

What Makes a Discovery Problem Complex?

When is the barrier computation, evidence, or the impossibility of identification?

Complexity, Information, and Cross-Domain Transfer

When perfect fit still leaves the question unanswered

Imagine unlimited observational records in which two binary variables always agree. Predicting either from the other is easy. Yet their causal direction remains unsettled.

Before buying more data or faster search, ask whether the available observations could ever separate the candidate explanations. This turns a performance problem into a question about information access.

More passive data may add zero identifying information

Model A: X is random and $Y=X$. Model B: Y is random and $X=Y$. With matching Bernoulli marginals, both produce exactly the same observational distribution.

Both models put all probability on $X=Y$. More records can estimate that shared distribution precisely without deciding which variable causes the other.

More data, faster search, or a different observation?

Candidate response	What it changes in this example
More passive records	Precision, but not the identical distributions
Faster model search	Search cost, but not the missing distinction
Intervene on X	Makes the models predict different Y outcomes

Set X to zero externally, replacing its usual cause: model A forces $Y=0$; model B leaves Y random. This intervention changes the evidence available.

Identifiability is about observational equivalence

$$P_{\theta} = P_{\theta'} \implies g(\theta) = g(\theta')$$

A target $g(\theta)$ is identifiable if identical allowed data distributions imply identical target values. Distinct parameters need not be distinguishable when the scientific target is coarser.

Identifiability is a property of model, target and allowed experiments together.

Worst-case time is a family-level statement

$$T_{\text{worst}}(n) = \max \text{ over } |x|=n \text{ of } T(x)$$

For input length n , worst-case time maximizes over all inputs of that length. Average-case time needs an explicit input distribution.

Neither quantity alone predicts how long a particular scientific instance will take. Representation, constants and structural restrictions matter.

Decision, search, counting and optimization

Task	Cap-set formulation	Output
Decision	Does a size- m cap set exist?	Yes / no
Search	Construct one of size m	A set
Counting	How many size- m sets exist?	An integer
Optimization	Find maximum achievable size	Optimum and a witness

A decision procedure answers whether a feasible object exists. Search returns one. Counting returns how many. Optimization returns the best according to an objective.

What P and NP mean

$$x \in L \iff \exists w, |w| \leq \text{poly}(|x|): V(x, w) = 1$$

P contains decision problems solvable in polynomial time. NP contains decision problems whose yes-instances have polynomial-length certificates checkable in polynomial time.

NP does not mean “not polynomial.” Whether P equals NP remains unresolved.

Why a verifier does not supply a finder

A certificate space of length m can contain 2^m candidates. Polynomial-time checking per certificate still permits exponential enumeration.

The cost of a proof checker is measured against the proof and statement lengths; short proofs are not guaranteed for every true statement.

A search-to-decision construction

n variables $\rightarrow$ at most n adaptive decision queries

Given a satisfiable Boolean formula, set the first variable to 0 and ask whether the residual formula is satisfiable. Keep 0 if yes; otherwise keep 1.

Repeat for every variable. At most n oracle calls recover an assignment, but each call assumes a satisfiability decision oracle.

Cap-set checking has a concrete cost

For a proposed set of m vectors of dimension n , examine distinct pairs x, y and test whether $-x-y$ is also in the set.

With a hashed representation, expected checking work is roughly $O(m^2n)$, including vector operations. Searching over subsets is a different resource problem.

A finite-class sample bound

$$n \geq [\ln M + \ln(1/\delta)] / \varepsilon$$

Assume M candidate classifiers, realizability, and independent identically distributed examples. A hypothesis with true error at least ε survives n examples with probability at most $\exp(-n\varepsilon)$.

A union bound over candidates makes the probability of any bad consistent hypothesis at most $M \exp(-n\varepsilon)$.

What the sample bound does and does not say

$$\text{ceil}[(6.931 + 2.996)/0.05] = 199$$

For $M=1,024$, $\epsilon=0.05$ and $\delta=0.05$, the sufficient bound is $n \geq 199$. It controls future error under the same sampling distribution.

It does not prove exact recovery of the true mechanism, tolerate arbitrary label noise, or supply an efficient search algorithm.

Agnostic learning: decompose excess risk

$$R(\hat{h}) - R(h^*) \leq 2 \sup_{h \in H} |R(h) - \hat{R}(h)|$$

Let $R(h) = E[\ell(h, Z)]$ and $\hat{R}(h)$ be the sample mean, with loss in $[0, 1]$. Let $\hat{h}$ minimize $\hat{R}$ and h^* minimize R in H .

For noisy program tests, empirical fit selects $\hat{h}$. One gap concerns $\hat{h}$; the other concerns the best catalogue predictor h^* . Empirical optimality removes the middle term.

Uniform convergence has a capacity price

$$R(\hat{h}) - R(h^*) \leq \sqrt{[2 \log(2|H|/\delta) / n]}$$

For finite H , Hoeffding plus a union bound gives the displayed guarantee with probability at least $1-\delta$.

An infinite class cannot be inserted by writing $\log|H|$. Its capacity must be controlled, for example by VC dimension or covering numbers.

Information gives a different lower bound

$$P_e \geq 1 - [I(H;D)+1] / \log_2 M$$

Let H be uniform over M hypotheses and let D be the complete data. If an estimator has error probability P_e , Fano's inequality bounds how little information can support identification.

The result allows arbitrary computation. It constrains evidence rather than algorithm speed.

A feedback channel limits identification

$$0.1 \geq 1 - (0.2n+1)/10 \Rightarrow n \geq 40$$

Suppose $M=1,024$ equiprobable hypotheses and each permitted observation provides at most 0.2 bits, conditional on previous history.

The information chain rule gives $I(H;D) \leq 0.2n$. Error at most 0.1 requires at least 40 observations by the displayed Fano bound.

One finite family, three different questions

Target among 1,024 hypotheses	Access and assumptions	Conclusion
Predict with error $< .05$	Realizable class, iid labeled data	199 examples suffice at confidence .95
Identify the exact index	Uniform prior; $\leq .2$ bits per observation	Error $\leq .10$ requires at least 40 observations
Separate two mechanisms	Identical laws for all allowed data	Impossible with that access, for any n

These use the same family size but different targets and evidence channels. The numbers are not competing sample-size estimates.

Detecting a small effect

$$\Pr(|\text{mean} - \mu| \geq \varepsilon) \leq 2 \exp(-2n\varepsilon^2)$$

For independent observations in $[0,1]$, Hoeffding bounds sample-mean error. A sufficient sample size for error at most ε with probability $1-\delta$ is $n \geq \ln(2/\delta)/(2\varepsilon^2)$.

Halving ε multiplies this bound by four. Repeating a biased measurement does not remove the bias.

Planted clique separates two kinds of difficulty

$$P_0(\text{any } k\text{-clique}) \leq \text{binom}(m,k) \cdot 2^{-k(k-1)/2}$$

Compare $G(m, 1/2)$ with the same random graph containing a planted k -clique. An exhaustive test asks whether any k -clique exists.

For $k \approx 3 \log_2 m$, a union bound makes a false detection unlikely as m grows. Enumerating all k -subsets is computationally expensive.

The computational gap is a conjecture

At $k=m^{1/3}$, the clique is statistically detectable for sufficiently large m , yet lies below the familiar $\sqrt{m}$ spectral scale.

The planted-clique assumption rules out a randomized polynomial-time detector in an appropriate sub- $\sqrt{m}$ regime. This is an assumption, not $P \neq NP$ proved for this task.

Representation can move the practical boundary

A good representation may expose a low-dimensional parameter, a compositional program or a symmetry. This can reduce both search cost and estimation burden.

If the representation is learned from extensive prior data, that cost belongs in the comparison.

Specification ambiguity in synthesis

The two polynomial programs from Lecture 1 agree on three inputs. No learner can select the intended one from those observations alone without additional bias.

A query at $x=3$ provides discriminating information; a restriction to degree at most two provides an alternative assumption.

Molecular difficulty has multiple sources

A huge sequence space is a search issue; uncertain property measurements are an information issue; unmeasured cellular context can be an identification or transport issue.

Counting sequences alone does not predict the number of designs that must be tested.

The same program task exposes all three barriers

Barrier	Access and target	What changes it
Computation	Fixed finite catalogue; costly evaluation	A faster valid search
Information	Noisy tests; low predictive risk	More informative tests
Identification	Three tests; unrestricted polynomials	Degree bound or richer access

A numerical contrast: computation versus evidence

Case A: one million exact candidates, each checked in one microsecond, gives one second of serial checking before overhead.

Case B: 100 noisy assays, each requiring a day and a previous result, gives 100 days of sequential latency even if proposal generation is free.

What remains unresolved

Worst-case complexity rarely predicts scientific instance difficulty directly. Information bounds often assume a known model family and fixed target.

Finding useful, measurable descriptors that bridge these theories to real discovery episodes remains a research problem.

Conclusion: more resources solve different barriers

More compute, more observations and more informative experiments are different interventions. Their benefits depend on the active bottleneck.

Next: how representations, symmetries and reusable components change the search space without abolishing fundamental limits.

References

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

Leslie G. Valiant (1984). A Theory of the Learnable. Communications of the ACM 27, 1134–1142; journal article

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

Wassily Hoeffding (1963). Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association 58, 13–30; journal article

References · continued

[Tsachy Weissman, Idoia Ochoa and Kedar Tatwawadi \(2016\). EE376A Information Theory lecture notes. Stanford course notes, Tsachy Weissman; teaching reference](#)

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

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

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

References · continued

Quentin Berthet and Philippe Rigollet (2013). Computational Lower Bounds for Sparse PCA. arXiv:1304.0828v2, 26 April 2013; primary preprint as inspected

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

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

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