Conversational Explanations

Explainable AI through human-machine conversation

Dave Braines
CTO Emerging Technology, IBM Research UK

Industry Technical Area Leader, DAIS ITA Research Program

PhD student @
Cardiff University Crime & Security Research Institute

Full material from the original 3 hour tutorial can be downloaded from: bit.ly/conv_exp

Original Agenda

- Introductions *[10]*
- Explanations
  - Scene setting for Explainable AI (XAI) *[20]*
  - Philosophy & Social Science *[20]*
- Collaborative XAI research examples *[10]*

(Coffee break)

- Deep learning – black box explanations *[20]*
- The role of the user *[20]*
- Conversational Explanations *[20]*
- Visual Exploration of Deep Learning *[20]*

Agenda for today

- Introductions ~~[10]~~ [3]
- Explanations [10]
  - Scene setting for Explainable AI (XAI) ~~[20]~~
  - Philosophy & Social Science ~~[20]~~
- ~~Collaborative XAI research examples~~ ~~[10]~~

~~(Coffee break)~~

- Deep learning – black box explanations ~~[20]~~ [5]
- The role of the user ~~[20]~~ [2]
- Conversational Explanations [20] [10]
- ~~Visual Exploration of Deep Learning~~ ~~[20]~~

Introductions

About me

dave_braines@uk.ibm.com

davebraines

davebraines

bit.ly/dbpubs

Active researcher in Artificial Intelligence.

Currently focused on Machine Learning, Deep Learning and Network Motif analysis.

Published 100+ conference/journal papers.

Interested in human-machine cognitive interfaces for deep interactions between human users and machine agents.

Likes kayaking, walking and camping.

Senior Certified Technical Specialist.

Part-time PhD student.

Emerging Technology Researcher.

Emerging Technology, IBM Research

Delivering leading edge innovation for our clients

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

Crime and Security Research Institute
About us

About us | Research | People | News | Publications | Executive education | Events

About us

About us

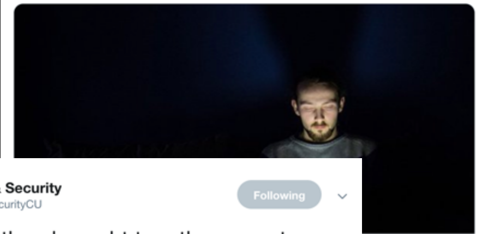
The Crime and Security Research Institute brings together Cardiff University's significant interdisciplinary research expertise in the fields of crime and security.

The effective management of crime and security is one of the biggest challenges we face in today's world. Our response to this challenge is to conduct research on a local and global scale, combining existing academic excellence from within the Universities Police Science Institute, the Violence Research Group and the Informatics and Visual Computing Research Groups in a dynamic new initiative.

We will foster creative and innovative conceptual and methodological approaches to shape policy and practice development in relation to crime and security challenges locally, nationally and internationally; we are committed to sustaining a record of achieving real-world impact as well as addressing community concerns.

Crime & Security @CrimeSecurityCU Following

Our researchers have identified three prominent techniques used on social media in the aftermath of terrorist violence to influence public perceptions, reactions and values. Read their recent @LSEpoliticsblog to find out more



...r terrorist attacks
...l communications
...or constrain their ...

Crime & Security @CrimeSecurityCU Following

Our hackathon brought together experts from police, computer science and other agencies to address real security issues. If your organisation would like to run a hackathon, check out our new video:

Policing Futures: An Evidence Based Policing Programme
The Policing Futures Masterclass Series is a unique collaboration between the Universities' Police Science Institute (UPSI) and South Wales Police (SWP), des...
youtube.com

9:51 AM - 21 Mar 2019

1 Like

Crime & Security @CrimeSecurityCU Following

The 'Cardiff Model' enables intelligence led policing which reduces violent crime, but more support is needed from government - Last night @BBCMarkEaston highlighted our initiative #KnifeCrime

BBC News at Ten - 06/03/2019
Latest national and international news, with reports from BBC correspondents worldwide.
bbc.co.uk

9:42 AM - 7 Mar 2019

2 Retweets 5 Likes

2 5

Improving Situational Understanding for Human/Machine Hybrid Teams

Dave Braines (BrainesDS@cardiff.ac.uk), 1st year PhD (part time)

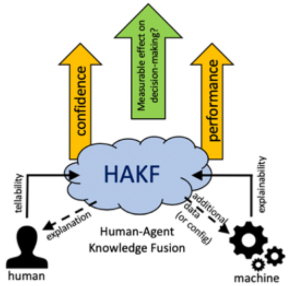
Supervisors: Prof. Alun Preece, Prof. Ian Taylor

Background

Machine-agent performance & human-agent confidence are increased in hybrid human-machine systems with dynamic feedback between human & machine agents.

Human Agent Knowledge Fusion (HAKF) is the mechanism proposed to facilitate this dynamic feedback exchange, with:

- ***Explainability*** providing feedback from machine agents to human users. Specifically, a description of the reasoning or processing used to reach the conclusion. This can relate to the algorithms and processes used, or can be post-hoc explanation in cases where the processing is "black box" or the algorithm details should not be shared.



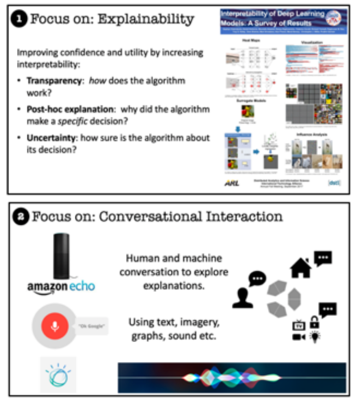
- ***Tellability*** from the human users to the machine agents. For example to provide additional local knowledge or guidance, especially in sparse data situations which may be common in rapidly evolving situational understanding environments. This is greater than simply enhancing the training data as the situation unfolds.

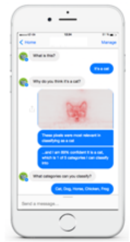
All of the above is in the context of *rapidly formed small coalition teams* with human and machine agents, operating at the *edge of the network*, with *limited connectivity, bandwidth* and *compute resources* in a *decision-making* role.

Hypothesis

Systems with *explainability will increase human-agent confidence*, and systems with *tellability will increase machine-agent performance*.
Hybrid systems with improved confidence and performance will have a measurable effect on decision making.

Narrowing the scope





Conversational explanations

Bringing together: ***explainability*** which is provided by the machine agents in the conversation, and ***tellability*** through the human agents correcting, configuring, and providing contextual information or local knowledge to improve the system.

Key 2018 Publications

All publications are collaborative, sponsored by the DAIS ITA research program. See http://sl.dais-ita.org for full details.

1. Braines, D., Preece, A., & Harborne, D. (2018). **Multimodal Explanations for AI-based Multisensor Fusion.** In *NATO SET-262 RSM on Artificial Intelligence for Military Multisensor Fusion Engines in Budapest, Hungary.*
2. Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). **Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems.** In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018) in Stockholm, Sweden.*

Next steps

We conducted a workshop in Nov 2018 with military experts using the Design Thinking method to elicit multiple use cases for AI Explainability.

1. Complete workshop write up
2. Extend meta-model for AI Explanations
3. Refine experimental user interface
4. Plan and get approval for human trials

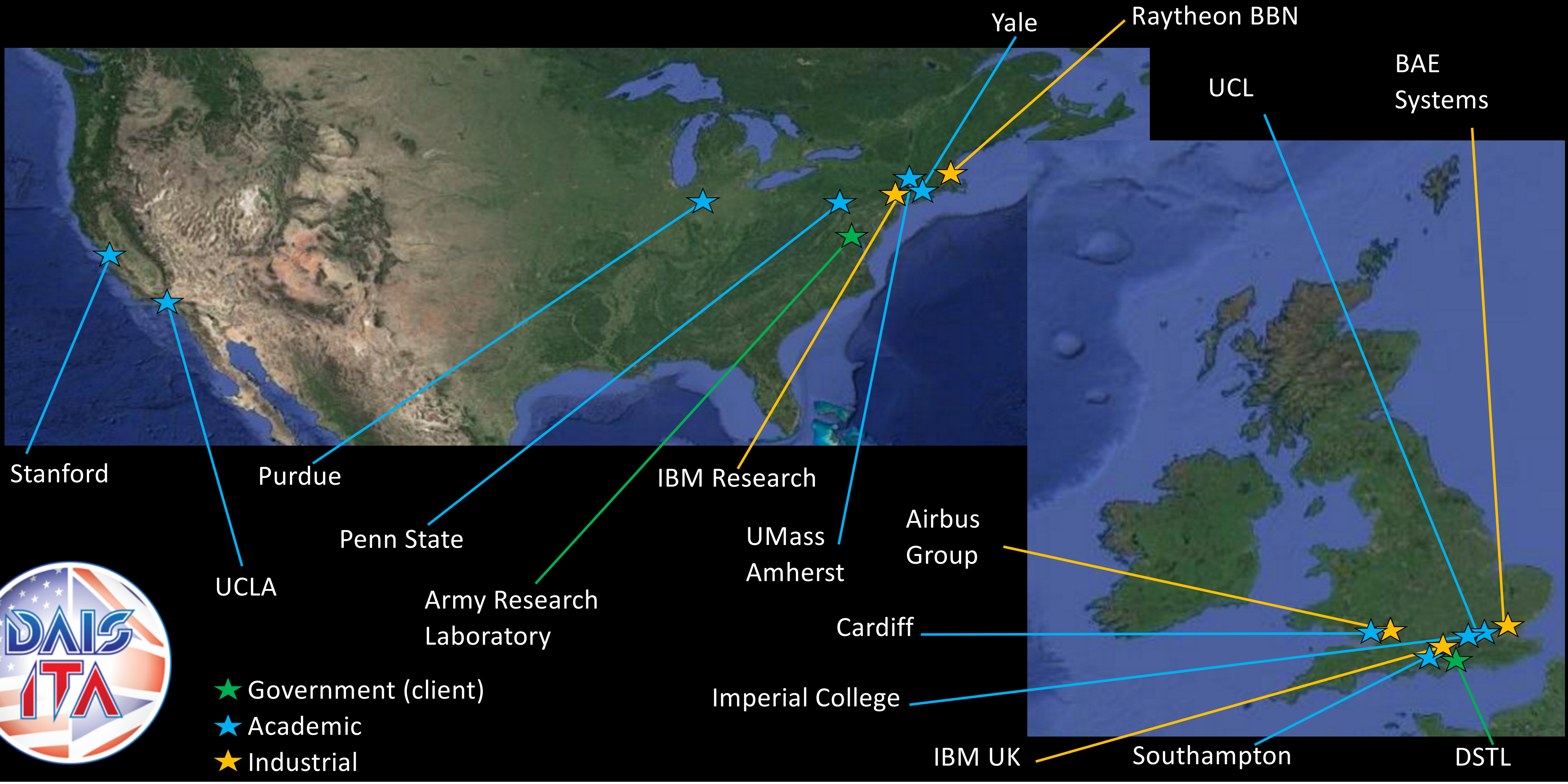
Design Thinking Workshop for AI Explanations with military stakeholders at IBM Hursley, Nov-2018

Acknowledgement

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Distributed Analytics and Information Science

International Technology Alliance

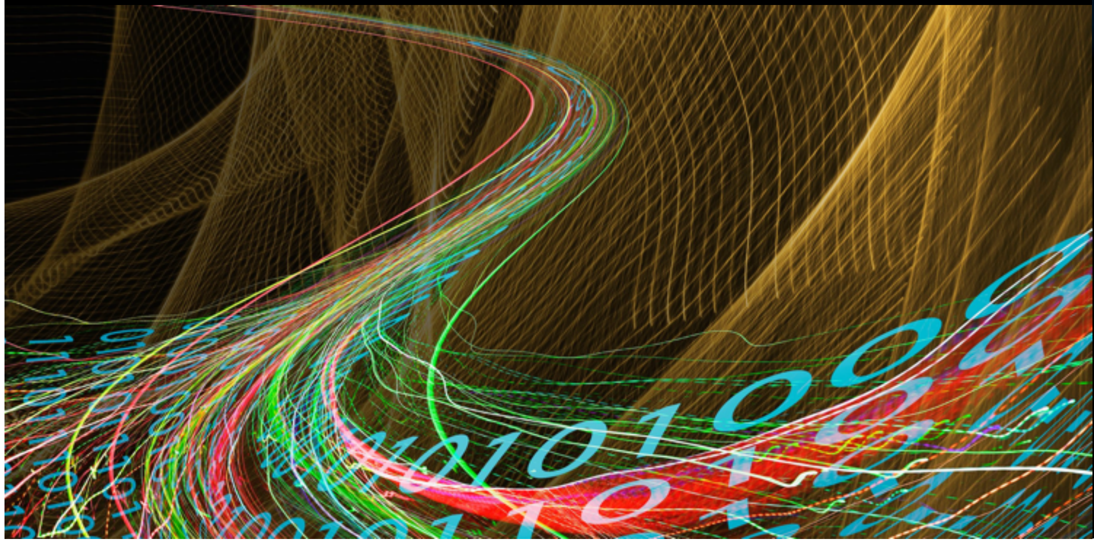


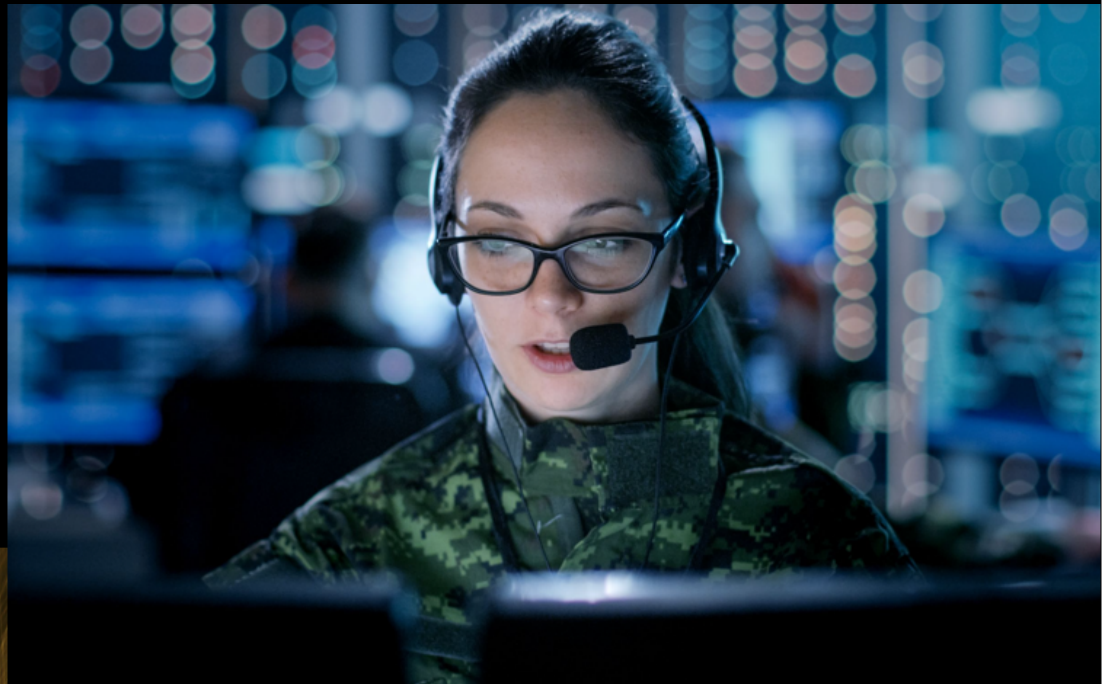
Focused on rapidly formed coalitions

Running at the edge of the network

Two Technical Areas:

Dynamic, Secure Coalition Information Infrastructures



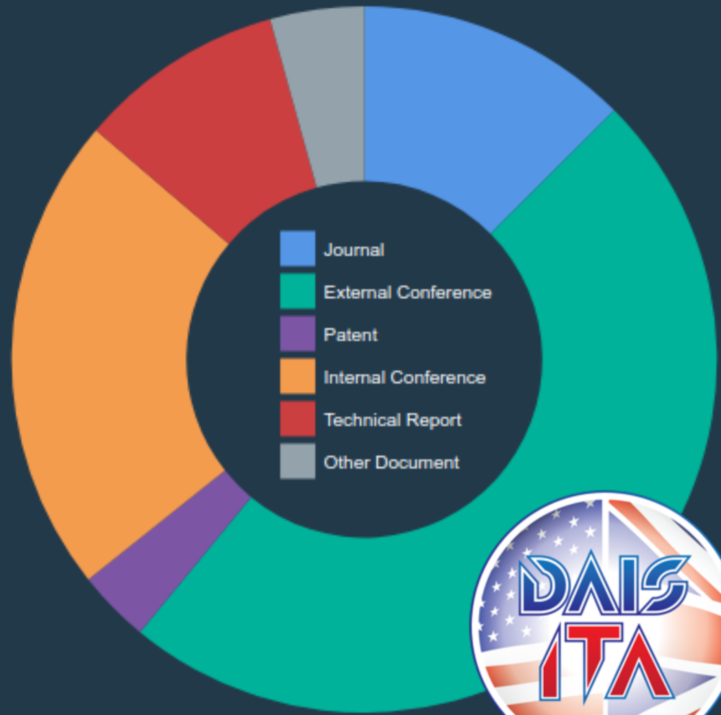


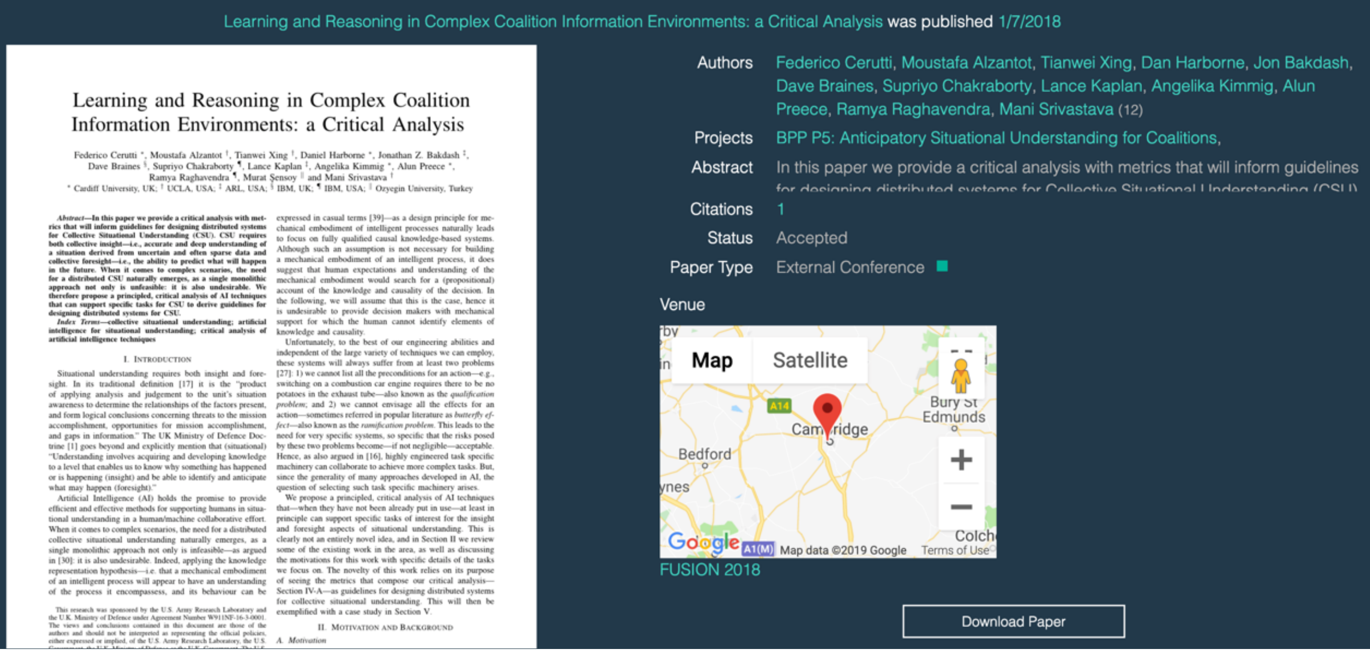
Coalition Distributed Analytics & Situational Understanding

All DAIS publications available online

bit.ly/sciencelibrary

Total (External)	1061
Journals	207
External Conferences	799
Patents	55
Internal Conferences	362
Technical Reports	156
Other Documents	71





Explainable AI

If we want to use AI
does it need to
explain itself?

Defining AI

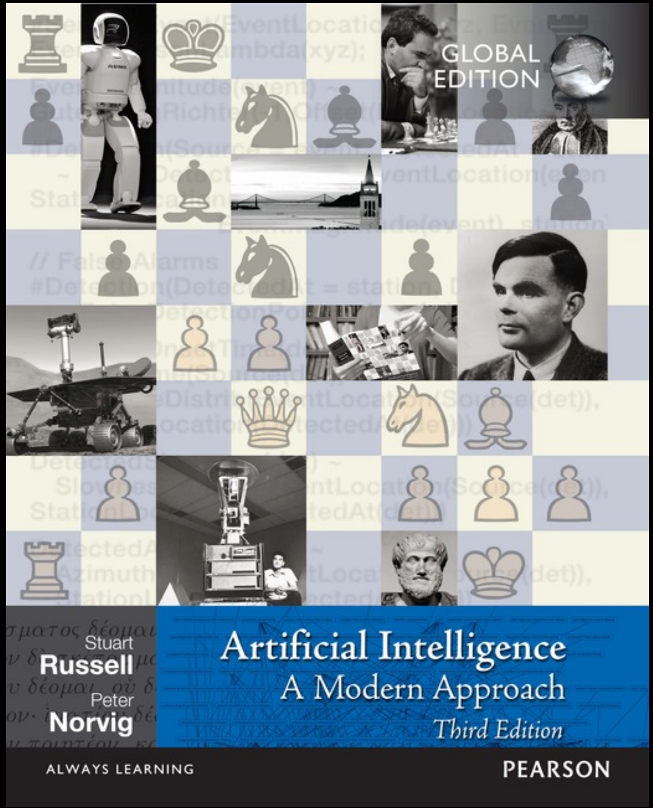
Artifacts that act like humans

Artifacts that think like humans

Artifacts that act rationally

Artifacts that think rationally

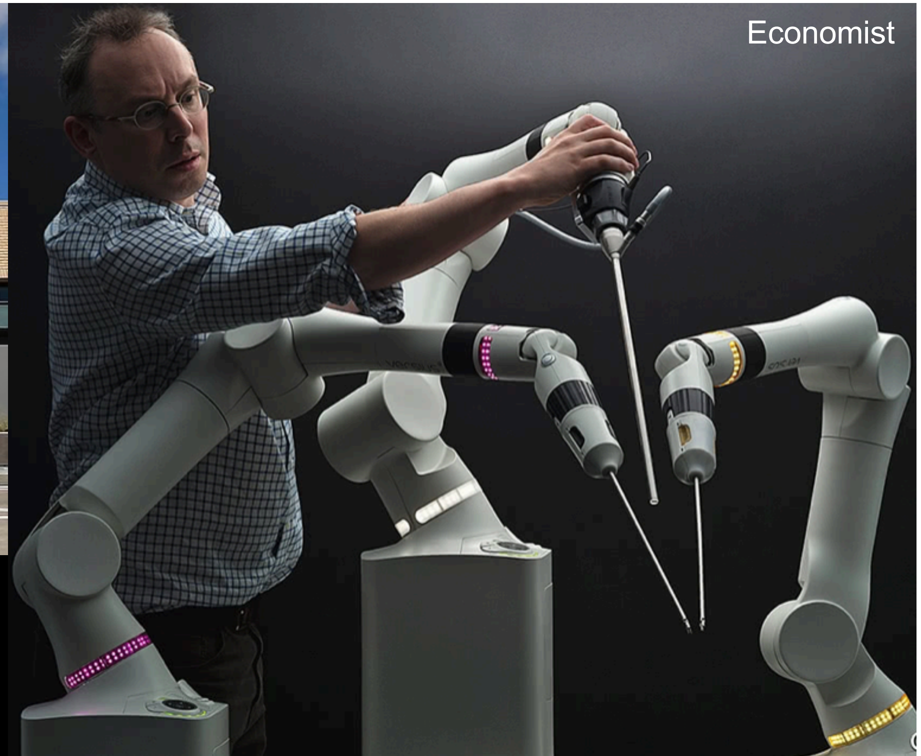
…but we're not considering Artificial General Intelligence (AGI) today



S Russell & P Norvig, **Artificial Intelligence: A Modern Approach** (3rd ed), Prentice Hall, 2009.

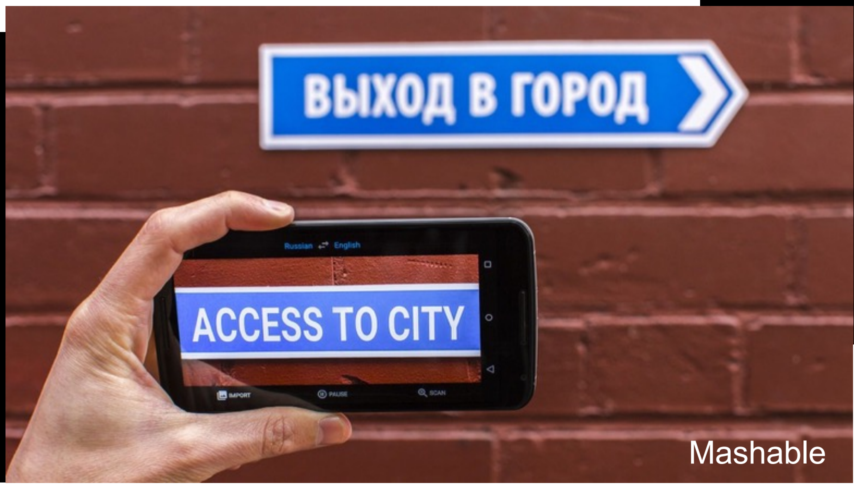
Technology Intelligence

Google computer becomes first non-human to officially qualify as car driver

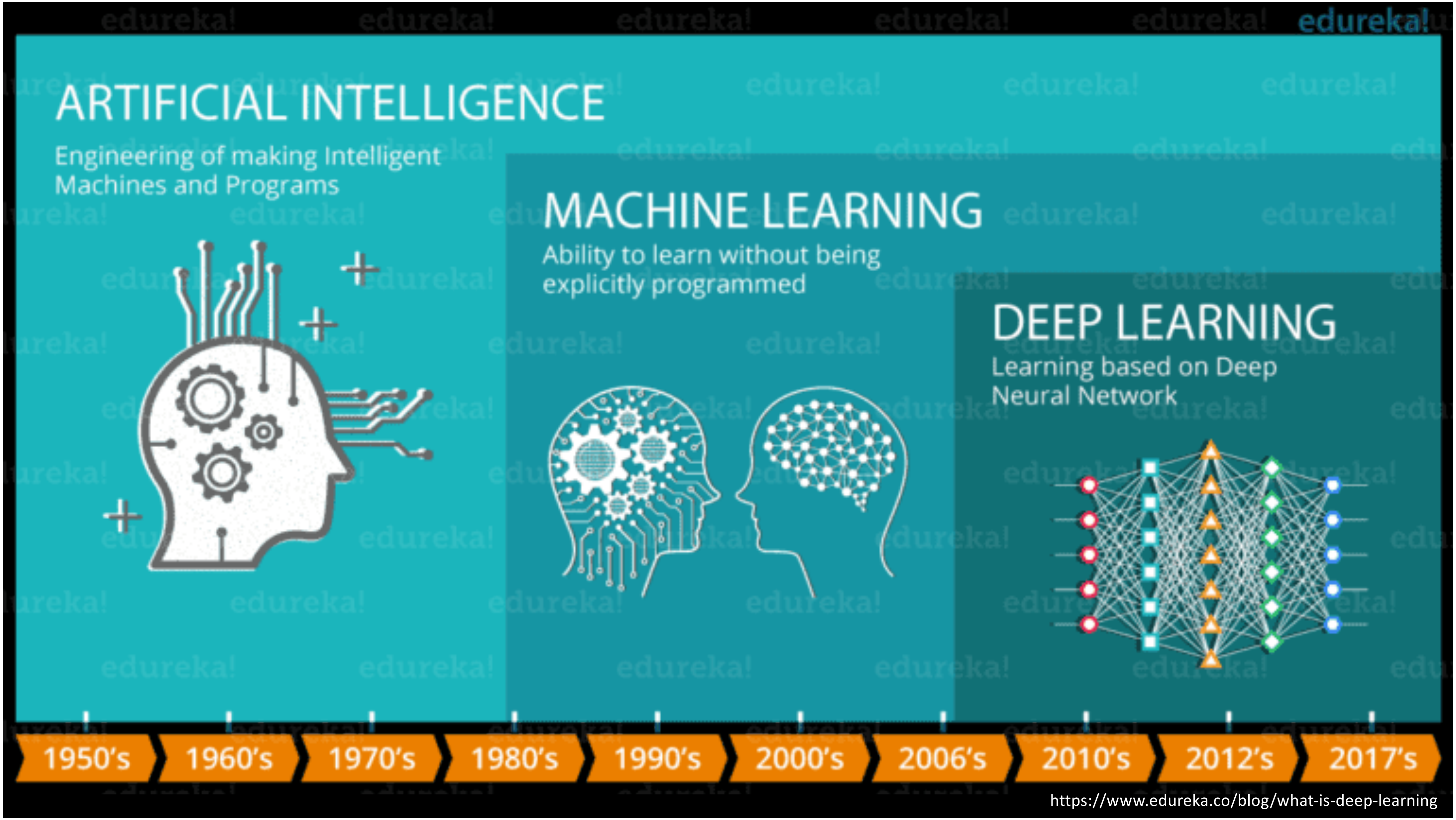


Medicine

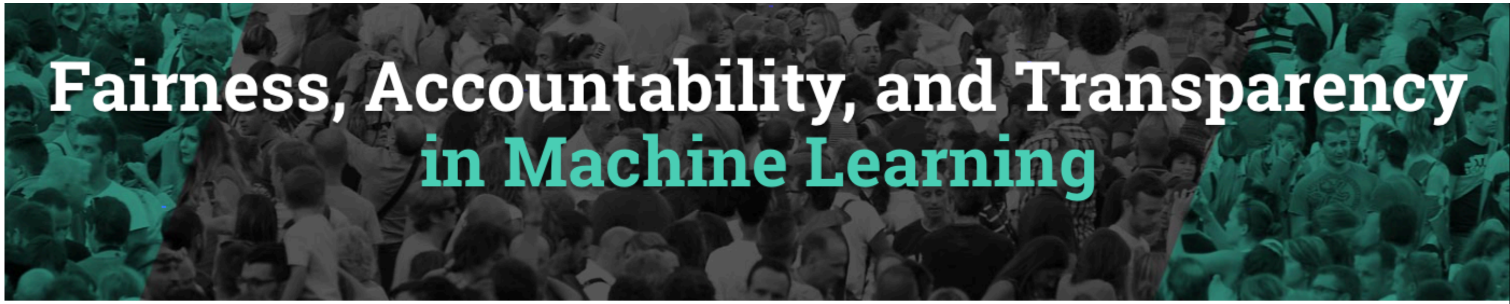
New surgical robots are about to enter the operating theatre



Google Translate gets smarter with language detection, Word Lens



ARTIFICIAL INTELLIGENCE
Engineering of making Intelligent Machines and Programs
MACHINE LEARNING
Ability to learn without being explicitly programmed
DEEP LEARNING
Learning based on Deep Neural Network
1950's
1960's
1970's
1980's
1990's
2000's
2006's
2010's
2012's
2017's
https://www.edureka.co/blog/what-is-deep-learning

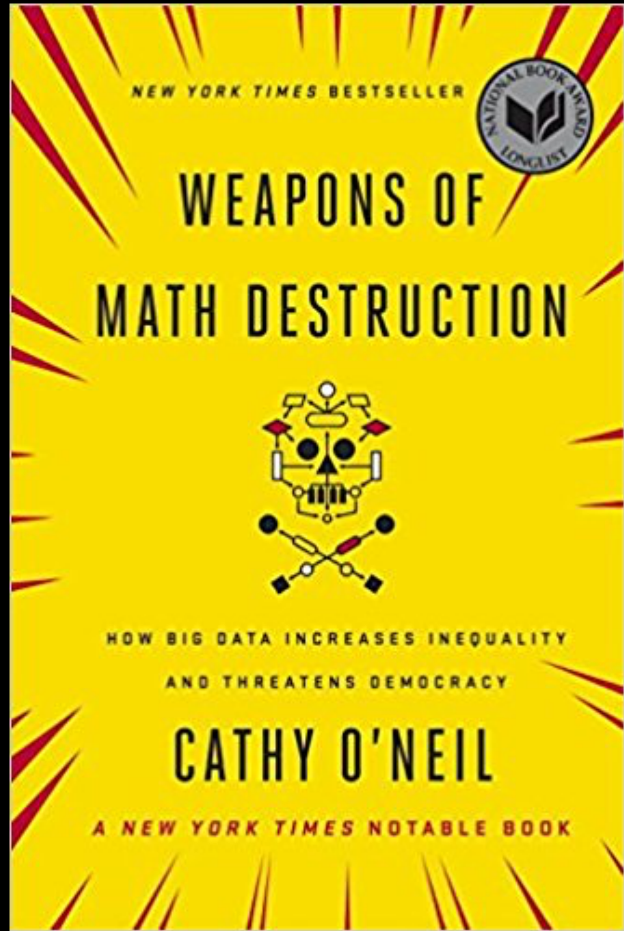


http://www.fatml.org

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

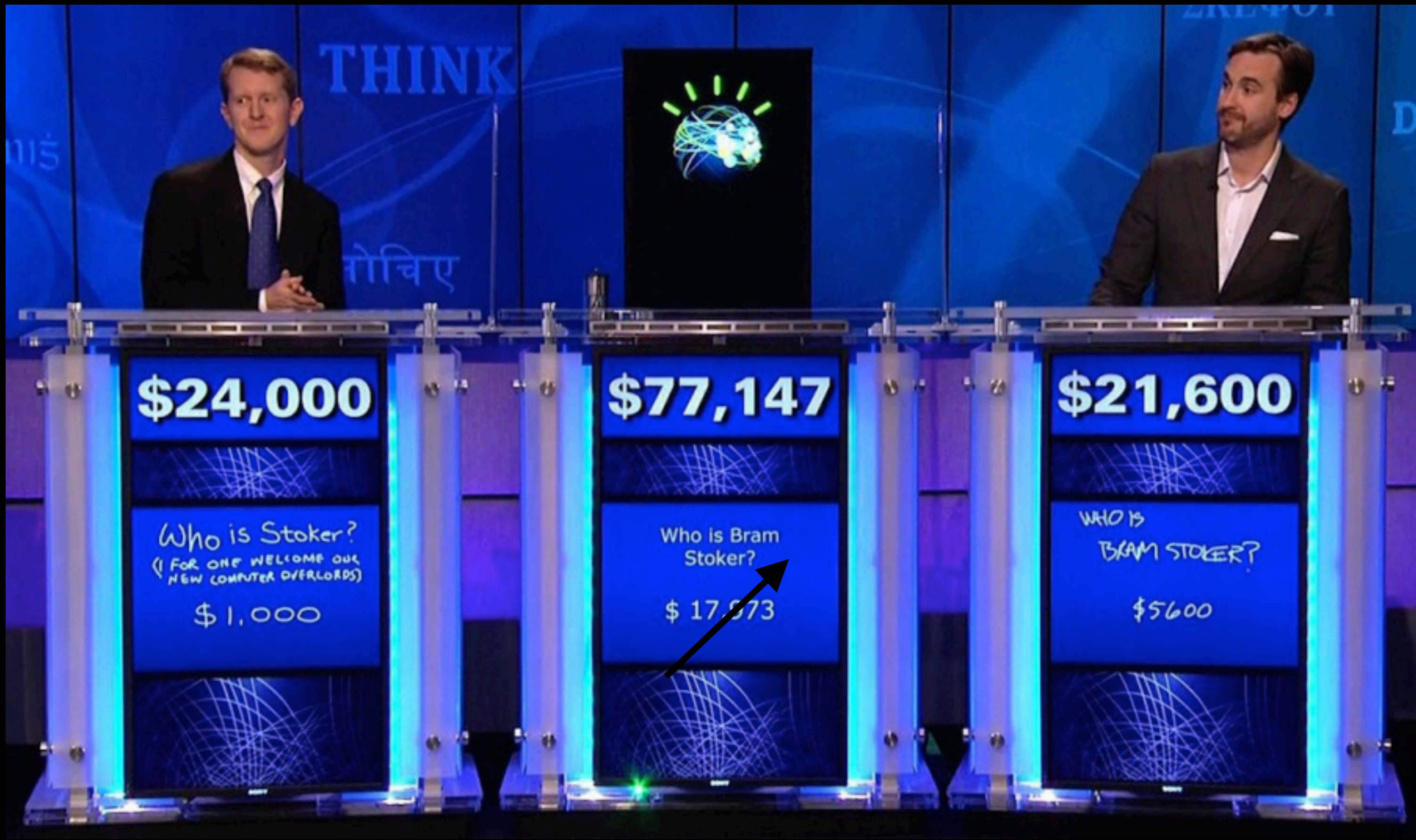
At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to "the algorithm made me do it."



C O'Neill, **Weapons of Math Destruction**, Crown, 2016.

Watson (2011)

Breakthrough in "deep" question-answering via an ensemble of methods including NLP, ML, KRR …



IBM Research, 2011

A key idea was that Watson tackled input questions using multiple strategies and needed a method to weigh up its certainty.

Watson (2011)

Breakthrough in "deep" question-answering via an ensemble of methods including NLP, ML, KRR …

IBM Research, 2011

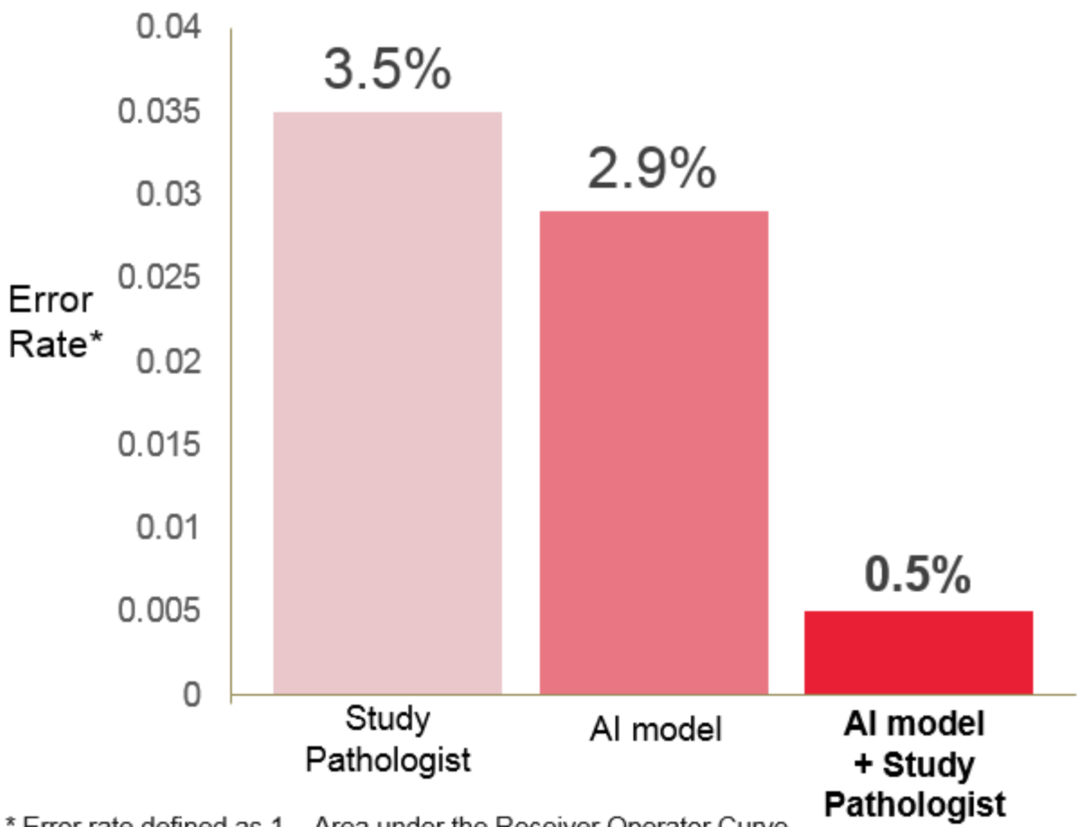
A key idea was that Watson tackled input questions using multiple strategies and needed a method to weigh up its certainty.

In chess, as in so many things, what computers are good at is where humans are weak, and vice versa. This gave me an idea for an experiment. What if instead of human versus machine we played as partners?

Garry Kasparov, *NY Review of Books*, 2010

"Centaur chess"

(AI + Pathologist) > Pathologist



* Error rate defined as 1 – Area under the Receiver Operator Curve
** A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.

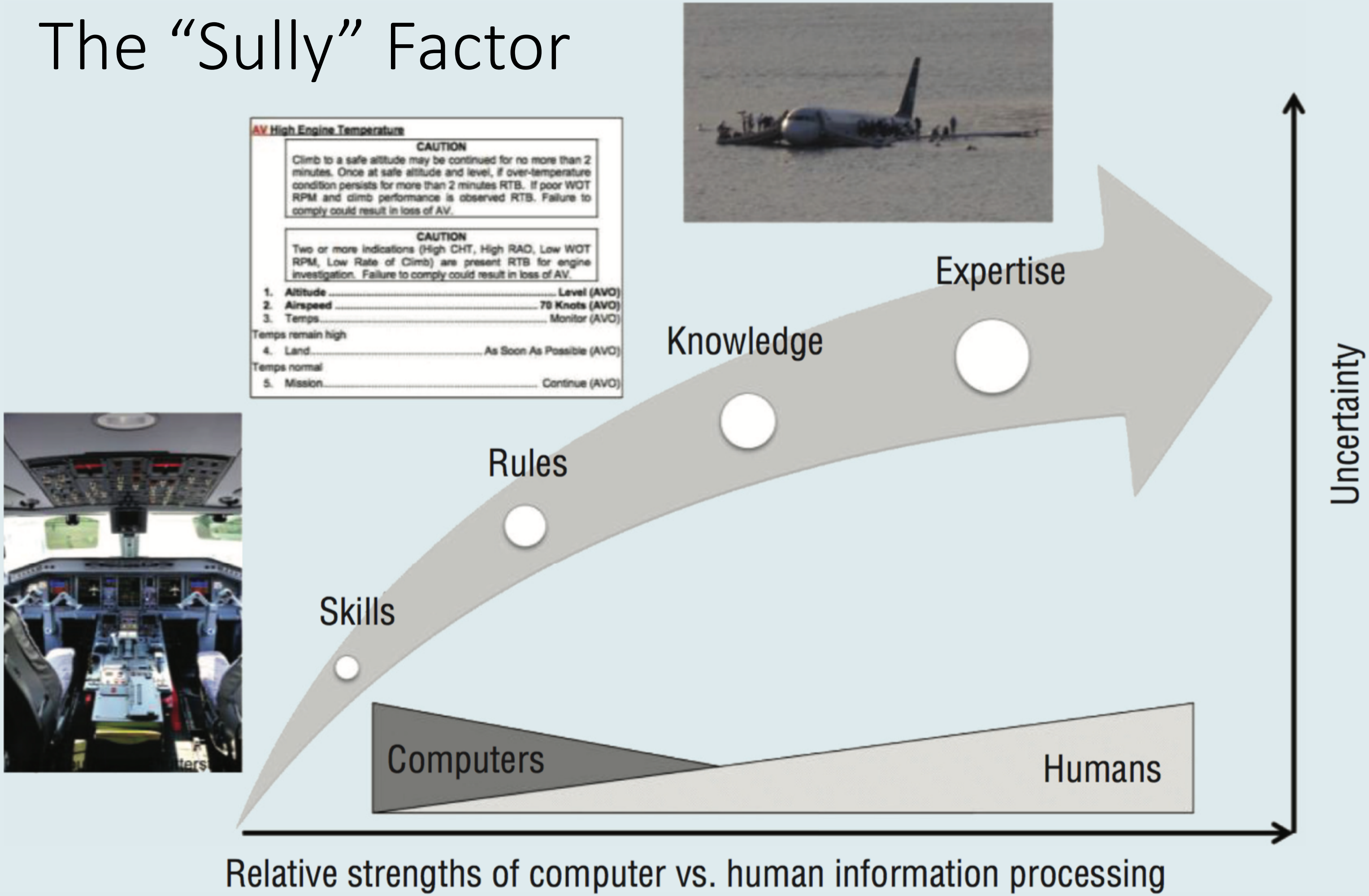
LIVESCIENCE

AI Boosts Cancer Screens to Nearly 100 Percent Accuracy

By Christopher Wanjek | June 21, 2016 01:54pm ET

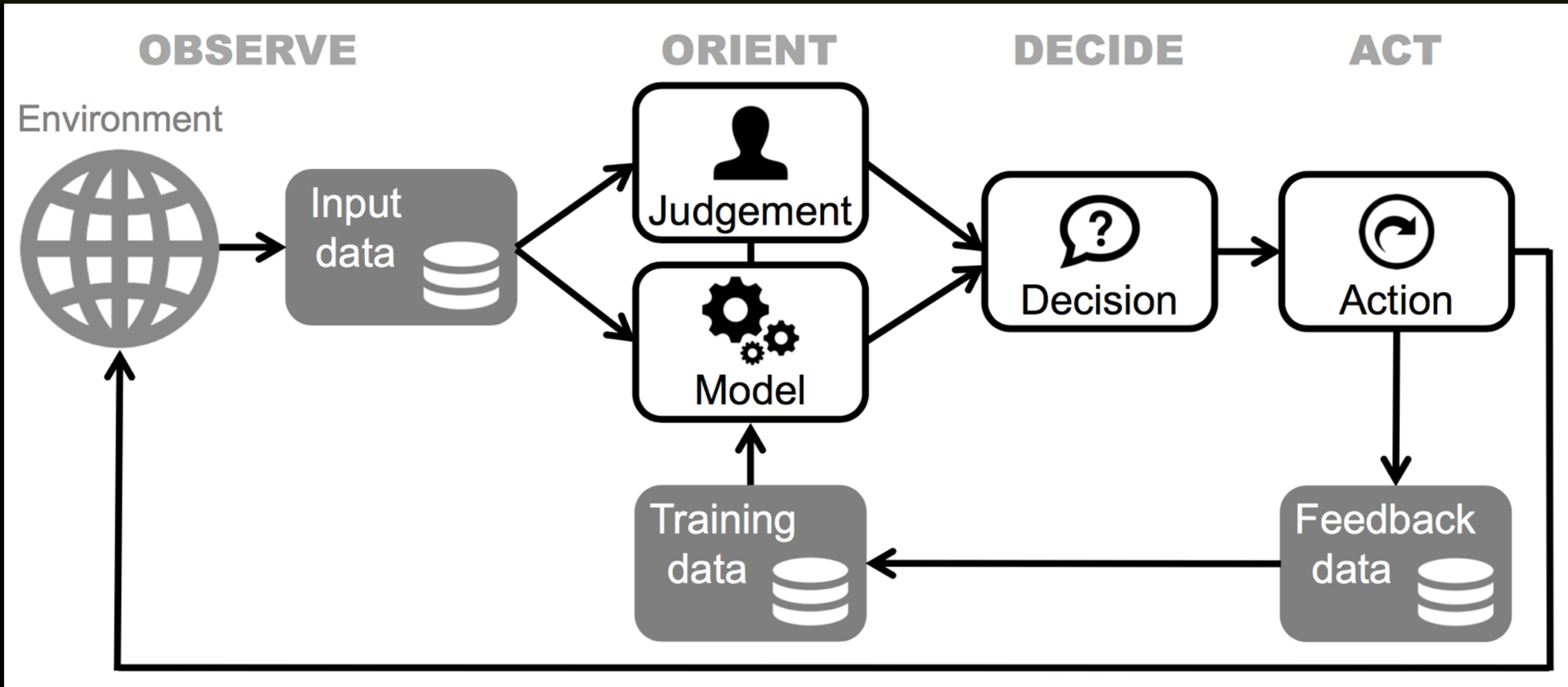
But the real surprise came when pathologists were teamed up with the Harvard team's AI. Together, the artificial intelligence and good, ole human intelligence identified 99.5 percent of the cancerous biopsies.

The “Sully” Factor

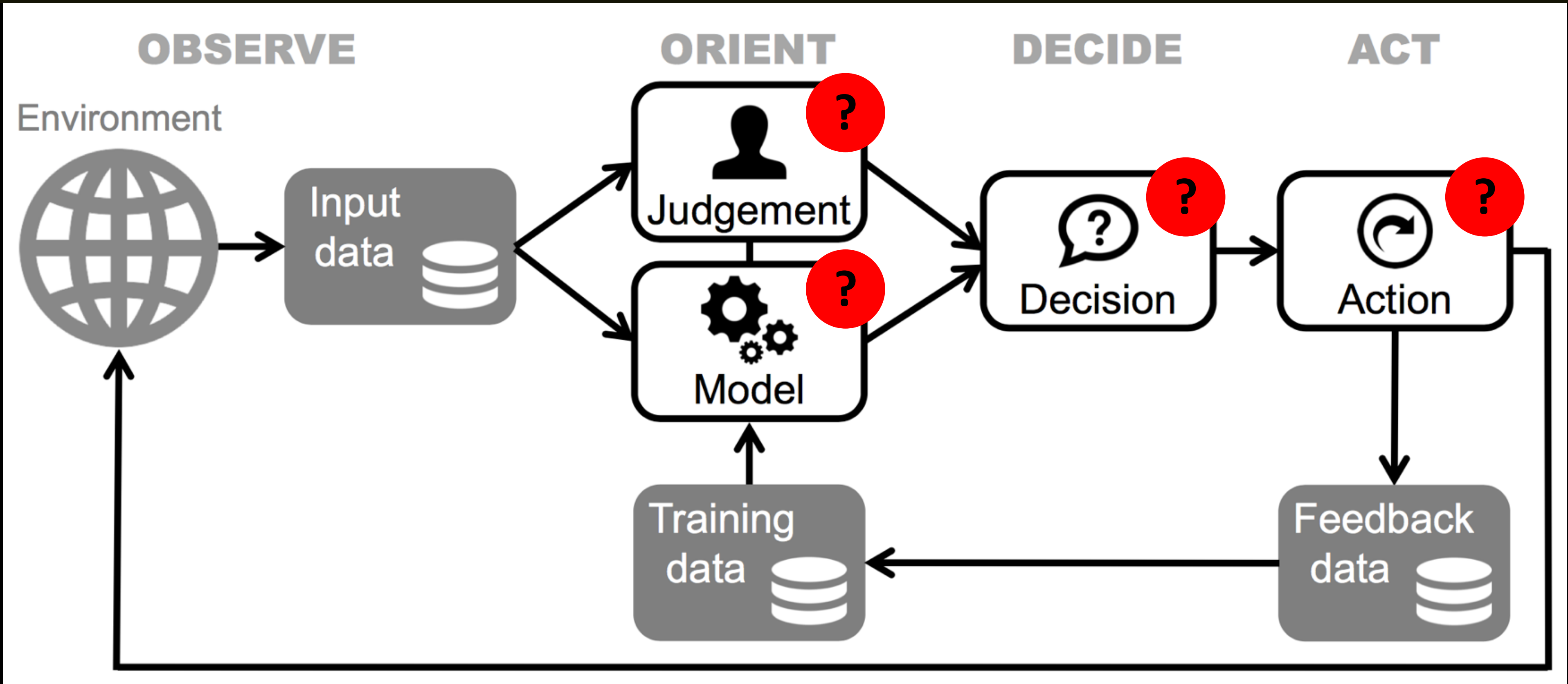


https://ieeexplore.ieee.org/document/6949509/

Human+machine decision loop



Explanation points



Explanations:

Philosophy and Social Science

Key publications

- Molnar, Christoph. "*Interpretable machine learning. A Guide for Making Black Box Models Explainable*", 2019. https://christophm.github.io/interpretable-ml-book/

- Miller, Tim. "*Explanation in artificial intelligence: Insights from the social sciences.*" *Artificial Intelligence* (2018).

Insights from the social sciences (Miller 2018)

- Humans prefer short explanations (1 or two causes)
- Contrastive explanations are best
  - Why this and not some other plausible outcome?
  - Abnormal causes are the best contrastive cases
- Explanations are selected
  - No need for a complete thorough list of causes
  - Beware: Selecting explanations can be inconsistent or contradictory
- Explanations are social interactions
  - The social context will drive the explanation content
- Explanations are truthful
  - …and match with prior beliefs
  - …and are generable and probable

Interpretability definitions

- "*Interpretability is the degree to which a human can understand the cause of a decision*" – Miller (2018)
- "*Interpretability is the degree to which a human can consistently predict the models result*"
- *"Interpretability: the level to which an agent gains, and can make use of, both the information embedded within explanations given by the system and the information provided by the system's transparency level."*

Interpretability considerations

- Importance/risk of a decision drives the need for interpretability
- There may be substantial additional costs for interpretability
  - As well as increased risks for privacy or adversarial attacks
- Interpretable models may be needed in cases where audit is required
  - These may be less powerful than “black box” alternatives
- Interpretation may be needed as part of the “answer”
  - In some cases the explanation qualifies the answer itself
- Decisions affecting humans or their wellbeing deserve explanations
  - GDPR has a right to explanation
- Not needed for well studies problems
- “Explanations in the wild” are becoming more commonplace

Related to interpretability

- Bias detection and mitigation
- Adversarial attacks; and defending against them
- Debugging and auditing
- Social acceptance
  - Especially of machine agents that are present in our lives
- Key considerations for interpretability:
  - Fairness
  - Privacy
  - Reliability
  - Causality
  - Trust

Interpretability methods

- Intrinsic (transparent) vs post-hoc
- Result types
  - Feature summary statistic
  - Feature summary visualization
  - Model internals
  - Data point
  - Intrinsically interpretable model
- Model specific or model agnostic
- Local or global

Interpretability techniques

- Supervised learning
  - Categorical -> classification
  - Numerical -> regression
- Interpretable models
- Model-agnostic methods
  - Surrogate models
  - LIME
  - Shapley/Shap
- Example-based explanations
- Ensemble models

Parting comment from Molnar (2019)

Robots and programs will explain themselves
We need more intuitive interfaces to machines and programs that make heavy use of machine learning. Some examples:

- A self-driving car that reports why it stopped abruptly
  ("70% probability that a kid will cross the road")
- A credit default program that explains to a bank employee why a credit application was rejected
  ("Applicant has too many credit cards and is employed in an unstable job")
- A robot arm that explains why it moved the item from the conveyor belt into the trash bin
  ("The item has a craze at the bottom")

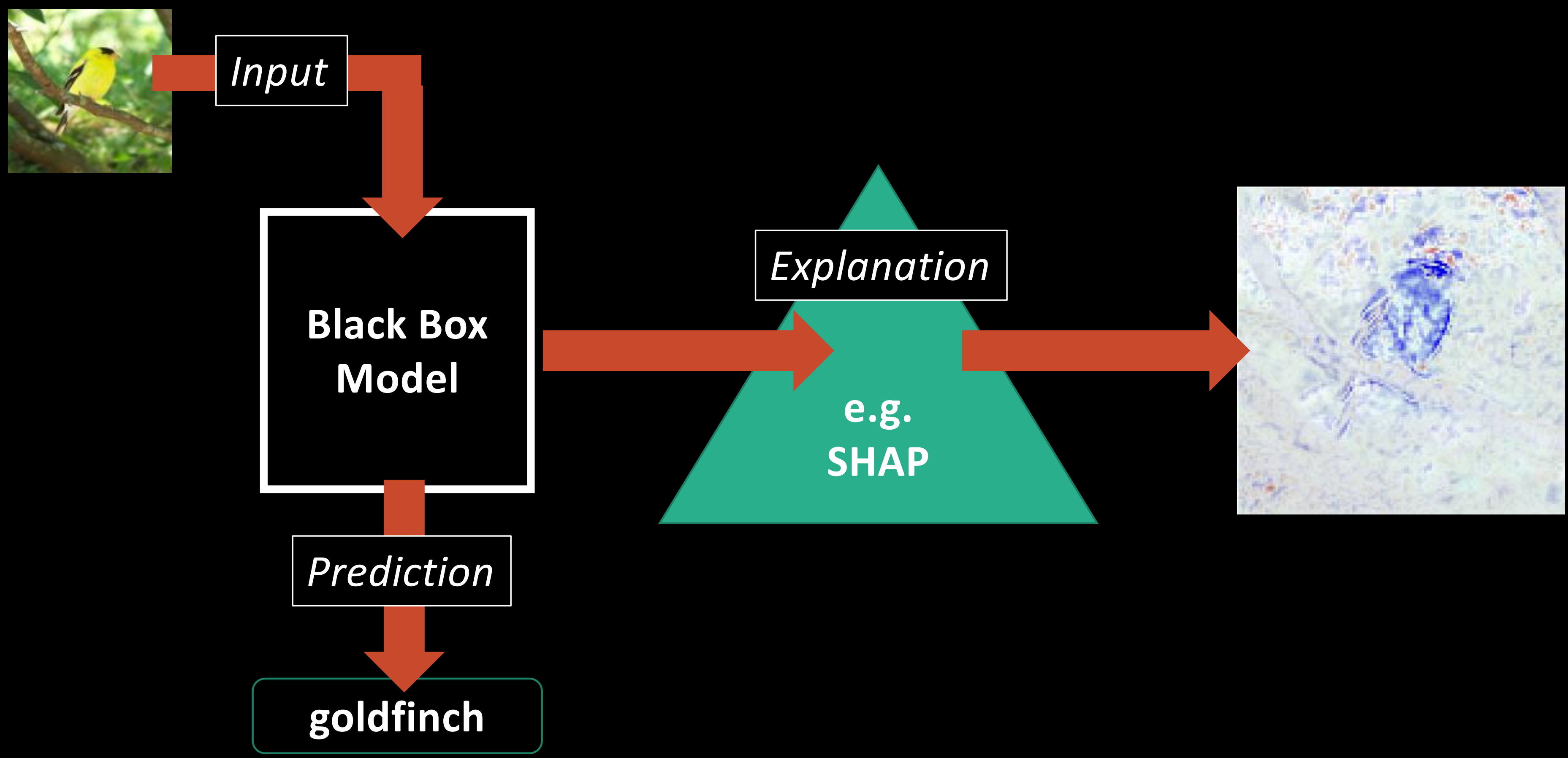
These examples and more are motivating our Conversational Explanation research – a simple unified interface to support any kind of explanation…

Deep Learning

Black Box Explanations

Deep Learning - Explainability

Accuracy & Comprehensiveness



Recap: Explanation Types and Techniques

Explanation Types:

- *Local* vs *Global* Explanations - The Mythos of Model Interpretability – Lipton 2016
- *Transparency* vs *Post-Hoc* - The Mythos of Model Interpretability – Lipton 2016
  (Molnar uses “intrinsic” instead of “transparent”)

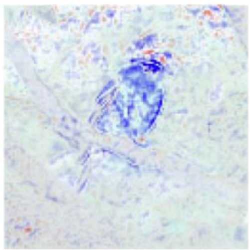
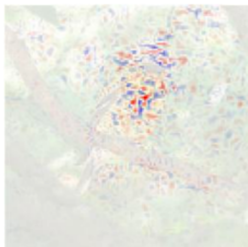
Categories:

(with reference & expansion : Personalized explanation in machine learning – Schneider et al. 2019)

- Feature Importance (Attribution)
- Counterfactual
- Component Data
- Model Internals
- Feature Visualisation
- Explanation by Example

Explanation Types and Techniques

Feature Importance (Attribution)

 **8) 05765_goldfinch.JPEG goldfinch**	**LIME**	**Shap**	**LRP**
vgg16_imagenet	 goldfinch Evidence towards predicted class shown in green	 goldfinch Evidence towards predicted class shown in blue, evidence against shown in red.	 goldfinch Evidence towards predicted class shown in blue, evidence against shown in red.

LIME:
"Why Should I Trust You?": Explaining the Predictions of Any Classifier – Ribeiro et al. 2016

Shap:
A Unified Approach to Interpreting Model Predictions - Lundberg et al. 2017

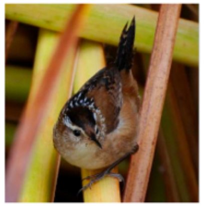
LRP:
On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation – Bach et al. 2015

(Explanation Table Generated Using DAIS Interpretability Framework)

Explanation Types and Techniques

Feature Importance

*This is a **Marsh Wren** because...*



Definition: this bird is brown and white in color with a skinny brown beak and brown eye rings.

Explanation: this is a small brown bird with a long tail and a **white eyebrow**.

*This is a **Downy Woodpecker** because...*

Definition: this bird has a white breast black wings and a red spot on its head.

Explanation: this is a black and white bird with a **red spot** on its crown.

*This is a **Shiny Cowbird** because...*

Definition: this bird is black with a long tail and has a very short beak.

Explanation: this is a black bird with a **long tail feather** and a pointy black beak.

*This is a **Marsh Wren** because...*

Definition: this bird is brown and white in color with a skinny brown beak and brown eye rings.

Explanation: this is a small bird with a long bill and brown and black wings.

*This is a **Downy Woodpecker** because...*

Definition: this bird has a white breast black wings and a red spot on its head.

Explanation: this is a white bird with a black wing and a black and white striped head.

*This is a **Shiny Cowbird** because...*

Definition: this bird is black with a long tail and has a very short beak.

Explanation: this is a black bird with a small black beak.

Generating Visual Explanations - Hendricks et al. 2016

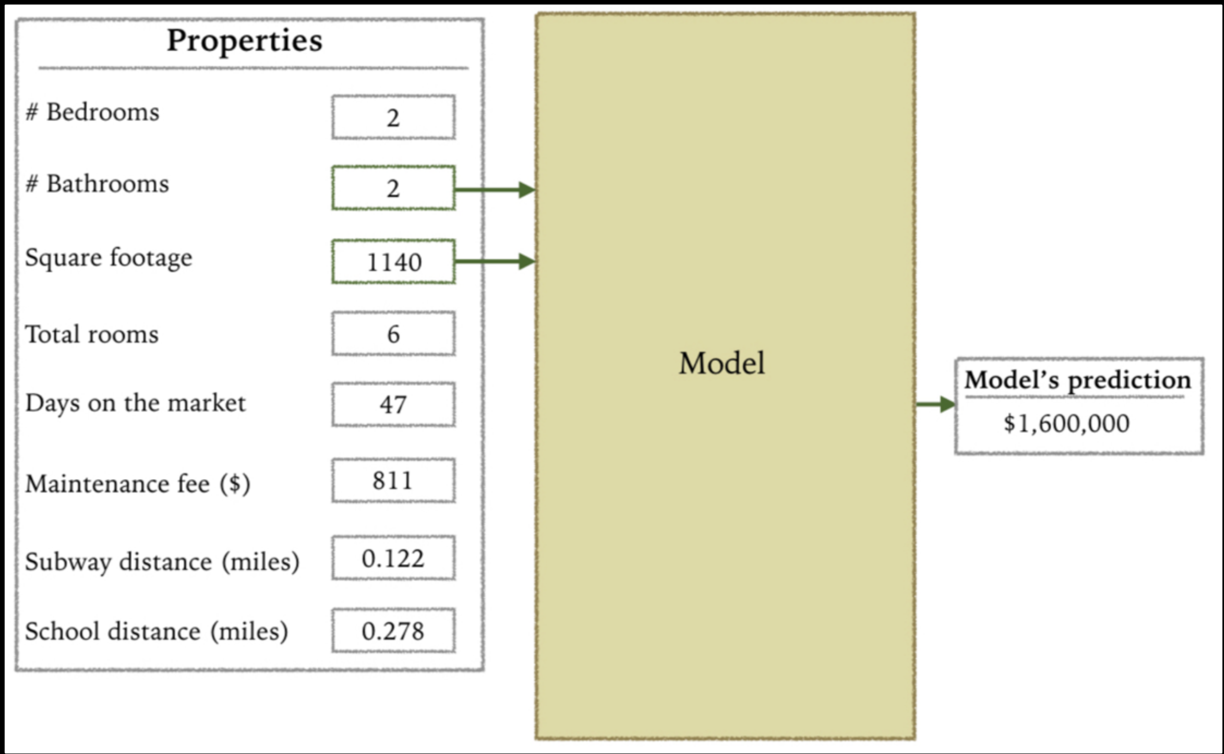
Explanation Types and Techniques

Counterfactual

Generating Counterfactual Explanations with Natural Language – Hendricks et al. 2018

Explanation Types and Techniques

Component Data



Output To the User
Model's Prediction: $1,600,000

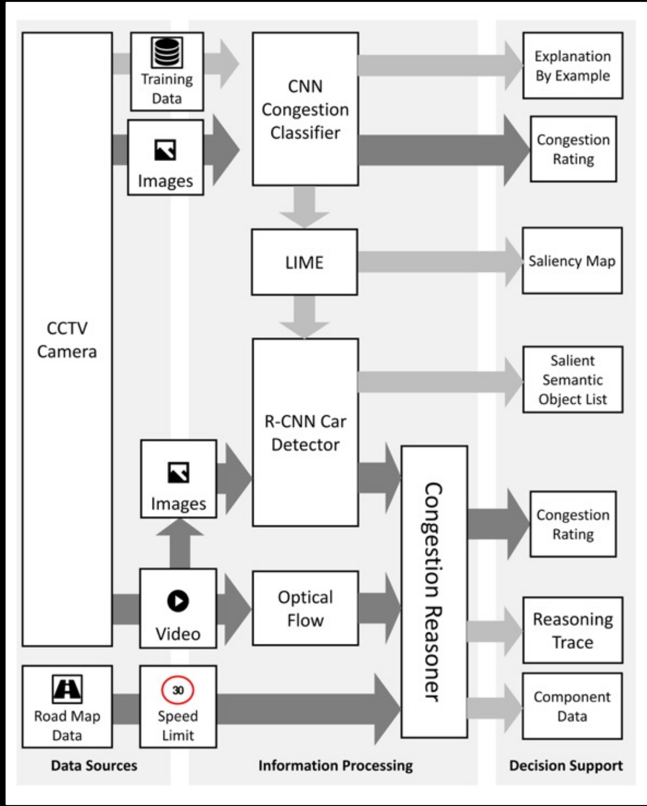
Data:
- Bathrooms: 2
- Square Footage: 1140

Manipulating and Measuring Model Interpretability - Poursabzi-Sangdeh 2018

Explanation Types and Techniques

Component Data

Detecting Traffic Congestion Using a Distributed System



System Output

Prediction:
Road is Congested

Component Data:
CNN CLASSIFIER
- *CNN Prediction*: 0.79 Congested

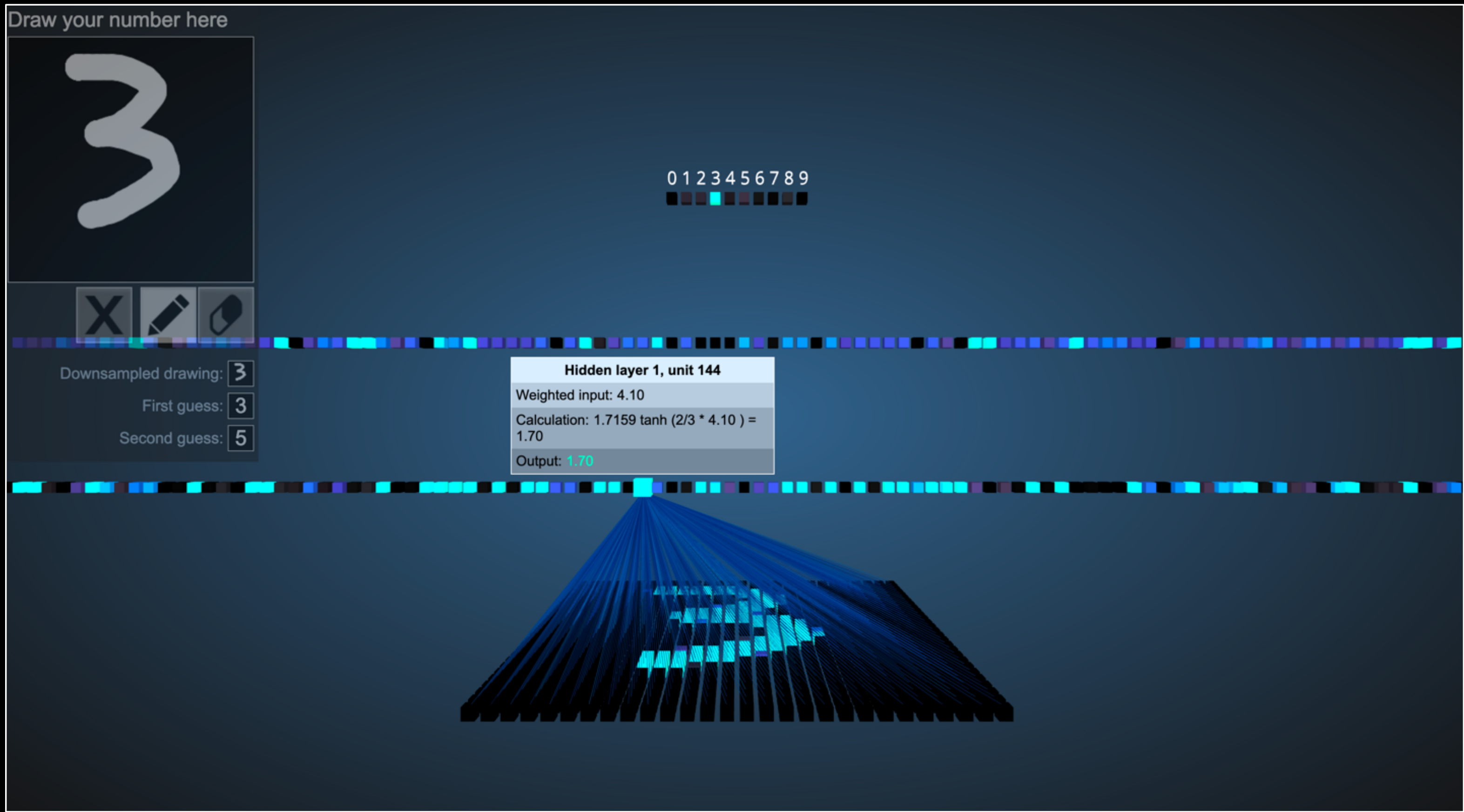
Congestion Reasoner
- *Congestion Rating*: 0.67
---- *Optical Flow*: 2.3
---- *Speed Limit*: 30 MPH

...

Integrating Learning and Reasoning Services for Explainable Information Fusion – Harborne et al. 2017

Explanation Types and Techniques

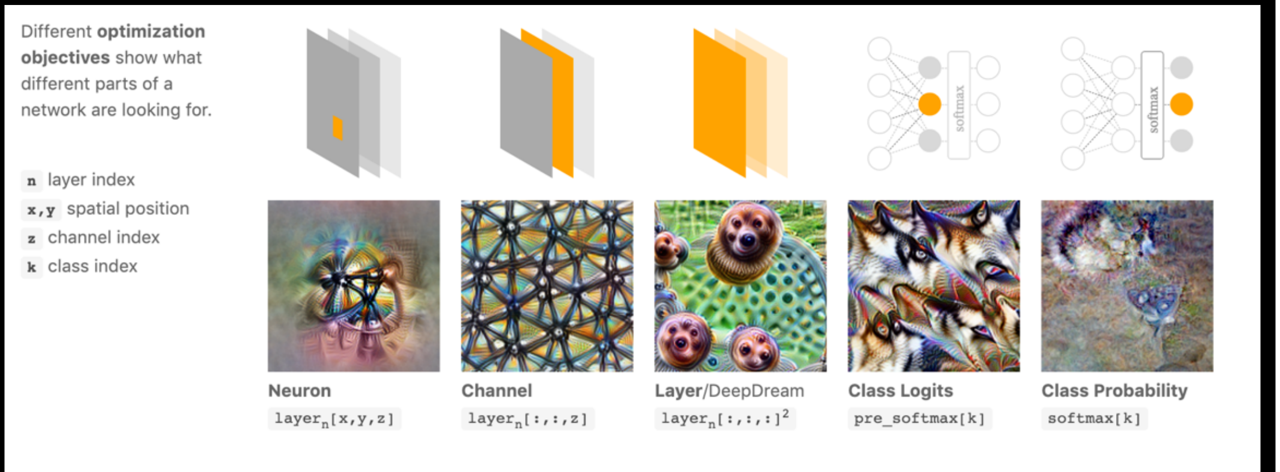
Model Internals

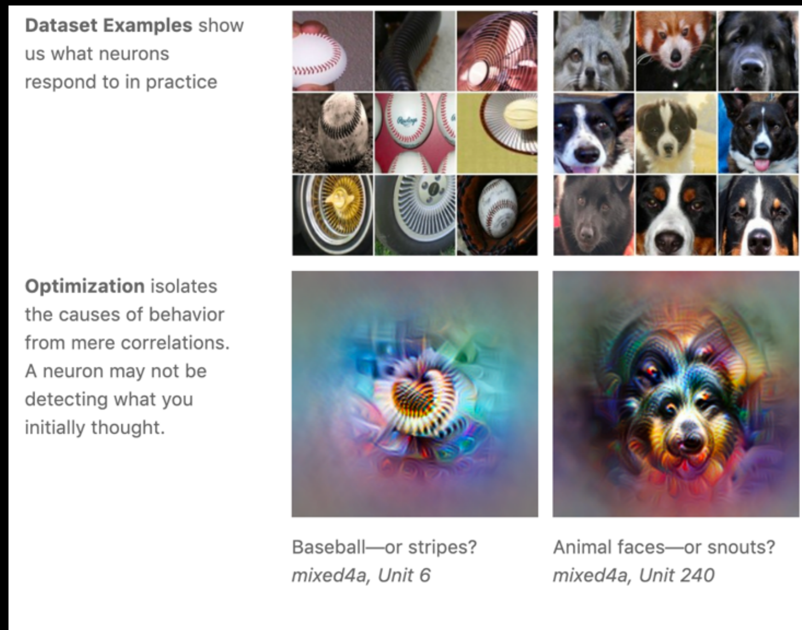


3D visualization of a Convolution Neural Network - http://scs.ryerson.ca/~aharley/vis/fc/

Explanation Types and Techniques

Feature Visualization

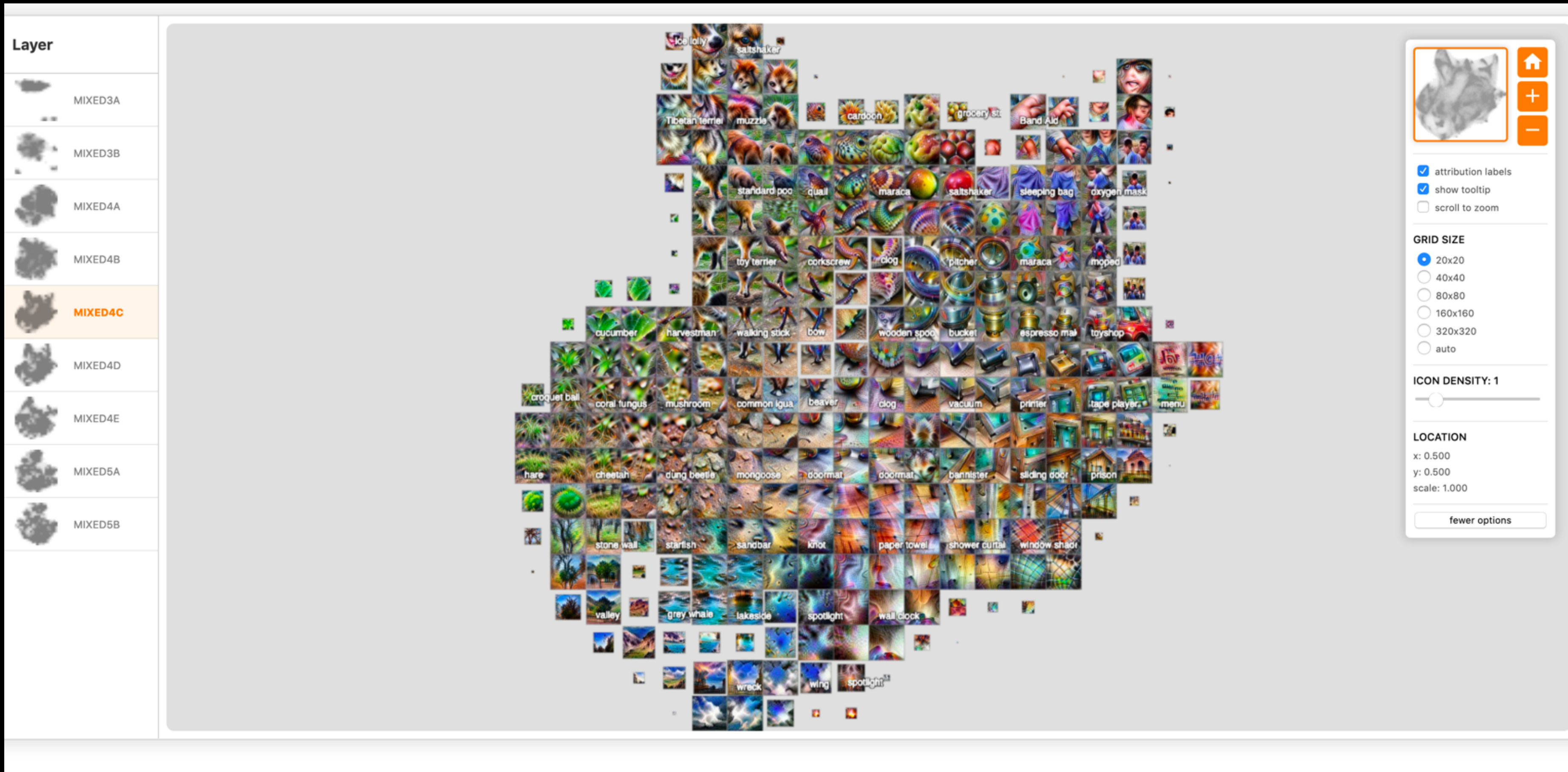




Feature Visualization - Olah, et al. 2017

Explanation Types and Techniques

Feature Visualization



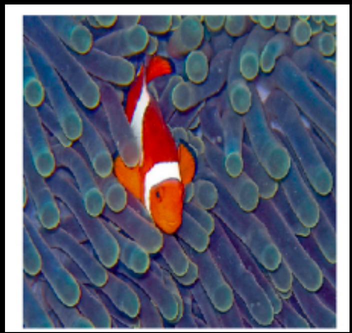
Exploring Neural Networks with Activation Atlases - Carter, et al. 2019 (March 6, 2019)

Explanation Types and Techniques

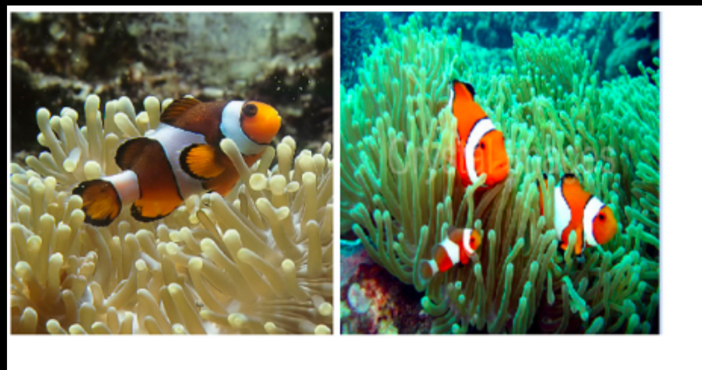
Explanation by Example

Understanding Dog Vs Fish Classification Using Influence Functions

Test Image



Helpful ("influential") Images from Training Data



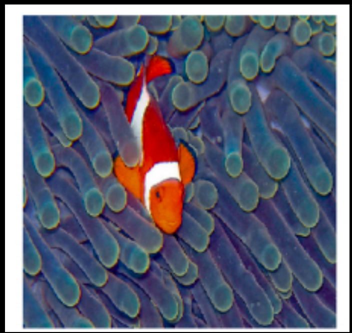
Understanding Black-box Predictions via Influence Functions - Koh et al. 2017

Explanation Types and Techniques

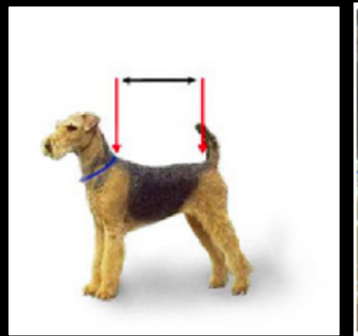
Counterfactual Explanation by Examples

Understanding Dog Vs Fish Classification Using Influence Functions

Test Image



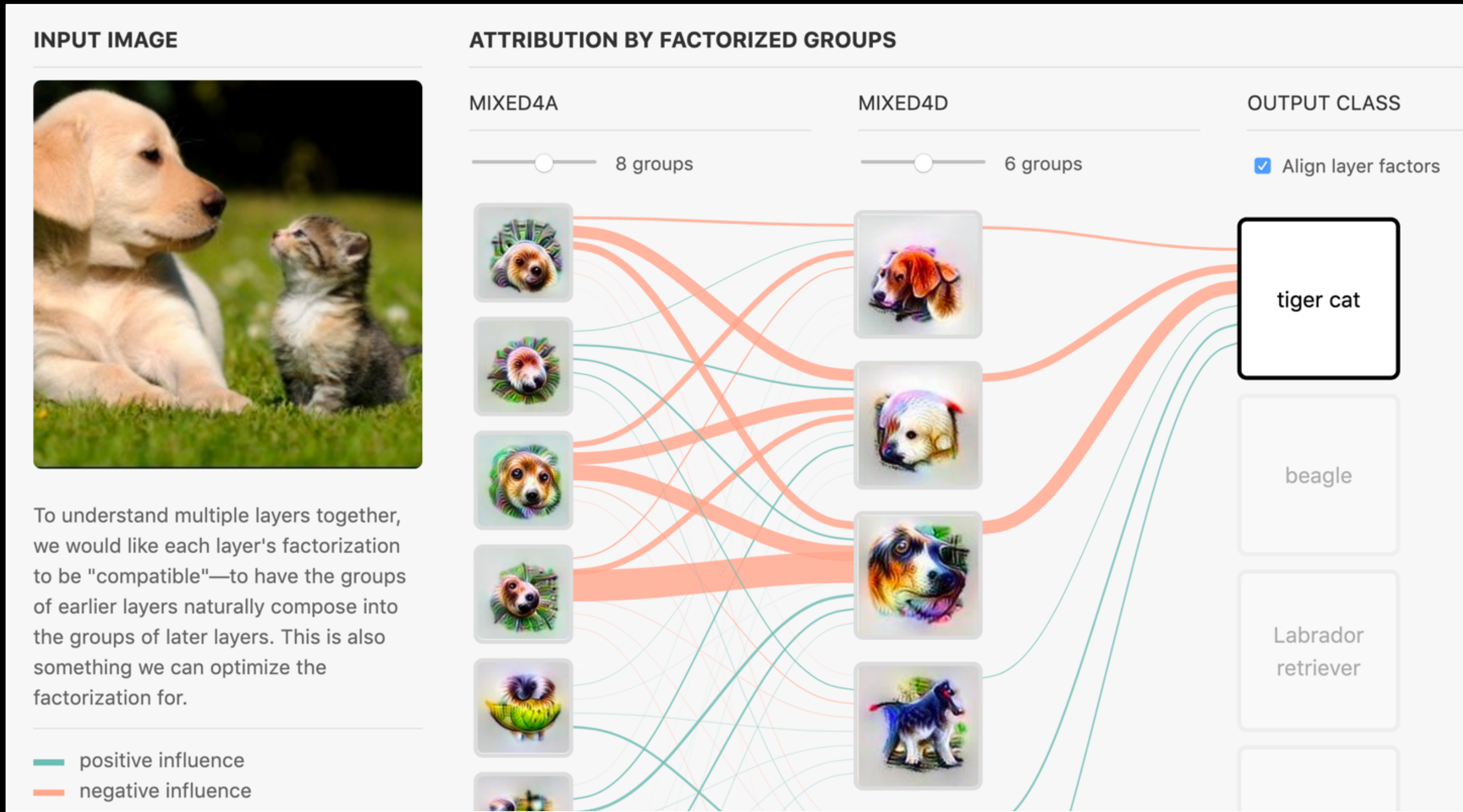
Helpful ("influential") Images from Training Data



Understanding Black-box Predictions via Influence Functions - Koh et al. 2017

Explanation Types and Techniques

Combinations



The Building Blocks of Interpretability - Olah, et al. 2018

Explanation Properties

- Complexity
- Prioritization of decision information
- Visualisation of Data
- Interactivity

What makes a good explanation technique?

Desirables of Explanations

Effectiveness:

- Explainability (Accuracy & Comprehensiveness)
- Interpretability

Versatility:

- Generalizability (how many models does it work for?)
- Explanatory Power (How many questions can it answer?)

Constraints:

- Privacy
- Resources
- Timely
- Information Collection Effort [for personalisation]

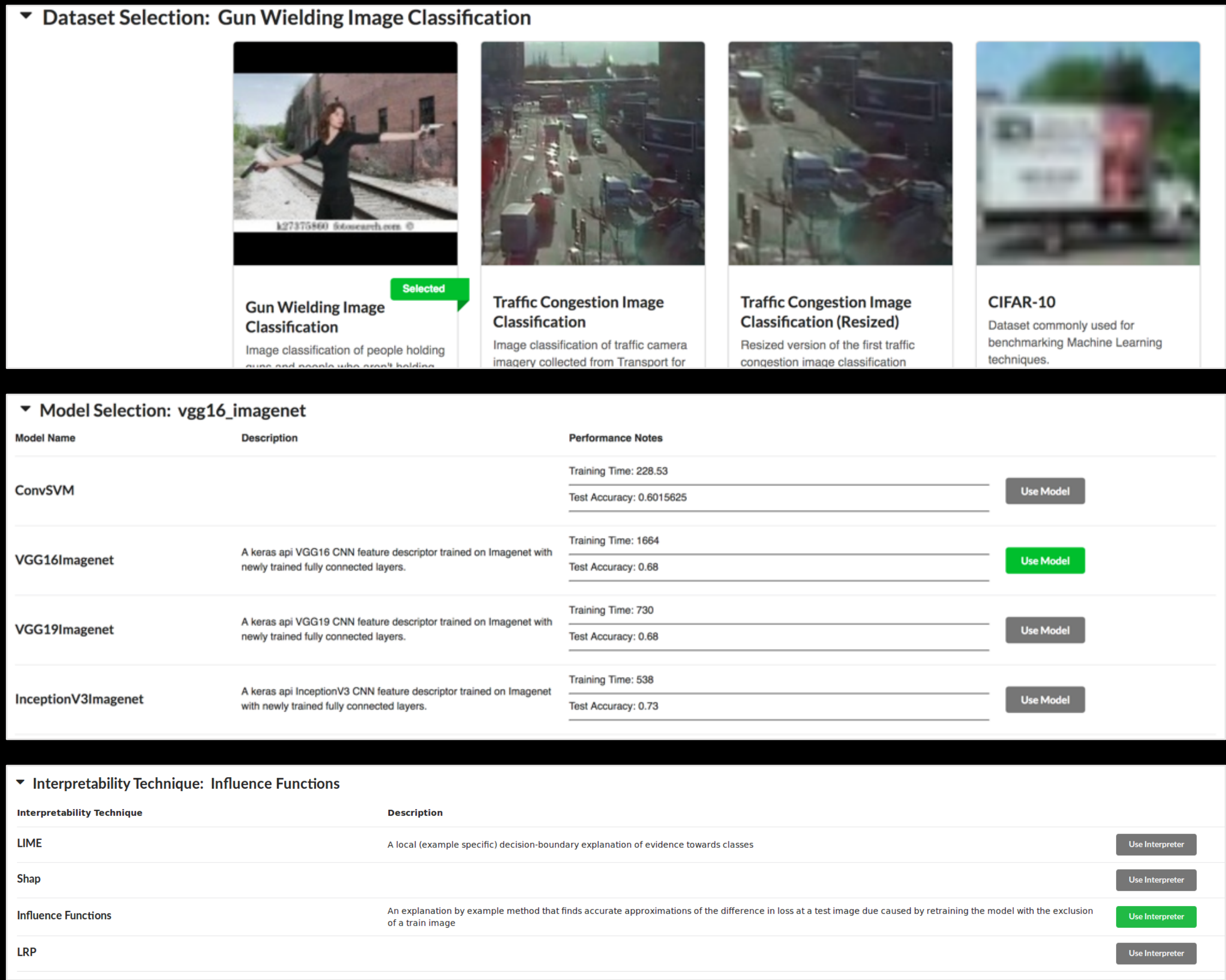
with reference & expansion : Personalized explanation in machine learning – Schneider et al. 2019

Interpretability

Aspects of a User

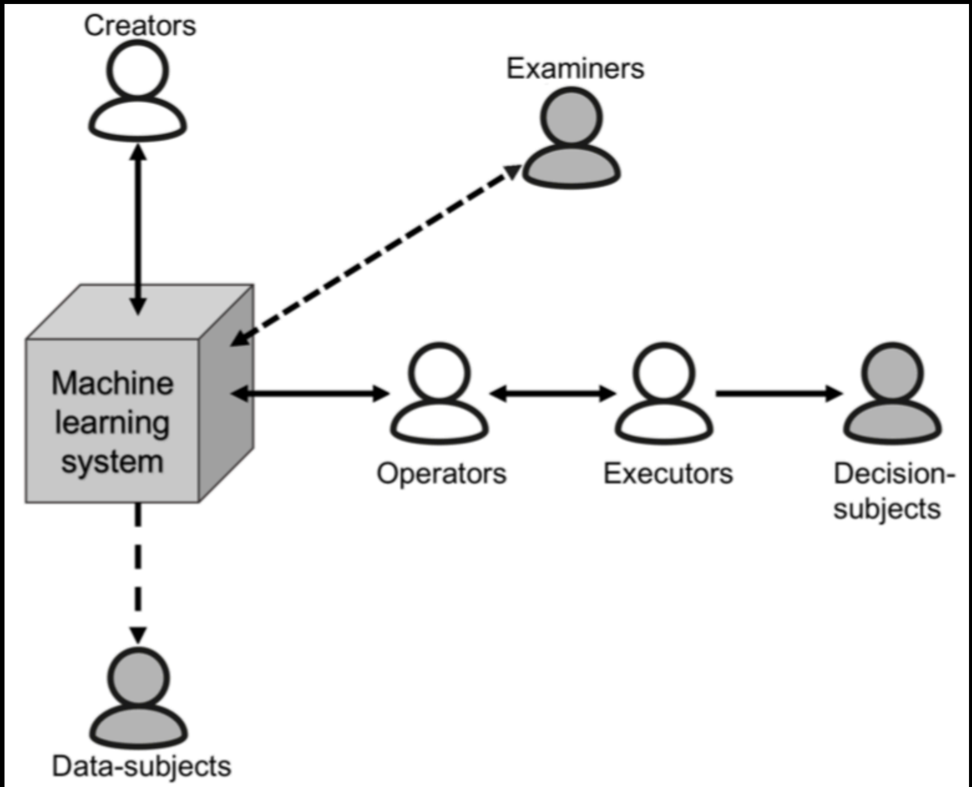
- Prior Knowledge
  - Machine Learning Knowledge
  - Task Domain Knowledge
- Decision Information
- Preference
- Purpose

Experimentation Framework – Our Interface



The role of the user

"Interpretable to Whom?" framework

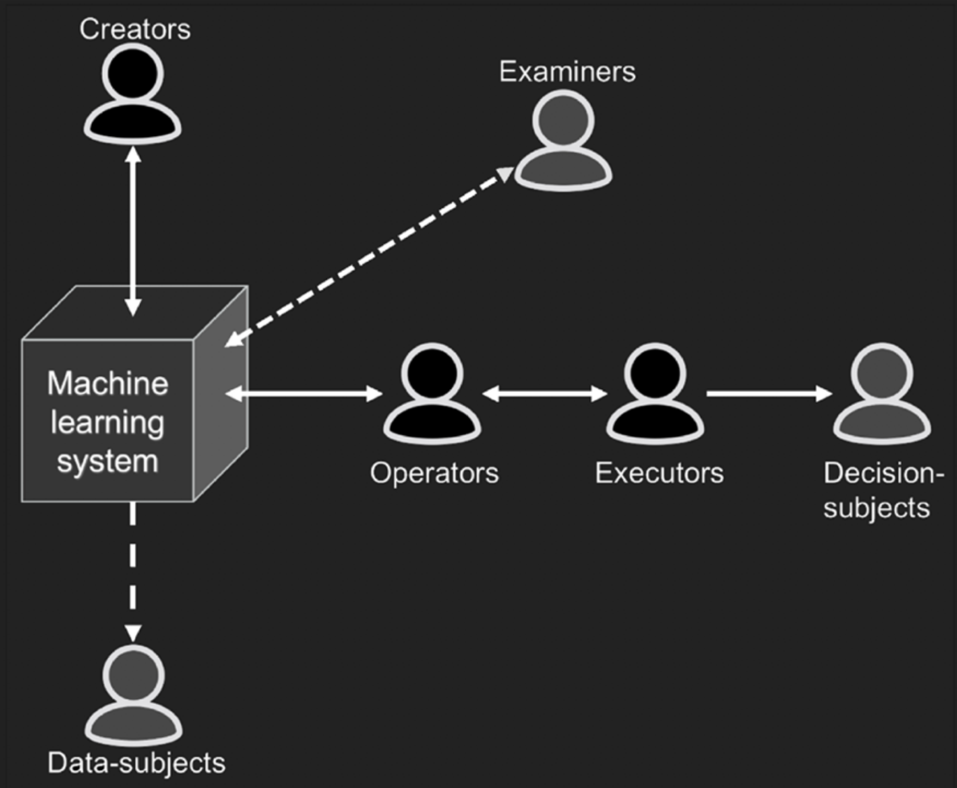


WHI workshop at ICML 2018

https://arxiv.org/abs/1806.07552

Argues that a machine learning system's interpretability should be defined in relation to a specific agent or task: we should not ask if the system is interpretable, but to whom is it interpretable.

Applied to six real-world example scenarios



- Web Advertising
- Route planning on a smartphone
- Loan application
- Medical advice for clinicians
- Releasing defendants on bail
- No-go order in a military operation

...with the various roles defined in detail for each

Impact of this work

- A useful framework for assessing AI/ML system development plans and architectures
- Interest from the UK Financial Conduct Authority (FCA)
  - Invited guest lecture
  - Panel session on Ethics in AI
  - Interest in DAIS ITA research more widely
- Future plans
  - To integrate the role-based model deeper into our meta-model to support conversational explanations
  - To cross-reference against more recent work (Miller, Molnar) to standardize terminology

Conversational Explanations

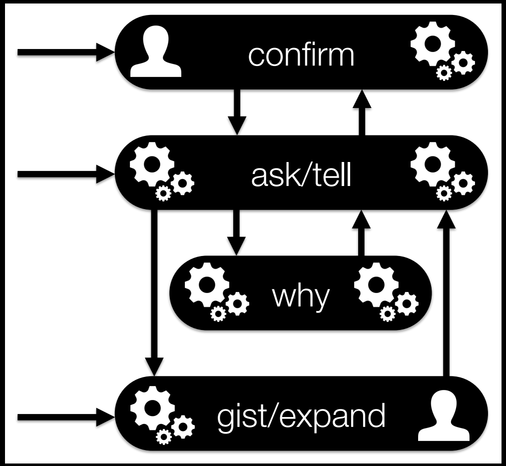
Earlier Research: Conversational Interaction

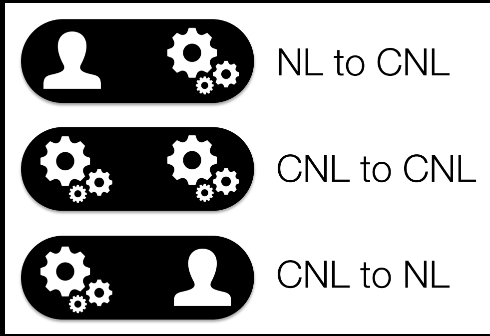
- Talking to machines in natural language is ideal but hard
- Controlled Natural Language as a compromise: “easy to read, harder to write”
- Let’s bring the two together:
  - Human users write NL sentences [easy to write]
  - Machine users convert to NL [easy to process]
  - Machine users respond in CNL by default [easy to read]

there is a person named p1
that is known as ‘John Smith’
and is a high value client.

Our conversational model

- We built a model of conversations in CNL
  - to enable interactions that flow freely between NL and CNL

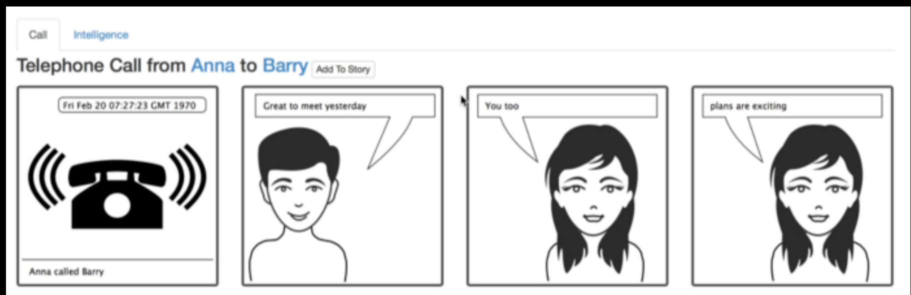


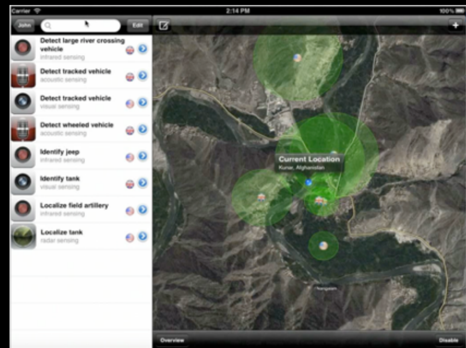


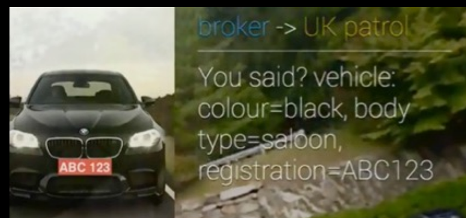
Draws on research in agent communication languages and philosophical linguistics (speech acts)

We carried out evaluations

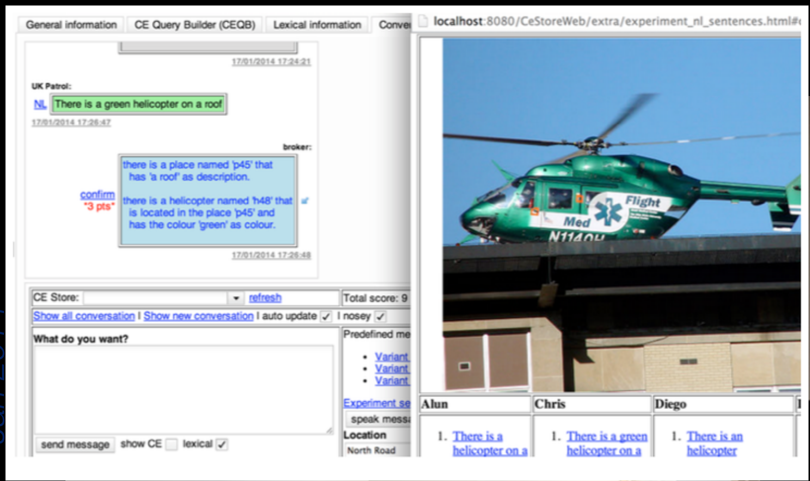
- Field trials
- Asset allocation
- Intelligence analysis
- Coalition planning
- Crowd-sourced intelligence
- Publication analytics



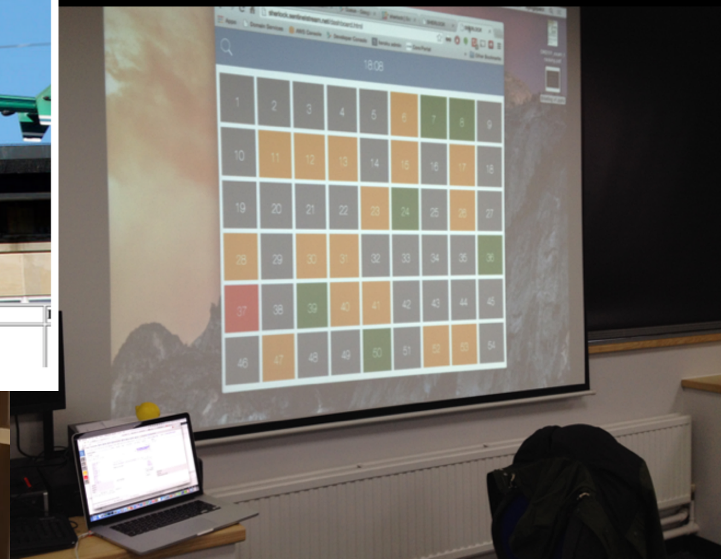




We analyzed student experiments



Jan 2014



Dec 2014

Oct 2015

…and worked with practitioners

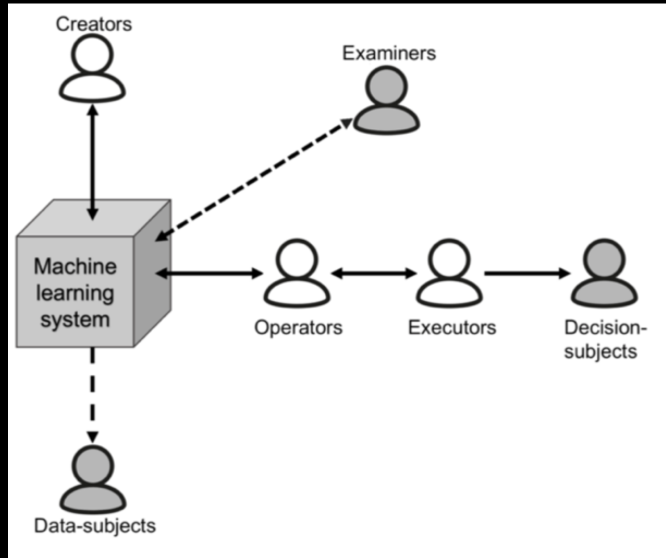
Oct 2016

May 2016

Applying conversation to explanations

- We gained key insights from this previous research
  - Conversations are social and experiential
  - They can apply in a broad set of domains
  - A single interface methodology to traverse numerous systems
  - The ability to converse across domain or system boundaries
  - Multi-modal conversations are possible
- This leads to our use of conversations for our Explainable AI research
- We hope to build a robust framework and meta-model
  - …and carry out a series of tests with human users

Conversational Explanations

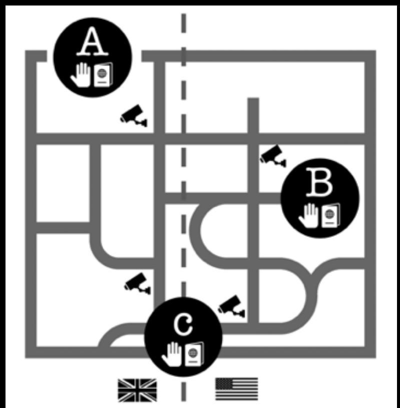


Scenario and dataset

- Real-time London CCTV imagery
- Coalition context & edge processing
- Many derivative datasets possible

Explanation-oriented architecture (XOA)

- Rapid ensemble services
- Trust and confidence



Explanation types

- Transparent, post-hoc
- Multiple modalities

Conversation and roles

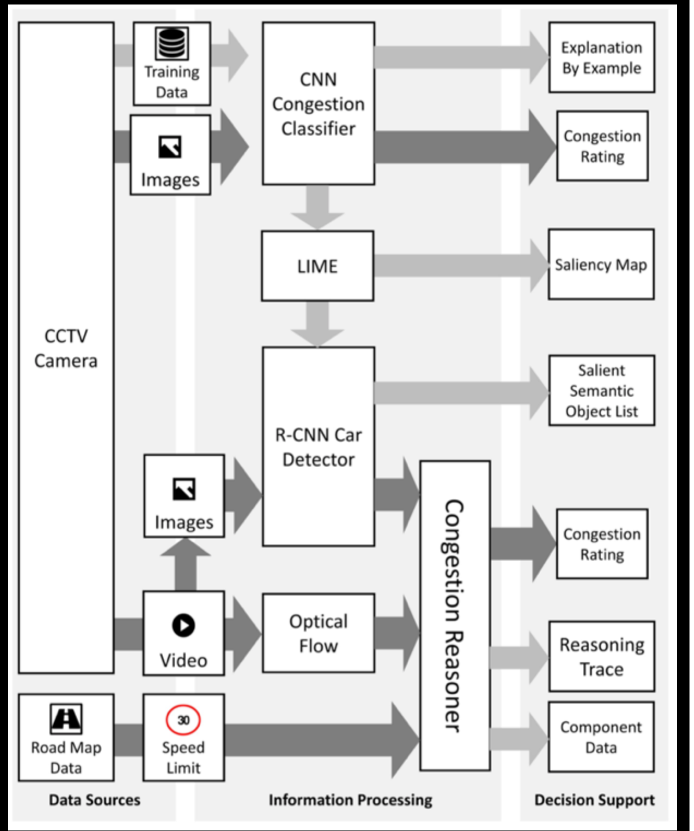
- We treat explanation as a conversation
- User role and task context are key

Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:



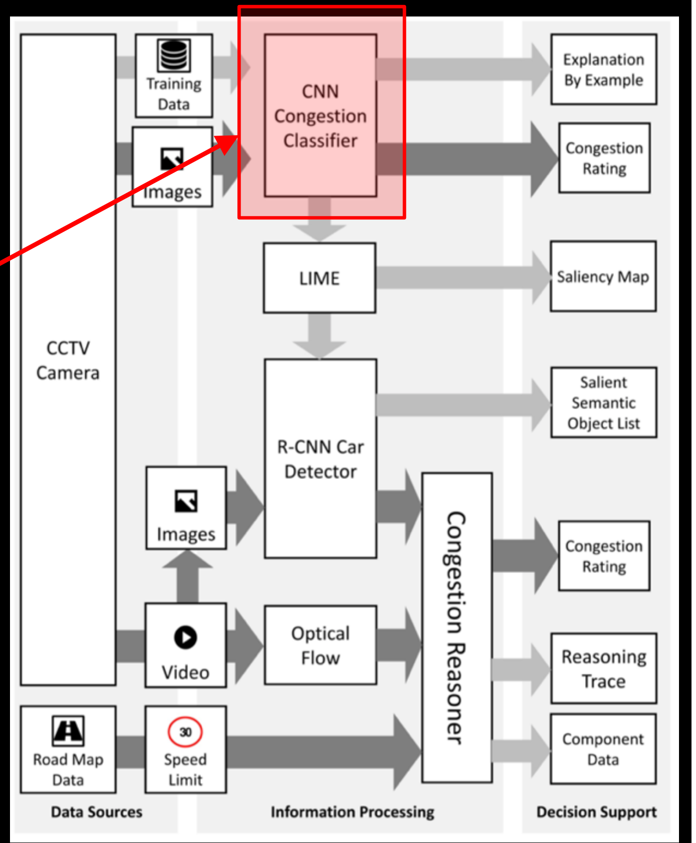
Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:

1. Congestion Image Classifier (CIC)



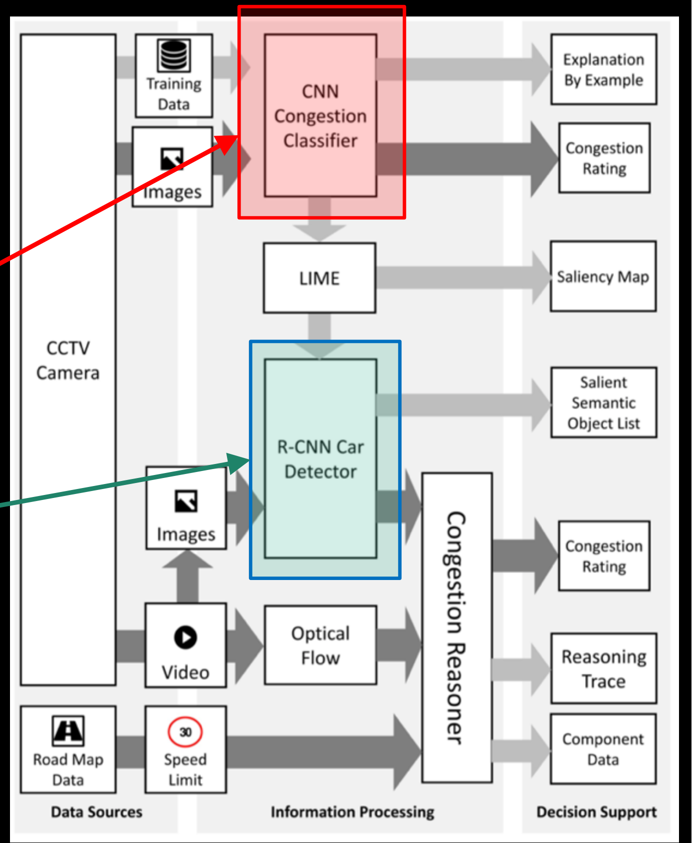
Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:

1. Congestion Image Classifier (CIC)
2. Entity detector (ED)



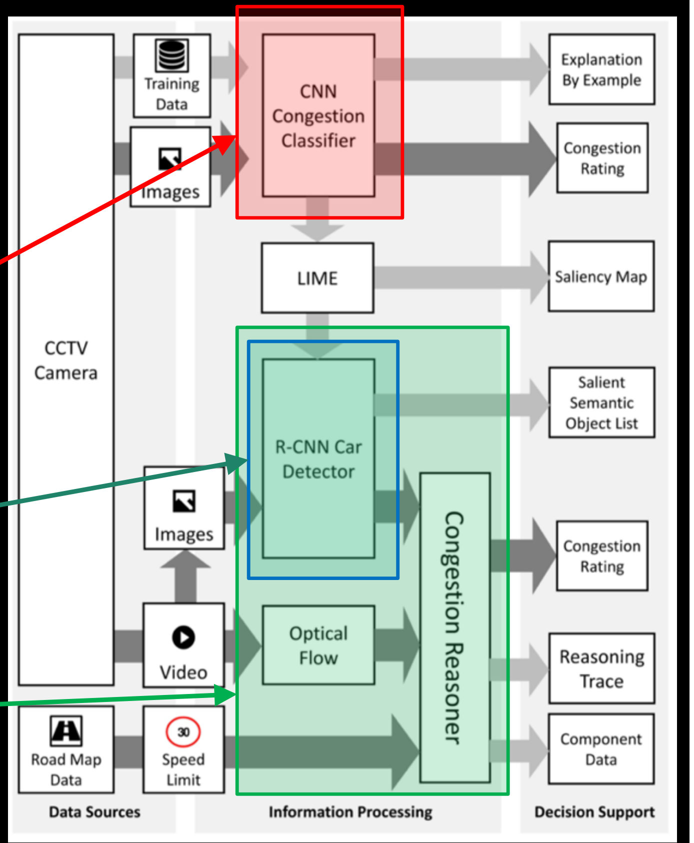
Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:

1. Congestion Image Classifier (CIC)
2. Entity detector (ED)
3. Congestion Speed Classifier (CSC)



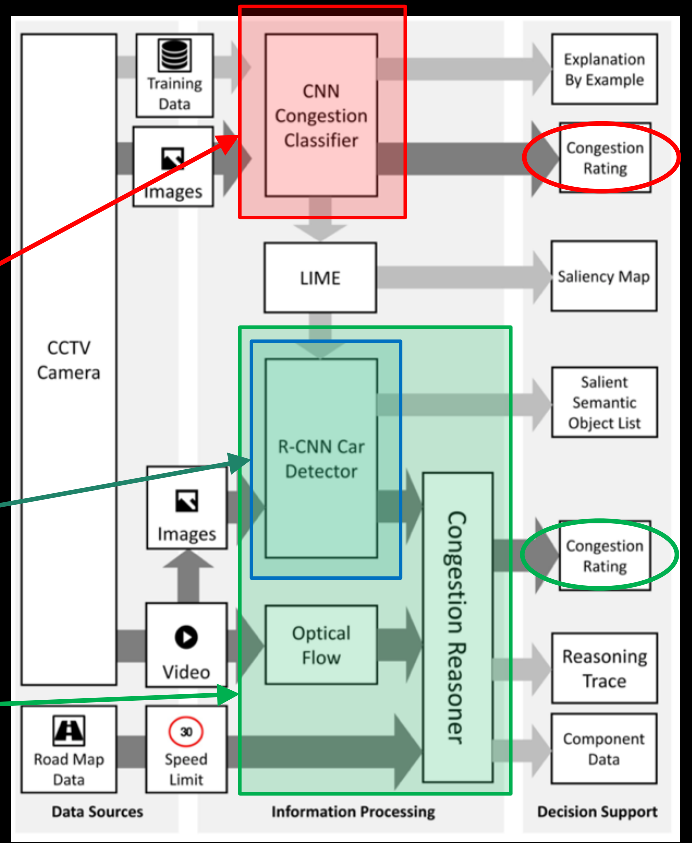
Worked Example

Using our Explanation Oriented Architecture

- Detect or infer traffic congestion
- Congestion & explanation services and flows
- Information fusion from multi-modal data sources

Three types of congestion services:

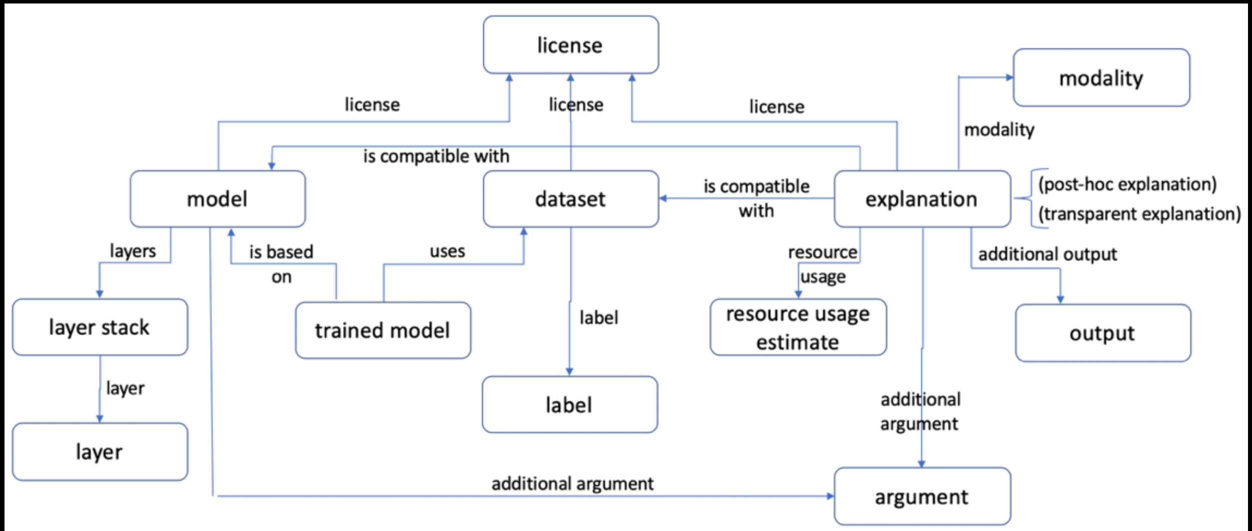
1. Congestion Image Classifier (CIC)
2. Entity detector (ED)
3. Congestion Speed Classifier (CSC)



Conversations for Explanation

Explanation takes the form of a conversation

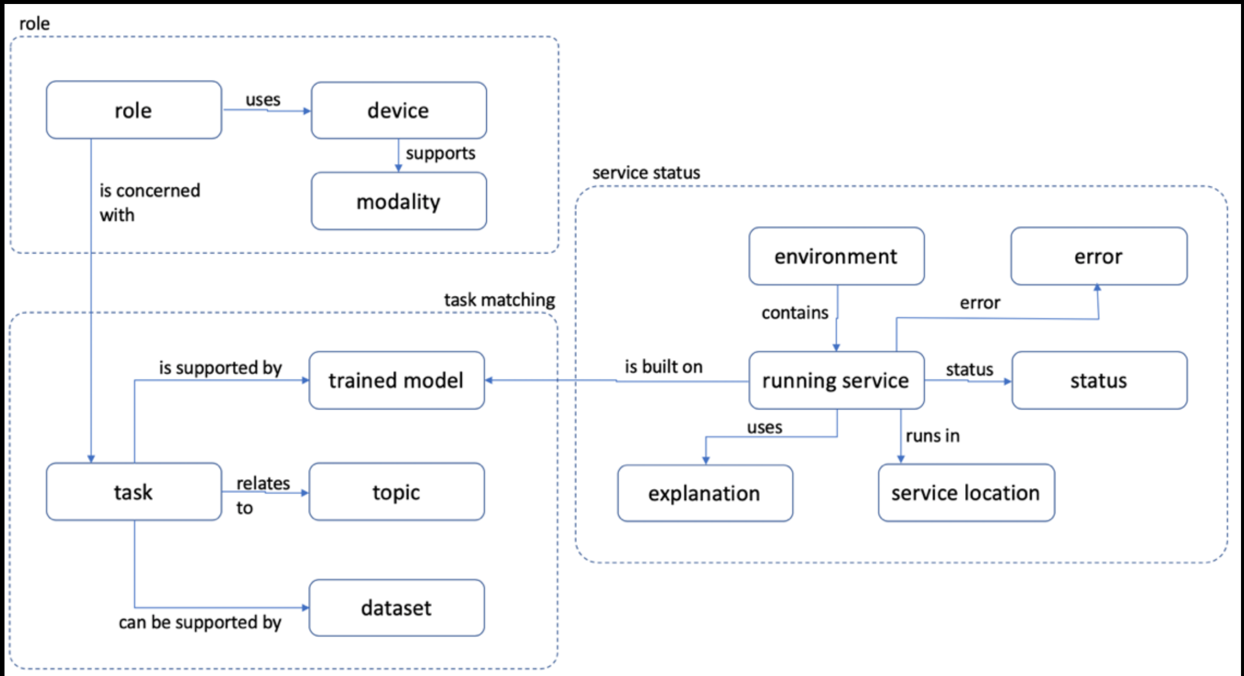
- Can be in many forms
- Focus on the abstract space
- Text-chat format
- Embedded additional modalities



Conceptual model defines key concepts and relationships

- Ontology, but for machine and human processing
- Key explanation-related concepts in “core” model

Conversations for Explanation



Additional models easily integrated

- To define the context for the explanations
- Role, task, services, etc

*The **ad-hoc** and **real-time** operational environment require **agile context modelling**.*

Conversations for Explanation

Use case 1: Fully transparent

Conversations for Explanation
Use case 1: Fully transparent

Is there congestion at Checkpoint A?

Conversations for Explanation
Use case 1: Fully transparent

Is there congestion at Checkpoint A?

No, I am confident that checkpoint A is not congested.

Conversations for Explanation

Use case 1: Fully transparent

Is there congestion at Checkpoint A?

No, I am confident that checkpoint A is not congested.

Why?

Conversations for Explanation

Use case 1: Fully transparent

Is there congestion at Checkpoint A?

No, I am confident that checkpoint A is not congested.

Why?

Live CCTV shows 12 moving objects at 80% of the speed limit which indicates it is not congested.

Conversations for Explanation

Use case 2a: Post-hoc by Saliency

Conversations for Explanation

Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?

Conversations for Explanation

Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?

I am not confident, but Checkpoint B is congested.

Conversations for Explanation

Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?

I am not confident, but Checkpoint B is congested.

Why?

Conversations for Explanation

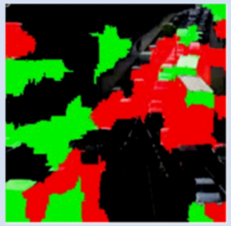
Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?

I am not confident, but Checkpoint B is congested.

Why?

This live CCTV image shows the situation and I believe Checkpoint B is congested (with 62% confidence).



The green areas show the parts of the image that most indicate congestion.

Conversations for Explanation

Use case 2b: Post-hoc by Example

Conversations for Explanation

Use case 2b: Post-hoc by Example

Is there congestion at Checkpoint C?

Conversations for Explanation

Use case 2b: Post-hoc by Example

Is there congestion at Checkpoint C?

I am not confident, but Checkpoint C is congested.

Conversations for Explanation
Use case 2b: Post-hoc by Example

Is there congestion at Checkpoint C?

I am not confident, but Checkpoint C is congested.

Why?

Conversations for Explanation

Use case 2b: Post-hoc by Example

Is there congestion at Checkpoint C?

I am not confident, but Checkpoint C is congested.

Why?

I can't show you the live CCTV image but these other images are rated as being similarly congested.

Conversations for Explanation

Use case 3: Disagreement within services

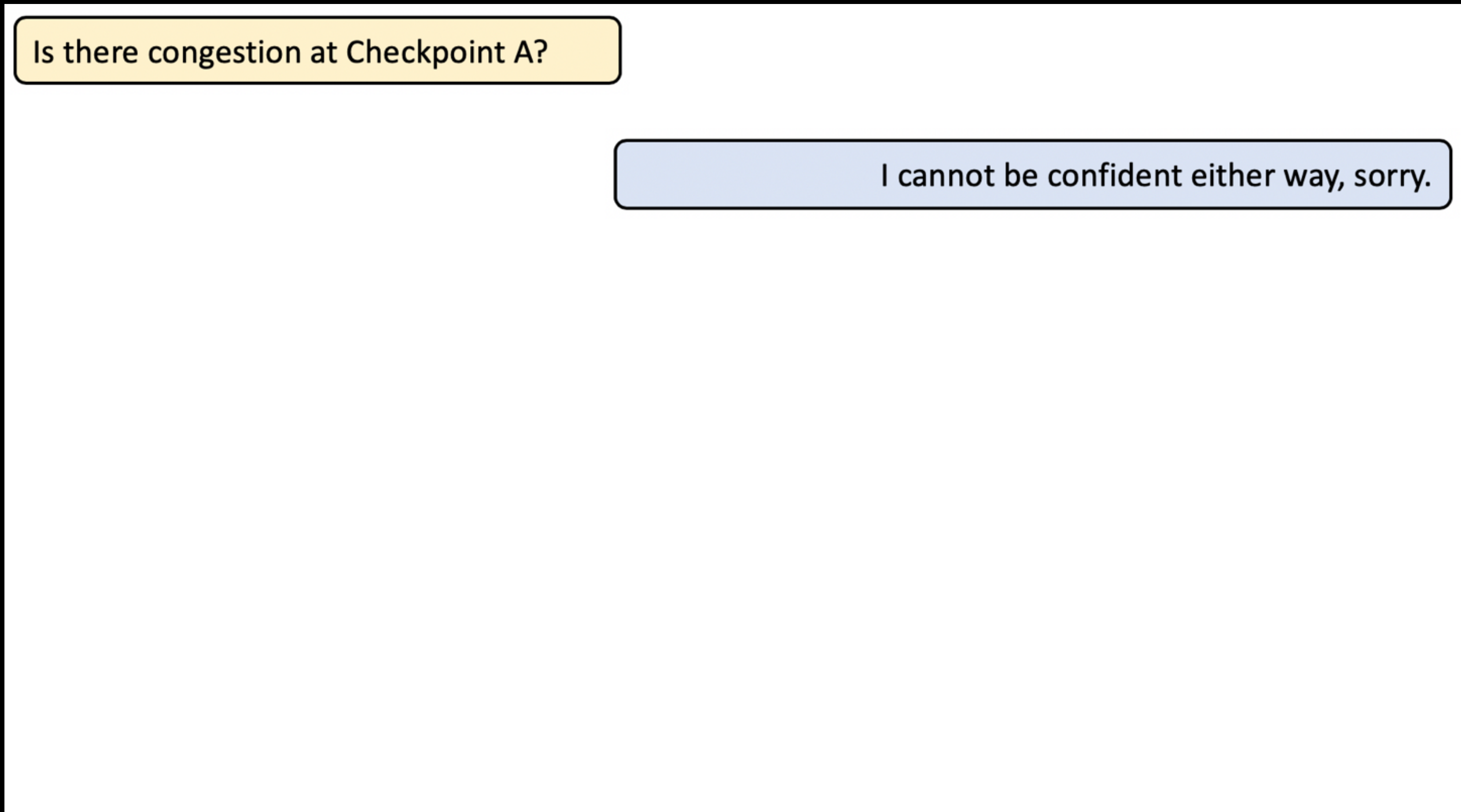
Conversations for Explanation

Use case 3: Disagreement within services

Is there congestion at Checkpoint A?

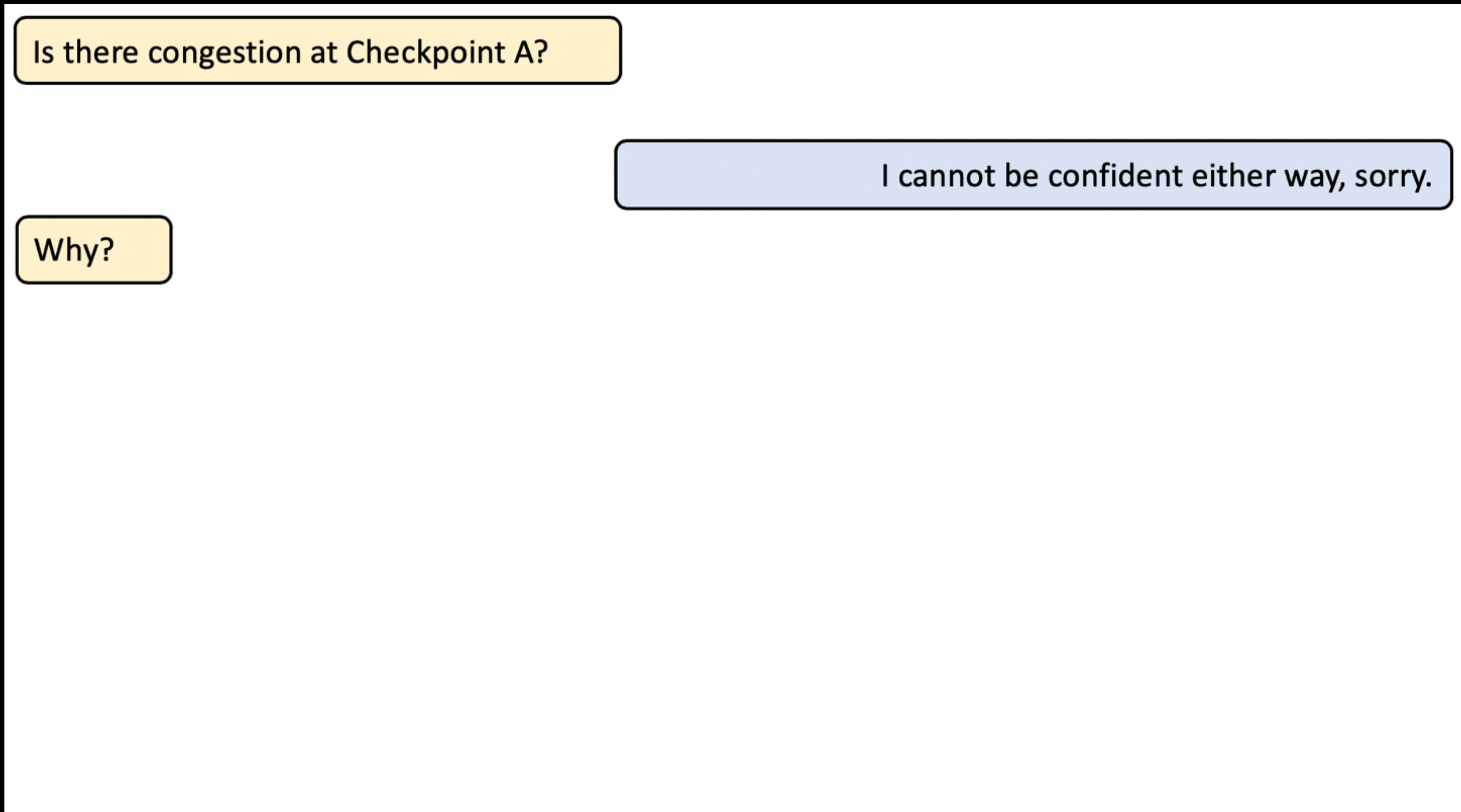
Conversations for Explanation

Use case 3: Disagreement within services



Conversations for Explanation

Use case 3: Disagreement within services



Conversations for Explanation

Use case 3: Disagreement within services

Is there congestion at Checkpoint A?

I cannot be confident either way, sorry.

Why?

Live CCTV shows 2 moving objects at 80% of the speed limit which indicates it is not congested. But the live CCTV image is classified as congested.

These outcomes are inconsistent.

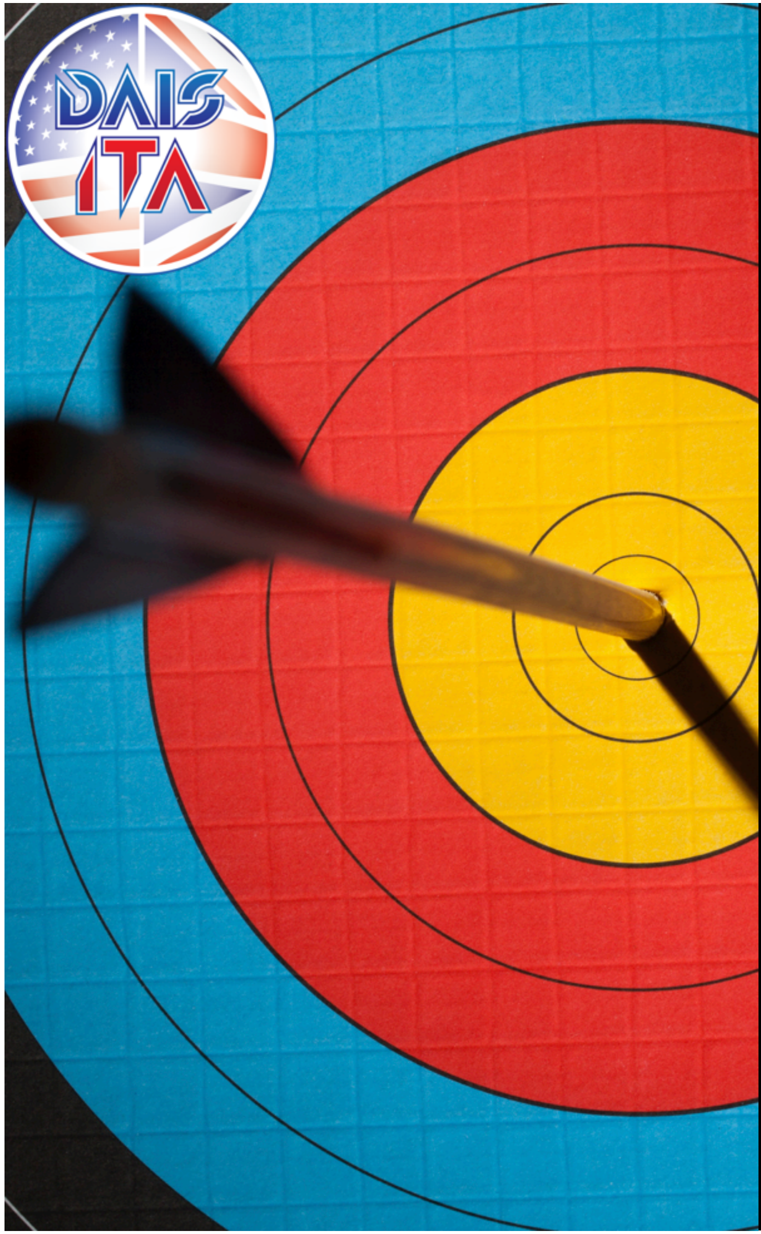
The green areas show the parts of the image that most indicate congestion.

Related work

- **Insight from Social Sciences**
  Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. arXiv preprint arXiv:1706.07269.
- **A grammar for the development of conversational explanations?**
  Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. Distill, 3(3), e10.
- **Affordances – the strengths of human and machine agents**
  Crouser, R. J., & Chang, R. (2012). An affordance-based framework for human computation and human-computer collaboration. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2859-2868.
- **Human-Computer Collaboration to drive our conversational principles**
  L. Terveen, "Overview of human-computer collaboration," Knowledge Based Systems, vol. 8(2), pp. 67–81, 1995.

Future plans

- Complete version 1 development of the conversational meta-model
- Build the experimental conversational explanation capability
  - Aligned against the conversational meta-model
- Choose a domain of interest for experimentation
- Design a user-focused experiment
  - Conversational Explanations
  - Measure some impact across multiple groups
    to test the effectiveness of conversational explanation



Thank you for listening!

Conversational Explanations

Explainable AI through human-machine conversation

Dave Braines
dave_braines@uk.ibm.com