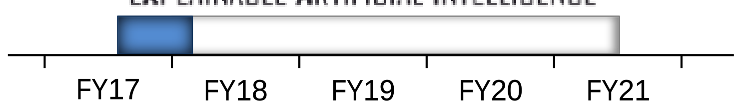
Explainable Artificial Intelligence (XAI)

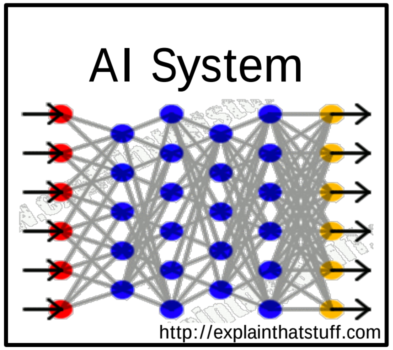


David Gunning

DARPA/I2O

Program Update November 2017

The Need for Explainable AI



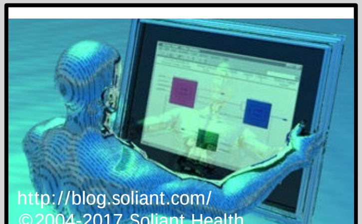
- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

Transportation

Finance

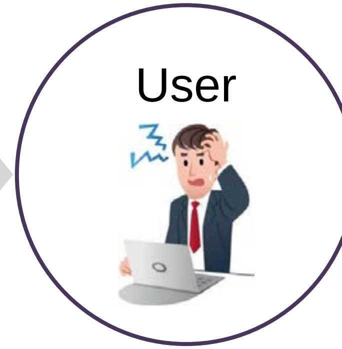
Security

Legal



Medicine

Military



- 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?

- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to users
- Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners

XAI In the News

MIT Technology Review

The Dark Secret at the Heart of AI
Will Knight
April 11, 2017

THE WALL STREET JOURNAL. WSJ

Inside DARPA's Push to Make Artificial Intelligence Explain Itself
Sara Castellanos and Steven Norton
August 10, 2017

The New York Times Magazine

Can A.I. Be Taught to Explain Itself?
Cliff Kuang
November 21, 2017

FT FINANCIAL

Intelligent Machines Are Asked to Explain How Their Minds Work
Richard Waters
July 11, 2017

The Register

You better explain yourself, mister: DARPA's mission to make an accountable AI
Dan Robinson
September 29, 2017

ExecutiveBiz

Charles River Analytics-Led Team Gets DARPA Contract to Support Artificial Intelligence Program
Ramona Adams
June 13, 2017

Entrepreneur

Elon Musk and Mark Zuckerberg Are Arguing About AI -- But They're Both Missing the Point
Artur Kiulian
July 28, 2017

Military EMBEDDED SYSTEMS

Team investigates artificial intelligence, machine learning in DARPA project
Lisa Daigle
June 14, 2017

FAST COMPANY

Why The Military And Corporate America Want To Make AI Explain Itself
Steven Melendez
June 22, 2017

NOVA NEXT

Ghosts in the Machine
Christina Couch
October 25, 2017

Jane's

DARPA's XAI seeks explanations from autonomous systems
Geoff Fein
November 16, 2017

COMPUTERWORLD

Oracle quietly researching 'Explainable AI'
George Nott
May 5, 2017

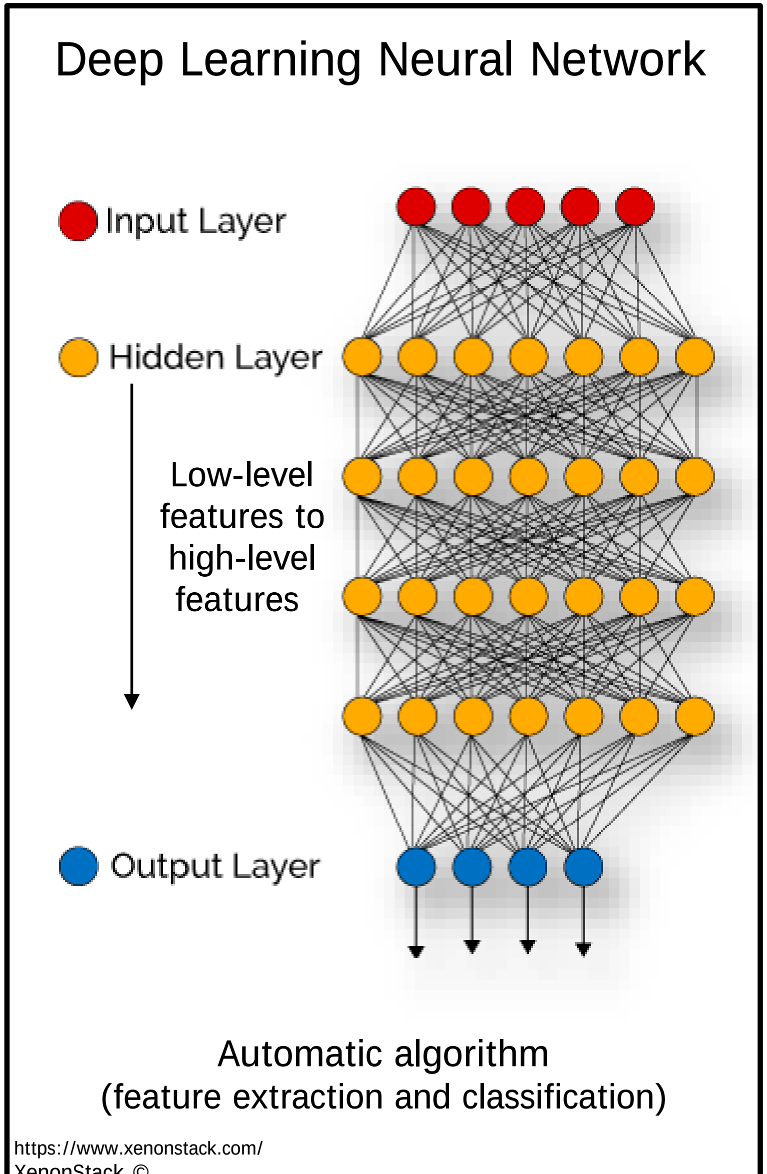
SCIENTIFIC AMERICAN

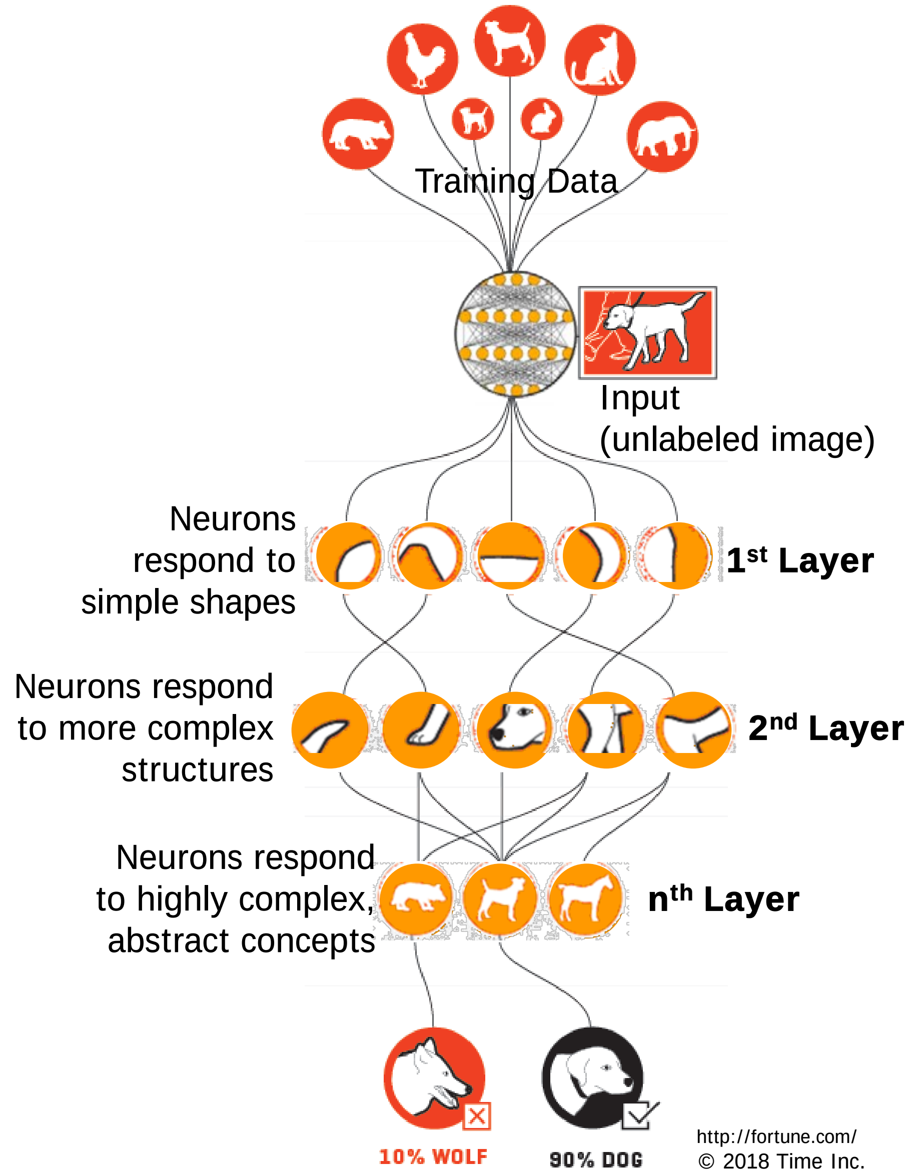
Demystifying the Black Box That Is AI
Ariel Bleicher
August 9, 2017

Science AAAS

How AI detectives are cracking open the black box of deep learning
Paul Voosen
July 6, 2017

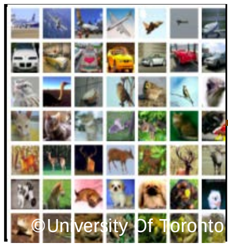
Deep Learning Neural Networks
Architecture and How They Work





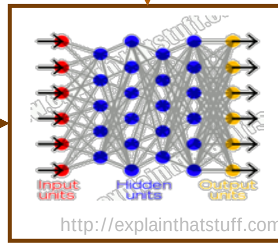
What Are We Trying To Do?

Today



Training Data

Learning Process



Learned Function

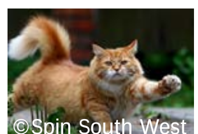
This is a cat
(p = .93)

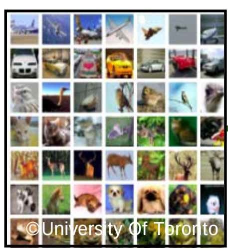
Output

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?

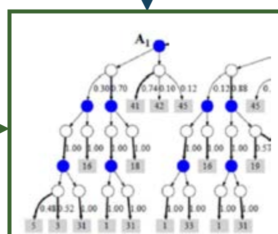
Tomorrow





Training Data

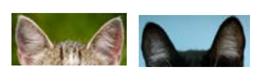
New Learning Process



Explainable Model

This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:



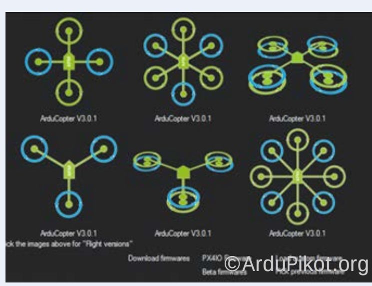
Explanation Interface

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

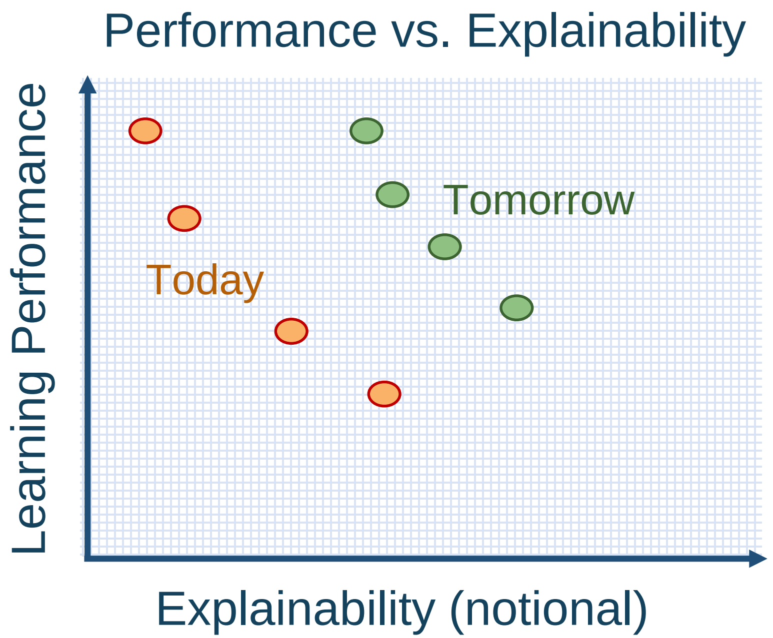
Challenge Problems

EXPLAINABLE ARTIFICIAL INTELLIGENCE

	Learn a model	Explain decisions	Use the explanation	
Data Analytics Classification Learning Task	 Multimedia Data	Explainable Model → Explanation Interface	Recommend / Explanation 	An analyst is looking for items of interest in massive multimedia data sets
	Classifies items of interest in large data set	Explains why/why not for recommended items	Analyst decides which items to report, pursue	
Autonomy Reinforcement Learning Task	 ArduPilot & SITL Simulation	Explainable Model → Explanation Interface	Actions / Explanation 	An operator is directing autonomous systems to accomplish a series of missions
	Learns decision policies for simulated missions	Explains behavior in an after-action review	Operator decides which future tasks to delegate	

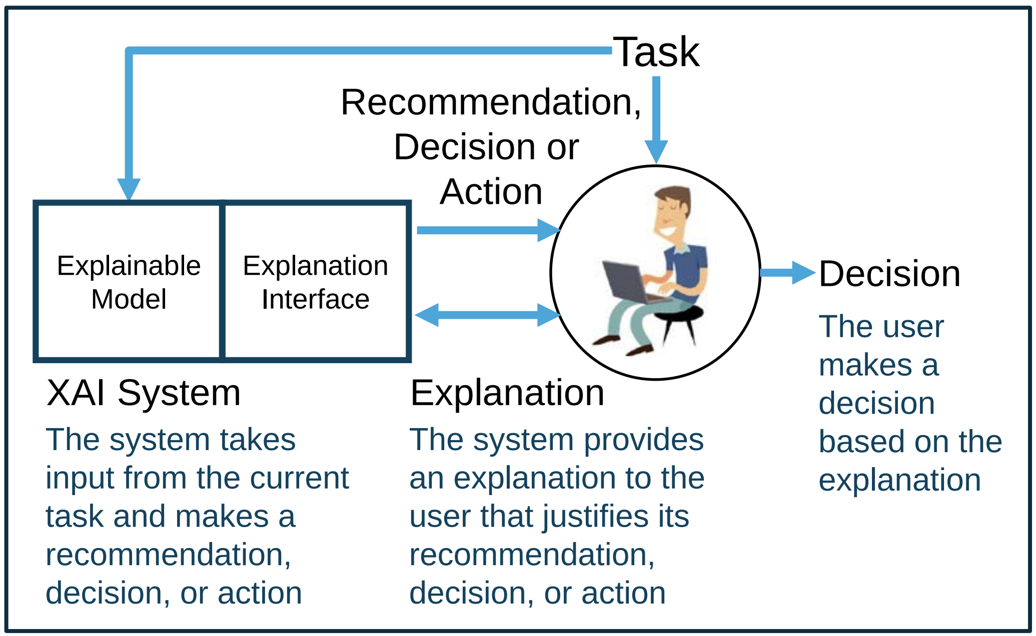
Goal: Performance and Explainability

- XAI will create a suite of machine learning techniques that
  - Produce more explainable models, while maintaining a high level of learning performance (e.g., prediction accuracy)
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners



Measuring Explanation Effectiveness

Explanation Framework



Measure of Explanation Effectiveness

User Satisfaction

- Clarity of the explanation (user rating)
- Utility of the explanation (user rating)

Mental Model

- Understanding individual decisions
- Understanding the overall model
- Strength/weakness assessment
- ‘What will it do’ prediction
- ‘How do I intervene’ prediction

Task Performance

- Does the explanation improve the user’s decision, task performance?
- Artificial decision tasks introduced to diagnose the user’s understanding

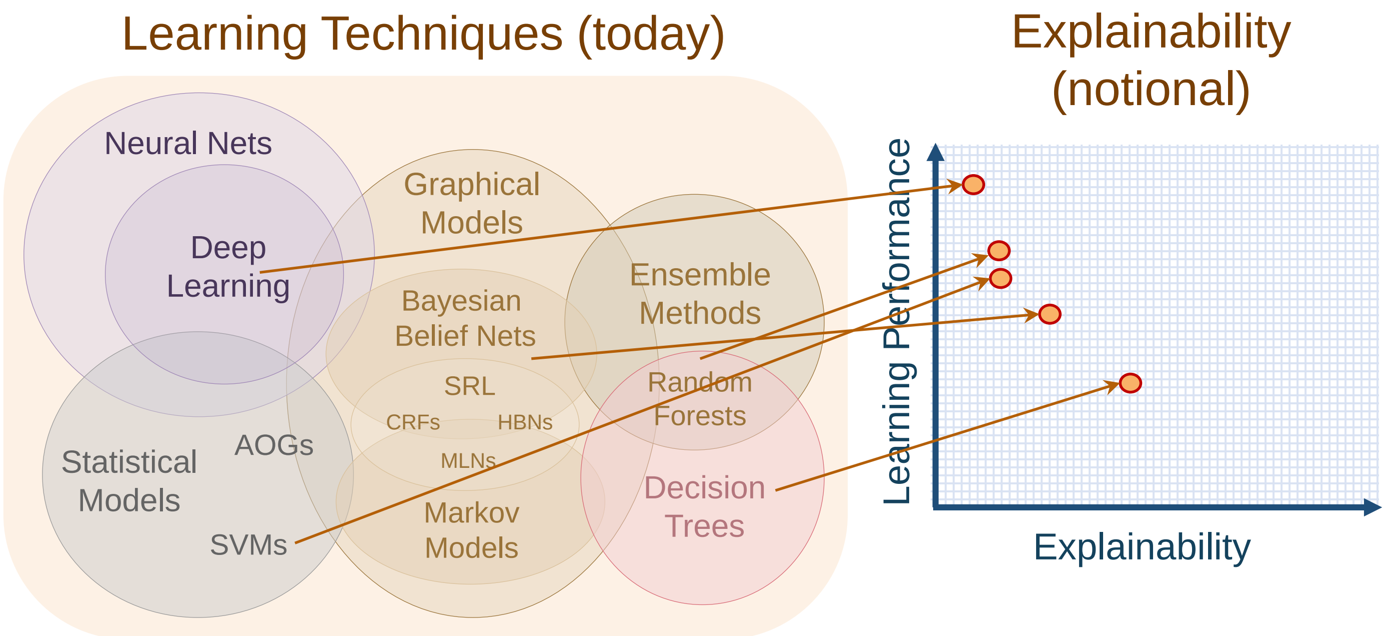
Trust Assessment

- Appropriate future use and trust

Correctability (Extra Credit)

- Identifying errors
- Correcting errors
- Continuous training

Performance vs. Explainability

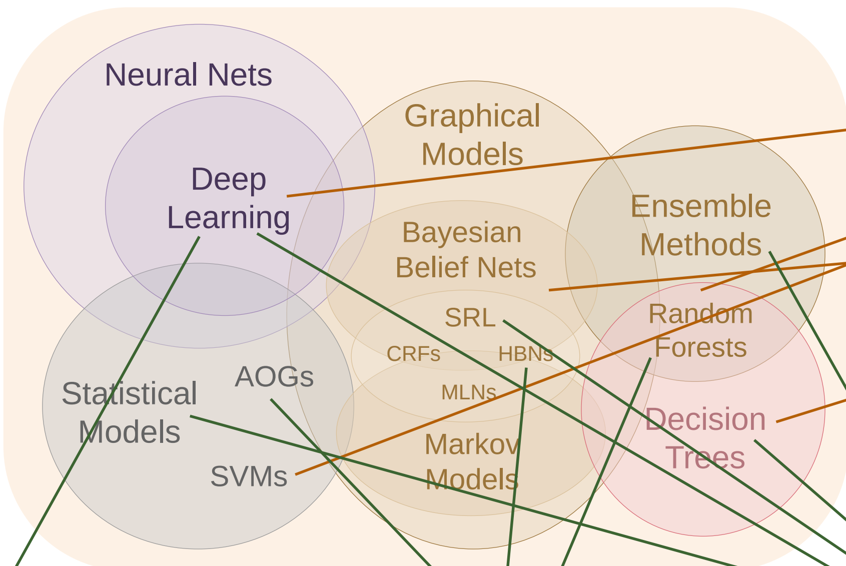


Performance vs. Explainability

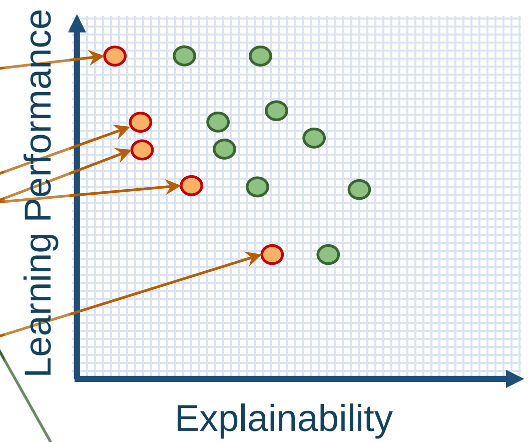
New Approach

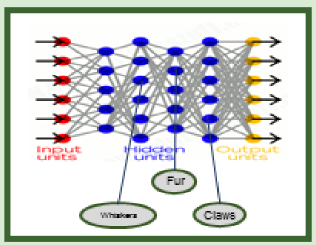
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



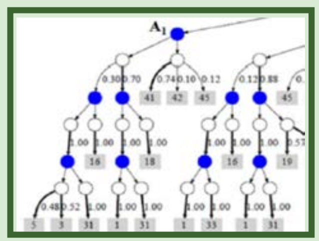
Explainability (notional)





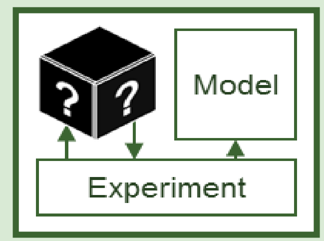
Deep Explanation

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

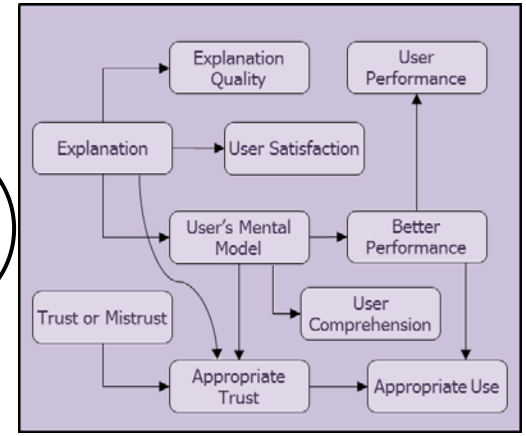
XAI Concept and Technical Approaches

Training Data

New Learning Process

Explainable Model

Explanation Interface



IHMC

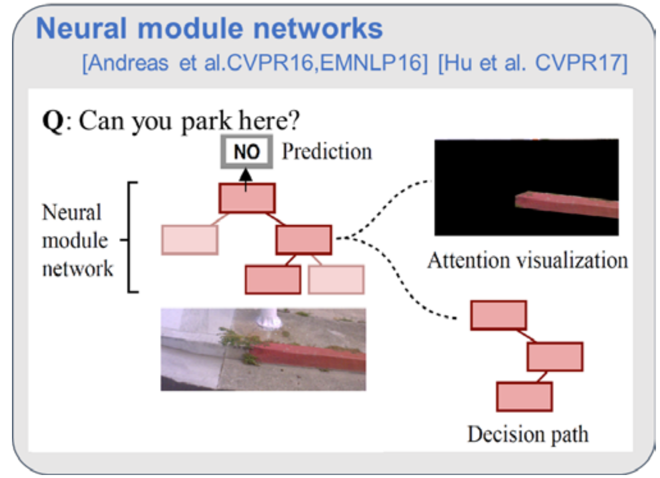
Psychological Model of Explanation

UC Berkeley	Deep Learning	Reflexive and Rational
Charles River Analytics	Causal Modeling	Narrative Generation
UCLA	Pattern Theory+	3-Level Explanation
Oregon State	Adaptive Programs	Acceptance Testing
PARC	Cognitive Modeling	Interactive Training
CMU	Explainable RL (XRL)	XRL Interaction
SRI International	Deep Learning	Show and Tell Explanations
Raytheon BBN	Deep Learning	Argumentation and Pedagogy
UT Dallas	Probabilistic Logic	Decision Diagrams
Texas A&M	Mimic Learning	Interactive Visualization
Rutgers	Model Induction	Bayesian Teaching

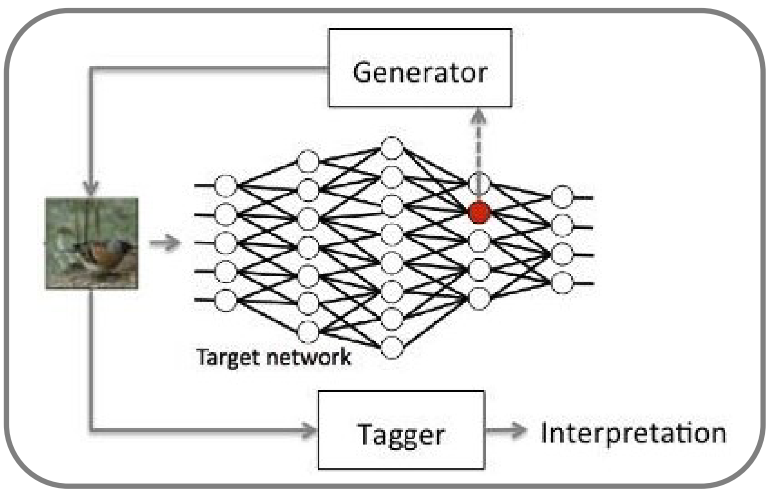
Approaches to Deep Explanation
(Berkeley, SRI, Raytheon BBN, OSU, CRA, PARC)

Attention Mechanisms

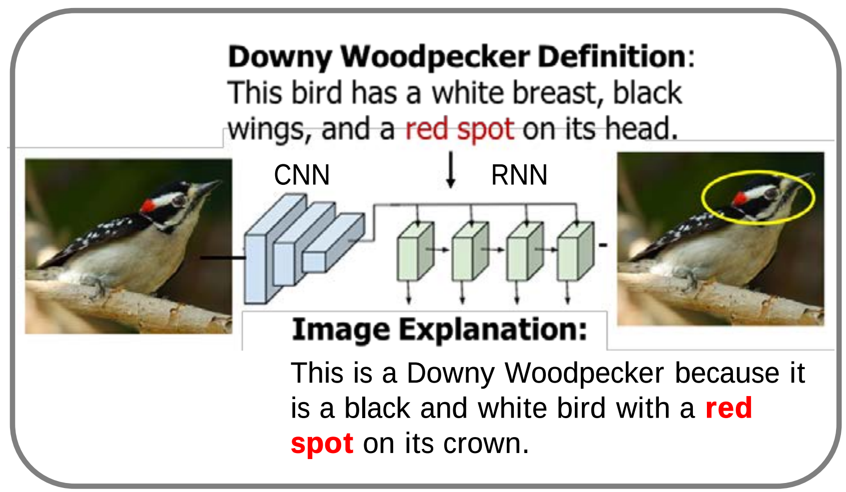
Modular Networks



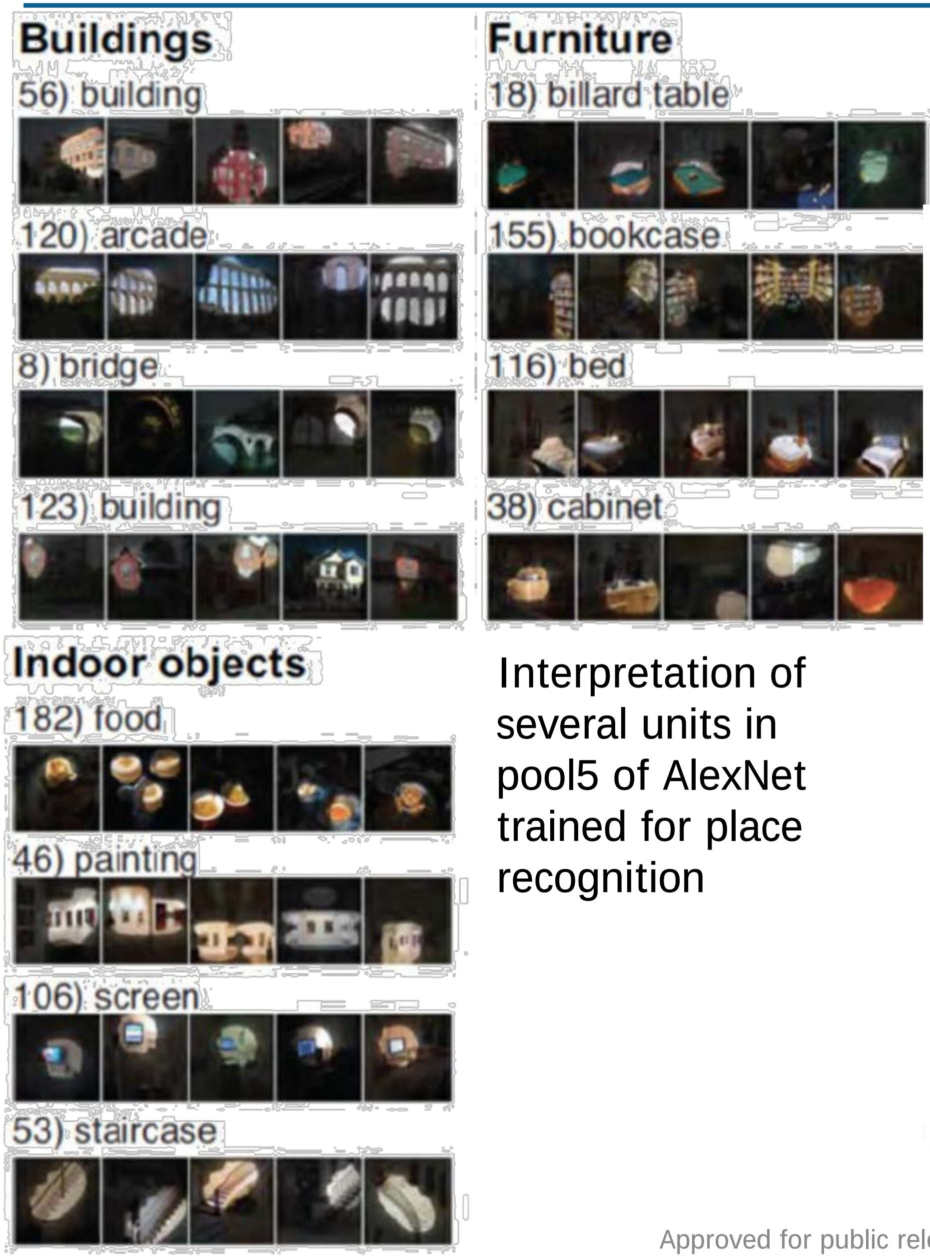
Feature Identification



Learn to Explain

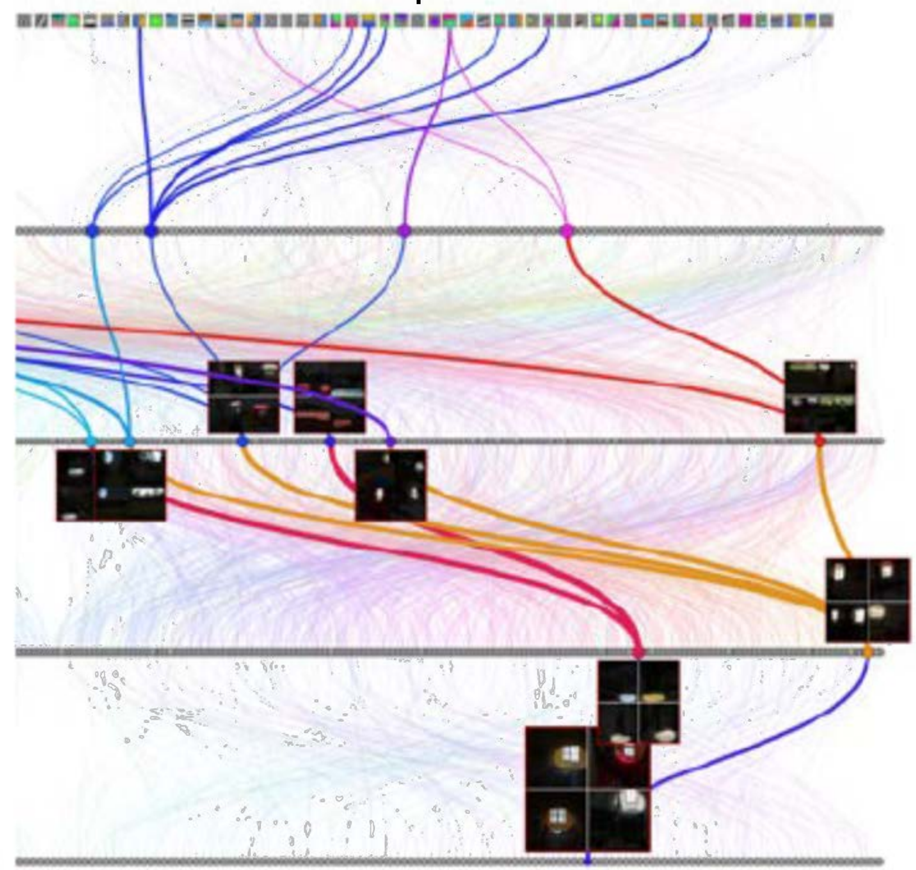


Network Dissection Quantifying Interpretability of Deep Representations (MIT)

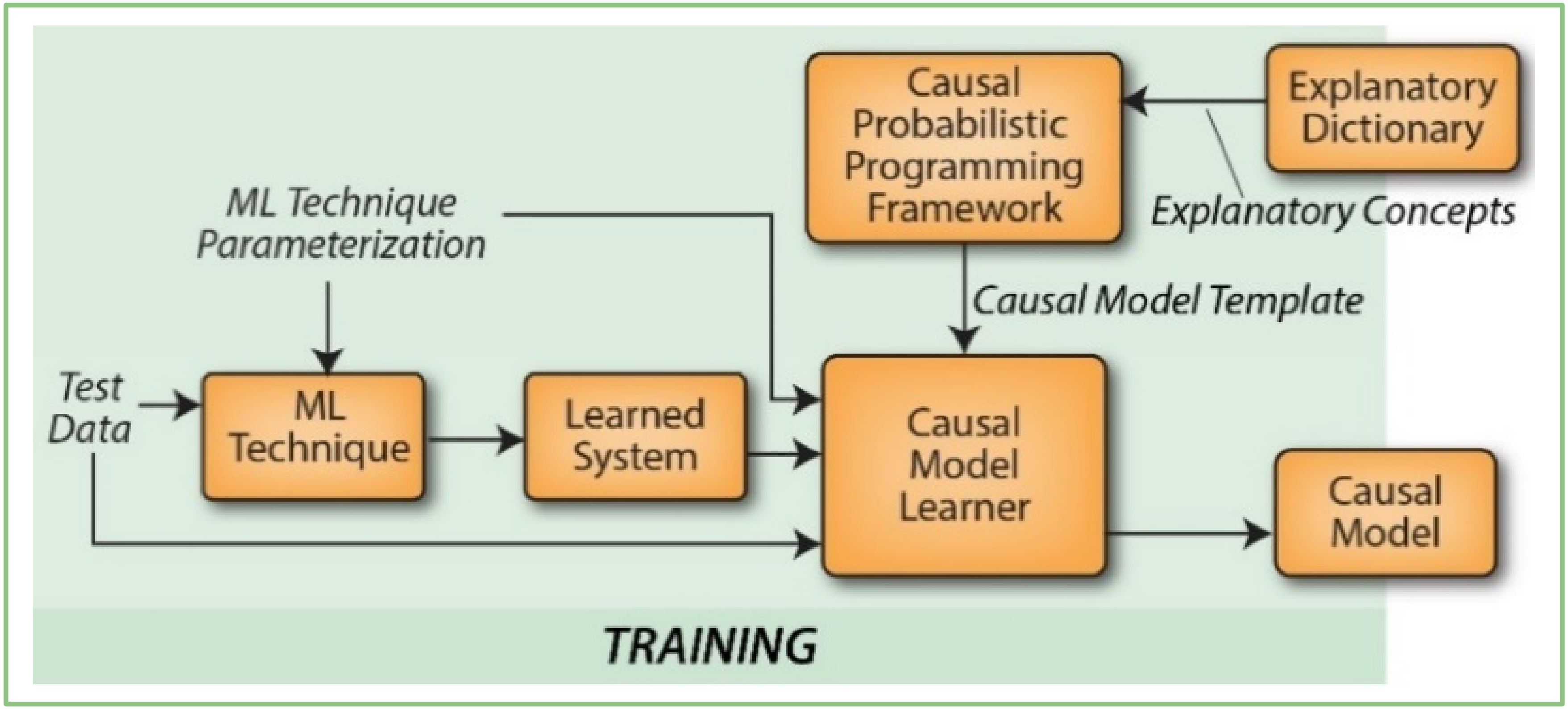


Interpretation of several units in pool5 of AlexNet trained for place recognition

Audit trail: for a particular output unit, the drawing shows the most strongly activated path

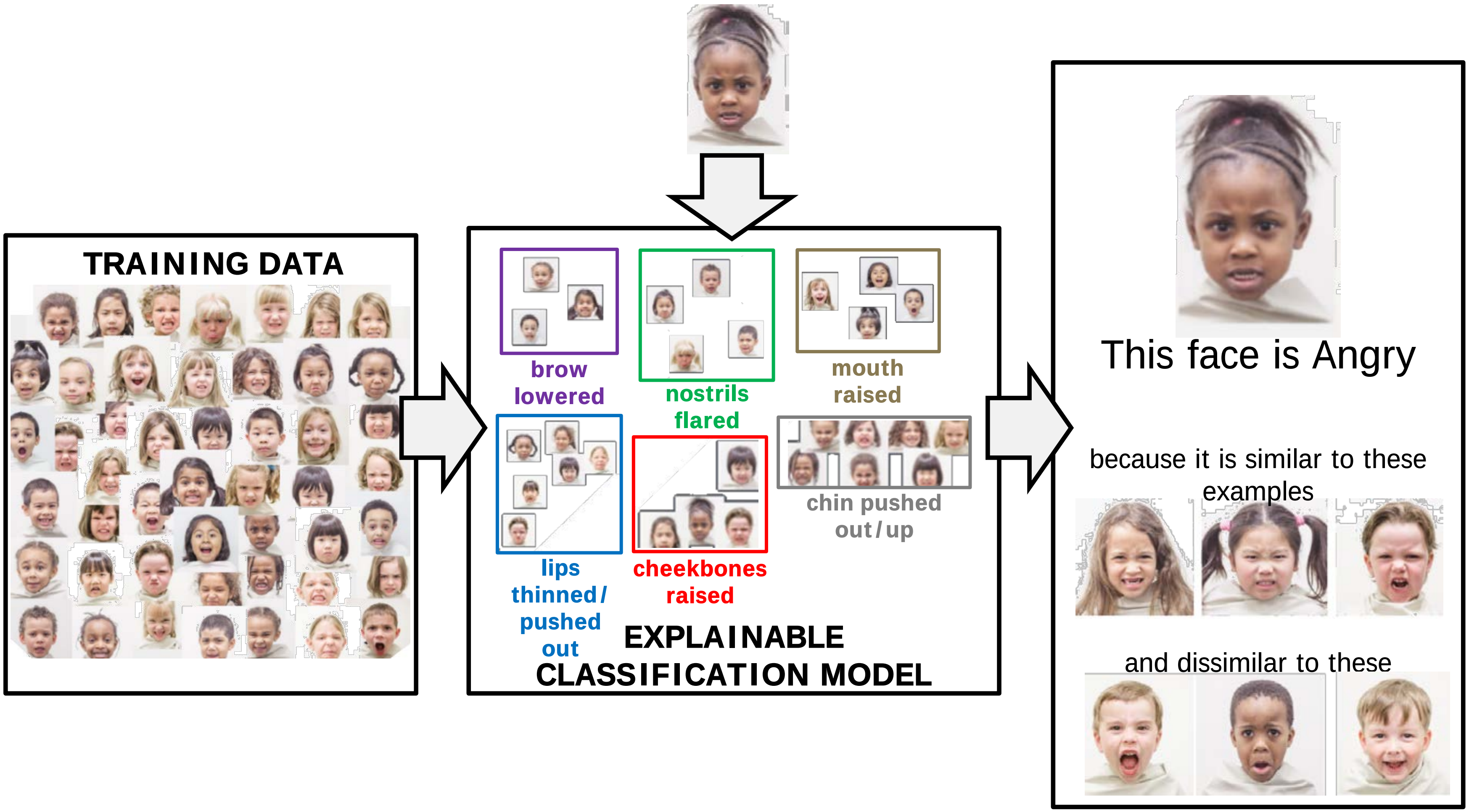


Causal Model Induction (CRA)



<u>Causal Model Induction</u>: Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model

Explanation by Selection of Teaching Examples (Rutgers)

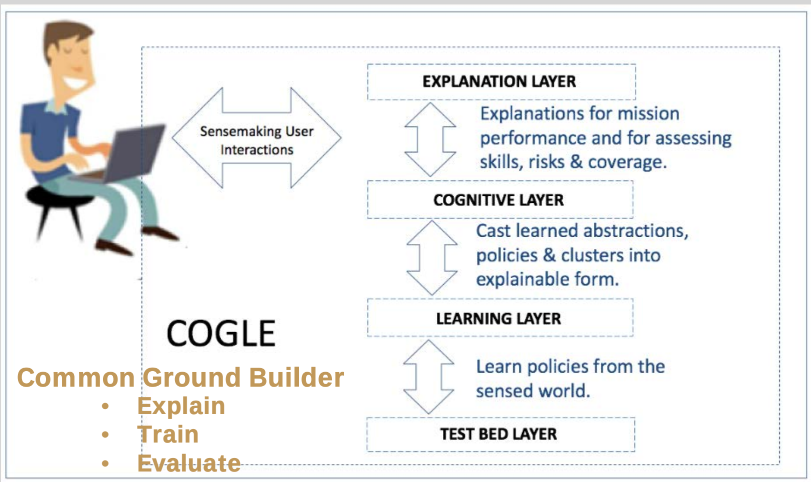


BAYESIAN TEACHING for optimal selection of examples for machine explanation

Autonomy (PARC, OSU)

Common Ground Learning and Explanation (COGLE)

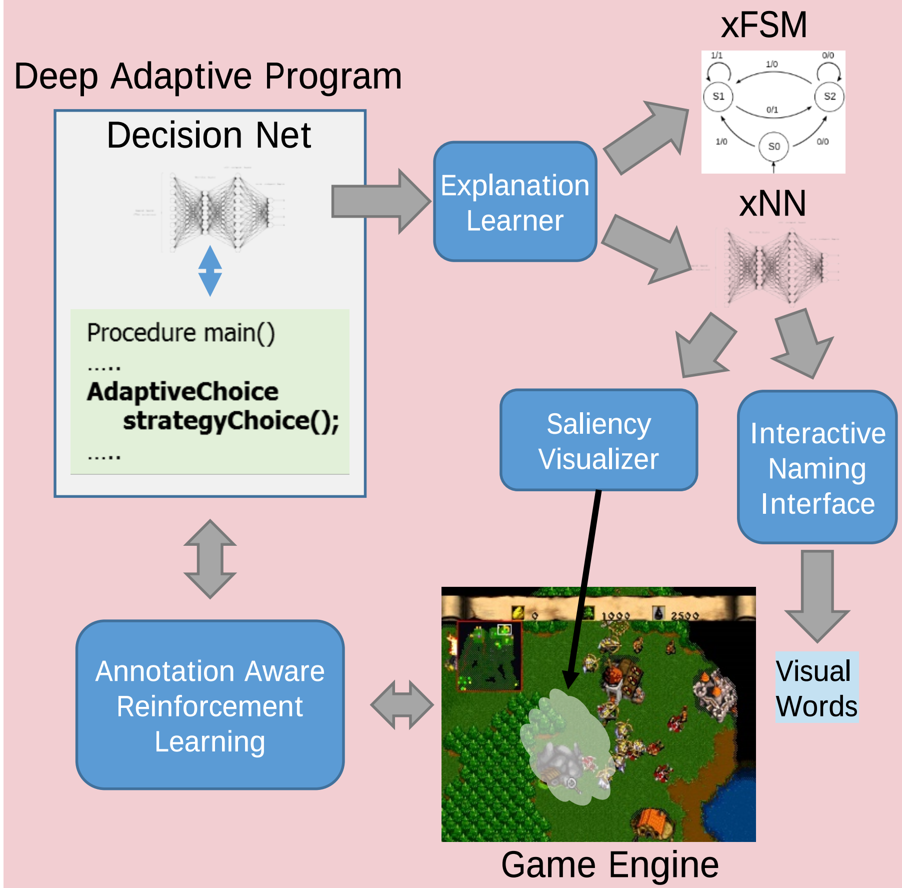
An interactive sensemaking system to explain the learned performance capabilities of a UAS flying in an ArduPilot simulation testbed



Robotics Curriculum

Explanation-Informed Acceptance Testing of Deep Adaptive Programs (xACT)

Tools for explaining deep adaptive programs and discovering best principles for designing explanation user interfaces



Four Modes of Explanation (Raytheon BBN)

Explanation Modes

Analytic (didactic) statements

in natural language that describe the elements and context that support a choice

Visualizations

that directly highlight portions of the raw data that support a choice and allow viewers to form their own perceptual understanding

Cases

that invoke specific examples or stories that support the choice

Rejections of alternative choices

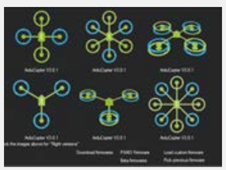
(or "common misconceptions" in pedagogy) that argue against less preferred answers based on analytics, cases, and data

XAI Program Structure

XAI
EXPLAINABLE ARTIFICIAL INTELLIGENCE

Challenge Problem Areas

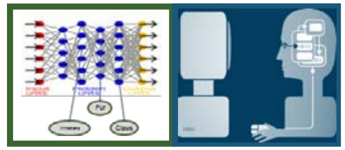
Data Analytics
Multimedia Data



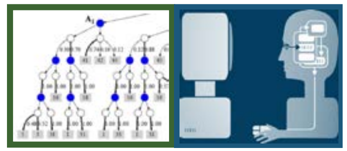
Autonomy
ArduPilot & SITL Simulation

TA1: Explainable Learners

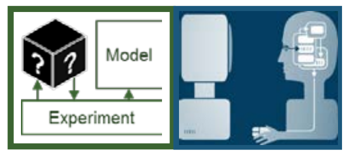
Teams that provide prototype systems with both components
- Explainable Model
- Explanation Interface



Deep Learning Teams



Interpretable Model Teams

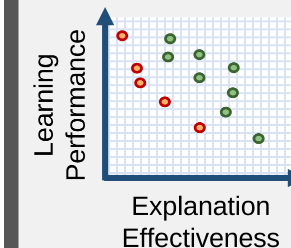


Model Induction Teams

TA2: Psychological Model of Explanation

- Psychological Theory of Explanation
- Computational Model Consulting

Evaluation Framework



Explanation Measures
- User Satisfaction
- Mental Model
- Task Performance
- Trust Assessment
- Correctability

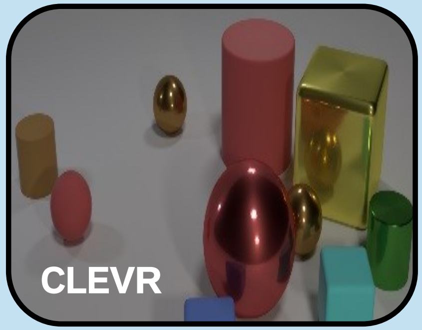
Evaluator
Naval Research Laboratory

- TA1: Explainable Learners
  - Multiple TA1 teams will develop prototype explainable learning systems that include both an explainable model and an explanation interface
- TA2: Psychological Model of Explanation
  - At least one TA2 team will summarize current psychological theories of explanation and develop a computational model of explanation from those theories

Challenge Problem Candidates

Analytics

Visual Question Answering

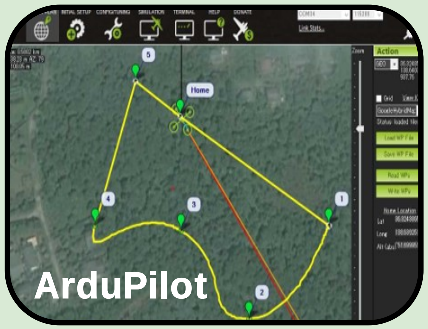


Activity Recognition

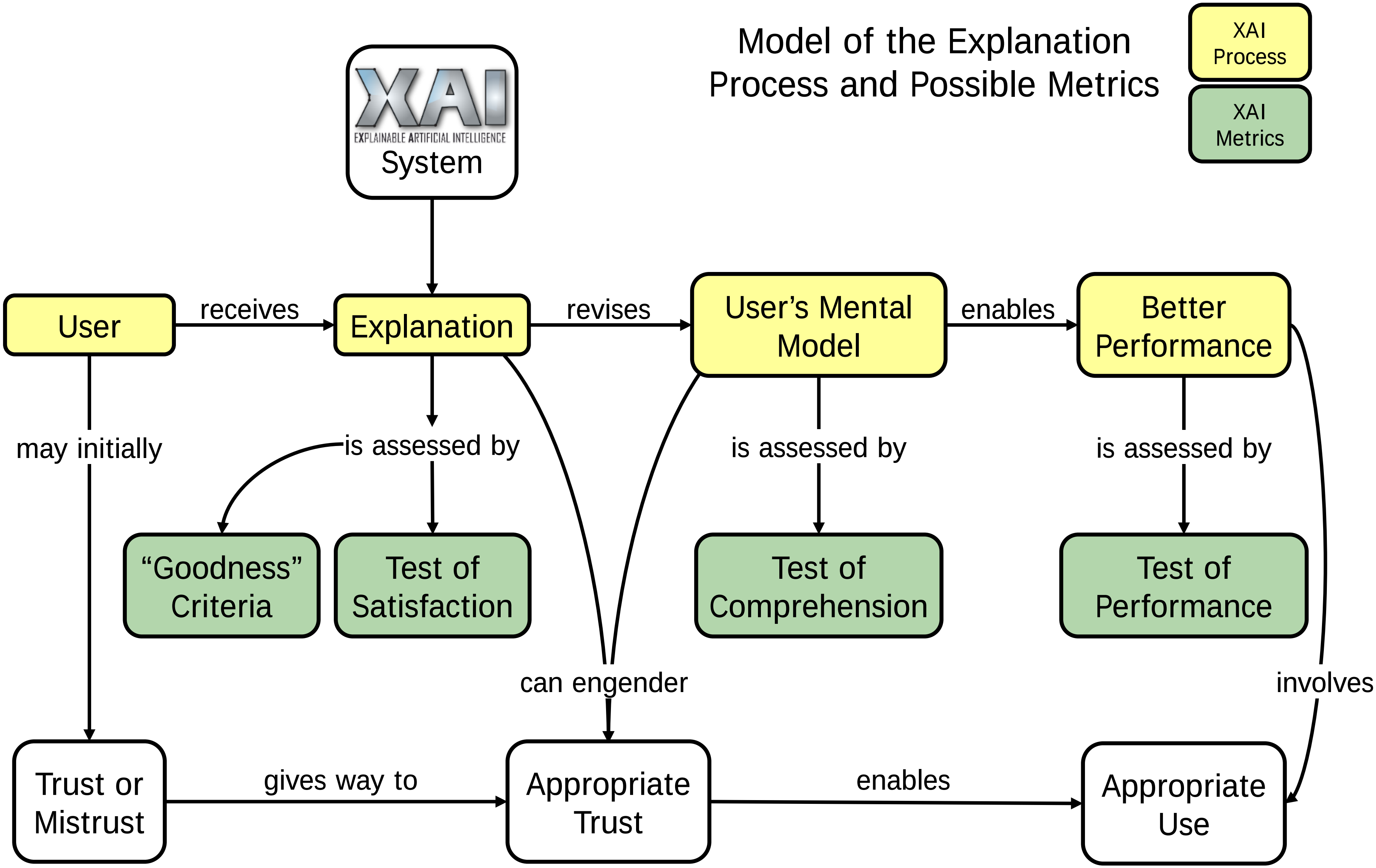
Autonomy

Strategy Games

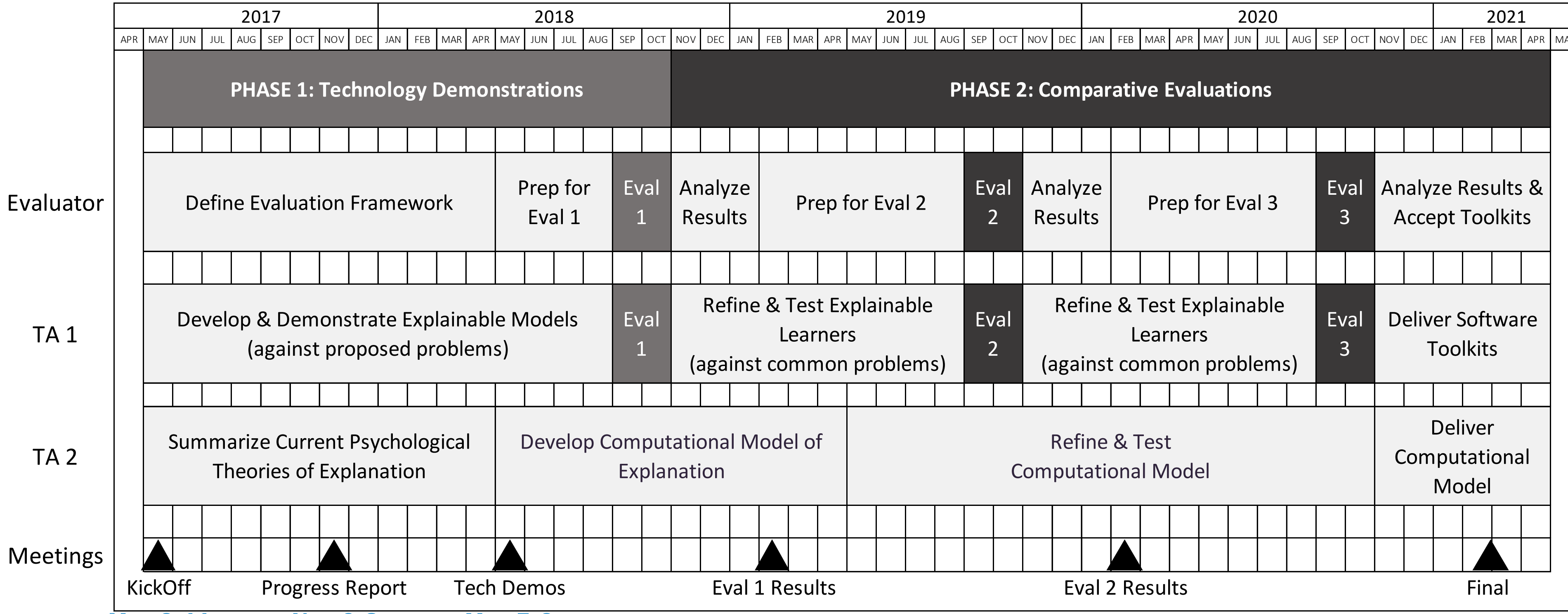
Vehicle Control



Psychological Model of Explanation (IHMC)



Schedule and Milestones



May 9-11 **Nov 6-8** **May 7-9**

- Technical Area 1 (Explainable Learners) Milestones
  - Demonstrate the explainable learners against problems proposed by the developers (Phase 1)
  - Demonstrate the explainable learners against common problems (Phase 2)
  - Deliver software libraries and toolkits (at the end of Phase 2)
- Technical Area 2 (Psychology of Explanation) Milestones
  - Deliver an interim report on psychological theories (after 6 months during Phase 1)
  - Deliver a final report on psychological theories (after 12 months, during Phase 1)
  - Deliver a computational model of explanation (after 24 months, during Phase 2)
  - Deliver the computational model software (at the end of Phase 2)

XAI Developers (TA1)

CP	Performer	Explainable Model	Explanation Interface
Both	UC Berkeley	Deep Learning	Reflexive and Rational
	Charles River	Causal Modeling	Narrative Generation
	UCLA	Pattern Theory+	3-level Explanation
Autonomy	Oregon State	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL (XRL)	XRL Interaction
Analytics	SRI International	Deep Learning	Show and Tell Explanation
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UT Dallas	Probabilistic Logic	Decision Diagrams
	Texas A&M	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

Deeply Explainable Artificial Intelligence

Explainable Model	Explanation Interface	Challenge Problem
Deep Learning • Explain *implicit* (latent) nodes by training additional DL models • Explain *explicit* nodes thru Neural Module Networks (NMNs)	**Reflexive & Rational** • Reflexive explanations (that arise directly from the model) • Rational explanations (that come from reasoning about user's beliefs)	**Autonomy** • ArduPilot and OpenAI Gym Simulations **Data Analytics** • Visual QA and Multimedia Event QA

- **PI**: Trevor Darrell (Berkeley)

- Pieter Abbeel (Berkeley)
- Tom Griffiths (Berkeley)
- Kate Saenko (BU)
- Zeynep Akata (U. Amsterdam)
- Dan Klein (Berkeley)
- John Canny (Berkeley)
- Anca Dragan (Berkeley)
- Anthony Hoogs (Kitware)

CAMEL: Causal Models to Explain Learning

Explainable Model

Model Induction Causal Models

- Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model

Explanation Interface

Narrative Generation

- Interactive visualization based on the generation of temporal, spatial narratives from the causal, probabilistic models

Challenge Problem

Autonomy

- Minecraft, Starcraft

Data Analytics

- Pedestrian Detection (INRIA), Activity Recognition (ActivityNet)

- **PI**: Brian Ruttenberg (CRA)

- Avi Pfeffer (CRA)
- David Jensen (U. Mass)
- Michael Littman (Brown)
- James Niehaus (CRA)
- Emilie Roth (Roth Cognitive Engineering)
- Joe Gorman(CRA)
- James Tittle (CRA)

Learning and Communicating Explainable Representations for Analytics and Autonomy

Explainable Model

Pattern Theory+

- Integrated representation across an entropy spectrum:
  - Deep Neural Nets
  - Stochastic And-Or-Graphs (AOG)
  - Predicate Calculus

Explanation Interface

3-Level Explanation

- Integrate 3 levels of explanation:
  - Concept compositions
  - Causal and counterfactual reasoning
  - Utility explanations

Challenge Problem

Autonomy

- Humanoid robot behavior and VR simulation platform

Data Analytics

- Understanding complex multimedia events

- **PI**: Song-Chun Zhu (UCLA)

- Ying Nian Wu (UCLA)
- Sinisa Todorovic (OSU)
- Joyce Chai (Michigan State)

xACT: Explanation-Informed Acceptance Testing of Deep Adaptive Programs

Explainable Model | Explanation Interface | Challenge Problem

Adaptive Programs

- Explainable Deep Adaptive Programs (xDAPs) – a new combination of Adaptive Programs, Deep Learning and explainability

Acceptance Testing

- Provides a visual & NL explanation interface for acceptance testing by test pilots based on Information Foraging Theory

Autonomy

- Real-Time Strategy Games based on custom designed game engine designed to support explanation
- Possible use of Starcraft

- **PI**: Alan Fern (OSU)

- Tom Dietterich (OSU)
- Fuxin Li (OSU)
- Prasad Tadepalli (OSU)
- Weng-Keen Wong (OSU)
- Margaret Burnett (OSU)
- Martin Erwig (OSU)
- Liang Huang (OSU)

COGLE: Common Ground Learning and Explanation

Explainable Model

Cognitive Model

- 3-layer architecture:
  - Learning Layer (DNNs)
  - Cognitive Layer (ACT-R Cog. Model)
  - Explanation Layer (HCI)

Explanation Interface

Interactive Training

- Interactive visualization of states, actions, policies & values
- Includes a module for test pilots to refine and train the system

Challenge Problem

Autonomy

- ArduPilot simulation environment
- *Value of Explanation* (VoE) framework for measuring explanation effectiveness

- **PI**: Mark Stefik (PARC)

- Honglak Lee (U. Mich.)
- Subramanian Ramamoorthy (U. Edinburgh)
- Christian Lebiere (CMU)
- John Anderson (CMU)
- Robert Thomson (USMA)
- Michael Youngblood (PARC)

XRL: Explainable Reinforcement Learning for AI Autonomy

Explainable Model

XRL Models

- Create a new scientific discipline for Explainable Reinforcement Learning with work on new algorithms and representations

Explanation Interface

XRL Interaction

- Interactive explanations of dynamic systems
- Human-machine interaction to improve performance

Challenge Problem

Autonomy

- Open AI Gym
- Autonomy in the electrical grid
- Mobile service robots
- Self-improving educational software

- **PI**: Geoff Gordon (CMU)

- Zico Kolter (CMU)
- Pradeep Ravikumar (CMU)
- Manuela Veloso (CMU)
- Emma Brunskill (Stanford)

DARE: Deep Attention-based Representations for Explanation

Explainable Model

Deep Learning

- Multiple deep learning techniques:
  - Attention-based mechanisms
  - Compositional NMNs
  - GANs

Explanation Interface

Show-and-Tell Explanations

- DNN visualization
- Query evidence that explains DNN decisions
- Generate natural language justifications

Challenge Problem

Data Analytics

- Visual Question Answering (VQA) using Visual Gnome, Flickr30
- MovieQA

- **PIs**: Giedrius Burachas (SRI), Mohamed Amer (SRI)

- Shalini Ghosh (SRI)
- Avi Ziskind (SRI)
- Michael Wessel (SRI)
- Richard R. Zemel (U. Toronto)
- Sanja Fidler (U. Toronto)
- David Duvenaud (U. Toronto)
- Graham Taylor (U. Guelph)
- Jürgen Schulze (UCSD)

EQUAS: Explainable QUestion Answering System

Explainable Model

Deep Learning

- Semantic labelling of DNN neurons
- DNN audit trail construction
- Gradient-weighted Class Activation Mapping

Explanation Interface

Argumentation Theory

- Comprehensive strategy based on argumentation theory
- NL generation
- DNN visualization

Challenge Problem

Data Analytics

- Visual Question Answering (VQA), beginning with images and progressing to video

- **PI**: William Ferguson (Raytheon BBN)

- Antonio Torralba (MIT)
- Ray Mooney (UT Austin)
- Devi Parikh (GA Tech)
- Dhruv Batra (GA Tech)

Tractable Probabilistic Logic Models: A New, Deep Explainable Representation

Explainable Model

Probabilistic Logic

- Tractable Probabilistic Logic Models (TPLMs) – an important class of (non-deep learning) interpretable models

Explanation Interface

Probabilistic Decision Diagrams

- Enables users to explore and correct the underlying model as well as add background knowledge

Challenge Problem

Data Analytics

- Infer activities in multimodal data (video and text)
- Using the Wetlab (biology) and TACoS (cooking) datasets

- **PI**: Vibhav Gogate (UT Dallas)

- Adnan Darwiche (UCLA)
- Guy Van Den Broeck (UCLA)
- Nicholas Ruozzi (UT Dallas)
- Eric Ragan (Texas A&M)
- Parag Singla (IIT-Delhi)

Transforming Deep Learning to Harness the Interpretability of Shallow Models: An Interactive End-to-End System

Explainable Model

Mimic Learning

- Develop a mimic learning framework that combines deep learning models for prediction and shallow models for explanations

Explanation Interface

Interactive Visualization

- Interactive visualization over multiple views, using heat maps & topic modeling clusters to show predictive features

Challenge Problem

Data Analytics

- Multiple tasks using data from Twitter, Facebook, ImageNet, UCI, NIST and Kaggle
- Metrics for explanation effectiveness

- **PI**: Xia Hu (Texas A&M)

- Shuiwang Ji (Wash. State)
- Eric Ragan (Texas A&M)

Model Explanation by Optimal Selection of Teaching Examples

Explainable Model

Model Induction

- Select the optimal training examples to explain model decisions based on Bayesian Teaching

Explanation Interface

Bayesian Teaching

- Example-based explanation of:
  - the full model
  - user-selected sub-structure
  - user submitted examples

Challenge Problem

Data Analytics

- Movie descriptions
- Image processing
- Caption data
- Movie events
- Human motion events

- **PI**: Patrick Shafto (Rutgers)

- Scott Cheng-Hsin Yang (Rutgers)

Naturalistic Decision Making Foundations of Explainable AI

Literature Review

Naturalistic Theory

- Extensive review of relevant psychological theories
- Extend the theory of Naturalistic Decision Making to cover explanation

Computational Model

Bayesian Framework

- Represent reductionist mental models that humans develop as part of the explanatory process
- Including mental simulation

Model Validation

Experiments

- Conduct interactive assessment and formal human experiments
- Validate the model
- Develop metrics of explanation effectiveness

- **PI**: Robert R. Hoffman (IHMC)

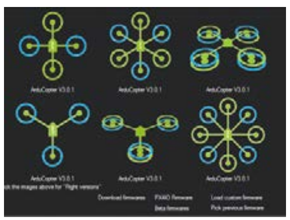
- Gary Klein (MacroCognition)
- Shane T. Mueller (Michigan Tech)
- William J. Clancey (IHMC)
- COL Timothy M. Cullen (SAASS)
- Jordan Litman (IHMC Psychometrician)
- Simon Attfield (Middlesex University-London)
- Peter Pirolli (IHMC)

XAI Evaluation

Challenge Problems

Analytics

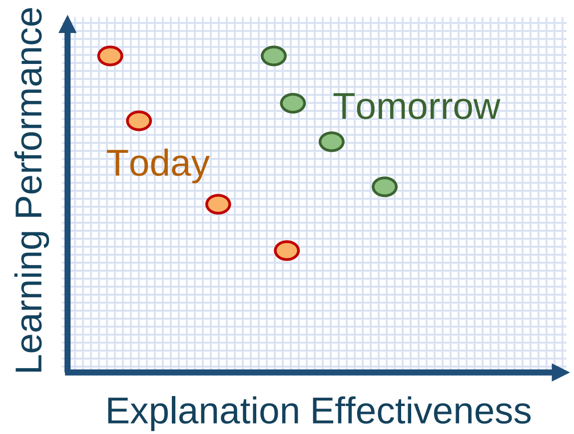
Autonomy



Evaluation Framework

- Evaluation protocols
- Training environment
  - Training data
  - Simulation environment
- Testing environment
  - Subjects
  - Web infrastructure
- Baseline systems

Measurement



- **PI**: David Aha (NRL)

- Justin Karneeb (Knexus)
- Matt Molineaux (Knexus)
- Leslie Smith (NRL)
- Mike Pazzani (UC Riverside)

www.darpa.mil