DEVOXX France

DigitalOcean

Explaining the Unexplainable

Python Tools for AI Transparency Using Captum

David vonThenen

@davidvonthenen

David vonThenen

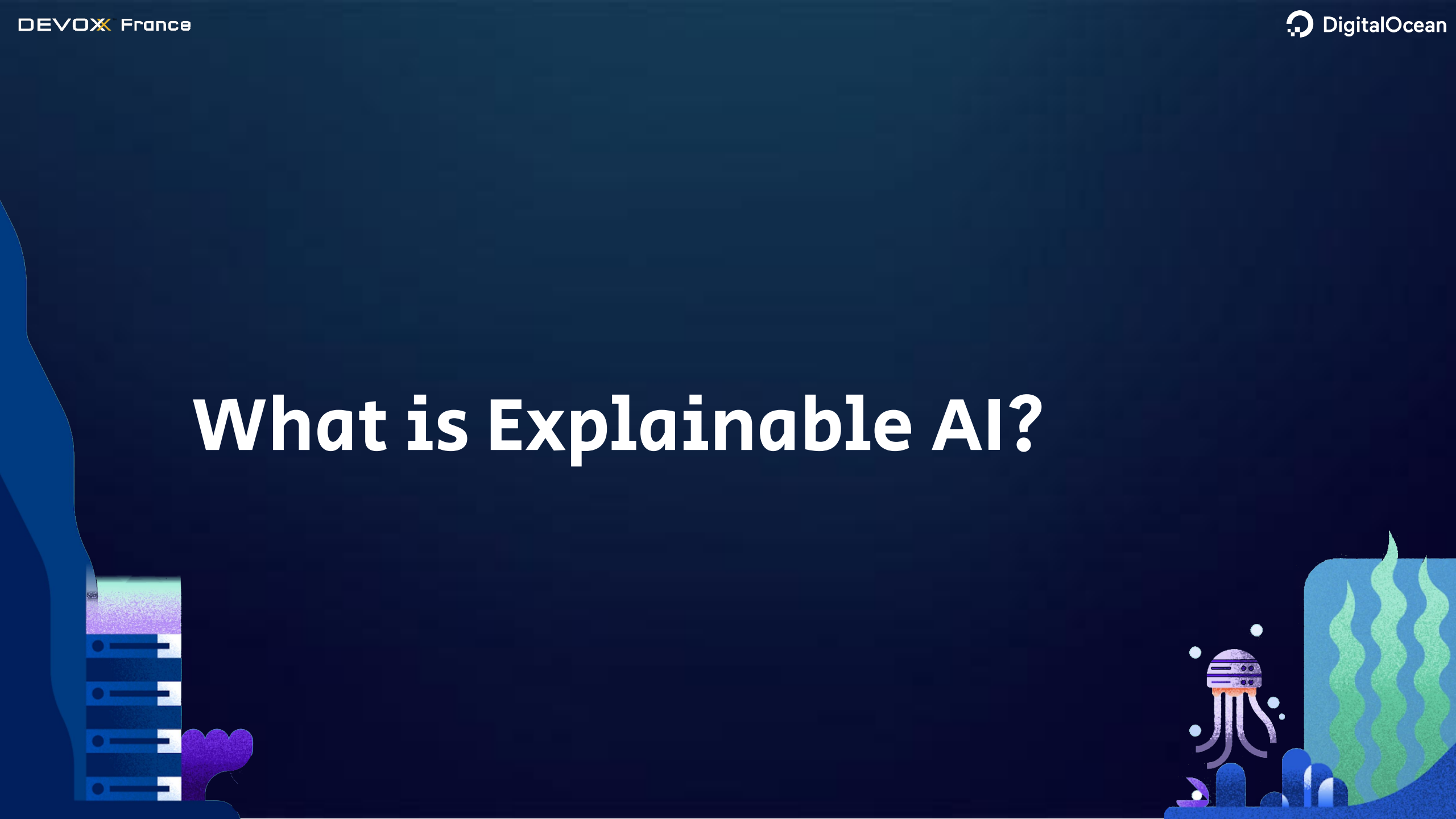
- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Cloudy, cloudy, cloudy...
- There is storage for that!

@davidvonthenen

Agenda

- What is Explainable AI?
- Understanding Data Inconsistencies
- Dataset Observability and Diagnostics
  - Demos, Demos, Demos
- Adversarial Attacks for Good… & Bad
  - Demos, Demos, Demos
- Q&A

What is Explainable AI?



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
  - Many, Many, Many More

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)

Explainable AI

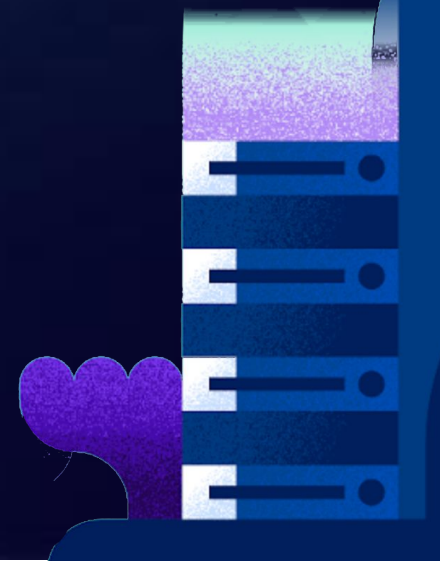
Why Do We Care?

- Transparency Build Trust
- Debugging –> Improvement
- Compliance and Ethics

Key Goals:

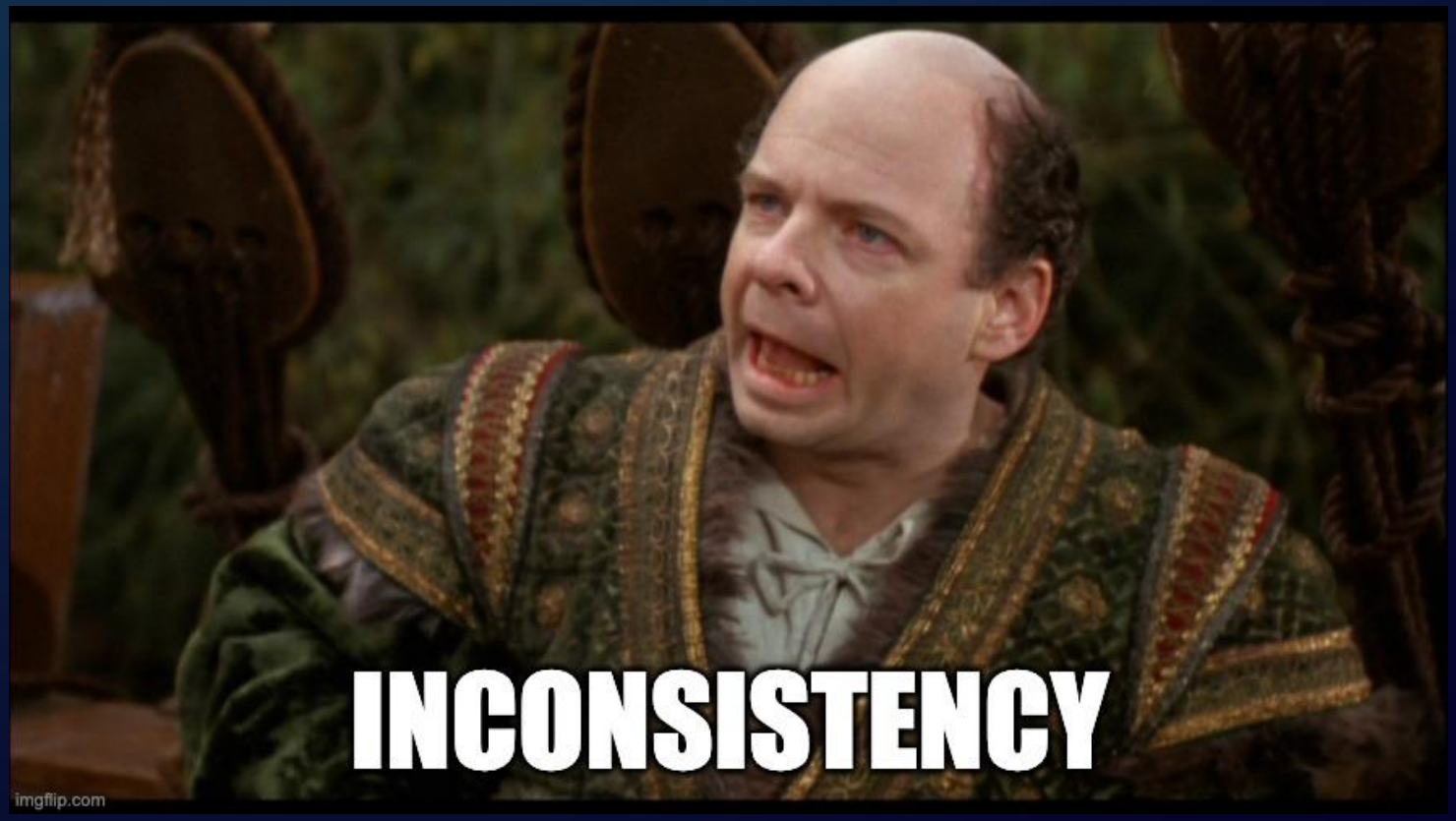
- Interpretability
- Accountability
- Fairness + Bias Detection

Understanding Data Inconsistencies



Data Inconsistencies Matter

- AI "Decision Making" Directly Shaped By Data
  - Annotation Errors
  - Data Bias
  - Distribution Drift
  - Adversarial Input
- Real World – Accidental, Unintended Consequences



Annotation Errors

Data Bias

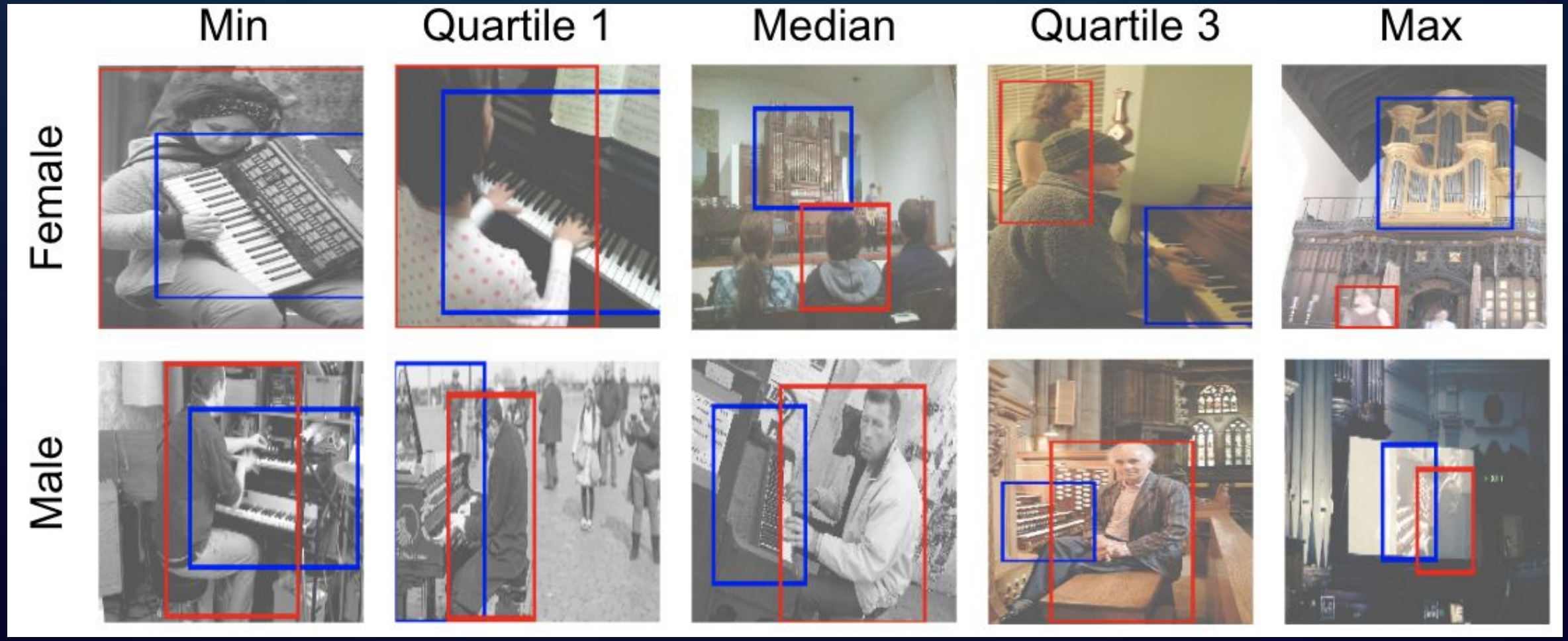


Image Attribution
Princeton University Researcher - Tool helps clear biases from computer vision
Molly Sharlach on October 1, 2020

Data Imbalance

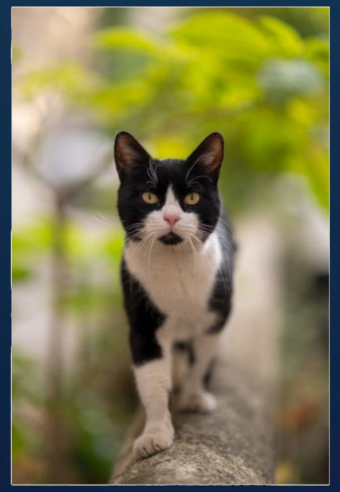
Unbalanced Dataset

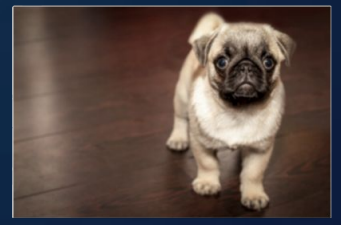
CATS

DOGS

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

Distribution Shifts



Let's Input These




Image Attribution
Dive Into Deep Learning
Nikolaj Buhl on November 11, 2022

Adversarial Samples

$+ .007 \times$

$=$

x

"panda"
57.7% confidence

$\text{sign}(\nabla_x J(\theta, x, y))$

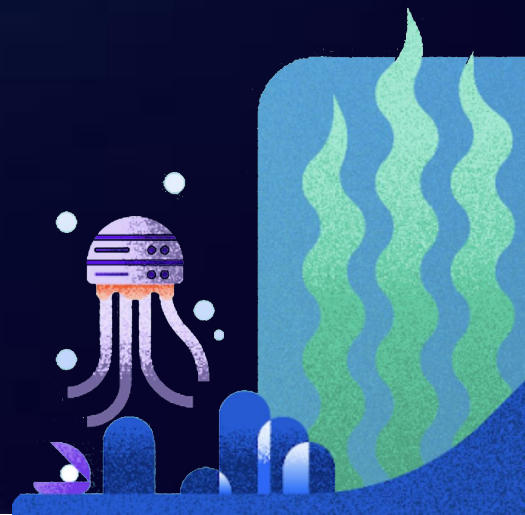
"nematode"
8.2% confidence

$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$

"gibbon"
99.3 % confidence

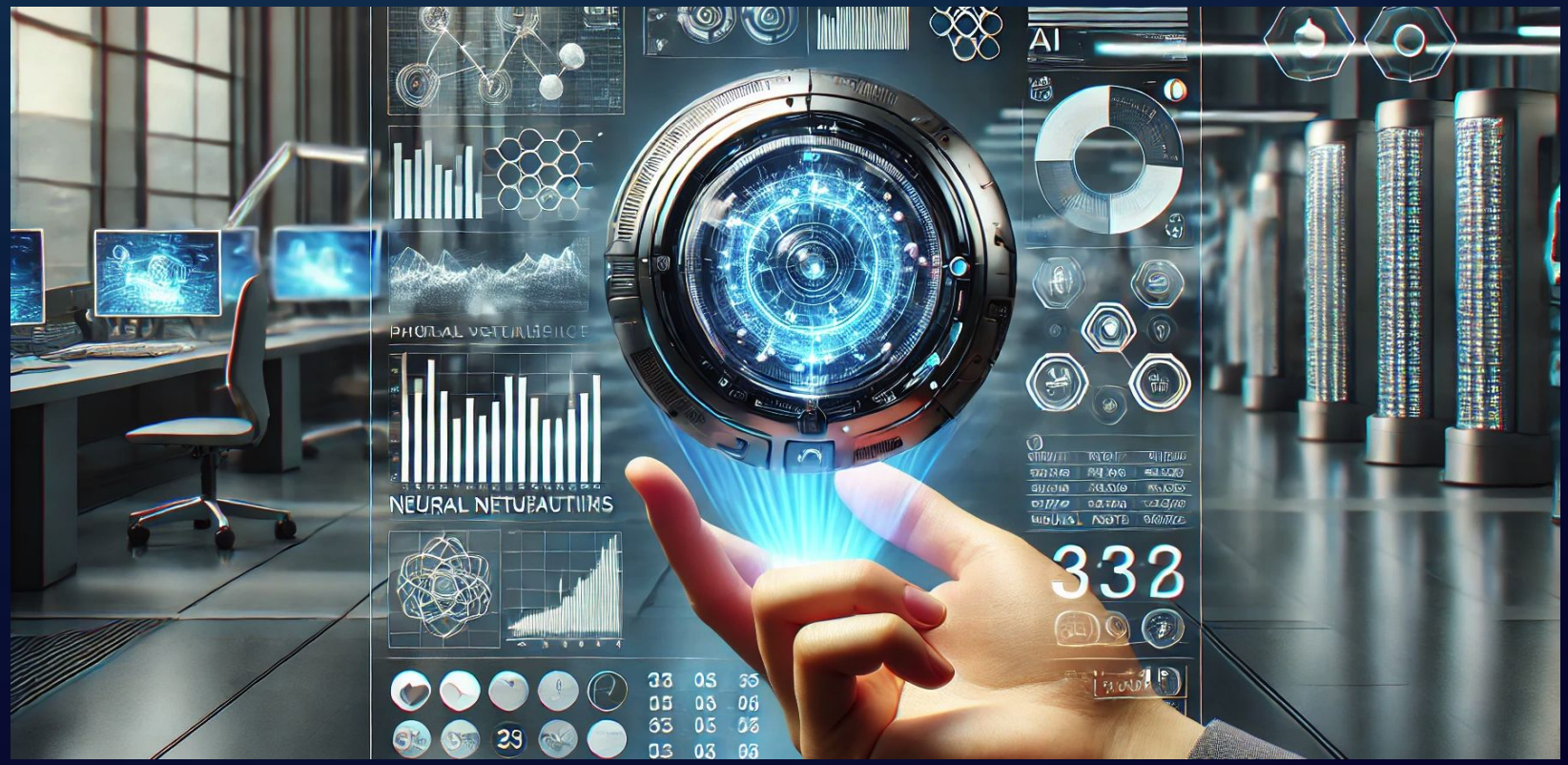
Image Attribution:
PyTorch Documentation - Adversarial Example Generation

Dataset Observability And Diagnostics



What Tools Can I Use?

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

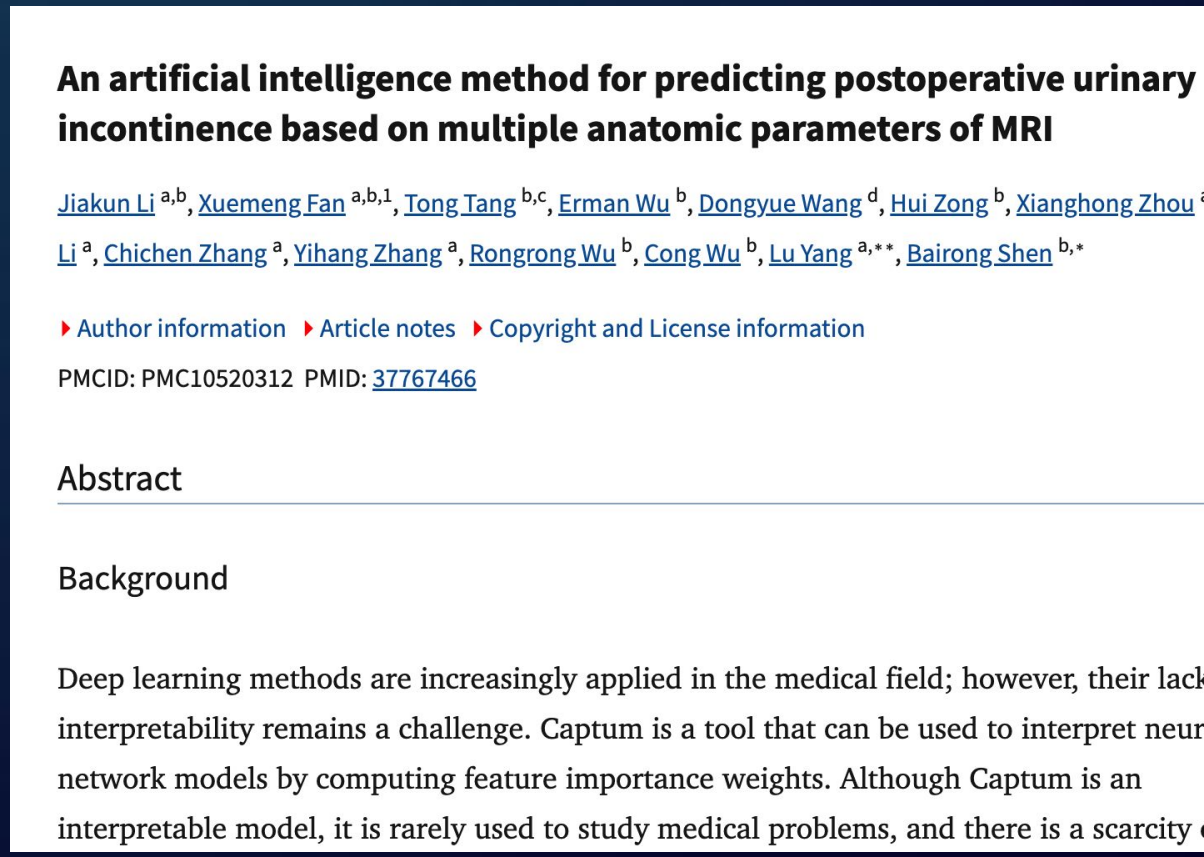


Let's Take a Look at Captum

- Open Source PyTorch Library
  - Gradients, Saliency Maps, SHAP
  - Layer/Neuron Contributions
  - NLP, Vision
- Detects:
  - Biases
  - Inconsistency
  - Hidden Patterns

Captum: Case Study

- Study: Urinary Incontinence
- Captum Revealed Findings:
  - Validated Contributions
  - Discovered 3 Features
- Future Application:
  - Update Surgical Protocols
  - Improved Techniques
  - Post-Op Therapy



An artificial intelligence method for predicting postoperative urinary incontinence based on multiple anatomic parameters of MRI

Jiakun Li [a,b], Xuemeng Fan [a,b,1], Tong Tang [b,c], Erman Wu [b], Dongyue Wang [d], Hui Zong [b], Xianghong Zhou [a], Li [a], Chichen Zhang [a], Yihang Zhang [a], Rongrong Wu [b], Cong Wu [b], Lu Yang [a,**], Bairong Shen [b,*]

Author information · Article notes · Copyright and License information

PMCID: PMC10520312 PMID: 37767466

Abstract

Background

Deep learning methods are increasingly applied in the medical field; however, their lack interpretability remains a challenge. Captum is a tool that can be used to interpret neur network models by computing feature importance weights. Although Captum is an interpretable model, it is rarely used to study medical problems, and there is a scarcity

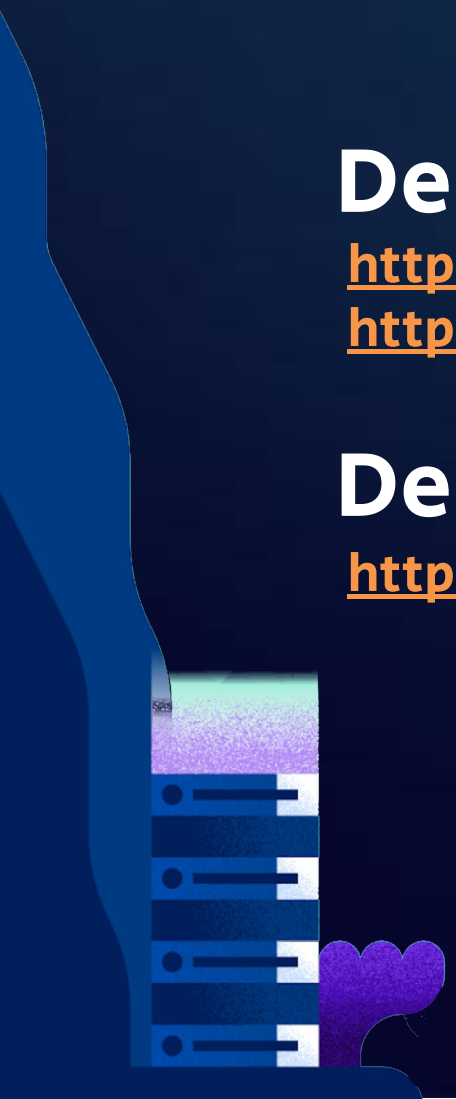
Case Study Paper
An artificial intelligence method for predicting postoperative urinary incontinence based on multiple anatomic parameters of MRI

Demo: Captum + NLP Classifier

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

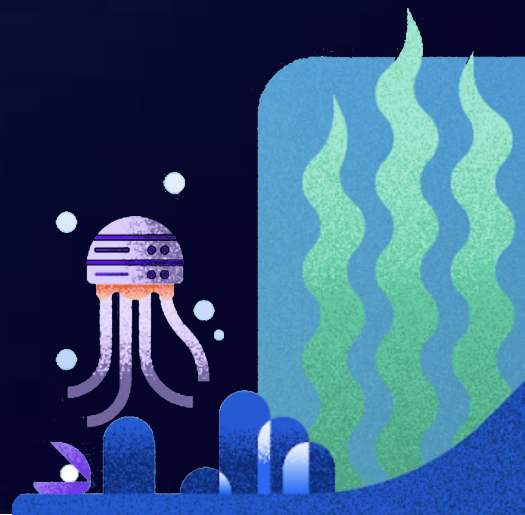
Demo: Captum + Vision Classifier

https://youtu.be/5J2sGIU0RV4



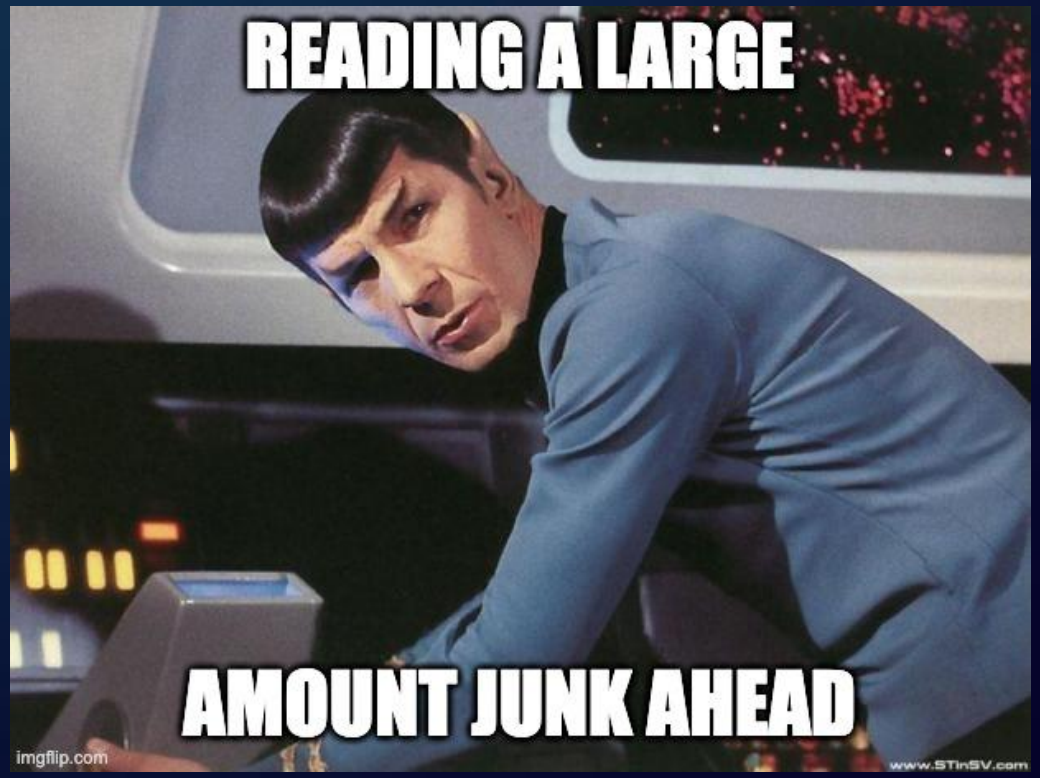

Adversarial Attacks: For Good... and Bad

Building Better Models via Intentional Disruption



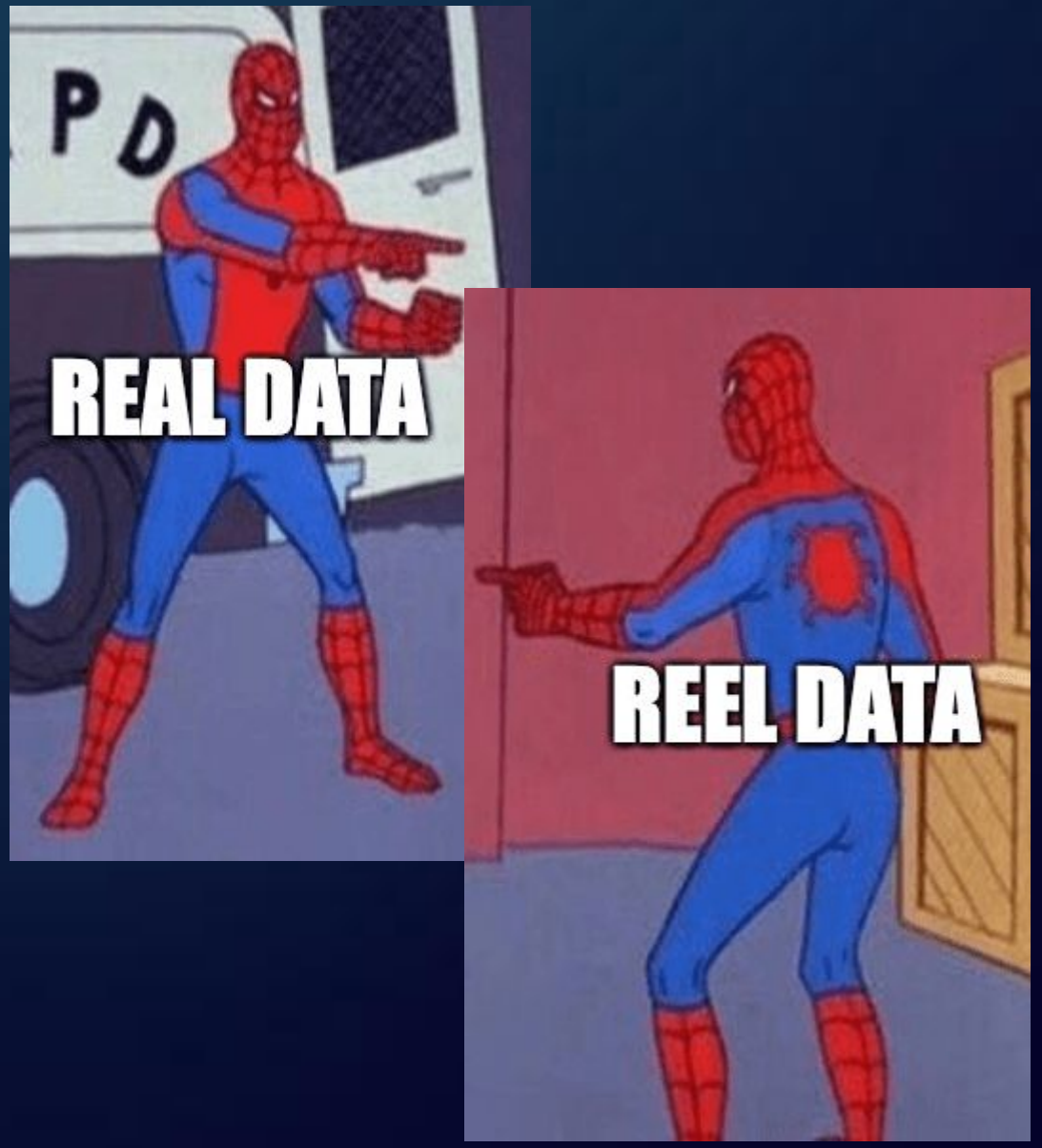
Turning Insights Into Action

- Why Explainable AI?
  - Question Rigid Assumptions
  - Finding Data Flaws
  - Expose Ethical Scenarios
  - Adversarial Testing
- Result
  - Why Exclude Data
  - Fix Problematic Data
  - Under Representation
  - Fairness



What Else...

- 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

DigitalOcean

Creative Communication

Demo: Read That Sentiment Wrong

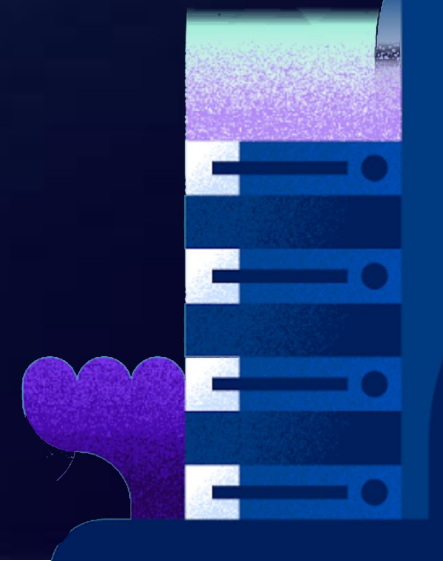
https://youtu.be/CoLnvqHHN_M

Demo: One Pixel Attack

https://youtu.be/s8SHeXXAWjQ

Demo: Spoofing Real-Time Vision

https://youtu.be/b_T448UXaHw

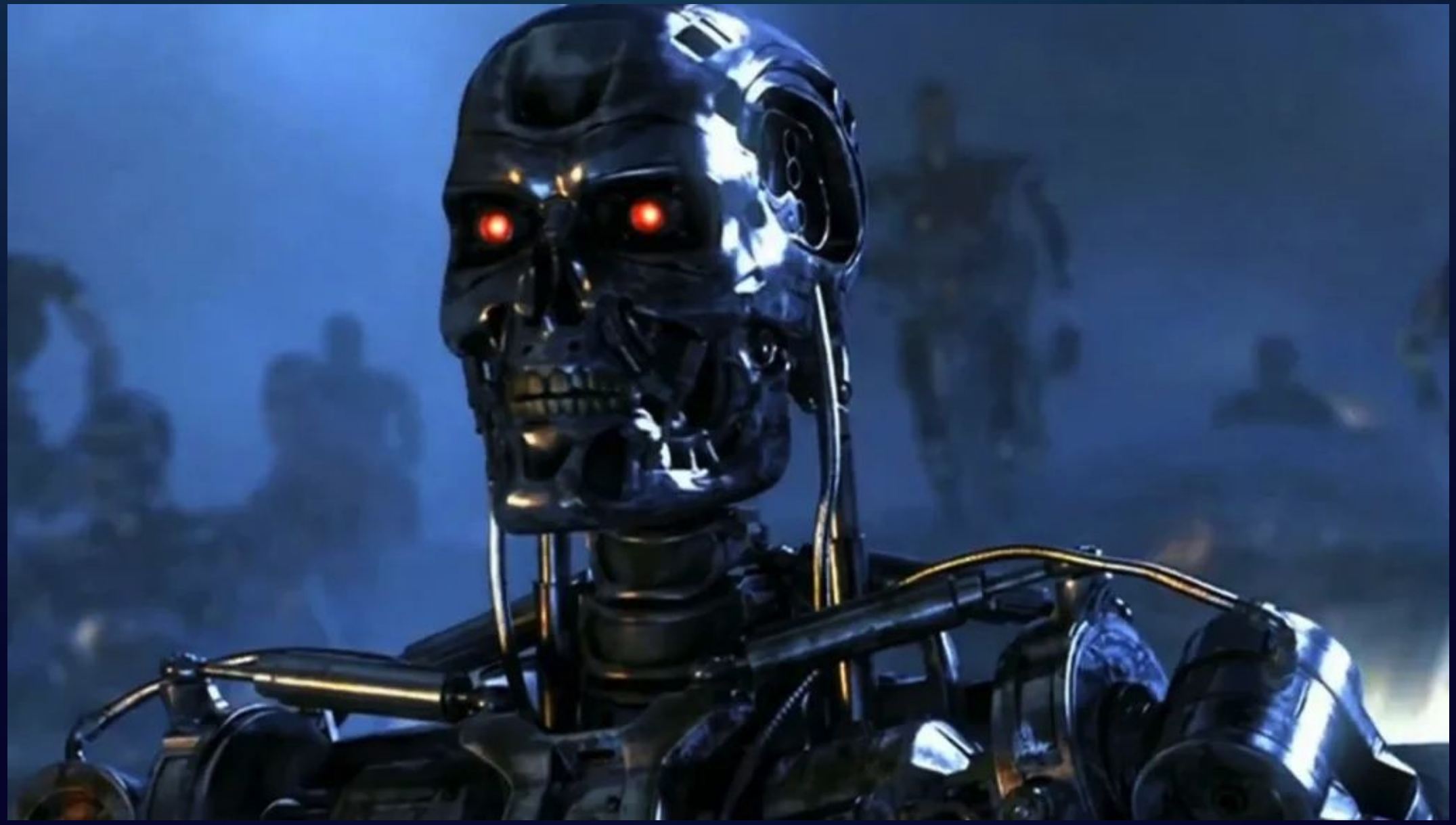


DigitalOcean

Creative Communication

Just In Case...

DigitalOcean



Resources

DigitalOcean

Resources

All Materials/Demos:: github.com/davidvonthenen/2025-devoxx-france

DigitalOcean AMD Bare Metal GPUs (MI300X) Availability
https://www.digitalocean.com/blog/now-available-bare-metal-amd-instinct-mi300x-gpus

Continue the Conversation – DigitalOcean Discord
https://discord.com/invite/digitalocean

- 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

DEVOXX France

DigitalOcean

Thank You!

David vonThenen
Senior AI/ML Engineer
@davidvonthenen