

Causality Uncertainty Quantification

two aspects of RDS

Aad van der Vaart
Mathematical Institute
Leiden University

RDS meeting, Amsterdam, October 20, 2016

Causality



Eating a lot of fish during pregnancy is linked to children who grow more quickly and have a later risk of becoming overweight

15 February 2016

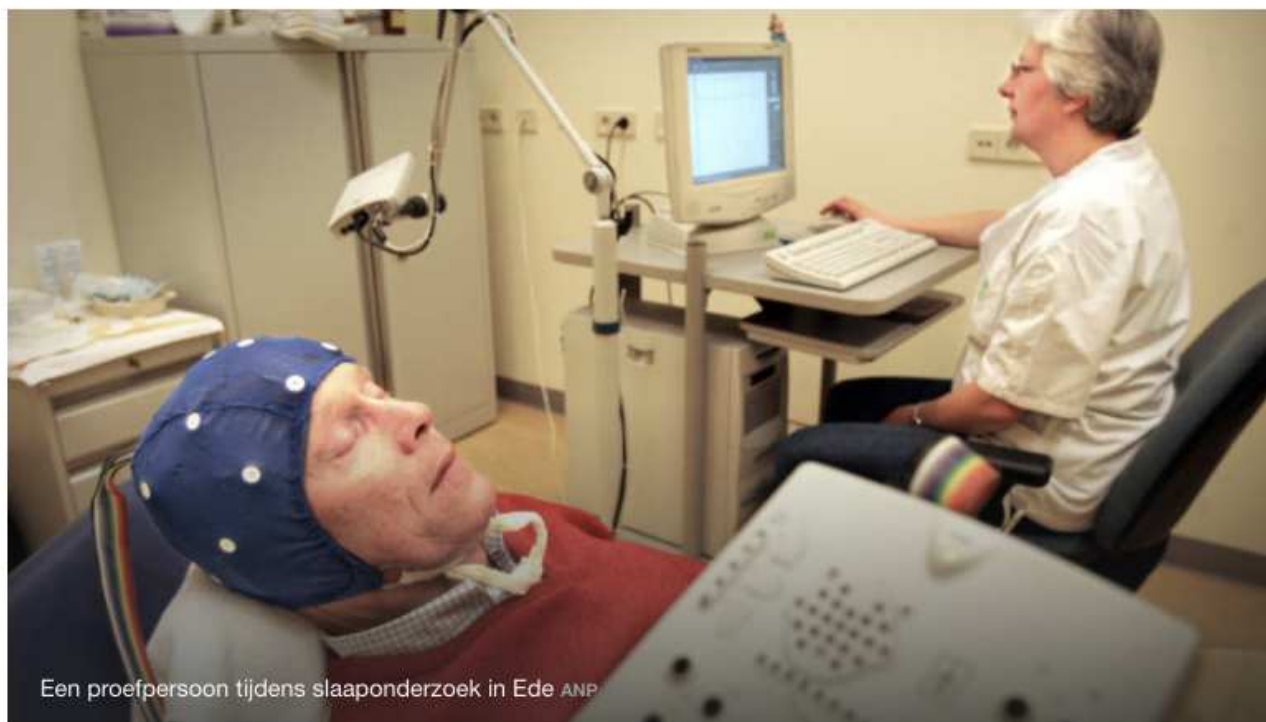
Pregnant women who eat fish more than three times a week give birth to babies who grow more quickly in the first years of life, and have a higher risk of being overweight as pre-schoolers. This is the result of international research in which many Dutch families also participated, and which was published in the scientific journal *JAMA Pediatrics*.

Fish is a known source of persistent organic pollutants (POPs) in our diet. POPs do not biodegrade or do so very poorly. These substances accumulate in the food chain and ultimately in humans. Some of these substances have endocrine-disrupting effects that may contribute to the development of overweight. To limit the intake of these




'Betere slaap kan depressie voorkomen'

MA 8 FEBRUARI, 19:25 AANGEPAST OP MA 8 FEBRUARI, 21:12 BINNENLAND



Een proefpersoon tijdens slaaponderzoek in Ede ANP

Onderzoekers zijn er achter waarom chronische slapeloosheid kan leiden tot



NIEUWS

BORSTSPARENDE THERAPIE BIJ VROEGE BORSTKANKER LEIDT TOT BETERE OVERLEVING

🕒 donderdag 10 december 2015

NEWS

Home

Video

World

UK

Business

Tech

Science

Magazine

Entertainment & Arts

Magazine

How maths can get you locked up

By Simon Maybin
BBC News

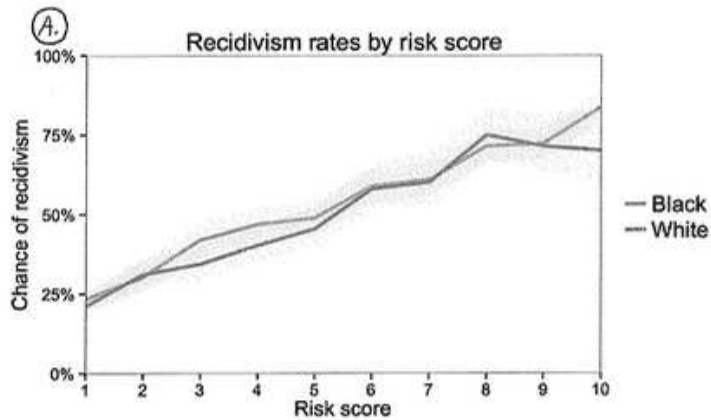
17 October 2016 | Magazine

Share

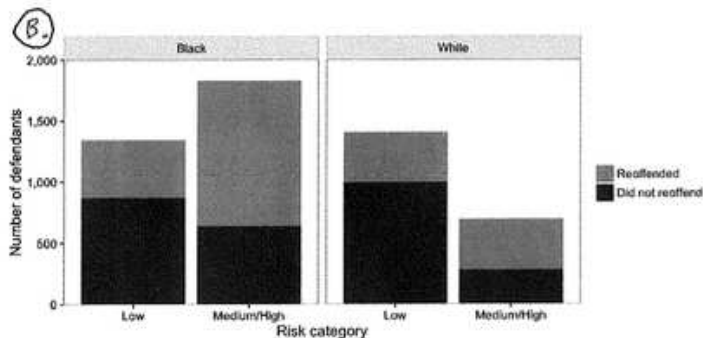


Criminals in the US can be given computer-generated "risk scores" that may affect their

In today's Magazine



Recidivism rate by risk score and race. White and black defendants with the same risk score are roughly equally likely to reoffend. The gray bands show 95 percent confidence intervals.



Distribution of defendants across risk categories by race. Black defendants reoffended at a higher rate than whites, and accordingly, a higher proportion of black defendants are deemed medium or high risk. As a result, blacks who do not reoffend are also more likely to be classified higher risk than whites who do not reoffend.

Risk of General Recidivism Cox Model (with Interaction Term)

Black	0.279*** (0.061)
Asian	-0.777 (0.502)
Hispanic	-0.064 (0.097)
Native American	-1.255 (1.001)
Other	0.014 (0.110)
High Score	1.284*** (0.084)
Medium Score	0.843*** (0.071)
Black:High	-0.190* (.100, p: 0.0574)
Asian:High	1.316* (0.768)
Hispanic:High	-0.119 (0.198)
Native American:High	1.956* (.083)
Other:High	0.415 (0.259)
Black:Medium	-0.173* (.091, p: 0.0578)
	0.986 (0.711)
	0.065 (0.164)
Native American:Medium	1.390 (1.120)
Other:Medium	-0.334 (0.232)
Observations	13,344
R2	0.072
Max. Possible R2	0.990
Log Likelihood	-30,280.410

From discussion of analysis by ProPublica of Compas Algorithm, Washington post, 2016, Corbett-Davies et al.

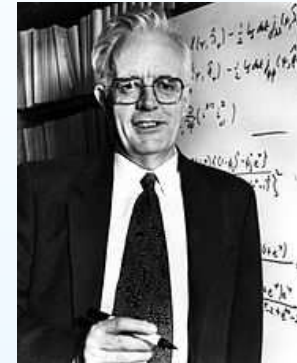
From analysis by ProPublica of Compas Algorithm by Northpointe

International Prize in Statistics Awarded to Sir David Cox for Survival Analysis Model Applied in Medicine, Science, and Engineering

ALEXANDRIA, VA (October 18, 2016) – Prominent British statistician Sir David Cox has been named the inaugural recipient of the International Prize in Statistics. Like the acclaimed Fields Medal, Abel Prize, Turing Award and Nobel Prize, the International Prize in Statistics is considered the highest honor in its field. It will be bestowed every other year to an individual or team for major achievements using statistics to advance science, technology and human welfare.

Cox is a giant in the field of statistics, but the International Prize in Statistics Foundation is recognizing him specifically for his 1972 paper in which he developed the proportional hazards model that today bears his name. The Cox Model is widely used in the analysis of survival data and enables researchers to more easily identify the risks of specific factors for mortality or other survival outcomes among groups of patients with disparate characteristics. From disease risk assessment and treatment evaluation to product liability, school dropout, reincarceration and AIDS surveillance systems, the Cox Model has been applied essentially in all fields of science, as well as in engineering.

“Professor Cox changed how we analyze and understand the effect of natural or human-induced risk factors on survival outcomes, paving the way for powerful scientific inquiry and discoveries that have impacted human health worldwide,” said Susan Ellenberg, chair of the International Prize in Statistics Foundation. “Use of the ‘Cox Model’ in the physical, medical, life, earth, social and other sciences, as well as engineering fields, has yielded more robust and detailed information that has helped researchers and policymakers address some of society’s most pressing challenges.”



Fitting a function $y = f(x)$ to data (x_i, y_i) goes under many names:

- regression analysis (statistics)
- input-output analysis (systems theory)
- (deep) learning (computer science)

In the news f is now called an *algorithm*.

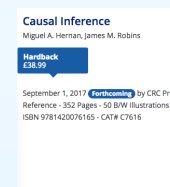
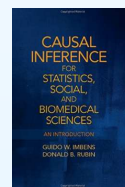
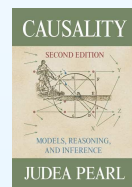
An algorithm may describe the data well
and can be useful for prediction (if not overfitted),
but does not give a causal model

A causal model

- explains (need theory?)
- predicts after **intervention**
- refers to counterfactuals

Finding a causal model typically needs a **designed experiment**

Observational data may be used together with **theoretical modelling** and **dedicated analysis techniques**



A crucial quantity is the **propensity score**:

Probability (person belongs to data **GIVEN** confounding variables).

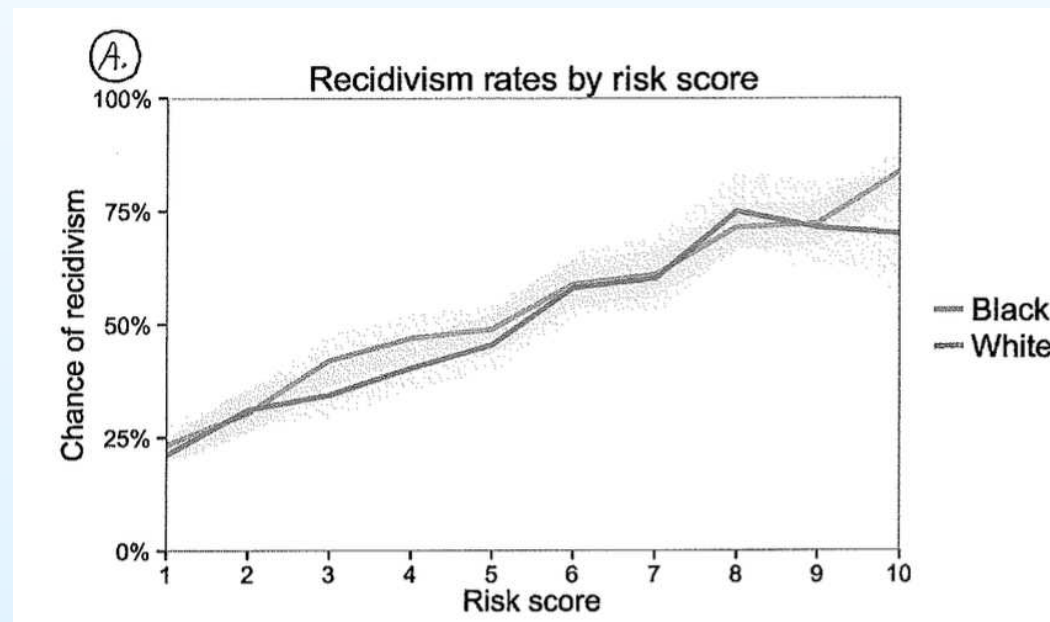
Big data allows to estimate this accurately?

Uncertainty Quantification

Most data-analysis gives **uncertain** conclusions

Probabilistic modelling gives error bands, confidence statements, significance levels, posterior distributions, generalization error, ...

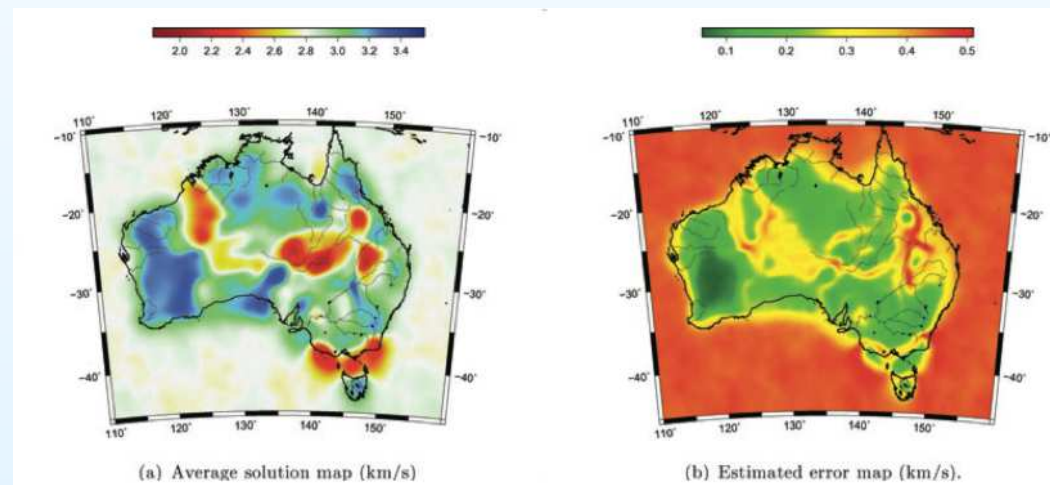
How to do this with big data?



Most data-analysis gives **uncertain** conclusions

Probabilistic modelling gives error bands, confidence statements, significance levels, posterior distributions, generalization error, ...

How to do this with big data?

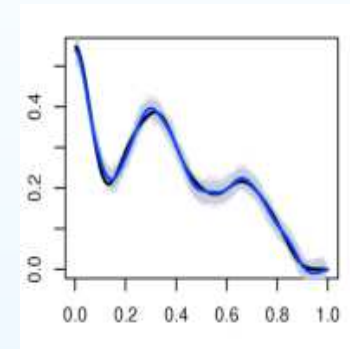
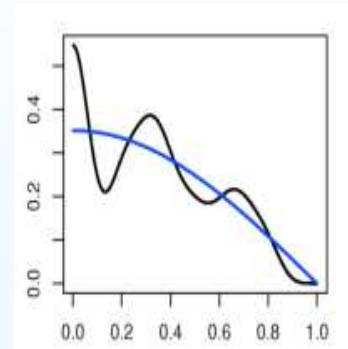
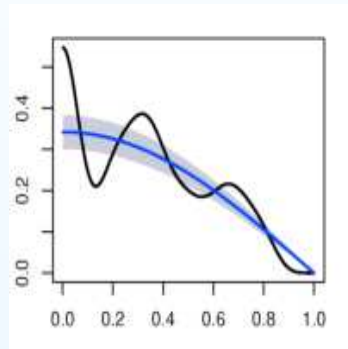
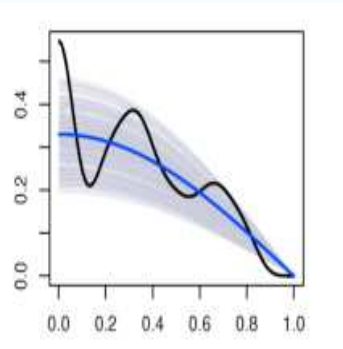


Travel times of surfaces waves: nonparametric Bayesian analysis in *earth science*. Left: posterior mean (a two-dimensional surface shown by colour coding); right: uncertainty quantification by the posterior spread. From Bodin and Sambridge, *Geophys. J. Int.* 178, 2009, 1411–1436.

VUmc claimt doorbraak in kankeronderzoek: druppel bloed volstaat



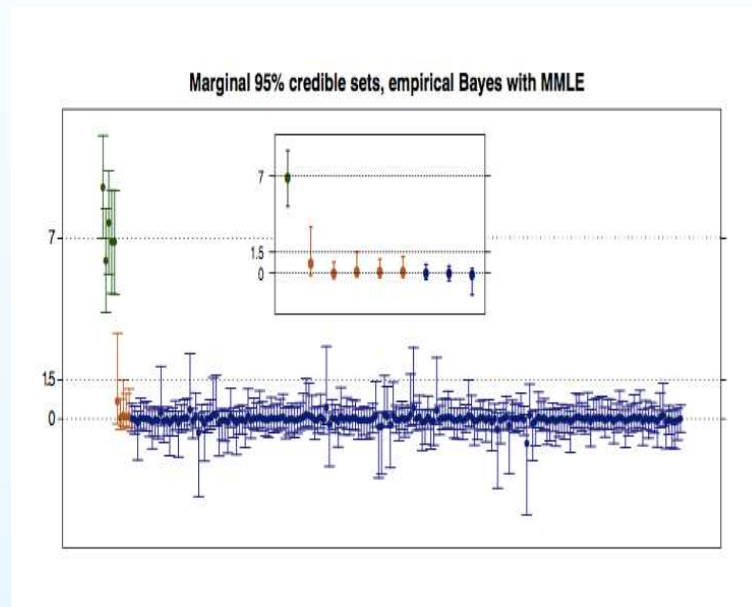
Natural algorithms may fail



Left to right: increasing amount of data

(Problems arise because the algorithm is trying the impossible)

When making **many discoveries simultaneously**, being **conservative** is wise



**Selection Bias,
Censoring,
Accuracy,
Modelling**

Responsible Data Science

genomics,
proteomics,
metabolomics,
systems biology,
medical signal analysis,
astronomy,
neuro science,
geo-science,
criminology,

....

