

Barriers to Preventing False Discovery in Interactive Data Analysis




Jonathan Ullman (Northeastern University)

Based on joint works with Moritz Hardt and Thomas Steinke,
and conversations with Adam Smith

False discovery occurs when you make conclusions based on your data that don't generalize to the population.



 OPEN ACCESS

ESSAY

1,140,912

VIEWS

1,413

CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: [10.1371/journal.pmed.0020124](https://doi.org/10.1371/journal.pmed.0020124)

Popular and academic articles report on an increasing number of false discoveries in empirical science.

Today

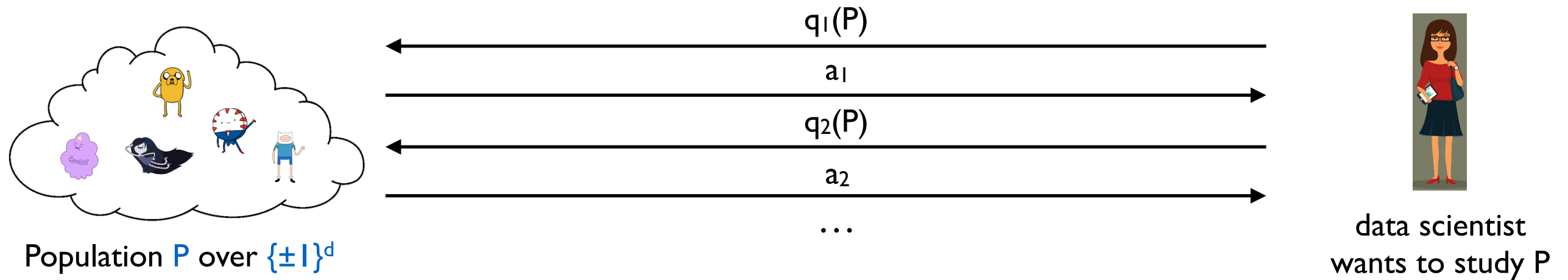
- Computational barriers to preventing false discovery in interactive data analysis
 - Computational hardness results
 - Information-theoretic (minimax) lower bounds
- An adversarial perspective on false discovery in interactive data analysis

Today

- Computational barriers to preventing false discovery in interactive data analysis
 - Computational hardness results
 - Information-theoretic (minimax) lower bounds
 - A language barrier?
- An adversarial perspective on false discovery in interactive data analysis

Step one: admit you have a problem...and formalize it.

Statistical Query Model (Kearns '93)



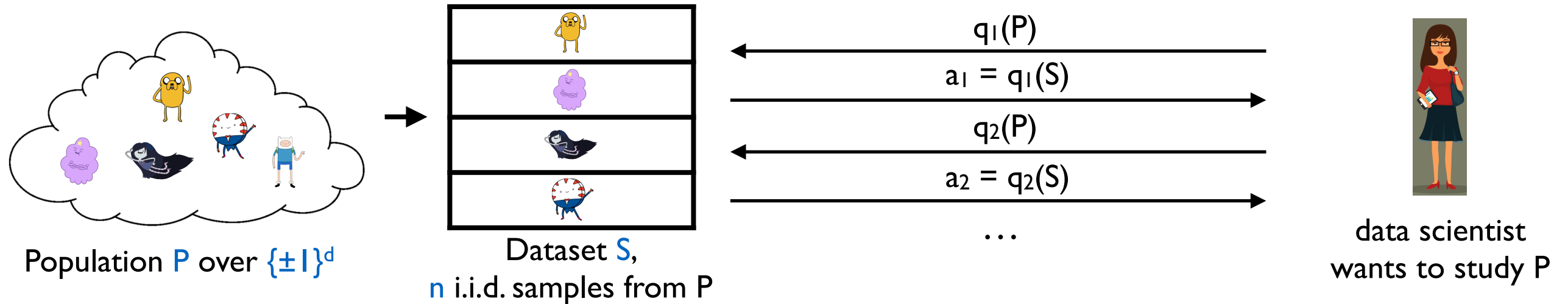
Goal: estimate the answers to k ,
adaptively chosen **statistical queries** on P

“false discovery” occurs when the answer is inaccurate

Statistical Queries

- specified by a **predicate** $q: \{\pm 1\}^d \rightarrow \{\pm 1\}$
- **true answer** $q(P) = \text{mean of } q \text{ over } P$
- an answer a is **accurate** if $|a - q(P)| \leq \epsilon$

Statistical Query Model (Kearns '93)



Goal: estimate the answers to k ,
adaptively chosen **statistical queries** on P

“false discovery” occurs when the answer is inaccurate

Statistical Queries

- specified by a **predicate** $q: \{\pm 1\}^d \rightarrow \{\pm 1\}$
- **true answer** $q(P) = \text{mean of } q \text{ over } P$
- an answer a is **accurate** if $|a - q(P)| \leq \epsilon$

What if we use the empirical answer
from the sample?

- **empirical answer** $q(S) = \text{mean of } q \text{ over } S$

Non-Interactive Queries are Easy

Easy Theorem (well known)

If the queries q_1, \dots, q_k are fixed before S is drawn, then whp over S

$$|q_i(S) - q_i(P)| \leq O\left(\sqrt{\frac{\log k}{n}}\right)$$

- Can answer nearly 2^n queries with non-trivial accuracy

Proof Sketch:

Apply your favorite tail bound for sums of independent random variables
+ Union Bound.

*Can fail spectacularly when the queries
can be chosen adaptively!*

Overfitting with Adaptive Queries

Fact

If the queries q_1, \dots, q_k can be chosen adaptively, then it could be that

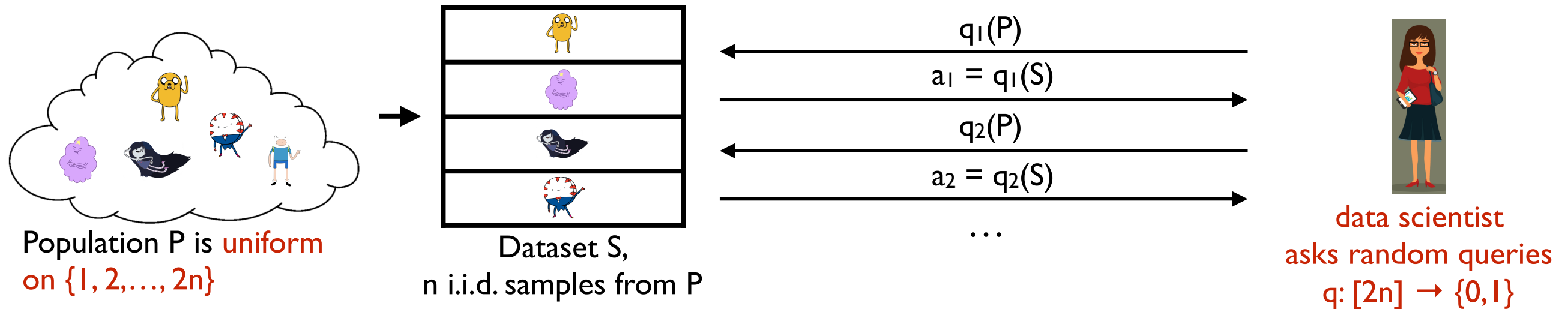
$$|q_i(S) - q_i(P)| > \Omega\left(\sqrt{\frac{k}{n}}\right)$$

- Cannot guarantee non-trivial accuracy for more than $k = O(n)$ queries.

Proof Sketch:

Next slide.

Overfitting with Adaptive Queries



Adversary can ask $O(n)$ random statistical queries, get the empirical answer to each one, then reconstruct S exactly.

Once you recover S exactly, ask the query

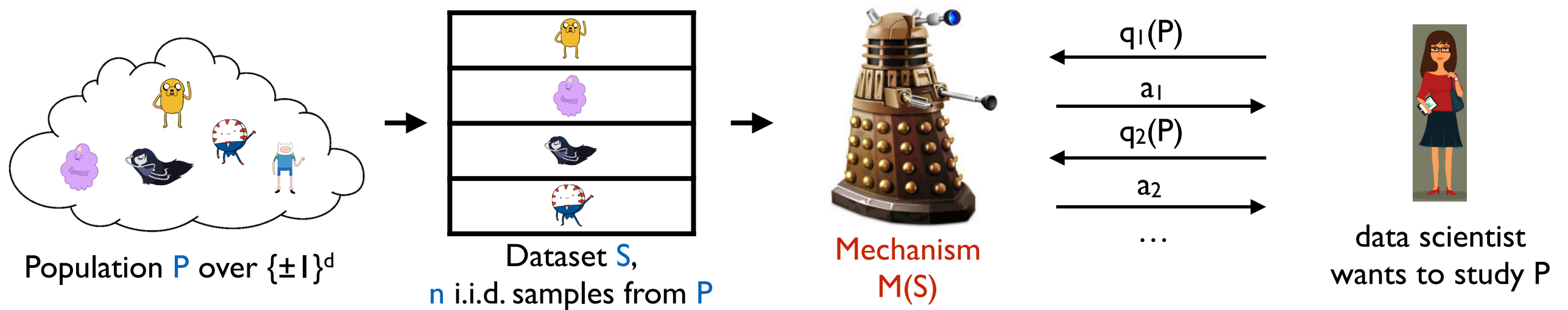
$$q(x) = \begin{cases} -1 & \text{if } x \in S \\ 1 & \text{if } x \notin S \end{cases}$$

Note, $q(S) - q(P) = 1$.

Can only answer $k \lesssim n$ queries!

Step two: appeal to a higher power for help.

Statistical Query Model (Kearns '93)



Goal: estimate the answers to k ,
adaptively chosen **statistical queries** on P

“false discovery” occurs when the answer is inaccurate

Statistical Queries

- specified by a **predicate** $q: \{\pm 1\}^d \rightarrow \{\pm 1\}$
- **true answer** $q(P) = \text{mean of } q \text{ over } P$
- **empirical answer** $q(S) = \text{mean of } q \text{ over } S$
- an answer a is **accurate** if $|a - q(P)| \leq \epsilon$
- M is **accurate** if, for every population P , every analyst, every sequence q_1, \dots, q_k , every answer a_1, \dots, a_k is accurate for P

Today's Goal: “universal mechanisms” to prevent false discovery

Differential Privacy and Adaptive Queries

Theorem (DFHPRR'15a, BNSSSU'15) Let M be a mechanism such that

1) M is ϵ -accurate with respect to the empirical answer

for every S , every adaptive sequence of k queries, q_1, \dots, q_k ,
 $M(S)$ answers with a_1, \dots, a_k such that $a_i = q_i(S) \pm \epsilon$ for $i=1, \dots, k$

2) M is (ϵ, δ) -differentially private for the sample

Then,

$$\Pr(\max_i |q_i(P) - a_i| \leq O(\epsilon)) \geq 1 - O(\delta/\epsilon).$$

*Step four: make a searching and fearless
inventory of our DP algorithms*

Differential Privacy and Adaptive Queries

Theorem (DFHPRR'15, BNSSSU'15) Let M be a mechanism such that

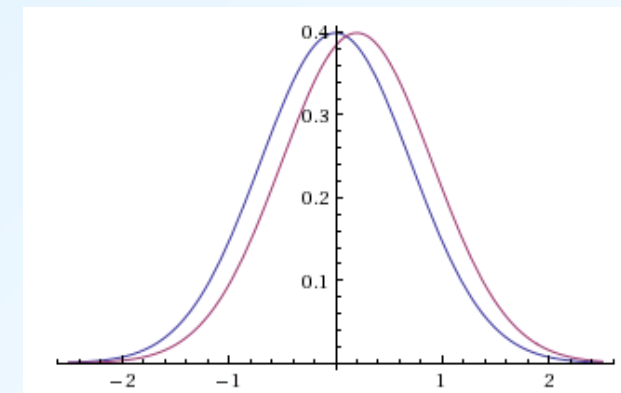
- 1) M is ϵ -accurate with respect to the empirical answer
- 2) M is (ϵ, δ) -differentially private for the sample.

Then, $\Pr(|q(P) - a| \leq O(\epsilon)) \geq 1 - O(\delta/\epsilon)$.

Gaussian Mechanism:

Answer $a_i(S) = q_i(S) + N(0, \epsilon^2)$, for $\epsilon \approx k^{1/4}/n^{1/2}$

- 1) is ϵ accurate wrt the empirical answer
- 2) is (ϵ, δ) -DP for negligible δ



$a_i(S)$ vs. $a_i(S')$

Corollary There is a simple, computationally efficient mechanism that is accurate for $k \gtrsim n^2$ queries.

Differential Privacy and Adaptive Queries

Theorem (DFHPRR'15, BNSSSU'15) Let M be a mechanism such that

- 1) M is ϵ -accurate with respect to the empirical answer
- 2) M is (ϵ, δ) -differentially private for the sample.

Then, $\Pr(|q(P) - a| \leq O(\epsilon)) \geq 1 - O(\delta/\epsilon)$.

Private Multiplicative Weights (HR'10)

There exists a $(1/100, \delta)$ -DP algorithm that is $(1/100)$ -accurate wrt to the empirical answer for $k \gtrsim \exp(n/d^{1/2})$ adaptively chosen queries. The mechanism runs in time polynomial in n , 2^d per query.

Corollary There is an accurate mechanism for $k \gtrsim \exp(n/d^{1/2})$ queries that runs in time polynomial in n , 2^d per query.

Step seven: ask the higher power to remove our shortcomings

Negative Results

Theorem (Computational Version) (HU'14, SU'15):

If one-way functions exist and $d = \omega(\log n)$, there is no computationally efficient mechanism* that answers more than $k = \tilde{O}(n^2)$ arbitrary adaptively chosen queries.

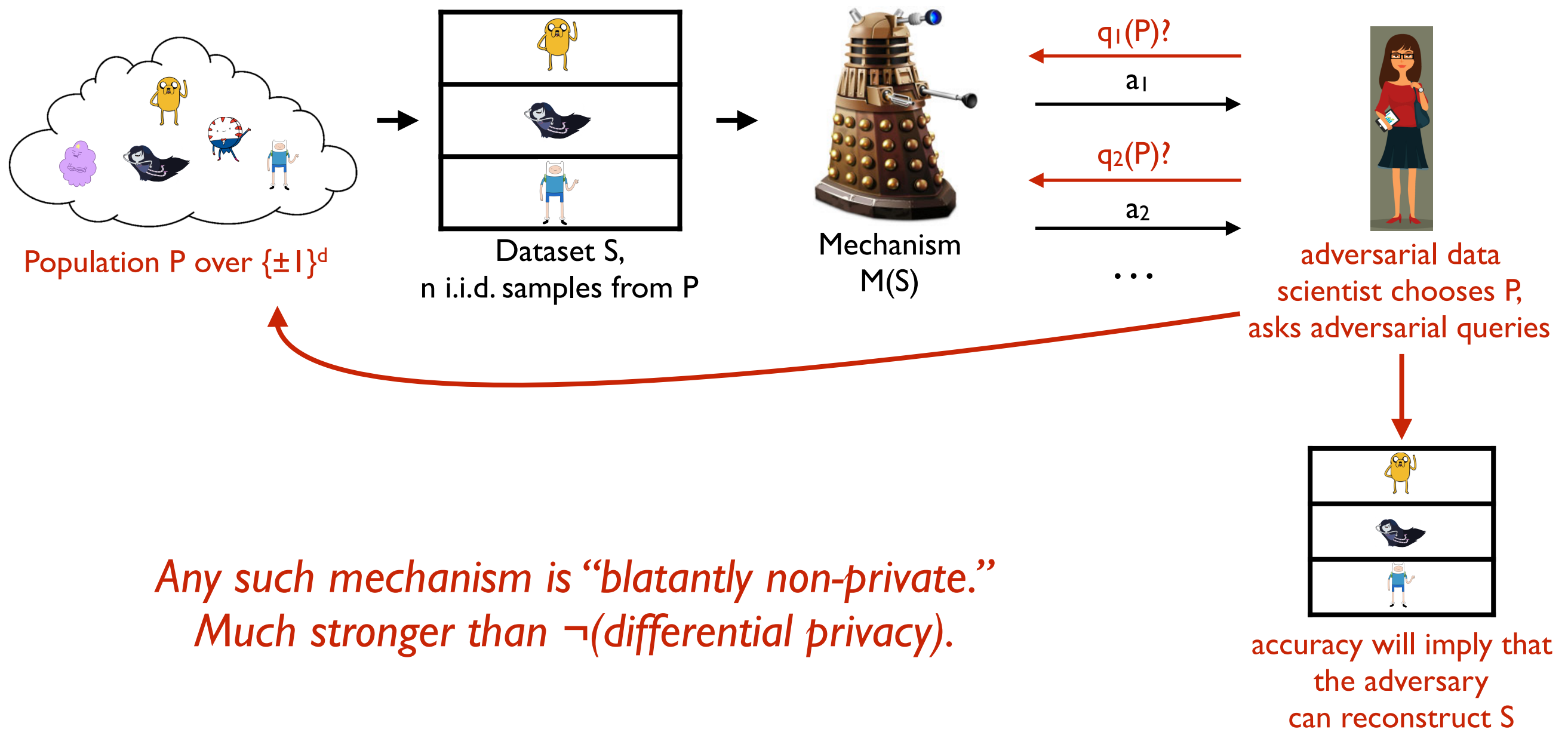
Theorem (Information-Theoretic Version) (HU'14, SU'15):

If $d > k$, then there is no mechanism, efficient or not, that answers more than $k = \tilde{O}(n^2)$ arbitrary adaptively chosen queries.

- Universal mechanisms are severely limited
- A “full employment theorem” for we who prevent false discovery!
 - Preventing false discovery will require detailed understanding

*computationally efficient \approx answers each query in time polynomial in $|S|=nd$

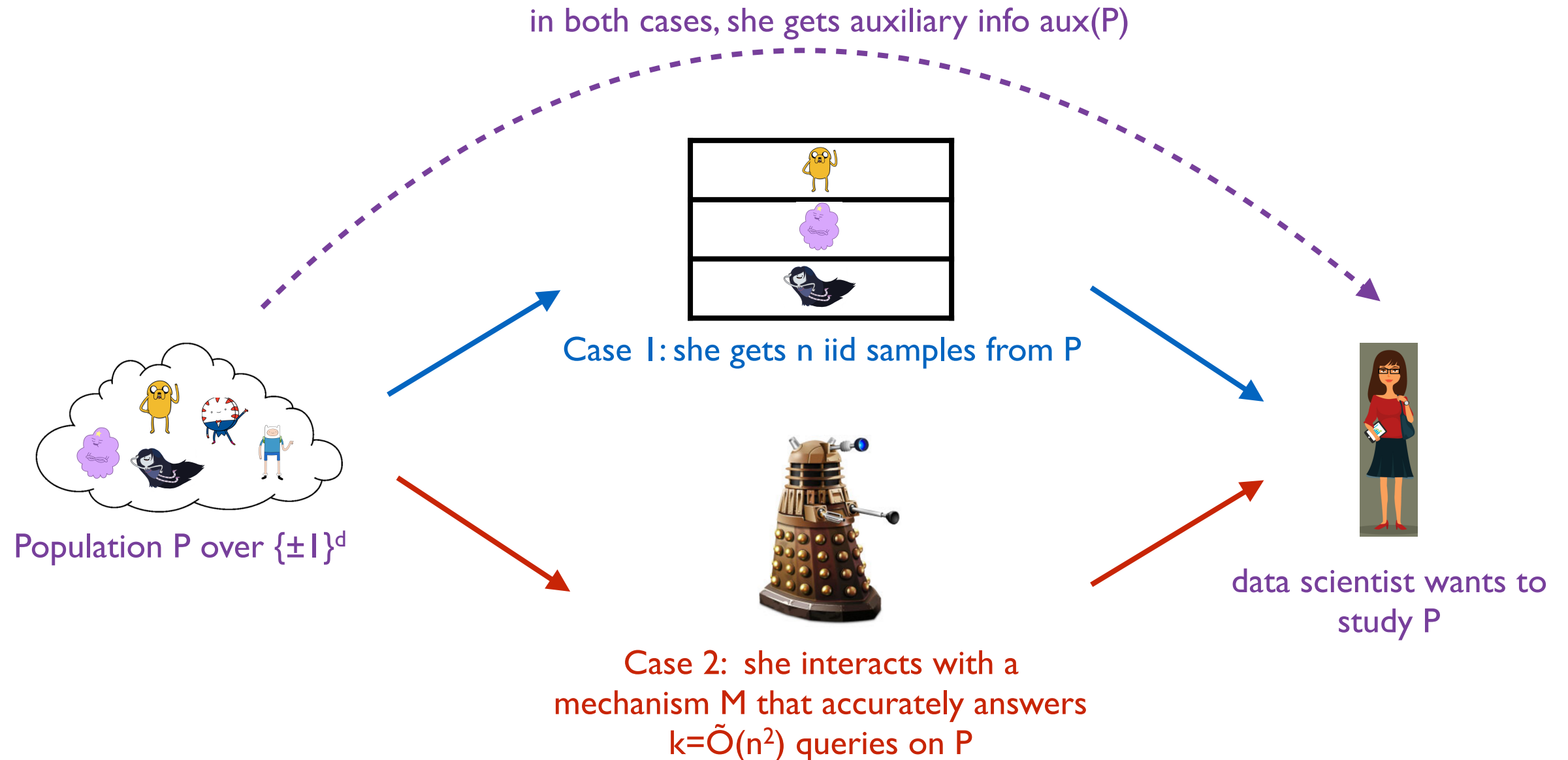
Our Approach (v.1): Blatant Non-Privacy



Once she has S , she can ask a “killer” query such that $|q(P) - q(S)|$ is large.

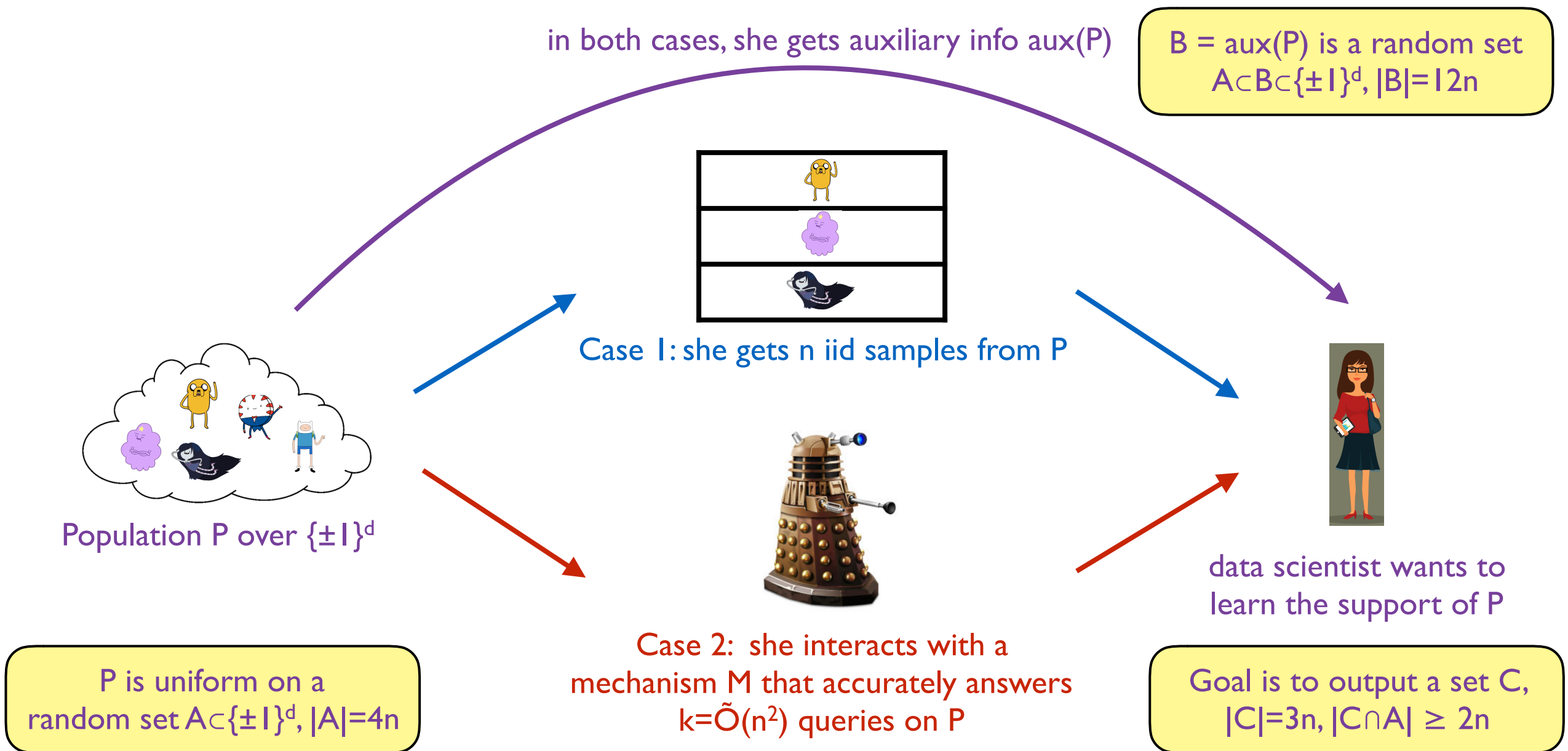
Not as trivial as it sounds, but I wouldn't call it non-trivial.

Our Approach (v.2): Estimation with Auxiliary Info



Approach: find a problem that she can solve in case 2, but not in case 1
 \implies cannot implement the mechanism given n samples.

Our Approach (v.2): Estimation with Auxiliary Info



- Case 1: $\Pr[\text{she succeeds}] \leq \exp(-n/100)$. Probability is over A, B, C
- Case 2: If M is computationally efficient, or $d > k$, $\Pr[\text{she succeeds}] \approx 1$. Probability is over A, B, C, M (hard half)

Negative Results

Theorem (Computational Version) (HU'14, SU'15):

If secure crypto exists* and $n = 2^{o(d)}$
then there is no computationally efficient mechanism*
that gives accurate answers* to more than $k = O(n^2)$ arbitrary adaptively
chosen queries.

Theorem (Information-Theoretic Version) (HU'14, SU'15):

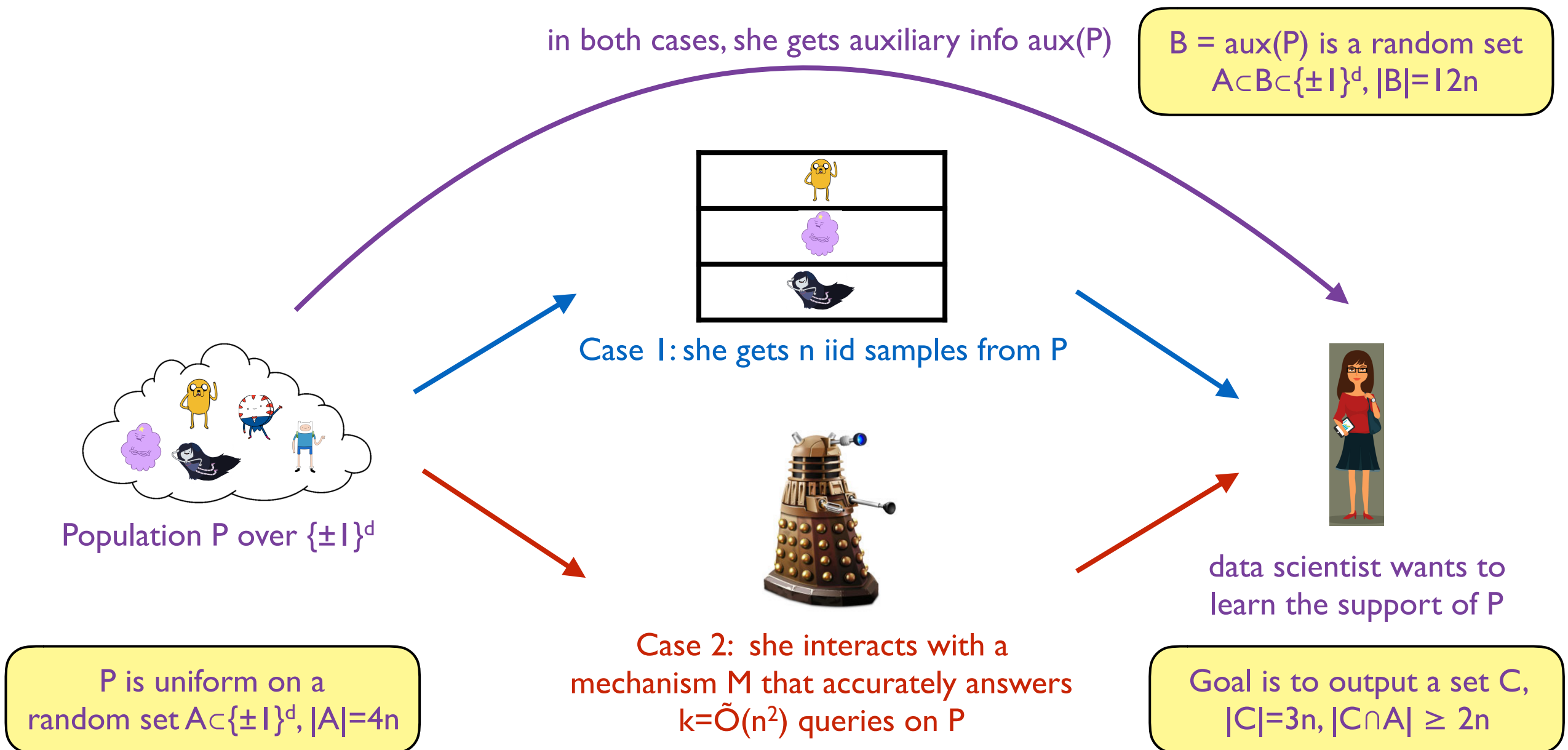
If $d > k$,
then there is no mechanism, efficient or not,
that gives accurate answers* to more than $k = O(n^2)$ arbitrary adaptively
chosen queries.

*secure crypto \approx exponentially hard one-way functions

*computationally efficient \approx answers each query in time polynomial in $|S|=nd$

*accurate answers \approx can distinguish $q(P) = 1$ from $q(P) = 0$ (very weak!)

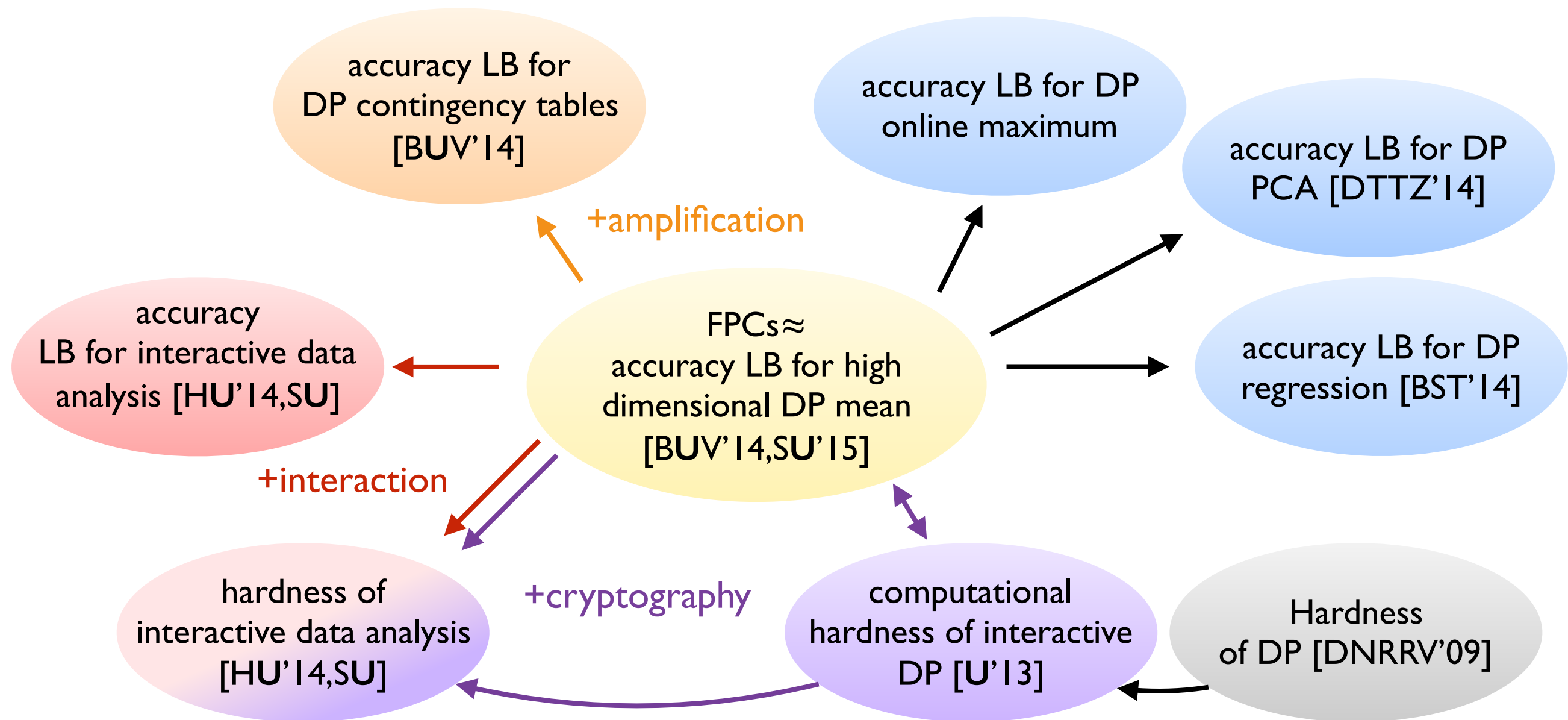
Our Approach (v.2): Estimation with Auxiliary Info



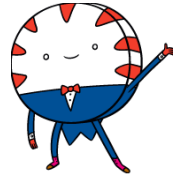
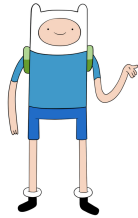
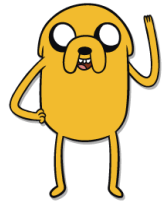
- Both approaches are very tailored to universal mechanisms.
 - Queries to the oracle are complex
 - Scientist gets auxiliary info that is unknown to the mechanism
- Open question: Can we prove negative results that don't rely on “secret” auxiliary information.

Main Tool: Fingerprinting Codes (Boneh-Shaw'95, Tardos'03)

A versatile tool for understanding the limits of learning in high dimensions.



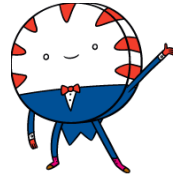
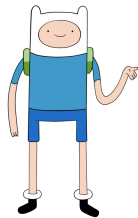
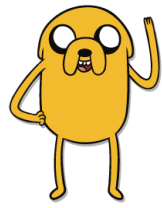
Main Tool: Fingerprinting Codes (Boneh-Shaw'95, Tardos'03)



I want to preview my new
movie: “The Fault in Our
Statistics”

...but I’m worried
about piracy

Main Tool: Fingerprinting Codes (Boneh-Shaw'95, Tardos'03)








(Gen, Trace)

Gen outputs
N patterns of
watermarks




Critics form
a coalition

Coalition releases
a pirated film

Trace outputs a
colluder in S

| | | | | | |
|--|---|---|---|---|---|
|  | 1 | 0 | 0 | 0 | 1 |
|  | 1 | 0 | 1 | 0 | 0 |
|  | 1 | 1 | 0 | 0 | 0 |
|  | 1 | 0 | 0 | 0 | 0 |
|  | 1 | 1 | 1 | 0 | 1 |

N users
 $\tilde{O}(n^2)$ marks

| | | | | | |
|---|---|---|---|---|---|
|  | 1 | 0 | 0 | 0 | 1 |
| | | | | | |
|  | 1 | 1 | 0 | 0 | 0 |
| | | | | | |
|  | 1 | 1 | 1 | 0 | 1 |

coalition S
of size n

users = support of P
one col = one query

coalition =
users in sample

F^*

.9 .5 0 .2 .8

Trace(F^*) =



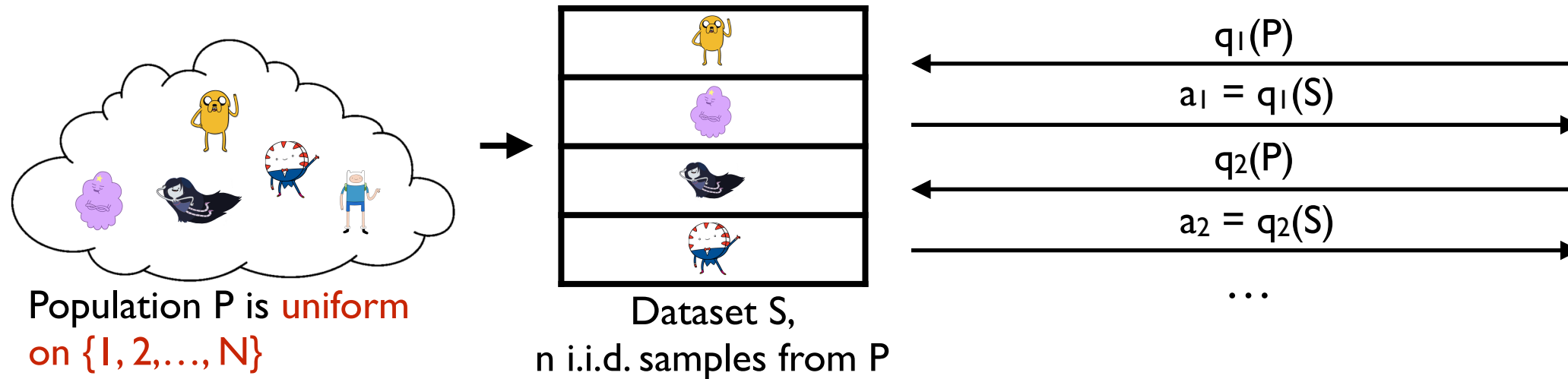
F^* close to the
“average”

$F^* =$
mechanism's
answers

F^* depends
only on S

↑
Ensured by
restrictions on M

Overfitting with Fingerprinting Codes



data scientist asks queries using the fingerprinting code $q: [N] \rightarrow \{\pm 1\}$

... applies Trace to the answers

Random fingerprinting code matrix
one query = one column

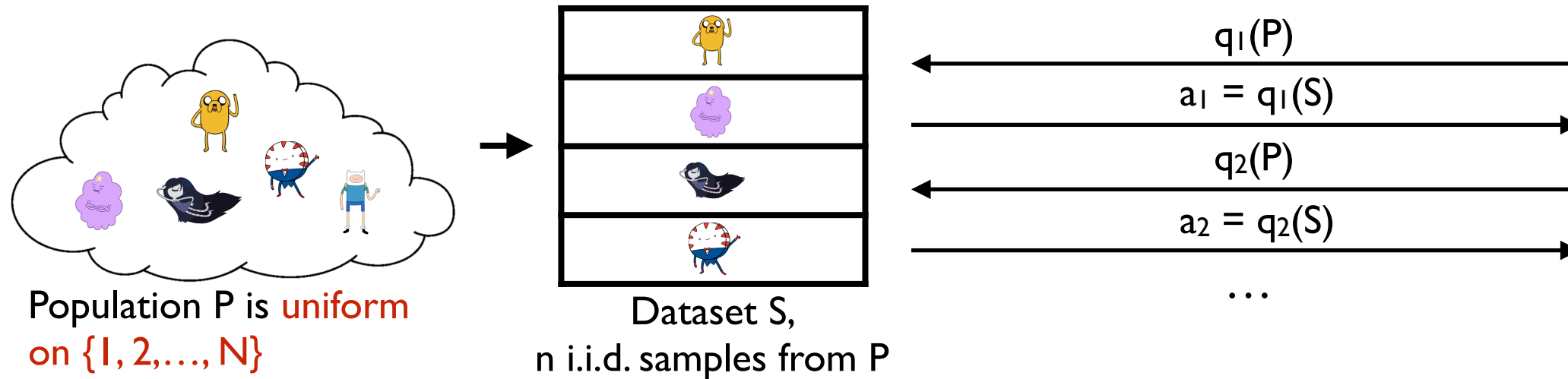
Query q_3

| | | | | | | | |
|--|---|---|---|---|---|--------------------------------------|---|
| | 1 | 0 | 0 | 0 | 1 | $q_3(\text{Jake}) =$ | 0 |
| | 1 | 0 | 1 | 0 | 0 | $q_3(\text{Finn}) =$ | 1 |
| | 1 | 1 | 0 | 0 | 0 | $q_3(\text{Marceline}) =$ | 0 |
| | 1 | 0 | 0 | 0 | 0 | $q_3(\text{BMO}) =$ | 0 |
| | 1 | 1 | 1 | 0 | 1 | $q_3(\text{Lumpy Space Princess}) =$ | 1 |

0.38

accurate answer is close to the average

Overfitting with Fingerprinting Codes



data scientist asks queries using the fingerprinting code $q: [N] \rightarrow \{\pm 1\}$

... applies Trace to the answers

Random fingerprinting code matrix
one query = one column

Query q_3

| | | | | | | | |
|--|---|---|---|---|---|------------------------|---|
| | 1 | 0 | 0 | 0 | 1 | $q_3(\text{Yellow}) =$ | 0 |
| | 1 | 0 | 1 | 0 | 0 | $q_3(\text{Blue}) =$ | 1 |
| | 1 | 1 | 0 | 0 | 0 | $q_3(\text{Black}) =$ | 0 |
| | 1 | 0 | 0 | 0 | 0 | $q_3(\text{Red}) =$ | |
| | 1 | 1 | 1 | 0 | 1 | $q_3(\text{Purple}) =$ | 1 |

0.38

Q: how do we ensure that the answers only depend on the sample?

A: use cryptography to “blind” the queries

accurate answer is close to the average

Negative Results

Theorem (Computational Version) (HU'14, SU'15):

If one-way functions exist and $d = \omega(\log n)$, there is no computationally efficient mechanism* that answers more than $k = \tilde{O}(n^2)$ arbitrary adaptively chosen queries.

Theorem (Information-Theoretic Version) (HU'14, SU'15):

If $d > k$, then there is no mechanism, efficient or not, that answers more than $k = \tilde{O}(n^2)$ arbitrary adaptively chosen queries.

*computationally efficient \approx answers each query in time polynomial in $|S|=nd$

Step eleven: thank your audience!

Step twelve: take questions!