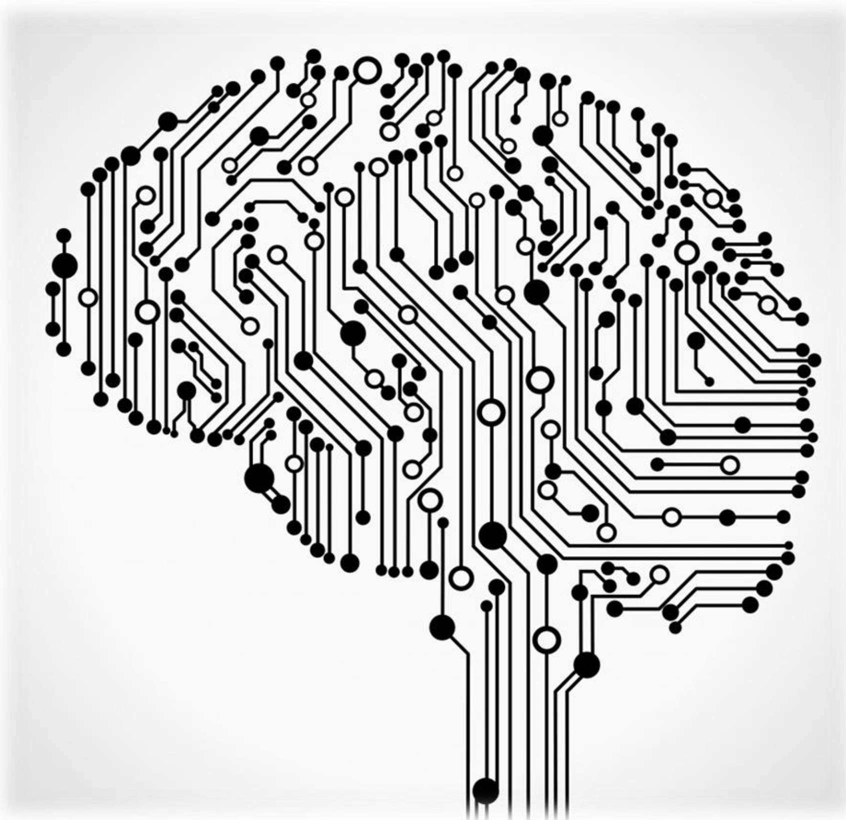
Explainable Artificial Intelligence Research at DARPA

David Gunning
DARPA/I2O

Components of Intelligence



perceive
rich, complex and subtle information

learn
within an environment

abstract
to create new meanings

reason
to plan and to decide

Three Waves of AI

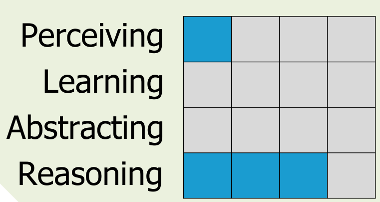
DESCRIBE

Symbolic Reasoning

engineers create sets of logic rules to represent knowledge in limited domains

reasoning over narrowly defined problems

no learning capability and poor handling of uncertainty



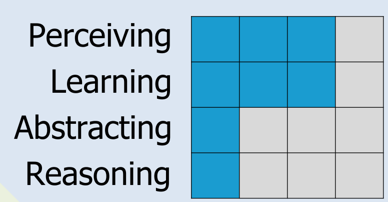
PREDICT

Statistical Learning

engineers create statistical models for specific problem domains and train them on big data

nuanced classification and prediction capabilities

no contextual capability and minimal reasoning ability



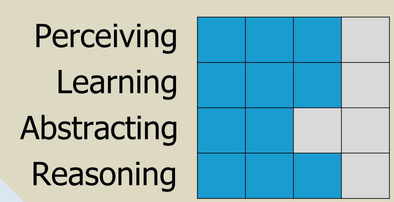
EXPLAIN

Contextual Adaptation

engineers create systems that construct explanatory models for classes of real world phenomena

natural communication among machines and people

systems learn and reason as they encounter new tasks and situations



Third wave technology: explanatory models

perceive

Adaptable input

abstract

Develop general principles from what is learned

explanatory model

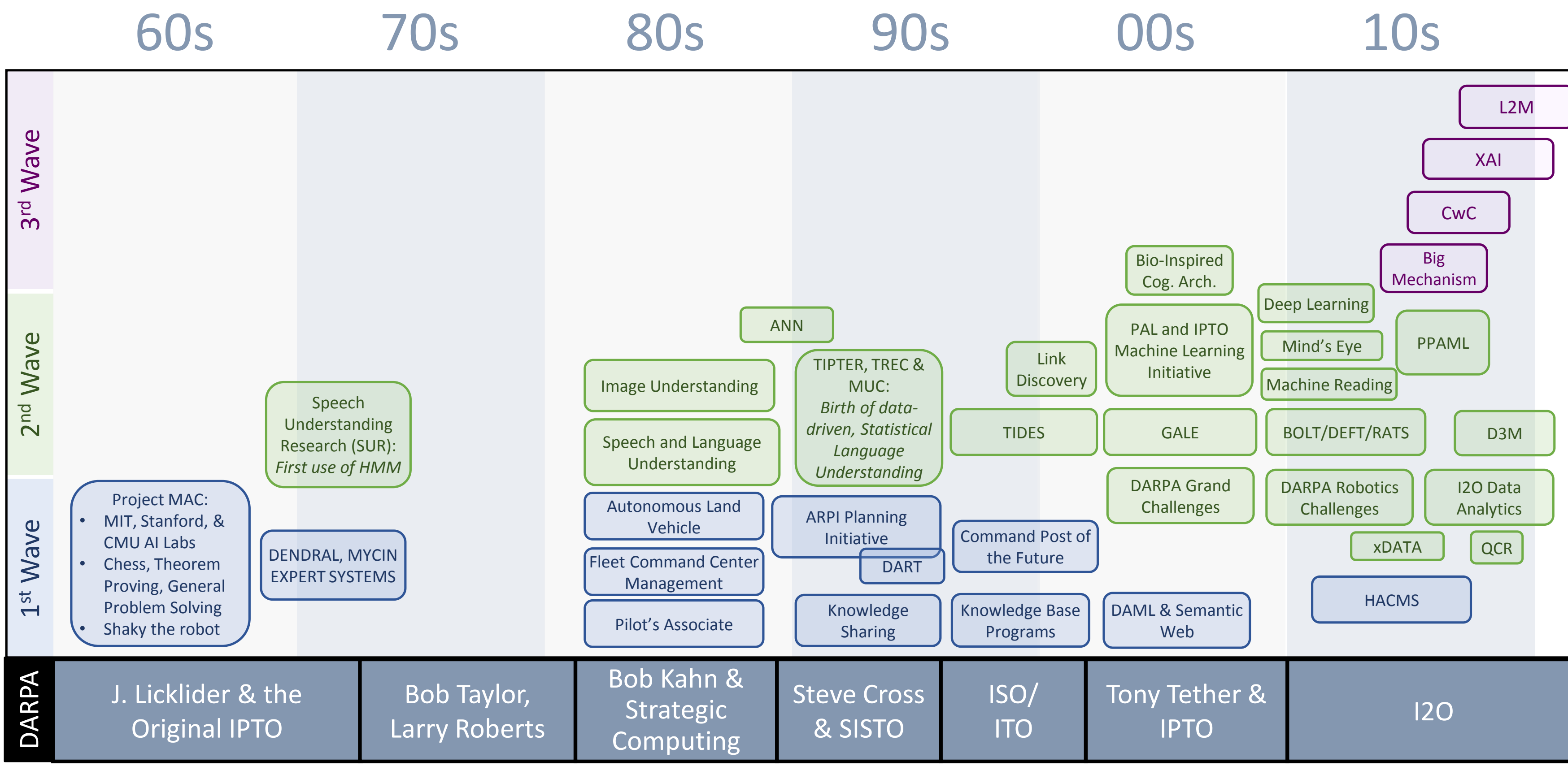
learn

Extract information from natural sources

reason

Explore models to predict and plan

DARPA Contributions to AI



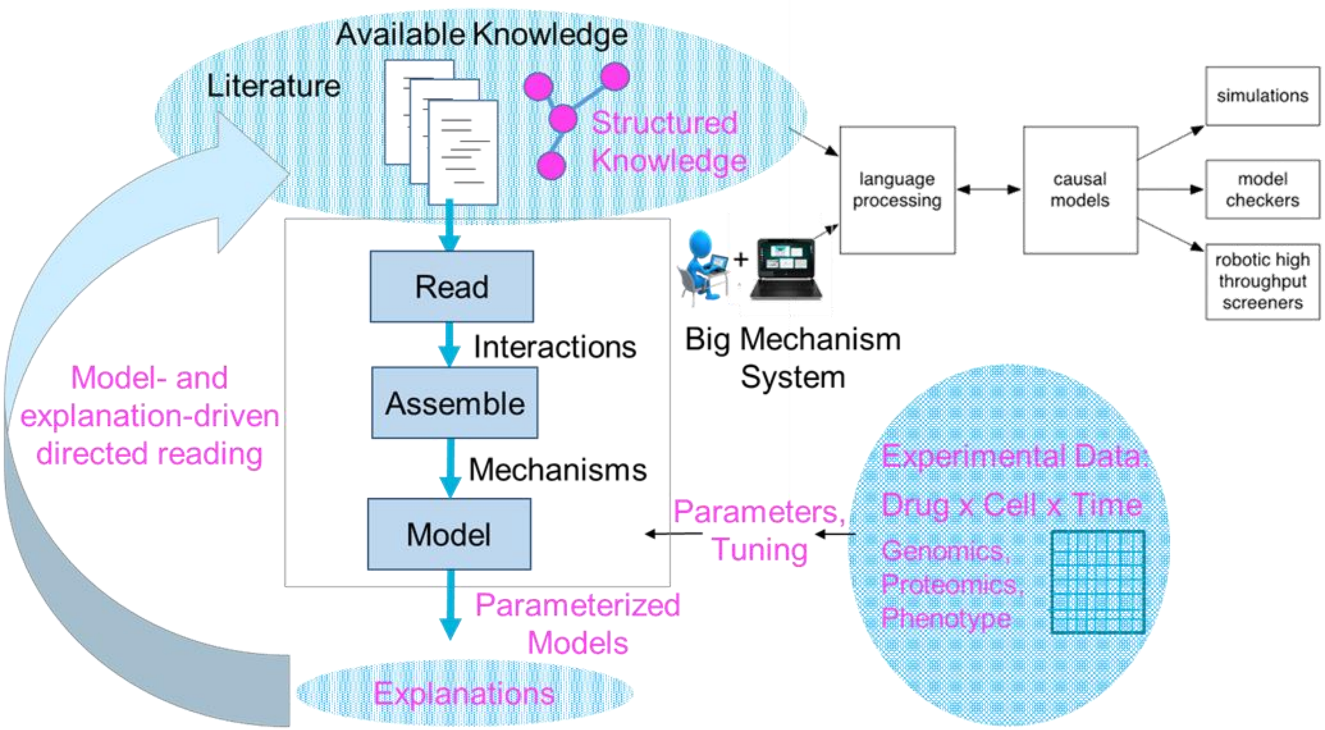
Big Mechanism
Steve Jameson

Automatically construct causal models of complicated systems to predict and explain the effects of system perturbations (cell biology)

READ publications and published data to extract fragments

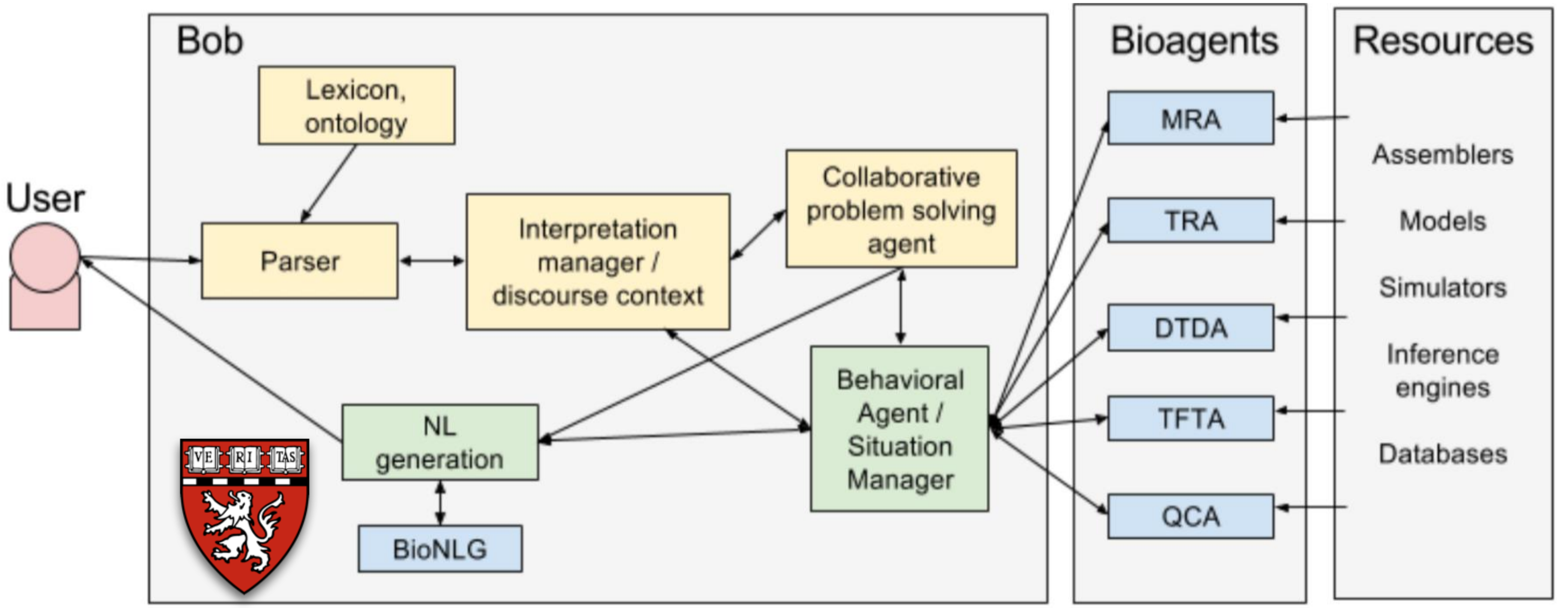
ASSEMBLE fragments into more comprehensive causal models

EXPLAIN and reason about causal phenomena in these models



Build causal, mechanistic, quantitative models to produce explanatory models of unprecedented completeness and consistency

Collaborative Problem Solving Agent "Bob"
Harvard Medical School

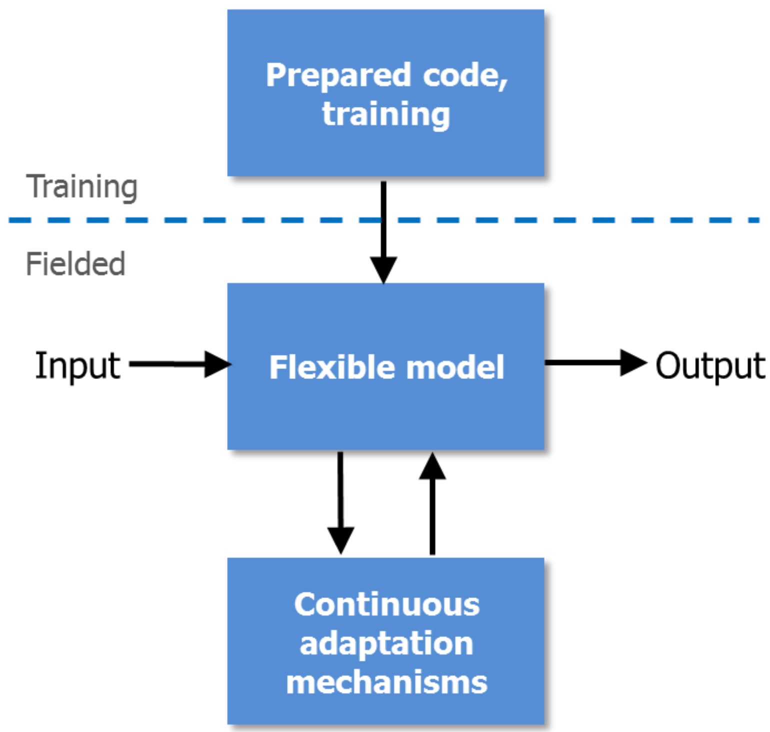


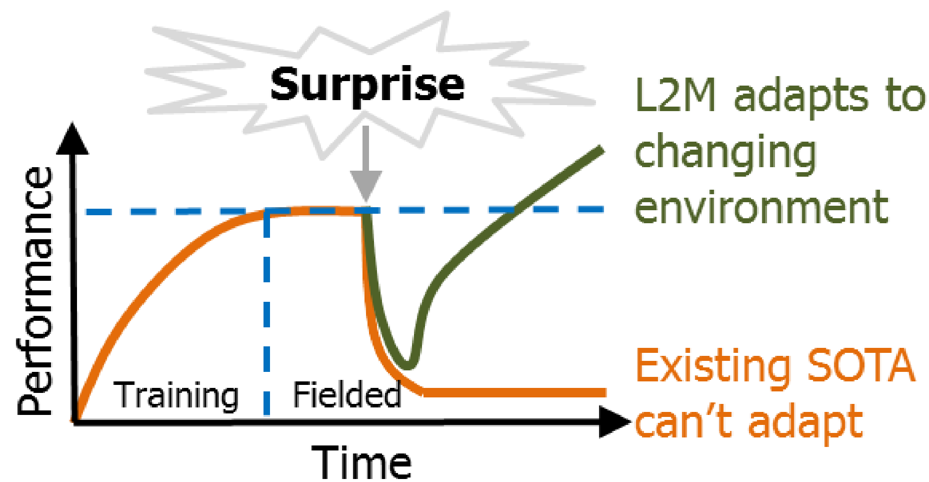
- Implementation of a generic **language understanding** system
- Working implementation of generic **collaborative problem solving and planning**
- **Biological problem solving agents (Bioagents)**, which are generic for their specific sub-tasks
- Integration into a working **communication-for-biocuration system**

Lifelong Learning Machines (L2M)

Hava Siegelmann

Develop fundamentally new machine learning mechanisms that will enable systems to improve their performance over their lifetimes



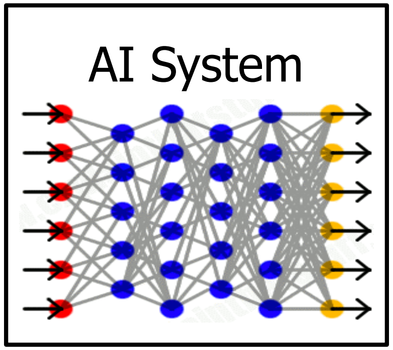


Focus on functional system development, taking inspiration from known biological properties

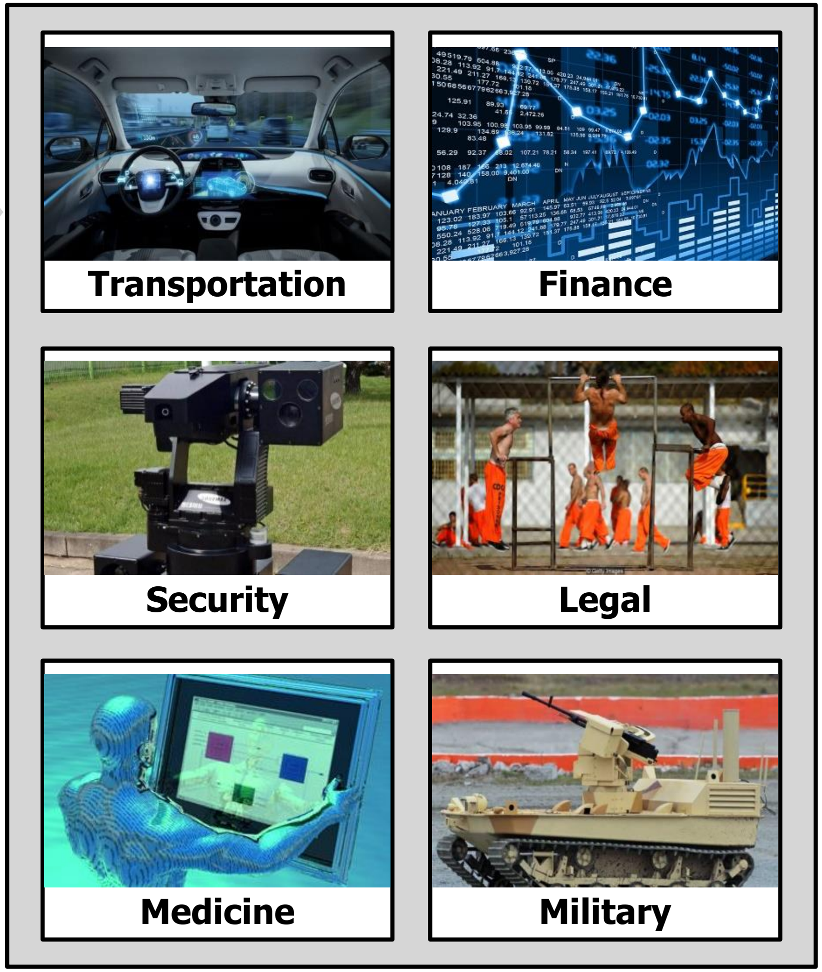
Dynamically evolve networks online
Use scalable approaches

Identify and explore biological mechanisms that underlie real-time adaptation for translation into novel algorithms

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

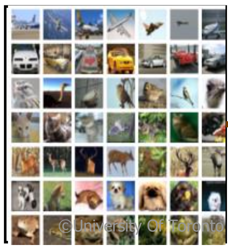


- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- 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.

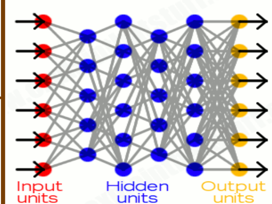
What Are We Trying To Do?

Today



Training Data

Learning Process



Learned Function

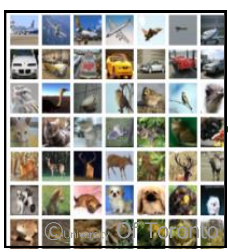
This is a cat
(p = .93)

Output

User with a Task

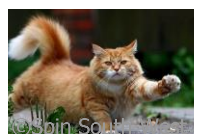
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

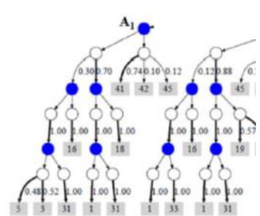
Tomorrow



Training Data

New Learning Process

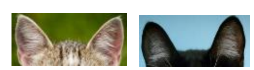




Explainable Model

This is a cat:

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



Explanation Interface

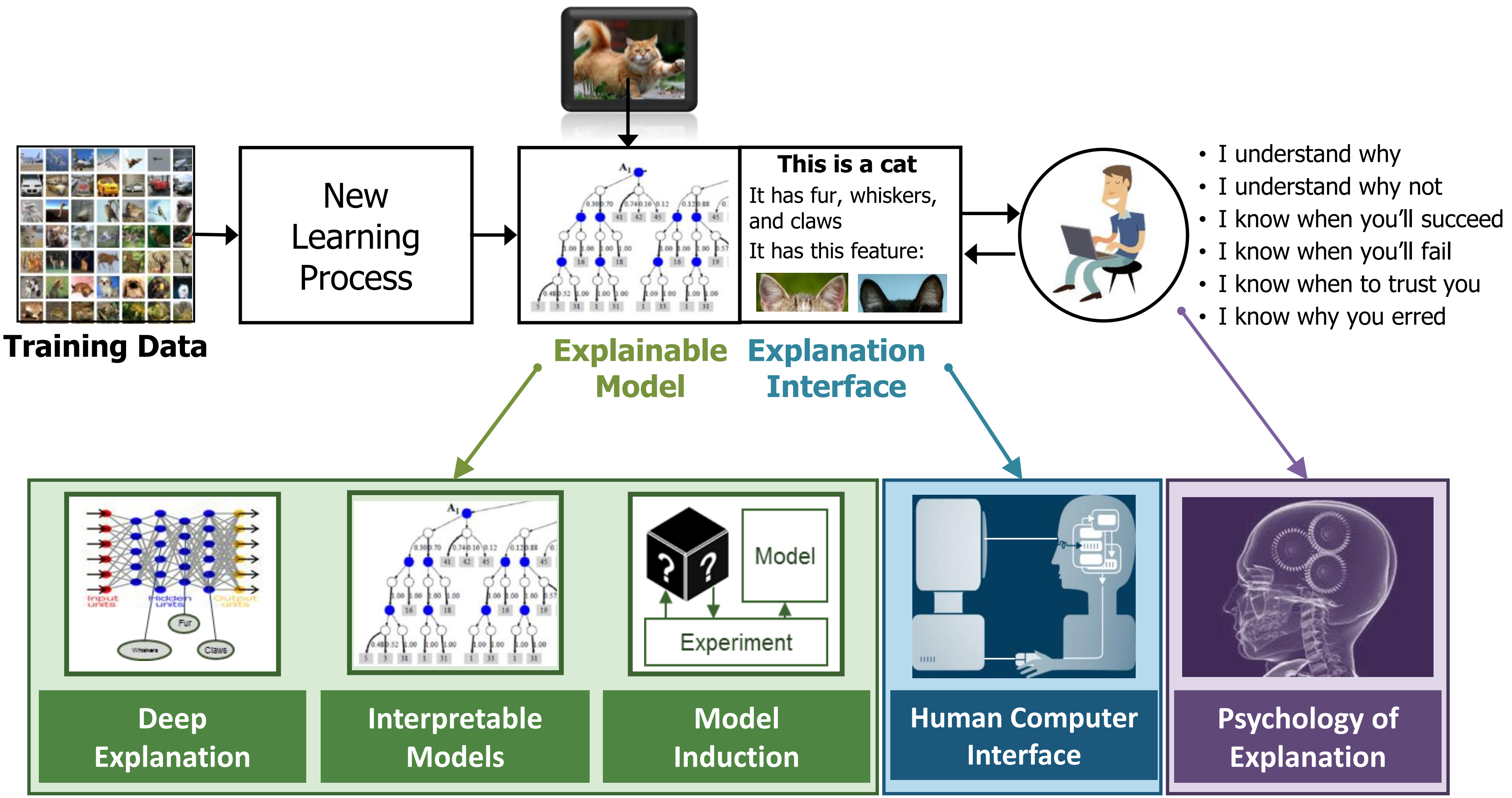
User with a Task

- 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

Explainable Artificial Intelligence (XAI)

David Gunning

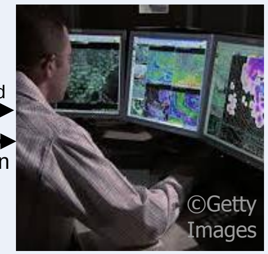
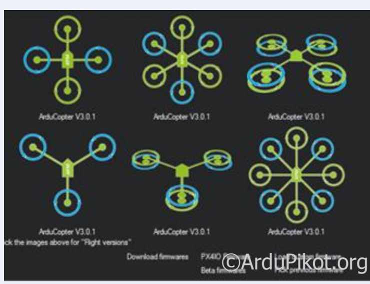
Create a suite of machine learning techniques to produce more explainable models and enable human users to understand, trust, and effectively manage the emerging generation of artificially intelligent partners



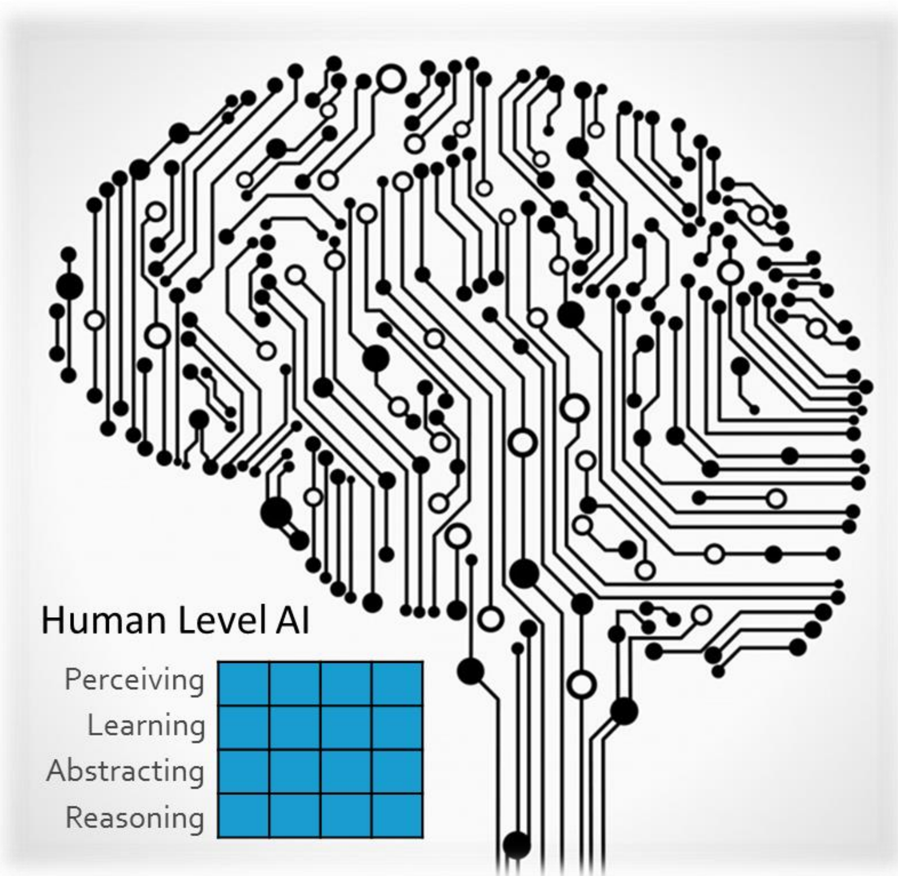
XAI Developers (TA1)

CP	Performer	Explainable Model	Explanation Interface
Both	UC Berkeley	Deep Learning	Reflexive & 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 & Tell Explanation
	Raytheon BBN	Deep Learning	Argumentation & Pedagogy
	UT Dallas	Probabilistic Logic	Decision Diagrams
	Texas A&M	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

Challenge Problems

	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	

Remaining Challenges for AI



- Learning
  - Unsupervised learning
  - One-shot learning
  - Lifelong learning
  - Learning from instruction
- Understanding
  - Explanation
  - Representation and abstraction
- Human-like cognition
  - Planning and action
  - Meta-reasoning
  - Common Sense

www.darpa.mil

XAI In the News

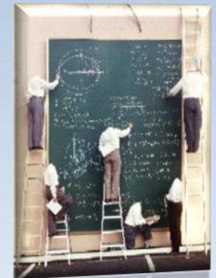
MIT Technology Review

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

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

FT FINANCIAL TIMES

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

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

Military EMBEDDED SYSTEMS

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

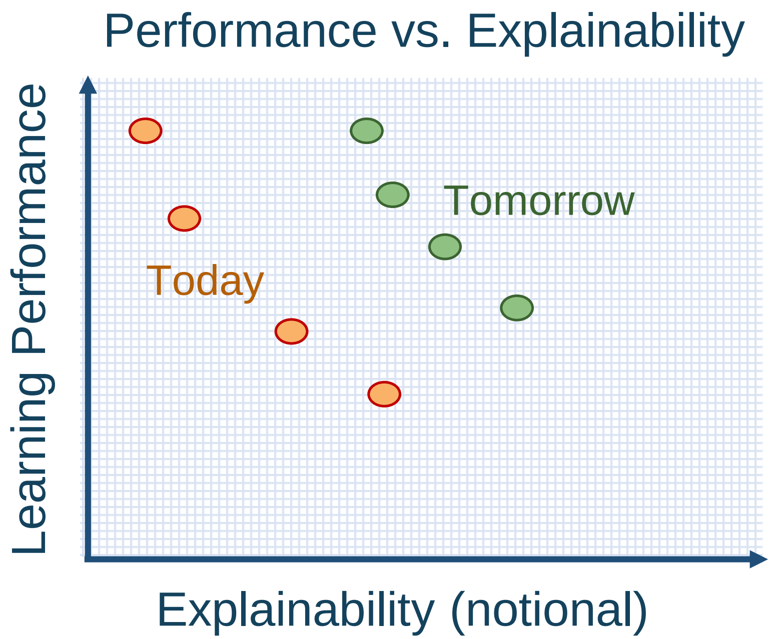
How AI detectives are cracking open the black box of deep learning
Paul Voosen
July 6, 2017

Science AAAS

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

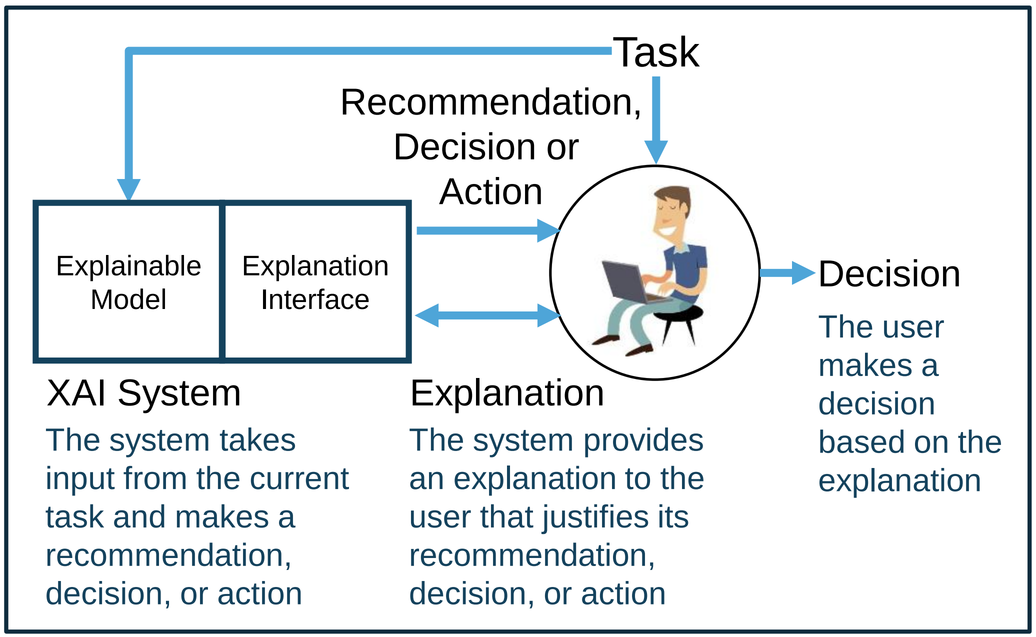
Goal: Performance and Explainability

- 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

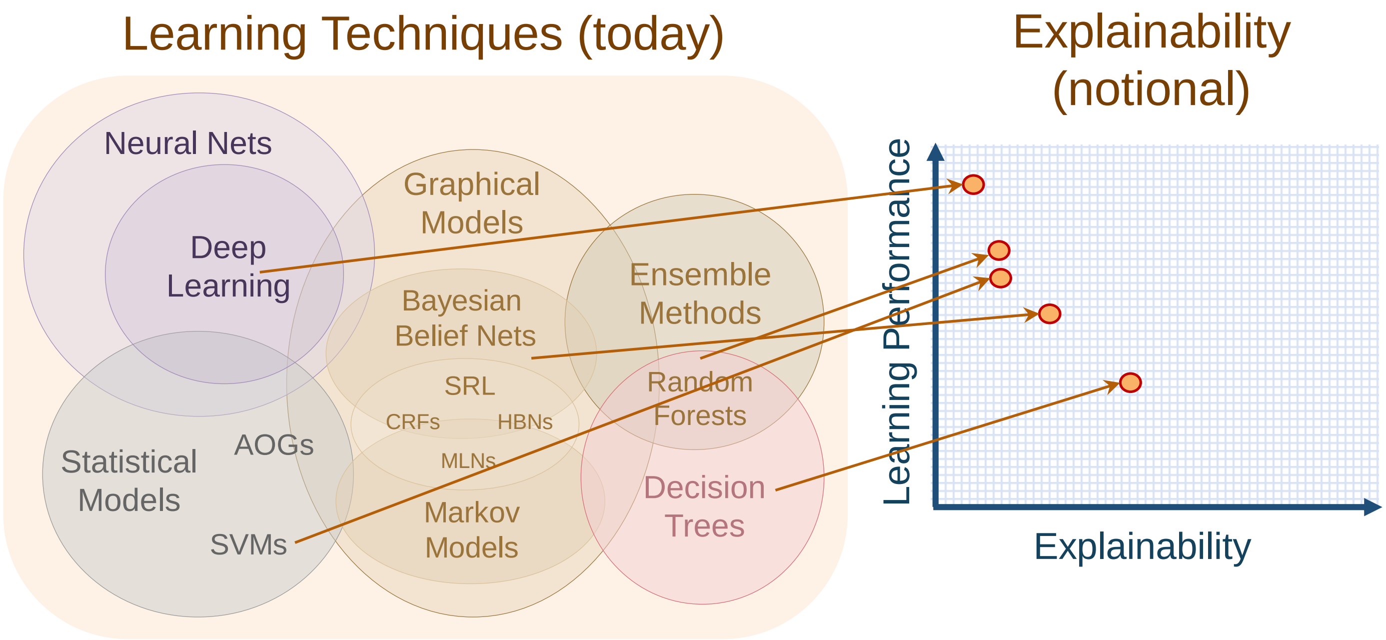
Trust Assessment

- Appropriate future use and trust

Correctability (Extra Credit)

- Identifying errors
- Correcting errors
- Continuous training

Performance vs. Explainability

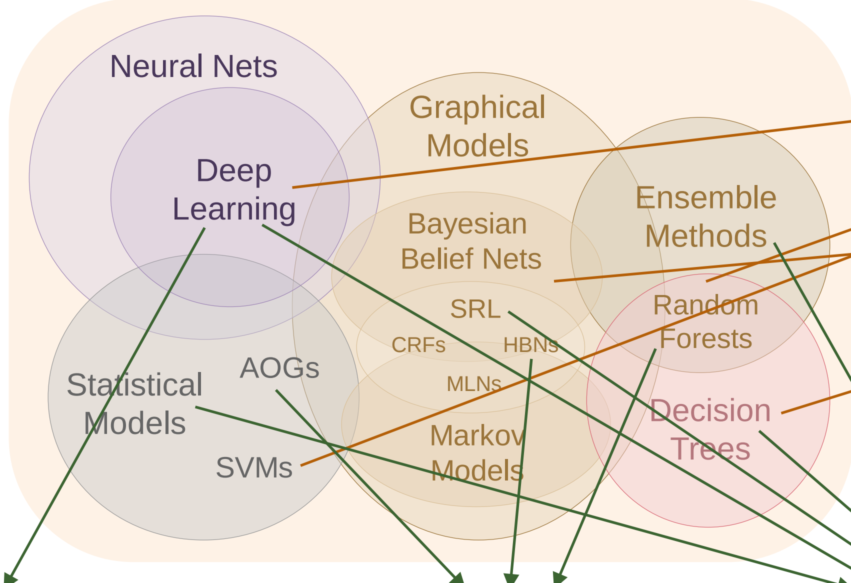


Performance vs. Explainability

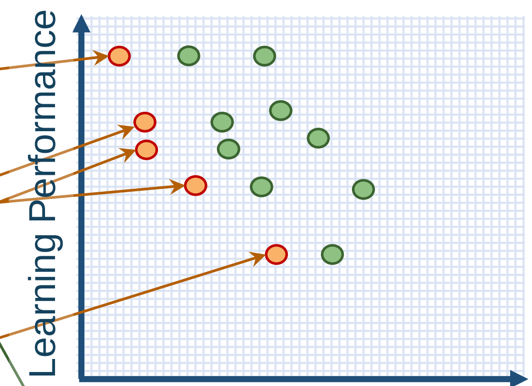
New Approach

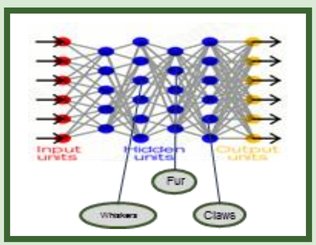
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



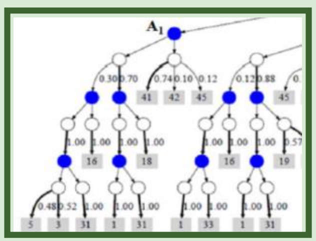
Explainability (notional)





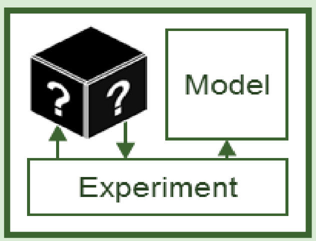
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models

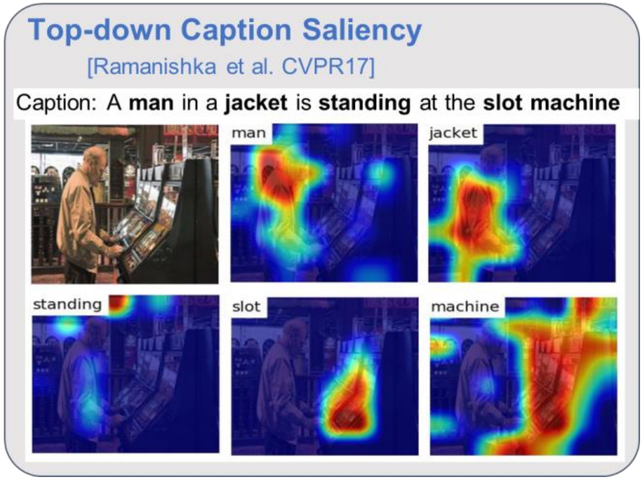


Model Induction

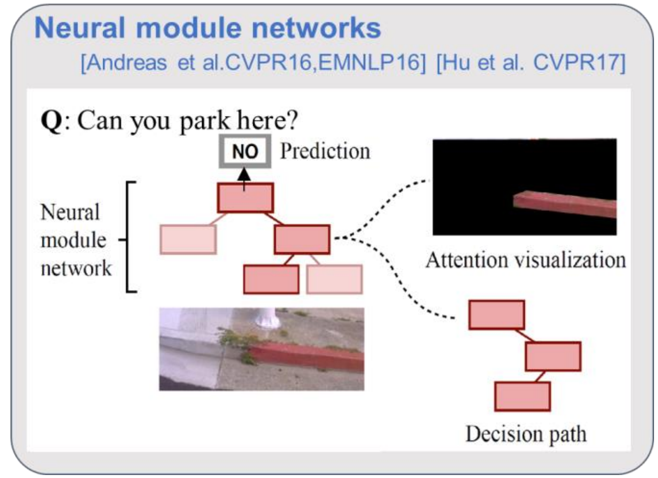
Techniques to infer an explainable model from any model as a black box

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

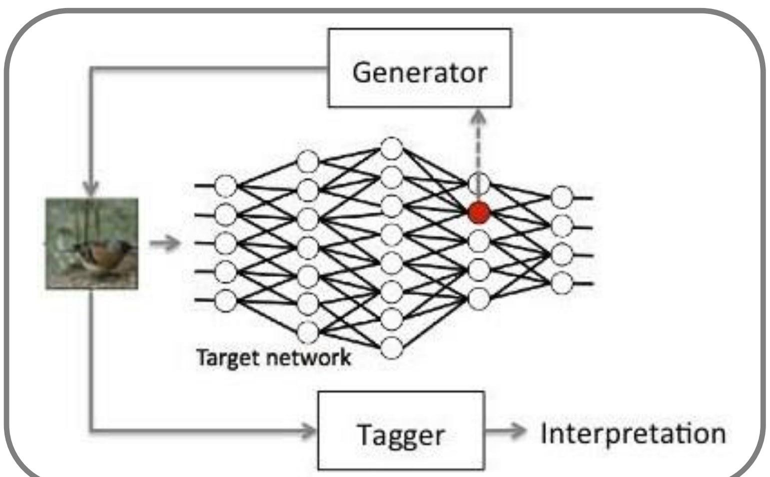
Attention Mechanisms



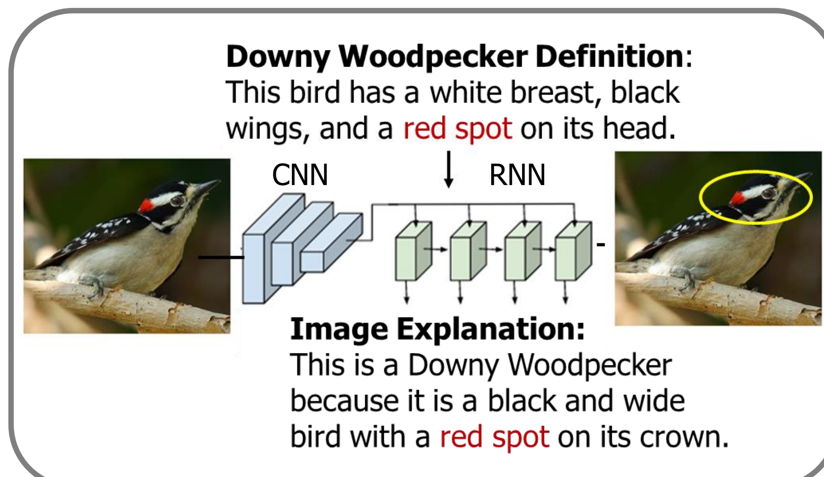
Modular Networks



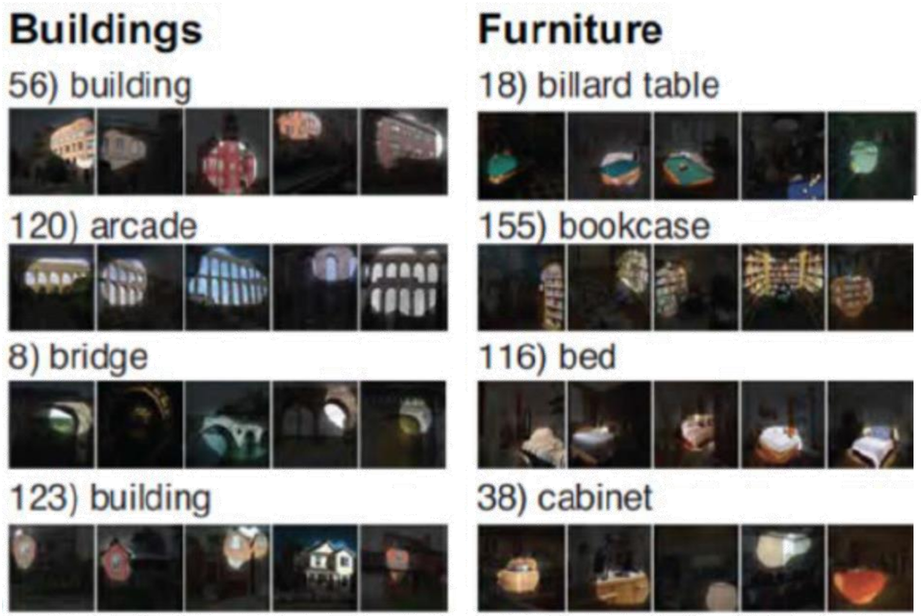
Feature Identification

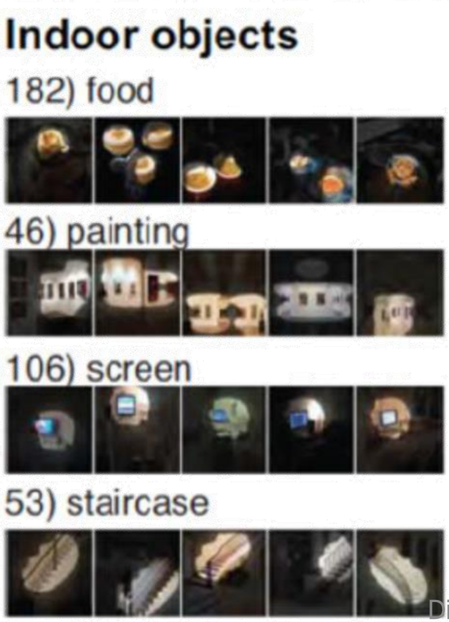


Learn to Explain



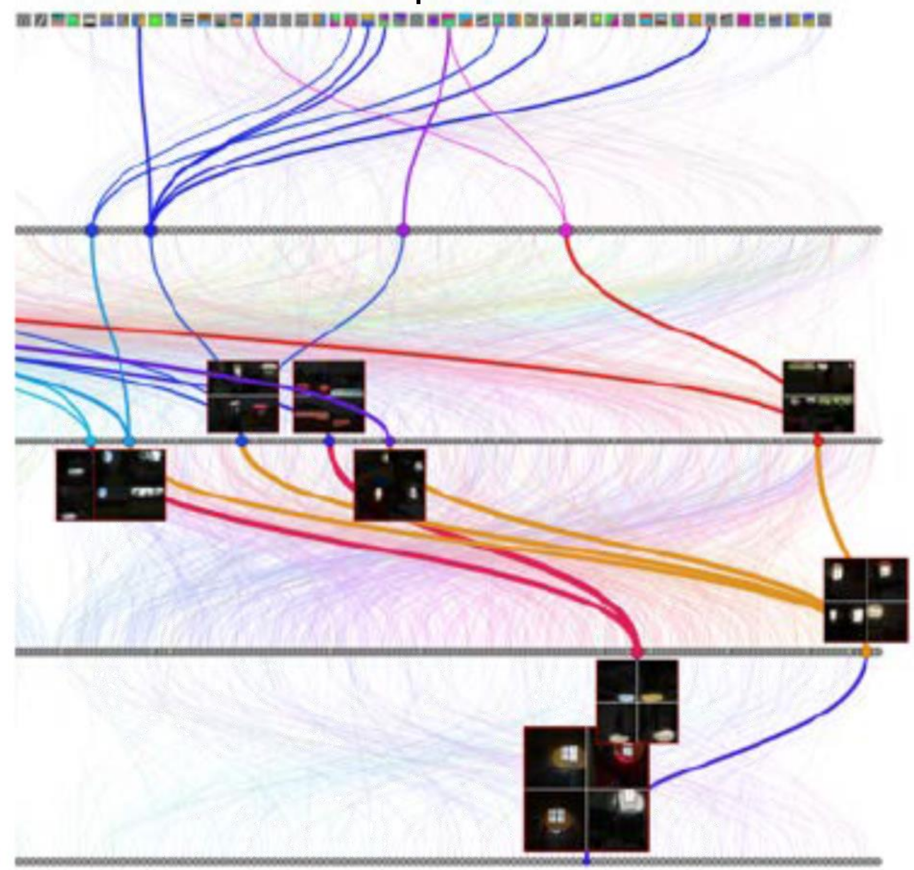
Network Dissection Quantifying Interpretability of Deep Representations (MIT)



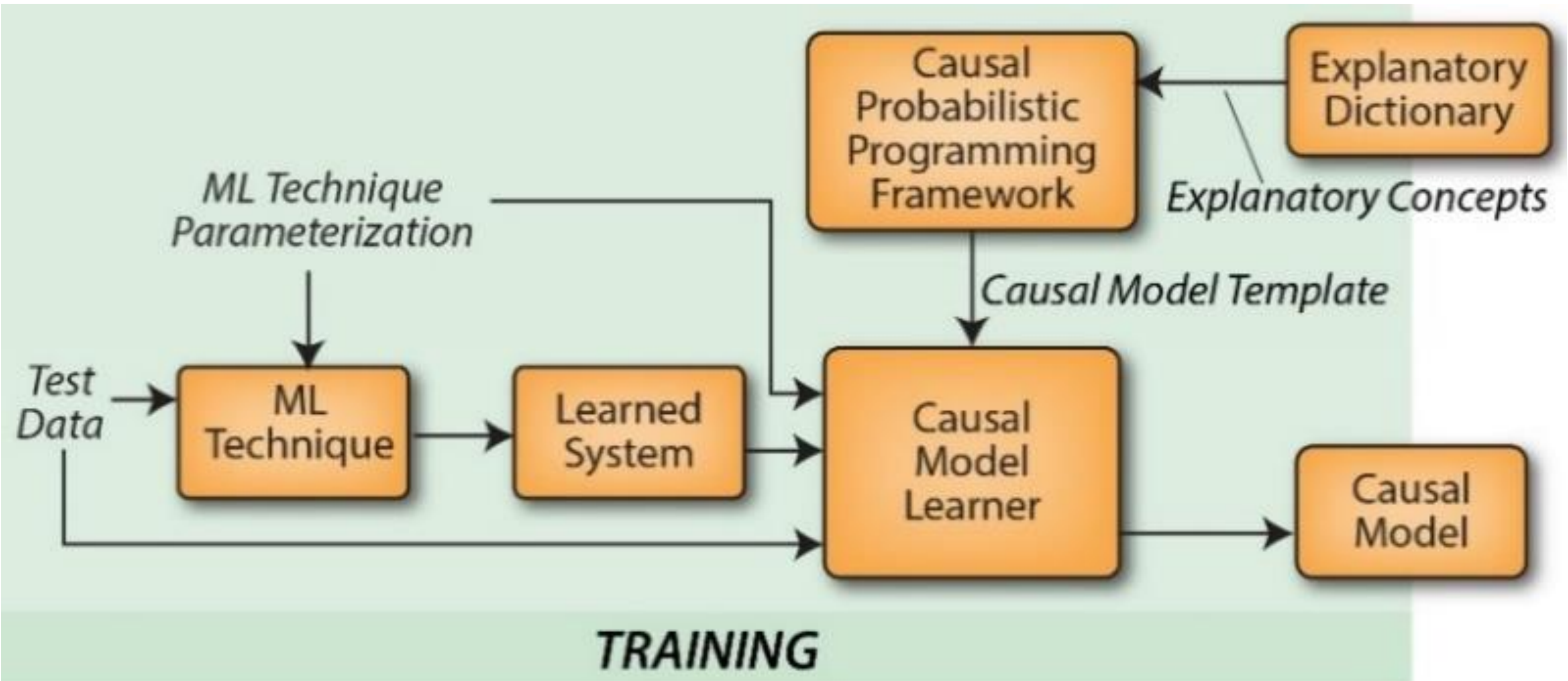


Interpretation of several units in pool5 of AlexNet trained for place recognition

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



Causal Model Induction (CRA)

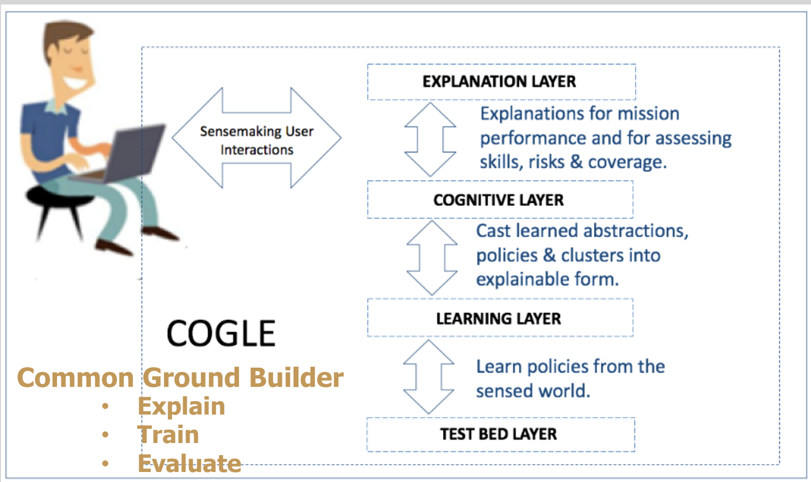


<u>Causal Model Induction</u>: Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model

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