Allier protection de données personnelles et open data

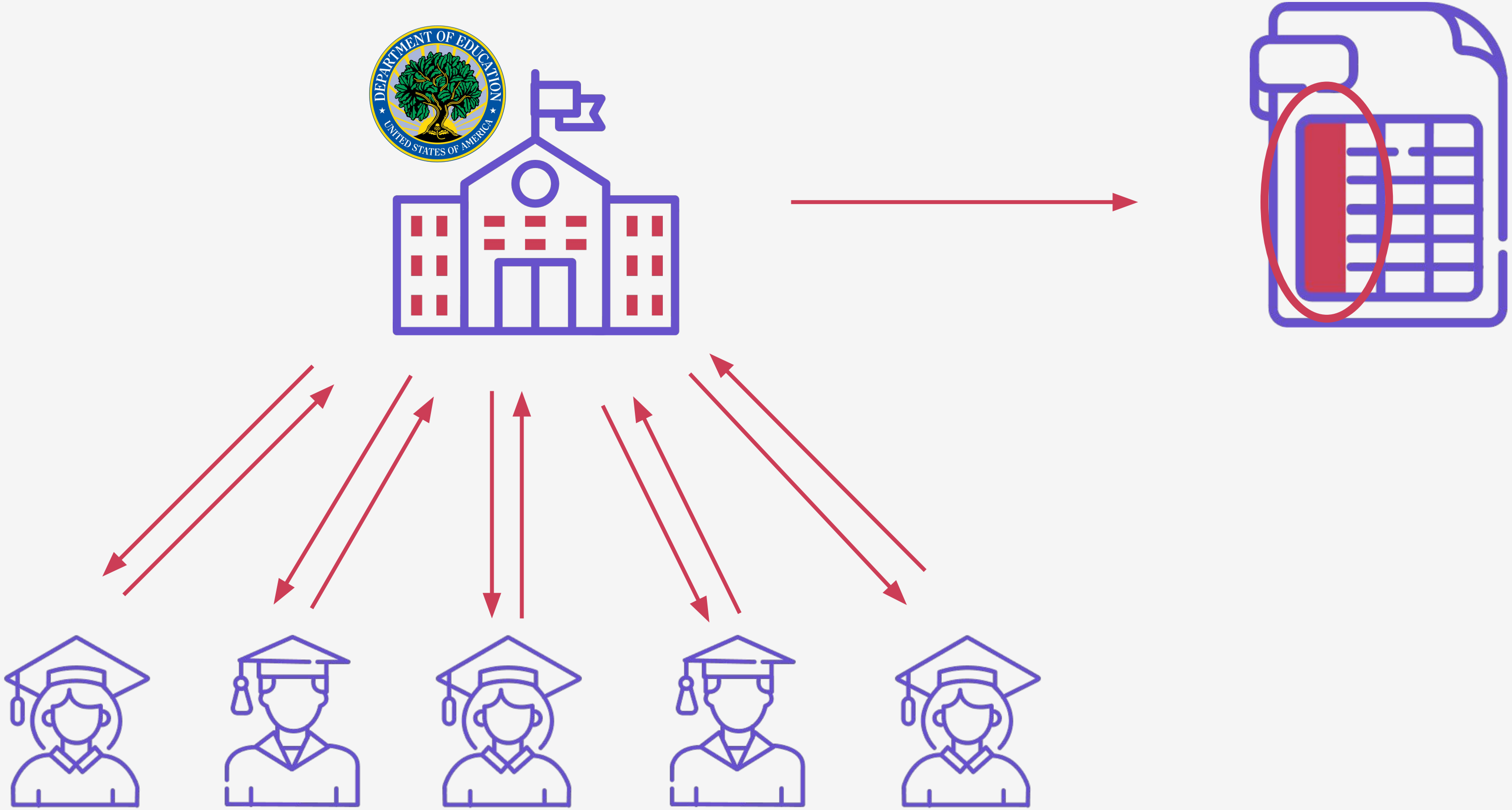
Damien Desfontaines
@TedOnPrivacy

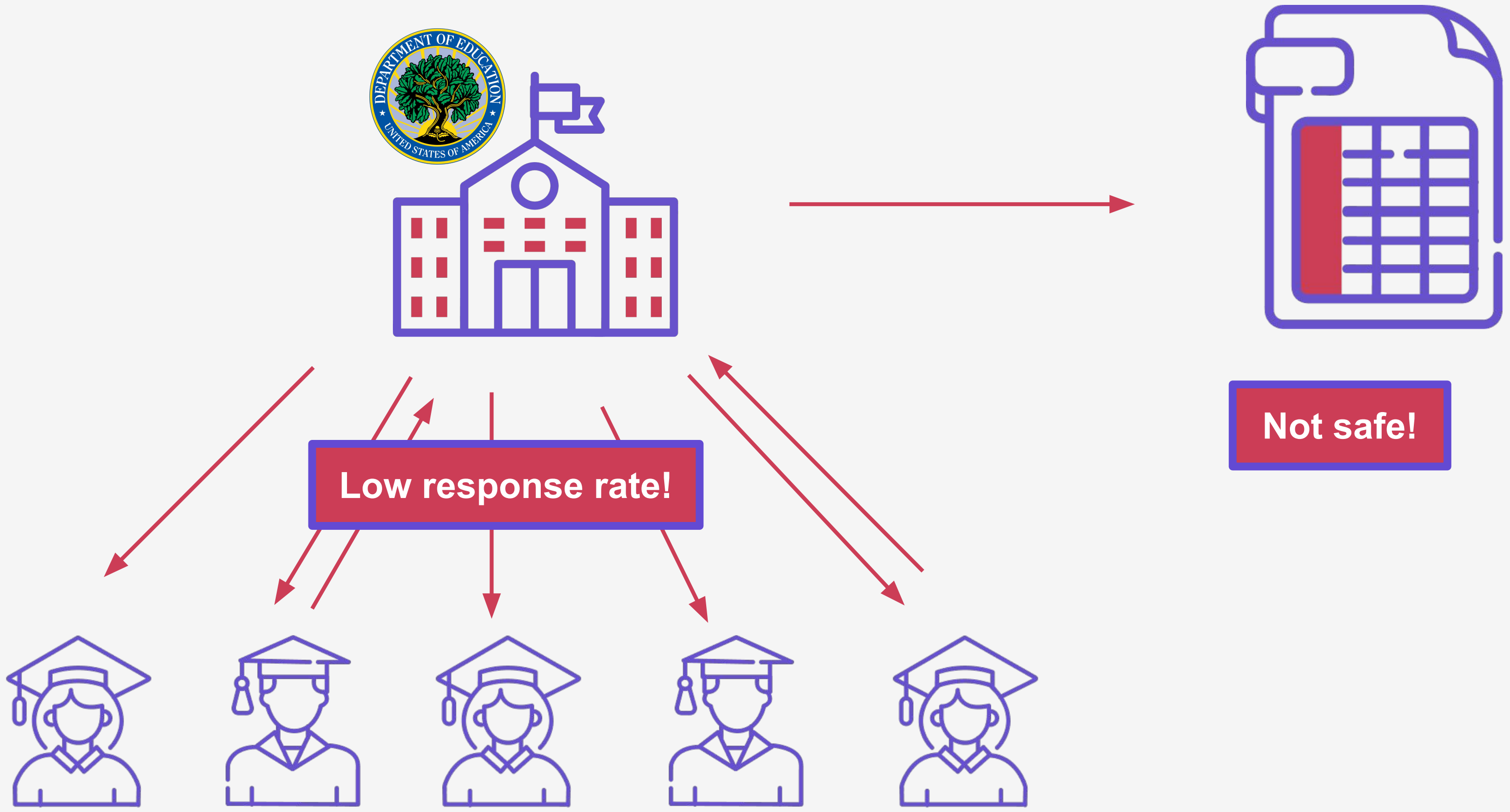
How much do new graduates earn after completing their degree?

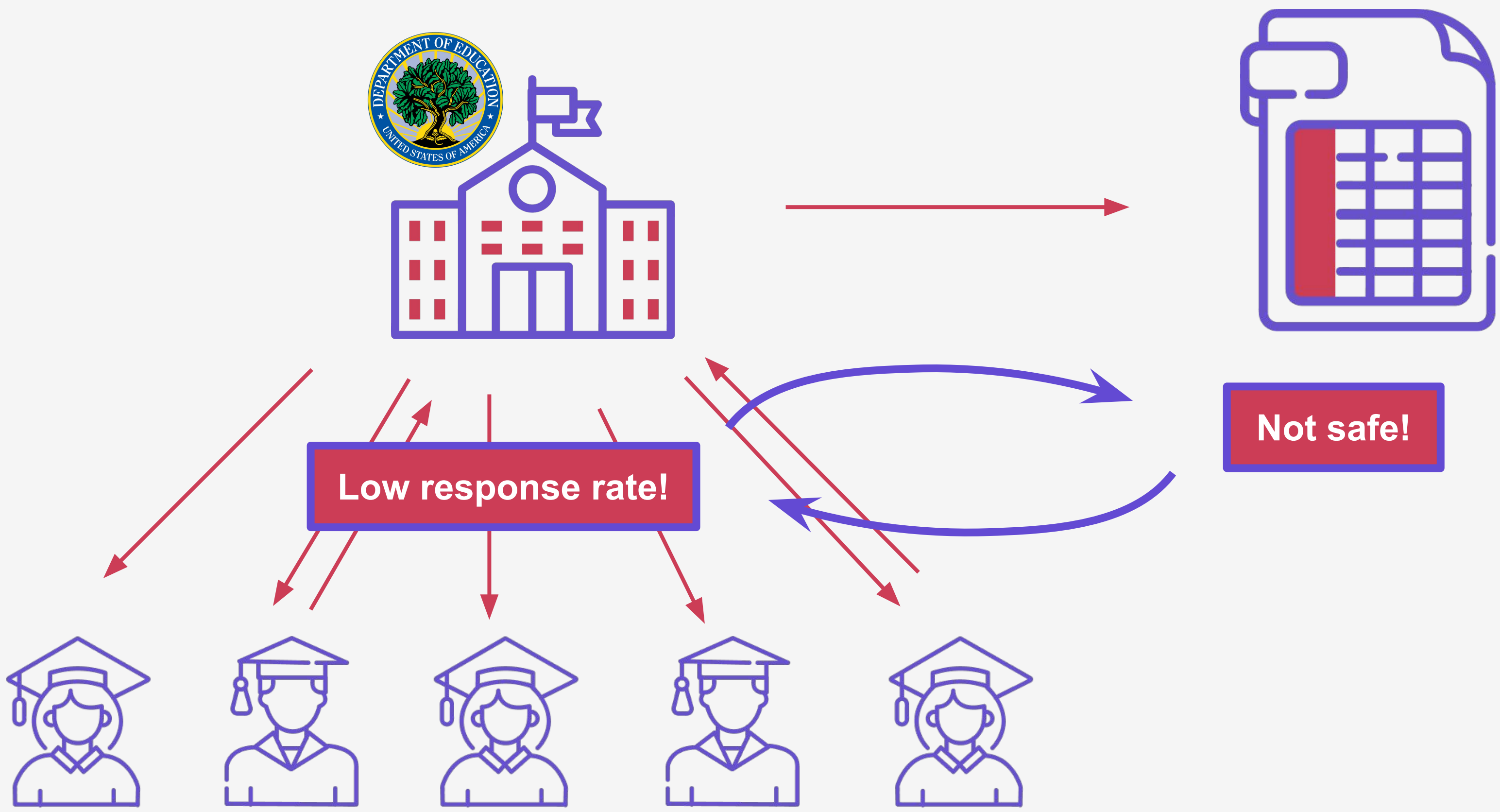
Scientists

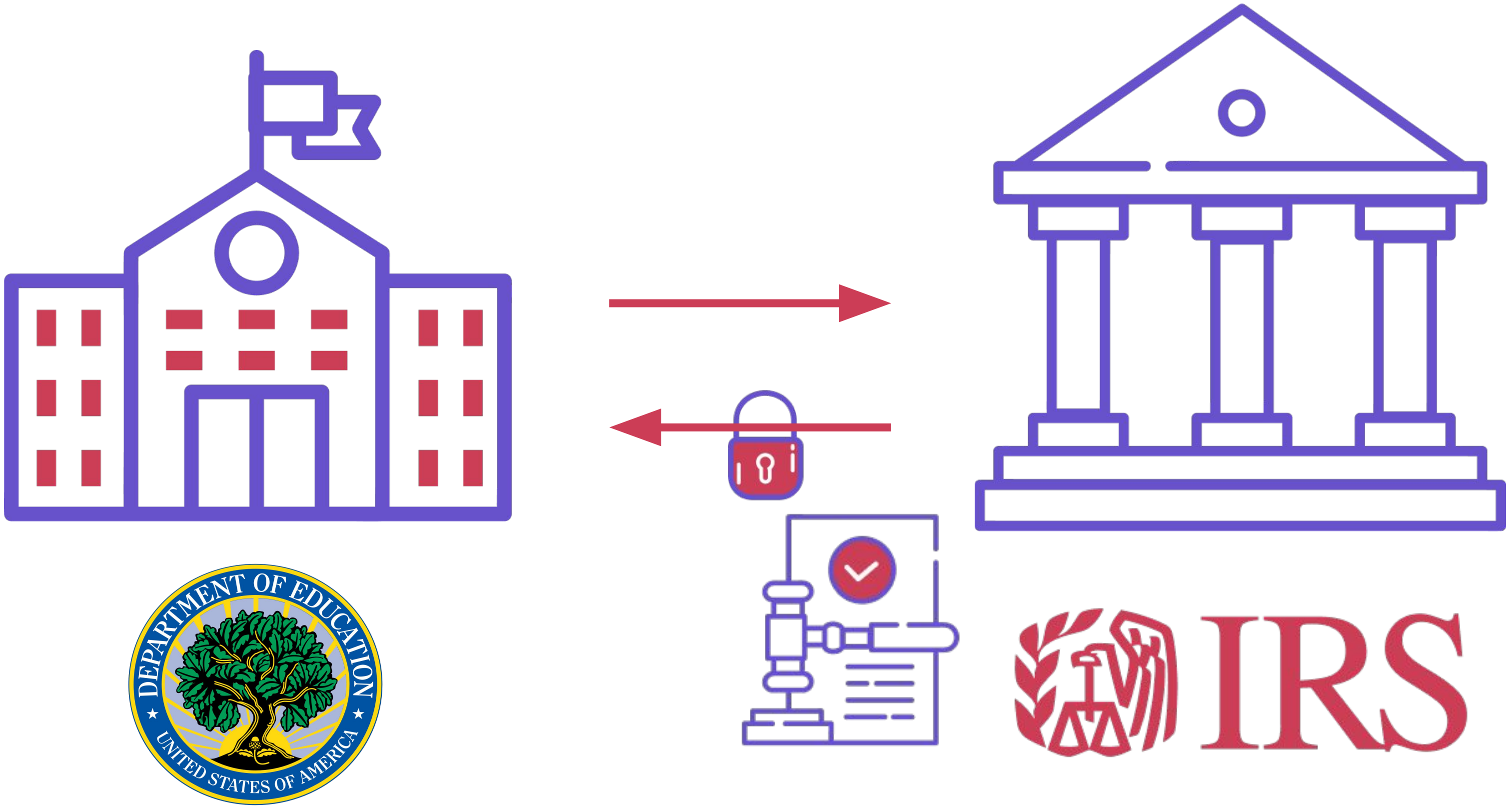
Government agencies

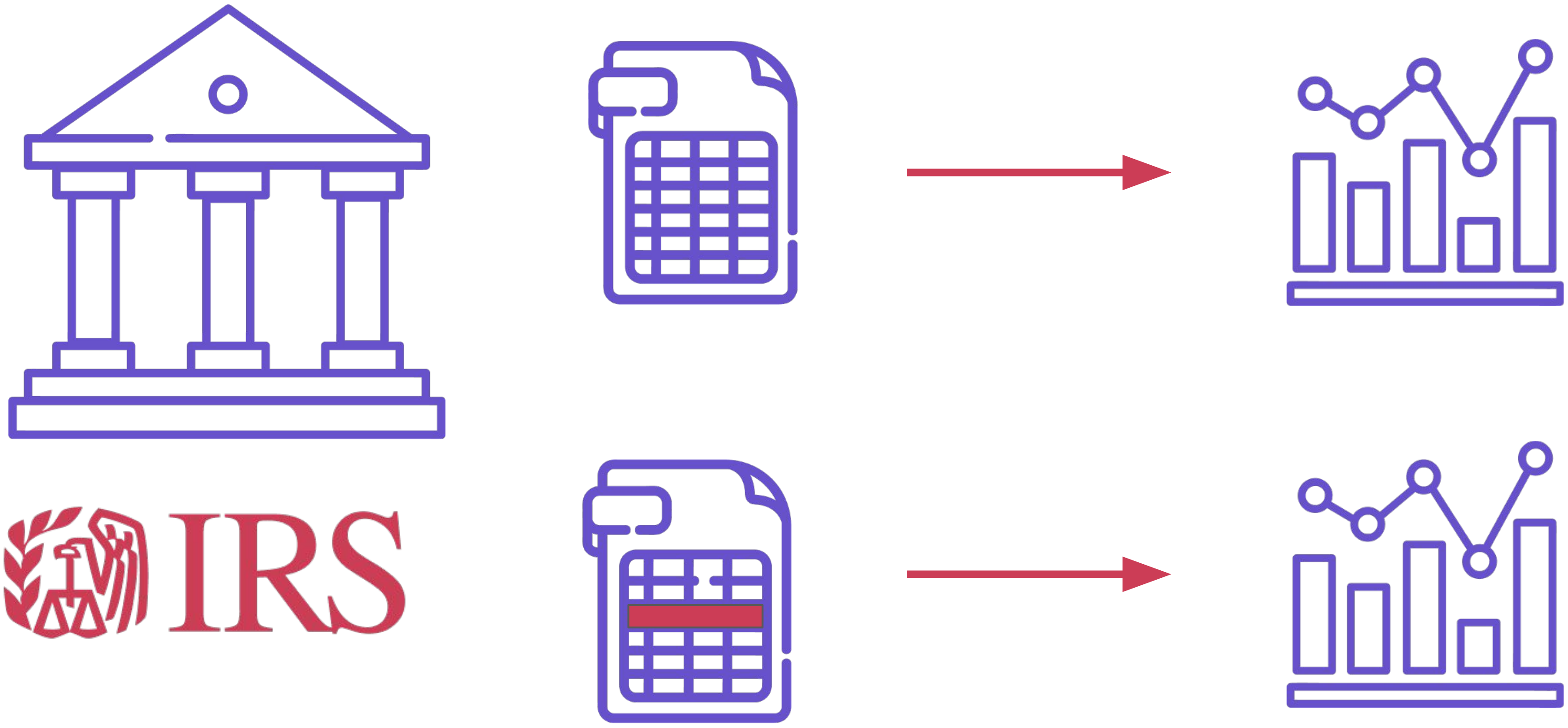
Students

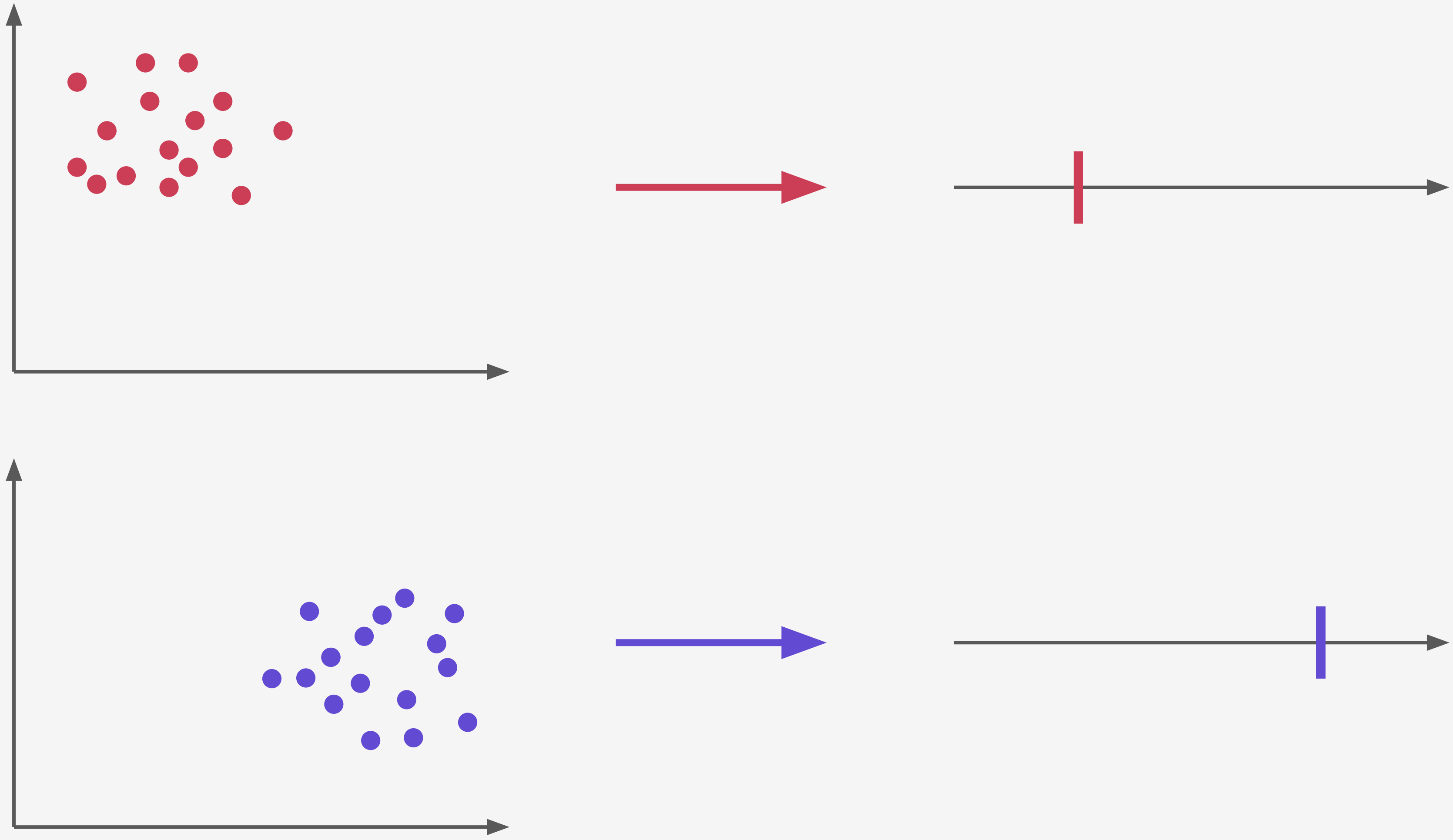

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA

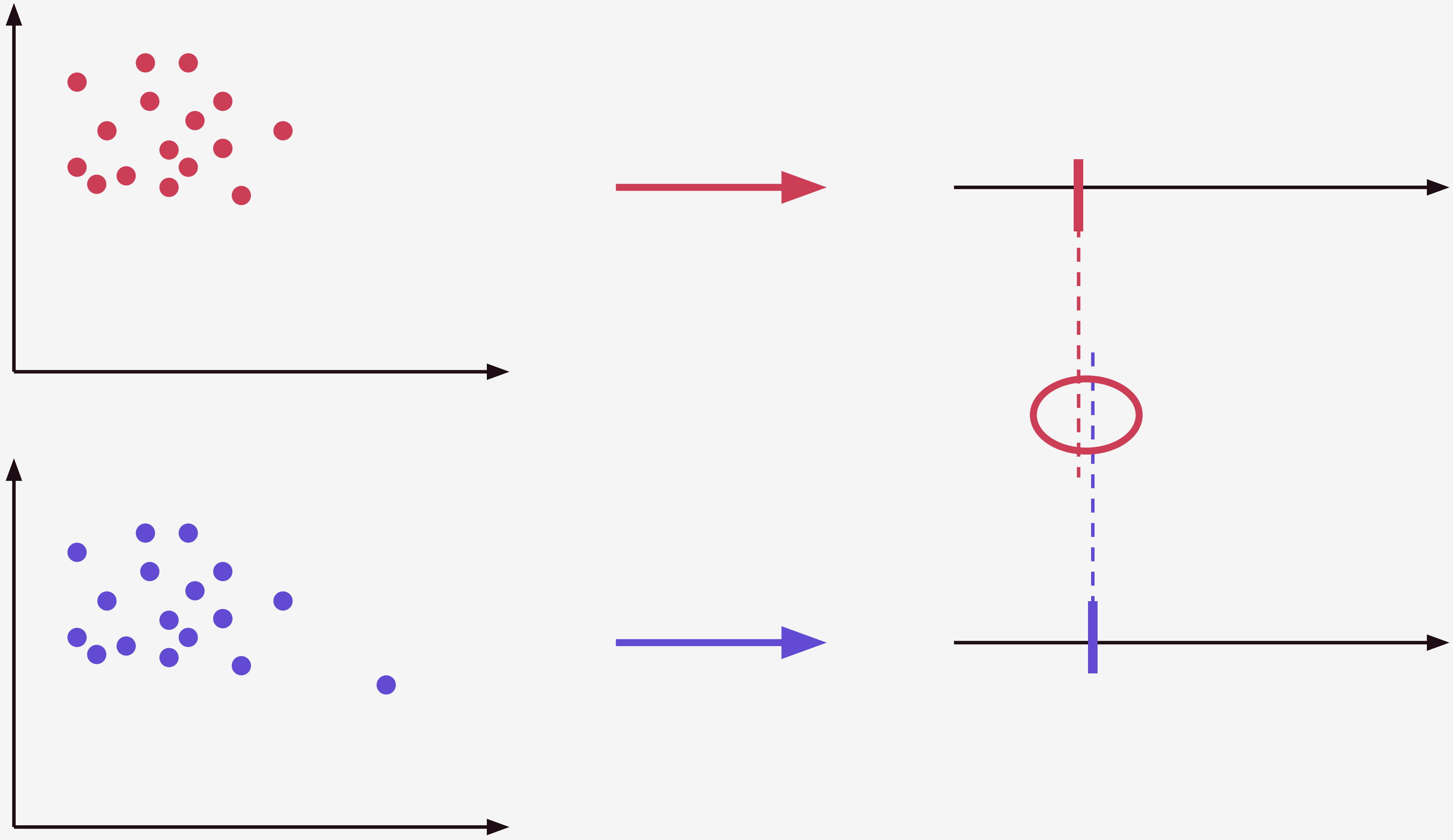

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA
Low response rate!
Not safe!

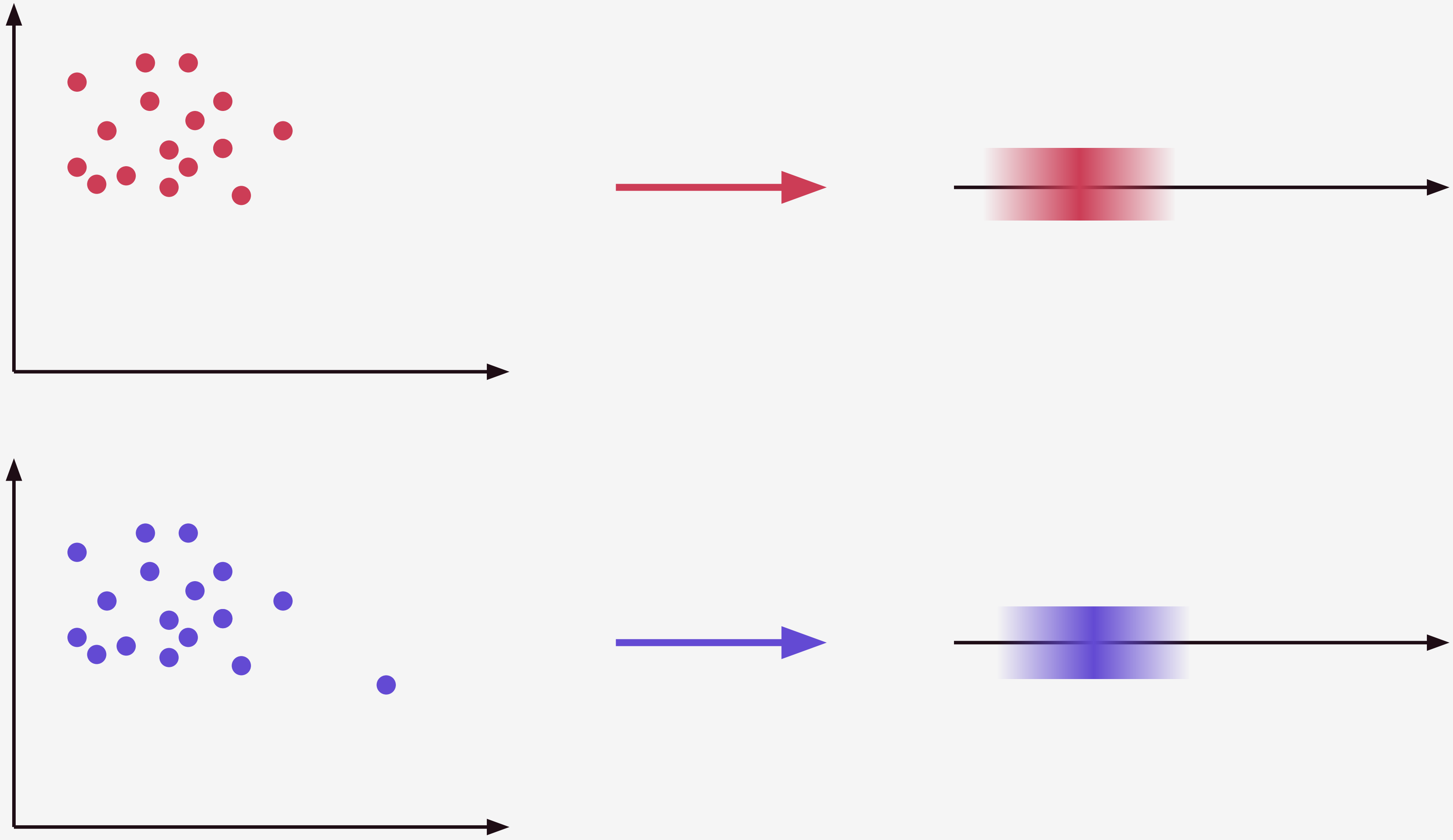

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA
Low response rate!
Not safe!

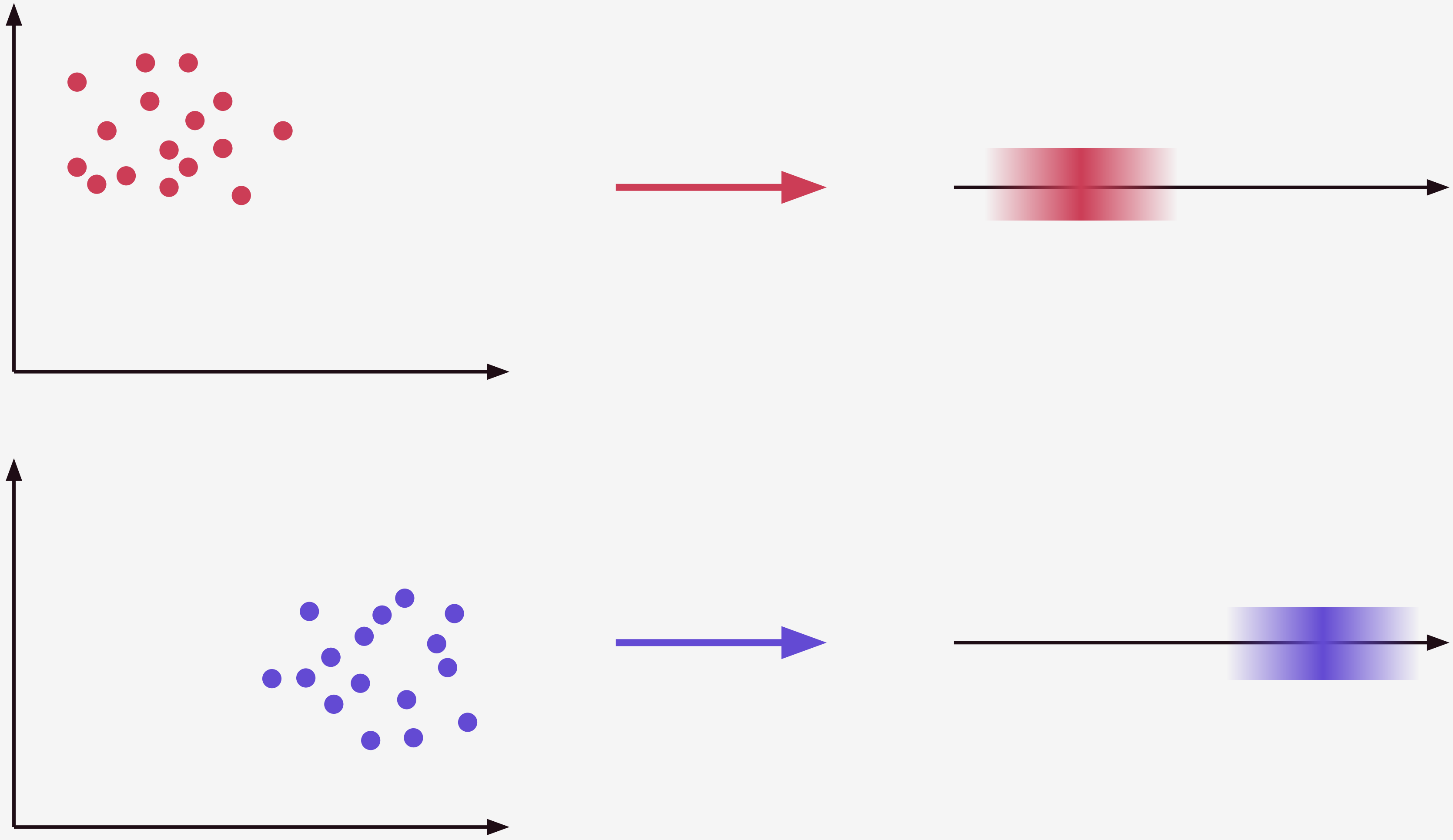

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA
IRS


IRS



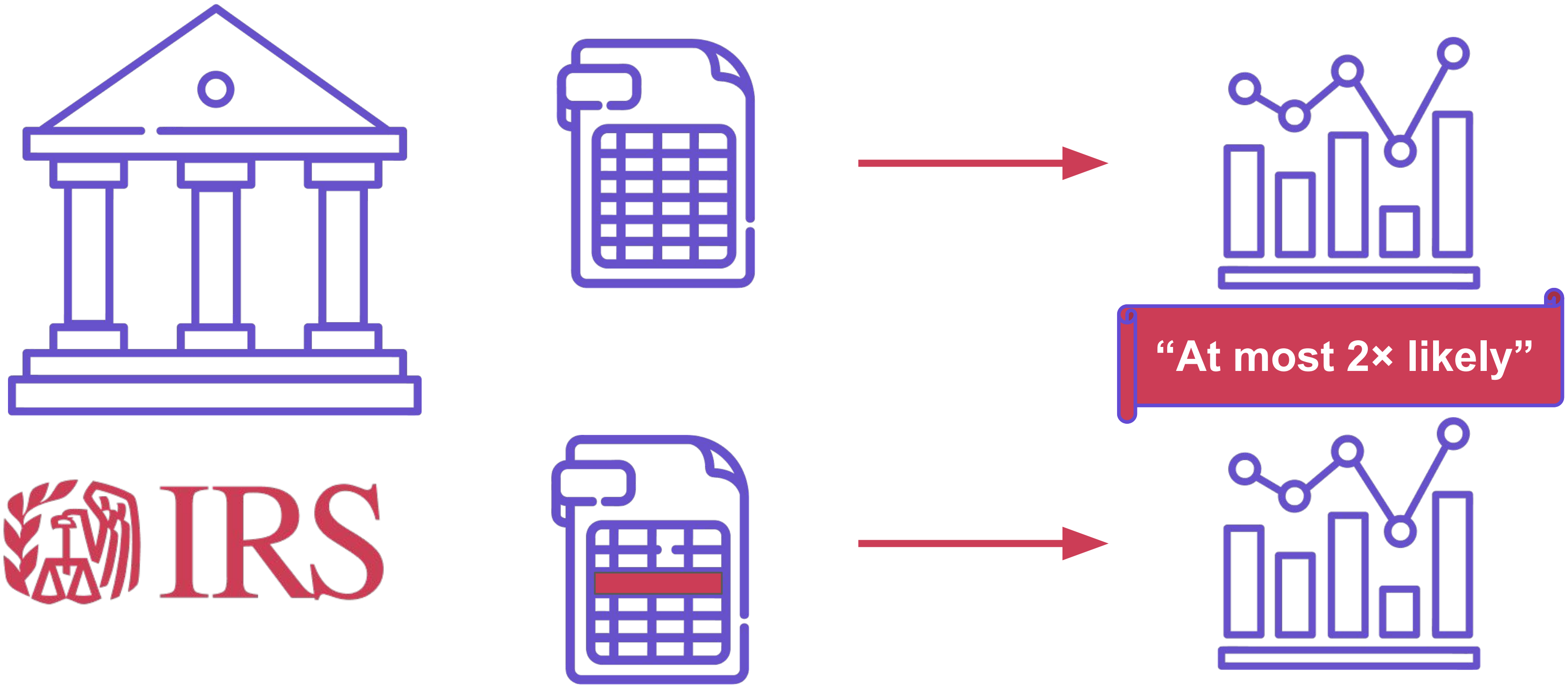


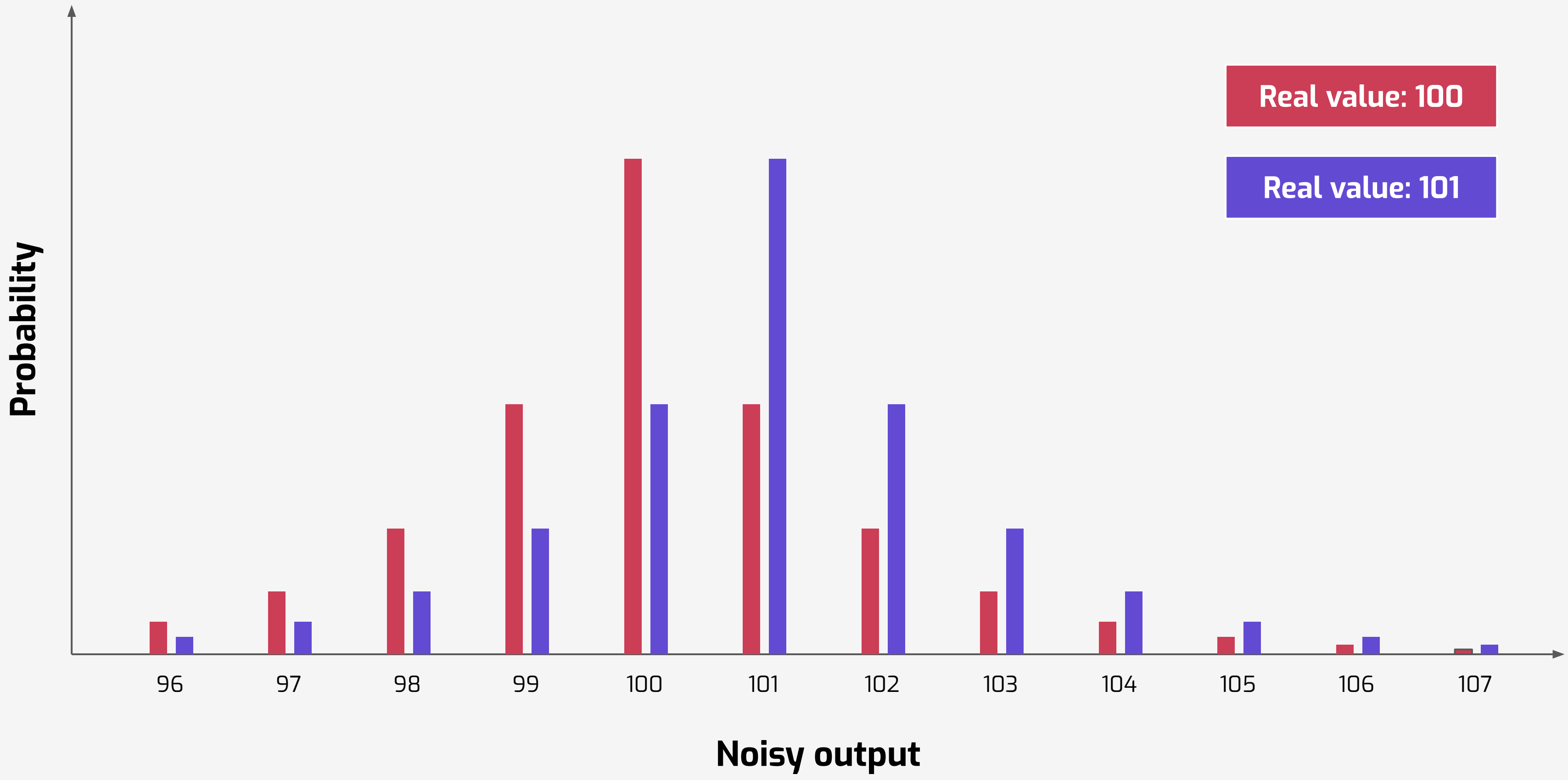




Differential privacy

- Transparent
- **Quantifiable** guarantee: ε
- Composition: $\varepsilon_1 + \varepsilon_2 = \varepsilon_{total}$

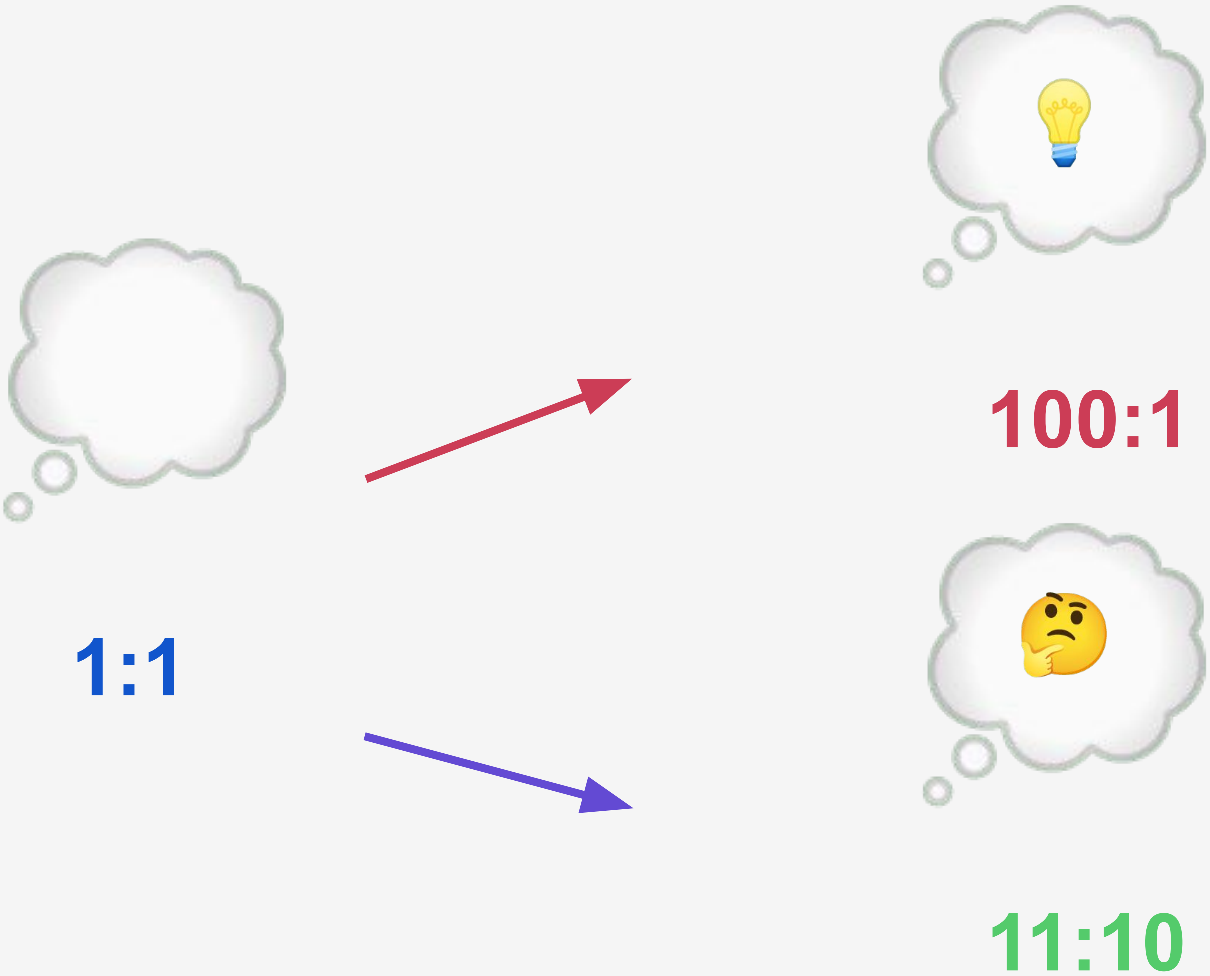

IRS
“At most 2× likely”

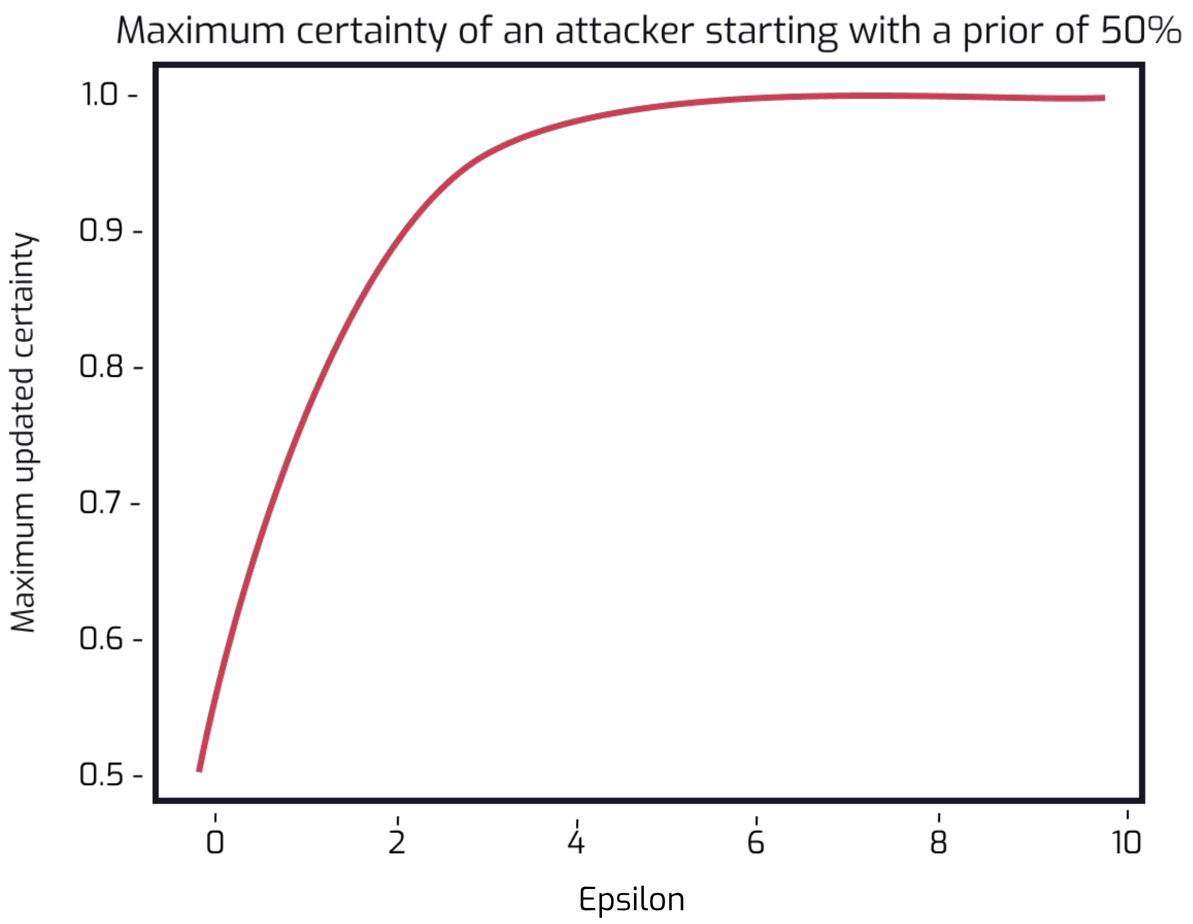

Real value: 100
Real value: 101
Probability
96
97
98
99
100
101
102
103
104
105
106
107
Noisy output

Differential privacy

- Transparent
- **Quantifiable** guarantee: ε
- Composition: $\varepsilon_1 + \varepsilon_2 = \varepsilon_{total}$

- You can't ignore the noise
- You have to choose ε
- Implementation can be tricky


1:1
100:1
11:10



Maximum certainty of an attacker starting with a prior of 50%
Maximum updated certainty
1.0
0.9
0.8
0.7
0.6
0.5
0
2
4
6
8
10
Epsilon

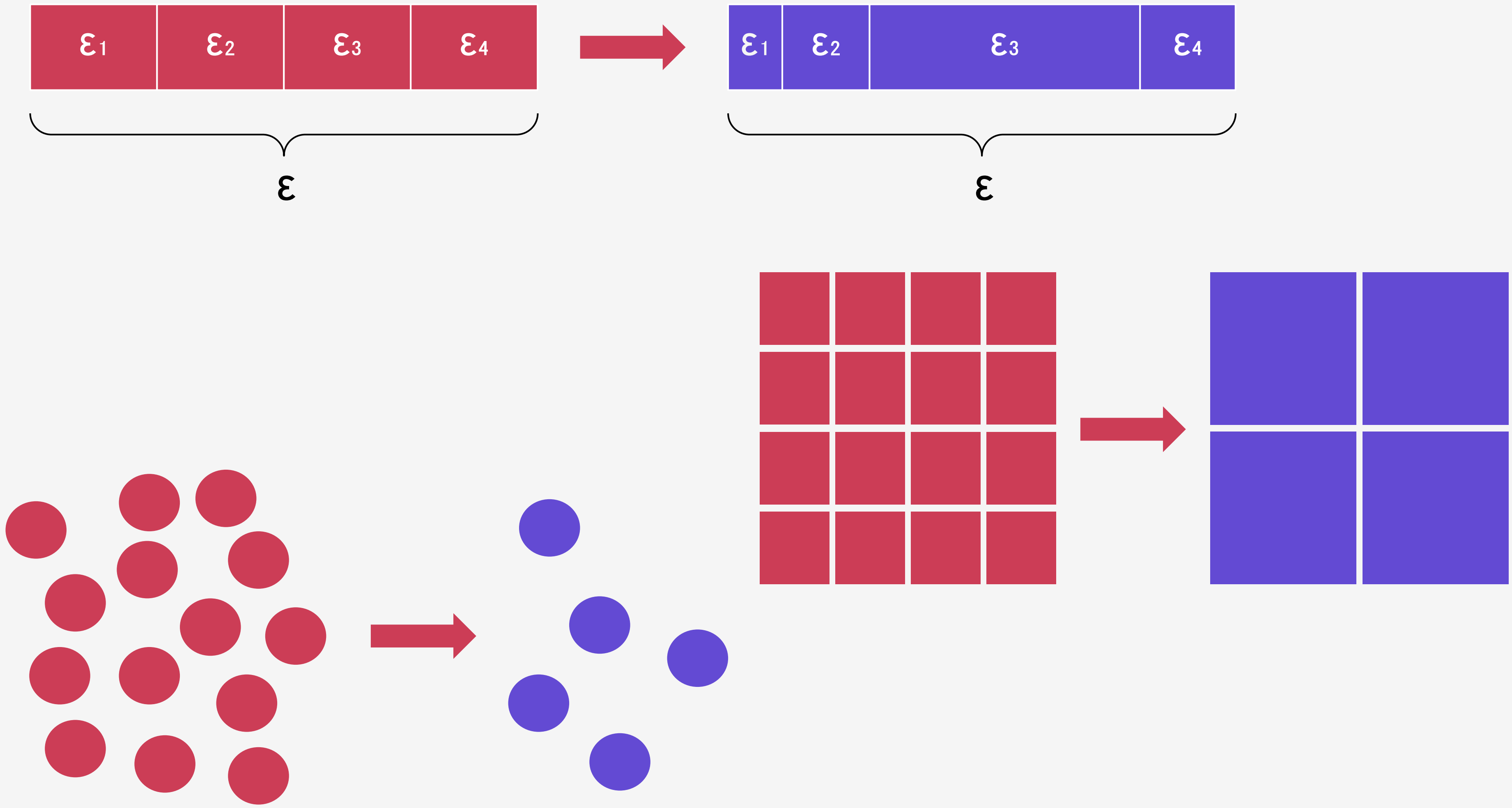
At the ε we chose, the expected error will be ≈3 for counts, and $1500 to $15000 for quantiles.

That's fine for counts, but the error is too great for quantiles. Can you increase ε?

Not beyond a certain limit.
But we could devote more of the privacy budget to quantiles: error would become ≈6 for counts, but $1000 to $10000 for quantiles.

That's better. But we care more about medians than 25th and 75th percentiles, could we optimize accordingly?

...



ε_1
ε_2
ε_3
ε_4
ε
ε_1
ε_2
ε_3
ε_4
ε

At the ε we chose, the expected error will be ~3 for counts, and $1500 to $15000 for quantiles.

That’s fine for counts, but the error is too great for quantiles. Can you increase ε?

...

OK, if we spend more of the budget towards quantiles, focus on medians, and drop some breakdowns, error comes down to between $500 and $5000.

That works for us!

Duke University

Durham, NC
6,650 undergraduate students
duke.edu

4 Year | Private Nonprofit | City | Medium

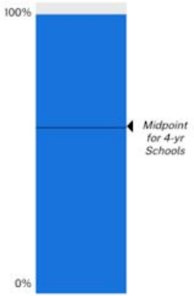
United Methodist

Midpoint for 4-yr Schools | Midpoint for All Schools

Graduation Rate

96%

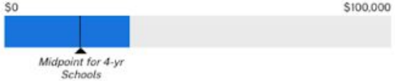
Midpoint for 4-yr Schools: 57%



Average Annual Cost

$32,459

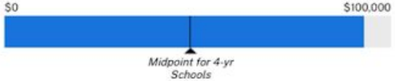
Midpoint for 4-yr Schools: $19,534



Median Earnings

$93,115

Midpoint for 4-yr Schools: $47,891



SEARCH FOR:

- School
- Field of Study

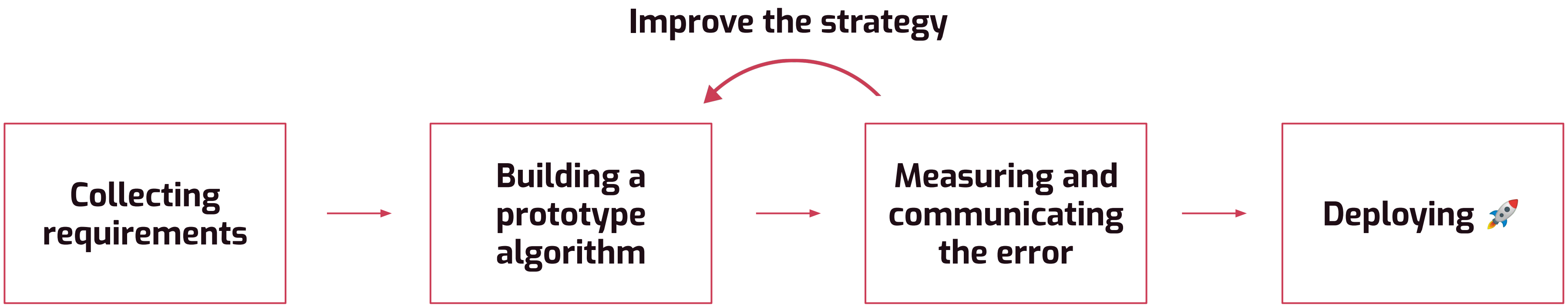
Type to search

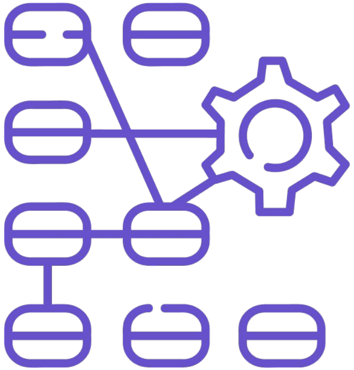
Start Your FAFSA® Application

To receive financial aid, you must complete the *Free Application for Federal Student Aid* (FAFSA®) form. You can use Federal Student Aid Estimator to see how much aid may be available to you.

START YOUR FAFSA® APPLICATION

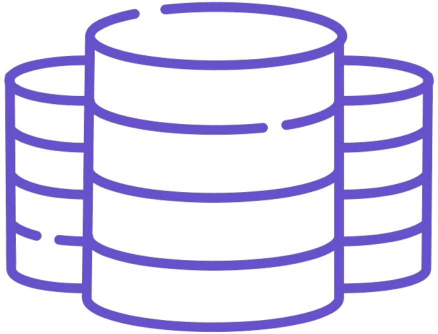
Don't forget: Do fill out the FAFSA® form, but also look into other programs such as GI Bill Benefits that may also help you pay for school.


Improve the strategy
Collecting requirements
Building a prototype algorithm
Measuring and communicating the error
Deploying 🚀



Complexity

Robustness



Scale

The Tumult Platform

Open source
≈ July 2022

- **Easy-to-use API for data scientists**
- Familiar interface, similar to Pandas/Spark
- Hides away the complexity of DP
- Includes optimizations for greater accuracy
- Many aggregations & transformations

Extensible framework for power users

- Framework based on peer-reviewed research
- Built for scale, on top of Apache Spark
- User composes DP “building blocks” …
- … and obtains an end-to-end privacy proof

WIKIMEDIA FOUNDATION

IRS

United States Census Bureau

Tumult **Analytics**

Tumult **Core**

APACHE Spark

Thanks 💚

Damien Desfontaines
damien@tmlt.io
@TedOnPrivacy

tmlt.io/connect
tmlt.io/careers

```
session = Session.from_dataframe(
    dataframe=private_data,
    source_id="my_data",
    privacy_budget=PureDPBudget(1.7),
)

query = (
    QueryBuilder("my_data")
    .filter("age > 42")
    .groupby(zip_codes)
    .median("income", low=0, high=10**6)
)

result = session.evaluate(query, PureDPBudget(0.8))
```

Differential privacy & regulation

- Privacy regulations have a carve-out for **fully anonymized** data
- The scientific community recognizes DP as a **gold standard**
- Research suggests alignment between legal concepts & DP

The Role of Differential Privacy in GDPR Compliance

Position Paper

Rachel Cummings
Georgia Institute of Technology
School of Industrial and Systems Engineering
rachelc@gatech.edu

Deven Desai
Georgia Institute of Technology
Scheller College of Business
deven.desai@scheller.gatech.edu

ABSTRACT

The EU General Data Protection Regulation (GDPR) empowers individuals with the right to control erasure of their personal data held by firms. GDPR also allows firms to retain anonymized aggregate data and statistical results. Unfortunately, most recommender systems (and many other types of machine learning models) memoize individual data entries as they are trained, and thus are not

specific person" [9]. This view connects to the language of
26 which states, "The principles of data protection should t
not apply to anonymous information, namely informatio
does not relate to an identified or identifiable natural pers
personal data rendered anonymous in such a manner that
subject is not or no longer identifiable." Recital 26 concludes
GDPR "does not therefore concern the processing of such

Towards formalizing the GDPR's notion of singling out

Aloni Cohen[a,b,c,1,2] and Kobbi Nissim[d,1,2]

[a]Rafik B. Hariri Institute for Computing and Computational Science & Engineering, Boston University, Boston, MA 02215; [b]School of Law, Boston University, Boston, MA 02215; [c]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; and [d]Department of Computer Science, Georgetown University, Washington, DC 20007

Edited by Moshe Y. Vardi, Rice University, Houston, TX, and approved March 3, 2020 (received for review August 21, 2019)

There is a significant conceptual gap between legal and mathematical thinking around data privacy. The effect is uncertainty as to which technical offerings meet legal standards. This uncertainty is exacerbated by a litany of successful privacy attacks demonstrating that traditional statistical disclosure limitation techniques often fall short of the privacy envisioned by reg-

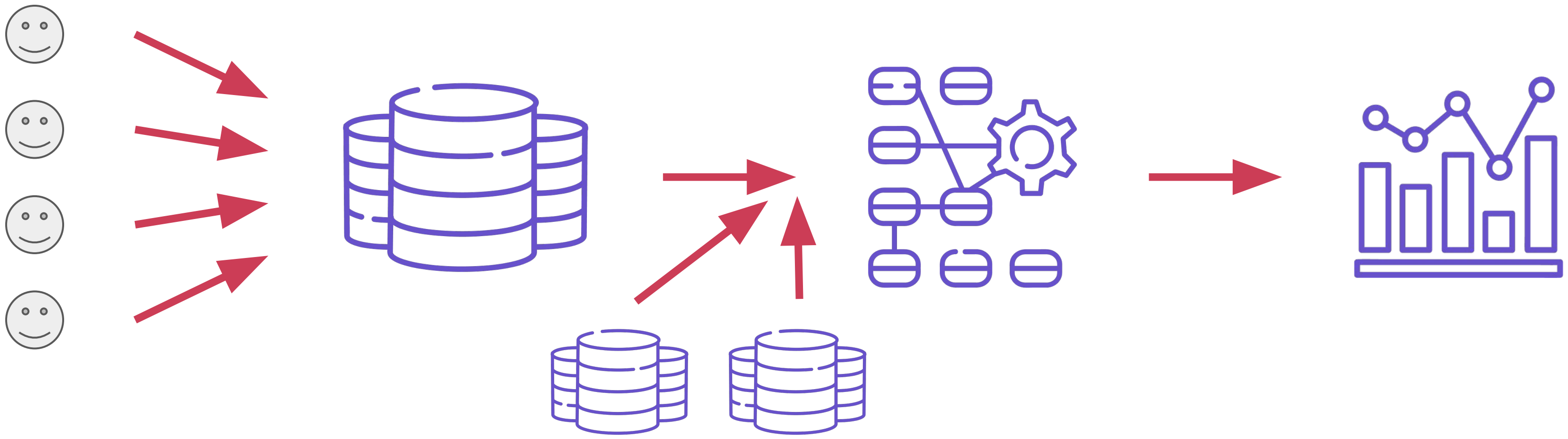
mented in the academy, industry, and government, there is a lack of discourse between the legal and mathematical conceptions. The effect is uncertainty as to which technical offerings adequately match expectations expressed in legal standards (3).

Privacy-enhancing technologies

Collecting data privately:
secure aggregation,
local differential privacy

Computing data privately:
secure enclaves,
homomorphic encryption

Sharing data privately:
differential privacy



Joining data privately:
multi-party computation