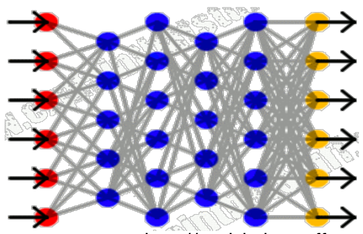
The Need for Explainable AI

AI System



http://explainthatstuff.com

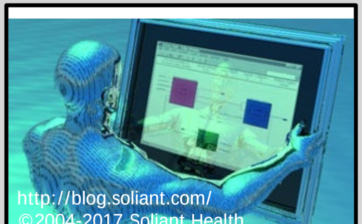
- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

Transportation

Finance

Security

Legal



Medicine

Military

User

- 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

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

NOVA NEXT

Ghosts in the Machine
Christina Couch
October 25, 2017

FAST COMPANY

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

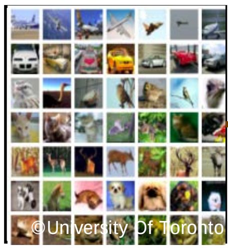
Demystifying the Black Box That Is AI
Ariel Bleicher
August 9, 2017

Science AAAS

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

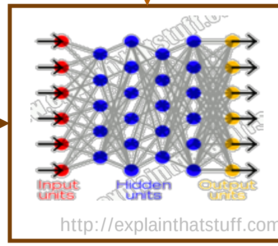
What Are We Trying To Do?

Today



Training Data

Learning Process



Learned Function

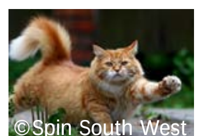
This is a cat
(p = .93)

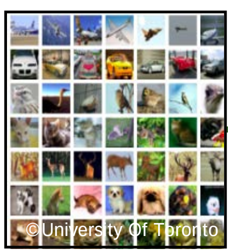
Output

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?

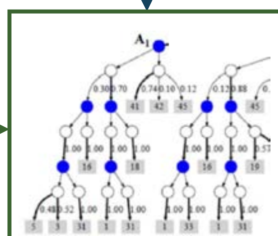
Tomorrow





Training Data

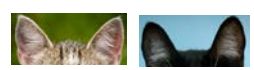
New Learning Process



Explainable Model

This is a cat:

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



Explanation Interface

User with a Task

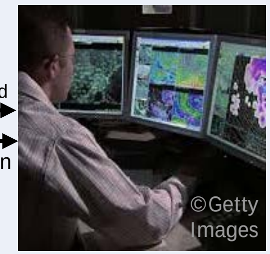
- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

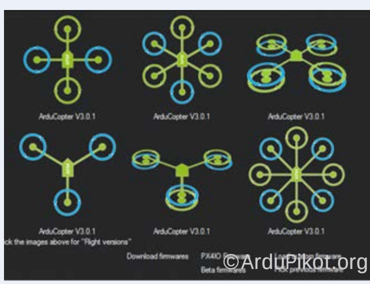
Challenge Problems

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	

Multimedia Data

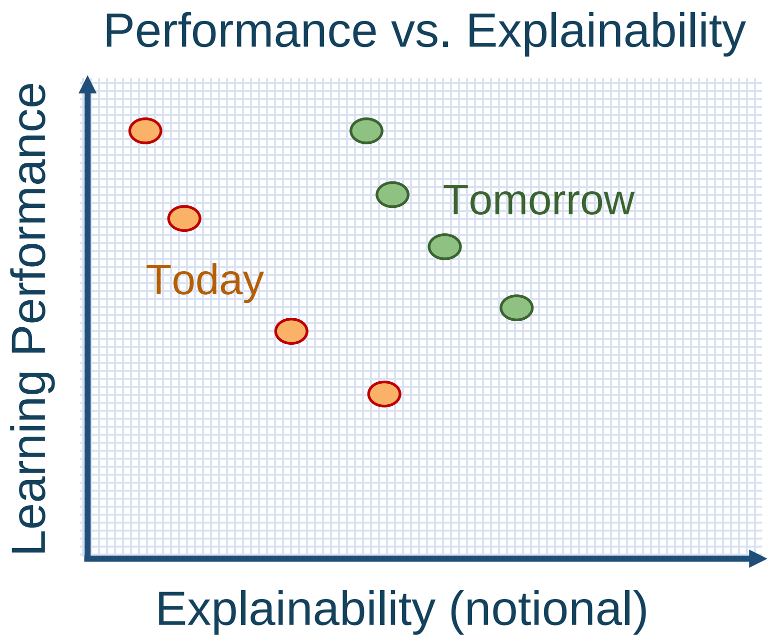




ArduPilot & SITL Simulation

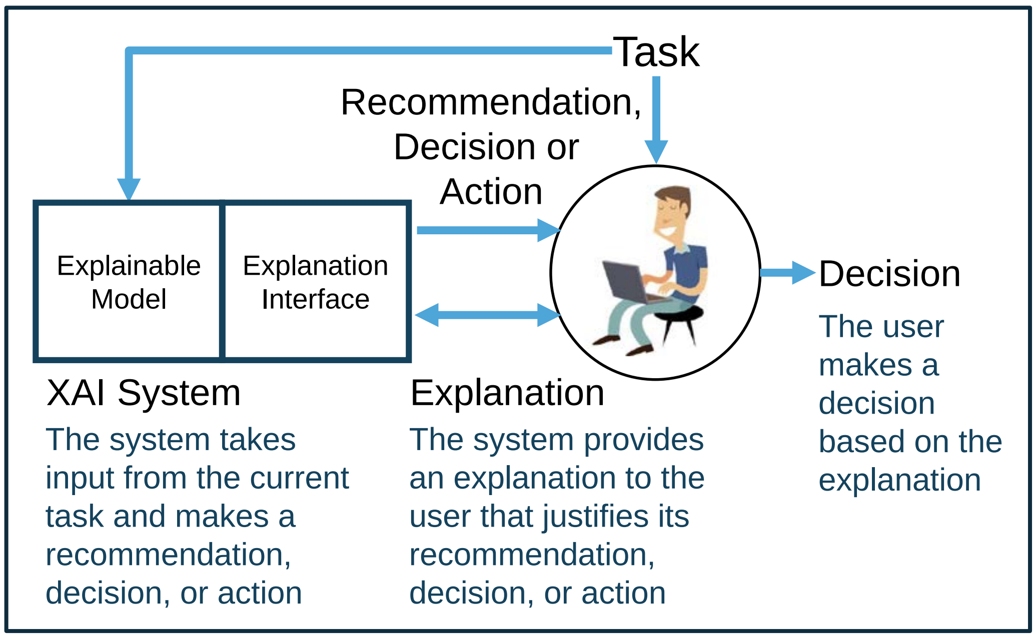
Goal: Performance and Explainability

- 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

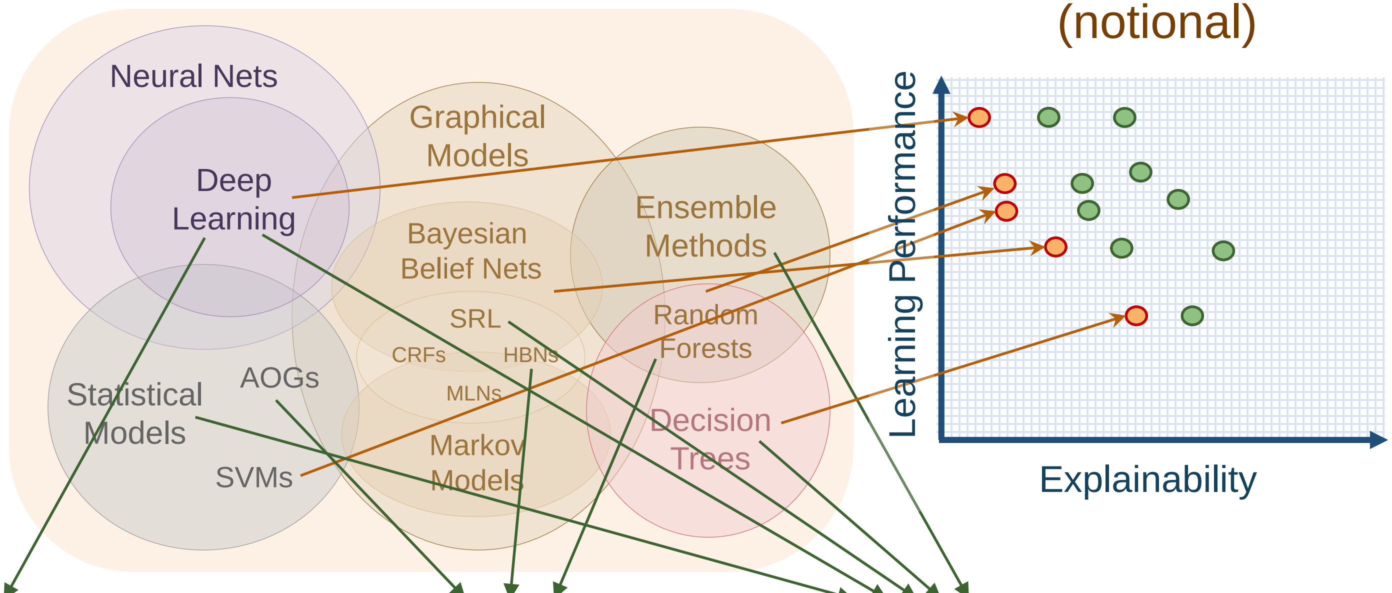
Performance vs. Explainability

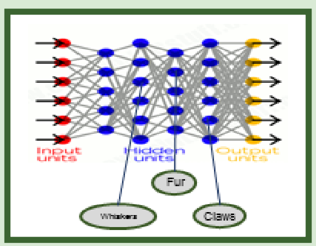
New Approach

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

Learning Techniques (today)

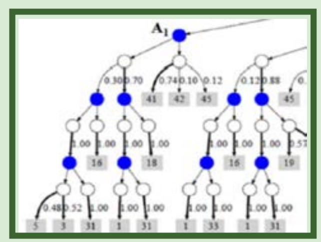
Explainability (notional)





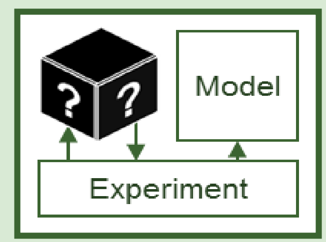
Deep Explanation

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

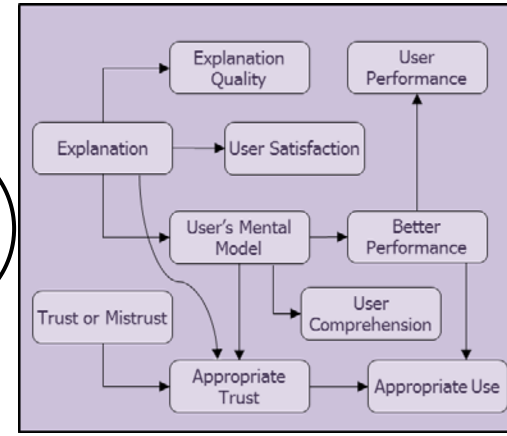
XAI Concept and Technical Approaches

Training Data

New Learning Process

Explainable Model

Explanation Interface

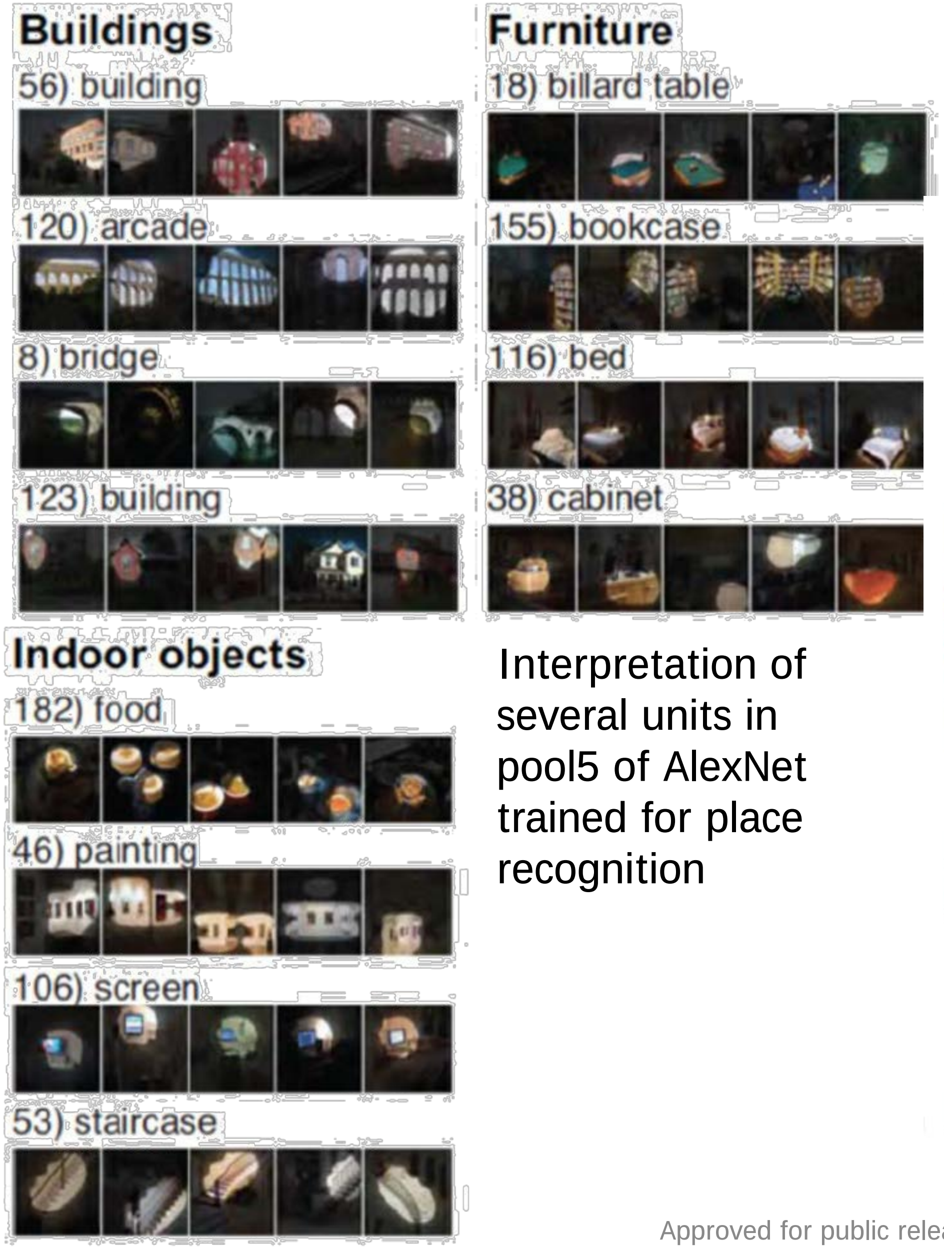


IHMC

Psychological Model of Explanation

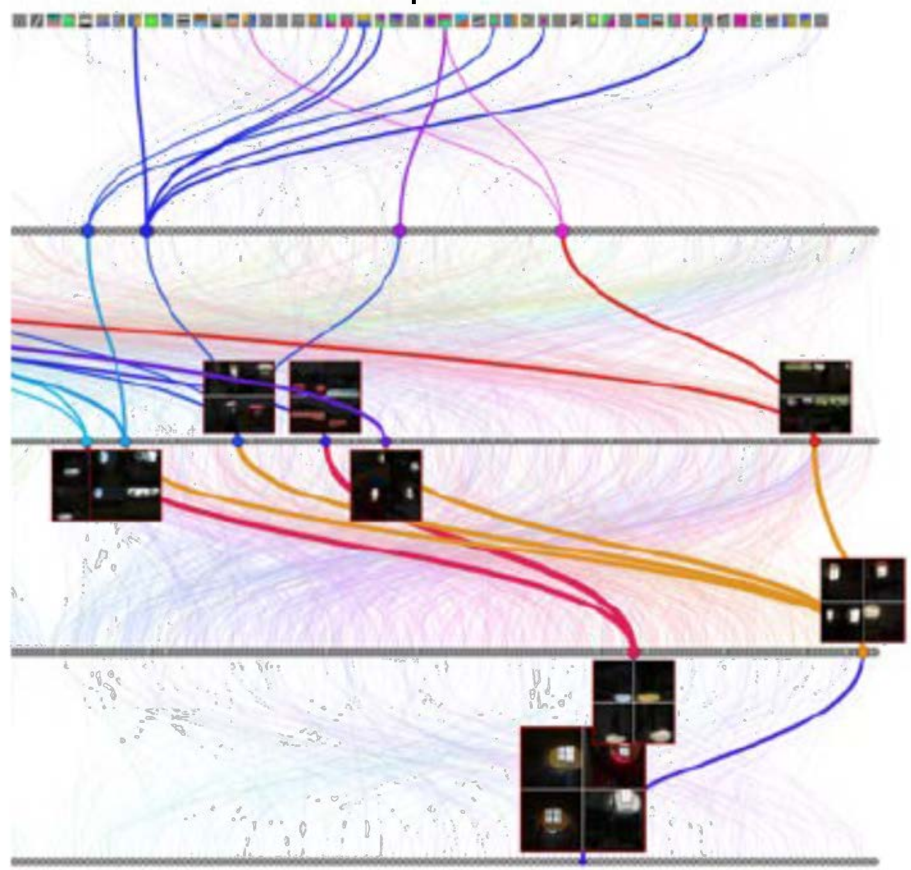
UC Berkeley	Deep Learning	Reflexive and Rational
Charles River Analytics	Causal Modeling	Narrative Generation
UCLA	Pattern Theory+	3-Level Explanation
Oregon State	Adaptive Programs	Acceptance Testing
PARC	Cognitive Modeling	Interactive Training
CMU	Explainable RL (XRL)	XRL Interaction
SRI International	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

Network Dissection Quantifying Interpretability of Deep Representations (MIT)

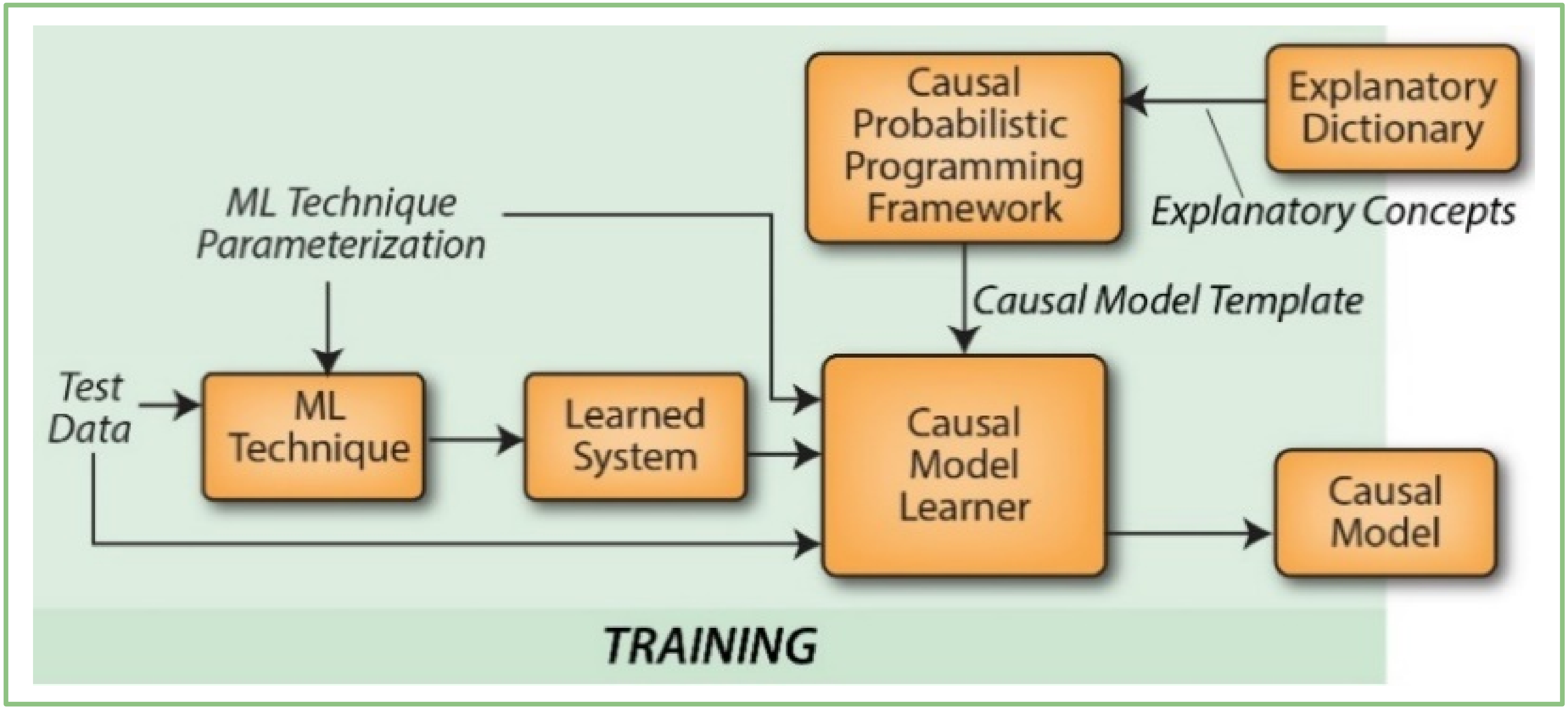


Interpretation of several units in pool5 of AlexNet trained for place recognition

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



Causal Model Induction (CRA)



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

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

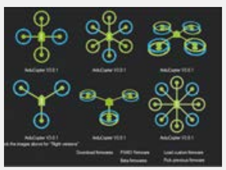
(or "common misconceptions" in pedagogy) that argue against less preferred answers based on analytics, cases, and data

XAI Program Structure

EXPLAINABLE ARTIFICIAL INTELLIGENCE

Challenge Problem Areas

Data Analytics
Multimedia Data

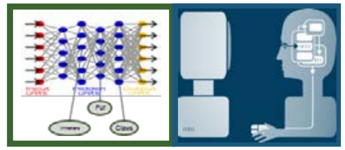


Autonomy
ArduPilot & SITL Simulation

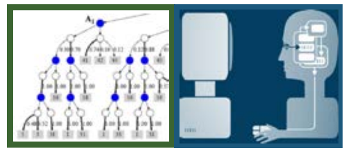
TA1: Explainable Learners

Teams that provide prototype systems with both components

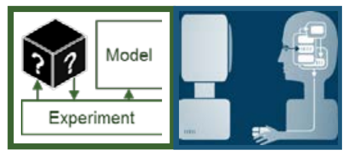
- Explainable Model
- Explanation Interface



Deep Learning Teams



Interpretable Model Teams

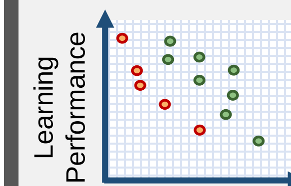


Model Induction Teams

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