Interactive Fingerprinting Codes and the Hardness of Preventing False Discovery

Thomas Steinke & Jonathan Ullman

Harvard & Northeastern

Motivation: False Discovery

PLOS | MEDICINE

Browse Publish About Search

advanced search

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

13,238 Saves | 1,673 Citations | 3,571

Article | Authors | Metrics | Comments

Abstract

Summary

There is increasing concern that most current published research findings are false. T probability that a research claim is true may depend on study power and bias, the nun other studies on the same question, and, importantly, the ratio of true to no relationshi the relationships probed in each scientific field. In this framework, a research finding is likely to be true when the studies conducted in a field are smaller; when effect sizes ar smaller; when there is a greater number and lesser preselection of tested relationship there is greater flexibility in designs, definitions, outcomes, and analytical modes; whe greater financial and other interest and prejudice; and when more teams are involved scientific field in chase of statistical significance. Simulations show that for most study

The Economist

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

Like 9k Tweet 1,066

Problem: Data Over-Use

- “Ideally,” data is used once and discarded. In practice, data is reused.

- Data analysis may depend on previous analysis of data.

- e.g. Freedman’s paradox: Use same dataset for variable selection and model fitting.

Model: Statistical Queries

Oracle(S)

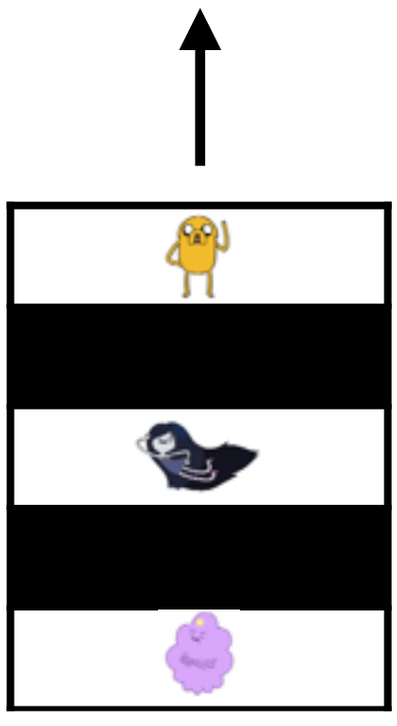
query q_1

answer a_1

query q_k

answer a_k

Interactive Analyst



Sample S of size n

Unknown Distribution D

Key Question [DFHPRR15]

- Q: How many samples n do I need to answer k *adaptive* statistical queries?

- Alternatively: Suppose I have a SQ learning algorithm that makes k statistical queries. If I simulate this as a PAC learning algorithm (in the obvious way) how many samples n does it need?

Our Results

- Q: How many samples n do I need to answer k *adaptive* statistical queries?

- For *non-adaptive* queries: n=O(log k) suffices

- **Main Result: n = $\Omega(\sqrt{k})$** *

- Previously: n = $\tilde{\Omega}(k^{1/3})$ [HU14]

- Almost Tight: n=$\tilde{O}(\sqrt{k})$ [DFHPRR15,BSSU15,NS15]

- *Assuming either dimension $O(n^2)$ or a computationally bounded oracle, super-logarithmic dimension, and OWFs.

Intuition

- If the analyst knows the sample, she can easily over-fit it.

- Answering lots of queries reveals enough information to identify the sample, which means the analyst may over-fit.

- Key tool: Interactive Fingerprinting Codes. [FT01,T03,LDRSdW13,this]