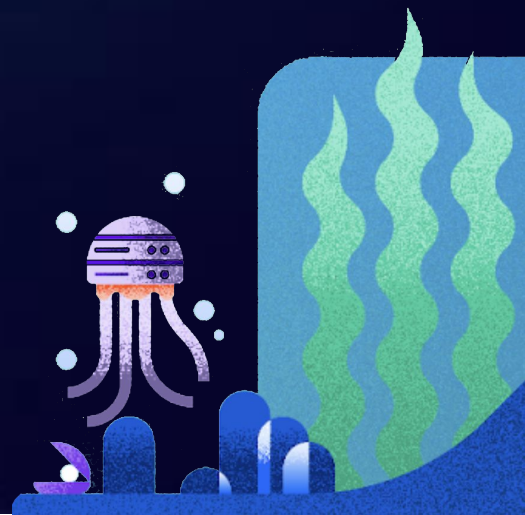
DigitalOcean

Crack the AI Black Box

Practical Techniques for Explainable AI

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!

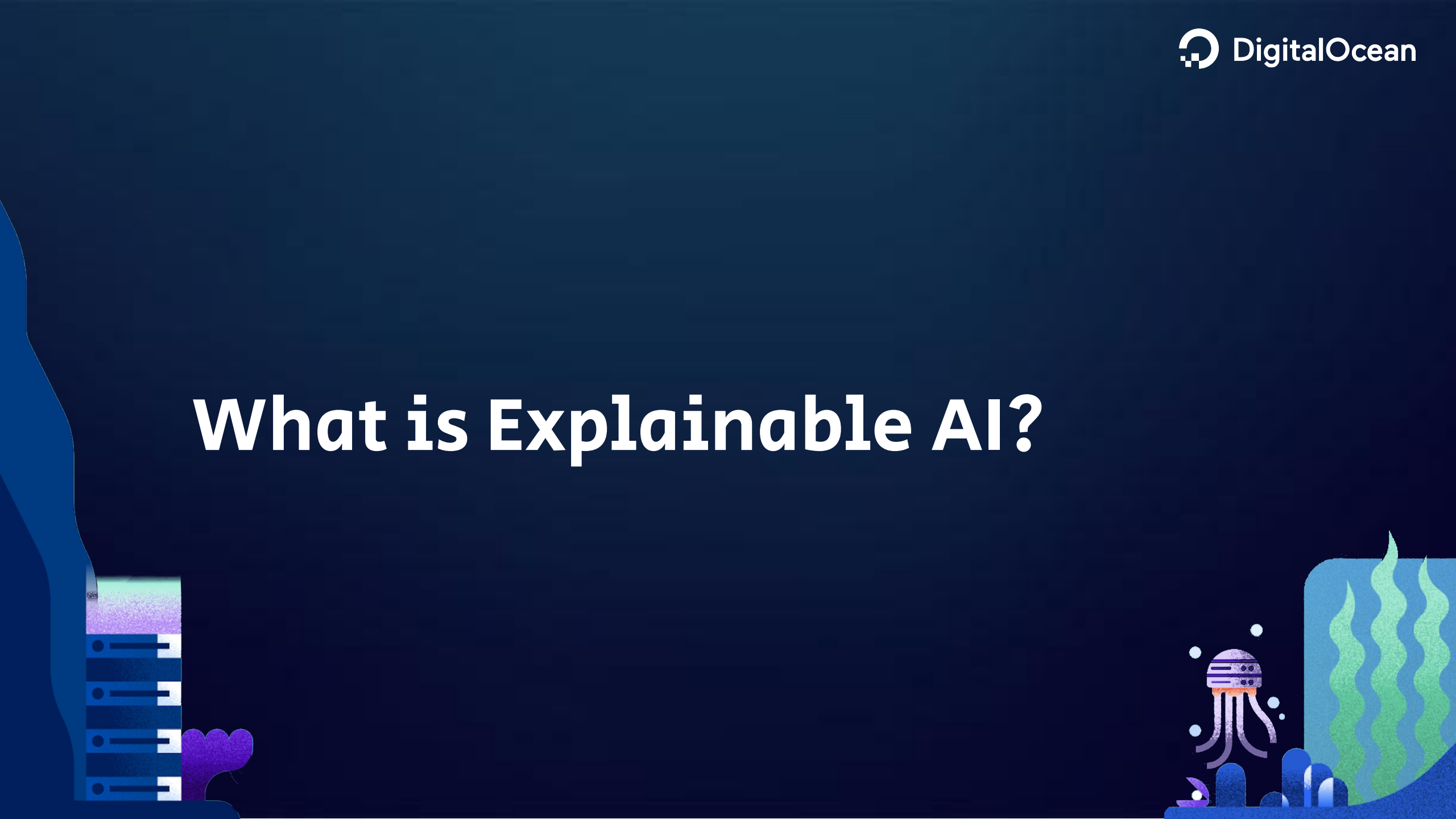
@davidvonthenen

Agenda

- What is Explainable AI?
- Understanding Data Inconsistencies
- Dataset Observability and Diagnostics
  - Demo: NLP
  - Demo: Vision
- Adversarial Attacks for Good... & Bad
- Let's Continue the Discussion!

DigitalOcean

What is Explainable AI?



Flawed Data

- AI/ML Only As Good As the Data
  - What If the Data is Flawed?
- Real-World Examples:
  - Recruiter AI + Male Skewed
  - Offensive AI Chatbot
  - Court Case Hallucinations
  - Home Purchasing Spree
  - Scan Sitting Up or Down
  - Many, Many, Many More

Explainable AI

Why Do We Care?

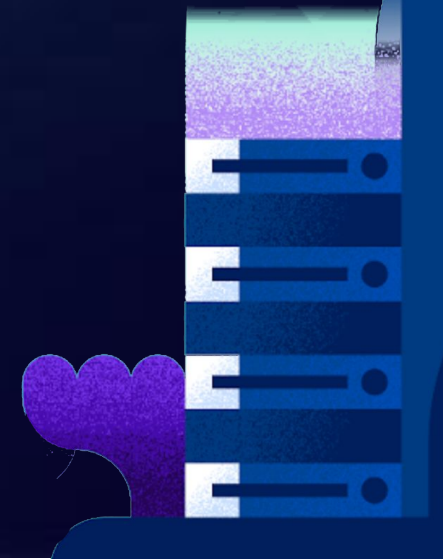
- Trust and Transparency
- Debugging + Improvement
- Compliance and Ethics

Key Goals:

- Interpretability
- Accountability
- Fairness + Bias Detection

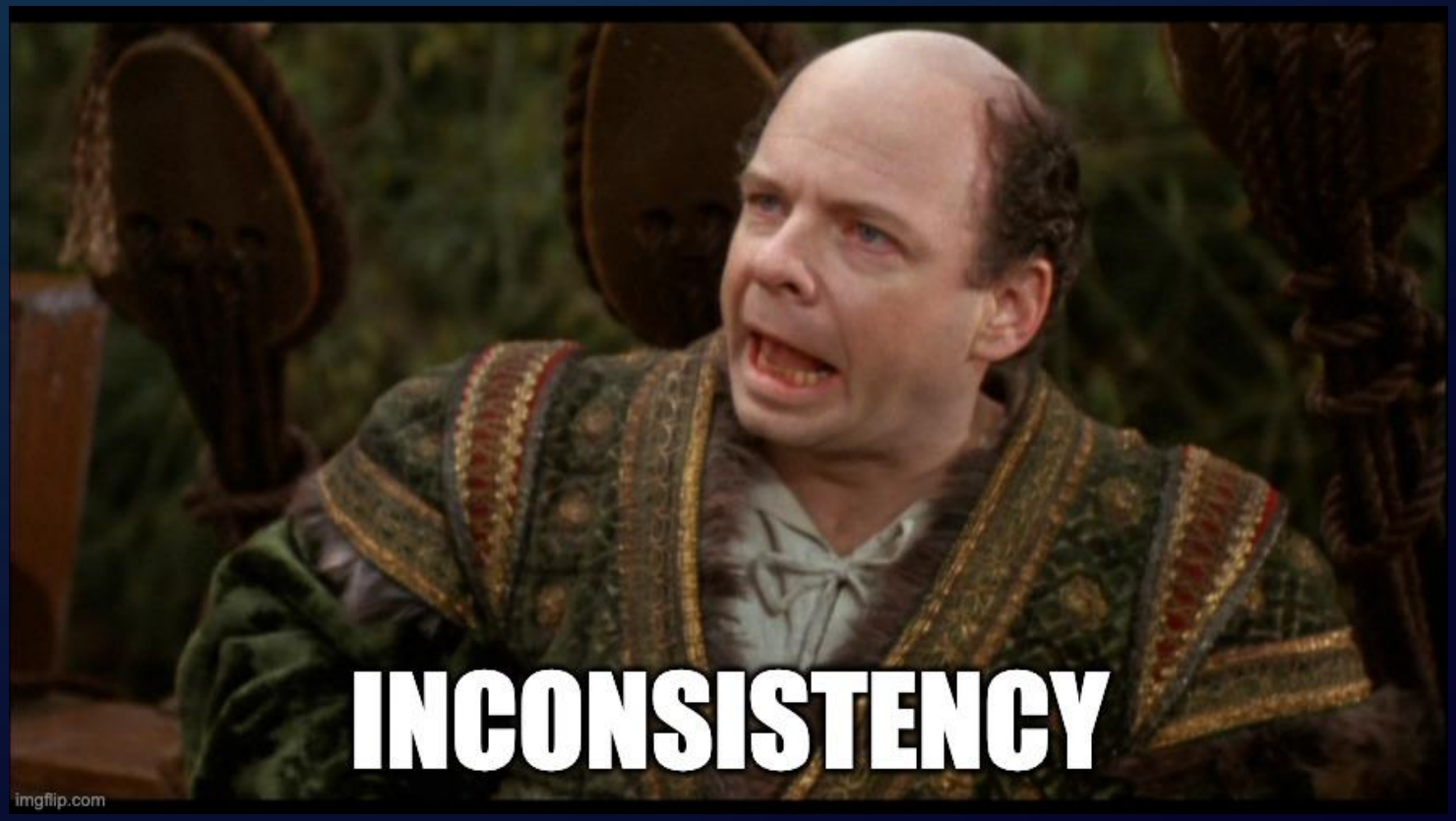
DigitalOcean

Understanding Data Inconsistencies



Data Inconsistencies Matter

- Impacts AI "Decision Making"
  - Annotation Errors
  - Data Bias
  - Distribution Drift
  - Adversarial Input
- Accidental, Unintended Consequences



Annotation Errors

Data Bias

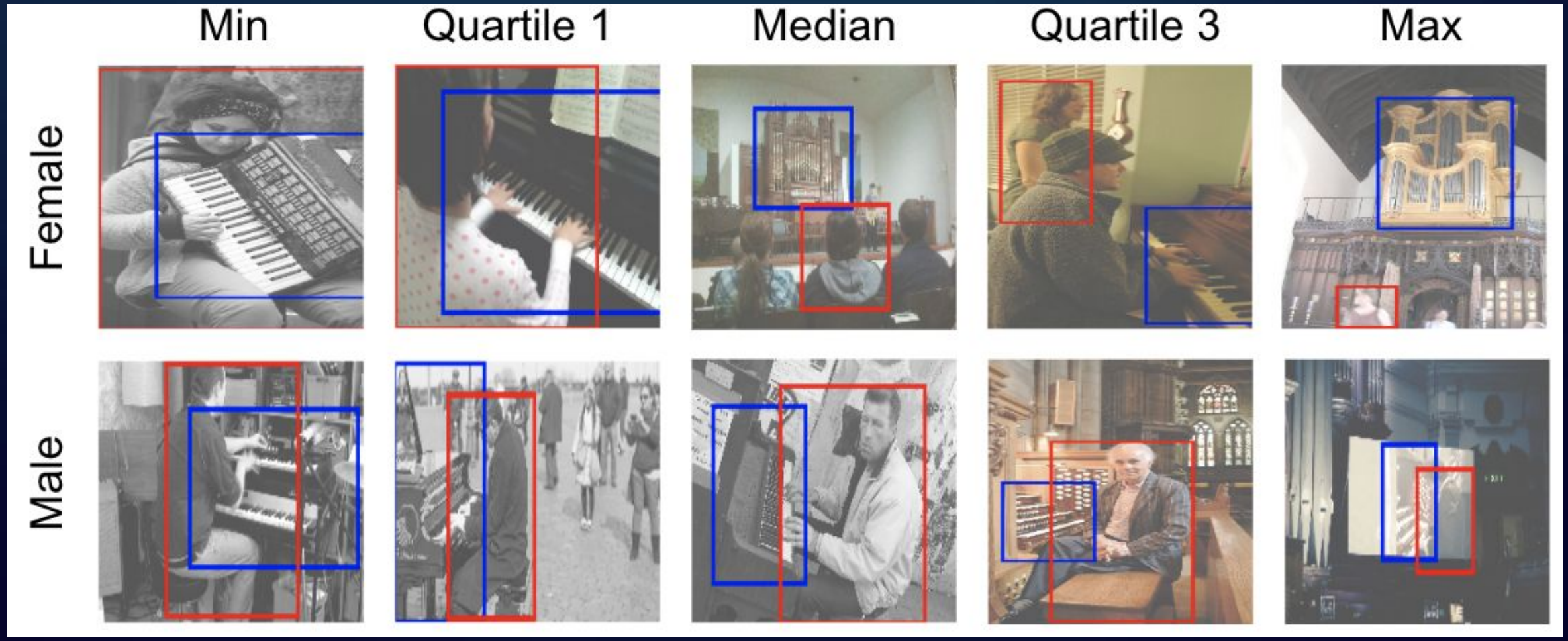


Image Attribution
Princeton University Researcher - Tool helps clear biases from computer vision
Molly Sharlach on October 1, 2020

Data Imbalance

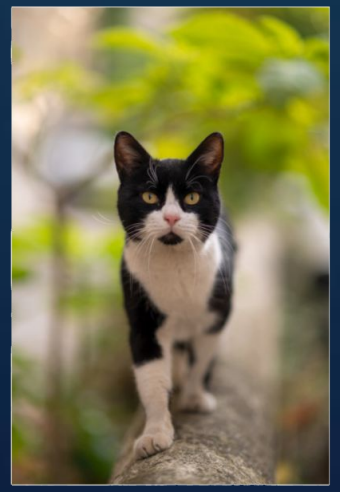
Unbalanced Dataset

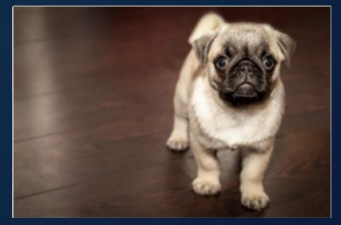
CATS

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

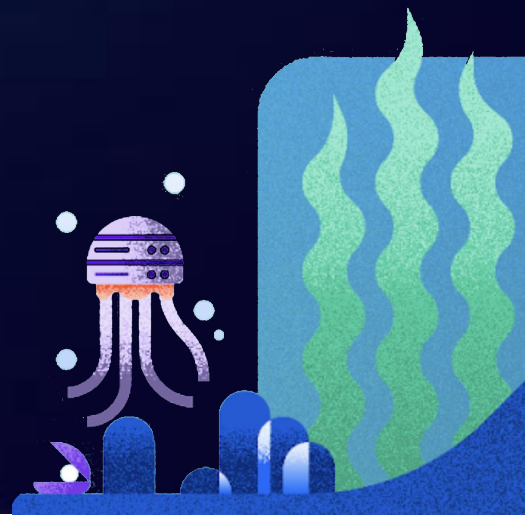
$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$

"gibbon"
99.3 % confidence

Image Attribution:
PyTorch Documentation - Adversarial Example Generation

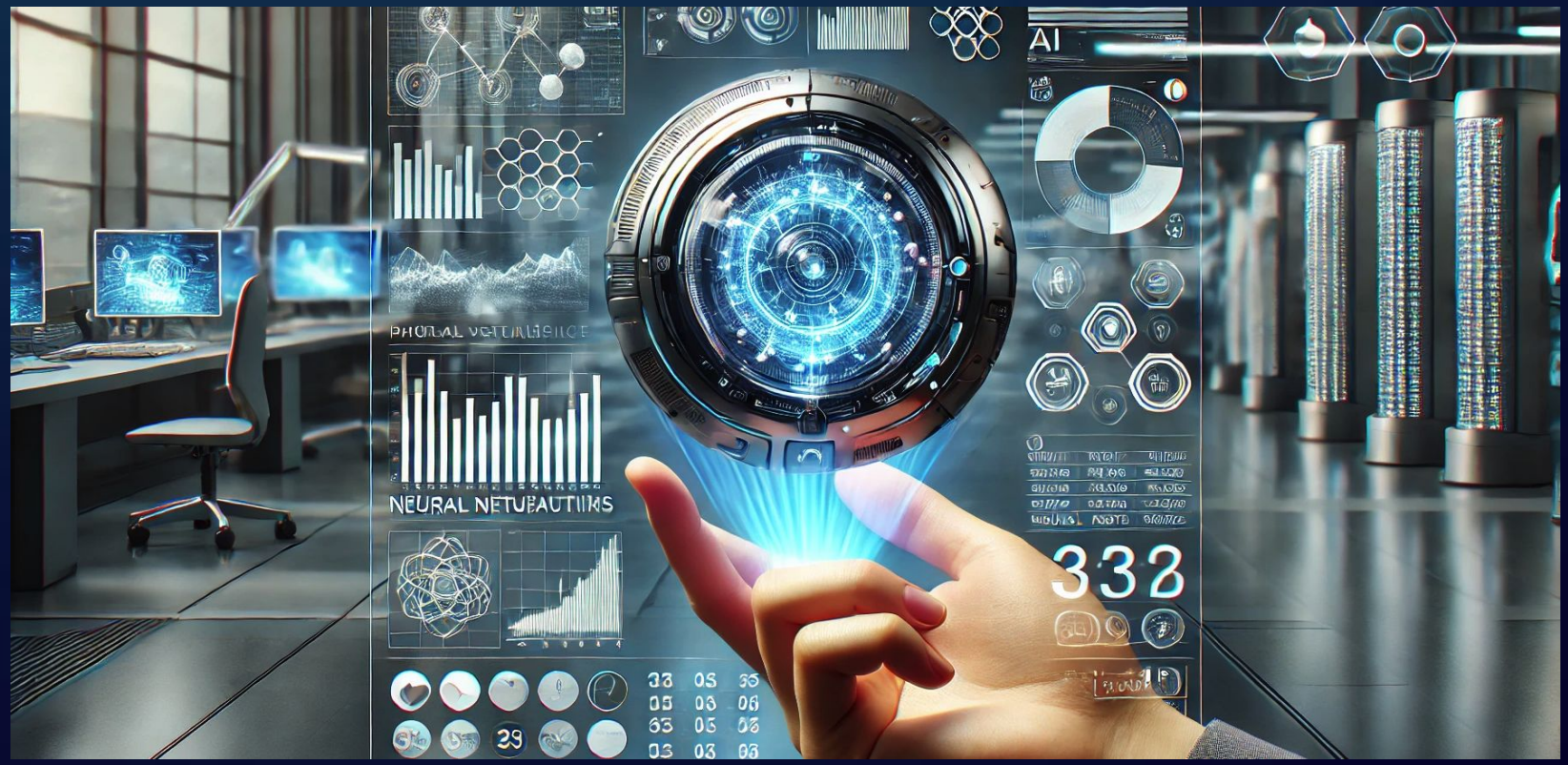
DigitalOcean

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 More

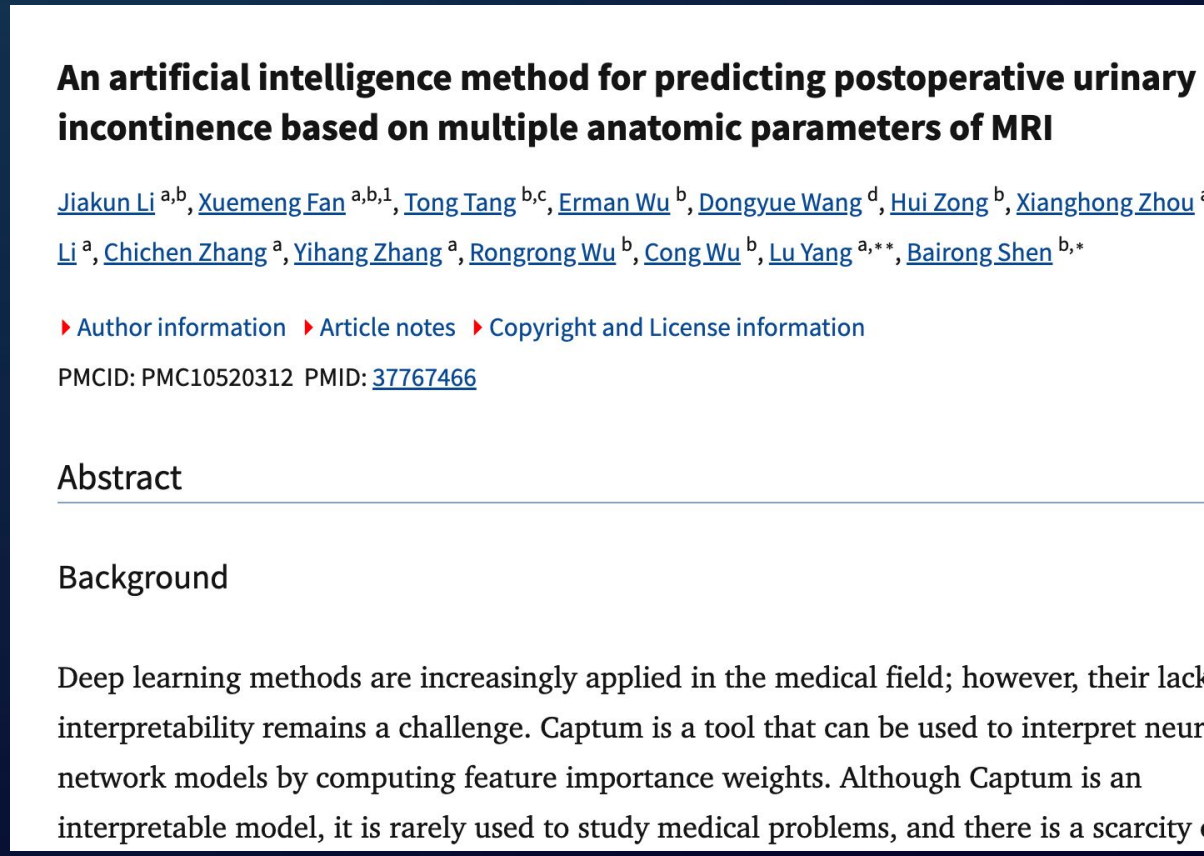


Let's Take a Look at Captum

- 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

DigitalOcean

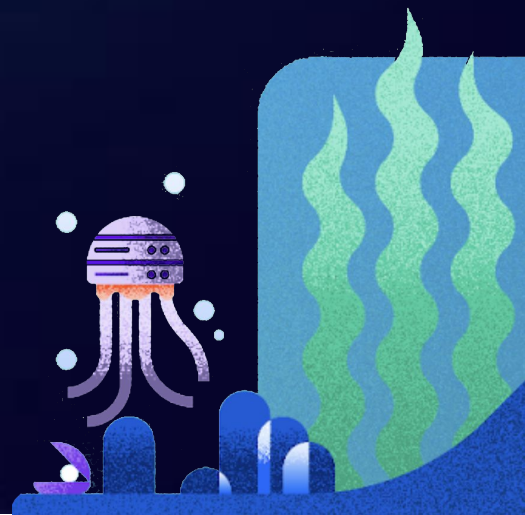
Demo: Captum + NLP Classifier

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

DigitalOcean

Demo: Captum + Vision Classifier

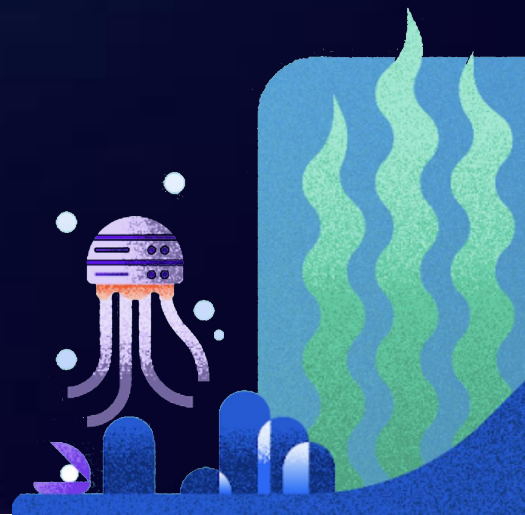
https://youtu.be/5J2sGIU0RV4



DigitalOcean

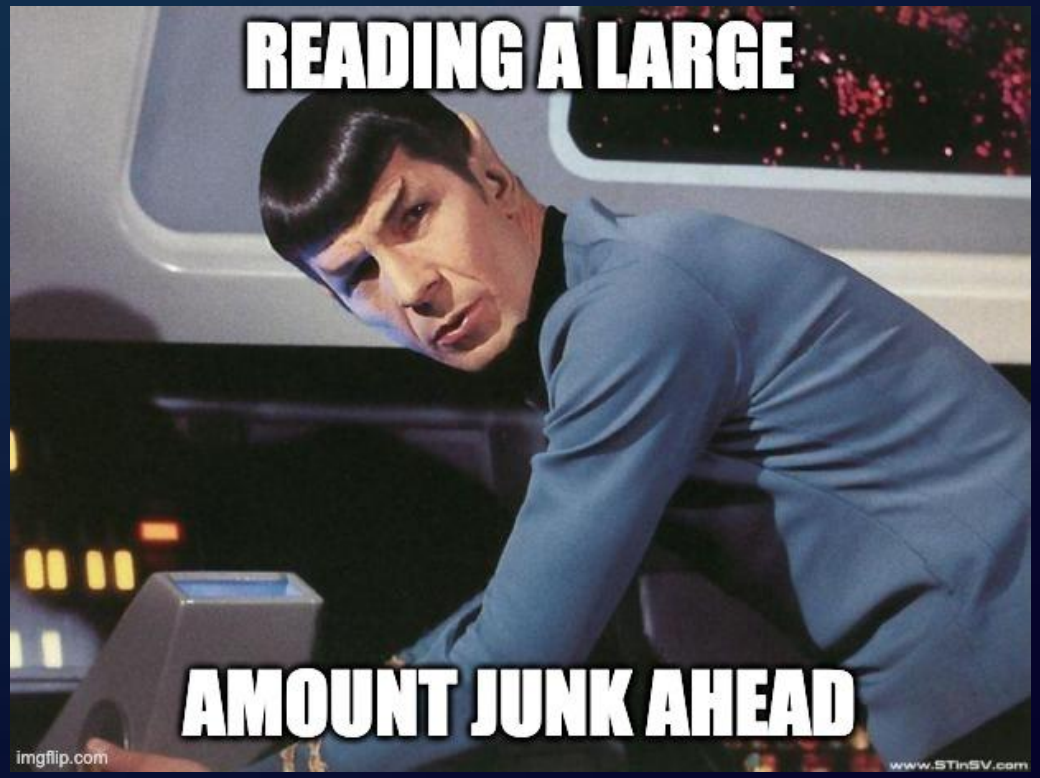
Adversarial Attacks: For Good... and Bad

Building Better Models via Intentional Disruption



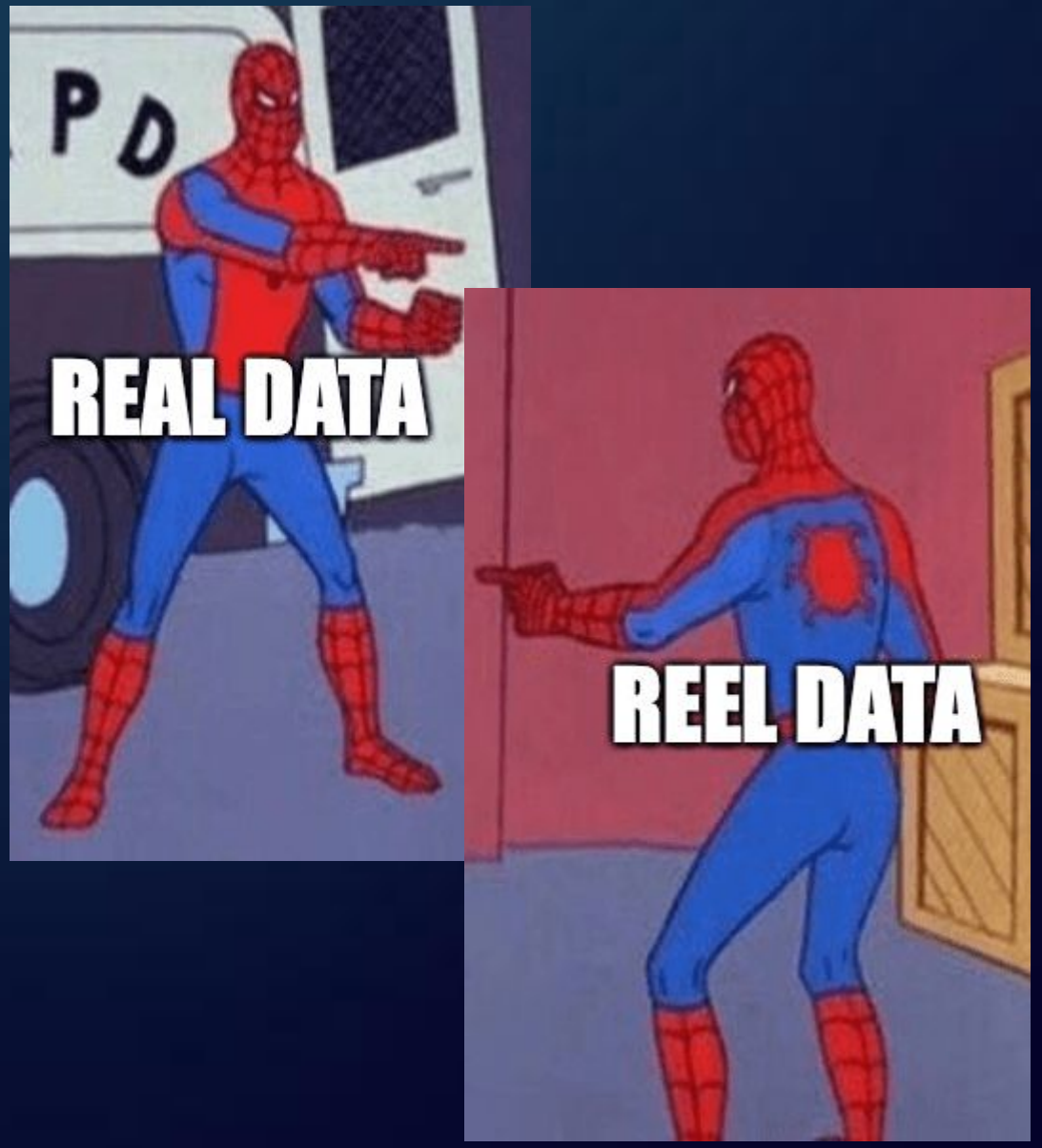
Turning Insights Into Action

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



What Else...

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



Adversarial Strategies

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

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

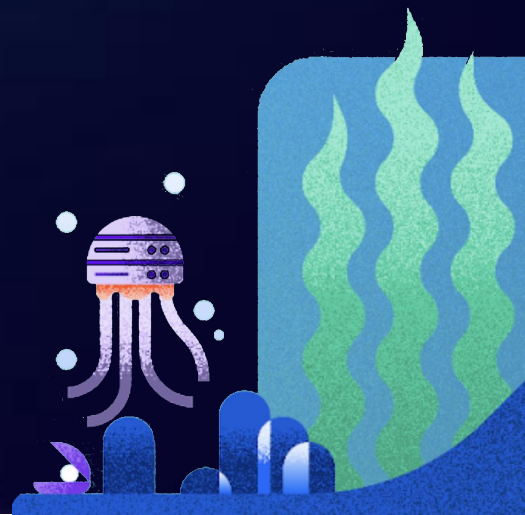
Creative Communication

DigitalOcean

DigitalOcean

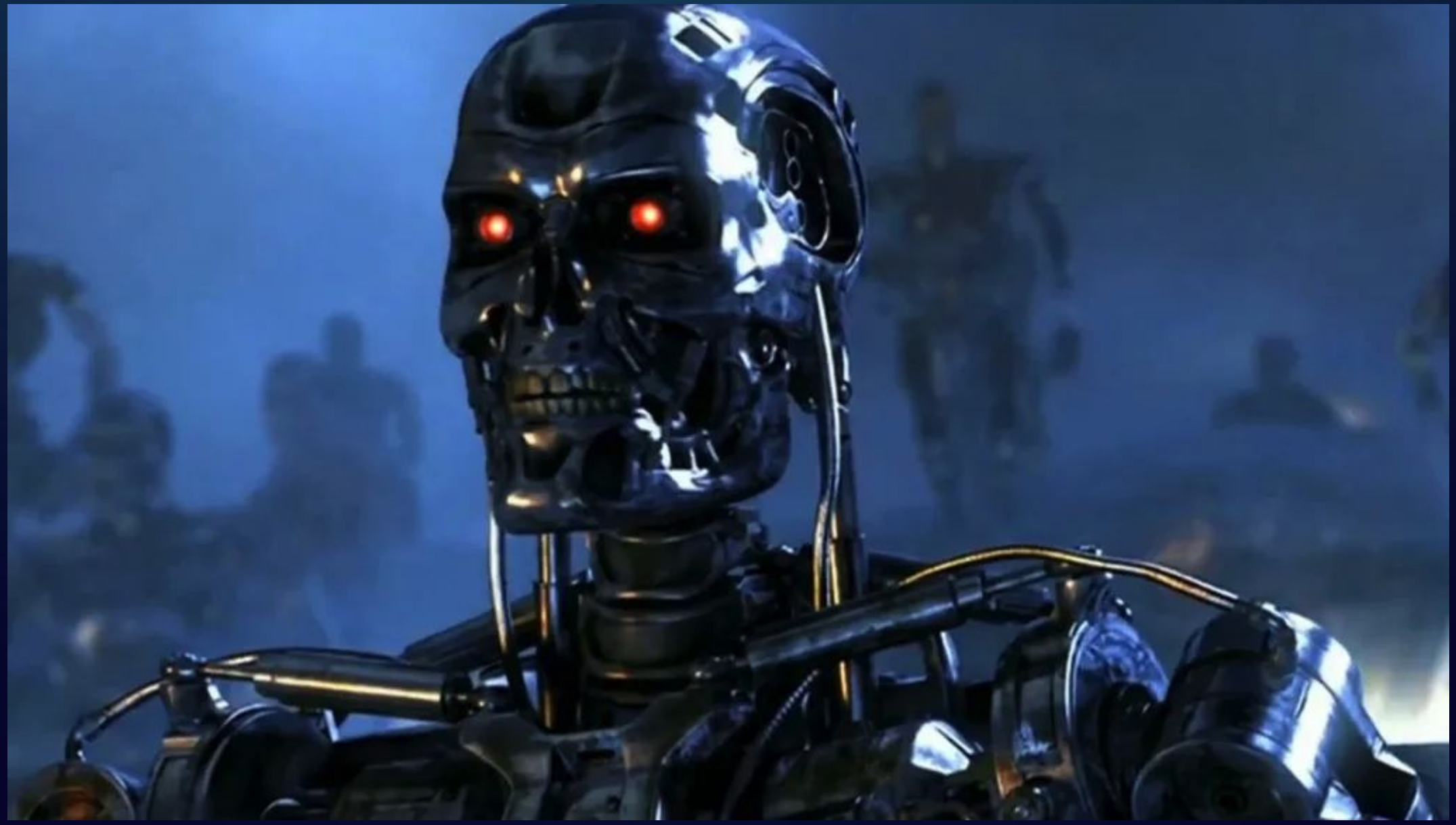
Demo: Spoofing Recognition

https://youtu.be/b_T448UXaHw



Just In Case...

DigitalOcean



DigitalOcean

Thank You!

David vonThenen
Senior AI/ML Engineer
@davidvonthenen