A Qualitative Analysis of Practical De-Identification Guides

Wentao Guo, Aditya Kishore, Adam Aviv,[1] Michelle Mazurek

University of Maryland

[1]*The George Washington University*

✉ wguo5@umd.edu

@wentaochirps

Sharing data can benefit the public good

Pharmaceutical companies publish **clinical trial** data.

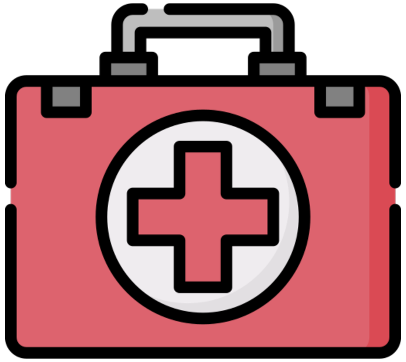
Scientists verify the **safety and effectiveness** of new treatments.

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

Journalists report on whether tax dollars are being spent **ethically and impactfully**.

all icons: flaticon.com

But data can also bring individuals harm



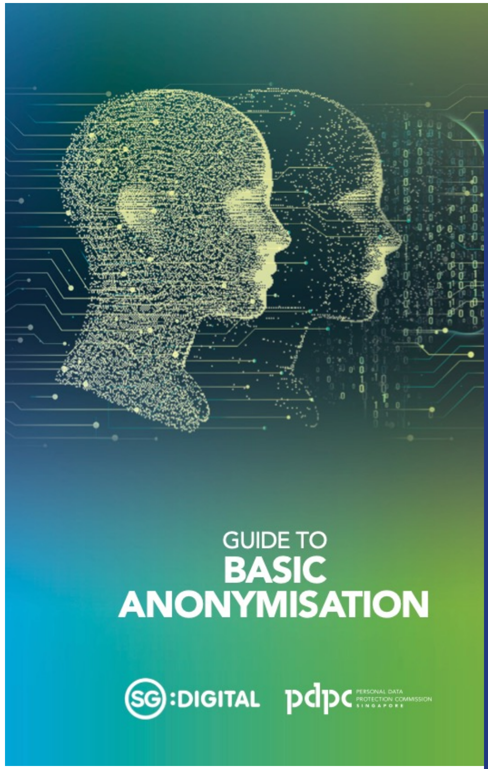
Clinical trial data could reveal participants' **physical and mental health** to employers and insurance companies.

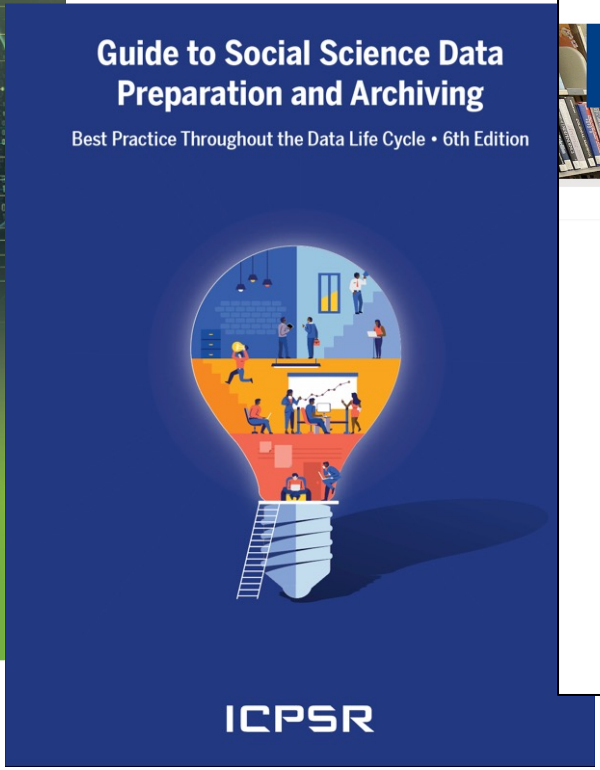
Foreign aid data could reveal participants' **political sentiments** to local organized crime and terrorism groups.

De-identifying data can protect individuals

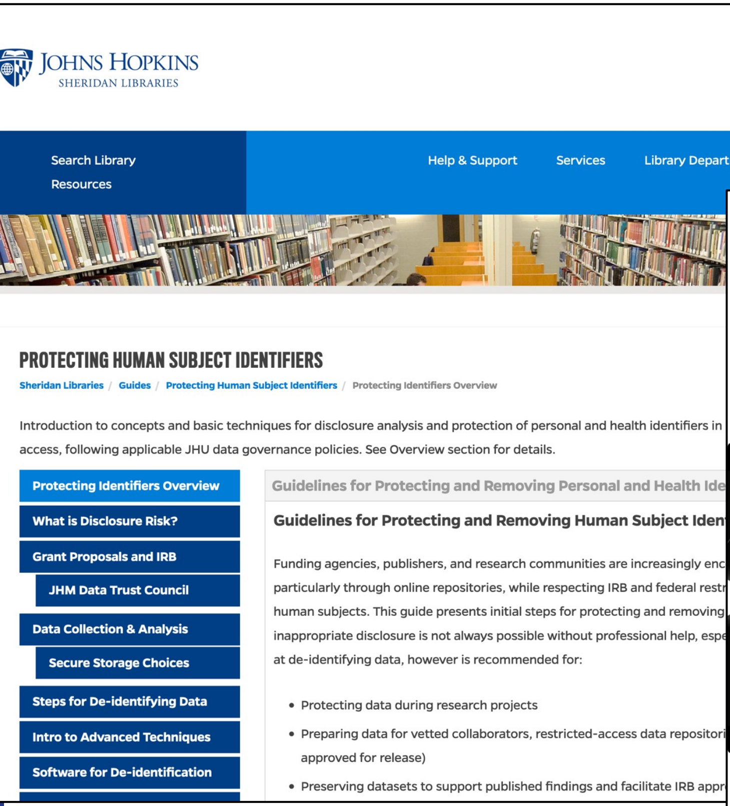
De-identification: modifying data to make it more difficult to re-identify or learn information about individuals

De-identifying data can protect individuals


GUIDE TO
BASIC
ANONYMISATION


Guide to Social Science Data Preparation and Archiving

Best Practice Throughout the Data Life Cycle • 6th Edition



PROTECTING HUMAN SUBJECT IDENTIFIERS

Sheridan Libraries / Guides / Protecting Human Subject Identifiers / Protecting Identifiers Overview

Introduction to concepts and basic techniques for disclosure analysis and protection of personal and health identifiers in access, following applicable JHU data governance policies. See Overview section for details.

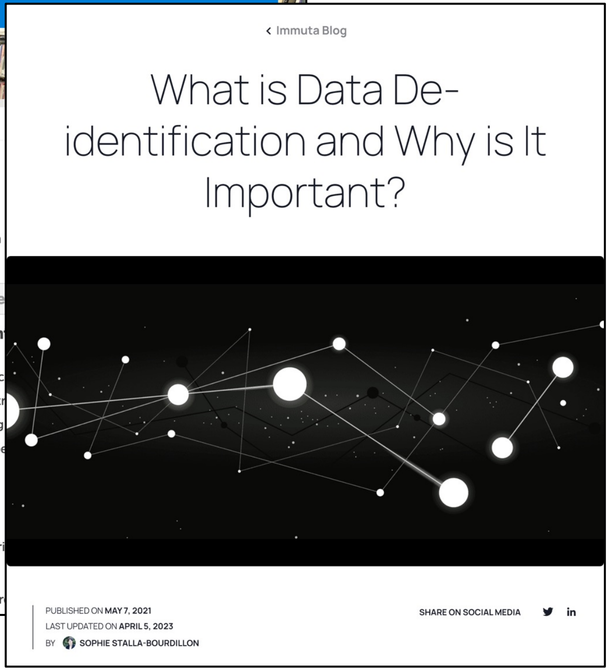
Protecting Identifiers Overview
What is Disclosure Risk?
Grant Proposals and IRB
JHM Data Trust Council
Data Collection & Analysis
Secure Storage Choices
Steps for De-identifying Data
Intro to Advanced Techniques
Software for De-identification

Guidelines for Protecting and Removing Personal and Health Ide

Guidelines for Protecting and Removing Human Subject Iden

Funding agencies, publishers, and research communities are increasingly enc particularly through online repositories, while respecting IRB and federal restr human subjects. This guide presents initial steps for protecting and removing inappropriate disclosure is not always possible without professional help, espe at de-identifying data, however is recommended for:

- Protecting data during research projects
- Preparing data for vetted collaborators, restricted-access data repositori approved for release)
- Preserving datasets to support published findings and facilitate IRB appr


‹ Immuta Blog

What is Data De-identification and Why is It Important?

PUBLISHED ON MAY 7, 2021
LAST UPDATED ON APRIL 5, 2023
BY SOPHIE STALLA-BOURDILLON

But practitioners need good guidance

Many de-id techniques and approaches

Delete data

Generalization

~~College Park~~

Maryland

Swap values

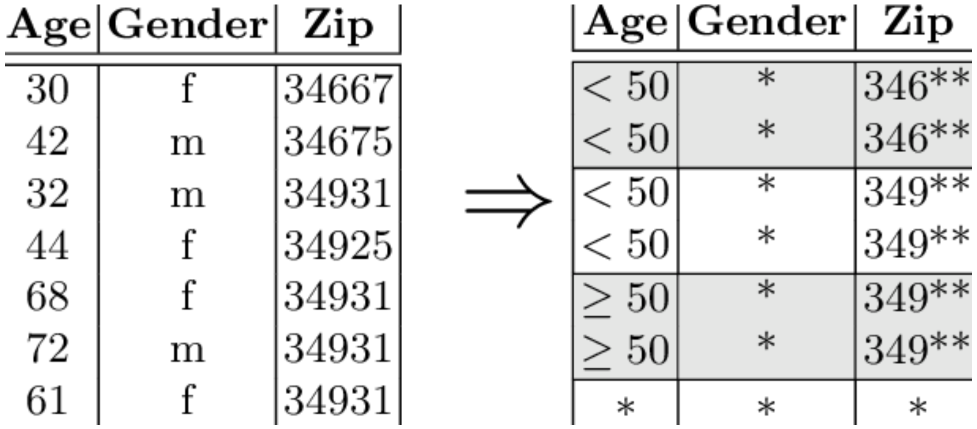
Add noise

2023-01-14

+ rand(n) =

2023-02-02

k-anonymity



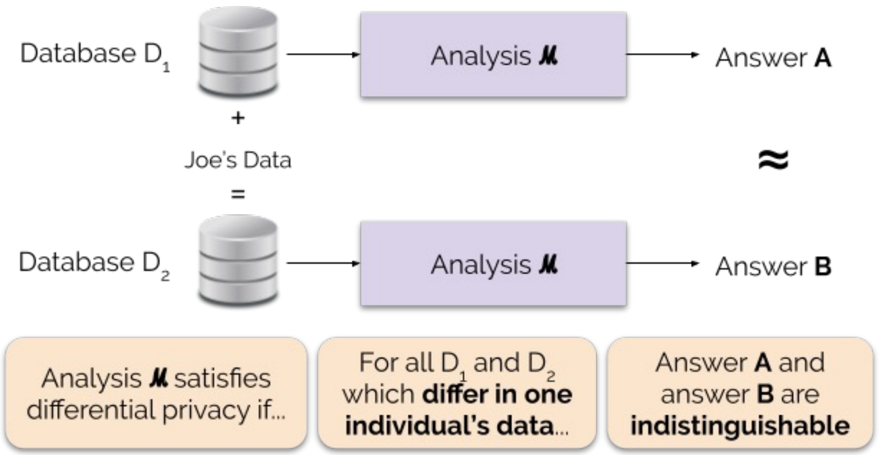
Age	Gender	Zip
30	f	34667
42	m	34675
32	m	34931
44	f	34925
68	f	34931
72	m	34931
61	f	34931

$\Rightarrow$

Age	Gender	Zip
< 50	*	346**
< 50	*	346**
< 50	*	349**
< 50	*	349**
≥ 50	*	349**
≥ 50	*	349**
*	*	*

Unal et al., https://doi.org/10.1016/B978-0-323-90570-1.00007-3

Differential privacy



Analysis $\mathcal{M}$ satisfies differential privacy if...

For all D_1 and D_2 which **differ in one individual's data**...

Answer **A** and answer **B** are **indistinguishable**

Gaikwad, https://doi.org/10.53469/jrse.2024.06(08).01

Achieving acceptable privacy is hard

Often involves significant technical expertise or manual effort

- Need to navigate various pitfalls that can undo intended protections

Balancing privacy with utility is complicated

- Impacts on downstream use cases are not well understood

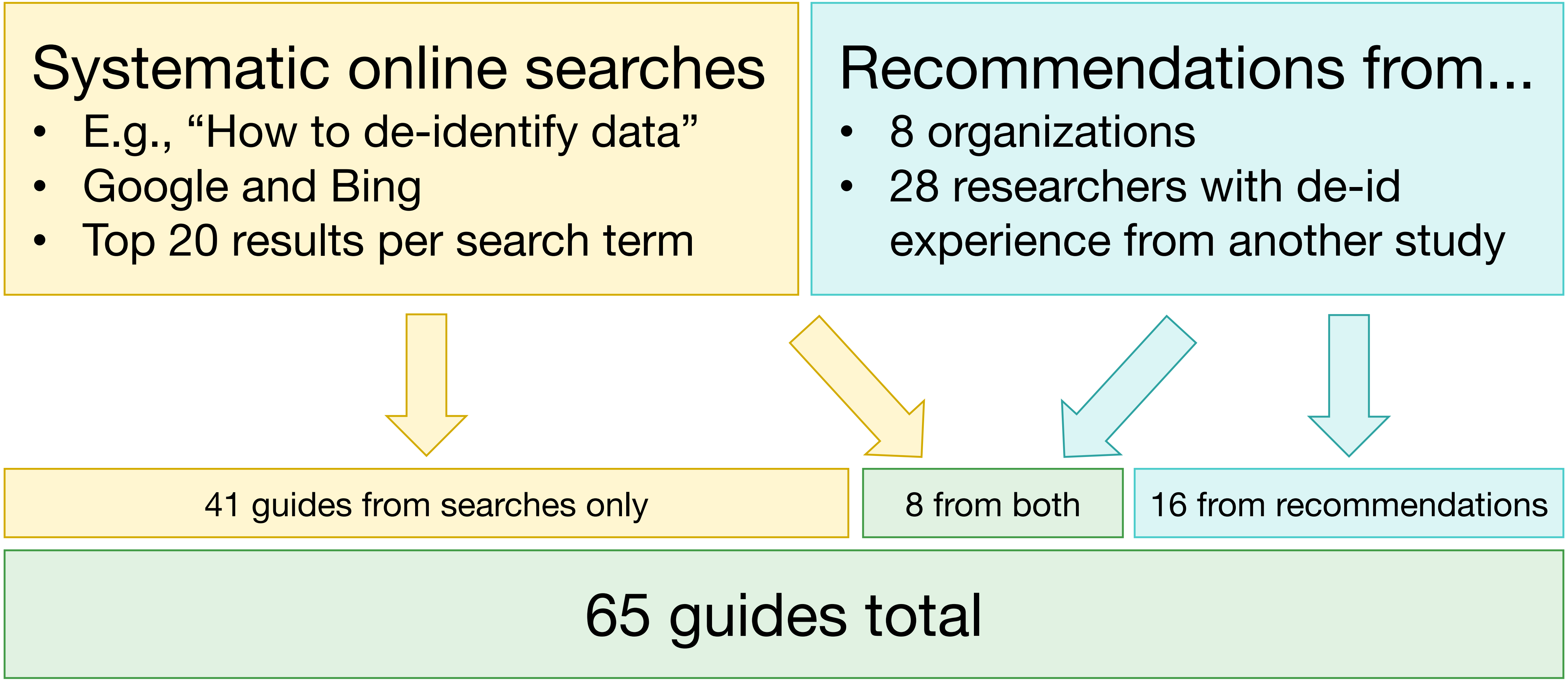
Research questions

1. What content do de-identification guides contain, particularly with regard to techniques and attacks?
2. Are guides designed to help readers decide on a de-identification strategy and carry it out?

Guide scope

- Updated 2018 or later
- Microdata (not aggregate statistics)
- For practitioners (not research papers)
- ...and more

Collecting de-id guides



Systematic online searches

- E.g., “How to de-identify data”
- Google and Bing
- Top 20 results per search term

Recommendations from...

- 8 organizations
- 28 researchers with de-id experience from another study

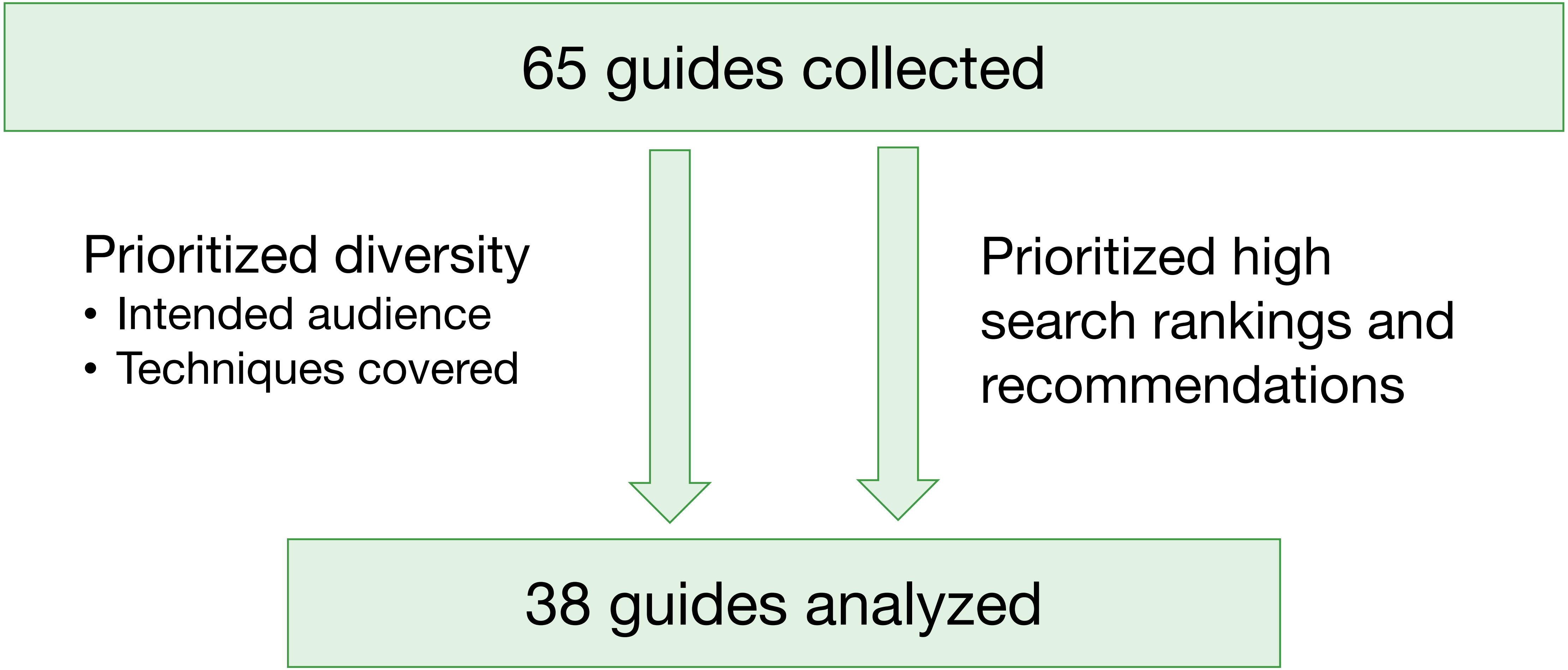
41 guides from searches only

8 from both

16 from recommendations

65 guides total

Sampling guides for analysis



65 guides collected

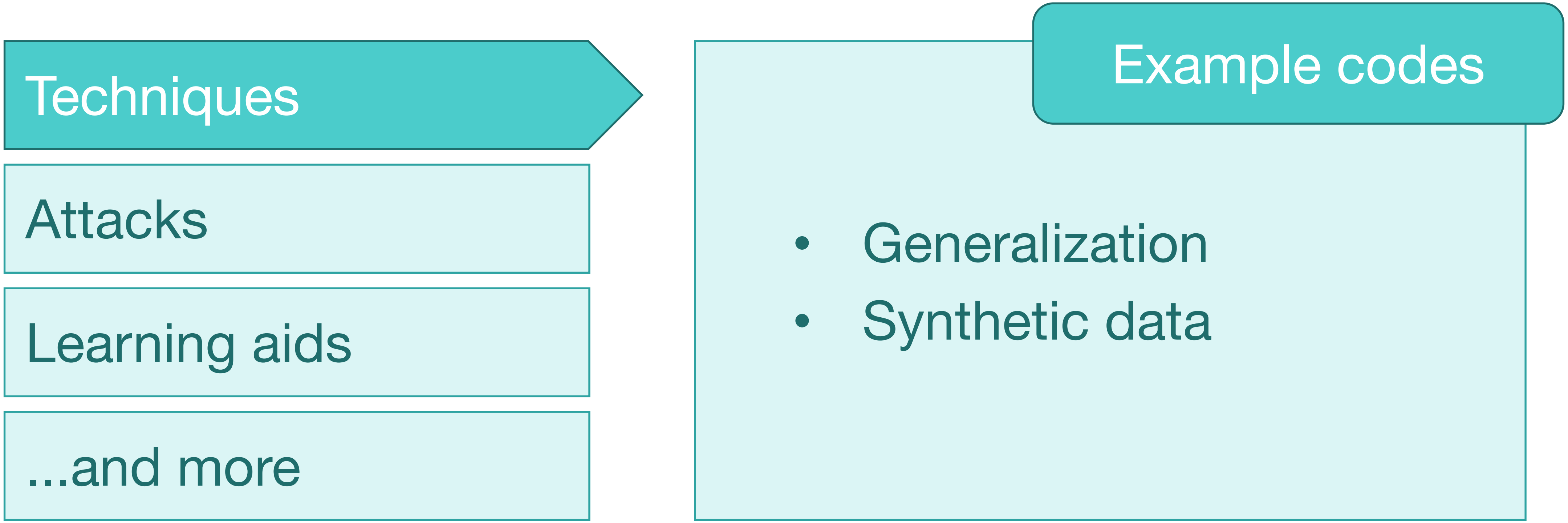
Prioritized diversity

- Intended audience
- Techniques covered

Prioritized high search rankings and recommendations

38 guides analyzed

Qualitative codebook



Techniques

Attacks

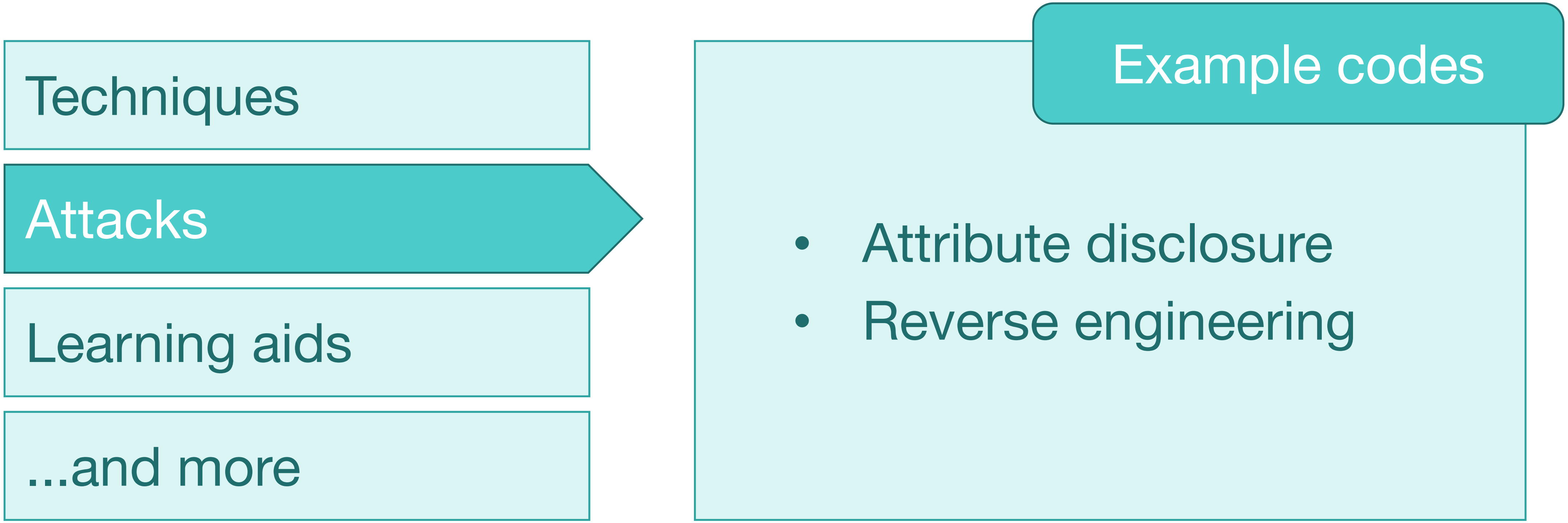
Learning aids

...and more

Example codes

- Generalization
- Synthetic data

Qualitative codebook

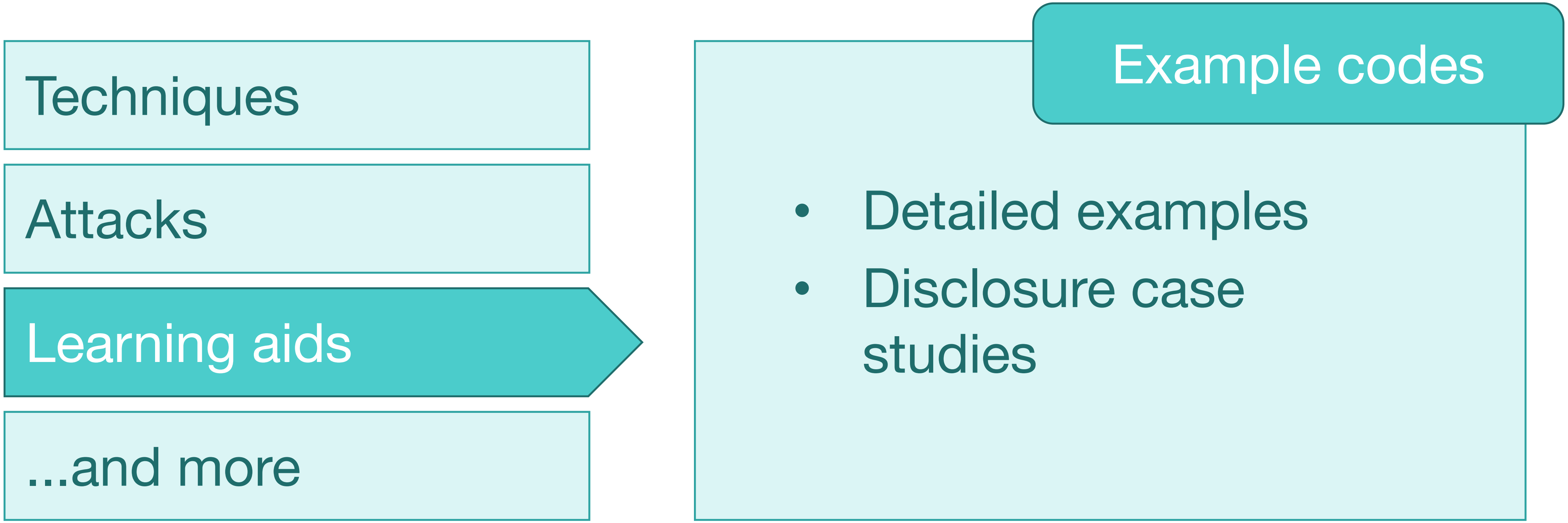


- Techniques
- Attacks
- Learning aids
- ...and more

Example codes

- Attribute disclosure
- Reverse engineering

Qualitative codebook



Techniques

Attacks

Learning aids

...and more

Example codes

- Detailed examples
- Disclosure case studies

Qualitative codebook

Techniques

Attacks

Learning aids

...and more

Example codes

- Access control
- Impossible to re-identify individuals

Coding process

Qualitative analysis with two coders

Coded one guide collaboratively to flesh out codebook structure

Double-coded all remaining guides separately

- Met regularly to resolve differences and update the codebook

RQ1: What content do guides contain (especially techniques and attacks)?

Different audiences get different content

	Researchers	Government agencies, businesses, and other
k-anonymity	2 out of 15 guides	15 out of 23 guides
Differential privacy	1 out of 15 guides	10 out of 23 guides

Different audiences get different content

	Researchers	Government agencies, businesses, and other
k-anonymity	■ ■ □ □ □ □ □ □ □ □ □ □ □ □ □	■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ □ □ □ □ □ □ □ □
Differential privacy	■ □ □ □ □ □ □ □ □ □ □ □ □ □ □	■ ■ ■ ■ ■ ■ ■ ■ ■ ■ □ □ □ □ □ □ □ □ □ □ □ □ □

Different audiences get different content

> There is as of yet **no easy to use, off-the-shelf tool that researchers can use to implement differential privacy**. Consequently, we do not recommend it at this stage, unless you are statistically proficient enough.
>
> – *Vrije Universiteit Brussel*

Different audiences get different content

There is as of yet **no easy to use, off-the-shelf tool that researchers can use to implement differential privacy.** Co

you

[Before our organization adopts differential privacy, we need to assess how well it applies to **the types of data we collect**, whether it is worth the **additional resources**, and if it matches **funders' expectations**.]

– anonymous guide

Inconsistent definitions

Anonymization . . . involves the **complete and irreversible** removal of any information from a dataset that could lead to an individual being identified.

– *The New School*

It is **not possible to say with certainty** that an individual will never be identified from a dataset which has been subjected to an anonymisation process.

– *Irish Data Prot. Comm.*

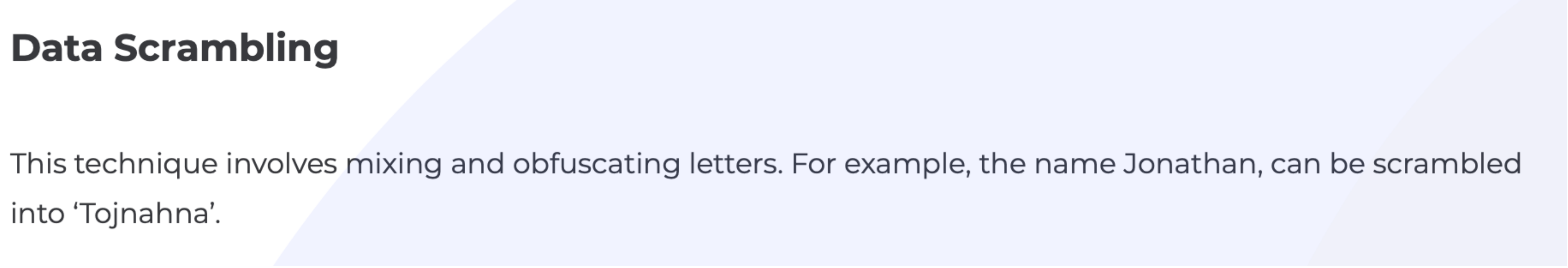
Also inference, aggregation, perturbation, and more

Gaps in threat coverage

Many guides cover *singling out* and *linking* as key concepts, but not *reverse engineering*

Guides lack details to help readers prevent reverse engineering

- Of 14 guides that discuss hashing, only 7 mention the importance of a salt
- Some suggest minimal randomness: e.g., shift all ages by the same offset



Data Scrambling

This technique involves mixing and obfuscating letters. For example, the name Jonathan, can be scrambled into ‘Tojnahna’.

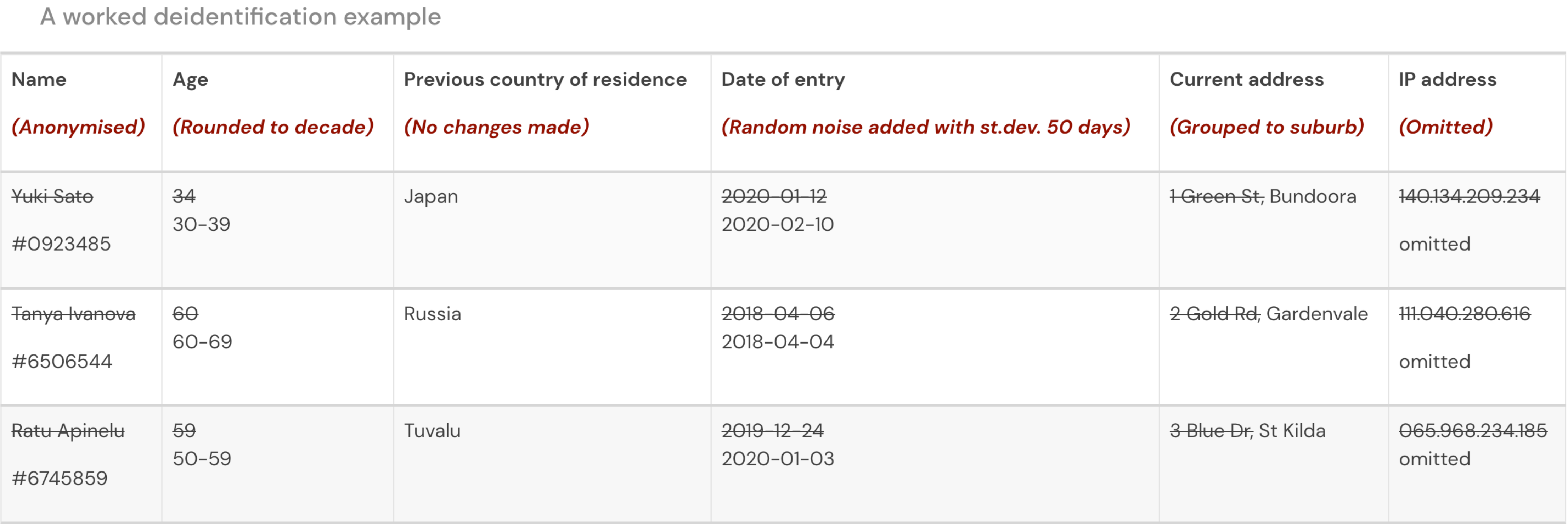
– Satori

RQ2: Are guides designed for usability?

Limited examples

Only 13 out of 38 guides contain *detailed examples*:

- Illustrating data both before and after de-id
- Meaningfully demonstrating de-id across multiple variables



A worked deidentification example

Name *(Anonymised)*	Age *(Rounded to decade)*	Previous country of residence *(No changes made)*	Date of entry *(Random noise added with st.dev. 50 days)*	Current address *(Grouped to suburb)*	IP address *(Omitted)*
~~Yuki Sato~~ #0923485	~~34~~ 30-39	Japan	~~2020-01-12~~ 2020-02-10	~~1 Green St,~~ Bundoora	~~140.134.209.234~~ omitted
~~Tanya Ivanova~~ #6506544	~~60~~ 60-69	Russia	~~2018-04-06~~ 2018-04-04	~~2 Gold Rd,~~ Gardenvale	~~111.040.280.616~~ omitted
~~Ratu Apinelu~~ #6745859	~~59~~ 50-59	Tuvalu	~~2019-12-24~~ 2020-01-03	~~3 Blue Dr,~~ St Kilda	~~065.968.234.185~~ omitted

– La Trobe University

A Qualitative Analysis of Practical De-Identification Guides

Wentao Guo, Aditya Kishore, Adam Aviv, Michelle Mazurek

We evaluated 38 de-id guides' content and usability.

We find notable differences in advice for different audiences, including discussion of barriers to differential privacy adoption.

We think de-id guides could be improved by...

- Explicitly noting potential confusion over inconsistent terms
- Discussing threats more systematically, especially reverse engineering
- Improving usability through more and better examples

✉ wguo5@umd.edu @wentaochirps