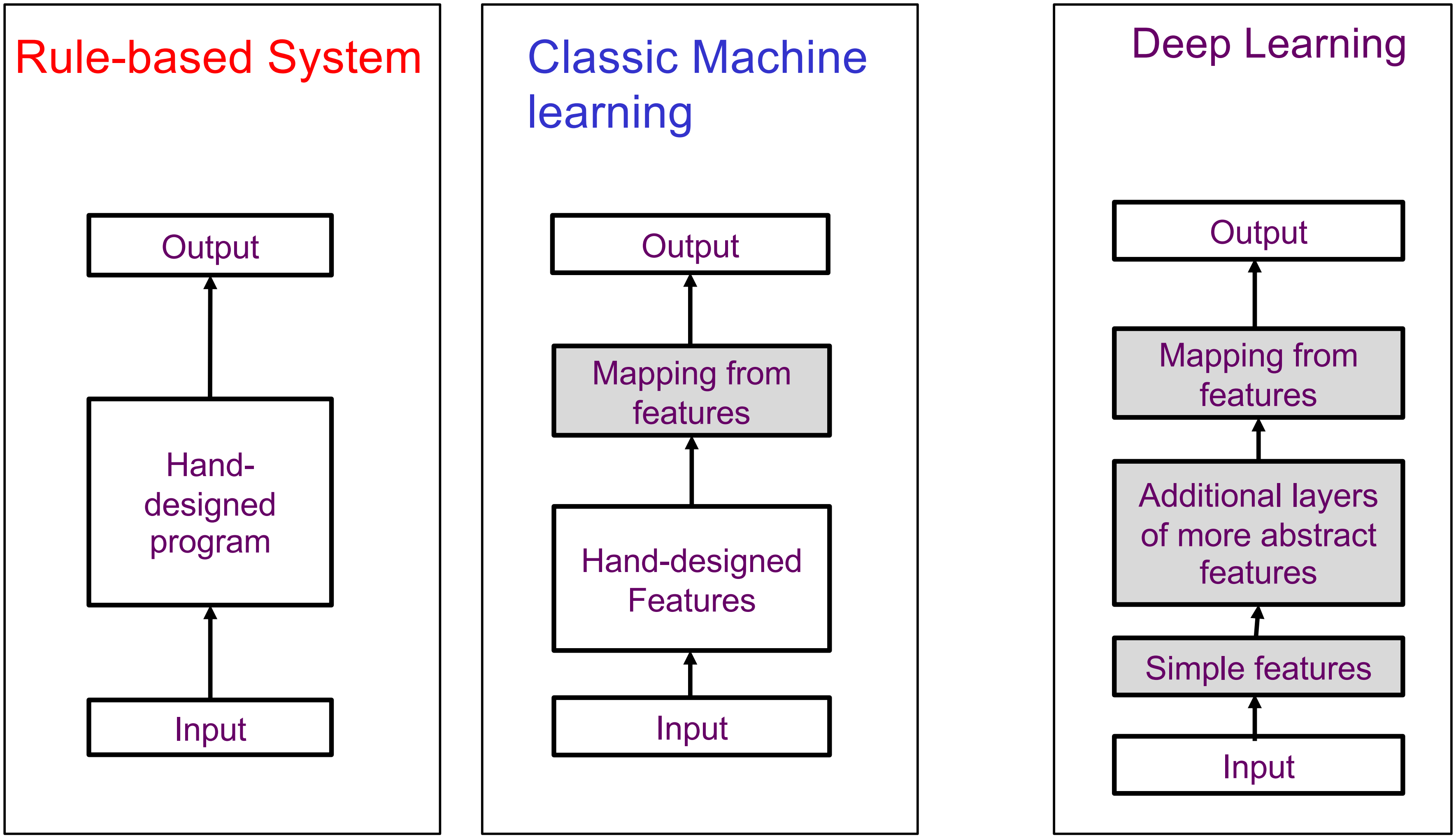
Explainable Artificial Intelligence

Sargur N. Srihari
srihari@buffalo.edu

Topics in Explainable AI (XAI)

1. Types of AI Models
2. Need for Explainability
3. Post-hoc and Ante-hoc Explainability
4. Sensitivity Analysis and Layerwise Relevance Propagation
5. Measures of Explanation Quality

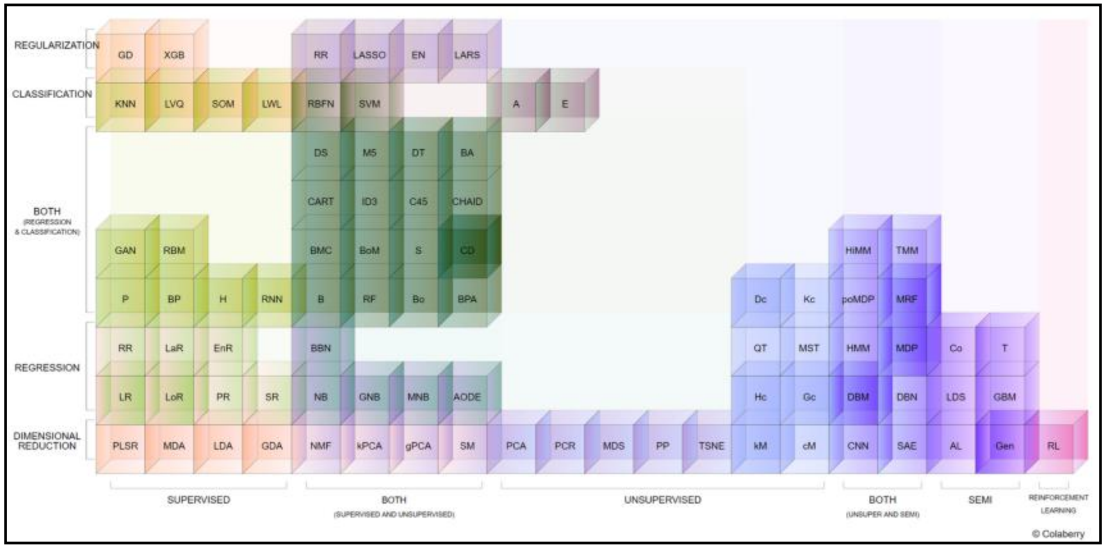
Summary of AI Models

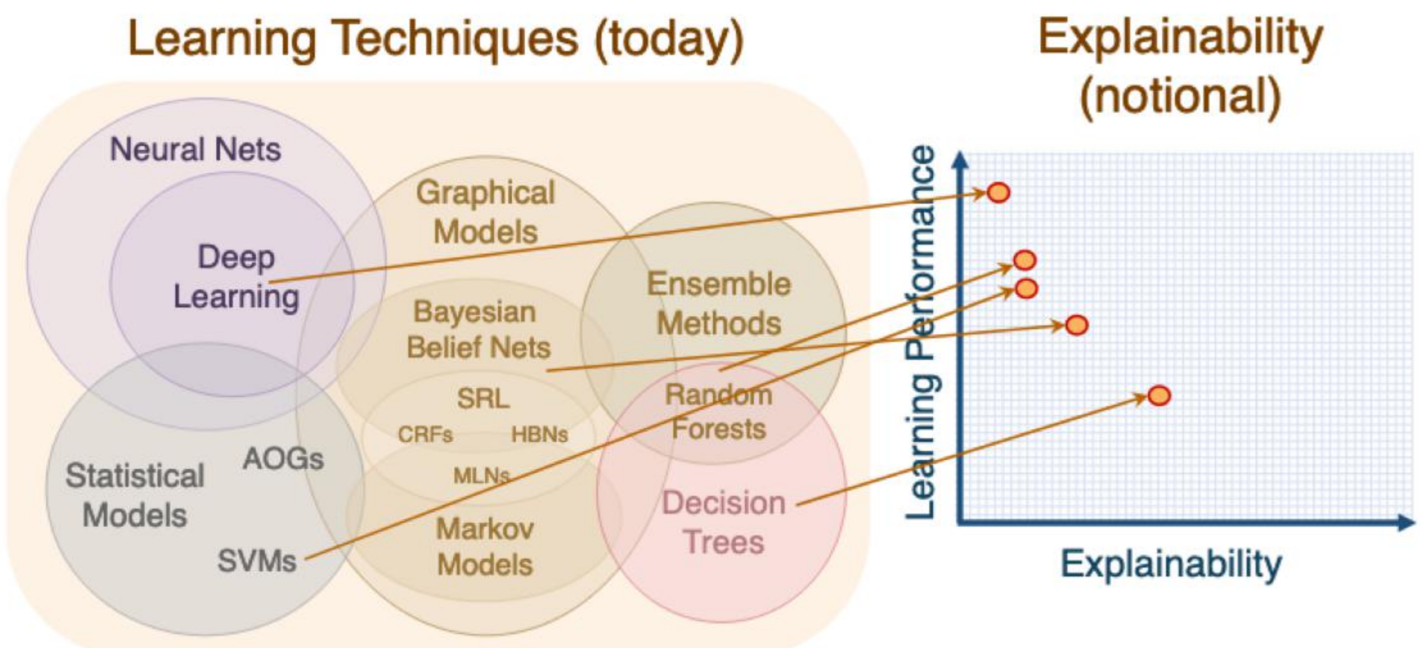


Shaded boxes indicate components that can learn from data

Explainability of AI Models

AI Models

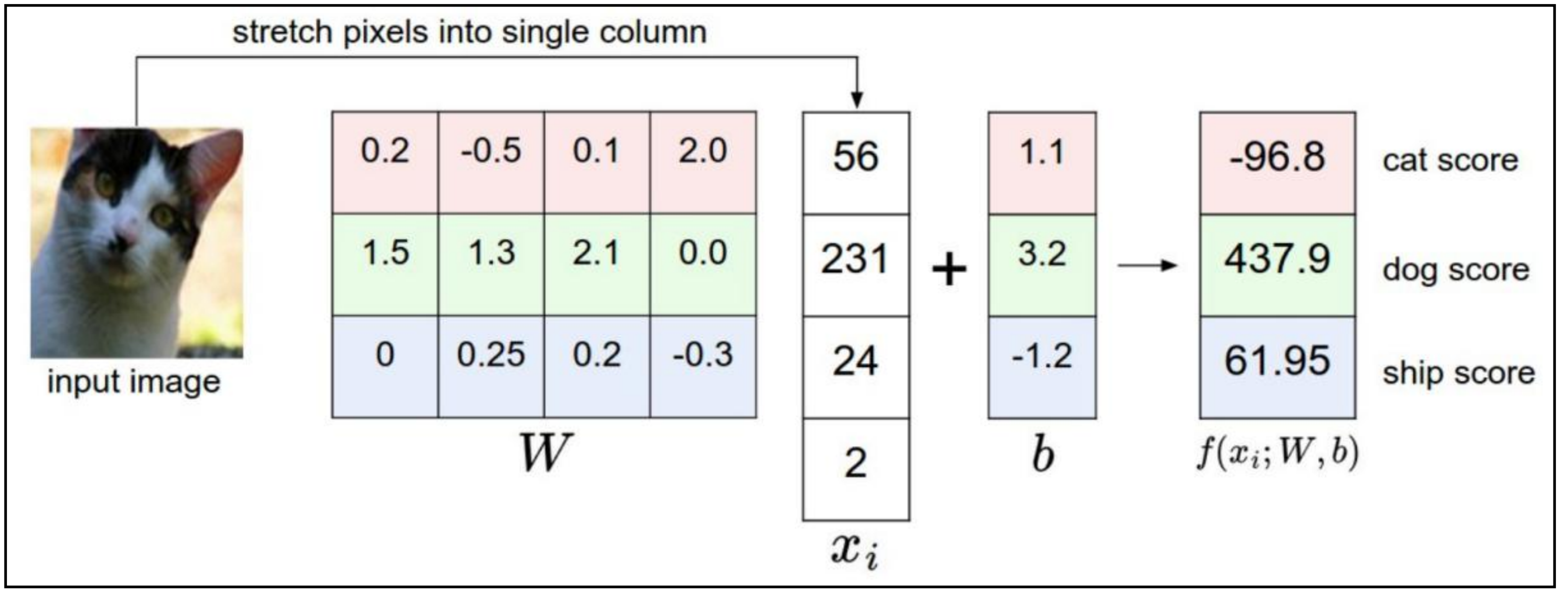


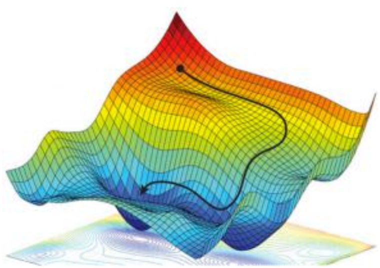


A neural network is defined by weights

Vector $\boldsymbol{x}$ is converted into vector $\boldsymbol{y}$ by multiplying $\boldsymbol{x}$ by a matrix $\boldsymbol{W}$

A linear classifier $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{b}$

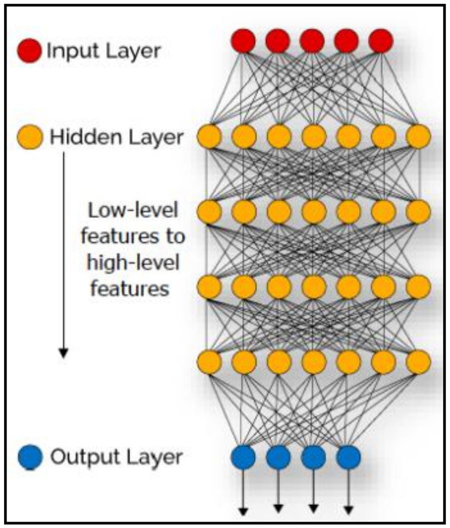




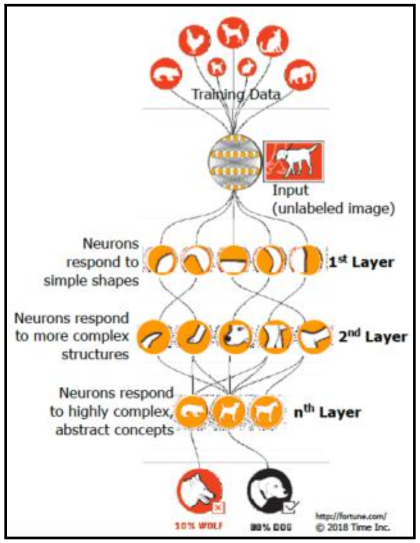
$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} - \varepsilon \nabla_{\boldsymbol{w}} f\left(\boldsymbol{w}^{t}\right)$$

Deep Neural Network

Supervised Deep Network

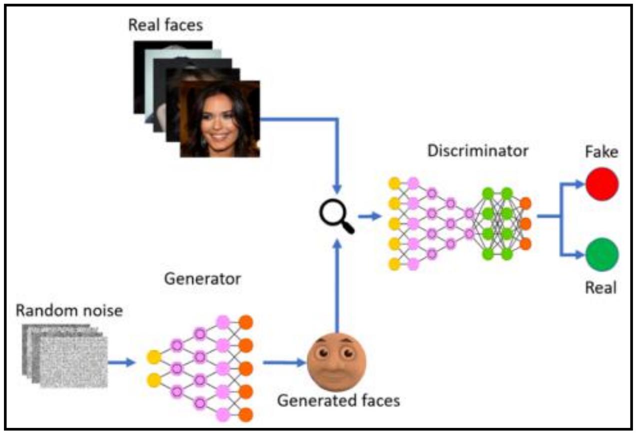


Source: XenonStack



Source: TIME

Unsupervised Deep Network



Deep Learning as Blackbox models

Blackbox:

- A function that is too complicated for any human to comprehend
- A function that is proprietary
- A model that is difficult to trouble-shoot

Deep Learning Models are blackbox models

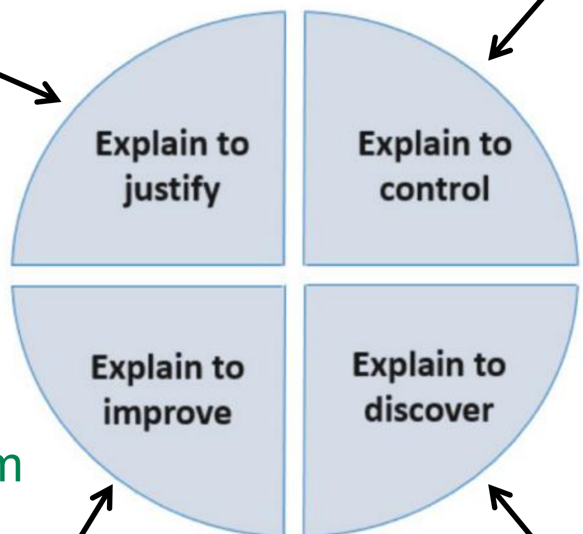
- Popular networks are recursive
- Opaque, non-intuitive and difficult for people to understand

Source Image: Wikipedia

Summary of need for explainable AI

Verification of the system
Ability to explain one's decision to other people is an important aspect of human intelligence
Ex: Due to lack of data no correlation of asthma-heart disease with death from pneumonia

Compliance to legislation
"Right to explanation" when Algorithm makes a loan decision



Improvement of the system
First step to improve system is to understand its weaknesses.
Detecting bias in system.

Learning from the system
AI systems are trained with millions of examples
They may observe patterns in data not available to humans

Explain to control

- Credit decisions:
  - How did DL model provide /deny credit to an individual? Was there bias: ethnicity, race religion?
  - In the US
    - lender must provide *reasons* for adverse decision
      - Take-home insufficient, insufficient collateral, poor credit rating
  - In the European Union,
    - GDPR (General Data Protection Regulation)
    - *right to explanation* for high-stakes automated decisions
- Healthcare:
  - Highly regulated due to HIPAA
  - How did AI predict grade 3 or grade 4 tumor?

Explain to justify: Ethical AI

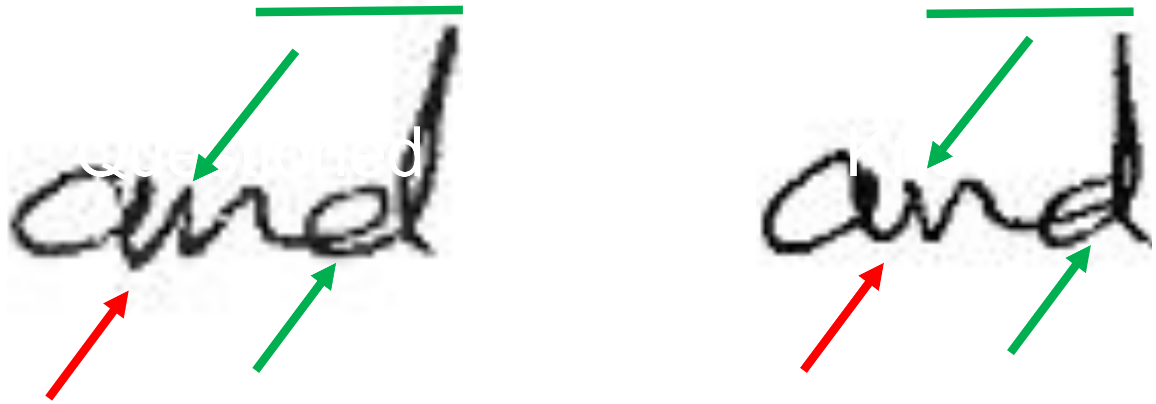
- Surveillance system:
  - Why did model interpret that individual in livestream video is suspicious? Are there biases?
- Autonomous vehicle:
  - Model decides on saving passenger or pedestrian

Explain to justify

1. Medical
   - Patient discharge to nursing home, Reading EKGs
2. Legal
   - People incorrectly denied parole
   - Bail decision leads to release of dangerous criminal
3. Environment
   - Pollution model states that dangerous is safe

- Poor use of limited valuable resources in
  - Medicine, Justice, energy reliability, finance

Explain to Justify in Forensics



This chart represents some of the handwriting features I used to reach my conclusion, and is representative of what I examined but is not meant to replicate my entire examination. There are several handwriting similarities noted using the green arrows/bar, and a dissimilarity noted using the red arrows.

Similarities include the pointed first arch of the letter “n”,
the counterclockwise motion of the formation of the front portion of the “d”
and the way the stroke connects from the “n” into the middle of the “d”,
and the relative height of the staff of the “d” compared to the other letters.

A dissimilarity is noted in the connecting stroke from the “a” to the “n”.
Is it below the baseline and has a sharp angle in the questioned “and”,
and at or above the baseline and rounded in the known “and”.

Conclusion would likely be “indications one writer wrote both words”.
A more definitive conclusion could not be reached because of the limited amount of writing for comparison, and the dissimilarity noted.

DARPA goals of XAI

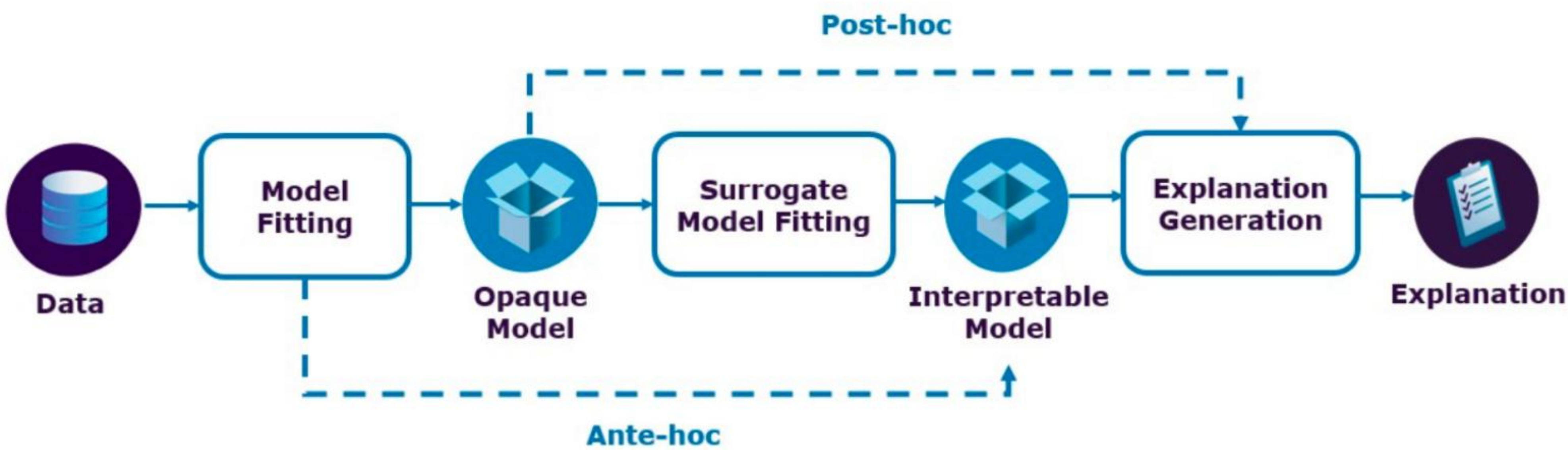
AI
SYSTEM

→

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?

Taxonomy of XAI

- Post-hoc (Explain the Blackbox)
  - Explainability based on test cases and results
- Ante-hoc (Build a new learning model)
  - Seeding explainability into model from the start



Source: Sogetilabs

Some examples of XAI

- Post-hoc (Blackbox)
  1. Sensitivity Analysis (SA)
  2. Layer-wise Relevance Propagation (LRP)
  3. Local Interpretable Model-Agnostic Explanations (LIME)
- Ante-hoc (New learning process)
  1. Reversed Time Attention Model (RETAIN)
  2. Bayesian Deep Learning (BDL)

Sensitivity Analysis (SA)

- Explains a prediction $f(\boldsymbol{x})$ based on the model's locally evaluated gradient (partial derivative)
  - It quantifies importance of each input x_i (e.g., image pixel) as

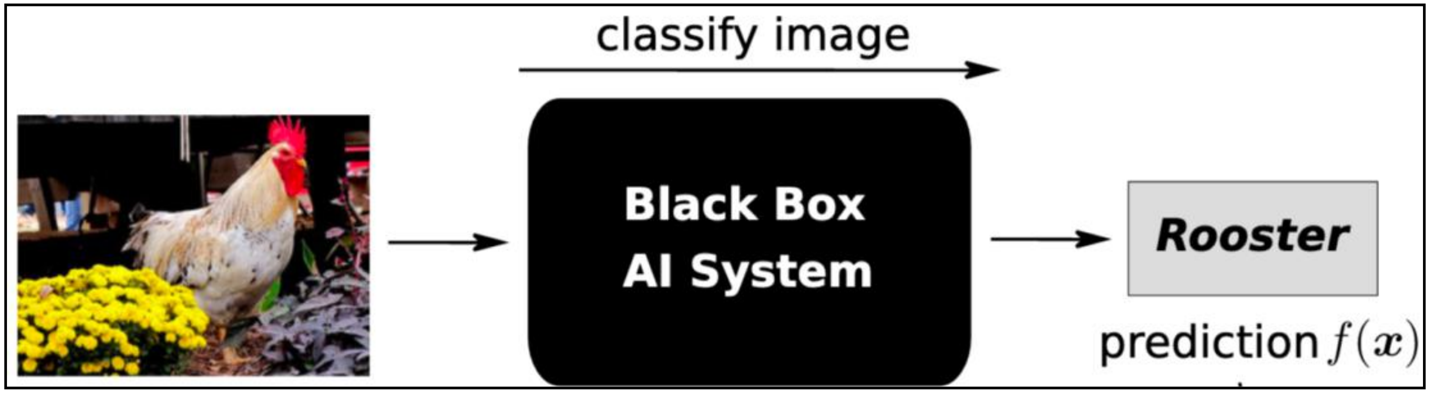
SA: Partial derivatives

$$R_i = \left\| \frac{\partial}{\partial x_i} f(\boldsymbol{x}) \right\|$$

(how much do changes in each pixel affect the prediction)

- Assumes that most relevant input features are those to which the output is most sensitive

Explanation using Sensitivity Analysis



Input x

SA: Partial derivatives

$$R_i = \left|\left|\frac{\partial}{\partial x_i} f(\boldsymbol{x})\right|\right|$$

(how much do changes in each pixel affect the prediction)

Assumes that most relevant features are those to which output is most sensitive.
Which pixels need to be changed to make image look more/less like the predicted class

e.g., changing yellow occluding pixels improves score, but does not explain rooster

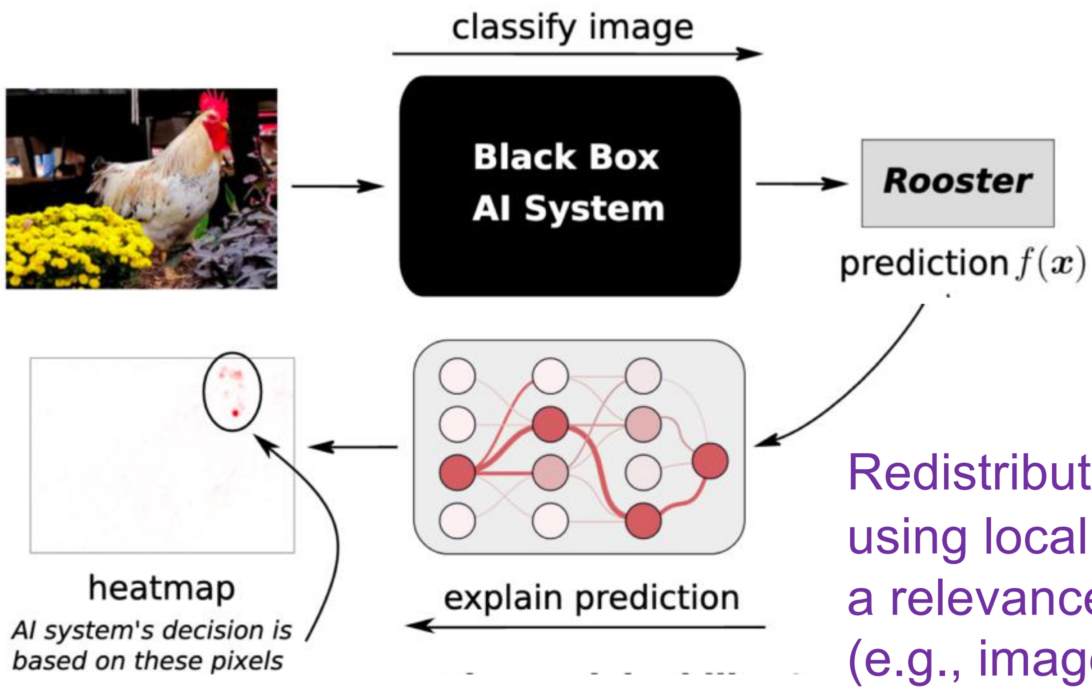
Explains prediction based on locally evaluated gradient
Does not explain $f(\boldsymbol{x})$ but a variation of input

Source:http://iphome.hhi.de/samek/pdf/SamITU18b.pdf

Layerwise Relevance Propagation

- A general framework for decomposing predictions of modern AI systems in terms of input variables
  - Applicable to feed-forwards networks, bag-of-words models, LSTM and FisherVector classifiers
- In contrast to SA, this method explains predictions relative to the state of maximum uncertainty
  - i.e., it identifies pixels which are pivotal for the prediction “rooster”

Explanation using LRP



Layerwise
Relevance
Propagation

Redistributes the prediction $f(x)$ backwards using local redistribution rules until it assigns a relevance score R_i to each input variable (e.g., image pixel)

Explains rooster by its head

LRP: Decomposition

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

(how much does each pixel contribute to prediction)

Simple LRP rule

$$R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \epsilon} R_k$$

x_j = neuron activations at layer l,
R_k=relevance scores of neurons at layer $l+1$
w_{jk}= weight connecting neuron j to neuron k

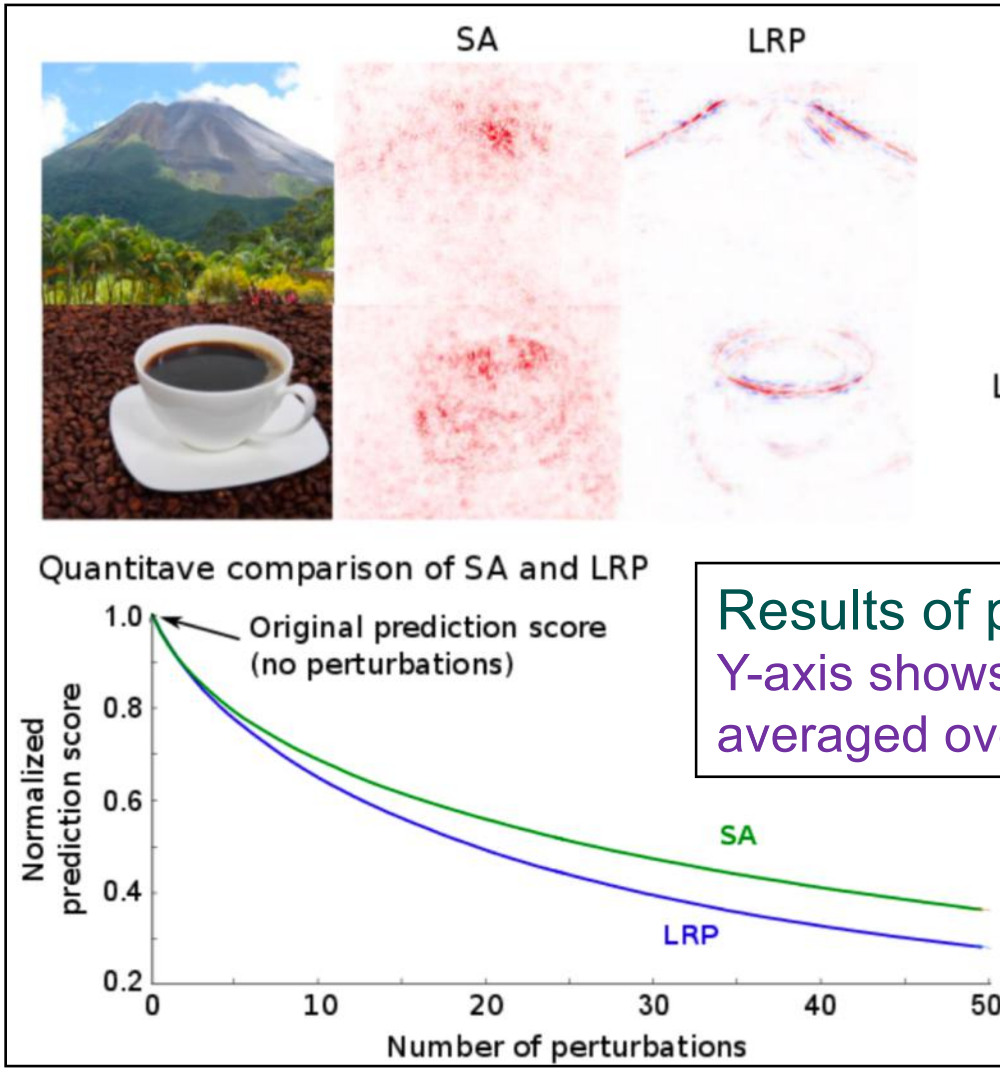
Source:http://iphome.hhi.de/samek/pdf/SamITU18b.pdf

Evaluating the quality of explanation

- Compare heatmaps of SA and LRP using *Perturbation analysis:*
  1. Perturbing important variables leads to a steeper decline of prediction score than lesser variables
  2. SA, LRP provide a score for each input. Thus, inputs can be sorted by this relevance score
  3. Iteratively perturb variables (starting from most relevant), track score after every perturbation
- Average decline of the prediction score is a measure of explanation quality
  - because a large decline indicates that explanation was successful in identifying truly relevant input variables

Comparison of SA and LRP

Two images correctly classified as *volcano, coffee cup*



Source:http://iphome.hhi.de/samek/pdf/SamITU18b.pdf

LRP detects shape of mountain and ellipsoidal shape as relevant features

SA does not indicate how much each pixel contributes to prediction

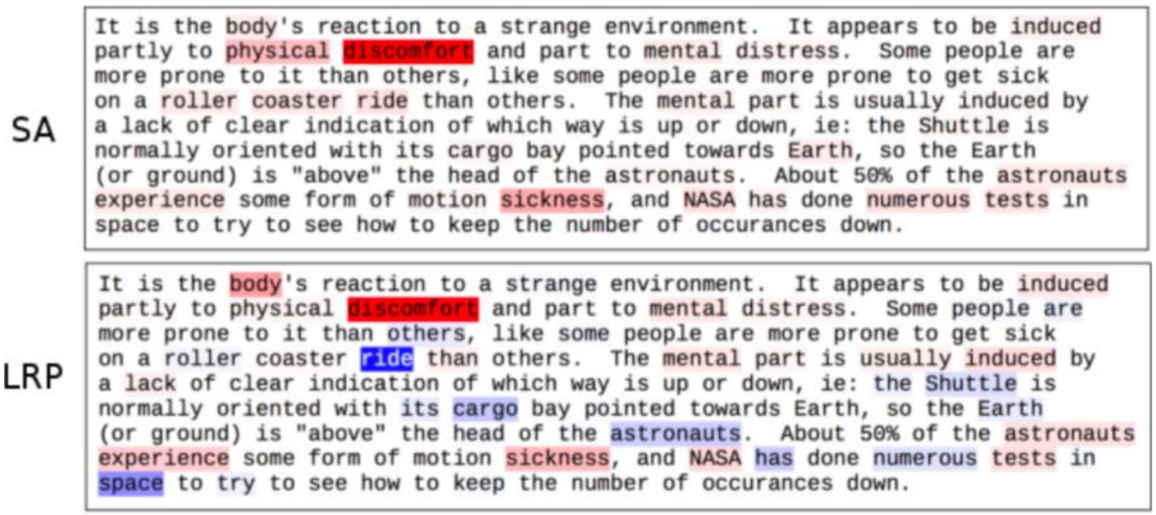
Results of perturbation analysis

Y-axis shows relative decrease of prediction score averaged over 5040 images of ILSVRC2012 dataset

At every perturbation step a 9x9 patch of the image (selected according to scores) is replaced by random values sampled from an uniform distribution

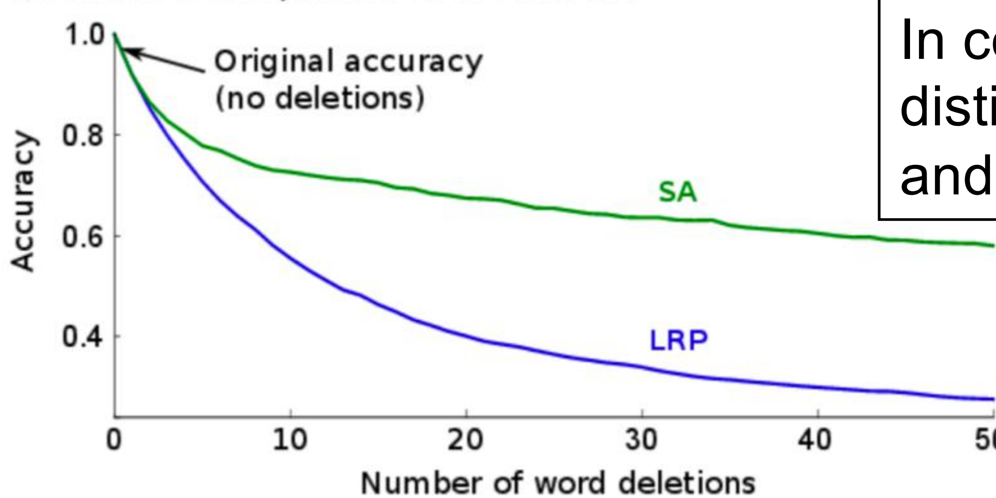
Explaining Text Classification

Explaining prediction *sci.med*



SA and LRP heatmaps identify words such as “discomfort”,“body” and “sickness” as relevant ones for explaining the prediction.

Quantitave comparison of SA and LRP

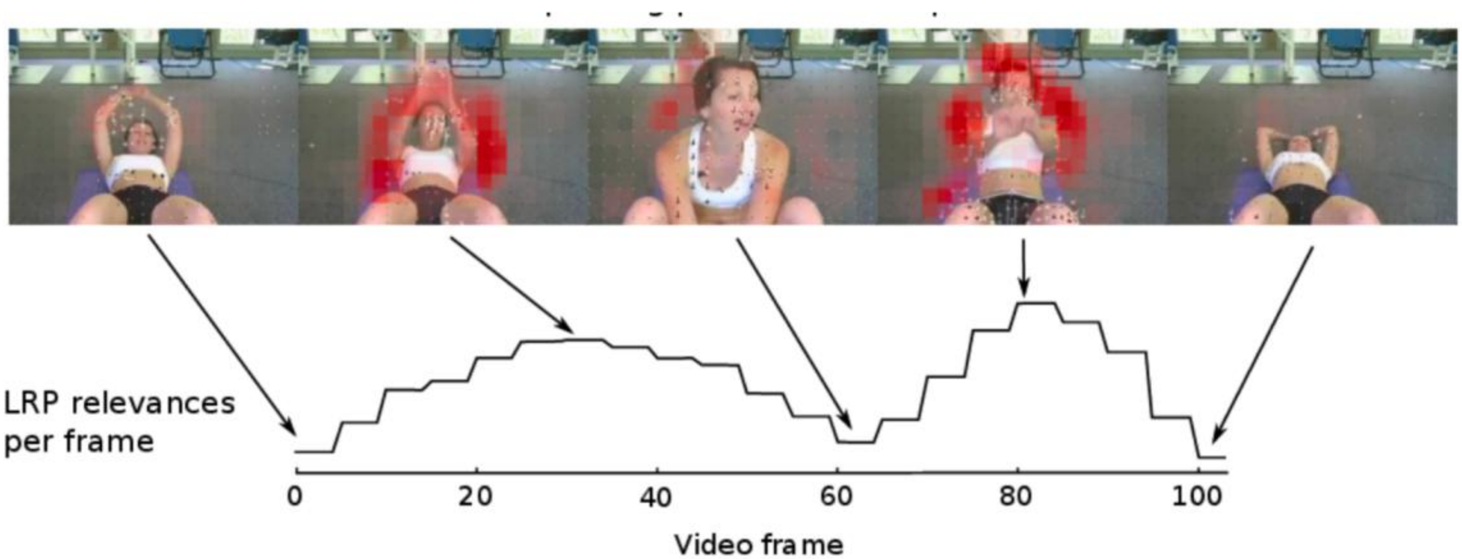


In contrast to SA, LRP distinguishes between positive (red) and negative (blue) relevances.

Source:http://iphome.hhi.de/samek/pdf/SamITU18b.pdf

Explaining Human Action Recognition

Explaining prediction *sit-up*

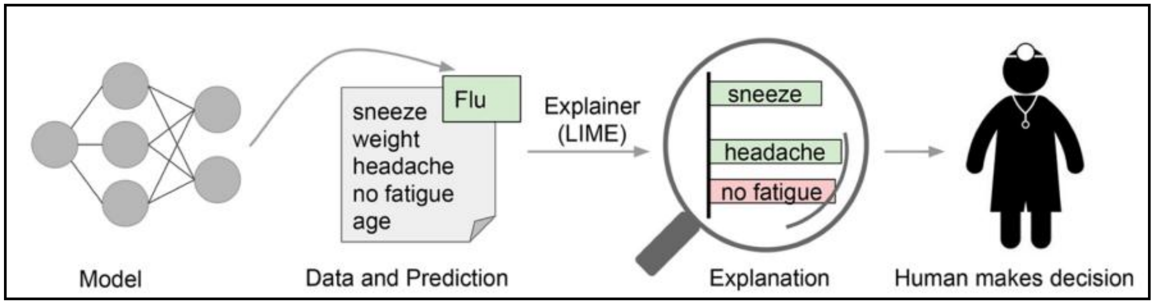


The LRP heatmaps of a video which was classified as “sit-up” show increased relevance on frames in which the person is performing an upwards and downwards movement..

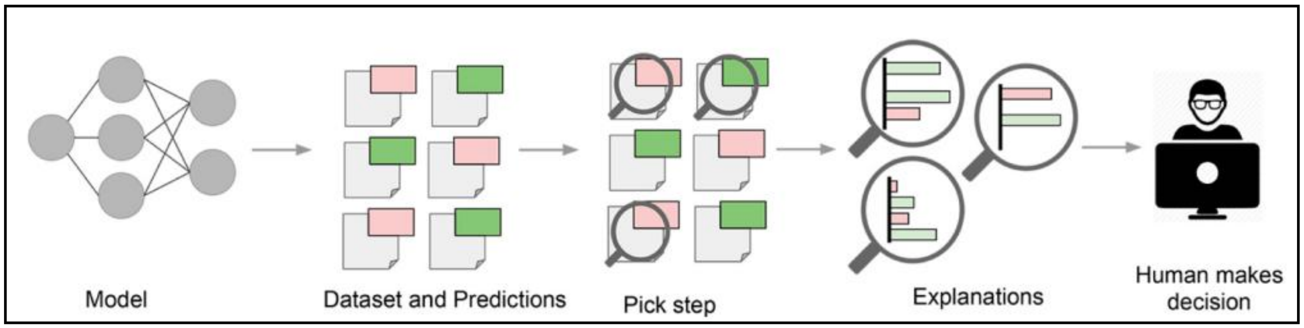
Source:http://iphome.hhi.de/samek/pdf/SamITU18b.pdf

A post-hoc system: LIME

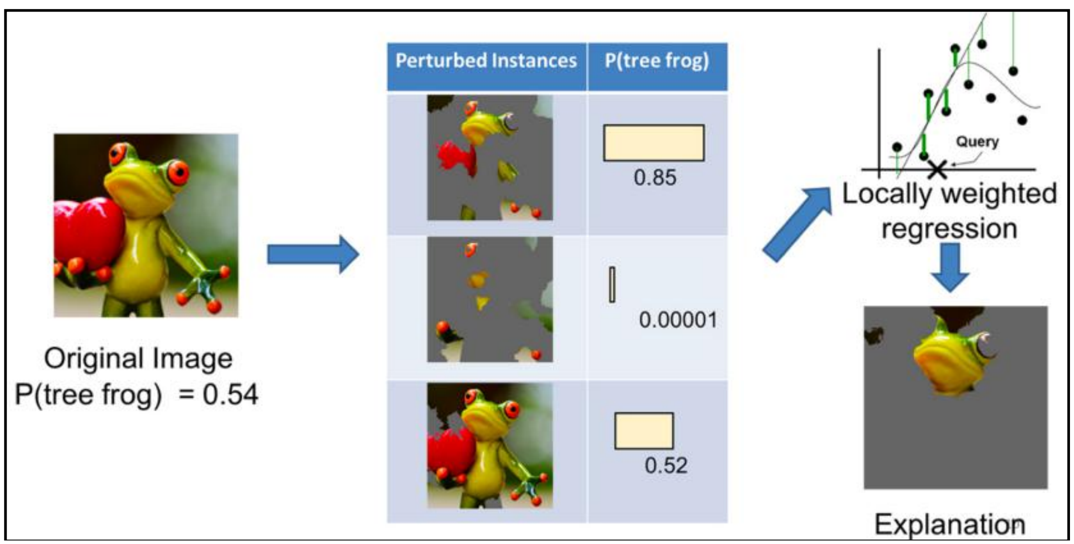
Local Interpretable Model-Agnostic Explanations (LIME)



Explaining individual prediction to a human decision-maker

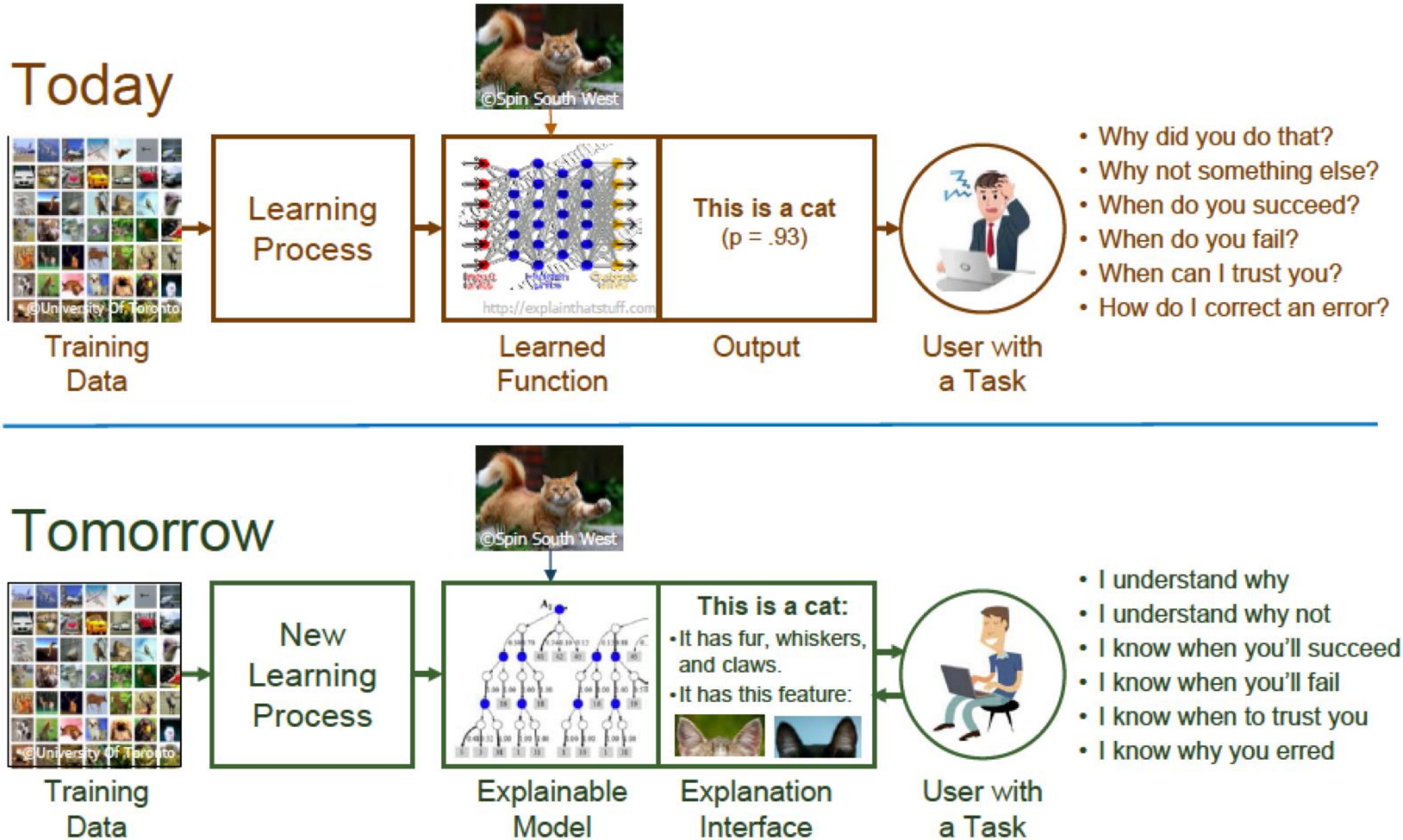


Explaining model to a human decision-maker



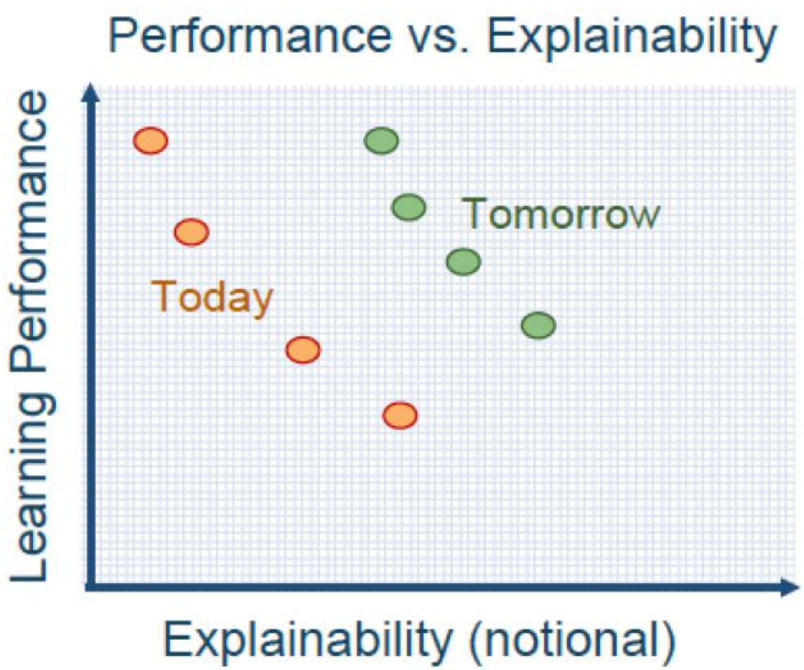
Source: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Ante-hoc systems: DARPA



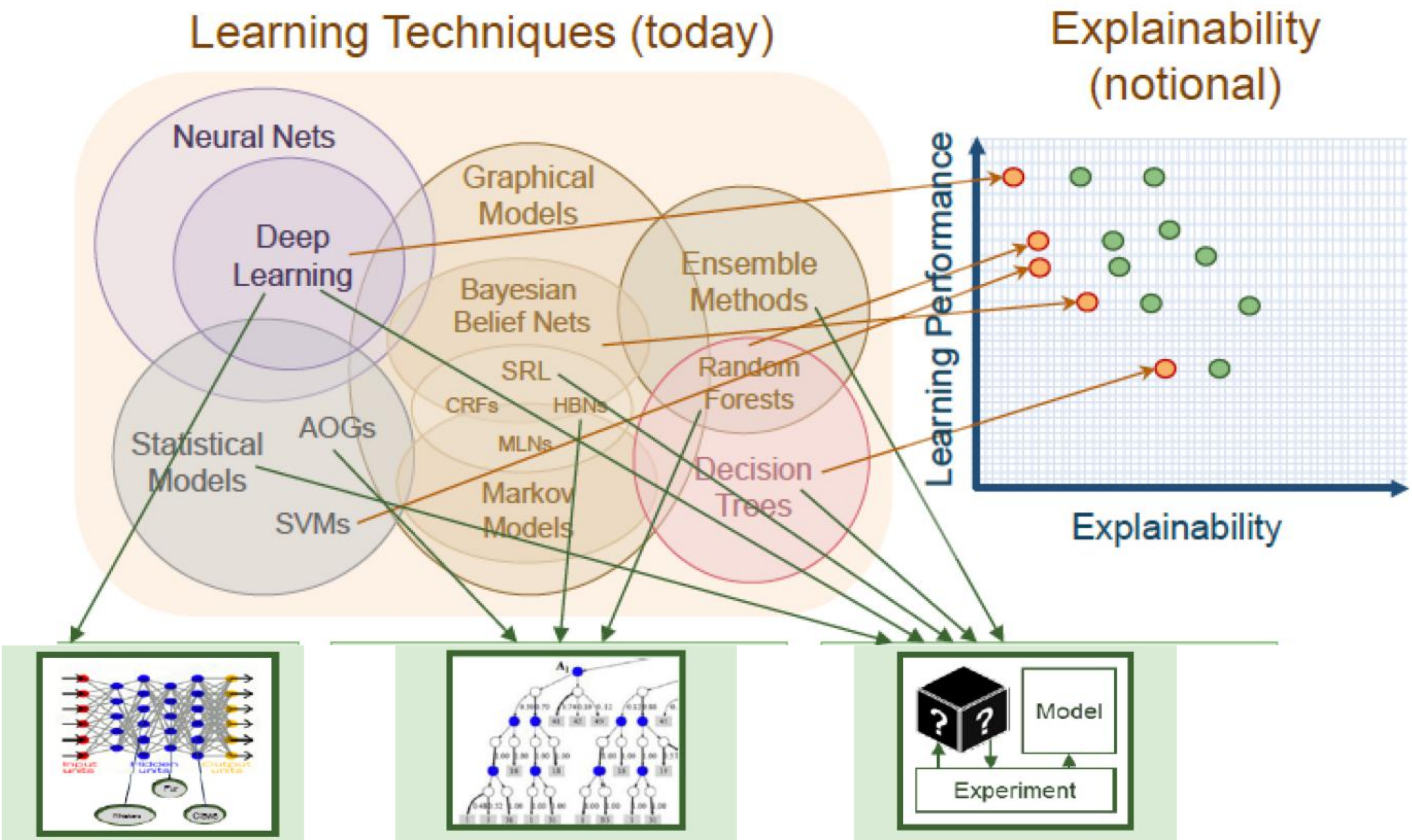
Performance vs Explainability

- Produce more explainable models, while maintain a high level of performance, e.g., prediction accuracy
- Enable human users to understand, trust and manage emerging AI partners



complex black box models such as RNN)
or less accurate traditional models with
better interpretation,e.g., logistic regression

Future AI Models



Deep Explanation
Modified deep learning techniques to learn explainable features

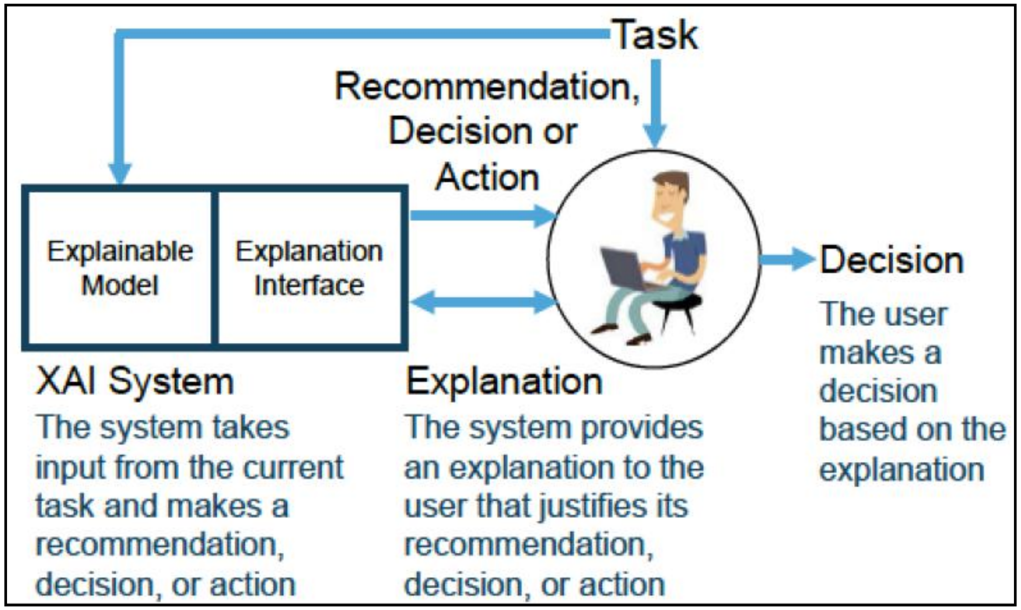
Interpretable Models
Techniques to learn more structured, interpretable, causal models

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

Measures of Explanation Effectiveness

- User satisfaction
  - Clarity of explanation (user rating)
  - Utility of explanation (user rating)
- Task performance
  - Does it improve user decision, performance
- Trust assessment
  - Future use and trust

Explanation Framework



DARPA XAI Program

1. Techniques to select the training examples most influential in a decision
2. Techniques to identify the most salient input features used in a decision
3. Network dissection techniques to identify meaningful features inside the layers of a deep net
4. Deep learning techniques to generate explanations