



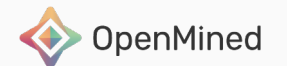
Privacy Conference, September 2020

Privacy-Preserving Natural Language Processing

Fatemehsadat Miresghallah

 @Fatemeh

fmireshg@ucsd.edu



Content

- ◆ NLP Applications
- ◆ ML in Natural Language Processing
- ◆ Threats and Attacks
- ◆ Existing Mitigations
- ◆ Conclusion

Applications of NLP



Translation



Error Detection



Email
Classification



Disease
Prediction



Task Planning



Fake News
Detection



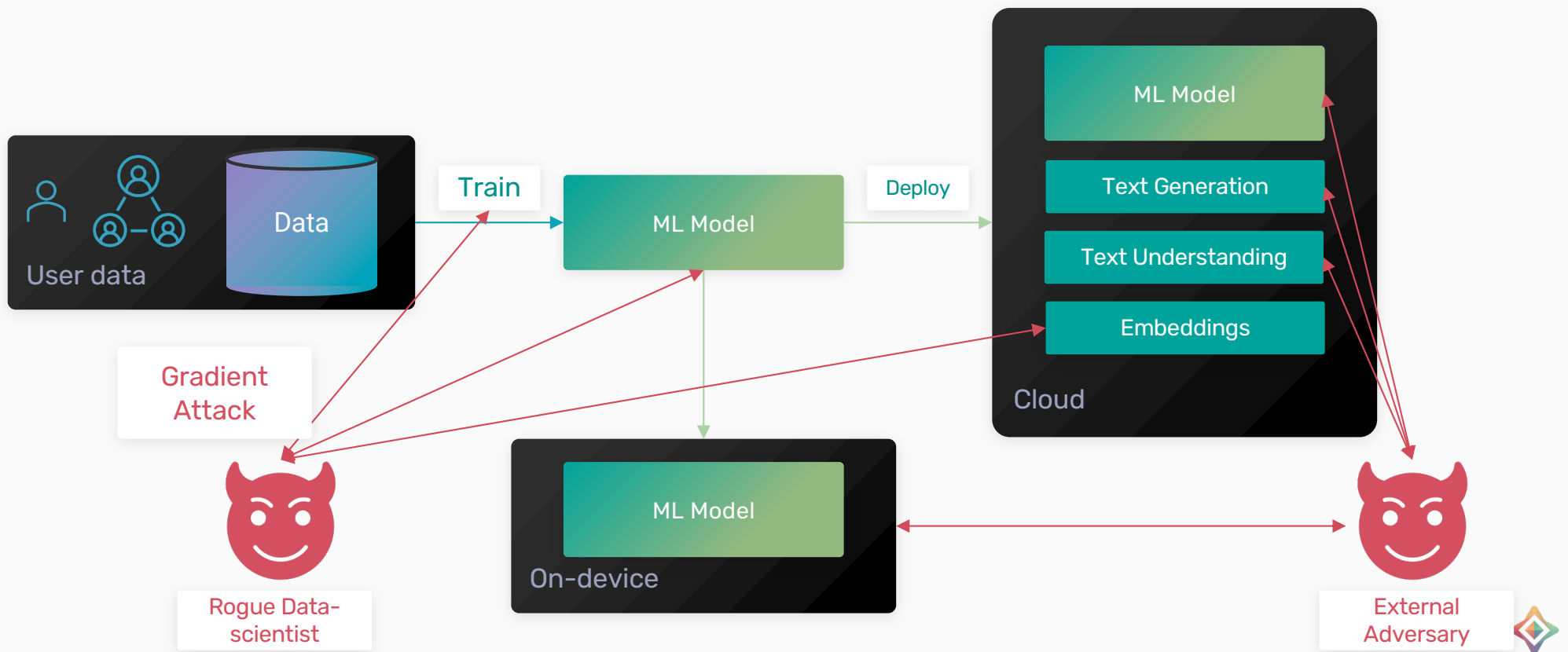
Email
Completion



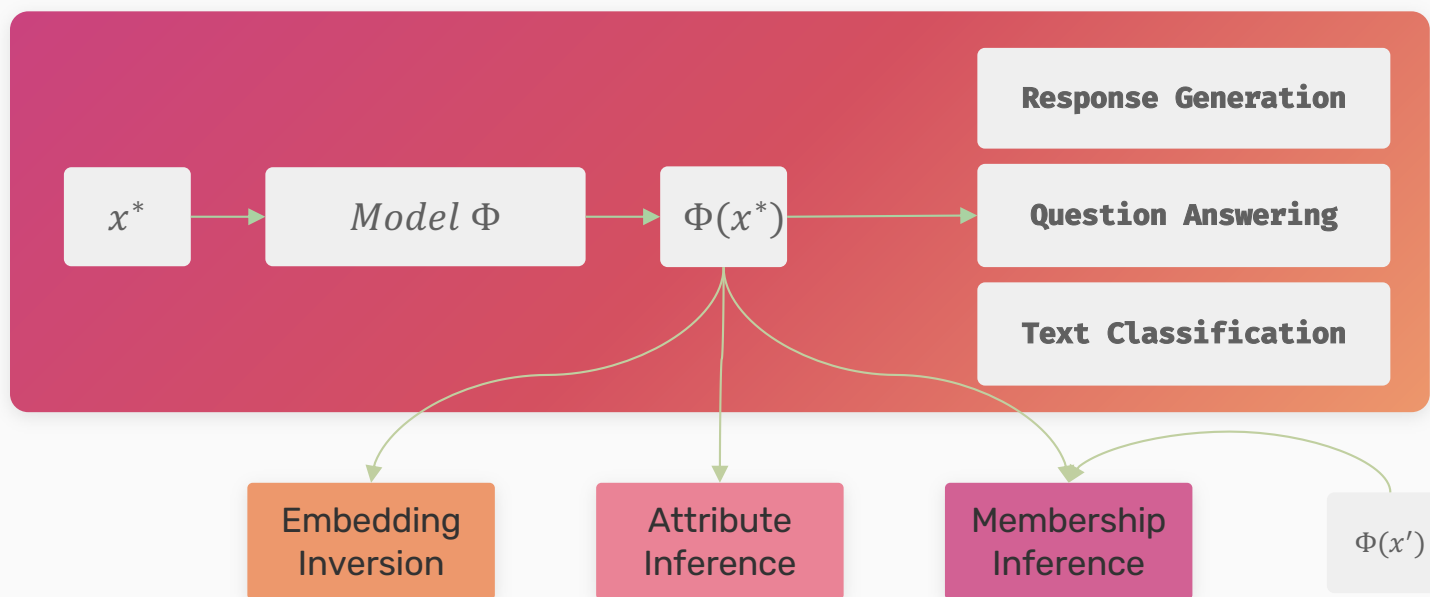
Sentiment
Analysis



ML in Natural Language Processing: Pipeline



Threat 1: Embedding Model Attacks



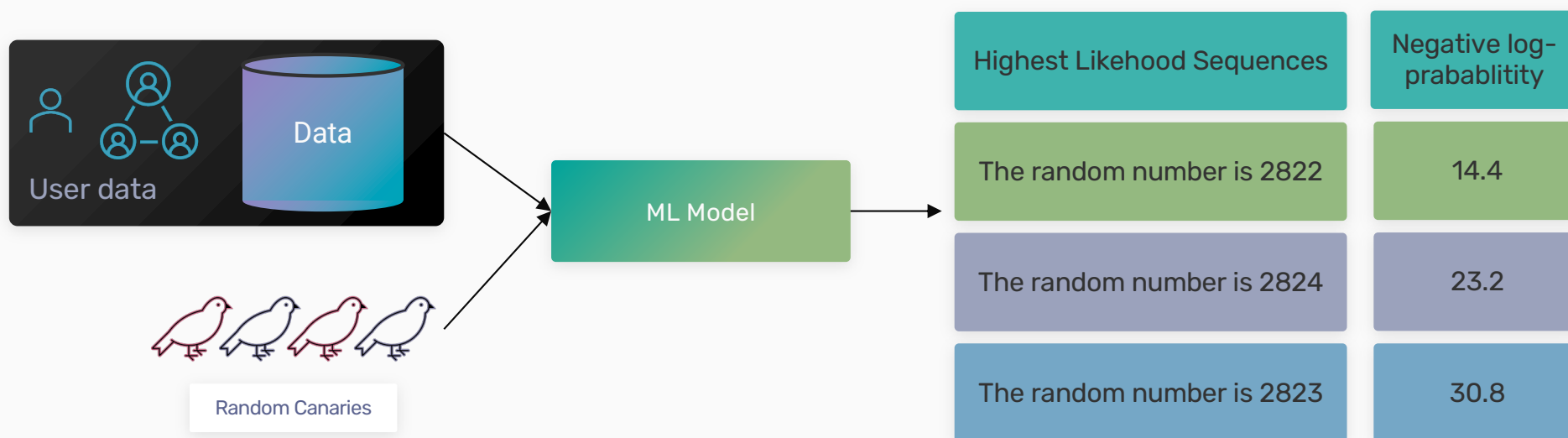
Threat 2: Text-Generation Model Attacks

Unintended Memorization of Secrets

My credit card number is 4403 2212
8563 2345



Threat 2: Text-Generation Model Attacks



“What information about an inserted canary is gained by access to the model?”

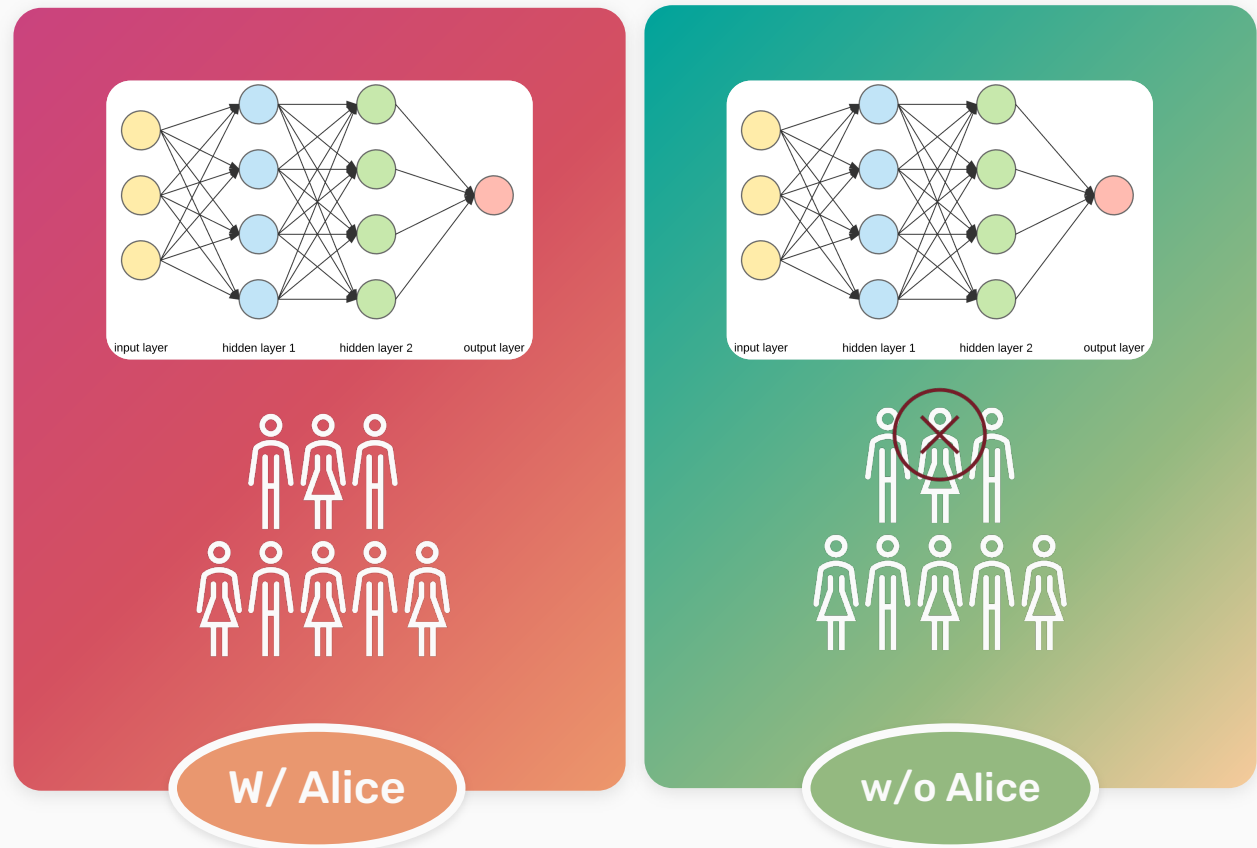
Carlini et al. The secret sharer: Evaluating and testing unintended memorization in neural networks. USENIX Security 2019



Preliminary: Differential Privacy

A randomized algorithm A satisfies ϵ -DP, if for all databases D and D' that differ in data pertaining to one user, and for every possible output value Y :

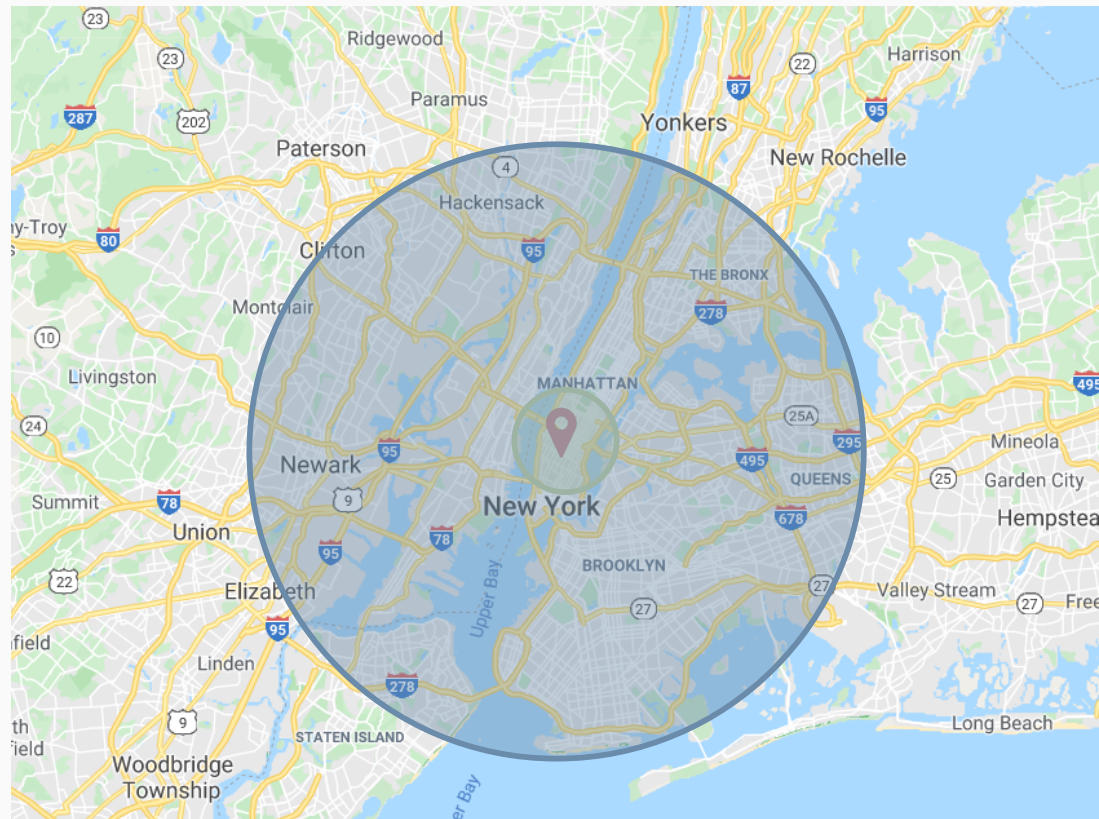
$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^\epsilon.$$



Preliminary: Geo-indistinguishability

Geo-indistinguishability

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^{\epsilon \cdot d(D, D')}.$$



Mitigation 1: Private Embeddings

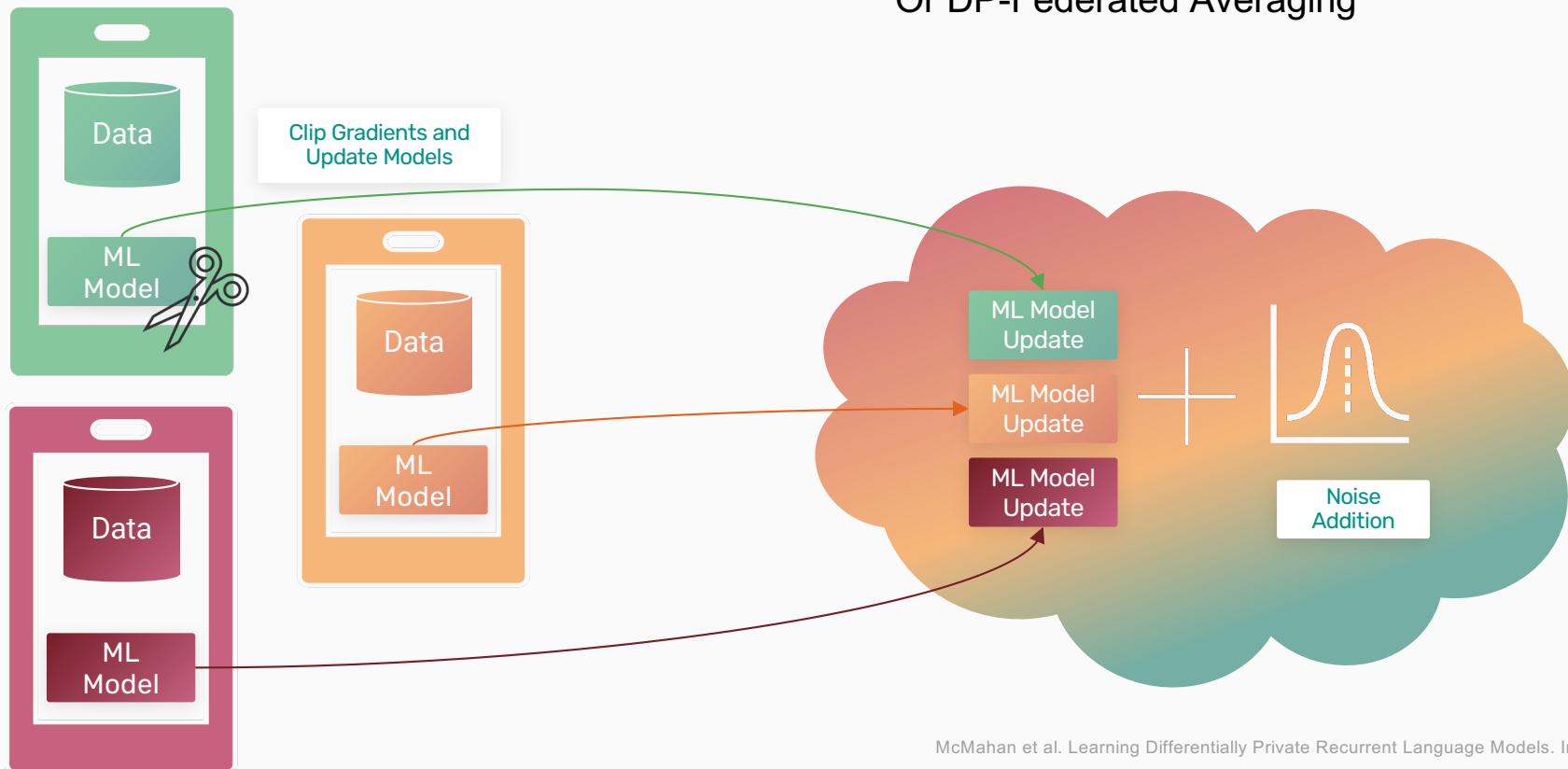


Perturb word embeddings with noise sampled from an exponential distribution.



Mitigation 2: Differentially Private RNNs

Or DP-Federated Averaging



Other Mitigations ...



OLYMPUS: Adversarial Learning

Raval et al. Olympus: sensor privacy through utility aware obfuscation Privacy Enhancing Technologies. 2019

Conclusion: Privacy-preserving for training of language and embedding models is under-explored!



Thank you!

@FatemeH on Slack
fmireshg@ucsd.edu

