DSC 180 – SECTION A01
EXPLAINABLE AI

Jürgen P. Schulze, Ph.D.
University of California, San Diego
Fall Quarter 2020

INSTRUCTOR

- Responsibilities:
  - Associate Research Scientist at Qualcomm Institute
  - Associate Adjunct Professor in Computer Science Department
  - Director of the Immersive Visualization Center
- Research Interests:
  - Virtual and augmented reality
  - Medical data visualization
  - Rendering algorithms
  - 3D human-computer interaction

EXPLAINABLE AI

- In this capstone domain we are going to study how we can make machine learning systems more user friendly by exploiting additional knowledge we can derive from the system and present it to the user.
- These types of systems are called Explainable AI (XAI).

MOTIVATION

- A recent incident, where Amazon has built an AI tool to automate hiring decisions only to discover few years later that the model was producing biased results because the underlying data was skewed to under-represent women, was widely publicized in the media.
- There are still situations where the transparency is not operationally or regulatory critical. Recommendation engines recipients for example do not have to know how Netflix chose the next movie to recommend. But even in those situations, a question "why" **is common on consumers' minds** and satisfying that curiosity will go a long way in building user's confidence in the recommendation.

EXAMPLE

VIDEO THREAT DETECTION

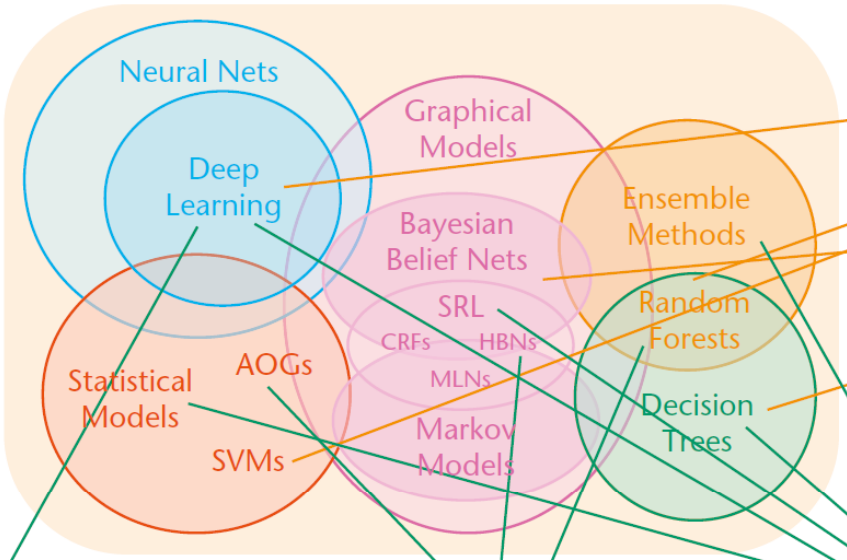
- **The Situation**
  Facilities need security to protect human and physical assets. A typical solution has been a combination of security personnel and use of video feeds. Personnel cannot monitor every entry point and every feed at all times, and due to error may miss some threats.
- **The ML/AI Solution**
  AI and Deep Learning Models evaluate the video feeds to detect threats which are then flagged for the security personnel. For example, AI image recognition models are trained to flag high risk individuals approaching an entrance (such as someone likely to carry a weapon or a known criminal), while ignoring a legitimate employee.
- **Why Explainable AI**
  Flagging an individual as a threat has a potential for significant legal implications. If the data used by AI has been trained on a sample containing disproportionate cases in some categories (say, certain minorities), a bias toward those categories is introduced.
  An individual humiliated by being intercepted, blocked or searched by security may challenge the legality of the incident. The company may be forced to justify its action in court, and disclose reasons for singling out that individual. Without the ability to explain the reason why the model flagged the individual, there would be no way to protect the company discrimination case.

DARPA XAI RESEARCH PROJECT

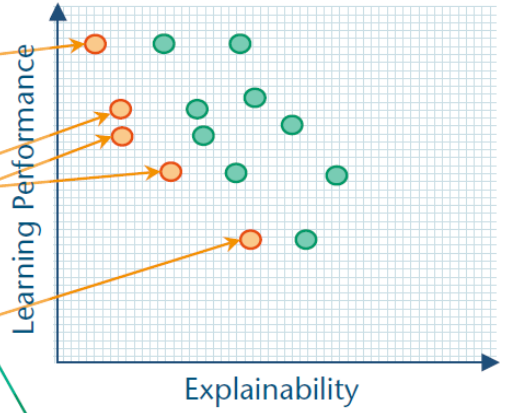
New Approach

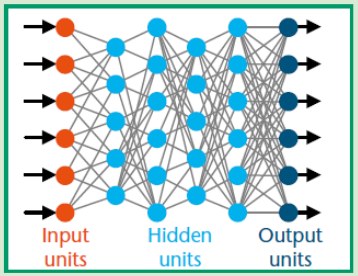
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques



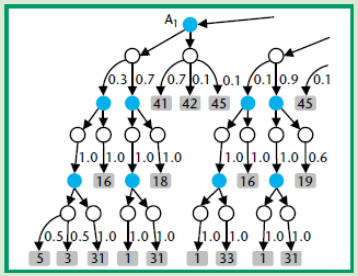
Performance vs. Explainability





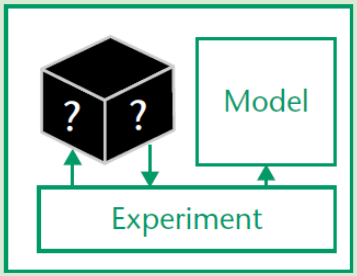
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

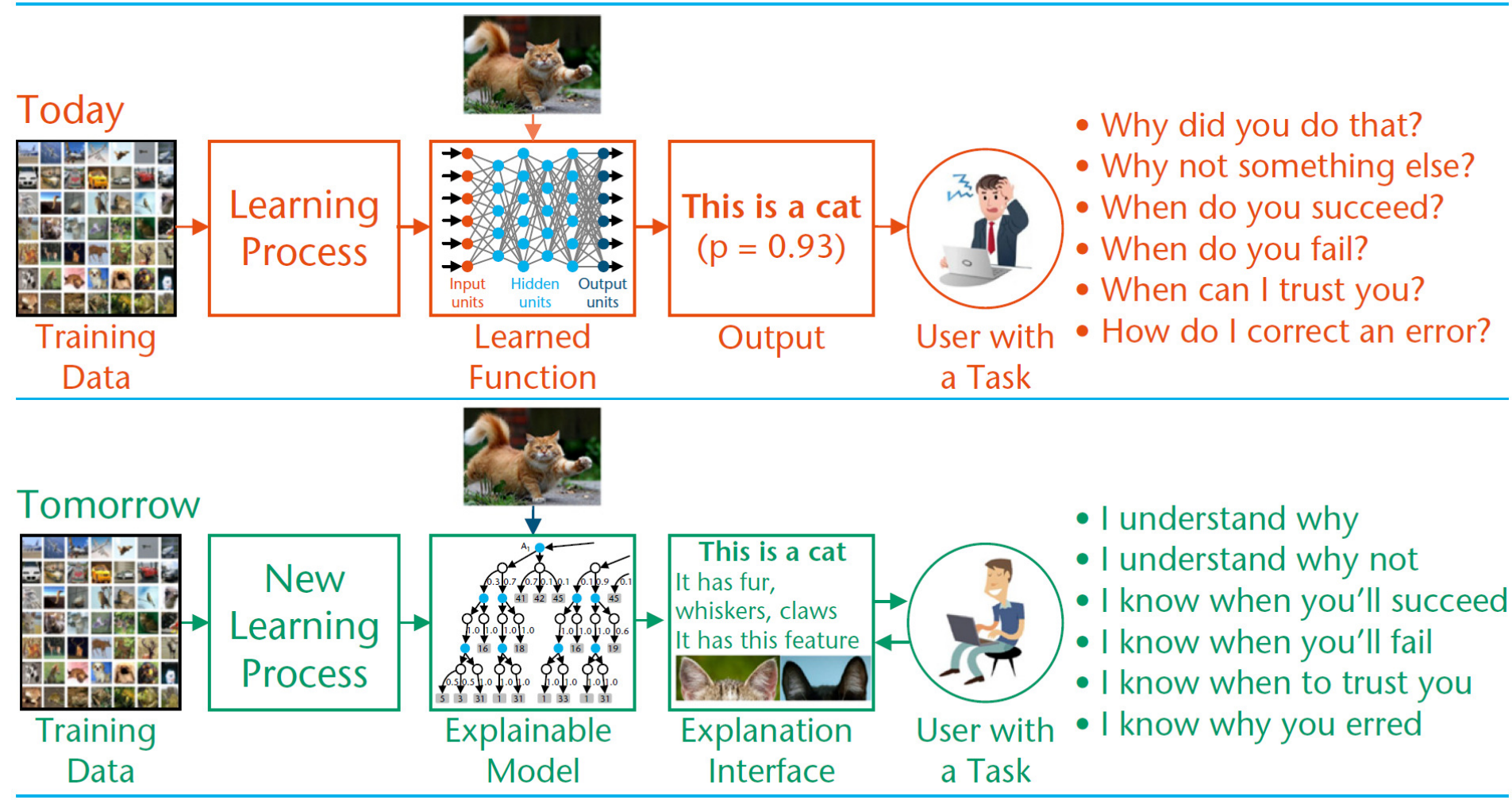
Techniques to learn more structured, interpretable, causal models



Model Induction

Techniques to infer an explainable model from any model as a black box

XAI EXAMPLE


Today
Training Data
Learning Process
Input units
Hidden units
Output units
Learned Function
This is a cat
(p = 0.93)
Output
User with a Task
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
Tomorrow
Training Data
New Learning Process
Explainable Model
This is a cat
It has fur, whiskers, claws
It has this feature
Explanation Interface
User with a Task
• I understand why
• I understand why not
• I know when you'll succeed
• I know when you'll fail
• I know when to trust you
• I know why you erred

COURSE PREREQUISITES

Familiarity with:

- Python
- Deep Learning
- Convolutional Neural Networks (CNNs)

TOPICS COVERED

- Direct visualization of CNNs
- Object recognition in images
- Attention maps

INFORMATION ON COURSE WEB SITE

URL:
http://ivl.calit2.net/wiki/index.php/DSC_Capstone2020

- Zoom and Piazza links
- Course overview
- Course Schedule
- Lecture slides
- Homework Assignments
- Deadlines
- Web links to papers and other resources

CANVAS

- For weekly assignments
- For grades
  - Let me know if a grade is missing or incorrect

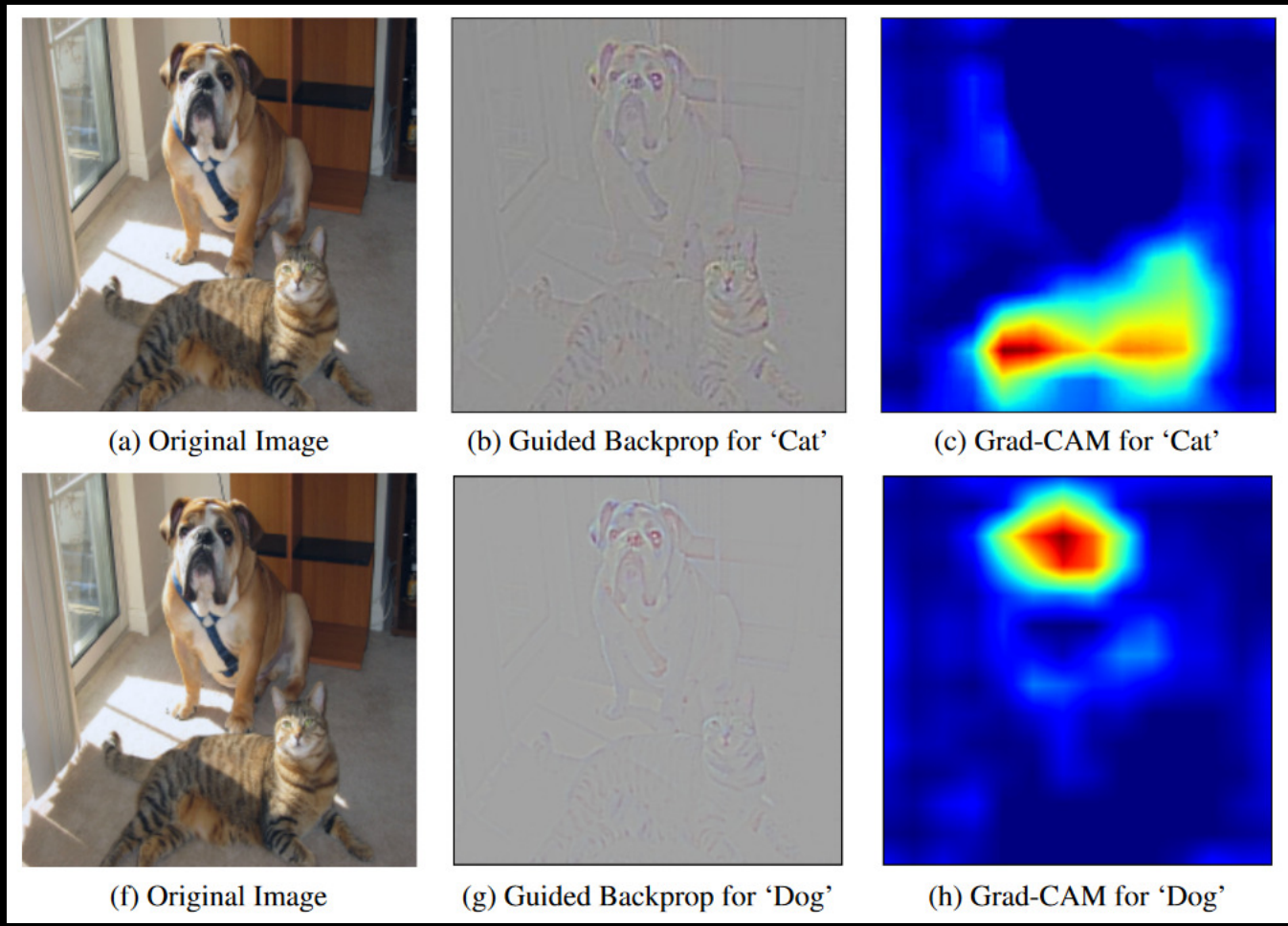
PIAZZA

- Discussion forums for
  - homework projects
  - other section related discussion

REPLICATION TASK

- The Grad-CAM technique (published 2016 by Virginia Tech researchers) aims at making CNN-based models more transparent by producing visual explanations
- The goal in this quarter is to replicate the work in the Grad-CAM (Gradient-weighted Class Activation Mapping) paper on image captioning:
  - https://arxiv.org/pdf/1610.02391v1.pdf
- The data set to apply it to is COCO
  - https://arxiv.org/abs/1405.0312

GRAD-CAM EXAMPLE



(a) Original Image (b) Guided Backprop for 'Cat' (c) Grad-CAM for 'Cat'

(f) Original Image (g) Guided Backprop for 'Dog' (h) Grad-CAM for 'Dog'

TASKS FOR NEXT WEEK

Reading

- The concept of saliency maps is going to be crucial to understand our replication paper.
- To get up to speed on these concepts, read the paper:
  - Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

Questions

- What is a saliency map?
- How is the saliency map obtained?
- What can a saliency map be used for?
- Which image data base did the authors of the paper use for training and inference?


PHILOSOPHY DEPT.
YOU ARE HERE
BUT WHY?
Coren Fishman