MANCHESTER 1824

The University of Manchester

Mining Science Data for Medicine

Steve Watts

School of Physics and Astronomy
The University of Manchester
Manchester, UK

Background to talk – Four connecting parts.

Q. How did a particle physicist get involved in ML for medicine ?

A. Network in Radiotherapy associated with a new proton beam therapy centre

Part 1) Spin Out from this has been ML projects with medical profession. Early days.

What has been learnt so far ?

Part 2) Data confidentiality – Role for **Distributed Machine Learning.**

Part 3) What on earth are the algorithms doing – can you explain please !

Not how they work but why they work and how to explain this to a non-expert – the patient ! **What is the black box doing ?**

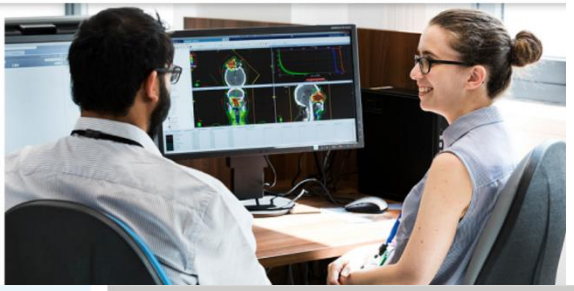
Part 4) Thoughts on how to solve 3) !

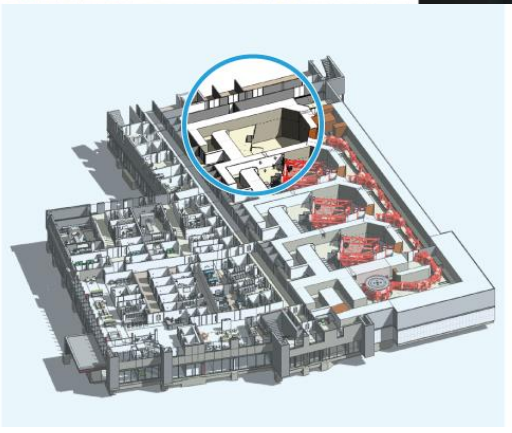
Finally - **Conclusions**

PART 1 - Proton Beam Therapy - The Christie Trust - Manchester

NHS The Christie NHS Foundation Trust

Home | Patients and Visitors | Professionals | Services | About Us | The Christie Charity





£125 m Investment
Open Dec. 2018
Treated its 1st patients.
750 per year planned

MANCHESTER 1824 The University of Manchester

Faculty of Biology, Medicine and Health

Search

Study | Research | Connect | About

Accelerator Physics, Detectors, Dosimetry, Modelling, GEANT4, Data Analysis, Machine Learning etc.

PBT research is undertaken by the PRECISE group (Proton research at The Christie and the University's Division of Cancer Sciences). The group has a dedicated research facility – known as the 'research room' within the clinical proton therapy centre aimed at addressing these challenges.
Led by Prof. Karen Kirkby

PART 1 - Global Challenge Advanced Radiotherapy Network+ STFC-UKRI

Led by Prof. Karen Kirkby – University of Manchester

Extracting meaning from Big Data – Sandpit Event – 19/20 Feb. 2018 Manchester

Big data is all around us and in the radiotherapy arena there is a lot of data (images, scans, outcomes, PROMs to name but a few). The STFC community really knows how to get the most out of big data, look at CERN and the Large hadron Collider, and the vast amount of data that has to be analysed before major discoveries such as the Higgs Boson can be made. Look at the astrophysics community that uses big data to probe distant galaxies, find new stars and "see" black holes.

Can we bring the STFC community, which works, on very small particles and very large suns to apply their expertise to develop innovative methods for analyzing the very large data sets that we have in radiotherapy?

This is what this proposed sandpit is all about:

Can we extract data biomarkers from radiotherapy data to act as early signals of the onset infiltration or reoccurrence of disease?
Can extract information on normal tissue toxicity study its impact on quality of life both in the short term and longer term?
Can we combine this information with genomics and proteomic data to select the patients who will benefit most from treatment?
Can we personalize treatments to improve both outcomes and quality of life?
Can we use the data we already have to improve the treatment of future patients?

This Sandpit aims to bring people from the STFC and clinical communities together to discuss how they might work together and develop innovative research ideas in the field of Big Data.

Pump-priming funding (typically £1500 to £15,000) is available to initiate the best projects resulting from collaborations developed during the Sandpit.

PART 1 Three of the projects funded involving particle physicists, astrophysicists, medical physicists, biologists, medical doctors…….

1) Image Registration

Coordinator: Yvonne Peters (PP)

Yvonne.Peters@manchester.ac.uk

Sarah Bannister, Denis Page

Convolutional Neural Network (CNN) shown to work well.

2.1 Image registration

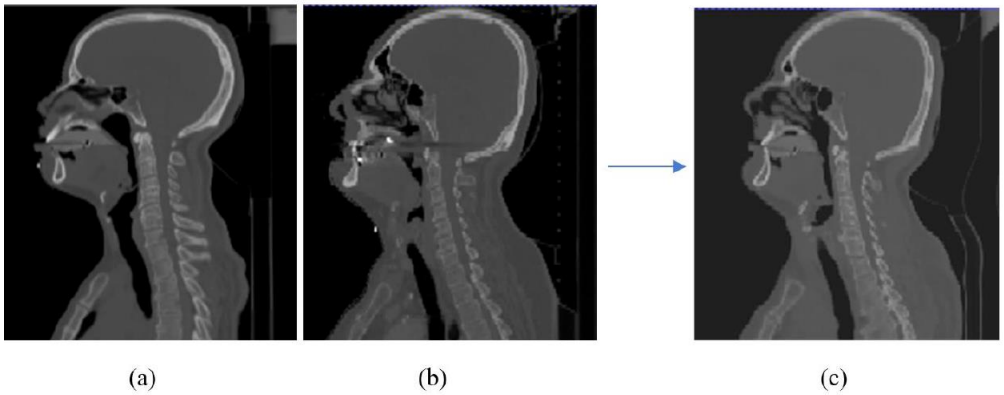


Figure 1. An example of inter-patient image registration. (a) is a patient scan used as a floating image, (b) is the reference image and (c) is the result of the deformable registration of the two.

2) RadioTherapy Machine Learning (RTML) Network

Coordinators

Dr Robert Lyon[1] (machine learning & astronomy)

Dr Tim Rattay[2] (domain expert, clinician)

contact@radiotherapymlnetwork.co.uk

Steering Group

Dr Andrew Green[1,3] (pipelines, data specialist)

Prof Nigel Mason[4] (data analysis, astronomical expertise)

Dr Sarah Osman[5] (domain expert, data analysis)

[1] University of Manchester, [2] University of Leicester, [3] The Christie NHS Foundation Trust, [4] The University of Kent, [5] Queens University Belfast,

2) Continued…….RTML Network

Aims

- To foster collaboration between the machine learning community within STFC and radiotherapy practitioners.
- Explore existing clinical data and address specific clinical questions using machine learning.
- Develop a network bringing together radiotherapy and STFC machine learning communities.
- Learn from existing data to improve patient outcomes and quality of life.

Some problems they are looking at

Group no	Group lead	Working title
~~1~~	~~Matt~~	~~PTV-GTV inhomogeneity and tumour recurrence~~
~~2~~	~~Penny~~	~~RT planning quality, tumour genetics to classify patients according to risk of recurrence~~
3	Emma / Joe	Can we predict treatment plan quality evaluation metrics from the structure outlines?
4	Alan	Predicting co-morbidities from imaging & data
5	Cheng/ Ian	Using AI/ML to outline plans
~~6~~	~~John & Gareth~~	~~Contour recognition of structures across different RT plans / distributed learning~~
7	Mahmoud	Predicting toxicity – genetic [and imaging] data
8	Mayur	Prediction of NSCLC using deep learning approaches based on NLST dataset

Therapy planning, morbidity, toxicity

GitHUB resource. Two workshops to date.

https://github.com/scienceguyrob/RTMLResources

3) Mining Science Data for Medicine - MisDAM

1st Mining Science Data for Medicine Workshop - 16 April 2019

This project resulted from a Sandpit Event organized by the STFC Global Challenges Network+ in Advanced Radiography [1]. The MisDaM project has three aims,

i) To obtain interesting medical science results with potential to apply to individual patients.
ii) Create a community of data miners to support the analysis of big data associated with medical science.
iii) Identify the algorithms and visualisations that are useful in this science area.

To achieve these aims, the project team will release challenges to the global community and invite anyone to solve a specific medical science problem using data mining and machine learning.

This handbook provides details on how anyone can get involved in the first challenge, called MiSDaM01.

http://www.hep.manchester.ac.uk/MiSDaM/

https://indico.hep.manchester.ac.uk/conferenceDisplay.py?ovw=True&confId=5452

Miriam Berry[1], Alfred Oliver, Ken Raj[2], Marina Romanchikova[3], Stephen Watts[4]

[1]National Physical Laboratory, University of Cambridge, [2] Public Health England, [3]National Physical Laboratory, Teddington, [4]School of Physics and Astronomy, The University of Manchester

MiSDAM01 - THE TWO CHALLENGES

Identification of DNA methylation–based markers of cellular senescence.

Ken Raj (Health England) gave a talk explaining the science and the data. Full and facinating talk at website

Big challenge – 850,000 variables with 48 samples.

"DNA methylation-based biomarkers and the epigenetic clock theory of ageing", Steve Horvath and Kenneth Raj. Nat Rev Genet. 2018 Jun;19(6):371-384. doi: 10.1038/s41576-018-0004-3

Explaining Machine Learning (ML) results to patients and doctors

David Watson Oxford Internet Institute
Full talk at website

See also article in British Medical Journal

Clinical applications of machine learning algorithms: beyond the black box

To maximise the clinical benefits of machine learning algorithms, we need to rethink our approach to explanation, argue **David Watson and colleagues**

David S Watson *doctoral student*[1 2 3], Jenny Krutzinna *postdoctoral researcher*[1], Ian N Bruce *professor of rheumatology and director*[4 5], Christopher EM Griffiths *foundation professor of dermatology*[5 6], Iain B McInnes *Muirhead professor of medicine*[7], Michael R Barnes *reader of bioinformatics*,[2 3], Luciano Floridi *professor of philosophy and ethics of information and director of the digital ethics lab*[1 3]

BMJ 2019;364:l886 doi: 10.1136/bmj.l886
(Published 12 March 2019)

Trying the Data Challenge approach – following on from the GREAT08 Challenge
Used by astrophysicists to get improved algorithms – won by a computer science group

A1.1 Challenge Rules

We are grateful to the GREAT08 Challenge Handbook,[2], from which these rules derive.

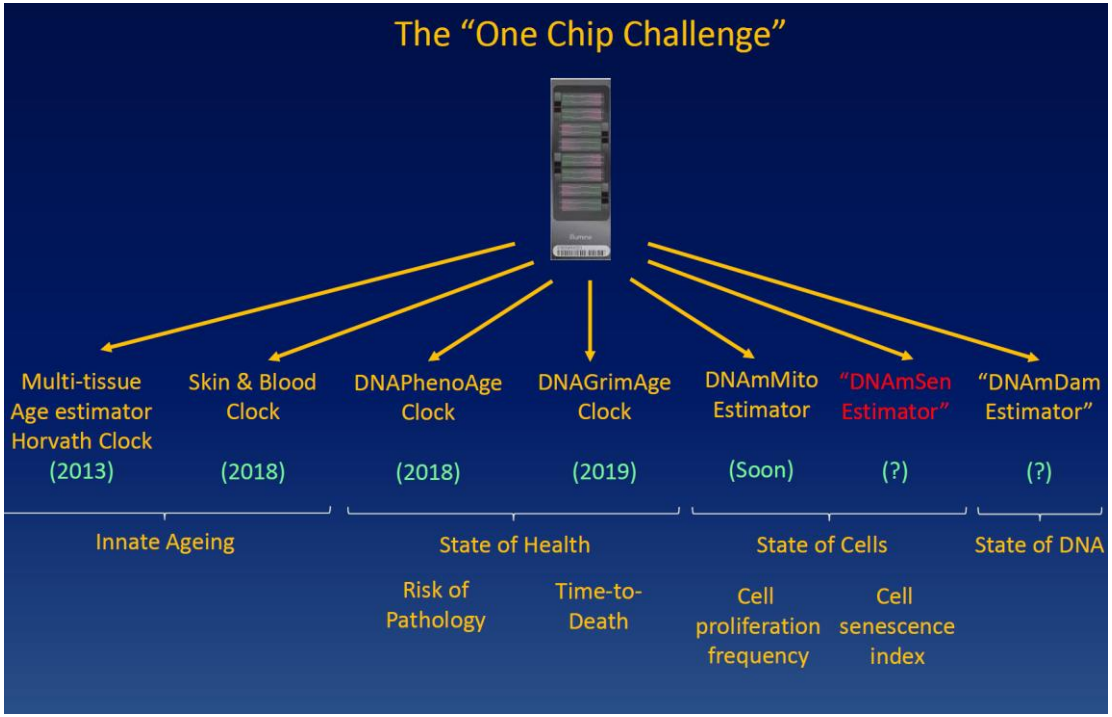
1) The data will be released publically at the end of the challenge.
2) The challenge end date will be shortly before the challenge " Results Workshop" in September 2019. To be specific, the end date is, 26 August 2019.
3) All registered teams are expected to present their results at the "Results Workshop" in September 2019. The date for the results workshop will be agreed at the kick-off workshop on 16 April 2019.
4) Teams signing up to the project will have access to the "DNA methylation–based marker of cellular senescence" data under condition that they abide by these rules.
5) Teams will be encouraged to publish if they wish, but commit not to do so until after the final challenge report has been issued in pre-print form to the arXiv or January 2020, whichever date is earliest.
6) Publications from individual teams should acknowledge the source of the data and the MiSDaM01 challenge. The fact that the data was provided as the result of UKRI/STFC funding should be acknowledged.
7) Participants may use a pseudonym or team name on the participants list, however real names (as used in publications) must be provided when requested during the result submission process.
8) Participants must provide a report detailing the results and methods used, at the challenge deadline. We would prefer that any code developed for this challenge is made public. We suggest the use of GitHuB.
9) We expect all participants to allow their results to be included in the final Challenge Report. We will however be flexible in cases where methods performed badly compared to other methods or if participants are strongly against publicising them.
10) Clarification concerning the data will be provided (if possible) when requested and will be made available to all participants.

Some additional competition rules apply to members of the Local Project Team who submit entries:

A) For the purpose of these rules, "Local Project Team" includes the authors of this document and staff or students associated with them.

B) Only information available to non-team participants may be used in carrying out the analysis.

See the handbook at the project website for more details



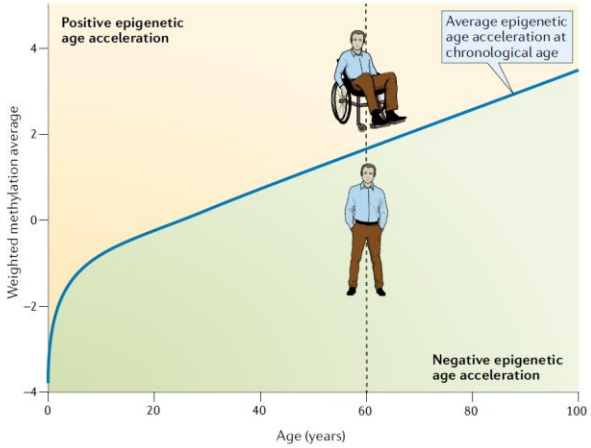


Fig. 2 | **Multi-tissue DNA methylation-based age and age acceleration.** The solid blue line shows how an uncalibrated version of the multi-tissue DNA methylation-based (DNAm) age estimate, weighted average of 353 CpGs, changes with age[8]. The rate is very

"DNA methylation-based biomarkers and the epigenetic clock theory of ageing", Steve Horvath and Kenneth Raj. Nat Rev Genet. 2018 Jun;19(6):371-384. doi: 10.1038/s41576-018-0004-3

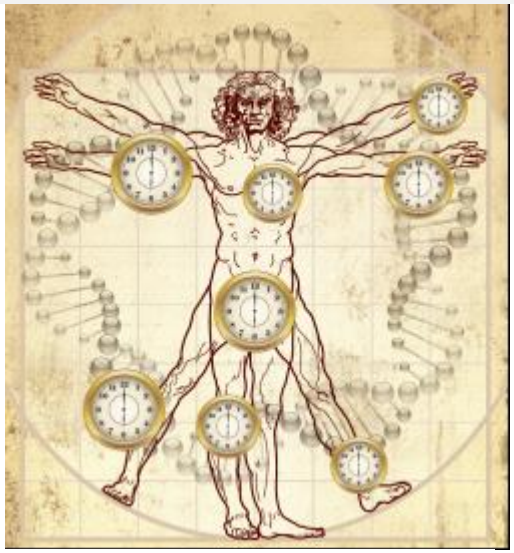
Our age is built into our DNA - various epigenetic clocks.

For the senescence clock, typically 850,000 DNAm markers.

24 Control + 24 Irradiated samples

Big Data may not be " deep " but may have large no. of variables **(Big Variates).**

Horvath Clock has ~ 350 markers linked to age out of 850K ! A very challenging problem……



Picture from Steve Horvath

PART 2 - Medical Data is sensitive ! There are clear rules in the UK for the medical area

https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology#Principle-1/

21 pages. Ten Key Principles

1. **Understand users, their needs and the context**
2. **Define the outcome and how the technology will contribute to it**
3. **Use data that is in line with appropriate guidelines for the purpose for which it is being used**
4. **Be fair, transparent and accountable about what data is being used**
5. **Make use of open standards**
6. **Be transparent about the limitations of the data used and algorithms deployed**
7. **Show what type of algorithm is being developed or deployed, the ethical examination of how the data is used, how its performance will be validated and how it will be integrated into health and care provision**
8. **Generate evidence of effectiveness for the intended use and value for money**
9. **Make security integral to the design**
10. **Define the commercial strategy**

Note: The "Black Box " also appears in this – more later

Data Confidentiality - Using Distributed (Machine) Learning

The Christie
NHS Foundation Trust

Research Using Data Mining

Clinical lead: Professor Corinne Faivre-Finn – Clinical oversight, trial methodology, outcomes collection, patient involvement.

Scientific lead: Professor Marcel van Herk – Scientific oversight, mathematical modelling techniques, model validation, advanced image processing.

Technical lead: Dr Gareth Price – System development, database management and administration, scientific computing and algorithm development, software processes and quality systems.

Scientific systems: Ian Porter – System and network management and security.

Modelling: Corinne Johnson – Lung and brain disease outcome modelling, use of dynamic Image Guided Radiotherapy data to assess impact of patient change.

Project management: Sally Falk - Project management and administration

Introduction

'Data mining' projects aim to learn from every patient treated at The Christie with the ultimate aim of being able to better target (or personalize) treatments to the individual disease and characteristics of newly presenting patients. The first steps towards this goal are to begin analyzing the large amounts of data about diagnosis, treatment, radiotherapy planning and outcomes which are already available at The Christie. Using mathematical models and new developments in computational technologies we can extract ANONYMOUS information from existing Christie databases and clinical systems and use this to understand where patient's treatments have been successful and where there have been unwanted side effects.

Team Objectives

We aim to use the power of 'big data' in the future to:

- Develop systems to support decision making in clinical practice via the Computer Aided Theragnostics project (ukCAT).
- Link outcomes (survival benefit and treatment toxicity) to individual patient characteristics, imaging results, blood & tissue biomarkers and treatment.
- Learn from previously treated patients and provide unbiased, individual predictions of potential help and harm of a particular therapy for each newly presenting patient.
- Enable this information, in the future, to be used together by patients and their doctors to reach shared, informed decisions about the most appropriate treatment.
- Extend 'big data' analytics to the complex, multi-dimensional clinical and imaging data collected every day during routine cancer treatment in the NHS.

For more information on Data Mining and the ukCAT project email the team's single point of contact:

UKCAT@christie.nhs.uk

https://www.christie.nhs.uk/professionals/research/research-themes/research-using-data-mining/distributed-learning/

The ukCAT Project

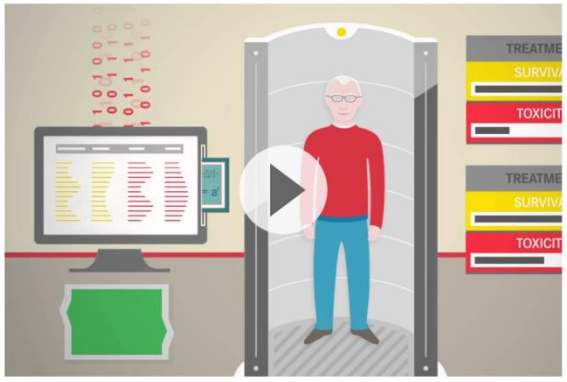
The Computer Aided Theragnostics (ukCAT) project aims to develop predictive models using routine clinical data which will, in the future, underpin decision support systems for use in clinics. The use of routine data from all patients, as opposed to selected groups of patients (for example, those that take part in clinical trials) ensures that the results of our analyses apply to everyone, including those patients who rarely take part in clinical trials (e.g. the very elderly and those with other serious medical conditions). We aim to extend 'big data' analysis to the complex clinical and imaging data collected every day during routine radiotherapy treatment in the NHS.

The Christie is the first UK partner in the CAT rapid learning oncology network (see www.eurocat.info for details of other international CAT partners) which aims to introduce decision support into clinical practice in the future. The ukCAT project uses a system called 'distributed learning' to analyse data.

Information learned from anonymous databases located at hospitals or other medical institutions are securely combined to allow knowledge to be shared in a secure and private environment. Safely pooling information allows models to benefit from using larger numbers of patients (thereby capturing rarer events) and to compare different methods of treatment around the world.

All data held in the ukCAT database is fully anonymised and held on secure servers in accordance with the law, and NHS Information Governance and Research Governance policies. The process of anonymising the data is performed before the data is sent to the ukCAT system meaning it is not possible for a research team member using the system to access any identifiable information.

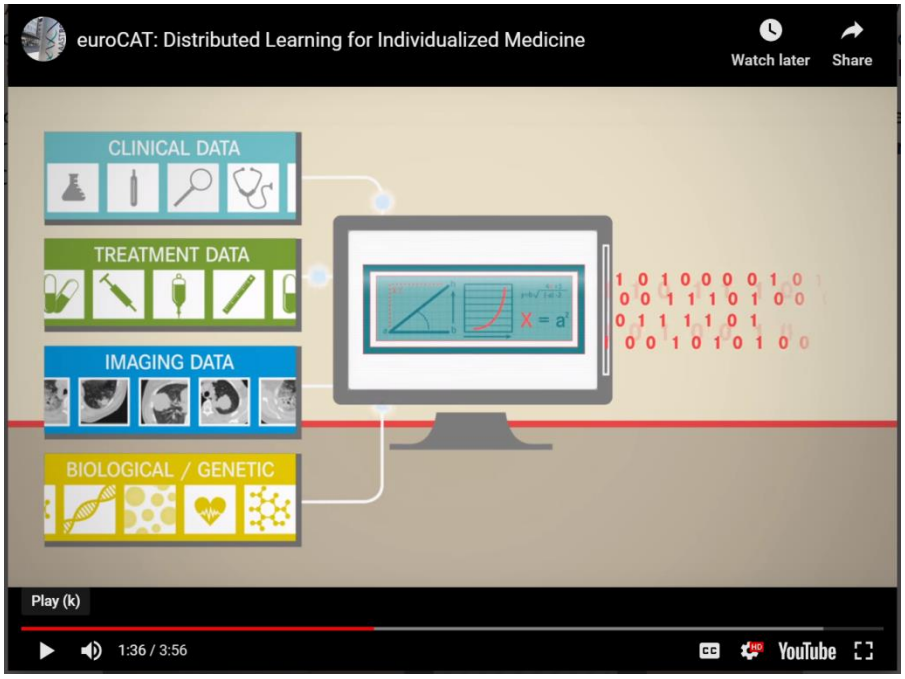
The central CAT data security policy is that all data on any CAT server is anonymous, and that during distributed learning no data ever leaves a hospital or institution's own servers - **Models go to the data, the data never leaves the institution.** The following video developed by our euroCAT partners explains how this works.

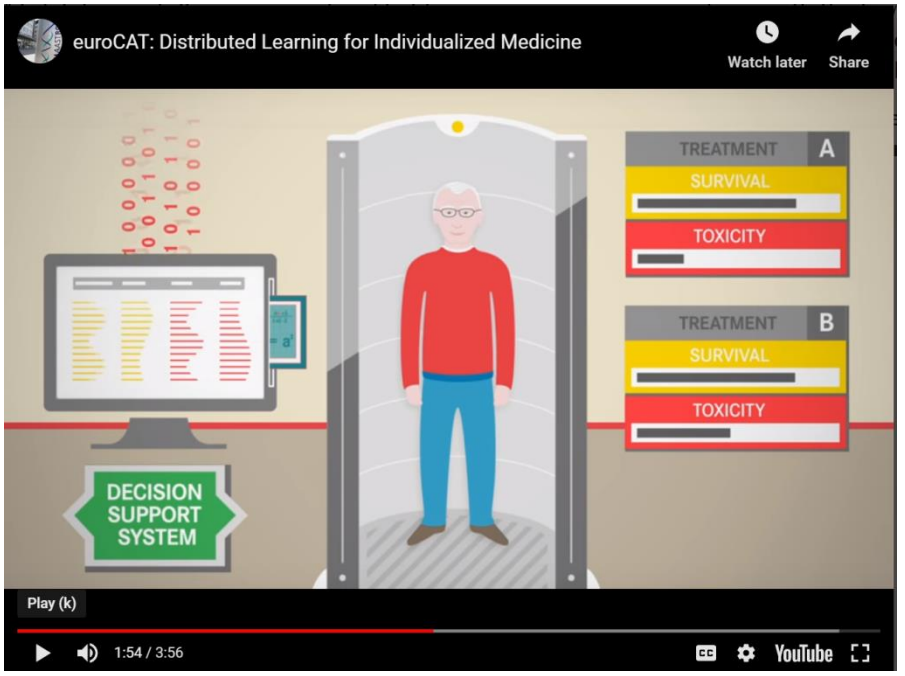


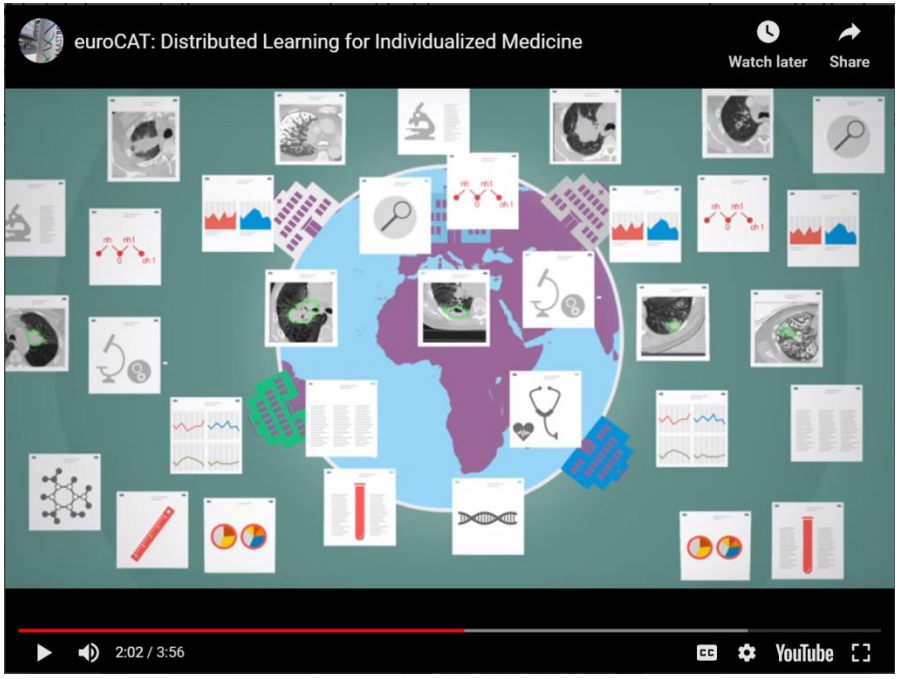
euroCAT: Distributed Learning for Individualized Medicine video

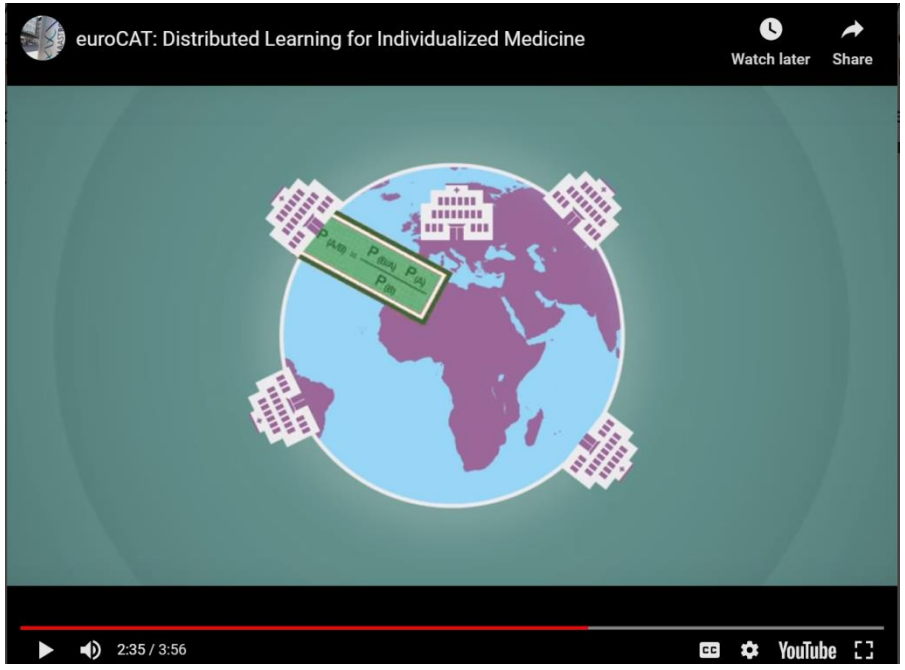
https://youtu.be/ZDJFOxpwqEA

Also go to www.eurocat.info


euroCAT: Distributed Learning for Individualized Medicine
Watch later
Share
CLINICAL DATA
TREATMENT DATA
IMAGING DATA
BIOLOGICAL / GENETIC
Play (k)
1:36 / 3:56
YouTube


euroCAT: Distributed Learning for Individualized Medicine
Watch later
Share
TREATMENT A
SURVIVAL
TOXICITY
TREATMENT B
SURVIVAL
TOXICITY
DECISION SUPPORT SYSTEM
Play (k)
1:54 / 3:56
YouTube


euroCAT: Distributed Learning for Individualized Medicine
Watch later
Share
2:02 / 3:56
YouTube


euroCAT: Distributed Learning for Individualized Medicine
Watch later
Share
2:35 / 3:56
YouTube

PART 2 - Lessons from ukCAT

Clinical Oncology 29 (2017) 814–817

Contents lists available at ScienceDirect

Clinical Oncology

journal homepage: www.clinicaloncologyonline.net

Editorial

Data Mining in Oncology: The ukCAT Project and the Practicalities of Working with Routine Patient Data

G. Price, M. van Herk, C. Faivre-Finn

The University of Manchester, Manchester Academic Health Science Centre, The Christie NHS Foundation Trust, Manchester, UK

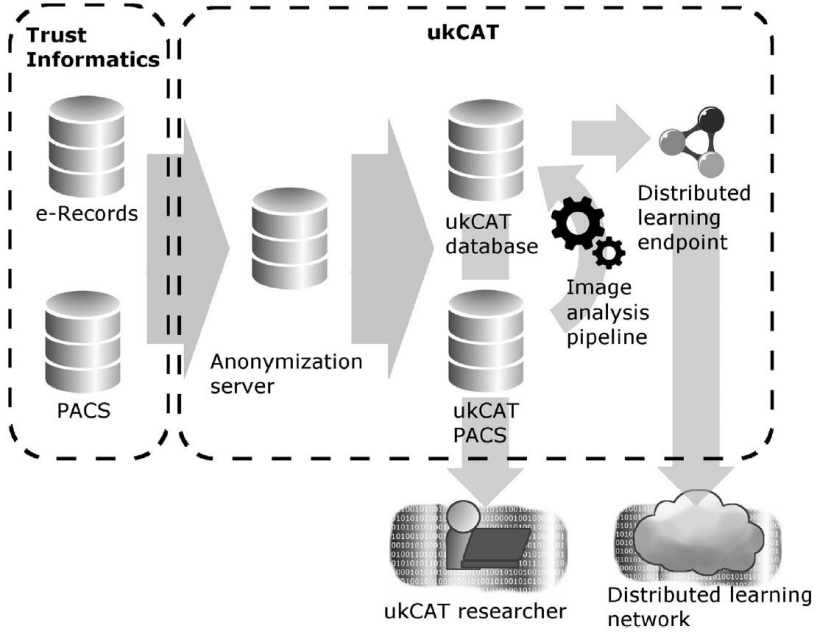


Fig 1. High-level schematic diagram of the ukCAT system showing data flow from Trust informatics via an anonymisation PACS and database. The data can then be learned upon locally (e.g. [14]) or published to the distributed learning end collaborative model training (e.g. [12]).

Machine learning in oncology, although quantitatively consolidating the existing knowledge base, and beginning to provide fantastic clinical insights, is still very much in its infancy. Over the next few years, once ukCAT and similar approaches become embedded into practice, we can expect that increasing numbers of data sources e from diagnostic, follow-up and image-guidance imaging, to biomarker measurements and genomic profiles will be added to the system. When combined with the possibilities of distributed learning, the information available to learn on will indeed be within the realms of the definition of 'big data' [15] with the associated potential to be clinically transformative.

The Achilles heel of the whole endeavour, however, is that of outcomes data. No matter how well we record presentation and treatment data, or how cleverly we interrogate patient images, without high-quality clinical outcomes - survival, local and distant control, and both acute and late treatment toxicities - it is quite simply impossible to model their occurrence. Unless we are able predict both the risk and benefit of treatment strategies, the dream of individualised treatment and comprehensive decision support systems will remain just that.

Technology can help- our team has recently had success using natural language processing to recover toxicity data from electronic patient records [16].

However, this is no substitute for the accurate, systematic, prospective recording of outcome information. The infrastructure and validated questionnaires needed to gather patient-reported outcomes are available, and have even been shown to deliver improved quality of life [17] and survival [18] in cancer patients. With the collective will to enhance such data collection efforts, we will have routine data covering the whole care pathway and be able to truly start learning from every patient treated.

PART 3 Algorithms (AI) , Ethical and Legal Implications, GDPR

Haphazard embrace of AI puts justice at risk

Tuesday June 4th 2019

4 June 2019

Print this page

The ad hoc use of complex algorithms in the justice system needs urgent oversight, the Law Society of England and Wales said as it released the results of a year-long investigation.

The Law Society Technology and Law Policy Commission publishes its report on algorithms in criminal justice alongside an interactive map that allows the public to see for the first time the beginnings of an overview of where algorithms are being used to assist decision-making across the justice system across England and Wales.

"Police, prisons and border forces are innovating in silos to help them manage and use the vast quantities of data they hold about people, places and events," said Law Society president Christina Blacklaws.

"Complex algorithms are crunching data to help officials make judgement calls about all sorts of things – from where to send a bobby on the beat to who is at risk of being a victim or perpetrator of domestic violence; who to pick out of a crowd, let out on parole or which visa application to scrutinise.

"While there are obvious efficiency wins, there is a worrying lack of oversight or framework to mitigate some hefty risks – of unlawful deployment, of discrimination or bias that may be unwittingly built in by an operator.

"These dangers are exacerbated by the absence of transparency, centralised coordination or systematic knowledge-sharing between public bodies. Although some forces are open about their use of algorithms, this is by no means uniform."

The Law Society

https://www.lawsociety.org.uk/

The Law Society's key recommendations:

- **Oversight**: A legal framework for the use of complex algorithms in the justice system. The lawful basis for the use of any algorithmic systems must be clear and explicitly declared
- **Transparency**: A national register of algorithmic systems used by public bodies
- **Equality**: The public sector equality duty is applied to the use of algorithms in the justice system
- **Human rights**: Public bodies must be able to explain what human rights are affected by any complex algorithm they use
- **Human judgement**: There must always be human management of complex algorithmic systems
- **Accountability**: Public bodies must be able to explain how specific algorithms reach specific decisions
- **Ownership**: Public bodies should own software rather than renting it from tech companies and should manage all political design decisions

Christina Blacklaws added: "Within the right framework algorithmic systems – whether facial recognition technology, predictive policing or individual risk assessment tools – can deliver a range of benefits in the justice system, from efficiency and efficacy to accountability and consistency.

thebmjopinion Latest Authors

Can AI fulfil its medical promise ?
Selected text below. April 26 2019

Can AI fulfil its medical promise?

April 26, 2019

AI technology has challenges to overcome, but it can be a force for good in medicine, say Luxia Zhang, Guilan Kong, Liwei Wang, and Qi-Min Zhan

Note: Medical profession decades ahead of the lawyers

- The first medical decision support system based on artificial intelligence (AI) was developed in the early 1970s.

- Studies have shown that the diagnostic accuracy of AI algorithms is comparable to experienced medical experts for diabetic retinopathy, heart disease, and certain cancers. An impressive example is detecting pulmonary nodules on lungs in computer tomography scans. It usually takes physicians several minutes to do this, while AI based systems only need a few seconds.

- Supporting decision making - Simple diagnosis and supporting doctors for complex clinical cases. " Physicians assistant"

- All those scenarios show the benefits of utilising AI in medicine. But one crucial question cannot and should not be ignored: "Is the information generated by an AI system trustworthy?" If we need to rely on an AI system to assist our decision making, we should care about its reliability and effectiveness.

- **Currently, AI algorithms based on deep learning act like a "black box": the inner logic behind most machine learning models is hard to explain, and the doctors using them are not given explanations for the advice they receive from these systems. This is not an intuitive way for doctors to practise and raises uncertainty about using AI, since the principle of identifying causality and treating causes is integral to medicine.**

- AI technology has challenges to overcome, but it can be a force for good in medicine. If we want to maximise the benefits of AI for the sake of patients and the public, then medical doctors, researchers, and AI scientists should work closely together.

Algorithms (AI) , Ethical and Legal Implications, GDPR

Clinical applications of machine learning algorithms: beyond the black box

To maximise the clinical benefits of machine learning algorithms, we need to rethink our approach to explanation, argue **David Watson and colleagues**

Unfortunately, many popular machine learning algorithms are essentially black boxes—oracular inference engines that render verdicts without any accompanying justification. This problem has become especially pressing with passage of the European Union's latest General Data Protection Regulation (GDPR), which some scholars argue provides citizens with a "right to explanation."

Now, any institution engaged in algorithmic decision making is legally required to justify those decisions to any person whose data they hold on request, a challenge that most are ill equipped to meet. We urge clinicians to link with patients, data scientists, and policy makers to ensure the successful clinical implementation of machine learning (fig 1).

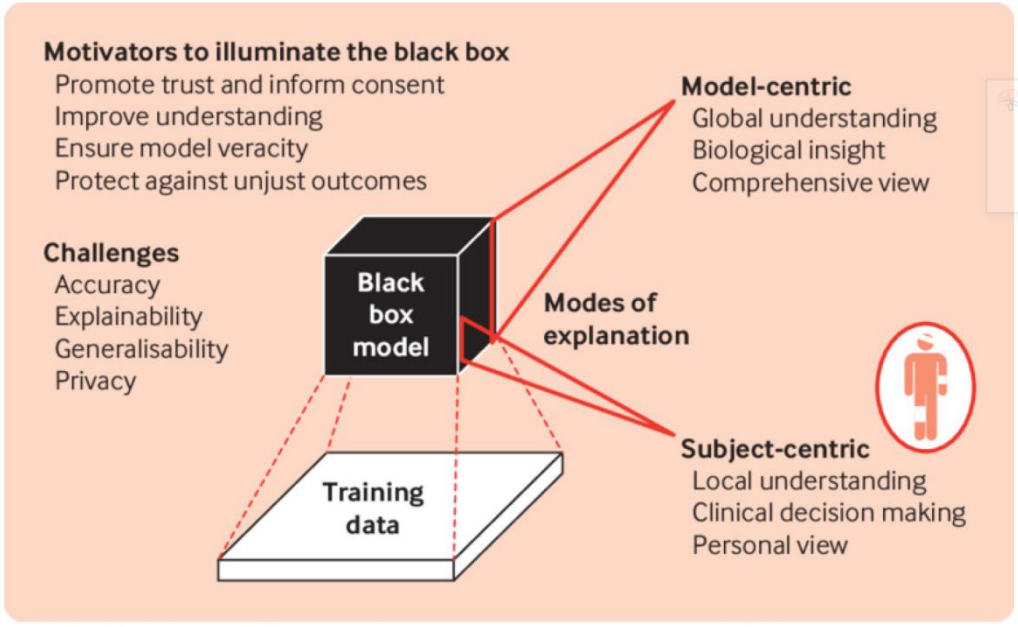


Fig 1 Overview of the opportunities and challenges associated with black box models in clinical decision making

Predictions versus explanations

Predictions tell us that x is true; explanations tell us why x is true. The past decade has seen enormous advances in our ability to predict complex phenomena using computational techniques. Explanatory breakthroughs, on the other hand, have been few and far between.

Key messages

- Machine learning algorithms may radically improve our ability to diagnose and treat disease
- For moral, legal, and scientific reasons, it is essential that doctors and patients be able to understand and explain the predictions of these models
- Scalable, customisable, and ethical solutions can be achieved by working together with relevant stakeholders, including patients, data scientists, and policy makers

Part 4 – Thoughts on explaining the Black Box

1) Use Visualization. e.g. CNN often have good explanations for what is a complicated box.
2) Unlike particle physics or cosmology, there is no " Standard Model " for most data.

However, key variables often known or good circumstantial evidence.

Q. How to analyse in a model independent way ?

A. Look at how the variables related to one another and the class variable.
(" Attribute analysis " – which is largely ignored in Machine Learning)

Use some statistics from information theory. Model independent.

a) Similarity Index (SI) – % of shared information between variables.
b) Class Distance Indicator (CDR) - estimate of the Kullback- Leibler Distance (KL)) between two classes of events. Stein's Lemma puts limit on how well one can separate two classes

Prob. False Alarm -> $2^{-\text{CDR}}$

Cohen's Kappa which is estimated from the confusion matrix - quantifies the ML performance.

CDR and Cohen's Kappa are related.

$$\kappa = 1 - 2^{-CDR}$$

1st Example - Wisconsin Breast Cancer

Relevant Information about the dataset:

W.H. Wolberg, W.N. Street, and O.L. Mangasarian.
Machine learning techniques to diagnose breast cancer from fine-needle aspirates.
Cancer Letters 77 (1994) 163-171.

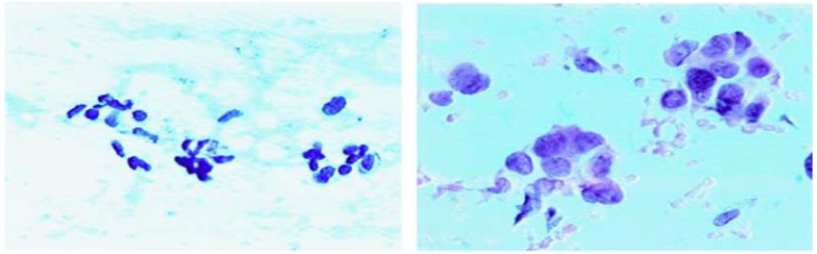


Fig. 1 Images taken using the FNA test: (a) Benign, (b) Malignant

What are the variables ?

a)radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)
Mean, Standard Error, Worst - so 3x10 = 30 variables

Number of events: class 0 = 357 ; class 1 = 212
Total = 569 0 = Benign, 1 = Malignant

Number of Attributes (Variates) = 30 + Class Variable

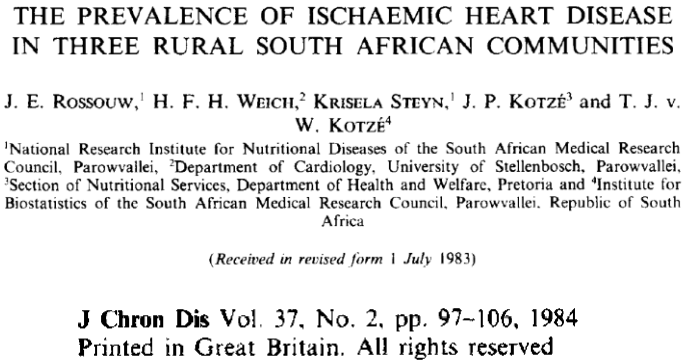
Second Example - CORIS Dataset

462 South African males evaluated for heart disease.

Outcome variable: Coronary heart disease (`chd`).

Covariates:

- Systolic blood pressure (`sbp`)
- Cumulative tobacco use (`tobacco`)
- LDL cholesterol (`ldl`)
- Adiposity (`adiposity`)
- Family history of heart disease (`famhist`)
- Type A behavior (`typea`)
- Obesity (`obesity`)
- Current alcohol consumption (`alcohol`)
- Age (age)



THE PREVALENCE OF ISCHAEMIC HEART DISEASE IN THREE RURAL SOUTH AFRICAN COMMUNITIES

J. E. Rossouw,[1] H. F. H. Weich,[2] Krisela Steyn,[1] J. P. Kotzé[3] and T. J. v. W. Kotzé[4]

[1]National Research Institute for Nutritional Diseases of the South African Medical Research Council, Parowvallei, [2]Department of Cardiology, University of Stellenbosch, Parowvallei, [3]Section of Nutritional Services, Department of Health and Welfare, Pretoria and [4]Institute for Biostatistics of the South African Medical Research Council, Parowvallei, Republic of South Africa

(*Received in revised form* 1 *July* 1983)

J Chron Dis Vol. 37, No. 2, pp. 97–106, 1984

Class 0 " Healthy", Class 1 " Coronary Heart Disease"

302 Healthy, 160 Diagnosed with CHD

Lets feed this data into a Machine Learning Package (WEKA)

Briefly - Use WEKA SVM (Breast Cancer) and Decision Tree (J48) CORIS – about the best one can do… Put in all the variables at this stage..

Variables	Success Rate	Confusion Matrix			
		S Found	B Found		
Breast Cancer	97.7 % correct	201	11	**S**	
	2.3 % wrong	2	355	**B**	**Input**
Cohen's Kappa 0.9507 – more about this later.					
CORIS	70.8 % correct	248	54		
	29.2 % wrong	81	79		
Cohen's Kappa 0.328					
Warning – Just above the random rate (65%) !!					

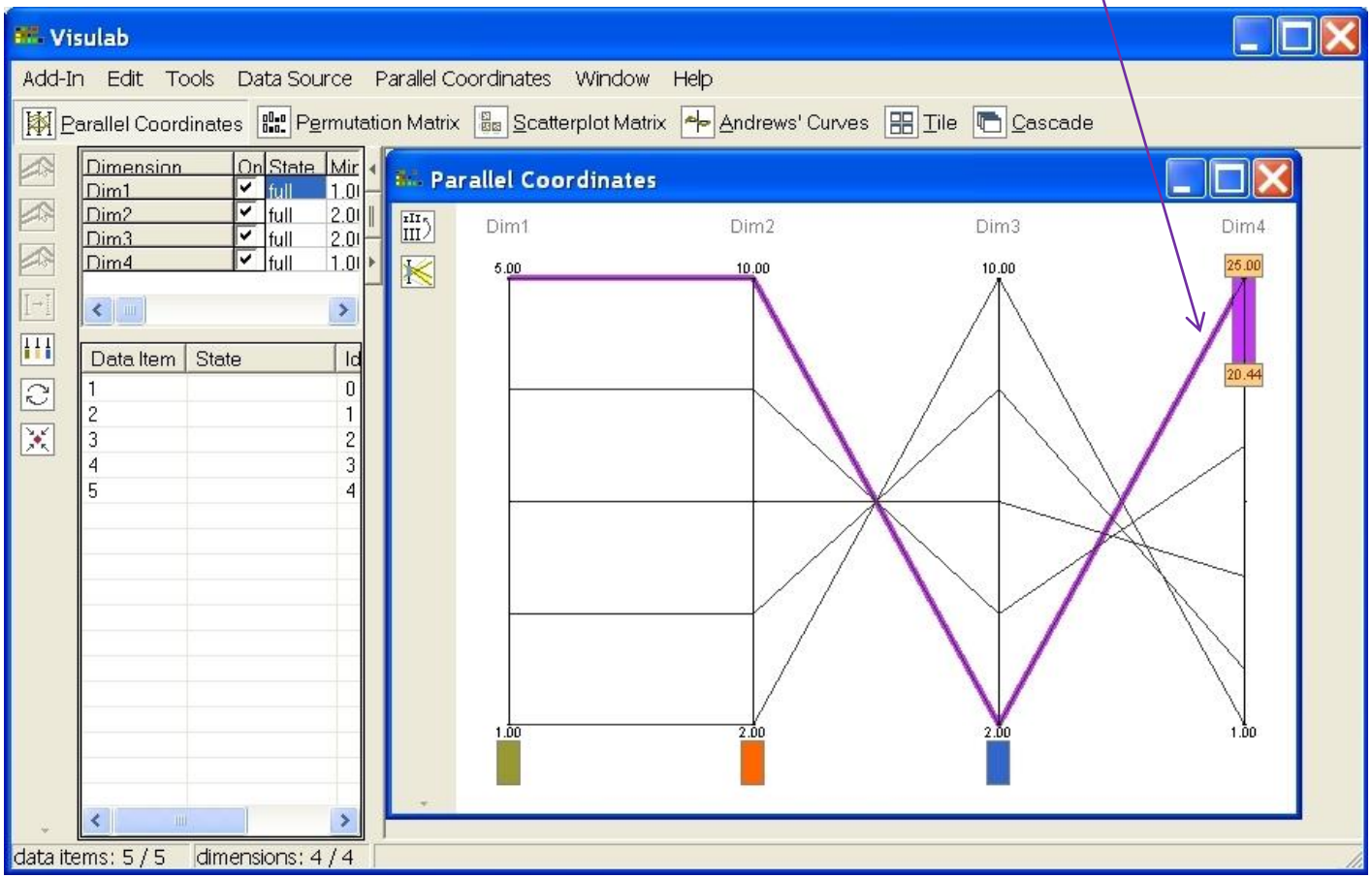
Is this the best one can do ? What variables matter ?
Why is the CORIS dataset worse than the Breast Cancer dataset ?

A multivariable visualisation - Parallel Coordinates

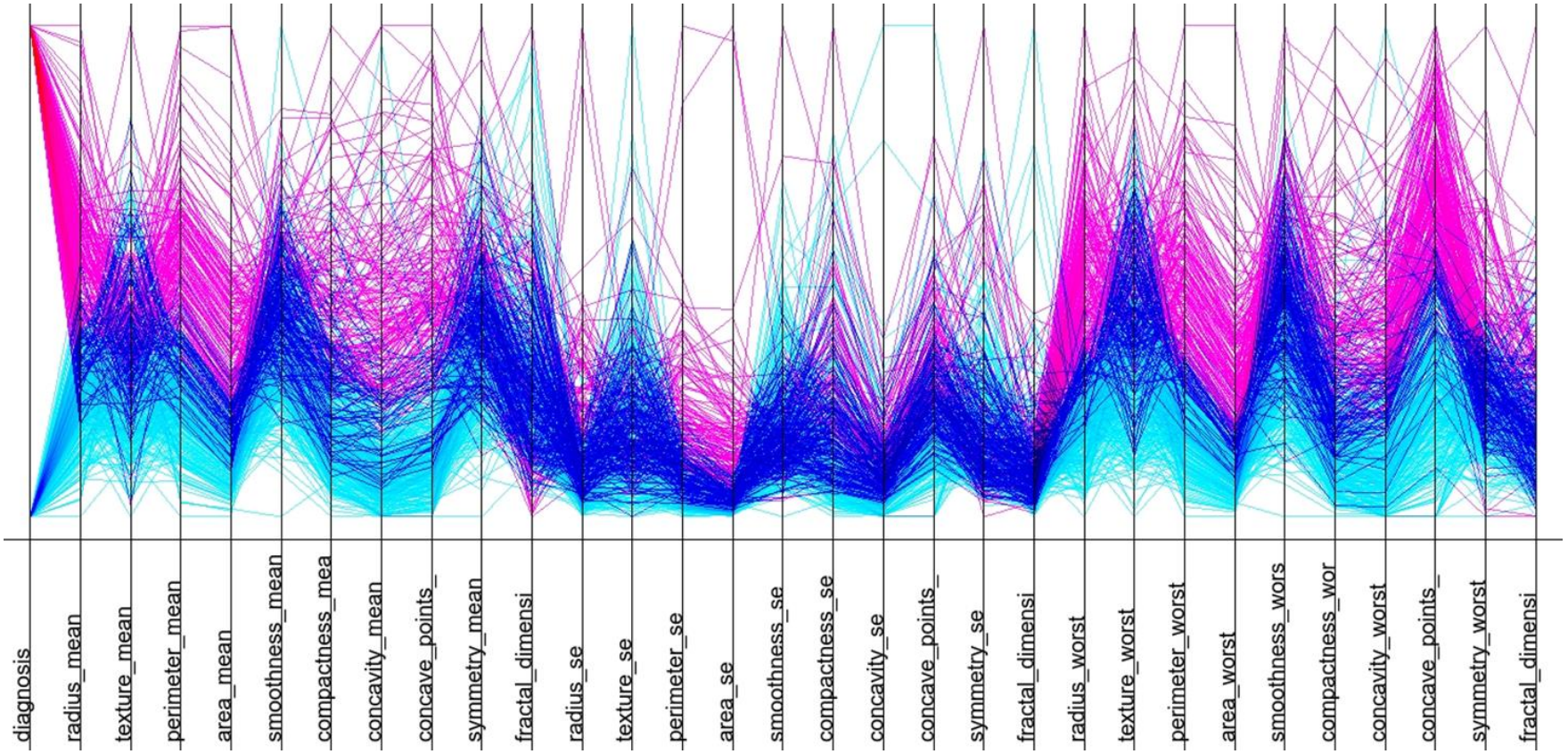
DataPoint	Dim1	Dim2	Dim3	Dim4
1	1	2	10	1
2	2	4	8	4
3	3	6	6	9
4	4	8	4	16
5	5	10	2	25

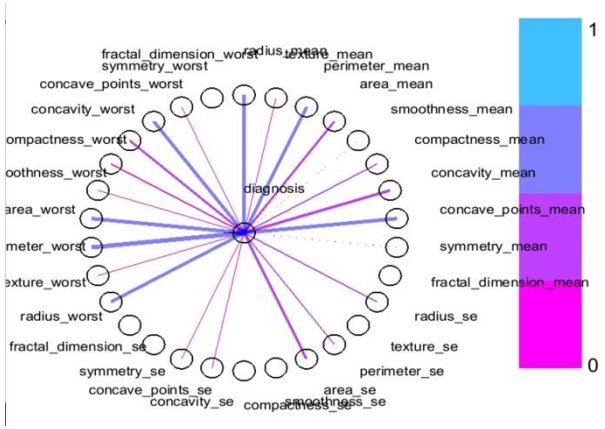
Simple Implementation with EXCEL plugin http://www.inf.ethz.ch/personal/hinterbe/Visulab/

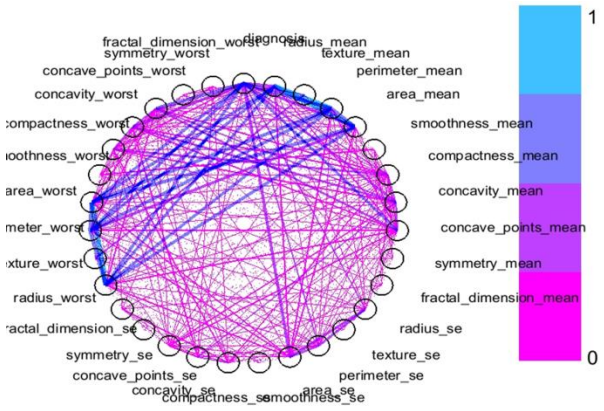
This also shows the idea of brushing!!!



Visualization→ parallel coordinates – breast cancer dataset

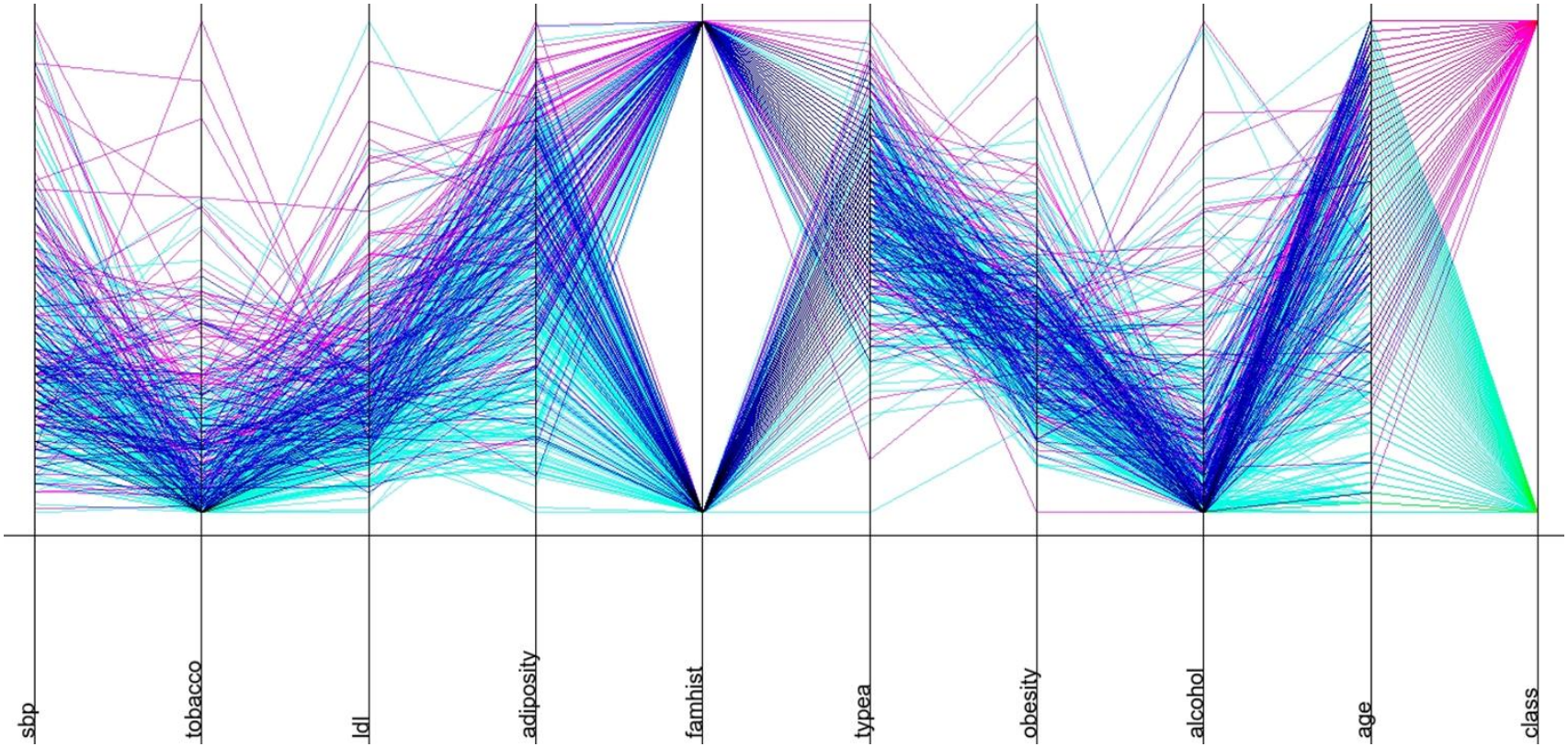


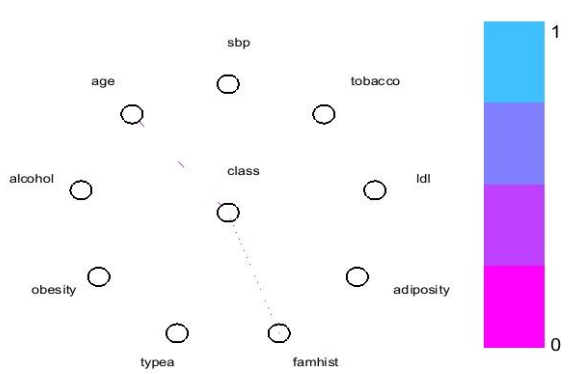


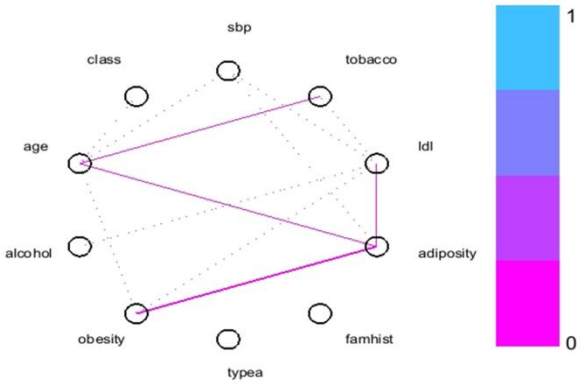


Variable Interaction Diagram

Discussion→ parallel coordinates – CORIS dataset







Variable Interaction Diagrams

Cohen's Kappa (ML) versus CDR (underlying PDF)

Kappa Scale

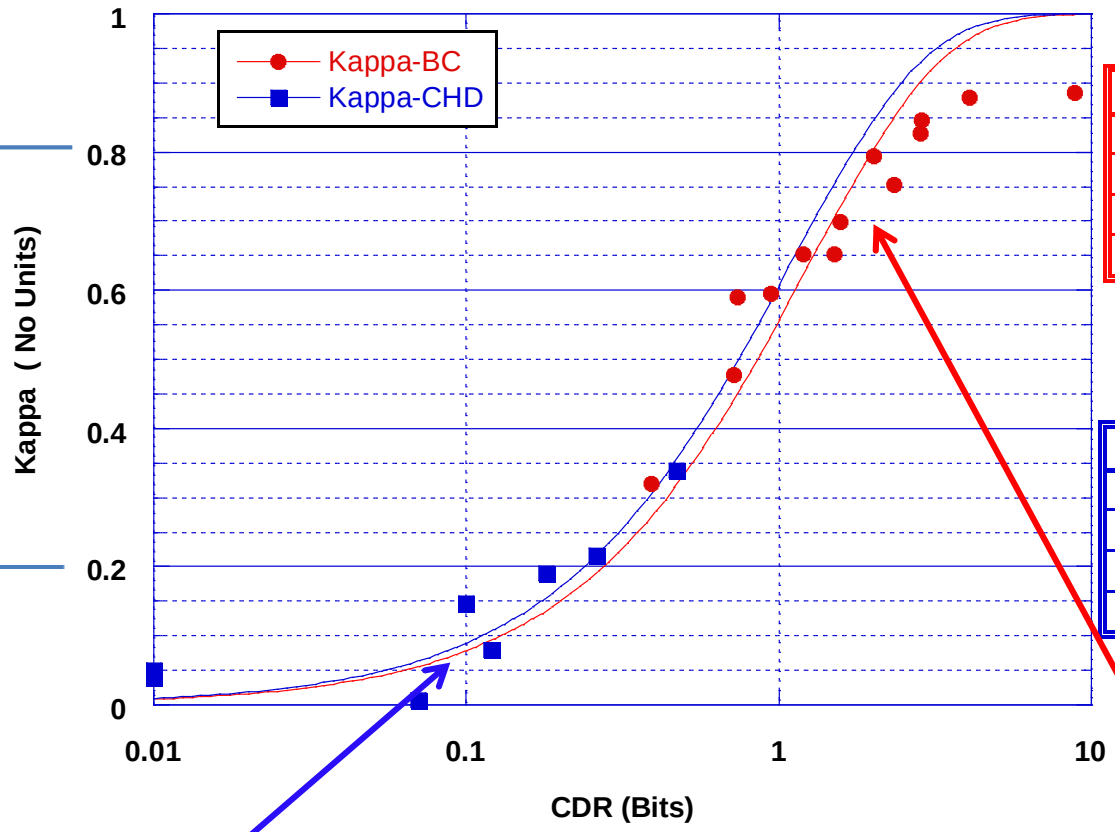
Very Good

Good

Moderate

Fair

Poor



y = 1- 2^(- M2*M0)		
	Value	Error
m2	1.1735	0.085073
Chisq	0.068431	NA
R	0.89262	NA

y = 1- 2^(- M2*M0)		
	Value	Error
m2	1.3474	0.15045
Chisq	0.017698	NA
R	0.91114	NA

CORIS-Heart Disease Data **Can identify the variables that are linked to CHD**
But population risk factor. Cannot identify individuals.

Breast cancer – Wisconsin **Can use ML to reliably identify health status of individual**

SUMMARY AND CONCLUSIONS

- **Collaboration and networking is vital.**
  All communities can learn much from one another.

- **Data confidentiality important. Aside – not only medical area. Most companies will not share their data for analysis. Want algorithms to work on their data which stays under their control.**

- **GDPR is getting to be more than data. The algorithms matter as well.**

- **Explaining the black box – matters to everyone.**

- **There are limits to what Machine Learning can do. Too much hype about,**