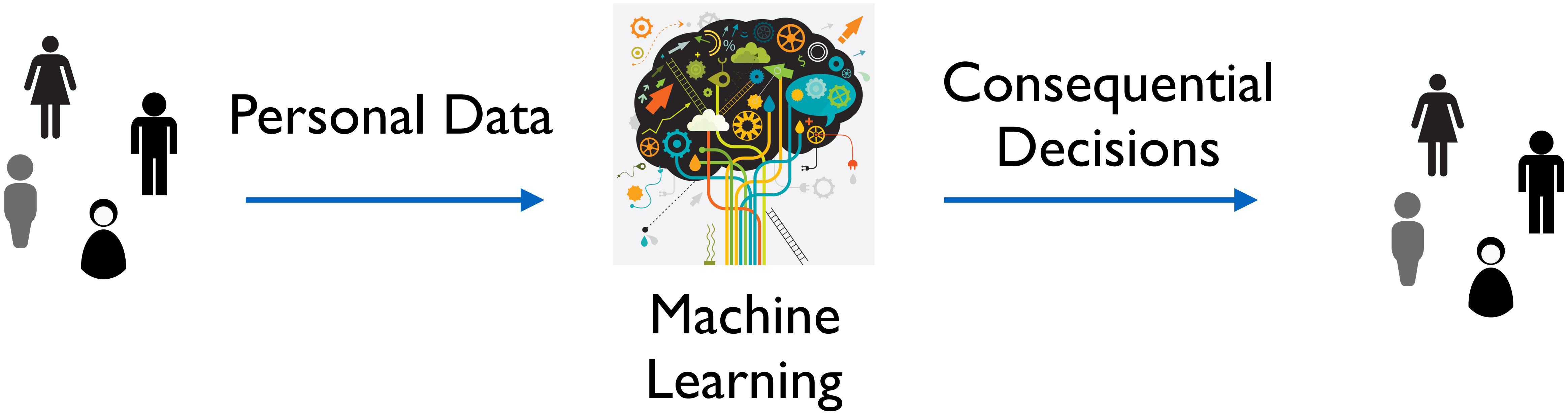
Privacy-Preserving Synthetic Data

Steven Wu

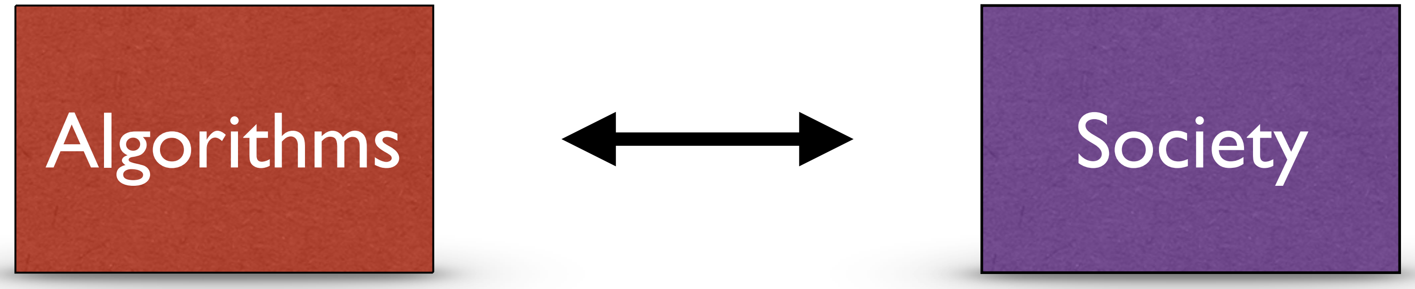
Assistant Professor
Institute for Software Research



Personal Data

Machine Learning

Consequential Decisions



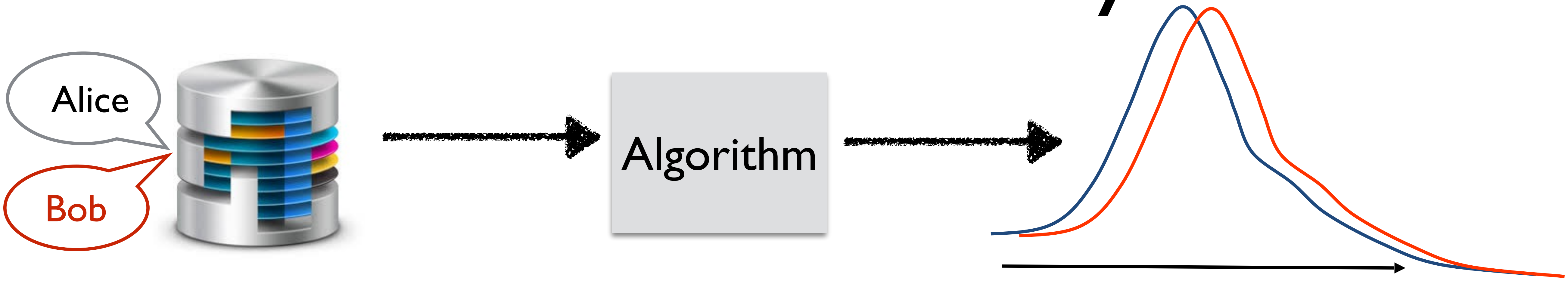
Algorithms

Society

How can we make machine learning better aligned with *societal values*?

Focus: *privacy* and *fairness*

Differential Privacy


Alice

Bob

Algorithm

Sensitive Data Set
(e.g., medical records)

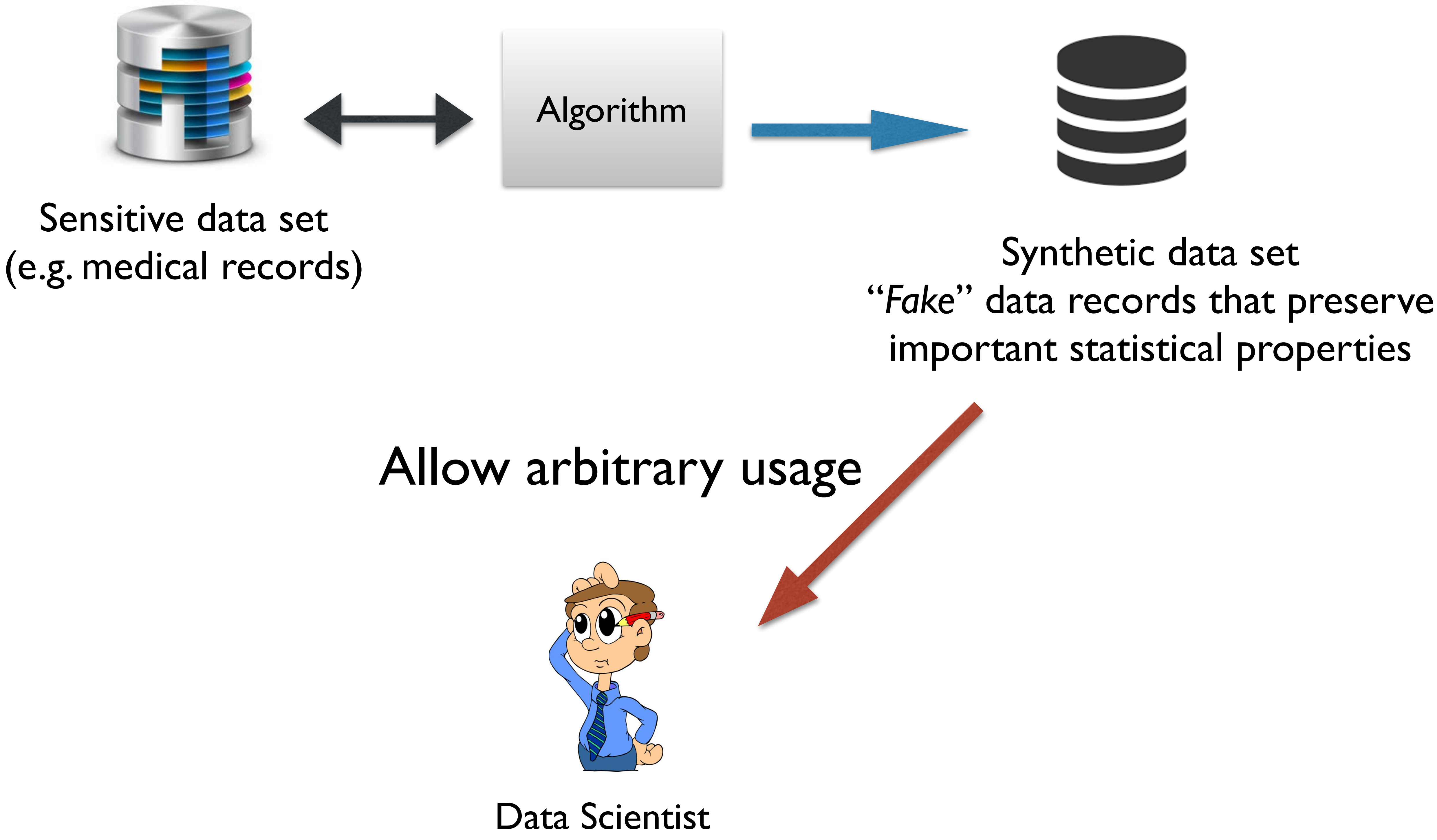
Output Distribution
(e.g., noisy statistics)

"An algorithm is *differentially private* if changing a single record does not alter its output distribution by much."
[DN03, DMNS06]

Challenge in Adoption:

How to facilitate non-privacy experts to work with differential privacy?

Differentially Private Synthetic Data



Algorithm

Sensitive data set
(e.g. medical records)

Synthetic data set
"*Fake*" data records that preserve
important statistical properties

Allow arbitrary usage

Data Scientist

Privacy-Preserving GANs Support Clinical Data Sharing

[BWWLBBG]

Published in *Circulation: Cardiovascular Quality and Outcomes 2019*

Data Set

Systolic Blood Pressure Intervention Trial (SPRINT)

- 9,361 patients (3 measurements over 12 periods)

Approach

Generative adversarial nets (GANs)

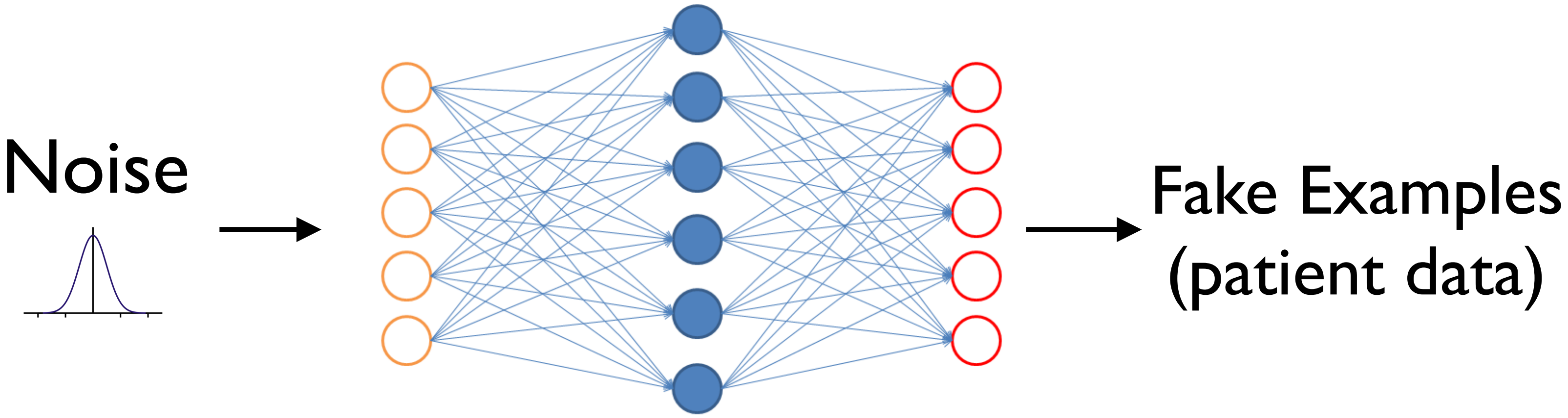
+ Differential privacy

Generative Adversarial Nets (GANs)

[GPM+14]

2-Player Zero-Sum Game

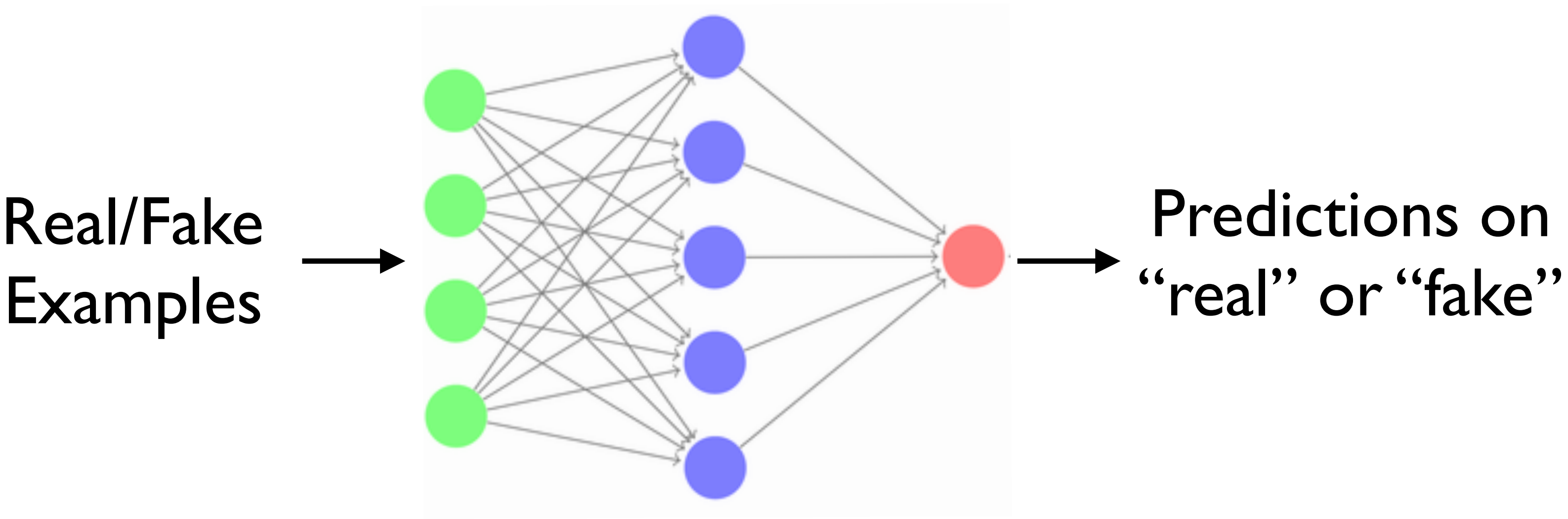
Generator:
mimic the real data



Noise

Fake Examples
(patient data)

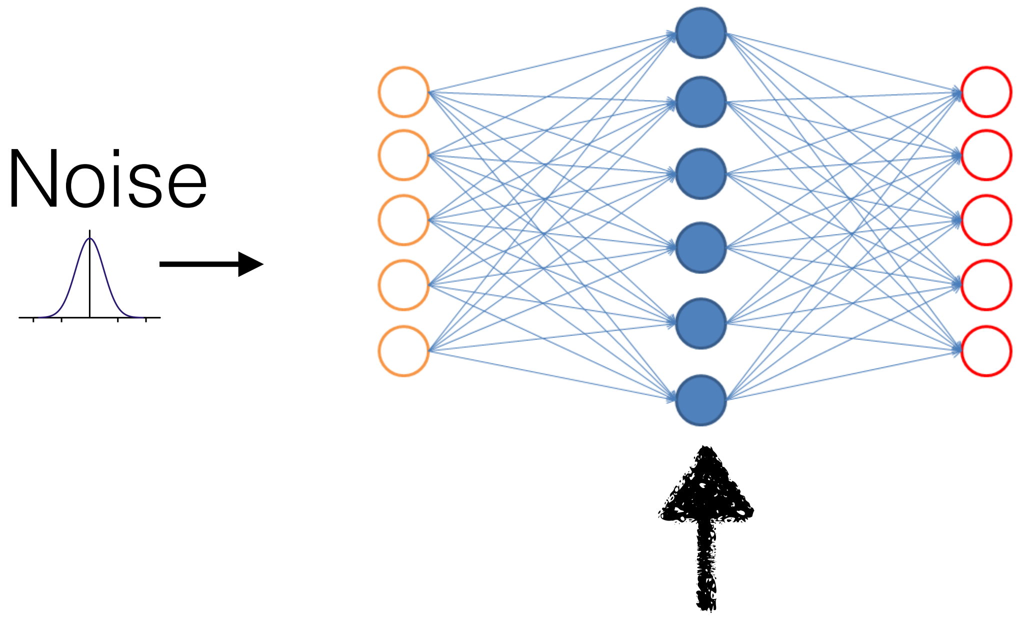
Discriminator:
distinguish real and fake data

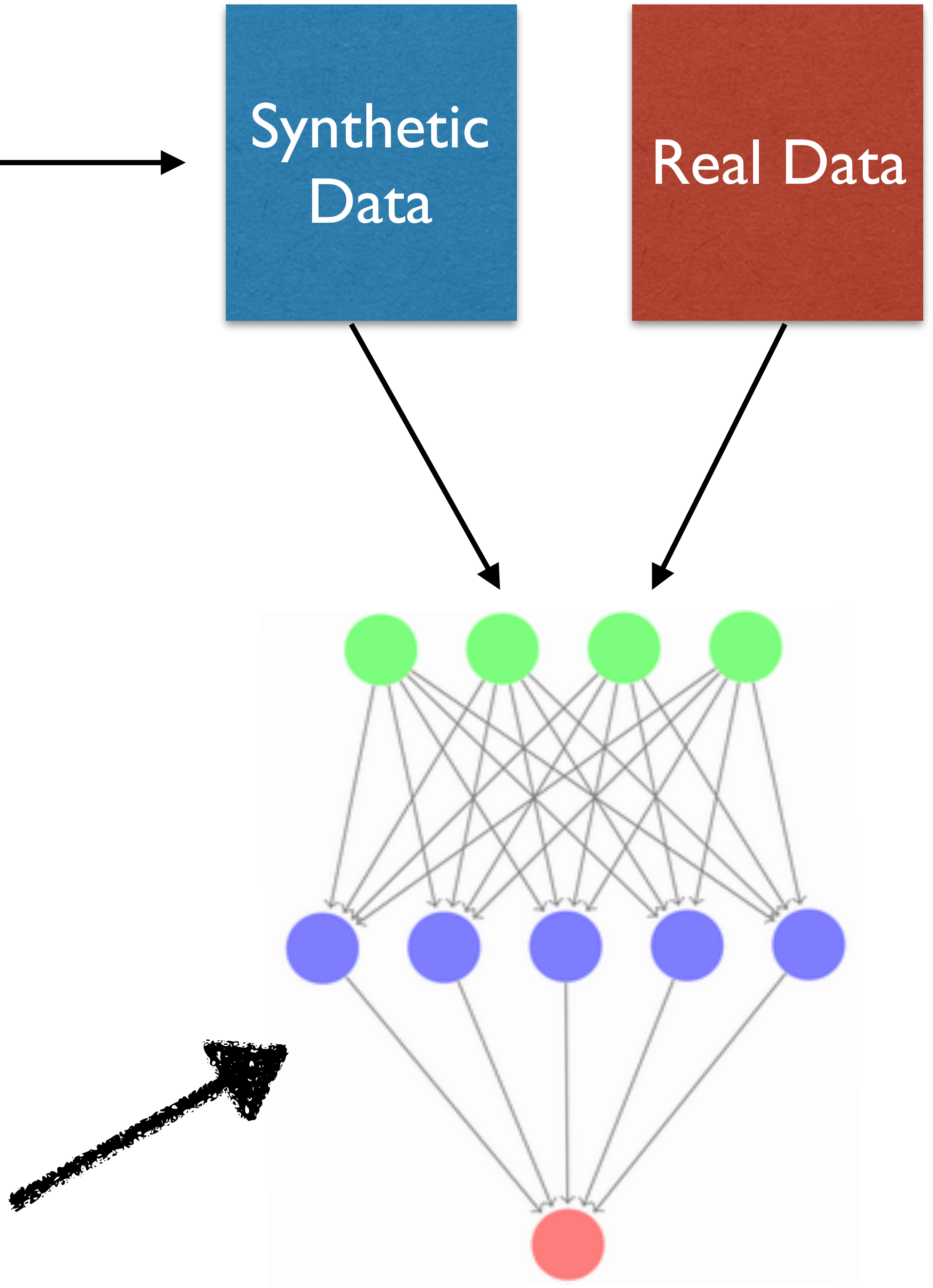


Real/Fake
Examples

Predictions on
“real” or “fake”

Private GAN Training [BWW+19]





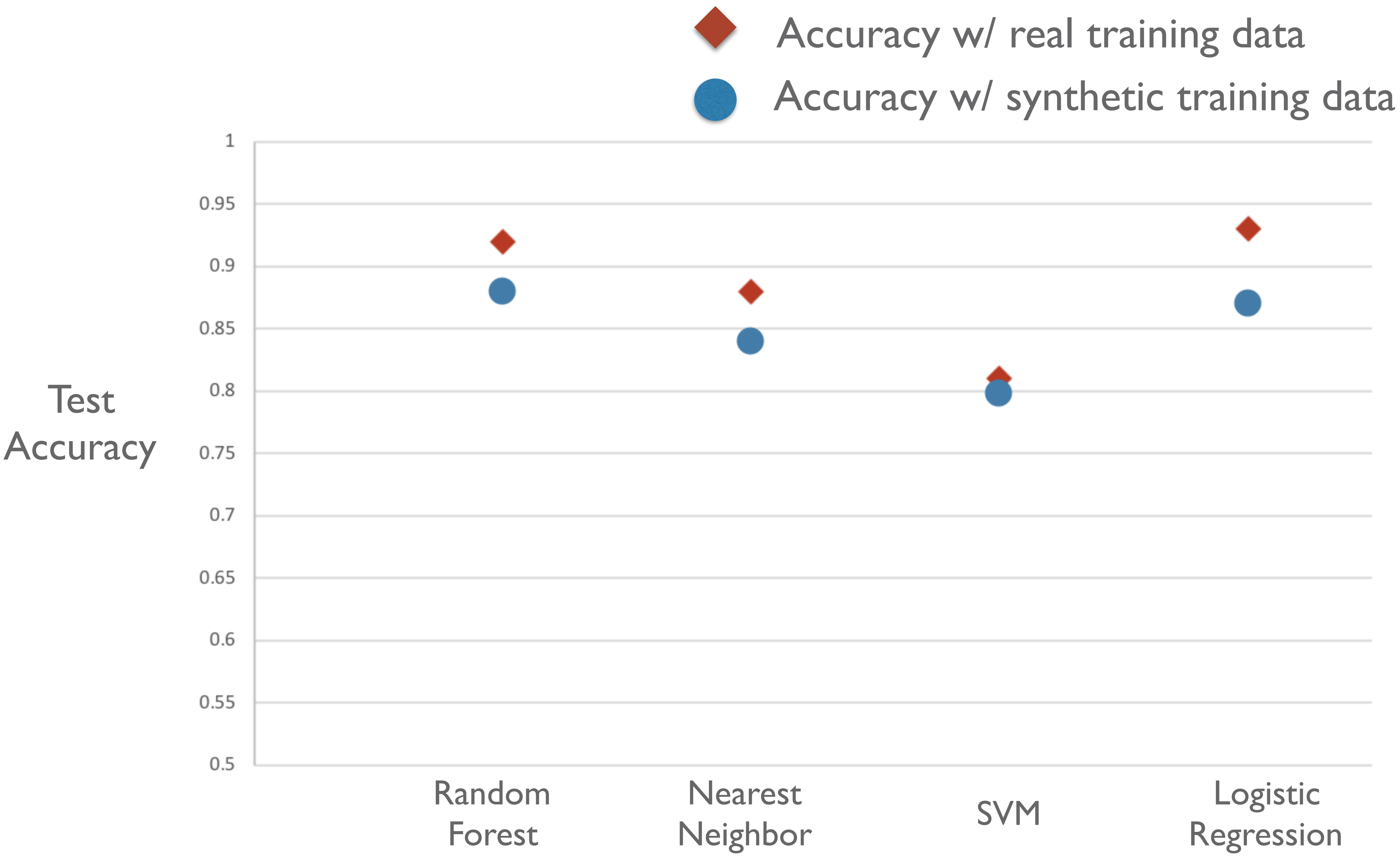
Training Generator:

- Does not directly interact with real data
- Train using standard (non-private) methods (e.g., SGD)

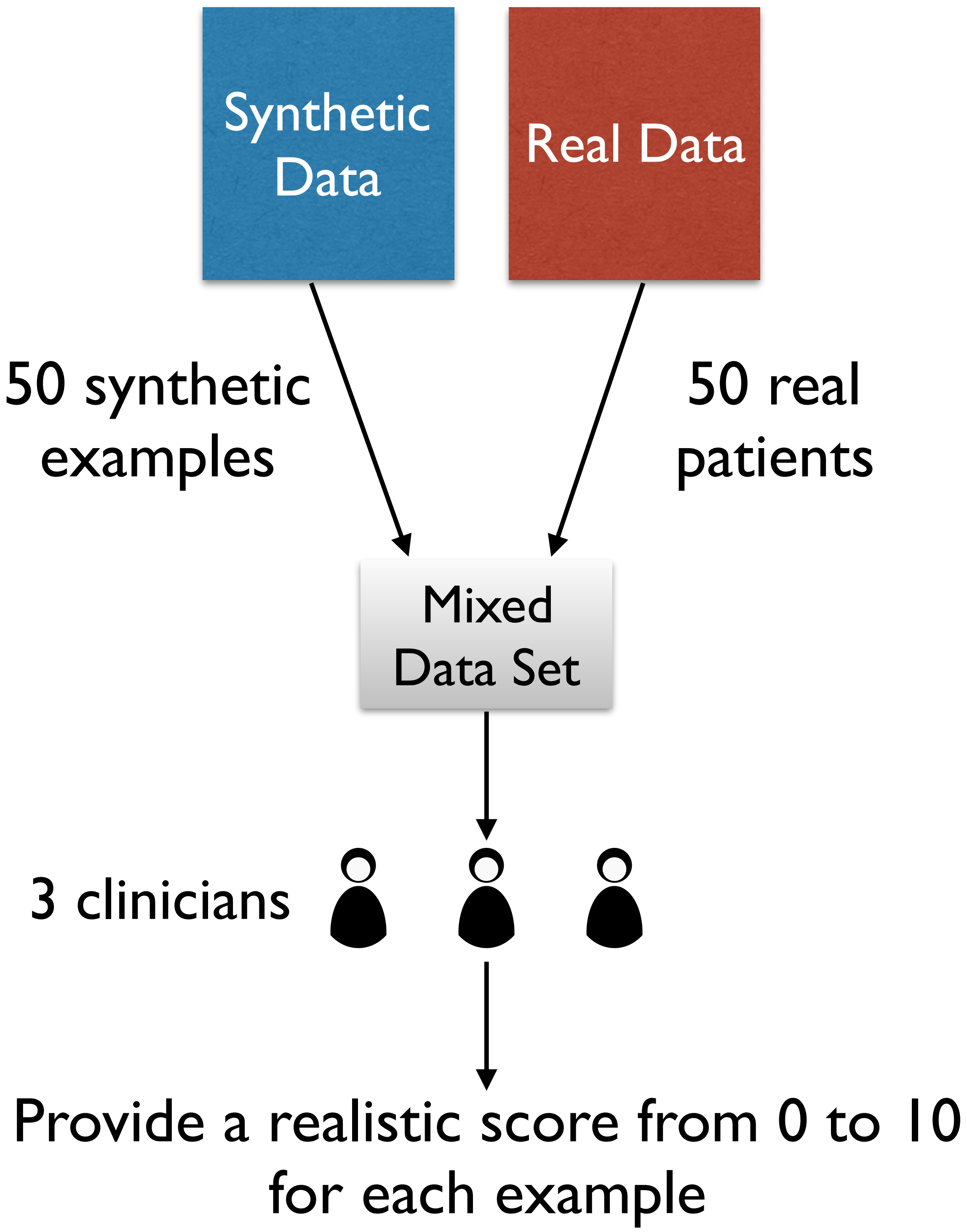
Privately Training Discriminator:

- Interacts with real data
- Train using *differentially private* SGD method
  - Gradient clipping + Gaussian Perturbation

Models Trained on Synthetic v.s. Real Data



Evaluation with Human (Discriminators)



Synthetic Data

Real Data

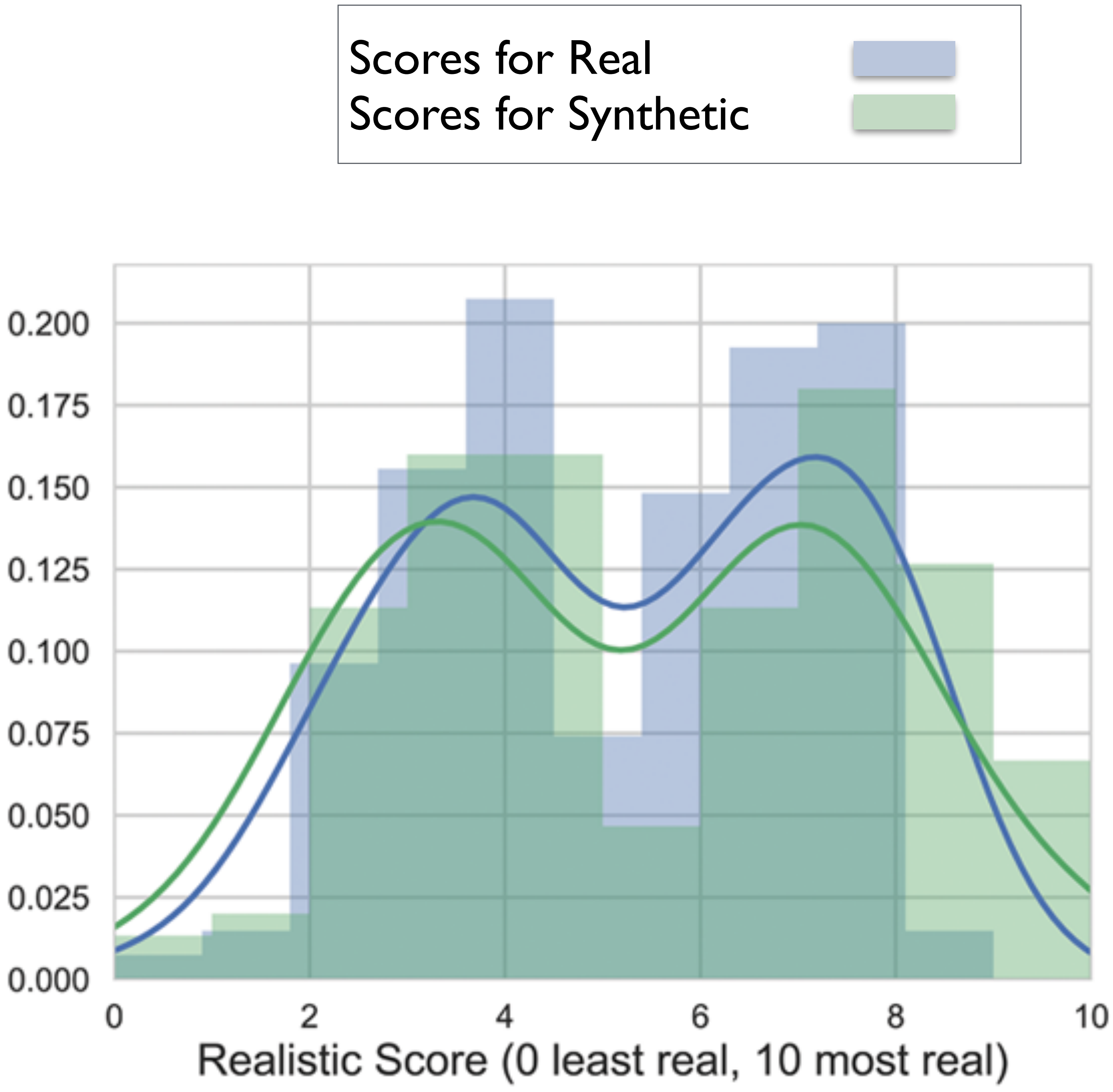
50 synthetic examples

50 real patients

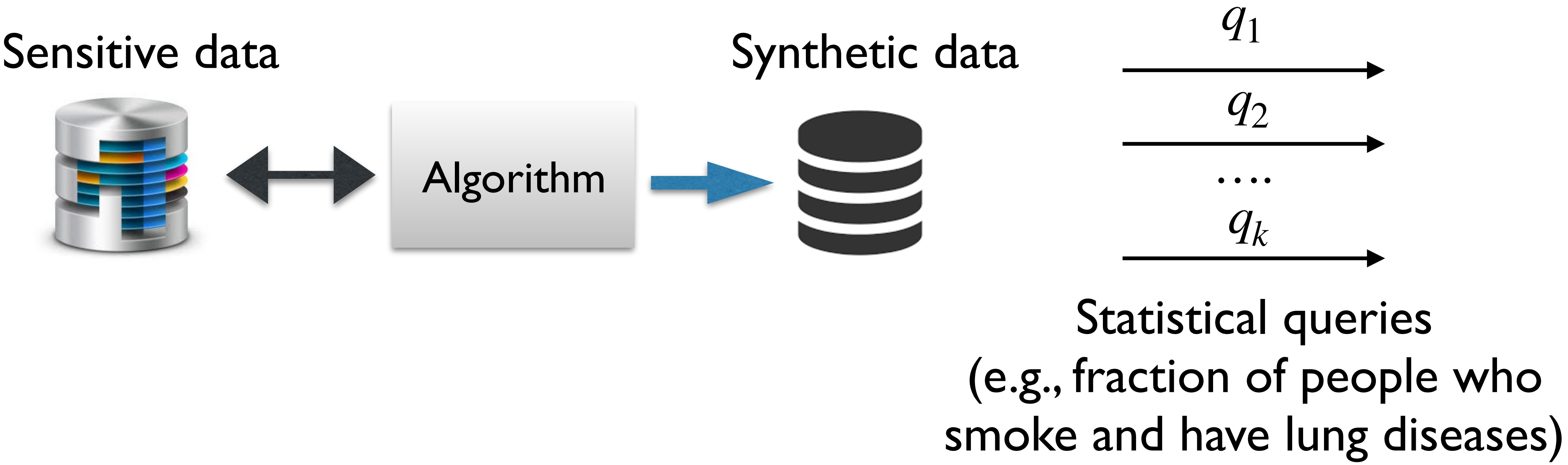
Mixed Data Set

3 clinicians

Provide a realistic score from 0 to 10 for each example



Synthetic Data for Query Release



Statistical queries
(e.g., fraction of people who
smoke and have lung diseases)

Fast algorithms by leveraging off-the-shelf solvers (e.g., Gurobi, CPLEX)

- [GGHRW] *ICML14*; [NRW] *FOCS19*
- [VTBSW] *ICML20*

Privacy-Preserving Synthetic Data

Steven Wu

zstevenwu.com