Ryan Rogers

Leveraging Privacy in Data Analysis

Data Analysis

$X \sim P^n$

t

$t(X)$

Analysis t

Science

AAAS

Data Analysis

$X \sim P^n$

t

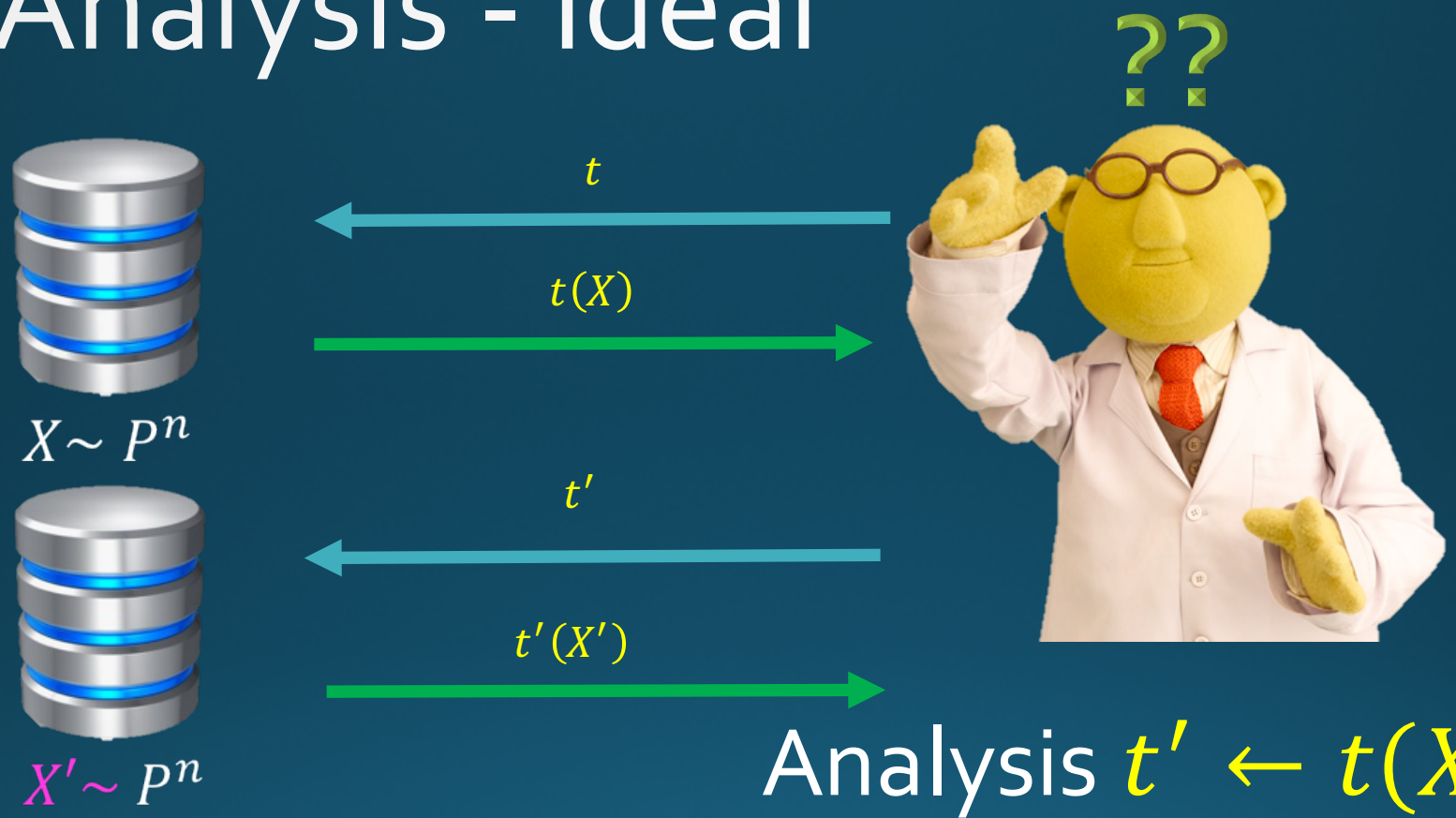
$t(X)$

??

Nothing Significant

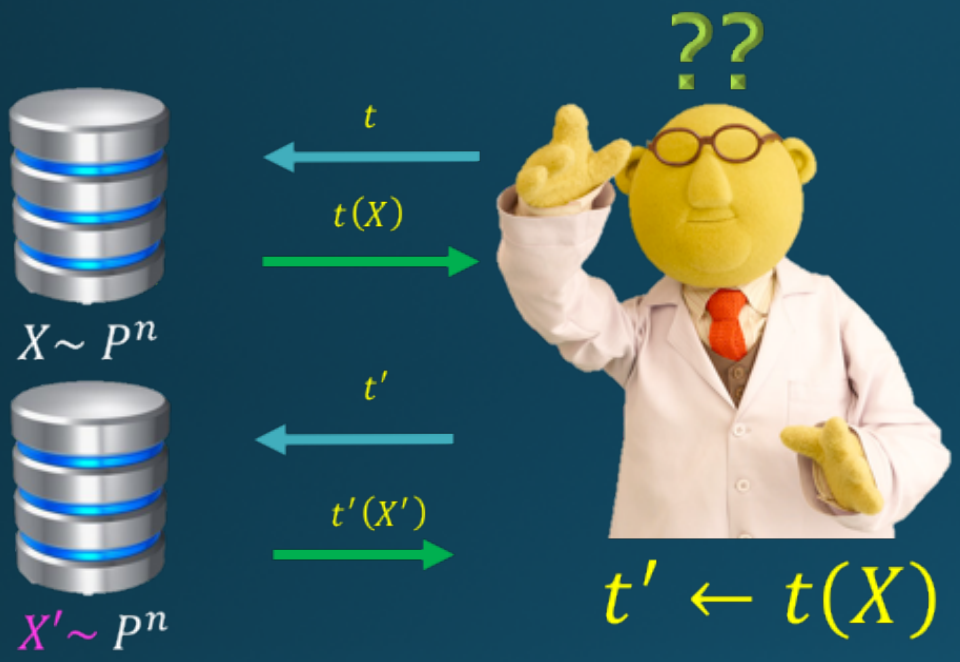
Analysis $t' \leftarrow t(X)$

Data Analysis - Ideal

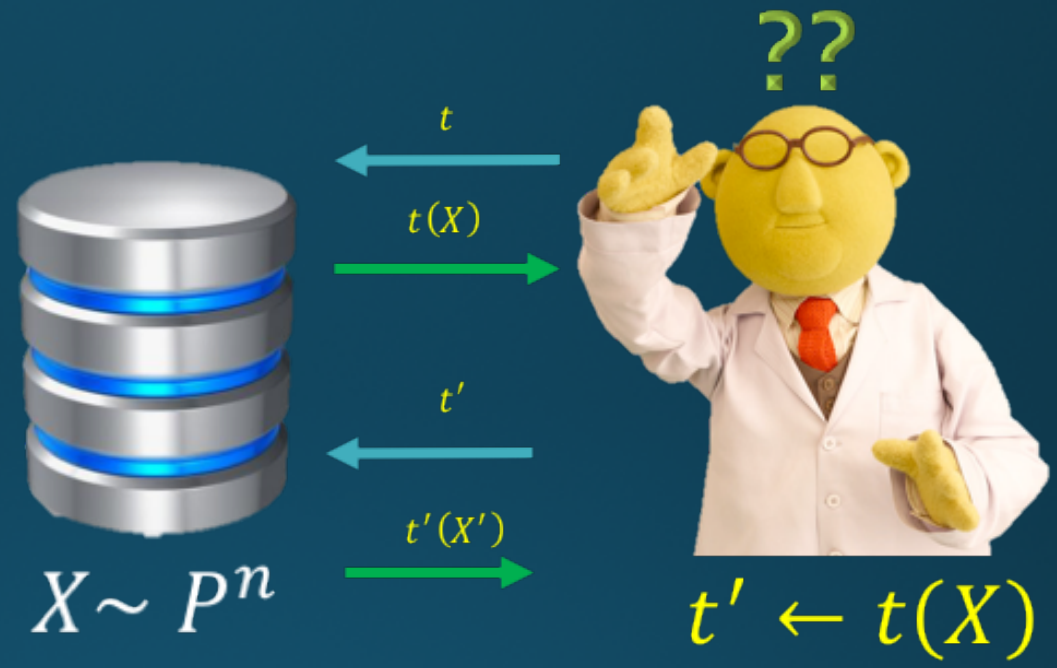


A lot of existing theory assumes tests are selected independently of the data.

Ideal World



Real World



How can we provide statistically valid answers to adaptively chosen analyses?

Ideal World

NOBA

Browse Content / The Replication Crisis in Psychology

The Replication Crisis in Psychology

By Edward Diener and Robert Biswas-Diener
University of Illinois at Urbana-Champaign, P

A Look into Mac... First Cheating S...

BY DANIEL WALTER
DECEMBER 08, 2015
DECEMBER 08, 2015
3 COMMENTS

At the end of May, the Imag...
(currently the word's top a...
competition policies by c...
(the Chinese search gia...
been banned from sub...
review called it "Ma...

The Economist

World politics | Business & finance | Economics | Scier

Unreliable research

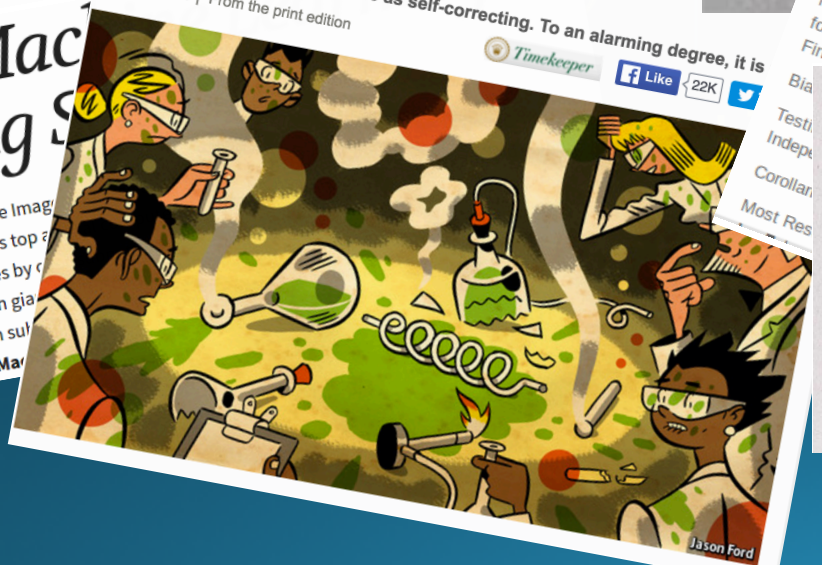
Trouble at the lab

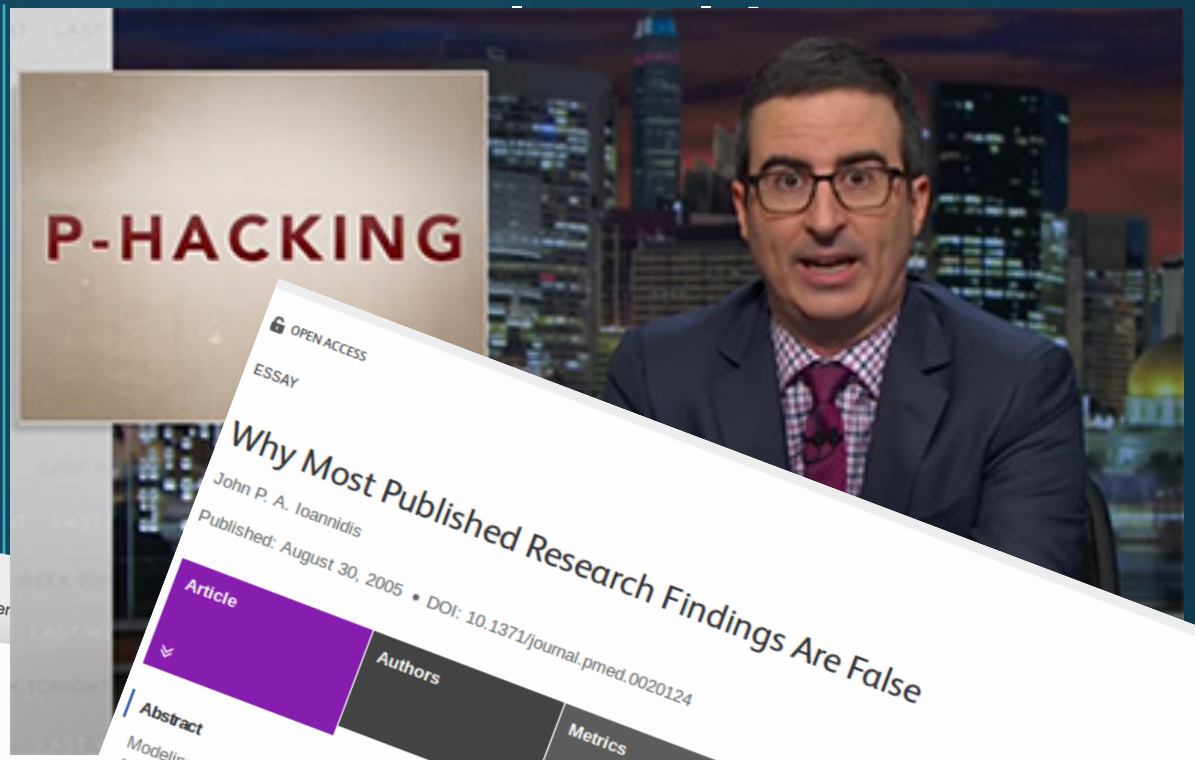
Scientists like to think of science as self-correcting. To an alarming degree, it is

Oct 19th 2013 | From the print edition

Timekeeper

Like 22K





OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

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

Article | Authors | Metrics

Abstract

Modeling the Framewo... for False Positi... Findin...

Bia... Testi... Indepe... Corollar... Most Res...

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

There is a growing realization that reported "statistically significant" claims in scientific publications are routinely mis-

a short mathematics test when it is expressed in two different contexts, involving either healthcare or the military. The question may be framed

This *multiple comparisons* issue is well known in statistics and has been called "*p*-hacking" in an influential 2011 paper by the psychology re-

...ction of tested relationships; where ...nomes, and analytical modes; when the...

Adaptive Data Analysis

$X \sim P^n$

t

$t(X)$

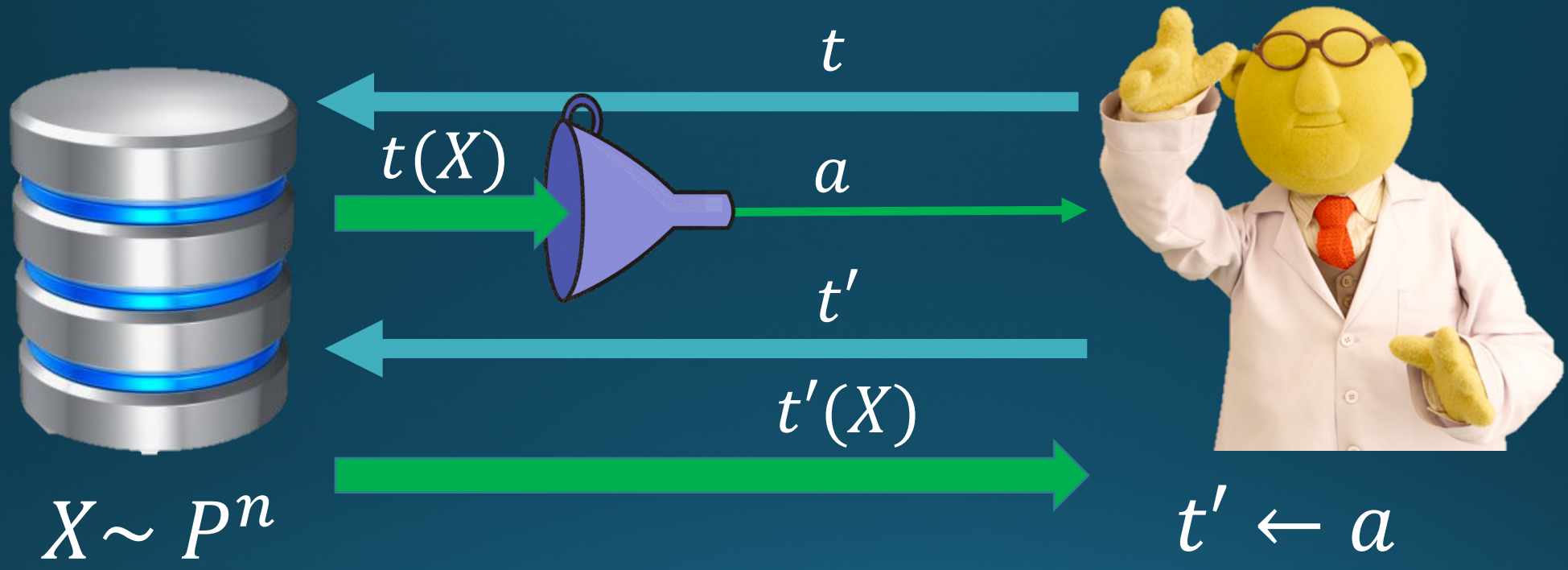
t'

$t'(X)$

$t' \leftarrow t(X)$

How can we provide statistically valid answers to adaptively chosen analyses?

Adaptive Data Analysis



Answer: Limit the info learned about the dataset with each analysis

[Dwork,Feldman,Hardt,Pitassi,Reingold,Roth'15].

Differential Privacy [Dwork,McSherry,Nissim,Smith'06]

- Limit information by making each analysis differentially private.
- A randomized algorithm $A: D^n \to Y$ is (ε, δ)-differentially private if for any neighboring data sets $x, x' \in D^n$ and for any $S \subseteq Y$,

$$P(A(x) \in S) \leq e^{\varepsilon} P(A(x') \in S) + \delta$$

- Ensures a stability guarantee on an algorithm.

Contributions

- Generalized and unified previous results on adaptive data analysis and its connection with DP [**R**, Roth, Smith, Thakkar'16].
  - Specifically with post-selection hypothesis testing.
- Differentially private hypothesis tests [Gaboardi, Lim, **R**, Vadhan'16], [Kifer,**R**'16].
  - Ensures statistical validity and privacy.
- Composition theorems for DP when privacy parameters can be selected adaptively [**R**,Roth,Ullman,Vadhan'16].

Thanks!