The Generic Holdout: Preventing False-Discoveries in Adaptive Data Science

Preetum Nakkiran, Jarosław Błasiok

Motivation

Classical science was **non-adaptive**:

1. Scientist fixes a [set of] hypothesis
2. Collects data to test the hypothesis.

Modern science is **adaptive:**

1. Scientist first collects data
2. Explores data to find plausible hypotheses
3. Tests hypotheses on the same data

Naively, "adaptive" science is NOT statistically valid:

- Hypotheses depend on the data used to test them.
- Hypothesis may be "overfit": appears true on data, but actually false.
- *Leads to false-discoveries -- key factor in the "reproducibility crisis" in science.*

Goal: Provide a *statistically-sound* methodology for adaptive science – to prevent scientists from publishing false discoveries.

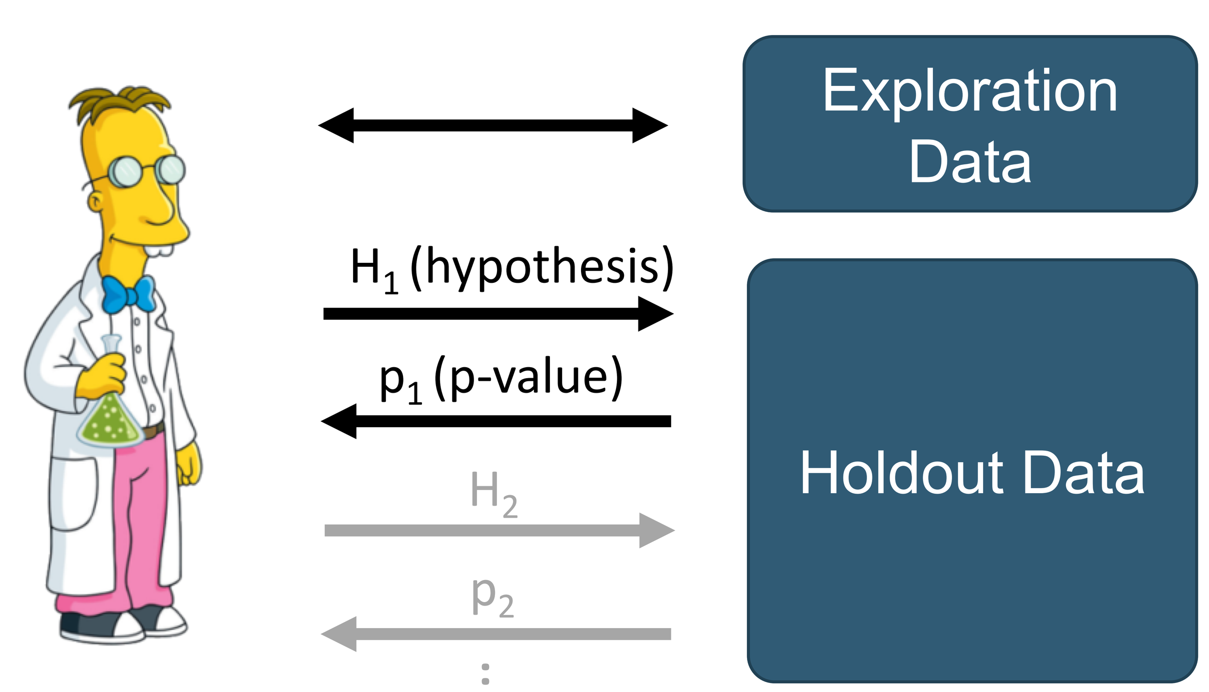
Prior Approaches

Prior approaches: statistically-invalid, or sample-inefficient.

Ex: "Naïve holdout"

Option 1: Collect new holdout set for each hypothesis (inefficient)

Option 2: "Reuse" same holdout set (invalid)



Method	Data Size (# samples)	Queries Possible
Naïve Holdout	n	$O(n)$
Reusable Holdout [Dwork et. al. '15]	n	$O(n^2)$
Our Method	n	$\exp(n)$

Our Results

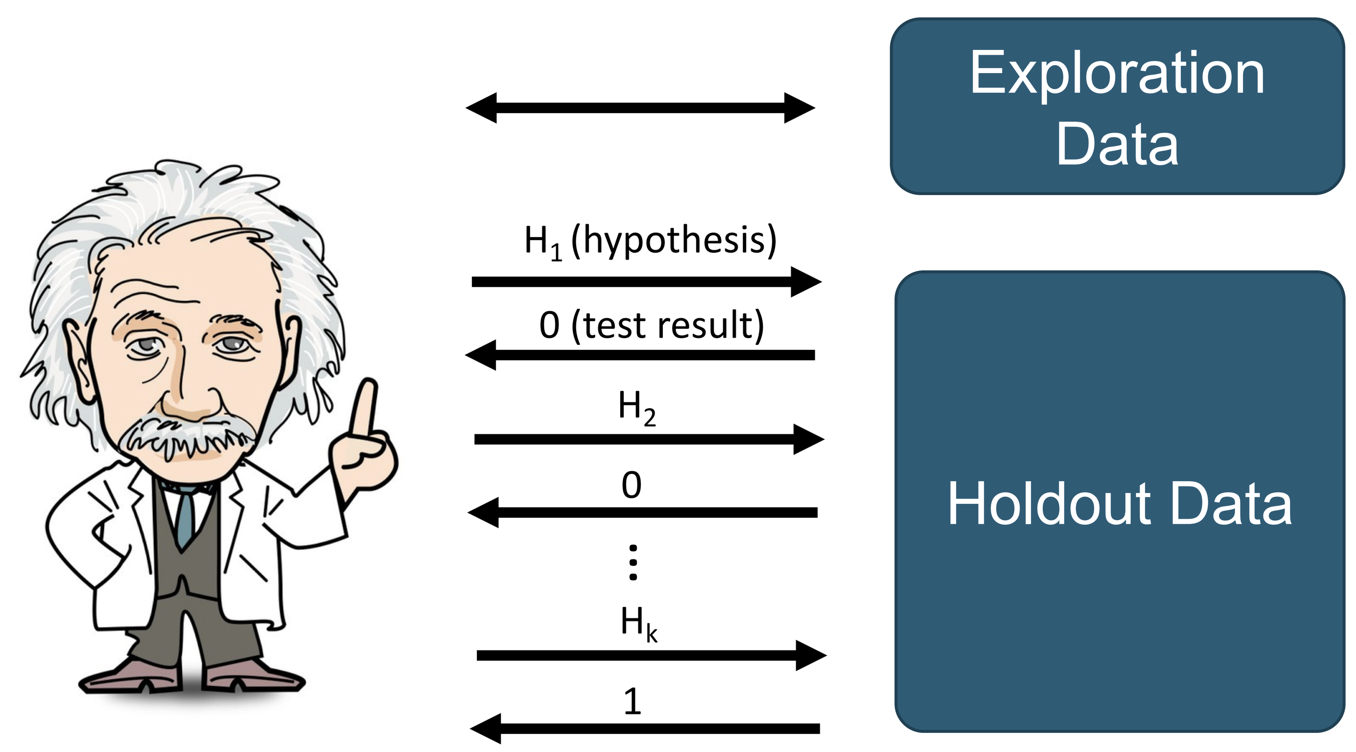
We propose a simple, statistically-sound, and sample-efficient framework for adaptive data science.

Our method allows the scientist to:

1. Explore part of the data to propose hypotheses
2. Adaptively propose & test hypotheses, based on previous hypothesis tests.
3. Test up to **exponentially-many** hypotheses in the size of the dataset, until discovering a true hypothesis (or several).
4. Bound the overall probability of a false-discovery

The Proposal: Generic Holdout

Key Idea: Holdout set only revels a **single bit** – whether hypothesis test passed or not. (NOT "how well" it fits, etc)



- Leaks minimal information from holdout set → prevents overfitting.
- Scientist can test many false-hypotheses before finding a true one.

Theorem: *Suppose a scientist interacting with the Generic Holdout generates a sequence of up to **m** adaptively-chosen hypotheses (H_1, H_2, …), and stops once **t** hypotheses are confirmed. If the false-positive probability of each hypothesis test H_i (on independent data) is bounded by **p**, then*

$$\Pr[\text{ scientist accepts a false hypothesis }] \leq m^t p$$

User Manual

How to use the Generic Holdout in your scientific process:

Assume we:

- Want to find a single true discovery
- Want to bound the probability of a false-discovery by **p**
- Will propose at most **m** hypotheses total (can be large).

Procedure:

1. Pick **n** large enough such that any hypothesis you pose can be tested with **n** iid samples, with false-positive probability $\leq \frac{p}{m}$.
   E.g. usually requires $n \sim \log(\frac{m}{p})$
2. Collect data (iid samples), and split it into a Holdout set of size **n**, and an Exploration set.
3. Use the Exploration Set as in your usual scientific process, to find plausible hypotheses.
4. Just before publishing a result, test the hypothesis against the holdout set, at false-positive level $\frac{p}{m}$ (seeing only the binary result).
5. If the test failed, you are free to repeat Steps 3-4 until finding a true hypothesis.

This controls the probability of false-discoveries, regardless of the method used to generate hypotheses.

Applications

1. **Main Application:** Preventing false discoveries for individual scientists/groups.
   - Alternative to pre-registration
2. Journal Application
   - Setting: Large public dataset collected once, many groups publish studies on it (eg, genomic data).
   - Proposal: Journals keep some of the data secret, as holdout. Use it to confirm every to-be-published study involving the public data.
   - Guarantee: Journal can confirm many true publications, before catching several false ones.

Full paper: https://arxiv.org/abs/1809.05596

Work supported in part by NSF GRFP Grant No. DGE1144152, ONR grant N00014-15-1-2388, and Madhu Sudan's Simons Investigator Award and NSF Award CCF 1715187.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences