FOUNDATIONS OF MACHINE LEARNING SYSTEMS

Robert Williamson

FOUNDATIONS OF MACHINE LEARNING SYSTEMS

machine learning
new perspectives
for science

Robert Williamson

FOUNDATIONS OF MACHINE LEARNING SYSTEMS

Robert Williamson

machine learning
new perspectives
for science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FOUNDATIONS OF MACHINE LEARNING SYSTEMS

Robert Williamson

machine learning
new perspectives
for science

Tübingen AI Center

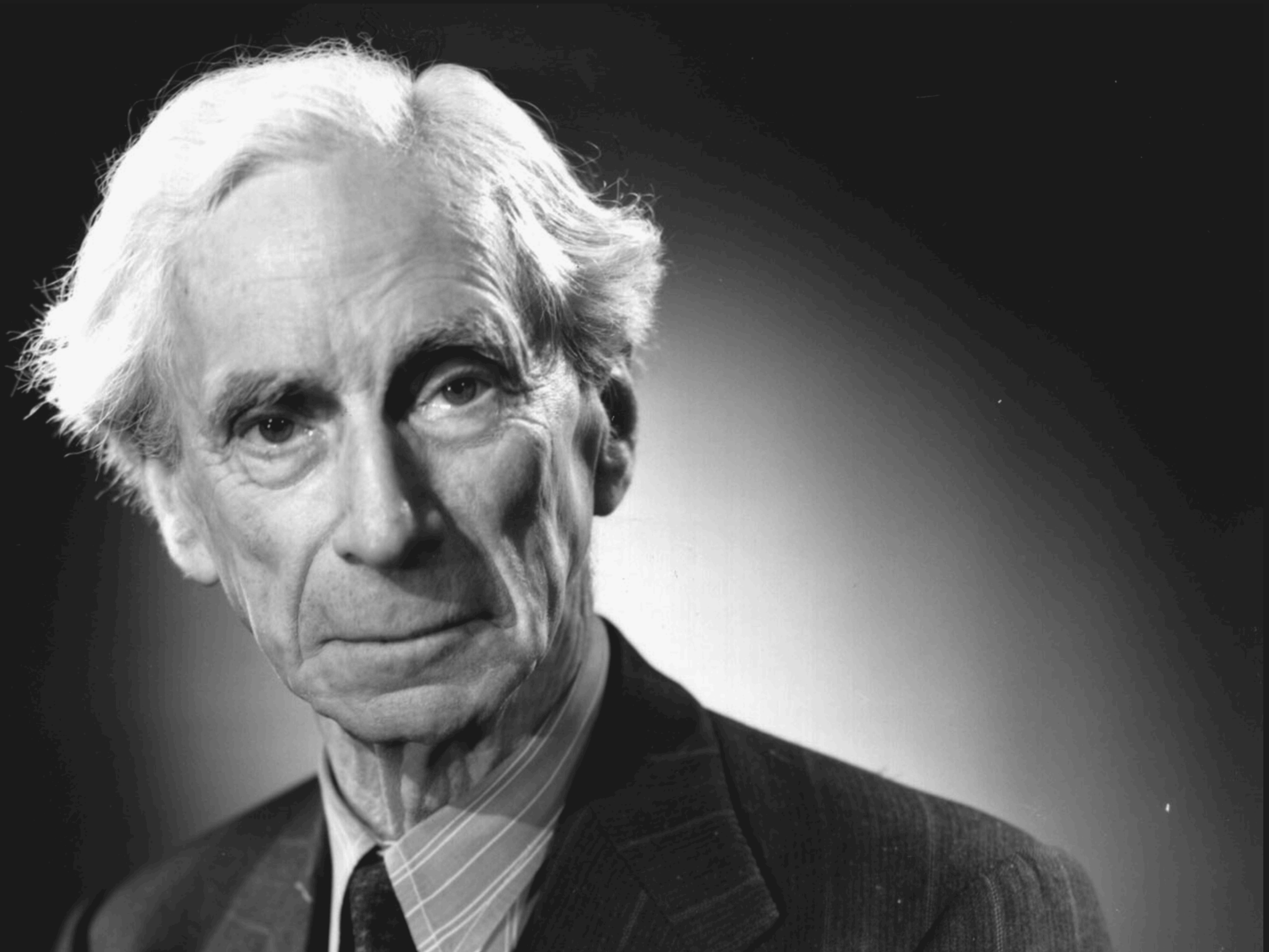
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN


OPTIMIST

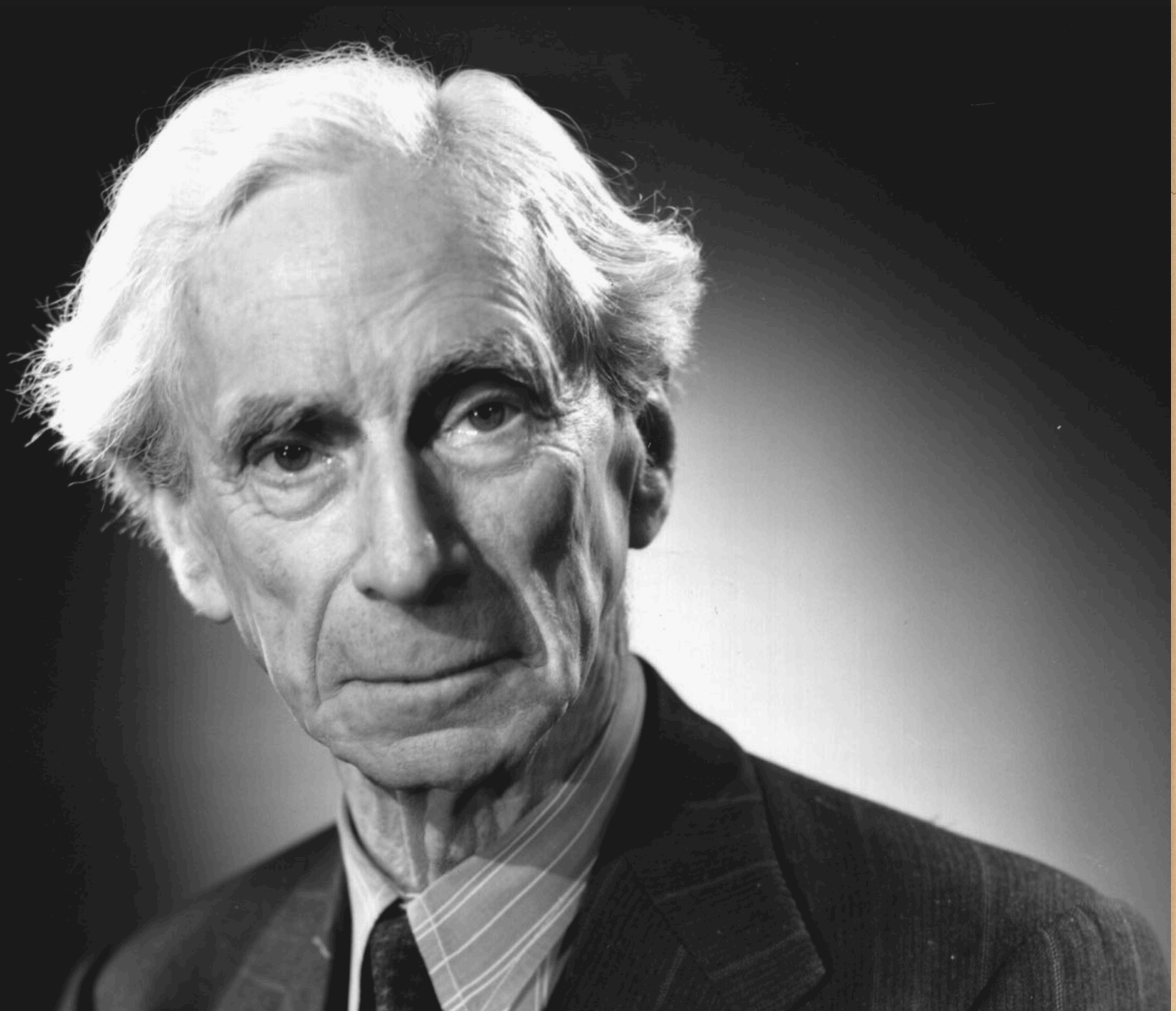
fm.ls

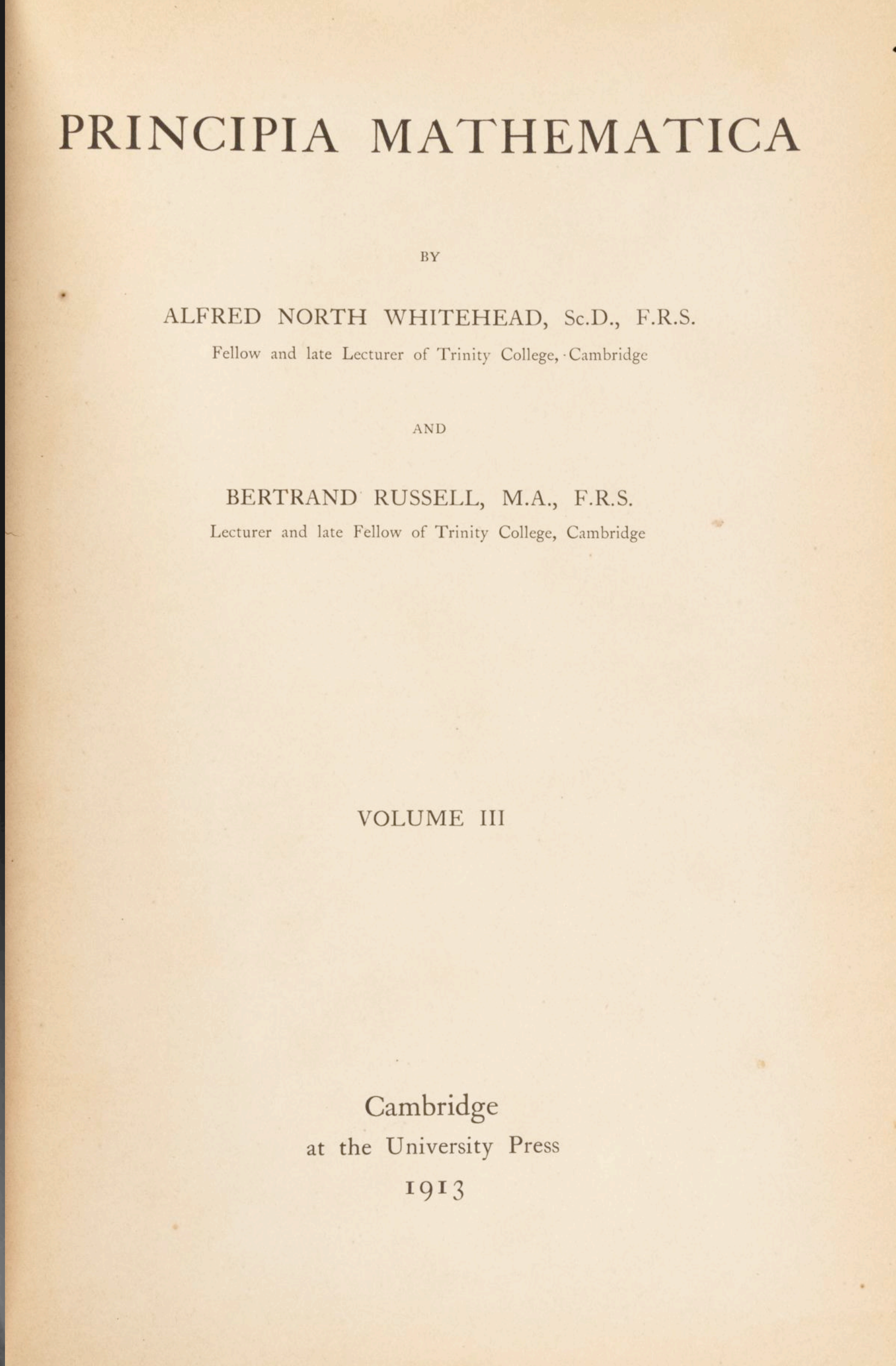

FOUNDATIONS of . . .

THE PHILOSOPHER'S VIEW



THE PHILOSOPHER'S VIEW





PRINCIPIA MATHEMATICA

BY

ALFRED NORTH WHITEHEAD, Sc.D., F.R.S.
Fellow and late Lecturer of Trinity College, Cambridge

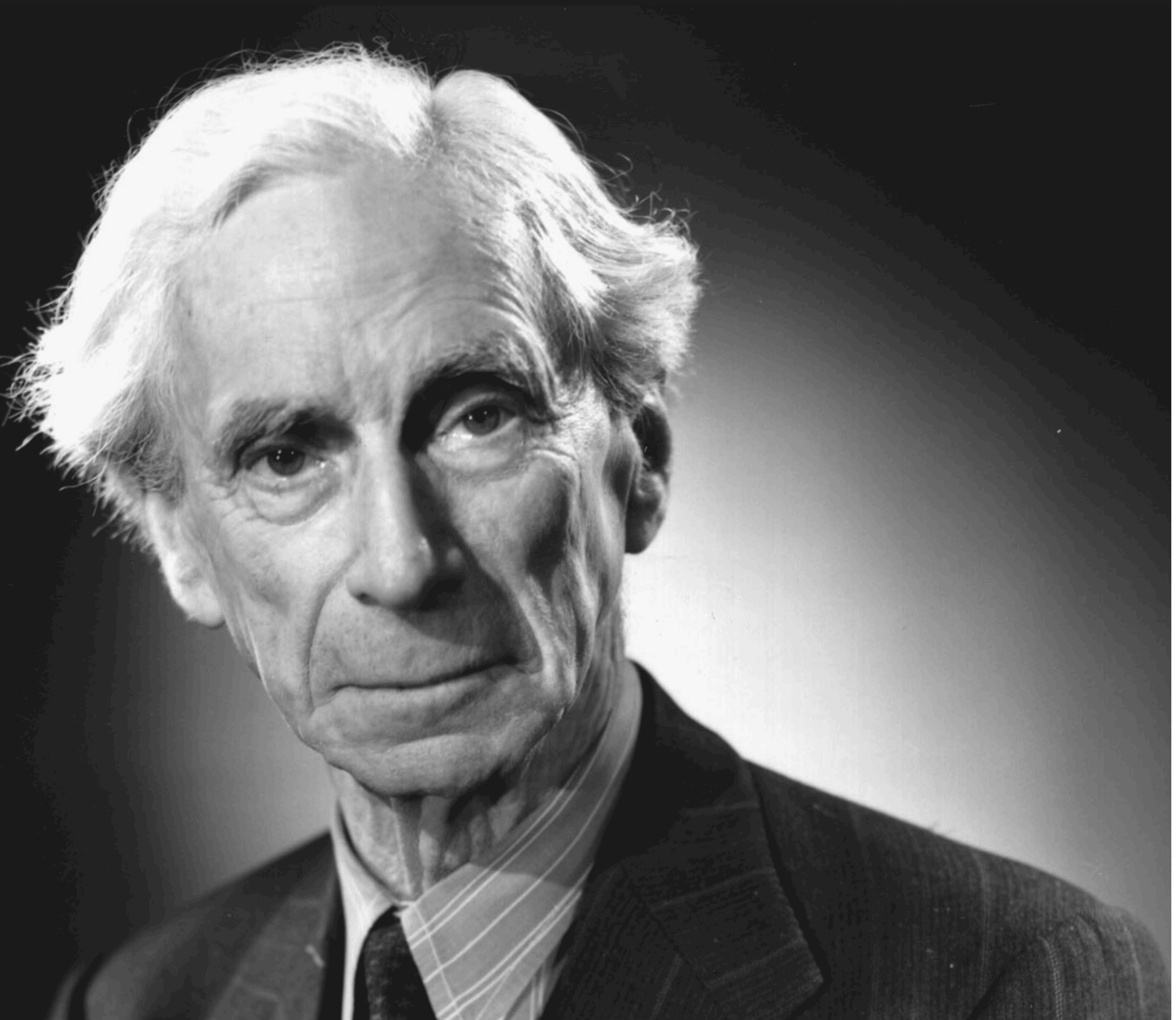
AND

BERTRAND RUSSELL, M.A., F.R.S.
Lecturer and late Fellow of Trinity College, Cambridge

VOLUME III

Cambridge
at the University Press
1913

THE PHILOSOPHER'S VIEW



∗10·2. $\vdash :. (x) . p \lor \phi x . \equiv : p . \lor . (x) . \phi x$

Dem.

$\vdash . *10\cdot1 . *1\cdot6 . \supset \vdash :. p . \lor . (x) . \phi x : \supset . p \lor \phi y :.$

$[*10\cdot11] \quad \supset \vdash :. (y) :. p . \lor . (x) . \phi x : \supset . p \lor \phi y :.$

$[*10\cdot12] \quad \supset \vdash :. p . \lor . (x) . \phi x : \supset . (y) . p \lor \phi y \quad (1)$

$\vdash . *10\cdot12 . \quad \supset \vdash :. (y) . p \lor \phi y . \supset : p . \lor . (x) . \phi x \quad (2)$

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

∗10·21. $\vdash :. (x) . p \supset \phi x . \equiv : p . \supset . (x) . \phi x \quad \left[*10\cdot2 \frac{\sim p}{p}\right]$

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.

$\vdash . *10\cdot1 . \quad \supset \vdash : (x) . \phi x . \psi x . \supset . \phi y . \psi y . \quad (1)$

$[*3\cdot26] \quad \supset . \phi y :$

$[*10\cdot11] \quad \supset \vdash :. (y) : (x) . \phi x . \psi x . \supset . \phi y :.$

$[*10\cdot21] \quad \supset \vdash :. (x) . \phi x . \psi x . \supset . (y) . \phi y \quad (2)$

$\vdash . (1) . *3\cdot27 . \quad \supset \vdash :. (x) . \phi x . \psi x . \supset . \psi z :.$

$[*10\cdot11] \quad \supset \vdash :. (z) : (x) . \phi x . \psi x . \supset . \psi z :.$

$[*10\cdot21] \quad \supset \vdash :. (x) . \phi x . \psi x . \supset . (z) . \psi z \quad (3)$

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

$\vdash . *10\cdot14\cdot11 . \quad \supset \vdash :. (y) :. (x) . \phi x : (x) . \psi x : \supset . \phi y . \psi y :.$

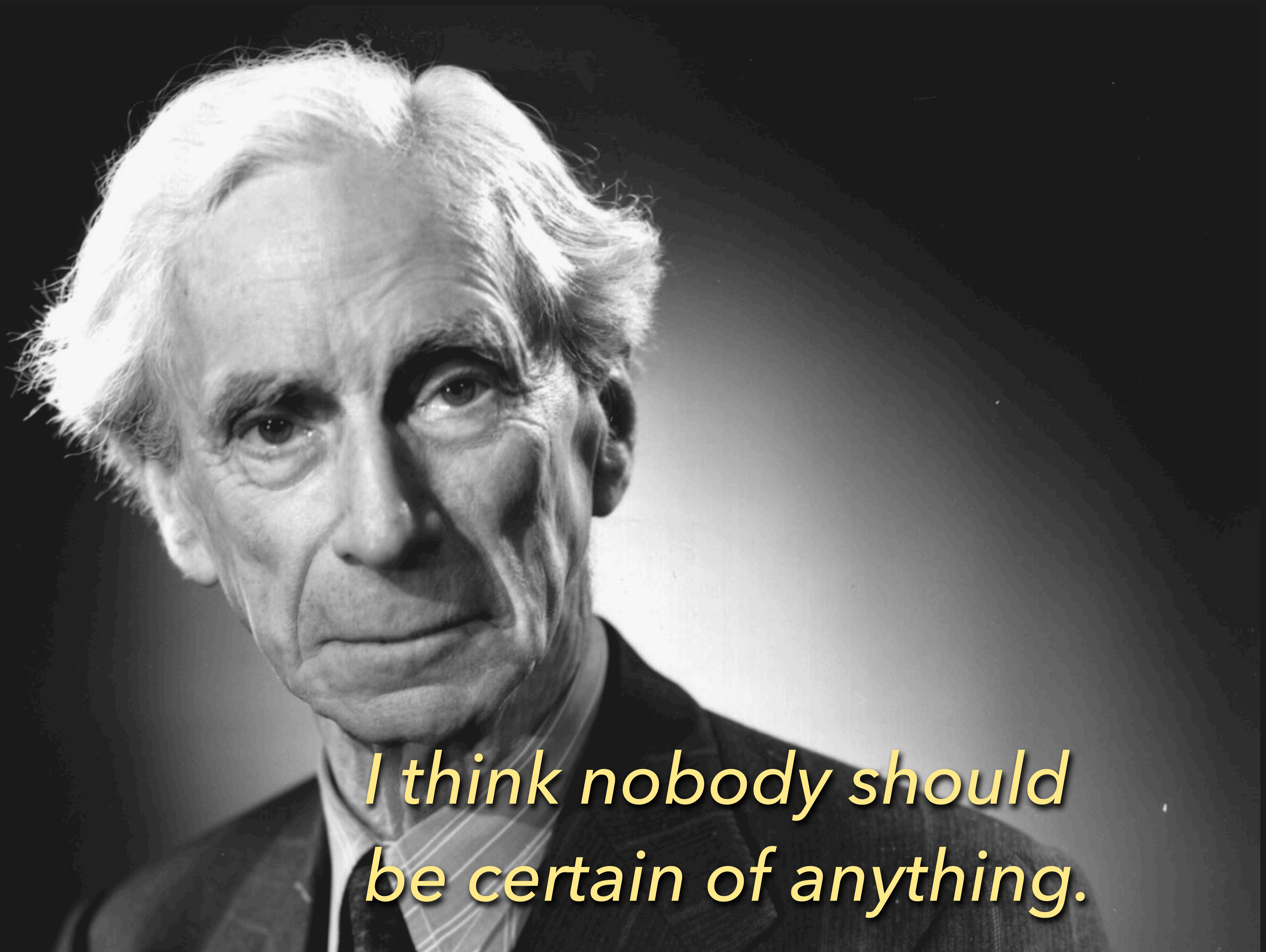
$[*10\cdot21] \quad \supset \vdash :. (x) . \phi x : (x) . \psi x : \supset . (y) . \phi y . \psi y \quad (5)$

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

∗10·221. If ϕx contains a constituent $\chi(x, y, z, \ldots)$ and ψx contains a constituent $\chi(x, u, v, \ldots)$, where χ is an elementary function and $y, z, \ldots u, v, \ldots$ 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 ϕ and ψ from χ, χ is only submitted to negations, disjunctions and generalizations. The process may 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 $(\exists y) . \sim \chi(x, y) \lor \theta x$, and ψx is $(y) . \sim fx \lor \chi(x, y)$. Hence since the primitive ideas $(x) . Fx$ and $(\exists x) . Fx$ only apply to functions, there are functions $\sim \chi(\hat{x}, \hat{y}) \lor \theta\hat{x}$, $\sim f\hat{x} \lor \chi(\hat{x}, \hat{y})$. Hence there is a proposition $\sim \chi(a, b) \lor \theta a$. Hence, since "$p \lor q$" and "$\sim p$" are only significant

THE PHILOSOPHER'S VIEW



THE ENGINEER'S VIEW

THE ENGINEER'S VIEW

- "an element of a structure which connects it to the ground" – Wikipedia

THE ENGINEER'S VIEW

- "an element of a structure which connects it to the ground" – Wikipedia
- Abstracting just slightly: An Interface to the World

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?

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?
    - Or if the world changes, and what was solid before no longer is…

FOUNDATIONS OF MACHINE LEARNING SYSTEMS

FOUNDATIONS OF MACHINE LEARNING SYSTEMS

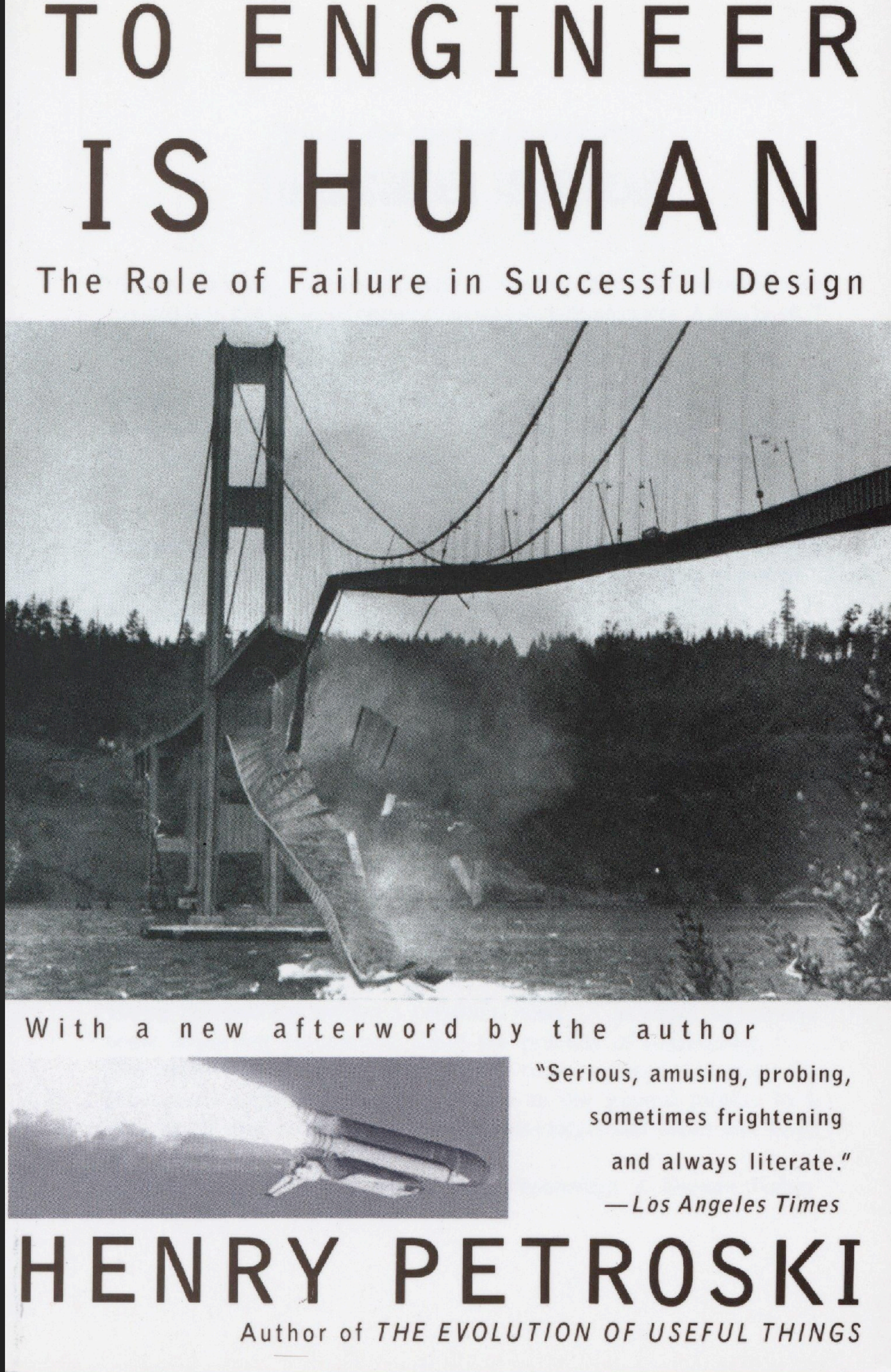
- Study the *interface* of ML systems to the world

FOUNDATIONS OF MACHINE LEARNING SYSTEMS

- Study the *interface* of ML systems to the world
- Pay attention to what we *assume* about the world

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 *failure*




...MACHINE LEARNING SYSTEMS

MACHINE LEARNING SYSTEMS

MACHINE LEARNING SYSTEMS

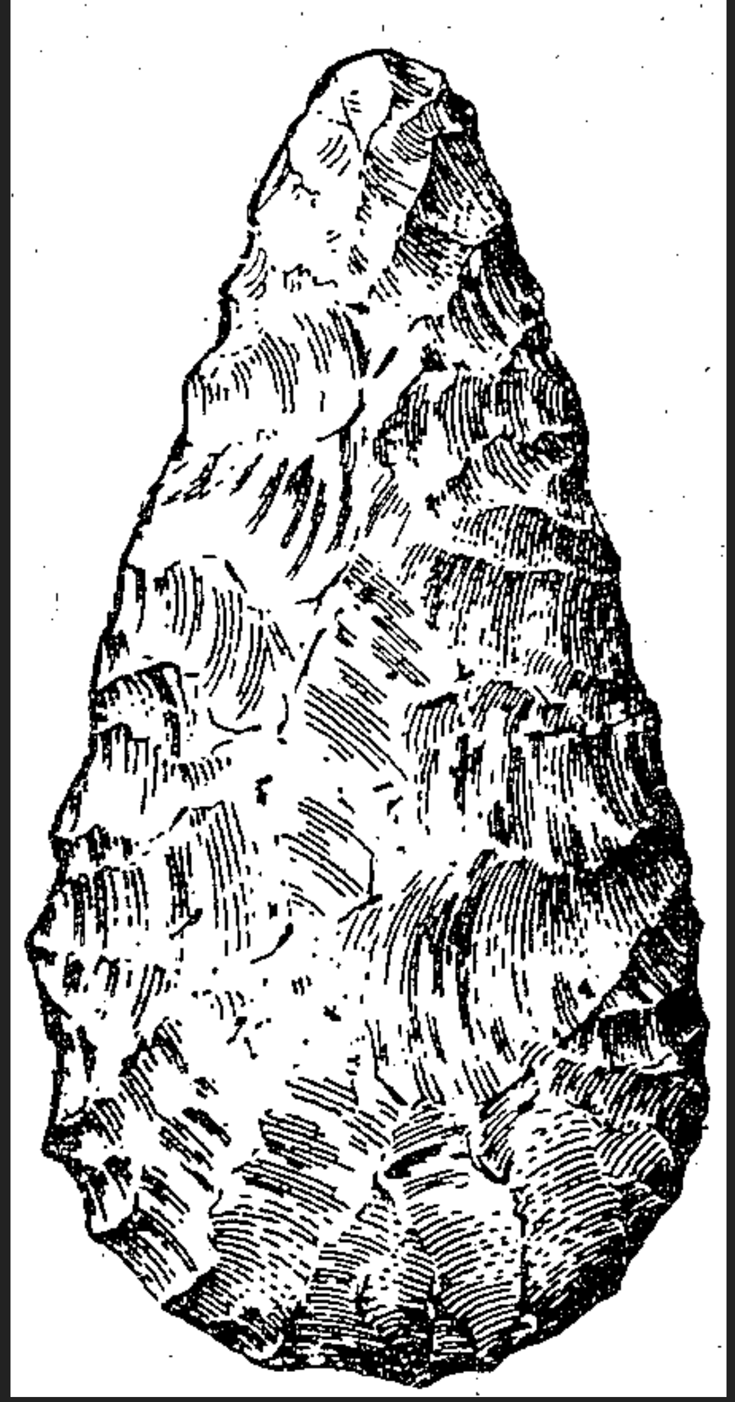
- Are everywhere; I hardly need justify interest in them…

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use



MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention
- But what do these systems *do*?

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention
- But what do these systems *do*?
- On the basis of data (symbolic views of part of the world)…

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention
- But what do these systems *do*?
- On the basis of data (symbolic views of part of the world)…
  which we choose (or take for granted)…

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention
- But what do these systems *do*?
- On the basis of data (symbolic views of part of the world)…
  which we choose (or take for granted)…
  they distill the data into a model (an approximation)…

MACHINE LEARNING SYSTEMS

- Are *everywhere*; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention
- But what do these systems *do*?
- On the basis of data (symbolic views of part of the world)…
  which we choose (or take for granted)…
  they distill the data into a model (an approximation)…
  in order to predict (which can be turned into an act) …

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention
- But what do these systems *do*?
- On the basis of data (symbolic views of part of the world)…
  which we choose (or take for granted)…
  they distill the data into a model (an approximation)…
  in order to predict (which can be turned into an act) …
  on our (or others) behalf…

MACHINE LEARNING SYSTEMS

- Are everywhere; I hardly need justify interest in them…
- Machine (or tool)?
    - Tools under our control, and require skill to use
    - Machines usually ascribed some autonomy
- Learning (sounds like knowledge…)
    - But what is that? Not certain. Not universal. Not objective. Not eternal.
- Systems - well *everything* is a system; the name just signals *context* … to which we should pay more attention
- But what do these systems *do*?
- On the basis of data (symbolic views of part of the world)…
  which we choose (or take for granted)…
  they distill the data into a model (an approximation)…
  in order to predict (which can be turned into an act) …
  on our (or others) behalf…
  according to goals we (or others) set…

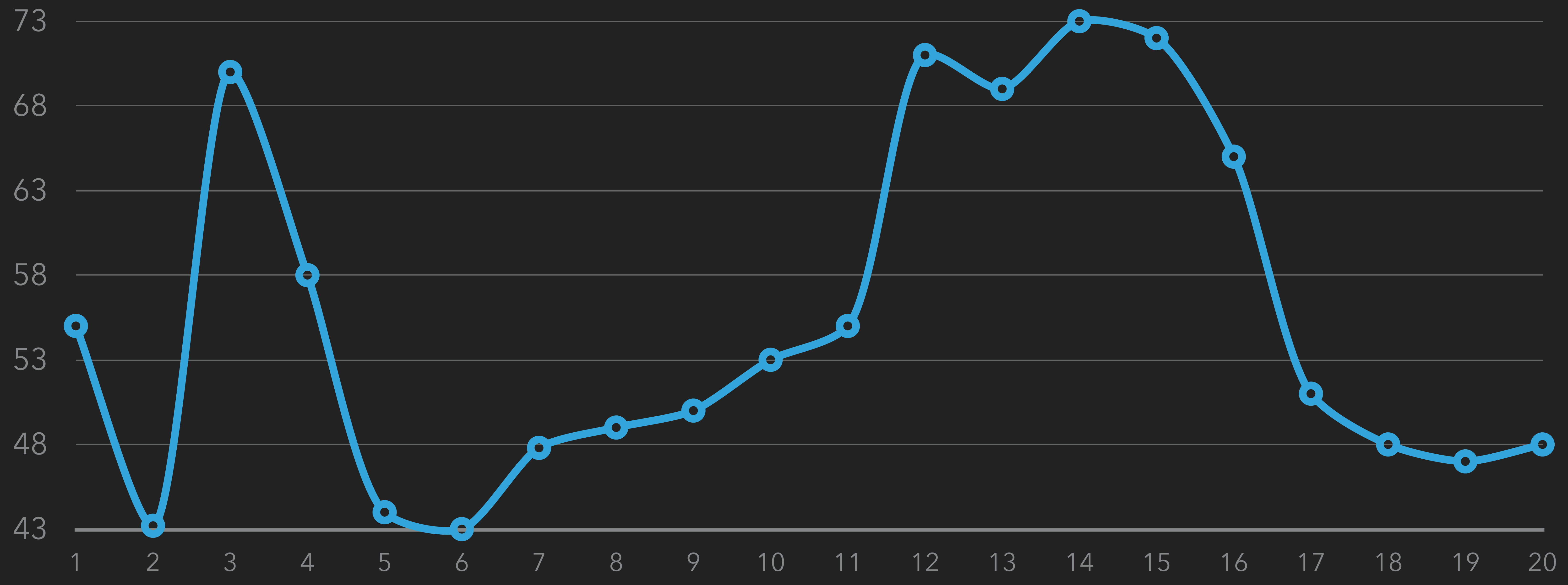
DATA AND THE ACTUARIAL TURN

DATA AND THE ACTUARIAL TURN

- Two styles of reasoning with data: direct, and actuarial

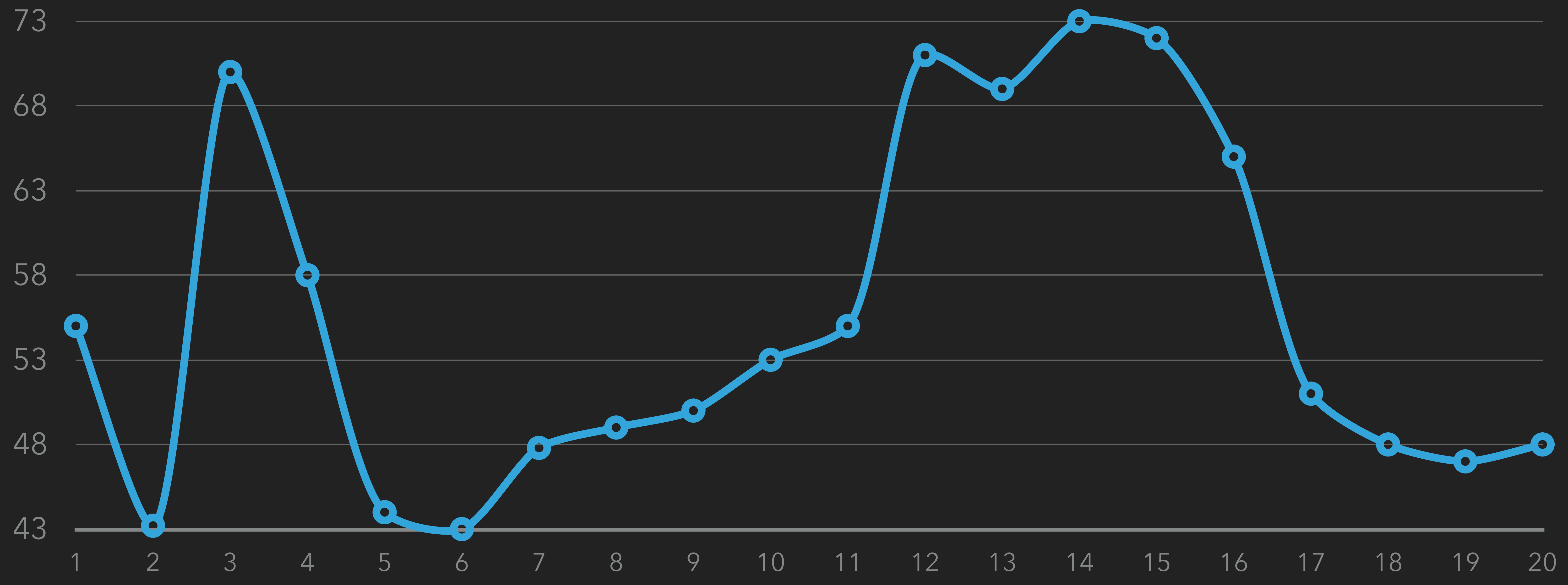
DATA AND THE ACTUARIAL TURN

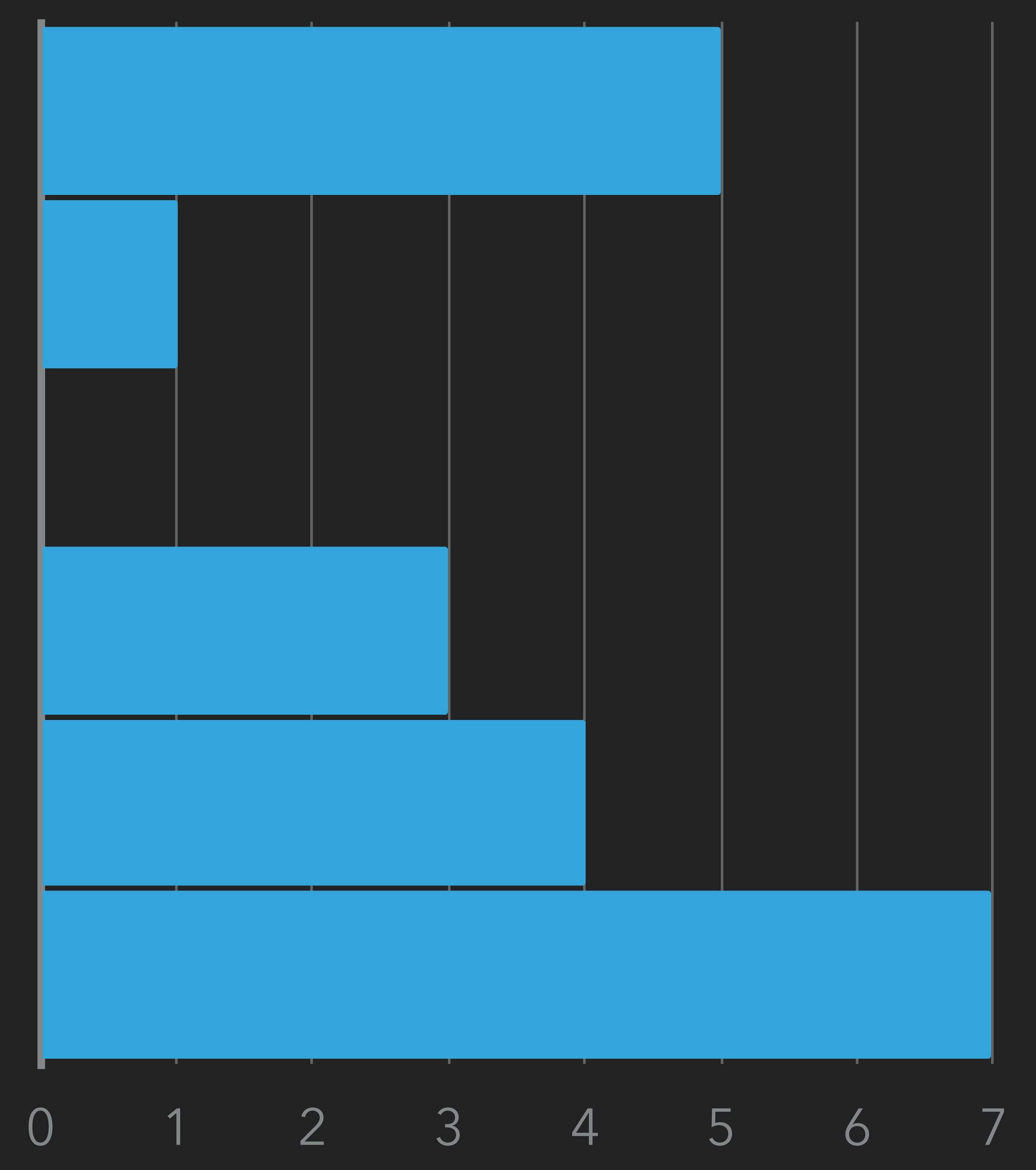
- Two styles of reasoning with data: direct, and actuarial



DATA AND THE ACTUARIAL TURN

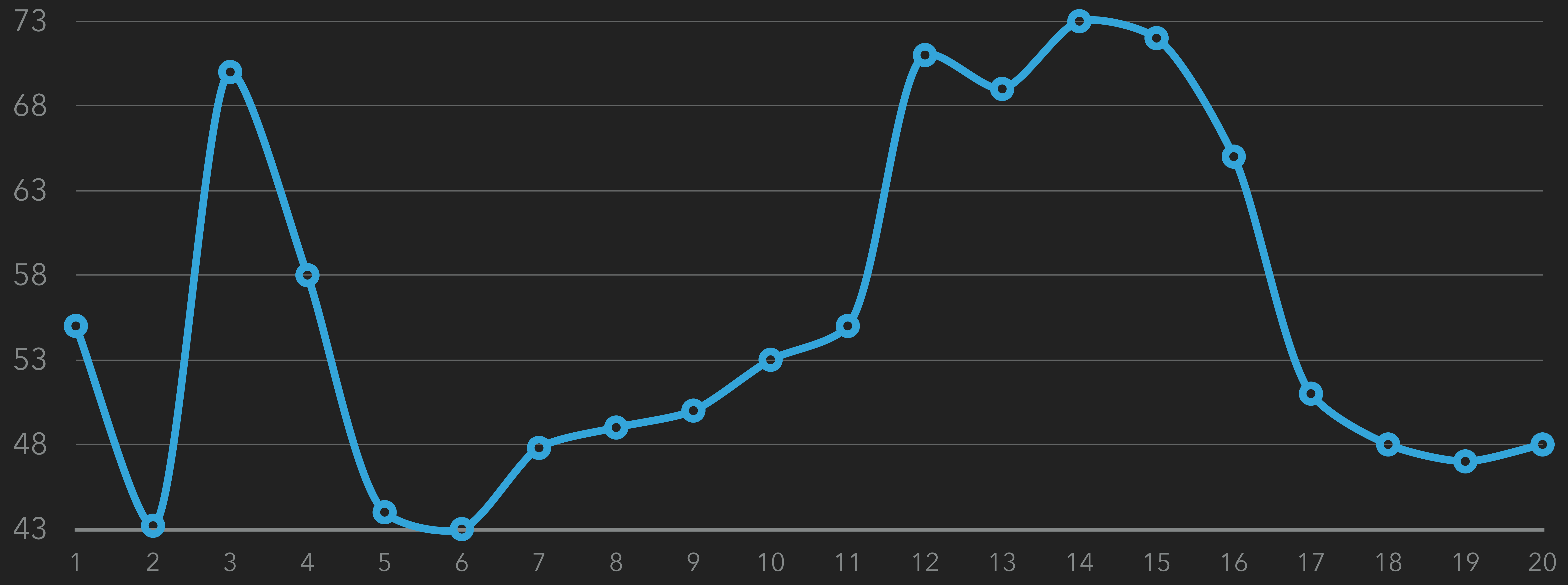
- Two styles of reasoning with data: direct, and actuarial

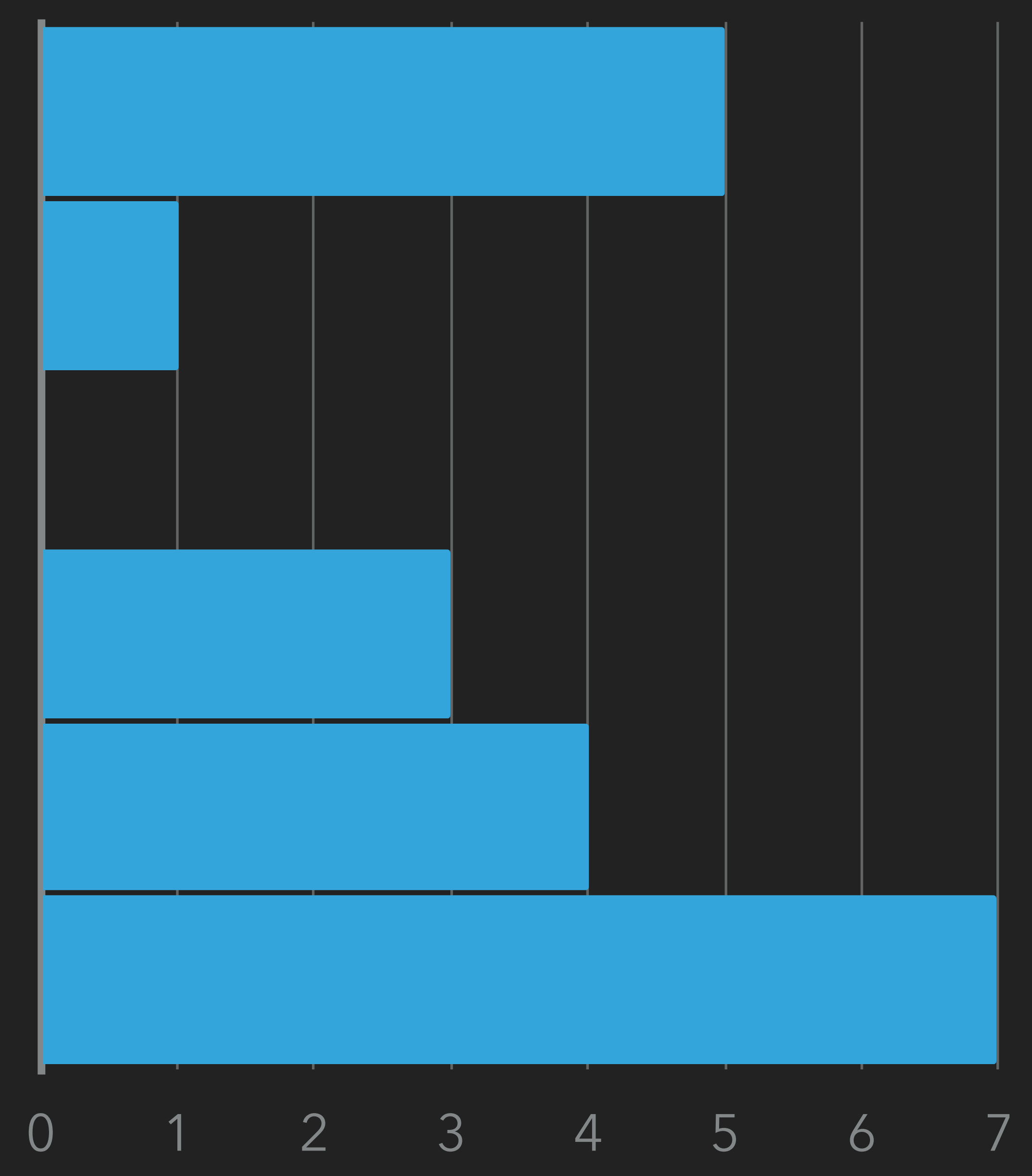




DATA AND THE ACTUARIAL TURN

- Two styles of reasoning with data: direct, and actuarial

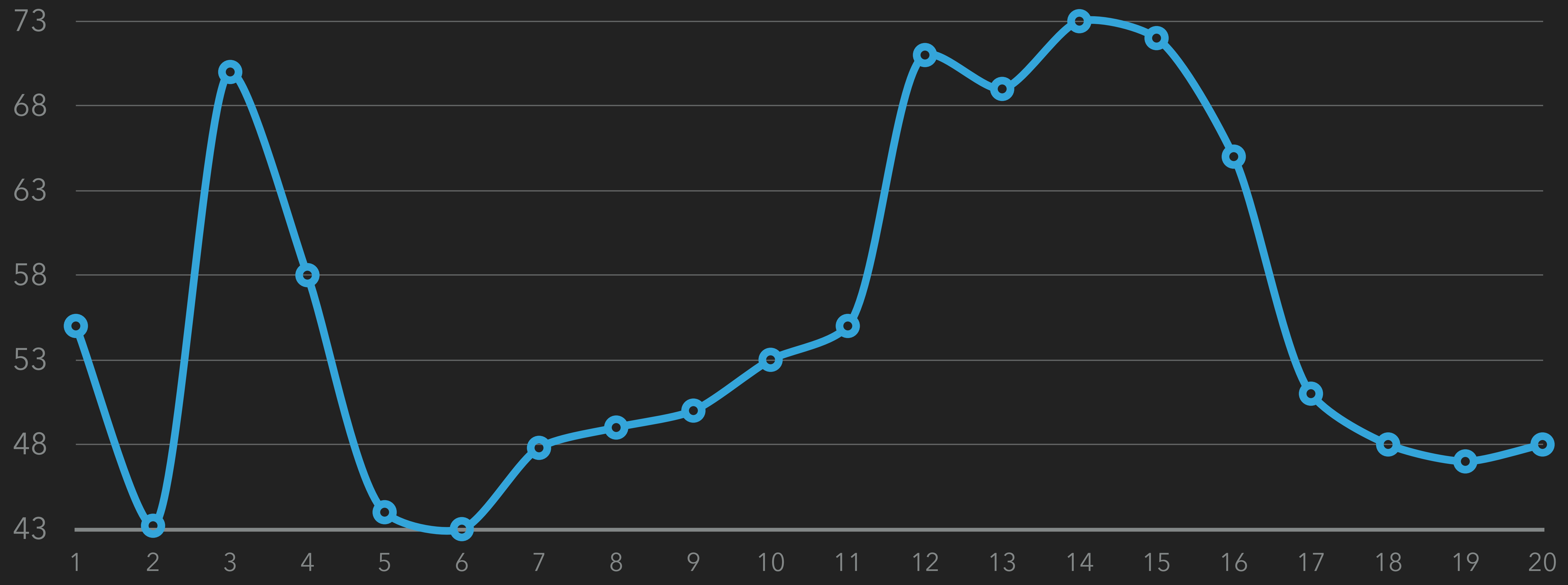


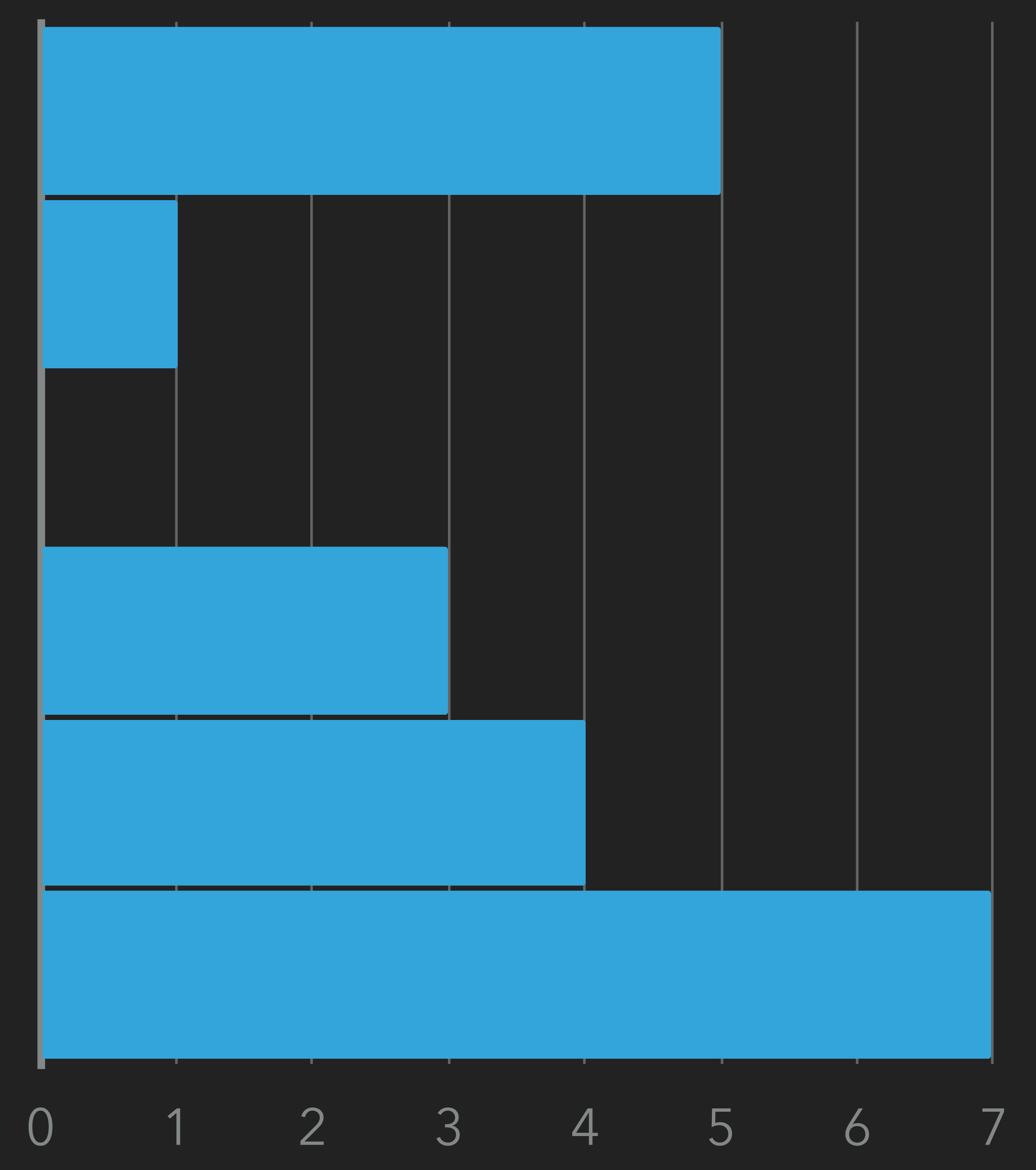


- Statistical Mechanics as an exemplar

DATA AND THE ACTUARIAL TURN

- Two styles of reasoning with data: direct, and actuarial

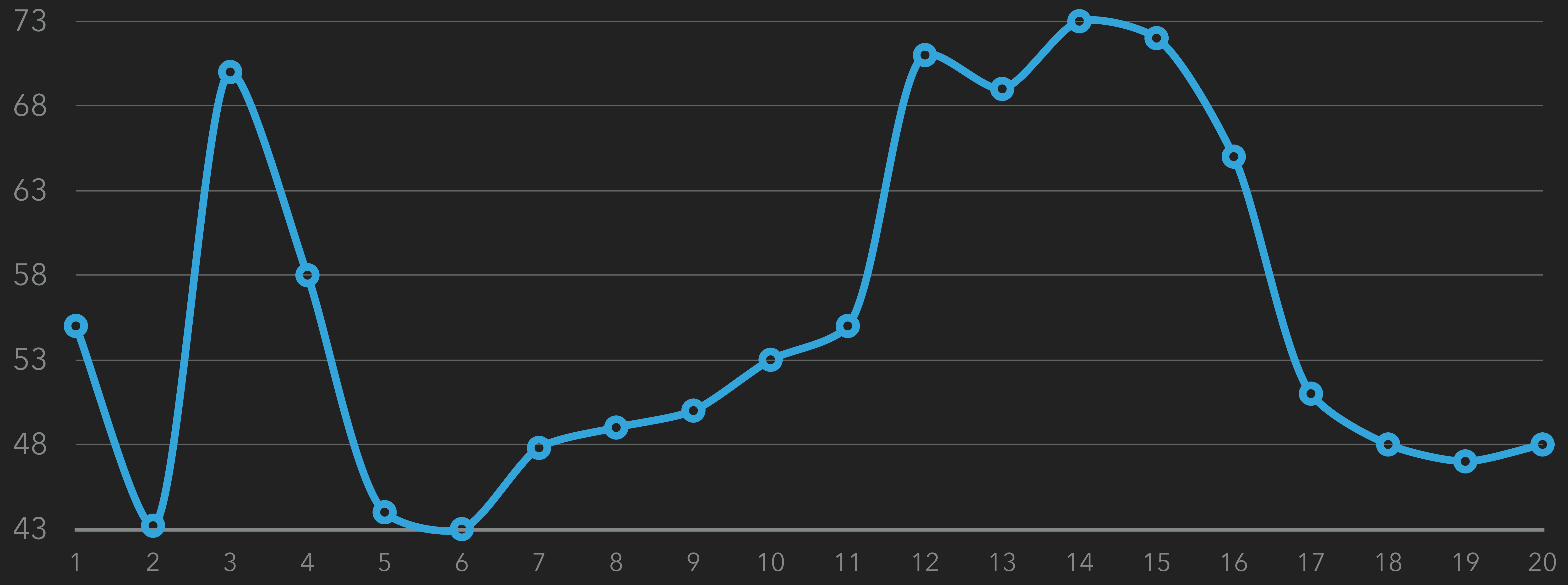


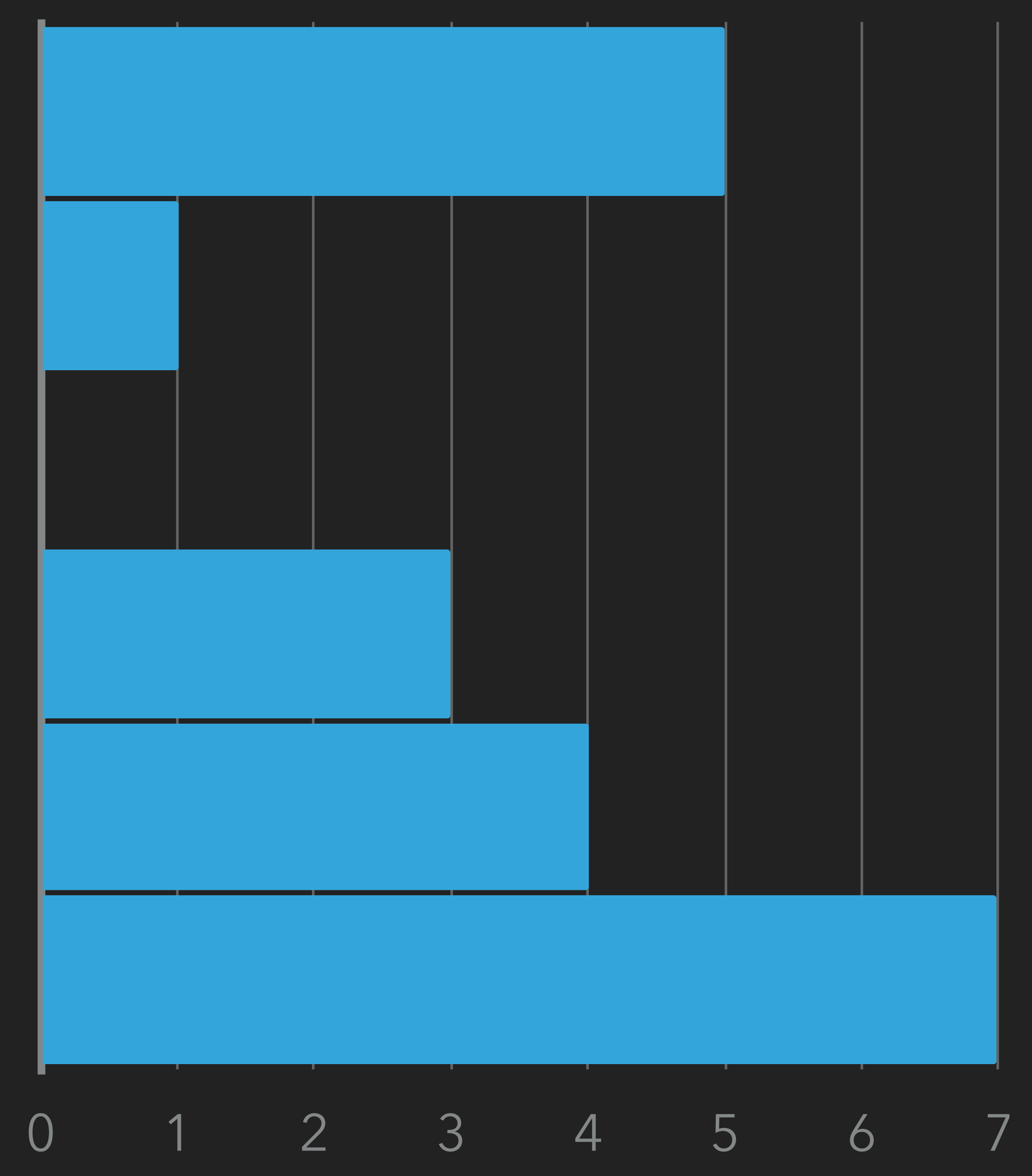


- Statistical Mechanics as an exemplar
- ML systems: an actuarial technology

DATA AND THE ACTUARIAL TURN

- Two styles of reasoning with data: direct, and actuarial





- Statistical Mechanics as an exemplar
- ML systems: an actuarial technology
  - (which means we can learn from insurance!)

Insights From Insurance for Fair Machine Learning: Responsibility, Performativity and Aggregates

FROM KNOWING TO INTERVENING

FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…

FROM KNOWING TO INTERVENING

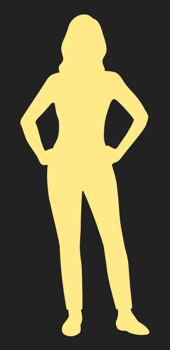
- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it

FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are

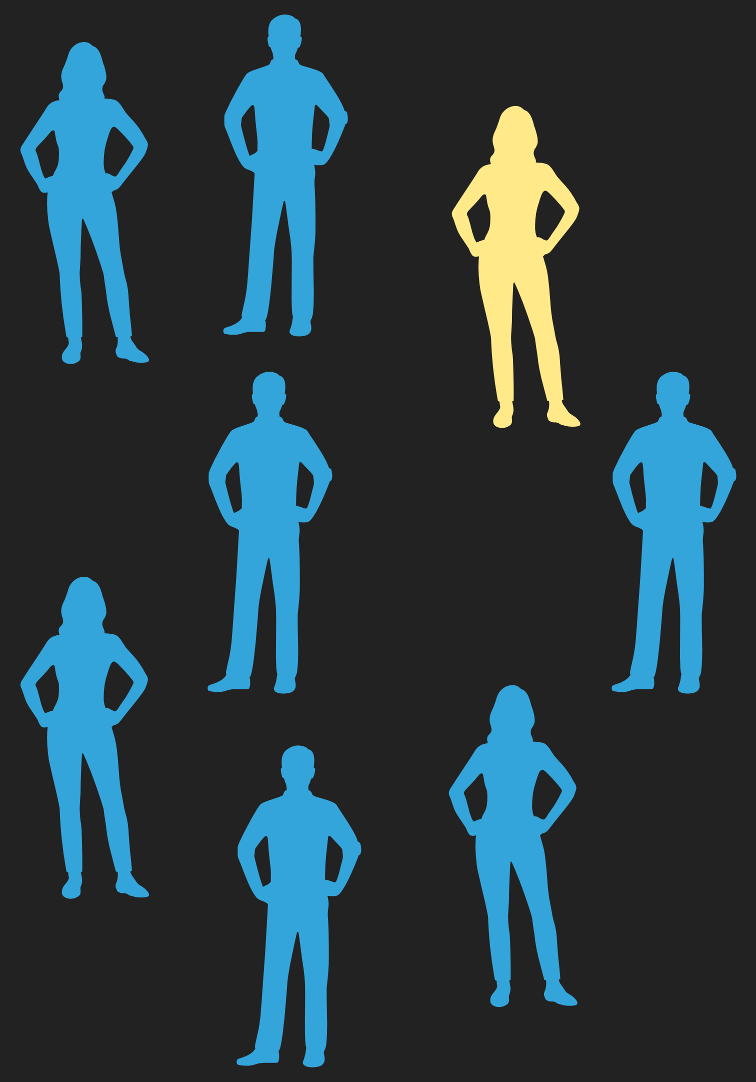
FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people



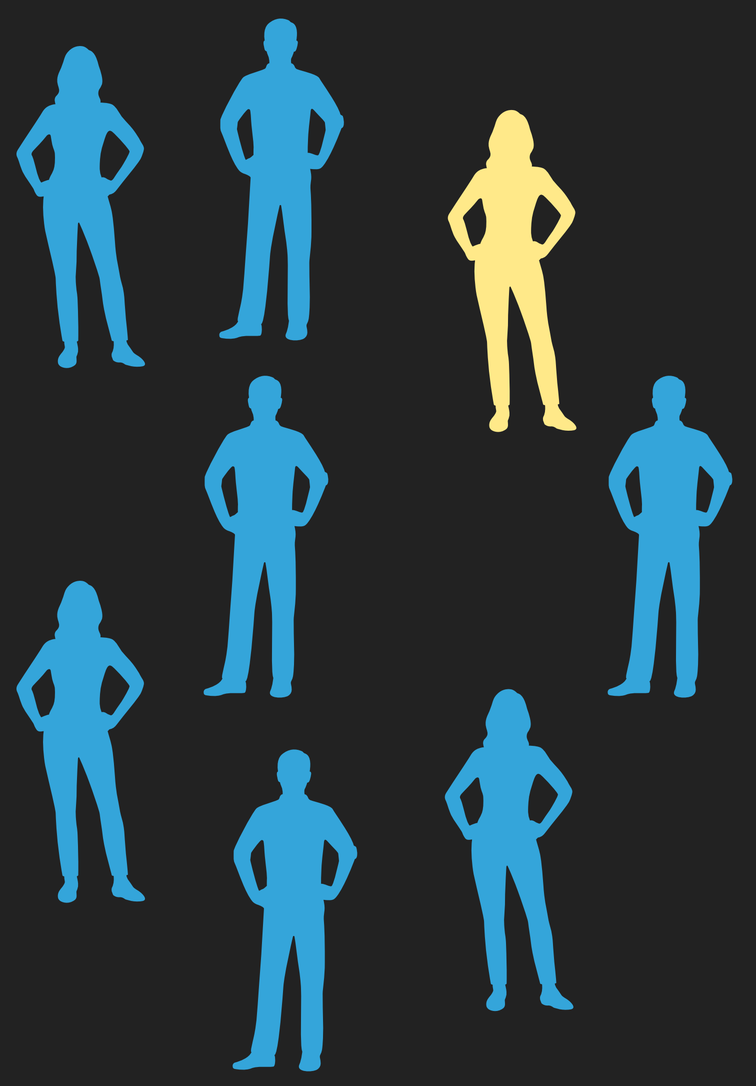
FROM KNOWING TO INTERVENING

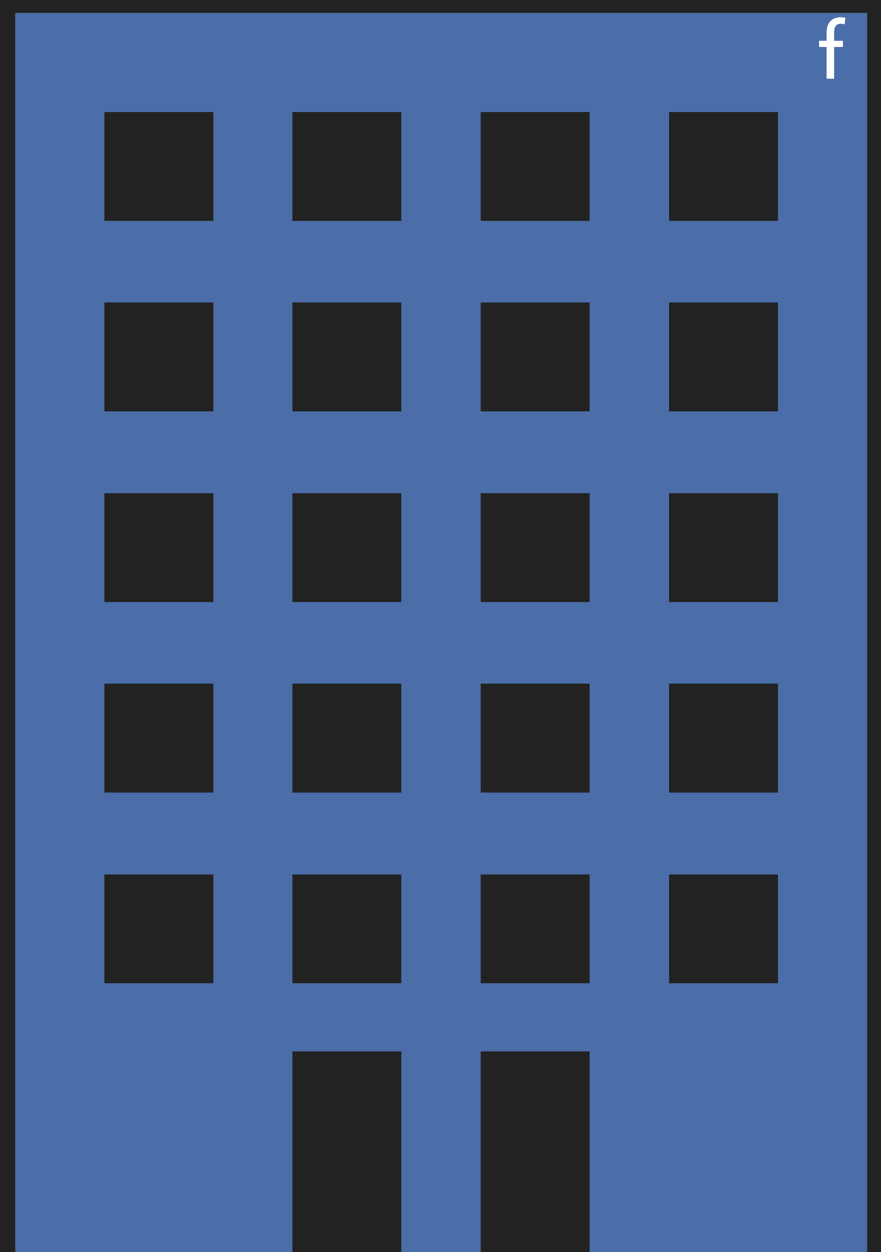
- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people



FROM KNOWING TO INTERVENING

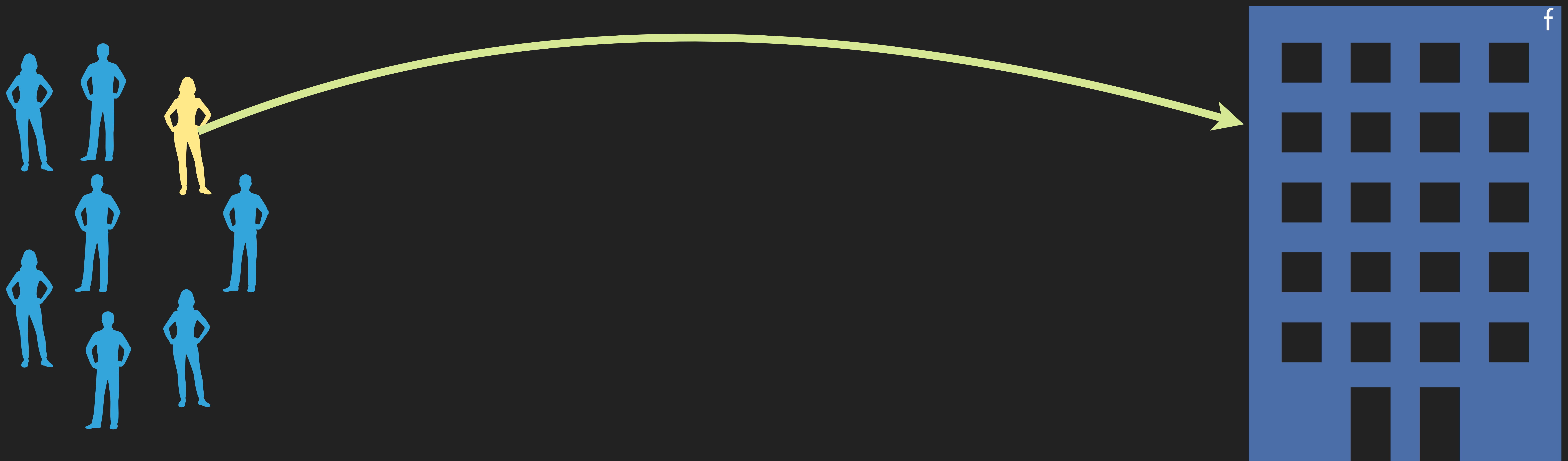
- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people





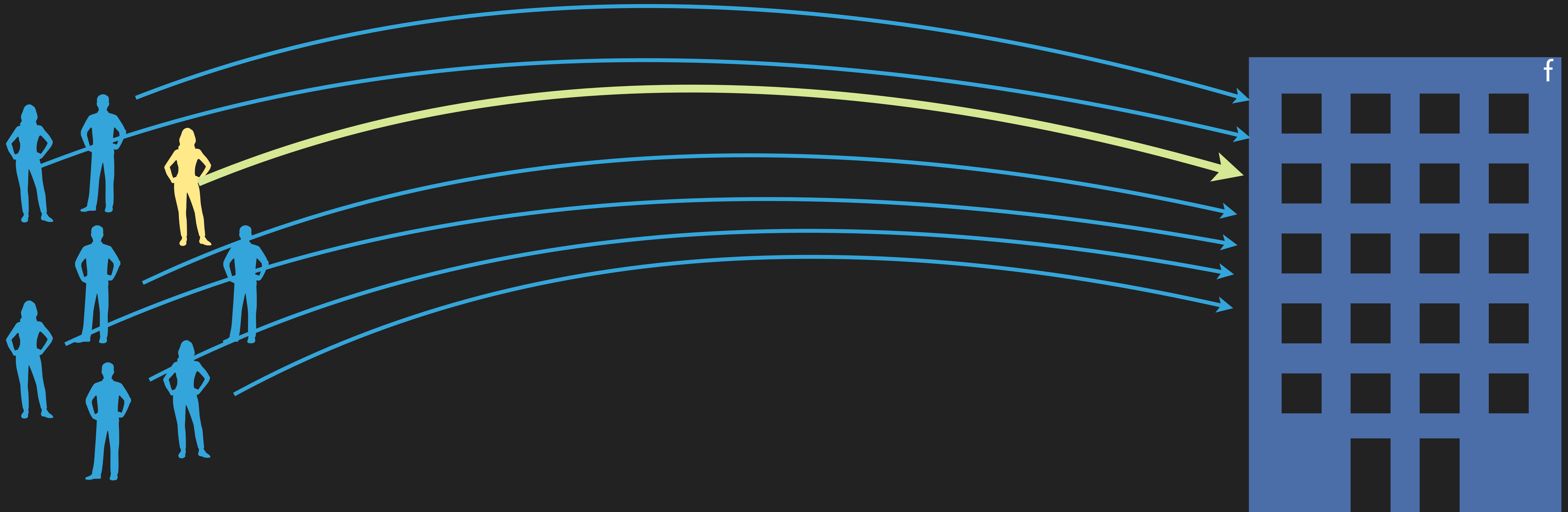
FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people



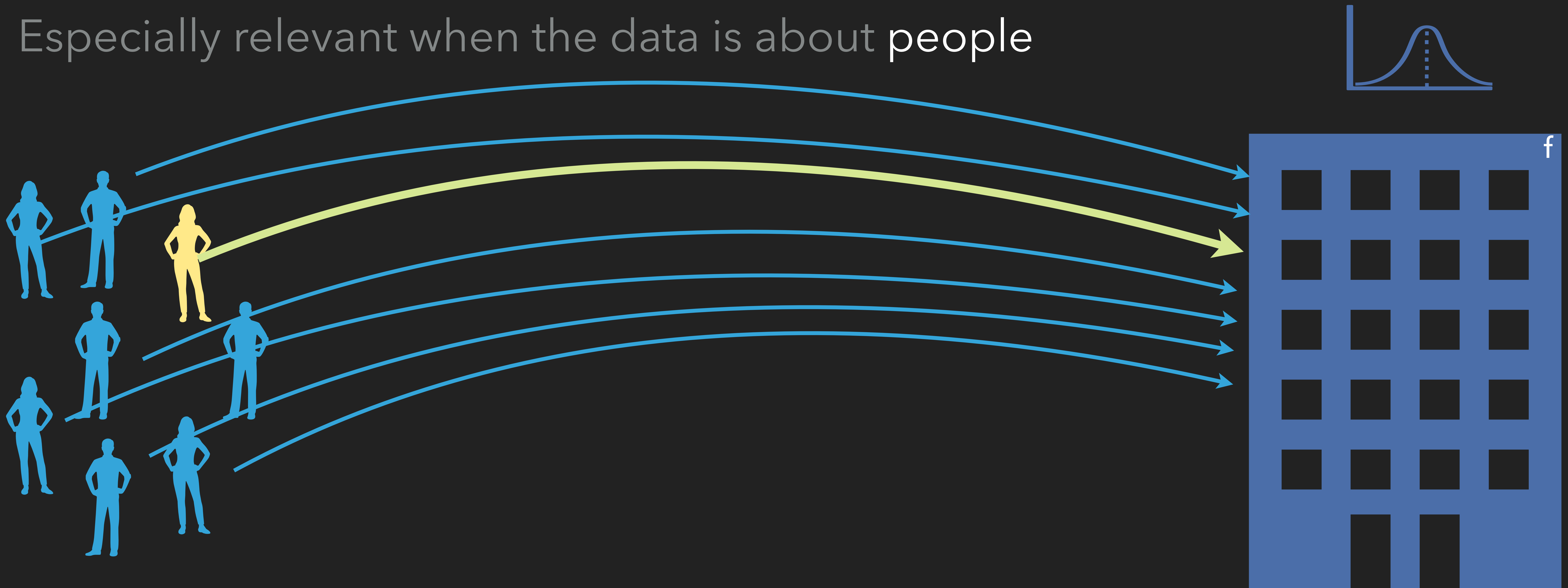
FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people



FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people



FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people

FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people

FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to **understand** the world, but to **change** it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about **people**

FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people

FROM KNOWING TO INTERVENING

- ML systems are now deployed in the world, taking Marx's exhortation to heart…
- Not just to understand the world, but to change it
- This necessitates a rethinking of what information and statistics are
- Especially relevant when the data is about people

FORMALISING ML: MINIMISE "EXPECTED RISK"

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y

Y X

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y
  - Goal: predict Y from X

Y X

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y
  - Goal: predict Y from X
  - Using an hypothesis h

$$Y \quad h(X)$$

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y
  - **Goal**: predict Y from X
  - Using an hypothesis h
  - X, Y are governed (drawn from) some fixed joint distribution P_{XY}

P_{XY} Y $h(X)$

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y
  - **Goal**: predict Y from X
  - Using an hypothesis h
  - X, Y are governed (drawn from) some fixed joint distribution P_{XY}
- The loss function ℓ judges how close h(X) is to Y; smaller is better

$$P_{XY} \quad \ell(Y, h(X))$$

FORMALISING ML: MINIMISE "EXPECTED RISK"

$$\mathbb{E}_{P_{XY}}\, \ell(Y, h(X))$$

- The world gives us X, Y
  - **Goal**: predict Y from X
  - Using an hypothesis h
  - X, Y are governed (drawn from) some fixed joint distribution P_{XY}
- The loss function ℓ judges how close h(X) is to Y; smaller is better
- Take the average of the loss with respect to P_{XY}

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y
  - Goal: predict Y from X
  - Using an hypothesis h
  - X, Y are governed (drawn from) some fixed joint distribution P_{XY}
- The loss function ℓ judges how close $h(X)$ is to Y; smaller is better
- Take the average of the loss with respect to P_{XY}

$$\arg\min_{h} \mathbb{E}_{P_{XY}} \ell(Y, h(X))$$

- Specific goal: find the h that minimises the average loss

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y
  - Goal: predict Y from X
  - Using an hypothesis h
  - X, Y are governed (drawn from) some fixed joint distribution P_{XY}
- The loss function ℓ judges how close $h(X)$ is to Y; smaller is better
- Take the average of the loss with respect to P_{XY}

$$\arg\min_{h} \mathbb{E}_{P_{XY}} \ell(Y, h(X))$$

- Specific goal: find the h that minimises the average loss
  - where h is a function $\mathscr{X} \to \mathscr{Y}$

FORMALISING ML: MINIMISE "EXPECTED RISK"

- The world gives us X, Y
  - Goal: predict Y from X
  - Using an hypothesis h
  - X, Y are governed (drawn from) some fixed joint distribution P_{XY}
- The loss function ℓ judges how close $h(X)$ is to Y; smaller is better
- Take the average of the loss with respect to P_{XY}

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

- Specific goal: find the h that minimises the average loss
  - where h is a function $\mathcal{X} \to \mathcal{Y}$
- Actually lower our sights; fix a hypothesis class $\mathcal{H}$ and consider…

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

WHAT TO STUDY?

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

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

WHAT TO STUDY?

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

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

- $\mathscr{H}$ the "model class" – the focus of much ML research

WHAT TO STUDY?

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

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

- $\mathscr{H}$ the "model class" – the focus of much ML research
- $\arg\min$ algorithms to optimise – the focus of most of the remainder

WHAT TO STUDY?

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

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

- $\mathscr{H}$ the "model class" – the focus of much ML research
- $\arg\min$ algorithms to optimise – the focus of most of the remainder
- h the hypothesis – focus of "explainability"

WHAT TO STUDY?

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

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

- $\mathscr{H}$ the "model class" – the focus of much ML research
- $\arg\min$ algorithms to optimise – the focus of most of the remainder
- h the hypothesis – focus of "explainability"
- ℓ the loss function, how performance is judged

The geometry and calculus of losses

WHAT TO STUDY?

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

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

- $\mathscr{H}$ the "model class" – the focus of much ML research
- $\arg\min$ algorithms to optimise – the focus of most of the remainder
- h the hypothesis – focus of "explainability"
- ℓ the loss function, how performance is judged
- $\mathbb{E}$ expectation to aggregate individual losses / encode fairness

The geometry and calculus of losses

WHAT TO STUDY?

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

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

- $\mathcal{H}$ the "model class" – the focus of much ML research
- $\arg\min$ algorithms to optimise – the focus of most of the remainder
- h the hypothesis – focus of "explainability"
- ℓ the loss function, how performance is judged

The geometry and calculus of losses

- $\mathbb{E}$ expectation to aggregate individual losses / encode fairness
- $\mathcal{S}$ the set system (usually a σ-algebra) – the set of measurable events

WHAT TO STUDY?

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

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

- $\mathcal{H}$ the "model class" – the focus of much ML research
- $\arg\min$ algorithms to optimise – the focus of most of the remainder
- h the hypothesis – focus of "explainability"
- ℓ the loss function, how performance is judged

The geometry and calculus of losses

- $\mathbb{E}$ 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"

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

WHAT TO STUDY?

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

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

- $\mathscr{H}$ the "model class" – the focus of much ML research
- $\arg\min$ algorithms to optimise – the focus of most of the remainder
- h the hypothesis – focus of "explainability"
- ℓ the loss function, how performance is judged

The geometry and calculus of losses

- $\mathbb{E}$ expectation to aggregate individual losses / encode fairness
- $\mathscr{S}$ 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"

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


BEYOND


BEYOND
Data, Information, Probability, Independence, Expectations


BENEATH

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

Considering what happens when we do not take data for granted

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

Considering what happens when we do not take data for granted

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

Considering what happens when we do not take data for granted

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

Considering what happens when we do not take data for granted

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

- Label noise - change the loss

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

Considering what happens when we do not take data for granted

- Label noise - change the loss
- Attribute noise - change the model class

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

Considering what happens when we do not take data for granted

- Label noise - change the loss
- Attribute noise - change the model class
- What about non-stochastic corruptions (who says the corruption process needs to be probabilistic?)

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

Considering what happens when we do not take data for granted

- Label noise - change the loss
- Attribute noise - change the model class
- What about non-stochastic corruptions (who says the corruption process needs to be probabilistic?)
- Selection bias – the stupidity of “big” data”:

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

Considering what happens when we do not take data for granted

- Label noise - change the loss
- Attribute noise - change the model class
- What about non-stochastic corruptions (who says the corruption process needs to be probabilistic?)
- Selection bias – the stupidity of "big" data":
  - Facebook survey on covid vaccine uptake.

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

Valerie C.Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. "Unrepresentative big surveys significantly overestimated US vaccine uptake." *Nature* 600, no. 7890 (2021): 695-700.

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

Considering what happens when we do not take data for granted

- Label noise - change the loss
- Attribute noise - change the model class
- What about non-stochastic corruptions (who says the corruption process needs to be probabilistic?)
- Selection bias – the stupidity of "big" data":
  - Facebook survey on covid vaccine uptake.
  - Sample size 250,000

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

Valerie C.Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. "Unrepresentative big surveys significantly overestimated US vaccine uptake." *Nature* 600, no. 7890 (2021): 695-700.

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

Considering what happens when we do not take data for granted

- Label noise - change the loss
- Attribute noise - change the model class
- What about non-stochastic corruptions (who says the corruption process needs to be probabilistic?)
- Selection bias – the stupidity of "big" data":
  - Facebook survey on covid vaccine uptake.
  - Sample size 250,000
  - Effective sample size due to selection bias: 10

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

Valerie C.Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. "Unrepresentative big surveys significantly overestimated US vaccine uptake." *Nature* 600, no. 7890 (2021): 695-700.

BEYOND DATA: BENIGN AND MALIGNANT CORRUPTION

If something is a fact, then it is incontrovertible, and thus is not to be questioned.

What I chose not to question, and treat as incontrovertible, I call a fact.

Considering what happens when we do not take data for granted

- Label noise - change the loss
- Attribute noise - change the model class
- What about non-stochastic corruptions (who says the corruption process needs to be probabilistic?)
- Selection bias – the stupidity of "big" data":
  - Facebook survey on covid vaccine uptake.
  - Sample size 250,000
  - Effective sample size due to selection bias: 10
- Open question: how to model selection bias?

A General Framework for Learning under Corruption: Label Noise, Attribute Noise, and Beyond

Valerie C.Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. "Unrepresentative big surveys significantly overestimated US vaccine uptake." *Nature* 600, no. 7890 (2021): 695-700.

BEYOND PROBABILITY

Suppose "the" probability does not exist. What to do instead?

BEYOND PROBABILITY

Suppose "the" probability does not exist. What to do instead?

It is now commonplace, in many domains, to see a general assumption that everything does have probability. … This unquestioning acceptance of mysterious probabilities may have many sources, but the authority and closed appearance of Kolmogorov's framework is surely one of them.

– Glenn Shafer (2015)

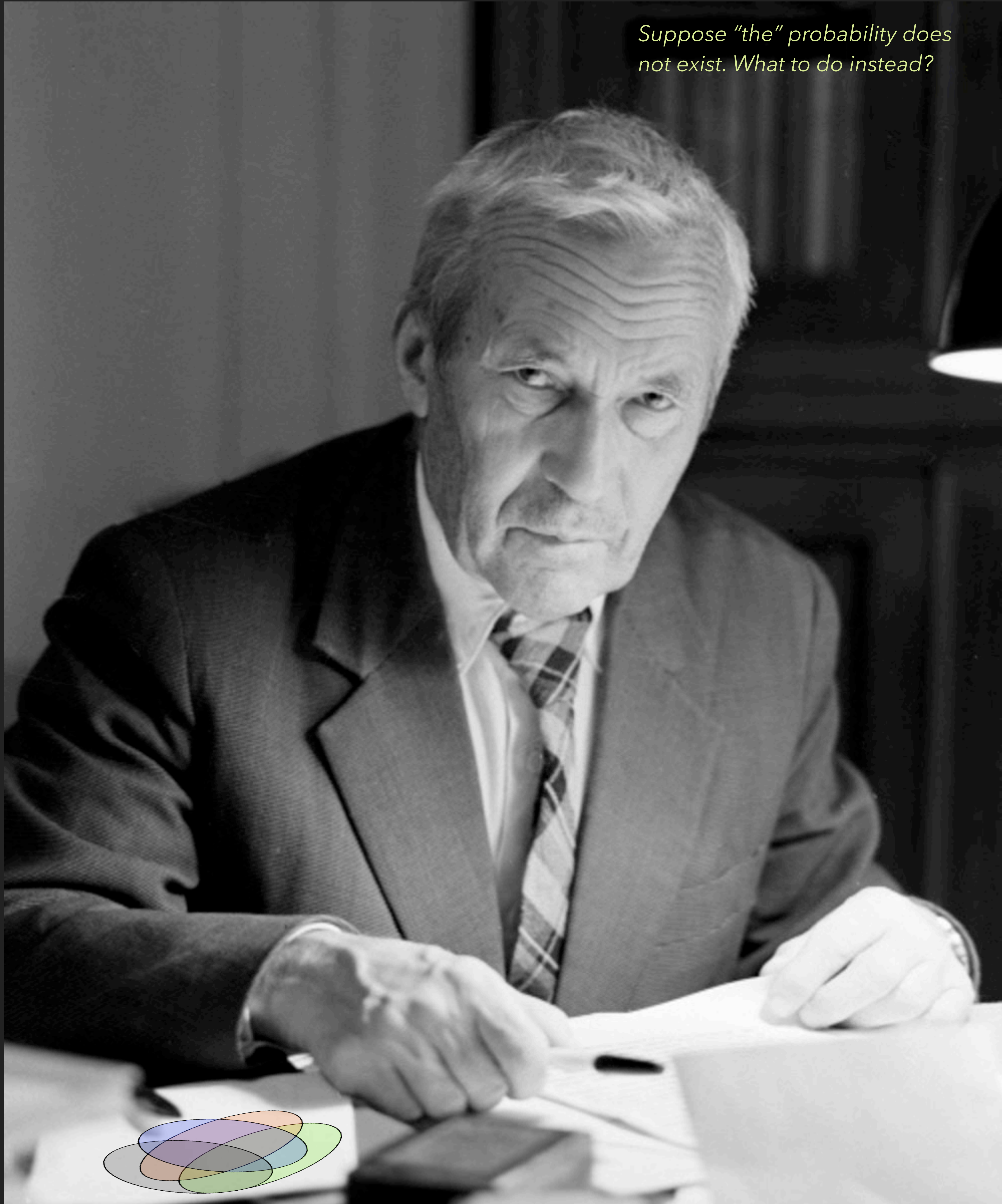
BEYOND PROBABILITY

It is now commonplace, in many domains, to see a general assumption that everything does have probability. … This unquestioning acceptance of mysterious probabilities may have many sources, but the authority and closed appearance of Kolmogorov's framework is surely one of them.

– Glenn Shafer (2015)

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)



BEYOND PROBABILITY

It is now commonplace, in many domains, to see a general assumption that everything does have probability. … This unquestioning acceptance of mysterious probabilities may have many sources, but the authority and closed appearance of Kolmogorov's framework is surely one of them.

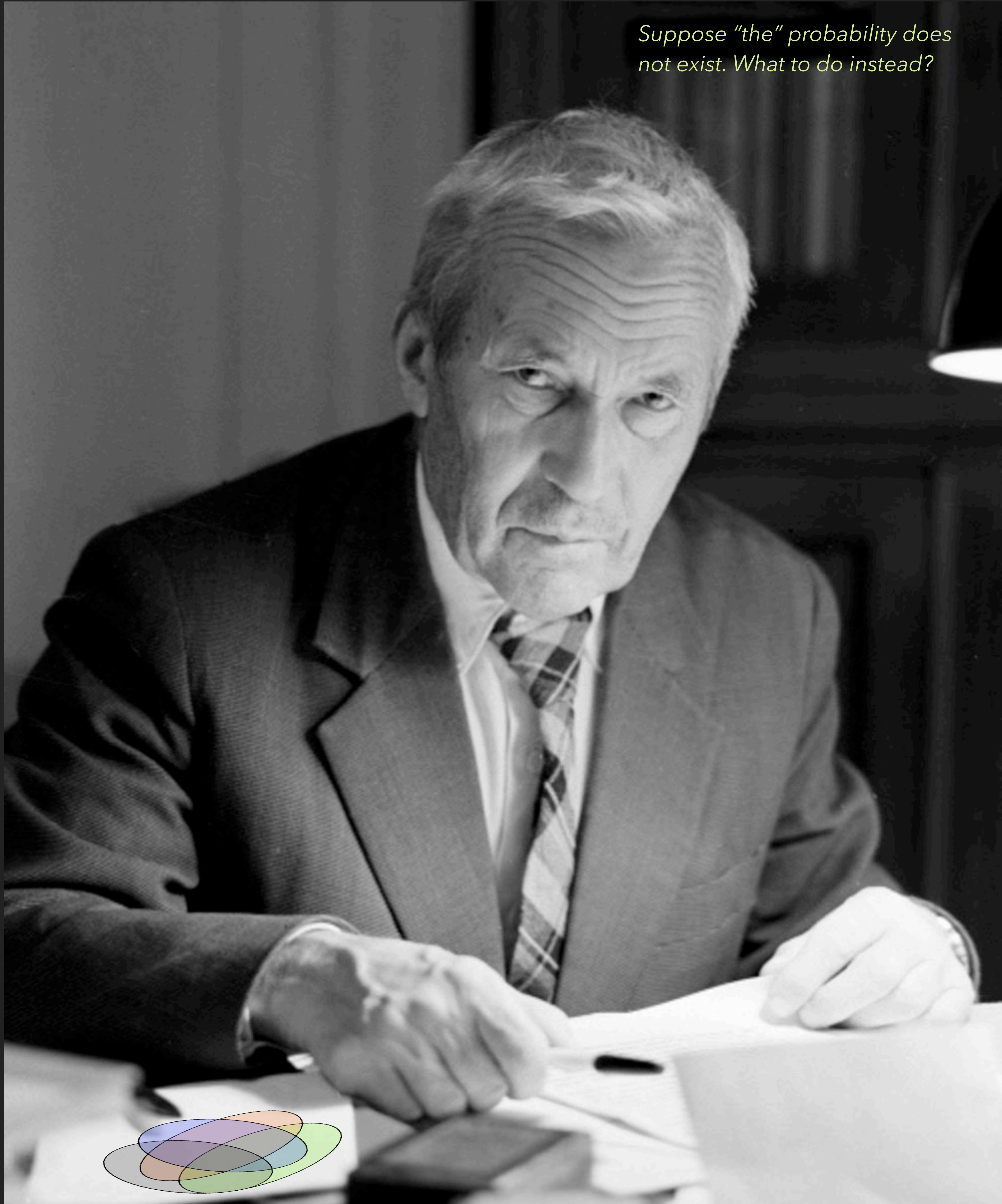
– Glenn Shafer (2015)

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)

The assumption that a definite probability … in fact exists for a given event under given conditions is a hypothesis which must be verified or justified in each individual case.

– Andrei Nikolaevich Kolmogorov (1951)



BEYOND PROBABILITY

It is now commonplace, in many domains, to see a general assumption that everything does have probability. … This unquestioning acceptance of mysterious probabilities may have many sources, but the authority and closed appearance of Kolmogorov's framework is surely one of them.

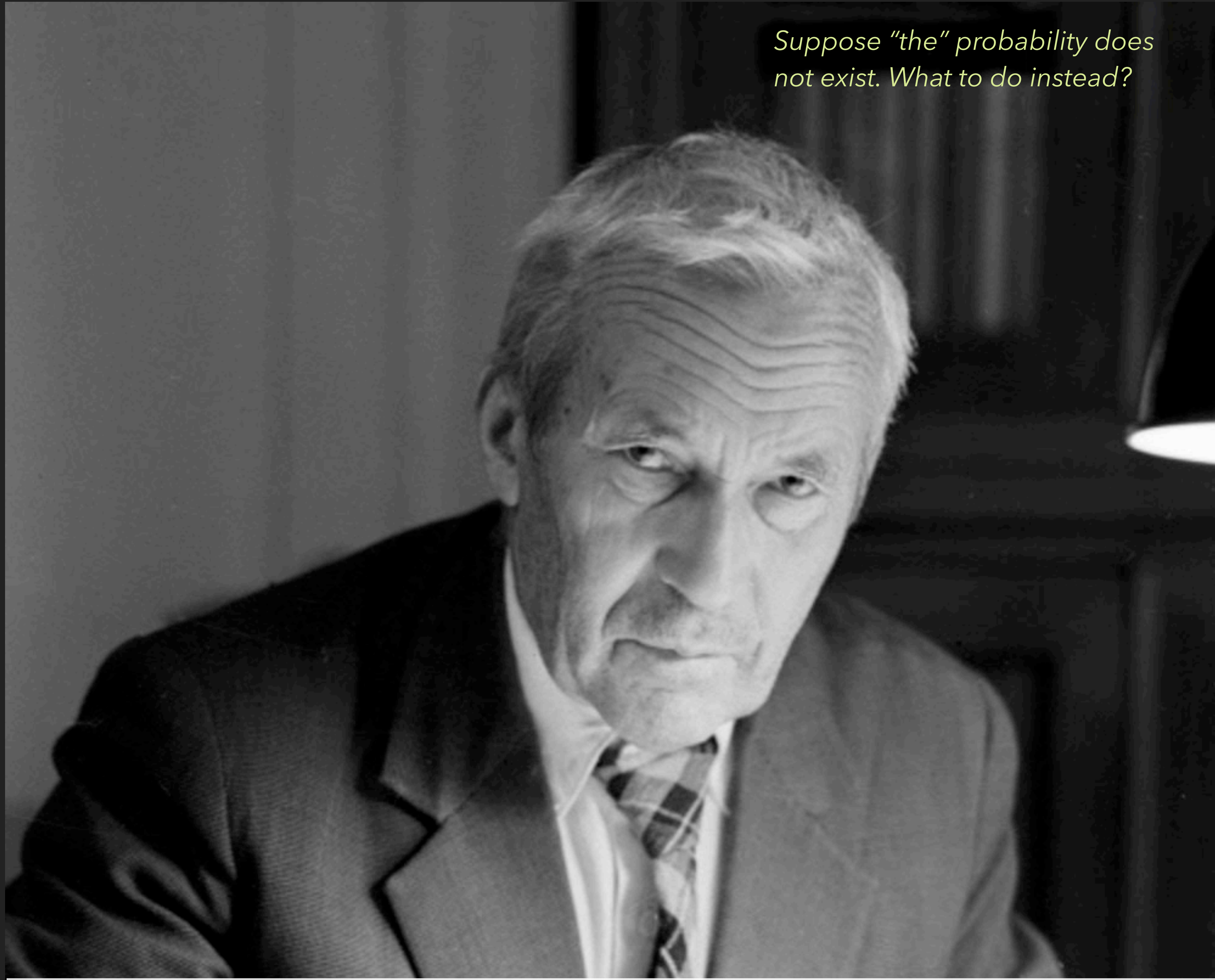
– Glenn Shafer (2015)

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)

The assumption that a definite probability … in fact exists for a given event under given conditions is a hypothesis which must be verified or justified in each individual case.

– Andrei Nikolaevich Kolmogorov (1951)



ON LOGICAL FOUNDATIONS OF PROBABILITY THEORY*) 1984

A.N.KOLMOGOROV

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).

BEYOND PROBABILITY

It is now commonplace, in many domains, to see a general assumption that everything does have probability. … This unquestioning acceptance of mysterious probabilities may have many sources, but the authority and closed appearance of Kolmogorov's framework is surely one of them.

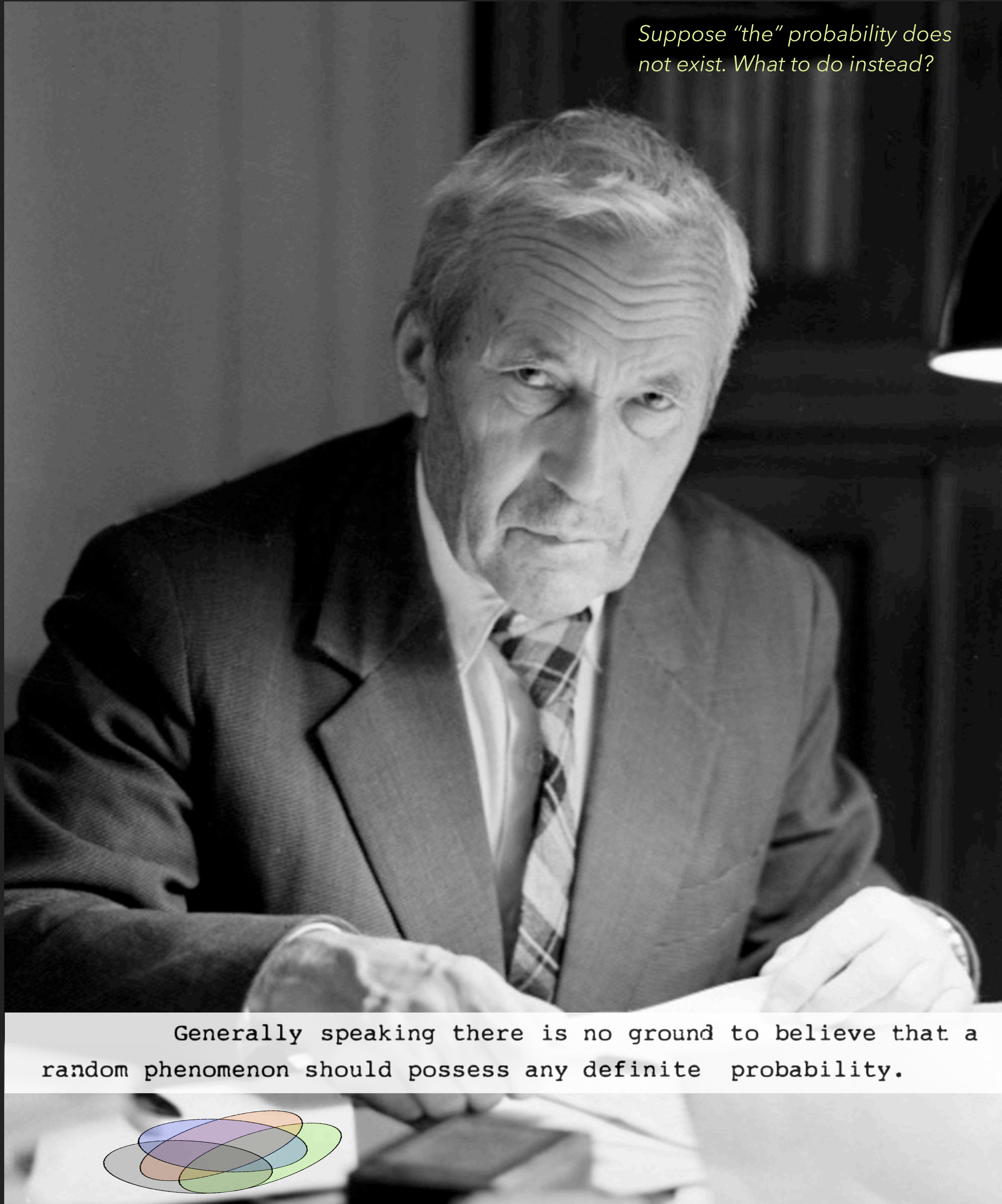
– Glenn Shafer (2015)

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)

The assumption that a definite probability … in fact exists for a given event under given conditions is a hypothesis which must be verified or justified in each individual case.

– Andrei Nikolaevich Kolmogorov (1951)



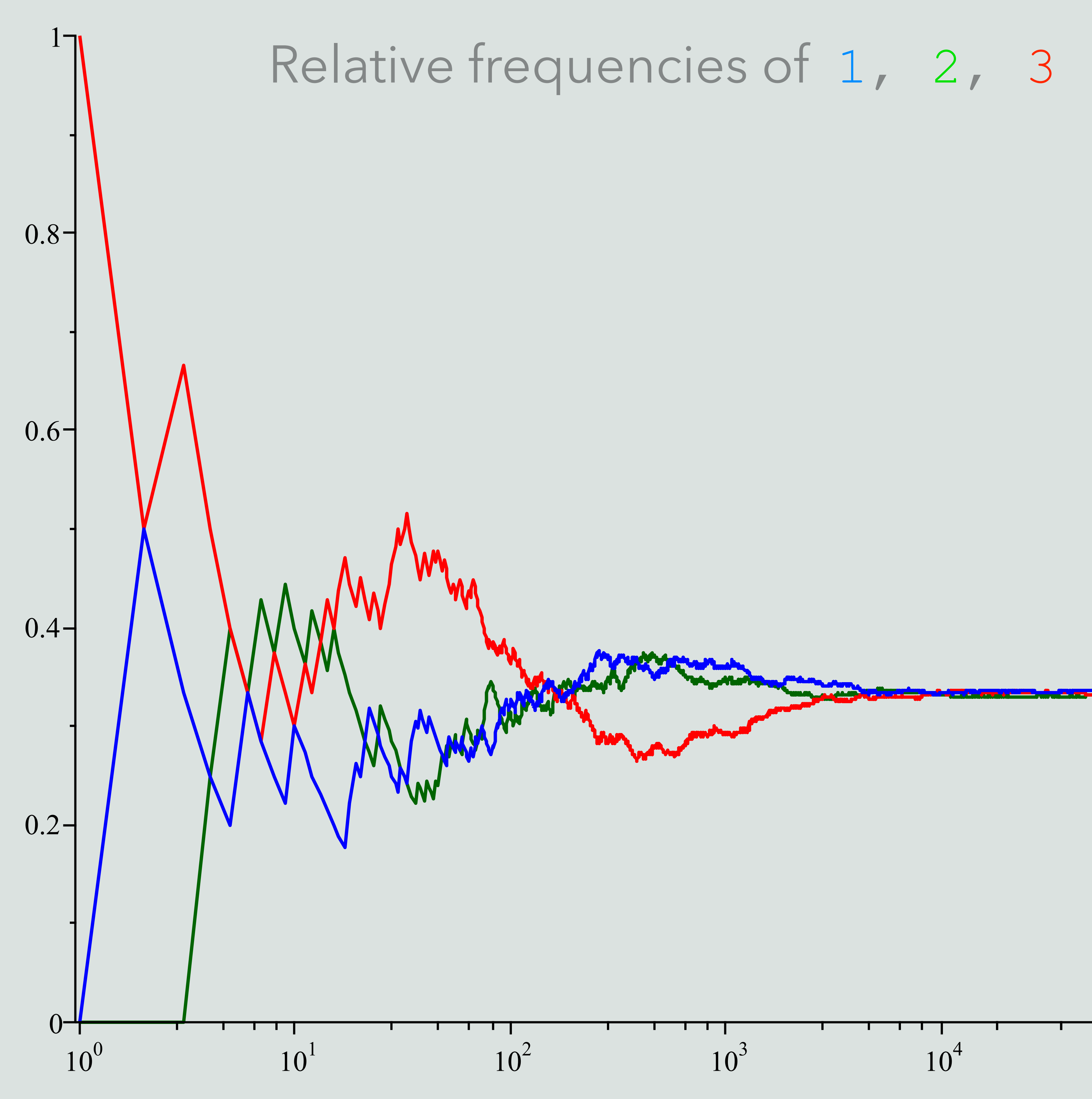
PROBABILITY AVERAGES AND LAWS

PROBABILITY AVERAGES AND LAWS

```
3, 1, 3, 2, 2, 1, 2, 3, 2, 1, 3, 2, 3, 3, 2, 3, 3, 1, 1, 3, 1, 1, 3, 2, 2, 3, 3, 3, 3, 3, 1, 3, 3, 1, 1
```

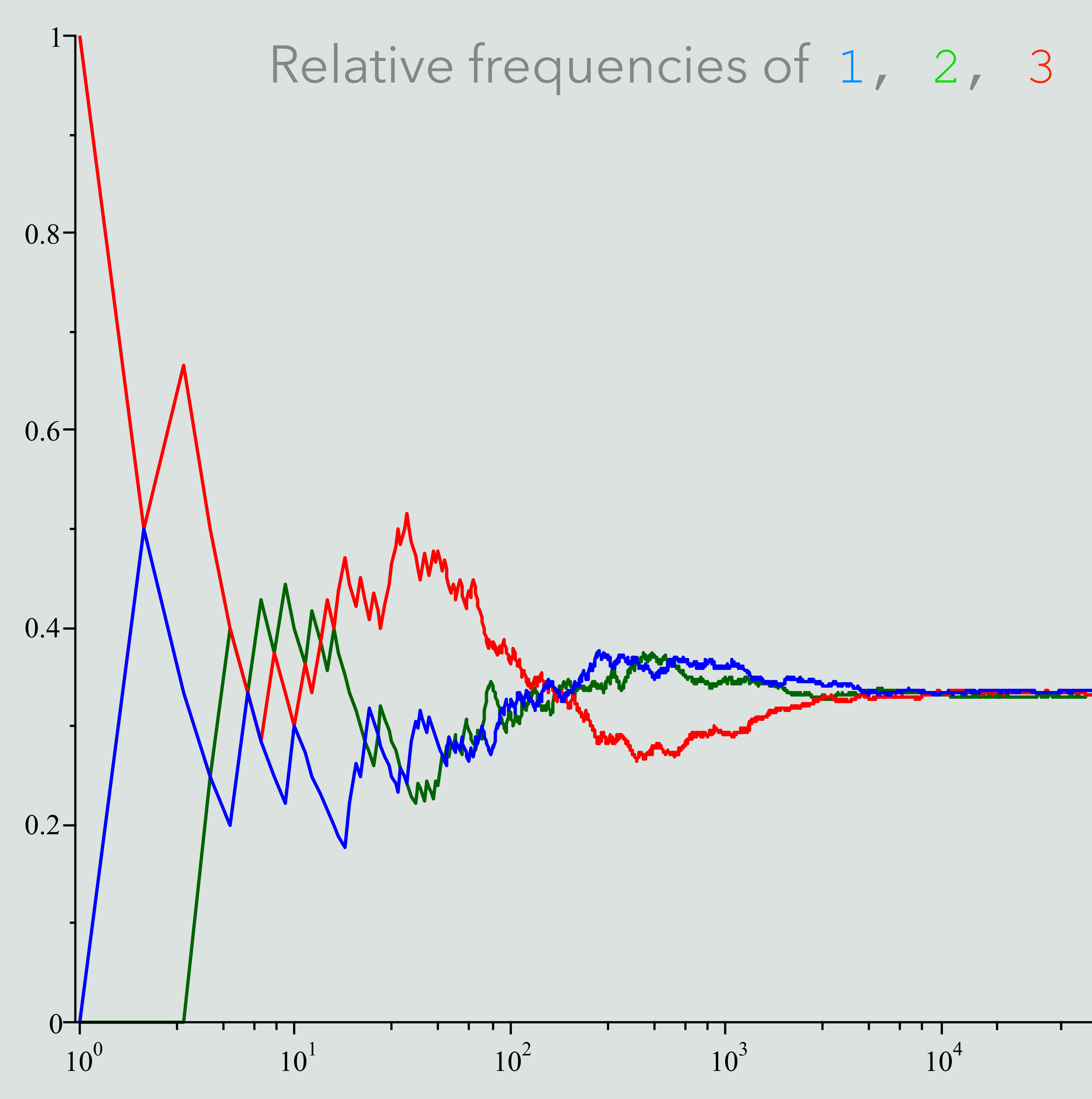
PROBABILITY AVERAGES AND LAWS

- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$



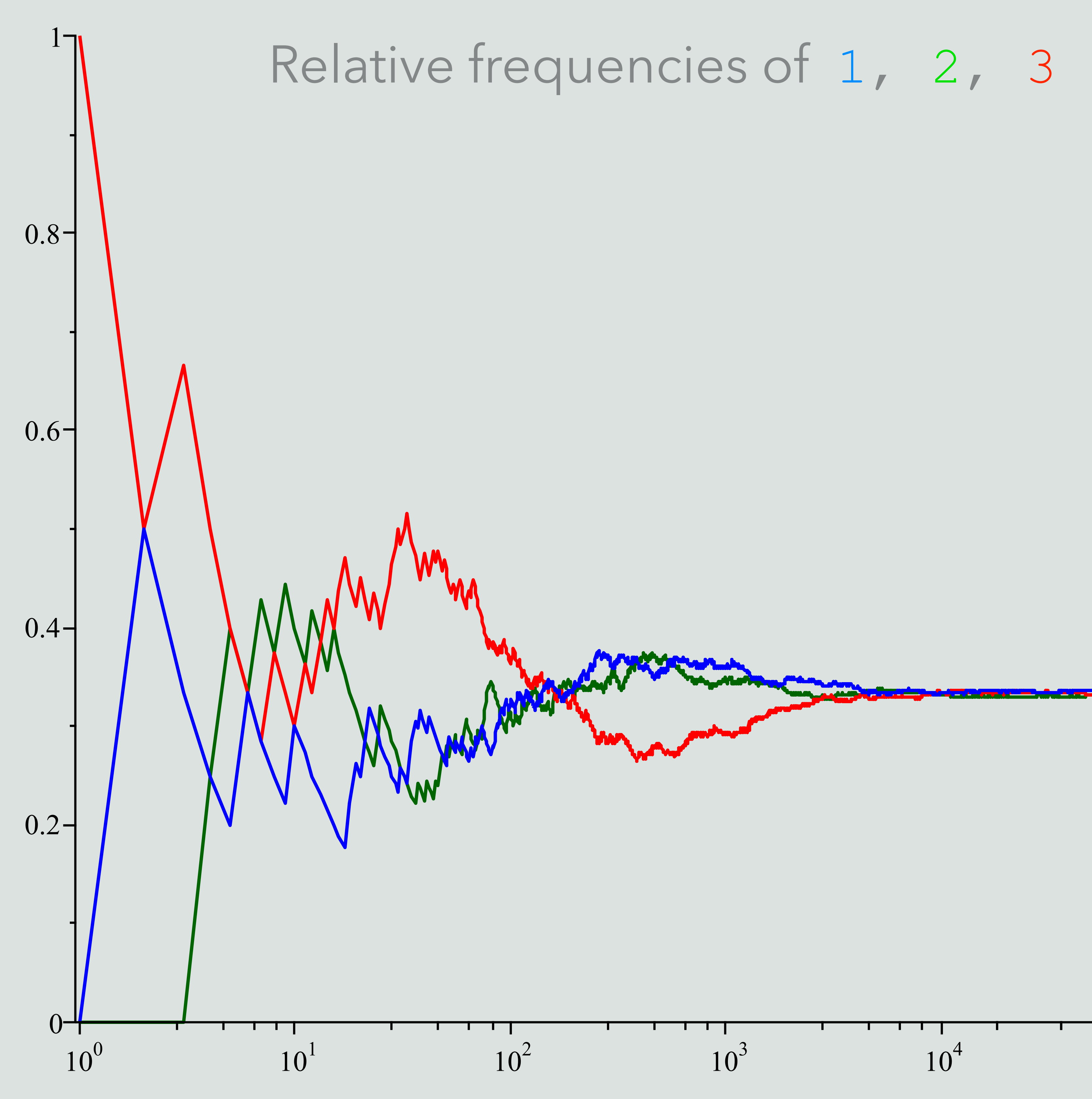
PROBABILITY AVERAGES AND LAWS

- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability



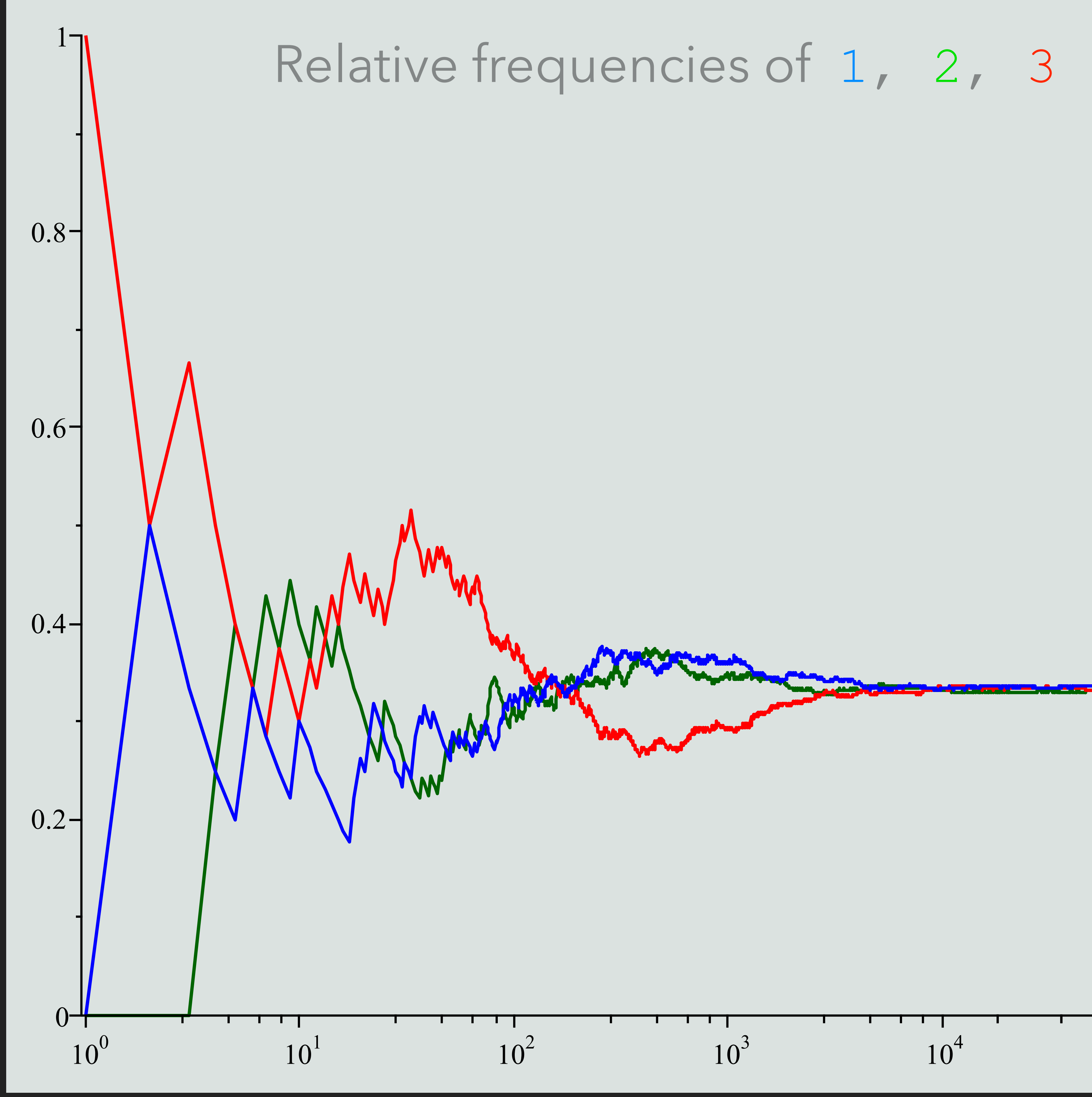
PROBABILITY AVERAGES AND LAWS

- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!

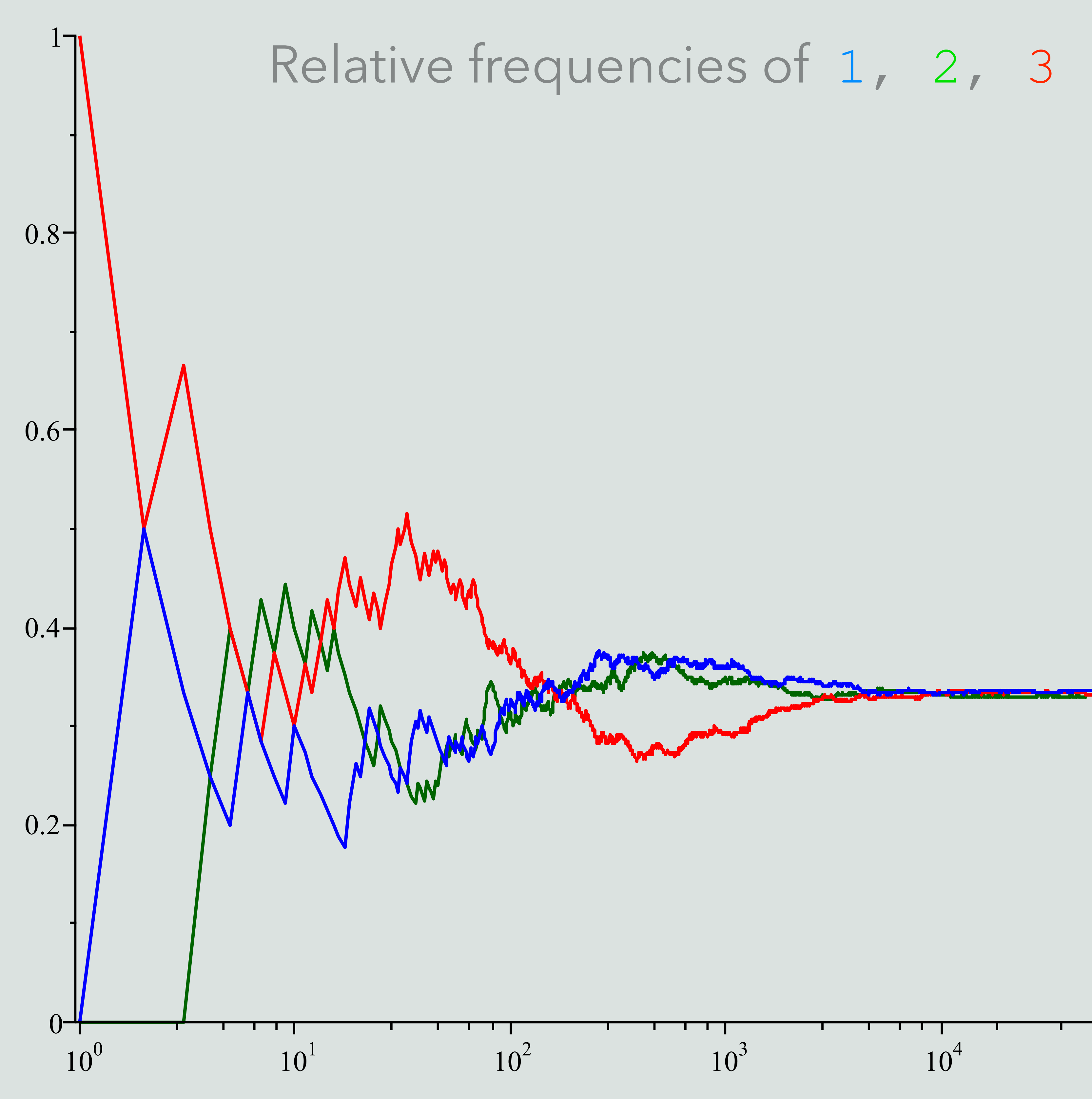


PROBABILITY AVERAGES AND LAWS

- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!
- What about all the other possible sequences?

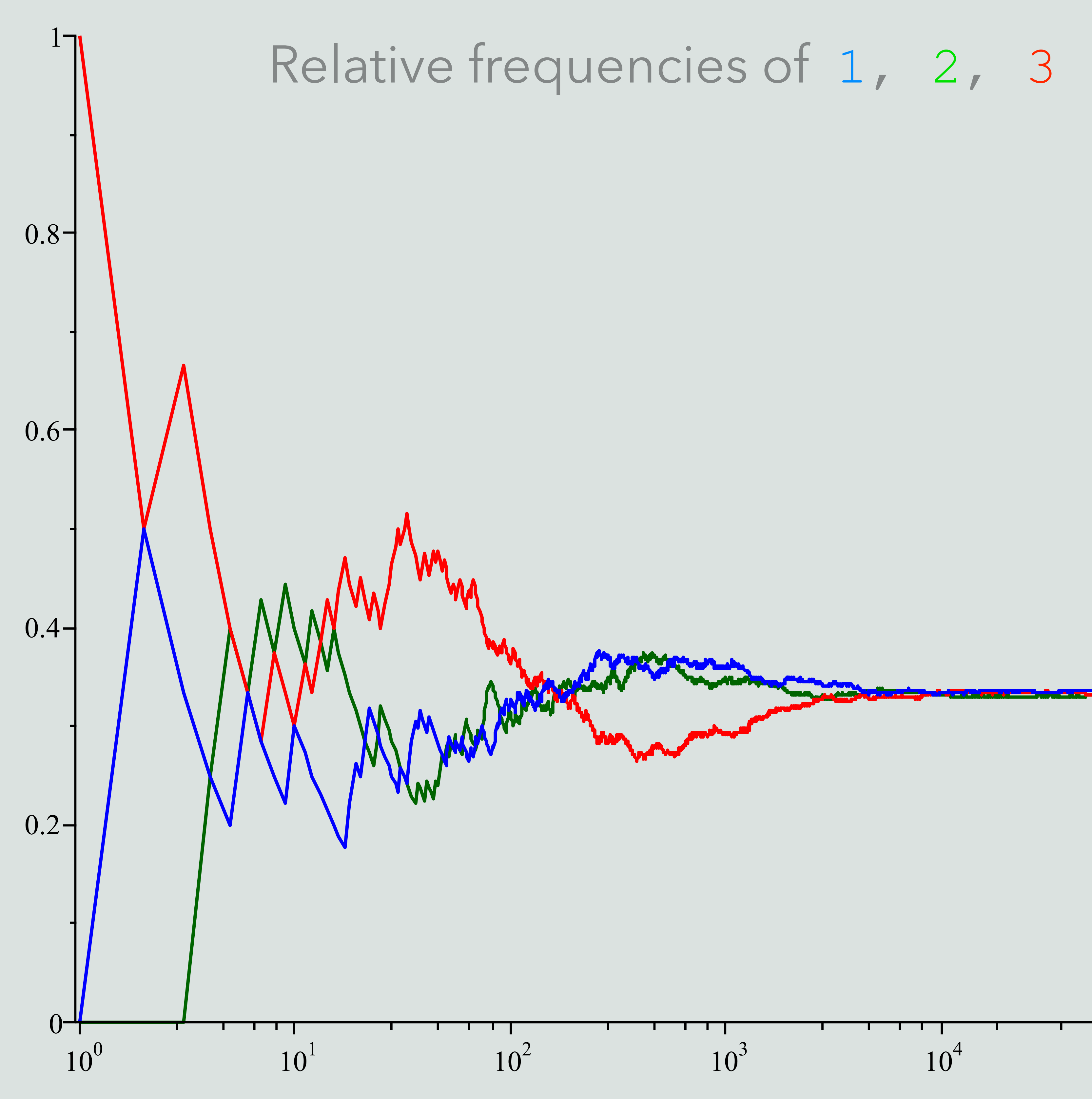


PROBABILITY AVERAGES AND LAWS



- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!
- What about all the other possible sequences?
  - "Most" of them behave similarly; just count for finite sequences

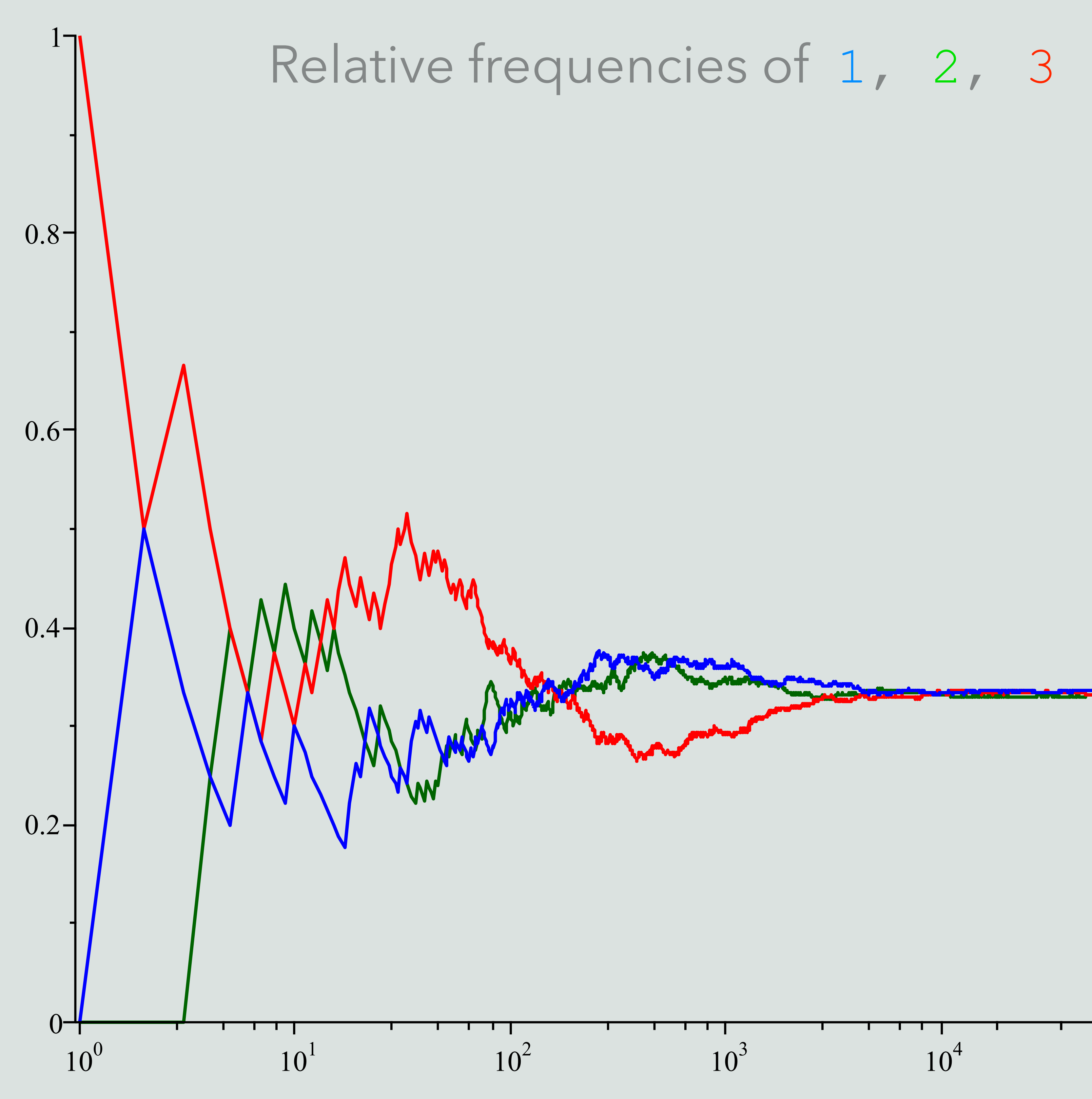
PROBABILITY AVERAGES AND LAWS



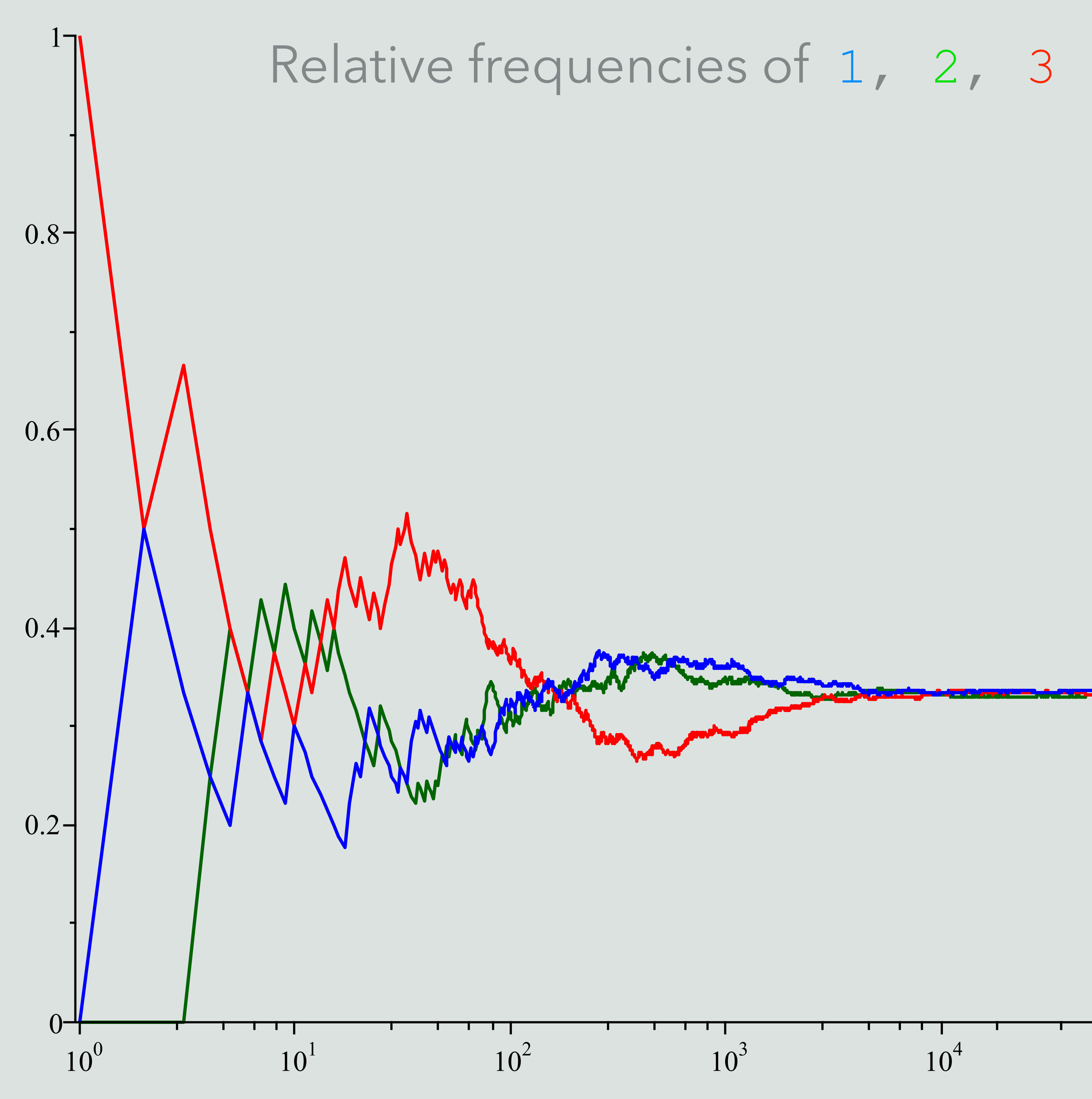
- The **relative frequencies** converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I **chose** the sequence to ensure this!
- What about all the other possible sequences?
  - **"Most"** of them behave similarly; just count for finite sequences
  - Then can argue about limits for infinite sequences

PROBABILITY AVERAGES AND LAWS

- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!
- What about all the other possible sequences?
    - "Most" of them behave similarly; just count for finite sequences
    - Then can argue about limits for infinite sequences
        - But there are many ways of quantifying "most", and the choice matters!

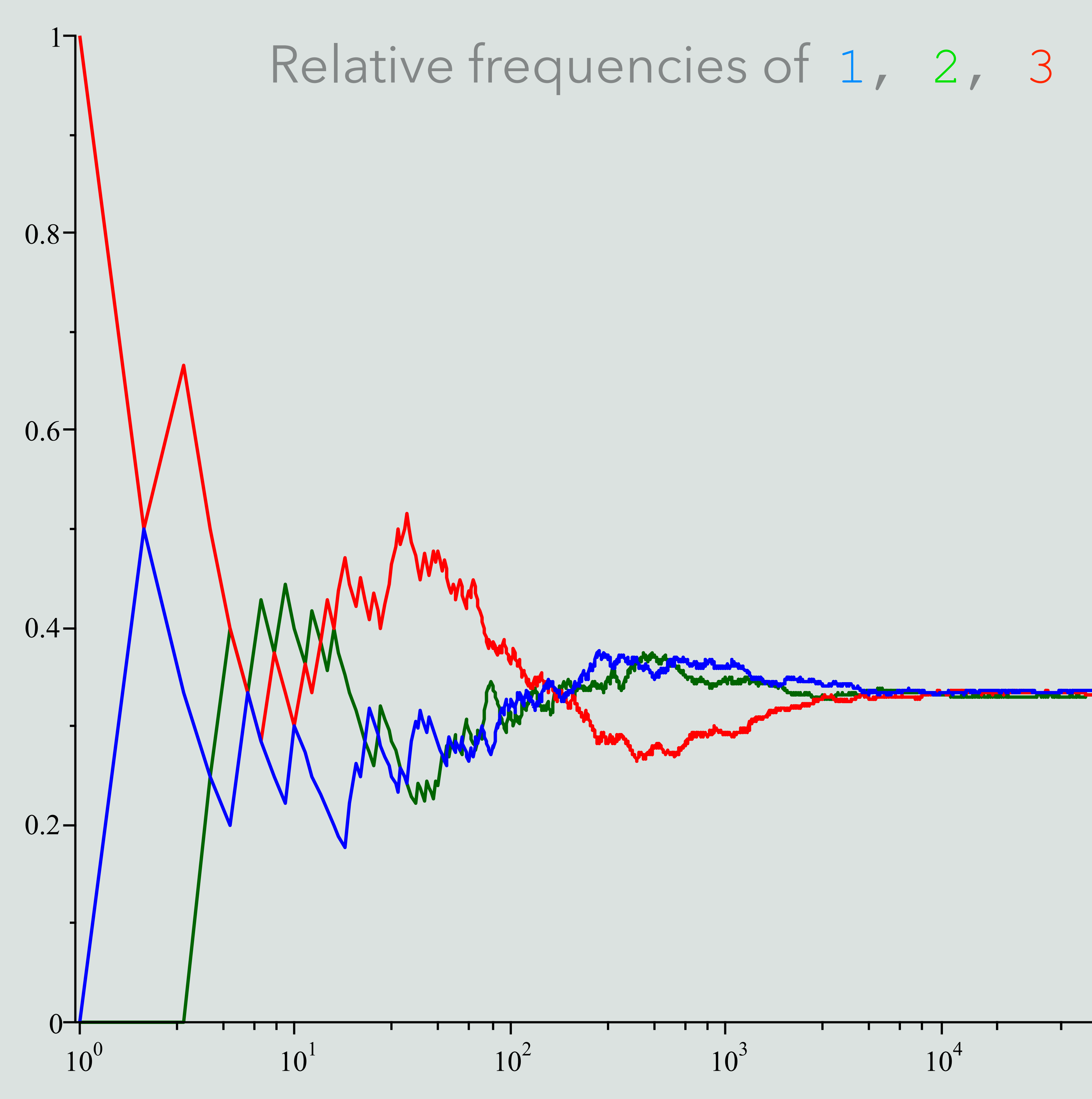


PROBABILITY AVERAGES AND LAWS



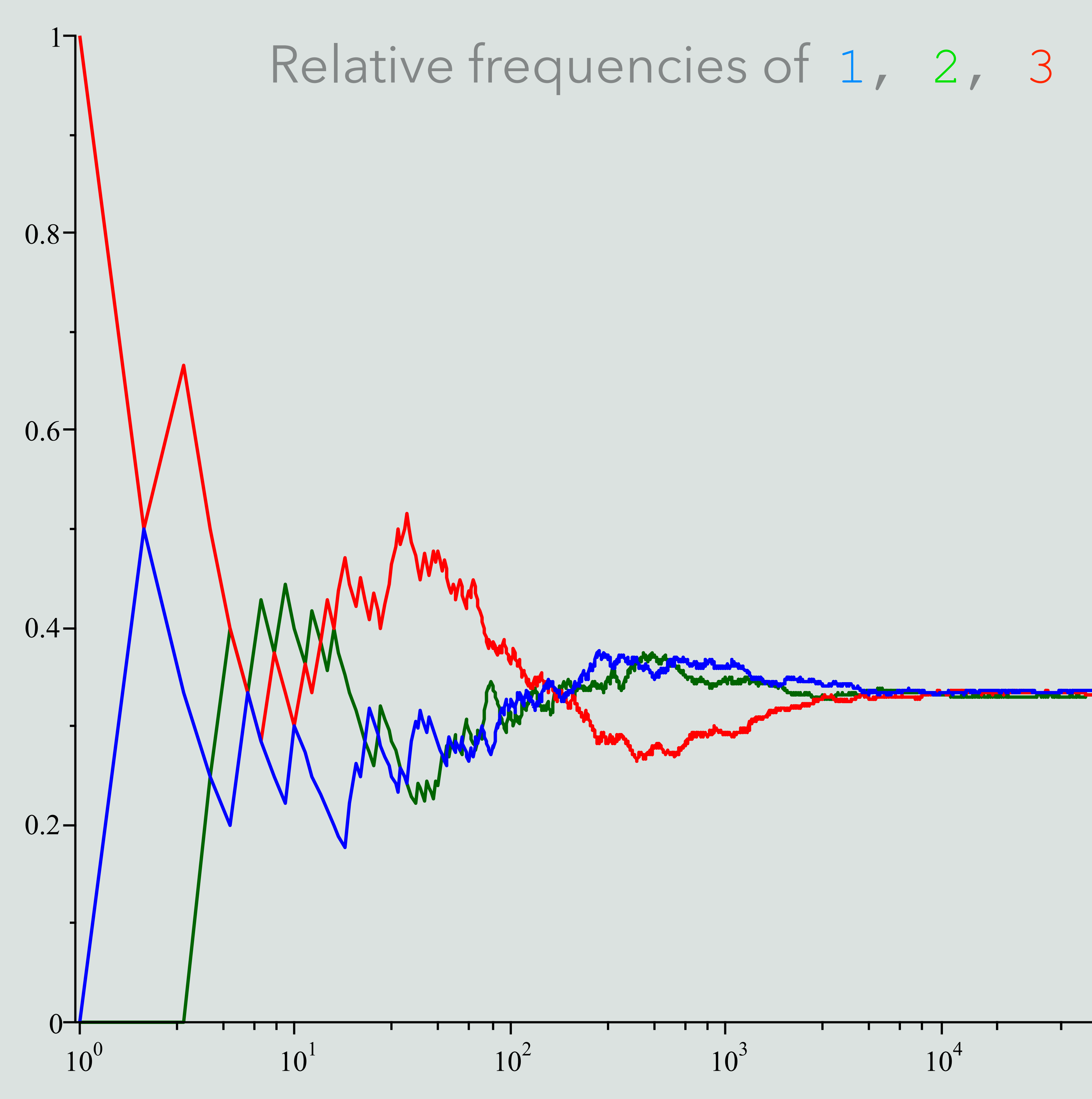
- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!
- What about all the other possible sequences?
  - "Most" of them behave similarly; just count for finite sequences
  - Then can argue about limits for infinite sequences
    - But there are many ways of quantifying "most", and the choice matters!
- And what does this say about the sequence you collect in the world?

PROBABILITY AVERAGES AND LAWS



- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!
- What about all the other possible sequences?
  - "Most" of them behave similarly; just count for finite sequences
  - Then can argue about limits for infinite sequences
    - But there are many ways of quantifying "most", and the choice matters!
- And what does this say about the sequence you collect in the world?
  - Nothing!!!!

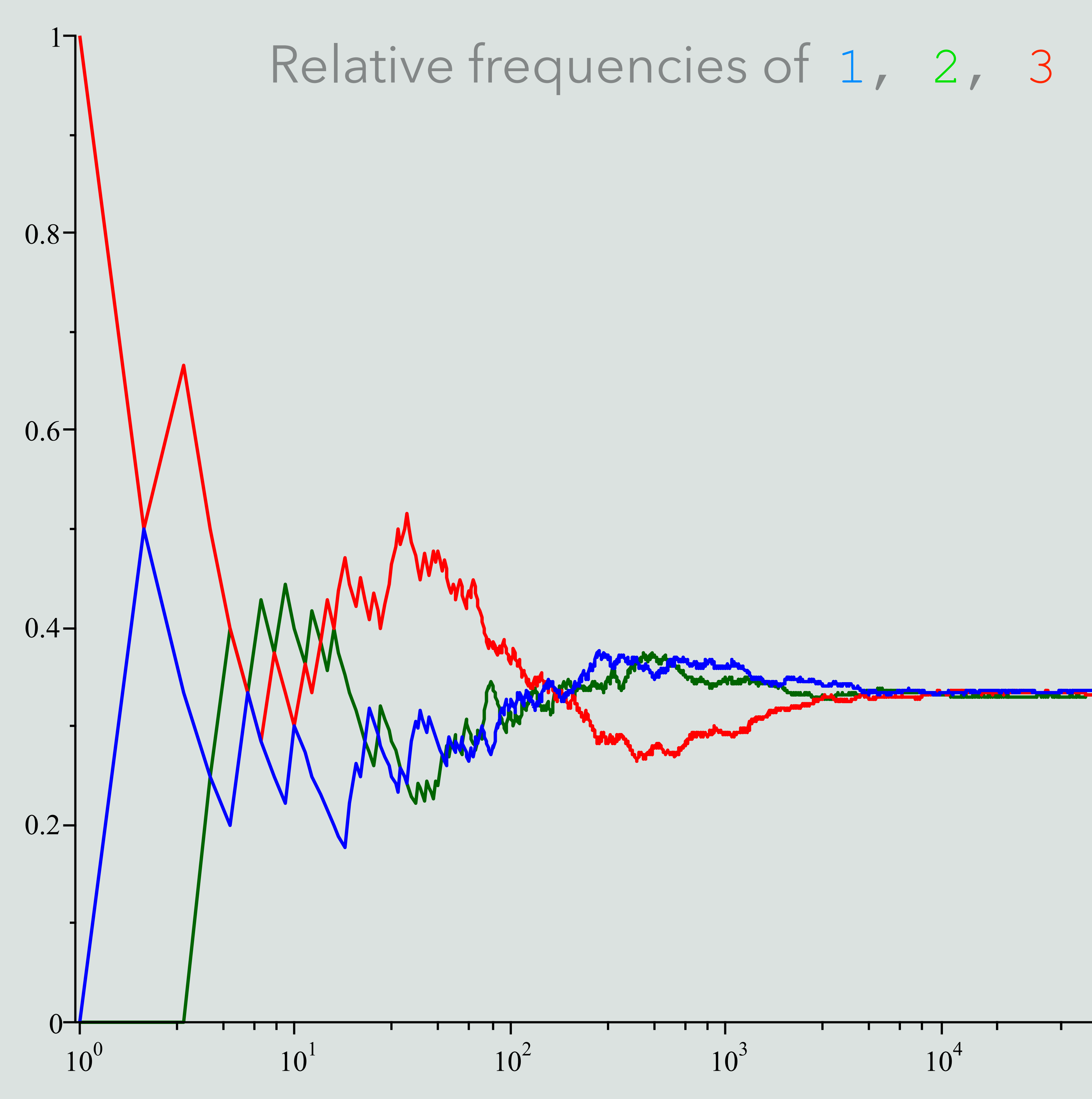
PROBABILITY AVERAGES AND LAWS



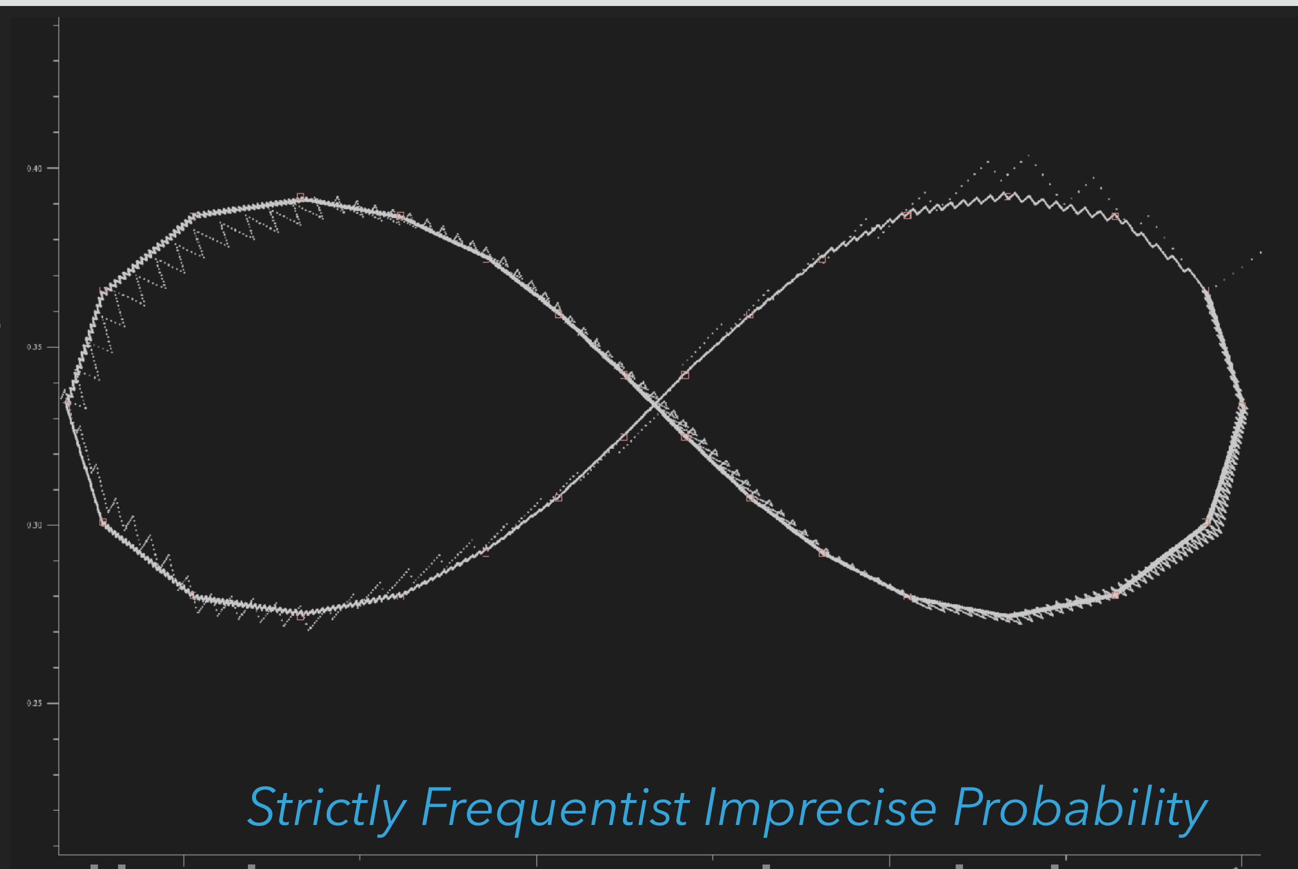
- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!
- What about all the other possible sequences?
    - "Most" of them behave similarly; just count for finite sequences
    - Then can argue about limits for infinite sequences
        - But there are many ways of quantifying "most", and the choice matters!
- And what does this say about the sequence you collect in the world?
    - Nothing!!!!
- It presumes some stability (the phenomenon that the averages converge is called "statistical stability")

Strictly Frequentist Imprecise Probability

PROBABILITY AVERAGES AND LAWS



- The relative frequencies converge to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- The limit exists, and is, *by definition*, "the" probability
- But there's a catch – I chose the sequence to ensure this!
- What about all the other possible sequences?
  - "Most" of them behave similarly; just count for finite sequences
  - Then can argue about limits for infinite sequences
    - But there are many ways of quantifying "most", and the choice matters!
- And what does this say about the sequence you collect in the world?
  - Nothing!!!!
- It presumes some stability (the phenomenon that the averages converge is called "statistical stability")
- *Most interesting stuff is not stable (non-equilibrium)*. Life, Society, Almost Everything!



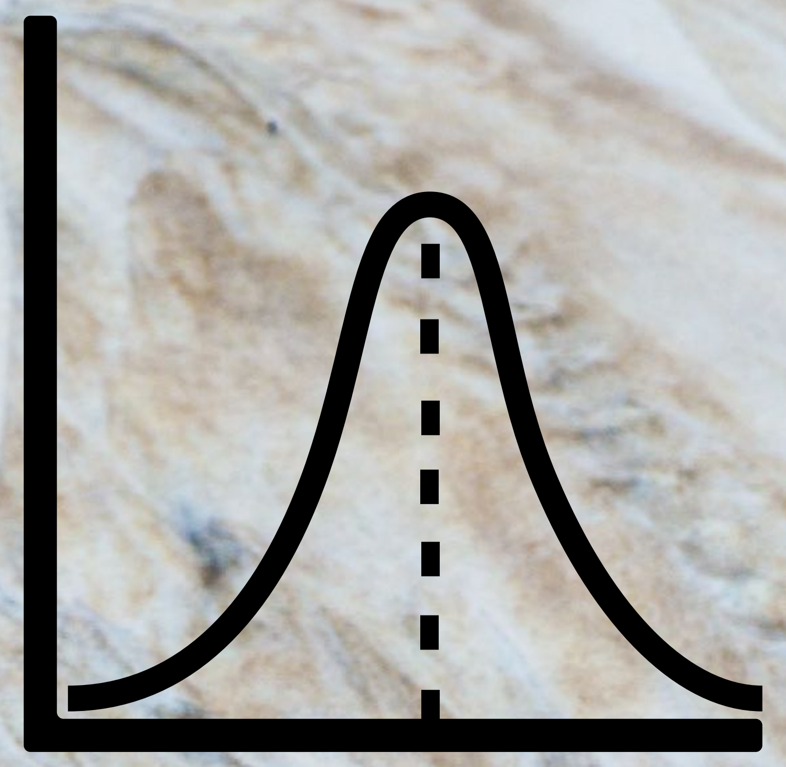
MASS PHENOMENA

MASS PHENOMENA

- Probability “mass”…

MASS PHENOMENA

- Probability “mass”…
- Imagined to be like sand



MASS PHENOMENA

- Probability “mass”…
- Imagined to be like sand

Ralph Alger Bagnold size-grading of sand by wind, Proc. R. Soc. Lond. **A163**, 250–264, 1937

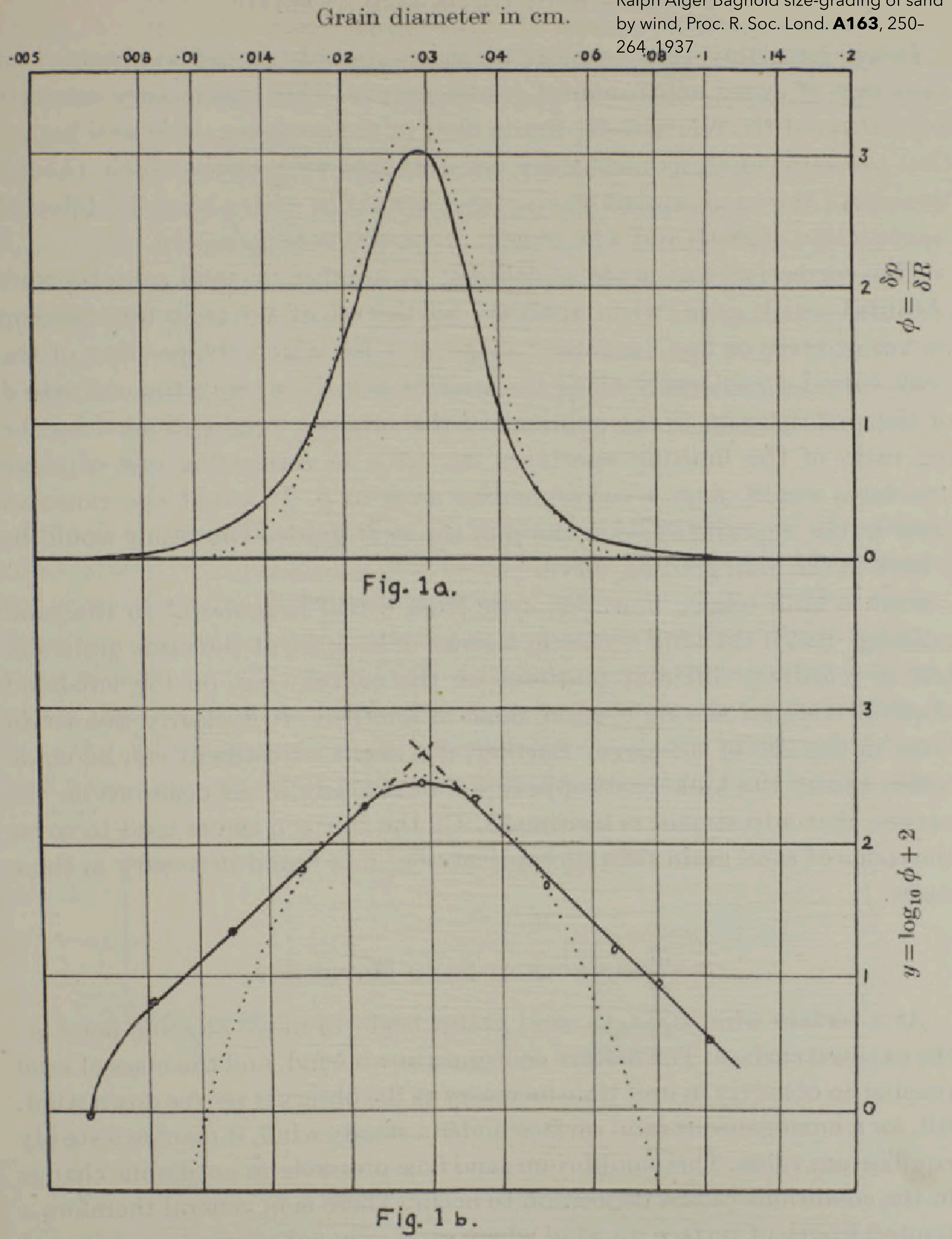


FIG. 1—Comparison of sand-grading curve (heavy line) with probability curve (dotted). *a*, ordinates on usual linear scale; *b*, ordinates on log scale.

MASS PHENOMENA

- Probability "mass"…
- Imagined to be like sand
- Data on people:
  treat people like sand…

```
462  % The true underlying data generating distribution
463  \newcommand{\pdata}{p_{\rm{data}}}
```

Ralph Alger Bagnold size-grading of sand by wind, Proc. R. Soc. Lond. **A163**, 250–264, 1937

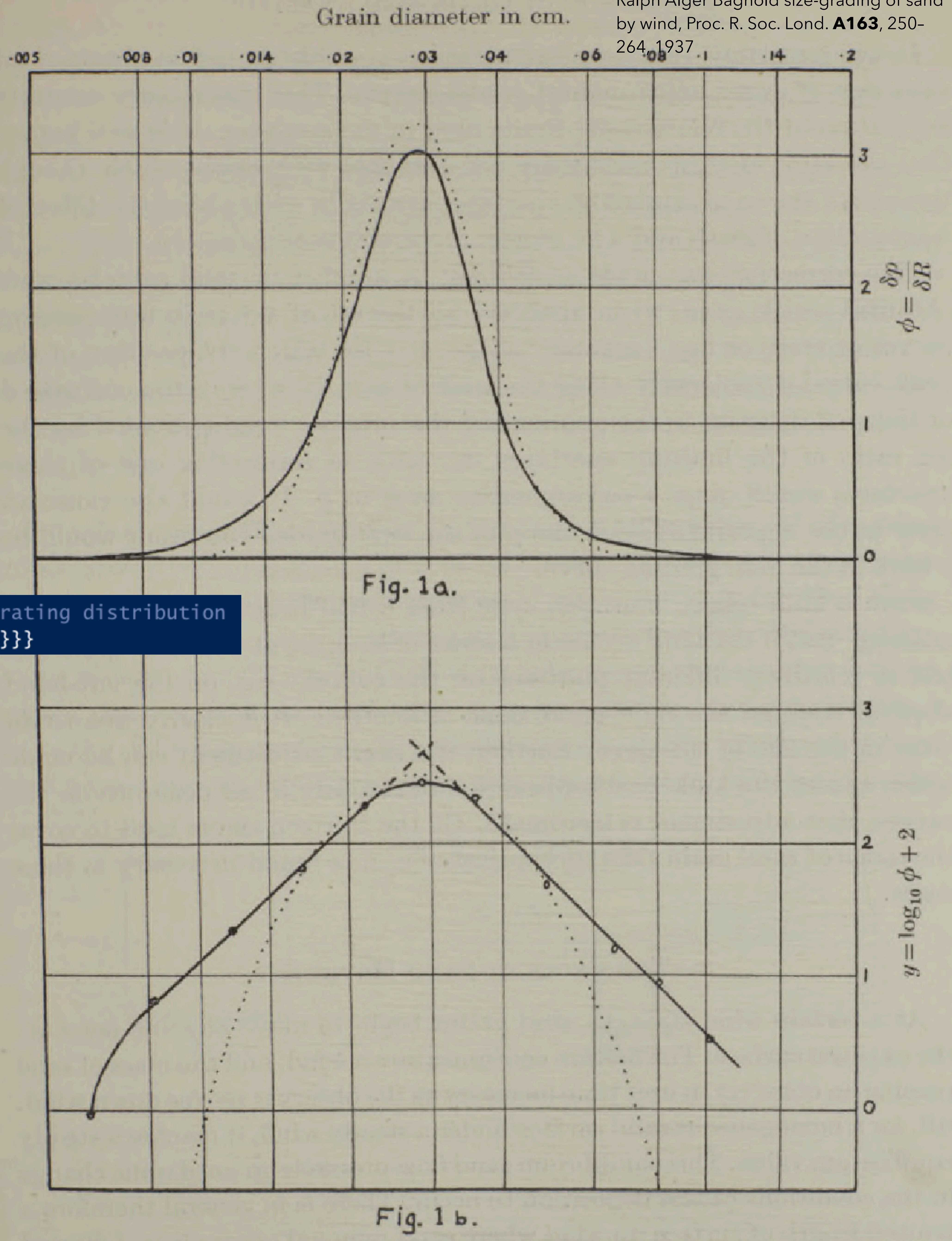


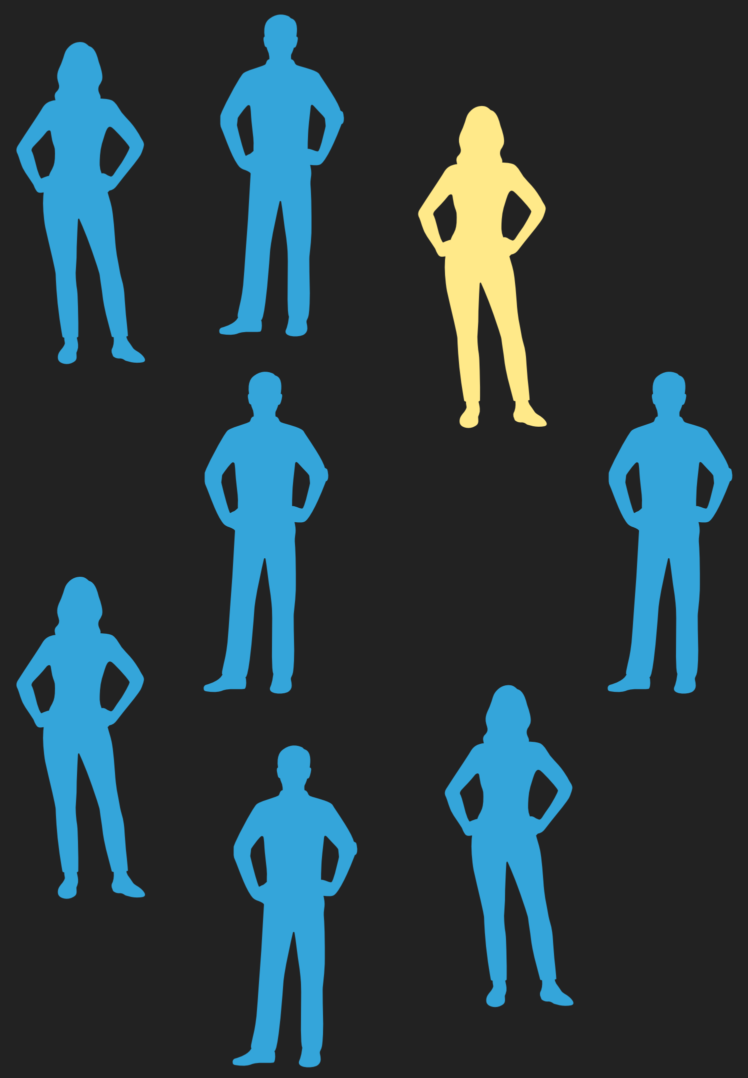
FIG. 1—Comparison of sand-grading curve (heavy line) with probability curve (dotted). *a*, ordinates on usual linear scale; *b*, ordinates on log scale.

MASS PHENOMEN

MASS PHENOMEN

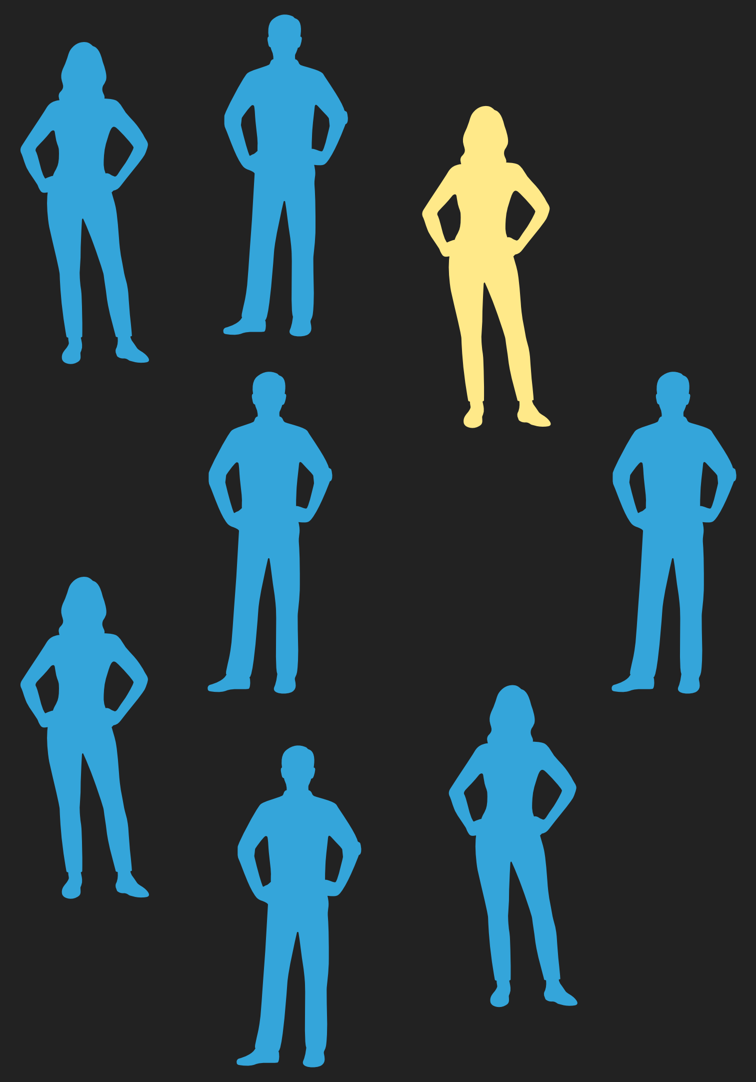
-
-
-
-
- What would a theory of mass phenomena that took account of individuals actually look like?

THE INDIVIDUAL AND THE AGGREGATE



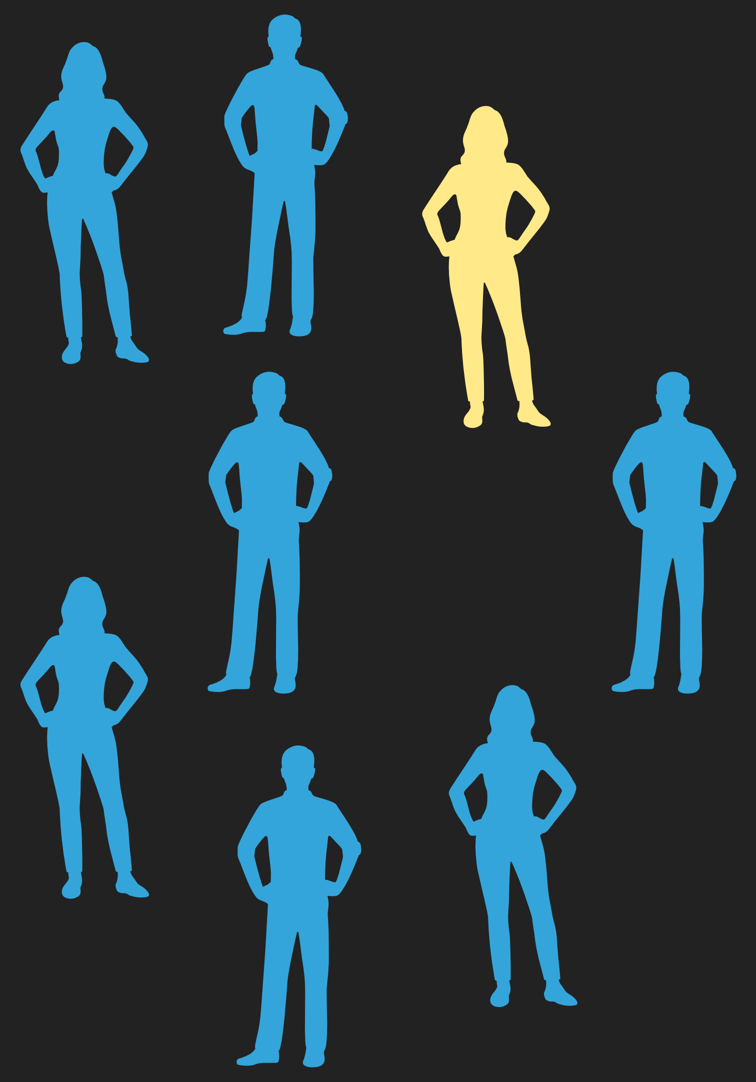
THE INDIVIDUAL AND THE AGGREGATE

- Starting with the data gives new insight into when something like probability exists



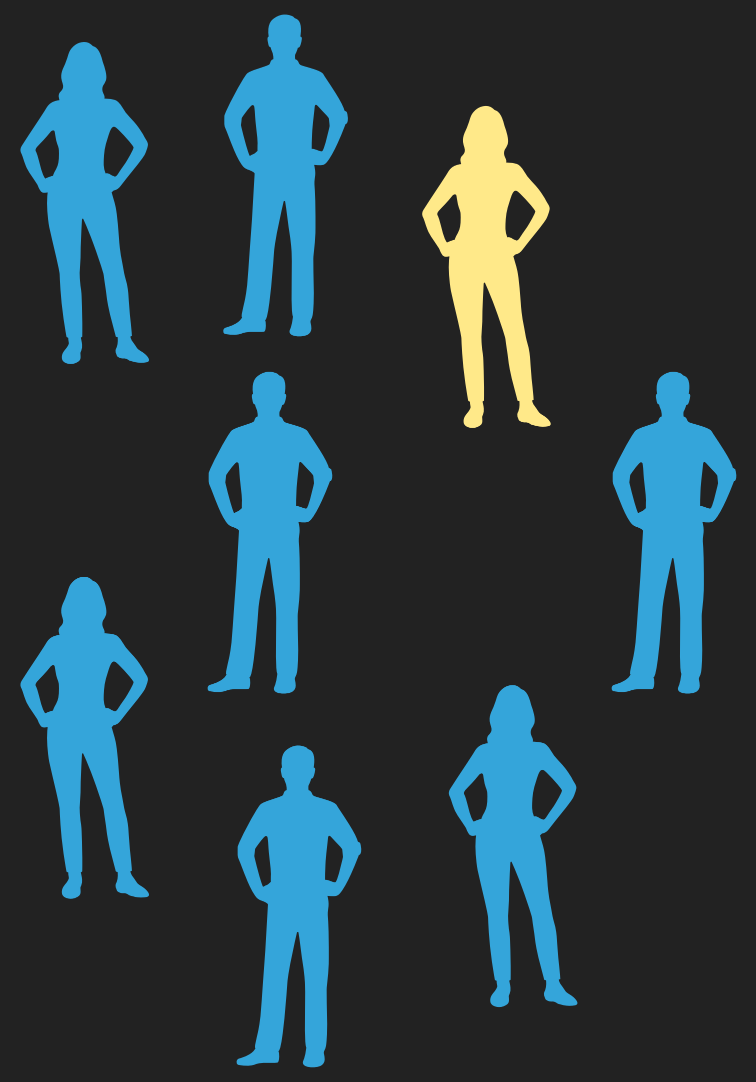
THE INDIVIDUAL AND THE AGGREGATE

- Starting with the data gives new insight into when something like probability exists
- But it does not answer the trickier question: what is the relationship between an individual and a probability?



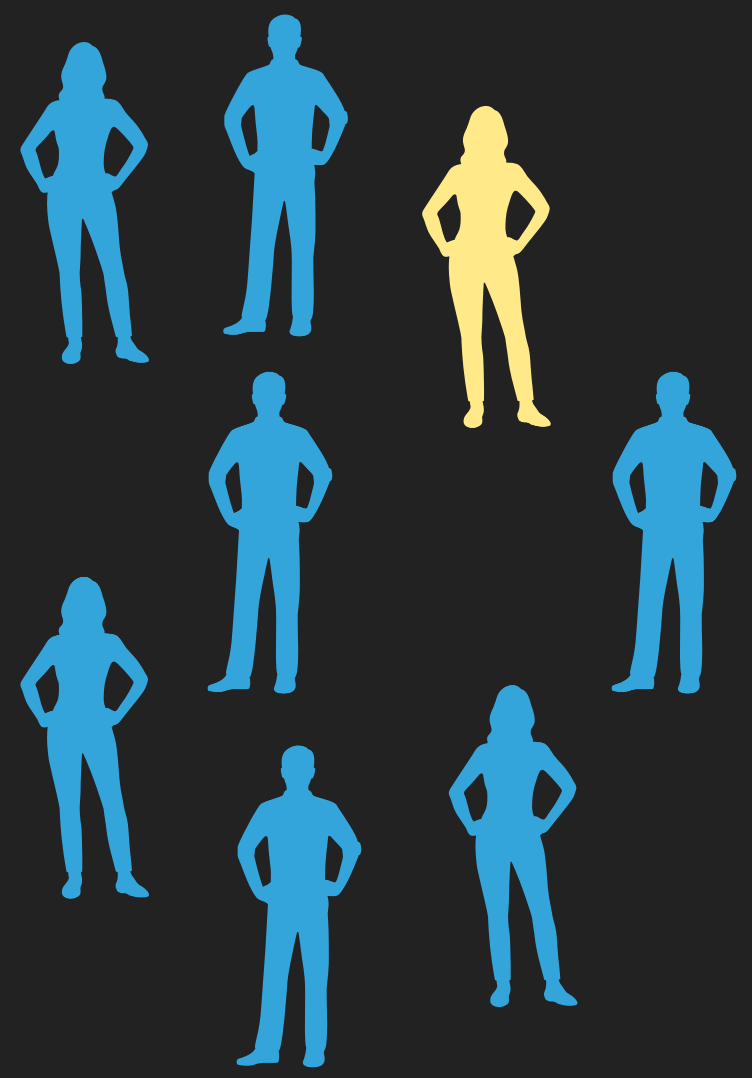
THE INDIVIDUAL AND THE AGGREGATE

- Starting with the data gives new insight into when something like probability exists
- But it does not answer the trickier question: what is the relationship between an individual and a probability?
- This matters, and is at the core of many people's anxiety with ML systems



THE INDIVIDUAL AND THE AGGREGATE

- Starting with the data gives new insight into when something like probability exists
- But it does not answer the trickier question: what is the relationship between an individual and a probability?
- This matters, and is at the core of many people's anxiety with ML systems



- Because statistics deals with aggregates, and ethics concerns the individual…

BEYOND EXPECTATIONS

Aggregating data in ways other than the average, and the connection to the earlier points

Risk Measures and Upper Probabilities: Coherence and Stratification

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

BEYOND EXPECTATIONS

Aggregating data in ways other than the average, and the connection to the earlier points

- How to aggregate?

Risk Measures and Upper Probabilities: Coherence and Stratification

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

BEYOND EXPECTATIONS

Aggregating data in ways other than the average, and the connection to the earlier points

- How to aggregate?
- Many more non-linear expectations than linear ones!

Risk Measures and Upper Probabilities: Coherence and Stratification

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

BEYOND EXPECTATIONS

Aggregating data in ways other than the average, and the connection to the earlier points

- How to aggregate?
- Many more non-linear expectations than linear ones!
- Turns out the sensible ones are combinations of expectations

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

Risk Measures and Upper Probabilities: Coherence and Stratification

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

BEYOND EXPECTATIONS

Aggregating data in ways other than the average, and the connection to the earlier points

- How to aggregate?
- Many more non-linear expectations than linear ones!
- Turns out the sensible ones are combinations of expectations
  - Very nice convex geometry

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

Risk Measures and Upper Probabilities: Coherence and Stratification

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

BEYOND EXPECTATIONS

Aggregating data in ways other than the average, and the connection to the earlier points

- How to aggregate?
- Many more non-linear expectations than linear ones!
- Turns out the sensible ones are combinations of expectations
  - Very nice convex geometry
  - Structure and stratification

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

Risk Measures and Upper Probabilities: Coherence and Stratification

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

BEYOND EXPECTATIONS

Aggregating data in ways other than the average, and the connection to the earlier points

- How to aggregate?
- Many more non-linear expectations than linear ones!
- Turns out the sensible ones are combinations of expectations
  - Very nice convex geometry
  - Structure and stratification
  - Useful for imposing fairness, robustness to perturbations, and controlling sensitivity to outliers

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

Risk Measures and Upper Probabilities: Coherence and Stratification

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

MANY PATHS TO ONE DESTINATION

MANY PATHS TO ONE DESTINATION

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)

Fairness Risk Measures

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)
   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)

   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)
4. Robust Bayes (imprecise prior)

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)

   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)
4. **Robust Bayes** (imprecise prior)
5. Rearrangement invariant norms

   Risk Measures and Upper Probabilities: Coherence and Stratification

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)

   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)
4. **Robust Bayes** (imprecise prior)
5. Rearrangement invariant norms

   Risk Measures and Upper Probabilities: Coherence and Stratification
6. de Finetti with a bid-ask spread

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)

   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)
4. **Robust Bayes** (imprecise prior)
5. Rearrangement invariant norms

   Risk Measures and Upper Probabilities: Coherence and Stratification
6. de Finetti with a bid-ask spread
7. Set systems for probability $(\Omega, \mathscr{S}, \mu)$

   Systems of Precision: Coherent Probabilities on Pre-Dynkin Systems and Coherent Previsions on Linear Subspaces

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)

   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)
4. **Robust Bayes** (imprecise prior)
5. Rearrangement invariant norms

   Risk Measures and Upper Probabilities: Coherence and Stratification
6. de Finetti with a bid-ask spread
7. Set systems for probability $(\Omega, \mathscr{S}, \mu)$

   Systems of Precision: Coherent Probabilities on Pre-Dynkin Systems and Coherent Previsions on Linear Subspaces
8. Generalised frequentism (von Mises)

   Strictly Frequentist Imprecise Probability

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)

   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)
4. **Robust Bayes** (imprecise prior)
5. Rearrangement invariant norms

   Risk Measures and Upper Probabilities: Coherence and Stratification
6. de Finetti with a bid-ask spread
7. Set systems for probability $(\Omega, \mathscr{S}, \mu)$

   Systems of Precision: Coherent Probabilities on Pre-Dynkin Systems and Coherent Previsions on Linear Subspaces
8. Generalised frequentism (von Mises)

   Strictly Frequentist Imprecise Probability

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

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

MANY PATHS TO ONE DESTINATION

1. Axiomatic approach to risk measures
2. Demanding fairness (f. risk measures)
   Fairness Risk Measures
3. Uncertainty + Ambiguity (economics)
4. **Robust Bayes** (imprecise prior)
5. Rearrangement invariant norms
   Risk Measures and Upper Probabilities: Coherence and Stratification
6. de Finetti with a bid-ask spread
7. Set systems for probability $(\Omega, \mathscr{S}, \mu)$
   Systems of Precision: Coherent Probabilities on Pre-Dynkin Systems and Coherent Previsions on Linear Subspaces
8. Generalised frequentism (von Mises)
   Strictly Frequentist Imprecise Probability

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

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

And they have already been used in ML (e.g. SVM via CVaR

The upshot: multiple compelling reasons to go "beyond expectations"

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

- How to reliably perform actuarial reasoning on data without assuming “iid”?

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

- How to reliably perform actuarial reasoning on data without assuming “iid”?
- How to model (and mitigate) various corruptions of data (including insidious ones)?

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

- How to reliably perform actuarial reasoning on data without assuming "iid"?
- How to model (and mitigate) various corruptions of data (including insidious ones)?
- How to think ethically about data (especially about people) which are not "drawn from a distribution" or are non-ergodic / non-equilibrium?

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

- How to reliably perform actuarial reasoning on data without assuming "iid"?
- How to model (and mitigate) various corruptions of data (including insidious ones)?
- How to think ethically about data (especially about people) which are not "drawn from a distribution" or are non-ergodic / non-equilibrium?
- What is information in a non-equilibrium situation?

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

- How to reliably perform actuarial reasoning on data without assuming "iid"?
- How to model (and mitigate) various corruptions of data (including insidious ones)?
- How to think ethically about data (especially about people) which are not "drawn from a distribution" or are non-ergodic / non-equilibrium?
- What is information in a non-equilibrium situation?
- How to reason about the effects of data (e.g. performativity) sans stochasticity?

SIX KEY QUESTIONS I WOULD LIKE TO ANSWER

- How to reliably perform actuarial reasoning on data without assuming "iid"?
- How to model (and mitigate) various corruptions of data (including insidious ones)?
- How to think ethically about data (especially about people) which are not "drawn from a distribution" or are non-ergodic / non-equilibrium?
- What is information in a non-equilibrium situation?
- How to reason about the effects of data (e.g. performativity) sans stochasticity?
- How to make better rhetorical practices when reasoning actuarially?


RHETORIC

THE RHETORIC OF MACHINE LEARNING

La Nouvelle Rhétorique: Traité de l'Argumentation
Presses Universitaires de France, 1958

THE RHETORIC OF MACHINE LEARNING

- Rhetoric: argumentation designed to persuade

La Nouvelle Rhétorique: Traité de l'Argumentation
Presses Universitaires de France, 1958

THE RHETORIC OF MACHINE LEARNING

- Rhetoric: argumentation designed to persuade
- The existing rhetoric of ML is that of “anti-rhetoric”

La Nouvelle Rhétorique: Traité de l'Argumentation
Presses Universitaires de France, 1958

THE RHETORIC OF MACHINE LEARNING

- Rhetoric: argumentation designed to persuade
- The existing rhetoric of ML is that of "anti-rhetoric"
- Data as fact; facts are what you call that which you wish not to discuss

La Nouvelle Rhétorique: Traité de l'Argumentation
Presses Universitaires de France, 1958

THE RHETORIC OF MACHINE LEARNING

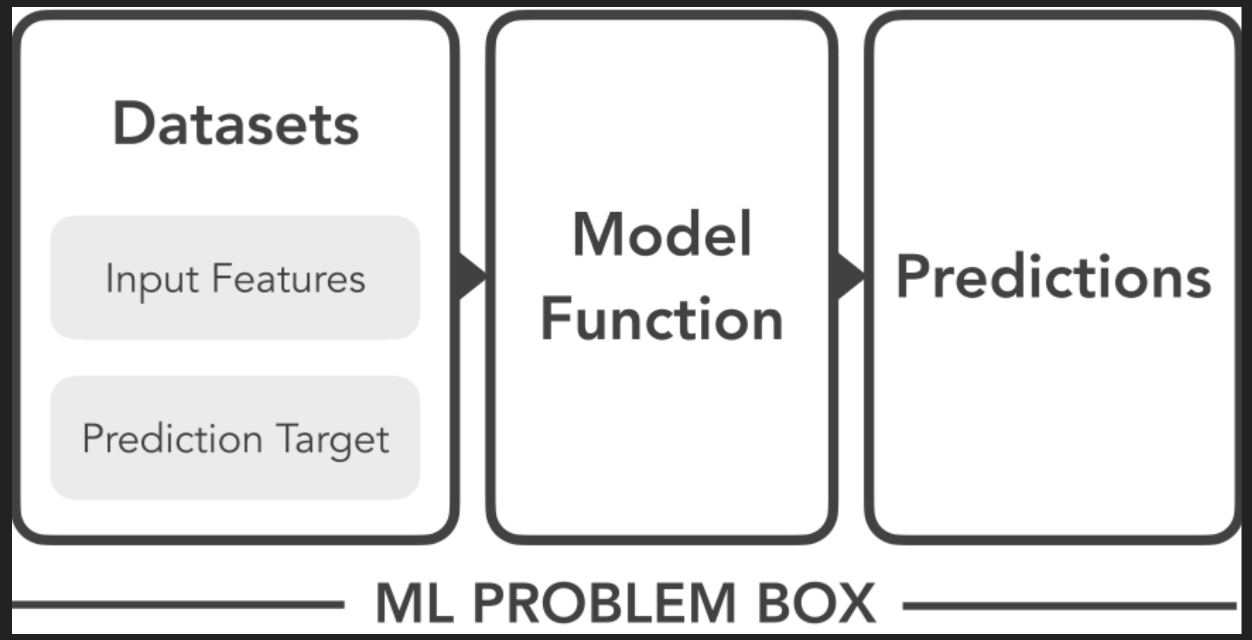
La Nouvelle Rhétorique: Traité de l'Argumentation
Presses Universitaires de France, 1958

- Rhetoric: argumentation designed to persuade
- The existing rhetoric of ML is that of "anti-rhetoric"
- Data as fact; facts are what you call that which you wish not to discuss
- But in the end you want to persuade through "chains of argument / reference"; think of scientific results, mathematical proofs and legal arguments…

THE RHETORIC OF MACHINE LEARNING

La Nouvelle Rhétorique: Traité de l'Argumentation
Presses Universitaires de France, 1958

- Rhetoric: argumentation designed to persuade
- The existing rhetoric of ML is that of "anti-rhetoric"
- Data as fact; facts are what you call that which you wish not to discuss
- But in the end you want to persuade through "chains of argument / reference"; think of scientific results, mathematical proofs and legal arguments…



THE RHETORIC OF MACHINE LEARNING

- Rhetoric: argumentation designed to persuade
- The existing rhetoric of ML is that of "anti-rhetoric"
- Data as fact; facts are what you call that which you wish not to discuss
- But in the end you want to persuade through "chains of argument / reference"; think of scientific results, mathematical proofs and legal arguments…

La Nouvelle Rhétorique: Traité de l'Argumentation
Presses Universitaires de France, 1958

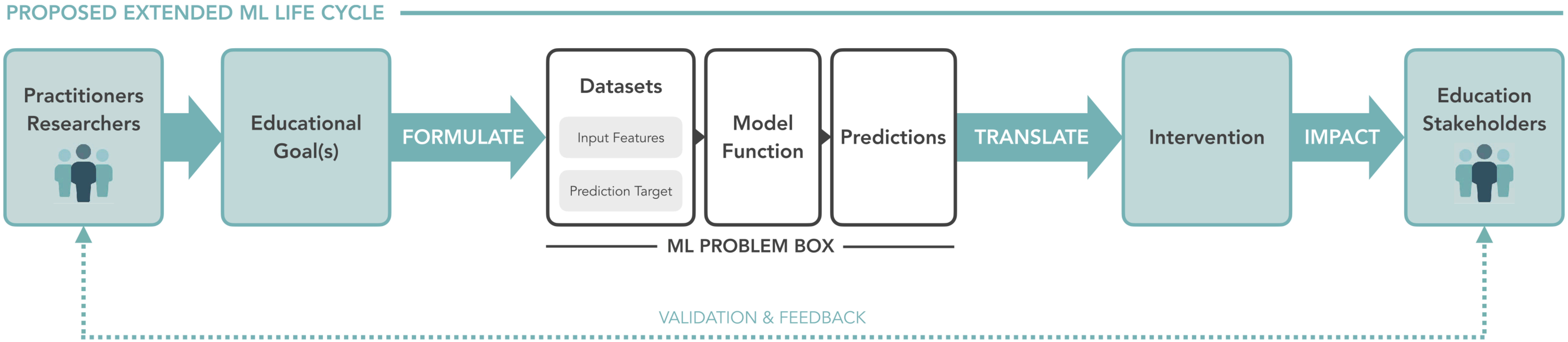


Fig. 1. An extended ML life cycle diagram. The inner "ML Problem Box" represents the typical aspects of the ML problem detailed in the surveyed ML research papers. Our interview findings reveal the need to consider an extended version of the ML life cycle in ML research, including the initial problem formulation stage by practitioners and researchers and the translation from predictions to interventions that eventually impact stakeholders.

Liu, L.T., Wang, S., Britton, T. and Abebe, R., 2023. Reimagining the machine learning life cycle to improve educational outcomes of students. Proceedings of the National Academy of Sciences, 120(9), p.e2204781120.


LOOKING AHEAD

SEVEN ASPECTS OF MY STYLE OF RESEARCH

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

- All knowledge is relational – so focus on the glue, not the wood

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

- All knowledge is relational – so focus on the glue, not the wood
- Foundations are an interface to the world – so pay attention to the world

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

- All knowledge is relational – so focus on the glue, not the wood
- Foundations are an interface to the world – so pay attention to the world
- Revolutions require creative destruction – so be explicit about what to tear down

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

- All knowledge is relational – so focus on the glue, not the wood
- Foundations are an interface to the world – so pay attention to the world
- Revolutions require creative destruction – so be explicit about what to tear down
- Much baggage is old and hidden – so follow problems to their roots

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

- All knowledge is relational – so focus on the glue, not the wood
- Foundations are an interface to the world – so pay attention to the world
- Revolutions require creative destruction – so be explicit about what to tear down
- Much baggage is old and hidden – so follow problems to their roots
- Seeking novelty leads to trivia – so seek to understand and take novelty as a gift

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

- All knowledge is relational – so focus on the glue, not the wood
- Foundations are an interface to the world – so pay attention to the world
- Revolutions require creative destruction – so be explicit about what to tear down
- Much baggage is old and hidden – so follow problems to their roots
- Seeking novelty leads to trivia – so seek to understand and take novelty as a gift
- One-way or approximate results are ephemeral – so seek exact characterisations

SEVEN ASPECTS OF MY STYLE OF RESEARCH

It ain't what you do but the way that you do it – that's what gets results!

- All knowledge is relational – so focus on the glue, not the wood
- Foundations are an interface to the world – so pay attention to the world
- Revolutions require creative destruction – so be explicit about what to tear down
- Much baggage is old and hidden – so follow problems to their roots
- Seeking novelty leads to trivia – so seek to understand and take novelty as a gift
- One-way or approximate results are ephemeral – so seek exact characterisations
- A professor's largest legacy is in people – so focus upon helping them grow

"HELPING THEM GROW"

"HELPING THEM GROW"

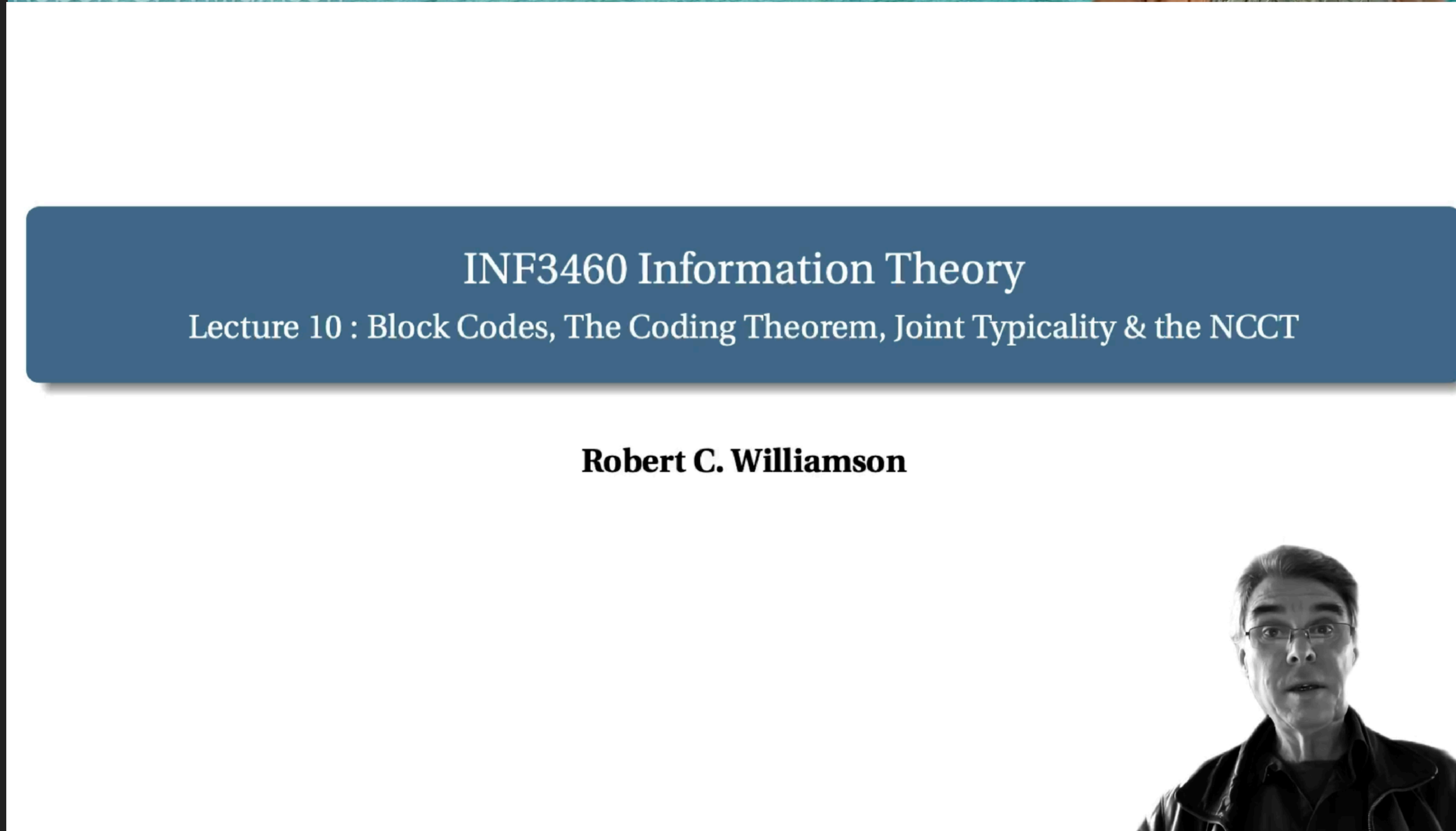
- a.k.a. "teaching"

"HELPING THEM GROW"

- a.k.a. "teaching"
- Formal ("courses")

"HELPING THEM GROW"

- a.k.a. "teaching"
- Formal ("courses")



"HELPING THEM GROW"

- a.k.a. "teaching"
- Formal ("courses")
- Informal ("tapas")

KYLIE CATCHPOLE AND ROBERT WILLIAMSON

BEING A SCIENTIST

INF3460 Information Theory

Lecture 10 : Block Codes, The Coding Theorem, Joint Typicality & the NCCT

Robert C. Williamson

BEING A SCIENTIST

Marie Skłodowska (age 16)


Marie Skłodowska (age 16)

EIGHT THINGS I DISAGREE WITH

EIGHT THINGS I DISAGREE WITH

- Data is *given,* and it represents the facts of the world, and is incontrovertible

EIGHT THINGS I DISAGREE WITH

- Data is *given,* and it represents the facts of the world, and is incontrovertible

- ML algorithms are *black boxes,* and they thus need opening & explaining

EIGHT THINGS I DISAGREE WITH

- Data is *given*, and it represents the facts of the world, and is incontrovertible
- ML algorithms are *black boxes*, and they thus need opening & explaining
- AI systems “make decisions” and are autonomous (and that’s ethically bad)

EIGHT THINGS I DISAGREE WITH

- Data is *given*, and it represents the facts of the world, and is incontrovertible
- ML algorithms are *black boxes*, and they thus need opening & explaining
- AI systems "make decisions" and are autonomous (and that's ethically bad)
- We (thus) need to regulate the *technology* of Machine Learning

EIGHT THINGS I DISAGREE WITH

- Data is *given*, and it represents the facts of the world, and is incontrovertible
- ML algorithms are *black boxes*, and they thus need opening & explaining
- AI systems "make decisions" and are autonomous (and that's ethically bad)
- We (thus) need to regulate the *technology* of Machine Learning
- The *more data the better*, and with enough data we don't need to think

EIGHT THINGS I DISAGREE WITH

- Data is *given*, and it represents the facts of the world, and is incontrovertible
- ML algorithms are *black boxes*, and they thus need opening & explaining
- AI systems "make decisions" and are autonomous (and that's ethically bad)
- We (thus) need to regulate the *technology* of Machine Learning
- The *more data the better*, and with enough data we don't need to think
- There is *"the probability"* for every event, and thus *"the probability distribution"*

EIGHT THINGS I DISAGREE WITH

- Data is *given*, and it represents the facts of the world, and is incontrovertible
- ML algorithms are *black boxes*, and they thus need opening & explaining
- AI systems "make decisions" and are autonomous (and that's ethically bad)
- We (thus) need to regulate the *technology* of Machine Learning
- The *more data the better*, and with enough data we don't need to think
- There is *"the probability"* for every event, and thus *"the probability distribution"*
- There is *one notion of information*, and it only concerns knowing

EIGHT THINGS I DISAGREE WITH

- Data is *given*, and it represents the facts of the world, and is incontrovertible
- ML algorithms are *black boxes*, and they thus need opening & explaining
- AI systems "make decisions" and are autonomous (and that's ethically bad)
- We (thus) need to regulate the *technology* of Machine Learning
- The *more data the better*, and with enough data we don't need to think
- There is *"the probability"* for every event, and thus *"the probability distribution"*
- There is *one notion of information*, and it only concerns knowing
- ML is *not rhetorical*; it is *objective* (it is "data driven" … and data is *fact*)


fm.ls

SPARE SLIDES

DATA

DATA AS FACT

DATA AS FACT

There is nothing more deceptive than an obvious fact.

– Arthur Conan Doyle, *The Boscombe Valley Mystery*

DATA AS FACT

There is nothing more deceptive than an obvious fact.

– Arthur Conan Doyle, *The Boscombe Valley Mystery*

- “Data as fact”

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:

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]

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"

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

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?

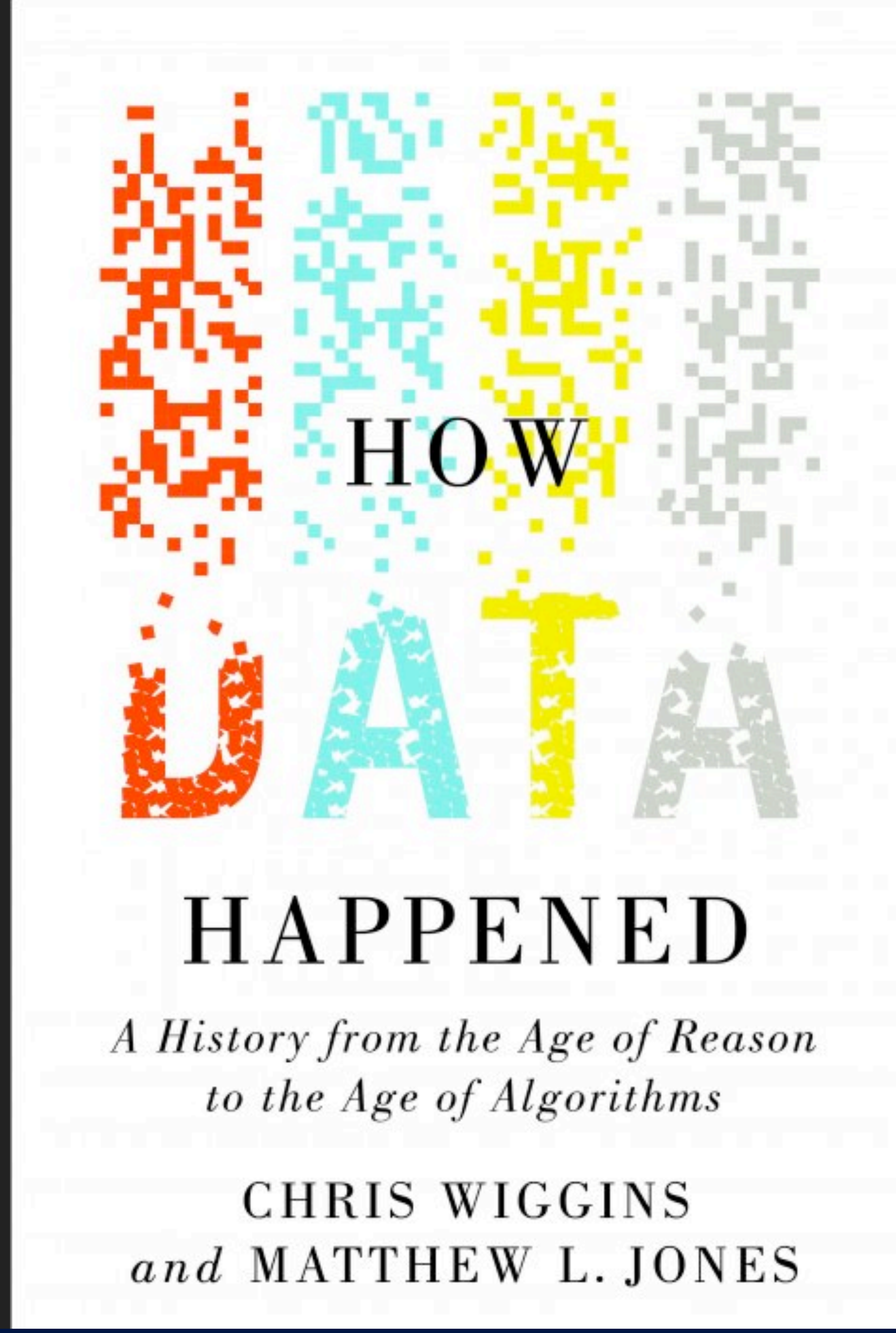
FROM INGESTING TO STUDYING DATA

FROM INGESTING TO STUDYING DATA

- Particular focus: failure of usual *models of data*

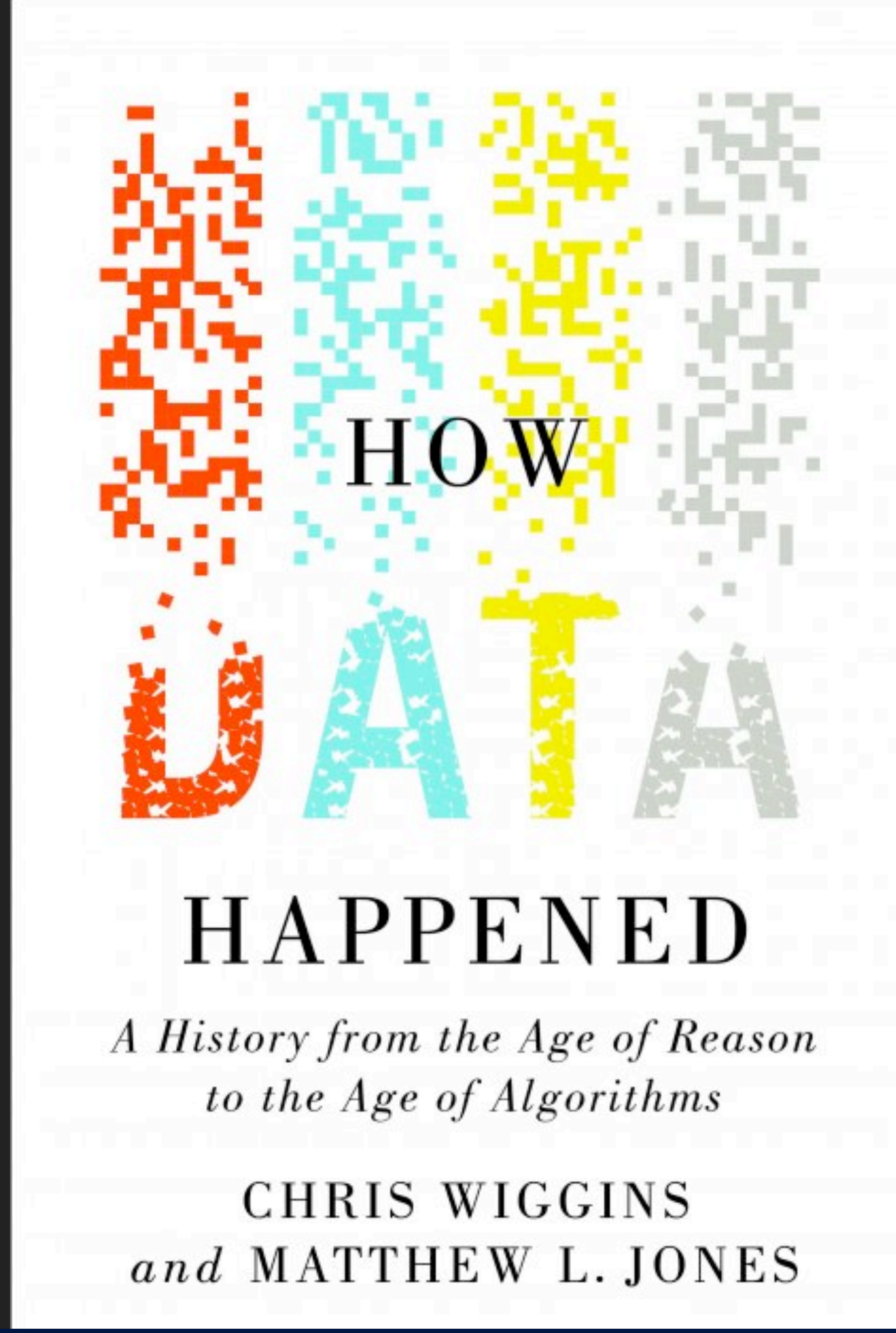
FROM INGESTING TO STUDYING DATA

- Particular focus: failure of usual *models of data*
- Need to pay attention to the data itself …



FROM INGESTING TO STUDYING DATA

- Particular focus: failure of usual *models of data*
- Need to pay attention to the data itself …
- ML perspective: Data is “drawn iid from some distribution”



JOHN VENN

- Grouping into "natural kinds" underpins statistical regularity (cf. the "reference class problem"!)

"Such regularity as we trace in nature is owing, much more than is often suspected, to the arrangement of things in natural kinds, each of them containing a large number of individuals.

…

A large number of objects in the class, together with that general similarity which entitles the objects to be fairly comprised in one class, seem to be important conditions for the applicability of the theory of Probability to any phenomenon."

John Venn, *The logic of chance* (3rd edition), MacMillan and Co., 1888, pages 55-56

THE (NEW) CATEGORICAL IMPERATIVE

THE (NEW) CATEGORICAL IMPERATIVE

- No matter how “big” your data, the classificatory problem remains

THE (NEW) CATEGORICAL IMPERATIVE

- No matter how "big" your data, the classificatory problem remains
- You are not "representing the world".

THE (NEW) CATEGORICAL IMPERATIVE

- No matter how "big" your data, the classificatory problem remains
- You are not "representing the world".
- At best you are representing how you represent the world…

THE (NEW) CATEGORICAL IMPERATIVE

- No matter how “big” your data, the classificatory problem remains
- You are not “representing the world”.
- At best you are representing how you represent the world…

THE (NEW) CATEGORICAL IMPERATIVE

- No matter how "big" your data, the classificatory problem remains
- You are not "representing the world".
- At best you are representing how you represent the world…

- A more significant problem:

THE (NEW) CATEGORICAL IMPERATIVE

- No matter how "big" your data, the classificatory problem remains
- You are not "representing the world".
- At best you are representing how you represent the world…

- A more significant problem:
  - You build a complex statistical model; it "works well"

THE (NEW) CATEGORICAL IMPERATIVE

- No matter how “big” your data, the classificatory problem remains
- You are not “representing the world”.
- At best you are representing how you represent the world…

- A more significant problem:
  - You build a complex statistical model; it “works well”
  - What does this say about an individual?

DATA AS "RANDOM VARIABLES"

DATA AS "RANDOM VARIABLES"

- The canonical model of "data"

DATA AS "RANDOM VARIABLES"

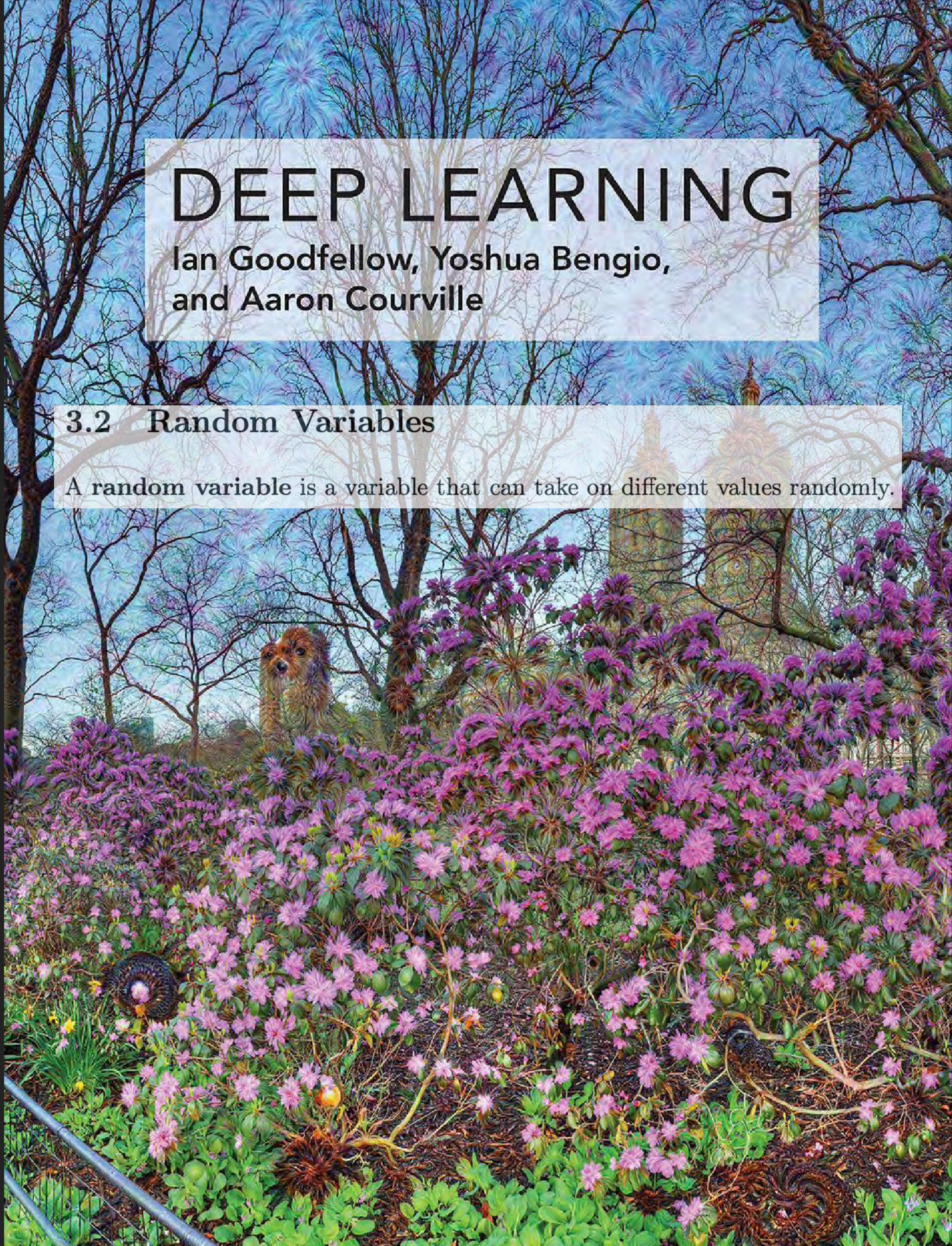
- The canonical model of "data"
- Two small difficulties from a mathematical perspective:
  - They are not "random" ; they do not "vary"
    - Because they are simply (measurable) <u>functions</u>!

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

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…



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) <u>functions</u>!
- <u>Interpreting</u> "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…

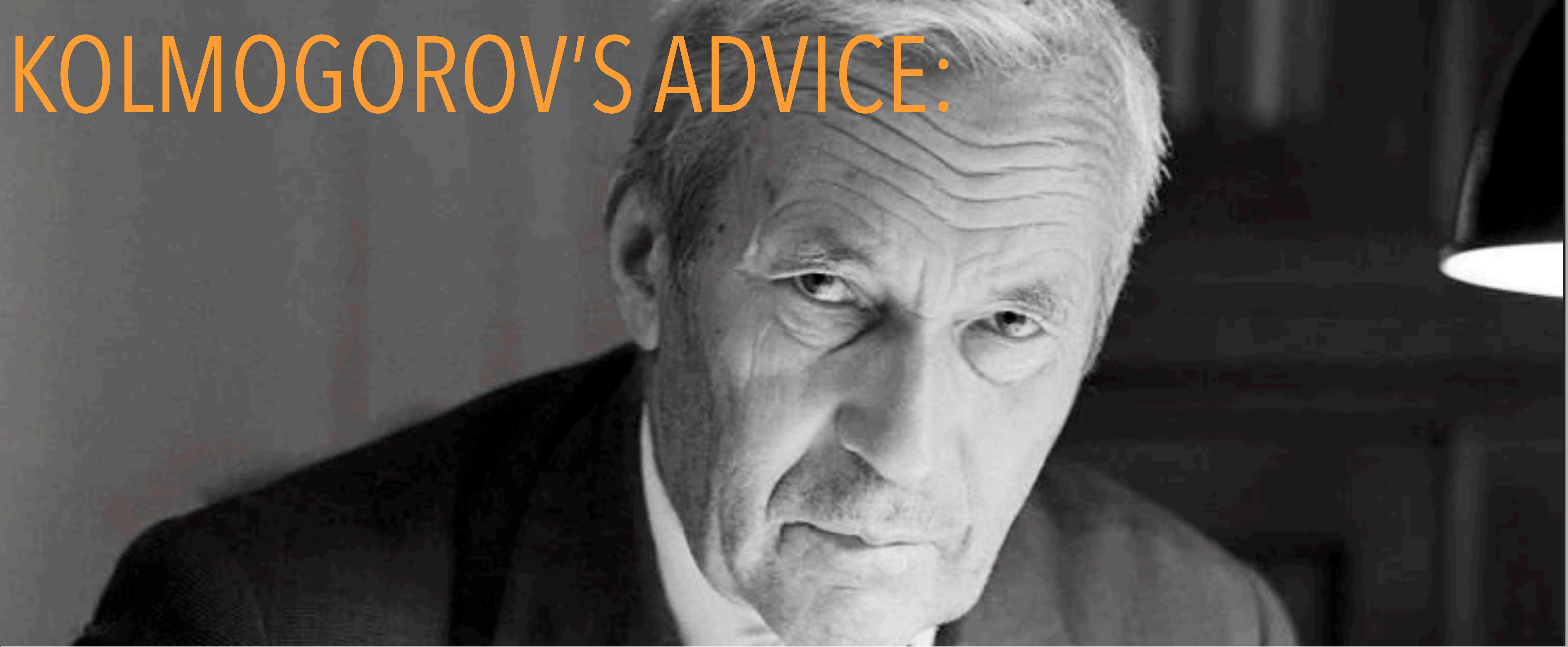
DATA AS "RANDOM VARIABLES"

- 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?*

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[1]

[1] 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.

BEYOND DATA

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*
- "Convenience samples" – data you found lying around somewhere

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*
- "Convenience samples" – data you found lying around somewhere
- How good a "representation" of the world is it?

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*
- "Convenience samples" – data you found lying around somewhere
- How good a "representation" of the world is it?
- *You can not answer this by just looking at your data!*

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*
- "Convenience samples" – data you found lying around somewhere
- How good a "representation" of the world is it?
- *You can not answer this by just looking at your data!*
- What to do?

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*
- "Convenience samples" – data you found lying around somewhere
- How good a "representation" of the world is it?
- *You can not answer this by just looking at your data!*
- What to do?
  - First: think of Data (given), Capta (taken), Constructa (made)

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*
- "Convenience samples" – data you found lying around somewhere
- How good a "representation" of the world is it?
- *You can not answer this by just looking at your data!*
- What to do?
  - First: think of Data (given), Capta (taken), Constructa (made)
  - Another possibility: study the multitude of ways data can be corrupted

BEYOND DATA

Considering what happens when we do not take data for granted

- "Data" means that which is *given*
- "Convenience samples" – data you found lying around somewhere
- How good a "representation" of the world is it?
- *You can not answer this by just looking at your data!*
- What to do?
  - First: think of Data (given), Capta (taken), Constructa (made)
  - Another possibility: study the multitude of ways data can be corrupted
  - Seek to understand the effects; not to just "fix" the problem

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

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

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

- There is a nice story relating loss functions to information

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

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

- There is a nice story relating loss functions to information

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) := \min_{h\in\mathscr{H}} \mathbb{E}\,\ell(\mathsf{Y}, h(\mathsf{X}))$$

BEYOND INFORMATION

- There is a nice story relating loss functions to information

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) := \min_{h\in\mathscr{H}} \mathbb{E}\, \ell(\mathsf{Y}, h(\mathsf{X}))$$

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) \text{"="} -\mathrm{I}_{\mathscr{F}}(P_{\mathsf{XY}})$$

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

- There is a nice story relating loss functions to information

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) := \min_{h \in \mathscr{H}} \mathbb{E}\, \ell(\mathsf{Y}, h(\mathsf{X}))$$

Information Processing Equalities and the Information-Risk Bridge

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) \text{"="} -\mathrm{I}_{\mathscr{F}}(P_{\mathsf{XY}})$$

$$(\ell, \mathscr{H}) \leftrightarrow \mathscr{F}$$

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

- There is a nice story relating loss functions to information

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) := \min_{h\in\mathscr{H}} \mathbb{E}\, \ell(\mathsf{Y}, h(\mathsf{X}))$$

Information Processing Equalities and the Information-Risk Bridge

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}})\text{"="}-\mathrm{I}_{\mathscr{F}}(P_{\mathsf{XY}})$$

$$(\ell,\mathscr{H}) \leftrightarrow \mathscr{F}$$

In words: the minimal risk of a learning problem (on given data) is (up to a sign change) equivalent to the "amount of information" in the data

(But there is no single notion of information!)

Thus knowing (information) and acting (prediction risk) are inextricably intertwined

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

- There is a nice story relating loss functions to information
  - Based on classical expectations $\mathbb{E}$

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) := \min_{h\in\mathscr{H}} \mathbb{E}\, \ell(\mathsf{Y}, h(\mathsf{X}))$$

Information Processing Equalities and the Information-Risk Bridge

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}})\text{"="}-\mathrm{I}_{\mathscr{F}}(P_{\mathsf{XY}})$$

$$(\ell,\mathscr{H}) \leftrightarrow \mathscr{F}$$

In words: the minimal risk of a learning problem (on given data) is (up to a sign change) equivalent to the "amount of information" in the data

(But there is no single notion of information!)

Thus knowing (information) and acting (prediction risk) are inextricably intertwined

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

- There is a nice story relating loss functions to information
  - Based on classical expectations $\mathbb{E}$
- What do you get when using *generalised* expectations?

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) := \min_{h\in\mathscr{H}} \mathbb{E}\, \ell(\mathsf{Y}, h(\mathsf{X}))$$

Information Processing Equalities and the Information-Risk Bridge

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}})\text{"="}-\mathrm{I}_{\mathscr{F}}(P_{\mathsf{XY}})$$

$$(\ell,\mathscr{H}) \leftrightarrow \mathscr{F}$$

In words: the minimal risk of a learning problem (on given data) is (up to a sign change) equivalent to the "amount of information" in the data

(But there is no single notion of information!)

Thus knowing (information) and acting (prediction risk) are inextricably intertwined

BEYOND INFORMATION

Supposing there is no single notion of information, that it is not just knowing but also doing, and asking what does information even mean in non-equilibrium situations

- There is a nice story relating loss functions to information
  - Based on classical expectations $\mathbb{E}$
- What do you get when using *generalised* expectations?
- Can this give analogous insights in situations where distributions are not stable (non-equilibrium)?

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) := \min_{h \in \mathscr{H}} \mathbb{E}\, \ell(\mathsf{Y}, h(\mathsf{X}))$$

Information Processing Equalities and the Information-Risk Bridge

$$\mathrm{BR}_{\ell,\mathscr{H}}(P_{\mathsf{XY}}) \text{"="} {-\mathrm{I}_{\mathscr{F}}(P_{\mathsf{XY}})}$$

$$(\ell, \mathscr{H}) \leftrightarrow \mathscr{F}$$

In words: the minimal risk of a learning problem (on given data) is (up to a sign change) equivalent to the "amount of information" in the data

(But there is no single notion of information!)

Thus knowing (information) and acting (prediction risk) are inextricably intertwined

BEYOND INDEPENDENCE

Supposing that just because A and B have a probability does not imply that $A \cap B$ does

BEYOND INDEPENDENCE

Supposing that just because A and B have a probability does not imply that $A \cap B$ does

- The "*casual* assumption of independence"

Miracles and Statistics: The Casual Assumption of Independence

WILLIAM KRUSKAL*

Journal of the American Statistical Association
December 1988, Vol. 83, No. 404, Presidential Address

The primary theme of this address is cautionary: Statistical independence is far too often assumed casually, without serious concern for how common is dependence and how difficult it can be to achieve independence (or related structures). After

BEYOND INDEPENDENCE

- The “*casual* assumption of independence”
- *Not* “the assumption of *causal* independence,”

Miracles and Statistics: The Casual Assumption of Independence

WILLIAM KRUSKAL*

Journal of the American Statistical Association
December 1988, Vol. 83, No. 404, Presidential Address

The primary theme of this address is cautionary: Statistical independence is far too often assumed casually, without serious concern for how common is dependence and how difficult it can be to achieve independence (or related structures). After

BEYOND INDEPENDENCE

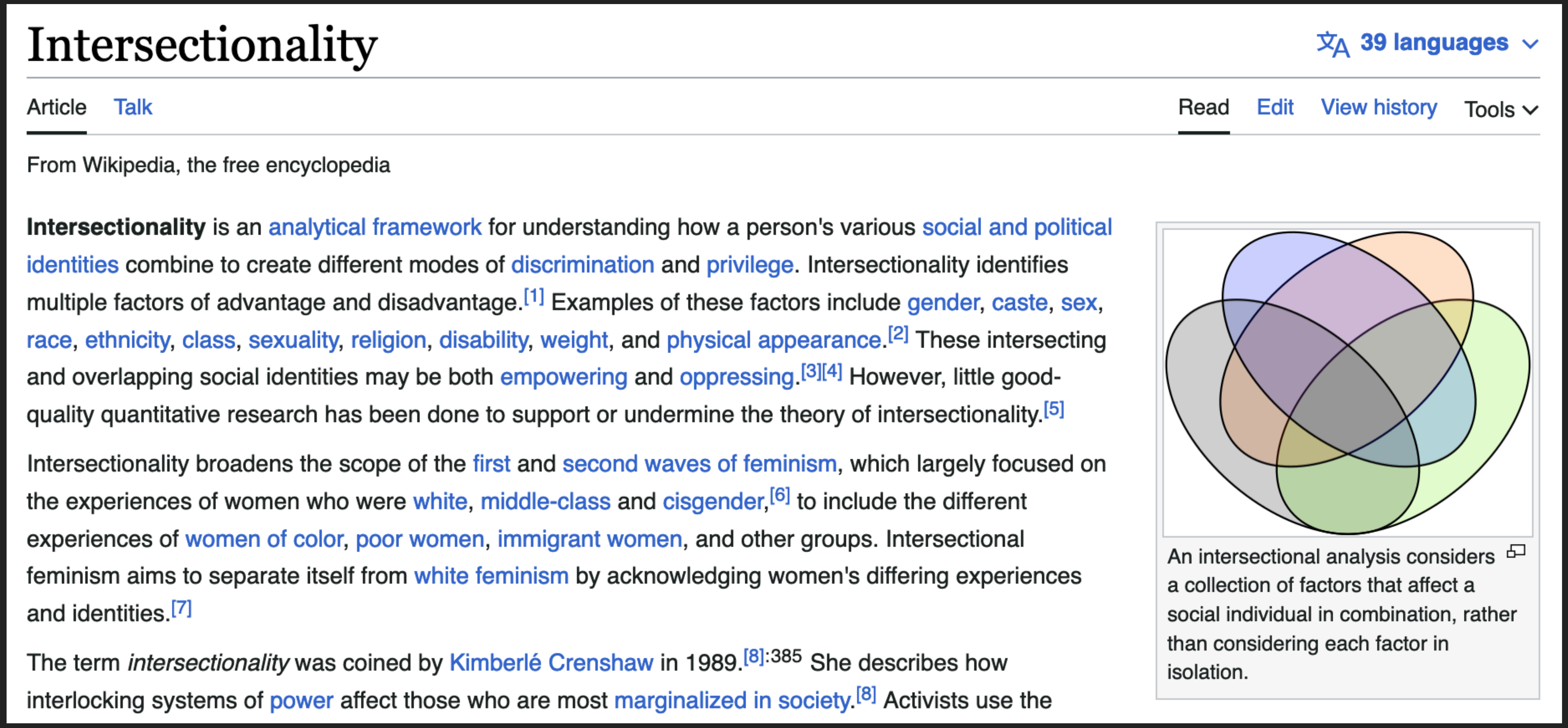
Supposing that just because A and B have a probability does not imply that $A \cap B$ does

- The "*casual* assumption of independence"
- *Not* "the assumption of *causal* independence,"
  - which is often also taken for granted and used as a justification for this…

BEYOND INDEPENDENCE

Supposing that just because A and B have a probability does not imply that $A \cap B$ does

- The "*casual* assumption of independence"
- *Not* "the assumption of *causal* independence,"
  - which is often also taken for granted and used as a justification for this…
- What if not all events have a probability?
  - "Intersectionality"
- Recall $A \perp B \Leftrightarrow P(A \cap B) = P(A) \times P(B)$



Intersectionality

39 languages

Article Talk

Read Edit View history Tools

From Wikipedia, the free encyclopedia

Intersectionality is an analytical framework for understanding how a person's various social and political identities combine to create different modes of discrimination and privilege. Intersectionality identifies multiple factors of advantage and disadvantage.[1] Examples of these factors include gender, caste, sex, race, ethnicity, class, sexuality, religion, disability, weight, and physical appearance.[2] These intersecting and overlapping social identities may be both empowering and oppressing.[3][4] However, little good-quality quantitative research has been done to support or undermine the theory of intersectionality.[5]

Intersectionality broadens the scope of the first and second waves of feminism, which largely focused on the experiences of women who were white, middle-class and cisgender,[6] to include the different experiences of women of color, poor women, immigrant women, and other groups. Intersectional feminism aims to separate itself from white feminism by acknowledging women's differing experiences and identities.[7]

The term *intersectionality* was coined by Kimberlé Crenshaw in 1989.[8]:385 She describes how interlocking systems of power affect those who are most marginalized in society.[8] Activists use the

An intersectional analysis considers a collection of factors that affect a social individual in combination, rather than considering each factor in isolation.

Supposing that just because A and B have a probability does not imply that $A \cap B$ does

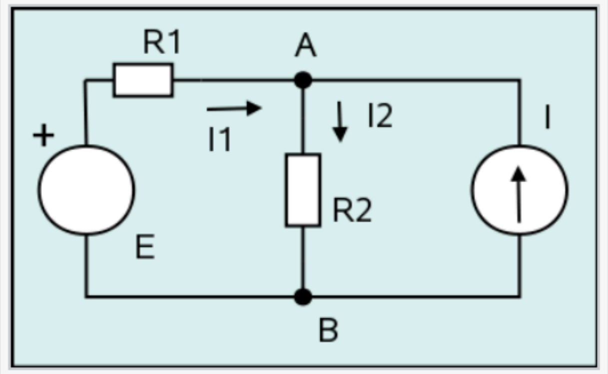
BEYOND INDEPENDENCE

- The "*casual* assumption of independence"
- *Not* "the assumption of *causal* independence,"
    - which is often also taken for granted and used as a justification for this…
- What if not all events have a probability?
    - "Intersectionality"
- Recall $A \perp B \Leftrightarrow P(A \cap B) = P(A) \times P(B)$

Hypergraph drawing [edit]

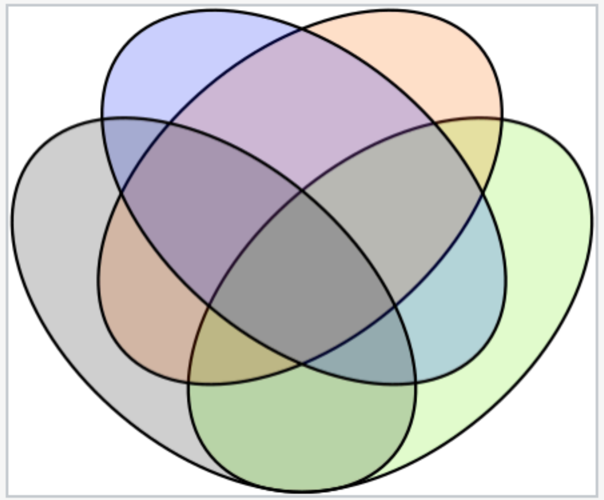
Although hypergraphs are more difficult to draw on paper than graphs, several researchers have studied methods for the visualization of hypergraphs.

In one possible visual representation for hypergraphs, similar to the standard graph drawing style in which curves in the plane are used to depict graph edges, a hypergraph's vertices are depicted as points, disks, or boxes, and its hyperedges are depicted as trees that have the vertices as their leaves.[19][20] If the vertices are represented as points, the hyperedges may also be shown as smooth curves that connect sets of points, or as simple closed curves that enclose sets of points.[21][22][23]



This circuit diagram can be interpreted as a drawing of a hypergraph in which four vertices (depicted as white rectangles and disks) are connected by three hyperedges drawn as trees.

In another style of hypergraph visualization, the subdivision model of hypergraph drawing,[24] the plane is subdivided into regions, each of which represents a single vertex of the hypergraph. The hyperedges of the hypergraph are represented by contiguous subsets of these regions, which may be indicated by coloring, by drawing outlines around them, or both. An order-n Venn diagram, for instance, may be viewed as a subdivision drawing of a hypergraph with n hyperedges (the curves defining the diagram) and $2^n - 1$ vertices (represented by the regions into which these curves subdivide the plane). In contrast with the polynomial-time recognition of planar graphs, it is NP-complete to determine whether a hypergraph has a planar subdivision drawing,[25] but the existence of a drawing of this type may be tested efficiently when the adjacency pattern of the regions is constrained to be a path, cycle, or tree.[26]



An order-4 Venn diagram, which can be interpreted as a subdivision drawing of a hypergraph with 15 vertices (the 15 colored regions) and 4 hyperedges (the 4 ellipses).

An alternative representation of the hypergraph called PAOH[1] is shown in the figure on top of this article. Edges are vertical lines connecting vertices. Vertices are aligned on the left. The legend on the right shows the names of the edges. It has been designed for dynamic hypergraphs but can be used for simple hypergraphs as well.

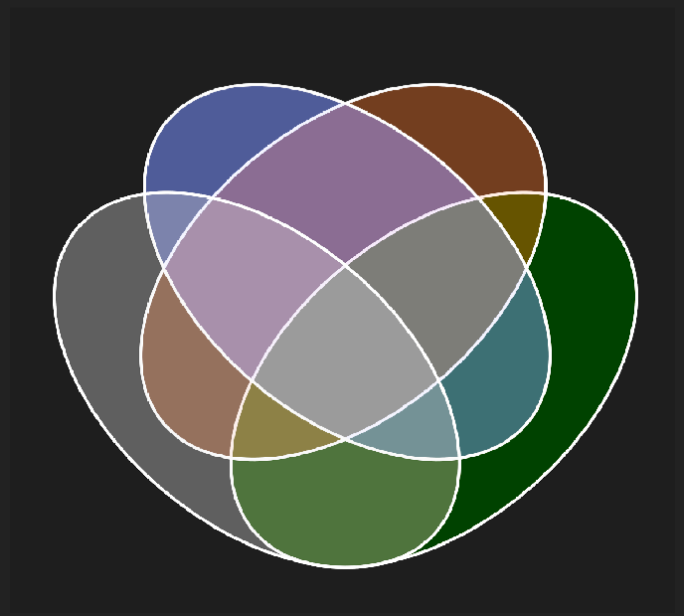
Hypergraph coloring [edit]

Classic hypergraph coloring is assigning one of the colors from set $\{1, 2, 3, \ldots, \lambda\}$ to every vertex of a hypergraph in such a way that each hyperedge contains at least two vertices of distinct colors. In other words, there must be no monochromatic hyperedge with cardinality at least 2. In this sense it is a direct generalization of graph coloring. Minimum number of used distinct colors over all colorings is called the chromatic number of a hypergraph.

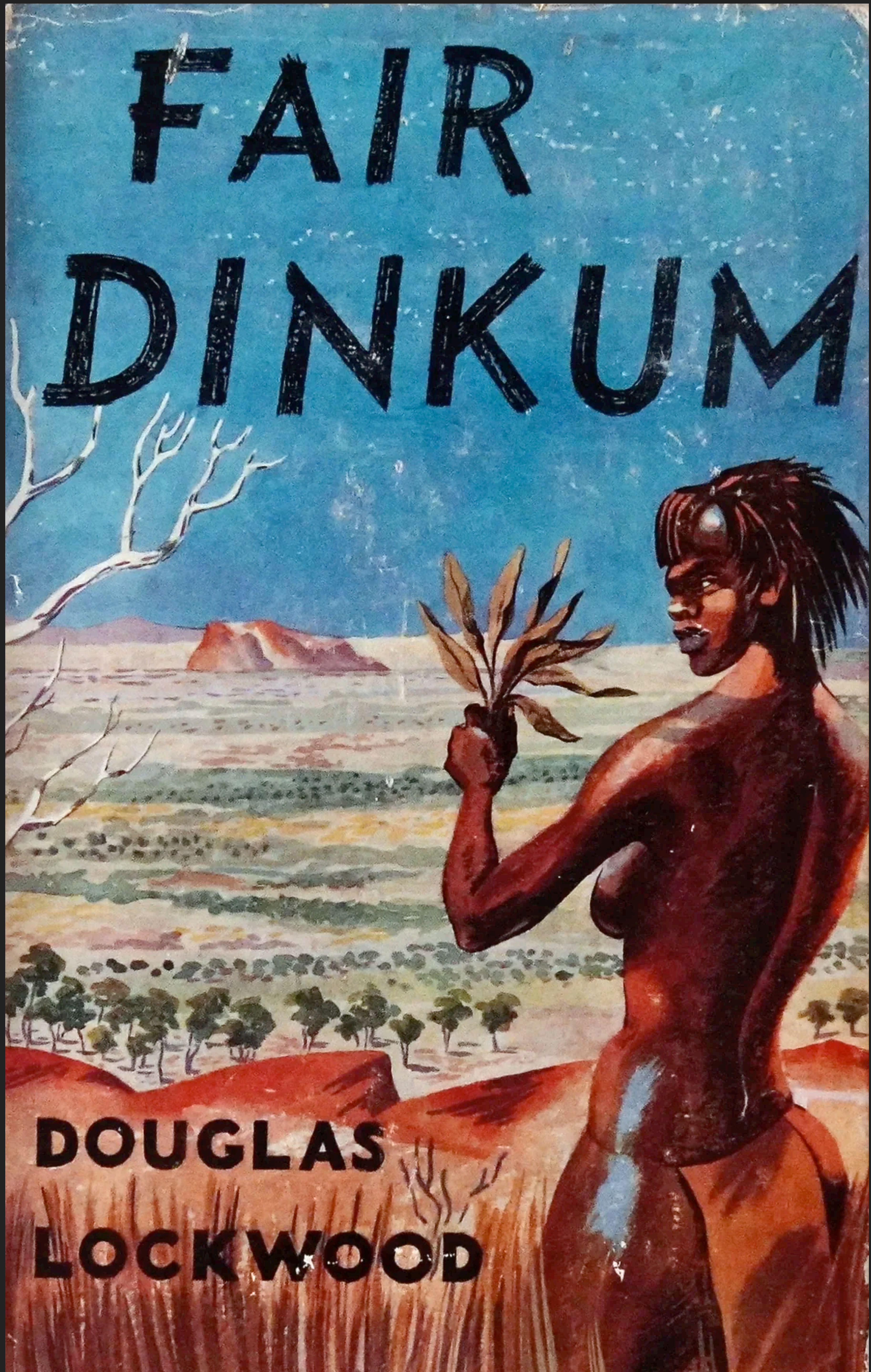
BEYOND INDEPENDENCE

Supposing that just because A and B have a probability does not imply that $A \cap B$ does

- The "*casual* assumption of independence"
- *Not* "the assumption of *causal* independence,"
  - which is often also taken for granted and used as a justification for this…
- What if not all events have a probability?
  - "Intersectionality"
- Recall $A \perp B \Leftrightarrow P(A \cap B) = P(A) \times P(B)$

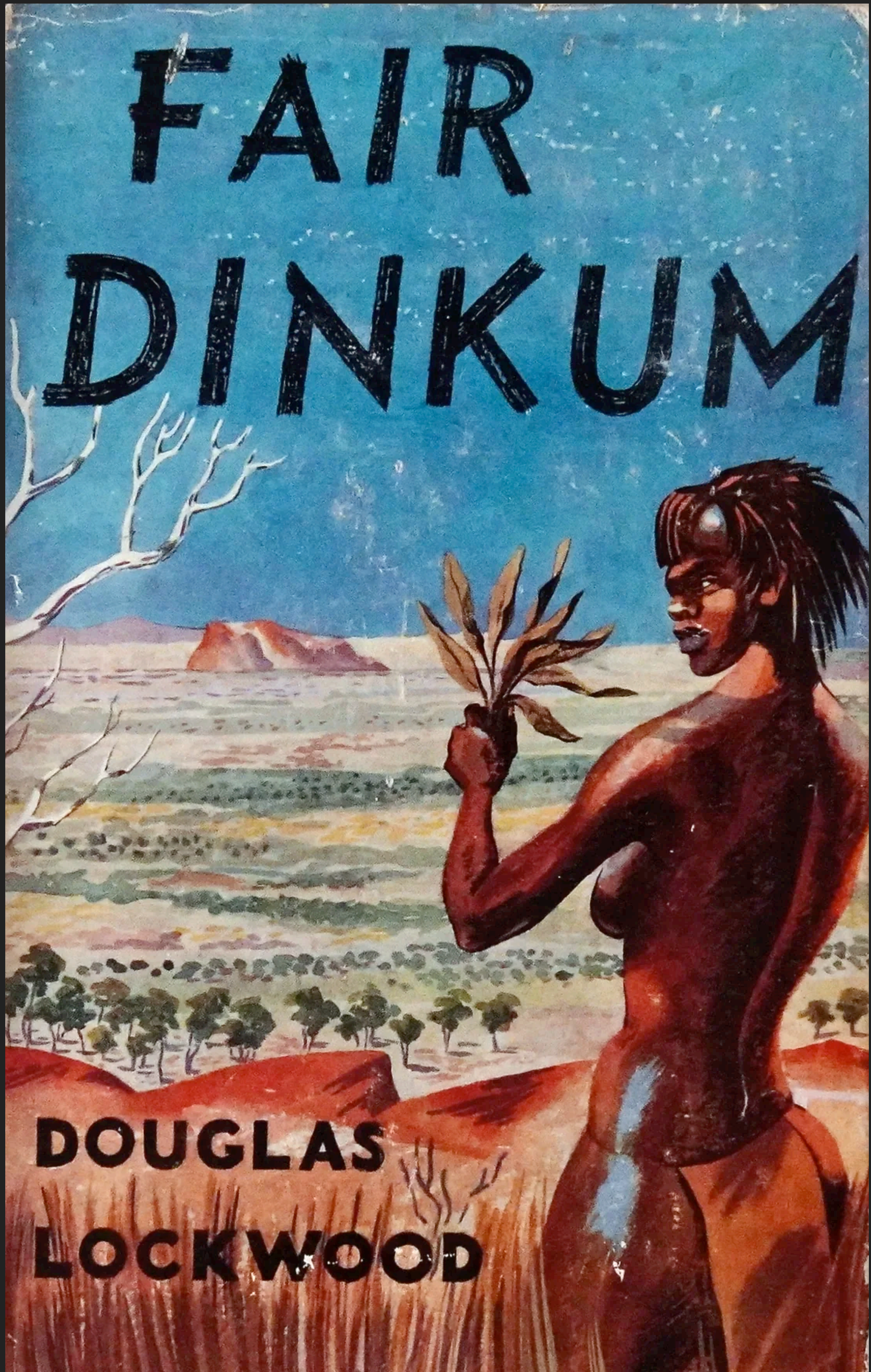


FAIR DYNKIN!



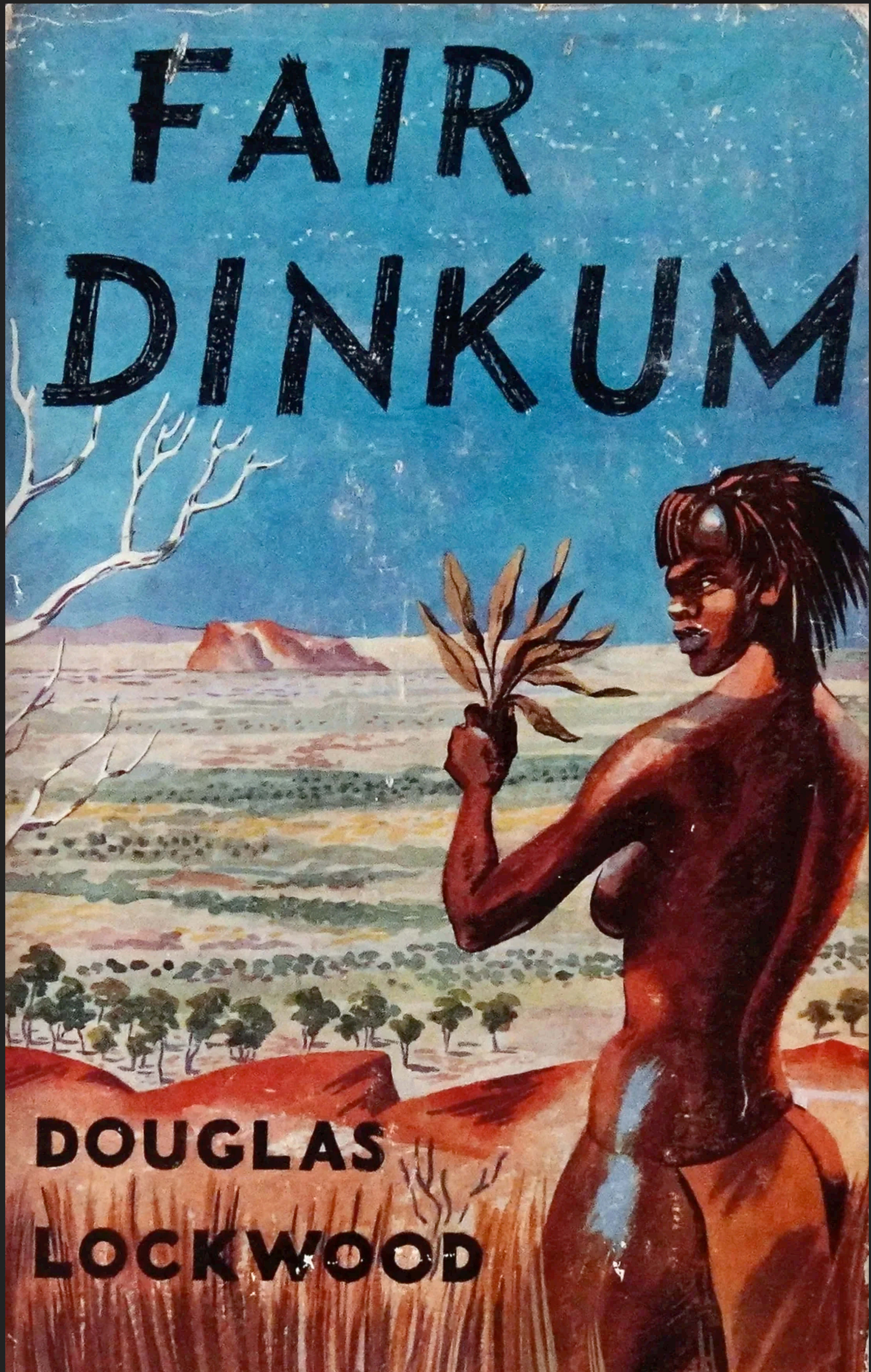
FAIR DYNKIN!

- Fairness as an actuarial problem



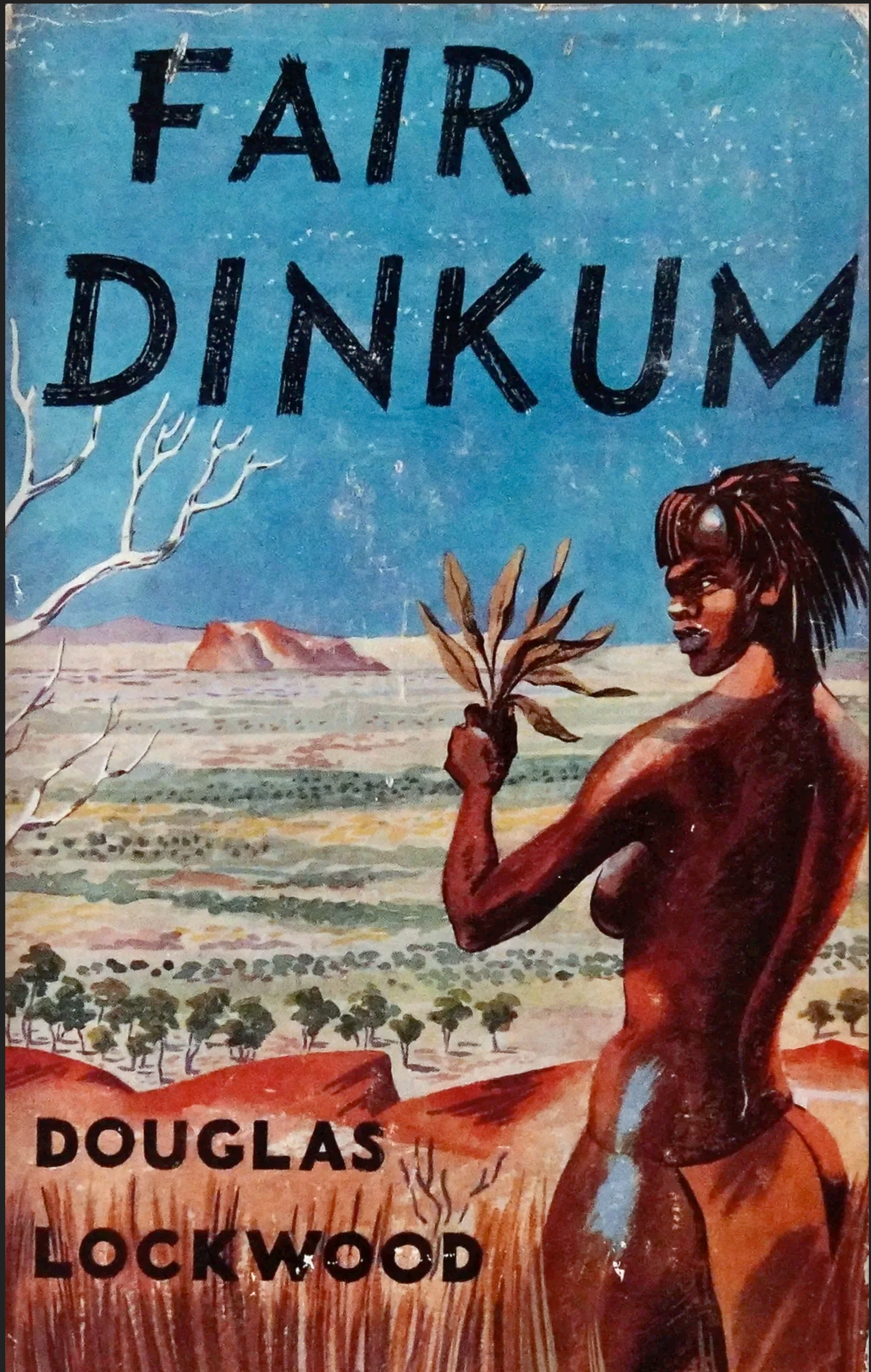
FAIR DYNKIN!

- Fairness as an actuarial problem
- Fairness = Independence



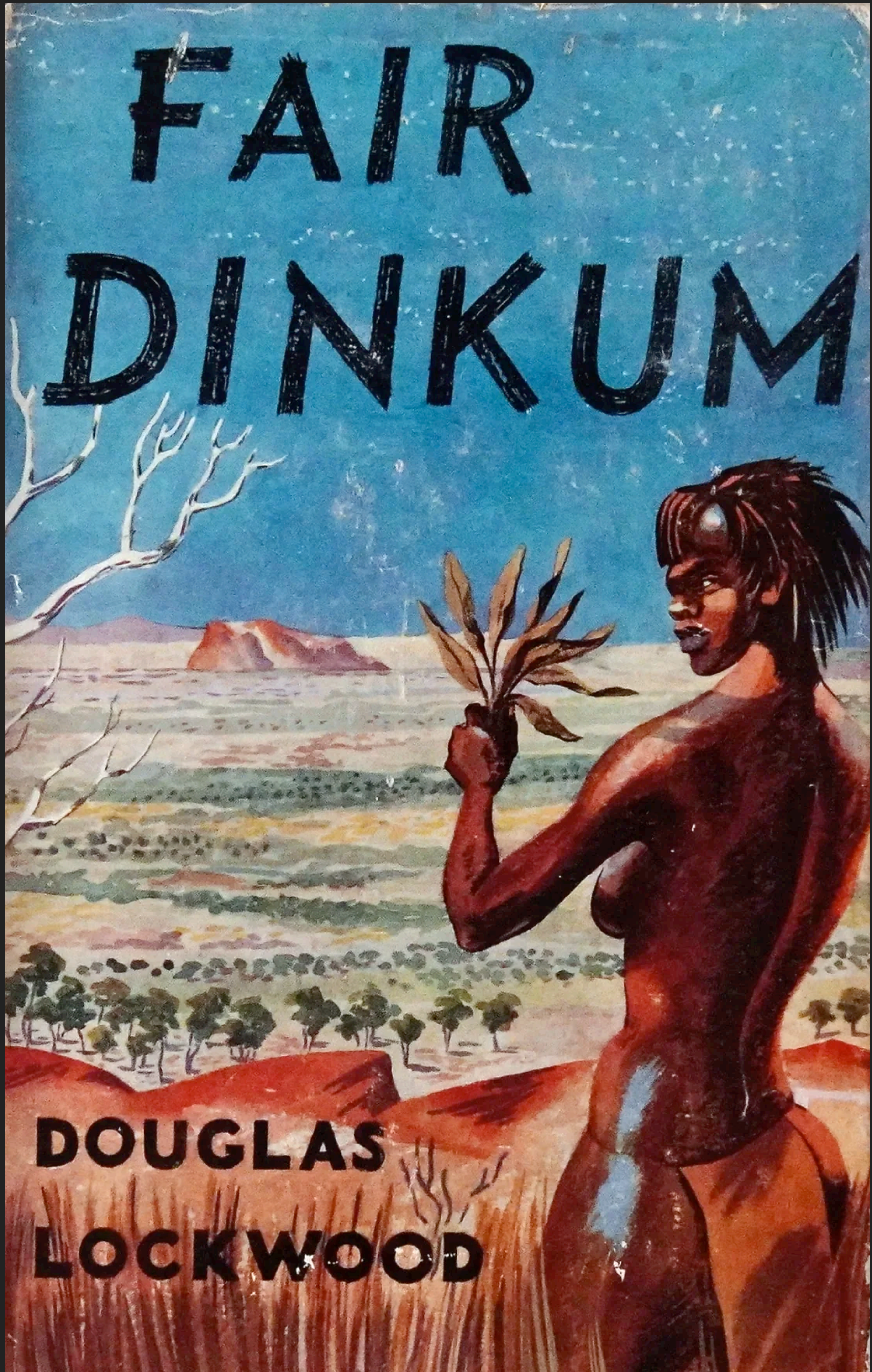
FAIR DYNKIN!

- Fairness as an actuarial problem
- Fairness = Independence
- Independence = Intersections



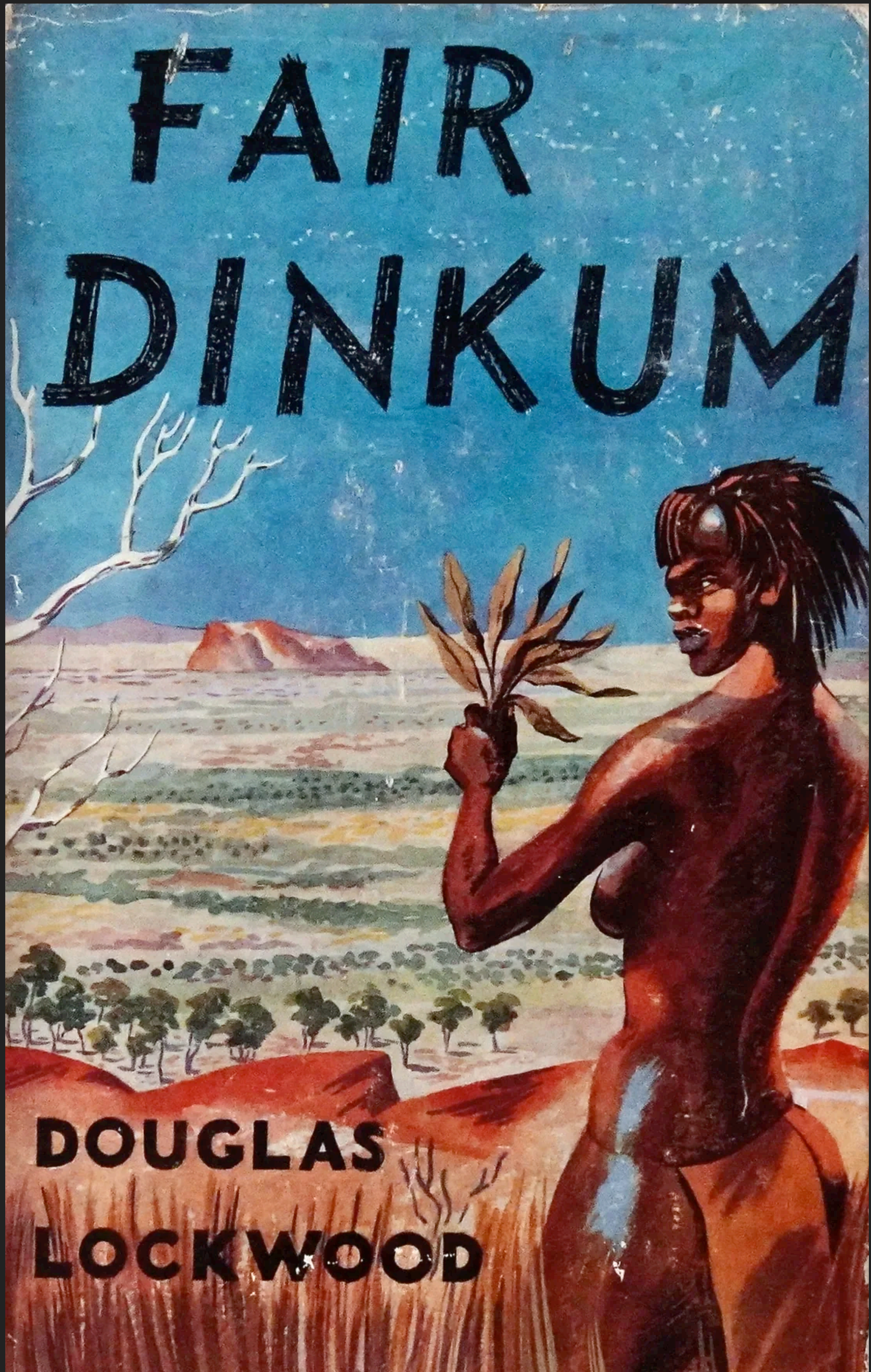
FAIR DYNKIN!

- Fairness as an actuarial problem
- Fairness = Independence
- Independence = Intersections
- Intersectionality = Dynkin systems



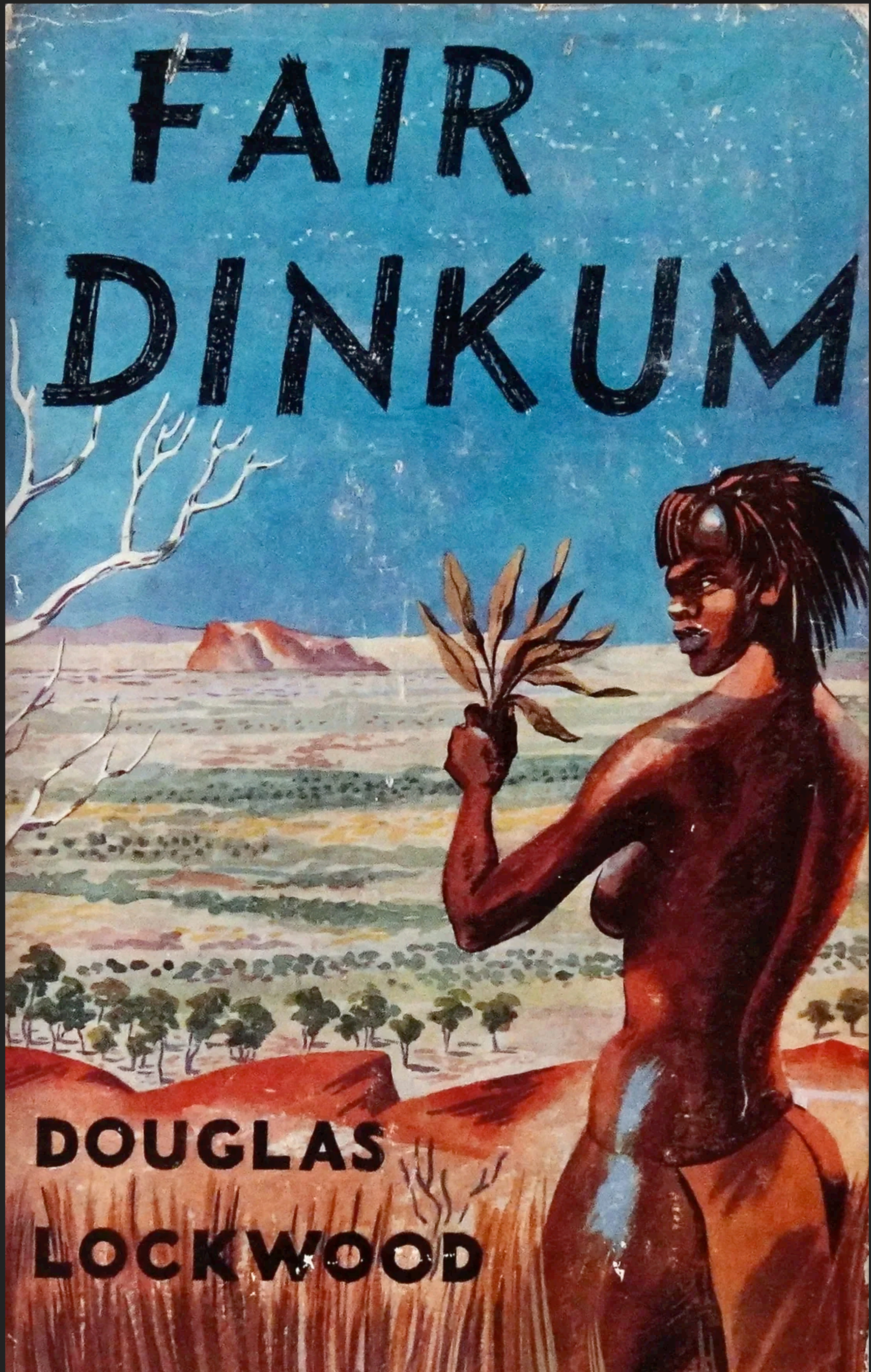
FAIR DYNKIN!

- Fairness as an actuarial problem
- Fairness = Independence
- Independence = Intersections
- Intersectionality = Dynkin systems
- Hence “Fair Dynkin”



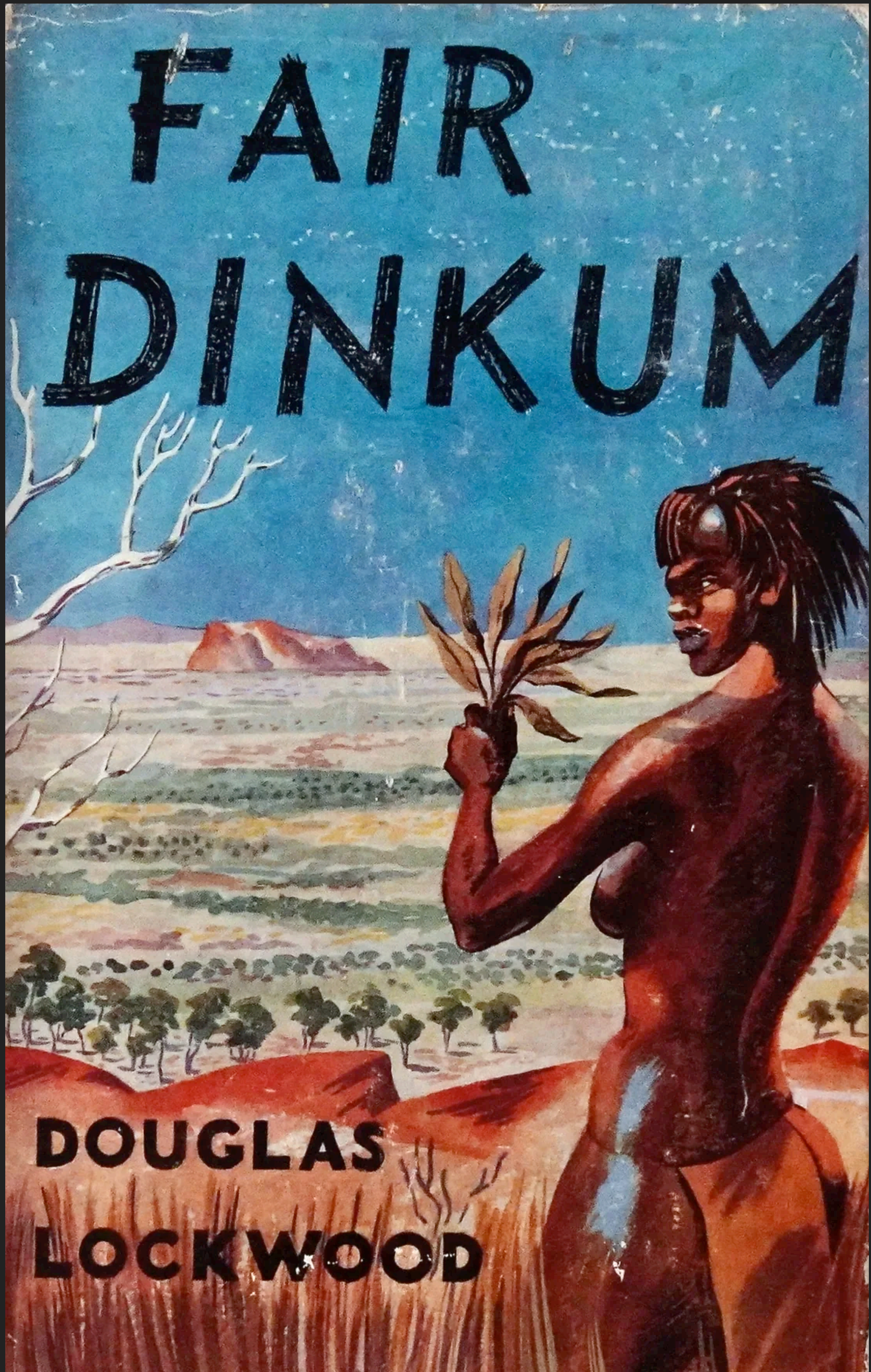
FAIR DYNKIN!

- Fairness as an actuarial problem
- Fairness = Independence
- Independence = Intersections
- Intersectionality = Dynkin systems
- Hence "Fair Dynkin"
- Also: Independence = Randomness



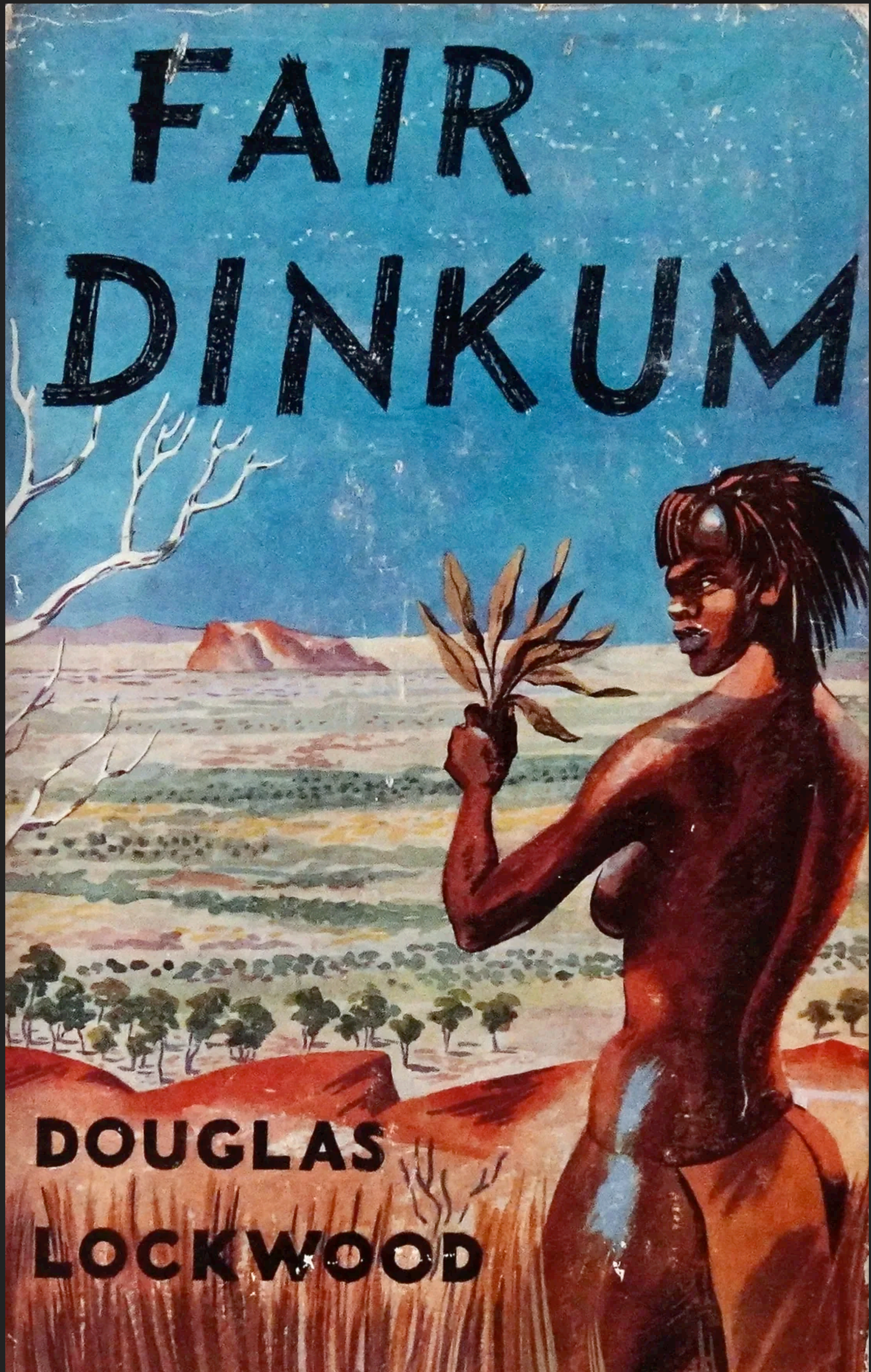
FAIR DYNKIN!

- Fairness as an actuarial problem
- Fairness = Independence
- Independence = Intersections
- Intersectionality = Dynkin systems
- Hence "Fair Dynkin"
- Also: Independence = Randomness
- And randomness inherently pluralistic or relative



FAIR DYNKIN!

- Fairness as an actuarial problem
- Fairness = Independence
- Independence = Intersections
- Intersectionality = Dynkin systems
- Hence "Fair Dynkin"
- Also: Independence = Randomness
- And randomness inherently pluralistic or relative
- Thus too for fairness (no surprise there really)



INTERSECTIONALITY AND IMPRECISION

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

Andrei Nikolaevich Kolmogorov. The general theory of measure and probability calculus. *Collected Works of the Mathematical Section, Communist Academy, Section for Natural and Exact Sciences,* 1:8–21, 1927/1929. In Russian. Translated to English in A.N. Shiryayev (Editor), *Selected Works of A.N. Kolmogorov, Volume II Probability and Mathematical Statistics,* pages 48–59, Springer 1992.

INTERSECTIONALITY AND IMPRECISION

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

Andrei Nikolaevich Kolmogorov. The general theory of measure and probability calculus. *Collected Works of the Mathematical Section, Communist Academy, Section for Natural and Exact Sciences,* 1:8–21, 1927/1929. In Russian. Translated to English in A.N. Shiryayev (Editor), *Selected Works of A.N. Kolmogorov, Volume II Probability and Mathematical Statistics,* pages 48–59, Springer 1992.

- Failure of intersectionality means the system of events is no longer an "algebra"

INTERSECTIONALITY AND IMPRECISION

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

Andrei Nikolaevich Kolmogorov. The general theory of measure and probability calculus. *Collected Works of the Mathematical Section, Communist Academy, Section for Natural and Exact Sciences,* 1:8–21, 1927/1929. In Russian. Translated to English in A.N. Shiryayev (Editor), *Selected Works of A.N. Kolmogorov, Volume II Probability and Mathematical Statistics,* pages 48–59, Springer 1992.

- Failure of intersectionality means the system of events is no longer an "algebra"
- Only closed under *disjoint* unions - a "Dynkin System"

INTERSECTIONALITY AND IMPRECISION

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

Andrei Nikolaevich Kolmogorov. The general theory of measure and probability calculus. *Collected Works of the Mathematical Section, Communist Academy, Section for Natural and Exact Sciences,* 1:8–21, 1927/1929. In Russian. Translated to English in A.N. Shiryayev (Editor), *Selected Works of A.N. Kolmogorov, Volume II Probability and Mathematical Statistics,* pages 48–59, Springer 1992.

- Failure of intersectionality means the system of events is no longer an "algebra"
- Only closed under *disjoint* unions - a "Dynkin System"
- Measure theory is not a technical annoyance to avoid by an incantation

INTERSECTIONALITY AND IMPRECISION

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

Andrei Nikolaevich Kolmogorov. The general theory of measure and probability calculus. *Collected Works of the Mathematical Section, Communist Academy, Section for Natural and Exact Sciences,* 1:8–21, 1927/1929. In Russian. Translated to English in A.N. Shiryayev (Editor), *Selected Works of A.N. Kolmogorov, Volume II Probability and Mathematical Statistics,* pages 48–59, Springer 1992.

- Failure of intersectionality means the system of events is no longer an "algebra"
- Only closed under *disjoint* unions - a "Dynkin System"
- Measure theory is not a technical annoyance to avoid by an incantation
- But a crucial part of one's modelling of the world

A COMMON STORY – SECTION 2, LINE 1…

arXiv > cs > arXiv:2101.02703

Computer Science > Machine Learning

[Submitted on 7 Jan 2021 (v1), last revised 4 Aug 2021 (this version, v3)]

Distribution-Free, Risk-Controlling Prediction Sets

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, Michael I. Jordan

1 Introduction

Black-box predictive algorithms have begun to be deployed in many real-world decision-making settings. Problematically, however, these algorithms are rarely accompanied by reliable uncertainty quantification. Algorithm developers often depend on the standard training/validation/test paradigm to make assertions of accuracy, stopping short of any further attempt to indicate that an algorithm's predictions should be treated with skepticism. Thus, prediction failures will often be silent ones, which is particularly alarming in high-consequence settings.

2.1 Setting and notation

Let $(X_i, Y_i)_{i=1,\ldots,m}$ be an independent and identically distributed (i.i.d.) set of variables, where the features

A COMMON STORY – SECTION 2, LINE 1...

arXiv > cs > arXiv:2101.02703

Computer Science > Machine Learning

[Submitted on 7 Jan 2021 (v1), last revised 4 Aug 2021 (this version, v3)]

Distribution-Free, Risk-Controlling Prediction Sets

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, Michael I. Jordan

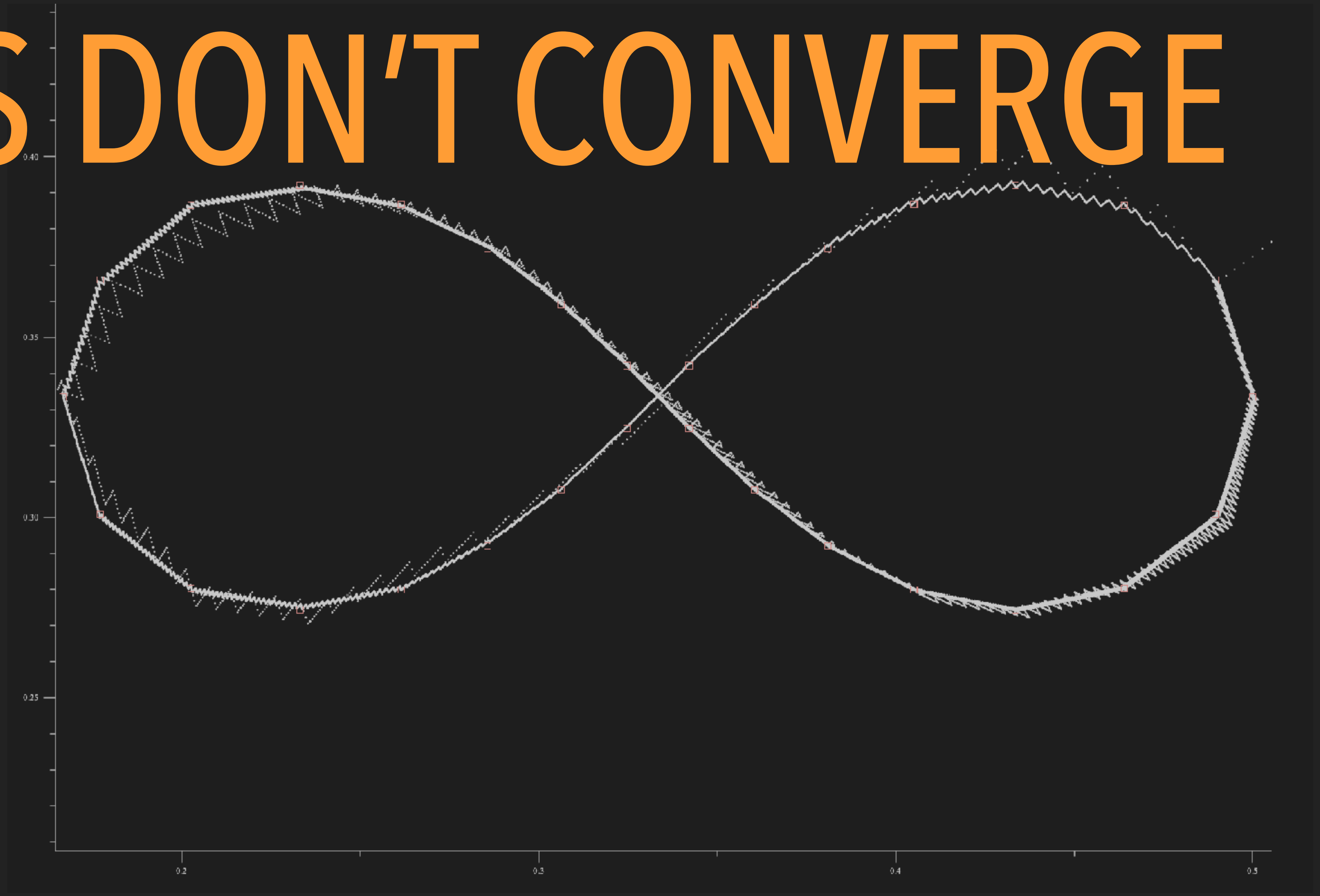
1 Introduction

Black-box predictive algorithms have begun to be deployed in many real-world decision-making settings. Problematically, however, these algorithms are rarely accompanied by reliable uncertainty quantification. Algorithm developers often depend on the standard training/validation/test paradigm to make assertions of accuracy, stopping short of any further attempt to indicate that an algorithm's predictions should be treated with skepticism. Thus, prediction failures will often be silent ones, which is particularly alarming in high-consequence settings.

2.1 Setting and notation

Let $(X_i, Y_i)_{i=1,\ldots,m}$ be an independent and identically distributed (i.i.d.) set of variables, where the features

WHEN RELATIVE FREQUENCIES DON'T CONVERGE



- "Non-stochastic randomness"
- Start with sequences (the data)
- Compute relative frequencies
- Von Mises assumes they converge to a limit – "the" probability
- What happens when they don't? (And no, there is no "law" that says they do)
- Multiple "cluster points" – generalisation of the mathematical limit
- *Every* sequence generates a sequence of relative frequencies with a set of cluster points
- *Any* connected set is the set of cluster points of the relative frequencies of some sequence