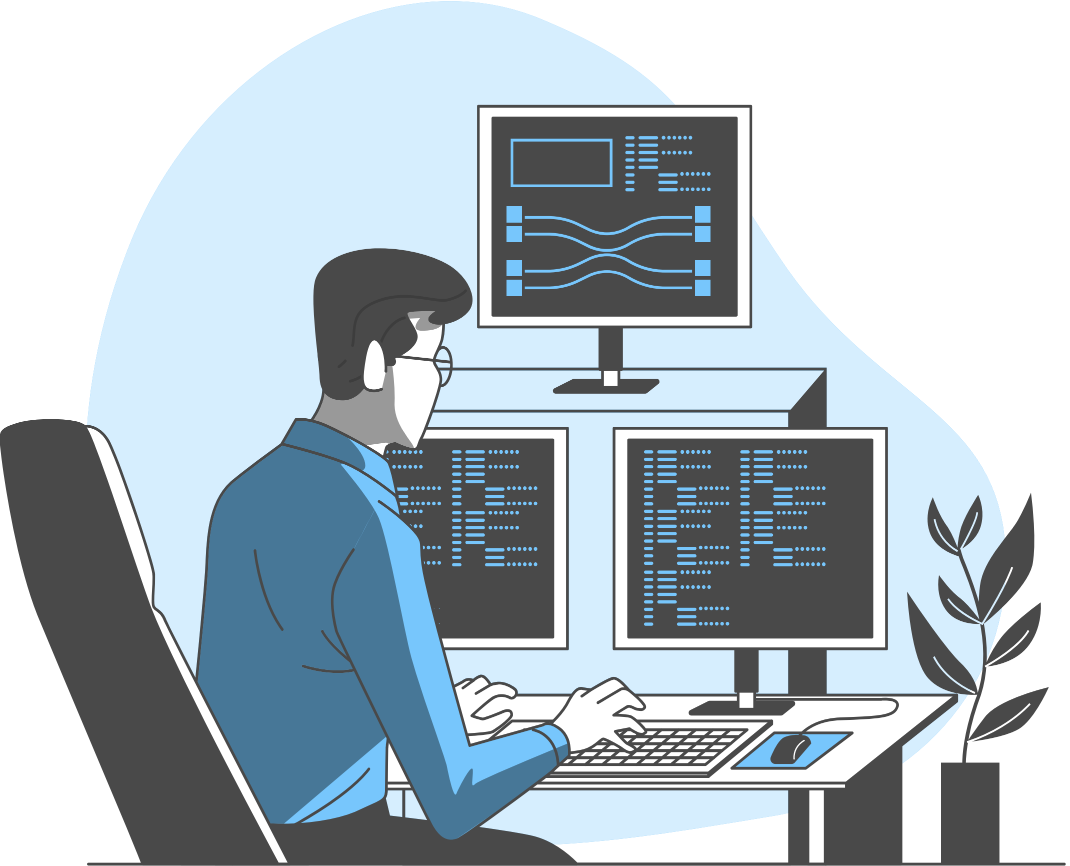


Preserving Privacy in Data Science

Faustina Maria Giaquinta

Table of Contents

01 Privacy & Methods

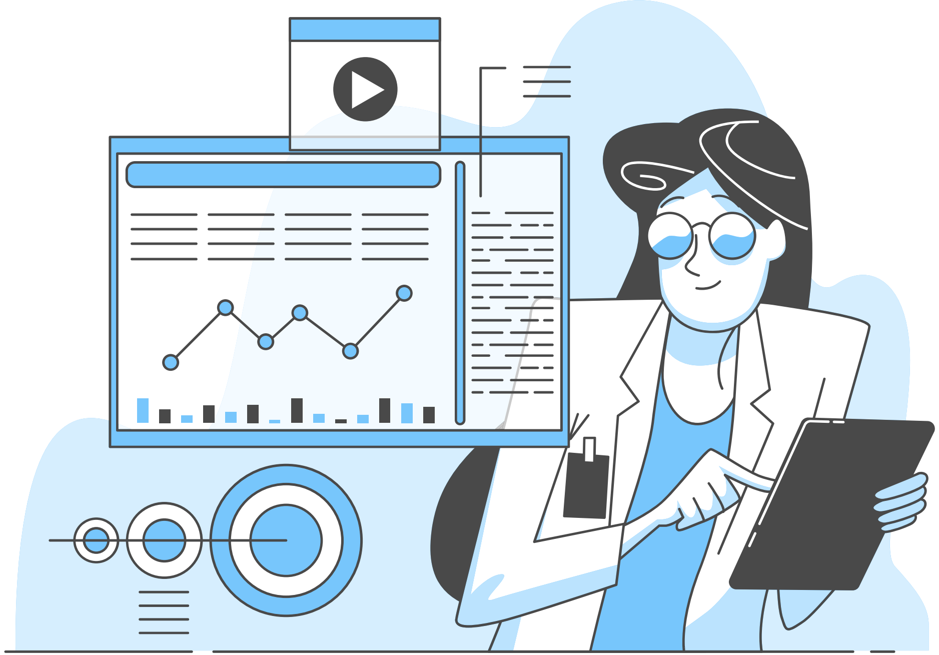
02 Data Anonymization

03 Federated Learning

04 Examples

05 Tools

06 Resources



What is Privacy?

Approaches to Protecting User Privacy

01

Protect the data

Data Anonymization
Differential Privacy

02

Build protection within the model

Federated Learning

Data Anonymization

Data processing technique that removes or modifies personally identifiable information; it results in anonymized data that cannot be associated with any one individual.

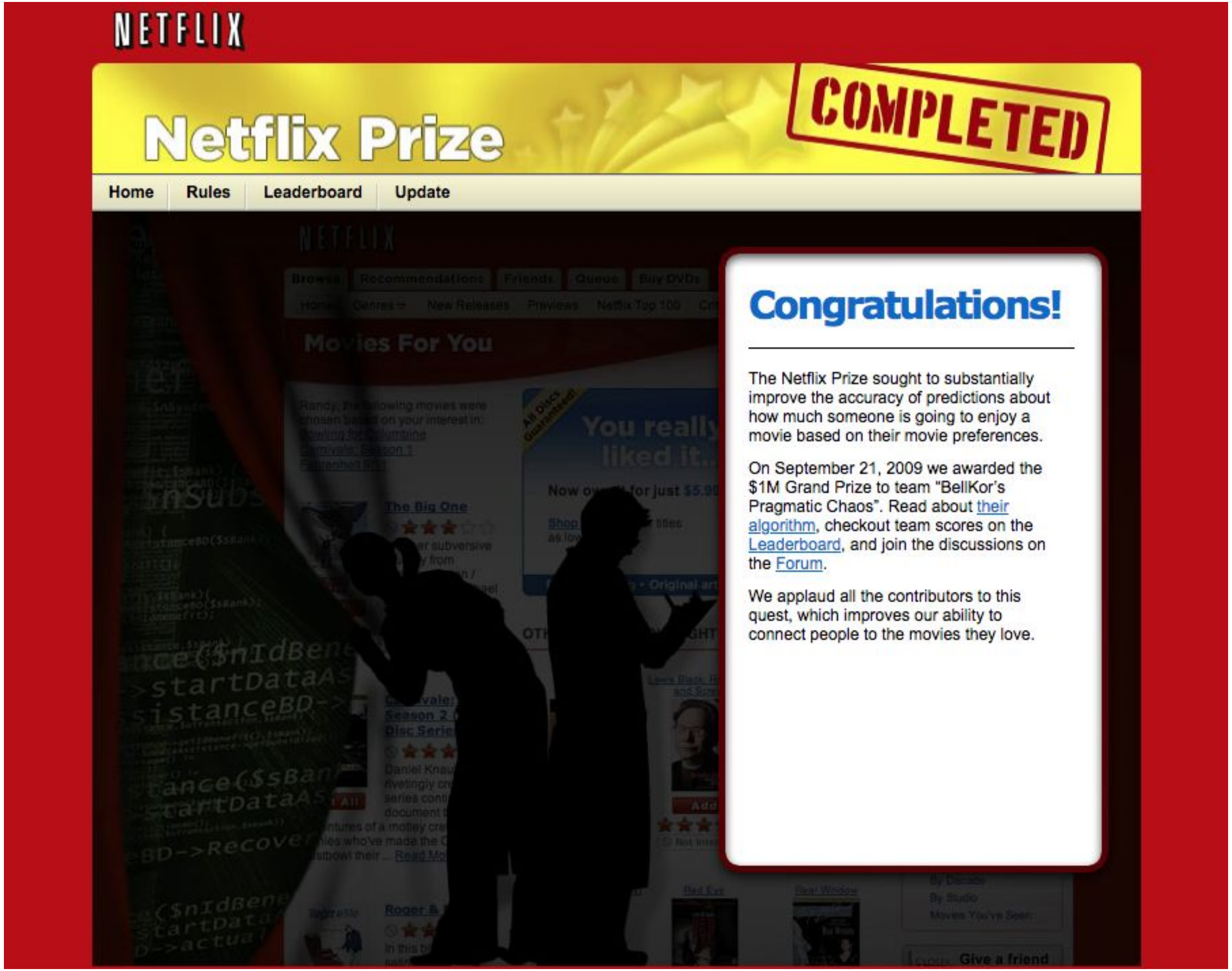
Differential Privacy

“You will not be affected, adversely or otherwise, by allowing your data to be used in any study, no matter what other studies, data sets, or information from other sources is available”

Federated Learning

Machine Learning technique for training models on data to which we don’t have access.
Type of remote execution where models are sent to remote data-holding machines (such as smartphones or IoT devices) for local training.

Data Anonymization



Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

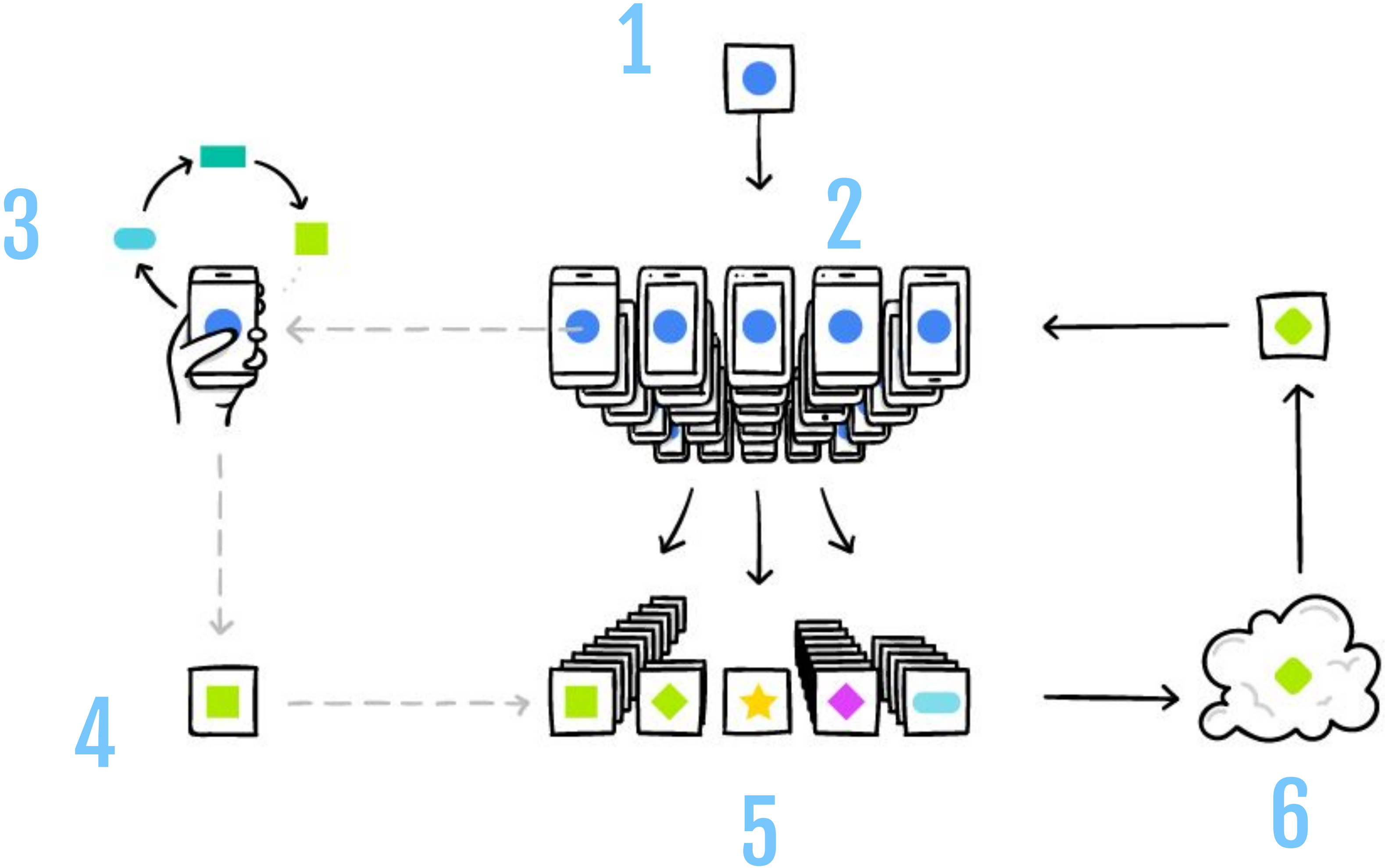
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

Federated Learning

Federated Learning
(Collaborative Learning)

Machine Learning technique for training models on data to which we don’t have access.

Type of remote execution where models are sent to remote data-holding machines (such as smartphones or IoT devices) for local training.


1
2
3
4
5
6

Procedure

1

Modeling

Main model ready to train and hosted in the cloud

2

Distribution

The cloud-hosted model is downloaded to each device

3

Training

The global model is trained on a local environment with user usage of the device

4

Upload

Trained local models are uploaded to the cloud

5

Aggregation

All trained updates are aggregated to form a consensus change

6

Update

Main model is updated with the aggregation of individual trained models

Protocols

Secure Aggregation

Interactive cryptographic protocol for computing sums of masked vectors, like model weights.
It works by coordinating the exchange of random masks among pairs of participating clients, such that the masks cancel out when a sufficient number of inputs are received.

Differential Privacy

Promise:
“You will not be affected, adversely or otherwise, by allowing your data to be used in any study, no matter what other studies, data sets, or information from other sources is available”.

“The Algorithmic Foundations of Differential Privacy” by Cynthia Dwork and Aaron Roth

Examples

Siri (Apple)

Voice recognition

Gboard (Google-Android)

Text recognition

Alexa (Amazon)

Voice recognition

Tools

TensorFlow

Keras

PyTorch

OpenMined

DO IT

Thanks!

Do you have any questions?

Resources

- The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now
- Robust De-anonymization of Large Sparse Datasets
- De-anonymization of Netflix Reviews using Amazon Reviews
- OpenMined
- Privacy-Preserving Data Science, Explained — OpenMined
- Federated Learning — Google
- Privacy Preserving AI (Andrew Trask) | MIT Deep Learning Series
- Differential Privacy + Federated Learning Explained (+ Tutorial)
- Private AI Mini Course - Udacity
- Practical Secure Aggregation for Privacy-Preserving Machine Learning
- The Algorithmic Foundations of Differential Privacy (Book)
- How Apple personalizes Siri without hoovering up your data