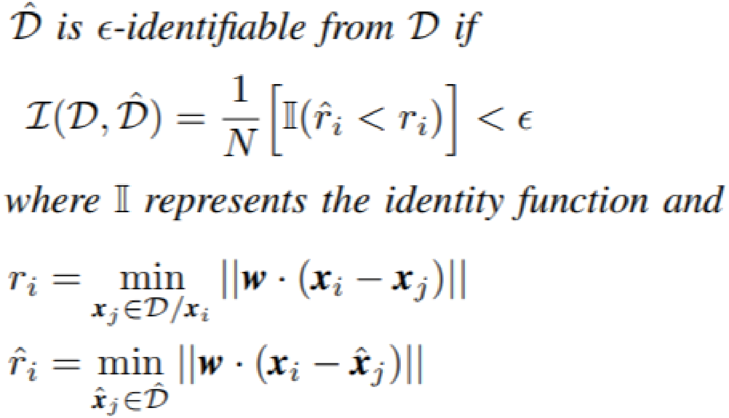
ADS-GAN: Synthetic data fulfilling GDPR privacy requirements

- Privacy
  - GDPR requirements are not formally/well-defined – descriptive, not formal
  - Formally defining requirements – introduce Identifiability Metric



ADS-GAN: Synthetic data fulfilling GDPR privacy requirements

- Privacy
  - GDPR requirements are not formally/well-defined – descriptive, not formal
  - Formally defining requirements – introduce Identifiability Metric



$\hat{\mathcal{D}}$ is ϵ-identifiable from $\mathcal{D}$ if

$$\mathcal{I}(\mathcal{D},\hat{\mathcal{D}}) = \frac{1}{N}\left[\mathbb{I}(\hat{r}_i < r_i)\right] < \epsilon$$

where $\mathbb{I}$ represents the identity function and

$$r_i = \min_{\boldsymbol{x}_j \in \mathcal{D}/\boldsymbol{x}_i} ||\boldsymbol{w} \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j)||$$

$$\hat{r}_i = \min_{\hat{\boldsymbol{x}}_j \in \hat{\mathcal{D}}} ||\boldsymbol{w} \cdot (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_j)||$$

ADS-GAN: Synthetic data fulfilling GDPR privacy requirements

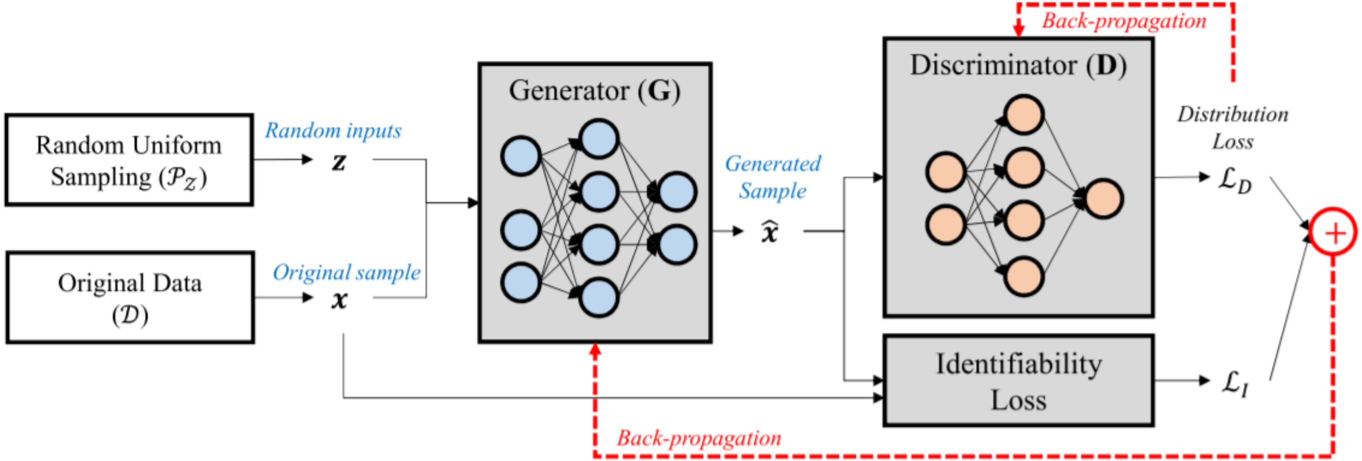
- Privacy
  - GDPR requirements are not formally/well-defined – descriptive, not formal
  - Formally defining requirements – introduce Identifiability Metric

- ε-identifiability = probability - that the distance to the closest synthetic observation is closer (“not different enough") than the distance from the closest real observation in D - is less than ε.

- ADS-GAN allows us to move away from differential privacy and define identifiability based on the probability of re-identification given the combination of all data on any individual patient.

ADS-GAN: Synthetic data fulfilling GDPR privacy requirements

- **Fidelity:** unlike conditional GANs (pre-determined variable set), ADS-GAN optimizes a conditioning set per patient and generates all components based on these



ADS-GAN: Synthetic data fulfilling GDPR privacy requirements

- **Fidelity:** unlike conditional GANs (pre-determined variable set), ADS-GAN optimizes a conditioning set per patient and generates all components based on these

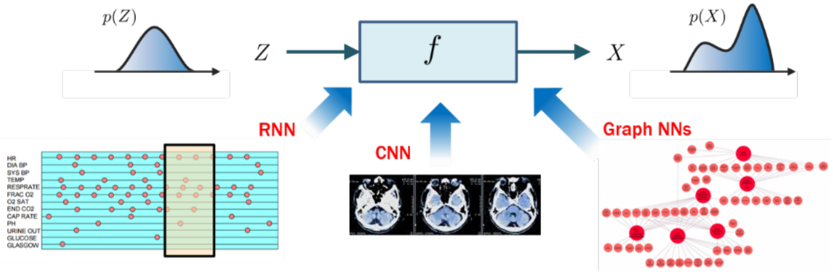
- **Advantages:**
  - Improve quality of synthetic data while ensuring that no feature combination could readily reveal a patient's identity.

Recipe for generating synthetic healthcare data

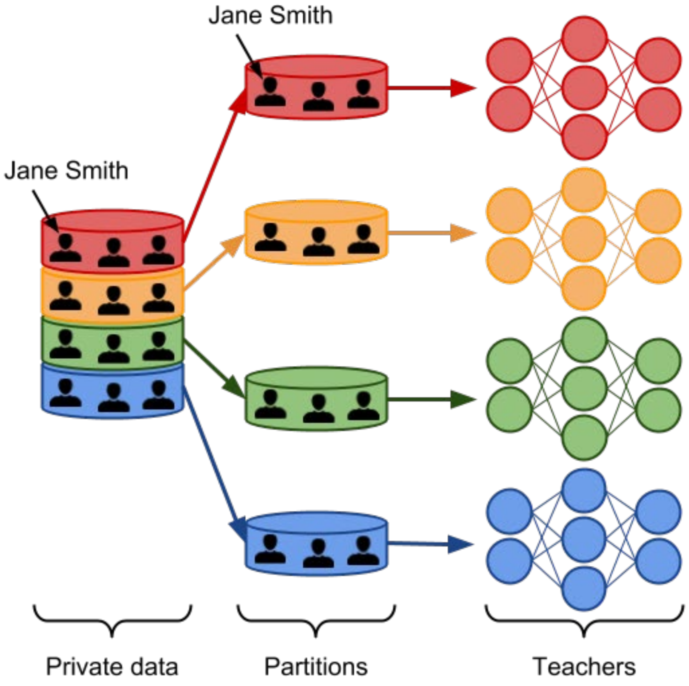
- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.

- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., NNs, RNN, CNN, transfer representations etc.

- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**

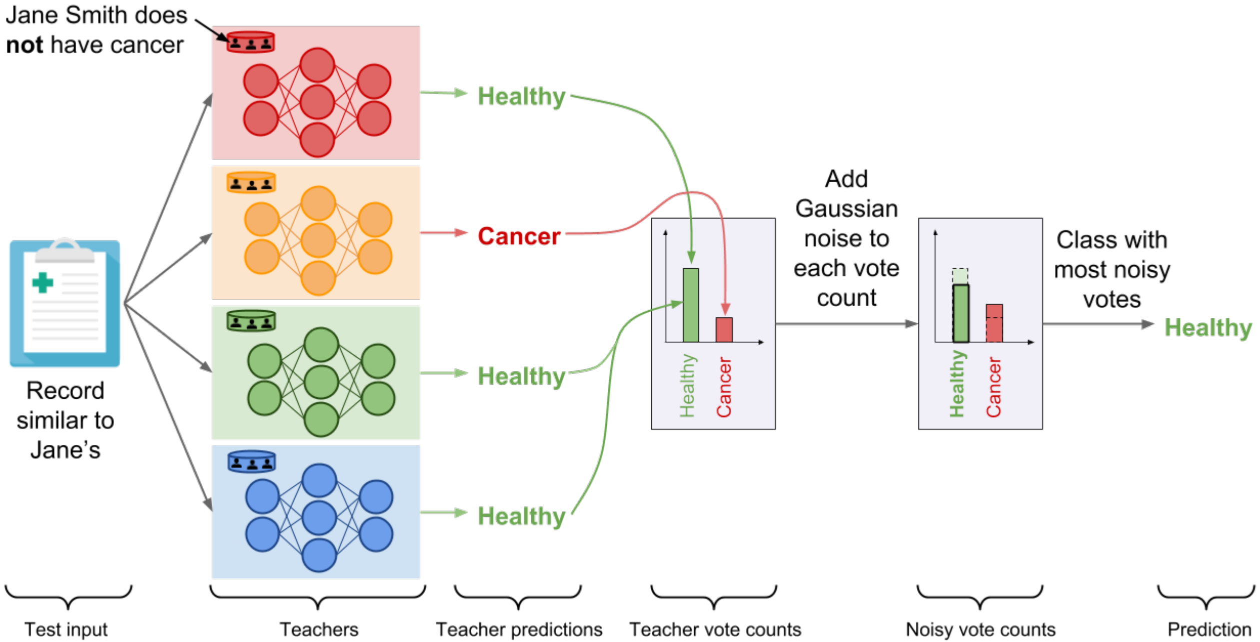


PATE: Private Aggregations of Teacher Ensembles

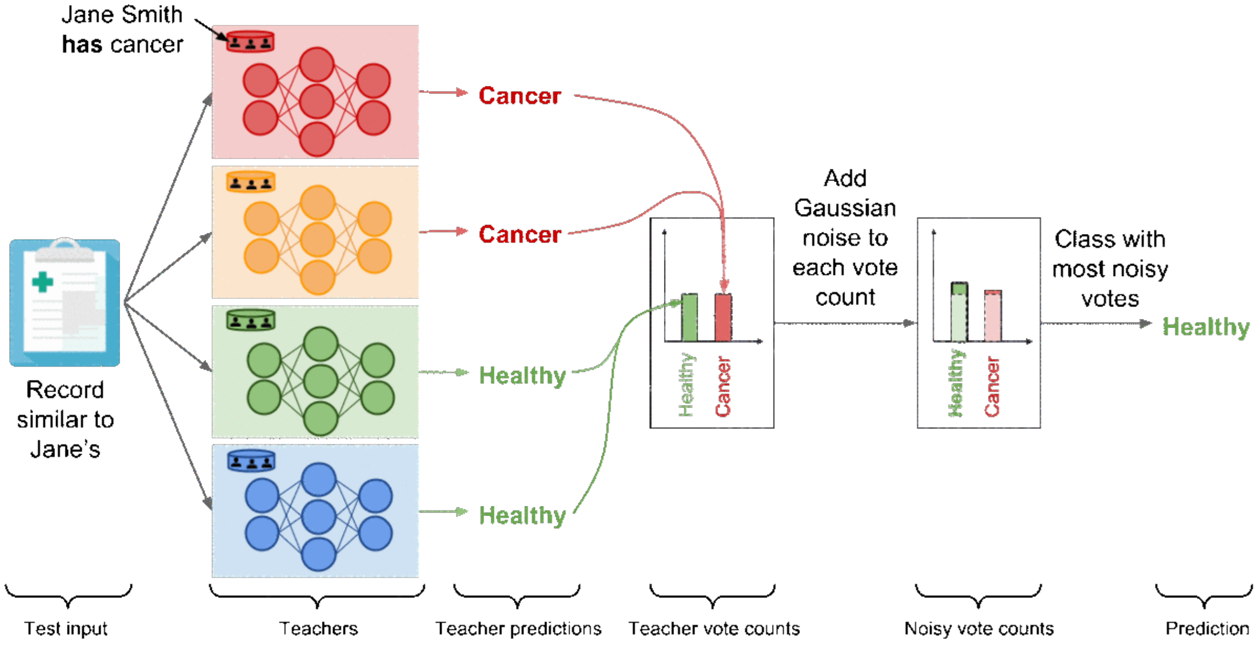


[17] http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.htmlReference: http://www.clev

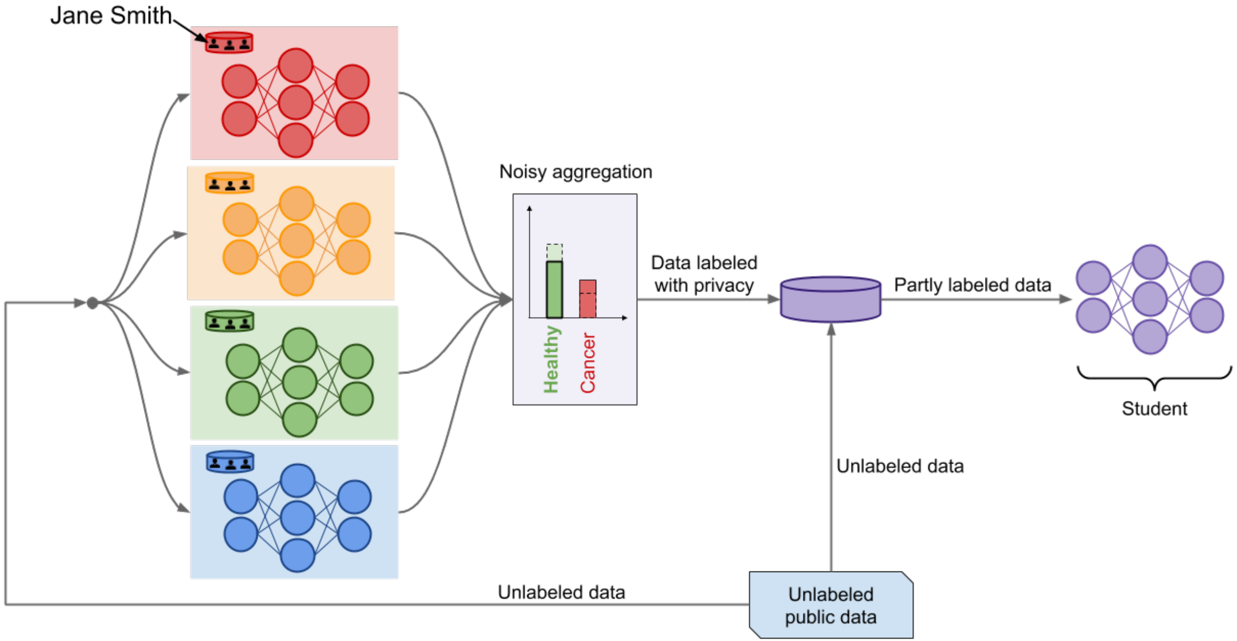
PATE: Private Aggregations of Teacher Ensembles



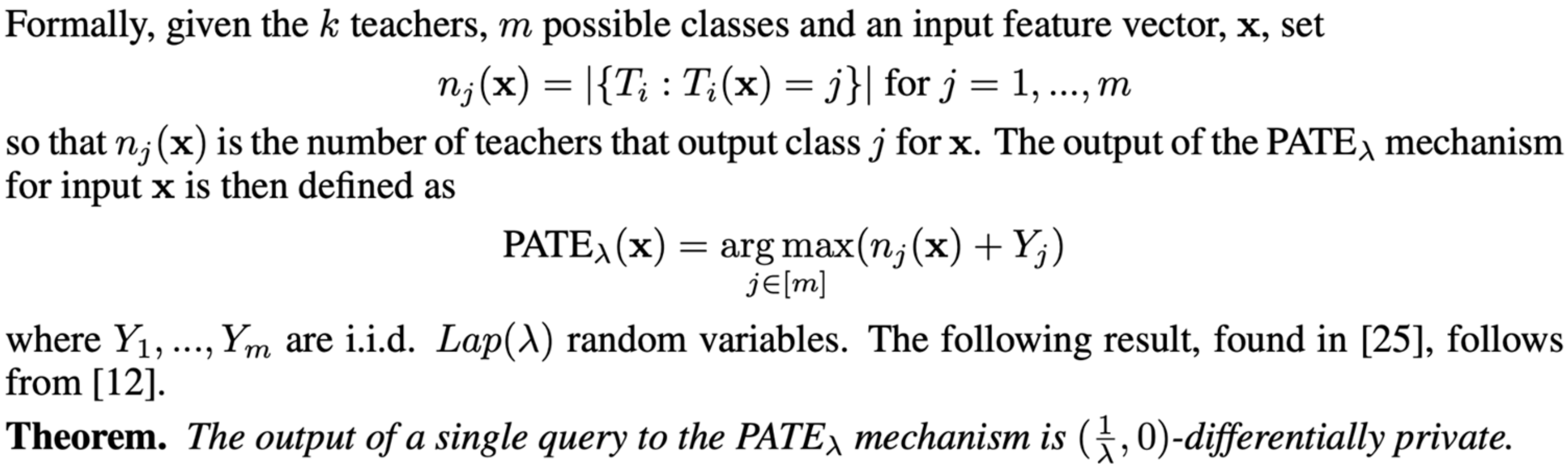
PATE: Private Aggregations of Teacher Ensembles



PATE: Private Aggregations of Teacher Ensembles



PATE: Differential Privacy



Formally, given the k teachers, m possible classes and an input feature vector, $\mathbf{x}$, set

$$n_j(\mathbf{x}) = |\{T_i : T_i(\mathbf{x}) = j\}| \text{ for } j = 1, ..., m$$

so that $n_j(\mathbf{x})$ is the number of teachers that output class j for $\mathbf{x}$. The output of the PATE$_\lambda$ mechanism for input $\mathbf{x}$ is then defined as

$$\text{PATE}_\lambda(\mathbf{x}) = \underset{j \in [m]}{\arg\max}(n_j(\mathbf{x}) + Y_j)$$

where $Y_1, ..., Y_m$ are i.i.d. $Lap(\lambda)$ random variables. The following result, found in [25], follows from [12].

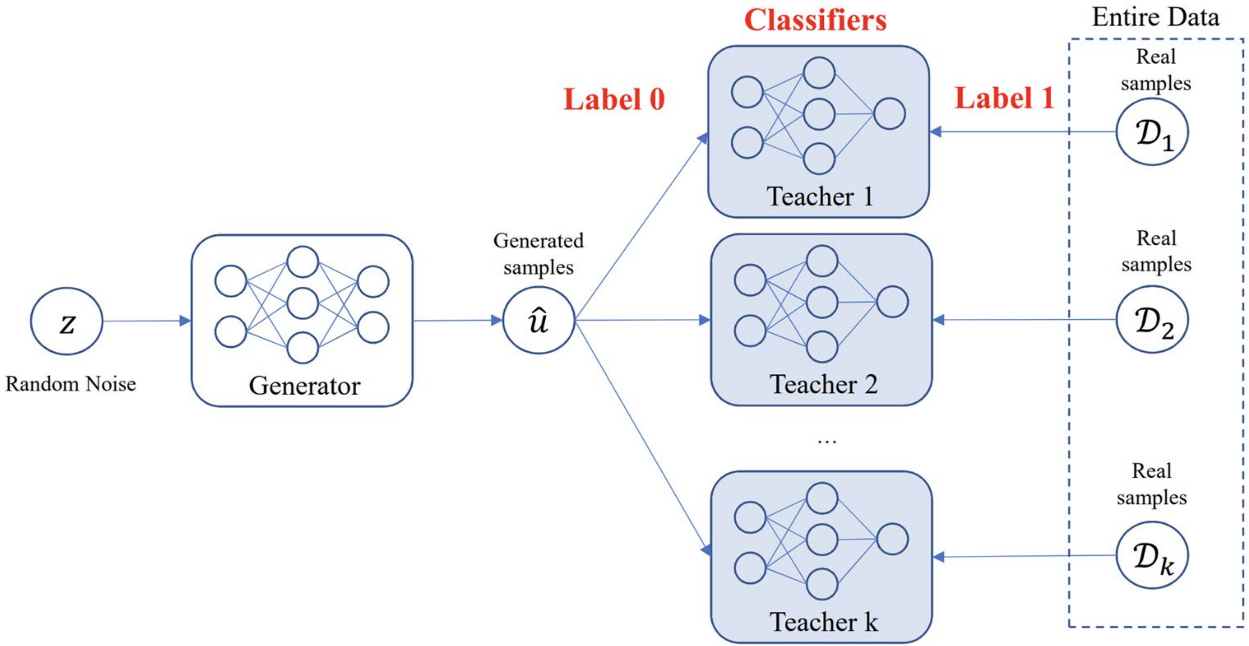
Theorem. *The output of a single query to the PATE$_\lambda$ mechanism is $(\frac{1}{\lambda}, 0)$-differentially private.*

PATE + GAN framework

- PATE framework is only applicable to **discriminative models**
- We need to extend the **PATE framework to generative models**
- We combine the state-of-the-art generative models **(GAN)** with the state-of-the-art DP framework **(PATE)**
- If we train the discriminator in a differentially private manner, then the generator will be differentially private by **post-processing theorem**
- A differentially private discriminator is **sufficient** but it may not be **necessary**.

[18] J. Jordon, J. Yoon and M. van der Schaar, ICLR, 2018

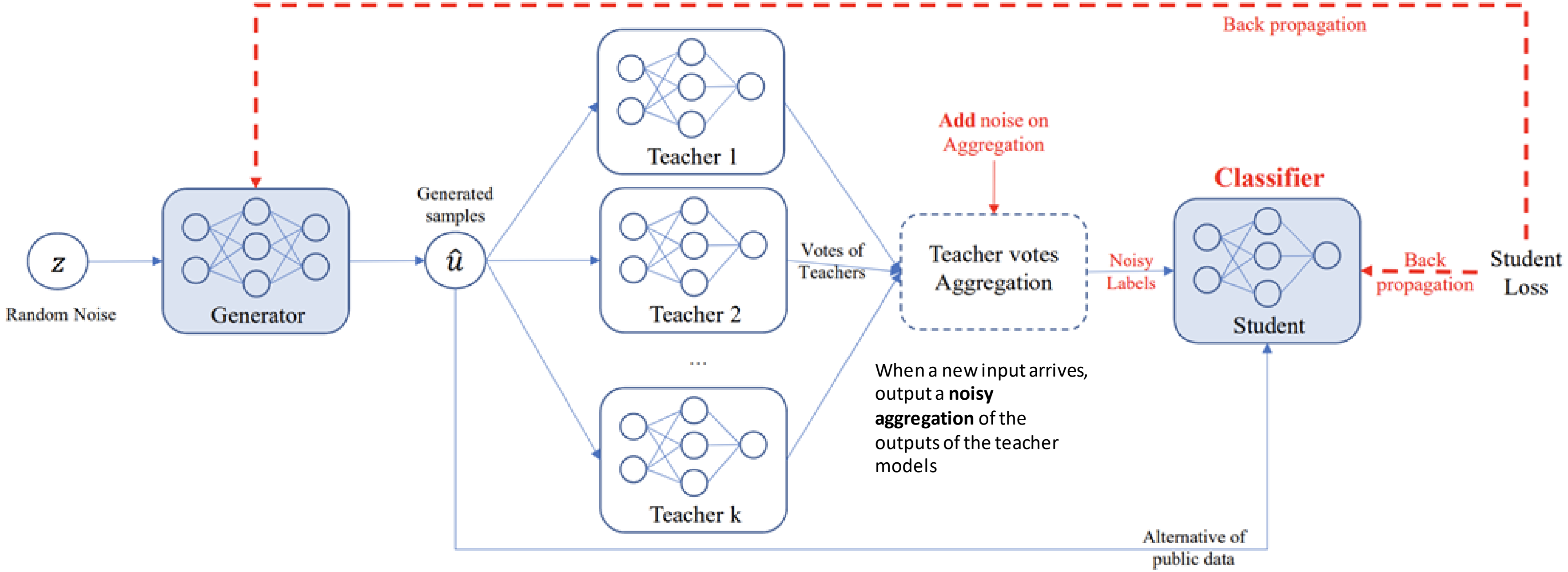
PATE-GAN teacher



Partition data so that no person lies in two partitions

[18] J. Jordon, J. Yoon and M. van der Schaar, ICLR, 2018

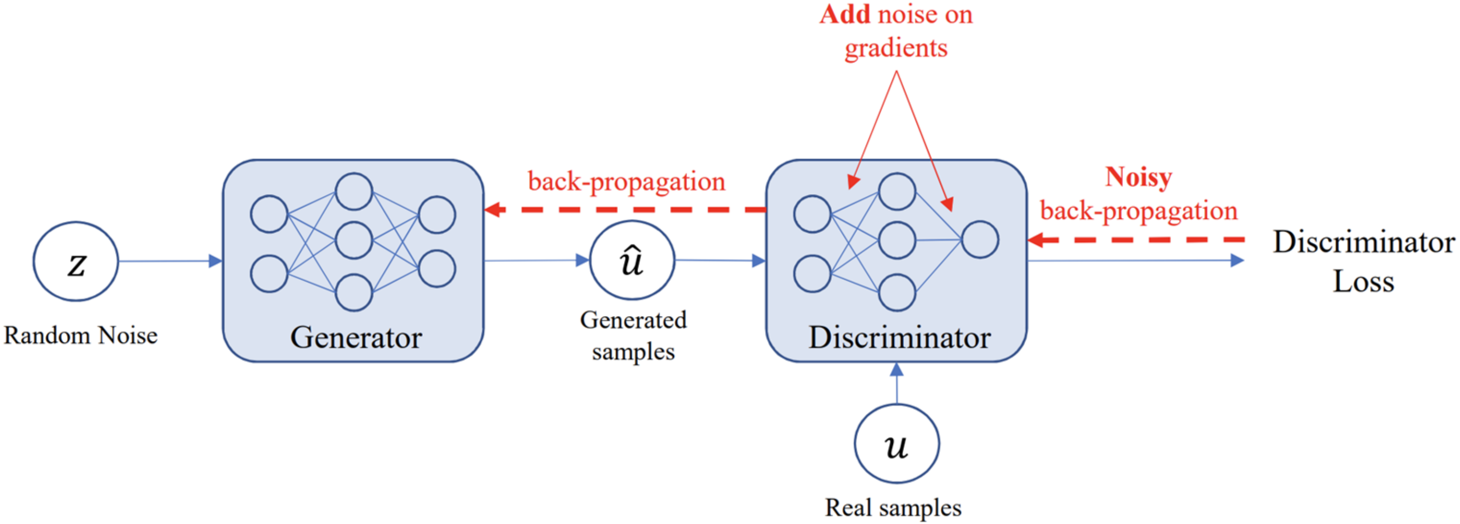
PATE-GAN framework



When a new input arrives, output a **noisy aggregation** of the outputs of the teacher models

[18] J. Jordon, J. Yoon and M. van der Schaar, ICLR, 2018

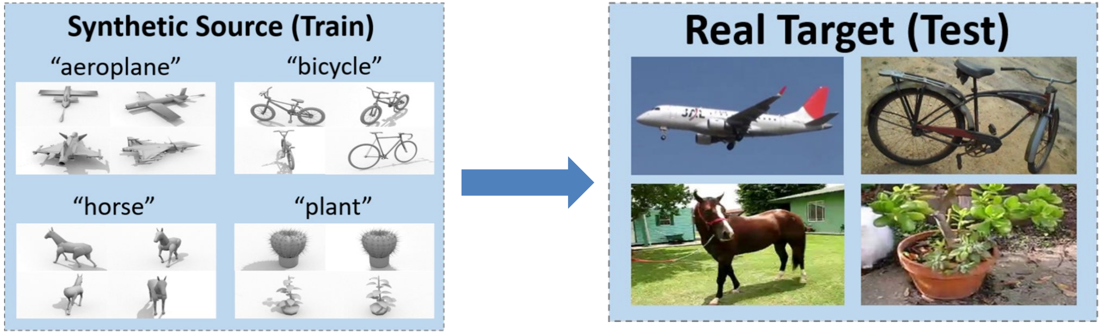
Baseline: DP-GAN



Evaluations: How to Evaluate Synthetic Data?

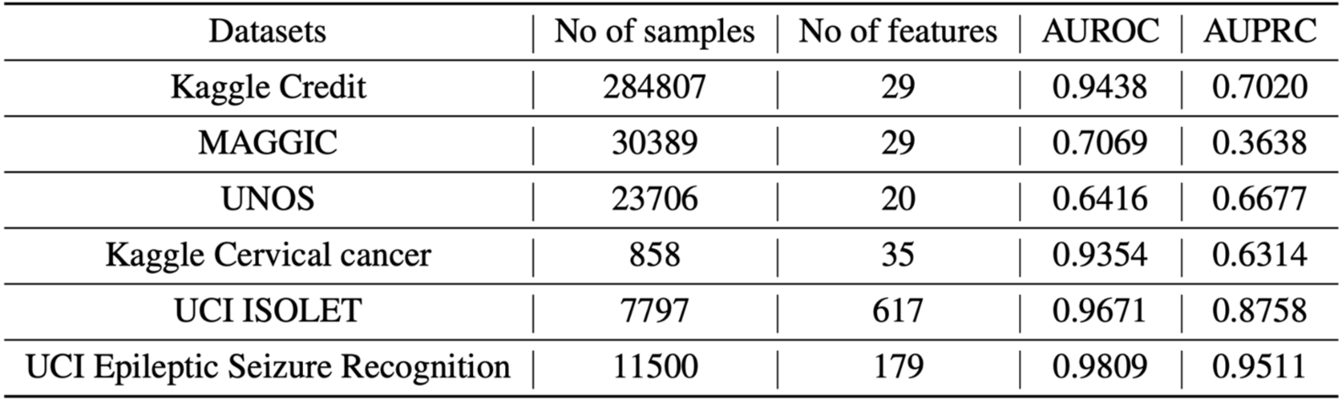
- **Predictivity**: samples should be **just as useful as** real data when used for the same predictive purposes (i.e. train on synthetic data, test on real data)

- **Experimental settings:**
  - **Setting A:** Train on Real, Test on Real
  - **Setting B:** Train on Synthetic, Test on Real
  - **Setting C:** Train on Synthetic, Test on Synthetic



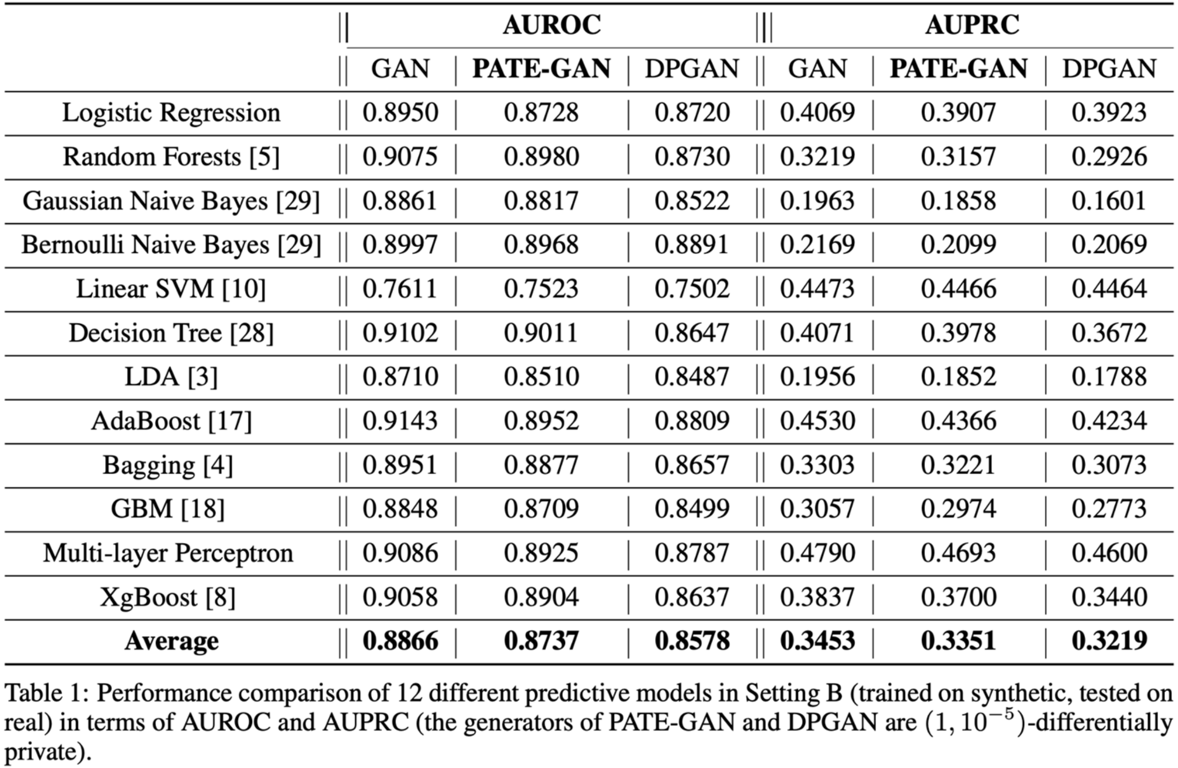
Evaluations: Datasets

- **Predictivity**: samples should be **just as useful as** real data when used for the same predictive purposes (i.e. train on synthetic data, test on real data)



Datasets	No of samples	No of features	AUROC	AUPRC
Kaggle Credit	284807	29	0.9438	0.7020
MAGGIC	30389	29	0.7069	0.3638
UNOS	23706	20	0.6416	0.6677
Kaggle Cervical cancer	858	35	0.9354	0.6314
UCI ISOLET	7797	617	0.9671	0.8758
UCI Epileptic Seizure Recognition	11500	179	0.9809	0.9511

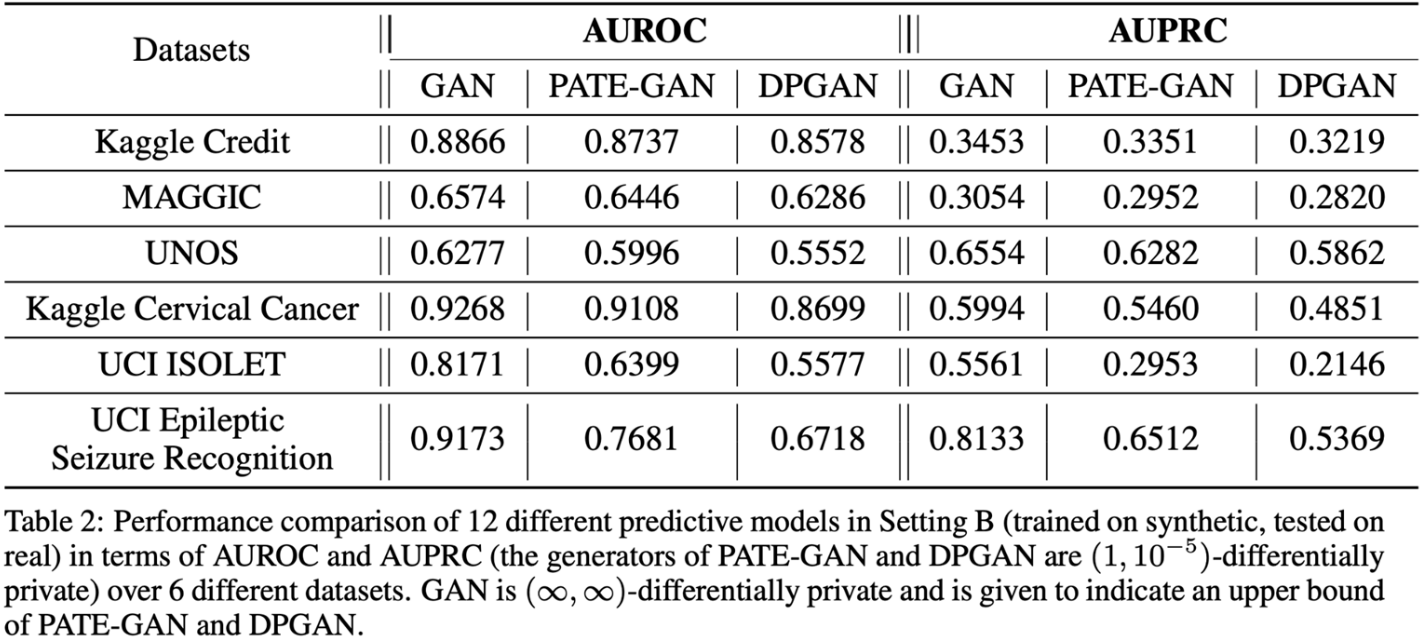
Evaluations: Setting B with different models



	AUROC			AUPRC		
	GAN	**PATE-GAN**	DPGAN	GAN	**PATE-GAN**	DPGAN
Logistic Regression	0.8950	0.8728	0.8720	0.4069	0.3907	0.3923
Random Forests [5]	0.9075	0.8980	0.8730	0.3219	0.3157	0.2926
Gaussian Naive Bayes [29]	0.8861	0.8817	0.8522	0.1963	0.1858	0.1601
Bernoulli Naive Bayes [29]	0.8997	0.8968	0.8891	0.2169	0.2099	0.2069
Linear SVM [10]	0.7611	0.7523	0.7502	0.4473	0.4466	0.4464
Decision Tree [28]	0.9102	0.9011	0.8647	0.4071	0.3978	0.3672
LDA [3]	0.8710	0.8510	0.8487	0.1956	0.1852	0.1788
AdaBoost [17]	0.9143	0.8952	0.8809	0.4530	0.4366	0.4234
Bagging [4]	0.8951	0.8877	0.8657	0.3303	0.3221	0.3073
GBM [18]	0.8848	0.8709	0.8499	0.3057	0.2974	0.2773
Multi-layer Perceptron	0.9086	0.8925	0.8787	0.4790	0.4693	0.4600
XgBoost [8]	0.9058	0.8904	0.8637	0.3837	0.3700	0.3440
Average	**0.8866**	**0.8737**	**0.8578**	**0.3453**	**0.3351**	**0.3219**

Table 1: Performance comparison of 12 different predictive models in Setting B (trained on synthetic, tested on real) in terms of AUROC and AUPRC (the generators of PATE-GAN and DPGAN are $(1, 10^{-5})$-differentially private).

Evaluations: Setting B with different datasets



Datasets	AUROC			AUPRC		
	GAN	PATE-GAN	DPGAN	GAN	PATE-GAN	DPGAN
Kaggle Credit	0.8866	0.8737	0.8578	0.3453	0.3351	0.3219
MAGGIC	0.6574	0.6446	0.6286	0.3054	0.2952	0.2820
UNOS	0.6277	0.5996	0.5552	0.6554	0.6282	0.5862
Kaggle Cervical Cancer	0.9268	0.9108	0.8699	0.5994	0.5460	0.4851
UCI ISOLET	0.8171	0.6399	0.5577	0.5561	0.2953	0.2146
UCI Epileptic Seizure Recognition	0.9173	0.7681	0.6718	0.8133	0.6512	0.5369

Table 2: Performance comparison of 12 different predictive models in Setting B (trained on synthetic, tested on real) in terms of AUROC and AUPRC (the generators of PATE-GAN and DPGAN are $(1, 10^{-5})$-differentially private) over 6 different datasets. GAN is (∞, ∞)-differentially private and is given to indicate an upper bound of PATE-GAN and DPGAN.

Evaluations: Setting B with Different Privacy Budget

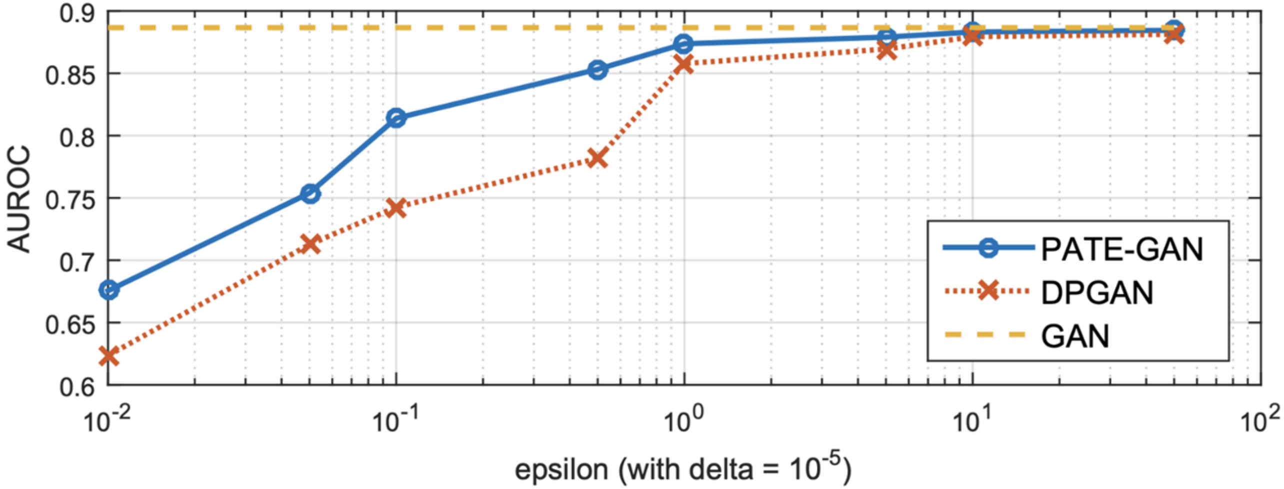
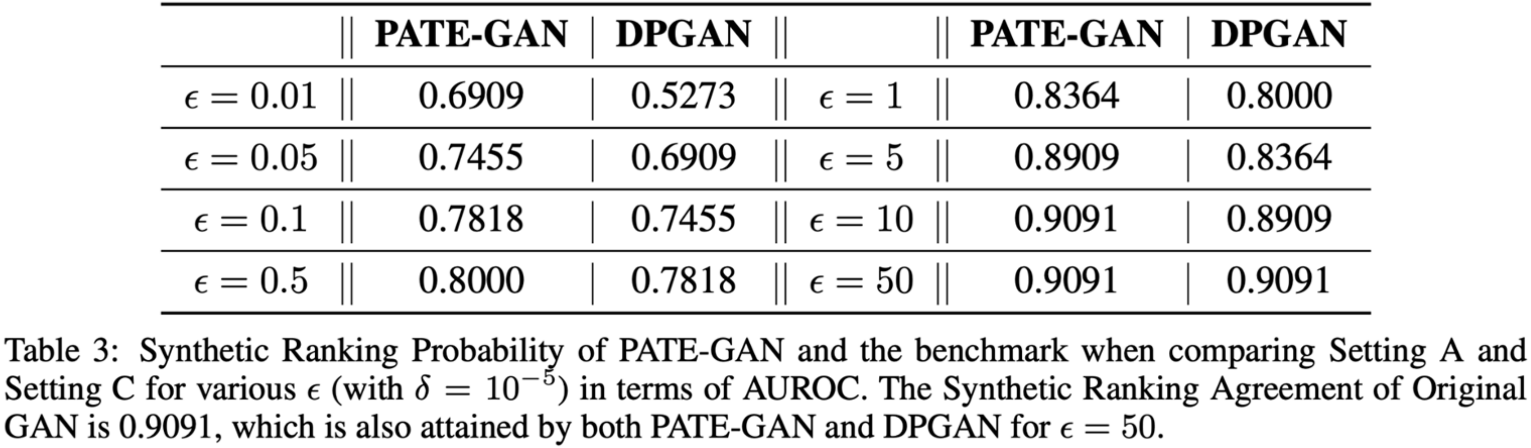


Figure 1: Average AUROC performance across 12 different predictive models trained on the synthetic dataset generated by PATE-GAN and DPGAN with various ϵ (with $\delta = 10^{-5}$) (Setting B).

Evaluations: Setting C for model ranking



	PATE-GAN	**DPGAN**		**PATE-GAN**	**DPGAN**
$\epsilon = 0.01$	0.6909	0.5273	$\epsilon = 1$	0.8364	0.8000
$\epsilon = 0.05$	0.7455	0.6909	$\epsilon = 5$	0.8909	0.8364
$\epsilon = 0.1$	0.7818	0.7455	$\epsilon = 10$	0.9091	0.8909
$\epsilon = 0.5$	0.8000	0.7818	$\epsilon = 50$	0.9091	0.9091

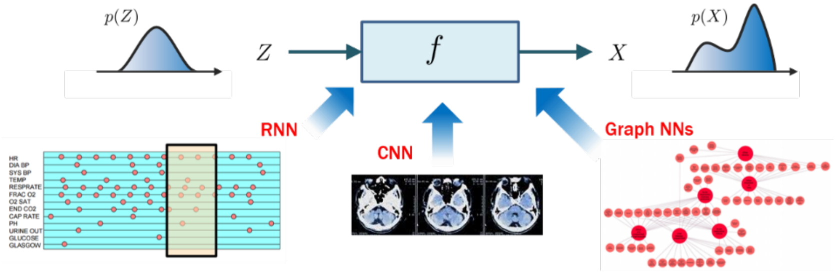
Table 3: Synthetic Ranking Probability of PATE-GAN and the benchmark when comparing Setting A and Setting C for various ϵ (with $\delta = 10^{-5}$) in terms of AUROC. The Synthetic Ranking Agreement of Original GAN is 0.9091, which is also attained by both PATE-GAN and DPGAN for $\epsilon = 50$.

Recipe for generating synthetic healthcare data

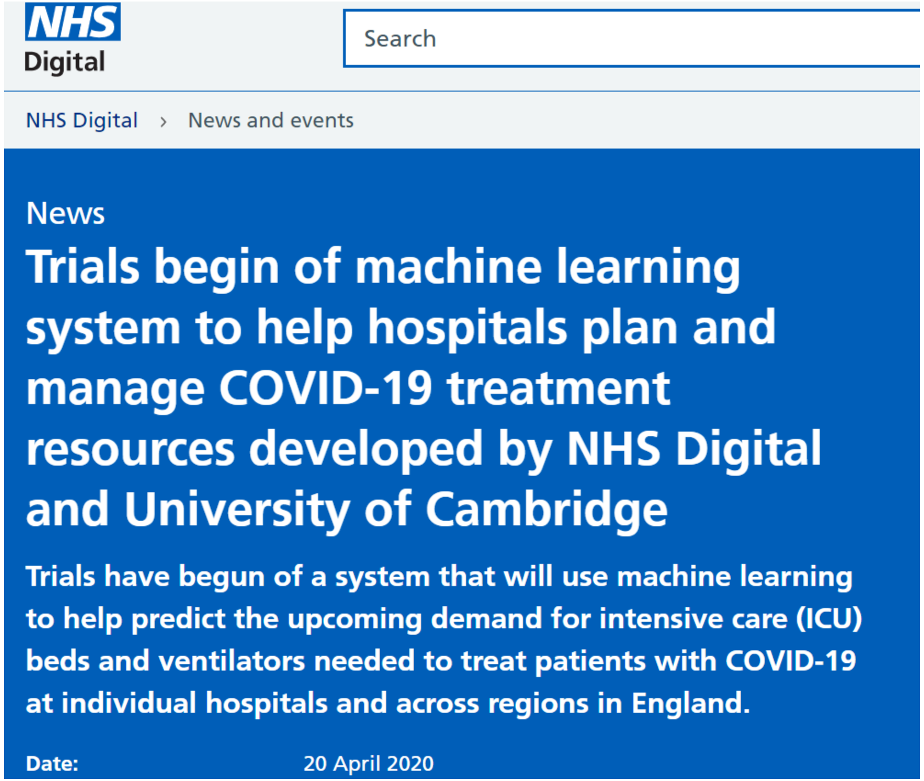
- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.

- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., NNs, RNN, CNN, transfer representations etc.

- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**



Generating synthetic data for multiple hospitals



NHS Digital

Search

NHS Digital › News and events

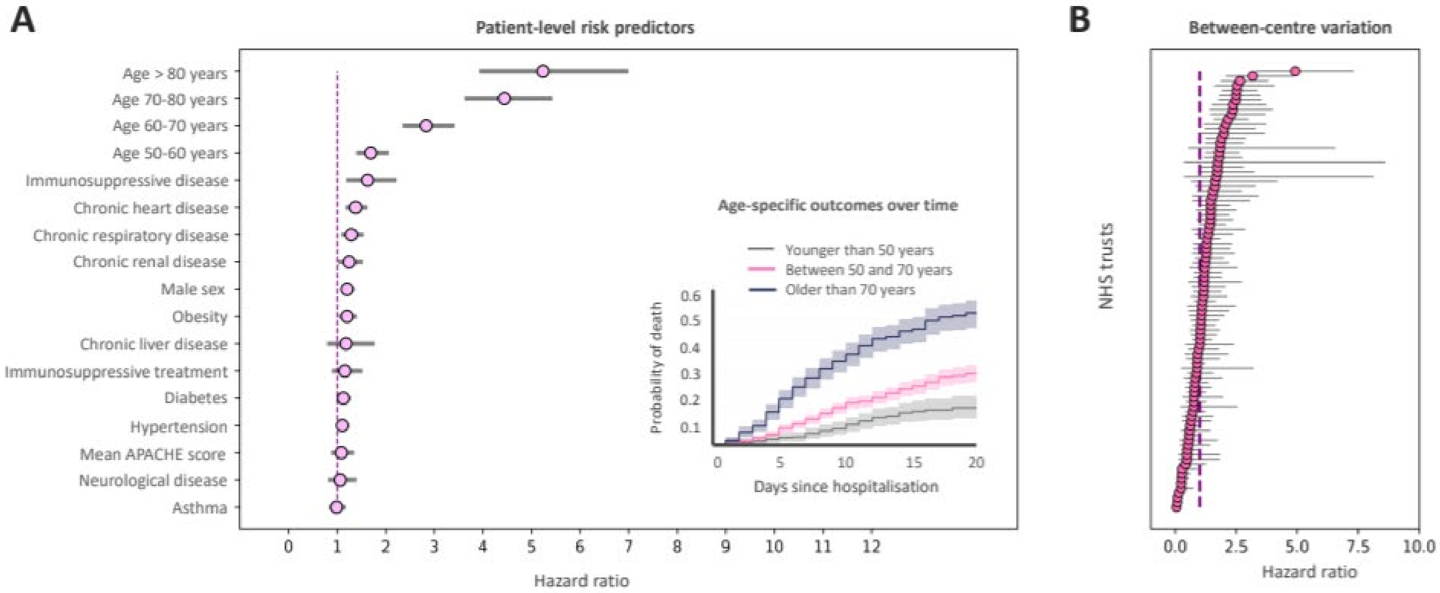
News

Trials begin of machine learning system to help hospitals plan and manage COVID-19 treatment resources developed by NHS Digital and University of Cambridge

Trials have begun of a system that will use machine learning to help predict the upcoming demand for intensive care (ICU) beds and ventilators needed to treat patients with COVID-19 at individual hospitals and across regions in England.

Date: 20 April 2020

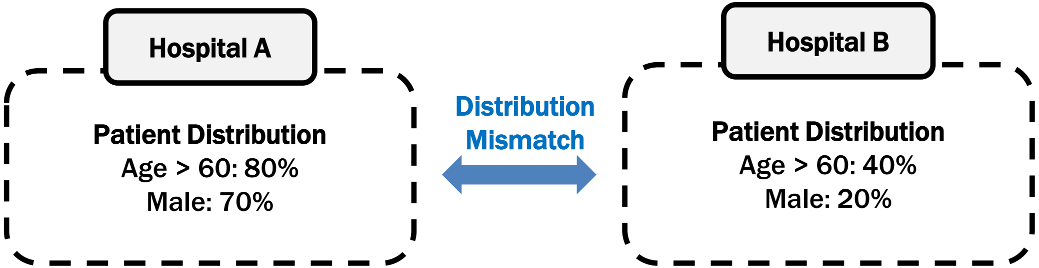
Between-center differences for COVID-19 ICU mortality from early data in England



[19] Z. Qian, A. Alaa, A. Ercole, M. van der Schaar, Intensive care medicine, 2020

Transferring what we learned to other jurisdictions

- **Distribution Mismatch**: Auxiliary data does not come from the **same distribution as target data**



Hospital A

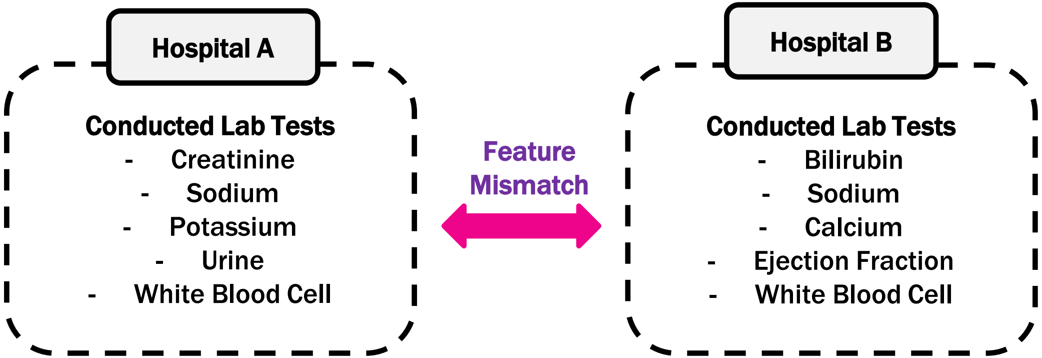
Patient Distribution
Age > 60: 80%
Male: 70%

Distribution Mismatch

Hospital B

Patient Distribution
Age > 60: 40%
Male: 20%

- **Feature Mismatch**: Different data collectors collect **different pieces of information** for each sample



Hospital A

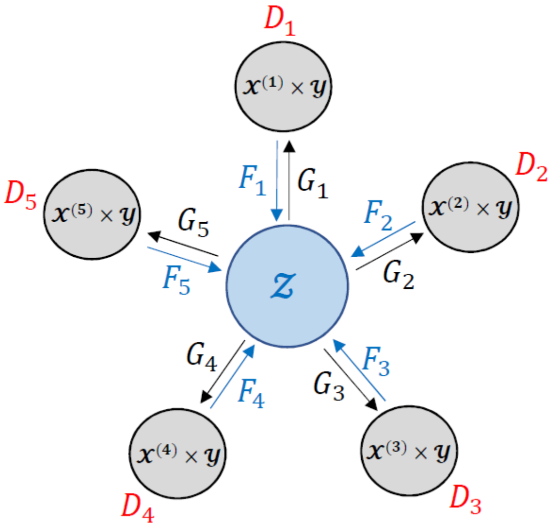
Conducted Lab Tests
- Creatinine
- Sodium
- Potassium
- Urine
- White Blood Cell

Feature Mismatch

Hospital B

Conducted Lab Tests
- Bilirubin
- Sodium
- Calcium
- Ejection Fraction
- White Blood Cell

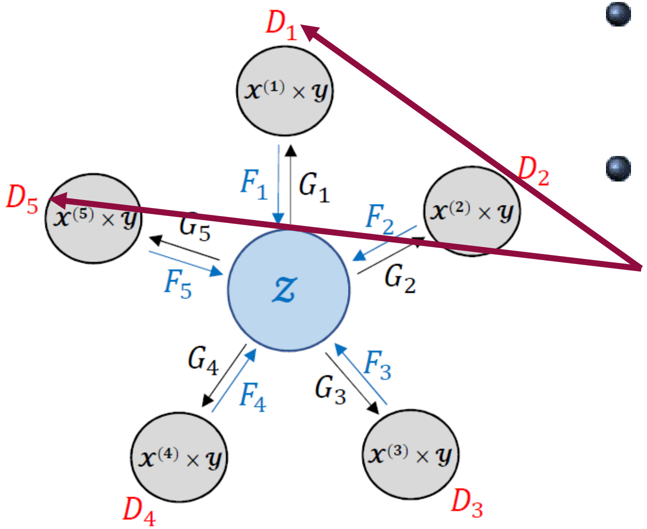
RadialGAN



- Use **multiple GAN architectures** to "**translate**" the data from one dataset to another

[20] J. Yoon, J. Jordon, M. van der Schaar, 2019

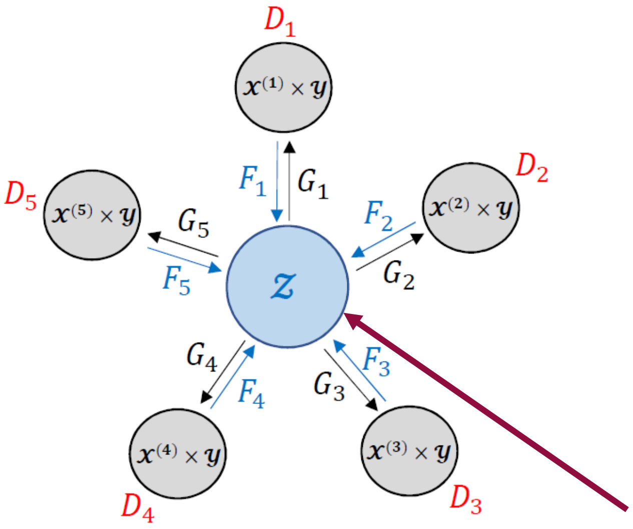
RadialGAN



- Use **multiple GAN architectures** to "**translate**" the data from one dataset to another

- **Distribution Mismatch** is dealt with by the **adversarial framework** which ensures that "translation" respects the **target distribution**

RadialGAN



- Use **multiple GAN architectures** to “**translate**” the data from one dataset to another

- **Distribution Mismatch** is dealt with by the **adversarial framework** which ensures that “translation” respects the **target distribution**

- **Feature mismatch** is dealt with by introducing a **latent space** through which all samples are mapped

RadialGAN

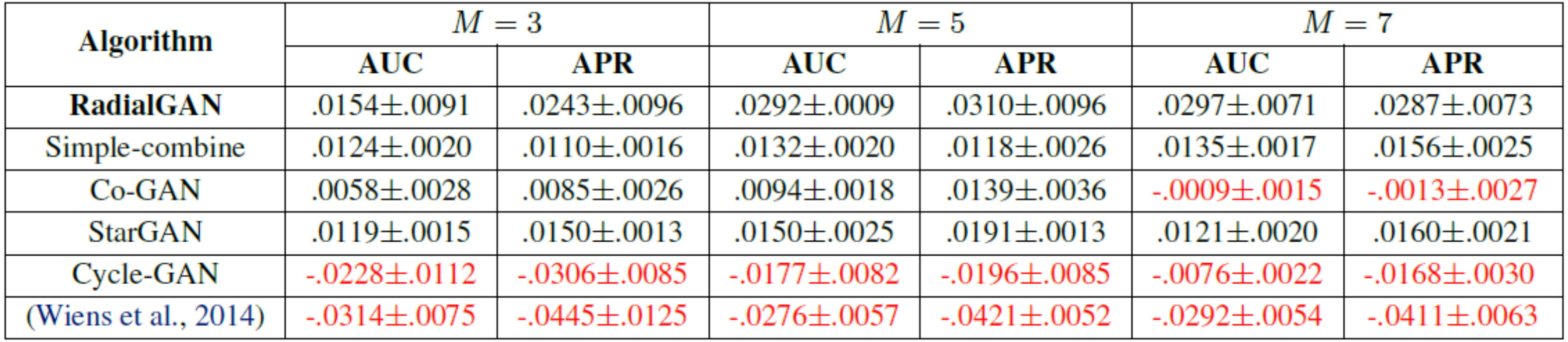
MAGGIC Dataset – Heart Failure

- A collection of **different medical studies**
- Label is set as **1-year all-cause mortality**

- Total number of **features** across all studies **is 216**

(1) Average number of features in each study: **66**

(2) Average number of **shared** features: **35 (53.8%)**



Algorithm	$M = 3$		$M = 5$		$M = 7$	
	AUC	**APR**	**AUC**	**APR**	**AUC**	**APR**
RadialGAN	.0154±.0091	.0243±.0096	.0292±.0009	.0310±.0096	.0297±.0071	.0287±.0073
Simple-combine	.0124±.0020	.0110±.0016	.0132±.0020	.0118±.0026	.0135±.0017	.0156±.0025
Co-GAN	.0058±.0028	.0085±.0026	.0094±.0018	.0139±.0036	-.0009±.0015	-.0013±.0027
StarGAN	.0119±.0015	.0150±.0013	.0150±.0025	.0191±.0013	.0121±.0020	.0160±.0021
Cycle-GAN	-.0228±.0112	-.0306±.0085	-.0177±.0082	-.0196±.0085	-.0076±.0022	-.0168±.0030
(Wiens et al., 2014)	-.0314±.0075	-.0445±.0125	-.0276±.0057	-.0421±.0052	-.0292±.0054	-.0411±.0063

Generative models for synthetic time-series data

Time-series generation

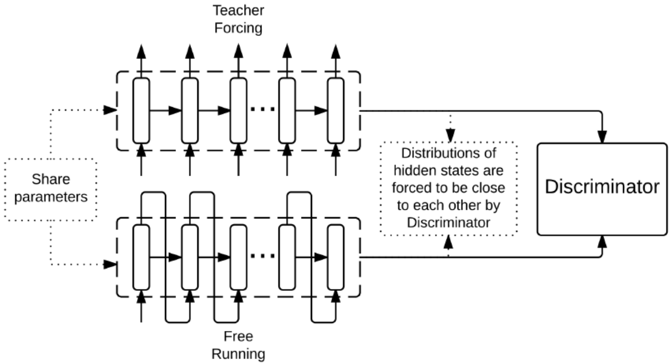
- Objective: To generate time-series data with preserving temporal dynamics

- Key Example: Synthetic time-series healthcare data generation

- Challenges: Capture the distributions of features within each time point as well as complex dynamics of those variables across time points

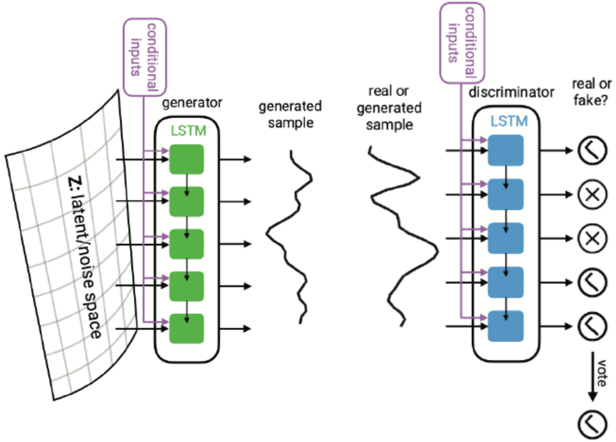
Time-series generation approaches

Autoregressive model



Teacher Forcing
Share parameters
Distributions of hidden states are forced to be close to each other by Discriminator
Discriminator
Free Running

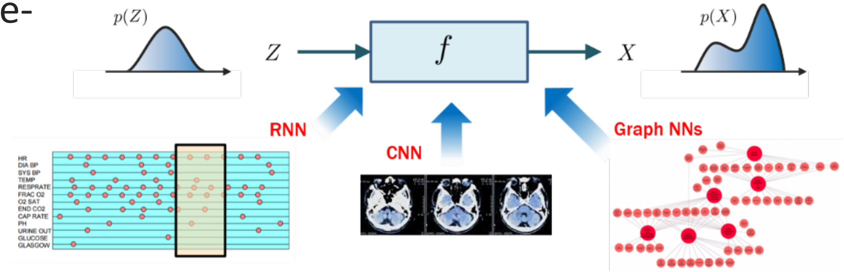
Recurrent generative model



Recipe for generating synthetic healthcare data

- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.

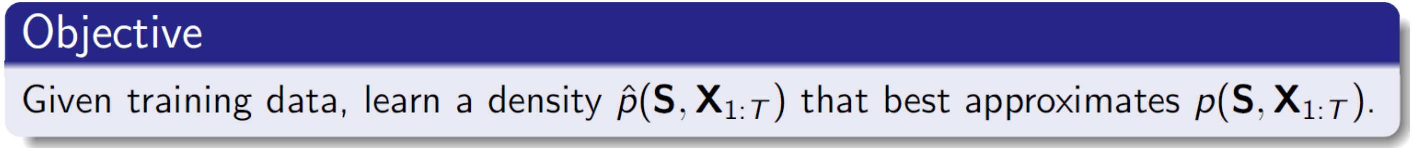
- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., Autoregressive, attention, latent state-space, RNN representations, spectral, etc.



- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**

Problem formulation

- Consider tuples of the form $(S, X_{1:T})$ with some joint distribution p, where
  - **Static features:** $S \in \mathcal{S}$
  - **Temporal features:** $X \in \mathcal{X}$



Objective

Given training data, learn a density $\hat{p}(\mathbf{S}, \mathbf{X}_{1:T})$ that best approximates $p(\mathbf{S}, \mathbf{X}_{1:T})$.

- Desiderata
  - **Fidelity:** samples should be **indistinguishable** from real data
  - **Diversity:** samples should be **distributed to cover** that of real data
  - **Predictivity:** samples should be **just as useful as** real data when used for the same predictive purposes (i.e. train on synthetic data, test on real data)

Two objectives

Global (sequence-level)

Matching the joint distribution...

$$\min_{\hat{p}} D\Big(p(\mathbf{S}, \mathbf{X}_{1:T}) \Big\| \hat{p}(\mathbf{S}, \mathbf{X}_{1:T})\Big) \tag{1}$$

...requires a perfect adversary.

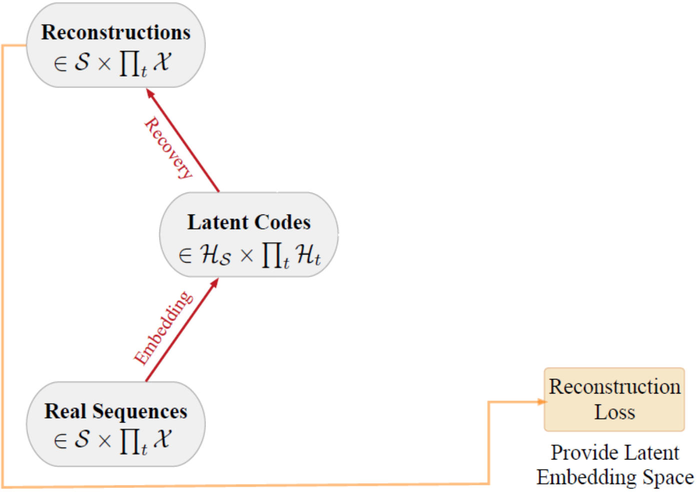
Local (stepwise)

Matching the conditionals...

$$\min_{\hat{p}} D\Big(p(\mathbf{X}_t | \mathbf{S}, \mathbf{X}_{1:t-1}) \Big\| \hat{p}(\mathbf{X}_t | \mathbf{S}, \mathbf{X}_{1:t-1})\Big) \tag{2}$$

...requires ground-truth sequences.

TimeGAN: Encode, Generate, and Iterate

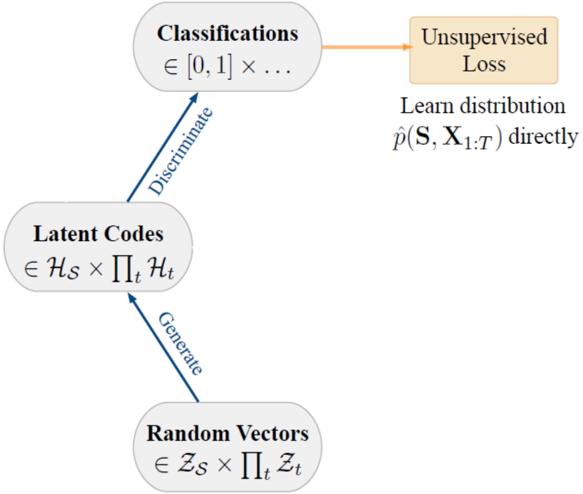

Reconstructions
$\in \mathcal{S} \times \prod_t \mathcal{X}$
Recovery
Latent Codes
$\in \mathcal{H}_{\mathcal{S}} \times \prod_t \mathcal{H}_t$
Embedding
Real Sequences
$\in \mathcal{S} \times \prod_t \mathcal{X}$
Reconstruction Loss
Provide Latent Embedding Space

- **Provide mapping between the feature space and latent space, where the adversarial network learns the underlying temporal dynamics of the data**

$$\mathcal{L}_{\mathrm{R}} = \mathbb{E}_{\mathbf{s},\mathbf{x}_{1:T}\sim p}\left[\|\mathbf{s} - \tilde{\mathbf{s}}\|_2 + \sum_t \|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|_2\right]$$

[21] J. Yoon, D. Jarrett and M. van der Schaar, NeurIPS 2019

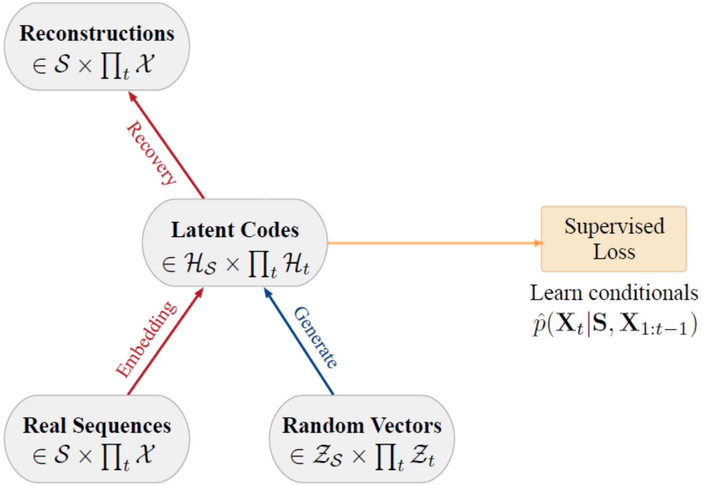
TimeGAN: Encode, Generate, and Iterate


Classifications
$\in [0,1] \times \dots$
Unsupervised Loss
Learn distribution
$\hat{p}(\mathbf{S}, \mathbf{X}_{1:T})$ directly
Discriminate
Latent Codes
$\in \mathcal{H}_{\mathcal{S}} \times \prod_t \mathcal{H}_t$
Generate
Random Vectors
$\in \mathcal{Z}_{\mathcal{S}} \times \prod_t \mathcal{Z}_t$

- **The generator produces synthetic outputs through the latent space, and the discriminator operates on the basis of real vs. synthetic embeddings.**

$$\mathcal{L}_{\mathrm{U}} = \mathbb{E}_{\mathbf{s},\mathbf{x}_{1:T}\sim p}\Big[\log y_{\mathcal{S}} + \sum_t \log y_t\Big] + \mathbb{E}_{\mathbf{s},\mathbf{x}_{1:T}\sim \hat{p}}\Big[\log(1-\hat{y}_{\mathcal{S}}) + \sum_t \log(1-\hat{y}_t)\Big]$$

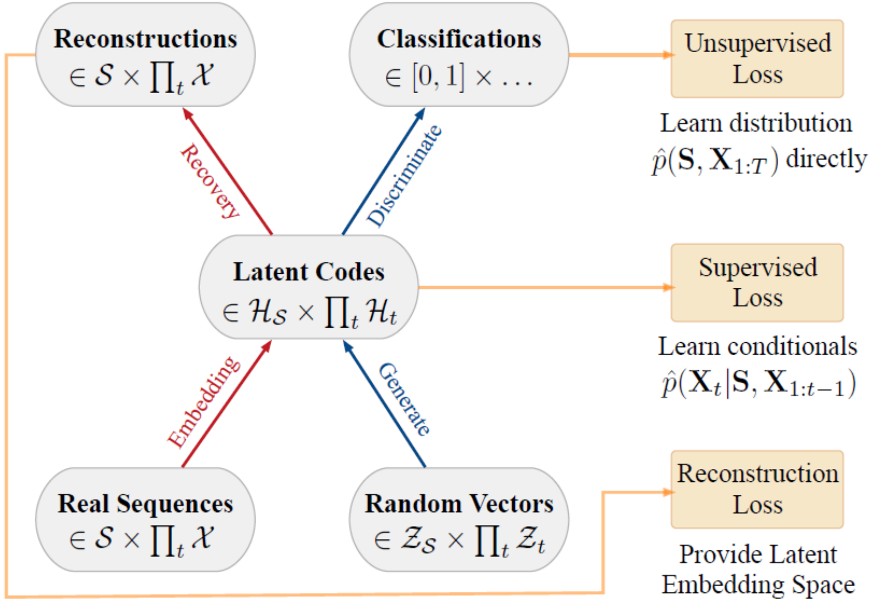
TimeGAN: Encode, Generate, and Iterate


Reconstructions
$\in \mathcal{S} \times \prod_t \mathcal{X}$
Recovery
Latent Codes
$\in \mathcal{H}_\mathcal{S} \times \prod_t \mathcal{H}_t$
Supervised Loss
Learn conditionals
$\hat{p}(\mathbf{X}_t|\mathbf{S}, \mathbf{X}_{1:t-1})$
Embedding
Generate
Real Sequences
$\in \mathcal{S} \times \prod_t \mathcal{X}$
Random Vectors
$\in \mathcal{Z}_\mathcal{S} \times \prod_t \mathcal{Z}_t$

- **The generator receives sequences of embeddings of actual data to generate the next latent vector.**

$$\mathcal{L}_\mathsf{S} = \mathbb{E}_{\mathbf{s},\mathbf{x}_{1:T}\sim p}\Big[\sum\nolimits_t \|\mathbf{h}_t - g_\mathcal{X}(\mathbf{h}_\mathcal{S}, \mathbf{h}_{t-1}, \mathbf{z}_t)\|_2\Big]$$

TimeGAN: Jointly optimize

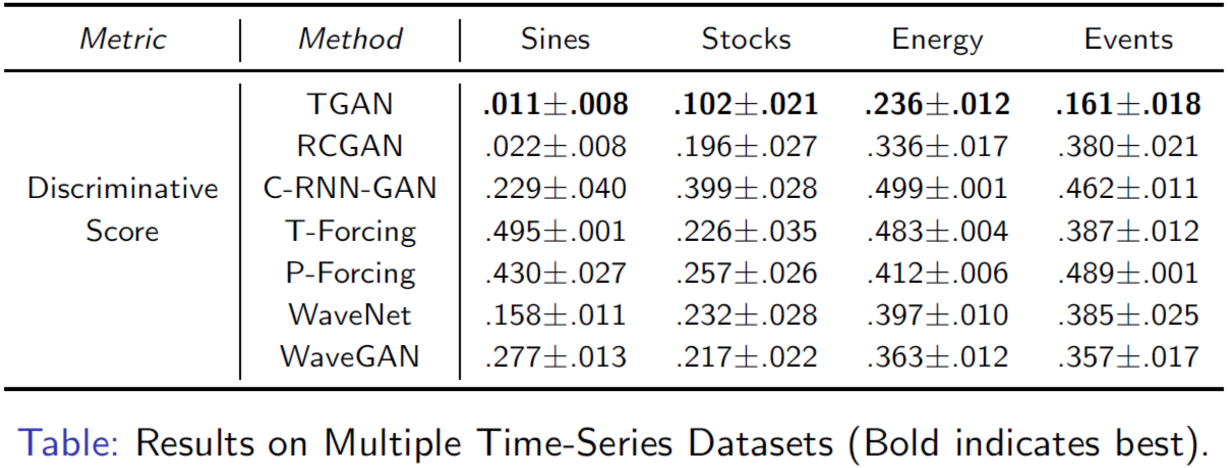


Reconstructions
$\in \mathcal{S} \times \prod_t \mathcal{X}$
Classifications
$\in [0,1] \times \ldots$
Unsupervised Loss
Learn distribution
$\hat{p}(\mathbf{S}, \mathbf{X}_{1:T})$ directly
Recovery
Discriminate
Latent Codes
$\in \mathcal{H}_\mathcal{S} \times \prod_t \mathcal{H}_t$
Supervised Loss
Learn conditionals
$\hat{p}(\mathbf{X}_t | \mathbf{S}, \mathbf{X}_{1:t-1})$
Embedding
Generate
Real Sequences
$\in \mathcal{S} \times \prod_t \mathcal{X}$
Random Vectors
$\in \mathcal{Z}_\mathcal{S} \times \prod_t \mathcal{Z}_t$
Reconstruction Loss
Provide Latent Embedding Space

$$\min_{\theta_e, \theta_r} (\lambda \mathcal{L}_\mathsf{S} + \mathcal{L}_\mathsf{R}) \qquad \text{and} \qquad \min_{\theta_g} (\eta \mathcal{L}_\mathsf{S} + \max_{\theta_d} \mathcal{L}_\mathsf{U})$$

How to evaluate synthetic time-series data?

- **Fidelity**: samples should be **indistinguishable** from real data

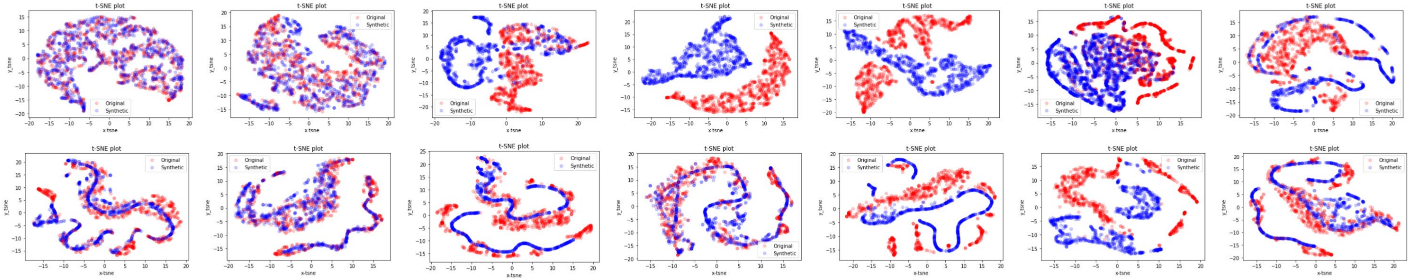


Metric	*Method*	Sines	Stocks	Energy	Events
Discriminative Score	TGAN	**.011±.008**	**.102±.021**	**.236±.012**	**.161±.018**
	RCGAN	.022±.008	.196±.027	.336±.017	.380±.021
	C-RNN-GAN	.229±.040	.399±.028	.499±.001	.462±.011
	T-Forcing	.495±.001	.226±.035	.483±.004	.387±.012
	P-Forcing	.430±.027	.257±.026	.412±.006	.489±.001
	WaveNet	.158±.011	.232±.028	.397±.010	.385±.025
	WaveGAN	.277±.013	.217±.022	.363±.012	.357±.017

Table: Results on Multiple Time-Series Datasets (Bold indicates best).

How to evaluate synthetic time-series data?

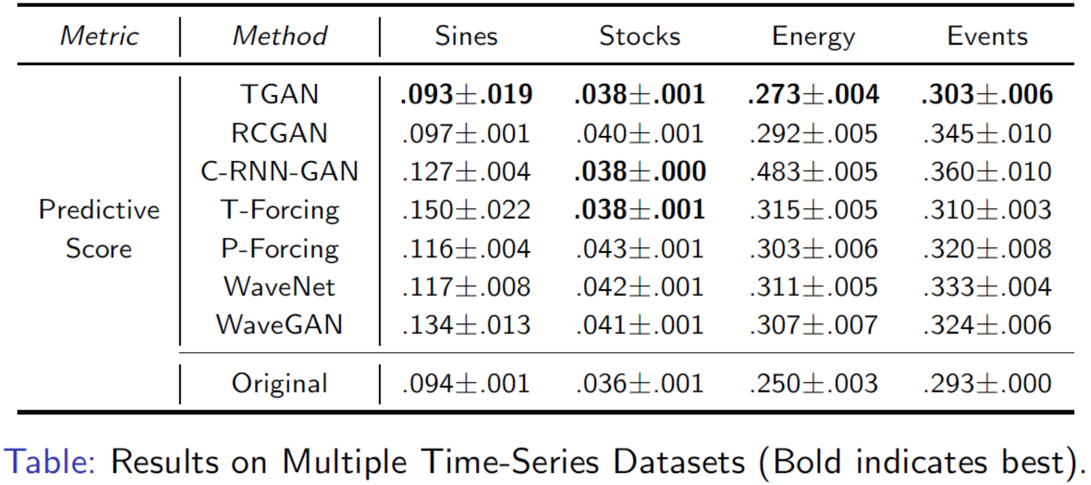
- **Diversity**: samples should be distributed to **cover** the real data



(a) TimeGAN (b) RCGAN (c) CRNNGAN (d) T-Forcing (e) P-Forcing (f) WaveNet (g) WaveGAN

How to evaluate synthetic time-series data?

- **Predictivity**: Samples should be just as useful as real data when used for the same predictive purposes (i.e. train on synthetic, test on real)

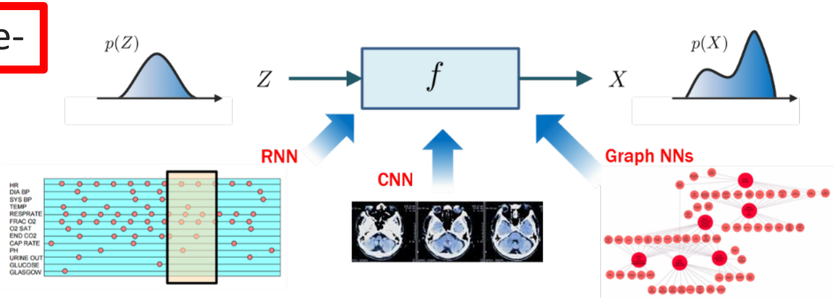


Metric	*Method*	Sines	Stocks	Energy	Events
Predictive Score	TGAN	**.093±.019**	**.038±.001**	**.273±.004**	**.303±.006**
	RCGAN	.097±.001	.040±.001	.292±.005	.345±.010
	C-RNN-GAN	.127±.004	**.038±.000**	.483±.005	.360±.010
	T-Forcing	.150±.022	**.038±.001**	.315±.005	.310±.003
	P-Forcing	.116±.004	.043±.001	.303±.006	.320±.008
	WaveNet	.117±.008	.042±.001	.311±.005	.333±.004
	WaveGAN	.134±.013	.041±.001	.307±.007	.324±.006
	Original	.094±.001	.036±.001	.250±.003	.293±.000

Table: Results on Multiple Time-Series Datasets (Bold indicates best).

Recipe for generating synthetic healthcare data

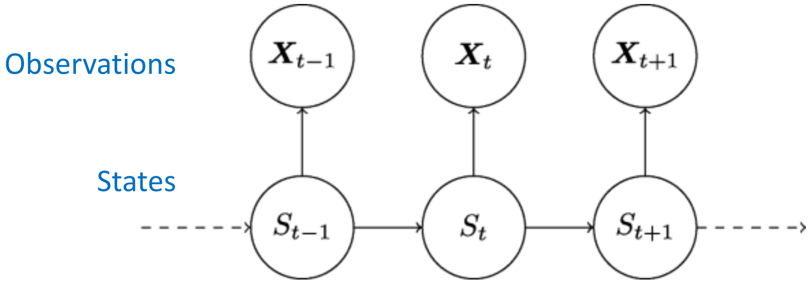
- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.
- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., Autoregressive, attention, latent state-space, RNN representations, spectral, etc.
- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**



Attentive state-space models

- **Latent variable models for sequential data → state-space models**
- **A sequence of variational autoencoders**
- **Rationale:**
  - Interpretable: can discover knowledge
  - Can incorporate structure and domain knowledge
- **Unlike other models: learning and inference need to be tailored to the model**

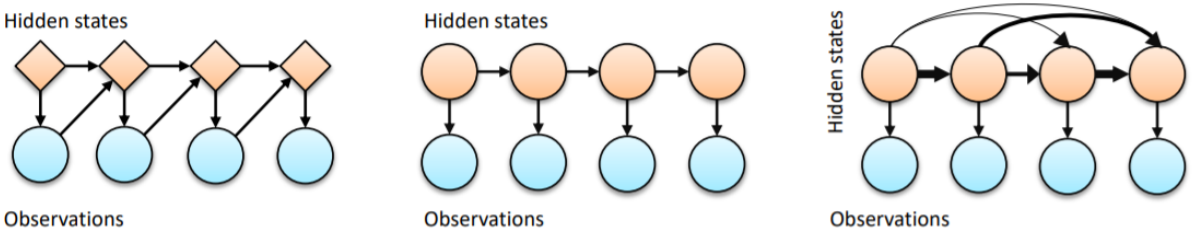
Patient health state = latent variable



[22] A. Alaa and M. van der Schaar, 2019

Attentive state-space models

- **Key idea:** non-Markovian, non-stationary state transition

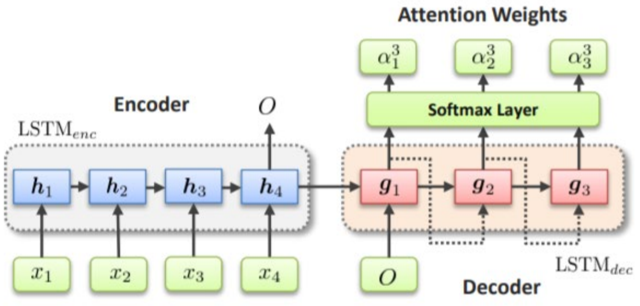


(a) RNN (b) HMM (c) Attentive state space

- **Attention mechanism:** assign different weights to previous states when computing transition probabilities to a new state at each time step

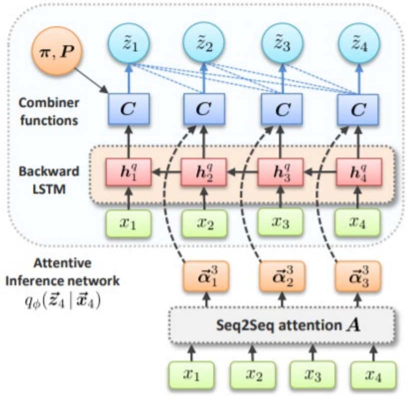
Attentive state-space models

- **Attention mechanism:** assign different weights to previous states when computing transition probabilities to a new state at each time step



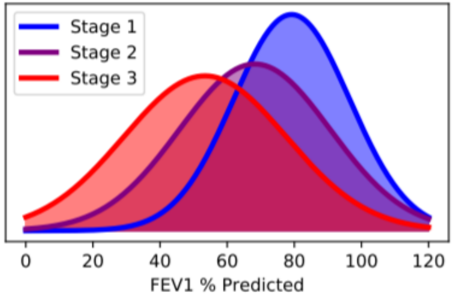
- **Encoder-decoder architecture**

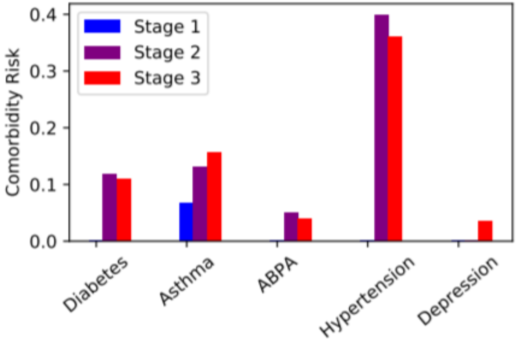
- **Variational inference:** inference network needs to be tailored to this particular model design…

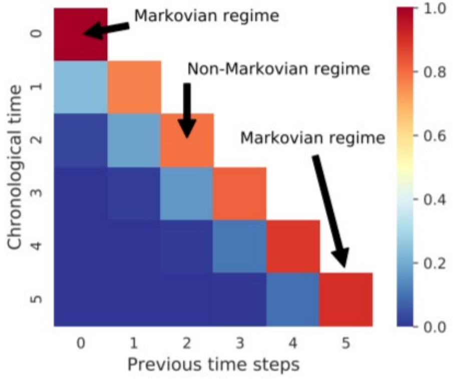


Attentive state-space models: Synthetic Cystic Fibrosis data

- **Not only able to generate synthetic data, but also explain disease progression dynamics both on a population level and patient level.**

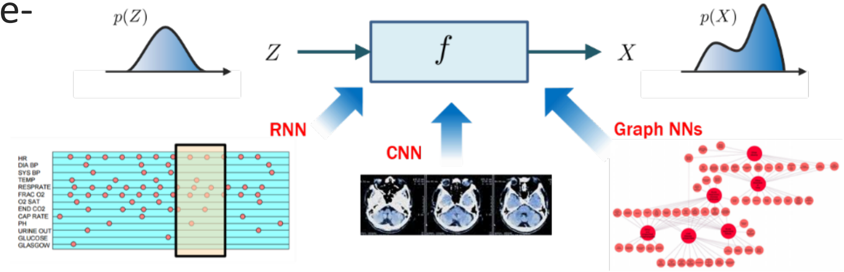






Recipe for generating synthetic healthcare data

- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.
- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., Autoregressive, attention, latent state-space, RNN representations, spectral, etc.
- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**



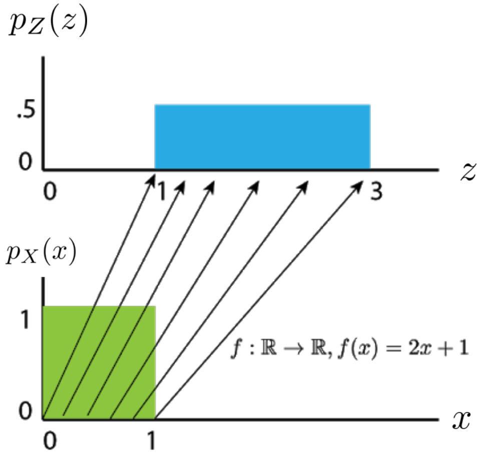
Fourier Flows

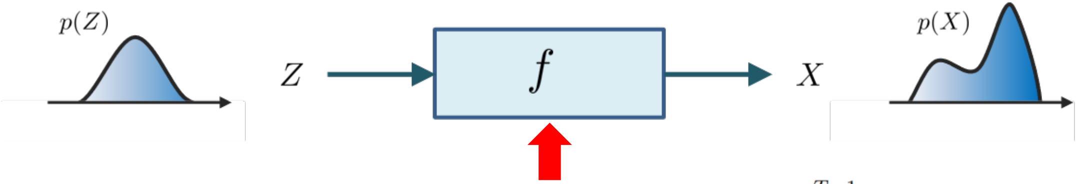
- **Recall: for normalizing flows, need to constrain the transformation so that:**
  - The Jacobean determinant is easy to compute

    Needed for efficient training

  - The forward and inverse map are easy to compute

    Needed for efficient sampling





Fourier transform

$$X_k = \sum_{t=0}^{T-1} x_t \cdot e^{-2\pi j \cdot \frac{kt}{T}}, \ \forall 1 \le k \le T-1$$

[23] A. Alaa and M. van der Schaar, 2021

Fourier Flows

- **Fourier transform has various favorable properties:**

  - DFT matrix

$$\bar{\boldsymbol{X}}^d = W\bar{\boldsymbol{x}}^d, \quad \omega = e^{-2\pi j/N}$$

$$W = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix}$$

  - Jacobean of the DFT matrix

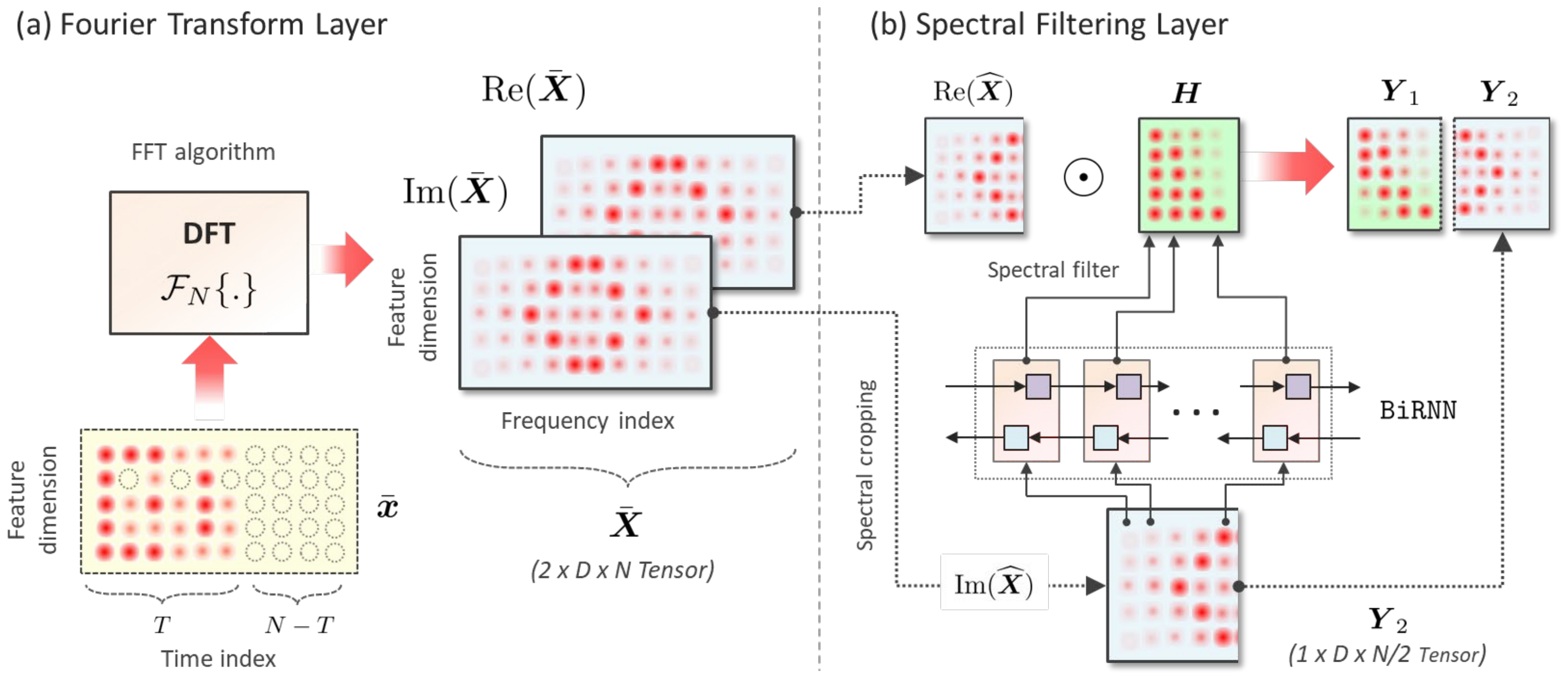
$$\begin{aligned} |\det(\boldsymbol{J}[W])| &= |\det(W)| \\ &= \left|\left(\frac{1}{\sqrt{N}}\right)^N \prod_{1<n<m\leq N} (\omega^m - \omega^n)\right| = 1 \end{aligned}$$

  - Convolution property

$$\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \iff \boldsymbol{X}_1 \odot \boldsymbol{X}_2$$

- **Handles variable-length time-series, variable sampling rates naturally**

Fourier Flows



Different modeling choices are appropriate for synthesizing different types of time-series data…

TimeGAN	Attentive State-space models	Fourier flows
Data-driven, high-dimensional data	Interpretable, handles domain knowledge	Periodic data (e.g., ECG), irregular and variable-length time-series

Evaluating synthetic healthcare data

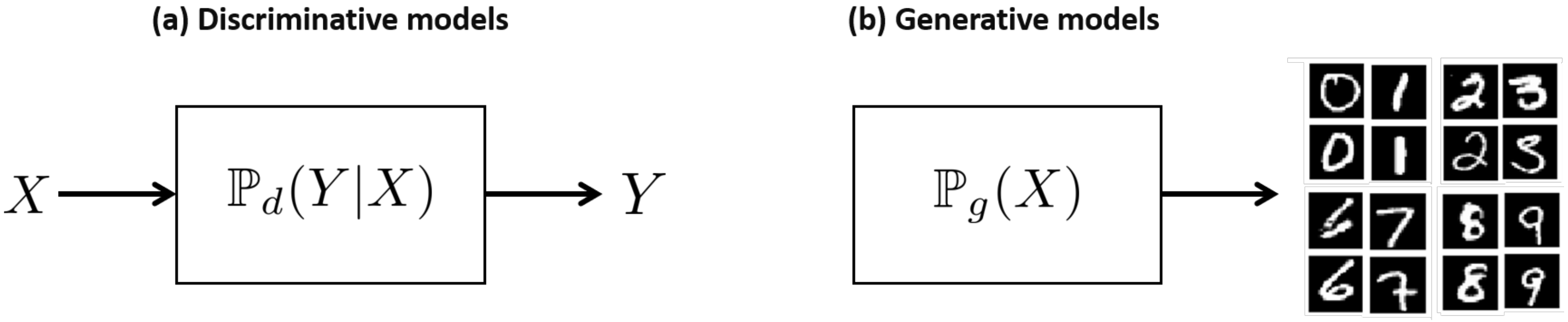
How to measure the quality of synthetic data sets?

- **We have followed the recipe and generated the desired synthetic data…**

- **How do we know if the synthetic data is of a high quality? What does "quality" mean?**
  - Inspect individual samples? We may have generated millions!
  - Train on synthetic, test on real? Does not tell us the full picture…
- **What makes a good synthetic dataset? How do we define "quality"?**

Evaluating generative models is tricky...

- Accuracy of **discriminative models** is easy to evaluate by simply assessing prediction accuracy on test data (e.g., cross-validation).
- **Generative models**: unsupervised, no ground-truth.



(a) Discriminative models

(b) Generative models

Model-dependent or domain-specific...

Likelihood and statistical divergence

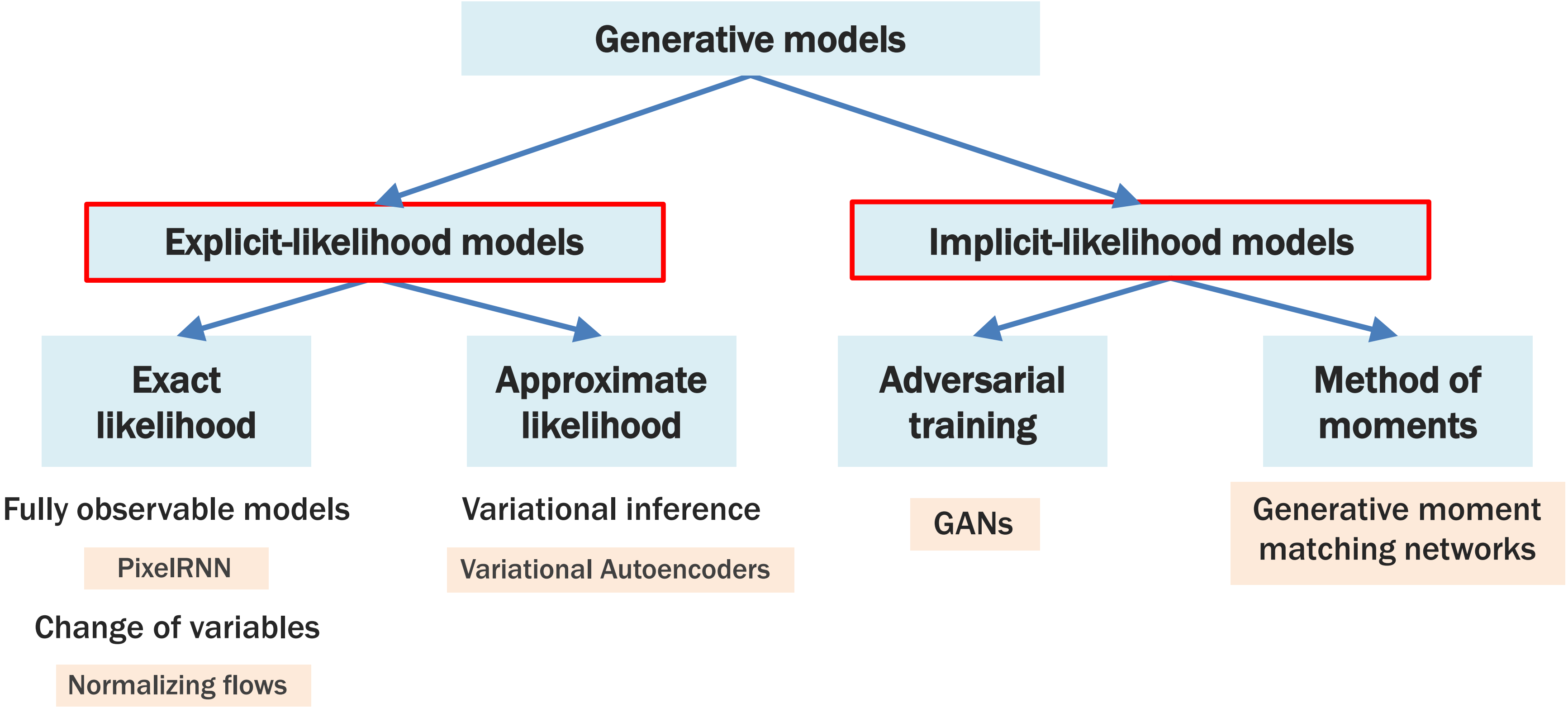
$$\log \mathbb{P}_g(X) \qquad D(\mathbb{P}_g || \mathbb{P}_r)$$

- State-of-the-art models do not have tractable likelihood (GANs, VAE, etc)
- Scale badly with high dimensions
- Single number that does not distinguish different modes of failure

Domain-specific scores

- Measures of the quality of images or audio
- Expert opinion
- Pre-trained representations on ImageNet (e.g. Fréchet Inception distance)

Recall: Not all models have explicit likelihoods!



Generative models

Explicit-likelihood models

Implicit-likelihood models

Exact likelihood

Approximate likelihood

Adversarial training

Method of moments

Fully observable models

PixelRNN

Change of variables

Normalizing flows

Variational inference

Variational Autoencoders

GANs

Generative moment matching networks

Model likelihood is not the right measure…

- **Likelihood is uninterpretable, does not evaluate individual samples and collapses all modes of failure into a single measure…**

- **Likelihood does not scale well with high-dimensional data…**

  - Great likelihood and poor samples: samples from a mixture of true model & noise

    $$0.01p(\mathbf{x}) + 0.99q(\mathbf{x})$$

    will have likelihood close to true model for high-dimensional data

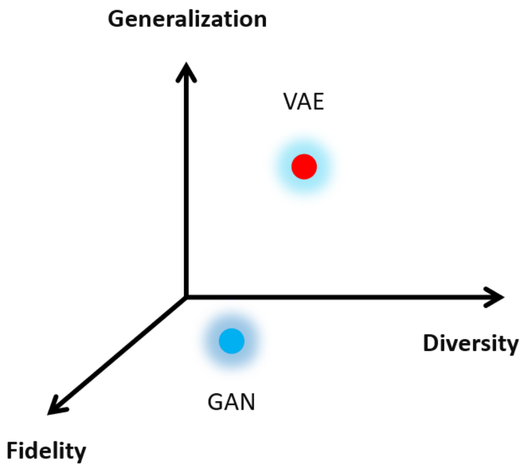
    $$\log\left[0.01p(\mathbf{x}) + 0.99q(\mathbf{x})\right] \geq \log\left[0.01p(\mathbf{x})\right] = \log p(\mathbf{x}) - \log 100$$

  - Poor likelihood and great samples: counter-examples exist.

L. Theis et al, 2016

Different ways in which a generative model may fail

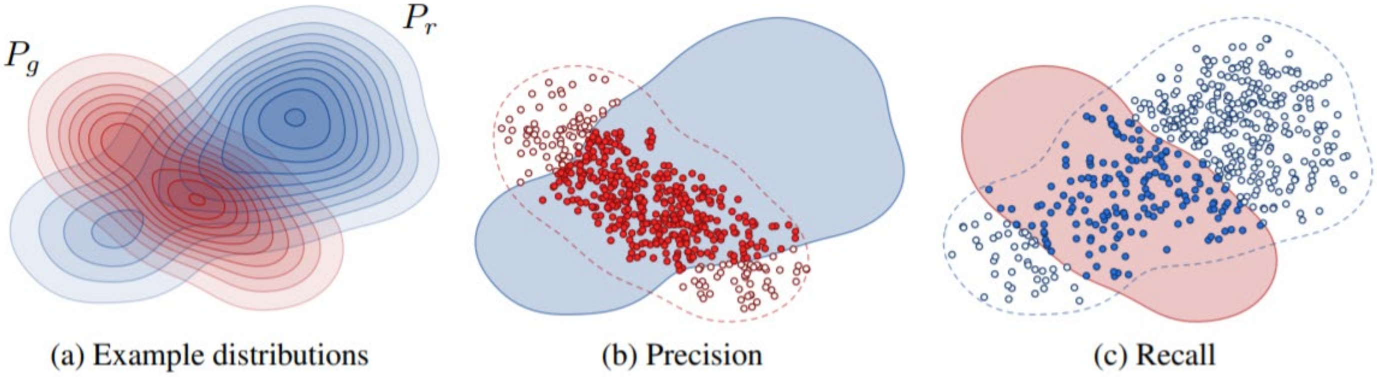
- A single-dimensional metric is not enough…
- Every model's performance can be viewed as a point in a 3D space
  - **Fidelity:** How "good" the synthetic samples are?
  - **Diversity:** How much of the real data is covered?
  - **Generalization:** How often does the model copy training data?
- **Need probabilistic, interpretable, multi-dimensional quantities**



Precision and Recall analysis

- **Precision:** the fraction of synthetic samples that look realistic (fidelity)
- **Recall:** fraction of real samples that the generative model can synthesize (diversity)

Need to estimate the support of real and synthetic distributions



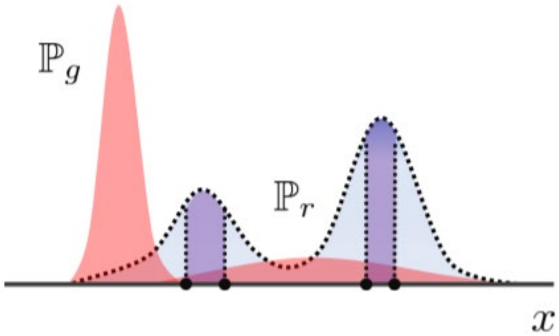
(a) Example distributions (b) Precision (c) Recall

[25] M. S. M. Sajjadi et al, 2018, Figure courtesy of: T. Kynkäänniemi et al, 2019

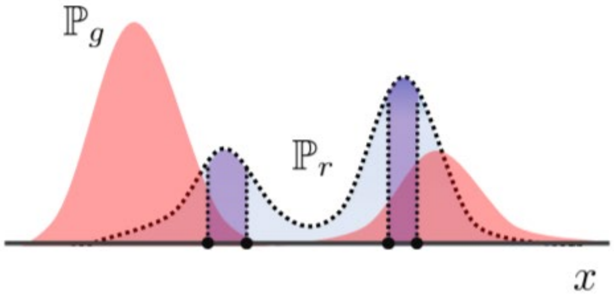
Refined Precision and Recall analysis

- **Problem with standard precision-recall analysis:**
  - What if the generative model only generates outliers in real distribution?
  - Can we detect mode collapse? Mode invention, etc?

Perfect precision score, poor generative distribution

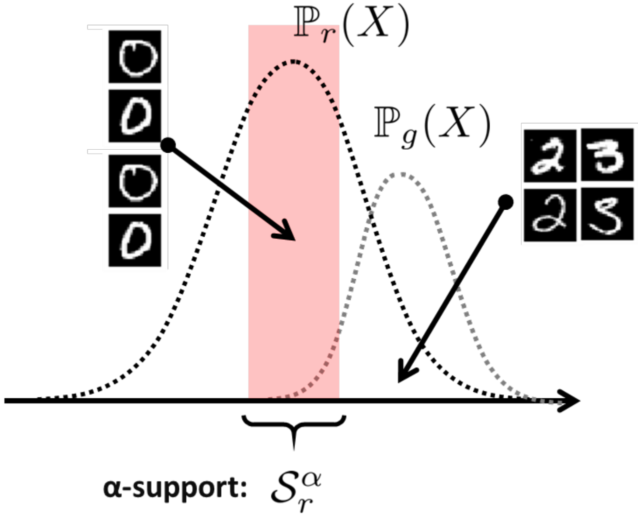


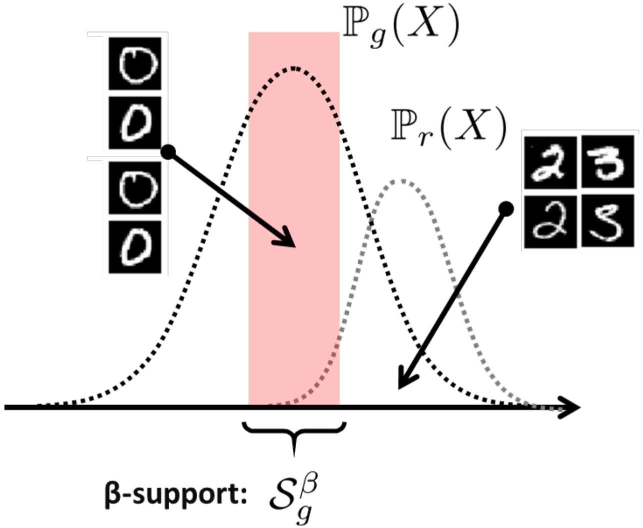
Poor precision score, better generative distribution



Refined Precision and Recall analysis

- Use a more refined definition of the support of a distribution…
  - **α-Precision** measures sample *fidelity*.
  - **β-Recall** measures sample *diversity*.

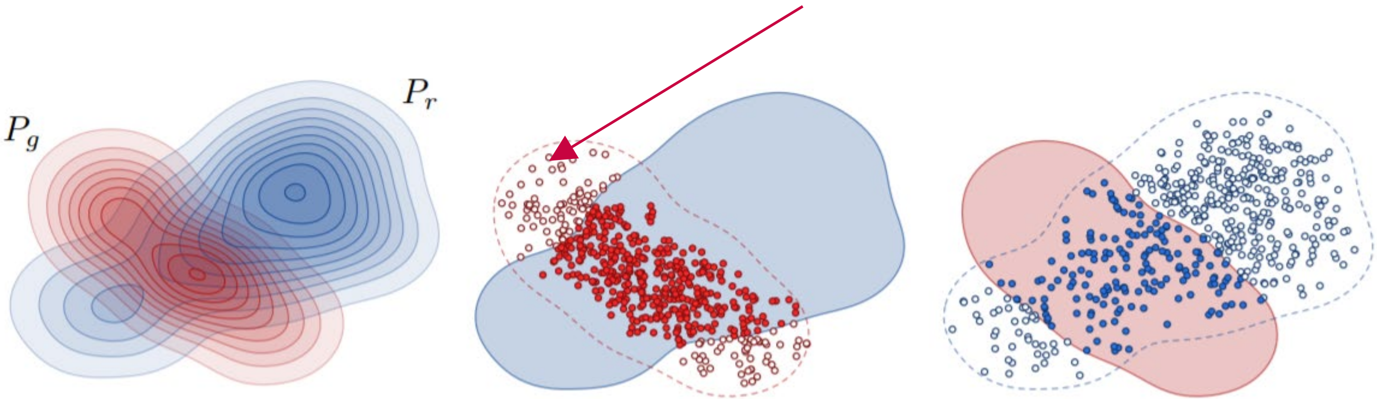




[26] A. Alaa, B. van Breugel, E. Saveliev, M. van der Schaar, 2021

Is my synthetic data clinically sensible?

- Precision automatically evaluates if each sample is “clinically realistic”

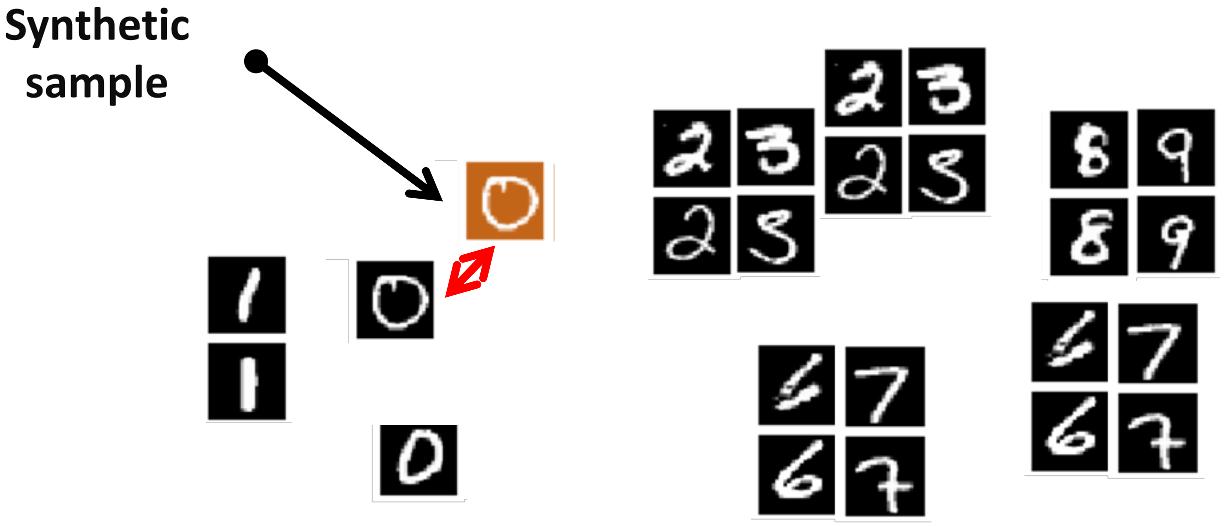


Age 120, systolic blood pressure 200

(a) Example distributions (b) Precision (c) Recall

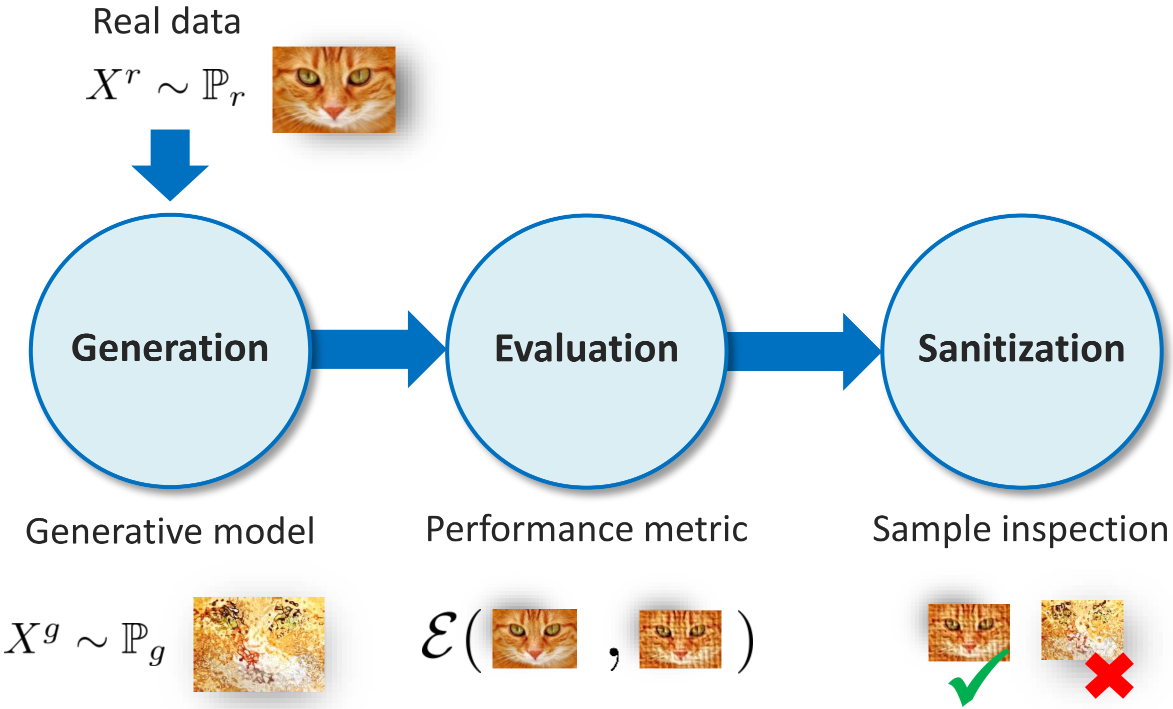
How to evaluate privacy preservation in synthetic data?

- Are their samples closer to training data than expected?



Sample-level metrics: a fourth step in the recipe

- **Sanitization:** **Remove samples that are copied or not precise**



Real data

$X^r \sim \mathbb{P}_r$

Generation

Evaluation

Sanitization

Generative model

Performance metric

Sample inspection

$X^g \sim \mathbb{P}_g$

$\mathcal{E}(\quad , \quad)$

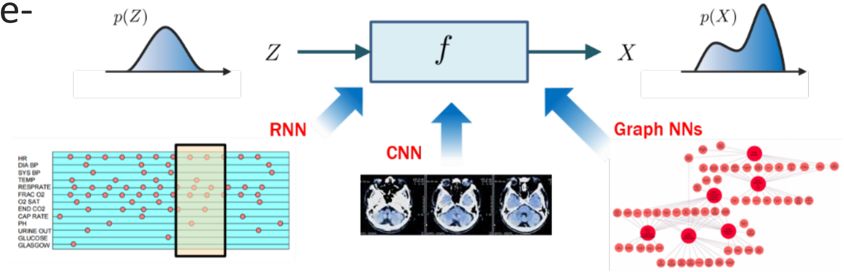
✓ ✗

Sample-level metrics: a fourth step in the recipe

- **Step 1: Decide the generative modeling approach to use…**
  - Variational auto-encoders, normalizing flows, GANs, etc.

- **Step 2: Construct an appropriate representation/structure for the type of healthcare data under consideration (time-series, images, notes, bio-markers, etc).**
  - E.g., Autoregressive, attention, latent state-space, RNN representations, spectral, etc.

- **Step 3: Incorporate differential privacy or other privacy notion/ sanitize the data to ensure privacy**



- **Step 4: Evaluate the model using sample-level metrics, remove "bad" samples…**

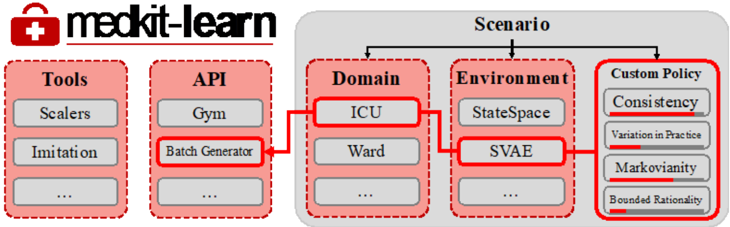
Future directions

Future directions

- Synthetic multi-modal data (genetic, images, time-series, etc.).
- Generative models for asynchronous, sparse follow-up clinic visits.
- Evaluating fairness and bias in synthetic data.
- Domain and disease-specific evaluation metrics.
- Disentangling clinical policies from real observational data.

Medkit – Synthetic Data for Decision Modelling

More than “just” synthetic data:
A full testing suite for decision modelling.



Challenges in healthcare:

- Fully offline with limited batch datasets
- Modern time-series generation methods don’t disentangle environment and policy

Desiderata:

- A collection of realistic environment models
- With a variety of expressive and customizable policy models

Achieved: Provide ground truth knowledge in complicated and realistic settings, ready to test decision making modelling tools (policies).

[26] A. Chan, I. Bica, A. Huyuk, D. Jarrett, M. van der Schaar, 2021

Why important?

Without ground-truth policy inference cannot be validated

Medkit provides policies learnt based on clinical behaviour but with customizable parameters, structured as:

$$Q_\pi^\Omega(y_t|\vec{x}_t, \vec{y}_{t-1}) = \sum_i w_i \frac{e^{\beta_i q_i(y_t|g_i(\vec{x}_t\langle\mathcal{X}'\rangle_i, \vec{y}_{t-1}))}}{\sum_{y\in\mathcal{Y}} e^{\beta_i q_i(y|g_i(\vec{x}_t\langle\mathcal{X}'\rangle_i, \vec{y}_{t-1}))}}$$

Where users can alter:

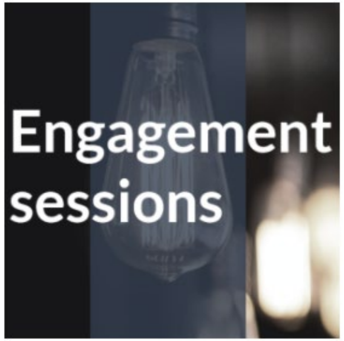
- Ground-truth Structure
- Markovianity
- Bounded Rationality
- Individual Consistency
- Variation in Practice

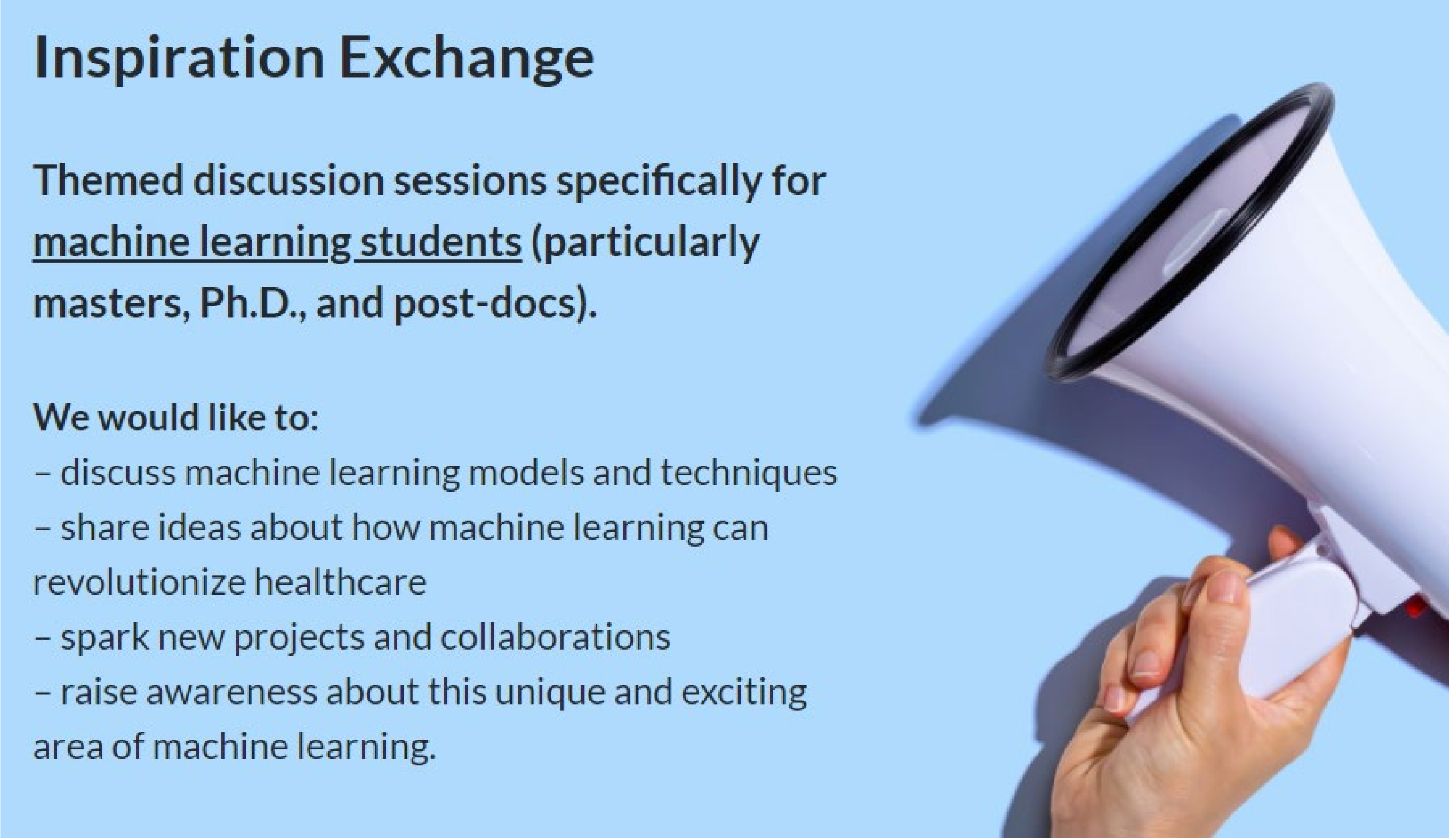
Or:
Bring your own completely custom policy!

Customizable Policies for Inference

Other Resources

- **Vision:** www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/
- Papers: www.vanderschaar-lab.com/publications/synthetic-data
- Software: www.vanderschaar-lab.com/software/





Inspiration Exchange

Themed discussion sessions specifically for machine learning students (particularly masters, Ph.D., and post-docs).

We would like to:
– discuss machine learning models and techniques
– share ideas about how machine learning can revolutionize healthcare
– spark new projects and collaborations
– raise awareness about this unique and exciting area of machine learning.

Join us!

www.vanderschaar-lab.com/engagement-sessions/inspiration-exchange/