Leveraging Privacy in Data Analysis

Ryan Rogers

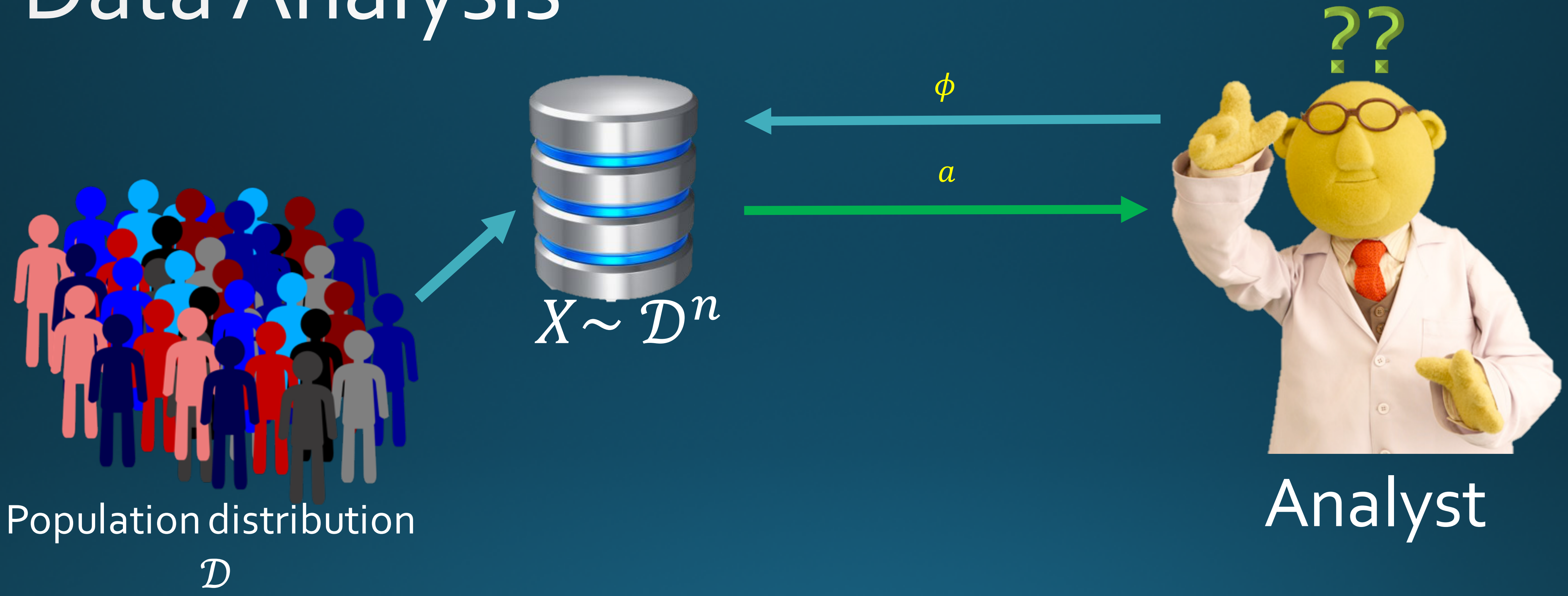
Applied Mathematics and Computational Sciences Department

Dissertation Defense

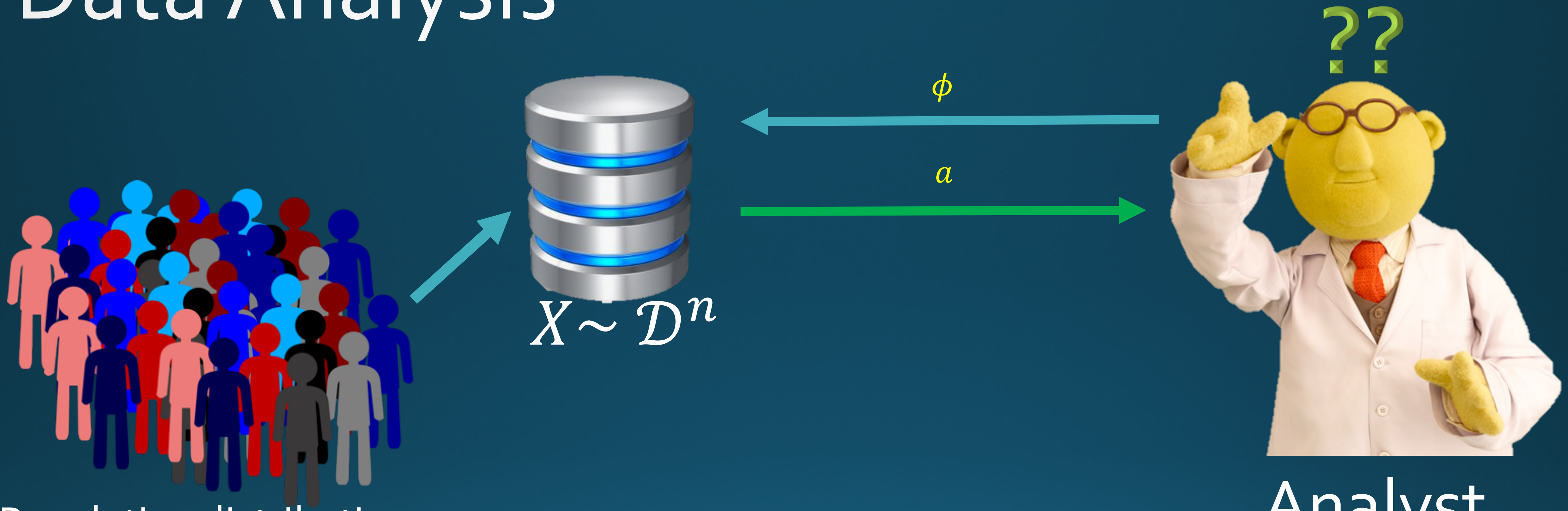
Advisors: Michael Kearns and Aaron Roth

Penn

Data Analysis



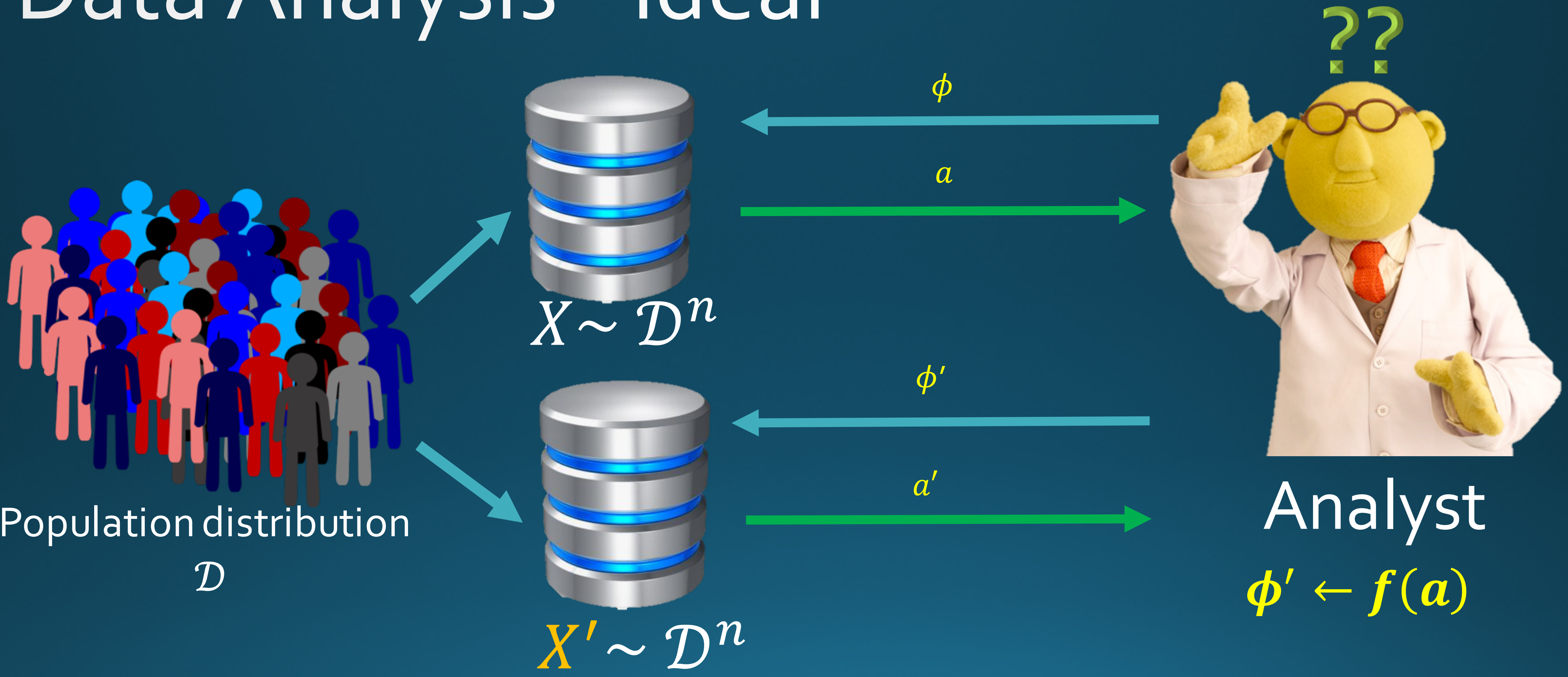
Data Analysis



Population distribution
$\mathcal{D}$

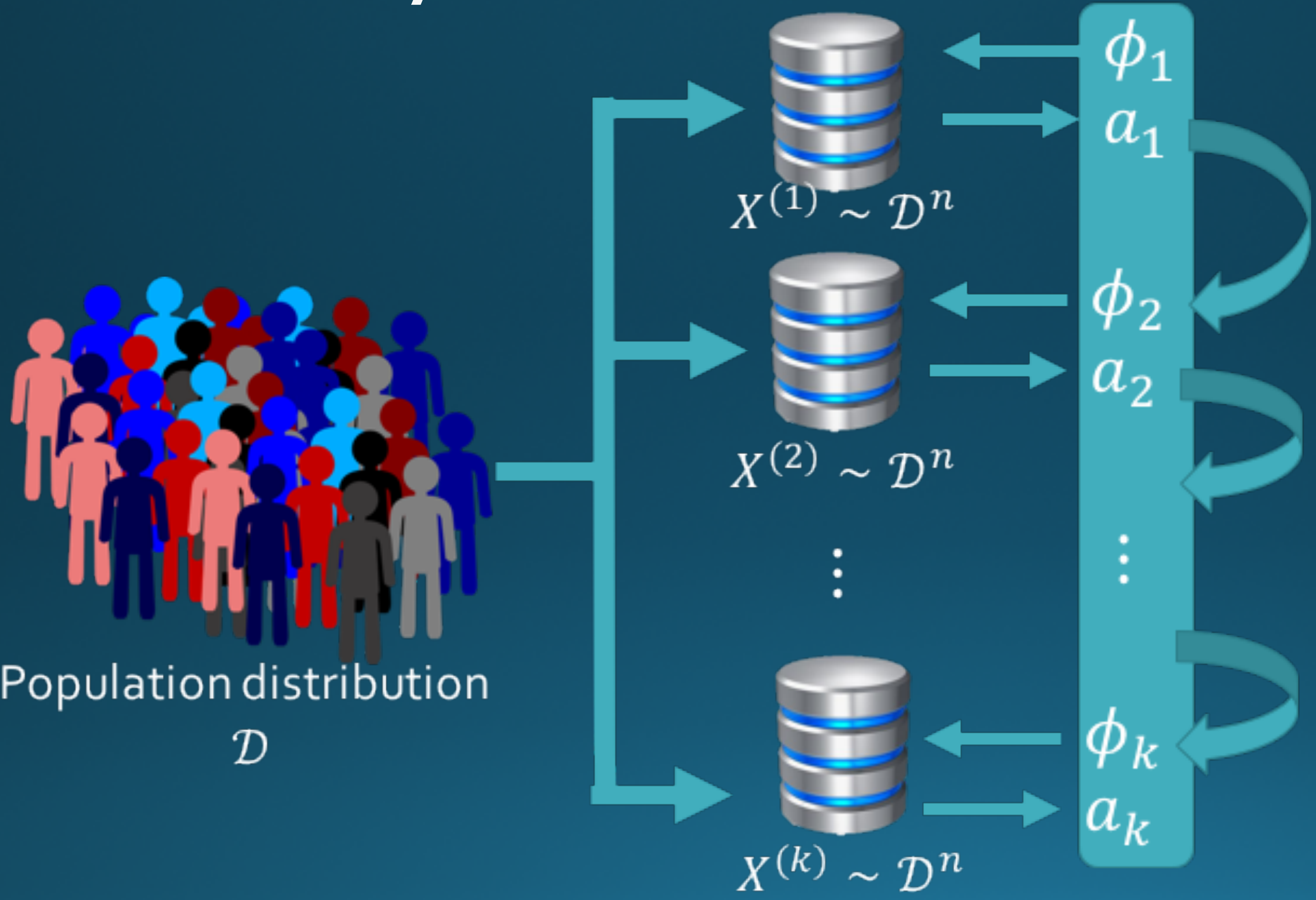
Analyst
$\boldsymbol{\phi' \leftarrow f(a)}$

Data Analysis - Ideal



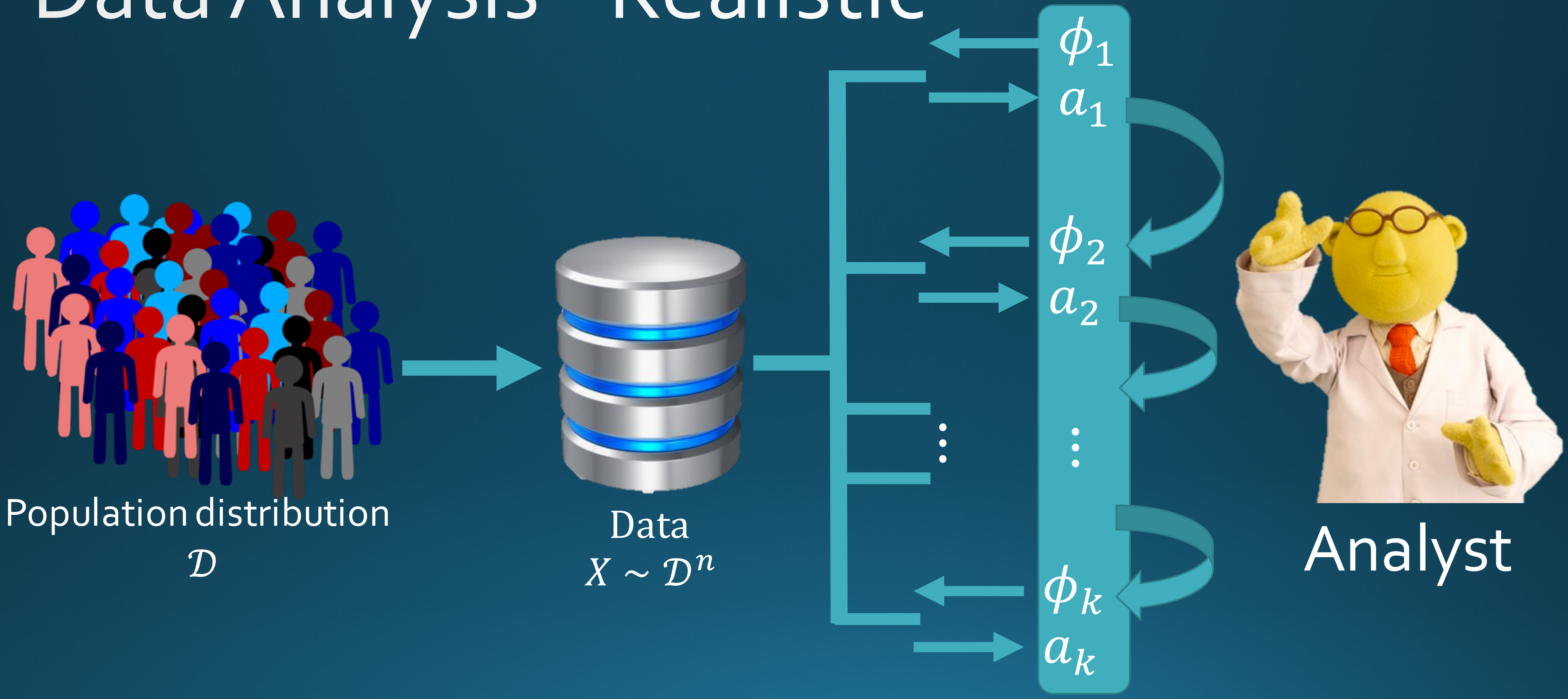
A lot of existing theory assumes tests are selected independently of the data.

Data Analysis - Ideal



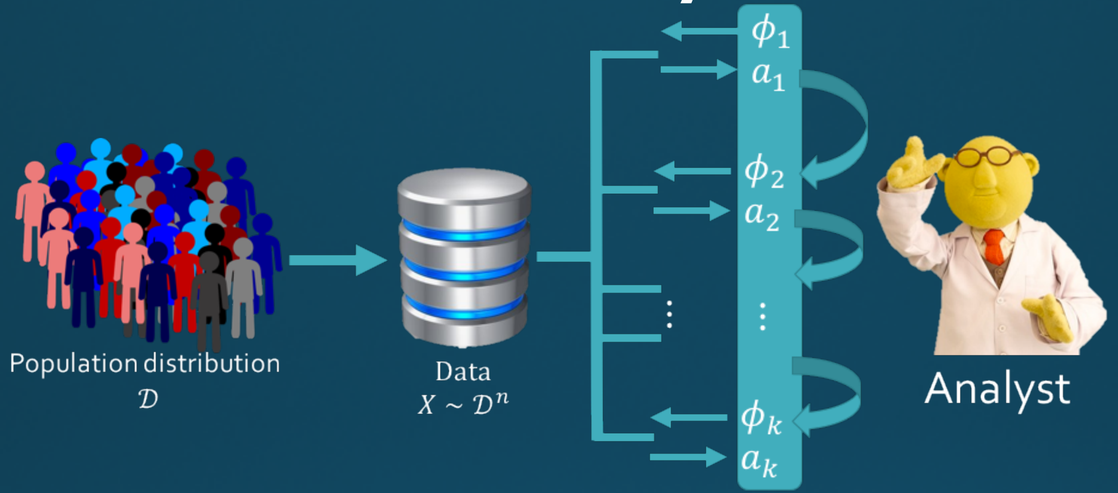
Analyst

Data Analysis - Realistic



Data Analysis - Realistic

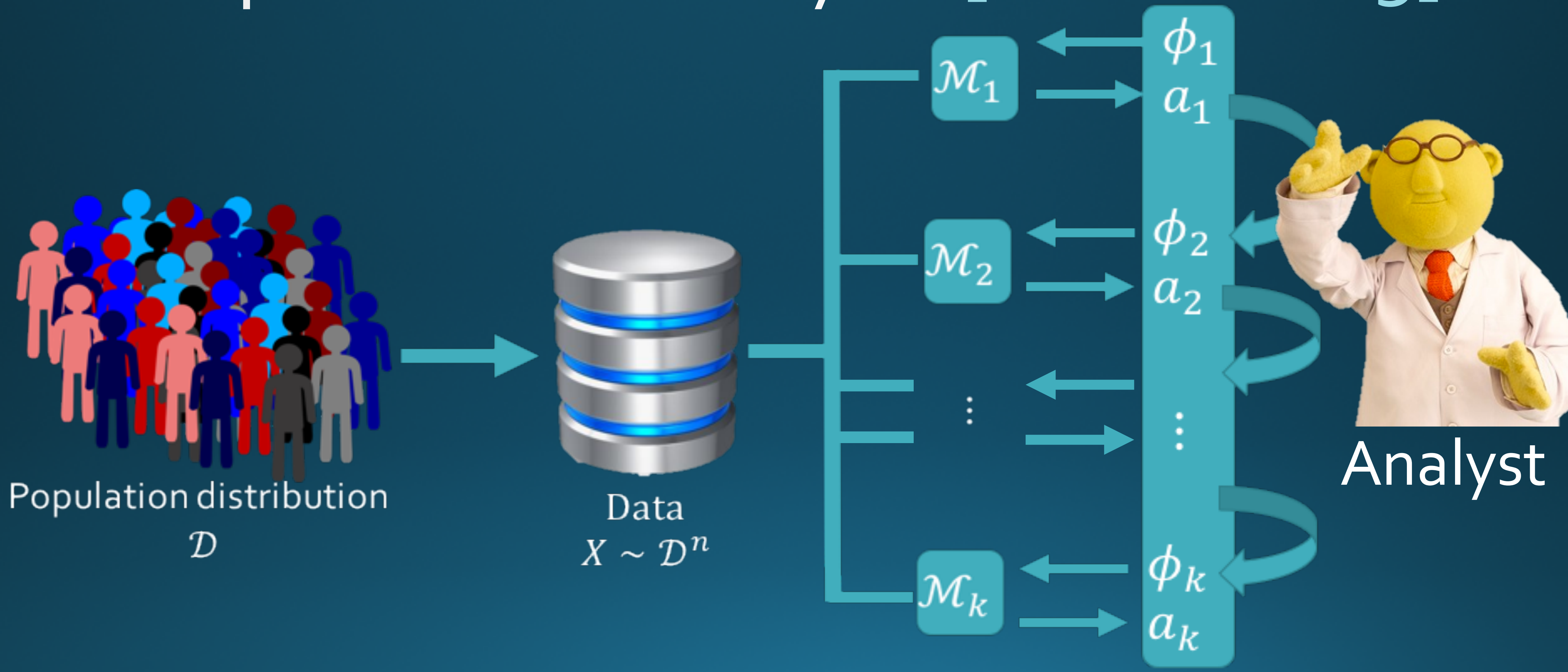
Adaptive Data Analysis



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

1. Traditional method – split the dataset into k chunks.
   - Requires $k \ll n$.
2. Limit the info learned about the dataset with each analysis [Dwork,Feldman,Hardt,Pitassi,Reingold,Roth'15].
   - Can handle $k \gg n$.

Adaptive Data Analysis [DFHPRR'15]



Example – Statistical Queries

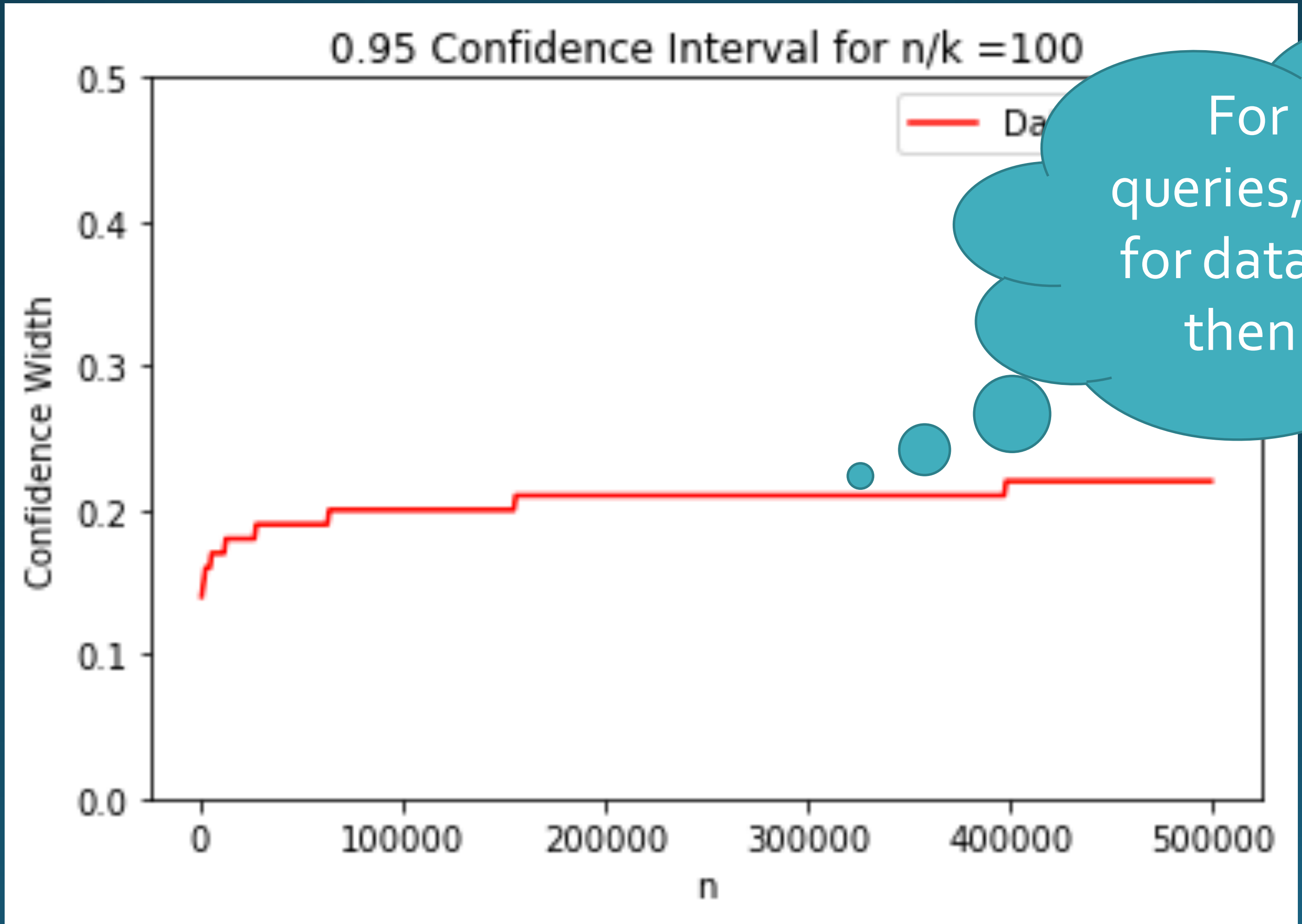
- Let each analysis $\phi_j : \mathcal{X} \to [0,1]$ for $j = 1, \cdots, k$. We want to estimate $\phi_j(\mathcal{D}) = \mathbb{E}_{Y \sim \mathcal{D}}[\phi_j(Y)]$
- Want to design algorithms $\mathcal{M}_j$ such that $\mathcal{M}_j(X) = a_j$ and

$$\max_{j \in \{1, \cdots, k\}} \{|a_j - \phi_j(\mathcal{D})|\} \leq \tau \qquad \text{w.p.} \geq 1 - \beta$$

- [DFHPRR'15]: $\mathcal{M}_j(X) = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) + N(0, \sigma^2)$

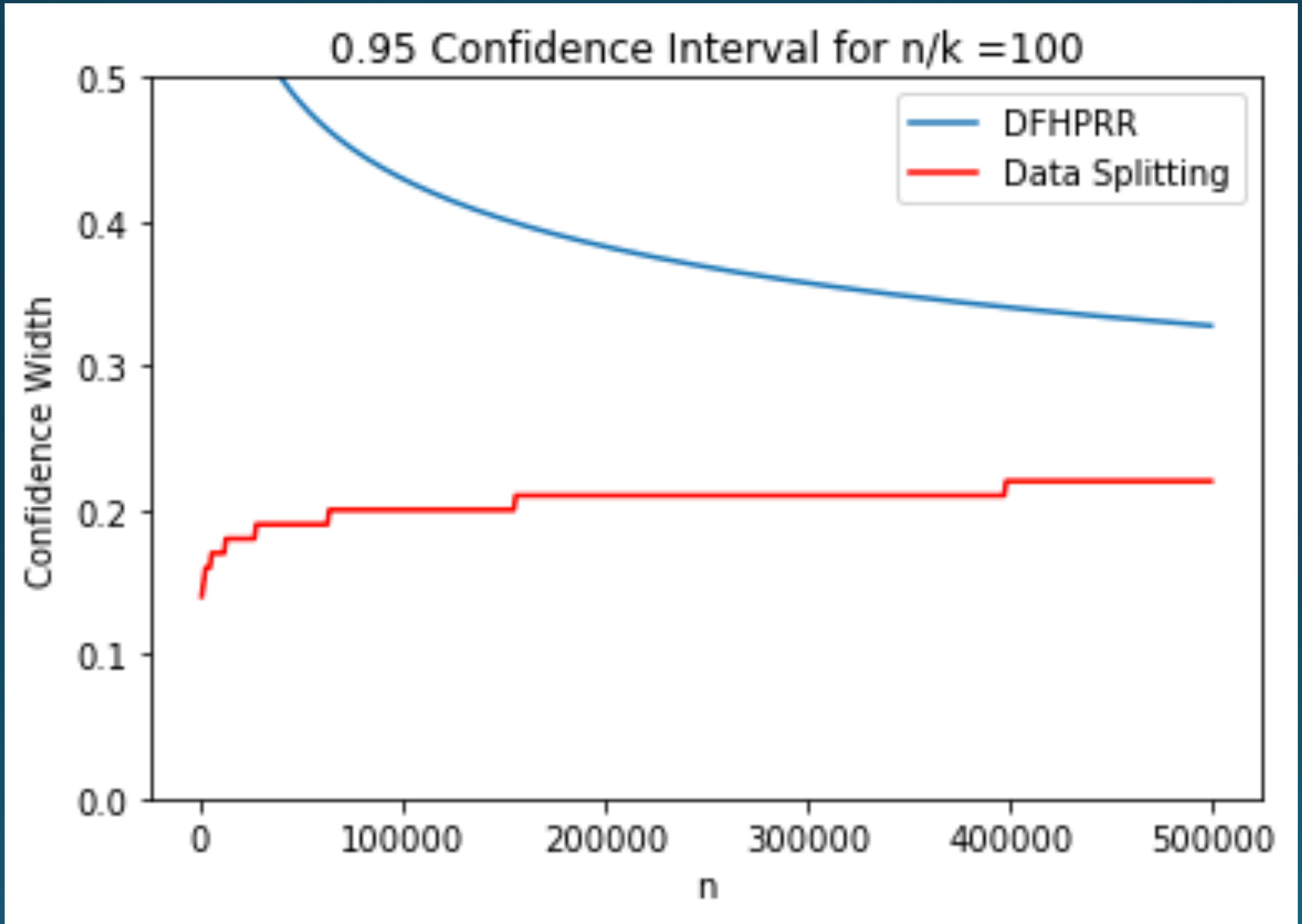
  so for properly chosen σ, we can have $k \sim n^2$

Comparison with Data Splitting

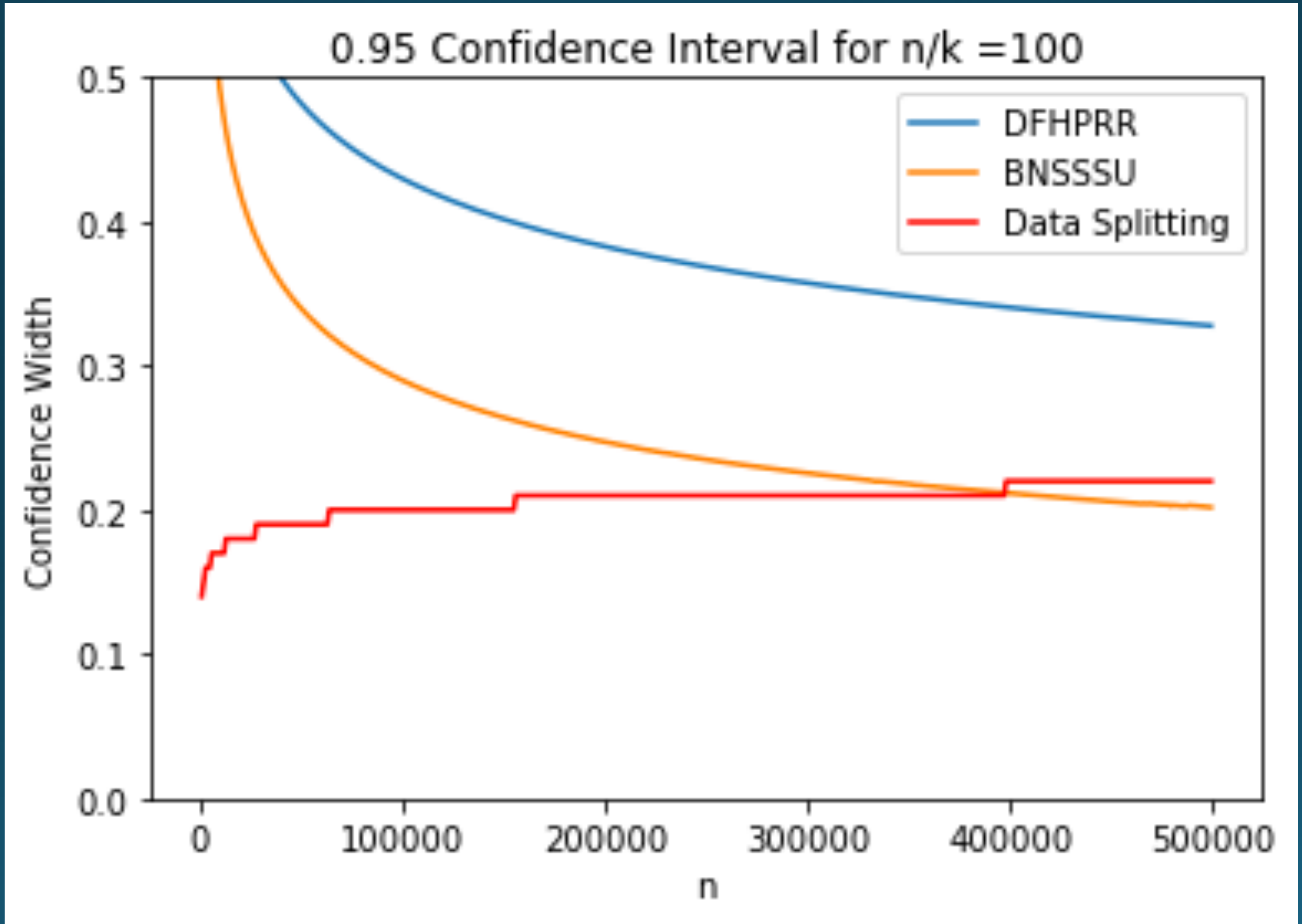


For each of the *k* queries, bound accuracy for dataset of size = *n*/*k*, then union bound.

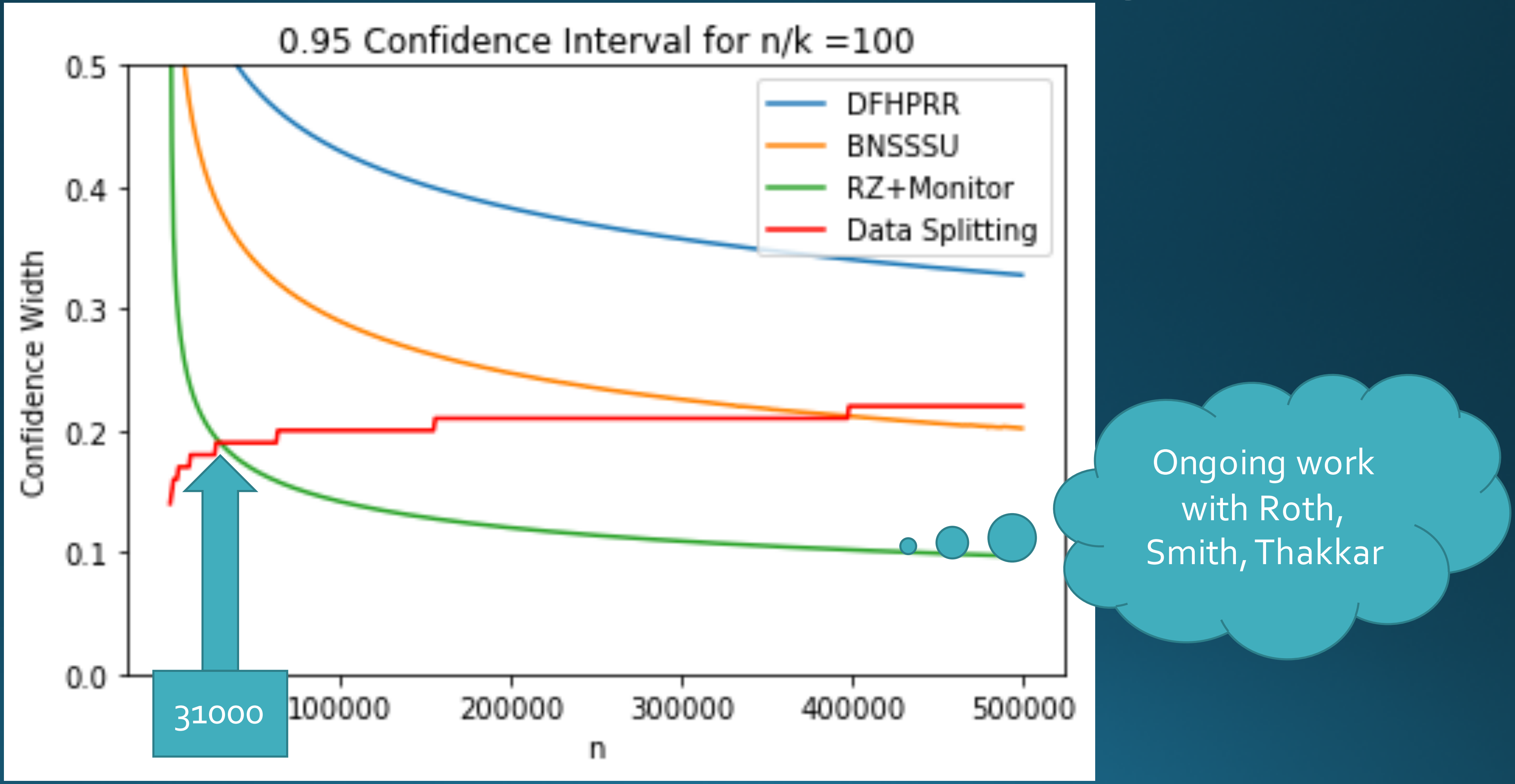
Comparison with Data Splitting



Comparison with Data Splitting

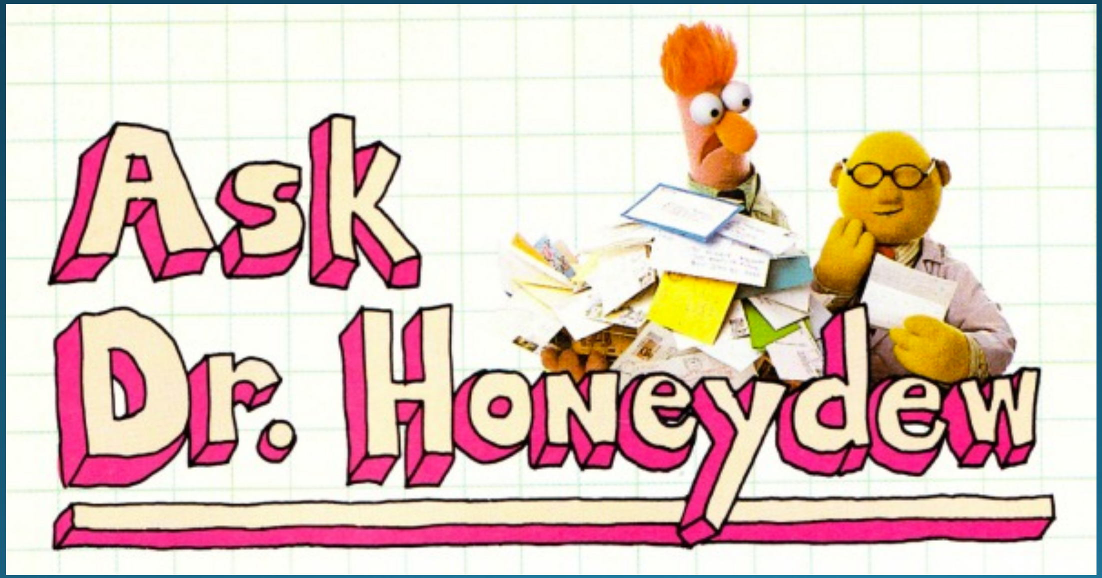


Comparison with Data Splitting



Questions

- What types of algorithms can we use to ensure validity?
- Can we do more general types of analyses, beyond linear queries?
- What existing techniques can we use, and what needs to be modified for adaptivity and the algorithms we use?



Related Work

- Lots of work in statistics community on selective 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]

Outline

- Post Selection Hypothesis Testing [**R**,Roth,Smith,Thakkar FOCS'16]
  - Connection between Max-Info and Differential Privacy
- DP composition with adaptively selected parameters [**R**,Roth,Ullman,Vadhan NIPS'16]
  - Privacy Odometers and Filters
- Private Hypothesis Tests [Gaboardi,Lim,**R**,Vadhan ICML'16],[Kifer,**R** AISTATS'16]
  - Chi-Square Tests
- Directions for Future Work

Max-Information [DFHPRR'15]

- Algorithm $\mathcal{M}$ has small max-info
  $\Rightarrow \mathcal{M}(X)$ and X are "close" to independent.

- The β-approximate max-info between $\mathcal{M}(X)$ and X:

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

Real World

Ideal World

Max-Information of Algorithms [DFHPRR'15]

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

The β-approximate max-info of an algorithm $\mathcal{M}$ on data sets of size n is:

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

Hypothesis Testing

- Hypothesis test $t: \mathcal{X}^n \to \{Inconclusive, Reject\ H_0\}$ is defined by
  - null hypothesis $H_0 \subseteq \Delta(\mathcal{X})$ and
  - statistic:

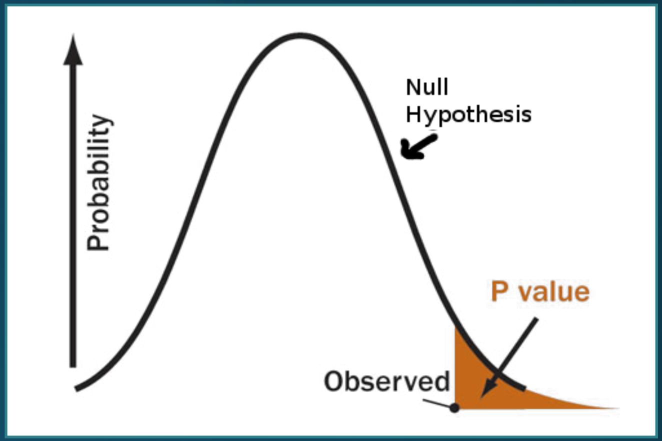
$$g: \mathcal{X}^n \to \mathbb{R}$$

	H_0 is True	H_0 is False
t rejects	α = False Discovery	Power
t fails to reject	1-α = Significance	Type II Error

- Want to bound $\mathbb{P}[$False Discovery$] \leq \alpha$ and get good Power.

p-Values

- The p-value associated with a value y and a distribution $\mathcal{D} \in H_0$ is given as
$$p(y) = \mathbb{P}_{X \sim \mathcal{D}^n}[g(X) > y]$$
- Denotes the prob of observing a value of the test statistic that is at least as extreme as y.



- Note that $p\big(g(X)\big) \sim U[0,1]$ if $X \sim \mathcal{D}^n$ where $\mathcal{D} \in H_0$.
- If we reject the model when $p\big(g(X)\big) < \alpha$ then
$$\mathbb{P}[\text{False Discovery}] < \alpha.$$
- No longer true when $t \leftarrow \mathcal{M}(X)$!

p-Value Corrections

- Even when we use the data to determine a test, we still want to be able to control the $\mathbb{P}$[False Discovery].
- A function $\gamma: [0,1] \rightarrow [0,1]$ is a valid *p*-value correction function for a selection procedure $\mathcal{M}: \mathcal{D}^n \rightarrow \mathcal{O}$ if for every α the procedure:
  1. Select test $t \leftarrow \mathcal{M}(X)$
  2. Reject H_0 if $p\big(g(X)\big) < \gamma(\alpha)$

  has probability at most α of false discovery.

Max-Info gives Valid *p*-Value Corrections

- If we have selection procedure $\mathcal{M}$ such that $I_{\infty,\Pi}^{\beta}(\mathcal{M}, n) \leq m$ then a valid *p*-value correction function is

$$\gamma(\alpha) = \frac{\alpha - \beta}{2^m}$$

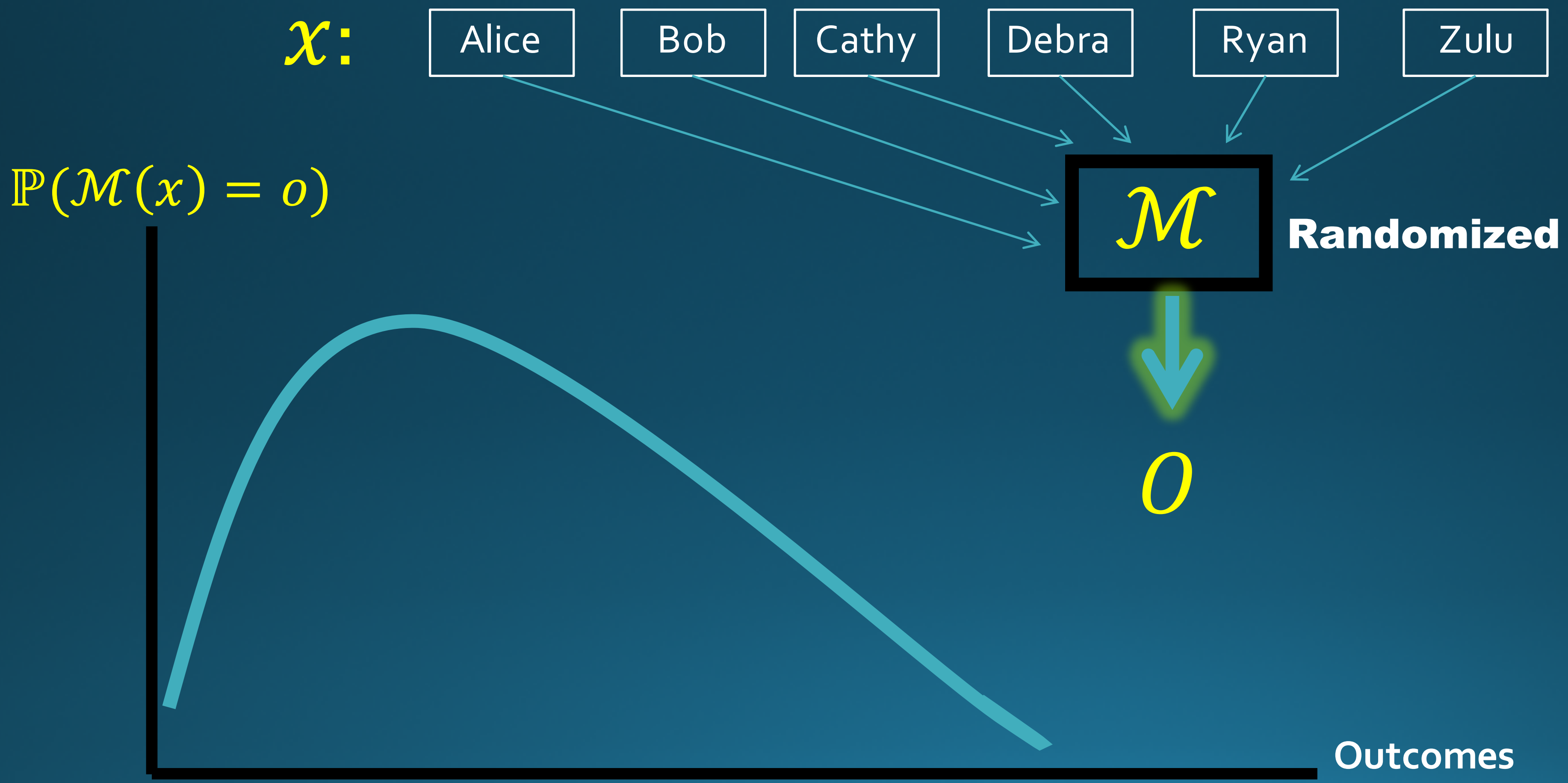
- Proof: Let $S \subseteq \mathcal{X}^n \times \mathcal{O}$ be the event that $\mathcal{M}$ selects a test statistic where the *p*-value is at most $\gamma(\alpha)$, but the null is true.

$$\mathbb{P}\left[p\left(g(X)\right) \leq \gamma(\alpha) \cap t \leftarrow \mathcal{M}(X) | H_0\right]$$
$$= \mathbb{P}\left[\left(X, \mathcal{M}(X)\right) \in S | H_0\right]$$
$$\leq 2^m \underbrace{\mathbb{P}\left[\left(X, \mathcal{M}(X')\right) \in S | H_0\right]}_{\leq \gamma(\alpha)} + \beta$$

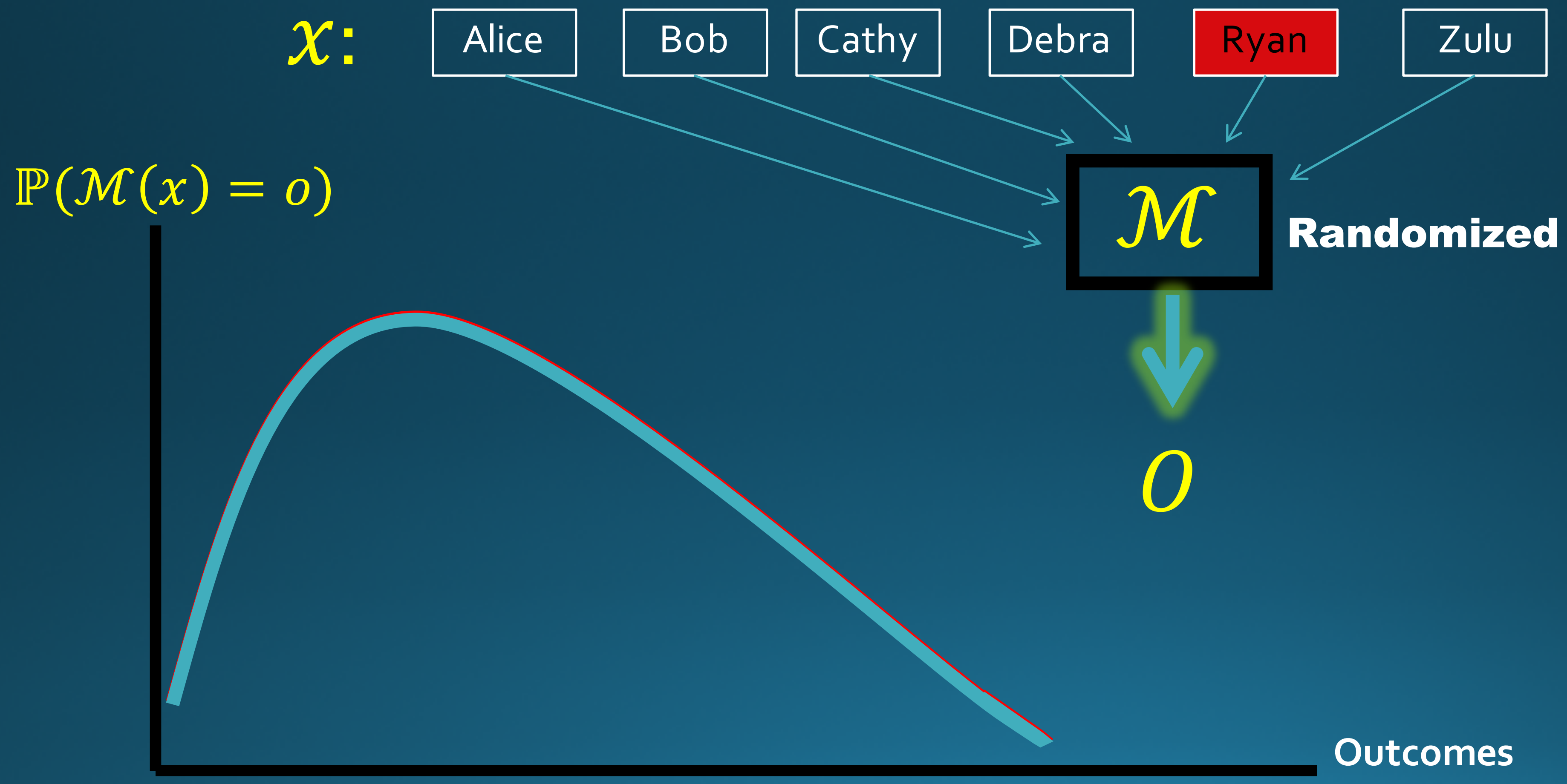
What procedures $\mathcal{M}$ have bounded max-info?

- [DFHPRR'15] Max-information bounds for:
  - (Pure) Differential Privacy – algorithmic stability condition.
  - Description Length – log(image size of $\mathcal{M}$)

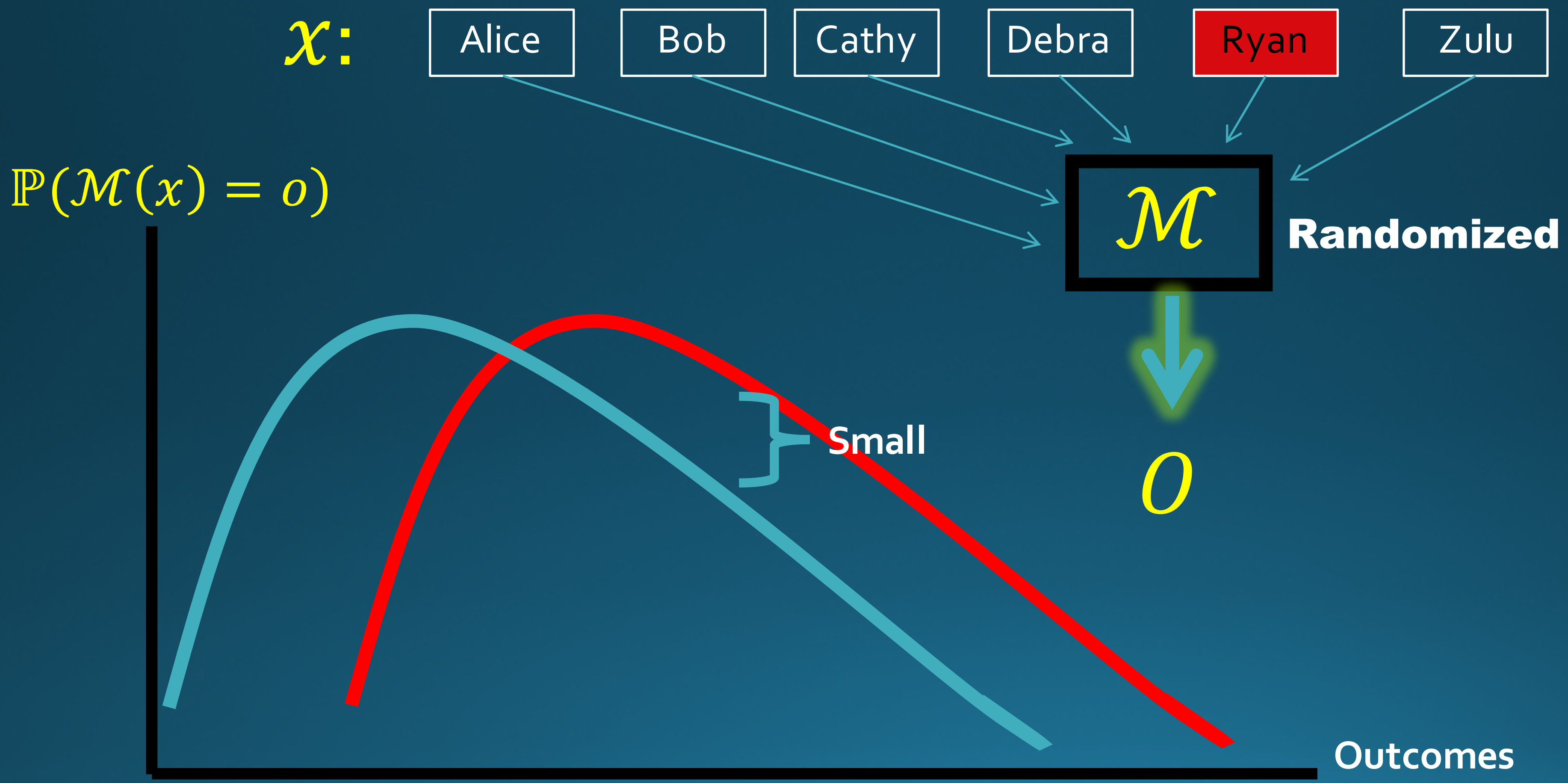
Differential Privacy [Dwork,McSherry,Nissim,Smith'06]



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



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



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

- A randomized algorithm $\mathcal{M}: \mathcal{D}^n \to \mathcal{O}$ is (ε, δ)-differentially private if for any neighboring data sets $x, x' \in \mathcal{X}^n$ and any outcome $S \subseteq \mathcal{O}$ we have

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

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

DP Composition

- If we run k many $(\epsilon, 0)$-DP algorithms $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_k$ on the same data set, then:
  - [DMNS'06]: The composed algorithm $\mathcal{M} = \mathcal{M}_k \circ \cdots \circ \mathcal{M}_1$ is:

$$(\epsilon k, 0)\text{-DP.}$$

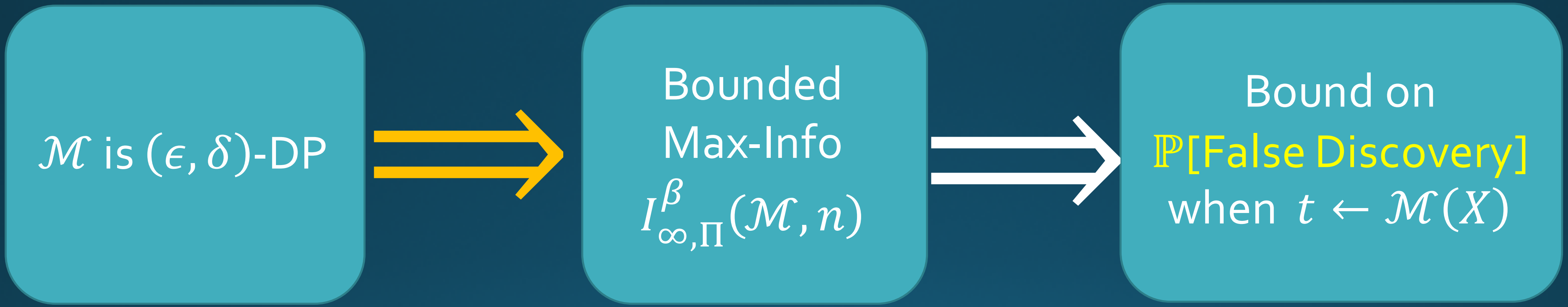
  - [Dwork,Rothblum,Vadhan'10]: The composed algorithm $\mathcal{M} = \mathcal{M}_k \circ \cdots \circ \mathcal{M}_1$ is:

$$\left(O\left(\epsilon\sqrt{k \log(1/\delta)}\right), \delta\right)\text{-DP.}$$

Quadratic Improvement with small $\delta > 0$!

Post-selection Hypothesis Testing

- [RRST'16] Connection between Max-Information and (approx)-differential privacy



Technical Contribution

- [DFHPRR'15] : If $\mathcal{M}: \mathcal{X}^n \rightarrow \mathcal{O}$ is $(\epsilon, 0)$-DP, then for $\beta >$

$$I_{\infty,\Pi}^{\beta}(\mathcal{M}; n) \leq \tilde{O}(\epsilon^2 n)$$

Similar Max-Info bounds

- [RRST'16]: If $\mathcal{M}: \mathcal{X}^n \rightarrow \mathcal{O}$ is (ϵ, δ)-DP, then

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

Consequences of Positive Result

Theorem: If $\mathcal{M}: \mathcal{X}^n \to \mathcal{O}$ is (ϵ, δ)-DP, then

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

- 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: $\mathcal{M}_1, \ldots, \mathcal{M}_k$
  - [DFHPRR'15]: $I^{\beta}_{\infty,\Pi}(\mathcal{M}_k \circ \cdots \circ \mathcal{M}_1; n) \le \tilde{O}(n\epsilon^2 k^2)$
  - [**R**RST'16]: $I^{\beta}_{\infty,\Pi}(\mathcal{M}_k \circ \cdots \circ \mathcal{M}_1; n) \le \tilde{O}(n\epsilon^2 k)$

Via strong composition theorem from [DRV'10]

Outline

- Post Selection Hypothesis Testing [**R**,Roth,Smith,Thakkar FOCS'16]
  - Connection between Max-Info and Differential Privacy
- DP composition with adaptively selected parameters [**R**,Roth,Ullman,Vadhan NIPS'16]
  - Privacy Odometers and Filters
- Private Hypothesis Tests [Gaboardi,Lim,**R**,Vadhan ICML'16],[Kifer,**R** AISTATS'17]
  - Chi-Square Tests
- Directions for Future Work

DP Composition

Data: x

$\mathcal{M}_1$ is ϵ_1-DP

a_1

$\mathcal{M}_2$ is ϵ_2-DP

a_2

$\vdots$

$\mathcal{M}_k$ is ϵ_k-DP

a_k

$$f(\epsilon_1, \cdots, \epsilon_k) \leq \epsilon_g$$

$\Rightarrow \mathcal{M}_k \circ \cdots \circ \mathcal{M}_1$ is ϵ_g-DP

DP Composition

Data: x

$\epsilon_1, \cdots, \epsilon_k$

$\mathcal{M}_1$ is ϵ_1-DP

a_1

$\mathcal{M}_2$ is ϵ_2-DP

a_2

$\vdots$

$\mathcal{M}_k$ is ϵ_k-DP

a_k

$$f(\epsilon_1, \cdots, \epsilon_k) \leq \epsilon_g$$

$$\Rightarrow \mathcal{M}_k \circ \cdots \circ \mathcal{M}_1 \text{ is } \epsilon_g\text{-DP}$$

DP Composition - Adaptive Privacy Parameters

- **Our focus**: Allow the analyst to allocate his privacy budget adaptively – also adaptively select the number of analyses.
  - Natural to allow the analyst to select parameters AND analyses adaptively – different DP analyses have different utility vs. privacy tradeoffs.
- **Questions**:
  - Which composition theorems still apply when we can select the parameters adaptively?
  - How can we even define differential privacy in this adaptively parameter setting?

Privacy Loss Random Variable

- Privacy loss for neighboring x, x' and for an algorithm $\mathcal{M}: \mathcal{X}^n \to \mathcal{O}$:

$$L(o) = \log\left(\frac{\mathbb{P}(\mathcal{M}(x)=o)}{\mathbb{P}(\mathcal{M}(x')=o)}\right) \text{ where } o \sim \mathcal{M}(x)$$

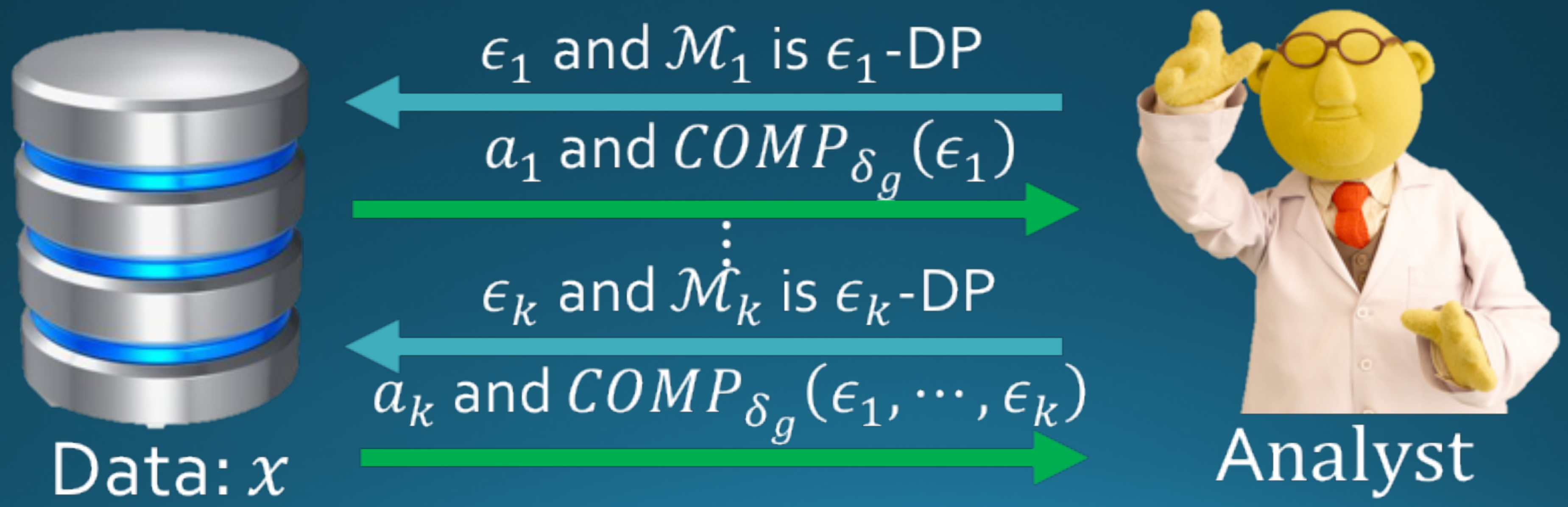
- Each round $i = 1, 2, \cdots, k$ the analyst selects $\epsilon_i \geq 0$ and an ϵ_i-DP algorithm $\mathcal{M}_i$ based on previous outcomes in an adversarial way.

$$L(o_1, \cdots, o_k) = \sum_{i=1}^{k} L_i(o_i) = \sum_{i=1}^{k} \log\left(\frac{\mathbb{P}(\mathcal{M}_i(x) = o_i \mid o_1, \cdots, o_{i-1})}{\mathbb{P}(\mathcal{M}_i(x') = o_i \mid o_1, \cdots, o_{i-1})}\right)$$

Privacy Odometer [R,Roth,Ullman,Vadhan'16]

- Privacy odometer provides a running upper bound on privacy loss.
- A valid privacy odometer $COMP_{\delta_g}: \mathbb{R}^* \to \mathbb{R}$ where an analyst selects $\epsilon_1, \cdots, \epsilon_k$ adaptively and w.p. $\geq 1 - \delta_g$

$$|L(o_1, \cdots, o_k)| \leq COMP_{\delta_g}(\epsilon_1, \cdots, \epsilon_k)$$



Privacy Odometer Results [RRUV'16]

- Basic composition applies – for any $\delta_g \geq 0$, the following is a valid privacy odometer:

$$COMP_{\delta_g}(\epsilon_1, \cdots, \epsilon_k) = \sum_{i=1}^{k} \epsilon_i$$

- For $\sum_i^k \epsilon_i^2 \geq \frac{1}{n^2}$, the following is a valid privacy odometer for $\delta_g > 0$:

$$COMP_{\delta_g}(\epsilon_1, \cdots, \epsilon_k) = O\left(\sqrt{\sum_{i=1}^{k} \epsilon_i^2 \log\left(\frac{\log(n)}{\delta_g}\right)}\right)$$

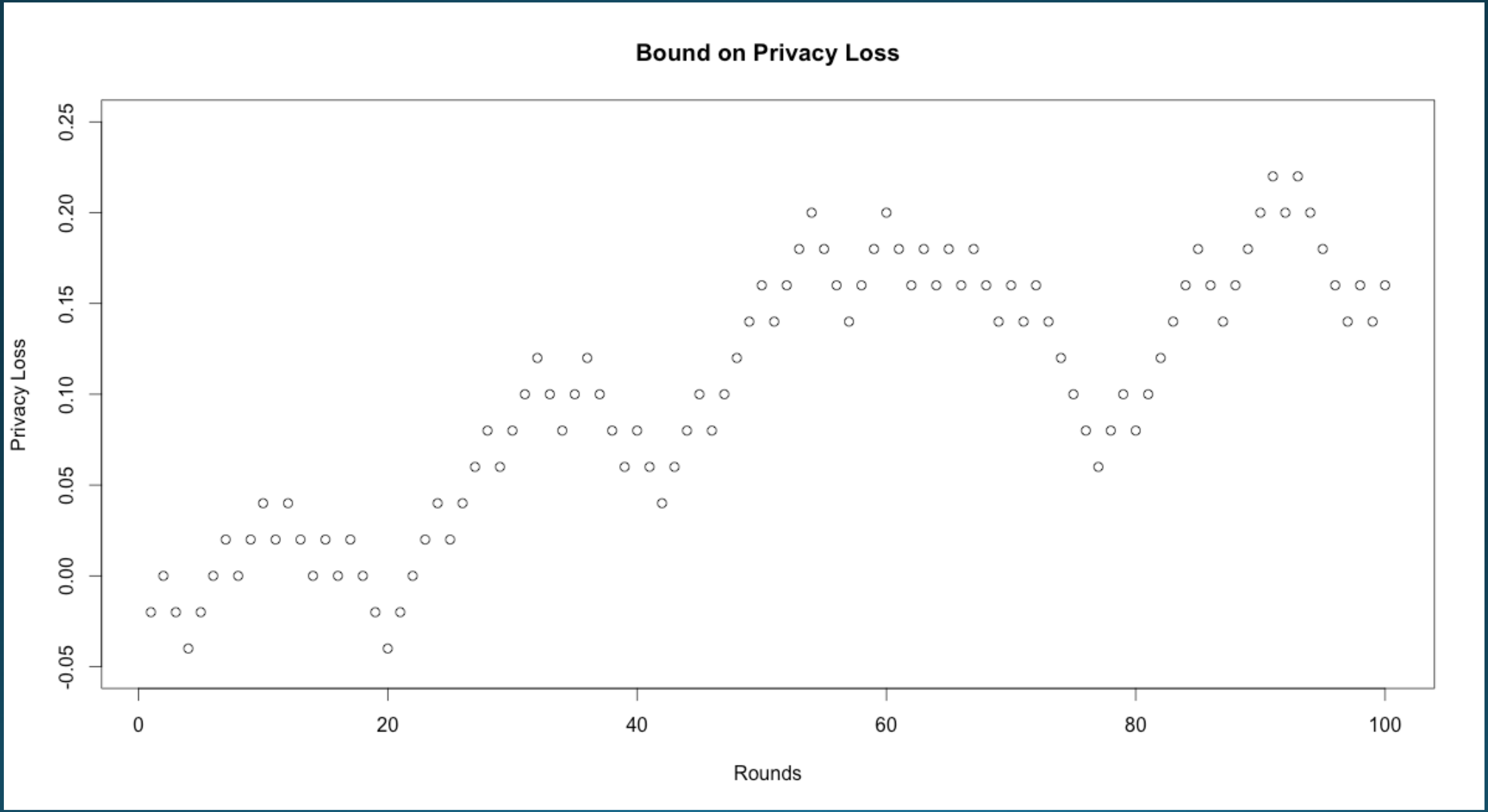
- There is no valid privacy odometer with $k > n$ and

$$COMP_{\delta_g}(\epsilon_1, \cdots, \epsilon_k) = o\left(\sqrt{\sum_{i=1}^{k} \epsilon_i^2 \log\left(\frac{\log(n)}{\delta_g}\right)}\right)$$

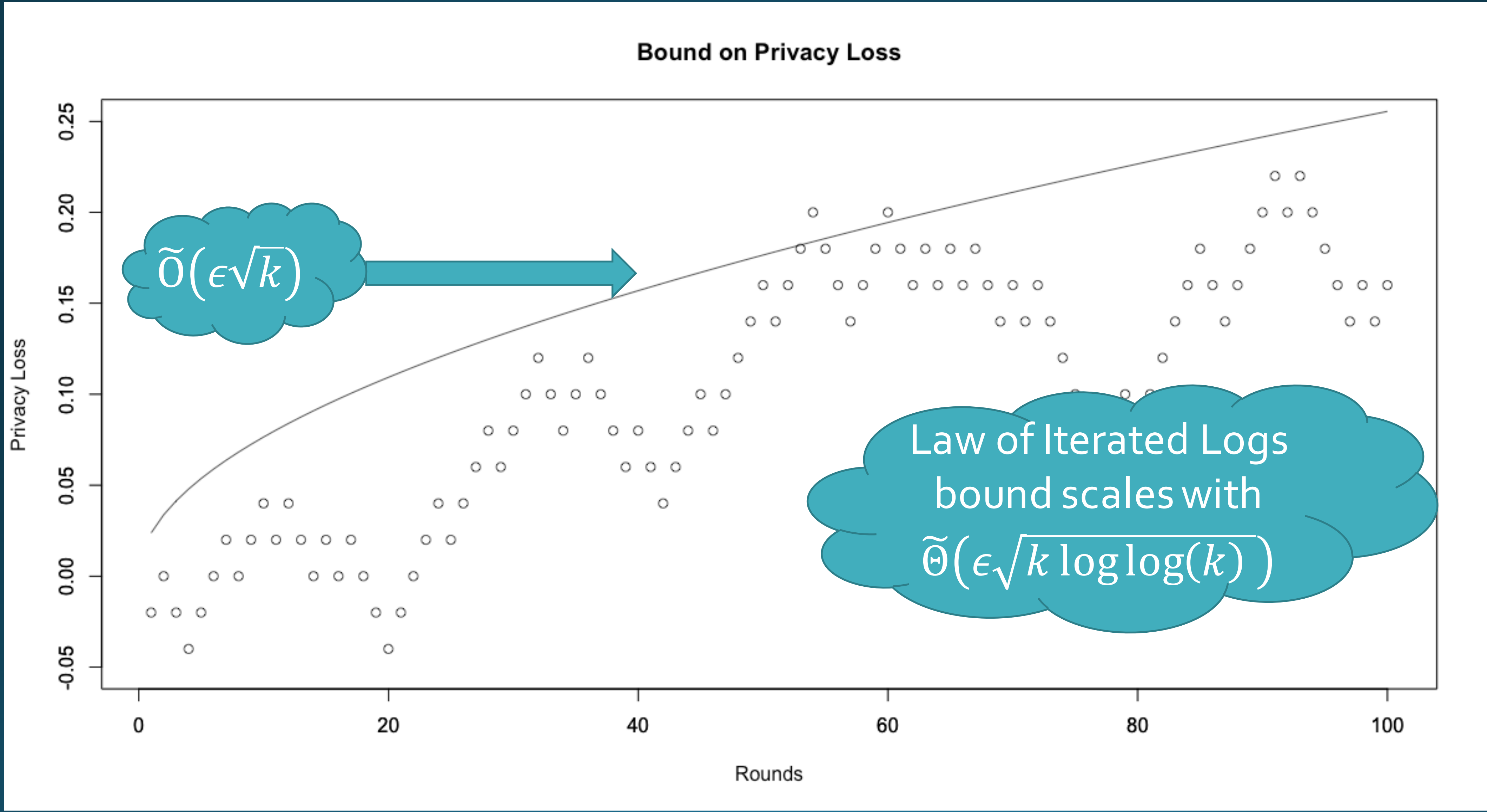
Proof Sketch

- Privacy Loss is a "biased" random walk with step size $\pm\epsilon$:

$$L(o_1, \cdots, o_k) = \sum_{i=1}^{k} L_i(o_i) = \sum_{i=1}^{k} \log\left(\frac{\mathbb{P}(\mathcal{M}_i(x) = o_i \mid o_1, \cdots, o_{i-1})}{\mathbb{P}(\mathcal{M}_i(x') = o_i \mid o_1, \cdots, o_{i-1})}\right)$$



Bound on Privacy Loss
Privacy Loss
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80
100
Rounds


Bound on Privacy Loss
Privacy Loss
Rounds
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80
100
$\tilde{O}(\epsilon\sqrt{k})$
Law of Iterated Logs bound scales with $\tilde{\Theta}(\epsilon\sqrt{k \log\log(k)})$

Outline

- Post Selection Hypothesis Testing [**R**,Roth,Smith,Thakkar FOCS'16]
  - Connection between Max-Info and Differential Privacy
- DP composition with adaptively selected parameters [**R**,Roth,Ullman,Vadhan NIPS'16]
  - Privacy Odometers and Filters
- Private Hypothesis Tests [Gaboardi,Lim,**R**,Vadhan ICML'16],[Kifer,**R** AISTATS'17]
  - Chi-Square Tests
- Directions for Future Work

Differentially Private Hypothesis Tests

- DP analyses ensure statistical validity over adaptive sequences of analyses.
- Thus, we aim to develop analyses which are each DP and produce valid conclusions.
- **GOAL**: Valid hypothesis testing while preserving privacy.

Classical Hypothesis Testing

- Want to design a test $t: \mathcal{X}^n \rightarrow \{Inconclusive, Reject\ H_0\}$ s.t.:

	H_0 is True	H_0 is False
t rejects	α = False Discovery	Power
t fails to reject	1-α = Significance	Type II Error

- Want to ensure test has $\mathbb{P}$[**False Discovery**] ≤α and has good **Power**.

Chi-Square Tests

- Categorical data X. Histogram: $D = (D_1, \dots, D_d) \sim Multinomial(n, \vec{p})$.

1. **Goodness of Fit**: $H_0 : \vec{p} = \vec{p}^0$
   - Simple Test - data distribution completely determined
2. **Independence Test**: $H_0 : Y^{(1)} \perp Y^{(2)}$
   - Composite Test – data distribution not completely determined

- Both classical tests use the Chi-Squared Statistic:

$$Q^2 = \sum \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

Private Goodness of Fit

- Add noise to each cell count to preserve differential privacy.
- Form the private chi-squared statistic:

$$Q_{DP}^2 = \sum_{i=1}^{d} \frac{\left(D_i + Z_i - np_i^0\right)^2}{np_i^0}$$ where we use either:

$$Z_i \sim N\left(0, O\left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)\right)$$ for $(\varepsilon, \delta) - DP$

or $Z_i \sim Lap\left(O\left(\frac{1}{\varepsilon}\right)\right)$ for $(\varepsilon, 0) - DP$

Naïve Approach – Use Classical Test

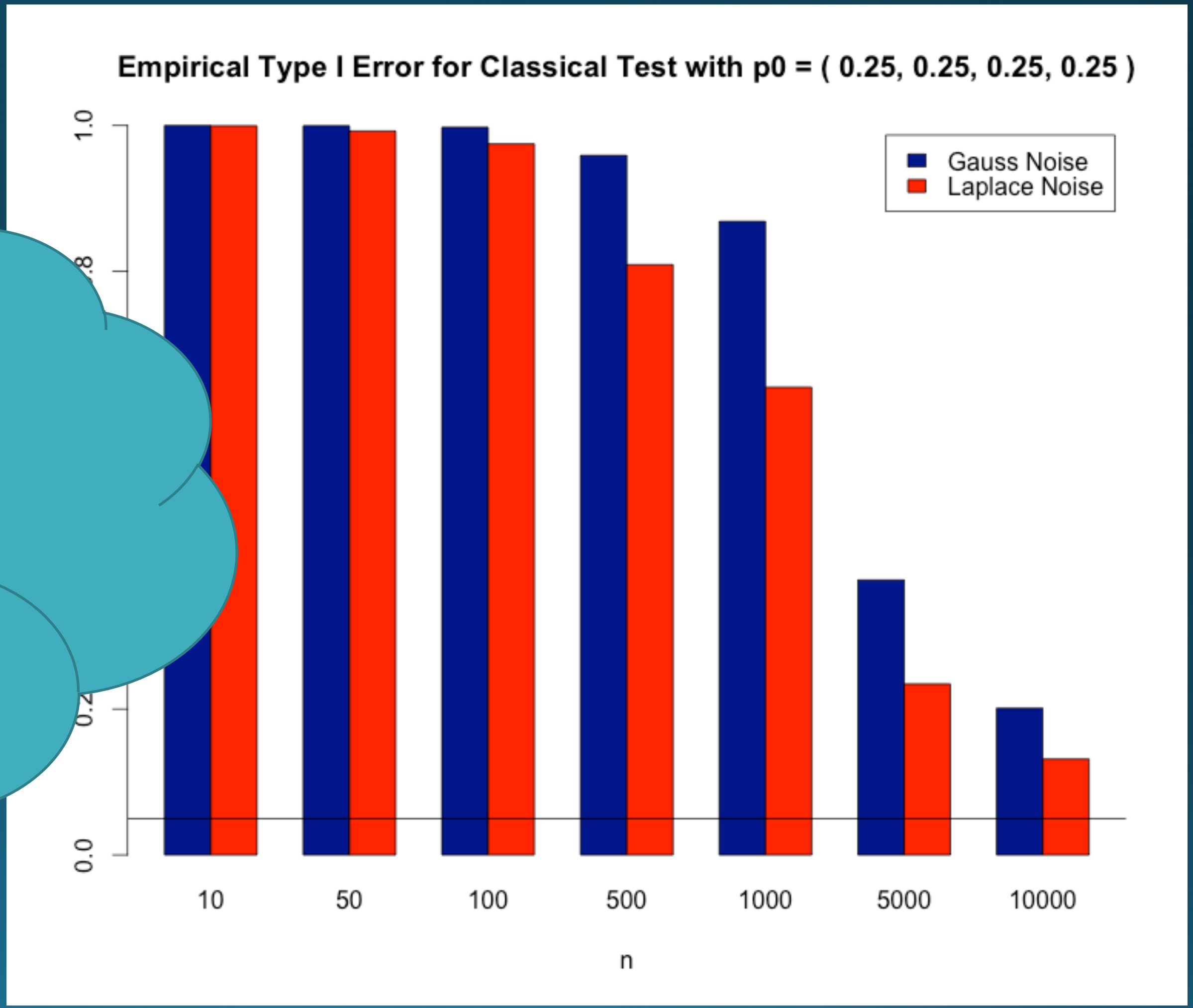
Noise is small as $n \to \infty$

[Johnson and Shmatikov '13],
[Vu and Slavkovic '09]

Se

H_0, c

Similar findings shown in [Fienberg, Rinaldo, Yang '10], [Karwa and Slavkovic '12, '16]

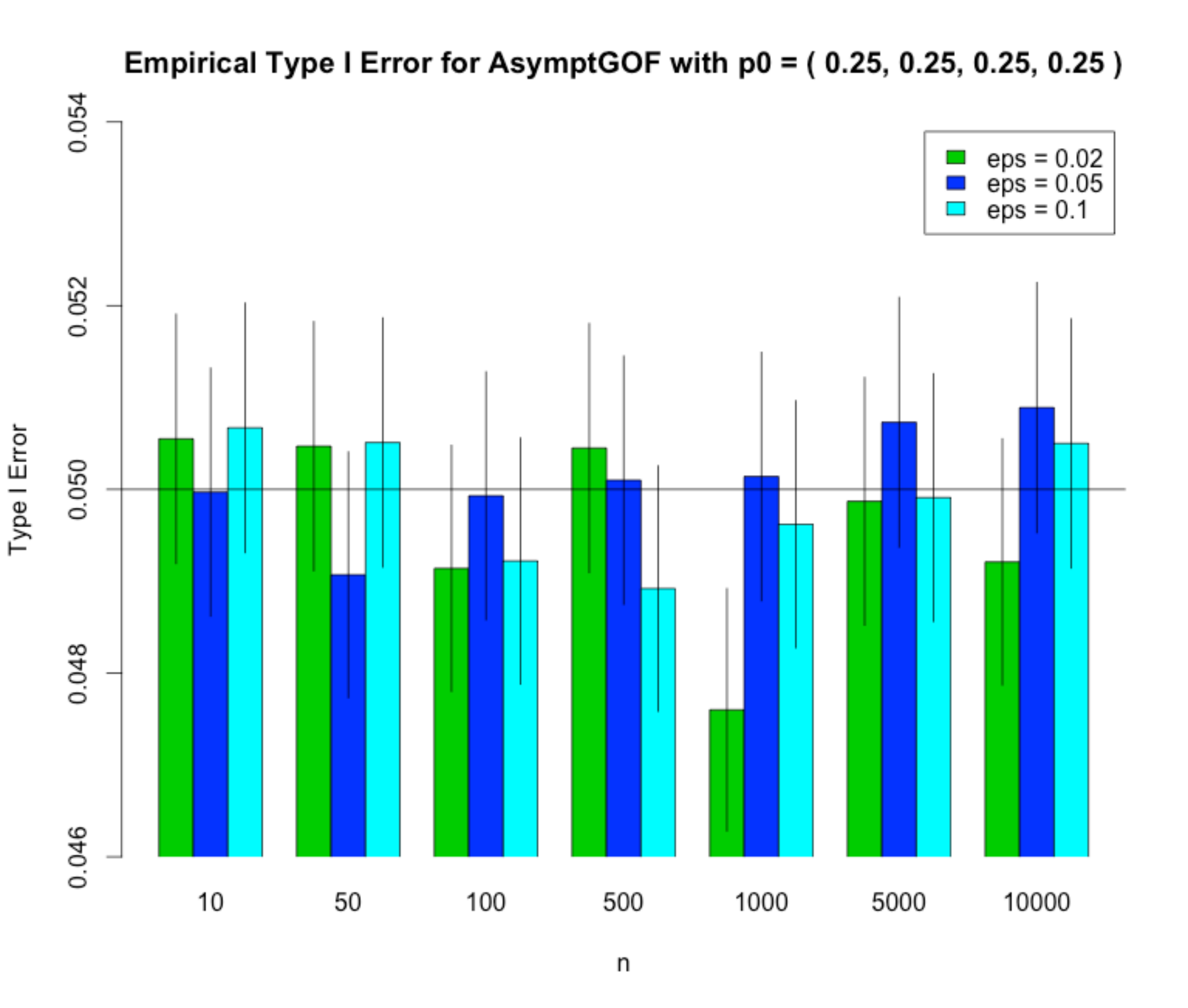


AsymptGOF – Private GOF Test [GLRV'16]

Take the (Gaussian) noise into account

Set $\delta = 10^{-6}$

Plotting the proportion of 100,000 trials that rejected H_0, despite it being true.

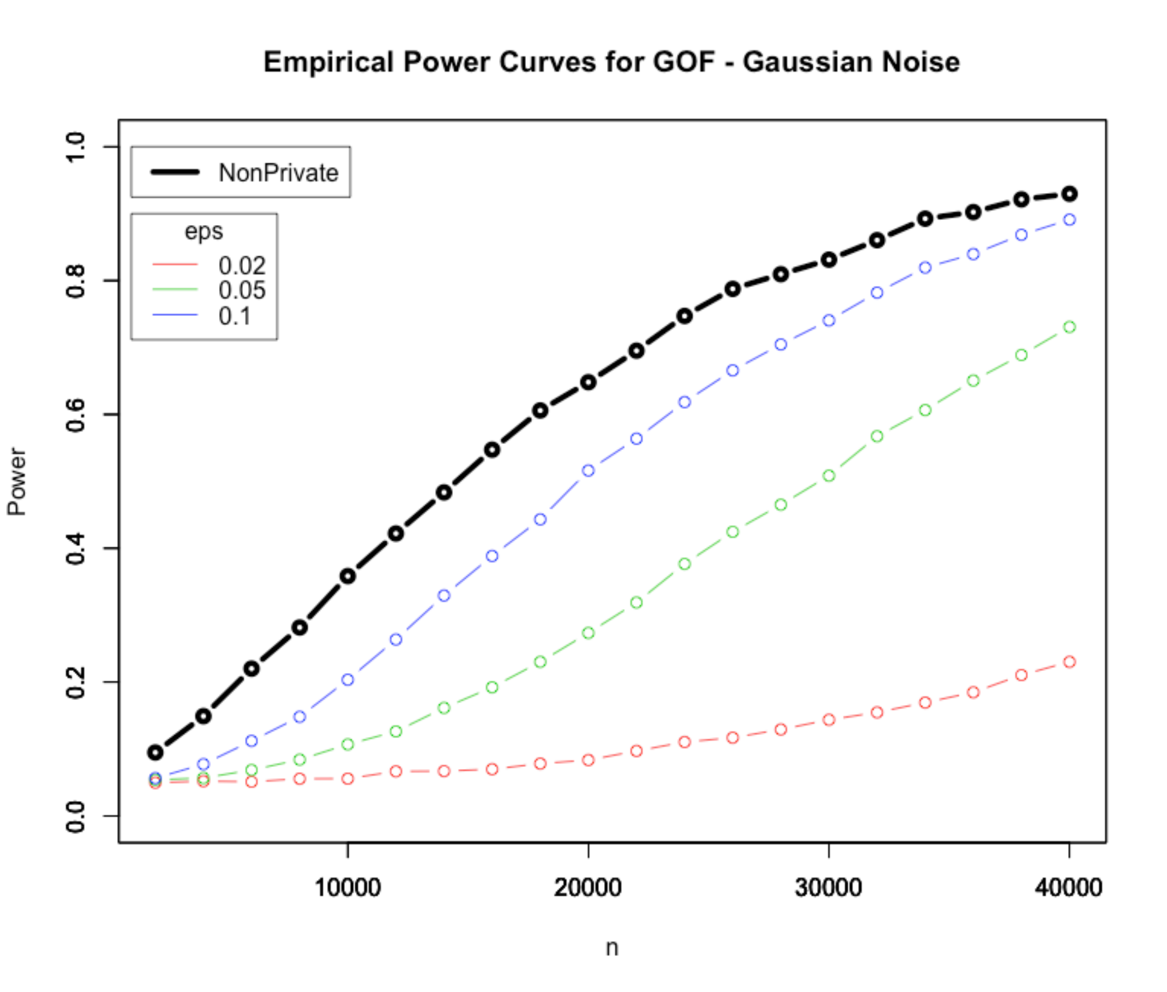


AsymptGOF – Private GOF Test [GLRV'16]

- Set $\delta = 10^{-6}$
- Test $\vec{p}^0 = \left(½, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$
- Generate data from $\vec{p}^1 =$

$$\vec{p}^0 + 0.01\left(1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}\right)$$

- Plot the proportion of 10,000 trials that correctly rejected null.

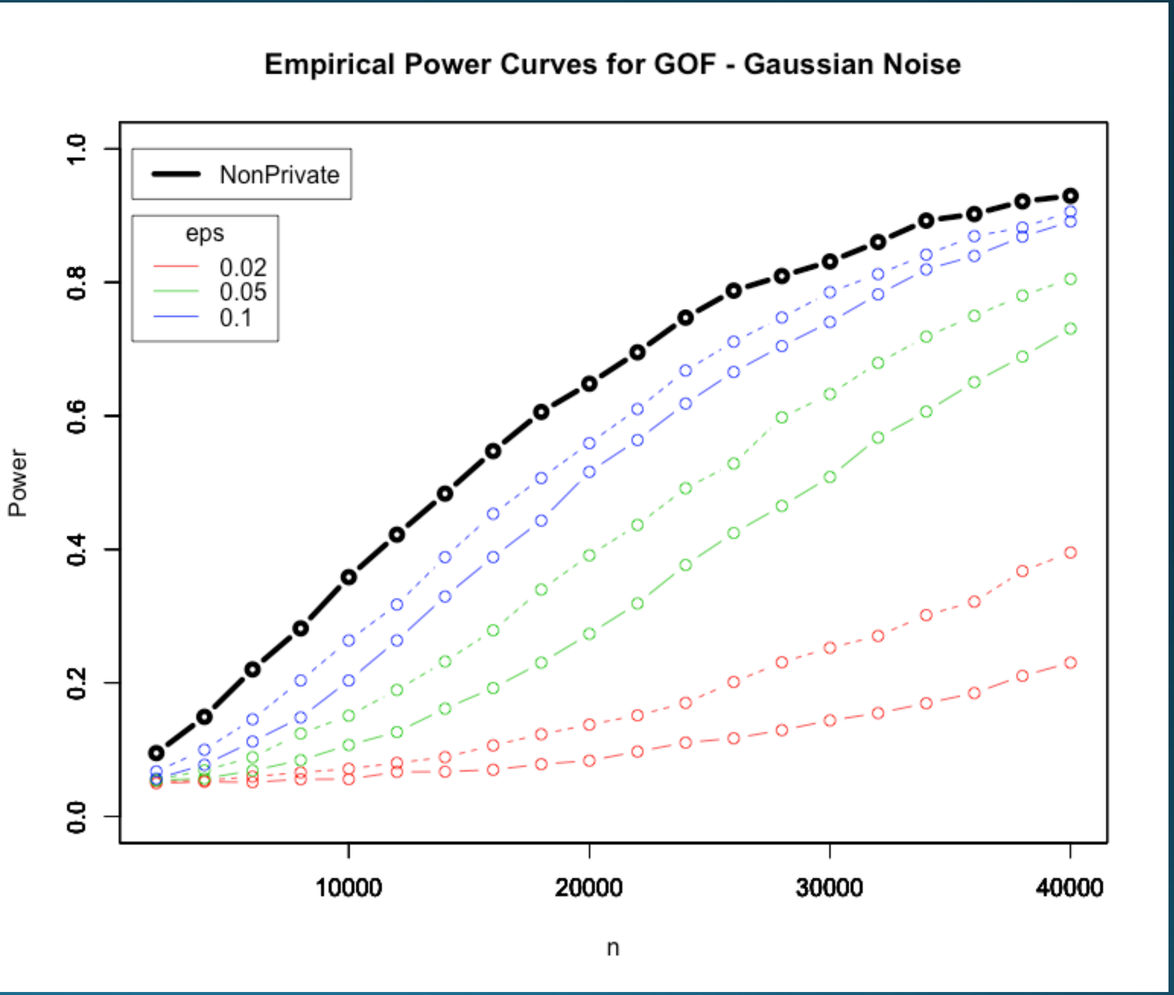


AsymptGOF vs NewStatAsymptGOF [KR'17]

- Set $\delta = 10^{-6}$
- Test $\vec{p}^0 = \left(½, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$
- Generate data from $\vec{p}^1 =$

$$\vec{p}^0 + 0.01\left(1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}\right)$$

- Plot the proportion of 10,000 trials that correctly rejected null.



Outline

- Post Selection Hypothesis Testing [**R**,Roth,Smith,Thakkar- FOCS'16]
  - Connection between Max-Info and Differential Privacy
- DP composition with adaptively selected parameters [**R**,Roth,Ullman,Vadhan- NIPS'16]
  - Privacy Odometers and Filters
- Private Hypothesis Tests [Gaboardi,Lim,**R**,Vadhan- ICML'16],[Kifer,**R** AISTATS'17]
  - Chi-Square Tests
- Directions for Future Work

Directions for Future Work

- Adaptive Data Analysis:
  - We do not fully understand adaptive data analysis.
  - Is there a unifying measure in adaptive data analysis?
  - Is differential privacy the right approach?
- Develop private hypothesis tests that incorporate the noise:
  - Local model.
  - Other tests, e.g. ANOVA and Regression.

Contributions

Thanks!

- **Adaptive Data Analysis:**
  - [**R**RST- FOCS'16] – Information and privacy
  - [**R**RUV- NIPS'16] – Adaptive parameter composition
  - [GL**R**V- ICML'16], [K**R**- AISTATS'17] – Private hypothesis tests
- **Algorithmic Game Theory:**
  - Incorporate privacy as a constraint:
    - [Kannan,Morgenstern,**R**,Roth- EC'15] - private allocations in kidney exchanges.
  - Leverage stability of DP to solve new problems
    - **[Kearns,Pai,R**,Roth,Ullman'15a**]** **,** [**R**,Roth,Ullman,Wu- EC'15] **–** Coordinate agents with incomplete information to desirable strategies
  - [Hsu,Morgenstern,**R**,Roth,Vohra-STOC'16] - Prices as coordination devices
  - [Jabbari,**R**,Roth,Wu-NIPS'16] - Revealed preferences
  - [Dudik,Lahaie,**R**,Vaughan'17] – Prediction markets