Explainable AI

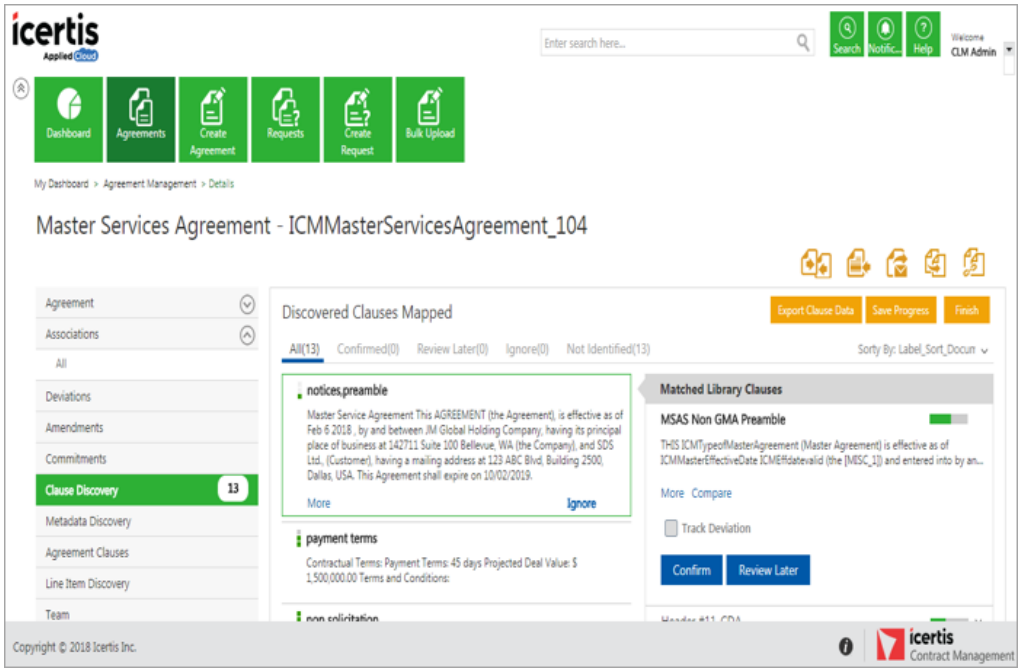
Yogesh Kulkarni

November 16, 2020

Introduction

Practical Scenario

Icertis Discover AI



(Ref: https://www.icertis.com/contract-management-software/ai-applications/discoverai/)

- Image/PDF Conversion
- Clause Discovery, by Delineation and Classification
- Attribute Discovery, by NER
- Tables Discovery, Image processing

A Sample Digitization Project

- Number of Contracts to be Digitized : 60K
- Number of Clauses to be discovered : 32 (“Term”, “Warranty”, . . .)
- Number of Attributes to be discovered: 12 (“Effective Date”, “Contract Value”,. . .)
- Time : 3 months

Steps

- Build AI Engine using annotated samples to extract attributes
- Show results of training to the Customer and get approval
- Start extractions (production) for all the contracts, in batches . . .

Halfway through the production . . .

A sample dialog

- **Customer**: The extractions are looking ok, but . . .
- **AIML team**: But??
- **Customer**: Why are you digitizing these ‘DELLA’ contracts?
- **AIML team**: Meaning?
- **Customer**: You should not digitize contracts having ‘DELLA Corporation’ in the footer.
- **AIML team**: But you never told us. No mention of any such rules in SOW also..
- **Customer**: We are just asking you to NOT process these, not adding to your work/scope!!
- **AIML team**: But . . .
- **Customer**: Shouldn’t be difficult, right? AI will find out . . .

(we keep hearing . . .)

AI **will** find out!!

(sometimes . . .)

AI **should** find out!!

(what if . . .)

AI **can not** find out!!

(and worse, if . . .)

AI **finds it and finds wrong stuff**!!

(bottom line... we need to understand that ...)

AI **is not Magic**!!

(But still, you need to explain me)

Why AI gave this result!!

(that's)
Explainable AI!!

Need for Explainable AI

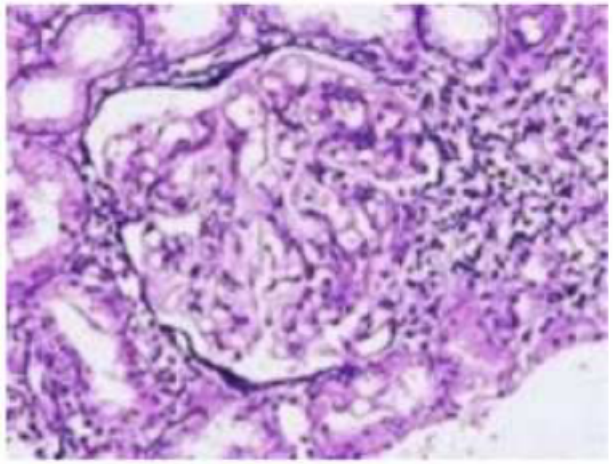
Explainable AI is essential for customers to understand and trust the decisions by AI.

(Ref: Explainable AI in Industry, KDD 2019 Tutorial,Sahin Cem Geyik, Krishnaram Kenthapadi & Varun Mithal)

Wrong decisions: Costly and dangerous

"Autonomous car crashes, because it wrongly recognizes ..."

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



(Ref: Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18)

AI as Black Box

- 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?

(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

Need for Explainable AI

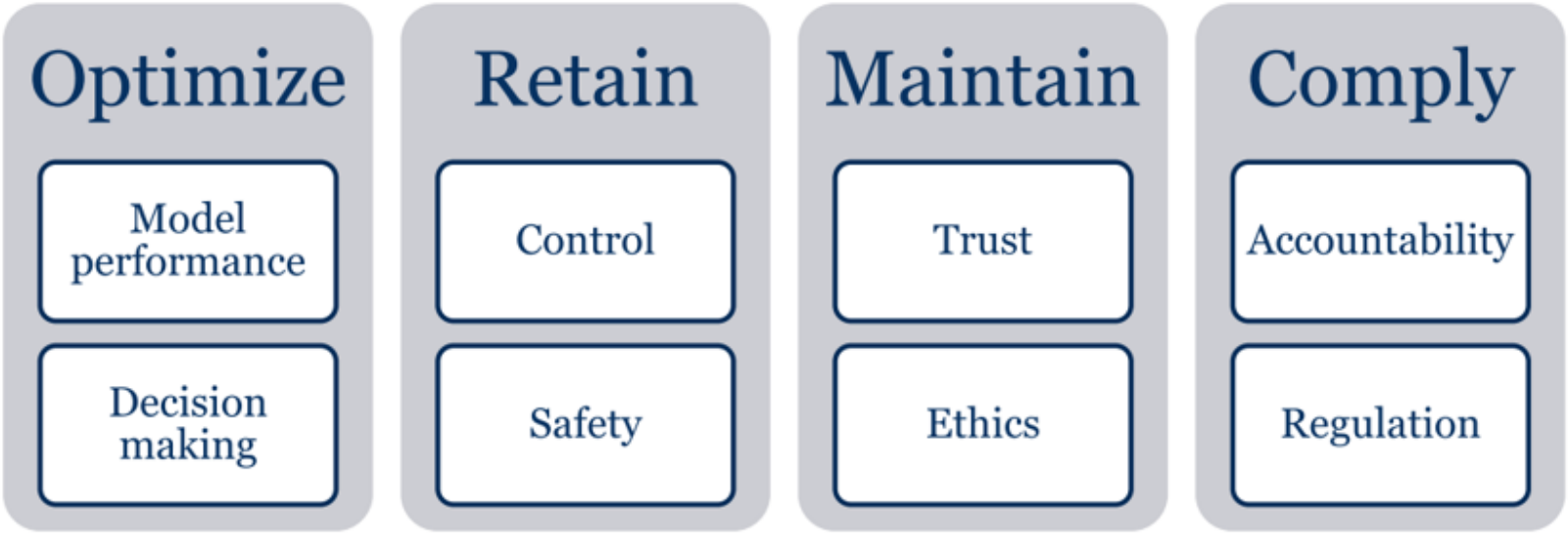
Three Key Pillars of Explainable AI

- **Reasonable AI**: The ability to understand the reasoning behind each individual prediction
- **Traceable AI**: The ability to trace prediction process from algorithm to data.
- **Understandable AI**: The ability to fully understand the AI decision-making is based

(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

Business Benefits

"Explainable AI is Responsible AI"



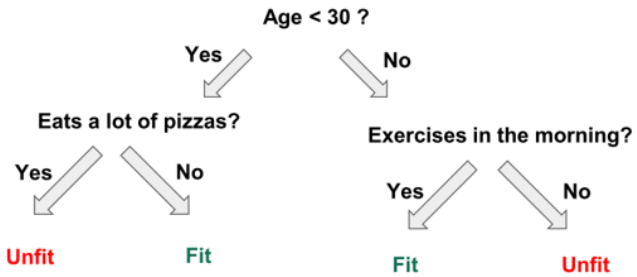
"Right to Explanation" – GDPR
(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

Achieving Explainable AI

Prediction explanations

- Important features
- Features weights
- Decision tree plotting

Build an interpretable model using libraries like LIME, SHAP, etc.



(Ref: Explainable AI in Industry, KDD 2019 Tutorial,Sahin Cem Geyik, Krishnaram Kenthapadi & Varun Mithal)

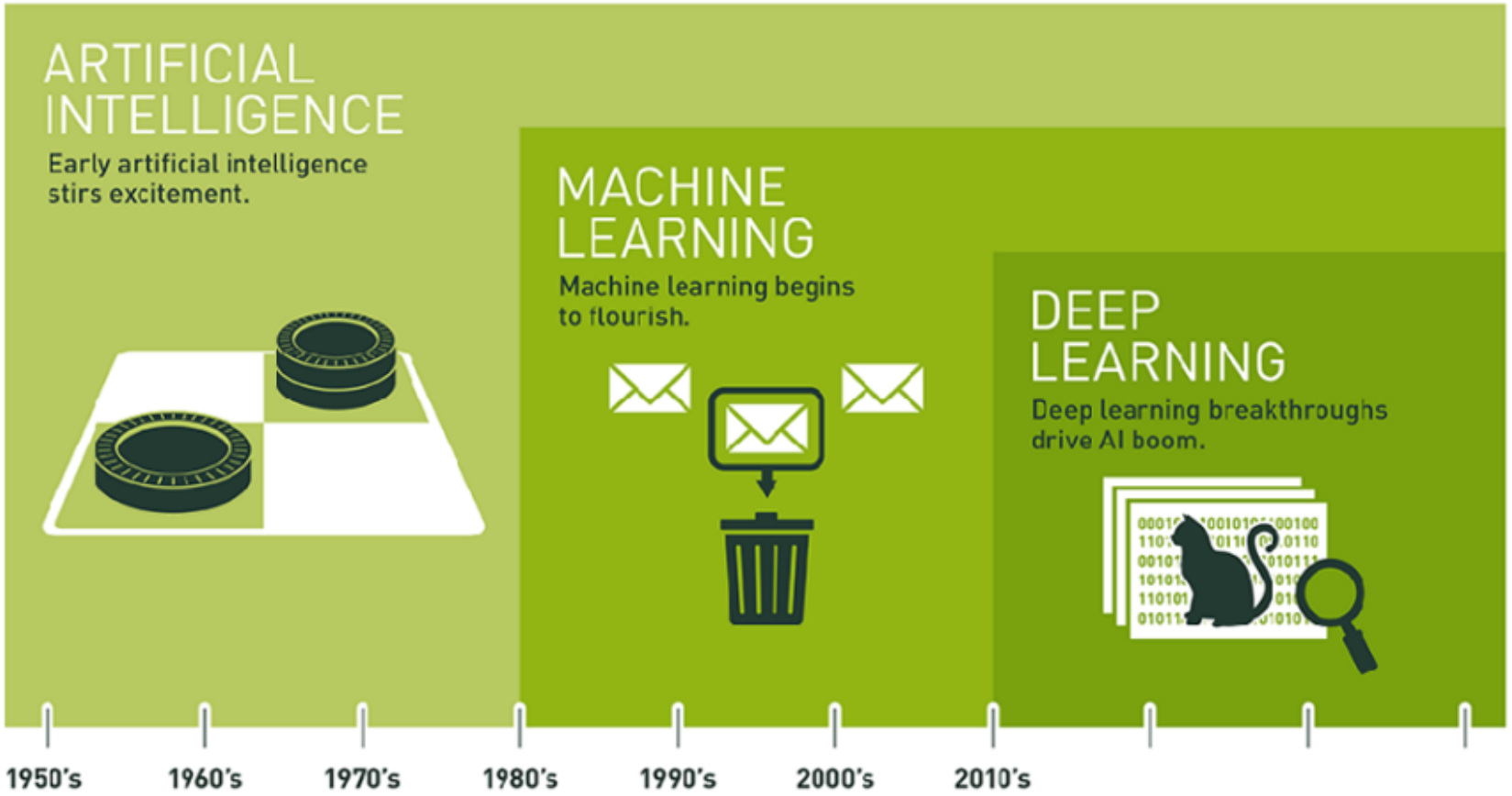
"Interpret-ability is the degree to which a human can understand the cause of a decision."

(Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences")

So, What (exactly) is AI?

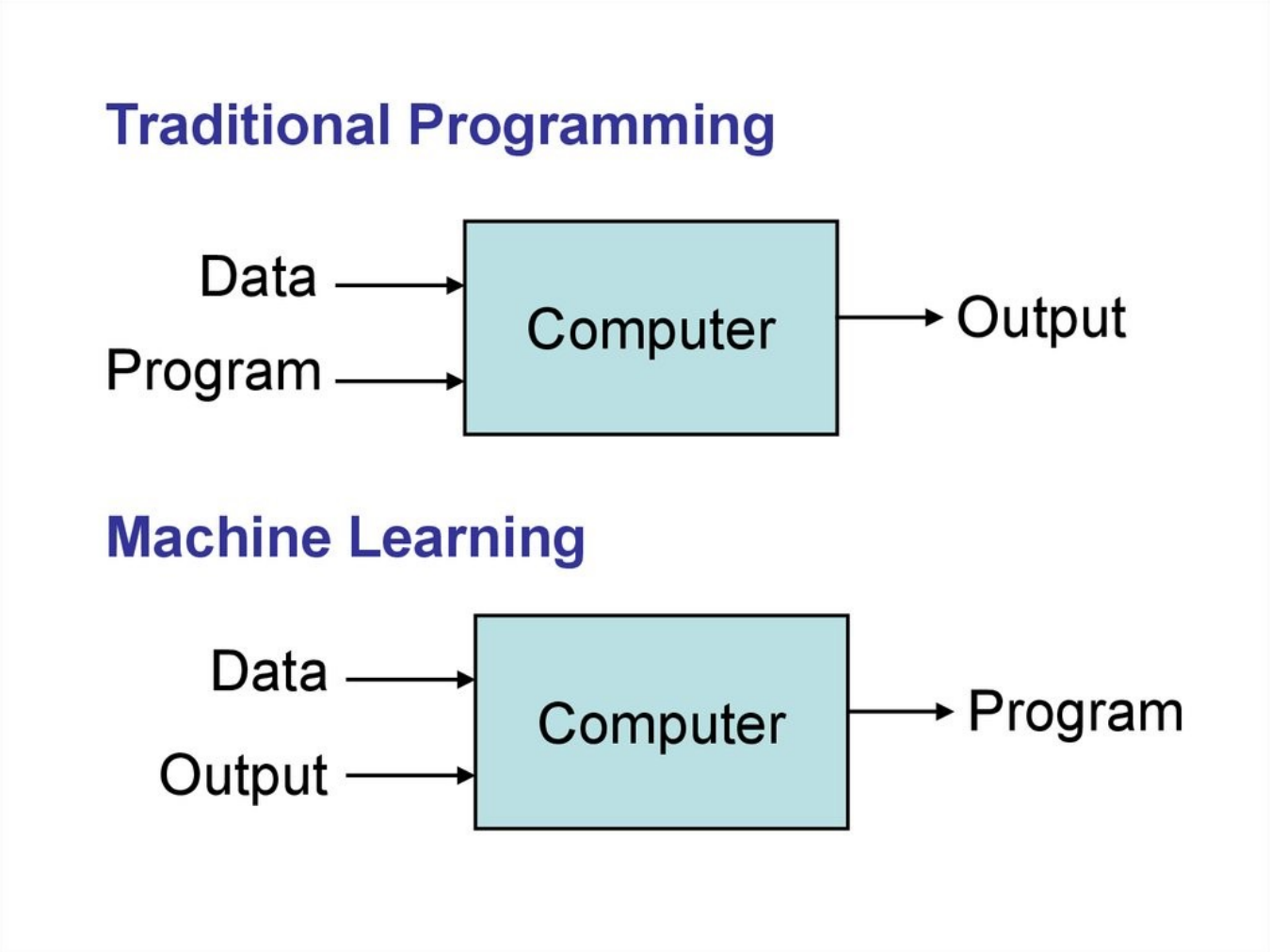
(typical understanding)

If Machines show intelligence, like Humans that's AI



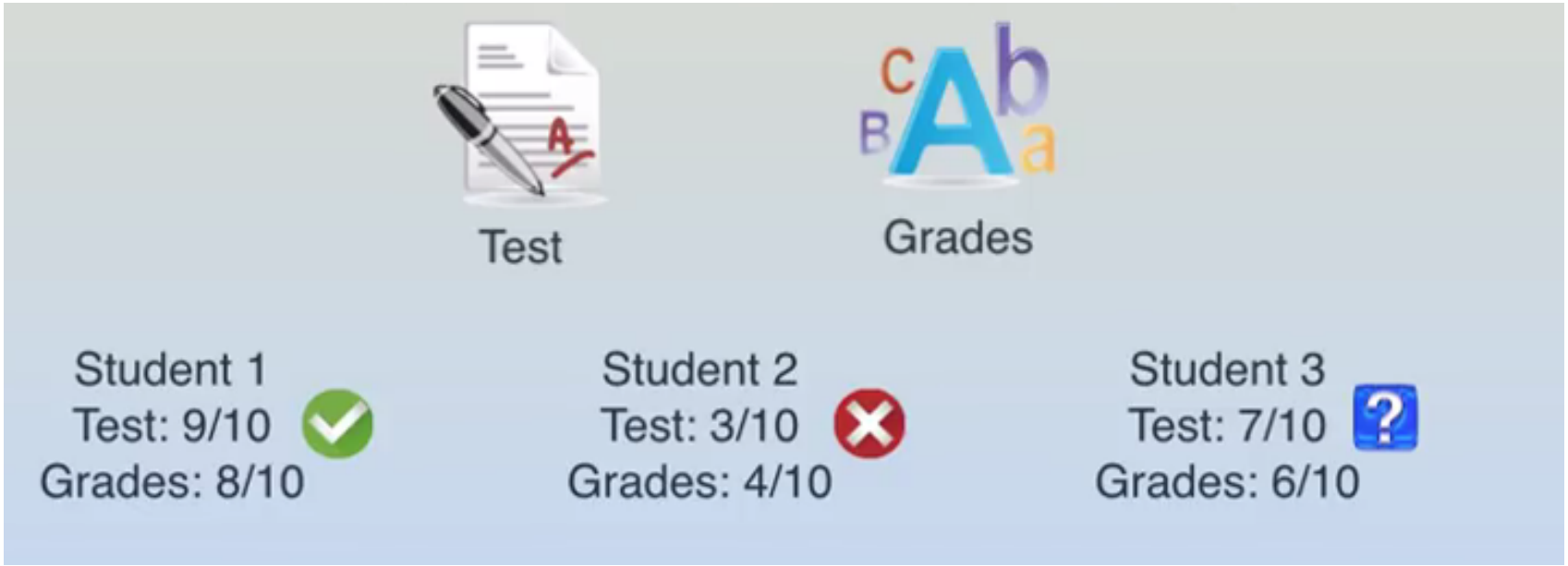
(Ref: Nvidia blog: Artificial Intelligence)

Difference



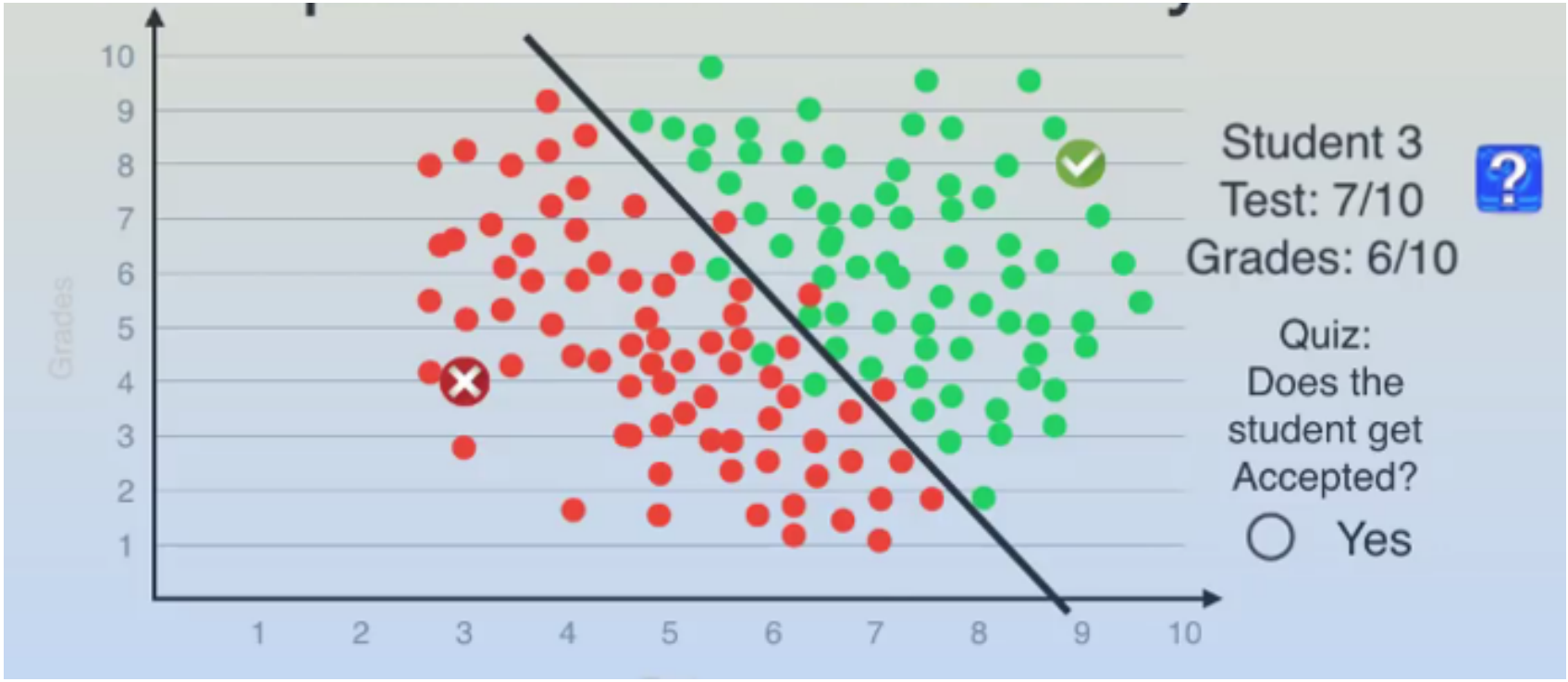
(Ref: Machine Learning - Luis Serrano - Youtube)

Supervised: Linear



(Ref: Machine Learning - Luis Serrano - Youtube)

Supervised: Linear



(Ref: Machine Learning - Luis Serrano - Youtube)

Supervised: Non-Linear

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Quiz: Between Gender and Age, which one seems more decisive for predicting what app will the users download?

- ○ Gender
- ○ Age

(Ref: Machine Learning - Luis Serrano - Youtube)

Supervised: Non-Linear

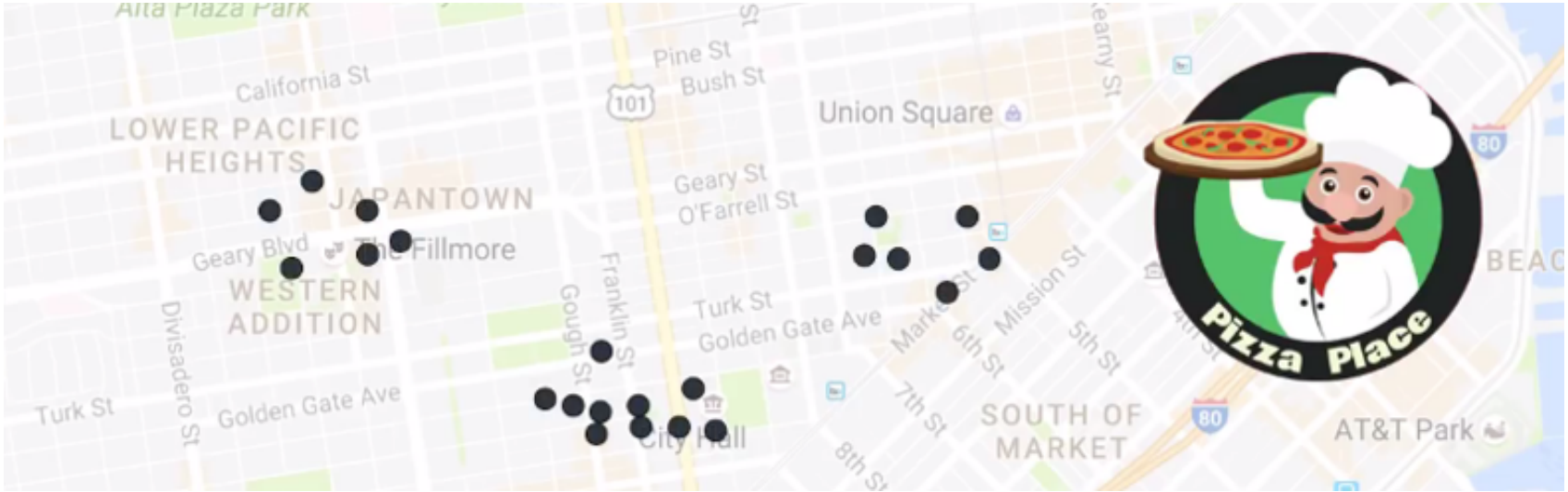
Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Quiz: Between Gender and Age, which one seems more decisive for predicting what app will the users download?

- ○ Gender
- ○ Age

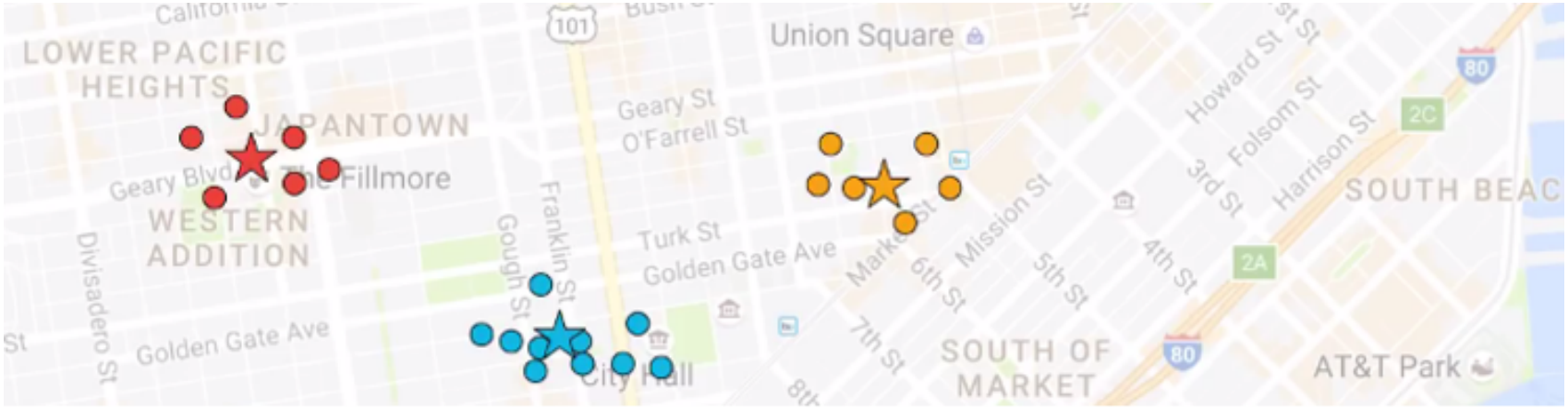
(Ref: Machine Learning - Luis Serrano - Youtube)

Unsupervised



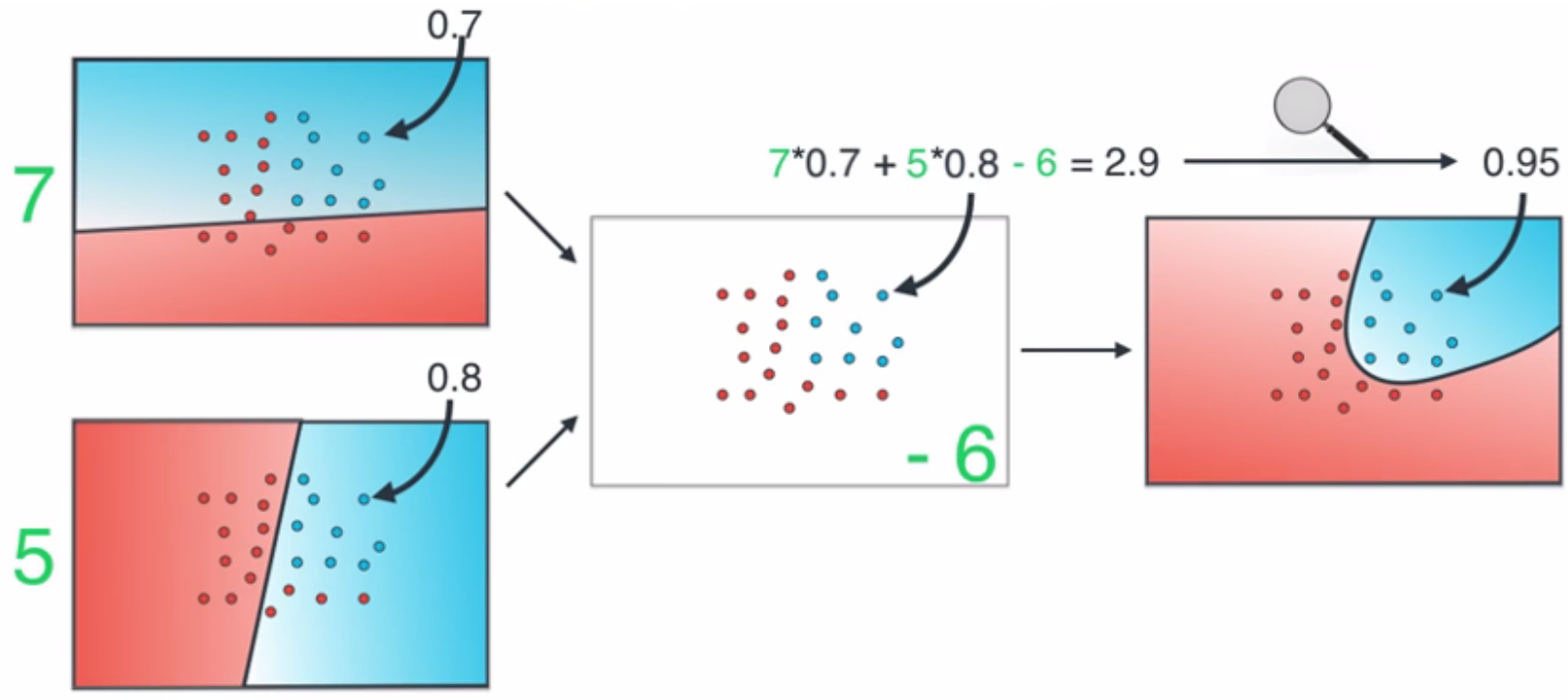
(Ref: Machine Learning - Luis Serrano - Youtube)

Unsupervised



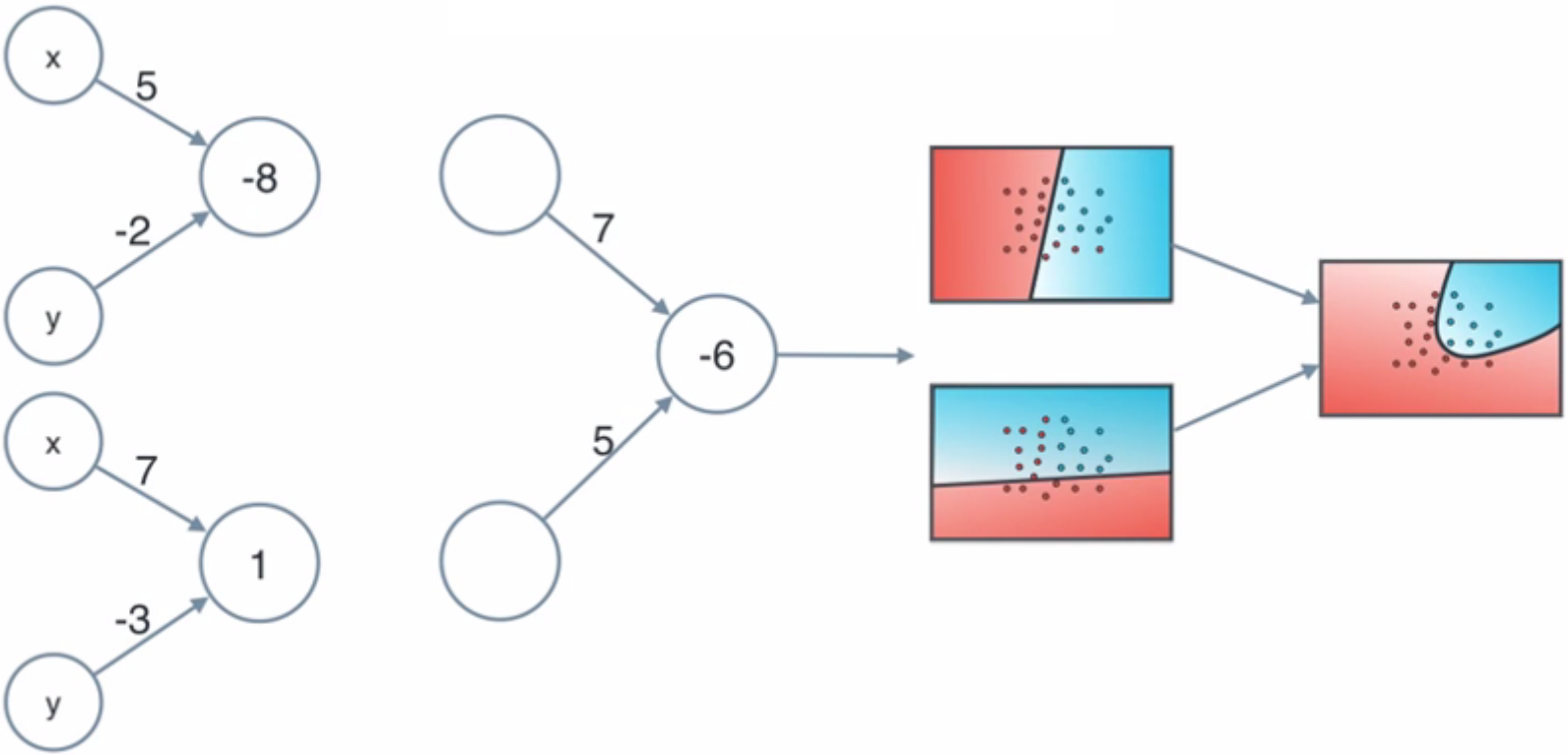
(Ref: Machine Learning - Luis Serrano - Youtube)

How Neural Networks capture Non-Linearity?



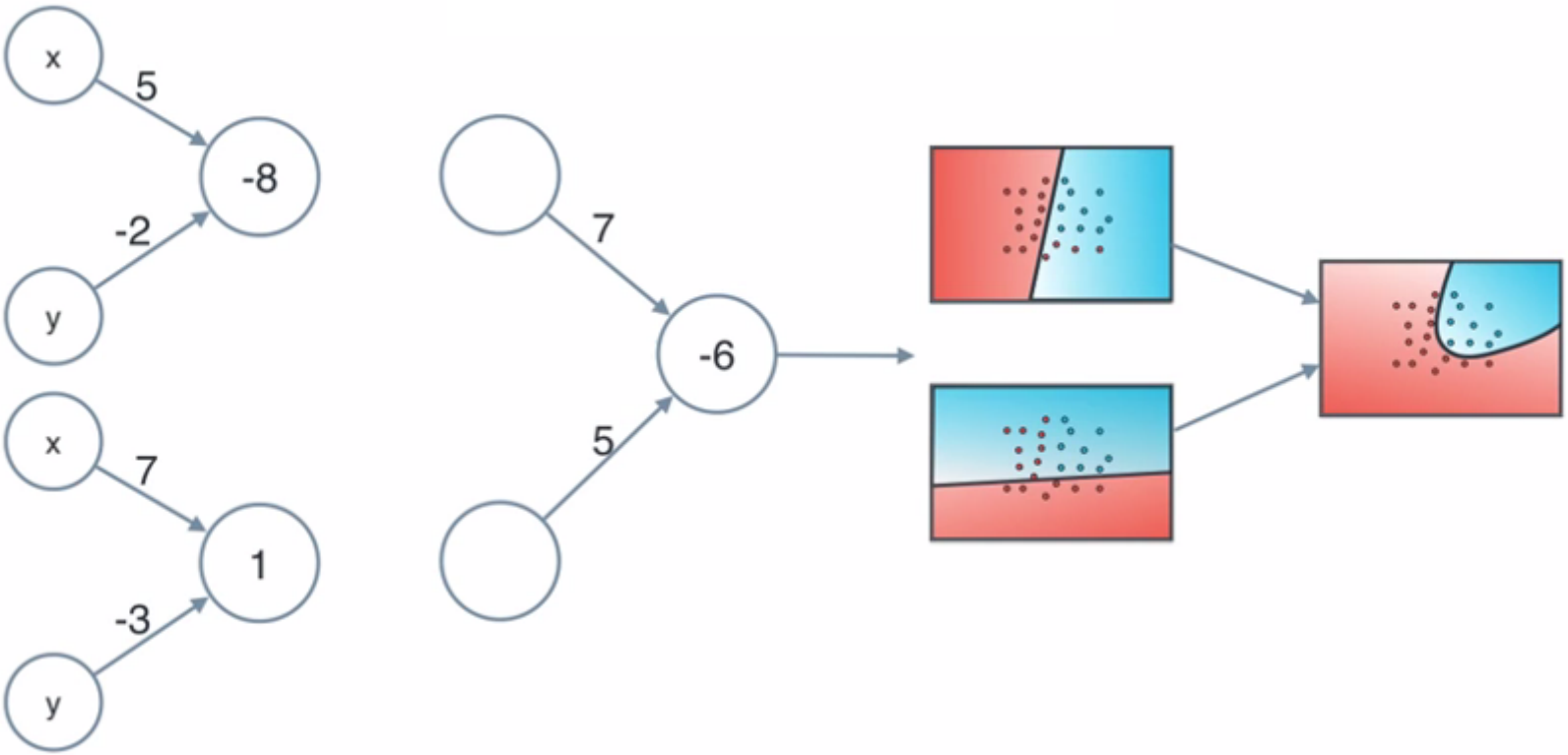
(Ref: Machine Learning - Luis Serrano - Youtube)

How Neural Networks capture Non-Linearity?



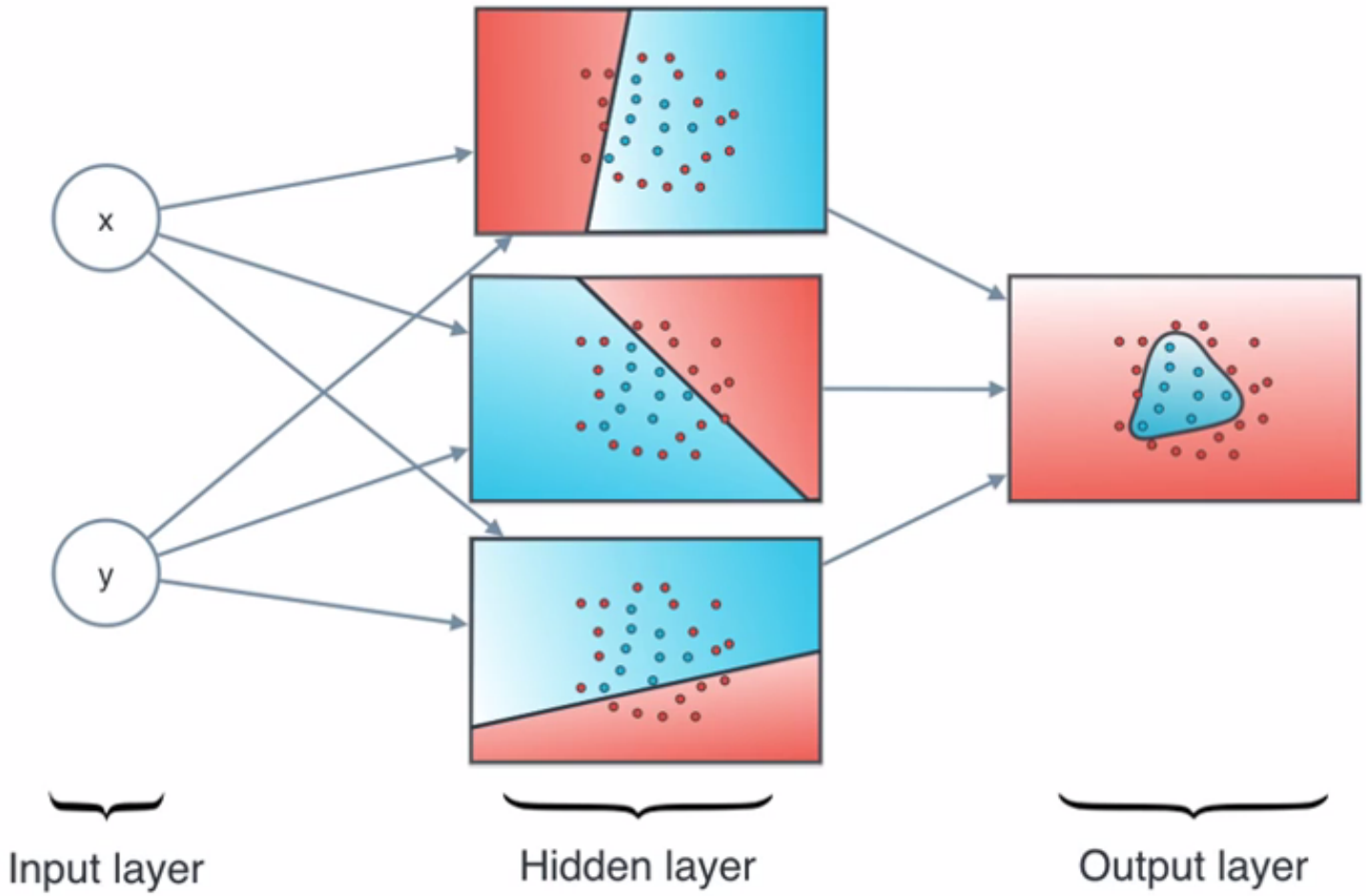
(Ref: Machine Learning - Luis Serrano - Youtube)

How Neural Networks capture Non-Linearity?



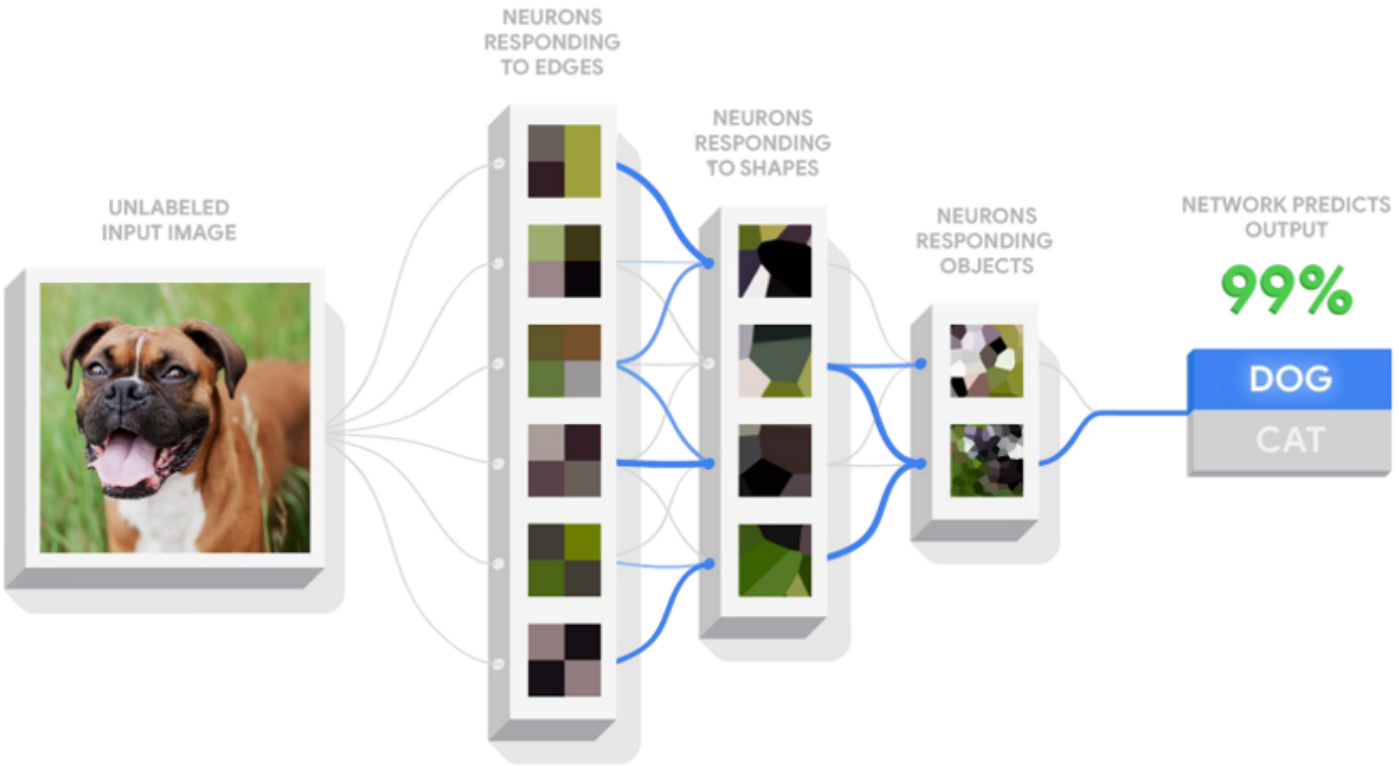
(Ref: Machine Learning - Luis Serrano - Youtube)

How Neural Networks capture Non-Linearity?



(Ref: Machine Learning - Luis Serrano - Youtube)

whole work-flow

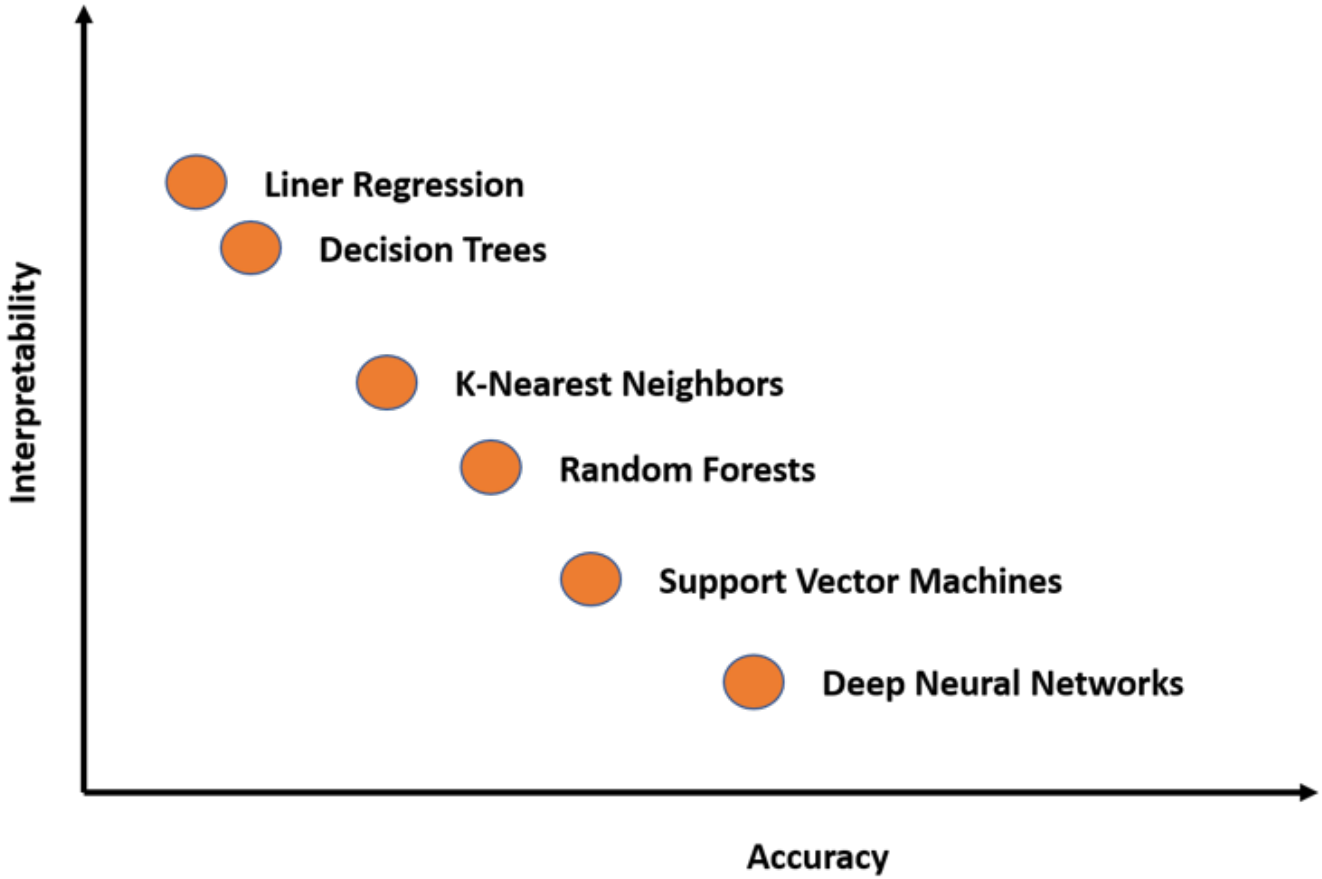


Explain-ability of AI Algorithms

- **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 Means**: Nature of Centroid don't describe cluster well.
- **NN**: Hidden Nodes and their way of creating features

(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

But ...



(Ref: Application of artificial intelligence in gastroenterology, April 2019,World Journal of Gastroenterology 25(14):1666-1683)

Techniques for Explain-ability

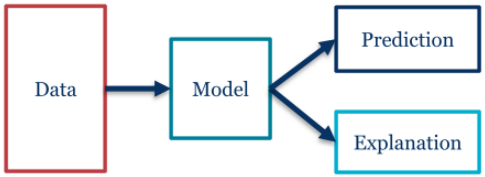
Explain-ability of AI Algorithms

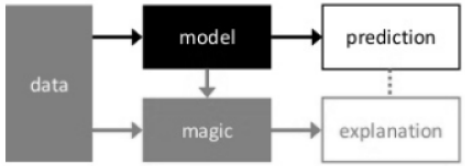
Model Specific Techniques

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

Model Agnostic Techniques

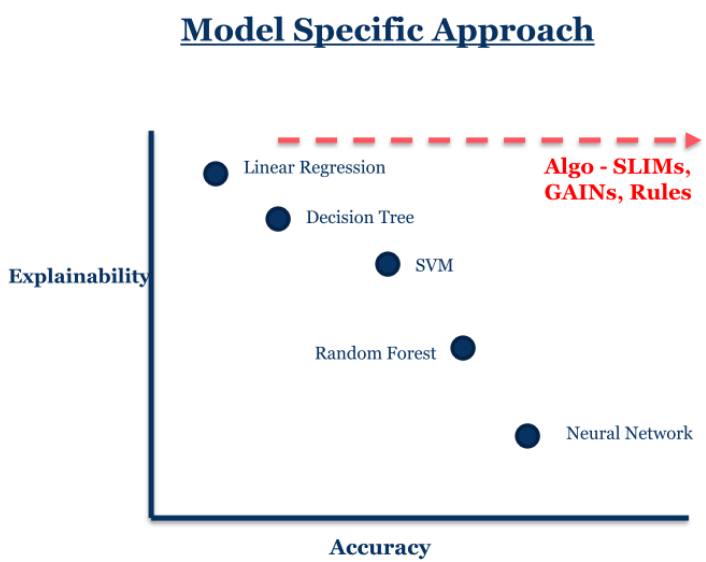
- Deals with analyzing the feature and its relationships with its output.

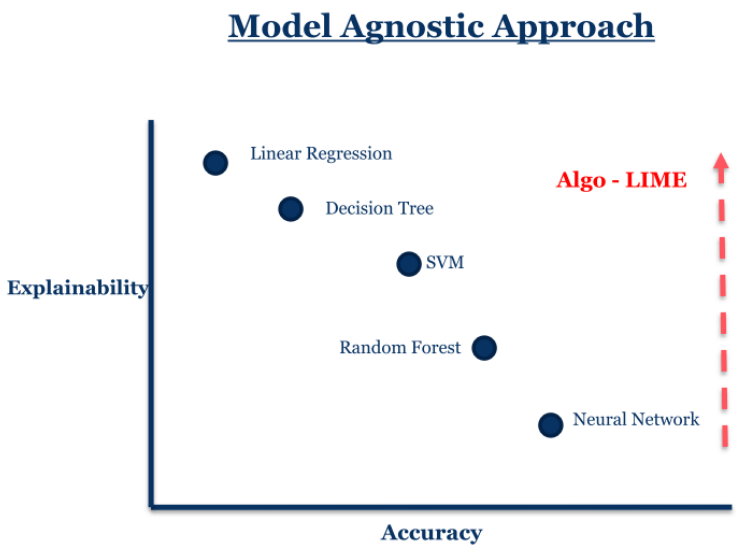




(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

Model Specific vs Agnostic – Approach



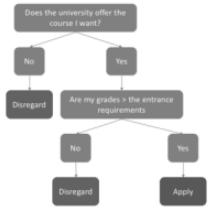


(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

Model Specific Techniques

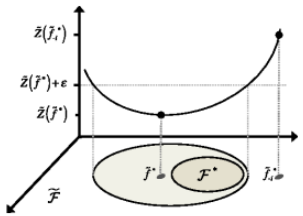
Tree Interpretation

- 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.



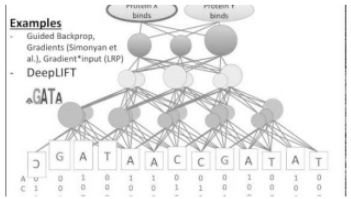
Supersparse Linear Integer Models (SLIM)

- 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.

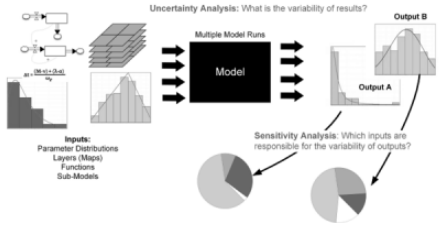


(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

Model Agnostic Techniques

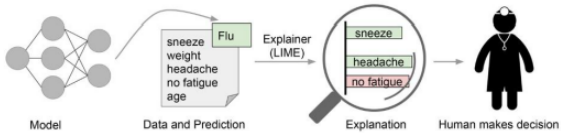
Sensitivity Analysis

- 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



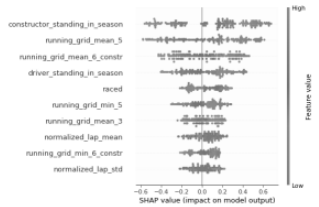
Local Interpretable Model Explanation – LIME

- 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

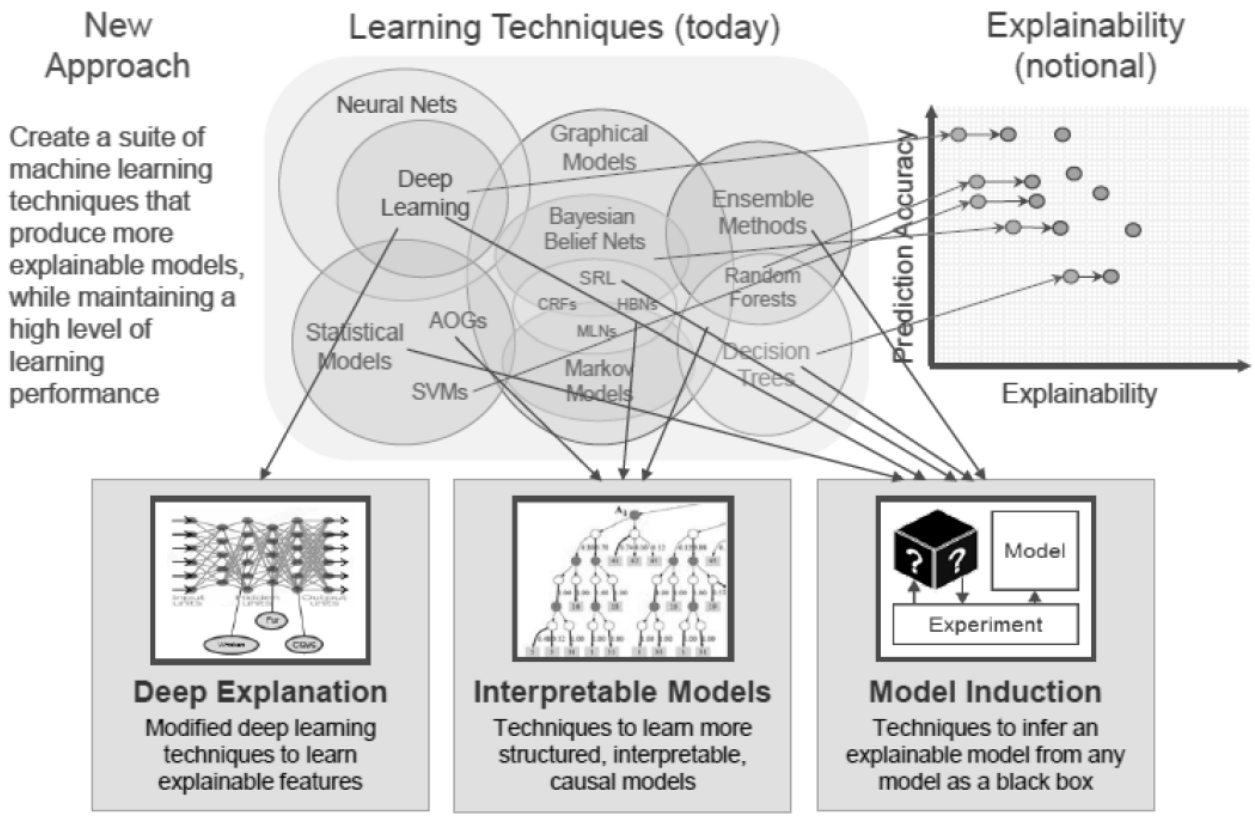
- SHAP is an Additive feature attribution methods which assigns a value to each feature for each prediction (i.e. feature attribution); the higher the value, the larger the feature's attribution to the specific prediction.



(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

Conclusion

Future Techniques



(Ref:Explainable AI (XAI) – A Perspective, Saurabh Kaushik)

So, Finally ...

AI is just a set of algorithms, which (predominantly) finds patterns from given Inputs as well as Outputs

For example: to find out clause "Term" ...

- Need to supply AI algorithm with 100s/1000s of "Term" clause samples
- Gets trained on patterns, word frequencies, context words
- Stores this information as "model"
- Can be used to classify unseen clause, whether "Term" or not.

Quiz

If you want to find something “new”, what would be needed?

Summary

- AI-ML-DL approaches are non-deterministic (but (. . .)?)
- For good results, need good annotated data and lots of it
- Annotations need to be perfect ("Gold"), else Garbage-In-(. . .)
- Data should cover all possible variations
- AI-ML-DL just fits the data, but just that, it does it automatically!!
- ML has better explain-ability than DL (why?)

Btw, thoughts to ponder on ...

- 'AI is biased'
- 'Explainable AI' is to understand how/why it is biased
- But then ...
- Human are biased too ...
- 'Explainable Humans'... the next topic?

Thanks ... yogeshkulkarni@yahoo.com