Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy

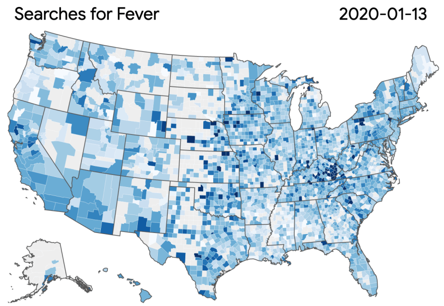
Lucas Rosenblatt[1], Bernease Herman[2], Anastasia Holovenko[5], Wonkwon Lee[1], Joshua Loftus[3], Elizabeth McKinnie[4], Taras Rumezhak[5], Andrii Stadnik[5], Bill Howe[2], Julia Stoyanovich[1]

[1] New York University
[2] University of Washington
[3] London School of Economics
[4] Microsoft
[5] Ukrainian Catholic University

Pervasively Deployed DP Mechanisms

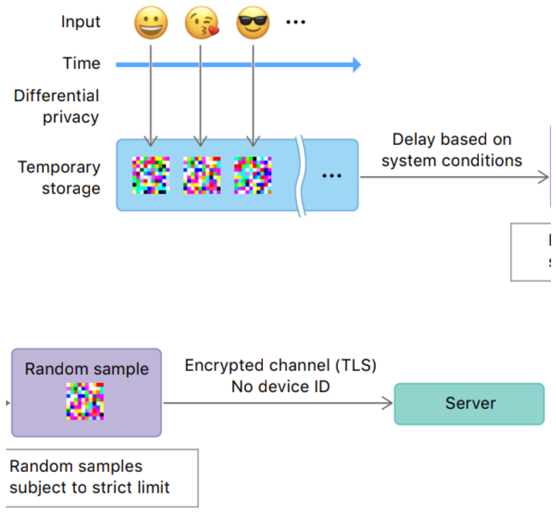
Big Tech Uses DP with all your data!

Search Trends Symptoms



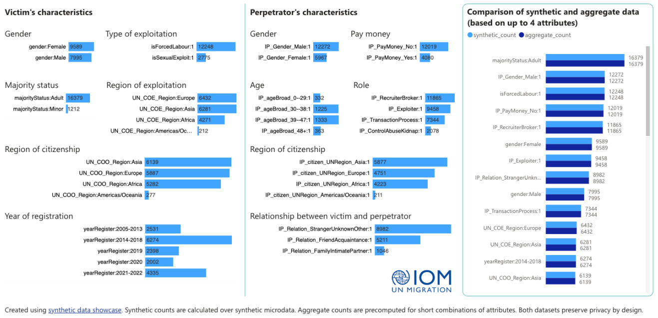
Apple

Emoji Suggestions + Health Type Usage

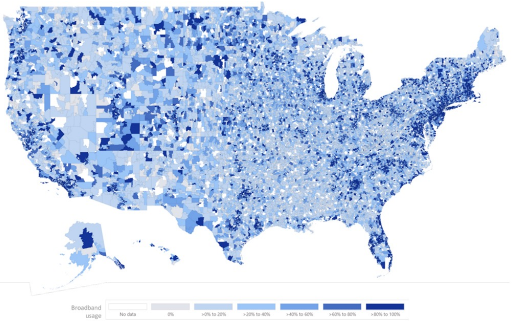


Microsoft

Global victim-perpetrator synthetic dataset



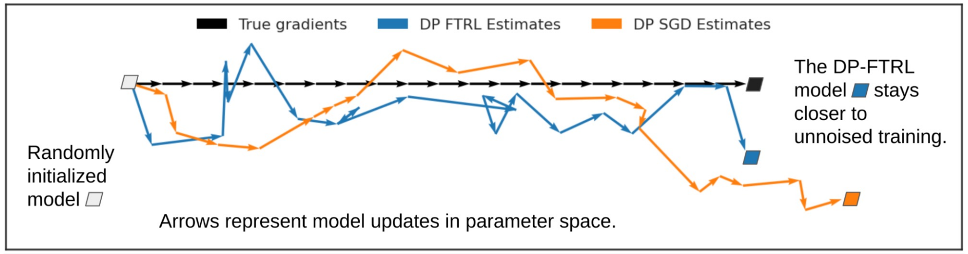
US Broadband Coverage Dataset



Google

Community Mobility Reports

Next-word prediction model on Gboard



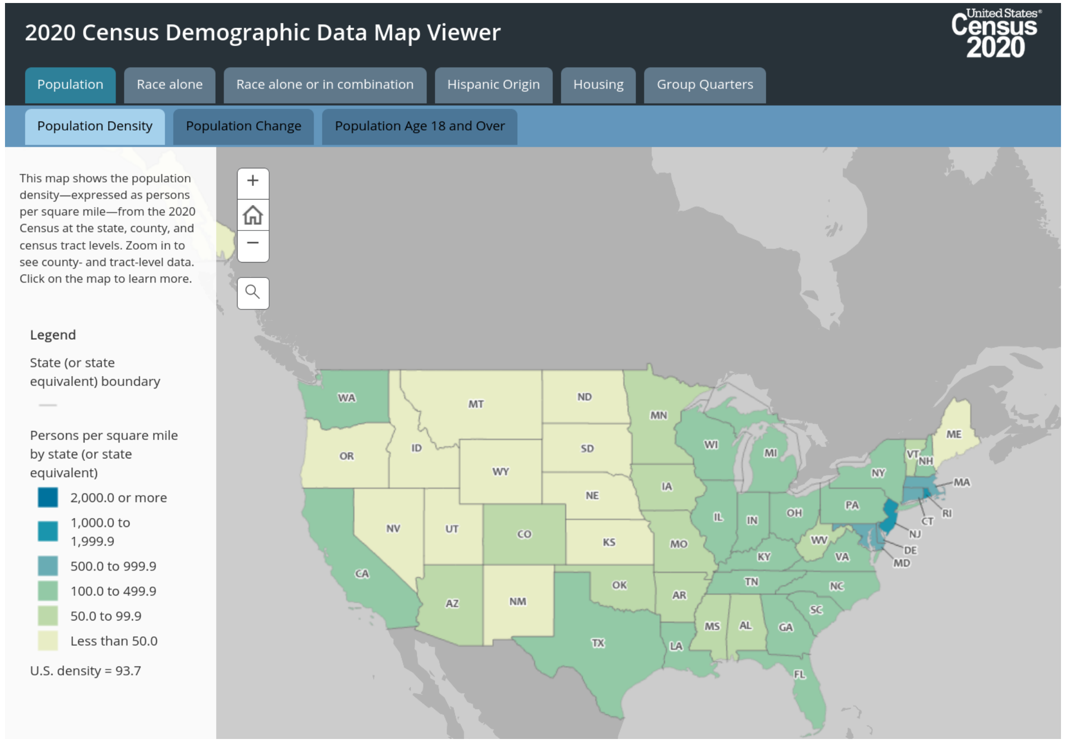
https://desfontain.es/privacy/real-world-differential-privacy.html

r/ai NYU

Pervasively Deployed DP Mechanisms

And so does the U.S. Government…

2020 Census Redistricting Data

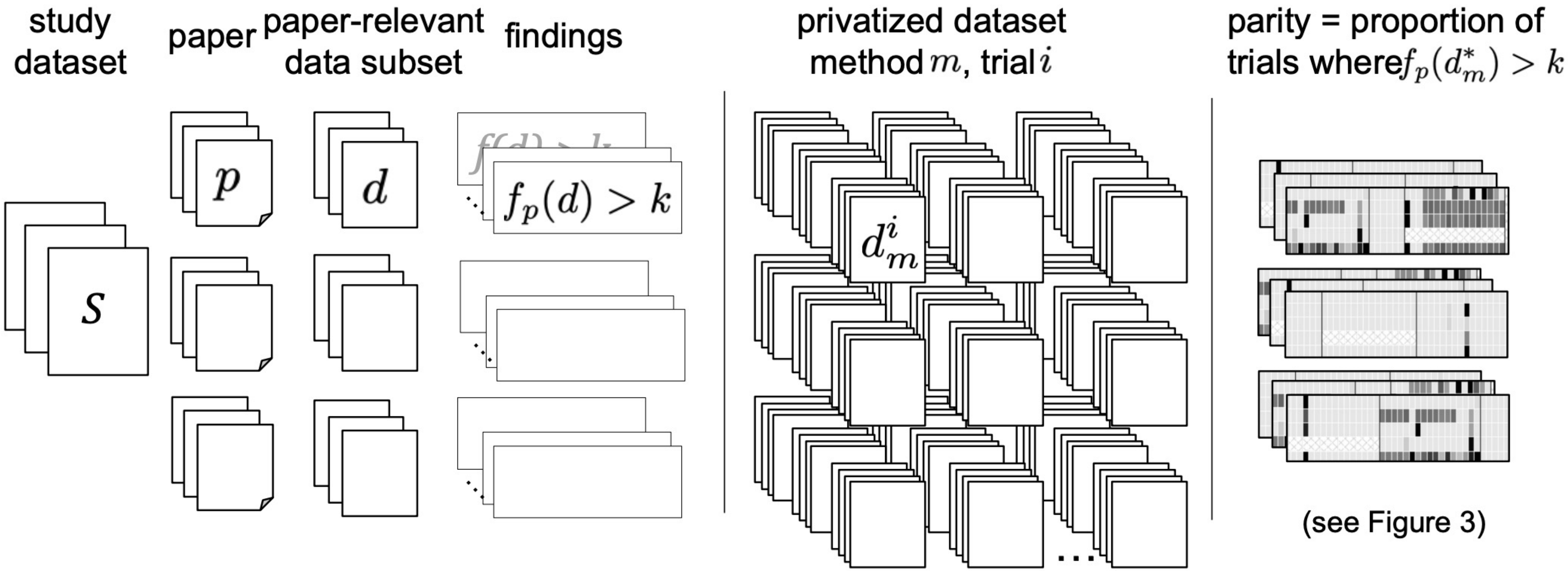


- A lot of these deployments rely on variations of **DP Synthetic Data**
- DP for the Census was met with resistance among many in the research community. They claim DP noise:
  - Affects demographic totals [Ruggles 2019]
  - Exacerbates underrepresentation of minorities [Ganev et. al 2021, Kenny et. al 2021]
- However, DP is still *probably better* than swapping in terms of the privacy/utility tradeoff [Christ et al 2022]

A Proposed Benchmark

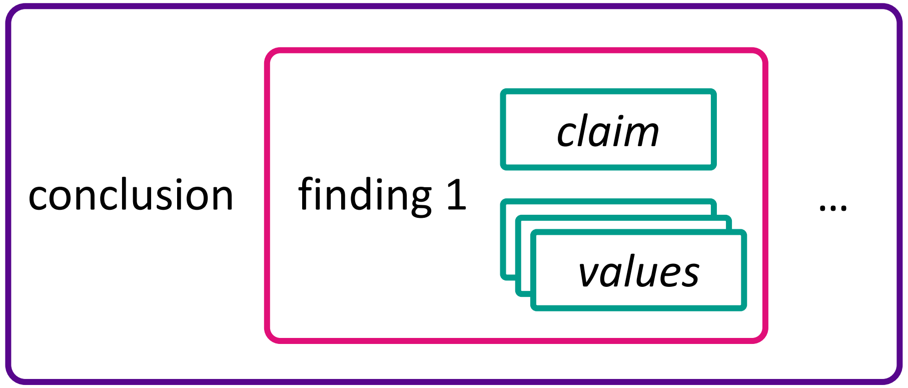
- Major challenge: **Evaluation!**
- How do we **convincingly** evaluate DP synthetic data?
  - Social scientists and practitioners don't trust *random linear que*ry workloads
  - Open questions: how do these synthesizers perform on a variety of data? What are their limitations?
- **SynRD: An "Epistemic Parity" Benchmark**
  1. Avoid assumptions about the representativeness of proxy tasks!
  2. Instead, measure likelihood that published conclusions (like those run on Census data) would ***change had the authors used DP synthetic data.***
  3. Make this an accessible benchmark and choose the "published conclusions" to be real, high-quality papers on impactful studies

SynRD: Benchmark for Evaluating "Epistemic Parity"



Challenge: Taxonomy over findings

- Problem with operationalizing "Epistemic Parity:" many scientific findings/conclusions are semantic!
- Solution: principled taxonomy over language of scientific literature (inspired by Cohen et. al, 2018)
- Means we can realize taxonomy (we do this in python)




```
def finding_6_9(self):
corr_df = self.get_corr()
corr_obesity_death = \
    corr_df['Obesity'].loc['Cerebrovascular death']
soft_finding = abs(corr_obesity_death) < 0.05
```

Challenge: significance of results

- Correctly done, experimental science relies on significance testing
- How does this work with DP synthetic data?

- Rubin's Rules for calculating uncertainty over results of synthetic data
  - (over estimated locations $q\_1, \ldots q\_m$ and variances $v\_1, \ldots v\_m$)

$$\hat{q} = \frac{1}{m}\sum_{i=1}^{m} q_i \qquad \hat{v} = \frac{1}{m}\sum_{i=1}^{m} v_i$$

$$b = \frac{1}{m-1}\sum_{i=1}^{m}(q_i - \hat{q})^2 \qquad T = \left(1 + \frac{1}{m}\right) b - \hat{v}, \qquad \mathrm{df} = \left(1 - \frac{1}{1 + \frac{1}{m}}\frac{\hat{v}}{b}\right)^2 (m-1).$$

- Problem: *Crucially* relies on a normality assumption for each $\tau\ (X_i) = q_$i

Solution: simplify finding statistics!

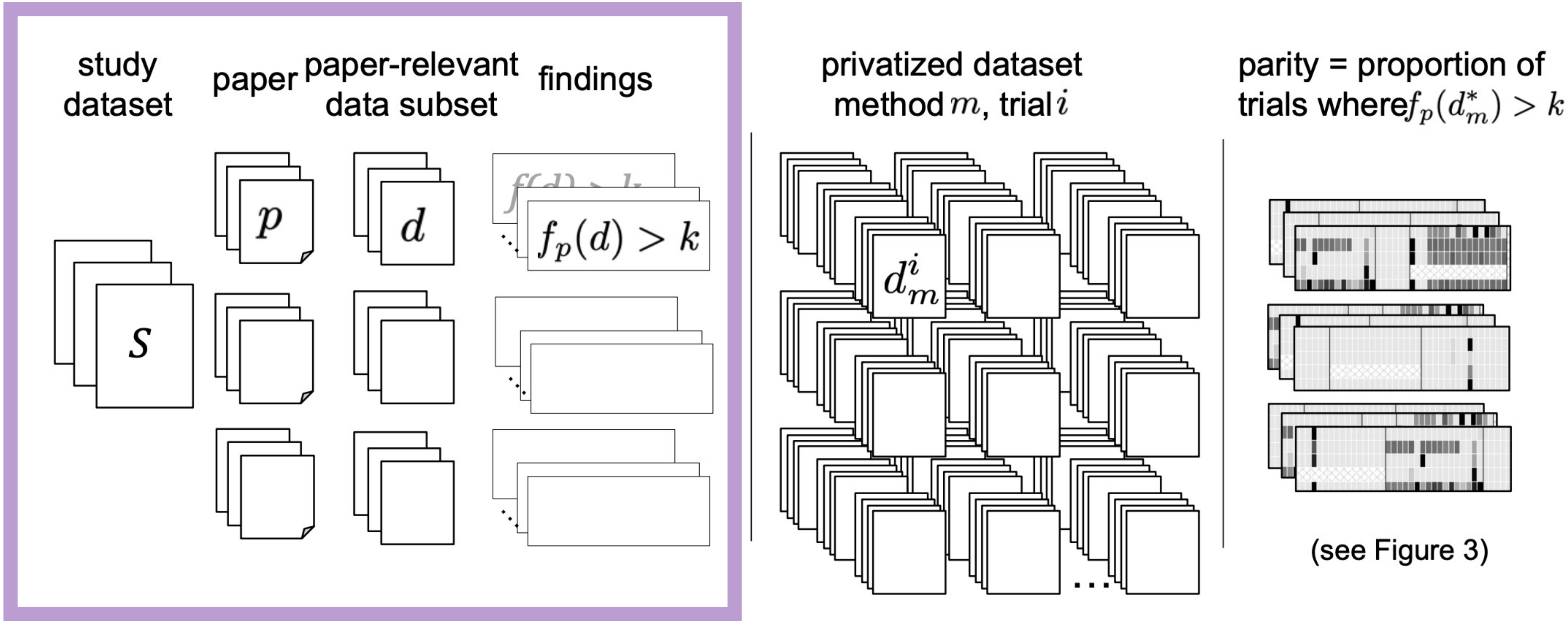
1. Findings are simply "reproduced or not"

$$finding(\tau, q_i^{synth}, q_i^{real}) = \mathbb{1}[|\tau(q_i^{synth}) - \tau(q_i^{real})| \leq \alpha]$$

2. Source of randomness 1 - Synthetic draw from fixed synthesizer
   Solution: Bootstrap over *B* samples from synthesizer (*B* is `big,' > 25).
3. Source of randomness 2 – fitting synthesizer (expensive!).
   Solution: Fit as many synthesizers as we can, aggregate and caveat that variance is underreported.

Thus, we report on the uncertainty relative to the real finding of the synthetic one, bootstrapping to estimate variance.

SynRD Composition



Four Studies => 8 Papers

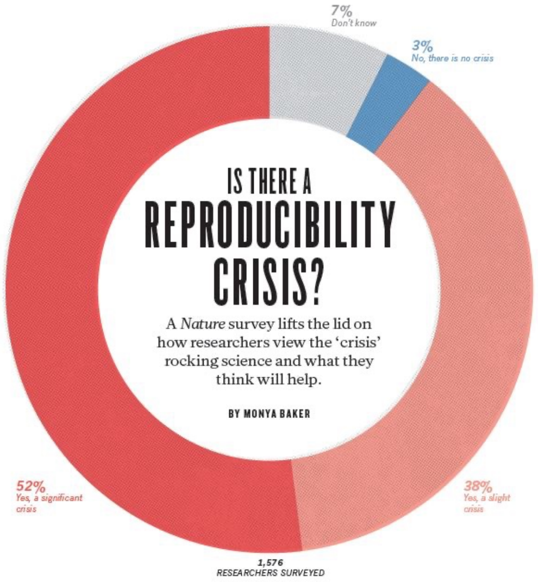
Studies

- **HSLS:09** (High School Longitudinal Study)
- **ACL** (Americans Changing Lives Survey)
- **AddHealth** (National Study of Adolescent and Adult Health)
- **NSDUH** (National Survey on Drug Use and Health)

8 Papers (8 different journals)

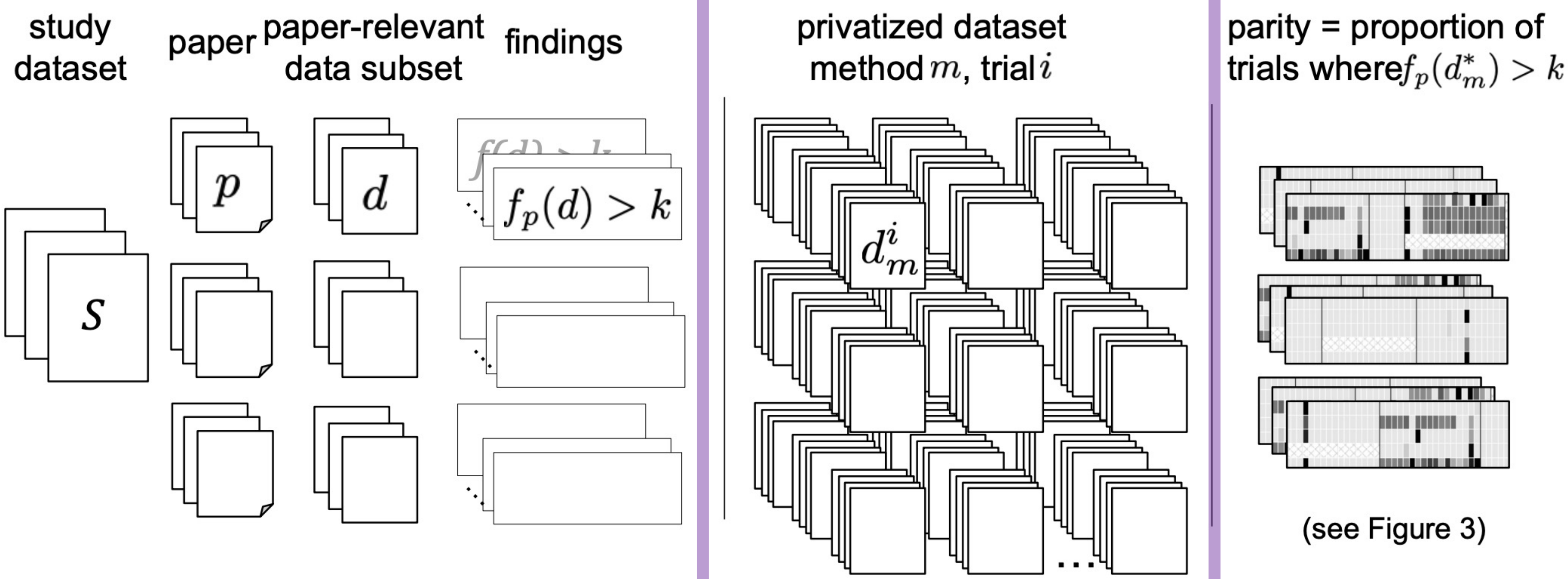
- Variety of methodologies
- Strict criteria for selection
- Reproducibility is hard!

	Descriptive Statistics	8
Regression	Between-Coefficients	4
	Fixed Coefficient (Sign)	2
Causal Paths	Variability	1
	Interaction	1
	Coefficient Difference	19
Logistic Regression	PBR	2
	FNR	2
	FPR	2
	Accuracy	2
Mean Difference	Between-Class	24
	Temporal (FC)	26
Correlation	Pearson	12
	Spearman	1



Most scientists agree, reproducible science isn't as common as it should be (Baker, Nature 2016)

SynRD Composition



Five Synthesizers => 4 ε regimes

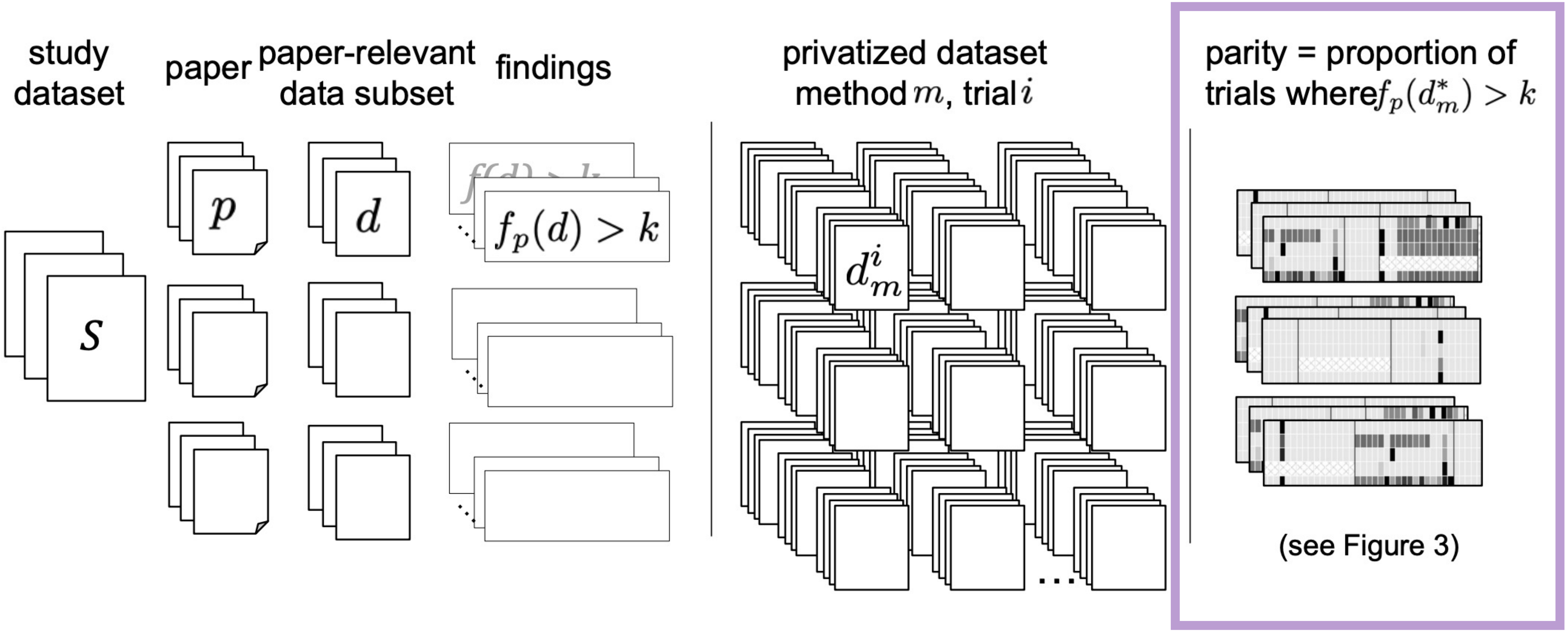
Synthesizers

	Budget-aware	Workload-aware	Data-aware	Efficiency-aware	Type?
PrivBayes	★		★	★	Bayesian
MST			★	★	Marginal (PGM)
PATECTGAN	★		★		GANs (Neural)
PrivMRF	★		★	★	Marginal (PGM)
AIM	★	★	★	★	Marginal (PGM)

ε = [e^-1, e^0, e^1, e^2]

- Here, e is scientific constant e (~2.72)
- Representative of “low to medium privacy” (informally)

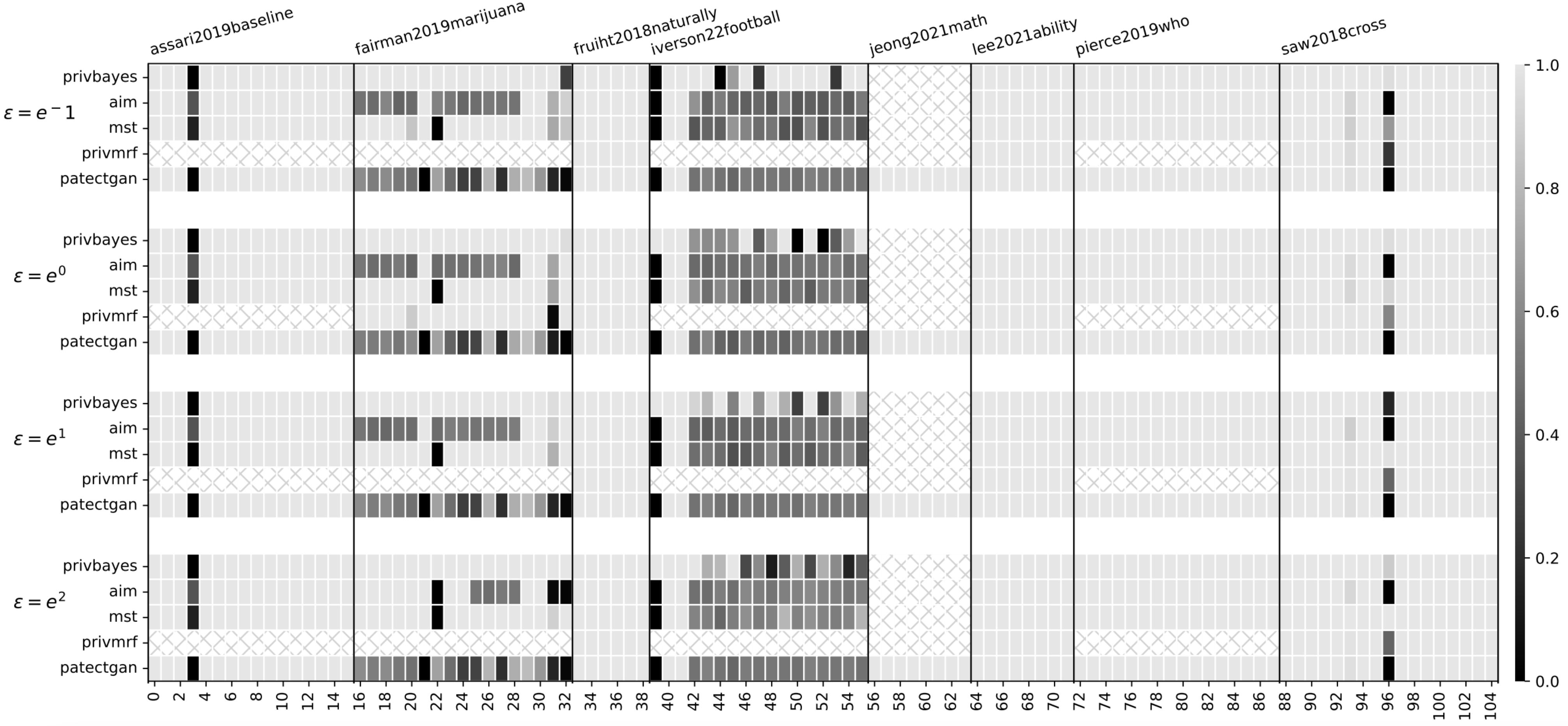
SynRD Composition



The benchmark!

```
from SynRD.papers import Saw2018Cross
from SynRD.benchmark import Benchmark
from SynRD.synthesizers import MSTSynthesizer
benchmark = Benchmark()
B = 25 # Bootstrap parameter
synth = MSTSynthesizer(epsilon=1.0)
papers = benchmark.initialize_papers([Saw2018Cross])
for paper in papers:
    synth.fit(paper.real_dataframe)
    dataset = synth.sample(len(paper.real_dataframe) * B)
    paper.set_synthetic_dataframe(dataset)
    benchmark.eval(paper, B=B)
```

Results



assari2019baseline
fairman2019marijuana
fruiht2018naturally
iverson22football
jeong2021math
lee2021ability
pierce2019who
saw2018cross
$\varepsilon = e^{-1}$
$\varepsilon = e^{0}$
$\varepsilon = e^{1}$
$\varepsilon = e^{2}$
privbayes
aim
mst
privmrf
patectgan
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98 100 102 104
0.0 0.2 0.4 0.6 0.8 1.0

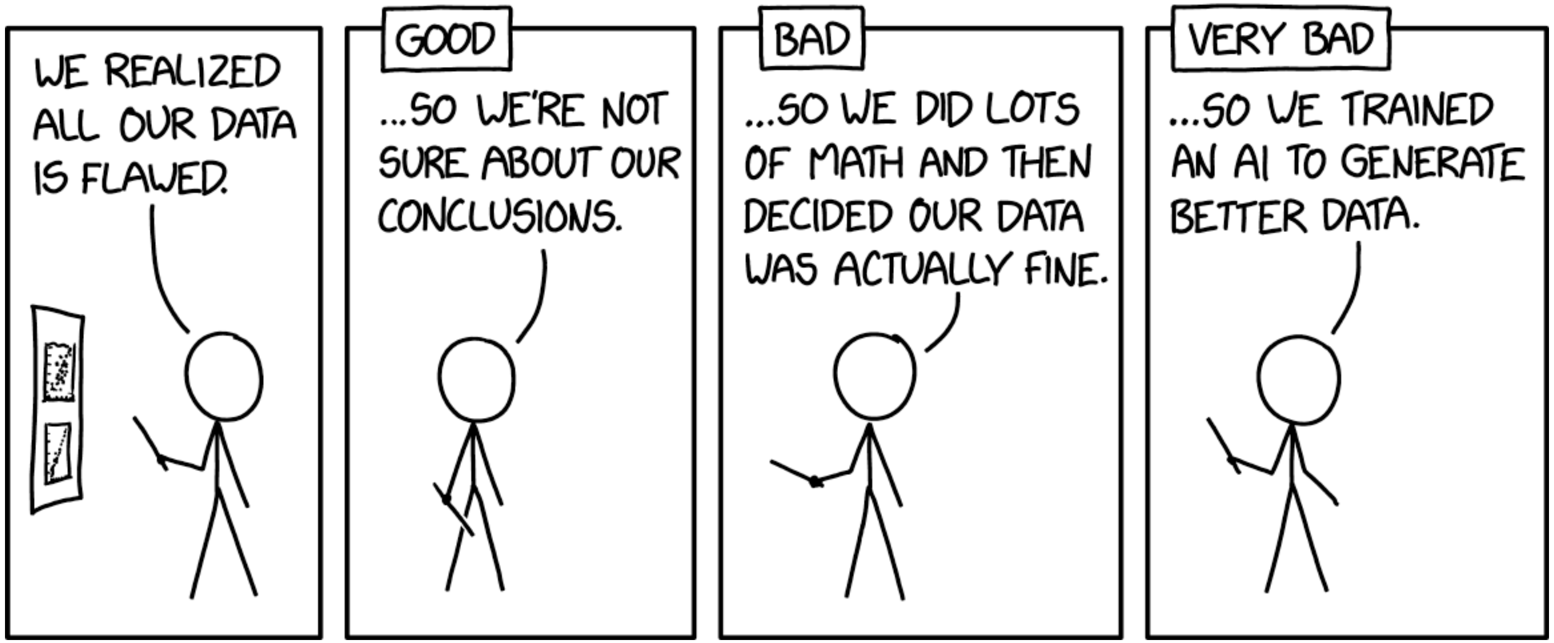
Results

- Overall performance of the synthesizers: impressive!
- Still, no synthesizer succeeded across all papers, and, remarkably, some findings were never reproduced by any of the synthesizers
- High number of findings across all our papers (even those that we were unable to replicate) relying only on 1- or 2-dimensional comparisons
- The low-dimensionality suggests that targeted improvements to the synthesizers may allow us to simultaneously support high utility for individual findings and their composition into broad conclusions

Future Work

- *Improving reproducibility and replicability of scientific discovery.*
  - File-drawer problem [29, 52] or publication bias - researchers publish positive results, negative results "end up in the researcher's drawer."
  - Epistemic parity could be extended to quantify the effect of DP noise in producing findings—which may or may not be false positives—that would not have been identified from the original data
- *Monte Carlo estimation of sample size for desired power for a particular finding*

Questions?



WE REALIZED ALL OUR DATA IS FLAWED.
GOOD
...SO WE'RE NOT SURE ABOUT OUR CONCLUSIONS.
BAD
...SO WE DID LOTS OF MATH AND THEN DECIDED OUR DATA WAS ACTUALLY FINE.
VERY BAD
...SO WE TRAINED AN AI TO GENERATE BETTER DATA.