Generalization from the Behavioral Perspective

Workshop on Emerging Generalization Settings

Maxim Raginsky (UIUC)

Device Modeling

separating experience from models

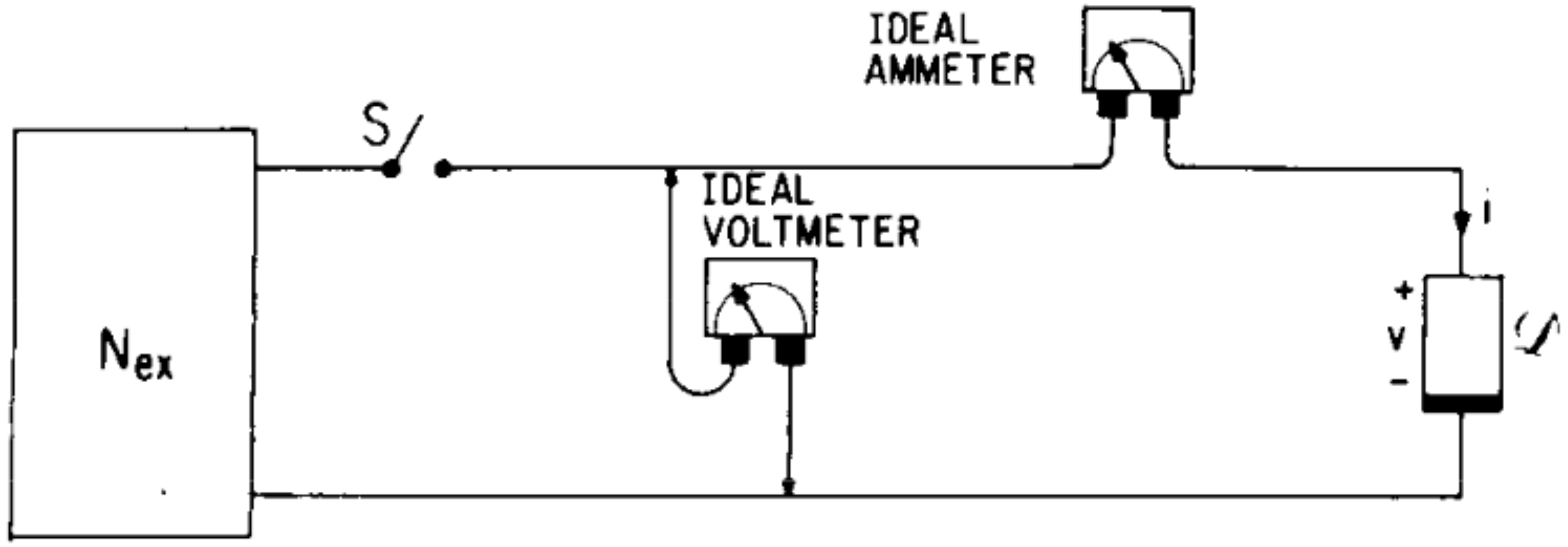


Fig. 1. Circuit for measuring admissible voltage–current signal pairs associated with a 2-terminal or one-port device $\mathscr{D}$.

Behavior of device $\mathscr{D}$: collection of all current and voltage waveforms that can be produced by connecting $\mathscr{D}$ to an arbitrary excitation network and closing the switch S at an arbitrary time t_0.

Model of device $\mathscr{D}$: a mathematical construct that allows us to make predictions about $\mathscr{D}$ in various circumstances of interest without having to enumerate or measure all the elements of its behavior.

L.O. Chua, "Device modeling via basic nonlinear circuit elements," IEEE Trans. on Circuits and Systems, 1980

Device Modeling

separating experience from models

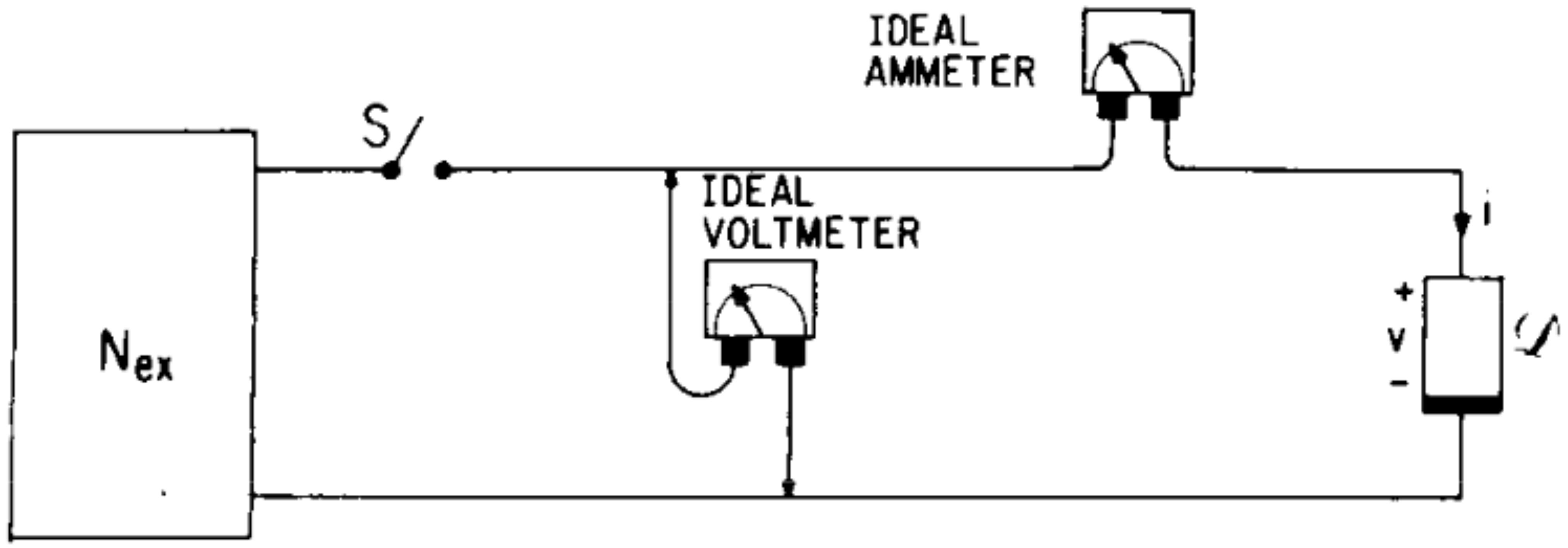


Fig. 1. Circuit for measuring admissible voltage–current signal pairs associated with a 2-terminal or one-port device $\mathscr{D}$.

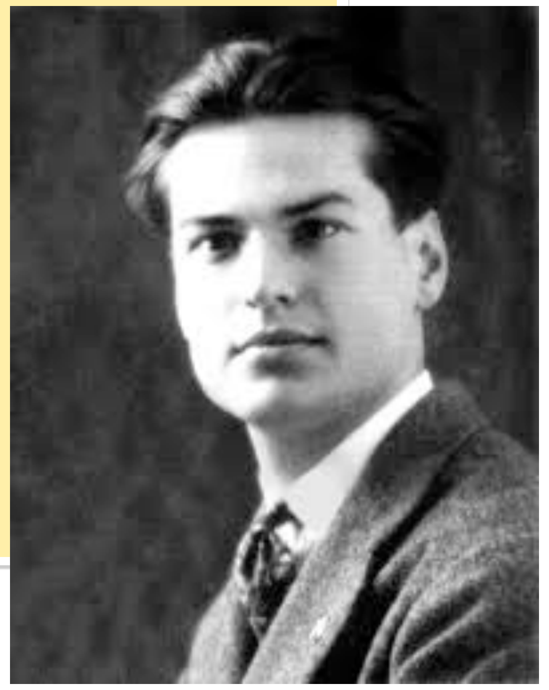
Desiderata for models:

1. Well-posedness
2. Simulation capability
3. Qualitative similarity
4. Predictive ability
5. Structural stability

A model obtained from a finite set of measurements should *generalize* to previously unseen operating conditions (= good simulation capability + good predictive ability).

<ins>But</ins>: Probabilistic assumptions about operating conditions are *highly artificial* and *not verifiable*!

"Probability does not exist." (B. de Finetti)



The Behavioral Approach

“by their fruits you will know them”

- Developed by Jan Willems in 1980s, but with important precursors in general systems theory from 1970s (Gaines, Klir, Mesarovic, Windeknecht, Zadeh, ...)
- Systems = behaviors (extensional description)
- Models = behavioral equations (intensional description)
- Goal of learning/modeling/system identification: proceeding from (samples of) behaviors to models

Generalities

- Basic object: $(W, \mathscr{B})$

  W is the space of outcomes (observable attributes)

  $\mathscr{B} \subseteq W$ is the behavior

- Models via behavioral equations:

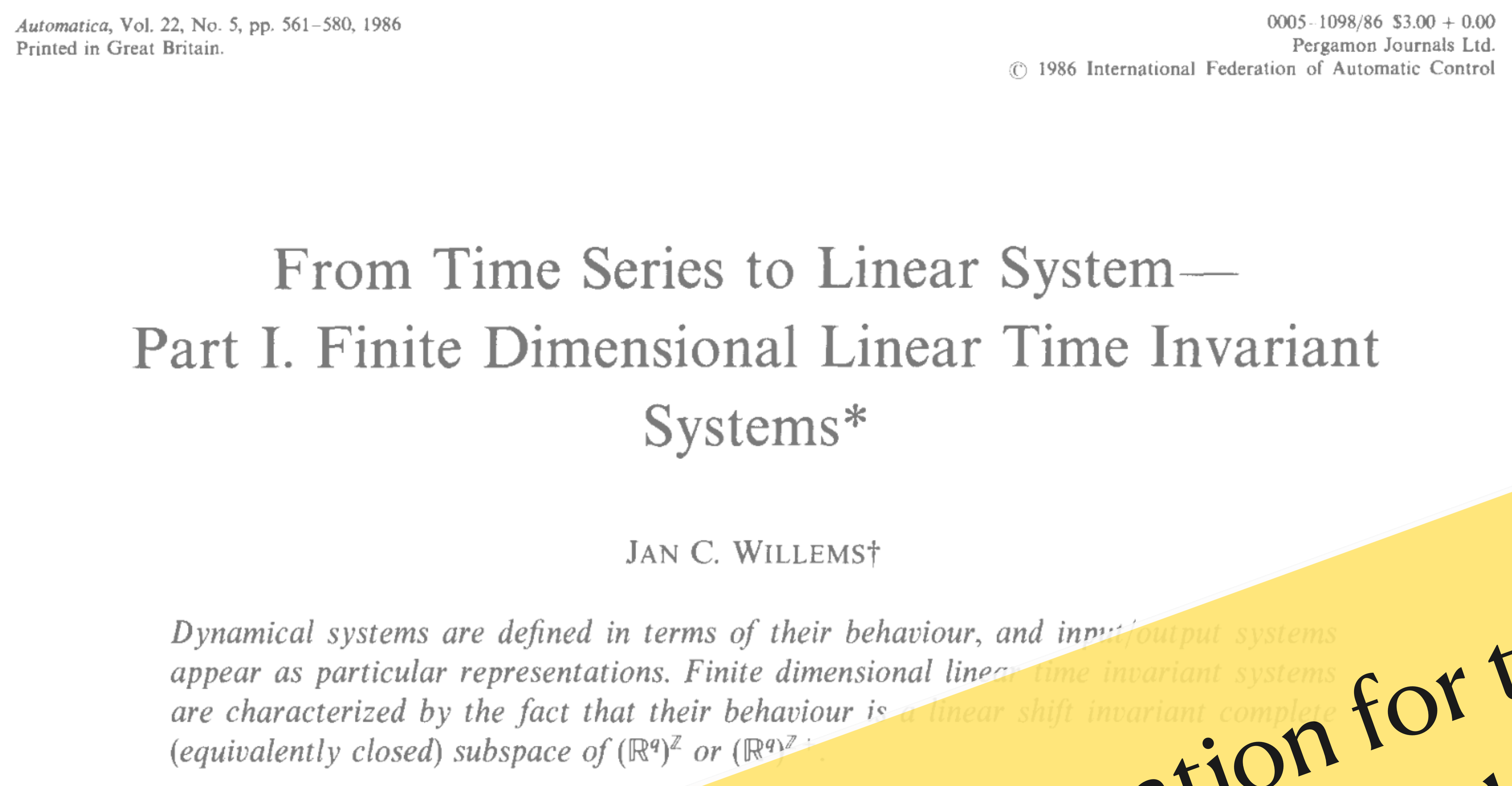
  $w \in \mathscr{B} \Leftrightarrow f(w) = 0$ — kernel representation

  $w \in \mathscr{B} \Leftrightarrow (\exists \lambda)\ w = g(\lambda)$ — image (or latent variable) representation

- *Example*: autoregressive models

  $W = \{(\ldots, w(-1), w(0), w(1), \ldots) : w(i) \in \Sigma\}$

  $w \in \mathscr{B} \Leftrightarrow w(i) = f(w(i-1), w(i-2), \ldots, w(i-L))$ for all i



Automatica, Vol. 22, No. 5, pp. 561–580, 1986
Printed in Great Britain.

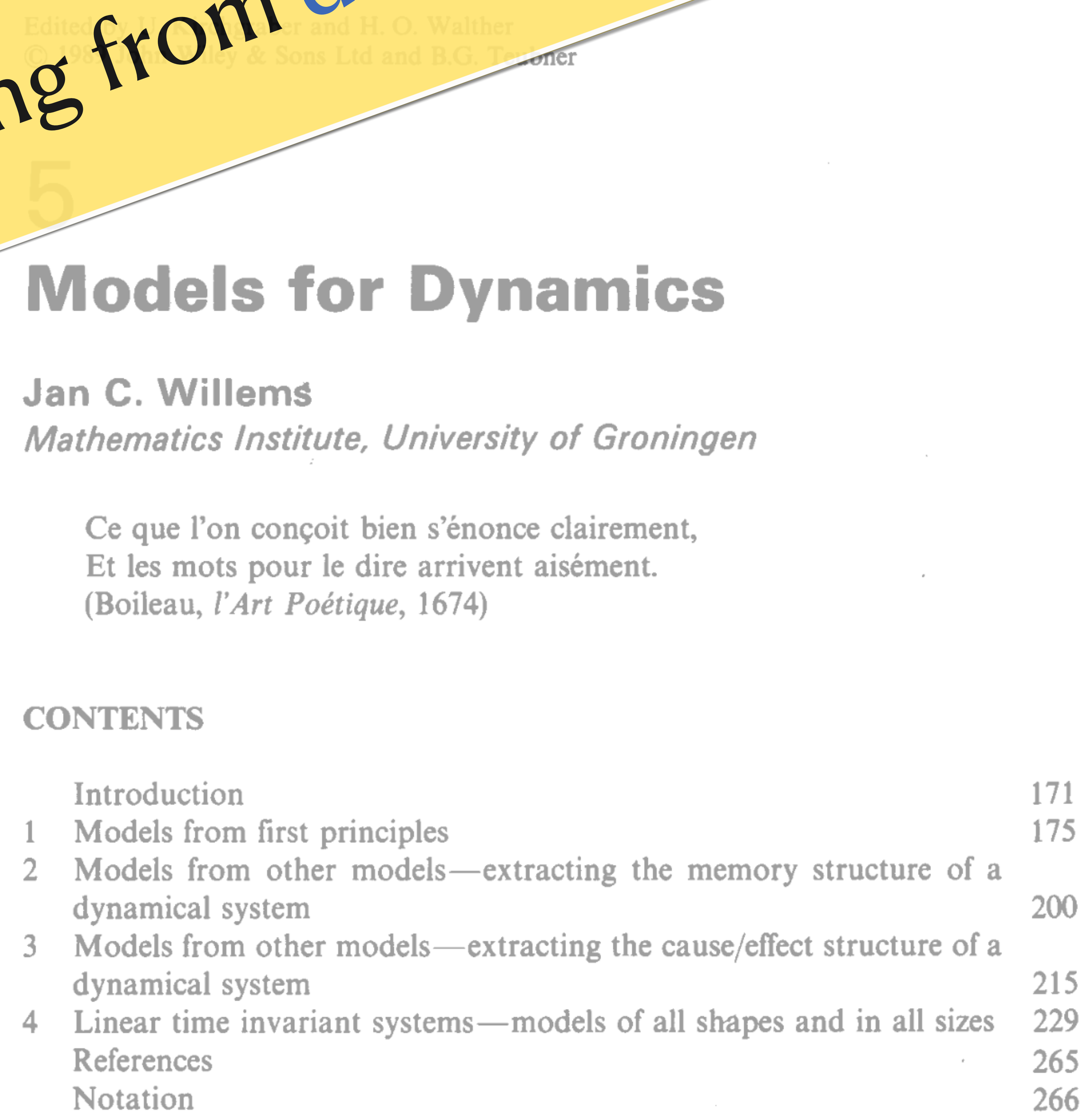
0005-1098/86 $3.00 + 0.00
Pergamon Journals Ltd.
© 1986 International Federation of Automatic Control

From Time Series to Linear System— Part I. Finite Dimensional Linear Time Invariant Systems*

JAN C. WILLEMS†

Dynamical systems are defined in terms of their behaviour, and input/output systems appear as particular representations. Finite dimensional linear time invariant systems are characterized by the fact that their behaviour is a linear shift invariant complete (equivalently closed) subspace of $(\mathbb{R}^q)^{\mathbb{Z}}$ or $(\mathbb{R}^q)^{\mathbb{Z}_+}$ ...

Original motivation for the behavioral approach: data-driven, nonparametric methods for going from data to models



Edited by U. Kirchgraber and H. O. Walther
© 1989 John Wiley & Sons Ltd and B.G. Teubner

5

Models for Dynamics

Jan C. Willems
Mathematics Institute, University of Groningen

Ce que l'on conçoit bien s'énonce clairement,
Et les mots pour le dire arrivent aisément.
(Boileau, *l'Art Poétique*, 1674)

CONTENTS

Modeling

$$\mathscr{B} \xrightarrow{\text{sampling}} \mathscr{B}_{\text{data}} = \{w_i : 1 \leq i \leq n\} \subseteq \mathscr{B} \xrightarrow{\text{modeling}} \hat{\mathscr{B}} = \{w \in W : \hat{f}(w) = 0\}$$

- Inductive bias: choice of $\mathscr{F}$ from which $\hat{f}$ is selected
- Interpolation: $\mathscr{B}_{\text{data}} \subseteq \hat{\mathscr{B}}$

  (Willems says that $\mathscr{B}_{\text{data}}$ is *unfalsified* by $\hat{\mathscr{B}}$)
- Complexity: $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots$

  $\hat{f}_i \in \mathscr{F}_i \mapsto \hat{\mathscr{B}}_i := \{w \in W : \hat{f}_i(w) = 0\}$ such that $\hat{\mathscr{B}}_1 \subseteq \hat{\mathscr{B}}_2 \subseteq \ldots$

Most Powerful Unfalsified Model: $\hat{\mathscr{B}}_{i*}$, where $i* := \min\{i : \mathscr{B}_{\text{data}} \subseteq \hat{\mathscr{B}}_i\}$

Example: 1-NN

$$\mathscr{B} \xrightarrow{\text{sampling}} \mathscr{B}_{\text{data}} = \{w_i : 1 \le i \le n\} \subseteq \mathscr{B} \xrightarrow{\text{modeling}} \hat{\mathscr{B}} = \{w \in W : \hat{f}(w) = 0\}$$

$\mathscr{B} = \{(x, f(x)) : x \in \mathscr{X}\}, \quad f : \mathscr{X} \to \{0,1\}$ unknown

$$\mathscr{B}_{\text{data}} = \{(x_i, y_i) : 1 \le i \le n\}, \qquad y_i = f(x_i)$$

1-NN classifier: $\hat{\mathscr{B}} = \{(x, \hat{f}(x)) : x \in \mathscr{X}\}, \qquad \hat{f}(x) = y_{i(x)}, i(x) = \arg\min_{1 \le i \le n} \operatorname{dist}(x, x_i)$

- $\mathscr{B}_{\text{data}}$ is unfalsified by $\hat{\mathscr{B}}$; this is the most we can say without imposing probabilistic assumptions
- under suitable probabilistic assumptions (which are *not verifiable*), guarantees like Cover & Hart can be given

Example: Interpolating Regression

$$\mathscr{B} \xrightarrow{\text{sampling}} \mathscr{B}_{\text{data}} = \{w_i : 1 \le i \le n\} \subseteq \mathscr{B} \xrightarrow{\text{modeling}} \hat{\mathscr{B}} = \{w \in W : \hat{f}(w) = 0\}$$

$\mathscr{B} = \{(x, f(x)) : x \in \mathbb{R}^d\}, \quad f : \mathbb{R}^d \to \mathbb{R}$ unknown

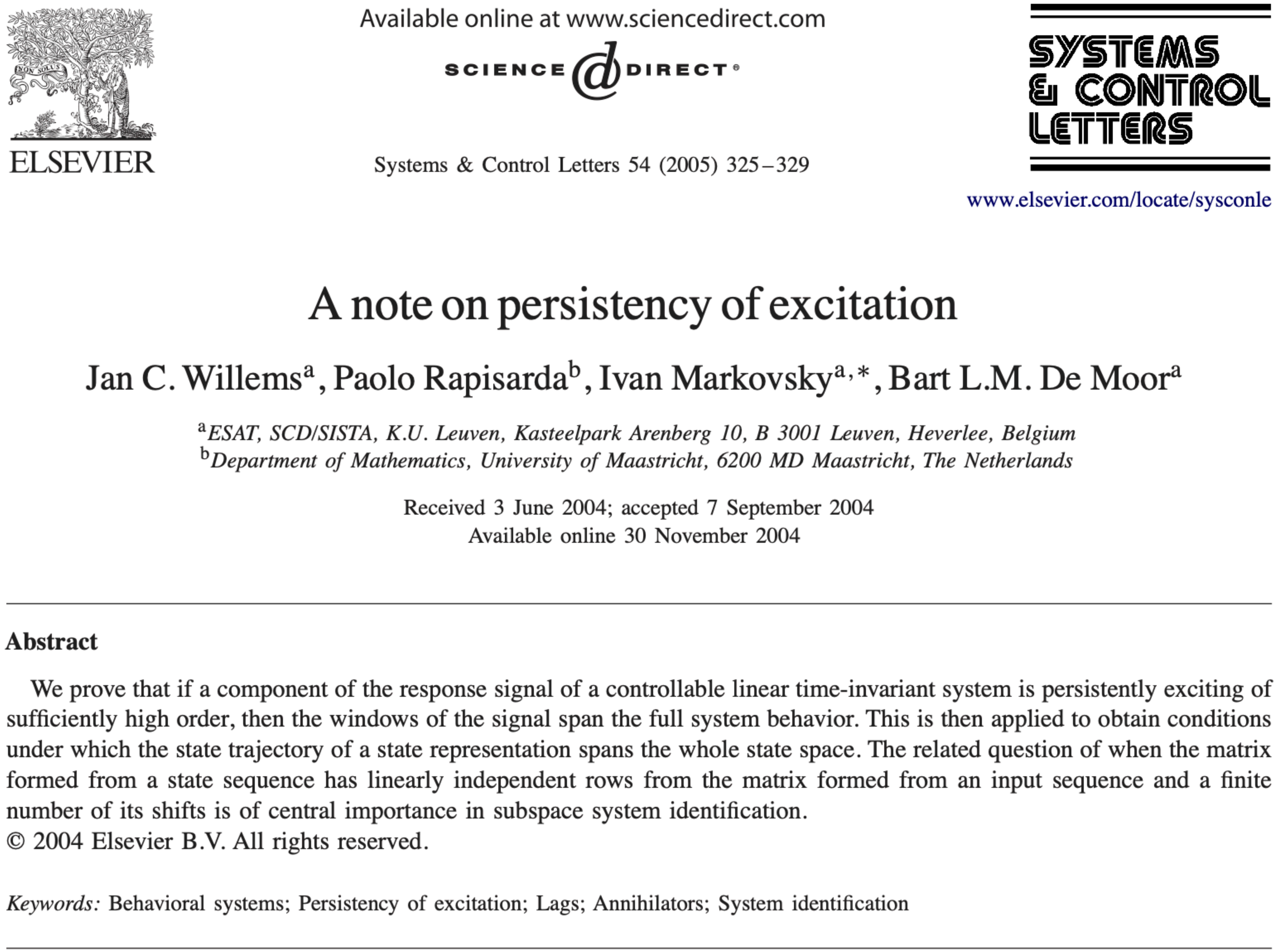
$\mathscr{B}_{\text{data}} = \{(x_i, y_i) : 1 \le i \le n\}, \qquad y_i = f(x_i)$

Nadaraya-Watson estimator: $\hat{\mathscr{B}} = \{(x, \hat{f}(x)) : x \in \mathscr{X}\}, \qquad \hat{f}(x) = \dfrac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$

where K is a singular kernel, i.e., $K(u) \to \infty$ as $u \to 0$

- $\mathscr{B}_{\text{data}}$ is unfalsified by $\hat{\mathscr{B}}$; this is the most we can say without imposing probabilistic assumptions
- under suitable probabilistic assumptions (which are *not verifiable*), we have minimax optimality guarantees (Belkin, Rakhlin, Tsybakov, 2018)

Example from Control: “Fundamental Lemma”



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Systems & Control Letters 54 (2005) 325–329

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

A note on persistency of excitation

Jan C. Willems[a], Paolo Rapisarda[b], Ivan Markovsky[a,*], Bart L.M. De Moor[a]

[a] ESAT, SCD/SISTA, K.U. Leuven, Kasteelpark Arenberg 10, B 3001 Leuven, Heverlee, Belgium
[b] Department of Mathematics, University of Maastricht, 6200 MD Maastricht, The Netherlands

Received 3 June 2004; accepted 7 September 2004
Available online 30 November 2004

Abstract

We prove that if a component of the response signal of a controllable linear time-invariant system is persistently exciting of sufficiently high order, then the windows of the signal span the full system behavior. This is then applied to obtain conditions under which the state trajectory of a state representation spans the whole state space. The related question of when the matrix formed from a state sequence has linearly independent rows from the matrix formed from an input sequence and a finite number of its shifts is of central importance in subspace system identification.

Keywords: Behavioral systems; Persistency of excitation; Lags; Annihilators; System identification

Background: Linear Systems

- Behavioral view of linearity: a system is linear iff its behavior $\mathscr{B}$ is a linear subspace of a vector space
- Time-invariance: $\mathscr{B}$ is invariant with respect to shifts
- *Example*: discrete-time linear time-invariant systems

$$W = \{w = (\ldots, w(-1), w(0), w(1), \ldots) : w(i)) \in \mathbb{R}^q\}$$
$$\mathscr{B} = \{w \in W : P(\sigma)w = 0\}$$

where $\sigma : W \to W$ is the shift operator and

$$P(\sigma) = R_\ell \sigma^\ell + R_{\ell-1}\sigma^{\ell-1} + \ldots + R_1\sigma + R_0, \qquad R_j \in \mathbb{R}^{p\times q}$$

The Fundamental Lemma

an informal statement

- Given a linear time-invariant system which is controllable, the restriction of its behavior $\mathscr{B}$ to sequences of length L can be reconstructed from a *single* observation trajectory $w_{\mathrm{d}} = (w_{\mathrm{d}}(1), \ldots, w_{\mathrm{d}}(T))$ for $T > L$, satisfying a certain *persistency of excitation* condition
- The length-L behavior $\mathscr{B}|_L$ is given by the image (column space) of the Hankel matrix

$$\mathscr{H}_L(w_{\mathrm{d}}) := \begin{pmatrix} w_{\mathrm{d}}(1) & w_{\mathrm{d}}(2) & \ldots & w_{\mathrm{d}}(T-L+1) \\ w_{\mathrm{d}}(2) & w_{\mathrm{d}}(3) & \ldots & w_{\mathrm{d}}(T-L+2) \\ \vdots & \vdots & \ldots & \vdots \\ w_{\mathrm{d}}(L) & w_{\mathrm{d}}(L+1) & \ldots & w_{\mathrm{d}}(T) \end{pmatrix}$$

- RKHS interpretation (recent work by O. Molodchyk and T. Faulwasser)

The Fundamental Lemma

generalization/induction perspective

- Data: $\mathscr{B}_{\text{data}} = \{w_{\text{d}} = (w_{\text{d}}(1), \ldots, w_{\text{d}}(T))\}$ — a single trajectory (designed to have PE)
- Data to model:

$$\hat{\mathscr{B}}|_L = \text{image}\, \mathscr{H}_L(w_{\text{d}}), \qquad \mathscr{H}_L(w_{\text{d}}) := \begin{pmatrix} w_{\text{d}}(1) & w_{\text{d}}(2) & \ldots & w_{\text{d}}(T-L+1) \\ w_{\text{d}}(2) & w_{\text{d}}(3) & \ldots & w_{\text{d}}(T-L+2) \\ \vdots & \vdots & \ldots & \vdots \\ w_{\text{d}}(L) & w_{\text{d}}(L+1) & \ldots & w_{\text{d}}(T) \end{pmatrix}$$

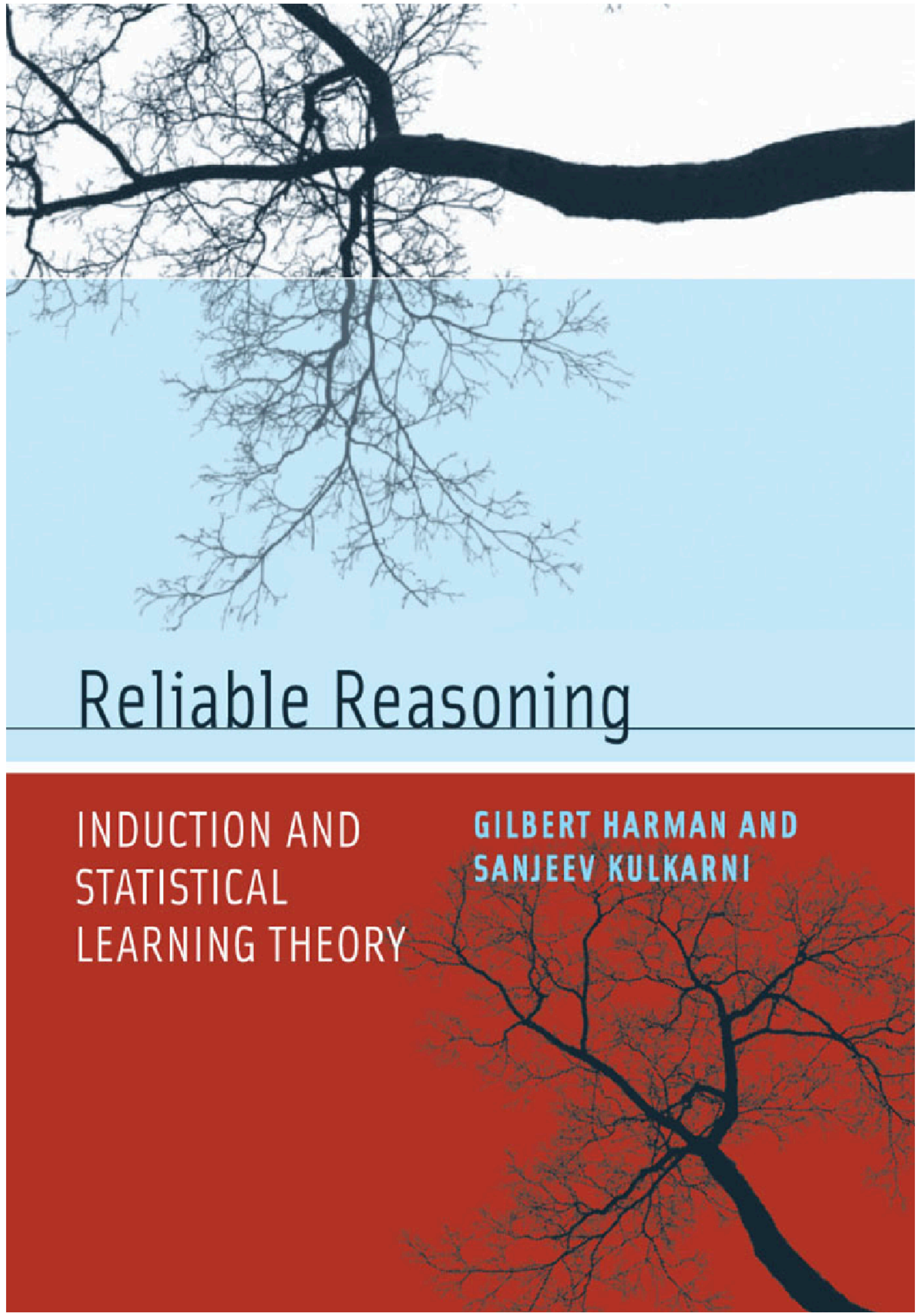
- Reconstruction is *exact*: $\hat{\mathscr{B}}|_L = \mathscr{B}|_L$
- Caveats: **strong** inductive bias (linearity, time-invariance, controllability)
- These assumptions are the *material conditions* for the validity of the inference prescribed by the Fundamental Lemma

The Problem of Induction

you can keep rethinking generalization all you want

- No universal justification that the inference from the particular ($\mathscr{B}_{\text{data}}$) to the general ($\hat{\mathscr{B}}$) is valid as far as $\mathscr{B}$ is concerned
- Any attempt to justify an inductive inference schema is either:
  - fallacious (using a deductive argument to justify an inductive schema)
  - circular (using the same inductive schema to justify itself)
  - or leads to an infinite regress (using meta-induction to justify the original induction, which will require using meta-meta-induction to justify meta-induction, which will require using meta-meta-meta-…)

The Statistical Learning Approach?


Reliable Reasoning
INDUCTION AND STATISTICAL LEARNING THEORY
GILBERT HARMAN AND SANJEEV KULKARNI

- Harman and Kulkarni argue that the “real” problem of induction is a problem of *reliability* of inductive inference
- Theoretical properties of Empirical and/or Structural Risk Minimization are interpreted as reliability certificates
- Inductive bias (choice of hypothesis class) formalizes Nelson Goodman’s notion of *projectible predicates*

This strategy fails for two reasons:

- uses a mathematical deductive argument to justify inductive inference
- relies on empirically unverifiable probabilistic assumptions (infinite regress)

Back to Device Modeling

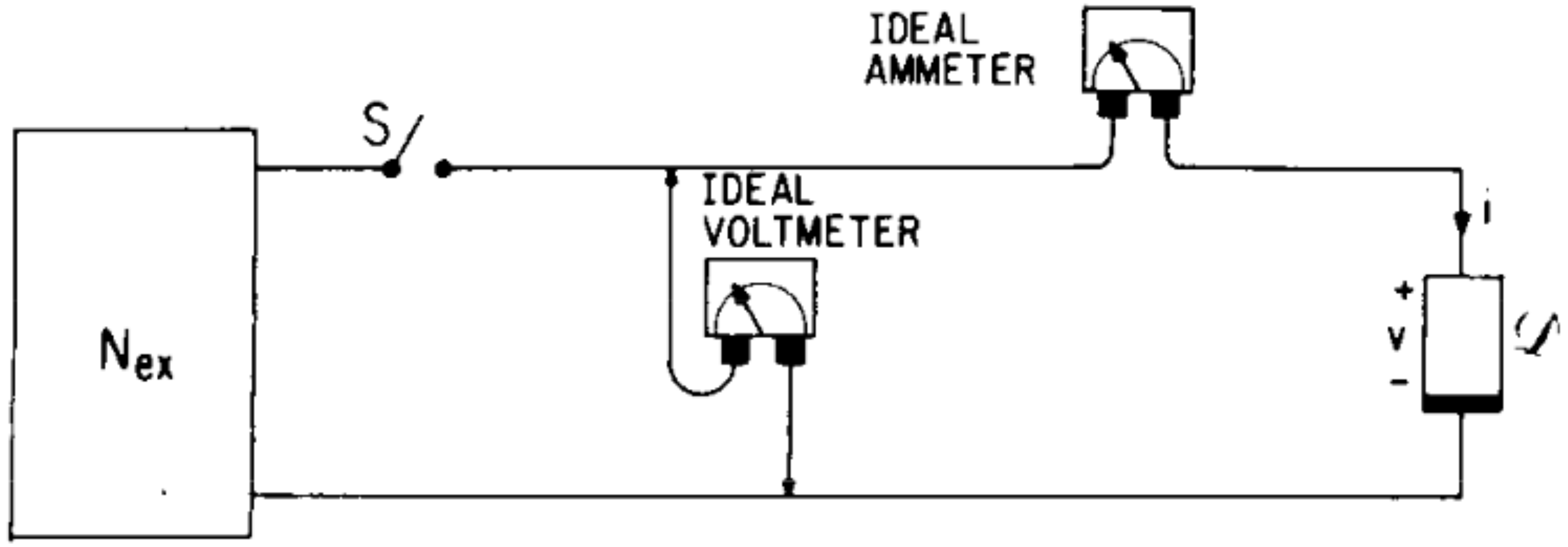
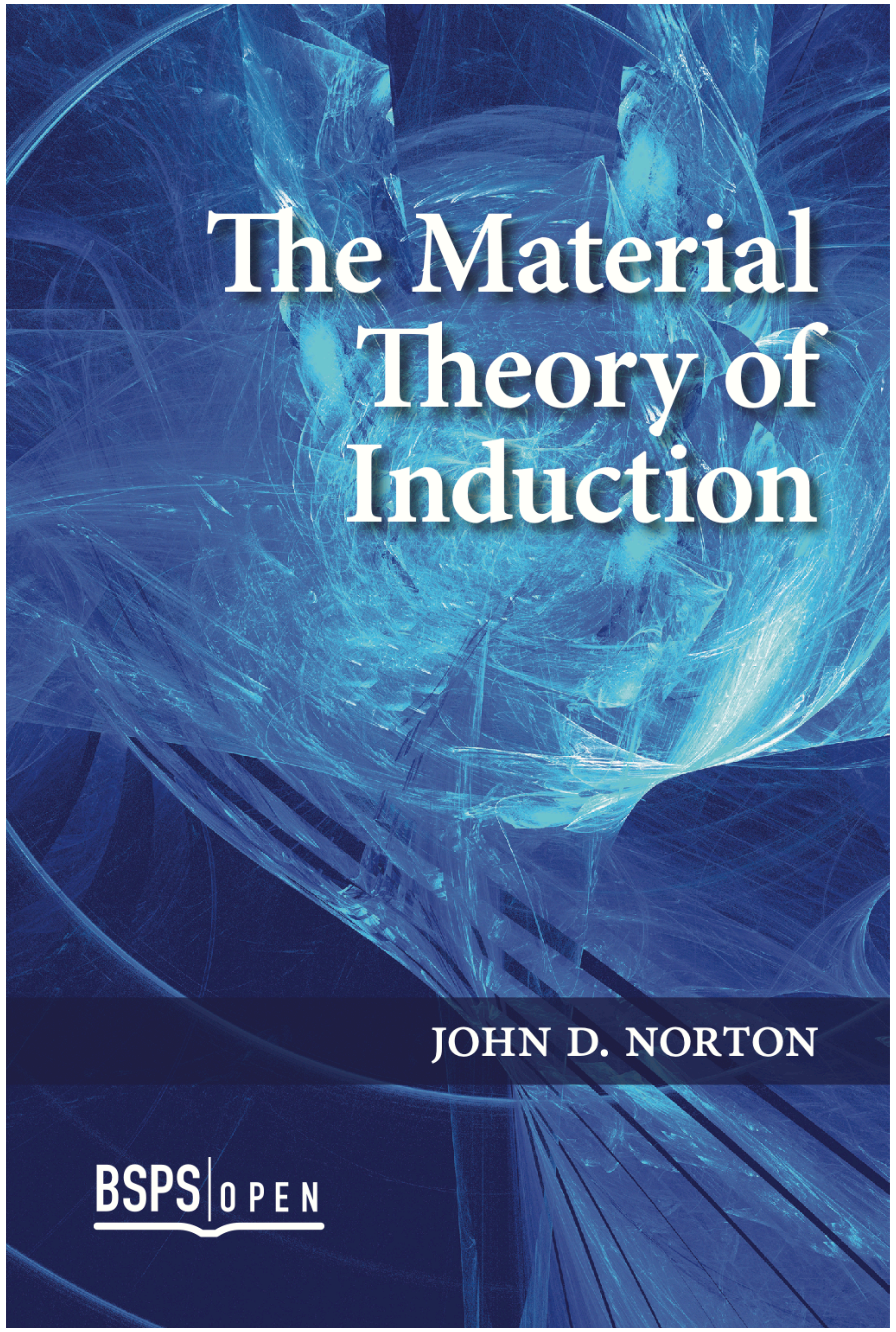


Fig. 1. Circuit for measuring admissible voltage–current signal pairs associated with a 2-terminal or one-port device $\mathscr{D}$.

- Everything comes down to the choice of N_{ex}
- E.g., it could be a "universal signal generator" with tunable parameters θ
- Training data — generate i.i.d. samples $\theta_1, \theta_2, \ldots$ from a well-chosen probability measure $\mathbb{P}$, use each θ_i to drive N_{ex}, collect measurements
- How can we guarantee good coverage?

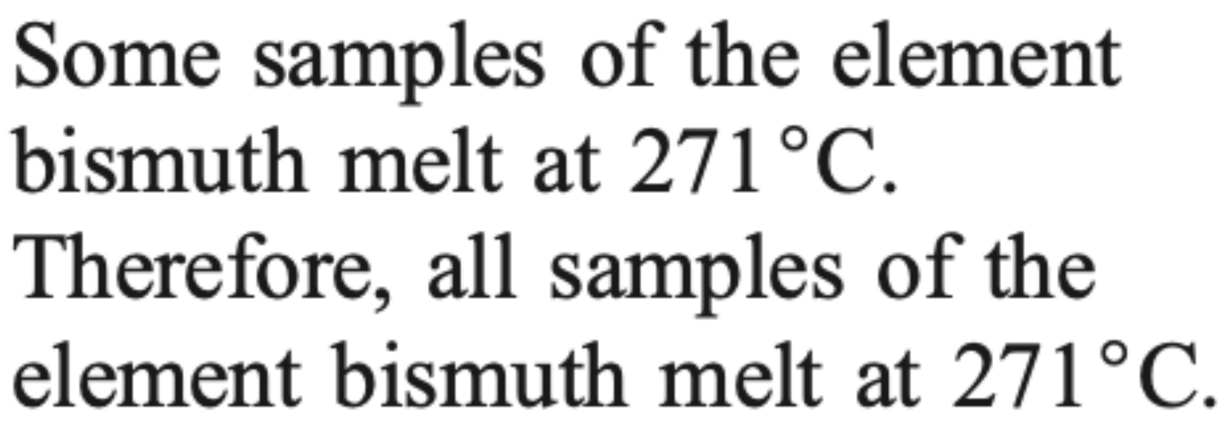
No universal procedure for completely characterizing the device behavior from finitely many measurements!

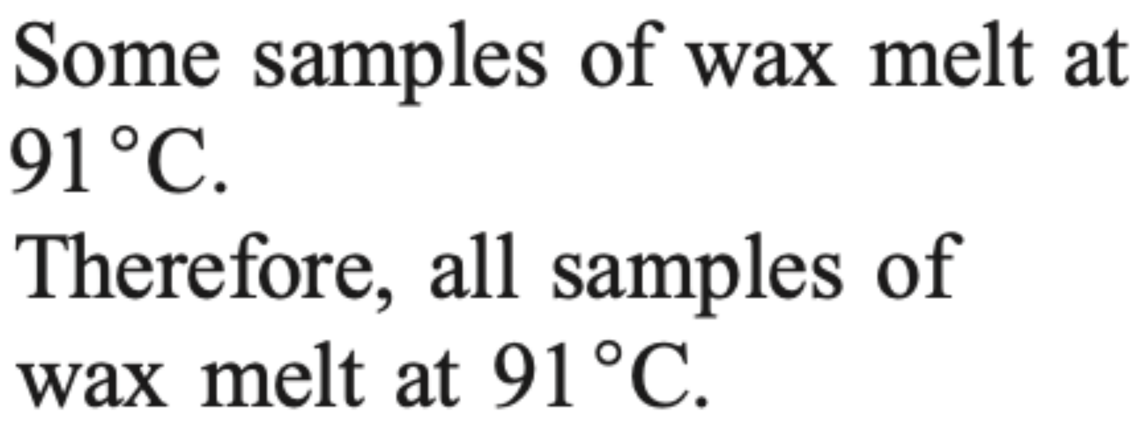
A Better Option: The Material View



Two slogans of Norton:

- all induction is local
- no universal rules of induction


Some samples of the element bismuth melt at 271 °C.
Therefore, all samples of the element bismuth melt at 271 °C.

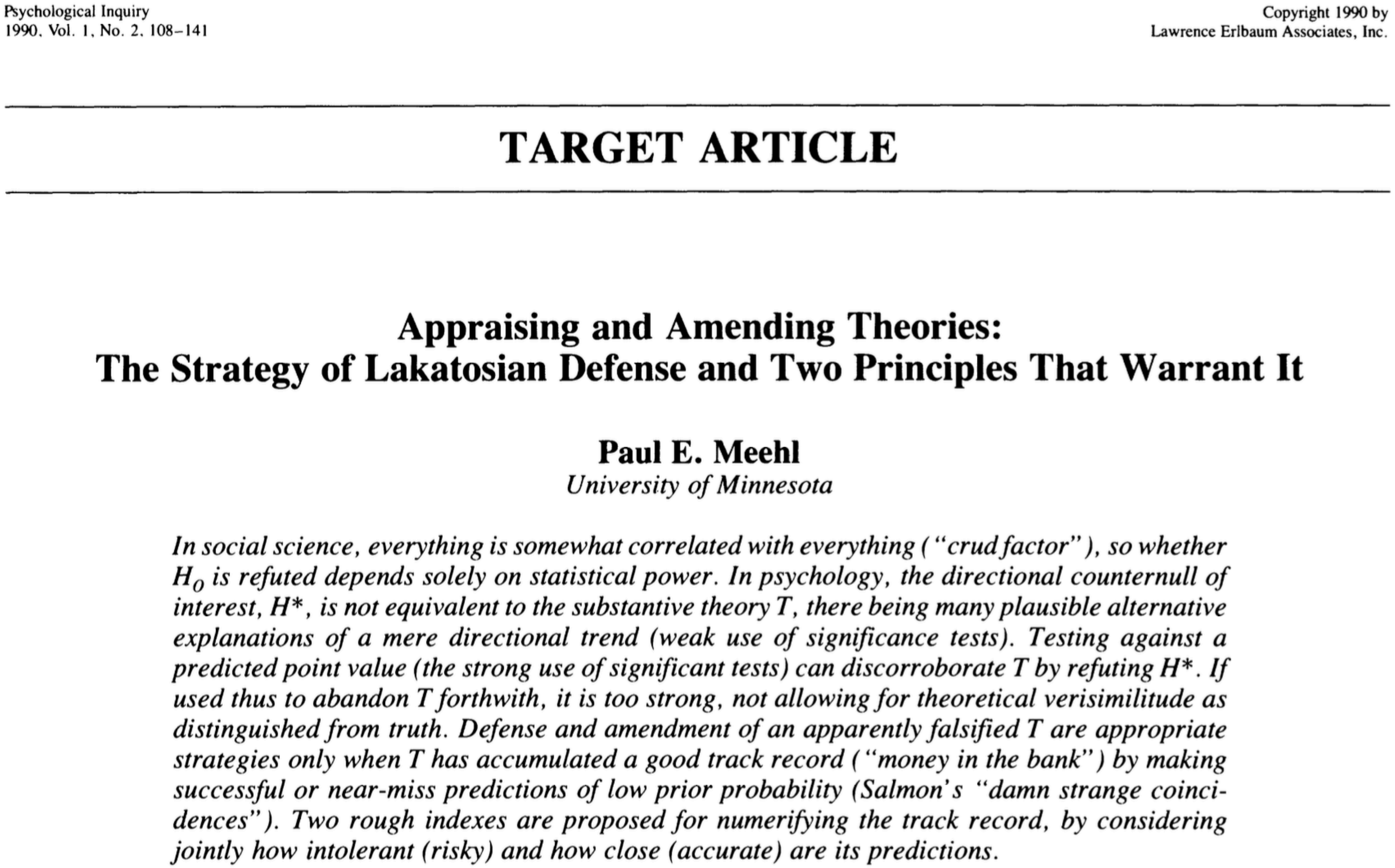

Some samples of wax melt at 91 °C.
Therefore, all samples of wax melt at 91 °C.

Any given inductive argument is warranted by a network of background *material facts* that justify it (provisionally and locally).

The Material Justification of Behavioral Modeling

- What are the relevant background facts?
- In system identification settings, we know from our experience as engineers that many systems in use in signal processing or control admit linear time-invariant approximations under typical (e.g., small-signal) operating conditions.
- In language modeling, we know from our experience (formalized by Shannon and Harris) that “anyone speaking a language possesses, implicitly, an enormous knowledge of the statistics of the language” (reliability of next-token prediction).
- Assumptions of this type are inevitably local and provisional, and have to be checked and re-checked against experience.

Material Theory of Induction, Revisited



Psychological Inquiry
1990, Vol. 1, No. 2, 108–141

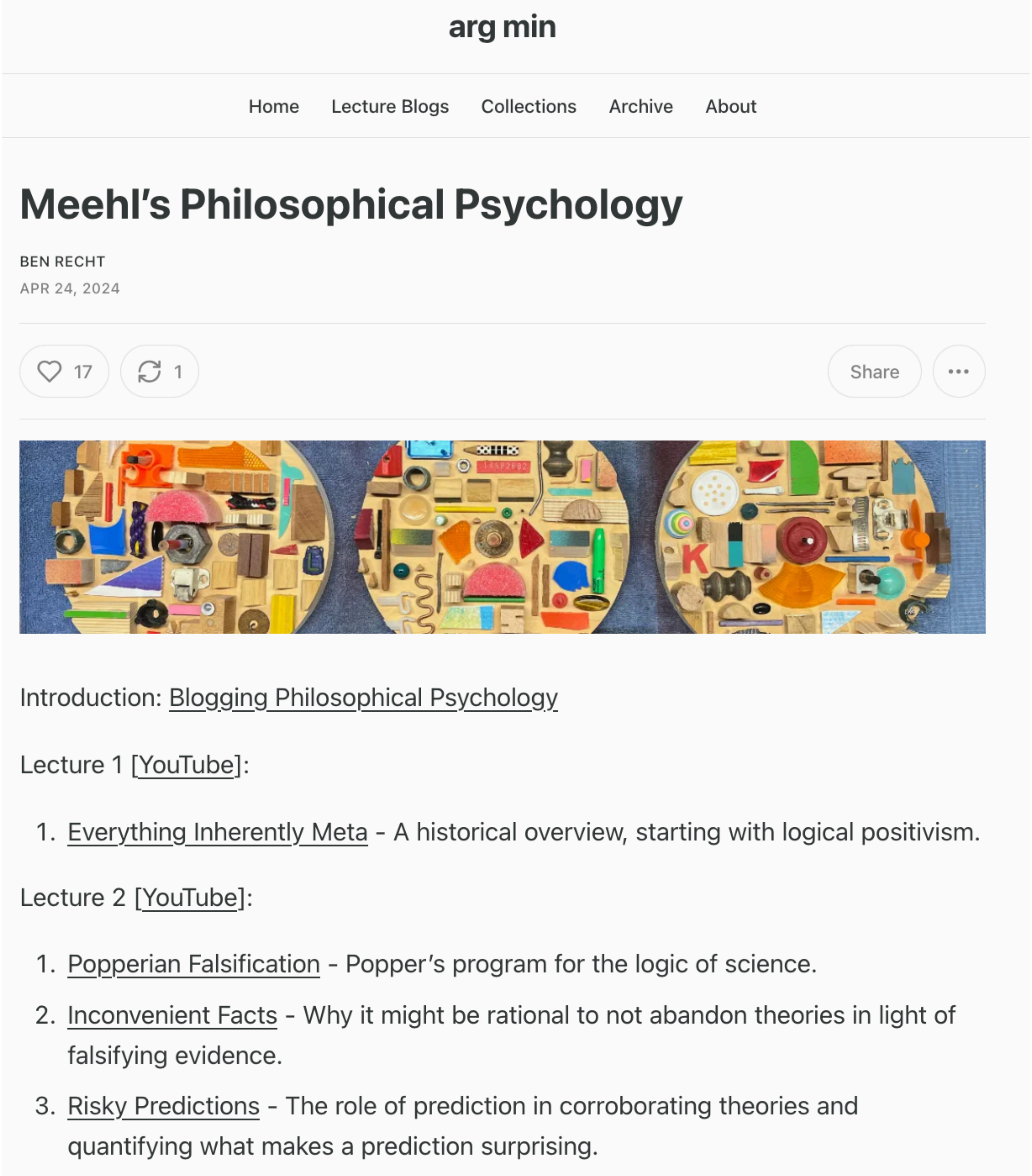
Copyright 1990 by
Lawrence Erlbaum Associates, Inc.

TARGET ARTICLE

Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant It

Paul E. Meehl
University of Minnesota

In social science, everything is somewhat correlated with everything ("crud factor"), so whether H_0 is refuted depends solely on statistical power. In psychology, the directional counternull of interest, H^*, is not equivalent to the substantive theory T, there being many plausible alternative explanations of a mere directional trend (weak use of significance tests). Testing against a predicted point value (the strong use of significant tests) can discorroborate T by refuting H^*. If used thus to abandon T forthwith, it is too strong, not allowing for theoretical verisimilitude as distinguished from truth. Defense and amendment of an apparently falsified T are appropriate strategies only when T has accumulated a good track record ("money in the bank") by making successful or near-miss predictions of low prior probability (Salmon's "damn strange coincidences"). Two rough indexes are proposed for numerifying the track record, by considering jointly how intolerant (risky) and how close (accurate) are its predictions.



Meehl's Philosophical Psychology

BEN RECHT
APR 24, 2024

Introduction: Blogging Philosophical Psychology

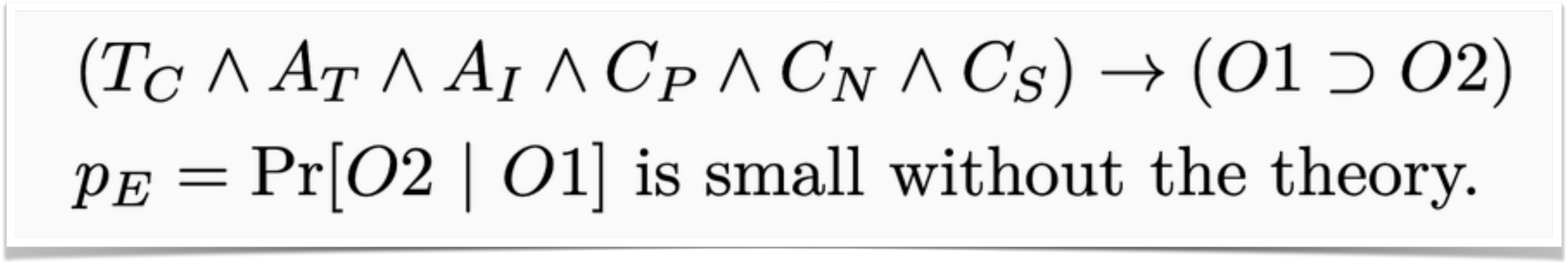
Lecture 1 [YouTube]:

1. Everything Inherently Meta - A historical overview, starting with logical positivism.

Lecture 2 [YouTube]:

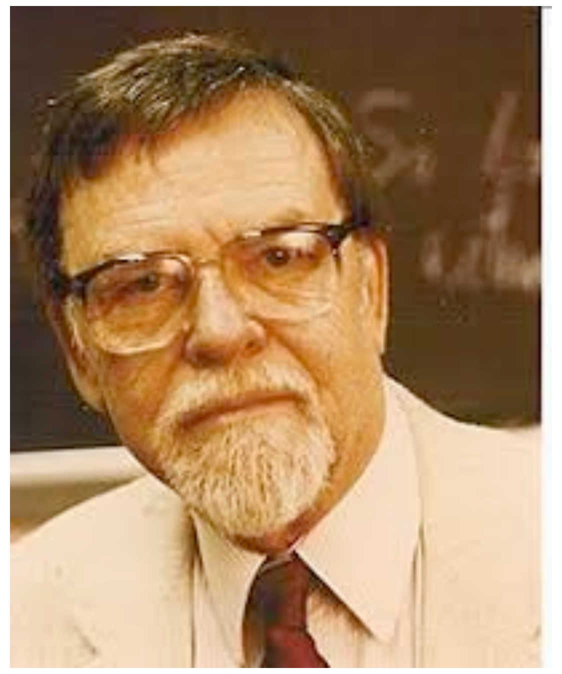
1. Popperian Falsification - Popper's program for the logic of science.
2. Inconvenient Facts - Why it might be rational to not abandon theories in light of falsifying evidence.
3. Risky Predictions - The role of prediction in corroborating theories and quantifying what makes a prediction surprising.

Model-Building as Theory-Building

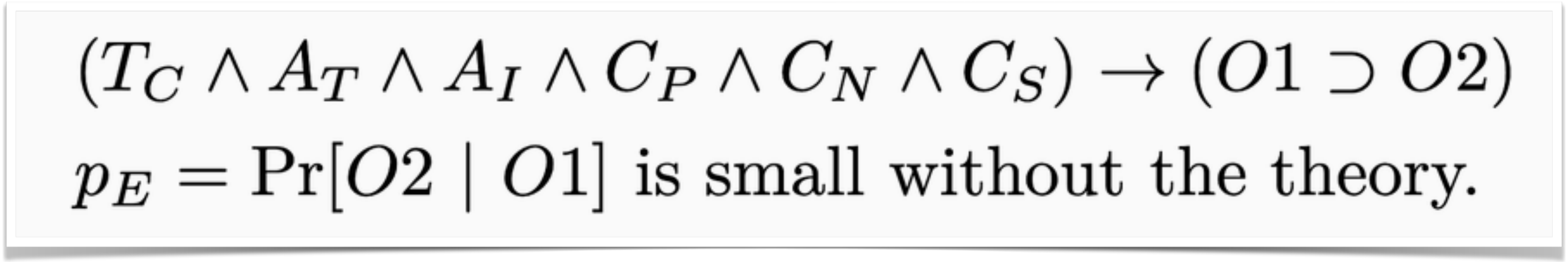


$$(T_C \wedge A_T \wedge A_I \wedge C_P \wedge C_N \wedge C_S) \to (O1 \supset O2)$$

$p_E = \Pr[O2 \mid O1]$ is small without the theory.



- T_C: core theory
- A_T: auxiliary theory
- A_I: auxiliary theories about the instruments
- C_T: ceteris paribus clause (“all else being equal”)
- C_N: particulars about experimental conditions
- C_S : software validity assumption (Ben Recht)



$$(T_C \wedge A_T \wedge A_I \wedge C_P \wedge C_N \wedge C_S) \rightarrow (O1 \supset O2)$$

$p_E = \Pr[O2 \mid O1]$ is small without the theory.

- T_C: core theory
  $\exists f \in \mathscr{F}$ s.t. $\mathscr{B} = \{(i, v) : f(i, v) = 0\}$

- A_T: auxiliary theory
  assumptions like linearity, time-invariance, passivity, etc.

- A_I: auxiliary theories about the instruments
  ammeter, voltmeter are "ideal" for all practical purposes

- C_T: ceteris paribus clause ("all else being equal")
  e.g., experimental stimuli are "similar" to typical use cases

- C_N: particulars about experimental conditions
  temperature, humidity don't matter, instrument calibration

- C_S : software validity assumption
  version of PyTorch used, etc.

$O1$: given $v(t), 0 \leq t \leq T$

$O2 : \hat{f}(i, v) \approx 0$

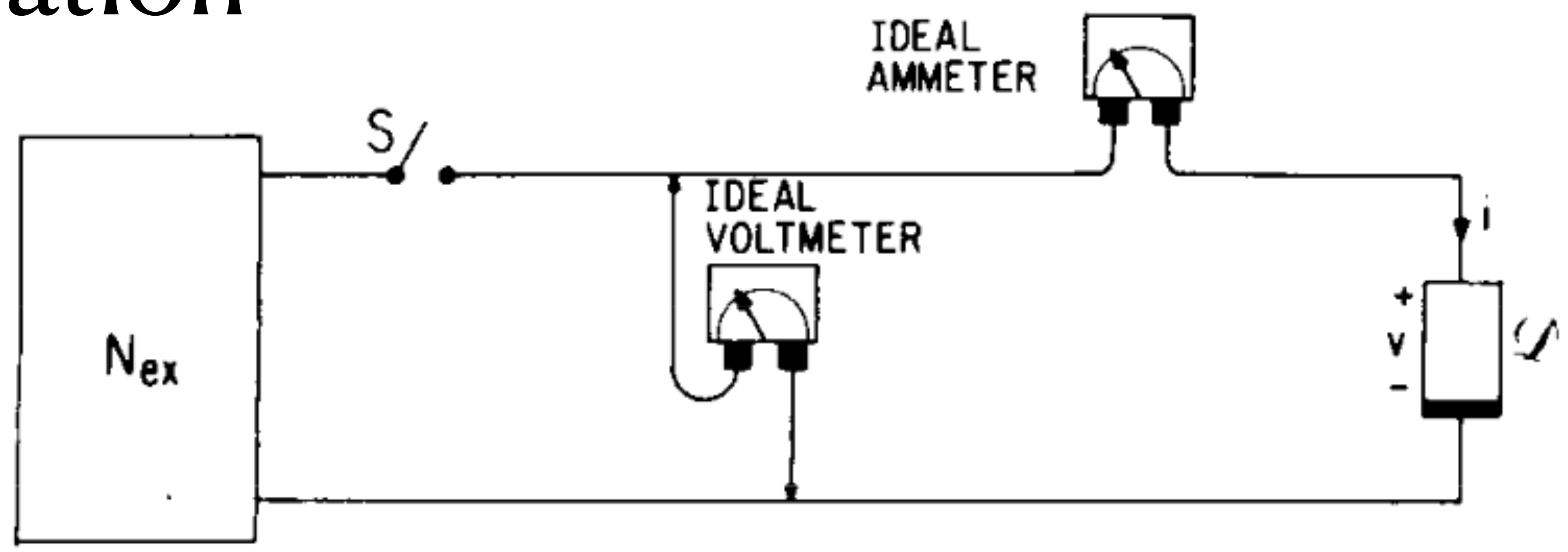
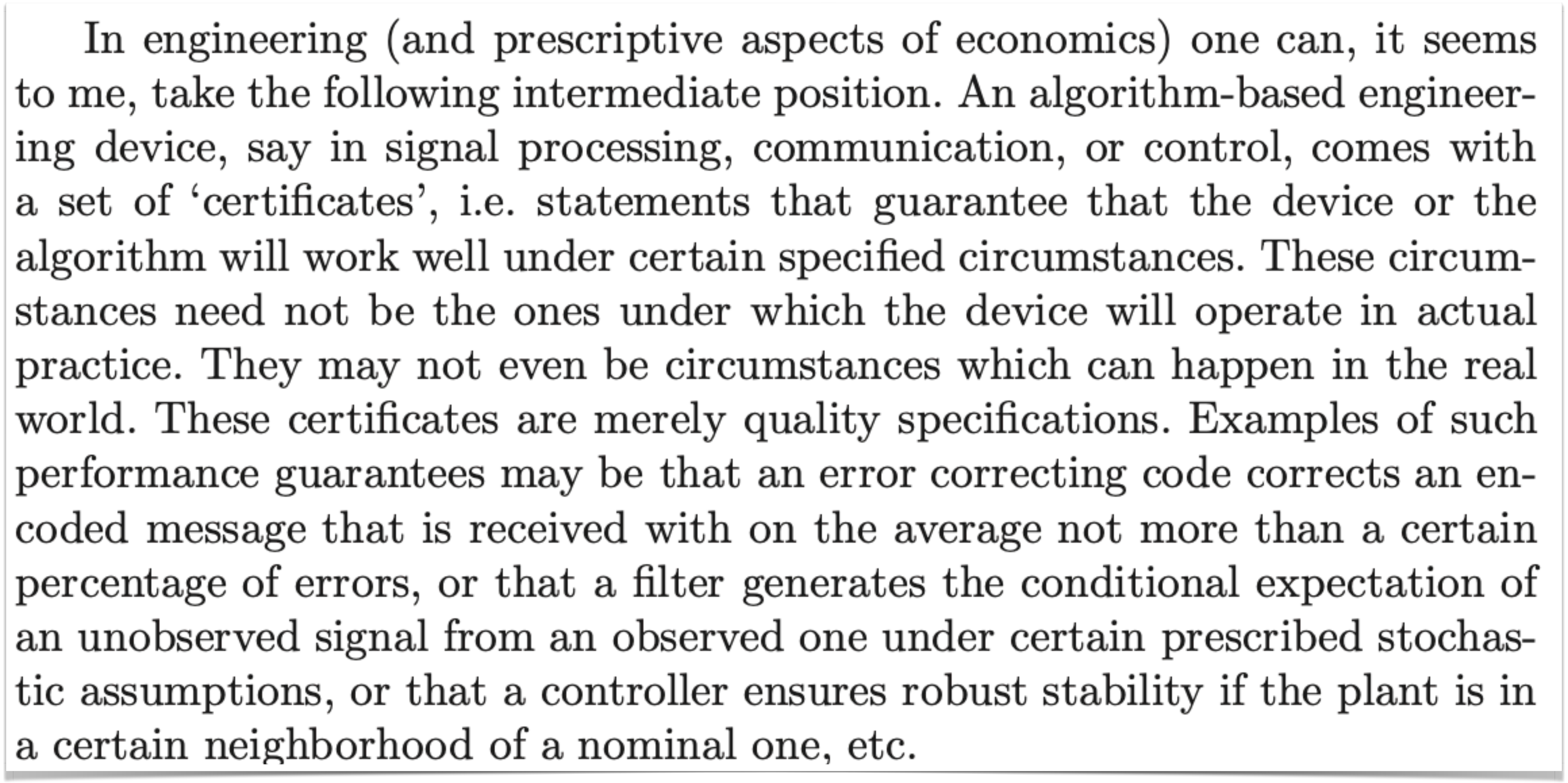


Fig. 1. Circuit for measuring admissible voltage–current signal pairs associated with a 2-terminal or one-port device 𝒟.

The Role of Probabilistic Assumptions



In engineering (and prescriptive aspects of economics) one can, it seems to me, take the following intermediate position. An algorithm-based engineering device, say in signal processing, communication, or control, comes with a set of ‘certificates’, i.e. statements that guarantee that the device or the algorithm will work well under certain specified circumstances. These circumstances need not be the ones under which the device will operate in actual practice. They may not even be circumstances which can happen in the real world. These certificates are merely quality specifications. Examples of such performance guarantees may be that an error correcting code corrects an encoded message that is received with on the average not more than a certain percentage of errors, or that a filter generates the conditional expectation of an unobserved signal from an observed one under certain prescribed stochastic assumptions, or that a controller ensures robust stability if the plant is in a certain neighborhood of a nominal one, etc.

J. Willems, “Thoughts on system identification,” 2006

Summary

- Behavioral view: systems = data, models = compressed descriptions of data
- Understanding generalization requires rethinking probability (hello Ben!)
- Lots of lessons from control and system identification: look for analogues of the "fundamental lemma" in machine learning
- We need a more nuanced modeling philosophy, with clear recognition that:
  - all generalizations are local
  - no universal justification for generalization