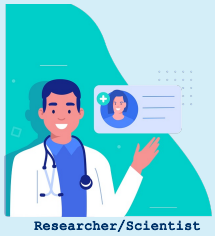
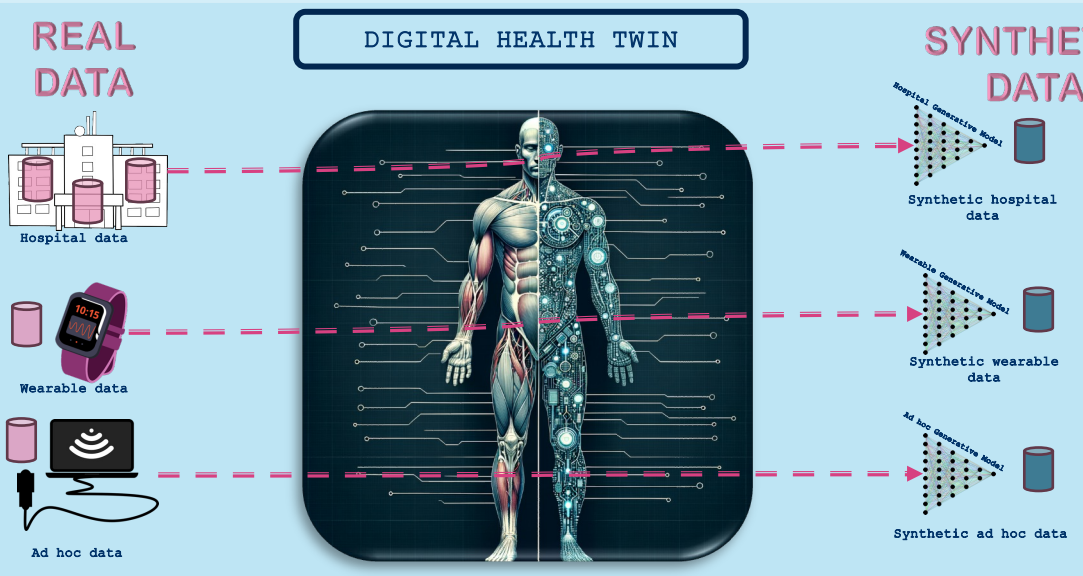
PROBLEM

- Researchers/Scientists need to access to patient data
  - Patient data is *distributed* and cannot be accumulated
  - Patient data is *private*
  - *Regulatory/Policy* requirements
- Federated learning solves some issues
  - *Privacy* is still not guaranteed
  - Collaborative nature of federated learning is susceptible to *security* risks
  - Federated learning has a *performance cost*
  - Non-IID distribution leads to *loss of performance/fairness*



SOLUTION

- Using synthetic data instead of real data
  - Train *private generative models* – e.g. VAEs, GANs, Normalizing Flows, etc. - on real data
  - *Validate* the model and its results
  - Generate private synthetic data to *replace the real data*
- Benefits over using real data
  - Guaranteed *privacy*
  - *Avoiding the complexity* of distributed/federated training
  - *Avoiding security risks* of federated training
  - Ability to *manipulate and augment the synthetic data* and fix some of its issues
  - Potential *probability evaluation* of the samples leading to uncertainty estimation, anomaly detection, etc.



ADVANCING TABULAR SYNTHETIC DATA GENERATION FOR CRITICAL DOMAINS

Synthesizing Heavy and Mixed Tail Data

- Real world data have a mix of heavy and light tail behavior
- The generative model needs to capture the tail behavior of the real data accurately
- It is necessary for correctly generate rare cases without generating non-existent anomalous data
- Most of the available literature do not address this issue directly

- Our solution: make Normalizing Flows tail adaptive without any pre-assumptions about the tail behavior of the target density

- Results:
  - Amiri, Saba, et al. "Compressive differentially private federated learning through universal vector quantization." AAAI Workshop on Privacy -Preserving Artificial Intelligence. 2021.
  - Amiri, Saba, et al. "On the impact of non-IID data on the performance and fairness of differentially private federated learning." 2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2022.
  - Amiri, Saba, et al. "Generating Heavy-Tailed Synthetic Data with Normalizing Flows." The 5th Workshop on Tractable Probabilistic Modeling. 2022.
  - Amiri, Saba, et al. " Practical Synthesis of Mixed-Tailed Data with Normalizing Flows", working paper

Synthesizing Semantically Correct Data

- Probabilistic generative models for synthesizing data
- These models could potentially generate samples that are in the support of their estimated distribution but are semantically incorrect, e.g. a male patient with pregnancy history
- The semantic rules are mostly either undocumented or unidentified
- A way to model the semantic boundaries of the variables in an unsupervised manner
- Use extracted boundaries to guide the generative model during the training/inference phase

- Our solution: add an independent validator component to the data synthesizer to model and enforce semantic rules

- Results:
  - Amiri, Saba, et al. "Differential Privacy vs Detecting Copyright Infringement: A Case Study with Normalizing Flows.", Gen Law 23, ICML
  - Amiri, Saba, et al. "Synthesizing Tabular Data with Regularized Latent Representations for Improved Semantic Integrity", working paper

Synthesizing Private Data

- Models trained on real data are vulnerable to adversarial attacks such as membership inference, could potentially leak training data
- Synthetic data could potentially let an adversary gain information about the training set and/or reconstruct it
- We need privacy preserving generative models with provable privacy guarantees
- We aim to make the generative models differentially private
- Differential privacy has a performance and/or fairness cost

- Our solution: Noiseless differentially private normalizing flows

- Results:
  - Amiri, Saba, et al. " Noise-less differentially private normalizing flows for tabular data synthesis", working paper