

THE NATURE AND SCALING OF SCIENTIFIC DISCOVERY

LECTURE 01

# What Is a Scientific Discovery?

What exactly must be found, and what would warrant believing it?

Complexity, Information, and Cross-Domain Transfer

Scholarly lectures · September 2026

# A field at the intersection of four questions

| Tradition                        | Question brought into discovery                    |
|----------------------------------|----------------------------------------------------|
| Learning and information         | What can observations distinguish?                 |
| Algorithms and optimization      | What can be found with available resources?        |
| Causal and experimental science  | What makes a claim warranted?                      |
| Science of science and economics | How do knowledge and institutions change progress? |

The series connects these traditions; it does not claim a new scientific discipline or a universal law.

# When an unreliable measurement becomes the question

---

Luria and Delbrück initially sought an accurate proportion of resistant bacteria. Large, irreproducible fluctuations became evidence to explain.

The question changed: not only “How many are resistant?” but “What growth history could produce this distribution?” Culture lineage became part of the model.

# From an unstable estimate to a new research direction

---

Suppose repeated cultures give very different resistant fractions. Better measurement is a sensible first response—but only if measurement is the source of variation.

If one early heritable event leaves many descendants, the count can record growth history. That possibility makes the distribution, rather than only its mean, worth investigating.

# Which explanation deserves the next experiment?

| Plausible source             | Comparison that could distinguish it               |
|------------------------------|----------------------------------------------------|
| Plating / counting variation | Repeated samples from one grown culture            |
| Unequal growth or exposure   | Comparable growth, size and plating conditions     |
| Shared lineage history       | Separately grown cultures versus one split culture |

The alternatives concern different sources of variation. They call for different comparisons, not just more precise averages.

# Novelty, validity and significance

---

Novelty compares an output with a dated body of knowledge. Validity concerns whether the evidence supports the claim. Significance concerns which questions or capabilities the result changes.

A novel but invalid claim is not an established discovery. A valid rediscovery may still be useful evidence about a method.

# A complete discovery episode: recovering a program

---

$$f(x)=ax^2+bx+c; \quad f(0)=0, f(1)=1, f(2)=4 \Rightarrow f(x)=x^2$$

Target: recover a polynomial of degree at most two. Available queries return its exact value at a chosen real input; coefficients are unknown.

The degree bound is prior knowledge. Three distinct queries identify the coefficients; noisy values or an unrestricted function class define different problems.

# A program can fit examples and still be wrong

---

$$f(x)=x^2; \quad g(x)=x^2+x(x-1)(x-2)$$

Suppose observed inputs are 0, 1 and 2, with outputs 0, 1 and 4. Both  $x^2$  and  $x^2+x(x-1)(x-2)$  fit every observation.

At  $x=3$  the programs return 9 and 15. One new discriminating input separates them.

# A common problem description

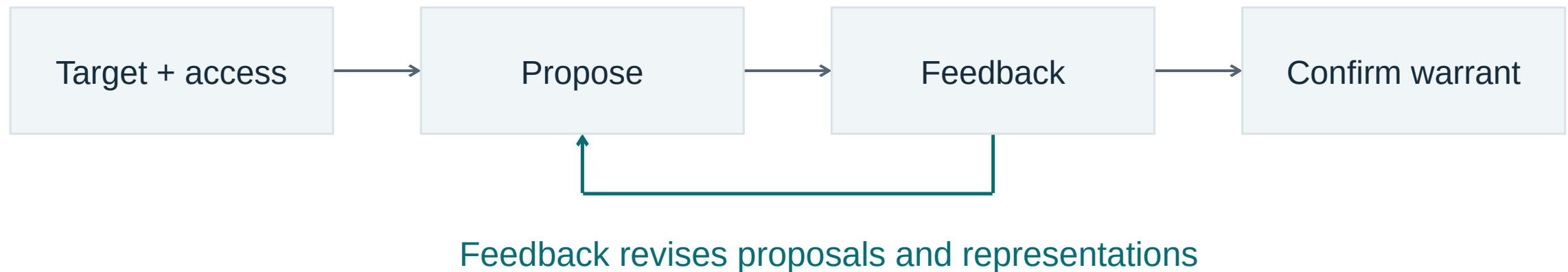
---

$$P = (X, E, Y, K, A, b)$$

Let  $X$  be candidate objects,  $E$  available experiments,  $Y$  observable outcomes and  $K$  prior knowledge. An acceptance rule  $A$  maps a candidate and evidence to a decision.

A resource vector  $b$  records compute, observations, experimental rounds, elapsed time and human effort.

# One process, four different warrants



The acceptance rule determines which observations count as evidence. Feedback can change the next proposal or its representation. Confirmation establishes the stated endpoint. An exact verifier can serve both feedback and confirmation roles.

# A construction with an exact certificate

---

$$S = \{0,1\}^n \subset \mathbb{F}_3^n; \quad |S| = 2^n$$

In one dimension,  $\{0,1\}$  is a cap set over arithmetic modulo 3: it contains no three distinct elements. Cartesian products give  $\{0,1\}^n$ , containing  $2^n$  points.

If three binary vectors sum to zero modulo 3 in every coordinate, each coordinate is either all zero or all one; the three vectors are equal.

# Finding an object versus proving optimality

---

$$\exists x \in F: \text{value}(x) = m \neq \forall x \in F: \text{value}(x) \leq m$$

A checker can establish that a particular set of size  $m$  is feasible. This supplies a lower bound on the best achievable size.

An upper bound ruling out every larger set is a different task. Search success and optimality certification have different quantifiers.

# Molecular design has a validation chain

---

A generated molecule can be syntactically valid yet unstable, unsynthesizable or inactive. A favorable predicted property is evidence about a model output.

Experimental activity, selectivity and functional effect are distinct endpoints. Each requires its own measurement and denominator.

# A correct prediction can leave the question open

---

Suppose a model predicts recovery from treatment records. We might improve its accuracy, ask whether treatment causes recovery, or ask how the effect occurs.

The last two questions concern changes to the world. Prediction on the same records cannot choose between explanations that agree on those records.

# Direct action or an intermediate mechanism?

| Candidate        | Outcome rule | With $T=1$ , set $M=0$ |
|------------------|--------------|------------------------|
| Direct action    | $Y=T$        | $Y=1$                  |
| Action through M | $Y=M$        | $Y=0$                  |

Toy variables: treatment  $T$ , intermediate state  $M$ , recovery score  $Y$ . Naturally  $M=T$ . Both models predict  $Y=T$ , even when  $T$  is randomized.

# Evidence for an effect leaves its route unresolved

---

In both toy models, changing  $T$  from 0 to 1 changes  $Y$  from 0 to 1. Evidence for that effect therefore leaves the direct-versus-mediated question open.

Setting  $M=0$  while  $T=1$  separates the predictions. This only identifies the route if the manipulation changes  $M$  without directly changing  $Y$  or other relevant processes.

# The acceptance rule determines the task

| Output       | Acceptance evidence               | Residual limitation                |
|--------------|-----------------------------------|------------------------------------|
| Formal proof | Kernel checks proof term          | Formal statement may miss intent   |
| Program      | Proof or exhaustive bounded tests | Specification may be incomplete    |
| Binder       | Measured binding under protocol   | Function and transfer unresolved   |
| Mechanism    | Discriminating intervention       | Assumptions and context dependence |

Changing the acceptance rule changes what counts as success, even if the candidate generator is identical. A useful comparison names the claim, the validation procedure and what remains uncertain.

# Success is a probability over an episode

---

$$Q(\pi, b; D, K) = \Pr[\text{accepted, valid outcome by budget } b]$$

A policy  $\pi$  chooses actions using its history. Its success probability depends on the task distribution  $D$ , knowledge  $K$  and resource budget  $b$ .

The target is a warranted outcome, not merely a plausible candidate. Different task distributions can reverse a method ranking.

# Prediction accuracy and discovery yield

---

Prediction evaluates outputs on a distribution of examples. Discovery often selects rare candidates from the extreme tail of a score distribution.

A small average prediction error can coexist with poor ranking near the selection threshold. Selection also changes the distribution that is tested.

# A simple yield calculation

---

$$E[Y] = 1000 \times 0.20 \times 0.05 = 10$$

Assume 1,000 independent proposals; 20% pass a screen, and 5% of screened proposals validate. Expected validated yield is 10.

If a new generator doubles the screened fraction but halves conditional validation, expected yield remains 10.

# Entropy needs a probability model

---

$$H(\Theta) = -\sum_i p_i \log_2 p_i$$

For a discrete target  $\Theta$  with probabilities  $p_i$ , entropy measures average surprise in bits.

A fair binary target has  $H=1$  bit; a known target has  $H=0$ . These are beliefs about an explicit target, not a measure of scientific importance.

# Information reduces uncertainty about a target

---

$$I(\Theta;Y) = H(\Theta) - H(\Theta | Y)$$

For a discrete target  $\Theta$ , entropy measures uncertainty under a specified probability distribution. Observing  $Y$  can reduce expected uncertainty by mutual information.

High information gain is useful only if  $\Theta$  represents the scientific distinction that matters.

# Searching and learning overlap

---

An evaluation can simultaneously reject a candidate and inform future proposals.  
Search benefits when the feedback reveals reusable structure.

An uninformative failure may eliminate only one object; a counterexample may eliminate an entire family.

# A discovery claim has a boundary

---

Every claim names a population or formal domain, a comparator, a budget and a success criterion.

A result about bounded integer programs cannot silently become a result about arbitrary programs; an in vitro effect cannot silently become a clinical claim.

# A hierarchy of evidence types

| Label             | What is established                 | Typical failure                      |
|-------------------|-------------------------------------|--------------------------------------|
| Theorem           | Conclusion under stated assumptions | Applying it outside its model        |
| Model             | Conditional consequences            | Treating assumptions as observations |
| Empirical finding | Pattern in a studied setting        | Extrapolating beyond support         |
| Conjecture        | A proposed relationship             | Reporting it as settled              |

Theorems constrain every object covered by their assumptions. Models simplify a process to make consequences calculable. Empirical findings describe measured populations. Conjectures propose unestablished relationships.

# No free lunch identifies the role of structure

---

Optimization advantages require assumptions about which objective functions are likely. Uniform averaging over all finite functions removes exploitable regularity.

Science supplies nonuniform structure: mechanisms, symmetries, compositionality and accumulated knowledge. The substantive question is whether the chosen bias matches the task.

# Knowledge changes the problem presented to search

---

A theorem library, a program primitive or an assay protocol can remove work from the next discovery episode.

Knowledge can also change the question, the variables and the available observations. A gain may come from reformulating the task, not only searching it faster.

# The limits of the formalization

---

The tuple  $P$  omits many institutional and historical determinants: incentives, communication, tacit knowledge and which questions researchers choose.

These omissions matter especially when moving from a discovery episode to aggregate scientific progress.

# Conclusion: a discovery requires a warrant

---

A discovery claim joins an object or explanation to evidence, a reference knowledge state and a scope.

Next: separate search difficulty, information requirements and nonidentifiability before asking what more resources can buy.

# References

---

Judea Pearl and Elias Bareinboim (2014). External Validity: From Do-Calculus to Transportability Across Populations. Statistical Science 29, 579–595; journal article; arXiv posted 2015

Nicholas Bloom, Charles I. Jones, John Van Reenen and Michael Webb (2020). Are Ideas Getting Harder to Find?. American Economic Review 110, 1104–1144; journal article

Claude E. Shannon (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423, 623–656; journal article

S. E. Luria and M. Delbrück (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28(6), 491–511; published primary research

# References · continued

---

Sumit Gulwani, Oleksandr Polozov and Rishabh Singh (2017). Program Synthesis. Foundations and Trends in Programming Languages 4, 1–119; survey monograph

Bernardino Romera-Paredes et al. (2024). Mathematical discoveries from program search with large language models. Nature 625, 468–475; online 14 December 2023

Stephen A. Cook (1971). The Complexity of Theorem-Proving Procedures. STOC 1971, 151–158; conference paper

Joseph L. Watson et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100; journal article

# References · continued

---

Martin Pacesa et al. (2025). One-shot design of functional protein binders with BindCraft. Nature 646, 483–492; online 27 August 2025

Jonas Peters, Peter Bühlmann and Nicolai Meinshausen (2016). Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society B 78, 947–1012; journal article

Phyllis Illari (2025–26). HPSC0084: Causality and Evidence in Science. UCL course syllabus; some internal headers retain 2024–25

Thomas Hubert et al. (2025). Olympiad-level formal mathematical reasoning with reinforcement learning. Nature; journal article, 2025

# References · continued

---

Constantin Ahlmann-Eltze, Wolfgang Huber and Simon Anders (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods 22, 1657–1661; journal article

Dennis V. Lindley (1956). On a Measure of the Information Provided by an Experiment. Annals of Mathematical Statistics 27, 986–1005; journal article

Burr Settles (2009). Active Learning Literature Survey. University of Wisconsin–Madison Technical Report 1648; survey

David H. Wolpert and William G. Macready (1997). No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1, 67–82; journal article

# References · continued

---

Kevin Ellis et al. (2021). DreamCoder: Bootstrapping Inductive Program Synthesis with Wake-Sleep Library Learning. PLDI 2021; conference paper; preprint has earlier title

Benjamin F. Jones (2009). The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder?. Review of Economic Studies 76, 283–317; journal article

Michael Park, Erin Leahey and Russell J. Funk (2023). Papers and patents are becoming less disruptive over time. Nature 613, 138–144; journal article