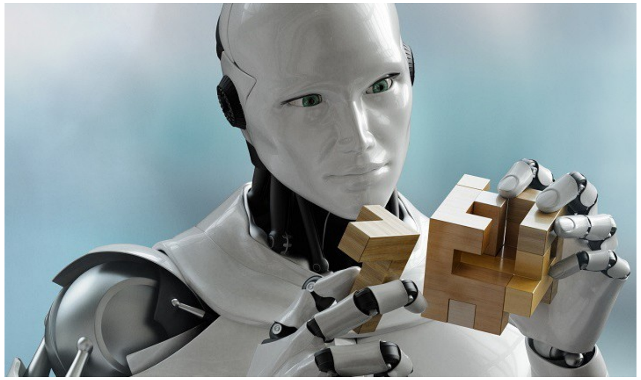
Machine Learning

CS60050

Explainable and Interpretable Machine Learning



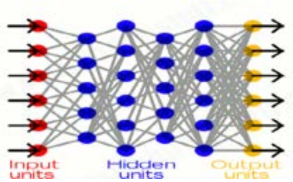
Explainable Artificial Intelligence (XAI)

David Gunning

DARPA/I2O

Explainable AI – What Are We Trying To Do?

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

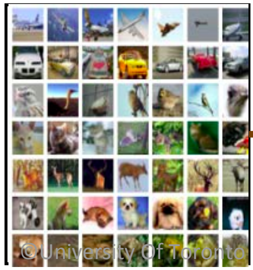
User

- 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?

Dramatic success in machine learning has led to an explosion of AI applications. Researchers have developed new AI capabilities for a wide variety of tasks. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems will be limited by the machine's inability to explain its thoughts and actions to human users. Explainable AI will be essential, if users are to understand, trust, and effectively manage this emerging generation of artificially intelligent partners.

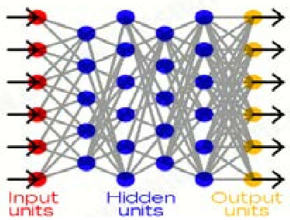
Explainable AI – What Are We Trying To Do?

Today



Training Data

Learning Process



Learned Function

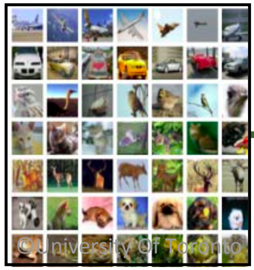
This is a cat
(p = .93)

Output

User with a Task

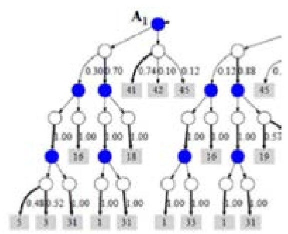
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

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

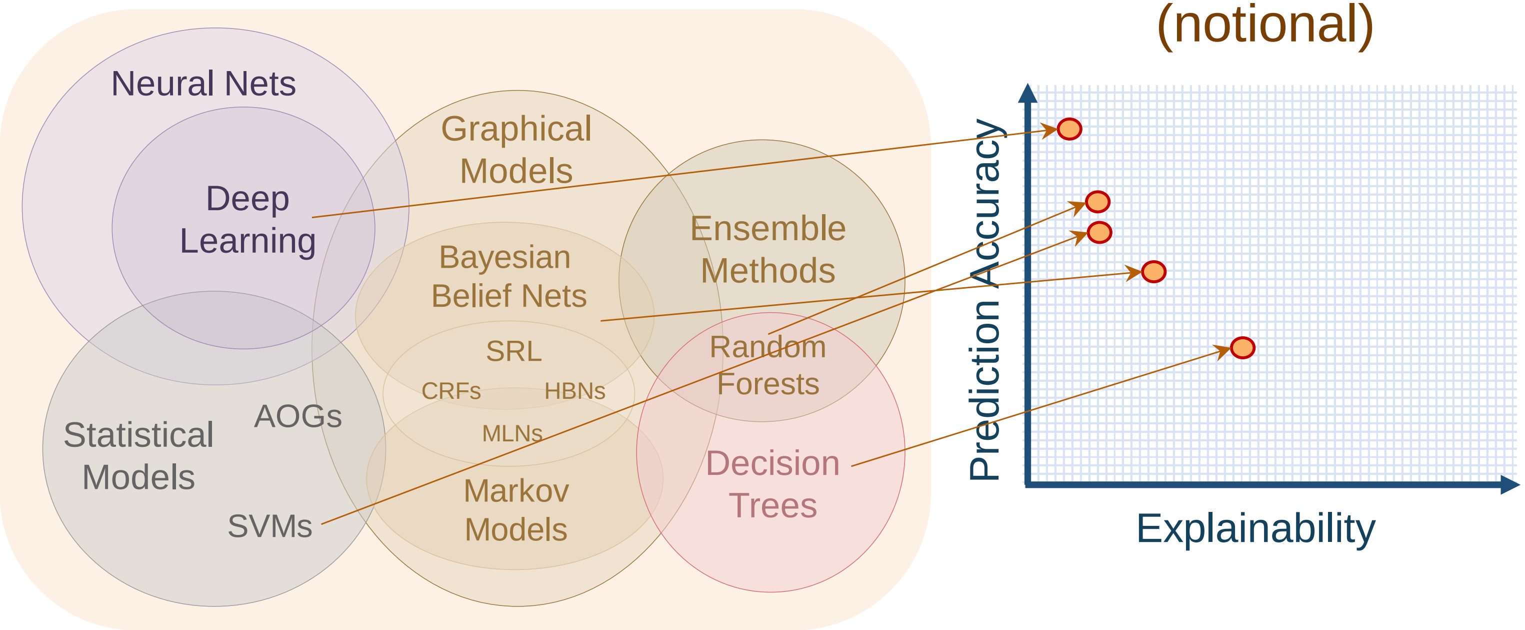
Explainable AI – Performance vs. Explainability

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)

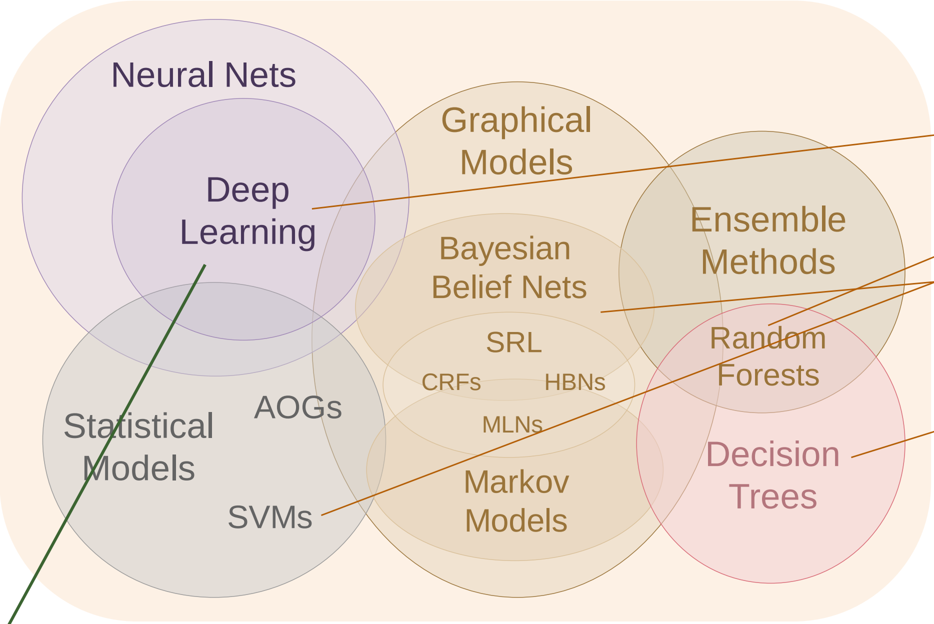


Explainable AI – Performance vs. Explainability

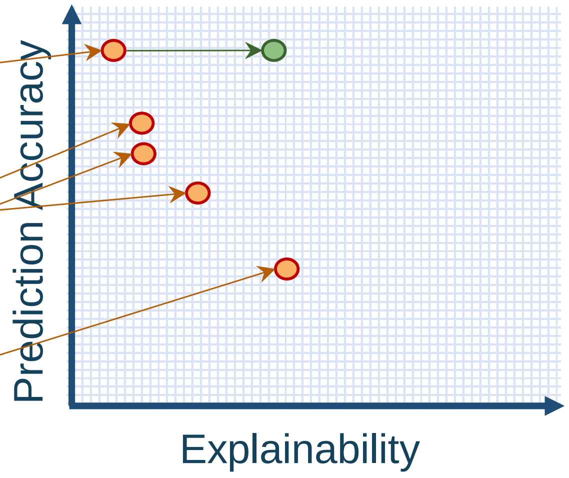
New Approach

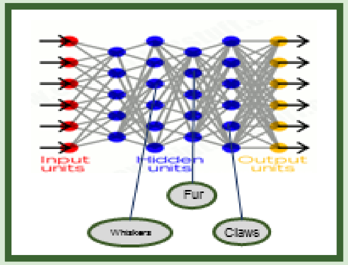
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)



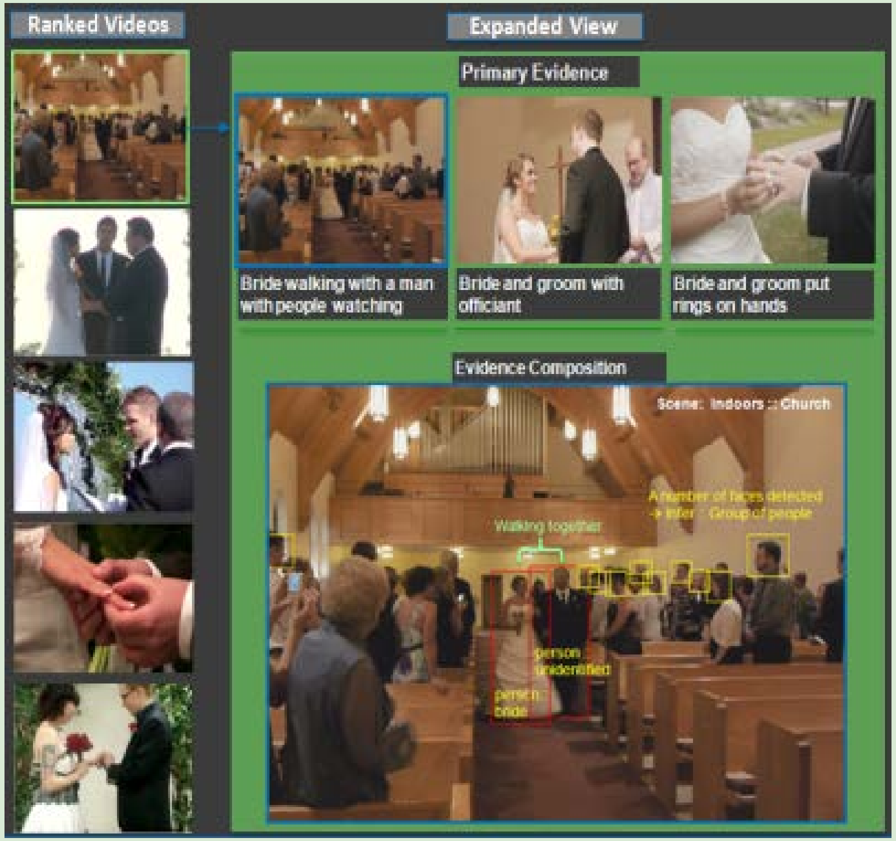


Deep Explanation

Modified deep learning techniques to learn explainable features

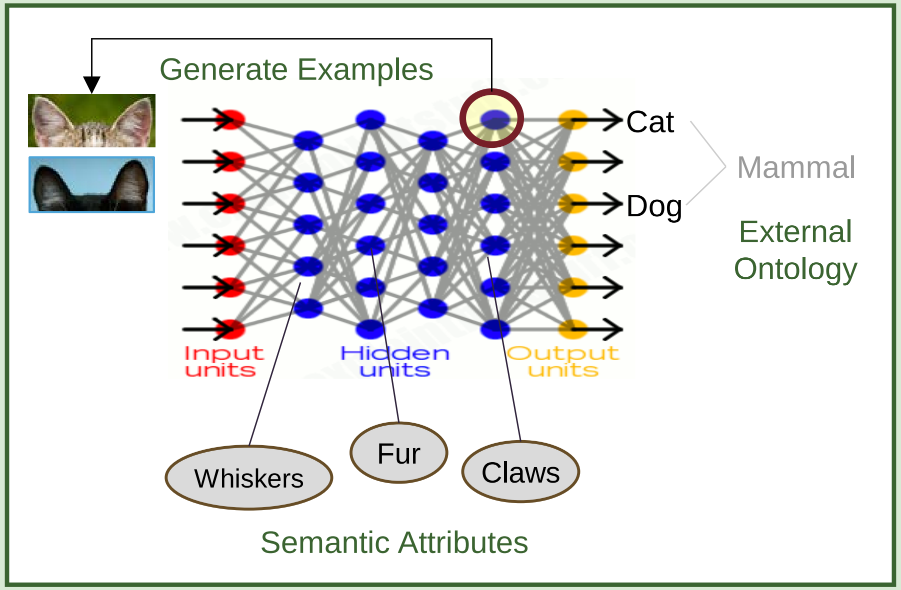
Learning Deep Explanations

Multimedia Event Recounting



- This illustrates and example of event recounting.
- The system classified this video as a wedding.
- The frames above show its evidence for the wedding classification

Learning Semantic Associations

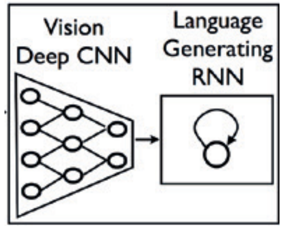


- Train the net to associate semantic attributes with hidden layer nodes
- Train the net to associate labelled nodes with known ontologies
- Generate examples of prominent but unlabeled nodes to discover semantic labels
- Generate clusters of examples from prominent nodes
- Identify the best architectures, parameters, and training sequences to learn the most interpretable models

Cheng, H., et al. (2014) SRI-Sarnoff AURORA at TRECVID 2014: Multimedia Event Detection and Recounting. http://www-nlpir.nist.gov/projects/tvpubs/tv14.papers/sri_aurora.pdf

Learning To Generate Explanations

Generating Image Captions

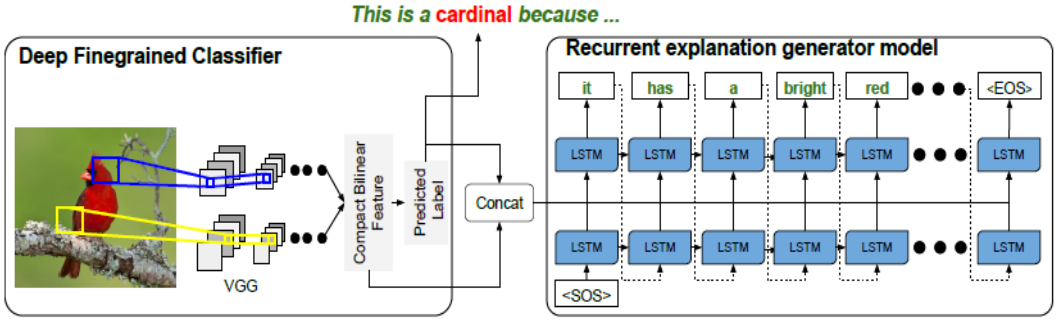


A group of people shopping at an outdoor market

There are many vegetables at the fruit stand

- A CNN is trained to recognize objects in images
- A language generating RNN is trained to translate features of the CNN into words and captions.

Generating Visual Explanations



Researchers at UC Berkeley have recently extended this idea to generate explanations of bird classifications. The system learns to:

- Classify bird species with 85% accuracy
- Associate *image descriptions* (discriminative features of the image) with *class definitions* (image-independent discriminative features of the class)

Example Explanations

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

Limitations

- Limited (indirect at best) explanation of internal logic
- Limited utility for understanding classification errors

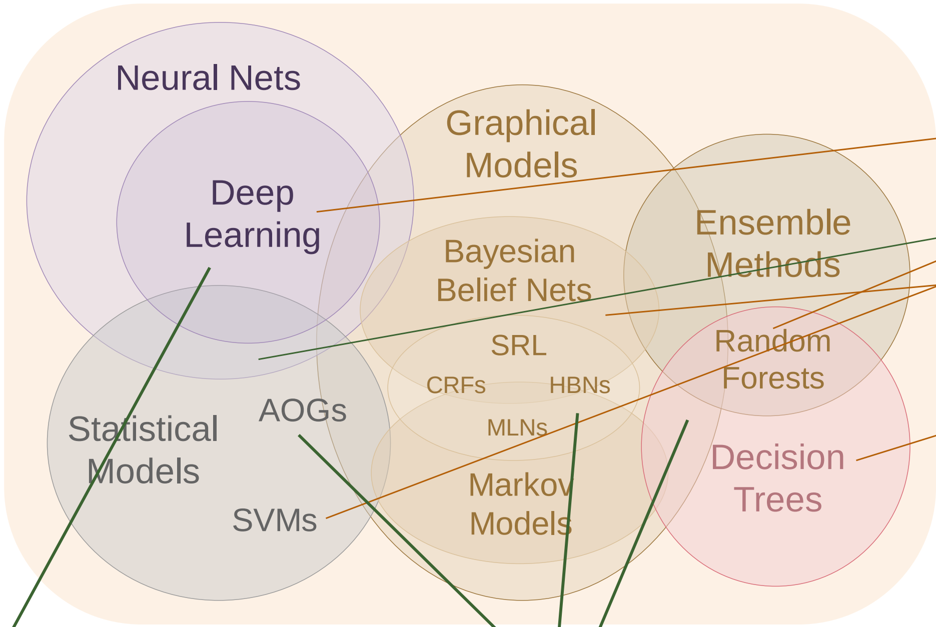
Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations, arXiv:1603.08507v1 [cs.CV] 28 Mar 2016

Explainable AI – Performance vs. Explainability

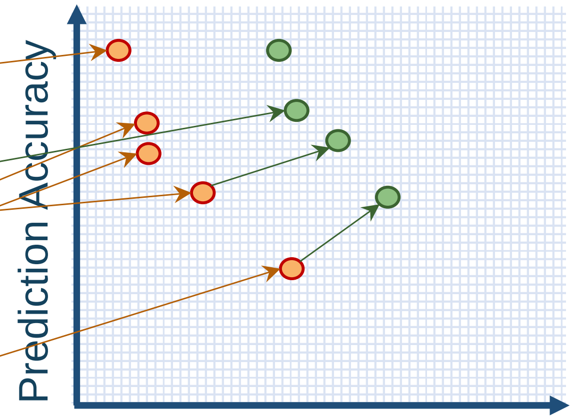
New Approach

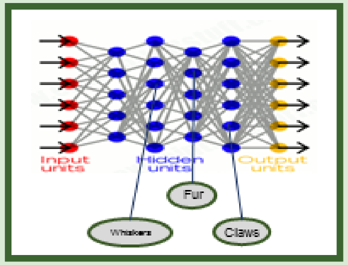
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



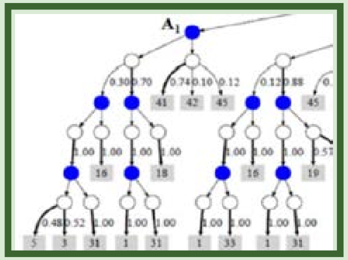
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features

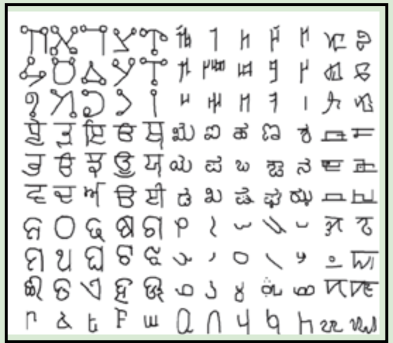


Interpretable Models

Techniques to learn more structured, interpretable, causal models

Learning More Interpretable Models

Training Data
1623 Characters

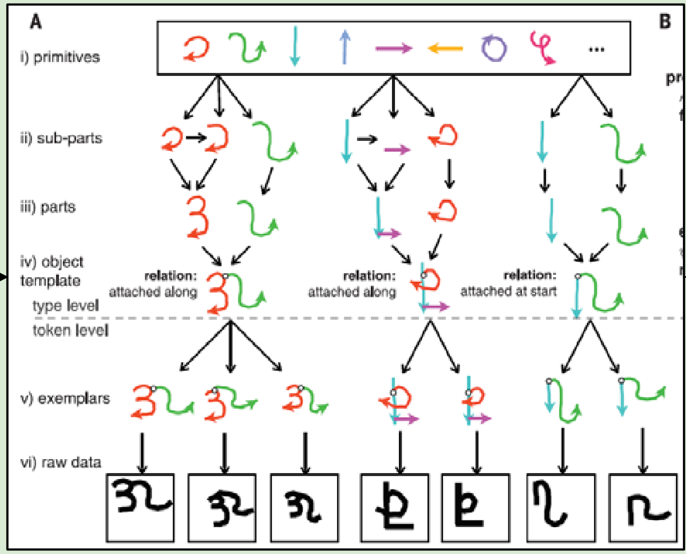


Bayesian Program Learning

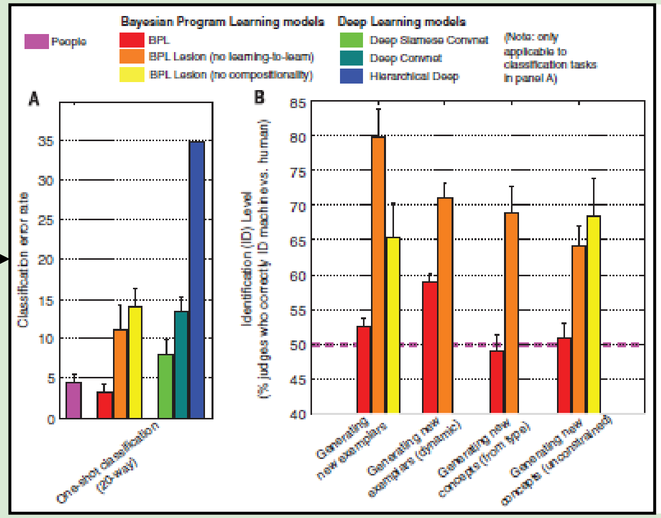
```
num_strokes = Poisson(2)
for i = 1 to num_strokes:
     num_substrokes_prior[i] = Discrete([0,1,1,1,1])
     num_substrokes[i] = Poisson(num_substrokes_prior[i])
     for j = 1 to num_substrokes[i]:
          substrokes[i][j] = substroke_transition_prob[i][j-1]
     relation[i] = relation_prob(substrokes)

for i = 1 to num_strokes:
     noised_substrokes[i][:] = stroke_noise(substrokes[i][:])
     stroke_start_position[i] = start_distribution(relation, s
     trajectory[i] = draw_trajectory(stroke_start_position[i],
AffineTransform = transform_distribution
image = render(AffineTransform(trajectory))
```

Seed Model
A simple Probabilistic Program that describes the parameters of character generation

Concept Learning Through Probabilistic Program Induction



Generative Model
Recognizes characters by generating an explanation of how a new test character might be created (i.e., the most probable sequence of strokes that would create that character)

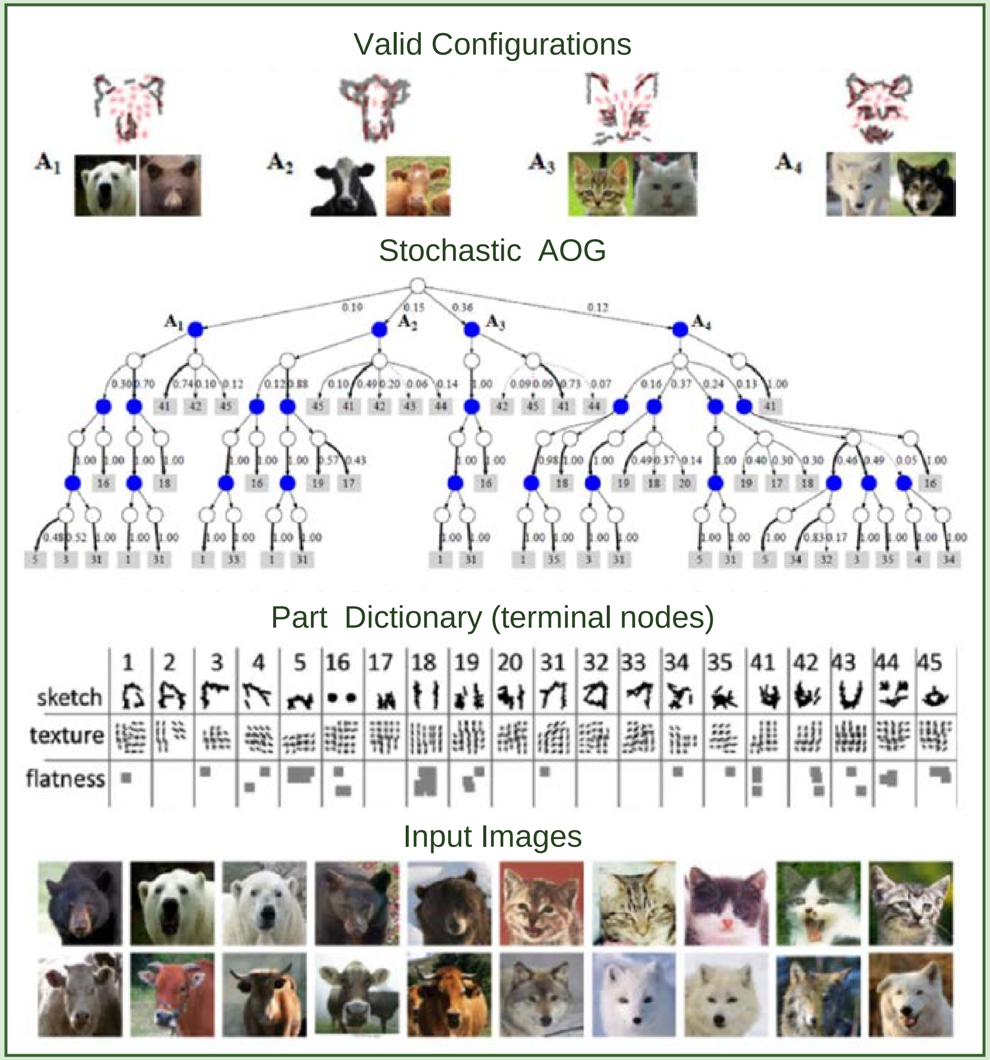


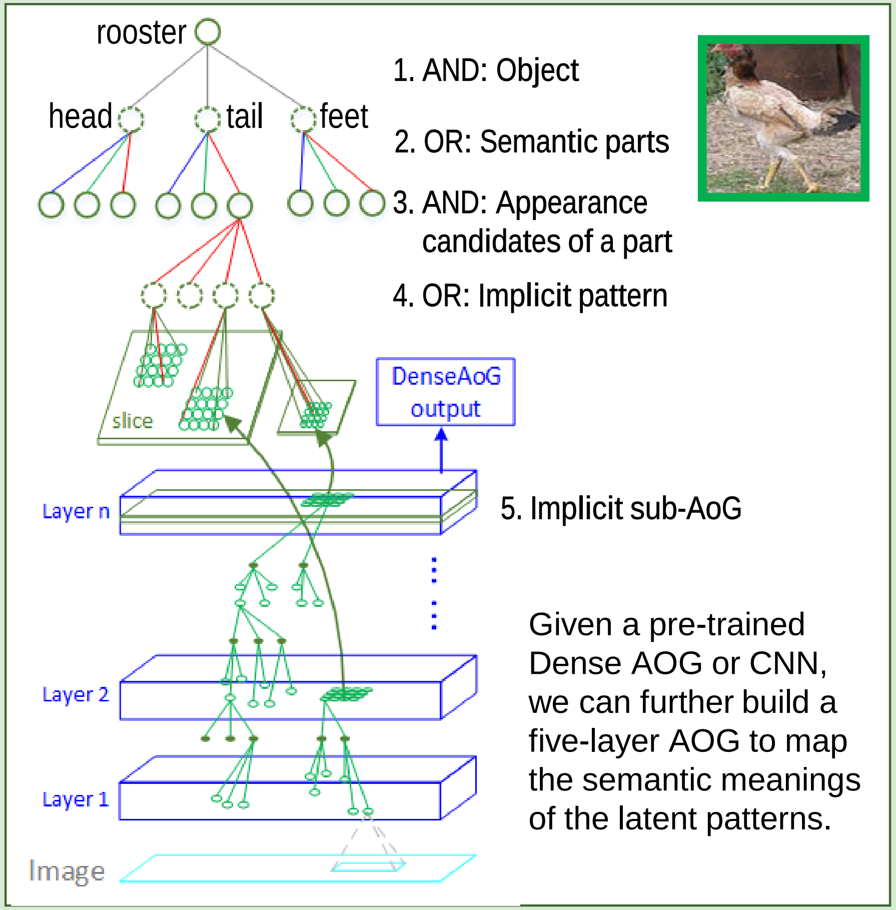
Performance
This model matches human performance and out performs deep learning

Lake, B.H., Salakhutdinov, R., & Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. *Science.* VOL 350, 1332-1338.

Learning More Interpretable Models

Stochastic And-Or-Graphs (AOG)





$$L(\theta) = \frac{1}{M}\sum_{m=1}^{M} \underbrace{\log P(I_m, \theta)}_{\text{generative}} + \underbrace{L(pg_m^*, \hat{pg}_m)}_{\text{discriminative}}$$

Si, Z. and Zhu, S. (2013). Learning AND-OR Templates for Object Recognition and Detection. *IEEE Transactions On Pattern Analysis and Machine Intelligence*. Vol. 35 No. 9, 2189-2205.

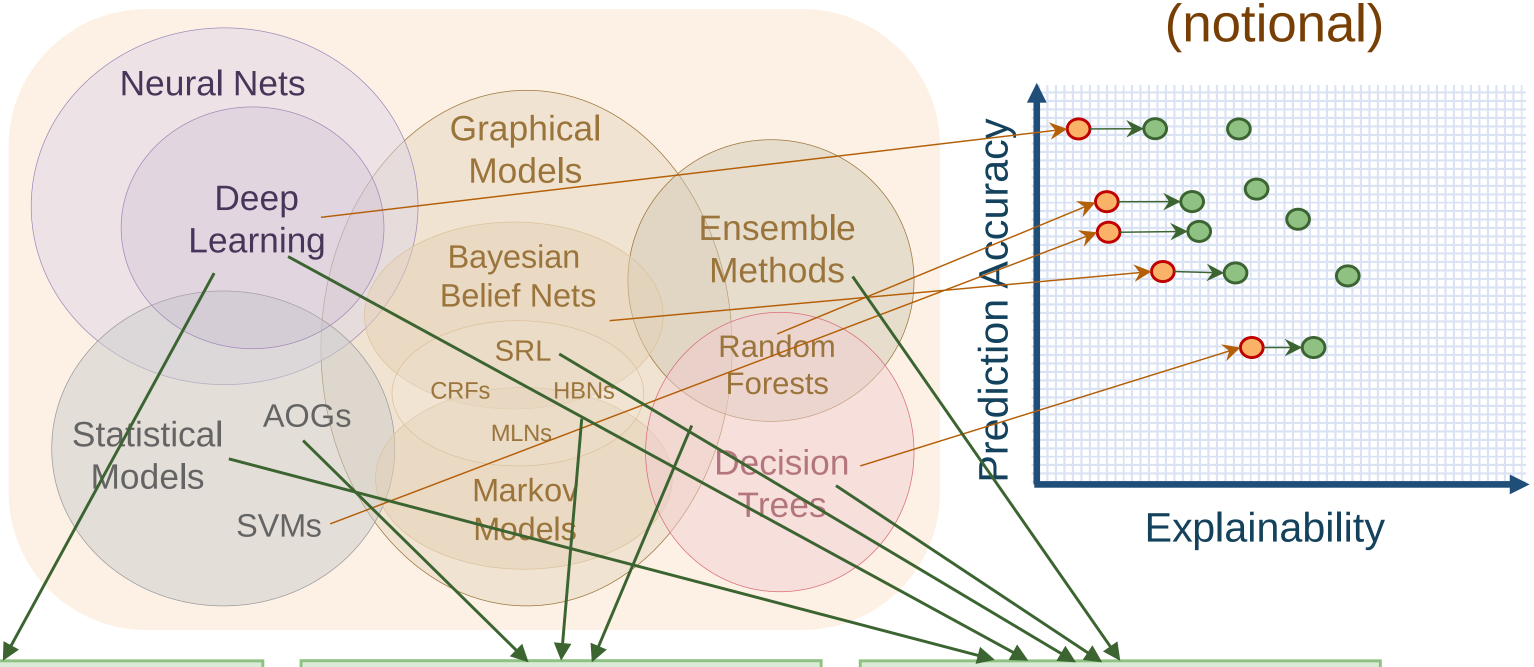
Explainable AI – Performance vs. Explainability

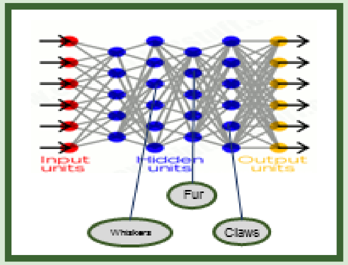
New Approach

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

Learning Techniques (today)

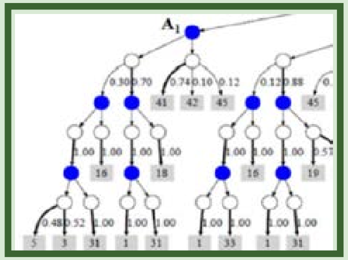
Explainability (notional)





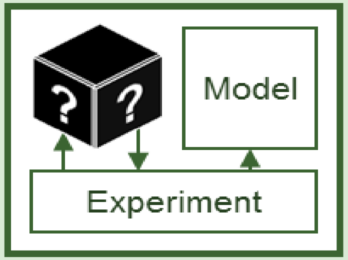
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models



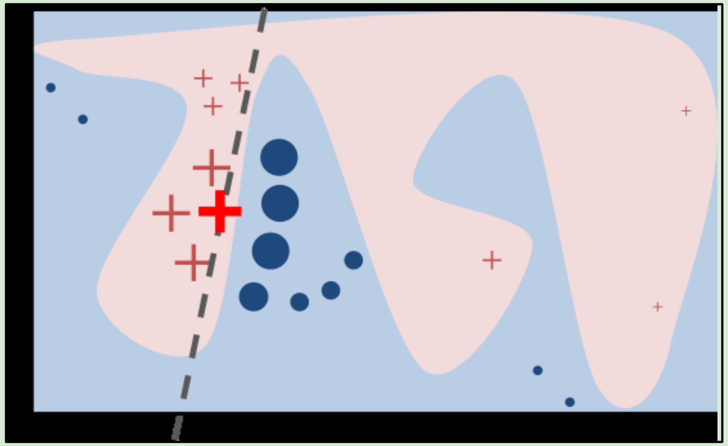
Model Induction

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

Model Induction

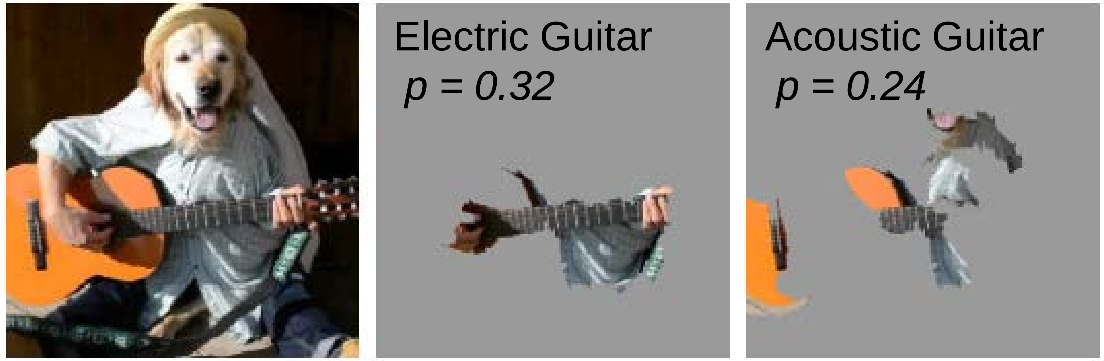
Local Interpretable Model-agnostic Explanations (LIME)

Black-box Induction



The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background. The bright bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful. .

Example Explanation



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar*

- **LIME** is an algorithm that can explain the predictions of any classifier in a faithful way, by approximating it locally with an interpretable model.

- **SP-LIME is** a method that selects a set of representative instances with explanations as a way to characterize the entire model.

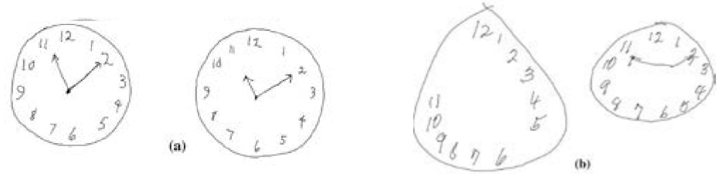
Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *CHI 2016 Workshop on Human Centered Machine Learning. (arXiv:1602.04938v1 [cs.LG] 16 Feb 2016)*

Model Induction

Bayesian Rule Lists (BRL)

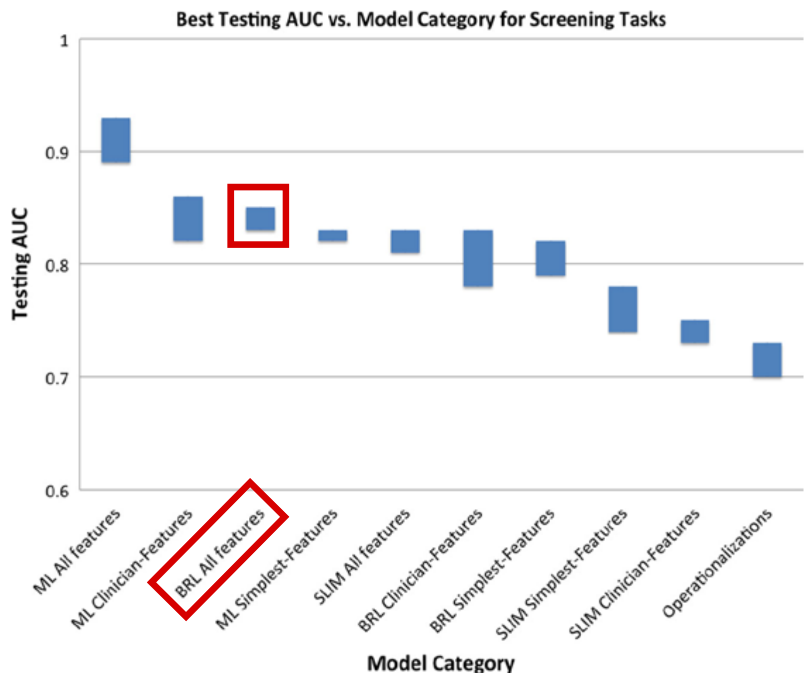
- **if** hemiplegia and age > 60
  - **then** stroke risk 58.9% (53.8%–63.8%)
- **else if** cerebrovascular disorder
  - **then** stroke risk 47.8% (44.8%–50.7%)
- **else if** transient ischaemic attack
  - **then** stroke risk 23.8% (19.5%–28.4%)
- **else if** occlusion and stenosis of carotid artery without infarction
  - **then** stroke risk 15.8% (12.2%–19.6%)
- **else if** altered state of consciousness and age > 60
  - **then** stroke risk 16.0% (12.2%–20.2%)
- **else if** age ≤ 70
  - **then** stroke risk 4.6% (3.9%–5.4%)
- **else** stroke risk 8.7% (7.9%–9.6%)

Clock Drawing Test



Normal Function

Cognitive Impairment



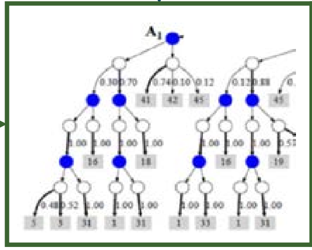
- BRLs are decision lists--a series of if-then statements
- BRLs discretize a high-dimensional, multivariate feature space into a series of simple, readily interpretable decision statements.
- Experiments show that BRLs have predictive accuracy on par with the current top ML algorithms (approx. 85-90% as effective) but with models that are much more interpretable

Letham, B., Rudin. C., McCormick, T., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics 2015*, Vol. 9, No. 3, 1350-137

Explainable AI – Why Do You Think It Will Be Successful?

Training Data

New Learning Process

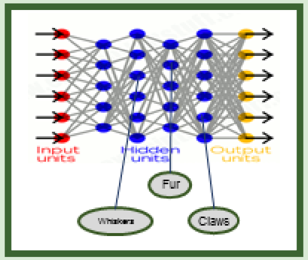


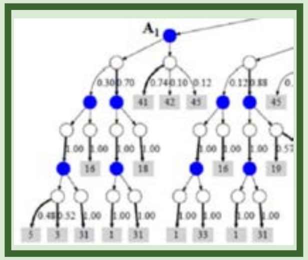
Explainable Model

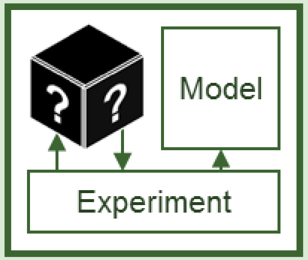
This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:

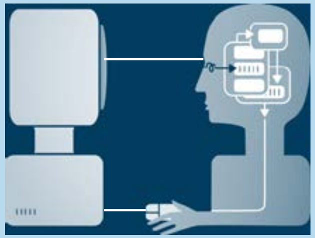
Explanation Interface

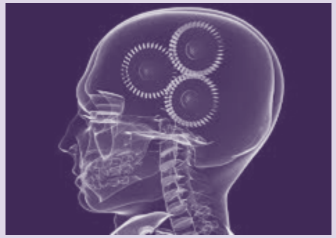
- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred











Deep Explanation	Interpretable Models	Model Induction	HCI	Psychology
Learning Semantic Associations H. Sawhney (SRI Sarnoff)	**Stochastic And-Or-Graphs (AOG)** Song-Chun Zhu (UCLA)	**Local Interpretable Model-agnostic Explanations (LIME)** C. Guestrin (UW)	**Prototype Explanation Interface** T. Kulesza (OSU/MSR)	**Principles of Explanatory Machine Learning** M. Burnett (OSU)
Learning to Generate Explanations T. Darrell, P. Abeel (UCB)	**Bayesian Program Learning** J. Tenenbaum (MIT)	**Bayesian Rule Lists** C. Rudin (MIT)	**UX Design, Language Dialog, Visualization** ENGINEERING PRACTICE	**Psychological Theories of Explanation** T. Lombrozo (UCB)

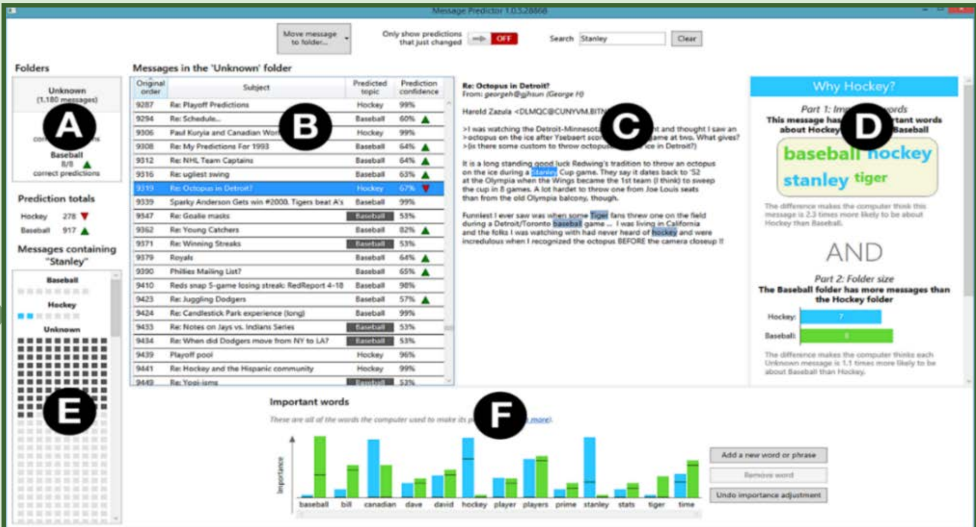
Explanation Interface – A Simple Example

Principles

Explainability
- Be Iterative
- Be Sound
- Be Complete
- Don't Overwhelm

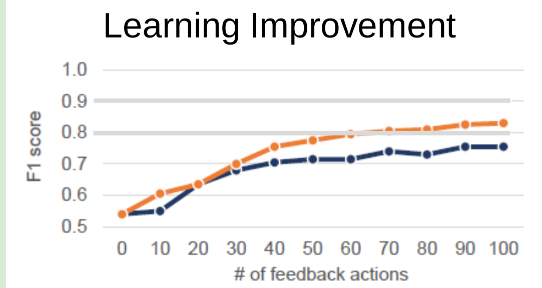
Correctability
- Be Actionable
- Always Honor User Feedback
- Incremental Changes Matter

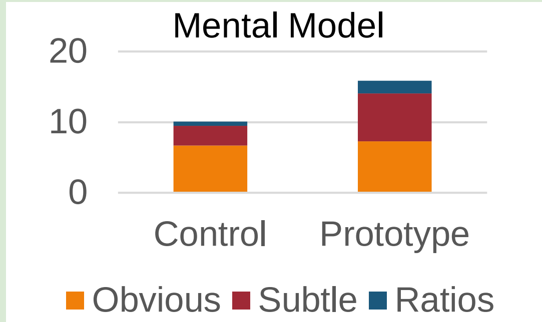
Prototype



(A) List of folders; (B) List of messages in the folder; (C) The selected message; (D) Explanation of the message's predicted folder; (E) Overview of messages; (F) Complete list of words the system used to make predictions

Results

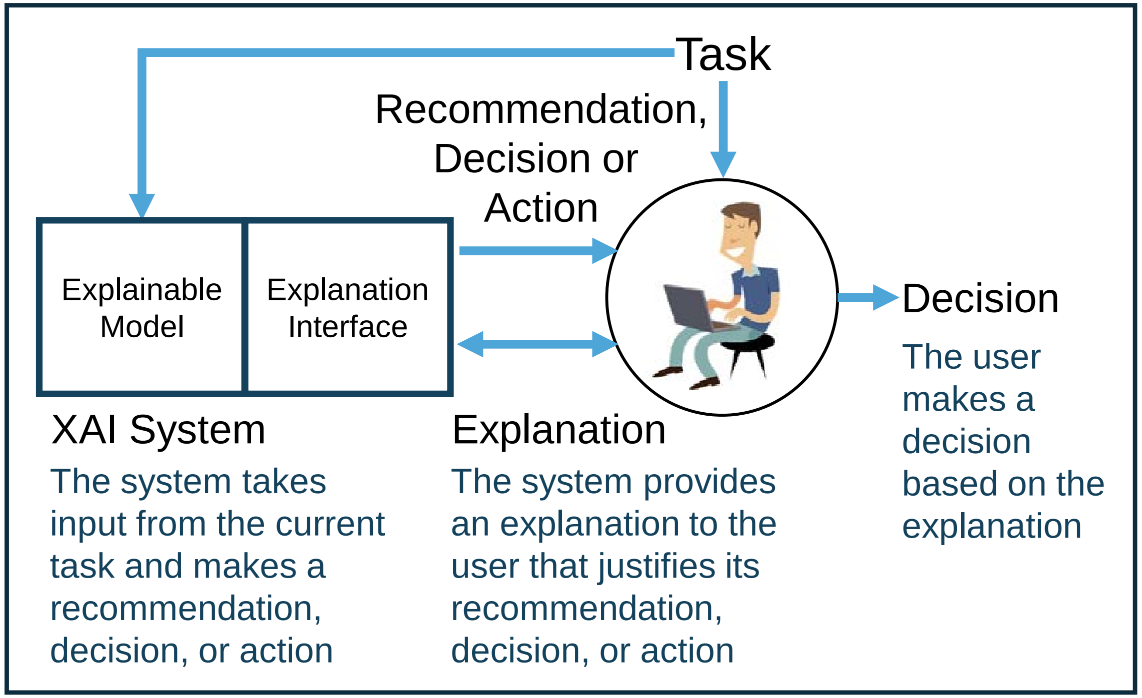




Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. *IUI 2015, Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126-137).

Explainable AI – Measuring Evaluation 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

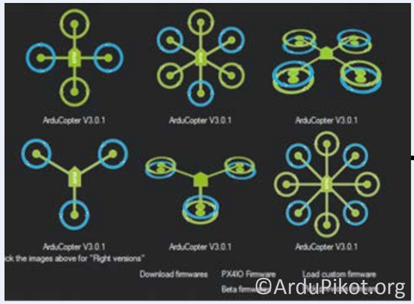
Trust Assessment

- Appropriate future use and trust

Correctablity

- Identifying errors
- Correcting errors
- Continuous training

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	

www.darpa.mil

Thank You!