Explainable AI

Presentation · September 2018

CITATIONS
0

READS
243

1 author:

Manojkumar Parmar
Robert Bosch GmbH
7 PUBLICATIONS **0** CITATIONS

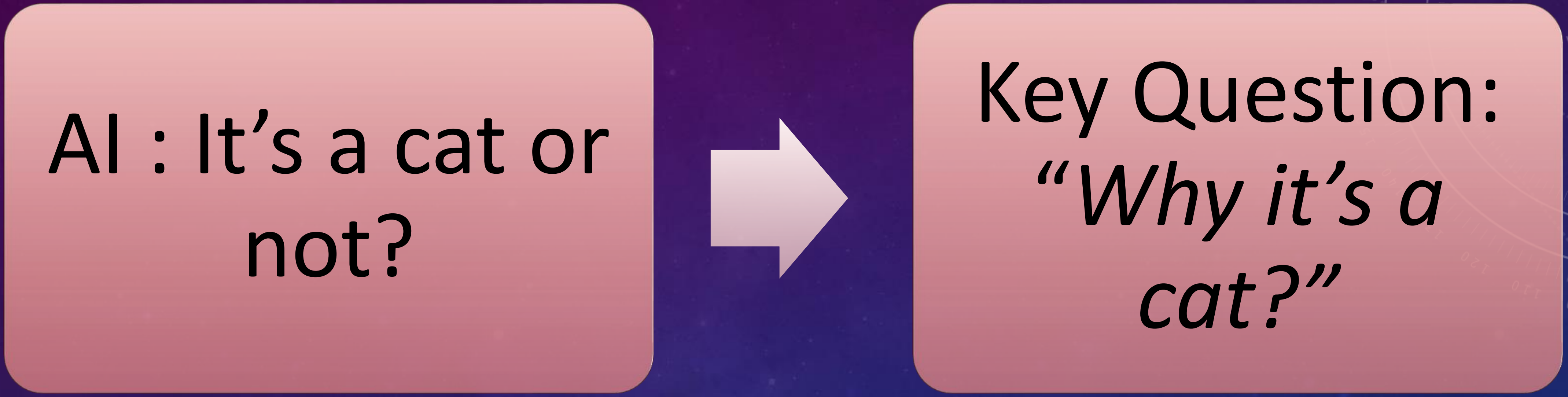
SEE PROFILE

EXPLAINABLE AI (XAI)

MANOJKUMAR PARMAR

ROBERT BOSCH ENGINEERING AND BUSINESS SOLUTIONS PRIVATE LIMITED, INDIA

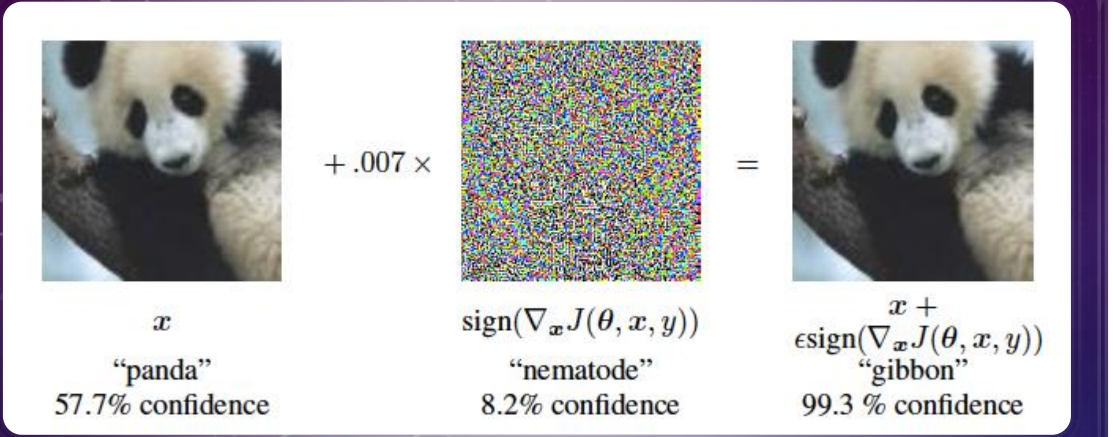
LIGHTNING TALK AT ICACCI'18 ON 21ST SEPTEMBER 2019



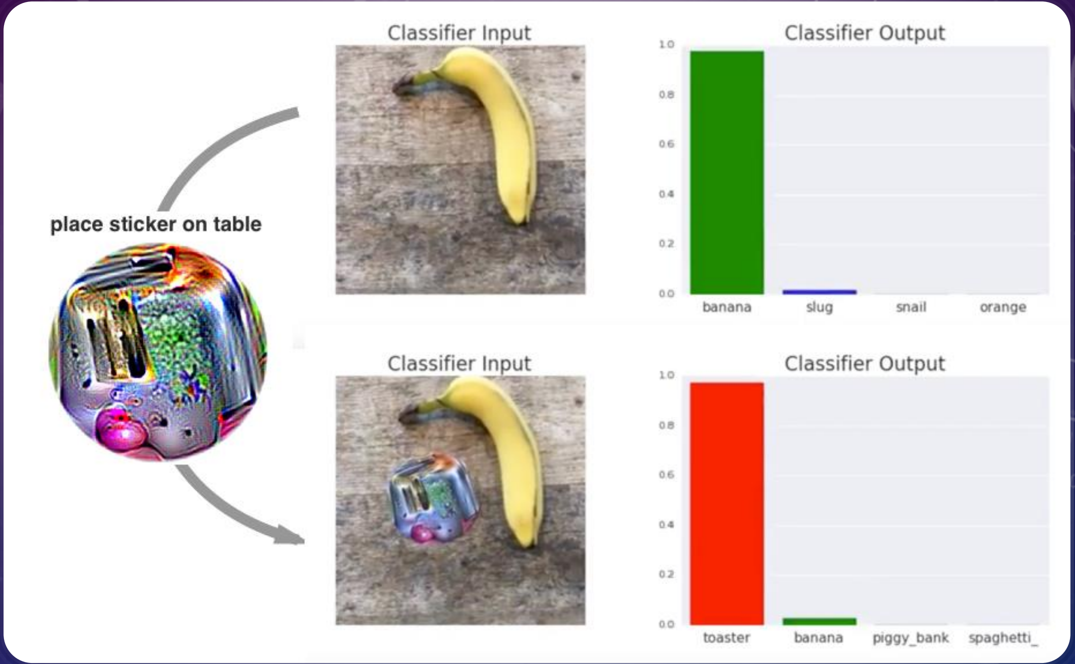
AI : It’s a cat or not?

Key Question: *“Why it’s a cat?”*

SIMPLE QUESTION BUT FAR REACHING IMPLICATION

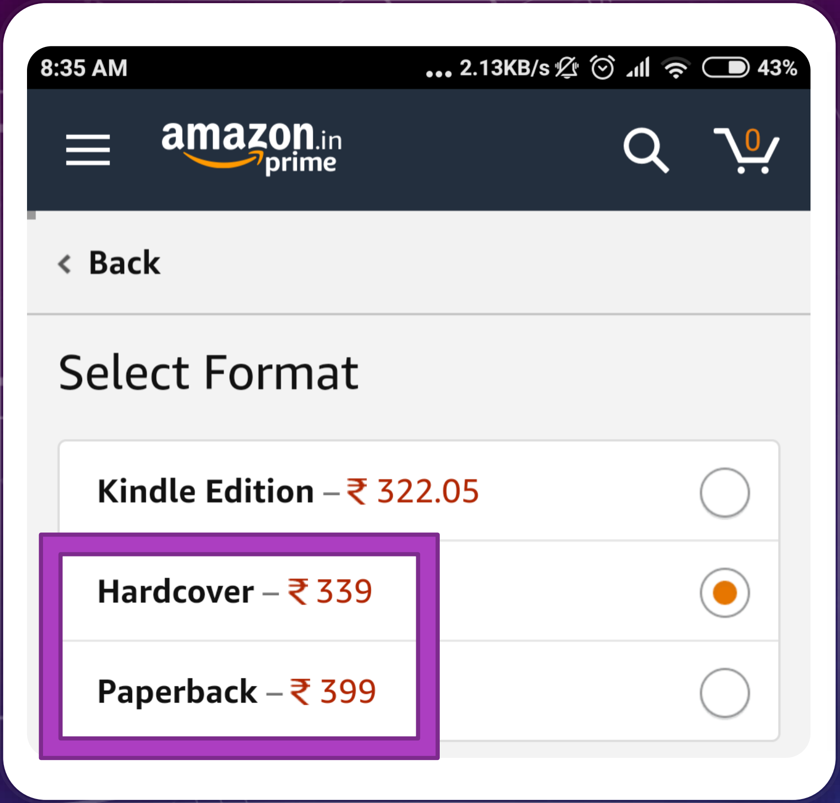

+ .007 ×
=
x
"panda"
57.7% confidence
$\text{sign}(\nabla_x J(\theta, x, y))$
"nematode"
8.2% confidence
$x + \epsilon\text{sign}(\nabla_x J(\theta, x, y))$
"gibbon"
99.3 % confidence

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.



Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. *arXiv preprint arXiv:1712.09665*.

FROM RESEARCH



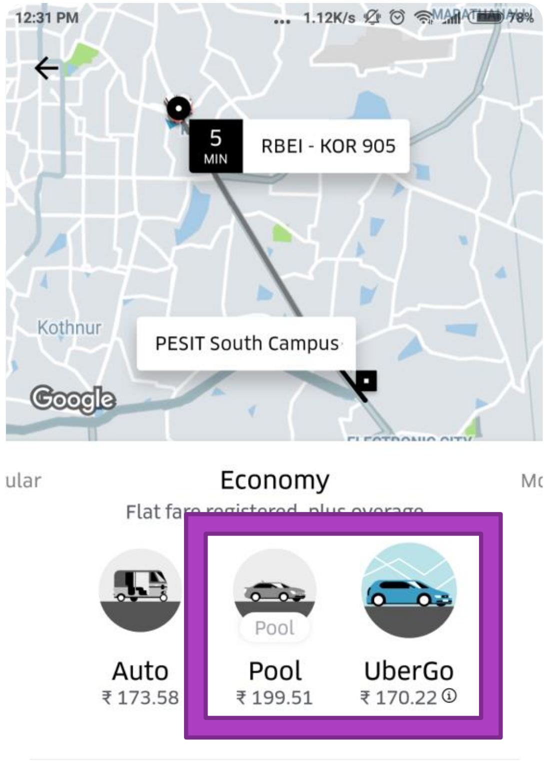
‹ Back

Select Format

Kindle Edition – ₹ 322.05 ○

Hardcover – ₹ 339 ●

Paperback – ₹ 399 ○



Auto ₹ 173.58

Pool ₹ 199.51

UberGo ₹ 170.22 ⓘ

REAL WORLD EXAMPLES: MY EXPERIENCES

QUESTIONS WE CARE

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?

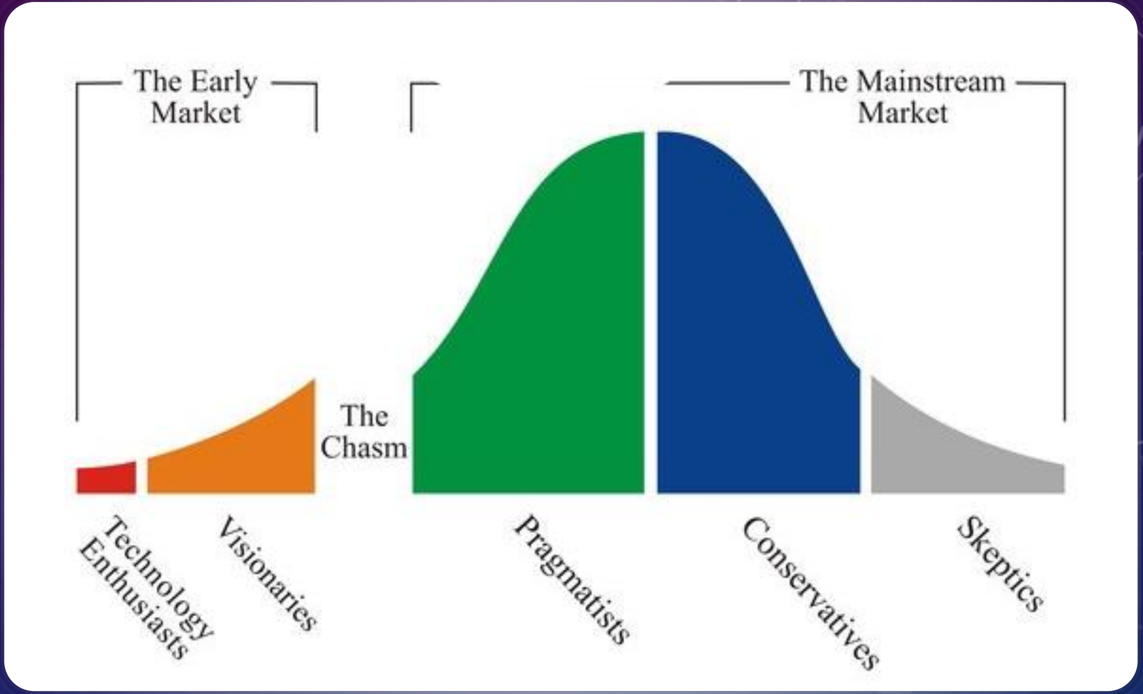
CHALLENGES TO ADOPTION

Understanding

Trust

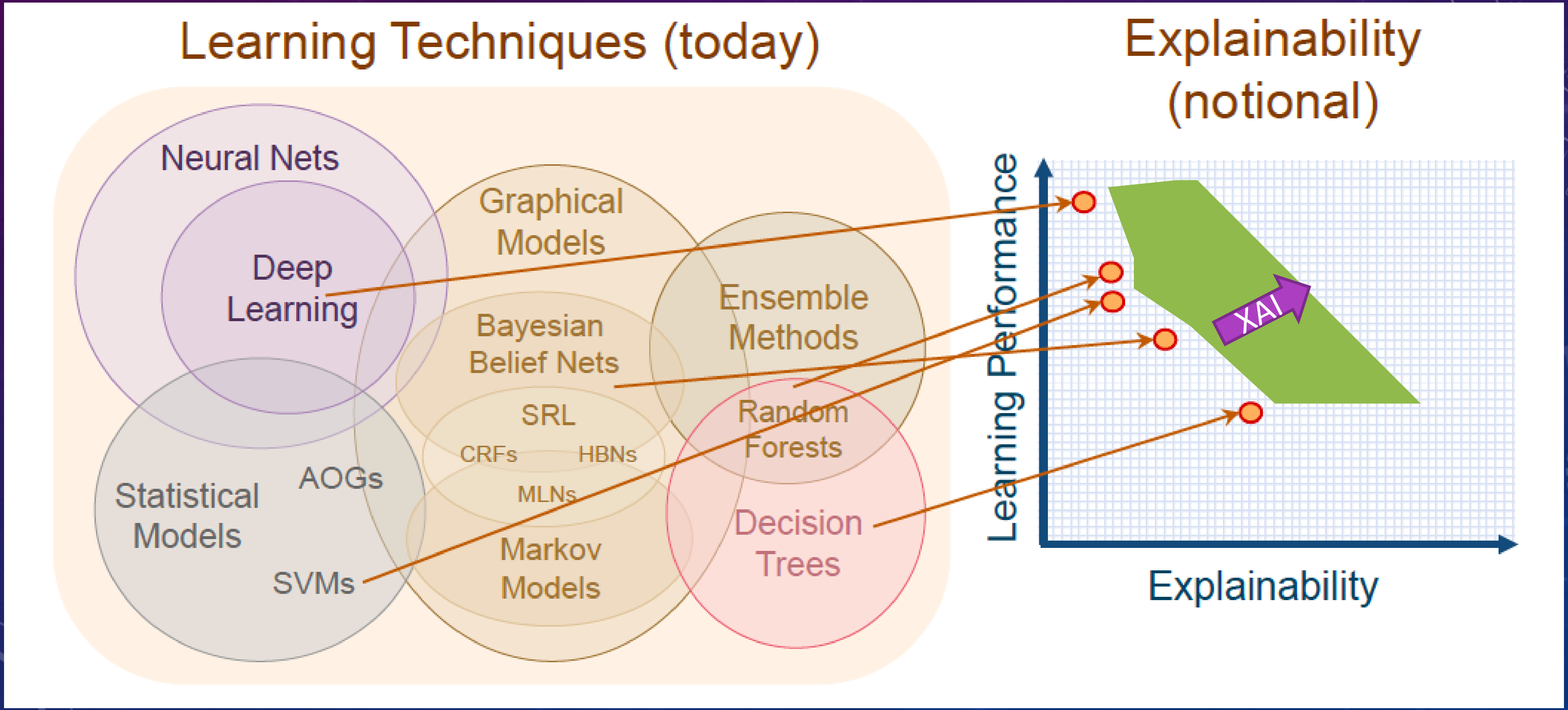
Transparency

Interaction



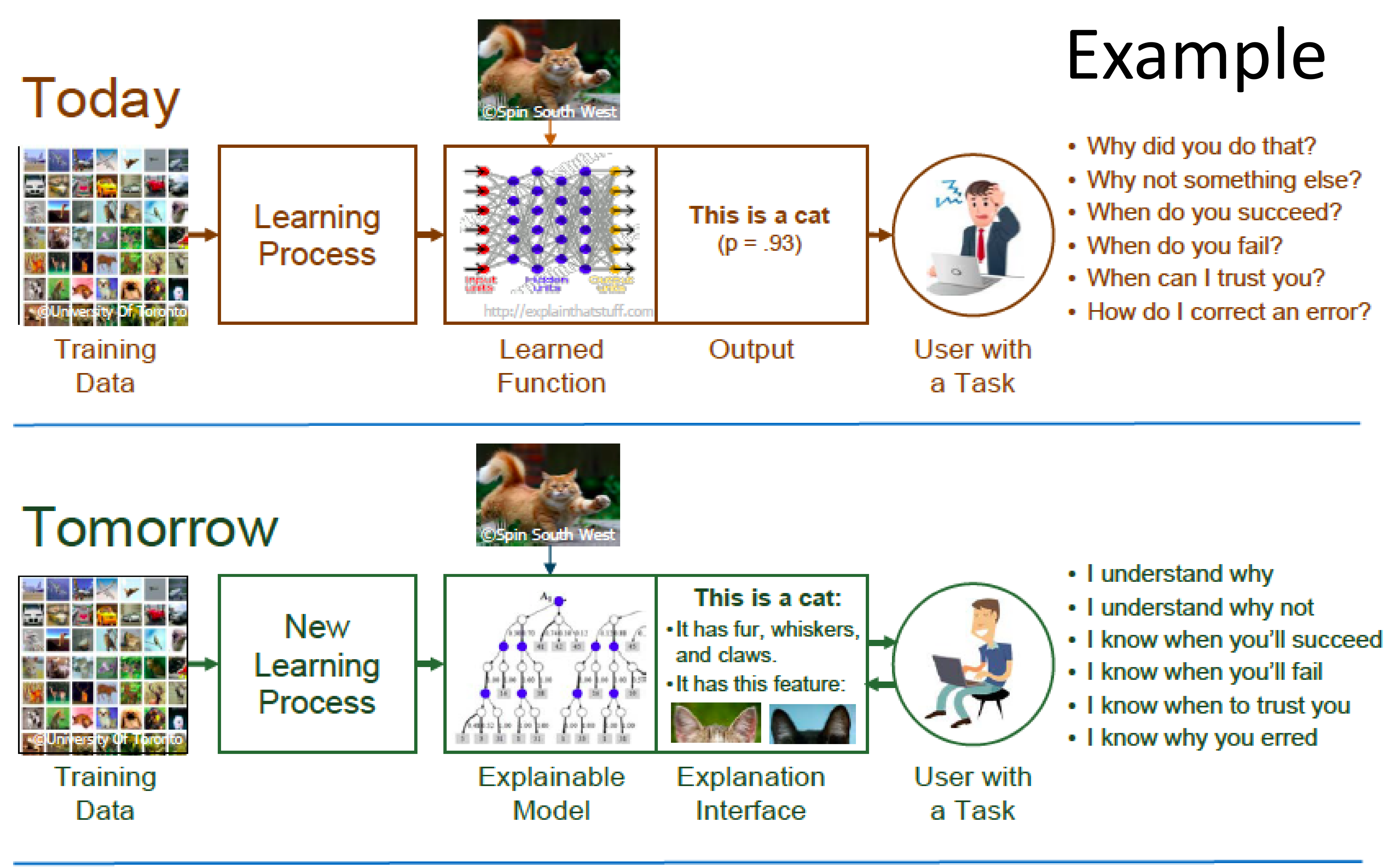
- ML/AI Models are Blackbox Models
- ML Models are Opaque, Non Intuitive and Difficult for people to understand
- Key Issue: Trustworthiness, reliability, rationality, and transparency of Models
  - Decisions of Machine or action thereby have far reaching impact on Individual, society or Government

PERFORMANCE VS. EXPLAINABILITY



http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Example



Today

Training Data → Learning Process → Learned Function → Output: This is a cat (p = .93) → User with a Task

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

Tomorrow

Training Data → New Learning Process → Explainable Model → Explanation Interface ⇄ User with a Task

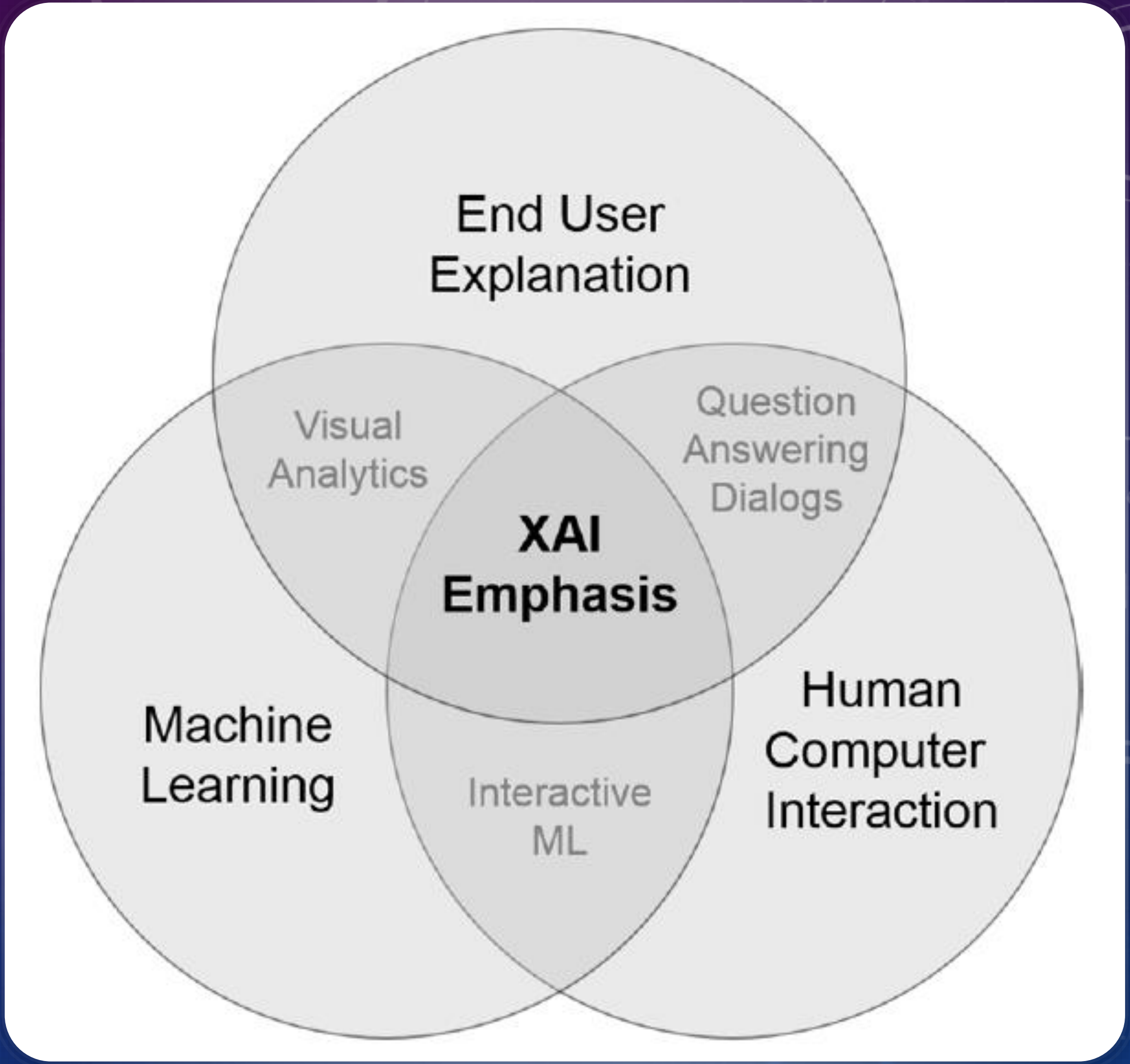
This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

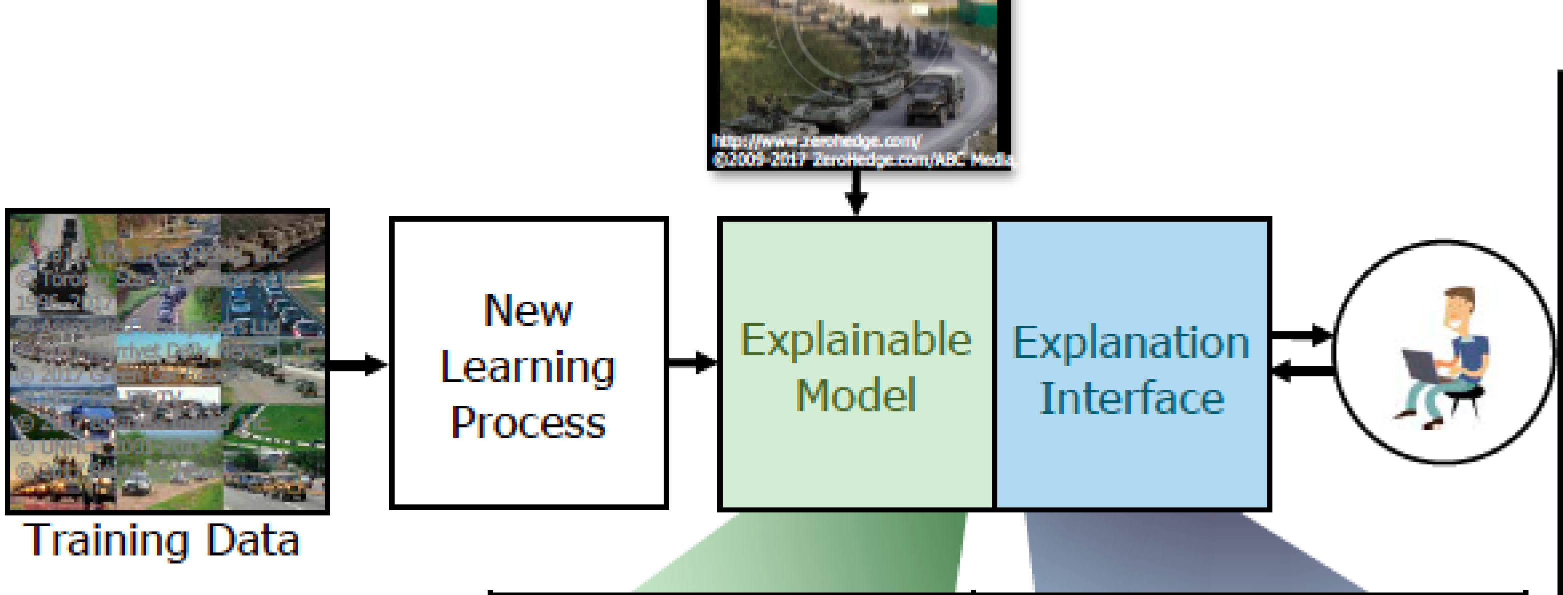
- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

XAI EMPHASIS



CURRENT RESEARCH FIELDS CONCEPTS AND APPROACHES



UC Berkeley	Deep Learning	Reflexive and Rational
Charles River Analytics	Causal Modeling	Narrative Generation
UCLA	Pattern Theory+	3-Level Explanation
Oregon State	Adaptive Programs	Acceptance Testing
PARC	Cognitive Modeling	Interactive Training
CMU	Explainable RL (XRL)	XRL Interaction
SRI International	Deep Learning	Show and Tell Explanations
Raytheon BBN	Deep Learning	Argumentation and Pedagogy
UT Dallas	Probabilistic Logic	Decision Diagrams
Texas A&M	Mimic Learning	Interactive Visualization
Rutgers	Model Induction	Bayesian Teaching

http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

REFERENCES

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM.
- Fox, M., Long, D., & Magazzeni, D. Explainable Planning. In IJCAI-17 Workshop on Explainable AI (XAI)
- Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.
- Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum. In IJCAI-17 Workshop on Explainable AI (XAI)

XAI: PATH TOWARDS FUTURE OF AI

THANK YOU!