What can we learn from Data Leakage and Unlearning for Law?
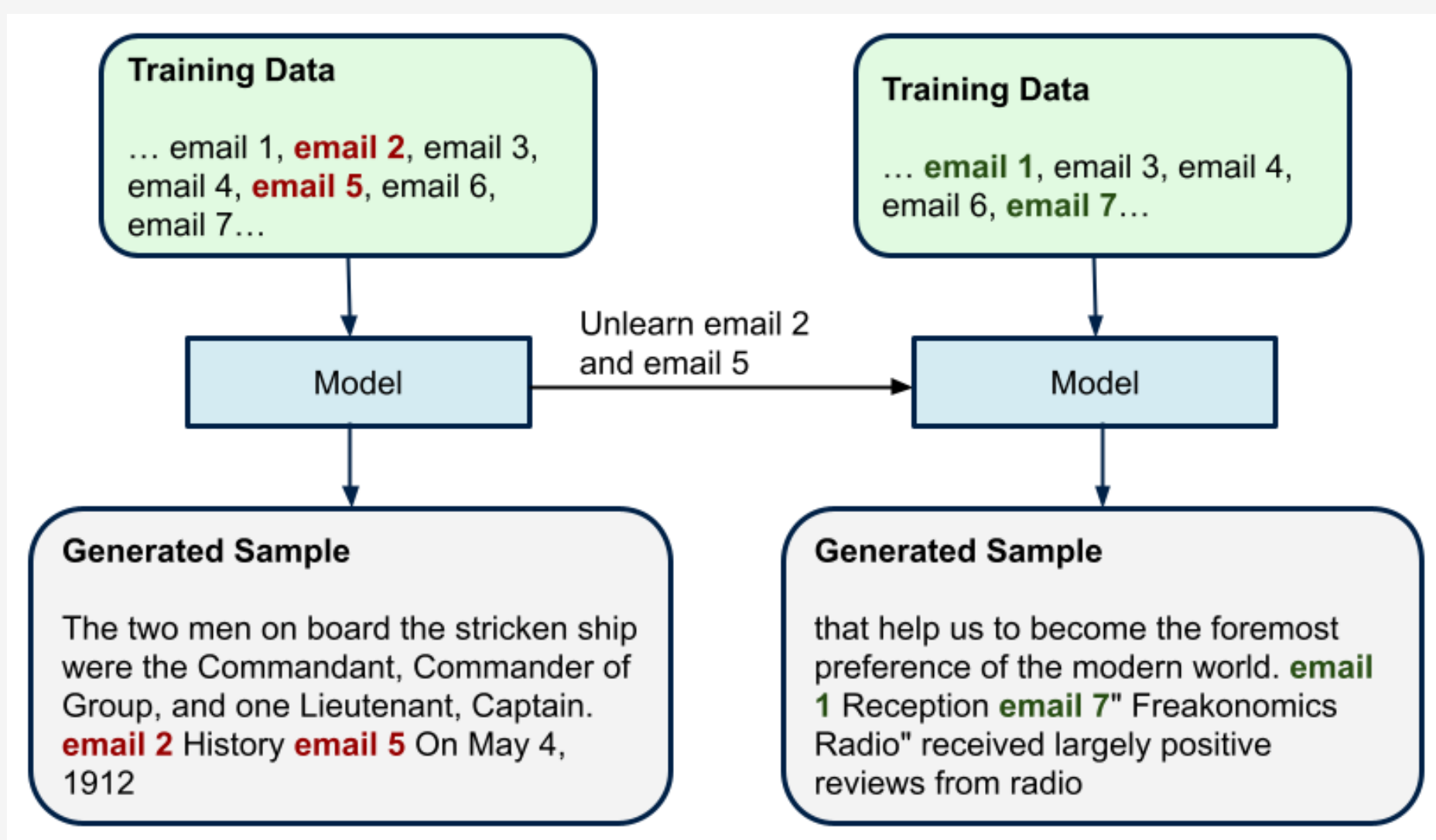
Jaydeep Borkar
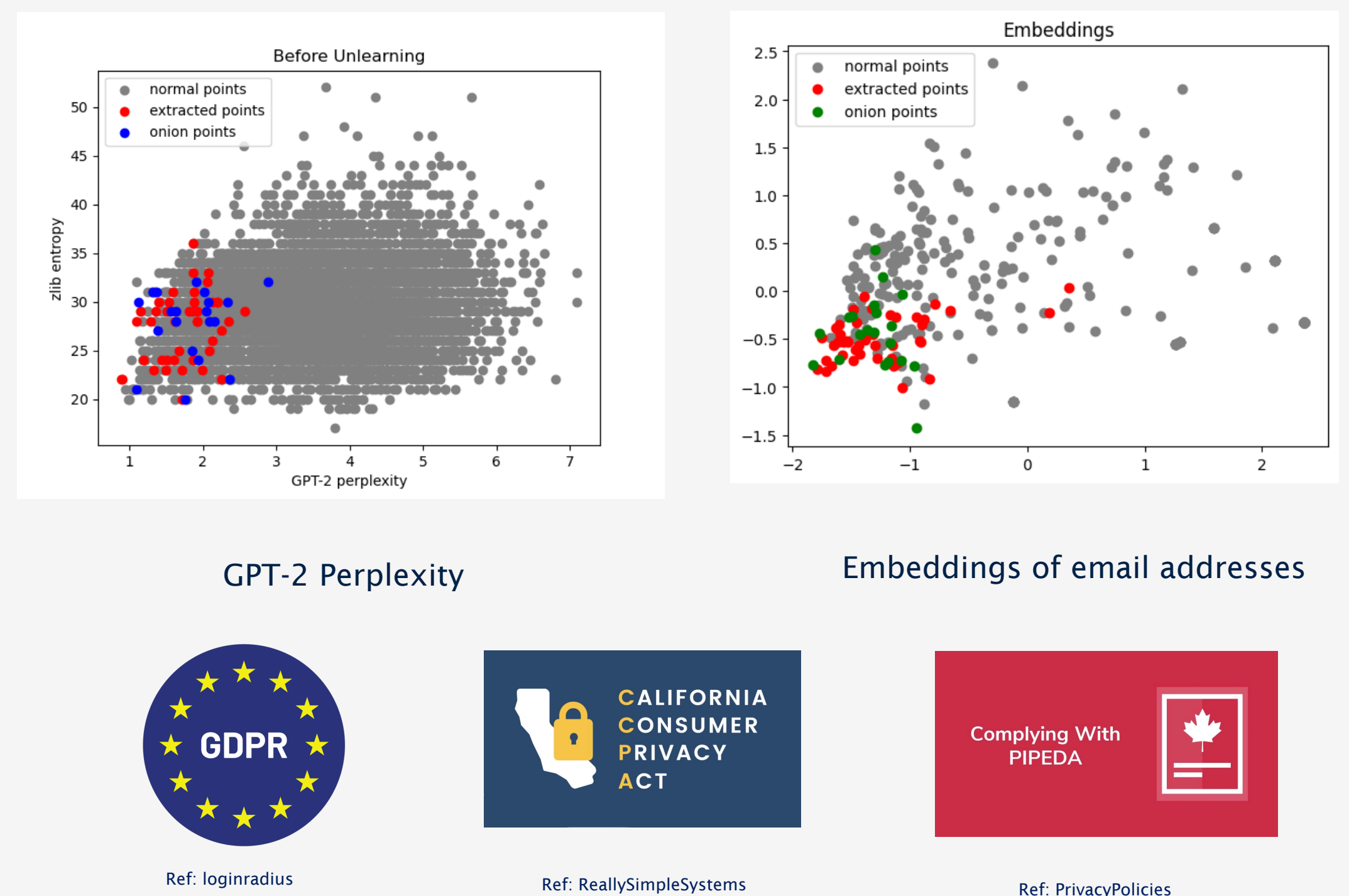Khoury College of Computer Sciences
Northeastern University

Unlearning can jeopardize the overall privacy

Why might we need Unlearning?

- Deleting users' data in order to comply with privacy policies such as *right to be forgotten* or removing data of highly vulnerable users due to safety concerns.
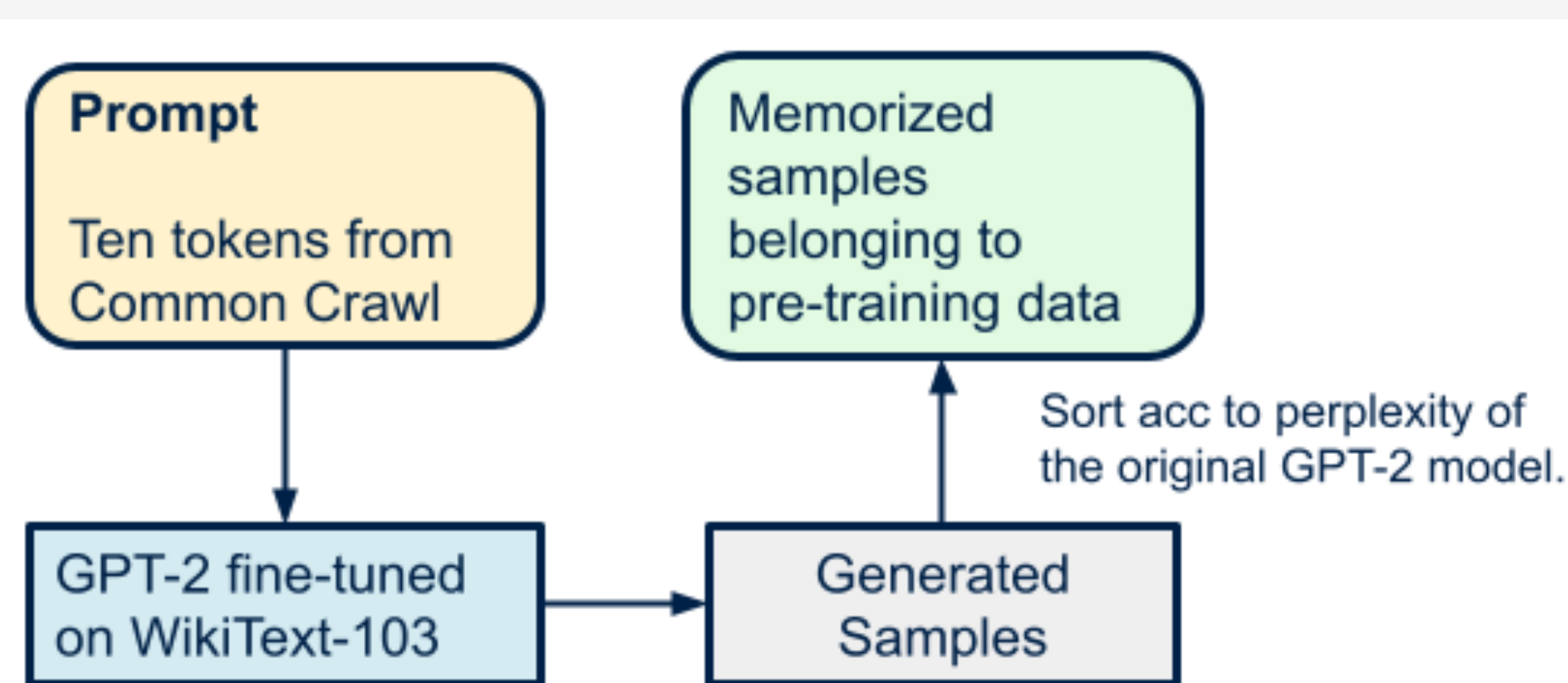

Unlearning specific points (email IDs in this case) can put other neighboring points at the risk of the leakage.

GPT-2 Perplexity

Embeddings of email addresses

Ref: loginradius

Ref: ReallySimpleSystems

Ref: PrivacyPolicies

1. Exact Unlearning is **not** the best way to implement *right to be forgotten*.

2. Fine-tuned models can leak **both** fine-tuning and pre-training data.

Extracting pre-training data from fine-tuned models

Fine-tuned models can leak pre-training data if the attacker has access to a dataset having similar distribution to craft prompts.

ICML International Conference On Machine Learning

GenLaw '23
Honolulu, Hawai'i

@JaydeepBorkar

borkar.j@northeastern.edu

Scan QR code
for the paper