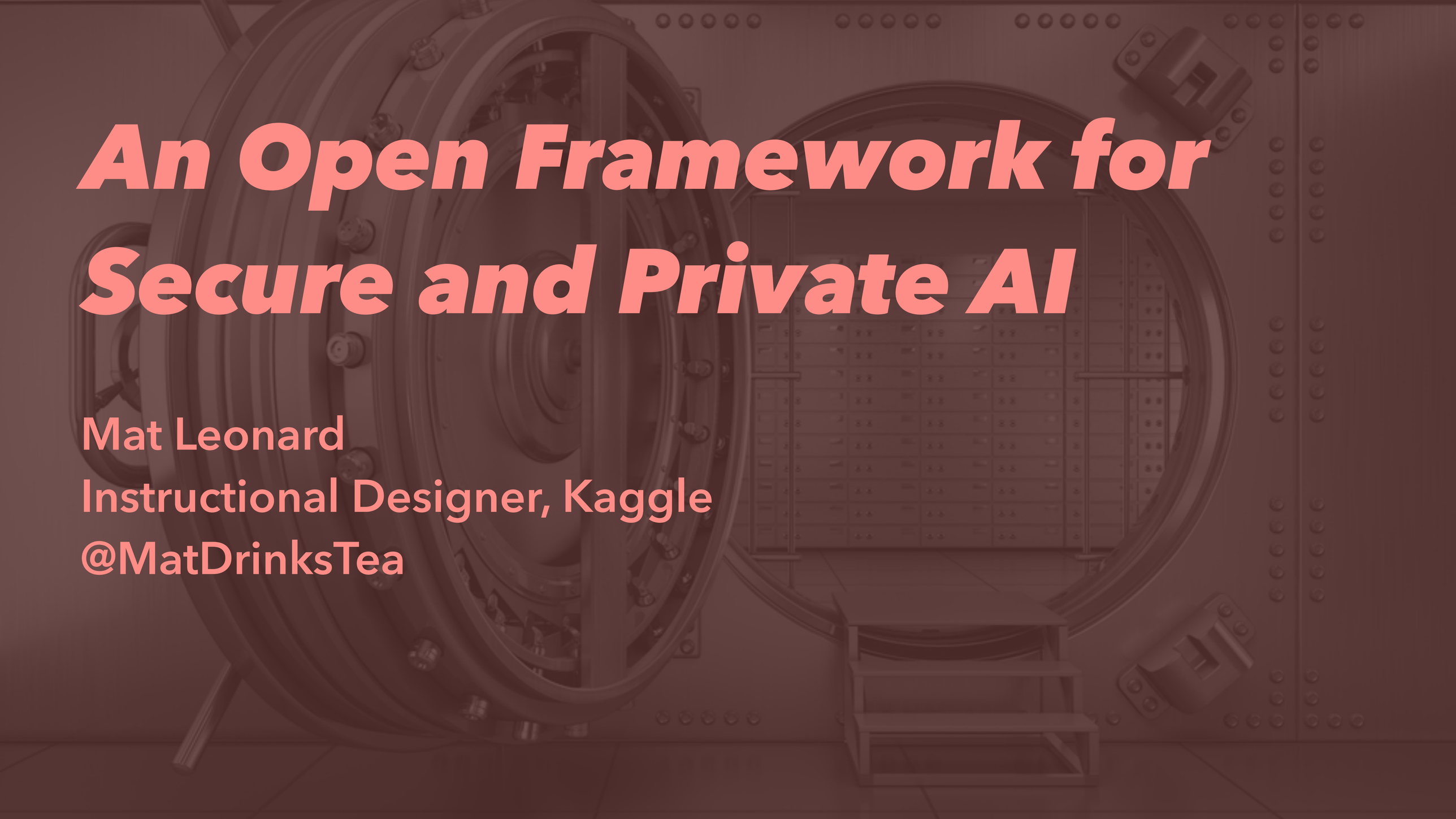
# OPEN DATA SCIENCE CONFERENCE

#ODSC

@ODSC

Boston | April 30 - May 4, 2019



# An Open Framework for Secure and Private AI

Mat Leonard
Instructional Designer, Kaggle
@MatDrinksTea

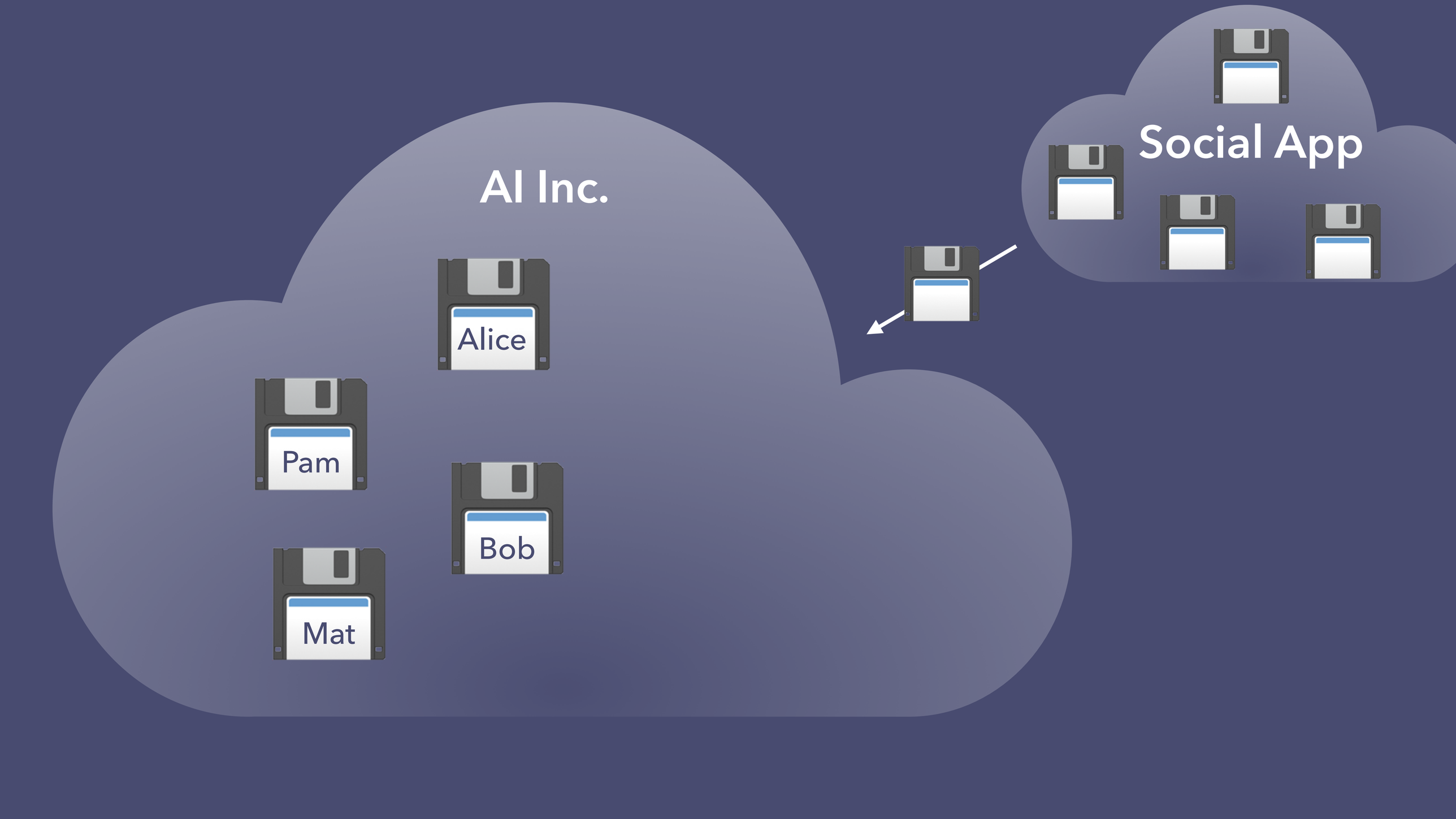
# Maintaining Privacy is an Ethical Obligation

# Maintaining Privacy is Good for Business

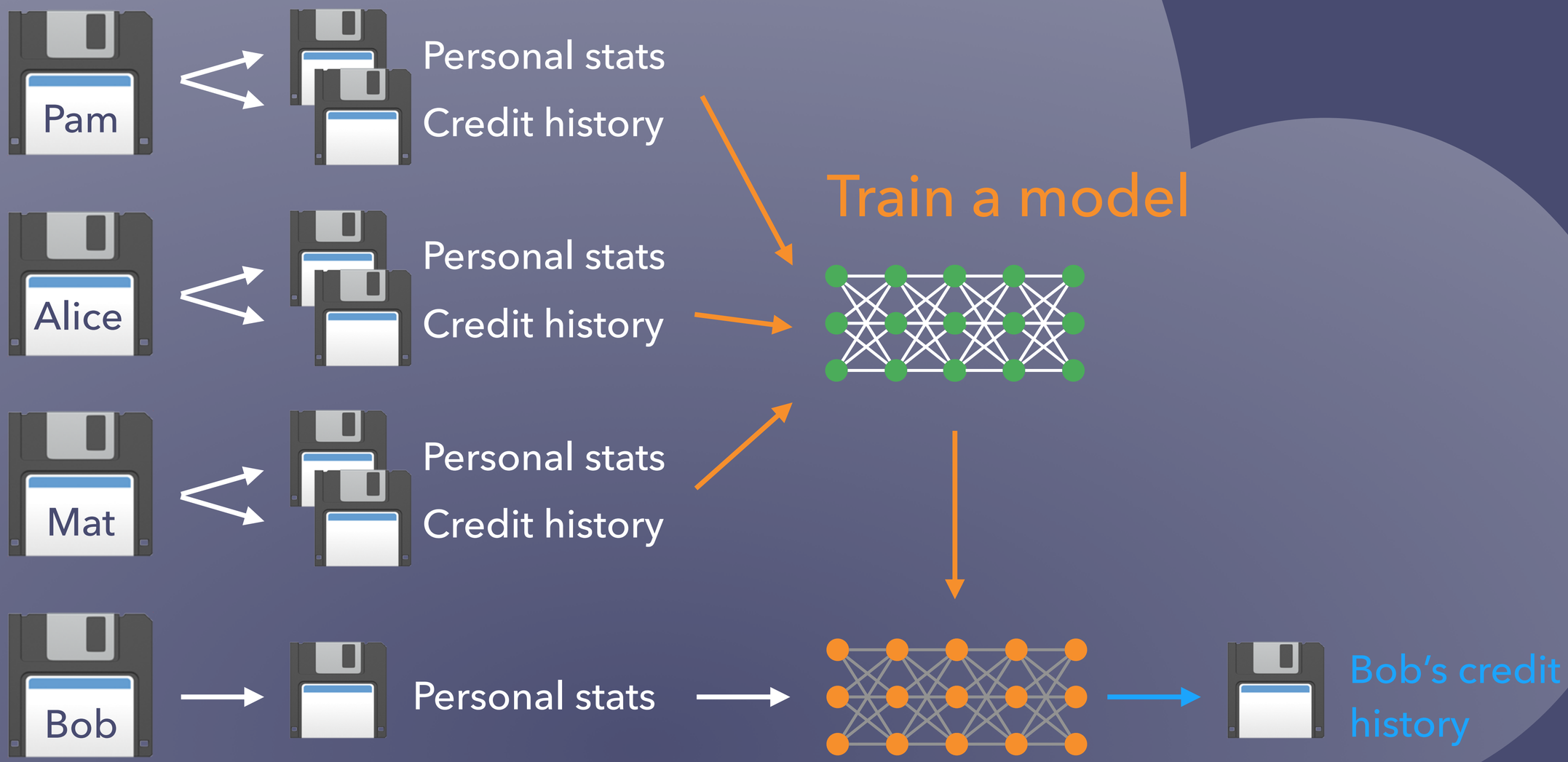
# *What Does Privacy Do For Us?*

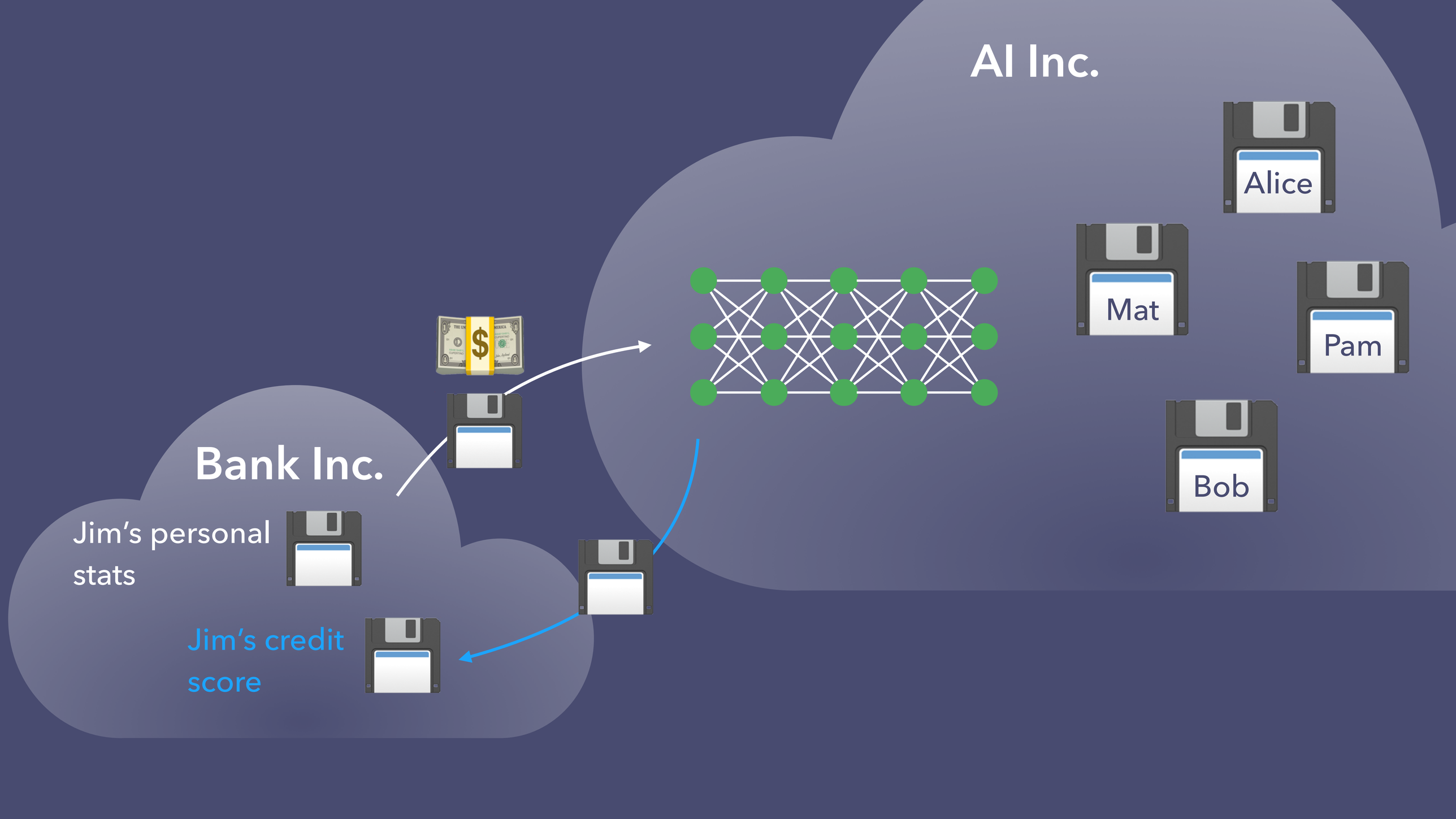
1. Builds user trust
2. More data with better quality
3. Innovative products
4. Compliance with GDPR and other regulations

# AI Inc.


AI Inc.
Alice
Pam
Bob
Mat
Social App

# AI Inc.


Pam
Alice
Mat
Bob
Personal stats
Credit history
Personal stats
Credit history
Personal stats
Credit history
Personal stats
Train a model
Bob's credit history


AI Inc.
Alice
Mat
Pam
Bob
Bank Inc.
Jim's personal stats
Jim's credit score

# *The AI Business Model*

1. Acquire data about people
2. Train a model that makes predictions
3. Sell the *use* of that model

# *Where this goes wrong*

1. Acquire data about people
   - **Privacy:** people lose control of their data
   - **Innovation:** “Sensitive Products” don’t get made
2. Train a model that makes predictions
   - **Privacy Loss:** models can leak private information from the training dataset
3. Sell the *use* of that model
   - **Lost Natural Income:** in practice, people are rarely compensated for their data

# How do we build machine learning products while preserving privacy?

1. Differential Privacy
2. Federated Learning
3. Multi-Party Computation

# Differential Privacy

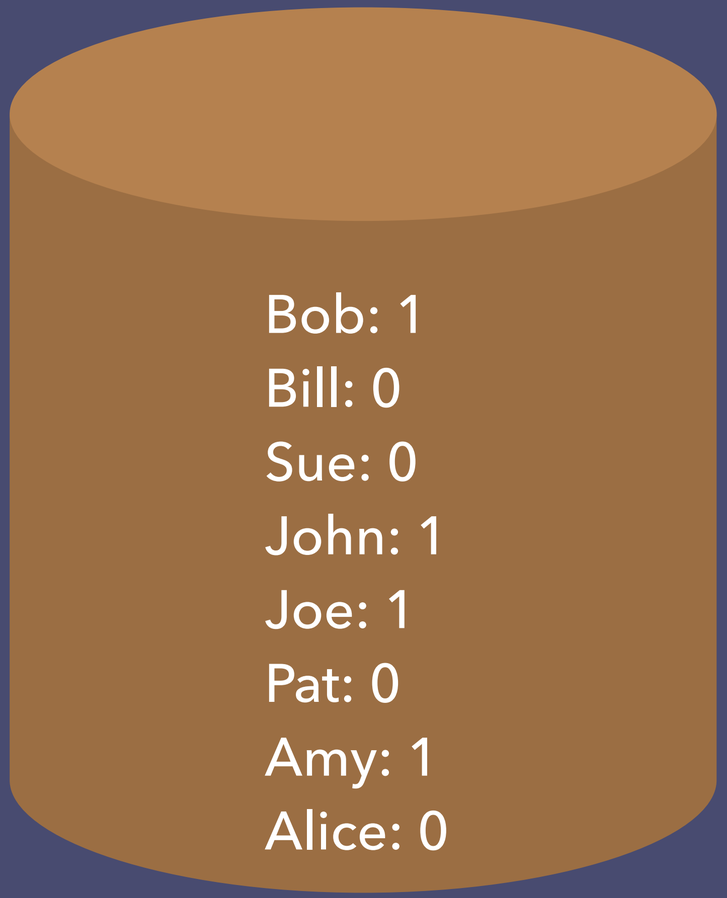
# Differential Privacy

- started with statistical database queries circa ‘03
- only more recently applied to Machine Learning
- Goal: ensure statistical analysis doesn’t compromise privacy

# *Privacy is preserved if...*

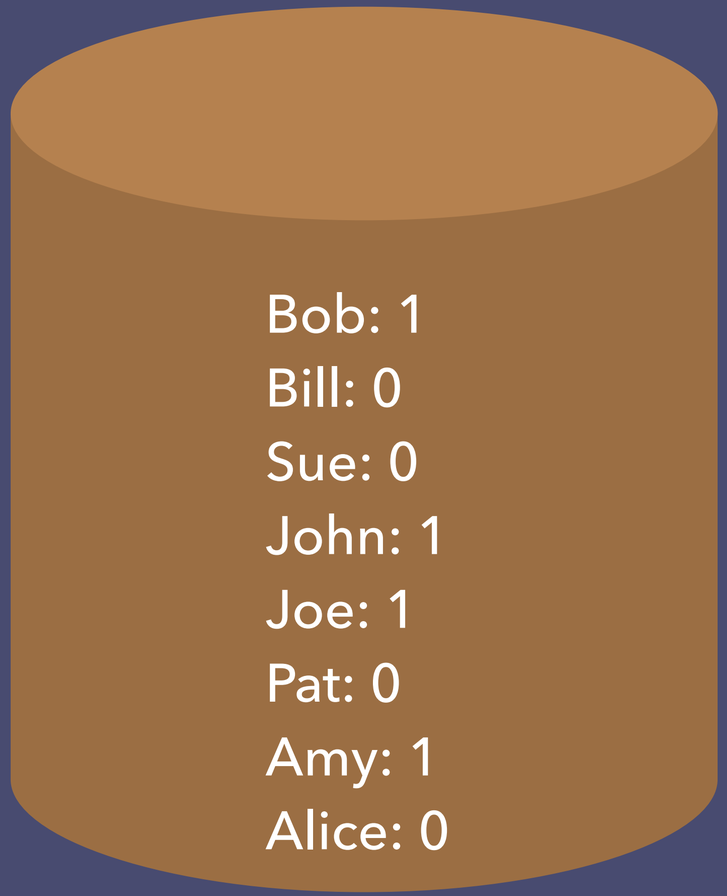
"You will **not be affected**, adversely or otherwise, by allowing your data to be used in any study or analysis, **no matter what other studies, data sets, or information sources, are available.**"

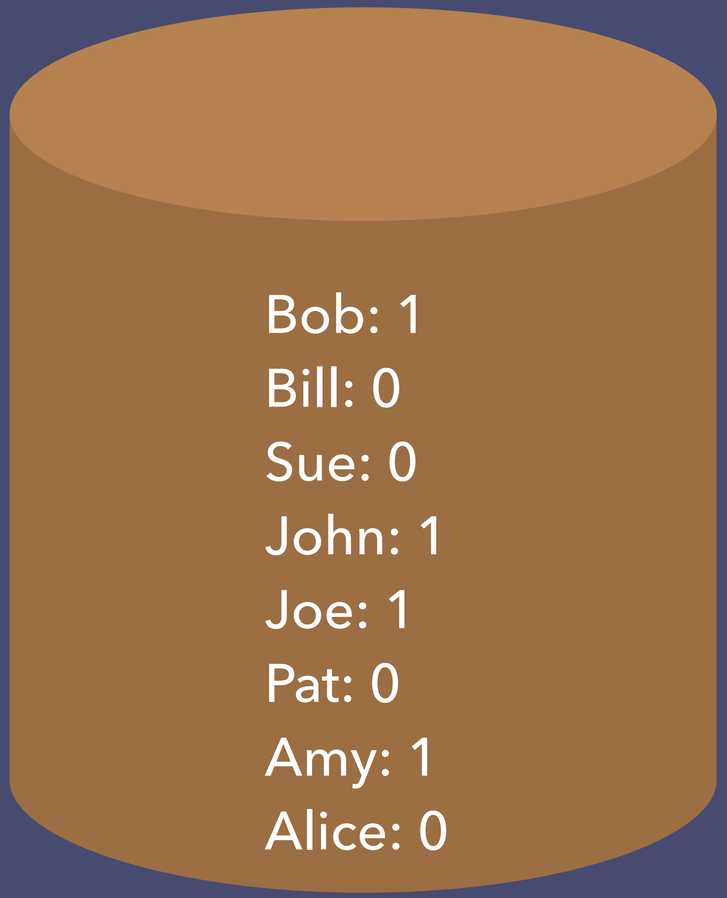
- Cynthia Dwork
  "The Algorithmic Foundations of Differential Privacy"



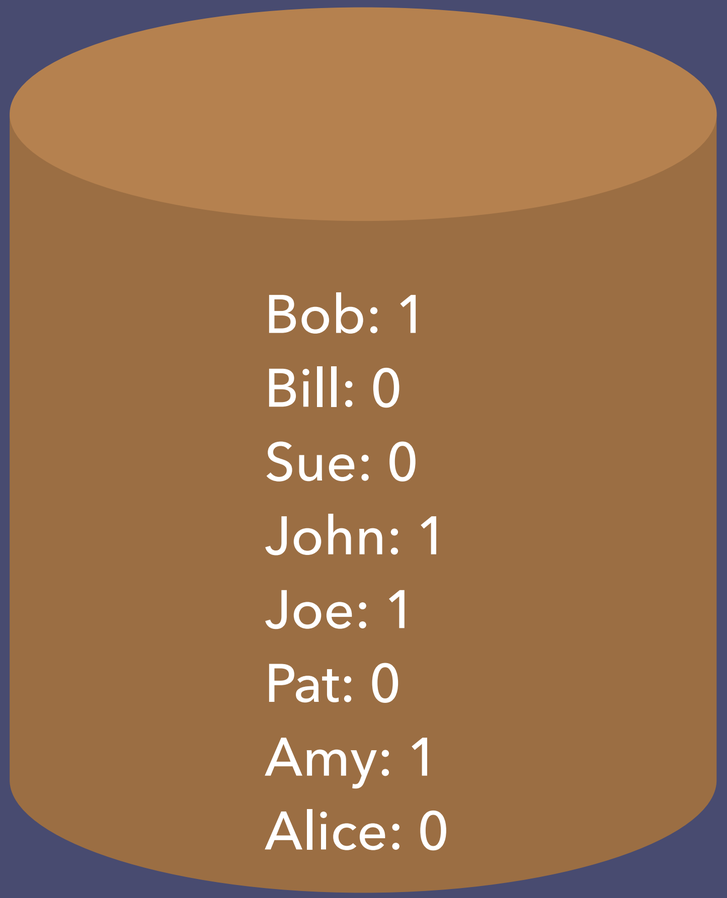

- **Query:** function(database)
- **Perfect Privacy:** the output of our query is the same between this database and any identical database with 1 person missing.
- **Exception:** you could still be affected by the results of the study even if you weren't included in the database (findings: smoking causes cancer)

# *Will your query be statistically different?*

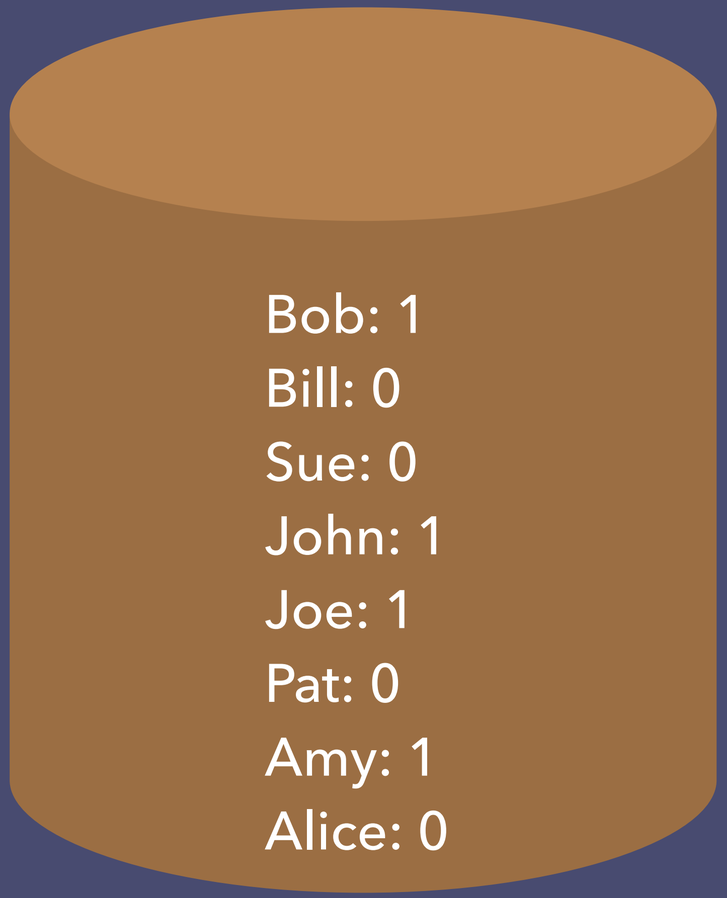


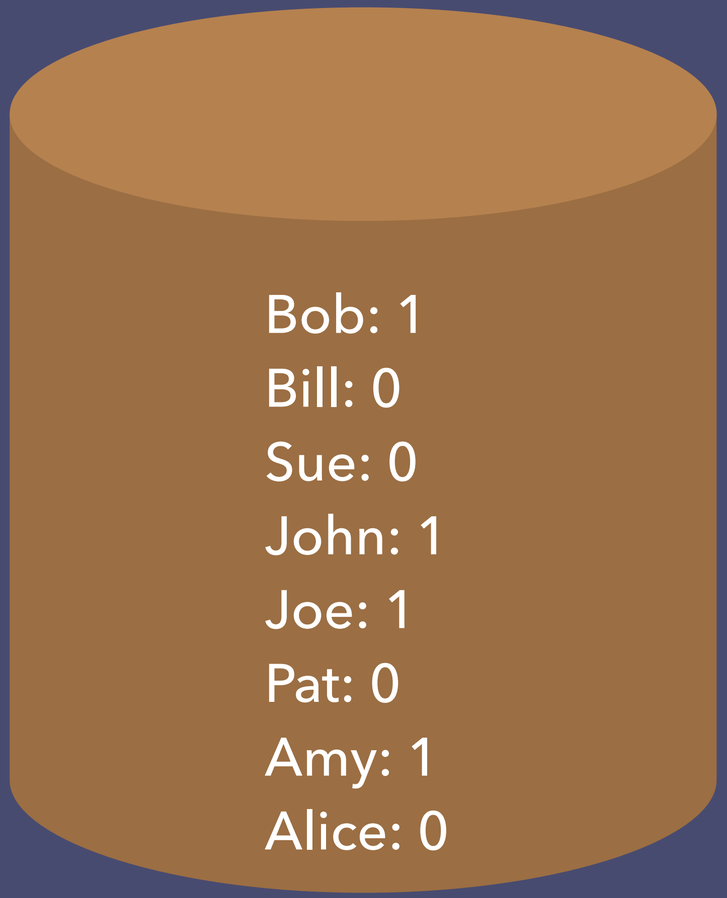

- **Query:** sum(database)
- **Perfect Privacy:** the output of our query is the same between this database and any identical database with 1 person missing.
- **Analysis:** query results are different if you remove a person from the study.




- **Query:** sum(database) + noise
- **Perfect Privacy:** the output of our query is the same between this database and any identical database with 1 person missing.
- **Analysis:** this one may work ok if the noise is great enough. How much is enough?




- **Query:** sum(database) > threshold
- **Perfect Privacy:** the output of our query is the same between this database and any identical database with 1 person missing.
- **Analysis:** this one may be perfect if the threshold is less than the max response by at least 1. Its success is dataset dependent.

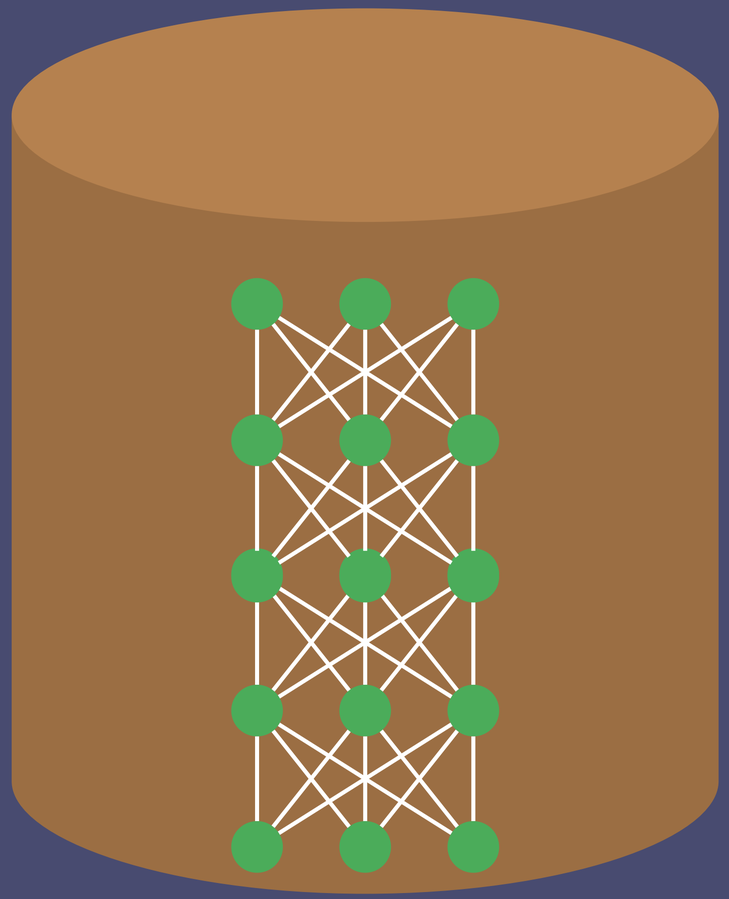



- **Query:** (sum(database) > threshold) + noise
- **Perfect Privacy:** the output of our query is the same between this database and any identical database with 1 person missing.
- **Analysis:** this one is likely to do well given that even if the threshold is equal to the max query, noise gives a “plausible explanation”

We will think of databases $x$ as being collections of records from a universe $\mathcal{X}$. It will often be convenient to represent databases by their histograms: $x \in \mathbb{N}^{|\mathcal{X}|}$, in which each entry $x_i$ represents the number of elements in the database $x$ of *type* $i \in \mathcal{X}$ (we abuse notation slightly, let-

**Definition 2.4** (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is $(\varepsilon, \delta)$-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$



- **Query:** model(database)
- **Perfect Privacy:** our model is identical regardless of whether a person is removed from the training dataset.
- **Best Case:** the model could still adversely affect people who weren't included in the dataset (recommendation systems)

# Options for DP in Machine Learning

- Train on the output of a differentially private query
  - Pro: removes private information
  - Con: removes lots of other information too

- Generate dataset from model - then perform DP-query on that dataset and retrain model
  - Pro: removes private information

"Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data" - Papernot et al. 2016

# *Private Aggregation of Teacher Ensembles (PATE)*

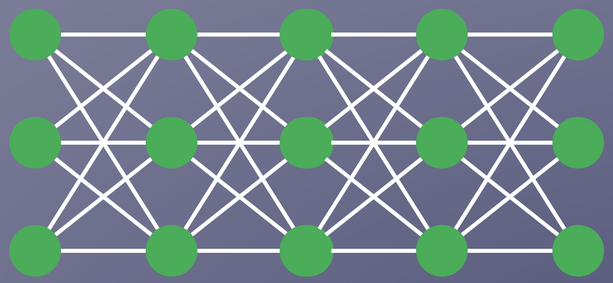
- **Step 1:** split data into buckets
- **Step 2:** train model on each bucket
- **Step 3:** predict models on public, unlabeled dataset
  - Sum the predictions over labels, take the max label, add noise
- **Step 4:** Train final model on public dataset with generated labels

"Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data" - Papernot et al. 2016

# Federated Learning

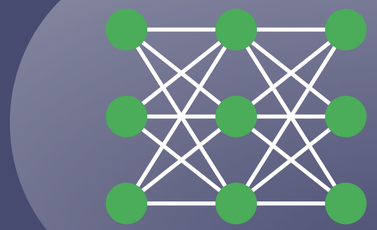
# Non-federated Learning

AI Inc.



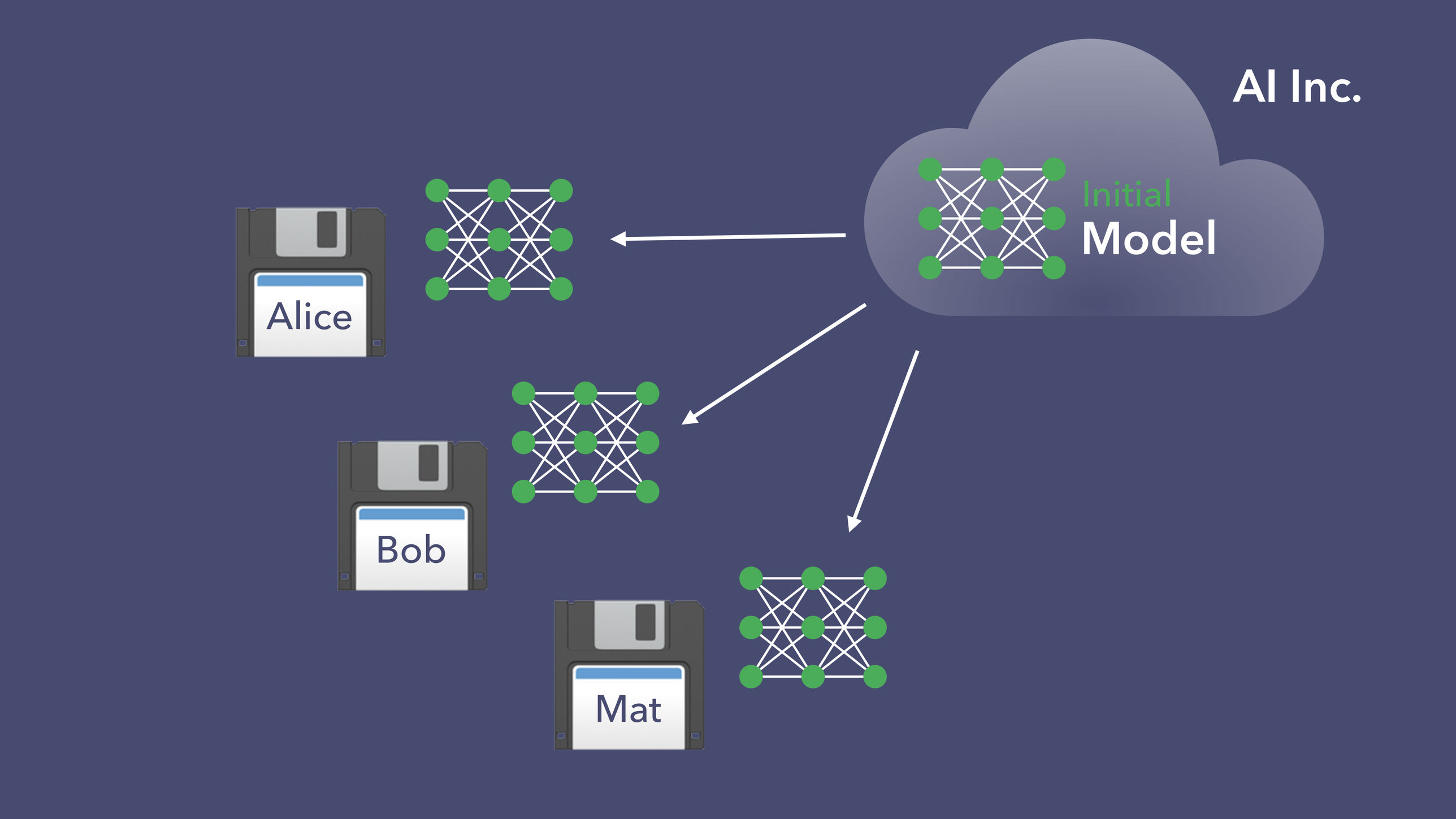



AI Inc.

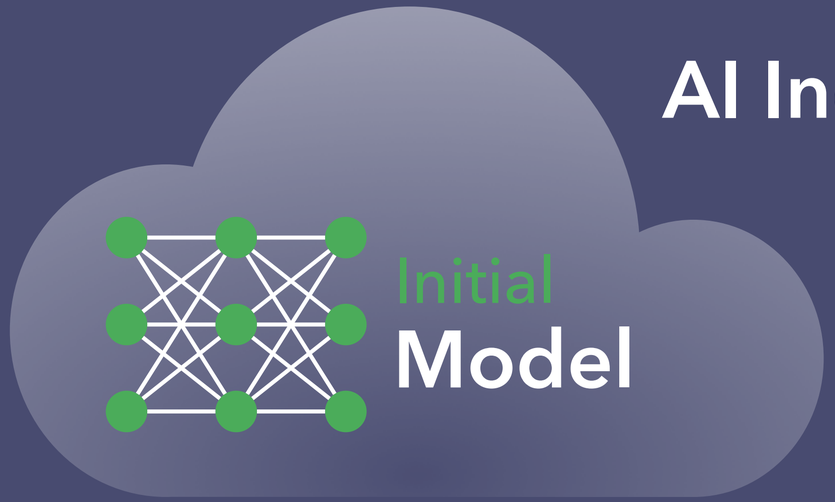


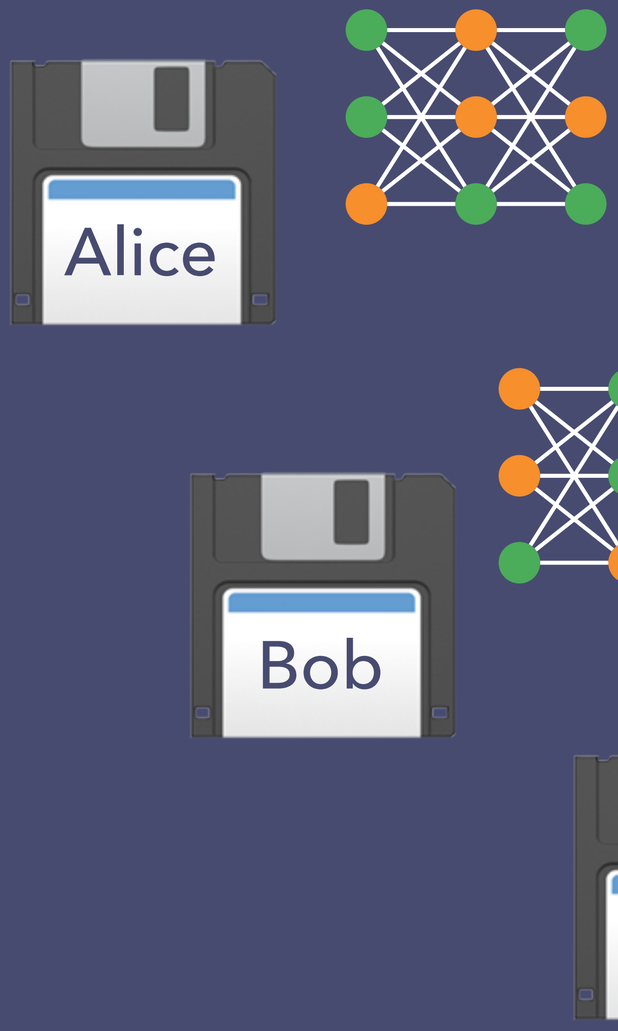
Initial
Model

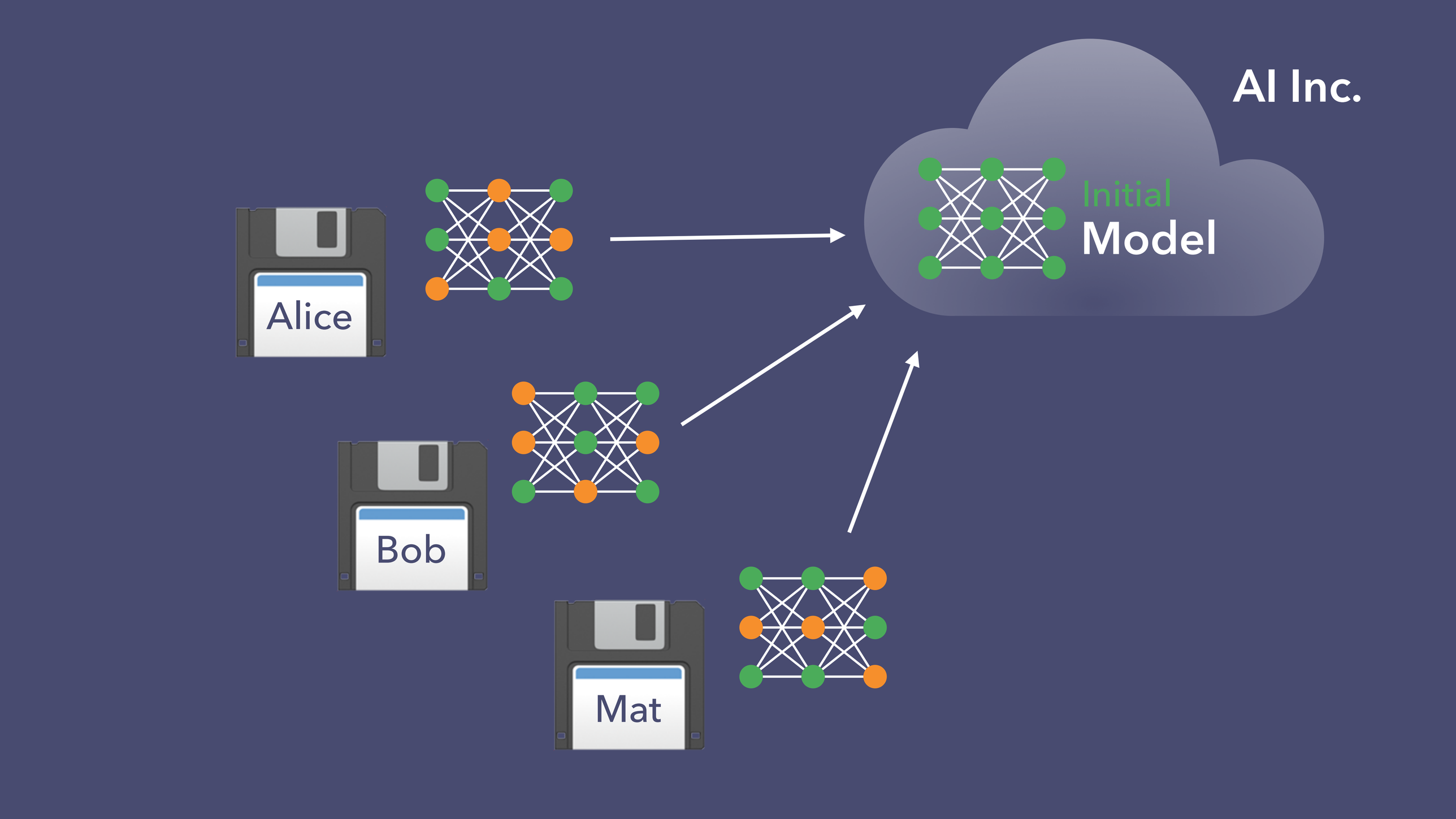



AI Inc.
Initial
Model
Alice
Bob
Mat

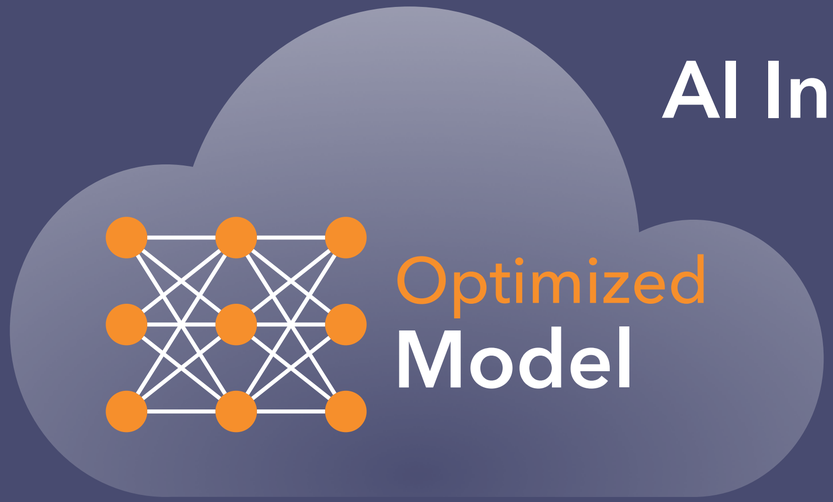
AI Inc.



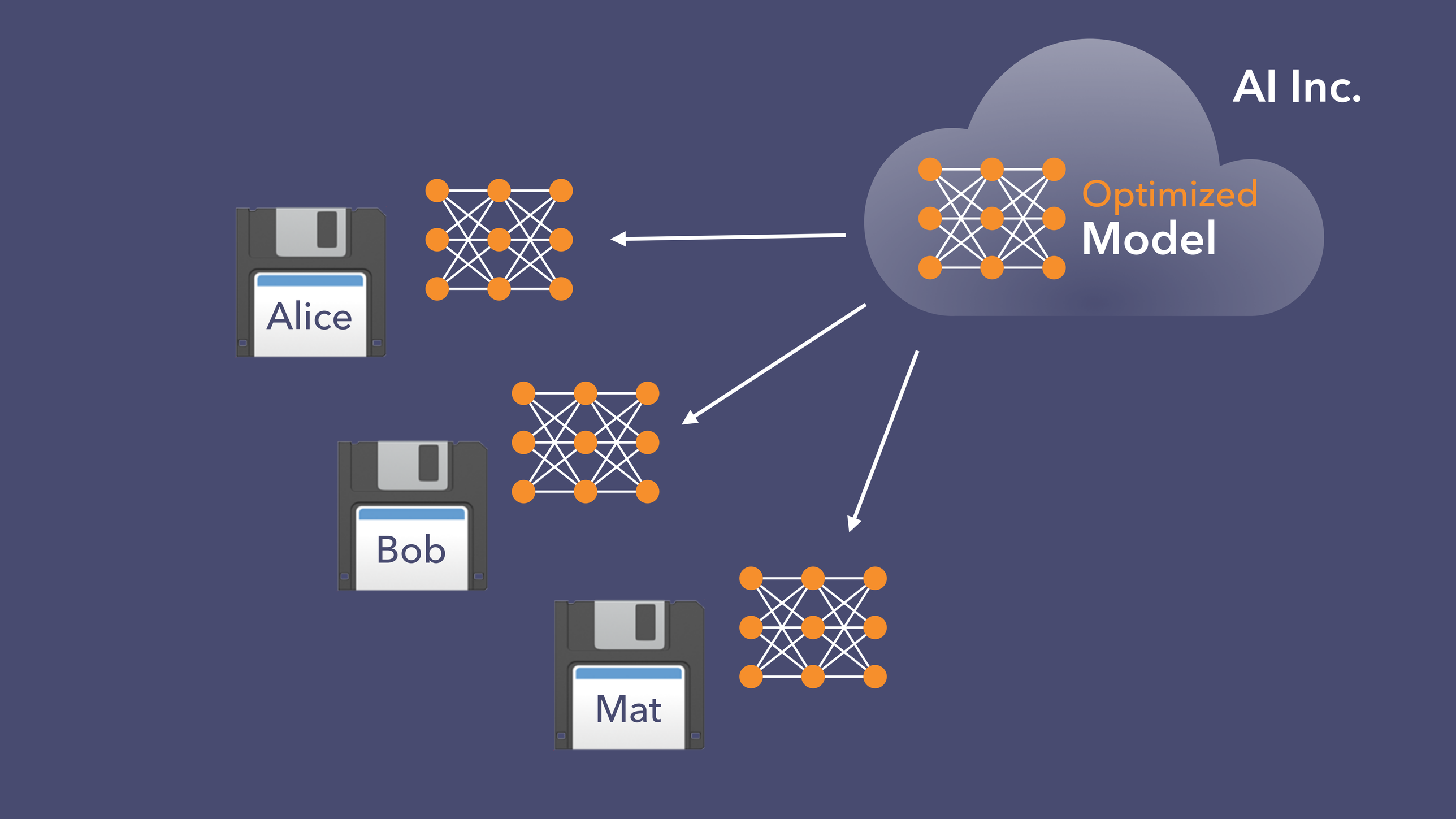




AI Inc.
Initial
Model
Alice
Bob
Mat

AI Inc.

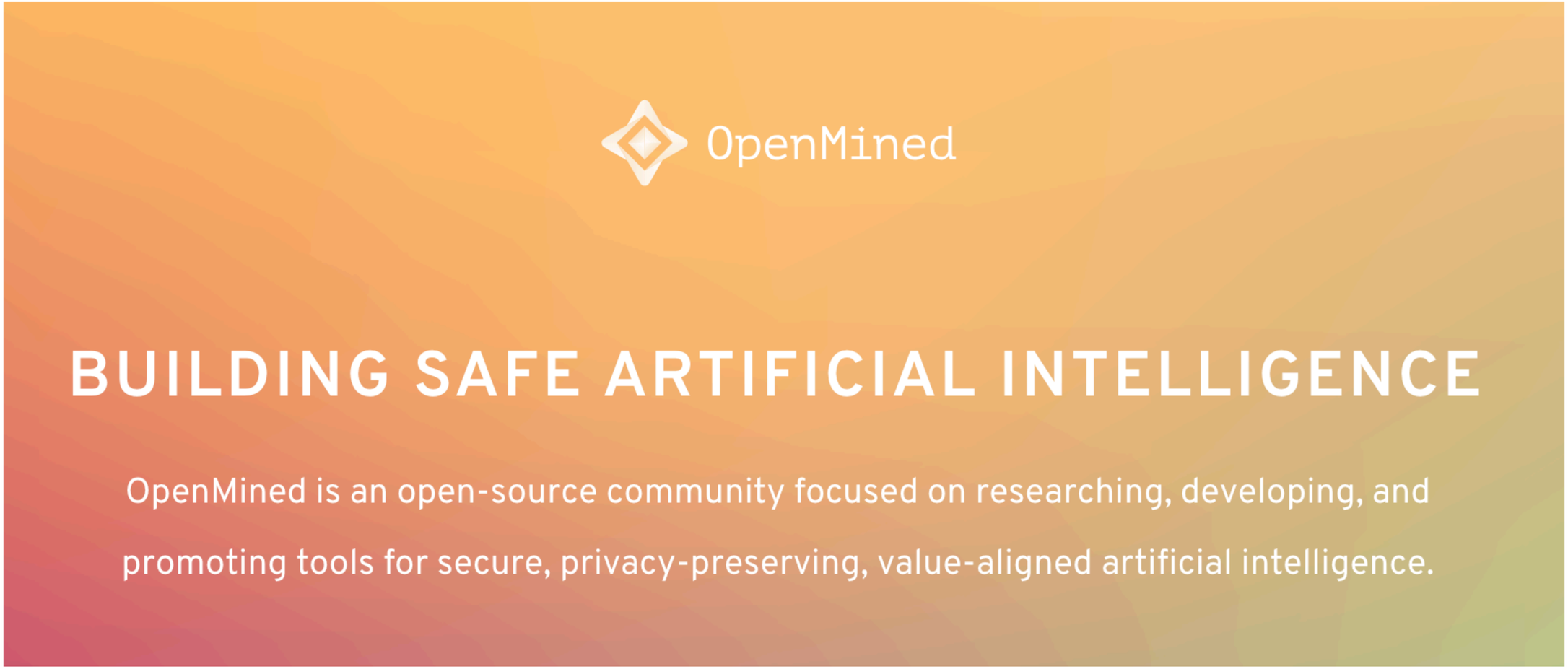






AI Inc.
Optimized
Model
Alice
Bob
Mat

# Demo!

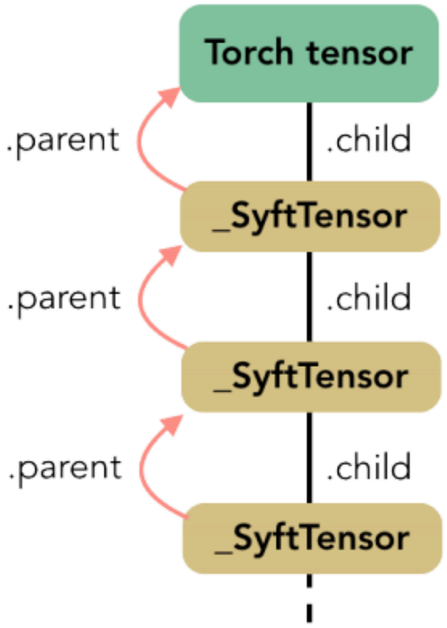
# OpenMined.org

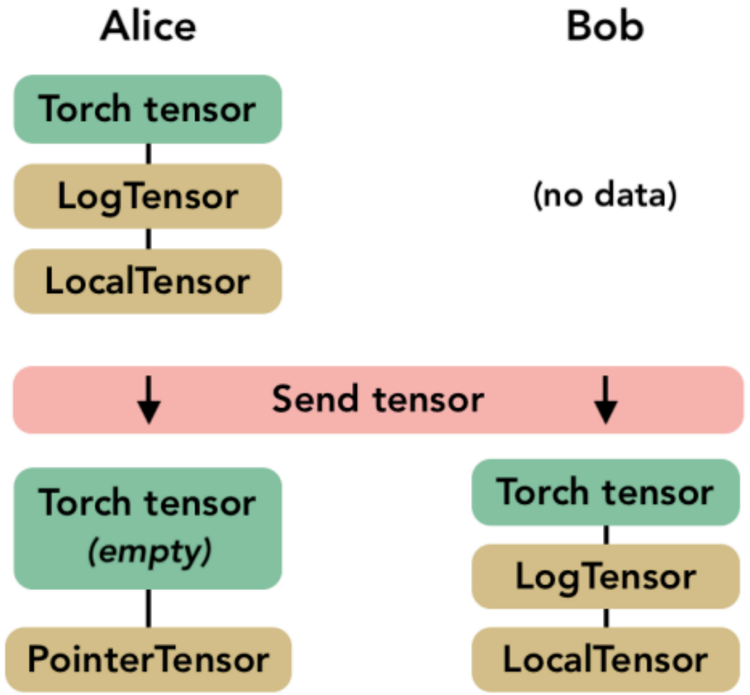



## Join the community at slack.openmined.org

# PySyft extends PyTorch

(and TensorFlow eventually)



# Demo!

# Great!

# Private data remains private

# But!

We're transmitting insecure models and gradients

# Secure Multi-party Computation

# Multi-Party Computation

# Multi-Party Computation

$a = 5$ → Share Splitter → $\text{shares}_a = [1, -3, 5, 0, 2]$

$b = 3$ → Share Splitter → $\text{shares}_b = [2, -5, 8, -3, 1]$

# Multi-Party Computation

$a = 5$ → Share Splitter → $\text{shares}_a = [1, -3, 5, 0, 2]$

$b = 3$ → Share Splitter → $\text{shares}_b = [2, -5, 8, -3, 1]$

| Alice | Bob | Chuck | Dora | Erika |
|---|---|---|---|---|
| $s_a = 1$ | $s_a = -3$ | $s_a = 5$ | $s_a = 0$ | $s_a = 2$ |
| $s_b = 2$ | $s_b = -5$ | $s_b = 8$ | $s_b = -3$ | $s_b = 1$ |
| $s_c = 3$ | $s_c = -8$ | $s_c = 13$ | $s_c = -3$ | $s_c = 3$ |

# Multi-Party Computation

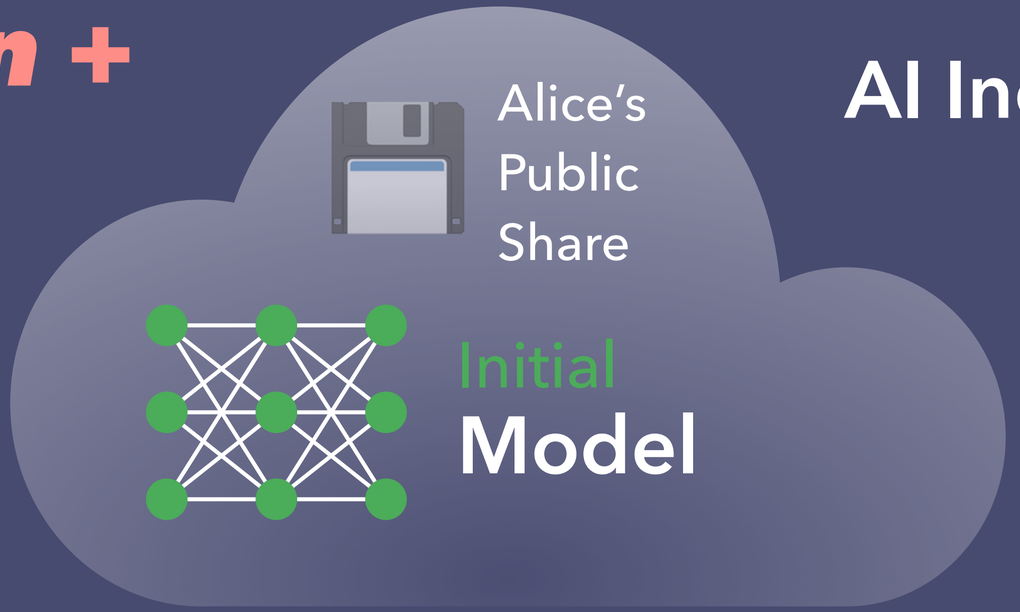
| Alice | Bob | Chuck | Dora | Erika |
|---|---|---|---|---|
| $s_a = 1$ | $s_a = -3$ | $s_a = 5$ | $s_a = 0$ | $s_a = 2$ |
| $s_b = 2$ | $s_b = -5$ | $s_b = 8$ | $s_b = -3$ | $s_b = 1$ |
| $s_c = 3$ | $s_c = -8$ | $s_c = 13$ | $s_c = -3$ | $s_c = 3$ |

$$\text{shares}_c = [3, -8, 13, -3, 3] \longrightarrow \boxed{\text{Share Combiner}} \longrightarrow c = 8$$

# Demo!

# *Multi-Party Computation + Federated Learning*

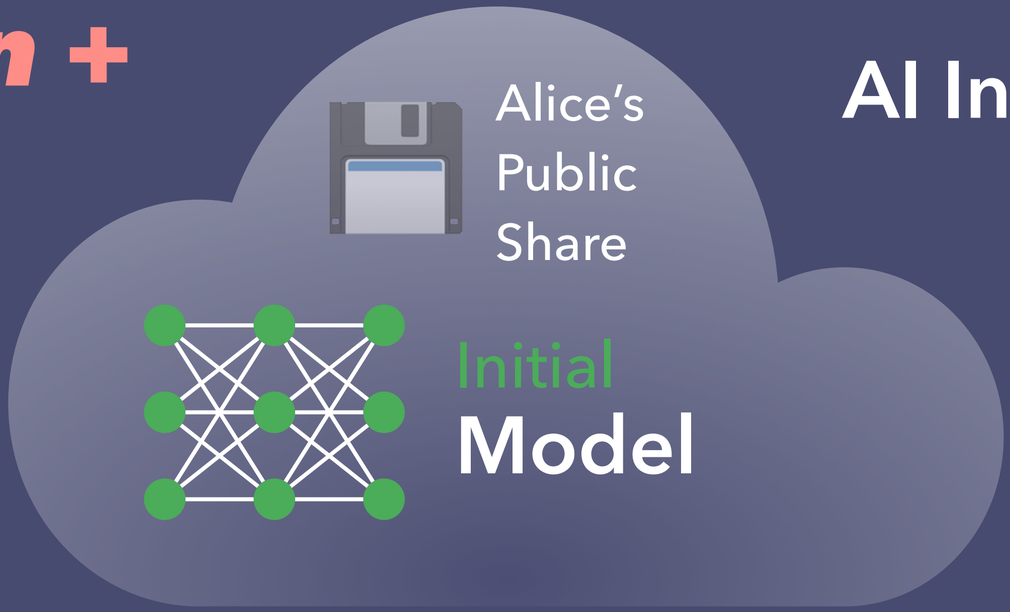
AI Inc.

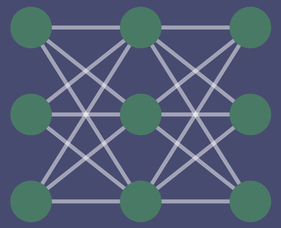





# *Multi-Party Computation + Federated Learning*

AI Inc.

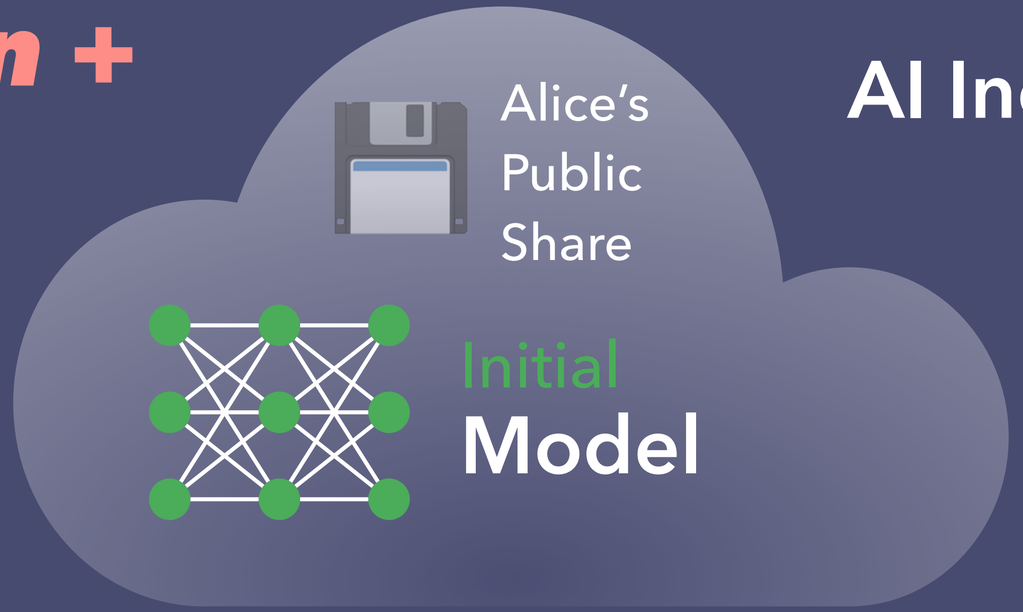


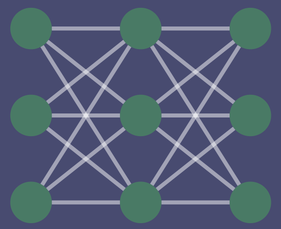



Model's
Public
Share



# *Multi-Party Computation + Federated Learning*

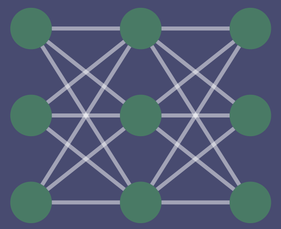


Model's
Public
Share


Training
Ensues


# *Multi-Party Computation + Federated Learning + Secure Aggregation*



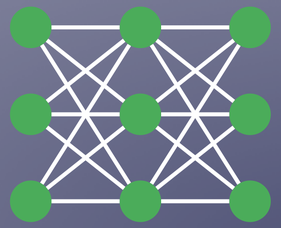
Model's Public Share


Training Ensues



AI Inc.

Alice's Public Share



Initial **Model**

# Udacity Free Course

FREE COURSE

## Secure and Private AI

by facebook Artificial Intelligence

Learn how to extend PyTorch with the tools necessary to train AI models that preserve user privacy.

NOTIFY ME

http://udacity.com/private-ai

### Course Leads

**Andrew Trask**

Leader of OpenMined, Research Scientist at DeepMind Oxford, PhD Student

# Thank you!

# Questions?