Explainable Artificial Intelligence (XAI)

XAI BAA Outline

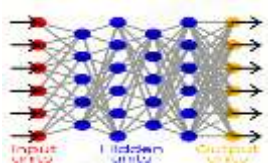
A. Introduction
B. Program Scope
    1. Explainable Models
    2. Explanation Interface
    3. Psychology of Explanation
    4. Emphasis and Scope of XAI Research
C. Challenge Problems and Evaluation
    1. Overview
    2. Data Analysis
    3. Autonomy
    4. Evaluation
D. Technical Areas
    1. Explainable Learners
    2. Psychological Model of Explanation
E. Schedule and Milestones
F. Deliverables

Questions

- What do you mean by explanation or explaining?

A. Introduction - The Need for Explainable AI

AI System



- 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

Watson

AlphaGo

Sensemaking

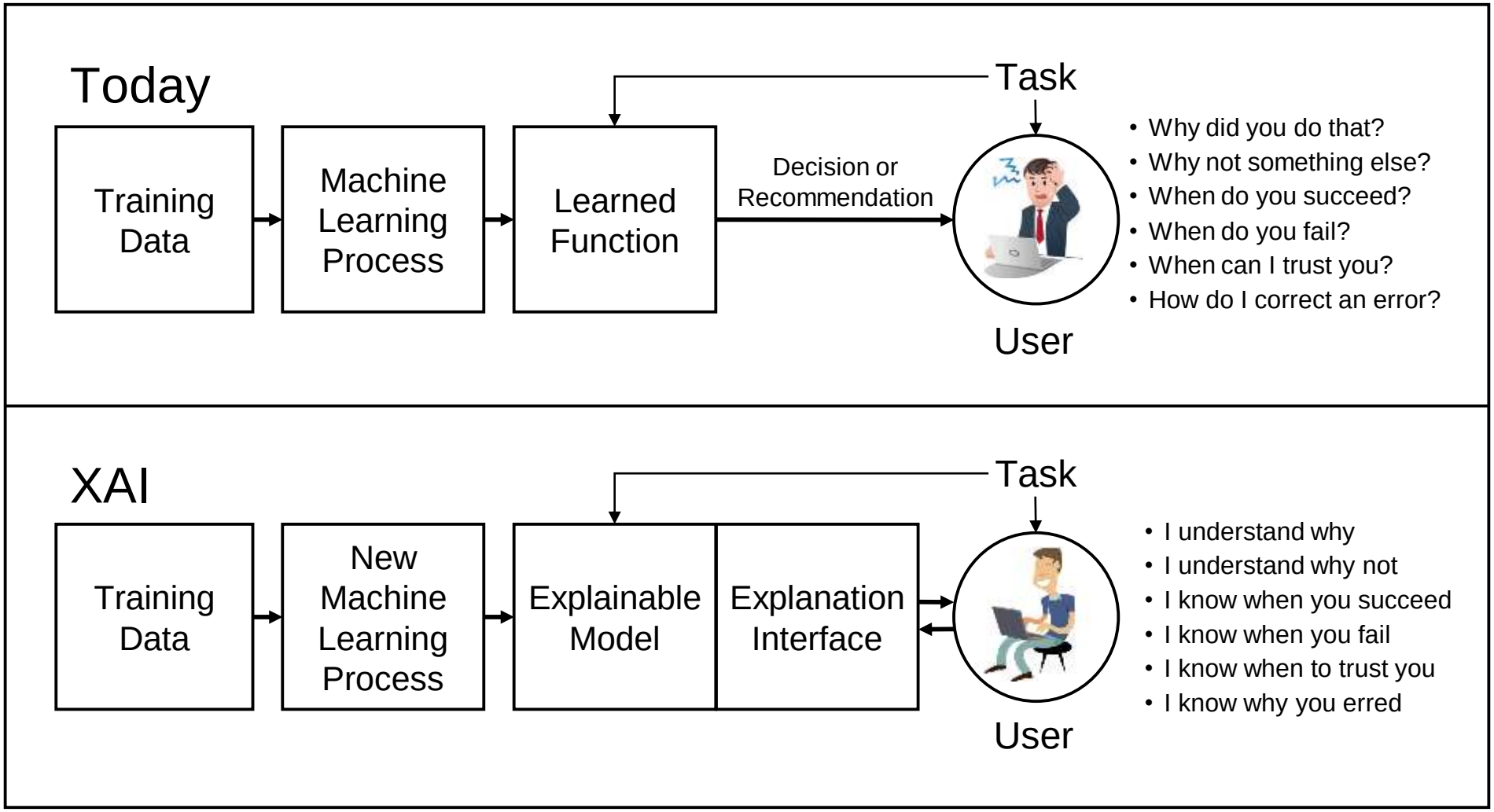
Operations

User

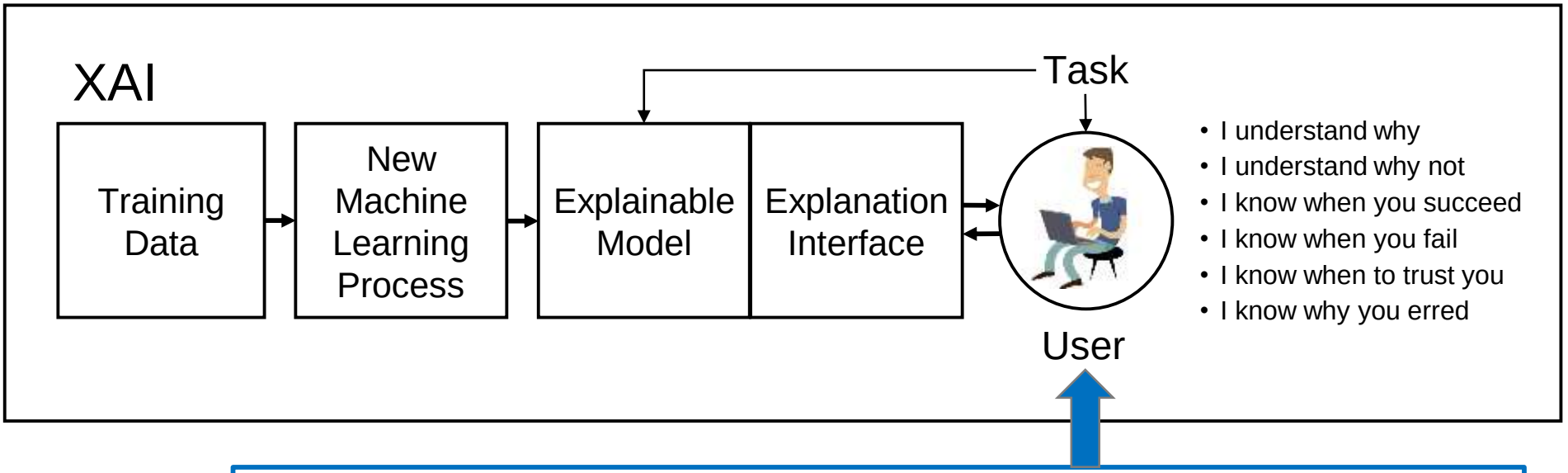
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- 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.

Concept

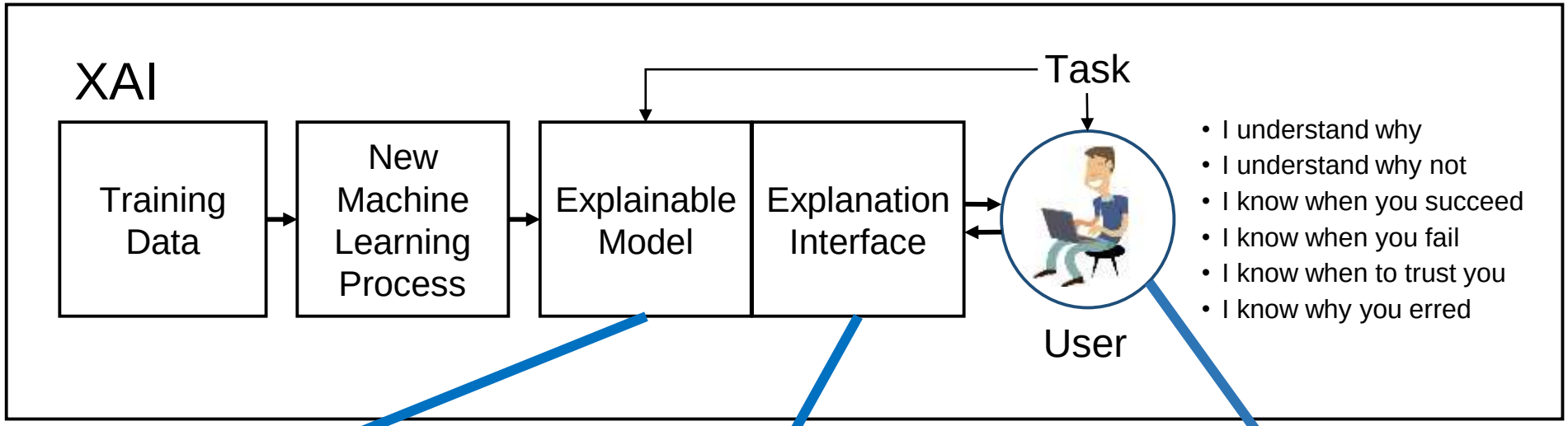

Today
Training Data
Machine Learning Process
Learned Function
Decision or Recommendation
Task
User
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
XAI
Training Data
New Machine Learning Process
Explainable Model
Explanation Interface
Task
User
• I understand why
• I understand why not
• I know when you succeed
• I know when you fail
• I know when to trust you
• I know why you erred

XAI Concept



- The target of XAI is an end user who:
  - depends on decisions, recommendations, or actions of the system
  - needs to understand the rationale for the system's decisions to understand, appropriately trust, and effectively manage the system
- The XAI concept is to:
  - provide an explanation of individual decisions
  - enable understanding of overall strengths & weaknesses
  - convey an understanding of how the system will behave in the future
  - convey how to correct the system's mistakes (perhaps)

B. Program Scope – XAI Development Challenges



Explainable Models

- develop a range of new or modified machine learning techniques to produce more explainable models

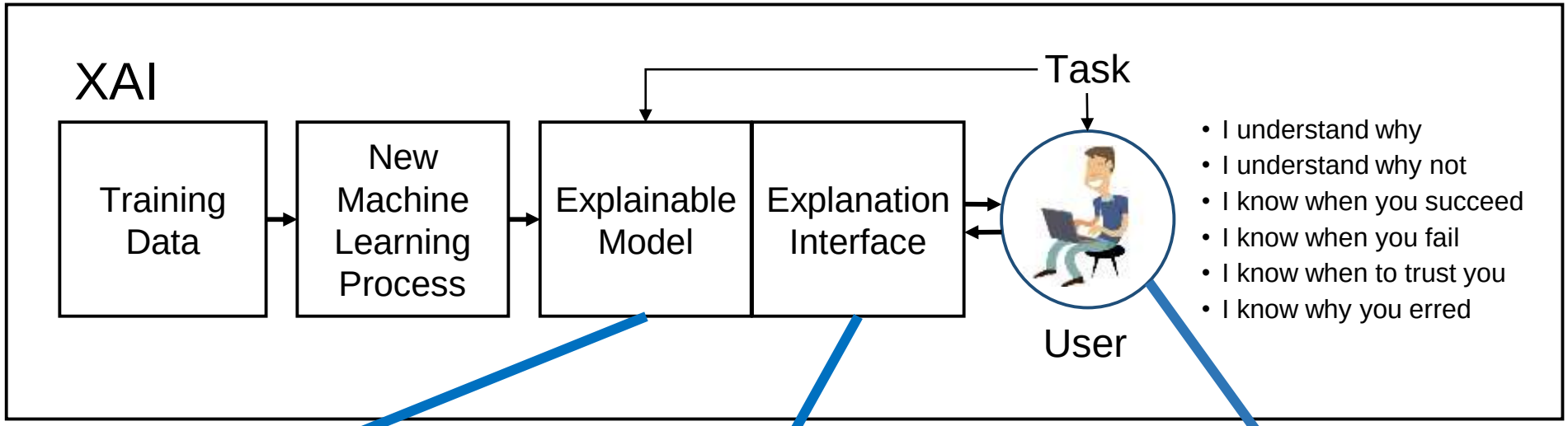
Explanation Interface

- integrate state-of-the-art HCI with new principles, strategies, and techniques to generate effective explanations

Psychology of Explanation

- summarize, extend, and apply current psychological theories of explanation to develop a computational theory

XAI Development Challenges



Explainable Models

- develop a range of new or modified machine learning techniques to produce more explainable models

Explanation Interface

- integrate state-of-the-art HCI with new principles, strategies, and techniques to generate effective explanations

Psychology of Explanation

- summarize, extend, and apply current psychological theories of explanation to develop a computational theory

TA 1: Explainable Learners

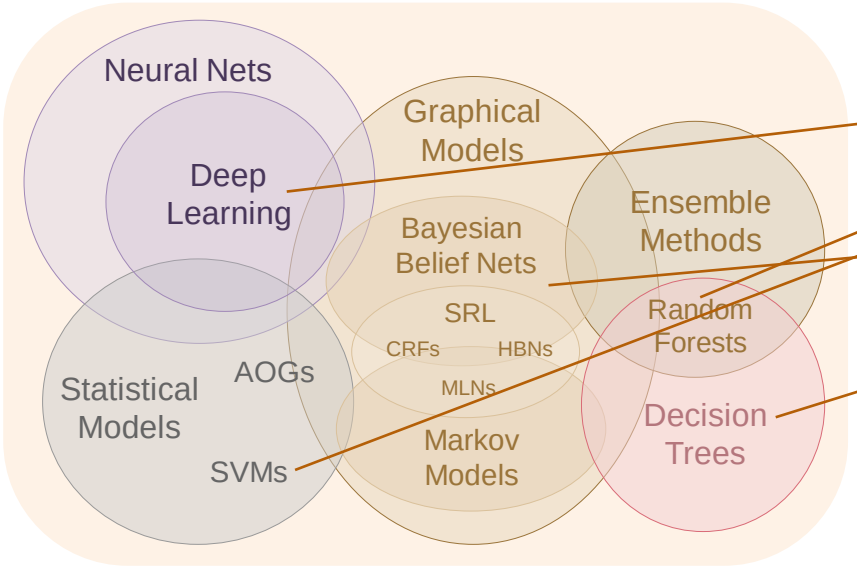
TA 2: Psychological Models

B.1 Explainable Models

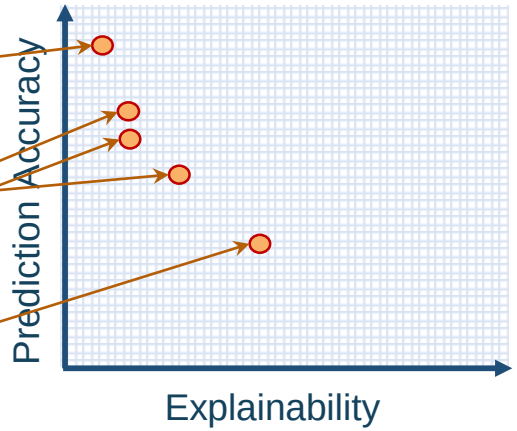
New Approach

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)

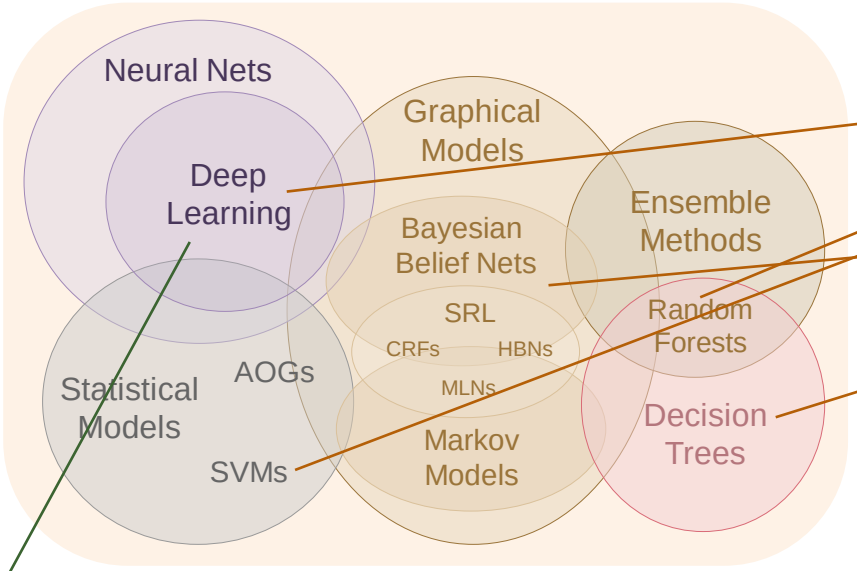


B.1 Explainable Models

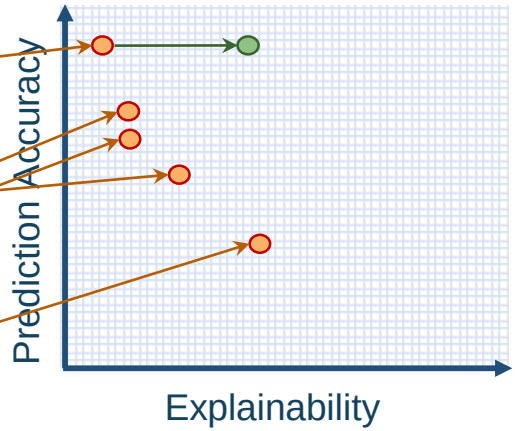
New Approach

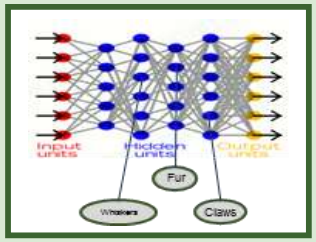
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)





Deep Explanation

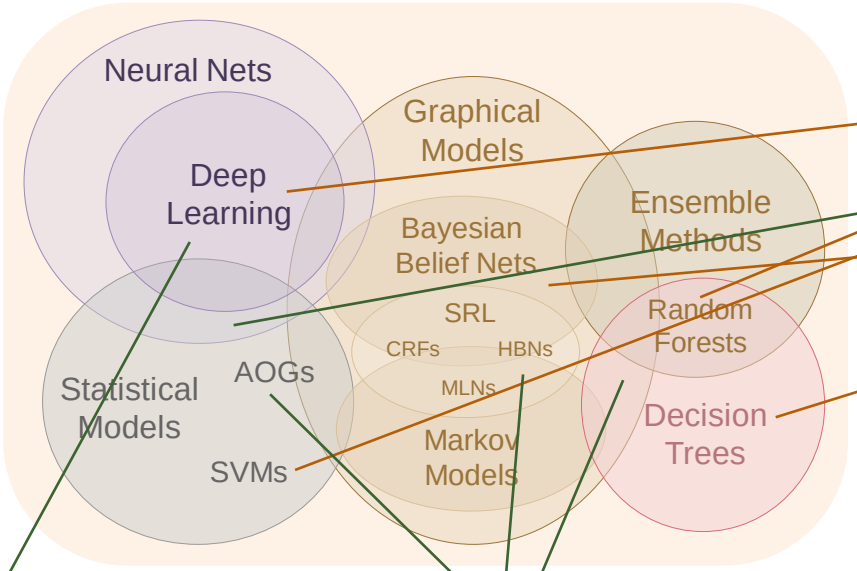
Modified deep learning techniques to learn explainable features

B.1 Explainable Models

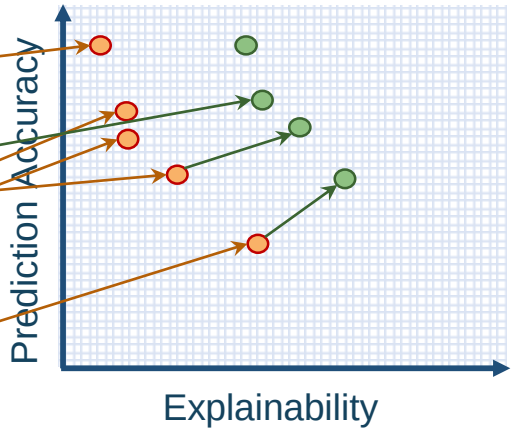
New Approach

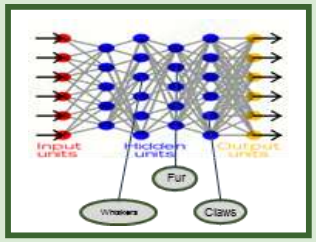
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



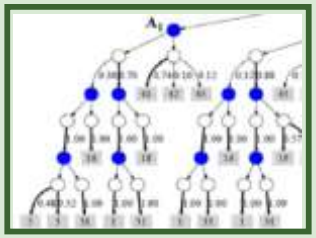
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

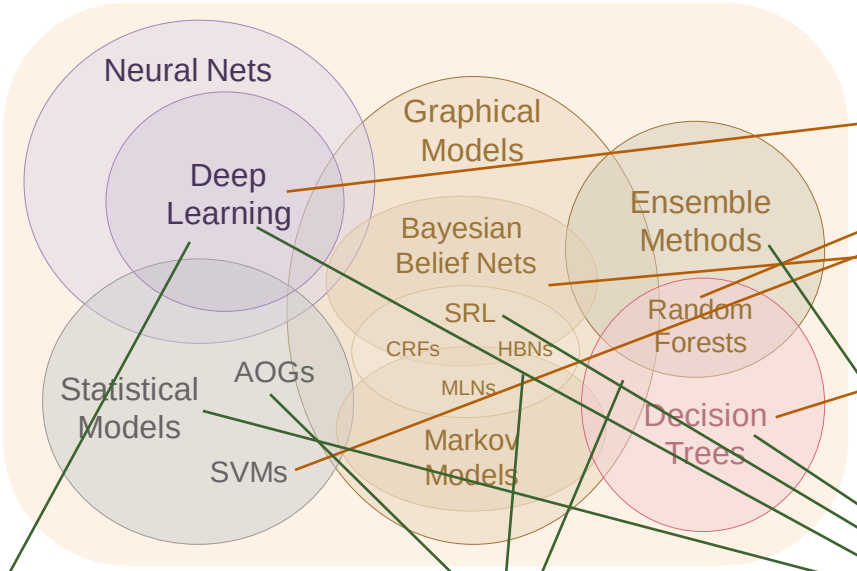
Techniques to learn more structured, interpretable, causal models

B.1 Explainable Models

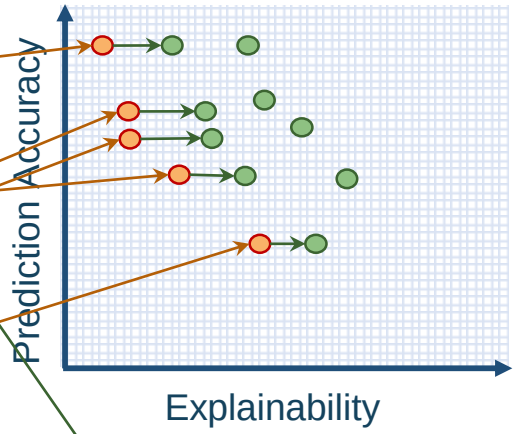
New Approach

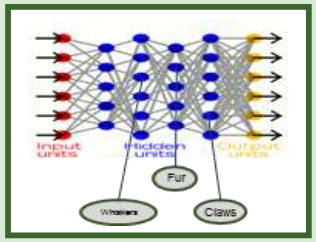
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



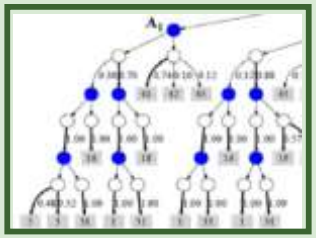
Explainability (notional)





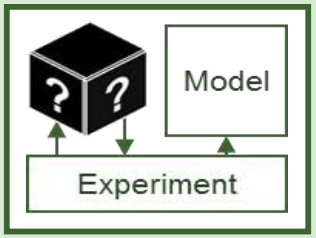
Deep Explanation

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

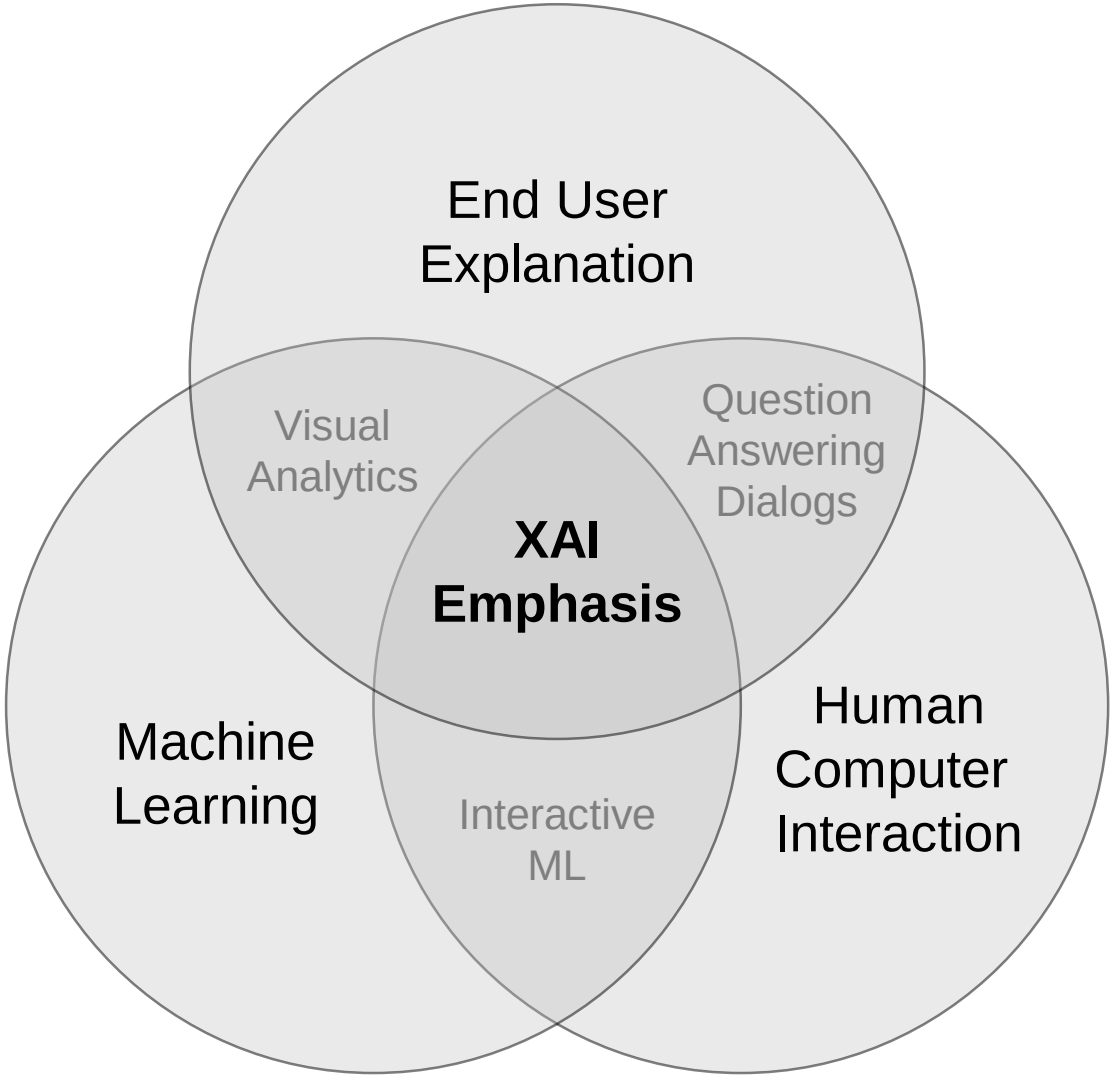
B.2 Explanation Interface

- State of the Art Human Computer Interaction (HCI)
  - UX design
  - Visualization
  - Language understanding & generation
- New Principles and Strategies
  - Explanation principles
  - Explanation strategies
  - Explanation dialogs
- HCI in the Broadest Sense
  - Cognitive science
  - Mental models
- Joint Development as an Integrated System
  - In conjunction with the Explainable Models
- Existing Machine Learning Techniques
  - Also consider explaining existing ML techniques

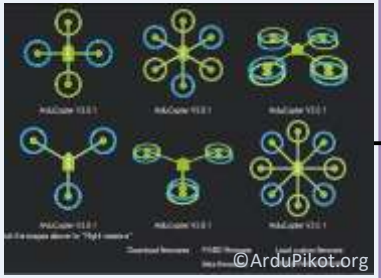
B.3 Psychology of Explanation

- Psychology Theories of Explanation
  - Structure and function of explanation
  - Role of explanation in reasoning and learning
  - Explanation quality and utility
- Theory Summarization
  - Summarize existing theories of explanation
  - Organize and consolidate theories most useful for XAI
  - Provide advice and consultation to XAI developers and evaluator
- Computational Model
  - Develop computational model of theory
  - Generate predictions of explanation quality and effectiveness
- Model Testing and Validation
  - Test model against Phase 2 evaluation results

B.4 Emphasis and Scope of XAI Research



Explainable AI – Challenge Problem Areas

	Learn a model to perform the task	Explain decisions, actions to the user	Use the explanation to perform a task	
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	

C.1 Data Analysis

- Machine learning to classify items, events, or patterns of interest
  - In heterogeneous, multimedia data
  - Include structured/semi-structured data in addition to images and video
  - Require meaningful explanations that are not obvious in video alone
- Proposers should describe:
  - Data sets and training data (including background knowledge sources)
  - Classification function to be learned
  - Types of explanations to be provided
  - User decisions to be supported
- Challenge problem progression
  - Describe an appropriate progression of test problems to support your development strategy

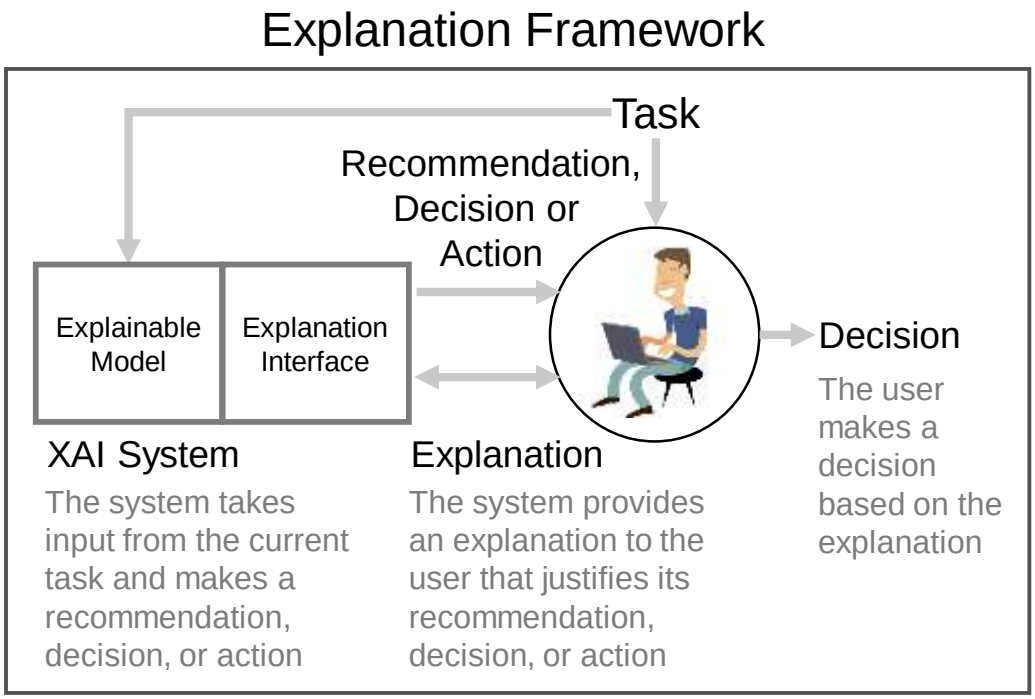
C.2 Autonomy

- Reinforcement learning to learn sequential decision policies
  - For a simulated autonomous agent (e.g., UAV)
  - Explanations may cover other needed planning, decision, or control modules, as well as decision policies learned through reinforcement learning
  - Explain high level decisions that would be meaningful to the end user (i.e., not low level motor control)
- Proposers should describe:
  - Simulation environment
  - Types of missions to be covered
  - Decision policies and mission tasks to be learned
  - Types of explanations to be provided
  - User decisions to be supported
- Challenge problem progression
  - Describe an appropriate progression of test problems to support your development strategy

C.3 Evaluation – Evaluation Sequence

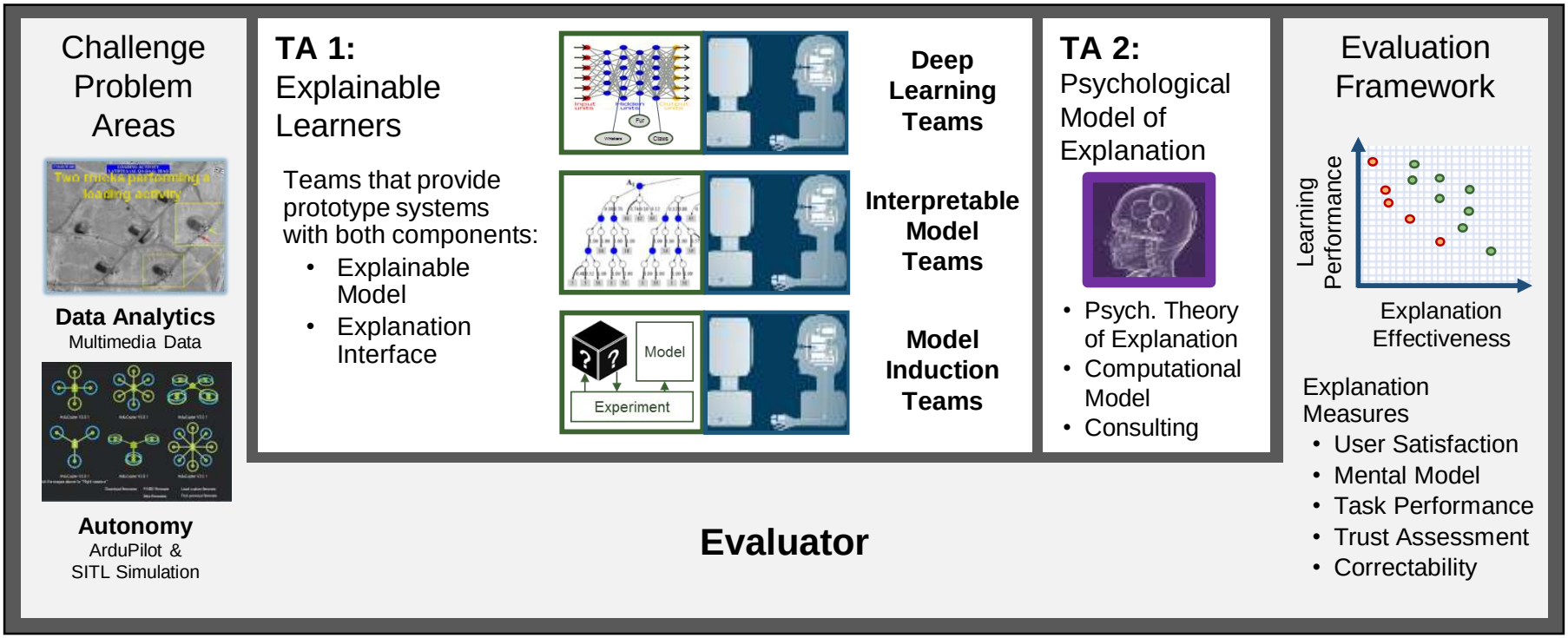
- XAI developers are presented with a problem domain
- Apply machine learning techniques to learn an explainable model
- Combine with the explanation interface to construct an explainable system
- The explainable system delivers and explains decisions or actions to a user who is performing domain tasks
- The system's decisions and explanations contribute (positively or negatively) to the user's performance of the domain tasks
- The evaluator measures the learning performance and explanation effectiveness
- The evaluator also conducts evaluations of existing machine learning techniques to establish baseline measures for learning performance and explanation effectiveness

C.3 Evaluation – Evaluation 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
Correctablity (Extra Credit)
• Identifying errors • Correcting errors, Continuous training

D. Technical Areas



- 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

Expected Team Characteristics

- TA1: Explainable Learners
  - Each team consists of a machine learning and a HCI PI/group
  - Teams may represent one institution or a partnership
  - Teams may represent any combination of university and industry researchers
  - Multiple teams (approximately 8-12 teams) expected
  - Team size ~ $800K-$2M per year
- TA2: Psychological Model of Explanation
  - This work is primarily theoretical (including the development of a computational model of the theory)
  - Primarily university teams are expected (but not mandated)
  - One team expected

D.1 Technical Area 1 – Explainable Learners

- Challenge Problem Area
  - Select one or both of the challenge problems areas: data analytics or autonomy
  - Describe the proposed test problem(s) you will work on in Phase 1
- Explainable Model
  - Describe the proposed machine learning approach(s) for learning explainable models
- Explanation Interface
  - Describe your approach for designing and developing the explanation interface
- Development Progression
  - Describe the development sequence you intend to follow
- Test and Evaluation Plan
  - Describe how you will evaluate your work in the first phase of the program
  - Describe how you will measure learning performance and explanation effectiveness

D.2 Technical Area 2 – Psychological Model

- Theories of Explanation
  - Describe how you will summarize the current psychological theories of explanation
  - Describe how this work will inform the development of the TA1 XAI systems
  - Describe how this work will inform the definition of the evaluation framework for measuring explanation effectiveness by the XAI evaluator
- Computational Model
  - Describe how you will develop and implement a computational model of explanation
  - Identify predictions that might be tested with the computational model
  - Explain how you will test and refine the model
- Model Validation
  - Describe how you will validate the computational model against the TA1 evaluation results in Phase 2 of the XAI program
  - The government evaluator will not conduct evaluation of TA2 models