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

Agenda

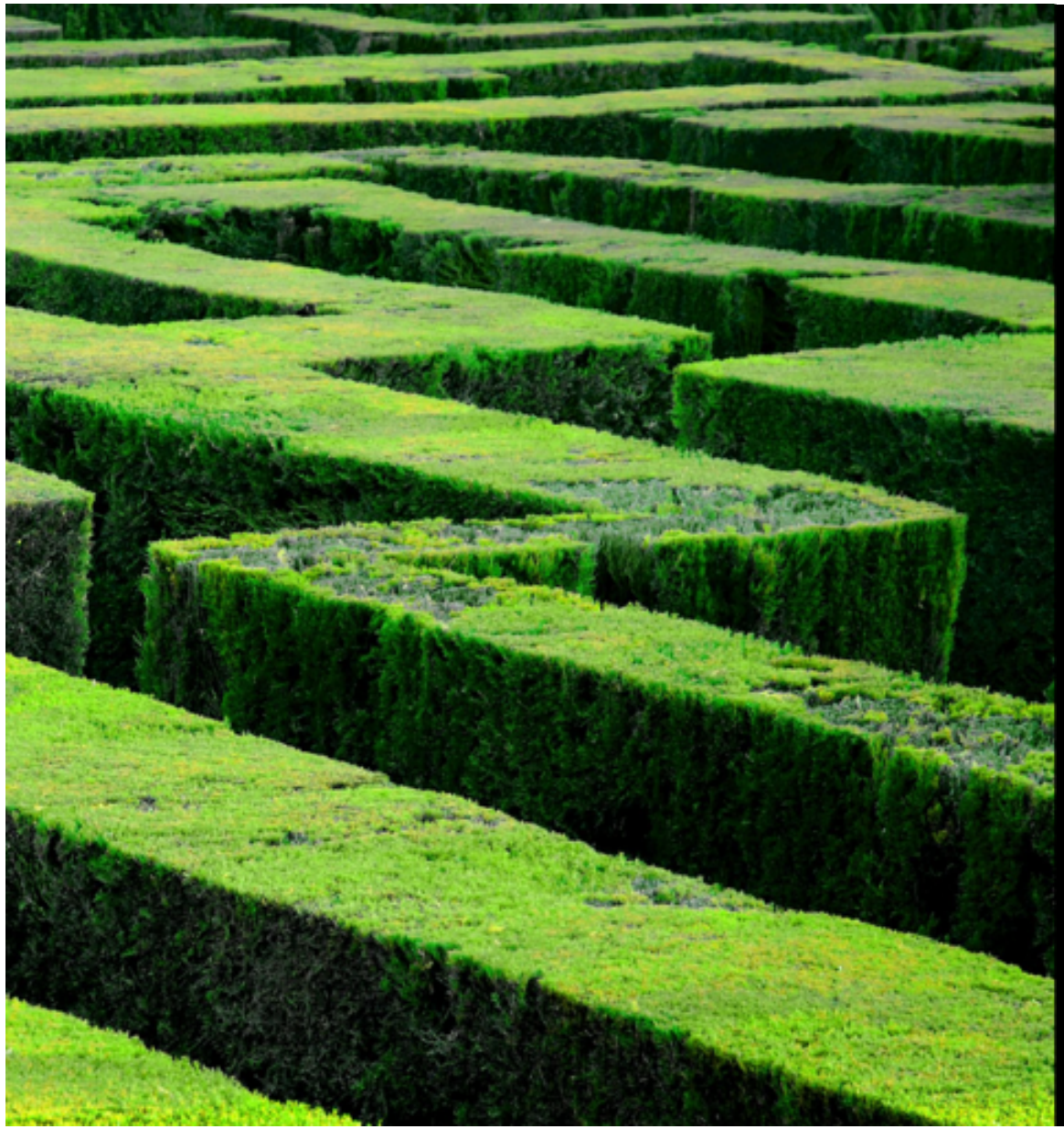
- Introductions *[10]*
- Explanations
  - Scene setting for Explainable AI (XAI) *[20]*
  - Philosophy & Social Science *[20]*
- Collaborative XAI research examples *[10]*

(Coffee break)

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

IEEE CogSIMA™

	Monday, April 8	
8:00 - 9:00 am	*Breakfast*	
9:00 - 9:10 am	T1: *Tutorial Session 1: Conversational Explanations - Explainable AI through Human-Machine Conversation*	
9:10 - 10:00 am		
10:00 - 10:30 am		*Coffee Break*
10:30 am - 12:00 pm		
12:00 - 1:30 pm	*Lunch (On your own)*	

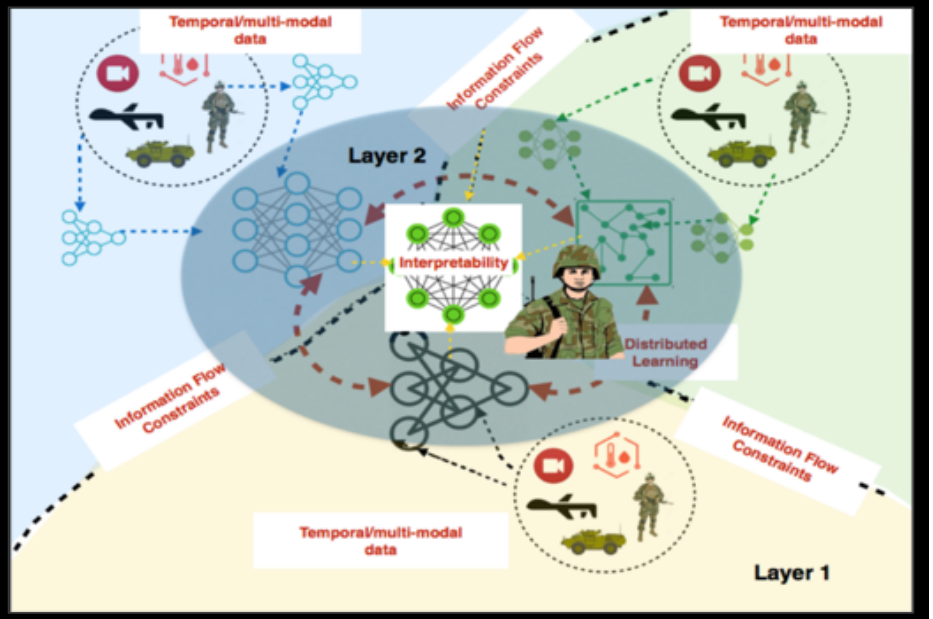


Introductions

- Relevance to CogSIMA
- Me
- IBM Research (Hursley)
- Cardiff University:
  Crime & Security
  Research Institute
- My PhD
- DAIS ITA

How do Conversational Explanations relate to CogSIMA?

- Our focus is coalition situational understanding:
  - Human-machine teams
  - Rapidly formed, ensembles of processes and services
  - No time to build customized experiences or user interfaces
  - Trust (especially in machine processes) is important to establish quickly



- In this tutorial I will cover:
  - Computer Science, Human factors and Cognitive Science
  - Situation sensing, Cognitive information fusion, models of human-machine collaboration with some aspects of ontology-based computing

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

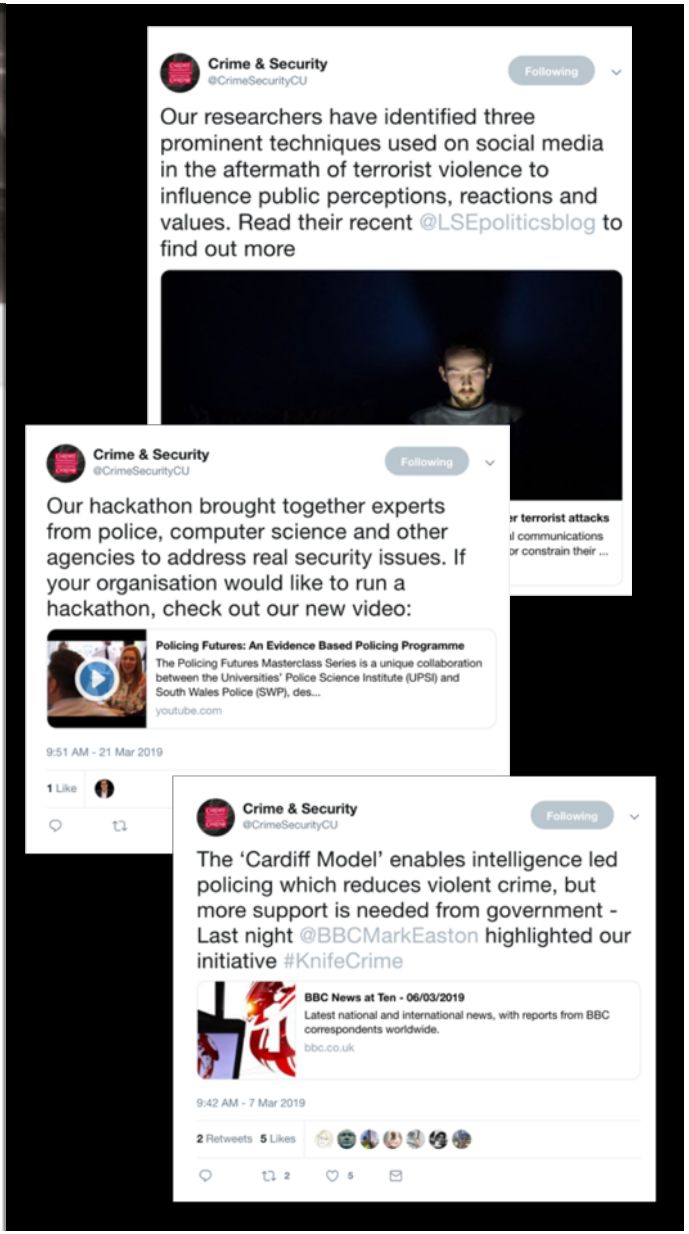
About us

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.



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.

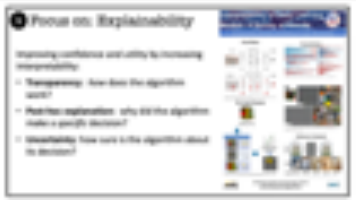
All of the above is in the context of *rapidly formed small coalition teams* with human and machine agents, operating at the *edge of the network*, with *limited connectivity, bandwidth and compute resources* in a *decision-making role*.

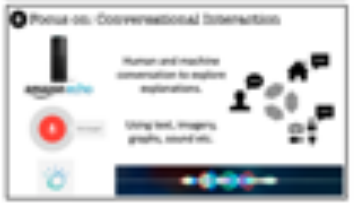
Hypothesis

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://isl.dais-ita.org for full details.

1. Braines, D., Preece, A., & Harborne, D. (2018). **Multimodal Explanations for AI-based Multisensor Fusion**. In *NATO SET-252 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

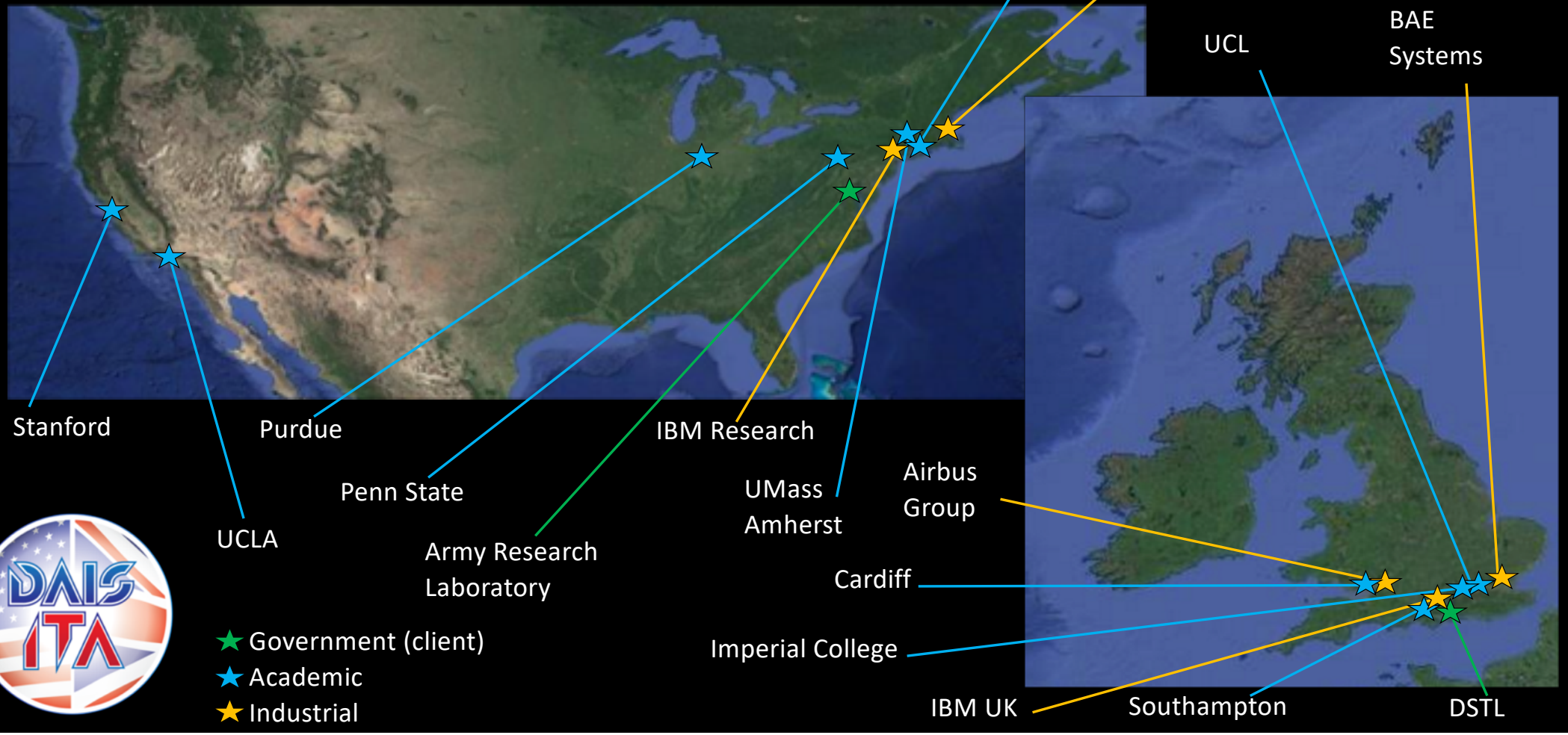

Design Thinking Workshop for AI Explanations with military stakeholders at IBM Hursley, Nov 2018

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 authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Distributed Analytics and Information Science

International Technology Alliance



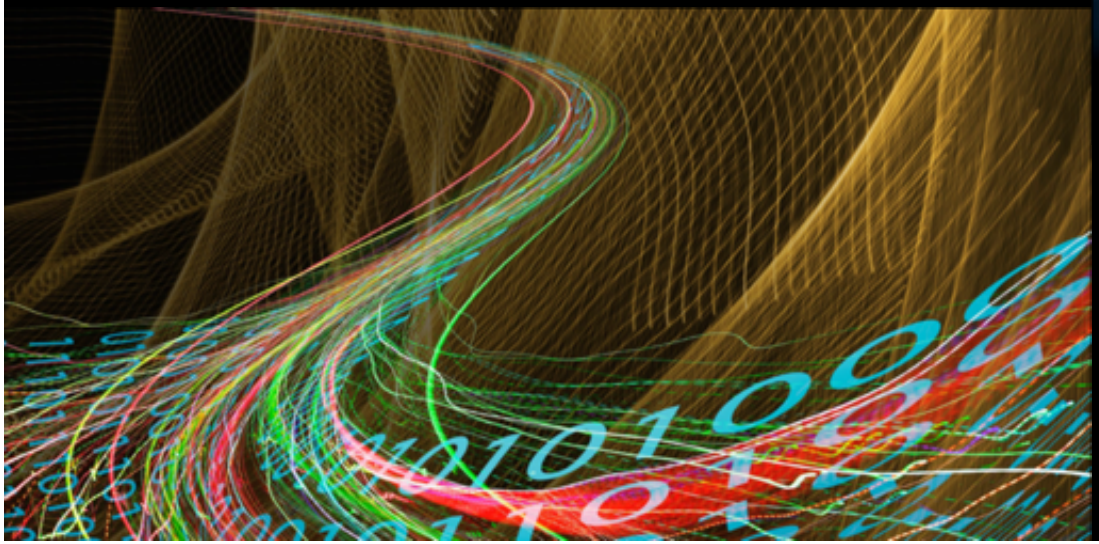
"Create a new and exciting US/UK collaborative community to lead the world in Distributed Analytics and Information Science research"

Focused on rapidly formed coalitions

Running at the edge of the network

Two Technical Areas:

Dynamic, Secure Coalition Information Infrastructures

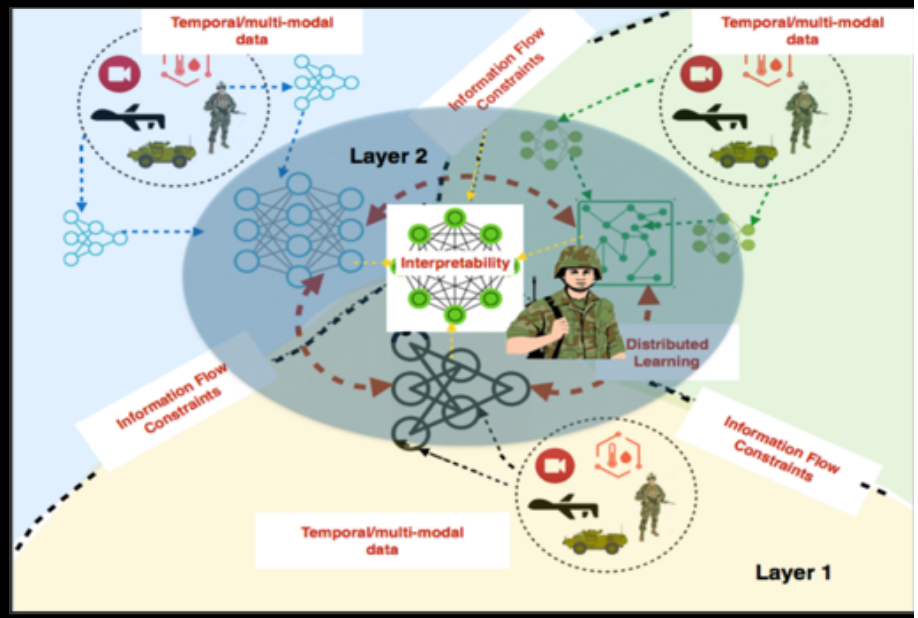


Coalition Distributed Analytics & Situational Understanding

Project 5: Anticipatory Situational Understanding

Problem & Goal:

Need to understand complex and rapidly changing situations given a limited number of personnel and an increasingly rich, varied and distributed data set. The goal is to improve the entire fusion process from signal processing to situational understanding, by synergistically leveraging machine learning and human insight (integrating reasoning and learning approaches).



Gap:

Ability to verify predictive analytics and exploit the synergies between user and machine learning & reasoning.

Challenges & Objectives:

Interpretability of Machine Learning in adversarial settings, with spatio-temporal data sets & with sparse data sets. **Objective**: Understand the efficacy of ML methods/tools in these settings.

Need for fusion techniques able to take account of bias within data sources (inc. learnt models). **Objective:** Understand best approach to handling uncertainty due to bias across entire fusion process.

Need for more expressive knowledge representations affording explainability and tellability to (non-data scientist) users. **Objective:** Step change in efficacy of such representations over current capabilities.

All DAIS publications available online

sl.dais-ita.org/science-library

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

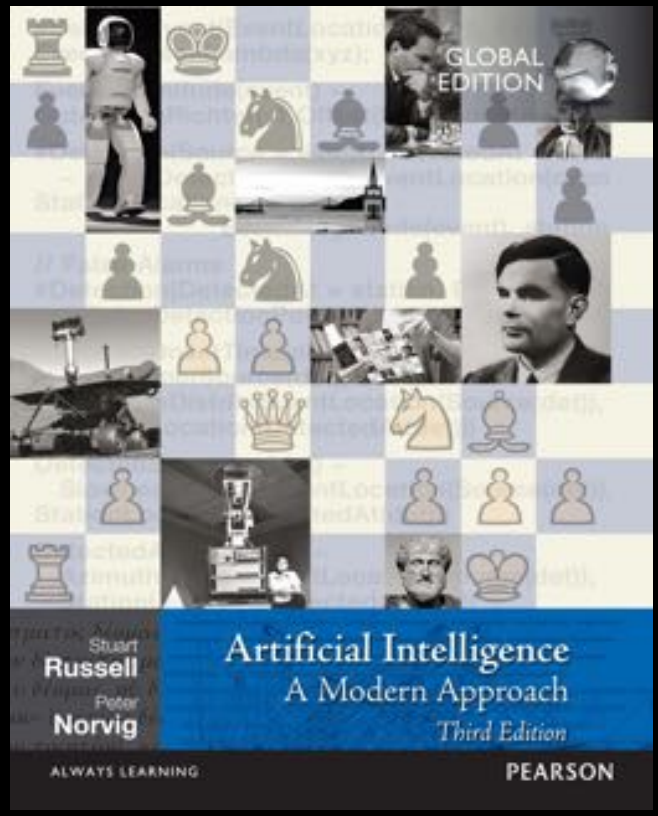
Artifacts that act like humans

Artifacts that think like humans

Artifacts that act rationally

Artifacts that think rationally

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

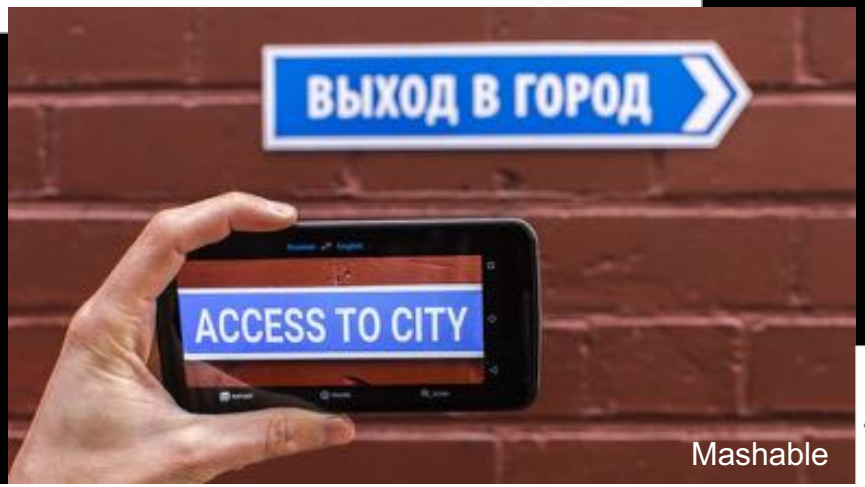


S Russell & P Norvig, **Artificial Intelligence: A Modern Approach** (3rd ed), Prentice Hall, 2009.

Telegraph

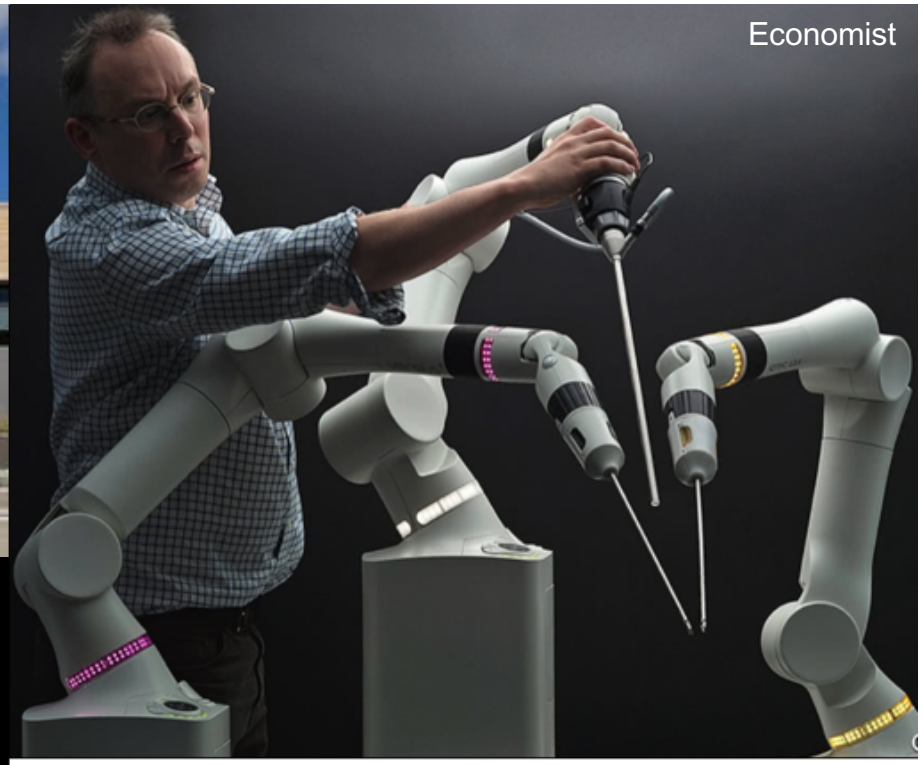
Technology Intelligence

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



Mashable

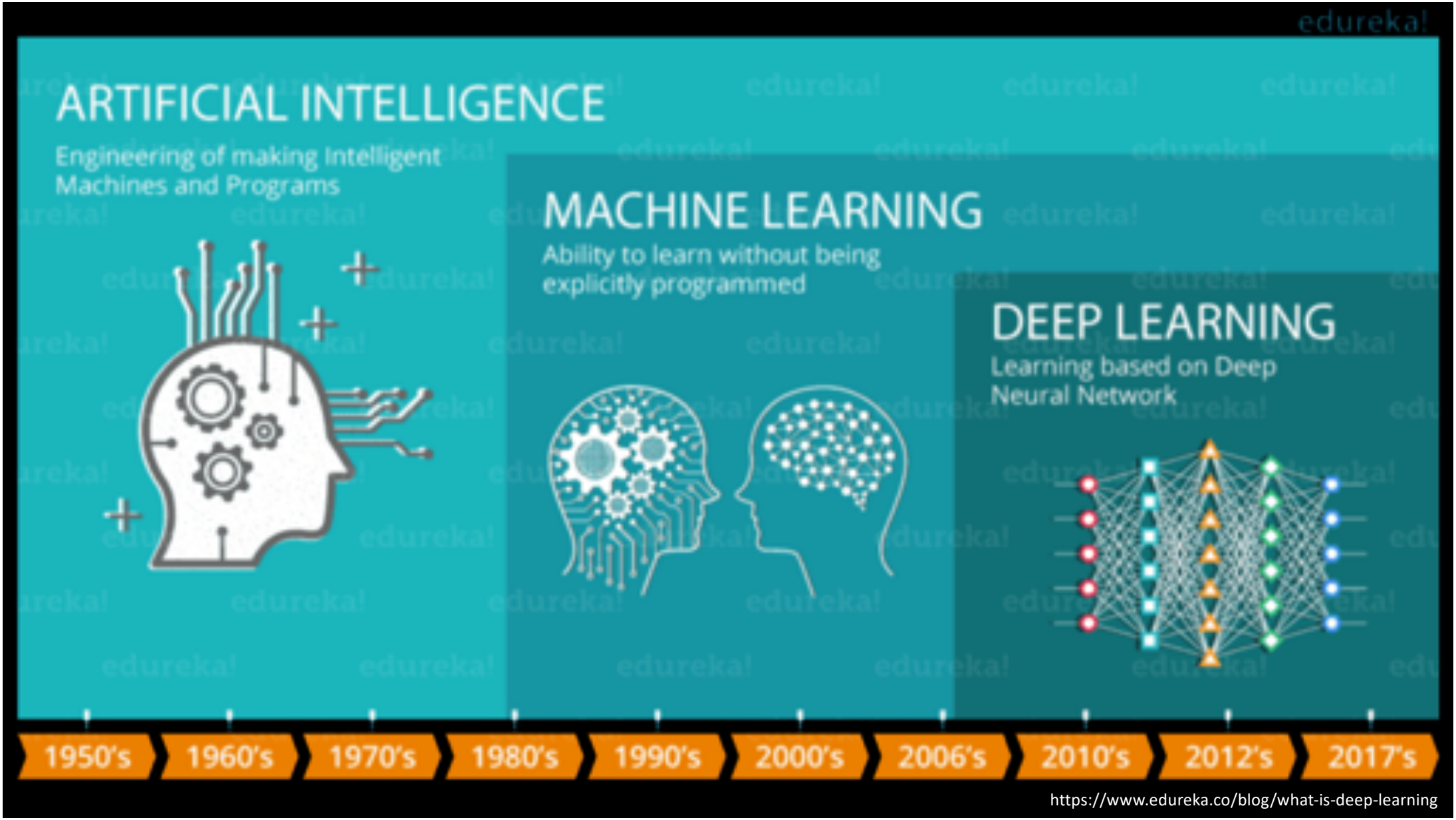
Economist



Medicine

New surgical robots are about to enter the operating theatre

Google Translate gets smarter with language detection, Word Lens


edureka!
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

How AI can help – 6 patterns

[New Orleans Office of Performance and Accountability]

- Finding the needle in a haystack
- Prioritizing work for impact
- Early warning tools
- Better, quicker decisions
- Optimizing resource allocation
- Experimenting for what works

Each of these (and more) have an impact on the need for explanations

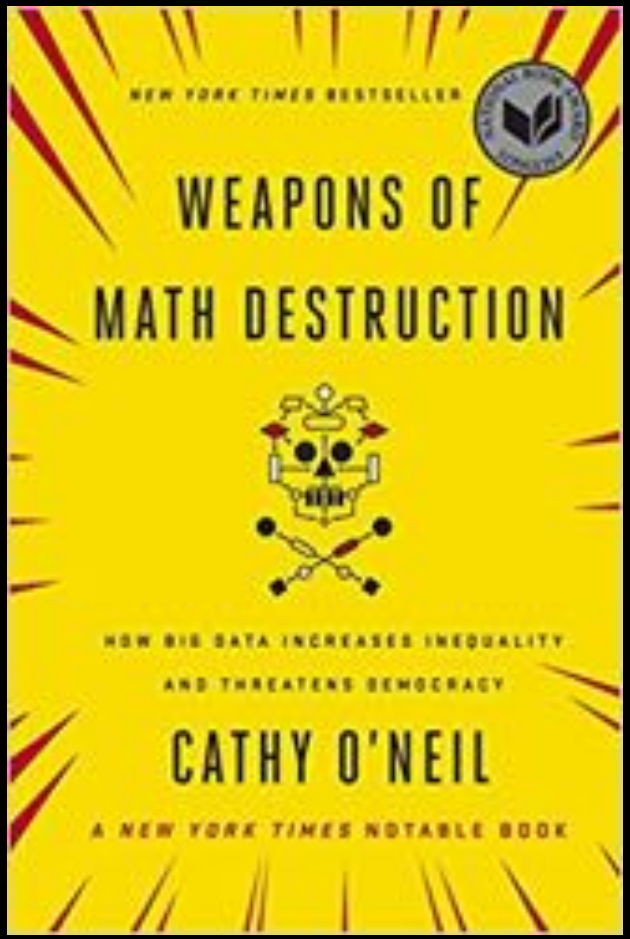
Fairness, Accountability, and Transparency in Machine Learning

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.

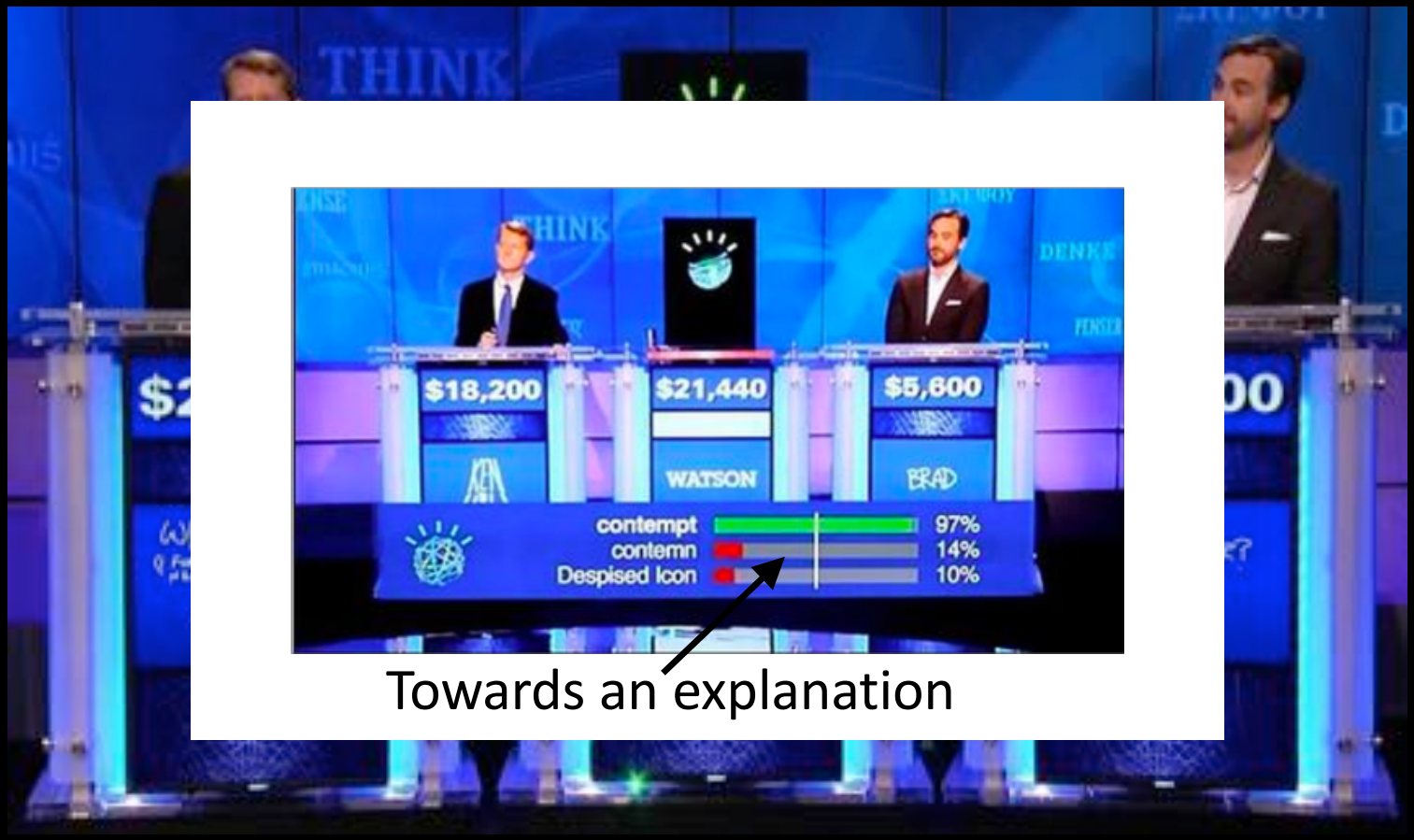
At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to "the algorithm made me do it."



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

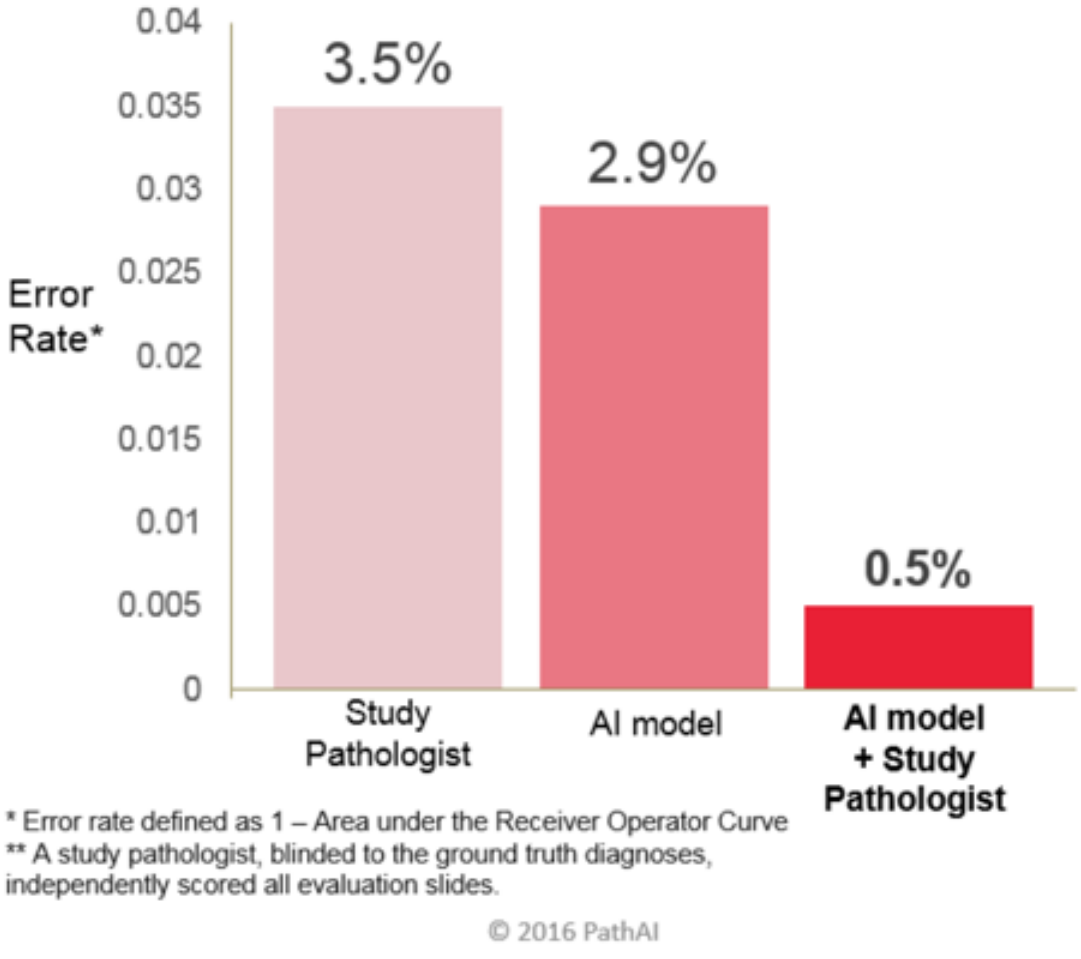
A key idea was that Watson tackled input questions using multiple strategies and needed a method to weigh up its certainty.

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.

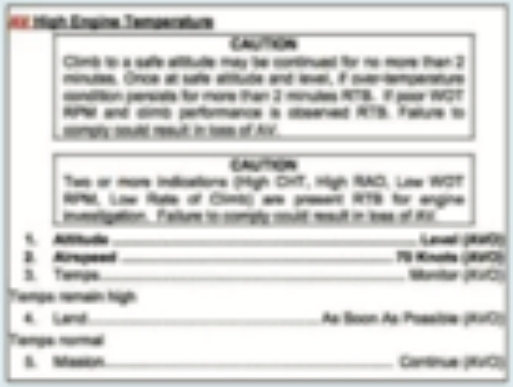
LIVESCIENCE

AI Boosts Cancer Screens to Nearly 100 Percent Accuracy

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

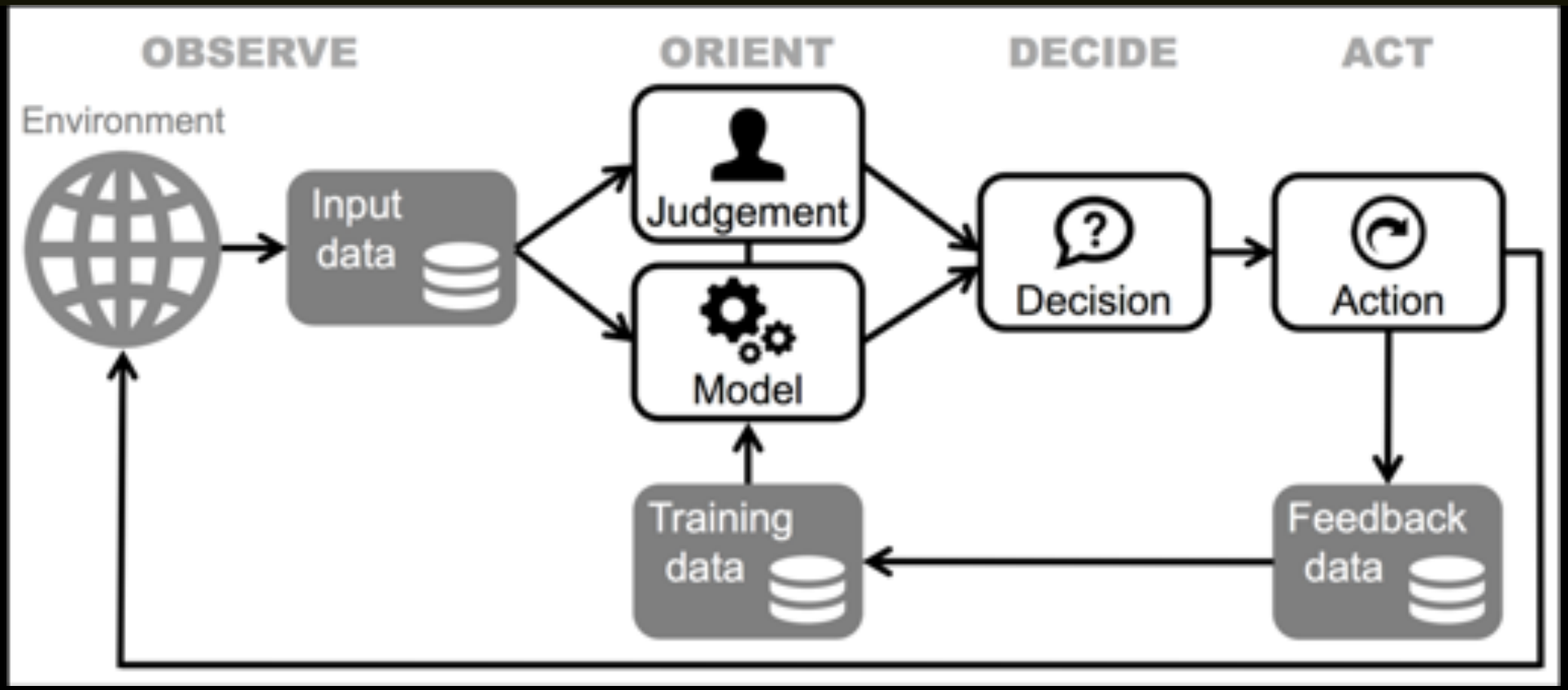
But the real surprise came when pathologists were teamed up with the Harvard team's AI. Together, the artificial intelligence and good, ole human intelligence identified 99.5 percent of the cancerous biopsies.

The “Sully” Factor

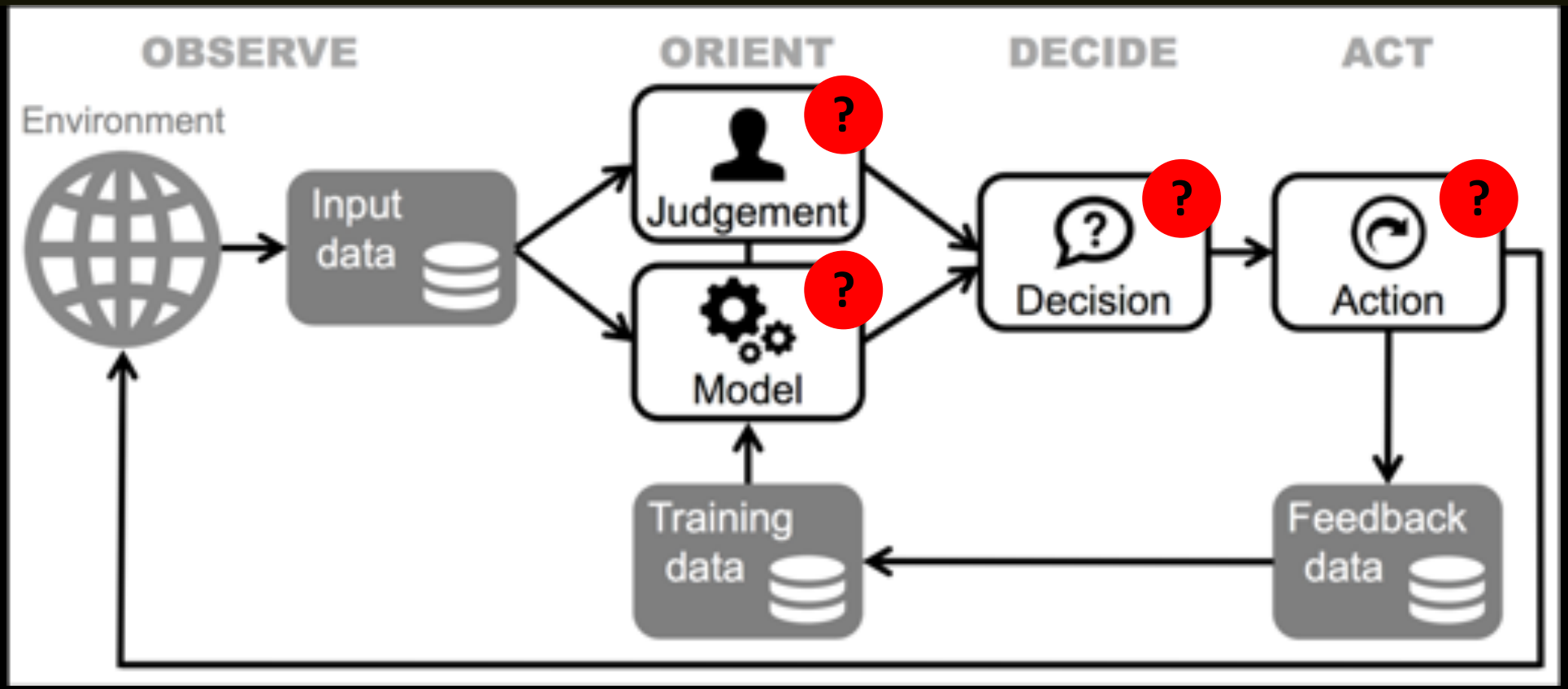


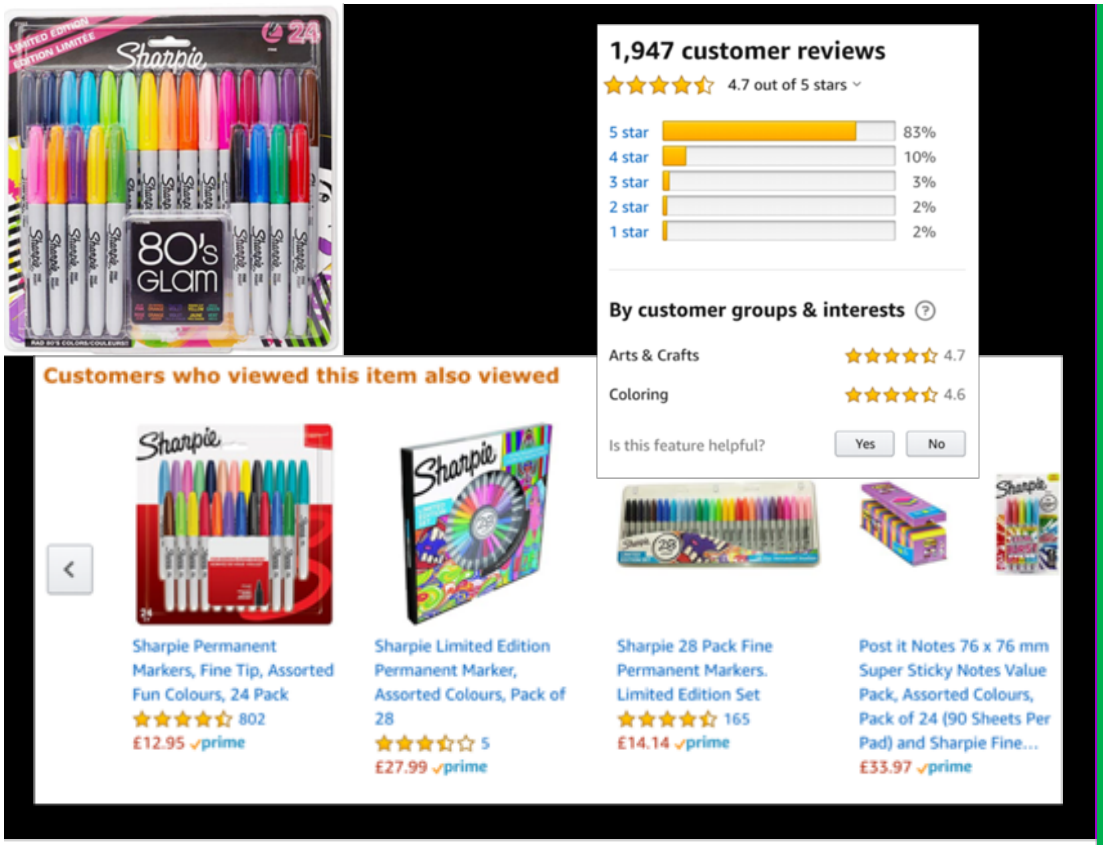
https://ieeexplore.ieee.org/document/6949509/

Human+machine decision loop



Explanation points





Explanations in the wild: Online shopping

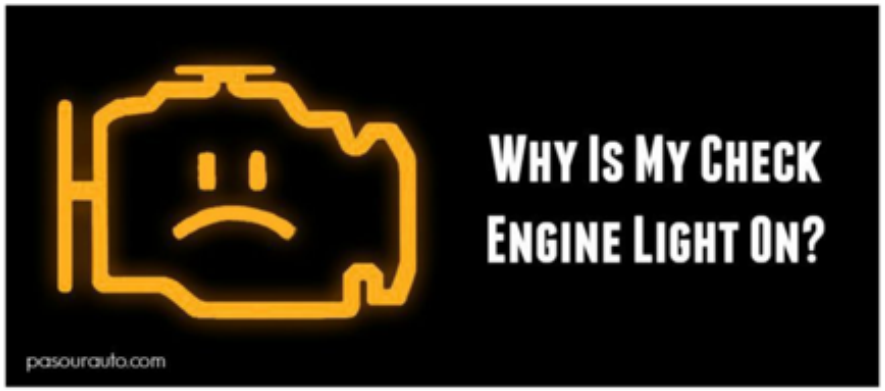
Online retailers (like Amazon) are very good at providing lots of resources to help answer the "*why should I buy this?*" question…

They are present because they are good for business.

Image: Screenshots from www.amazon.co.uk

Why Is My Check Engine Light On?

16 Sep, 2015



"The Check Engine Light strikes fear into the hearts of some and is totally ignored by just as many. Just what it means is a mystery to most drivers"

*"**Your [technician] will plug a scanner into the on-board diagnostic port and read the trouble code stored in the computer.** The trouble code will give the [technician] a starting place as he diagnoses the cause of the problem."*

Explanations in the wild: Warning/status lights

When something is broken it's great to be able to let you know…

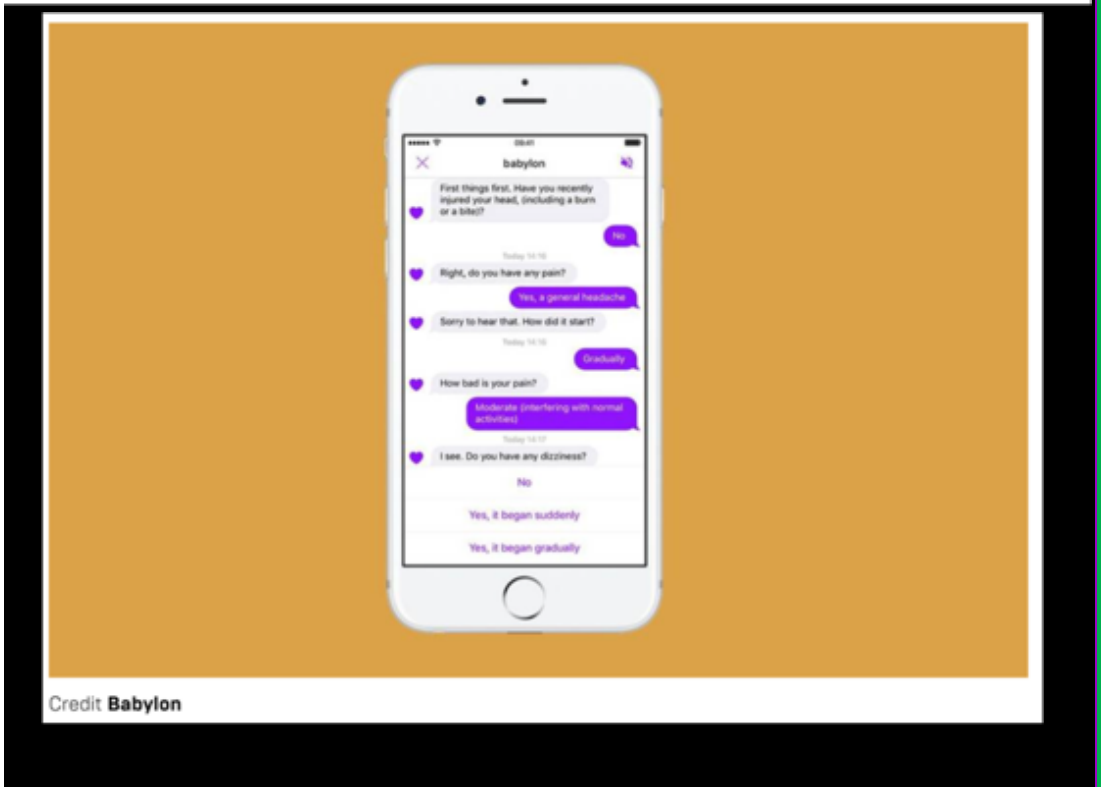
But it's also frustrating! Why can't I get more information. If only I could get inside the "black box" and ask "Why is the engine light on?" (that's what technicians are paid to do…)

Credit: http://www.pasourauto.com/why-is-my-check-engine-light-on

Artificial Intelligence

The NHS is trialling an AI chatbot to answer your medical questions

1.2 million people living in North London can use the app instead of calling the NHS 111 number



Credit **Babylon**

Explanations in the wild: Expert chatbots

Whether the chat-bot is machine or human it doesn't really matter… they are following a well trodden path down a pre-defined set of options. If you try asking "why?" you often won't get an answer.

Chatbots like this could be built entirely around "why?" questions…

Source: https://www.wired.co.uk/article/babylon-nhs-chatbot-app

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

Explanations

Explanation methods:

- Expressive power
- Translucency
- Portability
- Algorithmic complexity

“Explanations enable interpretations”

Individual explanations:

- Accuracy
- Fidelity
- Consistency
- Stability
- Comprehensibility
- Certainty
- Degree of Importance
- Novelty
- Representativeness

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…

Collaborative XAI Research

Collaborative XAI Topics

1. Working with Westpoint Military Academy cadets on Machine Learning Explanations

2. Applying “Design Thinking” techniques to Explainable AI with Military Advisors from UK forces

Working with Westpoint Military Academy

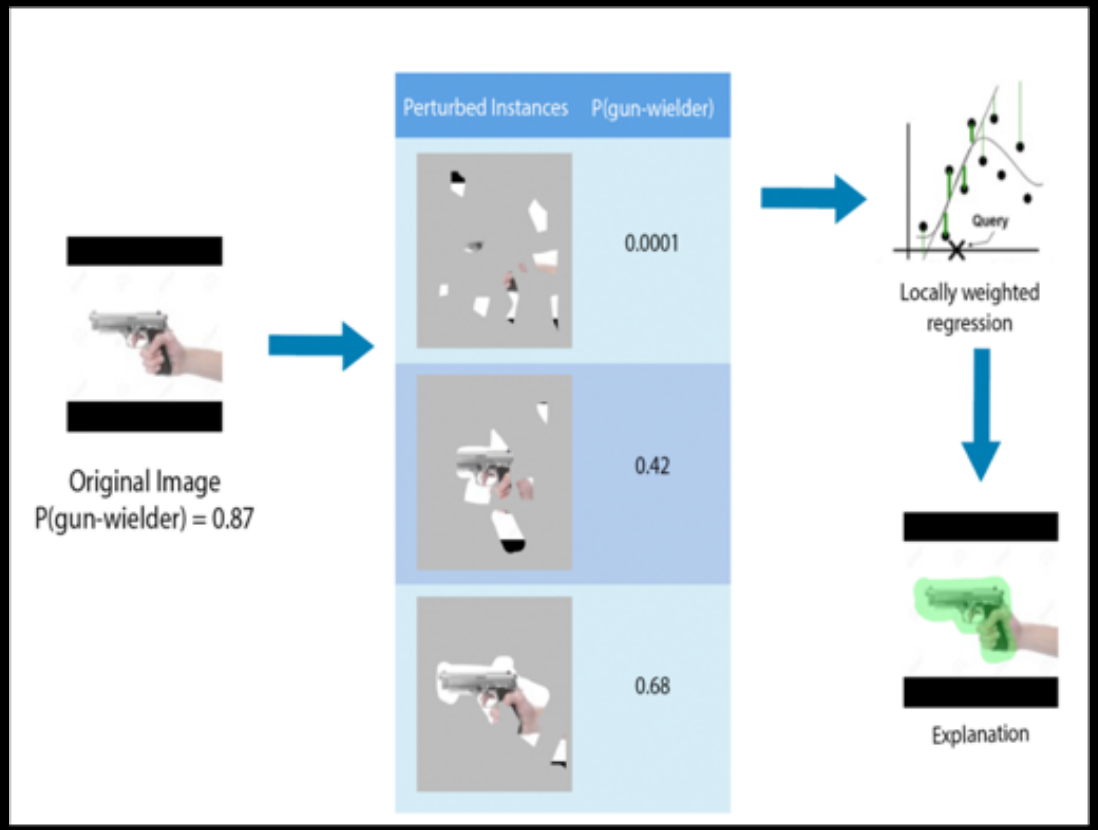
- Two visiting cadets from Westpoint Military Academy (US)
- 4 week rotation to IBM Hursley (UK)
- Wrote a paper for our internal DAIS conference (NY, US):
  - “An Analysis of Reliability Using LIME with Deep Learning Models” http://sl.dais-ita.org/science-library/paper/doc-2904
- Developed core ideas into more complete solution (IBM)
- Poster presented at SPIE DCS next week (Baltimore, US)
  - “Developing the sensitivity of LIME for better machine learning explanation” (will also be published to Science Library next week)

Exploring interpretability reliability for LIME

- LIME is one of the Model-Agnostic Methods
- It is perceived to be unstable for a number of reasons
- We tasked our visiting Westpoint cadets to explore this area
  - Is it unstable?
  - Can you do anything about it?
- Our traffic congestion dataset was not ideal for their task
  - They sourced a “gun wielder” dataset with clearer artifacts in the images

How does LIME work?

The LIME technique creates multiple different perturbed images in order to calculate a final explanation image in the form of a saliency map



LIME: Local Interpretable Model Agnostic Explanations

Calculated output weights

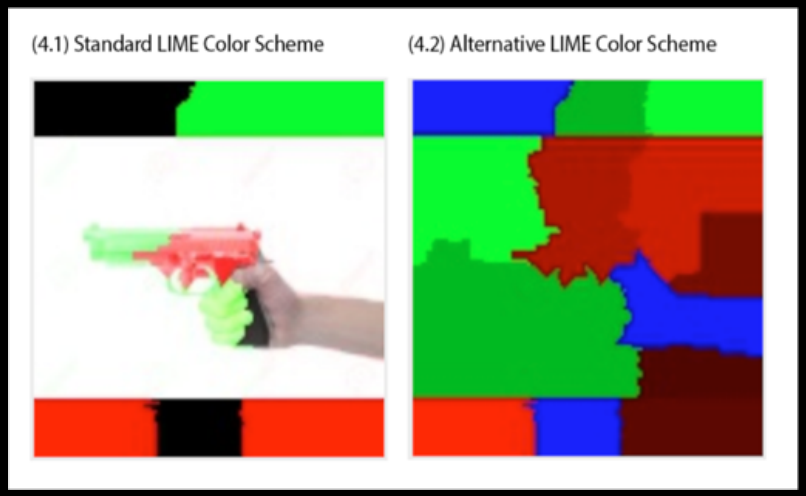
Image	Region 1	Region 2	**Region 3**	Region 4	**Region 8**	Region 9	Region 10
(2.1)	-0.05713165	0.05079833	**-0.31549571**	0.01499846	**-0.01942119**	0.06441398	0.01049571
(2.2)	-0.09951557	-0.09344236	**-0.31167949**	-0.02786685	**0.54578947**	0.07719415	-0.01167949
(2.3)	0.00512246	0.03044236	**-0.00293567**	0.05392633	**0.57944333**	-0.04223455	0.01293567

The 3 images above are separate explanations of the same image.
There is a lot of instability in the explanations.

Looking at averages and standard deviations

Set	Region	1	2	3	4	8	9	10
1	Average Weight	-0.02	0.06	**-0.38**	0.07	**0.46**	0.01	0.01
	Standard Deviation	0.08	0.07	**0.06**	0.06	**0.06**	0.06	0.07
2	Average Weight	-0.01	0.06	**-0.32**	0.07	**0.43**	0.02	0.01
	Standard Deviation	0.05	0.07	**0.06**	0.08	**0.07**	0.07	0.07
3	Average Weight	-0.02	0.05	**-0.35**	0.06	**0.39**	0.04	0.01
	Standard Deviation	0.08	0.06	**0.07**	0.06	**0.07**	0.06	0.08

Comparing alternative color schemes



The alternative color scheme added more information to the explanation image.
The darker the color of green or red, the more positive or more negative the weight was.
Blue is neutral.

Conclusions from the cadets

- LIME visual explanations of an image appear unstable
- By using averages and standard deviations across multiple explanations we showed improved stability
  - Some random elements had to be fixed, e.g. the choice of superpixels.
- The failure to convey the weight of the region in the visualization is one source of the perceived instability
- Using averages and a shading technique to convey weight help use confidence in the stability of the explanations.

- Poster presented at SPIE DCS next week (Baltimore, US)
  - “Developing the sensitivity of LIME for better machine learning explanation”

Design Thinking for Explainable AI

IBM Design Thinking

Eunjin Lee (Ellie)
Emerging Technology Specialist
& Senior Inventor

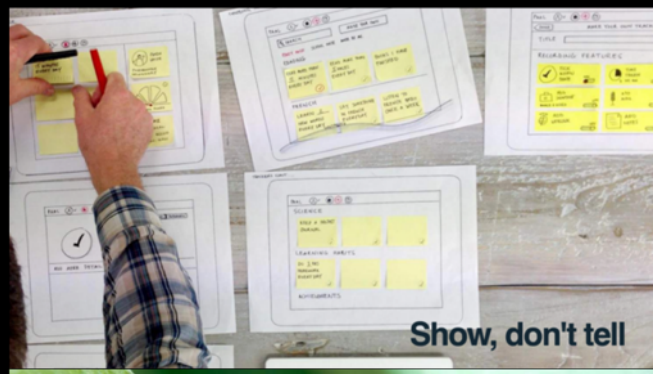
Graham White
Emerging Technology Specialist
& Master Inventor

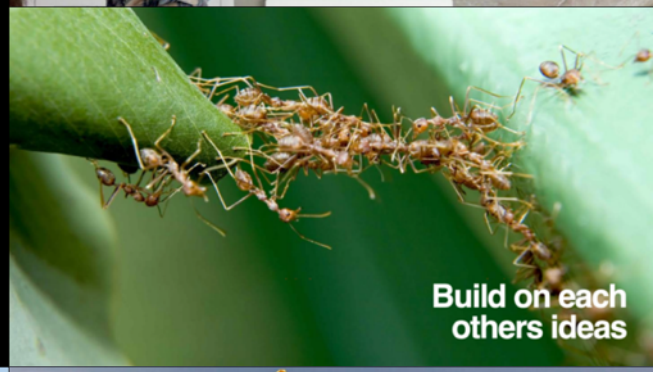
Emma Dawson
Emerging Technology Specialist

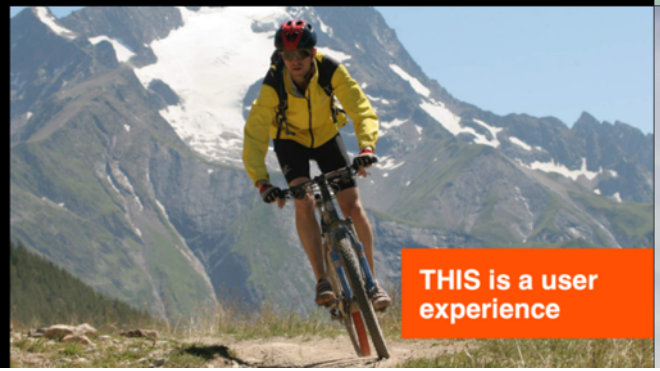
What is Design Thinking?

- A process for idea elicitation
- Widely used for software or product development
- Our team has lots of applied experience

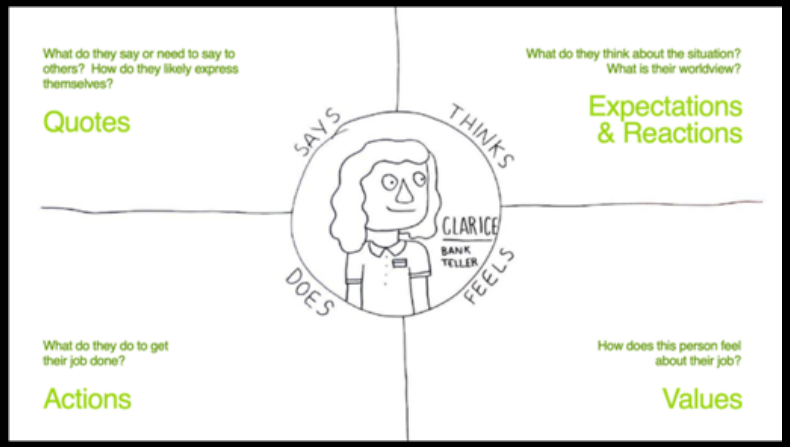
- But we've never used it for research before
- …nor with military professionals!







What did we do?



Empathy maps

1

Big ideas

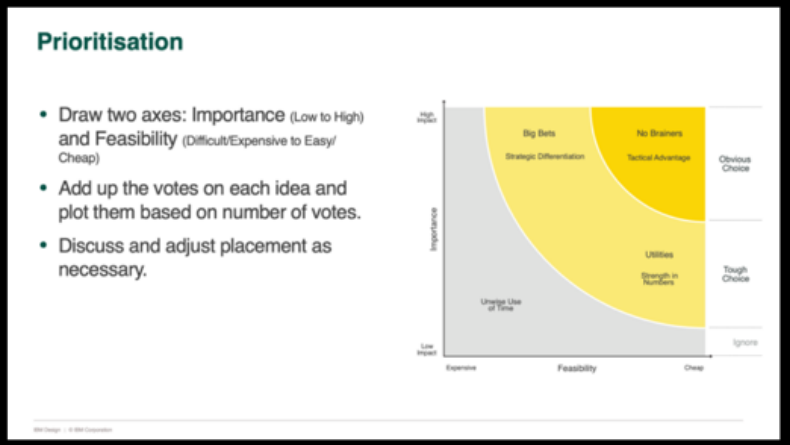
3

"As-is" scenario

2

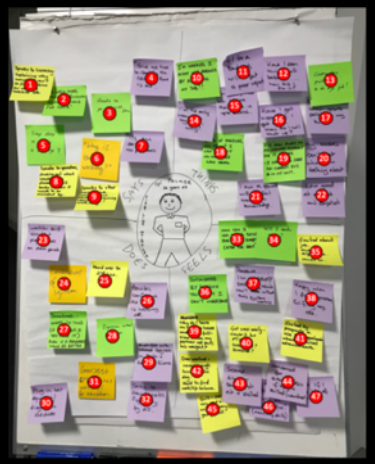
Prioritization grid

4



What happened?

- 3 teams
  - Military, Scientist, facilitator
- 1 day
  - Introduction
  - Compressed process
  - Subset of techniques
- Great results
  - Nearly 600 post-it comments
  - A number of emerging themes
  - Opportunities and concerns…
- Journal paper in progress
  - Further iteration planned



Deep Learning

Black Box Explanations

Deep Learning - Explainability

Accuracy & Comprehensiveness

Input

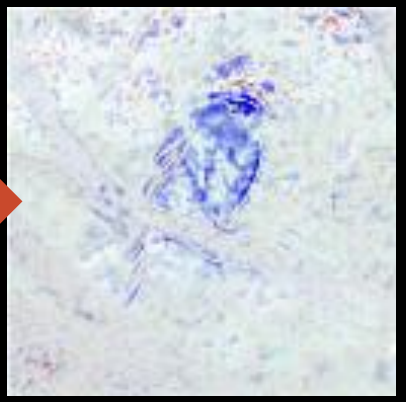
Black Box Model

Prediction

goldfinch

Explanation

e.g. SHAP



Why are interpretable models important?

- Debug at creation time
- Explain a specific prediction/model output at run time
- Investigate an issue after the fact (accountability)
- Gain further understanding in related domains

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”)

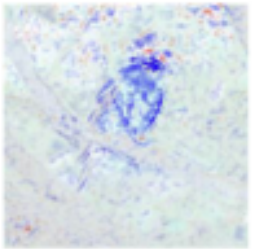
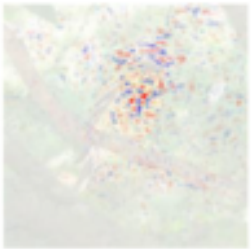
Categories:

(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

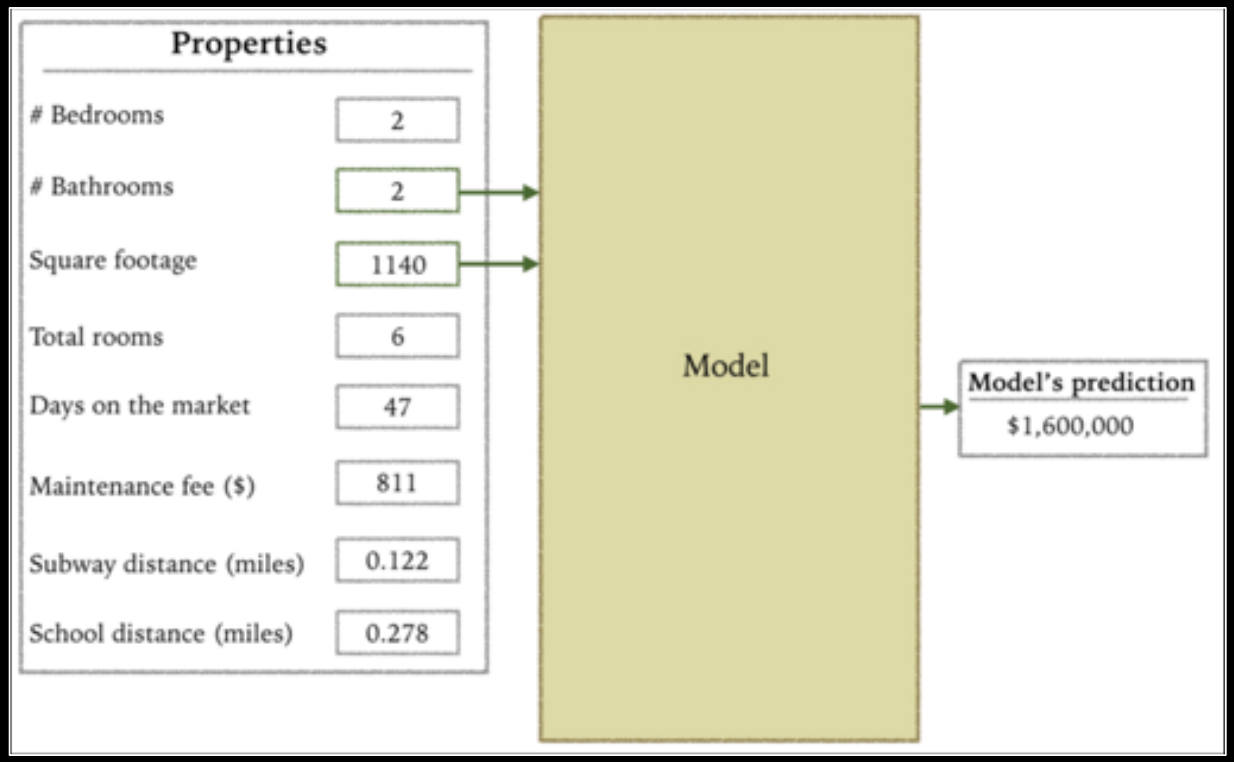
Explanation Types and Techniques

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

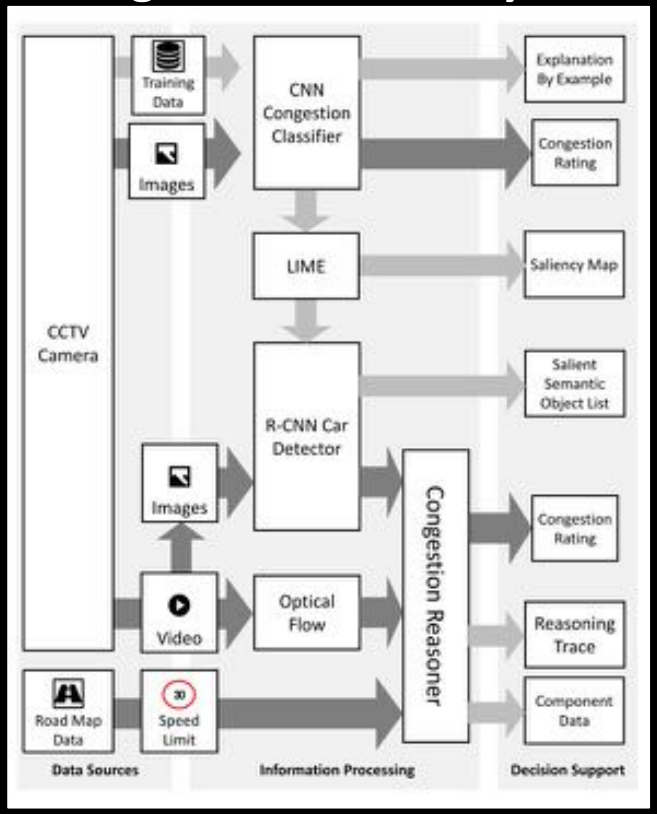
Data:
- Bathrooms: 2
- Square Footage: 1140

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

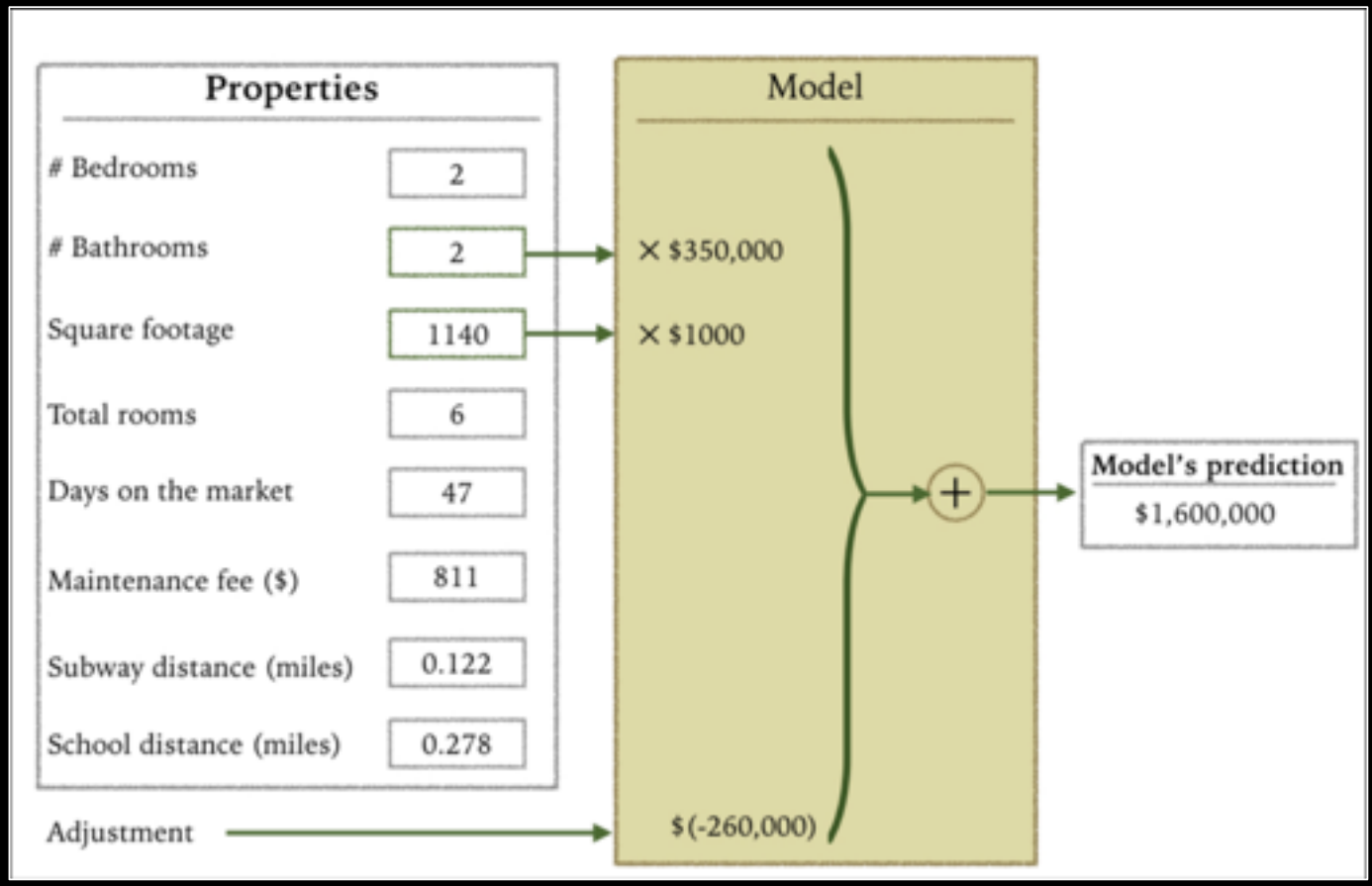
Congestion Reasoner
- *Congestion Rating*: 0.67
---- *Optical Flow*: 2.3
---- *Speed Limit*: 30 MPH

…

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

Explanation Types and Techniques

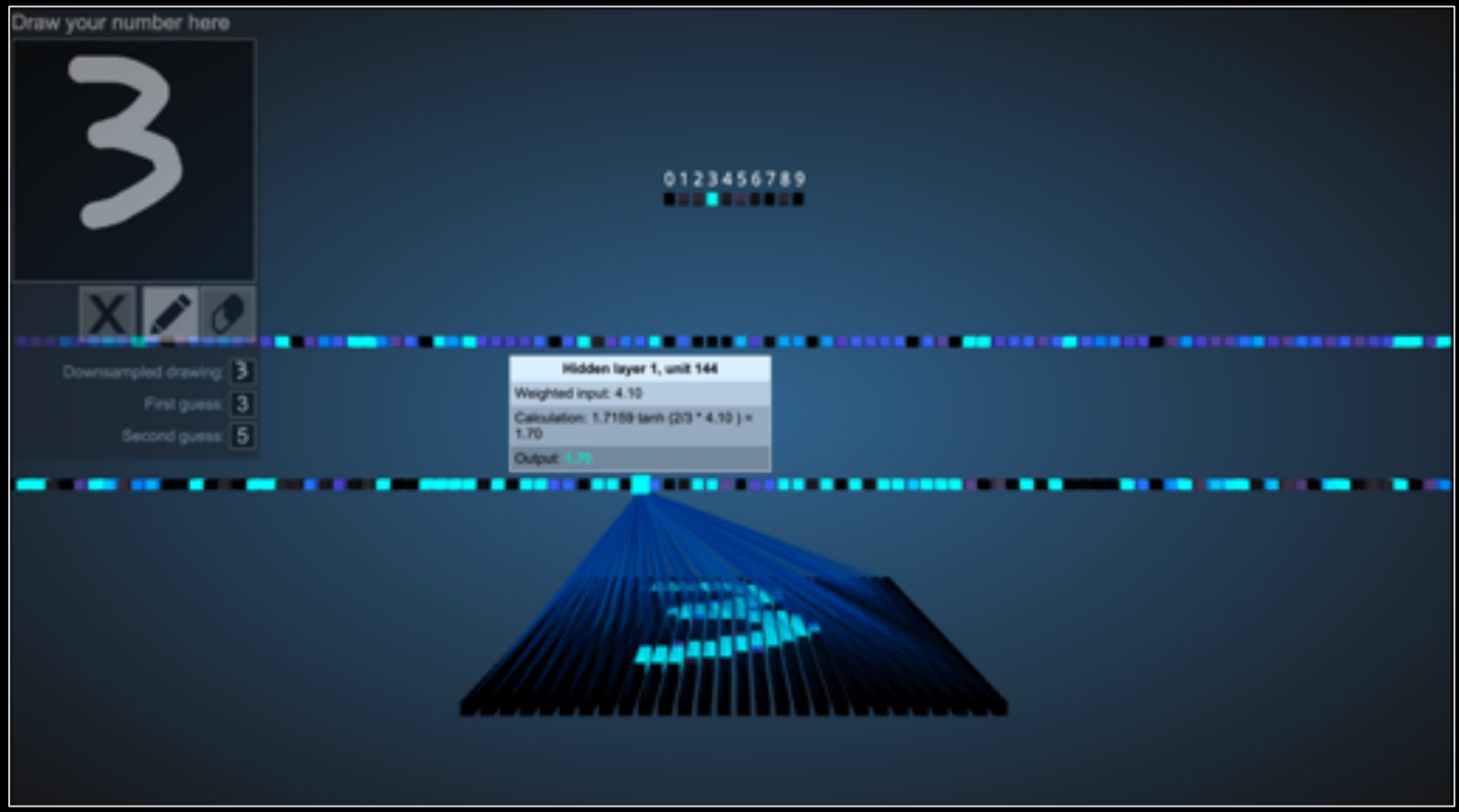
Model Internals



Manipulating and Measuring Model Interpretability - Poursabzi-Sangdeh 2018

Explanation Types and Techniques

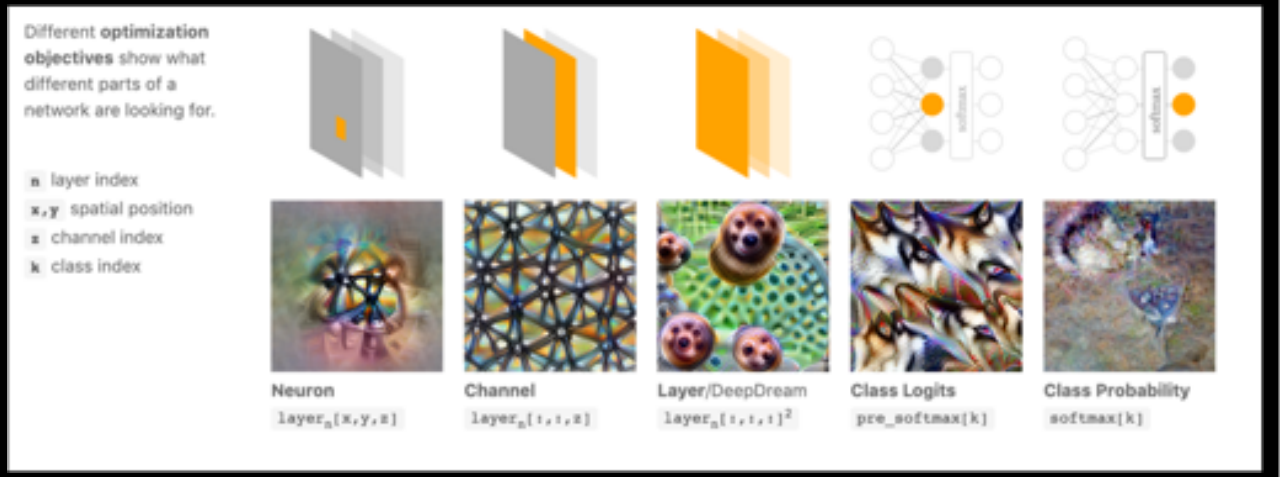
Model Internals

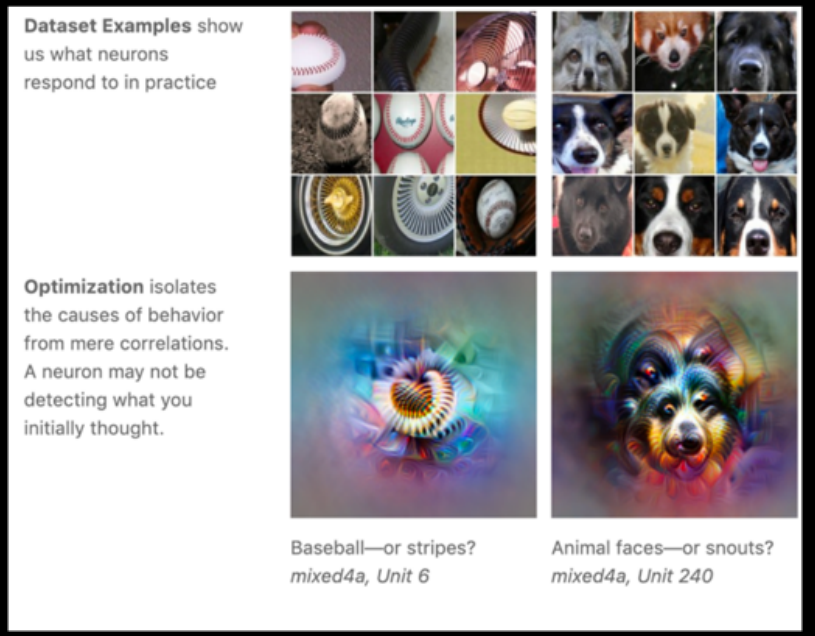


3D visualization of a Convolution Neural Network - http://scs.ryerson.ca/~aharley/vis/fc/

Explanation Types and Techniques

Feature Visualization

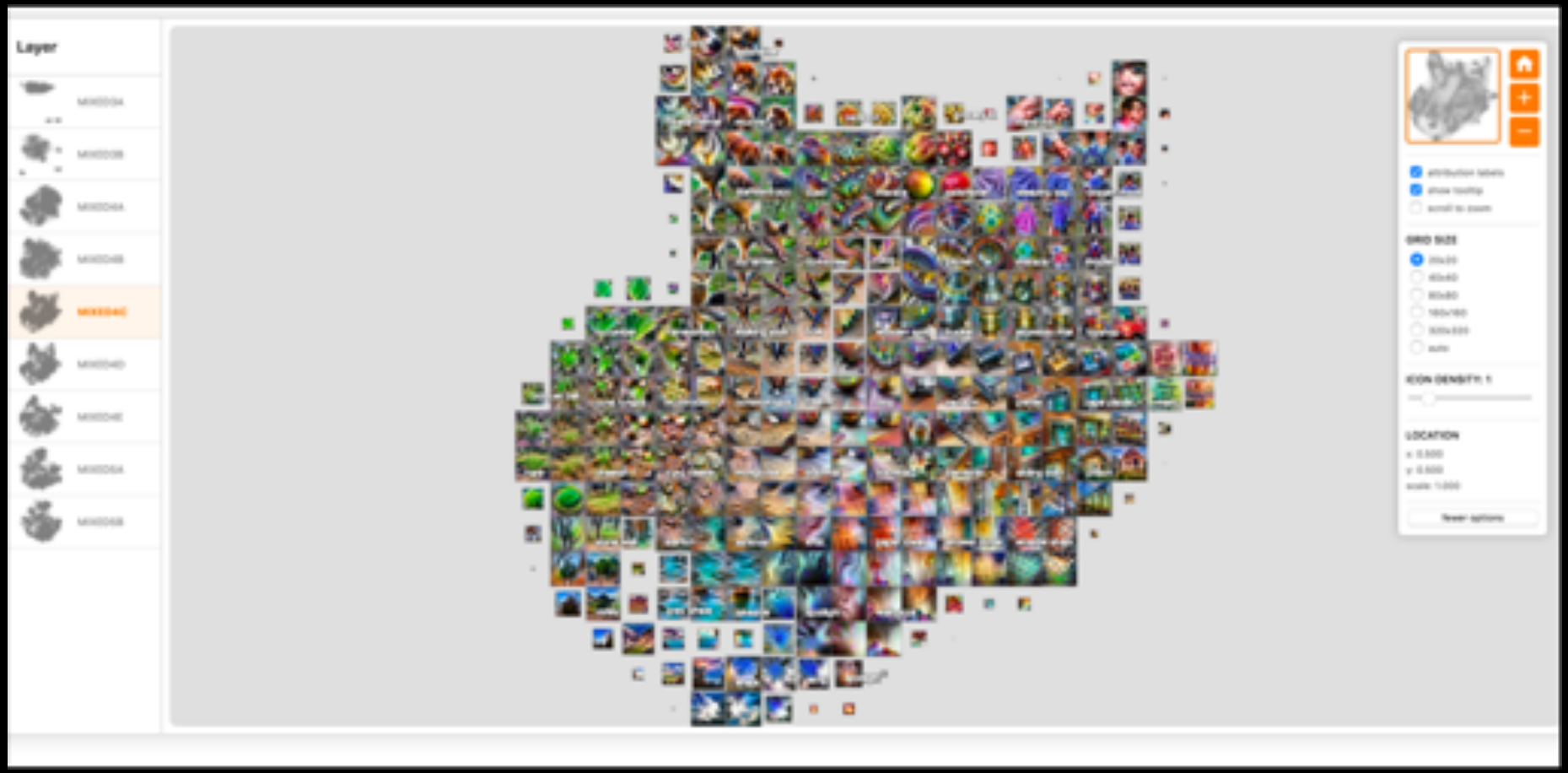




Feature Visualization - Olah, et al. 2017

Explanation Types and Techniques

Feature Visualization



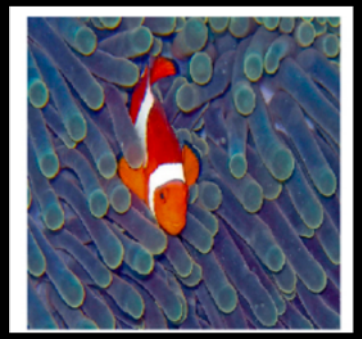
Exploring Neural Networks with Activation Atlases - Carter, et al. 2019 (March 6, 2019)

Explanation Types and Techniques

Explanation by Example

Understanding Dog Vs Fish Classification Using Influence Functions

Test Image



Helpful (“influential”) Images from Training Data

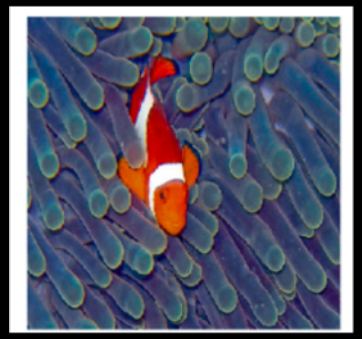
Understanding Black-box Predictions via Influence Functions - Koh et al. 2017

Explanation Types and Techniques

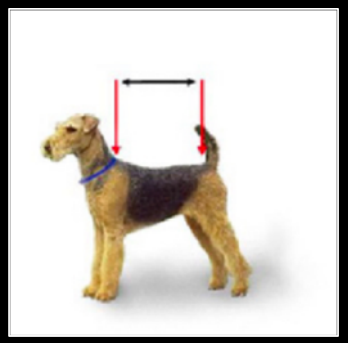
Counterfactual Explanation by Examples

Understanding Dog Vs Fish Classification Using Influence Functions

Test Image



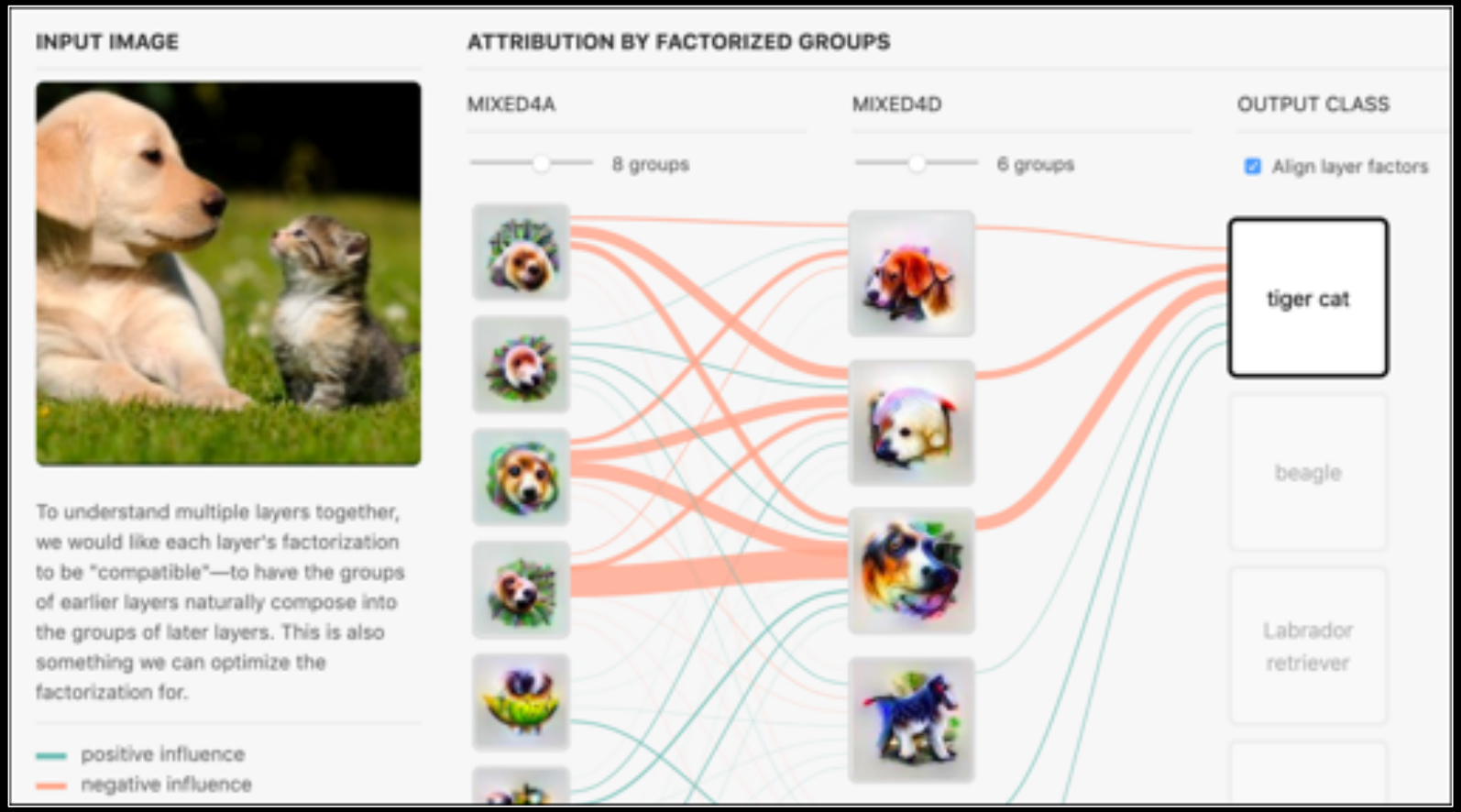
Helpful (“influential”) Images from Training Data



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

Issues in Current State of Interpretability

- Siloed Communities – Stakeholders in Explainable AI (Preece et al. 2018)
- Possible issues with many popular, widely used techniques
- **What makes a “good” explanation?**

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

Explainability

Accuracy & Comprehensiveness

Input

Black Box Model

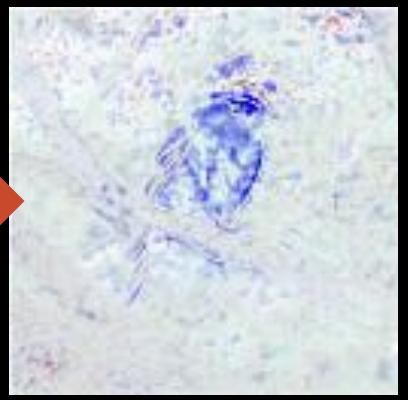
Prediction

goldfinch

Explanation Comprehensiveness

Explanation

e.g. SHAP

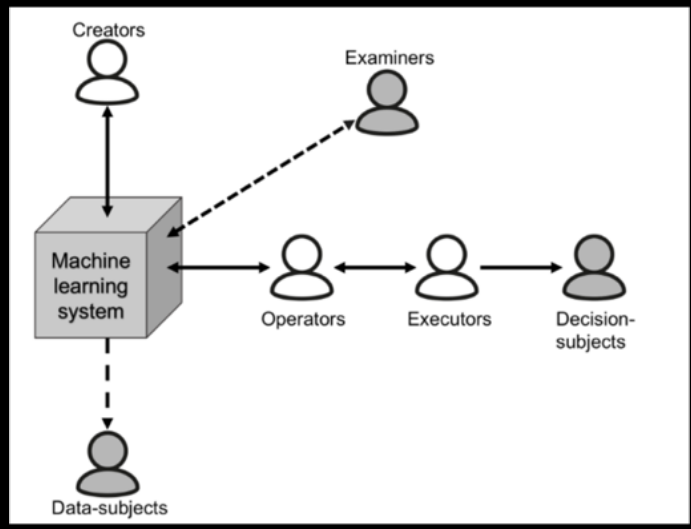


Explanation Accuracy

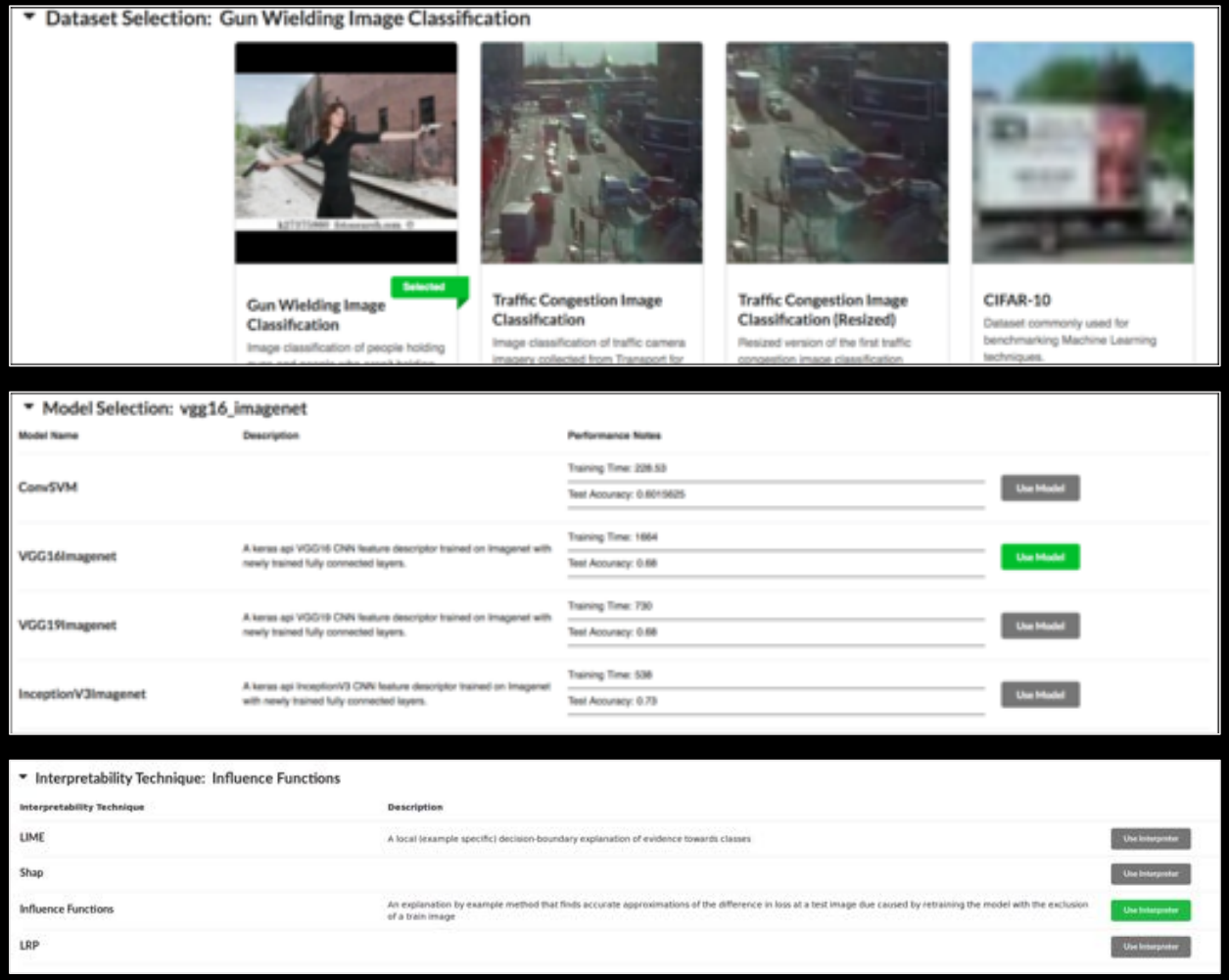
Interpretability

Aspects of a User

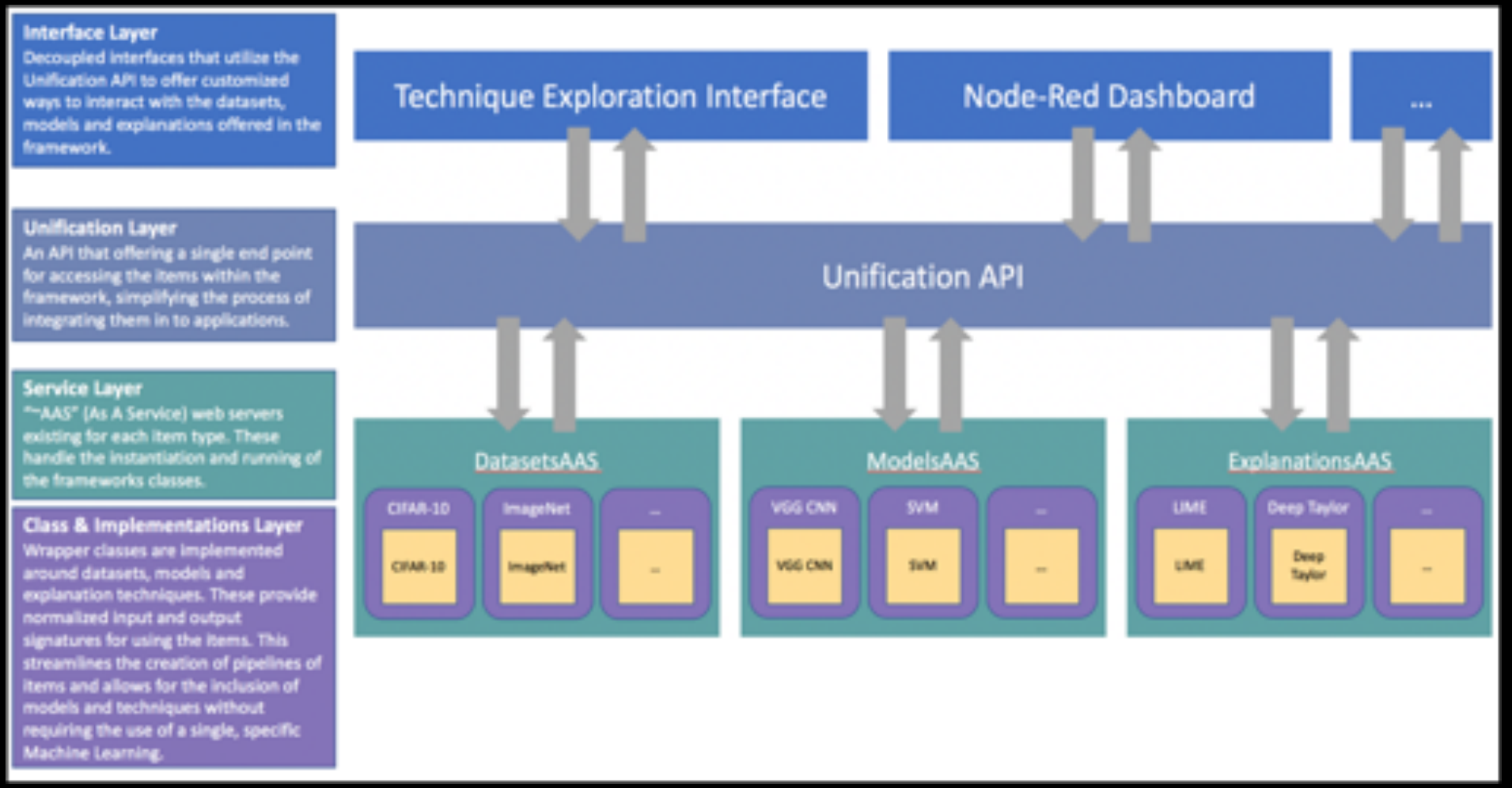
- Prior Knowledge
  - Machine Learning Knowledge
  - Task Domain Knowledge
- Decision Information
- Preference
- Purpose



Experimentation Framework – Our Interface

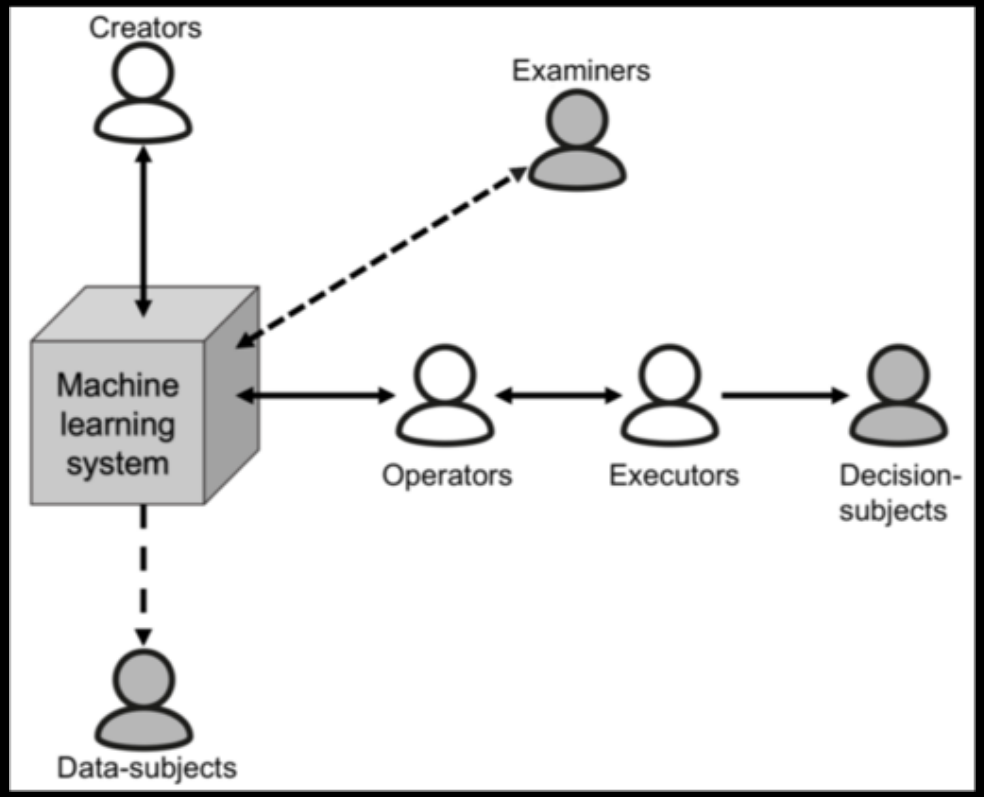


Experimentation Framework - Architecture



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.

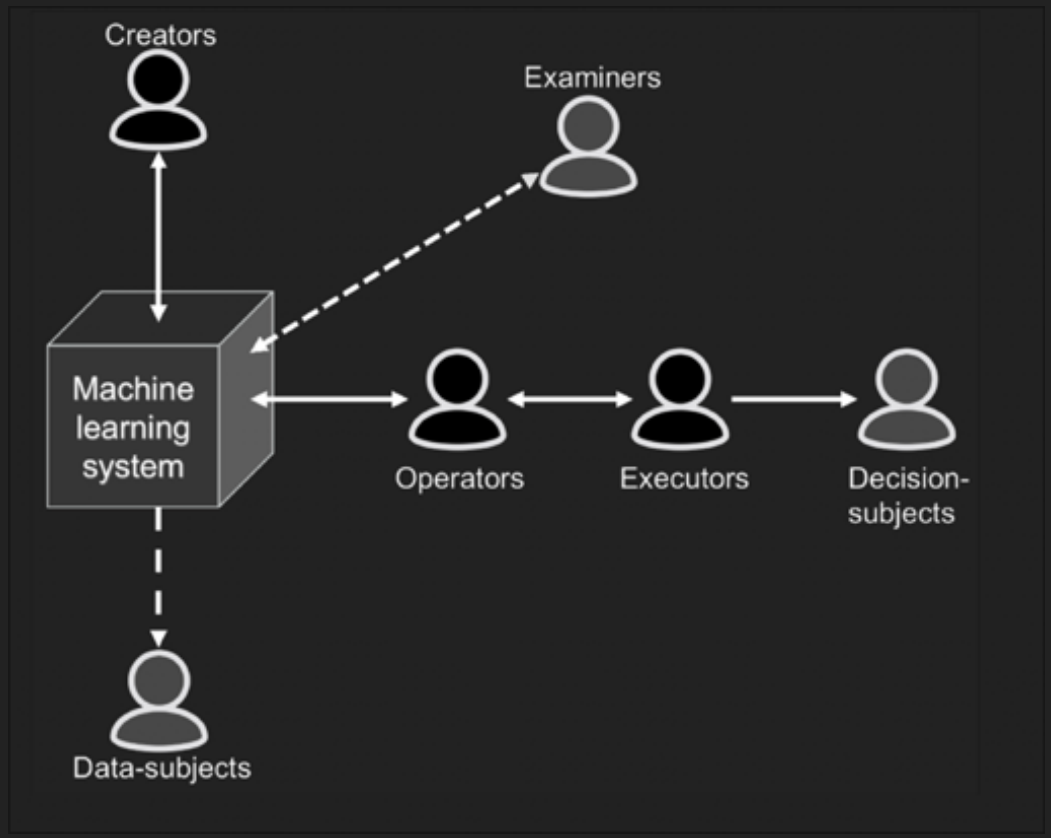
Terminology

- *Explanation*: the information provided by a system to outline the cause and reason for a decision or output for a performed task – a "post-hoc explanation" in Lipton's taxonomy (Lipton, 2016).
- *Interpretation*: the understanding gained by an agent with regard to the cause for a system's decision when presented with an explanation.

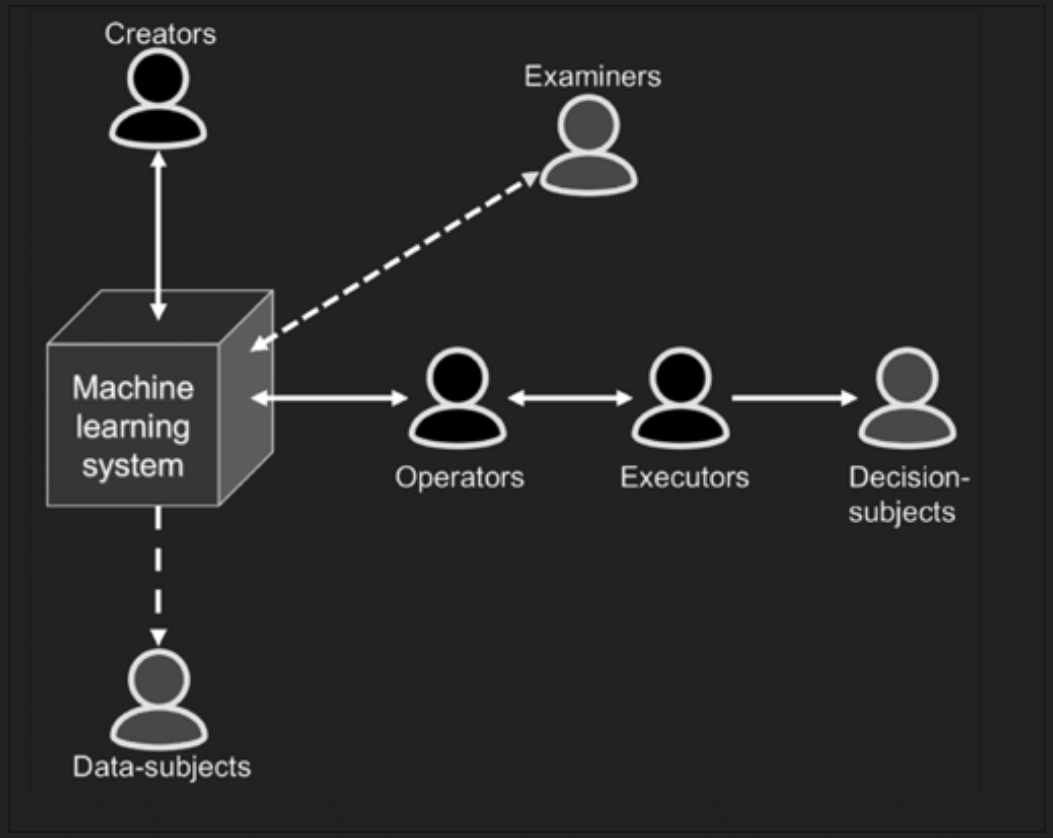
Terminology

- *Explainability*: the level to which a system can provide clarification for the cause of its decisions/outputs.
- *Transparency*: the level to which a system provides information about its internal workings or structure, and the data it has been trained with – this is similar to Lipton's definition of transparency (Lipton, 2016).
- *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.

Role-based Model

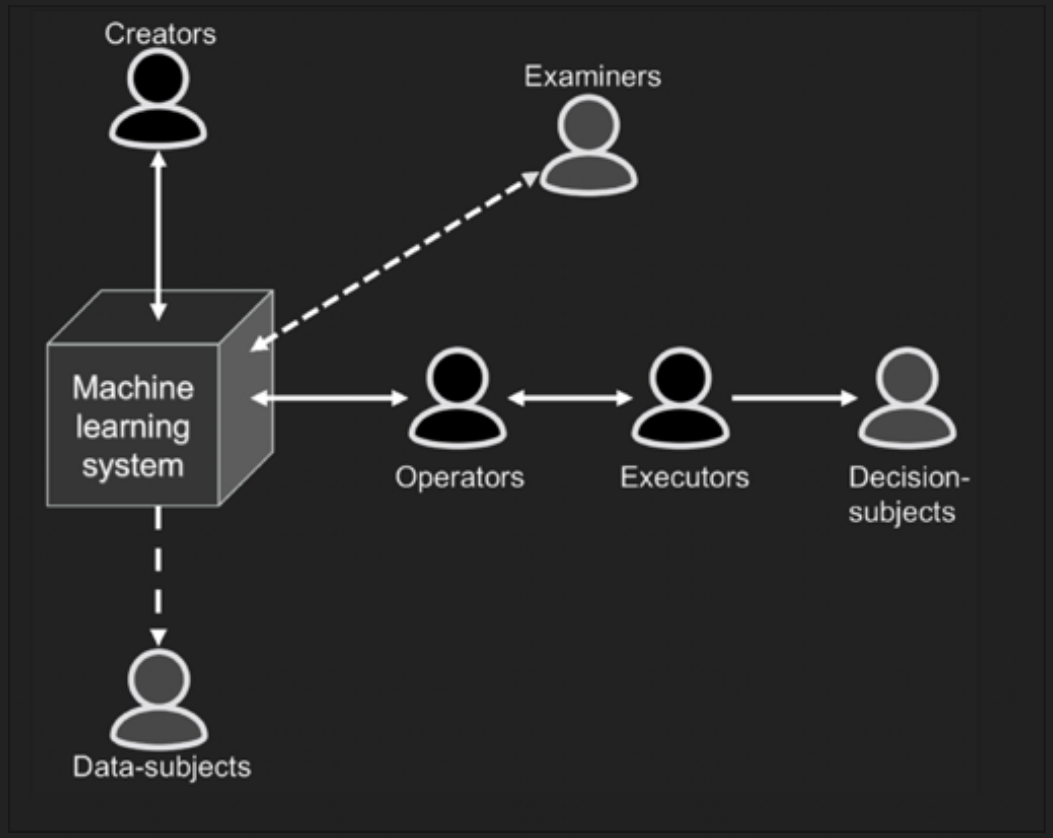


Scenario 1: Web advertising



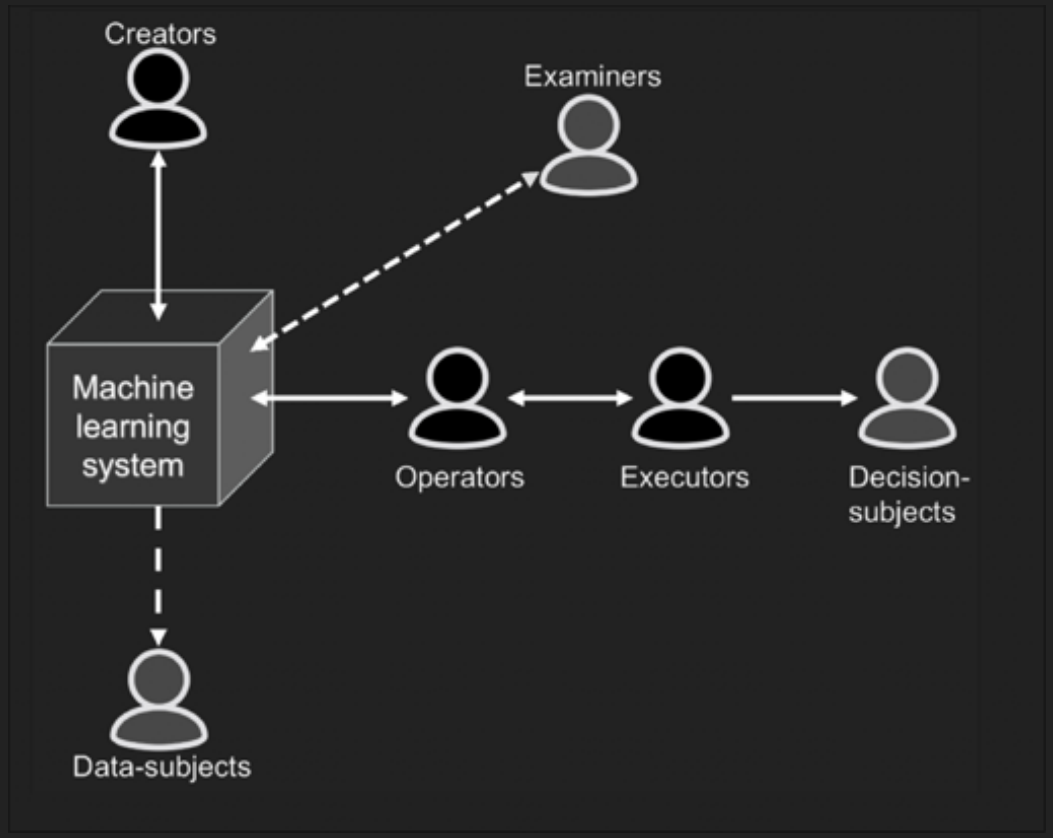
- **Creators**: The advertising company and its employees, any third-party development companies and their employees
- **Operator**: the host web-site
- **Executor**: the host web-site
- **Decision-subject**: the web-site user
- **Data-subjects**: any internet denizen whose data has been obtained by the advertising companies
- **Examiners**: relevant advertising standards body staff, “data commissioner” style authority staff (e.g., the UK’s Information Commissioner’s Office); usually, such authorities will only become examiners if a complaint or information request is made

Scenario 2: Route planning on a smartphone



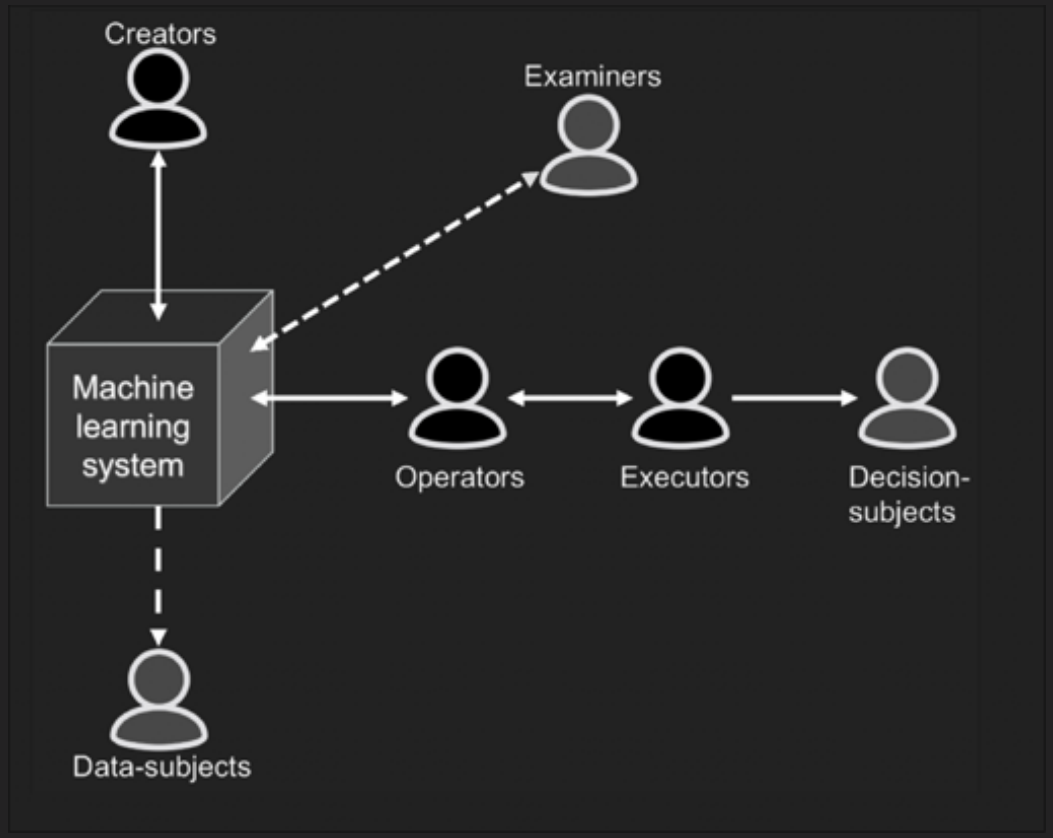
- **Creators**: the navigation app company and its employees
- **Operator**: the app user
- **Executor**: the app user
- **Decision-subject**: the app user
- **Data-subjects**: any road users whose location data has been obtained and used by the navigation app company
- **Examiners**: “data commissioner” style authority staff; usually, such authorities will only become examiners if a complaint or information request is made

Scenario 3: Loan application



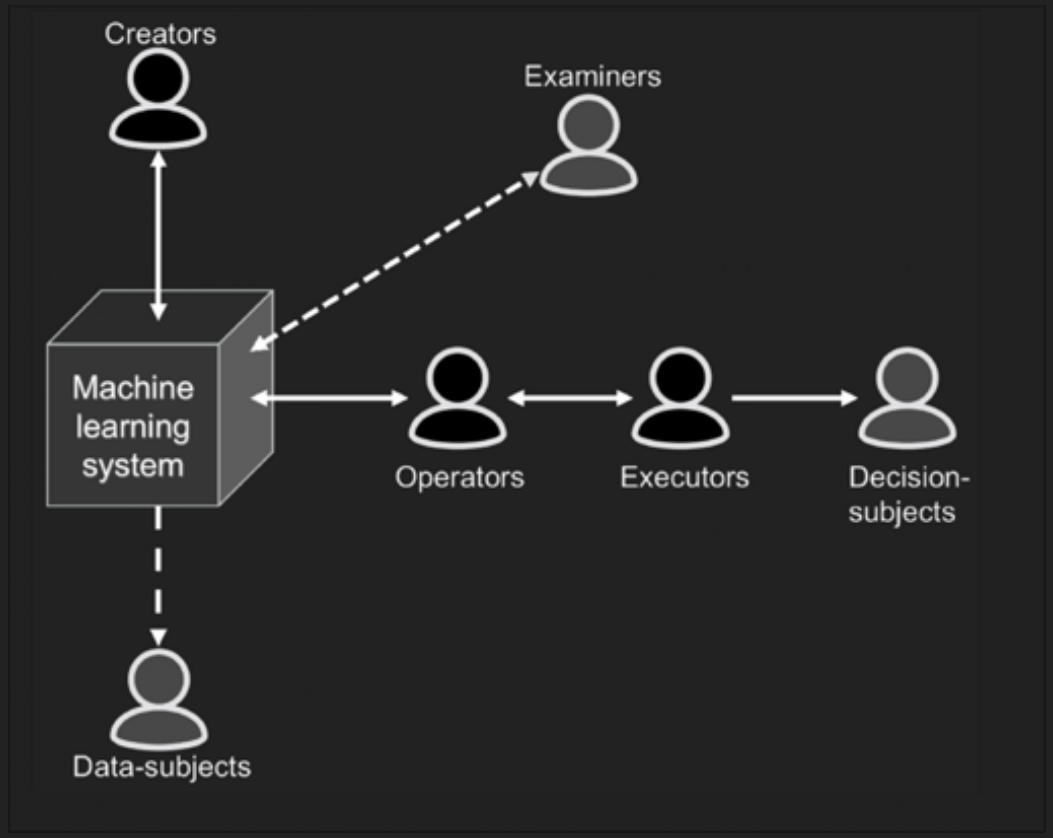
- **Creators**: The lender and its employees if they also developed the system, or any third-party development companies and their employees
- **Operators**: the lender's (customer-facing) employees
- **Executor**: the lender's (higher-ranking) employees
- **Decision-subject**: the loan applicant
- **Data-subjects**: prior loan applicants, any agent whose data has been obtained by the lender (likely most financial service users)
- **Executors**: financial regulation authority staff, financial ombudsman

Scenario 4: Medical advice for clinicians



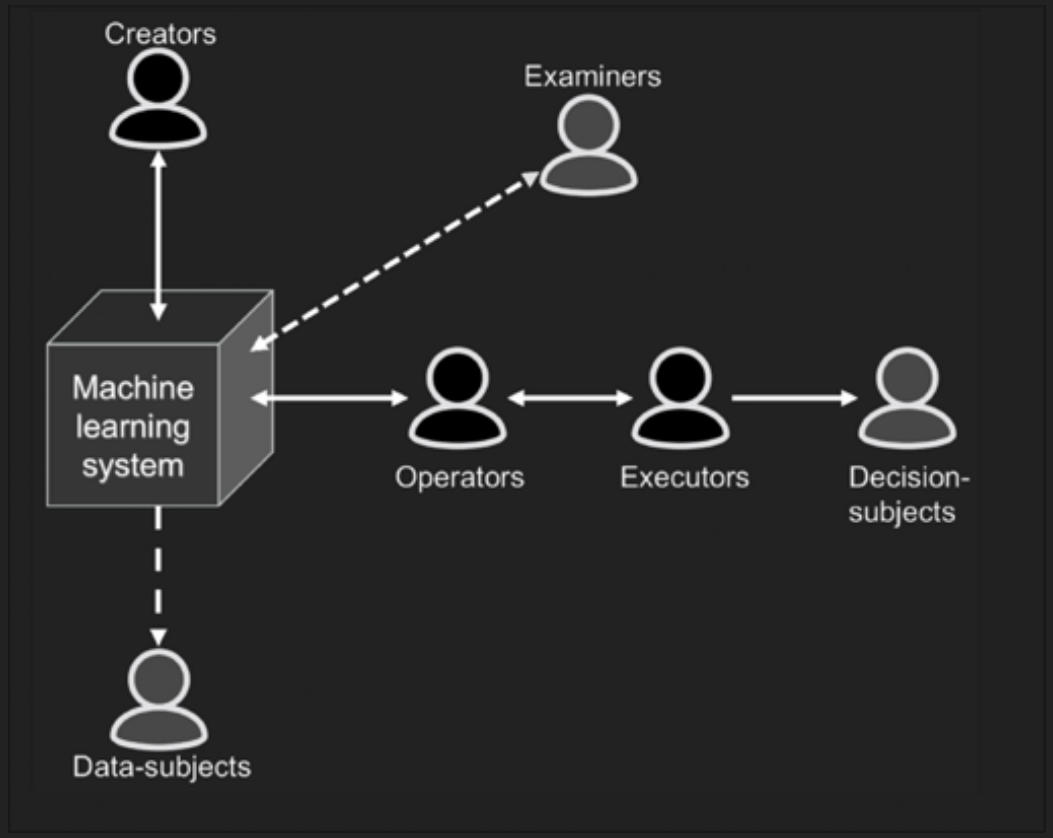
- **Creators**: the medical software company and its employees, any collaborating medical professionals and researchers
- **Operators**: medical professionals
- **Executors**: the patient, medical professionals
- **Decision-subject**: the patient
- **Data-subjects**: other patients, researchers and study subjects (e.g., data loaded from publications)
- **Examiners**: professional medical authority staff e.g., the UK's General Medical Council

Scenario 5: Releasing defendants on bail



- **Creators**: the legal software company and its employees
- **Operators**: the judge (or other court staff)
- **Executor**: the judge
- **Decision-subject**: the defendant
- **Data-subjects**: previous defendants
- **Examiners**: In the case of an appeal, the original decision may be scrutinized by, for example, the defendant's lawyers. In this scenario, these lawyers would become examiners.

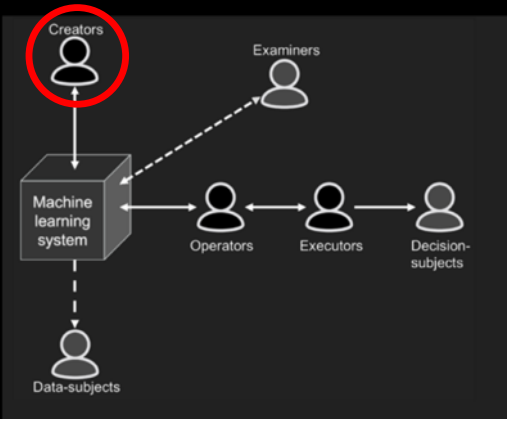
Scenario 6: No-go order in a military operation



- **Creators**: employees of the various coalition partners' militaries, employees of any military contractors involved
- **Operators**: military analysts
- **Executor**: the mission commander
- **Decision-subject**: the target, or agent identified as the target
- **Data-subjects**: other individuals of interest, their known associates
- **Examiners**: tribunal jurors

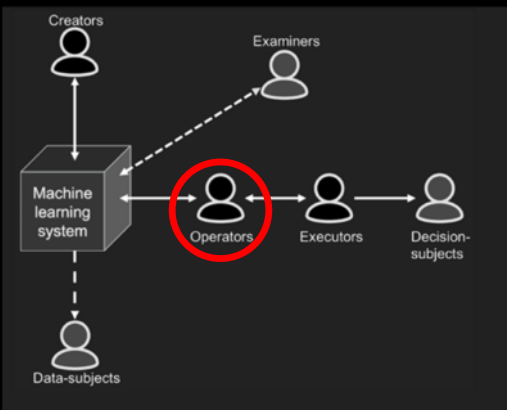
Role-based interpretability: Creators

- Architects, designers, writers, technical engineers
- Create the system
- Own the intellectual property and rights to the system
- May be a collaborative effort
- May include subject matter experts for training or data collection
- Performance and metric optimization
- Goal: Total understanding of the system (debug mode)
- Requires: Explainability and Transparency



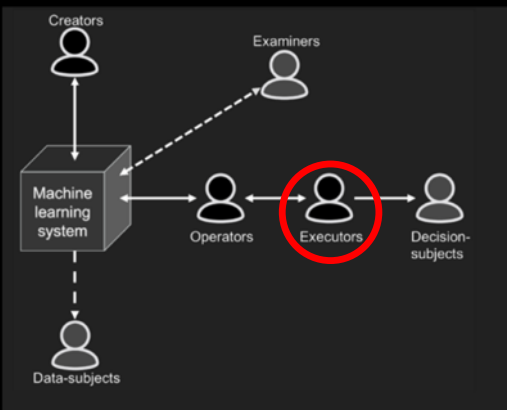
Role-based interpretability: Operators

- (Often the operator is also the executor)
- Need to validate inputs and sanity check outputs
- Pass outputs (or summary) on to executor
- May need to make multiple requests, or fuse information from other sources
- Useful techniques may include:
  - Summarization
  - Feature relevance highlighting
- Goal: Trust in the system behavior to support executor
- Requires: Explainability and possibly Transparency



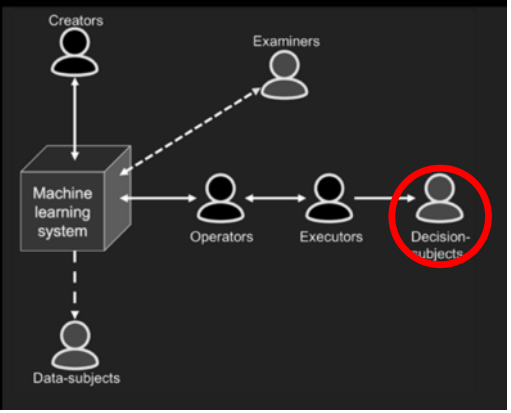
Role-based interpretability: Executors

- (Often the executor is also the operator)
- Responsible for decision-making and need to make “good” decisions
- Relies on understanding of system outputs and the operational context
- Useful techniques may include:
  - Summarization
  - Feature sensitivity
- Goal: Confidence that the right decision is being made
- Requires: Explainability



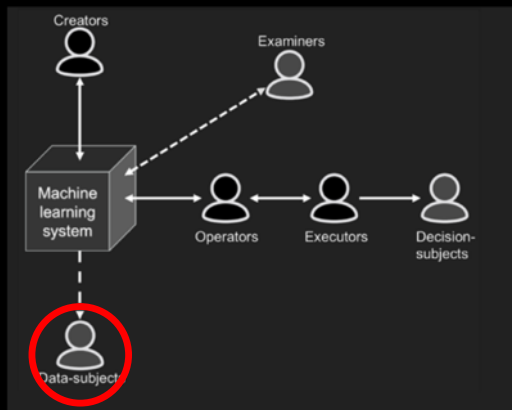
Role-based interpretability: Decision-subjects

- (May often also be the executor)
- Wants to know “why” the decision is being made… on what basis
  - May wish to challenge the decision (contestability)
- Can often be “adversaries” of the executor
  - We cannot assume the decision-subject will not try to “game” the system
- Useful techniques may include:
  - Summarization
  - Other techniques may be suppressed for protection
- Goal: Obtaining the decision they want, or disputing unfavourable decisions
- Requires: Explainability (not transparency)



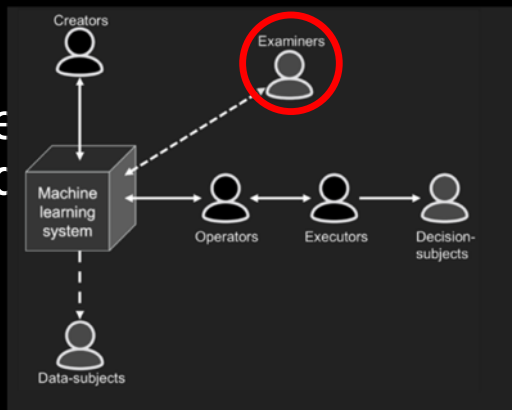
Role-based interpretability: Data-subjects

- Data subjects are not directly involved in the active processes
- They may be unaware their data was used to train the system
- They may have rights regarding their data
  - Including potential right to removal
- Removal of individual data may affect system performance



Role-based interpretability: Examiners

- During development testers examine possible future outputs
- Examiners are different:
  - They audit or investigate specific decisions after deployment
  - They may be external to the organization
- Requirements:
  - That the system at a point in time can be inspected
  - That the recommendations and explanations are the same as they would have been when asked previously
  - System-wide access (with protection for Intellectual Property)
- Examiners may try to improve model performance or mitigate
  may be simply interested in “forensic” examination of a past d



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

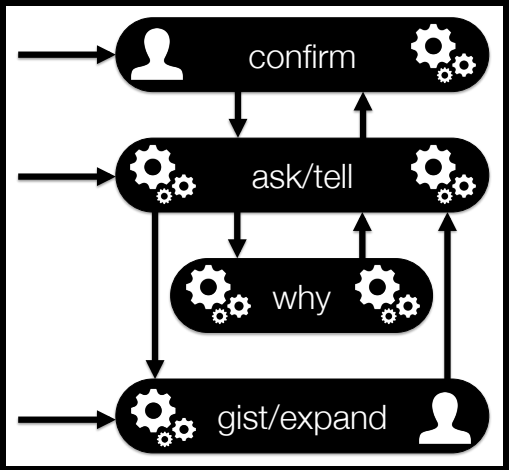
Earlier Research: Conversational Interaction

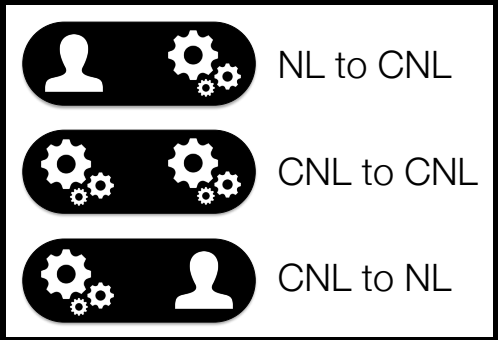
- Talking to machines in natural language is ideal but hard
- Controlled Natural Language as a compromise: “easy to read, harder to write”
- Let’s bring the two together:
  - Human users write NL sentences [easy to write]
  - Machine users convert to NL [easy to process]
  - Machine users respond in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

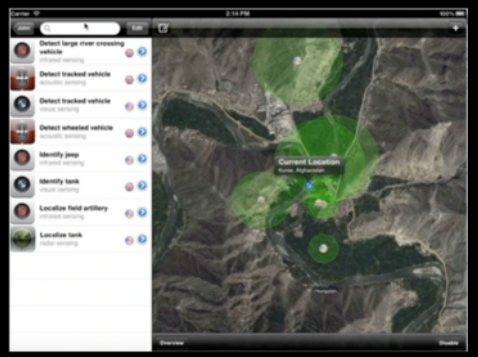


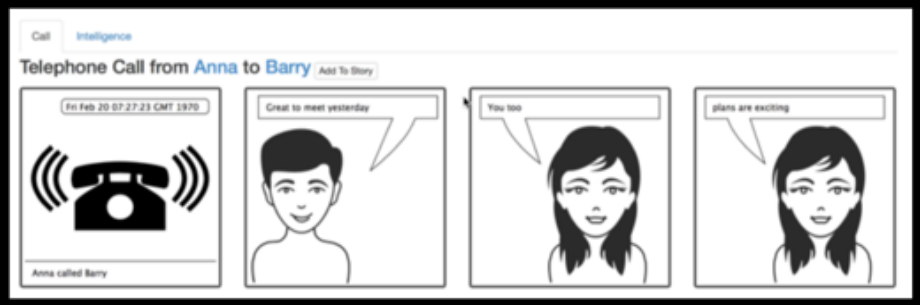


Draws on research in agent communication languages and philosophical linguistics (speech acts)

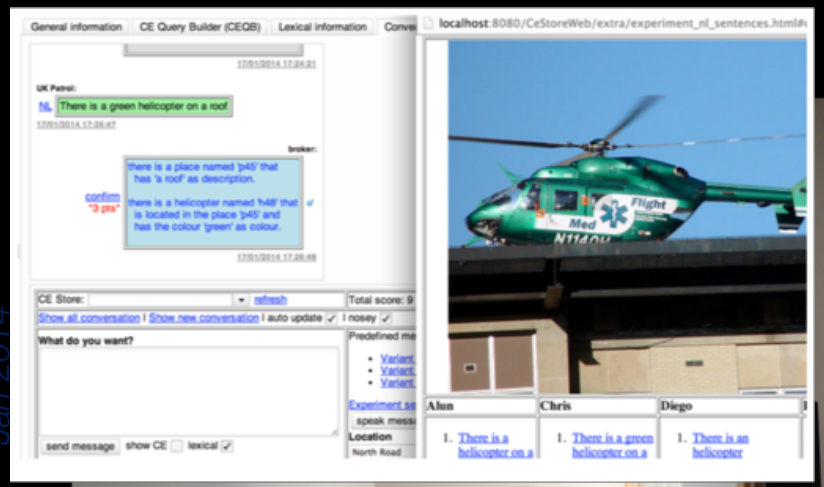
We carried out evaluations

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

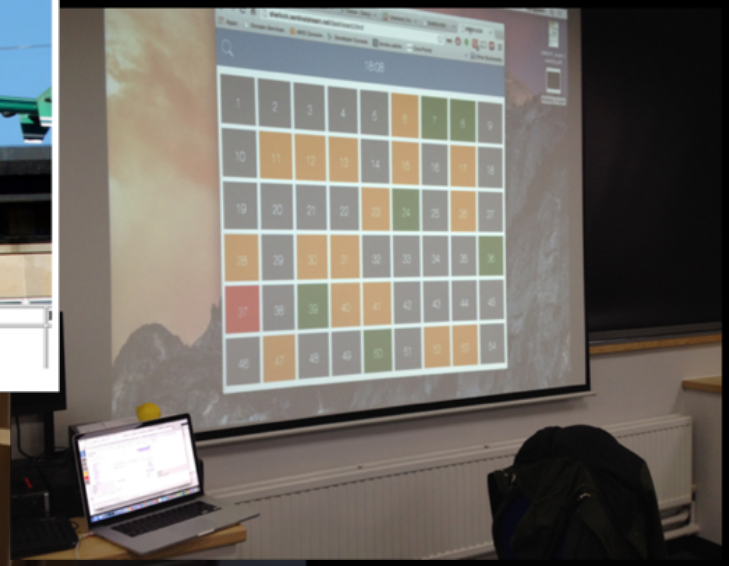




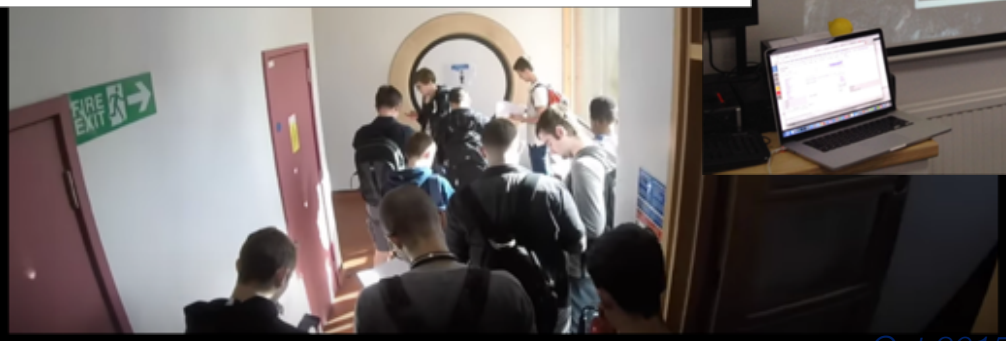
We analyzed student experiments



Jan 2014



Dec 2014



Oct 2015

...and worked with practitioners

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

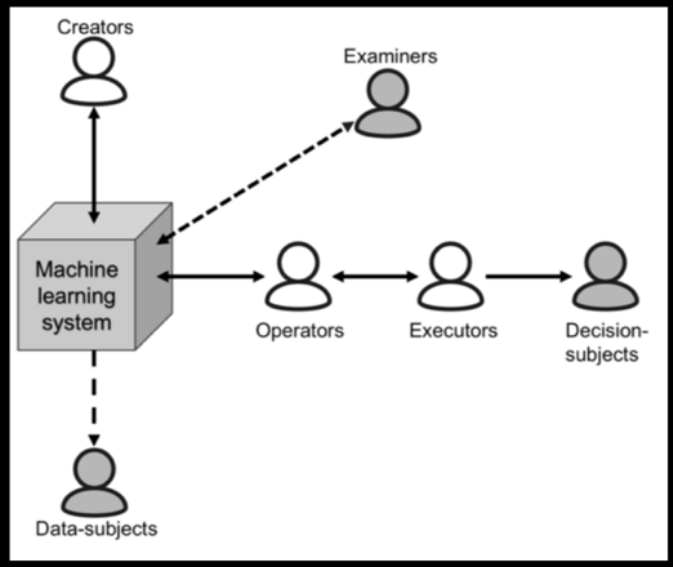
Conversational Explanations

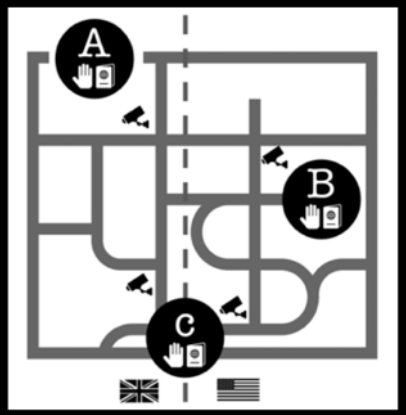
Scenario and dataset

- 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

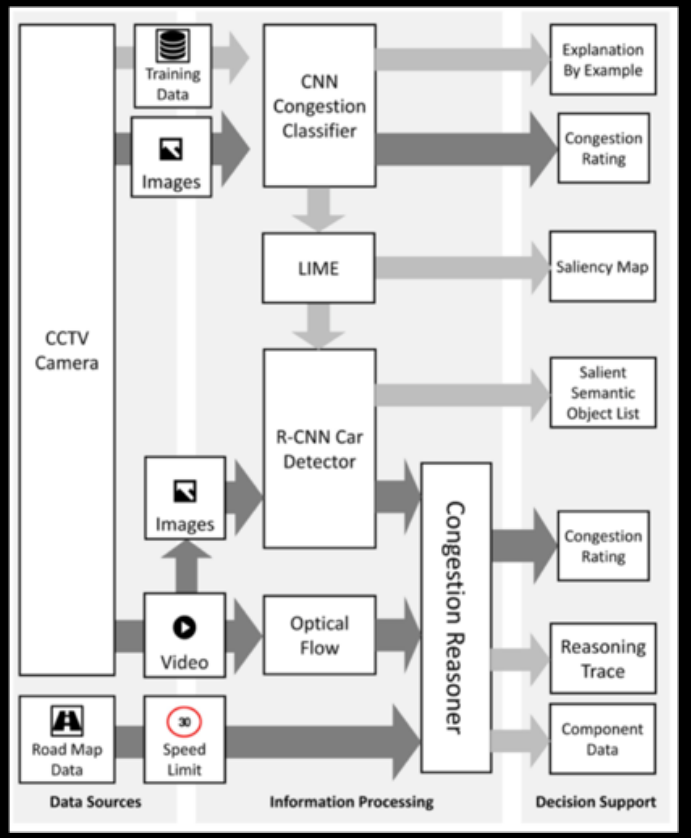
- We treat explanation as a conversation
- User role and task context are key

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:



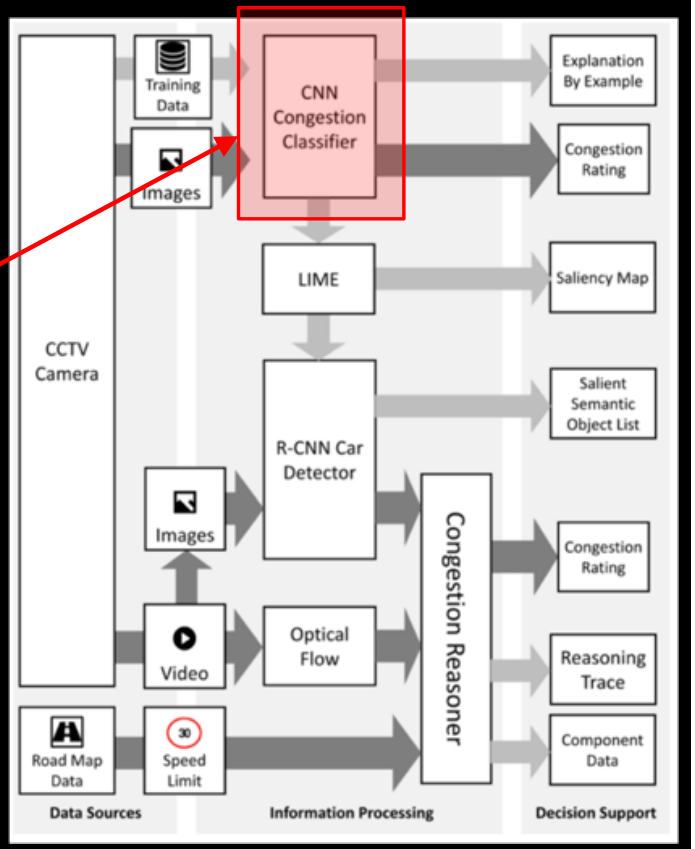
Worked Example

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)



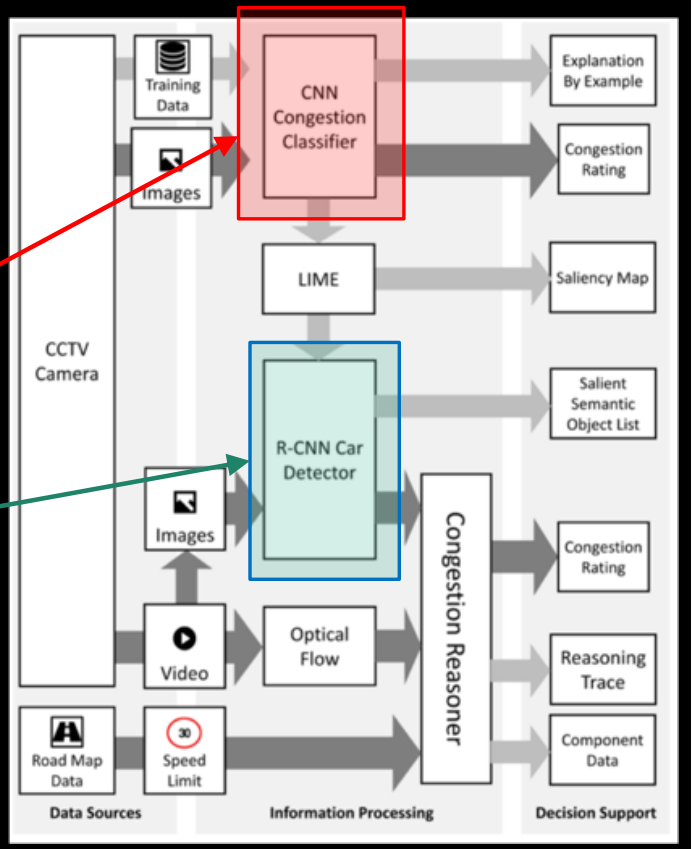
Worked Example

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)



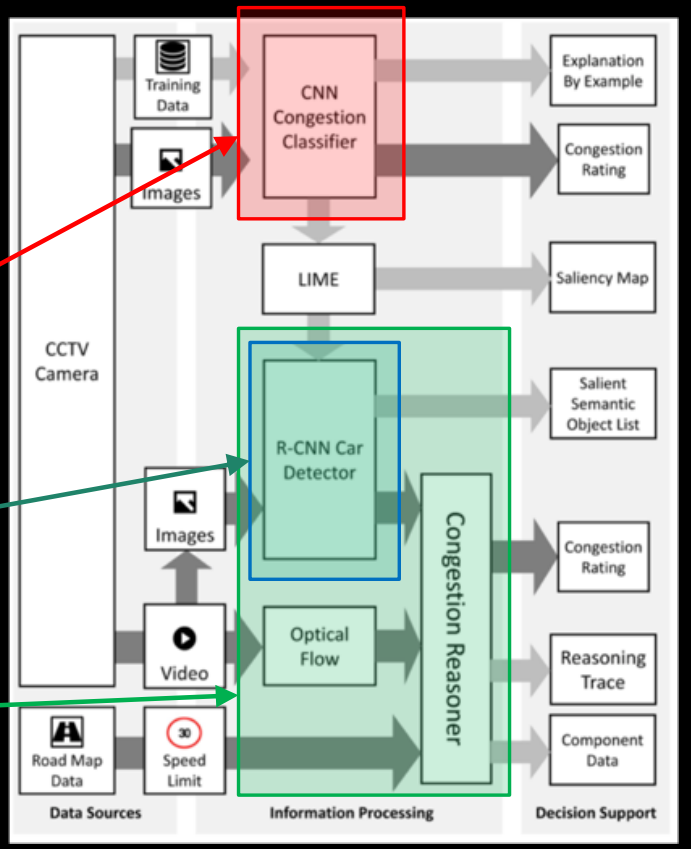
Worked Example

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)



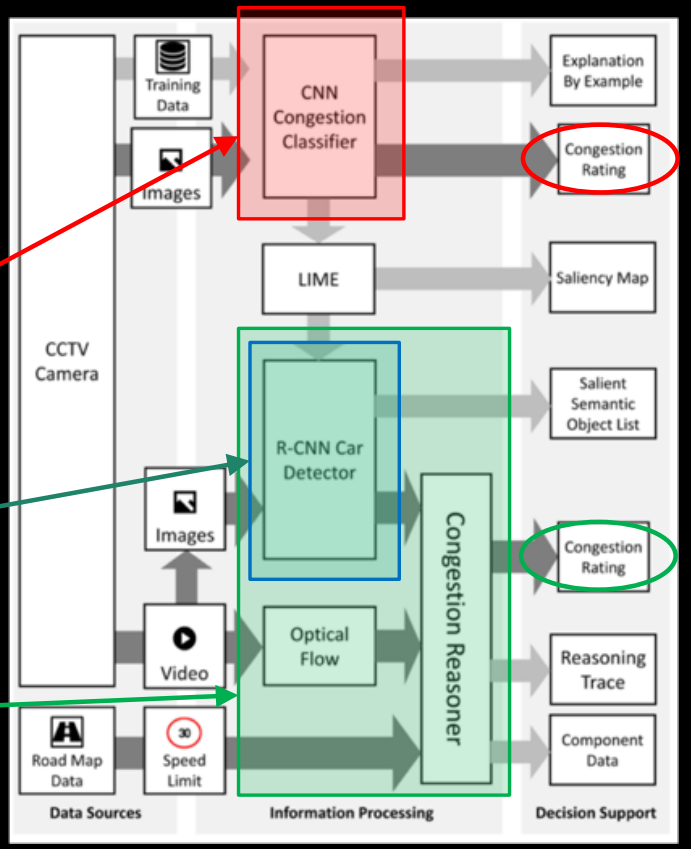
Worked Example

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:

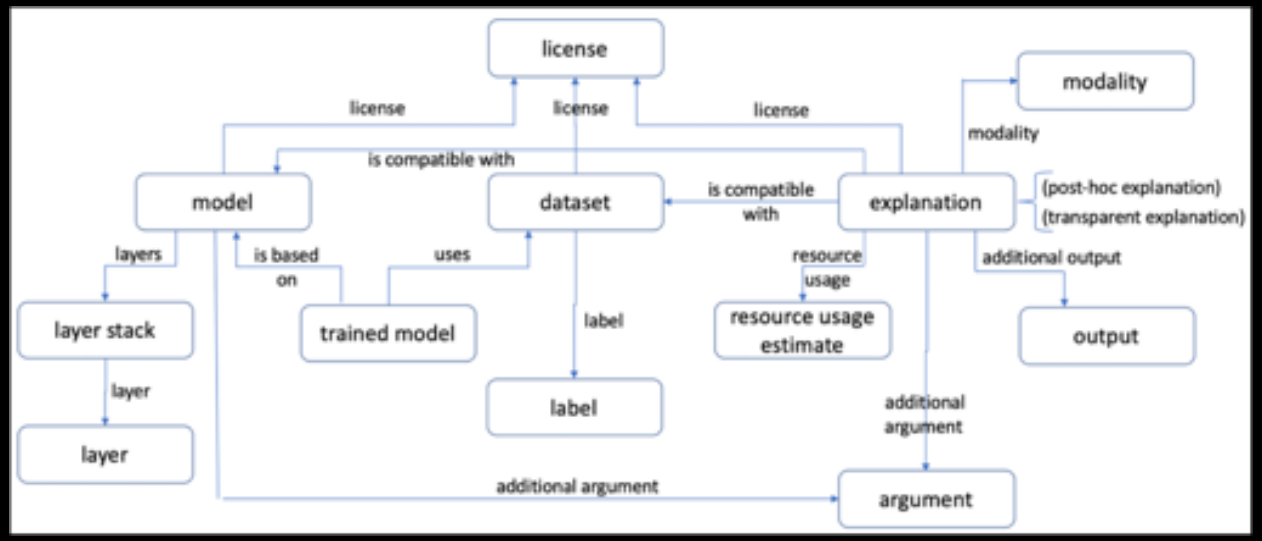
1. Congestion Image Classifier (CIC)
2. Entity detector (ED)
3. Congestion Speed Classifier (CSC)



Conversations for Explanation

Explanation takes the form of a conversation

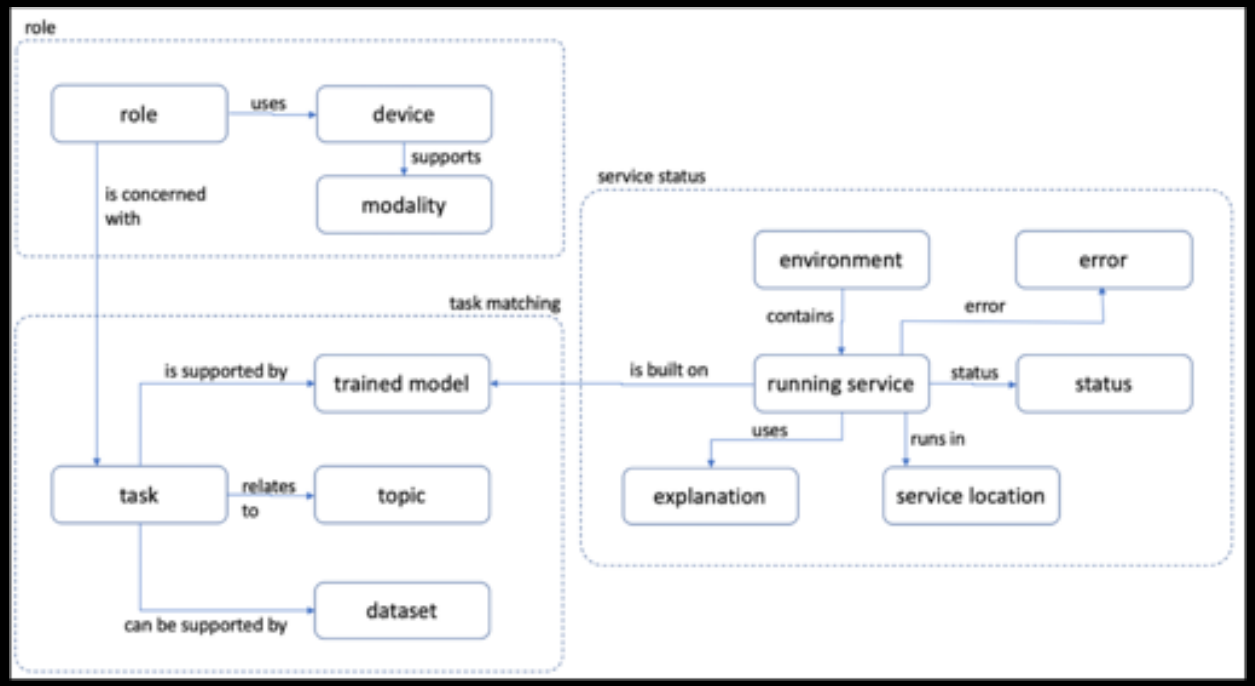
- Can be in many forms
- Focus on the abstract space
- Text-chat format
- Embedded additional modalities



Conceptual model defines key concepts and relationships

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

Conversations for Explanation

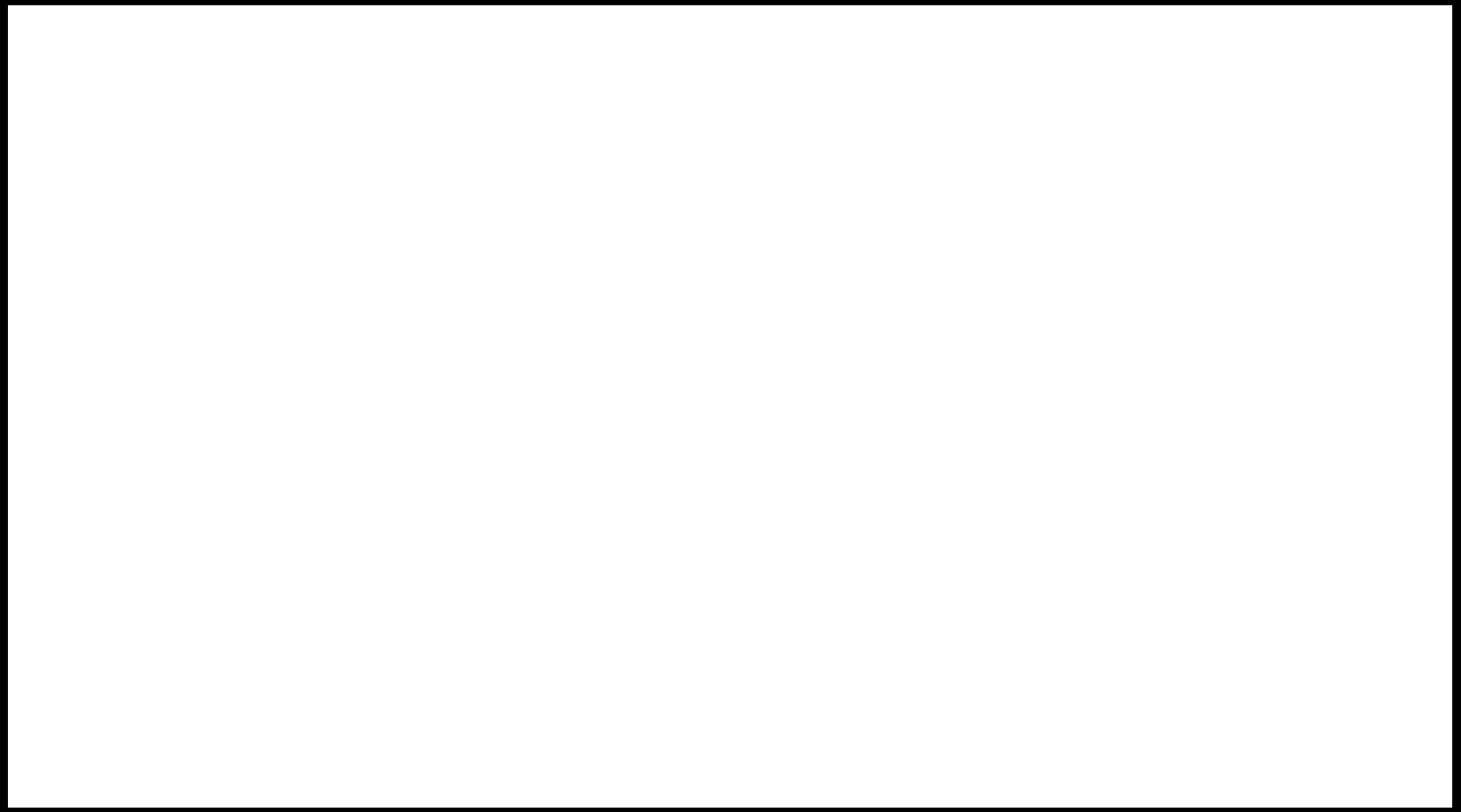


Additional models easily integrated

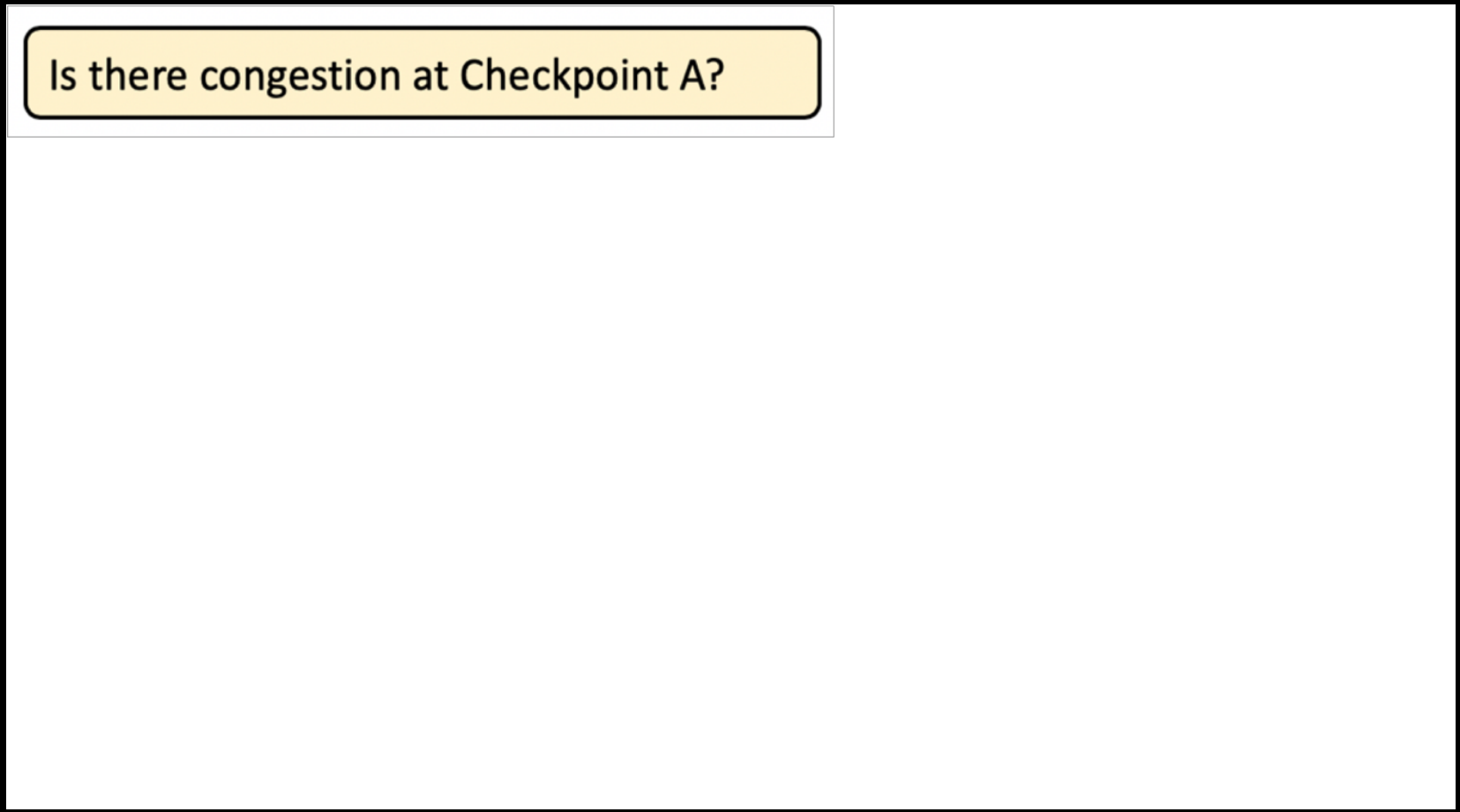
- To define the context for the explanations
- Role, task, services, etc

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

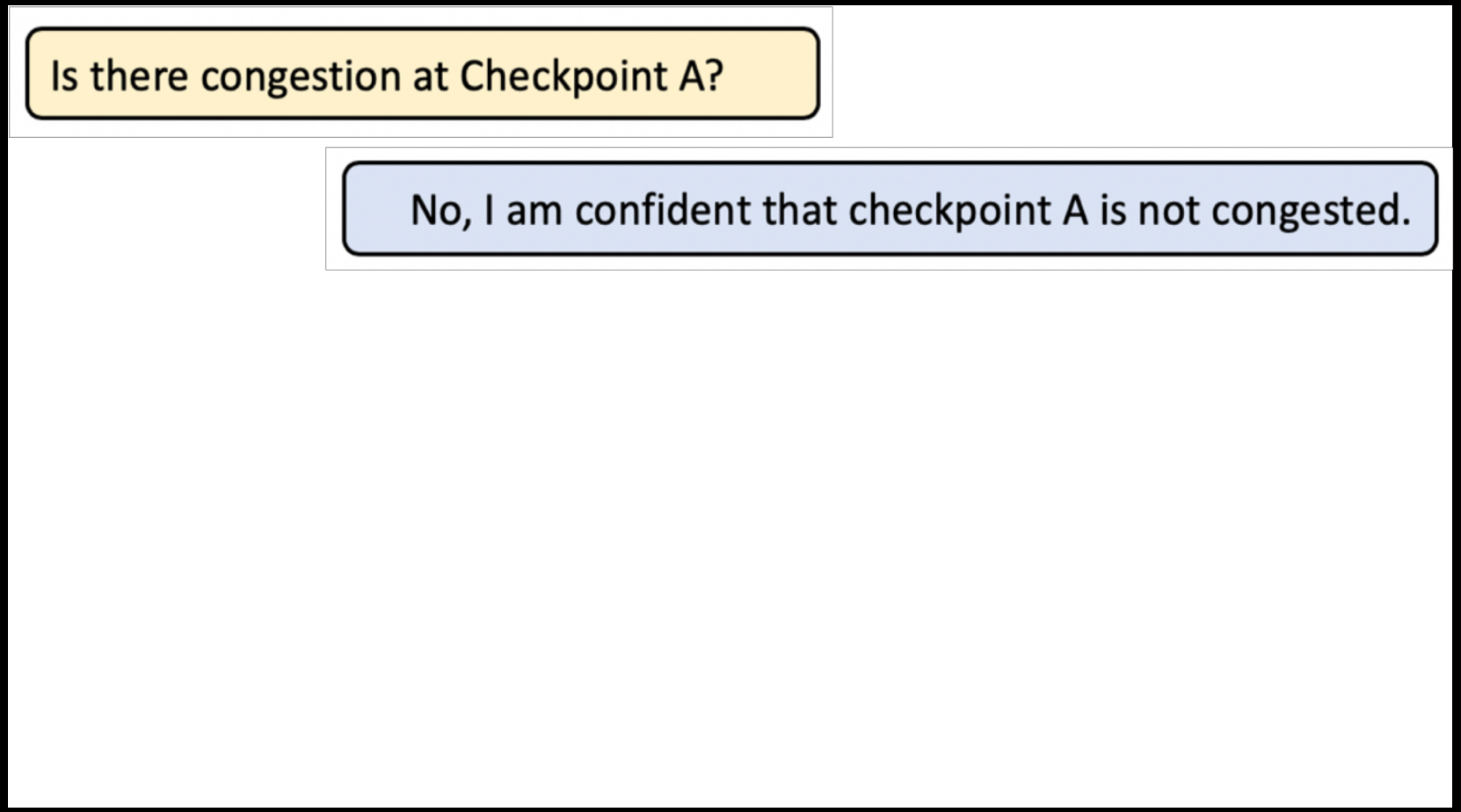
Conversations for Explanation
Use case 1: Fully transparent



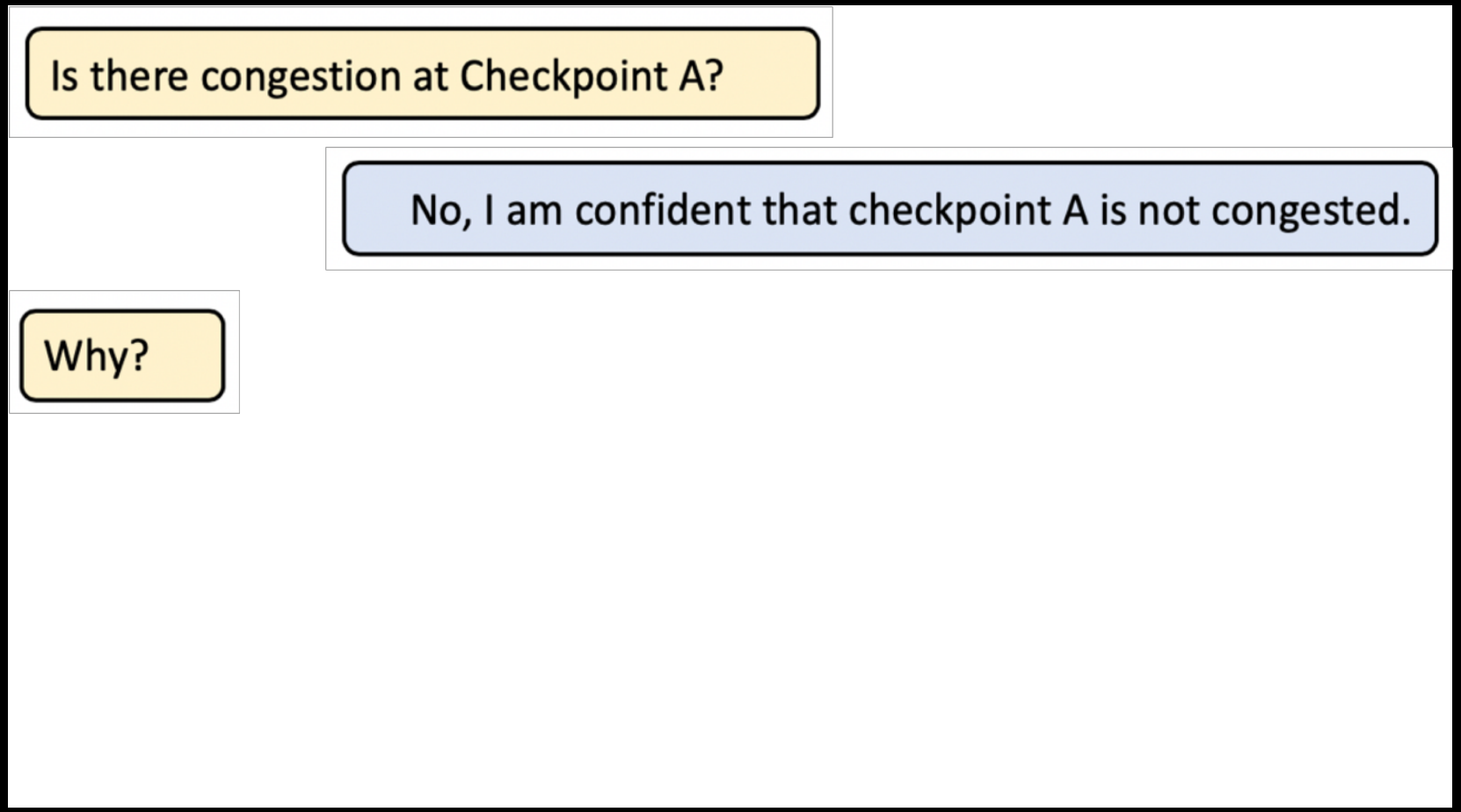
Conversations for Explanation
Use case 1: Fully transparent



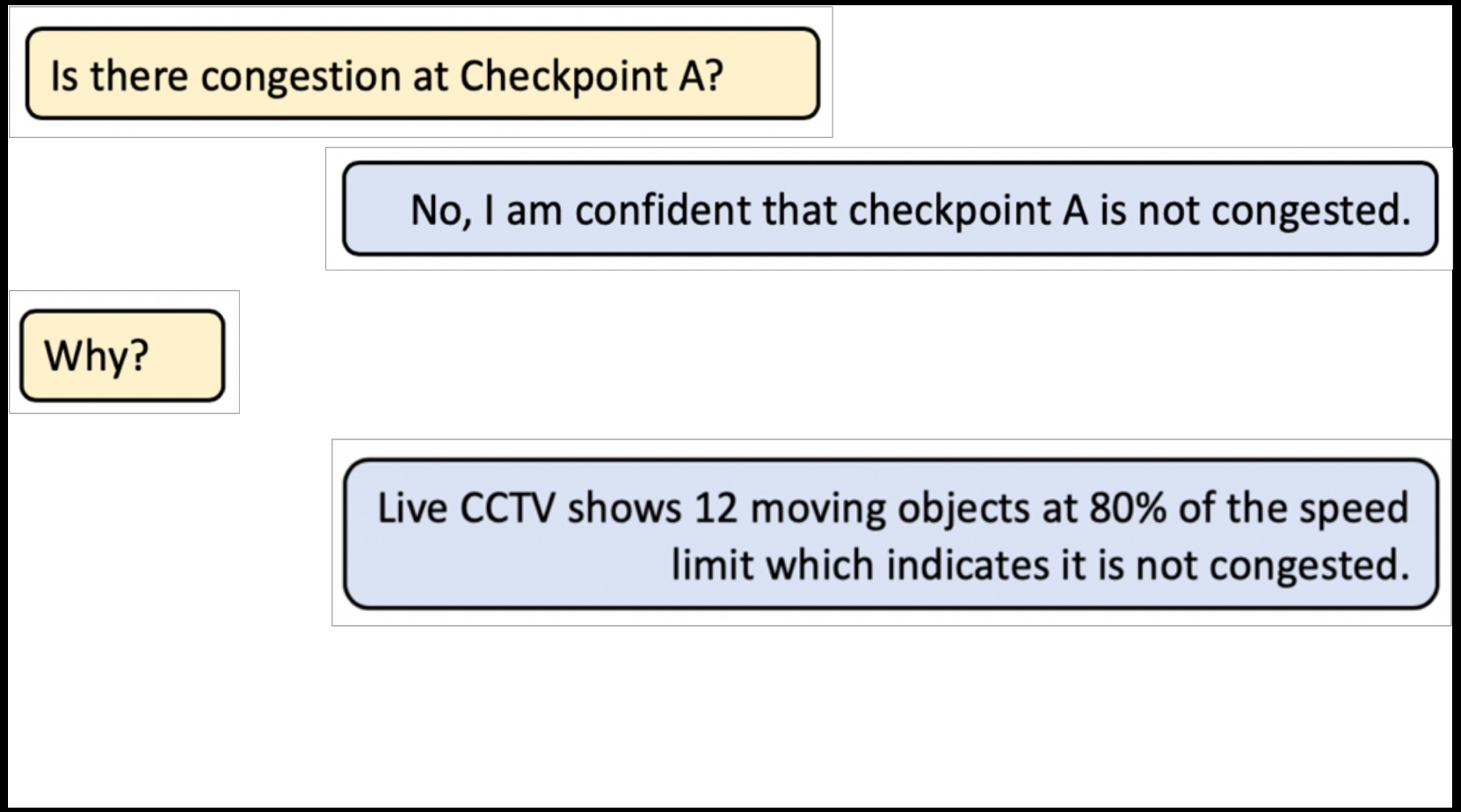
Conversations for Explanation
Use case 1: Fully transparent



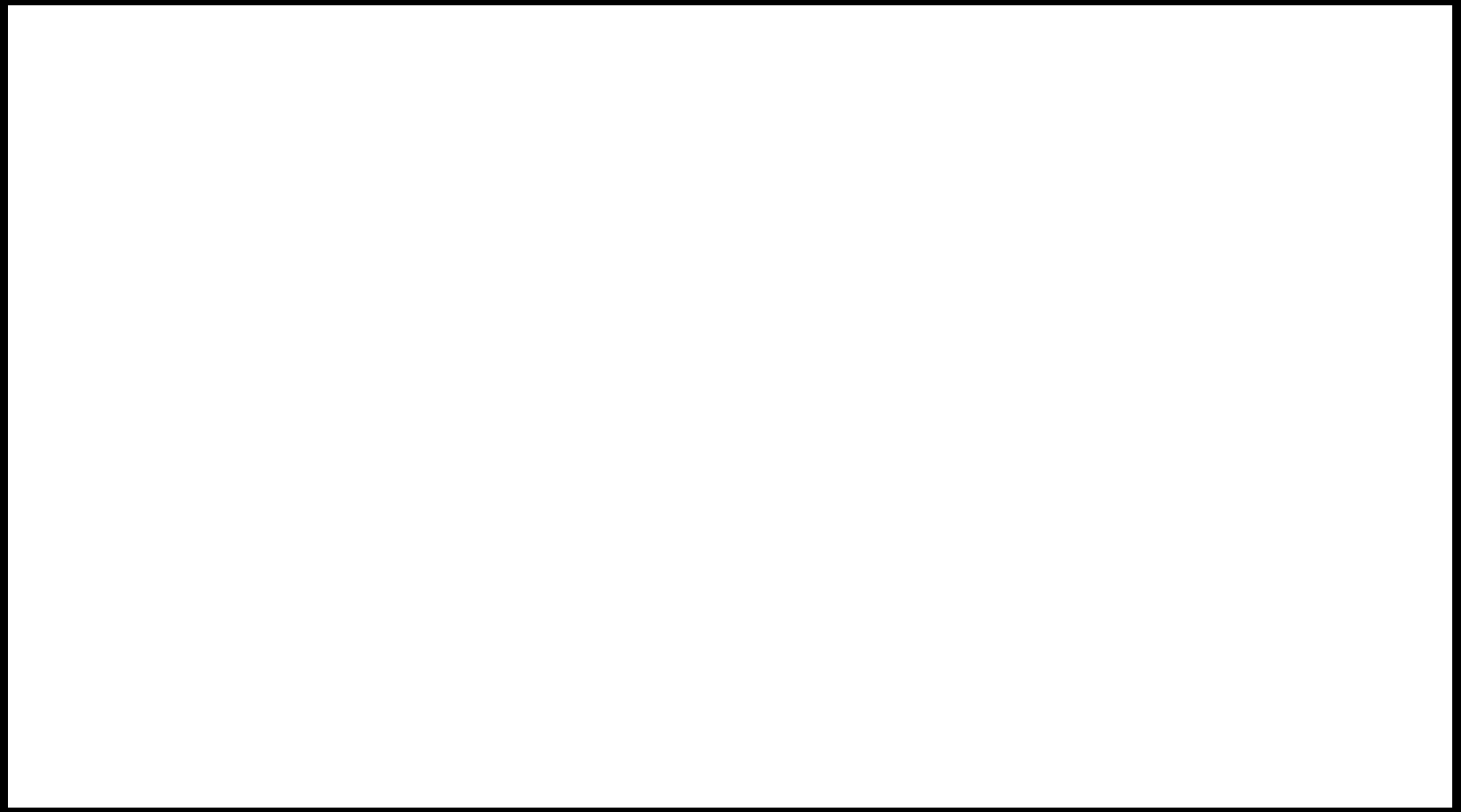
Conversations for Explanation
Use case 1: Fully transparent



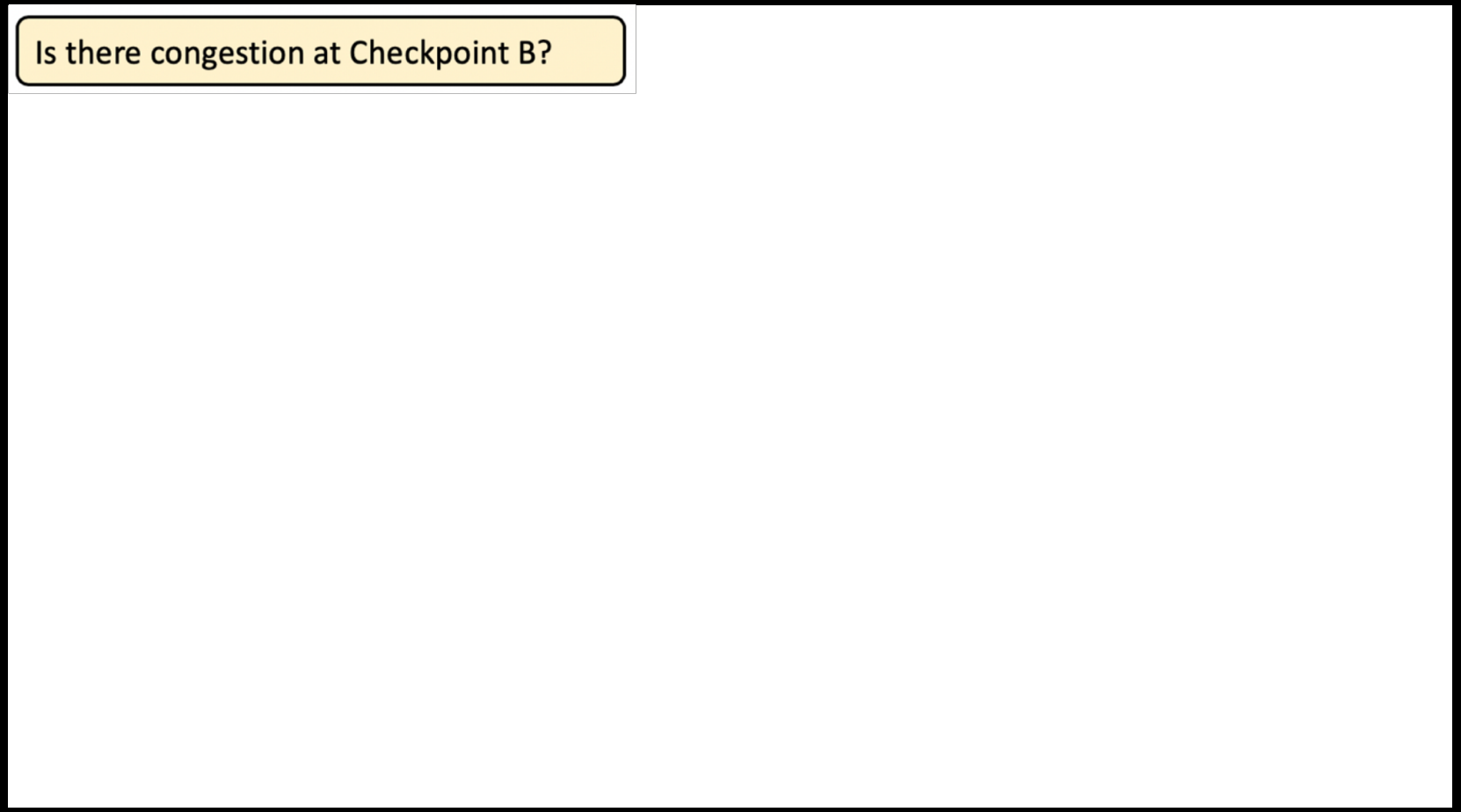
Conversations for Explanation
Use case 1: Fully transparent



Conversations for Explanation
Use case 2a: Post-hoc by Saliency



Conversations for Explanation
Use case 2a: Post-hoc by Saliency

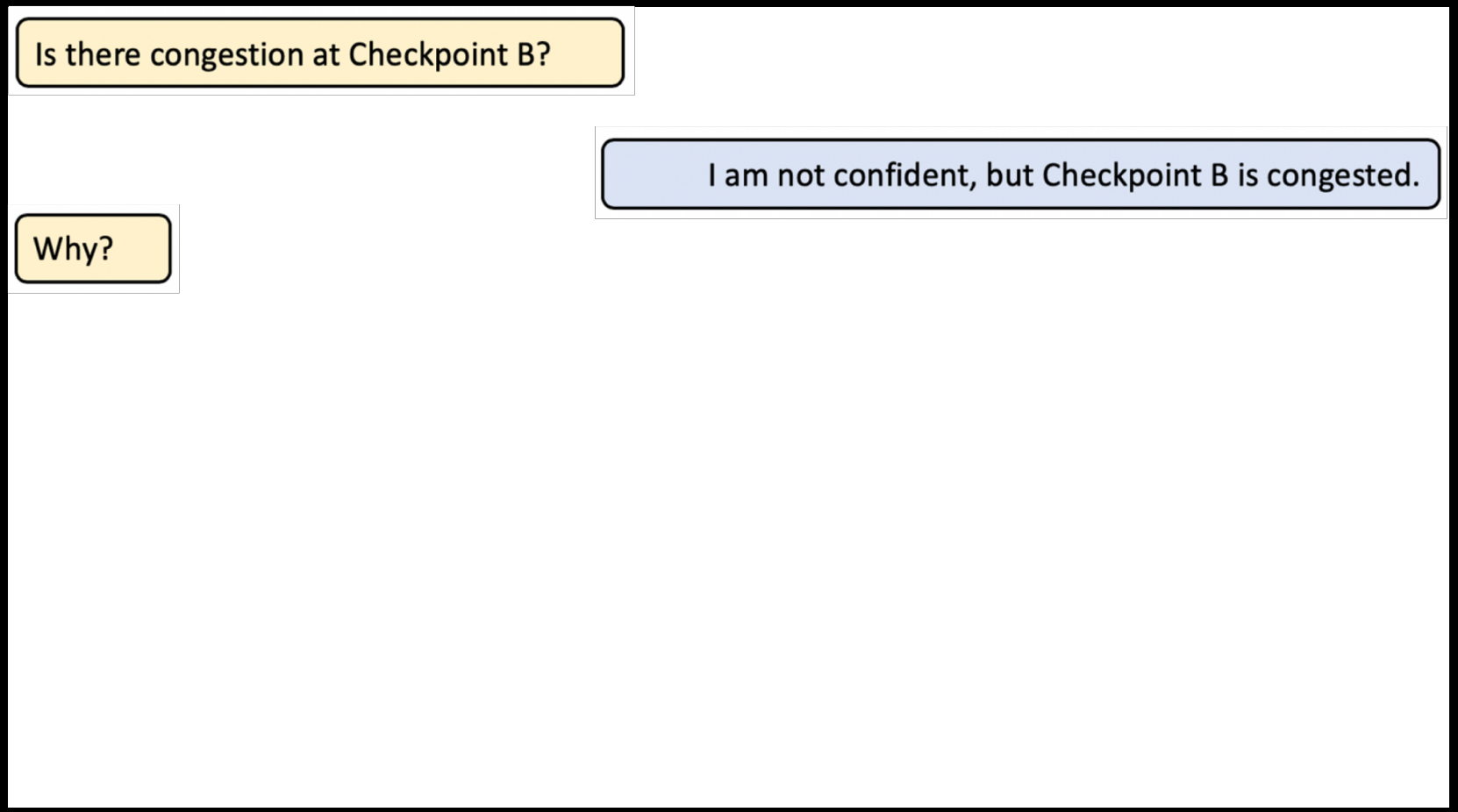


Conversations for Explanation
Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?

I am not confident, but Checkpoint B is congested.

Conversations for Explanation
Use case 2a: Post-hoc by Saliency



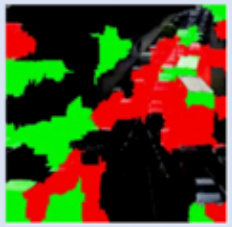
Conversations for Explanation
Use case 2a: Post-hoc by Saliency

Is there congestion at Checkpoint B?

I am not confident, but Checkpoint B is congested.

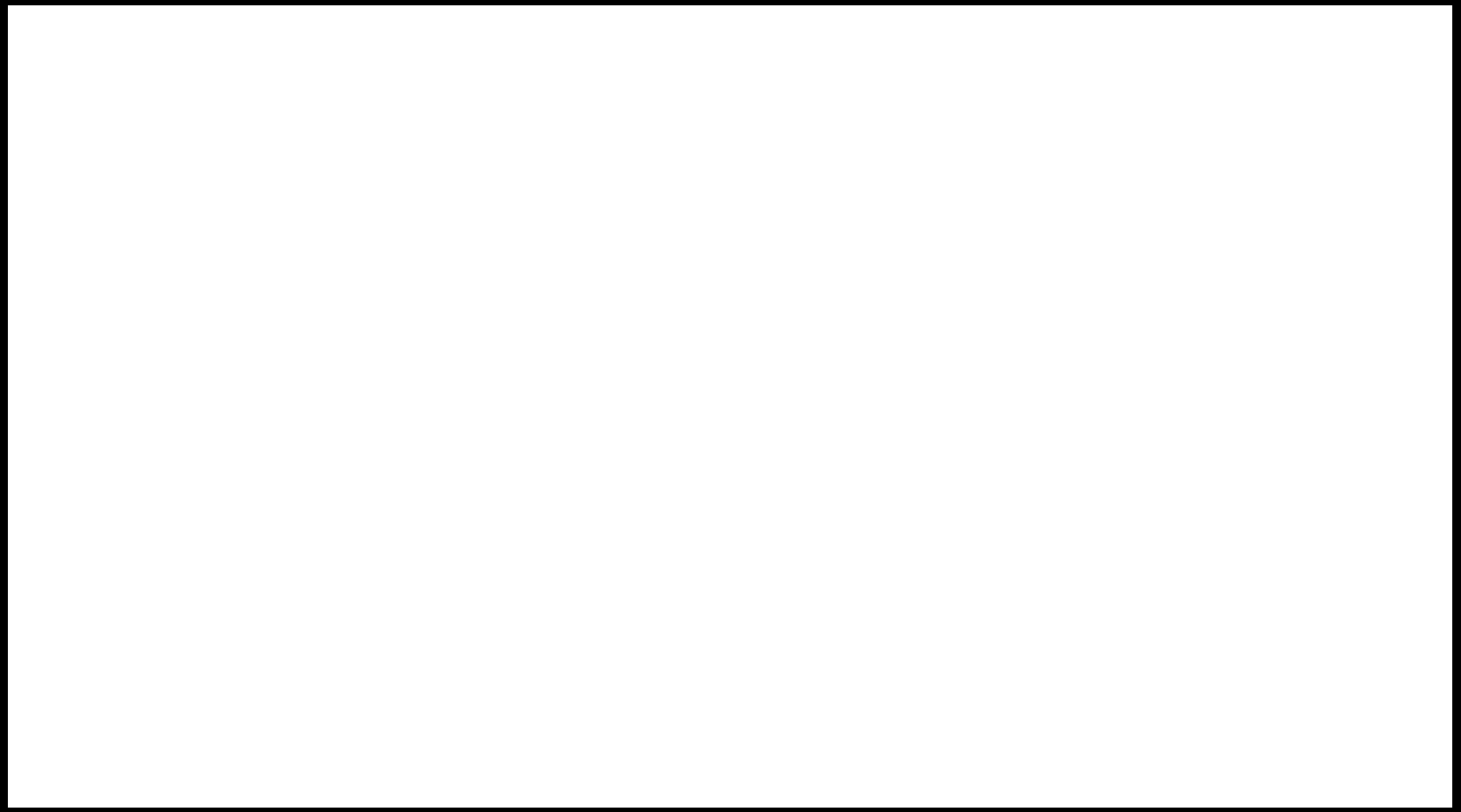
Why?

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

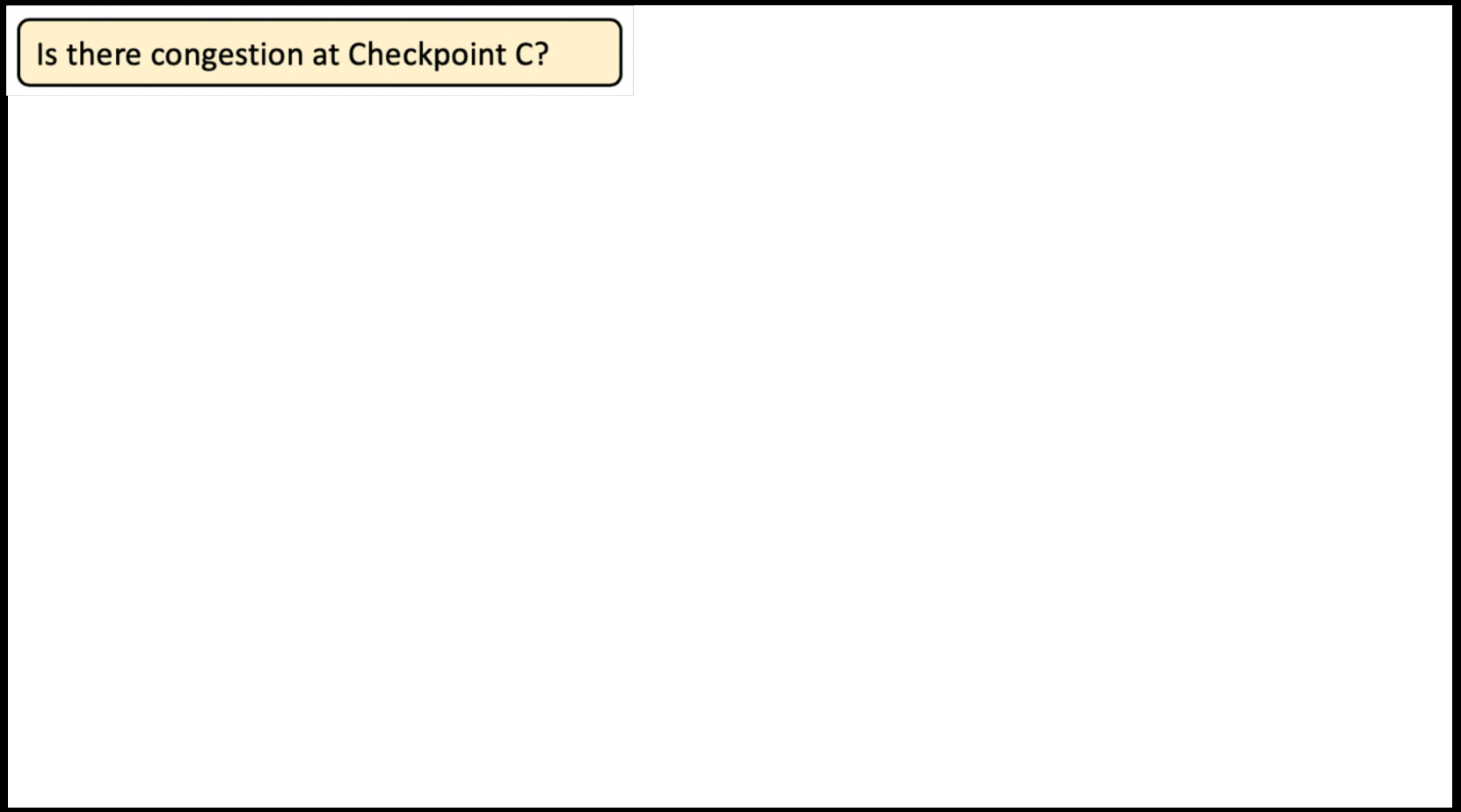


The green areas show the parts of the image that most indicate congestion.

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



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



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

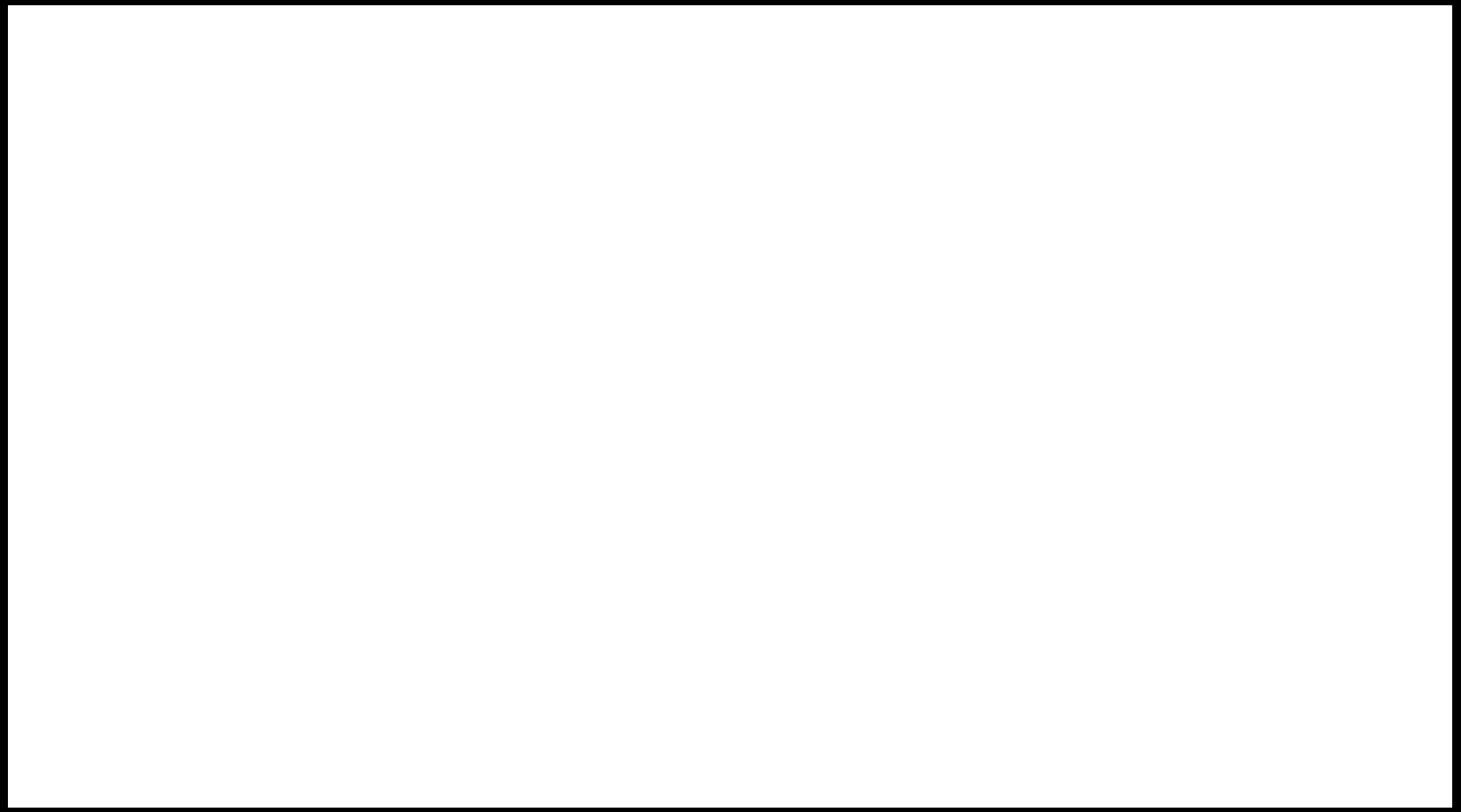
Is there congestion at Checkpoint C?

I am not confident, but Checkpoint C is congested.

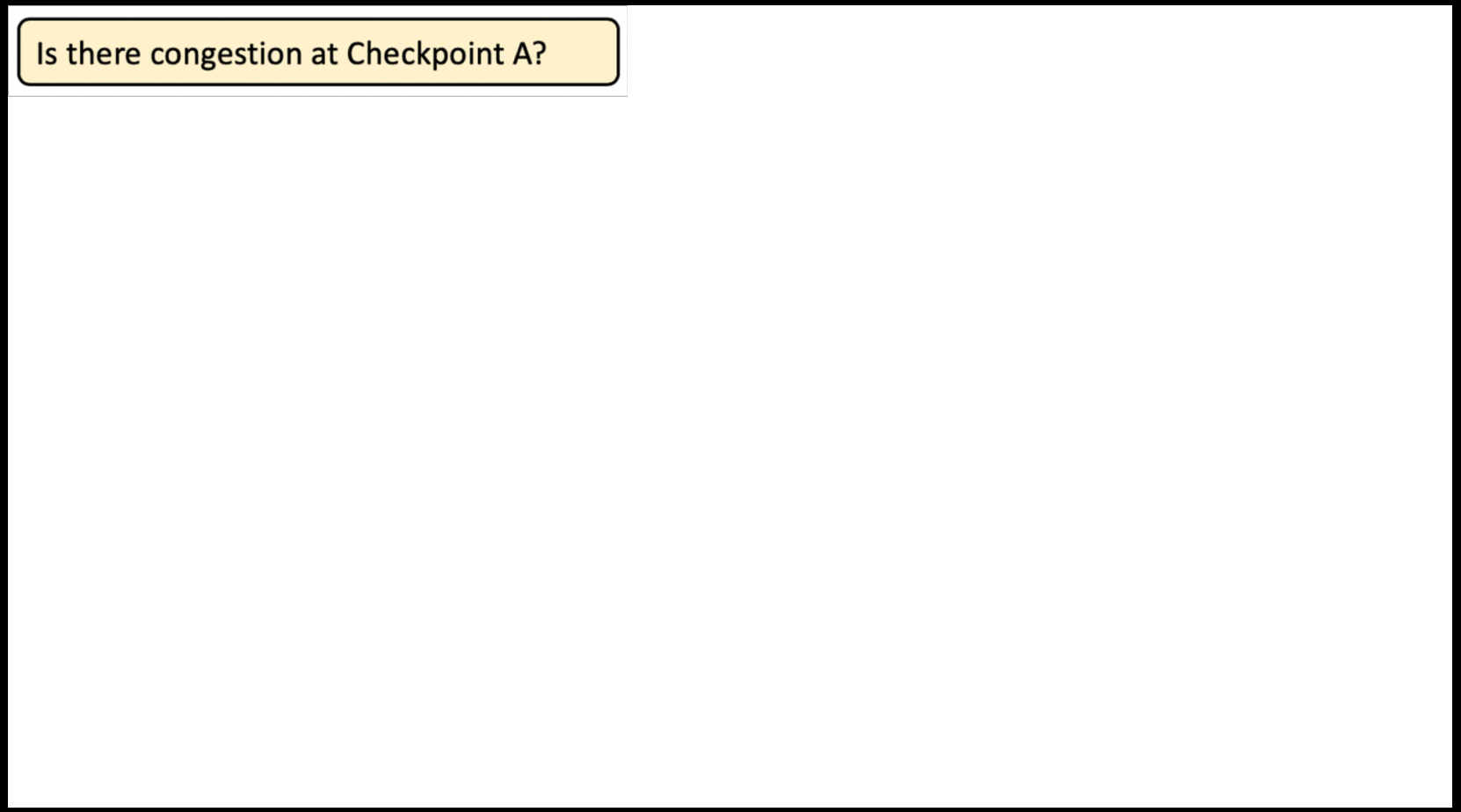
Why?

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

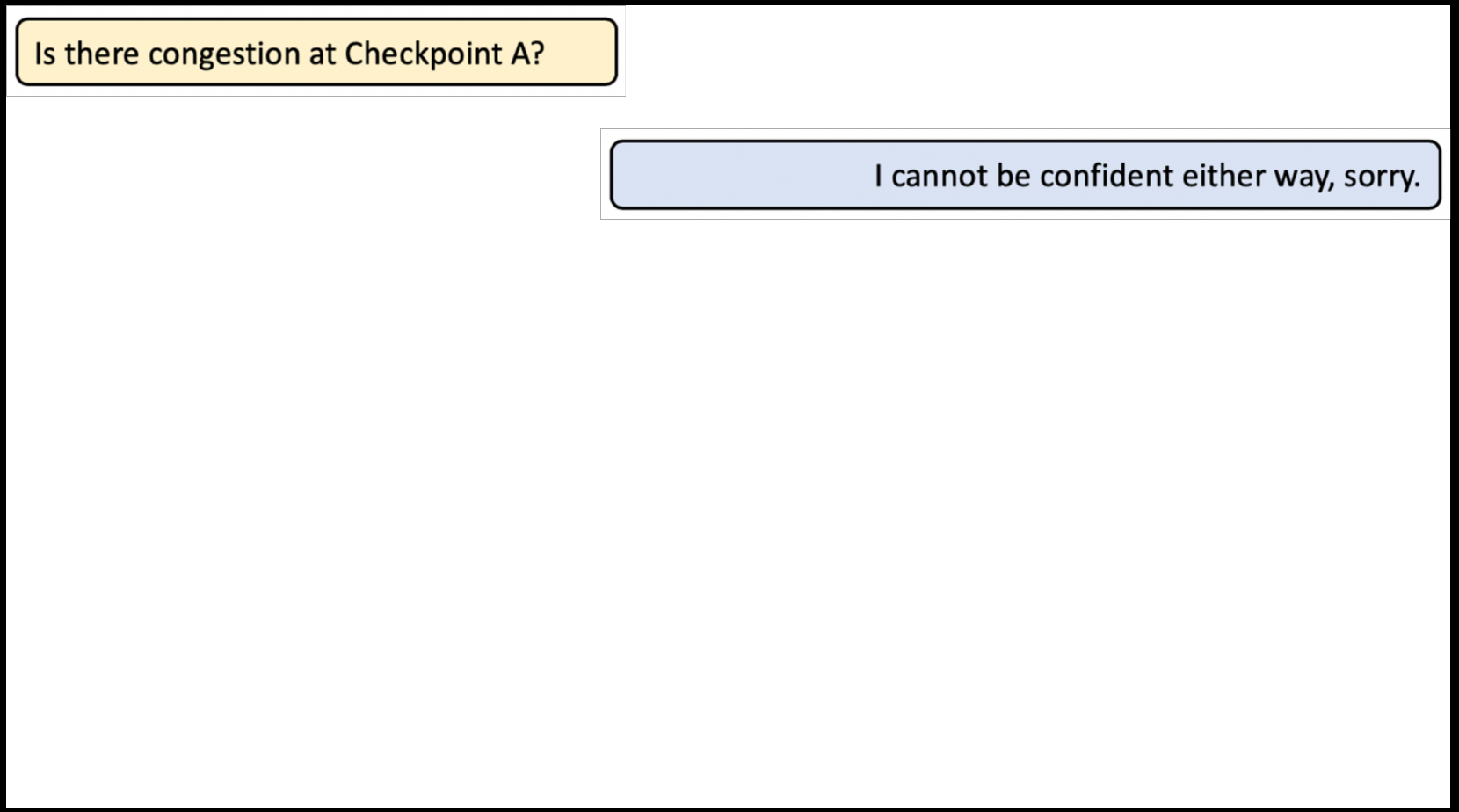
Conversations for Explanation
Use case 3: Disagreement within services



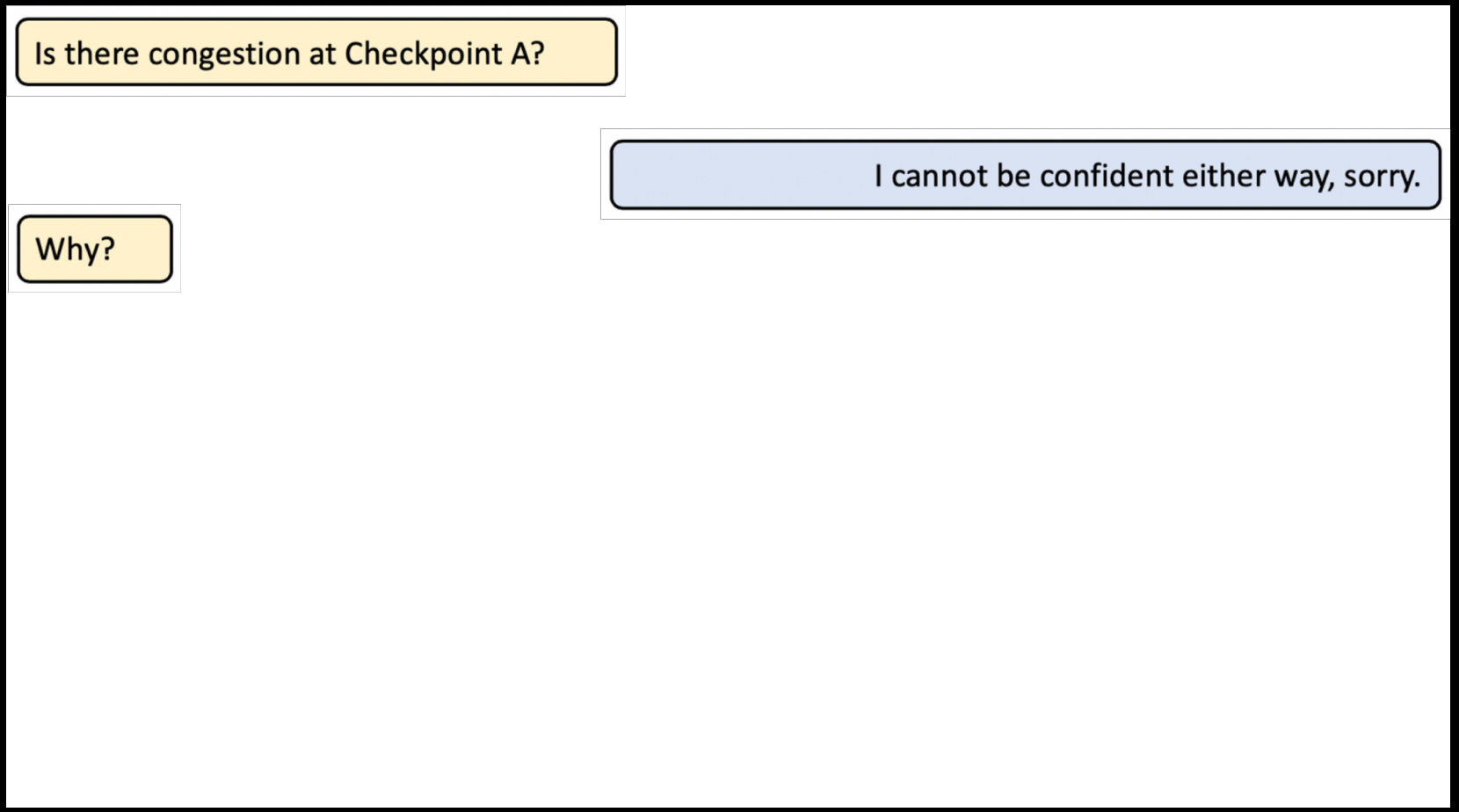
Conversations for Explanation
Use case 3: Disagreement within services



Conversations for Explanation
Use case 3: Disagreement within services



Conversations for Explanation
Use case 3: Disagreement within services



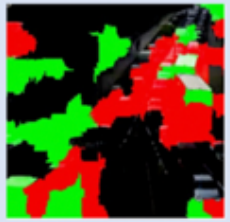
Conversations for Explanation
Use case 3: Disagreement within services

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.

Current plans for this work

- **Conversational interface**
  - To enable exploration of explanations
  - Substantially upgrade the explanation meta-model (accounting for Molnar and Miller)
- **Explanations for AI services**
  - Multi-modal data and explanations
  - Rapidly assembled ensembles
  - Coalition context
- **Improved conceptual model to underpin our approach**
- **Real-world scenario with 3 examples**
- **More interaction with Subject Matter Experts**
- **Experimental design and execution**

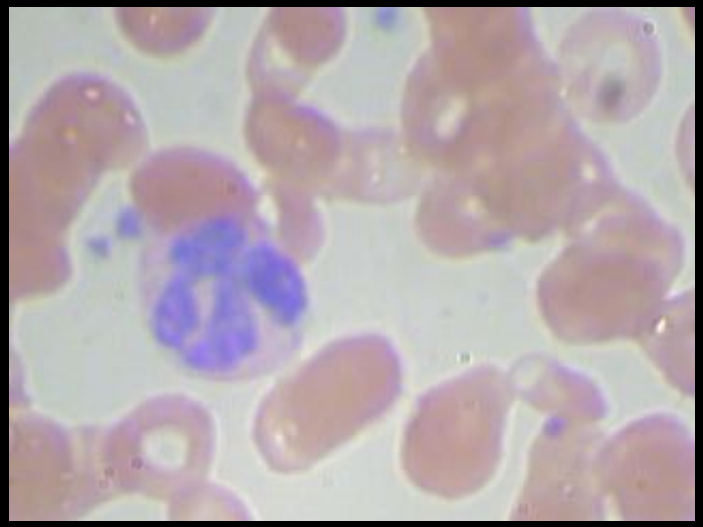
Visualization of Deep Learning

The Need for Explanation

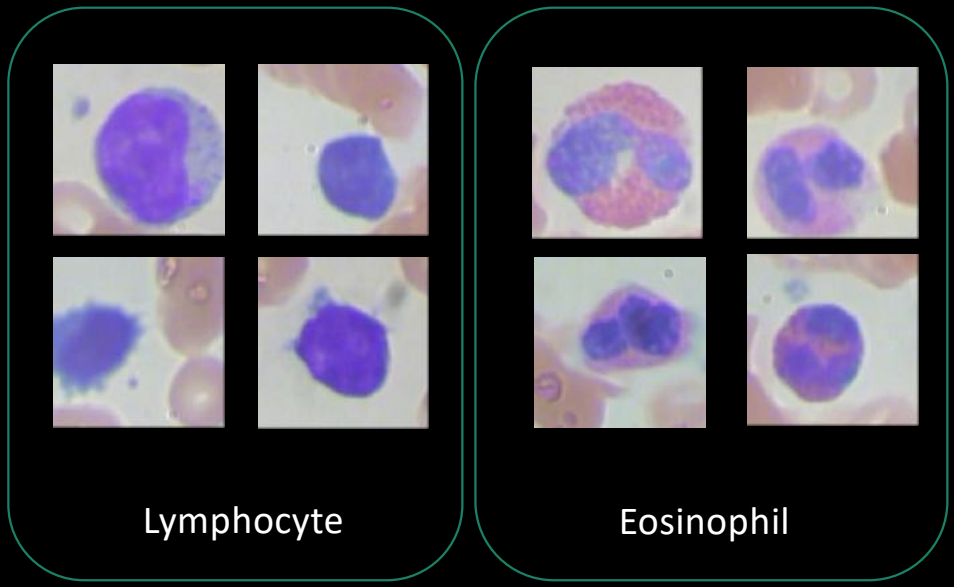
- **Deep Learning** networks can be effective in classification tasks
- However people need to understand the reasons for a particular conclusion
- We aim to build systems that can offer explanations

From "Going Deeper with Convolutions", Szegedy et al, https://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf

A typical task: classifying cells



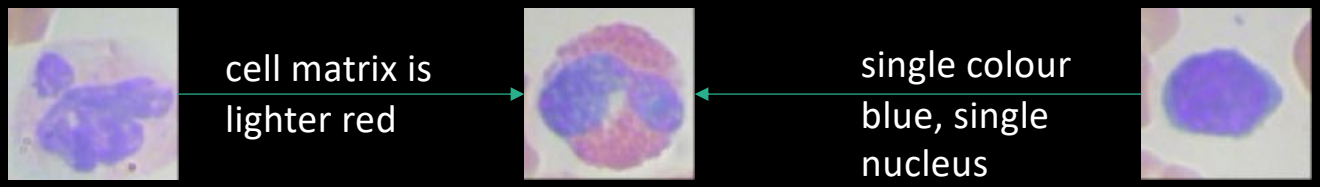
Segment and classify



KAGGLE dataset

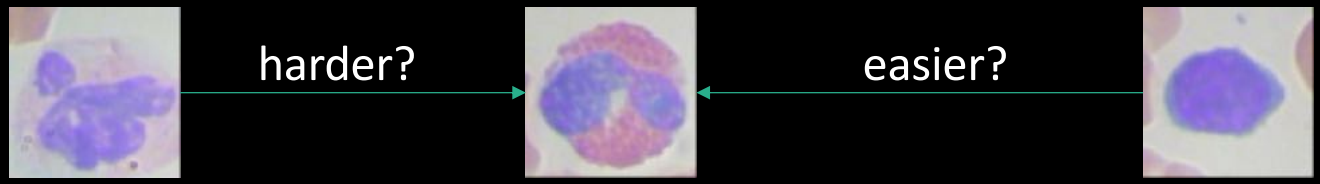
Several blood cell classification tasks

- Classification task LE is lymphocyte vs eosinophil
- Classification Task EN is eosinophil vs neutrophil
- By eye, LE seems an easier discrimination task than EN



First attempt

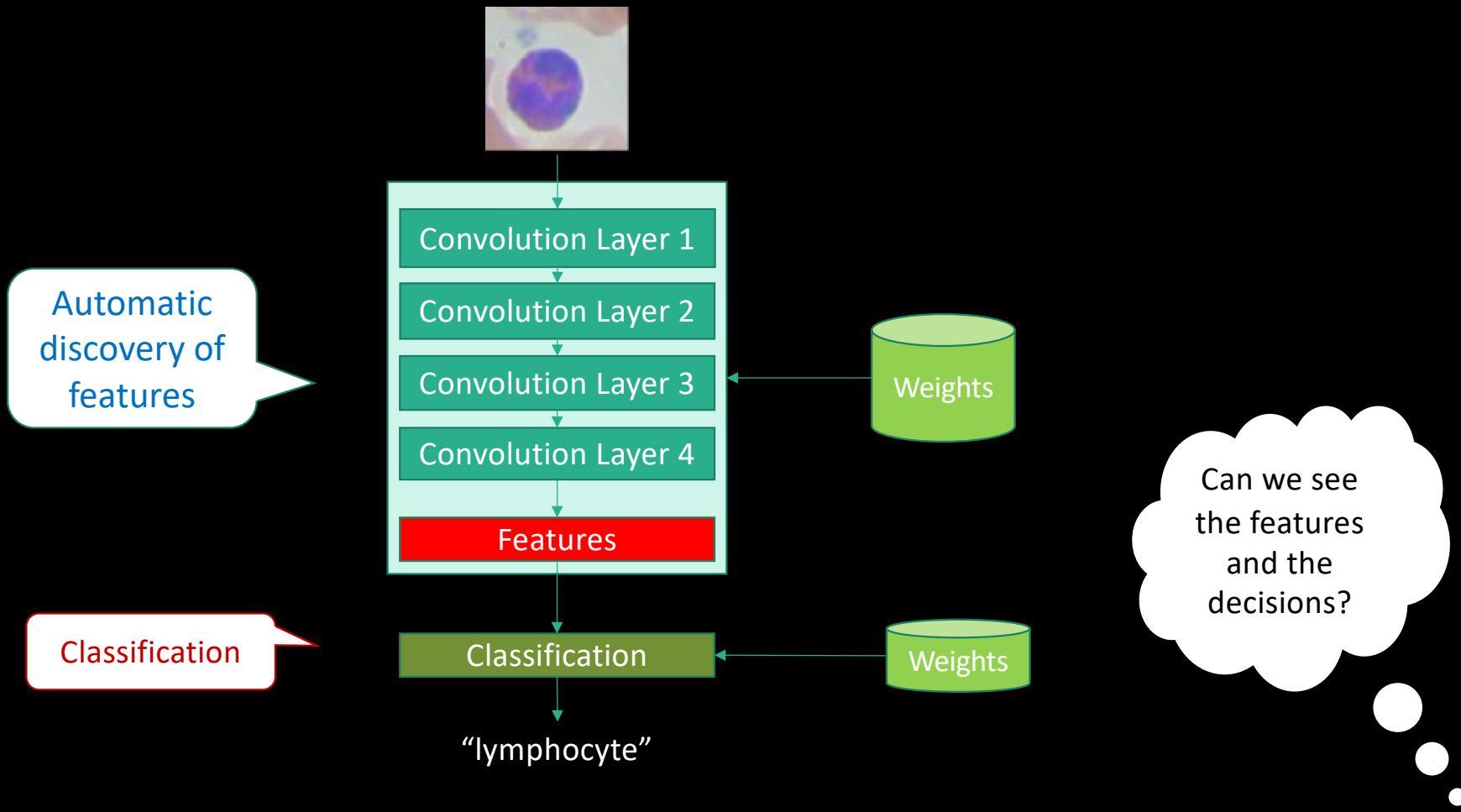
- Small Kaggle Dataset of 206 neutrophil, 88 eosinophil, 33 lymphocyte
- 4 layer Convolution Neural Network



neutrophil EN eosinophil LE lymphocyte

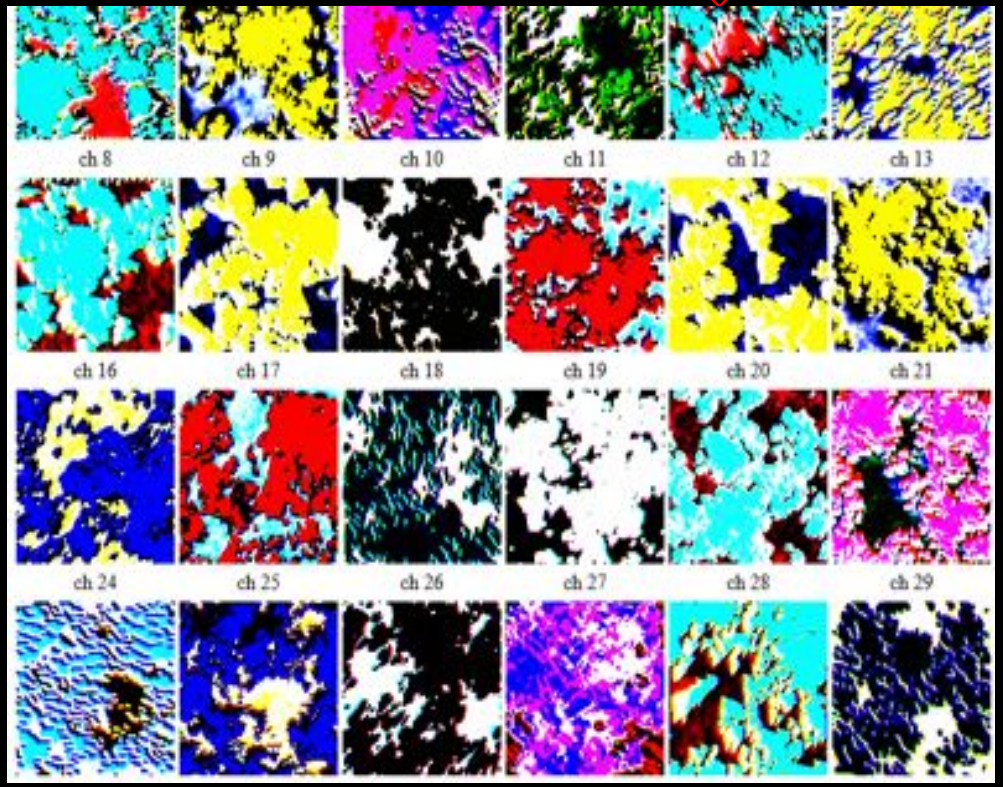
0.66 accuracy (“not good”) 0.875 accuracy (“fairly good”)

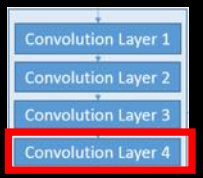
The Deep Learning Architecture



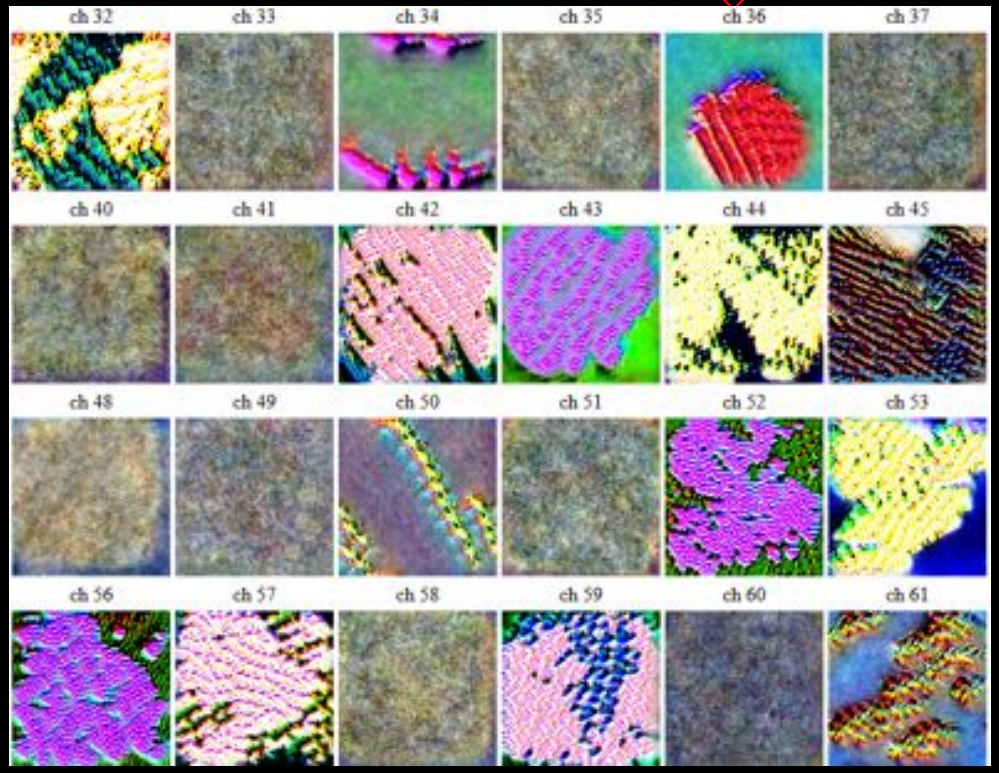
Visualising features

Edges, textures





More object-like



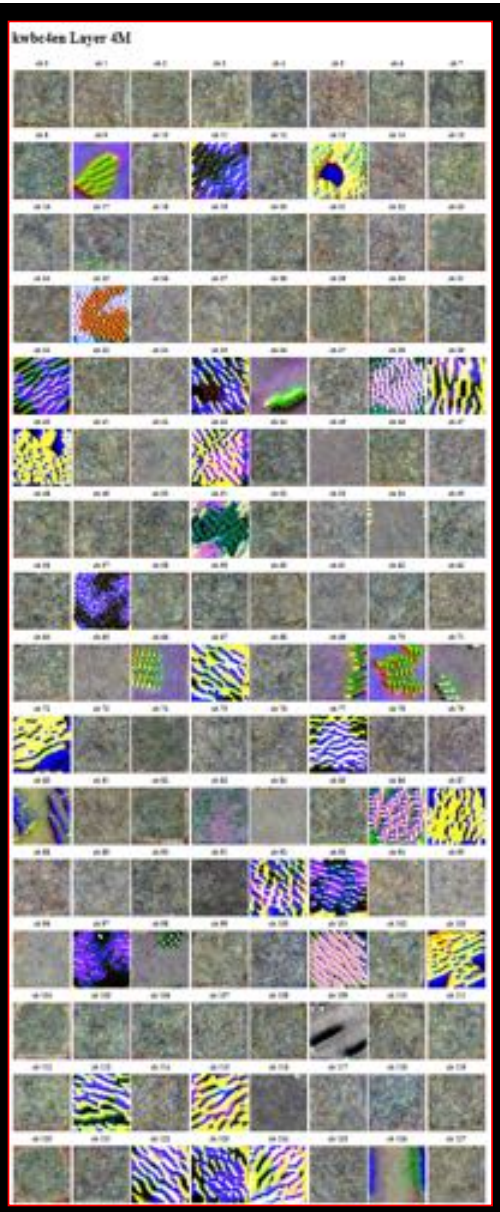
Comparing 4th layer features of both models

Noticeably different, by eye

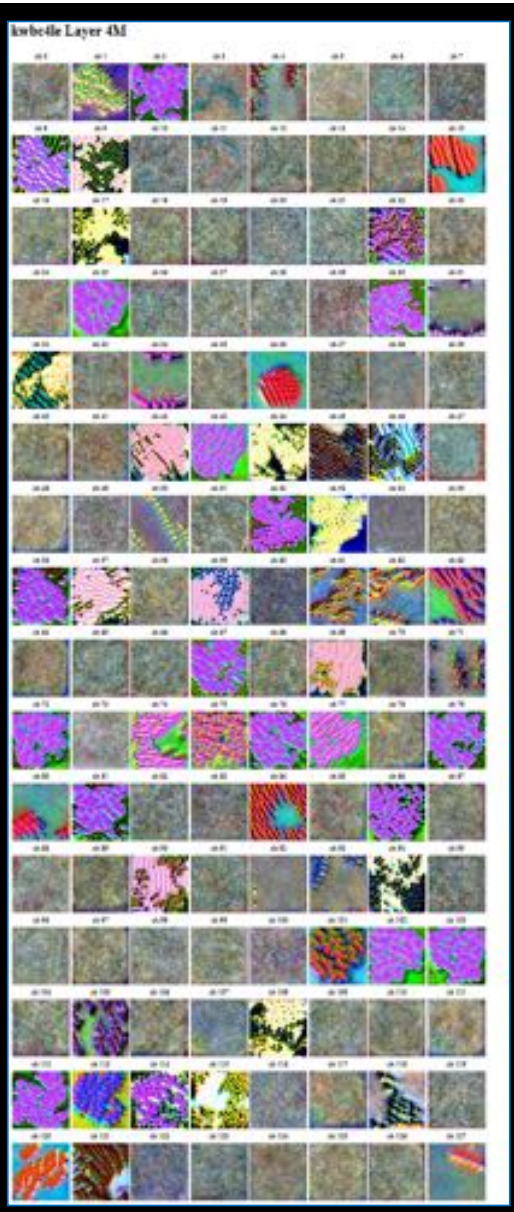
- EN less features (it's harder?) and mostly yellow-blue
- LE more blue-purple (the blue lymphocytes?)

Can we show the differences objectively?

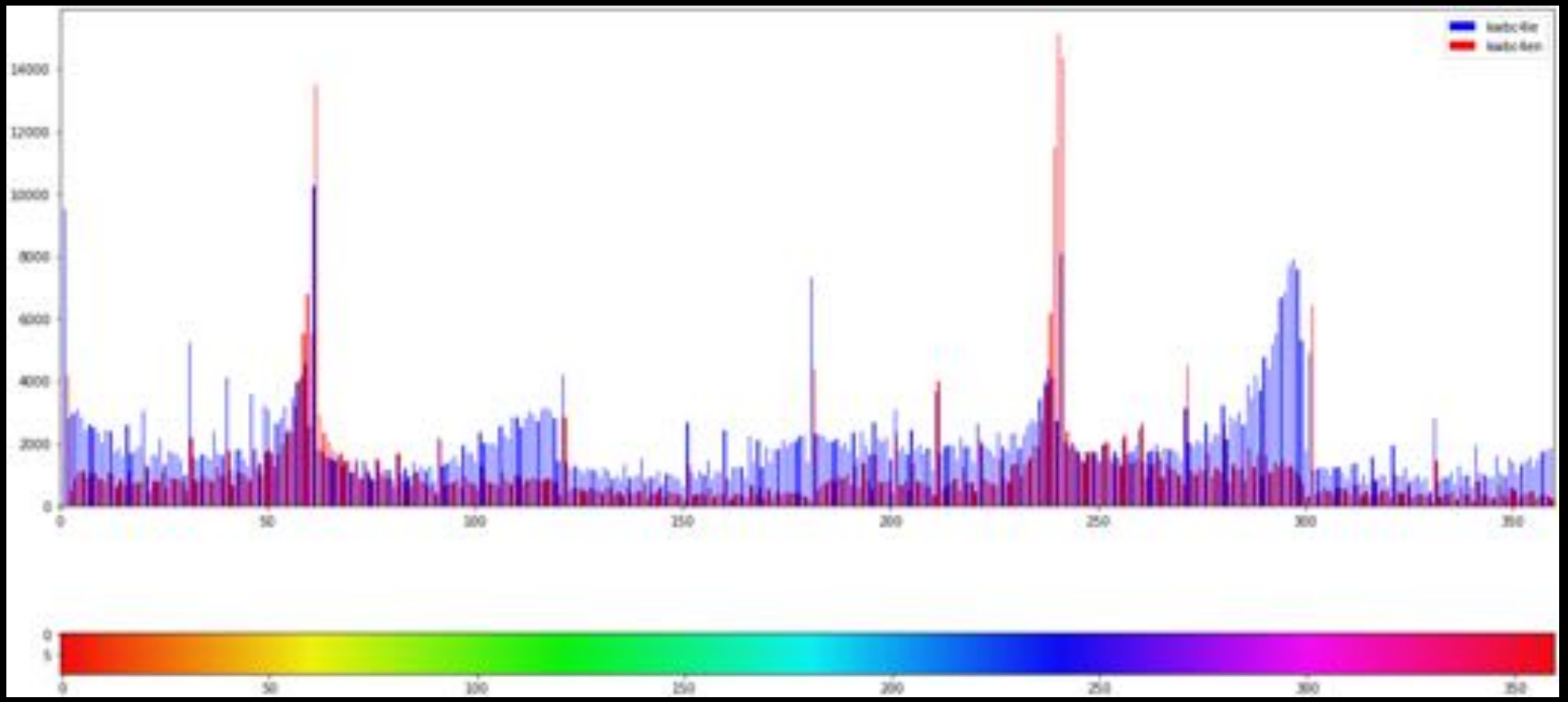
EN



LE

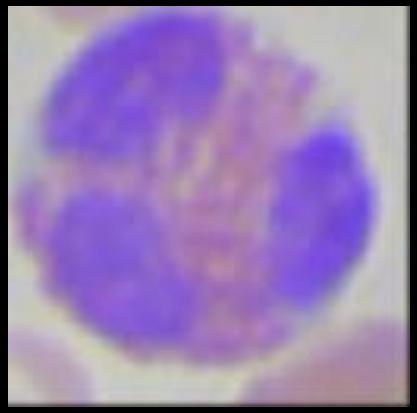


Accumulated Hue histogram – noisy images removed

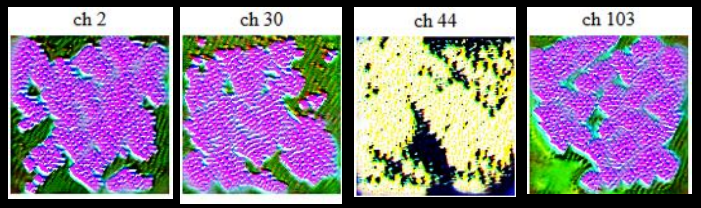


- LE more peaks than EN *[“more colourful”]*
- LE has red + cyan and purple + green; EN has mostly yellow+blue

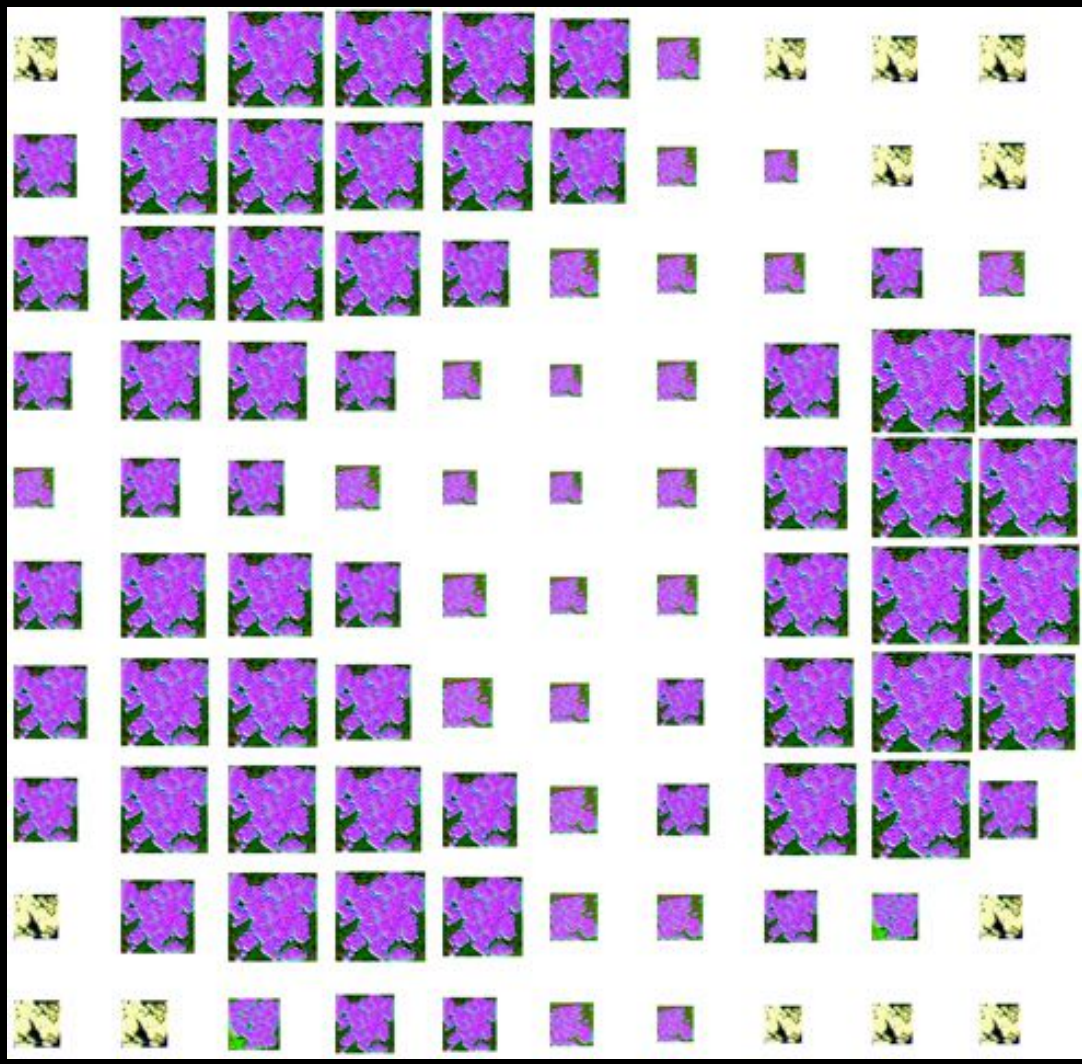
Strongest Features in an Image (BCI)



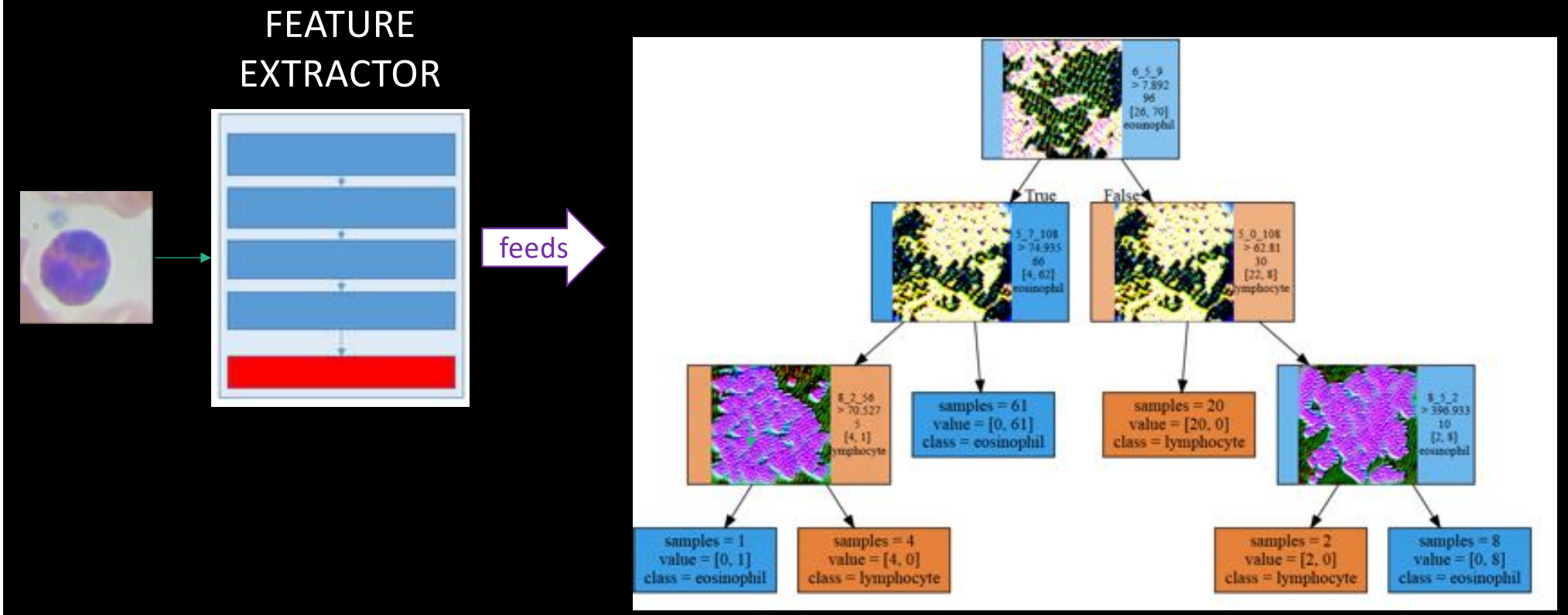
Eosinophil



Level 4 Features Used



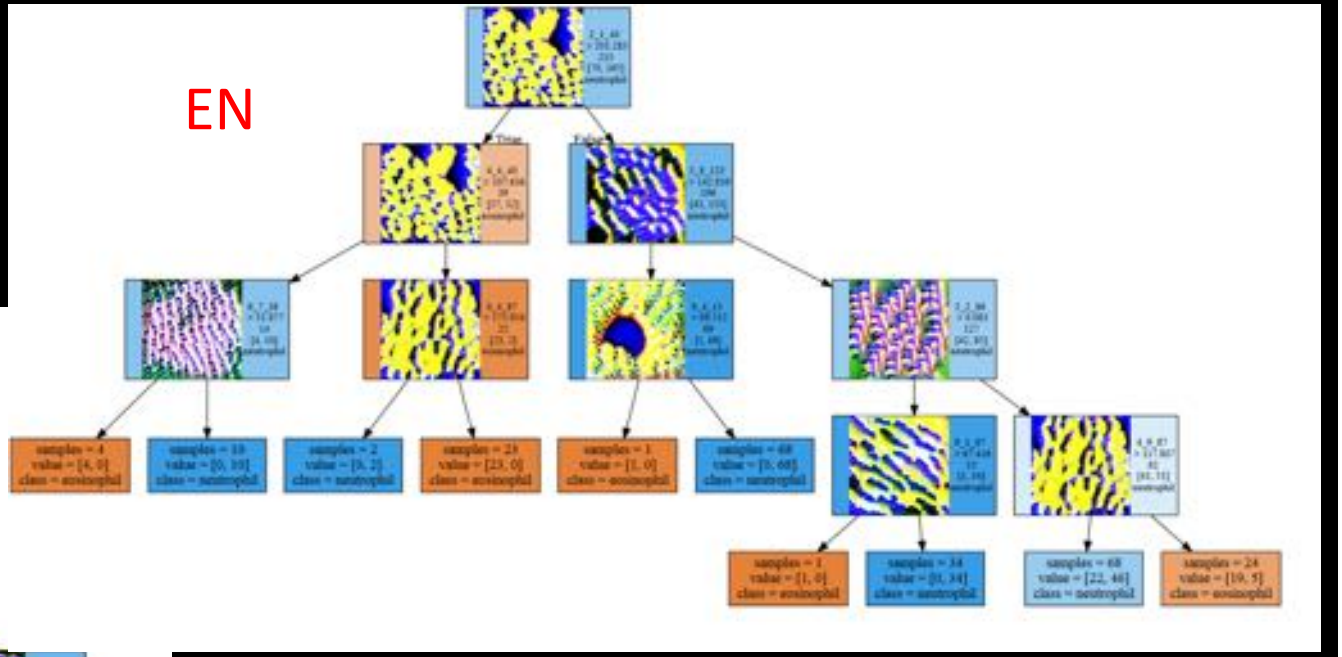
“Illuminated decision tree”



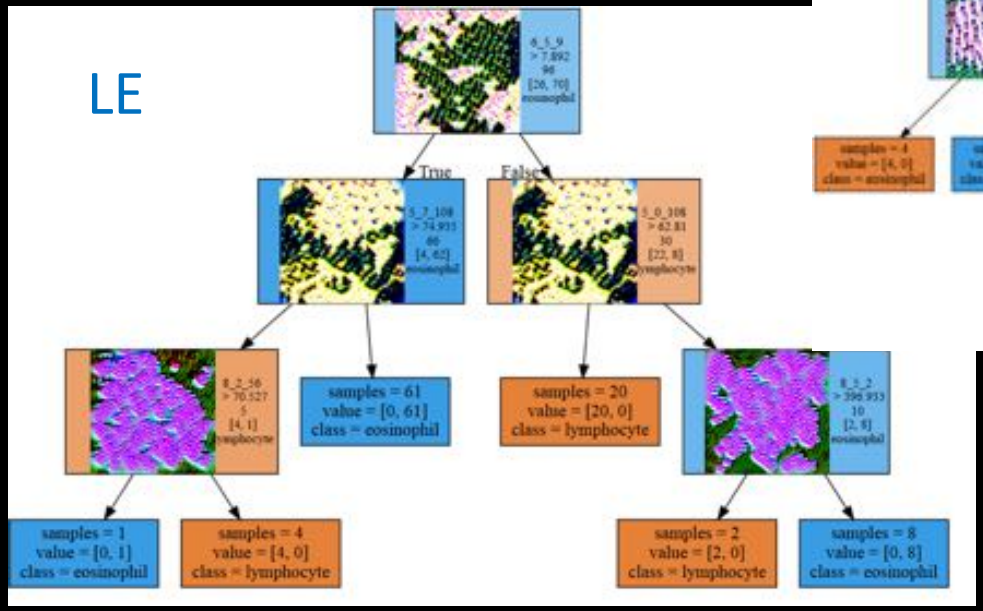
Visualises How Features are used in Decision Making

Comparing DTs of both models

EN



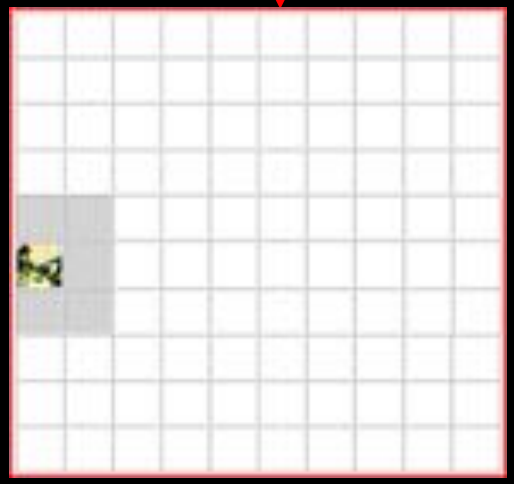
LE



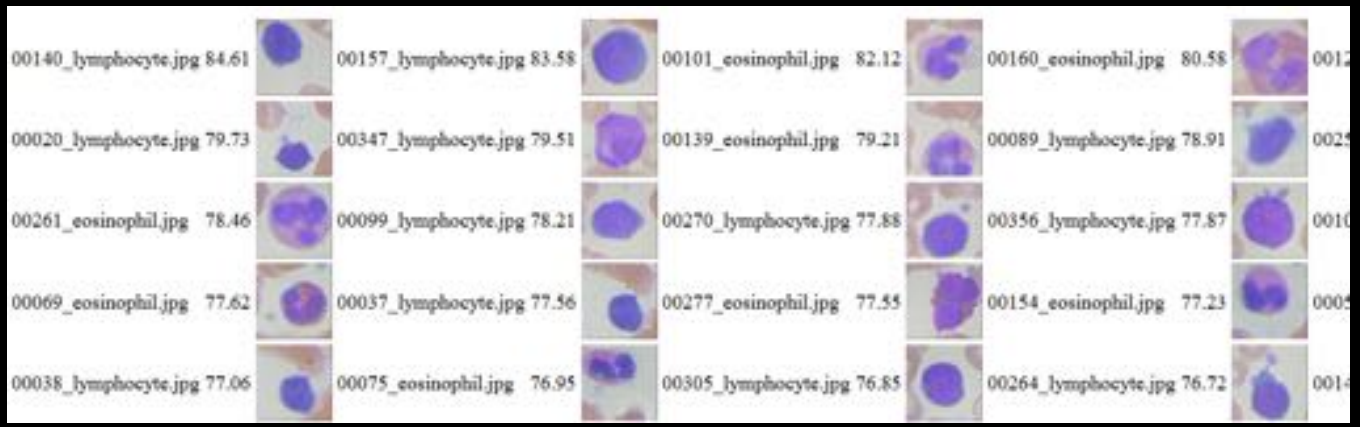
- More complex, first node fails to discriminate

- Simple decision tree; first node does the job; only three layers
- But a second level feature (5_0_108) is at the edge of the image

Visualising the Role of Feature (5_0_108)

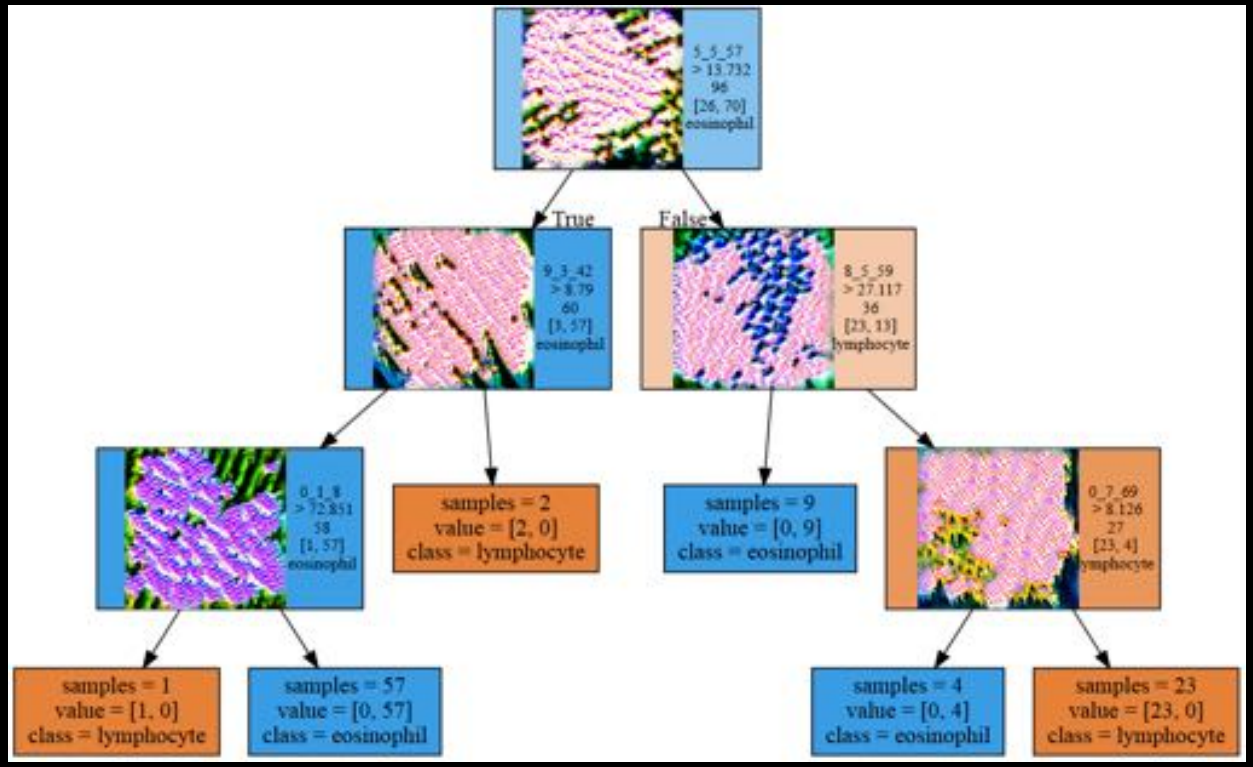


Images most matching this feature



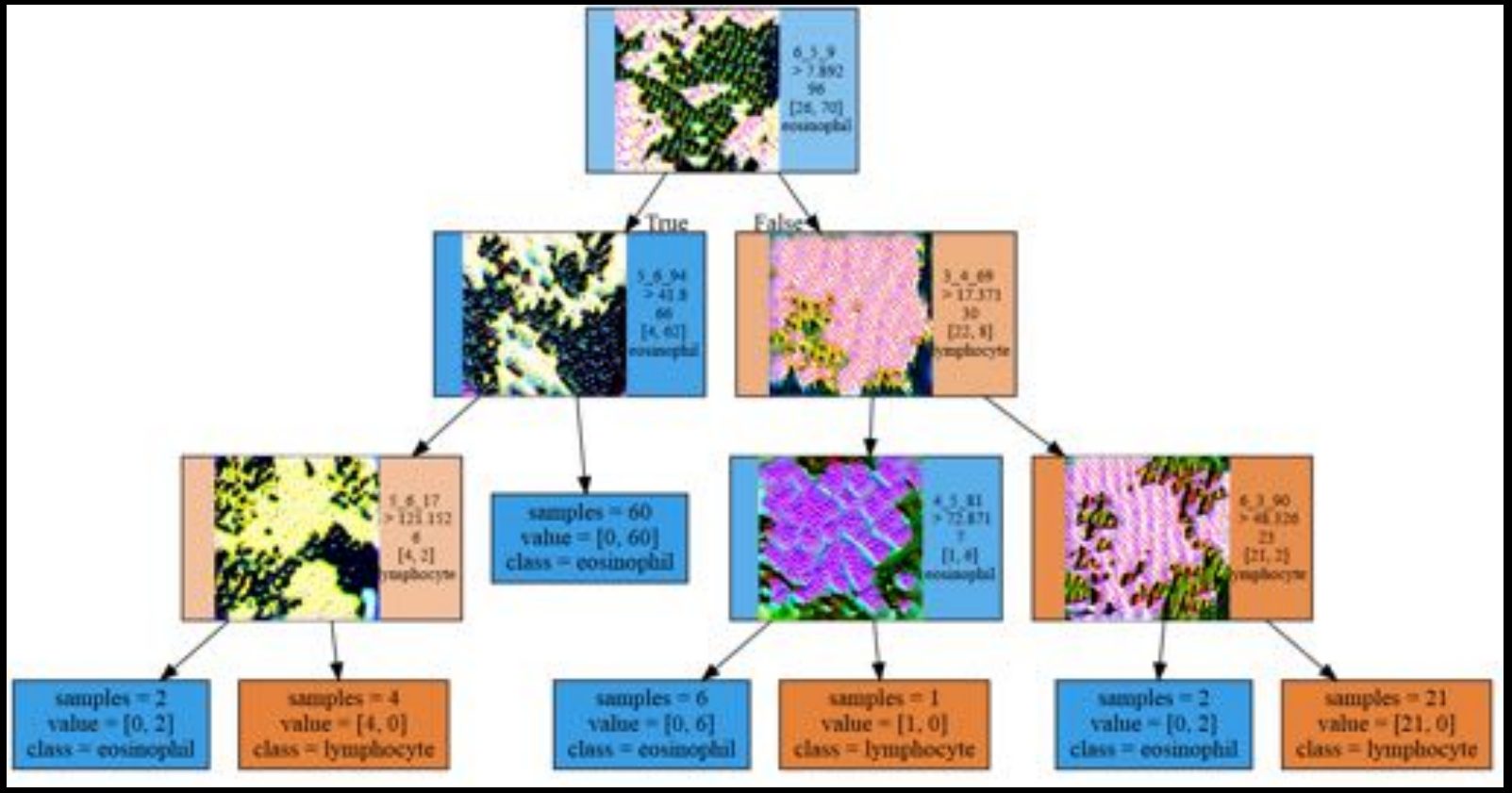
This feature is actually matching background on the left edge!

Switch-off all background features



All features look “cell-like” but accuracy is reduced

Switch-off all features at the picture edges



Some background in the middle, but accuracy close to its original level

Problems

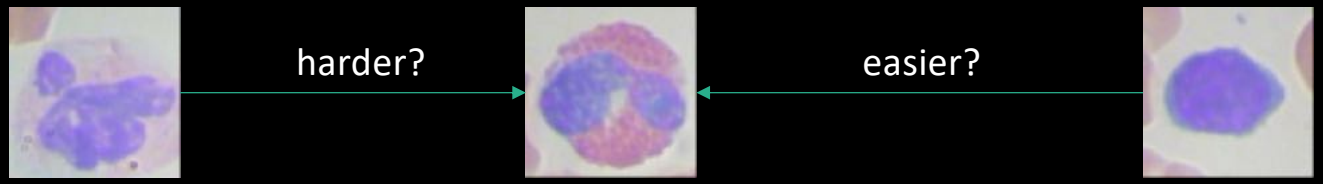
- The difficult EN has low accuracy (0.66)
- Even the easier LE is matching against the background

- One possibility is that there are insufficient training images
  - Use the large Kaggle database

- Another possibility is that there are insufficient layers in the model
  - Use a 6 layer model

Another attempt

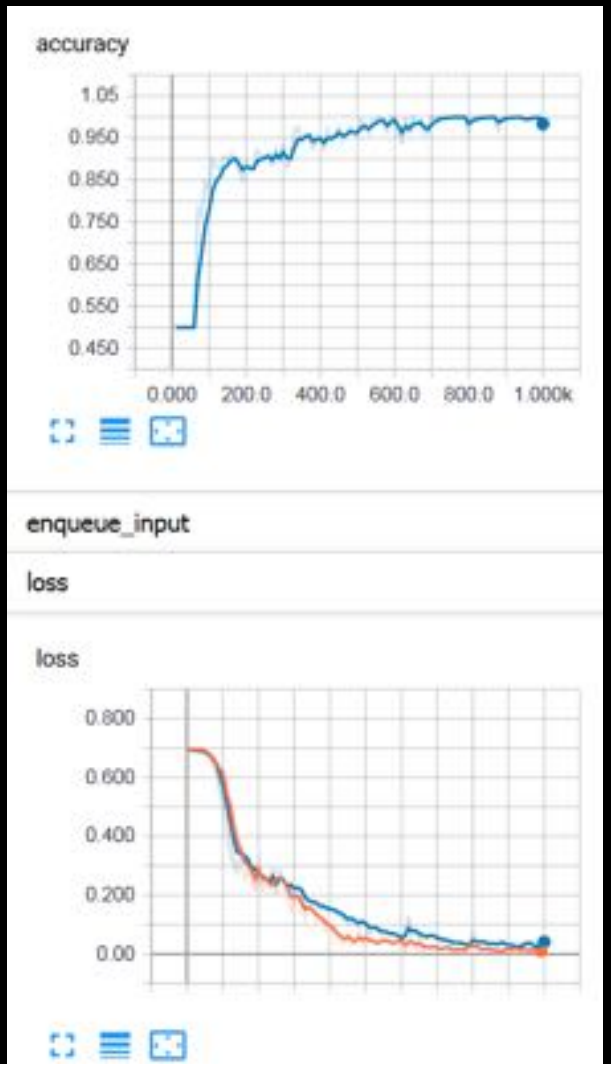
- Large Kaggle Dataset of 2537 neutrophil, 2506 eosinophil, 2487 lymphocyte
- 6 layer Convolution Neural Network



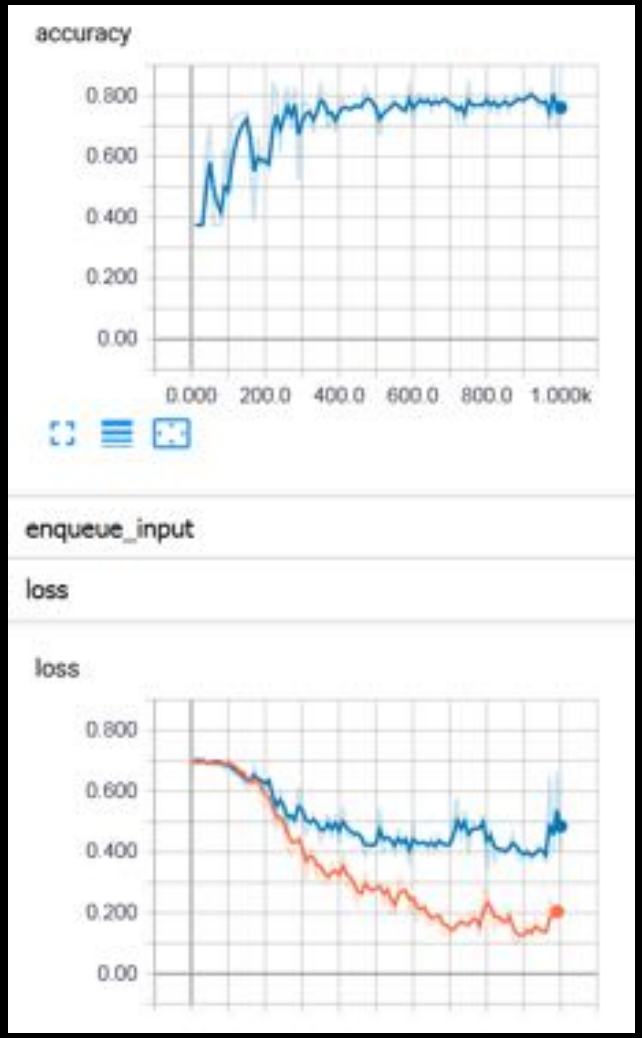
0.77 accuracy (“fairly good”) 0.97 accuracy (“good”)

Smoothed Accuracy and loss (to 1000 steps)

LE



EN

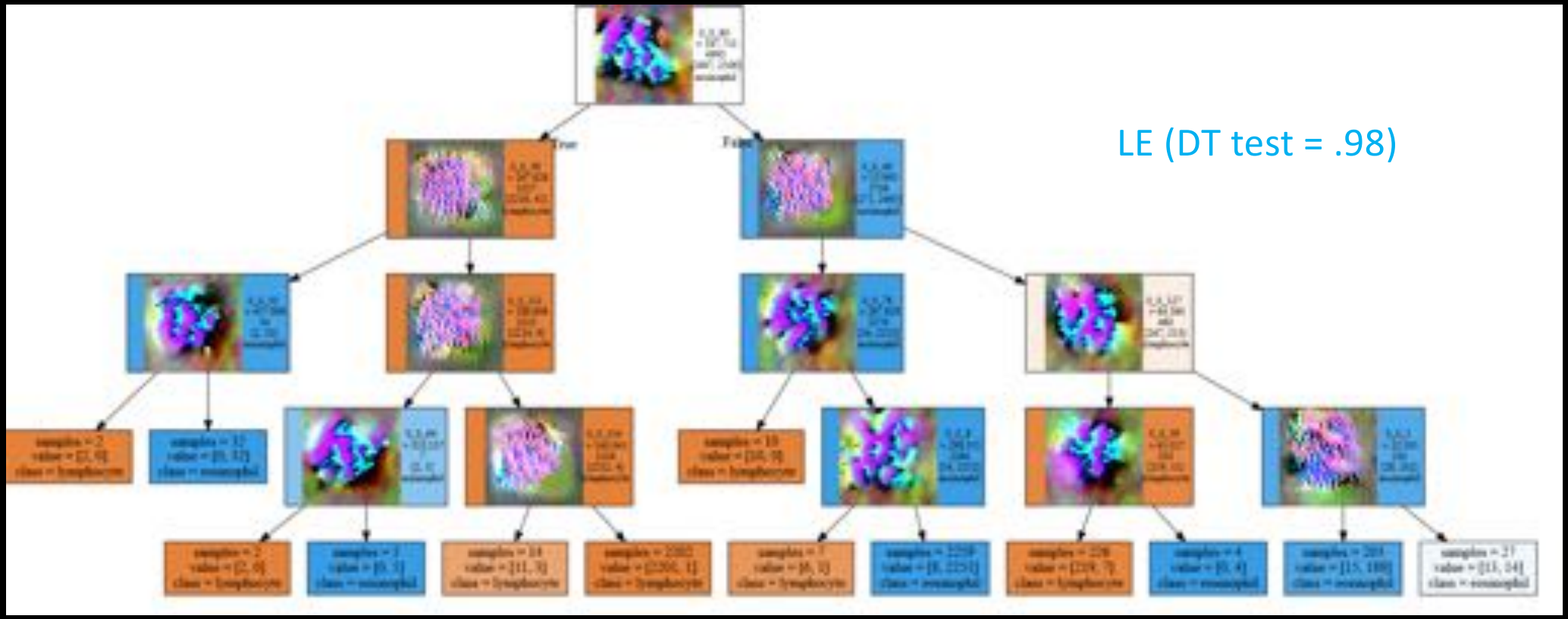


Layer 6M

LE – 1st 64 channels

EN – 1st 64 channels

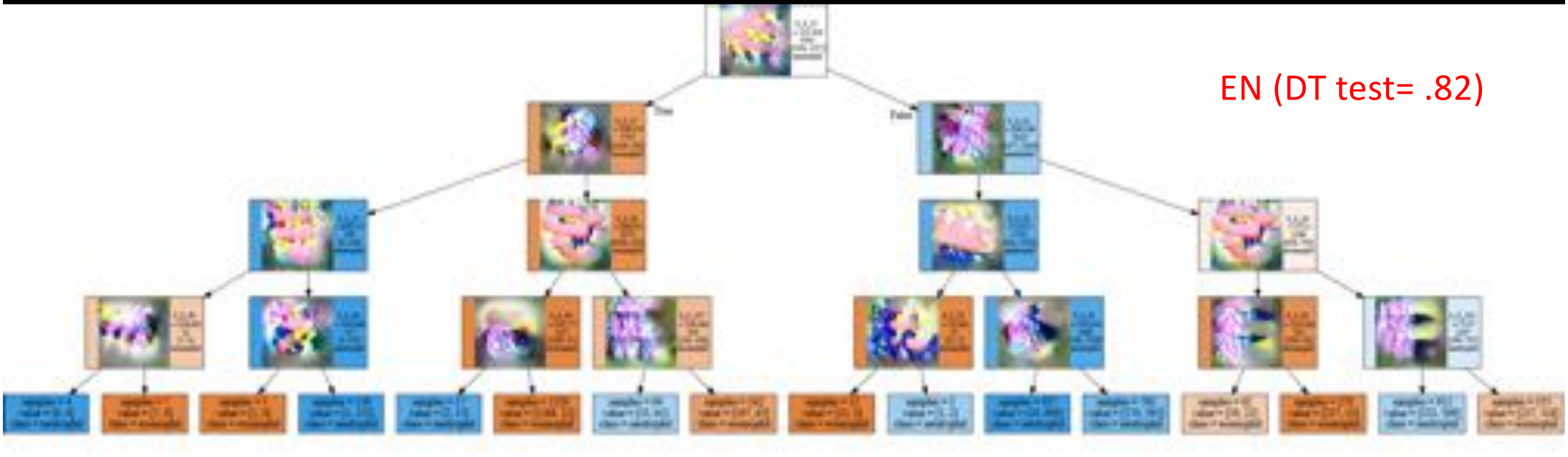
Decision Trees



LE (DT test = .98)

- Purple-blue features indicate lymphocytes, grainy pink-blue features indicate eosinophils
- 1st node makes a significant discrimination (89% of lymphocytes, 98% of eosinophils)
- Majority of each type end up in one leaf node, others are special cases
- No edge-like colours

Decision Trees

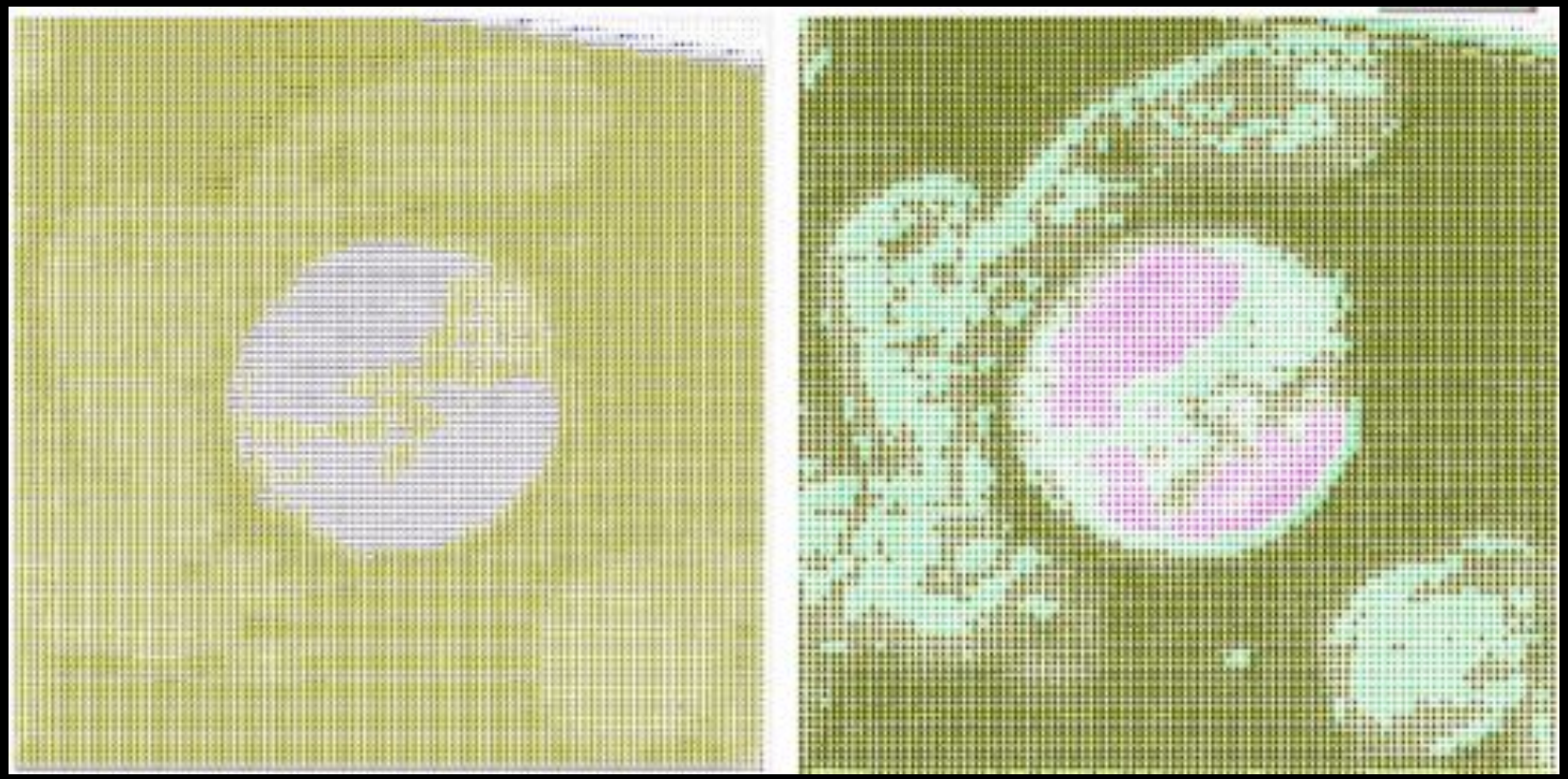


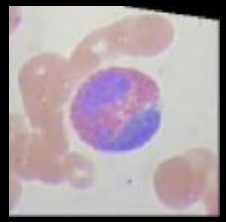
EN (DT test= .82)

- 1st node not so discriminating for eosinophils (57% eosinophils, 88% neutrophils)
- Pale pink features indicate eosinophil, grainy blue-pink features indicate neutrophil
- Many leaf nodes are failing to discriminate

BCI shows the focus of different classifiers

Layer 2



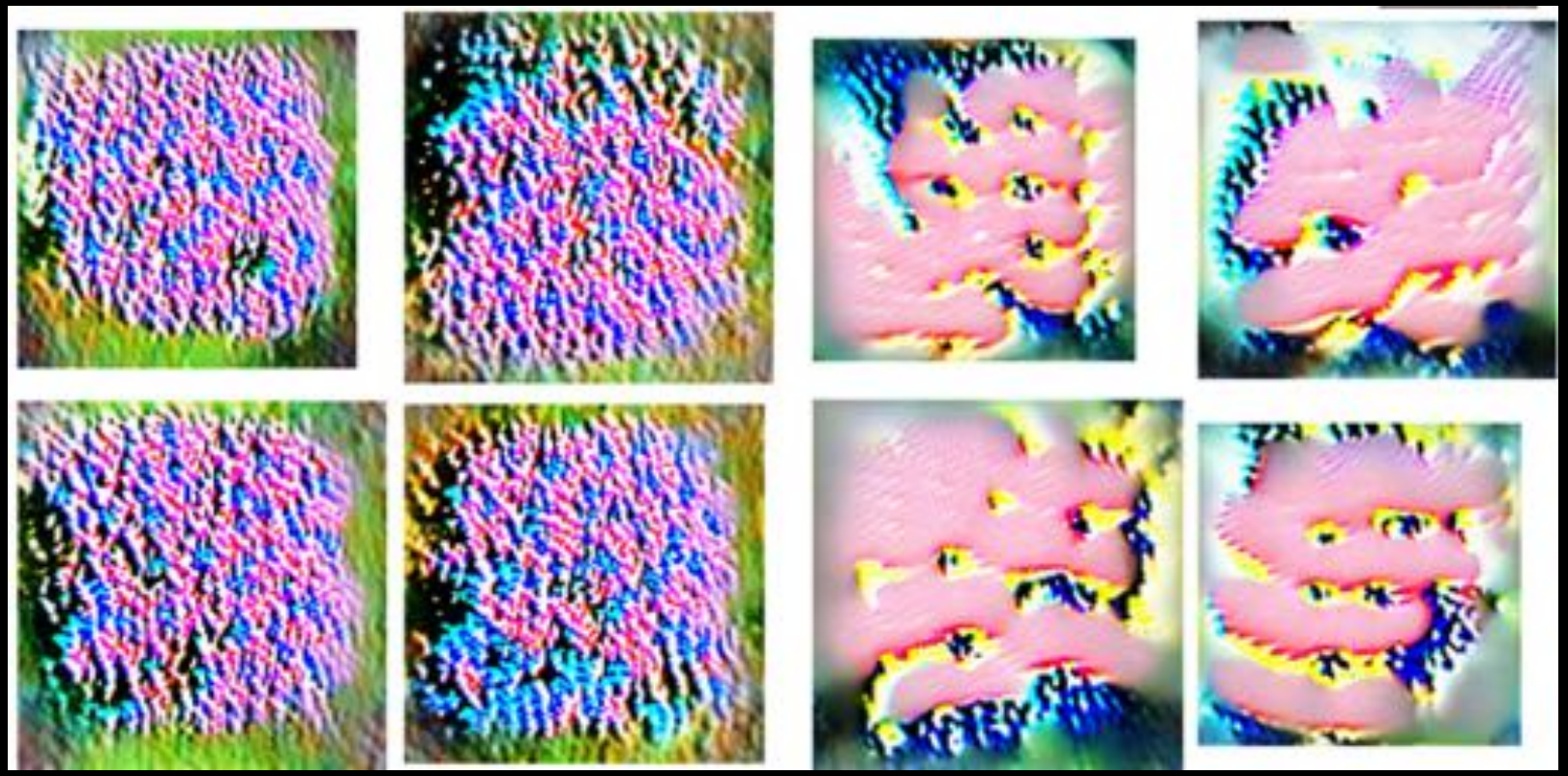


eosinophil

LE focuses on blueness

EN focuses on pinkness and dull-pink edges

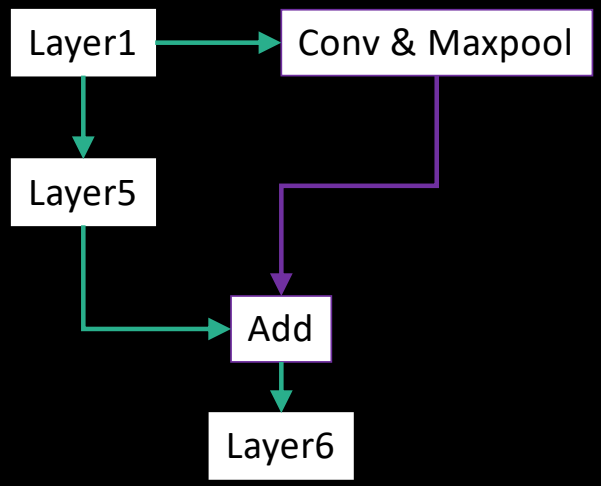
Layer 6 enhances these differences



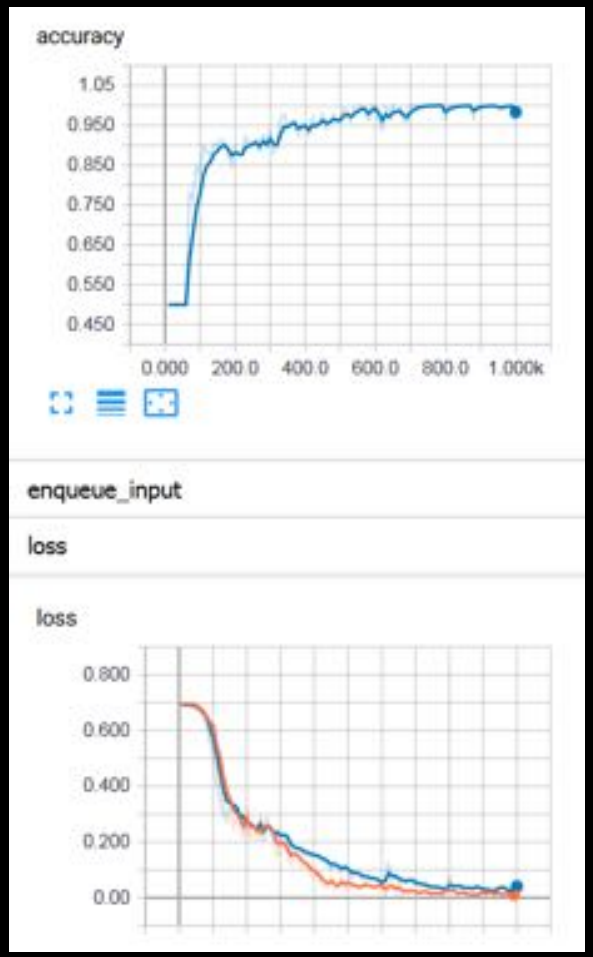
LE EN

Adding a “residual connection” to EN

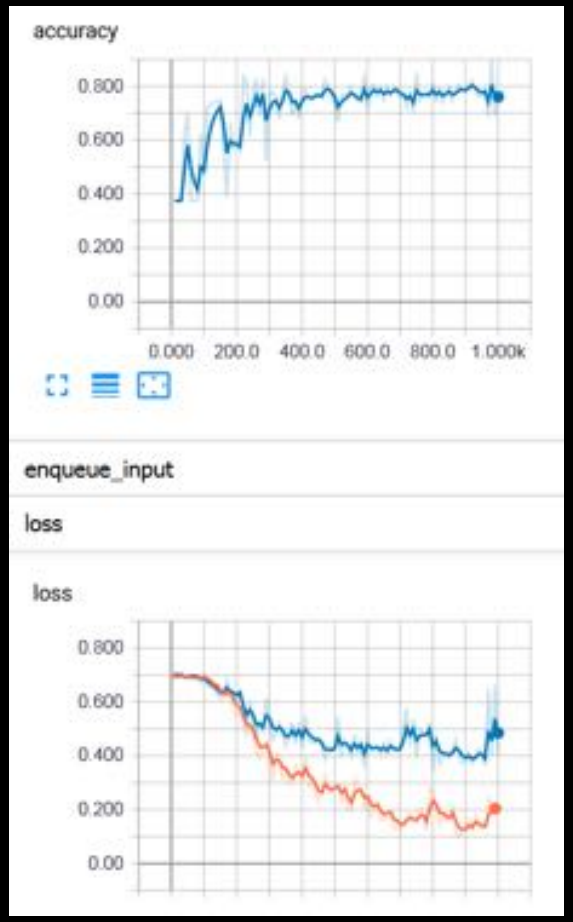
- To inject more information from the raw image, perhaps allowing more focus on the colour information
- For EN RC the residual connection added 0.05 to accuracy



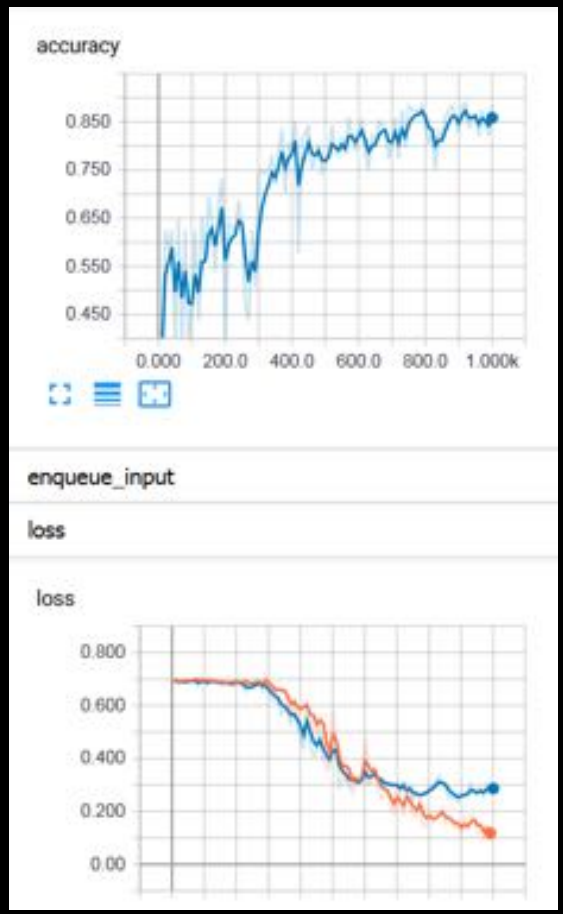
Comparing 3 Models



LE



EN



EN RC

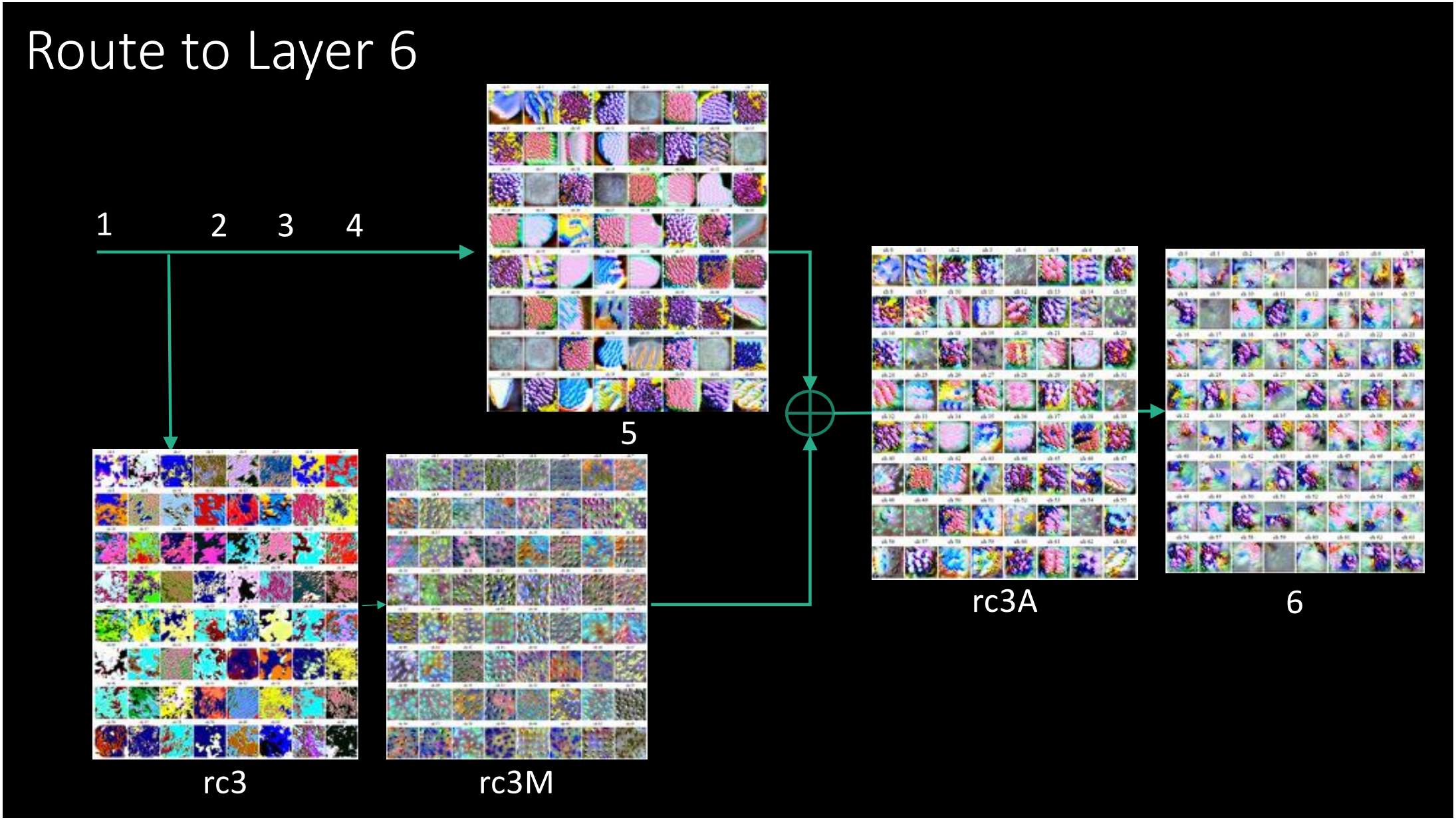
Layer 6 (1st 64 channels)

EN

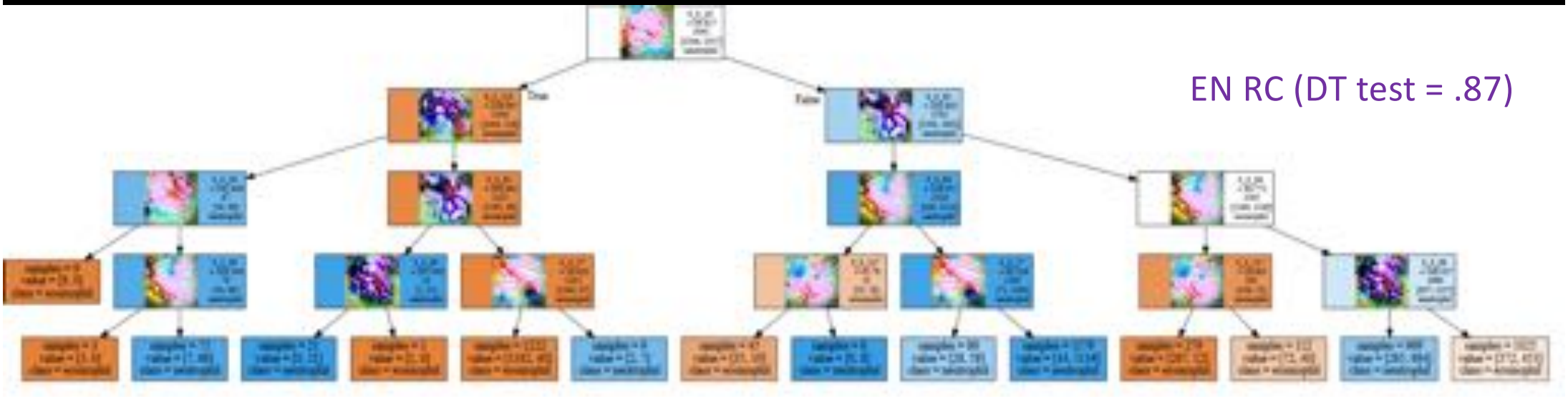
EN RC

Generally there is more colour in the EN RC version

Route to Layer 6



Decision Tree for EN RC

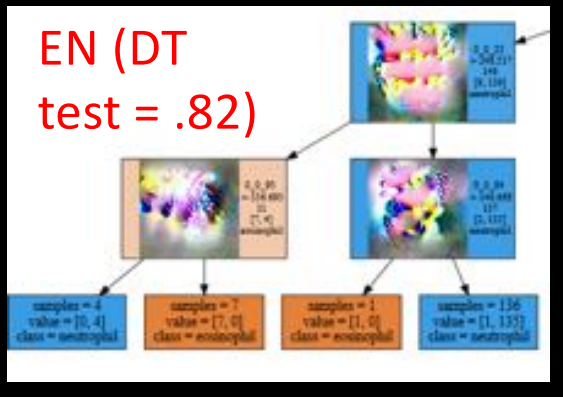


EN RC (DT test = .87)

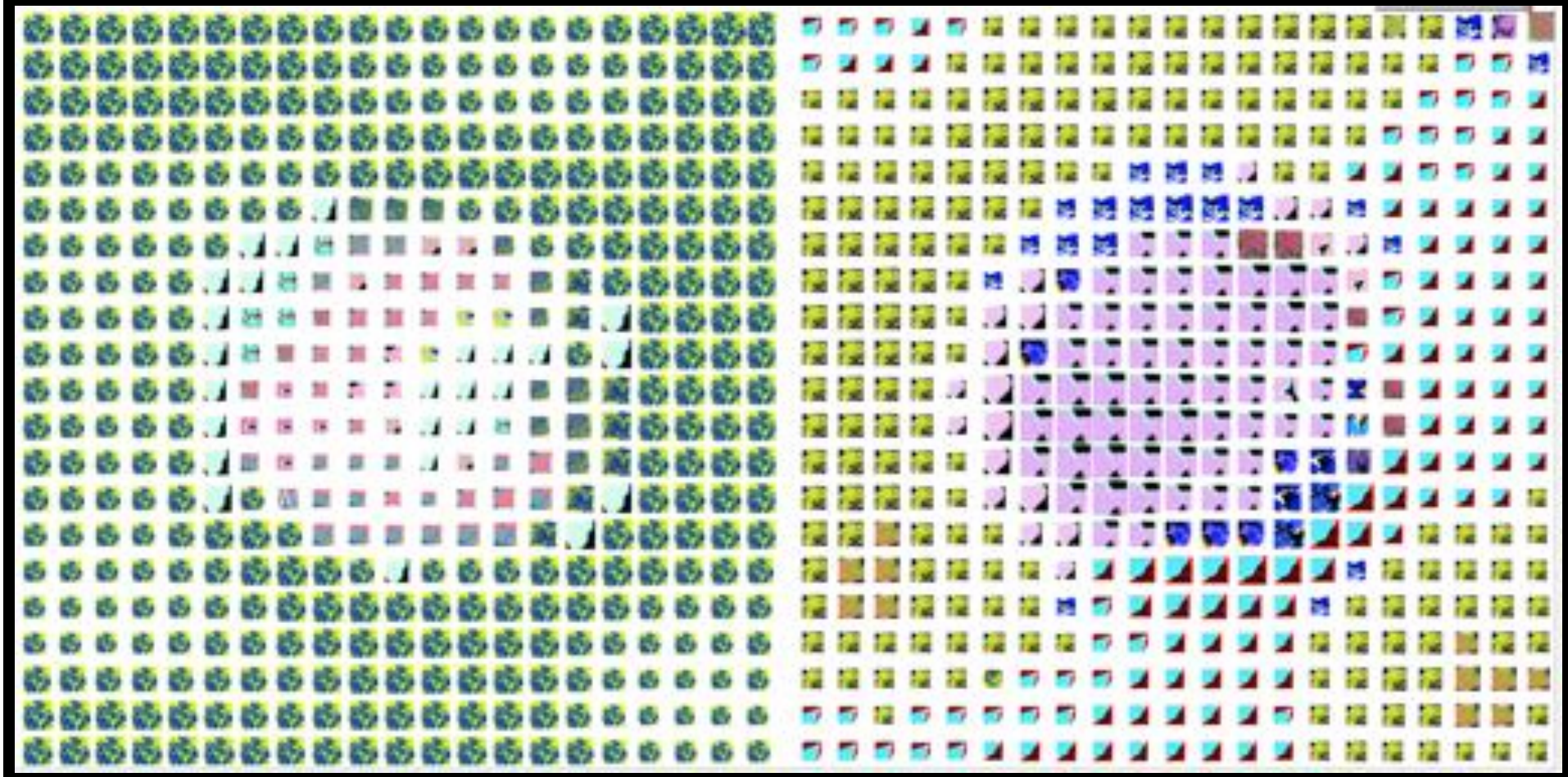
Features with pink areas are eosinophil, features with purple areas are neutrophil

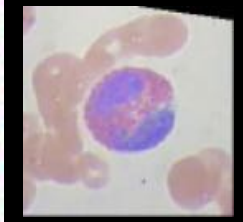
Compared to EN, this is more intense in its colours

EN (DT test = .82)



Classifier focus



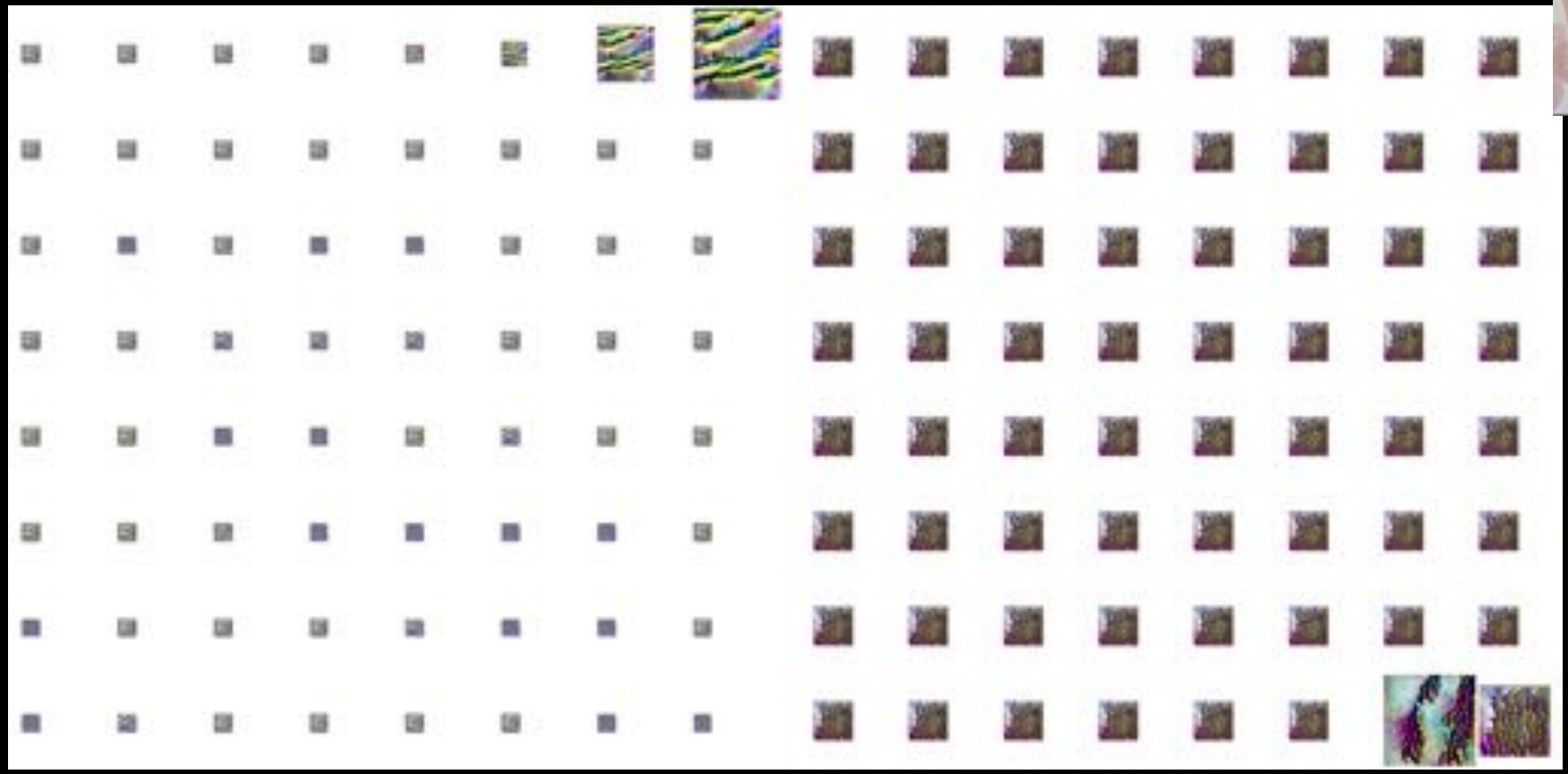


eosinophil

EN

EN-RC More components discriminated

Focus of classifiers with less precise Learning Rate



EN focuses on top right corner

EN-RC focuses on bottom right blob

Benefits

- Visualising features and decisions gives insight into the conclusions
- Exploring poor performance can help design of better models
- Deep Learning decisions could be integrated into other lines of reasoning

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 explanations
- Propose this research plan to the DAIS ITA program
- Complete the write up and submit the PhD!

Recap – here's what we covered

- 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]*

IEEE CogSIMA™

	Monday, April 8	
8:00 - 9:00 am	*Breakfast*	
9:00 - 9:10 am	T1: *Tutorial Session 1: Conversational Explanations - Explainable AI through Human-Machine Conversation*	
9:10 - 10:00 am		
10:00 - 10:30 am		*Coffee Break*
10:30 am - 12:00 pm		
12:00 - 1:30 pm	*Lunch (On your own)*	

Thank you for listening!

Conversational Explanations

Explainable AI through human-machine conversation

Dave Braines
dave_braines@uk.ibm.com

IEEE CogSIMA™