WeAreDevelopers
World Congress

NetApp

Confuse, Obfuscate, Disrupt

Using Adversarial Techniques for Better AI and True Anonymity

David vonThenen

@davidvonthenen

David vonThenen

- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Cloudy, cloudy, cloudy…
- There is storage for that!

@davidvonthenen

Agenda

- How Data Inconsistencies Happen
  - Demos, Demos, Demos
- Adversarial Attacks for Good... & Bad
  - Demos, Demos, Demos
- Adversarial Attack Defense
  - Demos, Demos, Demos
- Q&A

How Data Inconsistencies Happen

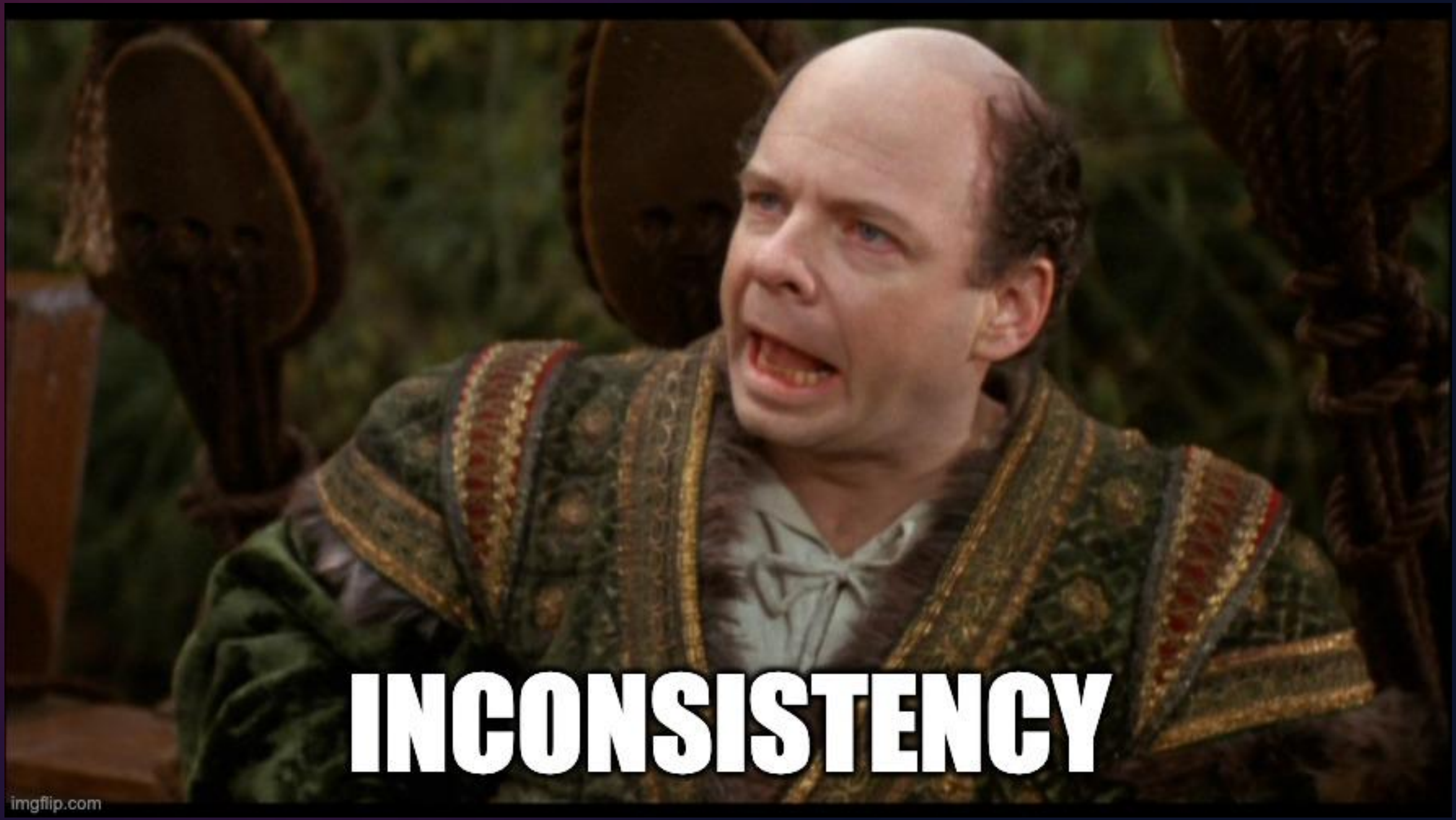
Flawed Data

- AI/ML Only As Good As the Data
  - Biased, Noise, Inaccuracies
- Real-World Examples:
  - Recruiter AI + Male Skewed
    - Not Representative Data
  - Offensive AI Chatbot
    - Using Racist Language
  - Court Case Hallucinations
    - ChatGPT fake cases
- Common Ways Of Flawed Data Getting Into Our Datasets…

1. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/
2. https://storage.courtlistener.com/recap/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.31.0.pdf
3. https://en.wikipedia.org/wiki/Tay_(chatbot)

Data Inconsistencies Matter

- AI "Decision Making" Directly Shaped By Data
  - Annotation Errors
  - Data Bias
  - Distribution Drift
  - Adversarial Data
  - Overfitting
  - Underfitting
  - Poor Feature Engineering
  - Noisy Data, etc...



Annotation Errors

RED

Data Imbalance

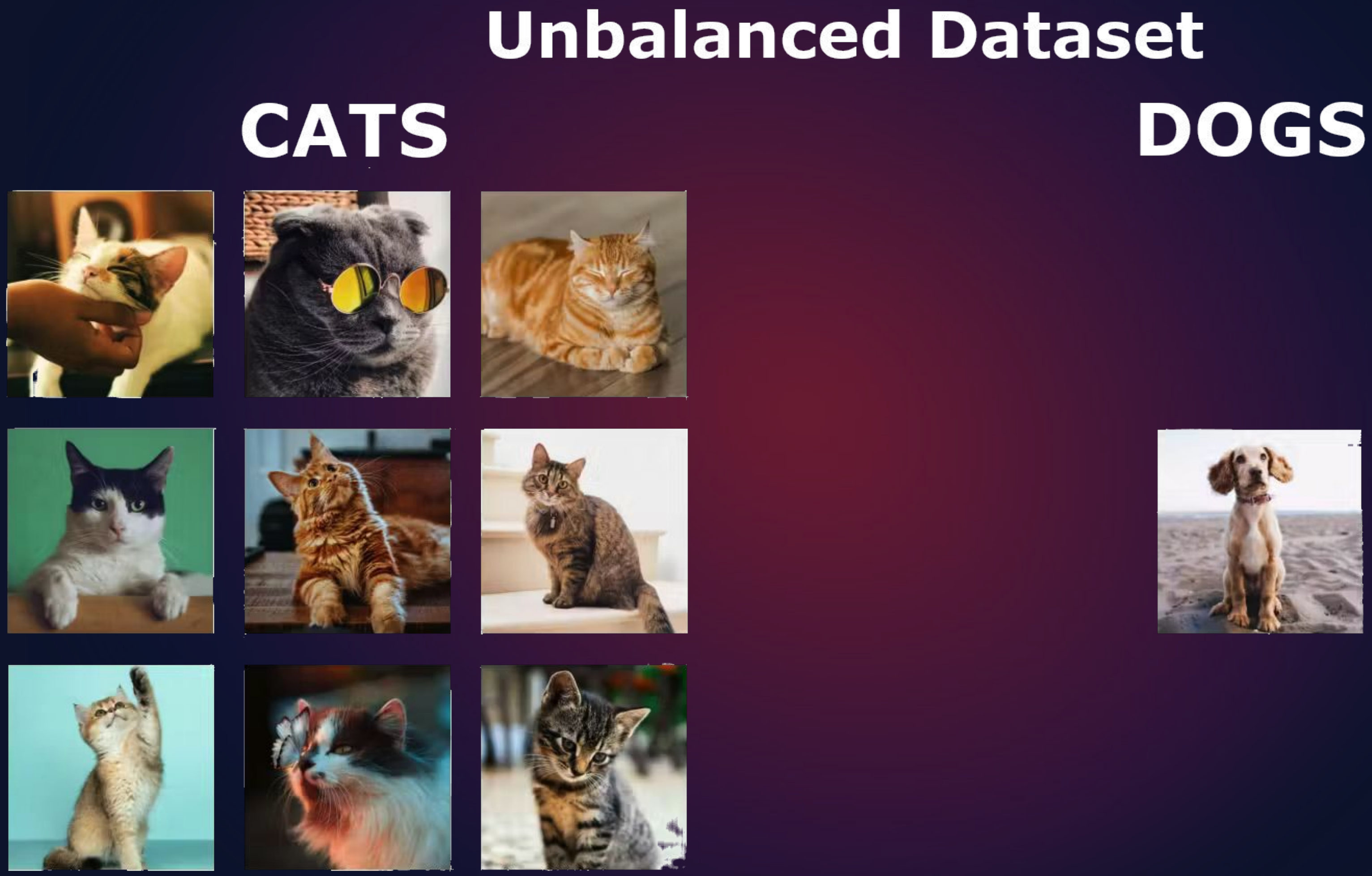


Image Attribution
Introduction to Balanced and Imbalanced Datasets in Machine Learning
Nikolaj Buhl on November 11, 2022

Adversarial Samples

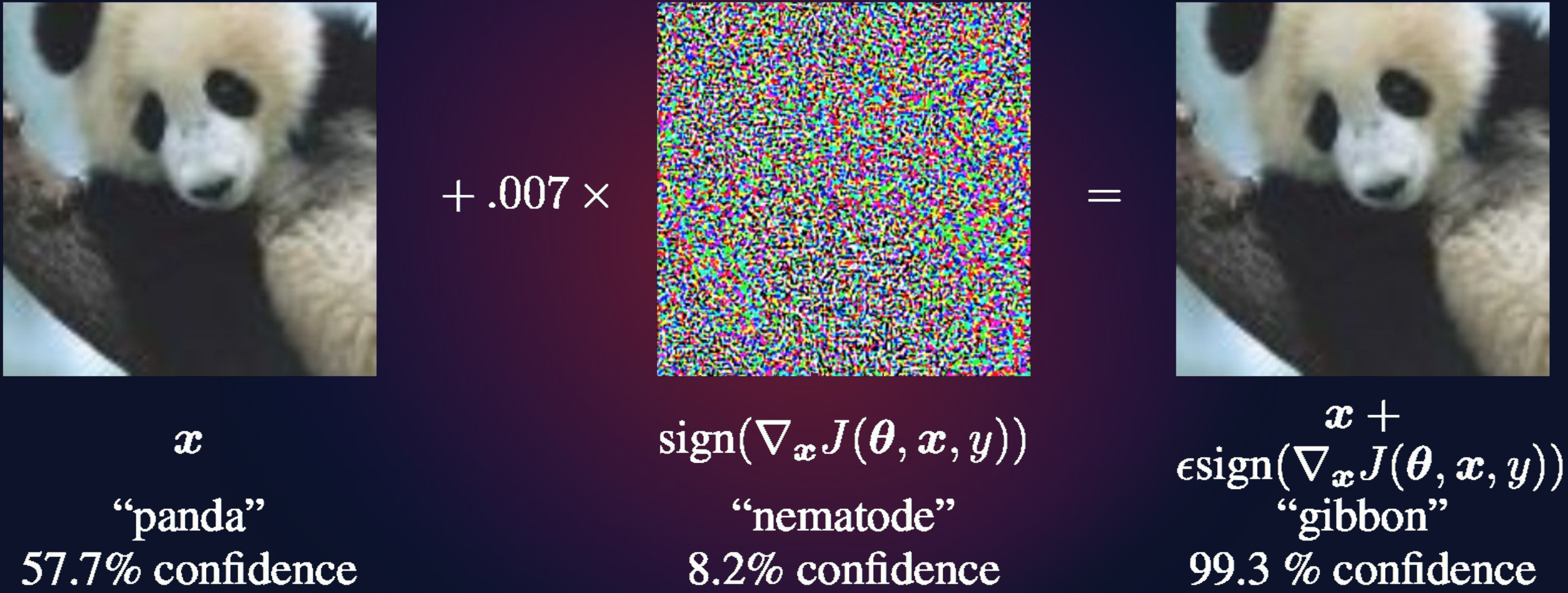
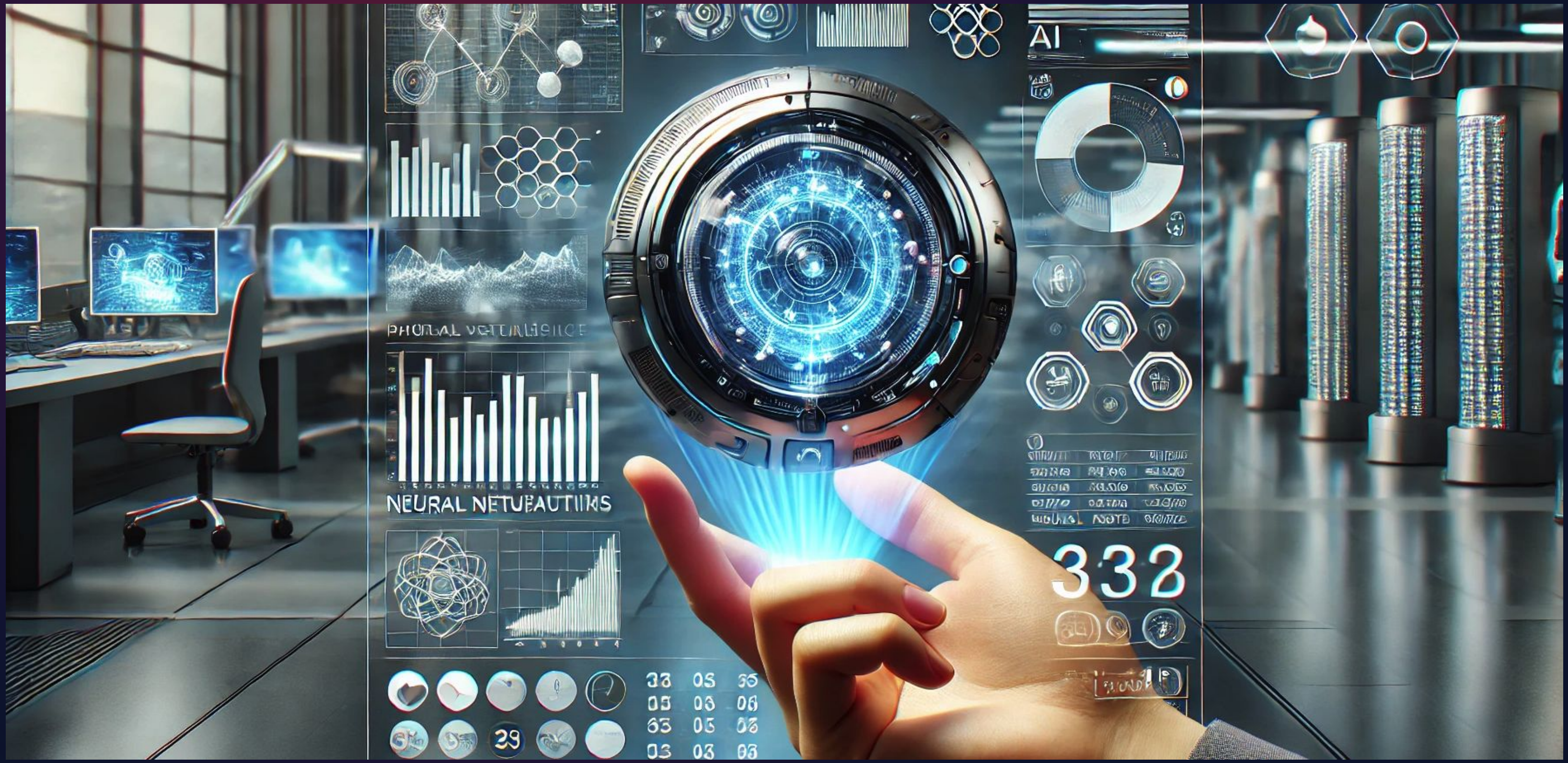


Image Attribution:
PyTorch Documentation - Adversarial Example Generation

What Tools Can I Use?

- Captum – https://github.com/pytorch/captum
- SHAP – https://github.com/shap/shap
- LIME
- ELI5
- AIX360
- Many…
- Many…
- More



Demo: Captum + NLP Classifier

https://youtu.be/geZNwLzoaT4
https://youtu.be/m0VxUAGhKcY

Demo: Captum + Vision Classifier

https://youtu.be/5J2sGIU0RV4

Adversarial Attacks: For Good… and Bad

Building Better Models via Intentional Disruption

Adversarial Attacks

- TODO
- Intentional Adversarial Attacks
    - Besides Finding Holes...
    - Disrupting Classification
        - Vision
        - NLP
- Why?
    - Unauthorized Surveillance
    - Protect Privacy
    - Obfuscation

Adversarial Strategies

Here Are Ideas/Concepts in NLP to Disrupt – Be Creative!!

- Encoding/Formatting
- Homophones and Phonetics
- Code Switching
- Low-Resource Languages
  - Navajo – "Code Talkers"
- Adversarial Spelling
- Polysemy/Multiple Meanings
- Speaking in Metaphors

Resources for Next Slide
Video Credit: Darmok, under 8 minutes (Star Trek Abridged)
Source: Star Trek: The Next Generator, Episode 102 - Darmok

Creative Communication

Demo: Read That Sentiment Wrong

https://youtu.be/CoLnvqHHN_M

Demo: One Pixel Attack

https://youtu.be/s8SHeXXAWjQ

Demo: Spoofing Real-Time Vision

https://youtu.be/b_T448UXaHw

Creative Communication

Adversarial Attack Defense

Protection Yourself From Bad Actors

Defending NLP Attacks

- Format Normalization
- Spell-Checker or Word Recognition
  - Morphology (or Subwords Tokens)
- Syntax/Grammar Checkers
- Semantic Similarity Checks
  - Synonym Encoding
- Phonetic Normalization
  - Text-to-Speech -> Speech-to-Text
- Adversarial Training:
  - Datasets w/ Noising and Typos, Synonyms, Phrase Diversity

Image Attribution
How Chess Works - https://entertainment.howstuffworks.com/leisure/brain-games/chess1.htm
Meredith Bower

Defending Vision Attacks

- Adversarial Training
  - Fast Gradient Sign Method (FGSM)
  - Projected Gradient Descent (PGD)
- Spatial Smoothing (Blurring)
  - Median Filtering (3x3 -> 1x1)
  - Gaussian Blur
  - Non-local Means, Bilateral Filters
- Feature Squeezing, Randomization
  - Bit-Depth Reduction
  - Random Resize/Pad, Add Noise

Image Attribution
https://hmmboutique.com - https://hmmboutique.com/products/bamboo-checkers-set

Non-Specific Defenses

- Adversarial Detection: Multiple Models
  - Use 2+ Different Models
- Voting Ensembles
  - Multi-Classifiers -> Majority Wins
- Reject On Low Confidence
  - Multi-Pass w/ Slight Variation
    - Drop Character
    - Swap Synonym
- EXPENSIVE and SLOW! -> More GPUs + Passes

Image Attribution
https://www.gamesforyoungminds.com/blog/2018/3/23/othello-review

Demo: Defending Adversarial NLP Attacks

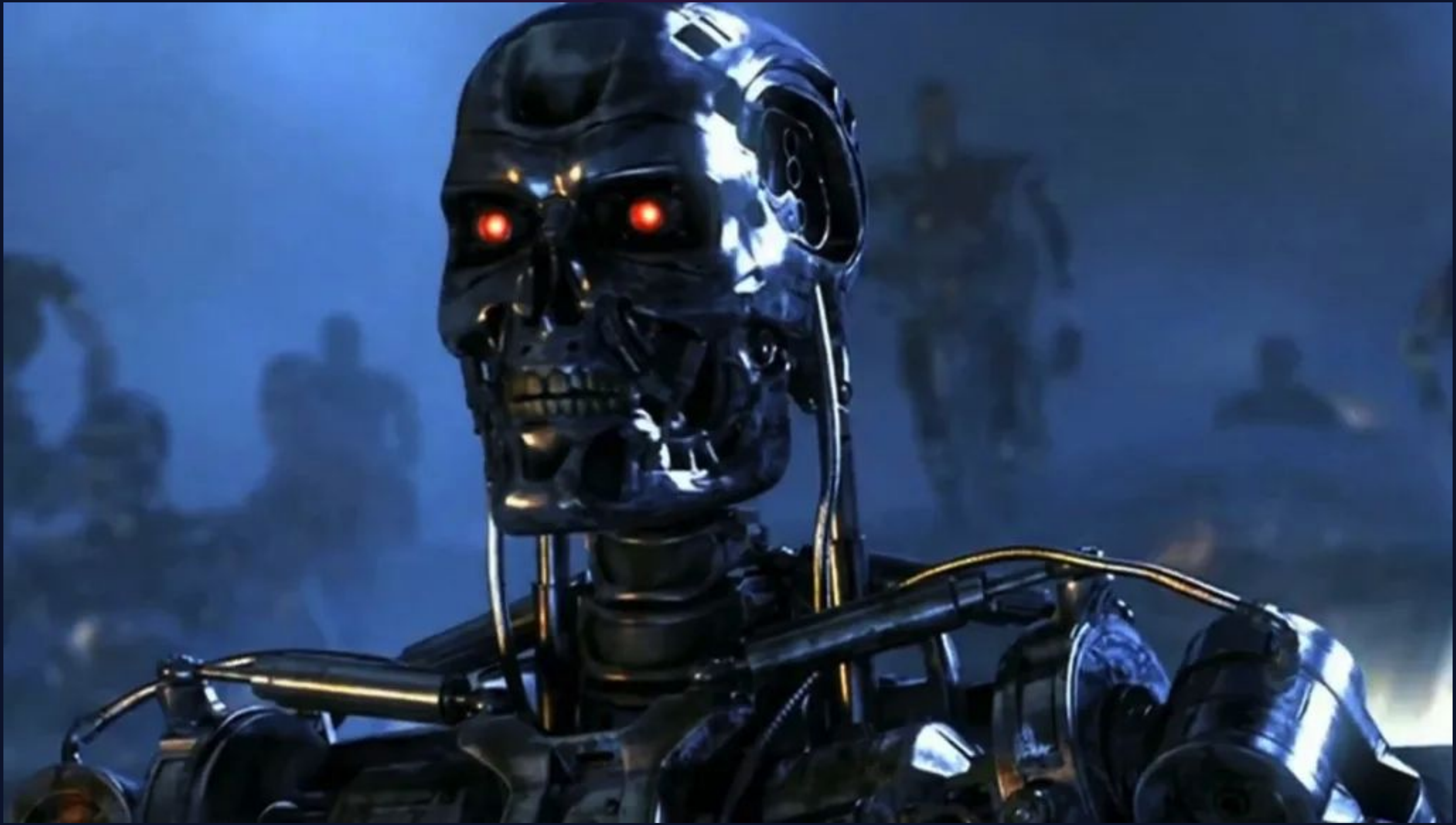
https://youtu.be/HB1RaL2OIQA

Demo: Defending Adversarial Vision Attacks

https://youtu.be/dLU5mBAt9qk

Why?

Just In Case…



Resources

Resources

All Materials/Code:
github.com/davidvonthenen/2025-we-are-developers

Let's Chat on Discord: discord.gg/NetApp

- Captum:
  - GitHub – https://github.com/pytorch/captum
  - Tutorials – https://captum.ai/tutorials/
- PyTorch:
  - GitHub – https://github.com/pytorch/pytorch
  - Tutorials – https://pytorch.org/tutorials/index.html

WeAreDevelopers
World Congress

NetApp

Thank You!

David vonThenen
Senior AI/ML Engineer
@davidvonthenen