Clinical prediction models in the age of artificial intelligence and big data

Ewout Steyerberg

Professor of Clinical Biostatistics and Medical Decision Making

<E.Steyerberg@ErasmusMC.nl /

E.W.Steyerberg@LUMC.nl >

Basel, Nov 1 2019

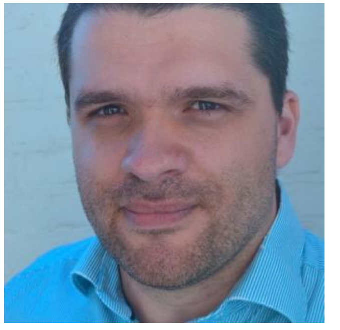
Leiden University Medical Center

Erasmus MC

University Medical Center Rotterdam

Thanks to co-workers; no COI

- LUMC: Maarten van Smeden



- Leuven: Ben van Calster

Both provided many of the slides shown

Main question

Where does Big Data / machine learning (ML) / artificial intelligence (AI) assist us in prediction research?

- Strengths and weaknesses of Big Data initiatives
- Consider links between classical statistical approaches, ML, AI for prediction

Prediction models; what for?

- Understanding nature:
  relative risks of different predictors
- Predicting outcomes:
  absolute risk by combinations of predictors

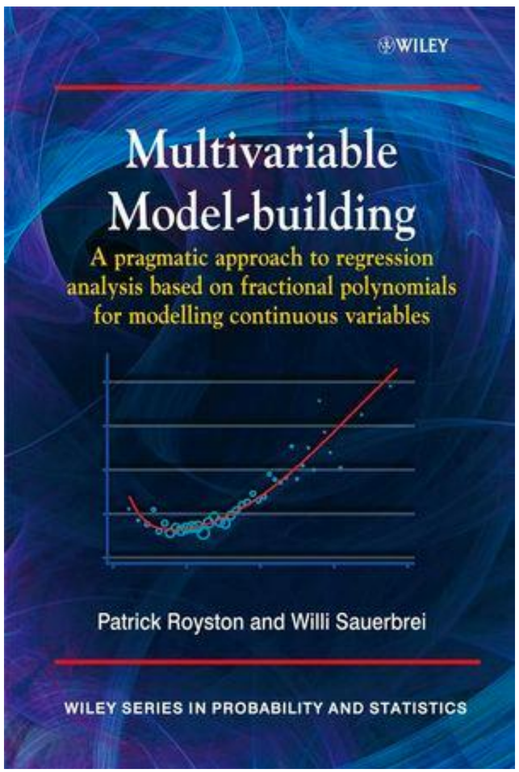
Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330
© Institute of Mathematical Statistics, 2010

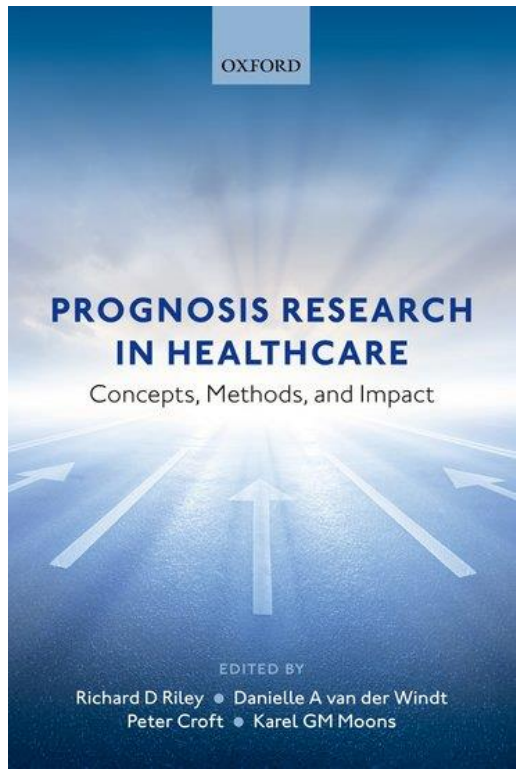
To Explain or to Predict?

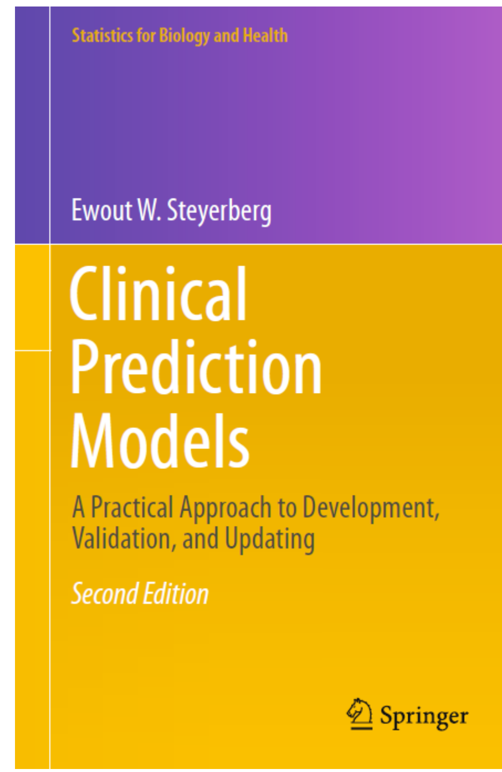
Galit Shmueli

Traditional regression modeling

Can well be used for explanation and prediction







Prediction models

- Diagnosis
  - Imaging findings, e.g. abnormal CT scan in trauma
  - Clinical condition, e.g. serious infection
  - …
- Prognosis
  - Mortality, e.g. < 30 days, over time, …
  - …

Prognostic / predictive models

Prognostic modeling

y ~ X	Prognostic factors
y ~ Tx	Treatment effect
y ~ X + Tx	Covariate adjusted tx effect

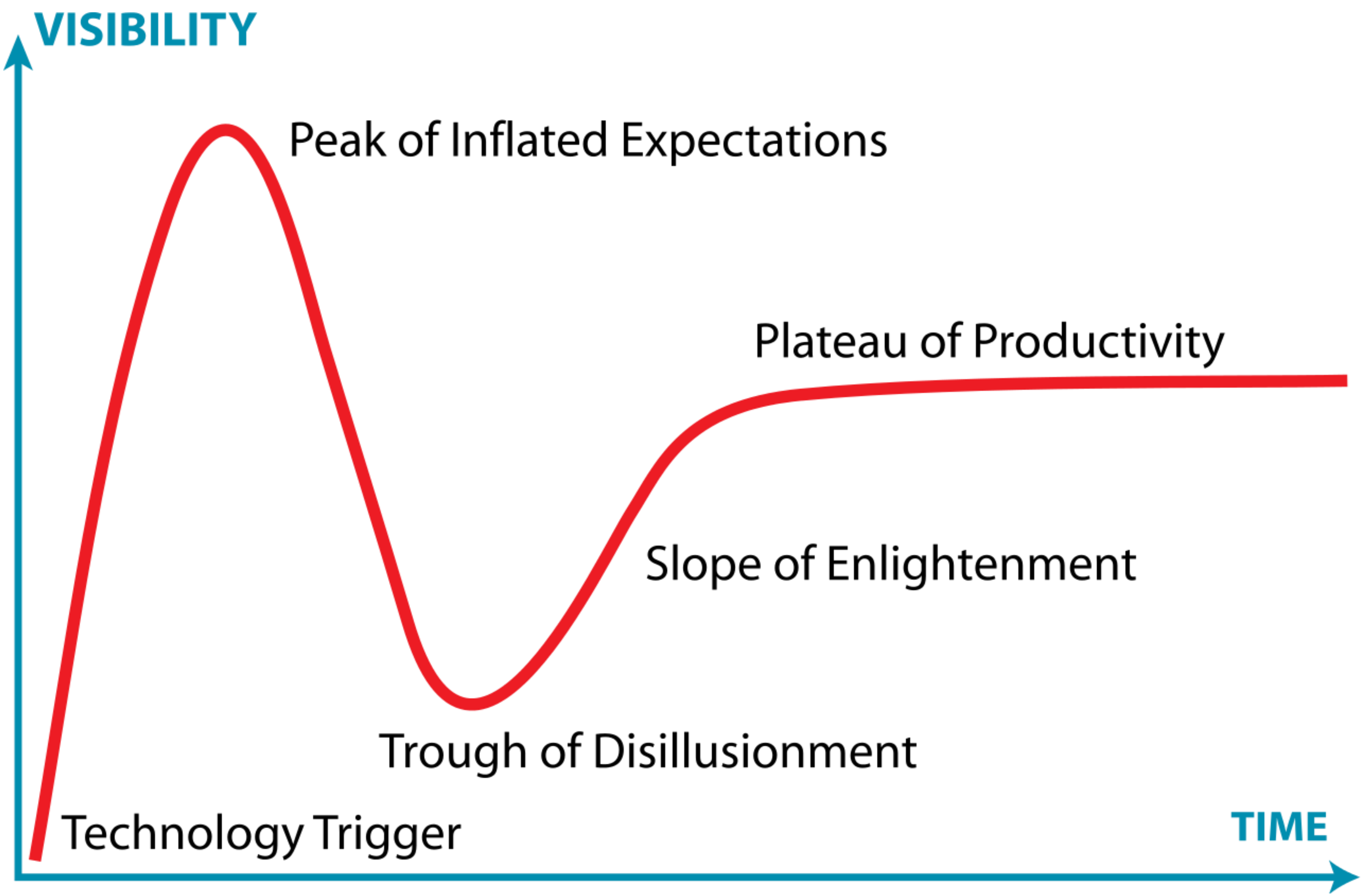
Predictive modeling

y ~ X * Tx	Predictive factors for differential tx effect

Opportunities in medical prediction

- More data
  - larger N
  - more variables
- More detail
  - biomarkers / omics / imaging / eHealth
- Novel methods
  - ML / AI / ..
  - Statistical methods
    - Dynamic prediction
    - Testing procedures for high dimensional data
    - …

Hype

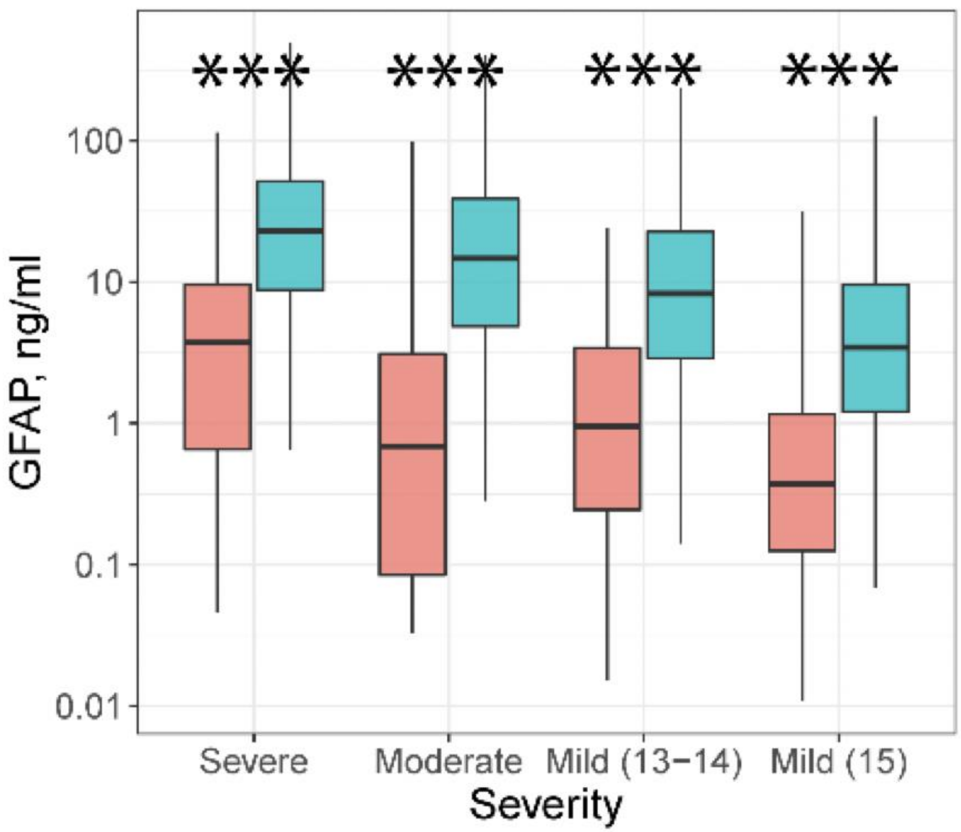

VISIBILITY
Peak of Inflated Expectations
Plateau of Productivity
Slope of Enlightenment
Trough of Disillusionment
Technology Trigger
TIME

Examples

- Biomarkers
- Imaging
- Omics

Positive example 1

- Biomarkers in diagnosing head trauma
  - Mild: AUC 0.89 [0.87-0.90] vs clinical 0.84 [0.83-0.86]



Positive example 2

- MRI Imaging in diagnosing prostate cancer

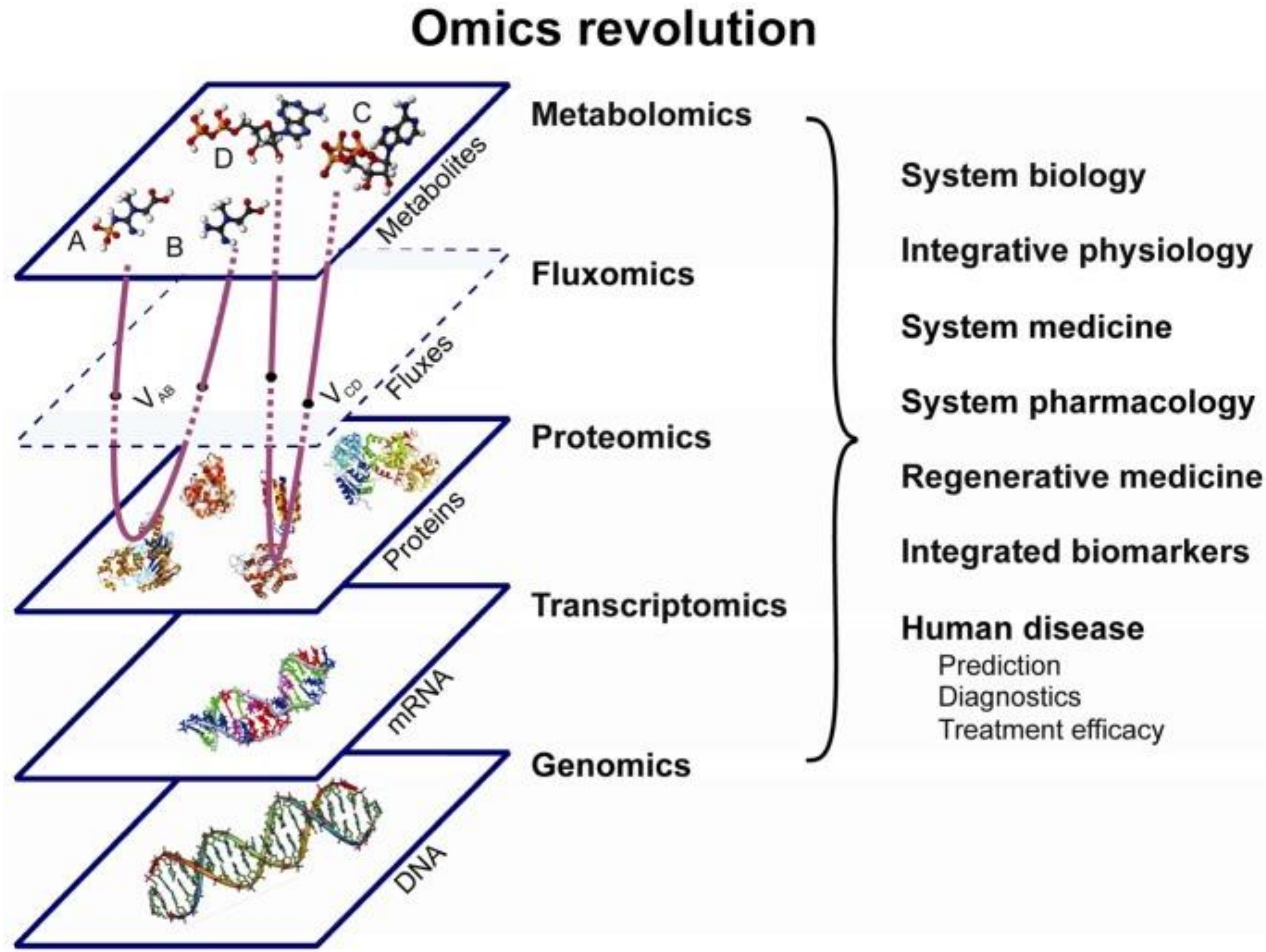
J Urol. 2019 Oct 25:101097JU0000000000000622. doi: 10.1097/JU.0000000000000622. [Epub ahead of print]

External Validation and Comparison of Prostate Cancer Risk Calculators Incorporating Multiparametric Magnetic Resonance Imaging for Prediction of Clinically Significant Prostate Cancer.

Saba K[1], Wettstein MS[1,2], Lieger L[1], Hötker AM[3], Donati OF[3], Moch H[4], Ankerst DP[5], Poyet C[1], Sulser T[1], Eberli D[1], Mortezavi A[1].

- MRI-PCa-RCs AUC **0.83 to 0.85** vs
  PCa-RCs AUC **0.69 to 0.74**

Positive example 3



Positive example 3

- Omics in diagnosing … / predicting … ??

- Because omics →
  clinical characteristics →
  outcome?

Examples

- Biomarkers
- Imaging
- Omics

- ML / AI

Success of ML / AI

Non-exhaustive list

Gaming

Natural Language Processing (Siri etc)

Fraud detection

Shoplifting

Object recognition (e.g. for driverless cars)

Facial recognition

Traffic predictions (e.g. Waze app)

Electrical load forecasting

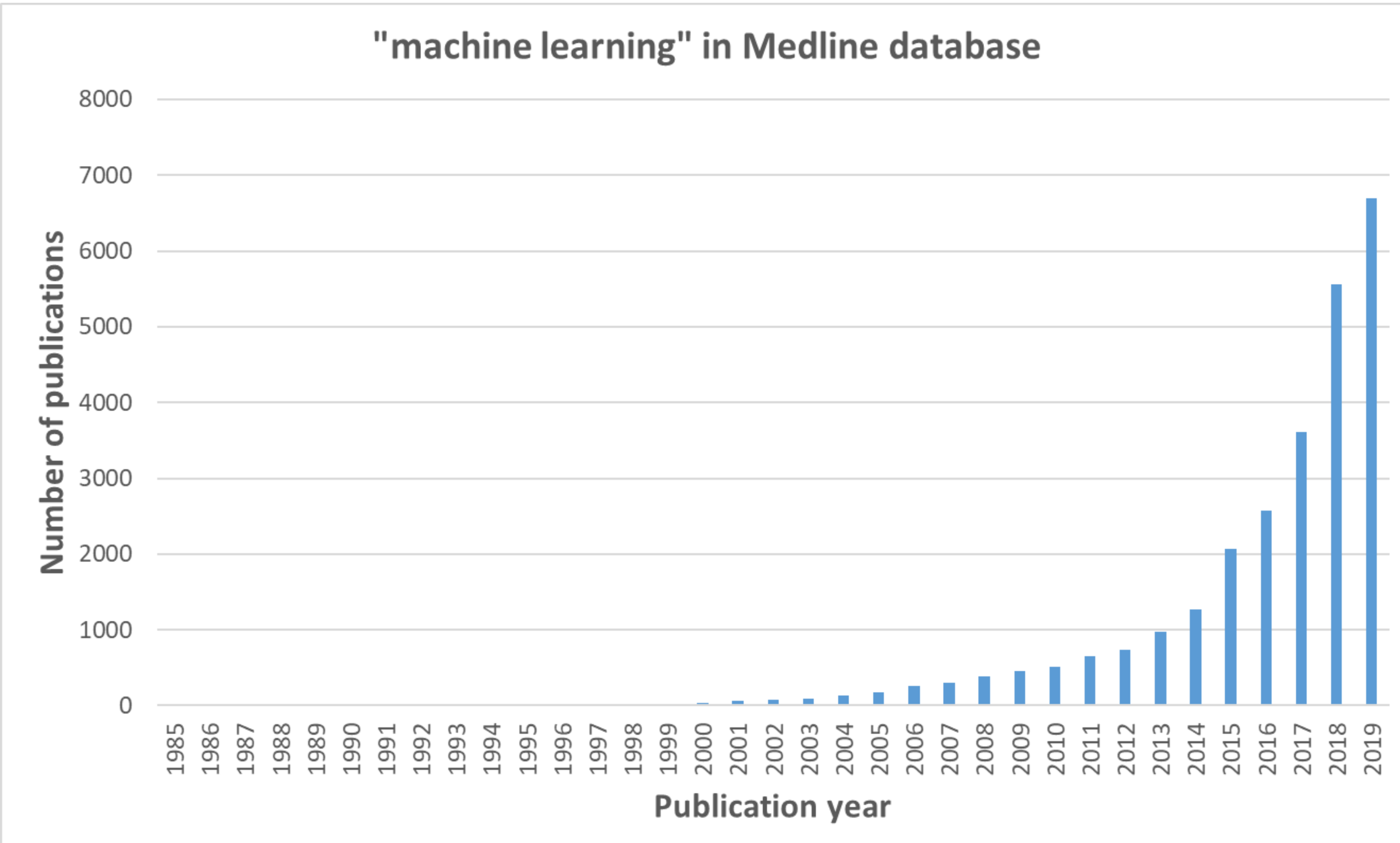
(Social) media and advertising (people you may know, movie suggestions,)

Spam filtering

Search engines (e.g. Google PageRank)

Handwriting recognition

Popularity skyrocketing



IBM Watson winning Jeopardy! (2011)

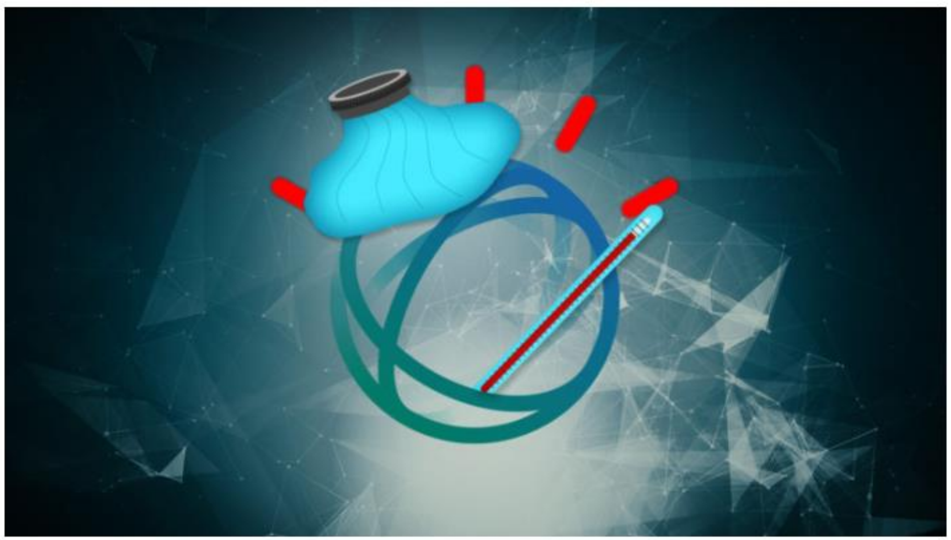
IBM Watson for oncology

EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By CASEY ROSS @caseymross *and* IKE SWETLITZ / JULY 25, 2018



ALEX HOGAN/STAT

Internal IBM documents show that its Watson supercomputer often spit out erroneous cancer treatment advice and that company medical specialists and customers identified "multiple examples of unsafe and incorrect treatment recommendations" as IBM was promoting the product to hospitals and physicians around the world.

https://bit.ly/2LxiWGj

Evidence

- Cochrane: ”We searched for RCTs and found 20 among ... papers”
- Dr Watson: “We searched 4 Million webpages in 1 second”

Five myths

1. Big Data will resolve the problems of small data
2. ML/AI is very different from classical modeling
3. Deep learning is relevant for all medical prediction problems
4. ML / AI is better than classical modeling for medical prediction problems
5. ML / AI leads to better generalizability

Myth 1: Big Data will resolve the problems of small data

Review Article | Published: 07 January 2019

High-performance medicine: the convergence of human and artificial intelligence

Eric J. Topol

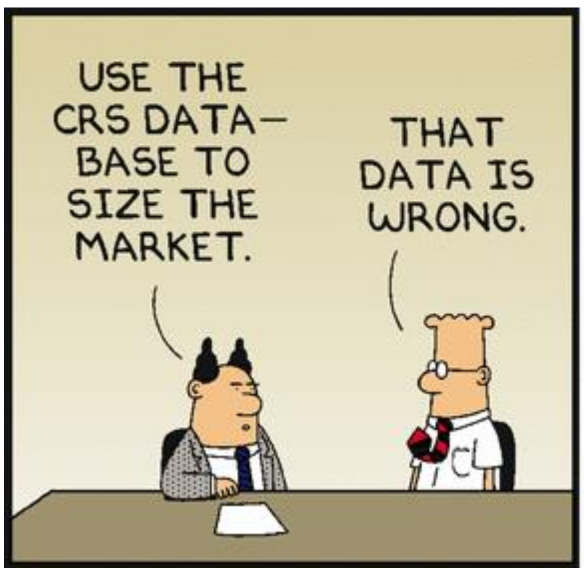
Nature Medicine **25**, 44–56 (2019) | Cite this article

Abstract

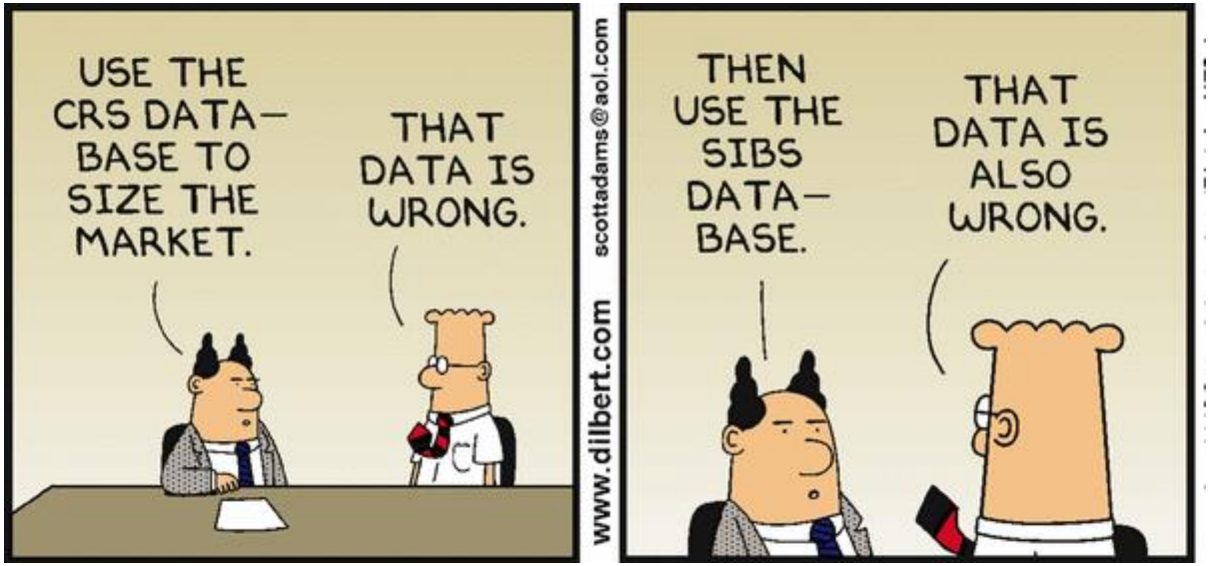
The use of artificial intelligence, and deep-learning in particular, has been enabled by the use of big data, along with markedly enhanced computing power and cloud storage, across all sectors.

In medicine, this is beginning to have an impact ...

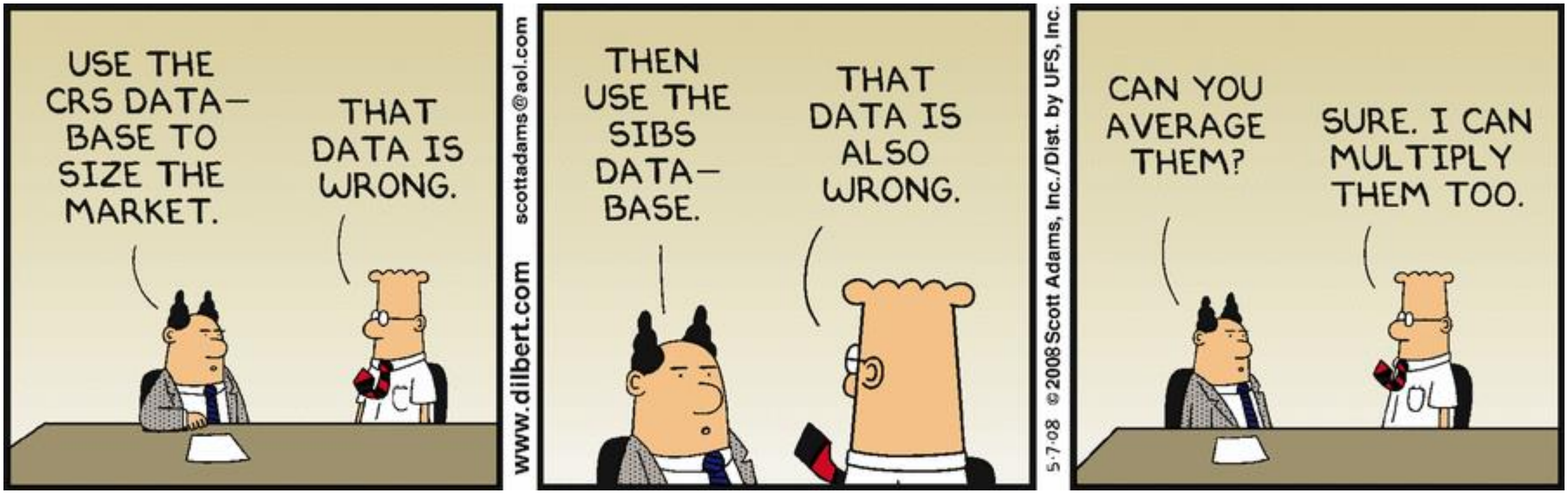
Do you have a clear research question?
Do you have data that help you answer the question?
What is the quality of the data?



Do you have a clear research question?
Do you have data that help you answer the question?
What is the quality of the data?



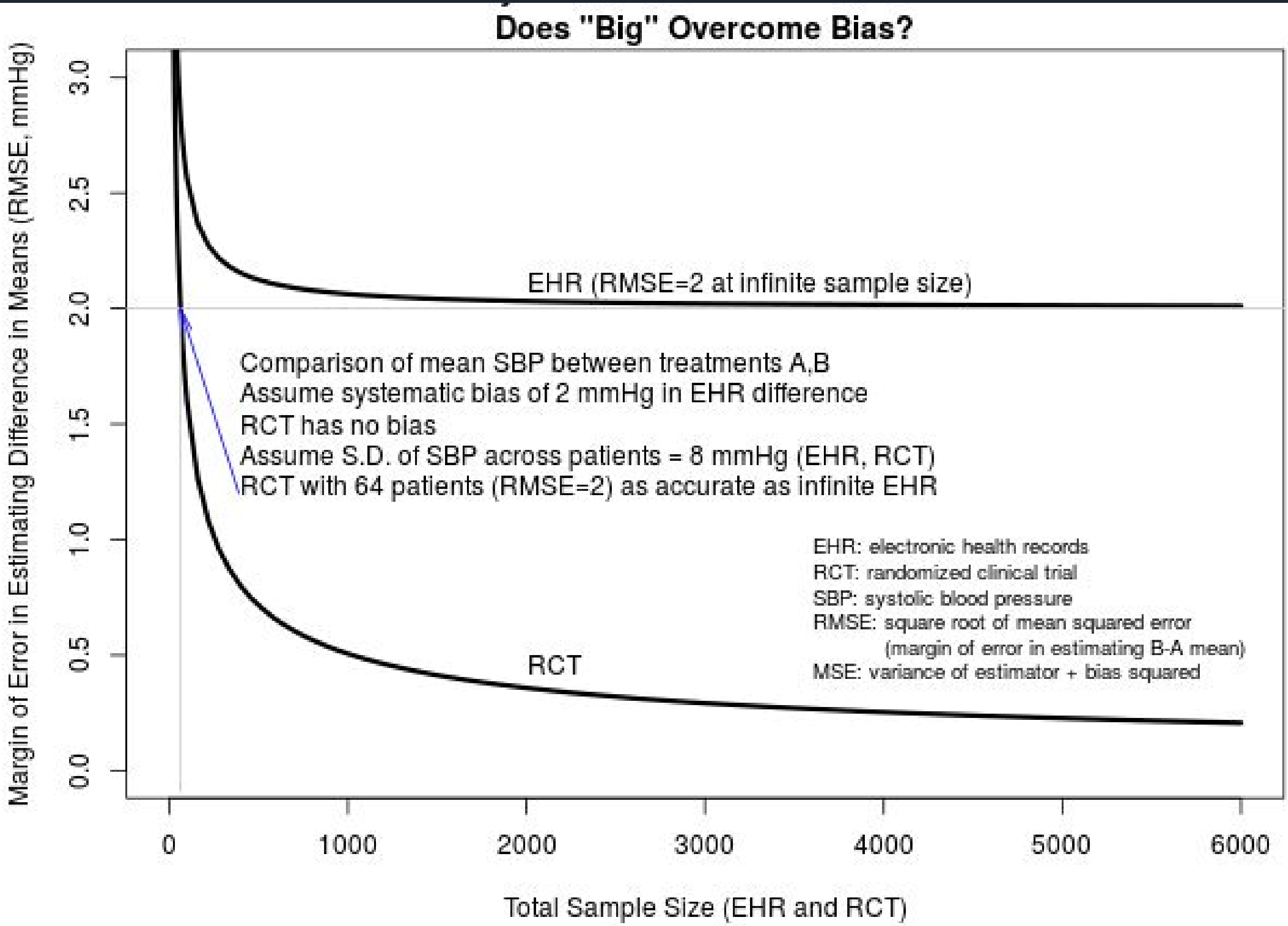
Do you have a clear research question?
Do you have data that help you answer the question?
What is the quality of the data?



Big Data, Big Errors

Frank Harrell @f2harrell · 23 jun. 2017

Example: RCT randomizing 64 patients as accurate as infinitely large EHR: fharrell.com/2017/06/ehrs-a... #StatThink #RCT #EHR **#BigData**



Myth 2: ML/AI is very different from classical modeling

“Everything is ML”

Scott H. Hawley
@drscotthawley

Replying to @JuliaHCox, @mikarv and @GSCollins

Logistic regression IS machine learning.

4:17 pm · 17 Feb 2019 · Twitter for iPhone

https://bit.ly/2lEVn33

Two cultures

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Breiman, Stat Sci, 2001, DOI:
10.1214/ss/1009213726

Traditional Statistics vs Machine Learning

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

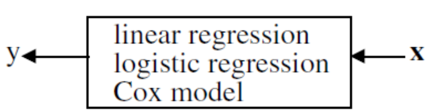
Leo Breiman

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

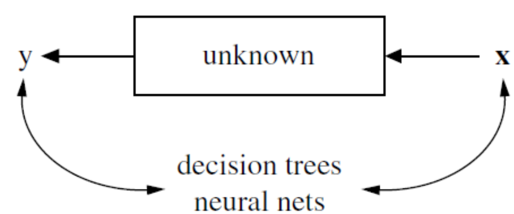
$$\text{response variables} = f(\text{predictor variables, random noise, parameters})$$

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

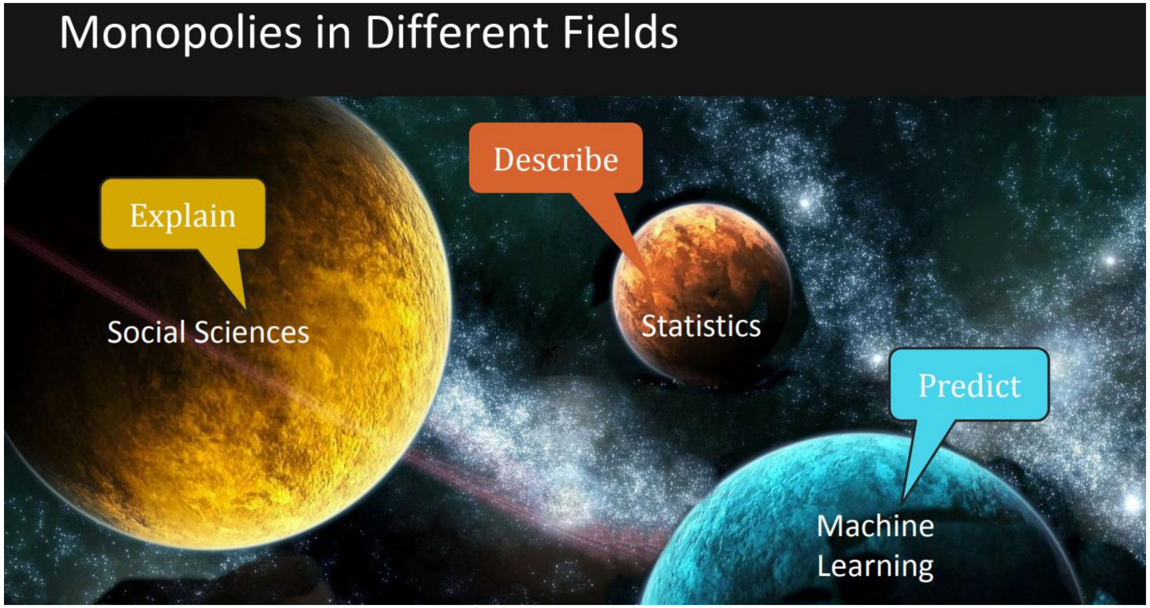


The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



Traditional Statistics vs Machine Learning



??

Statistics versus machine learning

Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.

Galit Shmueli. Keynote talk at 2019 ISBIS conference, Kuala Lumpur; taken from slideshare.net
Bzdok. Nature Methods 2018;15:233-4.

Example of exaggerating contrasts

PLOS | ONE

RESEARCH ARTICLE

Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease

Andrew J. Steele[1]*, Spiros C. Denaxas[2], Anoop D. Shah[2,3], Harry Hemingway[2], Nicholas M. Luscombe[1,4,5]

Table 1. The 27 expert-selected predictors used.

Category	Prognostic factors
Sociodemographic characteristics	Age, gender, most deprived quintile
CVD diagnosis and severity	SCAD subtype (stable angina, unstable angina, STEMI, NSTEMI, other CHD), PCI in last six months, CABG in last six months, previous/recurrent MI, use of nitrates
CVD risk factors	Smoking status (current, ex, never), hypertension, diabetes mellitus, total cholesterol, HDL
CVD comorbidities	Heart failure, peripheral arterial disease, atrial fibrillation, stroke
Non-CVD comorbidities	Chronic kidney disease, chronic obstructive pulmonary disease, cancer, chronic liver disease
Psychosocial characteristics	Depression at diagnosis, anxiety at diagnosis
Biomarkers	Heart rate, creatinine, white cell count, haemoglobin

Predicting mortality – the results

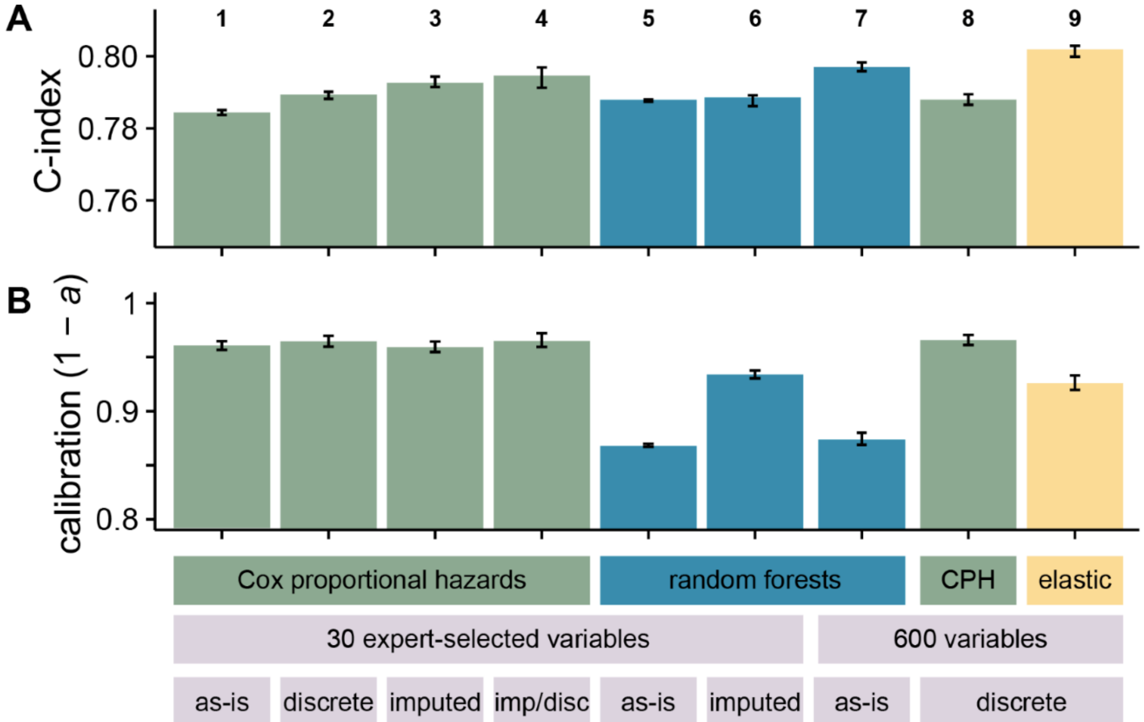


Fig 1. Overall discrimination and calibration performance for the different models and datasets used. (A) shows discrimination (C-

Elastic net, 586 ('600') variables: *c*=0.801

Traditional Cox, 27 ('30') expert-selected variables: *c*=0.793

PlosOne, 2018, DOI: 10.1371/journal.pone.0202344

Predicting mortality – the media

THE FRANCIS CRICK INSTITUTE

RESEARCH CAREERS AND STUDY PARTNERSHIPS WHAT'S ON NEWS ABOUT US

AI beats doctors at predicting heart disease deaths

4 SEPTEMBER 2018 HUMAN BIOLOGY HEALTH AND AGEING NEWS

AI NEWS RESEARCH —

Artificial Intelligence beats doctors at predicting heart disease deaths

BY SHACK15 - 5 SEPTEMBER, 2018

ScienceDaily®

Your source for the latest research news

SD Health Tech Enviro Society Quirky

Science News from research organization

AI beats doctors at predicting heart disease deaths

Date: September 4, 2018

PlosOne, 2018, DOI: 10.1371/journal.pone.0202344; https://bit.ly/2Q6H41R; https://bit.ly/2m3RLrn

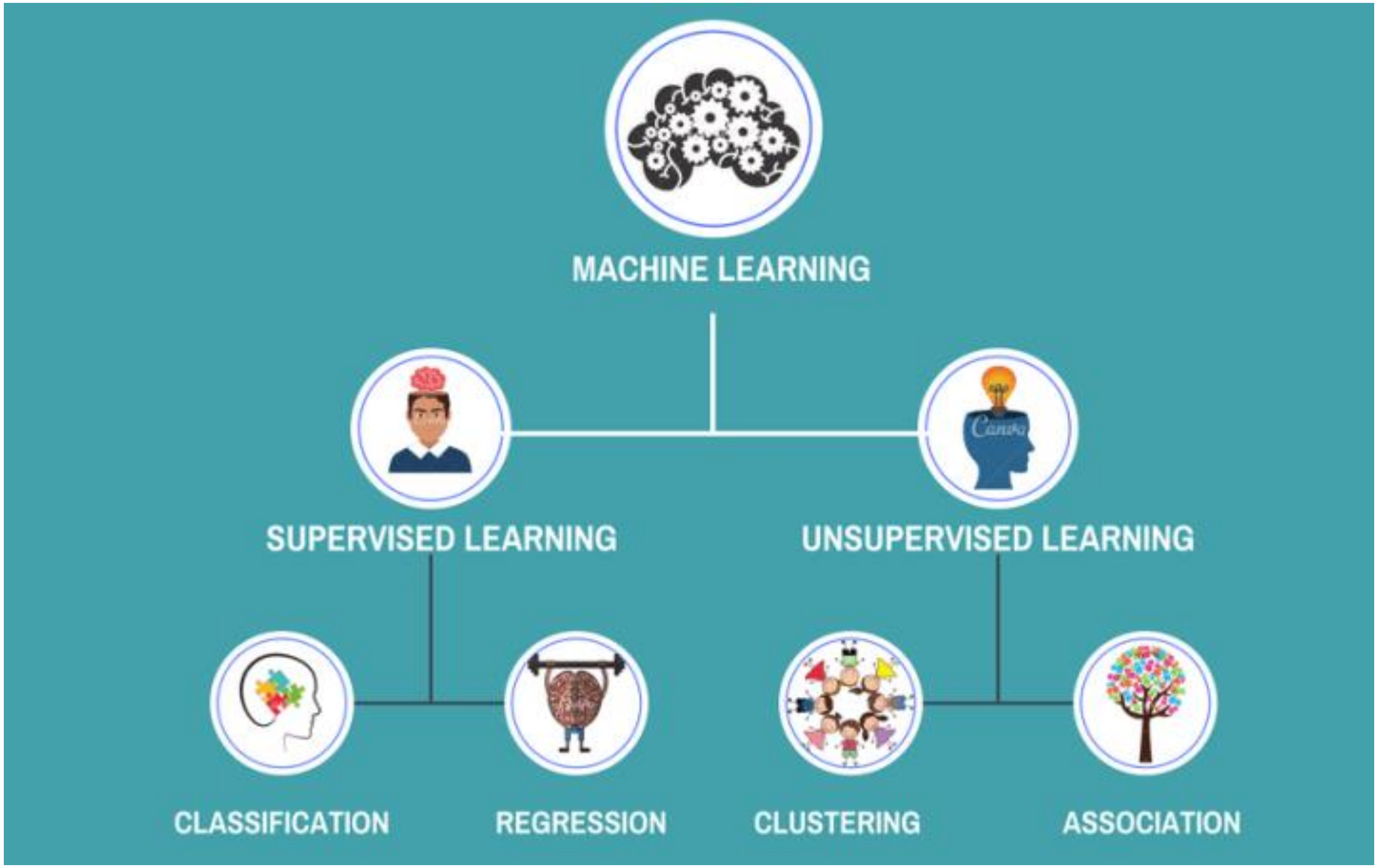
ML refers to a culture, not to methods

- Substantial **overlap methods** used by both cultures
- Substantial **overlap analysis goals**
- Attempts to separate the two frequently result in **disagreement**

Pragmatic approach:

"ML" refers to models roughly outside of the traditional regression types of analysis:
trees, SVMs, neural networks, boosting etc.

Machine learning: simple overview



Myth 3: Deep learning is relevant for all medical prediction


OBJECT
OBJECT
OBJECT
OBJECT
PERSON
PERSON
PERS
PERSON
PERSON

Example: retinal disease

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

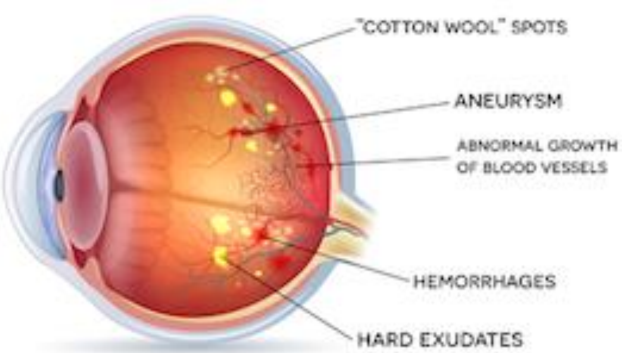
Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

Deep learning (= Neural network)

- 128,000 images
- Transfer learning (preinitialization)
- Sensitivity and specificity > .90
  - Estimated from training data

Diabetic retinopathy



Gulshan et al, JAMA, 2016, 10.1001/jama.2016.17216;
Picture retinopathy: https://bit.ly/2kB3X2w AS

Example: lymph node metastases

JAMA | Original Investigation

Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer

Babak Ehteshami Bejnordi, MS; Mitko Veta, PhD; Paul Johannes van Diest, MD, PhD; Bram van Ginneken, PhD; Nico Karssemeijer, PhD; Geert Litjens, PhD; Jeroen A. W. M. van der Laak, PhD; and the CAMELYON16 Consortium

Deep learning competition

But:

- 390 teams signed up, 23 submitted
- "Only" 270 images for training
- Test AUC range: 0.56 to 0.99

	Task 1: Metastasis Identification	Task 2: Metastases Classification
Codename[b]	FROC Score (95% CI)[c]	AUC (95% CI)[c]
HMS and MIT II	0.807 (0.732-0.889)	0.994 (0.983-0.999)
HMS and MGH III	0.760 (0.692-0.857)	0.976 (0.941-0.999
HMS and MGH I	0.596 (0.578-0.734)	0.964 (0.928-0.989)
VISILAB II	0.116 (0.063-0.177)	0.651 (0.549-0.742)
Anonymous I	0.097 (0.049-0.158)	0.628 (0.530-0.717)
Laboratoire d'Imagerie Biomédicale I	0.120 (0.079-0.182)	0.556 (0.434-0.654)

Bejnordi et al, JAMA, 2018, doi: 10.1001/jama.2017.14585.
See letter to the editor for a critical discussion: https://bit.ly/2kcYS0e

3. Deep learning is relevant for all medical prediction problems

NO: Deep learning excels in visual tasks

Myth 4: ML / AI is better than classical modeling for medical prediction

Reviewer #2, van Smeden submission 2019

used in this paper. Second, since the prediction performance of logistic regression models is often inferior to those of powerful machine learning algorithms such as random forest or boosting, focussing logistic regression models only can be boring. The detailed comments are given below.

Journal of Clinical Epidemiology 110 (2019) 12–22

REVIEW

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou[a], Jie Ma[b], Gary S. Collins[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

Poor methods and unclear reporting

What was done about missing data? 45% fully unclear, 100% poor or unclear

How were continuous predictors modeled? 20% unclear, 25% categorized

How were hyperparameters tuned? 66% unclear, 19% tuned with information

How was performance validated? 68% unclear or biased approach

Was accuracy of risk estimates checked? 79% not at all

Further observations:
- Prognosis: time horizon often ignored
- Patients matched on variables used a predictors
- 99% of patients excluded from modeling to obtain a balanced dataset
- First and last percentile of continuous predictors replaced with mean

Differences in discrimination

	Diff logit(AUC) (95% CI)	N
Overall		
– Any ML vs LR	0.25 (0.12;0.38)	282
– Tree vs LR	0.00 (–0.15;0.15)	42
– RF vs LR	0.33 (0.18;0.49)	59
– SVM vs LR	0.24 (0.10;0.39)	43
– ANN vs LR	0.47 (0.32;0.62)	52
– Other ML vs LR	0.22 (0.07;0.37)	86
Low risk of bias		
– Any ML vs LR	0.00 (–0.18;0.18)	145
– Tree vs LR	–0.34 (–0.65;–0.04)	16
– RF vs LR	0.06 (–0.15;0.26)	39
– SVM vs LR	0.03 (–0.20;0.26)	17
– ANN vs LR	–0.12 (–0.35;0.12)	27
– Other ML vs LR	–0.09 (–0.30;0.12)	46
High risk of bias		
– Any ML vs LR	0.34 (0.20;0.47)	137
– Tree vs LR	0.05 (–0.10;0.20)	26
– RF vs LR	0.41 (0.22;0.60)	20
– SVM vs LR	0.33 (0.19;0.48)	26
– ANN vs LR	0.71 (0.55;0.88)	25
– Other ML vs LR	0.31 (0.15;0.47)	40

Fig. 4. Differences in discriminative ability between LR and ML models, overall and according to risk of bias (n = 282 comparisons).

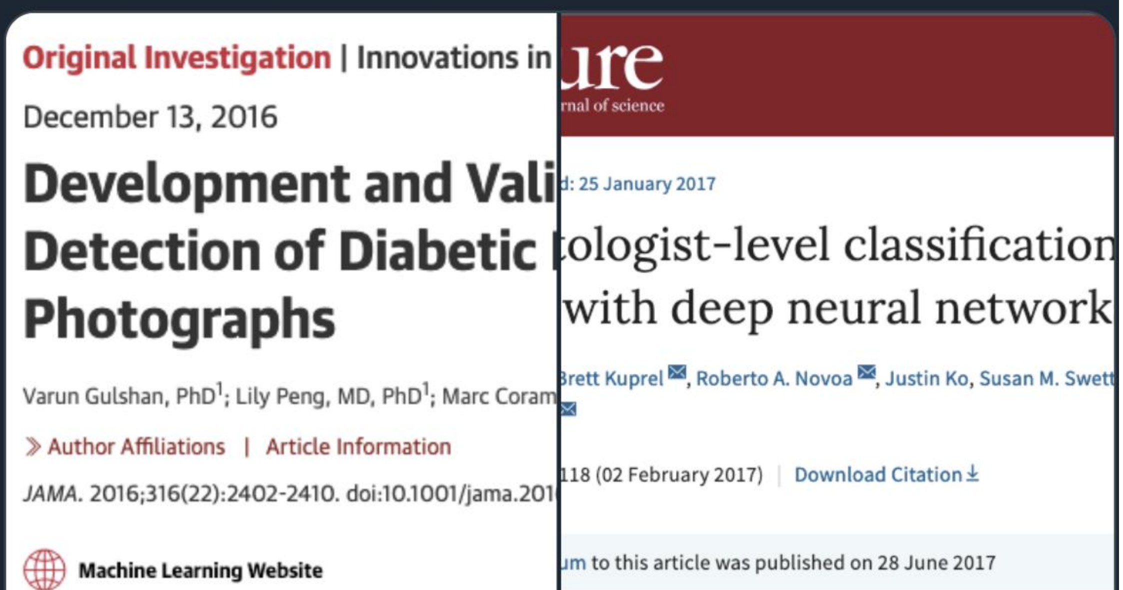
Christodoulou et al. Journal of Clinical Epidemiology, 2019,
doi: 10.1016/j.jclinepi.2019.02.004

Arjun (Raj) Manrai
@arjunmanrai

(Thread) The paper by Evangelia et al. in @JClinEpi on 'logistic regression = machine learning' for medicine has generated many reactions. This paper may be misinterpreted by #MachineLearning cynics and enthusiasts alike

Arjun (Raj) Manrai @arjunmanrai · 12 feb.

There are notable absences, such as many of the seminal contributions of deep learning to image analysis in medicine (e.g. Gulshan et al. JAMA 2016 and Esteva et al. Nature 2017). 7/n



Original Investigation | Innovations in

December 13, 2016

Development and Vali

Detection of Diabetic

Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram

» Author Affiliations | Article Information

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.201

Machine Learning Website

ure

rnal of science

d: 25 January 2017

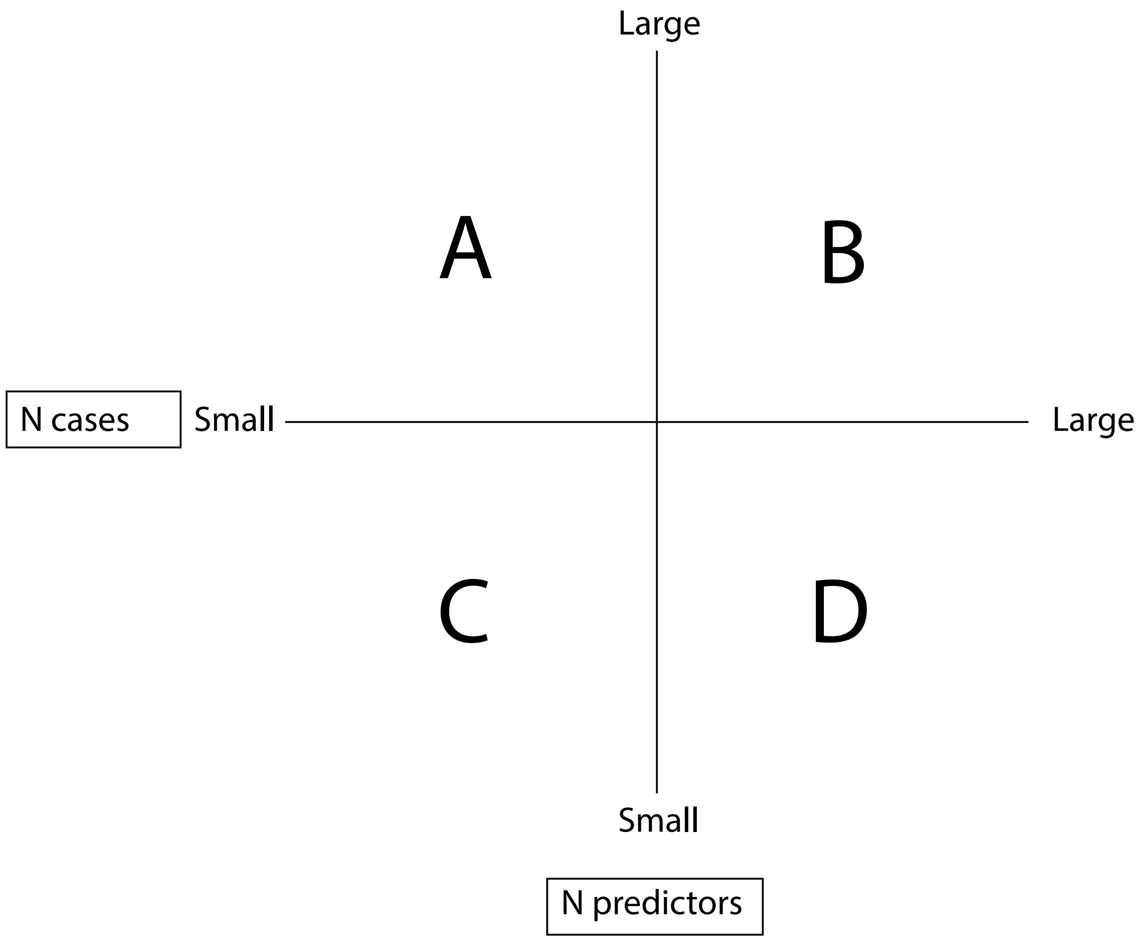
tologist-level classification with deep neural network

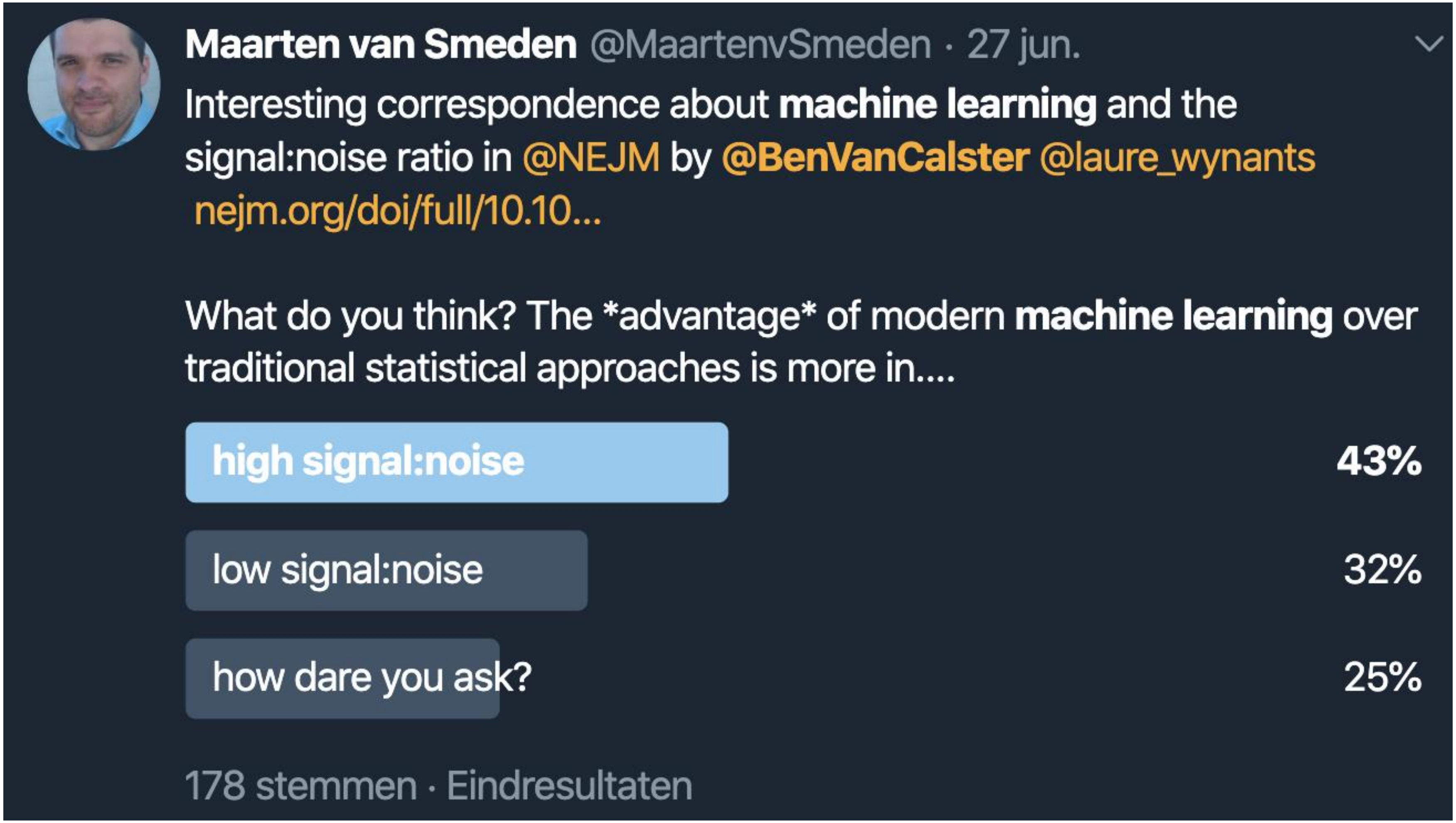
Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swett

118 (02 February 2017) | Download Citation

um to this article was published on 28 June 2017

Where is ML useful?





Maarten van Smeden @MaartenvSmeden · 27 jun.

Interesting correspondence about **machine learning** and the signal:noise ratio in @NEJM by **@BenVanCalster** @laure_wynants nejm.org/doi/full/10.10...

What do you think? The *advantage* of modern **machine learning** over traditional statistical approaches is more in....

high signal:noise	**43%**
low signal:noise	32%
how dare you ask?	25%

178 stemmen · Eindresultaten

The NEW ENGLAND JOURNAL of MEDICINE

Table 2. Key Questions to Ask When Deciding What Type of Model Is Necessary.
How complex is the prediction task?
Simple prediction tasks are defined as those that can be performed with high accuracy with a small number of predictor variables. For example, predicting the development of hyperkalemia might be possible from just a small set of variables, such as renal function, the use of potassium supplements, and receipt of certain medications.
Complex prediction tasks are defined as those that cannot be predicted accurately with a small number of predictor variables. For example, identification of abnormalities in a pathological slide requires evaluation of patterns that are not obvious over millions of pixels.
In general, simple prediction tasks can be performed with traditional models (e.g., logistic regression), and complex tasks require more complex models (e.g., neural networks).

Rajkomar et al. NEJM 2019;380:1347-58.

Myth 5: ML / AI leads to better generalizability

Calibration drift in regression and machine learning models for acute kidney injury FREE

Sharon E Davis, Thomas A Lasko, Guanhua Chen, Edward D Siew, Michael E Matheny

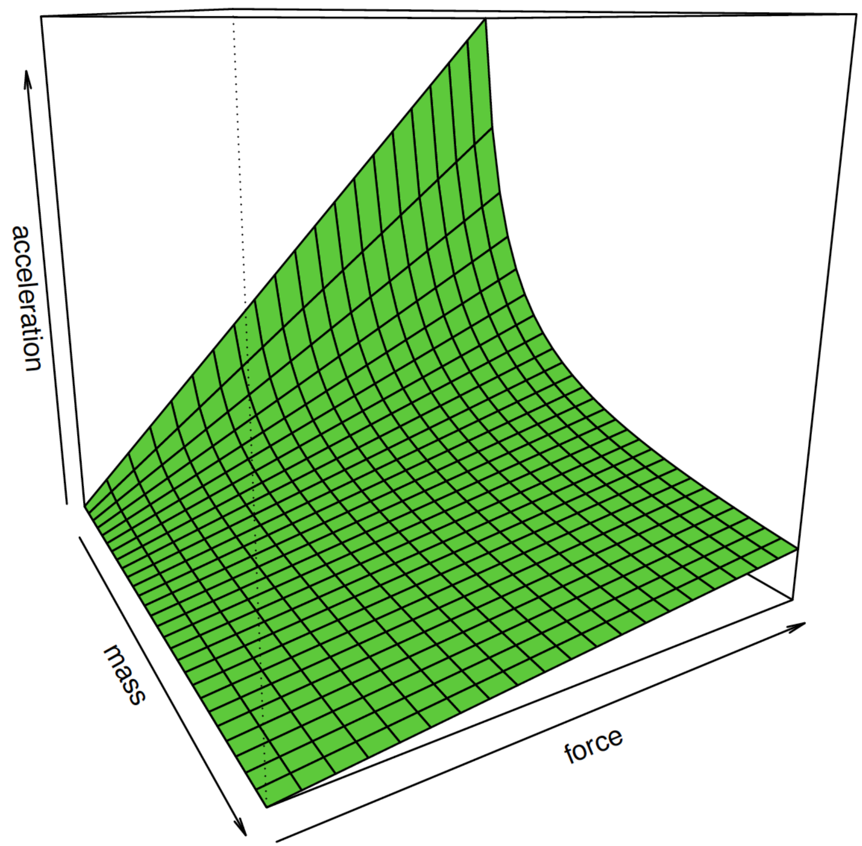
Journal of the American Medical Informatics Association, Volume 24, Issue 6, November 2017, Pages 1052–1061, https://doi.org/10.1093/jamia/ocx030

" … developed 7 parallel models for hospital-acquired acute kidney injury using common regression and machine learning methods, validating each over 9 subsequent years.":

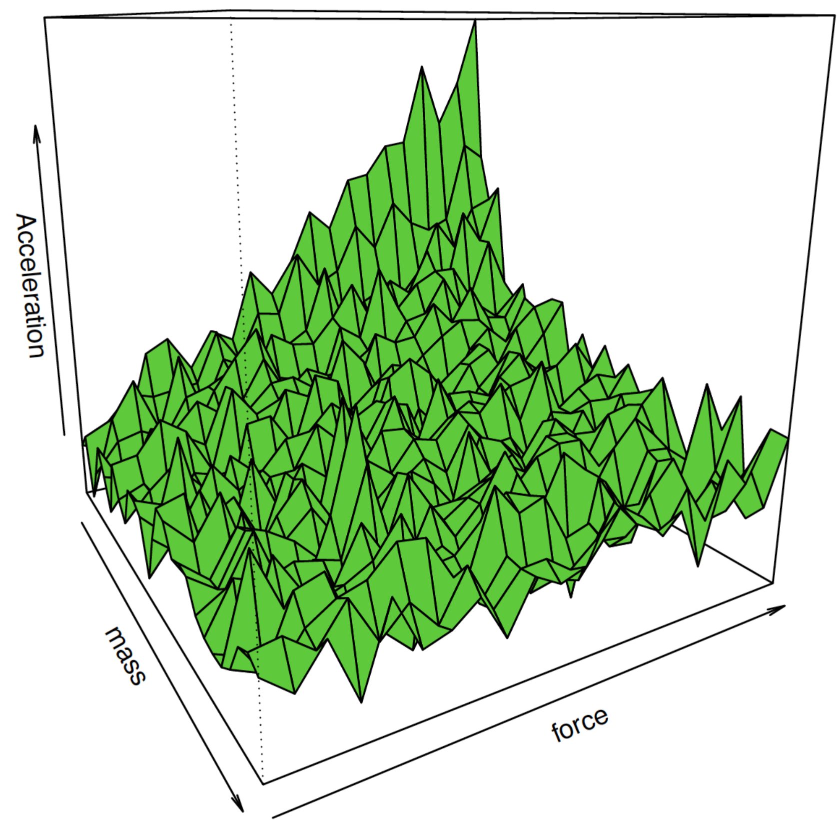
"Discrimination was maintained for all models. Calibration declined as all models increasingly overpredicted risk. **However, the random forest and neural network models maintained calibration** … "

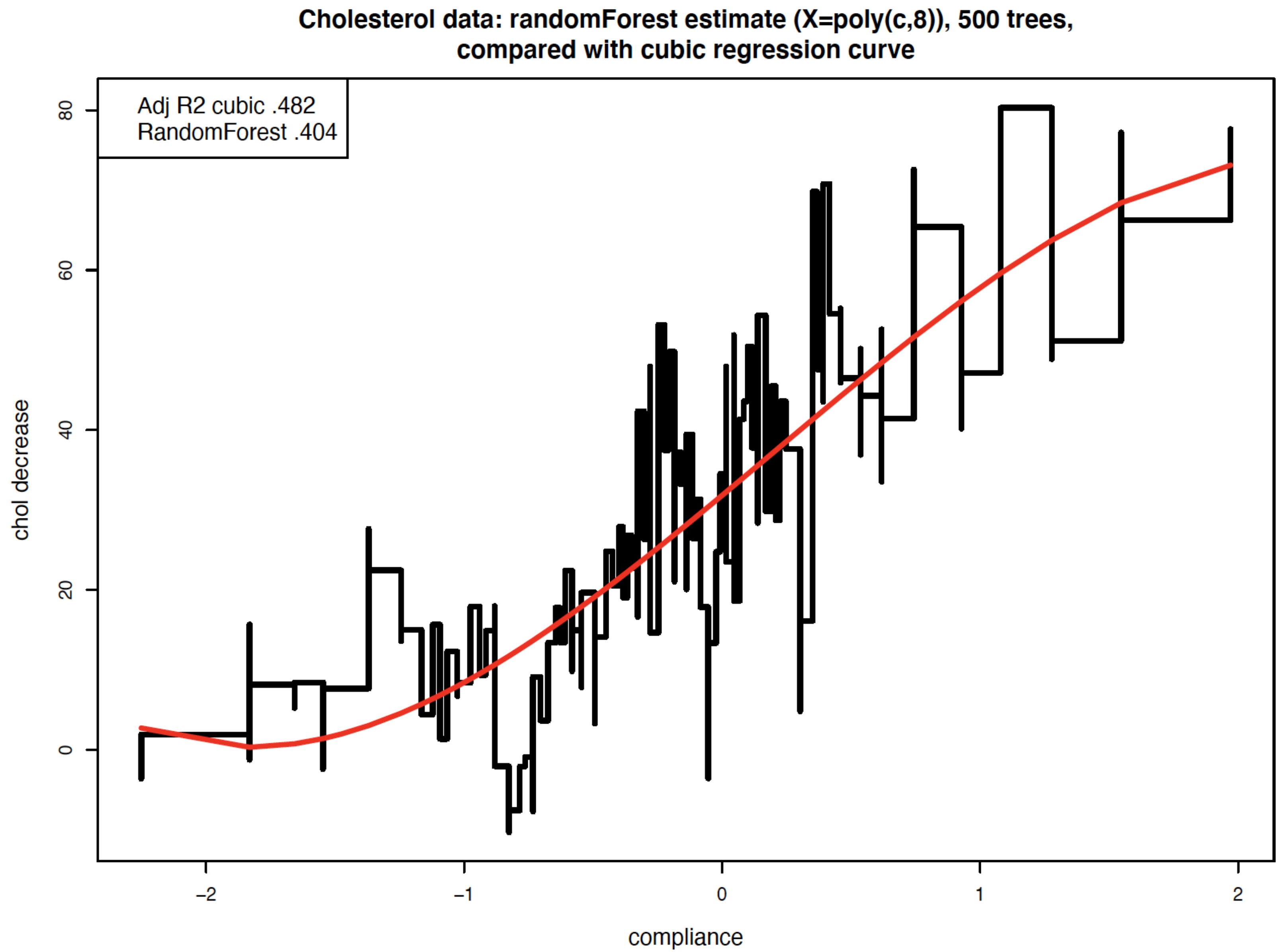
Efron talk Leiden

Newton's 2nd law: acceleration=force/mass



If Newton had done the experiment



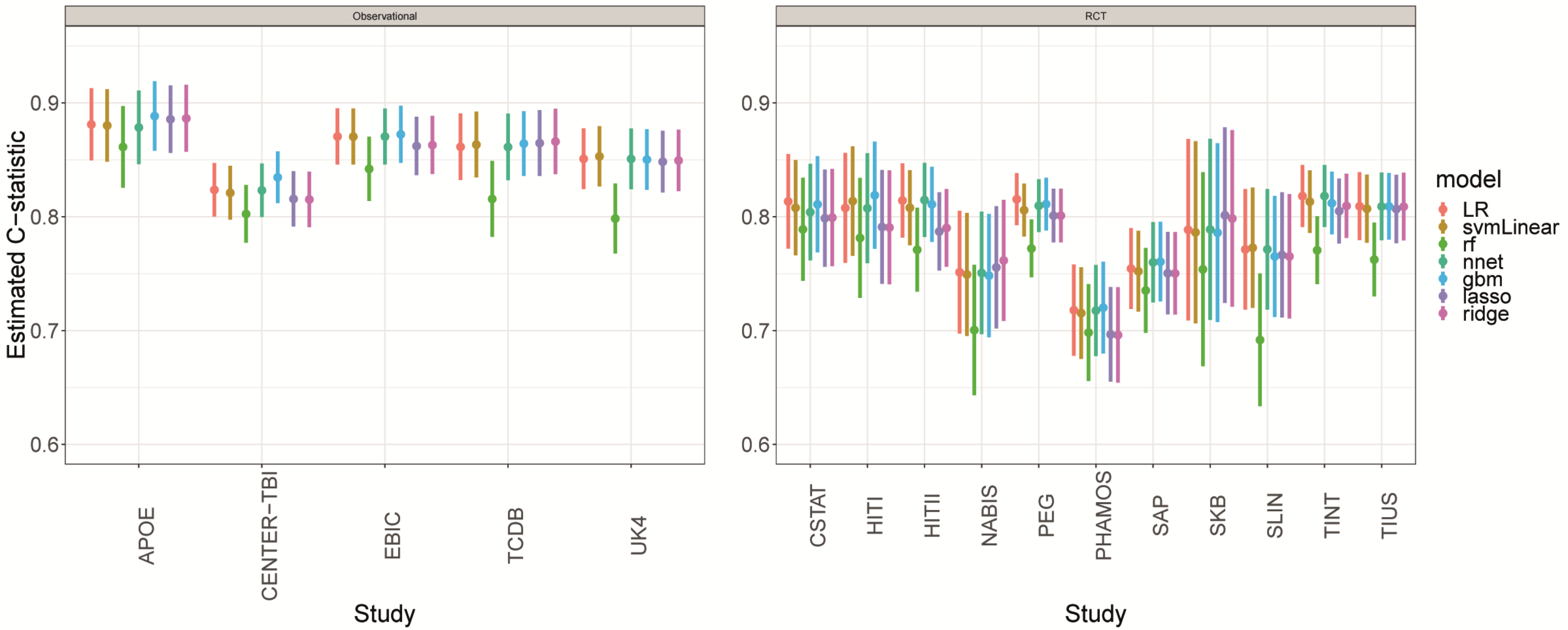

Cholesterol data: randomForest estimate (X=poly(c,8)), 500 trees,
compared with cubic regression curve
Adj R2 cubic .482
RandomForest .404
chol decrease
compliance
80
60
40
20
0
−2
−1
0
1
2

Empirical findings in TBI

– 16 cohorts: 5 observational, 11 RCTs
– Develop in 15, validate in 1
– 7 methods: LR; SVM; RF; nnet; gbm; LASSO; ridge

5 observational 11 RCTs

A



Variability between cohorts >> variability between methods

Prediction challenges

- There is no such thing as a validated prediction algorithm
- Algorithms are high maintenance
  - Developed models need **validation and updating** to remain useful over time and place
- Regulation and quality control of algorithms
  - What about proprietary algorithms?

Five myths

1. Big Data will resolve the problems of small data
   NO: Big Data, Big Errors
2. ML/AI is very different from classical modeling
   NO: a continuum, cultural differences
3. Deep learning is relevant for all medical prediction
   NO: Deep learning excels in visual tasks
4. ML / AI is better than classical modeling for prediction
   NO: some methods do harm (e.g. tree modeling)
5. ML / AI leads to better generalizability
   NO: any prediction model may suffer from poor generalizability