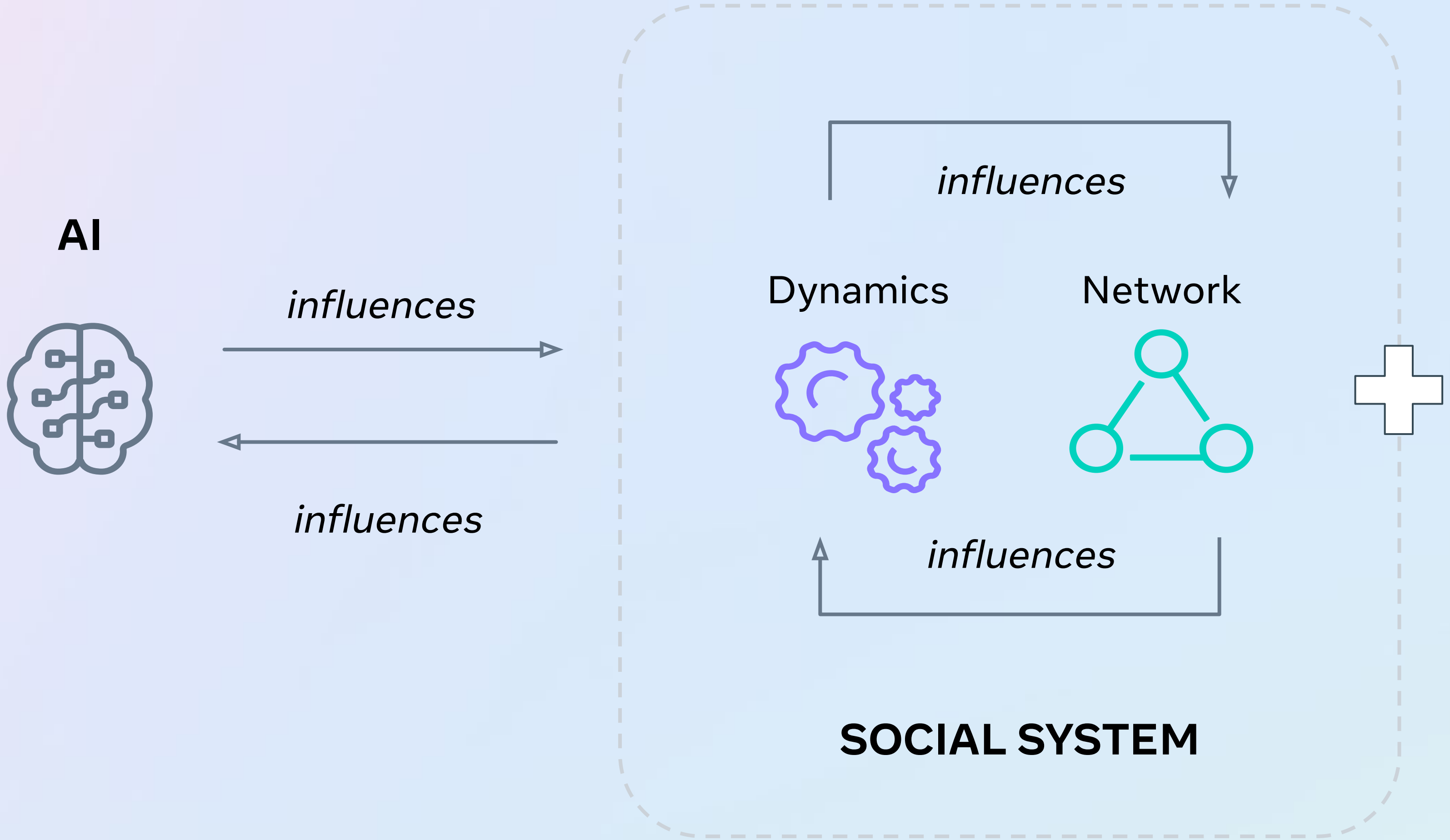
Agenda for today

Convince you that this equation holds and is relevant to your work



AI

influences

influences

influences

Dynamics

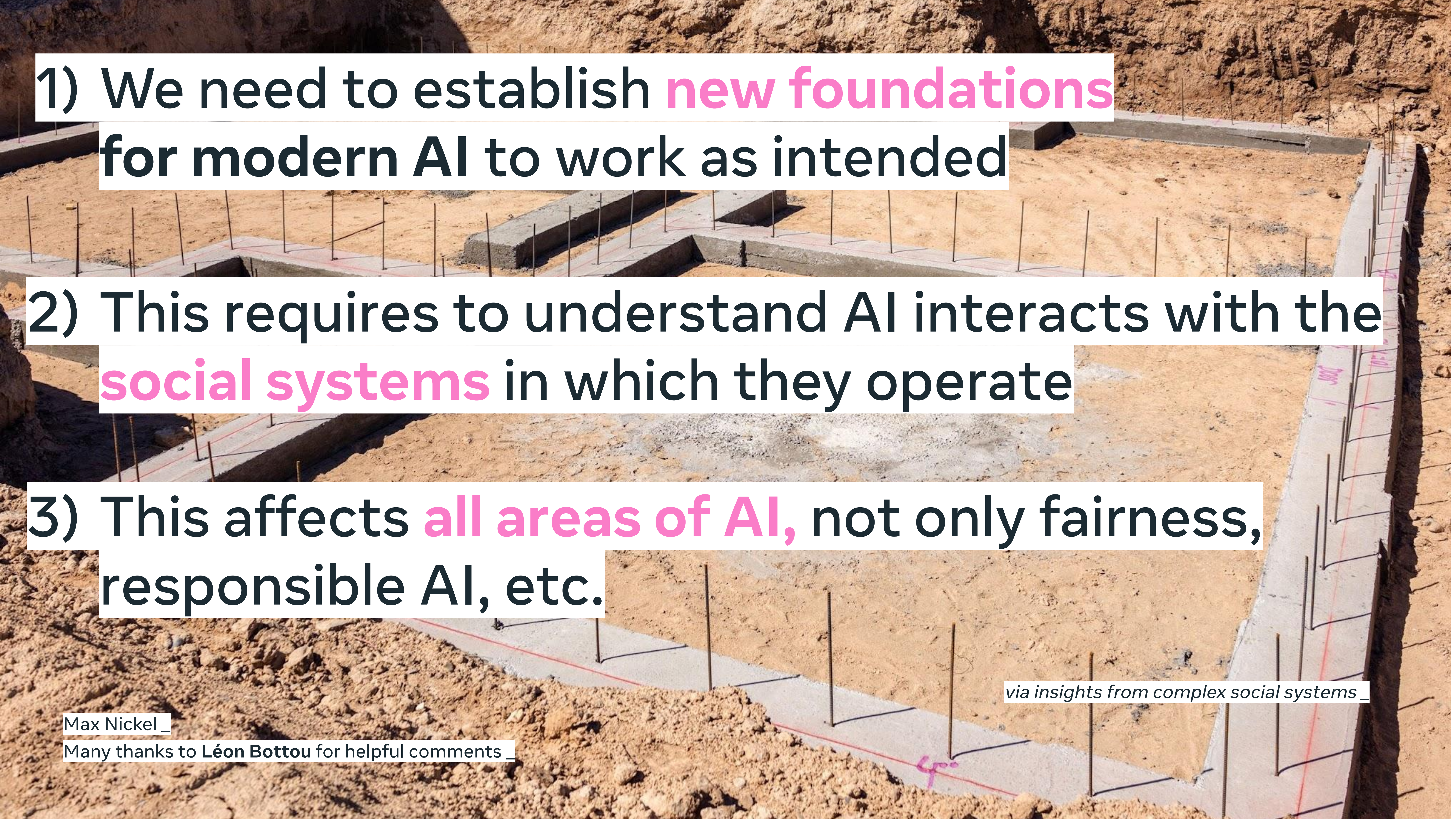
Network

influences

SOCIAL SYSTEM

+

=



1) We need to establish **new foundations** **for modern AI** to work as intended

2) This requires to understand AI interacts with the **social systems** in which they operate

3) This affects **all areas of AI,** not only fairness, responsible AI, etc.

via insights from complex social systems _

Max Nickel _

Many thanks to **Léon Bottou** for helpful comments _ can we **understand**

mpressive

ways at

the same

Are our **benchmarks** give insights into the intended tasks or do they project a false image of quality?

d to

Will naive **scaling** solve all our problems?

AI **Paradigm Shift_**

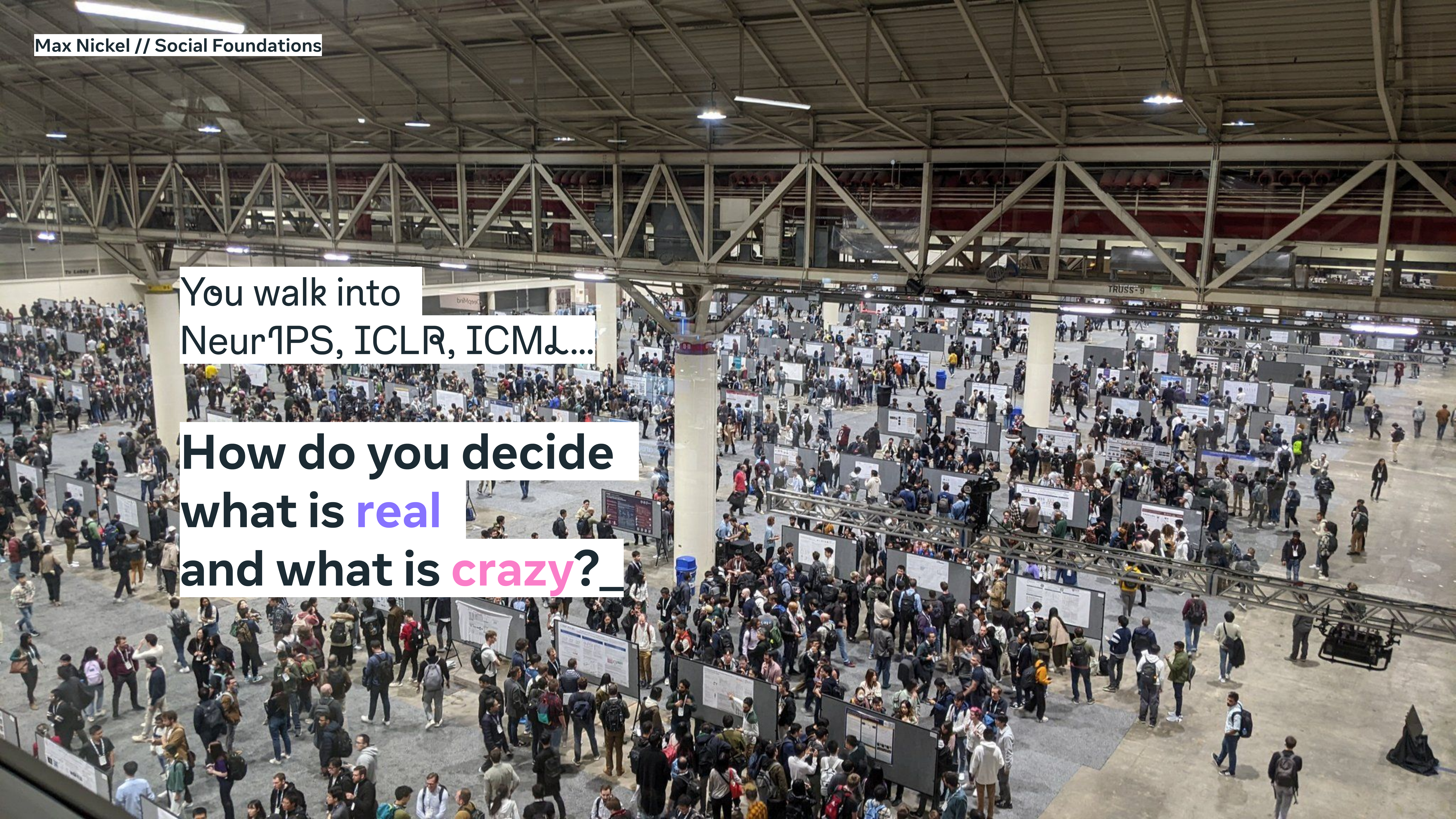
Tell me if this is an Iris or not.

Answer any question truthfully about any object in the known universe.

Our theory (read justification) is built for this case

Does it still hold for what we are doing now?
Easy: Obviously not…
Harder: CAN IT EVER be valid?

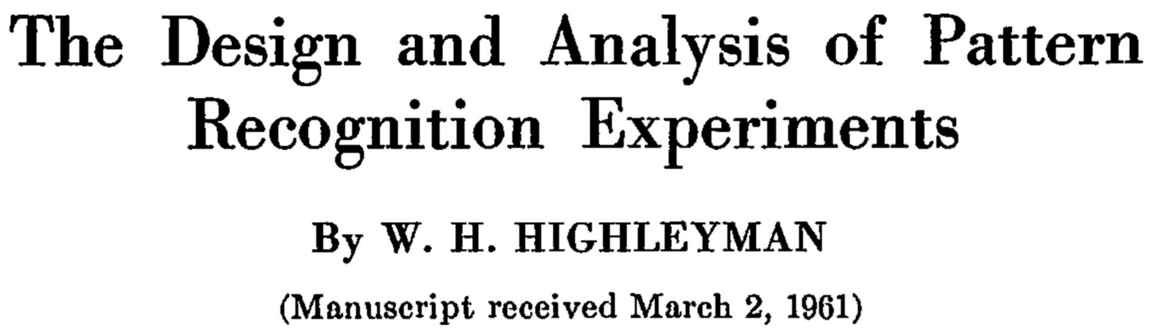
AI **Paradigm Shift_**



You walk into
NeurIPS, ICLR, ICML...

How do you decide what is real and what is crazy?_

Train-Test Paradigm



The Design and Analysis of Pattern Recognition Experiments

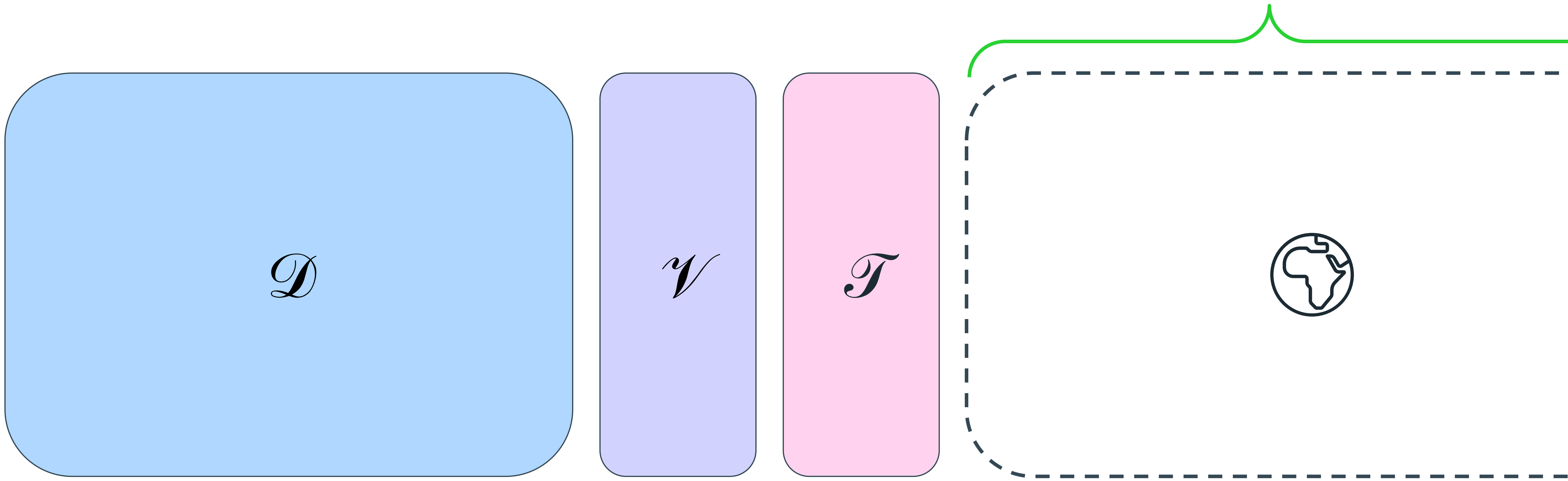
By W. H. HIGHLEYMAN

(Manuscript received March 2, 1961)

THE dominant / only approach to model validation in modern ML.

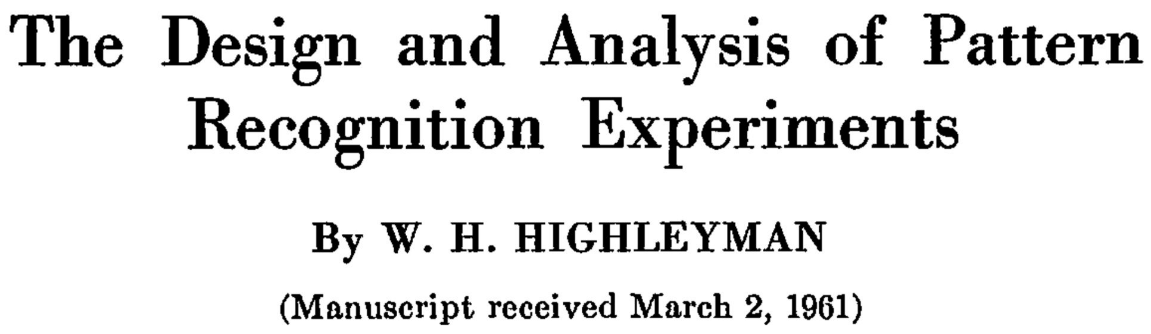
- Training set $\mathscr{D}$
- Validation set $\mathscr{V}$
- Test set $\mathscr{T}$

Rapid model validation via the train-test paradigm has been a key driver for the breathtaking progress in machine learning and AI (e.g. see *Bottou 2015*).



This is the only thing we care about.

Train-Test Paradigm



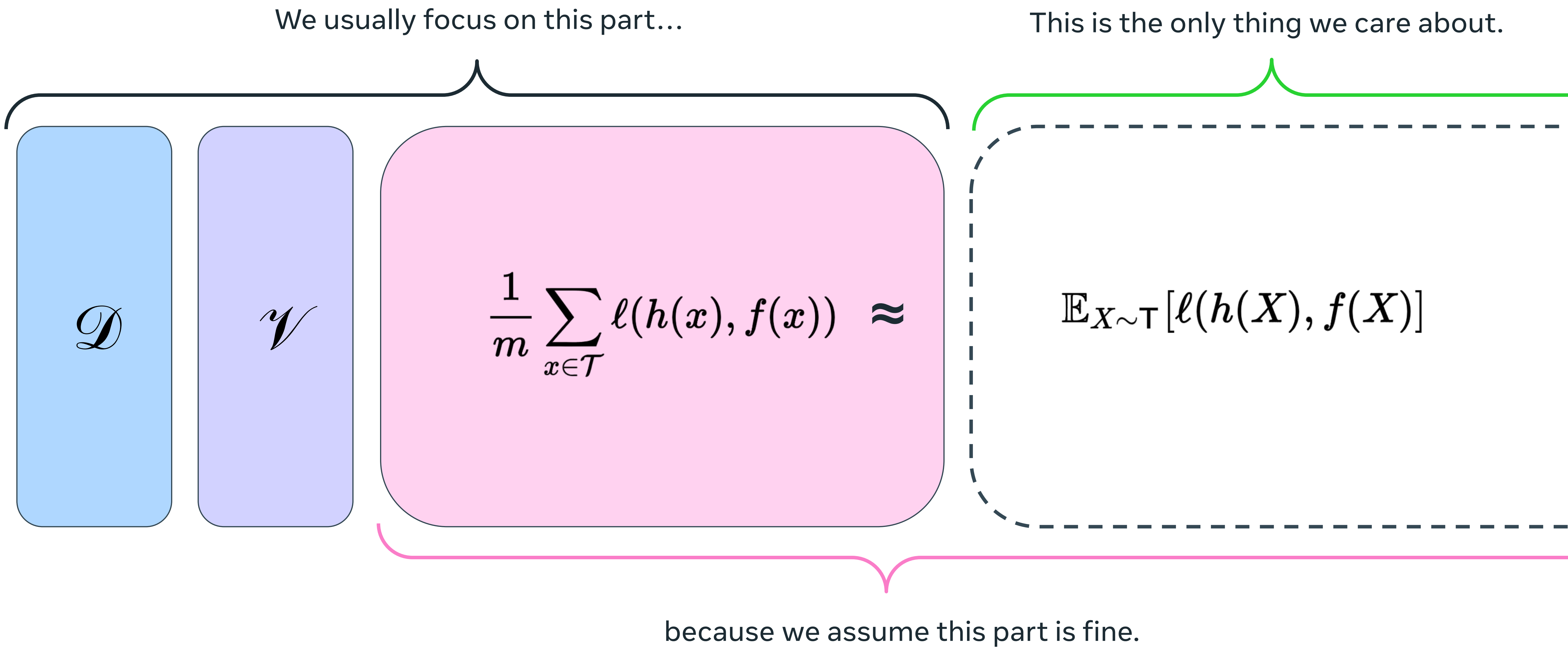
The Design and Analysis of Pattern Recognition Experiments

By W. H. HIGHLEYMAN

(Manuscript received March 2, 1961)

Rapid model validation via the train-test paradigm has been a key driver for the breathtaking progress in machine learning and AI (e.g. see *Bottou 2015*).

- Estimated model h
- True world f
- Loss function ℓ
- Target distribution T
- Test set $\mathscr{T}$


We usually focus on this part…
This is the only thing we care about.
$\mathscr{D}$
$\mathscr{V}$
because we assume this part is fine.

$$\frac{1}{m}\sum_{x\in\mathcal{T}}\ell(h(x), f(x)) \approx \mathbb{E}_{X\sim\mathsf{T}}[\ell(h(X), f(X))]$$

Train-Test Paradigm

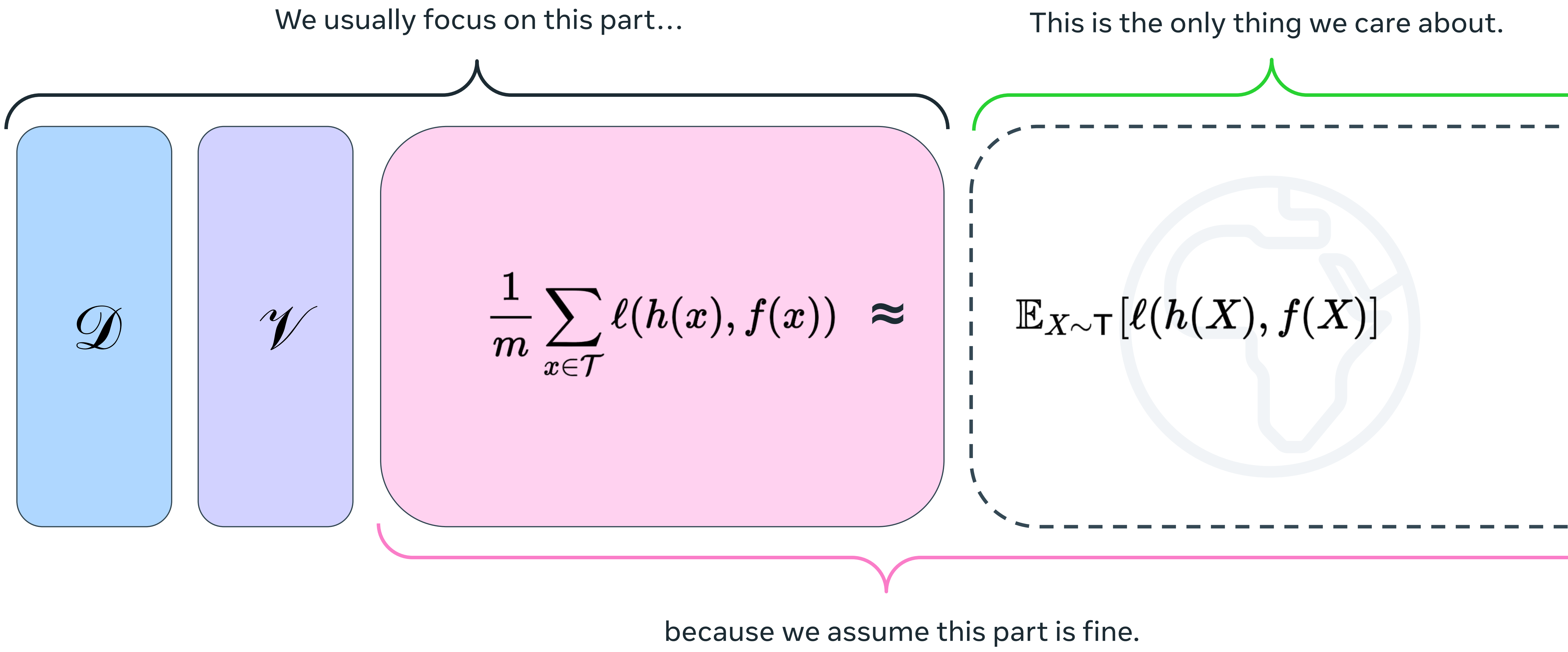
The Design and Analysis of Pattern Recognition Experiments

By W. H. HIGHLEYMAN

(Manuscript received March 2, 1961)

Rapid model validation via the train-test paradigm has been a key driver for the breathtaking progress in machine learning and AI (e.g. see *Bottou 2015*).

- Estimated model *h*
- True world *f*
- Loss function ℓ
- Target distribution *T*
- Test set $\mathscr{T}$



We usually focus on this part…
This is the only thing we care about.
$\mathscr{D}$
$\mathscr{V}$
because we assume this part is fine.

$$\frac{1}{m}\sum_{x\in\mathcal{T}} \ell(h(x), f(x)) \approx \mathbb{E}_{X\sim \mathsf{T}}[\ell(h(X), f(X)]$$

Train-Test Paradigm

Rapid model validation via the train-test paradigm has been a key driver for the breathtaking progress in machine learning and AI (e.g. see *Bottou 2015*).

- Estimated model h
- True world f
- Loss function ℓ
- Target distribution T
- Test set $\mathscr{T}$

We usually focus on this pa

g we care about.

$\mathscr{D}$

$\mathscr{V}$

$$\frac{1}{m}\sum_{x\in\mathcal{T}}\ell(h(x), f(x)) \approx \mathbb{E}_{X\sim T}[\ell(h(X), f(X))]$$

because we assume this part is fine.

Train-Test Paradigm

Rapid model validation via the train-test paradigm has been a key driver for the breathtaking progress in machine learning and AI (e.g. see *Bottou 2015*).

- ○ Estimated model *h*
- ○ True world *f*
- ○ Loss function ℓ
- ○ Target distribution *T*
- ○ Test set $\mathscr{T}$

We usually focus on this pa

g we care about.

$\mathscr{D}$

$\mathscr{V}$

$$\frac{1}{m}\sum_{x\in\mathscr{T}}\ell(h(x),f(x)) \approx$$

$$\mathbb{E}_{X\sim T}[\ell(h(X),f(X))]$$

Inductive inference NOT measurement

because we assume this part is fine.

Is **Induction** Possible?

Fundamental question in science, dating back at least to Hume's **problem of induction** *(1739)*:

"even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience"

In ML, Wolpert's **No-Free-Lunch** theorem *(1996)* established formally that statistical learning/prediction is **impossible without making assumptions** about the world.

Staying close to **Non-Uniformity** of Nature

“Domestic animals expect food when they see the person who usually feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that **more refined views as to the uniformity of nature would have been useful to the chicken.**”

Bertrand Russell — *The Problems of Philosophy*

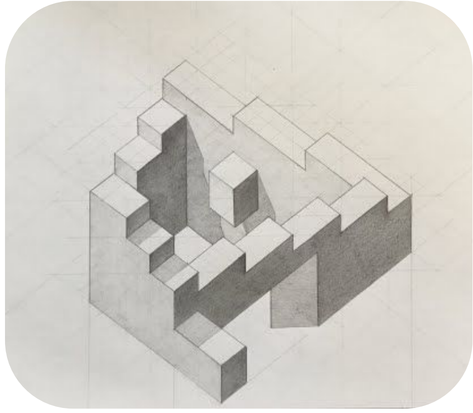
Ontological **Parsimony**

Hume's argument, or to say it w/ W. v. O. Quine (1969)

"The Humean condition is the human condition."

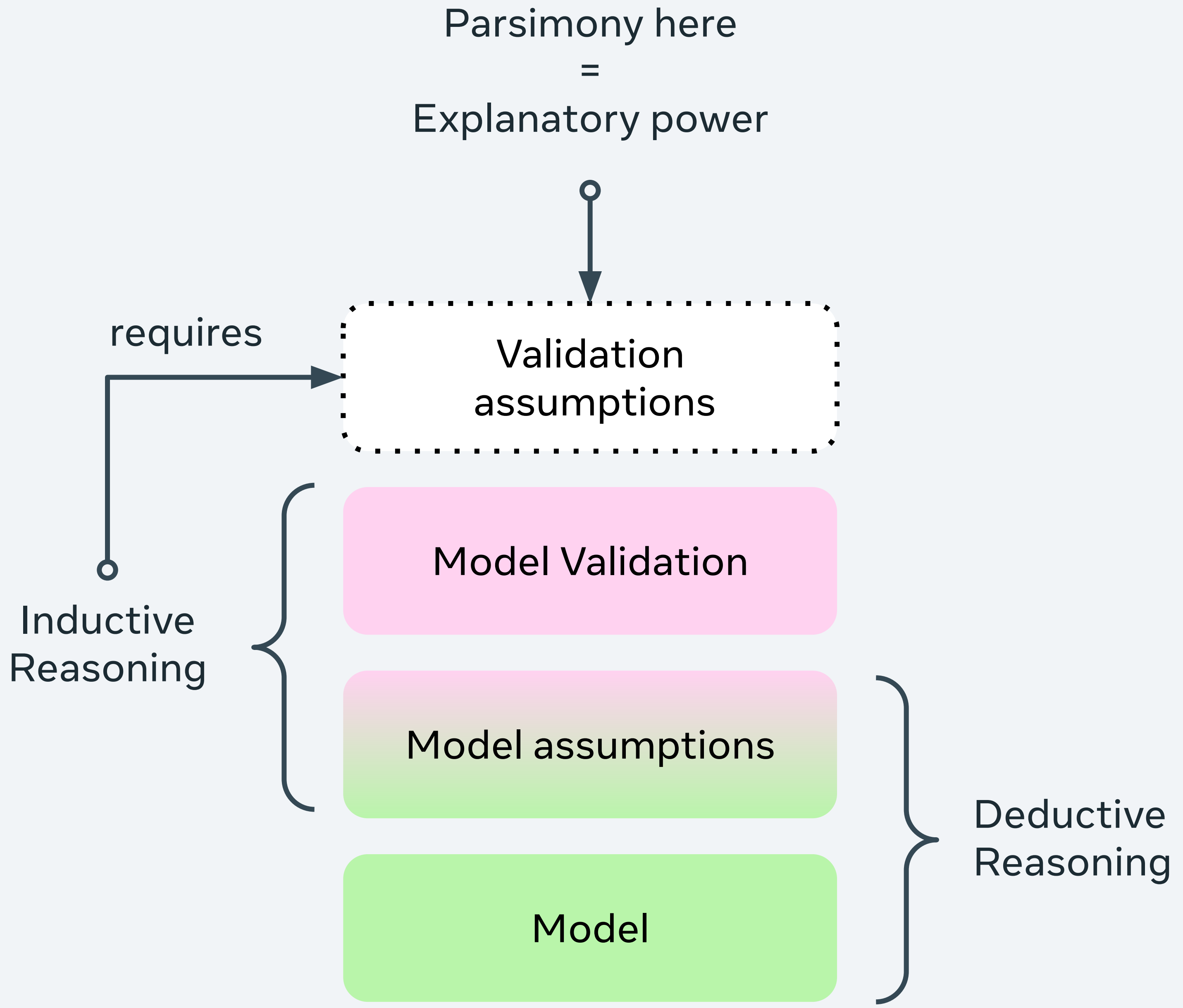
But we can ask: Are there some **reasonable assumptions** that we might be willing to make that can **neutralize the problem of induction**?

- We can validate model assumptions via model validation.
- We never can validate assumptions necessary to ensure the validity of the model validation itself w/o **circular reasoning or infinite regress.**



Preferring **parsimonious** hypotheses is rational — they have greater **explanatory power** than less parsimonious alternatives. *(Baker, 2003)*

Hierarchy of Validity



David Hume *hates* this one simple trick

$$\overbrace{\mathbb{E}_{X\sim\mathsf{T}}[\ell(h(X),f(X))]}^{\text{True risk } L^{\mathsf{T}}_{fh}} \approx \overbrace{\frac{1}{m}\textstyle\sum_{x\in\mathcal{T}}\ell(h(x),f(x))}^{\text{Empirical risk } \theta}$$

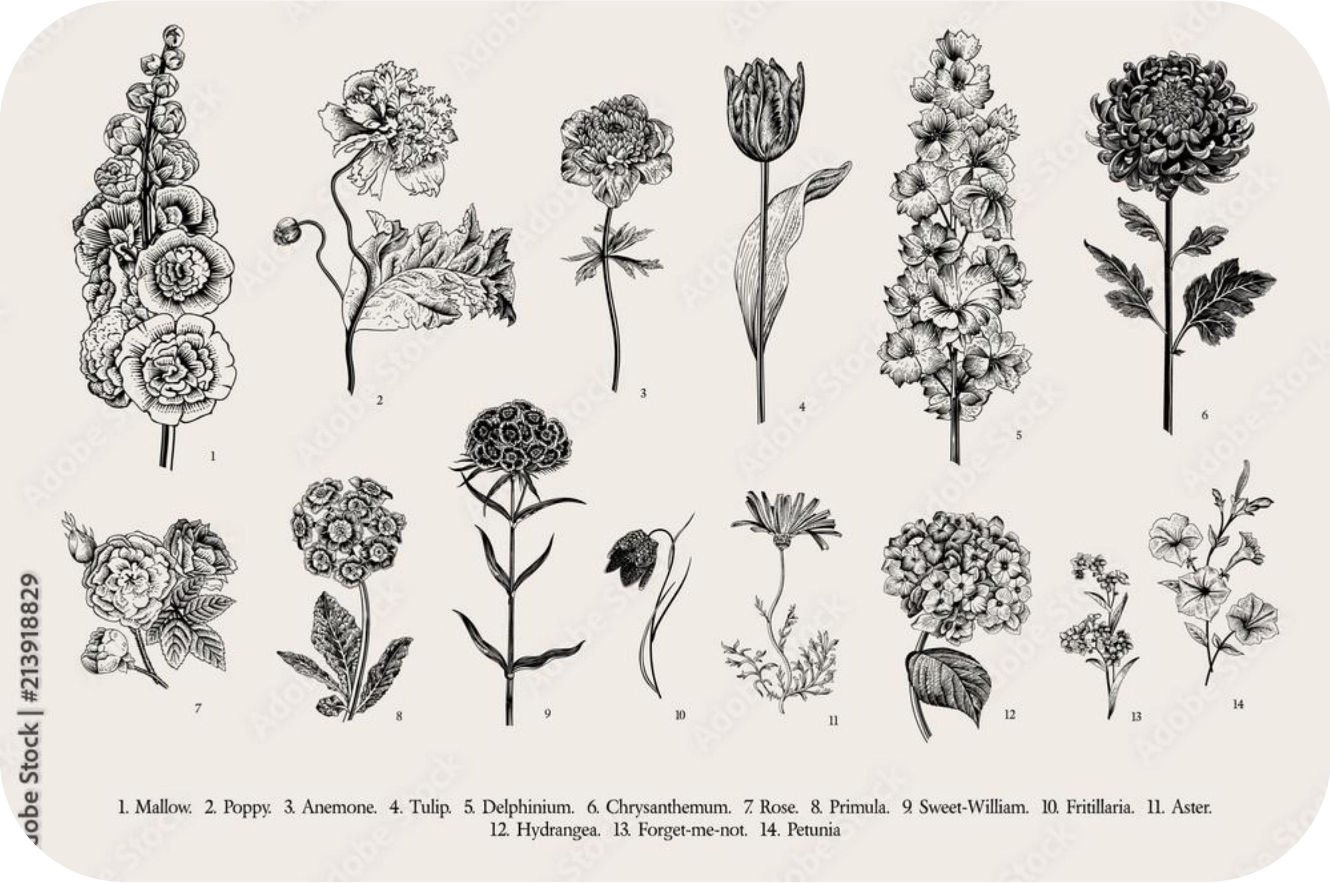
I.I.D. assumption
enables straightforward proof via Hoeffding's inequality

$$\underbrace{\mathbb{P}_{\mathcal{T}\sim\mathsf{T}^m}\left(|\theta - L^{\mathsf{T}}_{fh}\,| \le \sqrt{\frac{\log(2/\delta)}{2m}}\right) \ge 1-\delta}_{}$$

(ε,δ)-guarantee

With probability larger than 1 - δ, the error will be smaller than ε

- ○ ε = accuracy parameter
- ○ δ = confidence parameter


1. Mallow. 2. Poppy. 3. Anemone. 4. Tulip. 5. Delphinium. 6. Chrysanthemum. 7. Rose. 8. Primula. 9. Sweet-William. 10. Fritillaria. 11. Aster.
12. Hydrangea. 13. Forget-me-not. 14. Petunia
obe Stock | #213918829

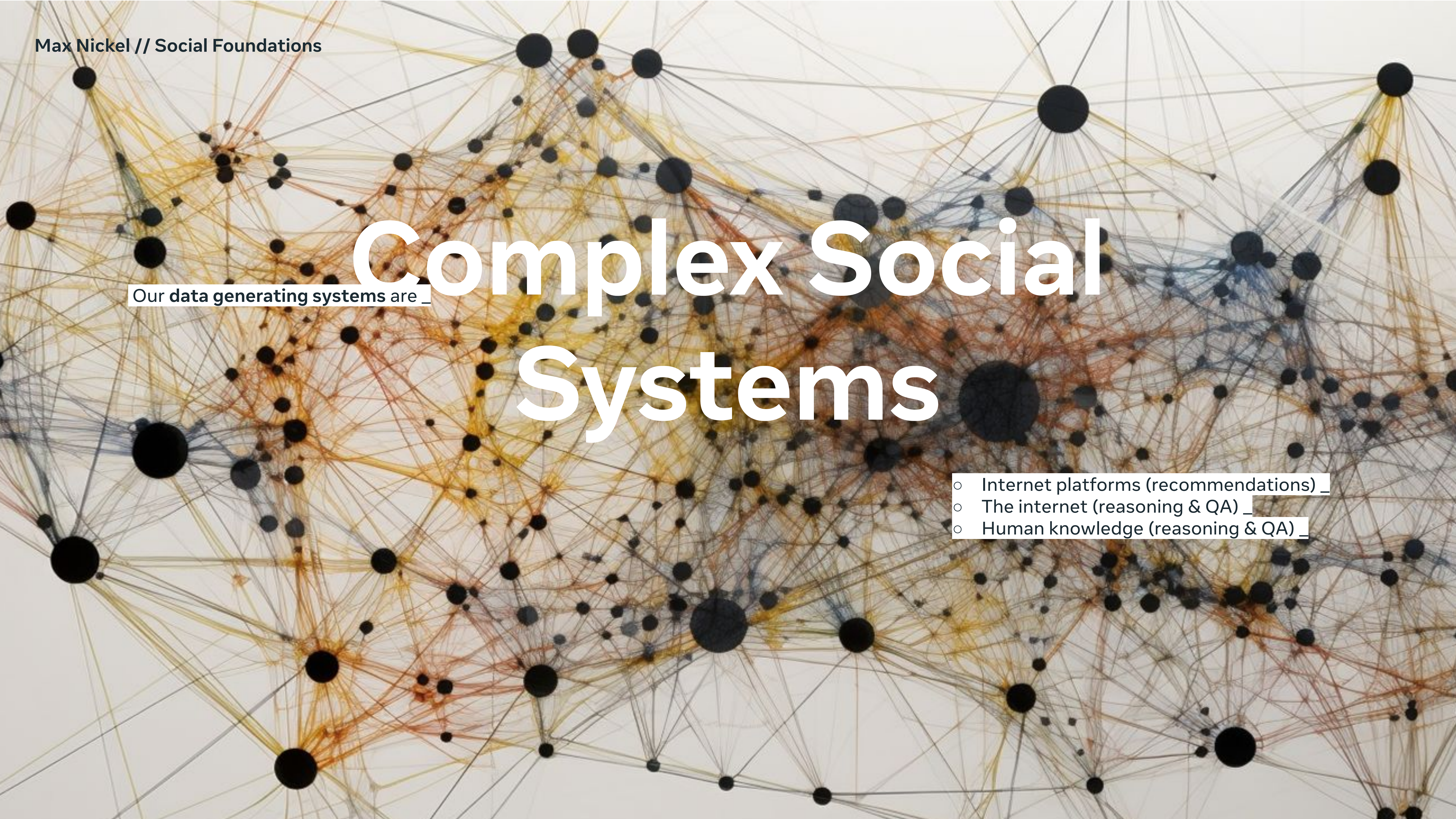
What we can **justify**

- **Actively collected data** to satisfy IID assumption
- Corresponds to target distribution
- Very costly, does not scale to large data sets
- **Scope:** closed domain

Data generating system
(Delivery service)

What we are **doing**

- **Passively collected data** from *some* data generating system
- Non-IID, does not correspond to target distribution
- Cheap, easy to scale when access to the system
- Provides massive datasets required for modern AI
- **Scope**: open domain



Our **data generating systems** are _

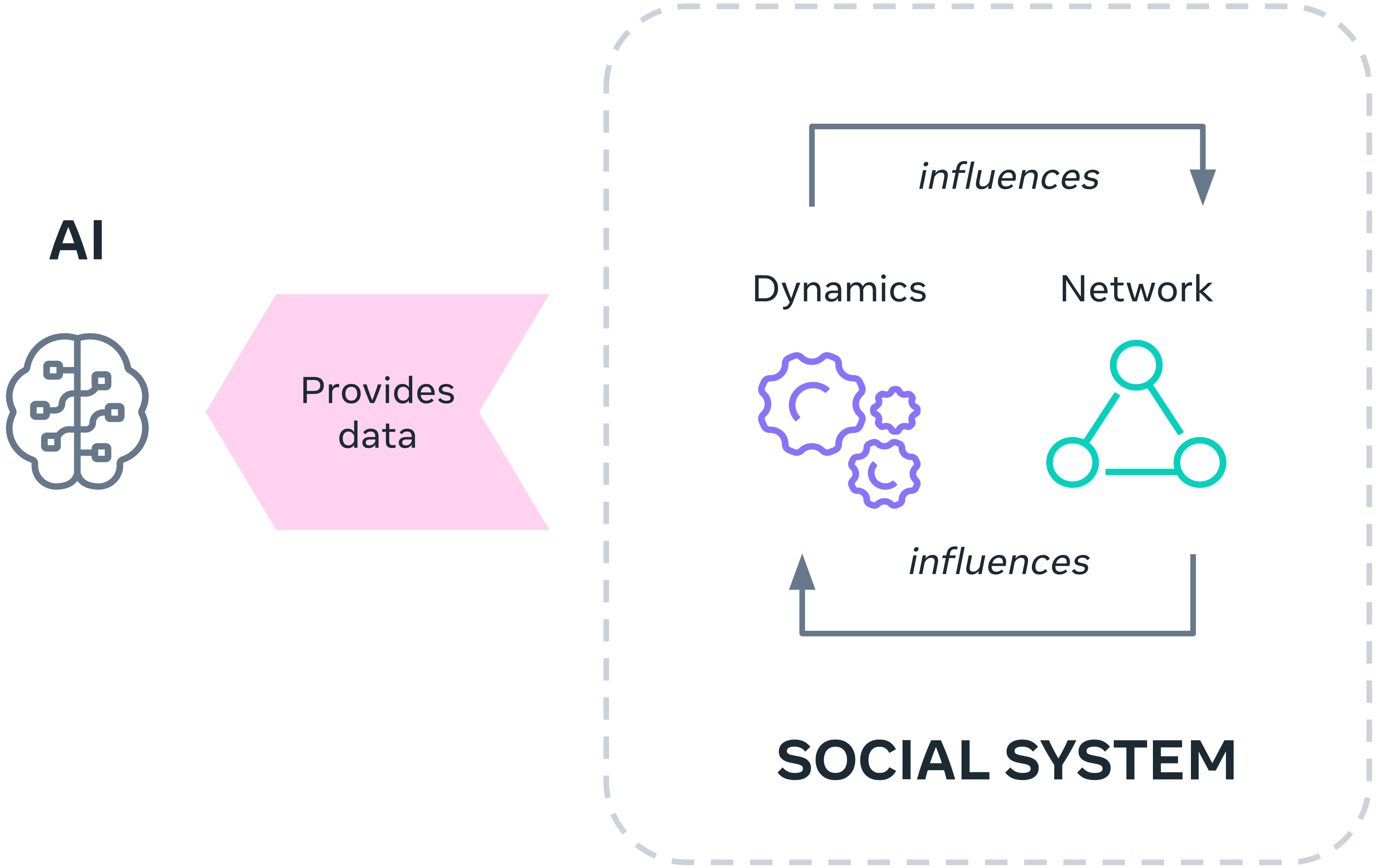
Complex Social Systems

- ○ Internet platforms (recommendations) _
- ○ The internet (reasoning & QA) _
- ○ Human knowledge (reasoning & QA) _

Social Systems
≈
Complex Systems

Often, we can understand social systems as complex systems, i.e., as systems with

- **Interactions** of their parts
- Internal **dynamics**
- **Non-linearities** and chaotic behavior
- Memory and **feedback**
- **Emergent** properties and behavior



AI

Provides data

influences

Dynamics

Network

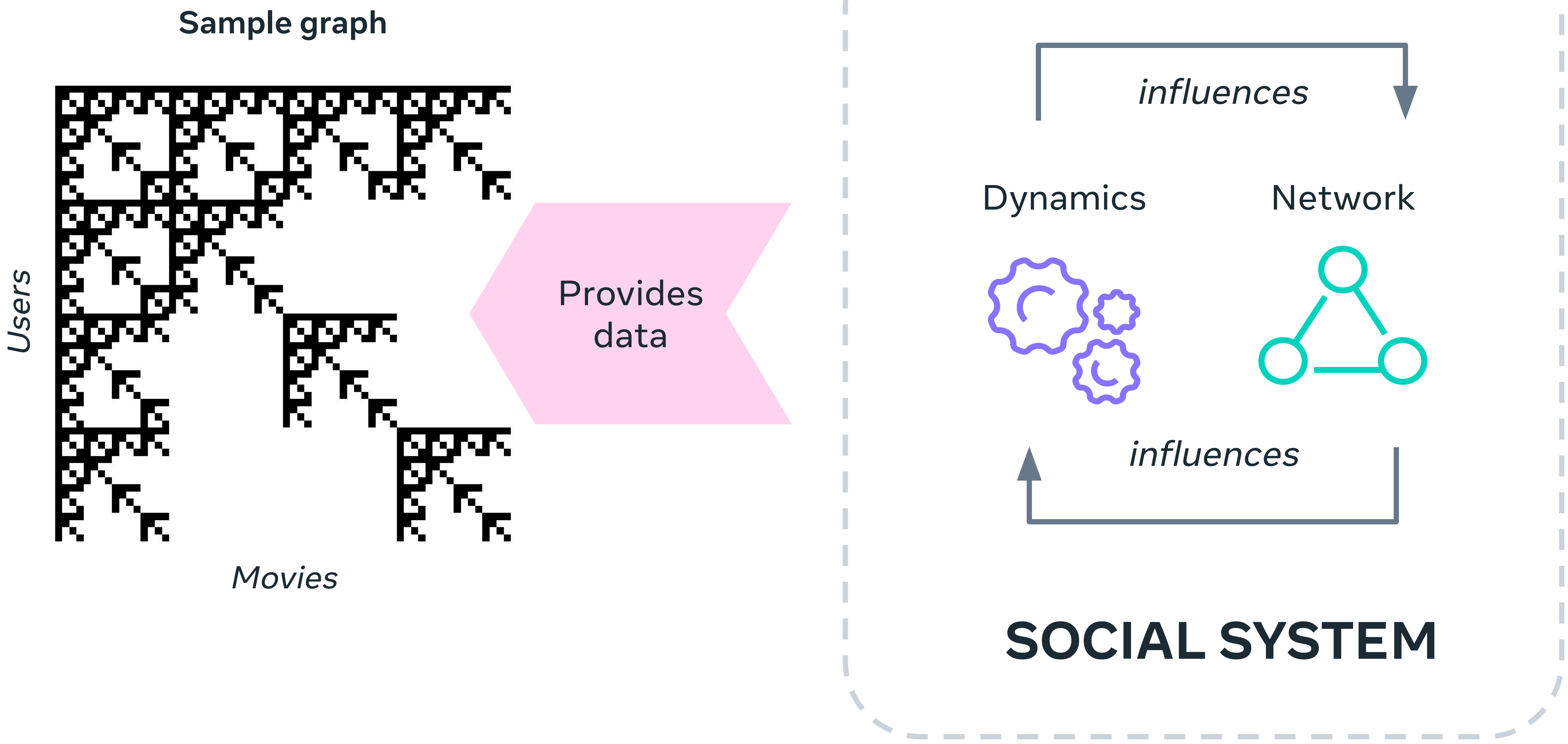
influences

SOCIAL SYSTEM

Modeling **passive data collection**

Formalize data collection via **sample graphs**

Edge in a sample graph denotes an observed data point (noise free)



Sample graph

Users

Movies

Provides data

influences

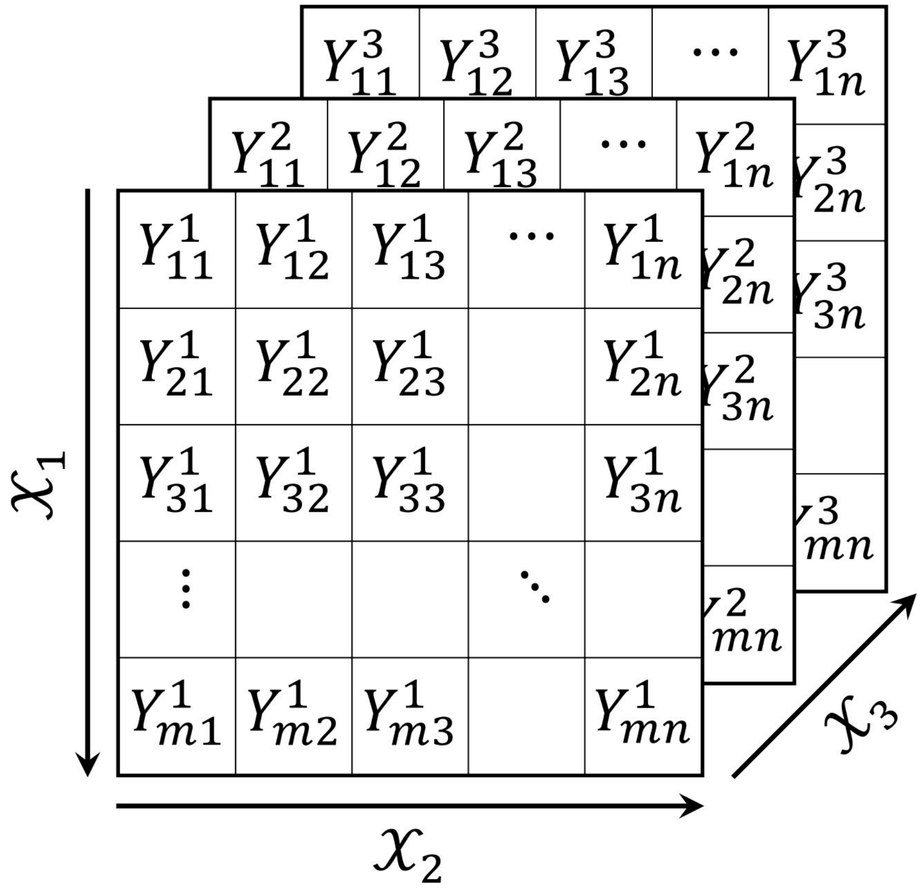
Dynamics

Network

influences

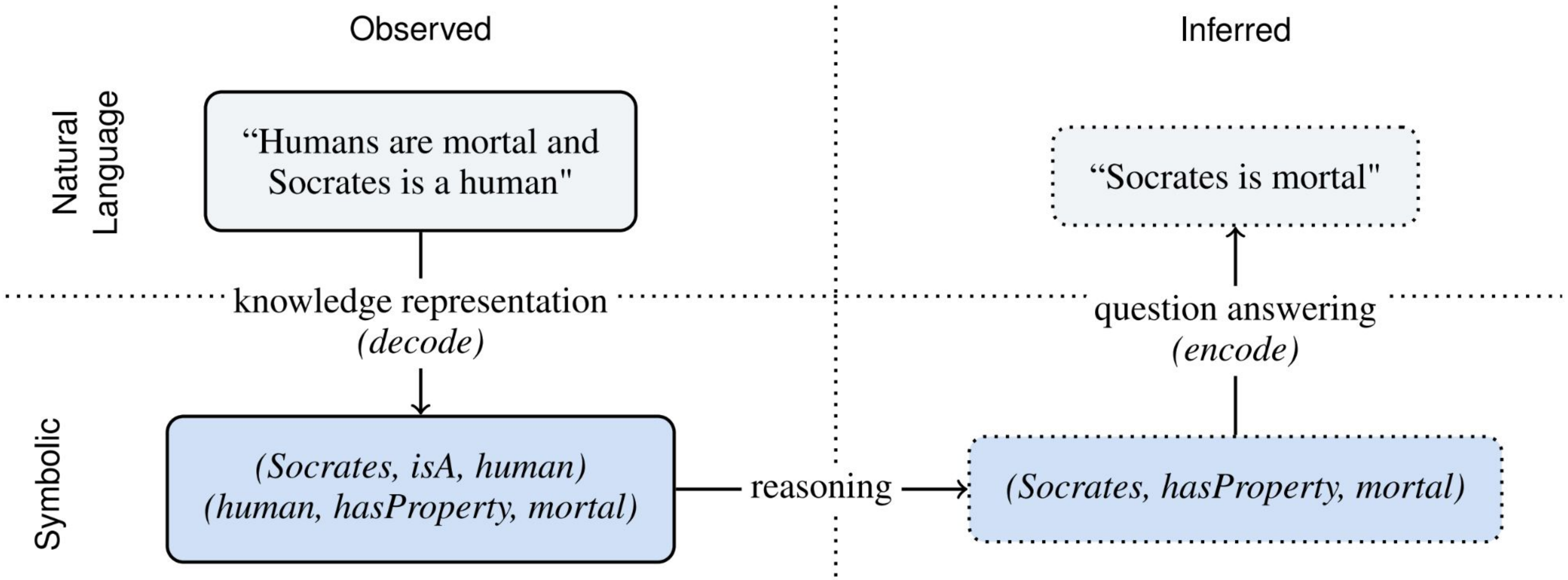
SOCIAL SYSTEM

Modeling passive data collection


Y^3_{11} Y^3_{12} Y^3_{13} $\cdots$ Y^3_{1n}
Y^3_{2n}
Y^3_{3n}
Y^3_{mn}
Y^2_{11} Y^2_{12} Y^2_{13} $\cdots$ Y^2_{1n}
Y^2_{2n}
Y^2_{3n}
Y^2_{mn}
Y^1_{11} Y^1_{12} Y^1_{13} $\cdots$ Y^1_{1n}
Y^1_{21} Y^1_{22} Y^1_{23} Y^1_{2n}
Y^1_{31} Y^1_{32} Y^1_{33} Y^1_{3n}
$\vdots$ $\ddots$
Y^1_{m1} Y^1_{m2} Y^1_{m3} Y^1_{mn}
$\mathcal{X}_1$
$\mathcal{X}_2$
$\mathcal{X}_3$

Formalize data collection via **sample graphs**

Edge in a sample graph denotes an observed data point (noise free)


Observed

Inferred

Natural Language

Symbolic

“Humans are mortal and Socrates is a human"

knowledge representation *(decode)*

(Socrates, isA, human)
(human, hasProperty, mortal)

reasoning

(Socrates, hasProperty, mortal)

question answering *(encode)*

“Socrates is mortal"

Modeling passive data collection

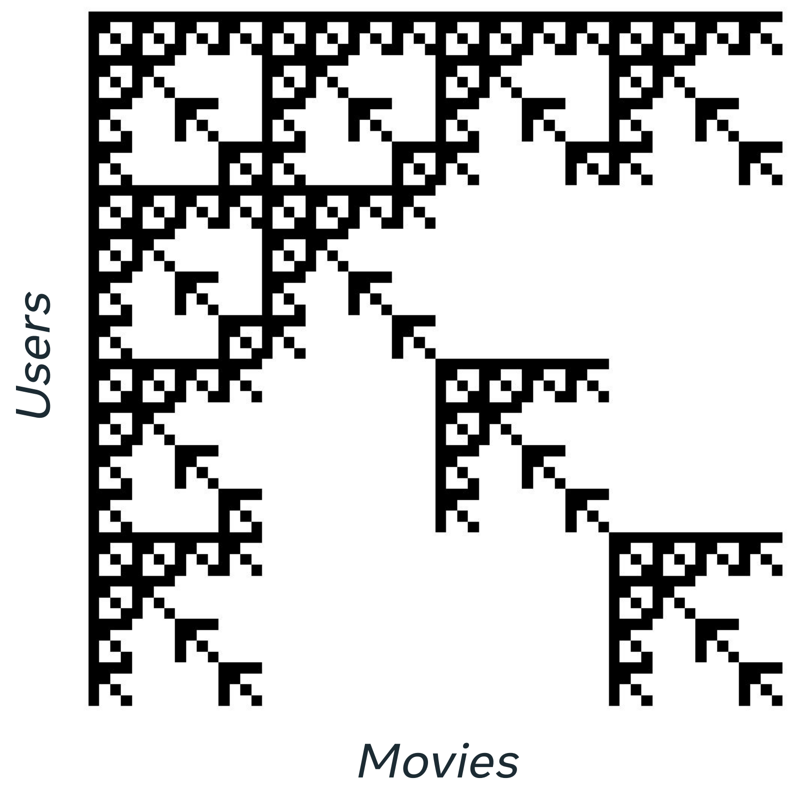
What kind of data do our sample generating systems generate? Since they are social systems, they usually exhibit

i) **Sample bias**
ii) **Heavy-tailed / Power-law distributed observations**

$$\mathbb{P}(K_1 > k) = u_1(k)k^{-\alpha_1} \quad \text{and} \quad \mathbb{P}(K_2 > k) = u_2(k)k^{-\alpha_2}$$

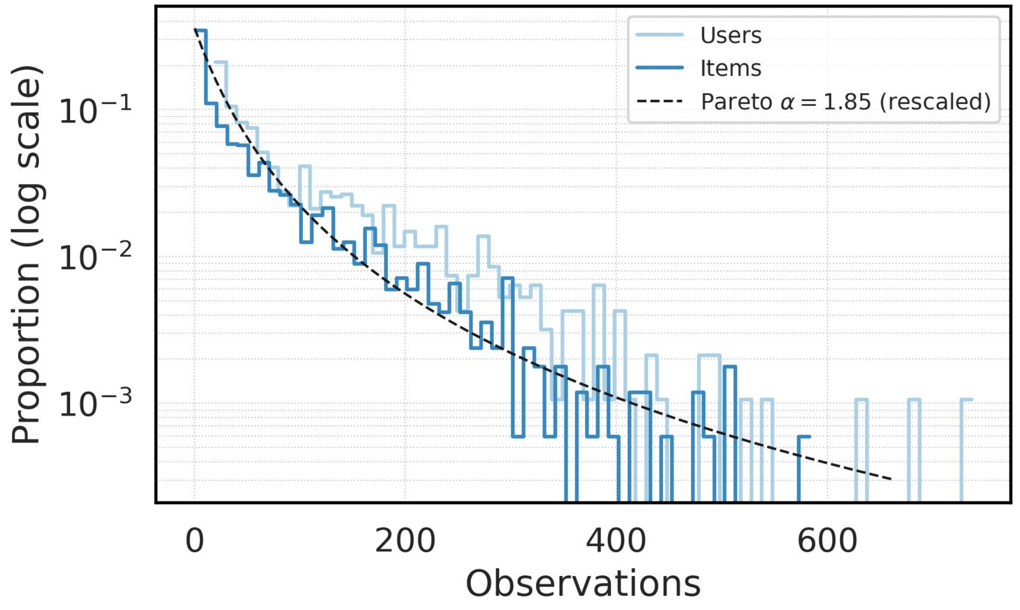
caused by well-documented processes such as popularity bias, homophily, feedback loops, etc

Sample graph

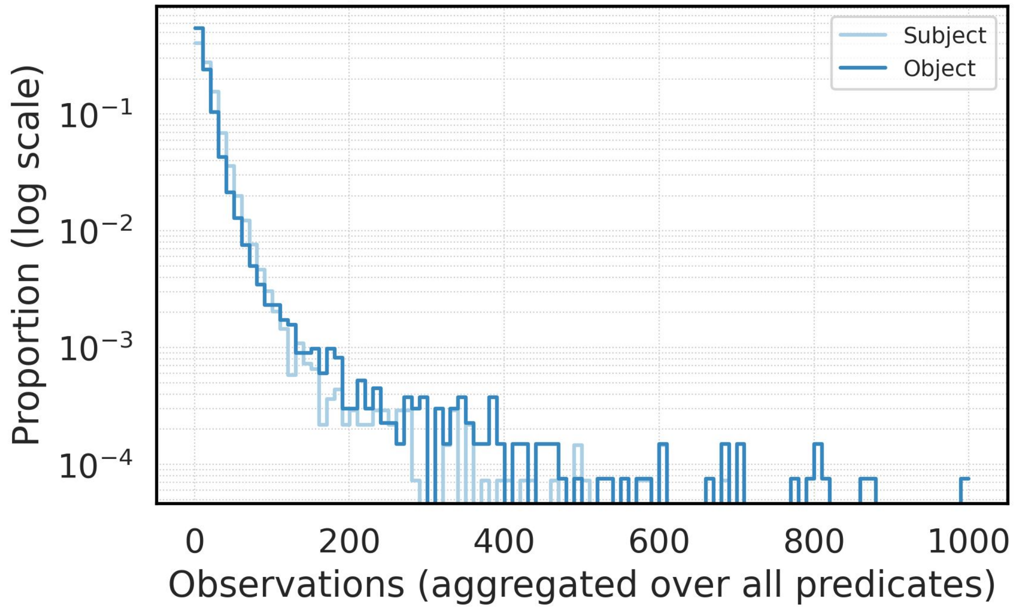


Has Structure

Recommender - MovieLens 100k



Reasoning - FB15k-237

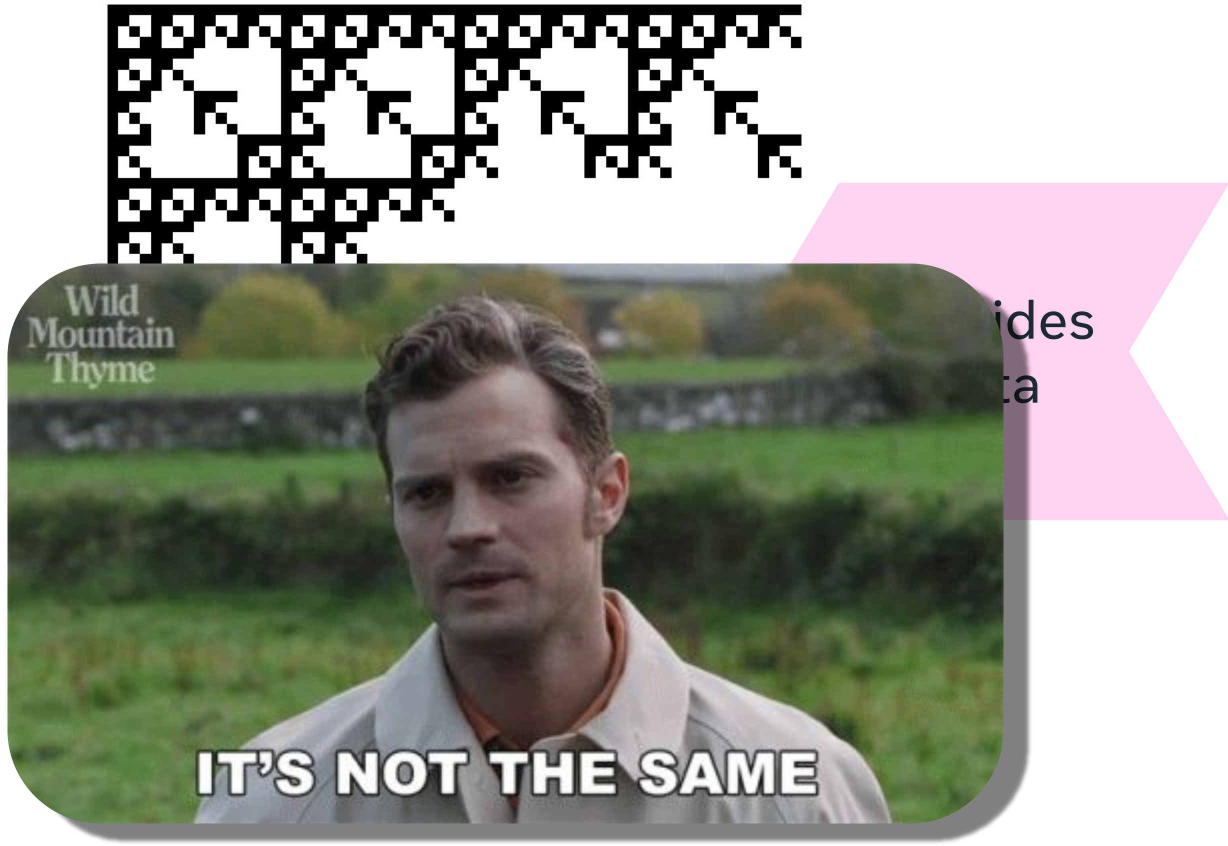


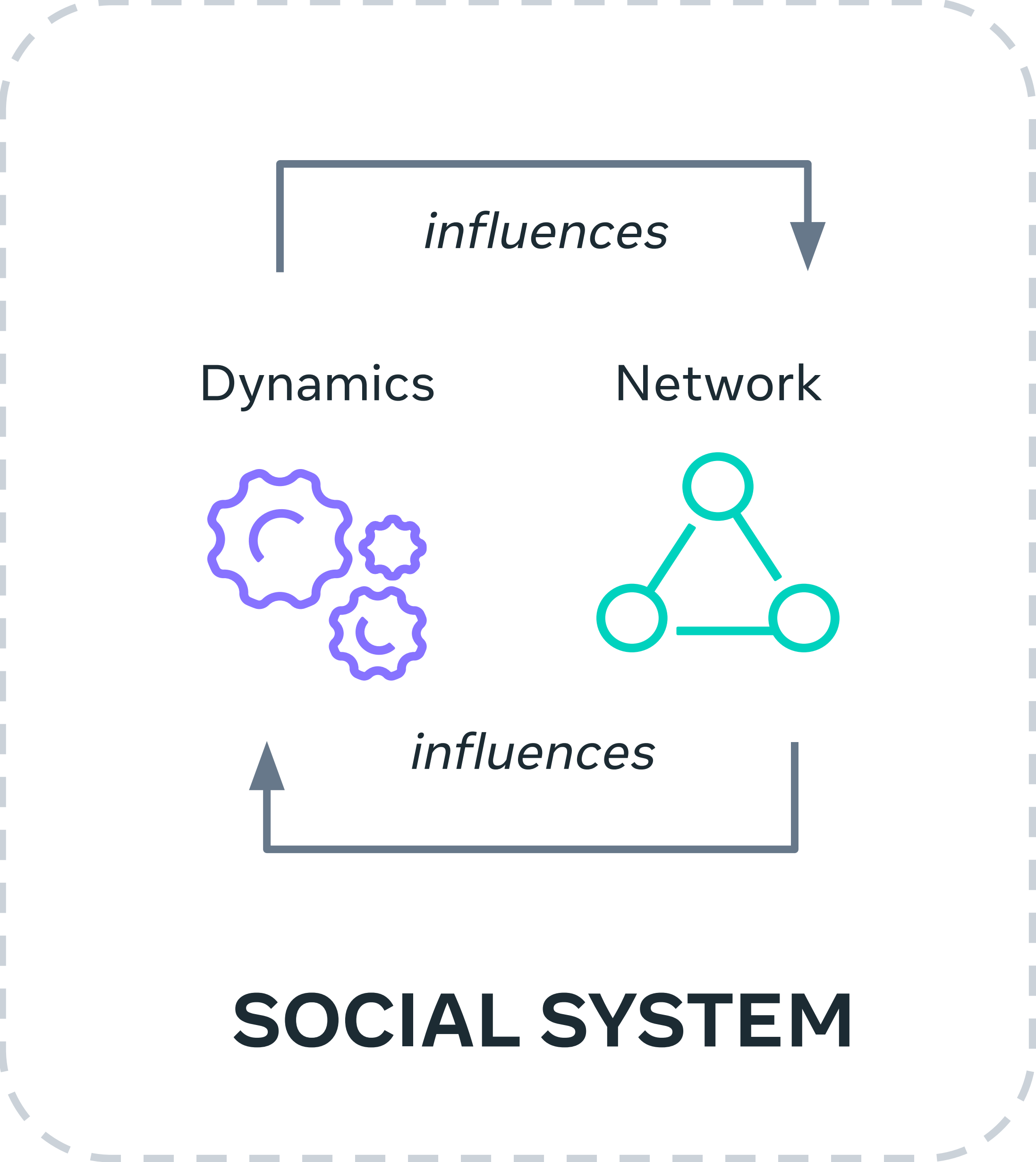
Modeling **passive data collection**

Formalize data collection via **sample graphs**

Edge in a sample graph denotes an observed data point (noise free)

Sample graph




influences

Dynamics

Network

influences

SOCIAL SYSTEM

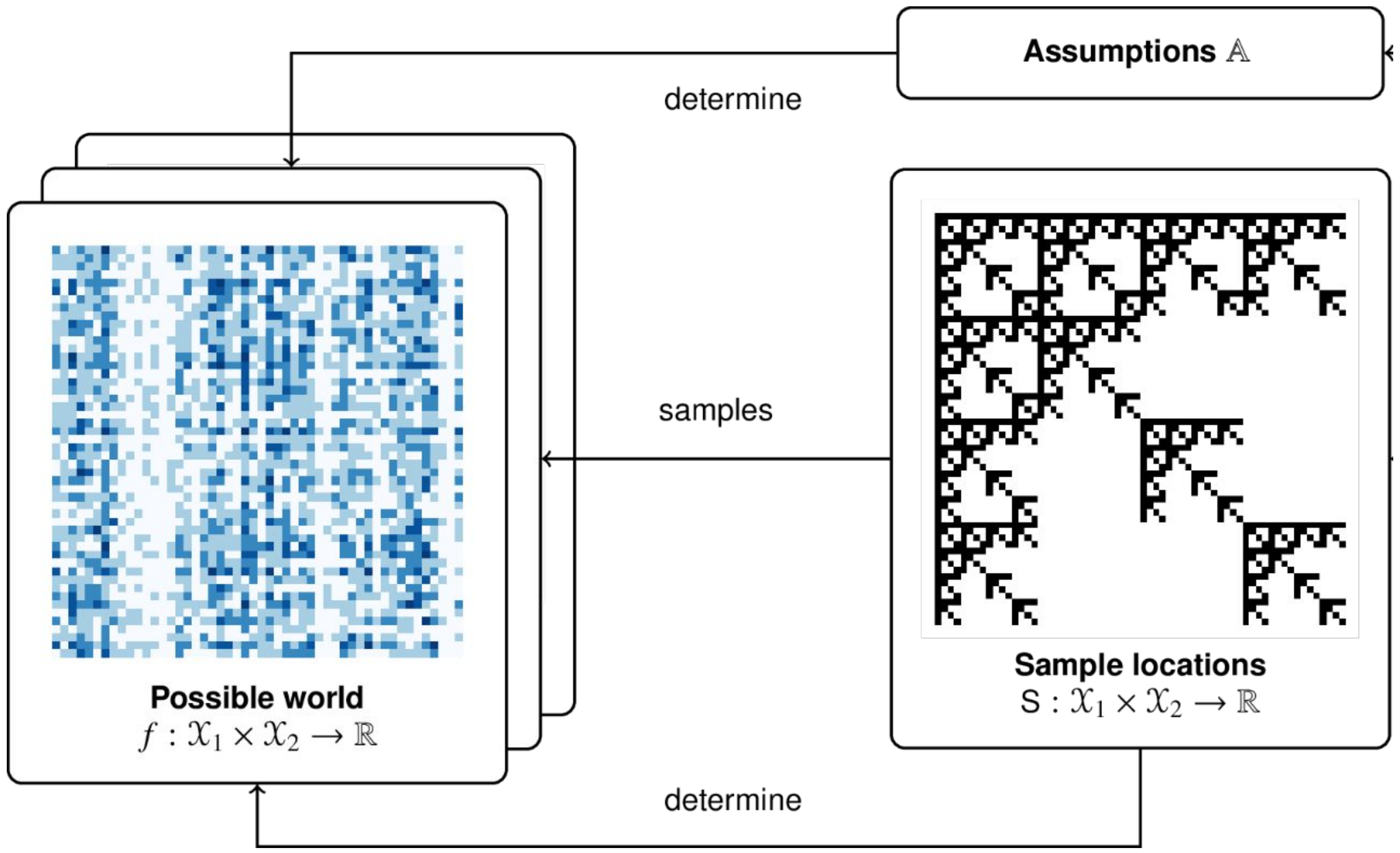
	Domain $\mathcal{X}$	**Possible world** f	**Sample distribution** S	**Target distribution** T
Recommender systems	$\mathcal{U} \times \mathcal{I}$	User preferences	Probability of user interacting with item, heavy-tailed in $\mathcal{U}$ and $\mathcal{I}$	Uniform, $p_T(u, i) = 1/\lvert\mathcal{U} \times \mathcal{I}\rvert$
Symbolic reasoning	$\mathcal{S} \times \mathcal{P} \times \mathcal{O}$	Truth value of factoids	Probability of observing factoid, heavy-tailed in $\mathcal{S}$, $\mathcal{P}$, and $\mathcal{O}$	Uniform, $p_T(s, p, o) = 1/\lvert\mathcal{S} \times \mathcal{P} \times \mathcal{O}\rvert$

Possible Worlds Semantics

We want to evaluate how well an

- estimated model *h*
- approximates the true world *f*

Observations + assumptions define possible worlds that are consistent with both.

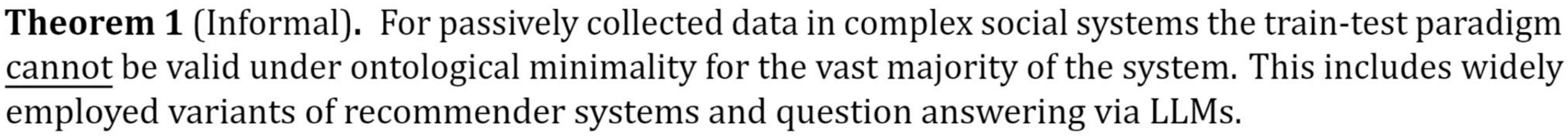

Assumptions $\mathbb{A}$
determine
samples
determine
Possible world
$f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$
Sample locations
$\mathsf{S} : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$

$$\mathbb{P}_{f\sim\mathsf{F}}\left(|\theta - L^{\mathsf{T}}_{fh}\,| \leq \epsilon\right) \geq 1 - \delta$$

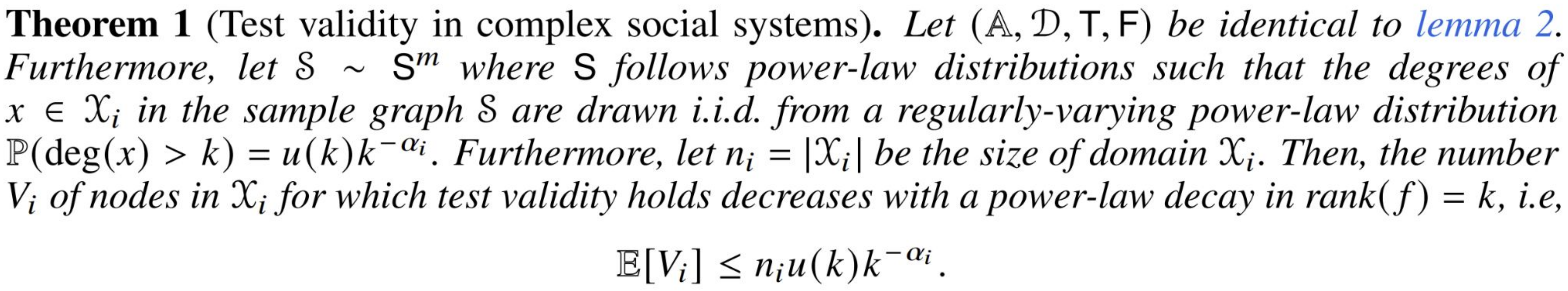
Test validity: can we bound the error of of *a risk estimator* **θ** compared to the true risk over all possible worlds?

A rigorous impossibility result
No Free Delivery Service

(Nickel, 2024)



Theorem 1 (Informal). For passively collected data in complex social systems the train-test paradigm cannot be valid under ontological minimality for the vast majority of the system. This includes widely employed variants of recommender systems and question answering via LLMs.



Theorem 1 (Test validity in complex social systems). *Let* $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ *be identical to lemma 2. Furthermore, let* $\mathcal{S} \sim \mathsf{S}^m$ *where* S *follows power-law distributions such that the degrees of* $x \in \mathcal{X}_i$ *in the sample graph* $\mathcal{S}$ *are drawn i.i.d. from a regularly-varying power-law distribution* $\mathbb{P}(\deg(x) > k) = u(k)k^{-\alpha_i}$. *Furthermore, let* $n_i = |\mathcal{X}_i|$ *be the size of domain* $\mathcal{X}_i$. *Then, the number* V_i *of nodes in* $\mathcal{X}_i$ *for which test validity holds decreases with a power-law decay in* $\operatorname{rank}(f) = k$, *i.e,*

$$\mathbb{E}[V_i] \leq n_i u(k) k^{-\alpha_i}.$$

Proof Sketch

Assumptions

$$\mathbb{A} = \{f \mid \mathrm{rank}(f) \leq k\}$$

Test validity

$$\mathbb{P}_{f\sim\mathsf{F}}(|\theta - L^{\mathsf{T}}_{fh}| \leq \epsilon) \geq 1 - \delta$$

$$\leqq$$

Necessary conditions

$$\mathbb{P}_{f\sim\mathsf{F}}(L^{\mathsf{T}}_{fh} \leq \epsilon + \theta) \geq 1 - \delta$$

$$\leqq$$

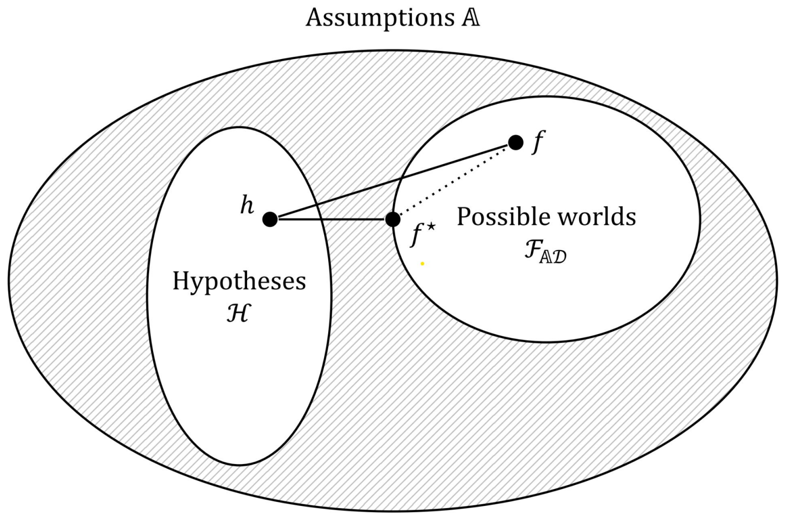
$$\mathbb{P}_{f\sim\mathsf{F}}(L^{\mathsf{T}}_{ff^\star} \leq \epsilon + \theta) \geq 1 - \delta$$

⚡

Grounding in

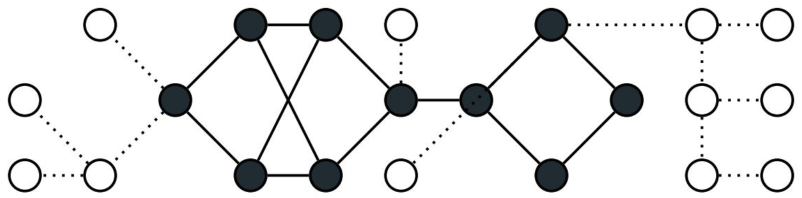
- $\mathscr{D}$ ~ complex social system
- Ontological parsimony

$$\nexists \epsilon : \mathbb{P}_{f\sim\mathsf{U}}(L_{ff^\star} \leq \epsilon) > 0$$



$$f^\star = \arg\min_f L^{\mathsf{T}}_{fh} : L^{\mathsf{T}}_{ff^\star} \leq L^{\mathsf{T}}_{fh}$$

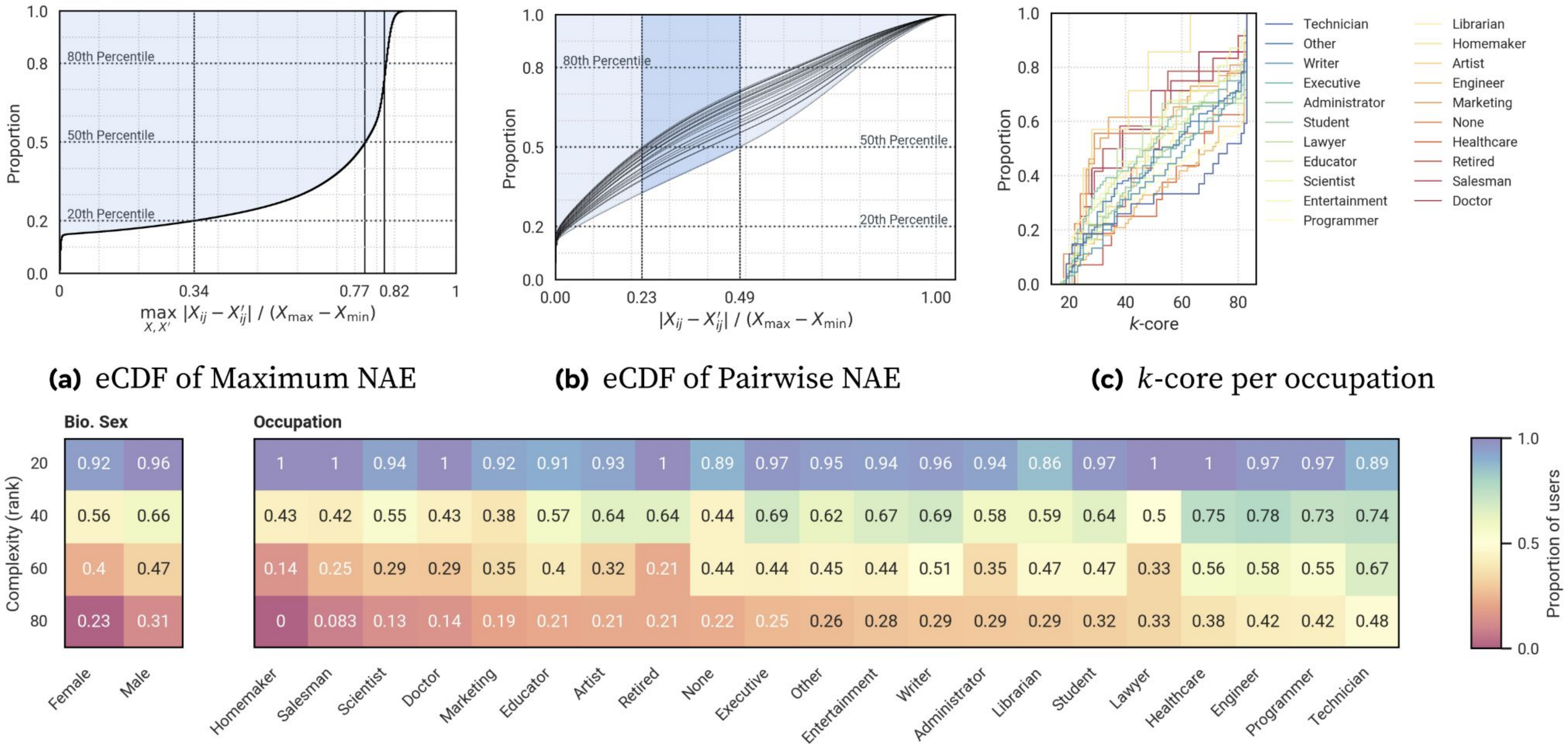
$\mathcal{F}_{\mathbb{A}\mathcal{D}}$ is a vector space if *k*-connectivity of $\mathscr{D}$ is smaller than *rank(f)*



MovieLens (100k)

THE recommender systems benchmark since 1998

Yet, it’s invalid for the evaluation tasks that we (typically) use it for… 😬



(a) eCDF of Maximum NAE

(b) eCDF of Pairwise NAE

(c) *k*-core per occupation

Bio. Sex

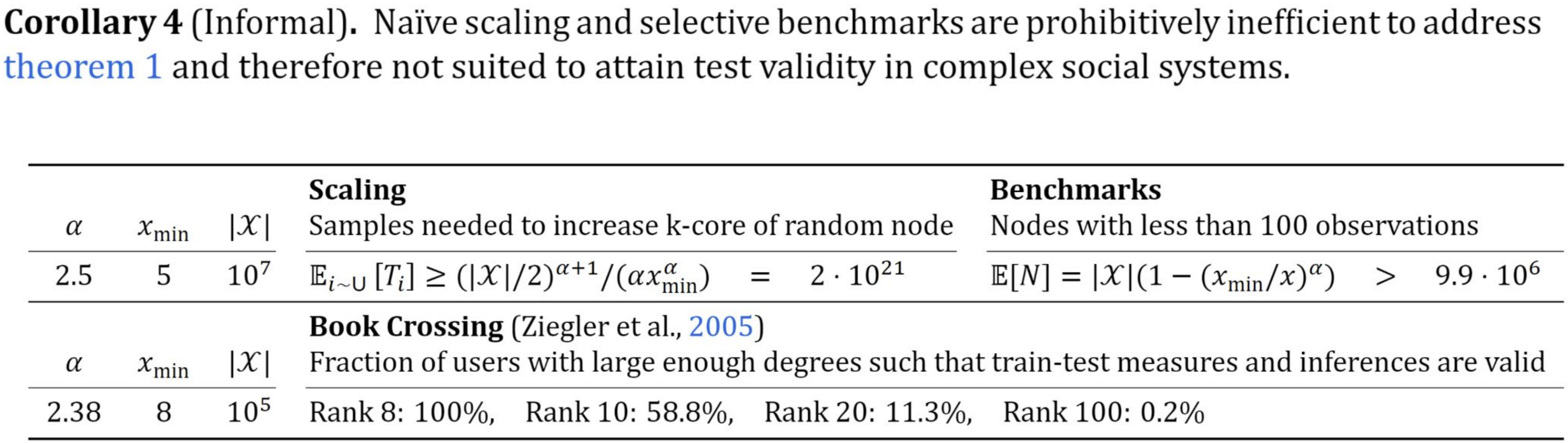
Complexity (rank)	Female	Male
20	0.92	0.96
40	0.56	0.66
60	0.4	0.47
80	0.23	0.31

Occupation

Complexity (rank)	Homemaker	Salesman	Scientist	Doctor	Marketing	Educator	Artist	Retired	None	Executive	Other	Entertainment	Writer	Administrator	Librarian	Student	Lawyer	Healthcare	Engineer	Programmer	Technician
20	1	1	0.94	1	0.92	0.91	0.93	1	0.89	0.97	0.95	0.94	0.96	0.94	0.86	0.97	1	1	0.97	0.97	0.89
40	0.43	0.42	0.55	0.43	0.38	0.57	0.64	0.64	0.44	0.69	0.62	0.67	0.69	0.58	0.59	0.64	0.5	0.75	0.78	0.73	0.74
60	0.14	0.25	0.29	0.29	0.35	0.4	0.32	0.21	0.44	0.44	0.45	0.44	0.51	0.35	0.47	0.47	0.33	0.56	0.58	0.55	0.67
80	0	0.083	0.13	0.14	0.19	0.21	0.21	0.21	0.22	0.25	0.26	0.28	0.29	0.29	0.29	0.32	0.33	0.38	0.42	0.42	0.48

(d) Test-validity per demographic group and model complexity

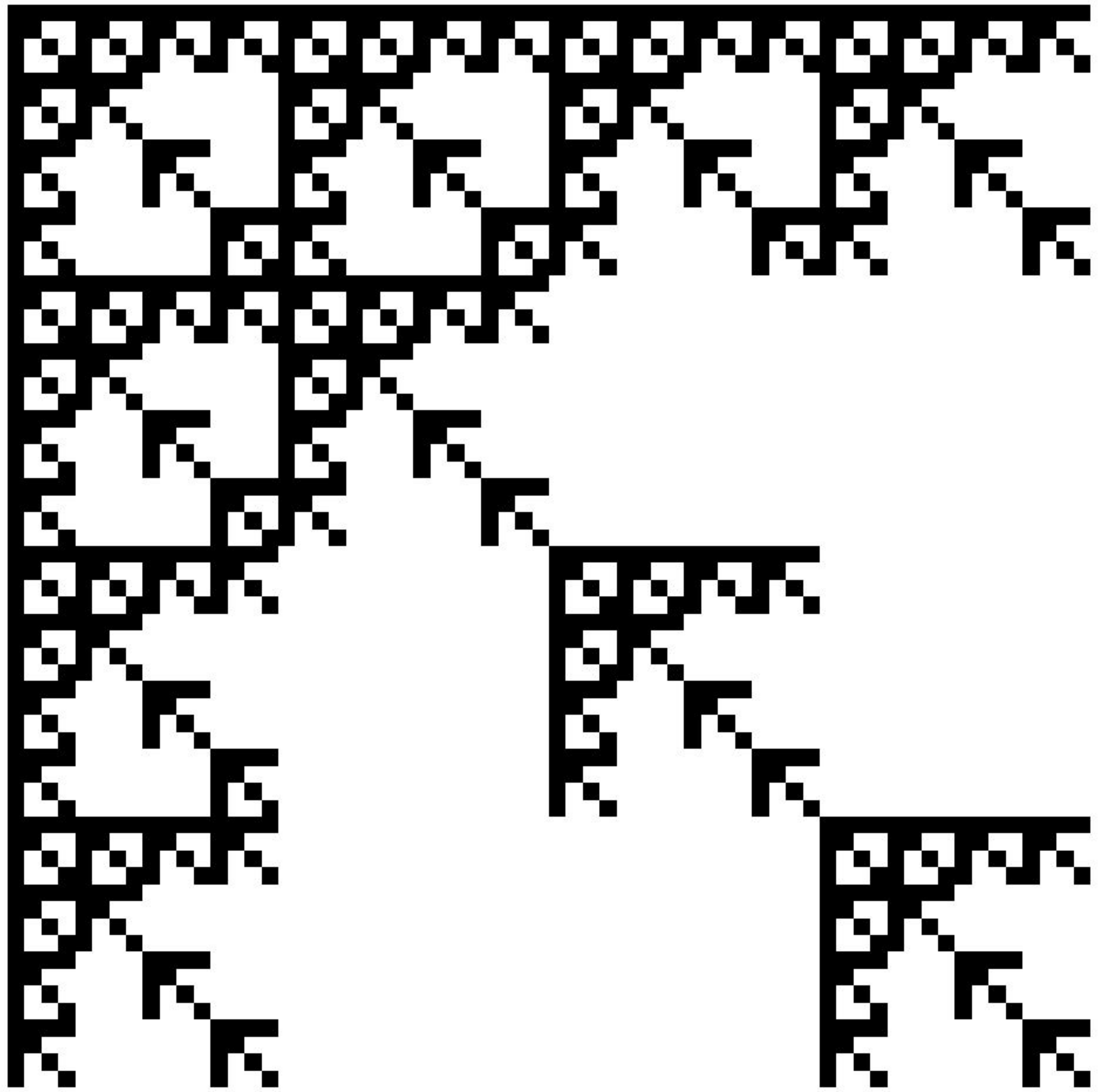
Naive scaling and **manual benchmarks** won’t fix it…



Corollary 4 (Informal)**.** Naïve scaling and selective benchmarks are prohibitively inefficient to address theorem 1 and therefore not suited to attain test validity in complex social systems.

			Scaling	**Benchmarks**
α	x_{min}	$\lvert\mathcal{X}\rvert$	Samples needed to increase k-core of random node	Nodes with less than 100 observations
2.5	5	10^7	$\mathbb{E}_{i\sim\mathrm{U}}\left[T_i\right] \geq (\lvert\mathcal{X}\rvert/2)^{\alpha+1}/(\alpha x_{\mathrm{min}}^{\alpha}) \quad = \quad 2\cdot10^{21}$	$\mathbb{E}[N] = \lvert\mathcal{X}\rvert(1-(x_{\mathrm{min}}/x)^{\alpha}) \quad > \quad 9.9\cdot10^{6}$
			Book Crossing (Ziegler et al., 2005)	
α	x_{min}	$\lvert\mathcal{X}\rvert$	Fraction of users with large enough degrees such that train-test measures and inferences are valid	
2.38	8	10^5	Rank 8: 100%, Rank 10: 58.8%, Rank 20: 11.3%, Rank 100: 0.2%	

No need for an existential crisis 😱 advances are real **but realism is needed**

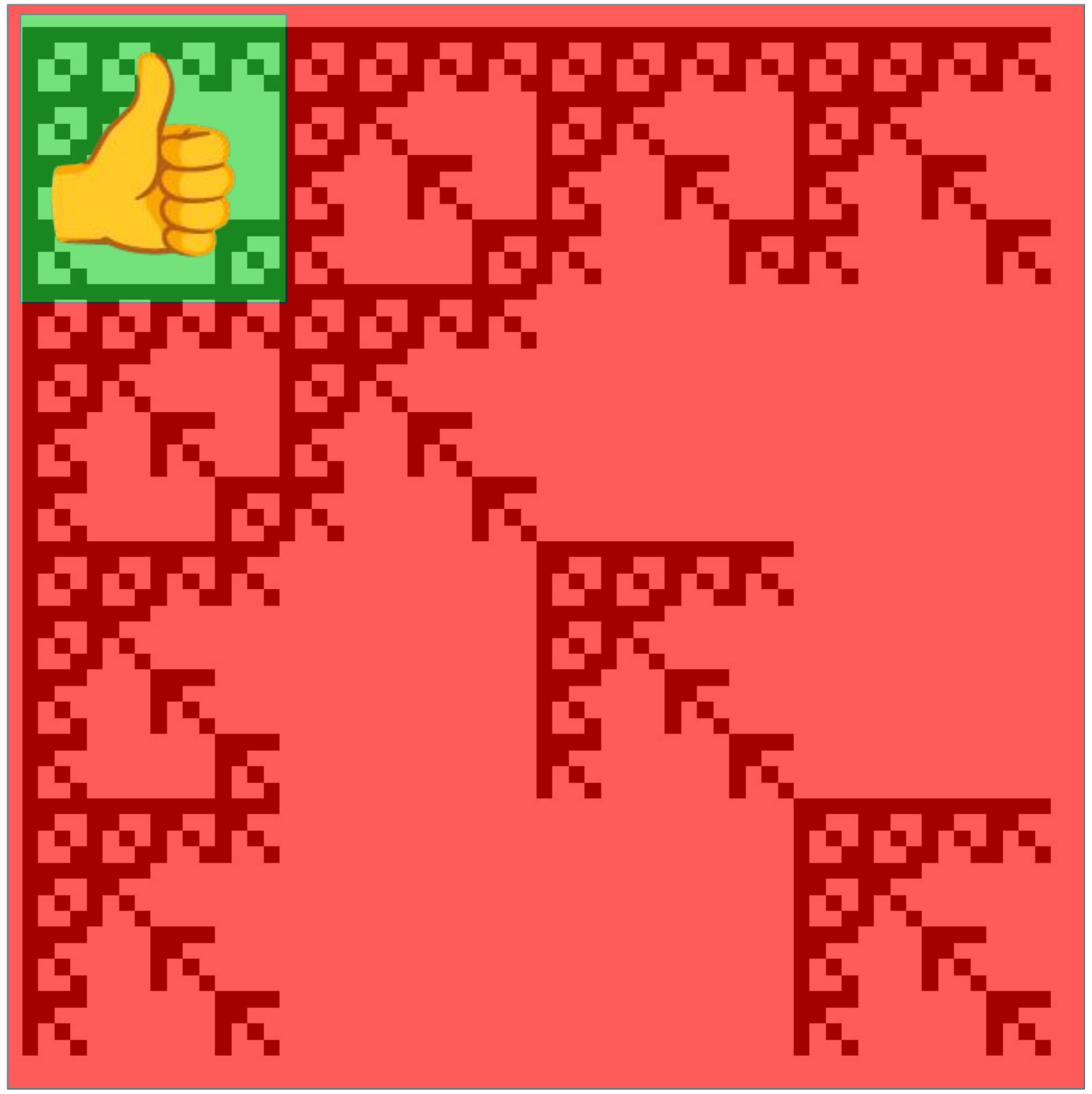


Sample Graph

👍 **Test Valid**
k-core(Sample Graph) ≥ Complexity of world

👎 **Test Invalid**
k-core(Sample Graph) < Complexity of world

No need for an existential crisis 😱 advances are real **but realism is needed**



Sample Graph

Test Valid
k-core(Sample Graph) ≥ Complexity of world

Test Invalid
k-core(Sample Graph) < Complexity of world

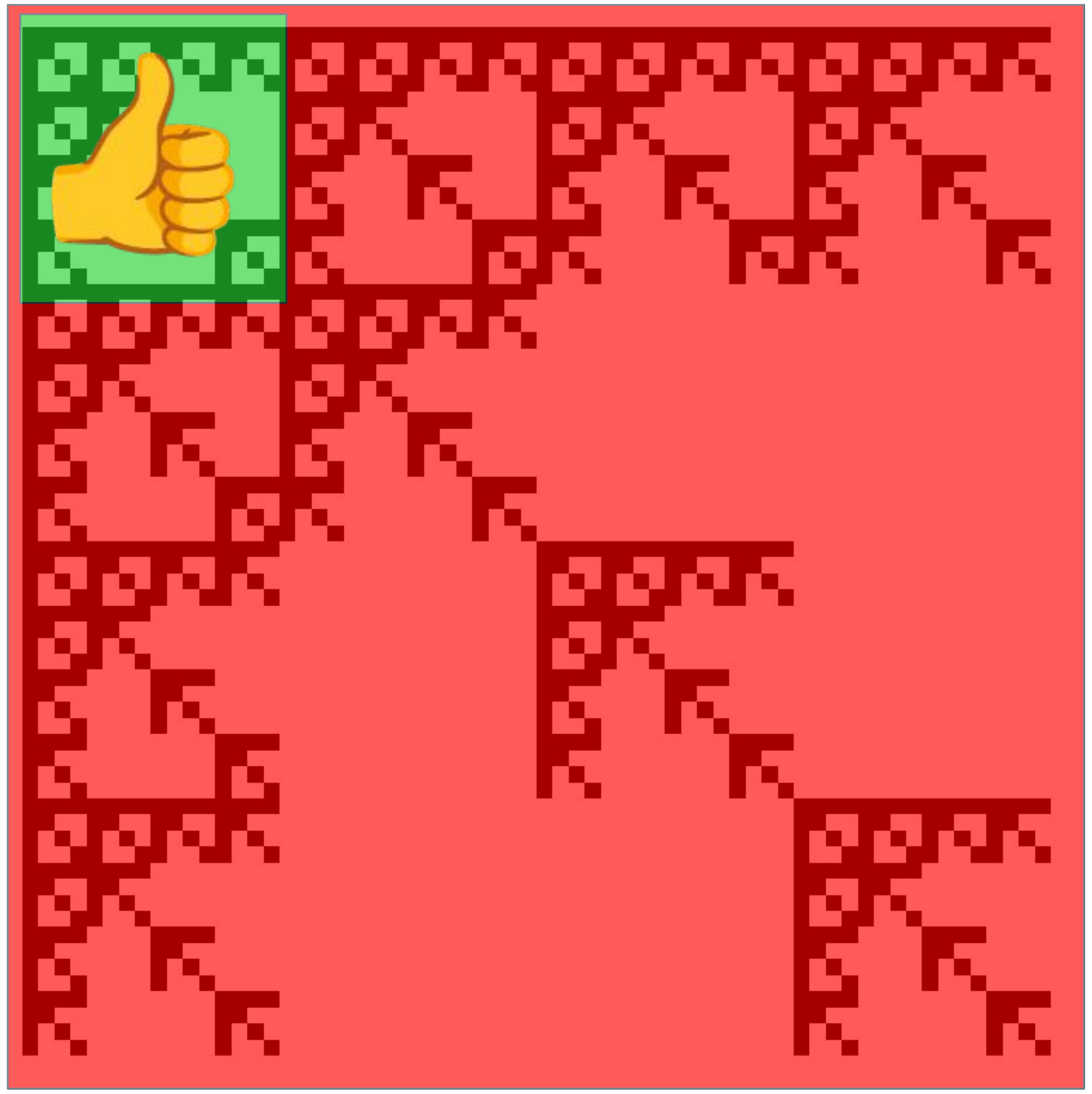
Hume is back…
back again!

Hume's **problem of induction** is now back in a slightly different form and **renders the train-test paradigm ineffective** for our **current data collection practices.**

Solving many complex AI tasks will not come for free through scaling or for cheap through extrapolating from small-scale benchmarks.

There is an **inherent trade-off between data quality, quantity, and task complexity.** If we want to avoid asking AI systems to solve simpler tasks (e.g., non-out-of-distribution or smaller scale), **new data curation efforts** are needed.

Future work
Provably fair cooperative data collection



Sample Graph

Test Valid
k-core(Sample Graph) ≥ Complexity of world

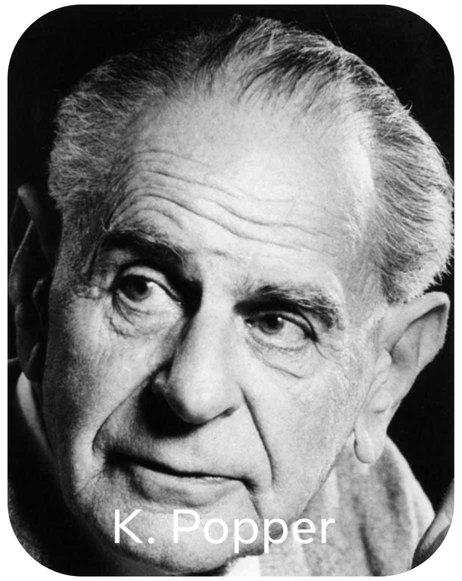
- Increase size of the green area.
- We know where to collect data via k-core condition!
- Number of test-valid data points is a **supermodular** function with regard to datasets
- Shapley value!
- Scalable at the level of organizations, e.g., in **open science**

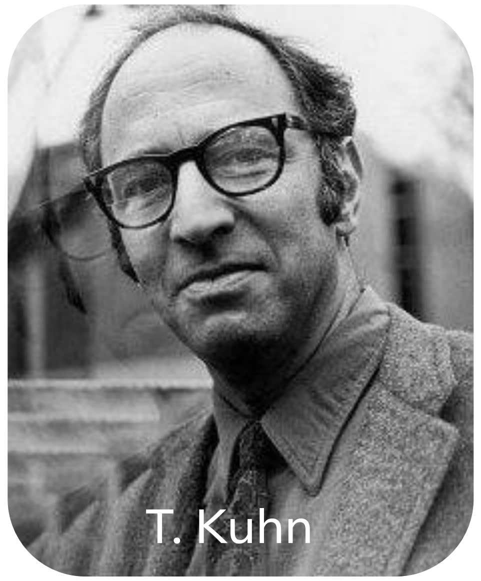
OUR MISSION

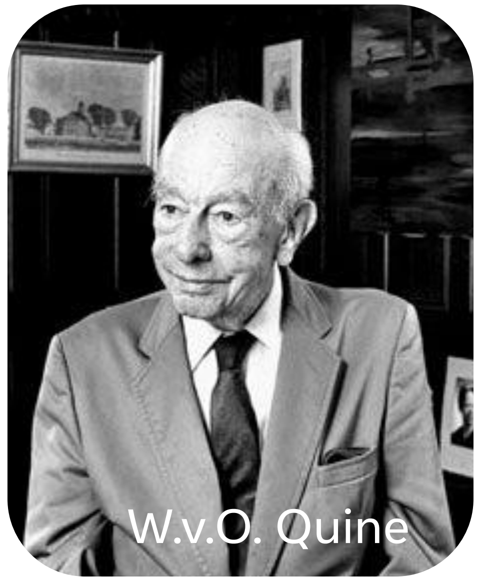
AI

~~Statistics~~ is the **science of induction**

Machine learning


K. Popper


T. Kuhn


W.v.O. Quine

Science

Trying to provide a definite answer to what science is, is a good way to get your philosopher friend upset.

The ultimate issue is “**how to determine which beliefs are epistemically [justified]**” *(Fuller 1985)*.

Demarcation

Demarcation

Alchemy, Phrenology, Astrology, … Us?

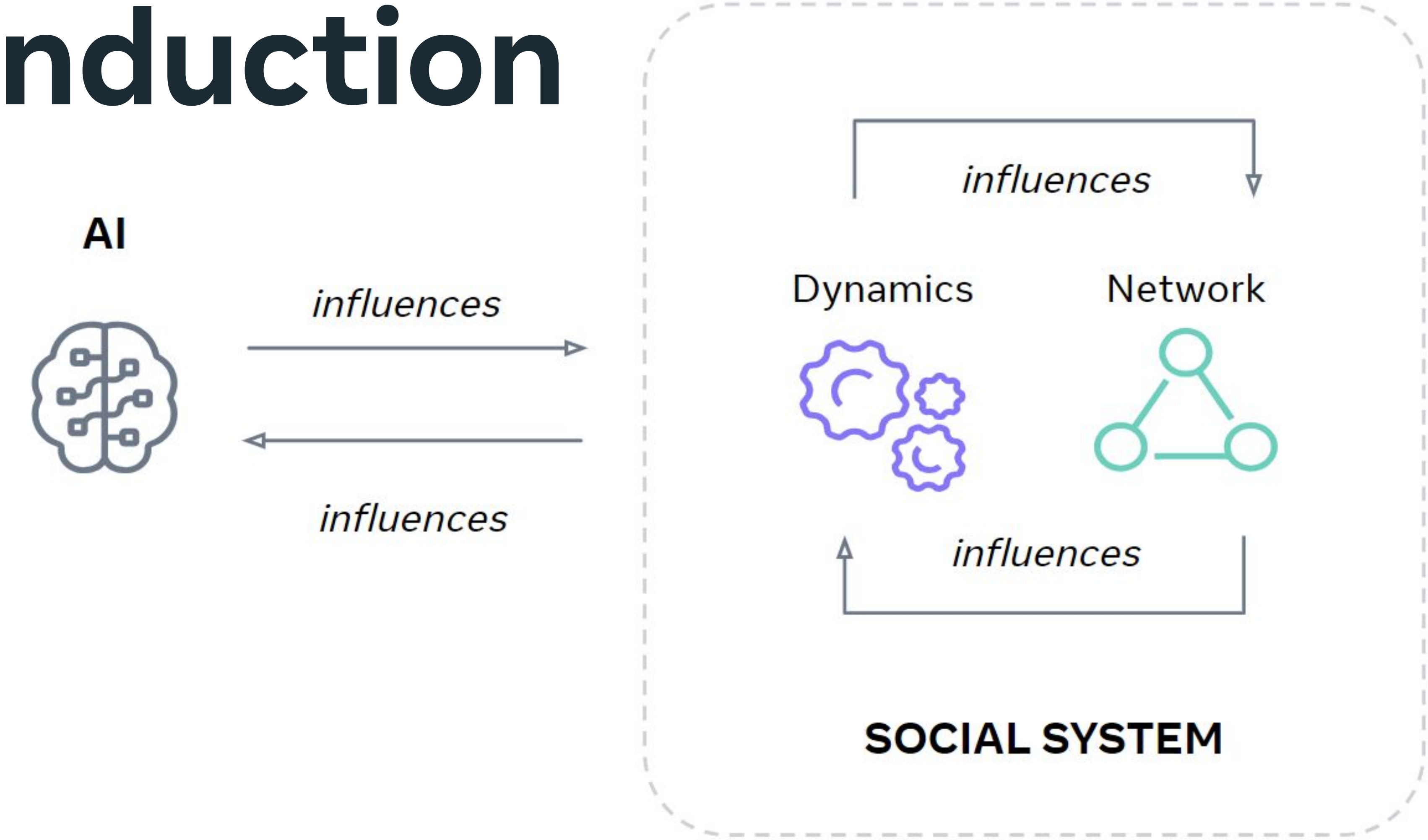
Pseudo-Science

What exactly constitutes pseudo-science is not clear either, but roughly, it amounts to *(Hansson 1996):*

1) it is **not scientific**, and
2) its major proponents try to **create the impression that it is scientific**.

OUR MISSION

Towards a **science of induction** in social systems



No Free Delivery Service

To appear at NeurIPS’24

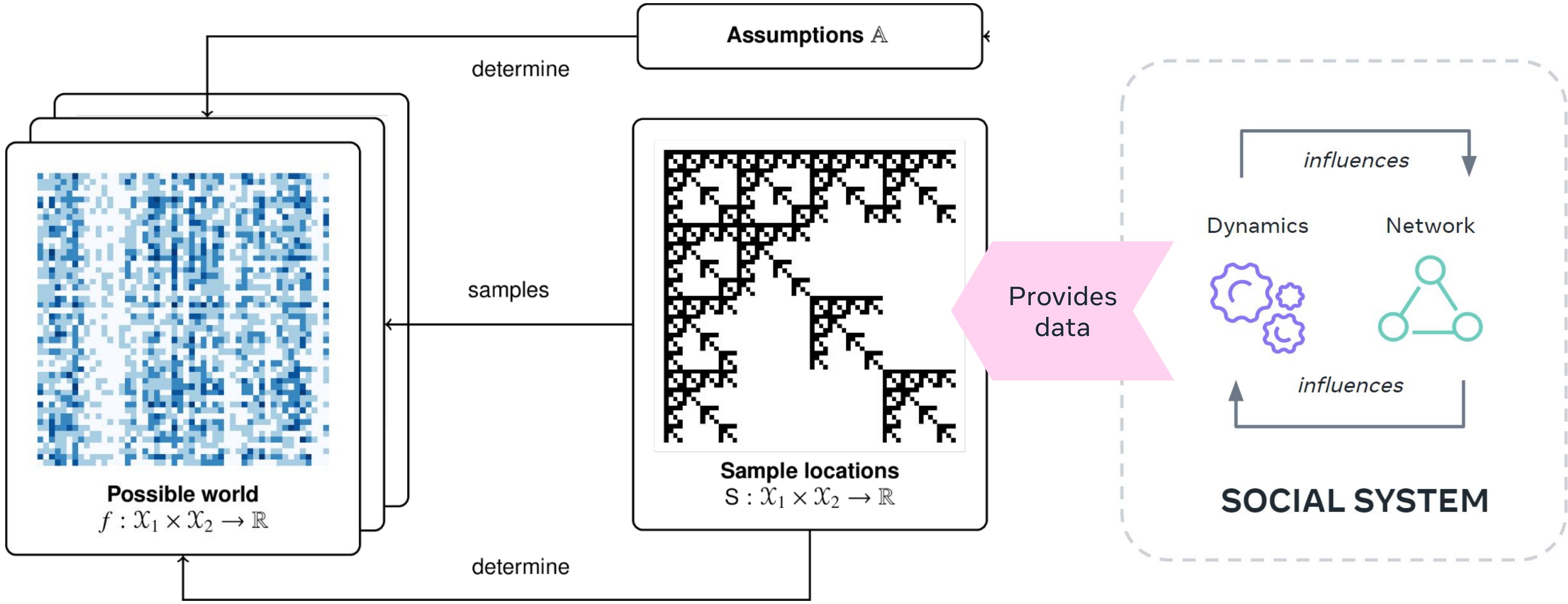
https://arxiv.org/abs/2411.13653

Theorem 1 (Informal) For passively collected data in complex social systems, the train-test paradigm cannot be valid under ontological parsimony for the vast majority of the system. This incluses widely employed variants of recommender systems and QA via LLMs.

How it works:

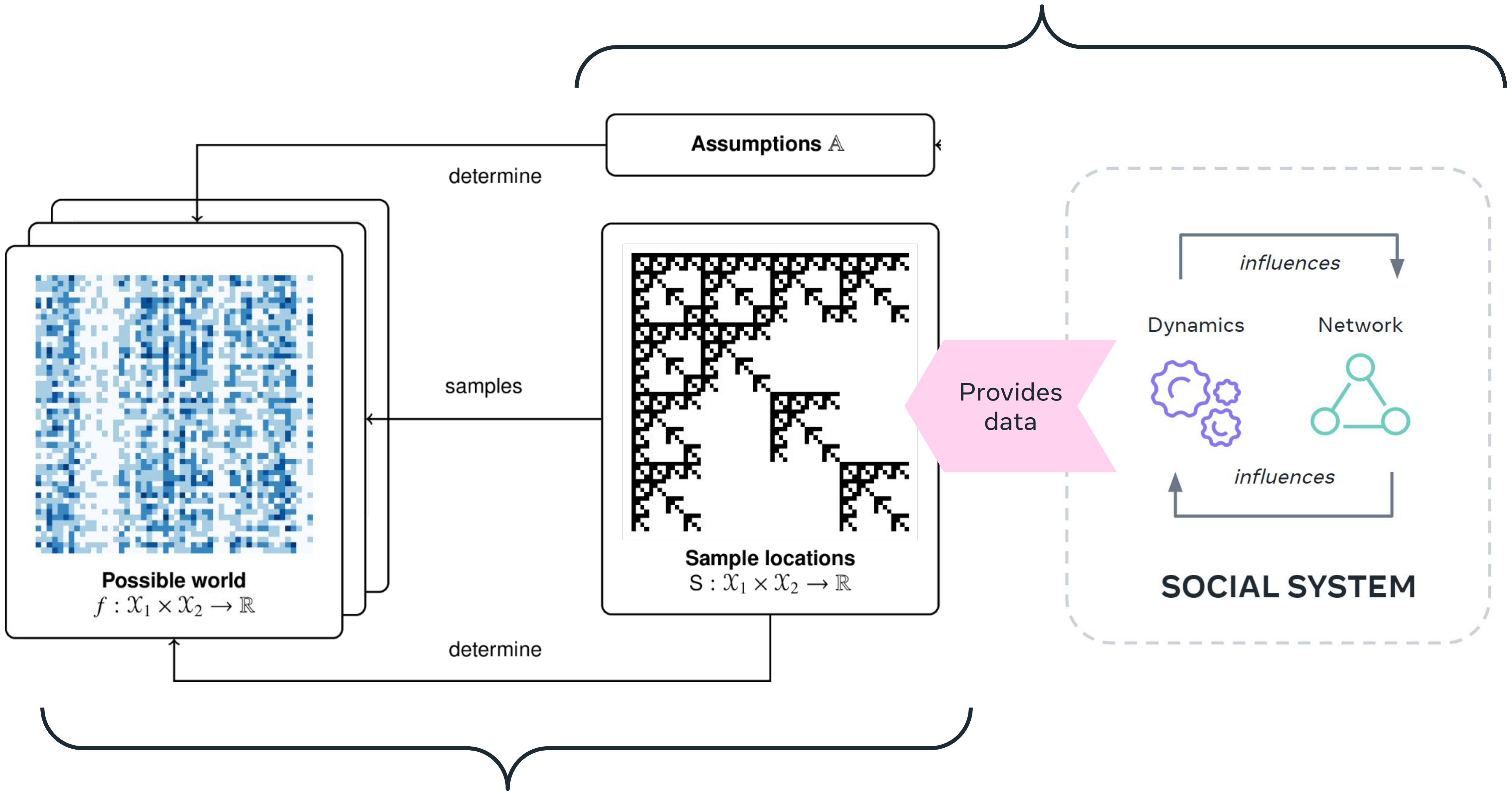
- Observations + assumptions define **possible worlds** that are consistent with both.
- The error of of *any risk estimator* $\boldsymbol{\theta}$ cannot be bounded whp over these possible worlds due to the structure of the data generating system.

$$\mathbb{P}_{f\sim \mathsf{F}}\left(|\theta - \boldsymbol{L}^{\mathsf{T}}_{fh}\, | \leq \epsilon\right) \geq 1-\delta$$



Meta

Formalization of data collection for validation of AGI tasks



Assumptions $\mathbb{A}$

determine

samples

determine

Possible world
$f : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \mathbb{R}$

Sample locations
$\mathrm{S} : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \mathbb{R}$

Provides data

influences

Dynamics

Network

influences

SOCIAL SYSTEM

Test validity: Can the test error be informative about the true generalization error?

$$\mathbb{P}_{f\sim \mathrm{F}}\left(|\theta - L^{\mathrm{T}}_{fh}\ | \leq \epsilon\right) \geq 1 - \delta$$