Distributed Machine Learning

Katharine Jarmul
PyLadies Amsterdam &
MLOps Community Amsterdam

/thoughtworks

What if users were able to keep their data and control over it?

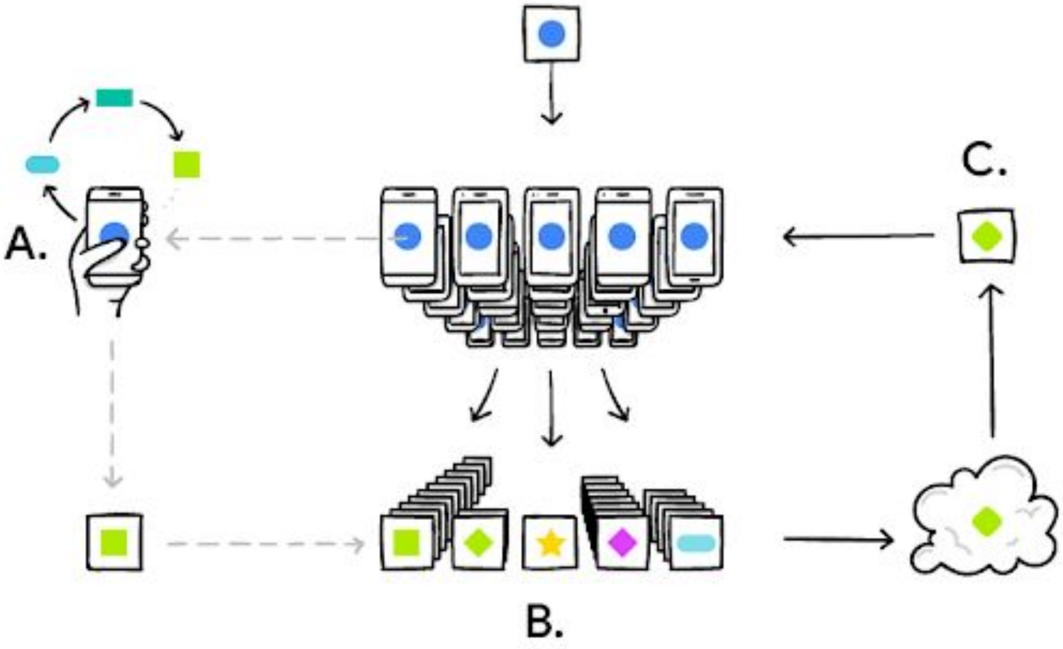
/thoughtworks

What is Federated Learning?

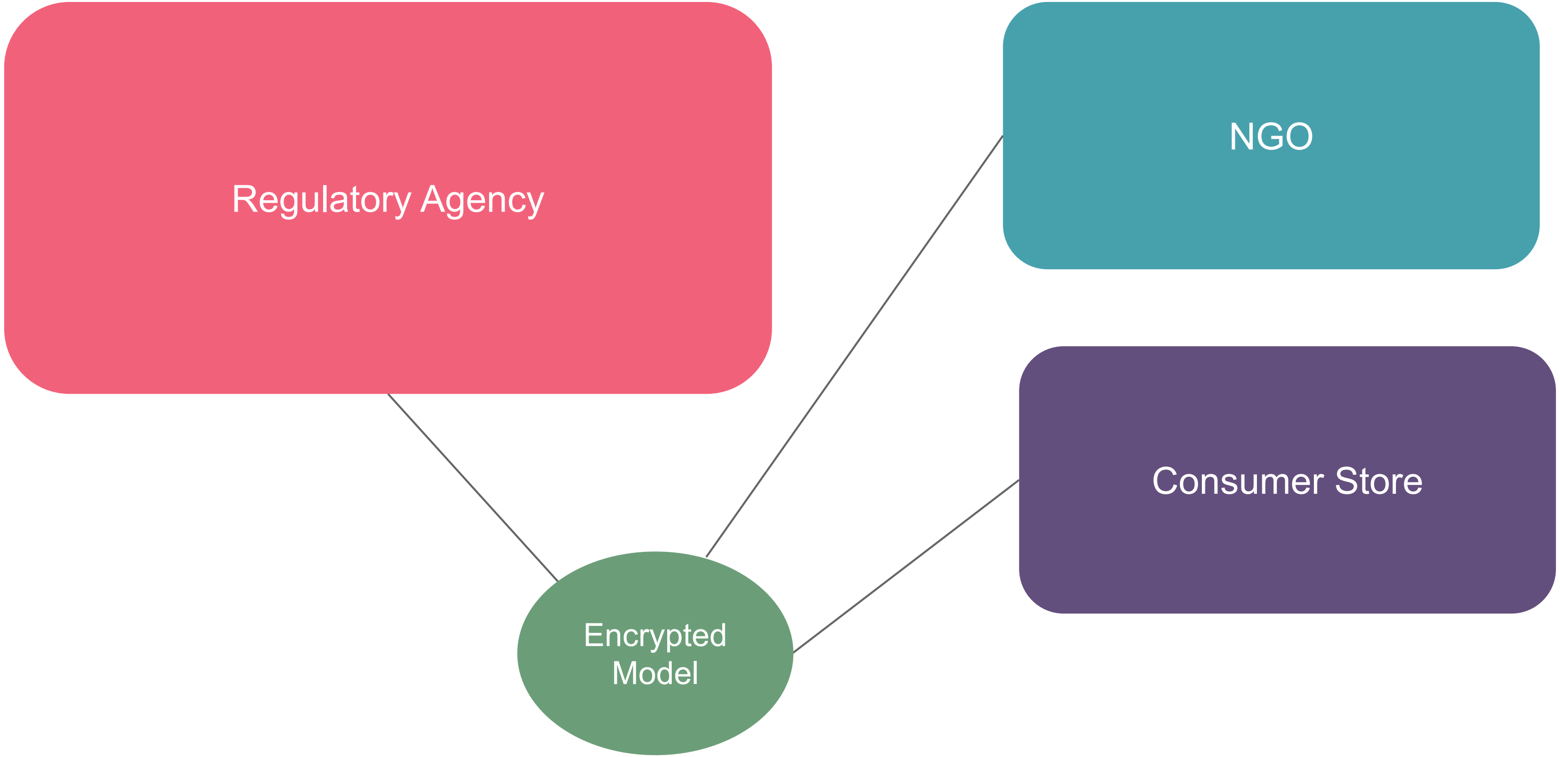
A: Your phone and your data are used to update a model (blue circle)

B: All updates from all participants are sorted and sent to the aggregator.

C: After aggregation, the global model updates are shared back to the participants when a new round can begin.



But it can also look like this...



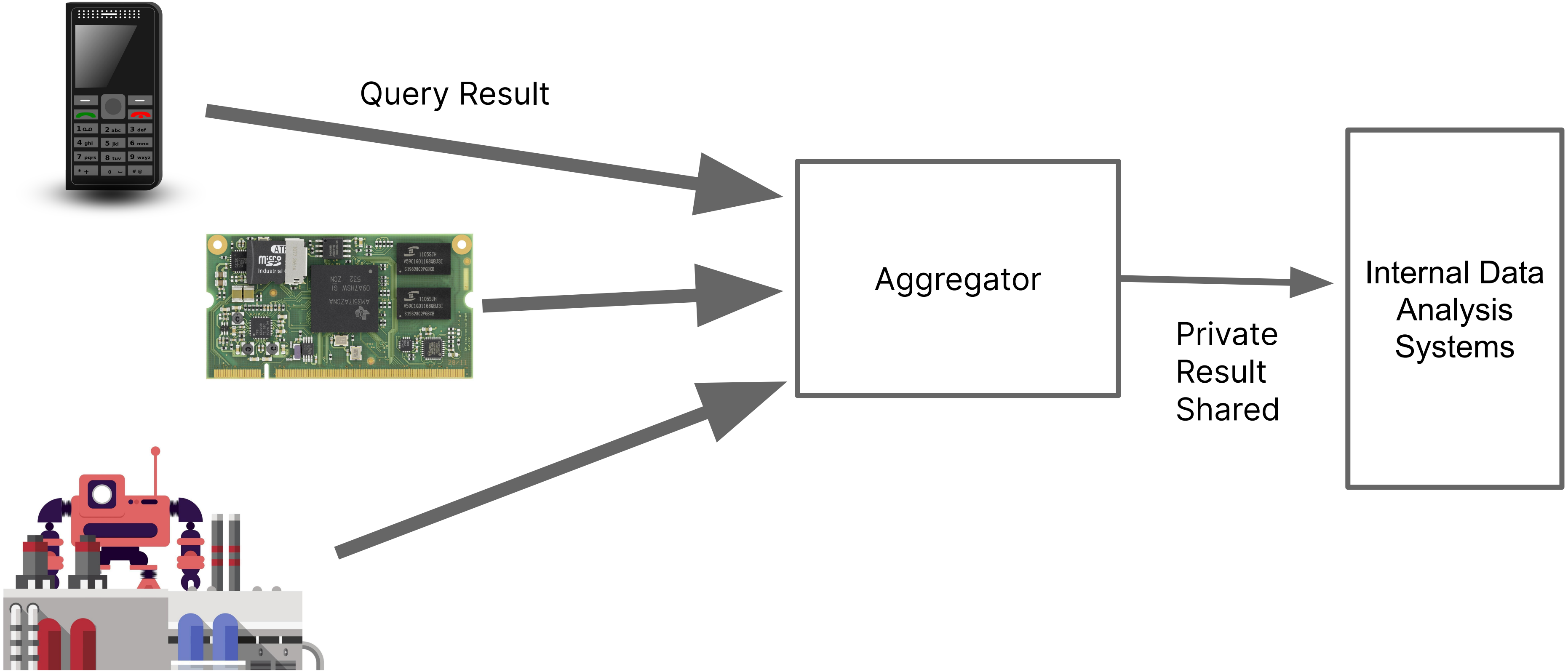
Regulatory Agency

NGO

Consumer Store

Encrypted Model

Or this... (for data analysis)



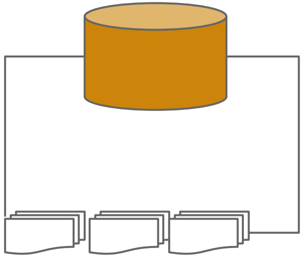
Query Result

Aggregator

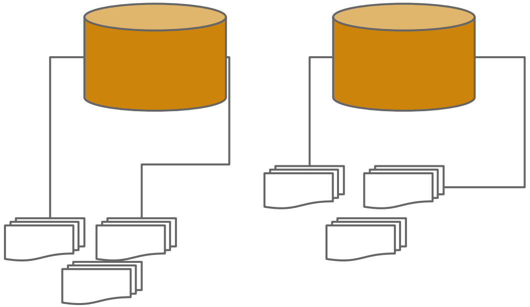
Private
Result
Shared

Internal Data
Analysis
Systems

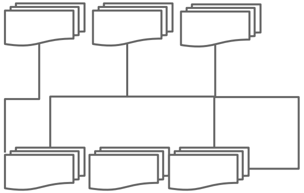
Distributed Learning Architectures



Classic: Centralized Aggregator with Distributed Participants



Clustered: Participants are clustered together and multiple centralized aggregators deploy different models



Fully Distributed: Participants connect to each other using different protocols to organize rounds and updates

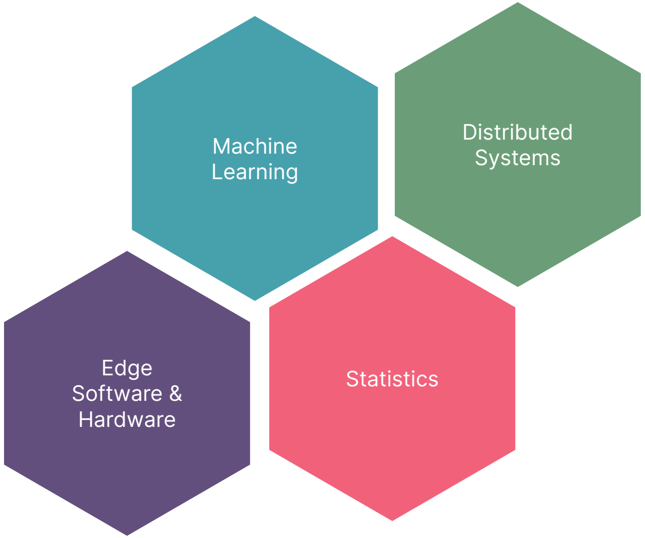
Distributed Learning: Benefits and Weaknesses

Benefit	Weakness
No Data Collection	Data Standardization Required
More Diverse Data	Unevenly Distributed Data
On-device ML	Shared Model
Privacy	Implementation Dependant

MLOps Challenges & Open Questions

- ***Thresholds***: making decisions about selected devices or players and training rounds.
- ***Software and Updates***: keeping players up-to-date. Making processing lightweight.
- ***Participants and Distributions***: population choices and managing divergent, non-i.i.d. data.
- ***Vertical Splits and Multi-Task Learning***: matching labels or features. Training with a variety of data sources.

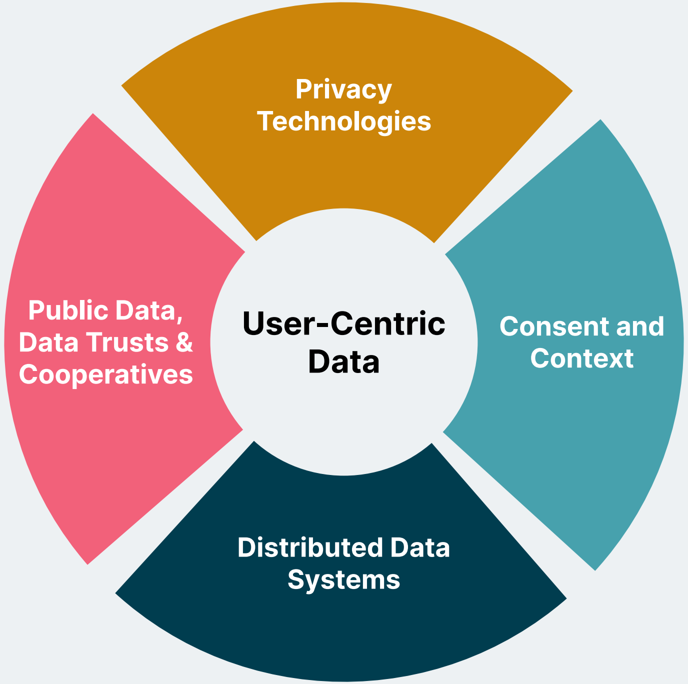
...and many more!



Distribute Data, Distribute Control

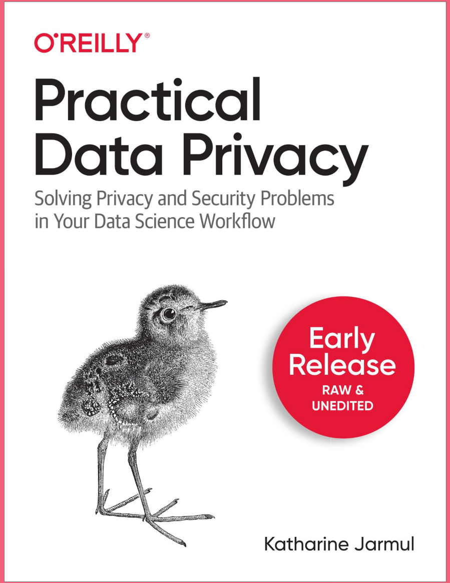
Keeping more data under the user's direct control and say can help create user-centricity in the current ways we manage data and open up new possibilities, like data for public good and fine-grained user consent and control.

It can also shift and challenge the power balances and control in today's ecosystem, creating more democratic AI systems and use cases.



Thank you! Questions? Comments? Thoughts?

Later: @kjam
katharine@probablyprivate.com



O'REILLY®

Practical Data Privacy

Solving Privacy and Security Problems in Your Data Science Workflow

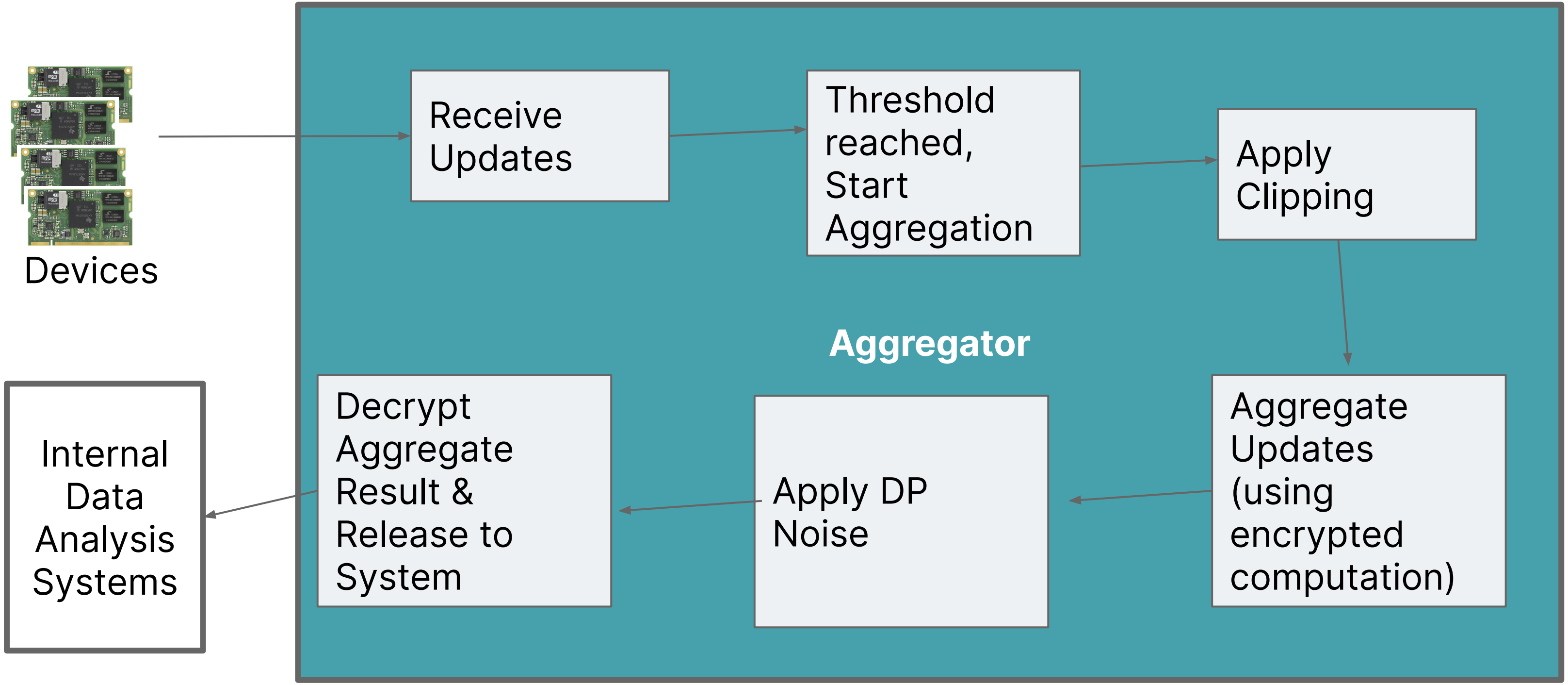
Early Release
RAW & UNEDITED

Katharine Jarmul

References and More Reading

- Google Announcement of Federated Learning: Federated Learning: Collaborative Machine Learning without Centralized Training Data
- [1912.04977] Advances and Open Problems in Federated Learning
- Flower: An interface for several different distributed learning OS libraries
- My InfoQ talk: Katharine Jarmul on Machine Learning at the Edge
- My book: Practical Data Privacy: Practical Data Privacy [Book] (early release) and Pre-Order on Amazon
- My mailing list: Probably Private

Secure & Private Aggregation


Devices
Receive Updates
Threshold reached, Start Aggregation
Apply Clipping
Aggregator
Aggregate Updates (using encrypted computation)
Apply DP Noise
Decrypt Aggregate Result & Release to System
Internal Data Analysis Systems