Métodos cuantitativos y estadísticos en las ciencias del lenguaje:

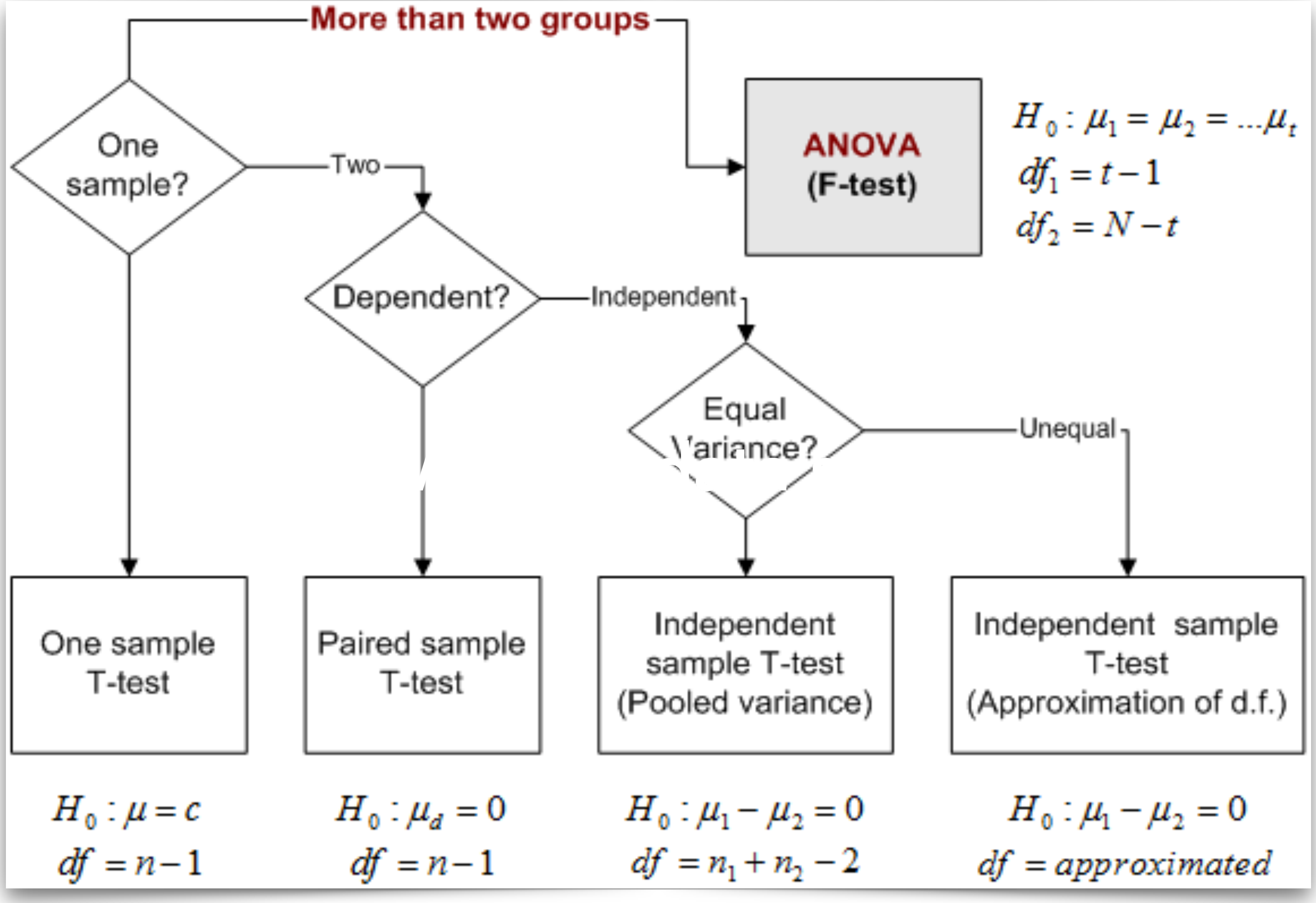
panorama, predicciones y recomendaciones

Damián E. Blasi

Department of Comparative Linguistics,
University of Zürich

Department of Linguistic and Cultural Evolution,
Max Planck Institute for the Science of Human History

Human Relation Area Files,
Yale University



More than two groups

One sample?

Two

Dependent?

Independent

Equal Variance?

Unequal

ANOVA (F-test)

$H_0: \mu_1 = \mu_2 = \ldots \mu_t$

$df_1 = t - 1$

$df_2 = N - t$

One sample T-test

$H_0: \mu = c$

$df = n - 1$

Paired sample T-test

$H_0: \mu_d = 0$

$df = n - 1$

Independent sample T-test (Pooled variance)

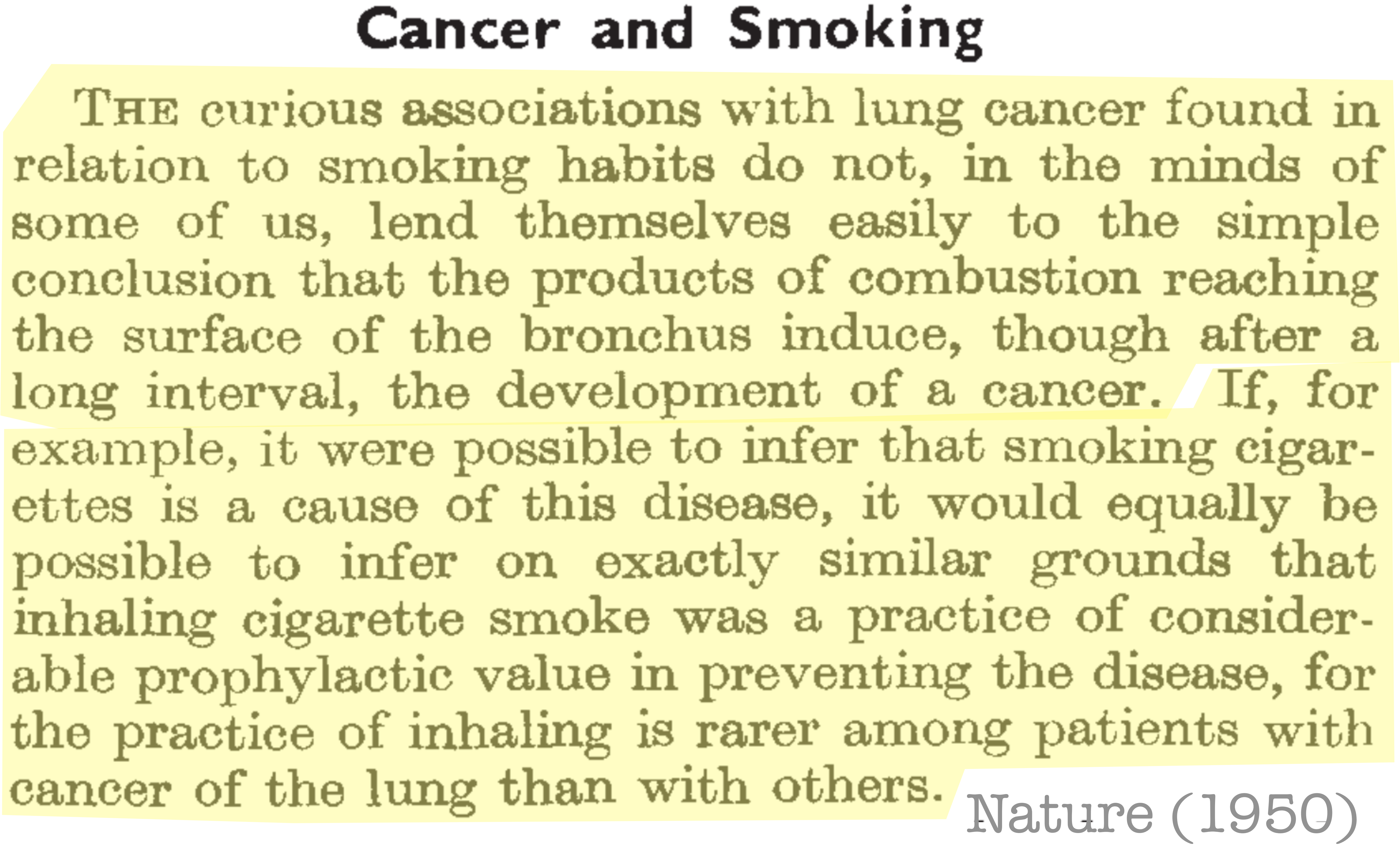
$H_0: \mu_1 - \mu_2 = 0$

$df = n_1 + n_2 - 2$

Independent sample T-test (Approximation of d.f.)

$H_0: \mu_1 - \mu_2 = 0$

$df = approximated$



Cancer and Smoking

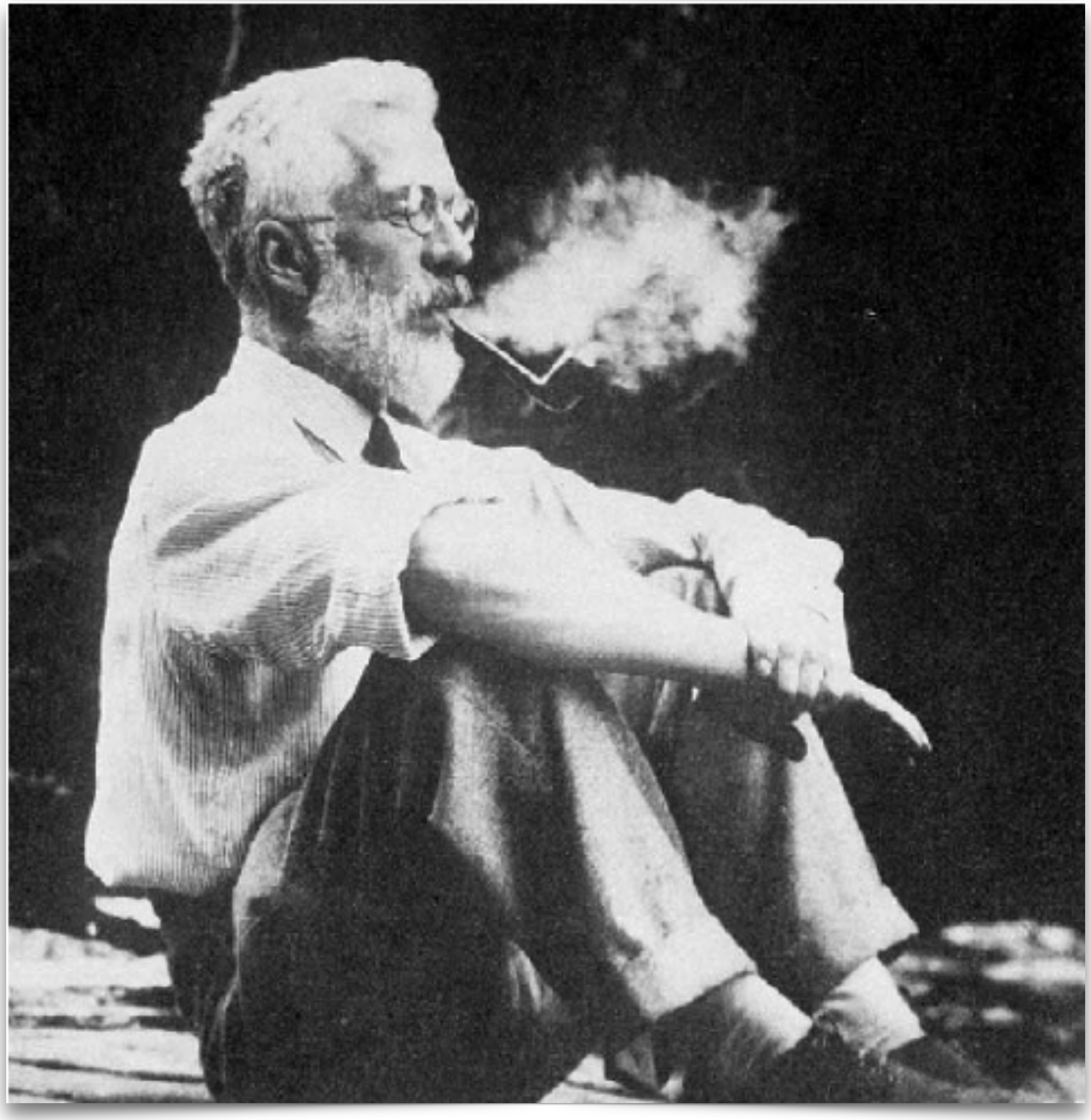
THE curious associations with lung cancer found in relation to smoking habits do not, in the minds of some of us, lend themselves easily to the simple conclusion that the products of combustion reaching the surface of the bronchus induce, though after a long interval, the development of a cancer. If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others.

Nature (1950)

What is data science?

Ronald Fisher

What is data science?

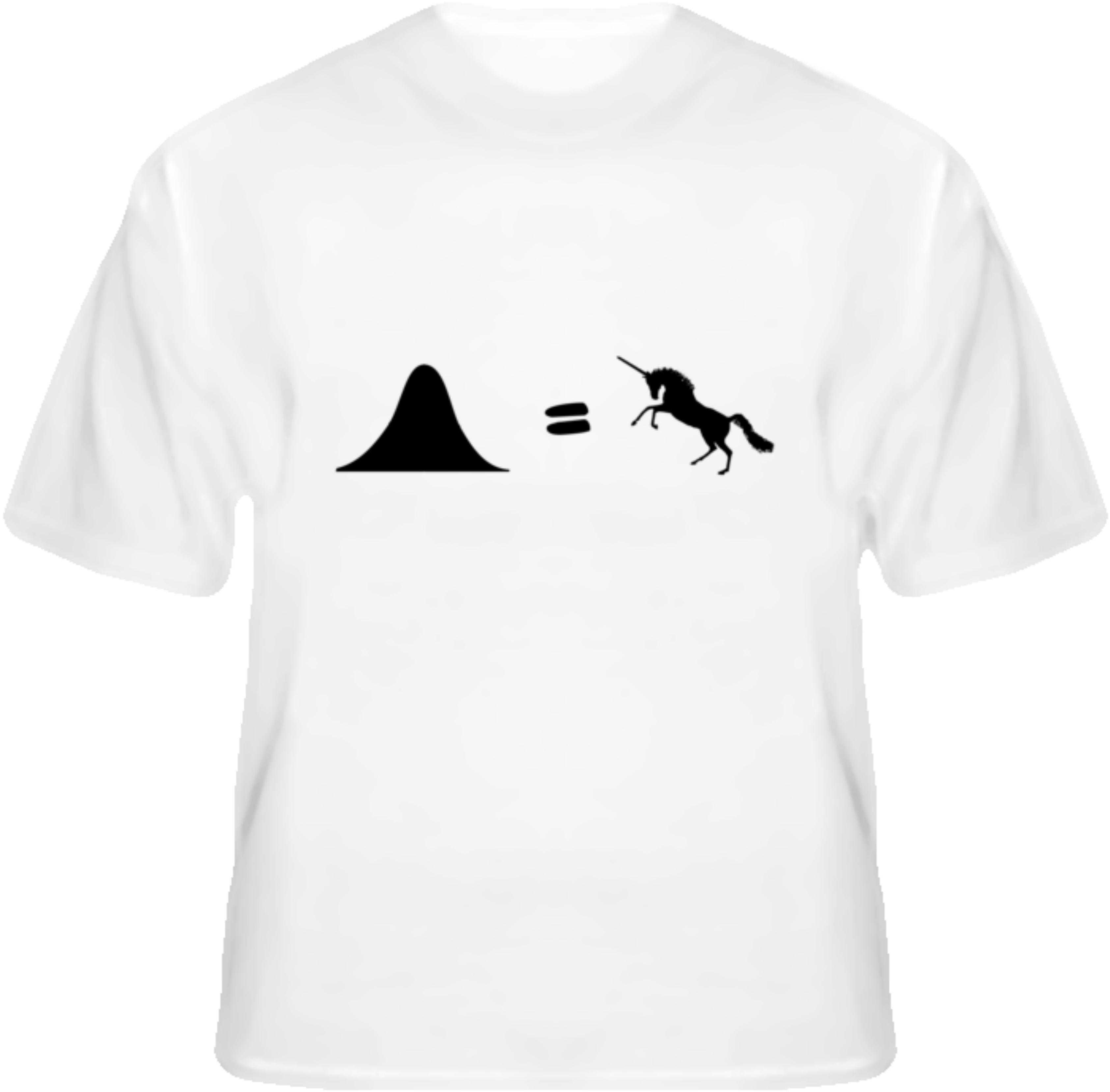


Sometimes the assumptions and the statements of statistical methods look *way too complex* to have anything to do with the real world.

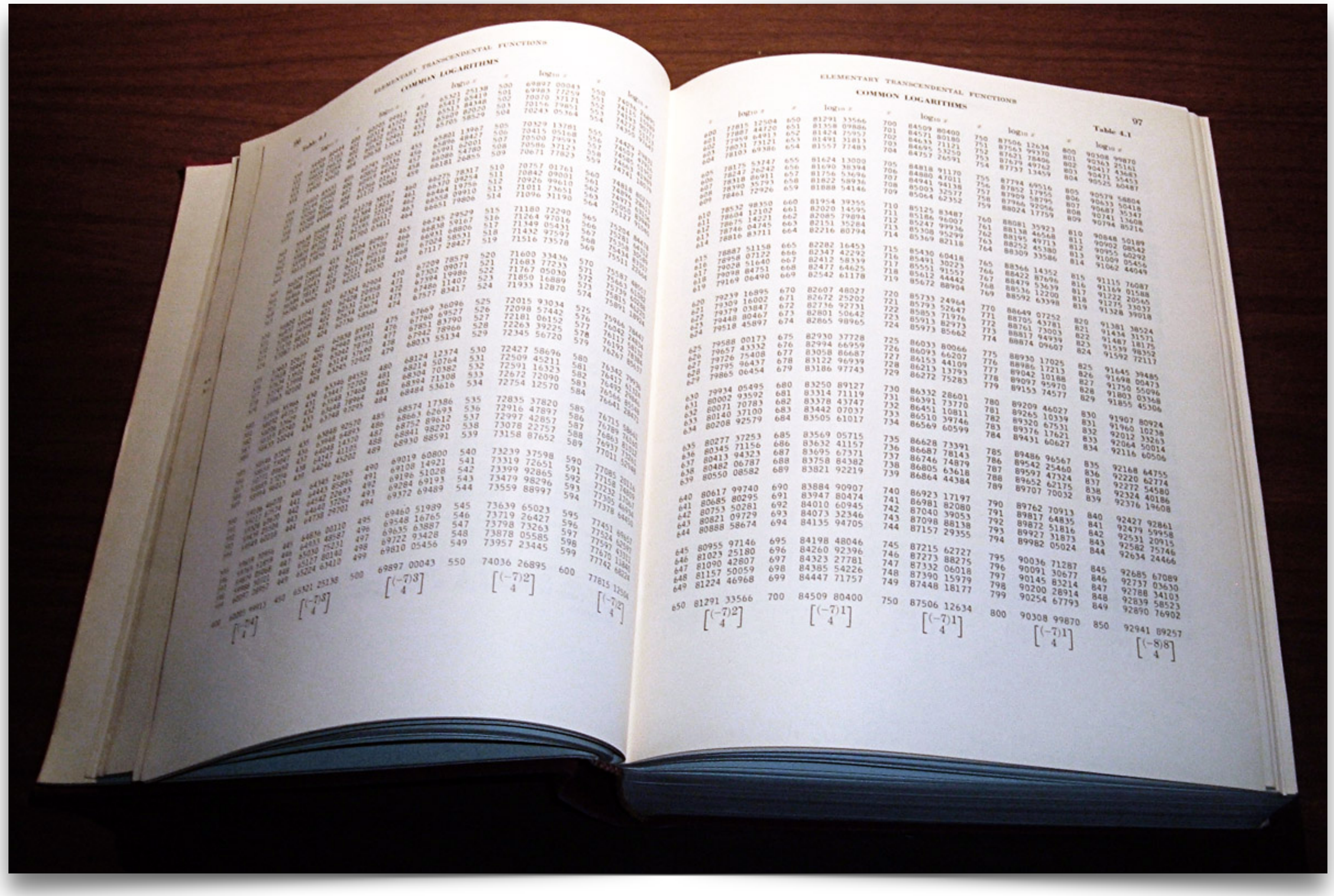
E.g. assumptions behind a basic linear regression:

homoskedastic,
independent,
normally distributed data

No data I know in linguistics/anthropology/cognitive sciences satisfy this!



These assumptions were motivated by sheer convenience



ELEMENTARY TRANSCENDENTAL FUNCTIONS

COMMON LOGARITHMS

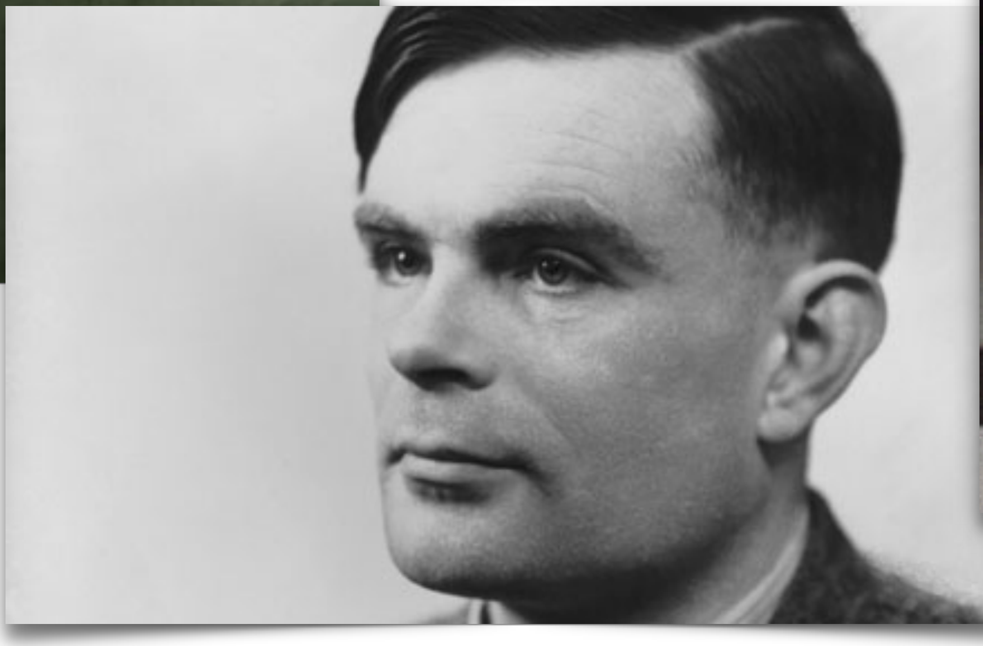
Table 4.1

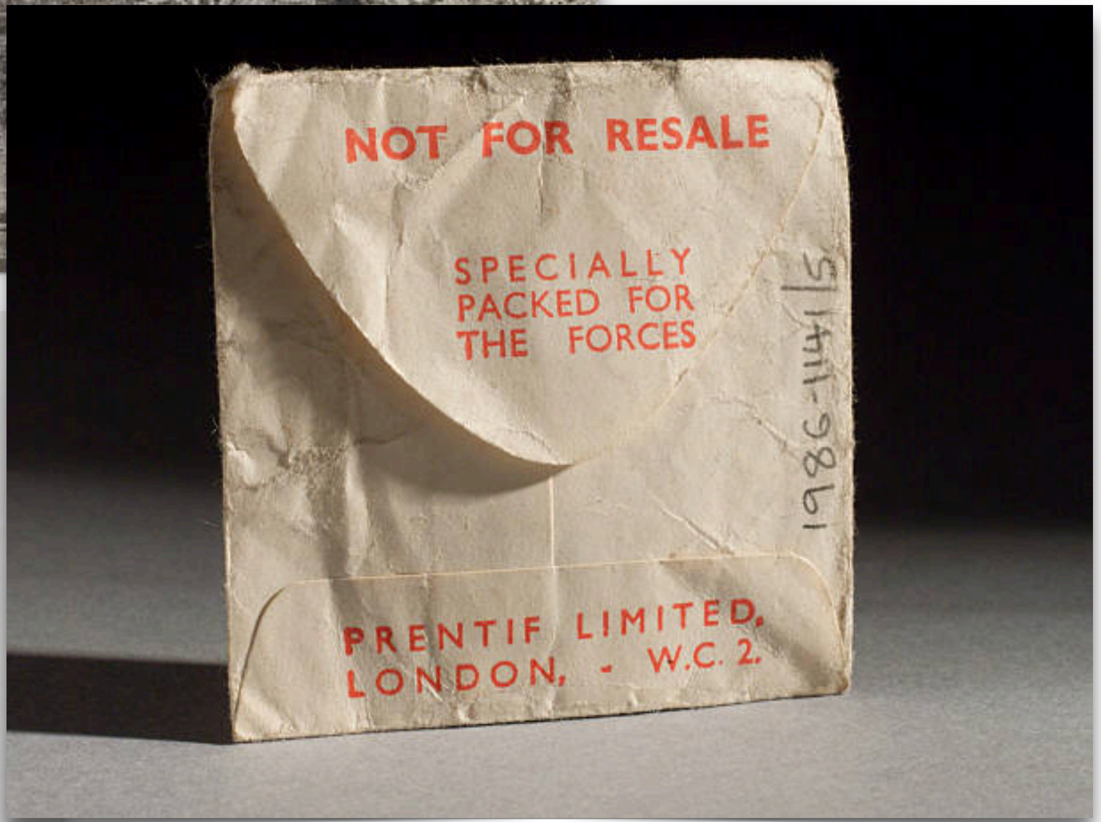
97

That explains the quirky math, but what about the recipe-like flavor of stats?

George Barnard

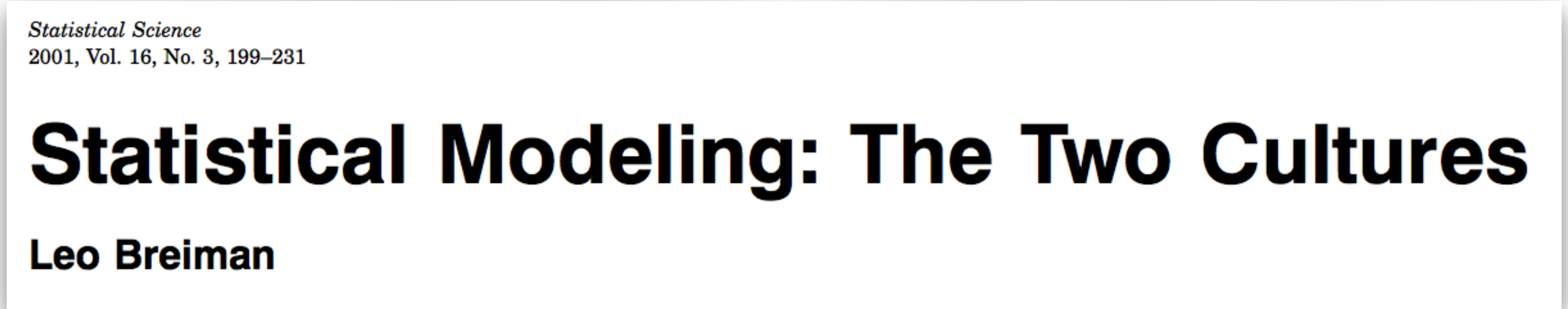
The development of early methods was **strongly** practically oriented




NOT FOR RESALE
SPECIALLY PACKED FOR THE FORCES
PRENTIF LIMITED, LONDON, - W.C. 2.
1986-1441/5

The usual statistics course in the social and health sciences results from the need of making standardized decisions

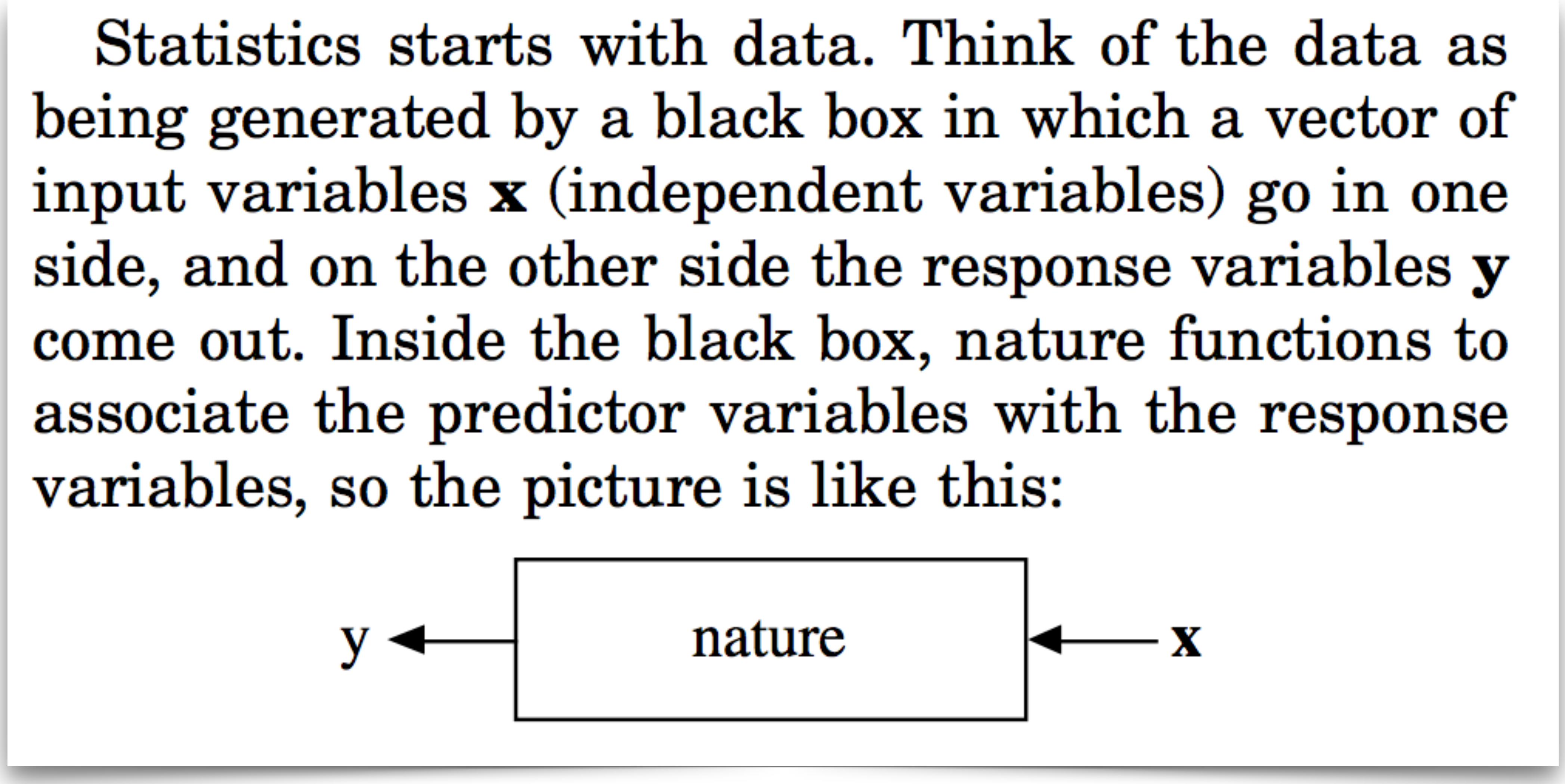
Computers have partially solved for us the first issue - we don’t require Procrustean assumptions anymore



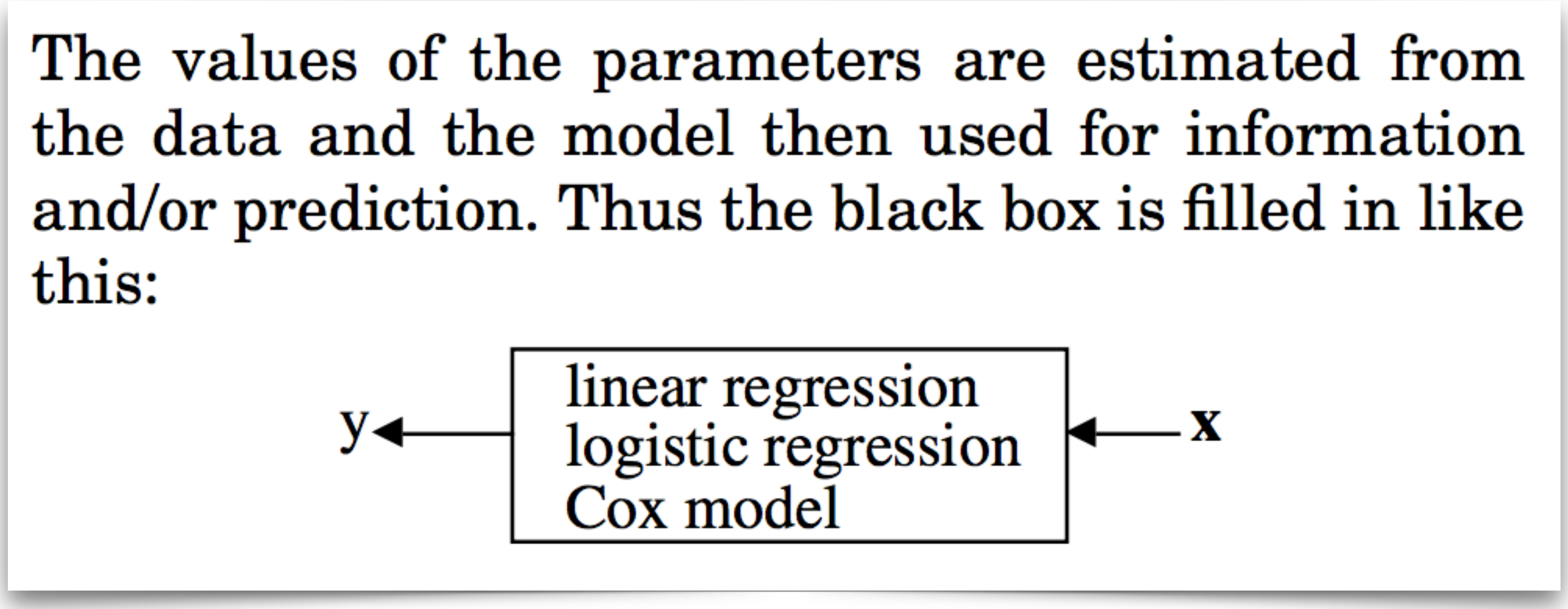
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

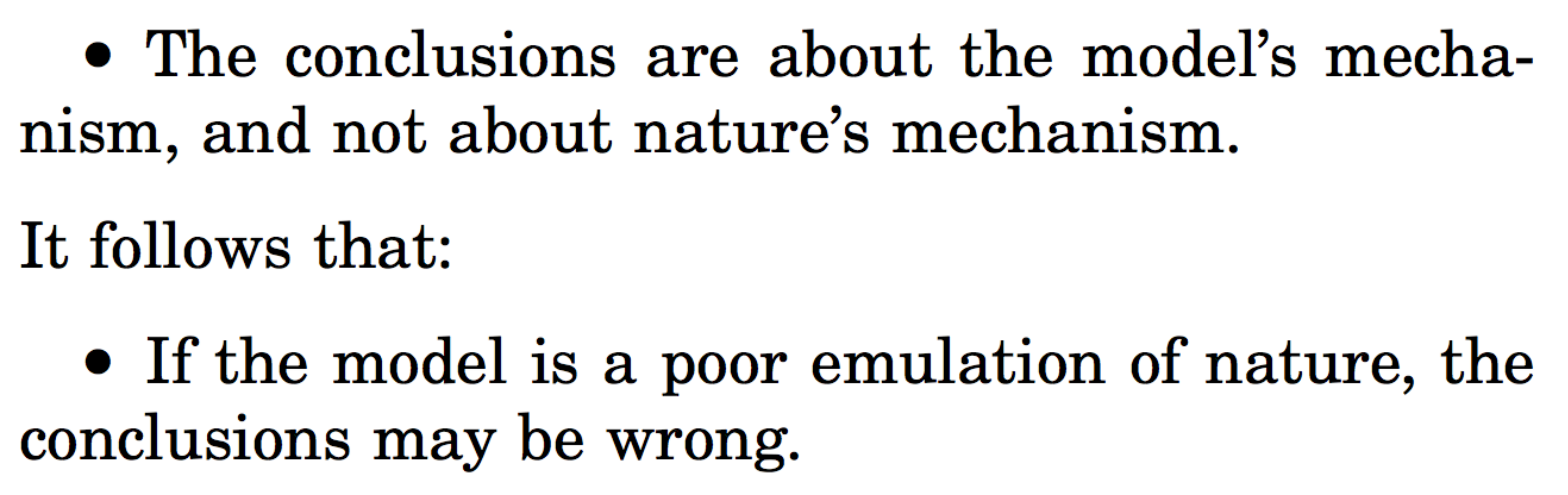


Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



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:

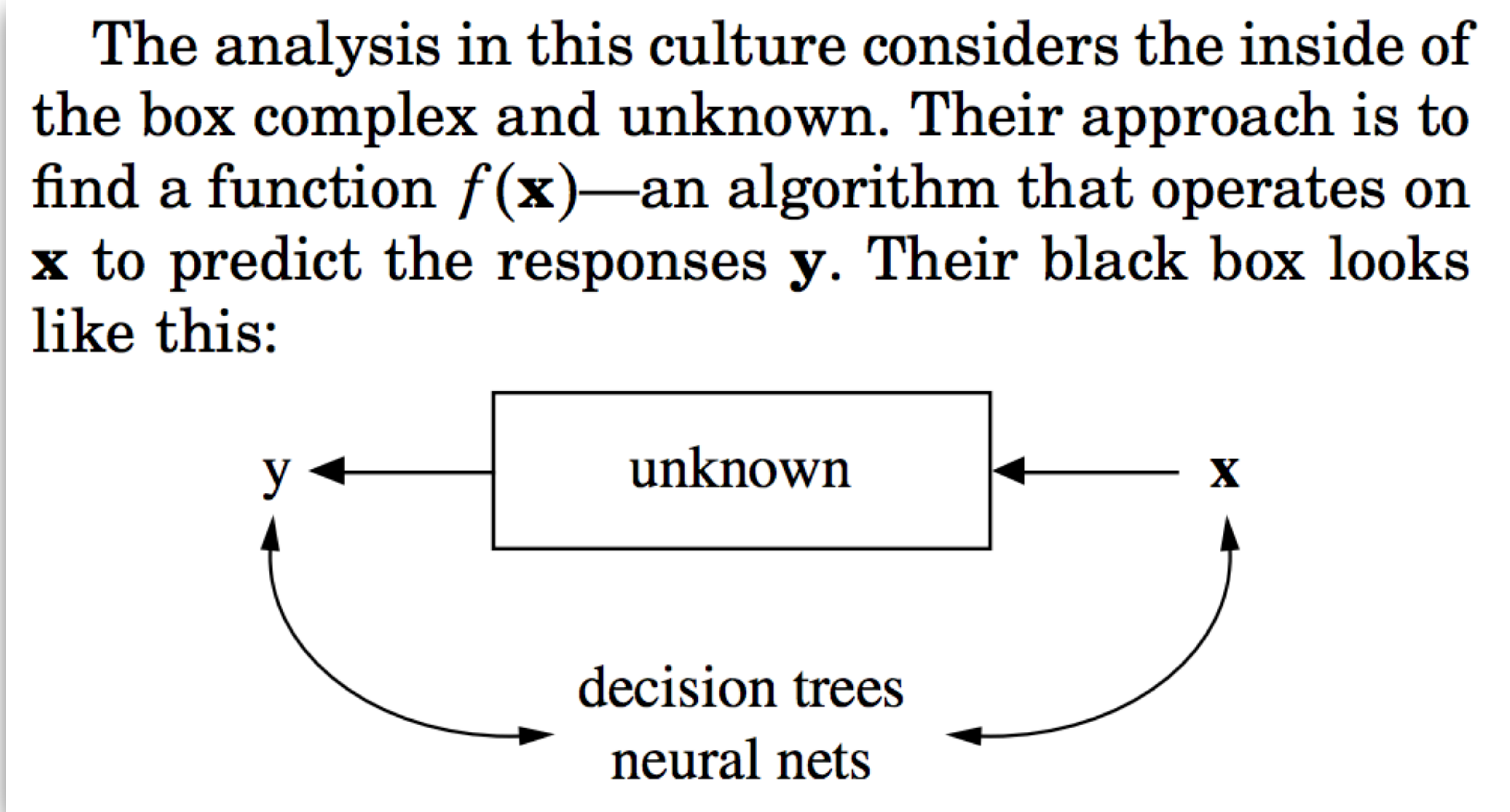
y
linear regression
logistic regression
Cox model
x



- The conclusions are about the model’s mechanism, and not about nature’s mechanism.

It follows that:

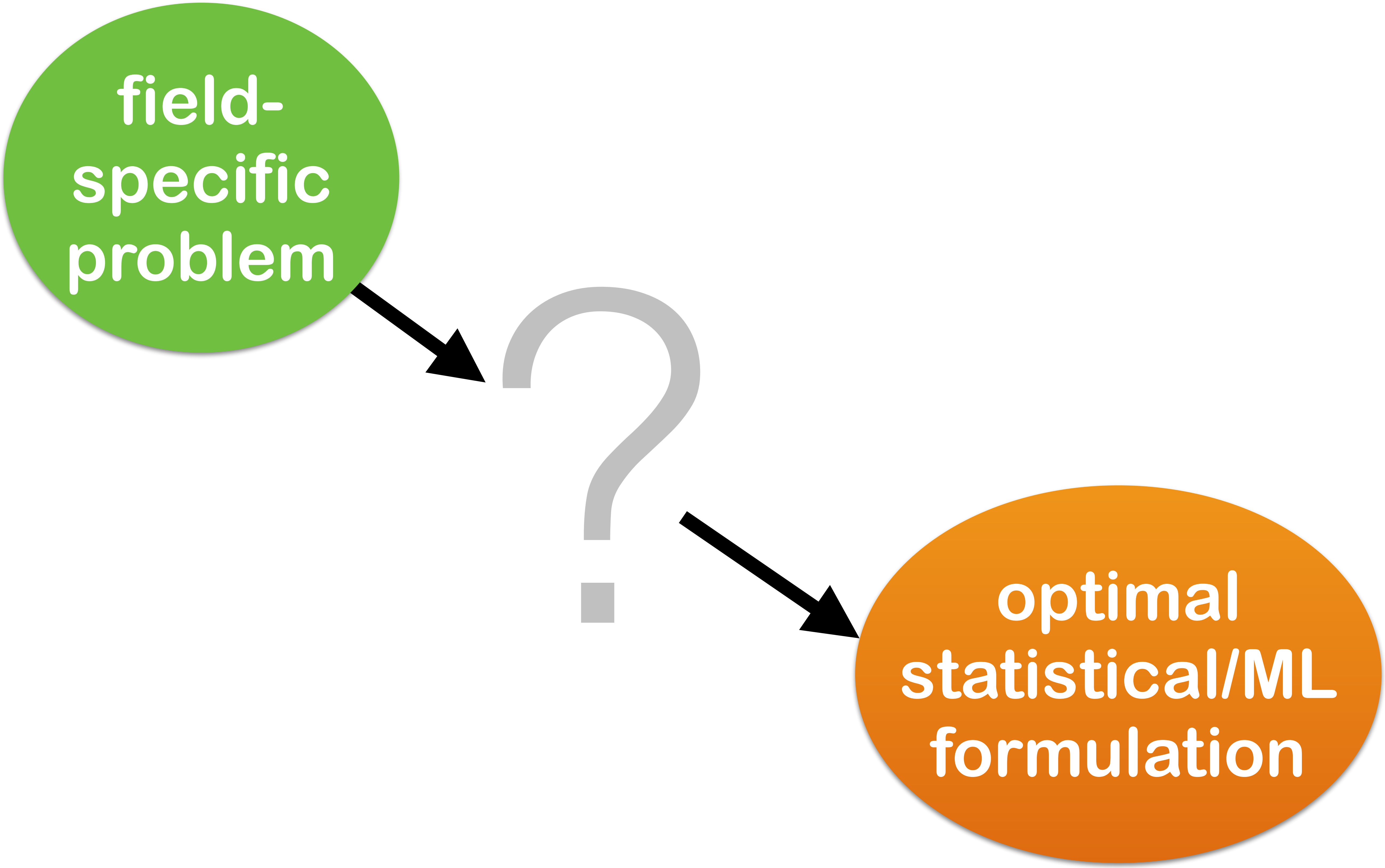
- If the model is a poor emulation of nature, the conclusions may be wrong.



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:

y
unknown
x
decision trees
neural nets

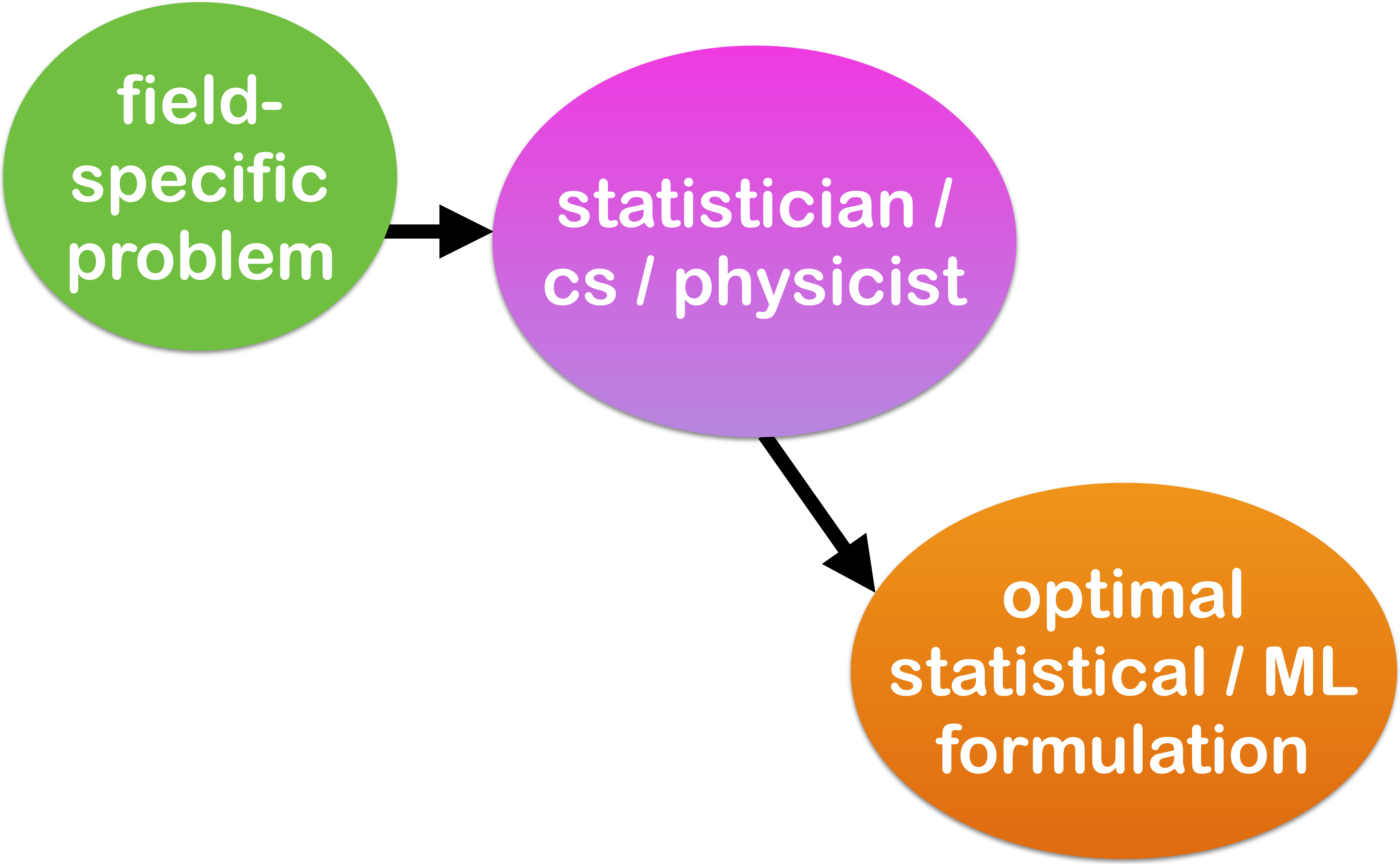
So now we have very
powerful models that are
able to describe (and predict)
subtle associations in data
without yielding an explicit
mechanistic pathway



field-specific problem

?

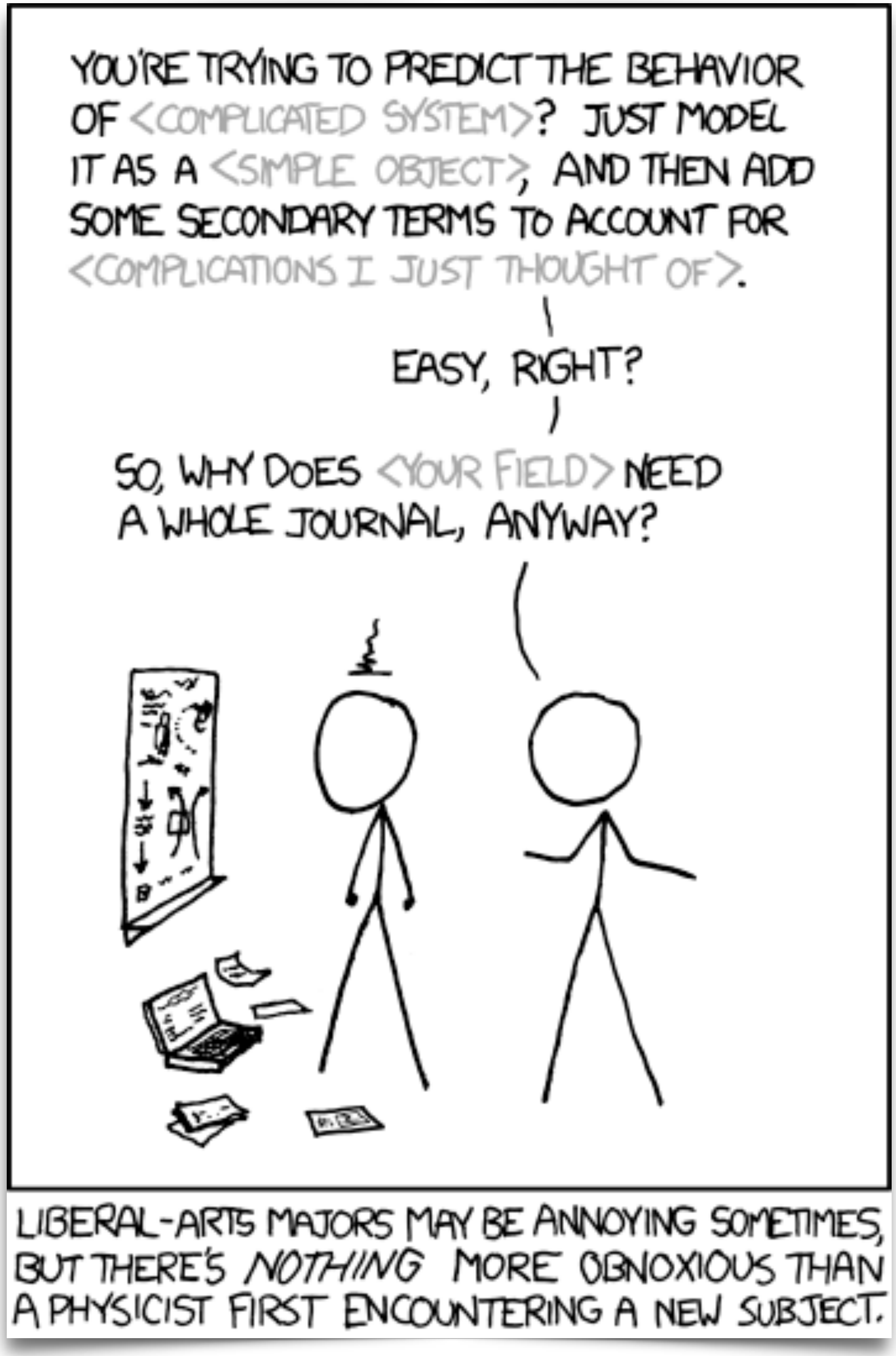
optimal statistical/ML formulation



field-specific problem

statistician / cs / physicist

optimal statistical / ML formulation



YOU'RE TRYING TO PREDICT THE BEHAVIOR OF <COMPLICATED SYSTEM>? JUST MODEL IT AS A <SIMPLE OBJECT>, AND THEN ADD SOME SECONDARY TERMS TO ACCOUNT FOR <COMPLICATIONS I JUST THOUGHT OF>.
EASY, RIGHT?
SO, WHY DOES <YOUR FIELD> NEED A WHOLE JOURNAL, ANYWAY?
LIBERAL-ARTS MAJORS MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A PHYSICIST FIRST ENCOUNTERING A NEW SUBJECT.

Linguistics

Modelling the dynamics of language death

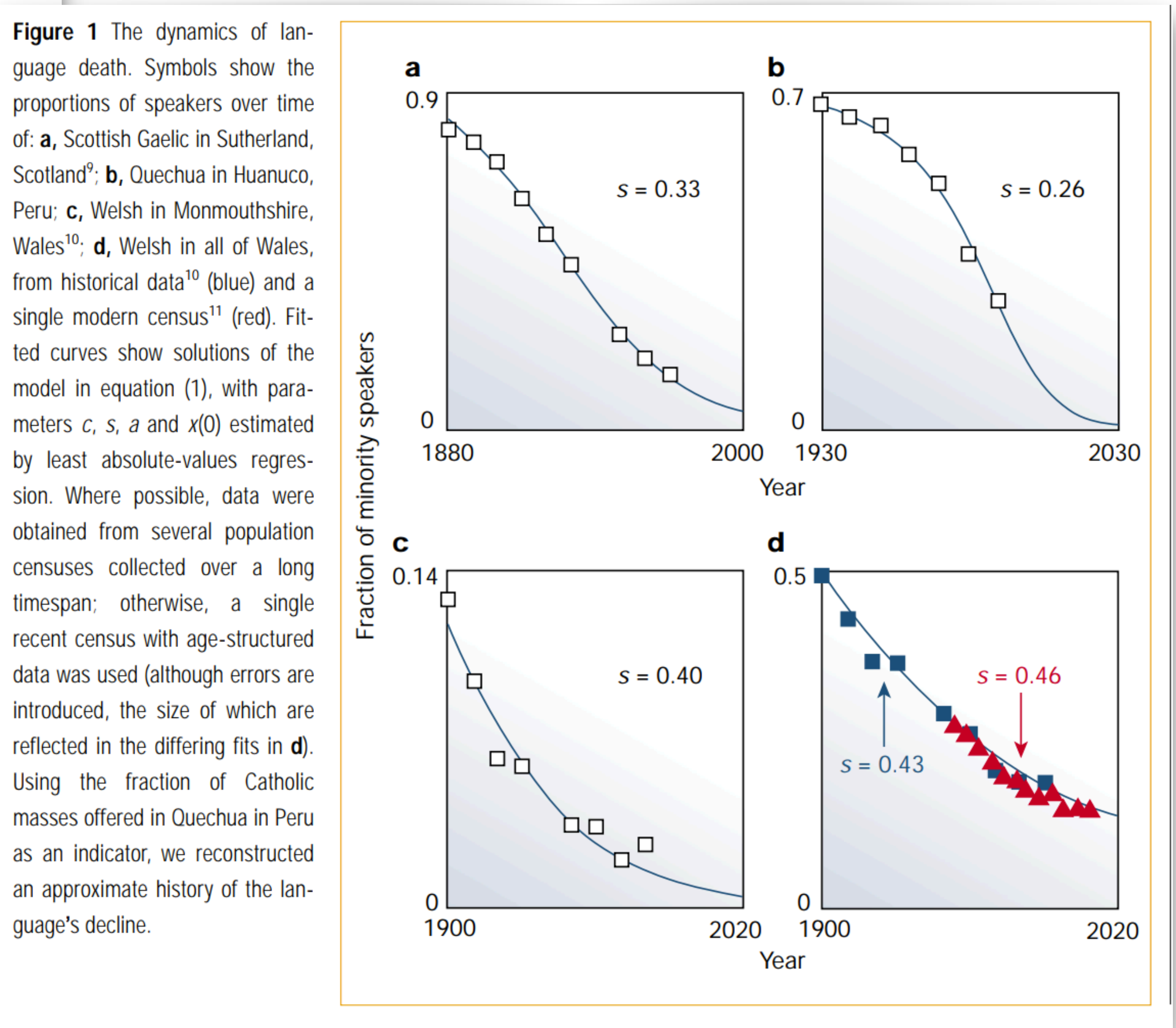
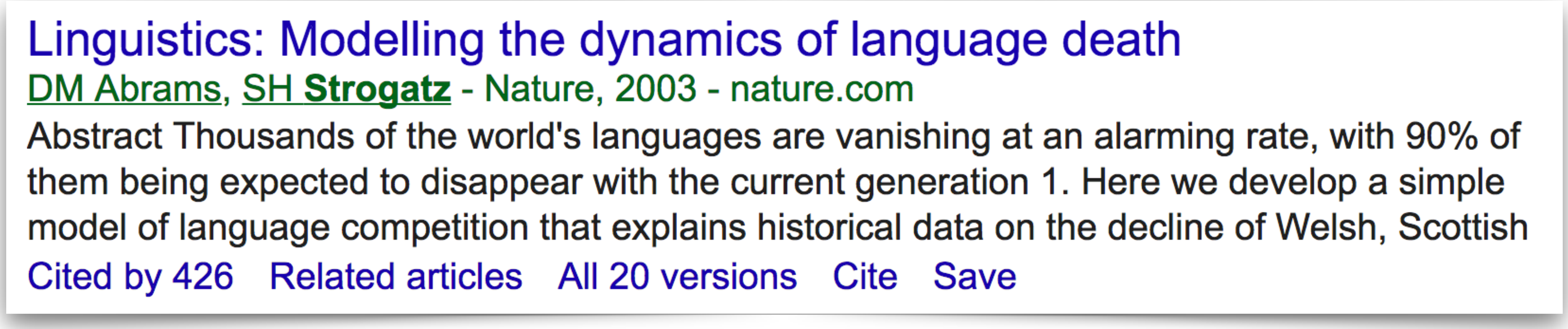


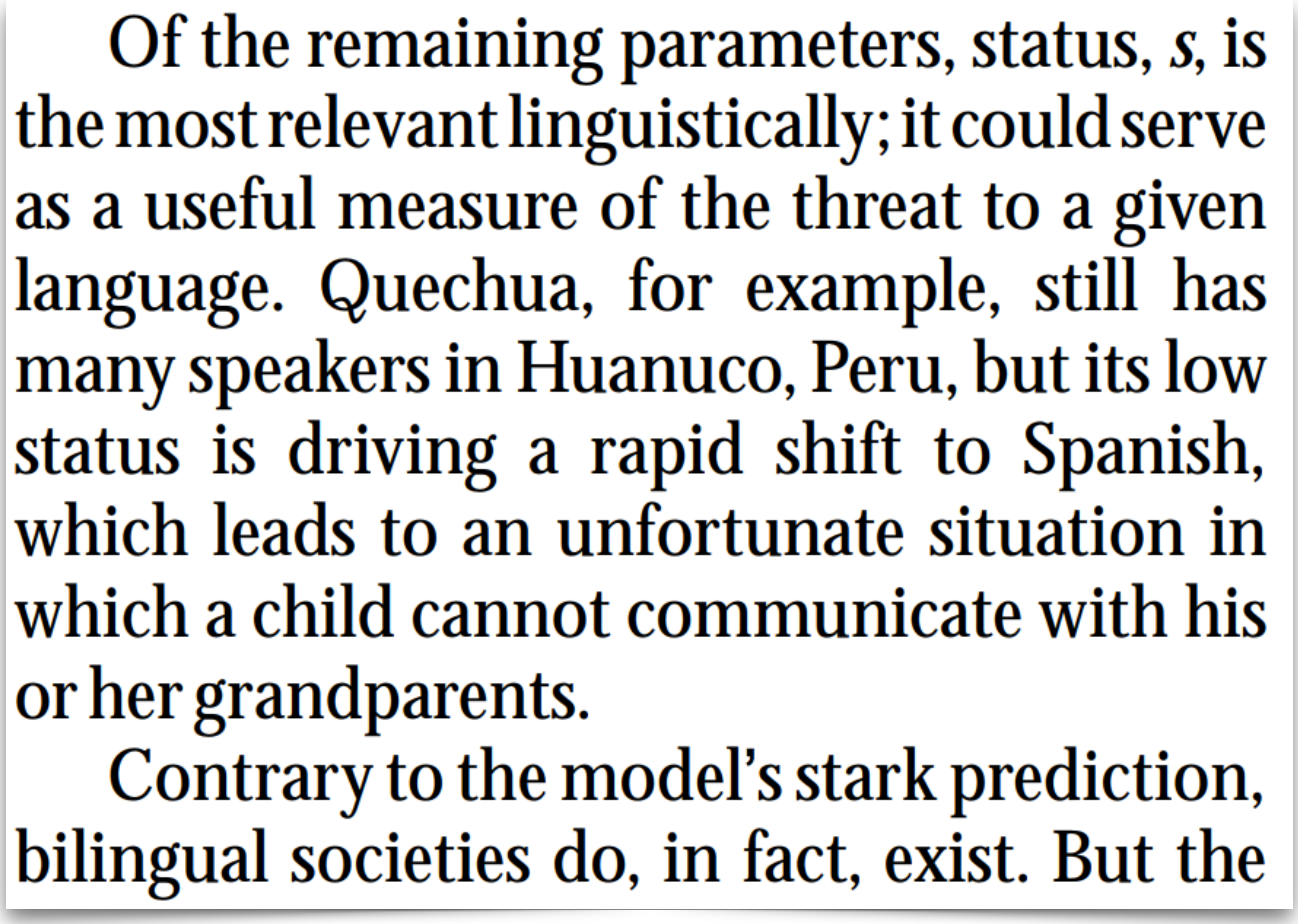
Figure 1 The dynamics of language death. Symbols show the proportions of speakers over time of: **a,** Scottish Gaelic in Sutherland, Scotland[9]; **b,** Quechua in Huanuco, Peru; **c,** Welsh in Monmouthshire, Wales[10]; **d,** Welsh in all of Wales, from historical data[10] (blue) and a single modern census[11] (red). Fitted curves show solutions of the model in equation (1), with parameters *c*, *s*, *a* and *x*(0) estimated by least absolute-values regression. Where possible, data were obtained from several population censuses collected over a long timespan; otherwise, a single recent census with age-structured data was used (although errors are introduced, the size of which are reflected in the differing fits in **d**). Using the fraction of Catholic masses offered in Quechua in Peru as an indicator, we reconstructed an approximate history of the language's decline.



Linguistics: Modelling the dynamics of language death

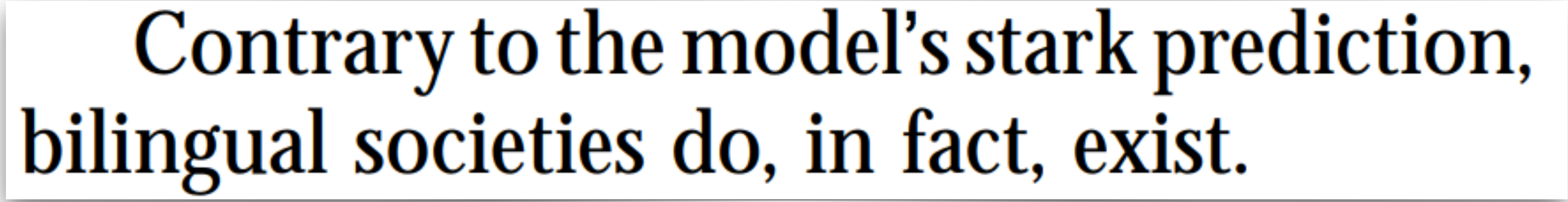
DM Abrams, SH **Strogatz** - Nature, 2003 - nature.com

Abstract Thousands of the world's languages are vanishing at an alarming rate, with 90% of them being expected to disappear with the current generation 1. Here we develop a simple model of language competition that explains historical data on the decline of Welsh, Scottish

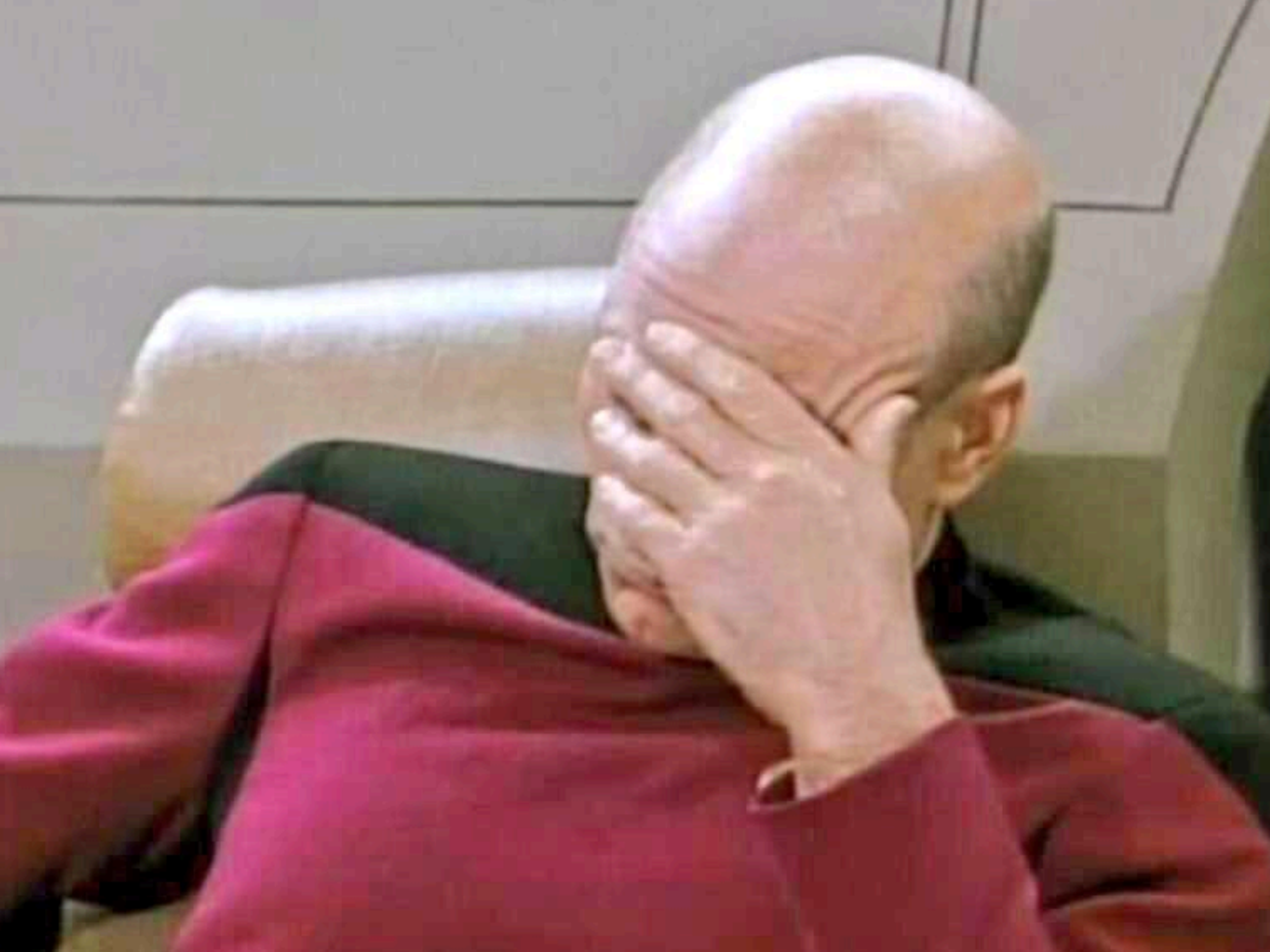


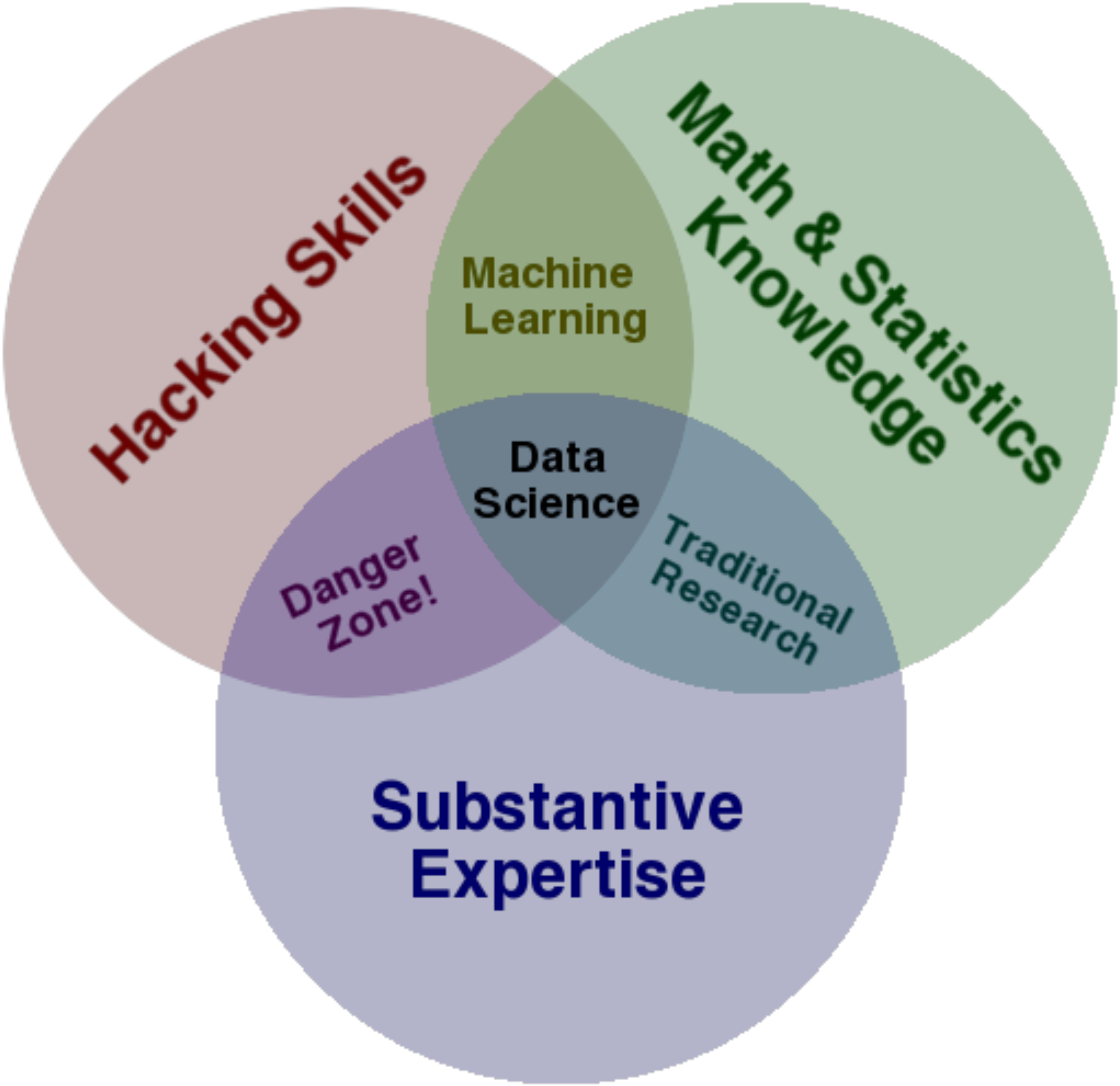
Of the remaining parameters, status, *s*, is the most relevant linguistically; it could serve as a useful measure of the threat to a given language. Quechua, for example, still has many speakers in Huanuco, Peru, but its low status is driving a rapid shift to Spanish, which leads to an unfortunate situation in which a child cannot communicate with his or her grandparents.

Contrary to the model's stark prediction, bilingual societies do, in fact, exist. But the



Contrary to the model's stark prediction, bilingual societies do, in fact, exist.



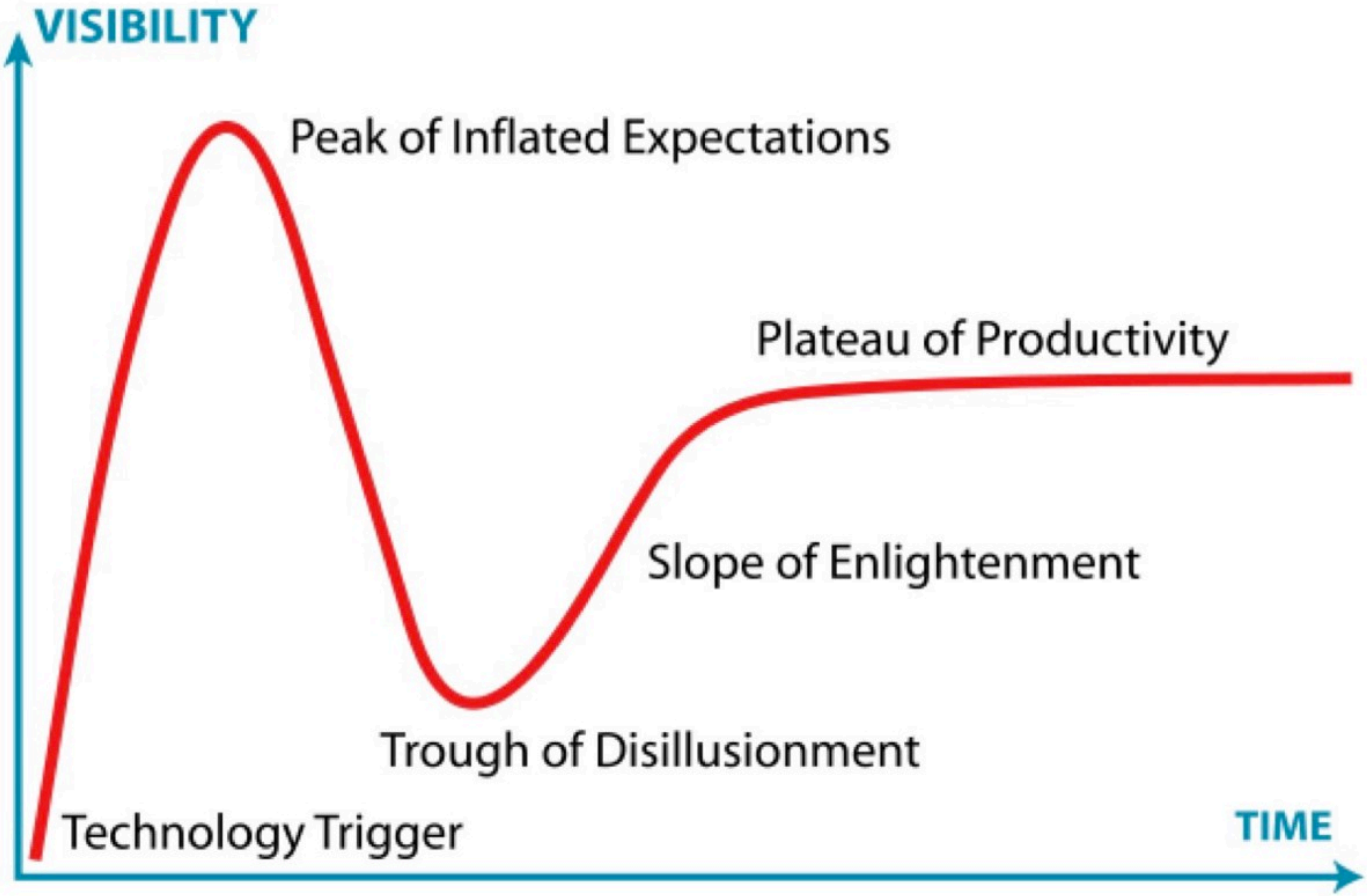

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

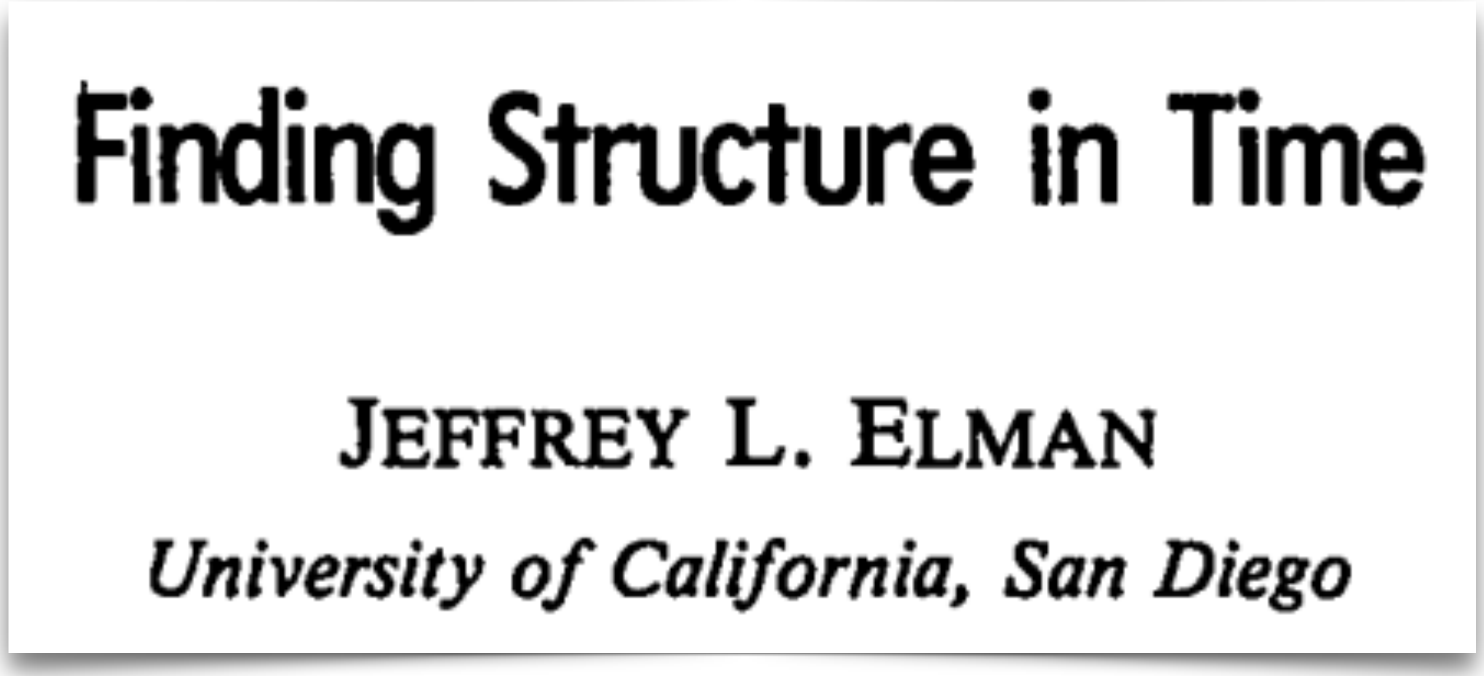
There is a blossoming data science of language that is *not* restricted to traditionally quantitative areas (such as e.g. psycholx, computational linguistics, phonetics, etc)

Importantly, more and more methods are informed by the special structure of the problem at hand

So, where are we *not* going from here?

Gartner Hype Cycle

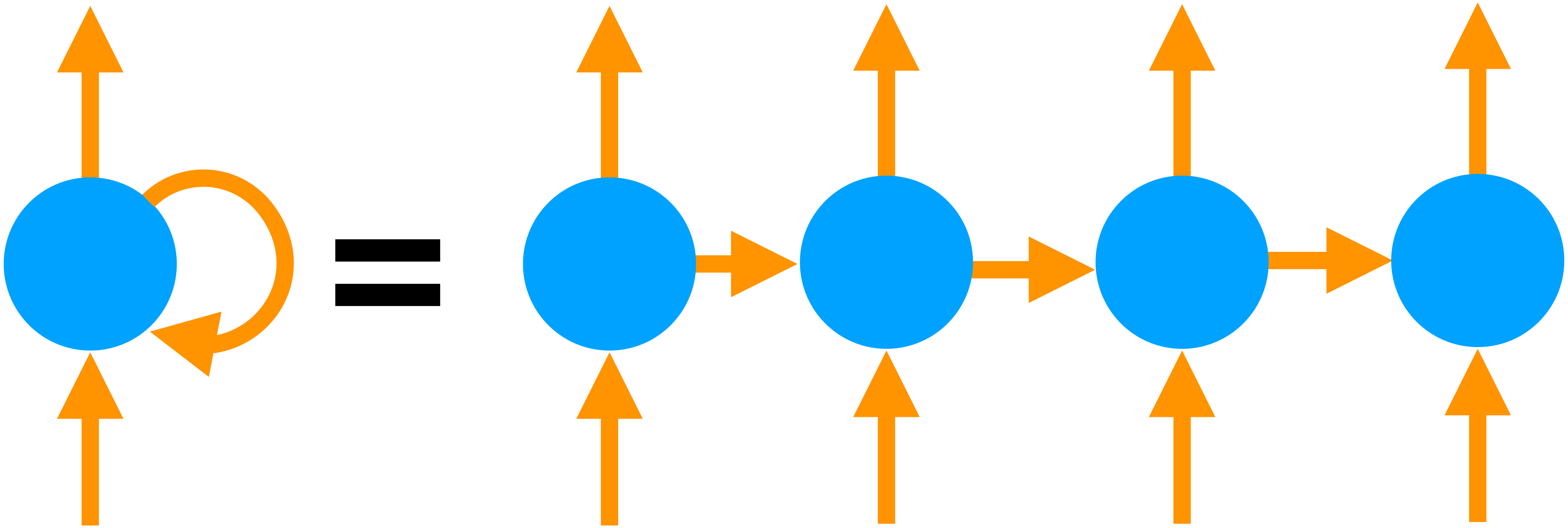


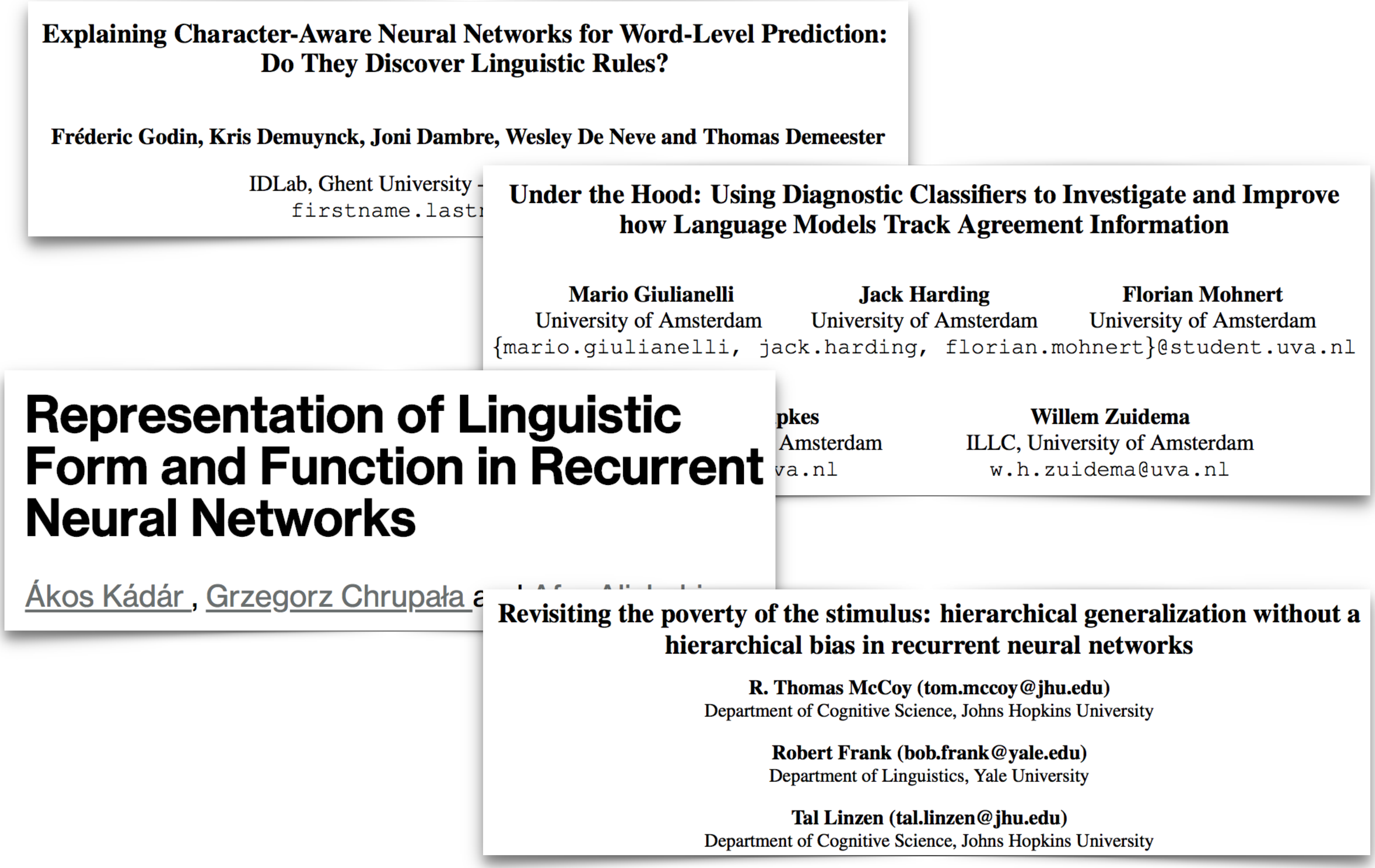


Finding Structure in Time

JEFFREY L. ELMAN

University of California, San Diego





Explaining Character-Aware Neural Networks for Word-Level Prediction: Do They Discover Linguistic Rules?

Fréderic Godin, Kris Demuynck, Joni Dambre, Wesley De Neve and Thomas Demeester

IDLab, Ghent University -
firstname.last

Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information

Mario Giulianelli University of Amsterdam
Jack Harding University of Amsterdam
Florian Mohnert University of Amsterdam
{mario.giulianelli, jack.harding, florian.mohnert}@student.uva.nl

pkes
Amsterdam
va.nl

Willem Zuidema
ILLC, University of Amsterdam
w.h.zuidema@uva.nl

Representation of Linguistic Form and Function in Recurrent Neural Networks

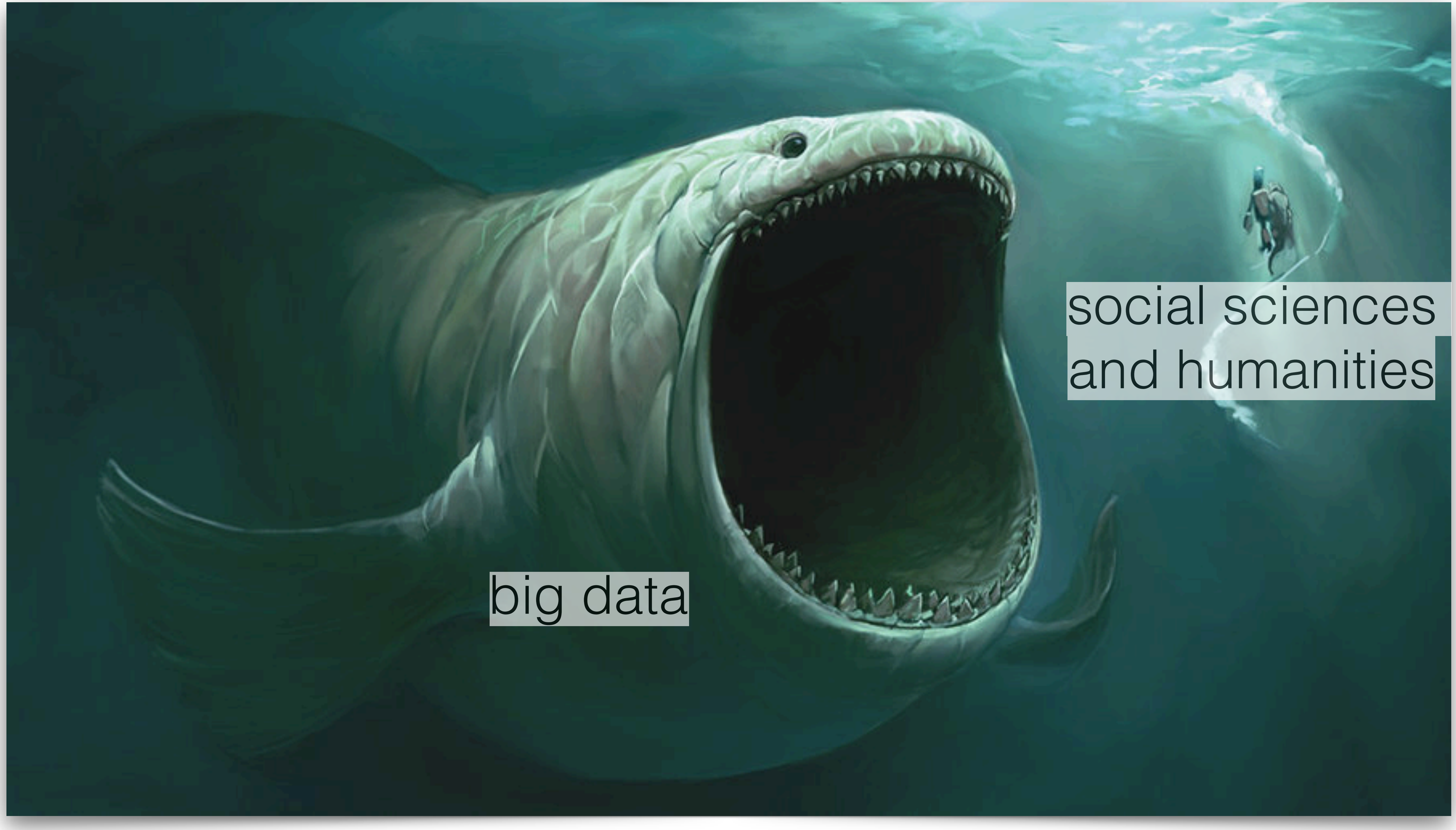
Ákos Kádár , Grzegorz Chrupała a

Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks

R. Thomas McCoy (tom.mccoy@jhu.edu)
Department of Cognitive Science, Johns Hopkins University

Robert Frank (bob.frank@yale.edu)
Department of Linguistics, Yale University

Tal Linzen (tal.linzen@jhu.edu)
Department of Cognitive Science, Johns Hopkins University


social sciences and humanities
big data

Fin