Machine Learning for Healthcare

6.871, HST.956

Lecture 21: Privacy

David Sontag

HEALTH SCIENCES
& TECHNOLOGY

Today's lecture

- Overview of privacy-utility tradeoff in ML for healthcare
  - Differential privacy: how to train ML models that do not leak patient data
  - Synthetic data generation and its limitations

- Guest lecture by Fei Wang on federated learning

ML in healthcare needs lots of data

- Why might institutions be hesitant to share health records for research and commercialization?
- Data governance and ownership vs. security and privacy

Re-identification attacks are possible

- We saw in PS2 how NLP can be used to identify and hide personal health information (PHI)
- Latanya Sweeney showed that it is possible, with side-information, to re-identify patient records:

ZIP Code	*Birth Date*	*Gender*	*Race*
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian

Table 2. Deidentified Data that Are Not Anonymous.

Looks anonymous, right? But 02657 is Provincetown, MA where (in '97) five black women live year-round.

L. Sweeney. *Maintaining Patient Confidentiality When Sharing Medical Data Requires a Symbiotic Relationship Between Technology and Policy*. Artificial Intelligence Laboratory, Massachusetts Institute of Technology, AIWP-WP344, May 1997

L. Sweeney. Weaving Technology and Policy Together to Maintain Confidentiality. *Journal of Law, Medicine & Ethics*, 25, nos. 2&3 (1997): 98-110.

What else can we do, instead of releasing original data?

- Release statistics derived from the data
  - Must fudge to not reveal anything
  - Of limited utility for machine learning. Non-starter.
- Put data in a secure data enclave for R&D
- Release ML models derived from the data
  - How do we know these models do not reveal anything about the training data?
- Release synthetic data
  - How do we know it doesn't just reproduce the original training data?

Classifiers can reveal information about training data

- An attack called *model inversion* can be used to reverse engineer training data

- Example: dataset from International Warfarin Pharmacogenetics Consortium
  - Linear regression to predict initial dose outperforms standard clinical regimen
  - But… when one knows a target patient's background and stable dosage, their genetic markers could be predicted 22% more accurately than guessing based on marginal distributions

M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In USENIX Security Symposium, pages 17–32, 2014.

Classifiers can reveal information about training data

- An attack called *model inversion* can be used to reverse engineer training data

1. Input: $\mathbf{z}_K = (x_1, \ldots, x_k, y), f, p_{1,\ldots,d,y}$
2. Find the *feasible set* $\hat{\mathbf{X}} \subseteq \mathbf{X}$, i.e., such that $\forall \mathbf{x} \in \hat{\mathbf{X}}$
   (a) $\mathbf{x}$ matches $\mathbf{z}_K$ on known attributes: for $1 \leq i \leq k, \mathbf{x}_i = x_i$.
   (b) f evaluates to y as given in $\mathbf{z}_K$: $f(\mathbf{x}) = y$.
3. If $|\hat{\mathbf{X}}| = 0$, return $\perp$.
4. Return x_t that maximizes $\sum_{\mathbf{x} \in \hat{\mathbf{X}}: \mathbf{x}_t = x_t} \prod_{1 \leq i \leq d} p_i(\mathbf{x}_i)$

M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In USENIX Security Symposium, pages 17–32, 2014.

Classifiers can reveal information about training data

- An attack called *model inversion* can be used to reverse engineer training data

Algorithm 1 Inversion attack for facial recognition models.

function MI-FACE($label, \alpha, \beta, \gamma, \lambda$)
  $c(\mathbf{x}) \stackrel{\text{def}}{=} 1 - \tilde{f}_{label}(\mathbf{x})$
  $\mathbf{x}_0 \leftarrow \mathbf{0}$
  for $i \leftarrow 1 \ldots \alpha$ **do**
    $\mathbf{x}_i \leftarrow \text{PROCESS}(\mathbf{x}_{i-1} - \lambda \cdot \nabla c(\mathbf{x}_{i-1}))$
    if $c(\mathbf{x}_i) \geq \max(c(\mathbf{x}_{i-1}), \ldots, c(\mathbf{x}_{i-\beta}))$ **then**
      break
    if $c(\mathbf{x}_i) \leq \gamma$ **then**
      break
  return $[\arg\min_{\mathbf{x}_i}(c(\mathbf{x}_i)), \min_{\mathbf{x}_i}(c(\mathbf{x}_i))]$

Figure 7: Reconstruction without using Process-DAE (Algorithm 2) (left), with it (center), and the training set image (right).

M. Fredrikson, S. Jha, T. Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. CCS '15.

Differentially private machine learning

- An algorithm is **differentially private** if its output is statistically indistinguishable when applied to two input datasets that differ by only one record in the dataset
- One way to achieve is via differentially private stochastic gradient descent (DP-SGD):

Algorithm 1 Differentially private SGD (Outline)

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

Initialize θ_0 randomly

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

Take a random sample L_t with sampling probability L/N

Compute gradient

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

Clip gradient

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

Add noise

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

Descent

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

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

Abadi et al., Deep Learning with Differential Privacy. CCS 2016.

The privacy-utility trade-off

- Evaluate using the following datasets:

Dataset	Data Type	Outcome Variable	n	d	Classification Task	Tail Size
health care						
mimic_mortality	Time Series	in-ICU mortality	21,877	(24,69)	Binary	Large
mimic_los_3	Time Series	length of stay > 3 days	21,877	(24,69)	Binary	Small
mimic_intervention	Time Series	vasopressor administration	21,877	(24,69)	Multiclass (4)	Small
NIH_chest_x_ray	Imaging	multilabel disease prediction	112,120	(256,256)	Multiclass multilabel (14)	Largest
Vision Baselines						
mnist	Imaging	number classification	60,000	(28,28)	Multiclass (10)	None
fashion_mnist	Imaging	clothing classification	60,000	(28,28)	Multiclass (10)	None

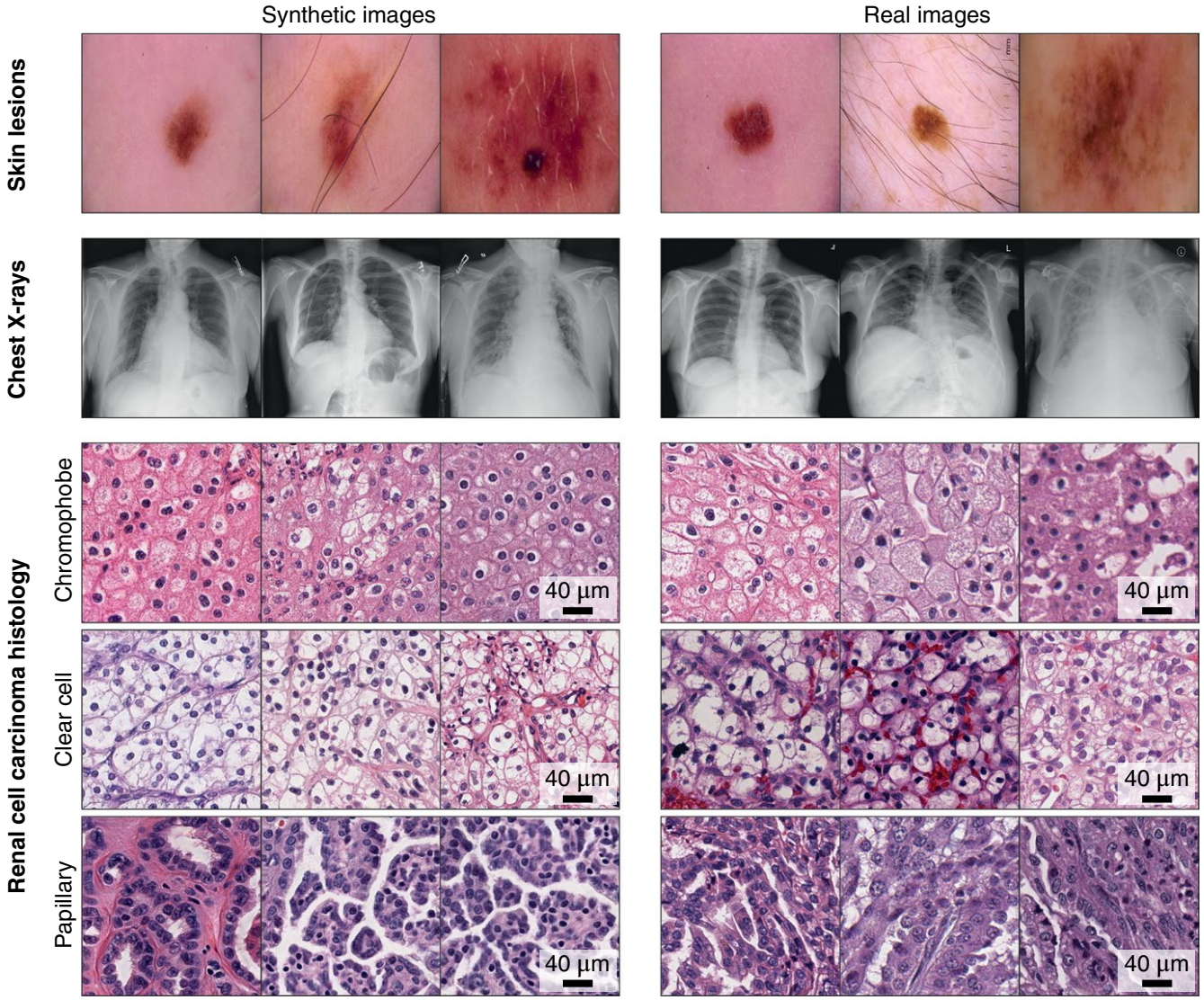
Suriyakumar, Papernot, Goldenberg, Ghassemi. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings, FAccT ‘21

The privacy-utility trade-off

Vision Baselines				
Dataset	Model	None (ϵ, δ)	Low (ϵ, δ)	High (ϵ, δ)
MNIST	CNN	98.83 ± 0.06 $(\infty, 0)$	98.58 ± 0.06 $(2.6 \cdot 10^5)$	93.78 ± 0.25 (2.01)
FashionMNIST	CNN	87.92 ± 0.19 $(\infty, 0)$	87.90 ± 0.16 $(2.6 \cdot 10^5)$	79.53 ± 0.10 (2.01)
MIMIC-III				
Task	Model	None (ϵ, δ)	Low (ϵ, δ)	High (ϵ, δ)
Mortality	LR	0.82 ± 0.03 $(\infty, 0)$	0.76 ± 0.05 $(3.50 \cdot 10^5, 10^{-5})$	0.60 ± 0.04 $(3.54, 10^{-5})$
Length of Stay > 3	LR	0.69 ± 0.02 $(\infty, 0)$	0.66 ± 0.03 $(3.50 \cdot 10^5, 10^{-5})$	0.60 ± 0.04 $(3.54, 10^{-5})$
Intervention Onset (Vaso)	LR	0.90 ± 0.03 $(\infty, 0)$	0.87 ± 0.03 $(1.63 \cdot 10^7, 10^{-5})$	0.77 ± 0.05 $(0.94, 10^{-5})$
NIH Chest X-Ray				
Metric	Model	None (ϵ, δ)	Low (ϵ, δ)	High (ϵ, δ)
Average AUC	DenseNet-121	0.84 ± 0.00 $(\infty, 0)$	0.51 ± 0.01 $(1.74 \cdot 10^5, 10^{-6})$	0.49 ± 0.00 $(0.84, 10^{-6})$
Best AUC	DenseNet-121	0.98 ± 0.00 (Hernia)	0.54 ± 0.04 (Edema)	0.54 ± 0.05 (Pleural Thickening)
Worst AUC	DenseNet-121	0.72 ± 0.00 (Infiltration)	0.48 ± 0.02 (Fibrosis)	0.47 ± 0.02 (Pleural Thickening)

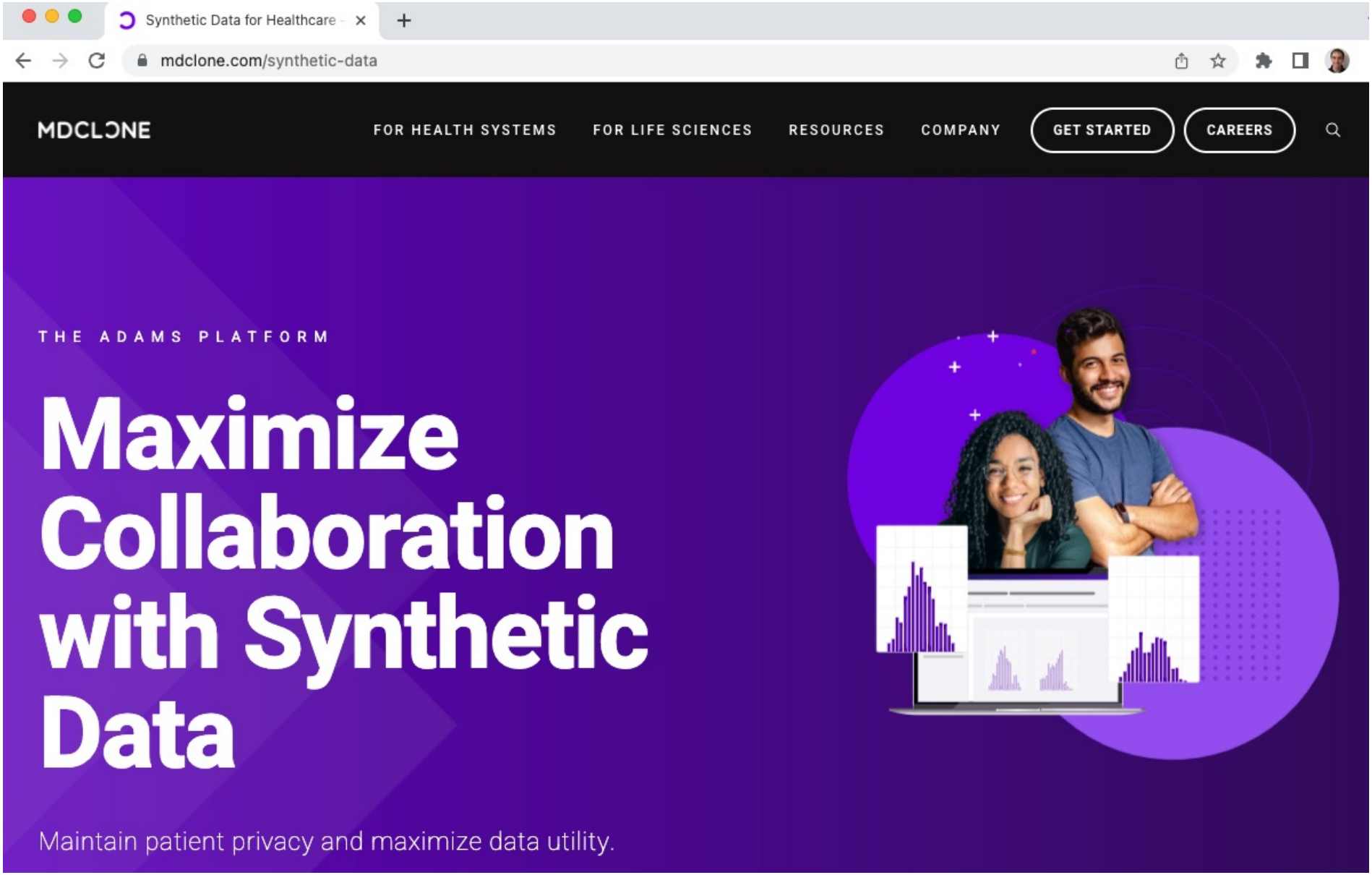
Suriyakumar, Papernot, Goldenberg, Ghassemi. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings, FAccT '21

Synthetic data generation



Chen, Lu, Chen, Williamson, Mahmood. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 2021.

Synthetic data generation



Synthetic data generation

- Key questions to ask are:
  - Can you do more with the synthetic data than you could have with just basic statistics derived from the data?
  - What does the synthetic data leak about the original training data?

- Many recent works applying differential privacy methods to training of generative adversarial networks
  - What are the privacy-utility trade-offs?