FOUNDATIONS OF MACHINE LEARNING SYSTEMS

Robert Williamson

Tübingen AI Center

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

machine learning new perspectives for science

"FOUNDATIONS OF MACHINE LEARNING SYSTEMS"

- It is the title of my chair… chosen to summarise my (planned) work
  - So a good title for a talk!
- I cover one aspect of what I am interested in (the mathematical side)
- I aim to convey *flavour* of work; perhaps a bit different to much ML research

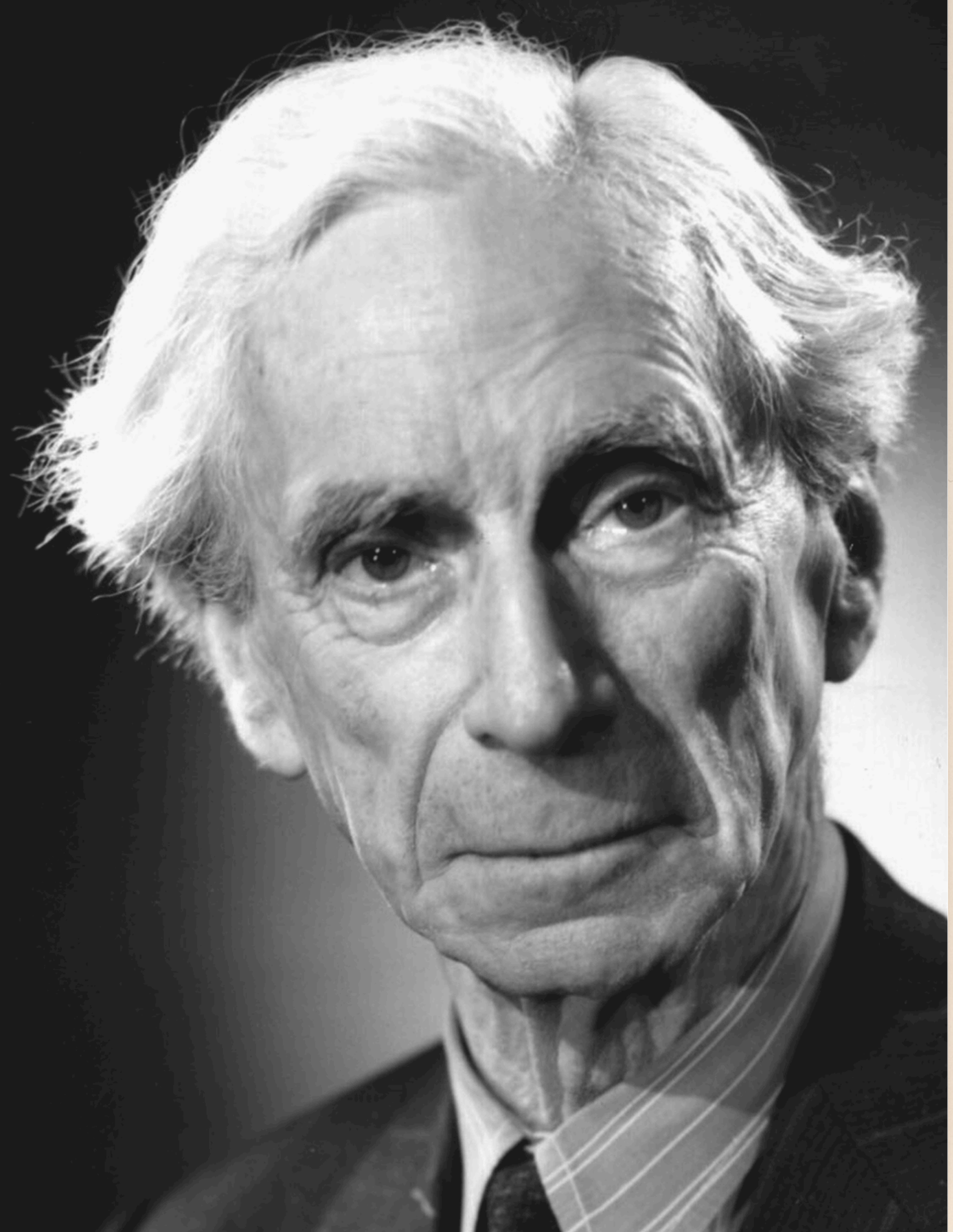
Details are in a few papers I refer to at the end

JOINT WORK WITH TWO PHD STUDENTS

Christian Fröhlich

Rabanus Derr


FOUNDATIONS



***10·2.** $\vdash :. (x) . p \vee \phi x . \equiv : p . \vee . (x) . \phi x$

Dem.

$\vdash . *10{\cdot}1 . *1{\cdot}6 . \supset \vdash :. p . \vee . (x) . \phi x : \supset . p \vee \phi y :.$

[*10·11] $\supset \vdash :. (y) :. p . \vee . (x) . \phi x : \supset . p \vee \phi y :.$

[*10·12] $\supset \vdash :. p . \vee . (x) . \phi x : \supset . (y) . p \vee \phi y$ (1)

$\vdash . *10{\cdot}12 .$ $\supset \vdash :. (y) . p \vee \phi y . \supset : p . \vee . (x) . \phi x$ (2)

$\vdash . (1) . (2) .$ $\supset \vdash .$ Prop.

***10·21.** $\vdash :. (x) . p \supset \phi x . \equiv : p . \supset . (x) . \phi x$ $\left[*10{\cdot}2 \frac{\sim p}{p}\right]$

This proposition is much more used than *10·2.

***10·22.** $\vdash :. (x) . \phi x . \psi x . \equiv : (x) . \phi x : (x) . \psi x$

Dem.

$\vdash . *10{\cdot}1 .$ $\supset \vdash : (x) . \phi x . \psi x . \supset . \phi y . \psi y .$ (1)

[*3·26] $\supset . \phi y :$

[*10·11] $\supset \vdash :. (y) : (x) . \phi x . \psi x . \supset . \phi y :.$

[*10·21] $\supset \vdash :. (x) . \phi x . \psi x . \supset . (y) . \phi y$ (2)

$\vdash . (1) . *3{\cdot}27 .$ $\supset \vdash :. (x) . \phi x . \psi x . \supset . \psi z :.$

[*10·11] $\supset \vdash :. (z) : (x) . \phi x . \psi x . \supset . \psi z :.$

[*10·21] $\supset \vdash :. (x) . \phi x . \psi x . \supset . (z) . \psi z$ (3)

$\vdash . (2) . (3) . \text{Comp} .$ $\supset \vdash :. (x) . \phi x . \psi x . \supset : (y) . \phi y : (z) . \psi z$ (4)

$\vdash . *10{\cdot}14{\cdot}11 .$ $\supset \vdash :. (y) :. (x) . \phi x : (x) . \psi x : \supset . \phi y . \psi y :.$

[*10·21] $\supset \vdash :. (x) . \phi x : (x) . \psi x : \supset . (y) . \phi y . \psi y$ (5)

$\vdash . (4) . (5) .$ $\supset \vdash .$ Prop

The above proposition is true whenever it is significant; but, as was pointed out in connexion with *10·14, it is not always significant when "$(x) . \phi x : (x) . \psi x$" is significant.

***10·22.** [illegible] constitu[illegible] y [illegible] constitu[illegible] ms a constituent χ ([illegible]), where χ is [illegible] $y, \ldots u, v, \ldots$ are either constants or apparent variables, then $\phi\hat{x}$ and $\psi\hat{x}$ take arguments of the same type. This can be proved in each particular case, though not generally, provided that, in obtaining [illegible] and [illegible] from χ, χ is only submitted to negations, disjunctions and generalizations. The proof can be illustrated by an example. Suppose ϕx is $(y) . \chi(x, y) . \supset . \theta x$, and ψx is $fx . \supset . (y) . \chi(x, y)$. By the definitions of *9, ϕx is $(\exists y) . \sim\chi(x, y) \vee \theta x$, and ψx is $(y) . \sim fx \vee \chi(x, y)$. Hence since the primitive ideas $(x) . Fx$ and $(\exists x) . Fx$ only apply to functions, there [illegible] of functions [illegible] $\vee$ [illegible] $(\hat{x})$. Hence there is a proposition [illegible] $(a, b) \vee$ [illegible]. Hence since "$p \vee$ [illegible]" and "p" are only significant

FOUNDATIONS - THE ENGINEER'S VIEW

- "an element of a structure which connects it to the ground" – Wikipedia
- Abstracting just slightly:
  An Interface to the World
- And what happens if your interface does not respect the properties of the world?

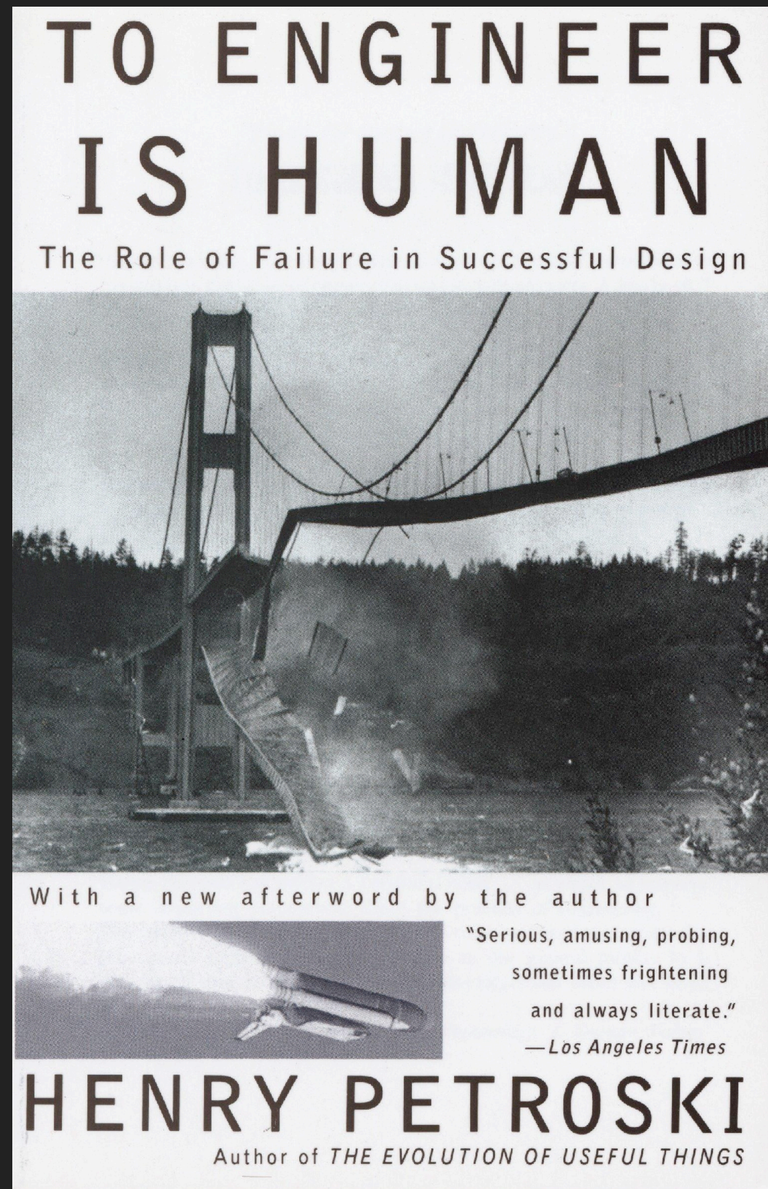
FOUNDATIONS OF MACHINE LEARNING SYSTEMS

- Study the interface of ML systems to the world
- Pay attention to what we assume about the world
- Like other areas of engineering, learn from failures
  - Particular focus: failure of usual *models of data*

Need to pay attention to the data itself …

Sabina Leonelli's remarkable "data journeys"

(Aside: watch her nice talk https://www.youtube.com/watch?v=XYNV1EuHVXk)



MACHINE LEARNING SYSTEMS

PROTOTYPE ML PROBLEM: MINIMISE EXPECTED RISK

$$\underset{h \in \mathcal{Y}^{\mathcal{X}}}{\arg\min} \; \mathbb{E}_{P_{\mathsf{XY}}} \, \ell(\mathsf{Y}, h(\mathsf{X}))$$

- The world gives us random variables X, Y with some fixed joint distribution P_{XY}
  - Actually it gives us samples "drawn i.i.d." from P_{XY}
- The loss function ℓ is given to us (by someone else?)
- Goal: minimise (over h) the expected value of the indexed family of random variables

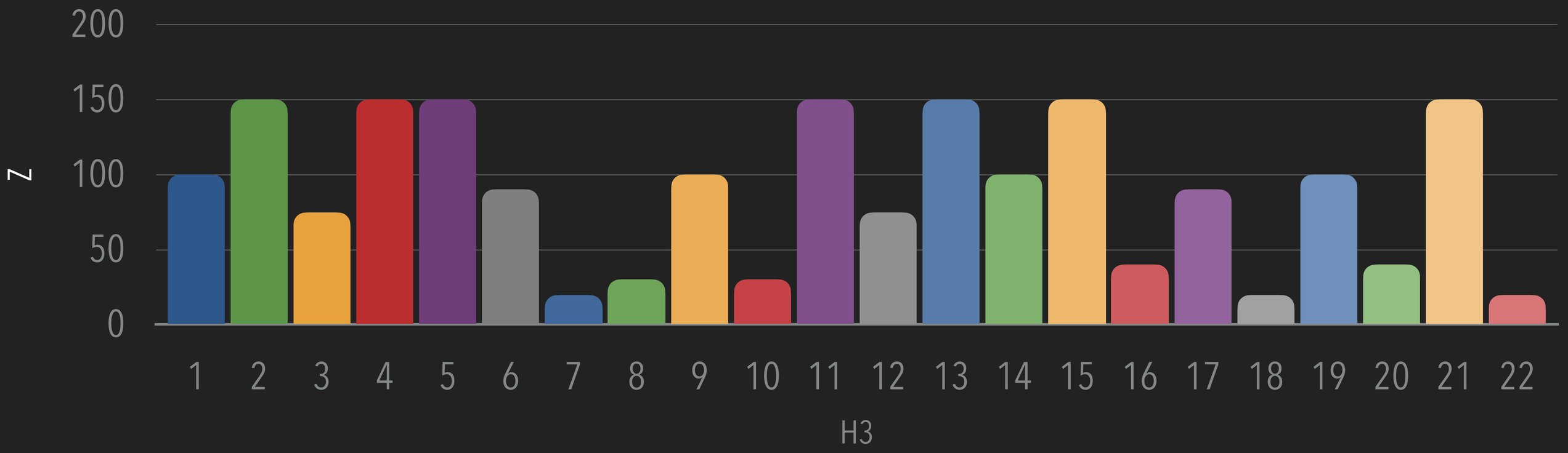
$$\mathsf{R}_h(\omega) := \ell(\mathsf{Y}(\omega), h(\mathsf{X}(\omega)))$$

LOWER ONE'S SIGHTS, AND FIX A MODEL CLASS $\mathcal{H}$

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}\, \ell(\mathsf{Y}, h(\mathsf{X}))$$

$$\arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathsf{Y}_i, h(\mathsf{X}_i))$$

MINIMISING EMPIRICAL RISK



$$\underset{h \in \mathcal{H}}{\arg\min} \; \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(\mathsf{Y}_i, h(\mathsf{X}_i))}_{\mathsf{Z}_i}$$

The question!

WHAT TO STUDY?

$$\underset{h \in \mathcal{H}}{\arg\min}\ \mathbb{E}\ \ell(\mathsf{Y}, h(\mathsf{X}))$$

$\mathbb{E}$ implies an underlying probability space $(\Omega, \mathcal{S}, \mu)$

- $\mathbb{E}$ expectation to aggregate individual losses / encode fairness
- $\mathcal{S}$ the set system (usually a σ-algebra) – the set of measurable events
- Y, X our real model of the world – $(\mathsf{X}_i, \mathsf{Y}_i)$ as iid "samples from a distribution"

$$\underset{h \in \mathcal{H}}{\arg\min}\ \frac{1}{n} \sum_{i=1}^{n} \ell(\mathsf{Y}_i, h(\mathsf{X}_i))$$

LOSS FUNCTIONS AND INFORMATION

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}\, \ell(Y, h(X))$$

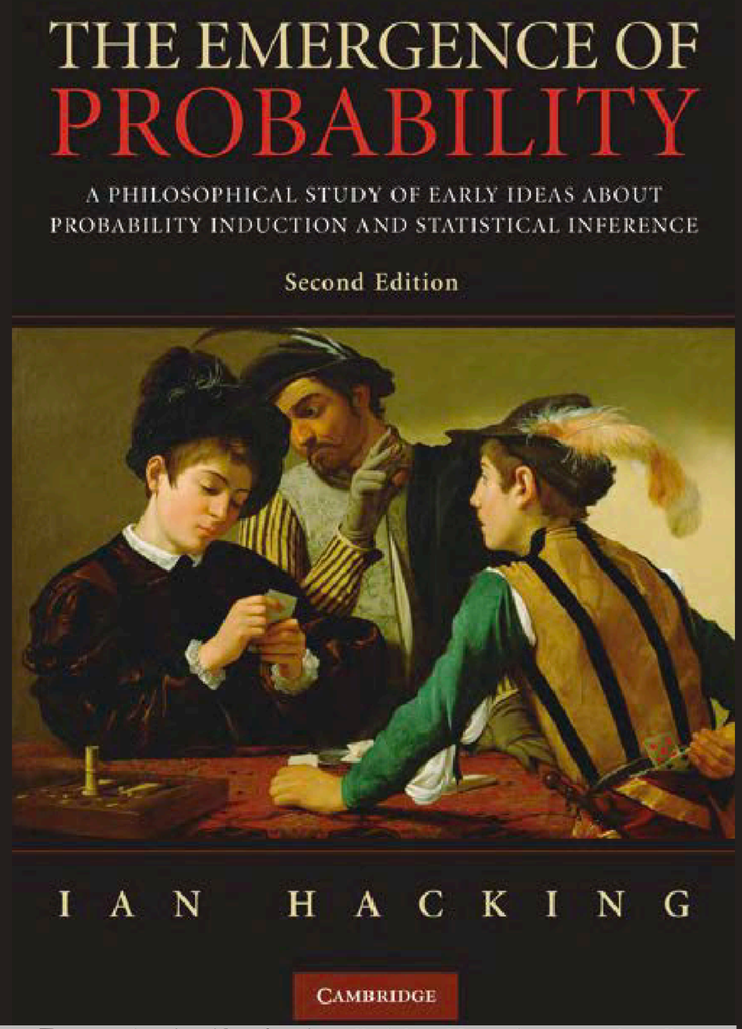
- Loss functions:
  - Traditionally $\ell(y, \hat{p})$; reframe as $\ell : \Delta^n \to \mathbb{R}^n$
  - Parametrise via a convex set $S = \mathrm{spr}(\ell) \subset \mathbb{R}^n$ as $\ell = \partial\sigma_S$, subgradient of support function of S
  - Provides new insights and notions; e.g. "inverse loss" $\ell^\diamond = \sigma_{S^\circ}$; have $\ell \circ \ell^\diamond \circ \ell = \ell$
- Information:
  - Not just Shannon; many information measures satisfy an "information processing inequality"
  - Not just binary $I(P\|Q)$, but a general experiment $E : [n] \to \Delta(\mathcal{X})$ (think multi-class class probability estimation)
  - Parametrise information by a convex set $D \subset \mathbb{R}^n$ or $\mathcal{F} \subset (\mathbb{R}^n)^{\mathcal{X}}$; $I_D(E)$ and $I_{\mathcal{F}}(E)$; averages of support functions
  - Bridge: $\mathrm{BayesRisk}_\ell(E, \pi) = -\, I_{-\pi \cdot \mathrm{spr}(\ell)}(E)$; $\quad \mathrm{BayesRisk}_{\ell, \mathcal{H}}(E, \pi) = -\, I_{\underbrace{-\pi \cdot \ell \circ \mathcal{H}}_{\mathcal{F}}}(E)$
    - Information depends upon *the use to which it will be put*
  - Information Processing *Equality*: $I_D(RE) = I_{RD}(E)$ and $I_{\mathcal{F}}(ET) = I_{T\mathcal{F}}(E)$
    - Information is not so much lost, as *transformed*
- If studying ℓ provides such insight, then what if we make a similar effort to study $\mathbb{E}$? Will convex sets prevail again?


BEYOND EXPECTATIONS

WHEN X, Y REPRESENT PEOPLE

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}\, \ell(Y, h(X))$$

- We are increasingly asked to be able to explain or justify the results of a ML based decision
- Much effort given to justifying $\mathcal{H}$ or $\arg\min$
- But justification of $\mathbb{E}$ is also called for!
- Probability theory was developed as a response to (very artificial) "games of chance".
- It works well there. But why believe the world is a casino?
- Confer the over-emphasis of games in AI research (Go)



Even today justification is called for. Nowadays people tell you that the expectation is the fair price, because if you repeatedly gambled on the same terms, the expectation is your average pay-off. If you paid more than the expectation you would tend to lose, and if you paid less you would tend to make a profit. This is undoubtedly a sound rationale for buying several hundred gross of tickets in successive lotteries, but what if you are going to buy only one ticket in one gamble? Why should long run average profit be the measure of fairness?

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)
3. Uncertainty + Ambiguity (economics)
4. **Robust Bayes** (imprecise prior)
5. Rearrangement invariant norms
6. de Finetti with a bid-ask spread
7. Set systems for probability $(\Omega, \mathcal{S}, \mu)$
8. Generalised frequentism (von Mises)

- The remainder of the talk:
  - Relations between 1, 5 and 6, and how to stratify the choices ($\mathscr{P}$)
  - How generalising $\mathcal{S}$ rather than μ leads to similar destination (7)
  - Can also reach the same destination by *starting with the data*, but not *assuming* probabilities (8)

All of these approaches lead to same object: nonlinear generalised expectation

$$\overline{R}(X) = \sup_{P \in \mathscr{P}} \mathbb{E}_P(X)$$

The nett result: multiple compelling reasons to go "beyond expectations"

RISK MEASURES AND COHERENT UPPER PROBABILITIES

IMPRECISE PROBABILITY

- $\omega \in \Omega$ represents a state of the world
- A gamble is a bounded function $X\colon \Omega \to \mathbb{R}$ yielding an uncertain loss $X(\omega)$ when the state ω is realised
- Gambles form a vector space $\mathscr{L}$
- Posit 4 axioms for the set $\mathscr{D}$ of desirable gambles:

1. $\sup X < 0 \Rightarrow X \in \mathscr{D}$
2. $\inf X > 0 \Rightarrow X \notin \mathscr{D}$
3. $X \in \mathscr{D},\ \lambda \in \mathbb{R}_{>0} \Rightarrow \lambda X \in \mathscr{D}$
4. $X \in \mathscr{D},\ Y \in \mathscr{D} \Rightarrow X + Y \in \mathscr{D}$

- The set $\mathscr{D}$ satisfying 1–4 above is a coherent set of desirable gambles (aka acceptance set)

Statistical Reasoning with Imprecise Probabilities

Peter Walley

COHERENT UPPER PREVISIONS

- The upper prevision of X is $\overline{P}(X) := \inf\{\alpha \in \mathbb{R} : X - \alpha \in \mathscr{D}\}$ – the smallest amount of certain loss α that, when subtracted from the uncertain loss X, makes the resulting gamble desirable (i.e. the certainty equivalent)
- Symmetrically, the lower prevision $\underline{P}(X) = -\overline{P}(-X) = \sup\{\alpha \in \mathbb{R} : \alpha - X \in \mathscr{D}\}$ – the largest certain loss α which we are willing to shoulder in exchange for giving away an uncertain X
- If $\overline{P}$ is defined as above on an $X \in \mathscr{D}$ then it satisfies

  1. $\overline{P}(X) \leq \sup X$
  2. $\overline{P}(\lambda X) = \lambda \overline{P}(X)$
  3. $\overline{P}(X + Y) \leq \overline{P}(X) + \overline{P}(Y)$

  and is said to be a coherent upper prevision

NATURAL EXTENSION AND UPPER PROBABILITIES

- Known that any upper prevision which avoids sure loss dominates at least one linear prevision (point wise)
- Thus $\mathbb{Q} := \{Q\colon\ Q(X) \leq \overline{P}(X) \quad \forall X \in \mathscr{L},\ \ Q \text{ linear}\}$ is non-empty
- Can construct a canonical coherent extension of $\overline{P}$ via $\overline{E} = \sup_{Q \in \mathbb{Q}} Q(X)$ known as the natural extension of $\overline{P}$.
- Upper probabilities arise by taking a prevision of an indicator gamble: $\overline{P}(A) := \overline{P}(\chi_A)$ and lower probabilities similarly: $\underline{P}(A) := 1 - \overline{P}(\chi_{\Omega \setminus A})$

COHERENT RISK MEASURES

- Yes, the same word used by different communities…
  - Who doesn't want to be coherent?
- Treat $X(\omega)$ and $Y(\omega)$as a random variables on some probability space $(\Omega, \mathcal{S}, \mu)$
  - $R(\lambda X) = \lambda R(X), \forall \lambda \in \mathbb{R}^+$ (positive homogeneity)
  - $R(X + Y) \leq R(X) + R(Y)$ (subadditivity)
  - $R(X + c) = R(X) + c, \forall c \in \mathbb{R}$ (translation equivariance)
  - $X(\omega) \leq Y(\omega)\ \forall \omega \Rightarrow R(X) \leq R(Y)$ (monotonicity)

COHERENT UPPER PREVISIONS = COHERENT RISK MEASURES

- In standard probability theory, there is a 1:1 correspondence between probabilities and expectations via Lebesgue integration
- However with imprecise probability there are many coherent upper previsions associated with a coherent upper probability
  - Hence Walley focussed upon coherent upper previsions
- Pelessoni and Vicig (2003):
  - Let $\mathscr{L}$ be a linear space of bounded real-valued random variables, containing all constant functions.
  - A functional R is a coherent risk measure on $\mathscr{L}$ if and only if it is a coherent upper prevision on $\mathscr{L}$

FROM UPPER PROBABILITIES TO SPECTRAL RISK MEASURES

- Walley does not start with a measure space but just a set of possibilities Ω; one can instead start with a probability space
- Then one can introduce the notion of law invariance – $R(X)$ only depends upon $\mathrm{Law}(X)$
- Spectral risk measures: Given a probability measure λ on $[0,1]$, define

$$R_\lambda(X) = \int_0^1 \mathrm{CVaR}_\alpha(X)\mathrm{d}\lambda(\alpha) \qquad \mathrm{CVar}_\alpha(X) := \mathbb{E}(X \,|\, X \geq q_\alpha(X))$$ (Conditional value-at-risk)

- Convenient to rewrite as $R_w(X) = \int_0^1 F_X^{-1}(q)w(q)\mathrm{d}q$
- Distortion risk measures: Given $\phi\colon [0,1] \to [0,1]$ which is monotonically increasing, concave and satisfiers $\phi(0) = 0$ and $\phi(1) = 1$ the distortion risk measure is given by the Choquet integral

$$R_\phi(X) = \int_{-\infty}^0 [\phi(S_X(x) - 1]\mathrm{d}x + \int_0^\infty \phi(S_X(x))\mathrm{d}x,$$ where $S_X(x) = 1 - F_X(x) = P(X > x)$

- Gzyl and Mayoral (2008): $R_W = R_\phi$ when $\phi'(t) = w(1-t)$

CONNECTION TO UPPER PROBABILITIES

- Work on a measurable space $(\Omega, \mathscr{F})$
- Given a distortion ϕ define $\overline{P}(A) = \phi(P(A))$.
  - This forms a capacity – a set function $\overline{\mu}\colon \mathscr{F} \to \mathbb{R}$ satisfying $\overline{\mu}(\varnothing) = 0$, $\overline{\mu}(\Omega) = 1$ and $A \subset B \Rightarrow \overline{\mu}(A) \leq \overline{\mu}(B)$.
- A submodular capacity also satisfies $\overline{\mu}(A \cup B) + \overline{\mu}(A \cap B) \leq \overline{\mu}(A) + \overline{\mu}(B)$ for all $A, B \in \mathscr{F}$
- It turns out (see paper) that for submodular capacities, Walley's natural extension corresponds with the Choquet integral.
- Thus spectral risk measures are natural extensions of coherent upper probabilities

THE FUNDAMENTAL FUNCTION

- Rearrangement invariant function norms R: just think of functionals of random variables that only depend upon their distribution
- For each measurable subset $E \subseteq \Omega$ with measure $\mu(E) = t$, we define the fundamental function $\phi_{\mathscr{R}} : [0,1] \to \mathbb{R}^+$ as $\phi_{\mathscr{R}}(t) = \|\chi_E\|_{\mathscr{R}} = R(\chi_E)$, where χ_E is the indicator function of E
- Fundamental functions are always "quasi-concave":
  - ϕ is non-decreasing and $\phi(0) = 0$
  - $t \mapsto \phi(t)/t$ is decreasing
  - ϕ is continuous except perhaps at the origin
- Little is lost if one demands concavity (and increasingness and anchored at 0)
- e.g. CVaR$_\alpha$: $\phi_\alpha(t) := \frac{t}{1-\alpha} \wedge 1$ for $\alpha \in [0,1)$

THE LORENTZ AND MARCINKIEWICZ NORMS

- Given any concave fundamental function ϕ, the Lorentz norm of X is
$$\|X\|_{\Lambda_\phi} = X^*(0)\phi(0+) + \int_0^1 F_{|X|}^{-1}(1-\omega)\phi'(\omega)\mathrm{d}\omega,$$
- The Marcinkiewciz norm is $\|X\|_{M_\phi} = \sup_{0<t\leq 1} \left\{\phi(t)\mathrm{CVaR}_{1-t}(|X|)\right\}$
- Known result: For all concave fundamental functions ϕ, all X, and all ri norms $\|\cdot\|_{\mathscr{R}}$ we have $\|X\|_{M_\phi} \leq \|X\|_{\mathscr{R}} \leq \|X\|_{\Lambda_\phi}$

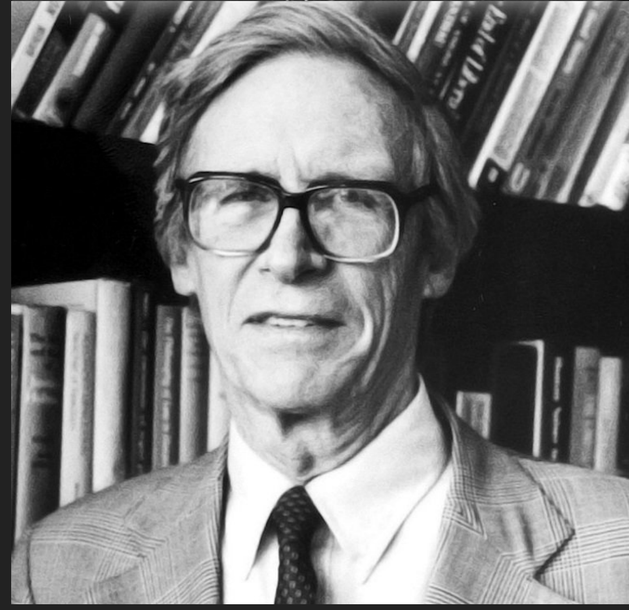
THE FUNDAMENTAL FUNCTION CONTROLS TAIL SENSITIVITY

- Recall $\|X\|_{M_\phi} \leq \|X\|_{\mathscr{R}} \leq \|X\|_{\Lambda_\phi}$
- If additionally we have $\phi'(0) < \infty$ then $\|X\|_{\Lambda_\phi} \leq K \cdot \|X\|_{M_\phi}$ with $K = \frac{1}{\phi(1/\phi'(0))}$.
  - Thus when $\phi'(0) < \infty$, all ri norms are equivalent
  - When $\phi'(0) = \infty$, $\|\cdot\|_{\Lambda_\phi}$ and $\|\cdot\|_{M_\phi}$ are not equivalent

FUNDAMENTAL FUNCTIONS AND UPPER PROBABILITIES

- Recall the definition of the fundamental function: $\phi_{\mathscr{R}}(t) = \|\chi_{[0,t]}\|_{\mathscr{R}}$
- This is the upper probability of the set $[0,t]$
- Provides new insight to coherent upper probabilities

ETHICAL INTERPRETATION



- Ethical interpretation: interpret as how ethically bad for a fraction t of the population to suffer loss 1.
  - $\mathbb{E}(X)$ corresponds to $\phi(t) = t$;
  - $\max X$ corresponds to $\phi(t) = \chi_{(0,1]}(t)$.
  - Harsanyi vs Rawls… (see *Fairness Risk Measures* for an elucidation)
- But there's still a problem: if you choose ϕ_1 and I choose ϕ_2, how might we come to some agreement about a compromise ϕ somehow "in between" ϕ_1 and ϕ_2? Want $\|\cdot\|_{\text{compr}} = \mathfrak{F}(\|\cdot\|_1, \|\cdot\|_2)$
  - Turns out that suitable functors $\mathfrak{F}$ also have "fundamental functions" $\psi_{\mathfrak{F}}$ and $\phi_{\text{compr}} = \psi_{\mathfrak{F}}(\phi_1, \phi_2)$, and $\{\psi_{\mathfrak{F}}\} \cong \{\phi\}$
  - The hard ethical choice remains – there is no cheap fix by going meta…

SYSTEM OF PRECISION

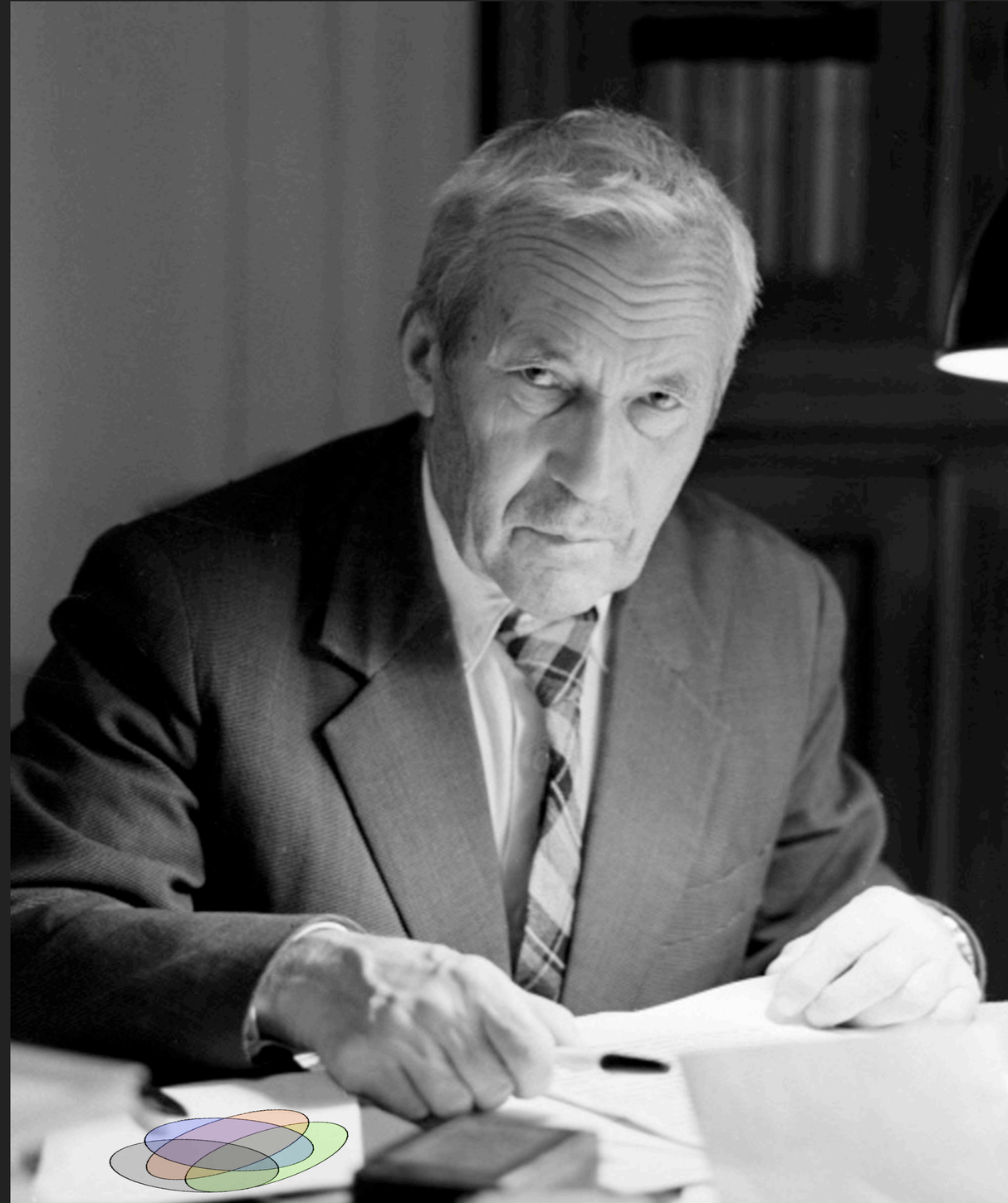
EVENTS $(\Omega, \mathcal{S}, \mu)$

When posing problems in probability calculus, it should be required to indicate for which events the probabilities are assumed to exist

– Andrei Nikolaevich Kolmogorov 1927

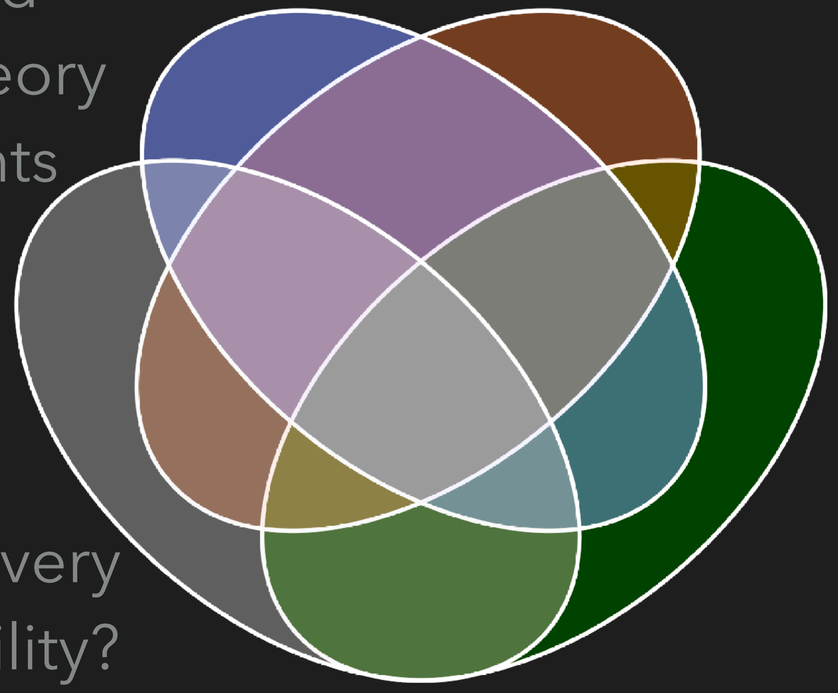
Traditionally $\mathcal{S}$ is taken to be a σ-algebra – a set system closed under complementation and arbitrary countable unions

This is rarely questioned (apart from the σ)



PRE-DYNKIN SYSTEMS

- It turns out that the "natural" set system to consider for precise events is a pre-Dynkin system which is only closed under finite disjoint unions
- Imprecise probabilities are precise on a pre-Dynkin system
  - Hence "the system of precision"
- When one defines a measure on a pre-Dynkin system and extends it to an algebra, one essentially recreates the theory of coherent upper previsions – newly "measurable" events have an imprecise probability
- Allows one to model intersectionality – think of what it means that all intersections of events have probabilities when one tries to reason with finite data – why assume every compound event should be ascribed a (precise) probability?



NON-STOCHASTIC SEQUENCES

DATA AS FACT

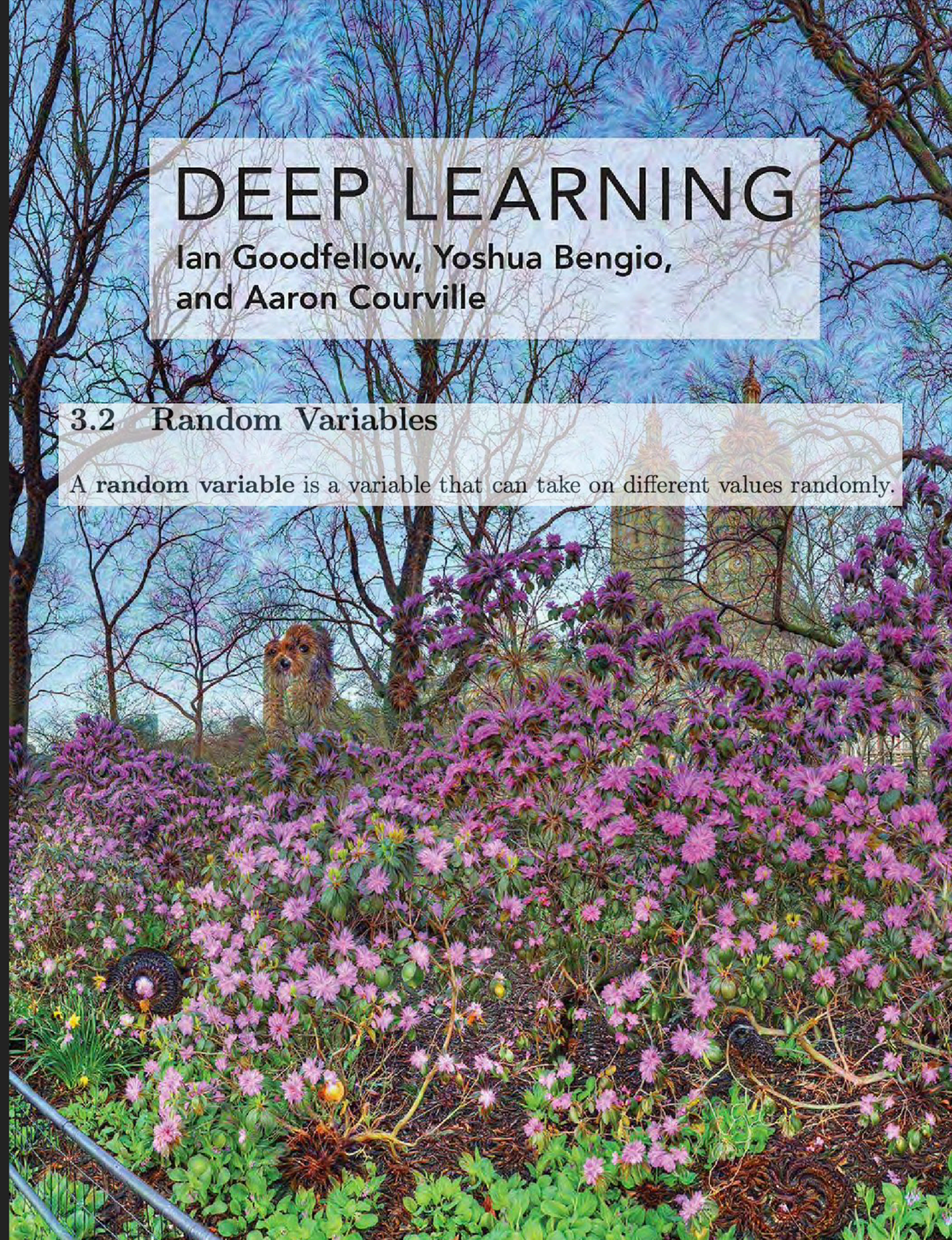
There is nothing more deceptive than an obvious fact.

– Arthur Conan Doyle, *The Boscombe Valley Mystery*

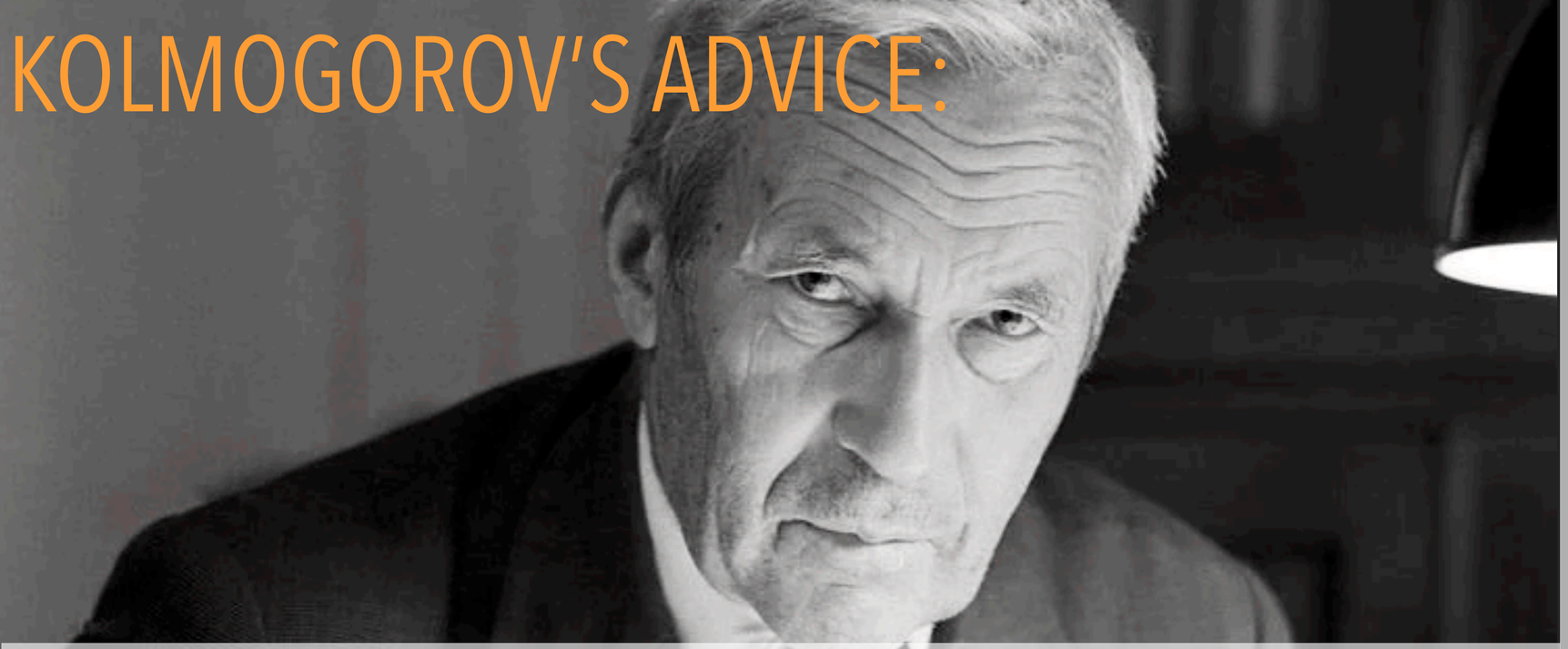
- "Data as fact"
  - The foundation of the "discipline" of statistics:
  - The prospectus of the Statistical Society of London (1838) stated "The Statistical Society will consider it the first and most essential rule of its conduct to exclude carefully opinions from transactions and publications."
    [page 47 of The Exclusion of Opinions, The London and Westminster Review, April-August 1838]
- Their motto was *aliis exterendum* – "to be threshed out by others"
- The wanted to sever any connection between the data and its use
- Nowadays: "benchmark data sets" ; but what gets lost in this view of data?

DATA AS "RANDOM VARIABLES"

- The canonical model of "data"
- Two small difficulties from a mathematical perspective:
  - They are not "random" ; they do not "vary"
    - Because they are simply (measurable) functions!
- Interpreting "random" and "variable" is hard!
  - Bertrand Russell reckoned the notion of a "variable" to be "one of the most difficult to understand" notions in mathematics
- Deep learning researchers are of little help…
- But we have a well accepted mathematical theory of probability. *Surely the answer is known?*
- Indeed! Based upon the Kolmogorov's axiomatisation. So what does he have to say?



KOLMOGOROV'S ADVICE:



FOUNDATIONS
OF THE
THEORY OF PROBABILITY

BY

A.N. KOLMOGOROV

Second English Edition

TRANSLATION EDITED BY
NATHAN MORRISON

WITH AN ADDED BIBLIOGRPAHY BY
A.T. BHARUCHA-REID

UNIVERSITY OF OREGON

CHELSEA PUBLISHING COMPANY
NEW YOURK

§ 2. The Relation to Experimental Data[4]

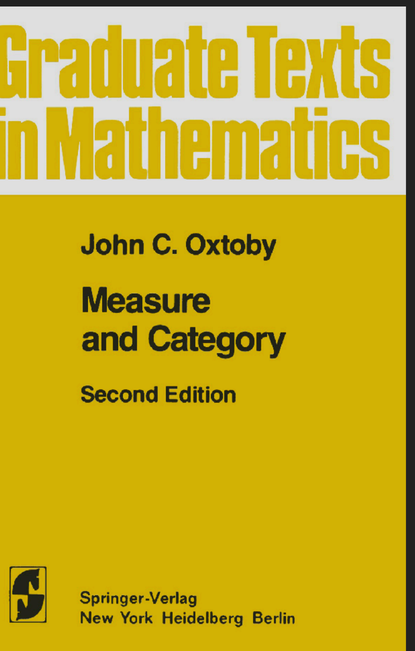
[4] The reader who is interested in the purely mathematical development of the theory only, need not read this section, since the work following it is based only upon the axioms in § 1 and makes no use of the present discussion. Here we limit ourselves to a simple explanation of how the axioms of the theory of probability arose and disregard the deep philosophical dissertations on the concept of probability in the experimental world. In establishing the premises necessary for the applicability of the theory of probability to the world of actual events, the author has used, in large measure, the work of R. v. Mises, [1] pp. 21-27.

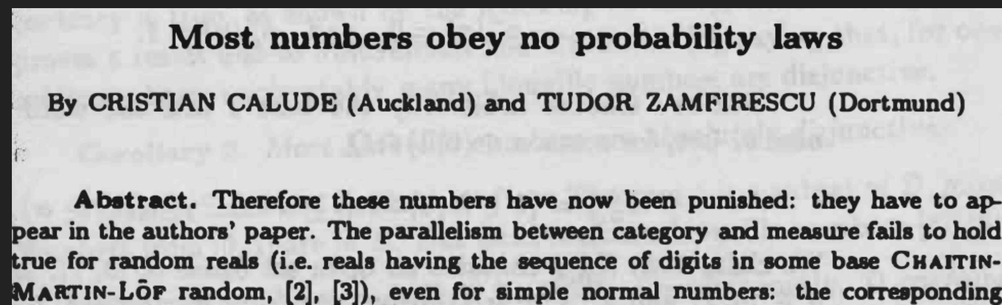
THE STANDARD JUSTIFICATION OF $\mathbb{E}$ – THE "LAW" OF LARGE NUMBERS

- Suppose $(X_i)_{i\in\mathbb{N}}$ is a iid sequence of Bernoulli random variables with $\mathbb{P}(X_i = 1) = p \in [0,1]$.
- Let $S_n := \sum_{i=1}^{n} X_i$ for $n \in \mathbb{N}$.
- Then $\mathbb{P}(\lim_{n\to\infty} S_n/n = p) = 1$

REPEAL OF THE "LAW" OF LARGE NUMBERS

- Let $L^{1/2}$ denote the set of binary sequences with limiting relative frequency $1/2$.
- SLLN: $\lambda(L^{1/2}) = 1.$ Thus $L^{1/2}$ is a very large subset of $\{0,1\}^\infty$
- Oxtoby: $L^{1/2}$ is meagre in $\mathcal{T}^\infty$. Thus $L^{1/2}$ is a very small subset of $\{0,1\}^\infty$.
- Many even stronger variants; bottom line: maths won't solve this…





Most numbers obey no probability laws

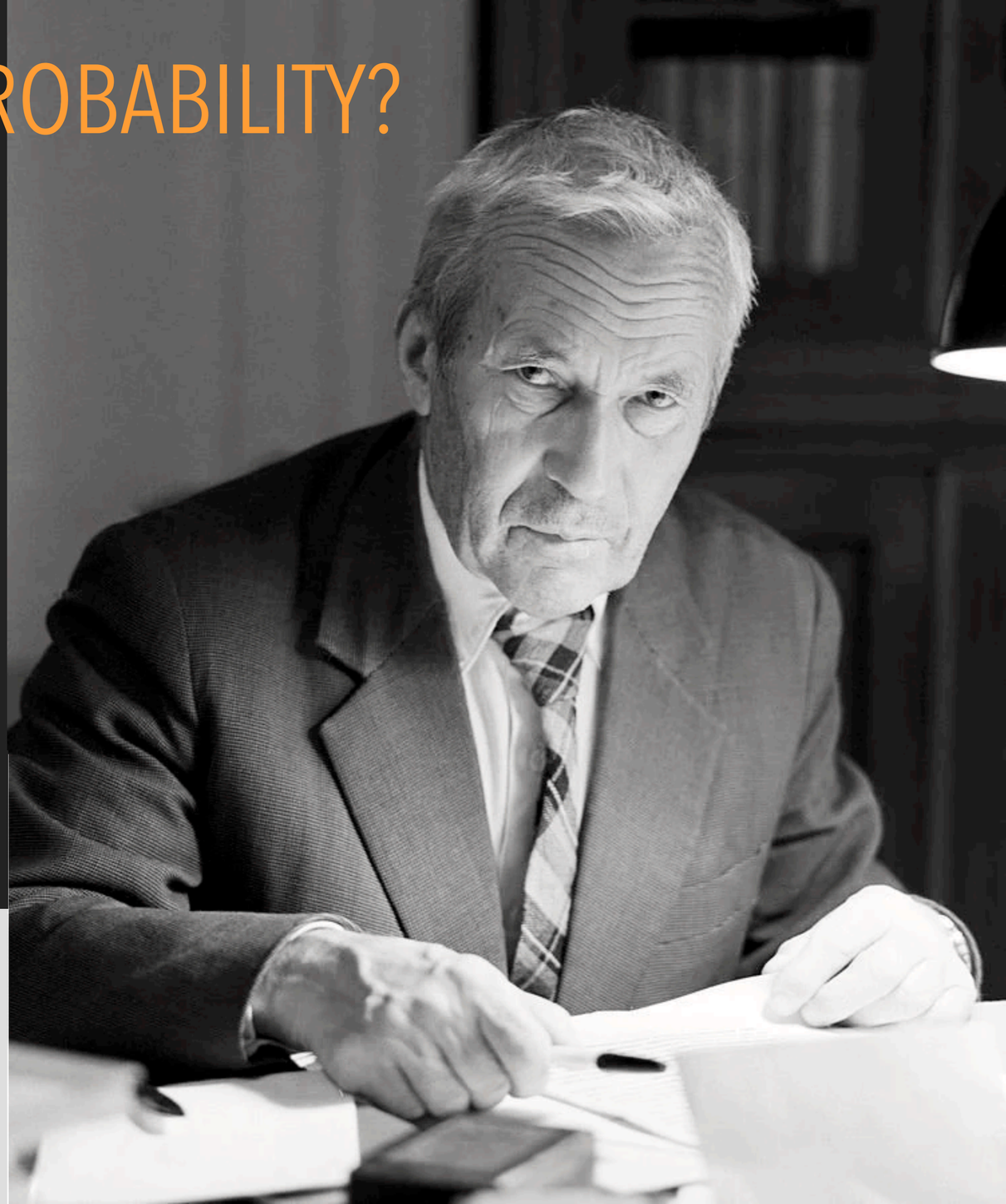
By CRISTIAN CALUDE (Auckland) and TUDOR ZAMFIRESCU (Dortmund)

Abstract. Therefore these numbers have now been punished: they have to appear in the authors' paper. The parallelism between category and measure fails to hold true for random reals (i.e. reals having the sequence of digits in some base CHAITIN-MARTIN-LÖF random, [2], [3]), even for simple normal numbers: the corresponding

BUT WON'T THIS UNDERMINE PROBABILITY?

- Yes, it undermines probability theory's *claim for universality* - for being the only theory of uncertainty you need
- But the grander theory that is needed for "nonstochastic" randomness is a strict generalisation of probability theory

In everyday language we call random these phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking there is no ground to believe that a random phenomenon should possess any definite probability. Therefore, we should have distinguished between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of the probability theory).



VON MISES' PROGRAM: DERIVE KOLMOGOROV'S AXIOMS FROM MORE BASIC ONES

- Von Mises was endeavouring to create a scientific theory of probability – not just a mathematical one
- Analogy: differential equations in physics
  - Axiom 1: relative frequencies converge
  - Axiom 2: They remain invariant under sub-selection
- What if axiom 1 is not satisfied?

THE ROOT OF THE PROBLEM: "THE" PROBABILITY

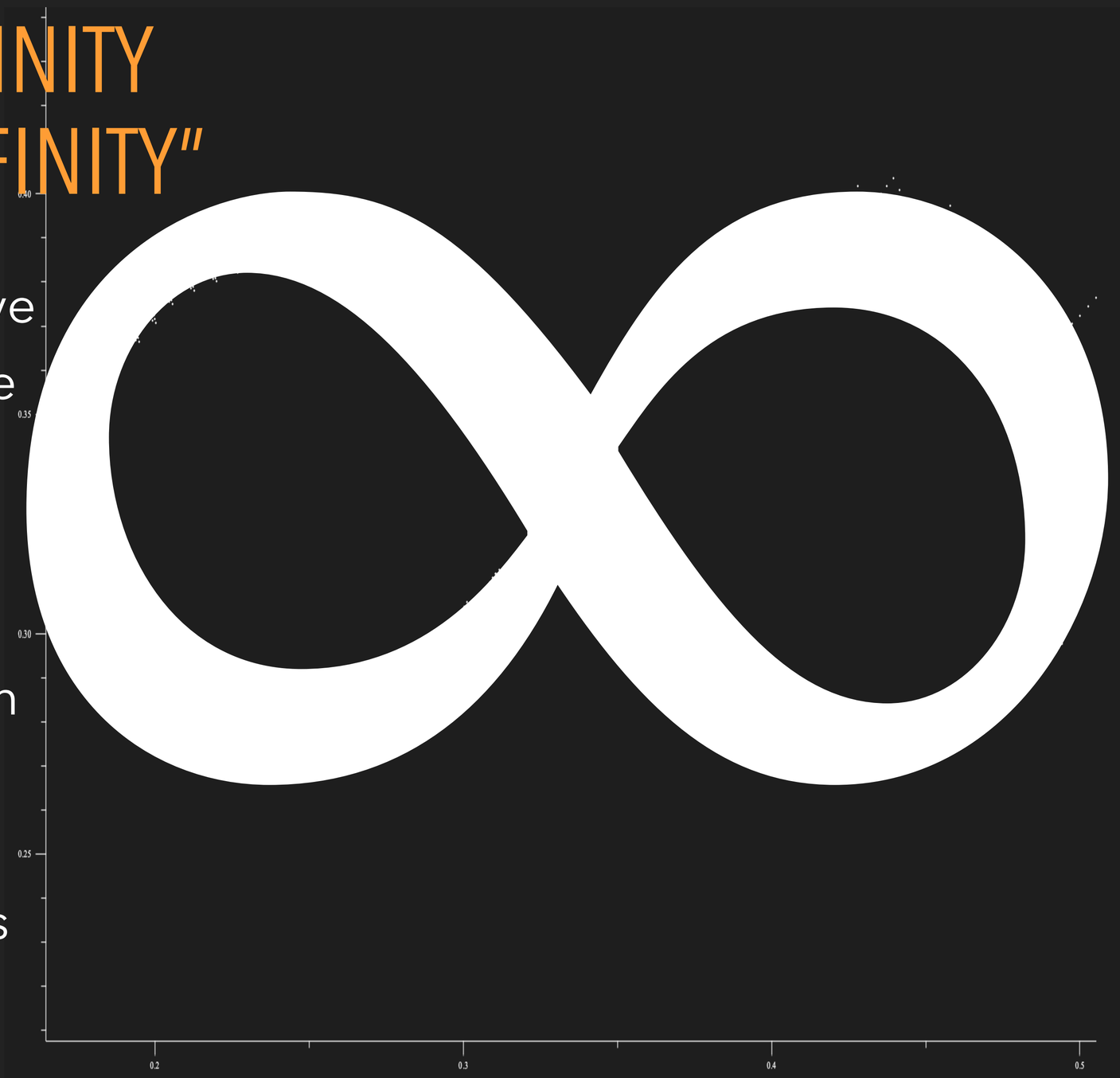
- Bruno De Finetti opened his book on probability by stating that:
  "Probability Does Not Exist"
  - (i.e. did not exist independent of observer).
  - "Thus" it is a wholly subjective notion…
- More accurately:
  "<u>The</u> Probability Does Not Exist"

WHEN THE RELATIVE FREQUENCIES DON'T CONVERGE…

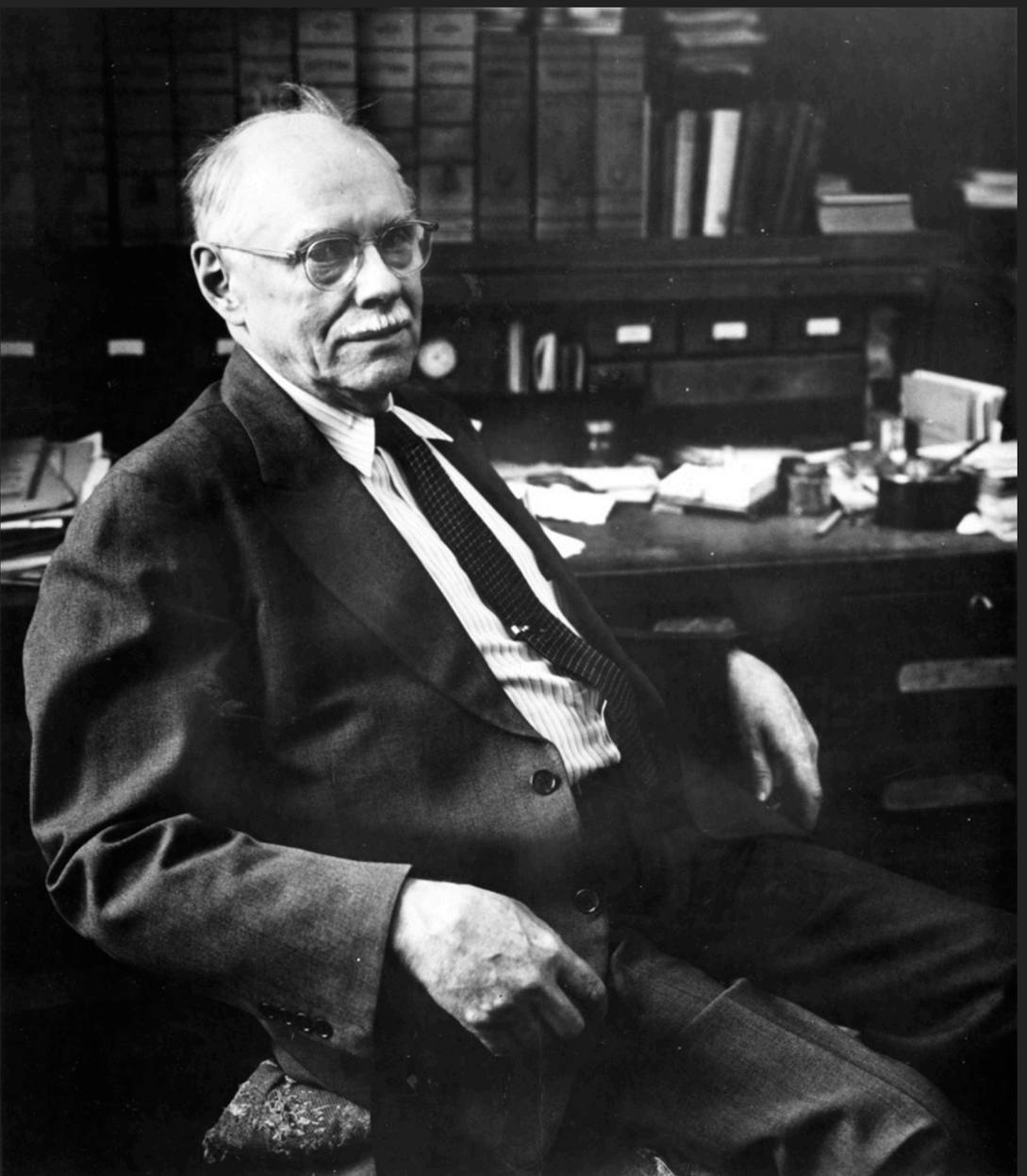
- Then there is no "probability" (or more precisely, no "the probability")
  - Instead there are multiple cluster points
- Have a sequence $x\colon \mathbb{N} \to [k]$
- Relative frequency of ith class: $r_i^x(n) := \frac{1}{n} |\{j \in [n] \colon x_j = i\}|$
- Stack all these into a k-vector to get the relative frequency $r^x(n) := r_{[k]}^x(n)$
- Let $\mathrm{CP}(r^x)$ denote the set of cluster points of r^x ; if a singleton, then limit exists

NON-CONVERGENCE AT INFINITY BUT CONVERGENCE TO "INFINITY"

- Suppose C is a rectifiable closed curve in Δ^k. Then there exists x such that the set of cluster points of r^x is C.
- From this one can naturally recover risk measures / upper previsions
- Every sequence can be reasoned with in terms of an upper prevision
- And every upper prevision can be generated by the relative frequencies of some sequence



IN ADDITION…



- Interpolation between Knightian risk and uncertainty
- Can also develop unstable conditional probability
  - And recover a "generalised Bayes rule"
- And define a nice notion of independence
  - Ends up being a bit tricksy to do (one needs to worry about the set-system)
- But this is early days
  - Yet to explore estimators and the use in practice
- Intriguing that one can completely recover Walley's theory (based as it is on the de Finetti subjective framework) in a strictly frequentist setting

LOOKING AHEAD

CURRENT / FUTURE QUESTIONS

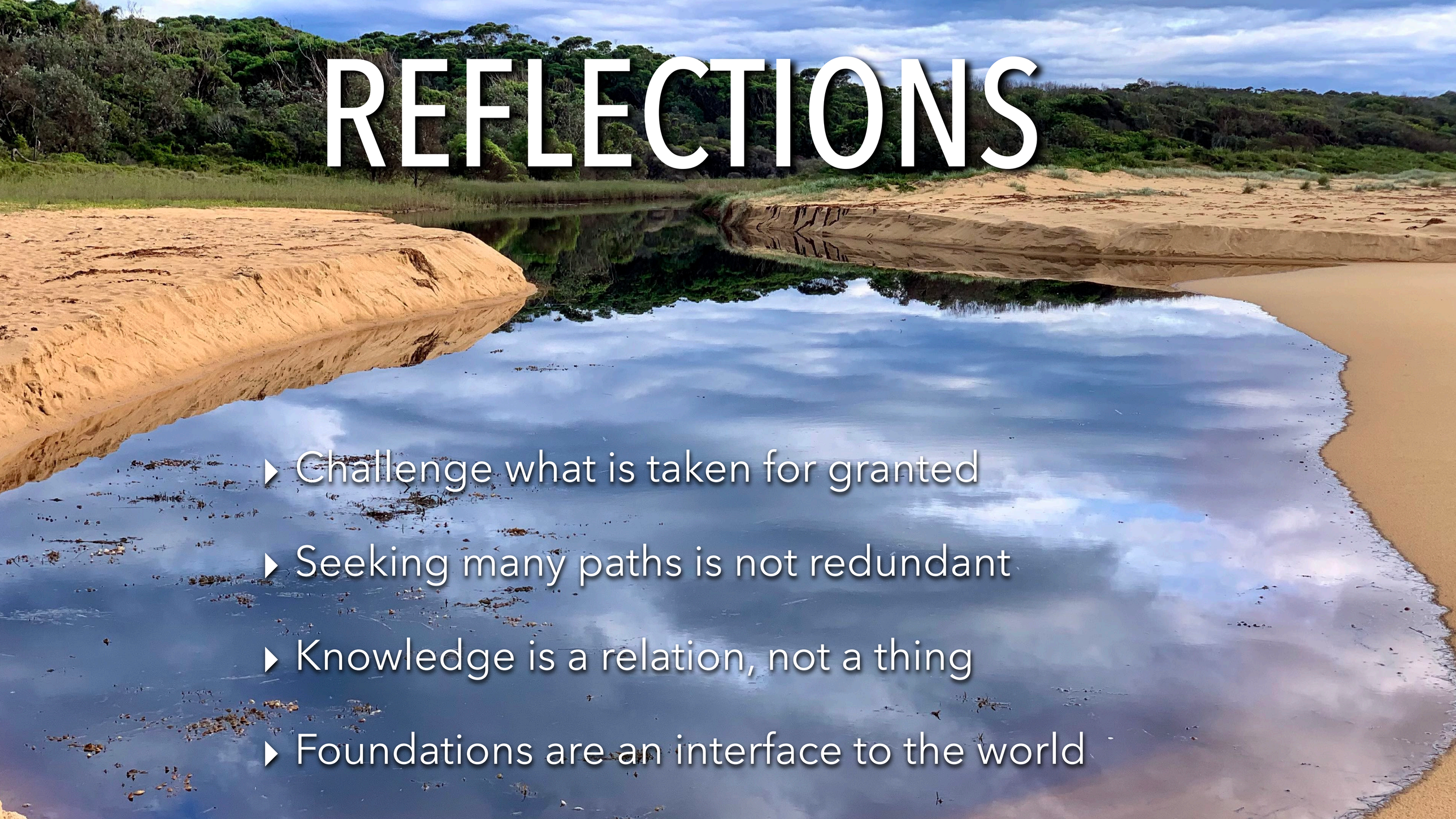
- Make it practical: estimating sets of cluster points of relative frequencies
- Applications in insurance – intersectionality and imprecision
- Relate to general measures of information
  - *Both* are support functions of convex sets…
- Non-stochastic sequences and traditional models
  - Distribution shift, non-stationarity, selection bias, etc
- Relation to other "black swan" theories
- Non-stochastic sequences and performative prediction

The foundations of statistics with black swans

Graciela Chichilnisky [a,b,c,*]

CONCLUSIONS

- Many roads lead to upper previsions $\overline{R}_{\mathscr{P}}(X) = \sup_{P \in \mathscr{P}} \mathbb{E}_P(X)$
  - Mathematically nice: a support function with $\langle X, P \rangle := \mathbb{E}_P(X)$; thus $\overline{R}_{\mathrm{co}\,\mathscr{P}}(X) = \overline{R}_{\mathscr{P}}(X)$
- Imprecise probability of an event A is $\overline{P}(A) = \overline{R}(\chi_A)$
- Fundamental function $\phi(t) = \overline{R}(\chi_{A_t})$, where $\mu(A_t) = t$, stratifies $\{\overline{R}\}$
  - Offers ethical insight and tail control
- System of precision provides a different parametrisation, and offers an approach to modelling intersectionality
- From "almost all" to "all": <u>every</u> sequence $x \colon \mathbb{N} \to [k]$ induces an $\overline{R}$, and vice versa
  - This provides strictly frequentist semantics for imprecise probability
- "The probability" does not necessarily exist; something richer <u>always</u> does



REFLECTIONS

- Challenge what is taken for granted
- Seeking many paths is not redundant
- Knowledge is a relation, not a thing
- Foundations are an interface to the world

Loss functions and Information

The Geometry and Calculus of Losses

Information Processing Equalities and the Information-Risk Bridge

Generalised Expectations

Risk Measures and Upper Probabilities: Coherence and Stratification

Tailoring to the Tails: Risk Measures for Fine-Grained Tail Sensitivity

The System of Imprecision

The Set Structure of Precision: Coherent Probabilities on Pre-Dynkin-Systems

Non-stochastic Sequences

A Strictly Frequentist Theory of Imprecise Probability

Independence*

Fairness and Randomness in Machine Learning: Statistical Independence and Relativization

Nonlinear Expectations & Fairness*

Fairness Risk Measures (ICML2019)

* Not covered in the talk


fm.ls