Explainable AI

Ricardo Baeza-Yates
CTO, NTENT
Director of Data Science, Northeastern Univ. at SV
USA

Data Council Barcelona, October 2, 2019

Agenda

- Machine Learning and Big Data
- What is XAI?
- Why do we need it?
- How do we tackle it?
- Examples
- Responsible AI
- The Future

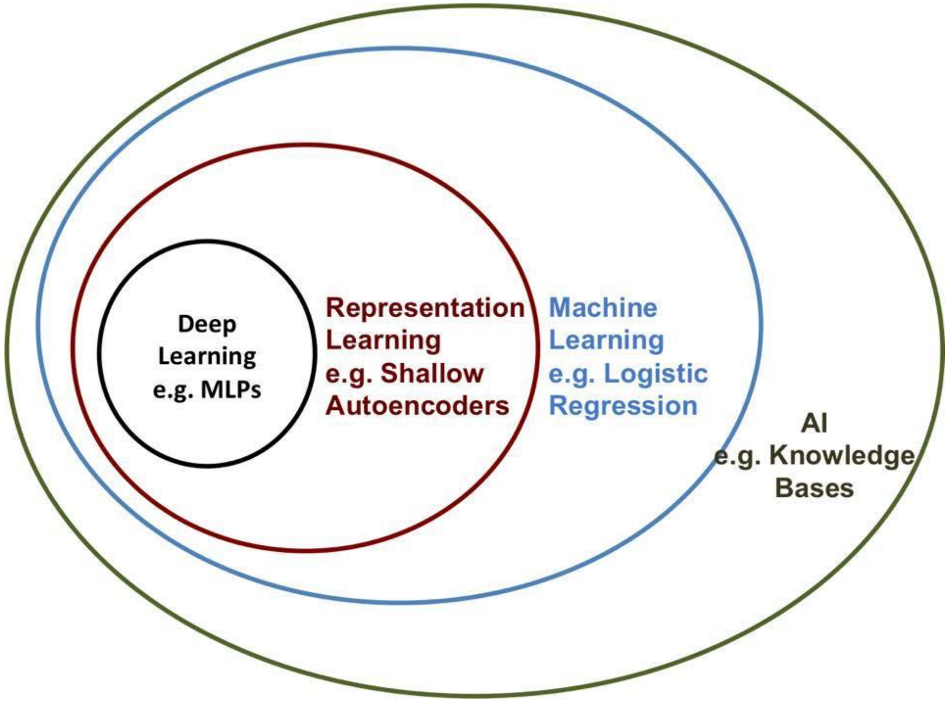
Machine Learning

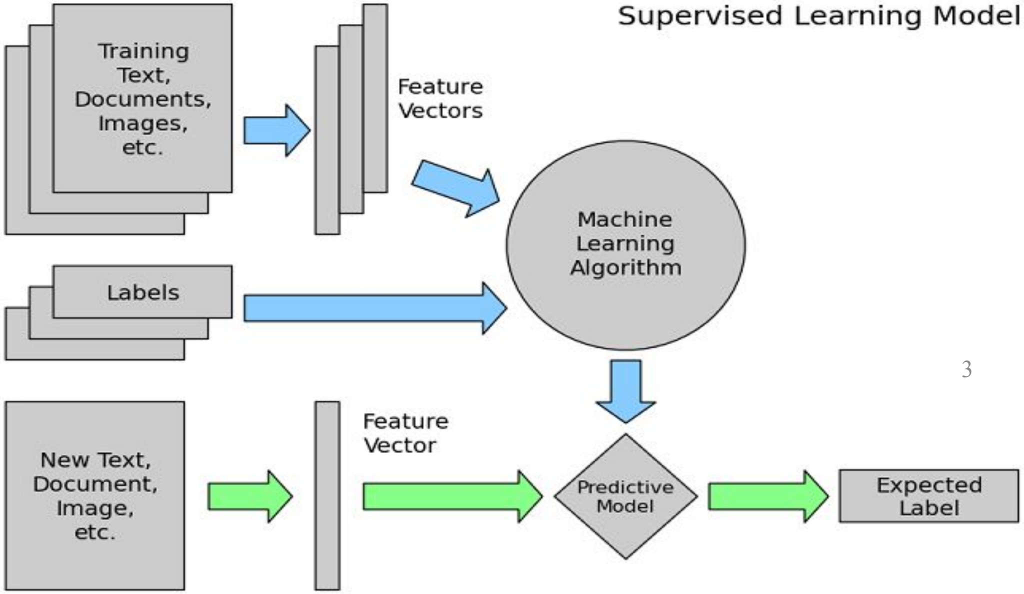
- AI is back!

- Why?
  - More data (Big data)
  - More processing power (GPUs)
  - Deep learning (neural networks)

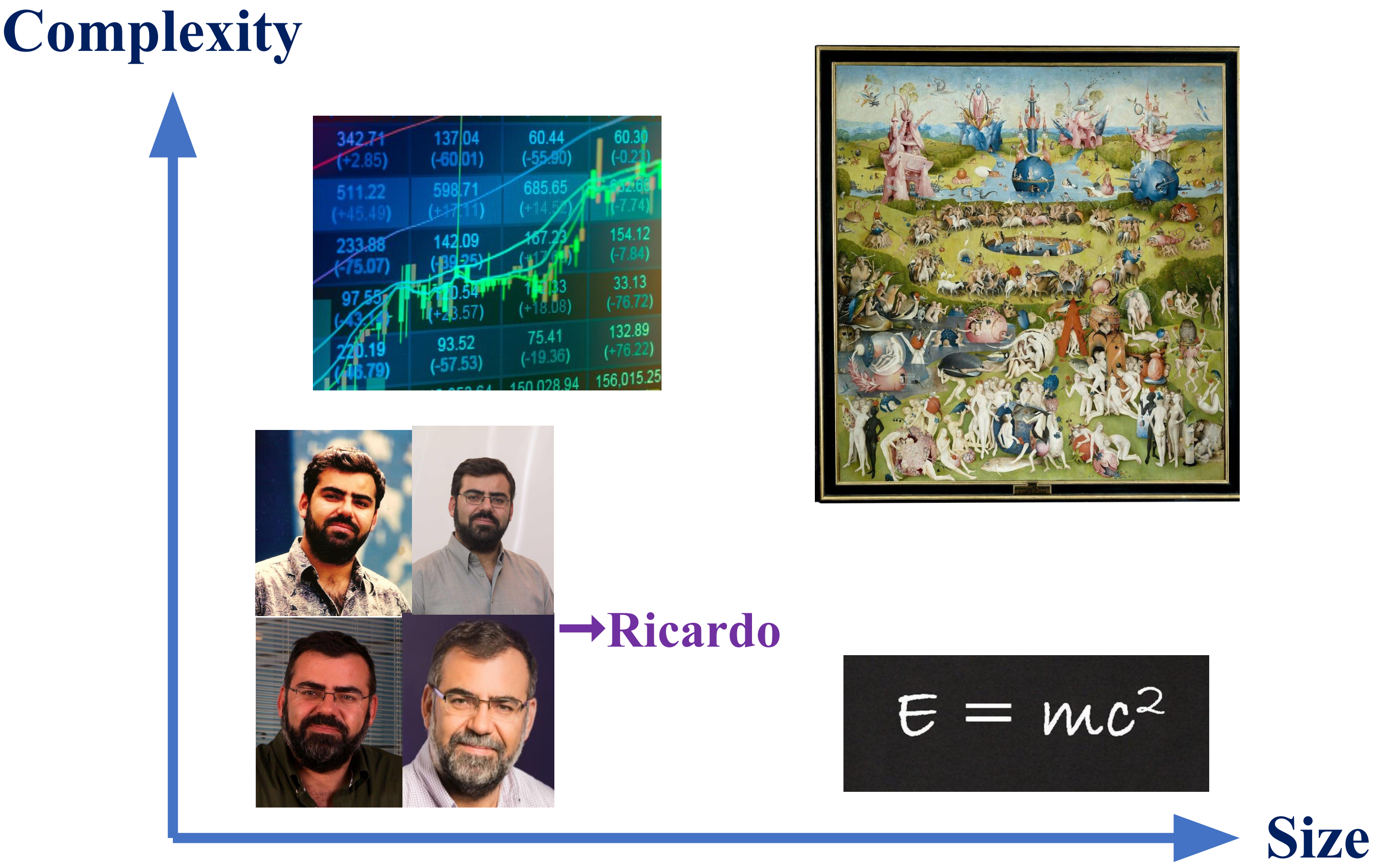
- Applications Everywhere
  - Shared economy
  - Driverless cars
  - Personalized Health
  - Improved Robots
  - YOUR Personal Data
  - …..
  - Learn, Predict, Prescribe
  - Truly Data-driven decisions

  - But in most cases is small data


Deep Learning e.g. MLPs
Representation Learning e.g. Shallow Autoencoders
Machine Learning e.g. Logistic Regression
AI e.g. Knowledge Bases


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Labels
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Big Data and Predictability



Complexity

→Ricardo

$E = mc^2$

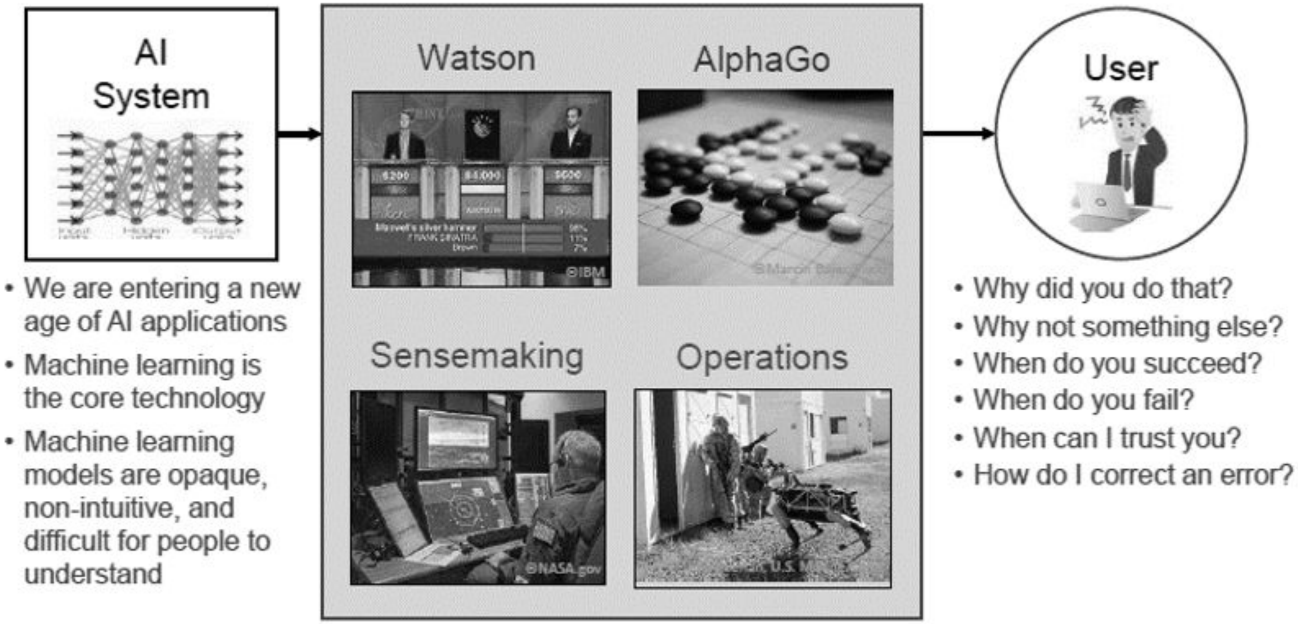
Size

Inspired by Claudia Perlich, 2018

Explainable AI

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?

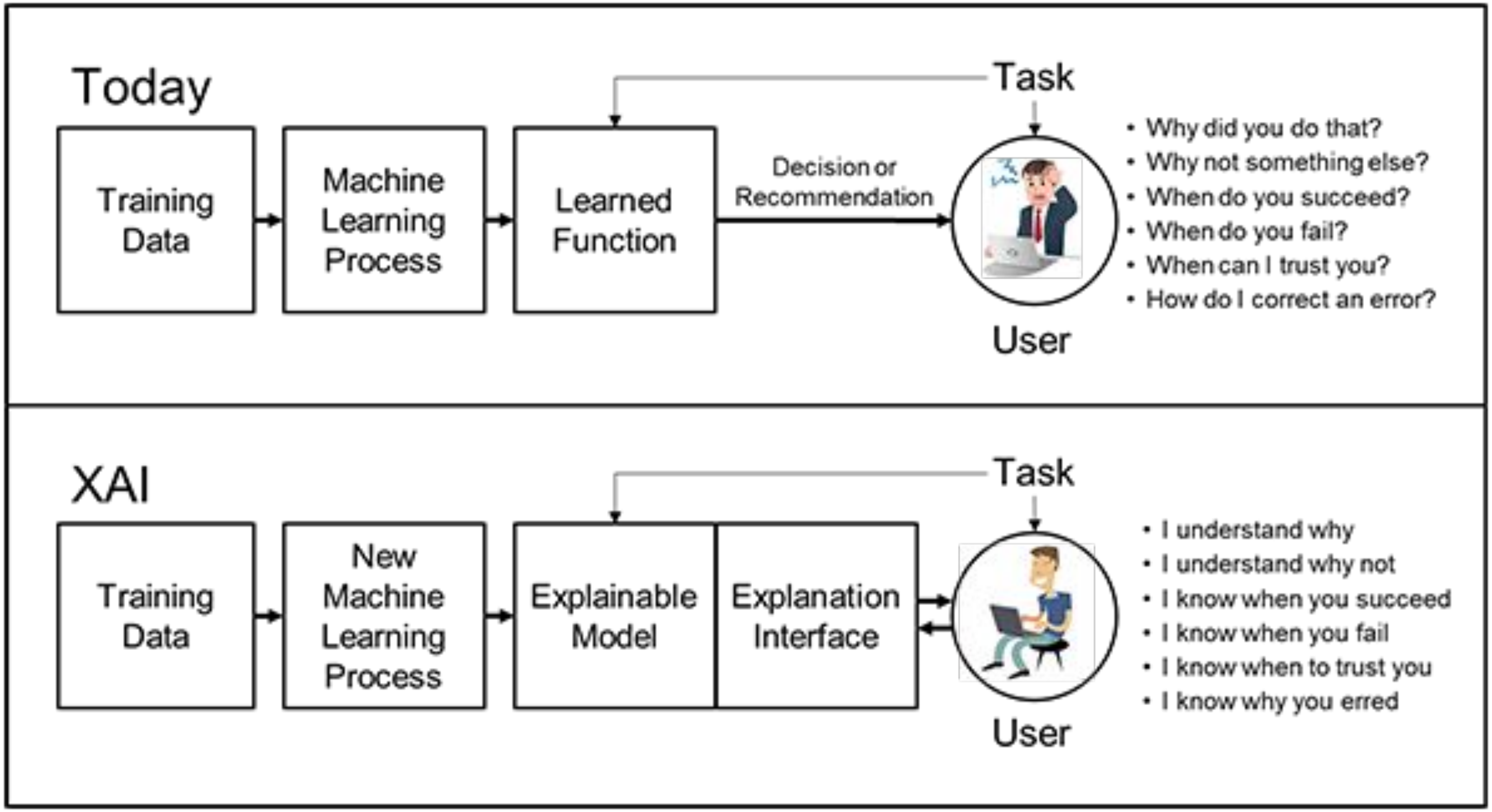


- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

- 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?

[DARPA XAI program, 2016]

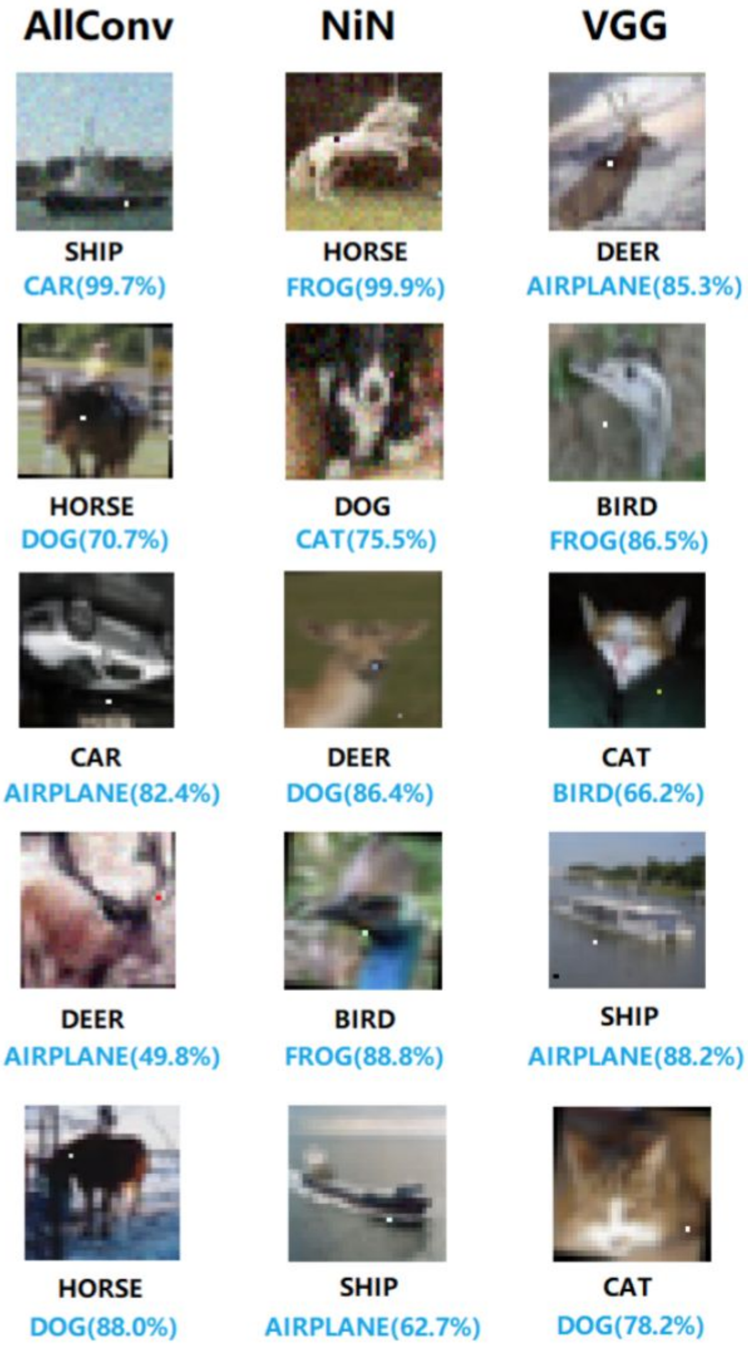
Goals



[DARPA XAI program, 2016]

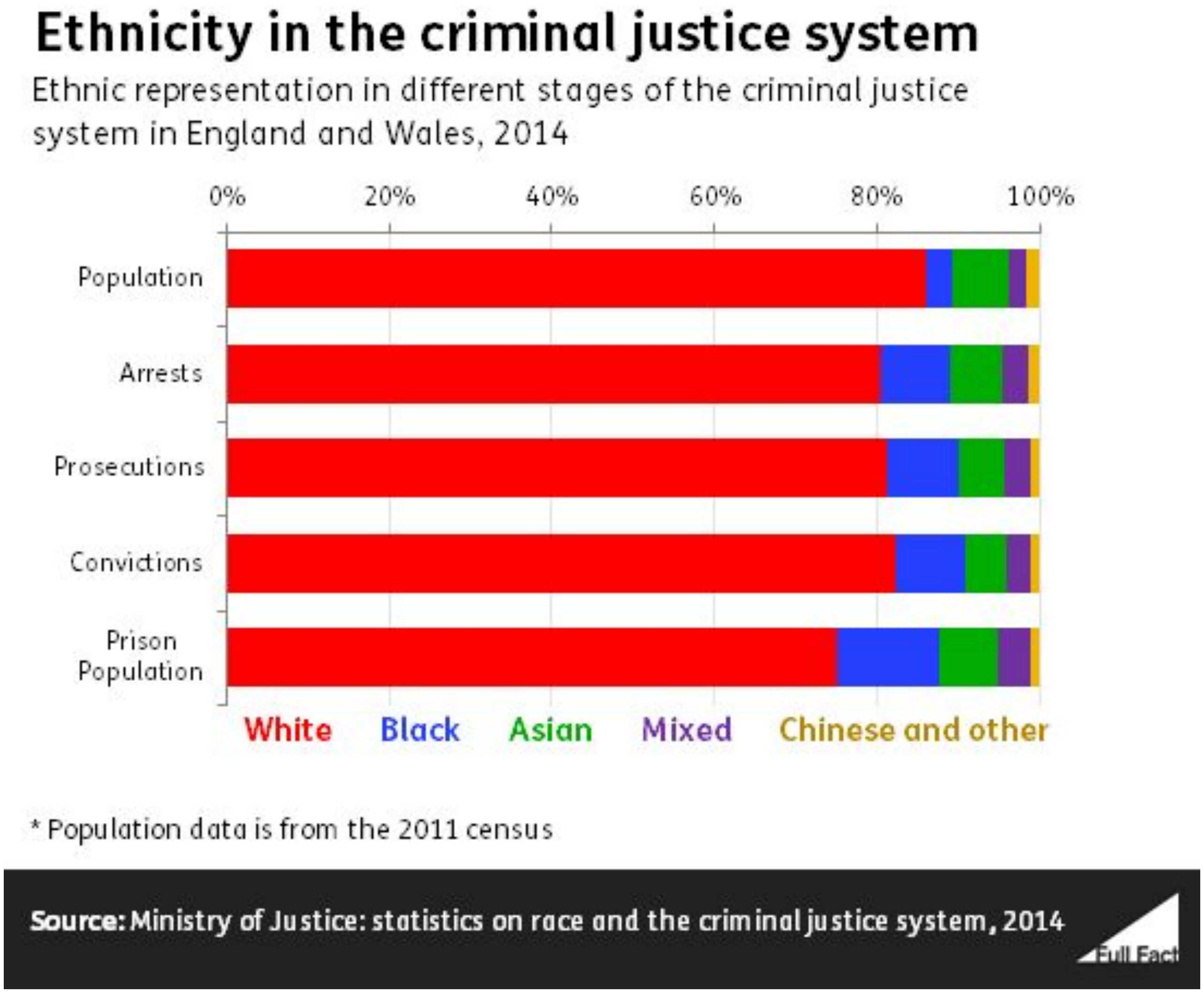
Why We Need XAI?

- Verification of the System
  - Transparency
- Improvement of the System
  - Mismatched objectives
  - Multi-objective trade-offs
  - Some systems are very sensitive
- Learning from the System
  - Causality
- Compliance to legislation
  - Safety
  - E.g., GDPR
- Ethical Issues
  - Transparency
  - Fairness: e.g., gender or race bias



One pixel attack
[Su et al, 2018]

It is Easy to Learn Bias from Data: Race



Ethnicity in the criminal justice system

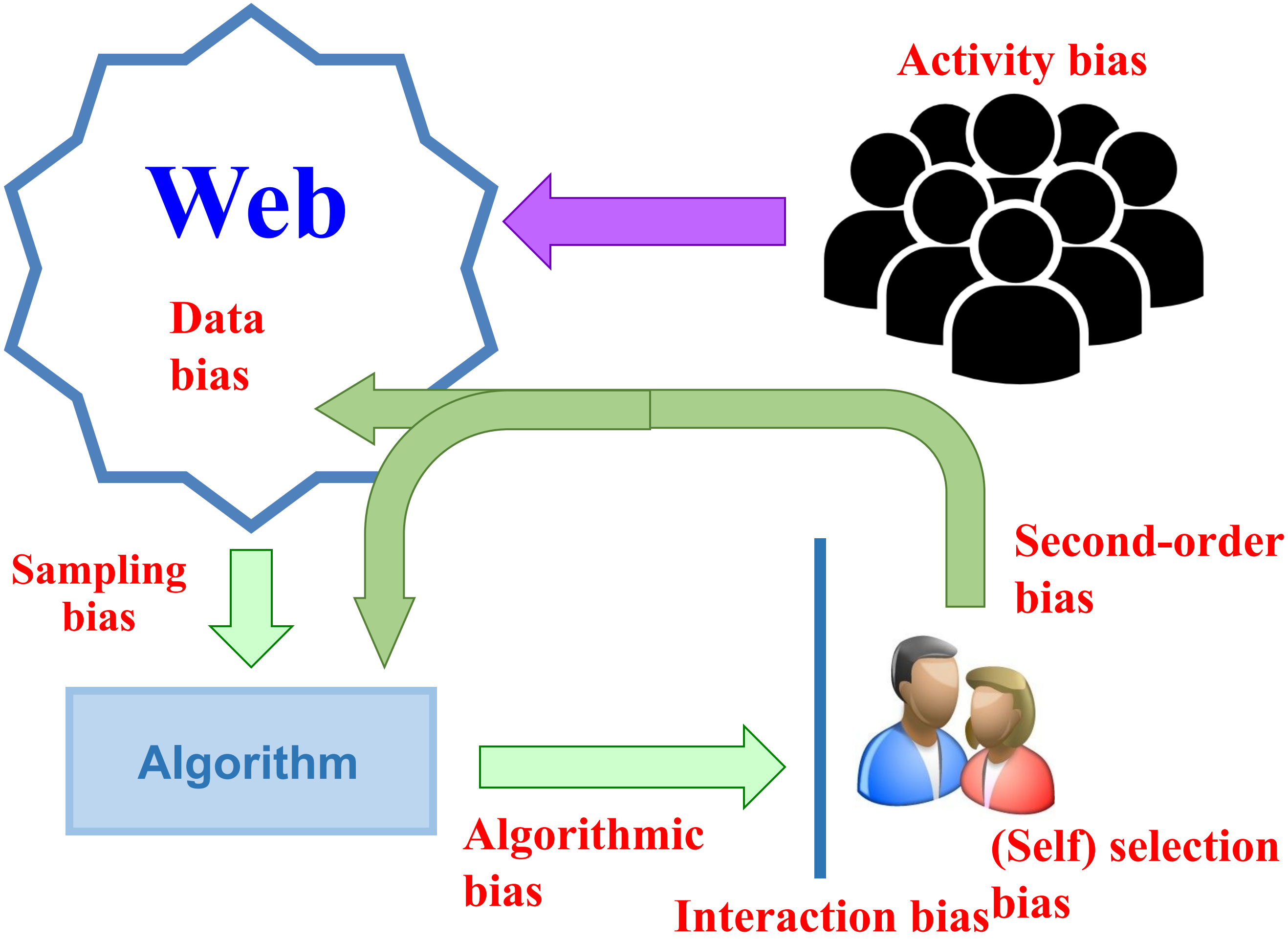
Ethnic representation in different stages of the criminal justice system in England and Wales, 2014

* Population data is from the 2011 census

Source: Ministry of Justice: statistics on race and the criminal justice system, 2014

Bias on the Web: A Vicious Cycle

Biases = Statistical + Cultural + Cognitive



[Baeza-Yates, Bias on the Web, *CACM*, June 2018]

Hinton’s interview in Wired answered by Forbes

(2018)

- ***Hinton:*** *People can’t explain how they work, for most of the things they do. When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them* ***to make up a story****.*
- Timothy Miller, Associate Professor in CS at the University of Melbourne, Australia answers: *His quoted paragraph is itself an explanation: an explanation of why he has reached the decision that explainability for AI would be a disaster.* ***Is he making up a story about this?*** *I imagine he would claim that he is not and that it is based on careful reasoning. But in reality, it is based on neurons in his brain firing in a particular way that nobody understands. The ability to communicate his reasons to others is a strength of the human brain. Philosopher Daniel Dennett claims that consciousness itself is simply our brain creating an `edited digest’ of our brains inner workers for precisely the purpose of communicating our thoughts and intentions (including explanations) to others.*

May Also Help Us!

POLICE
PAUL NOTH

"Does your car have any idea why my car pulled it over?"

New Yorker

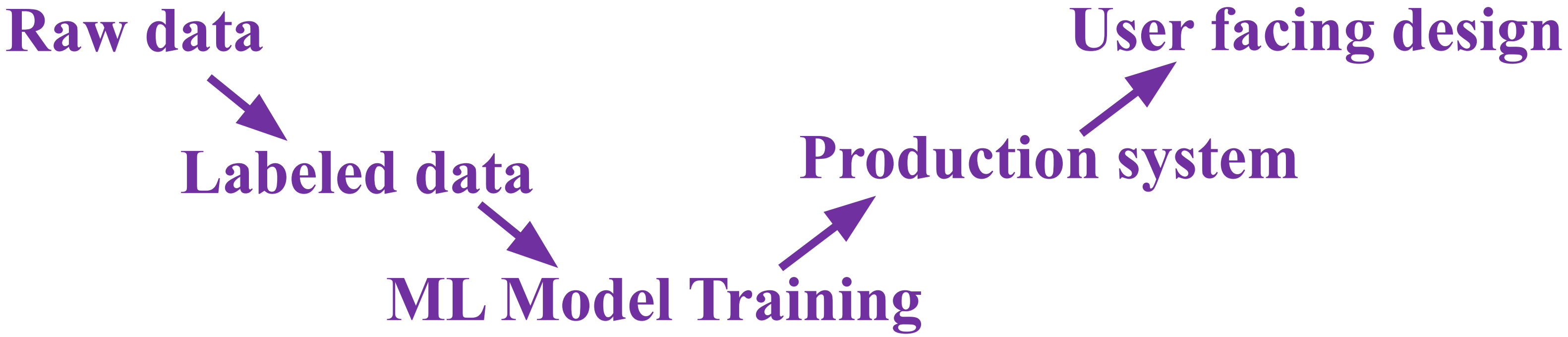
Accountability

- Can we hold an AI system accountable? (e.g., robot)
- Should an AI system have rights?

NO

- Legal entities: individuals, corporations, and idols
- Other animals and robots are not human beings

- Who then should be accountable?



Raw data

Labeled data

ML Model Training

Production system

User facing design

[Solaiman, 2016; Bryson et al., 2017]

Explaining AI Systems (Bryson, 2019)

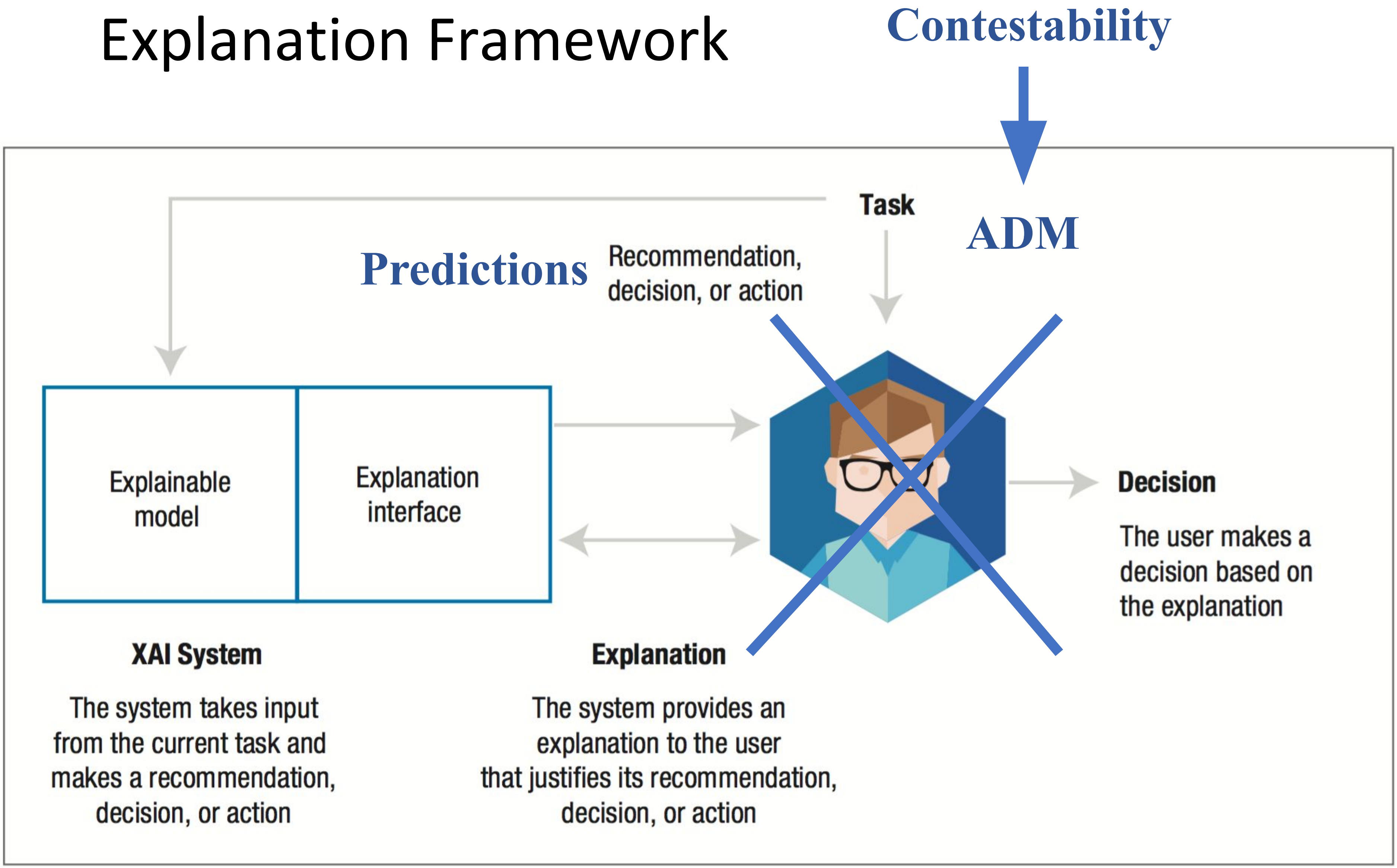
1. No explanation (too hard, in many cases it's impossible!) ✗
2. Explain human actions that led to the system (accountability, better understanding) ✓
3. Explaining what inputs resulted in what outputs
   a) Be able to experiment with a black box and see what changes (digital forensics)
   b) Record (secure) logs for later analysis (legal) ✓ People involved
4. Seeing exactly how the system works
   a) An explanation of the overall system (e.g., documentation)
   b) Making ML models more transparent

Do We Always Need XAI? No!

- ADM: Automatic Decision Making (not only ML)
- Many automatic systems do not have significant consequences for unacceptable results or
- Problem is sufficiently well-studied and validated in real applications that we trust the system’s decision
- Examples:
    - Web advertising
    - Postal code sorting
    - Aircraft collision avoidance systems

- Challenge: Which systems can/should be fully automated?

Explanation Framework



Contestability

Task

ADM

Predictions

Recommendation, decision, or action

Explainable model

Explanation interface

Decision

The user makes a decision based on the explanation

XAI System

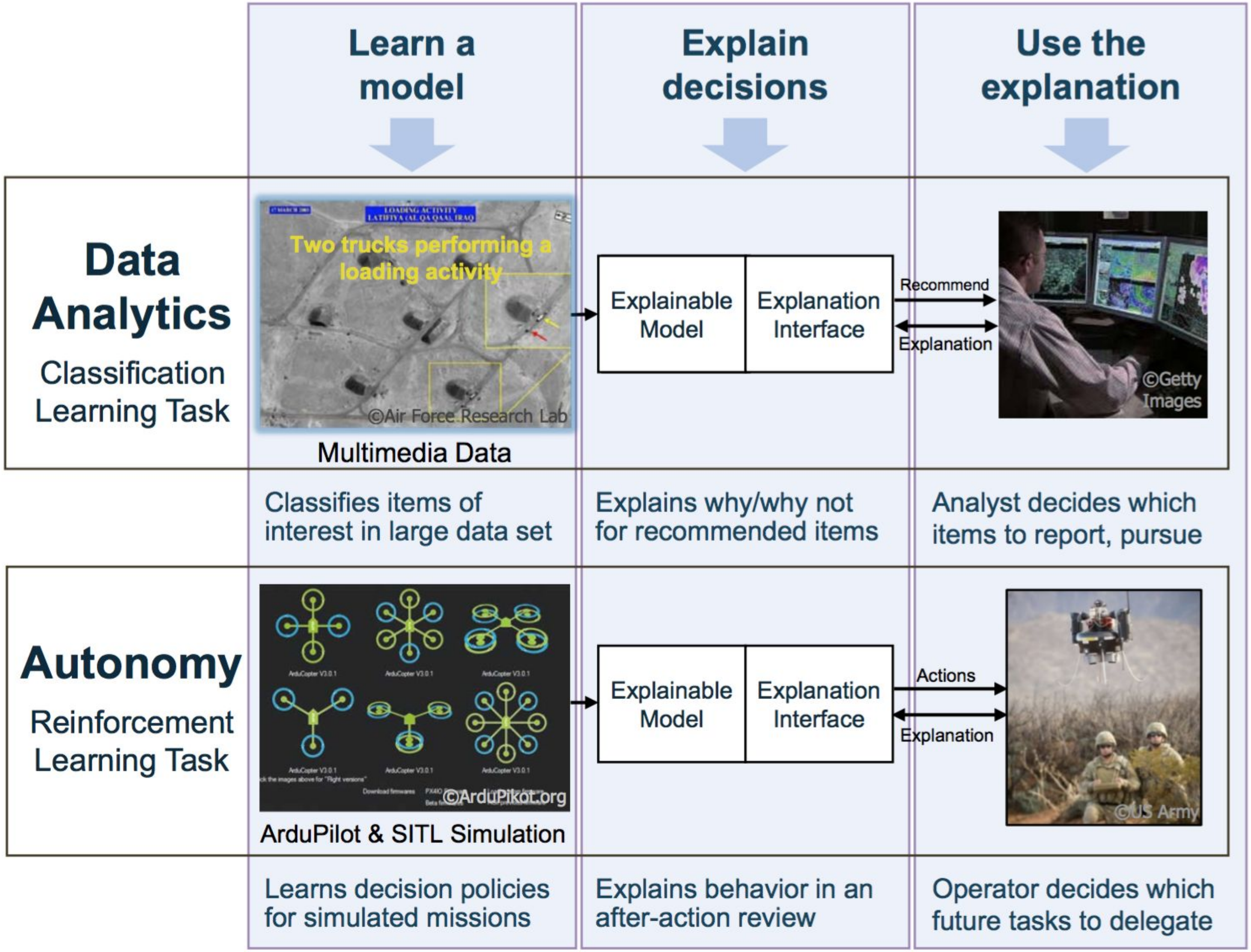
The system takes input from the current task and makes a recommendation, decision, or action

Explanation

The system provides an explanation to the user that justifies its recommendation, decision, or action

[DARPA XAI program, 2016]

Challenges

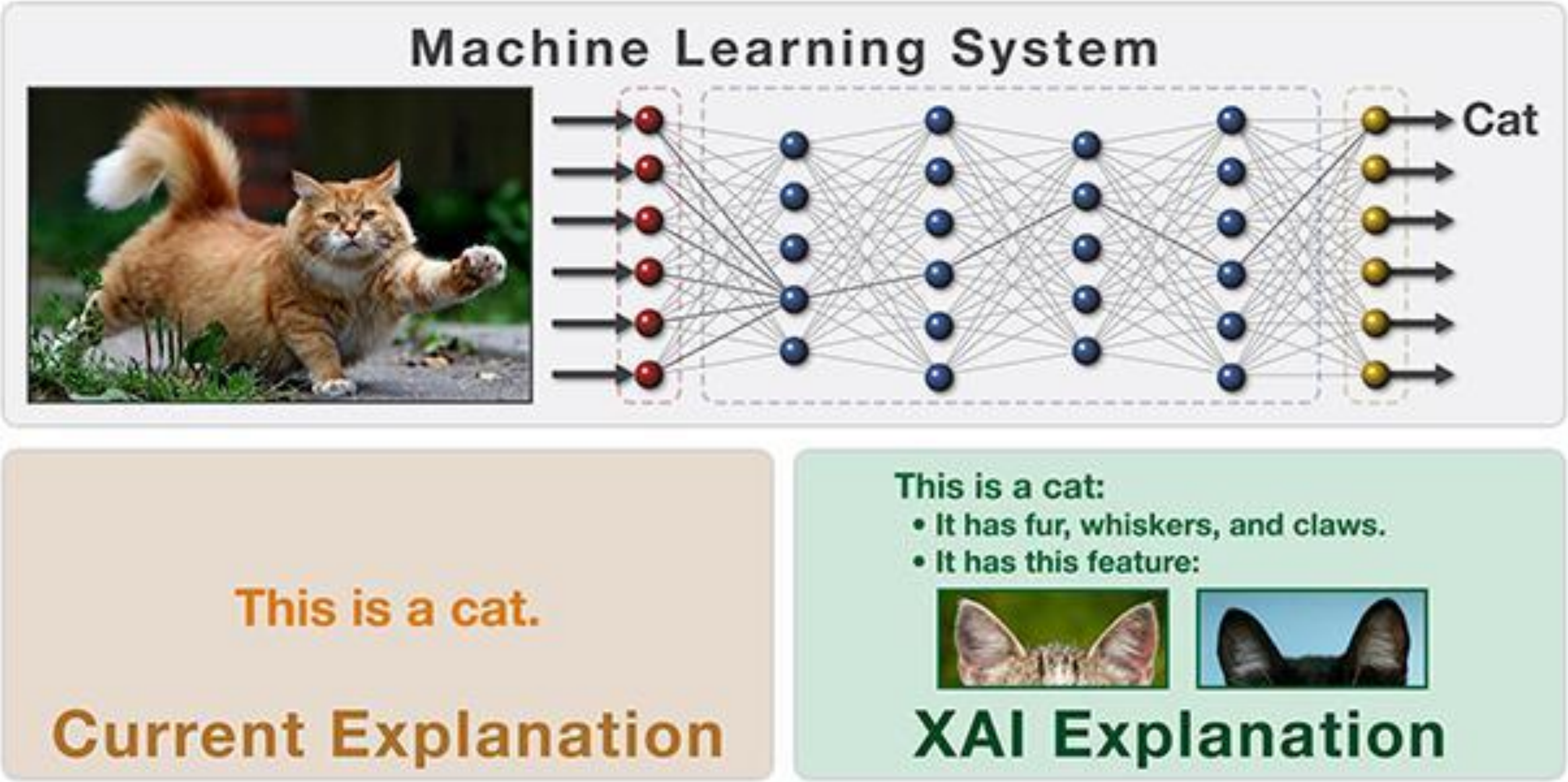


	Learn a model	Explain decisions	Use the explanation
Data Analytics Classification Learning Task	Multimedia Data	Explainable Model Explanation Interface	
	Classifies items of interest in large data set	Explains why/why not for recommended items	Analyst decides which items to report, pursue
Autonomy Reinforcement Learning Task	ArduPilot & SITL Simulation	Explainable Model Explanation Interface	
	Learns decision policies for simulated missions	Explains behavior in an after-action review	Operator decides which future tasks to delegate

An analyst is looking for items of interest in massive multimedia data sets

An operator is directing autonomous systems to accomplish a series of missions

Explanation Example



This is a cat.

Current Explanation

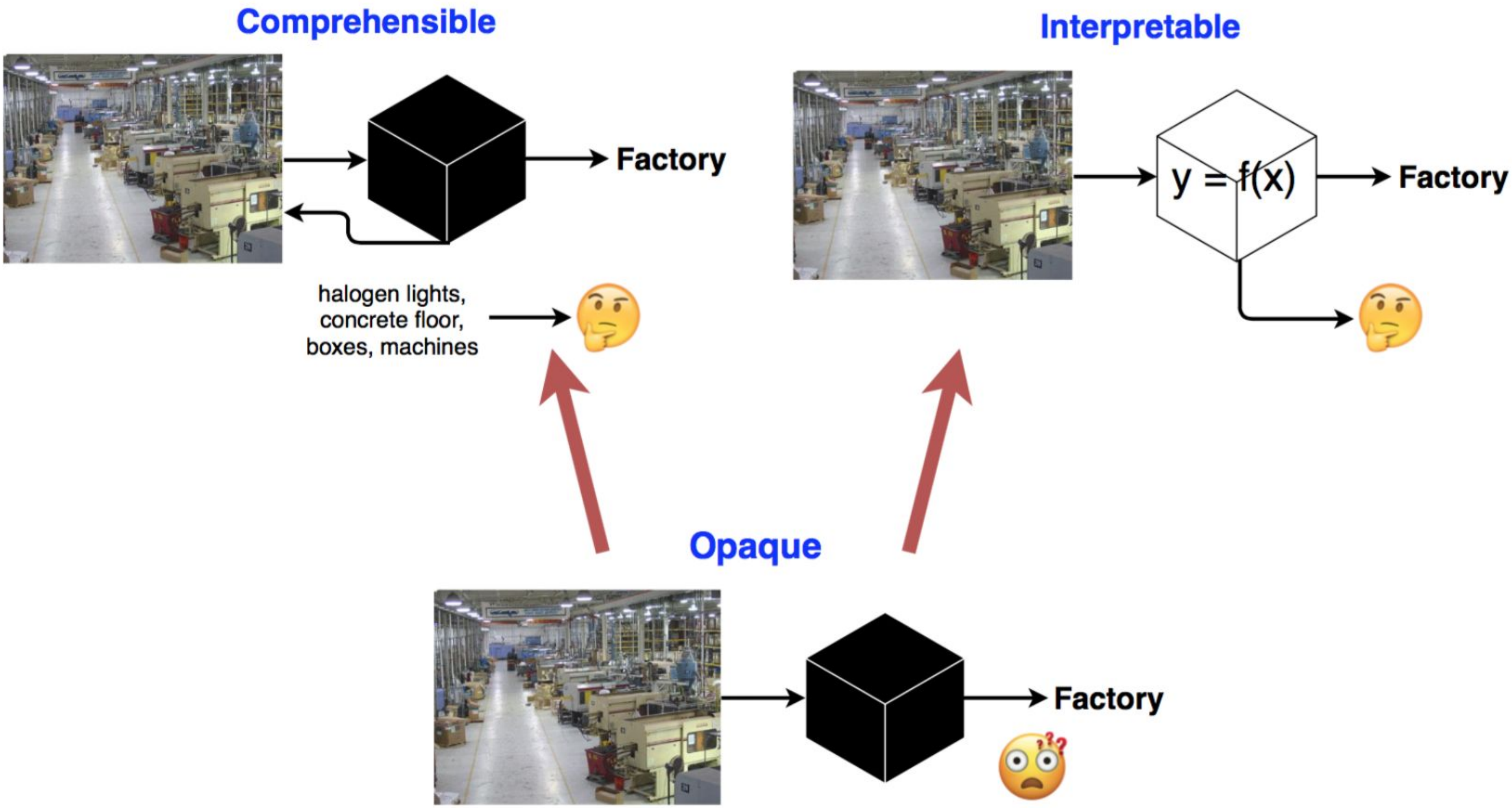
This is a cat:

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

XAI Explanation

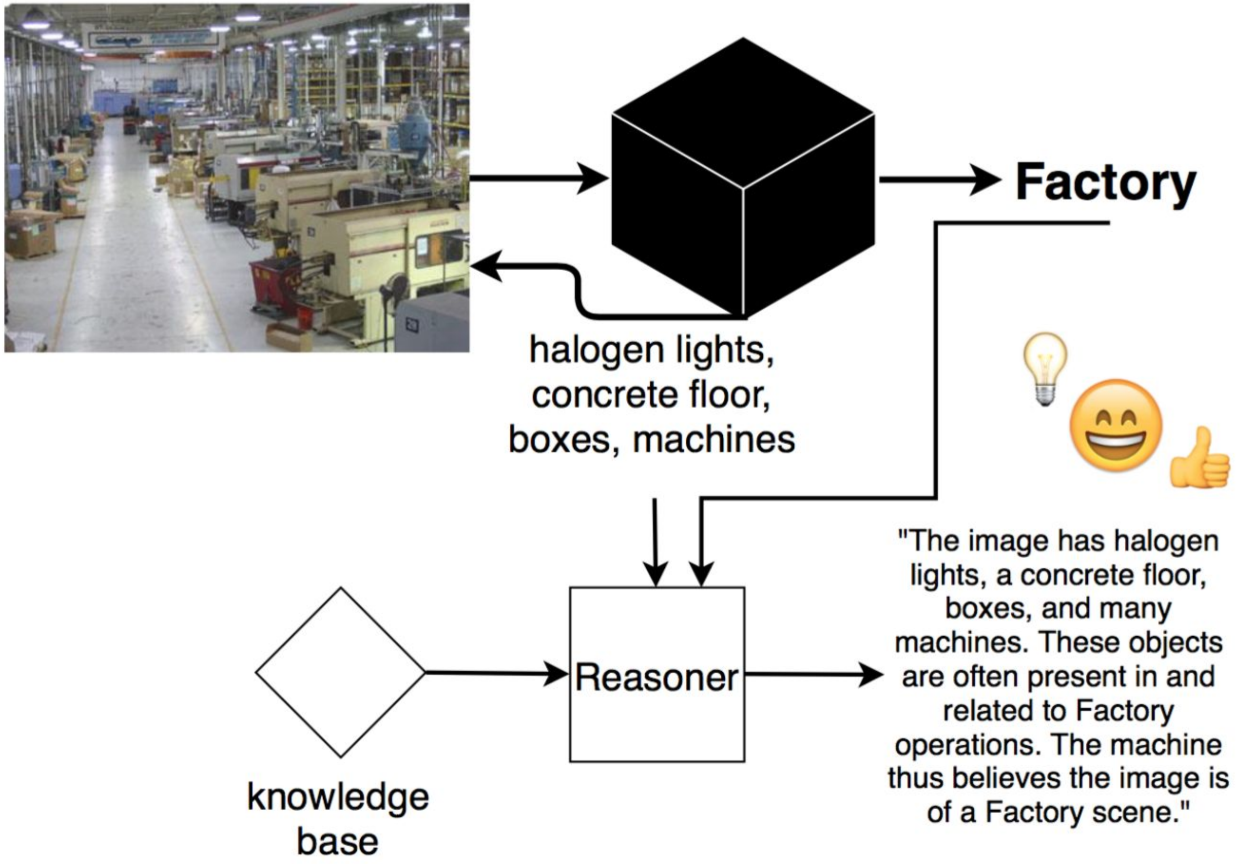
Types of explanations: descriptions or justifications (Vig et al., 2009)

Different Ways to Approach the Problem



Doran at al, 2017

Should XAI Include Reasoning?

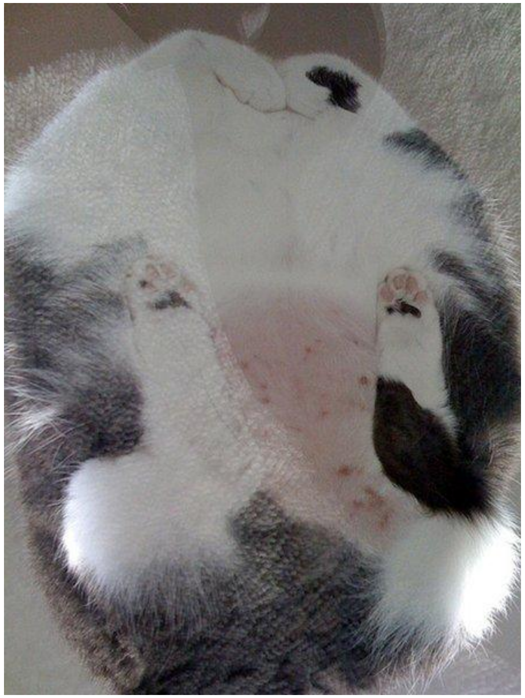


"The image has halogen lights, a concrete floor, boxes, and many machines. These objects are often present in and related to Factory operations. The machine thus believes the image is of a Factory scene."

[Doran at al, 2017]

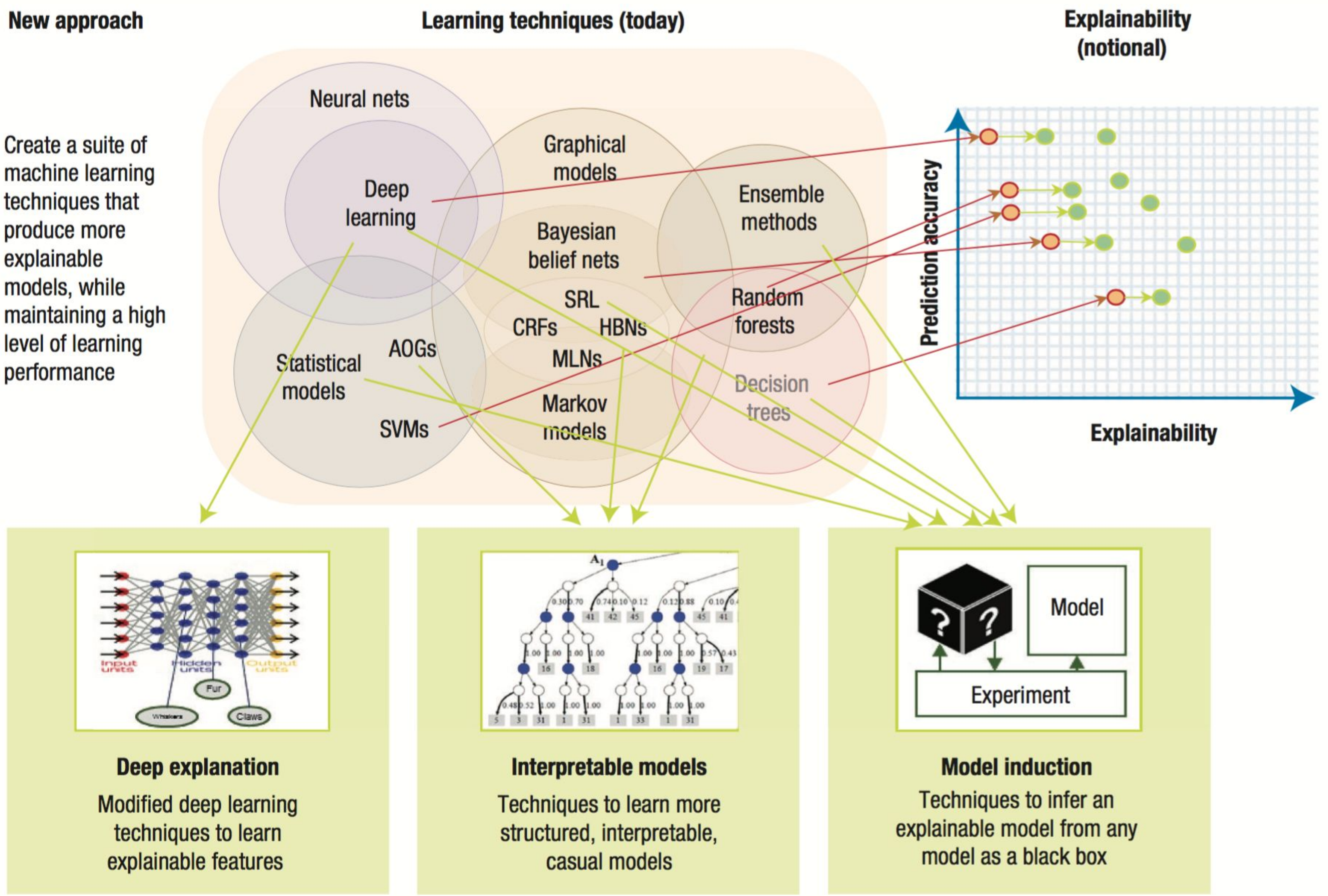
Explanations Might be Difficult

- What is this?
  - Systems are afraid to say “I don’t know”
  - “The Last Question”, Asimov (1956)



- If it is a cat, why does not have the pointing ears feature?
- Interpretability does not imply completeness
- Explanations may come from the data, the model, the process, etc.

Tackling the Problem



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)

Neural nets

Deep learning

Statistical models

AOGs

SVMs

Graphical models

Bayesian belief nets

SRL

CRFs

HBNs

MLNs

Markov models

Ensemble methods

Random forests

Decision trees

Explainability (notional)

Prediction accuracy

Explainability

Deep explanation

Modified deep learning techniques to learn explainable features

Interpretable models

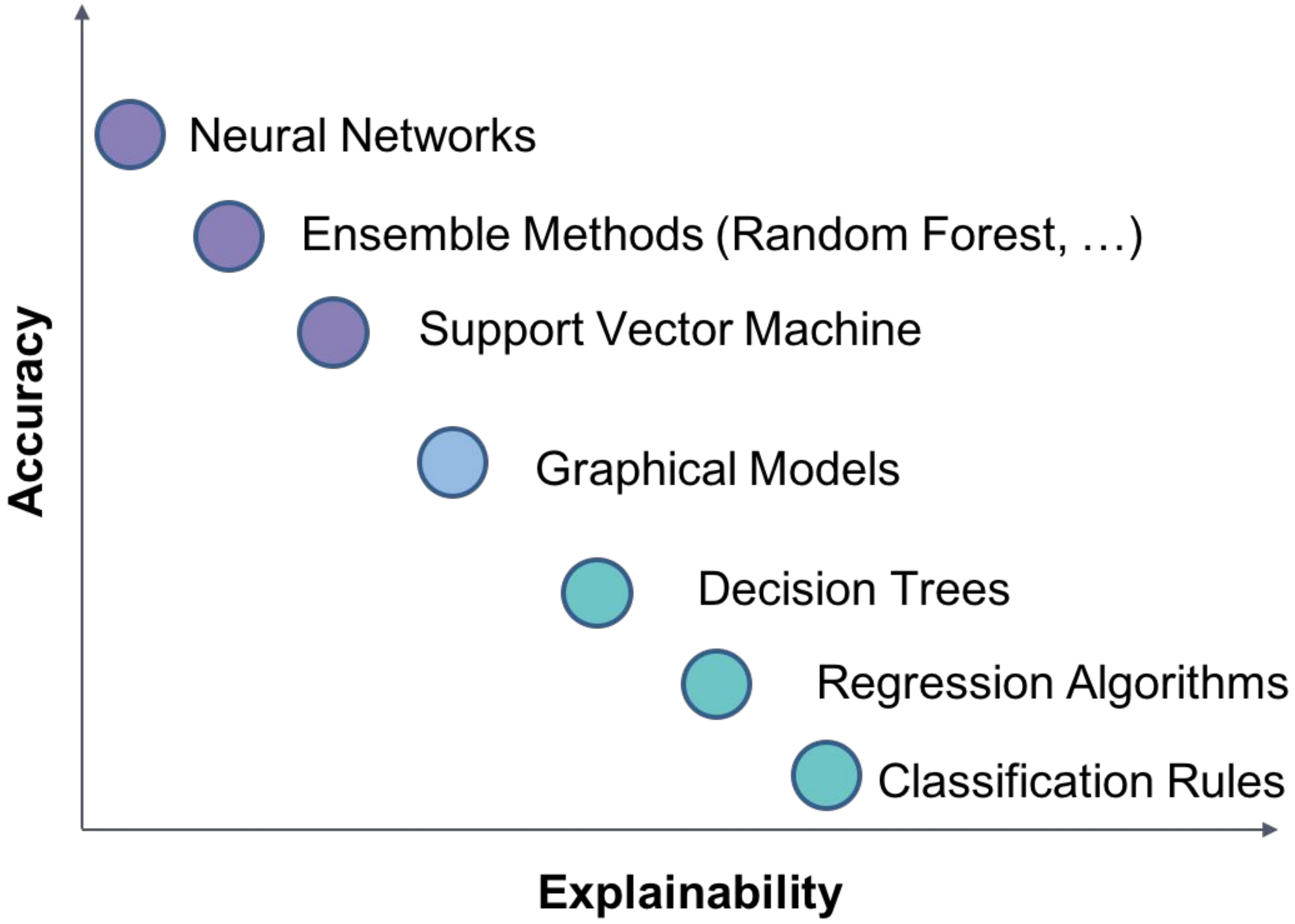
Techniques to learn more structured, interpretable, casual models

Model induction

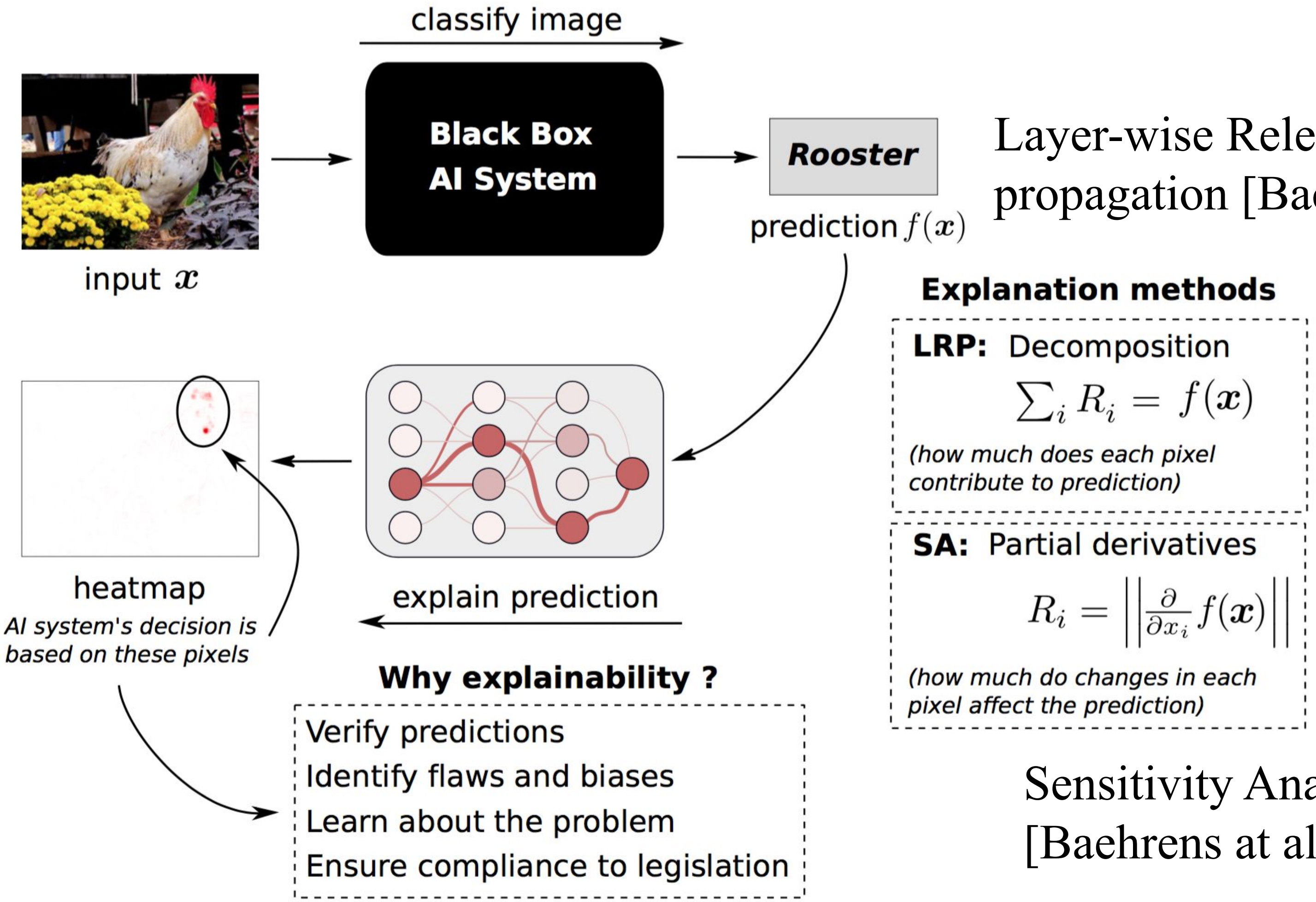
Techniques to infer an explainable model from any model as a black box

[DARPA XAI program, 2016]

ML Models are Easy or Hard to Explain



Explaining Deep Learning Models



classify image

Black Box AI System

Rooster

input $\boldsymbol{x}$

prediction $f(\boldsymbol{x})$

heatmap

AI system's decision is based on these pixels

explain prediction

Explanation methods

LRP: Decomposition

$$\textstyle\sum_i R_i = f(\boldsymbol{x})$$

(how much does each pixel contribute to prediction)

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)

Why explainability ?

Verify predictions
Identify flaws and biases
Learn about the problem
Ensure compliance to legislation

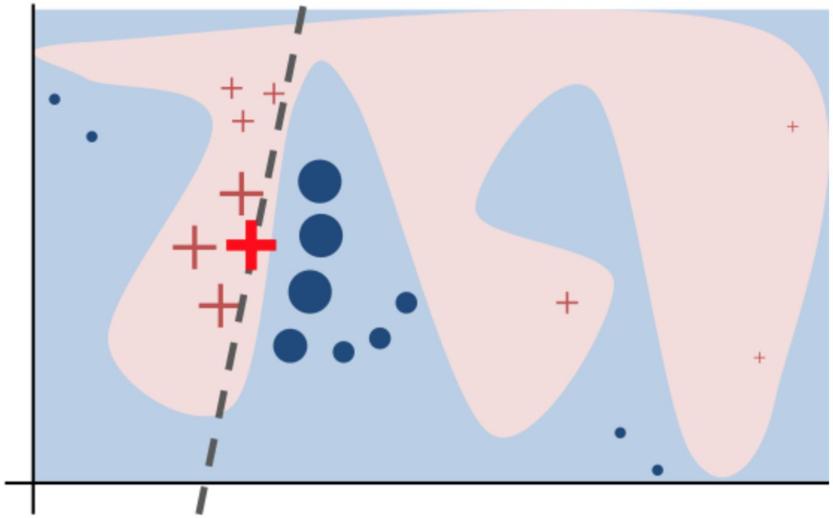
Layer-wise Relevant propagation [Bach et al, 2015]

Sensitivity Analysis [Baehrens at al, 2010]

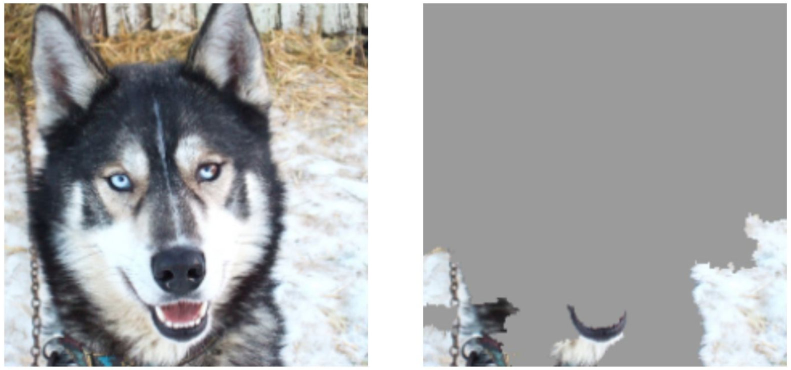
[Samek at al, 2017]

LIME: Local Interpretable Model-Agnostic Explanations

[Ribeiro et al, 2016]

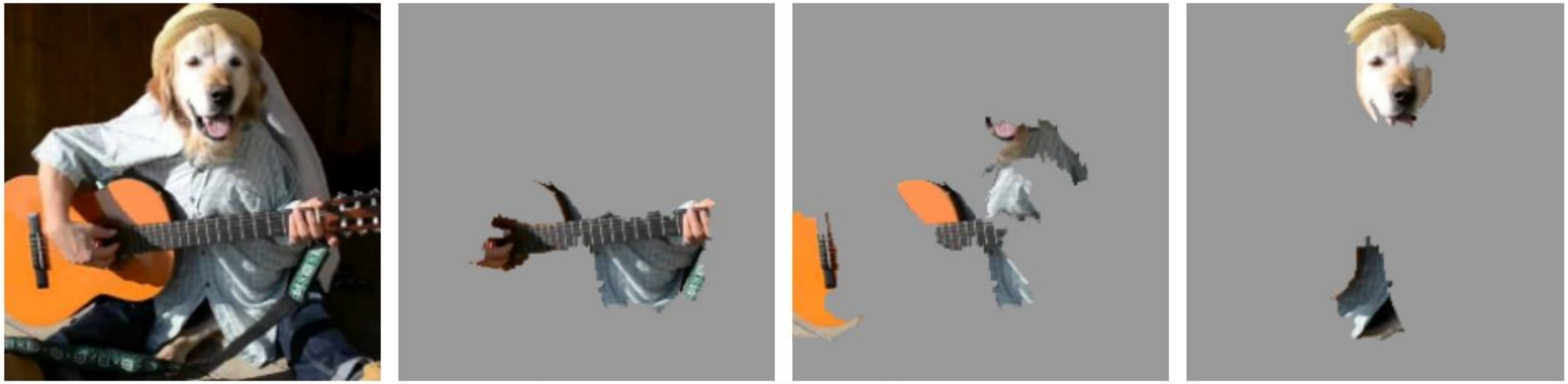


Impact of local perturbations



(a) Husky classified as wolf

(b) Explanation



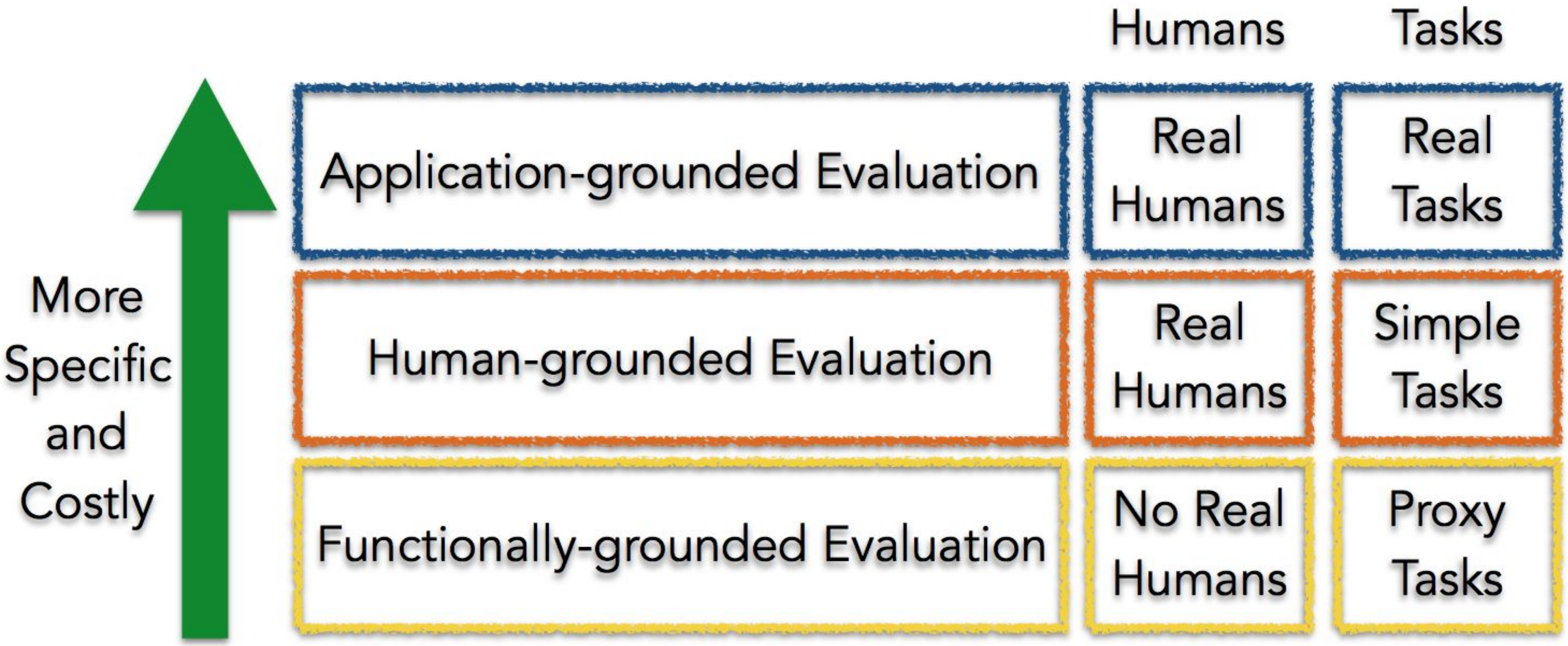
(a) Original Image

(b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*

Evaluating XAI



More Specific and Costly

	Humans	Tasks
Application-grounded Evaluation	Real Humans	Real Tasks
Human-grounded Evaluation	Real Humans	Simple Tasks
Functionally-grounded Evaluation	No Real Humans	Proxy Tasks

[Doshi-Velez & Kim, 2017]

Explaining Explanations [Gilpin et al, 2018]

- Explanations = Interpretability + Completeness

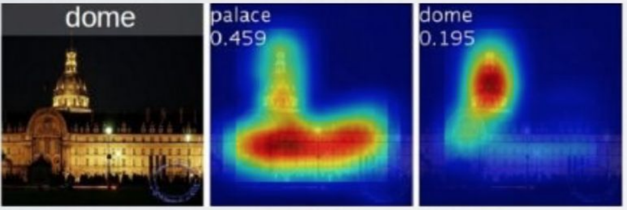
- Interpretability
  - GDPR
  - Liability for ADM

- Completeness
  - Explaining the wrong thing
  - Making decisions for the wrong reasons

- Taxonomy and Best Practices

Explaining Explanations [Gilpin et al, 2018]

Visual cues

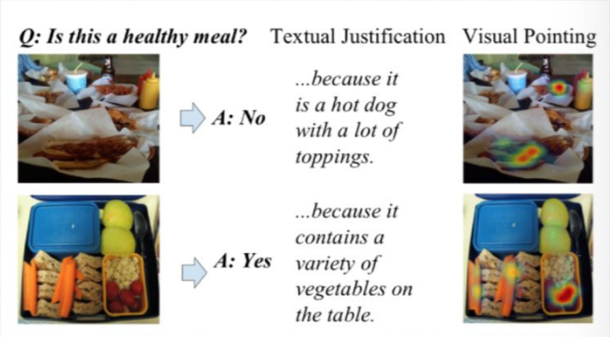


Completeness to model

Role of individual units

Completeness on other tasks

Attention based



Q: Is this a healthy meal? Textual Justification Visual Pointing

A: No *...because it is a hot dog with a lot of toppings.*

A: Yes *...because it contains a variety of vegetables on the table.*

Human evaluation

GDPR - Article 22 – **Automated individual decision-making, including profiling**

- The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
- Paragraph above shall not apply if the decision:
  - is necessary for entering into, or performance of, a contract between the data subject and a data controller;
  - is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
  - is based on the data subject's explicit consent.
- In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

What this Means?

You must identify whether any of your data processing falls under Article 22 and, if so, make sure that you:

- Give individuals information about the processing;
  - If you are using ML, you at least need interpretability
- Introduce simple ways for them to request human intervention or challenge a decision;
  - If you are using ML, you may need to explain
- Carry out regular checks to make sure that your systems are working as intended.

Legal and Ethical Issues

- Ethical codes for developers, companies and robots?
- No gray areas for legal accountability?
- Can we have an international consensus?

- Plenty of moral dilemmas
  - Example: Self-driving car having to choose between harming passengers or pedestrians

- Too many factors involved …..

- Ethical algorithms? Proving explanations?

ACM USA Statement on Algorithm Transparency and Accountability (Jan 2017)

1. Awareness
2. Access and redress
3. Accountability
4. Explanation
5. Data Provenance
6. Auditability
7. Validation and Testing

GDPR
EU New Copyright Directive

They do not have to be Perfect,
Just Better than Humans

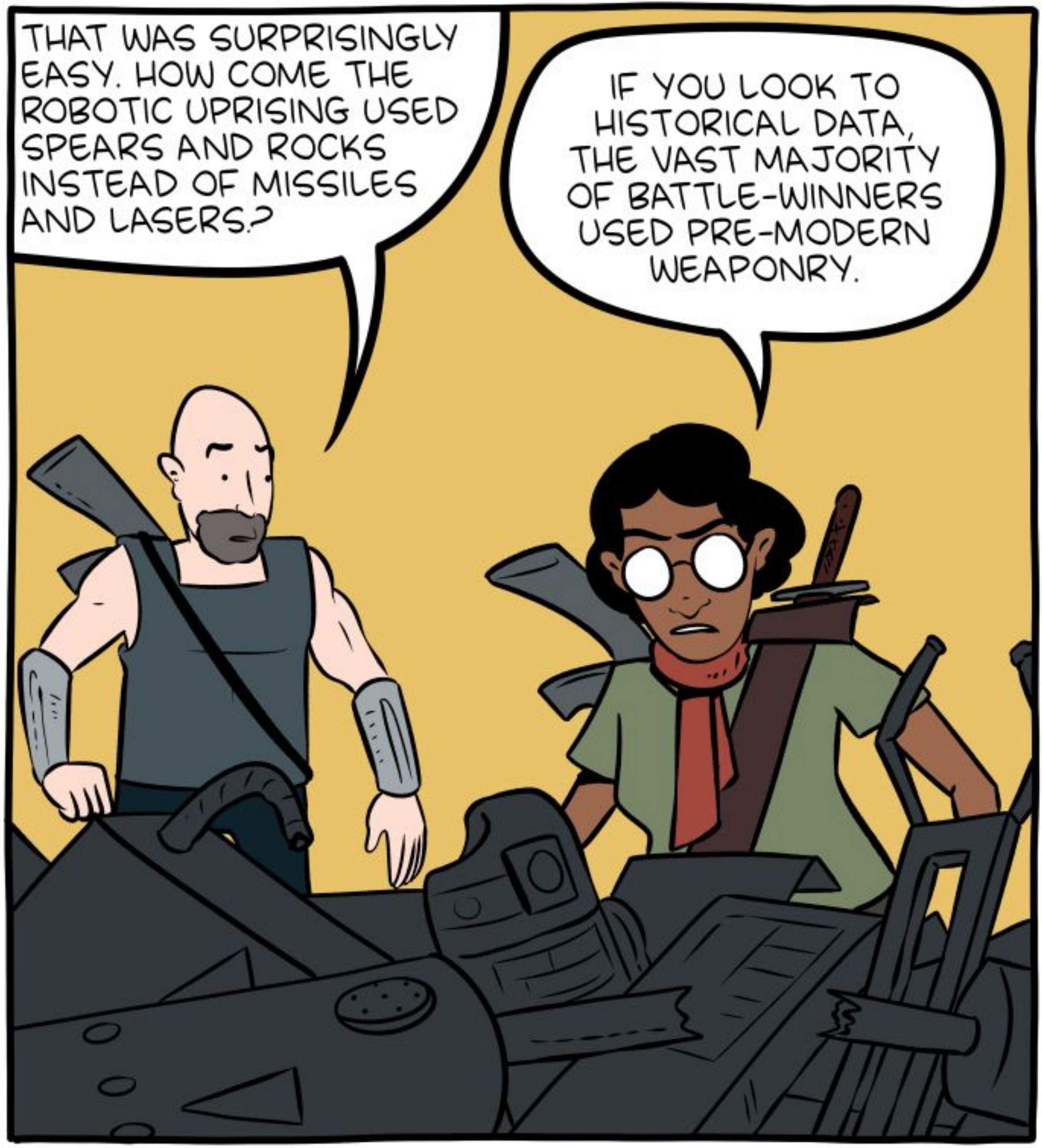
Many other similar initiatives

Bad (Human) Practices

- Learn from the Past Without Remembering the Context
- Learn from Humans Without Remembering Human Bias and the Possibility of Malicious Training
- Not Checking for Spurious Correlation/Proxies for Protected Information
- Code Reused in Unanticipated Contexts
- Tendency to Aggressively Resist Review
- Inappropriate Relationship of Human Decision Maker to System
- Failing to Measure Impact of Deployed System
- Individual Personalization instead of Personas
  - Trade-off with privacy
- Inaccurate Data or Just Data that you Have

Partially based in
[Matthews, 2019]

But Mistakes are Not Always Bad!

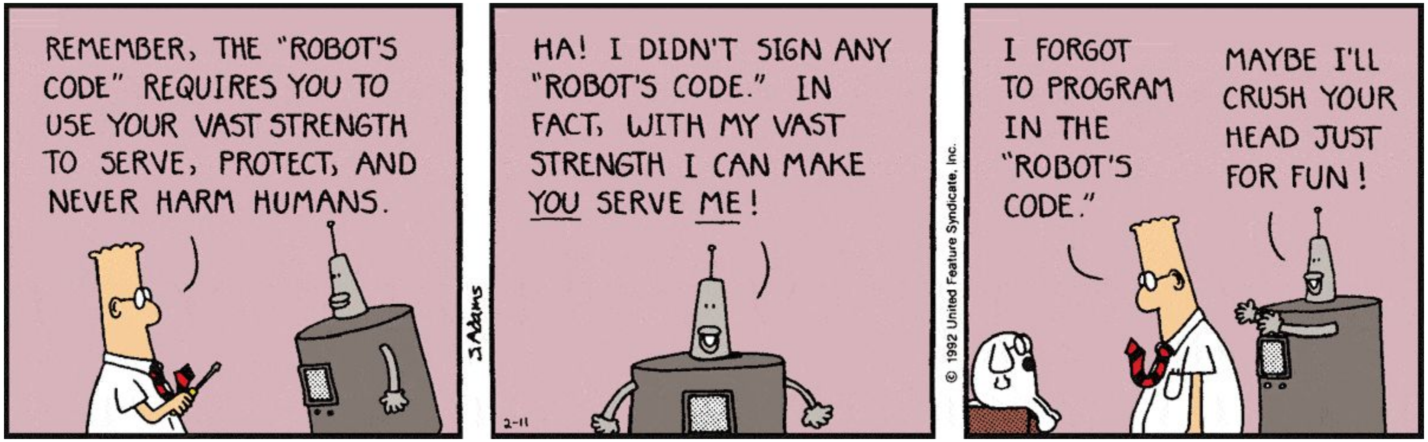

THAT WAS SURPRISINGLY EASY. HOW COME THE ROBOTIC UPRISING USED SPEARS AND ROCKS INSTEAD OF MISSILES AND LASERS?
IF YOU LOOK TO HISTORICAL DATA, THE VAST MAJORITY OF BATTLE-WINNERS USED PRE-MODERN WEAPONRY.

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.

Hard Learning Problems

- **Hard to Forget** what You Learn!
  - "Funes, The Memorious" (Borges, 1942-44)
  - The brain is very good at filtering

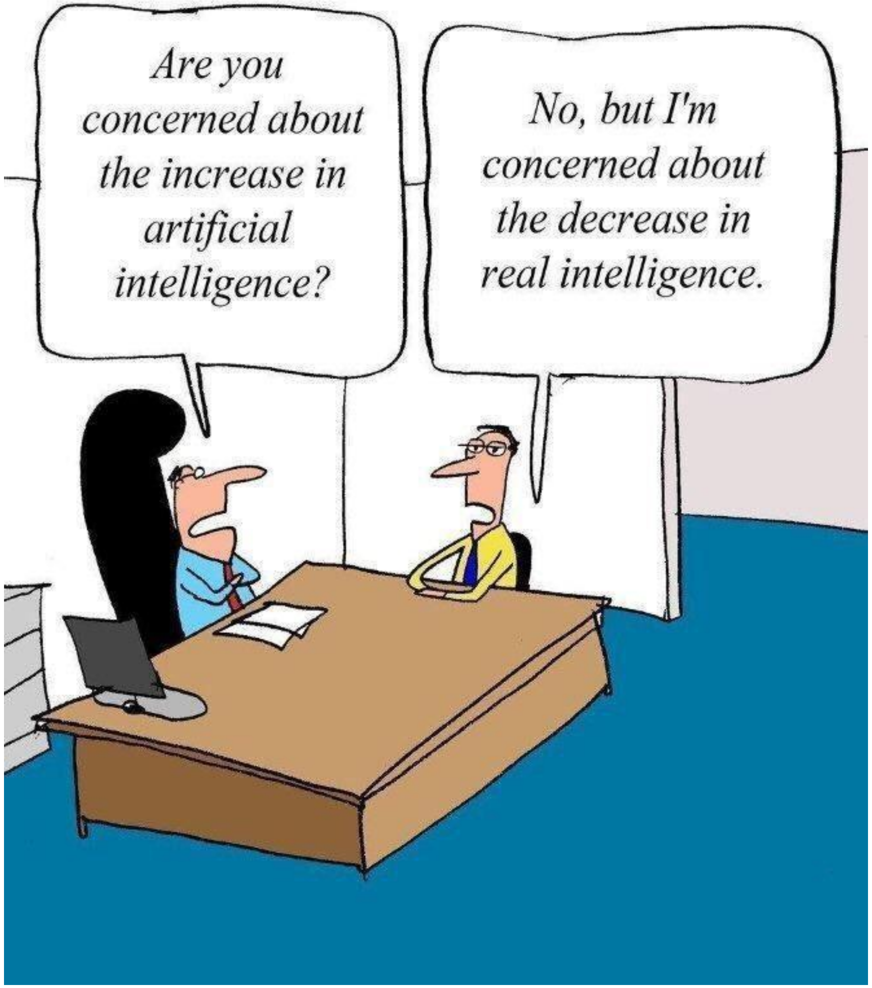
- You **Cannot Learn** what is not in the Data!

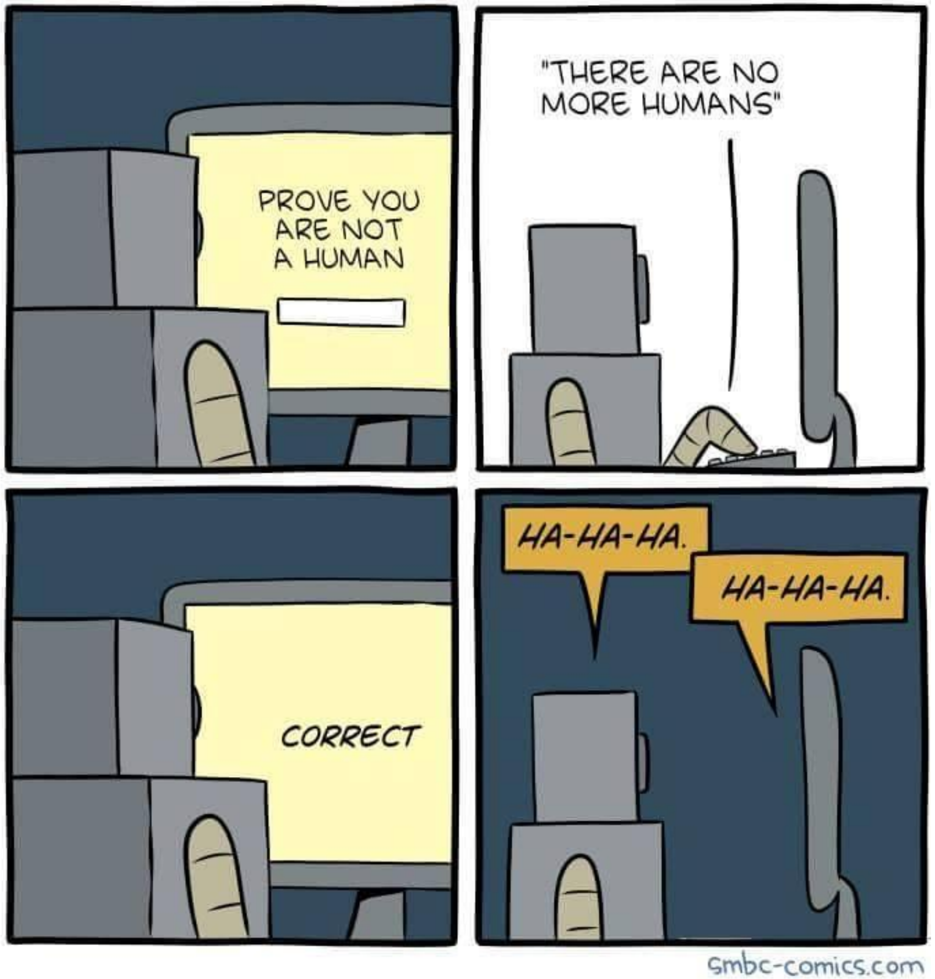

REMEMBER, THE "ROBOT'S CODE" REQUIRES YOU TO USE YOUR VAST STRENGTH TO SERVE, PROTECT, AND NEVER HARM HUMANS.
HA! I DIDN'T SIGN ANY "ROBOT'S CODE." IN FACT, WITH MY VAST STRENGTH I CAN MAKE YOU SERVE ME!
I FORGOT TO PROGRAM IN THE "ROBOT'S CODE."
MAYBE I'LL CRUSH YOUR HEAD JUST FOR FUN!
S.Adams
2-11
© 1992 United Feature Syndicate, Inc.

Recommendations for Us

- Design for People First!
- Deep Respect for Limitations of Our Systems
  - Assumptions, ethical risks, etc.
- Learning from the Past does not mean to Reproduce It
- Have and Enforce a Code of Ethics
- Improve Explainability (repeat 100 times)
- More evaluation and cross-discipline validation
- Research Best Practices with Humans in Control and **Machines in the Loop**
  - Don’t Use “Human in the Loop”!
- Study more Intelligence and Consciousness, not only AI
- Upgrade some Jobs like Teaching or Children Care

The Future?


Are you concerned about the increase in artificial intelligence?
No, but I'm concerned about the decrease in real intelligence.



"The real problem with robots is not their AI intelligence but the **natural stupidity** and **cruelty** of their human masters" [Harari, 2018]

A Dark Future?

- Infotech + Biotech [Harari, 2018]
- Free Will is an Illusion
- Humans can be hacked

Just Easy Parts (Politics?)

Emotions are predictions
[Feldman Barrett, 2017]

- Loss of Jobs

Leverage AI

More Literature & Art

When they are better than humans

- Loss of Skills
- Integrated Complex Machine Network versus Individuals
- Authority Switches to Algorithms and Owners of Our Data
- Even More Inequality
- No Sense of Purpose
- Irrelevance

What We can Do?

Borrowing from Harari (Stanford HCAI Institute, March 2019)

- You: Learn yourself better
  - Accept yourself earlier

- Us: Work in decentralized systems
  - Go beyond blockchain, we can do better!

- All: Individual AI armors
  - They warn us when we are being manipulated
  - They warn us about our biases
  - They help us to be better
  - Even though there will be more powerful AI, we have more data about ourselves

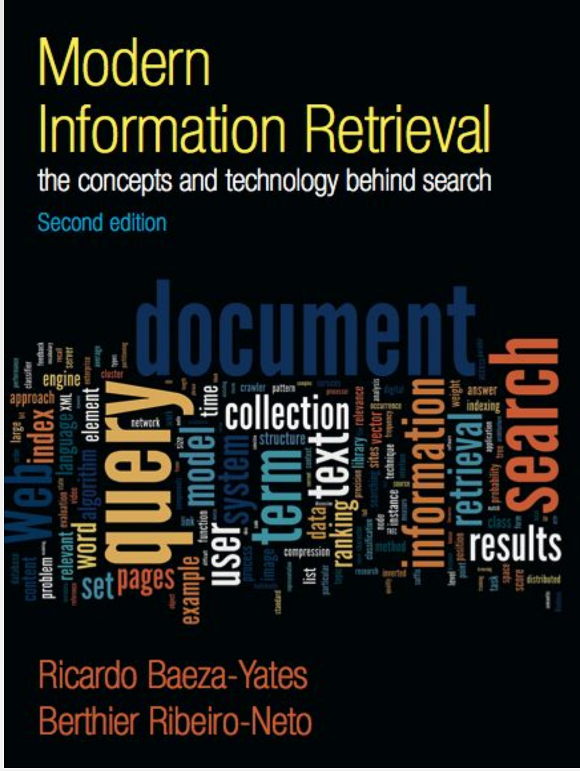
Epilogue

- BIG PICTURE: Integration
- No Privacy, e.g., Health
- Explainable/Transparent Algorithms?
- Software Insurance (my worst nightmare)
- Ethics for Robots?
- Remote Knowledge Workers
- Augmented Humanity?

"Either democracy will successfully reinvent itself in a radically new form or humanity will live in 'digital dictatorships'", Harari 2018

- Still, technological change is overall good!
  - USA 2016, Philippines 2017, Brazil 2019, China 2020?
- But, are we evolving towards Solaria?
  (*The Naked Sun*, Asimov)
- If there are nice aliens out there, please come soon!
  - See "Arrival"

ASIST 2012 Book of the Year Award



Biased Ad

Questions?

Biased Questions?

Contact: rbaeza@acm.org

Explanations?