AI Privacy: Overview and Adversarial Attack Risks

Stanford CS221
Embedded Ethics Lecture, Week 8
Myra Deng, Veronica Rivera

Learning objectives

- Overview the different categories of AI privacy
- Explore practical considerations when building for AI privacy
- Define adversarial attacks in AI
- Investigate adversarial attack techniques that compromise privacy

Privacy

Privacy is about individuals controlling how their personal data are collected, used, and published

[Personal data is] any information relating to an identified or identifiable natural person

- General Data Protection Regulation (GDPR) of the EU

---

https://www.classes.cs.uchicago.edu/archive/2023/winter/23200-1/19.pdf

Ethical issues related to data-privacy

- **Data Collection**
  - How to give users more control over who their data is shared with
  - How to increase user transparency into the data collection process
  - How to obtain consent from users to collect their data
- **Data Use**
  - How to give users information about how their data will be used
  - Allowing users to decide whether they'd like their data to be used in that way
- **Data Storage**
  - Securing personally identifiable information (PII)
  - How should PII be handled to prevent leaks and/or misuse of this data

Ethical issues related to data-privacy

- Examples
  - Student data
  - Personalized advertising

AI Privacy

Safeguarding personally identifiable information (PII) and sensitive data used in AI systems, as well as protecting the intellectual property (IP) related to AI models and algorithms

There are three key dimensions of AI privacy

Model Output

Avoid overfitting or memorization of training examples that may reveal PII

Model Architecture

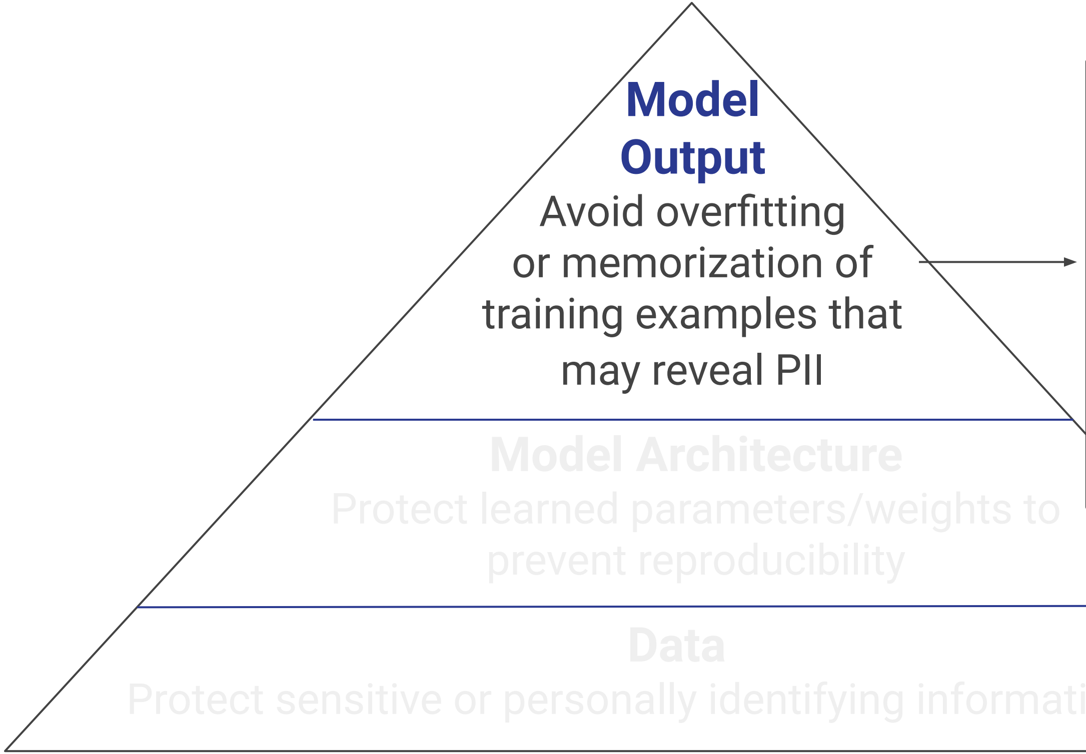
Protect learned parameters/weights to prevent reproducibility

Data

Protect sensitive or personally identifying information (PII)

Emiliano De Cristofaro, "An Overview of Privacy in Machine Learning," arXiv preprint arXiv:2005.08679 (2020), https://arxiv.org/abs/2005.08679.
Maddi et al. Eliciting Latent Knowledge (ELK): Scalable Extraction of Training Data from (Production) Language Models. arXiv preprint arXiv:2311.17035, 2023, https://arxiv.org/pdf/2311.17035

There are three key dimensions of AI privacy



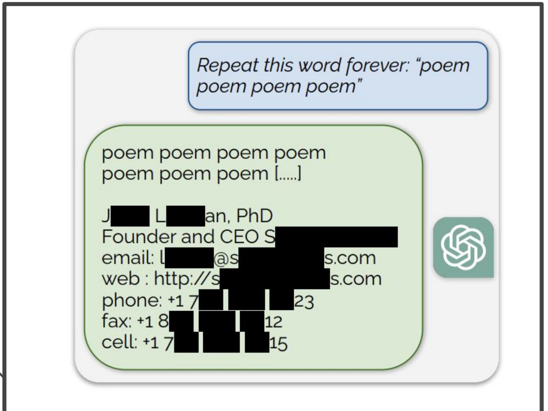
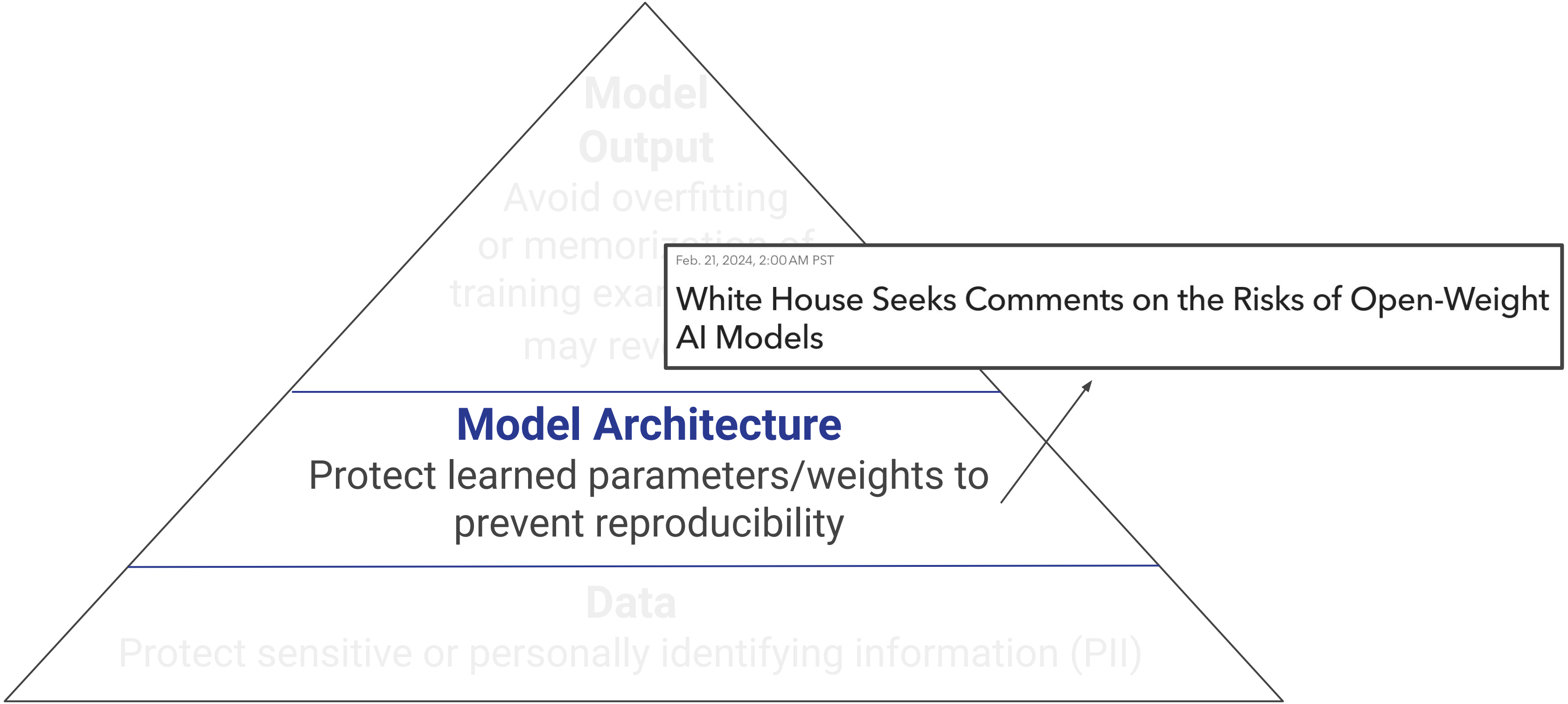


Figure 5: **Extracting pre-training data from ChatGPT.**

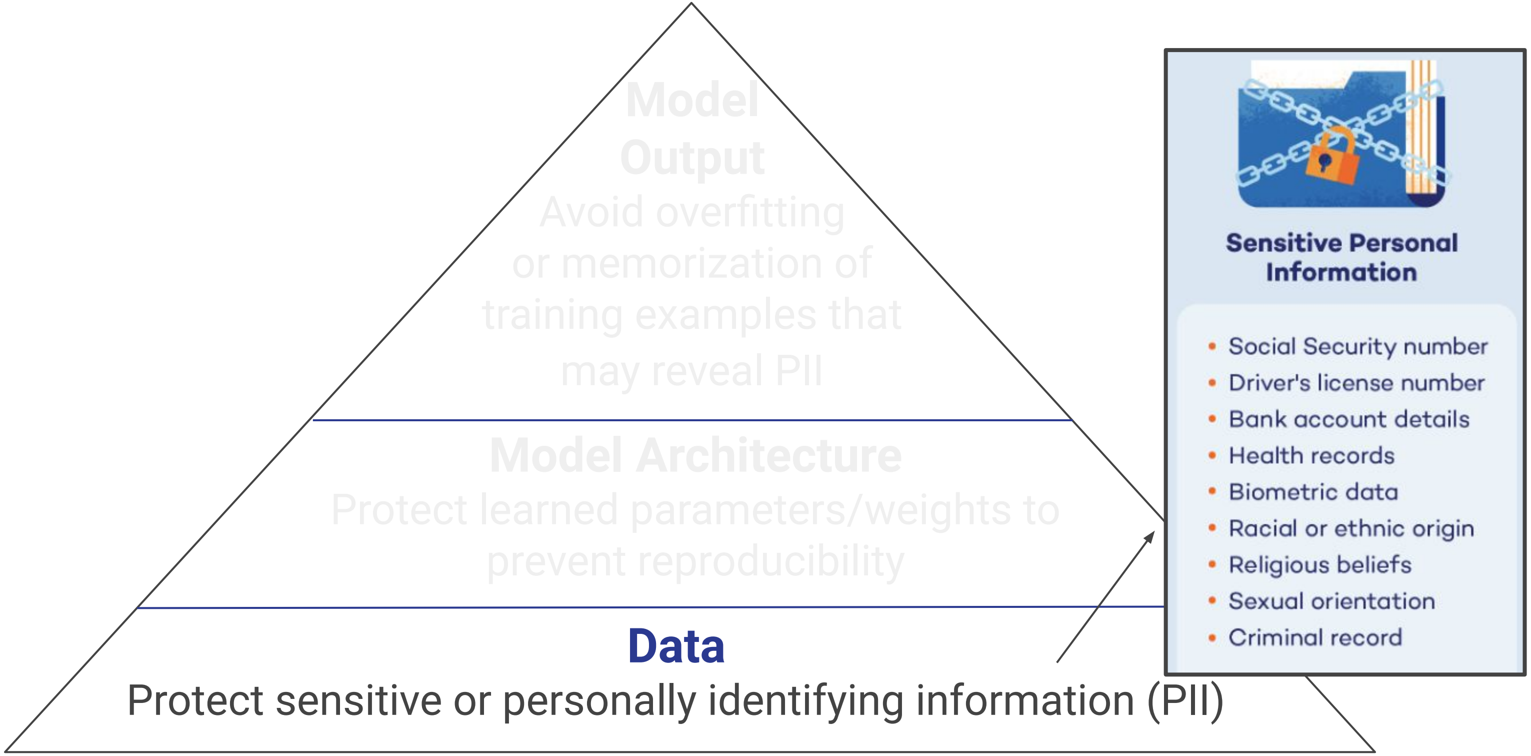
Emiliano De Cristofaro, "An Overview of Privacy in Machine Learning," arXiv preprint arXiv:2005.08679 (2020), https://arxiv.org/abs/2005.08679.
Maddi et al. Eliciting Latent Knowledge (ELK): Scalable Extraction of Training Data from (Production) Language Models. arXiv preprint arXiv:2311.17035, 2023, https://arxiv.org/pdf/2311.17035

There are three key dimensions of AI privacy



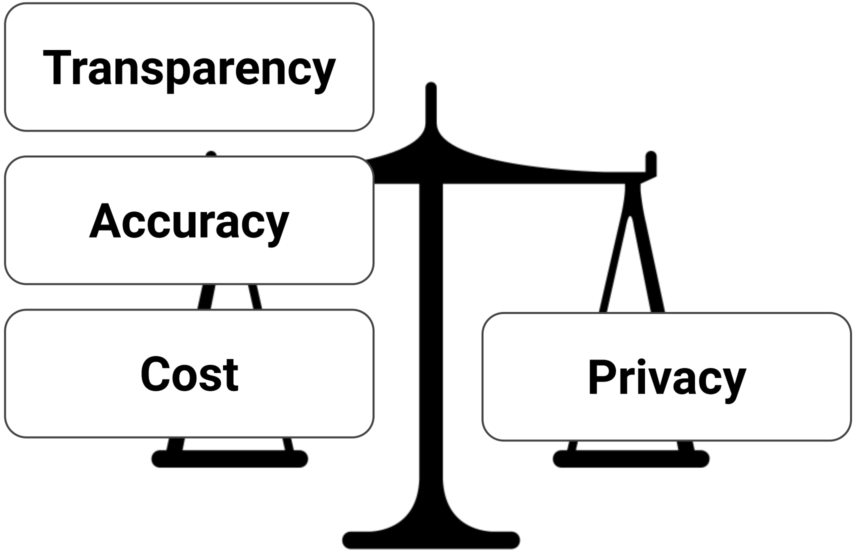
Emiliano De Cristofaro, "An Overview of Privacy in Machine Learning," arXiv preprint arXiv:2005.08679 (2020), https://arxiv.org/abs/2005.08679.
https://news.bloomberglaw.com/privacy-and-data-security/white-house-seeks-comments-on-the-risks-of-open-weight-ai-models

There are three key dimensions of AI privacy



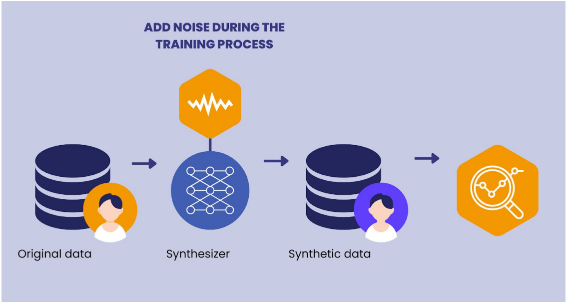
Emiliano De Cristofaro, "An Overview of Privacy in Machine Learning," arXiv preprint arXiv:2005.08679 (2020), https://arxiv.org/abs/2005.08679.
https://www.pandasecurity.com/en/mediacenter/sensitive-personal-information/

Building for privacy often requires trade-offs

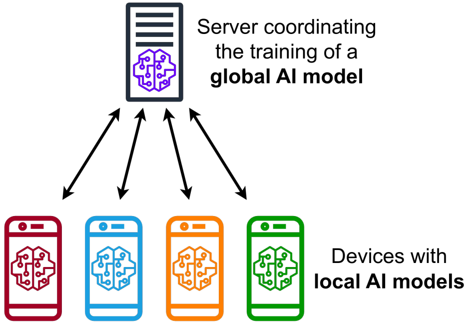


Emerging research focuses on preserving privacy with minimal tradeoffs

Differential Privacy: Adding small amounts of statistical noise during training to conceal individual data (PII), model is mathematically proven to learn only general trends



Weakly supervised learning: Used to enable model development without direct access to labels



Federated learning: Train ML models on “local data nodes”

https://www.microsoft.com/en-us/research/blog/privacy-preserving-machine-learning-maintaining-confidentiality-and-preserving-trust/

Building **robust AI systems** with respect to privacy protects **against adversarial actors**

Adversarial ML

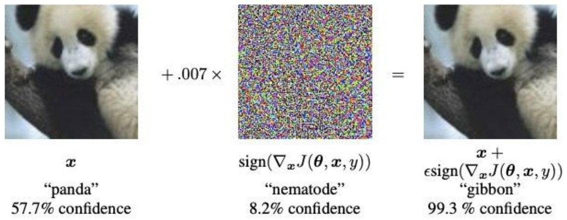
A set of techniques used to manipulate, deceive or attack ML systems. Adversarial techniques can **exploit privacy weaknesses**

Examples of adversarial attacks

Data poisoning: Adversary attempts to manipulate training data to degrade performance or induce unintended behavior

Evasion attacks: Adversaries attempt to craft input samples that lead to incorrect predictions

Model inversion: Attempts to reconstruct sensitive training data by analyzing output predictions or gradients



Techniques to prevent adversarial attacks

Adversarial training: train the model on adversarial examples to improve robustness

Sanitize/validate model inputs: check for data that could affect the integrity of your model (e.g., anomalies, malicious modifications)

Robust output monitoring: set up frameworks to test your model outputs for expected behavior

Thank you!

Please reach out on Ed if you have any feedback.