Computer says no! Explaining the decisions of machines

Sandy Brownlee

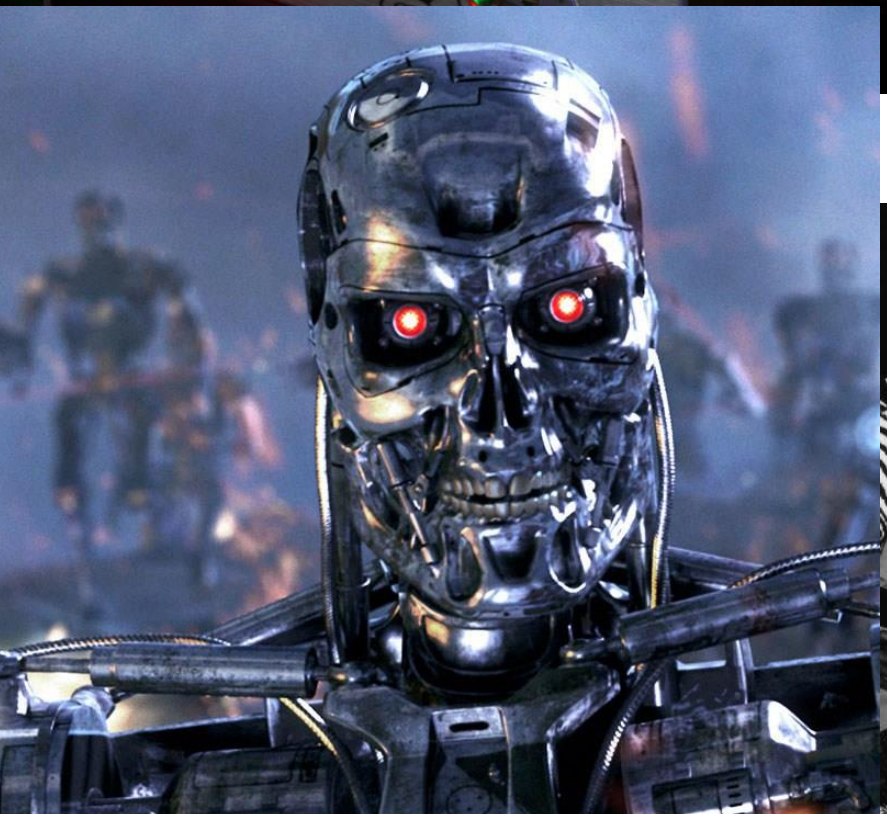
Content

- AI, what it’s good at and where it fails
- The basics – how do AI systems work?
- How do we explain the decisions of these systems?
- Is any of this enough?

Artificial Intelligence (AI) is on the march!


WOPR
War Operation Plan Response
CAUTION


I'm sorry Dave,
I'm afraid I can't do that.



GOOGLE'S AI WINS FIFTH AND FINAL GAME AGAINST GO GENIUS LEE SEDOL

AI can predict whether your relationship will last based on how you speak to your partner

September 29, 2017 10.25am BST

Lip-reading CCTV could soon capture shoppers' comments for big companies

Peter Swindon @PeterSwindon
Senior reporter, Sunday Herald

Google's 'worst' self-driving accident was still a human's fault

'Our light was green'

ndrew J. Hawkins | @andyjayhawk | Sep 26, 2016, 4:03pm EDT

SHARE TWEET LINKEDIN

Technology

Killer robots: Experts warn of 'third revolution in warfare'

5 hours ago | Technology 61

Would you trust a machine?

They often get decisions right…

Artificial intelligence can accurately predict future heart disease and strokes, study finds

Home

Press releases

2019

2018

2017

January

February

March

April

May

June

July

August

24 Apr 2017 08:45:00.000
PA77/17

Computers that can teach themselves from routine clinical data are potentially better at predicting cardiovascular risk than current standard medical risk models, according to new research at the University of Nottingham.

The team of primary care researchers and computer scientists compared a set of standard guidelines from the American College of Cardiology (ACC) with four 'machine-learning' algorithms – these analyse large amounts of data and self-learn patterns within the data to make predictions on future events – in this case, a patient's future risk having of heart disease or a stroke.

The results, published in the online journal PLOS ONE, showed that the self-teaching 'artificially intelligent' tools were significantly more accurate in predicting cardiovascular disease than the established algorithm. In computer science, the AI algorithms that were used are called 'random forest', 'logistic regression', 'gradient boosting' and 'neural networks'.

All News

Additional resources

No additional resources for this article

Related articles

Artificial intelligence can predict premature death, study finds

Wednesday 27th March 2019

New MyAsthma app can help relieve the stress of asthma management

Wednesday 12th April 2017

Almost half of stroke survivors suffer fatigue, study reveals

Friday 3rd March 2017

How technology is allowing police to predict where and when crime will happen

Report finds that British police have wealth of data but 'lack capability to use it'

Lizzie Dearden Home Affairs Correspondent | @lizziedearden |
Saturday 7 October 2017 22:40 | 21 comments


5:00
English
Spanish
English
Where is the train station?
Spanish
¿Dónde está la estación de tren?


amazon


amazon

Then again…

Who has ever clicked on the “you might like this” button? The systems underlying this, and other online advertising, are only so good…

This is what happens when you remove the humans and let the social media trackers reign supreme. Yes it has 'Carlsberg' and 'awesome' in the same sentence, but I think they may have missed the context on this one...

(Or it may be entirely deliberate?)

585 likes • 63 comments

Like Comment Share

amazon

Your Amazon.com | Today's Deals | See All Departments

Frequently Bought Together

You Bought

Waring CTT200BK Professional Cool...

by Waring

+

Customers Also Bought

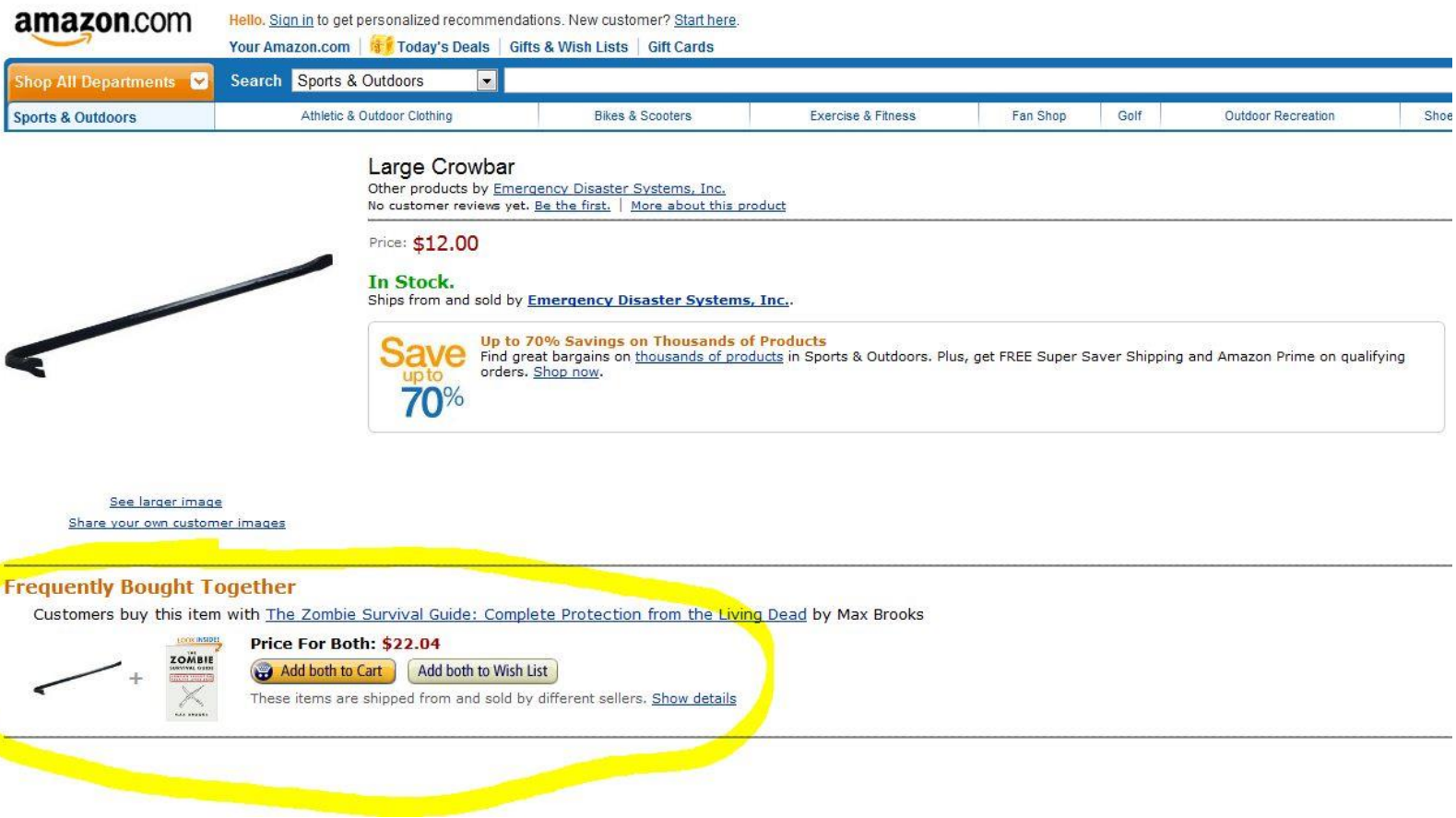
The Dark Knight Rises

Blu-ray ~ Christian Bale

(1,538 reviews)

$24.96

Learn more


amazon.com
Hello. Sign in to get personalized recommendations. New customer? Start here.
Your Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards
Shop All Departments
Search
Sports & Outdoors
Sports & Outdoors
Athletic & Outdoor Clothing
Bikes & Scooters
Exercise & Fitness
Fan Shop
Golf
Outdoor Recreation
Large Crowbar
Other products by Emergency Disaster Systems, Inc.
No customer reviews yet. Be the first. | More about this product
Price: $12.00
In Stock.
Ships from and sold by Emergency Disaster Systems, Inc..
Save up to 70%
Up to 70% Savings on Thousands of Products
Find great bargains on thousands of products in Sports & Outdoors. Plus, get FREE Super Saver Shipping and Amazon Prime on qualifying orders. Shop now.
See larger image
Share your own customer images
Frequently Bought Together
Customers buy this item with The Zombie Survival Guide: Complete Protection from the Living Dead by Max Brooks
Price For Both: $22.04
Add both to Cart
Add both to Wish List
These items are shipped from and sold by different sellers. Show details


bblica TV
Cruise Vacations--75% Off
Daily specials, last-minute cruises All cruises discounted up to 75%.
VacationsToGo.com
Ads by Google
0:13 / 0:29
360p

Photos Opinion Local News Odd News Comics Weather Full Coverage Video/Audio Kevin Sites

Search: All News Search Advanced

Over 250 sick after eating at Indiana Olive Garden

REUTERS

Reuters Photo: A plate of pasta from the Olive Garden is seen in an undated file photo....

50 minutes ago

LOS ANGELES (Reuters) - More than 250 people have reported becoming sick after eating at an Olive Garden restaurant in Indianapolis, Indiana, a county health official said on Friday, a day after an outbreak of E.coli at Taco Bell restaurants was declared over.

The news makes Olive Garden at least the third U.S. restaurant chain this month to be linked to widespread customer illnesses.

Some customers who ate at the Olive Garden restaurant in northeast Indianapolis between December 9 and December 13 have reported nausea, vomiting, diarrhea, and in some cases fever, said John Althardt, a spokesman for the Marion County Health Department.

Three of those people have been hospitalized.

ELSEWHERE ON THE WEB

CNN.COM

Tests of the sick peoples' stool and leftovers they took home from the restaurant will be conducted later today or Monday, Althardt said. He added that the tests would take about 48

Jolly fun, but this also has more serious implications

Police

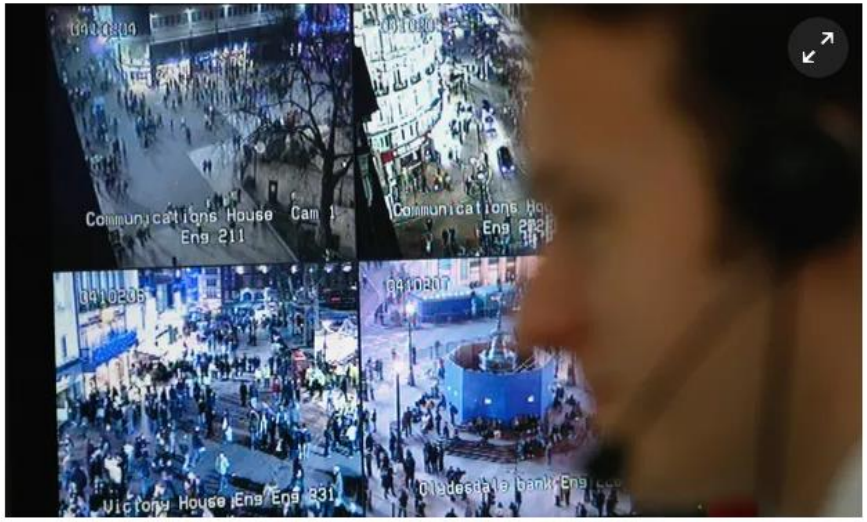
This article is more than **2 months old**

UK police use of computer programs to predict crime sparks discrimination warning

Human rights group claims the algorithms threaten a 'tech veneer to biased practices'

Sarah Marsh
@sloumarsh

Sun 3 Feb 2019
16.00 GMT



▲ A police officer watches a television monitor looking at London's CCTV camera network. Photograph: Daniel Berehulak/Getty Images

The rapid growth in the use of computer programs to predict crime hotspots and people who are likely to reoffend risks locking discrimination into the criminal justice system, a report has warned.

Amazon scrapped 'sexist AI' tool

10 October 2018 Share

The algorithm repeated bias towards men, reflected in the technology industry

An algorithm that was being tested as a recruitment tool by online giant Amazon was sexist and had to be scrapped, according to a Reuters report.

The artificial intelligence system was trained on data submitted by applicants over a 10-year period, much of which came from men, it claimed.

Reuters was told by members of the team working on it that the system effectively didates were preferable.

https://www.bbc.co.uk/news/world-africa-47913338

Top Stories

Sudan coup leader steps down
Army chief Awad Ibn Auf quits a day after staging a coup that toppled veteran leader Omar al-Bashir.
6 minutes ago

Nigel Farage launches Brexit Party
7 hours ago

Cross-party Brexit talks 'positive'
6 hours ago

Features

Squalor, defeat and the 'jihadist mind trick'

www.telegraph.co.uk/technology/2018/07/13/ai-doctor-app-babylon-fails-diagnose-heart-attack-complaint/

Home | My Feed | Saved

Search | Video | Rewards | Subscribe - 30 days free | Login | Register

The Telegraph

HOME | NEWS | SPORT | BUSINESS | ALL SECTIONS

Technology Intelligence

Gadgets | Innovation | Big Tech | Start-ups | Politics of Tech | Gaming | Podcast | Tech Jobs | Newsletter

Technology Intelligence

AI doctor app Babylon fails to diagnose heart attack, complaint alleges

share

Save 17

Babylon Health is facing criticism from an anonymous doctor. CREDIT: PA

So what happened? Uber discovered that its self-driving software decided not to take any actions after the car's sensors detected the pedestrian. Uber's autonomous mode disables Volvo's factory-installed automatic emergency braking system, according to US National Transportation Safety Board preliminary report on the accident.

Passport photo

Select photo

✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.
You have 9 attempts left.

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

Filename: Untitled.jpg

If you wish to contact us about the photo, you must provide us with the reference number given above.

Please print this information for your records.

Print

Image credit: Richard Lee | Facebook

EFF Asks Court: Can Prosecutors Hide Behind Trade Secret Privilege to Convict You?

PRESS RELEASE | SEPTEMBER 14, 2017

California Appeals Court Urged to Allow Defense Review of DNA Matching Software

If a computer DNA matching program gives test results that implicate you in a crime, how do you know that the match is correct and not the result of a software bug? The Electronic Frontier Foundation (EFF) has urged a California appeals court to allow criminal defendants to review and evaluate the source code of forensic software programs used by the prosecution, in order to ensure that none of the wrong people end up behind bars, or worse, on death row.

In this case, a defendant was linked to a series of rapes by a DNA matching software program called TrueAllele. The defendant wants to examine how TrueAllele takes in a DNA sample and analyzes potential matches, as part of his challenge to the prosecution's evidence. However, prosecutors and the manufacturers of TrueAllele's software argue that the source code is a trade secret, and therefore should not be disclosed to anyone.

https://www.eff.org/press/releases/eff-asks-court-can-prosecutors-hide-behind-trade-secret-privilege-convict-you

So how do we know when we've got a "good" AI system?

Experts?

Created by Chanut is Industries
from Noun Project

That kind of defeats the point.

- If the predictions **agree** with the experts, **what's the point?**
- If the predictions **disagree** with the experts, is the machine **incompetent**? Or is it spotting something humans **can't see**?

Wouldn’t it be great if the machine could answer…

why?

how?

Back to basics…

So, how do these things work?

Features

Numbers

Numbers

21 39 10 52 13 81 …

21 39 10 52 13 81 …

21 39 10 52 13 81 …

?

67%

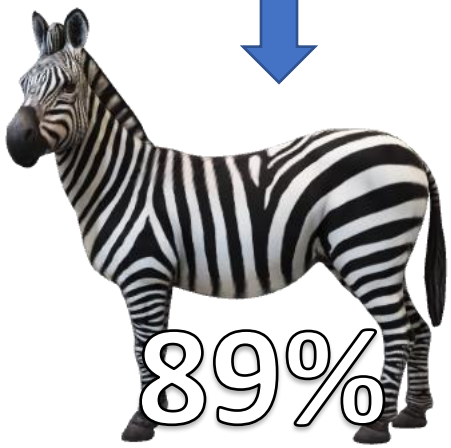
12%

23%

34 21 44 21 68 81 …

?

5%



89%

32%

34 21 44 21 68 81 …

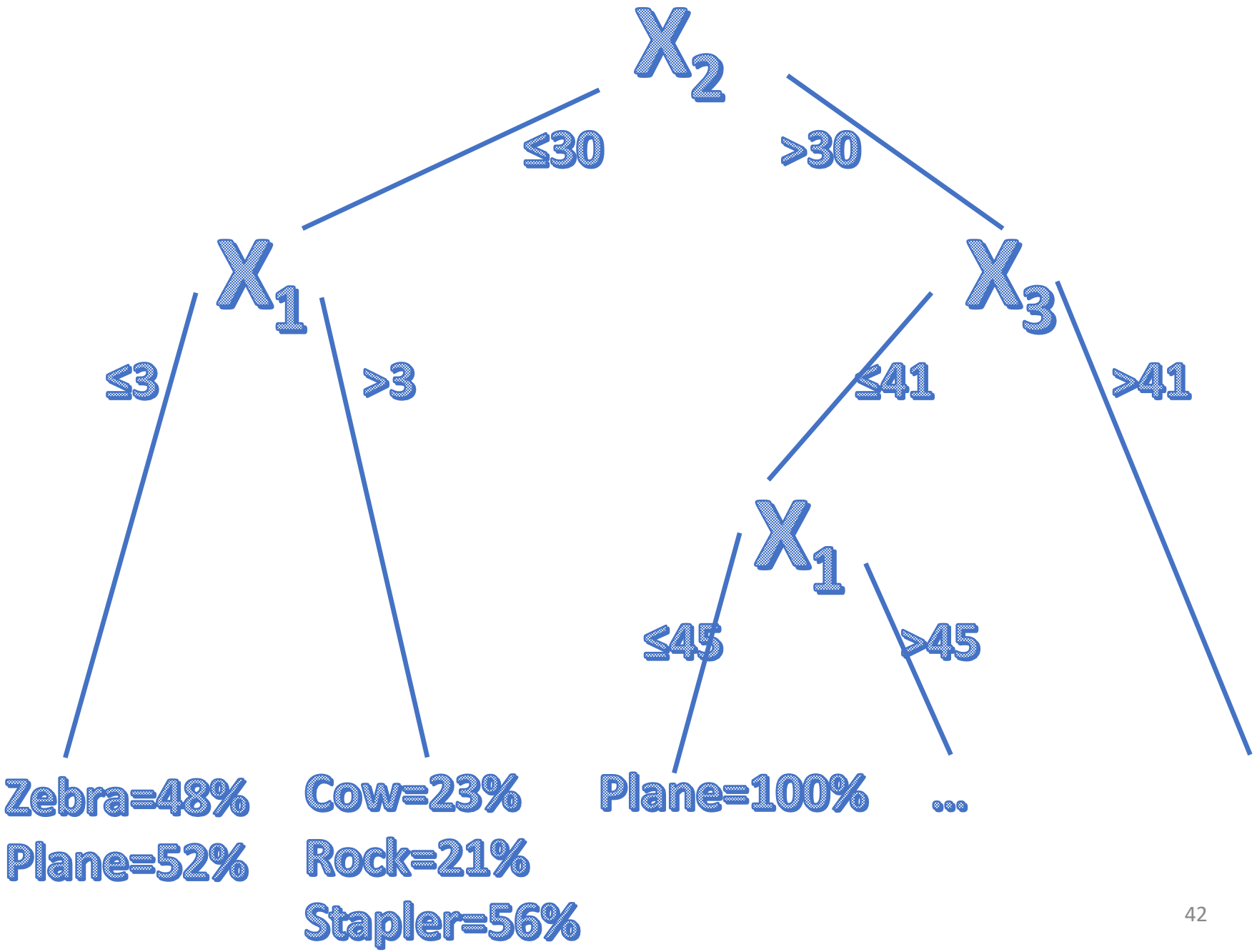
?

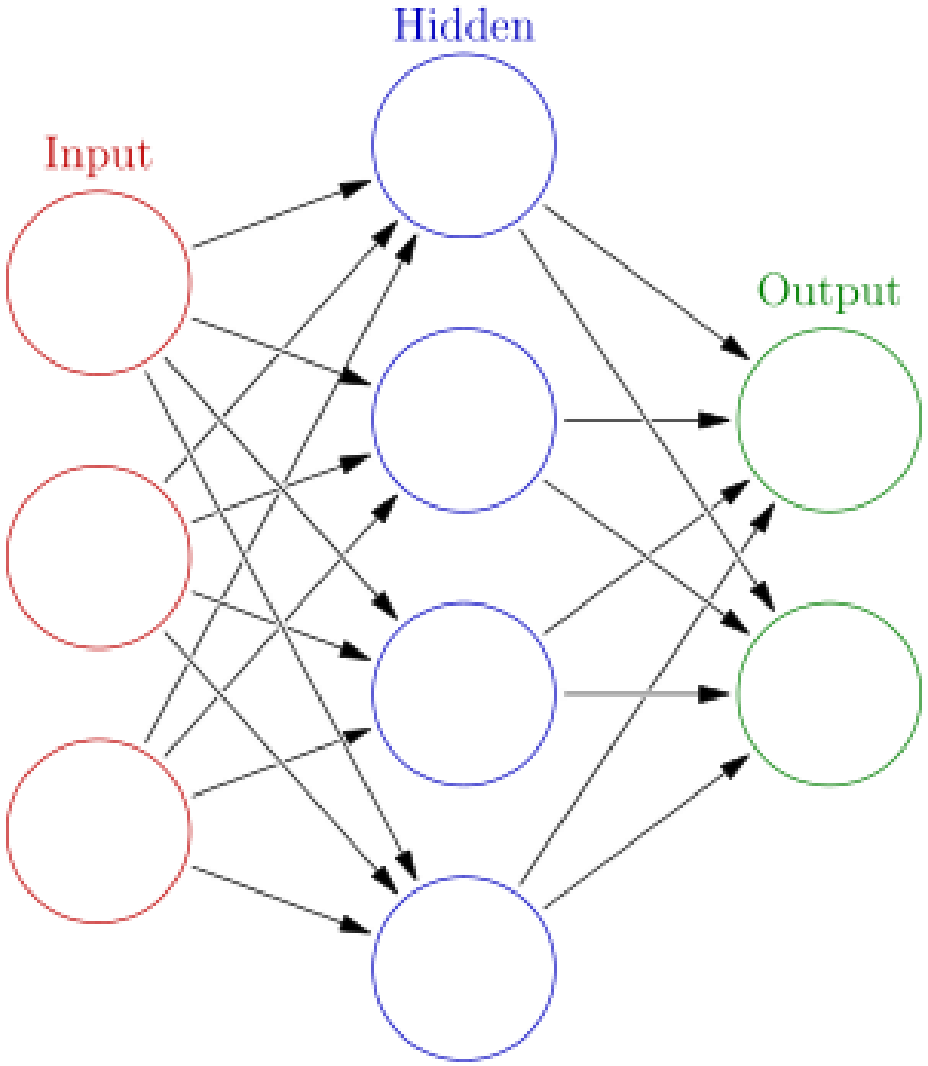
34%

7%

76%

$$\ell = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots$$


X2
≤30
>30
X1
X3
≤3
>3
≤41
>41
X1
≤45
>45
Zebra=48%
Plane=52%
Cow=23%
Rock=21%
Stapler=56%
Plane=100%
…



Okay, but…

where do these structures come from?

Minimising *cost:*

given some training data (examples),
find the parameters that get it right
most often (or some variation of this)

BUT:

The machine will only minimise cost!

However it is defined.

Machines are very good at finding shortcuts!

A simulated robot was supposed to evolve to travel as quickly as possible… what happened?

Why walk when you can flop? In one example, a simulated robot was supposed to evolve to travel as quickly as possible. But rather than evolve legs, it simply assembled itself into a tall tower, then fell over. Some of these robots even learned to turn their falling motion into a somersault, adding extra distance.

Lehman, Joel, et al. "The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities." *arXiv preprint arXiv:1803.03453* (2018).

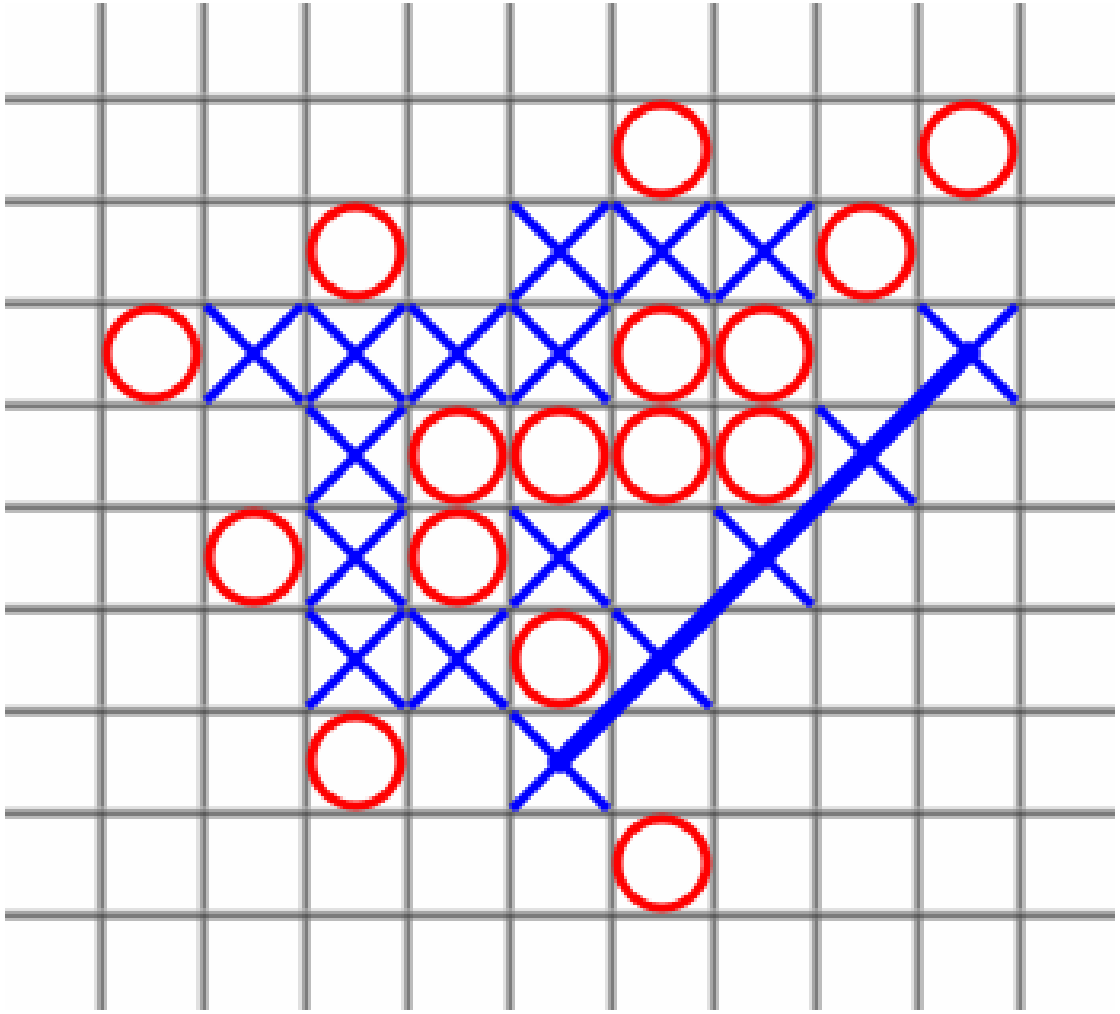


Image: DaBler [Public domain]

How to win at tic-tac-toe: … "a five-in-a-row Tic Tac Toe competition played on an infinitely large board … It turned out that the algorithm's strategy was to place its move very, very far away, so that when its opponent's computer tried to simulate the new greatly-expanded board, the huge gameboard would cause it to run out of memory and crash, forfeiting the game."

Lehman, Joel, et al. "The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities." *arXiv preprint arXiv:1803.03453* (2018).

Automatic software repair: the AI system was challenged to fix a program that sorted a sequence of numbers into ascending order.

The “cost function” checked whether the numbers were correctly sorted, giving a higher score to programs producing better sorted sequences.

```
2 5 8 14 18      cost = 0
18 14 8 5 2      cost = 4
8 2 14 5 18      cost = 2
```

What shortcut did the AI take?

It simply deleted all the numbers.

An empty list is, by its very nature, sorted.

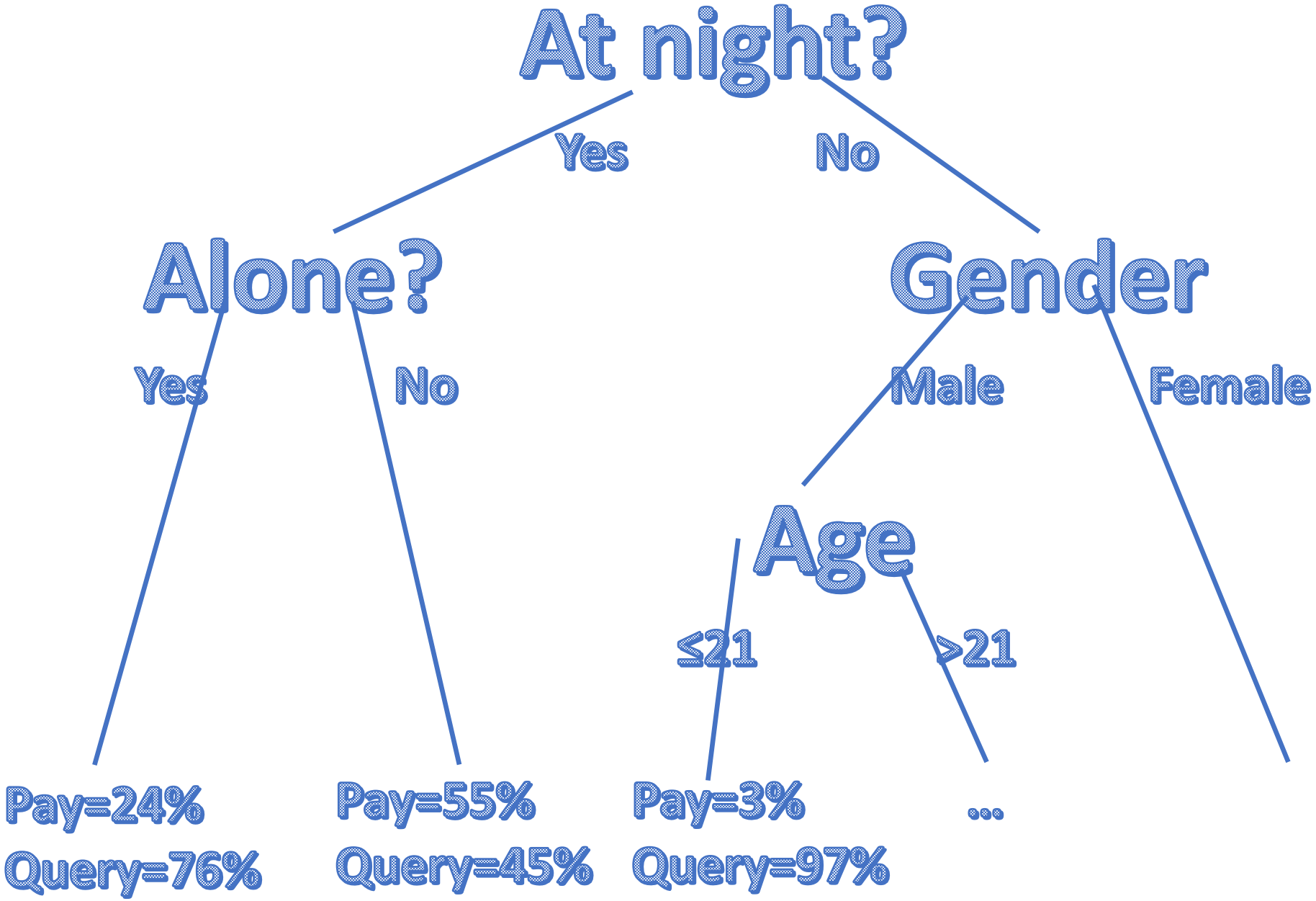
So:

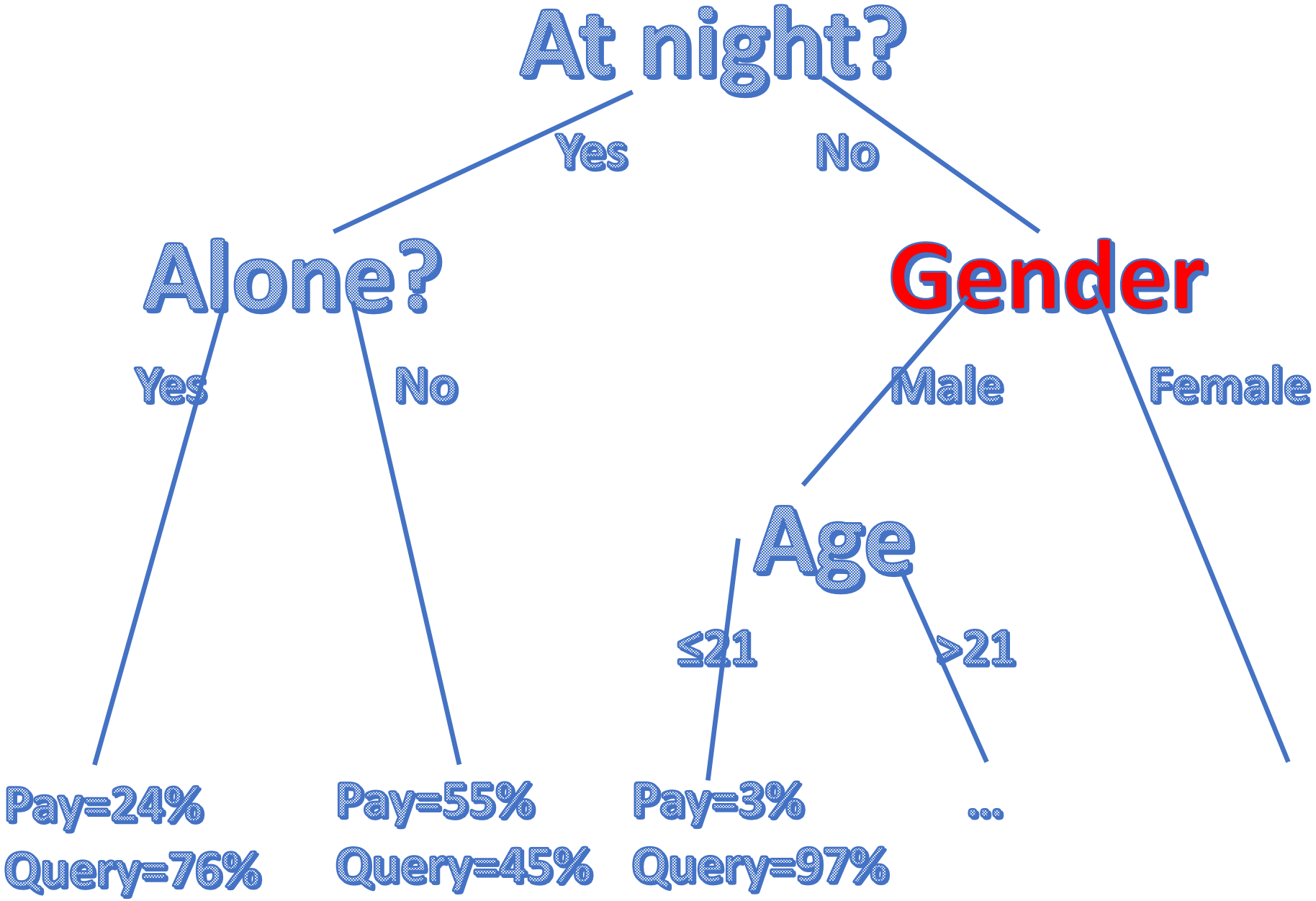
The learning process just looks for patterns… sometimes the patterns that are found are undesirable!

How can we explain the model?

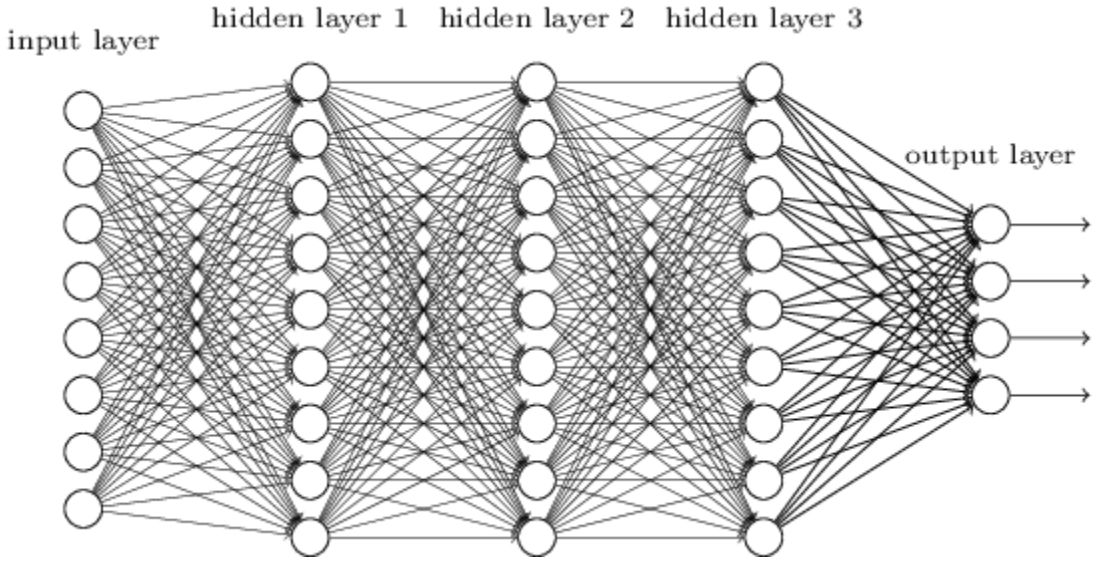
For the models we saw earlier, it's fairly simple.

These are *interpretable*.

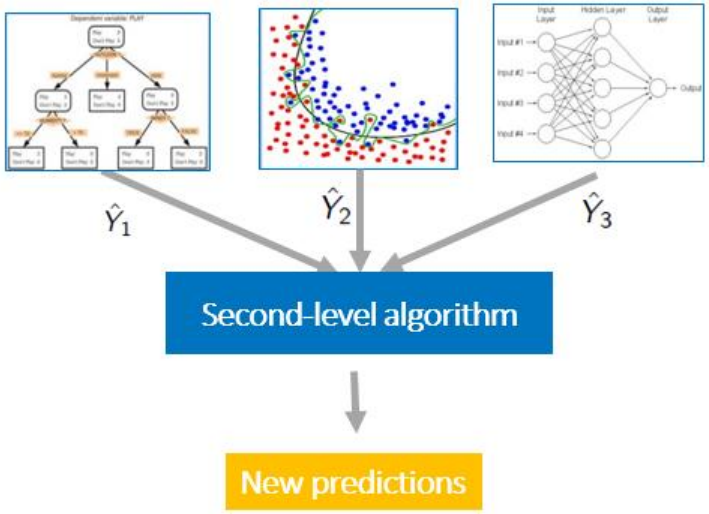

At night?
Yes
No
Alone?
Gender
Yes
No
Male
Female
Age
≤21
>21
Pay=24%
Query=76%
Pay=55%
Query=45%
Pay=3%
Query=97%
…


At night?
Yes
No
Alone?
Gender
Yes
No
Male
Female
Age
≤21
>21
Pay=24%
Query=76%
Pay=55%
Query=45%
Pay=3%
Query=97%
…

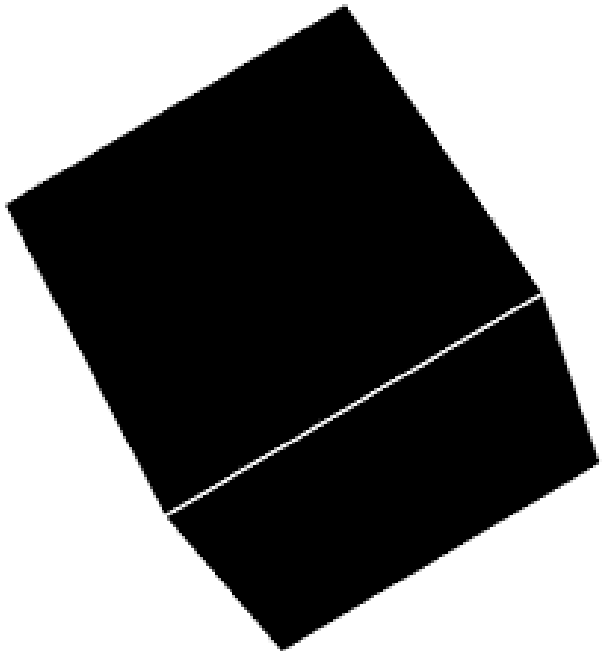
So what’s the problem?



Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015



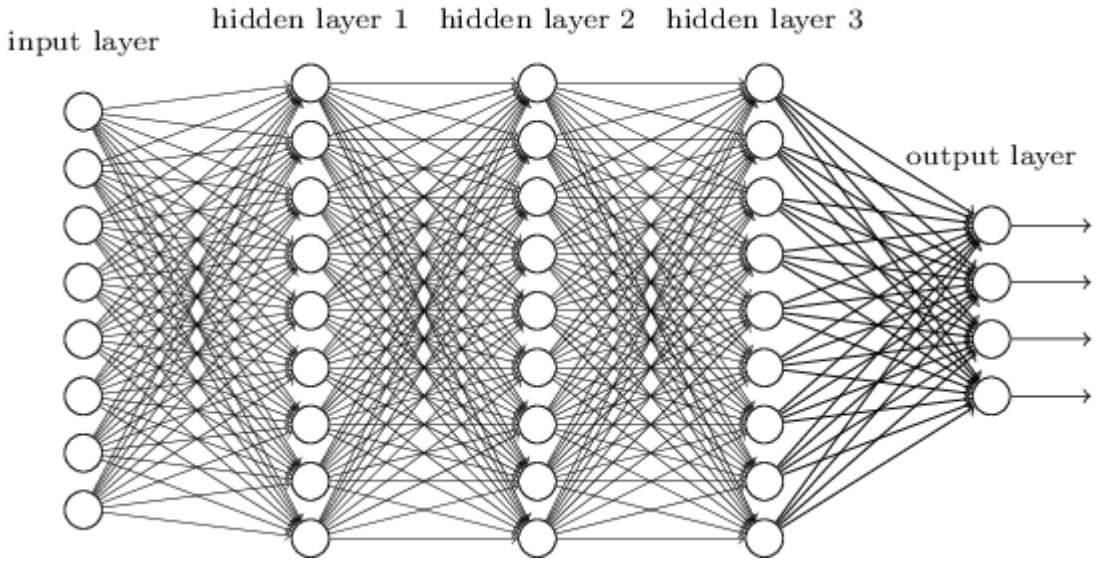
https://blogs.sas.com/content/subconsciousmusings/2017/05/18/stacked-ensemble-models-win-data-science-competitions/

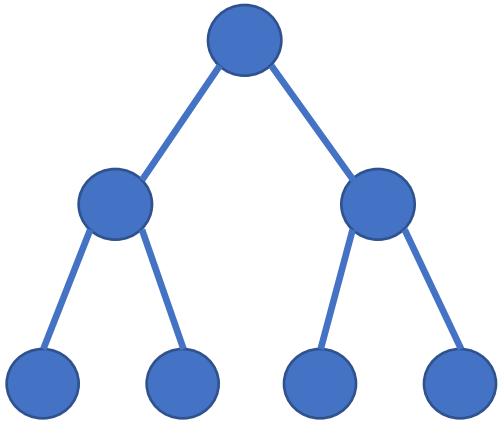


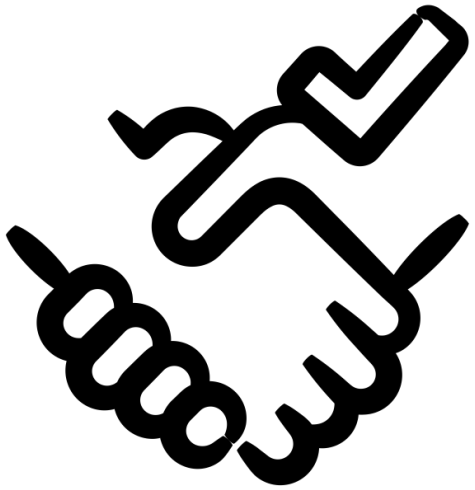
Interpretability: you **do understand** it but it **doesn't work** well.

Performance: you **don't understand** it but it **does work** well.

Combine interpretable models with non-interpretable ones to get explanations


input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer





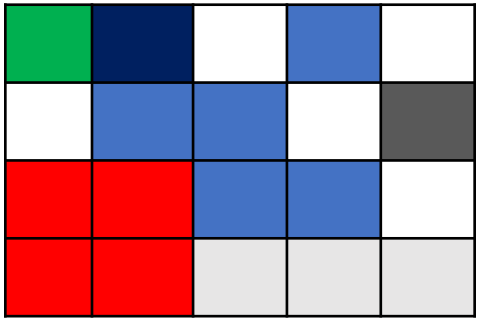
Created by Gregor Cresnar
from Noun Project

= explanation!

Created by CARLOS ALVAREZ LOPEZ
from Noun Project

= ???

Another way…

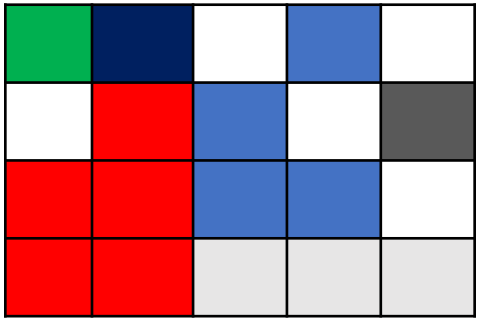


?

34%

7%

76%

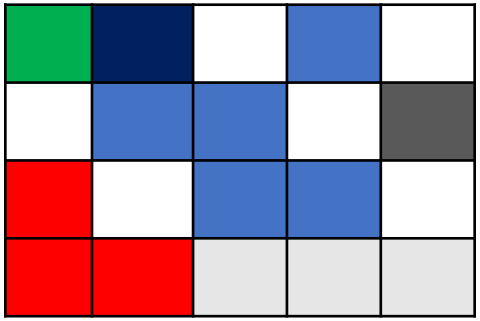


?

41%

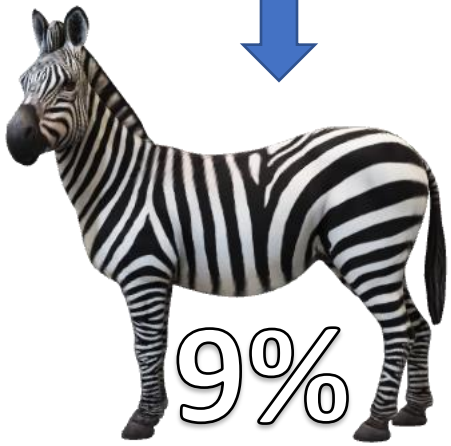
7%

71%



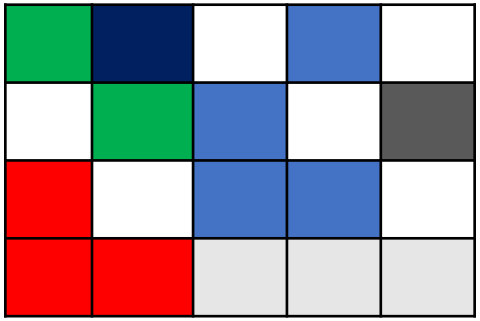
?

29%



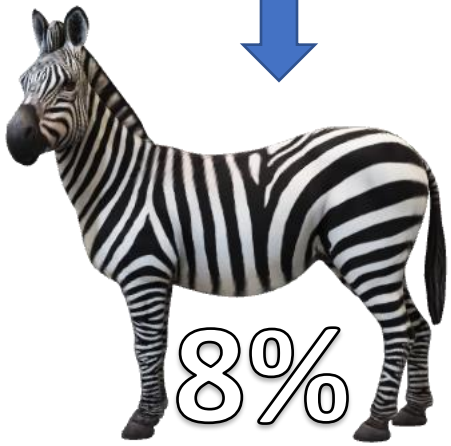
9%

78%



?

33%



8%

73%

We repeat this many times; it becomes a *search* problem where we try to find the smallest changes to the inputs that change the final classification (using the percentages to give some feedback, or rather direct the search)

Machines are also very good at this kind of task, and can try many millions of possibilities

(a) Original Image

(b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*

Is this really explaining how the algorithm made its decision?

Is it just a *plausible* explanation?

What are we really looking for?

- Level of detail – ideally a continuum specified by the user as different people want different levels of depth
- Consistency of results
- Competency – show confidence as percentage
- How do you show that reliable sources have been used and that the data is credible?

Issues Affecting User Confidence in Explanation Systems. David A. Robb, Stefano Padilla, Thomas S. Methven, Yibo Liang, Pierre Le Bras, Tanya Howden, Azimeh Gharavi, Mike J. Chantler, Ioannis Chalkiadakis (all at Heriot-Watt). RealX 2018 Proceedings

Summary

- We are really just scratching the surface of Explainable AI
- Striking a balance between good systems and explainable systems
- It might not be fully achievable!
  - Some patterns the machine finds could genuinely be too hard to explain, at least intuitively
  - Can be difficult for humans to understand their own decisions so what chance have machines got?

Thanks!

- www.cs.stir.ac.uk/~sbr --- sbr@cs.stir.ac.uk
- Article on mining models to explain optimisation

  Brownlee, A. E. I. Mining Markov Network Surrogates for Value Added Optimisation. GECCO Conference 2016. DOI:10.1145/2908961.2931711
- Some other relevant reading
  - https://medium.freecodecamp.org/an-introduction-to-explainable-ai-and-why-we-need-it-a326417dd000
  - https://theconversation.com/people-dont-trust-ai-heres-how-we-can-change-that-87129
  - https://hackernoon.com/explainable-ai-wont-deliver-here-s-why-6738f54216be
  - Lehman, Joel, et al. "The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities." *arXiv preprint arXiv:1803.03453* (2018)
  - Automated bug fixing: https://theconversation.com/computers-will-soon-be-able-to-fix-themselves-are-it-departments-for-the-chop-85632 (link from my home page above)

- **Next Lecture on 9th May: *What comes next?* Donald Smith**