university of groningen

NWO
Netherlands Organisation for Scientific Research

FISI conference
Singapore 2022

Data-driven science and undercover theory

The case of automated disease classification

⋆

Jan-Willem Romeijn
University of Groningen

Bacon’s epistemo-entomology

The theme of this talk is nicely captured in the following quote from Bacon’s Novum Organon:

> *[Scientists] have been either empirics or dogmatical. The former, like ants, only heap up and use their store, the latter like spiders spin out their own webs. The bee, a mean between both, extracts matter from the flowers of the garden and the field, but works and fashions it by its own efforts.*
>
> Francis Bacon, The New Organon [Book One], 1620.

Data science seems like the work of ants. It focuses on collecting data and “letting those data speak for themselves”.

Bees, not ants

As most data scientists will tell you, this popular understanding of data science methods is mistaken.

The general inevitability of inductive bias is well-known. But its precise nature is often hard to identify in concrete cases.

The logic of induction

My take on data science relies on the logical view on inductive inference, as captured in this quote from Colin Howson.

> *We have solved Hume's Problem in about the only way it could be solved, by divorcing the justification for inductive reasoning from a justification of its consequences. Inductive reasoning is justified to the extent that it is sound, given appropriate premises.*
>
> Colin Howson, Hume's Problem, 2000.

In this talk I will illustrate how this idea can be used to great effect when analyzing data science methods.

Plan of talk

1. Machine learning in science
2. Data-driven psychopathology
3. Reliability and accountability
4. Inductive logic
5. Uncovering inductive assumptions
6. The epistemology of data science

1 Machine learning in science

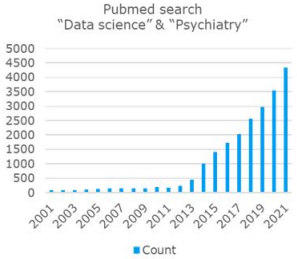
There are many examples of data-driven methods in the sciences, for prediction and automated model construction:

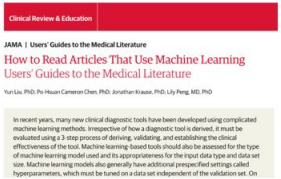
- Psychiatrists use hierarchical clustering to come up with subtypes of heterogeneous diseases like depression.
- Biomedical researchers employ methods of automated causal discovery to identify mechanisms of gene expression in the cell.
- Linguists employ latent Dirichlet analysis to disclose a corpus of texts and allocate them to thematic clusters.

In these cases the impact of theoretical starting points is difficult to trace. Does that matter?

Rapid uptake

The sciences see a rapid uptake of new data-scientific tools.




Clinical Review & Education

JAMA | Users' Guides to the Medical Literature

How to Read Articles That Use Machine Learning
Users' Guides to the Medical Literature

Yun Liu, PhD; Po-Hsuan Cameron Chen, PhD; Jonathan Krause, PhD; Lily Peng, MD, PhD

In recent years, many new clinical diagnostic tools have been developed using complicated machine learning methods. Irrespective of how a diagnostic tool is derived, it must be evaluated using a 3-step process of deriving, validating, and establishing the clinical effectiveness of the tool. Machine learning–based tools should also be assessed for the type of machine learning model used and its appropriateness for the input data type and data set size. Machine learning models also generally have additional prespecified settings called hyperparameters, which must be tuned on a data set independent of the validation set. On

This uptake goes hand-in-hand with increased interest in methodological guidelines for these methods.

Popular reception

The public perception of science is heavily impacted by new data science tools. Precision and personalized evidence-based medicine seems very promising.

At the same time the concerns over the accountability and intelligibility of evidence-based decision making are growing.

Methodological concerns

The different nature of the new methods puts the continuity with existing theory under pressure.




And the new methods are “black-boxed”. This makes it hard to hold machine-learning research accountable and motivate policy with it.

2 Data-driven psychopathology

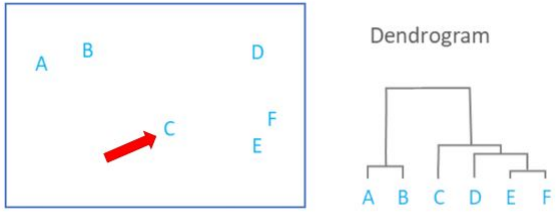
Psychiatric classification and sub-typing is assisted by automated clustering methods in the clinic.

Patient C.

Do the methods identify patient groups that are distinct for the purpose of prediction and intervention?

How does it work?

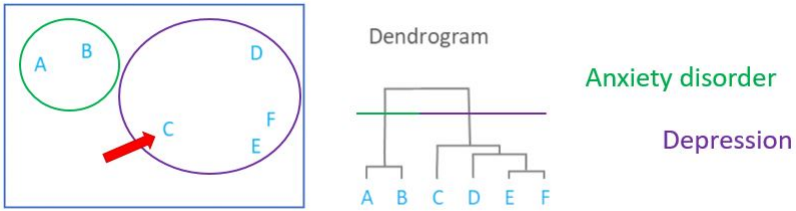
Here is a quick tour past the hierarchical clustering techniques that automated classification is based on.



The starting point is a space of patient characteristics and a tree structure expressing the proximity of the individuals in it.

Generating a classification

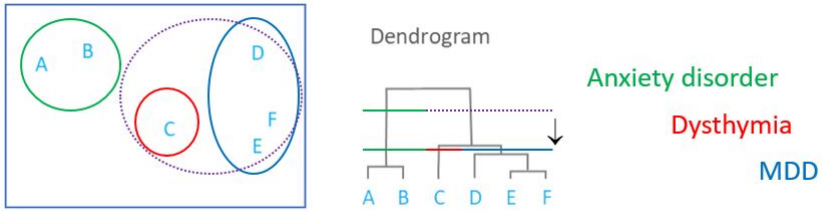
By choosing a certain granularity for the clusters we obtain a labelling for the patients.



This granularity can be determined by the classification system itelf, somewhat akin to model selection methods.

And generating a different one

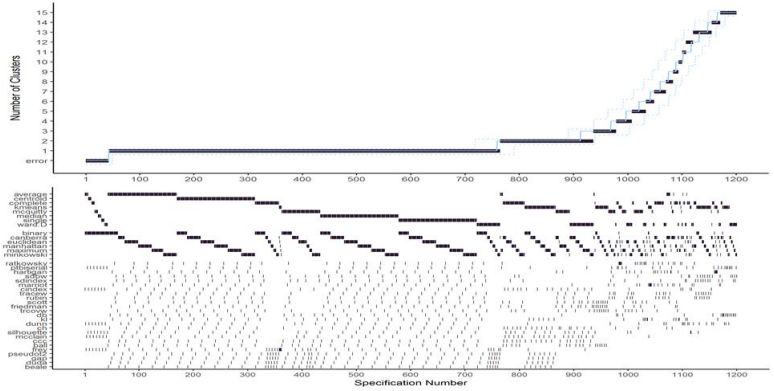
The resulting classification depends on many factors and parameter settings within the system.



Such settings are eventually, though often unreflectively, determined by the users of the system.

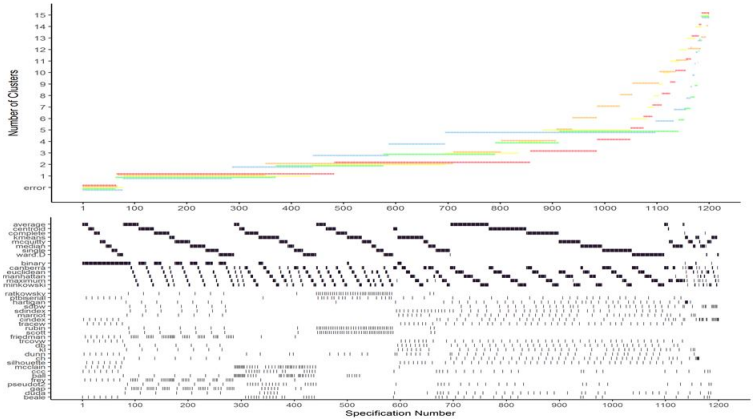
Specification curves

In a large comparison of clustering methods, Beijers et al. (manuscript) did not find much stability in the attempted clusterings.



Specification curves (continued)

When repeating the procedures for simulated data that were constructed to allow for easy detection, the same failures obtain.



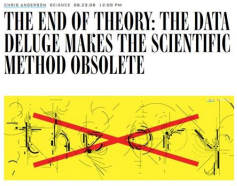
The defects of automated clustering

We must not write off the use of data-driven methods in psychopathology but there are serious problems.

- There is wide variation and little overlap among the results of clustering subtypes of mental disorders.
- The comparison does not point to any particular specifications as being most adequate.
- The theoretical choices do not relate to the clustering outcomes determined by them in a conspicuous way.
- Variance, noise variables, and outliers all contribute to the failure of the clustering.

3 Reliability and transparency

The “data science revolution” is arguably a rerun of a much older instrumentalist ideal of theory-free science.


THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Several machine learning researchers have optimistically claimed the “death of theory”.

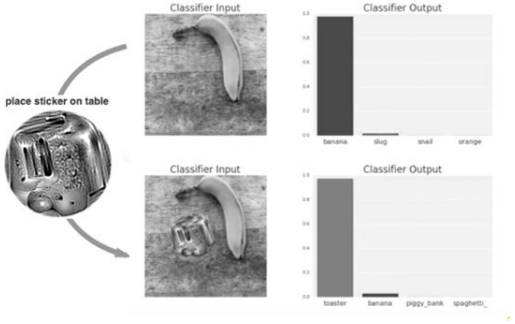
Hume’s curse

Of course all inductive methods are in some way dependent on theoretical starting points: “there is no free lunch”.

If we have no control over these implicit assumptions of our methods, we are liable to application errors.

Inevitable inductive bias

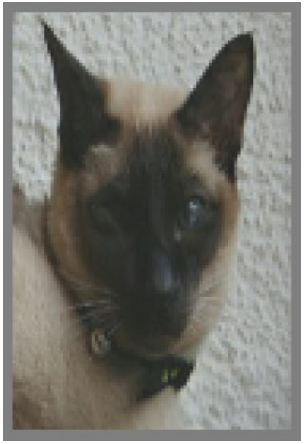
As illustrated by so-called adversarials, automatic classifiers are vulnerable to highly unexpected mistakes.



We have to gain insight into the workings of the classifiers to gain control over the reasons for misfiring and “debug”.

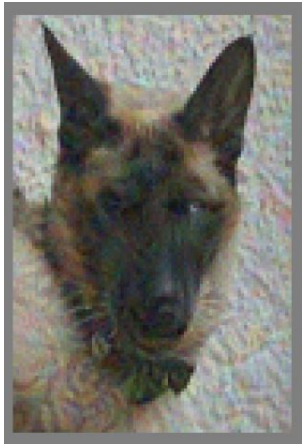
To illustrate adversarials. . .

What animal is this? Computer says “cat”.



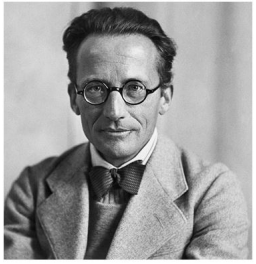
Adding a layer of noise

So what animal is this? Computer says “dog”.



“Anschaulichkeit”

The development of quantum mechanics offers an interesting parallel to this need for intelligibility.



Whether for epistemic, metaphysical or pragmatic reasons, scientists prefer theories that provide insights alongside predictions.

Wish list

In sum, despite the attractiveness of theory-free machine learning methods, we want methods to. . .

- allow continuity in research,
- facilitate accountability,
- be understandable and communicable,
- have clear application criteria,
- avoid erratic mistakes.

Transparency

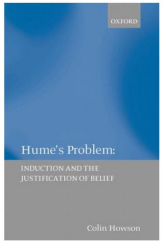
Continuity, intelligibility, accountability, applicability and reliability can all be linked to transparency. If the assumptions implicit in the machine learning methods are uncovered, we can. . .

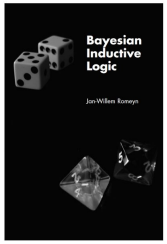
- relate them to earlier models,
- explain and critically assess their results,
- list application criteria and thereby avoid errors.

We have to uncover the inductive assumptions in machine learning methods.

4 Inductive logic

Seeing that probability theory is a form of logic, we can excavate the inductive assumptions that drive our procedures by writing them down in a probabilistic format.





In what follows I will offer several examples of this general idea.

Analogical predictions

Carnapian inductive logic is arguably a precursor of machine learning: data are the only input. Consider sampling pieces of fruit Q_i:

Carnapian predictions are made on the basis of data alone:

$$P(Q_{n+1} = a | Q_1 \ldots Q_n) = \frac{n_a + \lambda/k}{n + \lambda},$$

where the number of possible results $k = 4$ and we might choose $\lambda = k$.

Analogy effects

Carnap gradually admitted more flexibility in the prediction rules. A good example is analogical prediction, e.g.,

$$P(Q_{n+1} = a | Q_1 \ldots Q_n) \quad = \quad \frac{n_{\{a,c\}} + \mu/2}{n + \mu} \times \frac{n_a + \lambda/2}{n_{\{a,c\}} + \lambda}.$$

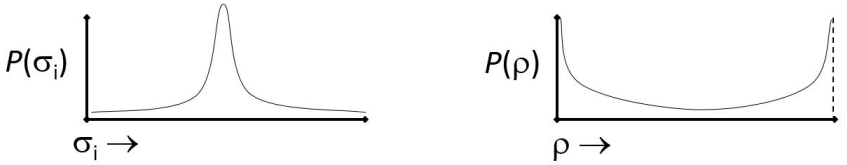
If $\mu < \lambda$, apples and bananas affect our expectation of cherries differently:

$$P(Q_{n+2} = c | Q_1 \ldots Q_n \wedge Q_{n+1} = a) \quad > \quad P(Q_{n+2} = c | Q_1 \ldots Q_n \wedge Q_{n+1} = b).$$

The literature offers numerous other systems that provide a handle on similarity in the data.

Using Bayesian statistics

Translating these prediction rules into Bayesian models is illuminating. We can redefine analogical prediction in fully Bayesian terms, by a prior over multinomial distributions: $P(H_\theta)$ where $\theta \in \langle \rho, \sigma_0, \sigma_1 \rangle$.



Here ρ is the probability for being round, and the σ's are the probabilities of having a stone conditional on being round or not.

Putnam’s adversarials

Notably, there is a striking parallel between adversarials and so-called unlearnable sequences in inductive logic.

- Putnam (1963) challenged Carnap’s project by constructing a sequence that, relative to a set of prediction rules, is not predictable.
- Once a rule assigns a high probability to an observation, the sequence will catch it by surprise and break the pattern.
- The formal learning theory developed after Putnam can shed light on the actively researched issue of adversarials in machine learning.

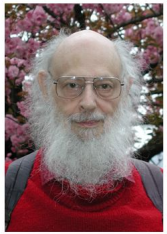
5 Uncovering inductive assumptions

Philosophy and statistics have seen many unsuccessful attempts to rid inductive inference from its theoretical starting points.

We can learn from these examples to inform our analysis of machine learning. Where did the implicit theoretical assumptions go to hide?

Universal prediction

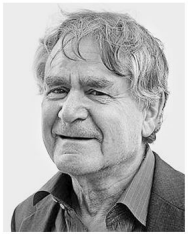
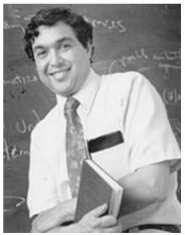
Sterkenburg (2017) offers an in-depth analysis of Solomonoff's idea of universal prediction, i.e., of considering all possible data patterns in prediction.



The predictions rest on the assumption of a highly skewed prior over all semi-computable measures. And in the end they fall prey to Putnam's curse.

Fiducial argument

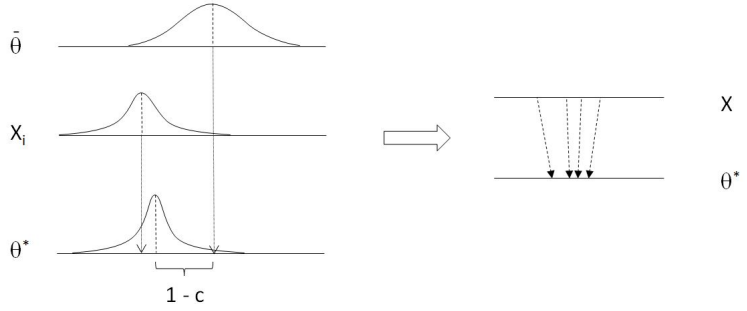
Fisher attempted to generate probabilistic conclusions about statistical hypotheses on the basis of data only.

But. . . his argument rests on the assumption of an improper implicit prior, projected onto the hypotheses via a functional model.

Shrinkage estimators

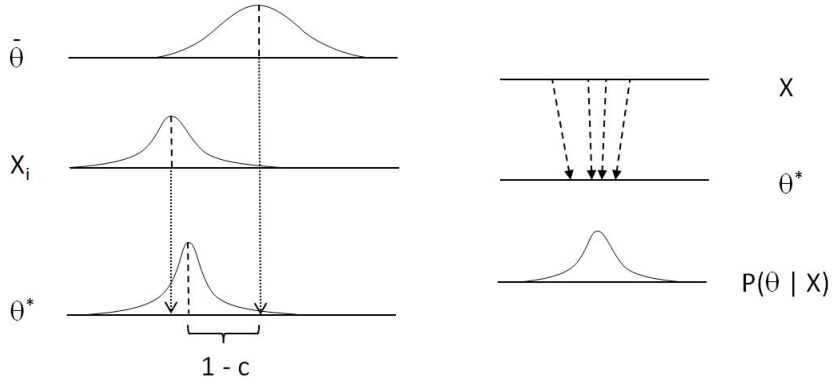
James and Stein (1957) derive that maximum likelihood estimators can be improved if we consider a collection of estimation problems.



As Efron and Morris (1977) show, the predictive improvement rests on an implicit empirical prior.

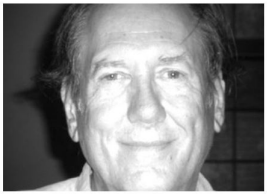
An empirical Bayesian model

Framed as a Bayesian method, Stein's shrinkage factor approximates the Kalman filter. The nudge towards the grand mean is the result of the specific prior that we chose for θ.



Bayesian logic

In all these cases the inductive assumptions are made explicit by converting the inductive procedures into Bayesian format.


The basis for this is in the view that the Bayesian format is a logic and as such epistemologically neutral.

6 An epistemology of data science

Philosophy of science can help to introduce data science methods into science in a responsible way.

- Data science will very likely transform our sciences so we will have to focus attention there.
- Preliminary studies suggest that the outcomes of these methods suffer from failures of reliability.
- To improve on the assistance, our primary goal should be to identify the assumptions inherent in the data science methods.

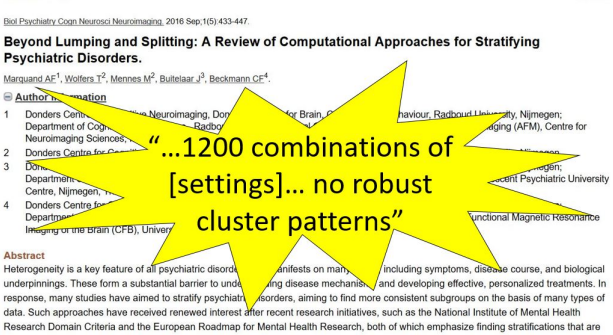
Making the assumptions explicit

The foregoing suggests how we can uncover inductive assumptions inherent in the new data science methods.

The idea is that we can identify modeling assumptions by translating data science into Bayesian logic.

Data-driven psychopathology

Automated psychiatric classification systems provide a good example of where transparency is badly needed.



Although methodological concerns are growing, the uptake of data science methods in the sciences cannot be stopped.

Towards a clarification of clustering

Framing the clustering methods in terms of a Bayesian logic helps us to see what assumptions might motivate the clustering method.

- Preliminary work suggests useful parallels between clustering and least-squares curve-fitting.
- Increasing the number of clusters is similar to increasing the number of parameters describing a family of curves.
- Any automated clustering method can be replicated by a hierarchical Bayesian model with distributional assumptions on the nature of a cluster.

Why again?

Uncovering the assumptions of machine learning is an important task for the philosophy of science.

- It will help to integrate the new methods into existing and more theoretical approaches.
- Similarly it will improve on the communicability and public acceptance of machine learning results.
- And it will make it easier to hold researchers accountable and critically scrutinize their conclusions.
- Most importantly, it will help to apply methods correctly and guard against unreliable inferences.

Beyond logic

There are more assumptions to take into account though, and they are not all covered by logical analysis.

- Machine learning also relies on how the sample space and the space of theoretical possibilities is constructed.
- Many machine learning methods include a form of model selection, and thereby a decision procedure, over and above model evaluation.
- The application of machine learning methods involves interpretations of their results.

Thanks for your attention

Help from Lian Beijers and Hanna van Loo is gratefully acknowledged. Slides of the talk will be available at http://www.philos.rug.nl/~romeyn. For comments and questions, email j.w.romeijn@rug.nl.