LEAP
A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

UKRI Engineering and Physical Sciences Research Council

Interpreting Artificial Intelligence Solutions for Healthcare

Organizers:

Dr Miquel Perelló Nieto & Dr Nawid Keshtmand

19 November 2025

About Dr Miquel Perelló Nieto

Research experience

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

sphere

TAILOR
Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

Research Interests

- Machine Learning
- Uncertainty quantification
- Optimal decision making
- Real-world applications
- Healthcare

About Dr Nawid Keshtmand

Research experience

Research Interests

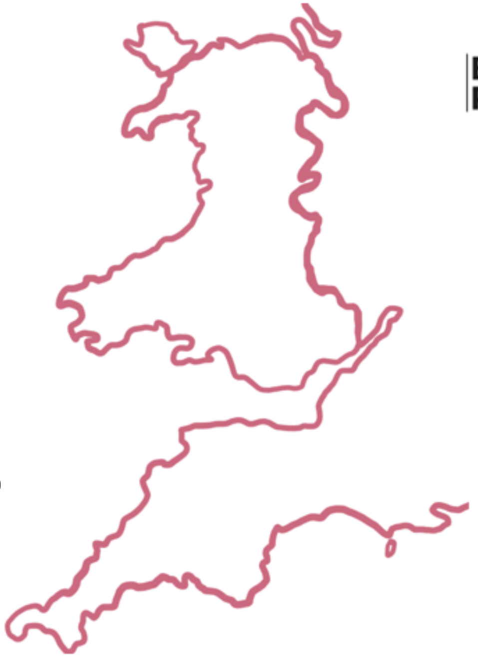
- Machine Learning
- Out-of-Distribution data
- Self-supervised learning
- Healthcare
- Climate

LEAP Digital Health Hub

Leadership **E**ngagement **A**cceleration & **P**artnership

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

- Hub for the South West of England and Wales
- **5** regional universities and HDR UK
- Led by the University of Bristol
- Network of **200+** organisations
- **£4M** of funding from EPSRC (2023-2026)
- **£1M** of research funding allocated
- Portfolio of **10** collaborative research pro
  1. Care outside of the hospital
  2. Service and resource planning
  3. Frailty, fall prediction and fall prevention
  4. Smartphone and wearable technologies



LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Agenda

10:30

1. What is AI?
2. AI in healthcare
3. Demistifying AI

11:15

Break

11:25

4. Ethics and Regulations
5. Explainability
6. Large Language Models
7. Discussion

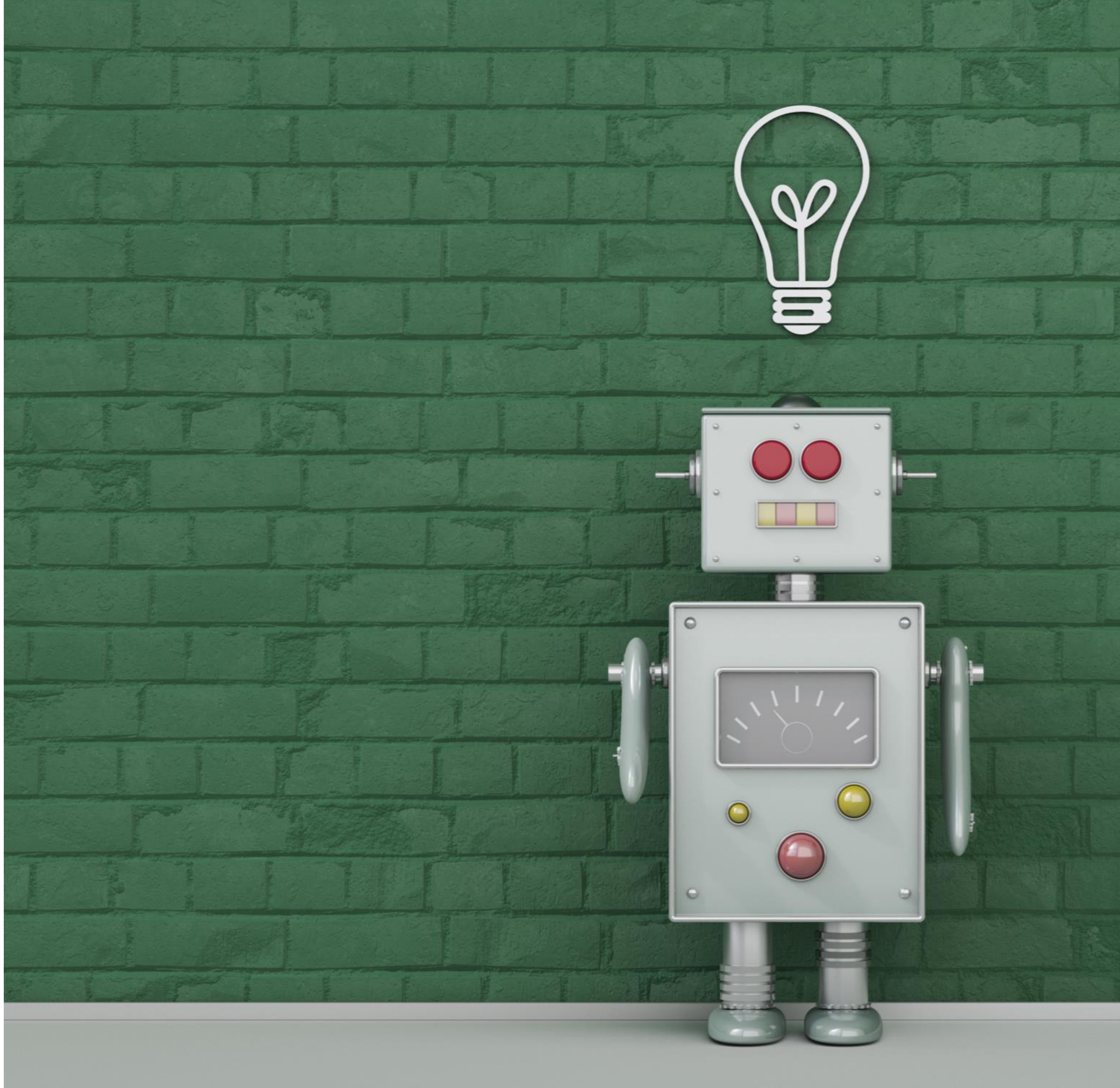
12:30

LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

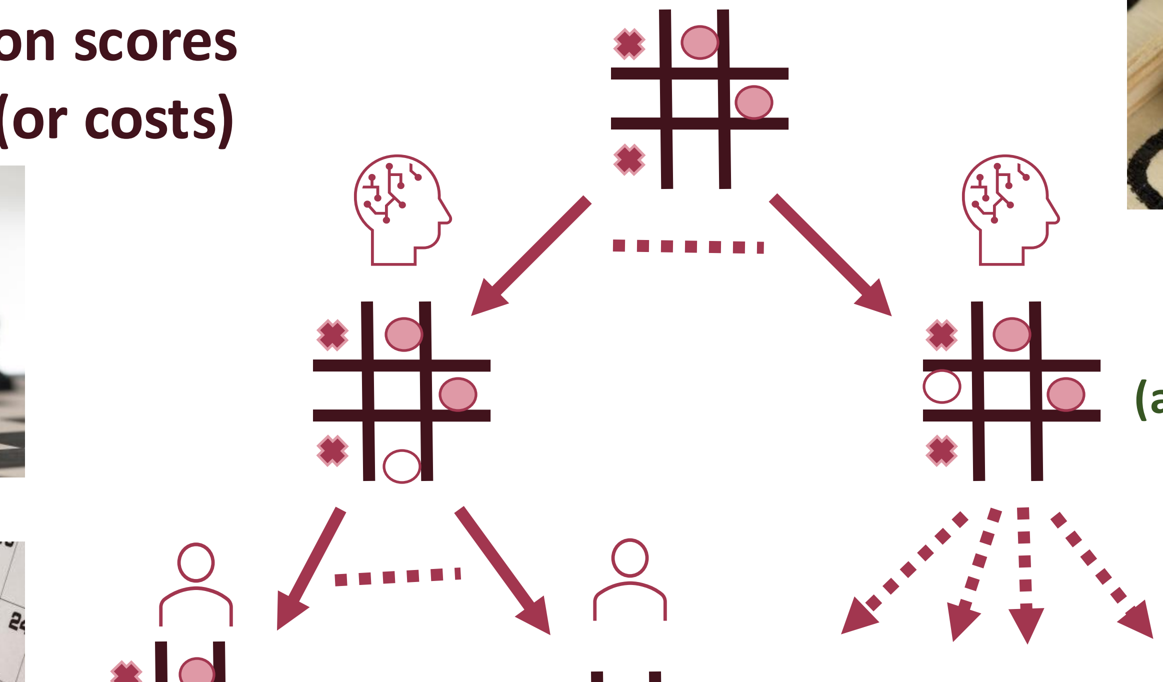
1. What is AI?

a. Examples
b. AI vs Machine Learning

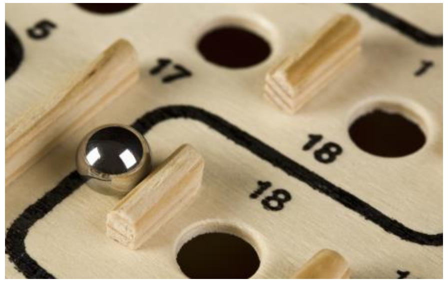


AI example: Search algorithms

Decisions based on scores (or costs)

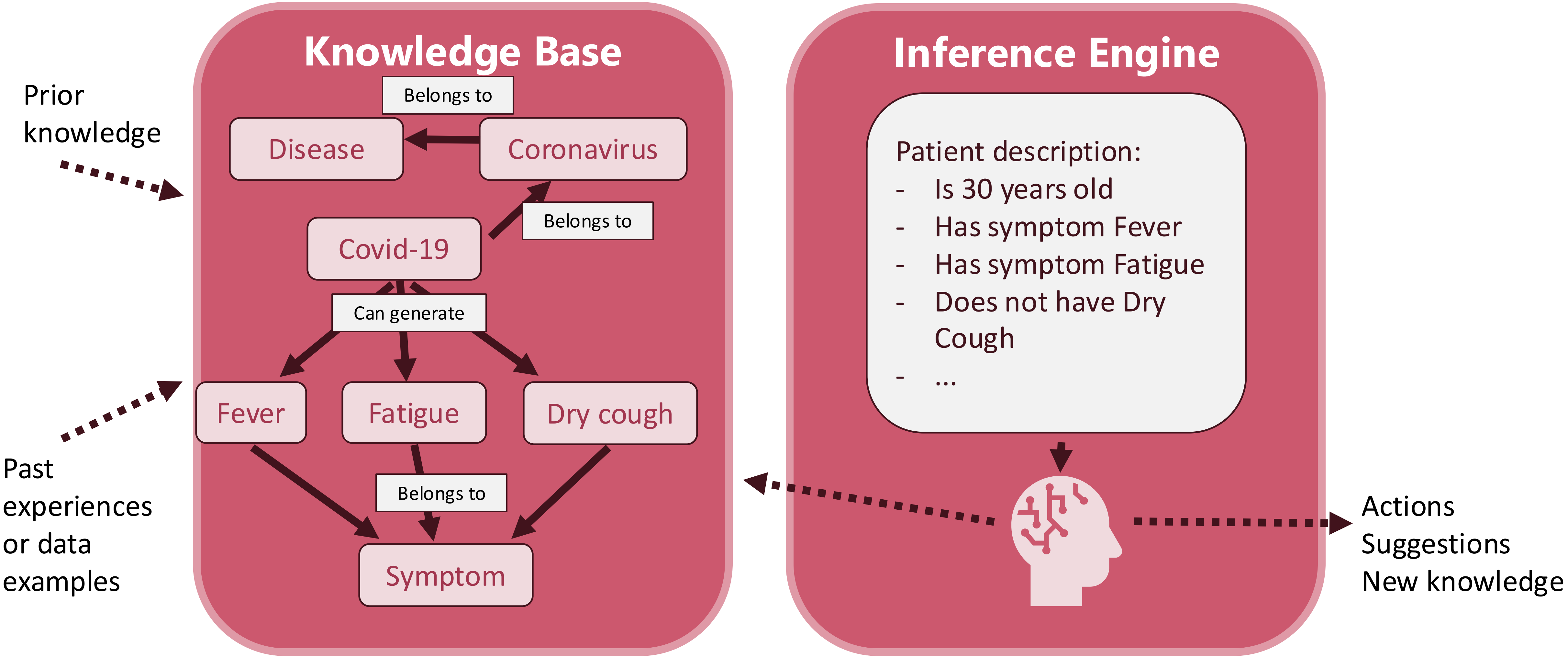


Good move (avoids human winning)



LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

AI example: Knowledge base



AI examples

Information retrieval

- Google searches
- Online articles
- Restaurants
- Travel agencies

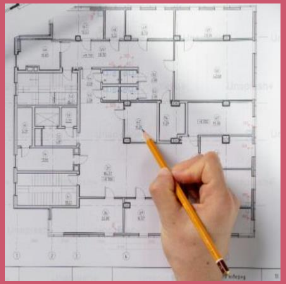
Recomender systems

- Similar to information retrieval but with personal profile
- Movie recommendations
- Music recommendations
- Online shopping

AI and Machine Learning

Artificial Intelligence (AI)

Less data requirements



Machine Learning (ML)

Deep Learning

More data

AI vs Machine Learning

Deep Blue

- AI brute force + heuristics (200M chess positions per second)
- Timeline:
  - 1985 at Carnegie Mellon University (called Chip Test)
  - 1989 at IBM (aka Deep Thought)
  - 1996 vs Garry Kasparov: 1 win, 2 draws, 3 loses
  - 1997 vs Garry Kasparov: 2 wins, 3 draws, 1 loss

Alpha Go

- AI + Deep Learning (CNNs 12 layers, reinforcement learning)
- Timeline
  - 2015 vs Fan Hui: 5 wins
  - 2016 vs Lee Sedol: 4 wins, 1 loss
  - 2017 vs Ke Jie: 3 wins

LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

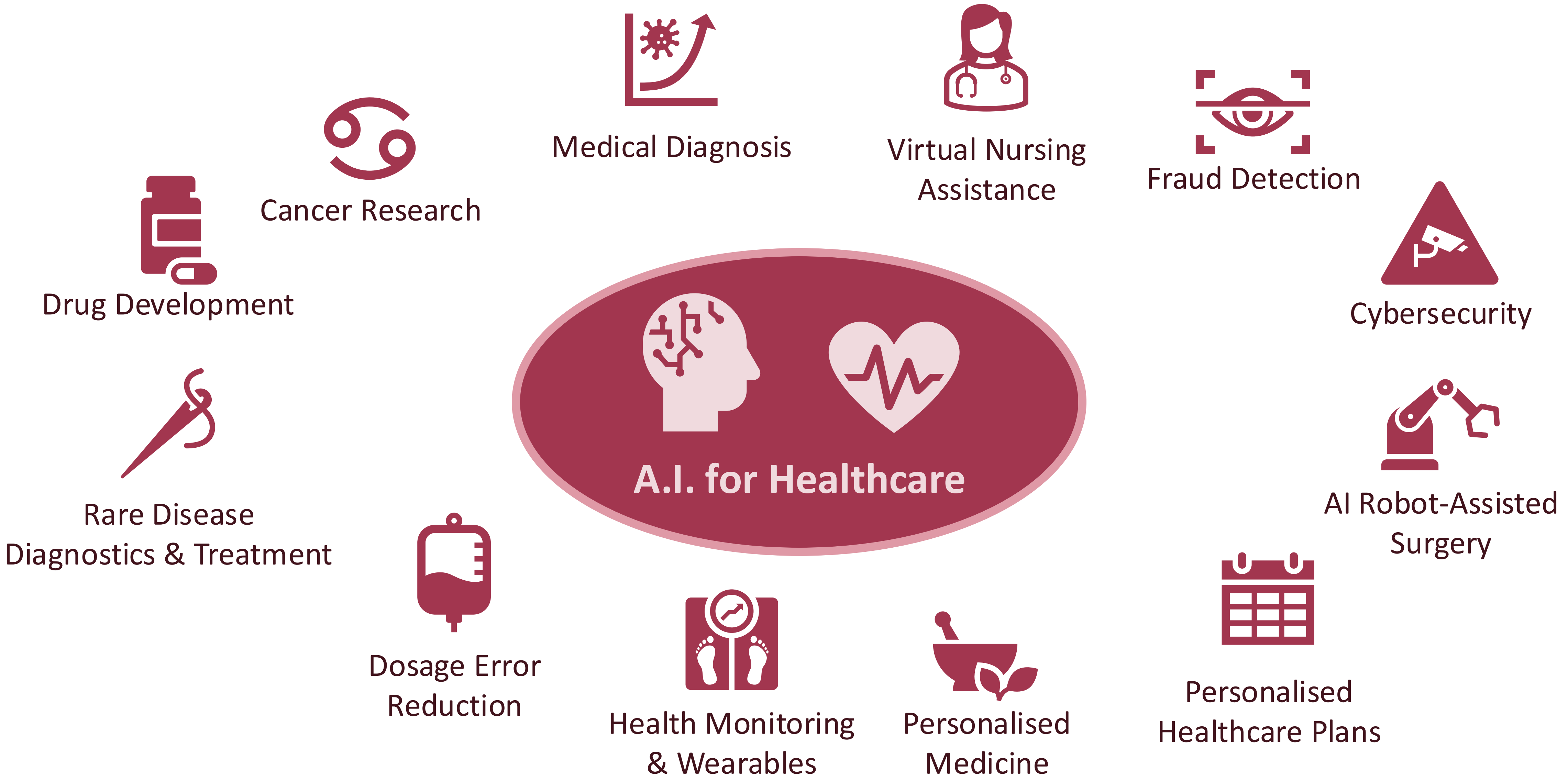
2. AI in Healthcare

a. Applications

b. Examples

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

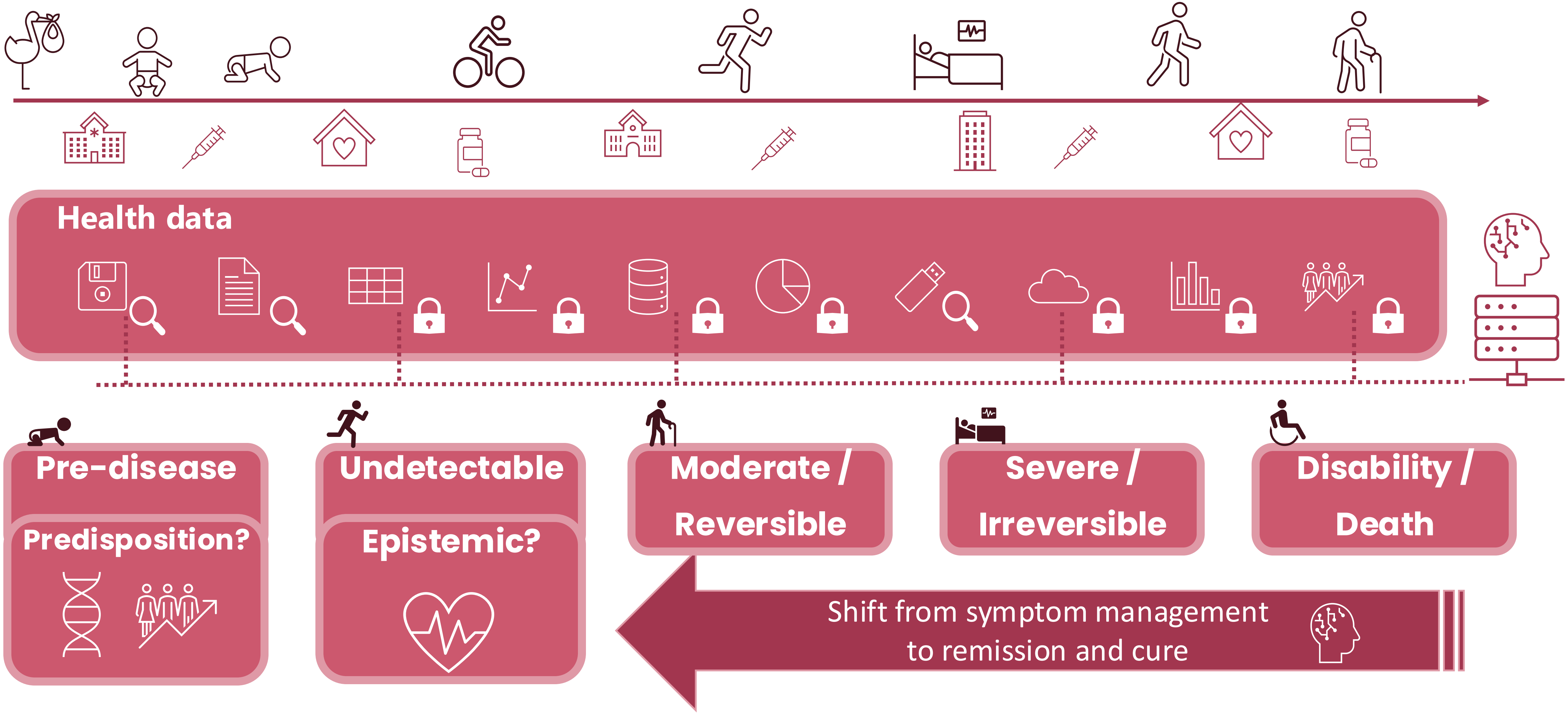
Applications of AI in Healthcare



LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Early diagnosis



LEAP
A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

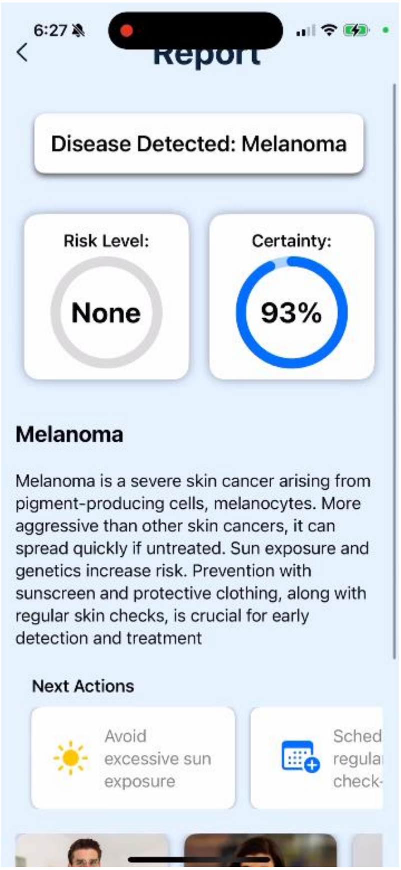
2. AI in Healthcare

a. Applications
b. Examples

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

DermAI: AI-Powered Skin Cancer Detection

- Diagnosis of non-melanoma and melanoma skin cancers and other skin conditions.
- Tailored reports including risk levels, recommended specialists, research findings, actionable next steps.
- AI chatbot functionality for real-time natural language interactions.



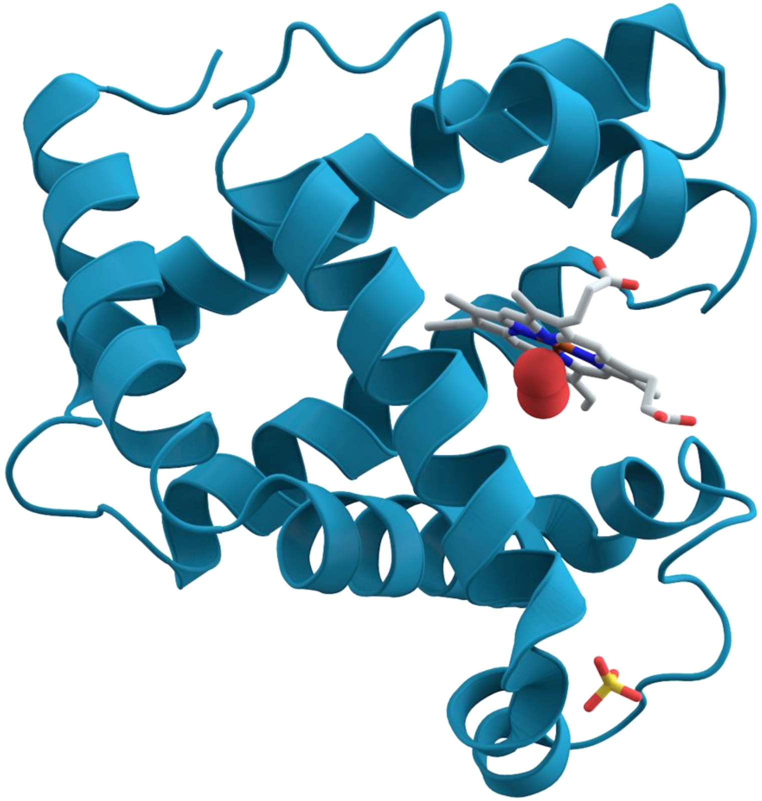
WHO Skin NTD (Neglected Tropical Diseases) mobile application

- App development "No Leprosy Remains"
- Used in clinical workflows in 35 FHWs and 5 Kenyan counties
- Computer vision to classify photos of skin lessions with special focus on leprosy and other NTDs

Emily E. V. Quilter, Cynthia Ruth Butlin, Carme Carrion, Jose-Antonio Ruiz-Postigo; The WHO Skin NTD mobile application – a paradigm shift in leprosy diagnosis through Artificial Intelligence?; Leprosy Review; 2024; 95; 2; e2024030; DOI: 10.47276/lr.95.2.2024030

Protein structure prediction

- Predicting the 3D structure of a protein based on its amino acid sequence
- One of the most important open research problem for more than 50 years
- AlphaFold: a neural network evaluated in the 14th Critical Assessment of protein Structure Prediction
- We now have 98.5% of the human proteome 3D structures
  - 36% of it with very high accuracy
  - 22% with high accuracy



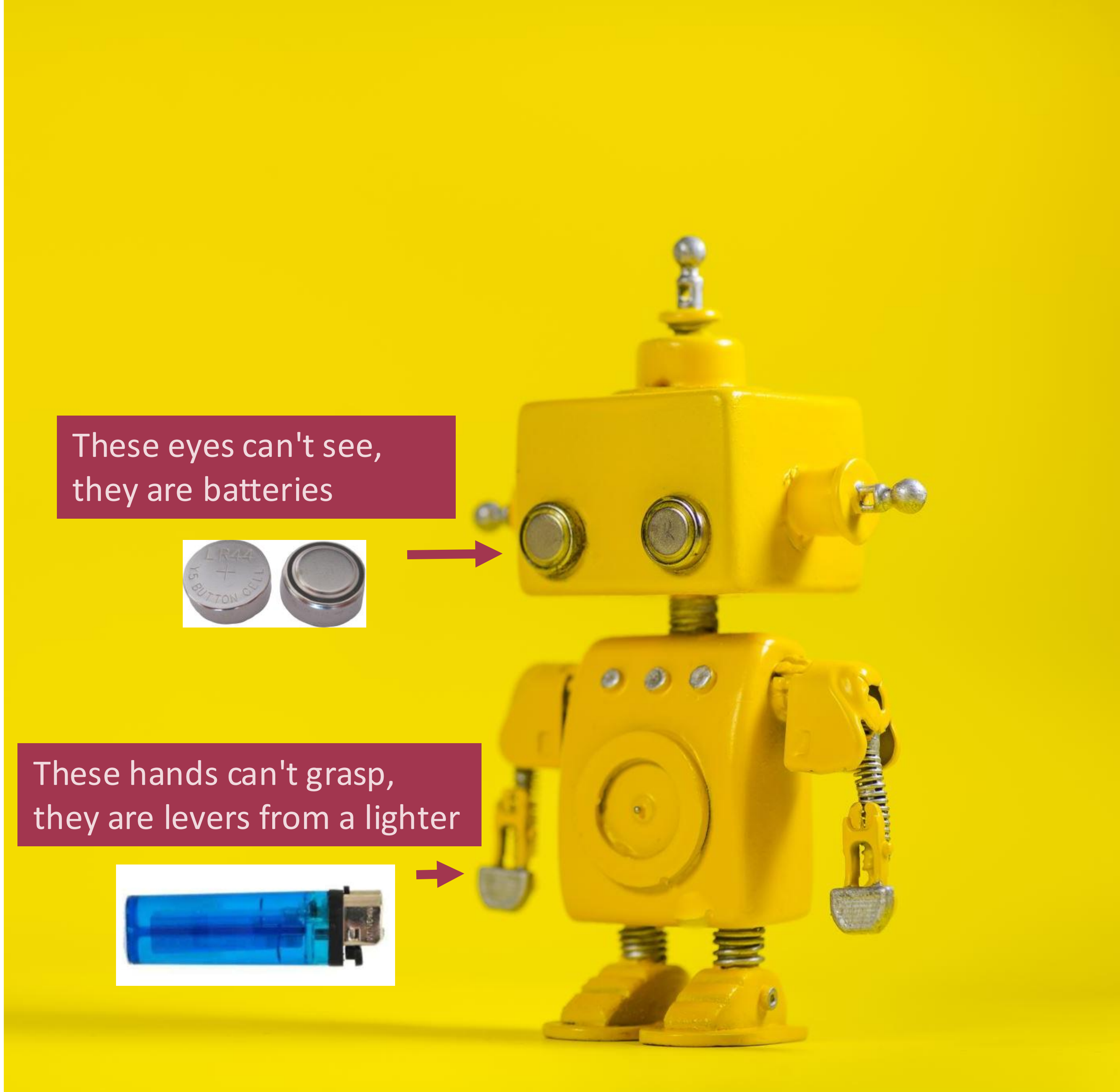
Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

3. Demystifying AI

a. ML Pipeline
b. Radiology
c. Scribes
d. Pitfalls



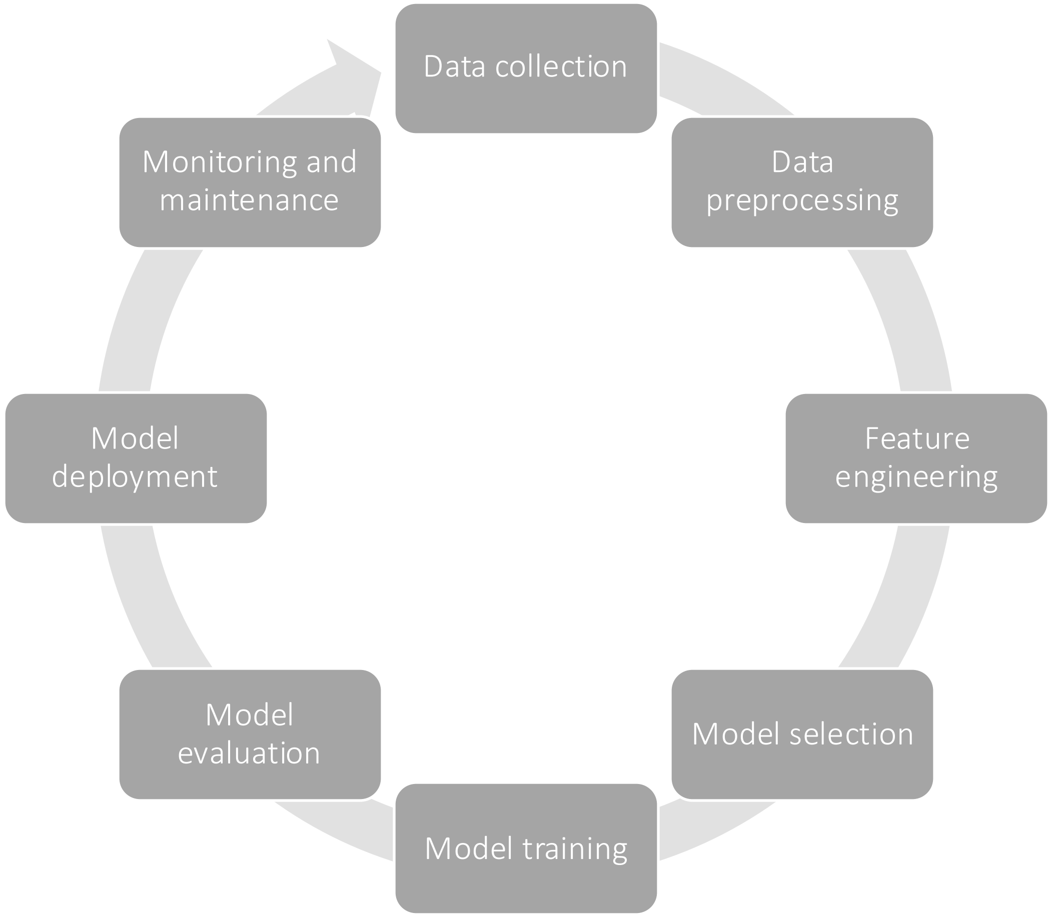
LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

3. Demystifying AI

a. ML Pipeline
b. Radiology
c. Scribes
d. Pitfalls

Machine Learning lifecycle



LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Data Preprocessing. Tabular

Quantitative

ID	Age	Sex	Weight	Height
1	26	0	65.2	155
2		2	78.5	178
3	18	1	58.1	
4	67	0	70.7	170
5	50			175

Qualitative

ID	Age	Sex	Weight	Height
6	young	male	normal	short
7	baby	female	underweight	small
8	adult	female		tall
9	teenager		normal	short
10	senior	male	overweight	

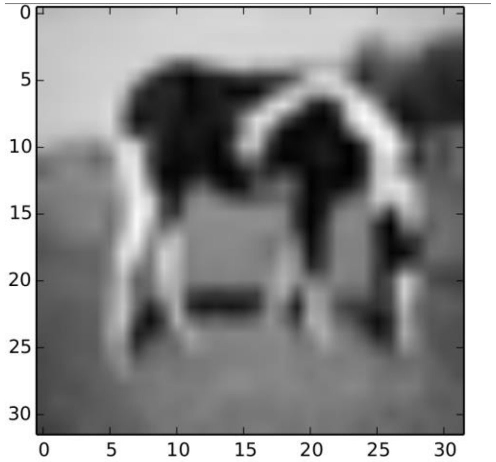
- Data is commonly converted into numeric data.
- Careful consideration for missing values.

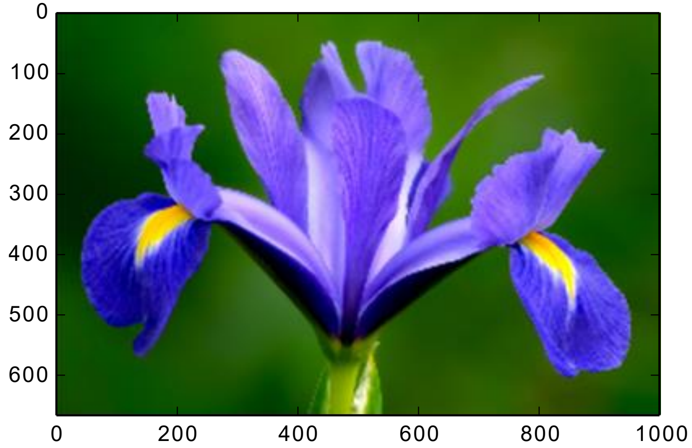
Data Preprocessing. Images

Grayscale

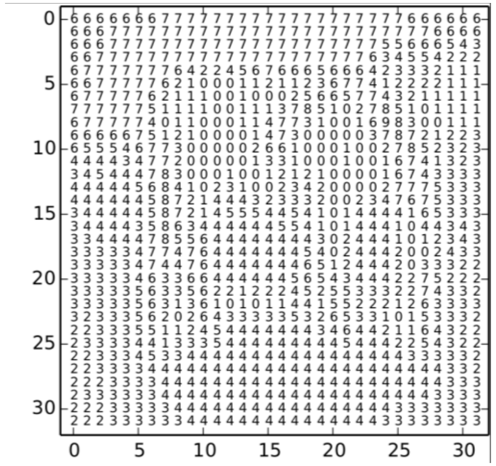
Colour

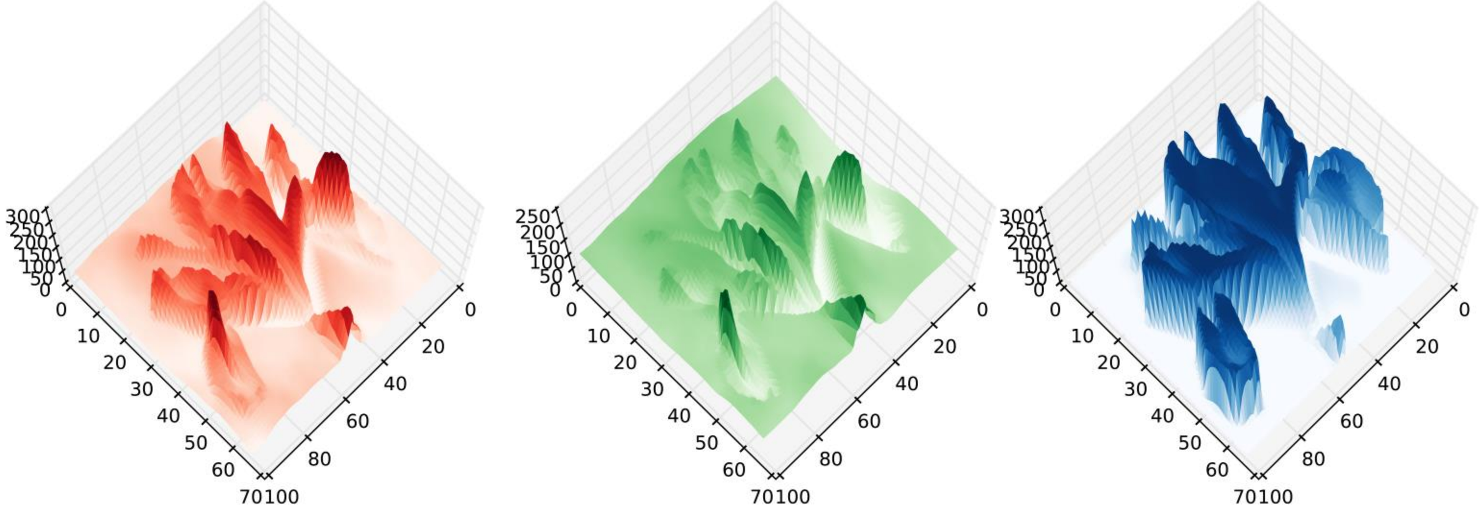
Original image





Computer representation





LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Data Preprocessing. Text

- *"Text can be encoded in numeric form in multiple ways, a simple example is in the form of bag of words"*

Words	Text	Can	Be	encoded	in	a	numeric	form	Multiple	Ways	Simple	...
Repetitions	1	1	1	1	3	1	1	2	1	1	1	

- This representation loses the order information.
- There are more complex approaches not covered here.

Data Preprocessing. Text

- *"Text can be encoded in numeric form in multiple ways, a simple example is in the form of bag of words"*

Words	Text	Can	Be	encoded	in	a	numeric	form	Multiple	Ways	Simple	...
Repetitions	1	1	1	1	3	1	1	2	1	1	1	

- N-gram: sequences of n adjacent symbols (words).

1-gram	Rep.
text	1
can	1
be	1
encoded	1
in	3
...	...

2-gram	Rep.
Text can	1
Can be	1
Be encoded	1
Encoded in	1
In numeric	1
...	...

3-gram	Rep.
Text can be	1
Can be encoded	1
Be encoded in	1
Encoded in numeric	1
In numeric form	1
...	...

Can we predict the new word in the sentence

A simple ...

And the next one?

In ...

LEAP

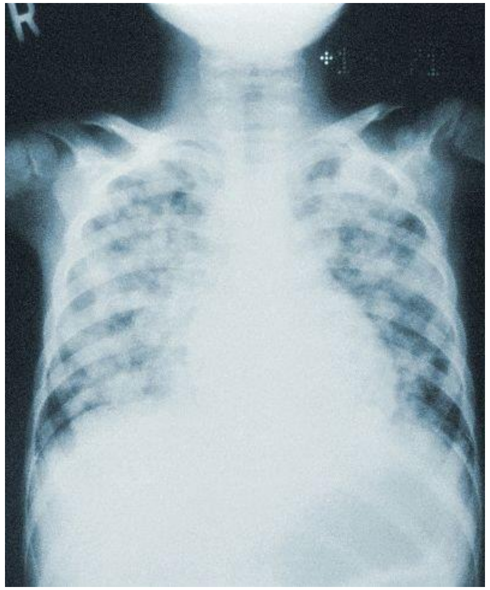
A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

3. Demystifying AI

a. ML Pipeline
b. Radiology
c. Scribes
d. Pitfalls

Radiology and image classification

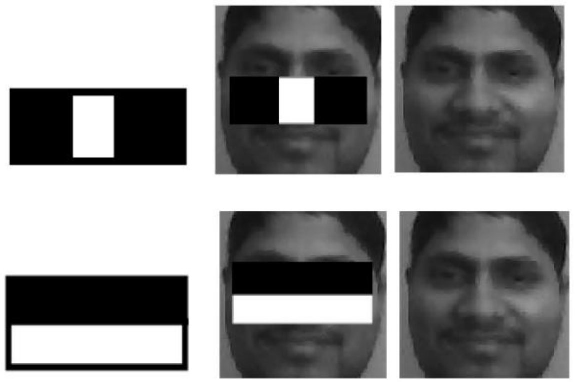
- The COVID-19 pandemic required a large amount of chest image analysis.
- AI could help speed up the process.
- But how could AI make a diagnosis?

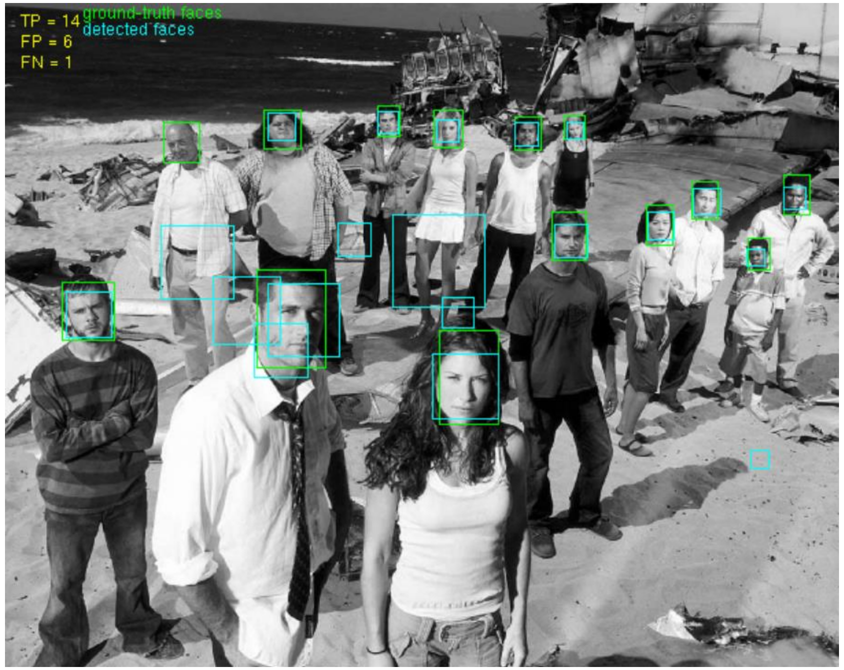


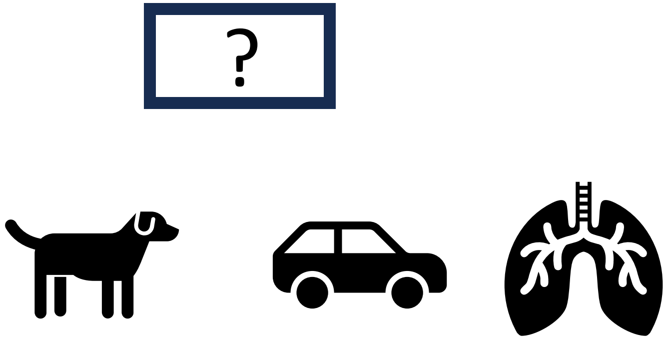
Feature engineering

- Haar filters and Viola-Jones face detection (Haar, A., Viola, P., and Jones, M.)
- How to create filters for every object?

Convolve the filter

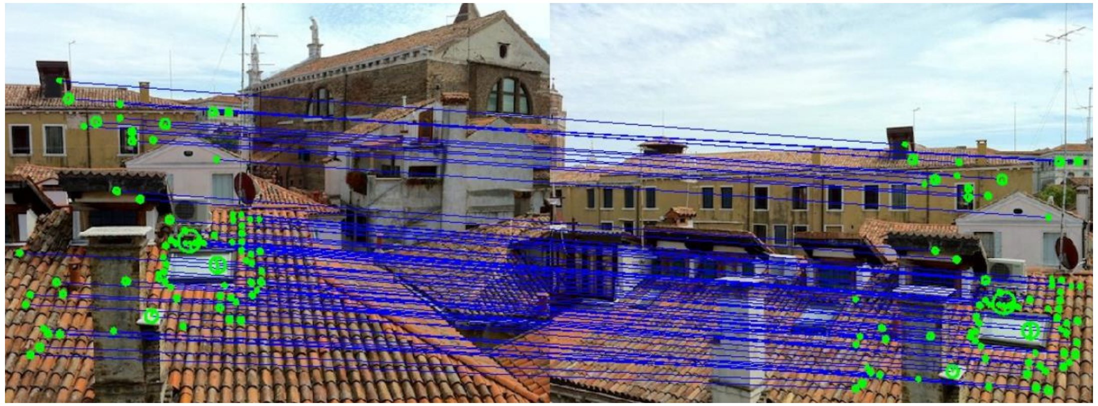






Automatic detection of Feature descriptors

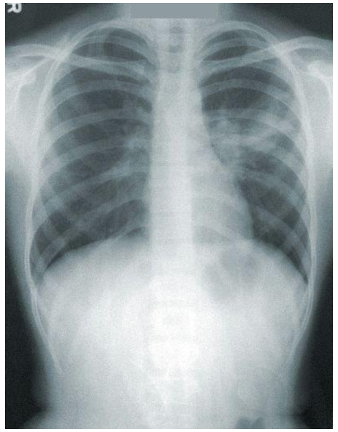
- Automatic detection of feature descriptors of an image (or sub-image)
- SIFT (Scale-Invariant Feature Transform)
  - Features of interest are robust to image translation, scaling, and rotation, and partially invariant to illumination changes, and local geometric distortion

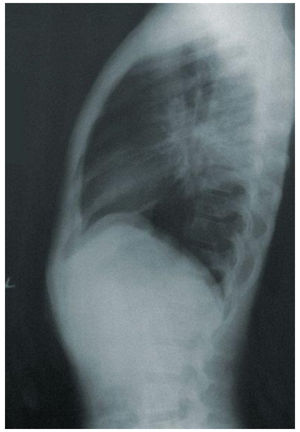


- Other methods:
  - SURF (Speeded Up Robust Features)
  - GLOH (Gradient Location and Orientation Histogram)
  - HOG (Histogram of Oriented Gradients)

Annotations for Machine Learning

Unlabelled data





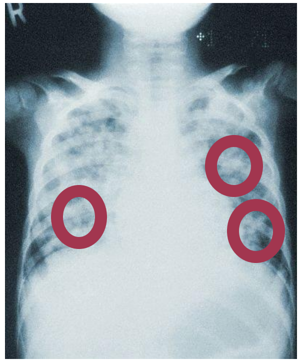
Annotators

Expert

Groups
Crowd

Labelled data

Fracture



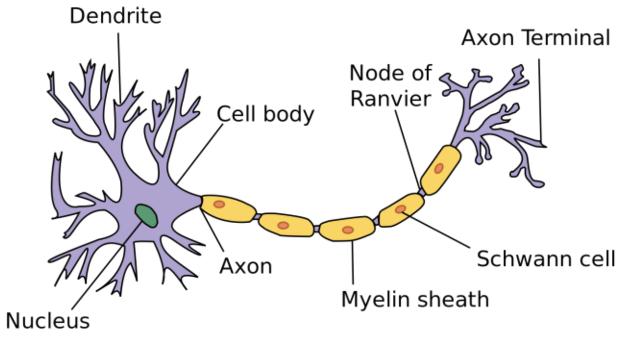
Pneumonia

- Classification
  - Detection
  - Segmentation
- Other tasks
  - Regression
  - Exploratory
    - Clustering
    - Dimensionality Reduction

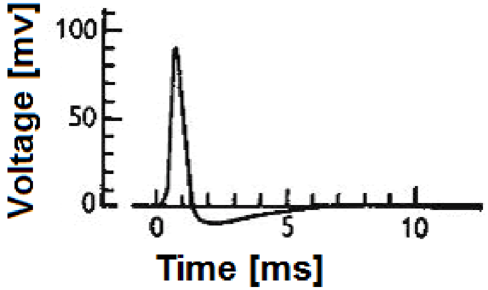
LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

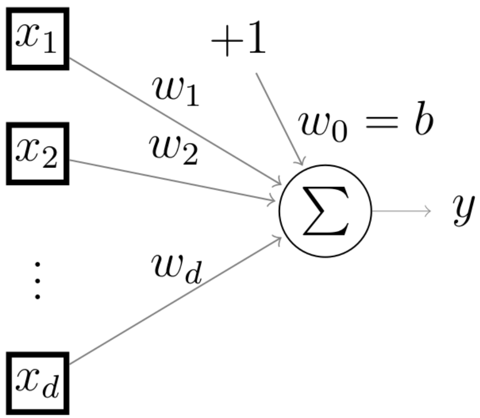
Neurons: bilogical vs artificial



1. Simplified schema of a biological neuron

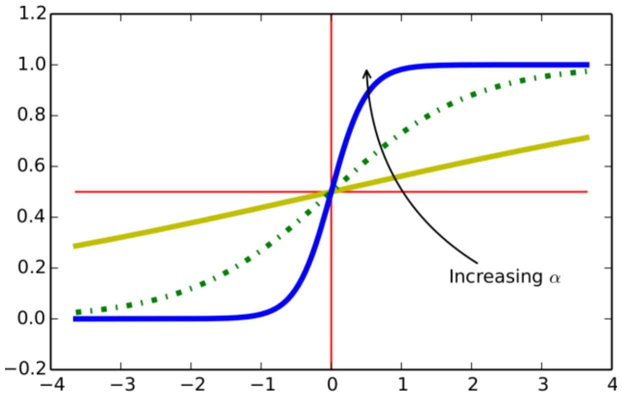


2. Neuronal action potential ("spike")

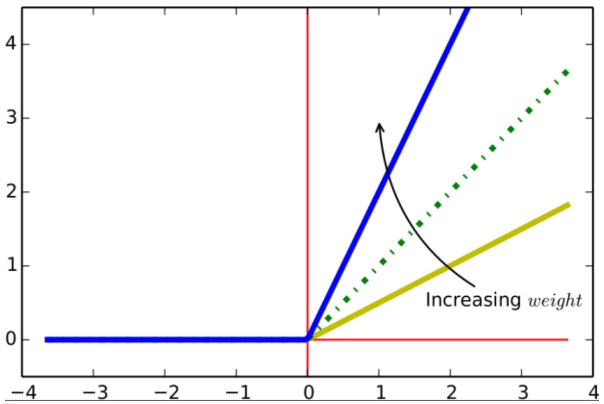


$$a(\mathbf{x}) = b + \sum_{i=1}^{D} w_i x_i$$

Mathematical simplification of a neuron as a weighted sum



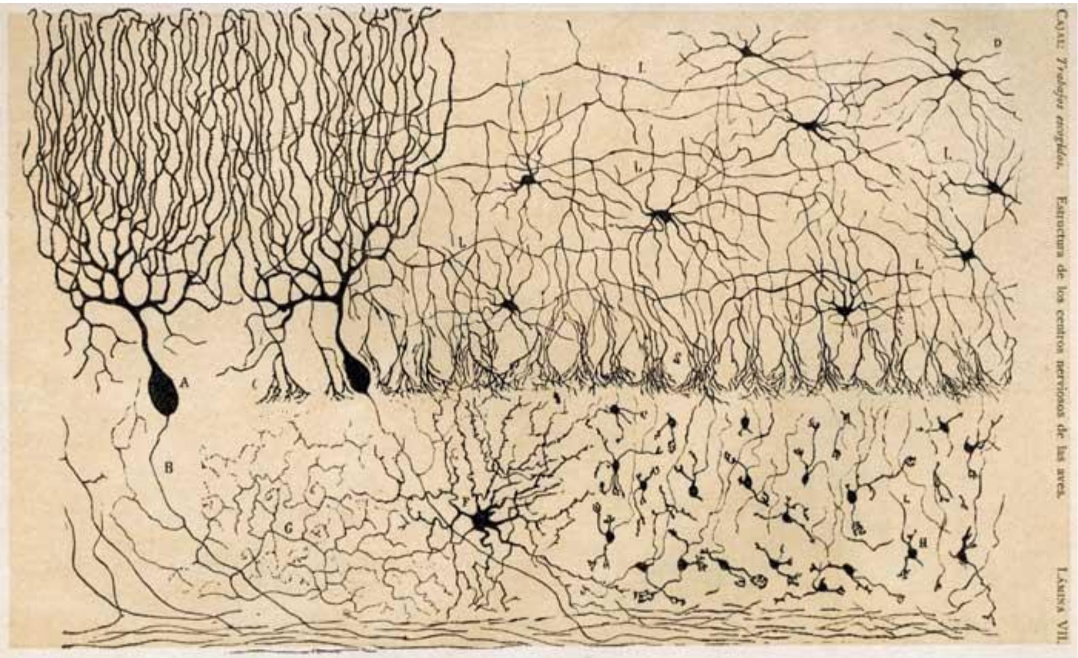
Logistic activation function



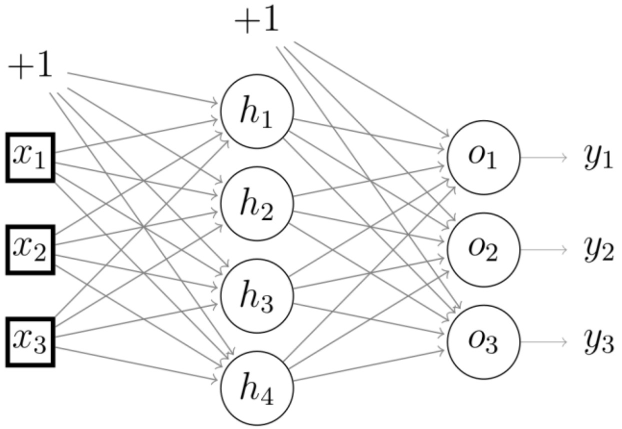
Rectified linear unit

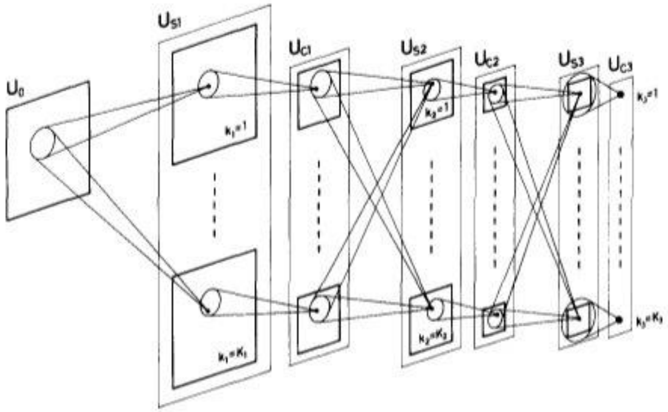
1. Originally Neuron.jpg taken from the US Federal (public domain) (Nerve Tissue, retrieved March 2007), redrawn by User:Dhp1080 in Illustrator. Source: “Anatomy and Physiology” by the US National Cancer Institute’s Surveillance, Epidemiology and End Results (SEER) Program.
2. By Nir.nossenson - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=48019779

Neural Networks



Drawing of neurons by Santiago Ramón y Cajal (around 1890s)

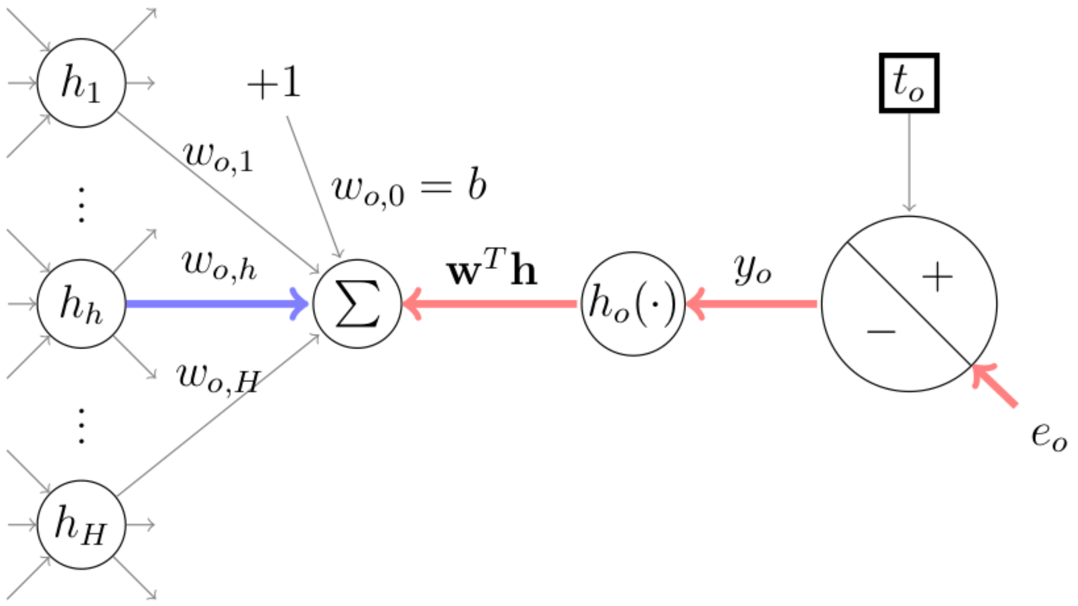




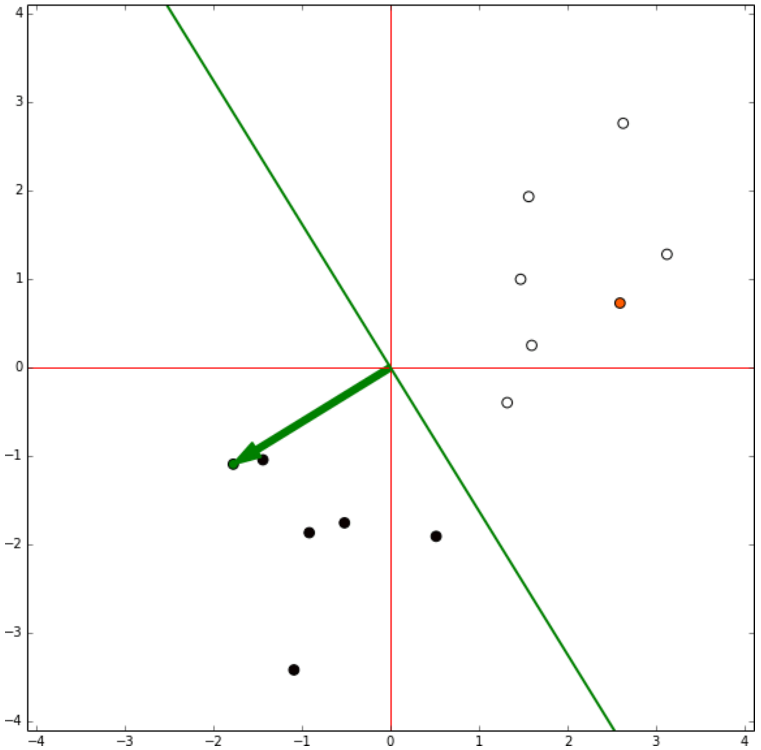
Neocognitron schematic diagram illustrating the interconections between multiple layers [Fukushima, 1980]

Learning procedure: Backpropagation

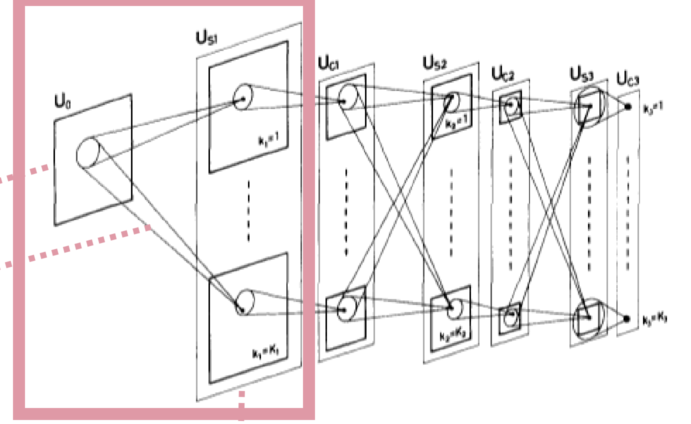
Example of how a prediction error during training propagates a loss that changes the strengths of the weights on a multilayer artificial neural network

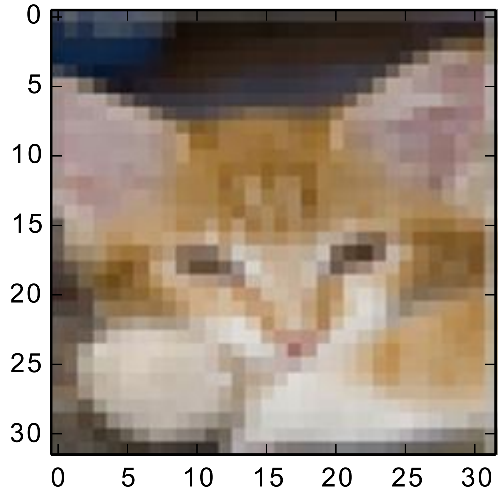


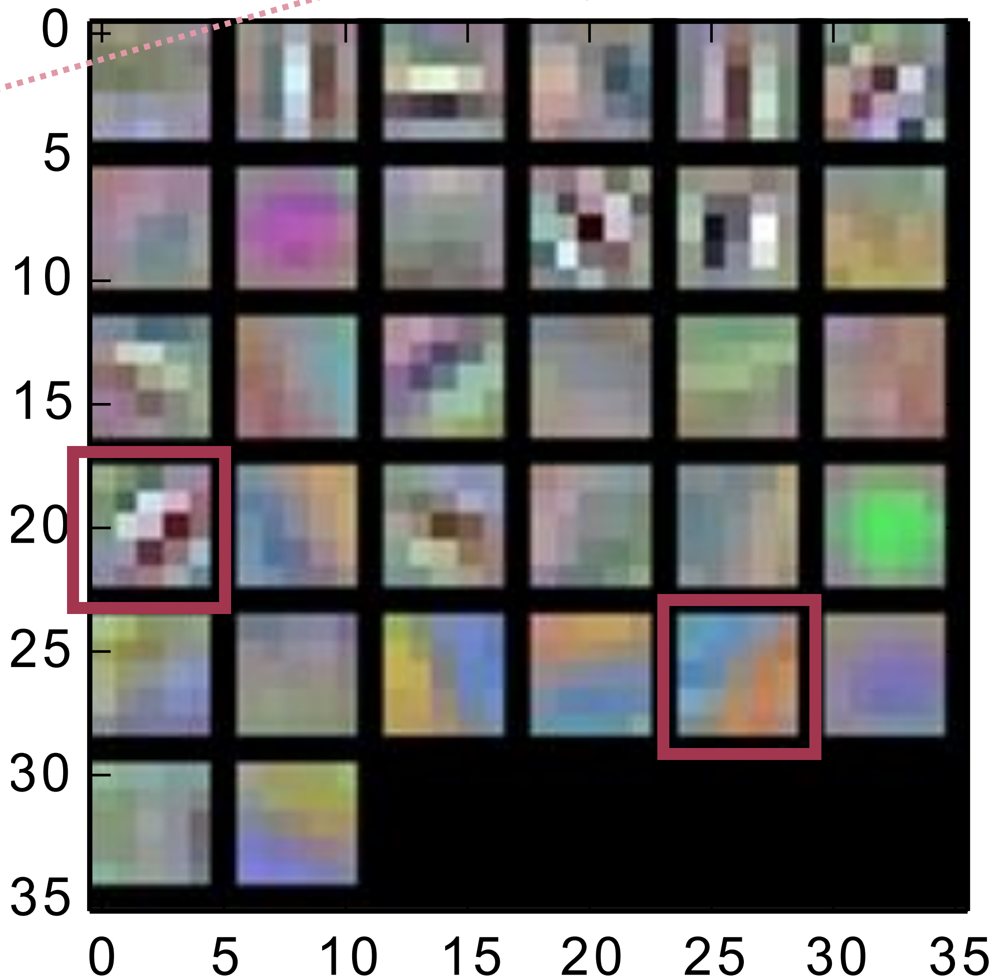
Toy example of the learning process to classify between white and black samples.

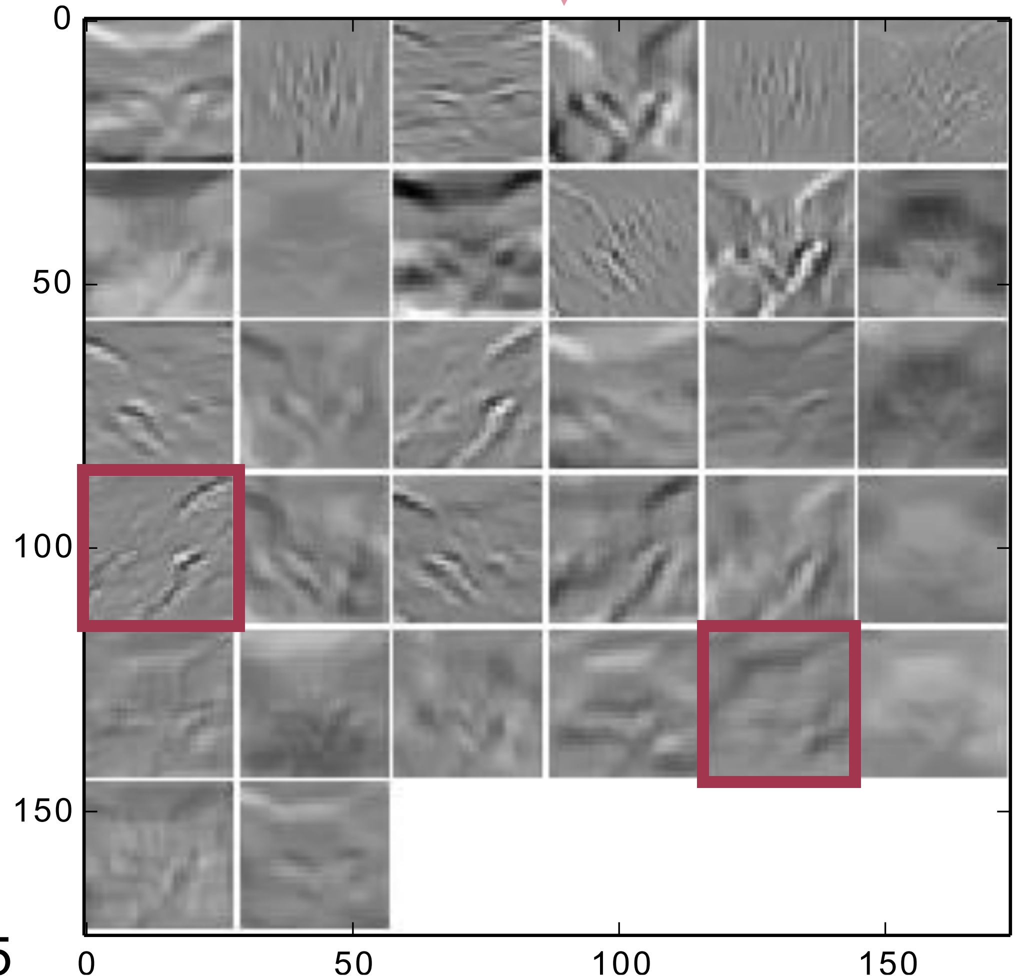


Learned filters



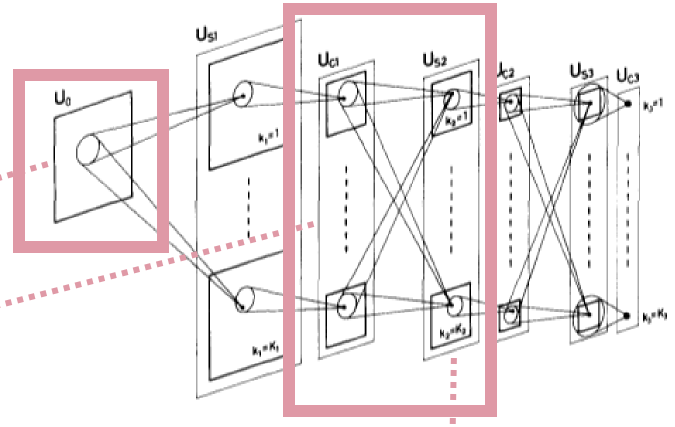


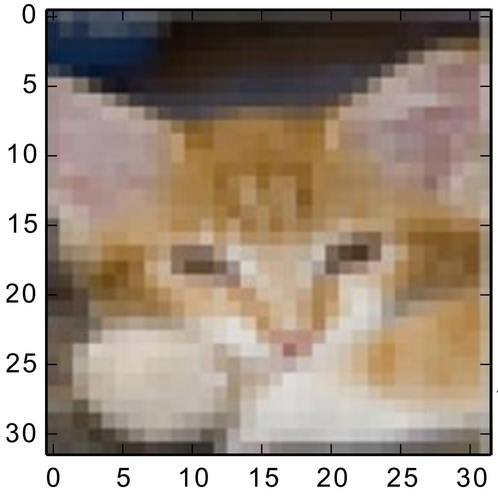


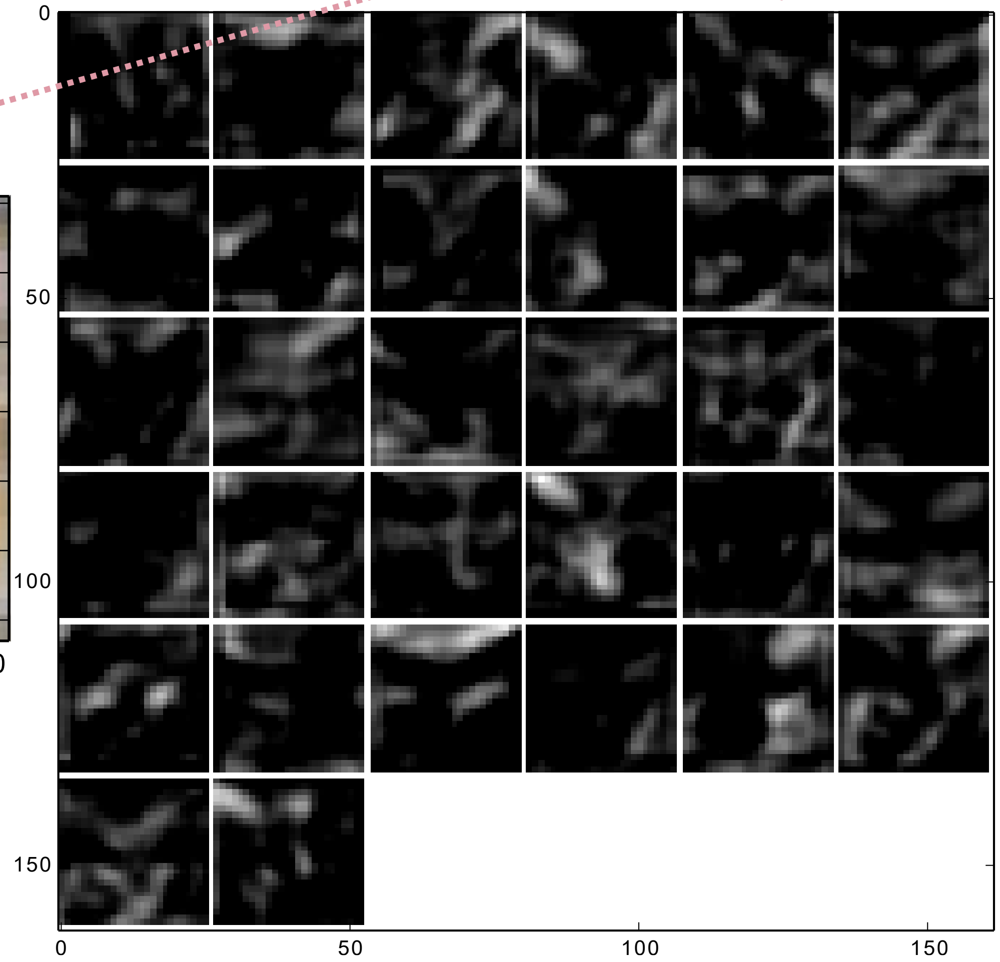


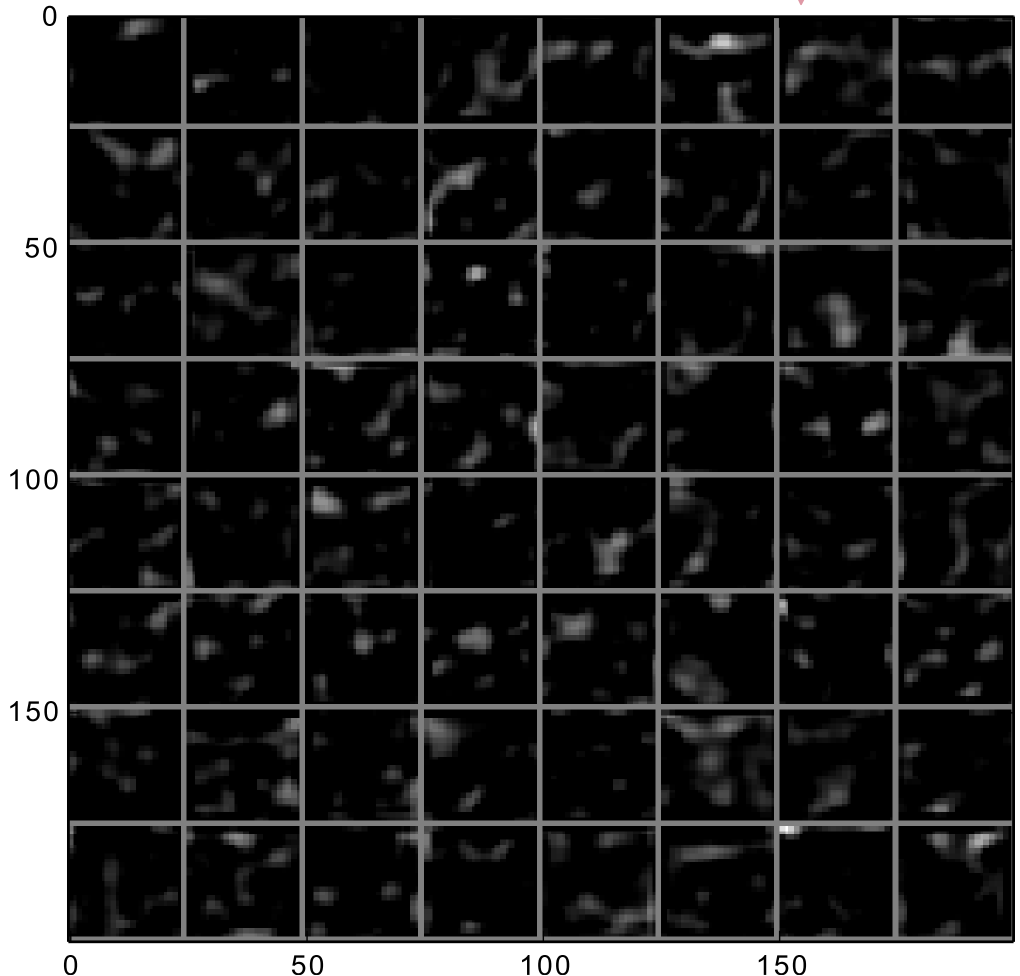
LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

More complex features

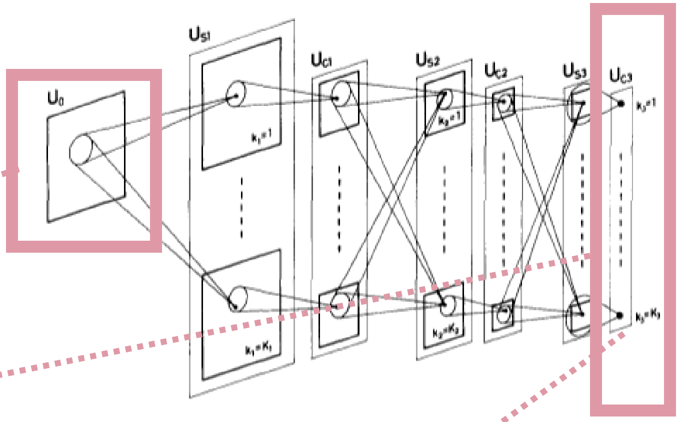


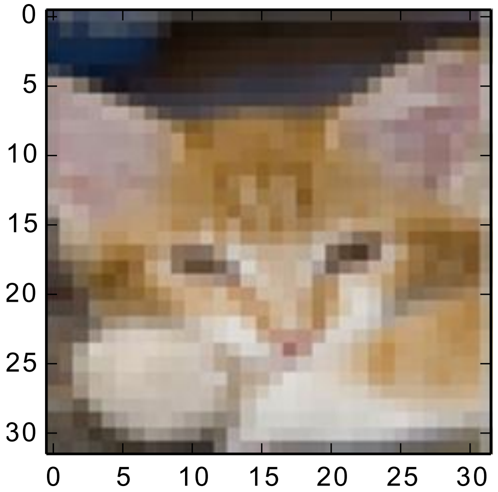


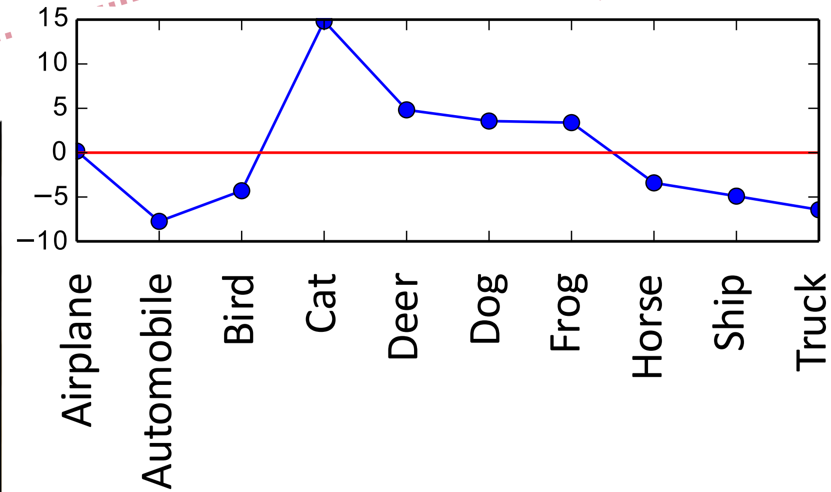




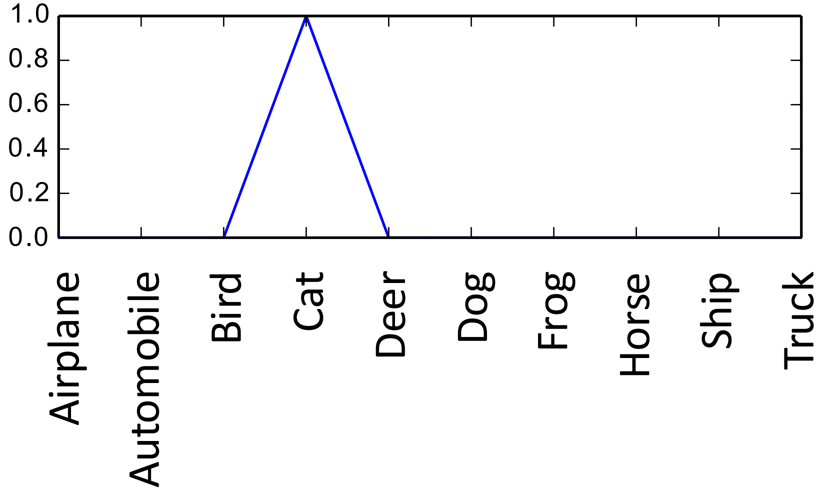
Final prediction





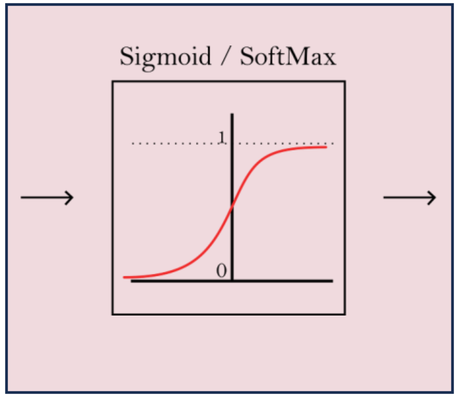


$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$



Input

$X \in \mathbb{R}$ $\longrightarrow$



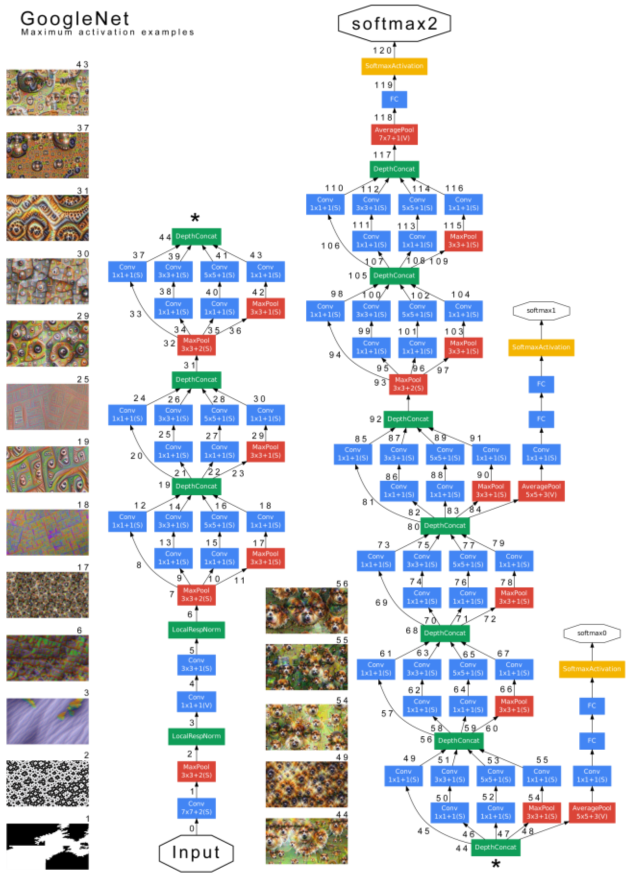
$\longrightarrow$

Output

$P(Y=k \mid X) \in [0,1]$

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

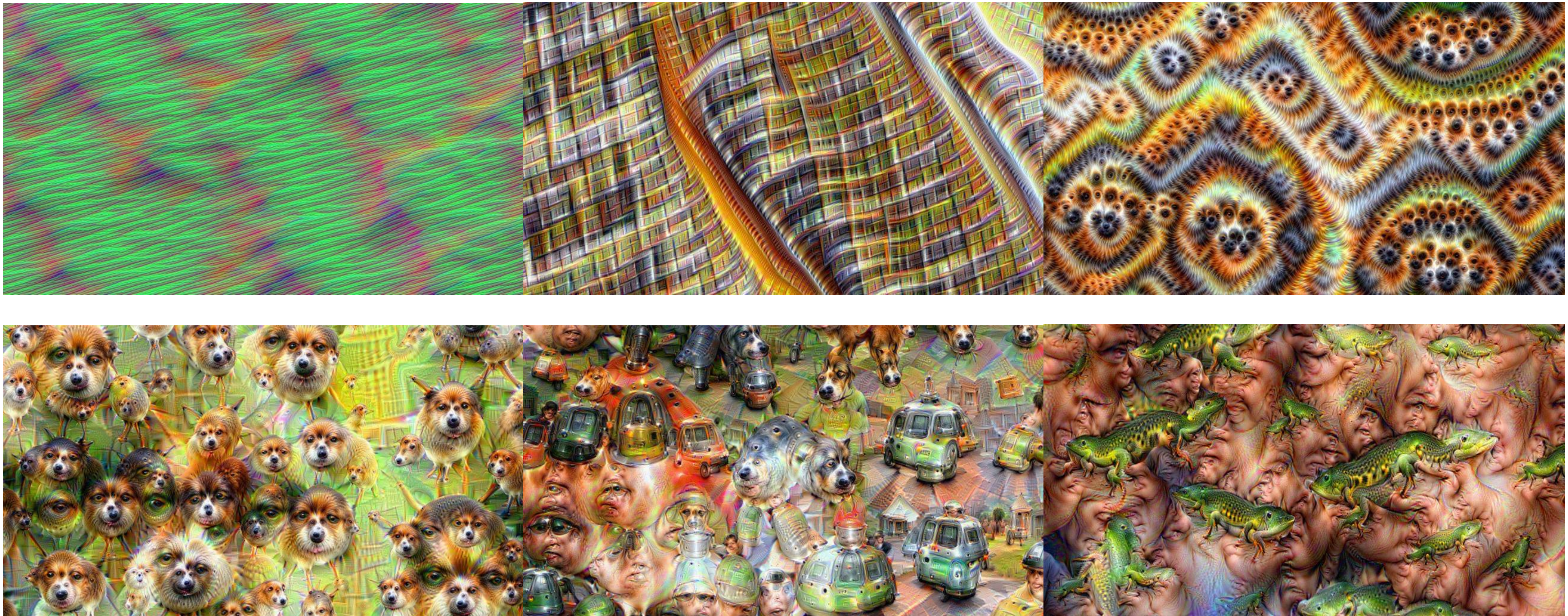
Deep Neural Network



- VGG-16: 16 layers
- GoogleNet: 22 layers DNN [Szegedy et al., 2014]
- ResNet: 34 layers
- ResNet-50: 50 layers
- …

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

What different layers represent



A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

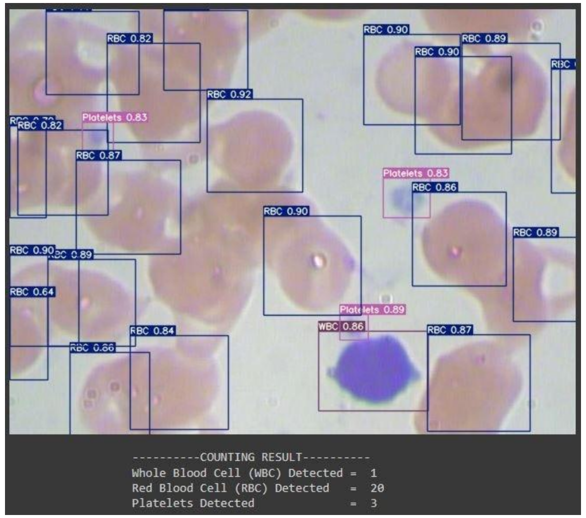
YouTube: DeepDream through all the layers of GoogleNet

LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Radiology and other image analysis

- Plenty of AI methods for image analysis
- Each with its own drawbacks
- AI should support experts
- Constant monitoring is necessary
- Provide explanations under request



Blood dell counter YoLo v7, MIT license, https://github.com/mmasdar/Blood-Cell-Counter-YoLo-v7

LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

3. Demystifying AI

a. ML Pipeline
b. Radiology
c. Scribes
d. Pitfalls

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

AI scribes

- Along side your Electronic Health Record (EHR), while you hold consutations with your patients
    - Listens and transcribes
    - Summarisation
    - Task creation
    - Drafts notes, letters and clinical codes
    - Some have a chatbot
- Lots of different tools:
    - Tortus: Self-registered as a class 1 medical device
    - Heidi
    - Anima:
    - Kiwipen
    - Antikit AI: Chatbot for GPS, uses the British Medical Association and the NHS guidance

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Scribes

- Voice to text in real time
- Summarisation of the text
- Use of additional text information (e.g., EHR)
- Help writing a reference letter
- Sometimes Chat functionalities

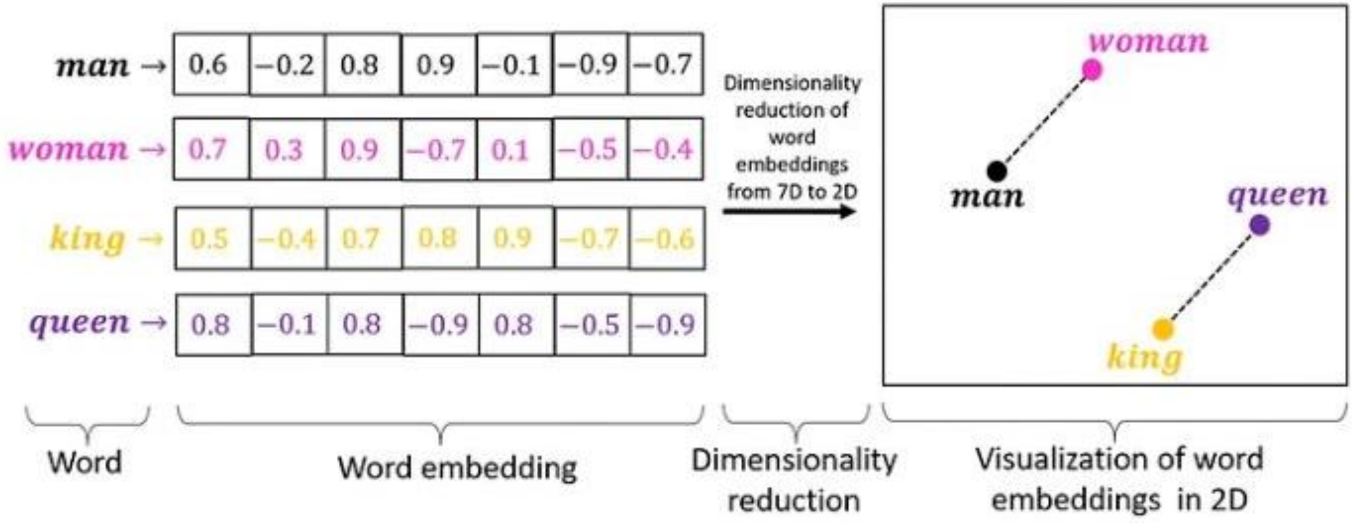
Most parts probably done with Large Language Models with prompt engineering

Large Language Models (Chat GPT)

- **G**enerative: Designed to **g**enerate the next "word".
- **P**re-trained: **P**re-trained with bast amounts of text.
- **T**ransformer: A **T**ransformer architecture to iteratively assign importance to the "words" in the text.

Learning the meaning of words

- Training the model to predict the new word with limited information forces the model to learn a representation of the meaning of words



king – man + woman ~= queen

sushi – japan + germany ~= bratwurst

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Text generation

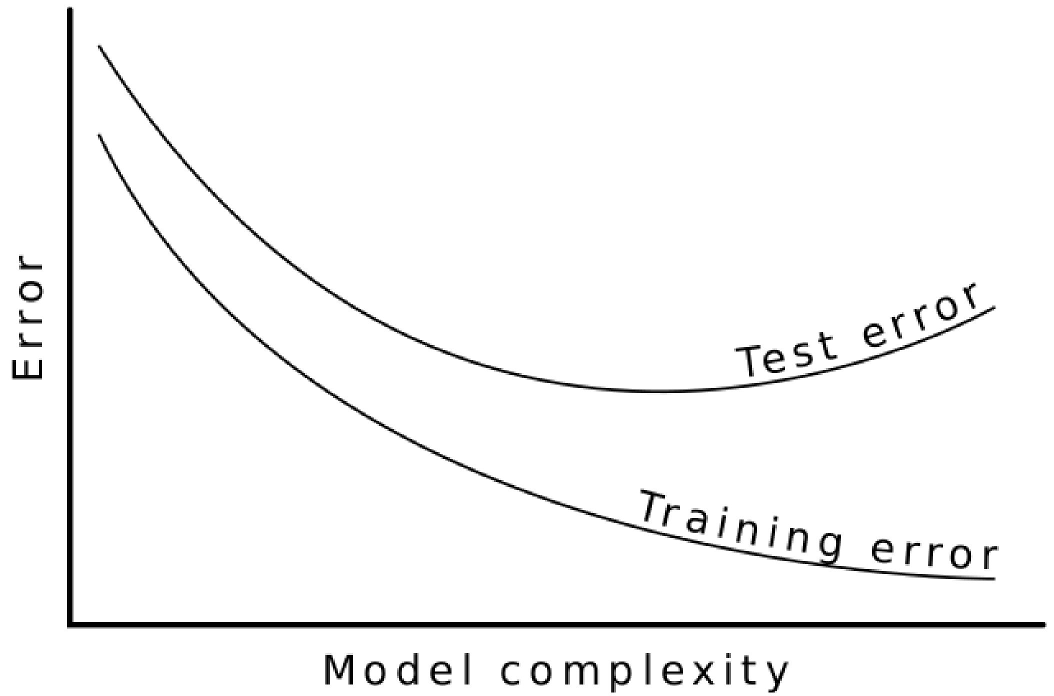
- The huge amount of training has a multitude of context
- Prompts provide a context to "reduce the search area"
- In any particular context the text generation will be different
- It does not need to find exact matches, as it has learned the "meaning" of words
- What it generates looks like natural language, but it is debatable if there is any type of thought process

3. Demystifying AI

a. ML Pipeline
b. Radiology
c. Scribes
d. Pitfalls

Generalization error

- A complex model can overfit the training data
- The model could memorise the available data
- However, it may not generalize to the deployment population



Misinterpreting the model's output

- How to interpret the model output?
  - Regression, class prediction, ranking, probabilities, sets

Example:

Calibrated probabilities approximate the proportion of correct predictions for each class

A weather forecast of 80% chance of rain, should be followed by rain 80% of the time.

It applies to multiclass problems (e.g., three medical outcomes):

A prediction of **s**=*[0.1, 0.2, 0.7], should be followed by*

10% from the first class, 20% from the second class, and 70% from the third class

Bennefits of calibrated probabilities

A calibrated classifier correctly quantifies the level of **uncertainty** or confidence associated with its predictions
(**Important:** assuming independent and identically distributed random variables (i.i.d.))

Cost Matrix	Predicted True	P. False
Actual True	-1£	20£
Actual False	5£	0£

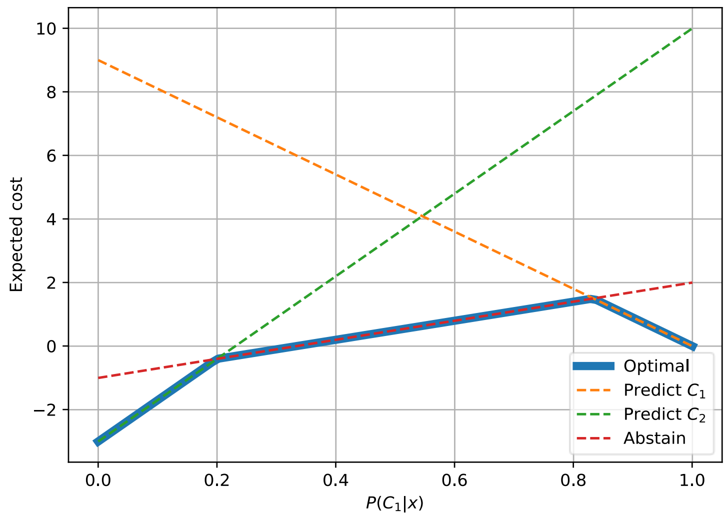
Population proportion
40%
60%

Changes on costs

Cost Matrix	Predicted True	P. False	**Abstain**
Actual True	0£	10£	2£
Actual False	9£	-3£	-1£

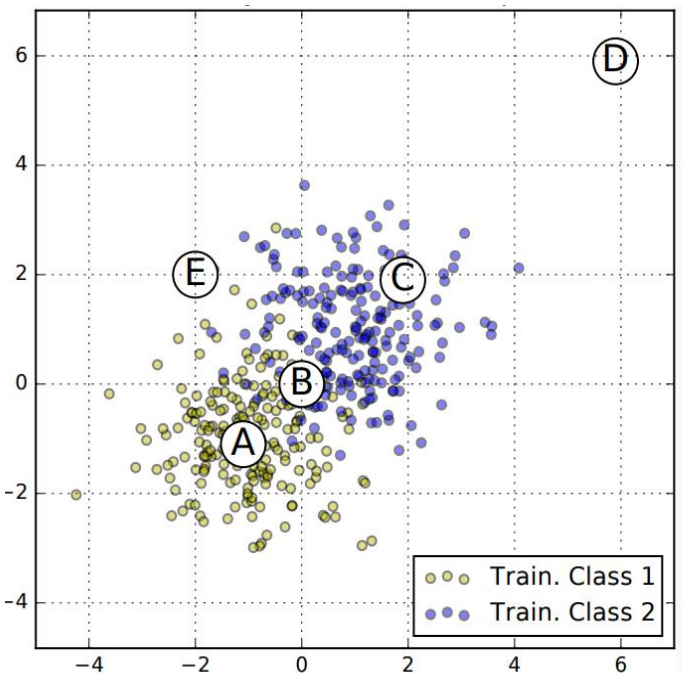
Changes on class proportions

Population proportion
50%
50%

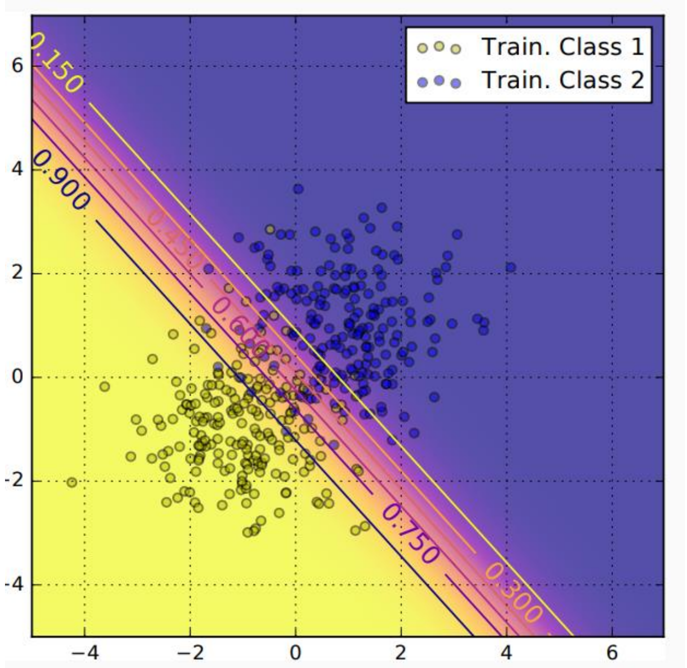


Can my model abstain?

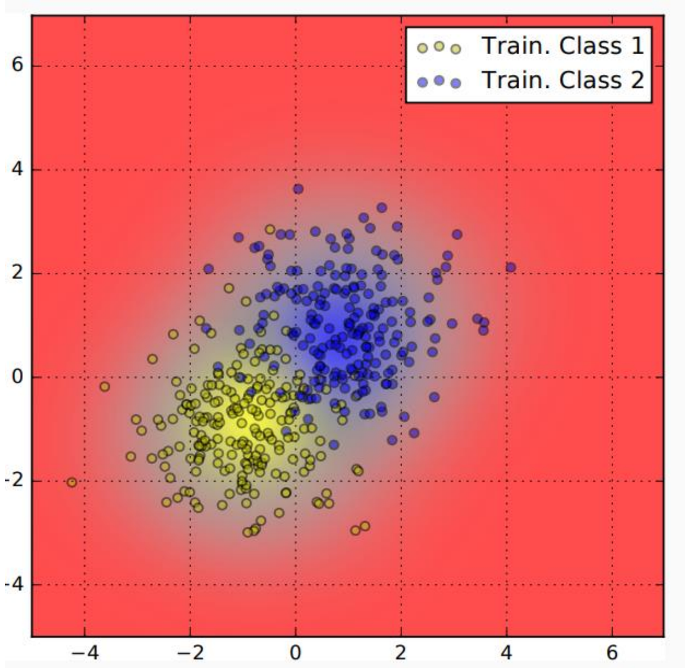
Imagine a simple classification task (e.g. height and weight by country)



Most classification models assume the same distribution during deployment



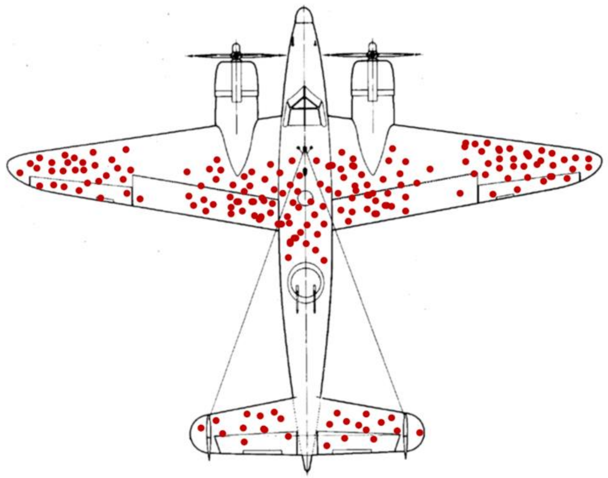
We may need models that have the option of abstaining if the query is too different



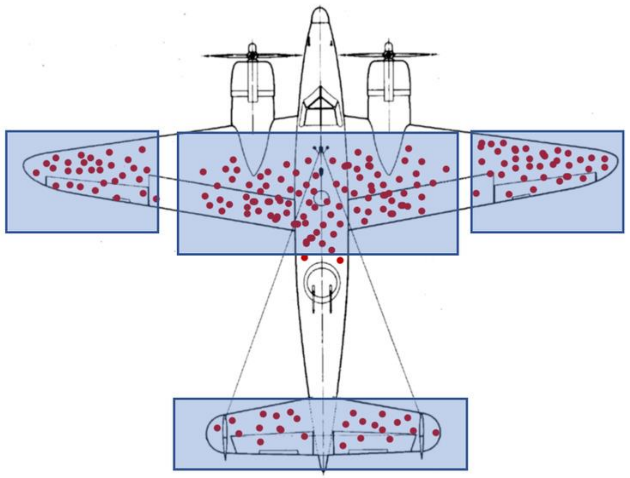
Perello-Nieto, Miquel, et al. "Background Check: A general technique to build more reliable and versatile classifiers." *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 2016.

The survivorship bias

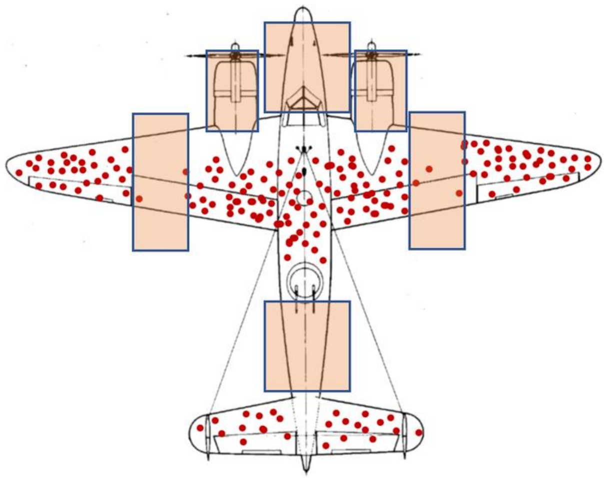
- Hypothetical damage patter on a WW2 bomber.



Bullet holes observed from returning flights.



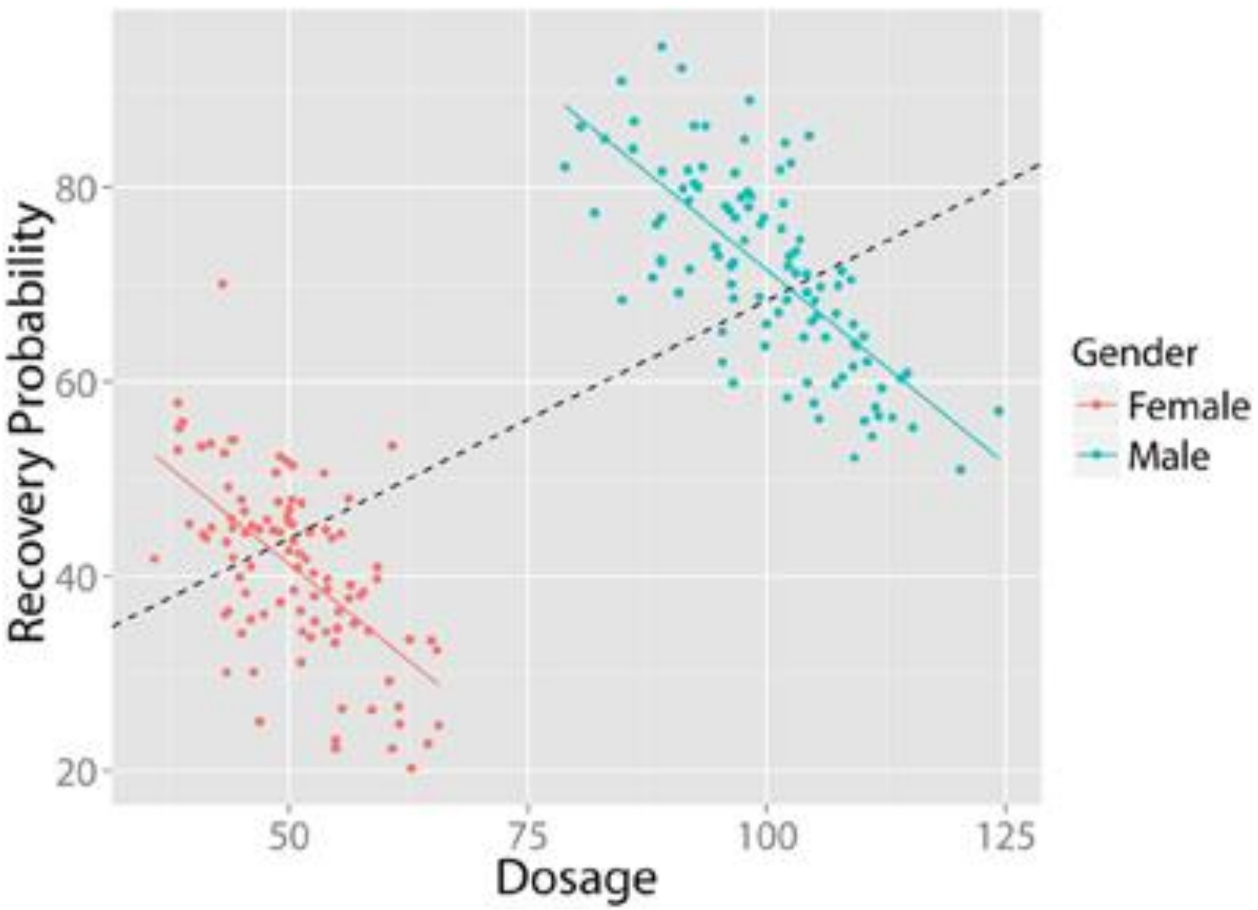
Potential areas to reinforce with additional armour covering the holes...



... the planes that didn't return probably got holes in more crucial areas.

- Are we using data from the patients that need the help?

The Simpson's paradox



Kievit Rogier , Frankenhuis Willem E., Waldorp Lourens , Borsboom Denny, Simpson's paradox in psychological science: a practical guide, Frontiers in Psychology, 2013

LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

4. Ethics and Regulations

a. Regulations
b. Privacy
c. Explainability
d. Fairness
e. Accountability
f. Contestability

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Healthcare regulations affecting AI

- Market regulations:
  - EU AI Act
  - EU Digital Services Act
  - EU Digital Markets Act
  - EU Cyber Resilience Act
- Biopharma regulations:
  - European Health Data Space
  - General Pharmaceutical Legislation
  - Clinical Trial Regulations
- Data regulations:
  - GDPR (General Data Protection Regulation)
  - Data Act
  - Data Governance Act

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

AI ethics and regulations

- Five key principles for regulatory use of AI for medical products

1
Safety, security and robustness

2
Transparency and explainability

3
Fairness

4
Accountability and governance

5
Contestability and redress

Policy paper. Impact of AI on the regulation of medical products. Medicines & Healthcare products Regulatory Agency, GOV.UK (Apr. 2024)

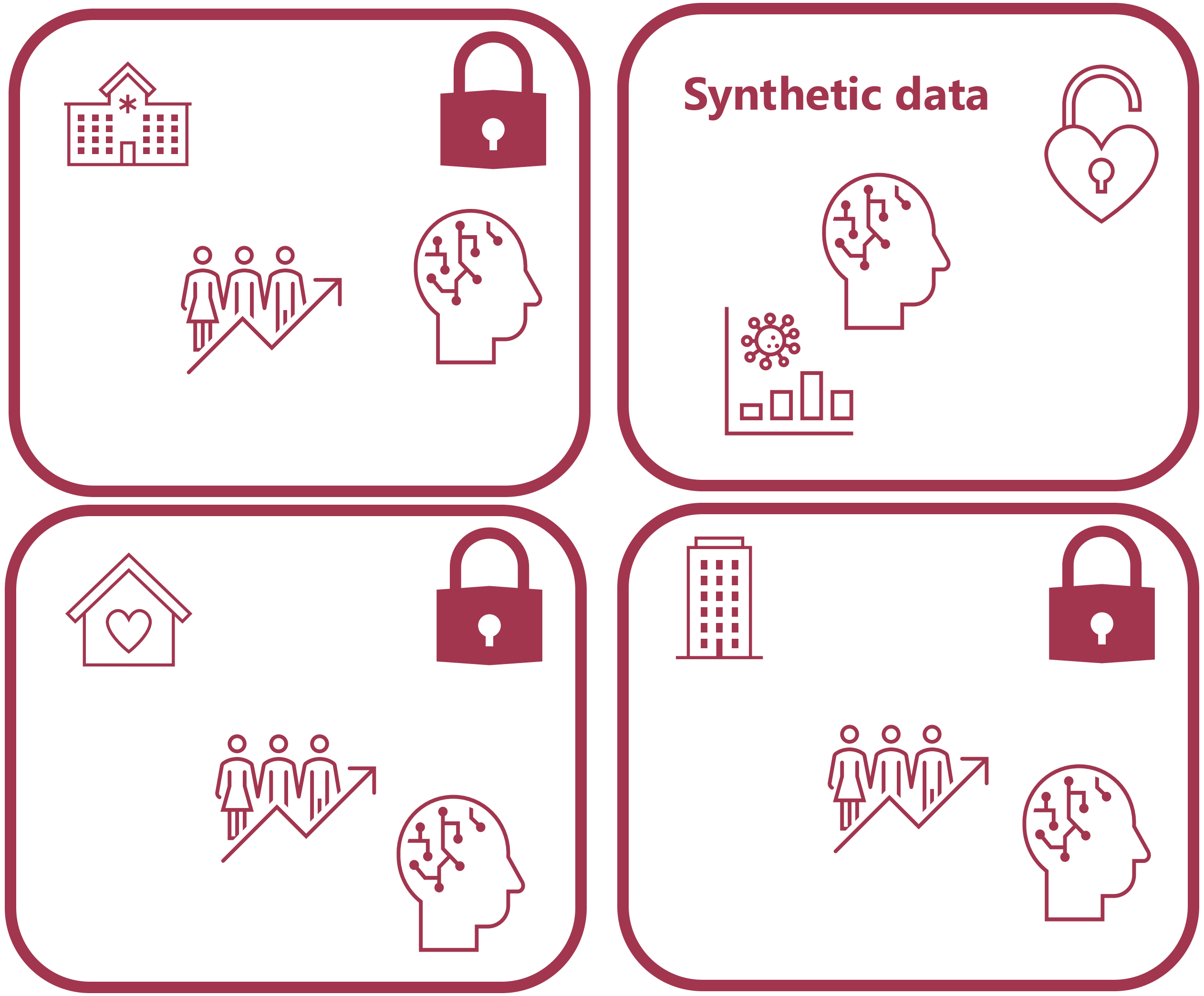
LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

4. Ethics and Regulations

a. Regulations
b. **Privacy**
c. Explainability
d. Fairness
e. Accountability
f. Contestability

Safety and security



- Federated learning
- Synthetic data
- Foundation models
- Differential privacy

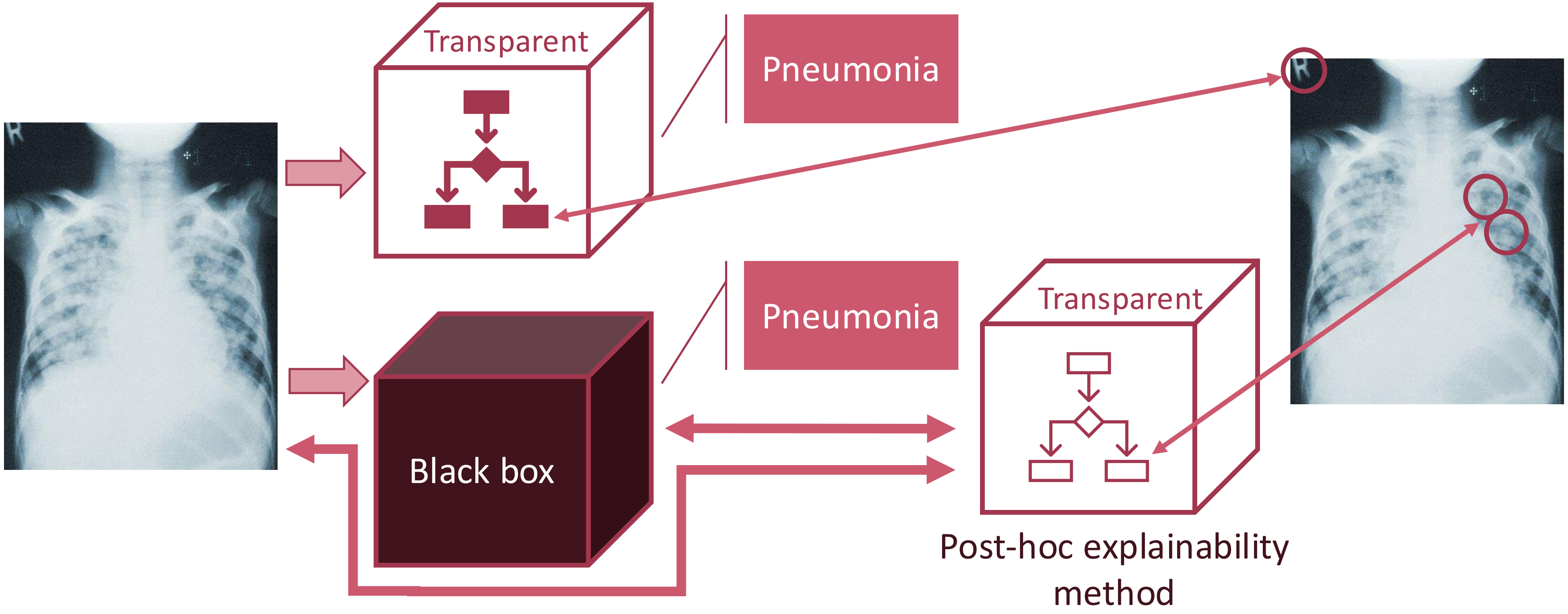
Requirements:

- Standardisation
- Infrastructure

4. Ethics and Regulations

a. Regulations
b. Privacy
c. **Explainability**
d. Fairness
e. Accountability
f. Contestability

Transparency and explainability



- Identify dataset biases, or model problems
- Model complexity vs transparency vs performance

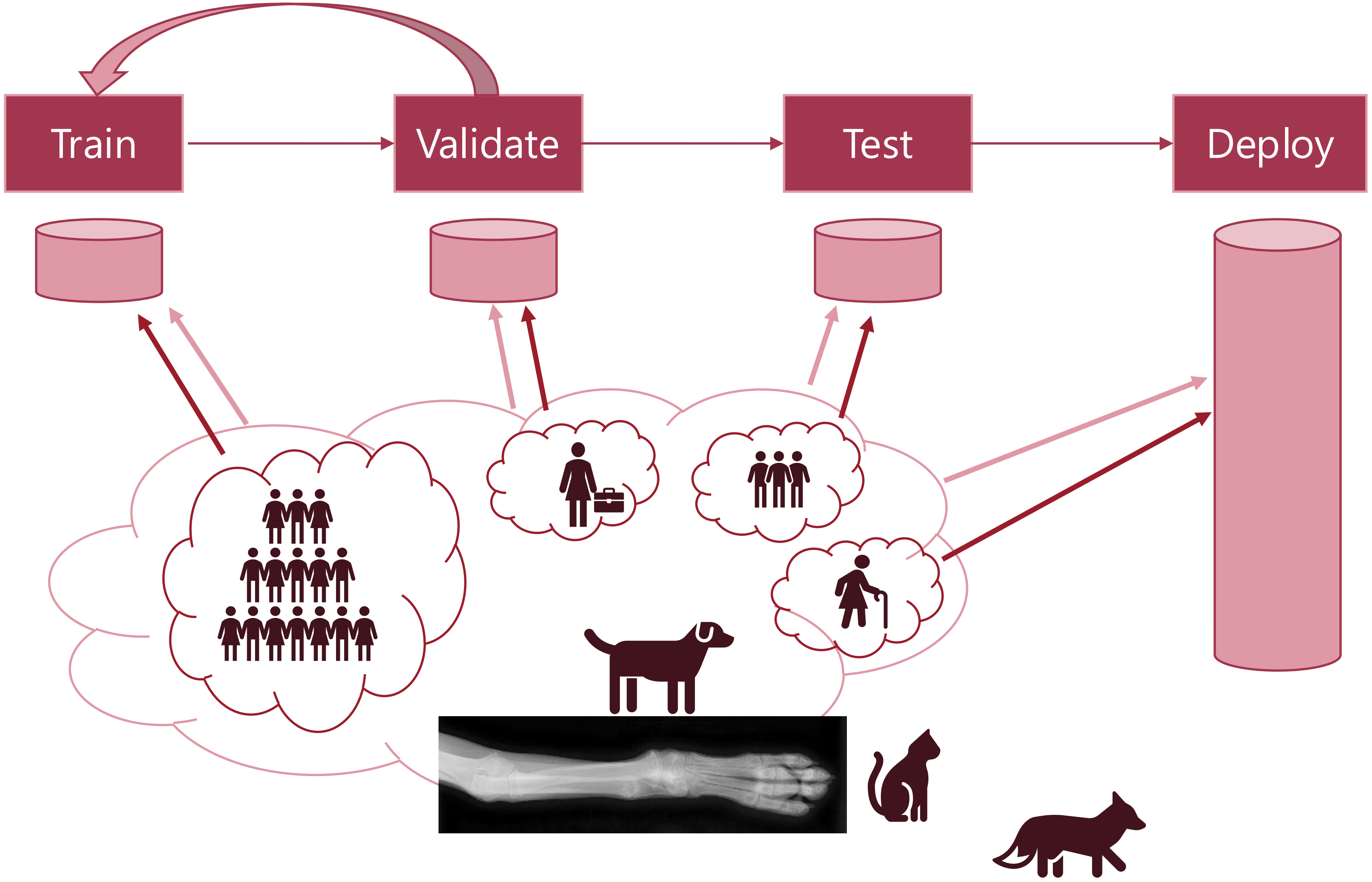
LEAP

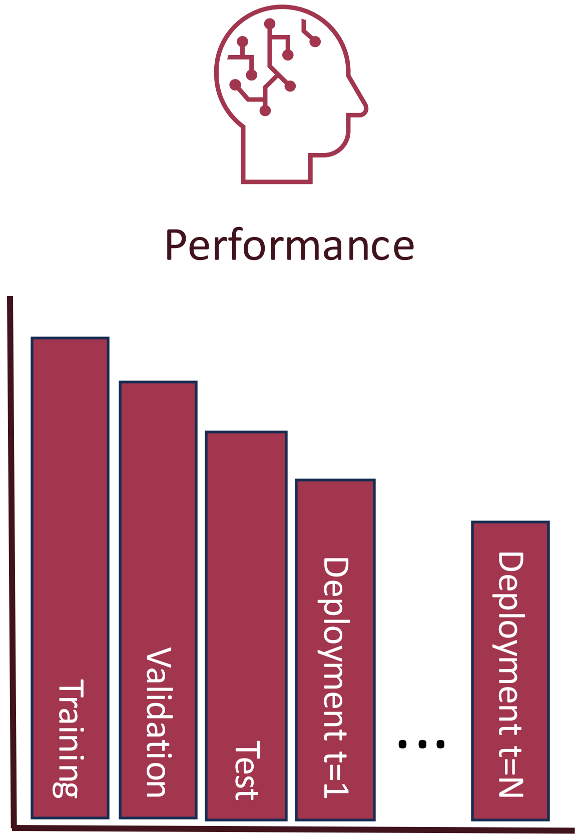
A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

4. Ethics and Regulations

a. Regulations
b. Privacy
c. Explainability
d. **Fairness**
e. Accountability
f. Contestability

Fairness and robustness





Requires constant monitoring

LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Fairness

Mail Online Science

Home | News | Royals | U.S. | Sport | TV | Showbiz | Femail | Health | Science | Money | Tra

Latest Headlines | Blue Origin | SpaceX | NASA | Apple | Google | Twitter | Microsoft

Is this soap dispenser RACIST? Controversy as Facebook employee shares video of machine that only responds to white skin

- A Facebook employee tweeted a soap dispenser that only works for white hands
- It's likely because the infrared sensor was not designed to detect darker skin
- Critics say tech's diversity problem causes this and other racist technology

By SAGE LAZZARO FOR DAILYMAIL.COM

PUBLISHED: 18:54, 17 August 2017 | **UPDATED:** 19:32, 18 August 2017

The New York Times

Does Your Teen Recognize A.I.? Art World Takes On A.I. Putting A.I. in Charge A.I. and Hollywood

Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's.

Desiree Rios/The New York Times

Eight years after a controversy over Black people being mislabeled as gorillas by image analysis software — and despite big advances in computer vision — tech giants still fear repeating the mistake.

By Nico Grant and Kashmir Hill

May 22, 2023

LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

4. Ethics and Regulations

a. Regulations
b. Privacy
c. Explainability
d. Fairness
e. **Accountability**
f. Contestability

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Accountability and Governance

- Effective oversight of the use of AI.
- Clear lines of accountability across the AI life cycle.
- Trustworthiness auditing

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

4. Ethics and Regulations

a. Regulations
b. Privacy
c. Explainability
d. Fairness
e. Accountability
f. Contestability

Contestability and redress

- A person affected by the outcomes or a decision from an AI should be able to contest the AI.
- How to rectify and address any harm resulting from an AI decision

LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

5. Explainability

a. Global vs local
b. Model-based
c. Surrogate model
d. Feature-based
e. Example-based

LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Explainable to whom?

- Different stakeholders, different explanations:
  - Patient (end-user):
    - confidence, clarity, fairness, contestability…
  - Doctor (domain expert):
    - support decision-making, trust, understanding, justification…
  - Developer / Scientists:
    - diagnosis, errors, design, model…
  - Researchers / Ethicists:
    - Fairness, biases, generalizability…
  - Executives:
    - Risk assessment, impact, investment, reputation, legal risks…

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

5. Explainability

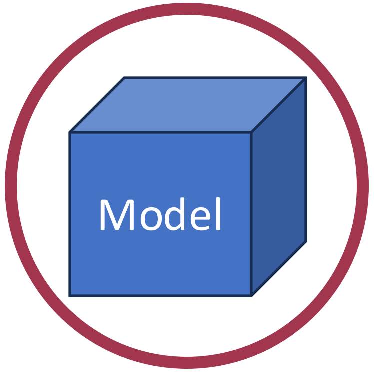
a. Global vs local
b. Model-based
c. Surrogate model
d. Feature-based
e. Example-based

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Global vs local explanations

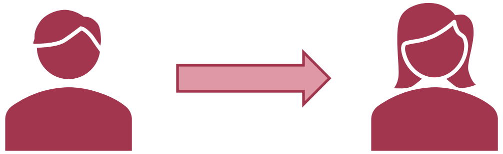
Global

- Explain the overall behaviour of a model
- Examples:
  - Overall feature importance
  - Overall performance of the model



Local

- Explain individual predictions
- Examples:
  - Feature attribution
  - Exemplars and prototypes
  - Counterfactual explanations



LEAP

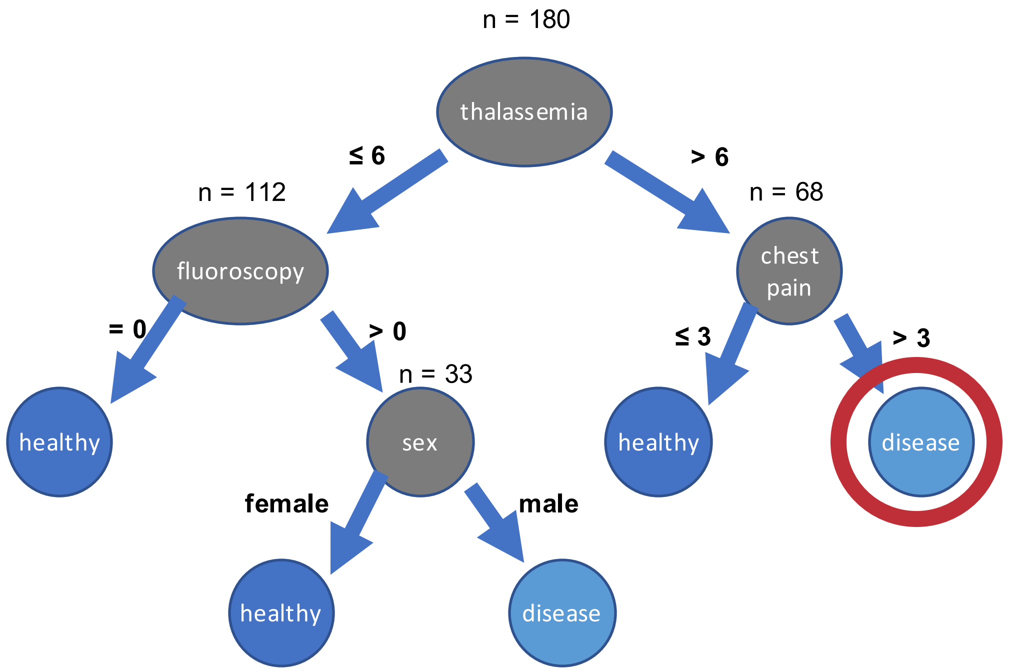
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

5. Explainability

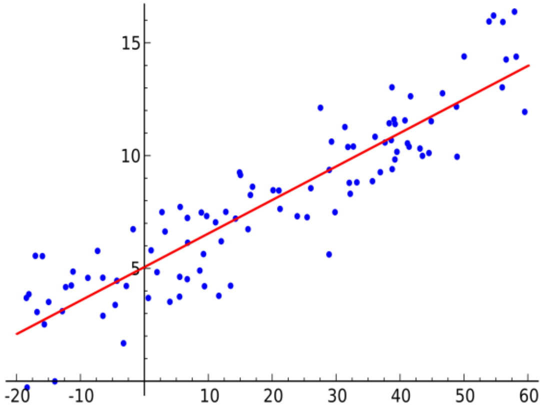
a. Global vs local
b. **Model-based**
c. Surrogate model
d. Feature-based
e. Example-based

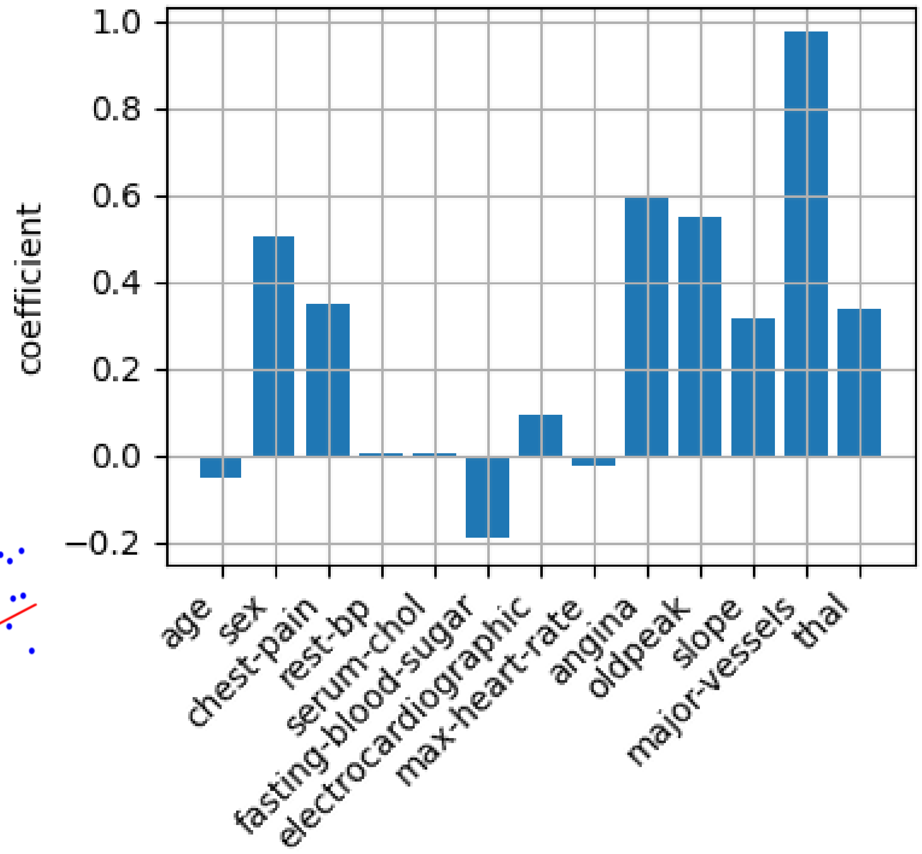
Transparent models

Decision Tree



Linear/Logistic Regression





LEAP
A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

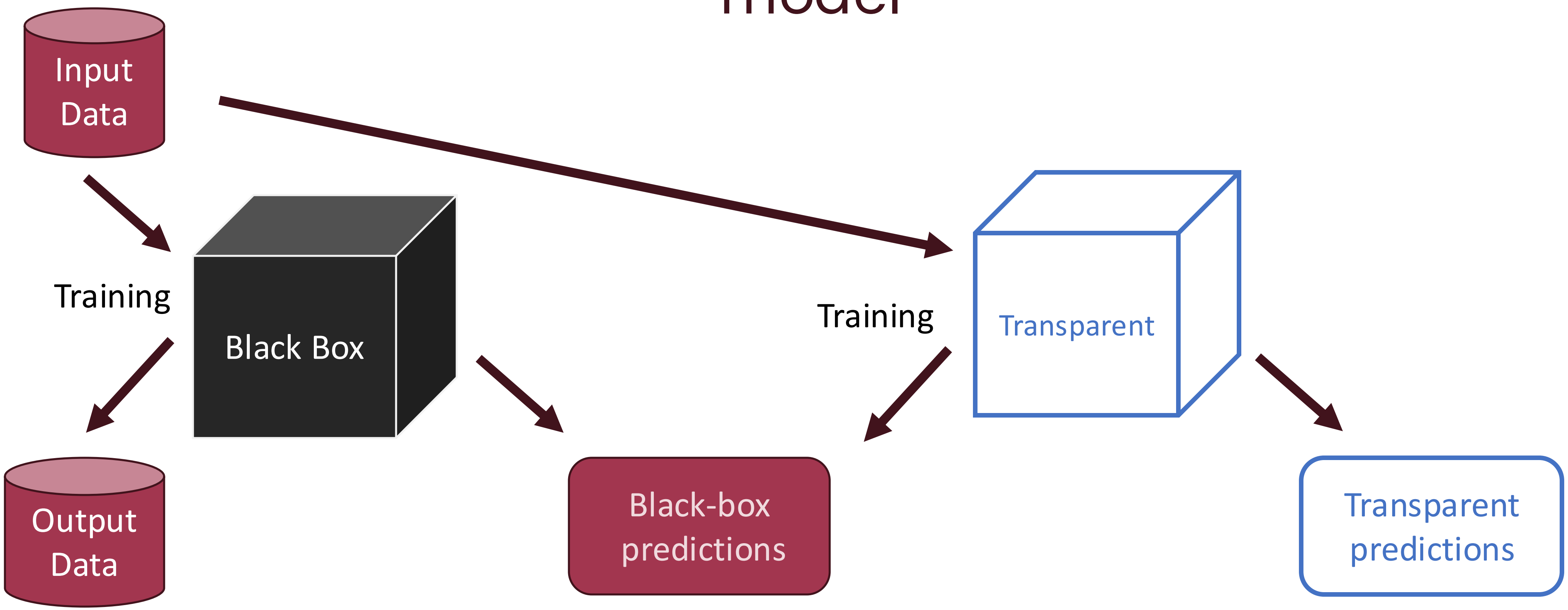
5. Explainability

a. Global vs local
b. Model-based
c. **Surrogate model**
d. Feature-based
e. Example-based

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Surrogate model

- Given a pre-trained black-box model
- Train a transparent model to simulate the black-box model



LEAP

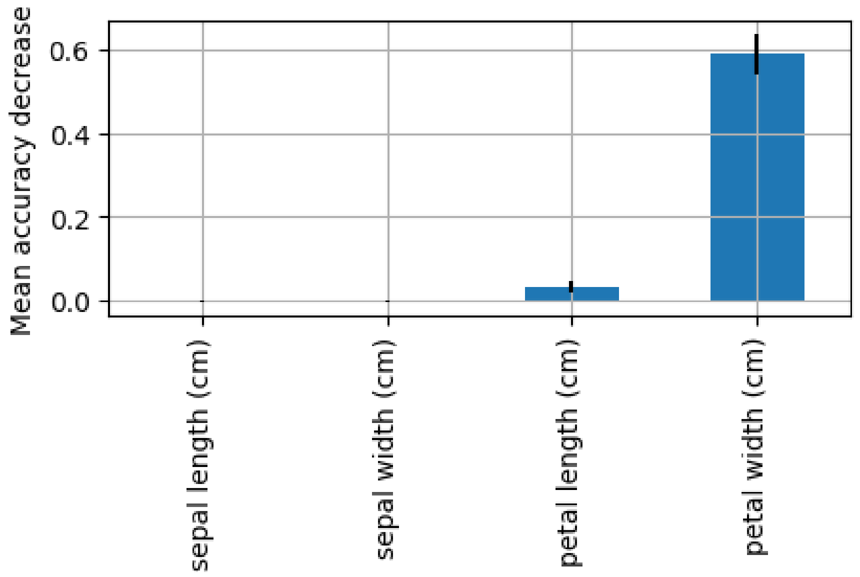
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

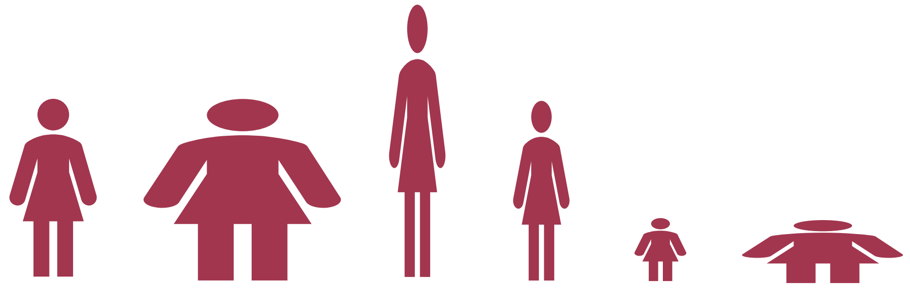
5. Explainability

a. Global vs local
b. Model-based
c. Surrogate model
d. **Feature-based**
e. Example-based

PFI (Permutation Feature Importance)

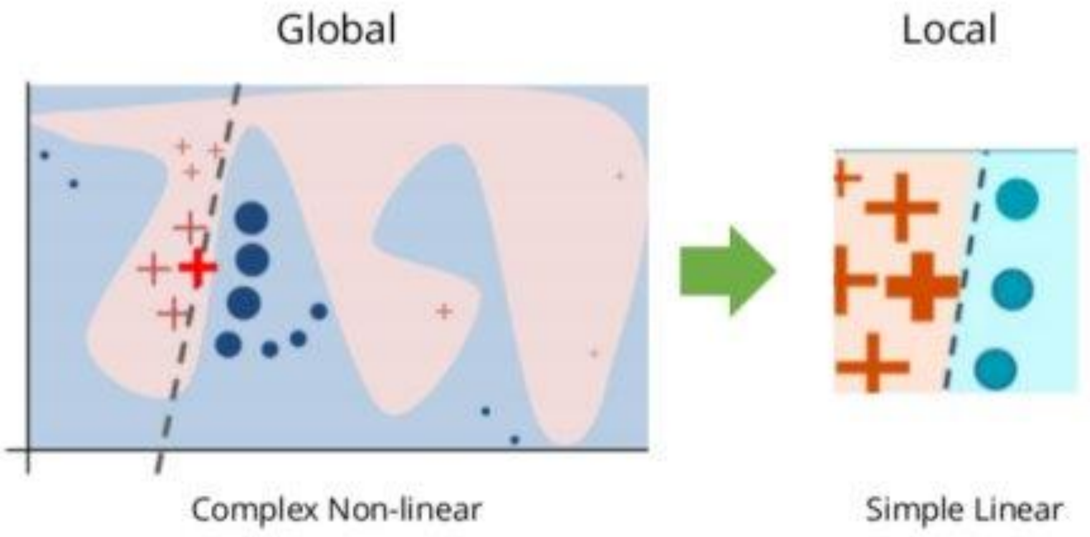
- Originally introduced for Random Forests, and made model agnostic in 2019 with the name Model Reliance.
- It requires a scoring metric to evaluate the model's performance.
- One feature to be assessed is selected and its values are randomly permuted.
- The performance of the model is compared between the original and the permuted dataset.



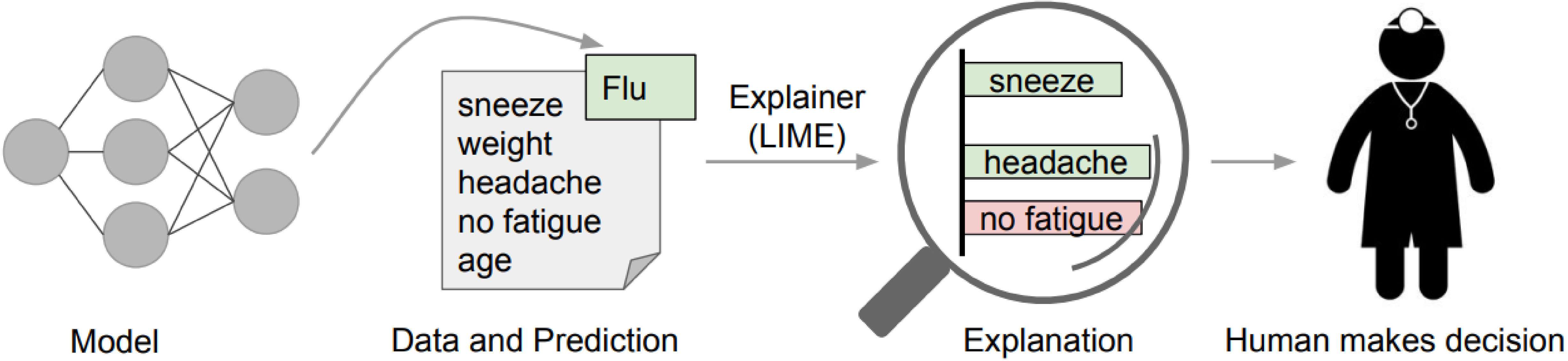


LIME (Local Interpretable Model-Agnostic Explanations)

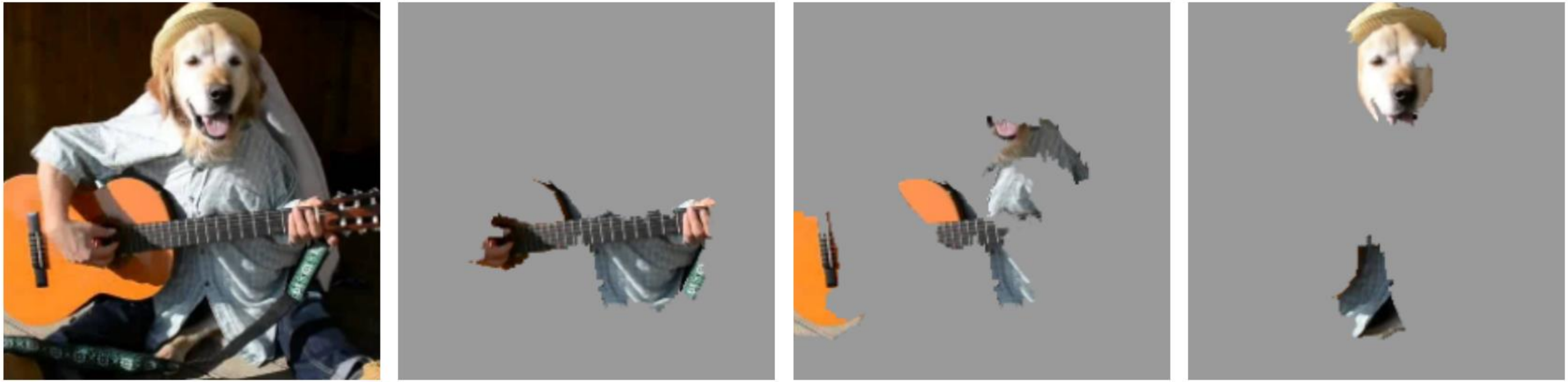
- Aims at producing locally consistent explanations
- Learns a model around the input and its prediction by the base model
- Uses a representation that is understood by humans:
  - For text: presence/absence of words
  - For images: presence/absence of superpixels (contiguous patch of similar pixels)
  - For tabular data: weighted combination of columns



LIME (Local Interpretable Model-Agnostic Explanations)



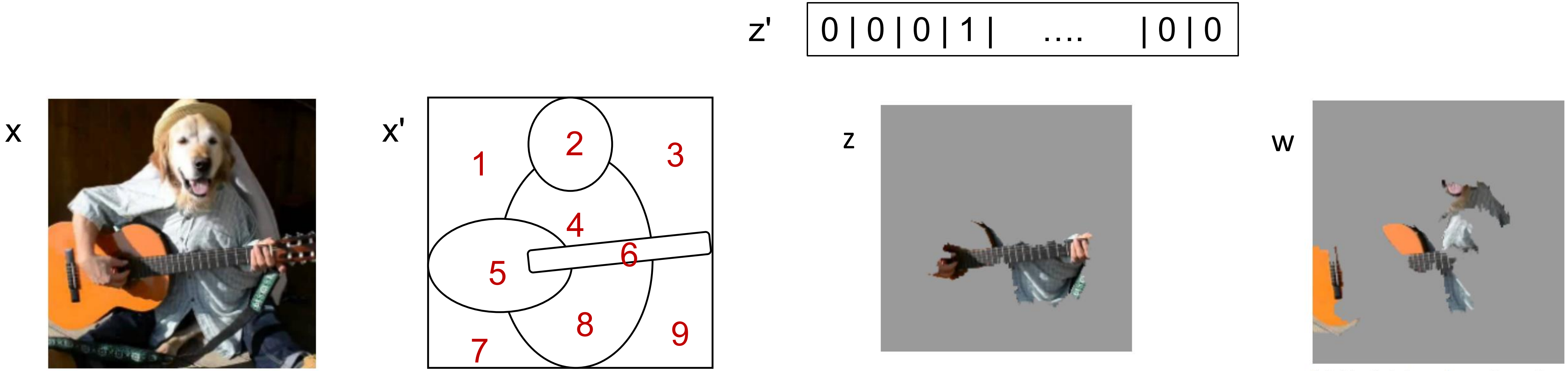
LIME (Local Interpretable Model-Agnostic Explanations)



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$)**, "Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

LIME (Local Interpretable Model-Agnostic Explanations)



(a) Original Image

(c) Explaining *Acoustic guitar*

SHAP (SHapley Additive exPlanations)

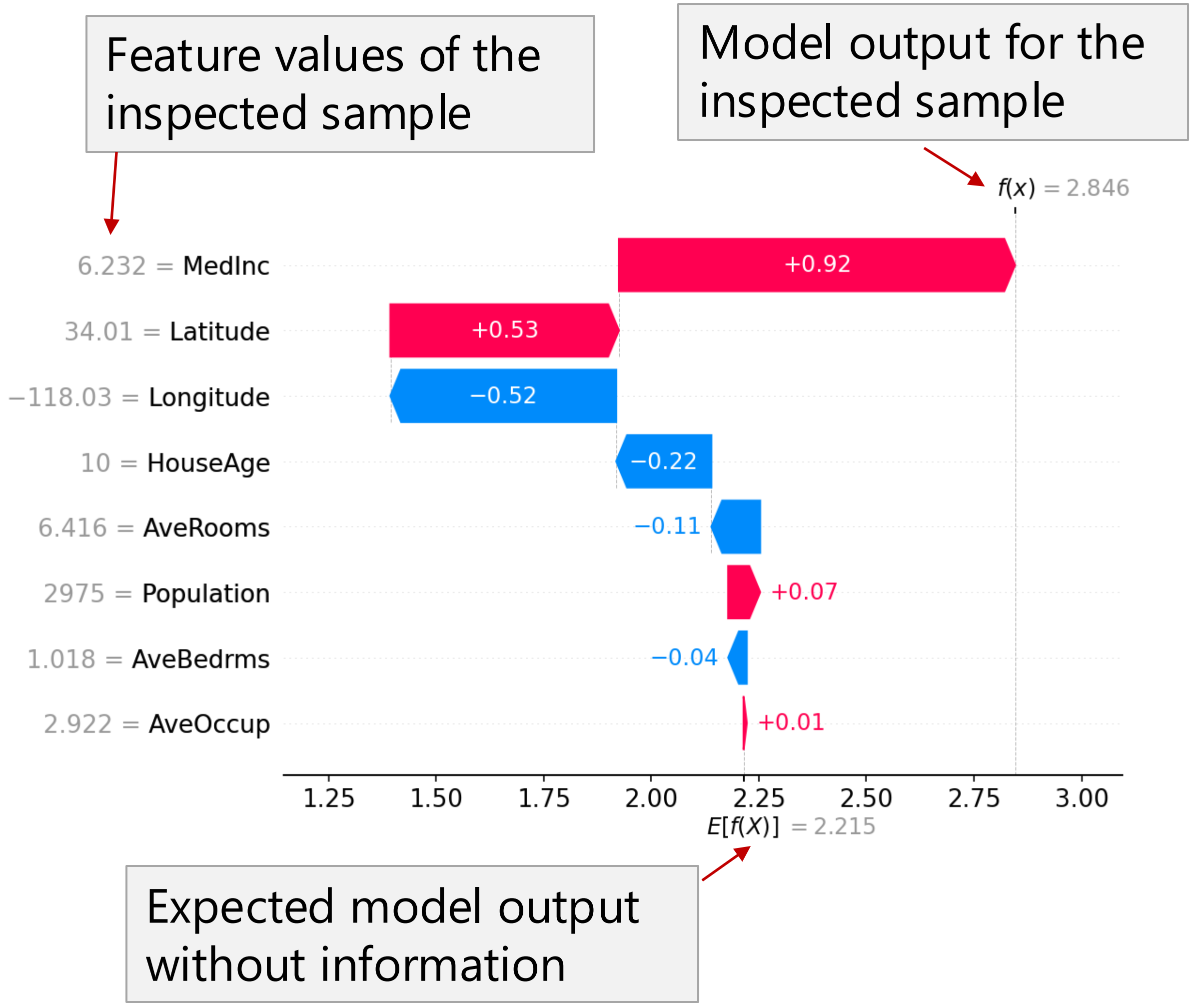
- Using Nobel-winning results from Game Theory, SHAP unifies previously proposed methods
  - Including LIME

- Treats input features as players forming coalitions to better explain the base model's prediction for ***x***

- **Shapley regression values** consists on retraining the model on all feature subsets and assessing how the model output changes with and without the feature being evaluated.

SHAP

- SHAP values of all the input features will always sum up to the difference between baseline model output and the model output for the prediction being explained
- For classification, baseline is the class proportion
- For regression, baseline is the average y

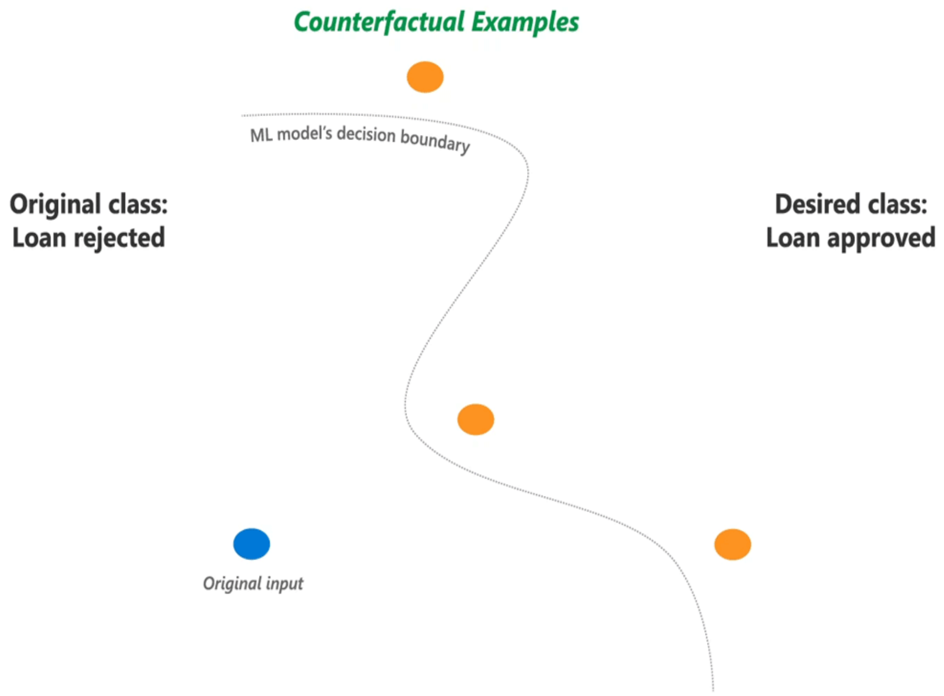


5. Explainability

a. Global vs local
b. Model-based
c. Surrogate model
d. Feature-based
e. Example-based

Counterfactual Explanations

- A counterfactual explanation for a specific prediction describes the closest possible input that would result in a *different* prediction from the model.
- **Why they are important:** They provide actionable insights, foster trust in AI systems, and help users understand the sensitive features influencing a decision.



LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

6. Large Language Models

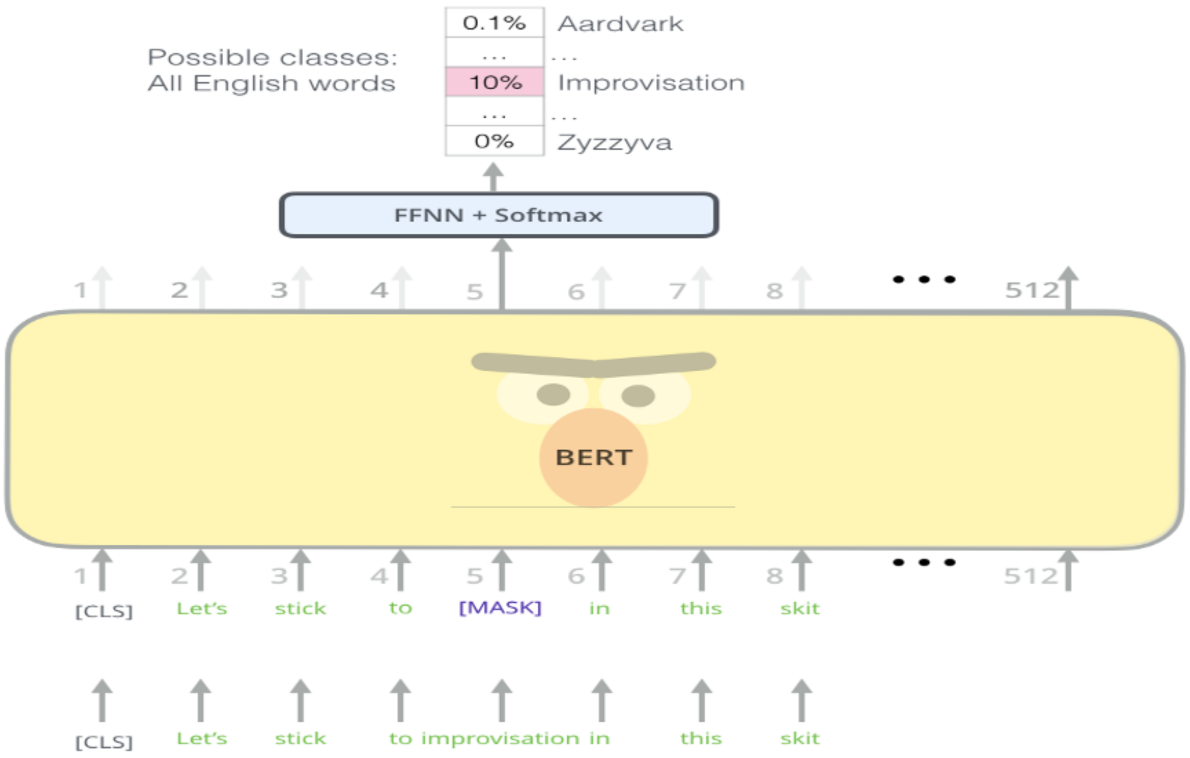
a. **Language Model**

b. Large Reasoning Models

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

What Is a Language Model?

- An AI system designed to understand and generate human language
- Predicts what comes next in a sentence.
- Predicts words within a sentence.



LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

What makes them 'Large'?

The "Large" in LLM refers to several key aspects:

Vast Training Data

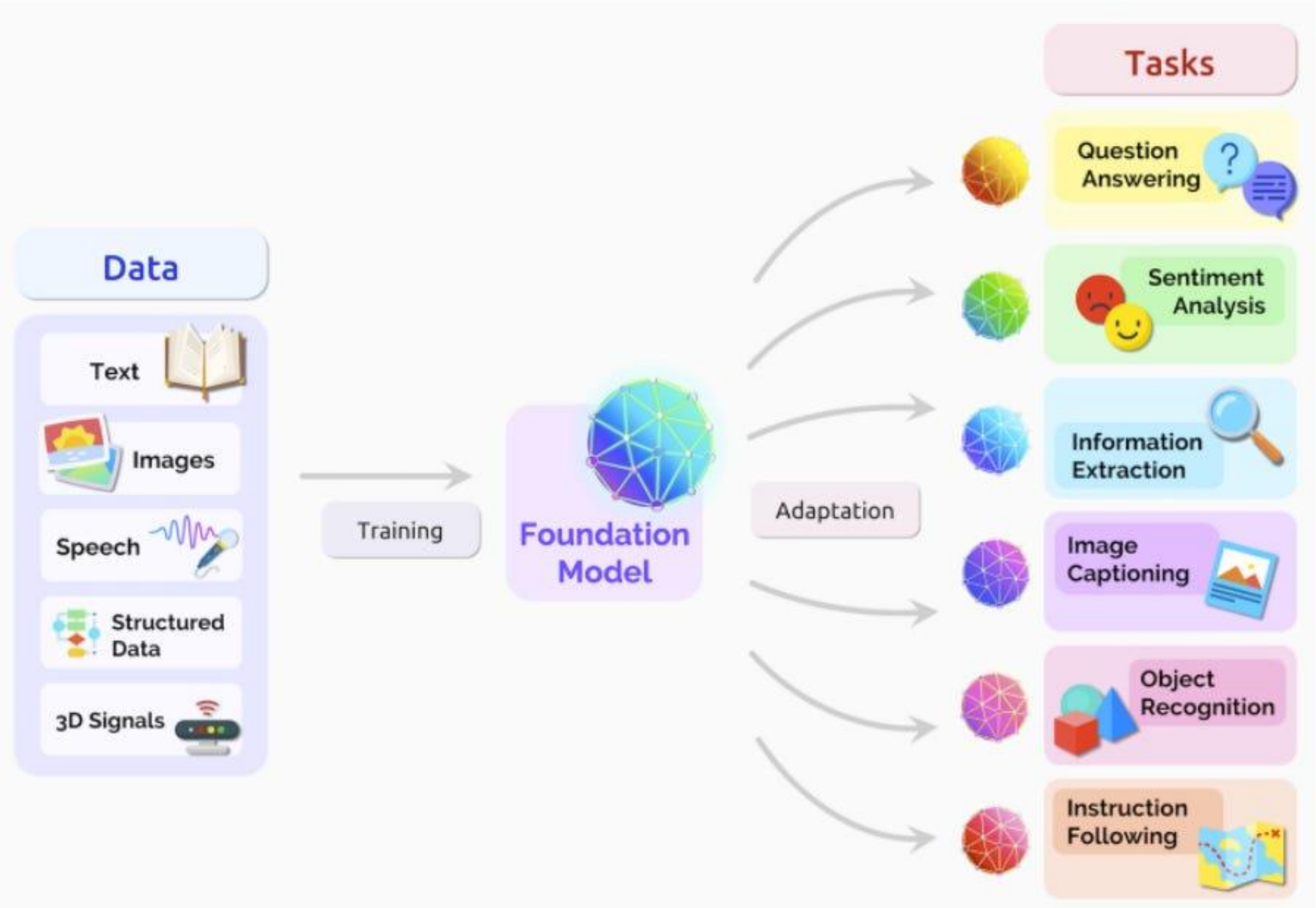
Billions of Parameters

Computational Power

What can they do ?

LLMs are versatile tools with many applications:

- **Text Generation**
- **Translation**
- **Code Generation & Debugging**
- **Chat and Dialogue**



Different language models

- OpenAI's GPT series
- Google's Gemini
- Anthtropic Claude
- Meta's LLaMA
- Mistral

LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

6. Large Language Models

a. Language Model

b. Large Reasoning Models

LEAP
A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Large Reasoning Models – Chain of Thought Reasoning

Encourages models to 'think through the problem'.

Example: "Let’s think step by step..."

Helps solve math and logic problems.

LEAP

A DIGITAL HEALTH HUB FOR
THE SOUTH WEST AND WALES

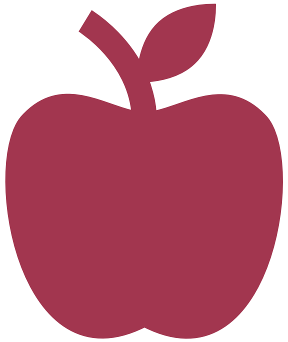
Chain of thought reasoning – No prompting

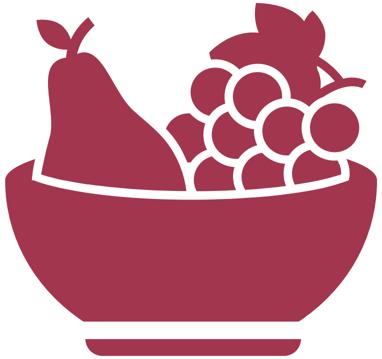
No Prompting

Q: If there are 5 baskets and each basket has 6 apples, how many apples are there in total?

A:30

Chain of thought reasoning – Zero Shot





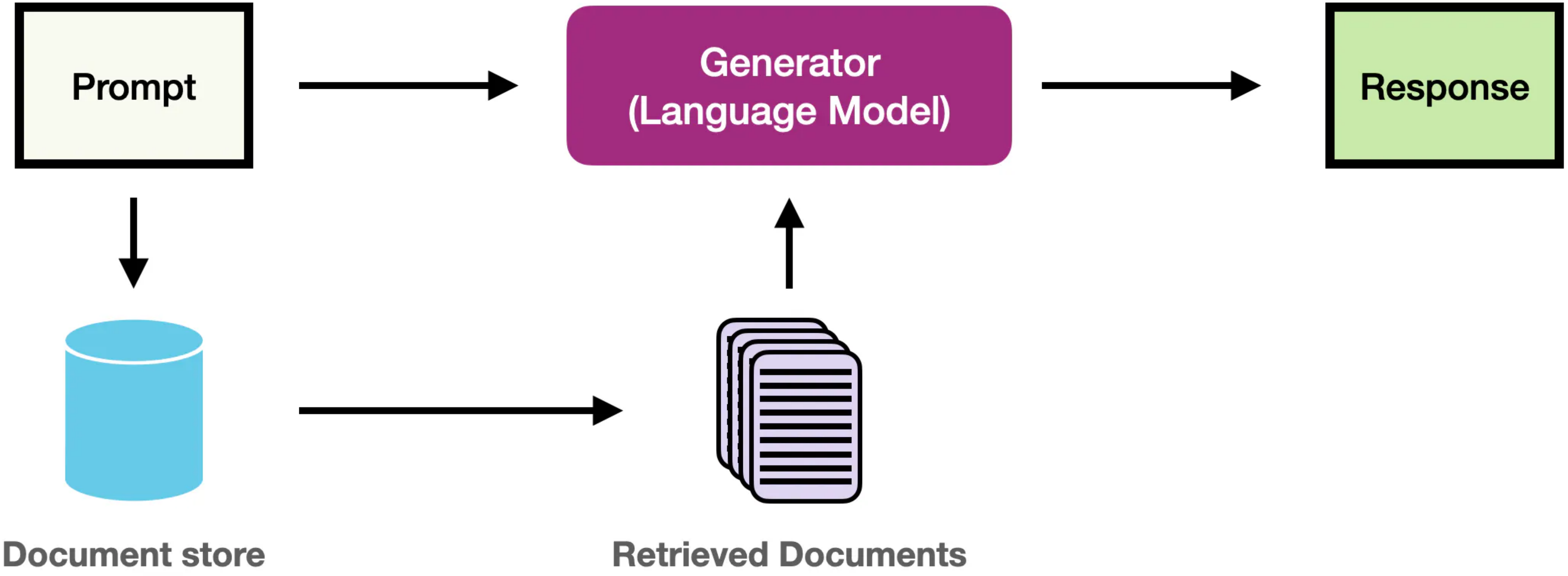
Q: If there are 5 baskets and each basket has 6 apples, how many apples are there in total? **Let's think step by step.**

A: Each basket has 6 apples. There are 5 baskets. So total apples = 5 × 6 = 30. The answer is 30.

Chain of thought reasoning – Few Shot

- **Q1: If a classroom has 6 rows with 7 chairs in each row, how many chairs are there?**
- **A1: Let's think step by step. Each row has 7 chairs. There are 6 rows. So total chairs = 6 × 7 = 42. The answer is 42.**
- Q2: If there are 5 baskets and each basket has 6 apples, how many apples are there in total?
- A2: Let's think step by step. Each basket has 6 apples. There are 5 baskets. So total apples = 5 × 6 = 30. The answer is 30.

Large Reasoning Models –Retrieval Augmented Generation (RAG)



Data privacy and security

- NHS Compliant, GDPR Compliant, HIPAA Compliant, ISO 27001 Accredited…
- Data security
- Not about the AI aspect

The limitations of AI

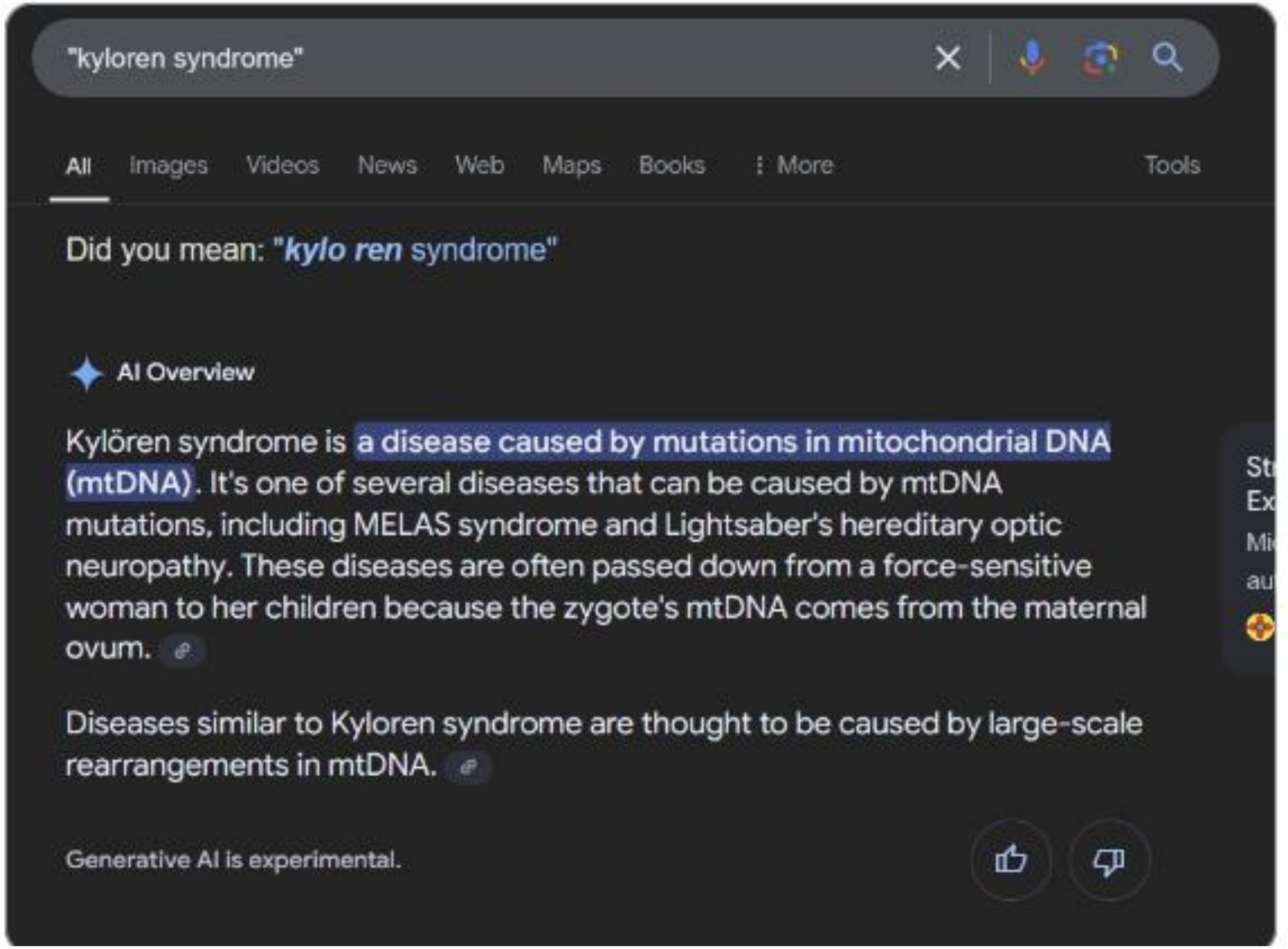
Neuroskeptic
@neuroskeptic.bsky.social
+ Follow

You know the " ◆ AI Overview" you get on Google Search?

I discovered today that it's repeating as fact something I made up 7 years ago as a joke.

"Kyloren syndrome" is a fictional disease I invented as part of a sting operation to prove that you can publish any nonsense in predatory journals...

- Some AI tools are designed to generate realistic text, but not true facts.
- The credibility of the sources may be difficult to assess.
- Google has fixed the issue.



LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Continuous improvement of AI

- I tried with Google but did not work.
- ChatGPT identifies that it is not a real syndrome.
- New online articles discussed the previous mistake.

Searched 3 sites

"Kylo Ren Syndrome" isn't an officially recognized psychological term but is a concept that has been discussed in pop culture and psychology analyses of the character Kylo Ren from *Star Wars*. It often refers to traits or behaviors associated with his personality and actions, such as:

1. **Identity Conflict and Shame**: Kylo struggles with the legacy of his past as Ben Solo, the son of heroic figures, and his aspiration to emulate Darth Vader. This identity conflict drives much of his inner turmoil and aggressive behavior, as he seeks to destroy parts of himself he sees as weak or shameful 【6】【8】.

LLMs hallucinations

Insights

- The AI generated plausible (but not real) references.
- The user didn't check the factuallity of the generated text.
- Used to fill a knowledge gap, instead of augmenting the experts habilities.

FINANCIAL REVIEW

Deloitte to refund government, admits using AI in $440k report

Edmund Tadros and Paul Karp

Oct 5, 2025 - 7.41pm

Deloitte Australia will issue a partial refund to the federal government after admitting that artificial intelligence had been used in the creation of a $440,000 report littered with errors including three nonexistent academic references and a made-up quote from a Federal Court judgement.

A new version of the report for the Department of Workplace Relations (DEWR) was quietly uploaded to the department's website on Friday, ahead of a long weekend across much of Australia. It features more than a dozen deletions of nonexistent references and footnotes, a rewritten reference list, and corrections to multiple typographic errors.

Q&A and Discussion

Interpreting Artificial Intelligence Solutions for Healthcare

Presenters:
Dr Miquel Perelló Nieto
Dr Nawid Keshtmand

LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Thank you!

Please take a moment to give us some feedback to help plan future events.

University of Bristol | UWE Bristol University of the West of England | University of Bath | Cardiff University Prifysgol Caerdydd | University of Exeter | HDRUK Health Data Research UK | UKRI Engineering and Physical Sciences Research Council leap-hub.ac.uk

leap-dh-hub@bristol.ac.uk

@LEAP_hub

leap-dh-hub

Sign up to the Hub's mailing list

LEAP

A DIGITAL HEALTH HUB FOR THE SOUTH WEST AND WALES

Connect with the Hub

University of Bristol | UWE Bristol University of the West of England | University of Bath | Cardiff University Prifysgol Caerdydd | University of Exeter | HDRUK Health Data Research UK | UKRI Engineering and Physical Sciences Research Council