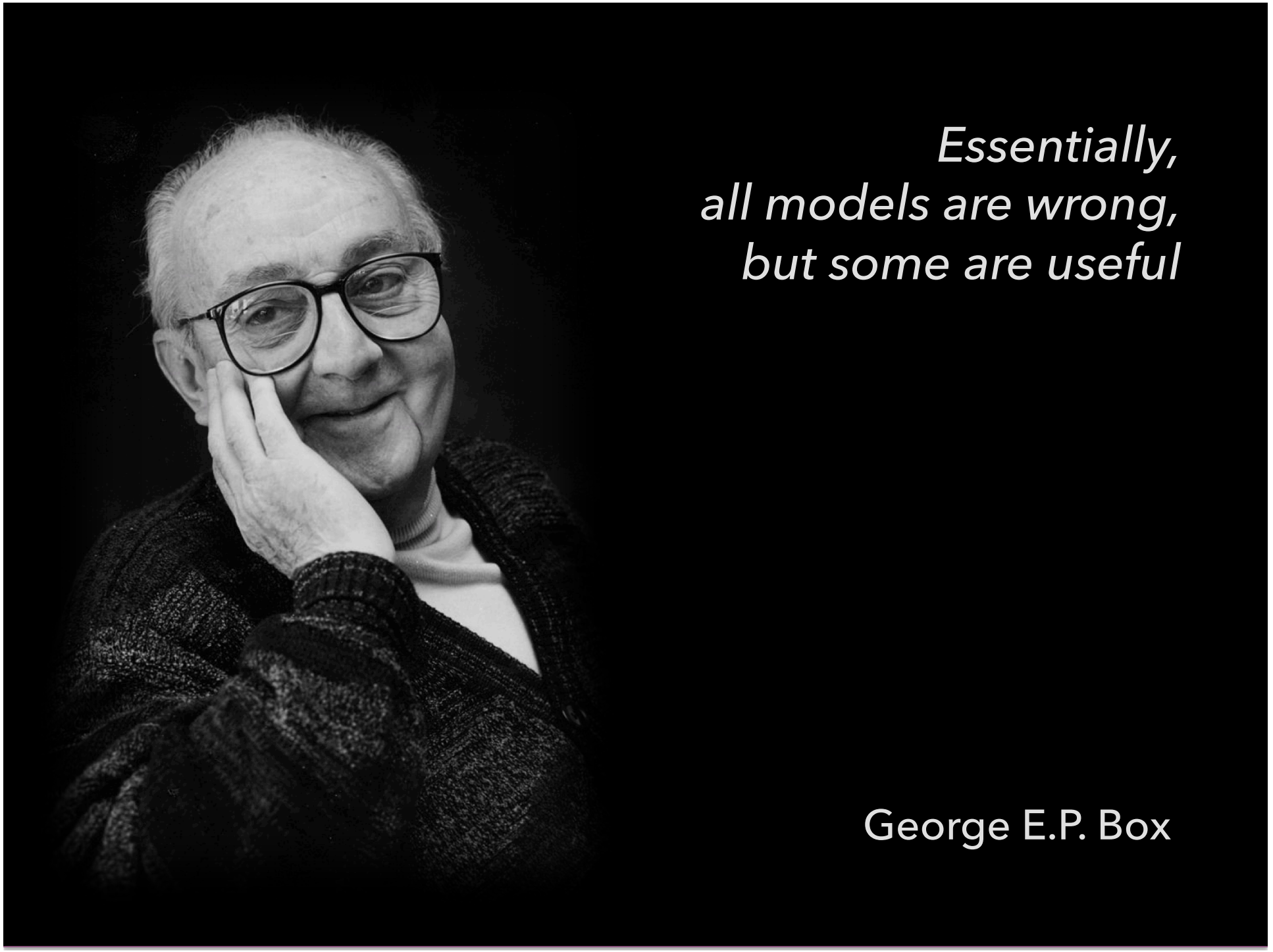
Transparency and interpretability of clinical prediction models

Niels Peek
Health eResearch Centre
School of Health Sciences
The University of Manchester

Opening the Black Box seminar
Brunel University, London, 21st November 2018


Essentially,
all models are wrong,
but some are useful
George E.P. Box

Menu

MANCHESTER 1824
The University of Manchester

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions

What are learning health systems?

Charles P. Friedman

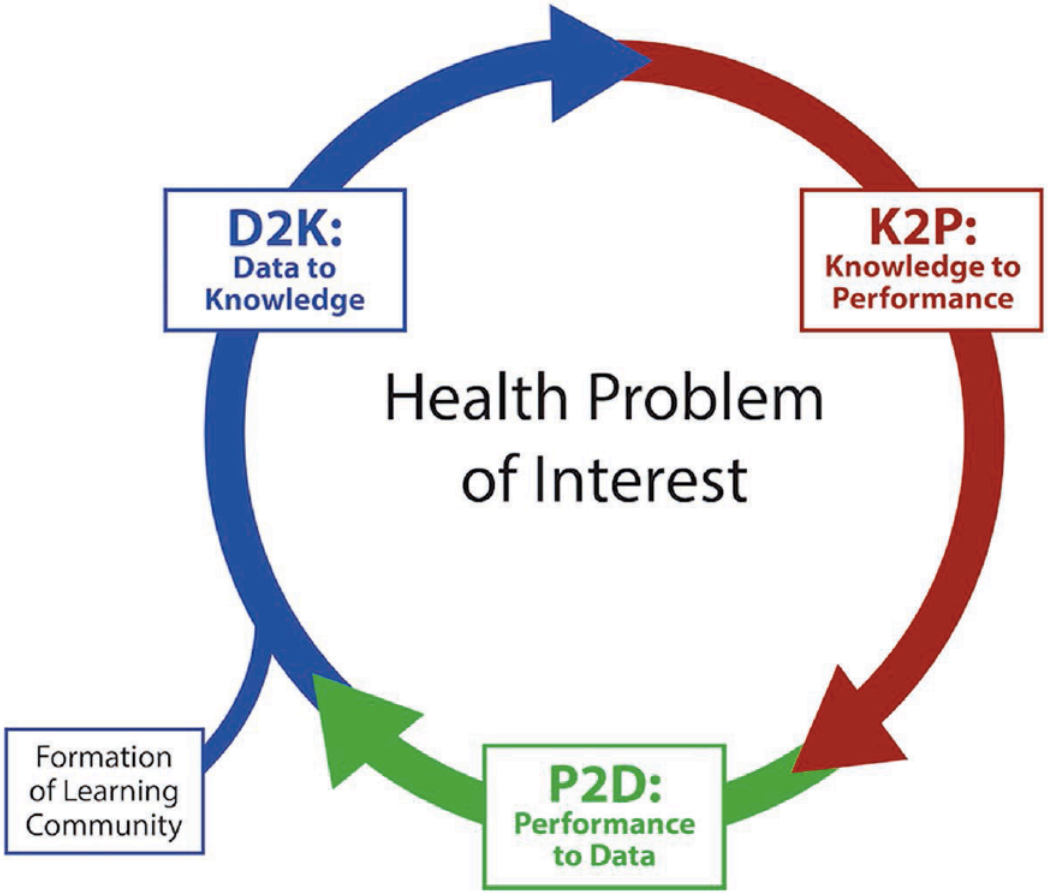
A system becomes a learning system when it can continuously and routinely improve itself by reflecting on its inputs, processes, and outputs.

A learning health system harnesses the power of data and technology to learn from every patient, and feed the knowledge of "what works best" back to clinicians and patients to create cycles of continuous improvement.

C. Friedman et al., Sci Trans Med 2010 Nov;2(57):57cm29.

The learning health cycle

The University of Manchester



Friedman et al., Yearb Med Inform 2017.

Example: community-acquired pneumonia

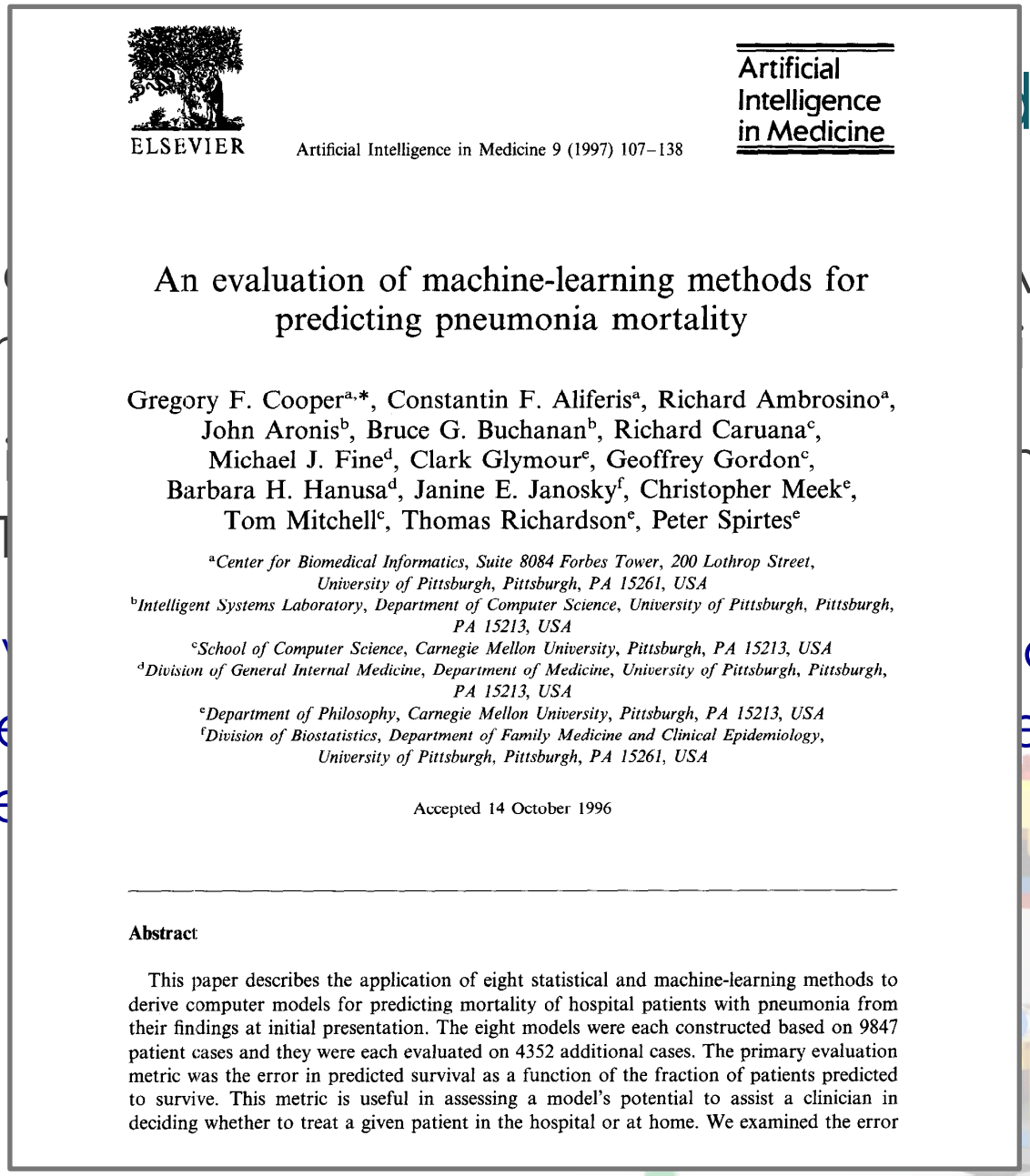
- Community-acquired pneumonia (CAP) is a common illness affecting >3m people annually in the US
- It is the 6th leading cause of death, and responsible for >1m hospital admissions per year

Example: community-acquired pneumonia

- Community-acquired pneumonia (CAP) is a common illness affecting >3m people annually in the US
- It is the 6th leading cause of death, and responsible for >1m hospital admissions per year
- If we can predict which CAP patients are at high risk of death, we can use these models to decide if a patient needs to be admitted to hospital

Exa… pneumonia

- C… AP) is a common illn… in the US
- It … nd responsible for >1…
- If … are at high risk of de… ecide if a patient ne…



ELSEVIER

Artificial Intelligence in Medicine 9 (1997) 107–138

Artificial Intelligence in Medicine

An evaluation of machine-learning methods for predicting pneumonia mortality

Gregory F. Cooper[a,*], Constantin F. Aliferis[a], Richard Ambrosino[a], John Aronis[b], Bruce G. Buchanan[b], Richard Caruana[c], Michael J. Fine[d], Clark Glymour[e], Geoffrey Gordon[c], Barbara H. Hanusa[d], Janine E. Janosky[f], Christopher Meek[e], Tom Mitchell[c], Thomas Richardson[e], Peter Spirtes[e]

[a]*Center for Biomedical Informatics, Suite 8084 Forbes Tower, 200 Lothrop Street, University of Pittsburgh, Pittsburgh, PA 15261, USA*
[b]*Intelligent Systems Laboratory, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15213, USA*
[c]*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[d]*Division of General Internal Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA*
[e]*Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[f]*Division of Biostatistics, Department of Family Medicine and Clinical Epidemiology, University of Pittsburgh, Pittsburgh, PA 15261, USA*

Accepted 14 October 1996

Abstract

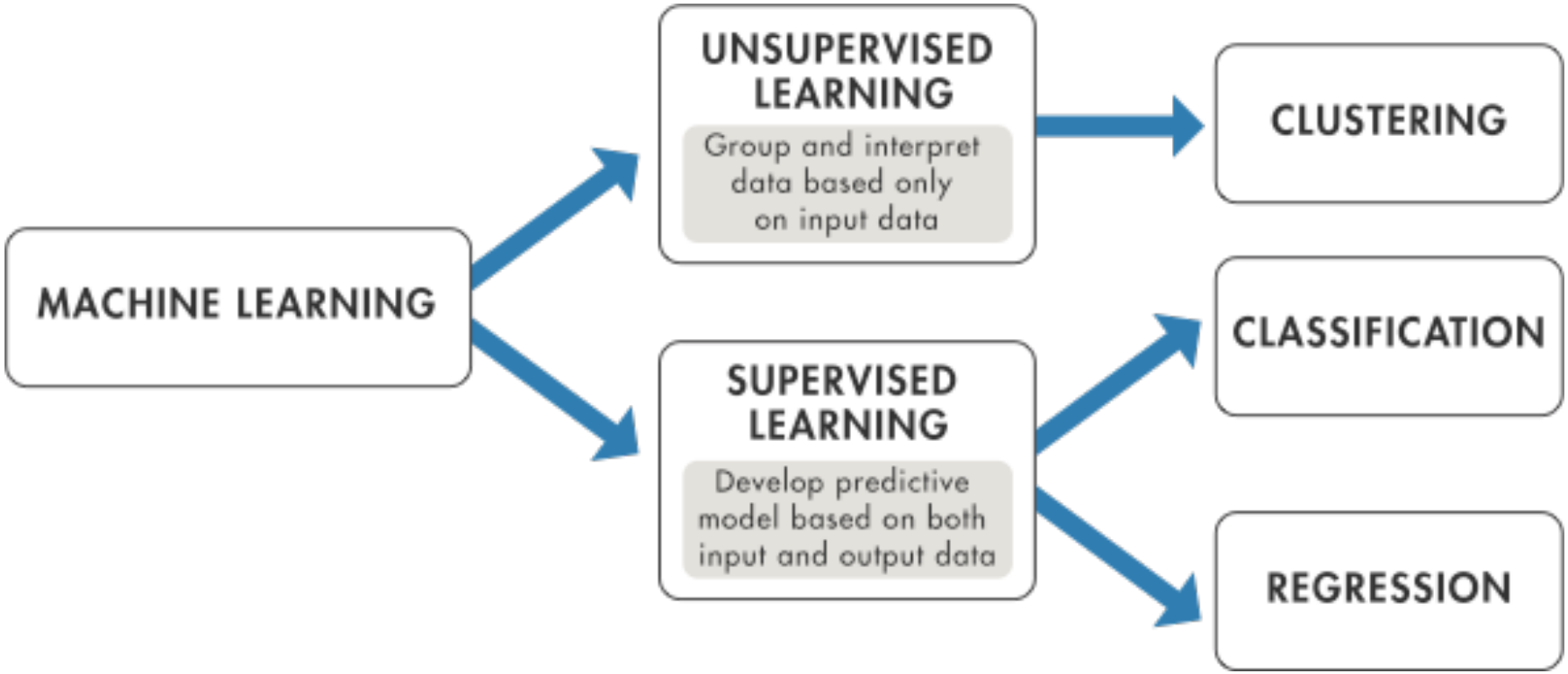
This paper describes the application of eight statistical and machine-learning methods to derive computer models for predicting mortality of hospital patients with pneumonia from their findings at initial presentation. The eight models were each constructed based on 9847 patient cases and they were each evaluated on 4352 additional cases. The primary evaluation metric was the error in predicted survival as a function of the fraction of patients predicted to survive. This metric is useful in assessing a model's potential to assist a clinician in deciding whether to treat a given patient in the hospital or at home. We examined the error

G. Cooper at al., Artificial Intelligence in Medicine 1997;3:107-38.

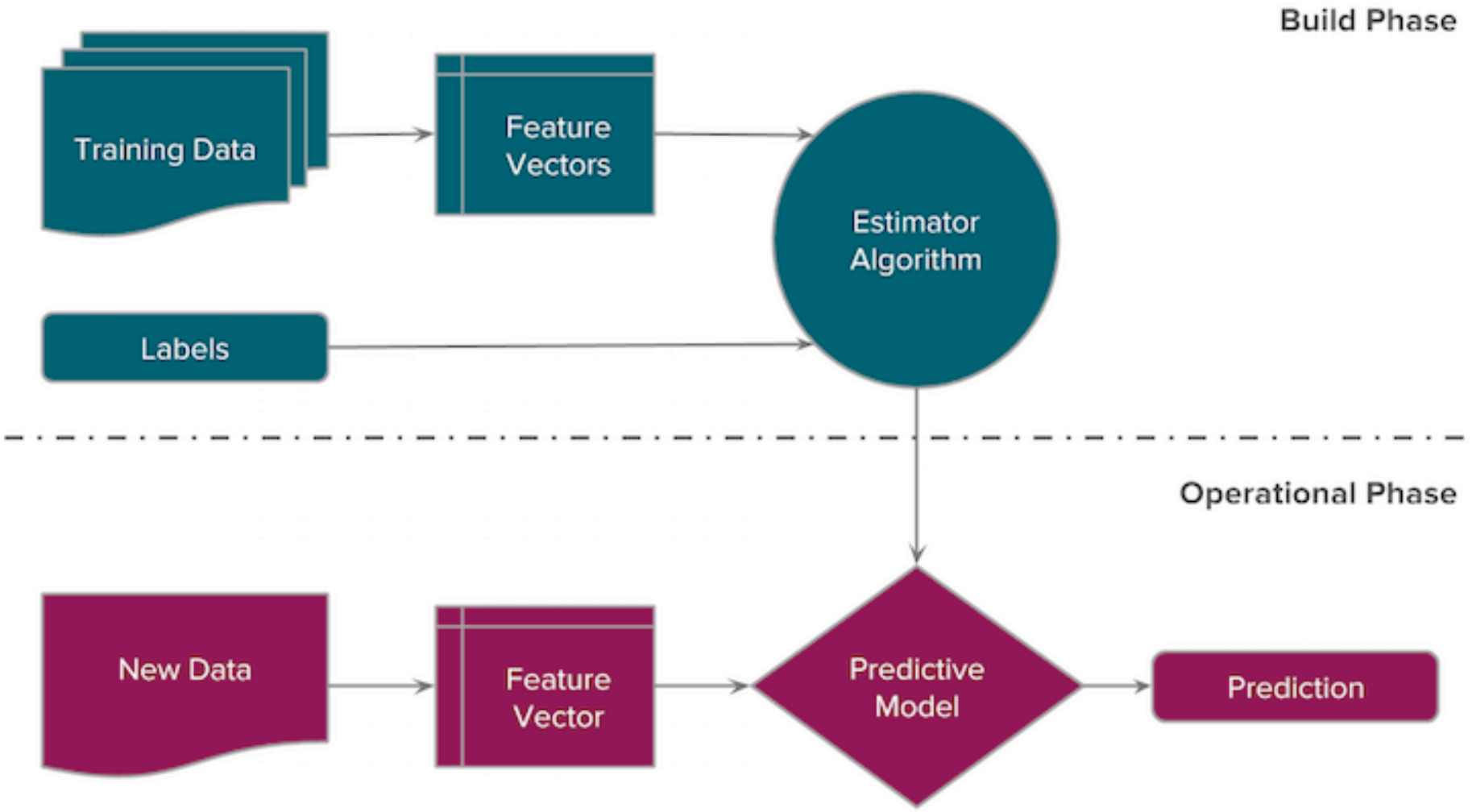
Menu

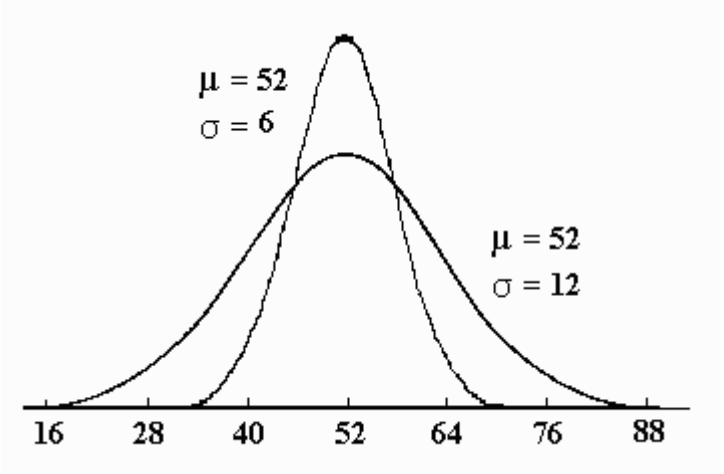
MANCHESTER 1824
The University of Manchester

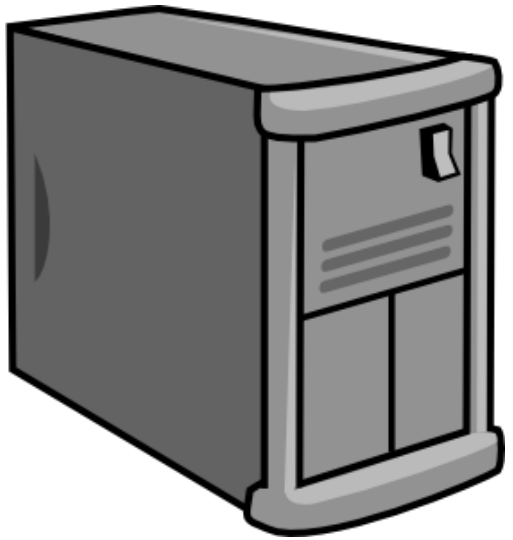
1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions


MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

Supervised learning







Statistics **Machine Learning**

T-Test

Logistic Regression

Elastic Net

Gradient Boosting

Deep Learning

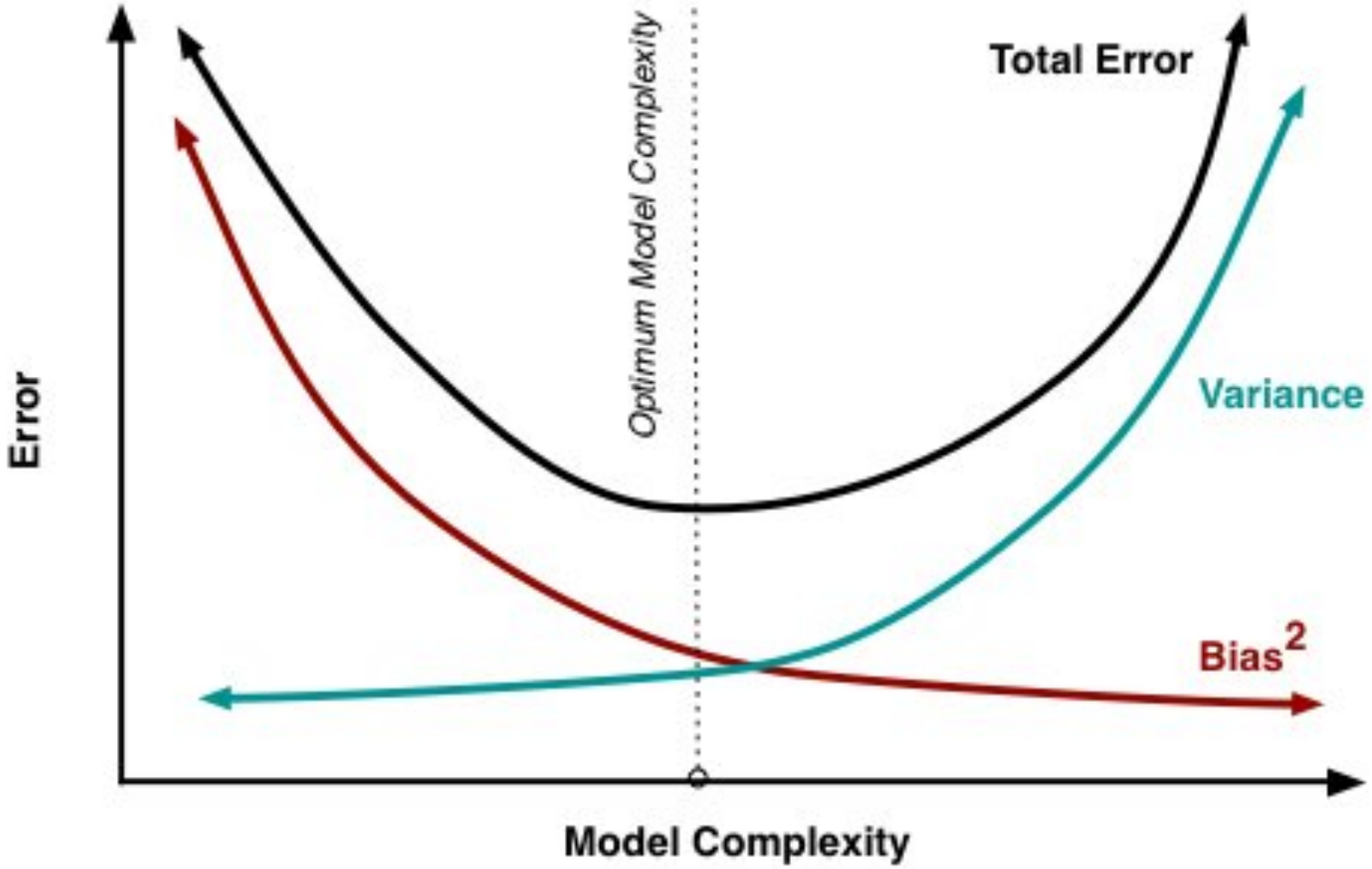
From a presentation by Tom Liptrot

Inductive bias

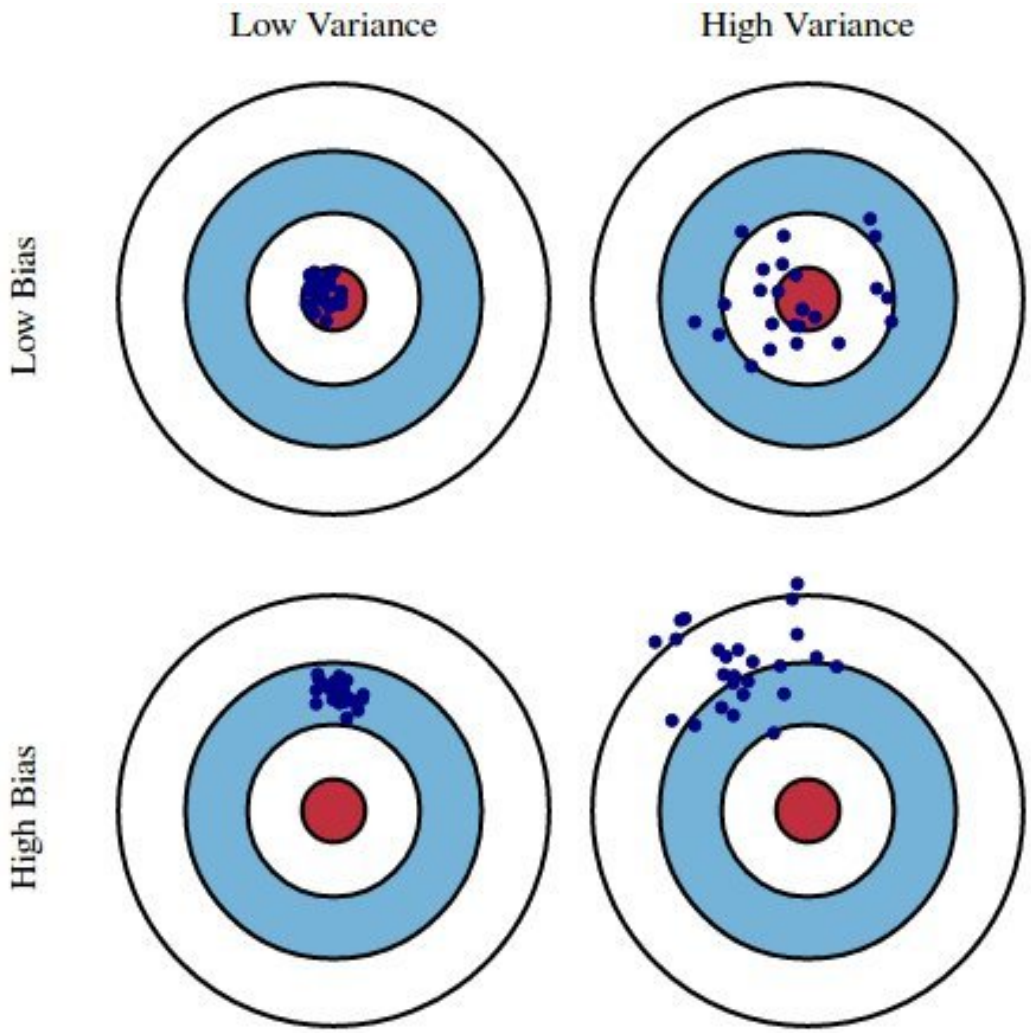
MANCHESTER 1824
The University of Manchester

- Inductive bias (= learning bias): The set of assumptions that a learning algorithm uses to construct a model from data
- Statistical models typically have a stronger inductive bias than machine learning methods, because they require prior specification of relevant features
- Assumption-free learning does not exist
- But we can reduce the impact of inductive bias by using more complex models
- ... at the expense of increasing variance

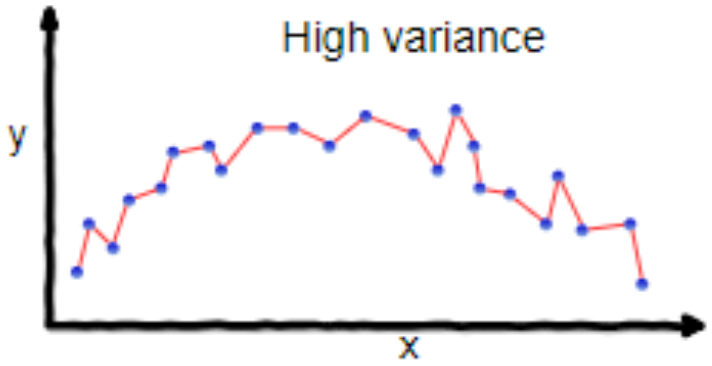
The bias-variance tradeoff (1)



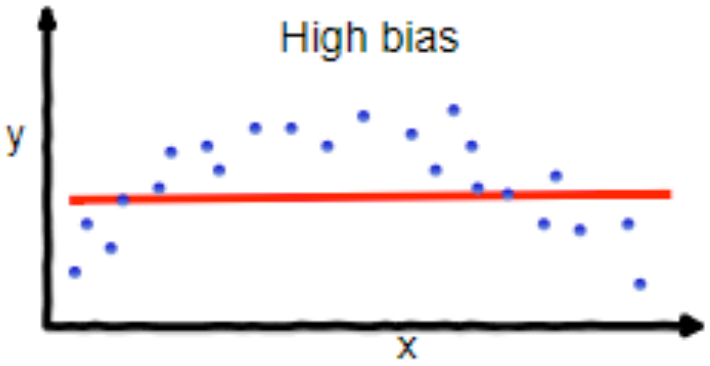
The bias-variance tradeoff (2)



The bias-variance tradeoff (3)



overfitting



underfitting

Menu

MANCHESTER 1824
The University of Manchester

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions

Explanatory models

MANCHESTER 1824
The University of Manchester

- Explanatory models are statistical models with high bias that are exclusively used for causal explanation
- They are used for testing causal hypotheses in observational data
- Predominant use of data in economics, psychology, education, and other social sciences
- In data science terms, they have a strong inductive bias

G. Shmueli, Statistical Science 2010;25(3):289-310.

Example: Unified Theory of Acceptance and Use of Technology (UTAUT)

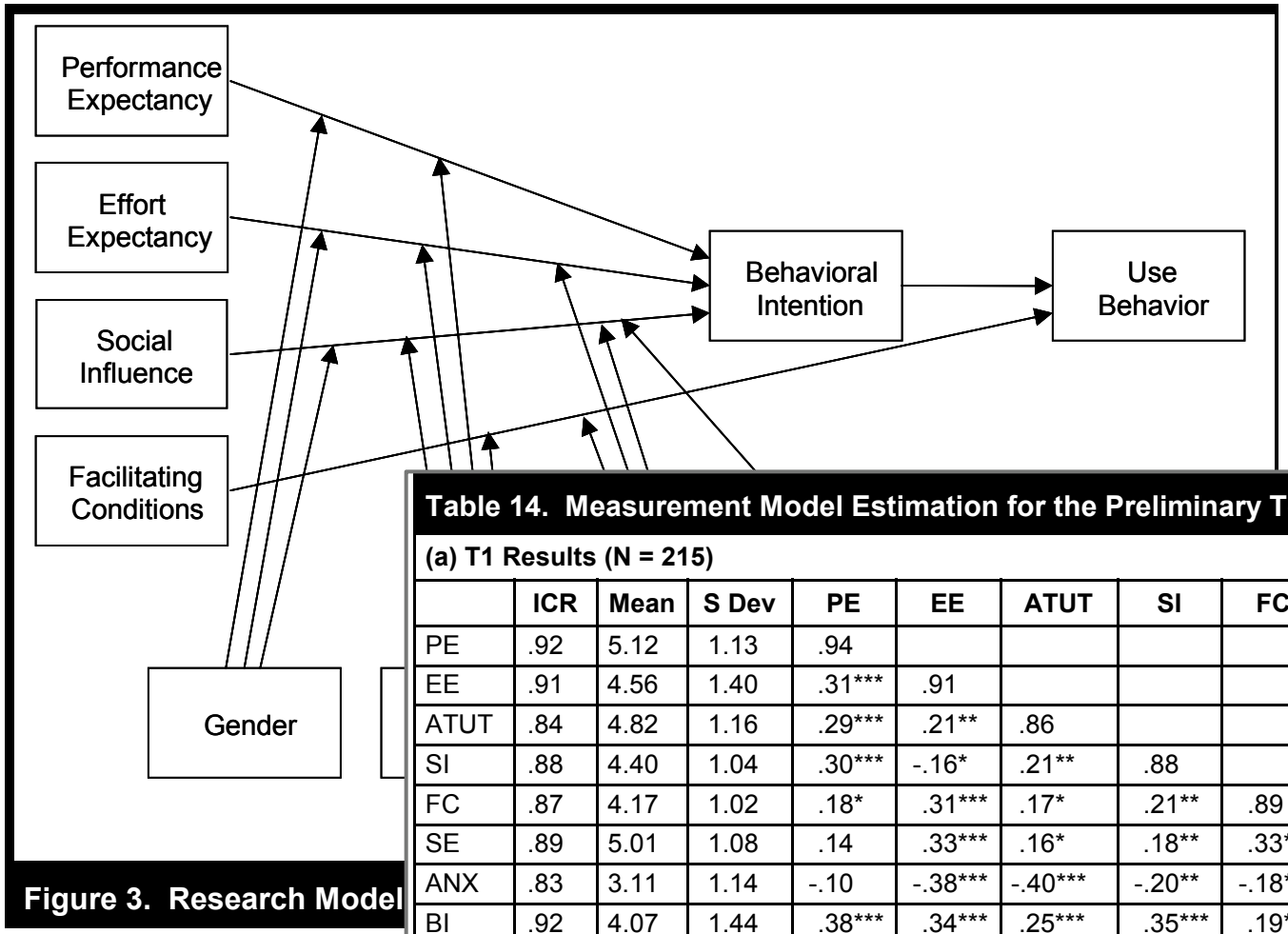


Figure 3. Research Model

Table 14. Measurement Model Estimation for the Preliminary Test of UTAUT

(a) T1 Results (N = 215)

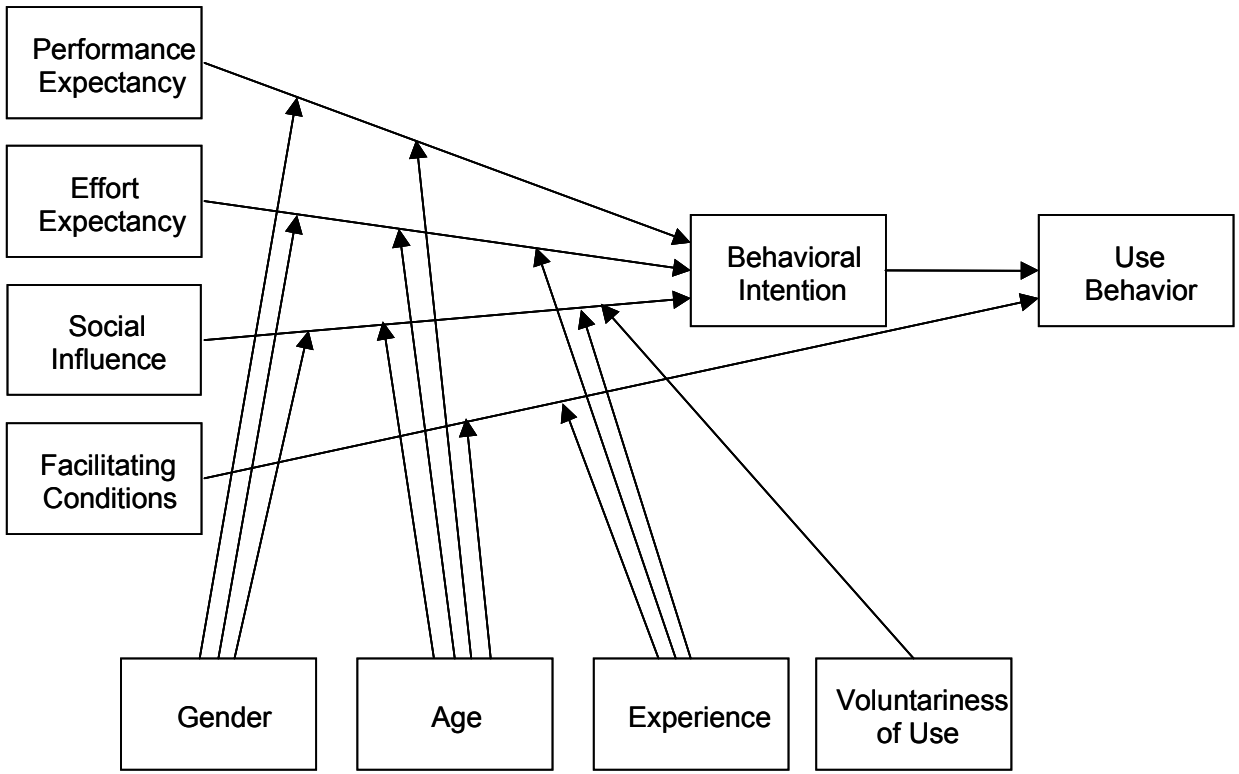
	ICR	Mean	S Dev	PE	EE	ATUT	SI	FC	SE	ANX	BI
PE	.92	5.12	1.13	.94							
EE	.91	4.56	1.40	.31***	.91						
ATUT	.84	4.82	1.16	.29***	.21**	.86					
SI	.88	4.40	1.04	.30***	-.16*	.21**	.88				
FC	.87	4.17	1.02	.18*	.31***	.17*	.21**	.89			
SE	.89	5.01	1.08	.14	.33***	.16*	.18**	.33***	.87		
ANX	.83	3.11	1.14	-.10	-.38***	-.40***	-.20**	-.18**	-.36***	.84	
BI	.92	4.07	1.44	.38***	.34***	.25***	.35***	.19**	.16*	-.23**	.84

V. Venkatesh et al., MIS Quarterly 2003; 27(3):425-78.

Question

The University of Manchester

Is it a problem if the UTAUT model is wrong?



WIRED

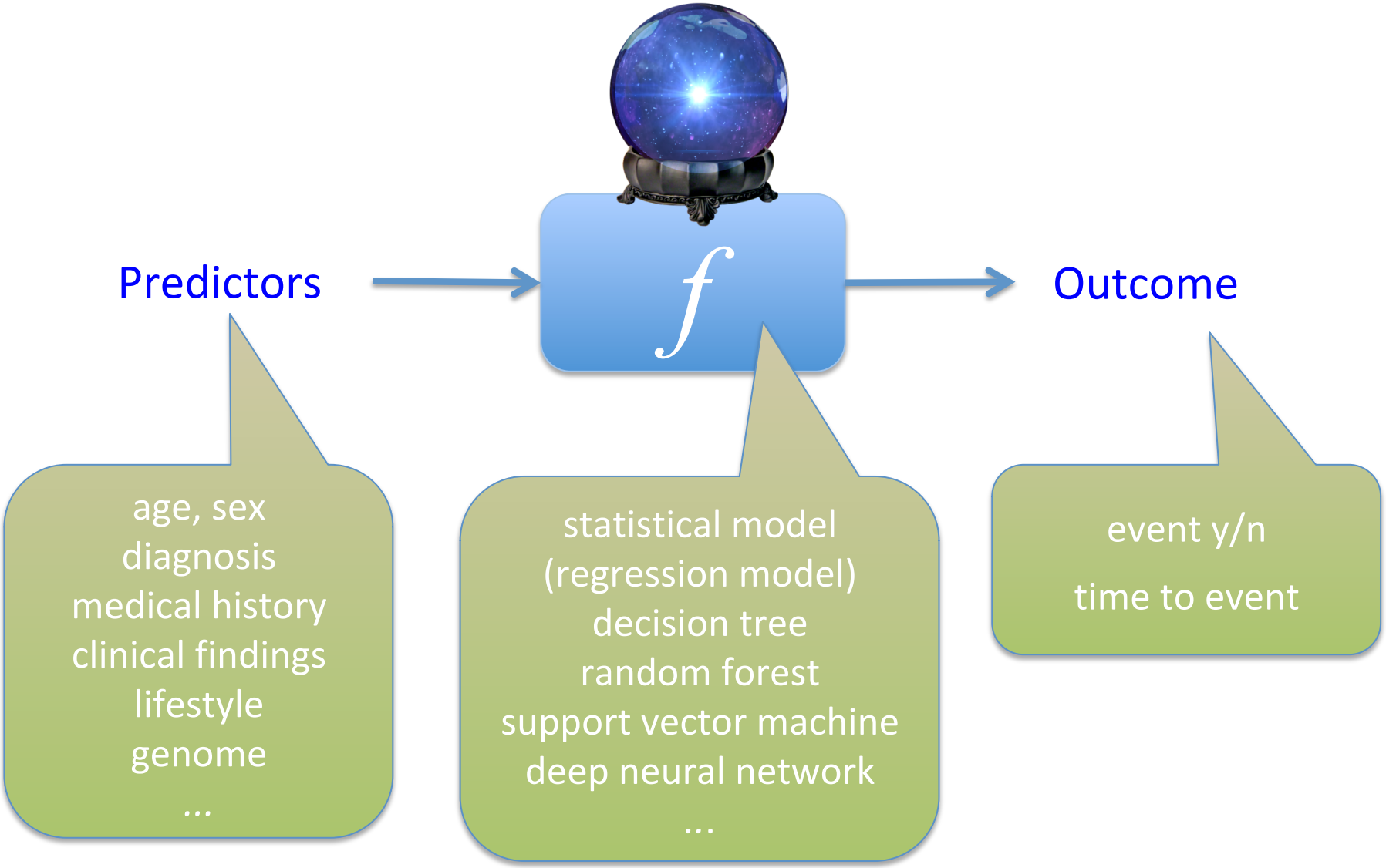
CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELU METH

The big target here isn't advertising, though. **It's science**. The scientific method is built around testable hypotheses. [...] Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). [...] But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.

Prediction models

The University of Manchester



Example revisited: hospital admissions

- Resources are too scarce to give preventive interventions to every patient
- Prediction models can help to deploy these resources in patients with the highest risks

Example revisited: pneumonia

- Community-acquired pneumonia (CAP) is a common illness affecting >3m people annually in the US
- It is the 6th leading cause of death, and responsible for >1m hospital admissions per year
- If we can predict which CAP patients are at high risk of death, we can use these models to decide if a patient needs to be admitted to hospital

Question

Is it a problem if the risk prediction model is wrong?

Is it important that we can interpret such a model or understand its predictions?

Menu

MANCHESTER 1824
The University of Manchester

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions

The Farr Institute of Health Informatics Research

The Mythos of Model Interpretability

IN MACHINE LEARNING, THE CONCEPT OF INTERPRETABILITY IS BOTH IMPORTANT AND SLIPPERY.

ZACHARY C. LIPTON

Supervised machine-learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? Models should be not only good, but also interpretable, yet the task of interpretation appears underspecified. The academic literature has provided diverse and sometimes non-overlapping motivations for interpretability and has offered myriad techniques for rendering interpretable models. Despite this ambiguity, many authors proclaim their models to be interpretable axiomatically, absent further argument. Problematically, it is not clear what common properties unite these techniques.

This article seeks to refine the discourse on interpretability. First it examines the objectives of previous

Z. Lipton, ACM Magazine Queue - Machine Learning 2018;16(3).

Model interpretability: What?

MANCHESTER 1824
The University of Manchester

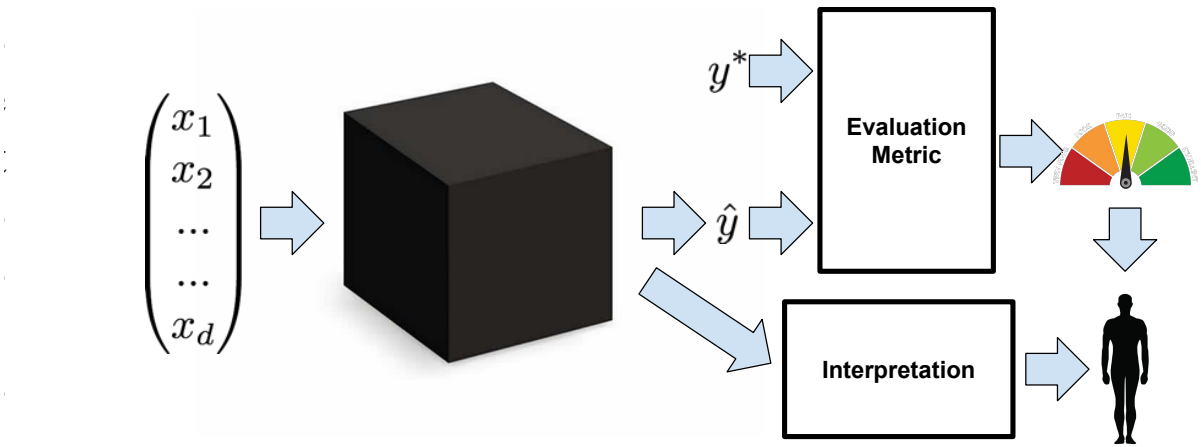
- The notion of interpretability is ill-defined, and has no formal meaning
- Claims regarding interpretability often exhibit a quasi-scientific character
- It is useful to make a distinction between:
  - model transparency
  - post-hoc interpretability of predictions

Z. Lipton, ACM Magazine Queue - Machine Learning 2018;16(3).

Model interpretability: Why?

The University of Manchester

- The formal objectives of supervised learning (test set performance) do not capture interpretability
- The demand for interpretability arises from additional objectives related to real-world deployment



Z. Lipton, ACM Magazine Queue - Machine Learning 2018;16(3).

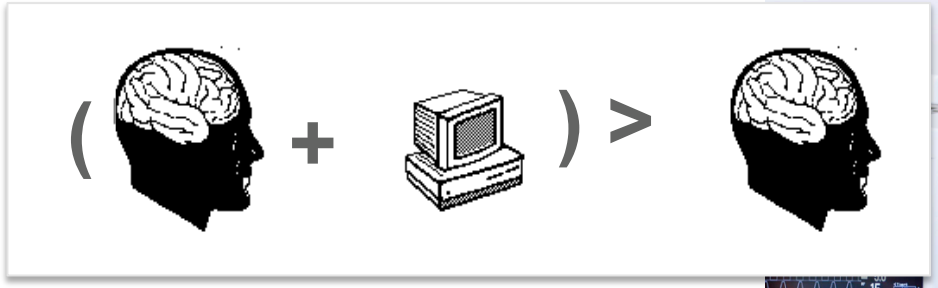
Model interpretability: Why?

MANCHESTER 1824
The University of Manchester

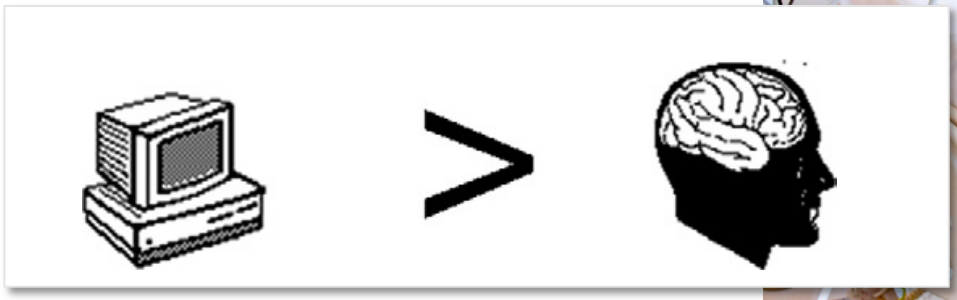
- Trust
- Causality
- Transferability
- Informativeness
- Fair and ethical decision making

Z. Lipton, ACM Magazine Queue - Machine Learning 2018;16(3).

Fundamental theorem of medical informatics



And not:



CP Friedman, JAMIA 2009;16(2):169-170

VIEWPOINT

Clinical Decision Support in the Era of Artificial Intelligence

Edward H. Shortliffe, MD, PhD
Biomedical Informatics, Columbia University, New York, New York; and Biomedical Informatics, Arizona State University, Phoenix.

Martin J. Sepúlveda, MD, ScD
Retired from IBM Research, Watson Research Laboratory, Yorktown Heights, New York.

Clinicians and researchers have long envisioned the day when computers could assist with difficult decisions in complex clinical situations. The first article on this subject appeared in the scientific literature about 60 years ago,[1] and the notion of computer-based clinical decision support has subsequently been a dominant topic for informatics research. Two recent Viewpoints in *JAMA* highlighted the promise of deep learning in medicine.[2,3] Such new data analytic methods have much to offer in interpreting large and complex data sets. This Viewpoint is focused on the subset of decision support systems that are designed to be used interactively by clinicians as they seek to reach decisions, regardless of the underlying analytic methodology that they incorporate.

With the evolution of digital and communication technologies plus innovative software methods, the ability to offer high-quality support to clinicians has resulted in impressive new capabilities and several commercial products. For example, many decision support tools are built into medical devices, creating new ways to visualize or interpret data that are provided to expert users. Artificial intelligence programs, which are increasingly based on a variety of machine learning and natural language processing methods, are especially prominent in these data interpretation and text mining settings.

> Despite the enthusiasm for exploring the potential of artificial intelligence and decision support in clinical settings, several complexities limit the ability to move ahead quickly.

Why, then, do clinical decision support systems (CDSSs) designed for direct interactive use by clinicians have challenges of credibility and adoption when the literature has been replete for 4 decades with studies that present computing systems demonstrating diagnostic accuracy that rivals the performance of expert clinicians?[4,5] The reasons are varied and reflect the realities and complexities of clinical practice. Biomedical informaticians have long understood those reasons, recognizing the spectrum of capabilities and characteristics that must be incorporated into a CDSS if it is to be accepted and integrated into routine workflow:

- Black boxes are unacceptable: A CDSS requires transparency so that users can understand the basis for any advice or recommendations that are offered.
- Time is a scarce resource: A CDSS should be efficient in terms of time requirements and must blend into the workflow of the busy clinical environment.
- Complexity and lack of usability thwart use: A CDSS should be intuitive and simple to learn and use so that major training is not required and it is easy to obtain advice or analytic results.
- Relevance and insight are essential: A CDSS should reflect an understanding of the pertinent domain and the kinds of questions with which clinicians are likely to want assistance.
- Delivery of knowledge and information must be respectful: A CDSS should offer advice in a way that recognizes the expertise of the user, making it clear that it is designed to inform and assist but not replace a clinician.
- Scientific foundation must be strong: A CDSS should have rigorous, peer-reviewed scientific publications establishing its safety, validity, reproducibility, and reliability.

Health care is a particularly challenging domain for decision support. A CDSS requires strong capabilities that can function in a domain where the understanding of causal mechanisms and relationships is still incomplete and where uncertainty, and an approach to dealing with it, is accordingly inevitable. To provide valid support while simultaneously addressing the list of requirements to help ensure adoption by clinicians. For example, effective decision support capabilities are those that avoid additional data entry tasks, such as those that acquires the bulk of the data needed to assess a case through integration with an electronic health record (EHR). Today's EHRs have not made this easy because they generally lack the cross-platform transparency and standards that would be required for a single CDSS to be tightly integrated with multiple EHR products or implementations.

Different decision-making tasks often pose different challenges for a CDSS. For example, a system designed to assist with clinical diagnosis is very different from one that is intended to assist with therapy planning. A CDSS for diagnosis can generally be built on linkages between clinical data and gold standards for accuracy (eg, biopsies, autopsies, biomolecular markers, or surgical findings). But in formulating a therapeutic plan, especially in complex settings, there is often no gold standard, and there may be disagreement, even among experts. For example, an early study evaluated a program designed to assist with the selection of antibiotic therapy

Corresponding Author: Edward H. Shortliffe, MD, PhD, Biomedical Informatics, Columbia University, 272 W 107th St, 5B, New York, NY 10025 (ted@shortliffe.net; ehs79@columbia.edu).

MANCHESTER 1824
The University of Manchester

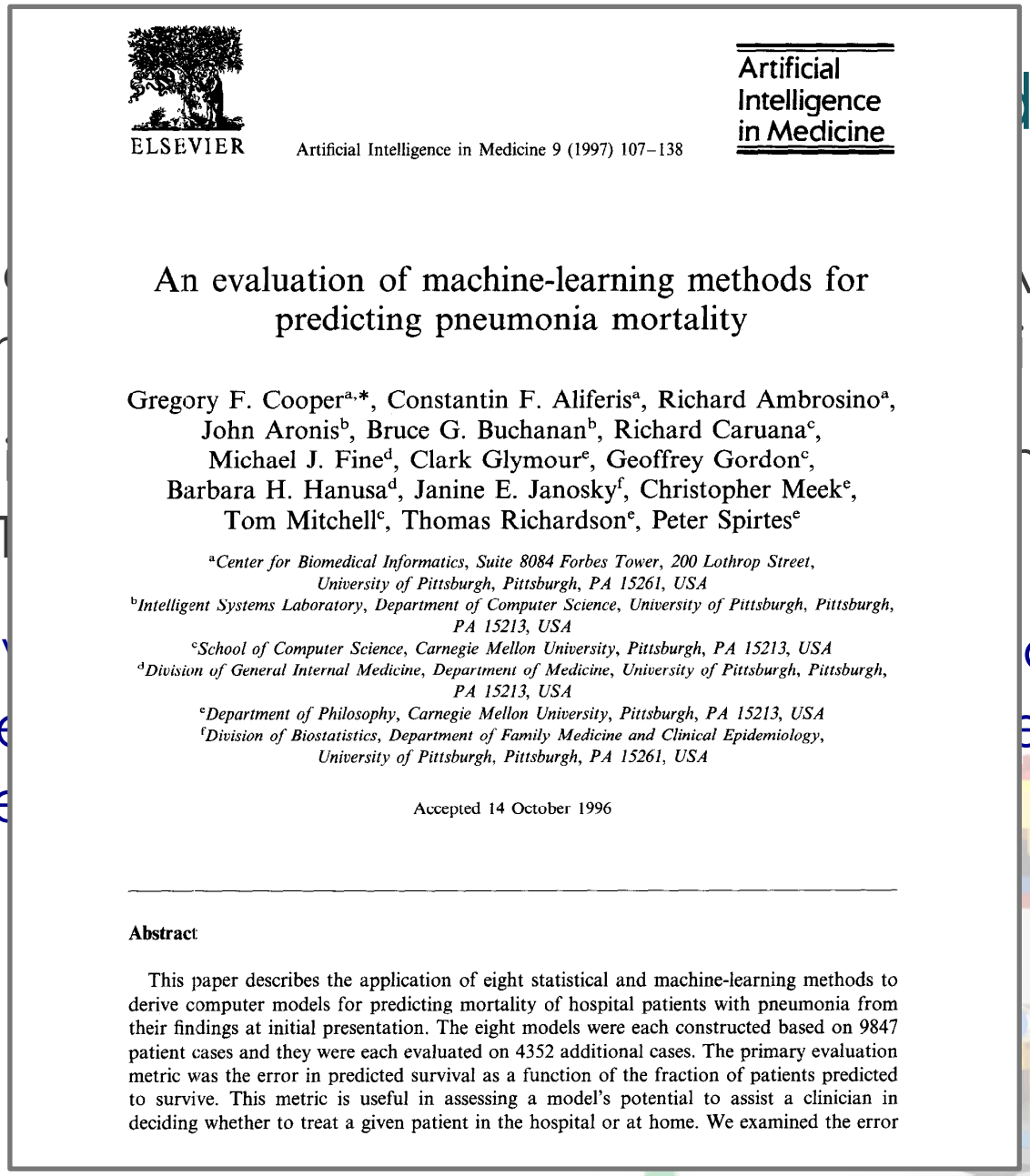
"Black boxes are unacceptable: A CDSS requires transparency so that users can understand the basis for any advice or recommendations that are offered."

"A CDSS should offer advice in a way that recognizes the expertise of the user, making it clear that it is designed to inform and assist but not to replace a clinician."

E.H. Shortliffe & M.J. Sepúlveda, JAMA 5th Nov 2018.

Exa… pneumonia

- C… AP) is a common illn… in the US
- It … nd responsible for >1…
- If … are at high risk of de… ecide if a patient ne…



ELSEVIER

Artificial Intelligence in Medicine 9 (1997) 107–138

Artificial Intelligence in Medicine

An evaluation of machine-learning methods for predicting pneumonia mortality

Gregory F. Cooper[a,*], Constantin F. Aliferis[a], Richard Ambrosino[a], John Aronis[b], Bruce G. Buchanan[b], Richard Caruana[c], Michael J. Fine[d], Clark Glymour[e], Geoffrey Gordon[c], Barbara H. Hanusa[d], Janine E. Janosky[f], Christopher Meek[e], Tom Mitchell[c], Thomas Richardson[e], Peter Spirtes[e]

[a]*Center for Biomedical Informatics, Suite 8084 Forbes Tower, 200 Lothrop Street, University of Pittsburgh, Pittsburgh, PA 15261, USA*
[b]*Intelligent Systems Laboratory, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15213, USA*
[c]*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[d]*Division of General Internal Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA*
[e]*Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[f]*Division of Biostatistics, Department of Family Medicine and Clinical Epidemiology, University of Pittsburgh, Pittsburgh, PA 15261, USA*

Accepted 14 October 1996

Abstract

This paper describes the application of eight statistical and machine-learning methods to derive computer models for predicting mortality of hospital patients with pneumonia from their findings at initial presentation. The eight models were each constructed based on 9847 patient cases and they were each evaluated on 4352 additional cases. The primary evaluation metric was the error in predicted survival as a function of the fraction of patients predicted to survive. This metric is useful in assessing a model's potential to assist a clinician in deciding whether to treat a given patient in the hospital or at home. We examined the error

G. Cooper at al., Artificial Intelligence in Medicine 1997;3:107-38.

What Cooper et al. found

MANCHESTER 1824
The University of Manchester

- The most accurate model was a neural network which outperformed other methods (e.g. logistic regression) by a wide margin
- One of the methods was a rule-based method that found the rule: HasAsthama(x) => LowerRisk(x)
- The authors chose the deploy the rule-based model, and left out this rule
- Several authors have since argued that prediction models must be intelligible and editable

G. Cooper at al., Artificial Intelligence in Medicine 1997;3:107-38.

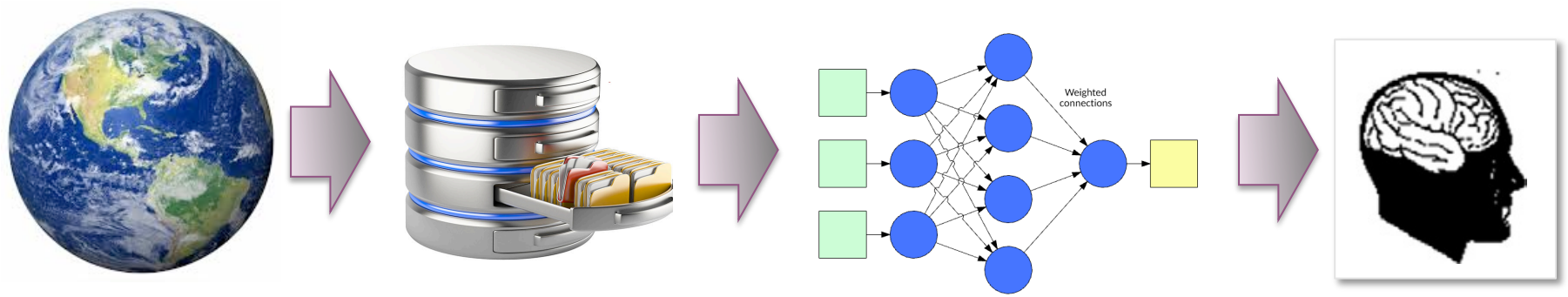
Question

MANCHESTER 1824
The University of Manchester

Why did the rule-based method infer that asthma patients were at low risk?

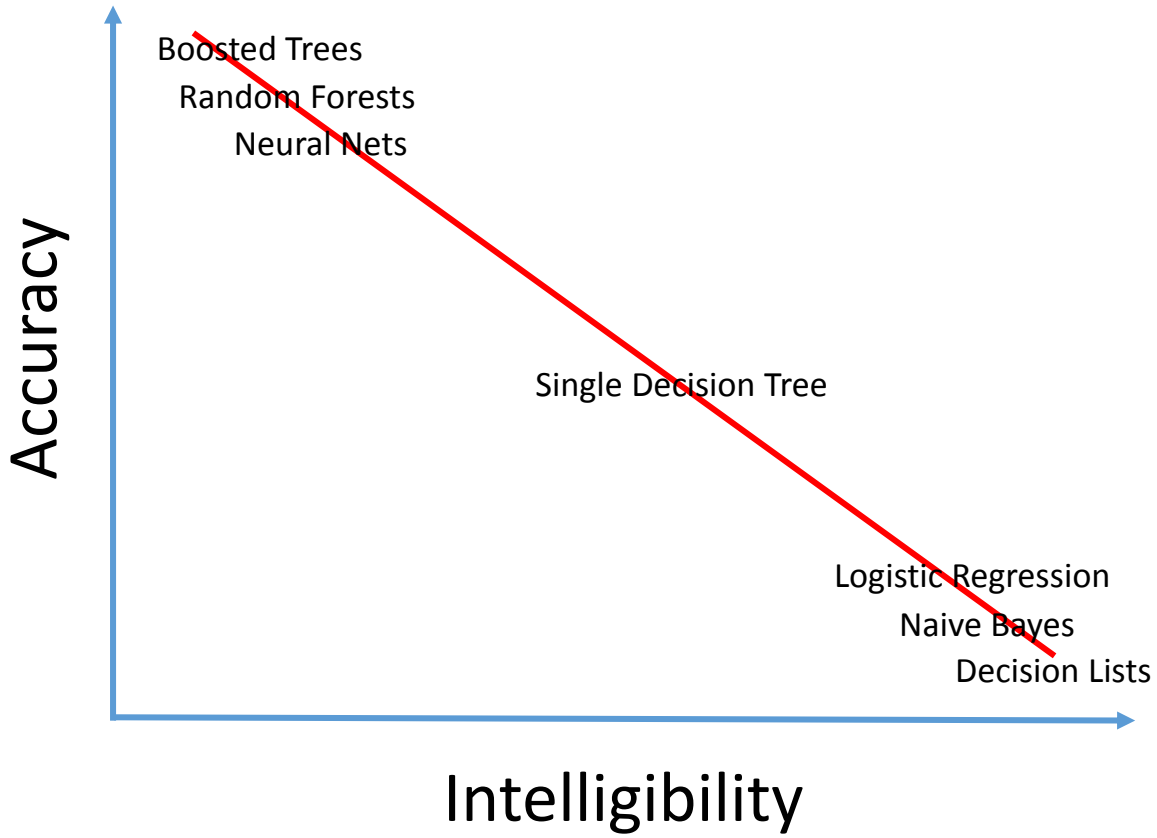
Opening the black box

The University of Manchester



- If managed in the same way, asthma patients with CAP would be at a higher risk than other patients
- In our current care system, this risk is recognised and therefore asthma patients are managed differently
- Their net risk is therefore lower

Accuracy vs intelligibility



From a presentation by Rich Caruana

Citizens Juries

The University of Manchester

- A "Citizens Jury" is a public engagement process that allows policy makers to hear thoughtful input from an informed microcosm of the public
- In Feb/Mar 2019 we will organise 2 Citizens Juries (5 days each) on explainable AI
- The juries will explore the trade-offs between performance and explainability of computer algorithms
- Scenarios in clinical medicine, criminal justice, and professional recruitment will be considered

Menu

The University of Manchester

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions

The Farr Institute of Health Informatics Research

Conclusions (1)

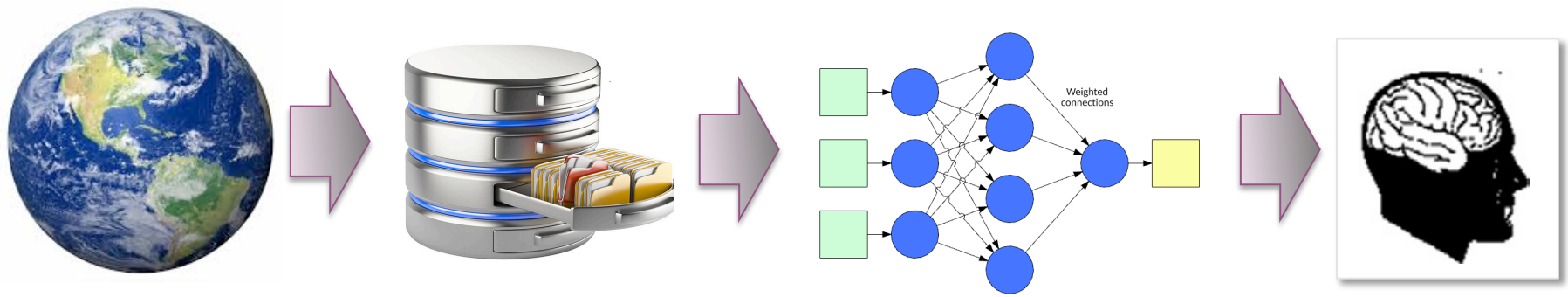
MANCHESTER 1824
The University of Manchester

- All models are wrong, but some are useful
- Prediction models are a radically pragmatic, "end-of-theory" use of data to engineer systems
- Their core purpose is to make predictions for future, unseen instances – not to increase our understanding
- But at the interface with humans, the need arises to provide interpretability

Conclusions (2)

MANCHESTER 1824
The University of Manchester

- Model interpretability is still a poorly defined notion
- It is ultimately something that should be studied by psychologists, not computer scientists
- To understand a model, we must understand its relationship with the real world



Niels Peek
MRC Health eResearch Centre
The University of Manchester, UK
niels.peek@manchester.ac.uk
@NielsPeek