PEPR'2020

PRIVACY IN DEPLOYMENT

PRIVACY IN DEPLOYMENT
PRIVACY IN DEPLOYMENT
PRIVACY IN DEPLOYMENT

PATRICIA THAINE, CO-FOUNDER & CEO, PRIVATE AI
PIETER LUITJENS, CO-FOUNDER & CTO, PRIVATE AI
DR. PARINAZ SOBHANI, HEAD OF APPLIED RESEARCH, GP

WHAT IS PRIVACY?

WHY PRIVACY?

PRIVACY TECH

INTEGRATION

WHAT IS PRIVACY?

"Data privacy means empowering your users to make their own decisions about who can process their data and for what purpose." - gdpr.eu

Signal was the eighth most downloaded social networking app on Tuesday, and ranked around the top 100 for apps overall.

App Annie

Source: https://www.vox.com/platform/amp/recode/2020/6/3/21278558/protest-apps-signal-citizen-twitter-instagram-george-floyd?__twitter_impression=true

Privacy is about maintaining customer trust and more

BUILD BETTER TECH

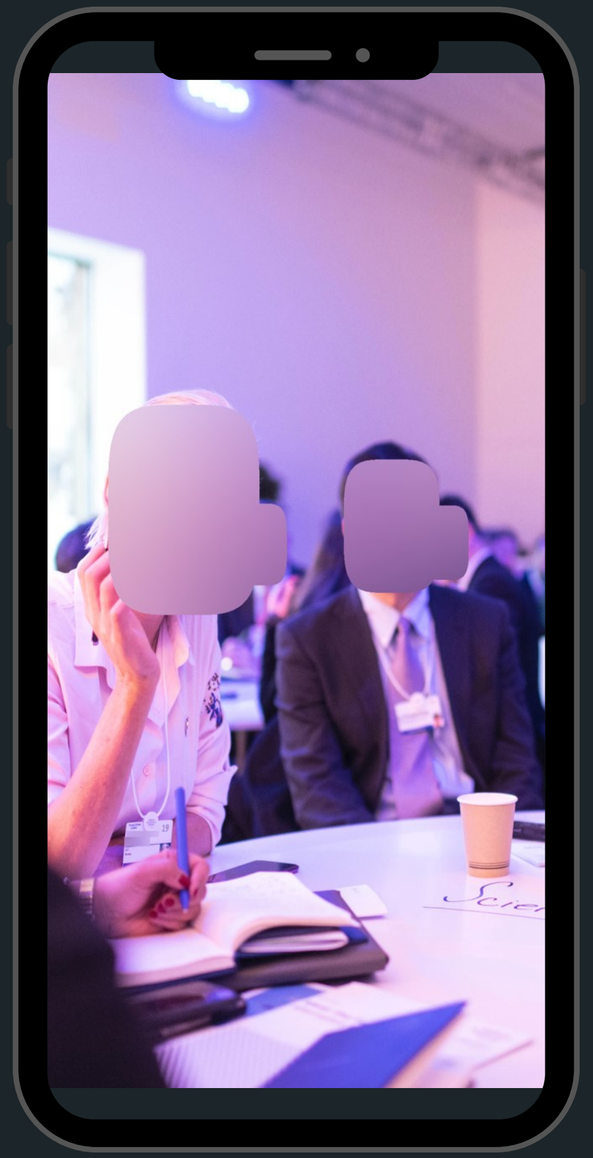
Privacy enhancing technologies unlock datasets that would otherwise be inaccessible.

PEACE OF MIND

Sleep better at night by doing the right thing and decreasing your chances of getting fined.

Privacy enhancing technologies

SECURE MULTIPARTY COMPUTATION
HOMOMORPHIC ENCRYPTION
DATA DE-IDENTIFICATION
DIFFERENTIAL PRIVACY
SECURE ENCLAVES
DATA SYNTHESIS



PRIVACY

PRIVACY

DO YOU HAVE THE USERS' CONSENT TO COLLECT AND PROCESS THEIR INFORMATION?

HAVE CONSENT

Are you sharing information with a 3rd party or directly with the user or group of users whose data you are making predictions on?

DON'T HAVE CONSENT

Get consent - it's required under regulations such as the GDPR.

ARE YOU SHARING INFORMATION WITH A 3RD PARTY OR DIRECTLY WITH THE USER OR GROUP OF USERS WHOSE DATA YOU ARE MAKING PREDICTIONS ON?

3RD PARTY

Do you have to share insights or share a dataset that needs to be visible?

USER(S)

Can predictions be done on user devices or do you have to collect the data in a central database?

DO YOU HAVE TO SHARE INSIGHTS OR SHARE A DATASET THAT NEEDS TO BE VISIBLE?

INSIGHTS

Are you making generalizations over a population or user-specific predictions?

DATASET

Do you need personally identifiable information to be in the dataset (e.g., names, social insurance numbers, faces) or can the data be useful without it?

DO YOU NEED PERSONALLY IDENTIFIABLE INFORMATION TO BE IN THE DATASET (E.G., NAMES, SOCIAL INSURANCE NUMBERS, FACES) OR CAN THE DATA BE USEFUL WITHOUT IT?

NEED PII

Have a data processor agreement, encrypt in transit, and keep track of who you shared the prediction with and what it is being used for. Potentially good candidate for MPC.

DON'T NEED PII

Are you dealing with structured (predefined format) or unstructured (e.g., text, images, video, speech) data?

ARE YOU DEALING WITH STRUCTURED (PREDEFINED FORMAT) OR UNSTRUCTURED (E.G., TEXT, IMAGES, VIDEO, SPEECH) DATA?

STRUCTURED

Data aggregation + differential privacy or risk-based data de-identification and/or data synthesis.

+

Encryption in transit and at rest,
Strict access controls.

UNSTRUCTURED

Risk-based data de-identification and/or data synthesis using AI.

+

Encryption in transit and at rest,
Strict access controls.

METHOD	RISK	IMPLEMENTATION & DEPLOYMENT COMPLEXITY
Differential Privacy	• Mathematical guarantee of privacy based on amount of noise one inserts.	• Need to manage accuracy/noise trade-off, while ensuring one can still guarantee privacy. • Integrated into TF and PyTorch. • Harvard's Open DP, Prof. Reza Shokri's ML Privacy Meter, IBM Diffprivlib.
Anonymization /Pseudonimization	• Based on strong re-identification risk metrics.	• Complex… that's why we're solving this for unstructured data at Private AI.
Data Synthesis	• No mathematical guarantee of privacy (unless perhaps combined with differential privacy), but rather based on strong re-identification risk metrics.	• Also complex… but 3rd parties working on this too (e.g., Replica Analytics under the direction of Professor Khaled El Emam).

DO YOU HAVE TO SHARE INSIGHTS OR SHARE A DATASET THAT NEEDS TO BE VISIBLE?

INSIGHTS

Are you making generalizations over a population or user-specific predictions?

DATASET

Do you need personally identifiable information to be in the dataset (e.g., names, social insurance numbers, faces) or can the data be useful without it?

ARE YOU MAKING GENERALIZATIONS OVER A POPULATION OR USER-SPECIFIC PREDICTIONS?

GENERALIZATIONS

Is latency a critical requirement or can computations take a little longer and be approximated using polynomial operations?

USER-SPECIFIC

Is/are the other party/parties contributing sensitive input data or can your compute the output without additional sensitive information?

IS LATENCY A CRITICAL REQUIREMENT OR CAN COMPUTATIONS TAKE A LITTLE LONGER AND BE APPROXIMATED USING POLYNOMIAL OPERATIONS?

LOW LATENCY

Trusted Execution Environments
(e.g., Intel SGX)

WAIT & APPROXIMATION

Homomorphic Encryption

ARE YOU MAKING GENERALIZATIONS OVER A POPULATION OR USER-SPECIFIC PREDICTIONS?

GENERALIZATIONS

Is latency a critical requirement or can computations take a little longer and be approximated using polynomial operations?

USER-SPECIFIC

Is/are the other party/parties contributing sensitive input data or can your compute the output without additional sensitive information?

IS/ARE THE OTHER PARTY/PARTIES CONTRIBUTING SENSITIVE INPUT DATA OR CAN YOUR COMPUTE THE OUTPUT WITHOUT ADDITIONAL SENSITIVE INFORMATION?

NEED OTHER PARTY'S INPUT

Can you afford higher communication costs and have repeatable algorithms to run?

OUTPUT WITHOUT MORE INFO

De-Identification, Data processor agreement, Encryption in transit and at rest, Strict access controls.

CAN YOU AFFORD HIGHER COMMUNICATION COSTS AND HAVE REPEATABLE ALGORITHMS TO RUN?

LOW COMM COSTS OK + REPEATABLE

Secure Multiparty Computation

LOW COMM COSTS NOT OK OR NOT REPEATABLE

Data processor agreement, Encryption in transit and at rest, Strict access controls.

METHOD	RISK	IMPLEMENTATION & DEPLOYMENT COMPLEXITY	COMPUTATIONAL COMPLEXITY
Homomorphic Encryption	• Can guarantee quantum-safety for inputs and outputs, depending on the scheme and security parameters. • An adversary can still reverse-engineer model weights given the outputs.	• Straightforward enough to implement polynomial operations. • Some alternatives exist for ReLU and sigmoid functions. • Libraries allow for easy on-premise and cloud deployment.	• 3-4 orders of magnitude slower that computing in plaintext, if algorithms are implemented efficiently. Might need more resource to increase parallelization to make up for the time.
Secure Multiparty Computation	• Quantum-safe. • An adversary can still reverse-engineer model weights given the outputs.	• When using garbled circuits, sometimes have to recreate the circuit to make a change. • OpenMined working on making a production-ready library.	• Comm. costs grow linearly. • Computational costs are approximately linear in the depth of the circuit. Can be as low as 1 order of magnitude more than computing on non-encrypted data.
Trusted Execution Environments (Secure Enclaves)	• No mathematical guarantees possible for hardware at this time. • An adversary can still reverse-engineer model weights given the outputs.	• SGX available in most modern computers with an Intel chip & cloud deployment possible. • Need to use oblivious RAM to avoid attacks based on code branching information.	• *If* batch size is small, increased computational cost can be lower than 1 order of magnitude.

PRIVACY IS POSSIBLE

We can protect users' private data while continuing to use them for practical tasks, if we use the right technologies!

PRIVACY IS PRACTICAL

Maintain customer trust, lower risk of data protection and data privacy fines, gain competitive advantage, and get access to more data.

PRIVACY IS THE RIGHT THING TO DO

PRIVACY IN DEPLOYMENT

PRIVACY IN DEPLOYMENT

PRIVACY IN

PRIVACY IN

Contact Us

PATRICIA THAINE

patricia@private-ai.ca

@PrivateNLP

PIETER LUITJENS

pieter@private-ai.ca

DR. PARINAZ SOBHANI

psobhani@georgianpartners.com

@PariAIML