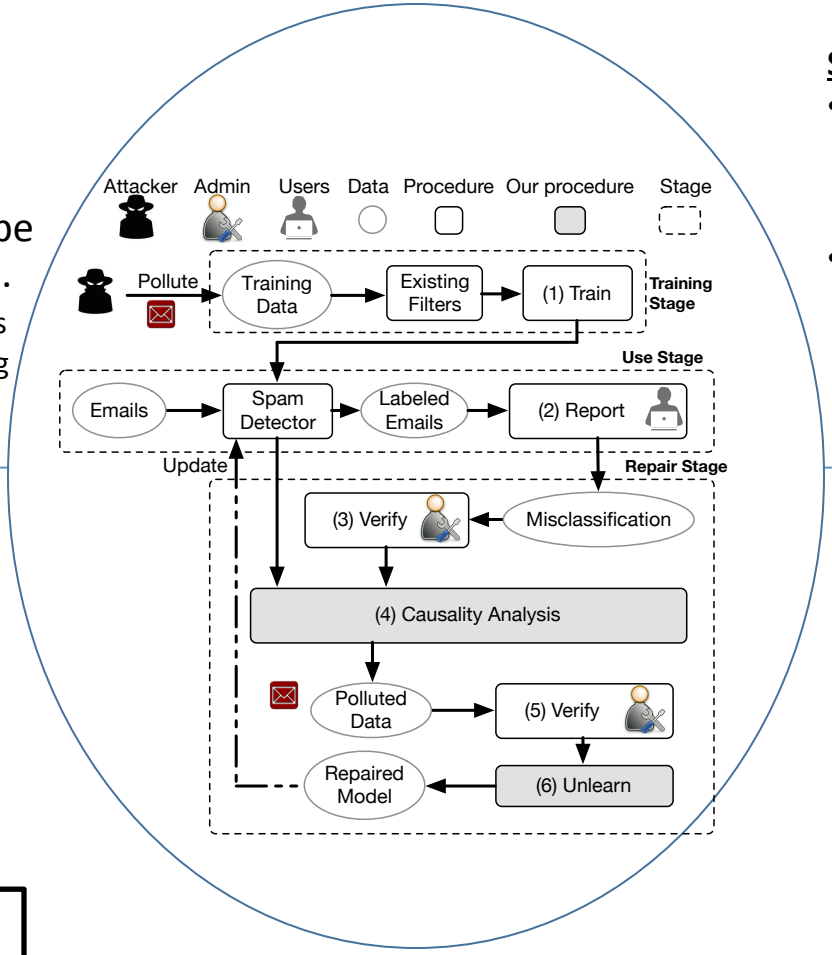
Efficient Repair of Learning Systems via Machine Unlearning

Challenge:

- Training set of a learning system can be polluted by attackers.
  - Inject crafted samples
  - Alter labels of training samples

Solution:

- Machine unlearning: Ask learning systems to forget data
  - Transform existing learning model to a summation form



Scientific Impact:

- We will change the long-standing doubt of the robustness of machine learning to adversaries.
- Machine learning techniques can be used more widely in security and privacy research without being worried about polluted training set.

Broader Impact:

- Microsoft may use machine unlearning to ask its AI-powered bot, Tay, to forget racism statements.
- Adversarial machine learning is covered in both PIs' courses.
- The Atlantic magazine features "machine unlearning" in an article.

CNS-1564055/1563843, Columbia/Lehigh University,
Contact: Yinzhi Cao (yinzhi.cao@lehigh.edu)
Junfeng Yang (junfeng@cs.columbia.edu)