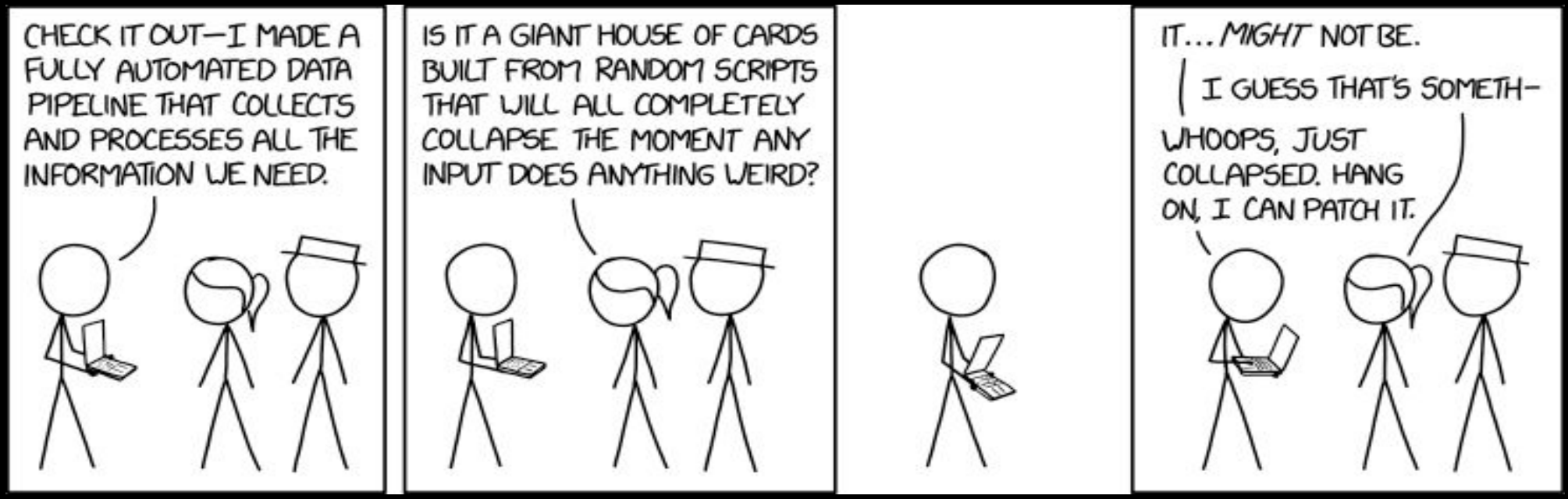
Week 13
AI Security

Anusha Ghosh

Announcements

- ACM Bar Crawl tonight after our meeting!
- b01lersCTF this weekend!
- CypherCon next Thursday
- T-shirts!!! Fill out the Google Form to order a SIGPwny T-Shirt!
  - **bit.ly/sigpwnyshirt**

sigpwny{ai_go_brrr}



Overview

- AI security as a field has two main branches
  - using AI for security purposes
  - exploring the security of AI itself
- Both are awesome fields, but be aware that “AI security” can mean either one!

AI for Security

AI for Malware

- Using AI to find malware and shut down execution
  - scans code before execution to determine malware potential
- AI for network monitoring
  - scans packets to monitor for attacks
- AI for anything else
  - pretty much anything that requires human foresight/monitoring

Pros/Cons

- Good things!
  - potential to find undiscovered malware and patch it
  - discover bugs before they can be exploited
  - continuous monitoring without continuous human involvement
- Bad things :(
  - potential to harmfully misclassify
  - black box
  - does accuracy generalize to the real world?

Security of AI

Security of AI

- How to find ways to make models behave incorrectly on purpose
  - key word here: adversarial
- Many different ways to get models to misbehave
  - small differences in input can completely change output!

Pros/Cons

- Good uses
  - Privacy preservation
  - Vulnerability detection
  - Explainability
  - Robustness
- Harmful uses
  - Attacking production systems

Privacy Preserving AI

Privacy Preserving AI/ML

- Data collection side
  - How do you select proper datasets?
- Data processing side
  - Can we make training data more private?
- Architecture side
  - Model federation
    - How to train on decentralized data
  - Model unlearning

Career Paths

- Research!!
  - Industrial research with a company
  - Staying in academia
- AI Engineering
- AI product teams

AMA

Next Meetings

Sunday Seminar: 2022-04-24

- Come play b01lersCTF!

Next Thursday: 2022-04-28

- Off for Cyphercon!