

Personalized Federated Training of Diffusion Models with Privacy Guarantees

Kumar Kshitij Patel, Institute for Foundations of Data Science (FDS), Yale University

Big Data Summer Immersion at Yale (BDSY) 2026

July 15, 2026

A Bit About Me

- I am Postdoctoral Associate at Yale University's Institute for Foundations of Data Science (FDS)

A Bit About Me

- I am Postdoctoral Associate at Yale University's Institute for Foundations of Data Science (FDS)
- Previously I was a fellow at the Simons Institute for the Theory of Computing, University of Berkeley

A Bit About Me

- I am Postdoctoral Associate at Yale University's Institute for Foundations of Data Science (FDS)
- Previously I was a fellow at the Simons Institute for the Theory of Computing, University of Berkeley
- I finished my PhD last year at Toyota Technological Institute, Chicago

A Bit About Me

- I am Postdoctoral Associate at Yale University's Institute for Foundations of Data Science (FDS)
- Previously I was a fellow at the Simons Institute for the Theory of Computing, University of Berkeley
- I finished my PhD last year at Toyota Technological Institute, Chicago
 - A philanthropically endowed computer science research institute on the University of Chicago campus

A Bit About Me

- I am Postdoctoral Associate at Yale University's Institute for Foundations of Data Science (FDS)
- Previously I was a fellow at the Simons Institute for the Theory of Computing, University of Berkeley
- I finished my PhD last year at Toyota Technological Institute, Chicago
 - A philanthropically endowed computer science research institute on the University of Chicago campus
- Before that I finished my BTech in Computer Science and Engineering at Indian Institute of Technology (IIT), Kanpur

A Bit About My Research

I work on the theory of Machine Learning

A Bit About My Research

I work on the theory of Machine Learning

- I study learning when raw data cannot simply be pooled together

A Bit About My Research

I work on the theory of Machine Learning

- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**

A Bit About My Research

I work on the theory of Machine Learning

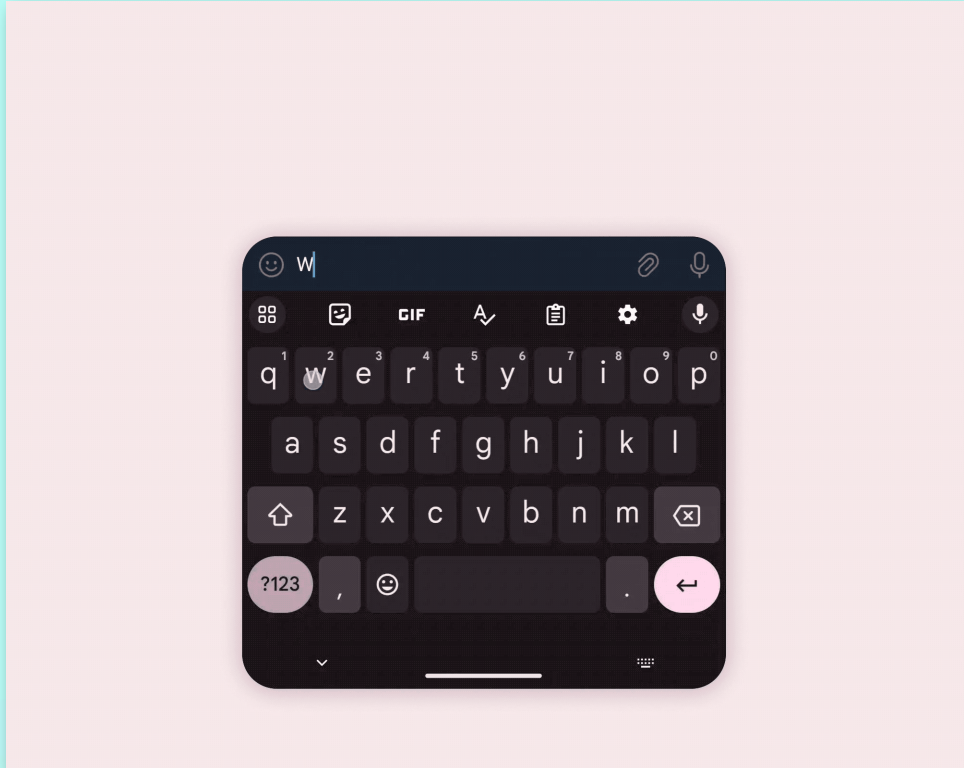
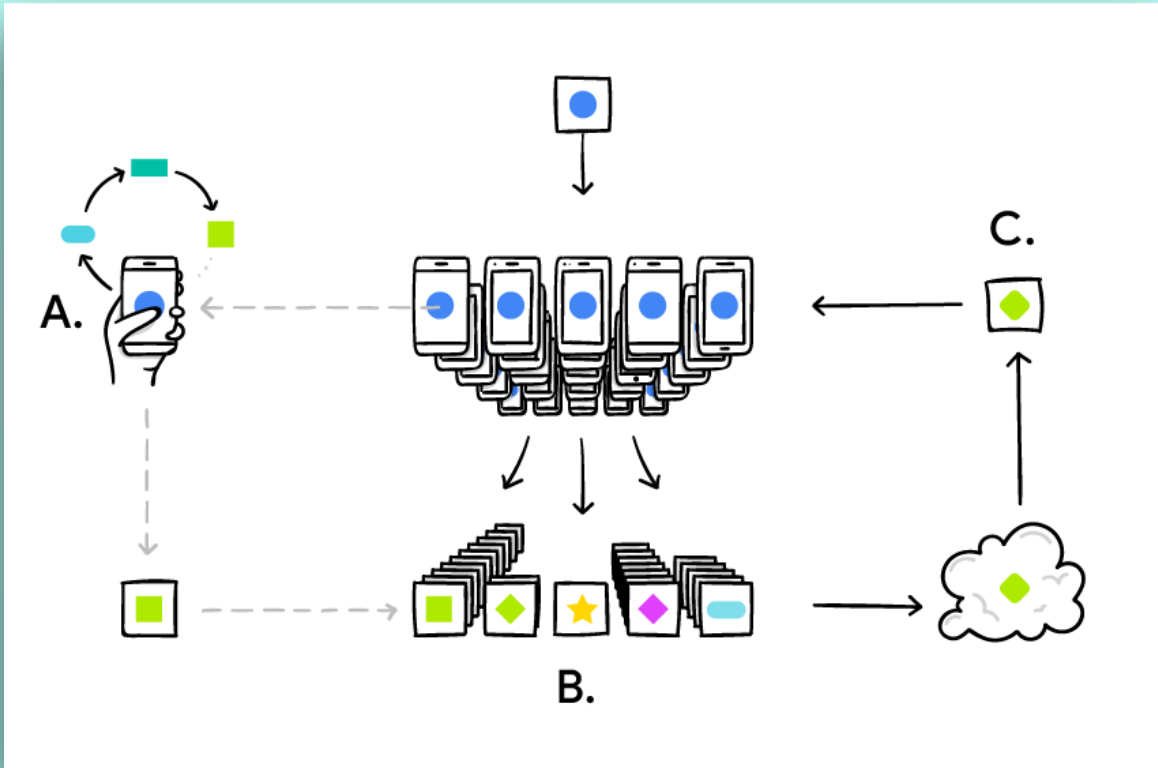
- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**
 - Since then it has been widely adopted in across many domains

Federated Learning Powers Several Consumer Electronics

[Submitted on 8 Nov 2018 (v1), last revised 28 Feb 2019 (this version, v2)]

Federated Learning for Mobile Keyboard Prediction

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, Daniel Ramage



[Submitted on 26 Jun 2023]

Private Federated Learning in Gboard

Yuanbo Zhang, Daniel Ramage, Zheng Xu, Yanxiang Zhang, Shumin Zhai, Peter Kairouz

MIT
Technology
Review

SIGN IN

SUBSCRIBE

ARTIFICIAL INTELLIGENCE

How Apple personalizes Siri without hoovering up your data

The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

By Karen Hao

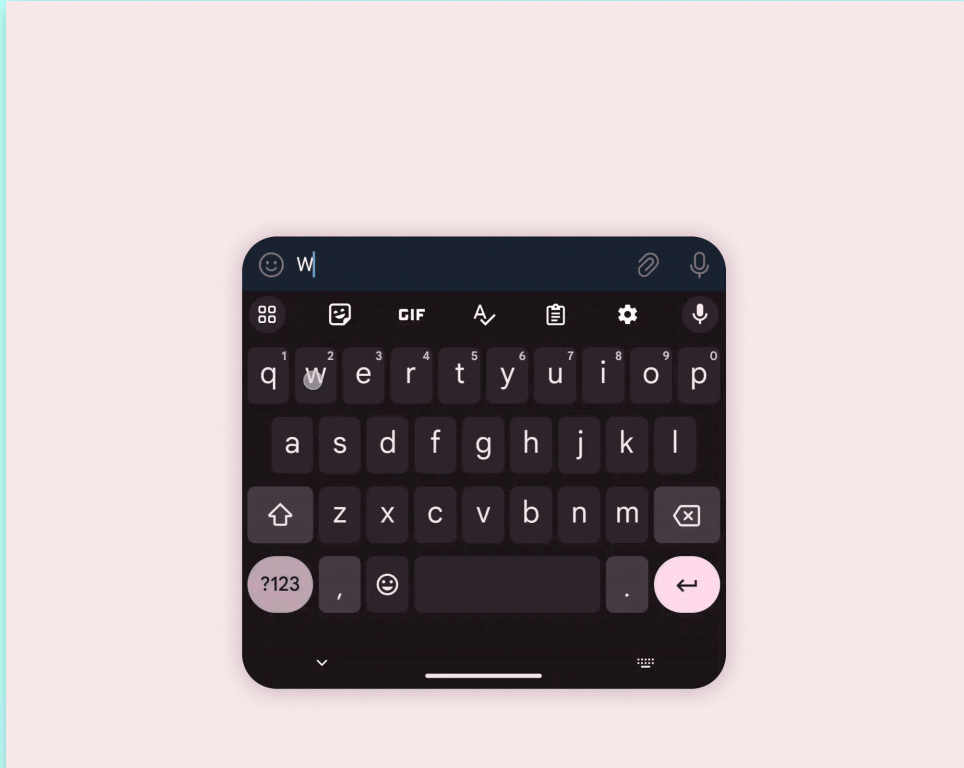
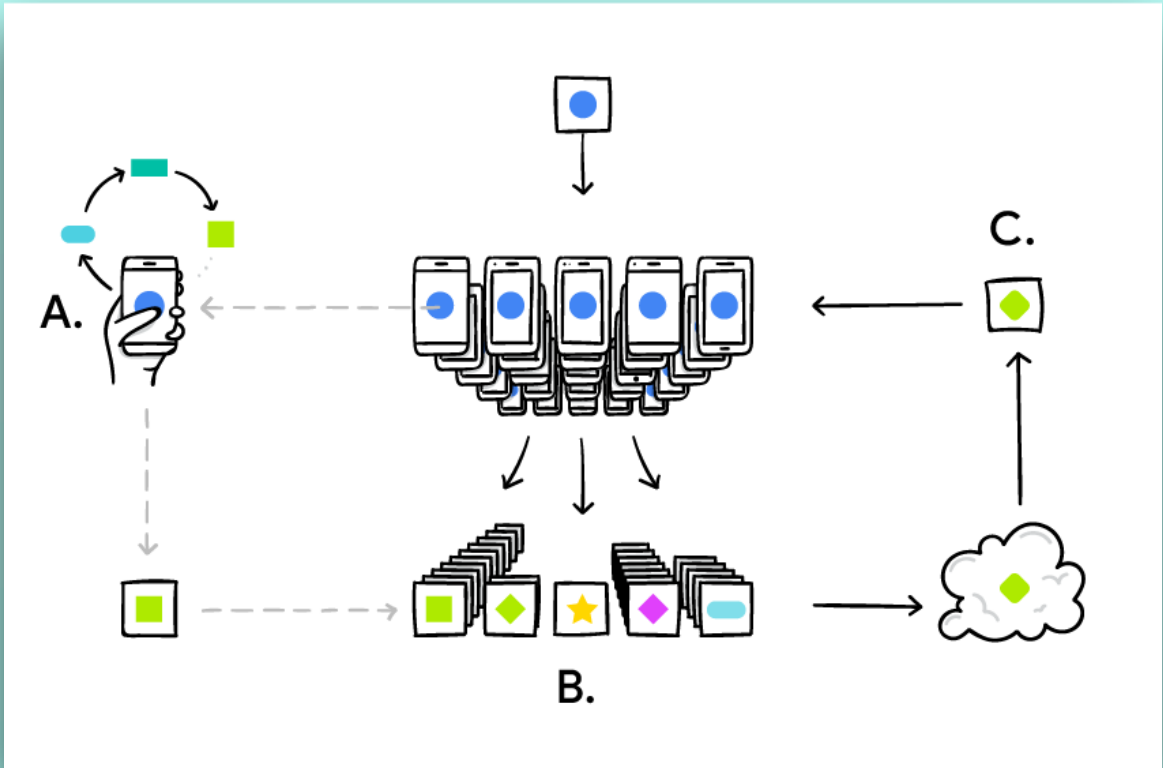
December 11, 2019

Federated Learning Powers Several Consumer Electronics

[Submitted on 8 Nov 2018 (v1), last revised 28 Feb 2019 (this version, v2)]

Federated Learning for Mobile Keyboard Prediction

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, Daniel Ramage



[Submitted on 26 Jun 2023]

Private Federated Learning in Gboard

Yuanbo Zhang, Daniel Ramage, Zheng Xu, Yanxiang Zhang, Shumin Zhai, Peter Kairouz

MIT
Technology
Review

SIGN IN

SUBSCRIBE

≡

ARTIFICIAL INTELLIGENCE

How Apple personalizes Siri without hoovering up your data

The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

By Karen Hao

December 11, 2019

Federated Learning can Enable Science and Medicine

Home > Machine Learning in Medical Imaging > Conference paper

Privacy-Preserving Federated Brain Tumour Segmentation

Conference paper | First Online: 10 October 2019
pp 133–141 | [Cite this conference paper](#)

Wenqi Li, Fausto Milletari, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M. Jorge Cardoso & Andrew Feng

Review

> Curr Opin Psychiatry. 2012 Mar;25(2):76–82. doi: 10.1097/YCO.0b013e32835035dd.

Genome-wide association studies of schizophrenia: does bigger lead to better results?

Sarah E Bergen¹, Tracey L Petryshen

Affiliations + expand

PMID: 22277805 PMCID: PMC3771358 DOI: 10.1097/YCO.0b013e32835035dd

Home > Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning > Conference paper

Federated Learning for Breast Density Classification: A Real-World Implementation

Conference paper | First Online: 26 September 2020
pp 181–191 | [Cite this conference paper](#)

Holger R. Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C. Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B. Min Yun, Elshaimaa Sharaf, Katharina V. Hoebel, Jay B. Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta D. Pisano, Laura Coombs, Daguang Xu, Keith J. Dreyer, Ittai Dayan, Ram C. Naidu, Mona Flores, Daniel Rubin & Jayashree Kalpathy-Cramer

nature communications

Explore content>About the journal>Publish with us

nature > nature communications > articles > article

Article | Open access | Published: 05 December 2022

Federated learning enables big data for rare cancer boundary detection

Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G. Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, Chiharu Sako, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Philipp Vollmuth, Gianluca Brugnara, Chandrakanth J. Preetha, Felix Sahm, Klaus Maier-Hein, Maximilian Zenk, Martin Bendszus, Wolfgang Wick, Evan Calabrese, Jeffrey Rudie, ... Spyridon Bakas

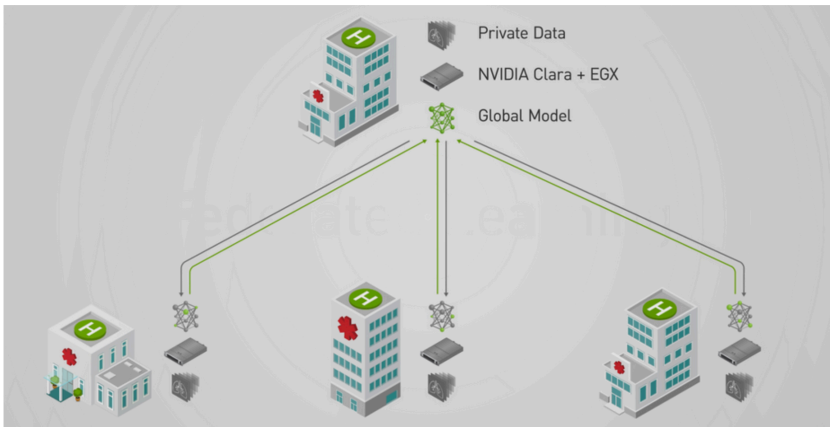
Nature Communications 13, Article number: 7346 (2022) | Cite this article

Computer Vision / Video Analytics

Federated Learning powered by NVIDIA Clara

Dec 01, 2019

By Yuhong Wen, Wenqi Li, Holger Roth and Prerna Dogra



Reinventing Anti-Money Laundering: How Federated Learning Can Power Fund Diversion Detection And GDPR Compliance



Chandra Gundlapalli Forbes Councils Member

Forbes Technology Council

COUNCIL POST | Membership (Fee-Based)



May 30, 2023, 09:45am EDT

A Bit About My Research

I work on the theory of Machine Learning

- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**
 - Since then it has been widely adopted in across many domains

A Bit About My Research

I work on the theory of Machine Learning

- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**
 - Since then it has been widely adopted in across many domains
- Federated Learning sits at the intersection of several research themes:

A Bit About My Research

I work on the theory of Machine Learning

- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**
 - Since then it has been widely adopted in across many domains
- Federated Learning sits at the intersection of several research themes:
 - Distributed optimization and systems

A Bit About My Research

I work on the theory of Machine Learning

- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**
 - Since then it has been widely adopted in across many domains
- Federated Learning sits at the intersection of several research themes:
 - Distributed optimization and systems
 - Multi-distribution learning/multi-objective optimization

A Bit About My Research

I work on the theory of Machine Learning

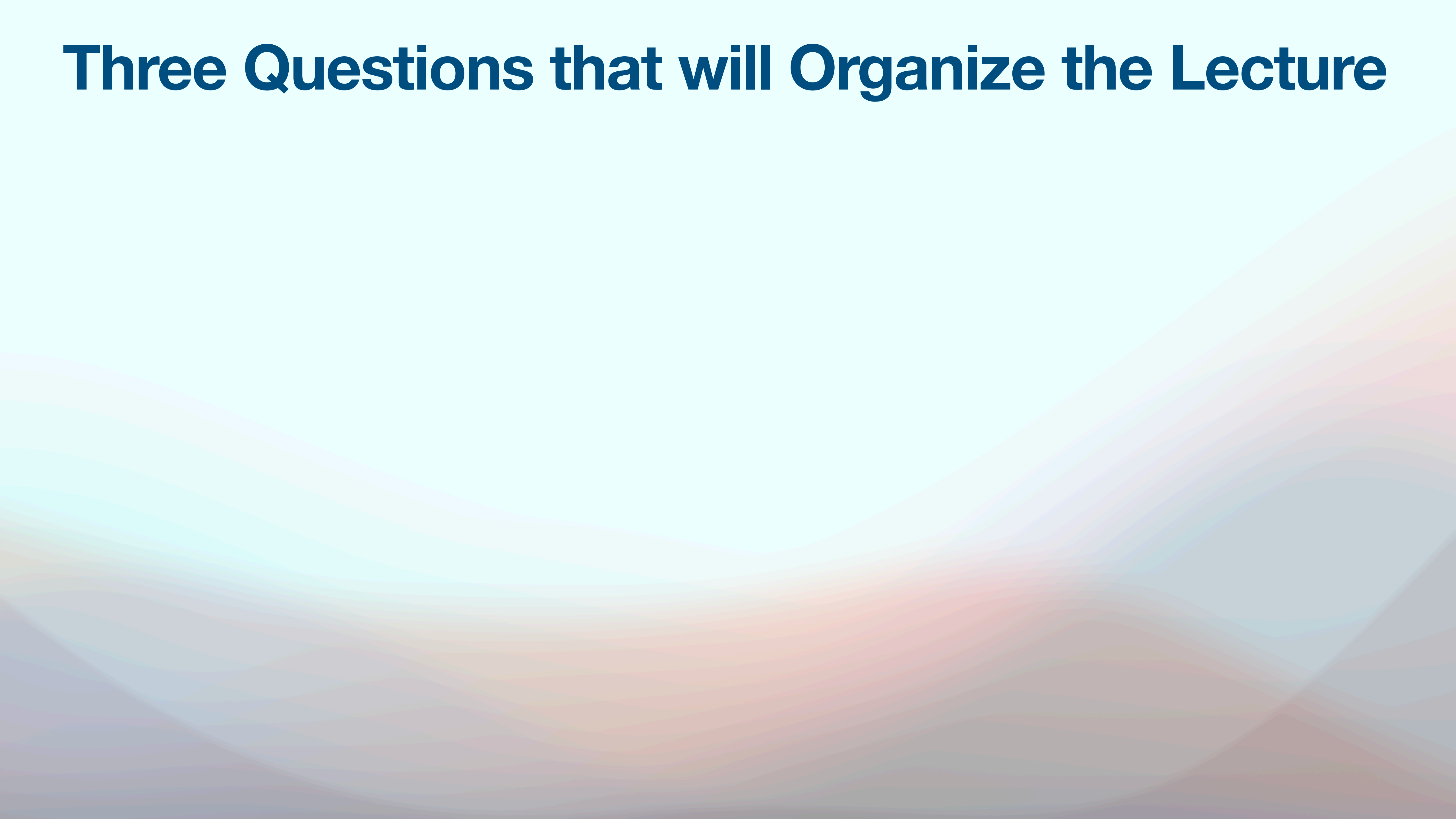
- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**
 - Since then it has been widely adopted in across many domains
- Federated Learning sits at the intersection of several research themes:
 - Distributed optimization and systems
 - Multi-distribution learning/multi-objective optimization
 - Differential privacy and security

A Bit About My Research

I work on the theory of Machine Learning

- I study learning when raw data cannot simply be pooled together
 - Circa 2017, Google proposed a term for this: **Federated Learning**
 - Since then it has been widely adopted in across many domains
- Federated Learning sits at the intersection of several research themes:
 - Distributed optimization and systems
 - Multi-distribution learning/multi-objective optimization
 - Differential privacy and security
 - Game theory and mechanism design for multi agent interactions

Three Questions that will Organize the Lecture



Three Questions that will Organize the Lecture

1. How can people reveal a population-level fact without revealing their individual answers? **Differential Privacy** (a short introduction)

Three Questions that will Organize the Lecture

1. How can people reveal a population-level fact without revealing their individual answers? **Differential Privacy** (a short introduction)
2. How can we learn to generate AI slop images from a “distribution”, a.k.a. generative modeling? **Diffusion Models** (a very short introduction)

Three Questions that will Organize the Lecture

1. How can people reveal a population-level fact without revealing their individual answers? **Differential Privacy** (a short introduction)
2. How can we learn to generate ~~AI-slop~~ images from a “distribution”, a.k.a. generative modeling? **Diffusion Models** (a very short introduction)

Three Questions that will Organize the Lecture

1. How can people reveal a population-level fact without revealing their individual answers? **Differential Privacy** (a short introduction)
2. How can we learn to generate ~~AI-slop~~ images from a “distribution”, a.k.a. generative modeling? **Diffusion Models** (a very short introduction)
3. How to combine both of these to enable data generation across multiple sites holding some private data?

Three Questions that will Organize the Lecture

1. How can people reveal a population-level fact without revealing their individual answers? **Differential Privacy** (a short introduction)
2. How can we learn to generate ~~AI-slop~~ images from a “distribution”, a.k.a. generative modeling? **Diffusion Models** (a very short introduction)
3. How to combine both of these to enable data generation across multiple sites holding some private data?
 - **Patel, K. K.**, Jiang, B., Kabir, A. F. M., Zhang, W., Zou, D., & Wang, L. (2026). Personalized federated training of diffusion models with privacy guarantees. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 31790-31801).

I want to estimate: What fraction of students in this classroom have ever cheated on an assignment?

I want to estimate: What fraction of students in this classroom have ever cheated on an assignment?

I will not ask you to answer. Instead, let us design a protocol that estimates the fraction without revealing anyone's individual answer.

We Want One Population Atatistic, not Individual Answers



We Want One Population Atatistic, not Individual Answers

- Student i holds a sensitive bit indicating if they have ever cheated

$$X_i \in \{0,1\}$$

We Want One Population Atatistic, not Individual Answers

- Student i holds a sensitive bit indicating if they have ever cheated

$$X_i \in \{0,1\}$$

- We want to estimate

$$p = \frac{1}{n} \sum_{i=1}^n X_i$$

We Want One Population Atatistic, not Individual Answers

- Student i holds a sensitive bit indicating if they have ever cheated

$$X_i \in \{0,1\}$$

- We want to estimate

$$p = \frac{1}{n} \sum_{i=1}^n X_i$$

- Each student sends me a message $Y_i \in \mathbb{R}$

We Want One Population Atatistic, not Individual Answers

- Student i holds a sensitive bit indicating if they have ever cheated

$$X_i \in \{0,1\}$$

- We want to estimate

$$p = \frac{1}{n} \sum_{i=1}^n X_i$$

- Each student sends me a message $Y_i \in \mathbb{R}$
- The protocol should make p estimable while preventing me to reliably infer any particular X_i

Truthful reporting gives accuracy without privacy

Protocol

Each student sends the truthful bit:

$$Y_i = X_i$$

I report the sample average of the messages

Truthful reporting gives accuracy without privacy

Protocol

Each student sends the truthful bit:

$$Y_i = X_i$$

I report the sample average of the messages

Consequence

Accuracy: perfect.

Privacy: none.

I learns p exactly, but also learns every individual X_i

Uniform Randomization Gives privacy Without Information

Protocol

Each student ignores the truth and reports a fair random bit, i.e.,

$$\mathbb{P}(Y_i = 1) = \frac{1}{2}$$

Uniform Randomization Gives privacy Without Information

Protocol

Each student ignores the truth and reports a fair random bit, i.e.,

$$\mathbb{P}(Y_i = 1) = \frac{1}{2}$$

Consequence

Privacy: perfect.

Accuracy: none.

The reports are independent of the sensitive bits, i.e., $X_i \perp Y_i$, so they contain no information about p .

Randomized Response Interpolates Between the Two Extremes



Randomized Response Interpolates Between the Two Extremes

- Choose a parameter $\gamma \in \left[0, \frac{1}{2}\right]$

Randomized Response Interpolates Between the Two Extremes

- Choose a parameter $\gamma \in \left[0, \frac{1}{2}\right]$
- With probability $\frac{1}{2} + \gamma$ report the true response

$$Y_i = X_i$$

Randomized Response Interpolates Between the Two Extremes

- Choose a parameter $\gamma \in \left[0, \frac{1}{2}\right]$
- With probability $\frac{1}{2} + \gamma$ report the true response

$$Y_i = X_i$$

- With probability $\frac{1}{2} - \gamma$ report the false response

$$Y_i = 1 - X_i$$

Randomized Response Interpolates Between the Two Extremes

- Choose a parameter $\gamma \in \left[0, \frac{1}{2}\right]$
- With probability $\frac{1}{2} + \gamma$ report the true response

$$Y_i = X_i$$

- With probability $\frac{1}{2} - \gamma$ report the false response

$$Y_i = 1 - X_i$$

- $\gamma = 0$: perfectly private but useless.

Randomized Response Interpolates Between the Two Extremes

- Choose a parameter $\gamma \in \left[0, \frac{1}{2}\right]$
- With probability $\frac{1}{2} + \gamma$ report the true response

$$Y_i = X_i$$

- With probability $\frac{1}{2} - \gamma$ report the false response

$$Y_i = 1 - X_i$$

- $\gamma = 0$: perfectly private but useless.
- $\gamma = \frac{1}{2}$: perfectly accurate but non-private.

At an Intermediate Value, Every Report Remains Deniable

Take $\gamma = \frac{1}{4}$. The student reports the truth with probability $\frac{3}{4}$ and flips it with probability $\frac{1}{4}$.

$$\mathbb{P}(Y_i = 1 \mid X_i = 1) = \frac{3}{4}$$

$$\mathbb{P}(Y_i = 1 \mid X_i = 0) = \frac{1}{4}$$

$$\mathbb{P}(Y_i = 0 \mid X_i = 1) = \frac{1}{4}$$

$$\mathbb{P}(Y_i = 0 \mid X_i = 0) = \frac{3}{4}$$

Either possible truth can produce either possible report. The report changes the odds, but it does not determine the truth.

Warner Designed Randomized Response for Sensitive Surveys



Warner Designed Randomized Response for Sensitive Surveys

- Stanley Warner introduced the method in 1965 to reduce evasive-answer bias
 - Stanley L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” JASA, 1965.

Warner Designed Randomized Response for Sensitive Surveys

- Stanley Warner introduced the method in 1965 to reduce evasive-answer bias
 - Stanley L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” JASA, 1965.
- Direct questions about stigmatized or illegal behavior often produce refusals or dishonest answers

Warner Designed Randomized Response for Sensitive Surveys

- Stanley Warner introduced the method in 1965 to reduce evasive-answer bias
 - Stanley L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” JASA, 1965.
- Direct questions about stigmatized or illegal behavior often produce refusals or dishonest answers
- Randomization gives each respondent plausible deniability: no reported answer proves the underlying truth

Warner Designed Randomized Response for Sensitive Surveys

- Stanley Warner introduced the method in 1965 to reduce evasive-answer bias
 - Stanley L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” JASA, 1965.
- Direct questions about stigmatized or illegal behavior often produce refusals or dishonest answers
- Randomization gives each respondent plausible deniability: no reported answer proves the underlying truth
- The analyst knows the randomization probabilities, so the aggregate distribution of reports still identifies the population prevalence

Warner Designed Randomized Response for Sensitive Surveys

- Stanley Warner introduced the method in 1965 to reduce evasive-answer bias
 - Stanley L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” JASA, 1965.
- Direct questions about stigmatized or illegal behavior often produce refusals or dishonest answers
- Randomization gives each respondent plausible deniability: no reported answer proves the underlying truth
- The analyst knows the randomization probabilities, so the aggregate distribution of reports still identifies the population prevalence
- Differential privacy later supplied a general mathematical language for this protection, circa 2003

The Randomized Report has a Simple Conditional Expectation



The Randomized Report has a Simple Conditional Expectation

- Condition on the private bit X_i and average over the student's randomization.

The Randomized Report has a Simple Conditional Expectation

- Condition on the private bit X_i and average over the student's randomization.

$$\mathbb{E}[Y_i \mid X_i] = X_i \left(\frac{1}{2} + \gamma \right) + (1 - X_i) \left(\frac{1}{2} - \gamma \right)$$

The Randomized Report has a Simple Conditional Expectation

- Condition on the private bit X_i and average over the student's randomization.

$$\mathbb{E}[Y_i \mid X_i] = X_i \left(\frac{1}{2} + \gamma \right) + (1 - X_i) \left(\frac{1}{2} - \gamma \right)$$

$$\mathbb{E}[Y_i \mid X_i] = 2\gamma X_i + \frac{1}{2} - \gamma$$

The Randomized Report has a Simple Conditional Expectation

- Condition on the private bit X_i and average over the student's randomization.

$$\mathbb{E}[Y_i \mid X_i] = X_i \left(\frac{1}{2} + \gamma \right) + (1 - X_i) \left(\frac{1}{2} - \gamma \right)$$

$$\mathbb{E}[Y_i \mid X_i] = 2\gamma X_i + \frac{1}{2} - \gamma$$

$$\mathbb{E} \left[\frac{Y_i - \frac{1}{2} + \gamma}{2\gamma} \mid X_i \right] = X_i$$

The Randomized Report has a Simple Conditional Expectation

- Condition on the private bit X_i and average over the student's randomization.

$$\mathbb{E}[Y_i \mid X_i] = X_i \left(\frac{1}{2} + \gamma \right) + (1 - X_i) \left(\frac{1}{2} - \gamma \right)$$

$$\mathbb{E}[Y_i \mid X_i] = 2\gamma X_i + \frac{1}{2} - \gamma$$

$$\mathbb{E} \left[\frac{Y_i - \frac{1}{2} + \gamma}{2\gamma} \mid X_i \right] = X_i$$

- The transformed report is an unbiased estimator of the private bit.

Averaging the Transformed Reports gives an Unbiased Estimator



Averaging the Transformed Reports gives an Unbiased Estimator

- Define the estimator from all n reports

Averaging the Transformed Reports gives an Unbiased Estimator

- Define the estimator from all n reports

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i - \frac{1}{2} + \gamma}{2\gamma}$$

Averaging the Transformed Reports gives an Unbiased Estimator

- Define the estimator from all n reports

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i - \frac{1}{2} + \gamma}{2\gamma}$$

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{Y_i - \frac{1}{2} + \gamma}{2\gamma} \right]$$

Averaging the Transformed Reports gives an Unbiased Estimator

- Define the estimator from all n reports

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i - \frac{1}{2} + \gamma}{2\gamma}$$

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{Y_i - \frac{1}{2} + \gamma}{2\gamma} \right]$$

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \sum_{i=1}^n X_i = p$$

Averaging the Transformed Reports gives an Unbiased Estimator

- Define the estimator from all n reports

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i - \frac{1}{2} + \gamma}{2\gamma}$$

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{Y_i - \frac{1}{2} + \gamma}{2\gamma} \right]$$

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \sum_{i=1}^n X_i = p$$

- Individual reports remain ambiguous, but their properly debiased average targets the true prevalence exactly in expectation.

The Report Variance is the Same for Either Private Answer



The Report Variance is the Same for Either Private Answer

- Y_i is Bernoulli with parameter $\frac{1}{2} + \gamma$

The Report Variance is the Same for Either Private Answer

- Y_i is Bernoulli with parameter $\frac{1}{2} + \gamma$

$$\text{Var}(Y_i \mid X_i = 1) = \left(\frac{1}{2} + \gamma\right) \left(\frac{1}{2} - \gamma\right)$$

The Report Variance is the Same for Either Private Answer

- Y_i is Bernoulli with parameter $\frac{1}{2} + \gamma$

$$\text{Var}(Y_i \mid X_i = 1) = \left(\frac{1}{2} + \gamma\right) \left(\frac{1}{2} - \gamma\right)$$

$$\text{Var}(Y_i \mid X_i = 0) = \left(\frac{1}{2} - \gamma\right) \left(\frac{1}{2} + \gamma\right)$$

The Report Variance is the Same for Either Private Answer

- Y_i is Bernoulli with parameter $\frac{1}{2} + \gamma$

$$\text{Var}(Y_i \mid X_i = 1) = \left(\frac{1}{2} + \gamma\right) \left(\frac{1}{2} - \gamma\right)$$

$$\text{Var}(Y_i \mid X_i = 0) = \left(\frac{1}{2} - \gamma\right) \left(\frac{1}{2} + \gamma\right)$$

$$\text{Var}(Y_i) = \frac{1}{4} - \gamma^2$$

The Report Variance is the Same for Either Private Answer

- Y_i is Bernoulli with parameter $\frac{1}{2} + \gamma$

$$\text{Var}(Y_i \mid X_i = 1) = \left(\frac{1}{2} + \gamma\right) \left(\frac{1}{2} - \gamma\right)$$

$$\text{Var}(Y_i \mid X_i = 0) = \left(\frac{1}{2} - \gamma\right) \left(\frac{1}{2} + \gamma\right)$$

$$\text{Var}(Y_i) = \frac{1}{4} - \gamma^2$$

- The exact variance does not depend on the unknown prevalence p .

The Estimator has a Closed Form for its Variance

The Estimator has a Closed Form for its Variance

- Assume the students randomize independently.

The Estimator has a Closed Form for its Variance

- Assume the students randomize independently.

$$\text{Var}(\hat{p}) = \frac{1}{4\gamma^2 n^2} \sum_{i=1}^n \text{Var}(Y_i)$$

The Estimator has a Closed Form for its Variance

- Assume the students randomize independently.

$$\text{Var}(\hat{p}) = \frac{1}{4\gamma^2 n^2} \sum_{i=1}^n \text{Var}(Y_i)$$

$$\text{Var}(\hat{p}) = \frac{1}{4\gamma^2 n^2} n \left(\frac{1}{4} - \gamma^2 \right)$$

The Estimator has a Closed Form for its Variance

- Assume the students randomize independently.

$$\text{Var}(\hat{p}) = \frac{1}{4\gamma^2 n^2} \sum_{i=1}^n \text{Var}(Y_i)$$

$$\text{Var}(\hat{p}) = \frac{1}{4\gamma^2 n^2} n \left(\frac{1}{4} - \gamma^2 \right)$$

$$\text{Var}(\hat{p}) = \frac{\frac{1}{4} - \gamma^2}{4\gamma^2 n}$$

The Estimator has a Closed Form for its Variance

- Assume the students randomize independently.

$$\text{Var}(\hat{p}) = \frac{1}{4\gamma^2 n^2} \sum_{i=1}^n \text{Var}(Y_i)$$

$$\text{Var}(\hat{p}) = \frac{1}{4\gamma^2 n^2} n \left(\frac{1}{4} - \gamma^2 \right)$$

$$\text{Var}(\hat{p}) = \frac{\frac{1}{4} - \gamma^2}{4\gamma^2 n}$$

- Because the estimator is unbiased, its mean-squared error equals this variance.

The Exact Root-mean-squared Scales as $1/(\gamma\sqrt{n})$

The Exact Root-mean-squared Scales as $1/(\gamma\sqrt{n})$

- To see this note that:

The Exact Root-mean-squared Scales as $1/(\gamma\sqrt{n})$

- To see this note that:

$$\sqrt{\mathbb{E}[(\hat{p} - p)^2]} = \frac{\sqrt{\frac{1}{4} - \gamma^2}}{2\gamma\sqrt{n}}$$

The Exact Root-mean-squared Scales as $1/(\gamma\sqrt{n})$

- To see this note that:

$$\sqrt{\mathbb{E}[(\hat{p} - p)^2]} = \frac{\sqrt{\frac{1}{4} - \gamma^2}}{2\gamma\sqrt{n}}$$

$$\sqrt{\mathbb{E}[(\hat{p} - p)^2]} \leq \frac{1}{4\gamma\sqrt{n}}$$

The Exact Root-mean-squared Scales as $1/(\gamma\sqrt{n})$

- To see this note that:

$$\sqrt{\mathbb{E}[(\hat{p} - p)^2]} = \frac{\sqrt{\frac{1}{4} - \gamma^2}}{2\gamma\sqrt{n}}$$

$$\sqrt{\mathbb{E}[(\hat{p} - p)^2]} \leq \frac{1}{4\gamma\sqrt{n}}$$

$$\mathbb{P}(|\hat{p} - p| \geq t) \leq \frac{\frac{1}{4} - \gamma^2}{4\gamma^2 n t^2}$$

The Exact Root-mean-squared Scales as $1/(\gamma\sqrt{n})$

- To see this note that:

$$\sqrt{\mathbb{E}[(\hat{p} - p)^2]} = \frac{\sqrt{\frac{1}{4} - \gamma^2}}{2\gamma\sqrt{n}}$$

$$\sqrt{\mathbb{E}[(\hat{p} - p)^2]} \leq \frac{1}{4\gamma\sqrt{n}}$$

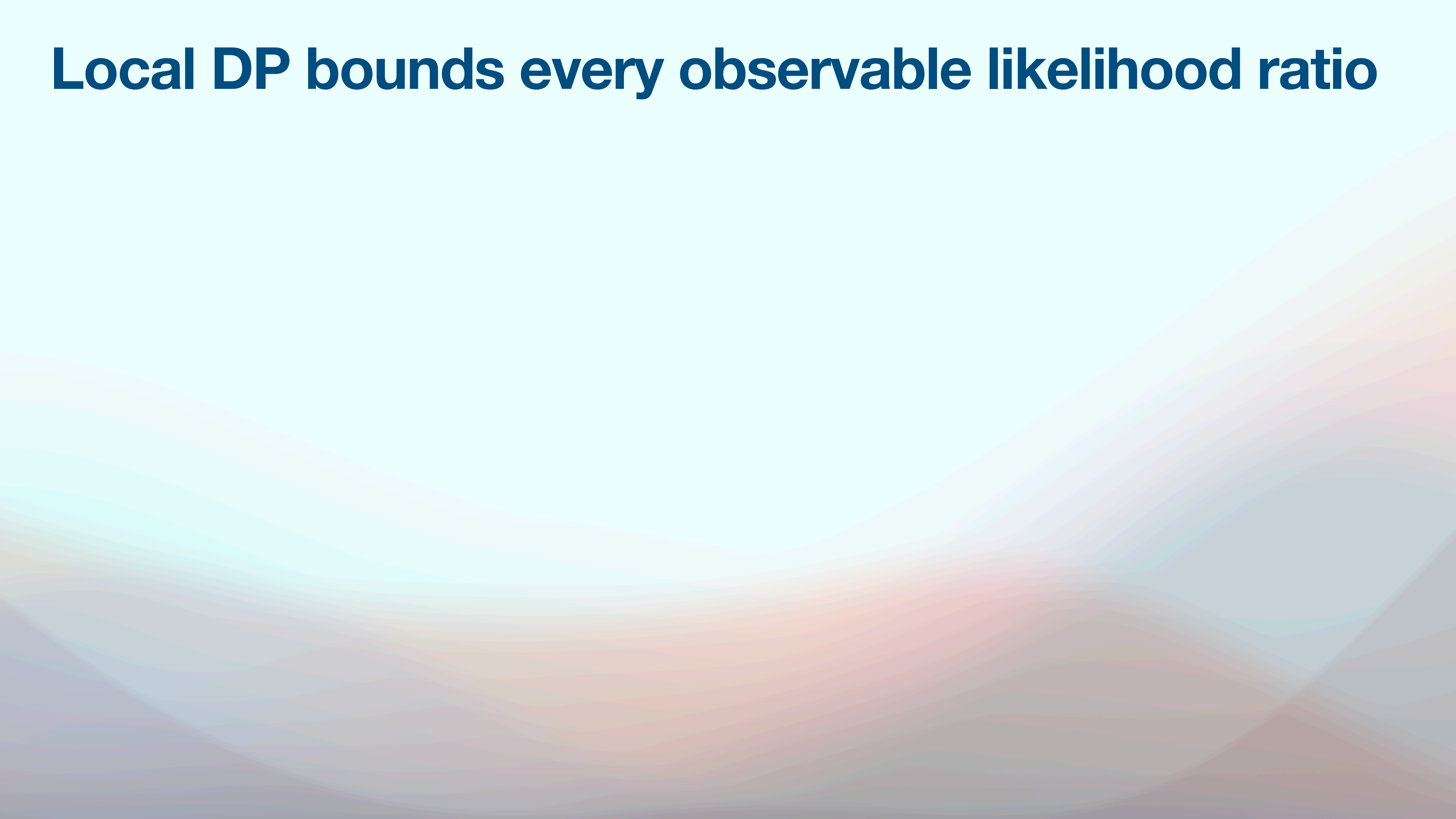
$$\mathbb{P}(|\hat{p} - p| \geq t) \leq \frac{\frac{1}{4} - \gamma^2}{4\gamma^2 n t^2}$$

- The last line follows from Chebyshev's inequality. Smaller gamma gives stronger privacy but larger statistical error.

How can we quantify how much an observed report distinguishes two possible truths?

The answer will compare the entire output distribution under every pair of possible private inputs.

Local DP bounds every observable likelihood ratio



Local DP bounds every observable likelihood ratio

- A randomizer R satisfies epsilon-local differential privacy if, for every two inputs and every output event:

Local DP bounds every observable likelihood ratio

- A randomizer R satisfies epsilon-local differential privacy if, for every two inputs and every output event:

$$\mathbb{P}(R(x) \in O) \leq e^\epsilon \mathbb{P}(R(x') \in O)$$

Local DP bounds every observable likelihood ratio

- A randomizer R satisfies epsilon-local differential privacy if, for every two inputs and every output event:

$$\mathbb{P}(R(x) \in O) \leq e^\epsilon \mathbb{P}(R(x') \in O)$$

$$x, x' \in \mathcal{X}, \quad O \subseteq \mathcal{Y}$$

Local DP bounds every observable likelihood ratio

- A randomizer R satisfies epsilon-local differential privacy if, for every two inputs and every output event:

$$\mathbb{P}(R(x) \in O) \leq e^\epsilon \mathbb{P}(R(x') \in O)$$

$$x, x' \in \mathcal{X}, \quad O \subseteq \mathcal{Y}$$

- The probability is over the randomizer.

Local DP bounds every observable likelihood ratio

- A randomizer R satisfies epsilon-local differential privacy if, for every two inputs and every output event:

$$\mathbb{P}(R(x) \in O) \leq e^\epsilon \mathbb{P}(R(x') \in O)$$

$$x, x' \in \mathcal{X}, \quad O \subseteq \mathcal{Y}$$

- The probability is over the randomizer.
- Because x and x' can be swapped, the two probabilities must be comparable in both directions.

Local DP bounds every observable likelihood ratio

- A randomizer R satisfies epsilon-local differential privacy if, for every two inputs and every output event:

$$\mathbb{P}(R(x) \in O) \leq e^\epsilon \mathbb{P}(R(x') \in O)$$

$$x, x' \in \mathcal{X}, \quad O \subseteq \mathcal{Y}$$

- The probability is over the randomizer.
- Because x and x' can be swapped, the two probabilities must be comparable in both directions.
- Smaller ϵ means stronger privacy.

Randomized response has an exact LDP parameter

Randomized response has an exact LDP parameter

- For binary outputs, it suffices to compare the probability of each possible report.

Randomized response has an exact LDP parameter

- For binary outputs, it suffices to compare the probability of each possible report.

$$\max_{x,x',y} \frac{\mathbb{P}(R(x) = y)}{\mathbb{P}(R(x') = y)} = \frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma}$$

Randomized response has an exact LDP parameter

- For binary outputs, it suffices to compare the probability of each possible report.

$$\max_{x,x',y} \frac{\mathbb{P}(R(x) = y)}{\mathbb{P}(R(x') = y)} = \frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma}$$

$$\varepsilon = \log \left(\frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma} \right)$$

Randomized response has an exact LDP parameter

- For binary outputs, it suffices to compare the probability of each possible report.

$$\max_{x,x',y} \frac{\mathbb{P}(R(x) = y)}{\mathbb{P}(R(x') = y)} = \frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma}$$

$$\varepsilon = \log \left(\frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma} \right)$$

- The worst case compares a report that agrees with one input and disagrees with the other. Because the pointwise ratio is bounded for every report y , summing over all y in an event O gives the event-level LDP inequality.

Randomized response has an exact LDP parameter

- For binary outputs, it suffices to compare the probability of each possible report.

$$\max_{x,x',y} \frac{\mathbb{P}(R(x) = y)}{\mathbb{P}(R(x') = y)} = \frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma}$$

$$\epsilon = \log \left(\frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma} \right)$$

- The worst case compares a report that agrees with one input and disagrees with the other. Because the pointwise ratio is bounded for every report y , summing over all y in an event O gives the event-level LDP inequality.
- At $\gamma = 1/4$, $\epsilon = \log 3$

**What if students trust the instructor with truthful answers,
but do not trust the people who see the published result?**

The instructor may now aggregate the clean data before adding randomness.

Neighboring datasets isolate one individual's contribution

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

$$D = (x_1, \dots, x_i, \dots, x_n)$$

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

$$D = (x_1, \dots, x_i, \dots, x_n)$$

$$D' = (x_1, \dots, x'_i, \dots, x_n)$$

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

$$D = (x_1, \dots, x_i, \dots, x_n)$$

$$D' = (x_1, \dots, x'_i, \dots, x_n)$$

$$D \sim D'$$

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

$$D = (x_1, \dots, x_i, \dots, x_n)$$

$$D' = (x_1, \dots, x'_i, \dots, x_n)$$

$$D \sim D'$$

- A randomized algorithm M satisfies pure epsilon-DP if, for all neighboring datasets and all output events:

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

$$D = (x_1, \dots, x_i, \dots, x_n)$$

$$D' = (x_1, \dots, x'_i, \dots, x_n)$$

$$D \sim D'$$

- A randomized algorithm M satisfies pure epsilon-DP if, for all neighboring datasets and all output events:

$$\mathbb{P}(M(D) \in O) \leq e^\epsilon \mathbb{P}(M(D') \in O)$$

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

$$D = (x_1, \dots, x_i, \dots, x_n)$$

$$D' = (x_1, \dots, x'_i, \dots, x_n)$$

$$D \sim D'$$

- A randomized algorithm M satisfies pure epsilon-DP if, for all neighboring datasets and all output events:

$$\mathbb{P}(M(D) \in O) \leq e^\epsilon \mathbb{P}(M(D') \in O)$$

$$D \sim D', \quad O \subseteq \mathcal{Y}$$

Neighboring datasets isolate one individual's contribution

- Under bounded adjacency, D and D' have the same size and differ in one entry.

$$D = (x_1, \dots, x_i, \dots, x_n)$$

$$D' = (x_1, \dots, x'_i, \dots, x_n)$$

$$D \sim D'$$

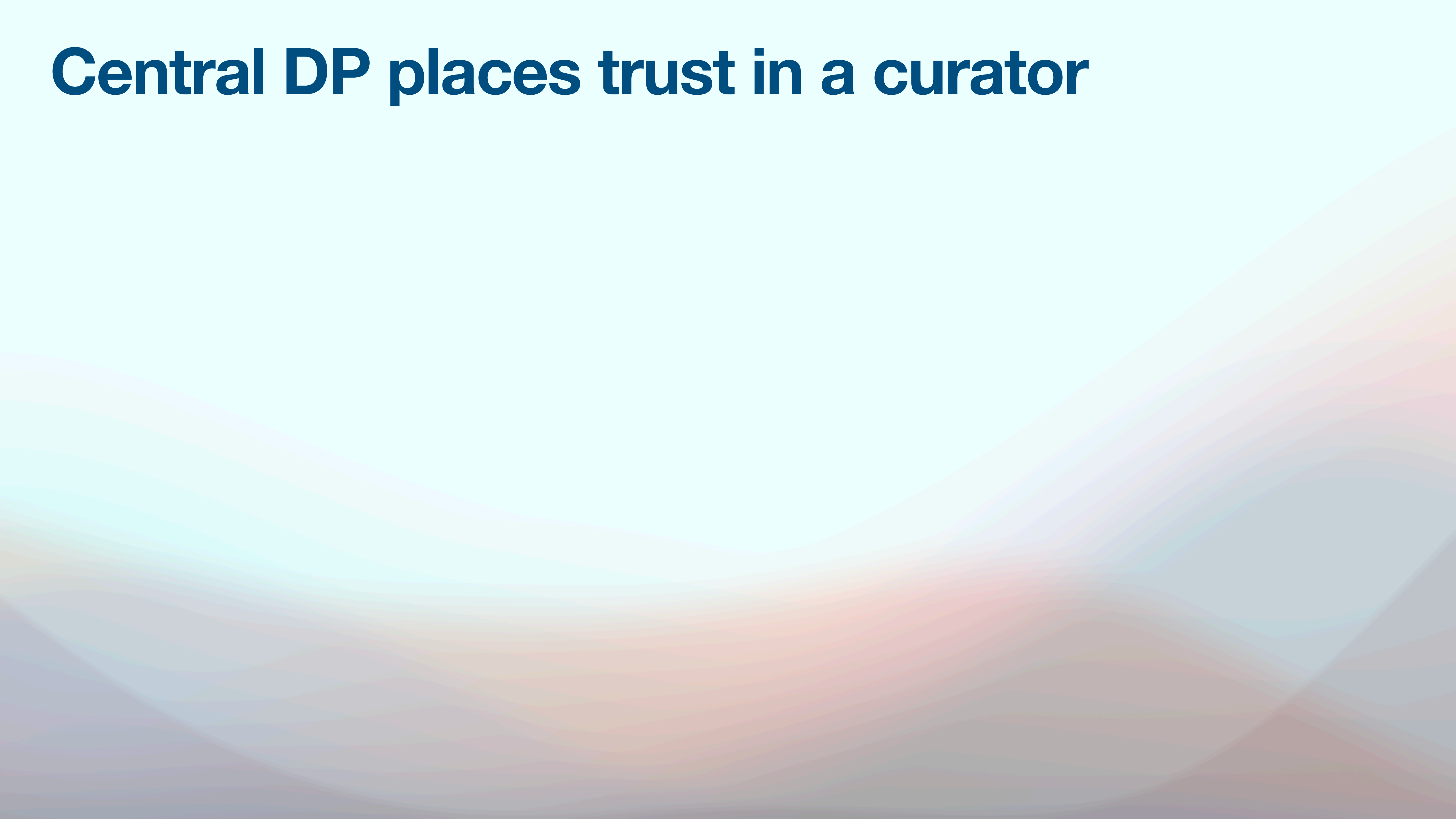
- A randomized algorithm M satisfies pure epsilon-DP if, for all neighboring datasets and all output events:

$$\mathbb{P}(M(D) \in O) \leq e^\epsilon \mathbb{P}(M(D') \in O)$$

$$D \sim D', \quad O \subseteq \mathcal{Y}$$

- DP says that one person cannot change the probability of any observable event by more than the multiplicative factor e^ϵ .

Central DP places trust in a curator



Central DP places trust in a curator

- Individuals send clean data to a trusted curator. The curator releases only a randomized result

Central DP places trust in a curator

- Individuals send clean data to a trusted curator. The curator releases only a randomized result
- The privacy guarantee constrains the public output, not the curator's internal view.

Central DP places trust in a curator

- Individuals send clean data to a trusted curator. The curator releases only a randomized result
- The privacy guarantee constrains the public output, not the curator's internal view.
- Because the curator can aggregate before randomizing, central DP generally provides substantially better accuracy than local DP.

Central DP places trust in a curator

- Individuals send clean data to a trusted curator. The curator releases only a randomized result
- The privacy guarantee constrains the public output, not the curator's internal view.
- Because the curator can aggregate before randomizing, central DP generally provides substantially better accuracy than local DP.
- The protected unit must still be specified: one record, one person, or an entire institution can define different neighboring relations.

Local and central DP use the same idea with different inputs

Local DP

Compared inputs: two possible individual values x and x' .

Randomization: before communication.

Clean data seen by: only the individual.

Trust: no curator is trusted.

Local and central DP use the same idea with different inputs

Local DP

Compared inputs: two possible individual values x and x' .

Randomization: before communication.

Clean data seen by: only the individual.

Trust: no curator is trusted.

Central DP

Compared inputs: neighboring datasets D and D' .

Randomization: before public release.

Clean data seen by: trusted curator.

Trust: curator sees raw data.

DP Limits the Effect of Participation, not Population Knowledge



DP Limits the Effect of Participation, not Population Knowledge

- DP does protect against conclusions that depend strongly on one individual's record. If an attack works when your data are present, it must work almost as well when your data are replaced or removed.

DP Limits the Effect of Participation, not Population Knowledge

- DP does protect against conclusions that depend strongly on one individual's record. If an attack works when your data are present, it must work almost as well when your data are replaced or removed.
- DP does not prevent learning genuine facts about a population. If a study concludes that smoking increases cancer risk, a smoker may be affected by that conclusion even when the study is private.

DP Limits the Effect of Participation, not Population Knowledge

- DP does protect against conclusions that depend strongly on one individual's record. If an attack works when your data are present, it must work almost as well when your data are replaced or removed.
- DP does not prevent learning genuine facts about a population. If a study concludes that smoking increases cancer risk, a smoker may be affected by that conclusion even when the study is private.
- The relevant guarantee is that one smoker's participation did not cause the conclusion.

Post-processing Cannot Weaken Differential Privacy



Post-processing Cannot Weaken Differential Privacy

- Suppose M is (ϵ, δ) -DP and F accesses only $M(D)$, not the original dataset.

Post-processing Cannot Weaken Differential Privacy

- Suppose M is (ϵ, δ) -DP and F accesses only $M(D)$, not the original dataset.

$$\mathbb{P}(F(M(D)) \in O) = \mathbb{P}(M(D) \in F^{-1}(O))$$

Post-processing Cannot Weaken Differential Privacy

- Suppose M is (ϵ, δ) -DP and F accesses only $M(D)$, not the original dataset.

$$\mathbb{P}(F(M(D)) \in O) = \mathbb{P}(M(D) \in F^{-1}(O))$$

$$\mathbb{P}(M(D) \in F^{-1}(O)) \leq e^\epsilon \mathbb{P}(M(D') \in F^{-1}(O)) + \delta$$

Post-processing Cannot Weaken Differential Privacy

- Suppose M is (ϵ, δ) -DP and F accesses only $M(D)$, not the original dataset.

$$\mathbb{P}(F(M(D)) \in O) = \mathbb{P}(M(D) \in F^{-1}(O))$$

$$\mathbb{P}(M(D) \in F^{-1}(O)) \leq e^\epsilon \mathbb{P}(M(D') \in F^{-1}(O)) + \delta$$

$$\mathbb{P}(F(M(D)) \in O) \leq e^\epsilon \mathbb{P}(F(M(D')) \in O) + \delta$$

Post-processing Cannot Weaken Differential Privacy

- Suppose M is (ϵ, δ) -DP and F accesses only $M(D)$, not the original dataset.

$$\mathbb{P}(F(M(D)) \in O) = \mathbb{P}(M(D) \in F^{-1}(O))$$

$$\mathbb{P}(M(D) \in F^{-1}(O)) \leq e^\epsilon \mathbb{P}(M(D') \in F^{-1}(O)) + \delta$$

$$\mathbb{P}(F(M(D)) \in O) \leq e^\epsilon \mathbb{P}(F(M(D')) \in O) + \delta$$

- For a randomized F , condition on its independent randomness and average.

Post-processing Cannot Weaken Differential Privacy

- Suppose M is (ϵ, δ) -DP and F accesses only $M(D)$, not the original dataset.

$$\mathbb{P}(F(M(D)) \in O) = \mathbb{P}(M(D) \in F^{-1}(O))$$

$$\mathbb{P}(M(D) \in F^{-1}(O)) \leq e^\epsilon \mathbb{P}(M(D') \in F^{-1}(O)) + \delta$$

$$\mathbb{P}(F(M(D)) \in O) \leq e^\epsilon \mathbb{P}(F(M(D')) \in O) + \delta$$

- For a randomized F , condition on its independent randomness and average.
- Thus training on or transforming a privatized release retains the privacy guarantee.

Approximate DP Allows a Small Additive Slack



Approximate DP Allows a Small Additive Slack

- A randomized algorithm M is (ϵ, δ) -DP if:

Approximate DP Allows a Small Additive Slack

- A randomized algorithm M is (ϵ, δ) -DP if:

$$\mathbb{P}(M(D) \in O) \leq e^\epsilon \mathbb{P}(M(D') \in O) + \delta$$

Approximate DP Allows a Small Additive Slack

- A randomized algorithm M is (ϵ, δ) -DP if:

$$\mathbb{P}(M(D) \in O) \leq e^\epsilon \mathbb{P}(M(D') \in O) + \delta$$

- The inequality must hold for every pair of neighboring datasets and every output event.

Approximate DP Allows a Small Additive Slack

- A randomized algorithm M is (ϵ, δ) -DP if:

$$\mathbb{P}(M(D) \in O) \leq e^\epsilon \mathbb{P}(M(D') \in O) + \delta$$

- The inequality must hold for every pair of neighboring datasets and every output event.
- The parameter delta permits a small amount of probability mass that the multiplicative comparison does not cover. We must set it smaller than $1/n$ to avoid catastrophe!

Approximate DP Allows a Small Additive Slack

- A randomized algorithm M is (ϵ, δ) -DP if:

$$\mathbb{P}(M(D) \in O) \leq e^\epsilon \mathbb{P}(M(D') \in O) + \delta$$

- The inequality must hold for every pair of neighboring datasets and every output event.
- The parameter delta permits a small amount of probability mass that the multiplicative comparison does not cover. We must set it smaller than $1/n$ to avoid catastrophe!
- This relaxation enables the Gaussian mechanism and often supports better composition and high-dimensional utility.

Gaussian Mechanism and L2 Sensitivity



Gaussian Mechanism and L2 Sensitivity

- For a vector-valued function f , its l2 sensitivity is:

Gaussian Mechanism and L2 Sensitivity

- For a vector-valued function f , its l2 sensitivity is:

$$\Delta_2 = \max_{D \sim D'} \|f(D) - f(D')\|_2$$

Gaussian Mechanism and L2 Sensitivity

- For a vector-valued function f , its l_2 sensitivity is:

$$\Delta_2 = \max_{D \sim D'} \|f(D) - f(D')\|_2$$

- A mechanism must add noise relative to the largest possible change caused by one protected unit, not relative to the average change.

Gaussian Mechanism and L2 Sensitivity

- For a vector-valued function f , its l_2 sensitivity is:

$$\Delta_2 = \max_{D \sim D'} \|f(D) - f(D')\|_2$$

- A mechanism must add noise relative to the largest possible change caused by one protected unit, not relative to the average change.
- The Gaussian mechanism adds isotropic noise calibrated to sensitivity:

Gaussian Mechanism and L2 Sensitivity

- For a vector-valued function f , its l_2 sensitivity is:

$$\Delta_2 = \max_{D \sim D'} \|f(D) - f(D')\|_2$$

- A mechanism must add noise relative to the largest possible change caused by one protected unit, not relative to the average change.
- The Gaussian mechanism adds isotropic noise calibrated to sensitivity:

$$M(D) = f(D) + Z \sim \mathcal{N}(0, \sigma^2 I)$$

Gaussian Mechanism and L2 Sensitivity

- For a vector-valued function f , its l_2 sensitivity is:

$$\Delta_2 = \max_{D \sim D'} \|f(D) - f(D')\|_2$$

- A mechanism must add noise relative to the largest possible change caused by one protected unit, not relative to the average change.
- The Gaussian mechanism adds isotropic noise calibrated to sensitivity:

$$M(D) = f(D) + Z \sim \mathcal{N}(0, \sigma^2 I)$$

$$\sigma \geq \frac{\Delta_2 \sqrt{2 \log(1.25/\delta)}}{\epsilon}$$

Gaussian Mechanism and L2 Sensitivity

- For a vector-valued function f , its L2 sensitivity is:

$$\Delta_2 = \max_{D \sim D'} \|f(D) - f(D')\|_2$$

- A mechanism must add noise relative to the largest possible change caused by one protected unit, not relative to the average change.
- The Gaussian mechanism adds isotropic noise calibrated to sensitivity:

$$M(D) = f(D) + Z \sim \mathcal{N}(0, \sigma^2 I)$$

$$\sigma \geq \frac{\Delta_2 \sqrt{2 \log(1.25/\delta)}}{\epsilon}$$

- For $0 < \epsilon < 1$ and $0 < \delta < 1/2$, the above noise calibration gives (ϵ, δ) -central-DP

The Same Gaussian Mechanism Supports Central or Local DP

Central Gaussian Mechanism

$$M(D) = f(D) + Z$$

Compare neighboring datasets D and D' .

Sensitivity: $\max_{D \sim D'} \|f(D) - f(D')\|_2$.

A trusted curator adds noise after collecting clean data.

The Same Gaussian Mechanism Supports Central or Local DP

Central Gaussian Mechanism

$$M(D) = f(D) + Z$$

Compare neighboring datasets D and D' .

Sensitivity: $\max_{D \sim D'} \|f(D) - f(D')\|_2$.

A trusted curator adds noise after collecting clean data.

Local Gaussian Mechanism

$$R(x) = f(x) + Z$$

Compare any two possible inputs x and x' .

Sensitivity: $\max_{x, x'} \|f(x) - f(x')\|_2$.

The individual adds noise before communicating.

Clipping Gives a Finite Local Sensitivity for Vectors



Clipping Gives a Finite Local Sensitivity for Vectors

- We define clipping as follows:

Clipping Gives a Finite Local Sensitivity for Vectors

- We define clipping as follows:

$$\text{clip}(x, C) = x \cdot \min \left\{ 1, \frac{C}{\|x\|_2} \right\}$$

Clipping Gives a Finite Local Sensitivity for Vectors

- We define clipping as follows:

$$\text{clip}(x, C) = x \cdot \min \left\{ 1, \frac{C}{\|x\|_2} \right\}$$

$$\Rightarrow \|\text{clip}(x, C)\|_2 \leq C$$

Clipping Gives a Finite Local Sensitivity for Vectors

- We define clipping as follows:

$$\text{clip}(x, C) = x \cdot \min \left\{ 1, \frac{C}{\|x\|_2} \right\}$$

$$\Rightarrow \|\text{clip}(x, C)\|_2 \leq C$$

- Clipping bounds the global sensitivity because:

Clipping Gives a Finite Local Sensitivity for Vectors

- We define clipping as follows:

$$\text{clip}(x, C) = x \cdot \min \left\{ 1, \frac{C}{\|x\|_2} \right\}$$

$$\Rightarrow \|\text{clip}(x, C)\|_2 \leq C$$

- Clipping bounds the global sensitivity because:

$$\|\text{clip}(x, C) - \text{clip}(x', C)\|_2 \leq \|\text{clip}(x, C)\|_2 + \|\text{clip}(x', C)\|_2$$

Clipping Gives a Finite Local Sensitivity for Vectors

- We define clipping as follows:

$$\text{clip}(x, C) = x \cdot \min \left\{ 1, \frac{C}{\|x\|_2} \right\}$$

$$\Rightarrow \|\text{clip}(x, C)\|_2 \leq C$$

- Clipping bounds the global sensitivity because:

$$\|\text{clip}(x, C) - \text{clip}(x', C)\|_2 \leq \|\text{clip}(x, C)\|_2 + \|\text{clip}(x', C)\|_2$$

$$\|\text{clip}(x, C) - \text{clip}(x', C)\|_2 \leq 2C$$

Clipping Gives a Finite Local Sensitivity for Vectors

- We define clipping as follows:

$$\text{clip}(x, C) = x \cdot \min \left\{ 1, \frac{C}{\|x\|_2} \right\}$$

$$\Rightarrow \|\text{clip}(x, C)\|_2 \leq C$$

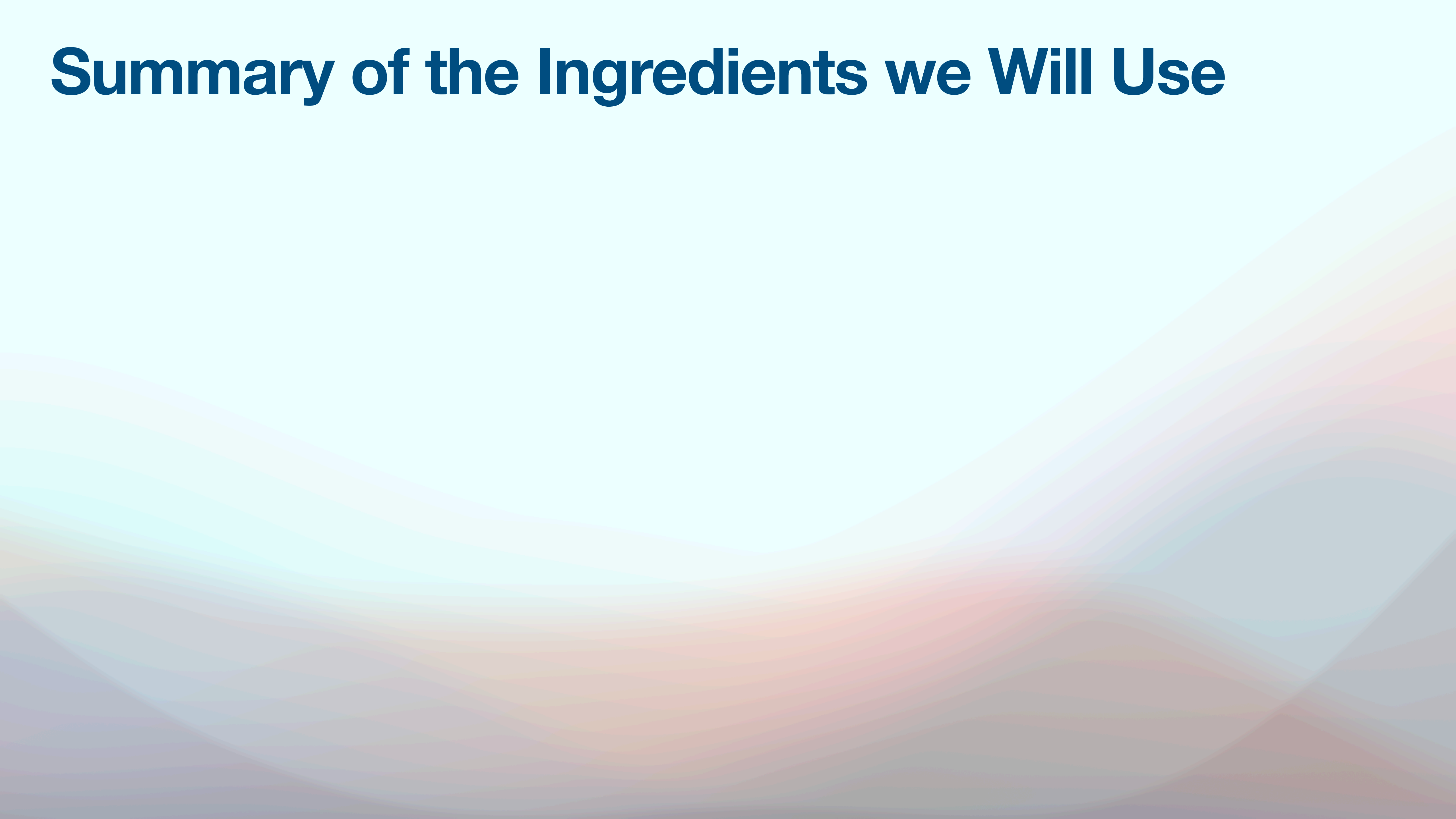
- Clipping bounds the global sensitivity because:

$$\|\text{clip}(x, C) - \text{clip}(x', C)\|_2 \leq \|\text{clip}(x, C)\|_2 + \|\text{clip}(x', C)\|_2$$

$$\|\text{clip}(x, C) - \text{clip}(x', C)\|_2 \leq 2C$$

- Thus the local Gaussian mechanism applied to a clipped vector has L2 sensitivity at most $2C$.

Summary of the Ingredients we Will Use



Summary of the Ingredients we Will Use

1. Randomization can hide individual values while preserving population information.

Summary of the Ingredients we Will Use

1. Randomization can hide individual values while preserving population information.
2. Local DP protects each input before communication; central DP trusts a curator with clean data.

Summary of the Ingredients we Will Use

1. Randomization can hide individual values while preserving population information.
2. Local DP protects each input before communication; central DP trusts a curator with clean data.
3. Approximate DP permits Gaussian noise while requiring delta to be extremely small.

Summary of the Ingredients we Will Use

1. Randomization can hide individual values while preserving population information.
2. Local DP protects each input before communication; central DP trusts a curator with clean data.
3. Approximate DP permits Gaussian noise while requiring delta to be extremely small.
4. Clipping bounds sensitivity, and the Gaussian mechanism calibrates noise to that bound.

Summary of the Ingredients we Will Use

1. Randomization can hide individual values while preserving population information.
2. Local DP protects each input before communication; central DP trusts a curator with clean data.
3. Approximate DP permits Gaussian noise while requiring delta to be extremely small.
4. Clipping bounds sensitivity, and the Gaussian mechanism calibrates noise to that bound.
5. Any later computation on a private release remains private by post-processing.

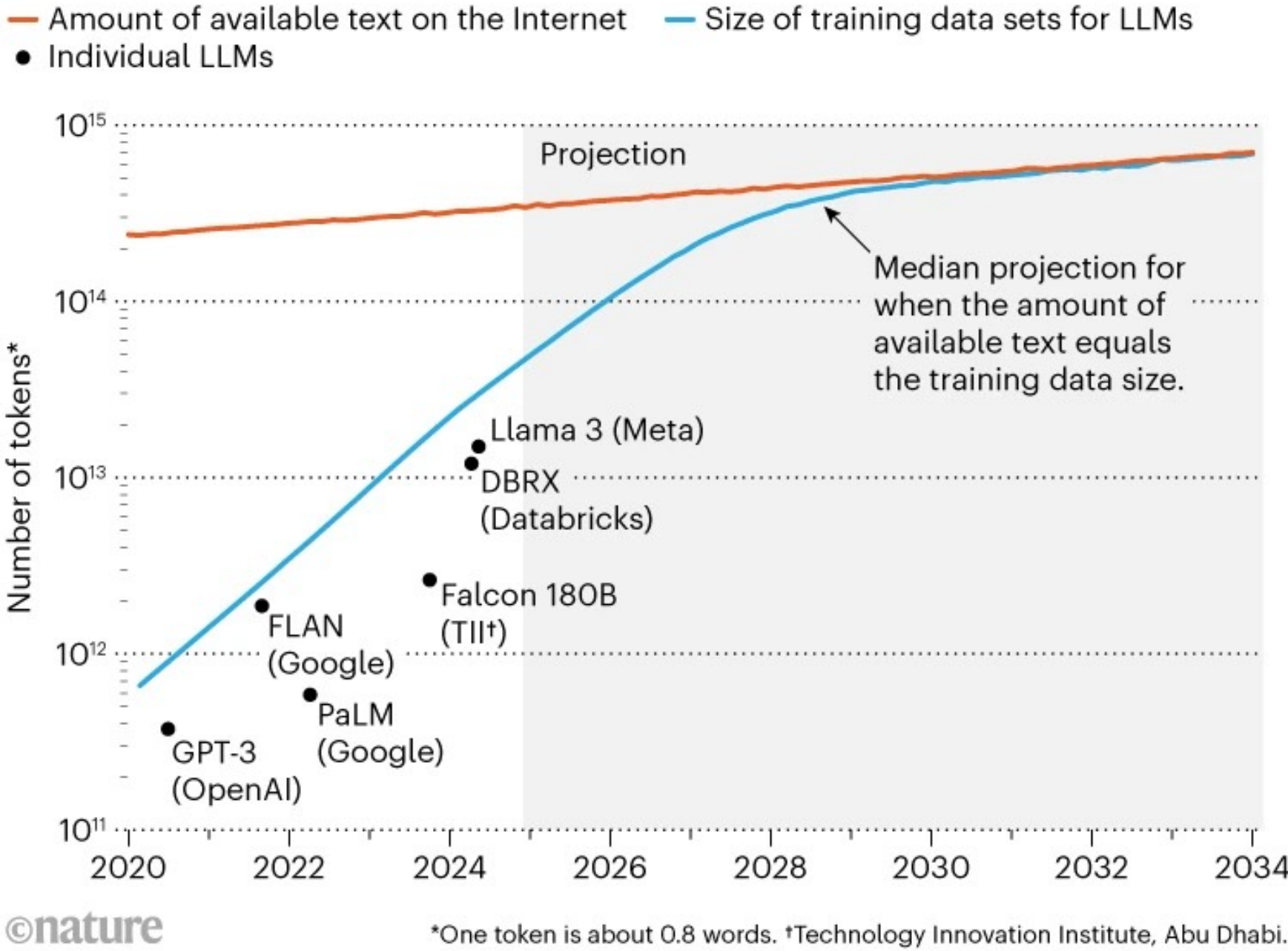
Five-minute break

When we return, we will combine privacy, diffusion models, and personalization.

The project starts from one question: what should remain global, and what should remain local?

RUNNING OUT OF DATA

The amount of text data used to train large language models (LLMs) is rapidly approaching a crisis point. An estimate suggests that, by 2028, developers will be using data sets that match the amount of text that is available on the Internet.



BUSINESS INSIDER

TECH

Why The New York Times' lawyers are inspecting OpenAI's code in a secretive room

Jacob Shamsian Oct 11, 2024, 11:28 AM CDT

Share Save



OpenAI and Microsoft are defending lawsuits from publishers, authors, and other creators alleging they ripped off their work. They can reshape the generative AI industry. Jack Guez/Getty Images; Jenny Chang-Rodriguez/BI

The Washington Post
Democracy Dies in Darkness

8 major newspapers join legal backlash against OpenAI, Microsoft

The Chicago Tribune, New York Daily News and six other dailies accused the tech companies of taking their work to train AI algorithms

Updated April 30, 2024

3 min ↗ 📌 🗨️ 78

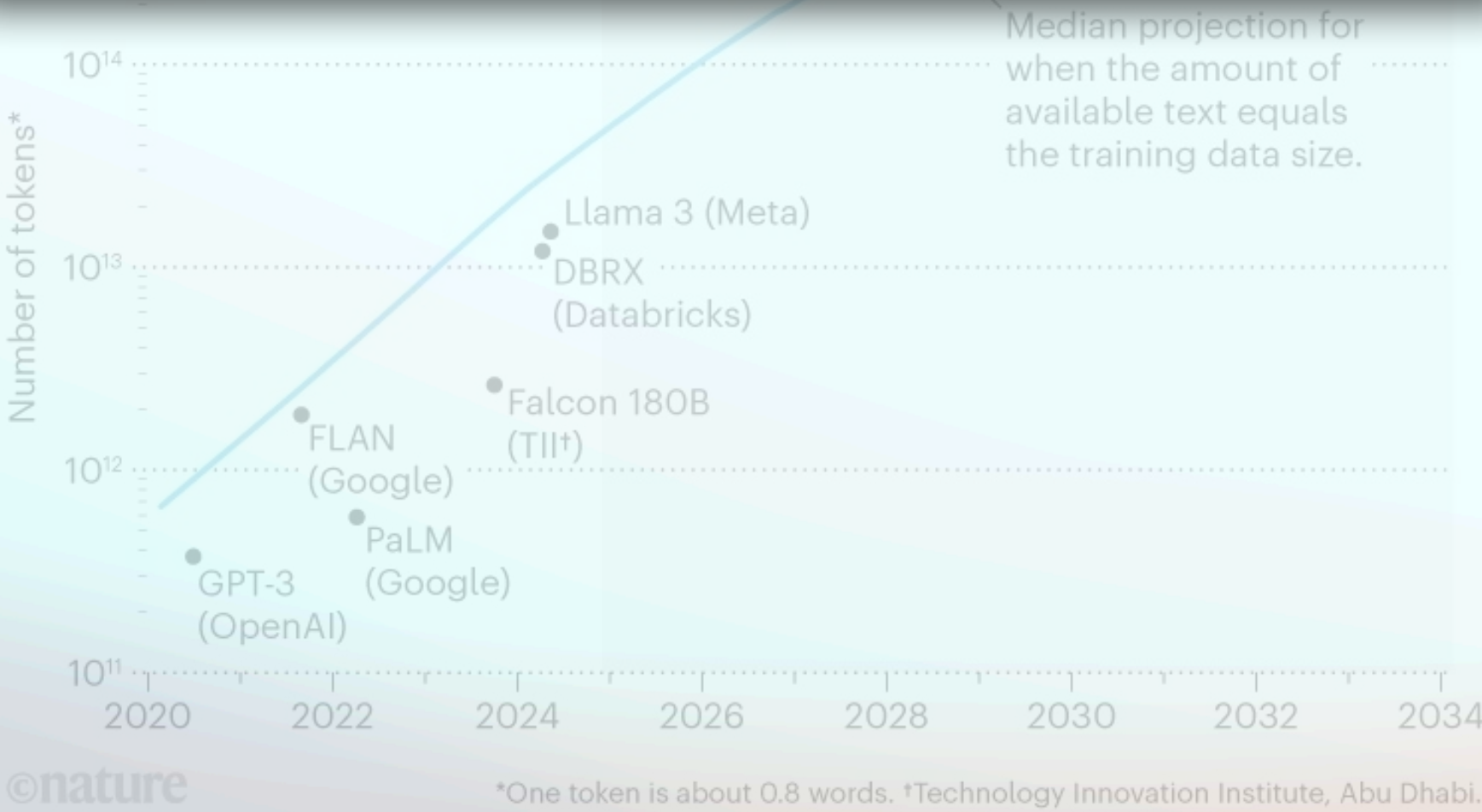
The amount of data that healthcare churns out with each passing second is almost incomprehensible. By 2020, it was estimated that globally the healthcare sector would generate **2.3 zettabytes of data** – that’s 2,300,000,000,000,000,000,000 bytes. Equivalent to 2.3 trillion DVDs of data.

RUNNING OUT OF DATA

The amount of text data used to train large language models (LLMs) is rapidly



- **97% of all data produced by hospitals each year goes unused, but that's beginning to change**



BUSINESS INSIDER

TECH

Why The New York Times' lawyers are inspecting OpenAI's code in a secretive room

Jacob Shamsian Oct 11, 2024, 11:28 AM CDT

Share Save

OpenAI and Microsoft are defending lawsuits from publishers, authors, and other creators alleging they ripped off their work. They can reshape the generative AI industry. Jack Guez/Getty Images; Jenny Chang-Rodriguez/BI

The Washington Post
Democracy Dies in Darkness

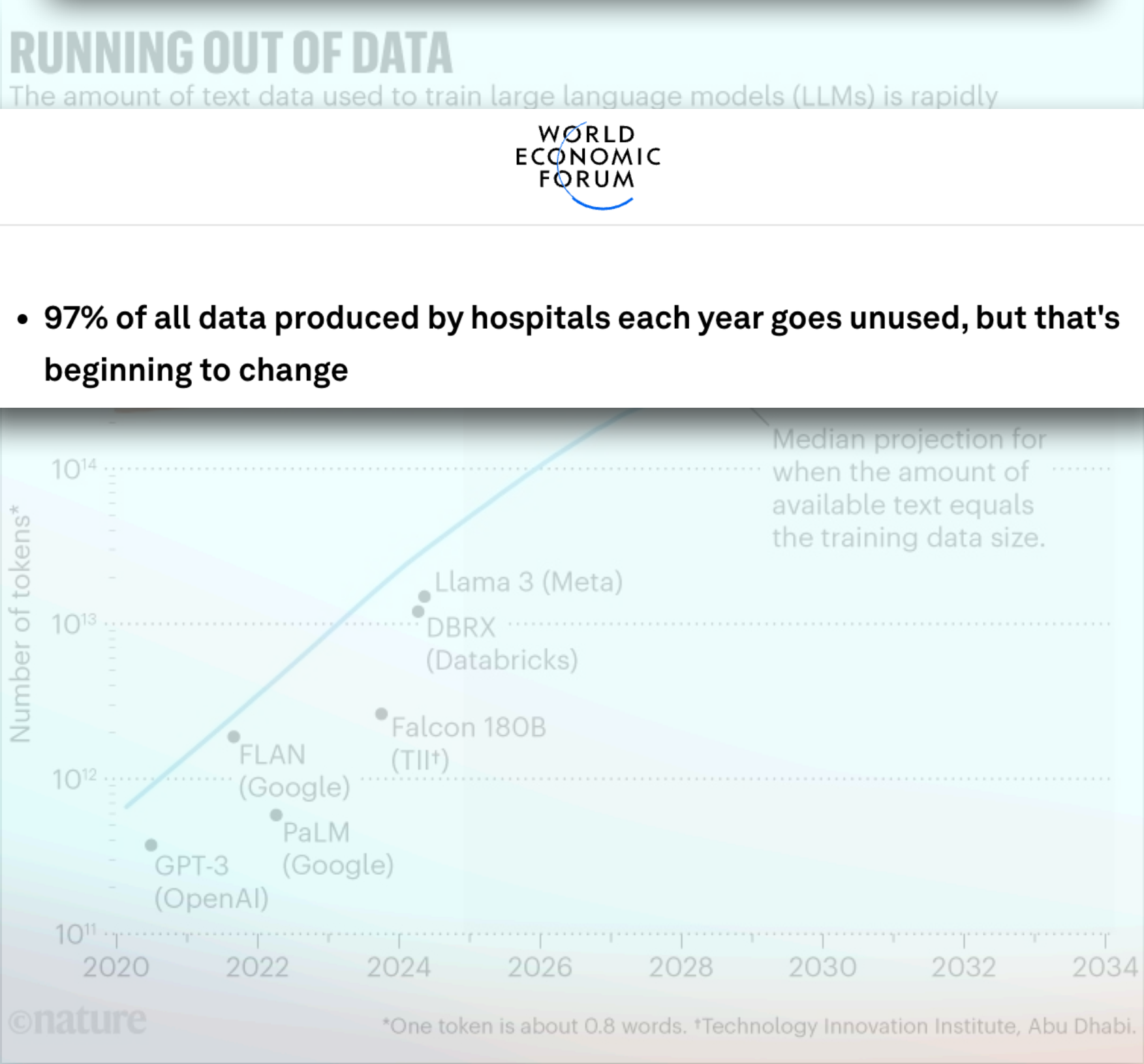
8 major newspapers join legal backlash against OpenAI, Microsoft

The Chicago Tribune, New York Daily News and six other dailies accused the tech companies of taking their work to train AI algorithms

Updated April 30, 2024

3 min Share Bookmark 78

The amount of data that healthcare churns out with each passing second is almost incomprehensible. By 2020, it was estimated that globally the healthcare sector would generate **2.3 zettabytes of data** – that’s 2,300,000,000,000,000,000,000 bytes. Equivalent to 2.3 trillion DVDs of data.



INSIDER

TECH

Why Th inspect

Jacob Shamsian Oct 11, 2024, 11:28 AM CDT

Share Save

OpenAI and Microsoft are defending lawsuits from publishers, authors, and other creators alleging they ripped off their work. They can reshape the generative AI industry. Jack Guez/Getty Images; Jenny Chang-Rodriguez/BI

The Washington Post
Democracy Dies in Darkness

8 major newspapers join legal backlash against OpenAI, Microsoft

The Chicago Tribune, New York Daily News and six other dailies accused the tech companies of taking their work to train AI algorithms

Updated April 30, 2024

3 min 78

The amount of data that healthcare churns out with each passing second is almost incomprehensible. By 2020, it was estimated that globally the healthcare sector would generate **2.3 zettabytes of data** – that’s 2,300,000,000,000,000,000,000 bytes. Equivalent to 2.3 trillion DVDs of data.

RUNNING OUT OF DATA

The amount of text data used to train large language models (LLMs) is rapidly



- **97% of all data produced by hospitals each year goes unused, but that's beginning to change**



Information Blocking

One of the key provisions in Title IV of the 21st Century Cures Act is the prevention of information blocking, or practices that restrict the access, exchange, or use of electronic health information (EHI). HIPAA-covered entities must ensure they are providing proper access to EHI without compromising HIPAA regulations for privacy and security of Protected Health Information (PHI).



Interoperability

The Interoperability clause of Title IV requires “the secure exchange of electronic health information with, and use of electronic health information from, other health information technology without special effort on the part of the user”. In 2020, HHS and ONC [issued a Final Rule](#) regarding how this clause was going to be implemented via the use of APIs.



OpenAI and Microsoft are defending lawsuits from publishers, authors, and other creators alleging they ripped off their work. They can reshape the generative AI industry. Jack Guez/Getty Images; Jenny Chang-Rodriguez/BI

The Washington Post
Democracy Dies in Darkness

8 major newspapers join legal backlash against OpenAI, Microsoft

The Chicago Tribune, New York Daily News and six other dailies accused the tech companies of taking their work to train AI algorithms

Updated April 30, 2024

3 min 78

The amount of data that healthcare churns out with each passing second is almost incomprehensible. By 2020, it was estimated that globally the healthcare sector would generate [2.3 zettabytes of data](#) – that’s 2,300,000,000,000,000,000 bytes. Equivalent to 2.3 trillion DVDs of data.

RUNNING OUT OF DATA

The amount of text data used to train large language models (LLMs) is rapidly



- **97% of all data produced by hospitals each year goes unused, but that's beginning to change**

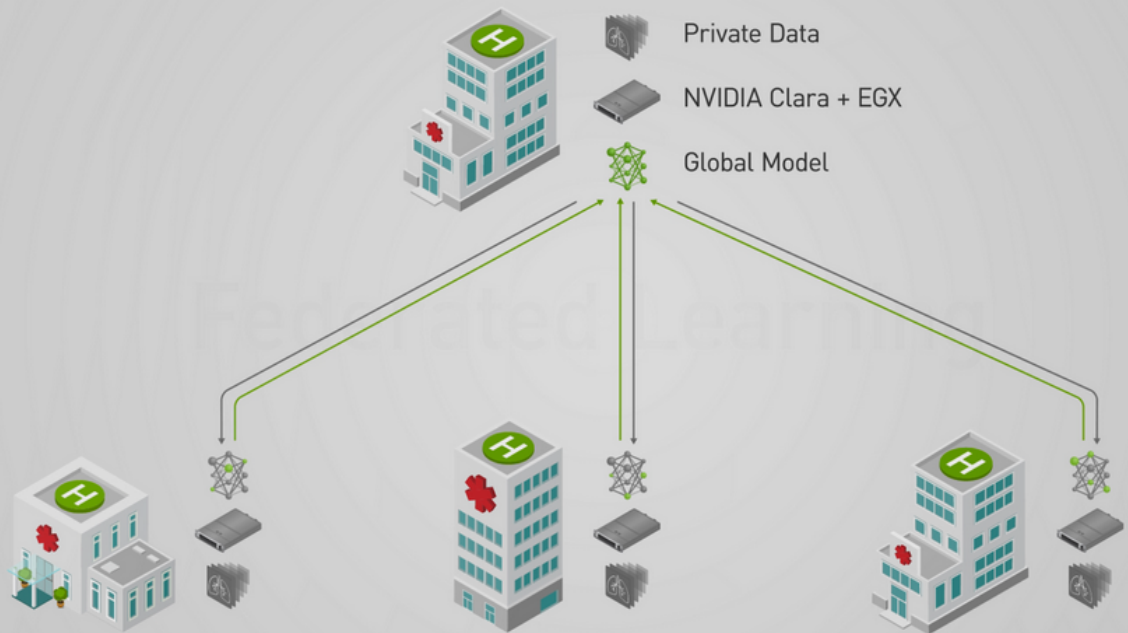


Information Blocking

One of the key provisions in Title IV of the 21st Century Cures Act is the prevention of information blocking, or practices that restrict the access, exchange, or use of electronic health information (EHI). HIPAA-covered entities must ensure they are providing proper access to EHI without compromising HIPAA regulations for privacy and security of Protected Health Information (PHI).

Interoperability

The Interoperability clause of Title IV requires “the secure exchange of electronic health information with, and use of electronic health information from, other health information technology without special effort on the part of the user”. In 2020, HHS and ONC [issued a Final Rule](#) regarding how this clause was going to be implemented via the use of APIs.



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter



4PB

of data created by Facebook, including

350m

photos

100m

hours of video watch time

Facebook Research

DEMYSTIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
b	bit	0 or 1
B	byte	8 bits
KB	kilobyte	1,000 bytes
MB	megabyte	1,000 ² bytes
GB	gigabyte	1,000 ³ bytes
TB	terabyte	1,000 ⁴ bytes
PB	petabyte	1,000 ⁵ bytes
EB	exabyte	1,000 ⁶ bytes
ZB	zettabyte	1,000 ⁷ bytes
YB	yottabyte	1,000 ⁸ bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020



4TB

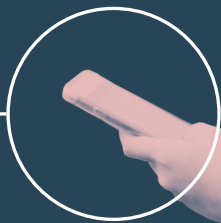
of data produced by a connected car

Intel

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



3.9bn

people use emails

4TB

of data produced by a connected car

Intel

Searches made a day

5bn

Searches made a day from Google

3.5bn

Smart insights



ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

44ZB

PwC

2019

2020

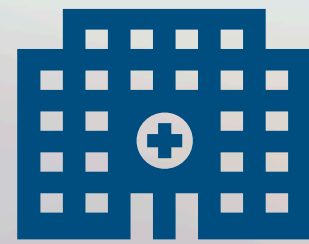
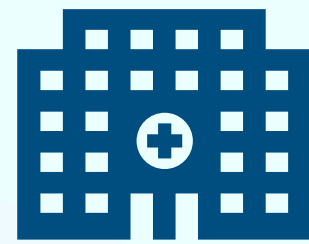
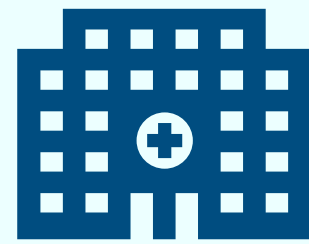
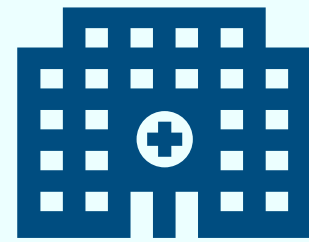
28PB

to be generated from wearable devices by 2020

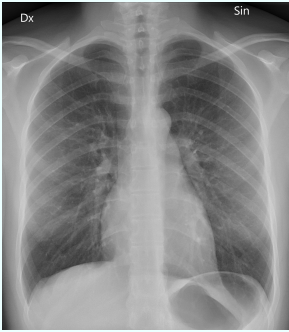
Statista



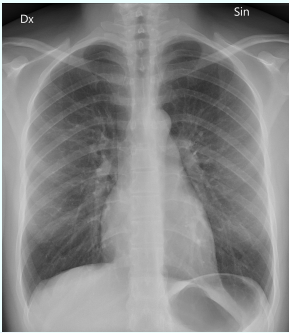
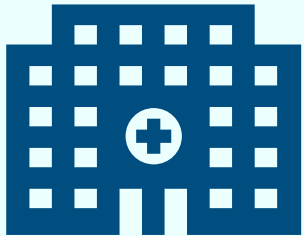
Multiple
hospitals



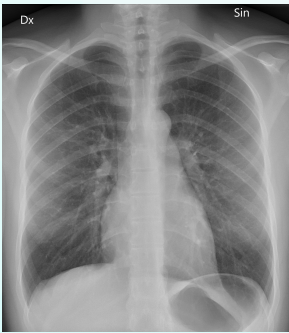
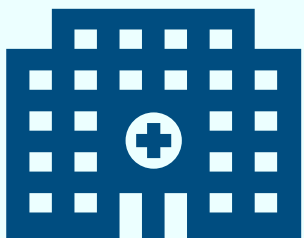
Multiple hospitals



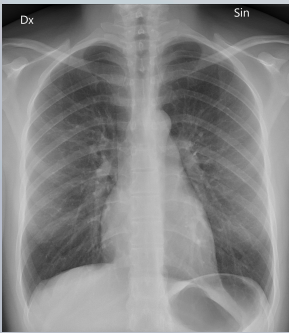
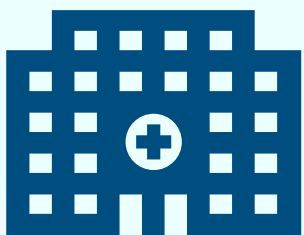
$$S_1 \sim \mathcal{D}^{\otimes n_1}$$



$$S_2 \sim \mathcal{D}^{\otimes n_2}$$



$$S_3 \sim \mathcal{D}^{\otimes n_3}$$

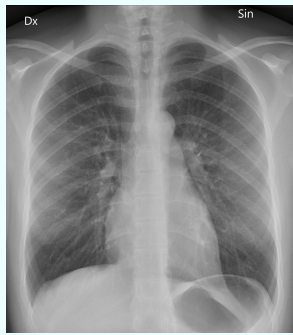


$$S_M \sim \mathcal{D}^{\otimes n_M}$$

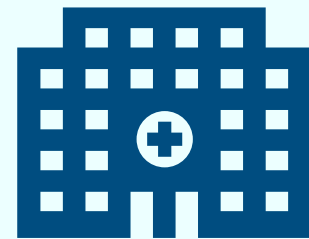


Chest X-ray images

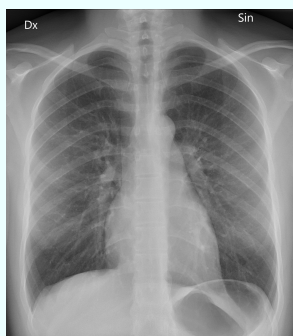
Multiple
hospitals



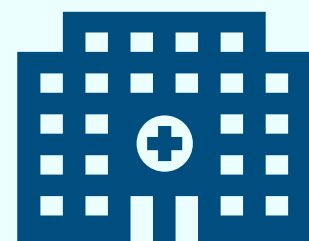
$$S_1 \sim \mathcal{D}^{\otimes n_1}$$



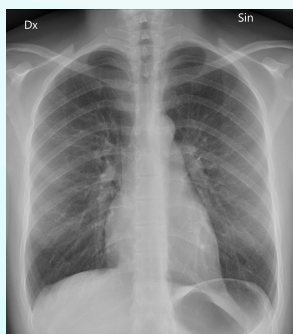
$$\min_w \mathbb{E}_{z \sim \mathcal{T}_1} [f_1(w; z)]$$



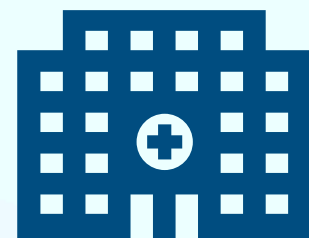
$$S_2 \sim \mathcal{D}^{\otimes n_2}$$



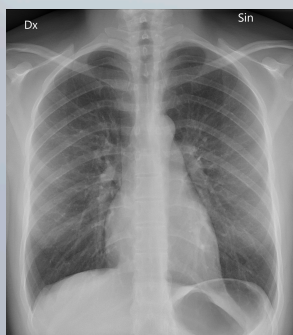
$$\min_w \mathbb{E}_{z \sim \mathcal{T}_2} [f_2(w; z)]$$



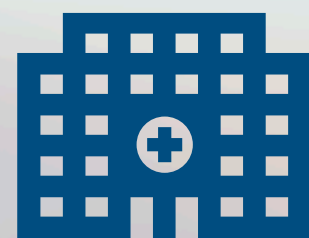
$$S_3 \sim \mathcal{D}^{\otimes n_3}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}_3} [f_3(w; z)]$$



$$S_M \sim \mathcal{D}^{\otimes n_M}$$



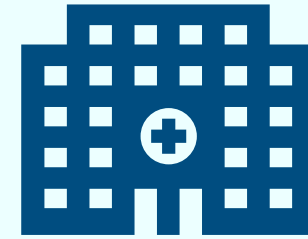
$$\min_w \mathbb{E}_{z \sim \mathcal{T}_M} [f_M(w; z)]$$

Chest X-ray images

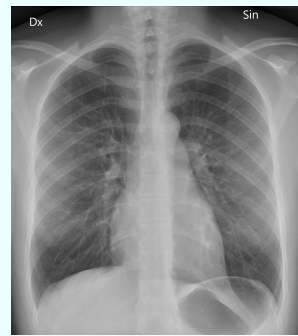
Multiple
hospitals



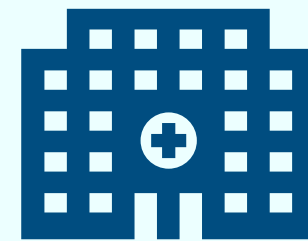
$$S_1 \sim \mathcal{D}^{\otimes n_1}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}_1} [f_1(w; z)]$$



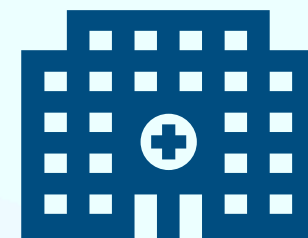
$$S_2 \sim \mathcal{D}^{\otimes n_2}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}_2} [f_2(w; z)]$$

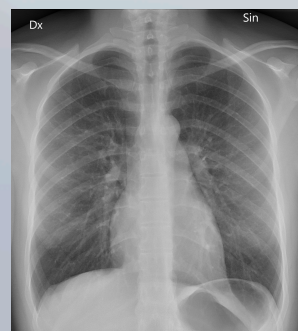


$$S_3 \sim \mathcal{D}^{\otimes n_3}$$

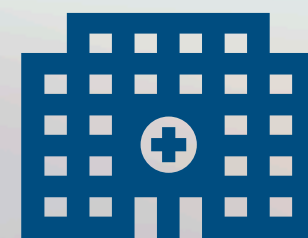


$$\min_w \mathbb{E}_{z \sim \mathcal{T}_3} [f_3(w; z)]$$

⋮



$$S_M \sim \mathcal{D}^{\otimes n_M}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}_M} [f_M(w; z)]$$

Chest X-ray images

Many possible downstream tasks:

- Disease classification
- Localizing infected regions
- Predicting severity/progression
- Policies for decision making
- Or, learning some general purpose embeddings

Multiple
hospitals

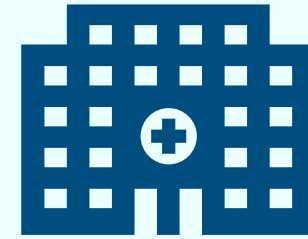
Task Heterogeneity
- Different losses
- Different targets

Many possible downstream tasks:

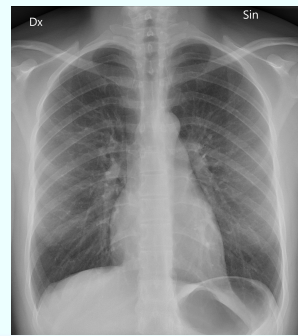
- Disease classification
- Localizing infected regions
- Predicting severity/progression
- Policies for decision making
- Or, learning some general purpose embeddings



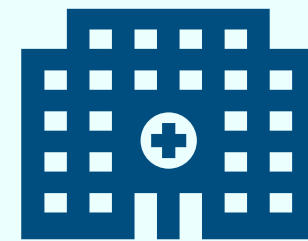
$$S_1 \sim \mathcal{D}^{\otimes n_1}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}_1} [f_1(w; z)]$$



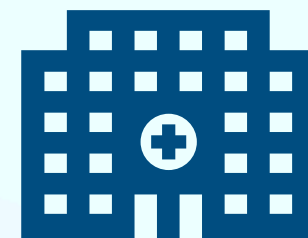
$$S_2 \sim \mathcal{D}^{\otimes n_2}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}_2} [f_2(w; z)]$$

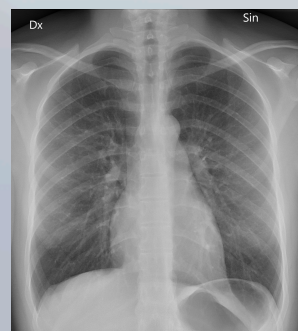


$$S_3 \sim \mathcal{D}^{\otimes n_3}$$

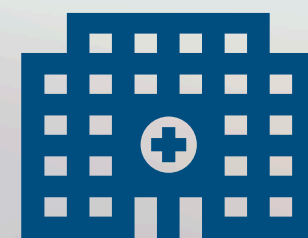


$$\min_w \mathbb{E}_{z \sim \mathcal{T}_3} [f_3(w; z)]$$

⋮



$$S_M \sim \mathcal{D}^{\otimes n_M}$$

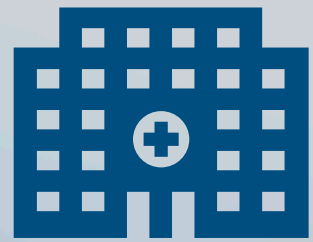
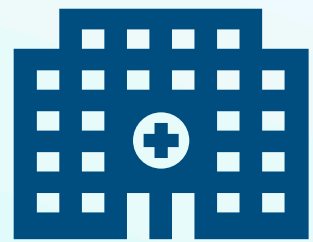
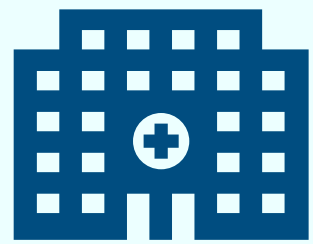
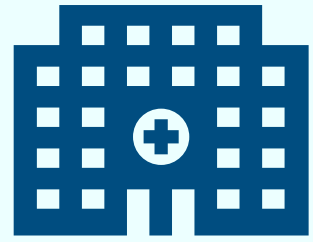


$$\min_w \mathbb{E}_{z \sim \mathcal{T}_M} [f_M(w; z)]$$

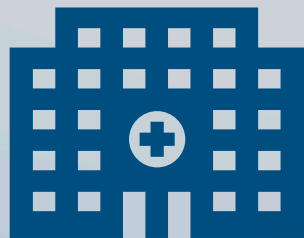
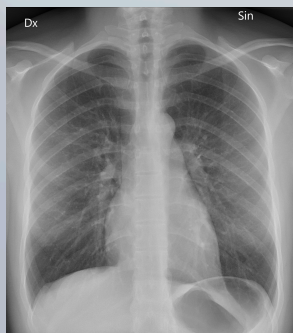
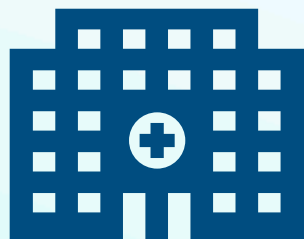
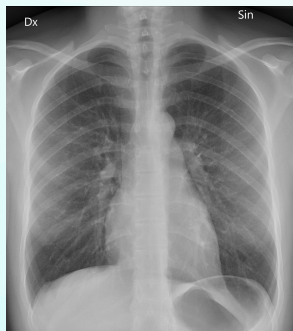
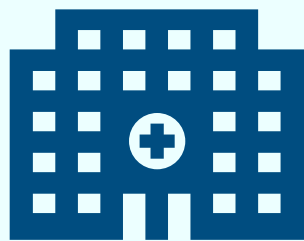
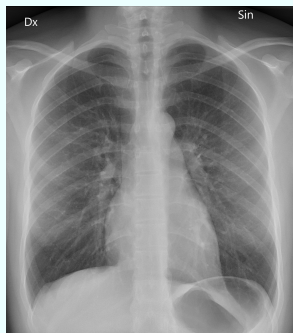
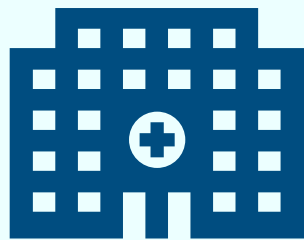
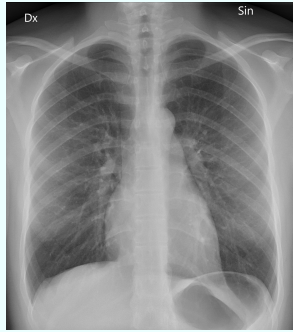
Chest X-ray images

One natural solution to deal with potentially unforeseeable task heterogeneity is to train a single model to **generate synthetic data** for all clients

Multiple
hospitals

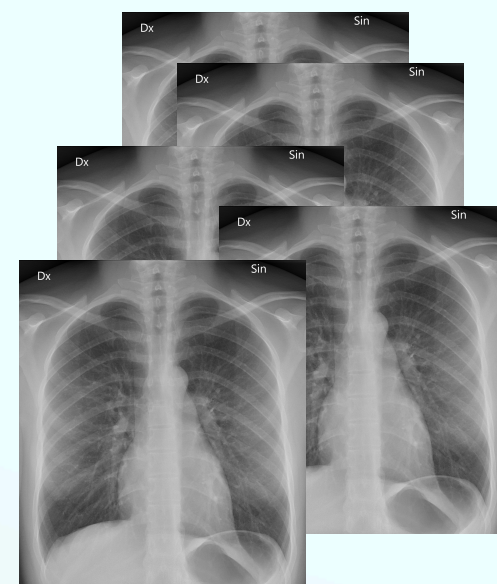
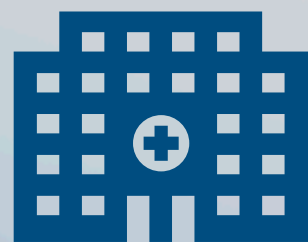
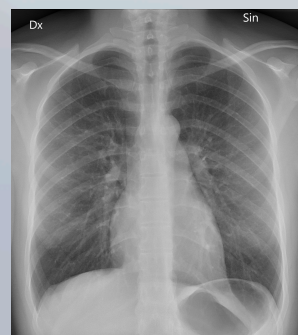
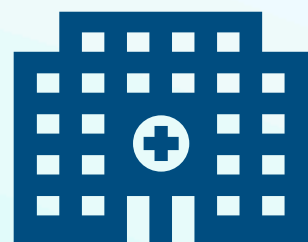
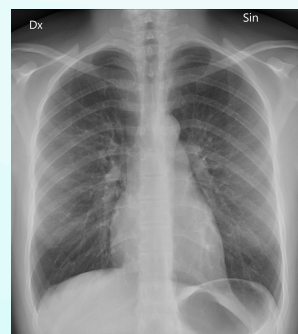
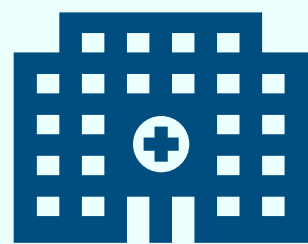
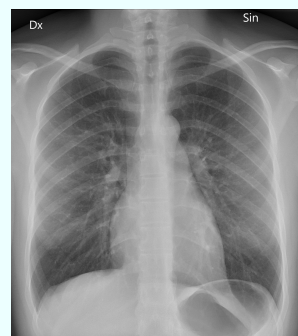
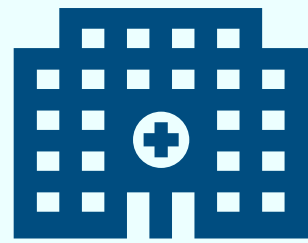
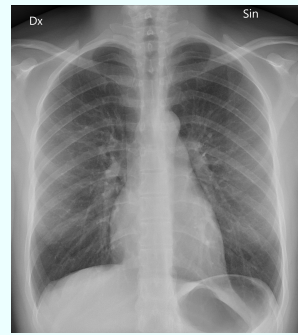


Multiple
hospitals



Clean chest
X-ray images

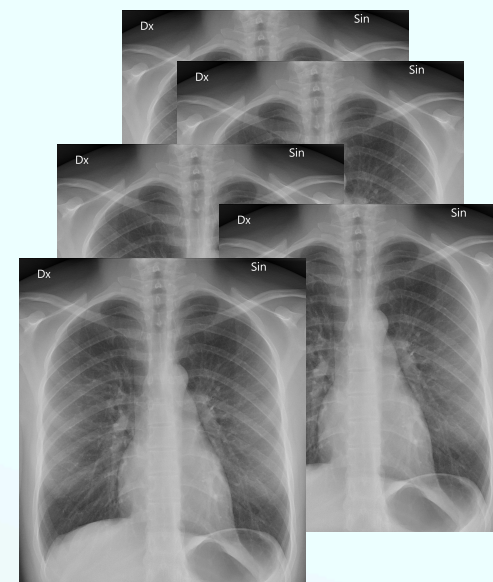
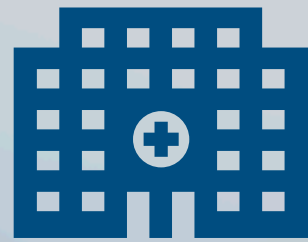
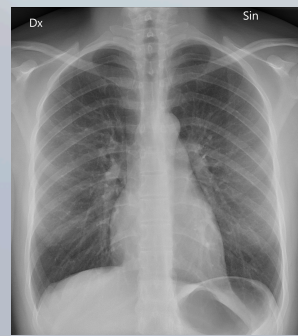
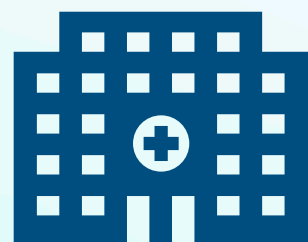
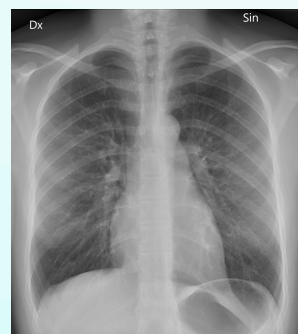
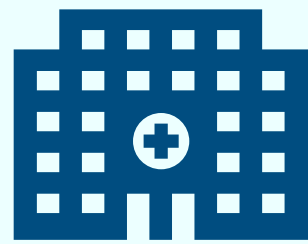
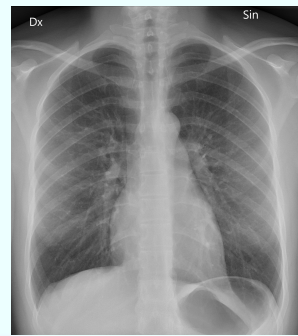
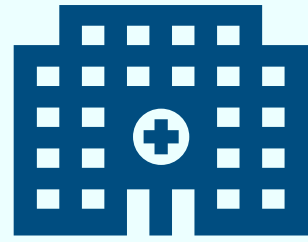
Multiple hospitals



Combined dataset

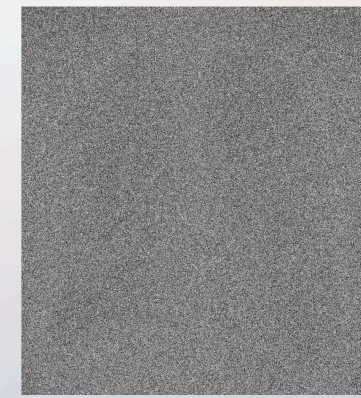
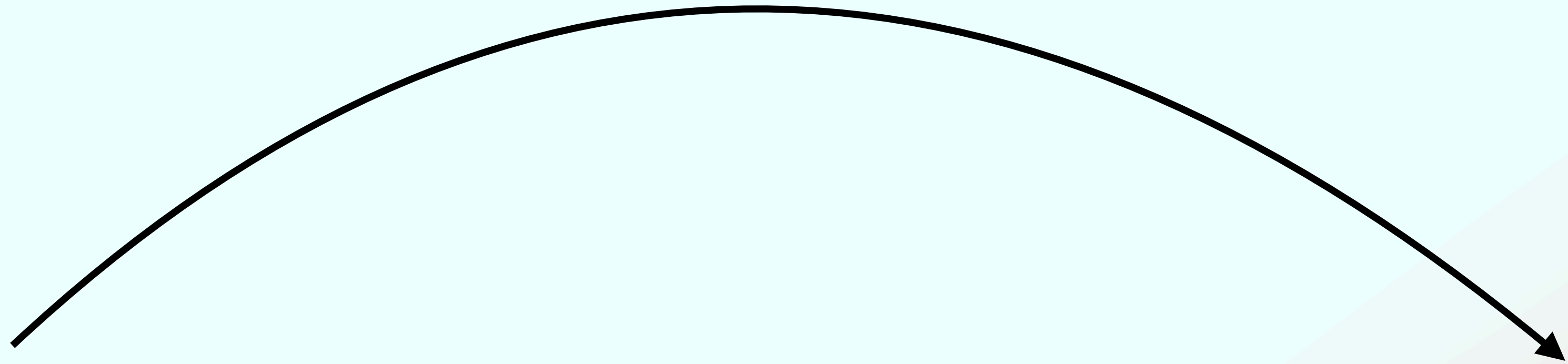
Clean chest X-ray images

Multiple hospitals



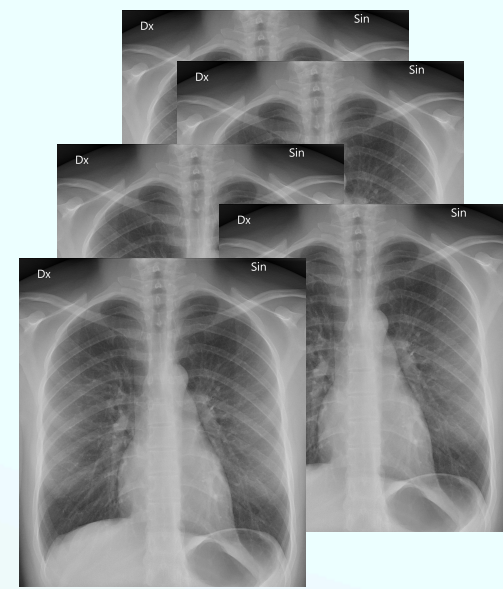
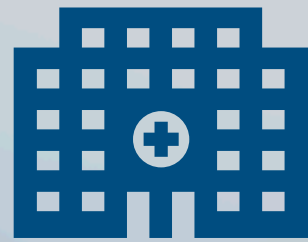
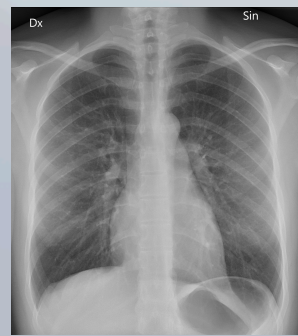
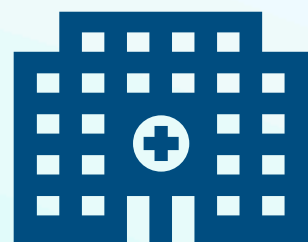
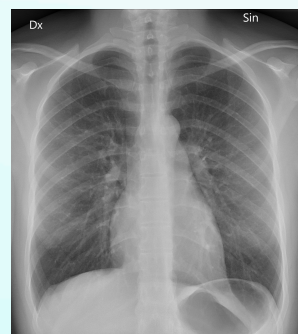
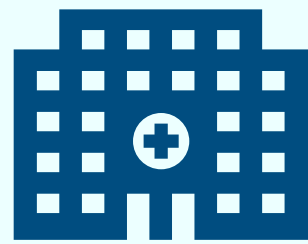
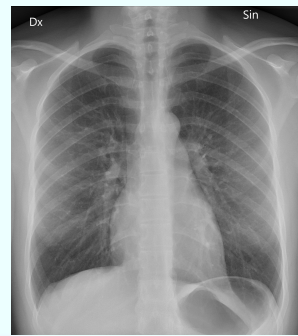
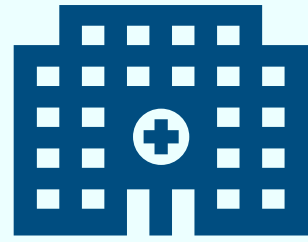
Combined dataset

Known forward diffusion process maps clean images to pure noise



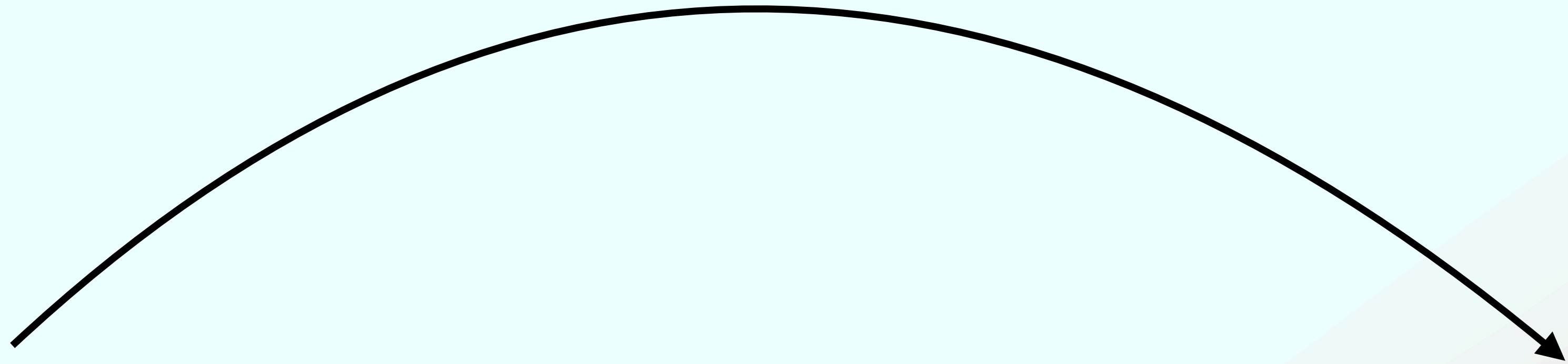
Clean chest X-ray images

Multiple hospitals

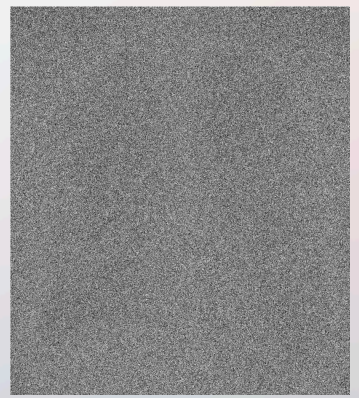


Combined dataset

Known forward diffusion process maps clean images to pure noise

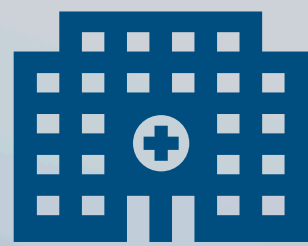
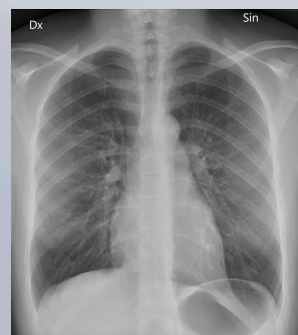
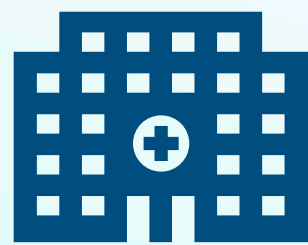
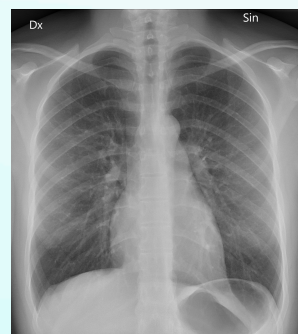
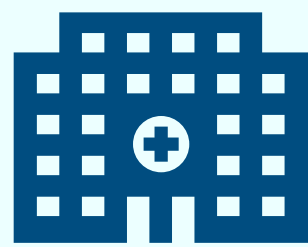
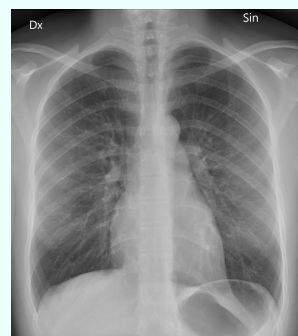
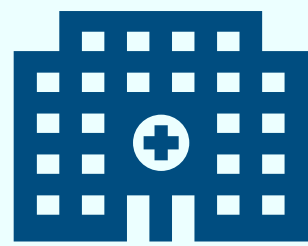
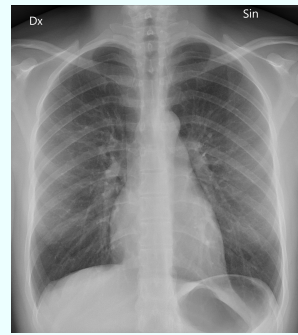


Some noise distribution that is easy to sample from



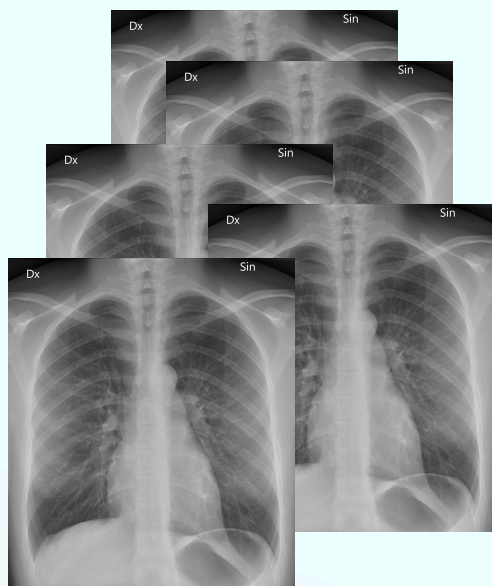
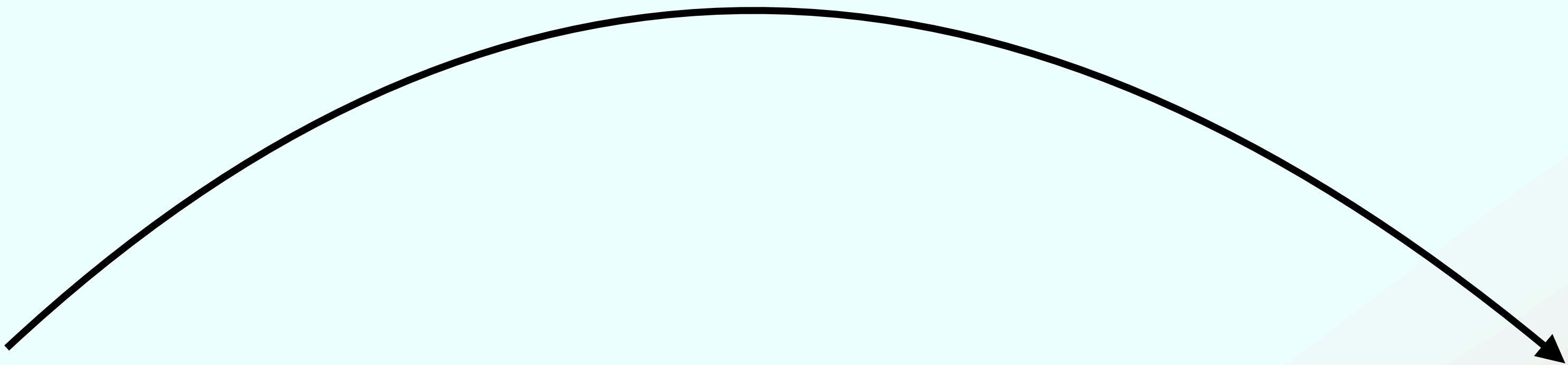
Clean chest X-ray images

Multiple hospitals



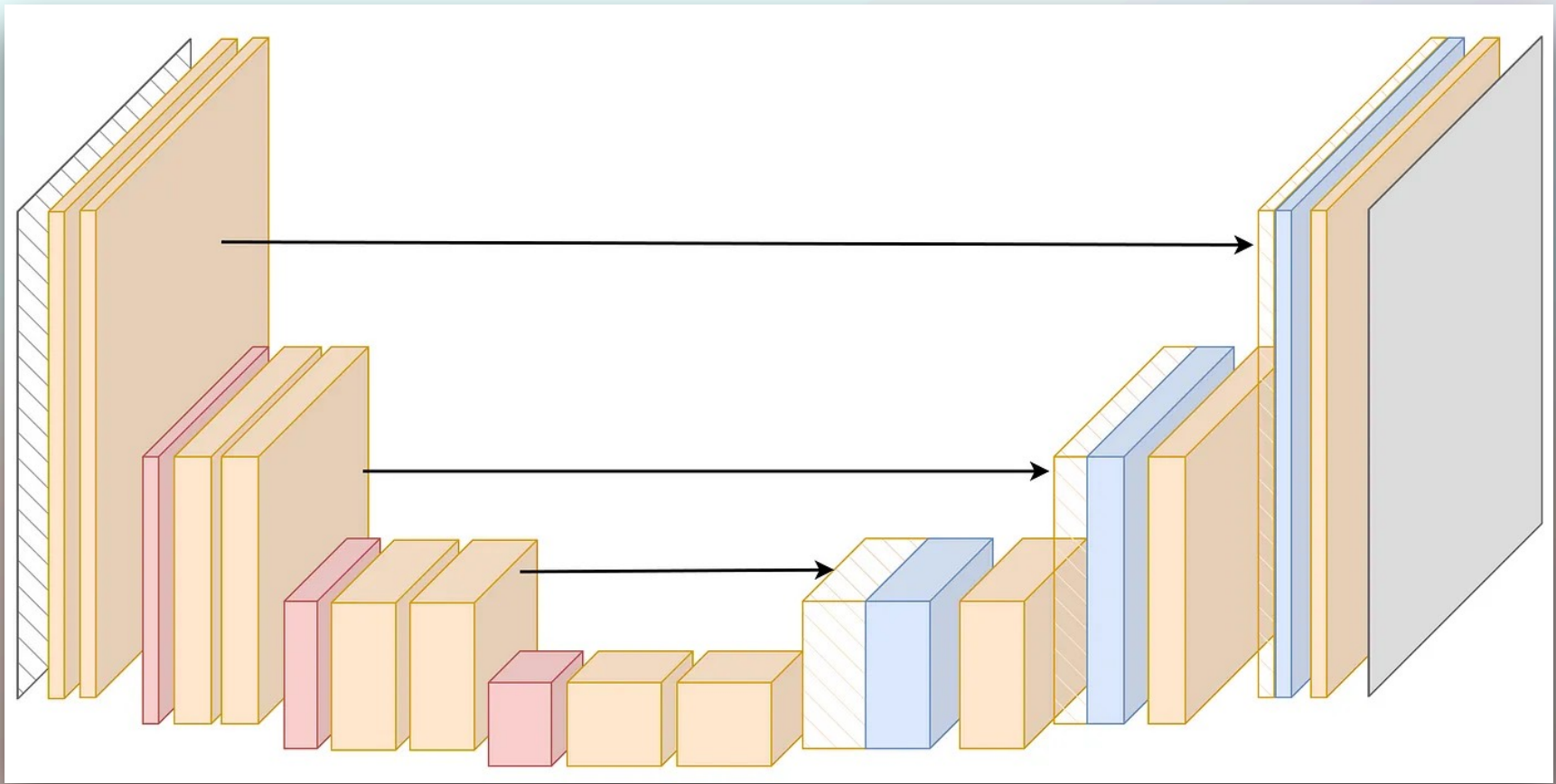
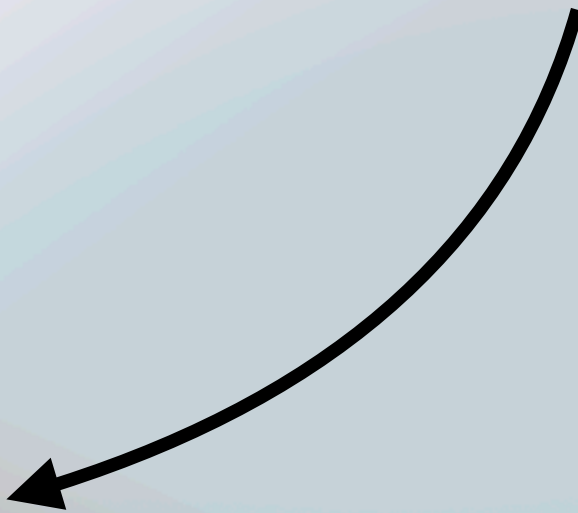
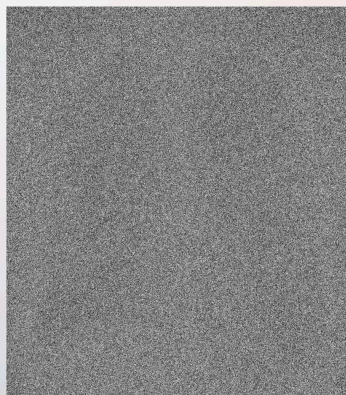
Clean chest X-ray images

Known forward diffusion process maps clean images to pure noise



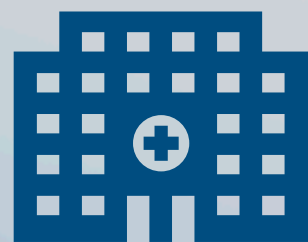
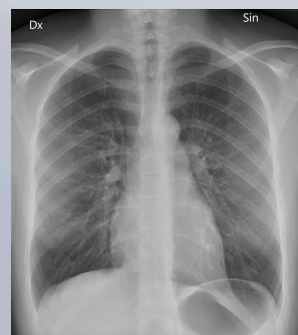
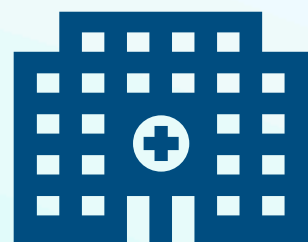
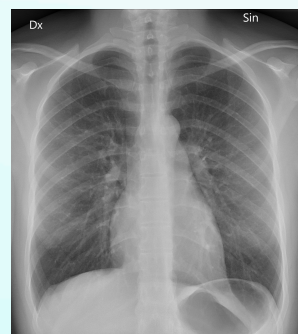
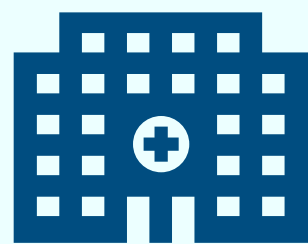
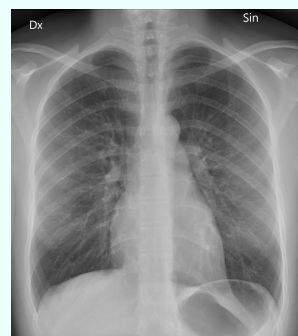
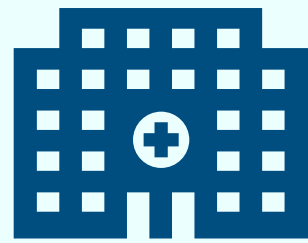
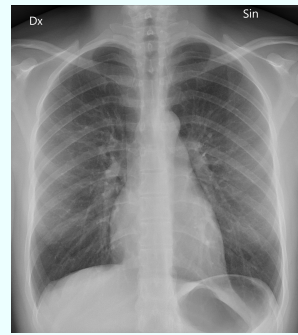
Combined dataset

Some noise distribution that is easy to sample from

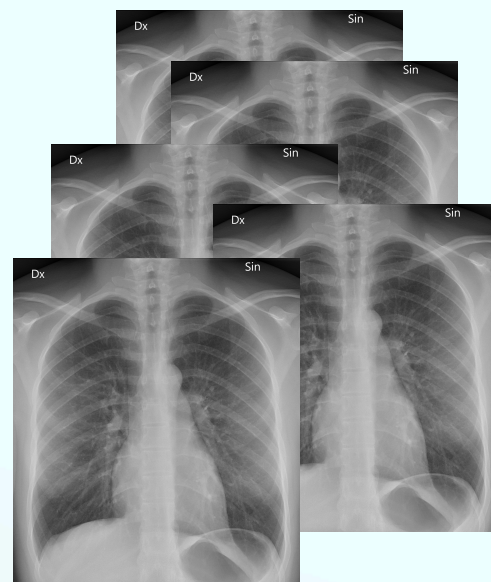


Diffusion model learns to predict this noise

Multiple hospitals



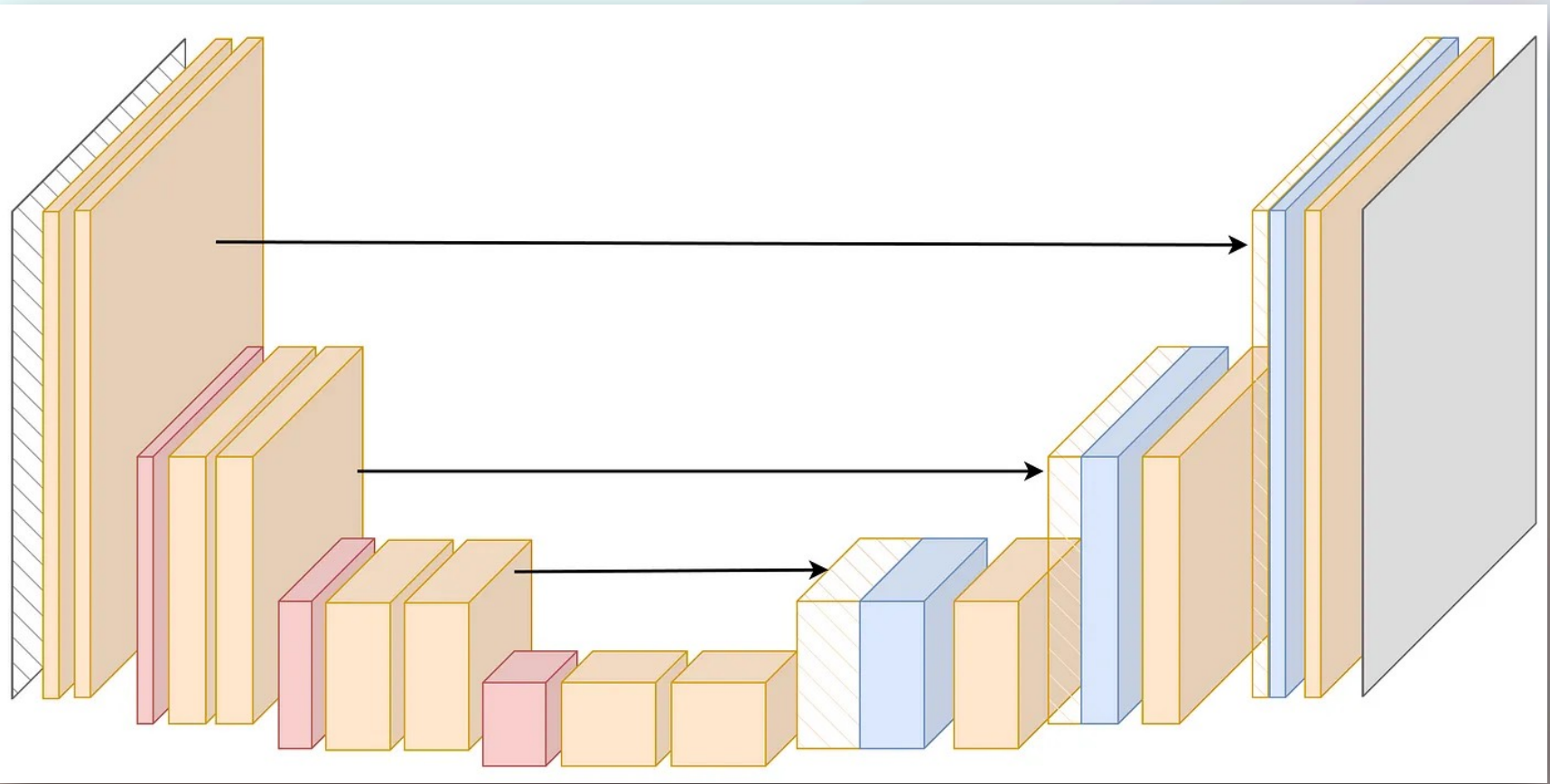
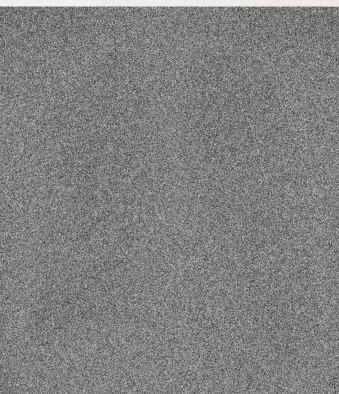
Clean chest X-ray images



Combined dataset

Known forward diffusion process maps clean images to pure noise

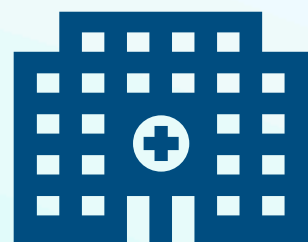
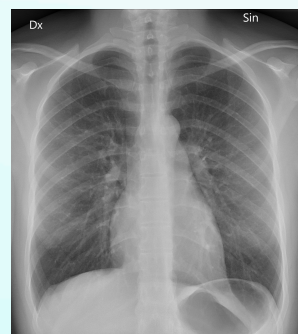
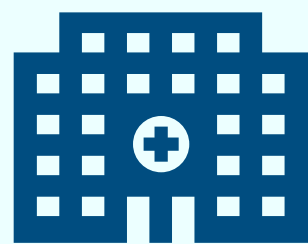
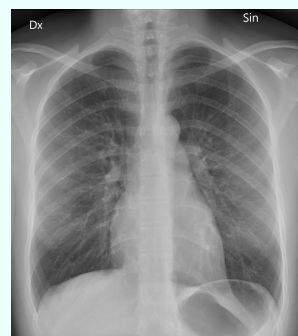
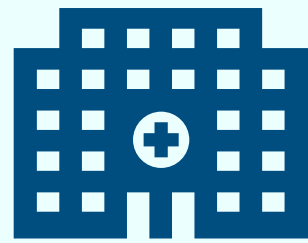
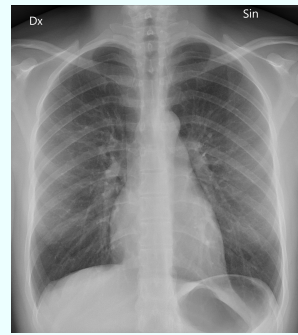
Some noise distribution that is easy to sample from



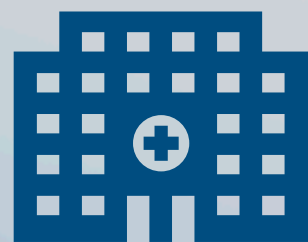
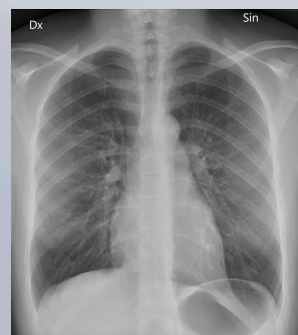
Diffusion model learns to predict this noise

Random Seed

Multiple hospitals

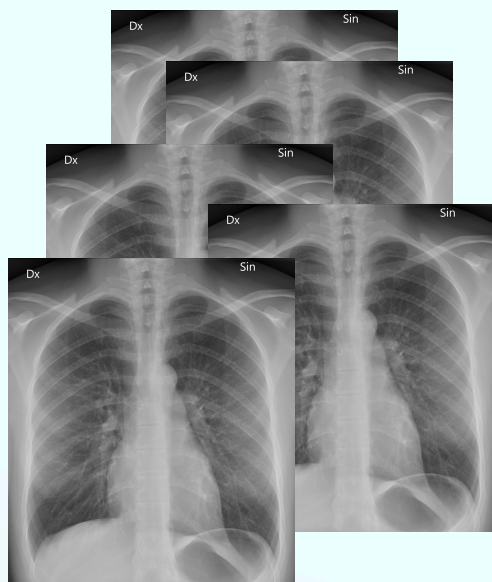
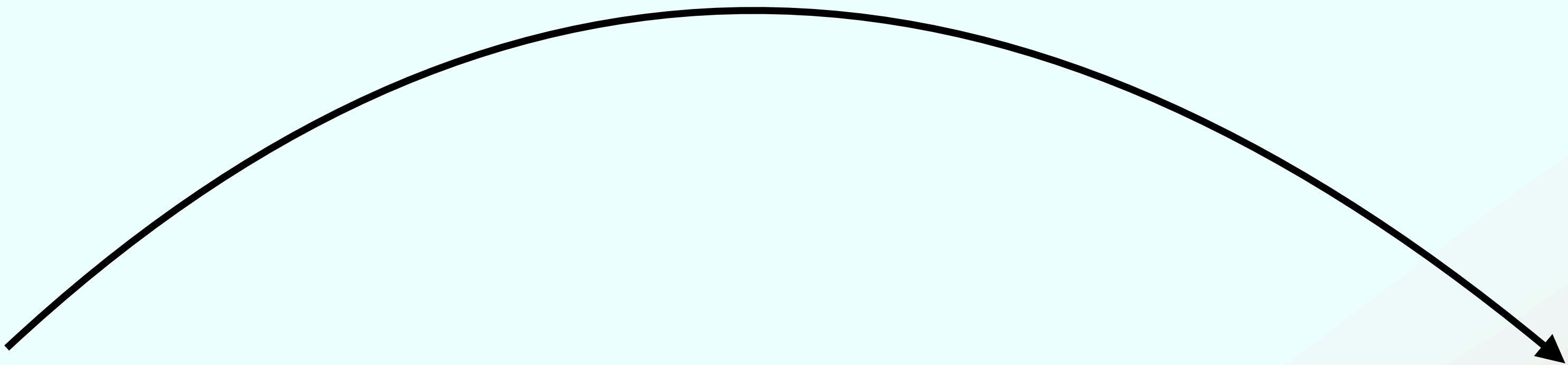


⋮



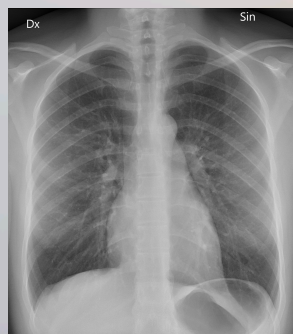
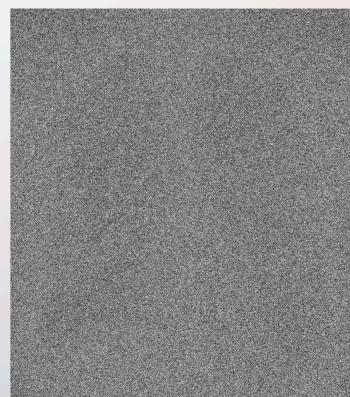
Clean chest X-ray images

Known forward diffusion process maps clean images to pure noise

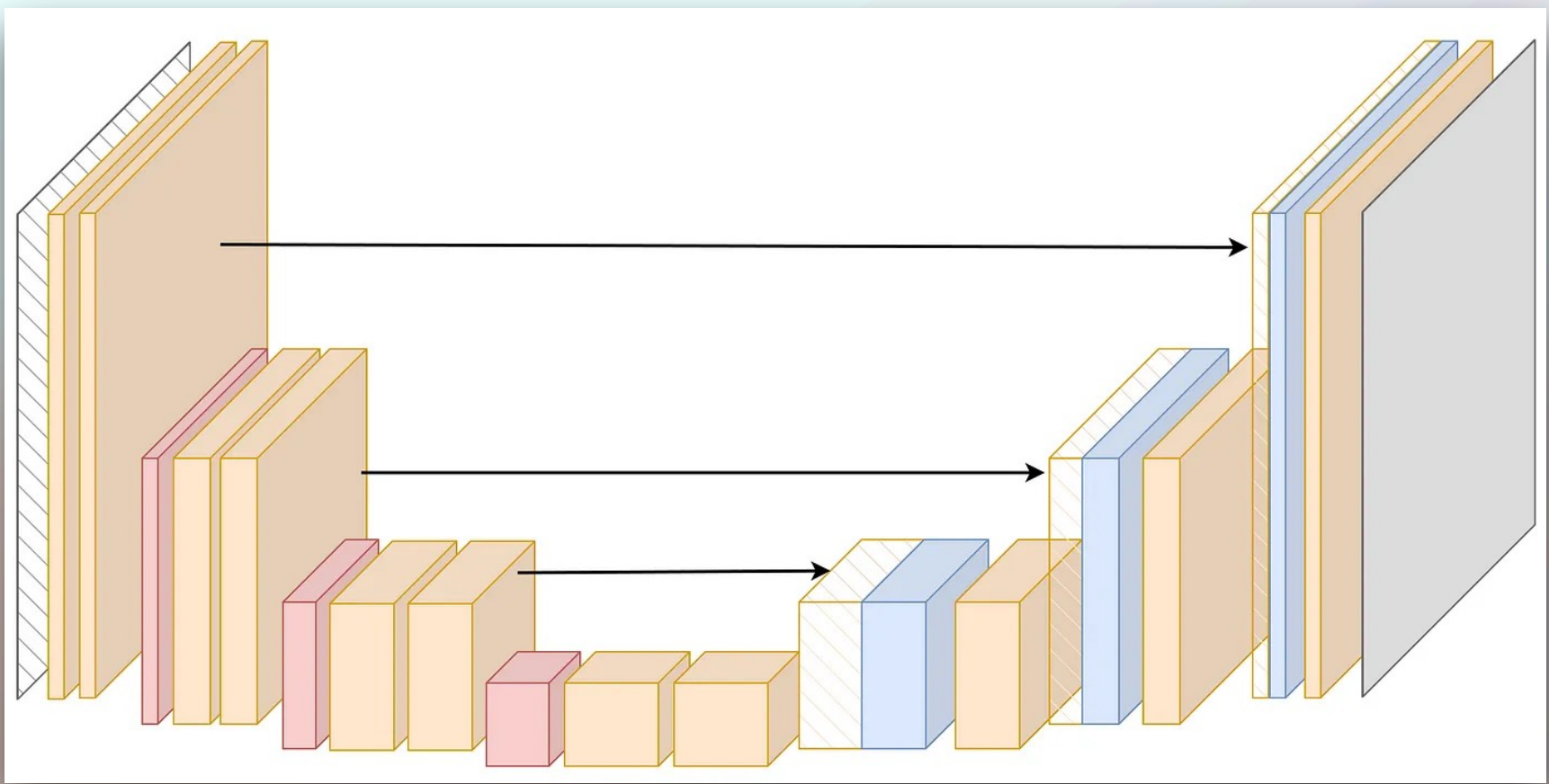


Combined dataset

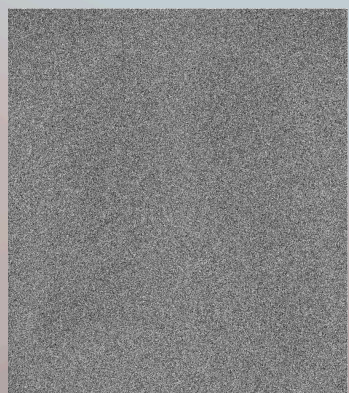
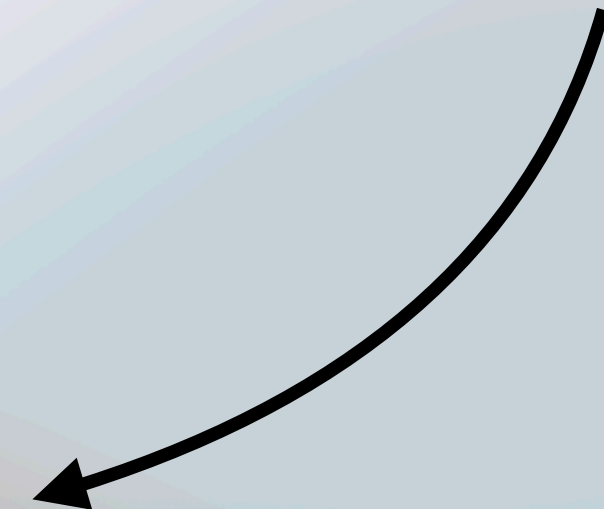
Some noise distribution that is easy to sample from



Synthetic image



Diffusion model learns to predict this noise



Random Seed

This has no privacy guarantee and the global model can be attacked to reveal sensitive information



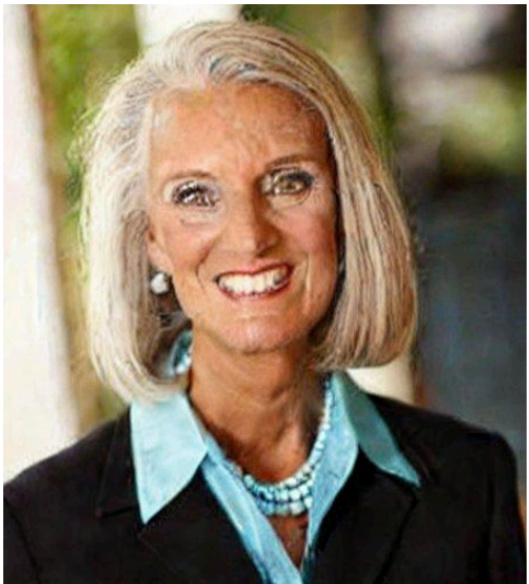
Extracting Training Data from Diffusion Models

Nicholas Carlini^{*1} *Jamie Hayes*^{*2} *Milad Nasr*^{*1}
Matthew Jagielski⁺¹ *Vikash Sehwal*⁺⁴ *Florian Tramèr*⁺³
Borja Balle^{†2} *Daphne Ippolito*^{†1} *Eric Wallace*^{†5}
¹Google ²DeepMind ³ETHZ ⁴Princeton ⁵UC Berkeley
^{*}Equal contribution ⁺Equal contribution [†]Equal contribution

Extracting Training Data from Diffusion Models

Nicholas Carlini*¹ Jamie Hayes*² Milad Nasr*¹
Matthew Jagielski⁺¹ Vikash Sehwal⁺⁴ Florian Tramèr⁺³
Borja Balle^{†2} Daphne Ippolito^{†1} Eric Wallace^{†5}
¹Google ²DeepMind ³ETHZ ⁴Princeton ⁵UC Berkeley
*Equal contribution +Equal contribution †Equal contribution

Training Set Generated Image



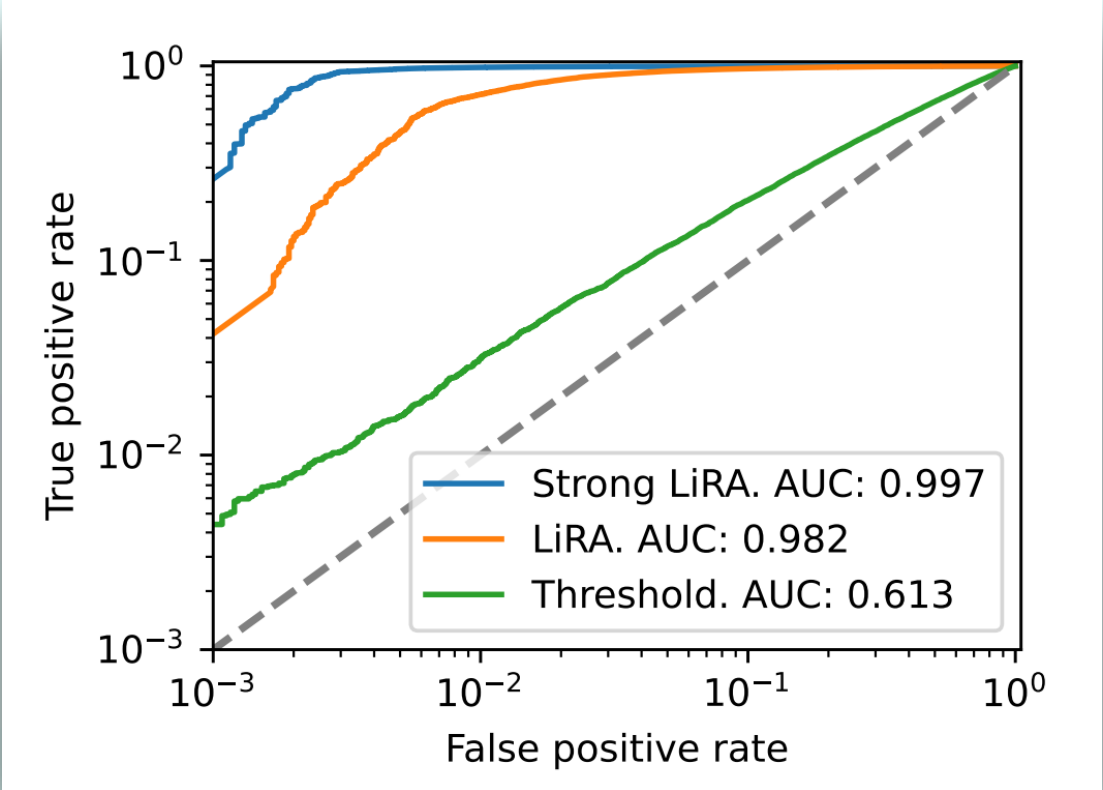
Caption: Living in the light
with Ann Graham Lotz

Prompt:
Ann Graham Lotz

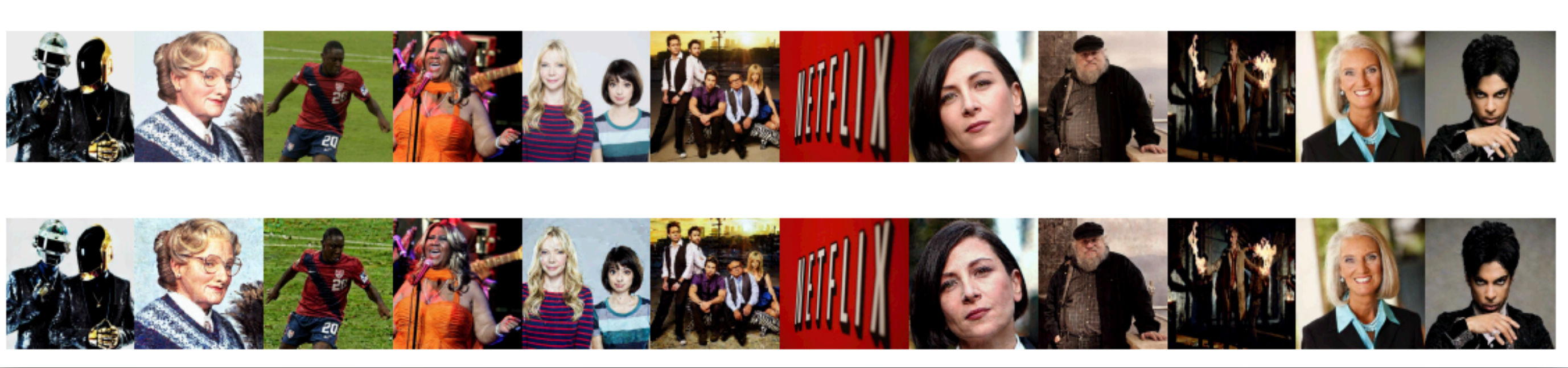
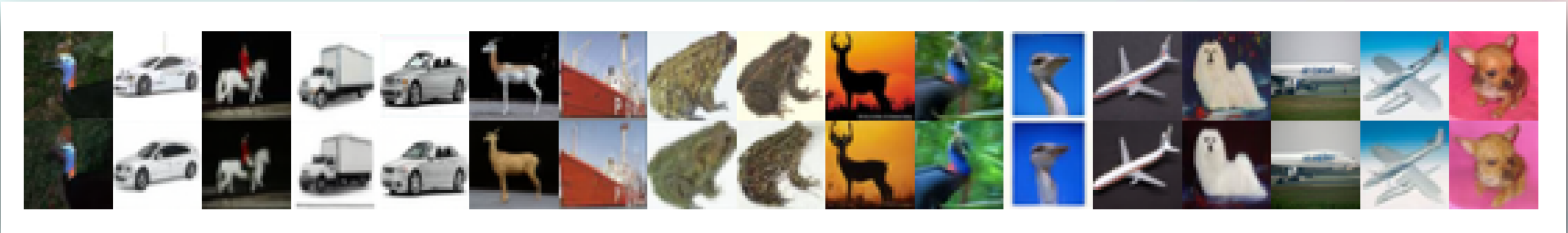


Extracting Training Data from Diffusion Models

Nicholas Carlini*¹ Jamie Hayes*² Milad Nasr*¹
Matthew Jagielski⁺¹ Vikash Sehwal⁺⁴ Florian Tramèr⁺³
Borja Balle^{†2} Daphne Ippolito^{†1} Eric Wallace^{†5}
¹Google ²DeepMind ³ETHZ ⁴Princeton ⁵UC Berkeley
*Equal contribution +Equal contribution †Equal contribution

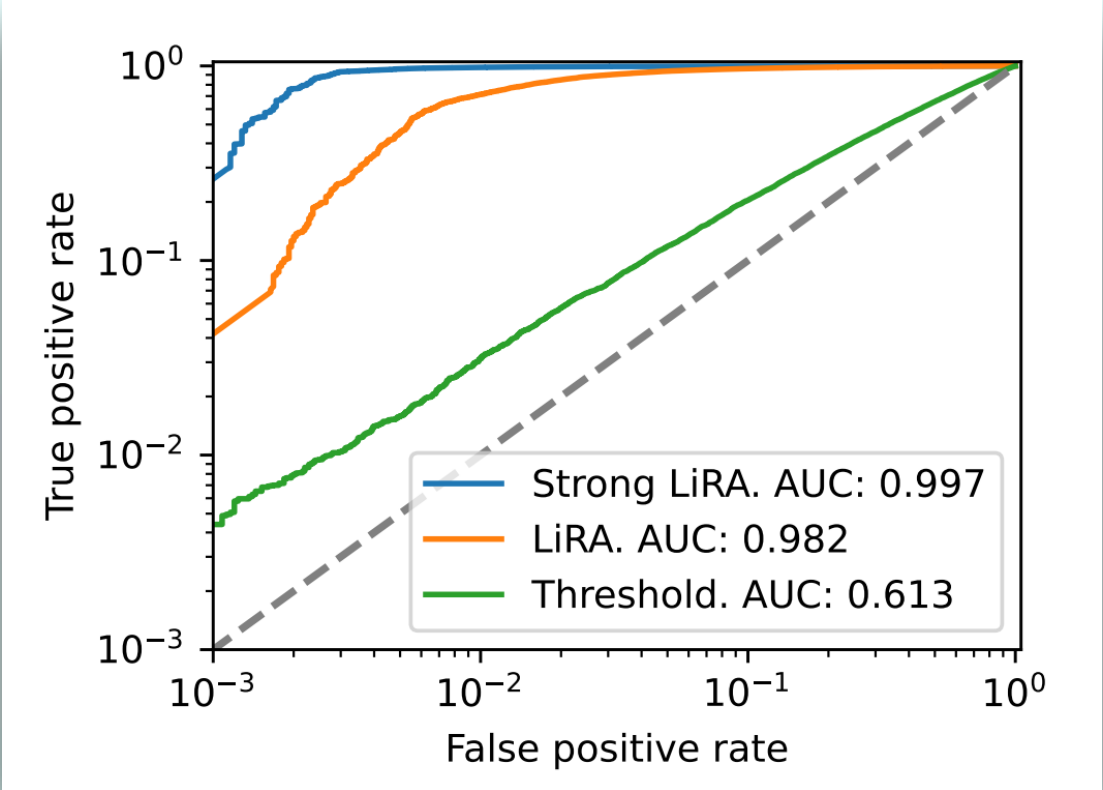


Training Set	Generated Image
	
Caption: Living in the light with Ann Graham Lotz	Prompt: Ann Graham Lotz



Extracting Training Data from Diffusion Models

Nicholas Carlini*¹ Jamie Hayes*² Milad Nasr*¹
Matthew Jagielski⁺¹ Vikash Sehwal⁺⁴ Florian Tramèr⁺³
Borja Balle^{†2} Daphne Ippolito^{†1} Eric Wallace^{†5}
¹Google ²DeepMind ³ETHZ ⁴Princeton ⁵UC Berkeley
*Equal contribution +Equal contribution †Equal contribution

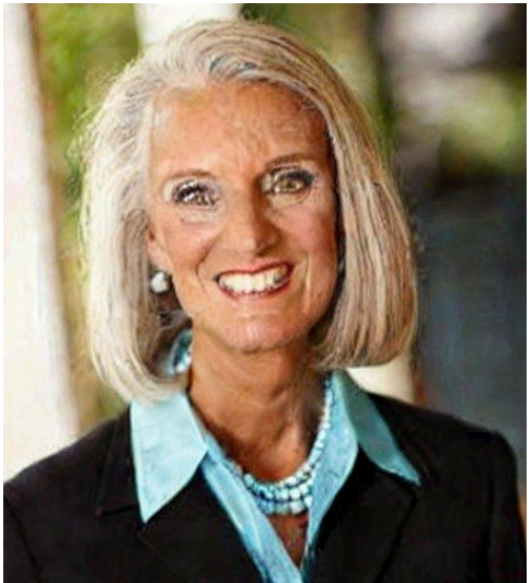


Training Set

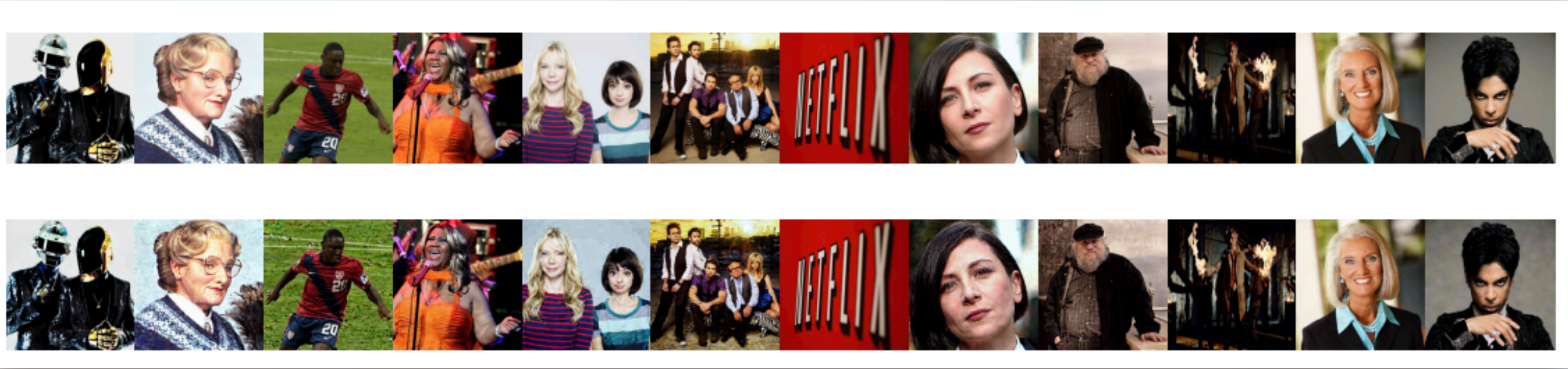
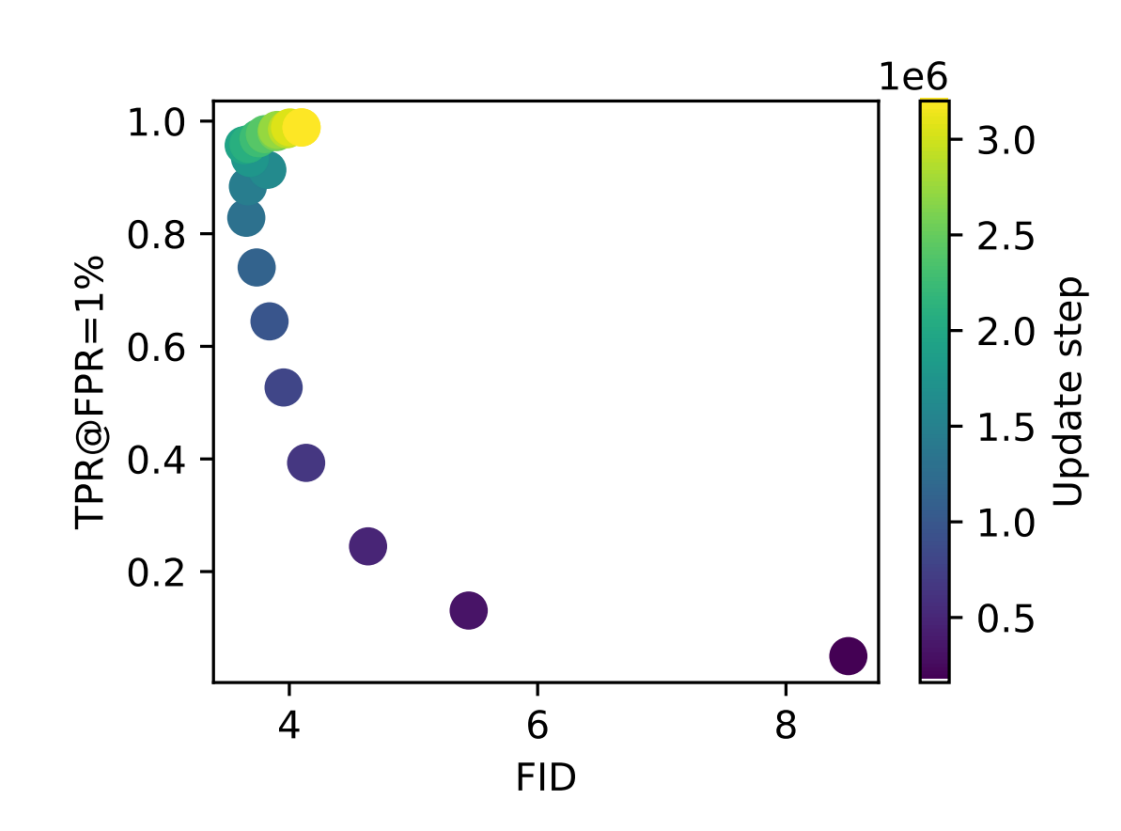


Caption: Living in the light with Ann Graham Lotz

Generated Image



Prompt: Ann Graham Lotz



Extracting Training Data from Diffusion Models

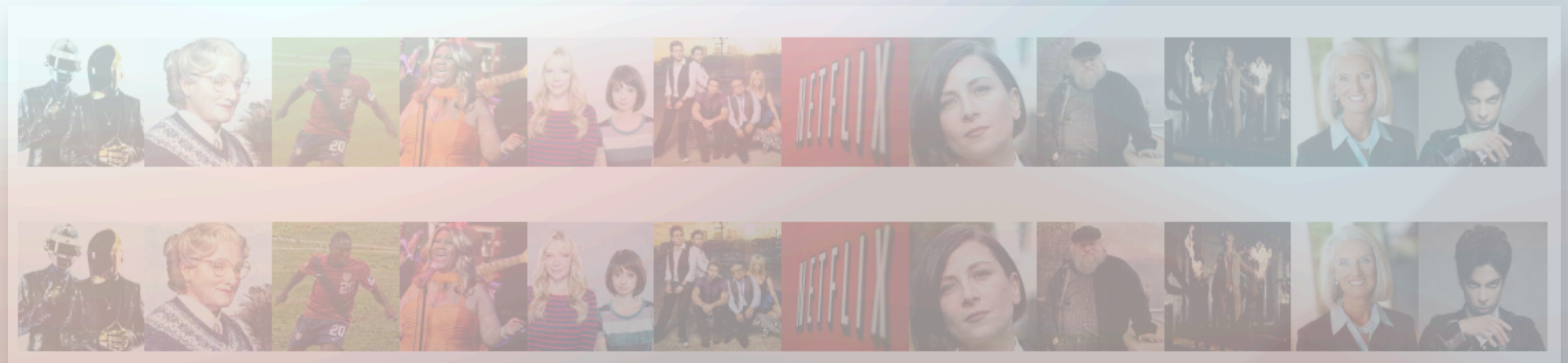
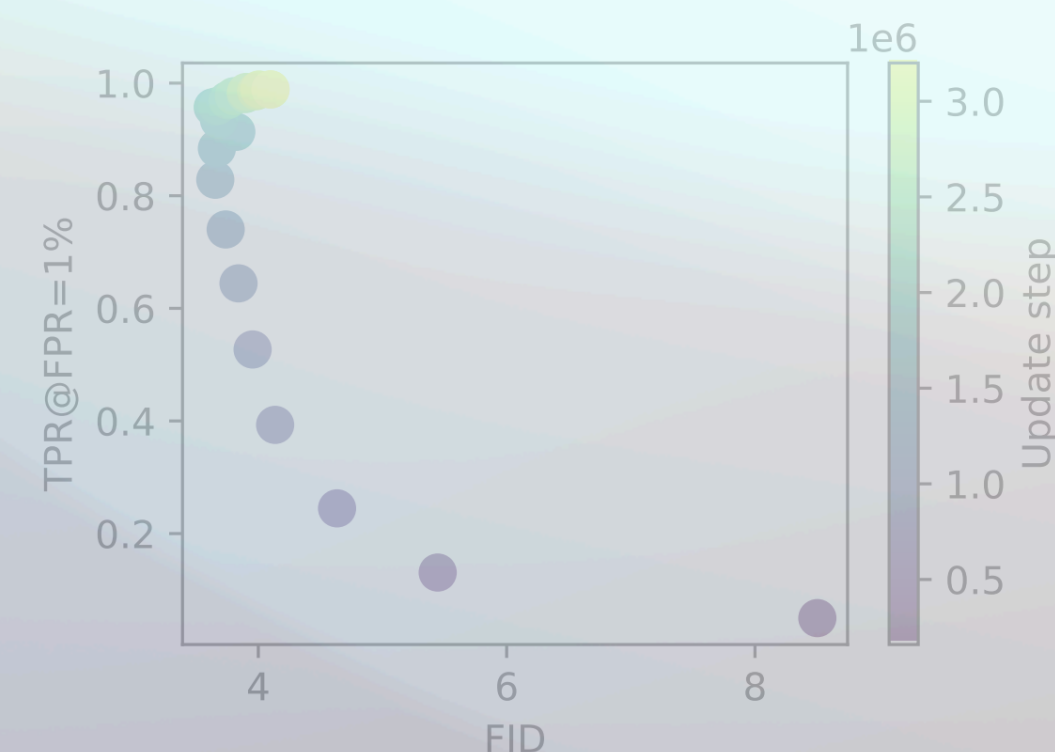
Nicholas Carlini^{*1} Jamie Hayes^{*2} Milad Nasr^{*1}
Matthew Jagielski⁺¹ Vikash Sehwal⁺⁴ Florian Tramèr⁺³
Borja Balle^{†2} Daphne Ippolito^{†1} Eric Wallace^{†5}
¹Google ²DeepMind ³ETHZ ⁴Princeton ⁵UC Berkeley
^{*}Equal contribution [†]Equal contribution

Training Set

Generated Image



“Unfortunately, we applied DP-SGD to our diffusion model codebase and found that it caused the **training on CIFAR-10 to consistently diverge**, even at high values for ϵ (the privacy budget, around 50). In fact, **even applying a non-trivial gradient clipping or noising on their own (both are required in DP-SGD) caused the training to fail**. We leave a further investigation of these failures to future work, and we believe that new advances in DP-SGD and privacy-preserving training techniques may be required to train diffusion models in privacy-sensitive settings.”

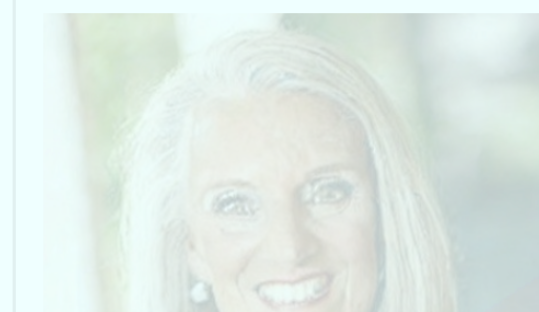


Extracting Training Data from Diffusion Models

Nicholas Carlini^{*1} Jamie Hayes^{*2} Milad Nasr^{*1}
Matthew Jagielski⁺¹ Vikash Sehwal⁺⁴ Florian Tramèr⁺³
Borja Balle^{†2} Daphne Ippolito^{†1} Eric Wallace^{†5}
¹Google ²DeepMind ³ETHZ ⁴Princeton ⁵UC Berkeley
^{*}Equal contribution [†]Equal contribution

Training Set

Generated Image



“Unfortunately, we applied DP-SGD to our diffusion model codebase and found that it caused the **training on CIFAR-10 to consistently diverge**, even at high values for ϵ (the privacy budget, around 50). In fact, **even applying a non-trivial gradient clipping or noising on their own (both are required in DP-SGD) caused the training to fail**. We leave a further investigation of these failures to future work, and we believe that new advances in DP-SGD and privacy-preserving training techniques may be required to train diffusion models in privacy-sensitive settings.”



But has there been no progress on DP training since the last 2 years?
Will come back to it towards the end.....



Multiple
hospitals

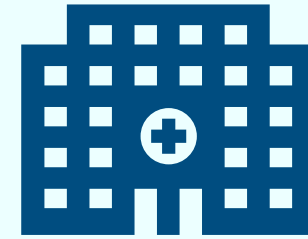
Task Heterogeneity
- Different losses
- Different targets

Many possible downstream tasks:

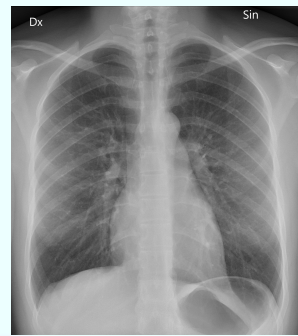
- Disease classification
- Localizing infected regions
- Predicting severity/progression
- Policies for decision making
- Or, learning some general purpose embeddings



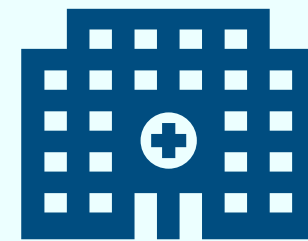
$$S_1 \sim \mathcal{D}^{\otimes n_1}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_1} [f_1(w; z)]$$



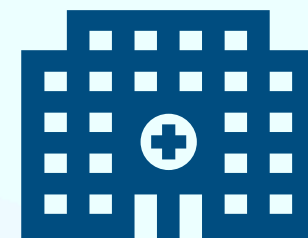
$$S_2 \sim \mathcal{D}^{\otimes n_2}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_2} [f_2(w; z)]$$

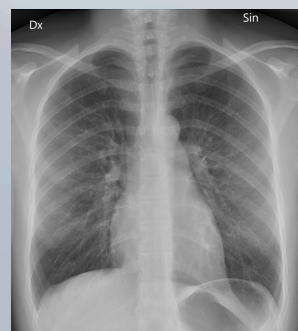


$$S_3 \sim \mathcal{D}^{\otimes n_3}$$

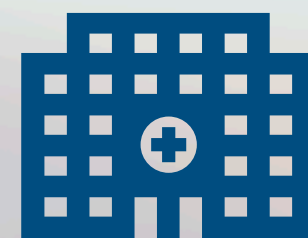


$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_3} [f_3(w; z)]$$

⋮



$$S_M \sim \mathcal{D}^{\otimes n_M}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_M} [f_M(w; z)]$$

Chest X-ray images

Data Heterogeneity
- Distributions
- Sample sizes

Multiple hospitals

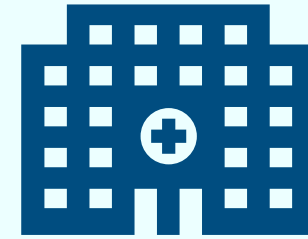
Task Heterogeneity
- Different losses
- Different targets

Many possible downstream tasks:

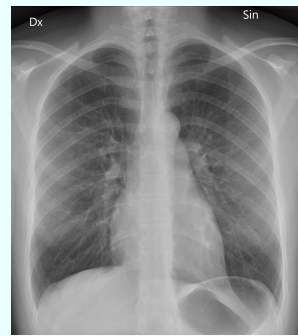
- Disease classification
- Localizing infected regions
- Predicting severity/progression
- Policies for decision making
- Or, learning some general purpose embeddings



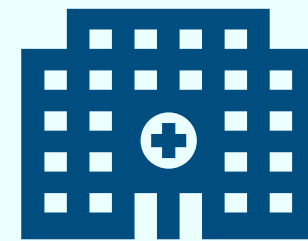
$$S_1 \sim \mathcal{D}_1^{\otimes n_1}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_1} [f_1(w; z)]$$



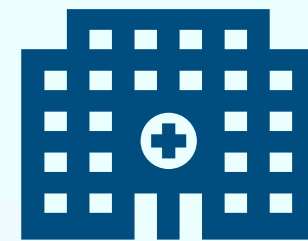
$$S_2 \sim \mathcal{D}_2^{\otimes n_2}$$



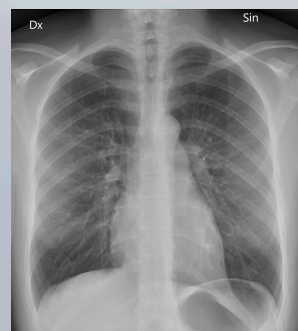
$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_2} [f_2(w; z)]$$



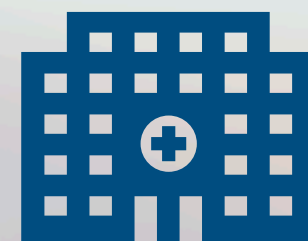
$$S_3 \sim \mathcal{D}_3^{\otimes n_3}$$



$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_3} [f_3(w; z)]$$

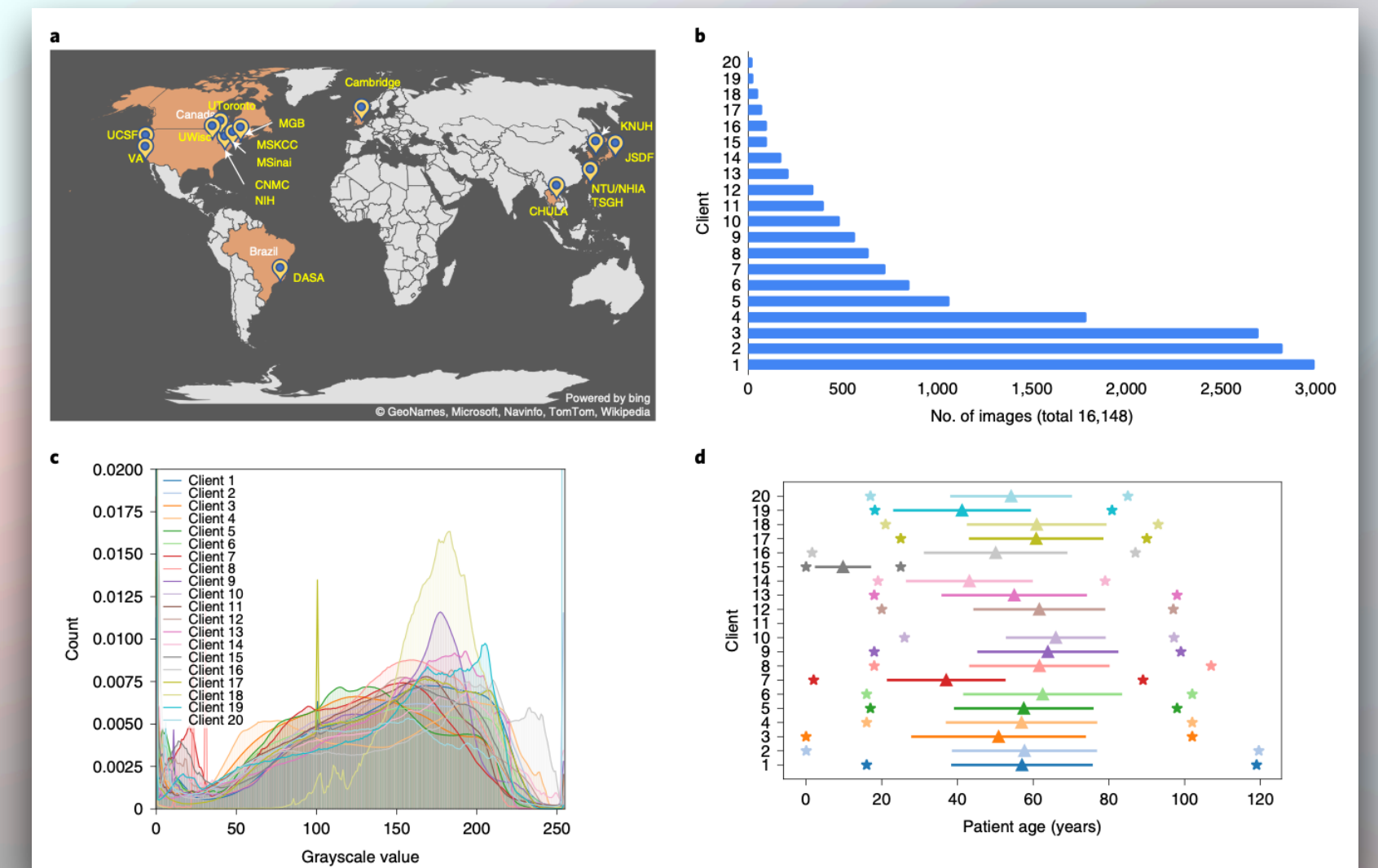


$$S_M \sim \mathcal{D}_M^{\otimes n_M}$$



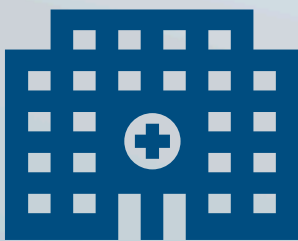
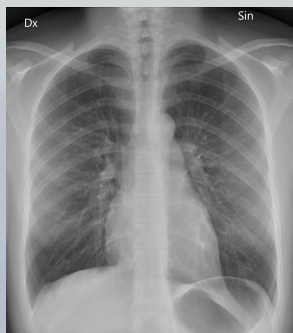
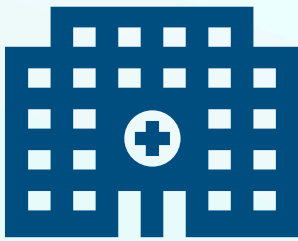
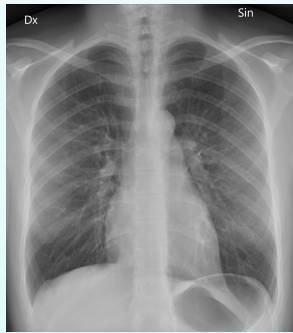
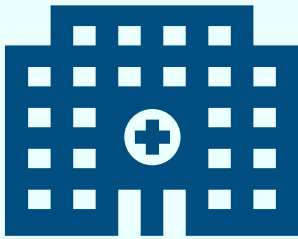
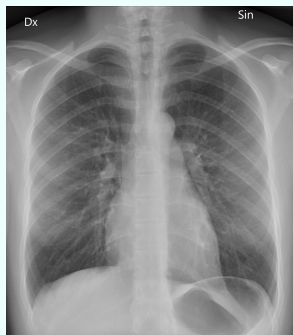
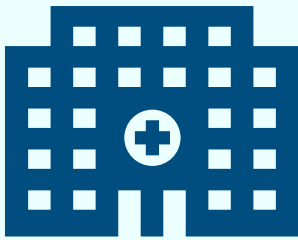
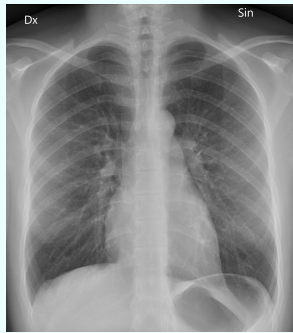
$$\min_w \mathbb{E}_{z \sim \mathcal{T}'_M} [f_M(w; z)]$$

Chest X-ray images



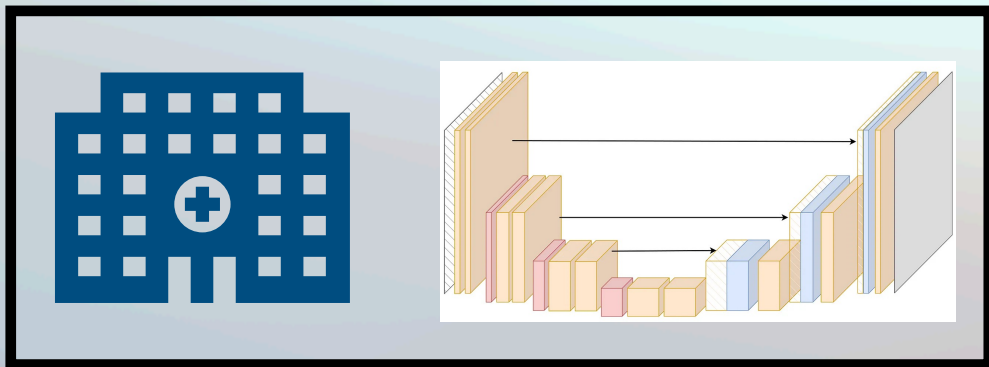
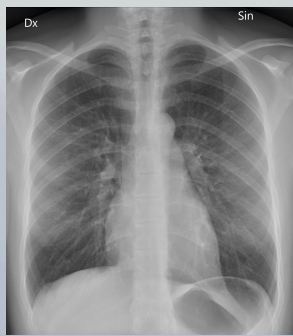
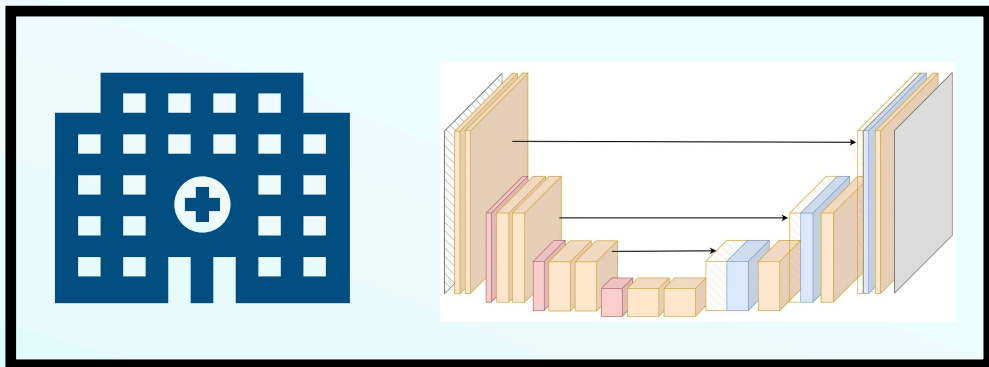
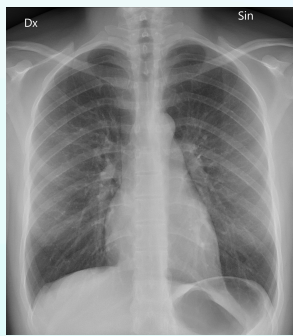
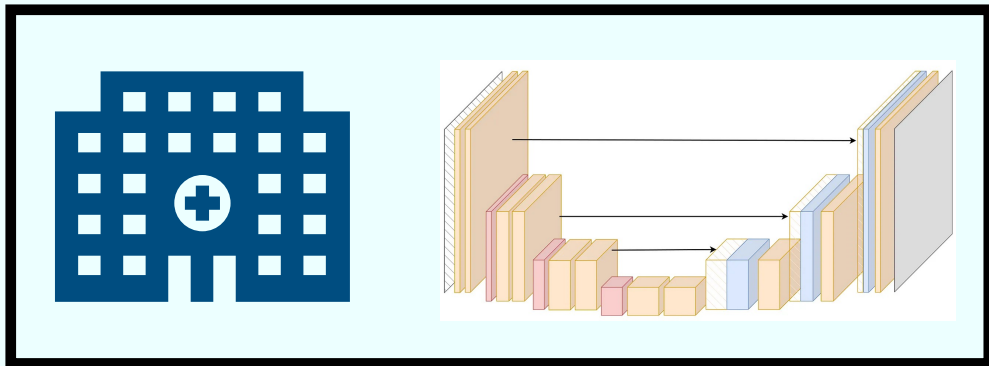
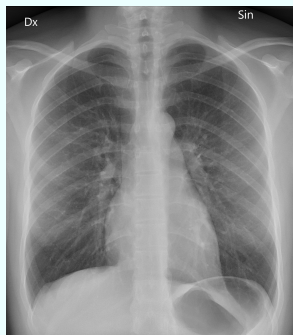
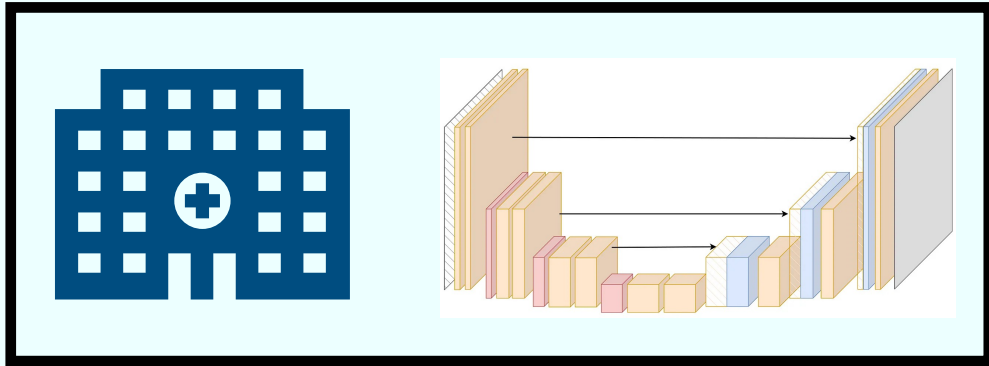
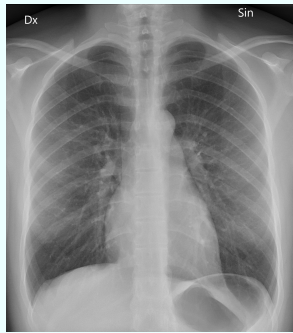
-Nvidia CLARA

An even better solution now would also deal with the data heterogeneity by generating M **different generative models** one for each hospital, i.e. do **personalization**



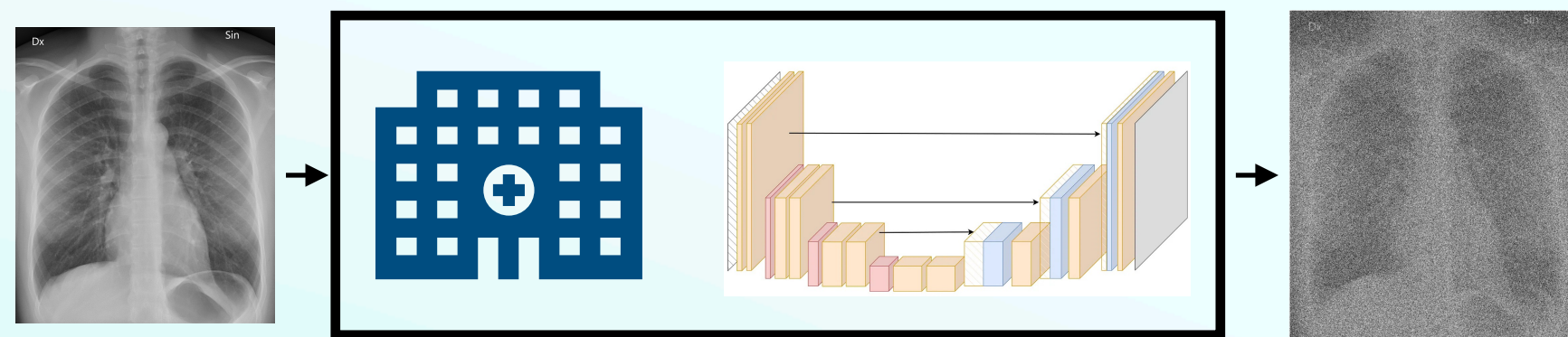
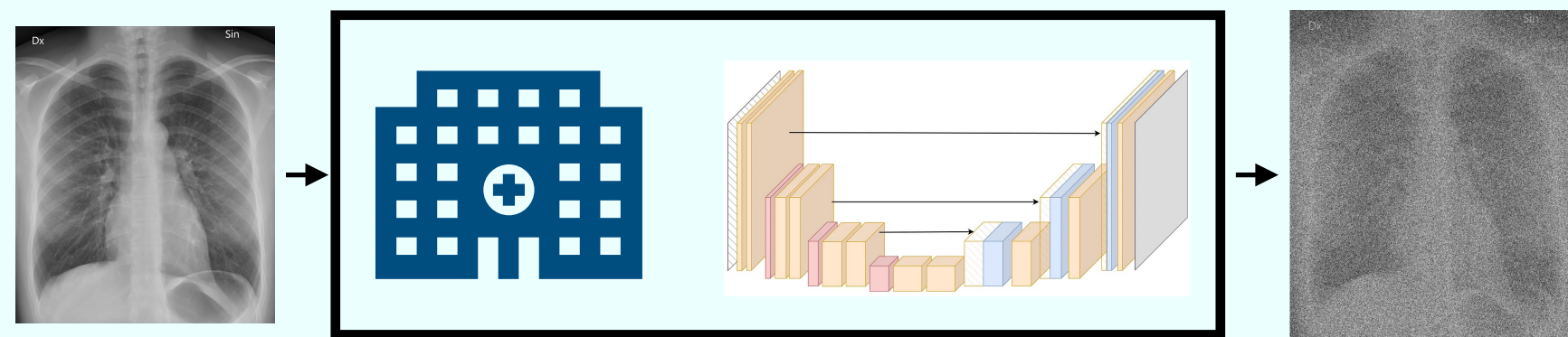
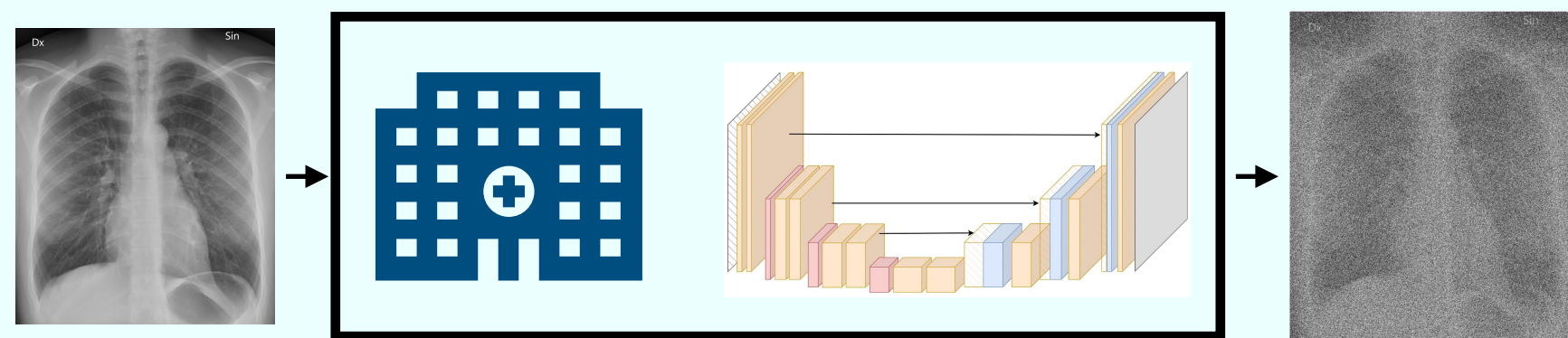
Clean chest
X-ray images

Each hospital with its personal
(and smaller) diffusion model

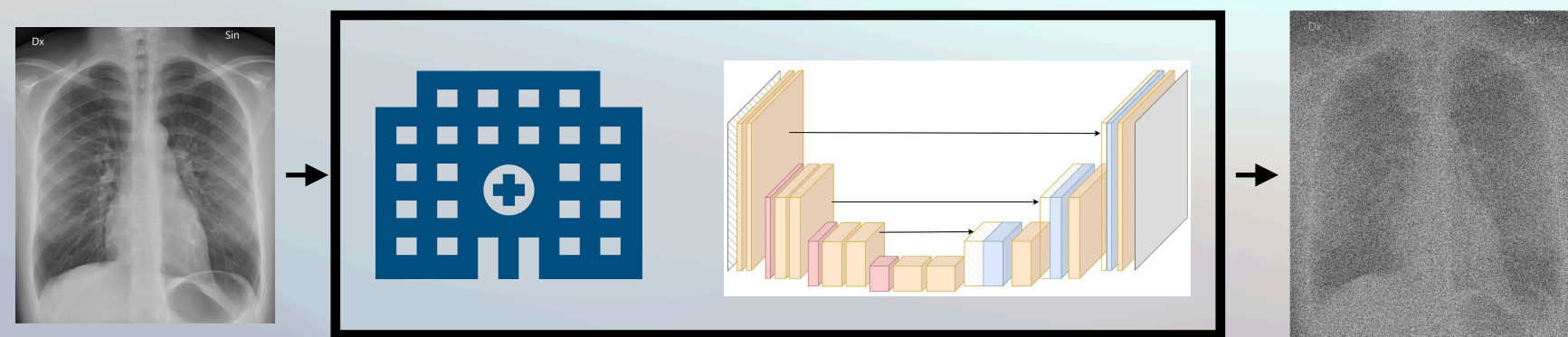


Clean chest
X-ray images

Each hospital with its personal
(and smaller) diffusion model



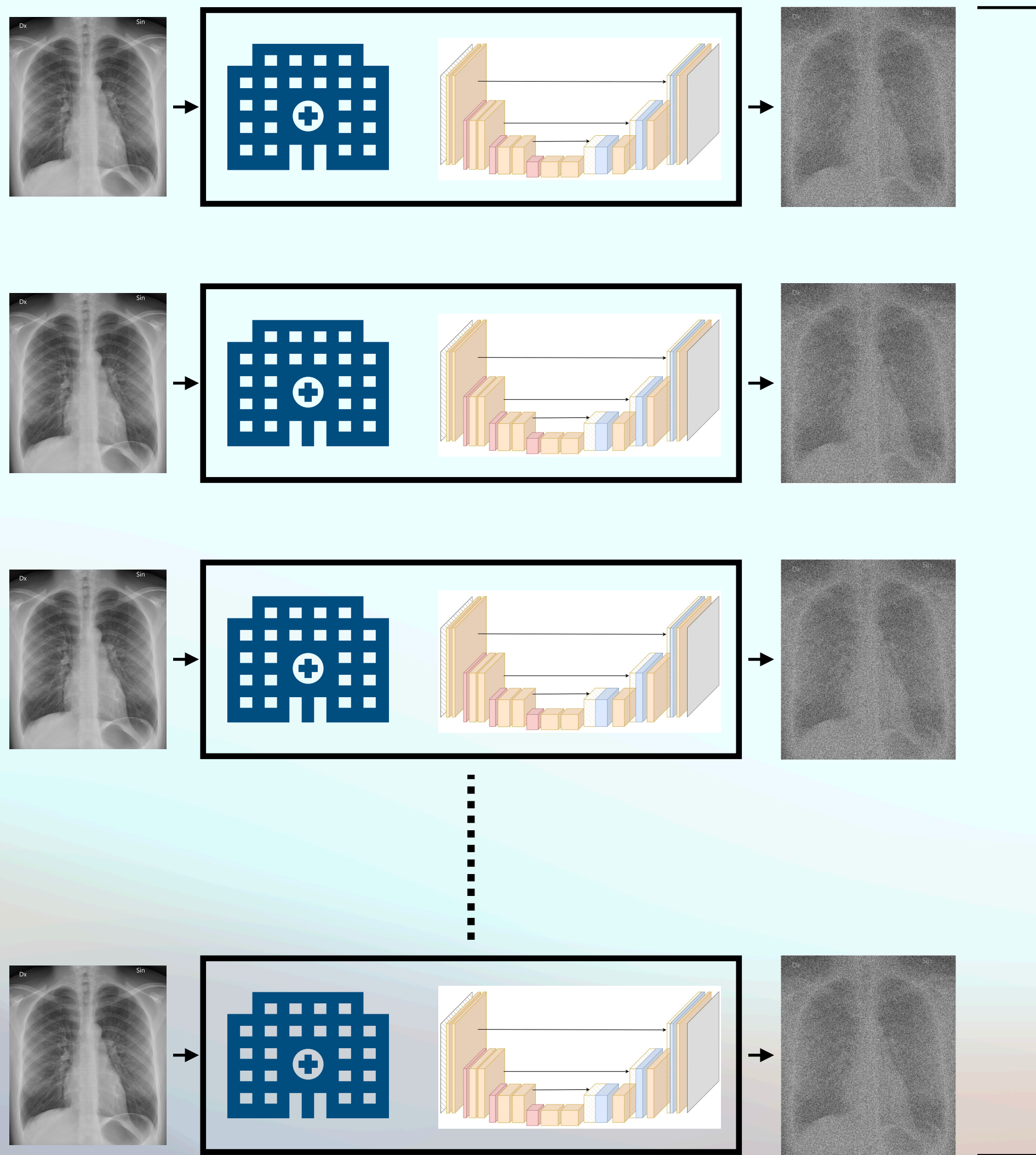
⋮



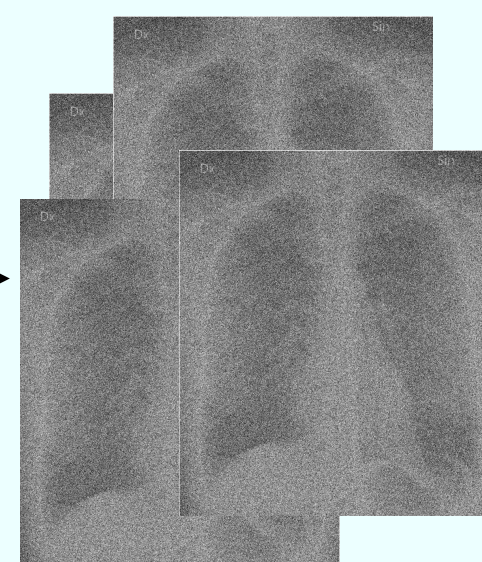
Clean chest
X-ray images

Noisy chest
X-ray images

Each hospital with its personal
(and smaller) diffusion model



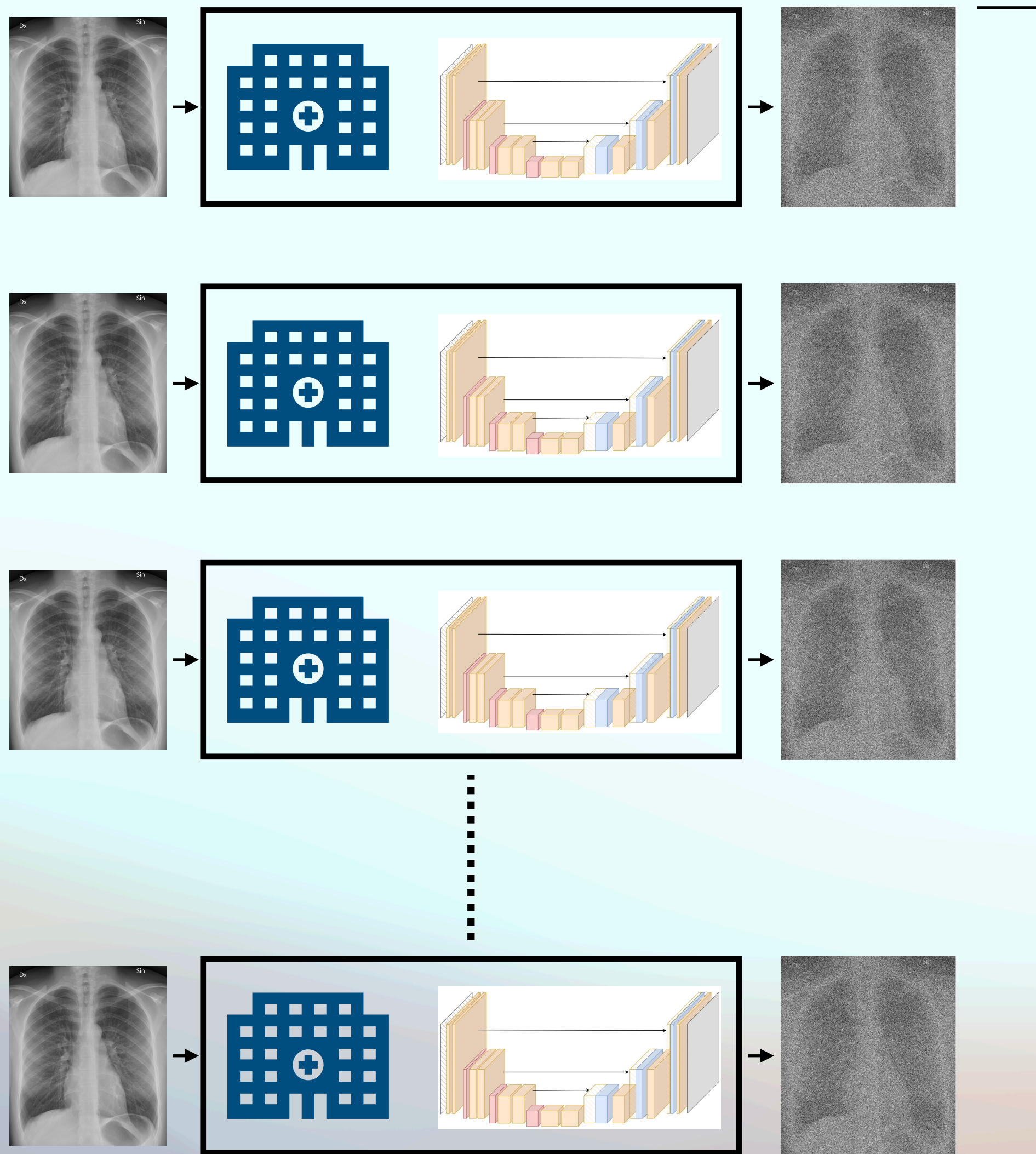
Aggregated
noisy dataset



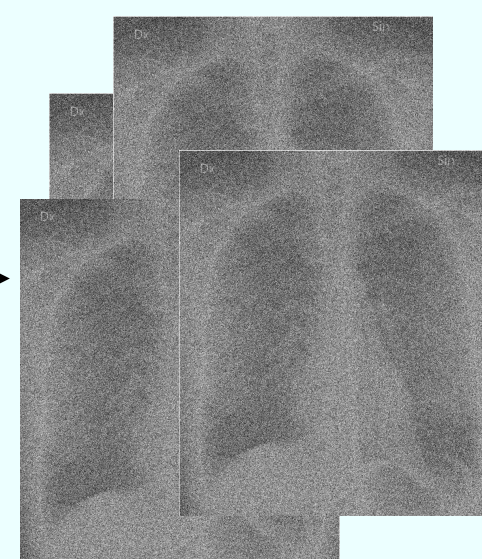
Clean chest
X-ray images

Noisy chest
X-ray images

Each hospital with its personal
(and smaller) diffusion model

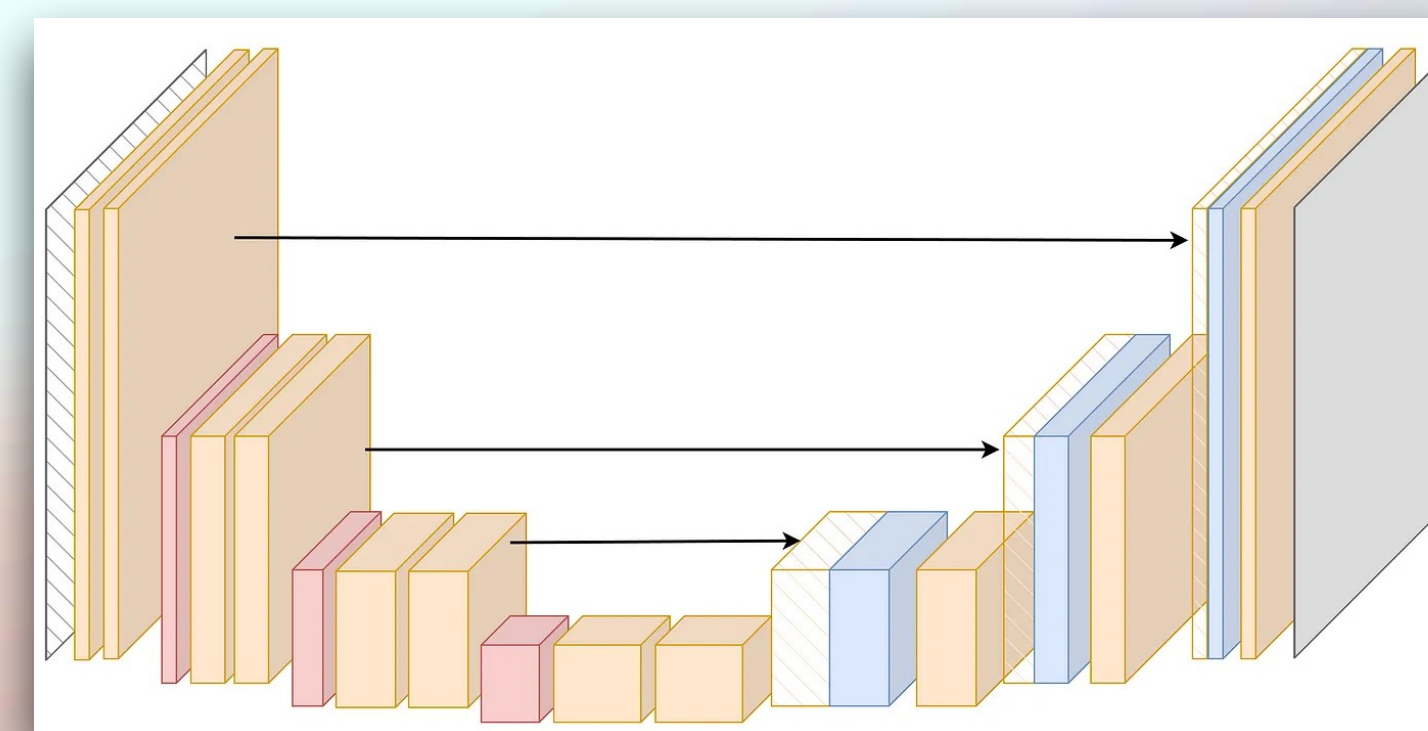
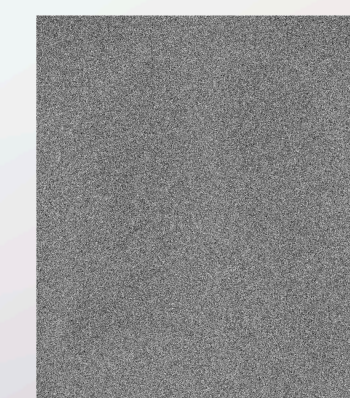


Aggregated
noisy dataset

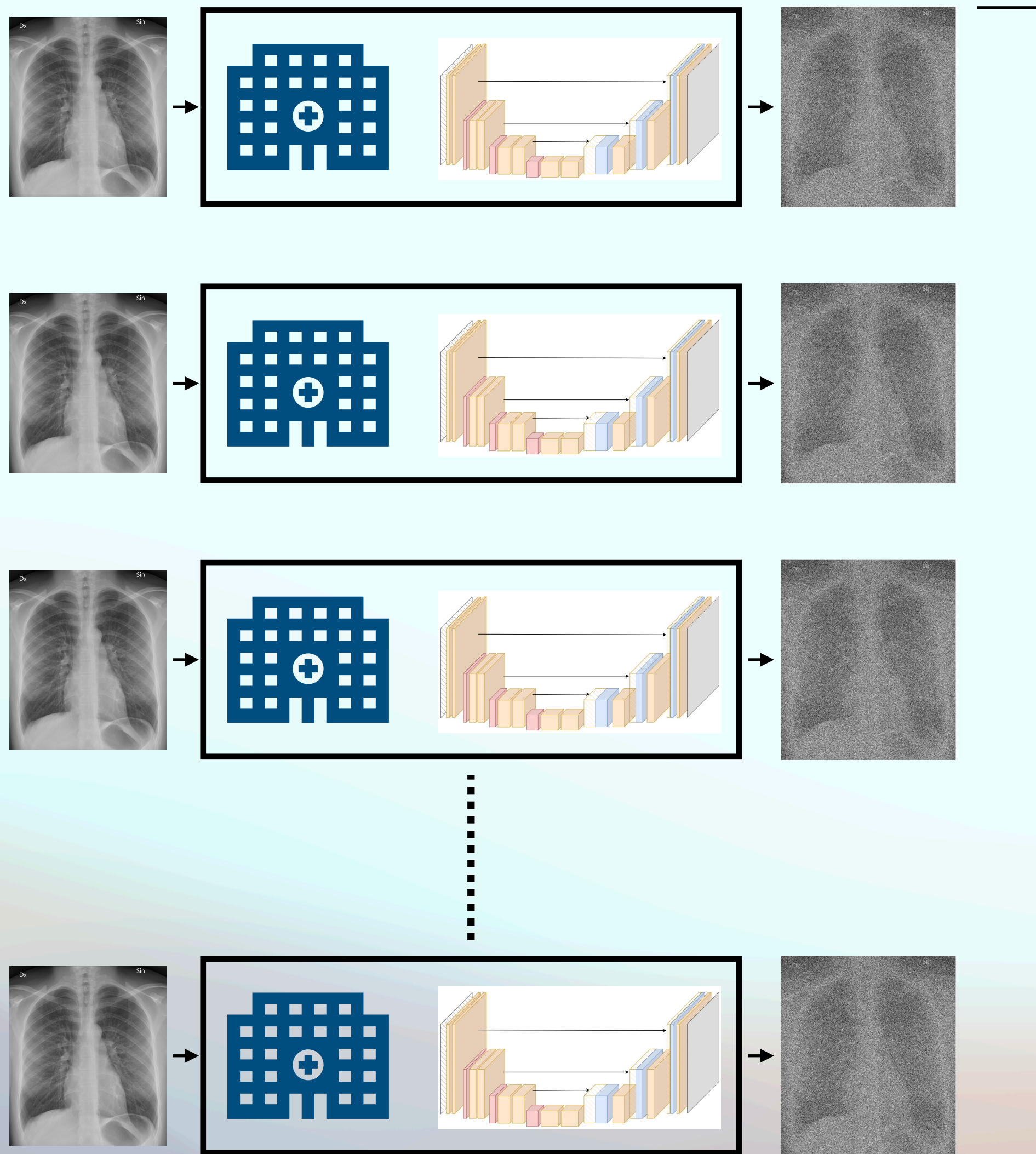


Forward
diffusion

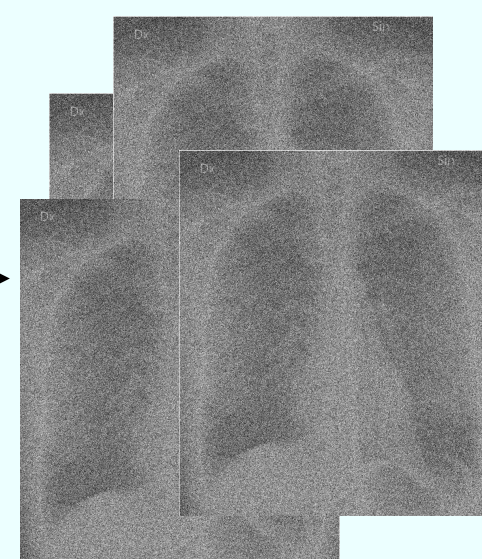
Pure noise



Each hospital with its personal
(and smaller) diffusion model

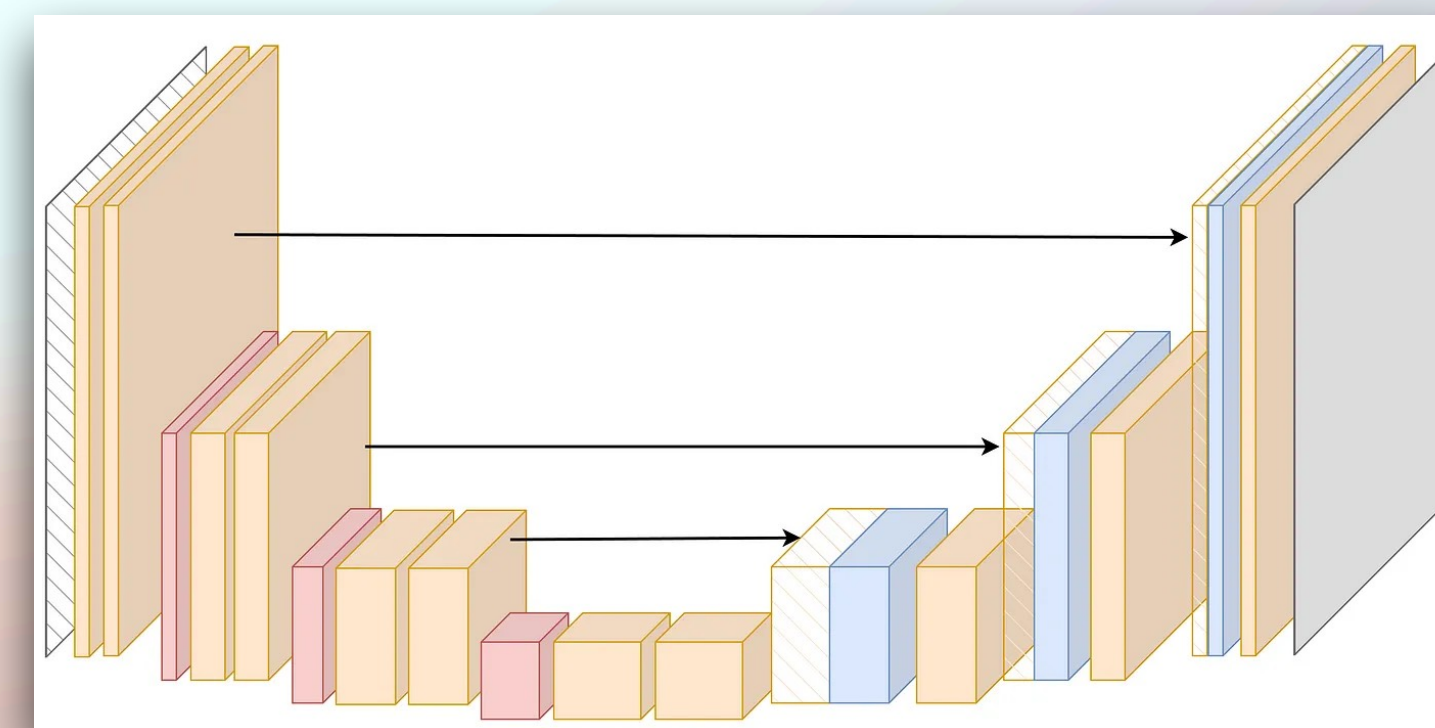
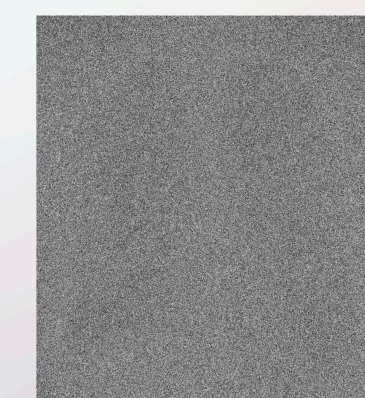


Aggregated
noisy dataset



Forward
diffusion

Pure noise

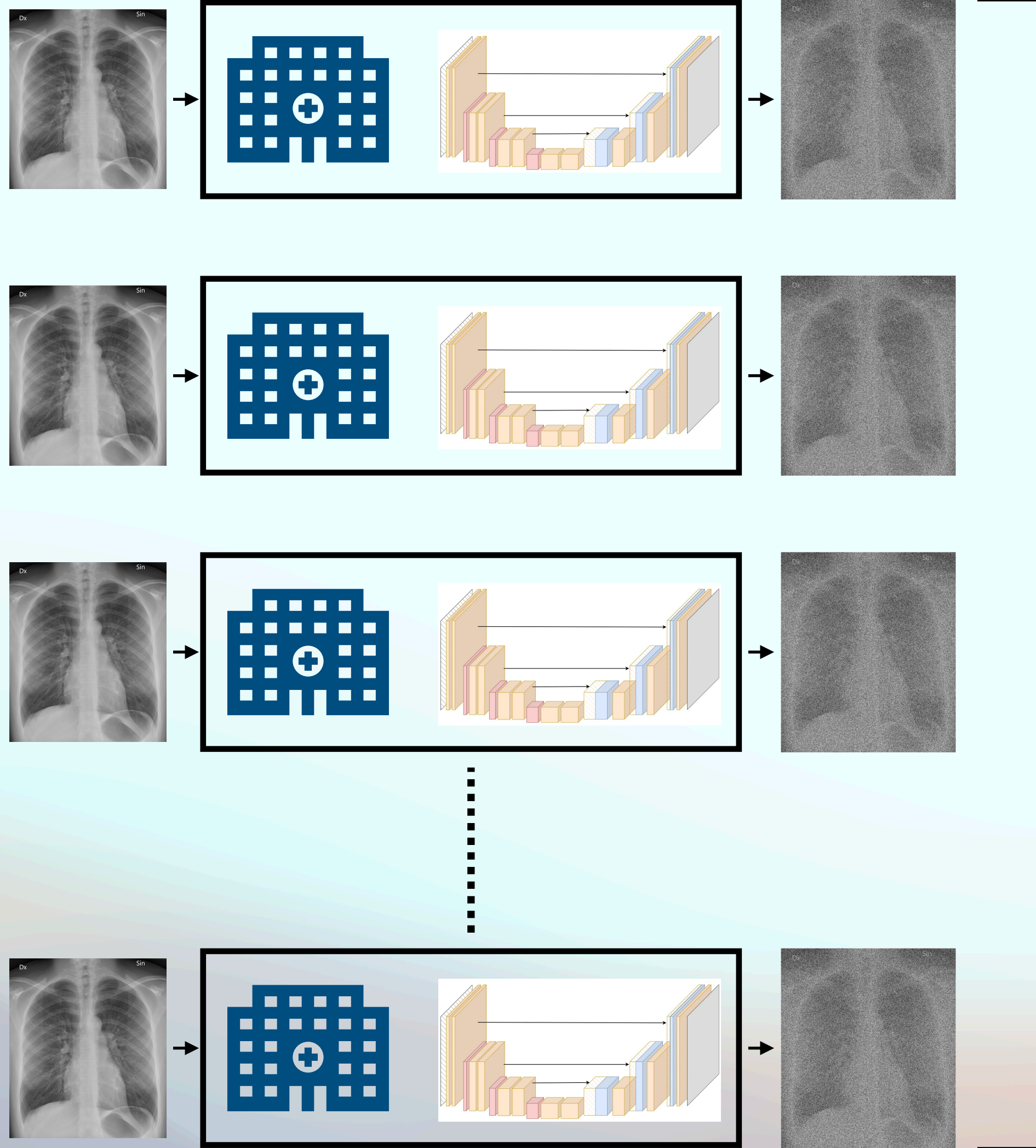


Large global (and potentially public)
diffusion model that only ever sees noisy images

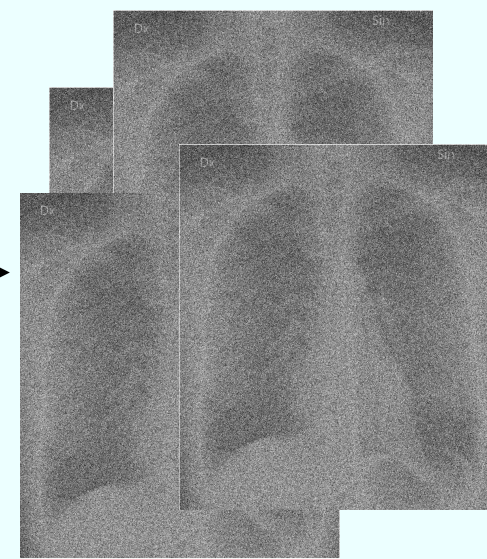
Clean chest
X-ray images

Noisy chest
X-ray images

Each hospital with its personal
(and smaller) diffusion model

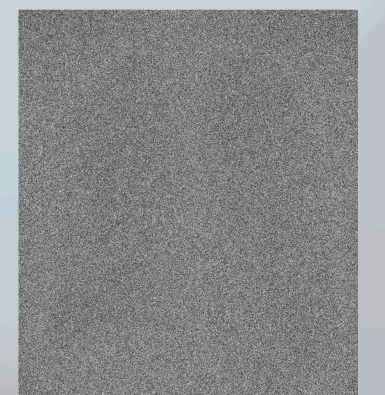
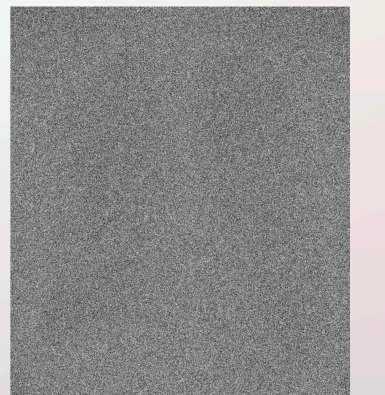


Aggregated
noisy dataset

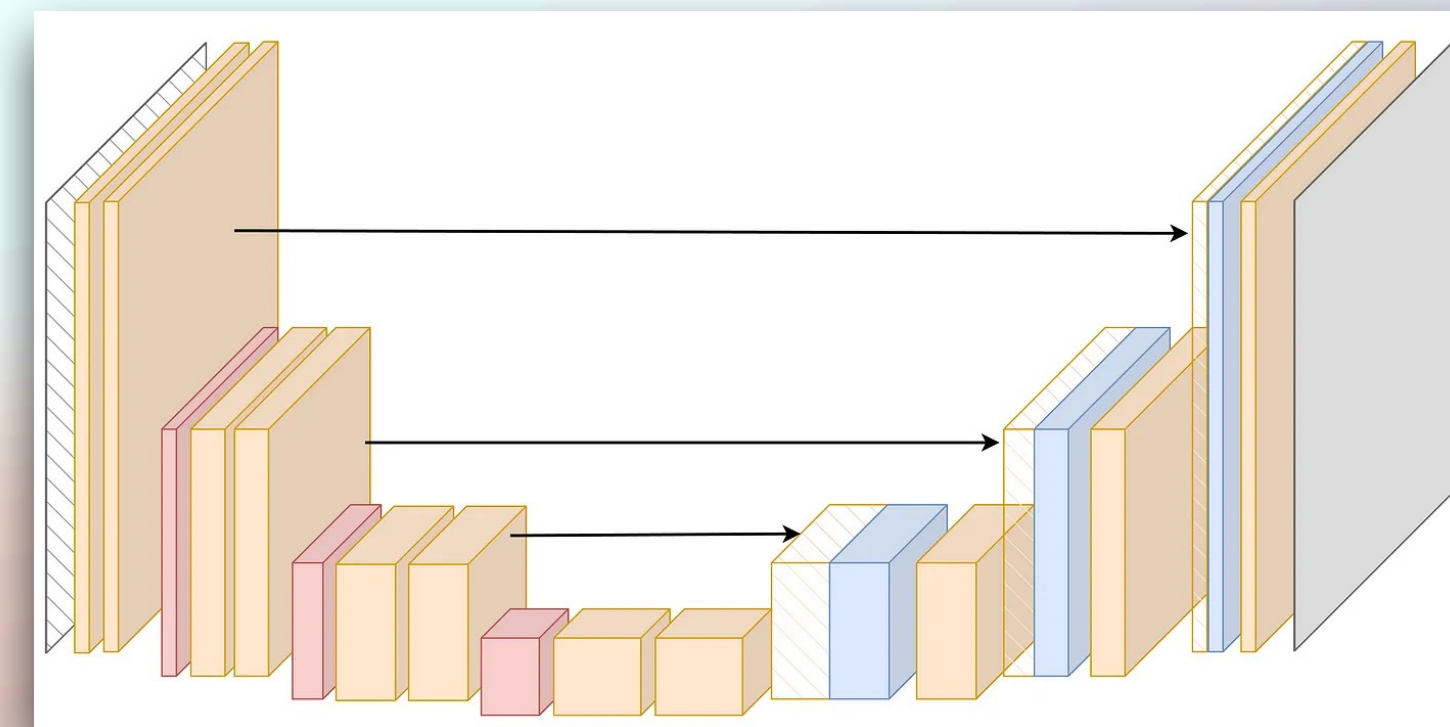


Forward
diffusion

Pure noise



Random seed

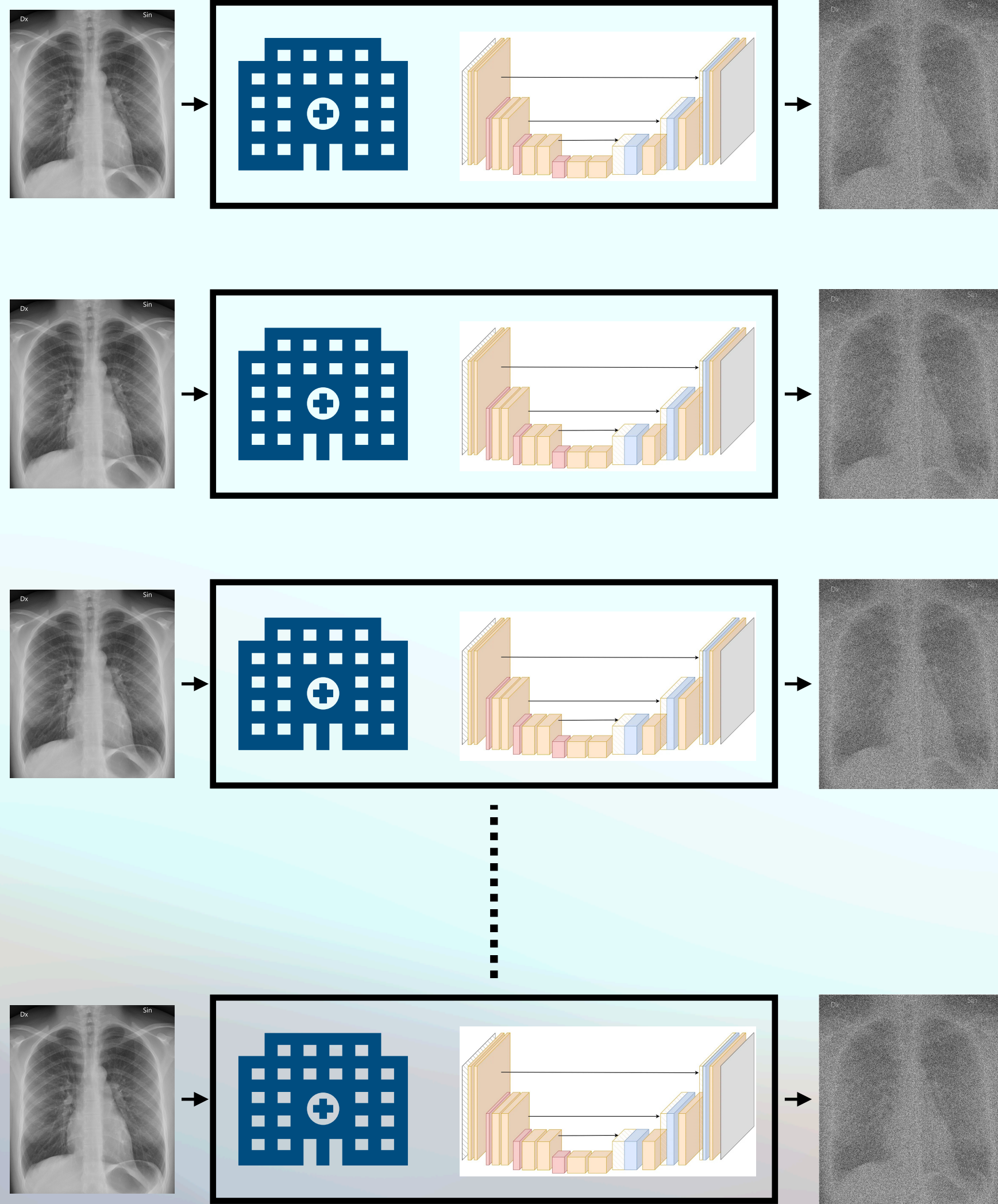


Large global (and potentially public)
diffusion model that only ever sees noisy images

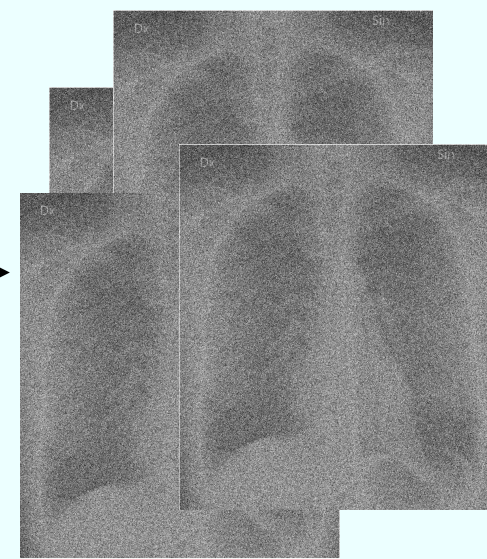
Clean chest
X-ray images

Noisy chest
X-ray images

Each hospital with its personal
(and smaller) diffusion model

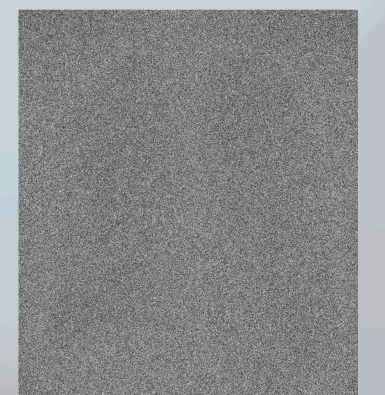
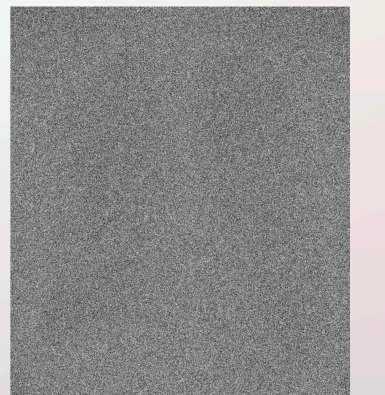


Aggregated
noisy dataset

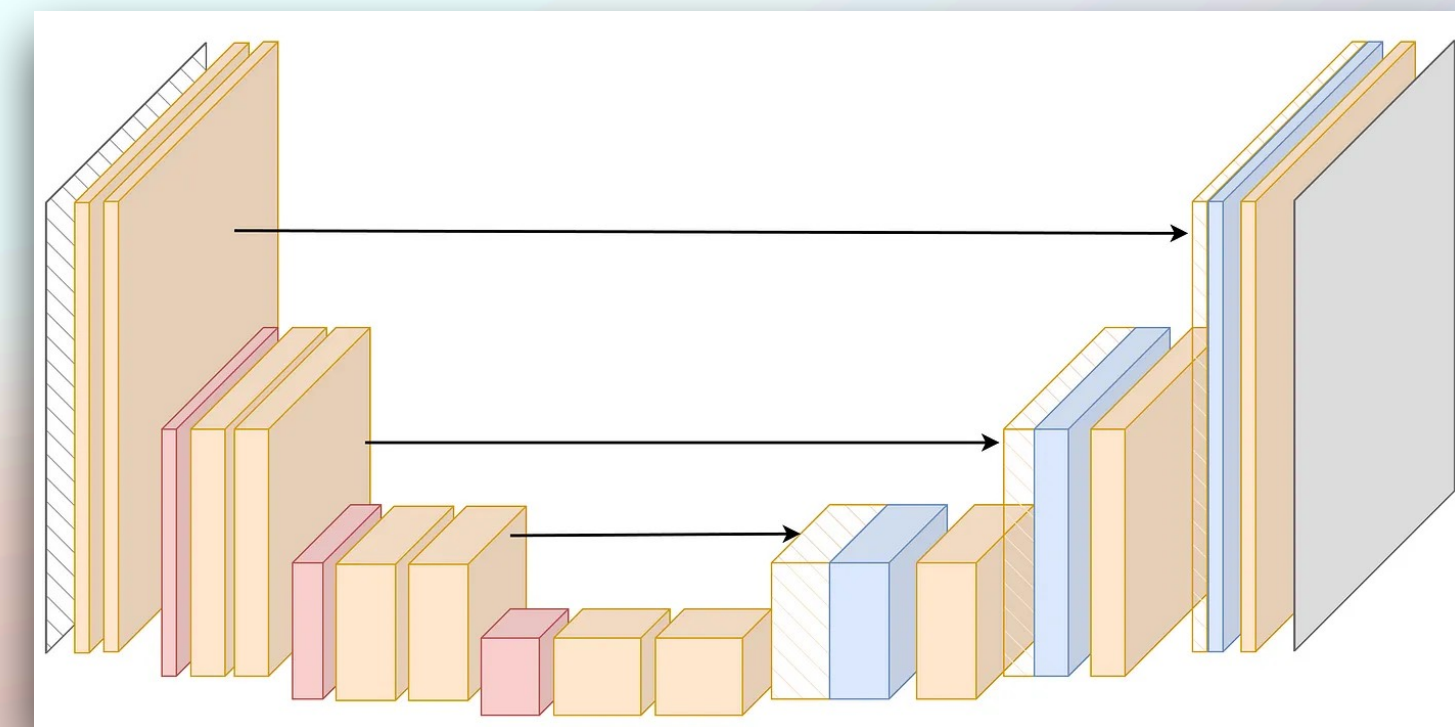


Forward
diffusion

Pure noise



Random seed

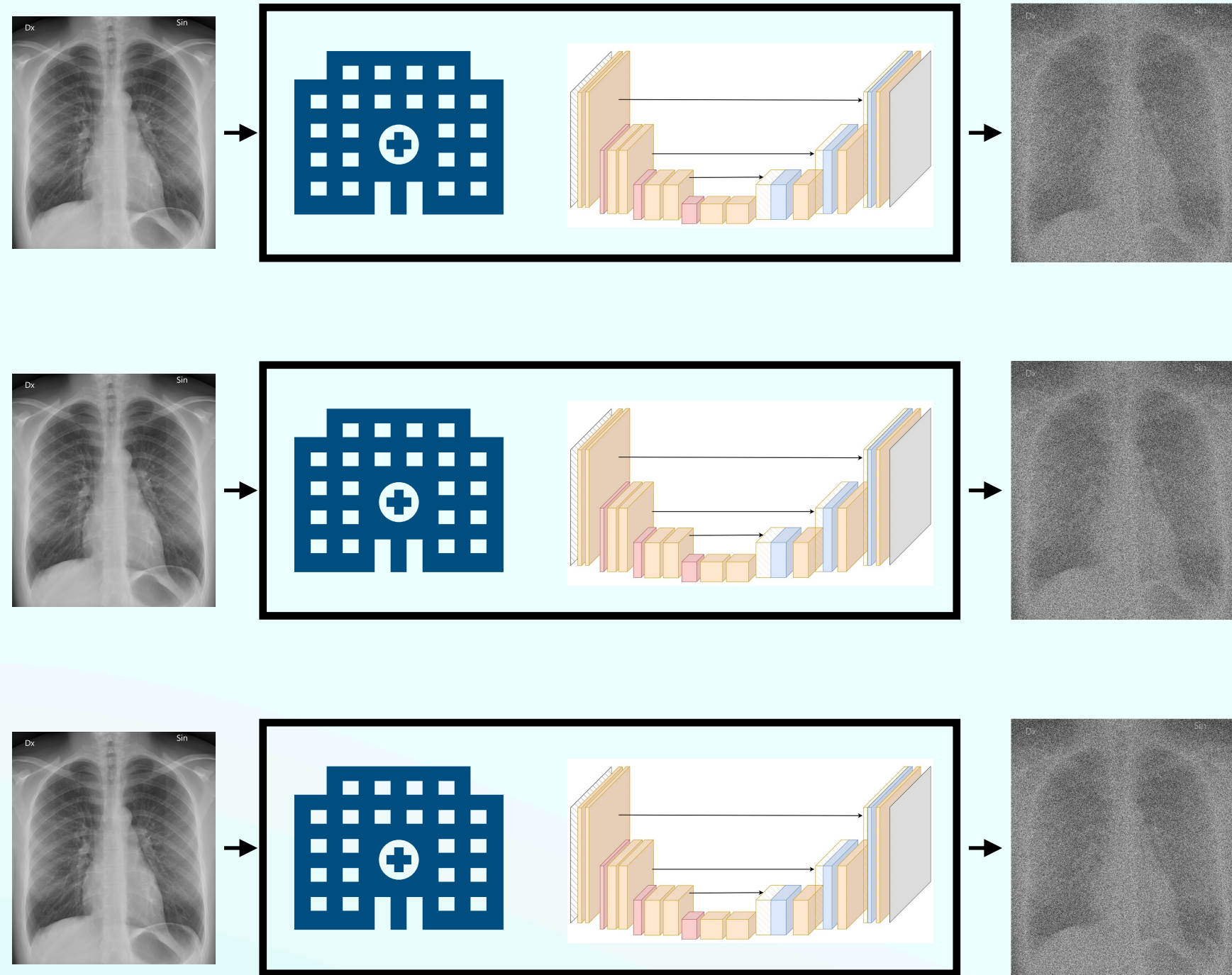


Large global (and potentially public)
diffusion model that only ever sees noisy images

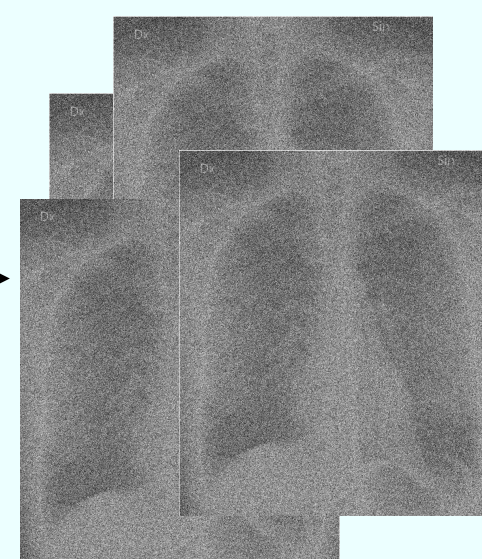
Clean chest
X-ray images

Noisy chest
X-ray images

Each hospital with its personal
(and smaller) diffusion model

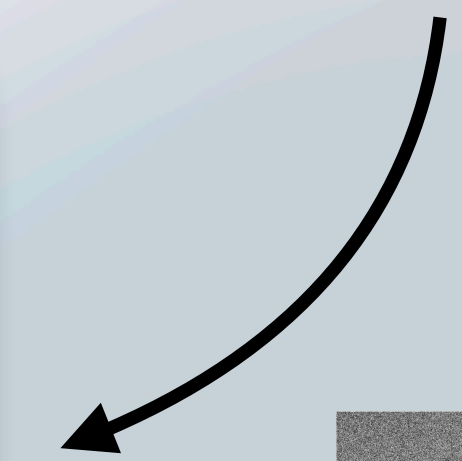
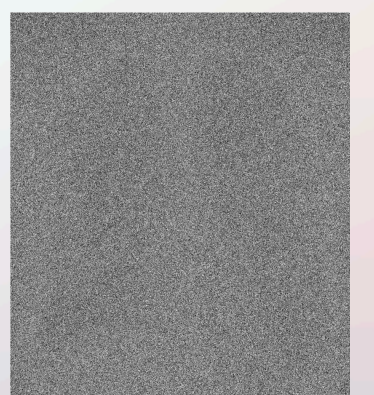


Aggregated
noisy dataset

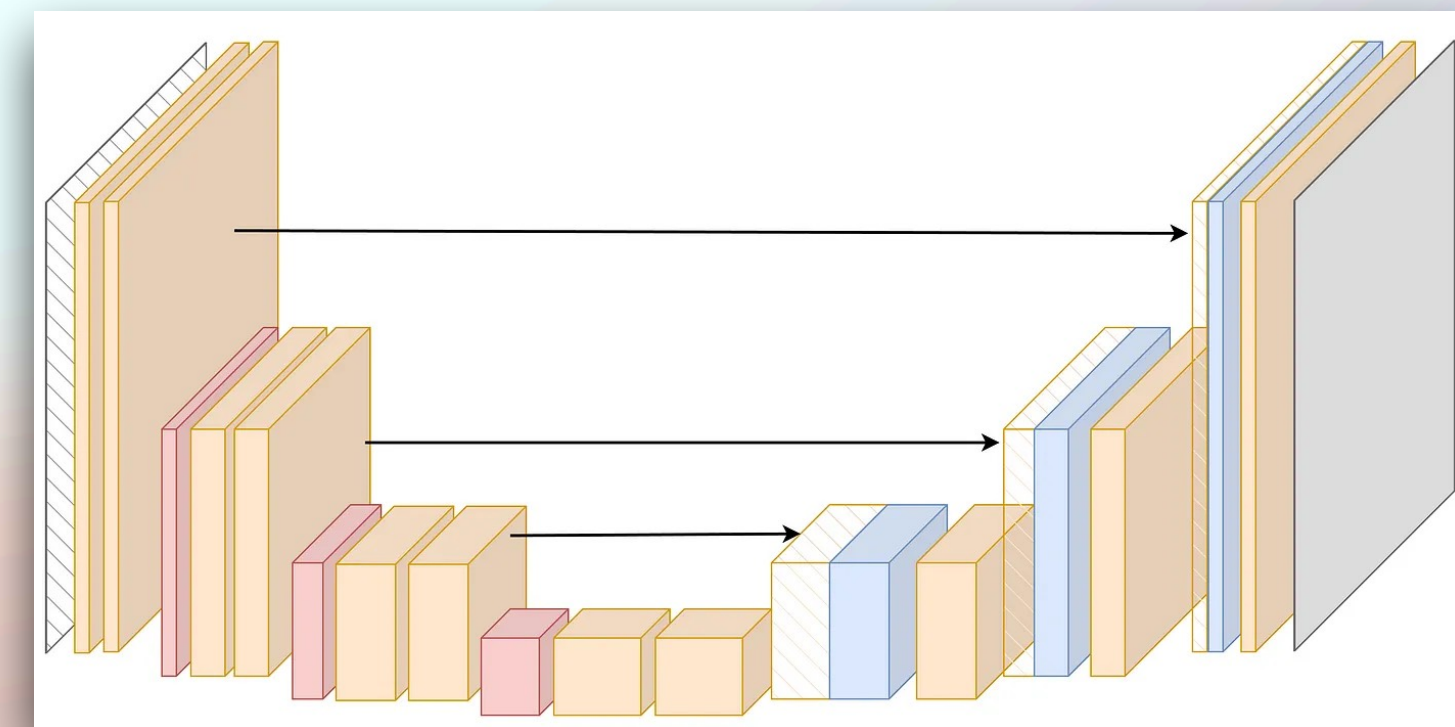
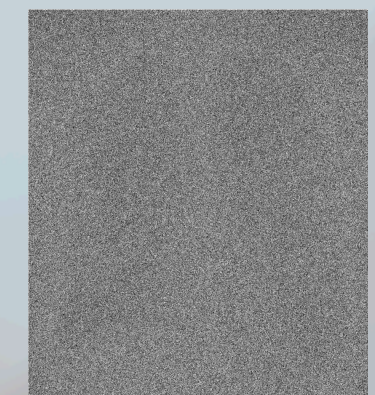


Forward
diffusion

Pure noise

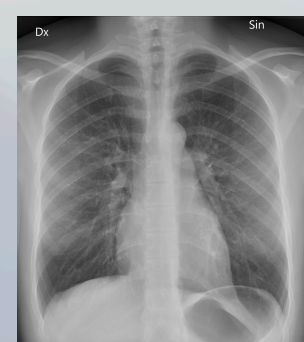


Random seed

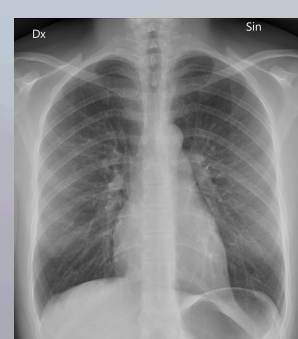


Large global (and potentially public)
diffusion model that only ever sees noisy images

Clean chest
X-ray images



Noisy chest
X-ray images



**This personalized approach satisfies Local DP
for each image in each hospital**

Canonical Training of Diffusion Models (DDPM)

Canonical Training of Diffusion Models (DDPM)

- Sample a clean image from your dataset: $x_0 \sim \mathcal{D}$

Canonical Training of Diffusion Models (DDPM)

- Sample a clean image from your dataset: $x_0 \sim \mathcal{D}$
- Progressively noise it for $t \in [T]$: $x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} z_t$, where $z_t \sim \mathcal{N}(0, I)$

Canonical Training of Diffusion Models (DDPM)

- Sample a clean image from your dataset: $x_0 \sim \mathcal{D}$
- Progressively noise it for $t \in [T]$: $x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} z_t$, where $z_t \sim \mathcal{N}(0, I)$
- For $\bar{\alpha}_t := \prod_{s=1}^t (1 - \beta_s)$ this implies, $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} z_t$

Canonical Training of Diffusion Models (DDPM)

- Sample a clean image from your dataset: $x_0 \sim \mathcal{D}$
- Progressively noise it for $t \in [T]$: $x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} z_t$, where $z_t \sim \mathcal{N}(0, I)$
- For $\bar{\alpha}_t := \prod_{s=1}^t (1 - \beta_s)$ this implies, $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} z_t$
- For a sufficiently large T and an appropriate noise schedule x_T approaches $\mathcal{N}(0, I)$

Canonical Training of Diffusion Models (DDPM)

- Sample a clean image from your dataset: $x_0 \sim \mathcal{D}$
- Progressively noise it for $t \in [T]$: $x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} z_t$, where $z_t \sim \mathcal{N}(0, I)$
- For $\bar{\alpha}_t := \prod_{s=1}^t (1 - \beta_s)$ this implies, $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} z_t$
- For a sufficiently large T and an appropriate noise schedule x_T approaches $\mathcal{N}(0, I)$
- **Key fact:** under certain assumptions on $\mathcal{D}$, $p(x_{t-1} \mid x_t) \approx \mathcal{N}(\mu(x_t, t), \Sigma(x_t, t))!!$

Canonical Training of Diffusion Models (DDPM)

- Sample a clean image from your dataset: $x_0 \sim \mathcal{D}$
- Progressively noise it for $t \in [T]$: $x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} z_t$, where $z_t \sim \mathcal{N}(0, I)$
- For $\bar{\alpha}_t := \prod_{s=1}^t (1 - \beta_s)$ this implies, $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} z_t$
- For a sufficiently large T and an appropriate noise schedule x_T approaches $\mathcal{N}(0, I)$
- **Key fact:** under certain assumptions on $\mathcal{D}$, $p(x_{t-1} \mid x_t) \approx \mathcal{N}(\mu(x_t, t), \Sigma(x_t, t))$!!
- Ho et al. reduce the “reverse” step to estimating $\mu(x_t, t)$ using a NN $\mu_\theta(x_t, t)$

Canonical Training of Diffusion Models (DDPM)

Algorithm 3 Training Procedure for DDPM (T-DDPM)

input Training dataset D , model parameters θ , time step T , noise scheduling parameters $\{\beta_t\}_{t=1}^T$

1: **repeat**

2: Sample x_0 from D

3: Sample $t \sim \text{Uniform}(\{1, \dots, T\})$

4: Sample random noise $z \sim \mathcal{N}(0, I)$

5: Set $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$

6: Take gradient descent step on

$$\nabla_{\theta} \left\| z - z_{\theta}(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}z, t) \right\|_2^2$$

7: **until** converged

output denoiser z_{θ}

Canonical Training of Diffusion Models (DDPM)

Algorithm 3 Training Procedure for DDPM (T-DDPM)

input Training dataset D , model parameters θ , time step T , noise scheduling parameters $\{\beta_t\}_{t=1}^T$

- 1: **repeat**
- 2: Sample x_0 from D
- 3: Sample $t \sim \text{Uniform}(\{1, \dots, T\})$
- 4: Sample random noise $z \sim \mathcal{N}(0, I)$
- 5: Set $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$
- 6: Take gradient descent step on

$$\nabla_{\theta} \|z - z_{\theta}(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}z, t)\|_2^2$$
- 7: **until** converged

output denoiser z_{θ}

Sample across the trajectory

Canonical Training of Diffusion Models (DDPM)

Algorithm 3 Training Procedure for DDPM (T-DDPM)

input Training dataset D , model parameters θ , time step T , noise scheduling parameters $\{\beta_t\}_{t=1}^T$

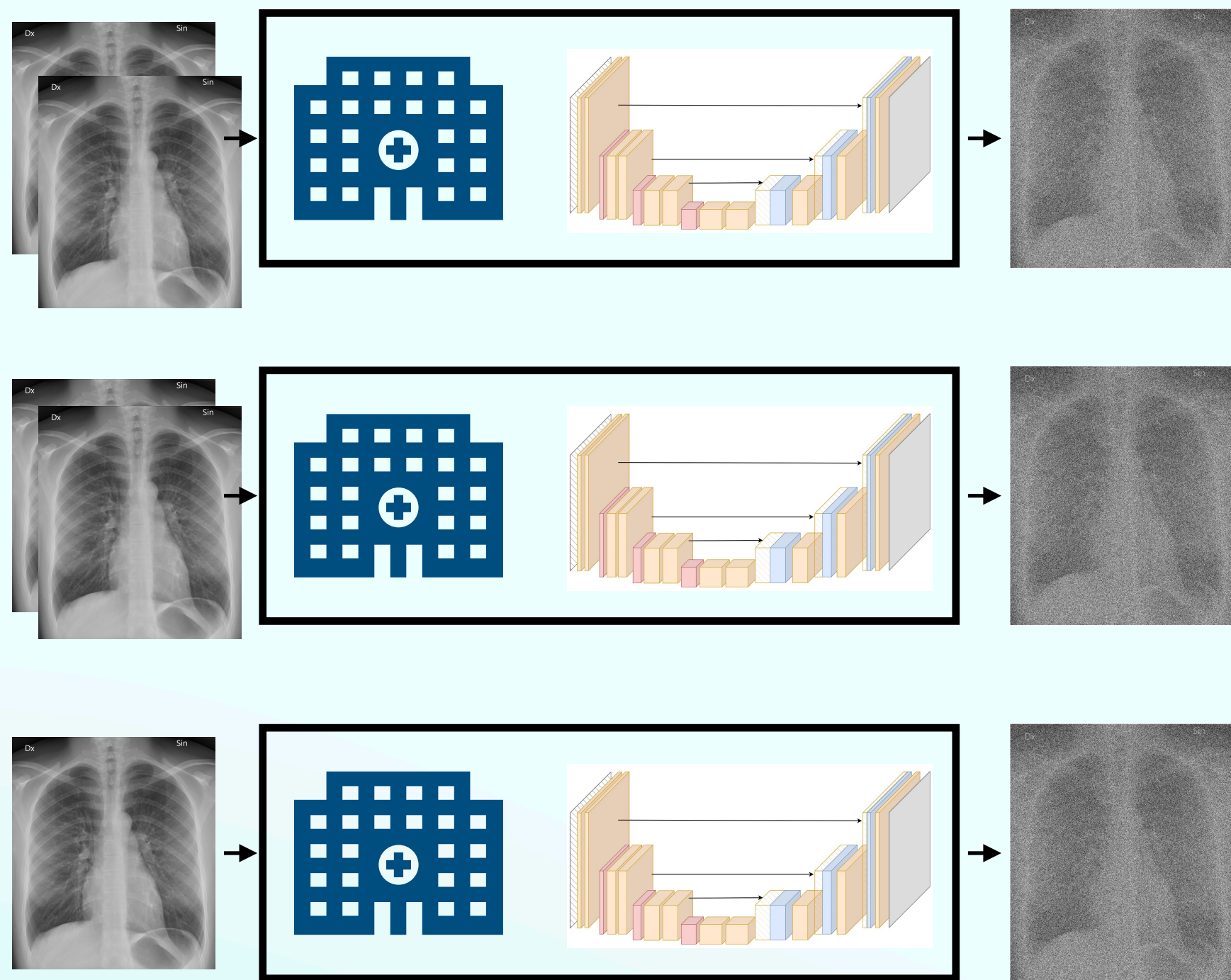
- 1: **repeat**
- 2: Sample x_0 from D
- 3: Sample $t \sim \text{Uniform}(\{1, \dots, T\})$
- 4: Sample random noise $z \sim \mathcal{N}(0, I)$
- 5: Set $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$
- 6: Take gradient descent step on
 $\nabla_{\theta} \|z - z_{\theta}(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}z, t)\|_2^2$
- 7: **until** converged

output denoiser z_{θ}

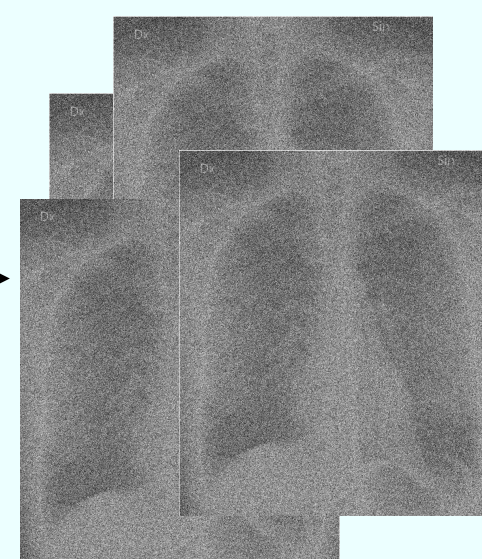
Sample across the trajectory

Predicting noised images or the noise itself are “equivalent”

Each hospital with its personal
(and smaller) diffusion model

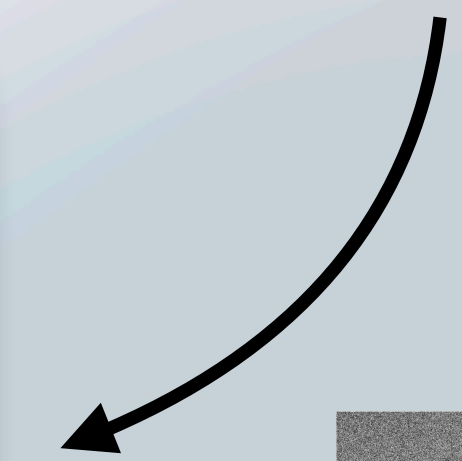
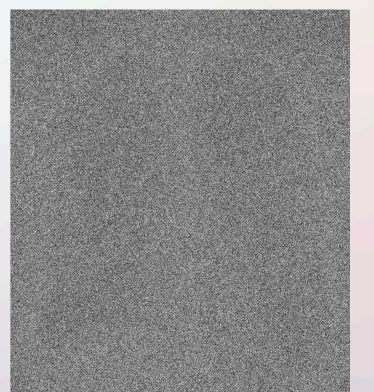


Aggregated
noisy dataset

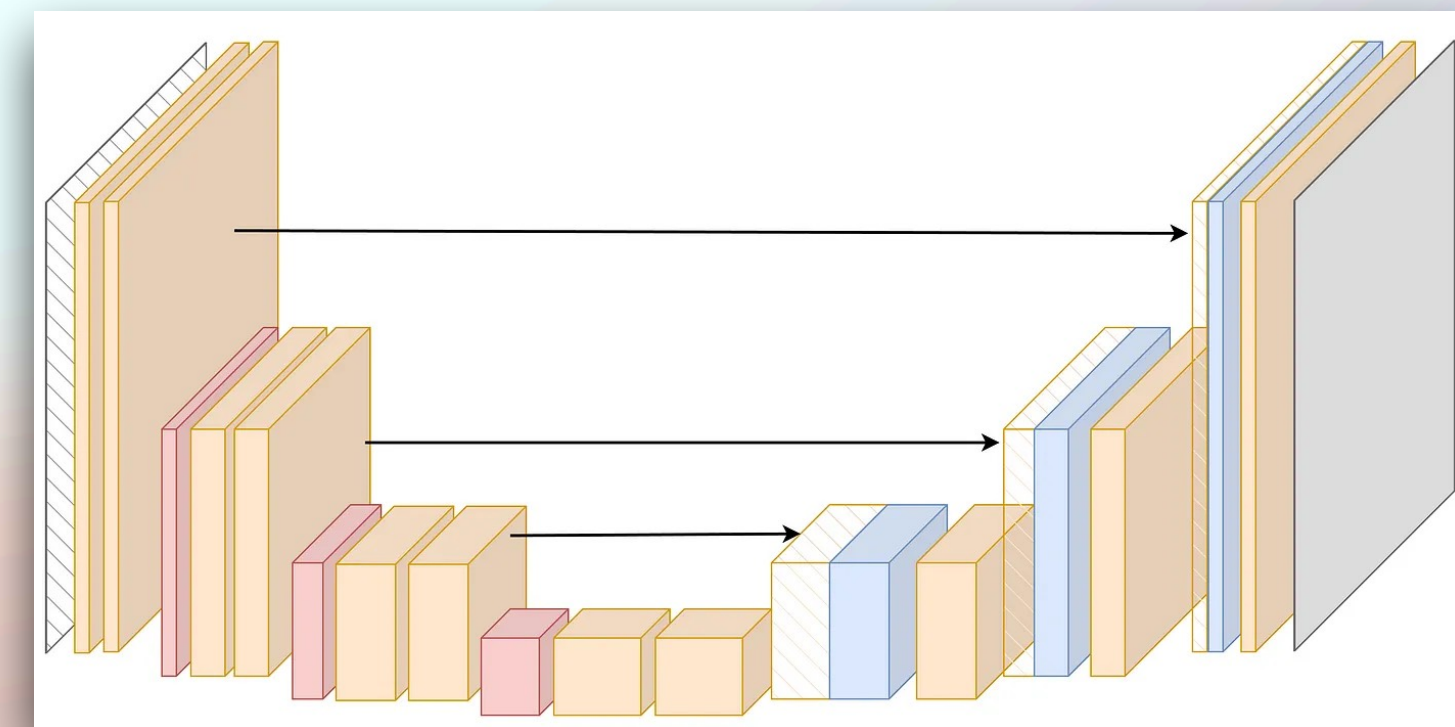
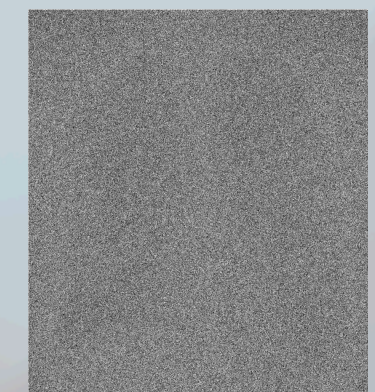


Forward
diffusion

Pure noise



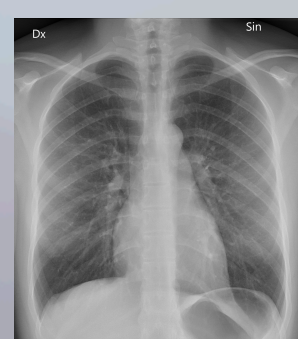
Random seed



Large global (and potentially public)
diffusion model that only ever sees noisy images

Clean chest
X-ray images

Noisy chest
X-ray images



Each hospital with its personal (and smaller) diffusion model

Each client performs t_0 steps of forward diffusion

Forward diffusion

Aggregated noisy dataset

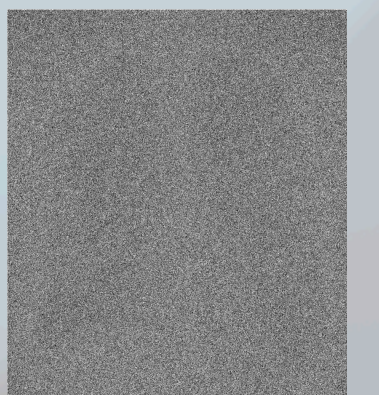
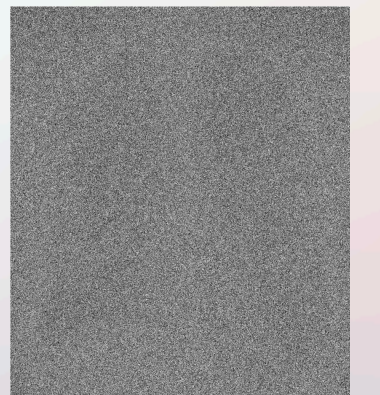
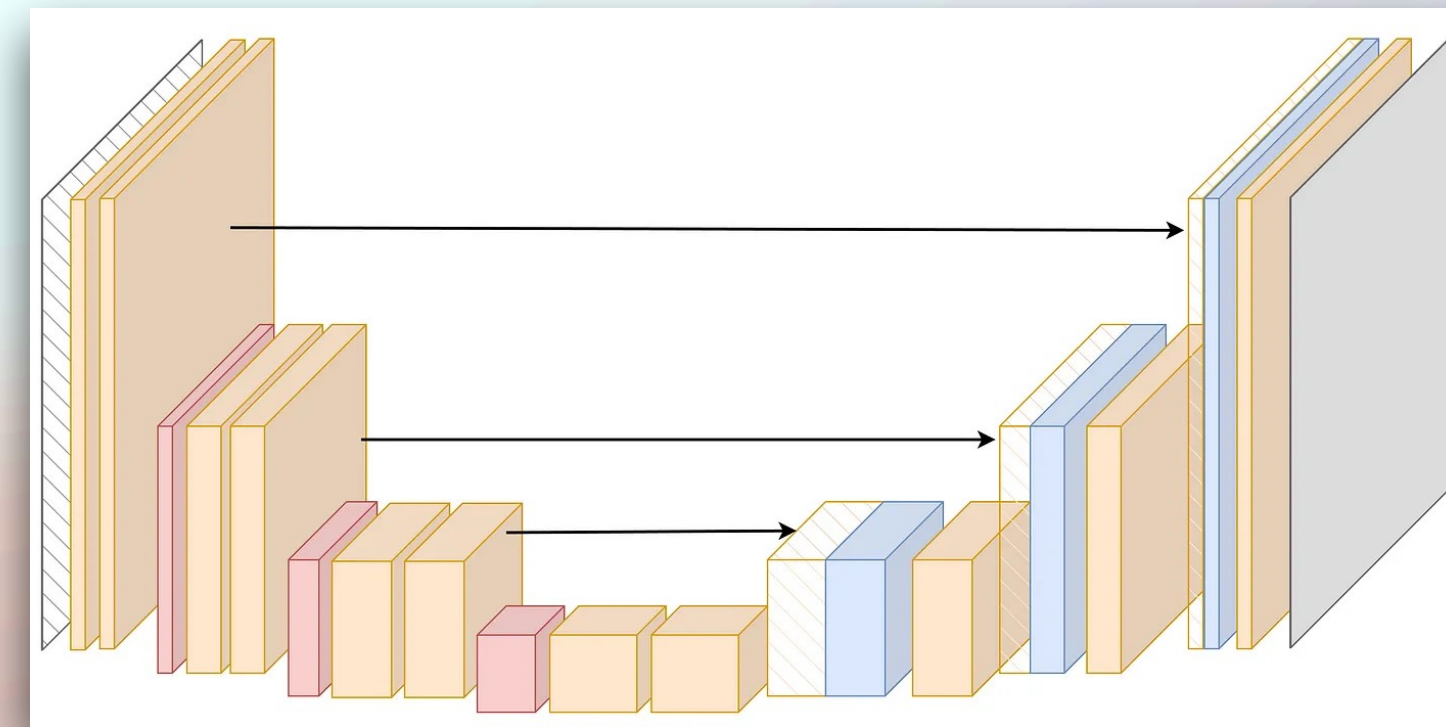
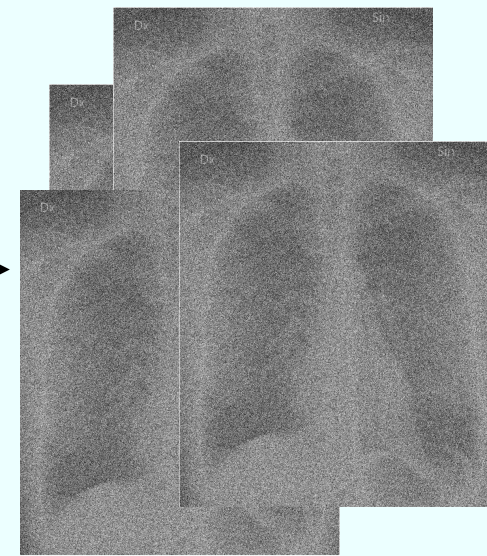
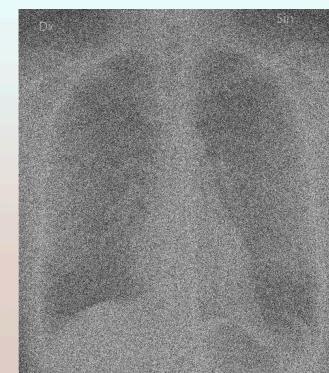
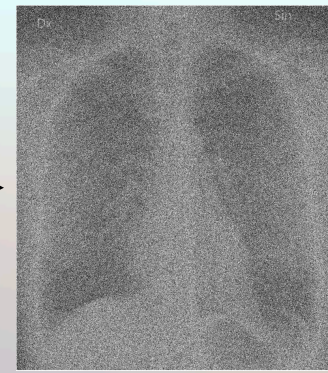
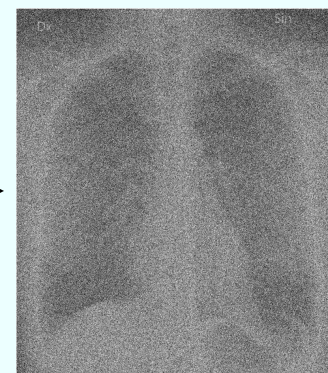
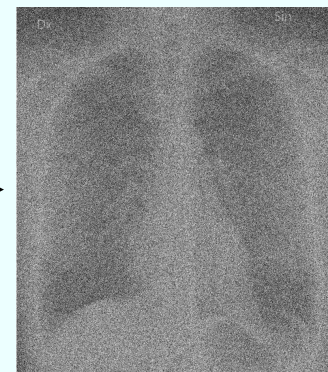
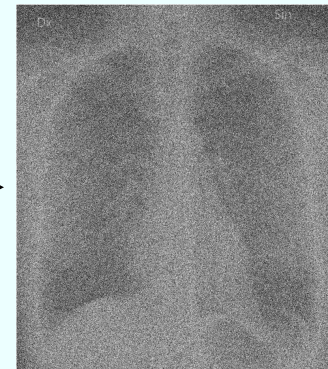
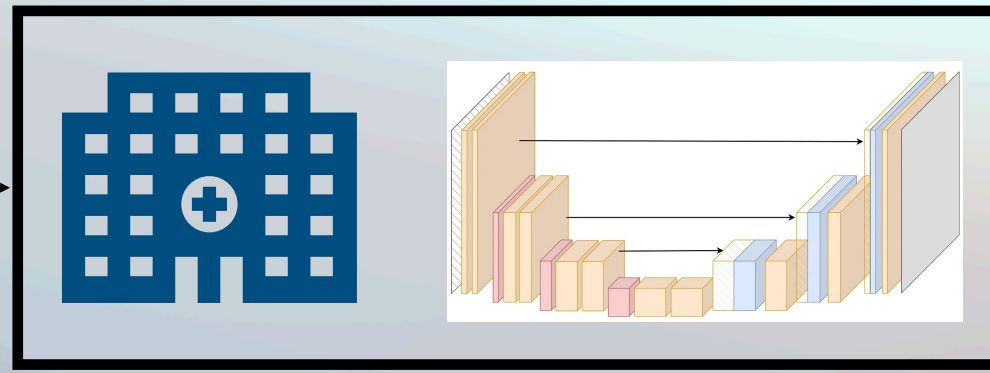
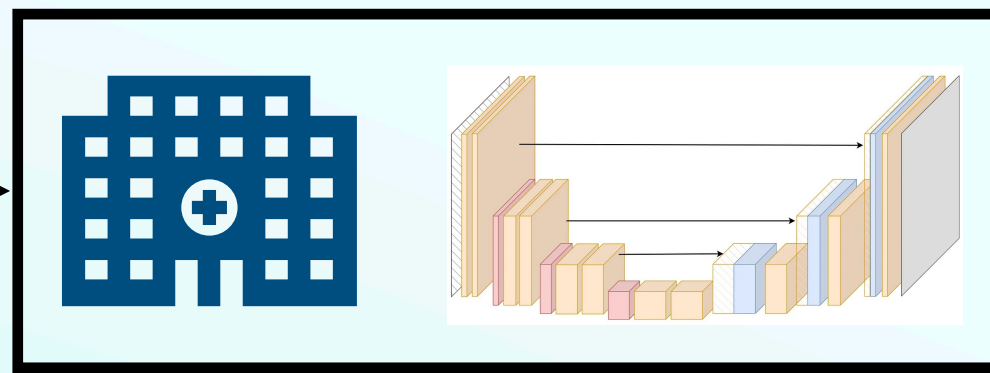
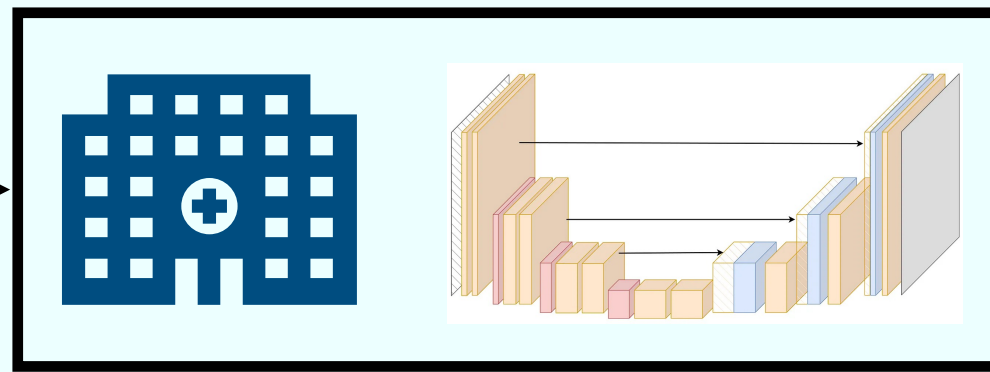
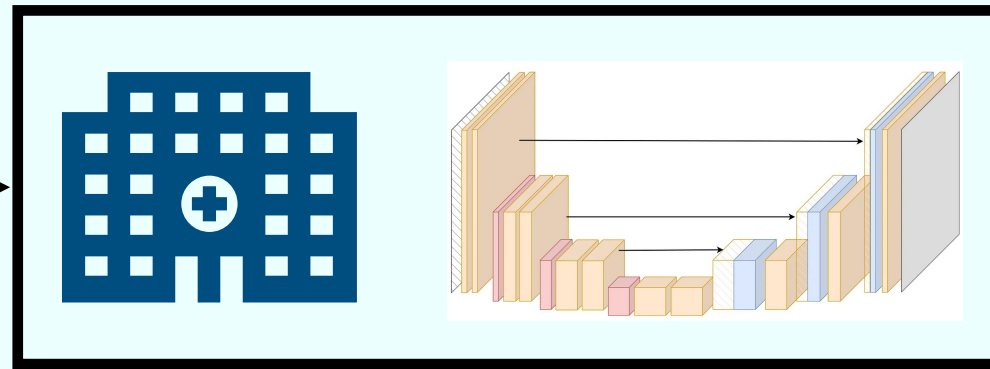
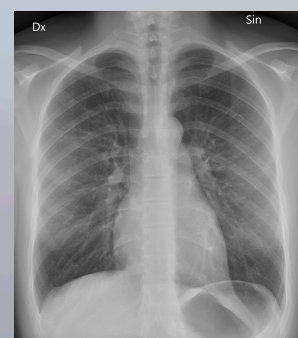
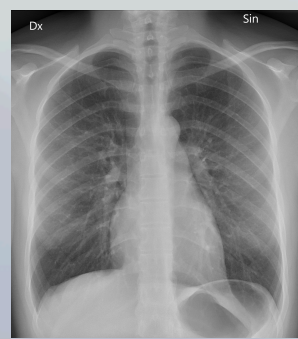
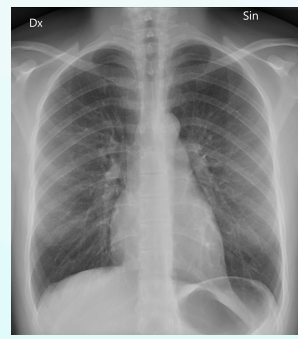
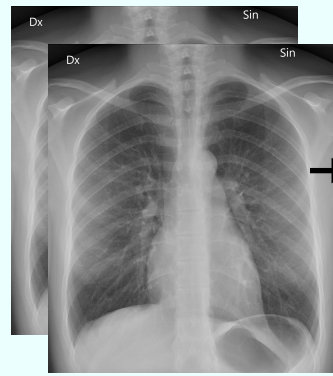
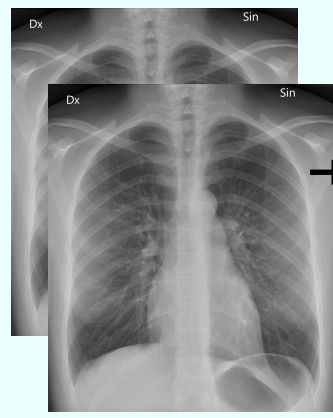
Pure noise

Random seed

Large global (and potentially public) diffusion model that only ever sees noisy images

Noisy chest X-ray images

Clean chest X-ray images



Each hospital with its personal (and smaller) diffusion model

Each client performs t_0 steps of forward diffusion

Forward diffusion

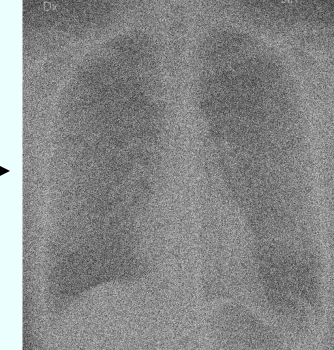
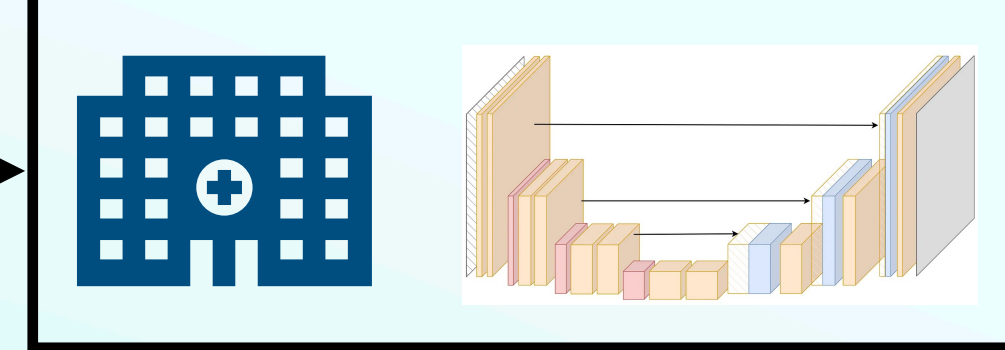
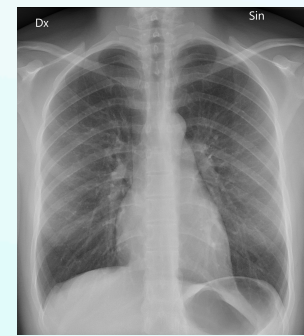
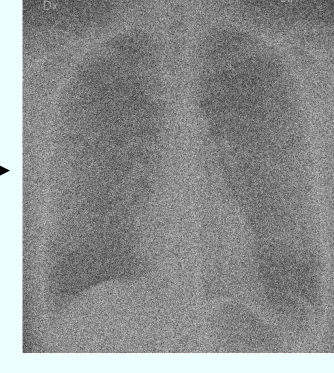
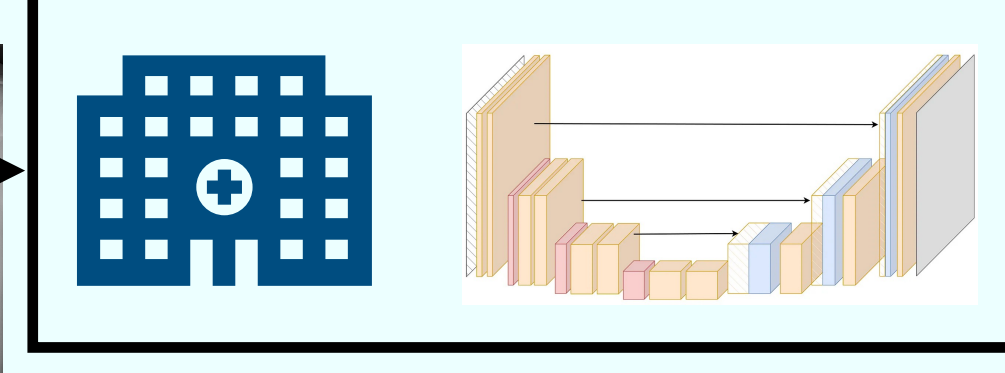
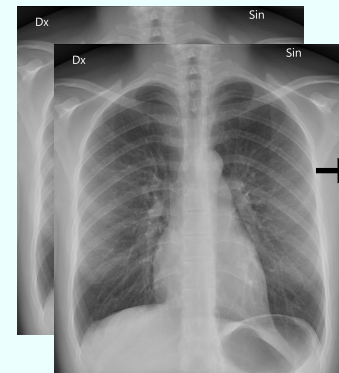
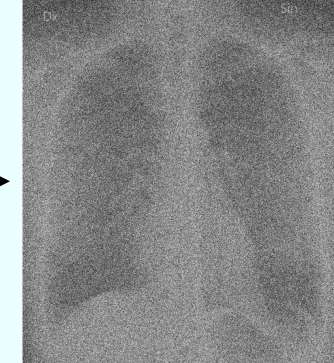
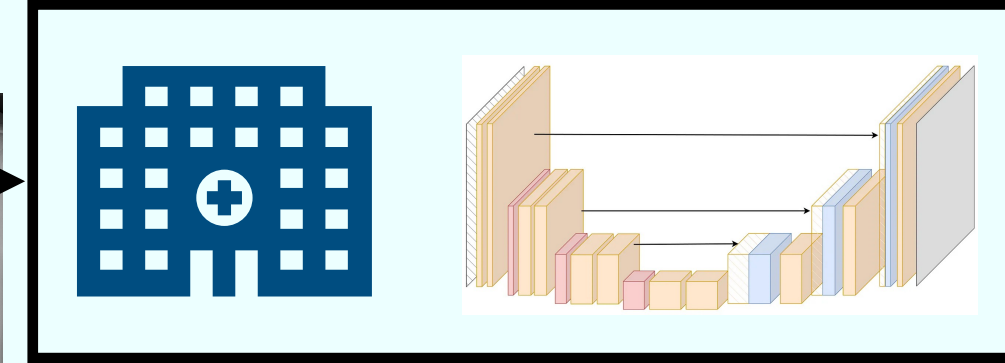
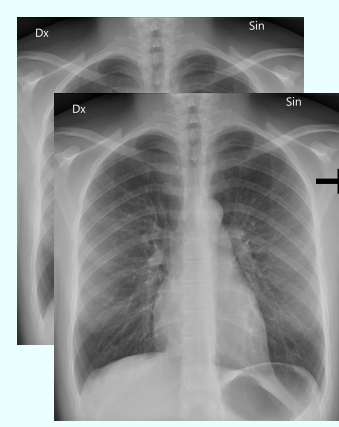
Time step T

Pure noise

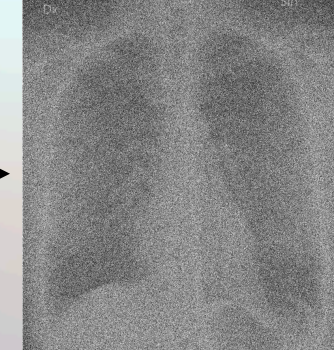
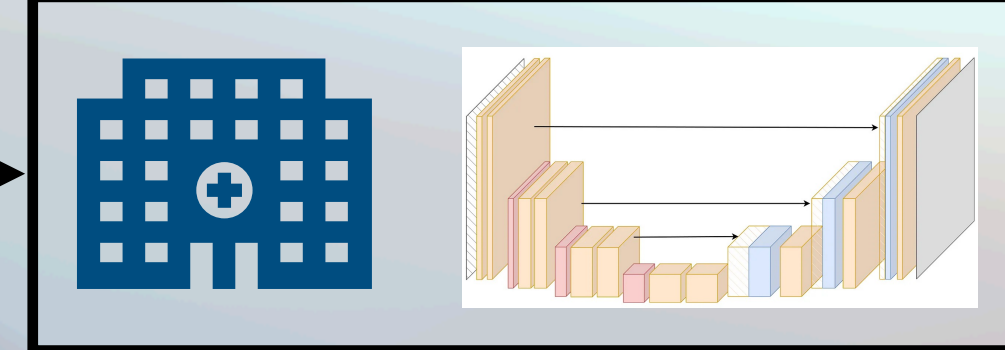
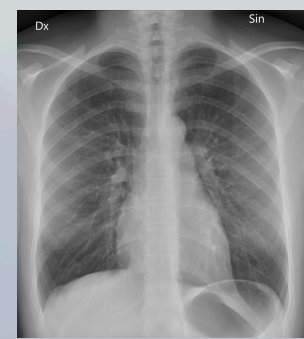
Aggregated noisy dataset

Large global (and potentially public) diffusion model that only ever sees noisy images

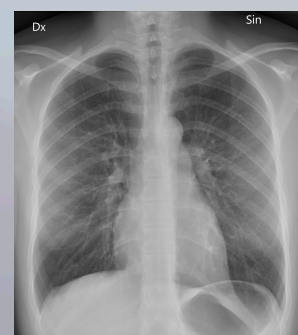
Random seed



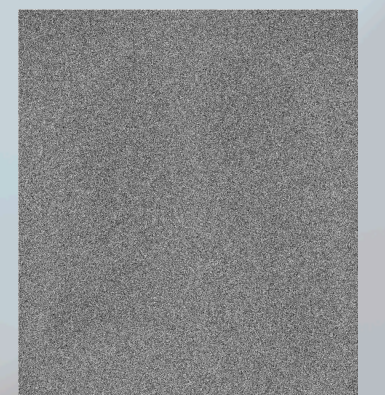
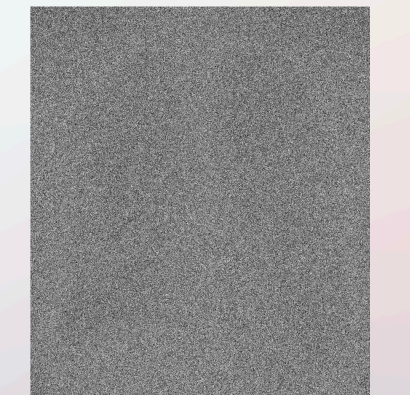
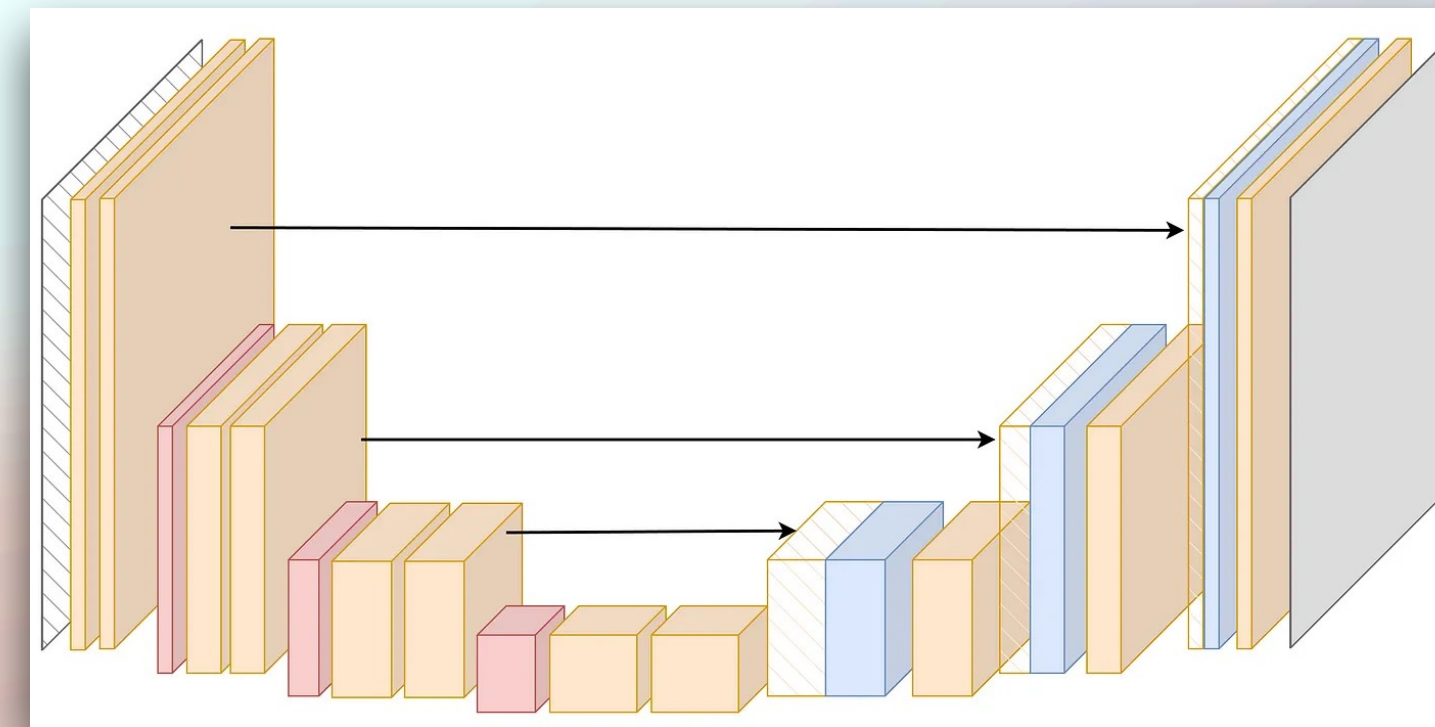
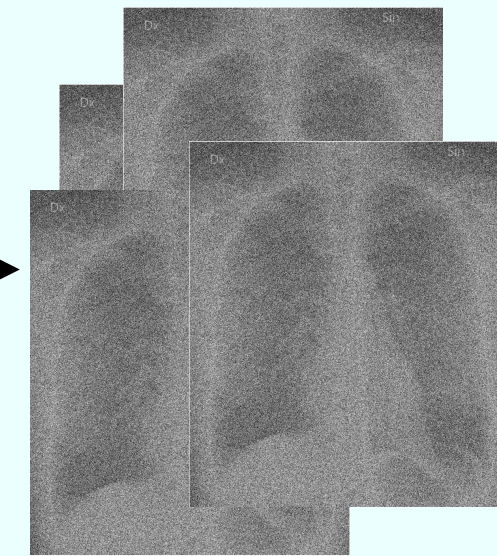
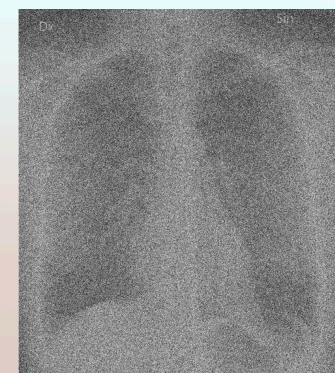
...



Clean chest X-ray images



Noisy chest X-ray images



Each hospital with its personal (and smaller) diffusion model

Each client performs t_0 steps of forward diffusion

Forward diffusion

Time step T

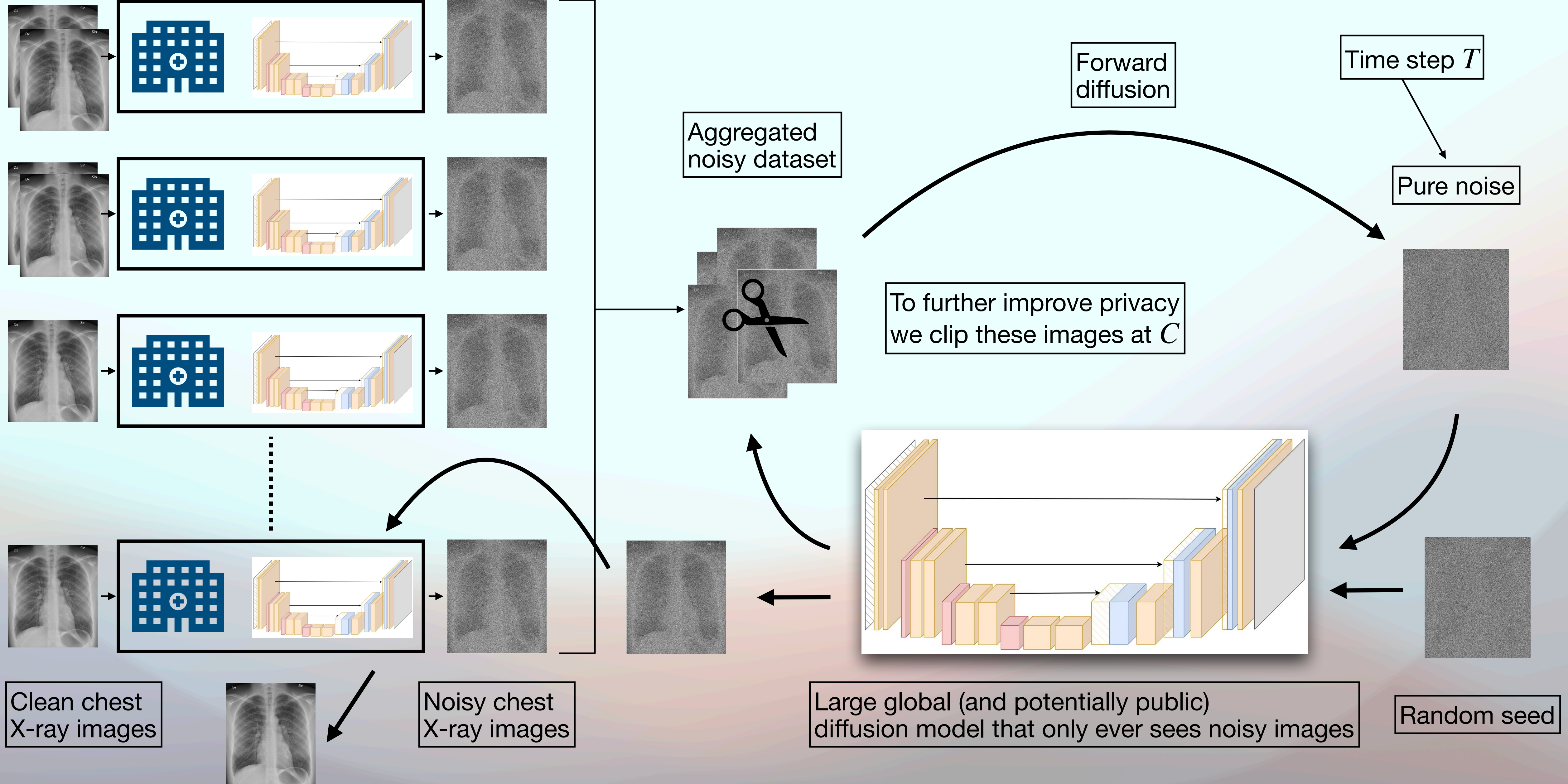
Pure noise

Aggregated noisy dataset

To further improve privacy we clip these images at C

Large global (and potentially public) diffusion model that only ever sees noisy images

Random seed



Each hospital with its personal (and smaller) diffusion model

Each client performs t_0 steps of forward diffusion

Forward diffusion

Time step T

Pure noise

Aggregated noisy dataset

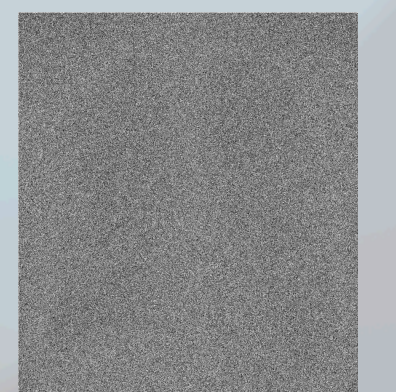
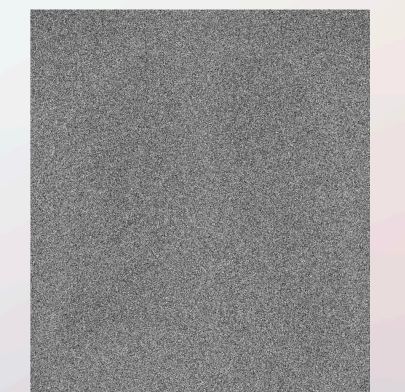
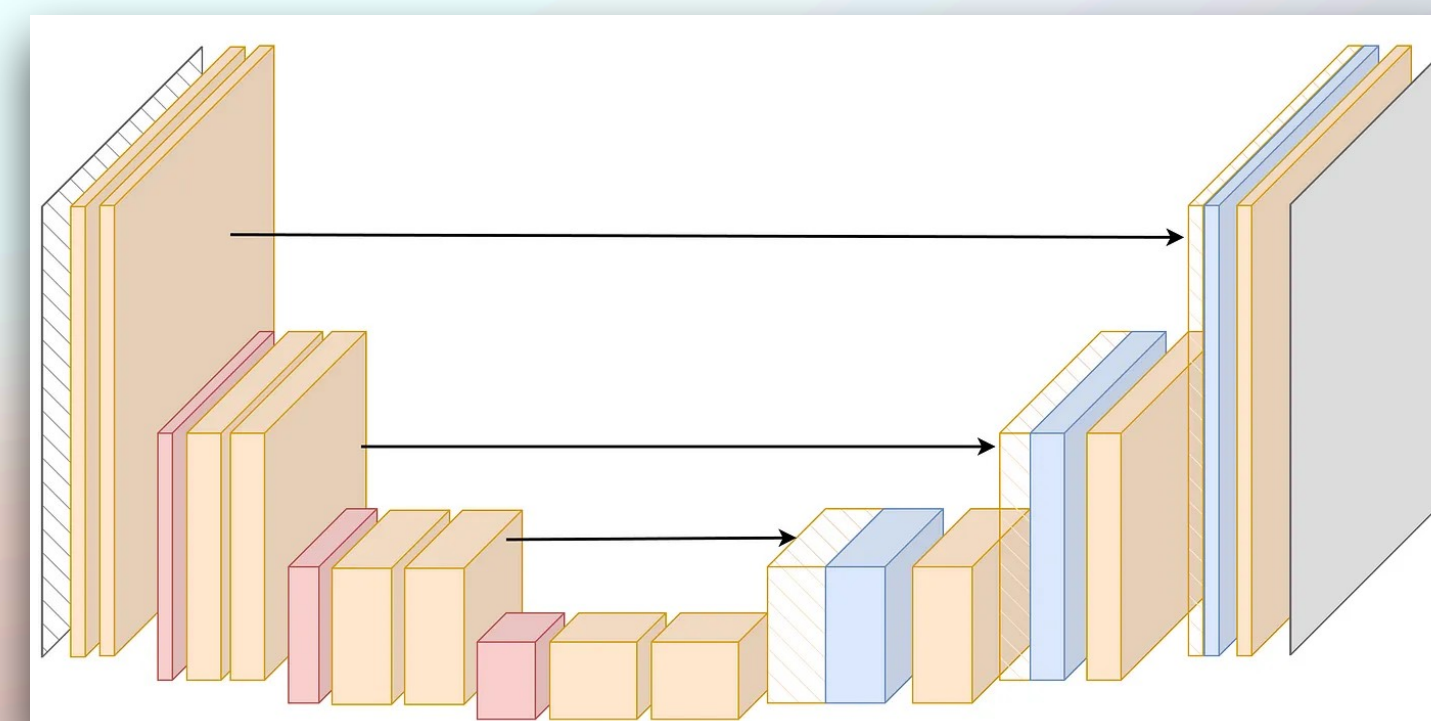
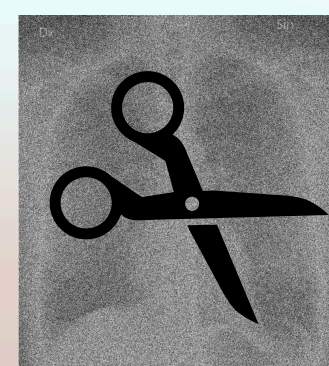
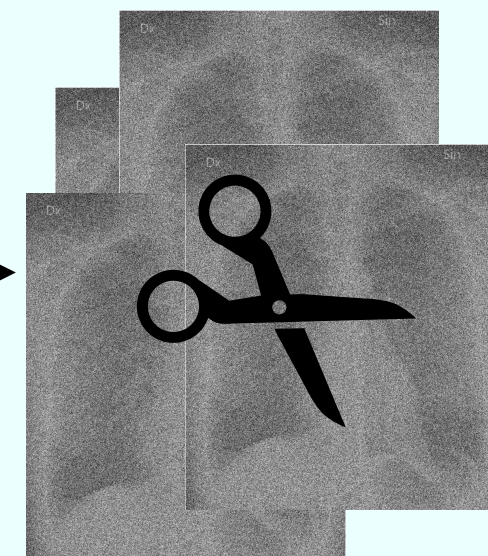
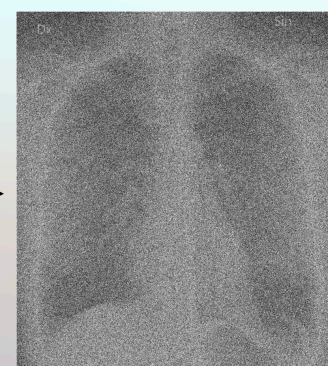
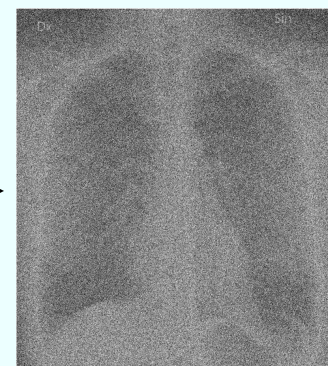
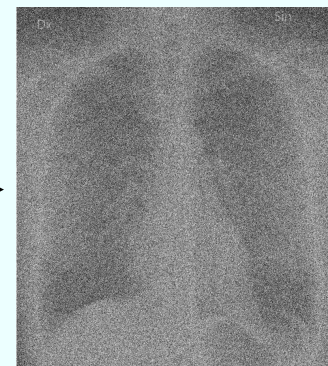
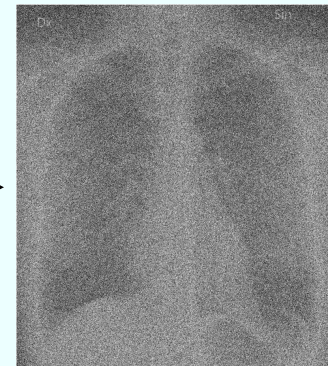
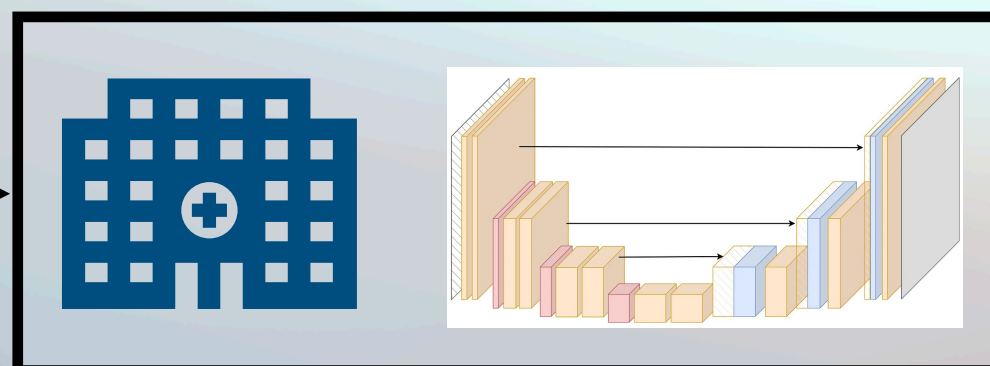
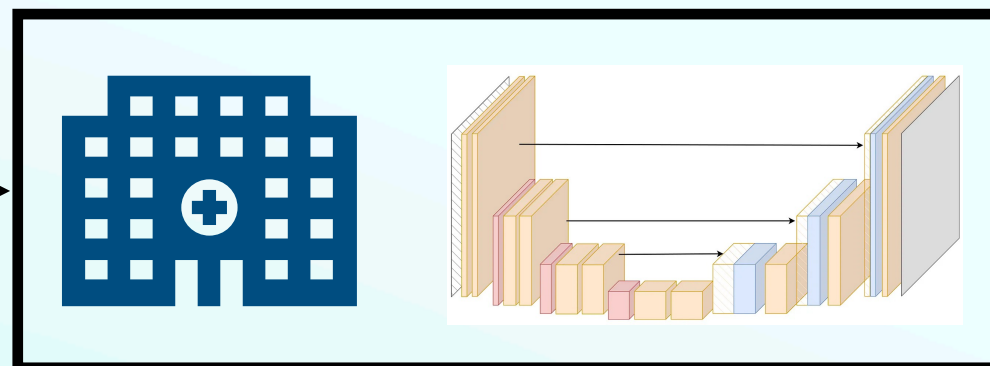
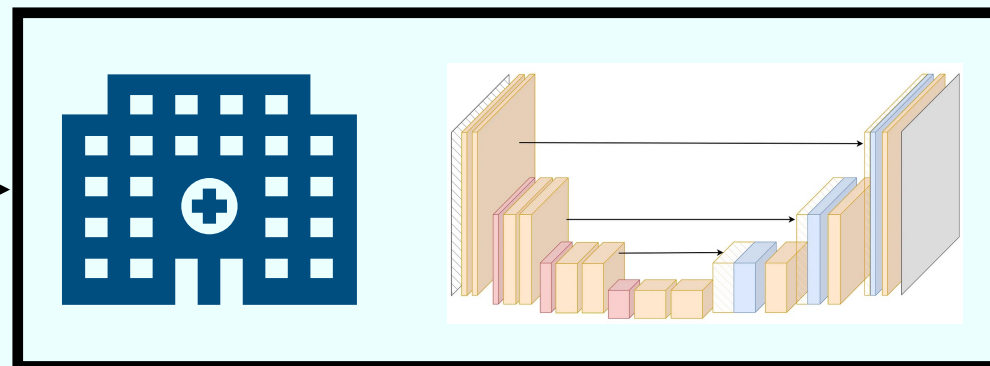
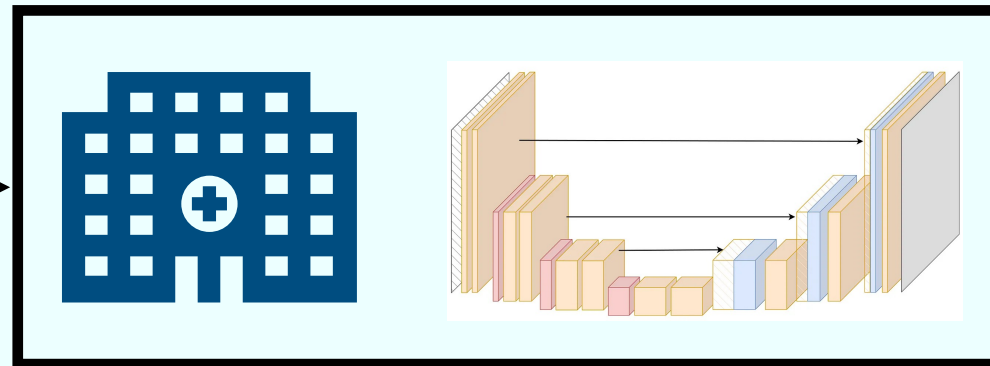
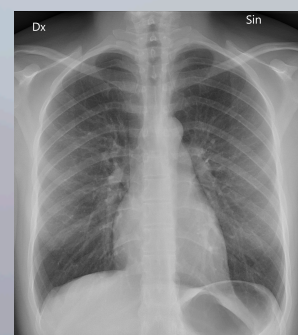
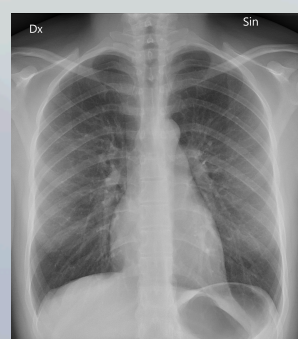
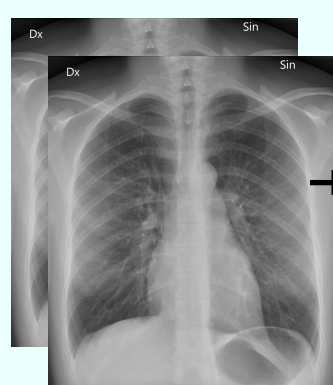
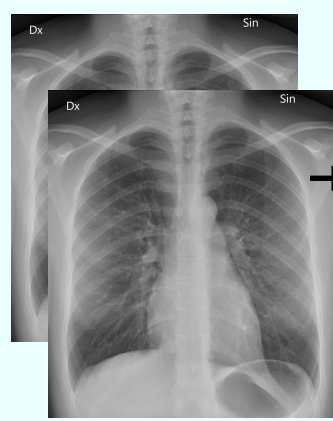
To further improve privacy we clip these images at C

Large global (and potentially public) diffusion model that only ever sees noisy images

Random seed

Noisy chest X-ray images

Clean chest X-ray images



Each hospital with its personal (and smaller) diffusion model

Each client performs t_0 steps of forward diffusion

We augment the local datasets as well but only share the clipped images

Forward diffusion

Time step T

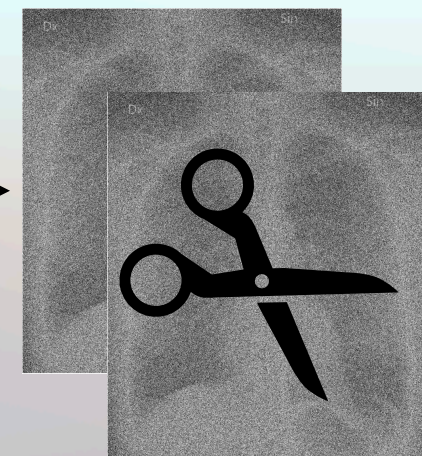
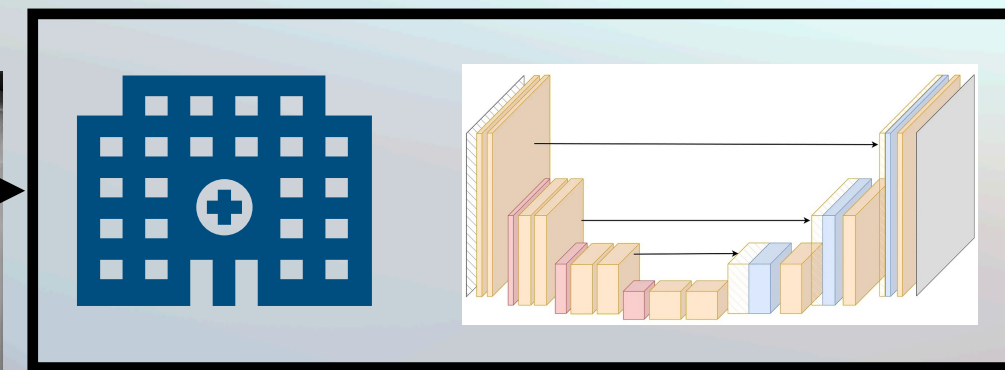
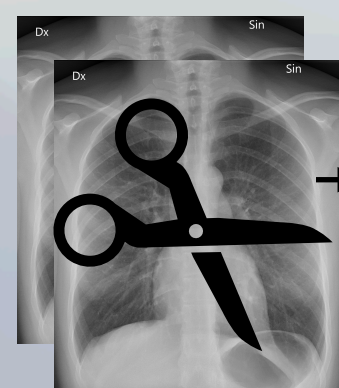
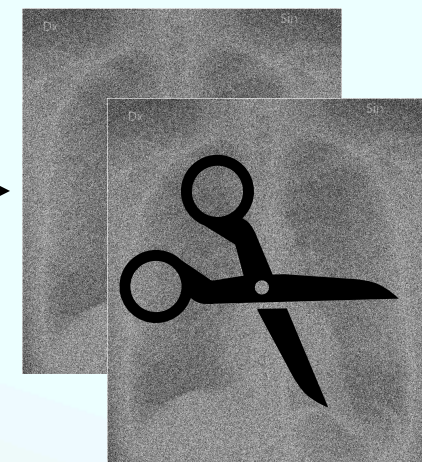
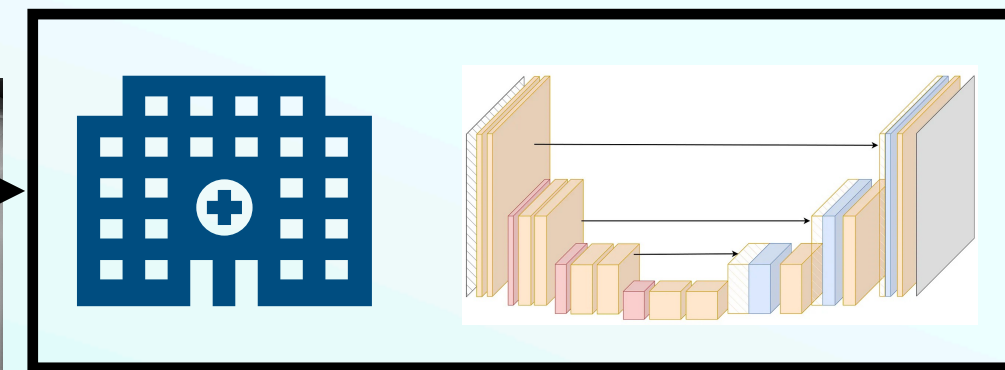
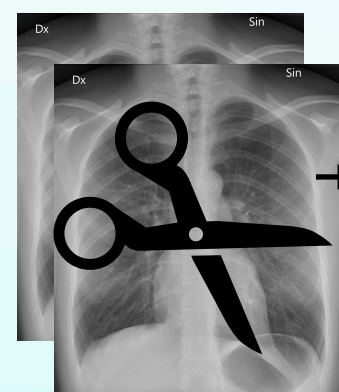
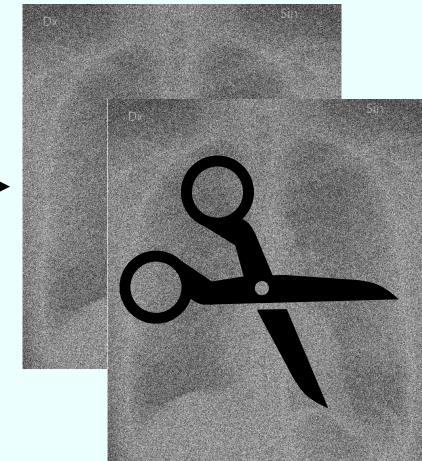
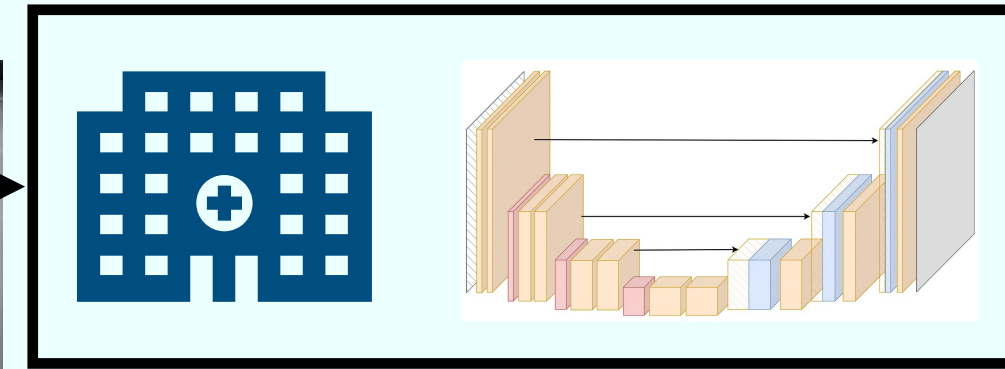
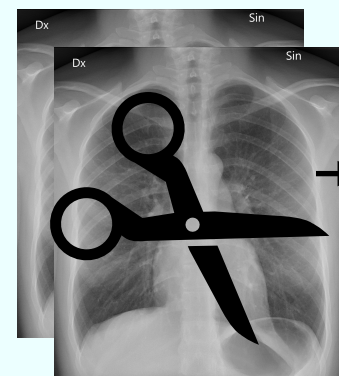
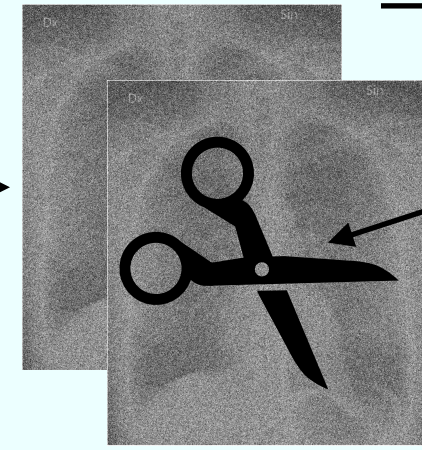
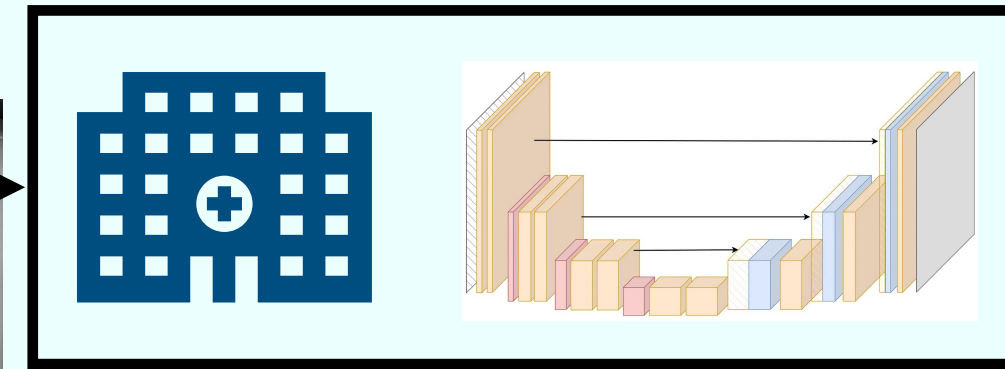
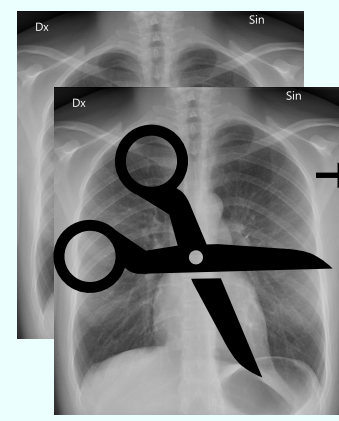
Pure noise

Aggregated noisy dataset

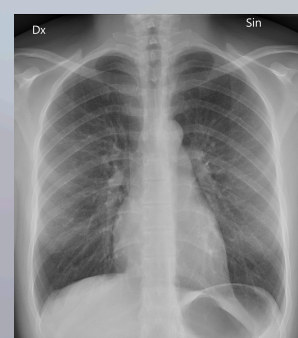
To further improve privacy we clip these images at C

Large global (and potentially public) diffusion model that only ever sees noisy images

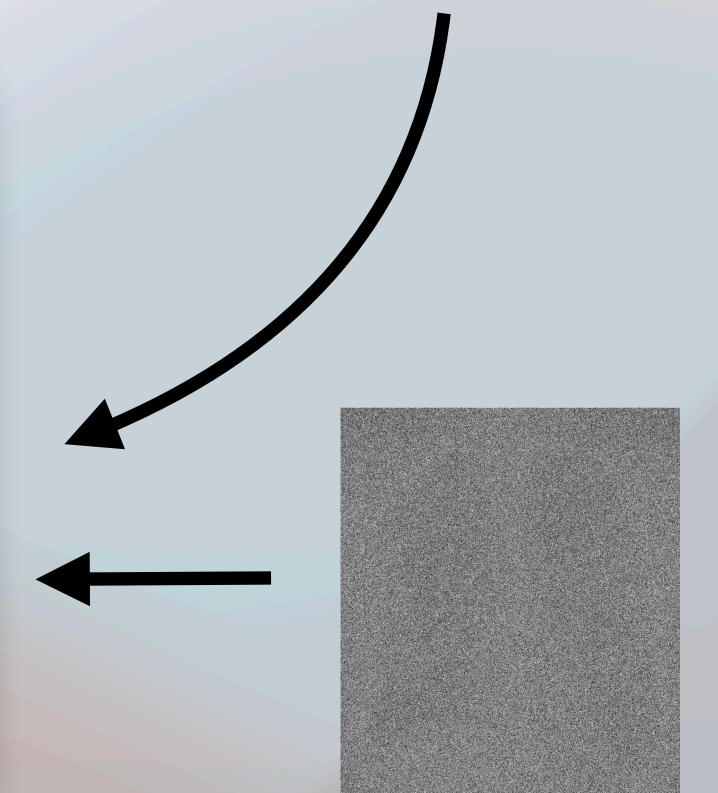
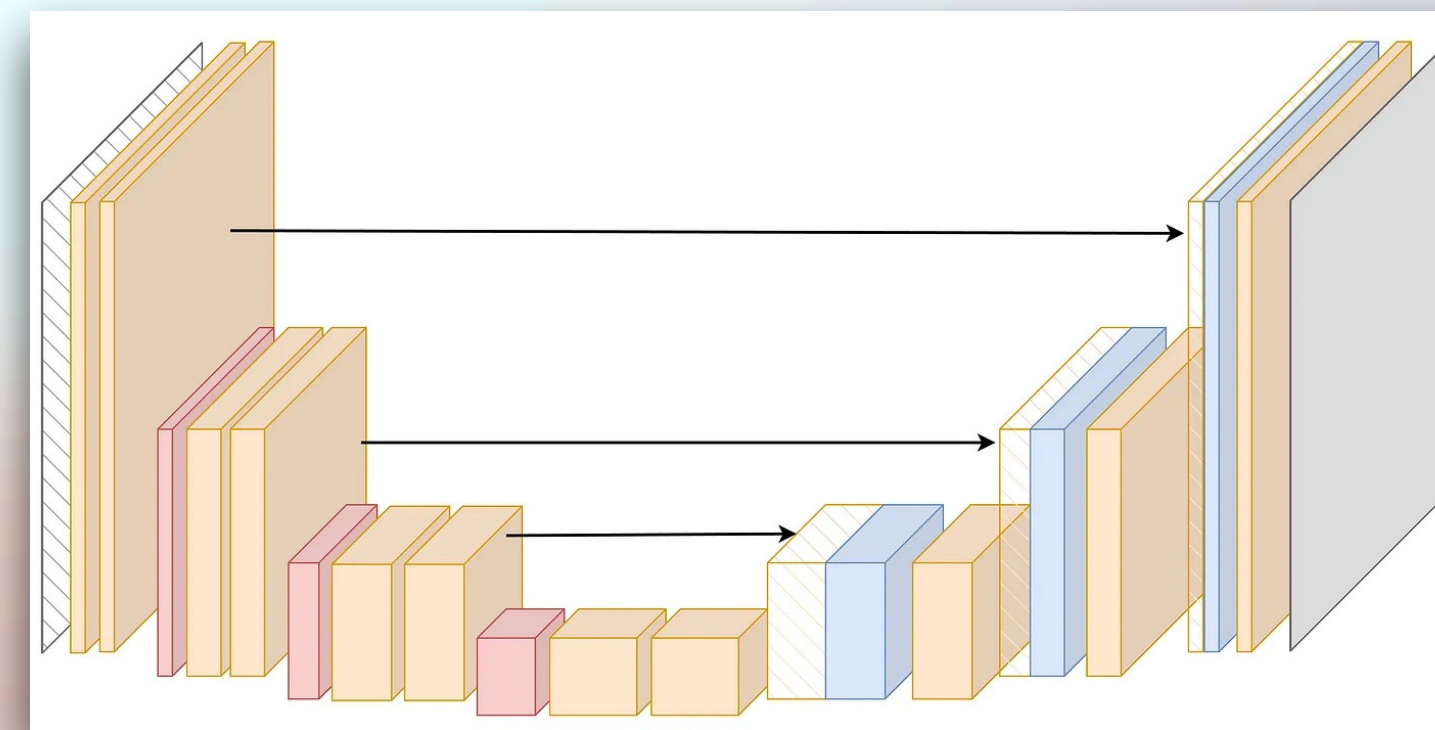
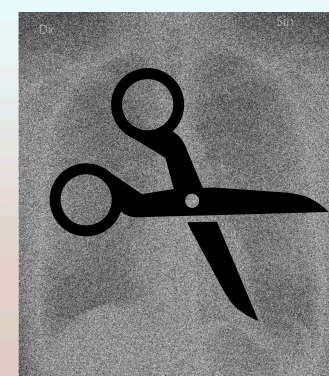
Random seed

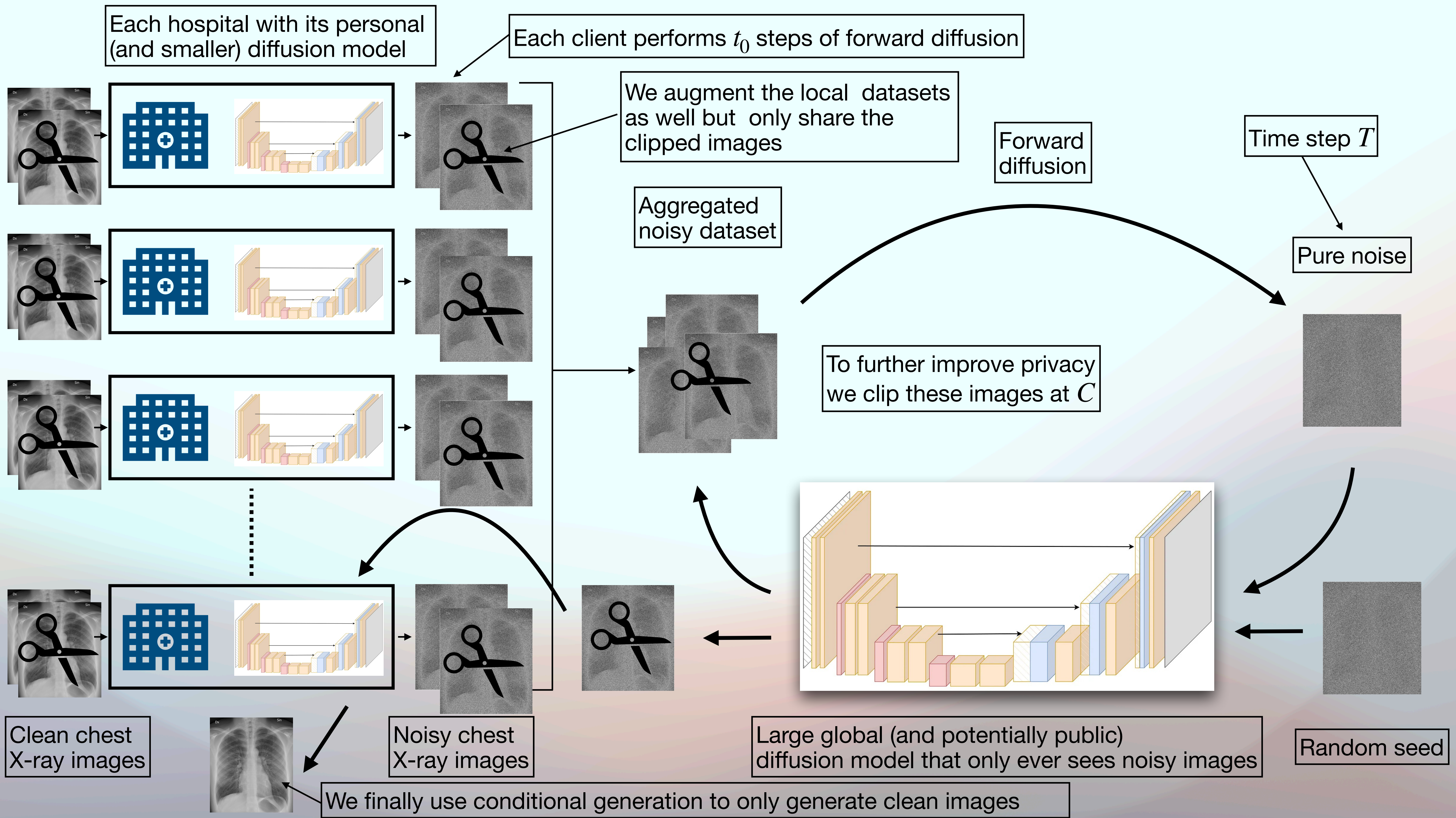


Clean chest X-ray images



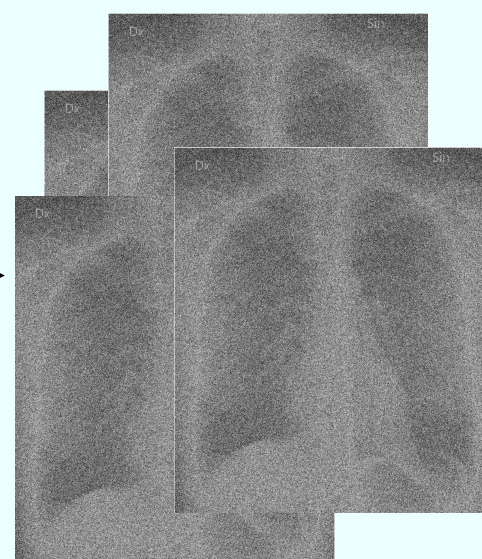
Noisy chest X-ray images





**The aggregate
distributions on
the clients don't
match the
distribution which
is used to train the
global model!!!**

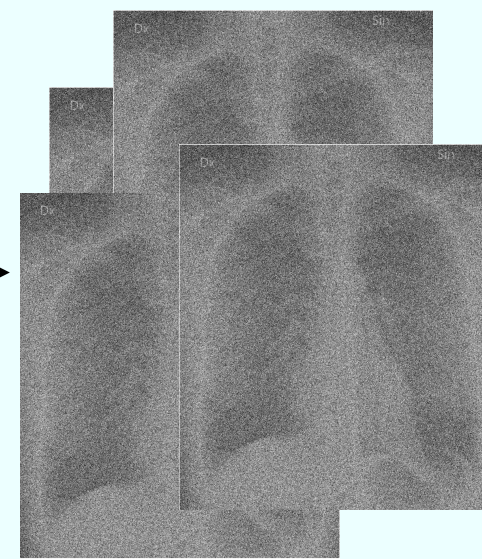
Aggregated
noisy dataset



The aggregate distributions on the clients don't match the distribution which is used to train the global model!!!



Aggregated
noisy dataset



If you squint enough the clipped and the unclipped images look the same upon adding noise :)

Our LDP Guarantee

Theorem: Given the training datasets $\bar{S} = \{S_m\}_{m \in [M]}$, if we choose the local diffusion depth as t_0 , total number of steps as T , some noise schedule $\{\beta_t\}_{t=1}^T$, and clipping parameter C , then, for all $x \in \bar{S}$, i.e., each image in the aggregate dataset the output of our satisfies,

$$\left(\frac{2C^2}{\sigma^2} + C \sqrt{\frac{8 \log(1/\delta)}{\sigma^2}}, \delta \right)\text{-Local DP}$$

where $\sigma^2 = (1 - \bar{\alpha}_{t_0})/\bar{\alpha}_{t_0}$ and $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$.

Our LDP Guarantee

Theorem: Given the training datasets $\bar{S} = \{S_m\}_{m \in [M]}$, if we choose the local diffusion depth as t_0 , total number of steps as T , some noise schedule $\{\beta_t\}_{t=1}^T$, and clipping parameter C , then, for all $x \in \bar{S}$, i.e., each image in the aggregate dataset the output of our satisfies,

$$\left(\frac{2C^2}{\sigma^2} + C \sqrt{\frac{8 \log(1/\delta)}{\sigma^2}}, \delta \right)\text{-Local DP}$$

where $\sigma^2 = (1 - \bar{\alpha}_{t_0})/\bar{\alpha}_{t_0}$ and $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$.

For example, with $T = 1000$, a linear noise schedule, and clipping parameter $C = 10$, choosing $t_0 = 690$ yields an $(10, 10^{-5})$ -LDP guarantee.

For the rest of the talk we will fix the privacy level to $(\epsilon = 10, \delta = 10^{-5})$ -LDP

Outputs of the Shared Model

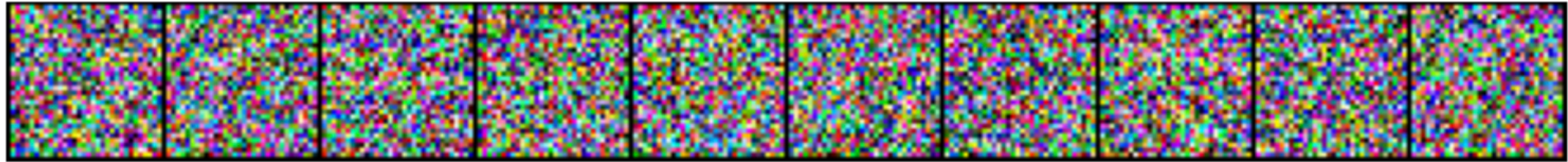
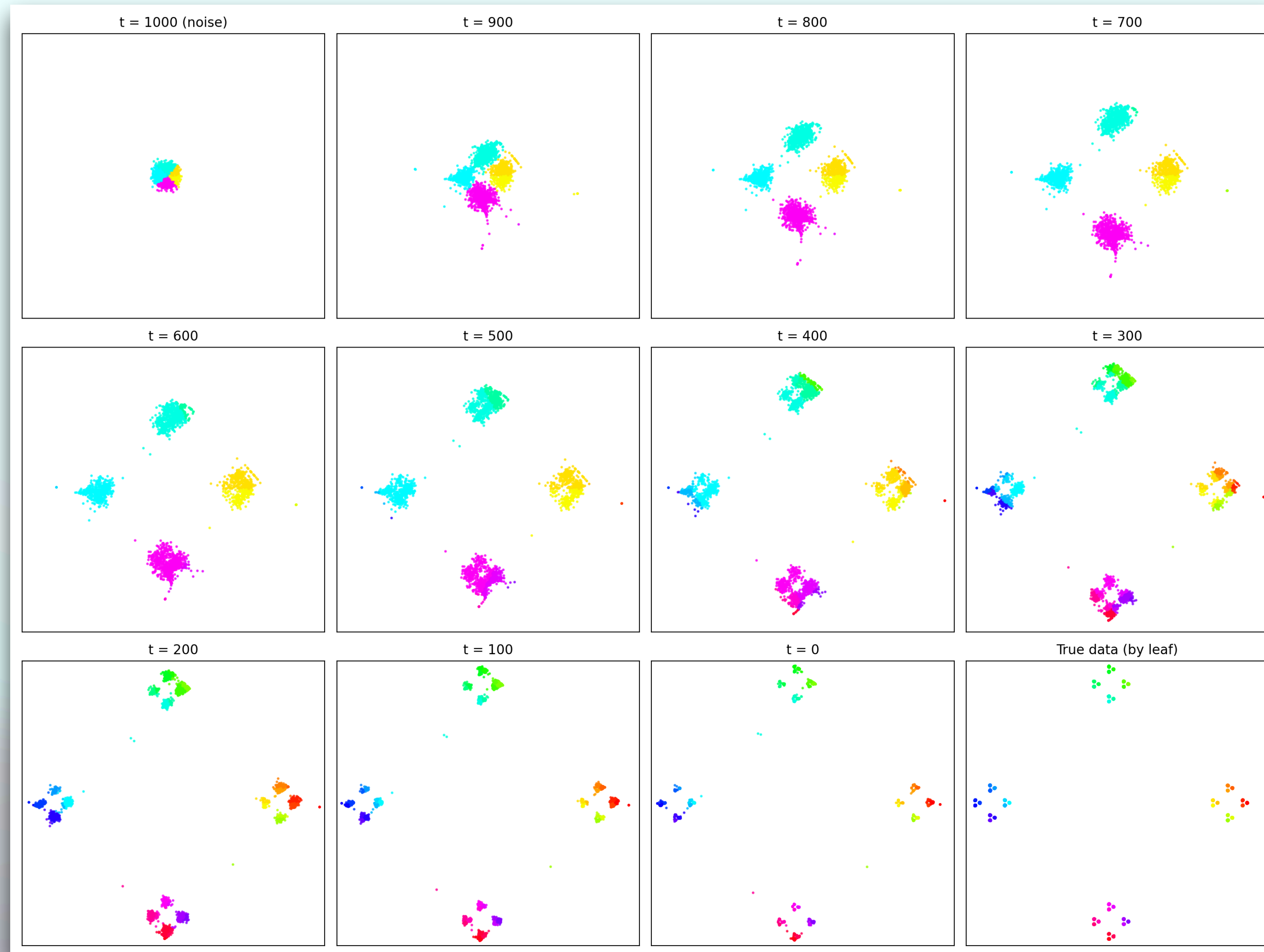


Figure 3. Samples generated by the global model used in our framework with red digits 0 to 9. The output digit shapes remain unrecognizable, indicating strong privacy protection of our method.

Exploiting the Progressive Refinement of Diffusion Models



Membership Inference (PIA) Attack

These are
white-box
attacks

	CIFAR-10	Colorized MNIST	CelebA
AUC	50.01	49.70	50.08
ASR	50.15	50.10	50.34
TPR@1% FPR	0.82	1.07	0.86

Our Model

	CIFAR10	Colorized MNIST	CelebA
MODELS	AUC /ASR/TPR@1% FPR	AUC /ASR/TPR@1% FPR	AUC /ASR/TPR@1% FPR
S1	53.03/52.44/1.23	62.71/59.44/1.51	61.59/58.36/3.28
S2	64.79/60.91/2.53	97.71/93.96/52.52	90.15/82.12/31.36
S3	82.13/75.14/8.75	99.62/97.72/93.39	99.59/97.74/93.34

Non-private Centralized Models

Membership Inference (PIA) Attack

These are
white-box
attacks

	CIFAR-10	Colorized MNIST	CelebA
AUC	50.01	49.70	50.08
ASR	50.15	50.10	50.34
TPR@1% FPR	0.82	1.07	0.86

Our Model

	CIFAR10	Colorized MNIST	CelebA
MODELS	AUC /ASR/TPR@1% FPR	AUC /ASR/TPR@1% FPR	AUC /ASR/TPR@1% FPR
S1	53.03/52.44/1.23	62.71/59.44/1.51	61.59/58.36/3.28
S2	64.79/60.91/2.53	97.71/93.96/52.52	90.15/82.12/31.36
S3	82.13/75.14/8.75	99.62/97.72/93.39	99.59/97.74/93.34

Non-private Centralized Models

Membership Inference (PIA) Attack

These are
white-box
attacks

	CIFAR-10	Colorized MNIST	CelebA
AUC	50.01	49.70	50.08
ASR	50.15	50.10	50.34
TPR@1% FPR	0.82	1.07	0.86

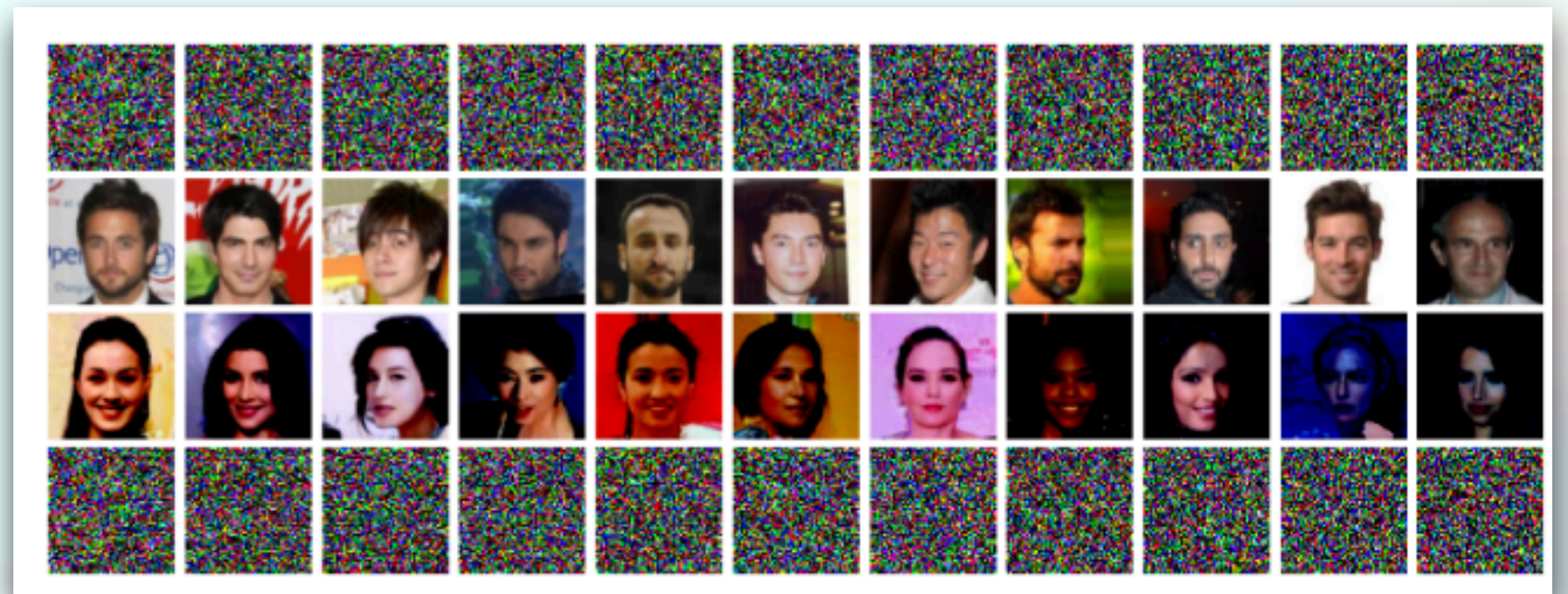
What about
DP training?

Our Model

	CIFAR10	Colorized MNIST	CelebA
MODELS	AUC /ASR/TPR@1% FPR	AUC /ASR/TPR@1% FPR	AUC /ASR/TPR@1% FPR
S1	53.03/52.44/1.23	62.71/59.44/1.51	61.59/58.36/3.28
S2	64.79/60.91/2.53	97.71/93.96/52.52	90.15/82.12/31.36
S3	82.13/75.14/8.75	99.62/97.72/93.39	99.59/97.74/93.34

Non-private Centralized Models

Reconstruction Attack

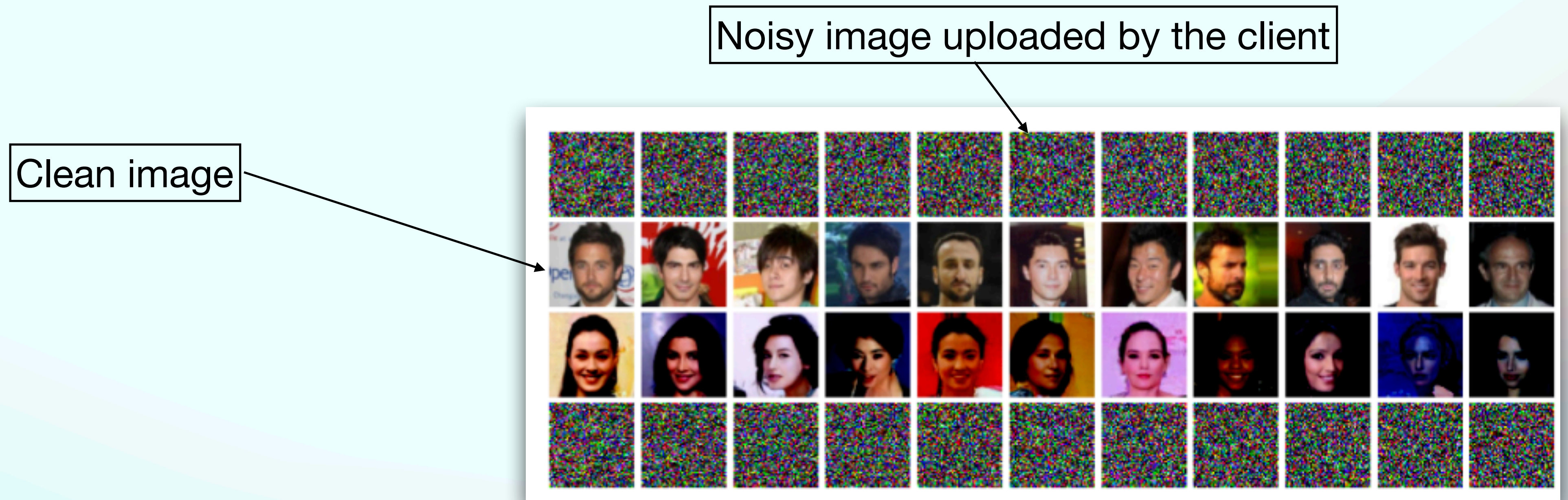


Reconstruction Attack

Noisy image uploaded by the client



Reconstruction Attack



Reconstruction Attack

Noisy image uploaded by the client

Clean image

Reconstruction attack #1 (can access the global model as well as a denser trained on a related data distribution)



Reconstruction Attack

Noisy image uploaded by the client

Clean image

Reconstruction attack #1 (can access the global model as well as a denser trained on a related data distribution)



For the CelebA dataset, the pre-trained attack model is trained on a dataset that consists of 5,000 female-black-hair, 5,000 female-non-black-hair, and 50 male-non-black-hair samples. For this dataset, we aim to reconstruct images with male and black hair.

Reconstruction Attack

Noisy image uploaded by the client

Clean image

Reconstruction attack #1 (can access the global model as well as a denser trained on a related data distribution)

For the CelebA dataset, the pre-trained attack model is trained on a dataset that consists of 5,000 female-black-hair, 5,000 female-non-black-hair, and 50 male-non-black-hair samples. For this dataset, we aim to reconstruct images with male and black hair.

Reconstruction attack #2 (can only access the global model)



Personalization Alleviates Heterogeneity Issues



Personalization Alleviates Heterogeneity Issues

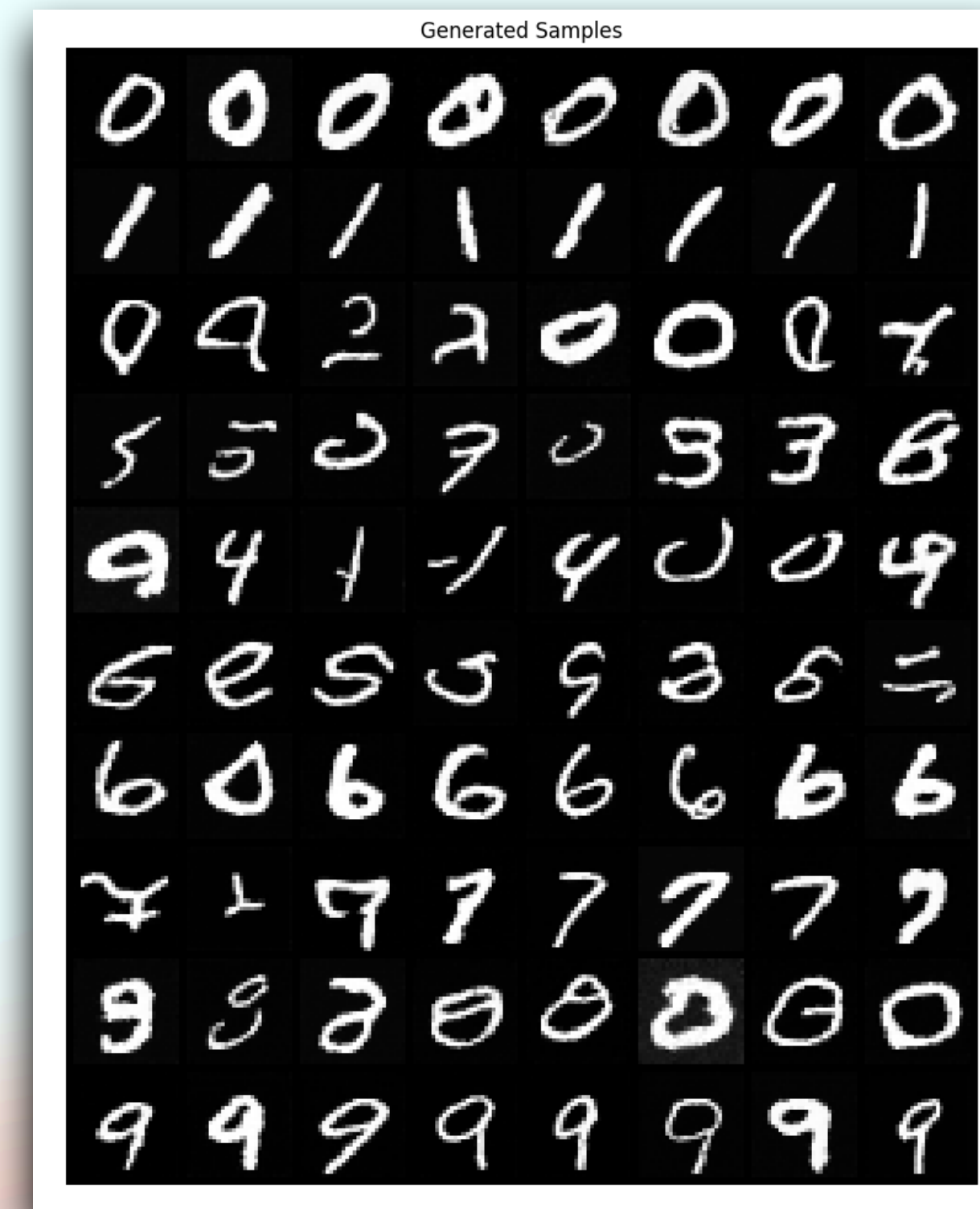


Personalized Training

Personalization Alleviates Heterogeneity Issues



Personalized Training



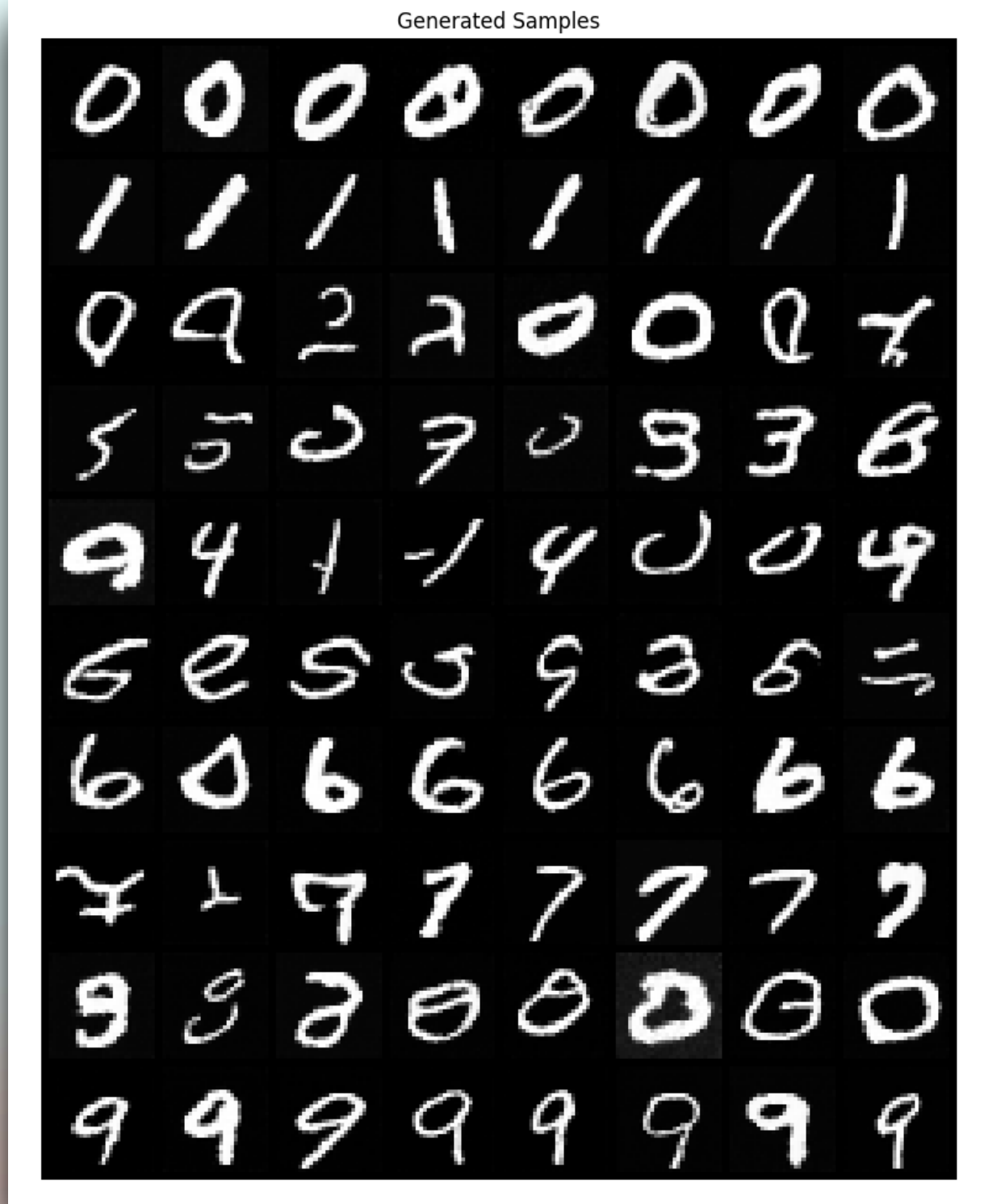
Pure Local Training

Personalization Alleviates Heterogeneity Issues

Minority Classes: 2,3,4,5,8



Personalized Training



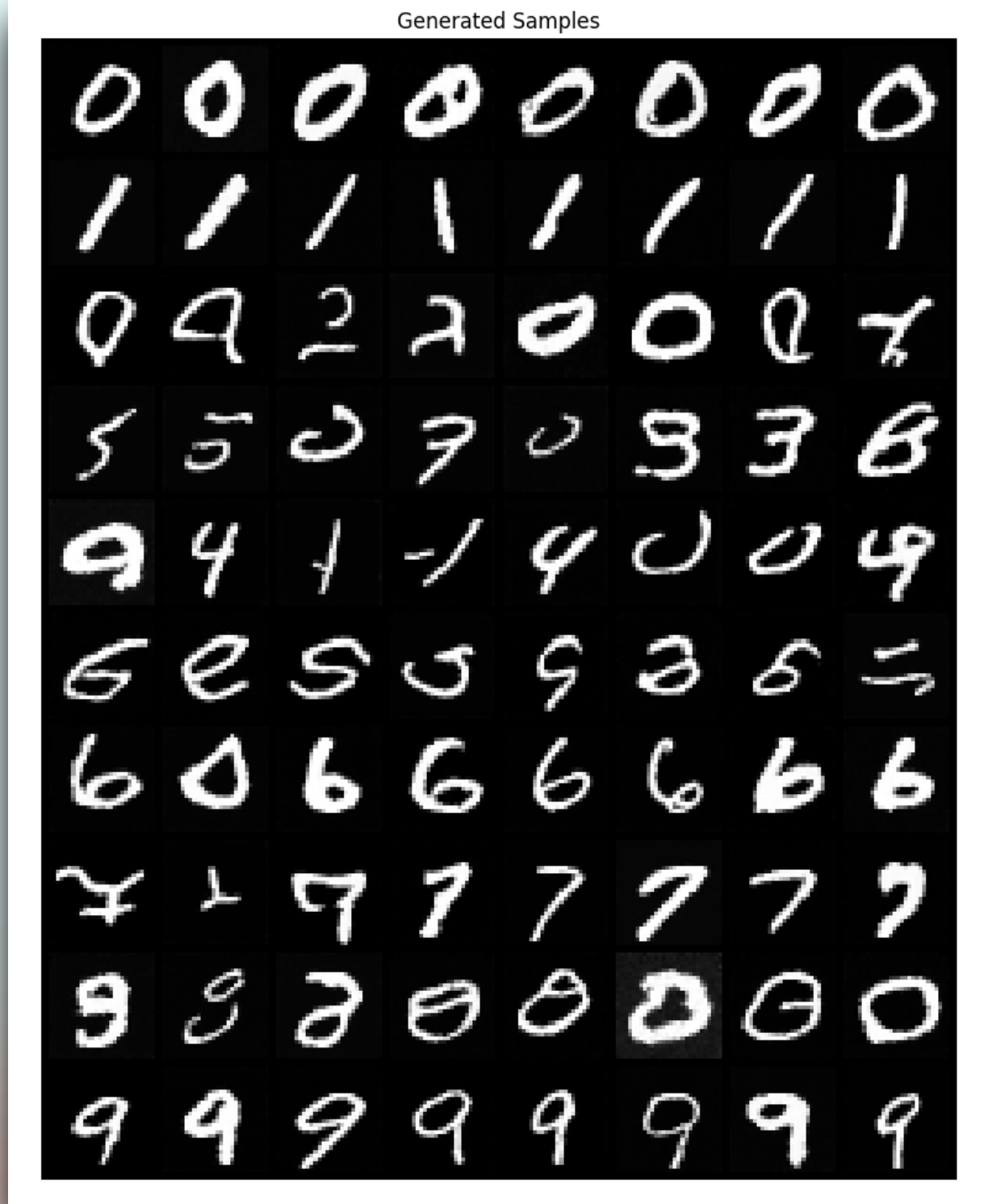
Pure Local Training

Personalization Alleviates Heterogeneity Issues

Minority Classes: 2,3,4,5,8



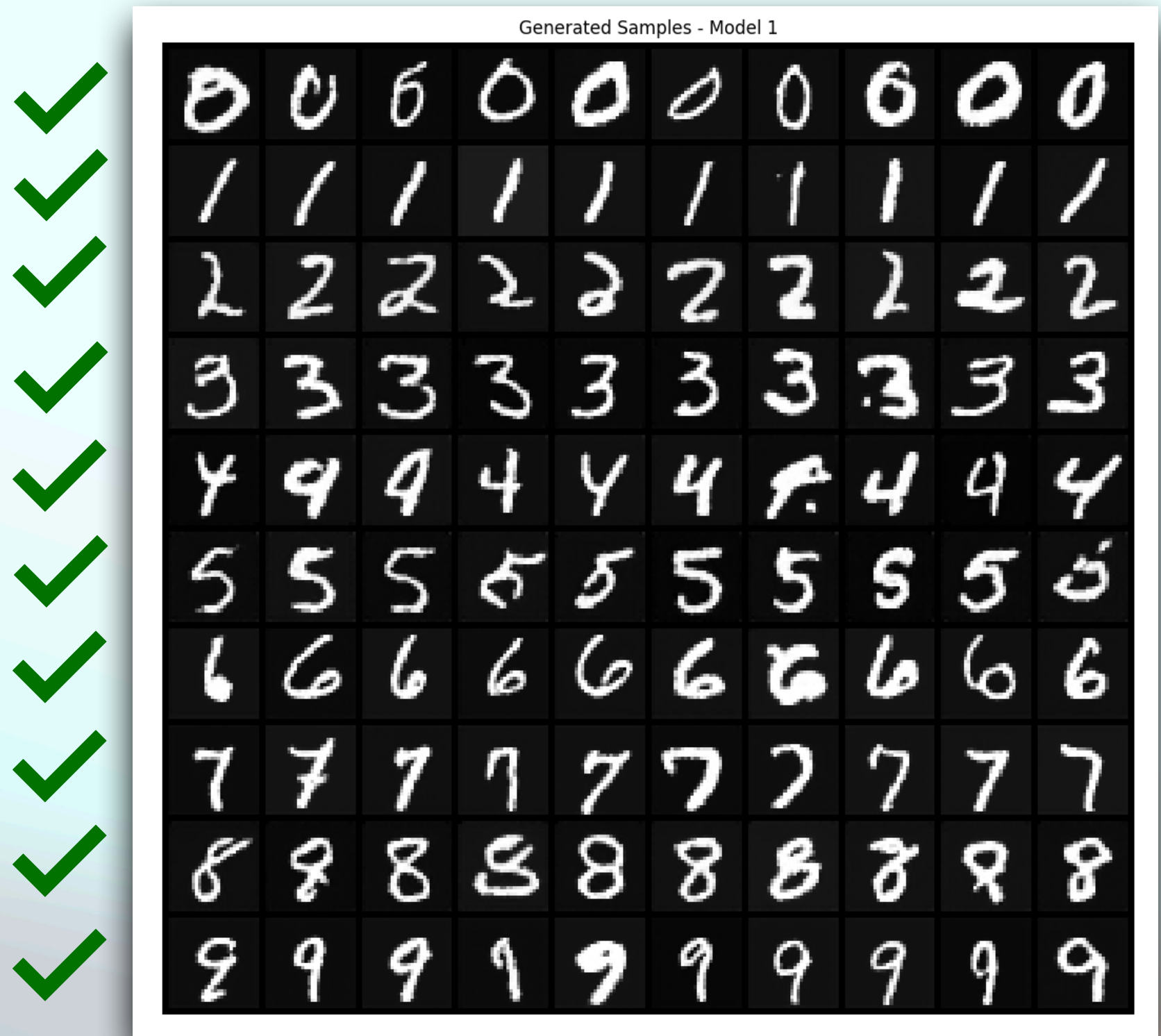
Personalized Training



Pure Local Training

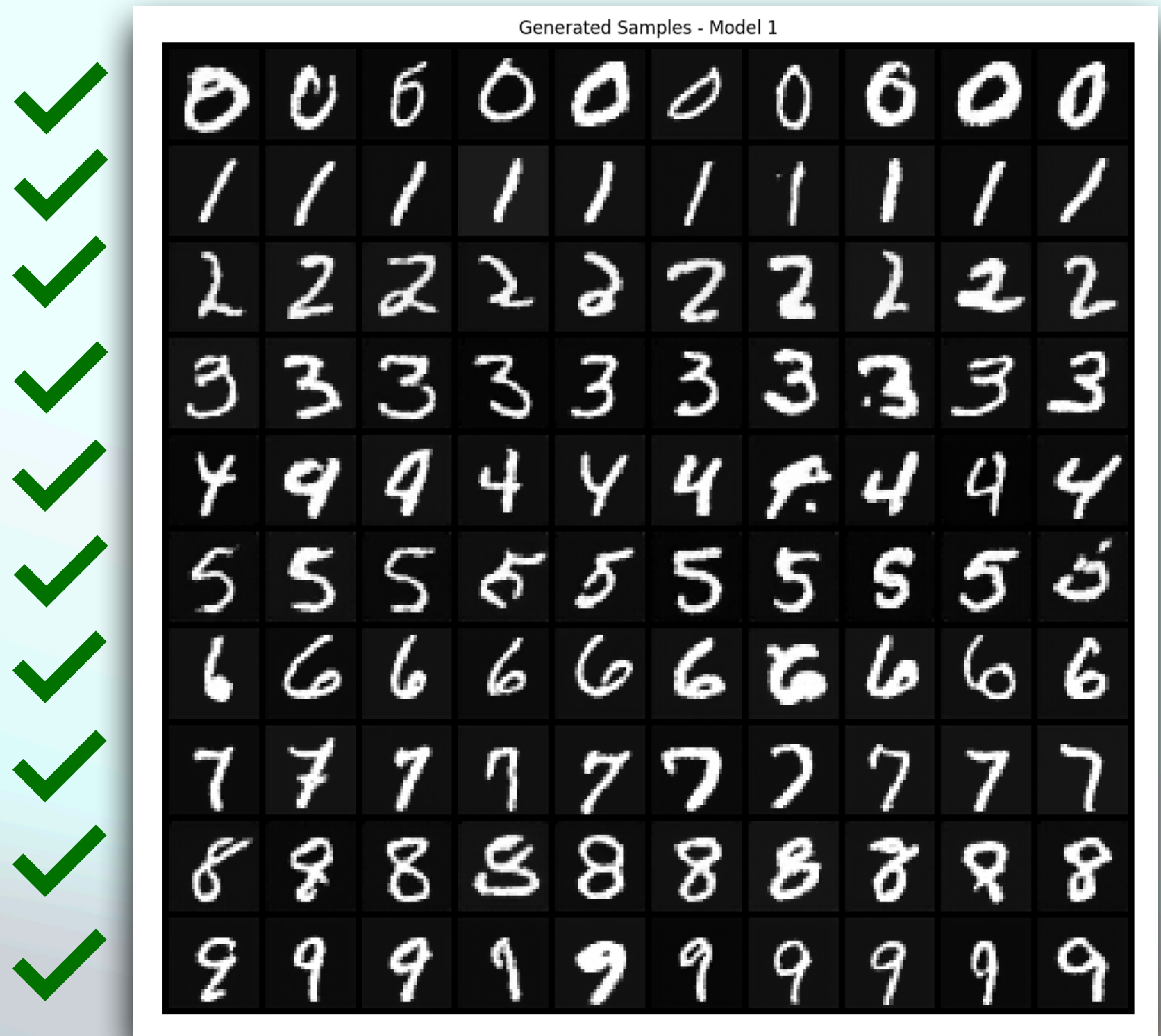
Personalization Alleviates Heterogeneity Issues

Minority Classes: 2,3,4,5,8

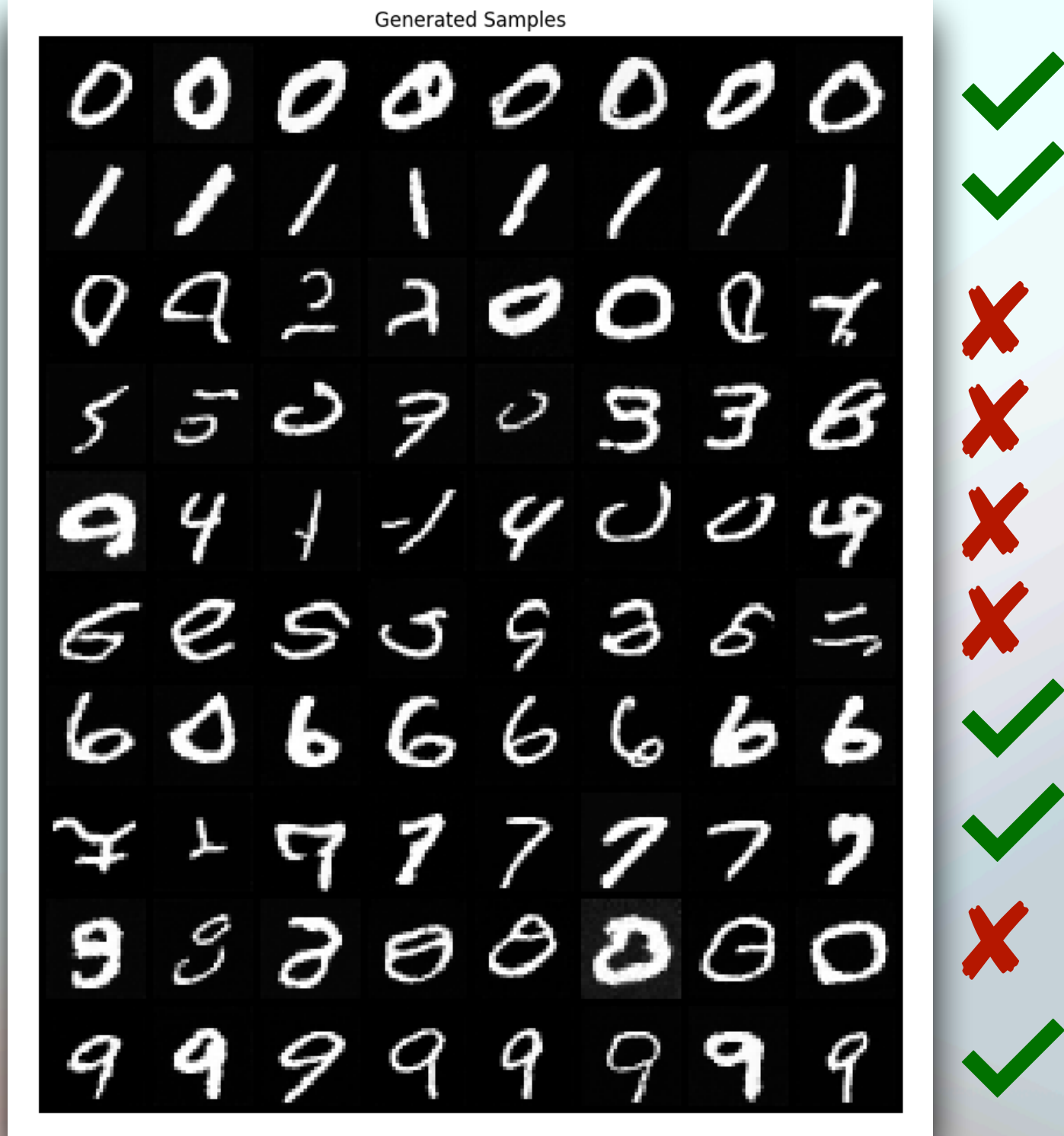


Personalization Alleviates Heterogeneity Issues

Minority Classes: 2,3,4,5,8



Personalized Training



Pure Local Training

Personalization Alleviates Heterogeneity Issues

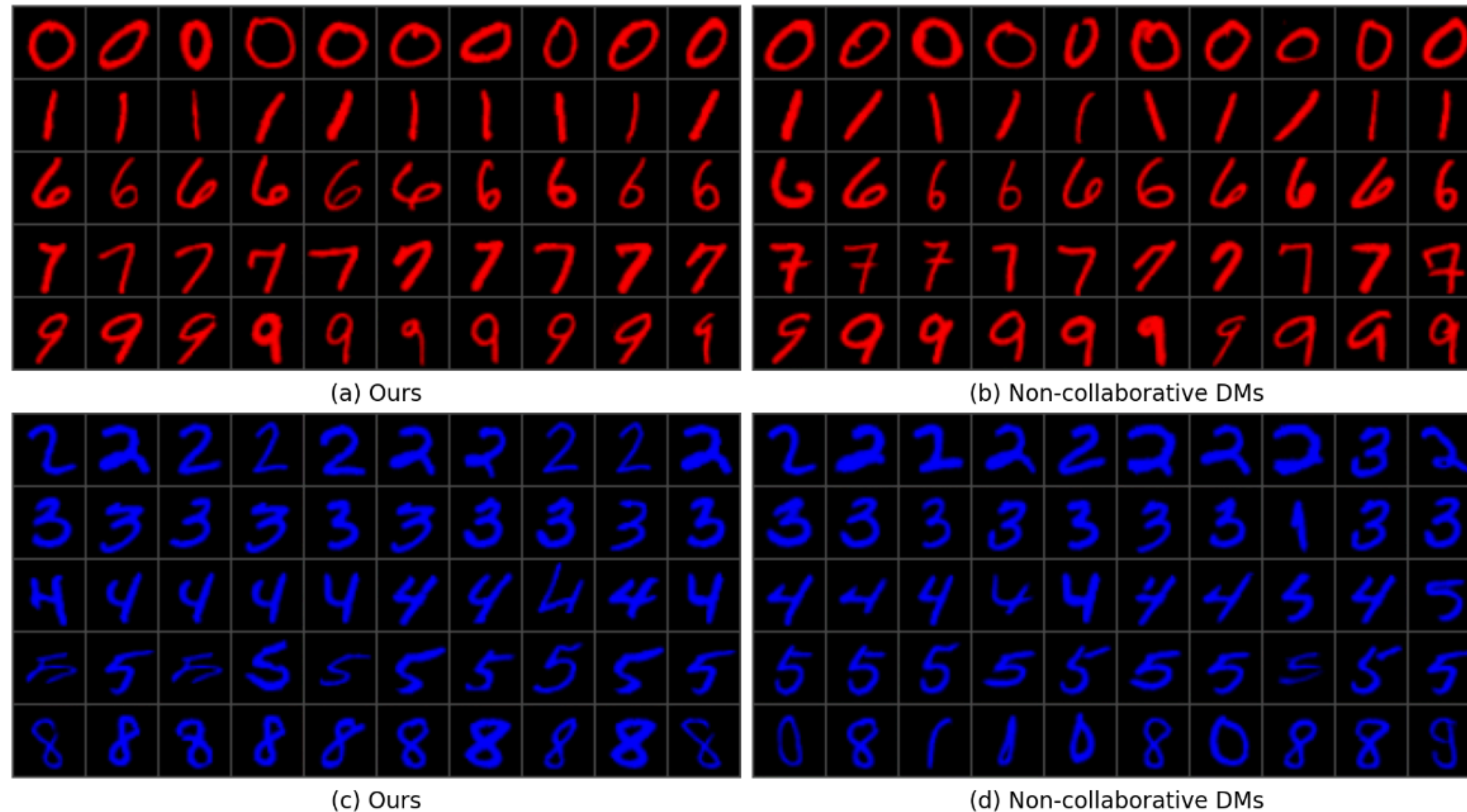
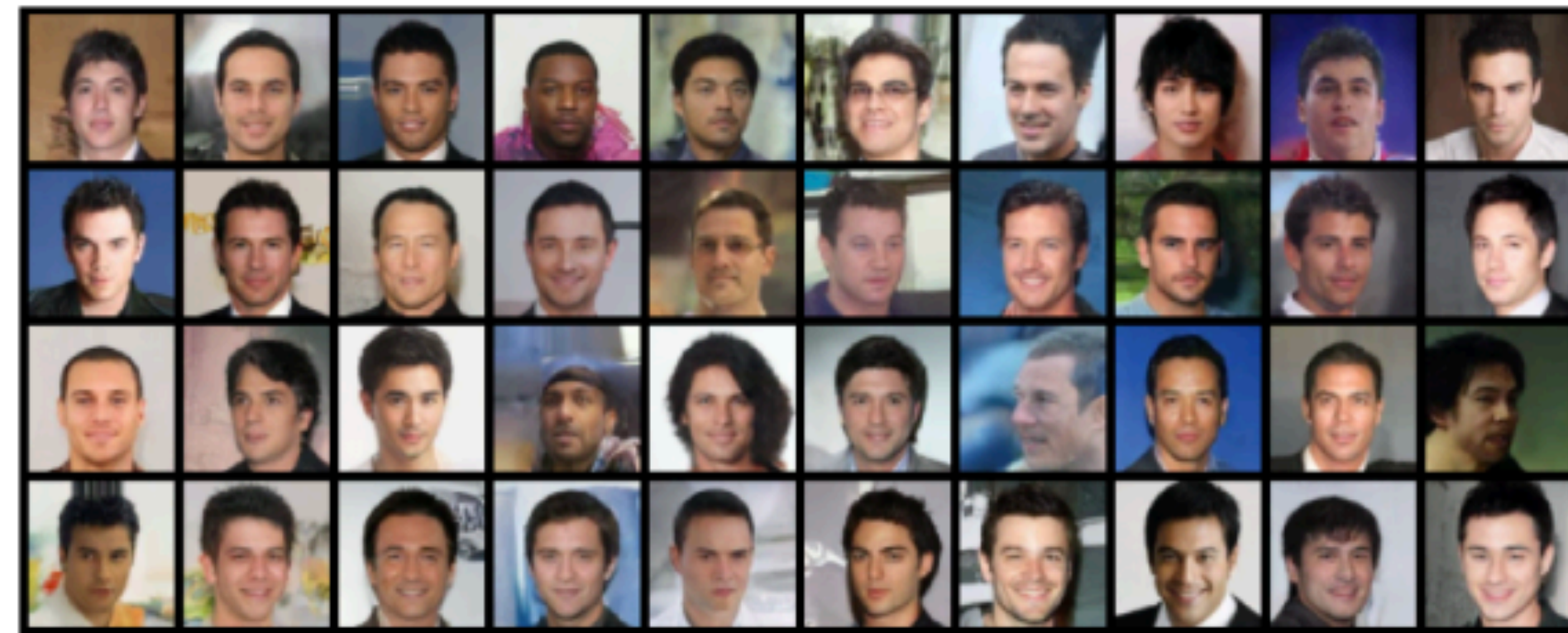


Figure 6. Colorized MNIST samples generated by different methods. (a), (b) correspond to samples generated for majority class group: red, 0,1,6,7,9. (c), (d) correspond to samples generated for minority class group: blue, 2,3,4,5,8. We report the results for the model trained by the first client.

Personalization Alleviates Heterogeneity Issues



(a) Ours



(b) Non-collaborative DMs



(c) Ours



(d) Non-collaborative DMs

Figure 1. CelebA samples generated by different methods. (a), (b) correspond to samples generated for majority class group: Black Hair and Male. (c), (d) correspond to samples generated for minority class group: NonBlack Hair and Female. We report the results for the model trained by the first client.

Personalization Alleviates Heterogeneity Issues

	CIFAR-10	Colorized MNIST	CelebA
METHODS	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)
Non-private (Centralized)	16.27/17.62/16.95	1.85/1.45/1.66	13.72/11.70/12.71
Non-private (FedDM)	18.0/19.1/18.6	1.9/1.5/1.7	14.5/11.8/13.1
(Central DP baseline) DP-LDM	44.0/36.5/38.4		
Non-collaborative	19.87/36.44/28.16	2.19/5.99/4.09	23.42/41.38/32.40
Ours	19.85/35.78/27.82	1.72/4.79/3.26	18.11/28.09/23.10

This is in fact
pre-trained
on ImageNet

Personalization Alleviates Heterogeneity Issues

This is in fact
pre-trained
on ImageNet

	CIFAR-10	Colorized MNIST	CelebA
METHODS	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)
Non-private (Centralized)	16.27/17.62/16.95	1.85/1.45/1.66	13.72/11.70/12.71
Non-private (FedDM)	18.0/19.1/18.6	1.9/1.5/1.7	14.5/11.8/13.1
(Central DP baseline) DP-LDM	44.0/36.5/38.4		
Non-collaborative	19.87/36.44/28.16	2.19/5.99/4.09	23.42/41.38/32.40
Ours	19.85/35.78/27.82	1.72/4.79/3.26	18.11/28.09/23.10

Personalization Alleviates Heterogeneity Issues

	CIFAR-10	Colorized MNIST	CelebA
METHODS	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)
Non-private (Centralized)	16.27/17.62/16.95	1.85/1.45/1.66	13.72/11.70/12.71
Non-private (FedDM)	18.0/19.1/18.6	1.9/1.5/1.7	14.5/11.8/13.1
(Central DP baseline) DP-LDM	44.0/36.5/38.4		
Non-collaborative	19.87/36.44/28.16	2.19/5.99/4.09	23.42/41.38/32.40
Ours	19.85/35.78/27.82	1.72/4.79/3.26	18.11/28.09/23.10

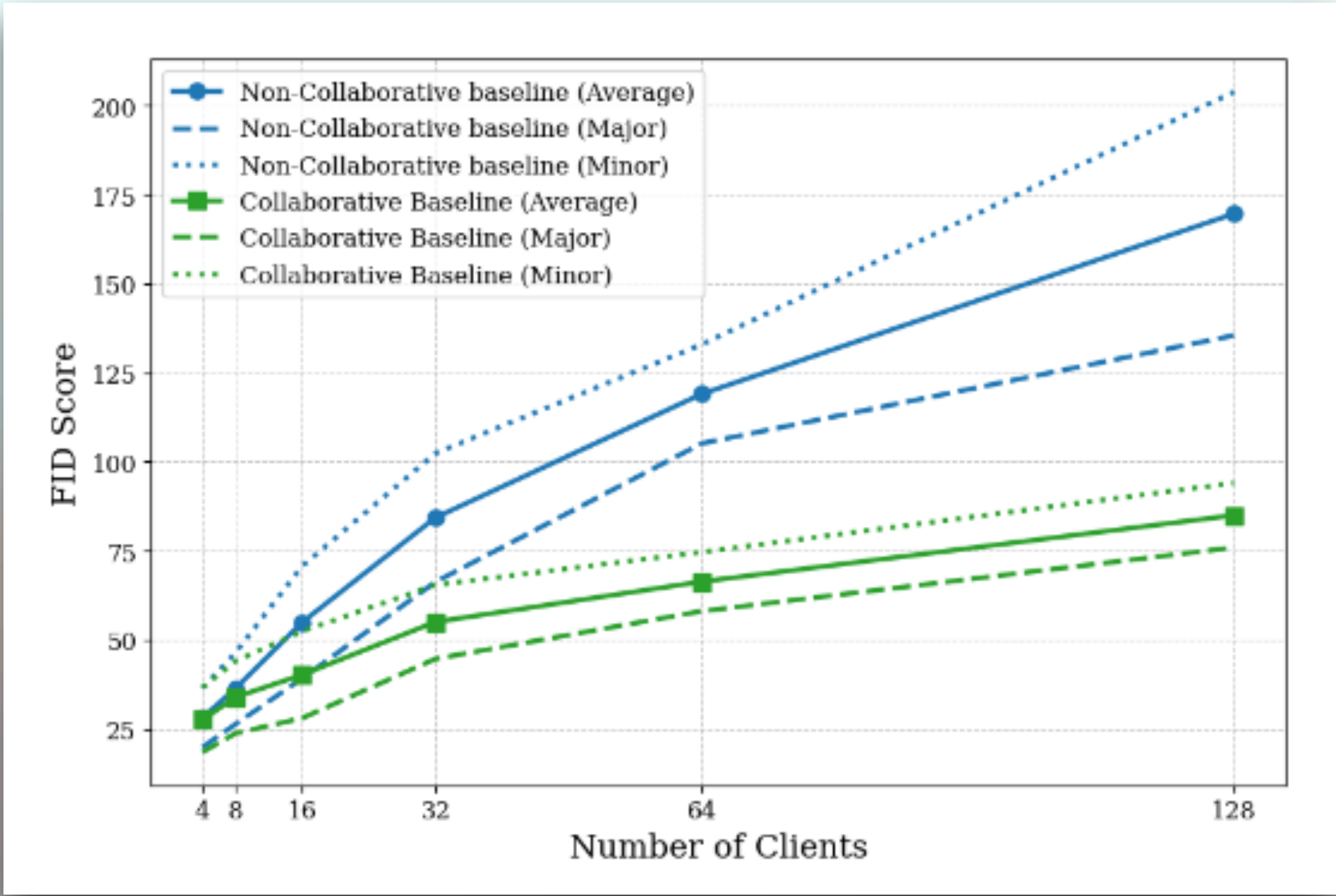
This is in fact
pre-trained
on ImageNet

Personalization Alleviates Heterogeneity Issues

	CIFAR-10	Colorized MNIST	CelebA
METHODS	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)	FID (Major/Minor/Avg)
Non-private (Centralized)	16.27/17.62/16.95	1.85/1.45/1.66	13.72/11.70/12.71
Non-private (FedDM)	18.0/19.1/18.6	1.9/1.5/1.7	14.5/11.8/13.1
(Central DP baseline) DP-LDM	44.0/36.5/38.4		
Non-collaborative	19.87/36.44/28.16	2.19/5.99/4.09	23.42/41.38/32.40
Ours	19.85/35.78/27.82	1.72/4.79/3.26	18.11/28.09/23.10

This is in fact pre-trained on ImageNet

The benefit of collaboration increases as we increase the number of clients for CIFAR-10.



Performance on a Downstream Task

Dataset	Our		Non-collaborative	
	CNN	ResNet18	CNN	ResNet18
CIFAR10	57.4	63.6	55.4	61.8
CelebA	69.9	69.0	49.2	49.1

Accuracy for a classification task using models trained using our synthetic data

Theoretical utility guarantee: We also have utility guarantees, for a GMM estimation task using a piece-wise linear model, where we show it is always individually rational to participate in the collaboration.