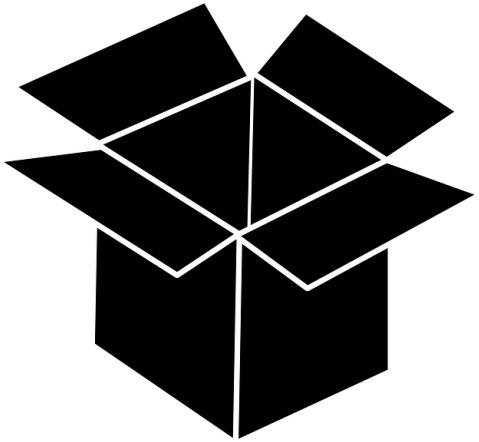
Seminar Series:

Opening the Black Box: Building Explainable AI Models



Allan Tucker
Intelligent Data Analytics
Computer Science

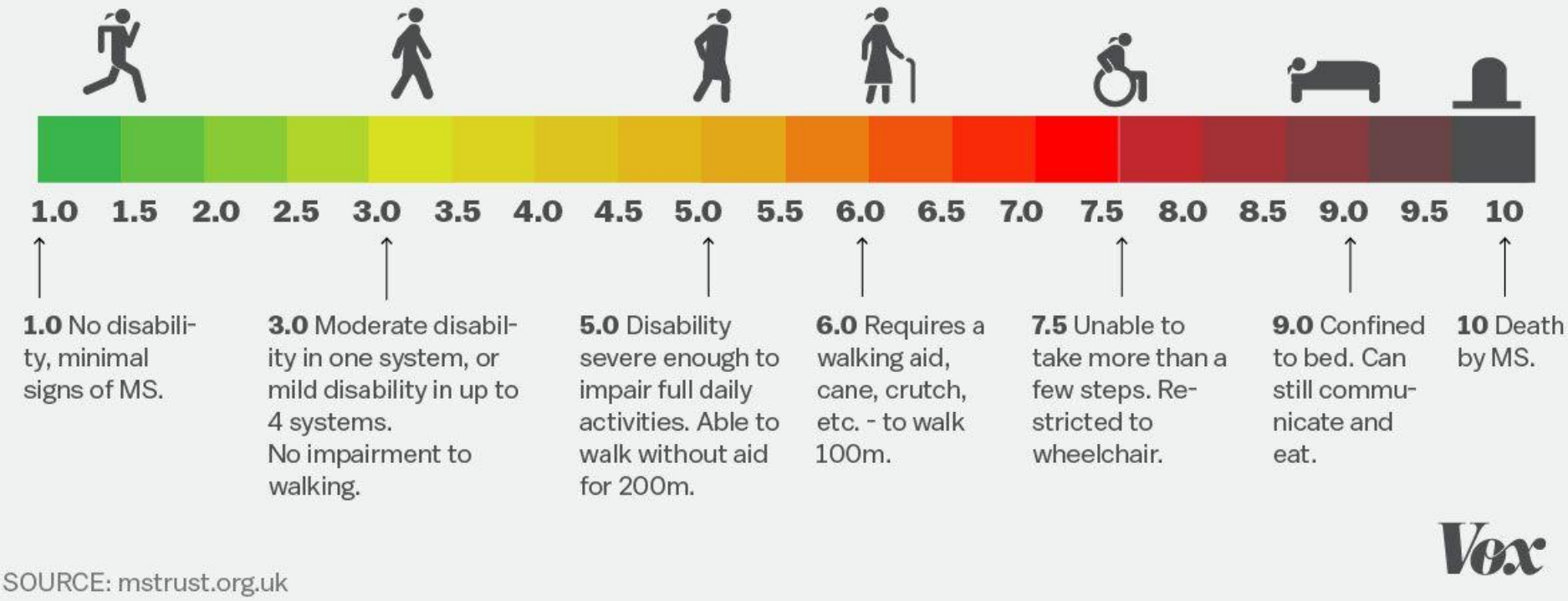
Brunel University London

Focus on “AI” in Healthcare

- Many new challenges in the health sector
- But also new technologies to deal with them:
  - Explosion in available data
  - Leading to a boom in computationally heavy analyses
  - Development of new “AI” / Machine Learning technologies

Brunel University London

“Better” Diagnose / Manage Disease

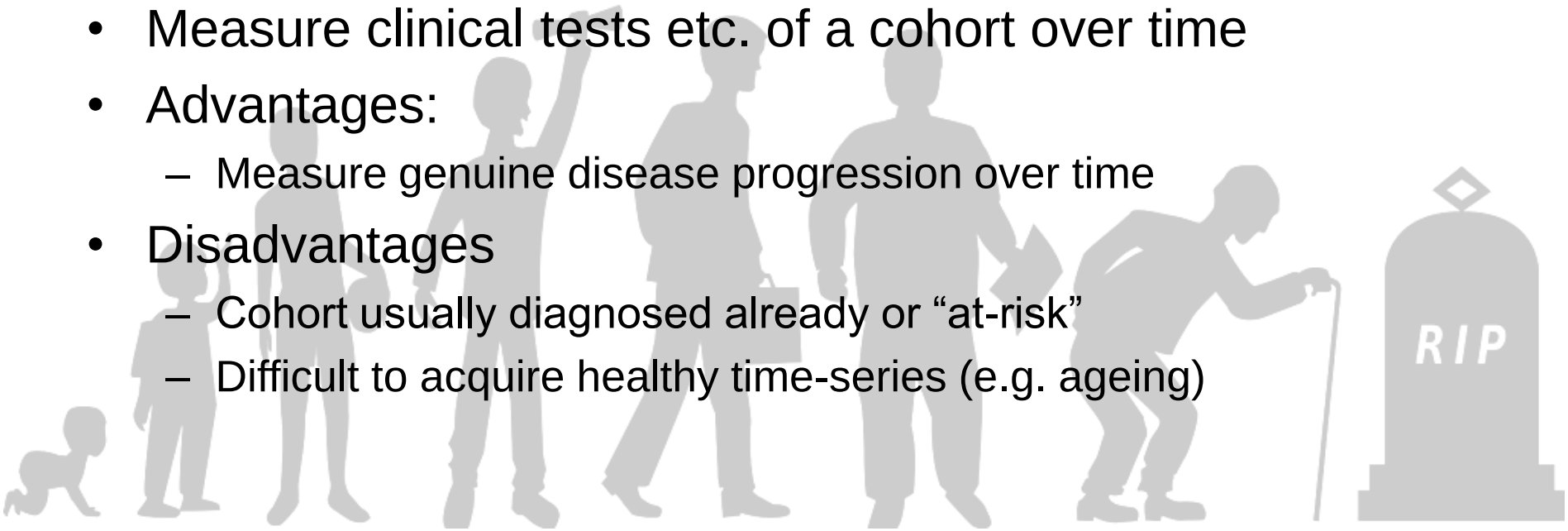


Longitudinal Studies

Brunel University London

Longitudinal Studies

- Measure clinical tests etc. of a cohort over time
- Advantages:
  - Measure genuine disease progression over time
- Disadvantages
  - Cohort usually diagnosed already or “at-risk”
  - Difficult to acquire healthy time-series (e.g. ageing)



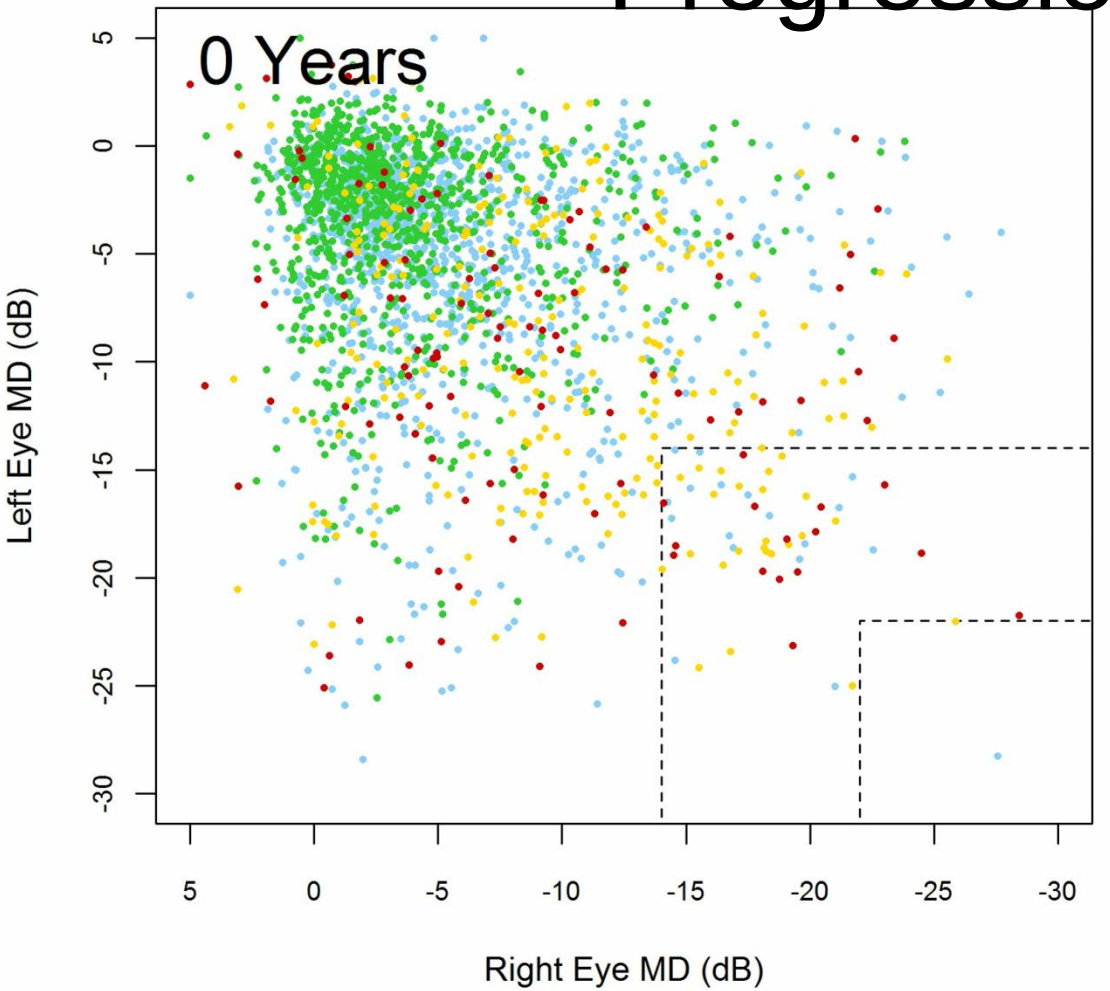
Brunel University London

Cross Section Studies

Cross Section Studies

- Record attributes across a (“large”) sample
- Carried out at a single window of time
- Gives a “snapshot” of a disease over the population
- Advantages:
  - No issues in following up
  - Captures diversity of disease in large cohort
  - Can capture genuine healthy and v early stages of disease
- Disadvantages:
  - No measure of temporal characteristics of disease

Degenerative Disease Progression



Saunders et al. IOVS 2014

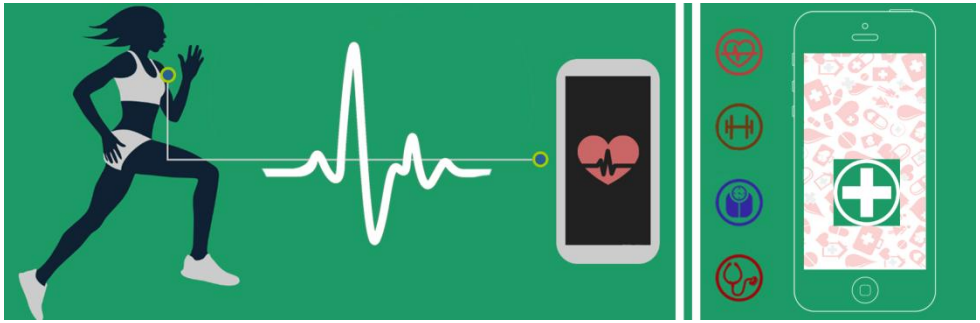
Brunel University London

New Sources of Data

- Larger studies e.g. Biobanks, Longitudinal
- Health Apps: Activity, heart rate, sleep, etc.
- Life Style / Environmental data
- But …

  Noisy, Bias etc.

biobank uk
Improving the health of future generations



Brunel University London

New Sources of Data

- Larger studies e.g. Biobanks, Longitudinal
- Health Apps: Activity, heart rate, sleep, etc.
- Life Style / Environmental data
- But …

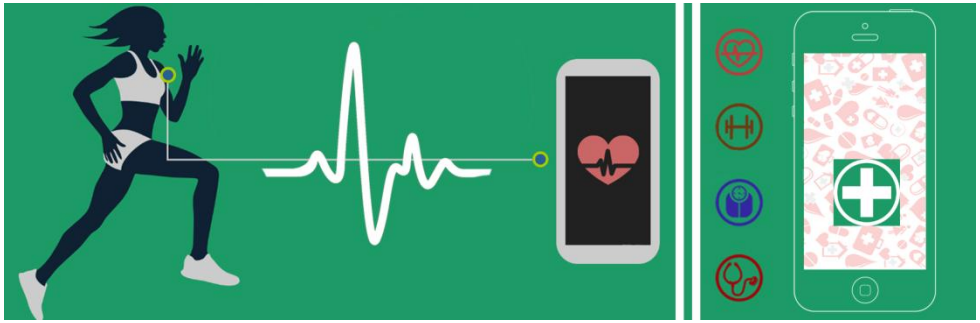
  Noisy, Bias etc.

biobank uk
Improving the health of future generations

HUFFPOST

The Poor Aren't Using Health Tracking Apps

01/17/2018 01:41 am ET



Brunel University London

“Artificial Intelligence” in Healthcare

“We think that machine learning technology, a type of artificial intelligence, can bring huge benefits to medical research. By using this technology to analyse medical data, we want to find ways to improve how illnesses are diagnosed and treated.”

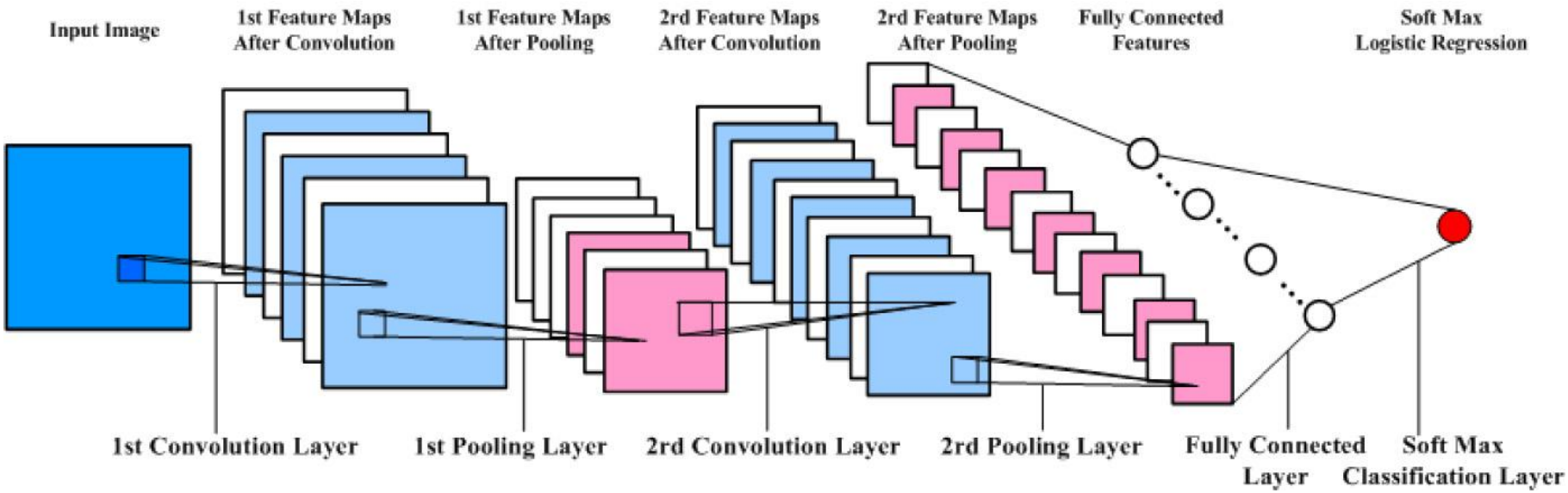
Google DeepMind

“Artificial Intelligence” in Healthcare

“Support better-informed, more effective patient care, health plans, wellness programs ... factors that influence a person’s health -including socioeconomic status, environment, social support and access to health care.”

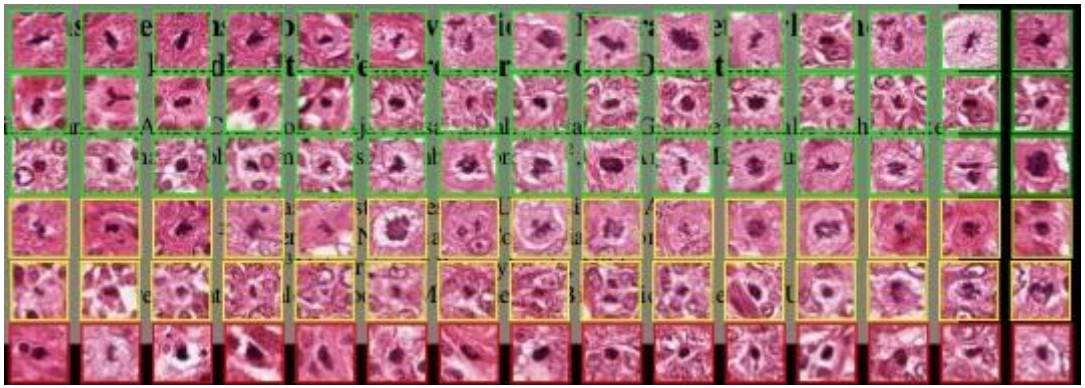
IBM Watson Health

“Deep” Methods & Hidden Variables



Brunel University London

“Deep” Methods & Hidden Variables



Artificial intelligence is now better at spotting deadly cancers than human doctors

Jasper Hamill Wednesday 30 May 2018 1:47 pm

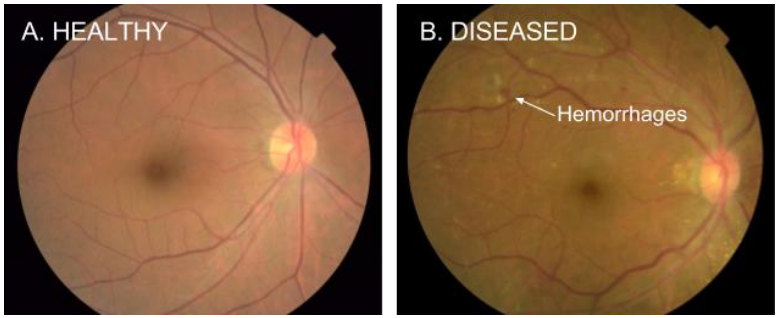
Computer trained to spot deadly malignant moles could save a lot of people's lives (Photo: Getty Images)

UK politics Environment Education **Society** Science Tech Global development Cities

London hospitals to replace doctors and nurses with AI for some tasks

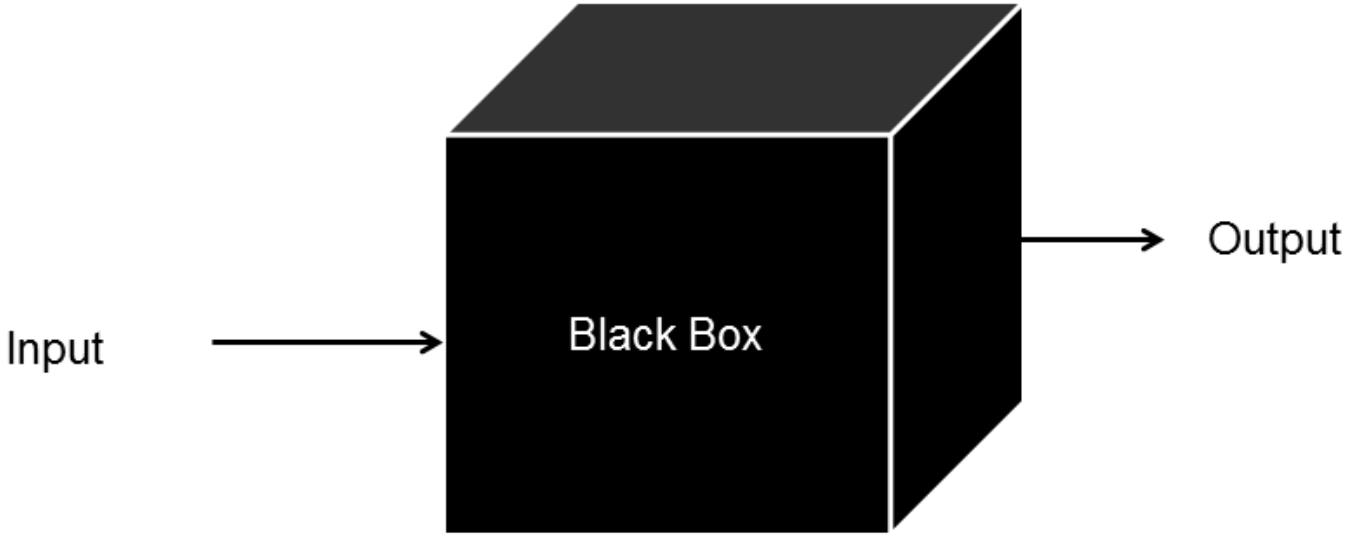
UCLH aims to bring ‘game-changing’ benefits of artificial intelligence to NHS patients, from cancer diagnosis to reducing wait times

▲ Machine learning could be applied to the analysis of patient scans that are usually checked by hospital staff. Photograph: Juice/REX/Shutterstock



Brunel University London

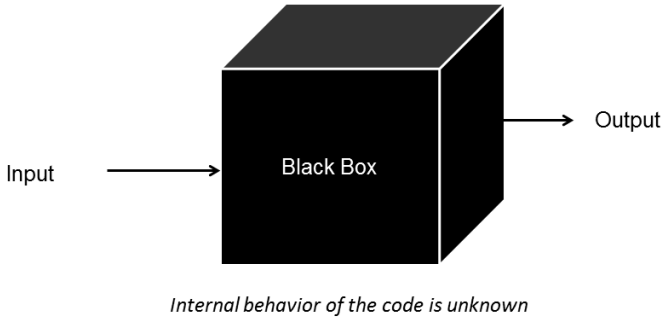
“Artificial Intelligence” in Healthcare



Internal behavior of the code is unknown

Brunel University London

Black Box – Two major reasons

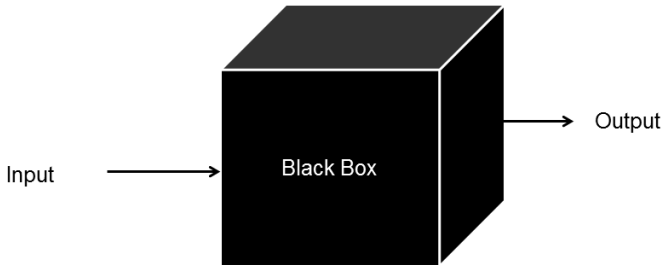


1. Commercially sensitive
   - Big business
   - Algorithms as commodities

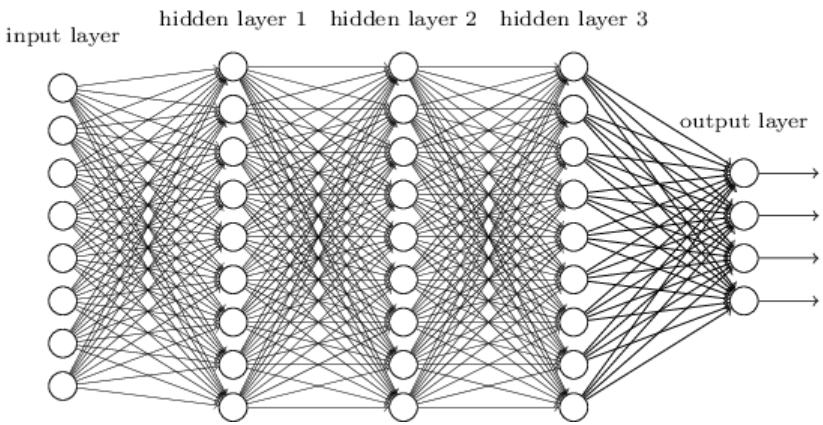
Brunel University London

Black Box – Two major reasons

2. Too complex for us to understand
   - Massively parallel
   - Huge numbers of parameters



Internal behavior of the code is unknown



Brunel University London

Do we care?

"I don't care if the decision cannot be explained if it is better than a human"

The Geoff Hinton "Is this a 2?" argument

Brunel University London

Do we care?

"I don't care if the decision cannot be explained if it is better than a human"

The Geoff Hinton "Is this a 2?" argument

Brunel University London

Automation Bias

Brunel University London

Automation Bias

Brunel University London

General Data Protection Reg. 2018

Automatic Zoom

Rights related to automated decision making and profiling

In brief...

The GDPR provides safeguards for individuals against the risk that a potentially damaging decision is taken without human intervention. These rights work in a similar way to existing rights under the DPA.

Identify whether any of your processing operations constitute automated decision making and consider whether you need to update your procedures to deal with the requirements of the GDPR.

In more detail...

When does the right apply?

Individuals have the right *not to be subject to a decision* when:

- it is based on automated processing; and
- it produces a legal effect or a similarly significant effect on the individual.

You must ensure that individuals are able to:

- obtain human intervention;
- express their point of view; and
- obtain an explanation of the decision and challenge it.

Does the right apply to all automated decisions?

No. The right does not apply if the decision:

- is necessary for entering into or performance of a contract between you and the individual;
- is authorised by law (eg for the purposes of fraud or tax evasion prevention); or
- based on explicit consent. (Article 9(2)).

Furthermore, the right does not apply when a decision does not have a legal or similarly significant effect on someone.

Brunel University London

Urgent need to open the black box

- We need to know the underlying mechanisms of the black box to
  - Gain trust of clinicians
  - Gain new insights
  - Make better decisions / interventions

"Transactions that are too complex to explain…may well be too complex to be allowed to exist" Tom Hamburger, Washington Post, Apr 12 2014.

Brunel University London

Urgent need to open the black box

- We need to know the underlying mechanisms of the black box to
  - Gain trust of clinicians
  - Gain new insights
  - Make better decisions / interventions

AI researchers allege that machine learning is alchemy

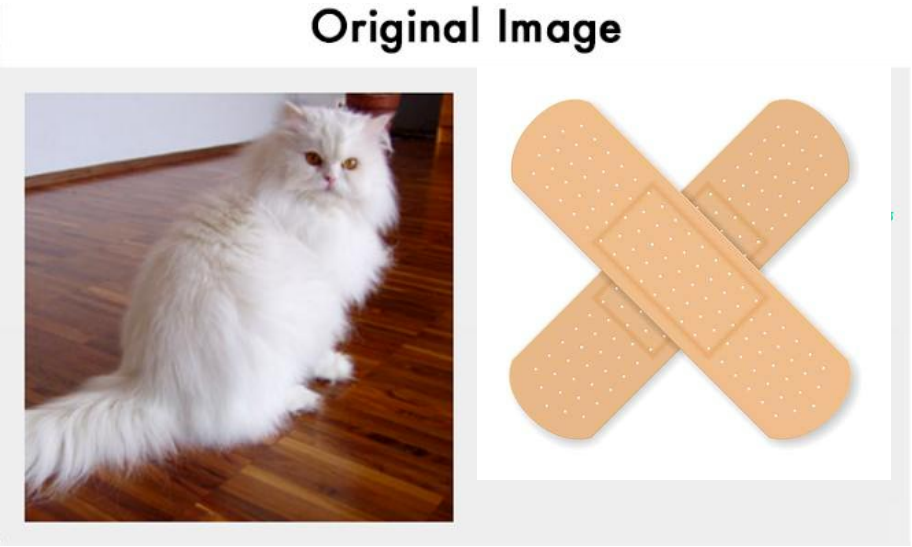
By Matthew Hutson | May. 3, 2018 , 11:15 AM

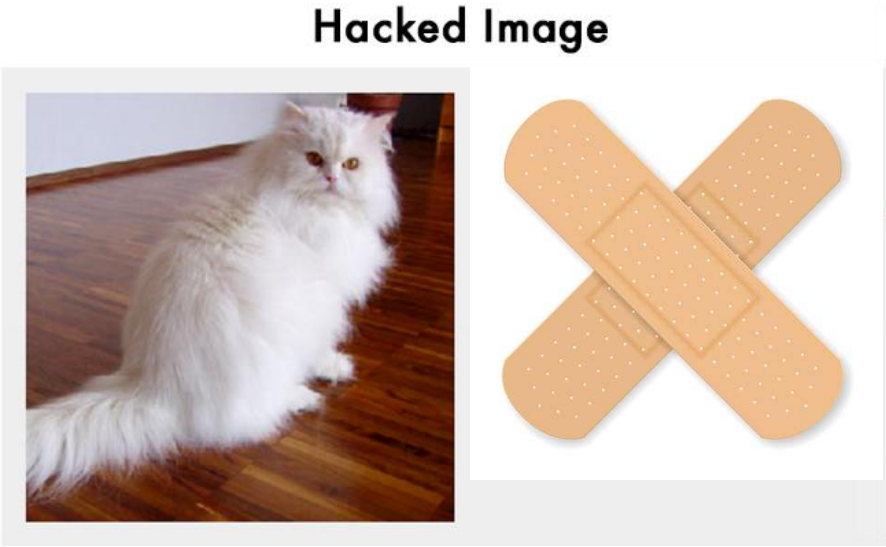
Ali Rahimi, a researcher in artificial intelligence (AI) at Google in San Francisco, California, took a swipe at his field last December—and received a 40-second ovation for it. Speaking at an AI conference, Rahimi charged that machine learning algorithms, in which computers learn through trial and error, **have become a form of "alchemy."** Researchers, he said, do not know why some algorithms work and others don't, nor do they have rigorous criteria for choosing one AI architecture over another. Now, in a paper presented on 30 April at the International Conference on Learning Representations in Vancouver, Canada, Rahimi and his collaborators **document examples** of what they see as the alchemy problem and offer prescriptions for bolstering AI's rigor.

"There's an anguish in the field," Rahimi says. "Many of us feel like we're operating on an alien technology."

Brunel University London

Easy to trick a DNN



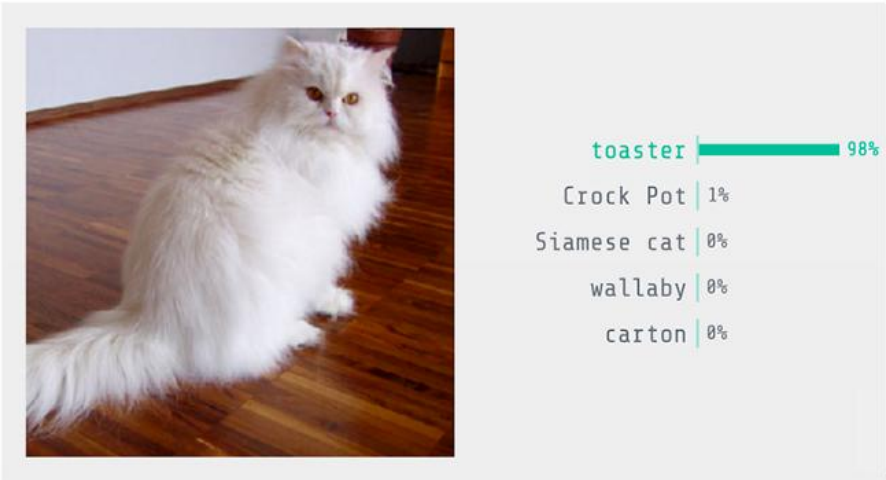


Brunel University London

Easy to trick a DNN

Original Image

Hacked Image

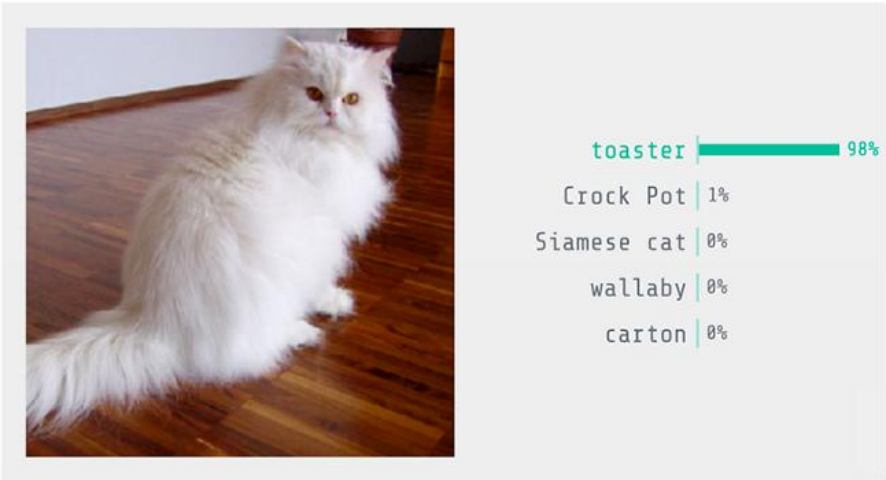


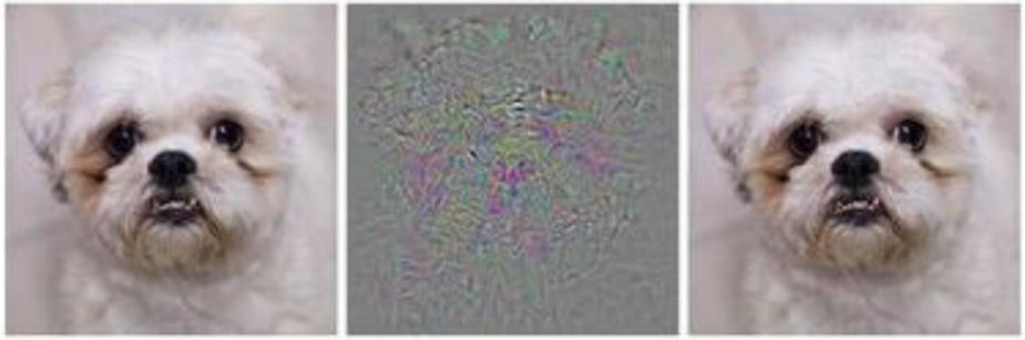
Brunel University London

Easy to trick a DNN

Original Image

Hacked Image



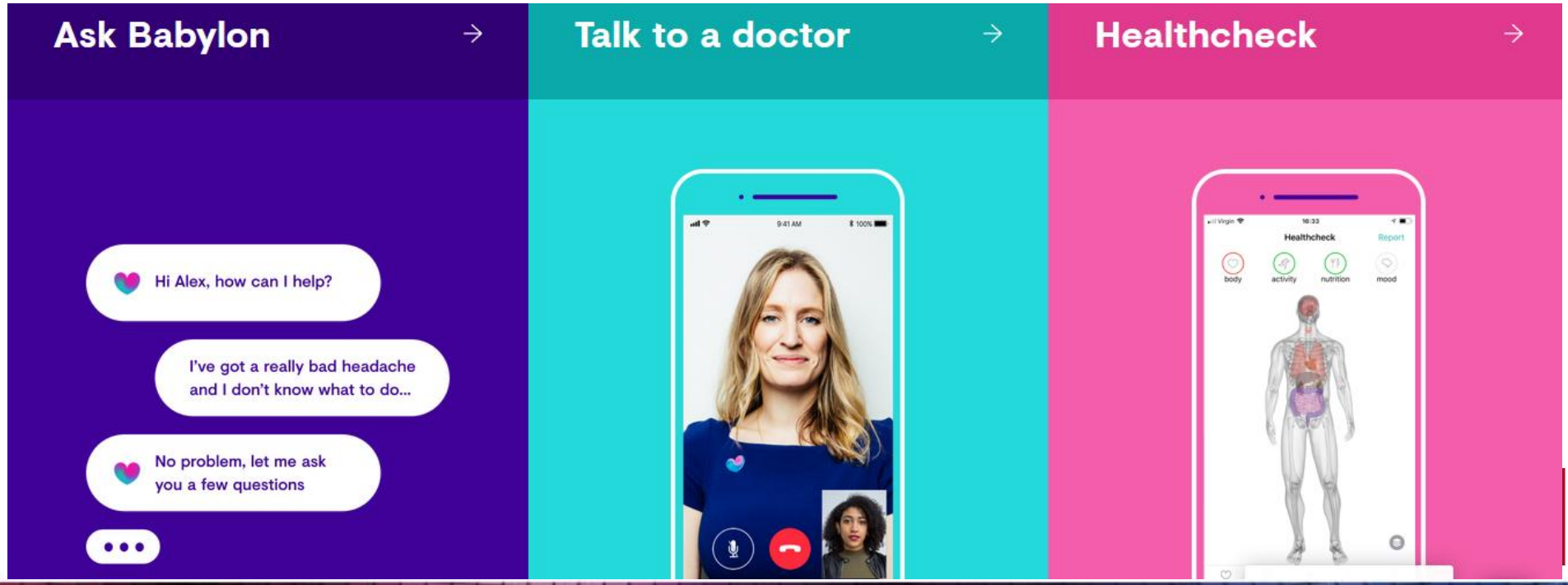


Prediction: **Dog** + Distortion Prediction: **Ostrich**

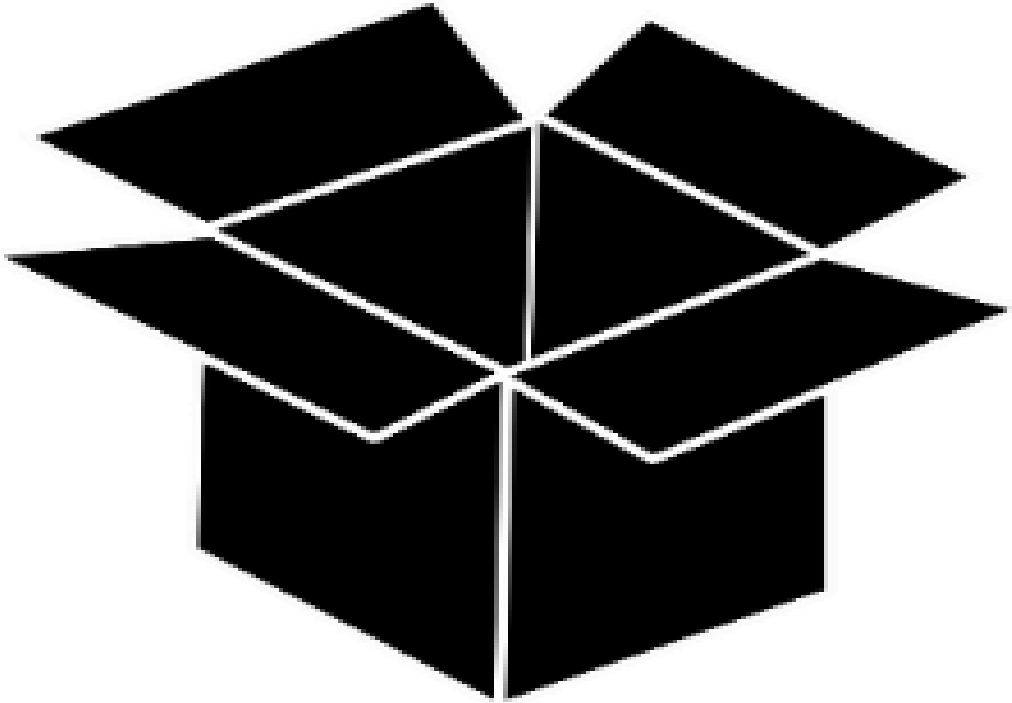
Brunel University London

The Importance of Validating Models

babylon



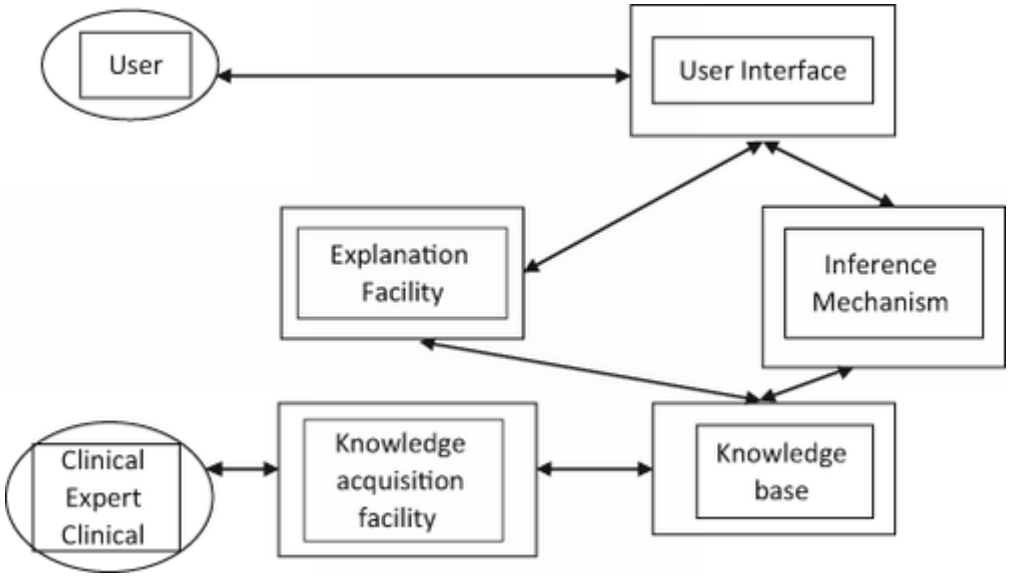
Opening the Black Box



Brunel University London

Existing Approaches: Rule Based ES

Rule-Based Expert Systems eg. MYCIN



Brunel University London

Existing Approaches: Rule Based ES

Rule-Based Expert Systems eg. MYCIN

1. Patient Information
1) Patient's name
PT538
2) Age
34 years
3) Sex
Male

Diagnosis
6) Please enter results of pending cultures in table:
SITE CULTURE# DATE EXAMINED
CSF 1234 30.1 YES
10) Do you suspect PT538 may have an infection at a site from which you have not obtained specimens?
No
15) Has PT538 recently had objective evidence of abnormal neurologic signs (e.g. seizures, coma) documented by physician?
Yes
The CSF culture will be considered to be associated with meningitis.
Cause
17) Does PT538 have an abnormal chest x-ray?
No
18) Has PT538 been exposed to any contagious diseases recently?

Existing Approaches: Argumentation

If it is believed that *belief*$_1$, ……, *belief*$_n$ is the case
Then we should do action a
Since this will result in effect *e* being the case
Which will realise our desired goal *g*.

Argument A1 =
If it is believed that the patient has had a myocardial infarct
Then we should administer aspirin
Since this will result in reduced platelet adhesion
Which will prevent blood clotting.

Argument A2 = If it is believed that the patient has had a myocardial infarct
Then we should administer chlopidogrel
Since this will result in reduced platelet adhesion
Which will prevent blood clotting.

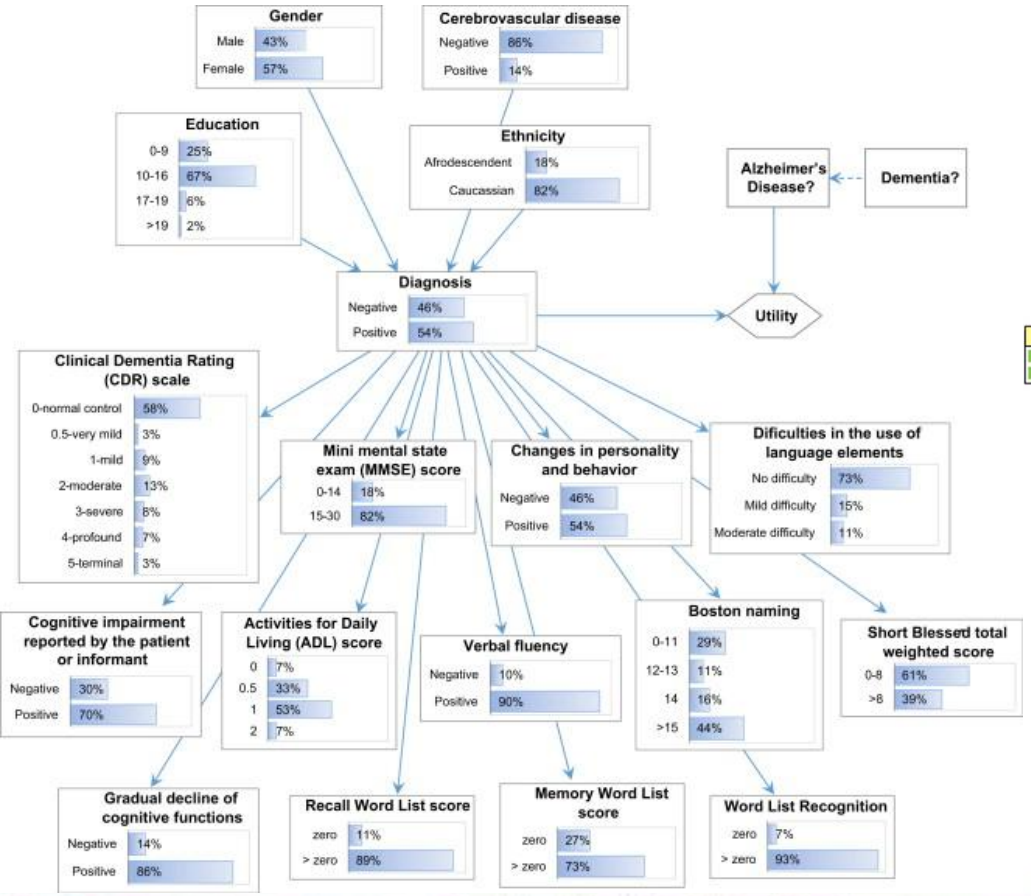
Argument A3 = If it is believed that the patient has a history of gastritis
So that administering aspirin
Will risk gastric bleeding.

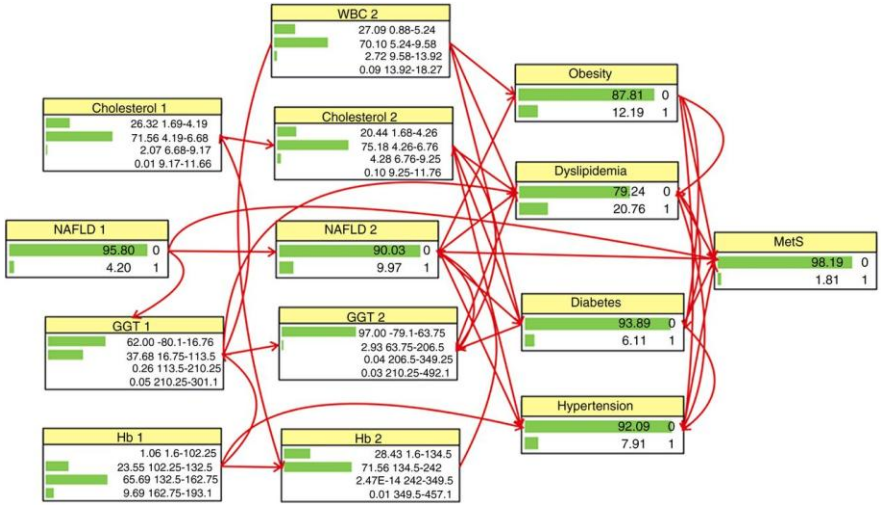
A1 and A2 conflict with each other, but A1 is a stronger argument than A2 since a clinical trial indicates that aspirin is more efficacious at preventing blood clotting than chlopidogrel. Hence A2 defeats A1. However, there is an argument A3 that defeats A1 on the grounds that aspirin results in an unwanted side-effect. Hence, argument A2 is reinstated as the winning argument, i.e., chlopidogrel is the preferred choice of action.

Brunel University London

Existing Approaches: Bayesian

eg. Bayesian Networks





Brunel University London

Existing Approaches: Recommenders

INPUT

4	3			5	
5		4		4	
4		5	3	4	
	3				5
	4				4
		2	4		5

Introduction to Recommender Systems
Machine learning Paradigms
Social Network-based Recommender Systems
Learning Spark
Recommender Systems Handbook
Recommender Systems and the Social Web

USER-BASED COLLABORATIVE FILTERING

1.00	0.75	0.63	0.22	0.30	0.00
0.75	1.00	0.91	0.00	0.00	0.16
0.63	0.91	1.00	0.00	0.00	0.40
0.22	0.00	0.00	1.00	0.97	0.64
0.30	0.00	0.00	0.97	1.00	0.53
0.00	0.16	0.40	0.64	0.53	1.00

(0.7 x) + (0.6 x) =

already rated by user

(0.7 x 4 + 0.6 x 5) / (0.7 + 0.6) = 4.5

(0.6 x 3) / 0.6 = 3.0

already rated by user

ITEM-BASED COLLABORATIVE FILTERING

1.00	0.27	0.79	0.32	0.98	0.00
0.27	1.00	0.00	0.00	0.34	0.65
0.79	0.00	1.00	0.69	0.71	0.18
0.32	0.00	0.69	1.00	0.32	0.49
0.98	0.34	0.71	0.32	1.00	0.00
0.00	0.65	0.18	0.49	0.00	1.00

(4 x) + (3 x) + (5 x) =

already rated by user

(0.8 x 4 + 0.7 x 5) / (0.8 + 0.7) = 4.5

already rated by user

(0.7 x 3) / 0.7 = 3.0

CONTENT-BASED FILTERING

1.00	0.00	0.58	0.00	0.67	0.58
0.00	1.00	0.00	0.41	0.00	0.00
0.58	0.00	1.00	0.00	0.58	0.75
0.00	0.41	0.00	1.00	0.00	0.00
0.67	0.00	0.58	0.00	1.00	0.58
0.58	0.00	0.75	0.00	0.58	1.00

(4 x) + (3 x) + (5 x) =

already rated by user

(0.4 x 3) / 0.4 = 3.0

already rated by user

(0.6 x 4 + 0.6 x 5) / (0.6 + 0.6) = 4.5

HYBRID

0.4 x UB CF) + (0.3 x IB CF) + (0.3 x CB) =

(0.4 x 4.5 + 0.3 x 4.5) / (0.4 + 0.3) = 4.5

(0.3 x 3.0 + 0.3 x 4.5) / (0.3 + 0.3) = 3.8

(0.4 x 3.0 + 0.3 x 3.0) / (0.4 + 0.3) = 3.0

OUTPUTS

FOUR RECOMMENDER ALGORITHMS ARE FED THE SAME INPUT AND PRODUCE DIFFERENT OUTPUTS.
THEIR DEFINITIONS OF USER AND/OR ITEM SIMILARITY ACCOUNT FOR MOST OF THE DIFFERENCES.

https://buildingrecommenders.wordpress.com

Existing Approaches: Counterfactuals

Counterfactuals:

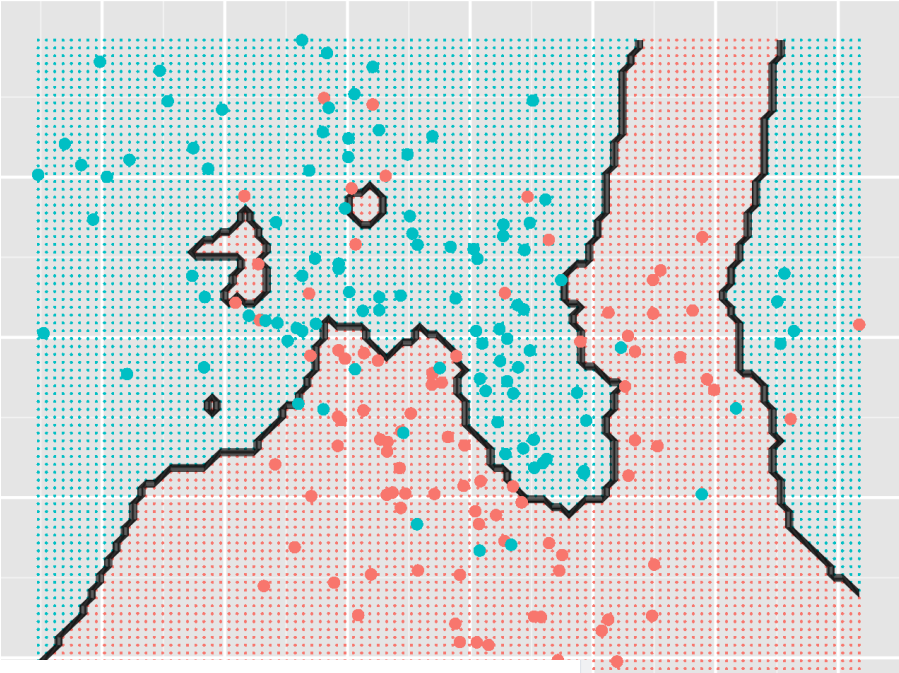
"You are classed as High-Risk of Disease A"

"If you want to be re-classed Low-Risk, then you must

i) Stop smoking

or

ii) Change Diet & ....."



Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR

Harvard Journal of Law & Technology, Forthcoming

52 Pages • Posted: 2 Nov 2017 • Last revised: 24 Mar 2018

Sandra Wachter
University of Oxford - Oxford Internet Institute

Brent Mittelstadt
University of Oxford - Oxford Internet Institute

Chris Russell
University of Surrey

Date Written: October 6, 2017

Brunel University London

Warning Advert!

IDA Research

Intelligent Data Analysis (est. 1995): focus on exploiting intelligence of:

data experts

analysts

within the algorithm

Not Black Box!

Brunel University London

Existing Approaches: Trajectory Analysis

NCBI Resources How To

PubMed.gov US National Library of Medicine National Institutes of Health — PubMed — Advanced

Format: Abstract — Send to

IEEE Trans Inf Technol Biomed. 2010 Jan;14(1):79-85. doi: 10.1109/TITB.2009.2023319. Epub 2009 Jun 12.

The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data.

Tucker A[1], Garway-Heath D.

Author information

1 School of Information Systems Computing andMaths, Brunel University, Uxbridge UB8 3PH, UK. allan.tucker@brunel.ac.uk

Abstract

Progressive loss of the field of vision is characteristic of a number of eye diseases such as glaucoma, a leading cause of irreversible blindness in the world. Recently, there has been an explosion in the amount of data being stored on patients who suffer from visual deterioration, including visual field (VF) test, retinal image, and frequent intraocular pressure measurements. Like the progression of many biological and medical processes, VF progression is processes are often cross sectional and the time dim address this issue by developing a method to build a involves building trajectories through all of the data th otherwise be impossible without longitudinal data). G there will be a number of key trajectories that are imp idea of pseudo time series by using resampling techr handles outliers and multiple possible disease traject present very promising results on VF data for predicti

ELSEVIER

Journal of Biomedical Informatics

Volume 46, Issue 2, April 2013, Pages 266-274

Modelling and analysing the dynamics of disease progression from cross-sectional studies

Yuanxi Li, Stephen Swift, Allan Tucker

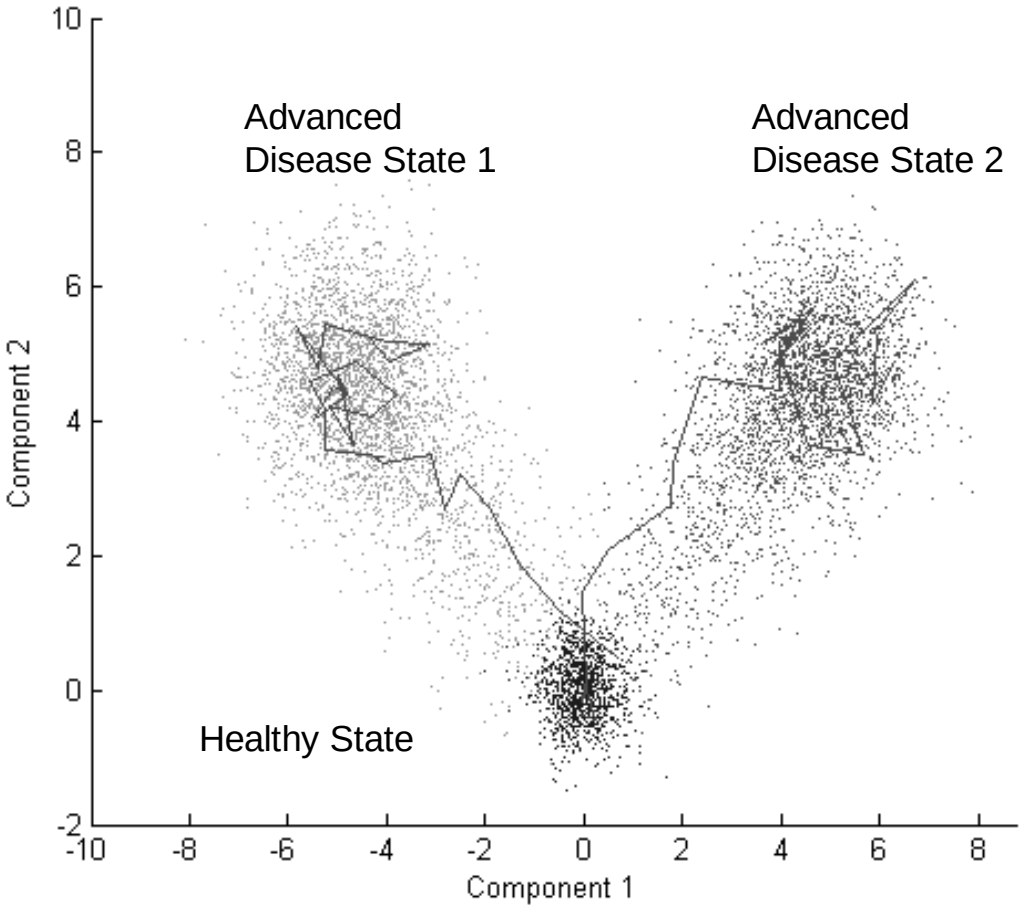
Show more

https://doi.org/10.1016/j.jbi.2012.11.003 — Get rights and content

Under an Elsevier user license — open archive

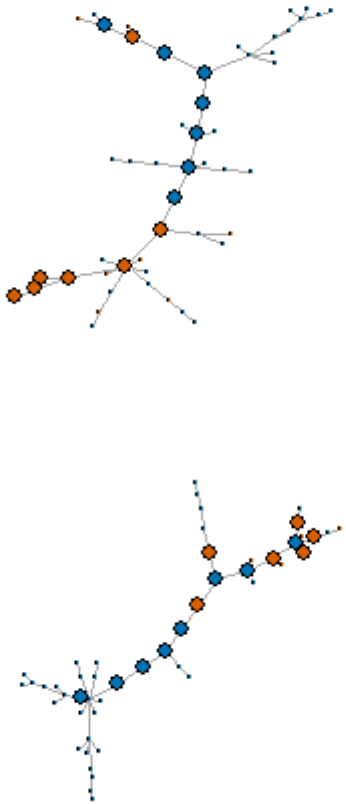
Abstract

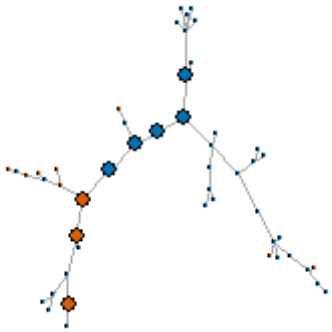
Clinical trials are typically conducted over a population within a defined time period in order to illuminate certain characteristics of a health issue or disease process. These cross-sectional studies give us a 'snapshot' of this disease process over a large number of people but do not allow us to model the temporal nature of disease, thereby allowing for modelling detailed prognostic predictions. The aim of this paper is to explore an extension of the temporal bootstrap to identify intermediate stages in a disease process and sub-categories of the disease exhibiting subtly different symptoms. Our approach is compared to a strawman method and investigated in its ability to explain the dynamics of progression on biomedical data from three diseases: Glaucoma, Breast Cancer and Parkinson's disease. We focus on creating reliable time-series models from large amounts of historical cross-sectional data using the temporal bootstrap technique. Two issues are explored: how to build time-series models from cross-sectional data, and how to automatically identify different disease states along these trajectories, as well as the transitions between them.

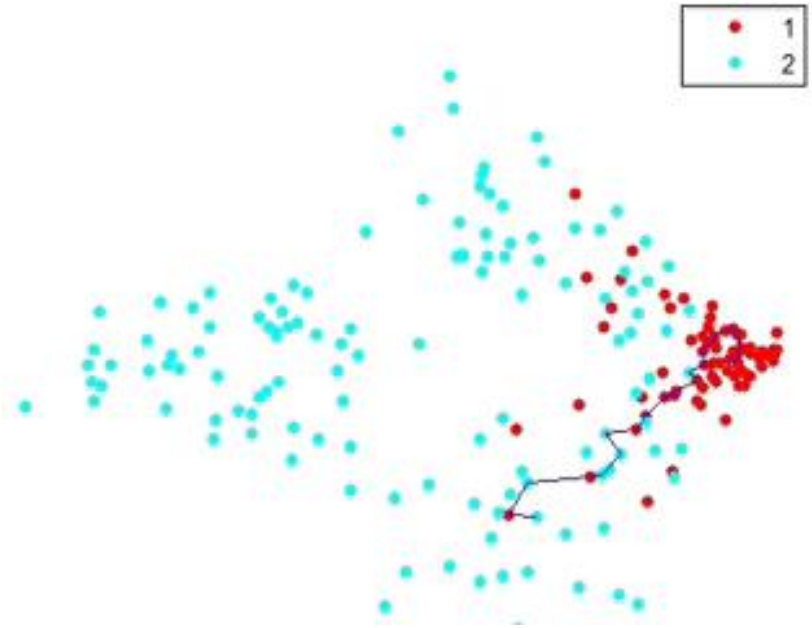


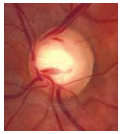
IDA Research

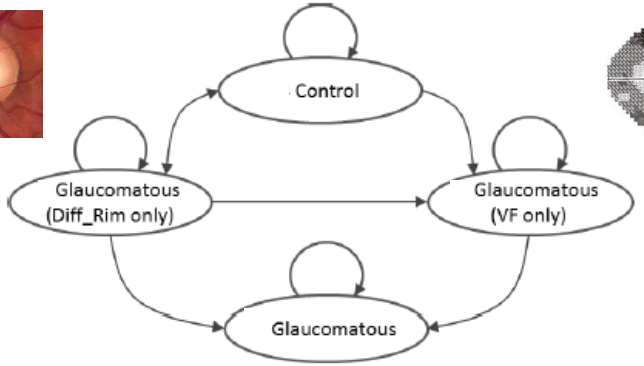
Existing Approaches: Trajectory Analysis

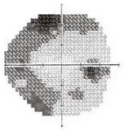










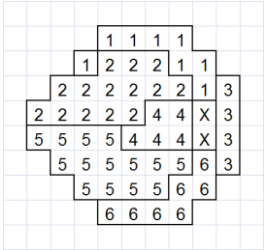


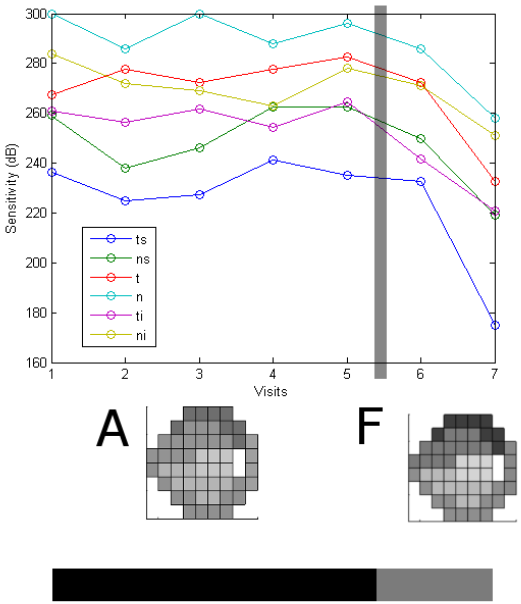
IDA Research

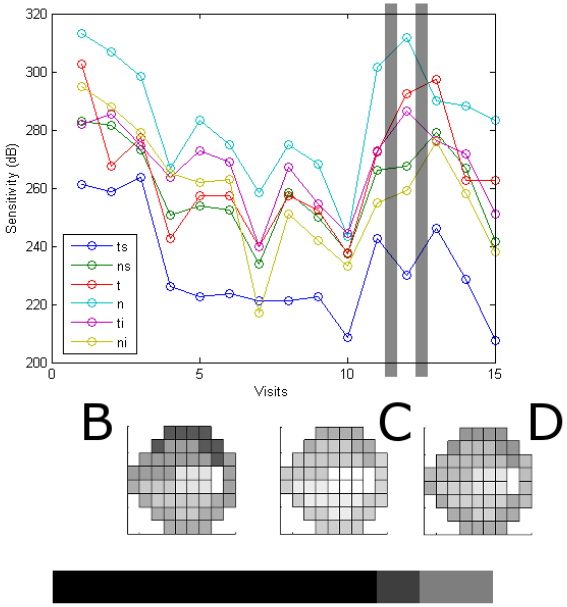
Brunel University London

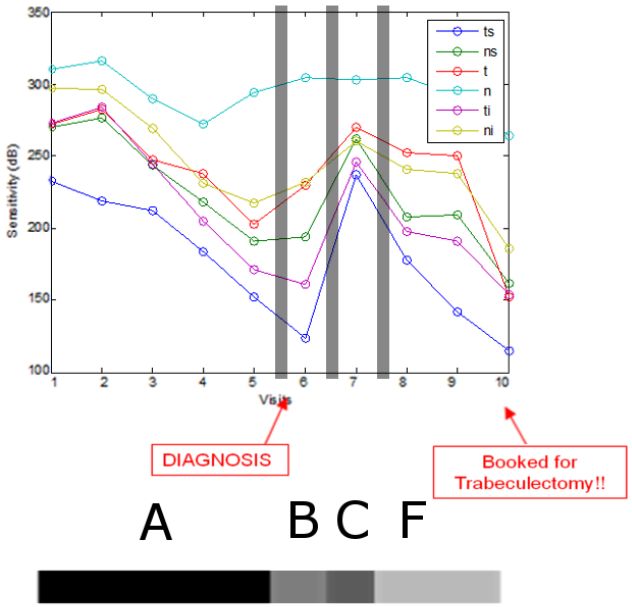
Existing Approaches: Latent Variables

ICML 2012 Workshop:
ML for Clinical Data Analysis





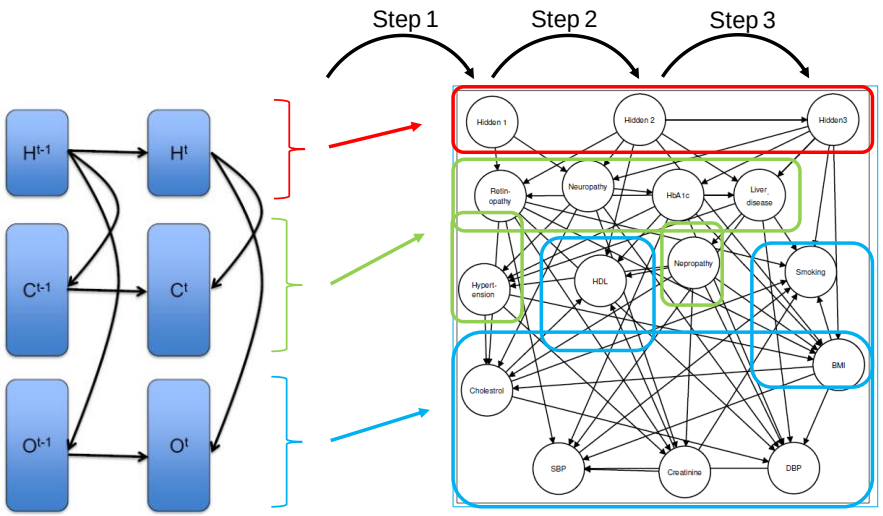


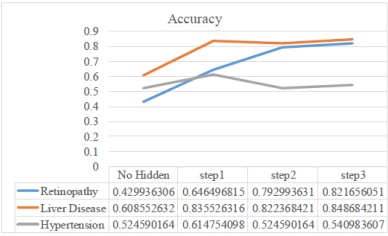


IDA Research

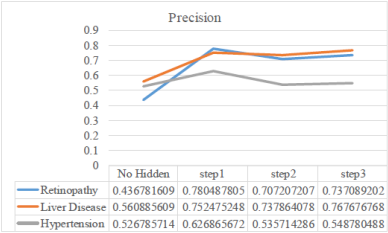
Brunel University London

Existing Approaches: Latent Variables

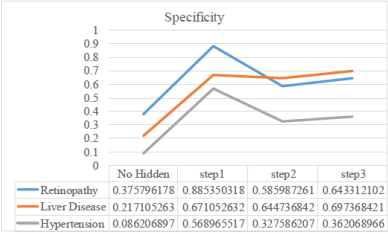




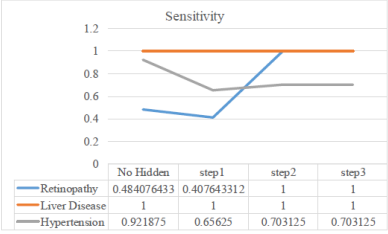
	No Hidden	step1	step2	step3
Retinopathy	0.429936306	0.646496815	0.792993631	0.821656051
Liver Disease	0.608552632	0.835526316	0.822368421	0.848684211
Hypertension	0.524590164	0.614754098	0.524590164	0.540983607



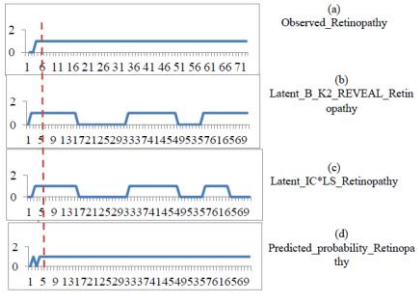
	No Hidden	step1	step2	step3
Retinopathy	0.436781609	0.780487805	0.707207207	0.737089202
Liver Disease	0.560885609	0.752475248	0.737864078	0.767676768
Hypertension	0.526785714	0.626865672	0.535714286	0.548780488

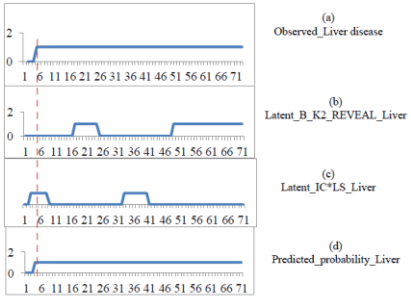


	No Hidden	step1	step2	step3
Retinopathy	0.375796178	0.885350318	0.585987261	0.643312102
Liver Disease	0.217105263	0.671052632	0.644736842	0.697368421
Hypertension	0.086206897	0.568965517	0.327586207	0.362068966



	No Hidden	step1	step2	step3
Retinopathy	0.484076433	0.407643312	1	1
Liver Disease	1	1	1	1
Hypertension	0.921875	0.65625	0.703125	0.703125





IDA Research

Brunel University London

Summary

- Modern AI methods offer great potential
- Run the risks of
  - Over hype (eg. “the self driving car”)
  - Over-reliance on models that have no “humans in the loop”: automation bias
  - Losing trust of the public
- Already technologies that can help …
- Research needed focussing on explanation

Brunel University London

Measuring Success

- Difficult to measure “explanation”
- Evidence that something has been explained if:
  - Some new piece of knowledge gained
  - Change in the way a process is implemented
  - Change in specific decisions

Brunel University London

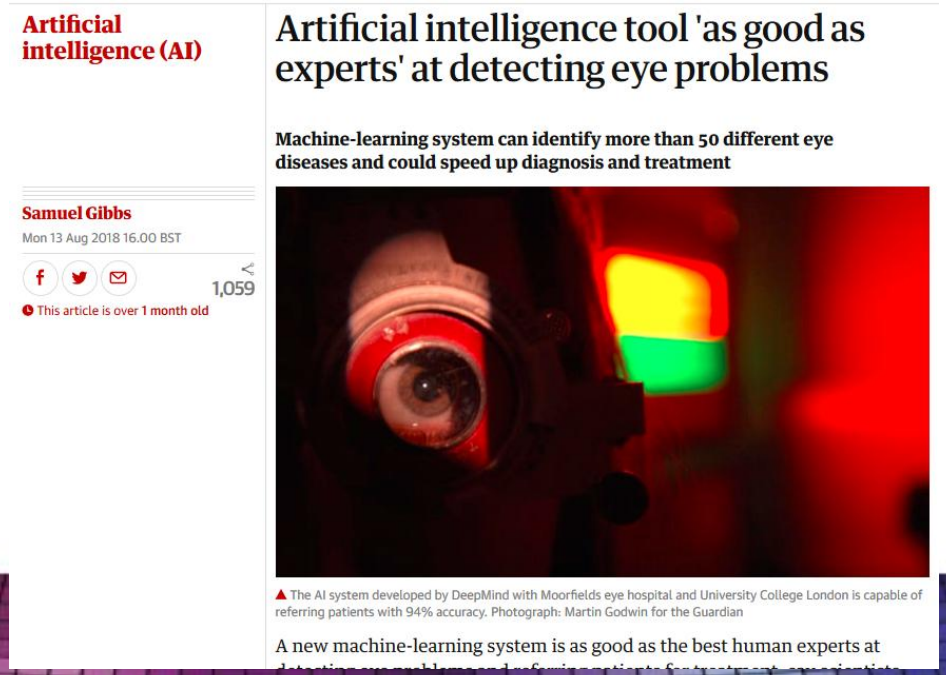
Seminar Series

- **October 17th**: Allan Tucker, “Opening the Black Box”, Brunel University London
- **November 21st**: Niels Peek, “Learning Health Systems”, University of Manchester
- **December 12th**: Pearse Keane, Moorfields Eye Hospital (in collaboration with Google Deepmind)
- **January 16th**: Norman Fenton, Queen Mary, University of London
- **TBC**: Pedro Rodrigues, University of Porto

Brunel University London

Seminar Series

- **October 17th**: Allan Tucker, “Opening the Black Box”, Brunel University London
- **November 21st**: Niels Peek, “Learning Health Systems”, University of Manchester
- **December 12th**: Pearse Keane, Moorfields Eye Hospital (in collaboration with Google Deepmind)
- **January 16th**: Norman Fenton, Queen Mary, University of London
- **TBC**: Pedro Rodrigues, University of Porto



Artificial intelligence (AI)

Samuel Gibbs

Mon 13 Aug 2018 16.00 BST

1,059

This article is over 1 month old

Artificial intelligence tool 'as good as experts' at detecting eye problems

Machine-learning system can identify more than 50 different eye diseases and could speed up diagnosis and treatment

The AI system developed by DeepMind with Moorfields eye hospital and University College London is capable of referring patients with 94% accuracy. Photograph: Martin Godwin for the Guardian

A new machine-learning system is as good as the best human experts at

Brunel University London

Seminar Series

- **October 17th**: Allan Tucker, “Opening the Black Box”, Brunel University London
- **November 21st**: Niels Peek, “Learning Health Systems”, University of Manchester
- **December 12th**: Pearse Keane, Moorfields Eye Hospital (in collaboration with Google Deepmind)
- **January 16th**: Norman Fenton, Queen Mary, University of London
- **TBC**: Pedro Rodrigues, University of Porto

Law

A formula for justice

Bayes' theorem is a mathematical equation used in court cases to analyse statistical evidence. But a judge has ruled it can no longer be used. Will it result in more miscarriages of justice?

Angela Saini
Sun 2 Oct 2011 21.30 BST

292 69

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

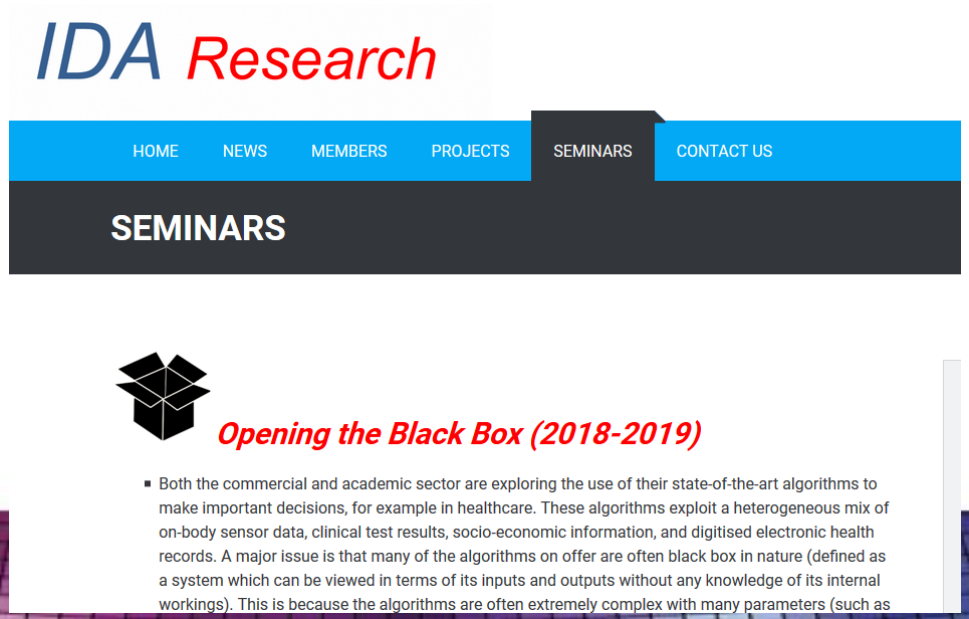
▲ Bayes' theorem. Photograph: guardian.co.uk

It's not often that the quiet world of mathematics is rocked by a murder case. But last summer saw a trial that sent academics into a tailspin, and has since swollen into a fevered clash between science and the law.

Brunel University London

Seminar Series

- **October 17th**: Allan Tucker, “Opening the Black Box”, Brunel University London
- **November 21st**: Niels Peek, “Learning Health Systems”, University of Manchester
- **December 12th**: Pearse Keane, Moorfields Eye Hospital (in collaboration with Google Deepmind)
- **January 16th**: Norman Fenton, Queen Mary, University of London
- **TBC**: Pedro Rodrigues, University of Porto



Brunel University London

Special Track at IEEE CBMS 2019

32th IEEE CBMS International Symposium on Computer-Based Medical Systems

Brunel University London

Special Track at IEEE CBMS 2019

Important dates for IEEE CBMS 2019:

- **Deadline for special track and tutorial proposal:** September 24, 2018
- **Special track and tutorial notification acceptance:** October 1, 2018
- **Paper submission:** January 14, 2019
- **Notification of acceptance:** March 1, 2019
- **Camera-ready due:** March 15, 2019
- **Author registration at the conference:** March 15, 2019 (same day as camera-ready due)
- **CBMS 2019**: June 5 – 7, 2019

These dates are for any kind of submission at CBMS, including special tracks that will have the same deadlines.

Brunel University London