How Researchers De-Identify Data in Practice

Wentao Guo, Paige Pepitone,[1] Adam Aviv,[2] Michelle Mazurek
University of Maryland, [1] NORC at the University of Chicago, [2] The George Washington University

✉ wguo5@umd.edu
@wentaoguo.bsky.social
@wentaochirps

Motivating examples

Medical researchers publish **clinical trial** data.

Scientists verify the **safety** of new treatments.

But data on **physical and mental health** could leak to insurance companies.

Aid organizations publish data about **program outcomes**.

Journalists cover the **impact** of taxpayer-funded programs.

But data on **political sentiments** could leak to local organized crime groups.

Social & medical scientists are increasingly **required** to de-identify and publish research data, despite the **difficulty of managing re-identification risk.**

Research scope

We interviewed…

- **18 practitioners** who had de-identified and published research data
- **6 curators** who review data submissions for repositories

research area

health	crime	int'l dev	other
8	4	9	3

# datasets de-identified

1–9	10–19	20+
6	6	6

*practitioners only

RQ1. How do researchers perceive **re-identification threats**?
RQ2. How do they **de-identify** data in practice?
RQ3. What **challenges** do they encounter?

icons: flaticon.com

RQ1 and RQ2. Mismatch between risk model and actual de-identification

Researchers are concerned about **combinations of indirect identifiers** that could link individuals to external data.

> You want to avoid putting clinicians into a **group of less than five similar clinicians**. Like, a 35-year-old Black endocrinologist from [a specific town]—there's probably just one.

> There might be a census block that **links back** to an external dataset. They're **one of now like 200** people.

quotes edited for brevity

In practice, researchers search for **distinctive values** and combinations of values. However, most only inspect **pairwise combinations of identifiers** (at most) and rely on **informal and social processes** for evaluating success.

1. Suppose we decide **age × occupation** is a particularly identifying combination.
2. Calculate crosstabs (2-way counts):

	18–24	25–29	30–34	⋯
Dentist	1	6	17	
Surgeon	0	2	7	
⋮				

3. Some counts are too low! Let's combine all three age categories into 18–34.
4. Repeat with different identifiers.

No evaluation of uniqueness by **age × occupation × race × gender × income × …**

> I get a bit into the weeds sometimes, and I'm like, "Ooh, they have two chickens, and **nobody else has two chickens**." And my boss is like, "Don't worry about it; there's a **very minute possibility** that somebody would go to this village, and they probably have more chickens now."

> You could crosstab all variables in theory, but that would be like millions of crosstabs. Maybe it's somebody's **position, crosstabbed with their age** or gender. It's not necessarily a scientific process. It's **more knowing what to look for**.

Why the mismatch?

1. Threats are seen as **unrealistic**.
2. Subsamples both mitigate risk and complicate de-ID.
3. Utility trade-offs are **unacceptable**.
4. Support and incentives are insufficient.

> I think **it is possible in many clinical datasets** to identify an individual, but the level of sophistication and effort you would need is **beyond the real threat**.

> We **really struggle with dates and time**. Every time you apply a date shift, you **severely limit the value** of your data.

Communication issues between curators & practitioners

- Practitioners experience minimal feedback on de-ID
- Practitioners often ghost curators after submitting data
- Curators generate distrust by asking for weaker measures

Curator:

> Data submitters can propose an access level, but it doesn't really matter, because **the repository has the final say**.

→

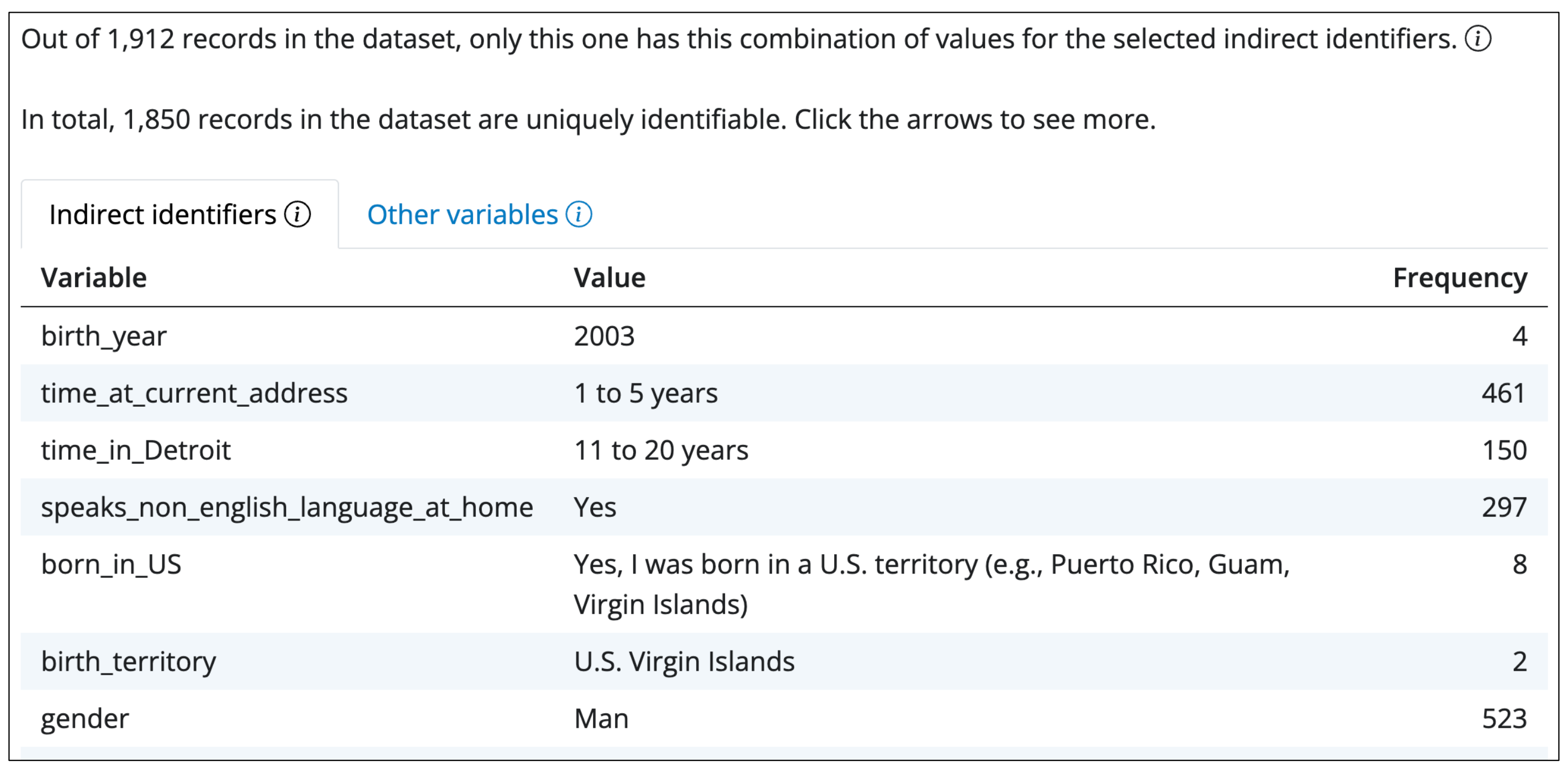
Practitioner:

> The data was basically **rendered useless** by the amount of de-identification we had to do. I could say I want the highest level of security, but **they don't have to do what I say**.

What next?

Researchers rarely assess risk across a whole set of identifiers…

➢ Build **design probes** to consider risk more comprehensively



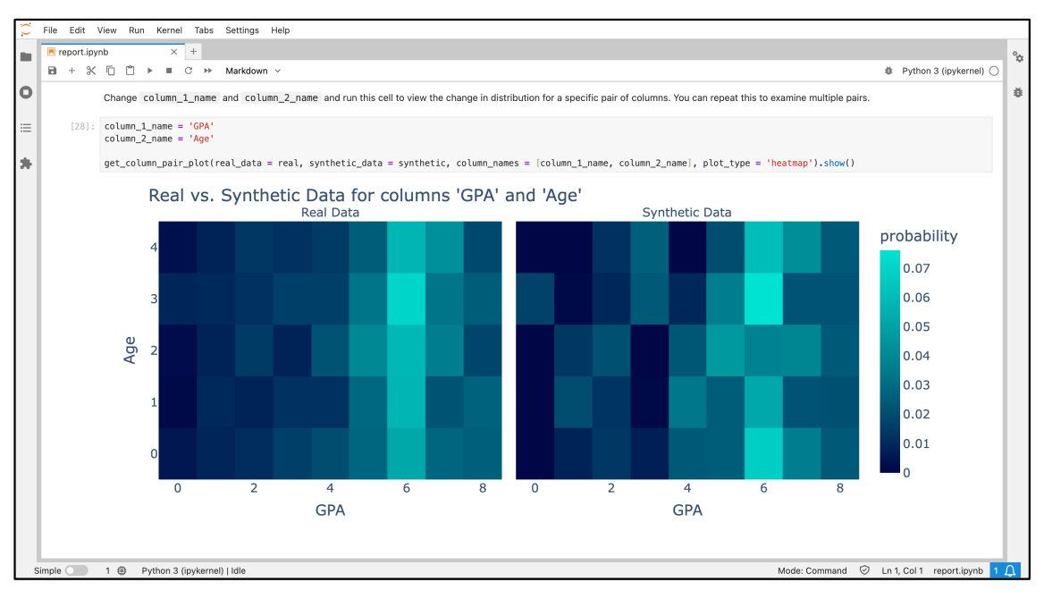
Out of 1,912 records in the dataset, only this one has this combination of values for the selected indirect identifiers. ⓘ

In total, 1,850 records in the dataset are uniquely identifiable. Click the arrows to see more.

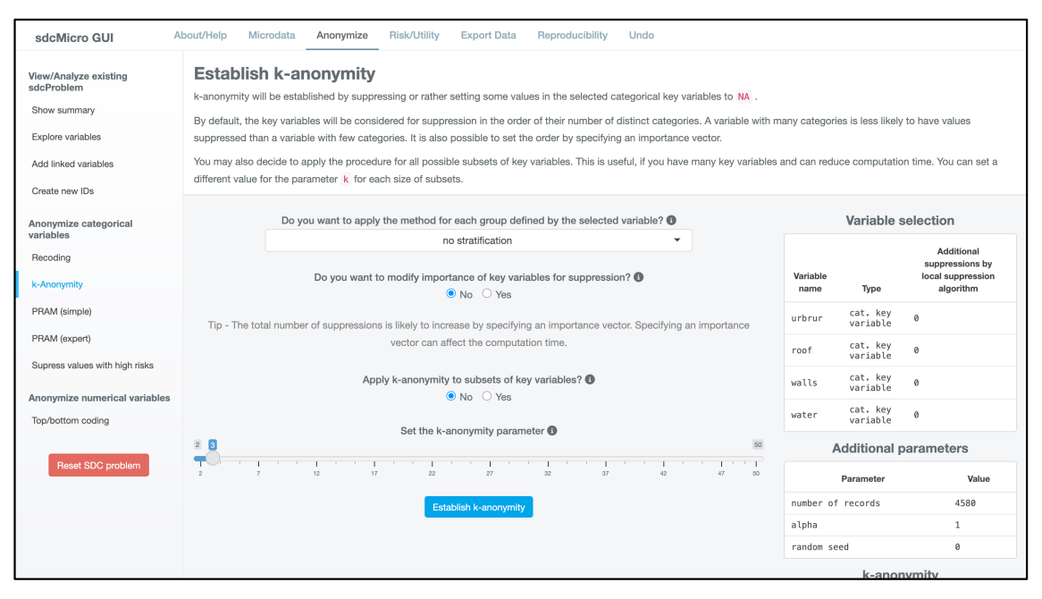
Variable	Value	Frequency
birth_year	2003	4
time_at_current_address	1 to 5 years	461
time_in_Detroit	11 to 20 years	150
speaks_non_english_language_at_home	Yes	297
born_in_US	Yes, I was born in a U.S. territory (e.g., Puerto Rico, Guam, Virgin Islands)	8
birth_territory	U.S. Virgin Islands	2
gender	Man	523

Researchers are open to more disruptive methods like differential privacy, but concerned about utility trade-offs…

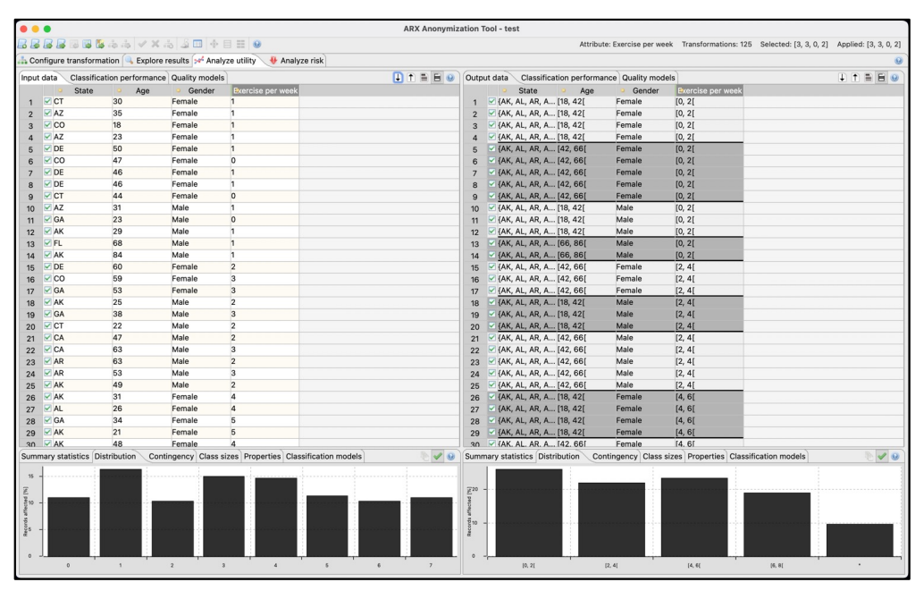
➢ Conduct exploratory **user studies** with existing tools



MST + SDMetrics



sdcMicro



ARX

This work was supported in part by NSF grant OAC-2232863.