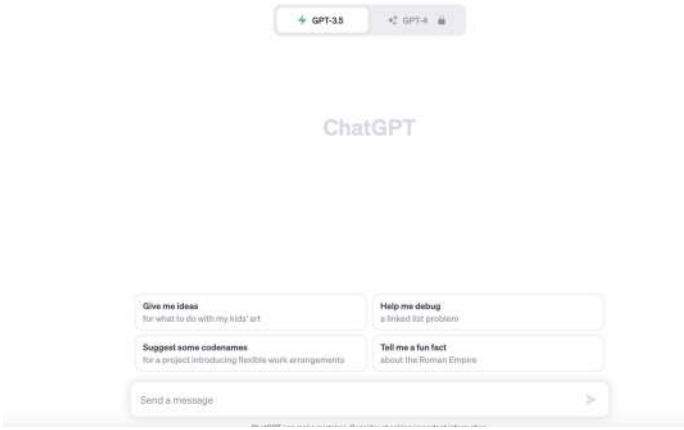
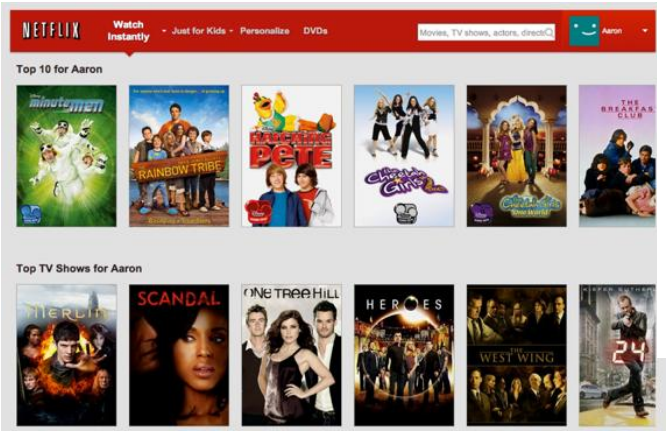
University of Glasgow
Oilthigh Ghlaschu
How intelligent is Artificial Intelligence, and is it really going to make us irrelevant?
Dr Simone Stumpf
Reader in Responsible and Interactive AI
Simone.Stumpf@glasgow.ac.uk
WORLD CHANGING GLASGOW
A WORLD TOP 100 UNIVERSITY

University of Glasgow
Oilthigh Ghlaschu

AI is starting to be everywhere…





University of Glasgow
Oilthigh Ghlaschu

What is AI anyway?

- AI has been around for a long time, since the 1950s
- It was mainly about using information to reason and solve problems
- Alan Turing was a founder of this discipline and we still use the Turing Test to think about what “intelligence” is

University of Glasgow
Oilthigh Ghlaschu

It used to be very different, back in the day (until 1990s)

- Mainly ”symbolic” with hand-crafted rules, ”expert systems”
- Simple neural networks
- Small amounts of data, low computing power
- It was often not correct
- Limited applications

University of Glasgow
Oilthigh Ghlaschu

But then…

- Greater computing power
- Lots more data
- Rise of machine learning and deep learning
- More applications

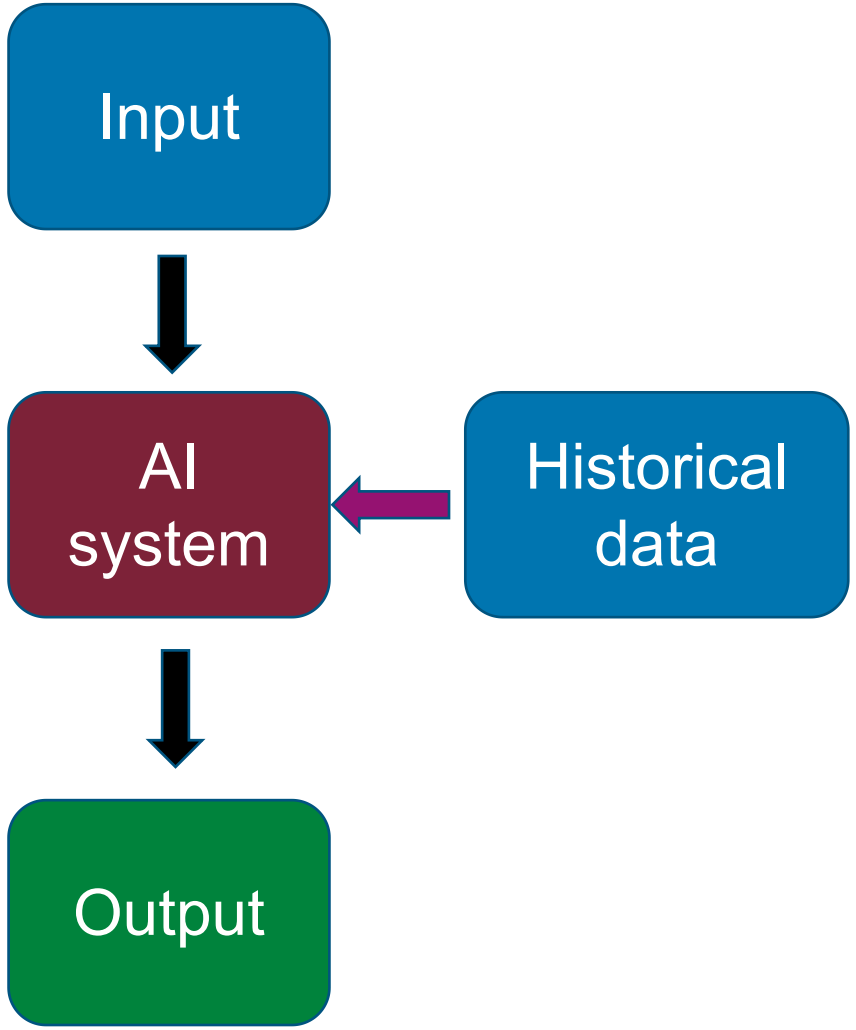
Chinese Go player Ke Jie has lost two games to AlphaGo

University of Glasgow
Oilthigh Ghlaschu

So how does it work?

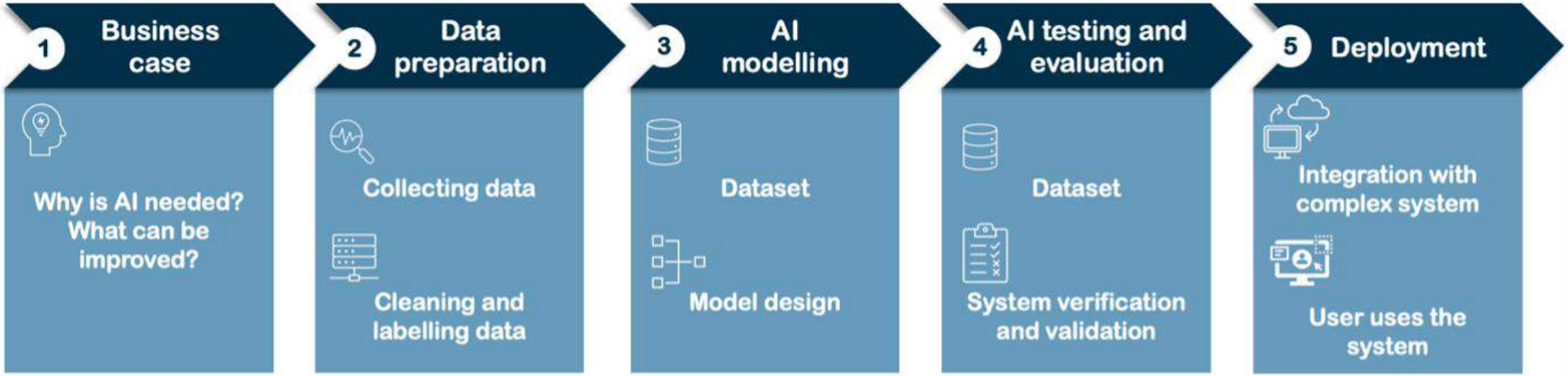
- Basically, you feed historical data into an AI system to ‘train’ a ‘model’ i.e. the machine learns a ‘pattern’ from this data

- Then, you give it new data as input and the model produces some output



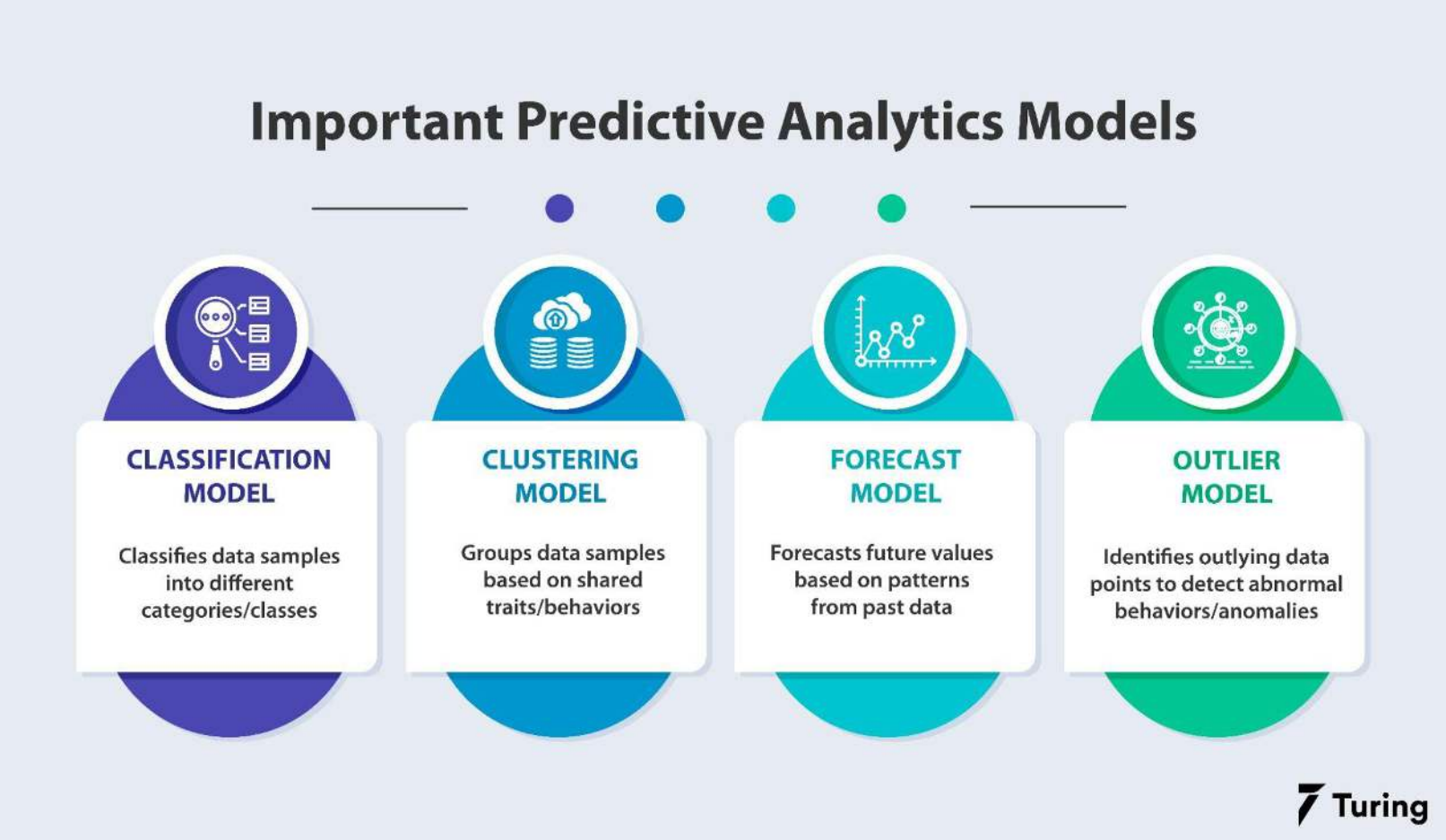
University of Glasgow
Oilthigh Ghlaschu

How are they developed?



University of Glasgow
Oilthigh Ghlaschu

Different kinds of predictive AI



University of Glasgow
Oilthigh Ghlaschu

An example: SPARRA

https://publichealthscotland.scot/our-areas-of-work/social-and-community-care/scottish-patients-at-risk-of-readmission-and-admission-sparra/overview/

- Scottish Patients at Risk of Readmission and Admission (SPARRA) is a risk prediction tool which predicts an individual's likelihood of being admitted to hospital as an emergency inpatient within the next year.
- SPARRA data can help health care professionals to prioritise patients with complex care needs who are likely to benefit most from anticipatory health care. SPARRA data can also be used in a service planning capacity by locating groups of patients who would benefit from specific interventions or services.
- Scores are calculated monthly for approximately 4.2 million patients i.e. about 80% of people living in Scotland.
- Was developed from patient-level data, including hospital admissions, prescriptions, A&E and outpatient attendances, long-term conditions, Scottish Index of Multiple Deprivation, and Age

University of Glasgow
Oilthigh Ghlaschu

Different kinds of generative AI

- Can be text, images, video, audio
- Works in principle the same as predictive AI – using data to train a model – but new AI technology called ‘transformers’ does not require features to be defined in advance or for data to be labeled
- Large Language Models (LLMs) are the most widely known of generative AI
  - OpenAI’s GPT-4, Google’s Bard, Claude – not cheap for serious stuff
  - Open source LLMs – free but some problems with amount of data used to train and performance
- Lots of implications for how things get created
  - Copyright? Authorship?

University of Glasgow
Oilthigh Ghlaschu

An example: DALL-e

University of Glasgow
Oilthigh Ghlaschu

When does it go wrong?

- Some AI is very “accurate” – it does the right thing very often – sometimes more often than humans

- However, sometimes it fails
  - Bad or incomplete input
  - Real-world doesn’t match the historical data
  - Very complex problem
  - No objective ground truth
  - Factually incorrect / “hallucinations”

Mr Parker and Mr Fugett were both arrested for drug possession

University of Glasgow
Oilthigh Ghlaschu

Fairness, accountability, transparency, explanations (FATE)

- Undercover the “black box”
- It needs to be robust and fair
- Design it responsibly, take the risks of harm into account

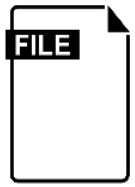
Transparency and explanations

Understanding facilitates

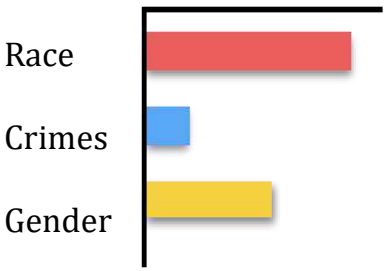
- debugging
- bias detection
- recourse

Model

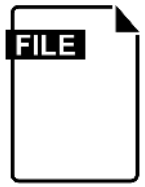
Prediction = Siberian Husky



Model



Prediction = Risky to Release



Model

Increase salary by 50K + pay credit card bills on time for next 3 months to get a loan

Prediction = Denied Loan

University of Glasgow
Oilthigh Ghlaschu

LIME: Local Interpretable Model-Agnostic Explanations

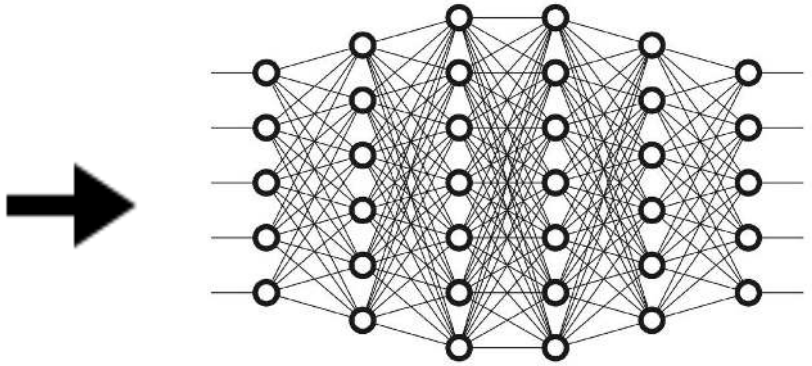
- Explains important feature that led to a decision
- Uses a post-hoc explanation on a simplified model
- Another popular method which outputs feature importances: SHAP

[Ribeiro et al. KDD 2016]

University of Glasgow
Oilthigh Ghlaschu

Saliency Maps

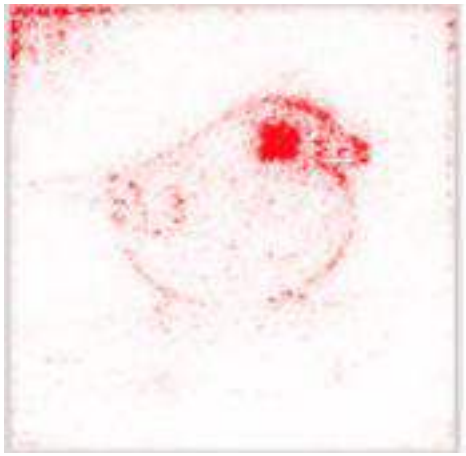
Input



Prediction

Junco Bird

What parts of the input are most relevant for the model's prediction: 'Junco Bird'?



Saliency Map

University of Glasgow
Oilthigh Ghlaschu

What is ‘accurate’?

- Depends on having ‘ground truth’ labels

- Quite often collected from human labeling

  - What if humans don’t know the true label?
  - What if humans don’t agree on a label?

  - A judge looks at offender profile and decides whether bail can be set or they should remain in prison
  - A loan officer looks at an applicant’s info and gives then a loan or not based on the risk profile

University of Glasgow
Oilthigh Ghlaschu

Bias and Fairness

- Biased humans produce biased data which gets trained into model
- AI can also go ‘rogue’ and produce a biased model
- Fairness is a human value
- Need transparency to understand if something is fair (or not)
- There are over 21+ different ways to measure fairness but they are context-dependent and people can’t agree what measure to use
  - Individual fairness: similar outcomes for similar individuals
  - Group fairness (statistical/demographic parity): people in both protected and unprotected groups have equal probability of having the positive outcome

University of Glasgow
Oilthigh Ghlaschu

Risks of harm - good or evil?

University of Glasgow
Oilthigh Ghlaschu

EU AI Act

- All AI will need to be evaluated as to its risk/harm
- If limited risk/harm, then go ahead!
- If high risk/harm ("AI systems that negatively affect safety or fundamental rights"), then might have to register with EU and be assessed throughout their lifecycle
  - AI systems that are used in products (toys, aviation, cars, medical devices and lifts)
  - Biometric identification and categorisation of natural persons
  - Management and operation of critical infrastructure
  - Education and vocational training
  - Employment, worker management and access to self-employment
  - Access to and enjoyment of essential private services and public services and benefits
  - Law enforcement
  - Migration, asylum and border control management
  - Assistance in legal interpretation and application of the law
- If unacceptable risk/harm ("threat to people"), then could be banned
- Generative AI will have to be transparent

University of Glasgow
Oilthigh Ghlaschu

Summary

- AI is starting to be everywhere and it has the real potential to transform our lives
- However there are also lots of problems with it
- Everyone should know what they are, how they work and when they can go wrong
- Everyone should be involved in some way in developing AI