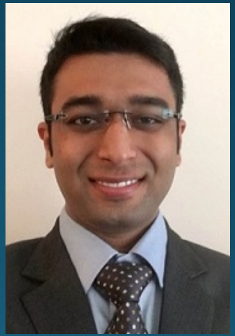
Max-Information, Differential Privacy, and Post-Selection Hypothesis Testing

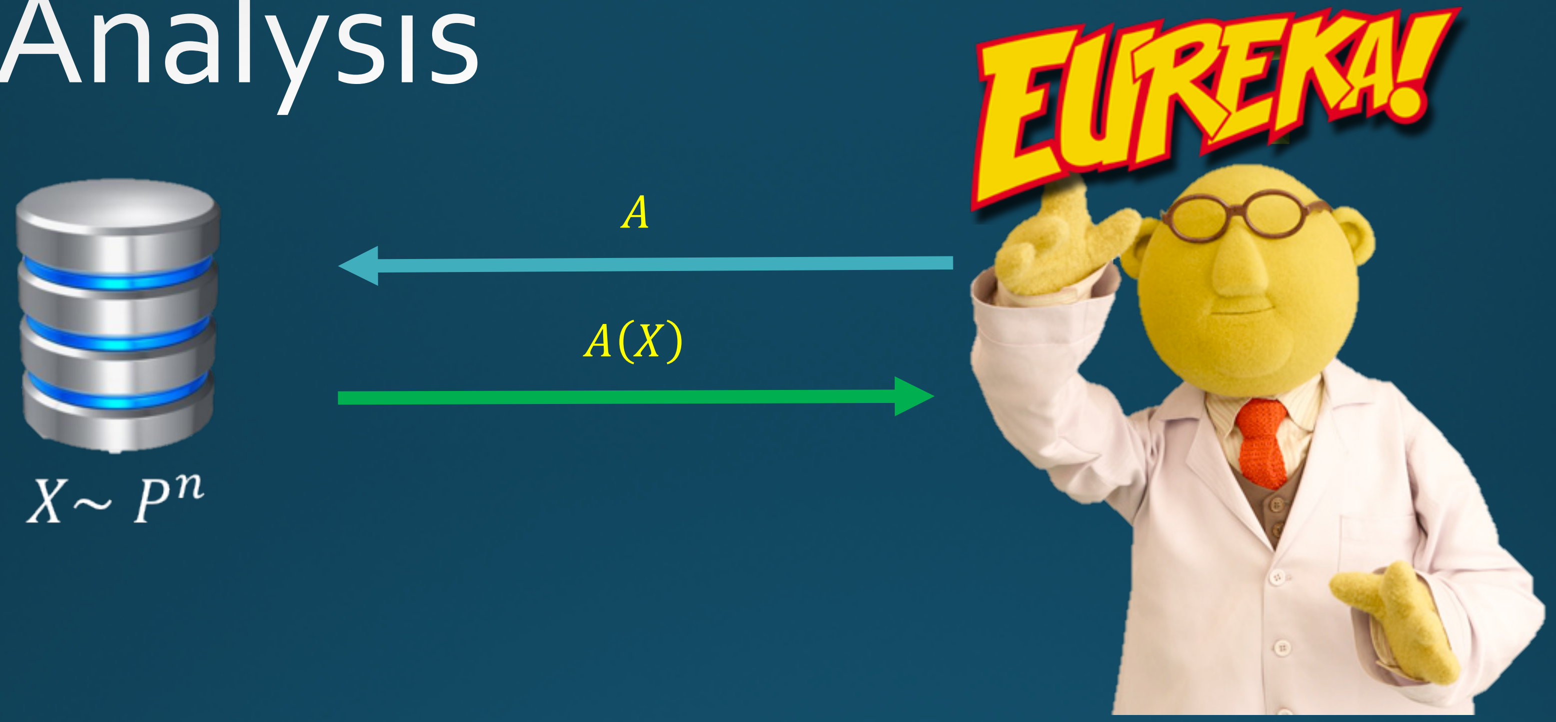
Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar



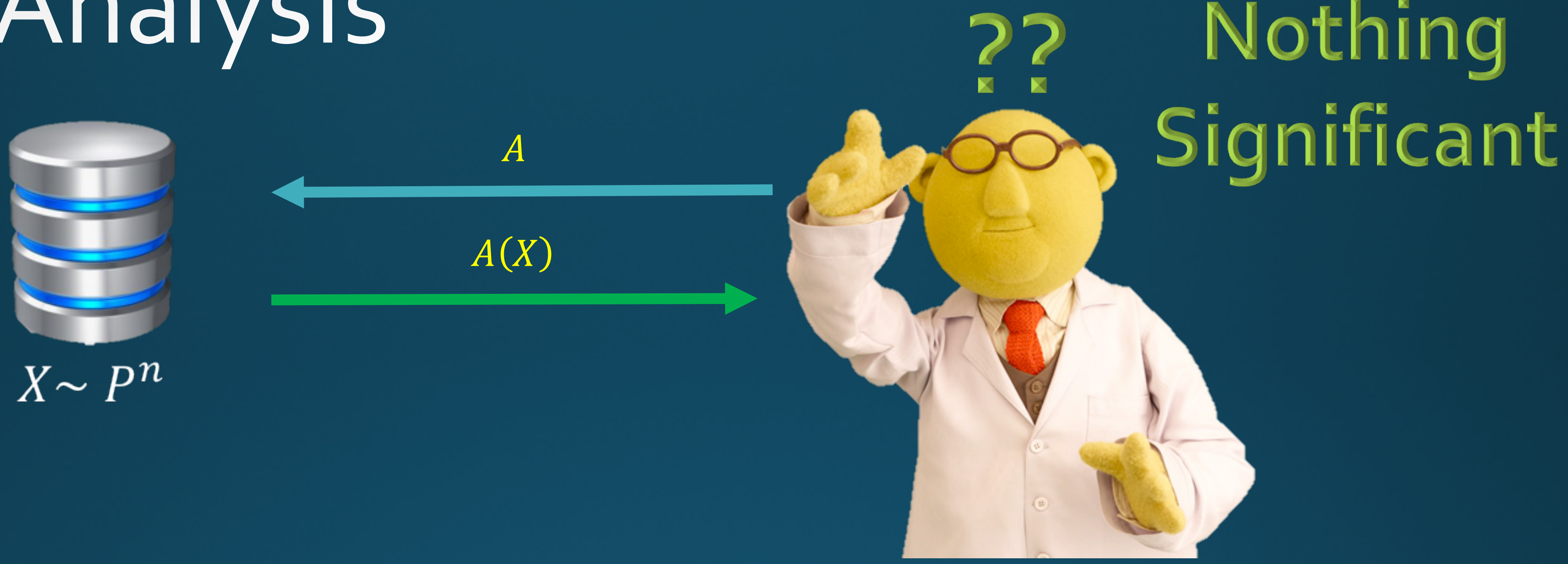
Supported by grants from the Sloan Foundation and NSF:
CNS-1253345, CNS-1513694, IIS-1447700.

Data Analysis



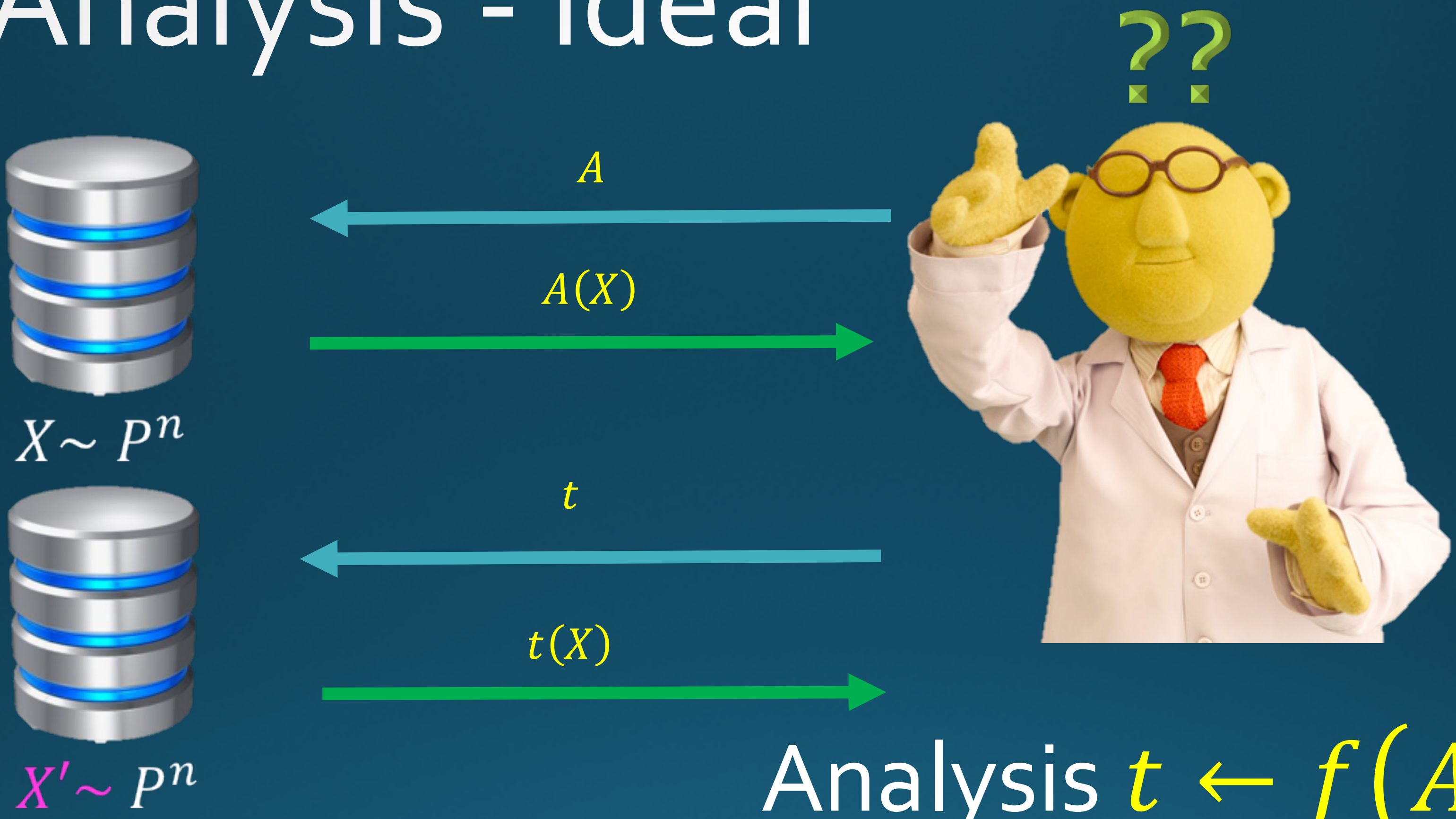
Analysis A

Data Analysis



Analysis $t \leftarrow f\big(A(X)\big)$

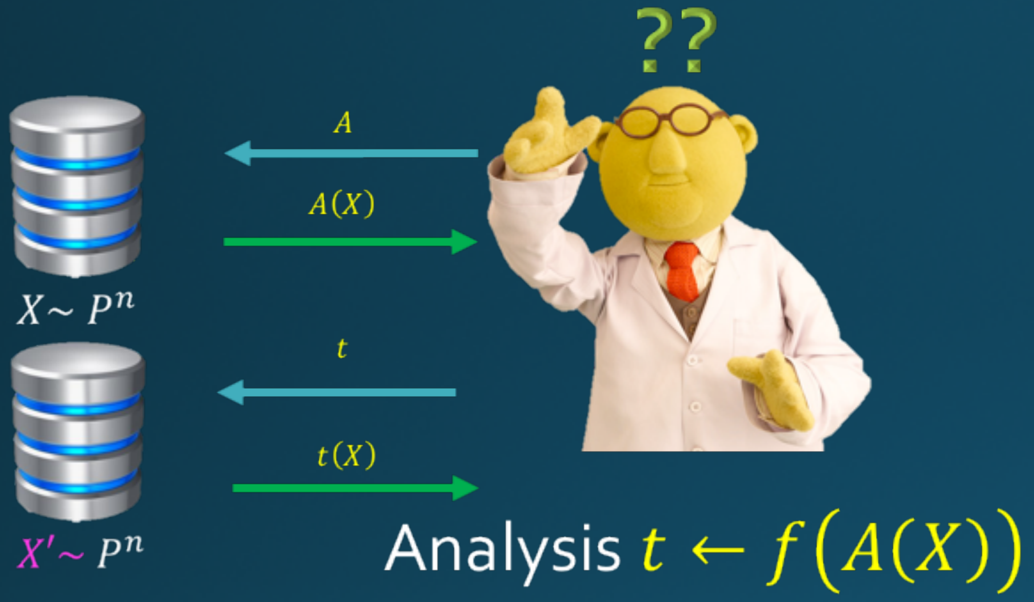
Data Analysis - Ideal



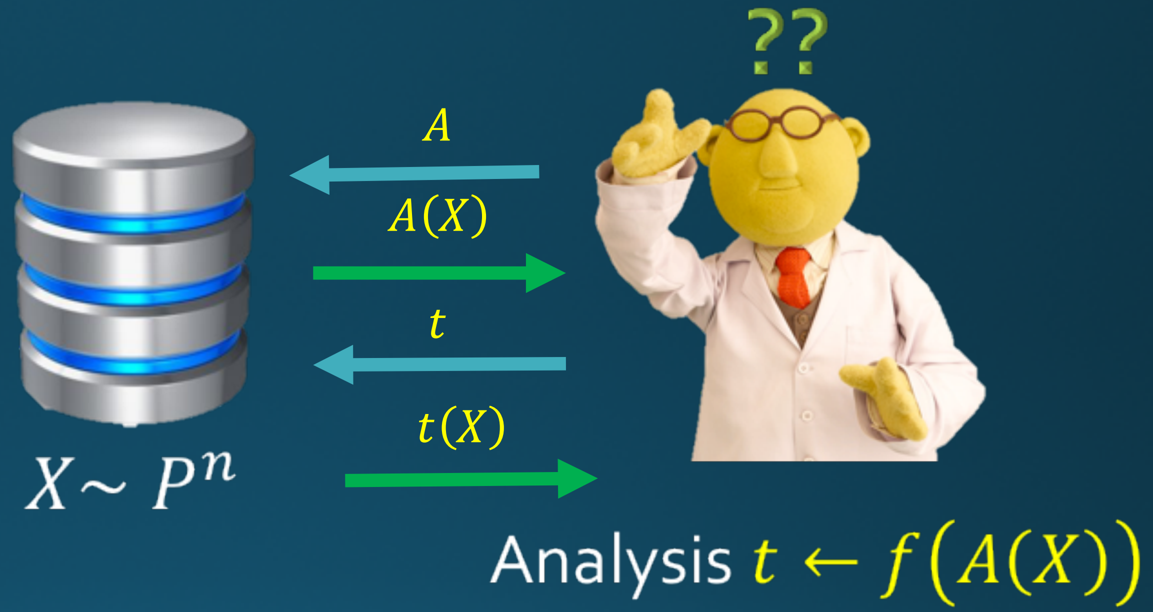
Analysis $t \leftarrow f\big(A(X)\big)$

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

Ideal World



Real World

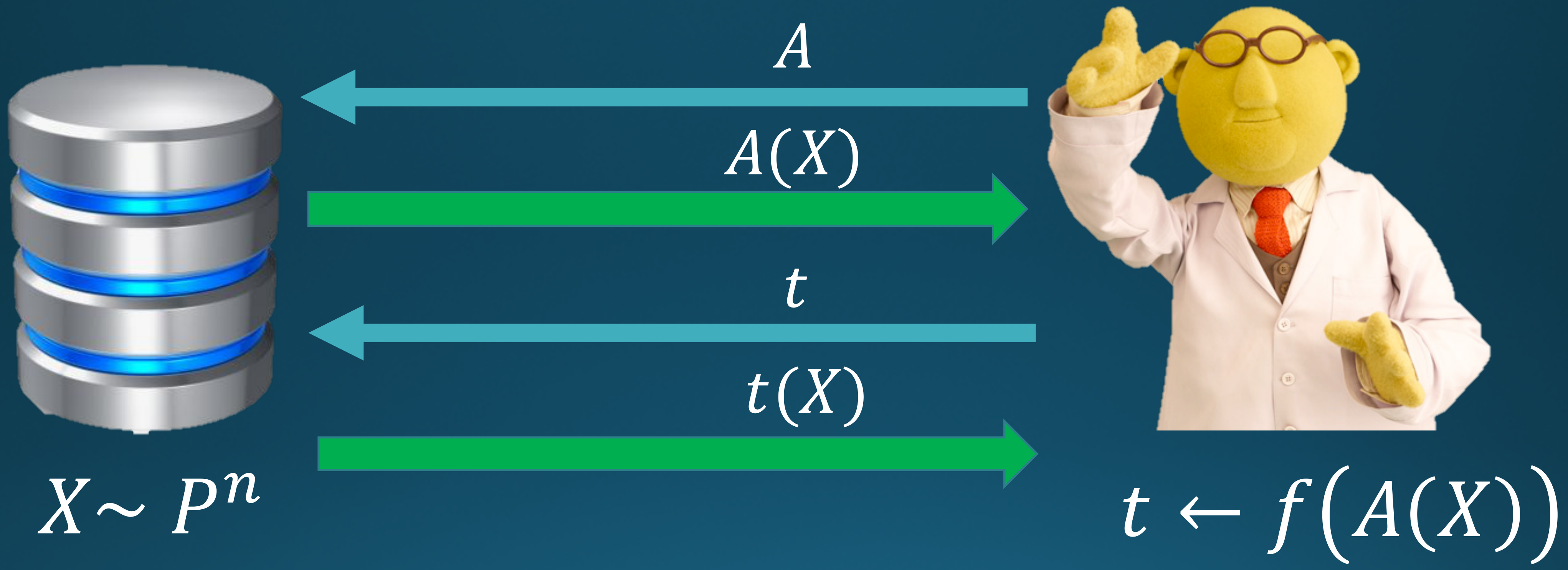


How can we provide statistically valid answers to adaptively chosen analyses?

Ideal World

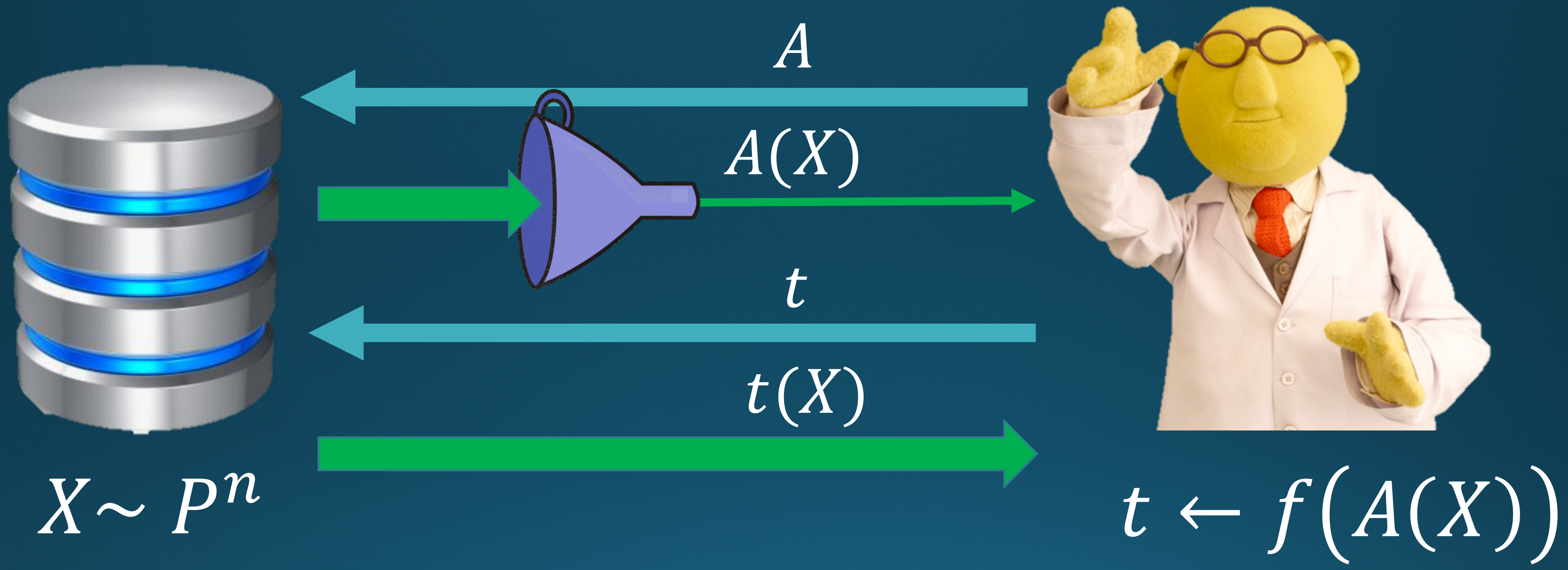
NOBA
Browse Content / The Replication Crisis in Psychology
The Replication Crisis in Psychology
By Edward Diener and Robert Biswas-Diener
University of Illinois at Urbana-Champaign,
A Look into Mac
First Cheating
BY DANIEL WALTER
DECEMBER 08, 2015
DECEMBER 08, 2015
3 COMMENTS
At the end of May, the Imag
(currently the word's top a
competition policies by
(the Chinese search gia
been banned from su
review called it "Ma
The Economist
World politics
Business & finance
Economics
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
Jason Ford
P-HACKING
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
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
a short mathematics test when it is expressed in two different contexts, involving either healthcare or the
This multiple comparisons issue is well known in statistics and has been called "p-hacking" in an influential

Adaptive Data Analysis



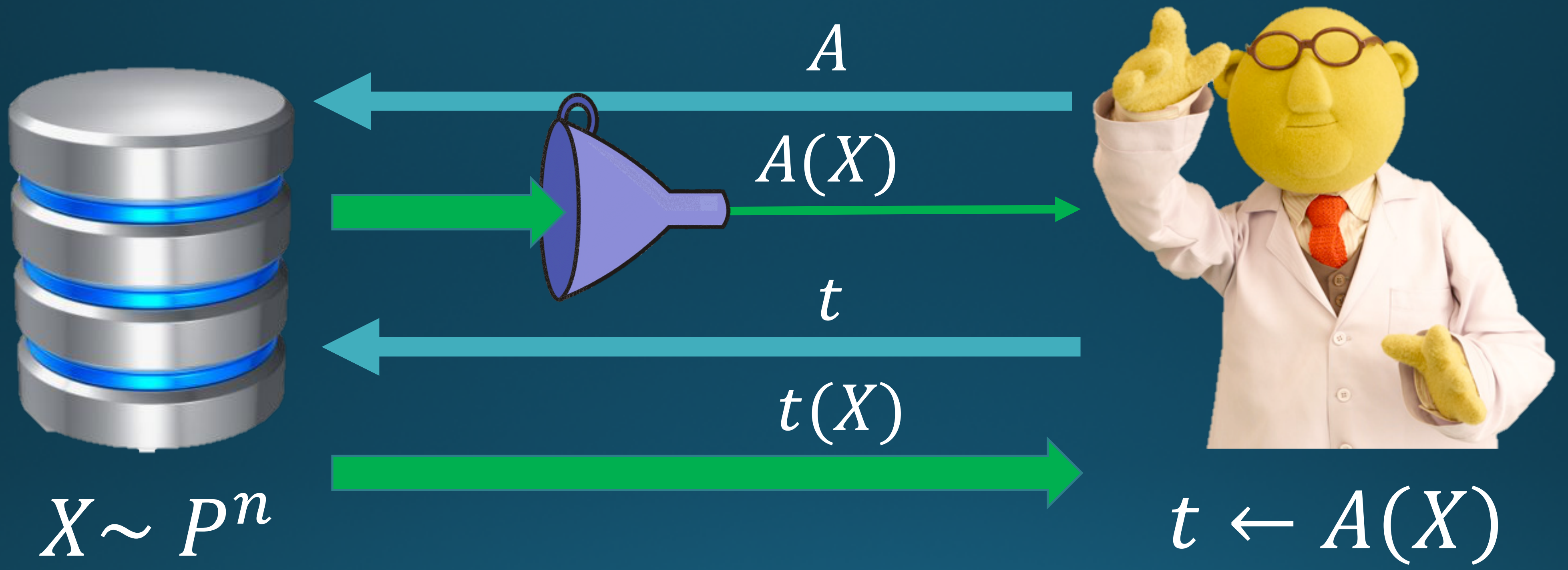
How can we provide statistically valid answers to adaptively chosen analyses?

Adaptive Data Analysis



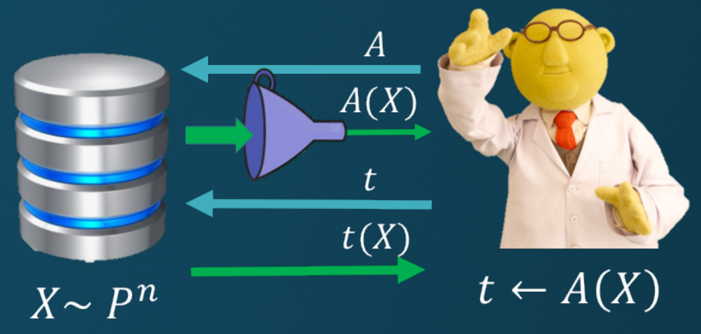
Answer: Limit the info learned about the dataset [Dwork,Feldman,Hardt,Pitassi,Reingold,Roth'15].

Adaptive Data Analysis



Answer: Limit the info learned about the dataset [Dwork,Feldman,Hardt,Pitassi,Reingold,Roth'15].

Contributions

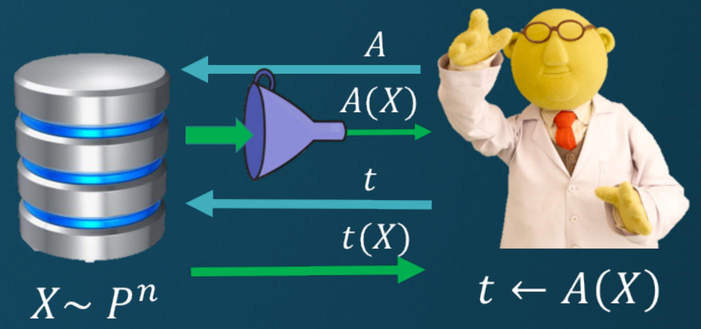


- Post-selection Hypothesis Testing
  - Bounded Max-Info $\implies$ Valid Tests
  - Tighter connection than previous results.

- Approximate Differential Privacy $\implies$ Bounded Max-Info
  - k rounds of adaptivity: max-info $\sim k$ rather than k^2.

Generalizes and unifies previous work

Related Work



- Lots of work in statistics community on post-selection inference [Freedman'83],[Leeb,Potscher'06],[Berk,Brown,Buja,Zhang,Zhao'13], …
  - Specific to type of analyses performed
- [DFHPRR](STOC'15,NIPS'15,Science'15)
  - Initial connections between information, privacy and adaptive analysis
- Accuracy for specific queries
  - [DFHPRR] (STOC'15,Science'15)
  - [Bassily,Nissim,Smith,Steinke,Stemmer,Ullman'16]
  - [Cummings,Ligett,Nissim,Roth,Wu'16]
  - [Russo,Zou'16]
  - [Wang,Lei,Fienberg'16]
- Impossibility results
  - [Hardt,Ullman'14], [Steinke,Ullman'15]

Hypothesis Testing

- Hypothesis test is defined by
  - null hypothesis $H_0 \subseteq \Delta(D)$ and
  - statistic:

$$t: D^n \to \{Inconclusive, Reject\}$$

- A *False Discovery* is when $X \sim P^n$ and $P \in H_0$ but $t(X) = Reject$

- Classical results apply when t is independent of X.
- Want to bound Pr[*False Discovery*] when $t \leftarrow A(X)$.

Max-Information [DFHPRR'15]

- Algorithm A has small max-info
  $\Rightarrow A(X)$ and X are "close" to independent.

- The β-approximate max-info between $A(X)$ and X i

$$I_\infty^\beta(A(X);X) = \log\left(\sup_{O} \frac{\Pr[(A(X),X) \in O] - \beta}{\Pr[(A(X'),X) \in O]}\right)$$

Real World

Ideal World

Max-Information of Algorithms [DFHPRR'15]

$$I_\infty^\beta(A(X);X) = \log\left(\sup_O \frac{\Pr[(A(X),X)\in O]-\beta}{\Pr[(A(X'),X)\ \in O]}\right)$$

An algorithm A has β-approximate max-info for data sets of size n if

any data distribution

$$I_\infty^\beta(A;n) = \sup_{\mathcal{S}:X\sim\mathcal{S}}\left\{I_\infty^\beta(A(X);X)\right\}$$

restrict to product distribution

$$I_{\infty,\Pi}^\beta(A;n) = \sup_{P:X\sim P^n}\left\{I_\infty^\beta(A(X);X)\right\}$$

Post-selection Hypothesis Testing

- [RZ'16]: When mutual info $I(X;A(X))$ is bounded, we can control Pr[False Discovery] for adaptively selected tests.

Bounded
Mutual Info
$I(X;A(X))$

$\Longrightarrow$

Bound on
Pr[False Discovery]
when $t \leftarrow A(X)$

Post-selection Hypothesis Testing

- [RZ'16]: When mutual info $I\big(X; A(X)\big)$ is bounded, we can control Pr[False Discovery] for adaptively selected tests.
- [This Paper]: We get a tighter connection via max-info

Bounded Mutual Info $I\big(X; A(X)\big)$ $\Longrightarrow$ Bounded Max-Info $I_\infty^\beta\big(X; A(X)\big)$ $\Longrightarrow$ Bound on Pr[False Discovery] when $t \leftarrow A(X)$

What procedures A have bounded max-info?

- [DFHPRR'15] Max-information bounds for:
  - (Pure) Differential Privacy – algorithmic stability condition.
  - Description Length – log(image size of A)

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

- 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 outcome $S \subseteq Y$ we have

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

If $\delta = 0$ we say **pure** DP, and otherwise **approximate** DP.

Technical Contributions

- [DFHPRR'15] : If $A: D^n \rightarrow T$ is $(\epsilon, 0)$-DP, then for $\beta > 0$,

$$I^{\beta}_{\infty,\Pi}(A; n) \leq \tilde{O}(\epsilon^2 n)$$

$$I^{0}_{\infty}(A; n) \leq O(\epsilon n)$$

- [This paper]: If $A: D^n \rightarrow T$ is (ϵ, δ)-DP, then

$$I^{\beta}_{\infty,\Pi}(A; n) \leq \tilde{O}(\epsilon^2 n) \text{ where } \beta \approx n\sqrt{\frac{\delta}{\epsilon}}$$

- [This paper] (based on [De'12]) : There exists an (ϵ, δ)-DP procedure A where,

$$I^{\beta}_{\infty}(A; n) \approx n \text{ for any } \beta < \frac{1}{2} - \delta$$

Consequences of Positive Result

Theorem: If $A: D^n \to T$ is (ϵ, δ)-DP, then

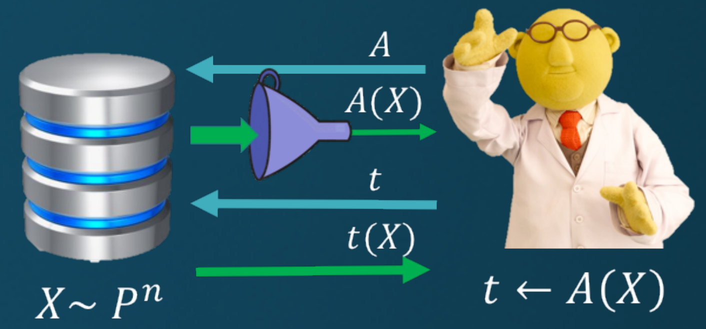
$$I^{\beta}_{\infty,\Pi}(A; n) \leq \tilde{O}(\epsilon^2 n) \text{ where } \beta \approx n\sqrt{\frac{\delta}{\epsilon}}$$

- Recover (optimal) results of [BNSSSU'16] for low sensitive queries.
  - However, our bounds apply more generally (e.g. adaptive hypothesis tests).

- Composition of k adaptively selected $(\epsilon, 0)$-DP procedures: $A_1, \ldots, A_k$
  - [DFHPRR'15]: $I^{\beta}_{\infty,\Pi}(A_k \circ \cdots \circ A_1; n) \leq \tilde{O}(n\epsilon^2 k^2)$
  - [This Paper]: $I^{\beta}_{\infty,\Pi}(A_k \circ \cdots \circ A_1; n) \leq \tilde{O}(n\epsilon^2 k)$

Via strong composition theorem from [Dwork,Rothblum,Vadhan'10]

Contributions



- Post-selection Hypothesis Testing
  - Max-Info Bound $\implies$ bound Pr[False Discovery] in adaptive settings
  - Improves on previous result of [RZ'16] that uses mutual info.
- (ϵ, δ)-DP $\implies$ Bounded Max–Info over product distributions
  - Recovers results from [BNSSSU'16] that dealt with specific analyses.
  - k rounds of adaptivity: we get max-info$\sim k$, where [DFHPRR'15] gives $\sim k^2$

Thanks!

Proof Sketch of Positive Result

Theorem: If $A: D^n \rightarrow T$ is (ϵ, δ)-DP, then

$$I^{\beta}_{\infty,\Pi}(A; n) \leq \tilde{O}(\epsilon^2 n) \text{ where } \beta \approx n\sqrt{\frac{\delta}{\epsilon}}$$

- Define the following random variable where $x \sim P^n, a \sim A(x)$ and

$$Z(a,x) = \log\left(\frac{\Pr[(A(X), X) = (a,x)]}{\Pr[(A(X'), X) = (a,x)]}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{\Pr[X_i = x_i \mid a, x_{1:i-1}]}{\Pr[X_i = x_i]}\right) = \sum_{i=1}^{n} Z_i(a, x_{1:i})$$

- Note that if we can bound this with high probability then we can bound approximate max-info.

Proof Sketch of Positive Result

- We want to apply a concentration bound (Azuma's inequality) to the following quantity: $\sum_{i=1}^{n} Z_i(a, x_{1:i})$
- We must then have:
  - A bound on the expectation of each $Z_i(a, x_{1:i})$
  - A bound on each $Z_i(a, x_{1:i})$
- Problem: Each $Z_i(a, x_{1:i})$ is NOT bounded.
- Although each term is bounded with high probability, conditioning on the same $A(X) = a$ and a prefix of data $X_{1:i-1} = x_{1:i-1}$ in every term complicates the argument.

Proof Sketch of Positive Result

For any $t > 0$

$$\Pr\left[\sum_{i=1}^{n} Z_i(A, X_{1:i}) \geq \epsilon^2 n + n\sqrt{\delta/\epsilon} + t\,\epsilon\sqrt{n}\right]$$

$$\leq \Pr\left[\sum_{i=1}^{n} Z_i(A, X_{1:i}) \geq \epsilon^2 n + n\sqrt{\delta/\epsilon} + t\,\epsilon\sqrt{n} \cap (A, X) \in GOOD\right]$$

$$+ \Pr[(A, X) \in BAD]$$

$$\leq e^{\frac{-t^2}{2}} + O\left(n\sqrt{\delta/\epsilon}\right)$$

Set $t = O(\epsilon\sqrt{n})$