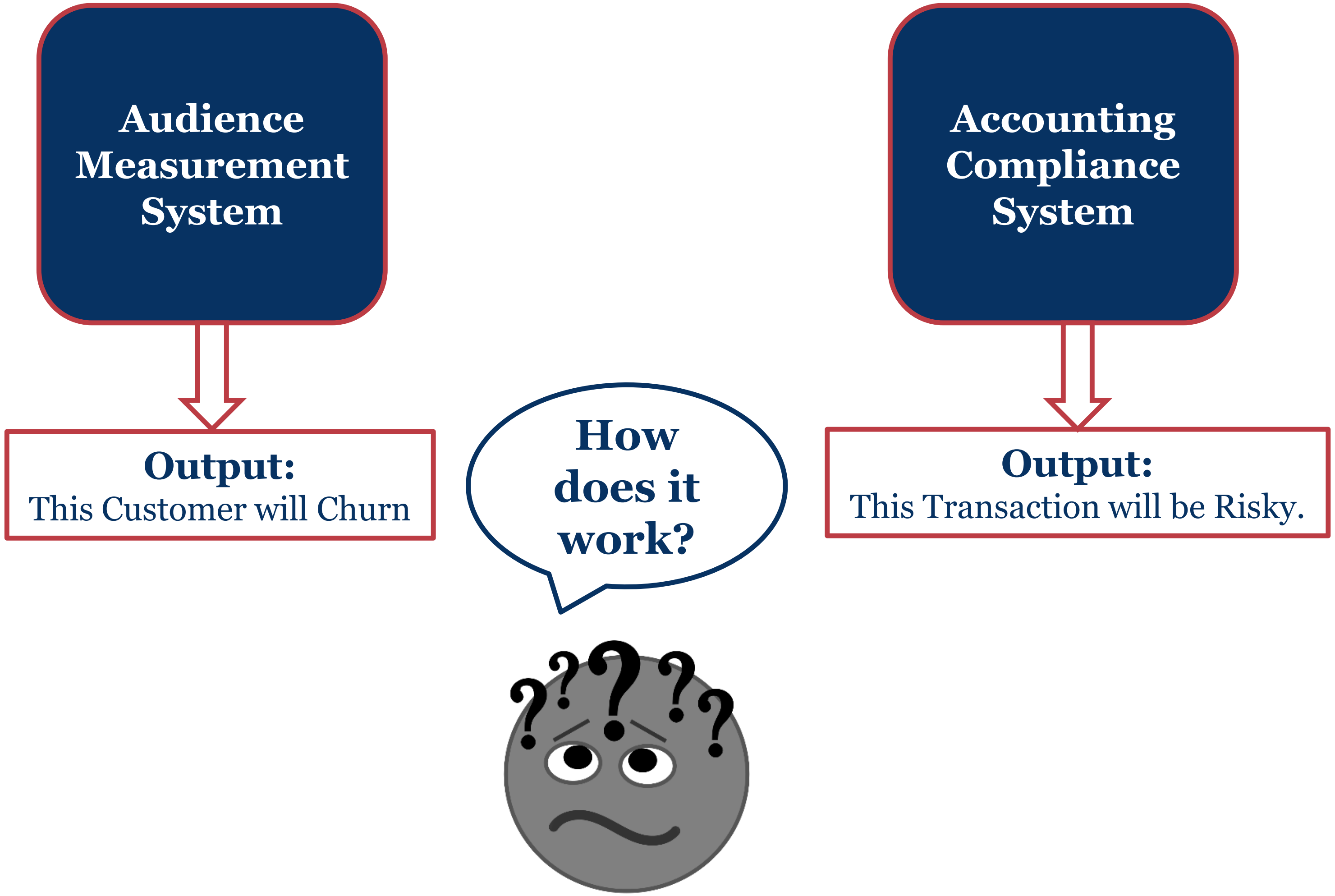
Explainable AI (XAI) – A Perspective

Author – Saurabh Kaushik
Twitter - @saurabhkaushik

AI - How does it Work?



Audience Measurement System

Output:
This Customer will Churn

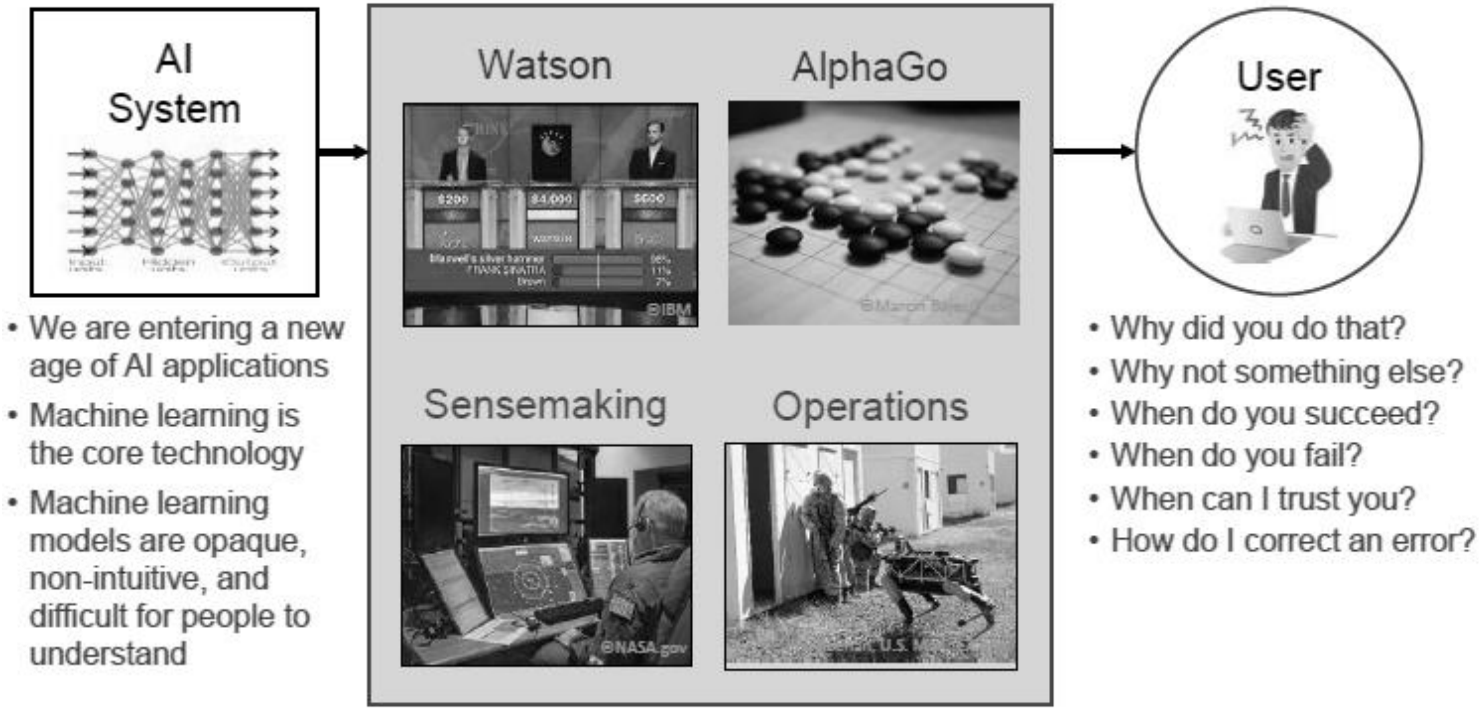
Accounting Compliance System

Output:
This Transaction will be Risky.

AI as Black Box

Black Box AI

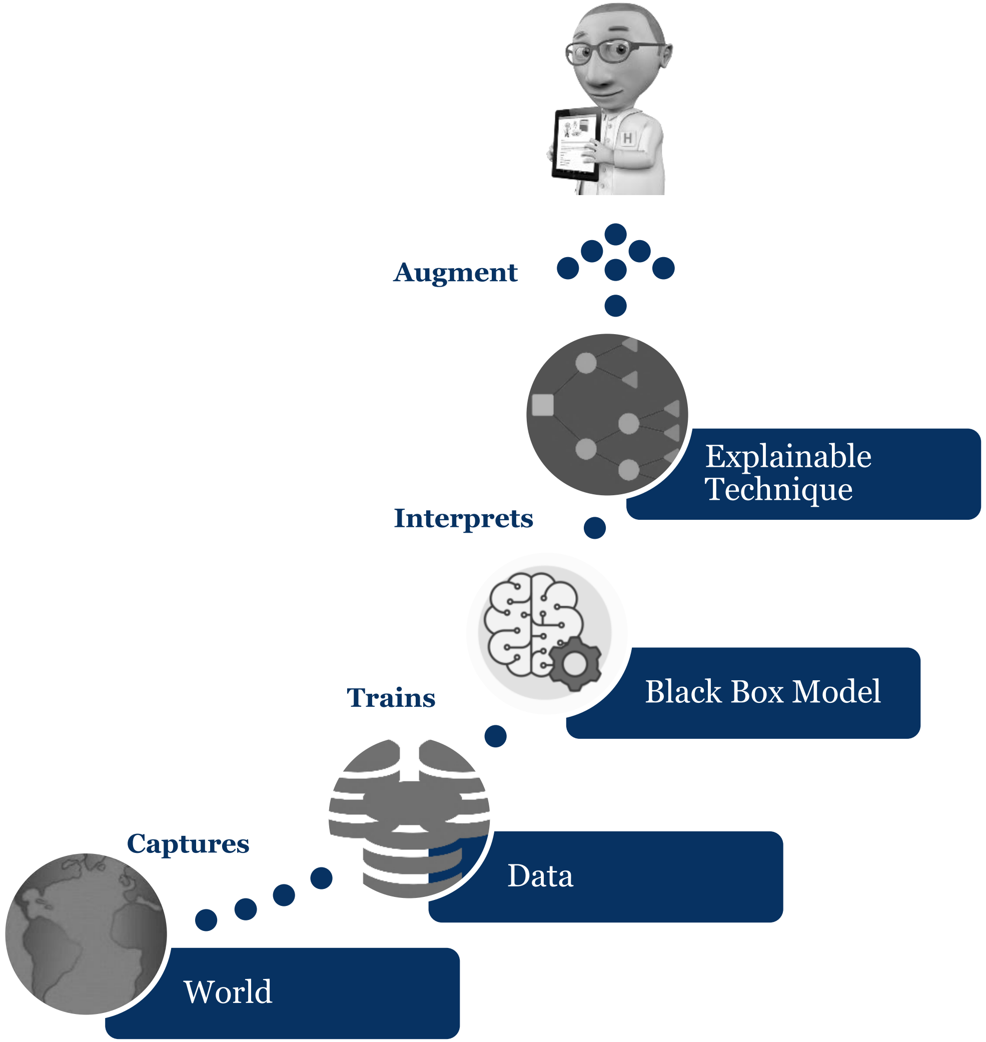
- Why did the AI system do that?
- Why didn't the AI system do something else?
- When did the AI system succeed?
- When did the AI system fail?
- When does the AI system give enough confidence in the decision that you can trust it?
- How can the AI system correct an error?



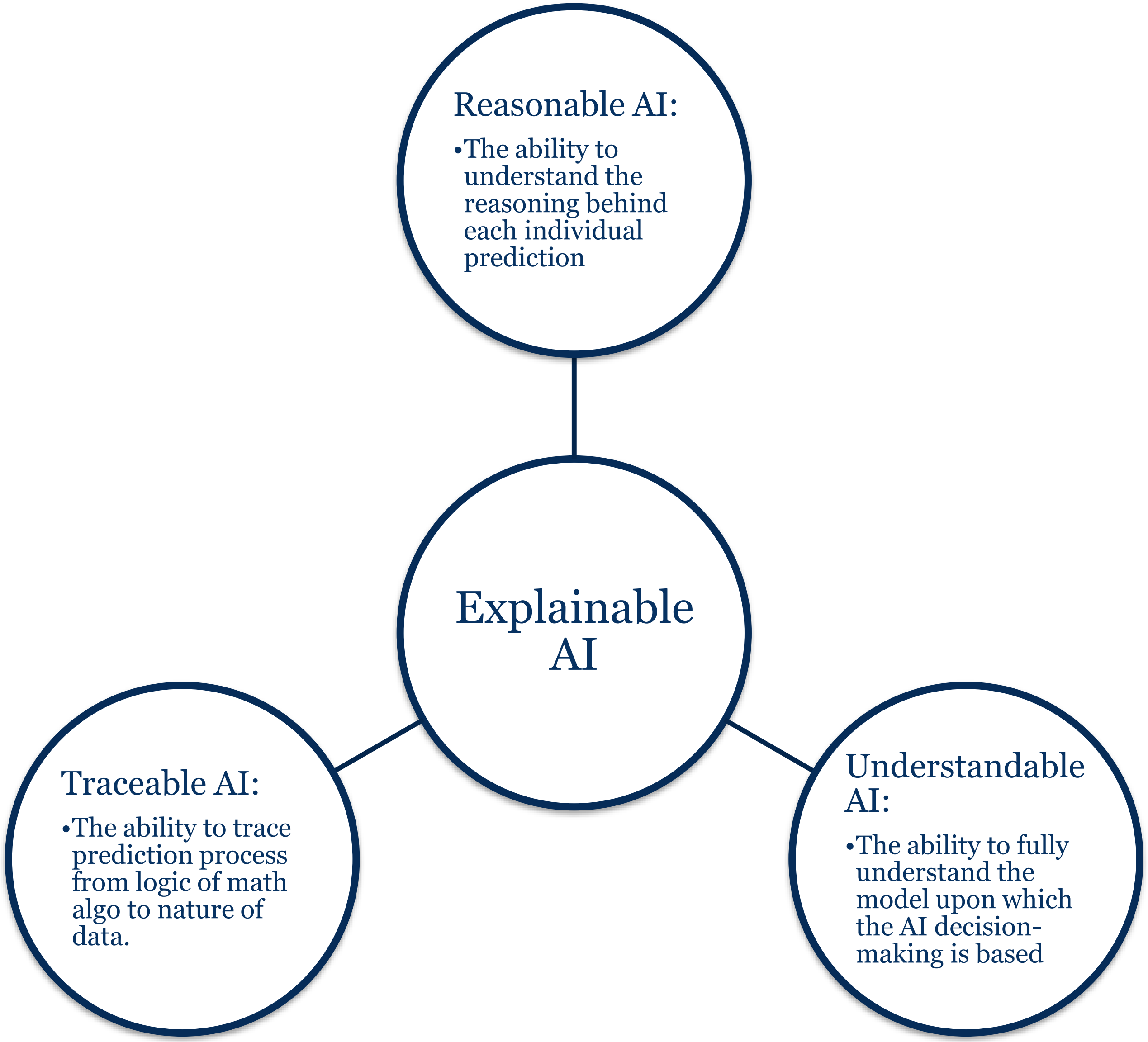
- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

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

XAI – Full Circle of Life



Explainable AI – Three Key Pillars



Explainable AI

Reasonable AI:

- The ability to understand the reasoning behind each individual prediction

Traceable AI:

- The ability to trace prediction process from logic of math algo to nature of data.

Understandable AI:

- The ability to fully understand the model upon which the AI decision-making is based

Principals for XAI

Designing	• AI with an eye to societal impact
Defining	• Standards for the provenance, use, and securing of data sets
Testing	• AI extensively before release
Using	• AI transparently
Monitoring	• AI rigorously after release
Fostering	• Workforce training and retraining
Protecting	• Data privacy
Establishing	• Tools and standards for auditing algorithms

Limitations for XAI

Feature importance

- Most model analysis stays in finding the important features, but they largely ignore the interactions between the features.

Problem domain

- Each business problem has its domain nuances and has defining contribution to the Explainability. If model analysis process has less of it, the outcome of explanation will be blurred or too generalized.

Data preprocessing

- Preprocessing (Dimensionality Reduction, Word Vectorization, etc.) obscure the original human meaning and makes data less informative.

Correlated input features

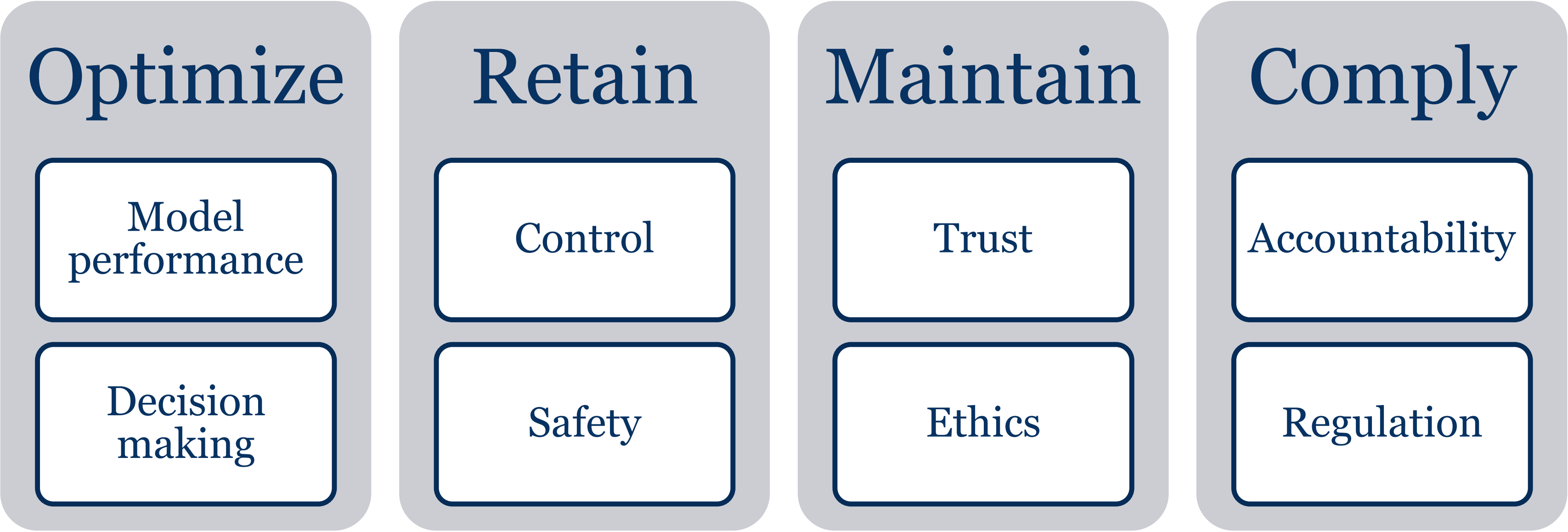
- During feature analysis and selection process, most of the correlations are either dropped to improve accuracy or swamped by other powerful features with more predictive powers. This hides the latent correlation factors from model and leads to incorrect explanation.

Type of prediction

- Straightforward Binary or Ordinal Classification and Linear Regression models are easier to explain as they have natural direction for decision. However, models like multi class classification do not have inherent order are difficult to explain, unless it comes down to ‘One vs All’ model.

Business Benefits from XAI

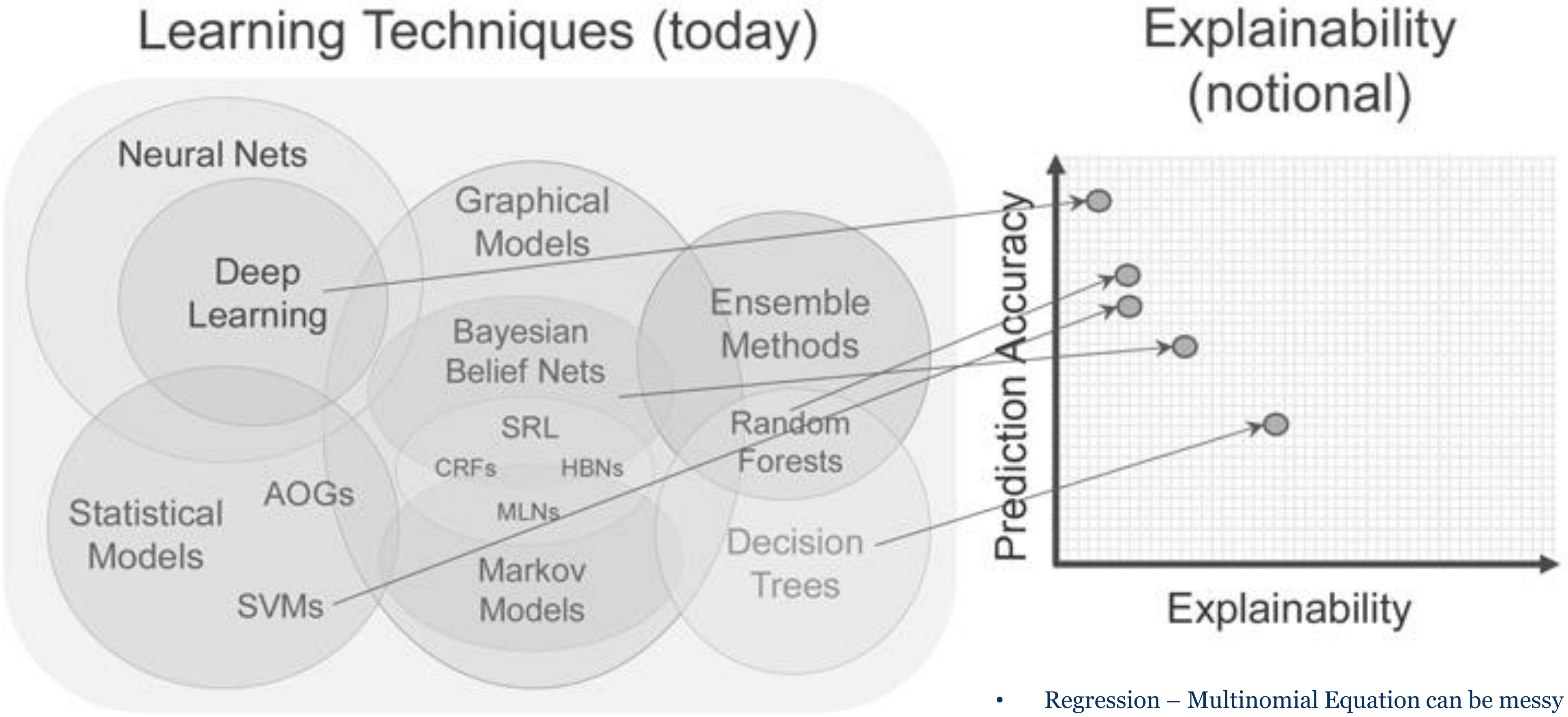
“Explainable AI is Responsible AI”



Optimize	Retain	Maintain	Comply
Model performance	Control	Trust	Accountability
Decision making	Safety	Ethics	Regulation

“Right to Explanation” – GDPR

Current State of AI Models w.r.t. XAI vs Performance

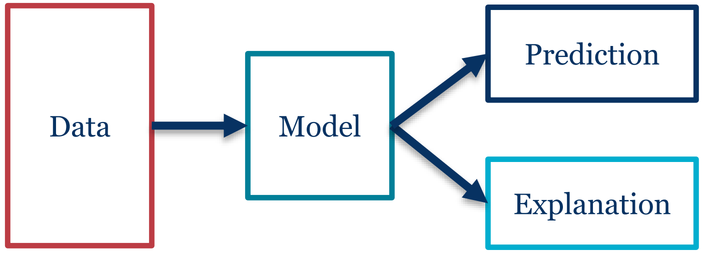


- Regression – Multinomial Equation can be messy
- Random Forrest – Multiple Tree and their Data Set and Voting
- SVM – Kernel and Data Partition effect on Feature
- K Mean – Nature of Centroid don’t describe cluster well.
- NN – Hidden Nodes and their way of creating features

Techniques for XAI

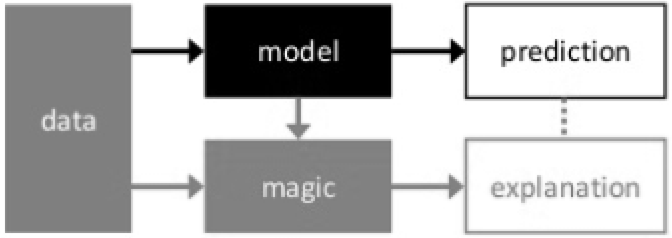
Model Specific Techniques

- Deals with inner working of Algo/Model to interpret its results



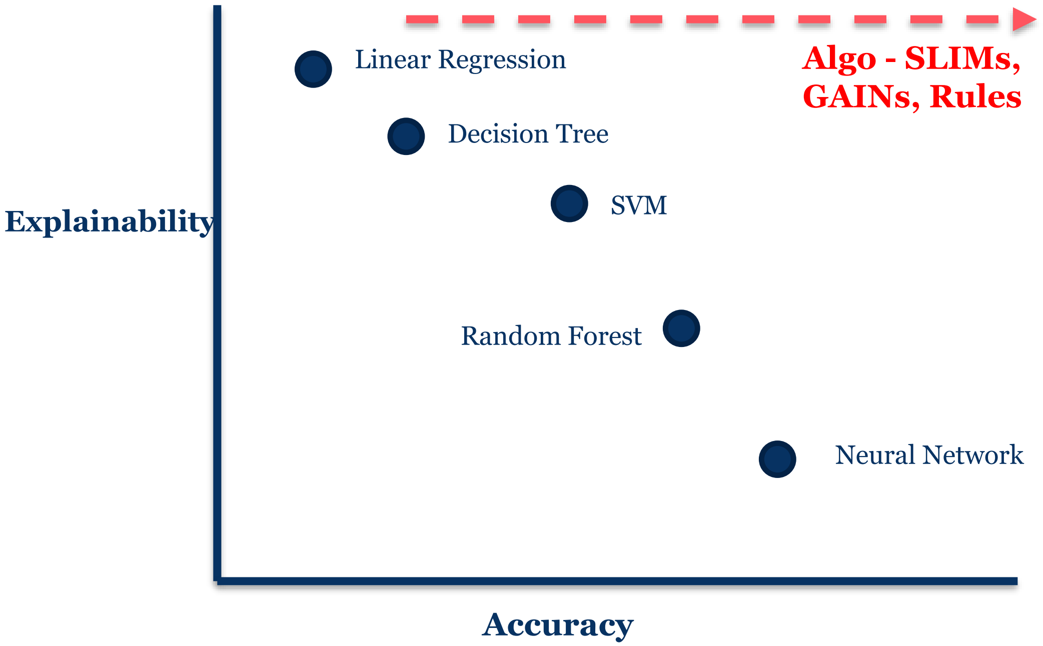
Model Agnostic Techniques

- Deals with analyzing the feature and its relationships with its output.

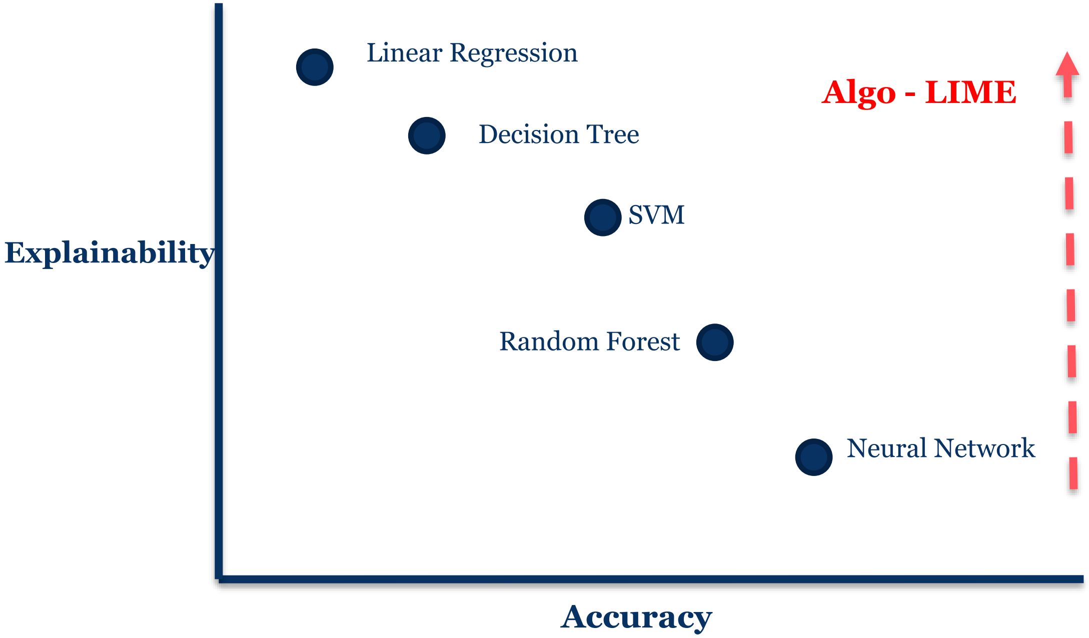


Model Specific vs Agnostic – Approach

Model Specific Approach



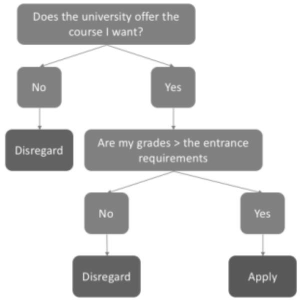
Model Agnostic Approach



XAI - Model Specific Techniques

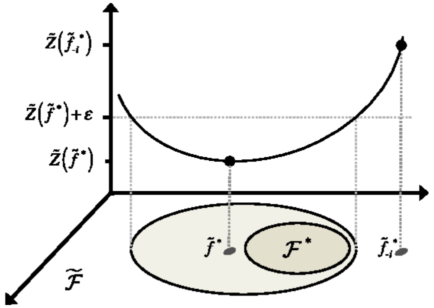
Tree Interpretation

- Decision Tree is most interpretable algorithm because its simple information gain mechanism of building the tree. Random Forest on other has multiple small tree with dataset variation and votes for final decision, which is based on majority.
- There are open source project – Tree Interpreter available but deep domain and deep algorithm knowledge are required to study its output.


Does the university offer the course I want?
No
Yes
Disregard
Are my grades > the entrance requirements
No
Yes
Disregard
Apply

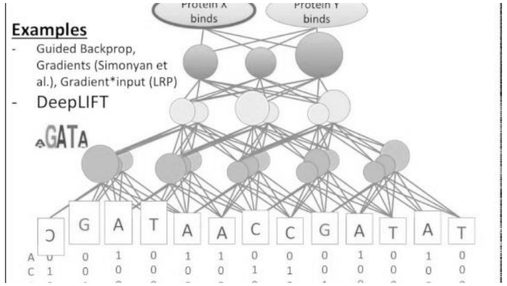
Supersparse Linear Integer Models (SLIM)

- SLIM is a discrete optimization problem that minimizes the 0-1 loss to encourage a high level of accuracy, regularizes the L0-norm to encourage a high level of sparsity, and constrains coefficients to a set of interpretable values.



Deep Neural Network

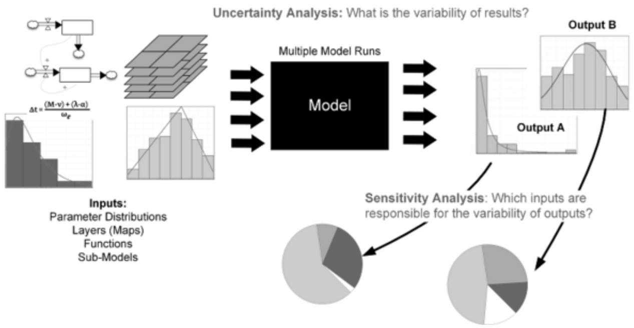
- Deep Lift - A method for decomposing the output prediction of a neural network on a specific input by backpropagation the contributions of all neurons in the network to every feature of the input. Deep LIFT compares the activation of each neuron to its 'reference activation' and assigns contribution scores according to the difference. Using this, This can give important sequence of input data.


Protein X binds
Protein Y binds
Examples
- Guided Backprop, Gradients (Simonyan et al.), Gradient*input (LRP)
- DeepLIFT
AGATA
C G A T A A C C G A T A T
A 0 0 1 0 1 1 0 0 0 1 0 1 0
C 1 0 0 0 0 0 1 1 0 0 0 0 0

XAI - Model Agnostic Techniques

Sensitivity Analysis

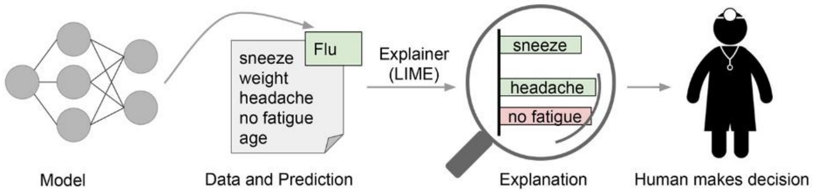
- It try to analyze the effect of single input feature alteration on its model output. This given linear approximation of model response.
- This approach is often extended to Partial Dependence Plots (PDP) or Individual Conditional Expectation (ICE) Plots to give global graphical representation of single



Uncertainty Analysis: What is the variability of results?
Multiple Model Runs
Model
Output A
Output B
$\Delta t = \frac{(M \cdot v) + (\lambda \cdot \alpha)}{\omega_\varepsilon}$
Inputs:
Parameter Distributions
Layers (Maps)
Functions
Sub-Models
Sensitivity Analysis: Which inputs are responsible for the variability of outputs?

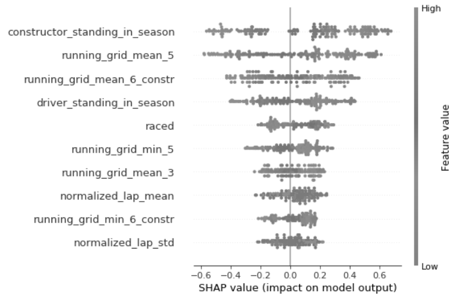
Local Interpretable Model Explanation – LIME

- It try to find out Feature Importance by capturing Feature Interaction (Correlation and Covariance Analysis) between features and output using a linear model between Features. It used Perturbance technique which observer Prediction deviation based on one feature modification.

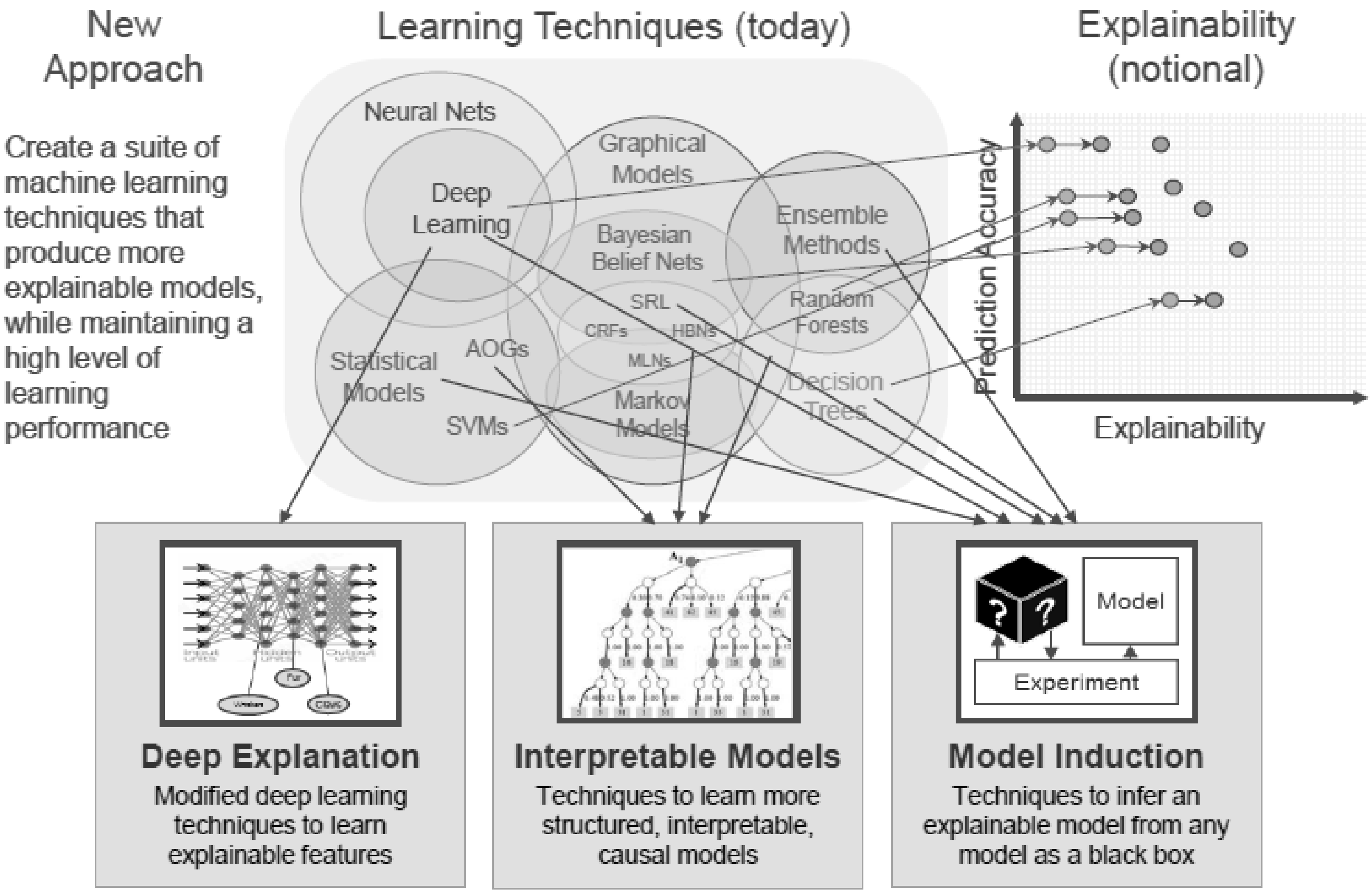


Shapley Addictive Explanations – SHAP

- SHAP is an Additive feature attribution methods which assigns a value to each feature for each prediction (i.e. feature attribution); the higher the value, the larger the feature's attribution to the specific prediction.



Future Status of XAI Systems



New Approach

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

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

Thank You

Author – Saurabh Kaushik
Twitter - @saurabhkaushik

References

- https://www.darpa.mil/program/explainable-artificial-intelligence
- https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/
- https://arxiv.org/pdf/1708.08296.pdf
- https://bons.ai/blog/explainable-ai-machine-learning-systems
- https://medium.com/@BonsaiAI/explainable-ai-3-deep-explanations-approaches-to-xai-1807e251e537
- https://www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions/explainable-ai.html
- https://aibusiness.com/explainable-algorithms-rainbird-ai/
- https://www.accenture.com/us-en/blogs/blogs-why-explainable-ai-must-central-responsible-ai
- https://www.accenture.com/us-en/blogs/blogs-responsible-ai