Statistical Machine Learning

Explainable AI – Interpretable ML
Interpretable Machine Learning
PART-I&II
Spring 2020

Part I: Two different approaches to AI

- Symbolic AI: It had dominated the field from 1950s to 1980s. It used mathematical symbols to represent objects and the relationship between objects.
- Coupled with extensive knowledge bases built by humans, such systems proved to be impressively good at reasoning and reaching conclusions about different domains.
- But by the 1980s, it was known that symbolic AI was impressively bad at dealing with the fluidity of symbols, concepts, and reasoning in real life.

Part I: Two different approaches to AI

- Connectionist AI: In response shortcomings of symbolic AI, researchers began advocating for artificial neural networks (ANN), the precursors of today's deep-learning networks.
- The idea in any such system is to process signals by sending them through a network of simulated nodes: analogs of neurons in the human brain.
- The signals pass from node to node along connections, or links: analogs of the synaptic junctions between neurons, and learn, by adjusting the weights that amplify or damp the signals carried by each connection.

Part I: Two different approaches to AI

- Most networks arrange the nodes as a series of layers that are roughly analogous to different processing centers in the cortex.
- Once activated, these nodes propagate their activation levels through the weighted connections to other nodes in the next level, which combine the incoming signals and are activated (or not) in turn.
- This continues until the signals reach an output layer of nodes, where the pattern of activation provides an answer.

Part I: Two different approaches to AI

- Finally, backpropagation algorithm works its way back down through the layers, adjusting the weights for a better outcome the next time.
- By the end of the 1980s, such neural networks had turned out to be much better than symbolic AI at dealing with noisy or ambiguous input.
- Yet the standoff between the two approaches still wasn't resolved, mainly because the AI systems that could fit into the computers of the time were so limited.

Part I: Two different approaches to AI

- With the advent of computers that were orders of magnitude more powerful and social media sites offering a tsunami of images, sounds, and other training data (Big Data), researchers started training networks that were considerably deeper.
- The number of layers were increasing from one or two to about half a dozen; commercial networks today often use more than 100 layers.
- By 2012, researchers showed that deep neural networks could be much better than standard vision systems at recognizing images. They almost halved the error rates: the revolution in deep learning applications took off.

Part I: Two different approaches to AI

- If you just happen to have a few hundred thousand carefully labeled training examples, supervised learning works great.
- But, That is not often the case, and it simply does not work for tasks such as playing a video game where there are no right or wrong answers; there are just strategies that succeed or fail.
- Thus, for most of real world problems you need reinforcement learning.
- A reinforcement learning system playing a video game learns to seek rewards and avoid punishments.

Part I: Two different approaches to AI

- The first successful implementation of reinforcement learning on a deep neural network came in 2015 when a group at DeepMind trained a network to play classic Atari 2600 arcade games.
- The network play equaled or surpassed that of human Atari players.
- In, 2016, DeepMind researchers used a more elaborate version of the same approach with AlphaGo, a network that mastered the complex board game Go, and beat the world-champion human player.

Part I: Deep Learning Drawbacks

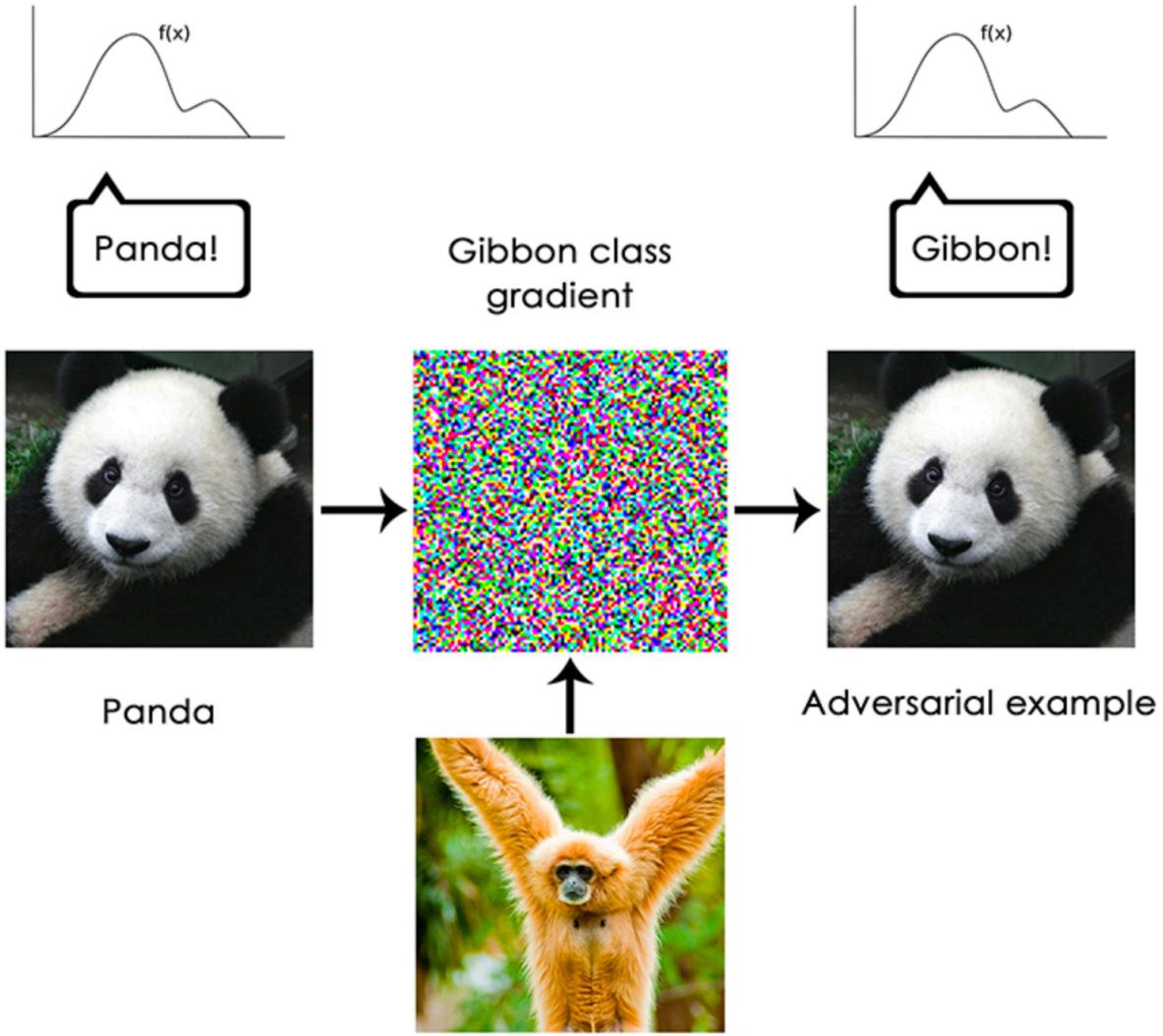
- Despite their high performance, many machine learning techniques remain black boxes because it is difficult to understand the role of each feature and how it combines with others to produce prediction or classification.
- Users need to understand and trust the decisions made by machine learning models, especially in sensitive fields such as medicine and autonomous driving.
- There is an increasing need of methods able to explain the individual predictions of a model; a way to understand what features made the model give its prediction for a specific instance.

Part I: Deep Learning Drawbacks

- Adversarial attack:
- Discovered by the Google Brain highlights just how far AI still has to go before it remotely approaches human capabilities.
- Adversarial examples are input samples to a deep learning network that are designed to trick the model into misclassifying them.
- Many robust training methods have been designed to combat adversarial attacks in deep learning networks.

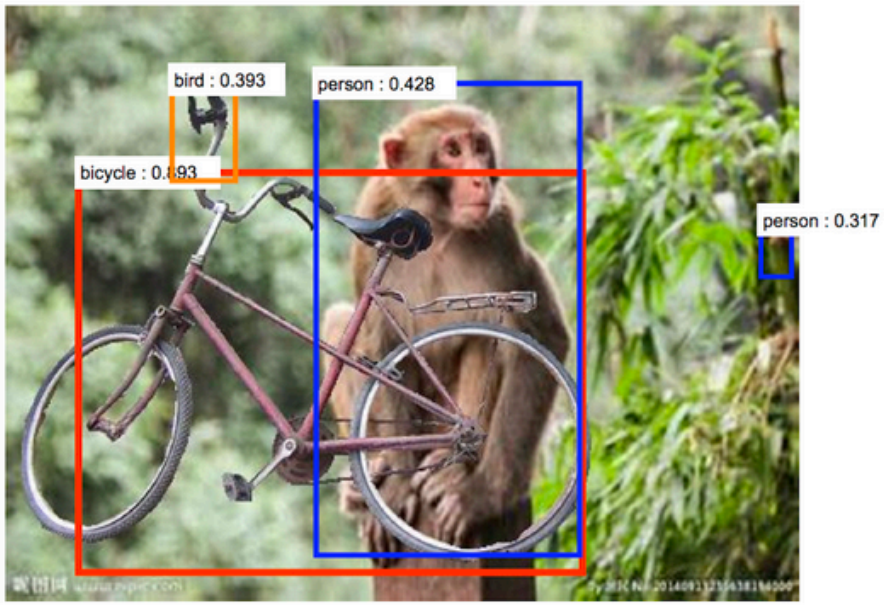
Part I: Deep Learning Drawbacks

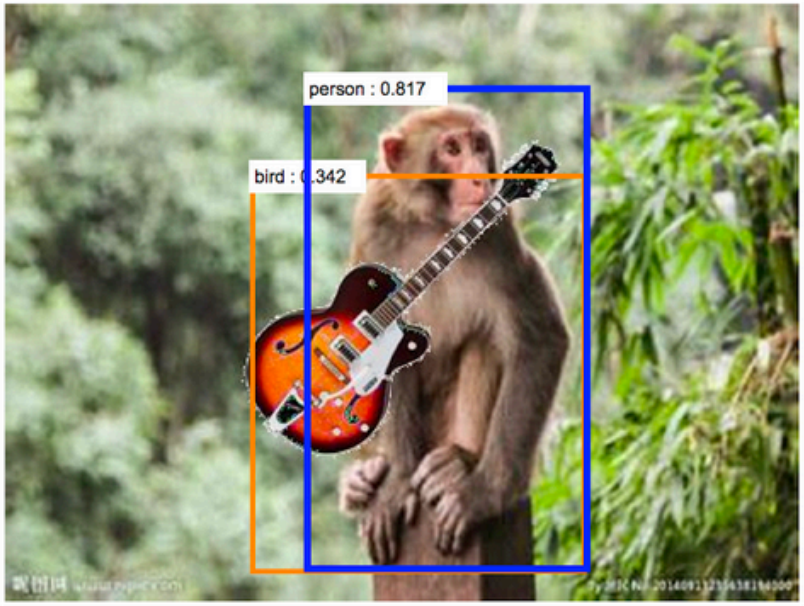
- Adversarial attack:



Part I: Deep Learning Drawbacks

- Adversarial attack:





Left: The occluding bicycle turns a monkey into a person and the jungle turns the bicycle handle into a bird.
Right: The occluding guitar turns the monkey into a person and the jungle turns the guitar into a bird.

Part I: Deep Learning Drawbacks

- Gross inefficiency:
- For a child to learn to recognize a cat, it is not necessary to cat 10,000 time, a number that is often required for deep-learning systems. Humans learn new concepts from just one or two examples.
- Opacity problem:
- Once a deep learning system has been trained, it is not always clear how it is making its decisions.
- In many contexts that is just not acceptable, even if it gets the right answer!

Part I: Deep Learning Drawbacks

- Lack of common sense:
- Deep-learning systems may be great at recognizing patterns in the pixels, but they cannot understand what the patterns mean, nor they can not reason reason about them.
- For example, the current deep learning networks would not be able to understand that chairs are for sitting (function of object).
- There had been a feeling that deep learning is magic. Now people are realizing that it is not magic in many instances!

Part I: Deep Learning Drawbacks

- Small Data:
- Many deep learning algorithms require having a large dataset for good performance. What if we only have access to small data.

	VGGNet	DeepVideo	GNMT
Used For	Identifying Image Category	Identifying Video Category	Translation
Input	Image 	Video 	English Text 
Output	1000 Categories	47 Categories	French Text
Parameters	140M	~100M	380M
Data Size	1.2M Images with assigned Category	1.1M Videos with assigned Category	6M Sentence Pairs, 340M Words
Dataset	ILSVRC-2012	Sports-1M	WMT'14

Part I: Beyond Deep Learning

- It has been shown that deep neural networks are mathematically equivalent to a universal computer, which means there is no computation they cannot perform, if you can ever find the right connection weights.
- However, given the many drawbacks, deep learning require some fundamentally new ideas:
- Data Augmentation: It is a strategy that enables practitioners to significantly increase the diversity of data available for training models, without actually collecting new data.
- Data augmentation techniques such as cropping, and horizontal flipping are commonly used to train large neural networks.

Part I: Beyond Deep Learning

- Basic augmentation techniques include flip, rotation, scale, crop, translation, and adding Gaussian noise.
- Advanced augmentation techniques include kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, generative adversarial networks, neural style transfer.
- Meta-learning; learning to learn: in this context, we may train the network on more than one task.
- As long as the network has enough recurrent connections running backward from later layers to earlier ones, a feature that allows the network to remember what it is doing from one instant to the next.

Part I: Beyond Deep Learning

- Therefore, it will automatically draw on the lessons it learned from earlier tasks to learn new ones faster; it is a big part of our ability to master things quickly.
- Generative Query Network architecture; to give up trying to training just one big network, and instead have multiple networks work in tandem.
- Here, two different networks are utilized to learn their way around complex virtual environments with no human input.
- One, dubbed the representation network, essentially uses standard image-recognition learning to identify what is visible to the AI at any given instant.

Part I: Beyond Deep Learning

- The generation network, learns to take the first network's output and produce a kind of 3D model of the entire environment, making predictions about the objects and features.
- For example, if a table only has three legs visible, the model will include a fourth leg with the same size, shape, and color.
- In other words, An agent that is trying to predict things gets feedback automatically on every time-step, since it gets to see how its predictions turned out.
- Hence, it can constantly update its models to make them better: the learning is selfsupervised.

Part I: Beyond Deep Learning

- Moreover, the researchers do not have to label anything in the environment for it to work or even provide rewards and punishments.
- Graph network: These are deep-learning systems that have an inductive bias toward representing things as objects and relations.
- It is known that infants are also beginning to learn the basics of intuitive psychology, which includes an ability to recognize faces and a realization that the world contains agents that move and act on their own, with many hardwired "inductive biases" that prime them to absorb certain core concepts at a prodigious rate.

Part I: Beyond Deep Learning

- Having this kind of built-in inductive biasing would help deep neural networks learn just as rapidly.
- For example, certain objects such as paws, tail, and whiskers might all belong to a larger object cat, with the relationship is-a-part-of.
- This concept, could be represented as an abstract graph in which the nodes correspond to objects and the links to relationships.
- A graph network, then, is a neural network that takes such a graph as input, as opposed to raw pixels or sound waves, then learns to reason about and predict how objects and their relationships evolve over time.

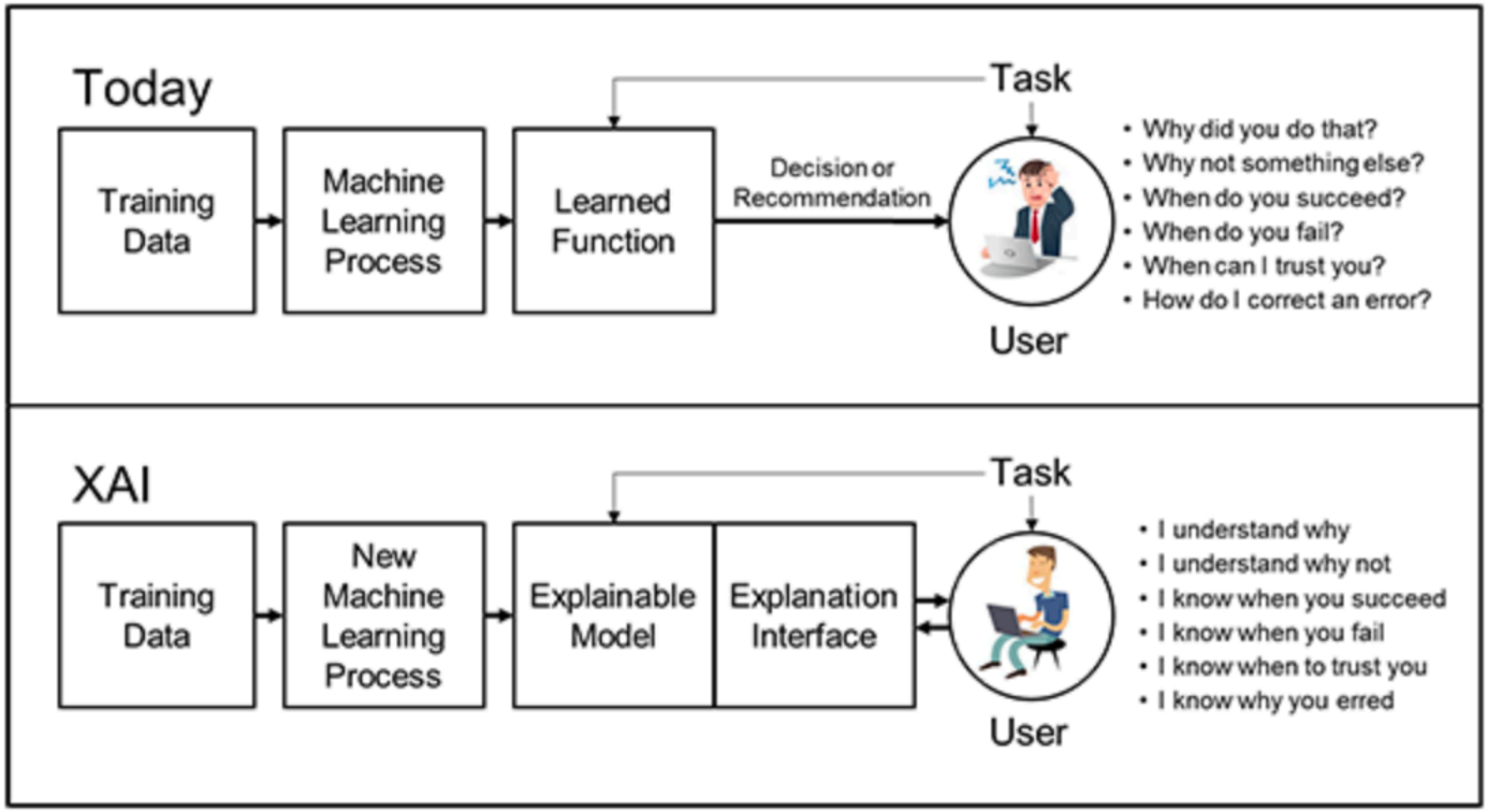
Part I: Beyond Deep Learning

- A graph network could make the networks far less vulnerable to adversarial attacks.
- This is because a system that represents things as objects, as opposed to patterns of pixels, is not going to be so easily thrown off by a little noise or an extraneous sticker (it is more robust).
- Explainable AI, Interpretable Machine Learning:
  - Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems.

Part I: Explainable vs Interpretable

- Explainability and interpretability are often used interchangeably!
- Interpretability is about the extent to which a cause and effect can be observed within a system.
- Interpretability is the extent to which you are able to predict what is going to happen, given a change in input or algorithmic parameters.
- Explainability is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.

Part I: Explainable vs Interpretable



- It's easy to miss the subtle difference with interpretability, but consider it like this: interpretability is about being able to discern the mechanics without necessarily knowing why. Explainability is being able to quite literally explain why it is happening.

Part I: Explainable & Interpretable AI

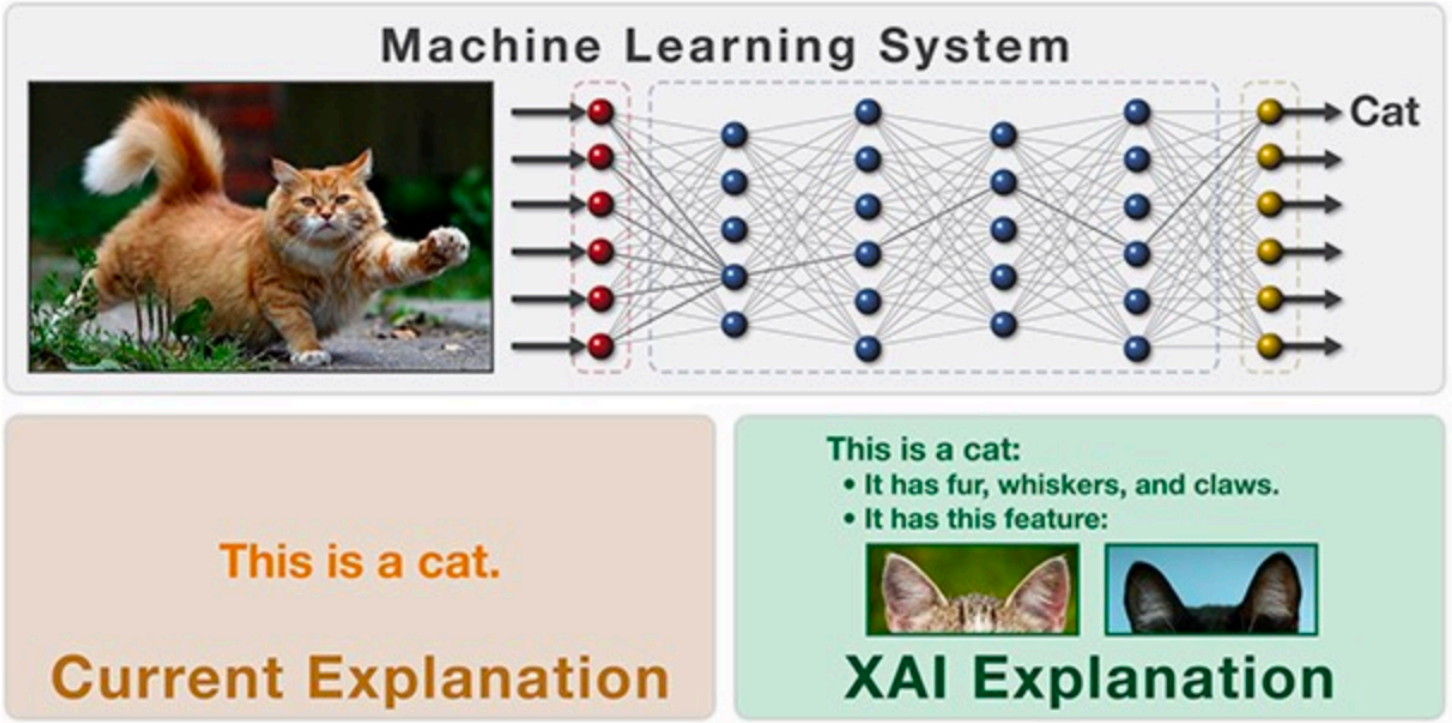


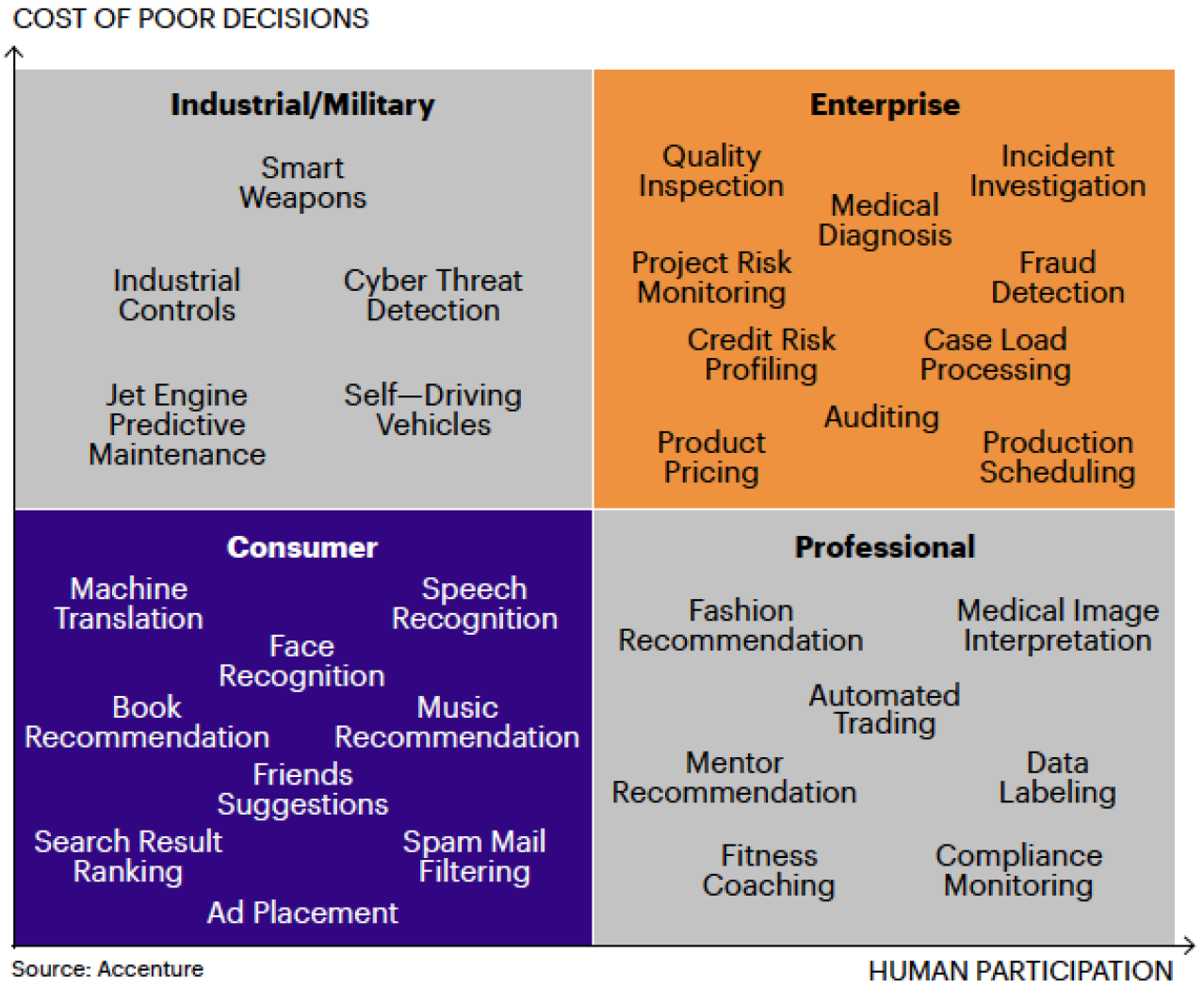
Image classifier: This deep model provides a probability of 0.98 that a cat appears in the picture (*observation*); hence "our model predicts that this is a cat with a probability of 0.98". But we wouldn't really know about the reasons behind this prediction!

Explainable approach: "our model predicts that this is a cat with a probability of 0.98 because it has fur, whiskers, claws and ears with a certain shape"? With that information, we would understand why our neural network (correctly) predicted that there is a cat and, also, we could decide whether to trust or not the prediction. This is the goal of interpretable (explainable) machine learning (AI).

Part I: Explainable & Interpretable AI

- It's a future of "Citizen AI", where AI is here and ready to work alongside its human counterparts.
- And where, by raising AI for responsibility, fairness, and transparency, businesses can create a collaborative, powerful new member of the workforce.
- Explainable and more responsible AI will be the backbone of the intelligent systems of the future that enable the intelligent enterprise.
- In playing this role, Explainable AI won't replace people, but will complement and support them so they can make better, faster, more accurate and more consistent decisions (Human Centered AI).

Part I: Explainable & Interpretable AI



The need for explainable AI rises with the potential cost of poor decisions.

Part I: Explainable & Interpretable AI

Three factors are accelerating progress towards Explainable AI (according to Accenture):

- The growing need for transparency, as required by laws such as the EU's GDPR, mandating how personal data is used for selection and other decision-making. Ethical values also demand transparency, so people can be sure decisions are fair and evenhanded.
- Second, trust. Before humans can act on a system's recommendations, people need to trust it. This trust will only be created if the system can explain the underlying model and process through which decisions are made.
- Third, better human-machine synergy. Machines and humans work differently in how they sense, understand and learn. Machines are better at recognizing low-level patterns in huge amounts of data, while people excel at connecting the dots among high-level patterns.

Part I: Explainable & Interpretable AI

- Explainable AI systems will play this pivotal role through their ability to:

Explain
their rationale;

the reasoning, whenever needed;

Characterize
their strengths and weaknesses

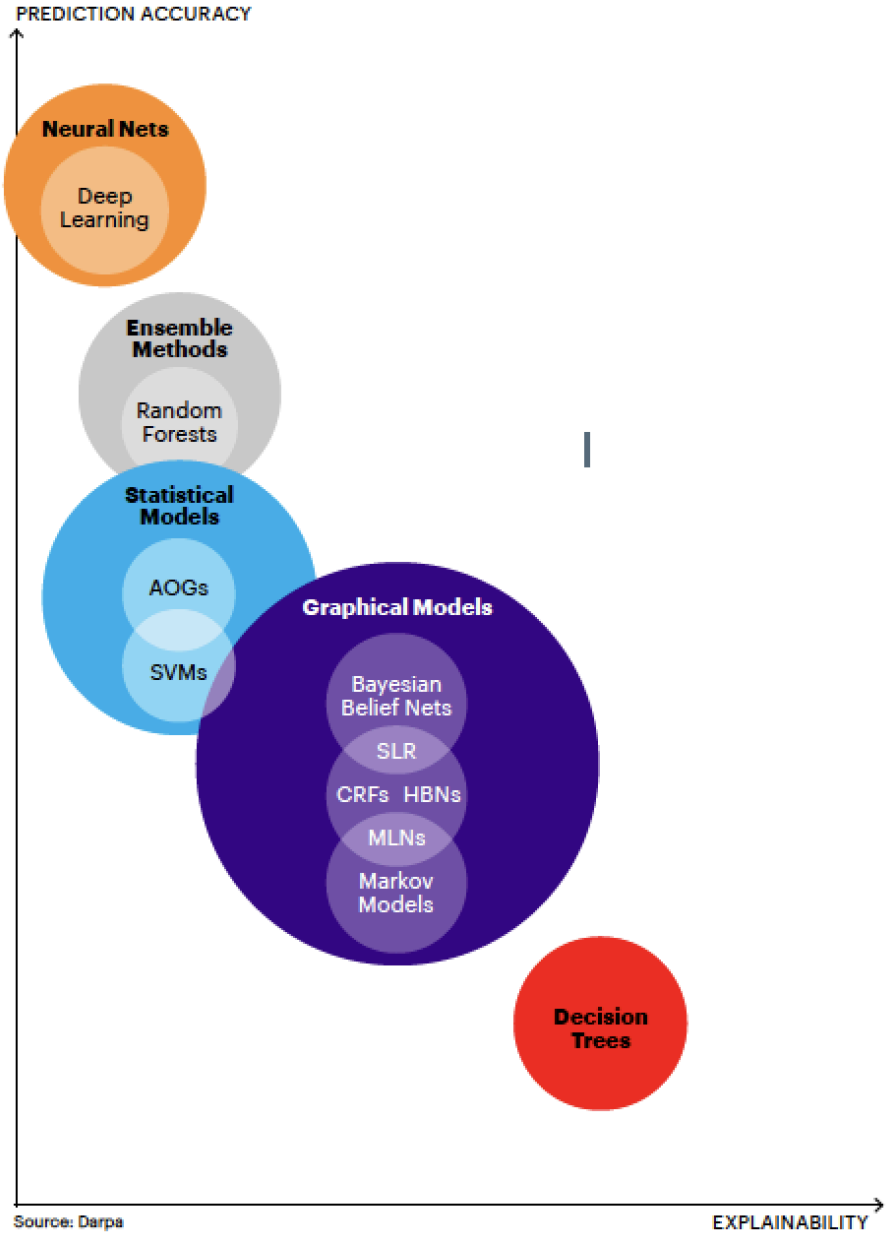
Compare
with other AI systems

Convey
an understanding of how they will behave in the future

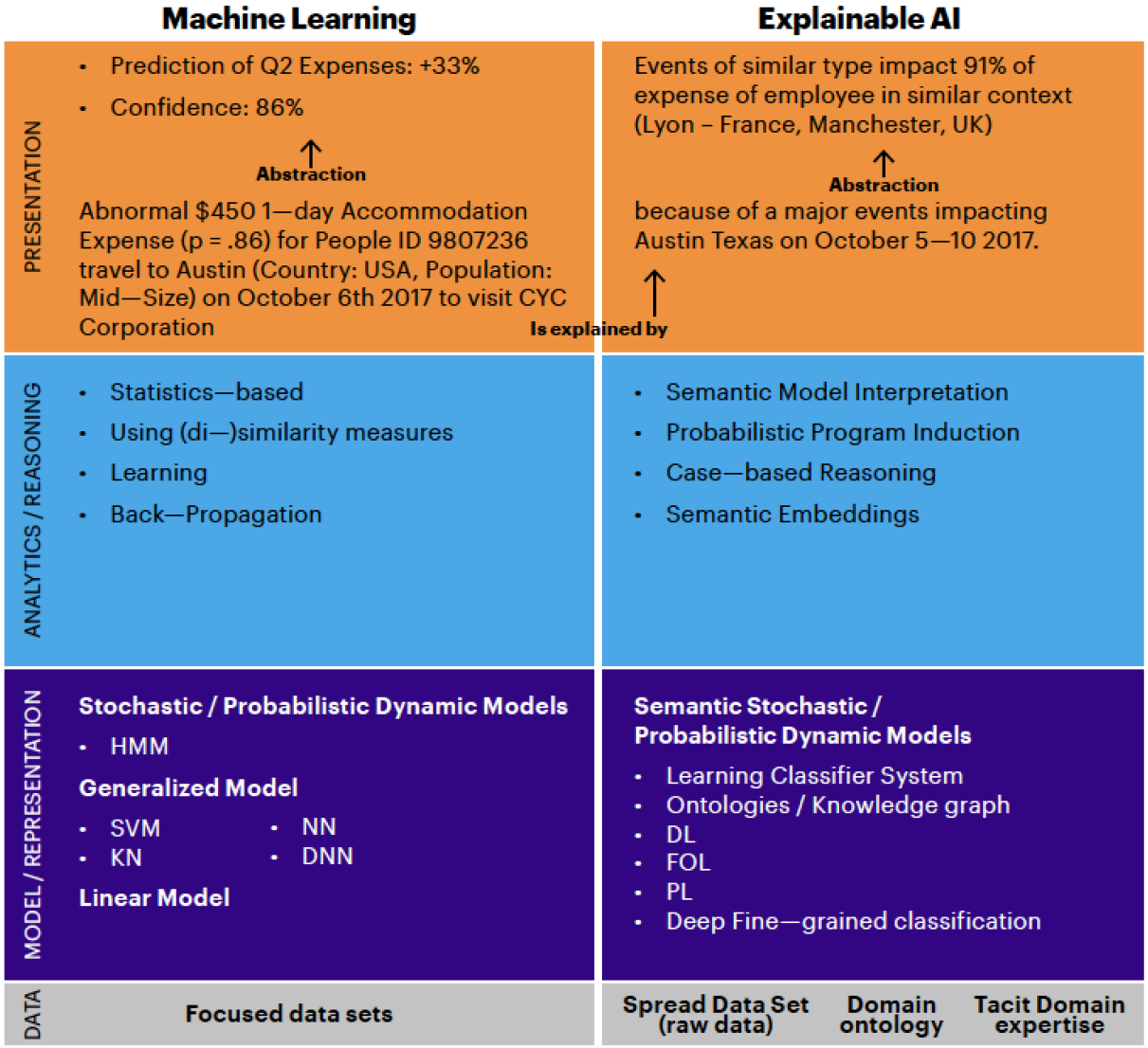
Make
the enterprise scalable through intelligent decisions;

decisions smarter by augmenting humans with machines.

Part I: Explainable & Interpretable AI



Part I: Explainable & Interpretable AI



Source: Accenture

Ways to manifest and convey the reasoning behind AI decisions

Part I: Explainable & Interpretable AI

- Data-level explanation: provides evidence of the modeling by using comparisons with other examples to justify the decisions around a particular classification, clustering or targeted prediction.
- In the case of a mortgage application, an explanation might look like: “The mortgage is not approved because the applicant case is similar to 82% of rejected cases”, thus showing how similar or dissimilar different instances or examples are.
- Model-level explanation: focuses more on the algorithmic basis of the Machine Learning approach.
- Explanations mimic the learning model by abstracting it through rules or combining it with semantics.
- Compared to the data-level and hybrid-level approaches, the model-level approach abstracts most from the data.

Part I: Explainable & Interpretable AI

- In the example of a mortgage application, an explanation might look like: “The mortgage is not approved, because any applicant who has held less than 5,000 Euros of savings for the past 20 months is rejected”, thus showing the logic of the model.
- Hybrid-level explanation: works at a higher level of abstraction, by refactoring data at a metadata level, meaning it’s a particularly useful method if data is big and very packed.
- Instead of providing data as evidence, this approach offers metadata and feature level explanations.
- In the example of a mortgage application, an explanation would look like: “The mortgage is not approved because both the amount and duration of savings are the most important factors”, explaining which factors have the greatest influence on the decision.

Part I: Explainable & Interpretable AI

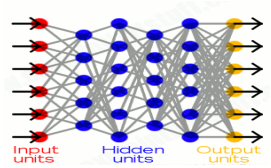
- The measures that people need in an explanation:

Comprehensibility		How much effort is needed for a human to interpret it?
Succinctness		How concise is it?
Actionability		How actionable is the explanation? What can we do with it?
Reusability		Could it be interpreted/reused by another AI system?
Accuracy		How accurate is the explanation?
Completeness		Does the "explanation" explain the decision completely, or only partially?

DARPA's Program on Explainable AI (XAI)

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

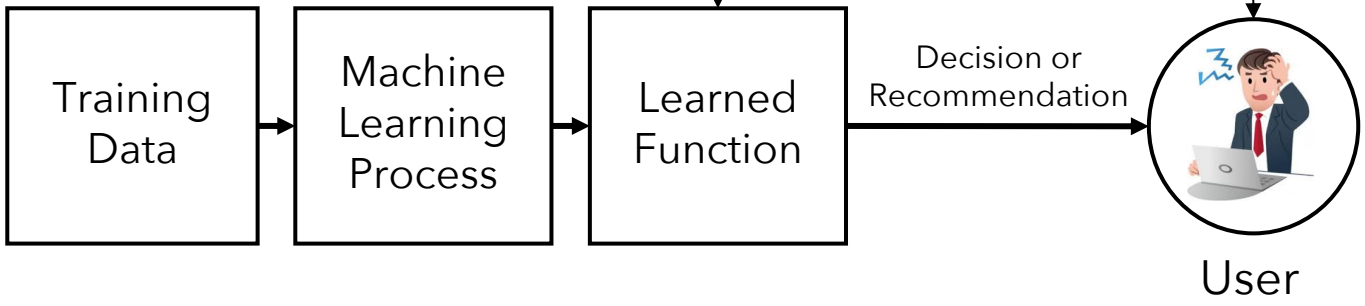
User

- 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.

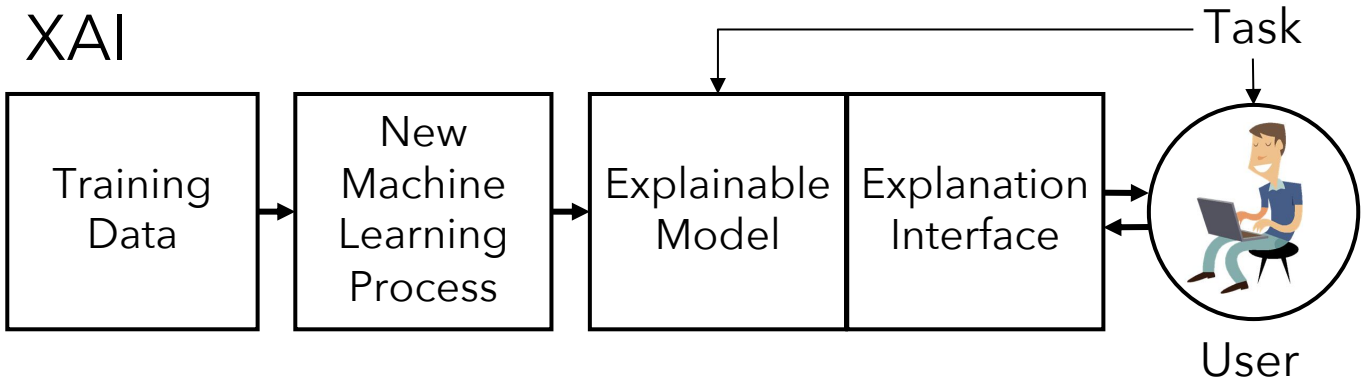
B. Program Scope – XAI Concept

Today



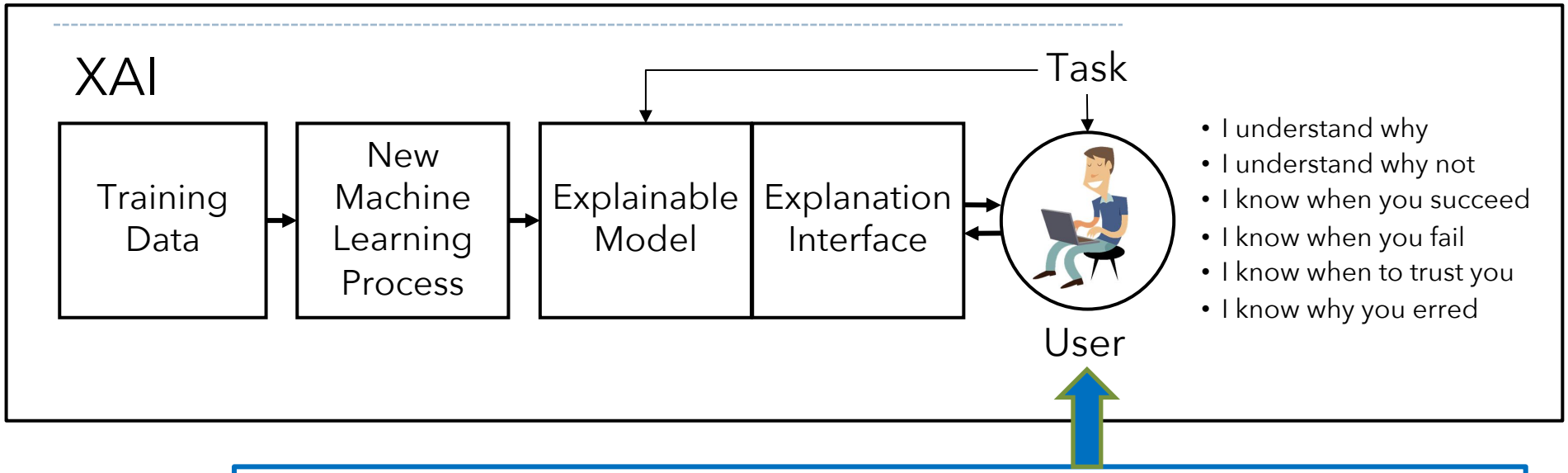
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

XAI



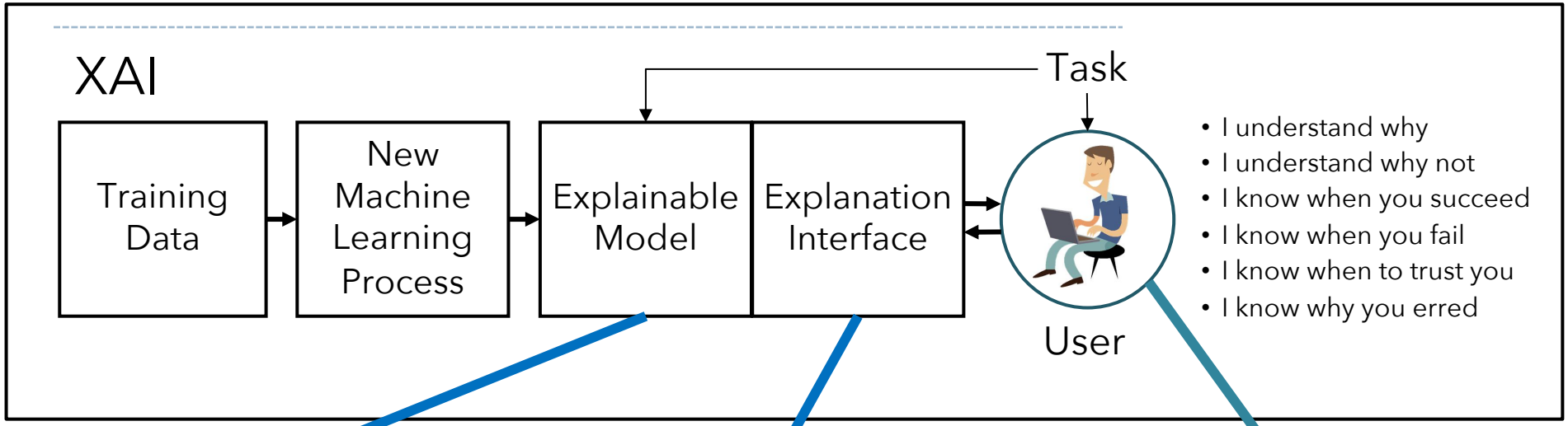
- I understand why
- I understand why not
- I know when you succeed
- I know when you fail
- I know when to trust you
- I know why you erred

B. Program Scope – 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

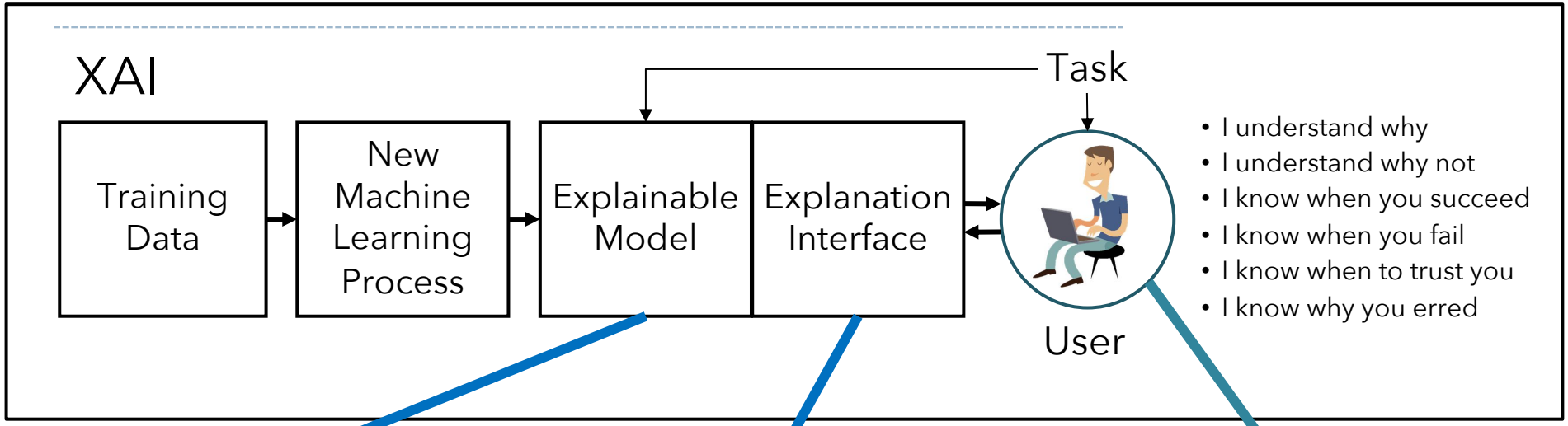
Explanation Interface

- 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

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

TA 1: Explainable Learners

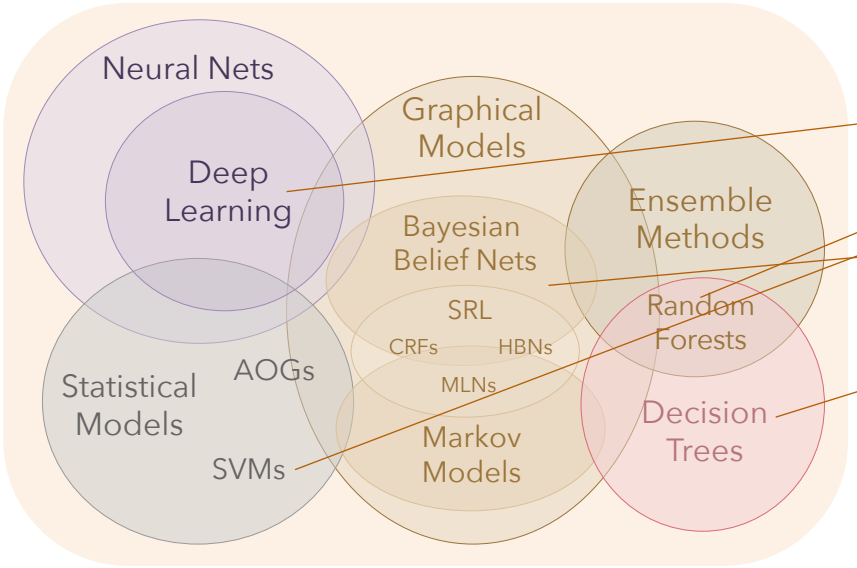
TA 2: Psychological Models

B.1 Explainable Models

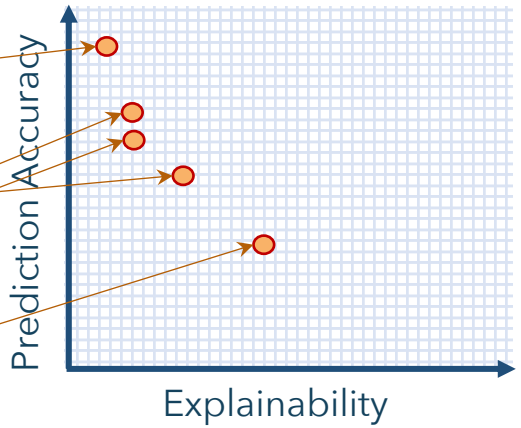
New Approach

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)

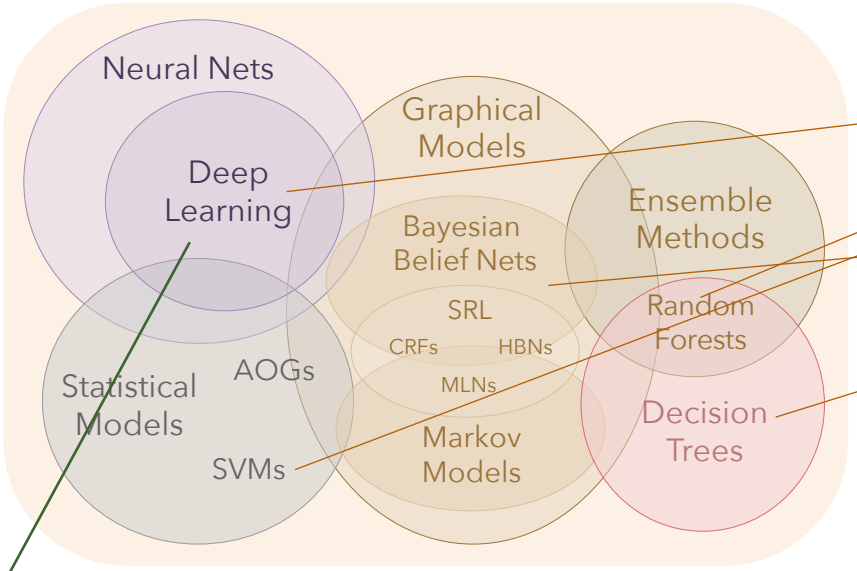


B.1 Explainable Models

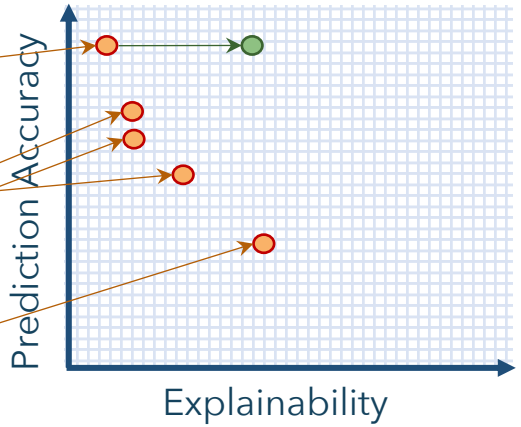
New Approach

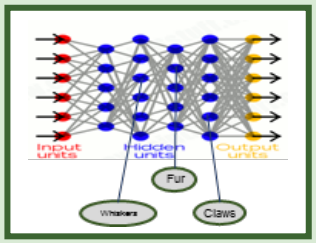
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)





Deep Explanation

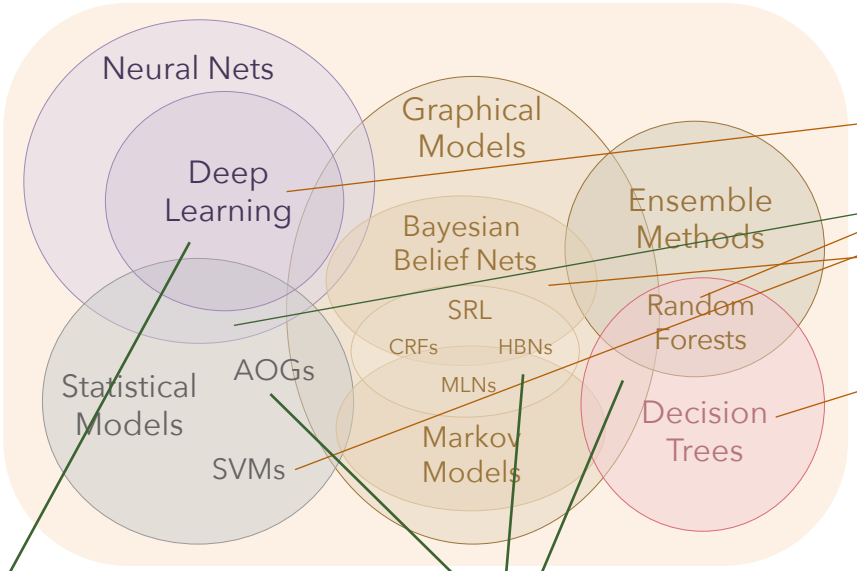
Modified deep learning techniques to learn explainable features

B.1 Explainable Models

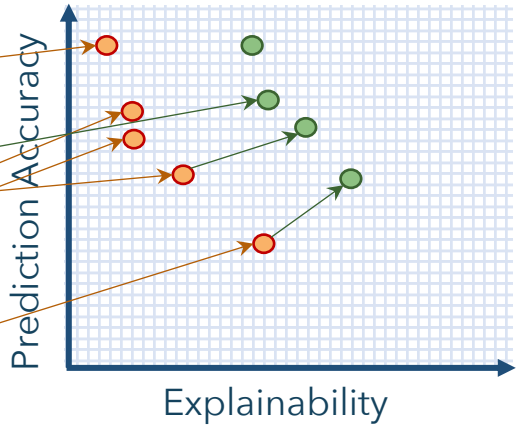
New Approach

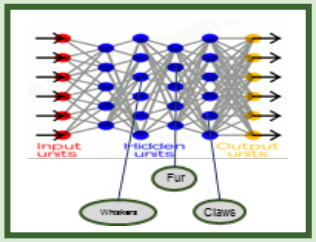
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



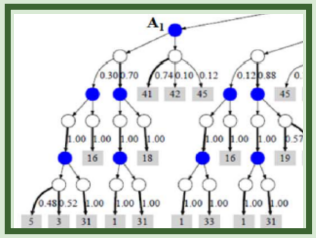
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

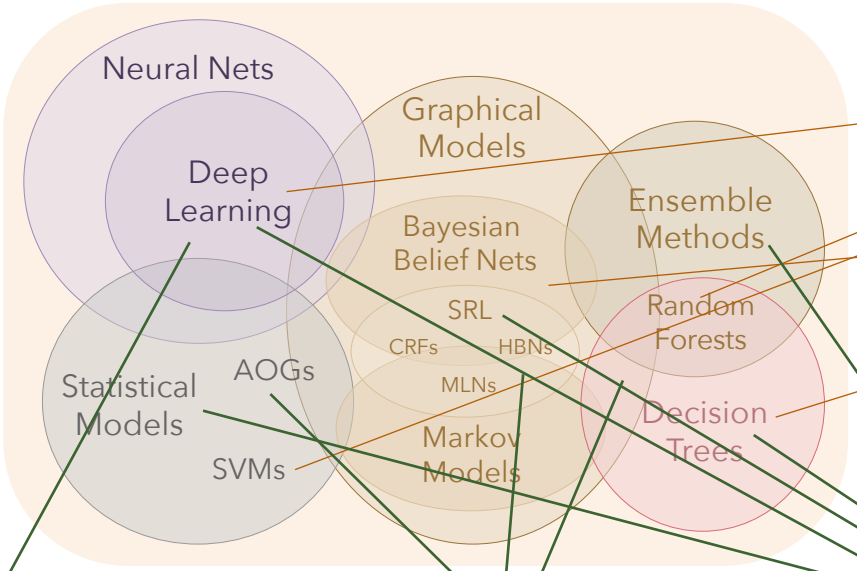
Techniques to learn more structured, interpretable, causal models

B.1 Explainable Models

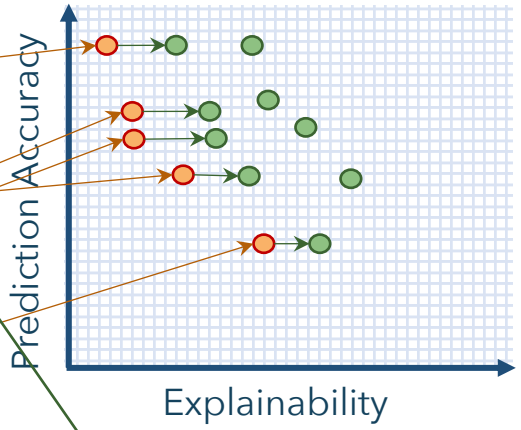
New Approach

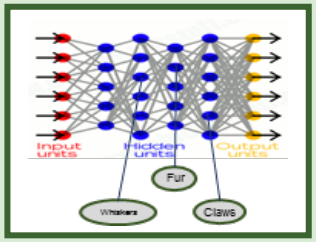
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



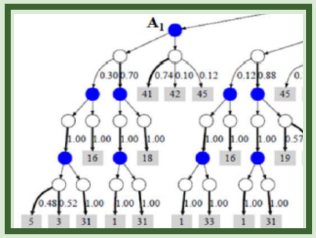
Explainability (notional)





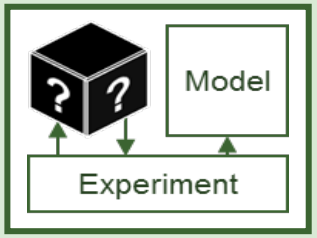
Deep Explanation

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

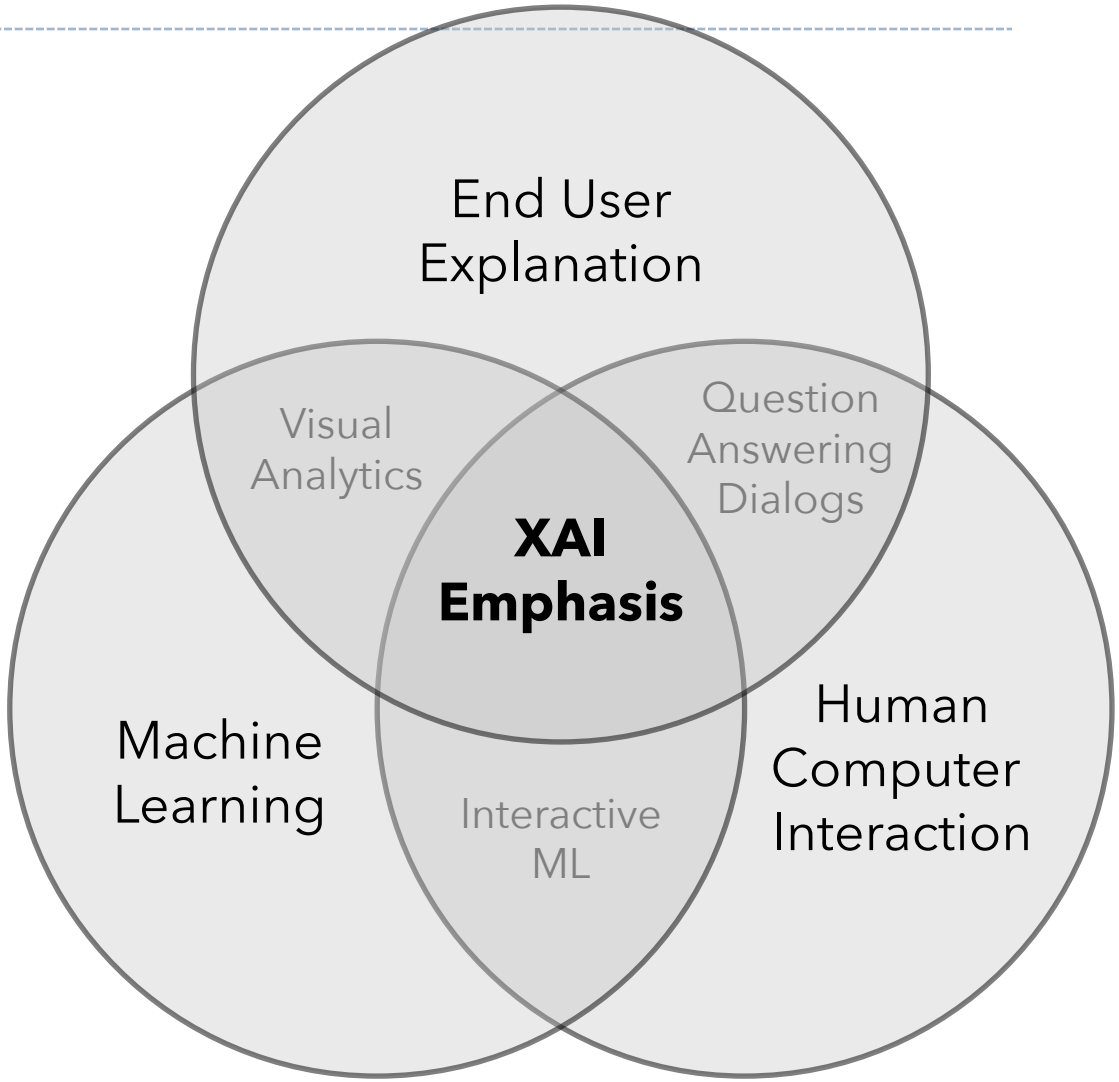
B.2 Explanation Interface

- 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

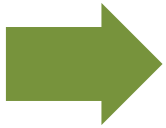
B.3 Psychology of Explanation

- 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

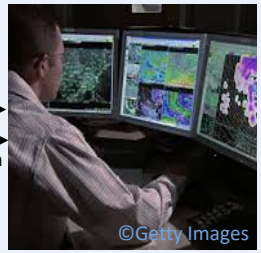
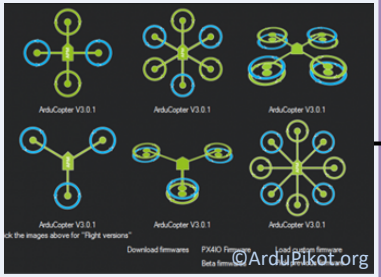


DoD Funding Categories

XAI 

Category	Definition
Basic Research (6.1)	Systematic study directed toward greater knowledge or understanding of the fundamental aspects of phenomena and/or observable facts without specific applications in mind.
Applied Research (6.2)	Systematic study to gain knowledge or understanding necessary to determine the means by which a recognized and specific need may be met.
Technology Development (6.3)	Includes all efforts that have moved into the development and integration of hardware (and software) for field experiments and tests.

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. Challenge Problems and Evaluation

- Developers propose their own Phase 1 problems
  - Within one or both of the two general categories (Data Analytics and Autonomy)
- During Phase 1, the XAI evaluator will work with developers
  - Define a set of common test problems in each category
  - Define a set of metrics and evaluation methods
- During Phase 2, the XAI developers will demonstrate their XAI systems against the common test problems defined by the XAI evaluator
- Proposers should suggest creative and compelling test problems
  - Productive drivers of XAI research and development
  - Sufficiently general and compelling to be useful for multiple XAI approaches
  - Avoid unique, tailored problems for each research project
  - Consider problems that might be extended to become an open, international competition

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

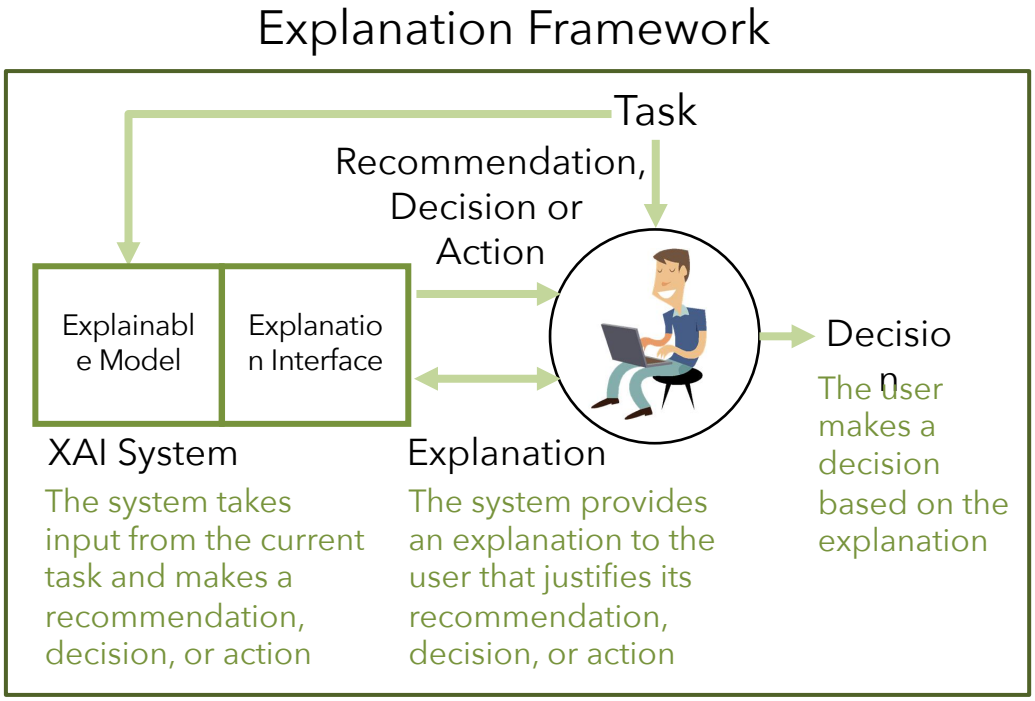
C.2 Autonomy

- 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

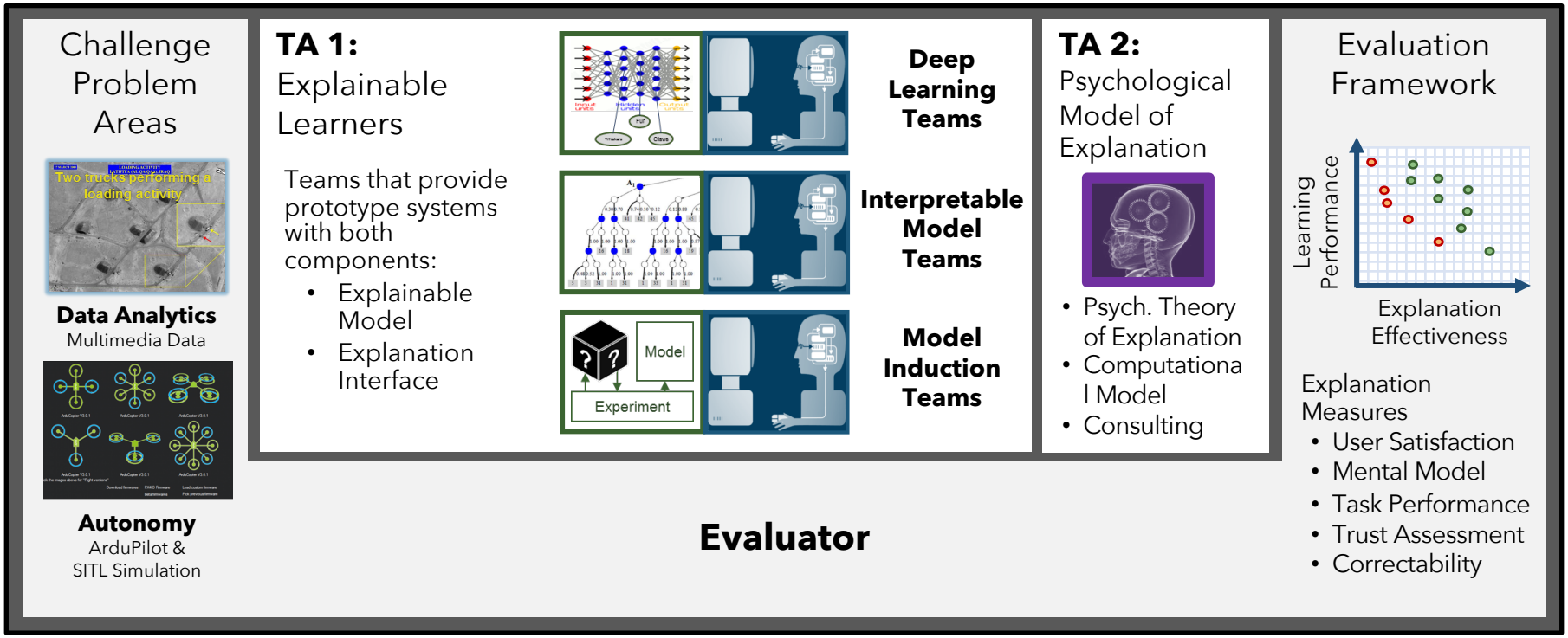
- XAI developers are presented with a problem domain
- Apply machine learning techniques to learn an explainable model
- Combine with the explanation interface to construct an explainable system
- The explainable system delivers and explains decisions or actions to a user who is performing domain tasks
- 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

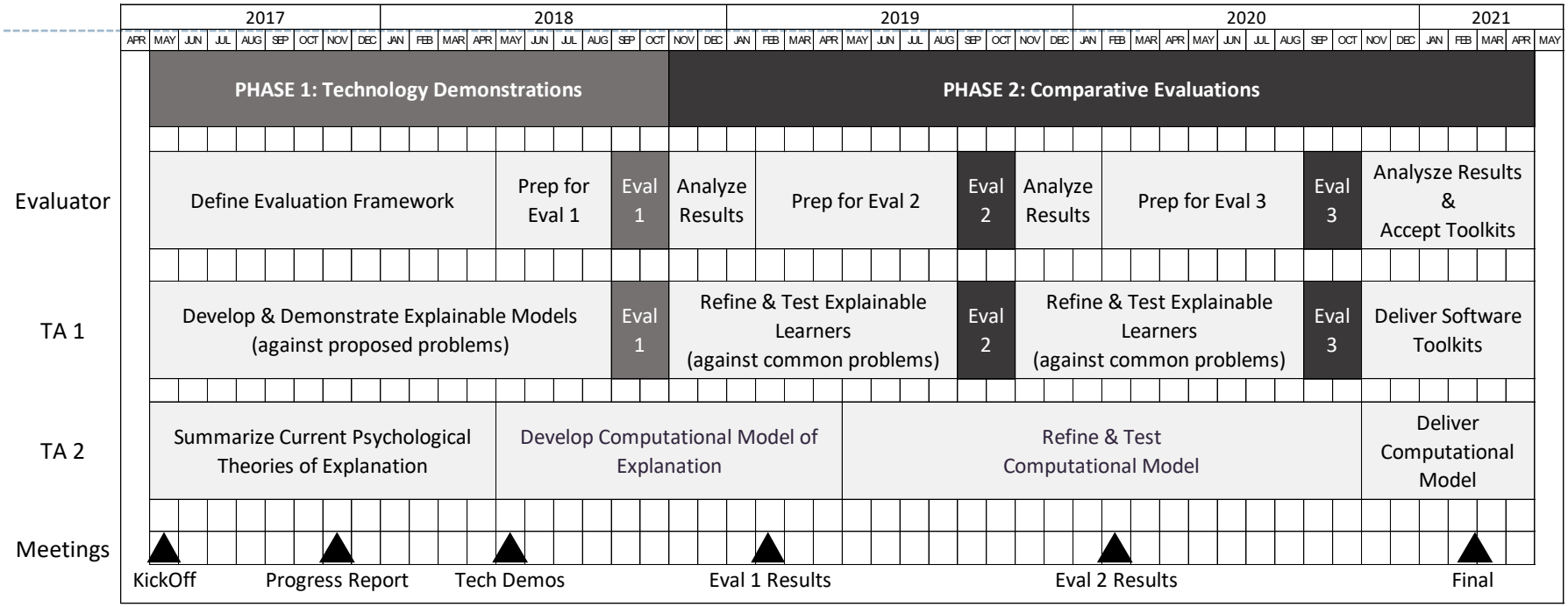
D.1 Technical Area 1 – Explainable Learners

- 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

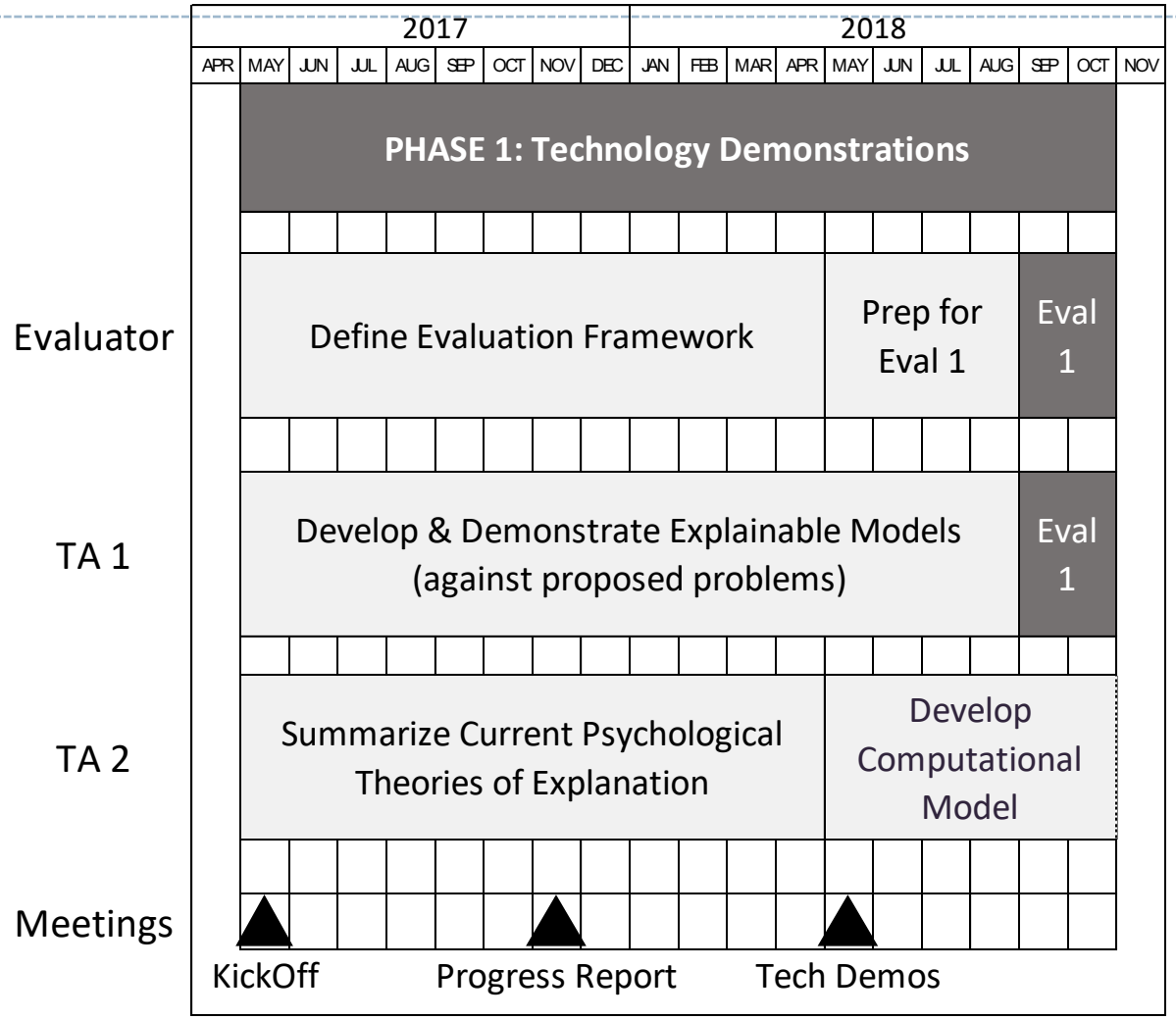
- Theories of Explanation
  - Describe how you will summarize the current psychological theories of explanation
  - Describe how this work will inform the development of the TA1 XAI systems
  - Describe how this work will inform the definition of the evaluation framework for measuring explanation effectiveness 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

E. Schedule and Milestones

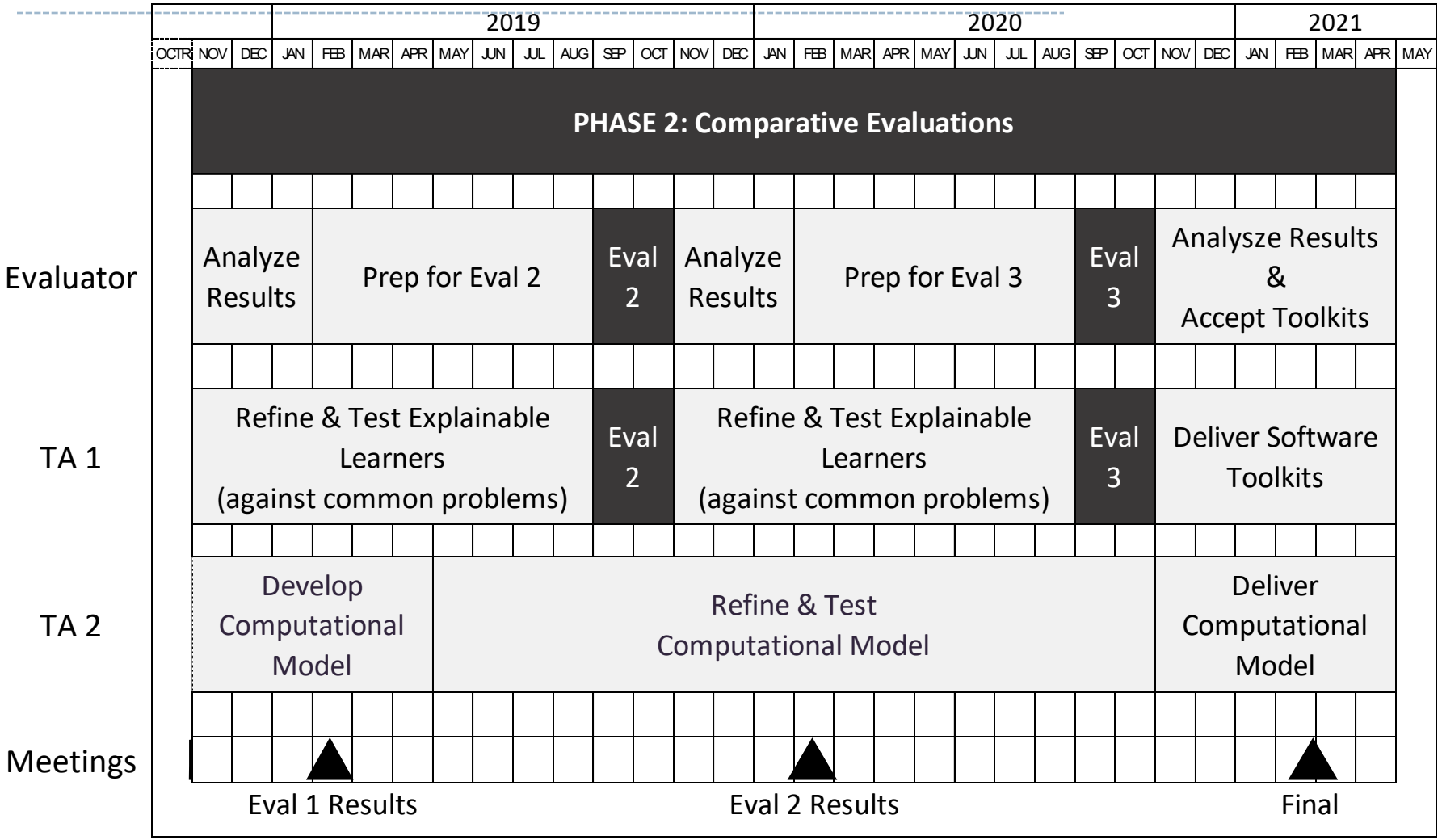


- Technical Area 1 Milestones:
  - Demonstrate the explainable learners against problems proposed by the developers (Phase 1)
  - Demonstrate the explainable learners against common problems (Phase 2)
  - Deliver software libraries and toolkits (at the end of Phase 2)
- Technical Area 2 Milestones:
  - 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)

E. Schedule and Milestones – Phase 1



E. Schedule and Milestones – Phase 2



Summary

- Goal: to create a suite of new or modified machine learning techniques
  - to produce explainable models that
  - when combined with effective explanation techniques
  - enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems
- XAI is seeking the most interesting and compelling ideas to accomplish this goal

Next Lecture: Part II
How to Achieve Interpretability
Measures of Interpretability

Part II: Explainable AI Requirements

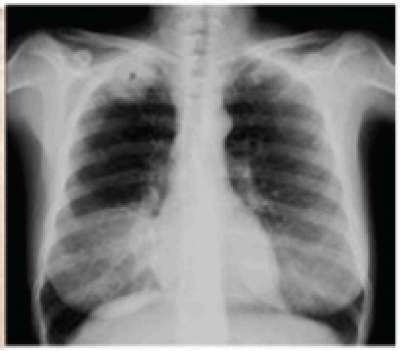
Noticeable Performance of AI & ML in recent years

Game GO

Traffic Sign Recognition

Skin cancer detection

Lung cancer detection

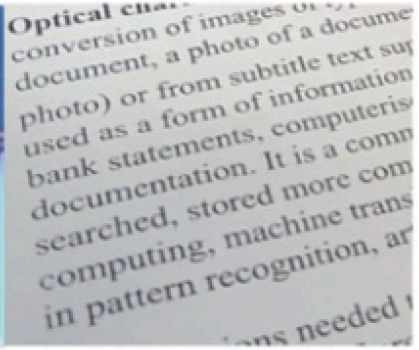


Poker

Computer games

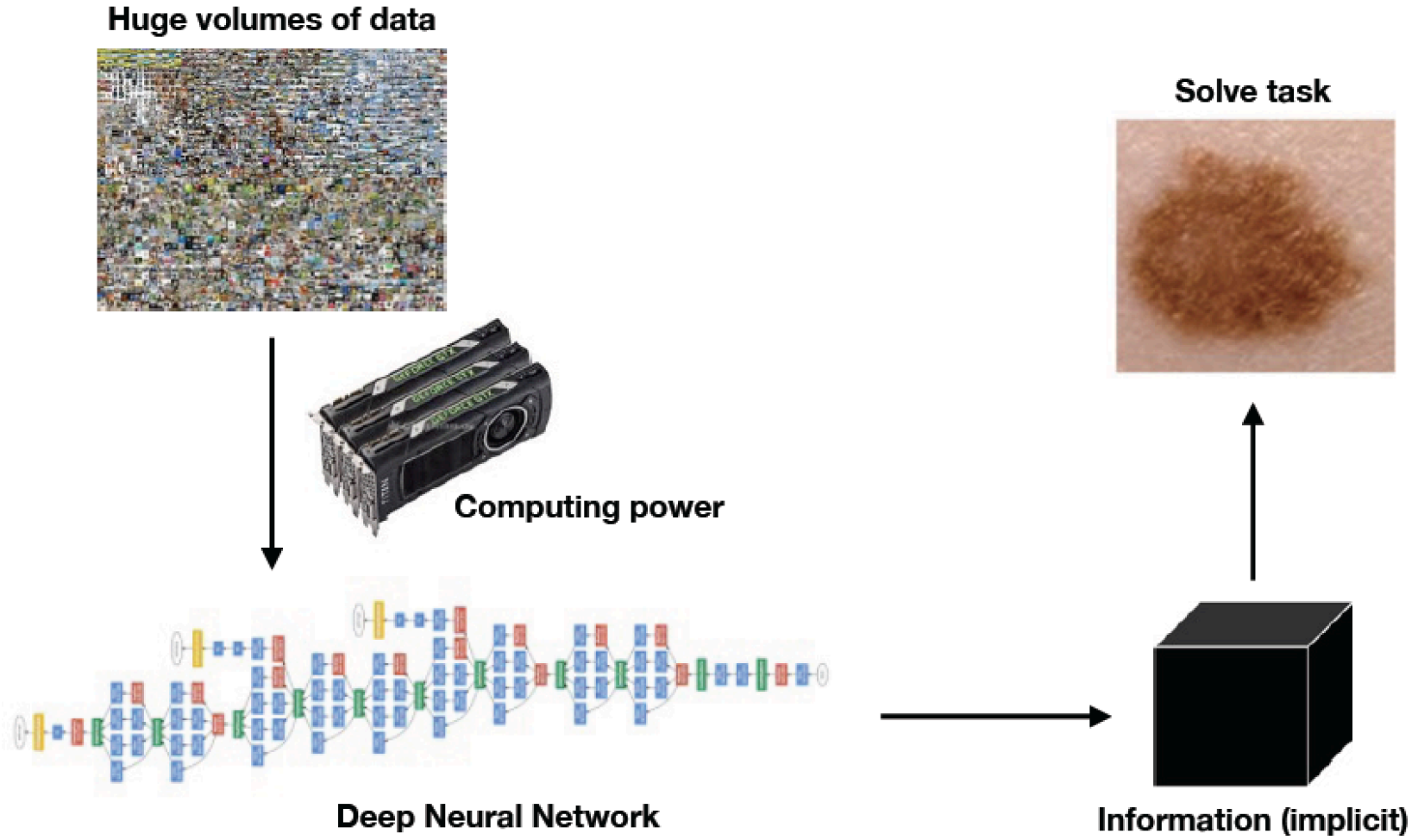
Jeopardy

OCR



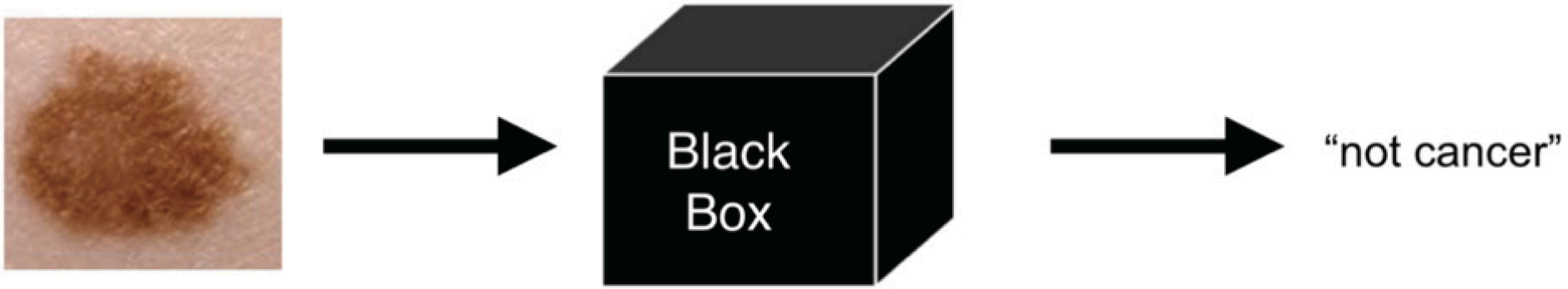
Part II: Explainable AI Requirements

Black Box Models not Sufficient for XAI



Part II: Explainable AI Requirements

Black Box Models not Sufficient for XAI



Is minimizing the error a guarantee for the model to work well in practice?

Part II: Explainable AI Requirements

Why XAI?

We need interpretability in order to:

verify system

legal aspects

learn new things from data

understand weaknesses

Part II: Explainable AI Requirements

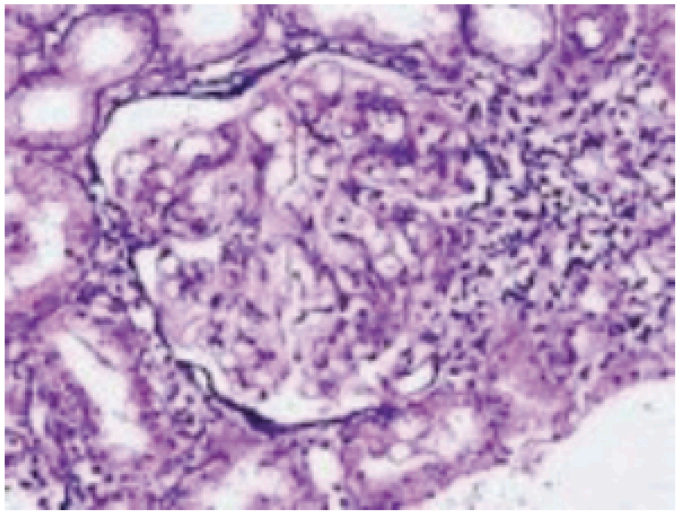
Why XAI?

1) Verify that classifier works as expected

Wrong decisions can be costly and dangerous

"Autonomous car crashes, because it wrongly recognizes …"

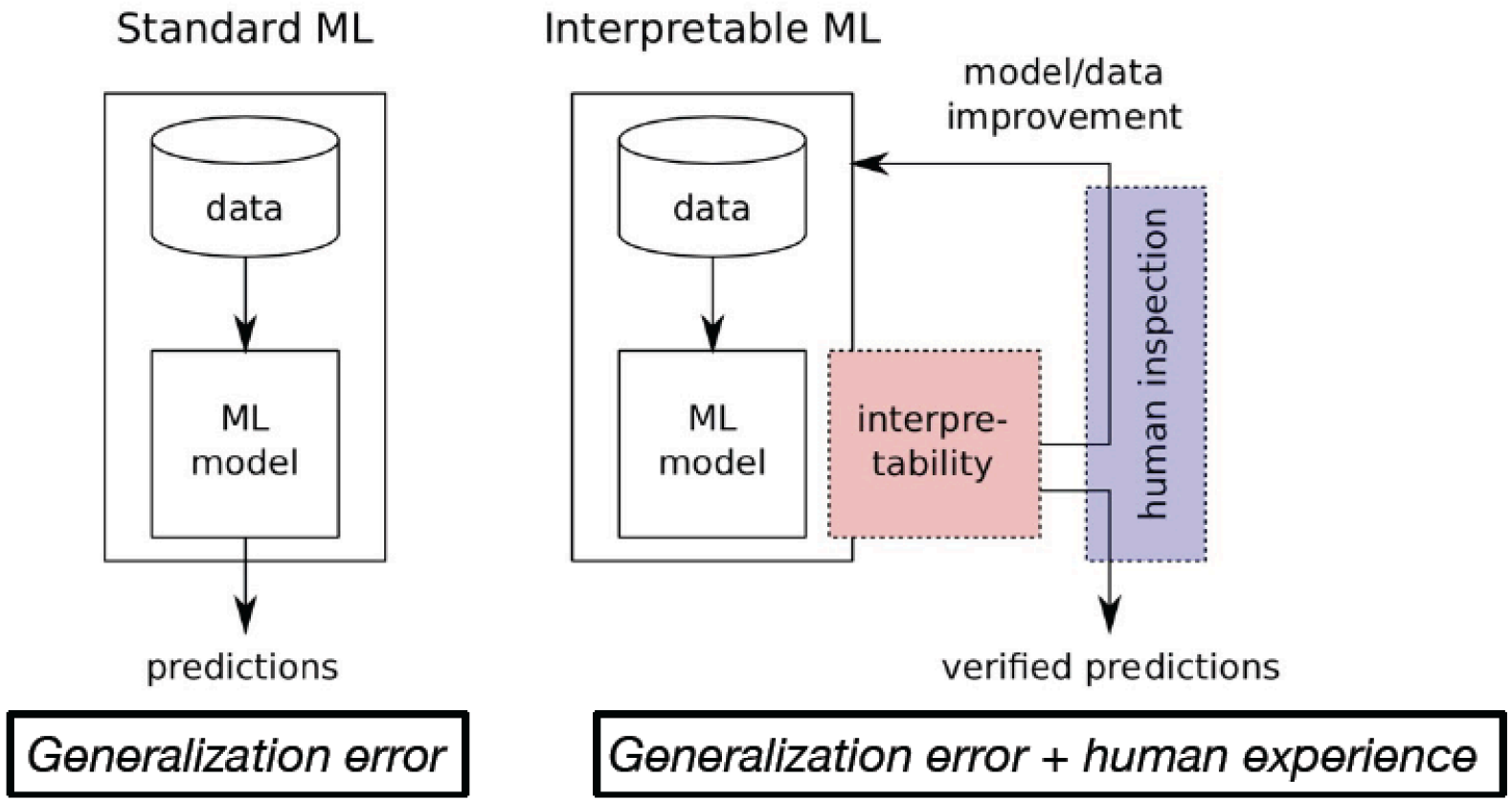
"AI medical diagnosis system misclassifies patient's disease …"



Part II: Explainable AI Requirements

Why XAI?

2) Understand weaknesses & improve classifier



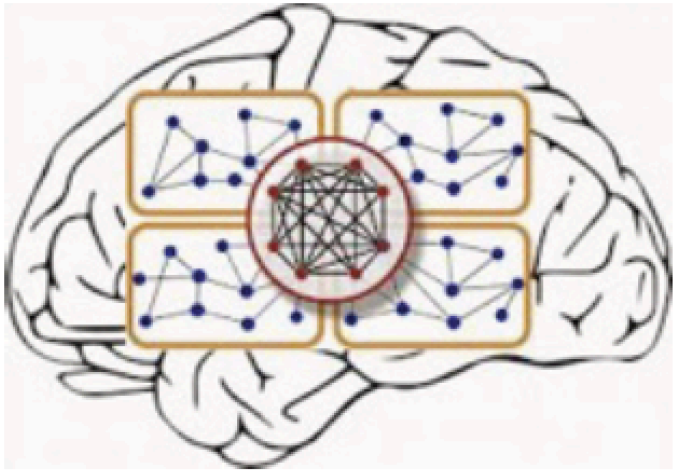
Part II: Explainable AI Requirements

Why XAI?

3) Learn new things from the learning machine

"It's not a human move. I've never seen a human play this move." (Fan Hui)

Old promise:
"Learn about the human brain."

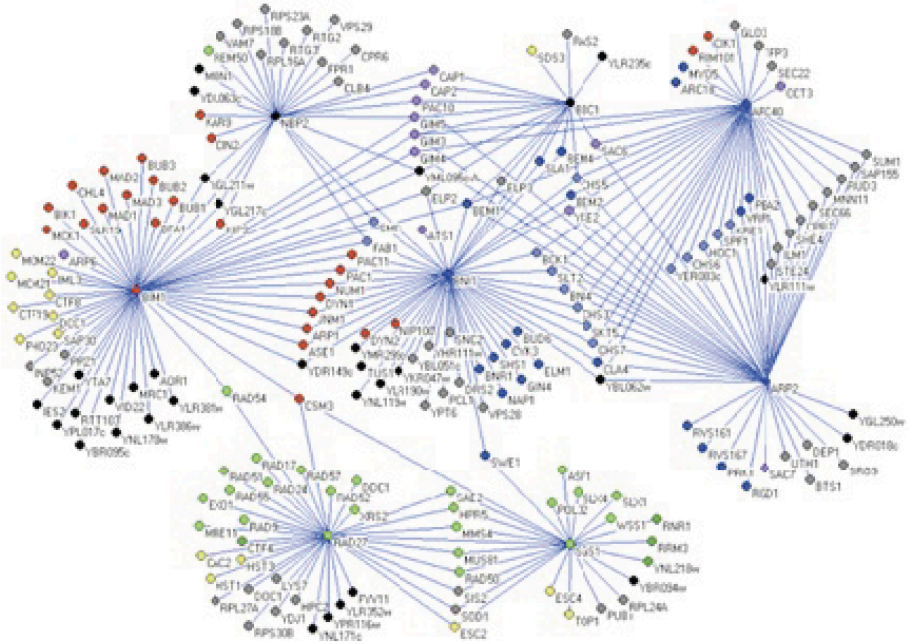


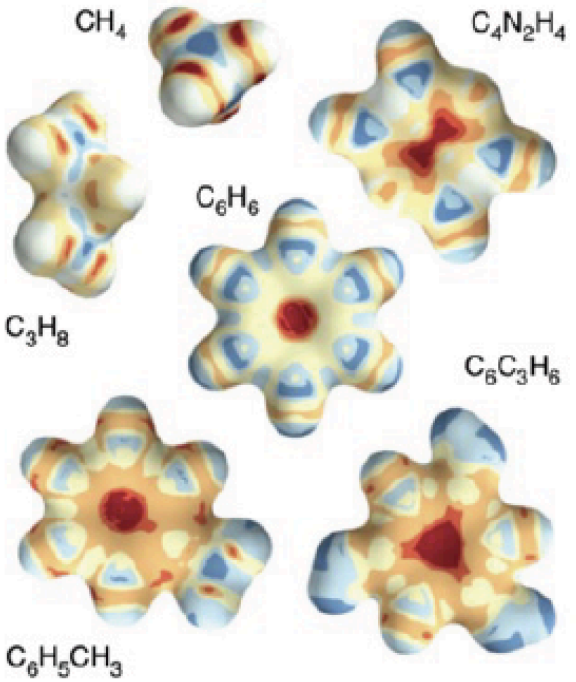
Part II: Explainable AI Requirements

Why XAI?

4) Interpretability in the sciences

Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites …)





Part II: Explainable AI Requirements

Why XAI?

5) Compliance to legislation

European Union's new General Data Protection Regulation → "right to explanation"

Retain human decision in order to assign responsibility.

"With interpretability we can ensure that ML models work in compliance to proposed legislation."

Part II: Explainable AI Requirements

ITU/WHO Focus Group on AI4Health

Focus Group on "Artificial Intelligence for Health" established by

ITU Workshop on Artificial Intelligence for Health
Geneva, Switzerland, 25 September 2018

More information about the group:

`https://www.itu.int/en/ITU-T/focusgroups/ai4h`

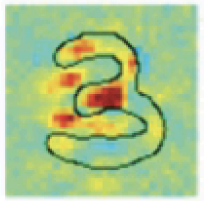
Part II: Explainable AI Requirements

Dimensions of Interpretability

Different dimensions of "interpretability"

prediction

"Explain why a certain pattern x has been classified in a certain way f(x)."



model

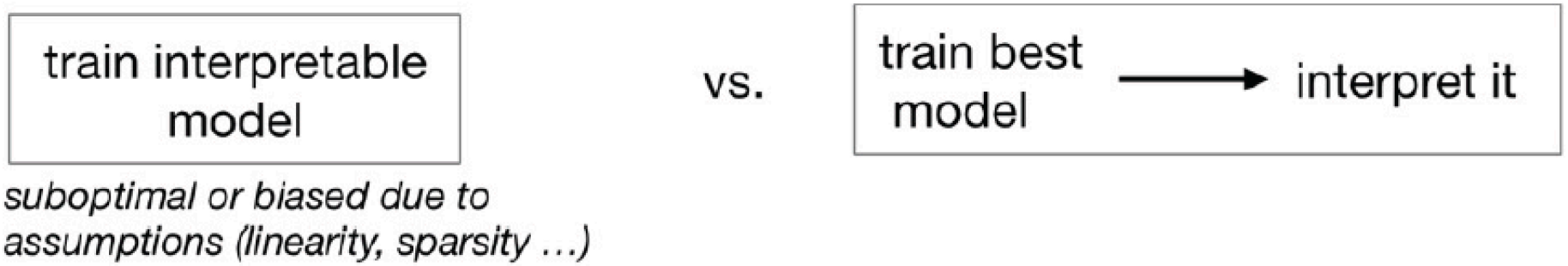
"What would a pattern belonging to a certain category typically look like according to the model."

data

"Which dimensions of the data are most relevant for the task."

Part II: Explainable AI Requirements

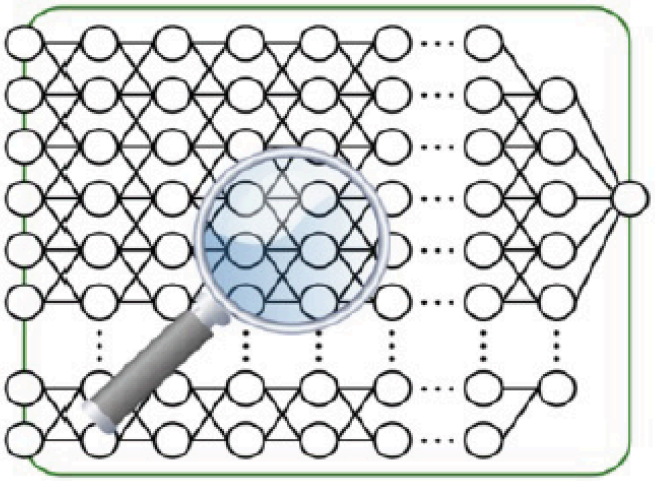
Dimensions of Interpretability



Part II: Explainable AI Requirements

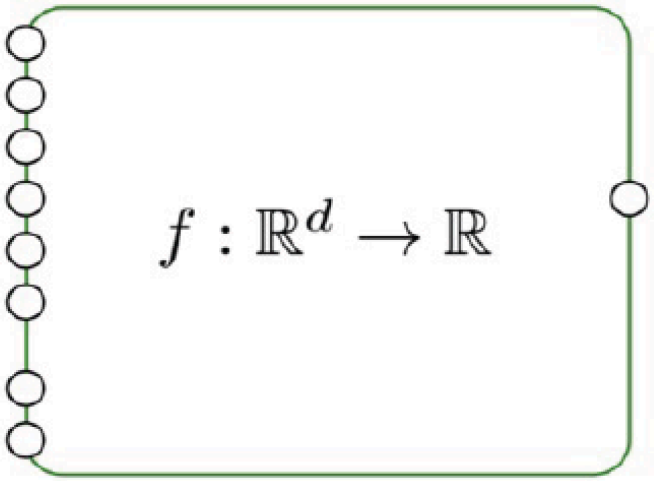
Techniques of Interpretability

mechanistic understanding



Understanding what mechanism the network uses to solve a problem or implement a function.

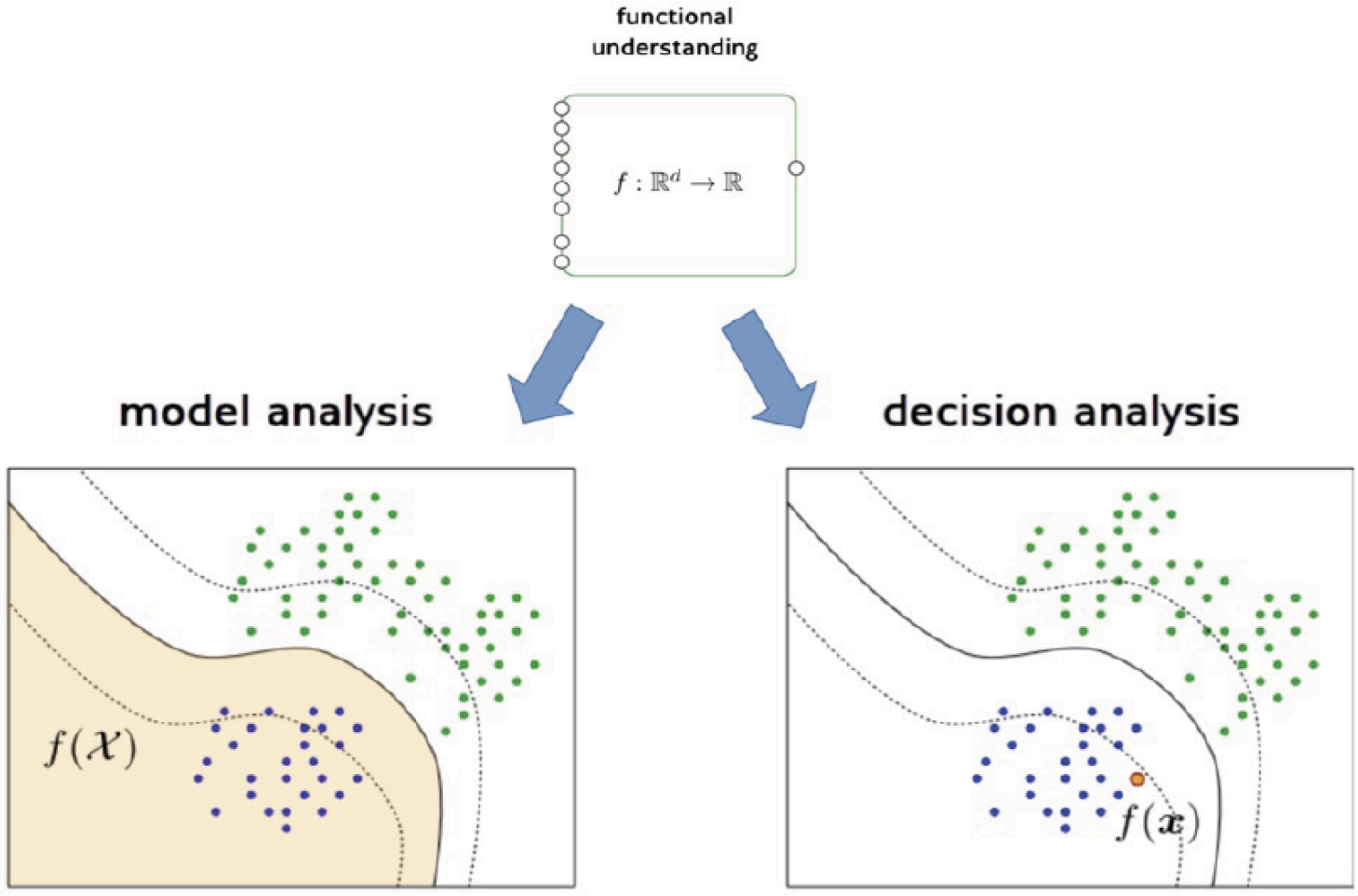
functional understanding



Understanding how the network relates the input to the output variables.

Part II: Explainable AI Requirements

Techniques of Interpretability

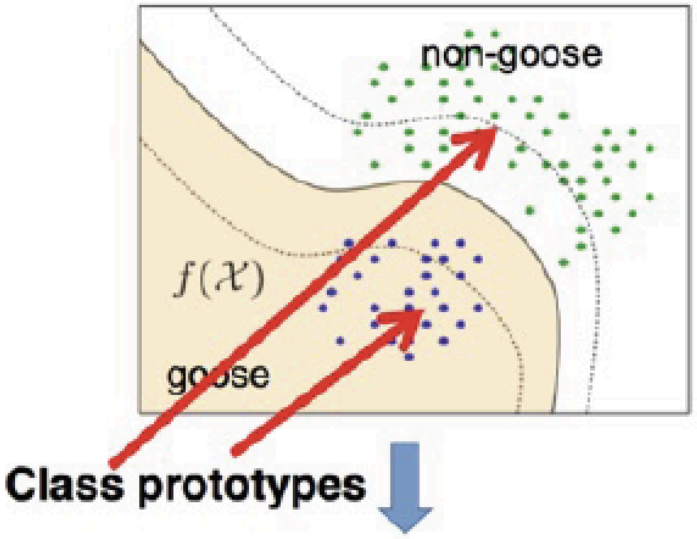


Model Interpretability

Interpreting the Model

Approach 1: Class Prototypes

"How does a goose typically look like according to the neural network?"

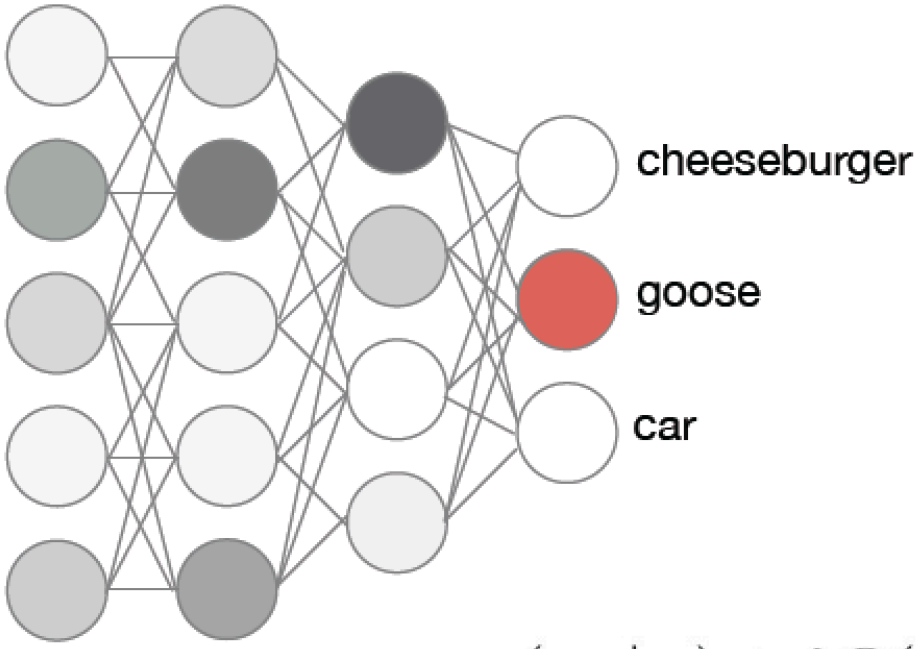


$$\arg\max_x f(x) + \text{reg.}$$

Interpreting the Model

Activation Maximization

- *find prototypical example of a category*
- *find pattern maximizing activity of a neuron*



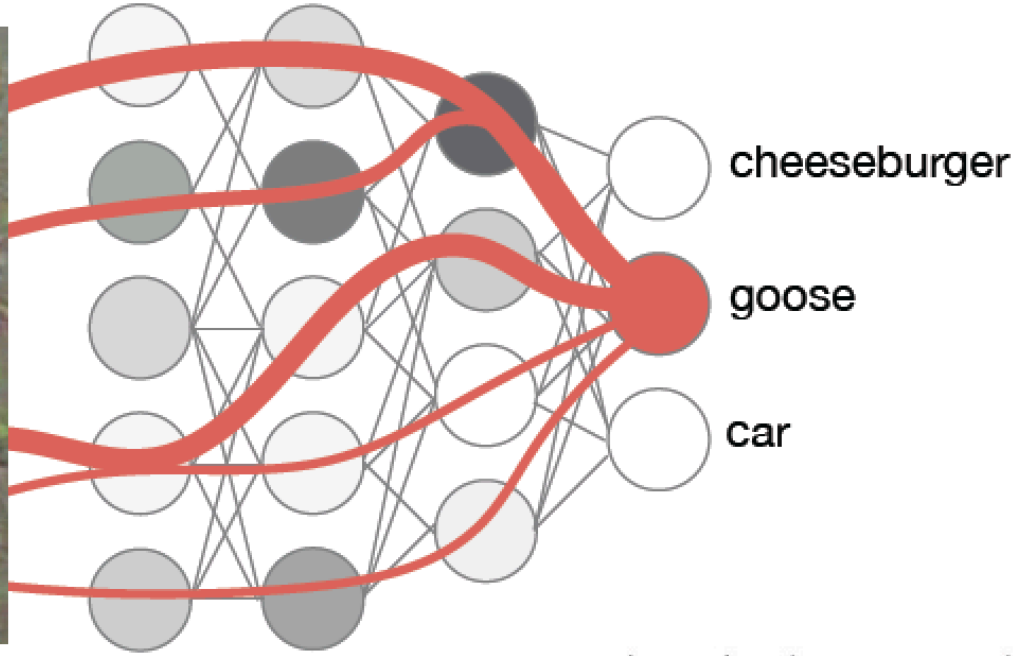
$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)$$

Interpreting the Model

Activation Maximization

- *find prototypical example of a category*
- *find pattern maximizing activity of a neuron*

simple regularizer
(Simonyan et al. 2013)

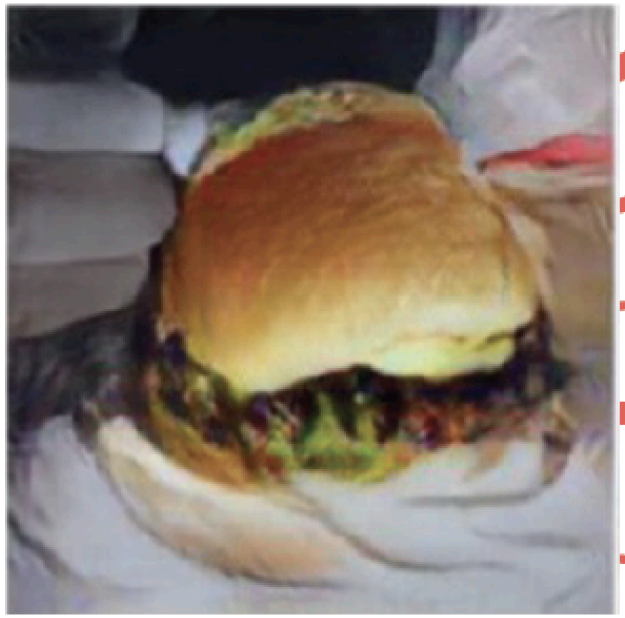


$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)$$

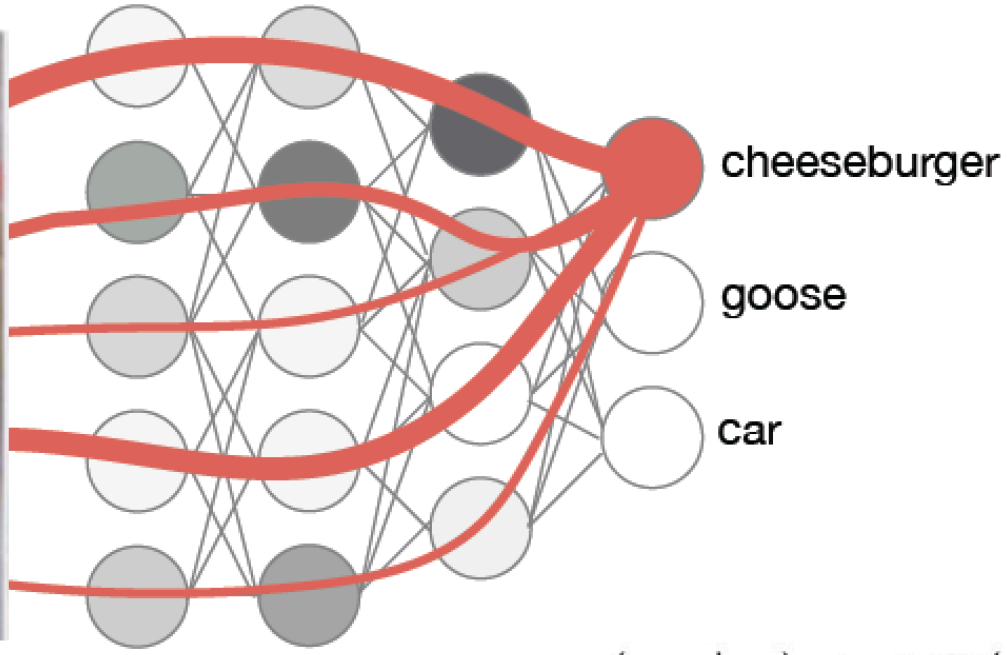
Interpreting the Model

Activation Maximization

- *find prototypical example of a category*
- *find pattern maximizing activity of a neuron*



complex regularizer
(Nguyen et al. 2016)

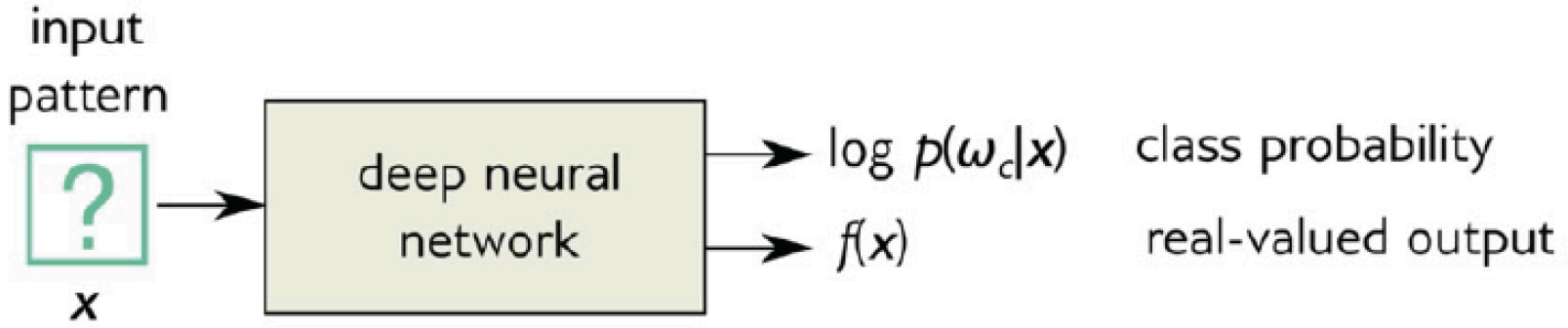


$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)$$

Interpreting the Model

Activation Maximization

Let us interpret a concept predicted by a deep neural net (e.g. a class, or a real-valued quantity):



Examples:

- Creating a class prototype: $\max_{x \in \mathcal{X}} \log p(\omega_c|x)$.
- Synthesizing an extreme case: $\max_{x \in \mathcal{X}} f(x)$.

Interpreting the Model

goose

ostrich

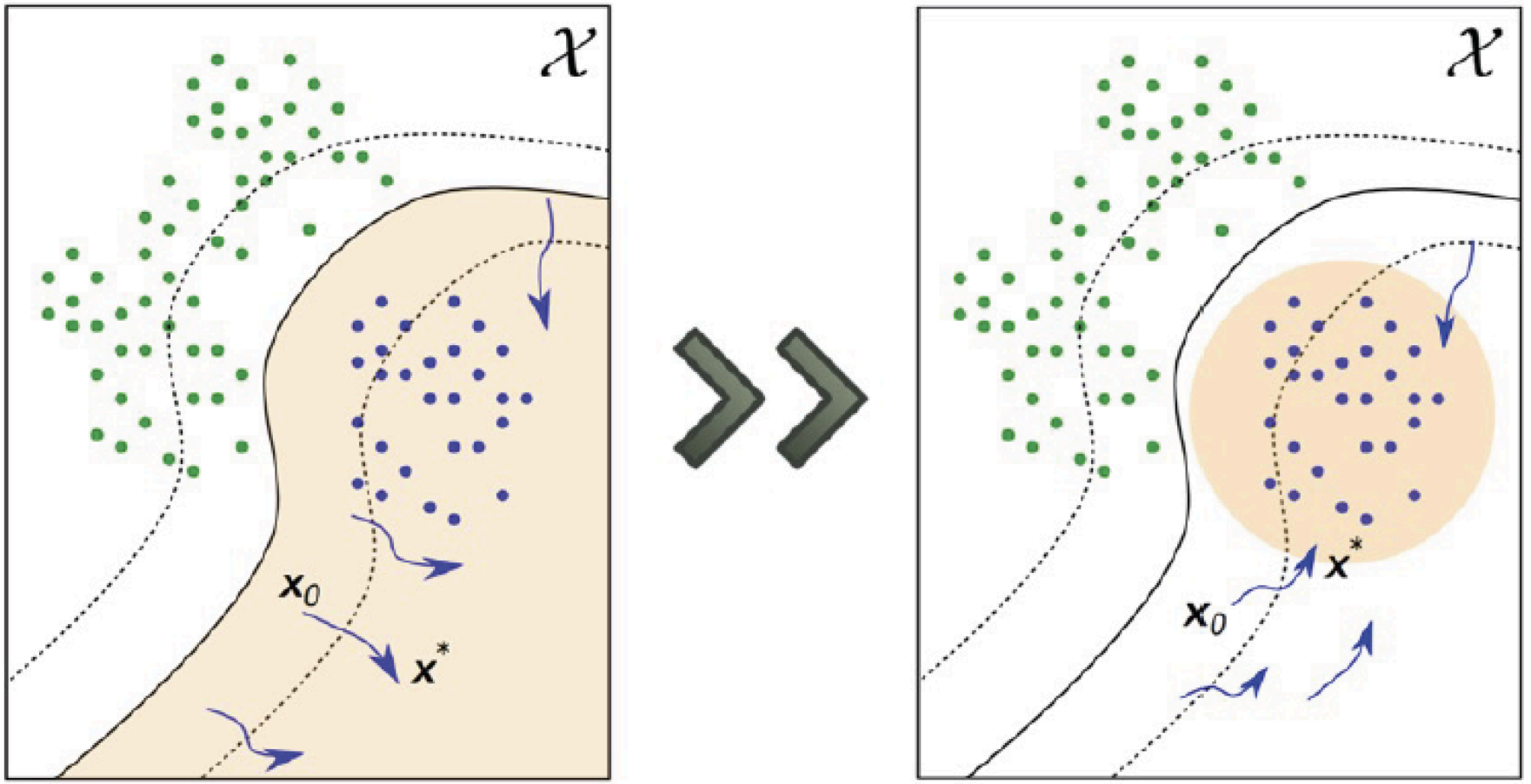
Images from **Simonyan et al. 2013** "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps"

Observations:

- AM builds typical patterns for these classes (e.g. beaks, legs).
- Unrelated background objects are not present in the image.

Enhancing Activation Maximization

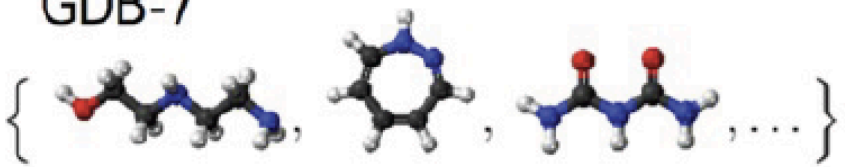
Find the input pattern that maximizes class probability. → Find the most likely input pattern for a given class.

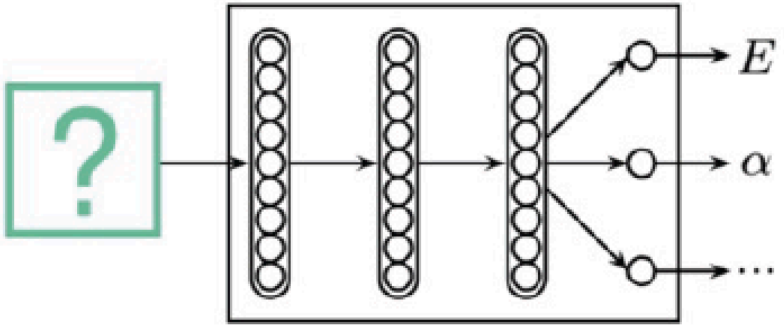


Another Application

Finding a prototype:

GDB-7





Question: How does a molecule with properties XYZ look like ?

Limitations of Global Interpretations

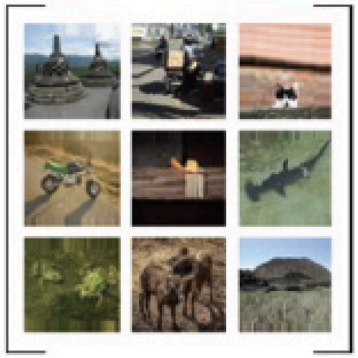
Question: Below are some images of motorbikes. What would be the best prototype to interpret the class "motorbike"?

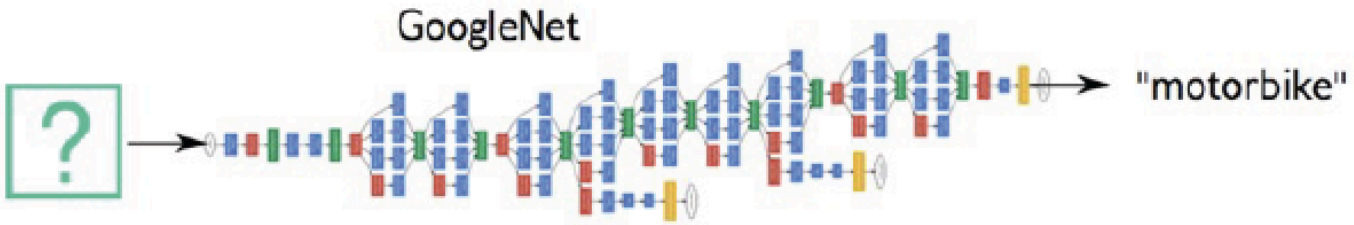
Observations:

- Summarizing a concept or category like "motorbike" into a single image can be difficult (e.g. different views or colors).
- A good interpretation would grow as large as the diversity of the concept to interpret.

Need Individual Explanation

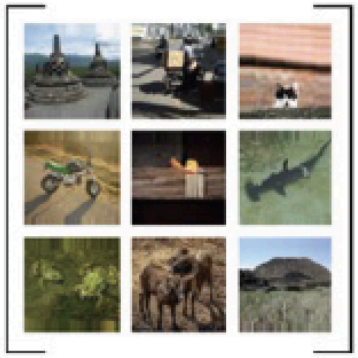
Finding a prototype:

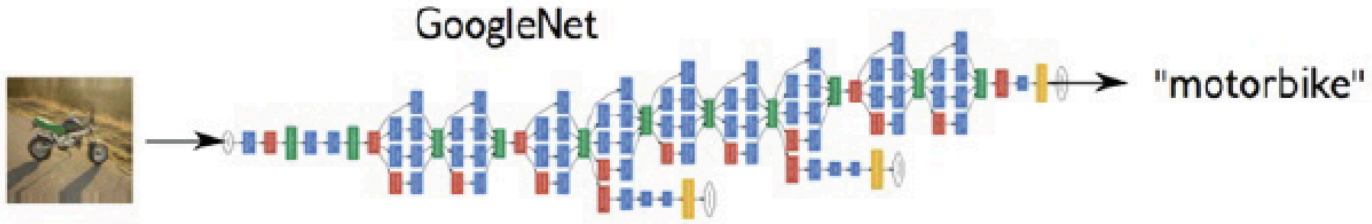




Question: How does a "motorbike" typically look like?

Individual explanation:





Question: Why is *this* example classified as a motorbike?

Need Individual Explanation

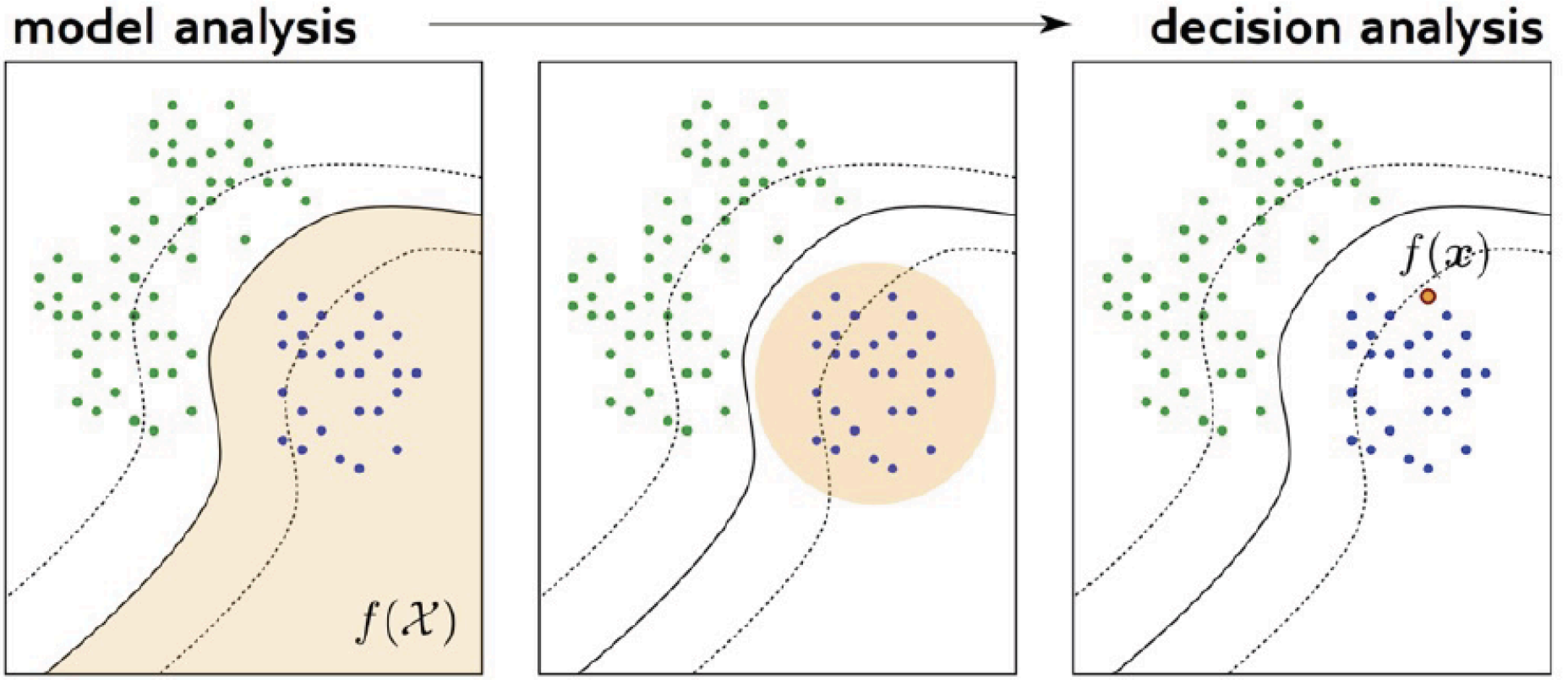
Personalized medicine: Extracting the relevant information about a medical condition for a *given* patient at a *given* time.

Each case is unique and needs its own explanation.

Population view: Which symptoms are most common for the disease

Both aspects can be important depending on who you are (FDA, doctor, patient).

Making Deep Networks Transparent



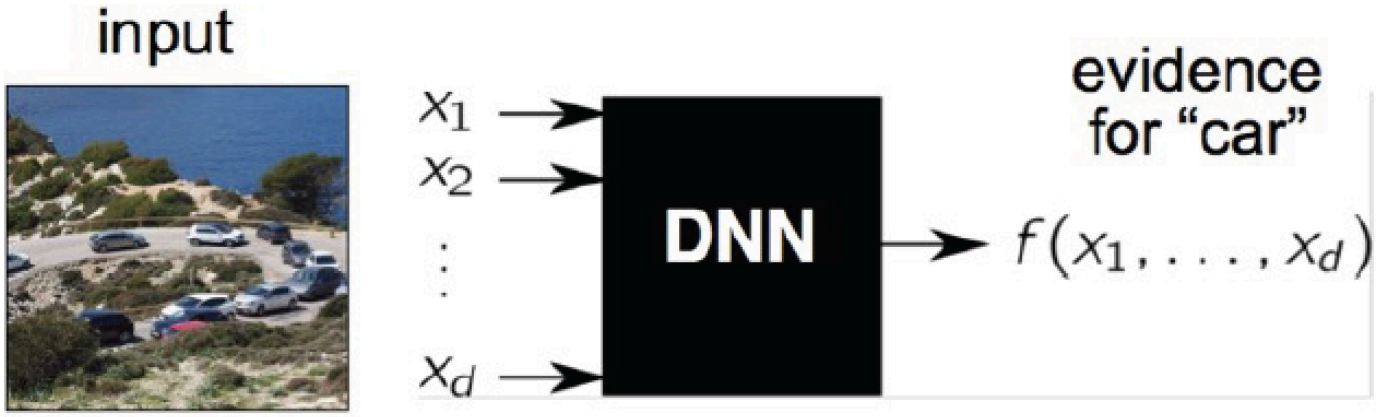
- visualizing filters
- max. class activation

- include distribution (RBM, DGN, etc.)

- sensitivity analysis
- decomposition

Decision Analysis

Decision Analysis: Sensitivity Analysis



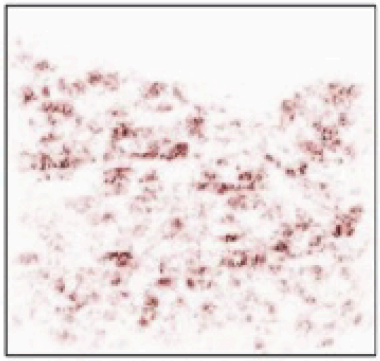
Sensitivity analysis: The relevance of input feature *i* is given by the squared partial derivative:

$$R_i = \left(\frac{\partial f}{\partial x_i}\right)^2$$

Decision Analysis: Sensitivity Analysis

Sensitivity analysis:

$$R_i = \left(\frac{\partial f}{\partial x_i}\right)^2$$



Problem: sensitivity analysis does not highlight cars

highlights parts, which (when changed) increase or decrease the prediction for "car".

Observation:

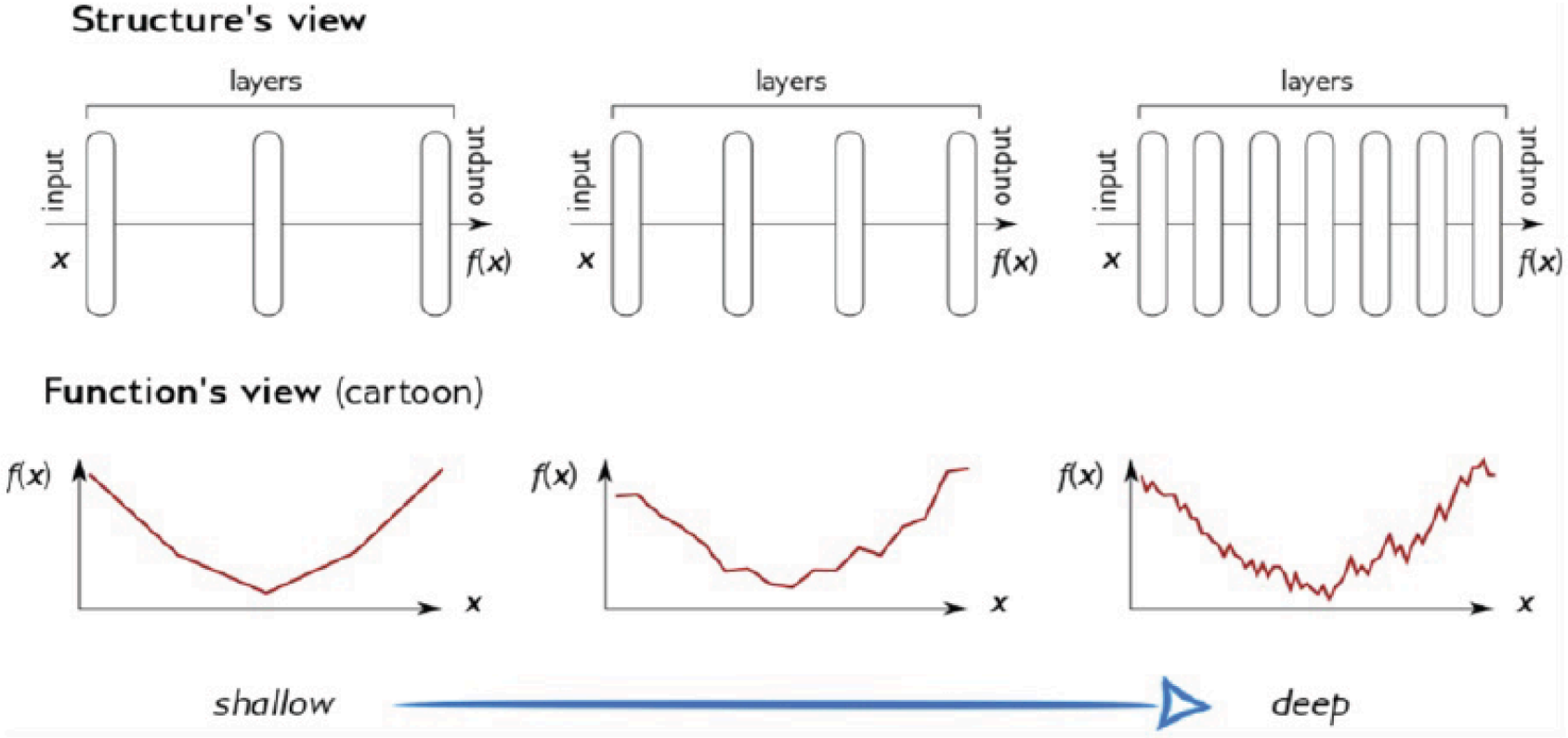
$$\sum_{i=1}^{d} \left(\frac{\partial f}{\partial x_i}\right)^2 = \|\nabla_x f\|^2$$

Sensitivity analysis explains a *variation* of the function, not the function value itself.

Decision Analysis: Sensitivity Analysis

Shattered Gradient Problem

Input gradient (on which sensitivity analysis is based), becomes increasingly highly varying and unreliable with neural network depth.



Layer-wise Relevance Propagation (LRP)

Decision Analysis: LRP

input x

Black Box

Rooster

prediction $f(x)$

Idea: Decompose function

$$\sum_i R_i = f(x)$$

heatmap

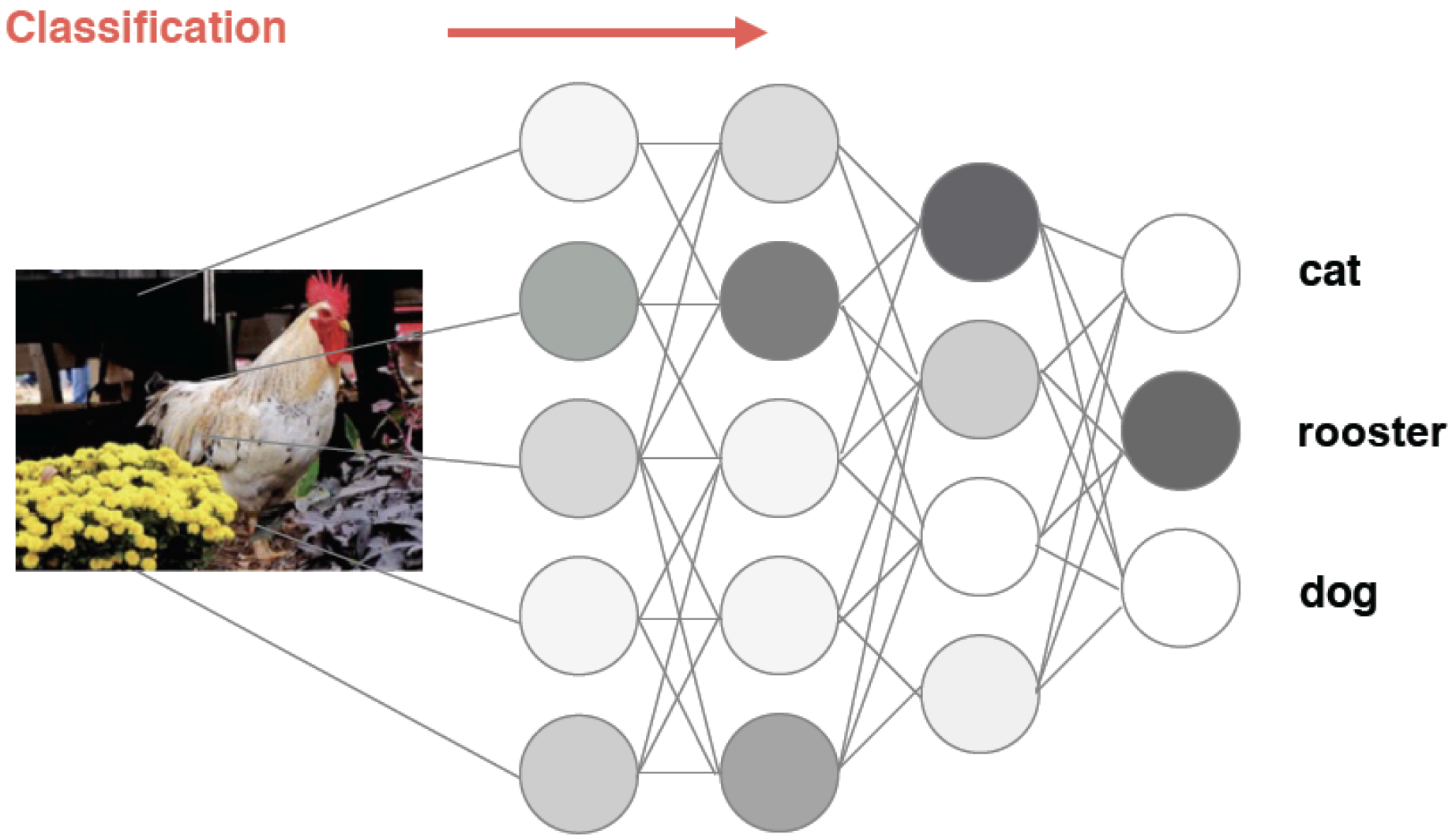
Explain prediction

(how much each pixel contributes to prediction)

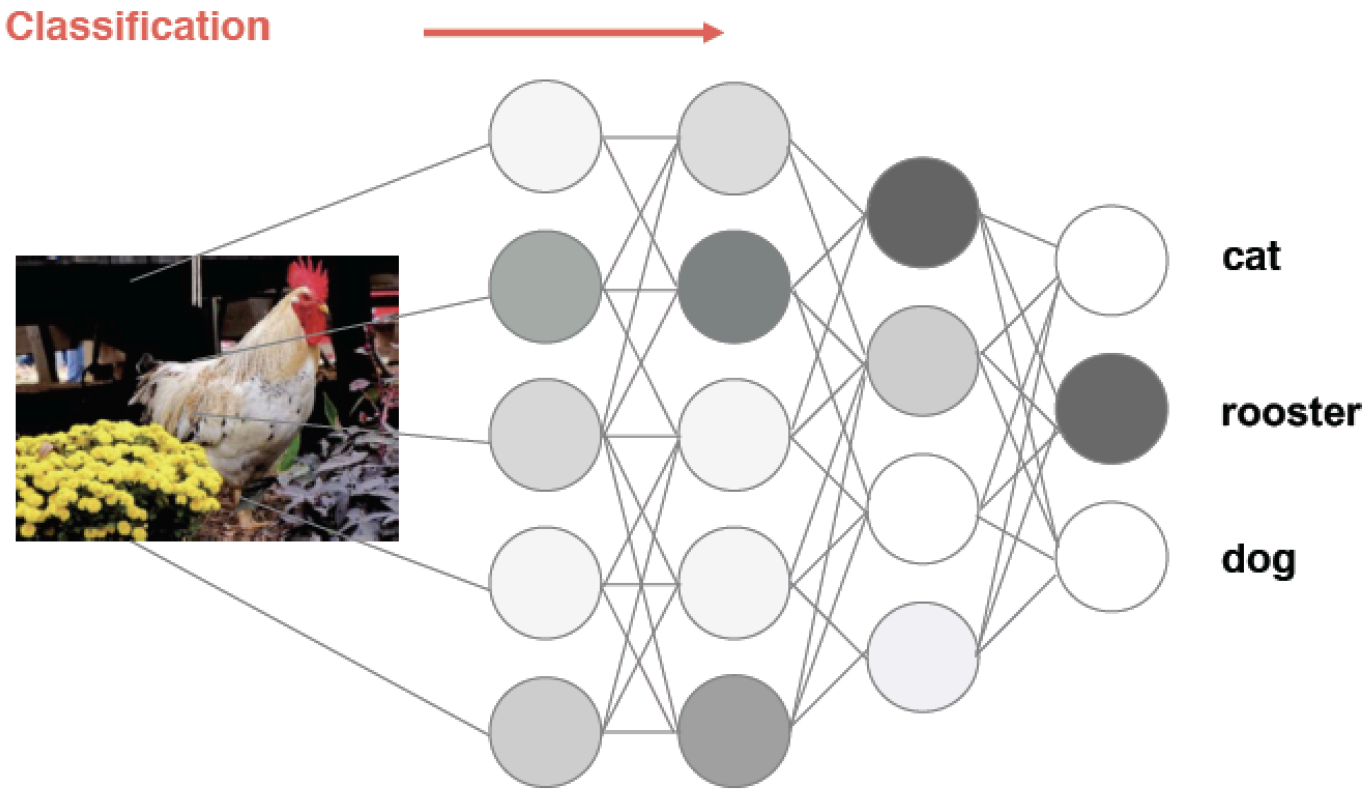
Layer-wise Relevance Propagation (LRP)
(Bach et al., PLOS ONE, 2015)

Explain prediction itself
(not the change)

Decision Analysis: LRP



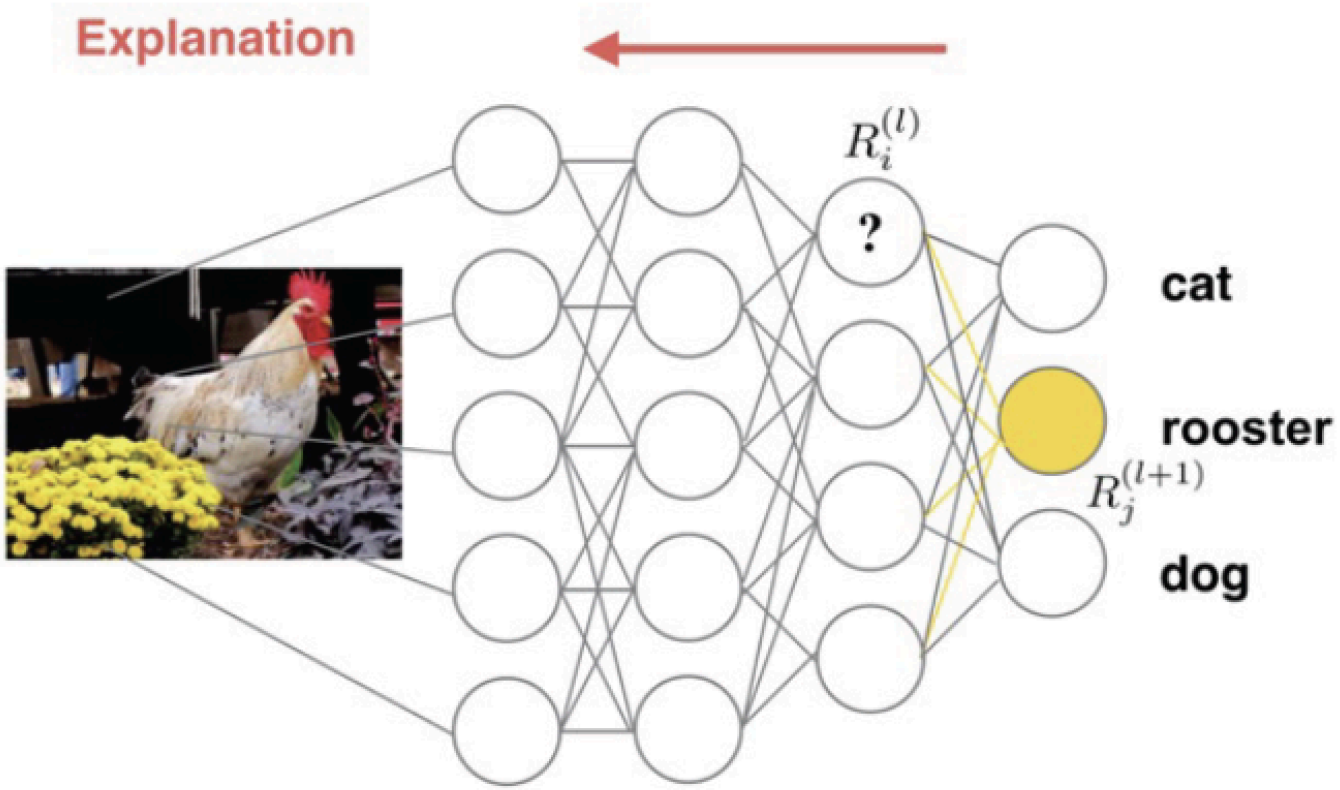
Decision Analysis: LRP



What makes this image a “rooster image” ?

Idea: Redistribute the evidence for class rooster back to image space.

Decision Analysis: LRP

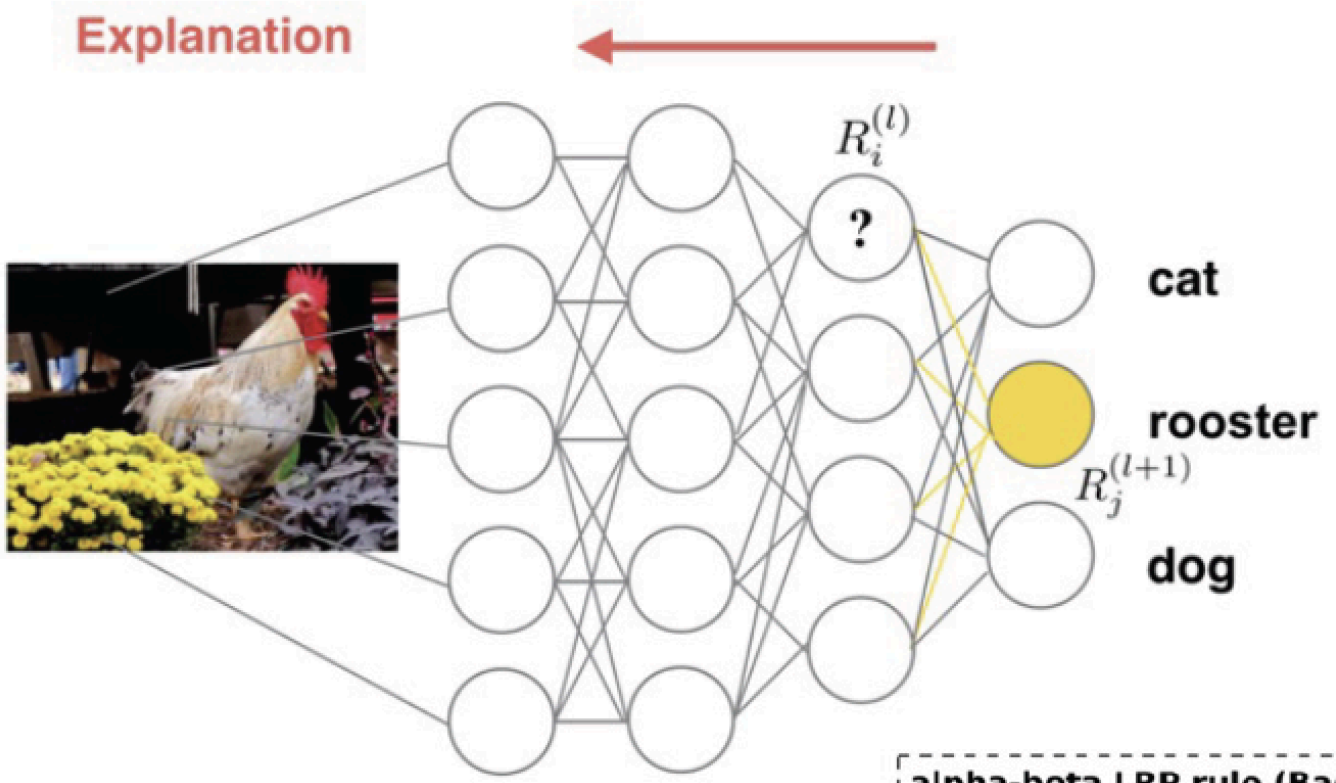


Simple LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}$$

Every neuron gets its "share" of the redistributed relevance

Decision Analysis: LRP



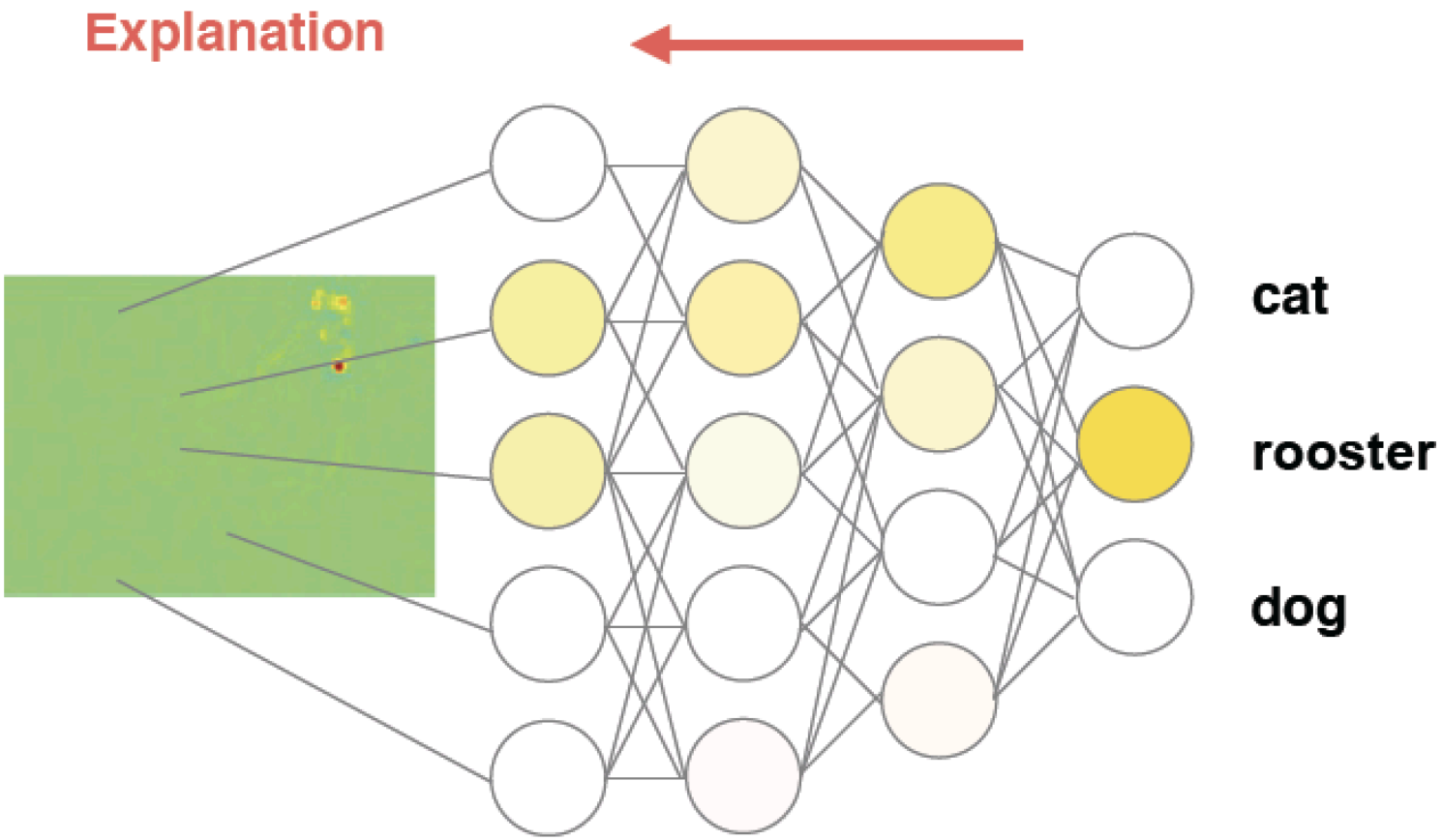
Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)
(no gradient shattering)

alpha-beta LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_{i'}(x_{i'} \cdot w_{i'j})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_{i'}(x_{i'} \cdot w_{i'j})^-}\right) R_j^{(l+1)}$$

where $\alpha + \beta = 1$

Decision Analysis: LRP



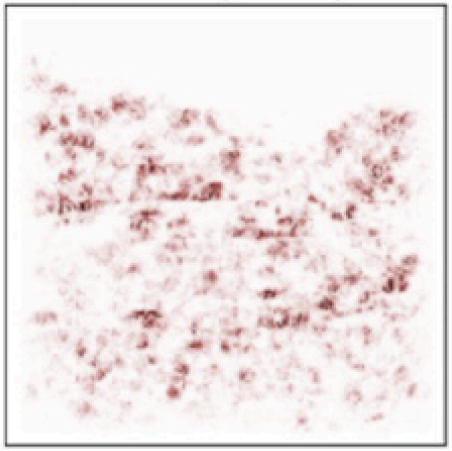
Layer-wise relevance conservation

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

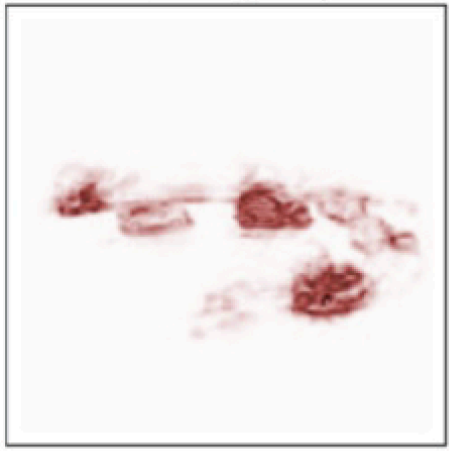
Decision Analysis: LRP

Image

Sensitivity Analysis



LRP / Deep Taylor



Explains what influences prediction "cars".

Explains prediction "cars" as is.

Slope decomposition

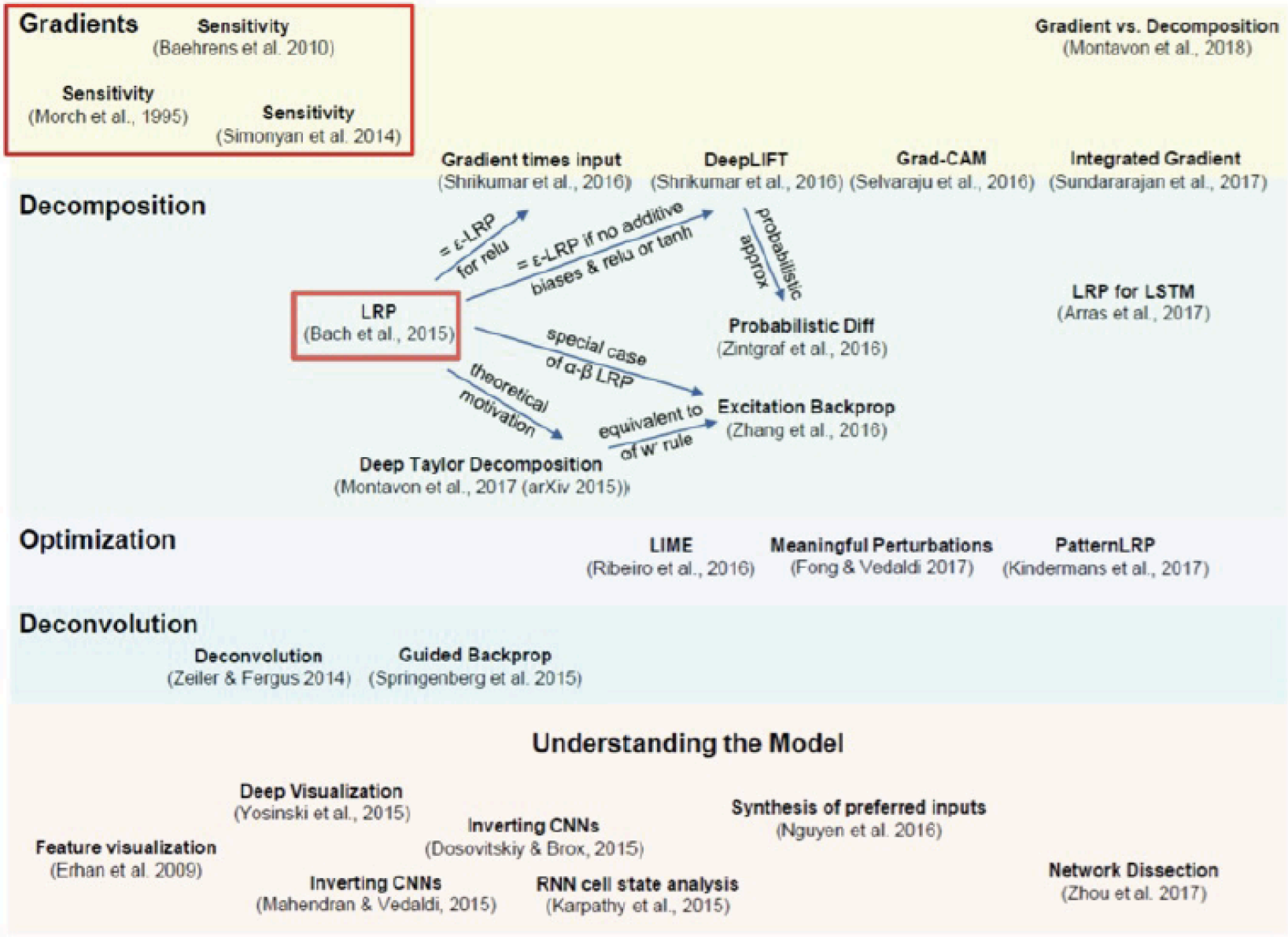
$$\sum_i R_i = \|\nabla_x f\|^2$$

Value decomposition

$$\sum_i R_i = f(x)$$

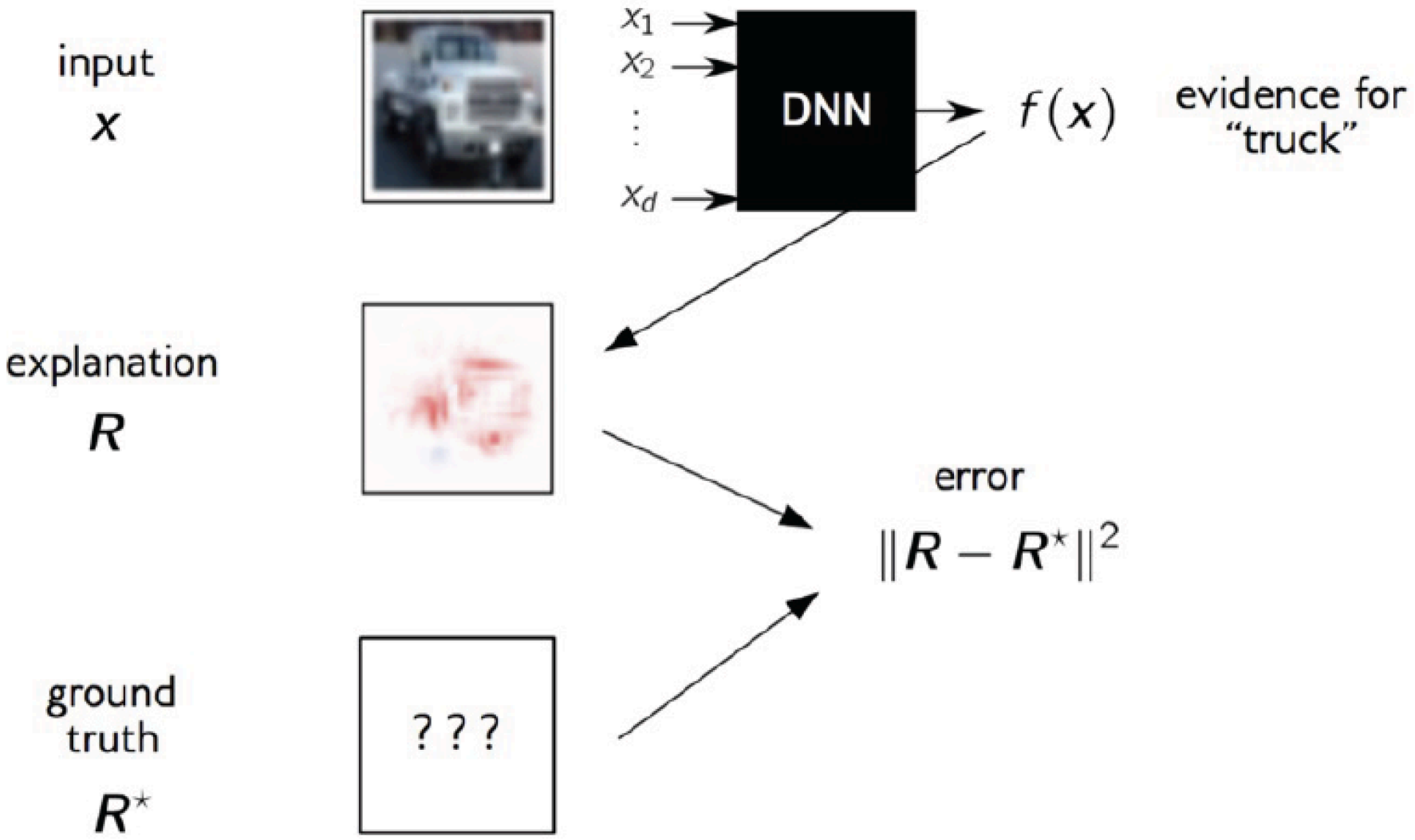
More information (Montavon et al., 2017 & 2018)

Other Explanation Methods



Axiomatic Approach to Interpretability

Distance to Ground Truth

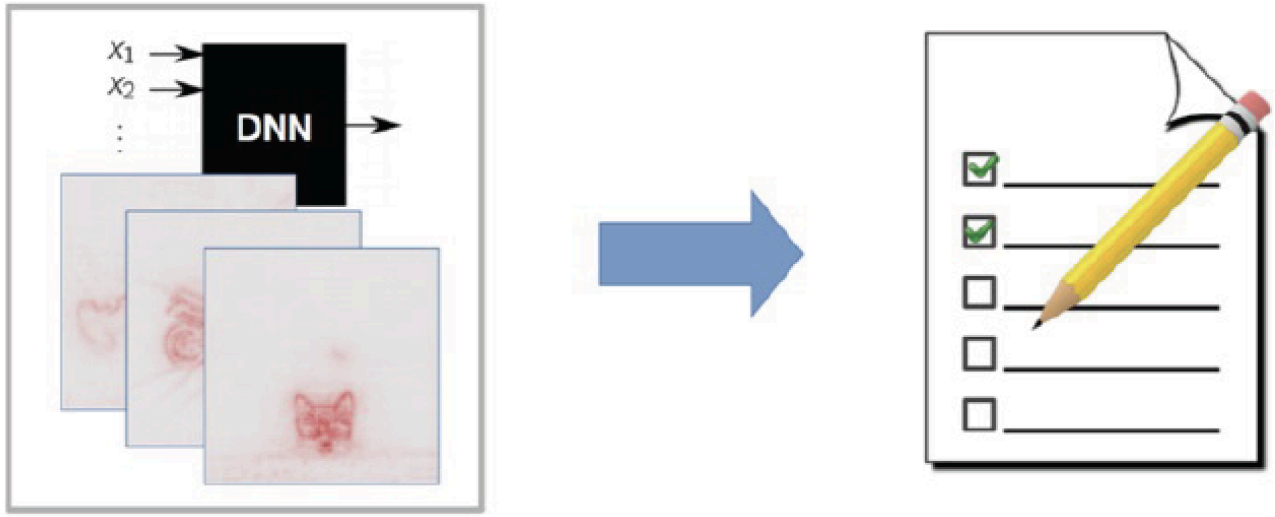


Axiomatic Approach

Idea: Evaluate the explanation technique axiomatically, i.e. it must pass a number of predefined "unit tests".

[Sun'11, Bach'15, Montavon'17, Samek'17, Sundarajan'17, Kindermans'17, Montavon'18].

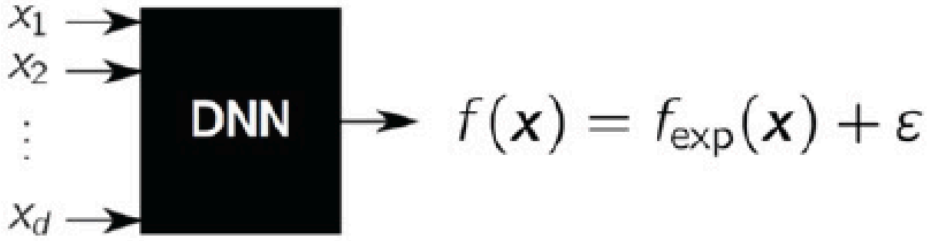
explanation technique



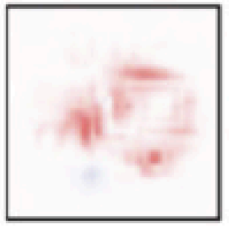
Axiomatic Approach

Properties 1-2: Conservation and Positivity

[Montavon'17, see also Sun'11, Landecker'13, Bach'15]



explanation



$R_1, \ldots, R_d$

Conservation: Total attribution on the input features should be proportional to the amount of (explainable) evidence at the output.

$$\sum_{p=1}^{d} R_p = f_{\text{exp}}(x)$$

Positivity: If the neural network is certain about its prediction, input features are either relevant (positive) or irrelevant (zero).

$$\forall_{p=1}^{d} : \; R_p \geq 0$$

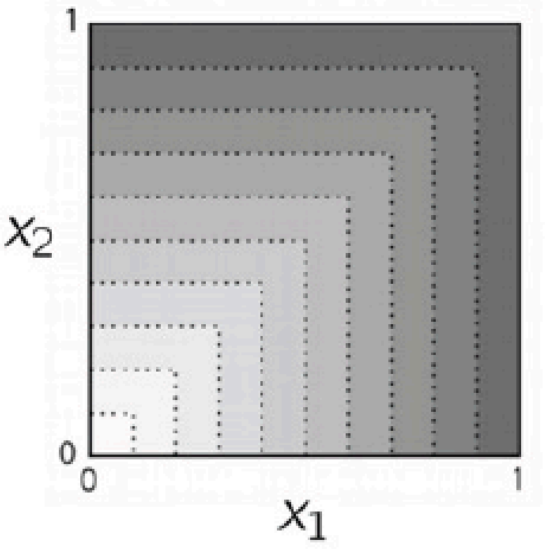
Axiomatic Approach

Property 3: Continuity [Montavon'18]

If two inputs are the almost the same, and the prediction is also almost the same, then the explanation should also be almost the same.

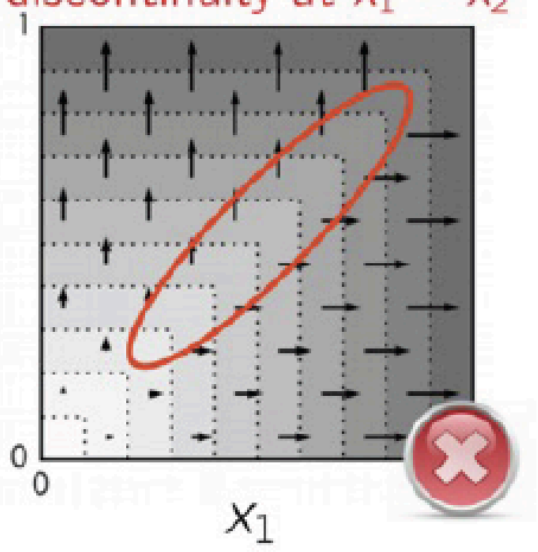
Example:

$f(x) = \max(x_1, x_2)$

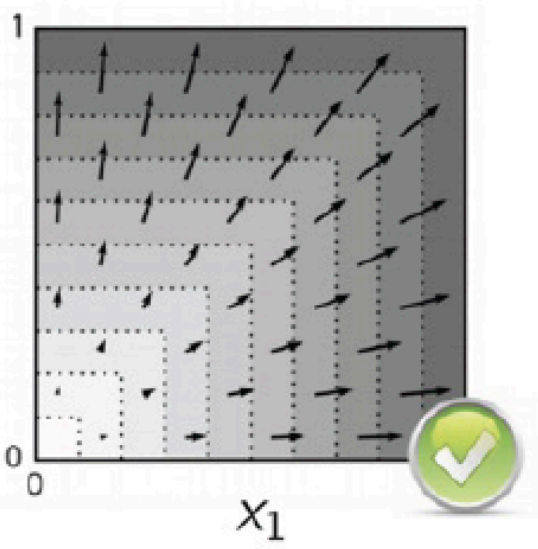


Method 1

discontinuity at $x_1 = x_2$



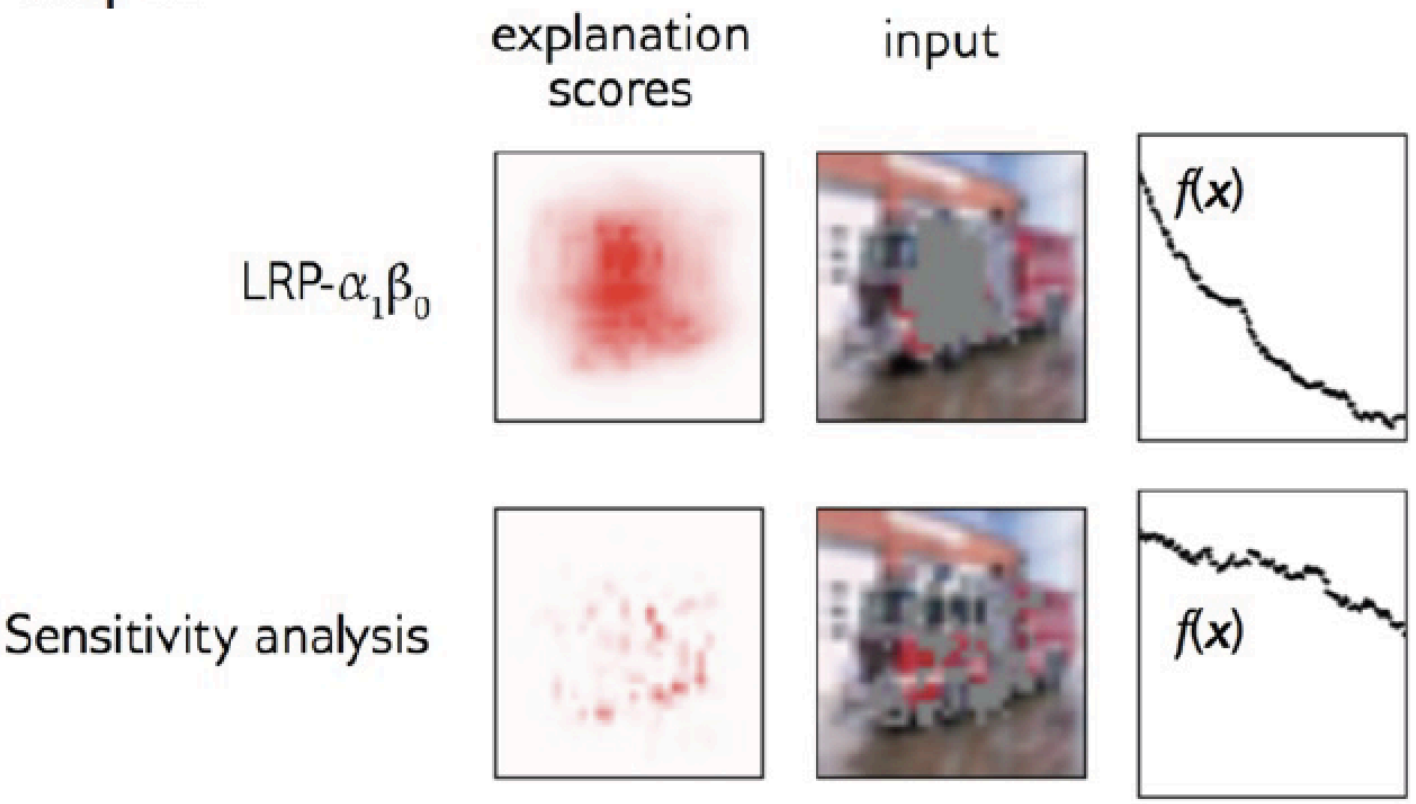
Method 2



Axiomatic Approach

Property 4: Selectivity [Bach'15, Samek'17]

Model must agree with the explanation: If input features are attributed relevance, removing them should reduce evidence at the output.

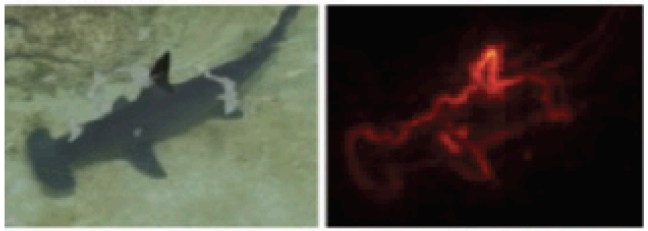


Axiomatic Approach

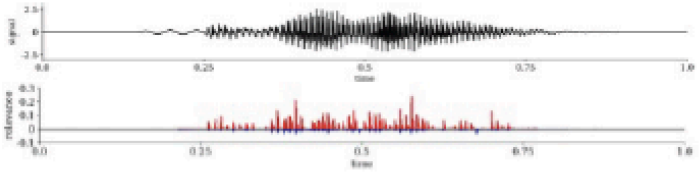
Explanation techniques		Uniform	$(\text{Gradient})^2$	$(\text{Guided BP})^2$	Gradient x Input	Guided BP x Input	LRP-$\alpha_1\beta_0$	...
Properties								
1. Conservation		✔			✔	✔	✔	
2. Positivity		✔	✔	✔		✔	✔	
3. Continuity		✔		✔		✔	✔	
4. Selectivity			✔	✔	✔	✔	✔	
...								

Axiomatic Approach

General Images (Bach' 15, Lapuschkin'16)



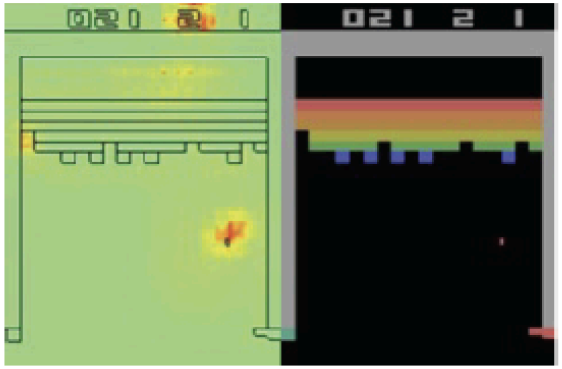
Speech (Becker'18)



Text Analysis (Arras'16 &17)

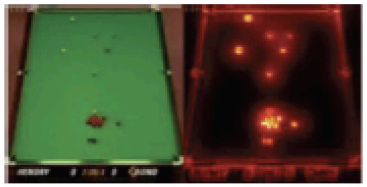
Morphing (Seibold'18)

Games (Lapuschkin'18)

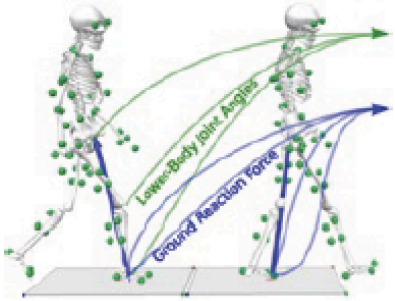


VQA (Arras'18)

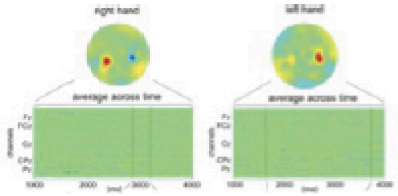
Video (Anders'18)



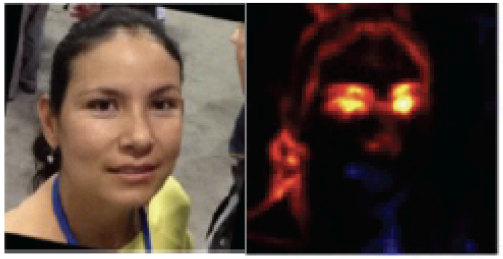
Gait Patterns (Horst'18)



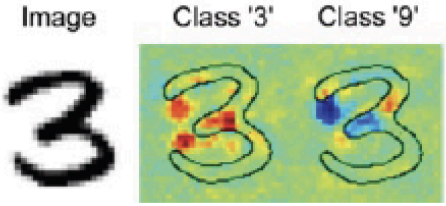
EEG (Sturm'16)



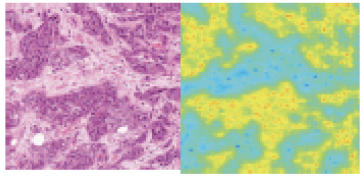
Faces (Lapuschkin'17)



Digits (Bach' 15)



Histopathology (Binder'18)



fMRI (Thomas'18)