University of Zurich[UZH]

EBPI Epidemiology, Biostatistics and Prevention Institute

Big Data–Big Knowledge?

Torsten Hothorn

2015-03-06

The end of theory

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete (Chris Anderson, Wired Magazine 16.07)

Petabytes allow us to say: "Correlation is enough."

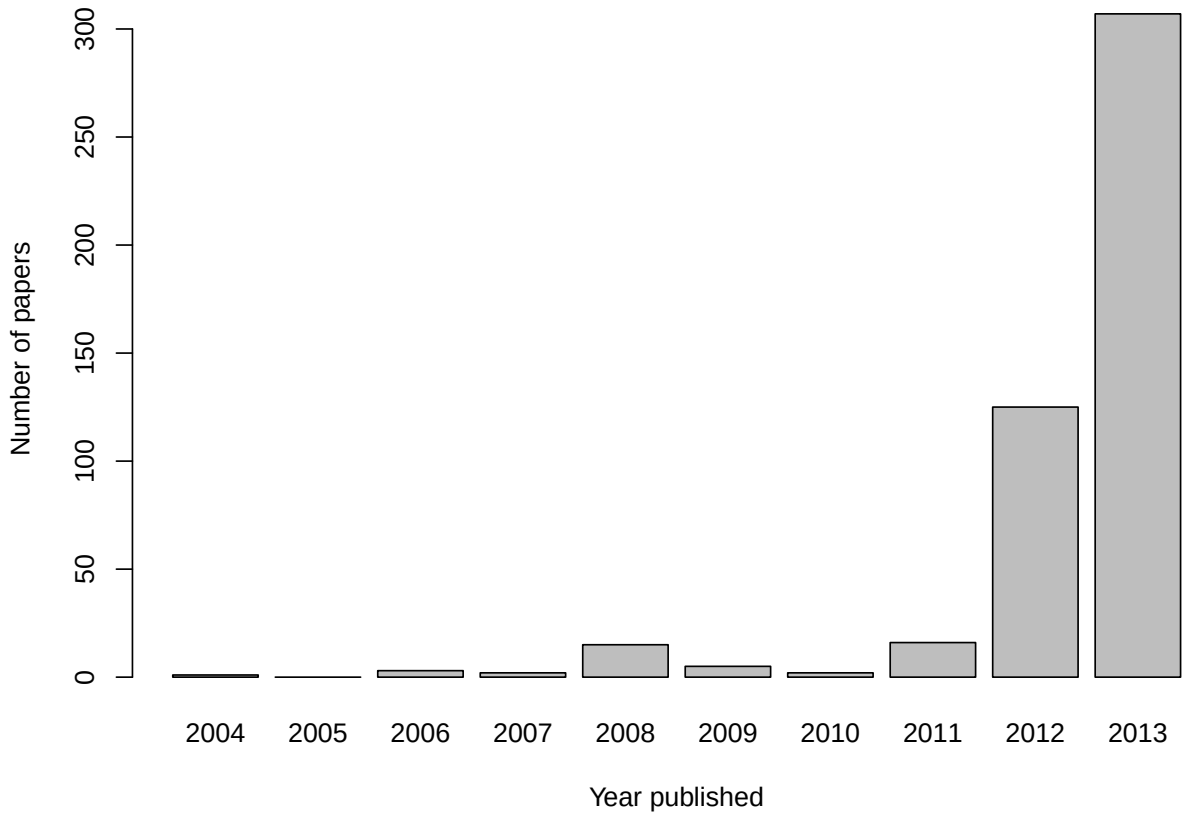
Big data science

- Big data
- Data science
-
-
-
-
-
-
-

Big data science

- Big data revolution
- Data science
-
-
-
-
-
-
-

“Big data” in journal titles



(Source: Web of Science)

Big data science

- Big data revolution
- Data science
- Predictive modelling
- Business intelligence
- Machine learning
- (parts of) Artificial intelligence; neural networks
- (parts of) Pattern recognition
- Knowledge discovery in data (KDD)
- ...

But what about ...

Statistics?

Interestingly, Andersons article starts with the famous quote of George Box

> *All models are wrong, but some are useful.*

Anderson uses 8 times the term "statistic*" in his 1336 words long article.

So, what is the connection between big data etc. and statistics now and what is the future of statistics?

> *Whoever wishes to foresee the future must consult the past. (Machiavelli)*

Statistics

Statistics is the science (the art?) of collecting, analysing, interpreting and communicating data.

The word "statistics" refers to "state"

- *statisticum* (lat) regarding the state
- *statista* (ital) statesman, politician

So, originally (and, to a large extend, still today), statistics is concerned with data describing the population, economy, administration etc. of a state. This is where the "bean counter" connotation comes from.

Statistics in academia

Scientists (working empirically) have

- a hypothesis/theory–and thus a (probabilistic) model
- an experiment–and thus data

Statistical methods

- use the data to estimate free parameters in the model
- assess their uncertainty
- and provide means to falsify a theory and/or to formulate a better theory

Estimation is performed by either optimisation (frequentists, this talk) or integration (Bayesians, not really today).

Models for conditional distributions

As an example, suppose a theory states that one or more explanatory variables X affect the distribution of a (so-called "response") variable Y.
We are interested if and how the conditional distribution of Y given $X = x$

$$(Y|X = x) \sim \mathbb{P}_{Y|X=x}$$

depends on x through a function $f(x)$:

$$\underbrace{\xi(Y|X = x) = f(x)}_{\text{statistical model}} := \underbrace{\arg\min_{f} \mathbb{E}_{Y,X} \rho(Y, f(X))}_{\text{minimisation problem}}$$

Statistical decision theory

Abraham Wald (1902-1950) established statistical decision theory; in a nutshell, a statistical model is defined by the minimal expected loss $\mathbb{E}_{Y,X}(\rho(Y, f(X)))$.

Statistical decision theory is the common foundation of statistics, machine learning, neural networks, pattern recognition, KDD, etc. But the language is different in comput[er,ational] science and statistics.

Same thing, different name

Machine learning	Statistics
supervised learning	regression

$$\xi(Y|X = x) = f(x)$$

target variable	response variable
Y	
attribute, feature	explanatory variable, covariate
X	
hypothesis	model, regression function
f	

Same thing, different name

instances, examples — samples, observations, realisations

$$(Y_i, X_i) \sim \mathbb{P}_{(Y,X)}, i = 1, \ldots, n$$

learning — estimation, fitting

$$\hat{f} = \arg\min_f \sum_{i=1}^{n} \rho(Y_i, f(X_i)) + \lambda \text{pen}(f)$$

classification — prediction

$$\hat{f}(x)$$

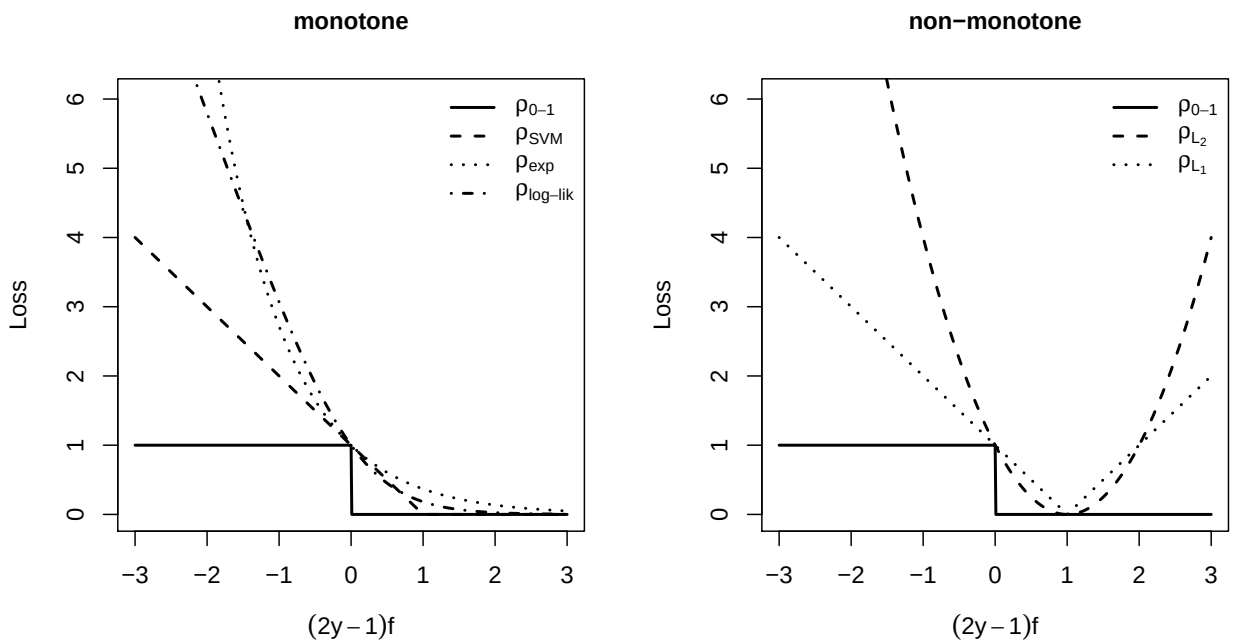
generalisation error — risk

$$\mathbb{E}_{Y,X} \rho(Y, \hat{f}(X))$$

So, what's the difference?

ρ (and thus ξ, the optimisation problem and optimiser) is often different causing much confusion. For binary Y, the loss ρ is

hinge loss, exponential loss log-density binomial distribution



So, what's the difference?

Traditionally, machine learners are more interested in black box classification, i.e. $\hat{f}(x)$ or even only $\hat{Y}$.
Statisticians focus on interpretation, i.e., look at

$$\hat{f}(x) = x^\top \hat{\beta} \quad \text{(linear model)}$$

or

$$\hat{f}(x) = \sum_{j=1}^{J} f_j(x) \quad \text{(additive model)}$$

Have a strong background in optimisation.
Have a strong background in modelling.

Some history

The median regression model

$$\rho(Y, f(X)) = |Y - f(X)| \Rightarrow f(x) = \text{Median}(Y|X = x)$$

was suggested by Boscovic and Laplace in the late 18th century.

The optimisation problem $\hat{f} = \arg\min_f \sum_{i=1}^{n} |Y_i - f(X_i)|$ is (was?) hard to solve.

Some history

The mean regression model

$$\rho(Y, f(X)) = |Y - f(X)|^2 \Rightarrow f(x) = \mathbb{E}(Y|X = x).$$

was suggested only a little later by Legendre and Gauß.

Why? Because $\hat{f} = \arg\min_f \sum_{i=1}^n |Y_i - f(X_i)|^2$ was relatively easy to compute with $f(x) = x^\top \beta$.

Some history

Carl-Friedrich Gauß (1777-1855), the great-grandfather of statistics, replaced a not-so-nice loss function with a nice one and suggested a fast optimisation algorithm (Gaussian elimination).
So he was actually a machine learner!

We see this pattern over and over again.

Same model, different optimiser

Machine learning	Statistics
artificially neural networks	additive/nonlinear logistic regression
support vector machines	generalised mixed/additive models
boosting	generalised additive models
decision trees	regression trees
random forests	random forests

random forests?

Working together

Machine Learning, 45, 5–32, 2001

Random Forests

LEO BREIMAN
Statistics Department, University of California, Berkeley, CA 94720

Editor: Robert E. Schapire

(cited 5189 times)

Talking to each other really helps.

What's different in big data?

Doug Laney (2001), a META Group/Gartner (!) employee:

> *Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimisation.*

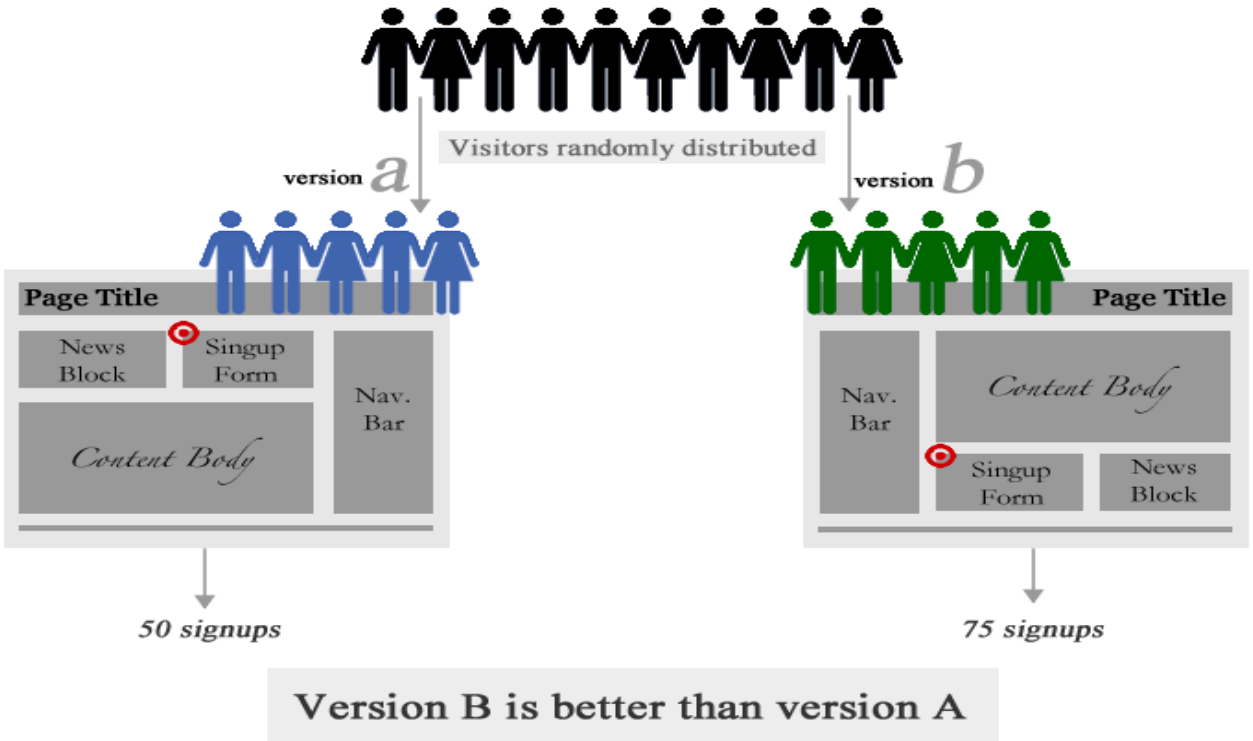
Wikipedia has

> *Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large data sets to reveal relationships, dependencies, and to perform predictions of outcomes and behaviours.*

In other words: Statistics for (large) data sets from multiple unplanned retrospective observational studies / sources.

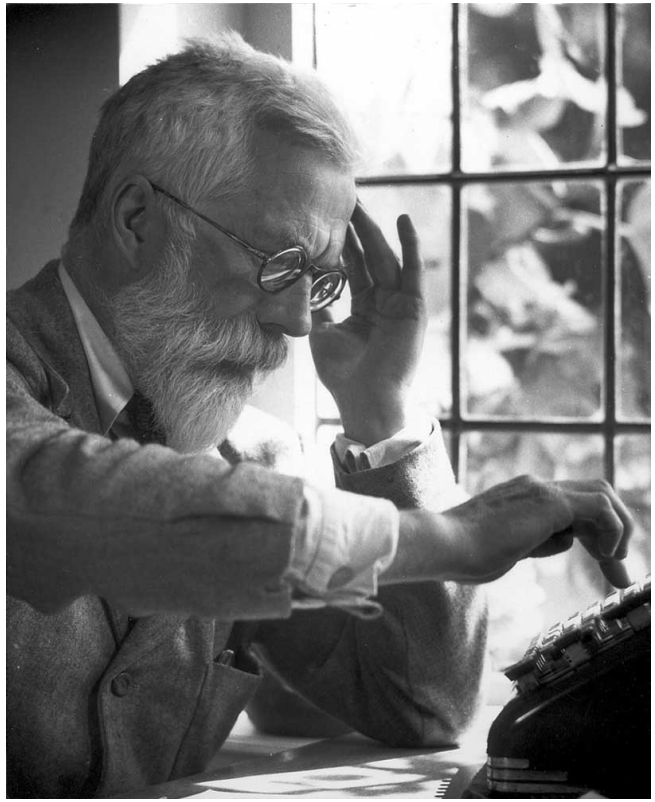
Not much!

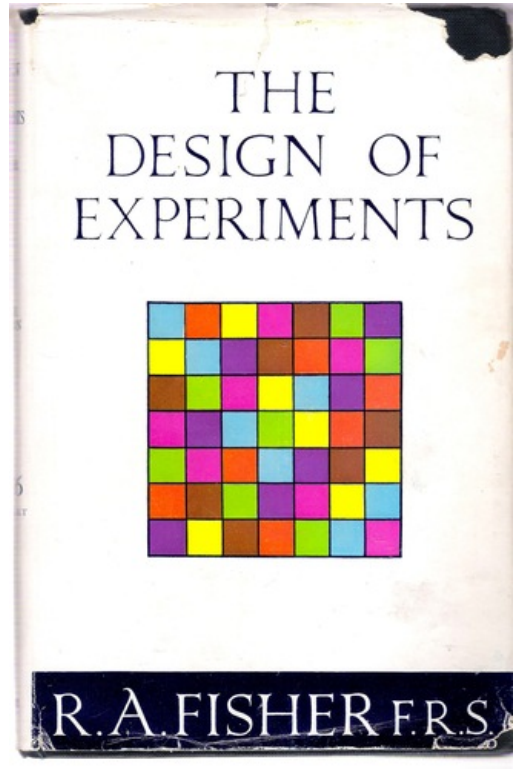
One of the most shattering examples of re-selling existing statistical technology under a new name is **A/B testing**.



(Source: `smashingmagazine.com`)

Not much!





This is a permutation test, most of the time applied incorrectly.
And with big data, the test will always be significant anyways.

What's the technical challenge?

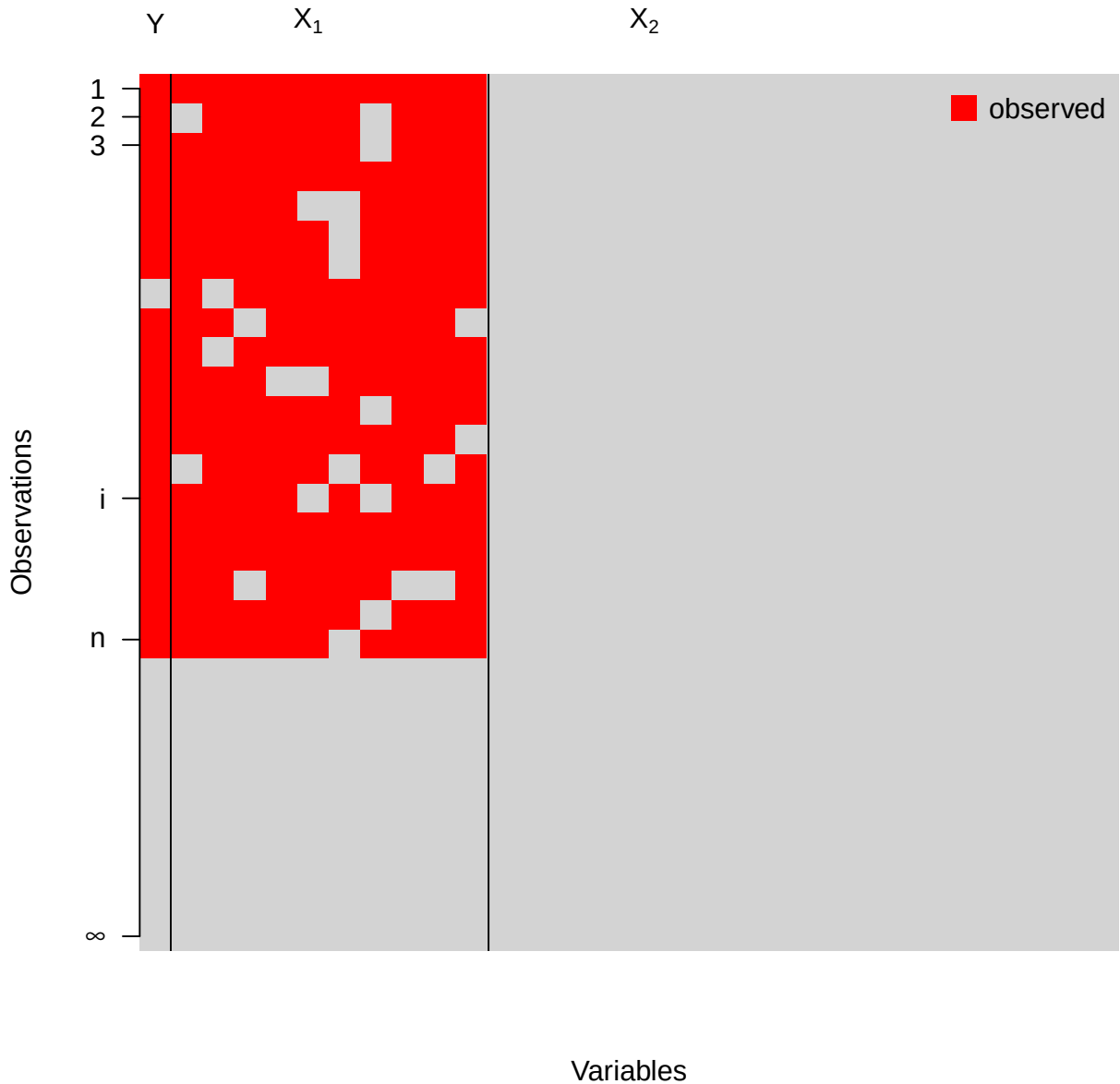
Problem:
RAM too small for data; can't load all the data to compute something.

This has been the rule with all data over the last 300 years, not the exception.

Solution:
(Finite) sampling and assessment of variability: go back to STA101.

Good news for statisticians: you can bootstrap from the true instead of the empirical distribution.

Bias & missings are much bigger problems



Opportunities

- We may have enough data to model the whole conditional distribution $\mathbb{P}_{Y|X=x}$ and not just some real-valued functional $\xi(\mathbb{P}_{Y|X=x})$ like the mean, for example by conditional transformation models (Hothorn, Kneib, Bühlmann, 2014).
- This allows probabilistic forecasts (Gneiting & Katzfuss, 2014).
- Funny: In biometry, Kaplan-Meier estimates and (to a certain extend) the Cox model for survival times always looked at the whole conditional distribution!

Opportunities

- Big data instead of meta-analysis: The PRO-ACT data base has time-course information of more than 8500 ALS patients from multiple clinical trials. Use this pooled data to model ALS disease progression (Hothorn & Jung, 2014) instead of somehow merging multiple analyses.
- Merge different data sources (police records, road information systems, weather records, satellite images, browsing surveys) to model spatial and temporal distribution of wildlife-vehicle collisions (Hothorn et al, 2012).

Can we learn something?

- Statisticians are rather hesitant to new models and techniques because partially educated and employed for policing science (sample size? power? analysis plan? significance?).
- In the 1990ies, statisticians lost track of microbiology; now there is bioinformatics.
- However, it seems statisticians are still needed. Think (lack of) reproducibility (Lancet Jan 11 series "Increasing value, reducing waste"; Hothorn & Leisch, 2011).
- Is $p < .05$ necessary and sufficient for reproducibility?
- Statistics needs better marketing. The trademark of my own field, biometry, was hijacked by people scanning fingerprints and irises.

References

Hothorn & Leisch (2011)
http://dx.doi.org/10.1093/bib/bbq084

Hothorn, Brandl & Müller (2012)
http://dx.doi.org/10.1371/journal.pone.0029510

Gneiting & Katzfuss (2014) http://dx.doi.org/10.1146/
annurev-statistics-062713-085831

Hothorn & Jung (2014)
http://dx.doi.org/10.3109/21678421.2014.893361

Hothorn, Kneib & Bühlmann (2014)
http://dx.doi.org/10.1111/rssb.12017