Visual Explainability in Machine Learning

Lecture 1: Introduction to Explainable AI

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu

Dec 5, 2023

Short Course Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/sps-education-short-course/

{alregib, mohit.p}@gatech.edu

Title: Visual Explainability in Machine Learning

Presented by: ***Ghassan AlRegib, and Mohit Prabhushankar***

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)

School of Electrical and Computer Engineering

Georgia Institute of Technology, Atlanta, USA

https://alregib.ece.gatech.edu/

Accessible Explainability for All

- Impress on the importance of Explainability in AI systems as a function of humans (users, engineers, researchers, and policymakers) requiring it
- Define Explainability and characterize it based on its required properties, methodologies and the intended audience it caters to
- Detail popular visual explanatory techniques across multiple data modalities including natural images, biomedical and seismic images, and videos
- Expand on subjective and objective techniques to evaluate explanations
- Discuss accepted proxies for Explainability – robustness and uncertainty
- Contrast against data-specific instantiations of Explainability
- Consider alternative data and explanation-centric training regimen
- Debate on the role of Visual Explainability through the lens of causality and Generative AI

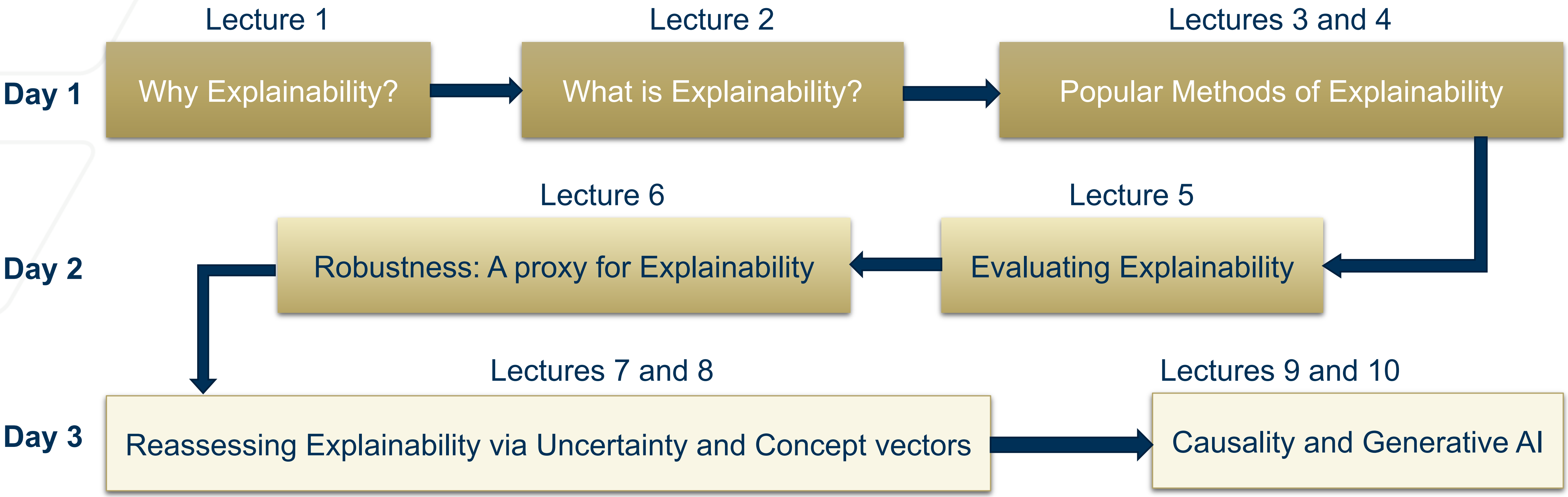
Short Course

Course Outline

- Lecture 1: Introduction to Explainable AI
- Lecture 2: Basics of Explainability in Deep Learning
- Lecture 3: Visual Explanations I
- Lecture 4: Visual Explanations II
- Lecture 5: Evaluating Visual Explanations
- Lecture 6: Robustness as Explanatory Proxy
- Lecture 7: Rethinking Explanations via Uncertainty
- Lecture 8: Concept Vectors: Utility in Training and Testing
- Lecture 9: Causality and Explainability
- Lecture 10: Generative AI and the Future of Visual Explainability

Short Course
Course Outline

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess



Short Course

Course Logistics

- 10 Lectures spanning three days
  - Day 1 (Tuesday, December 5, 2023): 4 Lectures
  - Day 2 (Wednesday, December 6, 2023): 2 Lectures
  - Day 3 (Thursday, December 7, 2023): 4 Lectures

- All course materials present at: https://alregib.ece.gatech.edu/sps-education-short-course/
- Presenter emails: {alregib, mohit.p}@gatech.edu

Lecture Outline

Lecture 1: Introduction to Explainable AI

- Artificial Intelligence
- Explainability
- Need for Explainability in AI systems
- Deep Learning
    - Training
- Foundation Models
    - Challenges in Foundation Models
- Challenges in Explainability
    - Technical Challenges
    - Functional Challenges
    - Operational Challenges
- Takeaways

Lecture Outline

Lecture 1: Introduction to Explainable AI

- Artificial Intelligence
- Explainability
- Need for Explainability in AI systems
- Deep Learning
  - Training
- Foundation Models
  - Challenges in Foundation Models
- Challenges in Explainability
  - Technical Challenges
  - Functional Challenges
  - Operational Challenges
- Takeaways

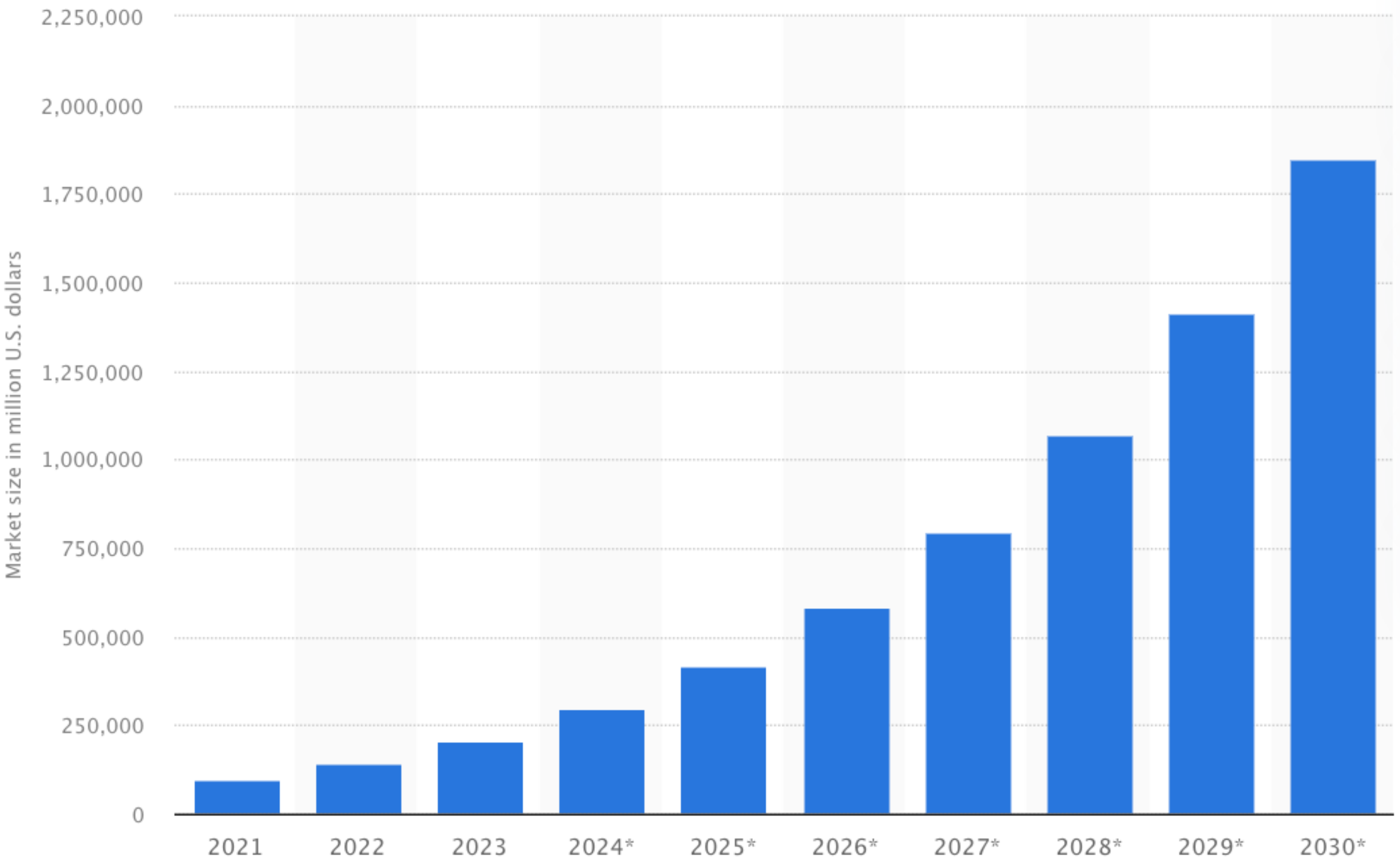
Artificial Intelligence

AI in Everyday Life

AI systems are bringing about the 4th industrial revolution

Artificial Intelligence
Projected Growth of AI Sectors

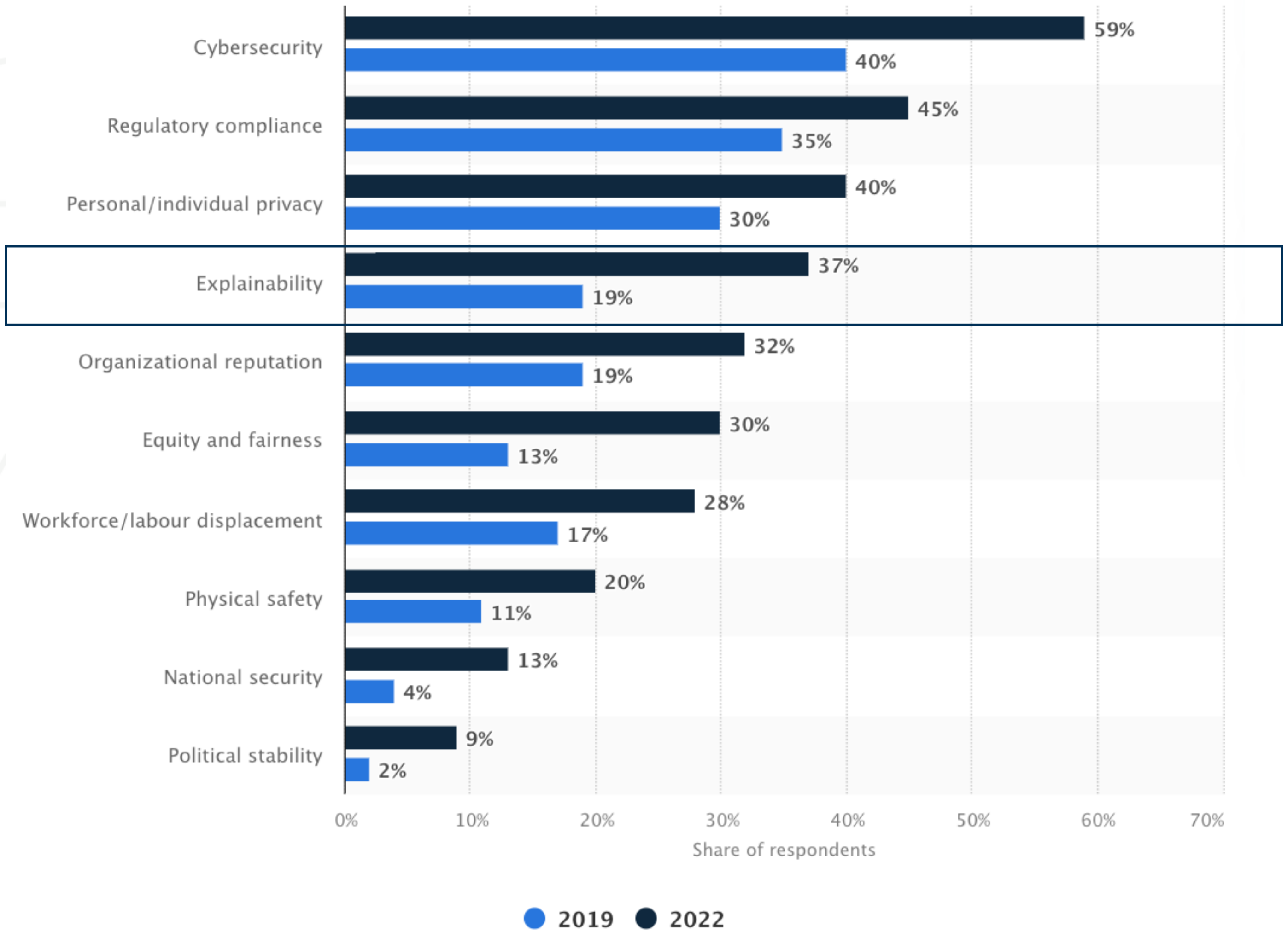
AI market size worldwide in 2021 with a forecast until 2030 *(in million U.S. dollars)*



https://www.statista.com/statistics/1365145/artificial-intelligence-market-size/

Artificial Intelligence

Public Perception of Risks in AI



- Bar graph shows public perception of risks when adopting AI systems
- 59% of respondents in 2022 believe AI adoption poses a security risk to cybersecurity as opposed to 40% in 2019
- **Explainability ranks fourth as a risk. Along with cybersecurity, it has seen the steepest increase in risk perception since 2019**
- **And since 2022, the billion parameter AI systems have nudged into the trillions**

https://www.statista.com/statistics/1365145/artificial-intelligence-market-size/

The ability of an entity to explain or justify its decisions or predictions in human-understandable terms

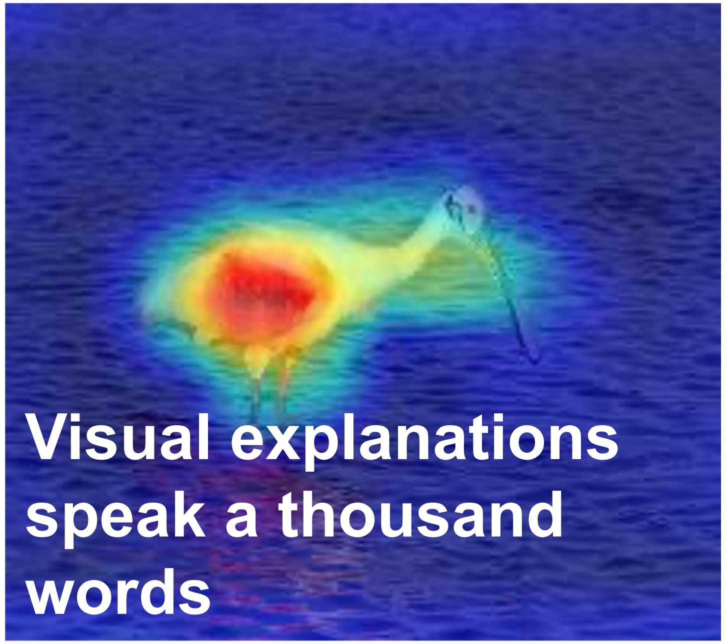
What is Visual Explainability?

Visual Explainability justifies decisions based on visual characteristics in a scene

This is a spoonbill because:

- It has a long, flat beak
- It is large and long-legged

Natural language explanation



Visual explanation

Lecture Outline

Lecture 1: Introduction to Explainable AI

- Artificial Intelligence
- Explainability
- **Need for Explainability in AI systems**
- Deep Learning
    - Training
- Foundation Models
    - Challenges in Foundation Models
- Challenges in Explainability
    - Technical Challenges
    - Functional Challenges
    - Operational Challenges
- Takeaways

Tesla driver dies in first fatal crash while using autopilot mode

The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky

Autopilot didn't detect the trailer as an obstacle (NHTSA investigation and Tesla statements)

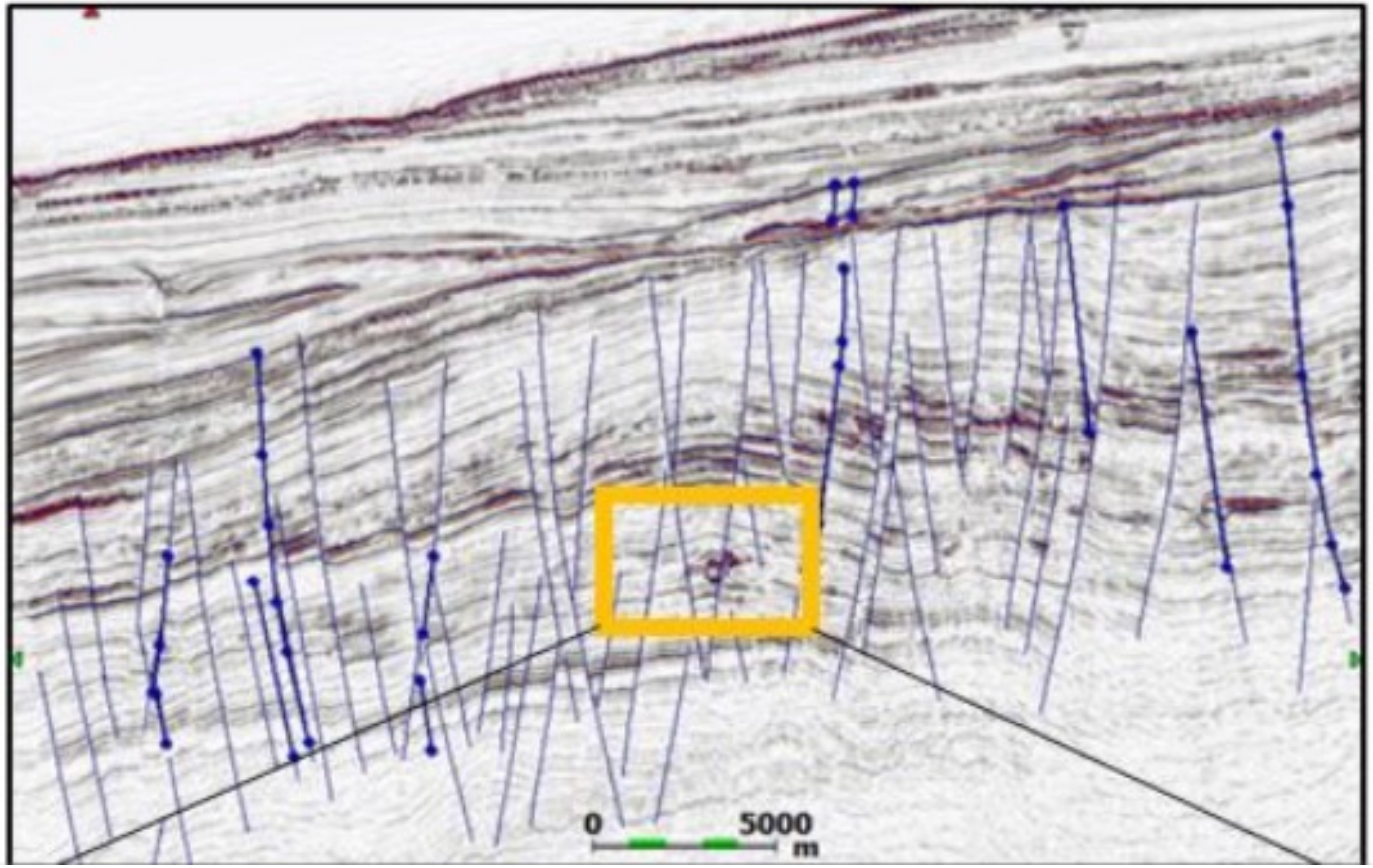
1. The National Highway Traffic Safety Administration (NHTSA) determined that a "lack of safeguards" contributed to the death
2. "Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied," Tesla said.

Explanation

Explanations in AI Systems

Case Study: Seismic Fault Interpretations

Seismic Fault interpretation is essential for earthquake monitoring and carbon capture

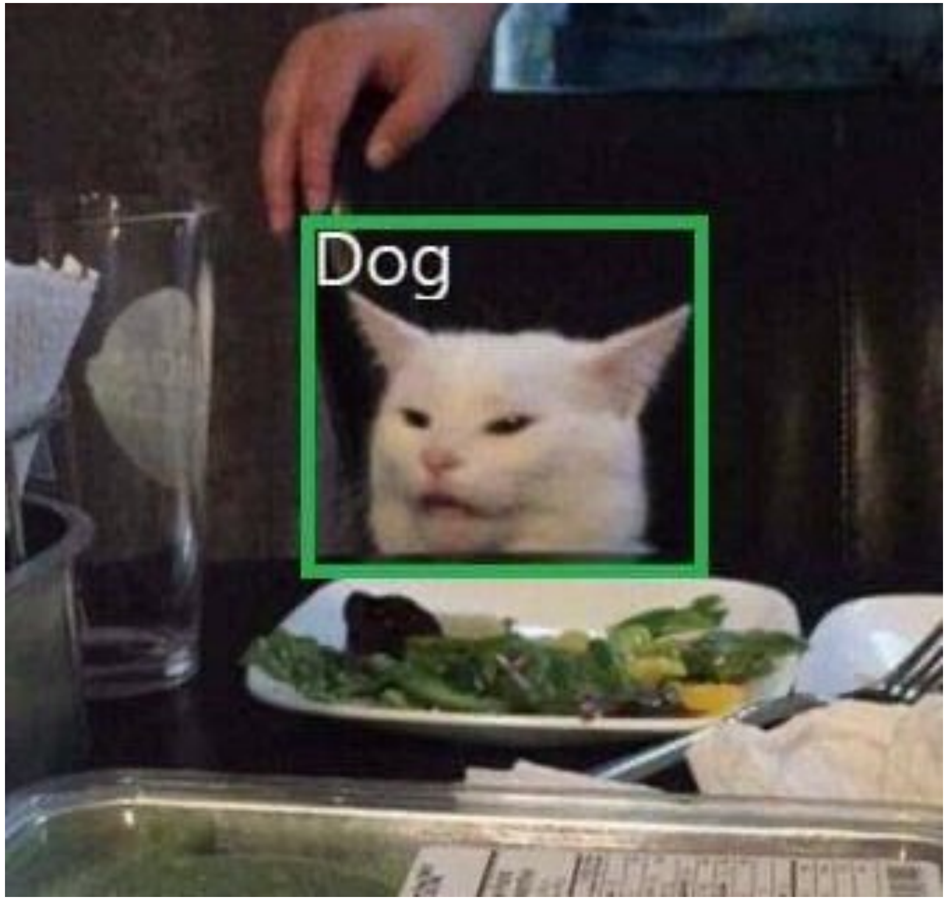


- The lines indicate faults
- The yellow box is roughly 25 km^2
- There are 5 interconnected faults within the box
- **Each pixel is worth 500m^2**

AI systems assist Geophysicists in fault interpretations. They must **explain their decisions** down to the **pixel-level** with high accuracy!

Adversarial attacks are engineered to intentionally mislead an AI system

- Widespread **face recognition** systems that use AI models can be attacked using **adversarial eyeglasses**
- **Infrared dots** act as adversaries for **face authentication** systems
- **Small patches and posters** on **traffic signs** cause autonomous vehicle perception modules to misclassify signs



Explanations in AI Systems

Case Study: Bias mitigation in Finance

AI Systems excel at integrating seemingly inconsequential data to harm protected groups

- According to CFPB:

"*A creditor employs facially neutral policies or practices that have an adverse effect or impact on a member of a protected class unless it meets a legitimate business need that cannot reasonably be achieved by means that are less disparate in their impact*"

- For people shopping on Wayfair on credit, the following variables were the most correlated to repayment[1]:
  - Borrower type of computer (Mac or PC)
  - Type of device (phone, tablet, PC)
  - Time of day you applied for credit (borrowing at 3am is not a good sign)
  - Your email domain (Gmail is a better risk than Hotmail)
  - Is your name part of your email (names are a good sign)

- **Each of the above variables are protected classes and using them is illegal to deny credit**

Berg, T., et al. "On the rise of the FinTechs—Credit scoring using digital footprints. Federal Deposit Insurance Corporation." *Center for Financial Research WP* 4 (2018): 2018

Explanations in AI Systems

Required Explanations

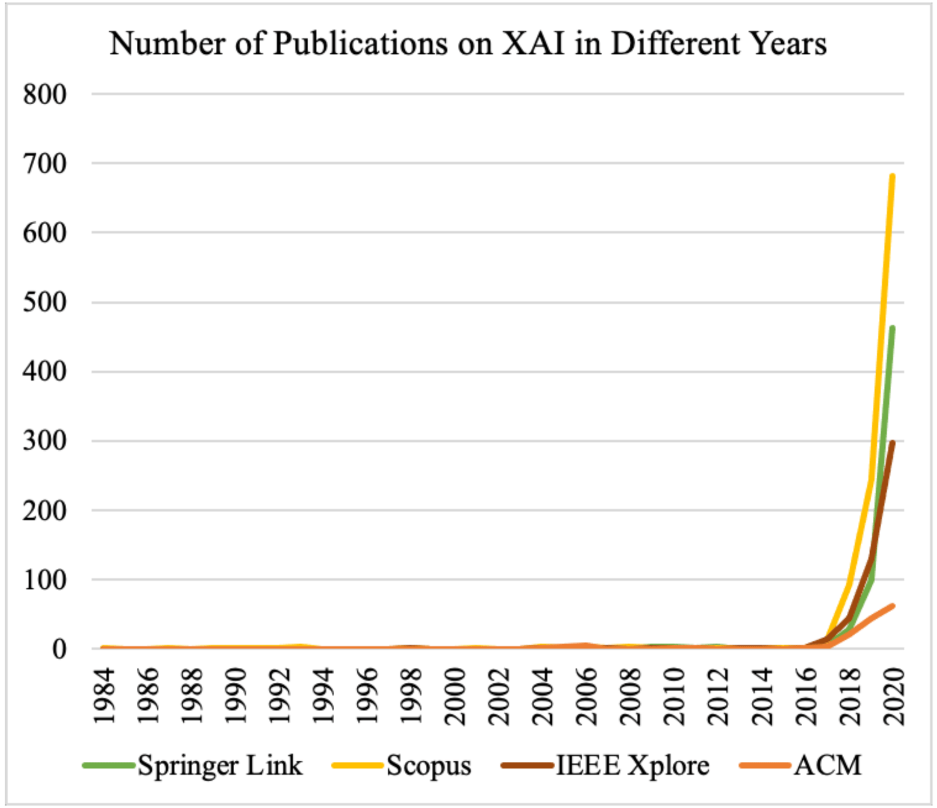
There is no “*One Size Fits All*” Explanation

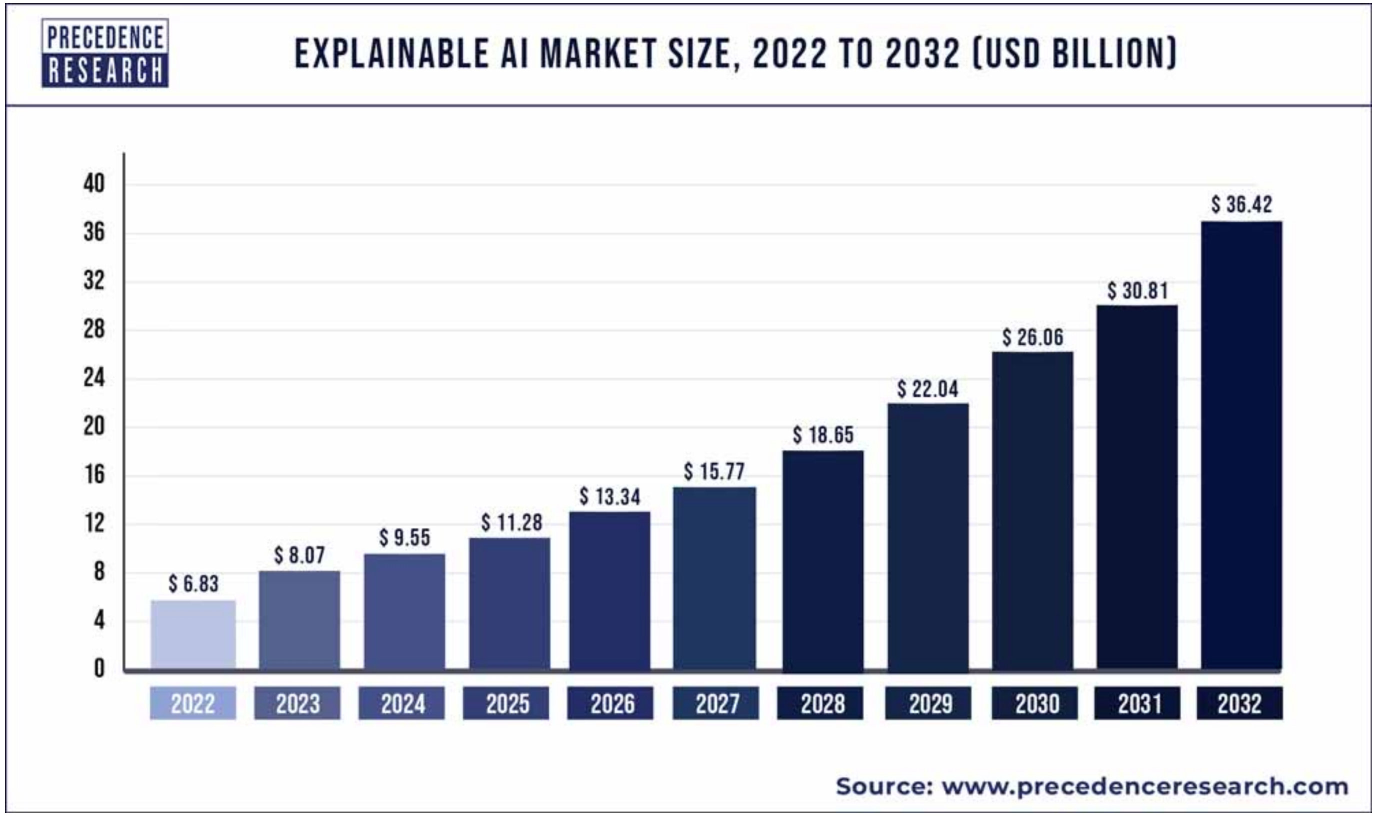
- Autonomous Vehicles require high-level semantic explanations
- Seismic interpretability requires low-level pixel explanations
- Medical images require structure-wise explanations
- Credit monitoring requires feature-based explanations
- Adversarial examples require ANY explanation!

Explanations in AI Systems

Research in Explainable AI

Research in Explainable AI has seen a tremendous growth and will continue to do so

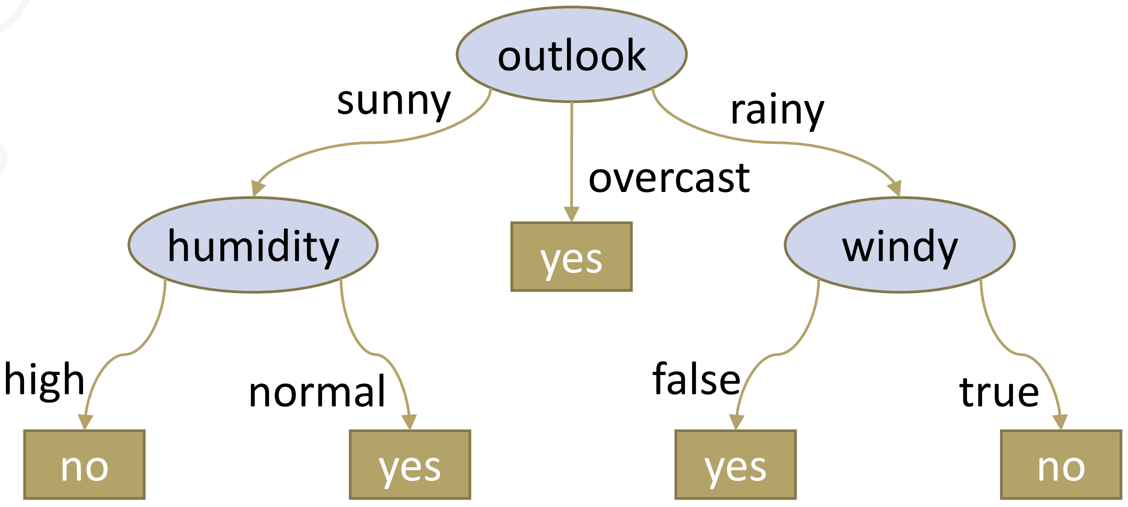




Explanations in AI Systems

Traditional Explainable AI

AI systems, traditionally, were logic-based handcrafted systems



Final decision tree computed based on data in table

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

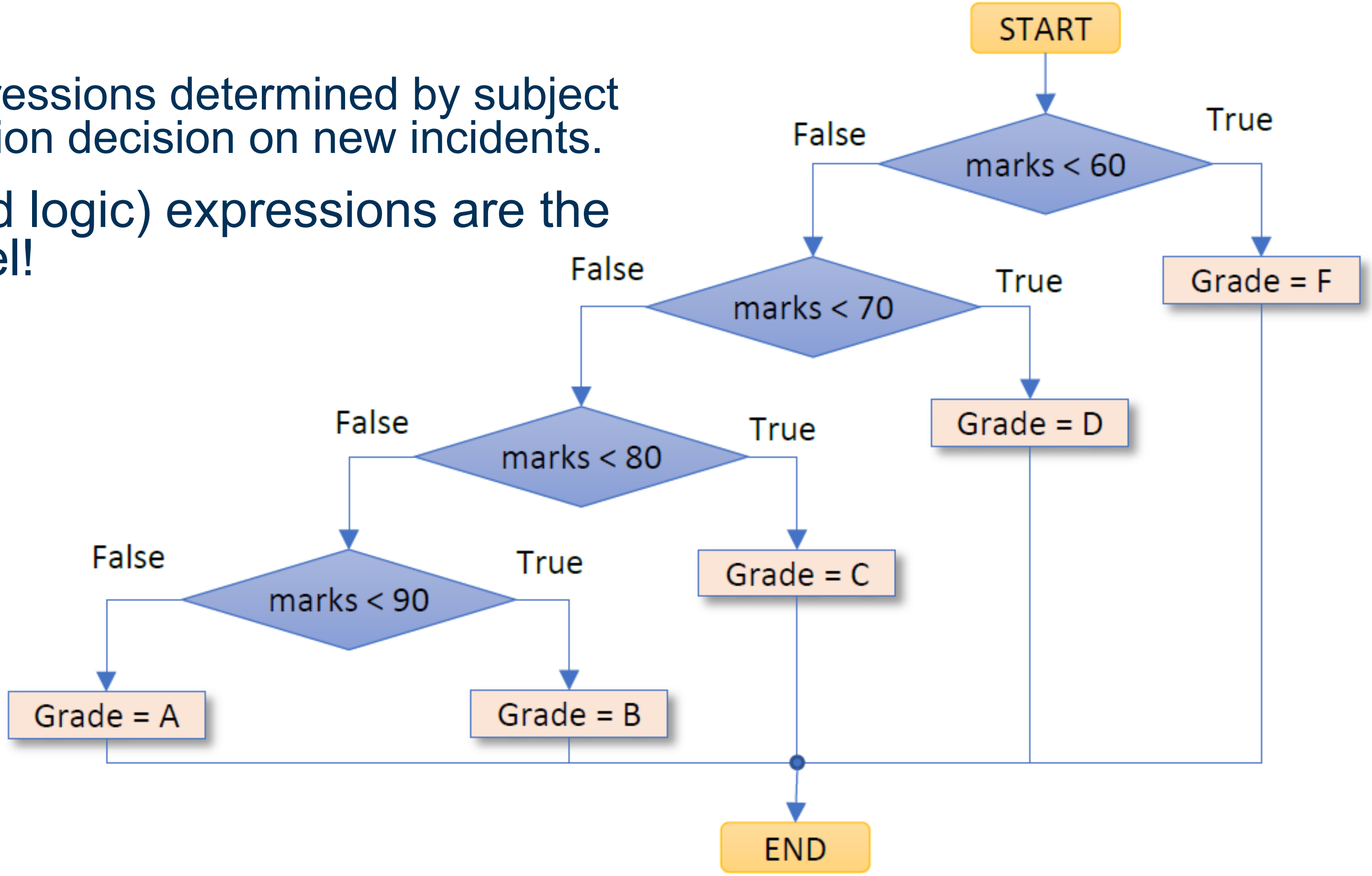
Explanations in AI Systems

Traditional Explainable AI

AI systems, traditionally, were logic-based handcrafted systems

- A set of rules and mathematical expressions determined by subject matter experts to arrive at classification decision on new incidents.
- The mathematical (arithmetic and logic) expressions are the model, but it is a hardwired model!

The method is the explanation!



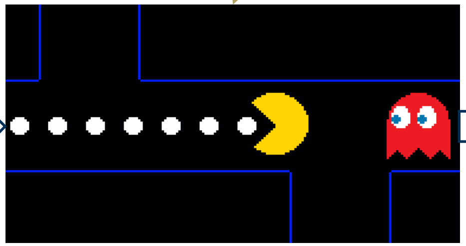
Explanations in AI Systems

Advent of Deep Learning

Deep Learning is an end-to-end trainable system with trillions of parameters

Traditional AI: Input → Hand-designed program → Output

Requires additional Explainability

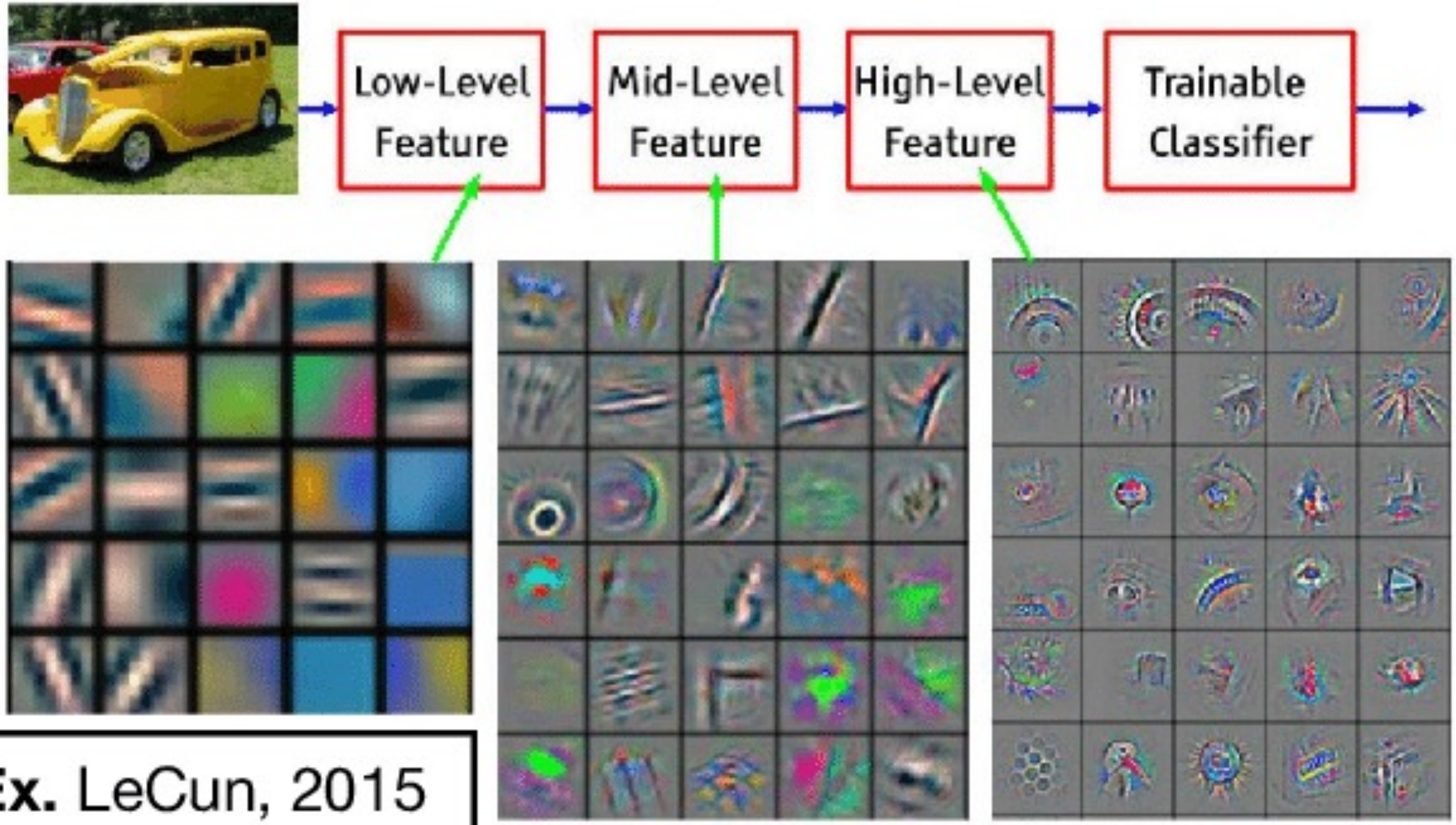
Deep Learning: Input →  → Output

Lecture Outline

Lecture 1: Introduction to Explainable AI

- Artificial Intelligence
- Explainability
- Need for Explainability in AI systems
- Deep Learning
  - Training
- Foundation Models
  - Challenges in Foundation Models
- Challenges in Explainability
  - Technical Challenges
  - Functional Challenges
  - Operational Challenges
- Takeaways

Deep Learning
Model Decomposition



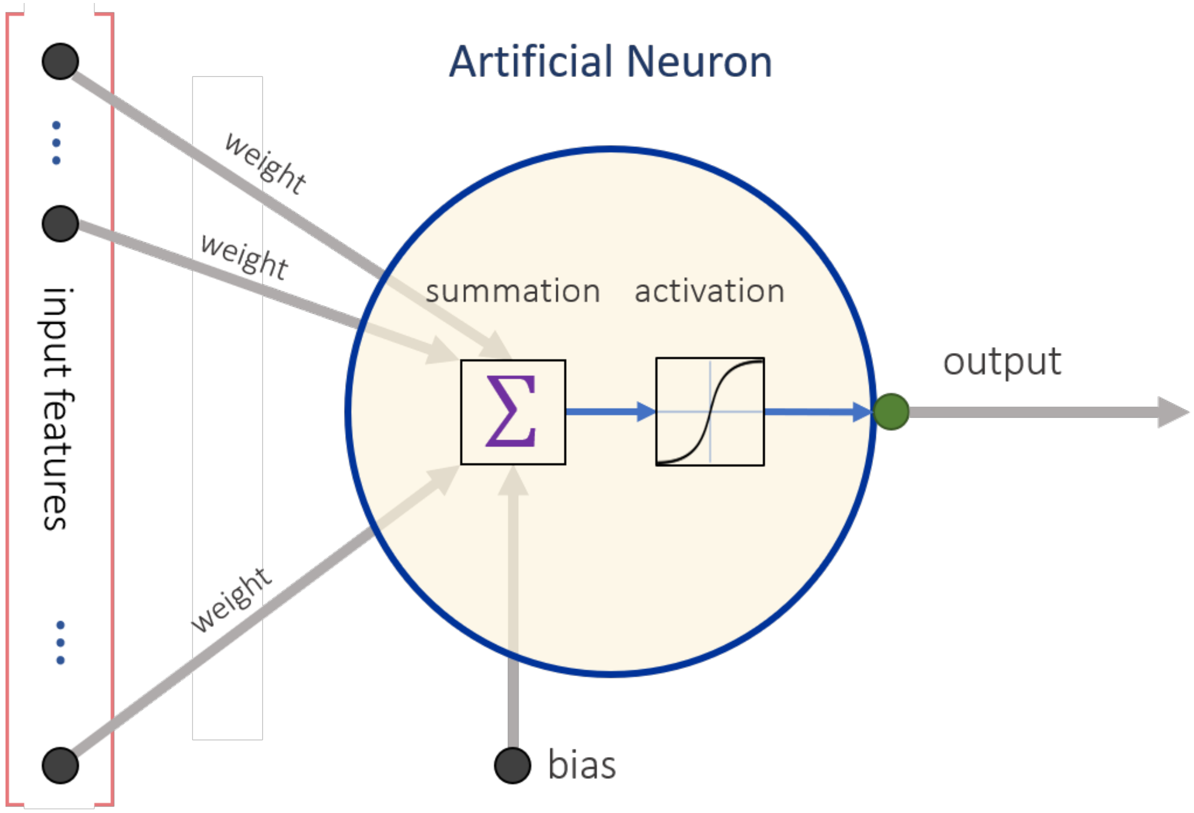
Ex. LeCun, 2015

Deep Learning
Model Decomposition

The underlying computational unit is the artificial neuron

Artificial neurons consist of:

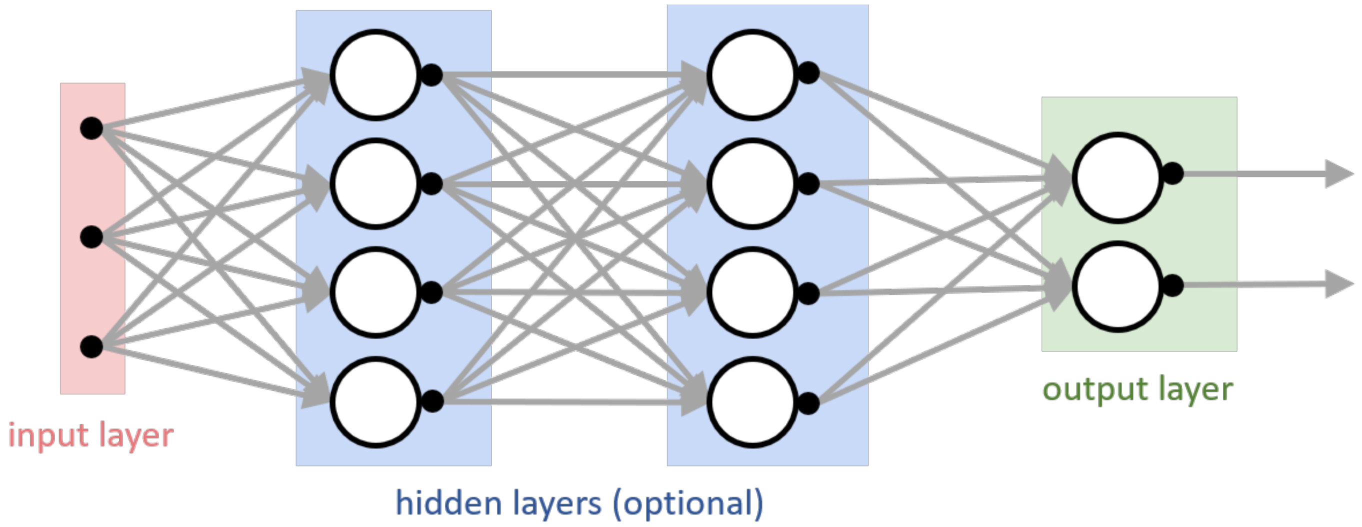
- A single **output**
- Multiple **inputs**
- Input **weights**
- A **bias** input
- An **activation function**



Deep Learning

Model Decomposition

The underlying computational unit is the artificial neuron



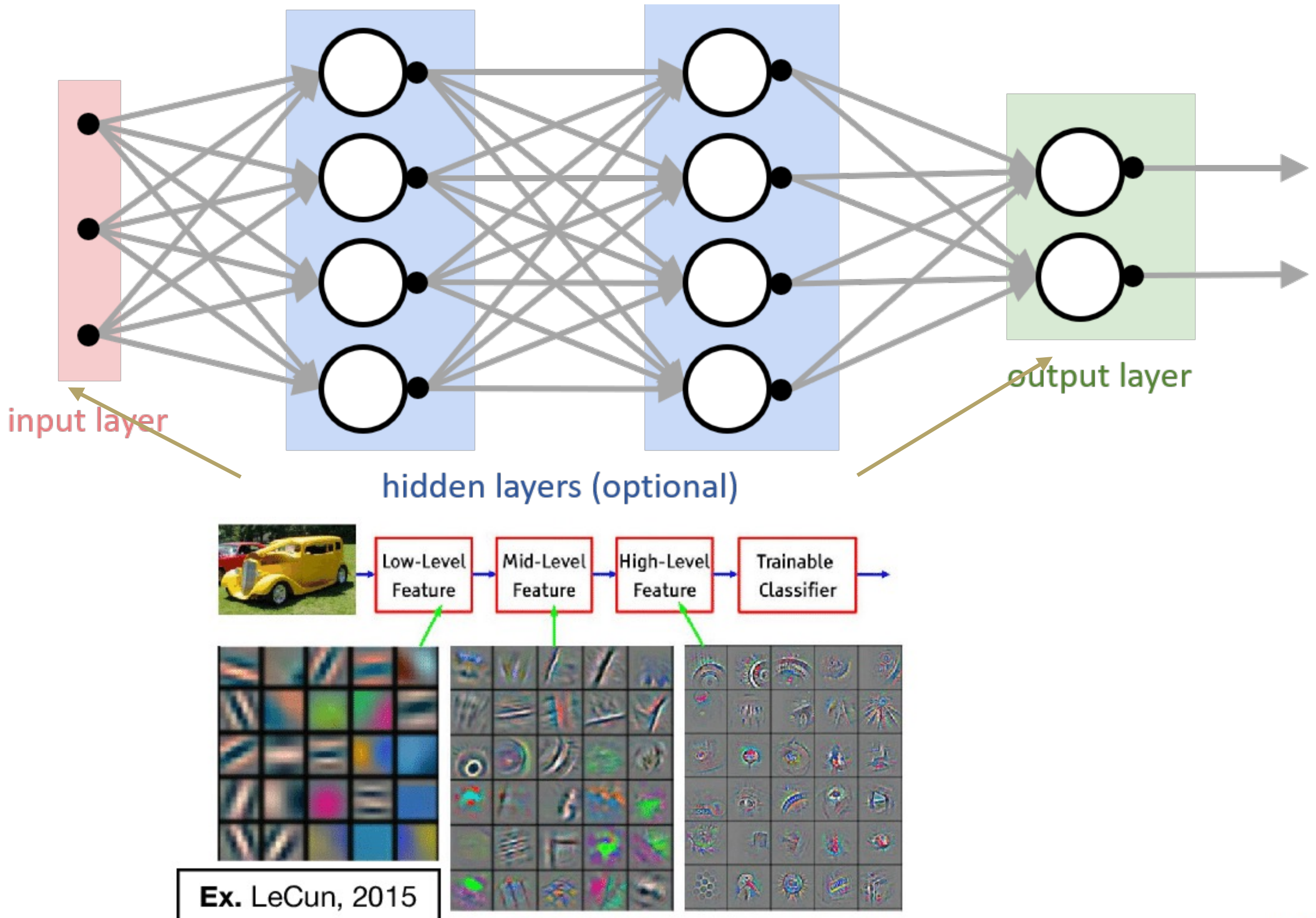
Typically, a neuron is part of a network organized in **layers**:

- An **input** layer (Layer 0)
- An **output** layer (Layer K)
- Zero or more **hidden** (middle) layers (Layers $1 \dots K-1$)

Deep Learning

Convolutional Neural Networks

Utilizes the stationary property of images to extract features via convolution filters

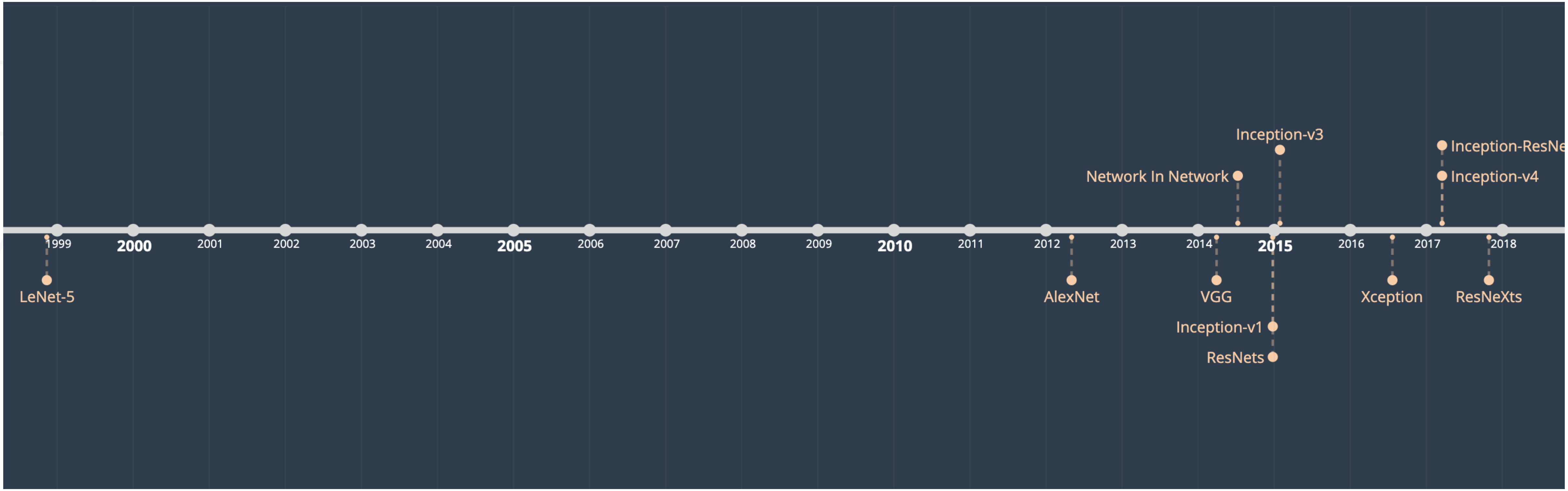


Ex. LeCun, 2015

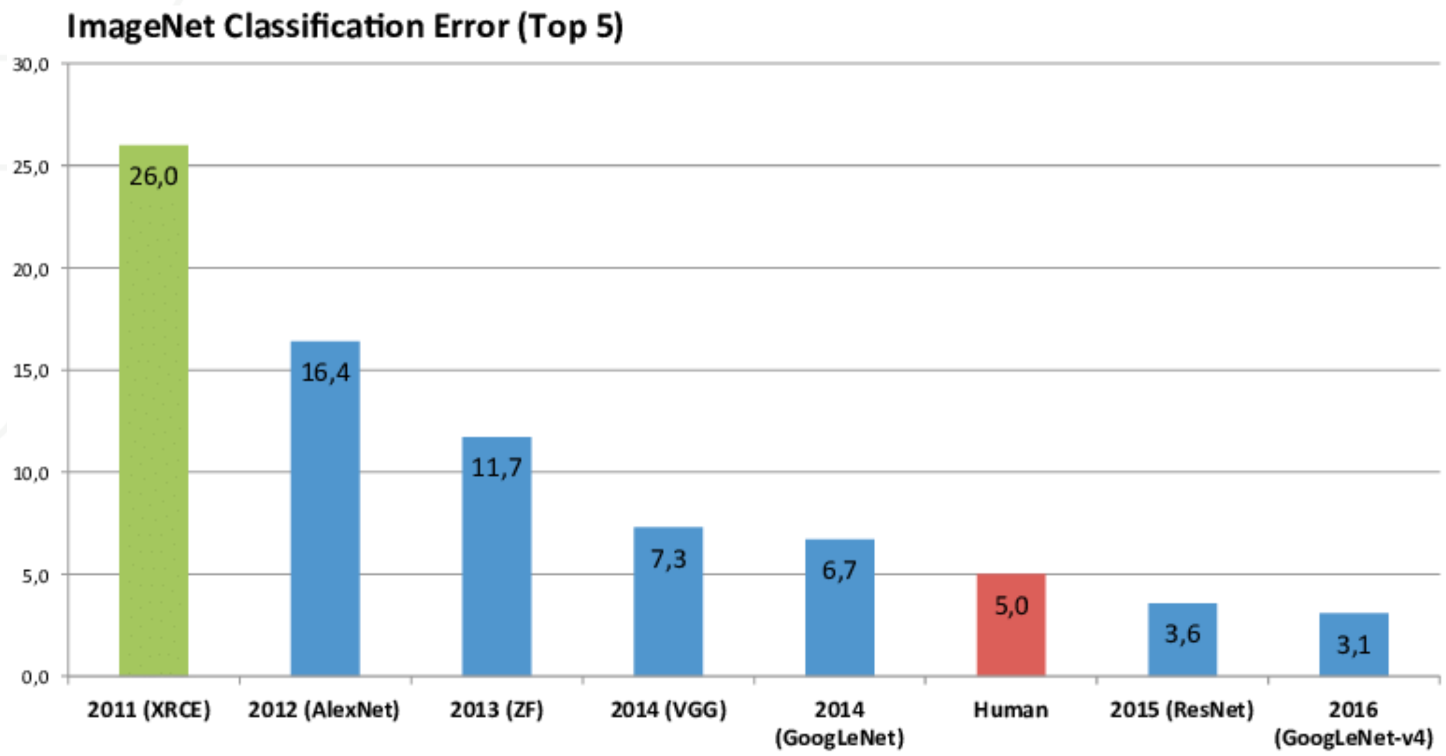
Deep Learning

Convolutional Neural Networks

Utilizes the stationary property of images to extract features via convolution filters



Access to largescale datasets like ImageNet and GPU acceleration aided CNN research

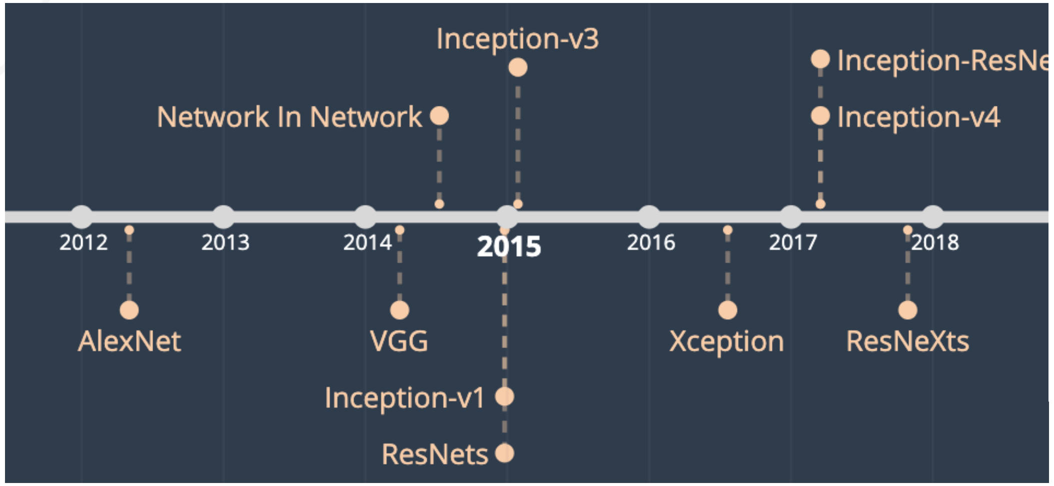


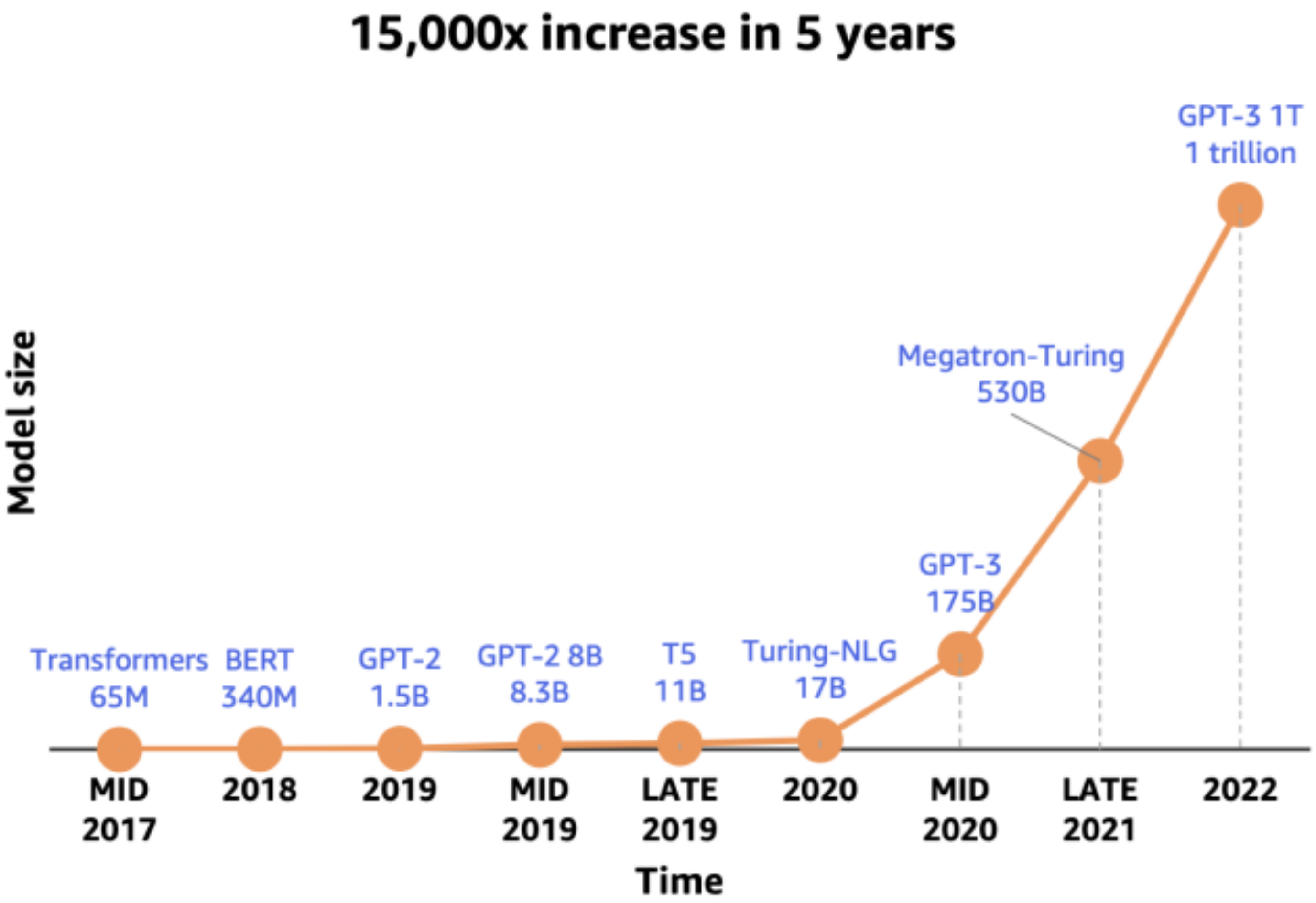
Imagenet:
1000 classes, 1.2M training images, 150K for testing

Deep Deep Deep Deep … Deep Learning

Recent Advancements

The number of parameters in models has increased exponentially





How to train such large networks?

Iteratively reduce a loss function $L(\boldsymbol{\theta})$ to find the optimal parameters $\boldsymbol{\theta}$

- θ is a combination of weights and biases
- Compute the gradients of a loss function iteratively and update the weights according to the update rule:

$$\theta(t+1) = \theta(t) - \alpha \frac{\partial\, L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

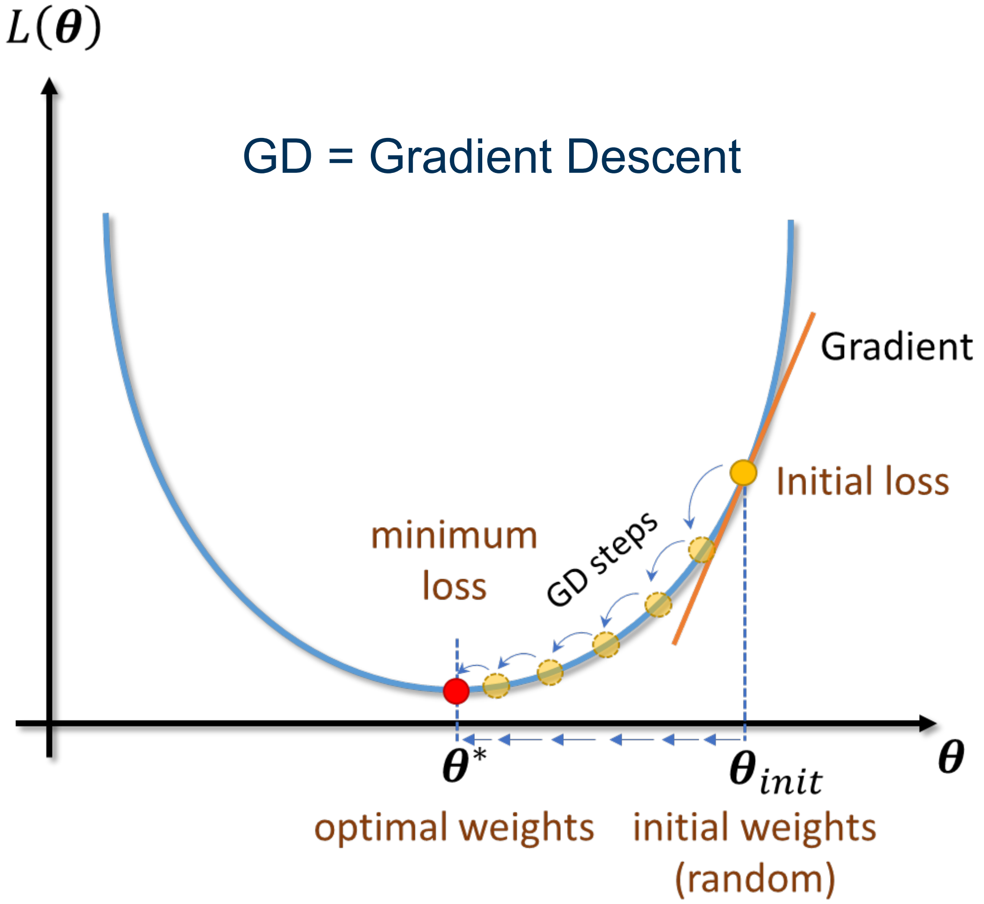
$\theta =$ Weights, biases

$t =$ Iteration step

$\alpha =$ Step Length

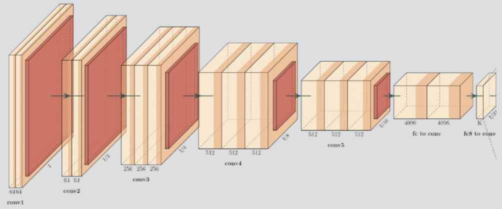
$L(\theta) =$ Loss function between prediction and ground truth

$\frac{\partial L(\theta)}{\partial \theta} =$ Gradient w.r.t weights and biases

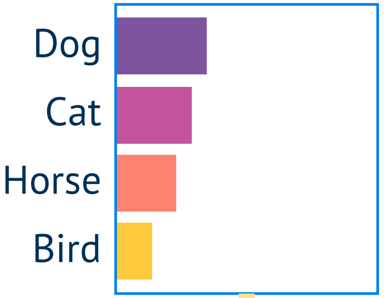


Gradients construct the manifold

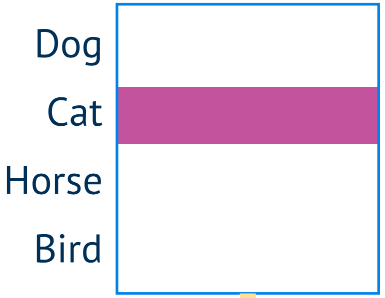
Network $\boldsymbol{f(\theta)}$



Predicted Class Probability



Ground-Truth Label

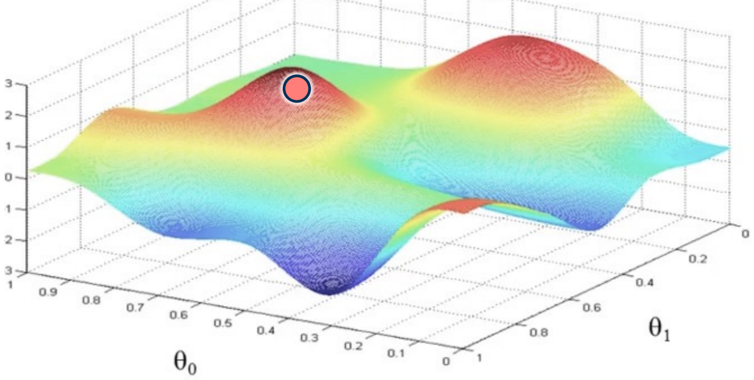


Loss $\boldsymbol{L(\theta)}$

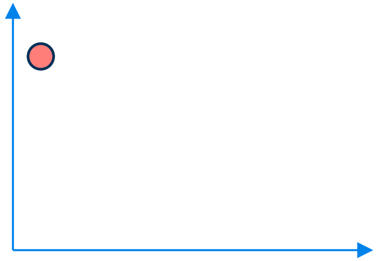
Backprop $\boldsymbol{L(\theta)}$ to generate **gradients** $\boldsymbol{\nabla_\theta L(\theta)}$

$$\boldsymbol{\theta(t+1) = \theta(t) - \alpha \nabla_\theta L(\theta)}$$

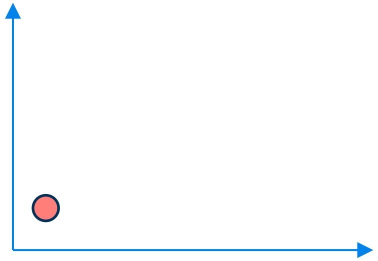
$\boldsymbol{L(\theta)}$



Loss



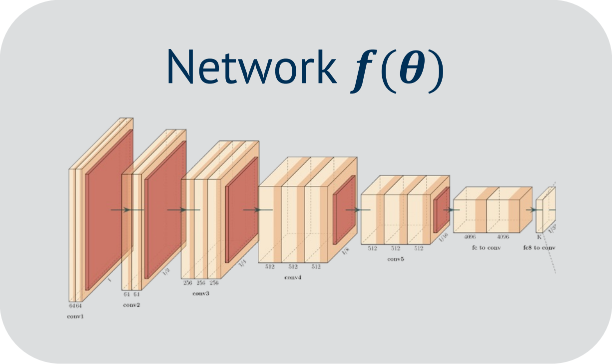
Accuracy

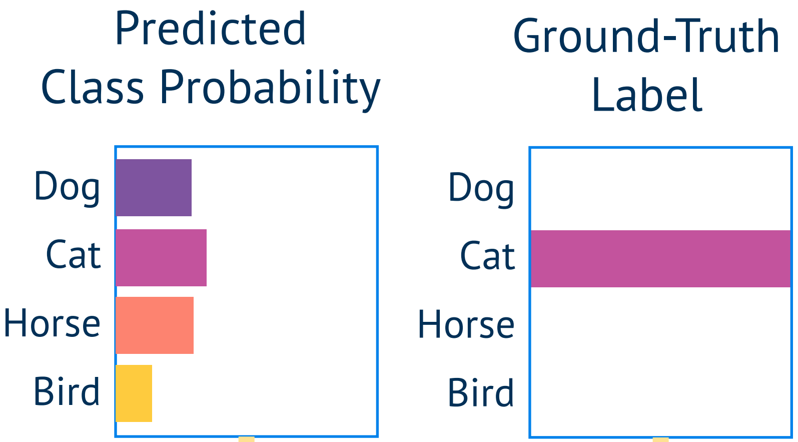


Deep Learning

Gradient Descent in Action

Gradients construct the manifold

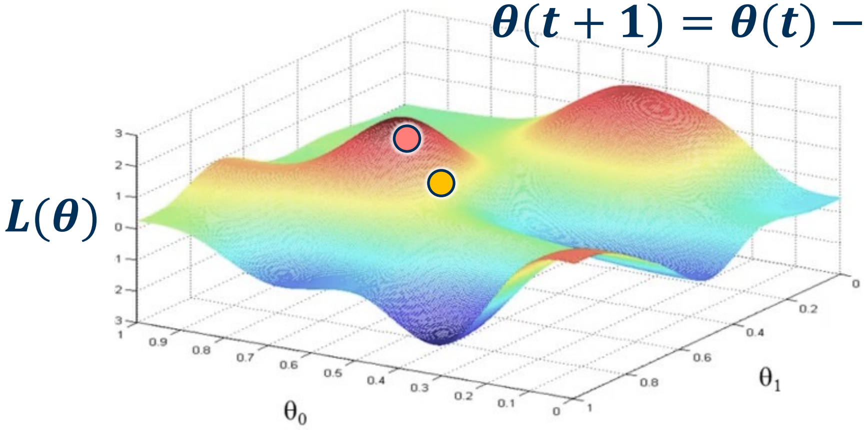


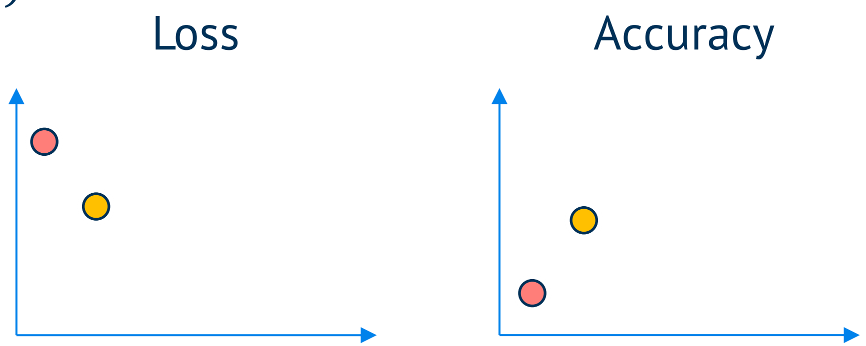


Backprop $\boldsymbol{L}(\boldsymbol{\theta})$ to generate **gradients** $\nabla_{\boldsymbol{\theta}}\, \boldsymbol{L}(\boldsymbol{\theta})$

Loss $\boldsymbol{L}(\boldsymbol{\theta})$

$$\boldsymbol{\theta}(\boldsymbol{t}+\mathbf{1}) = \boldsymbol{\theta}(\boldsymbol{t}) - \boldsymbol{\alpha}\, \nabla_{\boldsymbol{\theta}}\, \boldsymbol{L}(\boldsymbol{\theta})$$



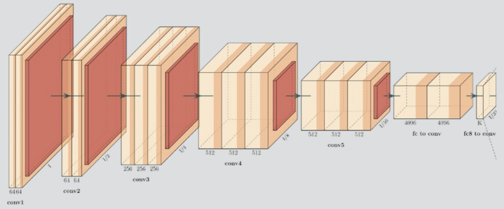


Deep Learning

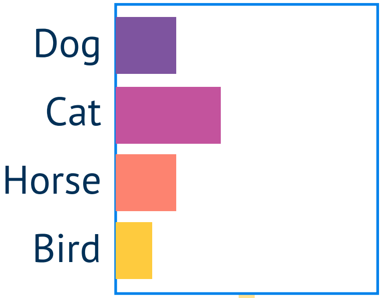
Gradient Descent in Action

Gradients construct the manifold

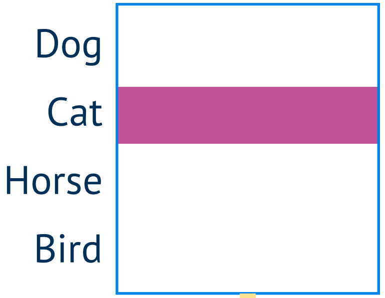
Network $f(\boldsymbol{\theta})$



Predicted Class Probability



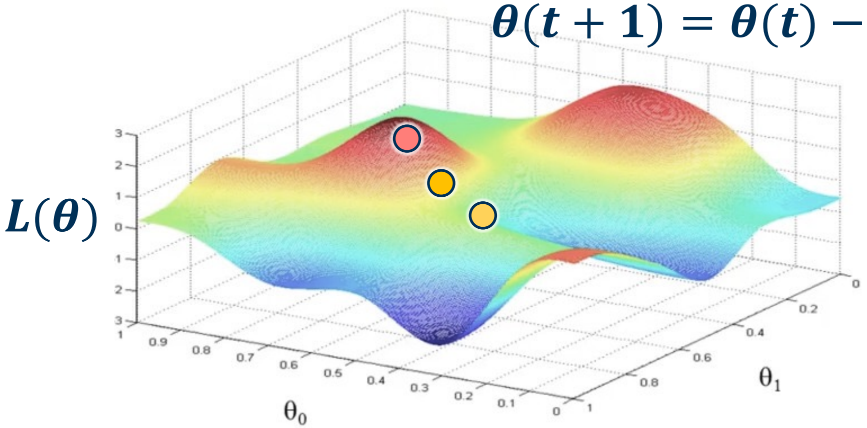
Ground-Truth Label



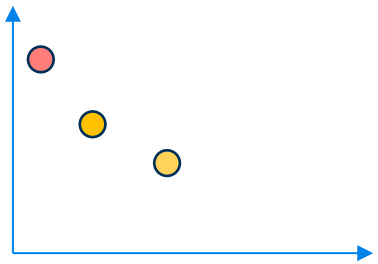
Loss $\boldsymbol{L}(\boldsymbol{\theta})$

Backprop $\boldsymbol{L}(\boldsymbol{\theta})$ to generate **gradients** $\nabla_{\theta}\, \boldsymbol{L}(\boldsymbol{\theta})$

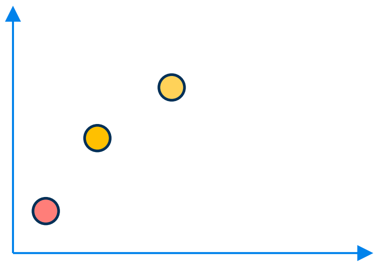
$$\boldsymbol{\theta}(\boldsymbol{t}+\mathbf{1}) = \boldsymbol{\theta}(\boldsymbol{t}) - \boldsymbol{\alpha}\, \nabla_{\theta}\, \boldsymbol{L}(\boldsymbol{\theta})$$



Loss



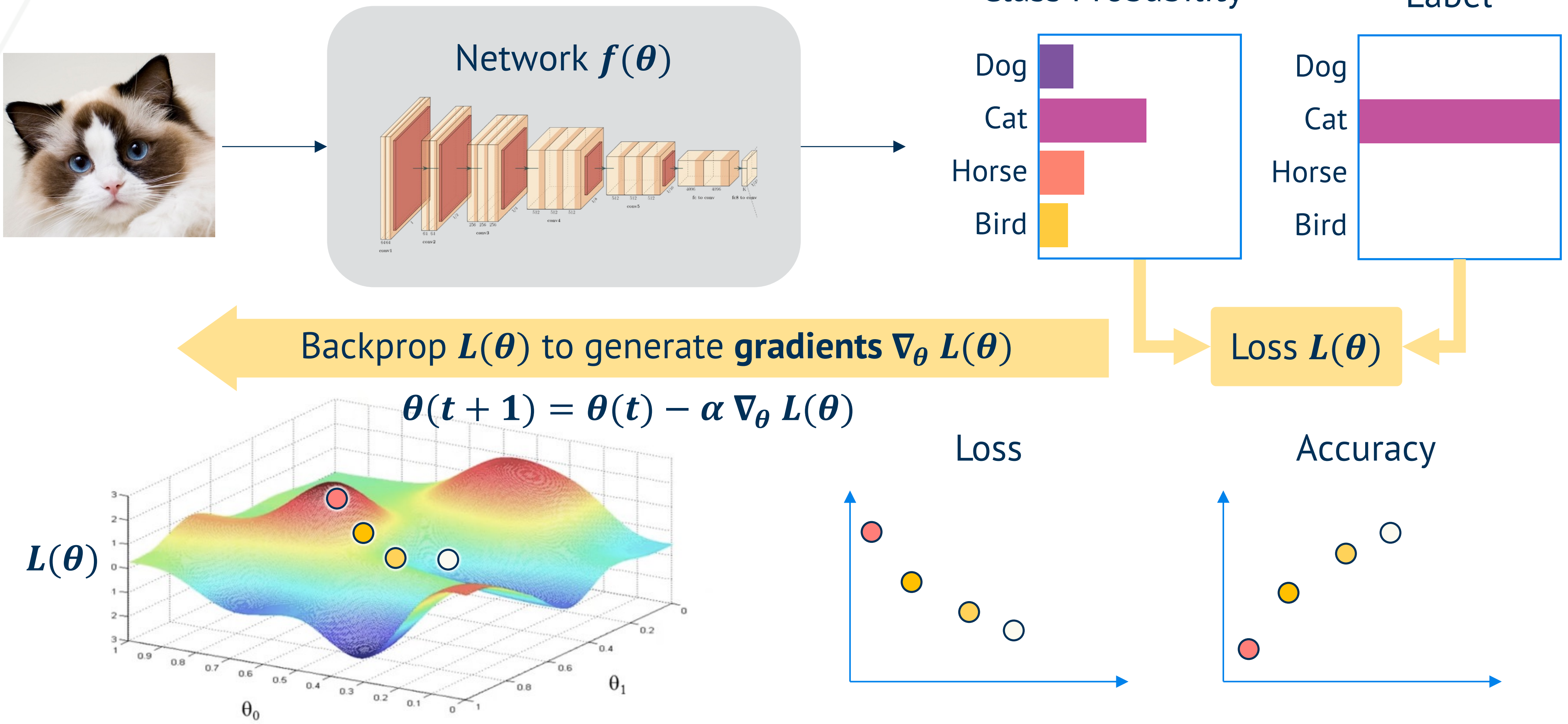
Accuracy



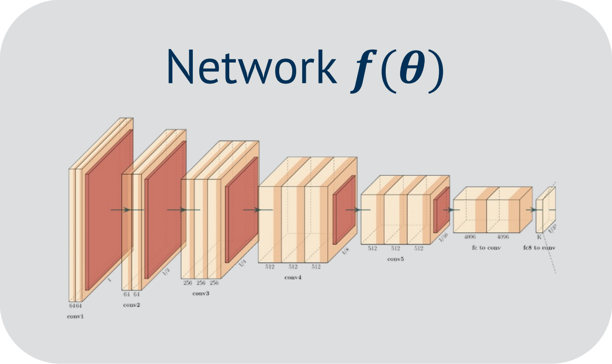
Deep Learning

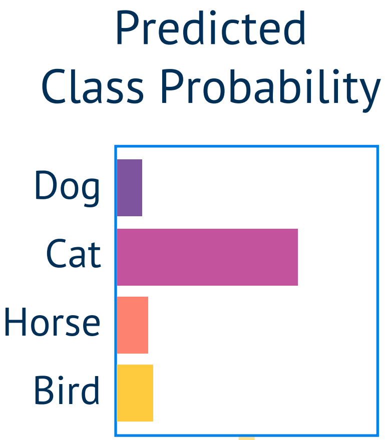
Gradient Descent in Action

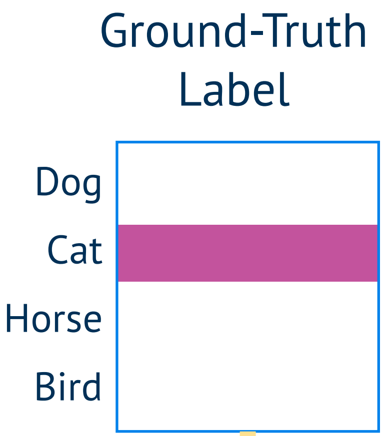
Gradients construct the manifold



Gradients construct the manifold



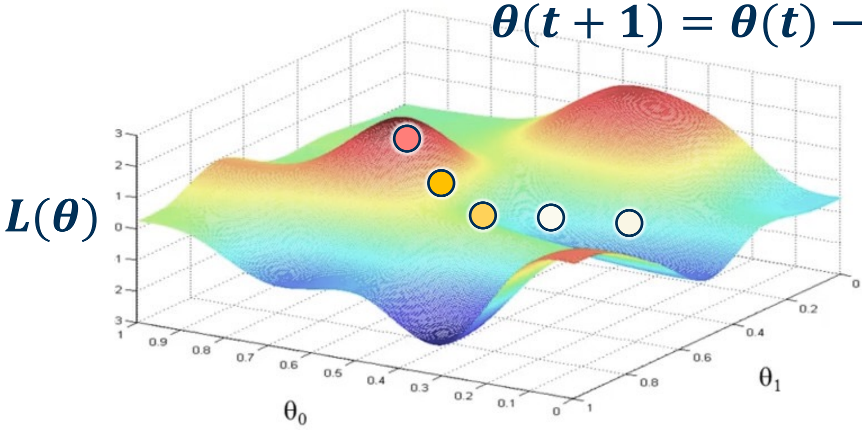


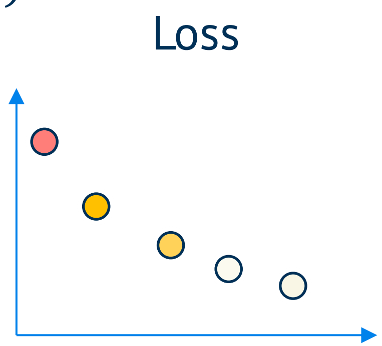


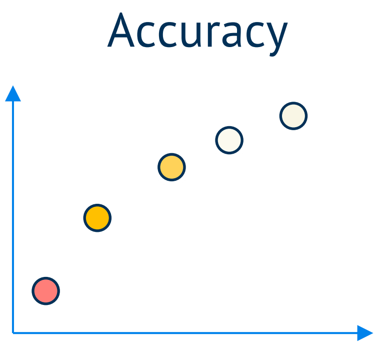
Loss $L(\boldsymbol{\theta})$

Backprop $L(\boldsymbol{\theta})$ to generate **gradients** $\nabla_{\theta}\, L(\boldsymbol{\theta})$

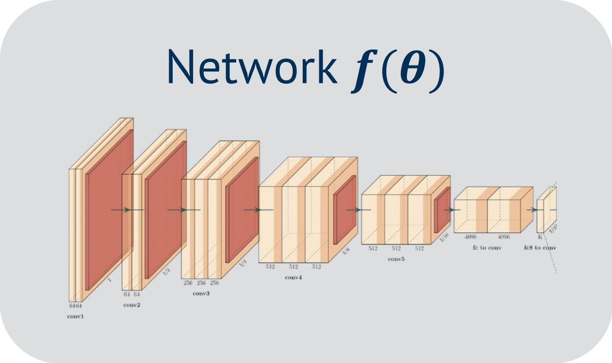
$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \alpha\, \nabla_{\theta}\, L(\boldsymbol{\theta})$$



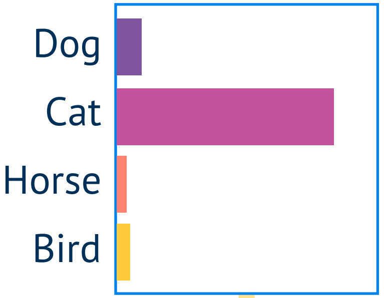




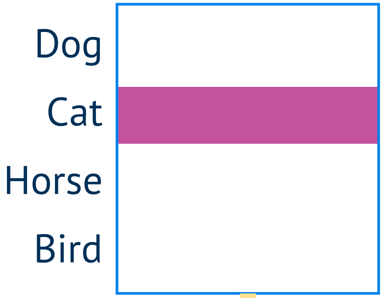
Gradients construct the manifold



Predicted Class Probability



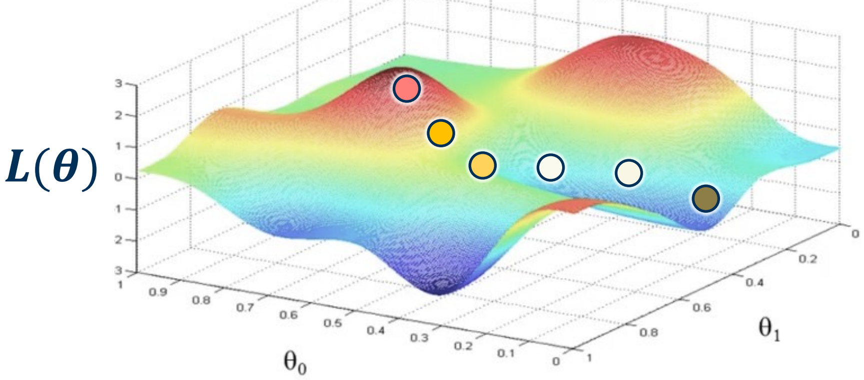
Ground-Truth Label



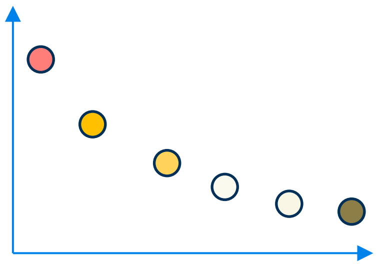
Loss $\boldsymbol{L(\theta)}$

Backprop $\boldsymbol{L(\theta)}$ to generate **gradients** $\boldsymbol{\nabla_\theta L(\theta)}$

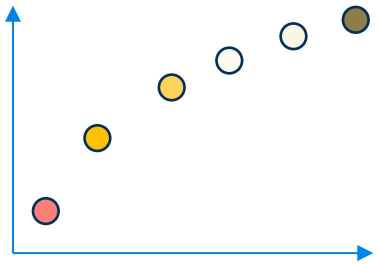
$$\boldsymbol{\theta(t+1) = \theta(t) - \alpha\, \nabla_\theta\, L(\theta)}$$



Loss



Accuracy



Lecture Outline

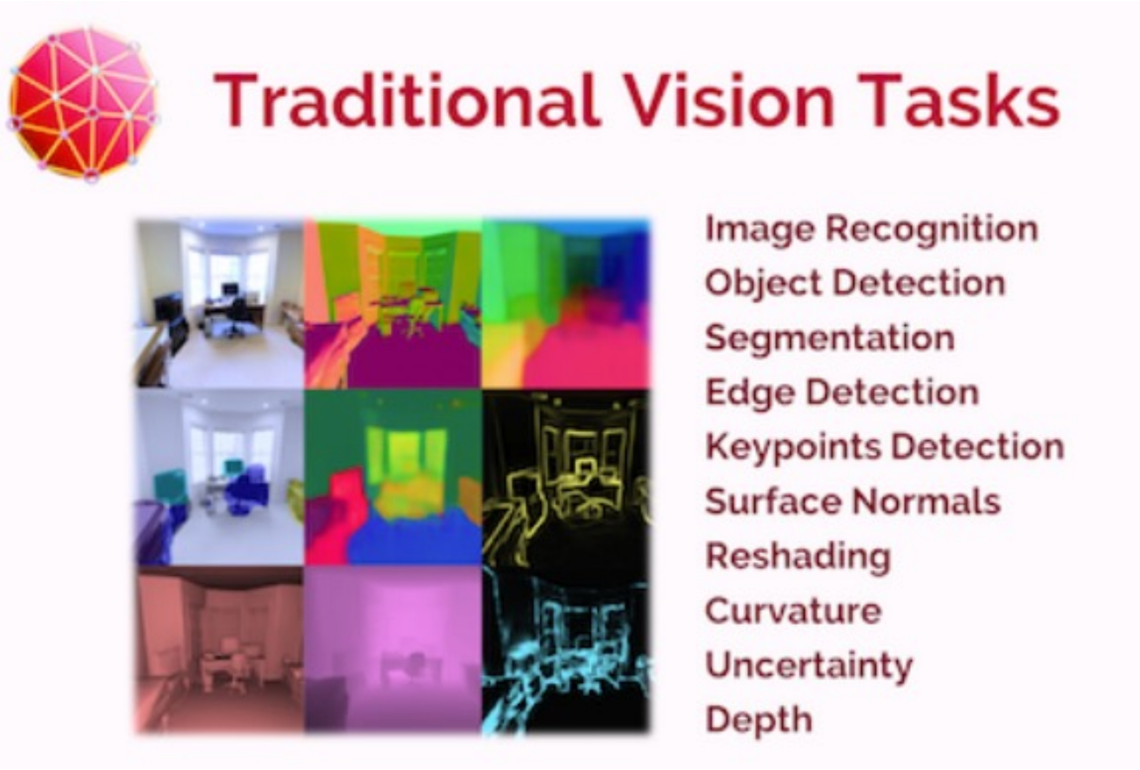
Lecture 1: Introduction to Explainable AI

- Artificial Intelligence
- Explainability
- Need for Explainability in AI systems
- Deep Learning
  - Training
- Foundation Models
  - Challenges in Foundation Models
- Challenges in Explainability
  - Technical Challenges
  - Functional Challenges
  - Operational Challenges
- Takeaways

Deep Deep Deep Deep … Deep Learning

Recent Advancements

Underlying features among different vision tasks are similar



This similarity leads to Transfer Learning

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Transfer Learning

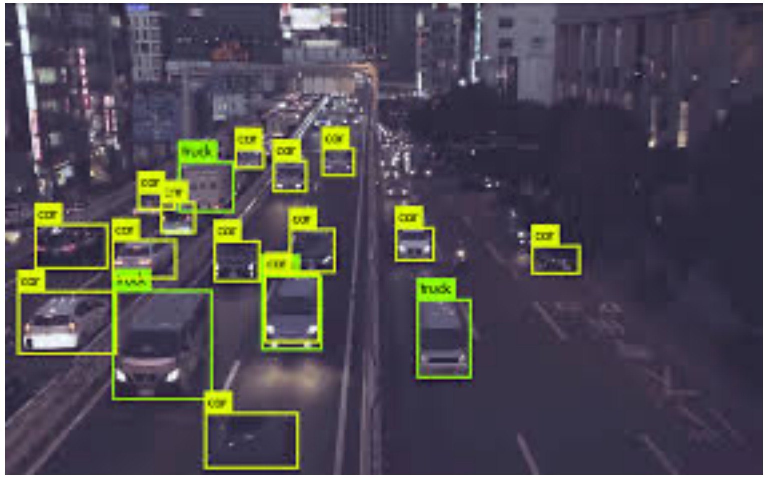
What is Transfer Learning?

- Deep networks tend to **learn common representations** for various tasks in their earlier layers
- Can be exploited **to transfer representations from networks trained on large datasets** on one task (i.e., Image Classification on ImageNet) called the *source* to a different task called the *target* task
- Usually done by **taking large pretrained network** and then **finetuning last layer** (with all other layers frozen) on target dataset
- **Pre-trained frozen backbone** acts as a **feature extractor** while **finetuned last layer** acts to project the representations into the **decision boundary for the target task**
- Utility depends on how closely related the source and target datasets and/or tasks are

Transfer Learning

Foundation Models

Source: https://gluon-cv.mxnet.io/



Source: https://www.move-lab.com/blog/tracking-things-in-object-detection-videos

Source: https://www.saagie.com/blog/object-detection-part1/

Foundation Models

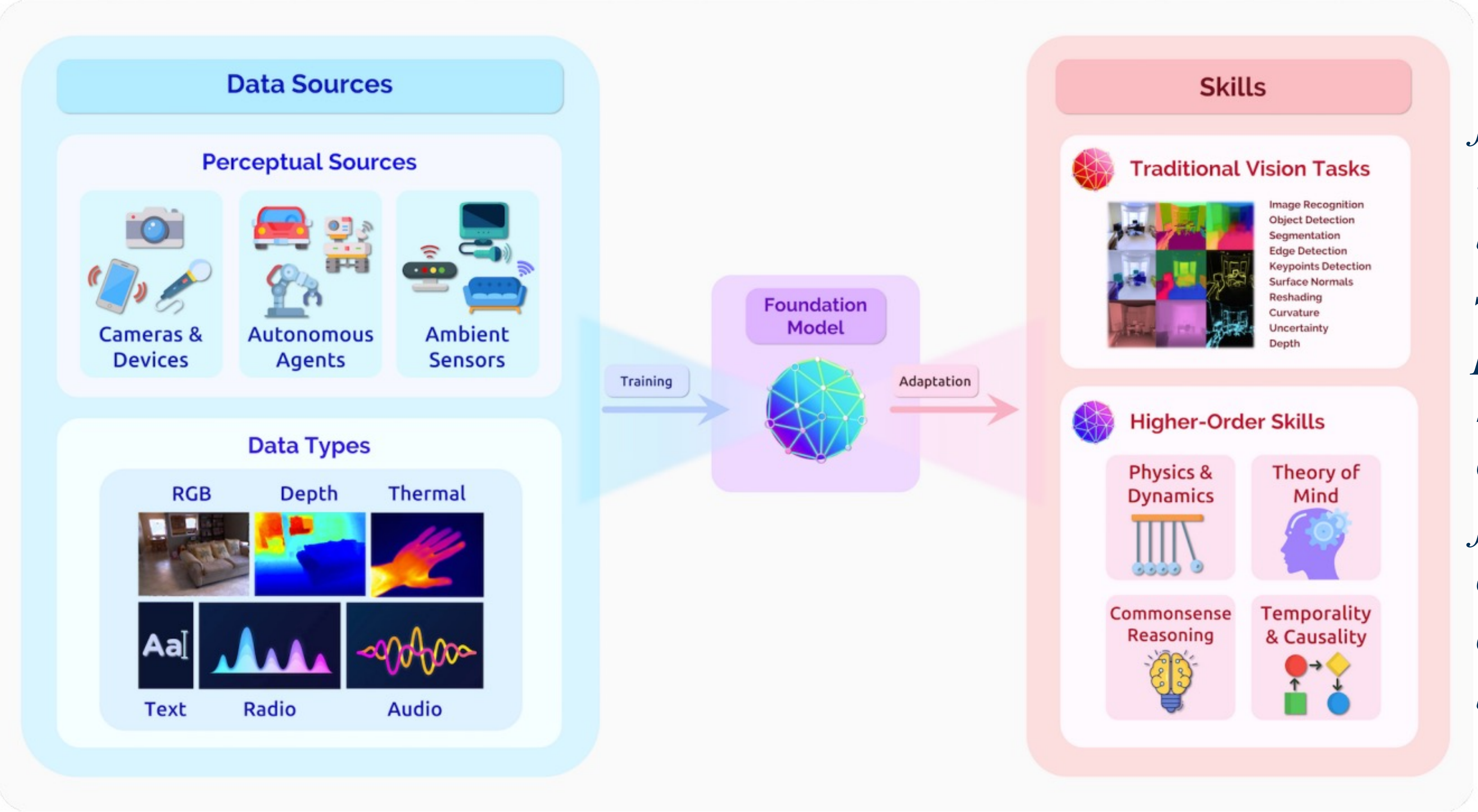
Origin of the term Foundation Models

- **Foundation models** are like any other deep network that have employed **transfer learning**, except **at *scale***
- ***Scale*** brings about ***emergent* properties** that are common between tasks
- **Before 2019:** Base architectures that powered multiple neural networks were **ResNets, VGG** etc.
- **Since 2019: BERT, DALL-E, GPT, Flamingo**
- Changes since 2019: **Transformer architectures and Self-Supervision**

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Foundation Models

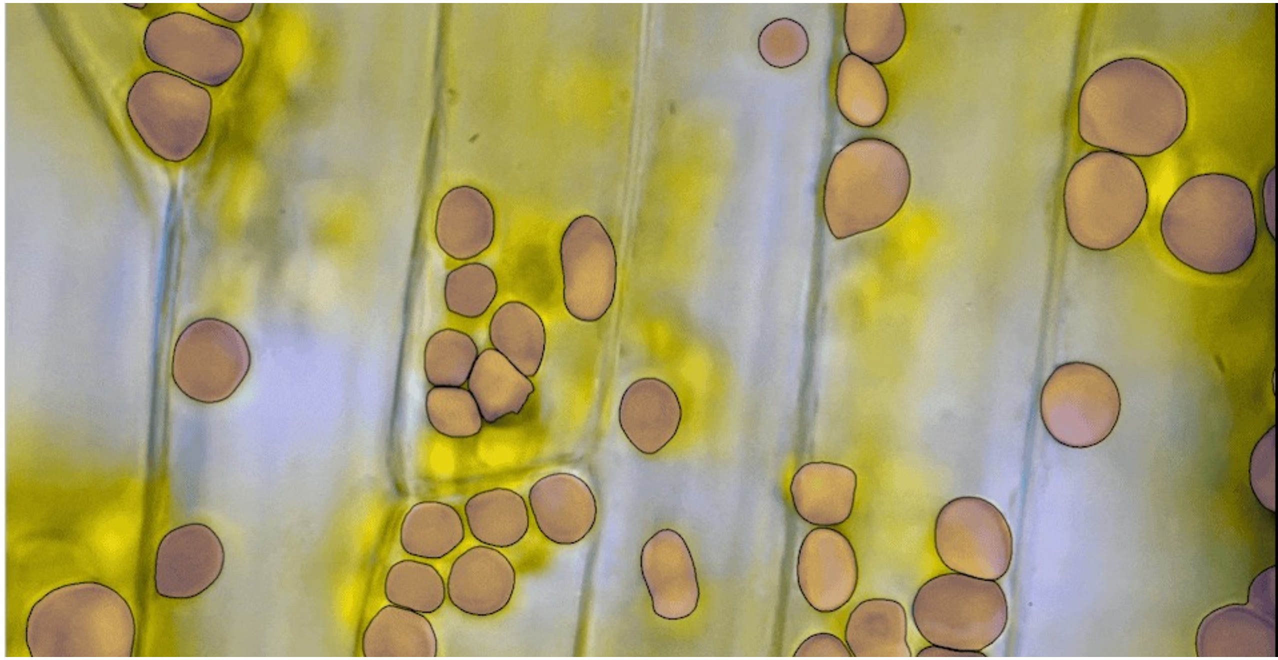
Origin of the term Foundation Models



*'By harnessing **self-supervision at scale**, foundation models for vision have the potential to **distill raw, multimodal sensory information into visual knowledge**, which may effectively support traditional **perception tasks** and possibly enable new progress on challenging higher-order skills like **temporal and commonsense reasoning** These inputs can come from a **diverse range of data sources** and application domains, suggesting promise for applications in **healthcare and embodied, interactive perception settings**'*

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Foundation Models
Segment Anything Model



Segment Anything Model (SAM) released by Meta on April 5, 2023 was trained on Segment Anything 1 Billion dataset with 1.1 billion high-quality segmentation masks from 11 million images

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

Foundation Models

Challenges in Segment Anything Model

Case study: SAM on Fisheye cameras

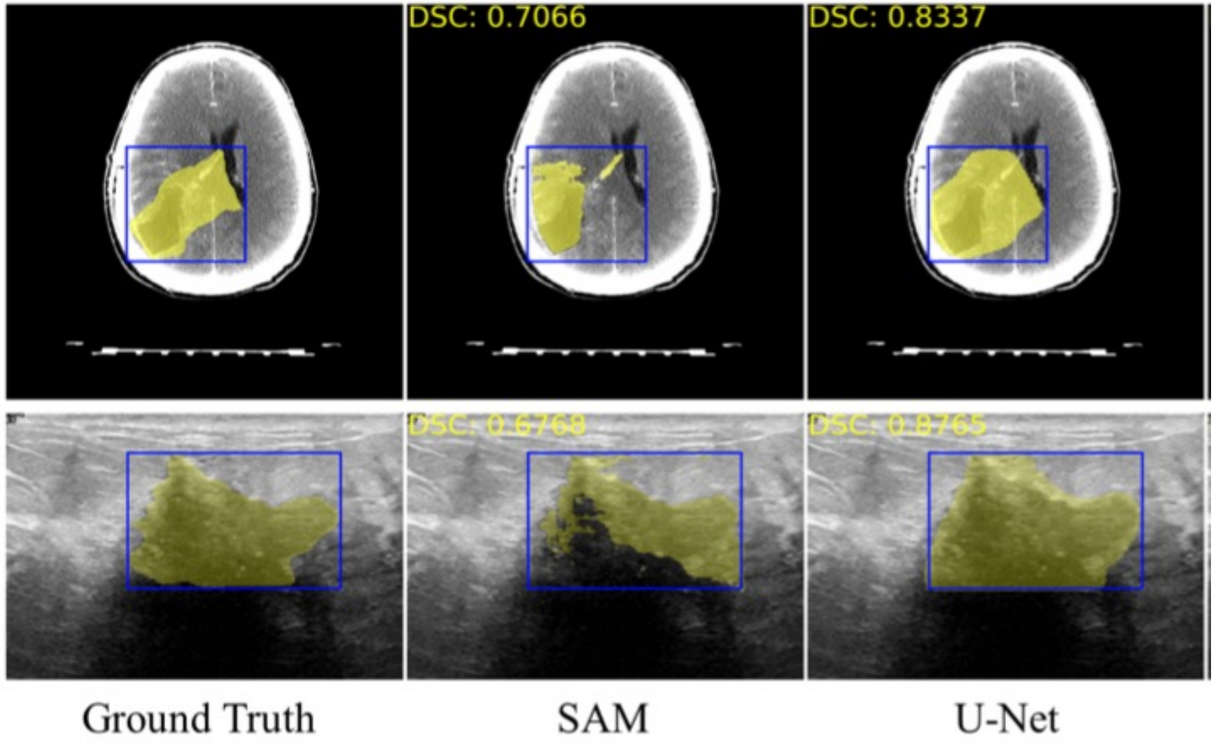
Results from Zero-shot (using the trained model out of the box) Segment Anything Model on Woodscape dataset

Important context and objects are not segmented

Kokilepersaud, K., Prabhushankar, M., Yarici, Y., AlRegib, G., & Parchami, A. (2023). Exploiting the Distortion-Semantic Interaction in Fisheye Data. *IEEE Open Journal of Signal Processing*.

Case study: SAM on Medical images

Results from Zero-shot Segment Anything Model on various segmentation datasets



U-Net outperforms existing SAM

Ma, Jun, and Bo Wang. "Segment anything in medical images." *arXiv preprint arXiv:2304.12306* (2023).

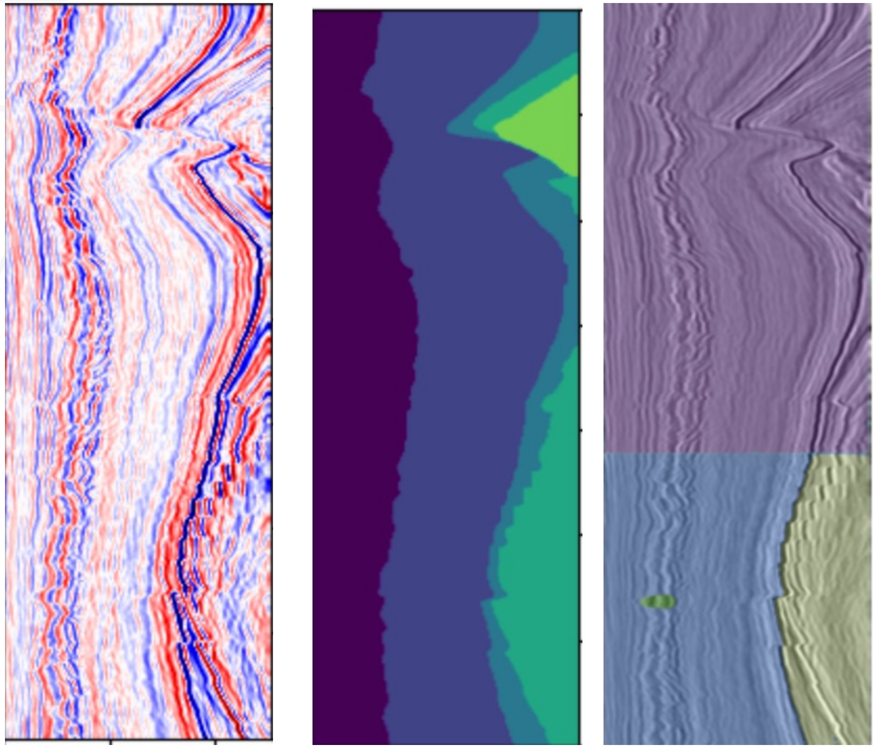
Foundation Models

Challenges in Segment Anything Model

Case study: SAM on Seismic Data

Results from Zero-shot (using the trained model out of the box) Segment Anything Model on F3 dataset

Seismic Image | Ground Truth | SAM Output



Faults are confused for boundaries

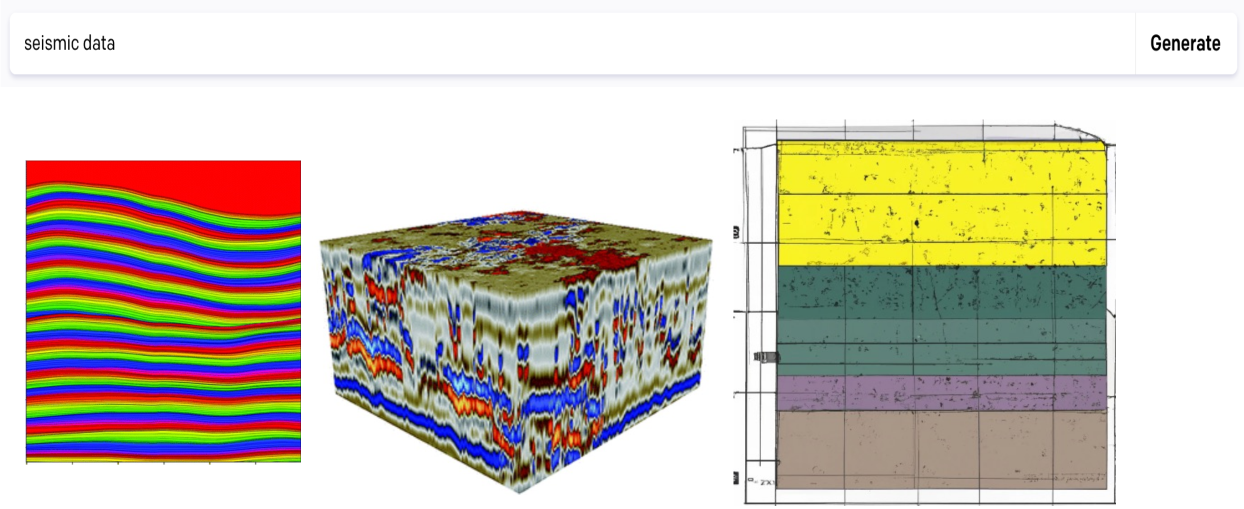


Image generation using DALLE-2

Foundation Models

Challenges in Segment Anything Model

Case study: SAM on Seismic Data

Results from prompting Segment Anything Models on natural images

Everything detection

Ideal Prediction after prompting

Point prompts generated every 4X4 pixels

All objects segmented

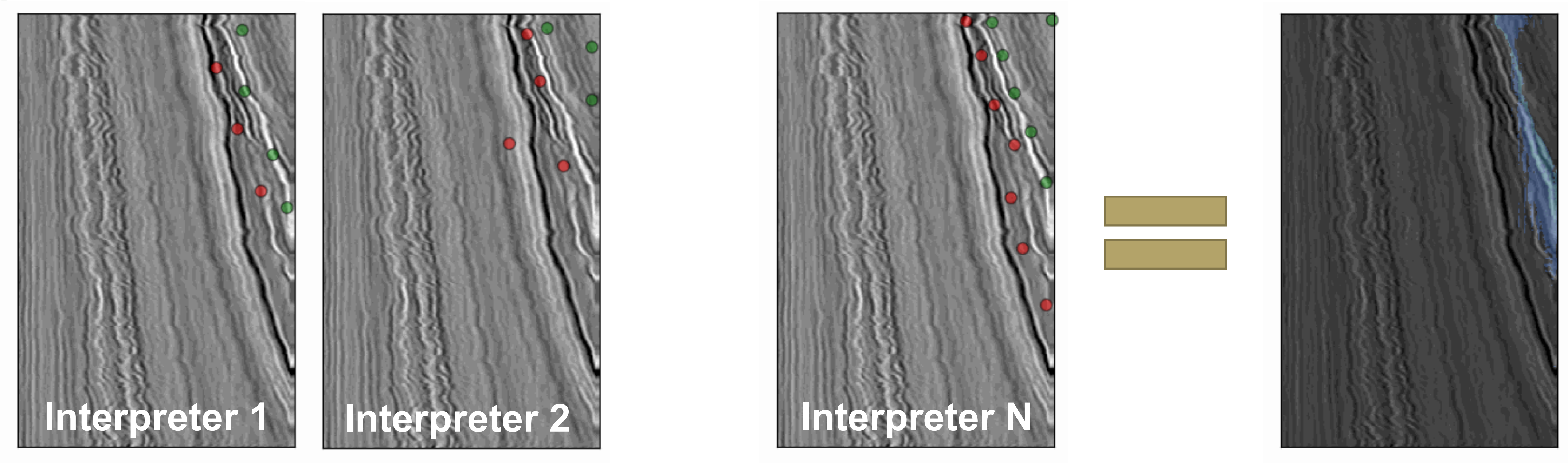
Manual prompting selects only one segment

Foundation Models

Challenges in Segment Anything Model

Since SAM is not understood, different people prompt differently and get different results

Results when prompting Segment Anything Models on seismic images



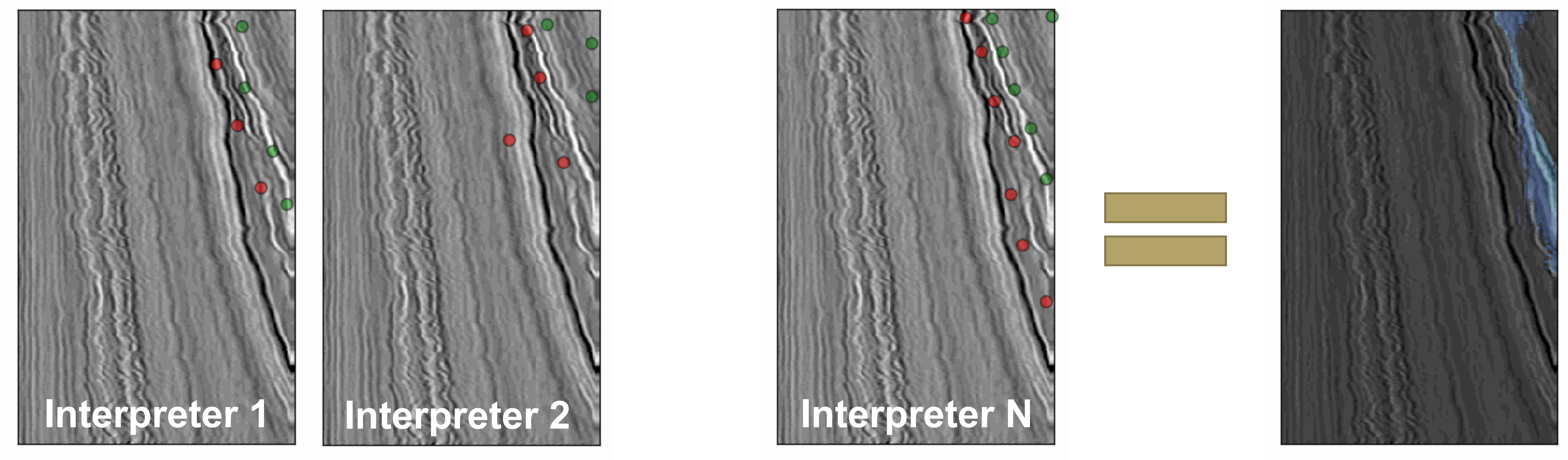
Variance of outputs from 6 prompters

Foundation Models

Challenges in Segment Anything Model

Since SAM is not understood, different people prompt differently and get different results

Results when prompting Segment Anything Models on seismic images



Variance of outputs from 6 prompters

Explanations are key to unlocking Neural Networks for Everybody!

Lecture Outline

Lecture 1: Introduction to Explainable AI

- Artificial Intelligence
- Explainability
- Need for Explainability in AI systems
- Deep Learning
    - Training
- Foundation Models
    - Challenges in Foundation Models
- Challenges in Explainability
    - Technical Challenges
    - Functional Challenges
    - Operational Challenges
- Takeaways

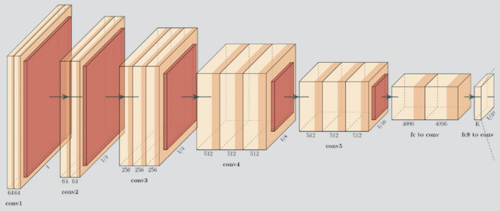
Challenges in Explainability

Technical challenges in Explainability

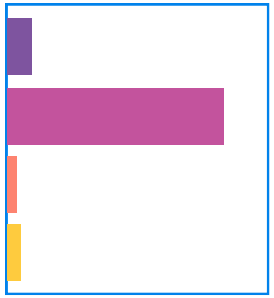
When explanations are required, we have access to a trained model, and a single data point

Given

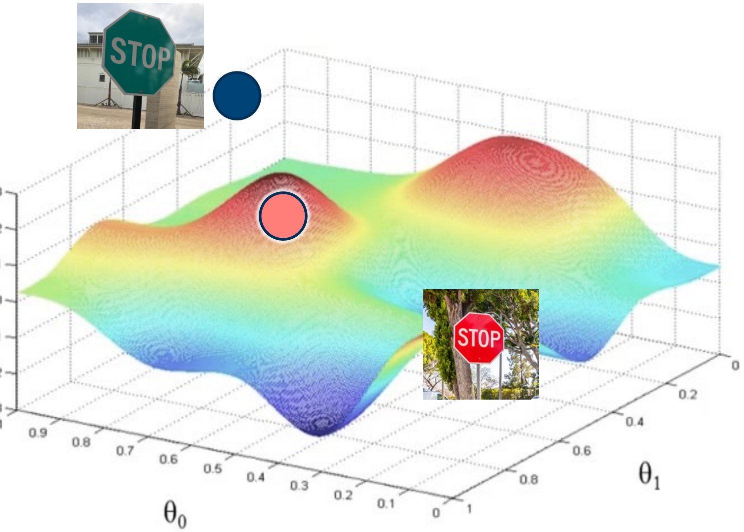
Network $\boldsymbol{f}(\boldsymbol{\theta})$



Predicted Class Probability



$\boldsymbol{L}(\boldsymbol{\theta})$



Explain the decision of the predicted class as a function of the learned manifold

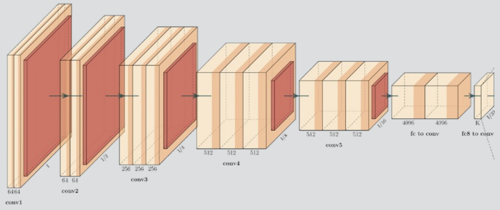
Challenges in Explainability

Technical challenges in Explainability

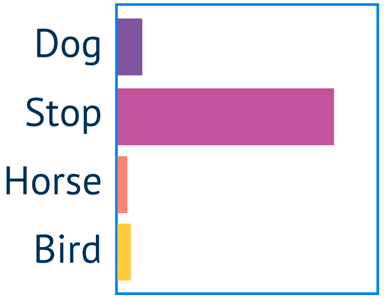
When explanations are required, we have access to a trained model, and a single data point

Given

Network $f(\theta)$



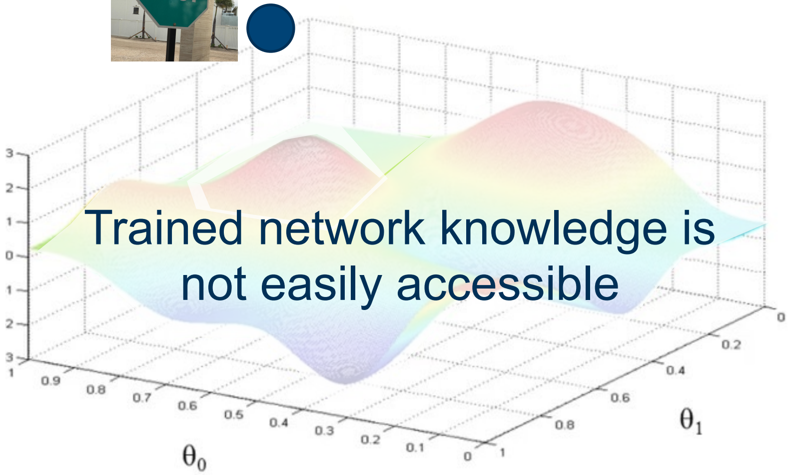
Predicted Class Probability



Explain the decision of the predicted class as a function of the learned manifold

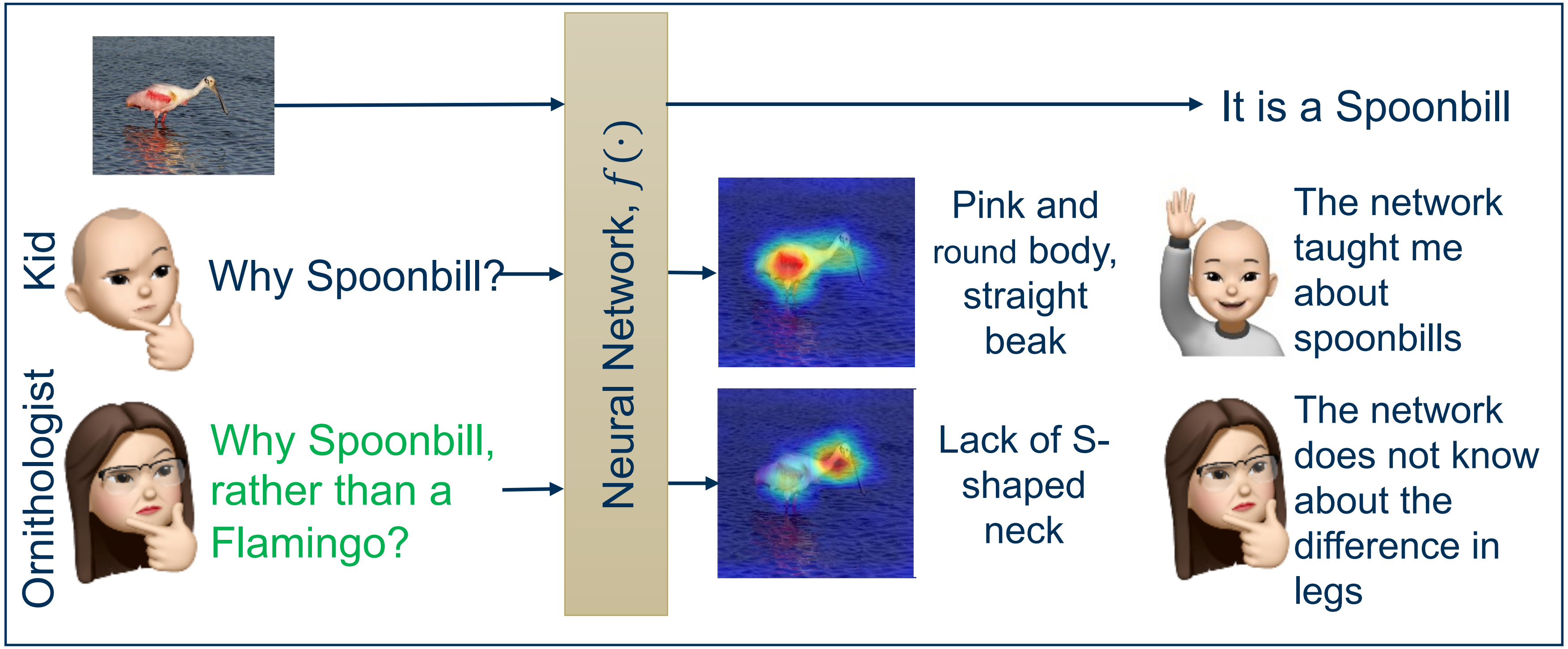
Challenge

$L(\theta)$



Functional challenges in Explainability

The requirements from explanations are contextual; These requirements are determined by the audience



AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Challenges in Explainability

Operational challenges in Explainability

Given a set of operational constraints, the goal is to find the best Explainability technique

I need a fast Explainabilty technique

Use GradCAM!

I need a contrastive technique

Use ContrastCAM!

I don’t have access to model

Use CEM!

I cannot retrain

Okay

Takeaways

Takeaways from Lecture 1

- Explainable AI is crucial for widespread adoption of Deep Learning based technologies
- Deep Learning architectures have far outpaced traditional models and Explainability techniques
- There are **no “one size fits all” explanations** and techniques
- The **technical challenge** in Explainability is to **extract relevant information** from trained neural networks
- The **functional challenge** is to **cater relevant explanations** to the audience
- The **operational challenge** is to **identify the goals** based on applications, requirements, and data

References

Lecture 1: Introduction to Explainable AI

- Berg, T., et al. "On the rise of the FinTechs—Credit scoring using digital footprints. Federal Deposit Insurance Corporation." *Center for Financial Research WP* 4 (2018): 2018
- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).
- Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).
- Kokilepersaud, K., Prabhushankar, M., Yarici, Y., AlRegib, G., & Parchami, A. (2023). Exploiting the Distortion-Semantic Interaction in Fisheye Data. *IEEE Open Journal of Signal Processing*.
- Ma, Jun, and Bo Wang. "Segment anything in medical images." *arXiv preprint arXiv:2304.12306* (2023).
- AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.