Empirically Measuring Privacy Over the NIST CRC Deidentified Data Archive

Christine Task, DJ Streat, Karan Bhagat
Knexus Research

Gary Howarth
NIST

Disclaimer

Portions of this talk are presented by a guest speaker. The contents of this presentation do not necessarily reflect the views or policies of the National Institute of Standards and Technology or the U.S. Government.

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Please note, unless mentioned in reference to a NIST Publication, all information and data presented is preliminary/in-progress and subject to change.

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

This talk focuses on a NIST metrology program for data **deidentification techniques**.

Deidentification includes any processing to microdata that produces microdata in the same schema and is *intended* to be resistant to individual reidentification: SDC, synthetic data, differential privacy.

Since February 2023 we've been running a massive community benchmarking and meta-analysis project, collecting metrics, algorithms and data samples from stakeholders, researchers and statistical agencies around the world— and making them all freely available and easy to use.

In this talk we'll be presenting our recent and upcoming work on **empirical privacy metrics**.

NIST’s Collaborative Research Cycle: Benchmarking Deidentification

Data: Diverse Community Excerpts Benchmark Data

Excerpts of 2019 American Community Survey Data
Tractable schema size for research: 22 Data Features + Weights
Curated to focus on geographies (PUMA) with challenging distributions
Recommended Feature Subsets provided for small schema approaches

Feature Name	Feature Description
PUMA	Public use microdata area code
AGEP	Person's age
SEX	Person's gender
MSP	Marital Status
HISP	Hispanic origin
RAC1P	Person's Race
NOC	Number of own children in household (unweighted)
NPF	Number of persons in family (unweighted)
HOUSING_TYPE	Housing unit or group quarters
OWN_RENT	Housing unit rented or owned
DENSITY	Population density among residents of each PUMA

INDP	Industry codes
INDP_CAT	Industry categories
EDU	Educational attainment
PINCP	Person's total income in dollars
PINCP_DECILE	Person's total income in 10-percentile bins
POVPIP	Income-to-poverty ratio (ex: 250 = 2.5 x poverty line)
DVET	Veteran service connected disability rating (percentage)
DREM	Cognitive difficulty
DPHY	Ambulatory (walking) difficulty
DEYE	Vision difficulty
DEAR	Hearing difficulty

The next few slides will use the "Demographic-focused" Feature Subset

https://github.com/usnistgov/SDNist/tree/main/nist%20diverse%20communities%20data%20excerpts

NIST’s Collaborative Research Cycle: Benchmarking Deidentification

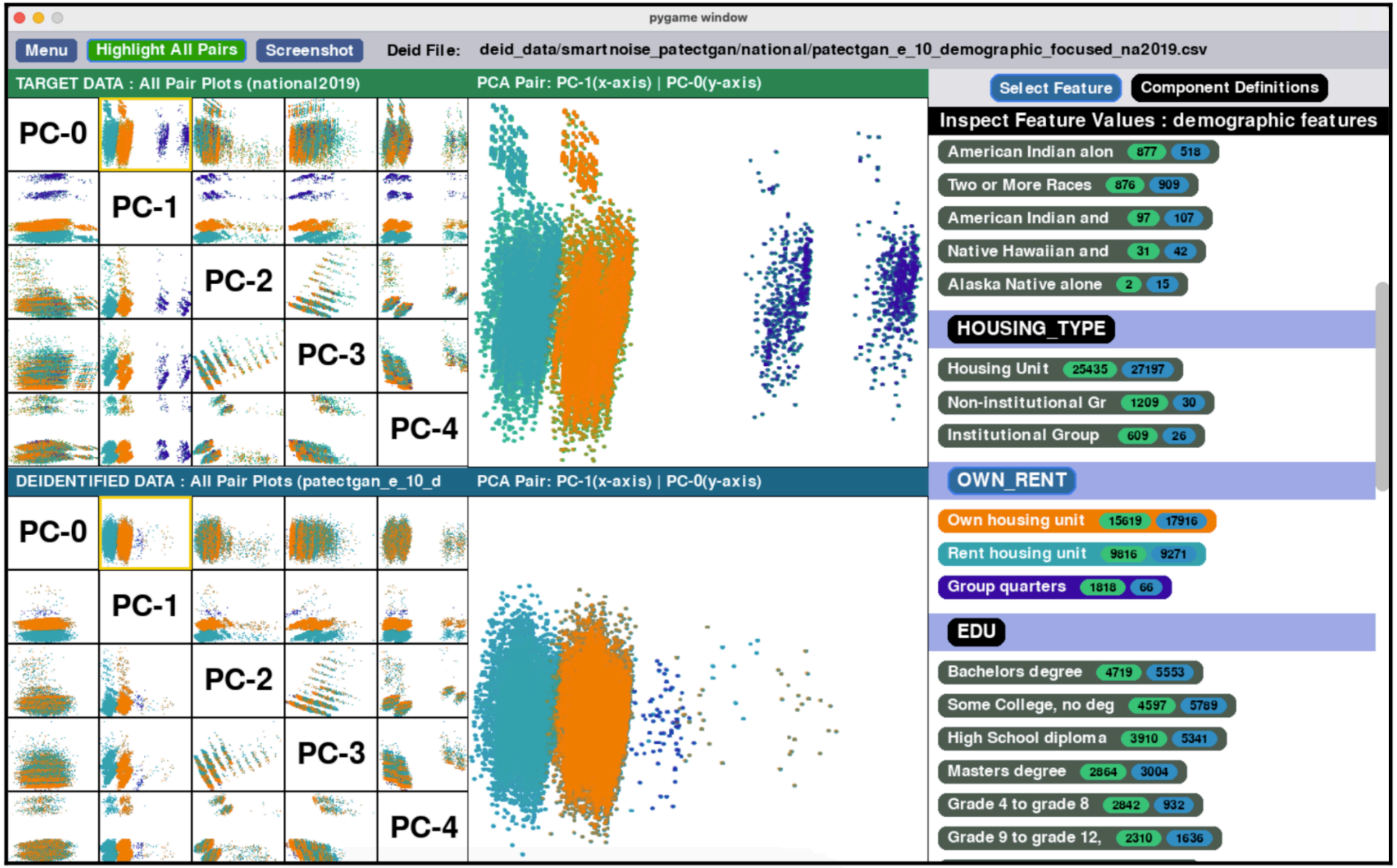
Pair-wise PCA Inspection Tool

Pairwise PCA is a relatively new visualization metric that was introduced by the IPUMS International team during the HLG-MOS Synthetic Data Test Drive.

It lets us look at the high dimensional data distribution using a set of 2D scatterplots along principle component axes. The plots look at the deidentified data and target data from the same angle (ie, using axes from the target data), so we can directly see where their distributions differ from each other.

The pairwise PCA tool lets you interactively explore these plots using a GUI interface.

You can install it by following the directions here:
https://github.com/usnistgov/pair-wise_PCA



NIST's Collaborative Research Cycle: Benchmarking Deidentification

The NIST Collaborative Research Cycle (CRC) Research Acceleration Bundle v1.1

- Direct download link for deidentified data and reports (537 MB)
- Direct download link for the metareports (484 MB)

Introduction

Welcome!

This repository contains deidentified data submitted to the CRC and their evaluation results as generated by SDNist v2.3.0. The CRC homepage provides more detailed information about the program, its goals, and how to participate.

In short, the CRC seeks to equip the research community with resources to explore, evaluate, and discuss deidentification approaches. The original data for this project are the NIST Diverse Communities Excerpts, curated data drawn from the American Community Survey.

There are three ground truth partitions, corresponding to three geographic regions (Boston-area (ma), Dallas-Fort Worth Area (tx), and a national sample (national). Submissions may include any or all of these partitions.

The original data contains 24 features. We also have a list of recommended reduced-size feature sets which can be found in the Excerpts Readme. Deidentified data may include any combination of feature set, though we have encouraged participants to use one of the recommended combinations to facilitate comparison of techniques.

What do we have here?

This repository contains the results of the first round of submissions. Additional submissions will be added with the next drop (expected in July 2023). The repository contains the navigable structure for the entire bundle. You can find all of the compressed data in Releases or you can use the links at the top of this readme.

The `crc-data-and-metrics-bundle` file contains:

- All of the deidentified data submissions and their evaluation metric results in the current release of our archive,
- An index.csv file that tracks all submission metadata, algorithm properties and definitions,
- A comprehensive set of tutorial jupyter notebooks and utilities that teach users how to programmatically explore the archive using the index file, and

Library	Algorithm	Team	#Entries	#Feature sets	Avg. Feat. Space Size	ε	Utility: SsE	Privacy Leak: UEM
rsynthpop	ipf_NonDP	Rsynthpop-categorical	1	1	3.405e+08		50.0	15.82
rsynthpop	catall_NonDP	Rsynthpop-categorical	1	1	2.270e+08		50.0	63.37
subsample_40pcnt	subsample_40pcnt	CRC	15	5	4.363e+25		40.67	39.93
rsynthpop	cart	CRC	12	4	3.457e+20		40.0	16.14
sdcmicro	pram	CRC	12	3	9.747e+10		38.33	56.27
MostlyAI SD	MostlyAI SD	MOSTLY AI	6	1	1.891e+26		30.0	0.01
rsynthpop	catall	Rsynthpop-categorical	6	1	2.270e+08	1, 10, 100	22.33	47.24
rsynthpop	cart	CBS-NL	3	1	2.270e+08		21.67	28.6
tumult	DPHist	CRC	5	2	5.732e+07	1, 2, 4, 10	18.8	92.14
smartnoise-synth	mst	CRC	36	5	3.781e+25	1, 5, 10	14.03	6.8
Genetic SD	Genetic SD	DataEvolution	19	2	9.454e+25	1, 10	11.84	0.11
LostInTheNoise	MWEM+PGM	LostInTheNoise	1	1	5.178e+26	1	10.0	0.0
synthcity	bayesian_network	CRC	12	4	5.672e+25		7.17	17.86
subsample_5pcnt	subsample_5pcnt	CRC	4	4	1.295e+26		5.0	4.97
Sarus SDG	Sarus SDG	Sarus	1	1	2.270e+08	10	5.0	13.99
sdv	ctgan	CBS-NL	6	1	1.891e+26		4.33	0.0

NIST's Collaborative Research Cycle: Benchmarking Deidentification

Deidentification Techniques Evaluated in this Talk:

Differentially Private Synthetic Data
- SmartNoise MST (E = 1, 10)
- SmartNoise AIM (E = 1, 10)

Non-DP Statistical Model Synthetic Data
- R Synthpop CART

Proprietary AI (non-DP) Synthetic Data
- AINDO
- MostlyAI
- Anonos

Statistical Disclosure Control
- SDCMicro Cell Suppression (k-6, small vs large quasi ID set)
- 40% Subsample

Sanity Checks: Complete ground truth data, complete withheld data

NIST’s Collaborative Research Cycle: Benchmarking Deidentification

Deidentification Techniques Evaluated in this Talk:

All synthetic data approaches shown here have reasonable utility on this data set

Library	Algorithm	Utility: Subsample Equivalence
aindo-synth	aindo-synth	30.0
Anonos Data Embassy SDK	Anonos Data Embassy SDK	20.0
MostlyAI SD	MostlyAI SD	30.0
rsynthpop	cart	40.0
smartnoise-synth	aim	20.0
smartnoise-synth	aim	60.0
smartnoise-synth	mst	10.0
smartnoise-synth	mst	10.0
sdcmicro	kanonymity all-features	1
sdcmicro	kanonymity-quasiID	20
subsample	subsample_40pcnt	40.0

The pieces of our problem

Data Record

The pieces of our problem

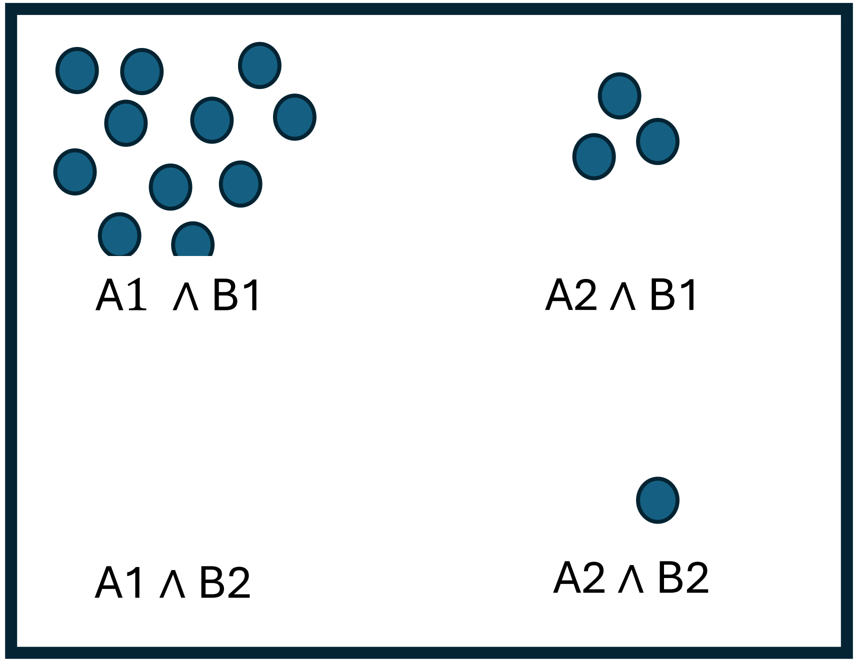
Features A

Features B

The pieces of our problem

Features A **Features B**

Input Real Data

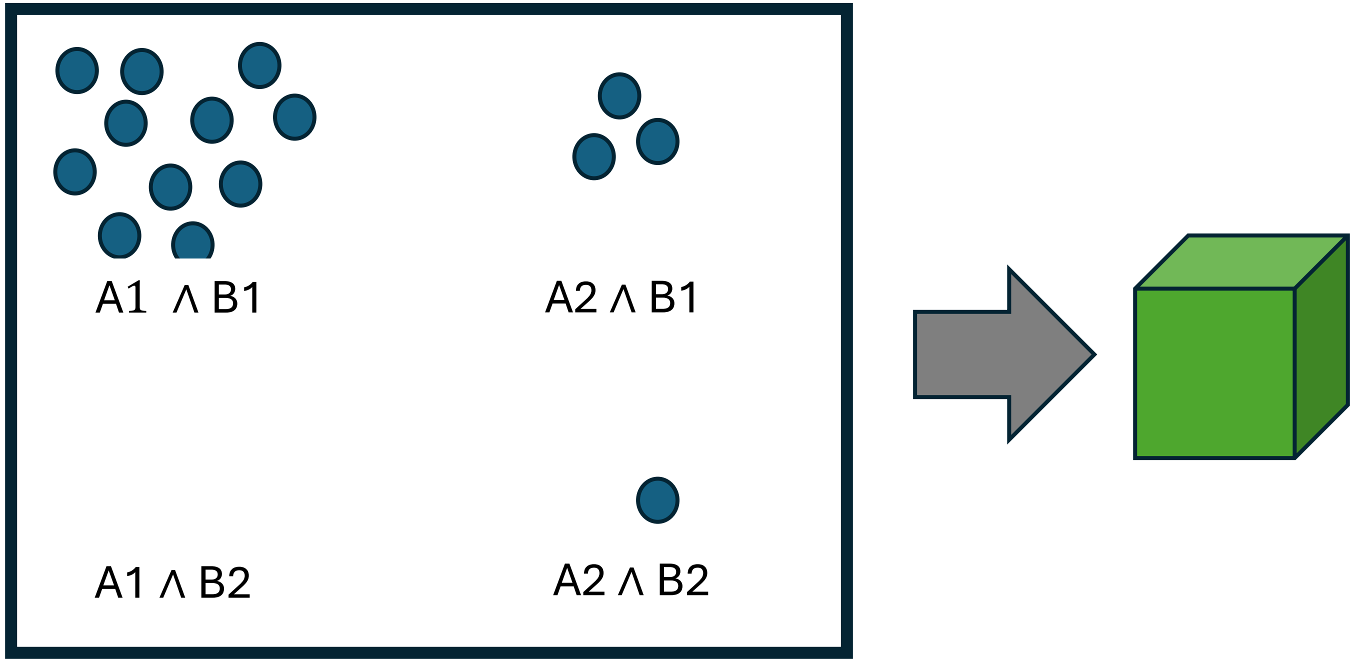


The pieces of our problem

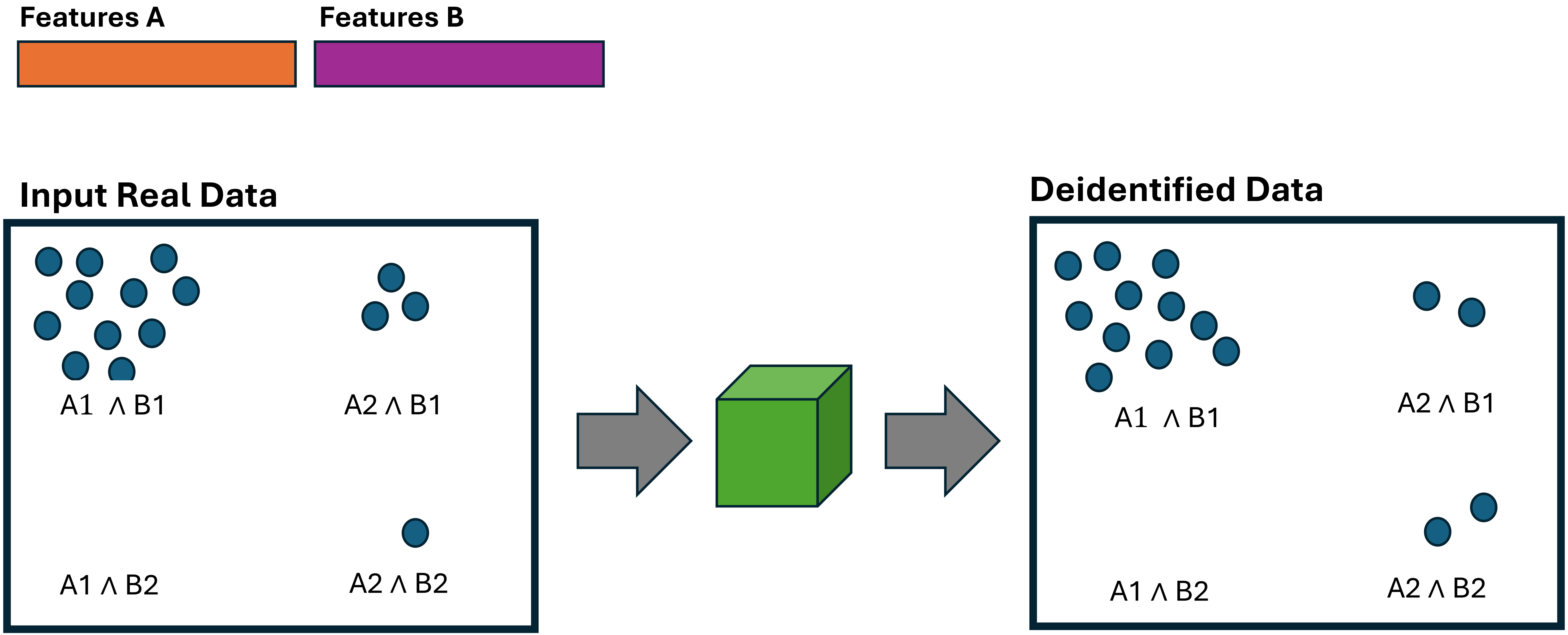
Features A

Features B

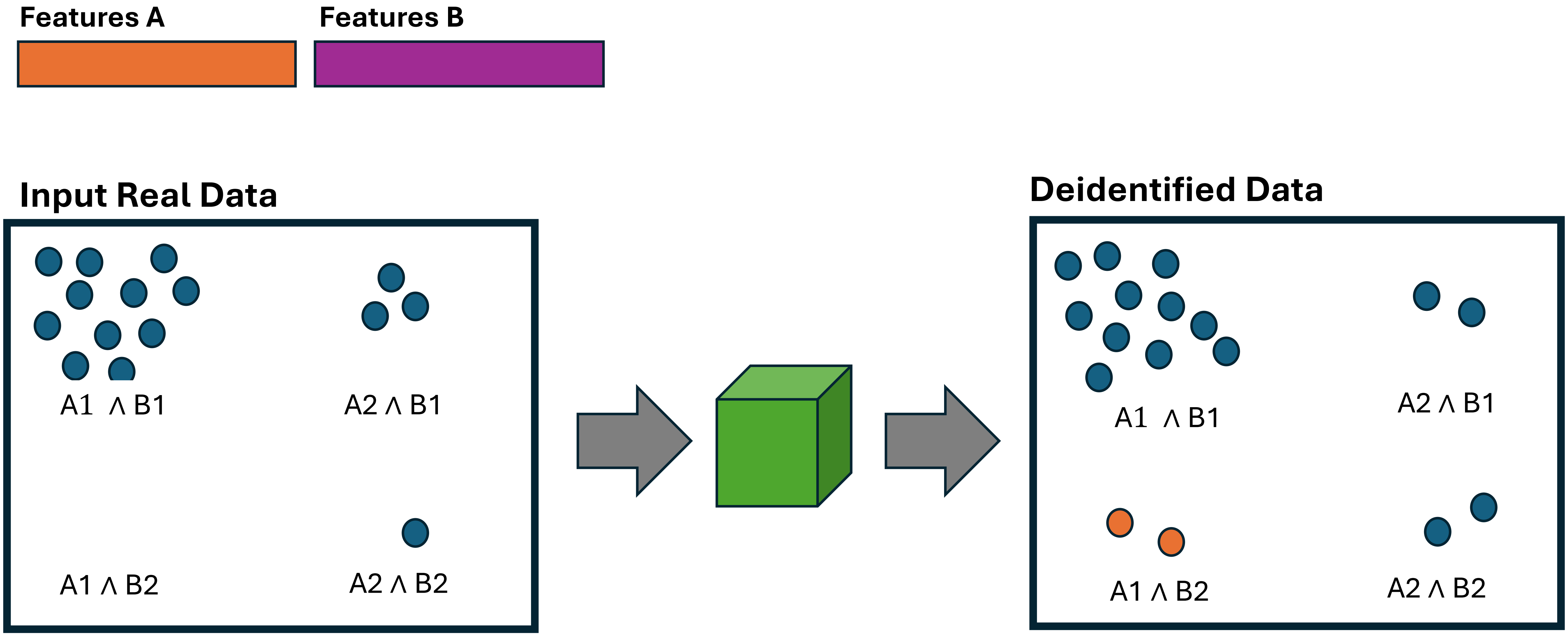
Input Real Data



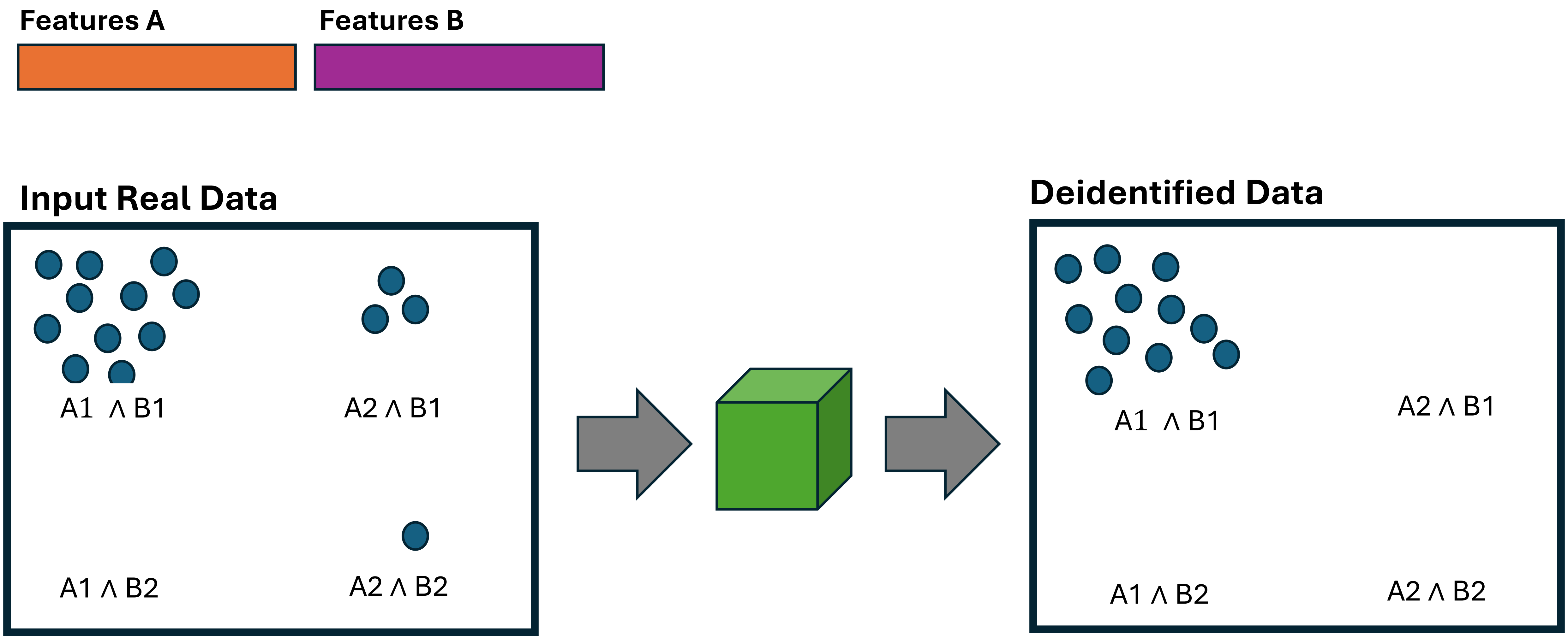
The pieces of our problem



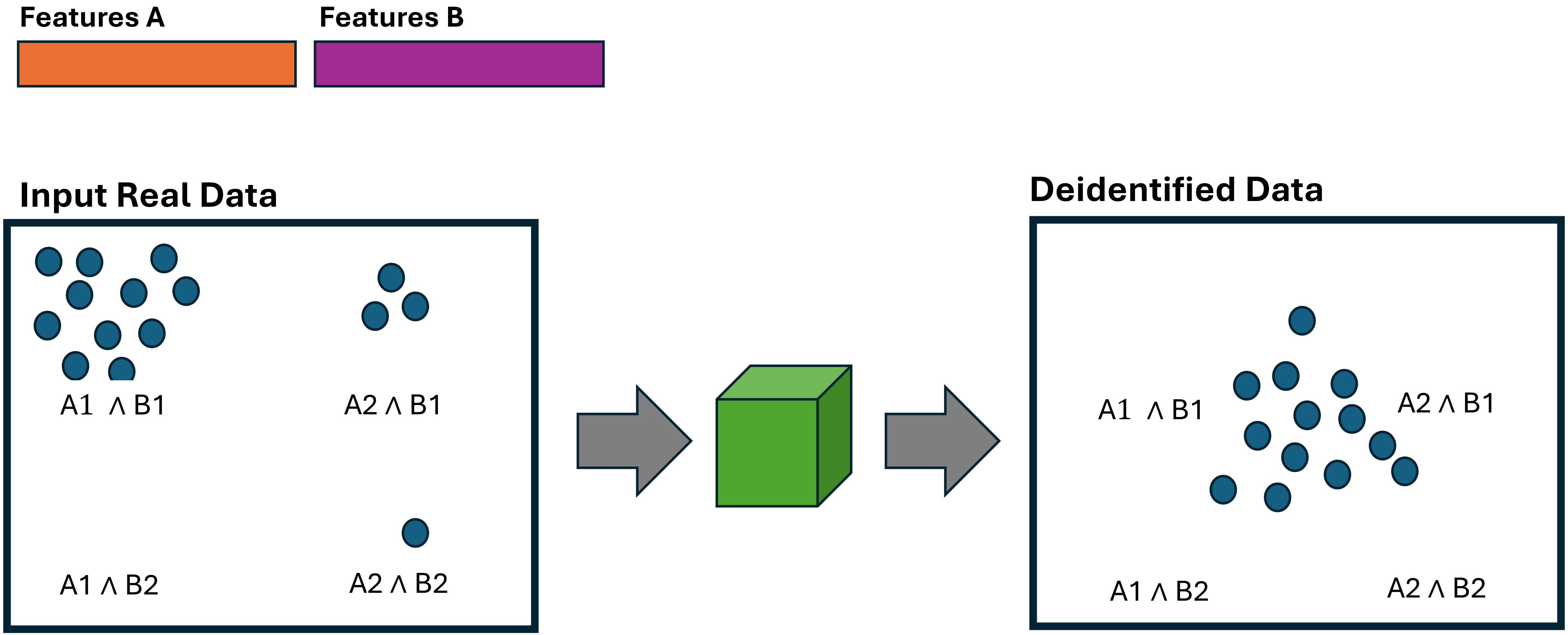
The pieces of our problem



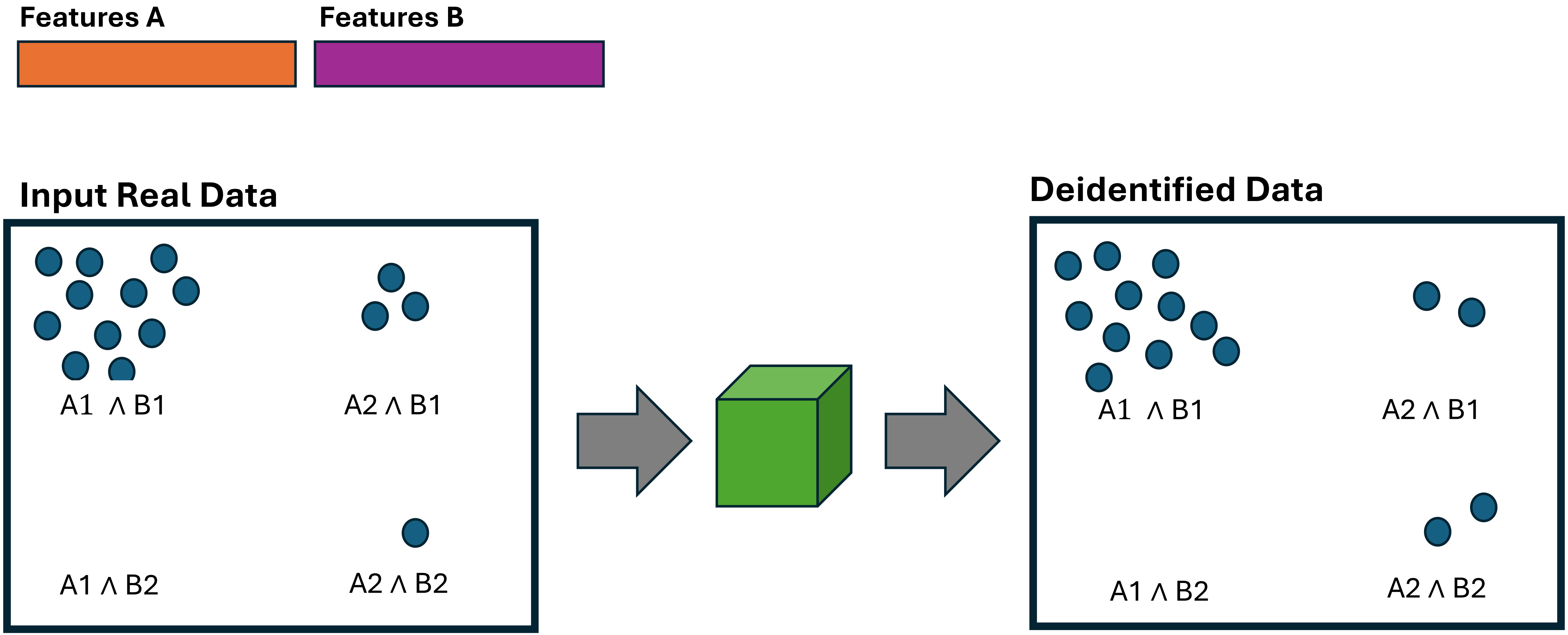
The pieces of our problem



The pieces of our problem

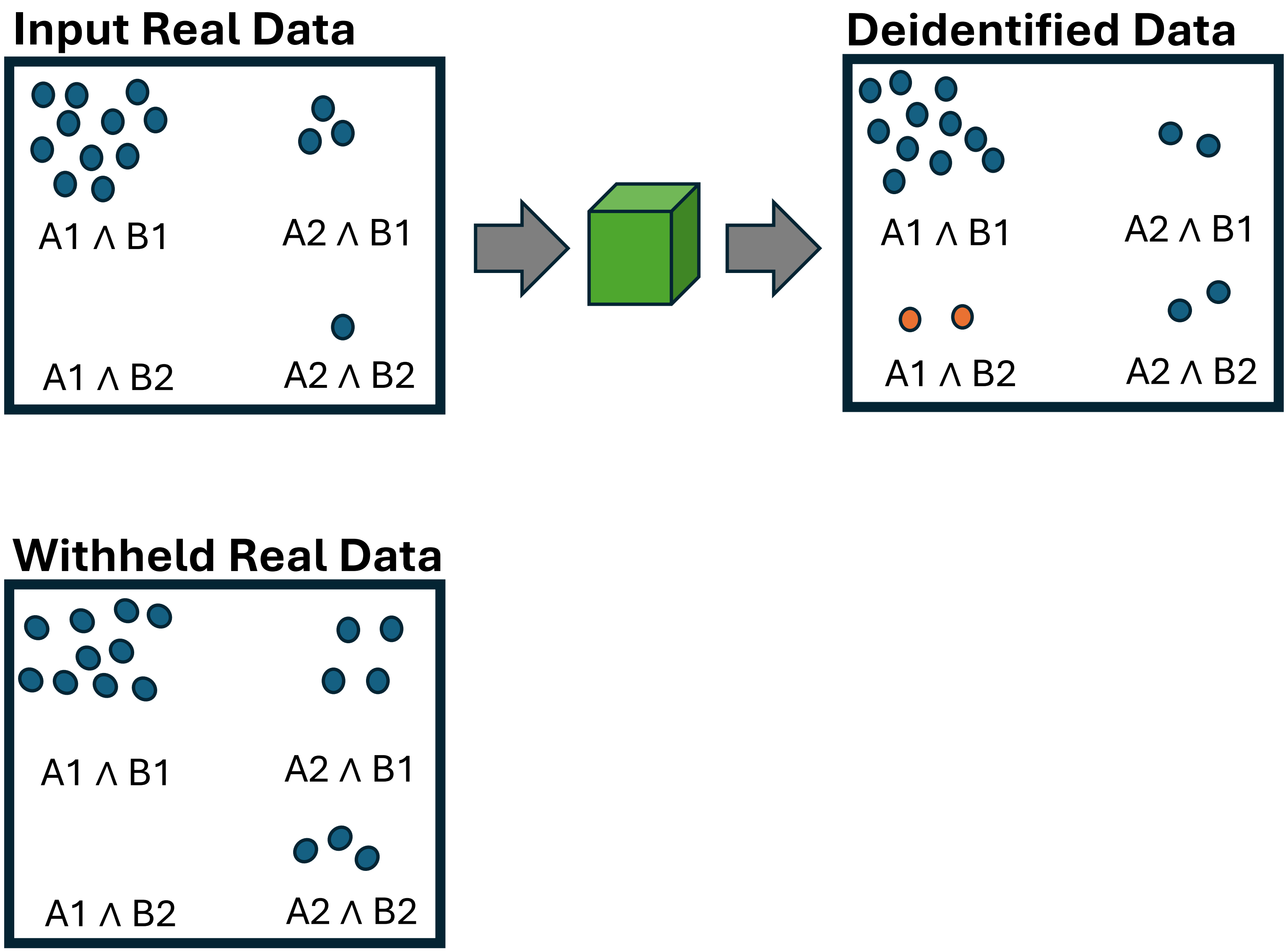


The pieces of our problem



The pieces of our problem

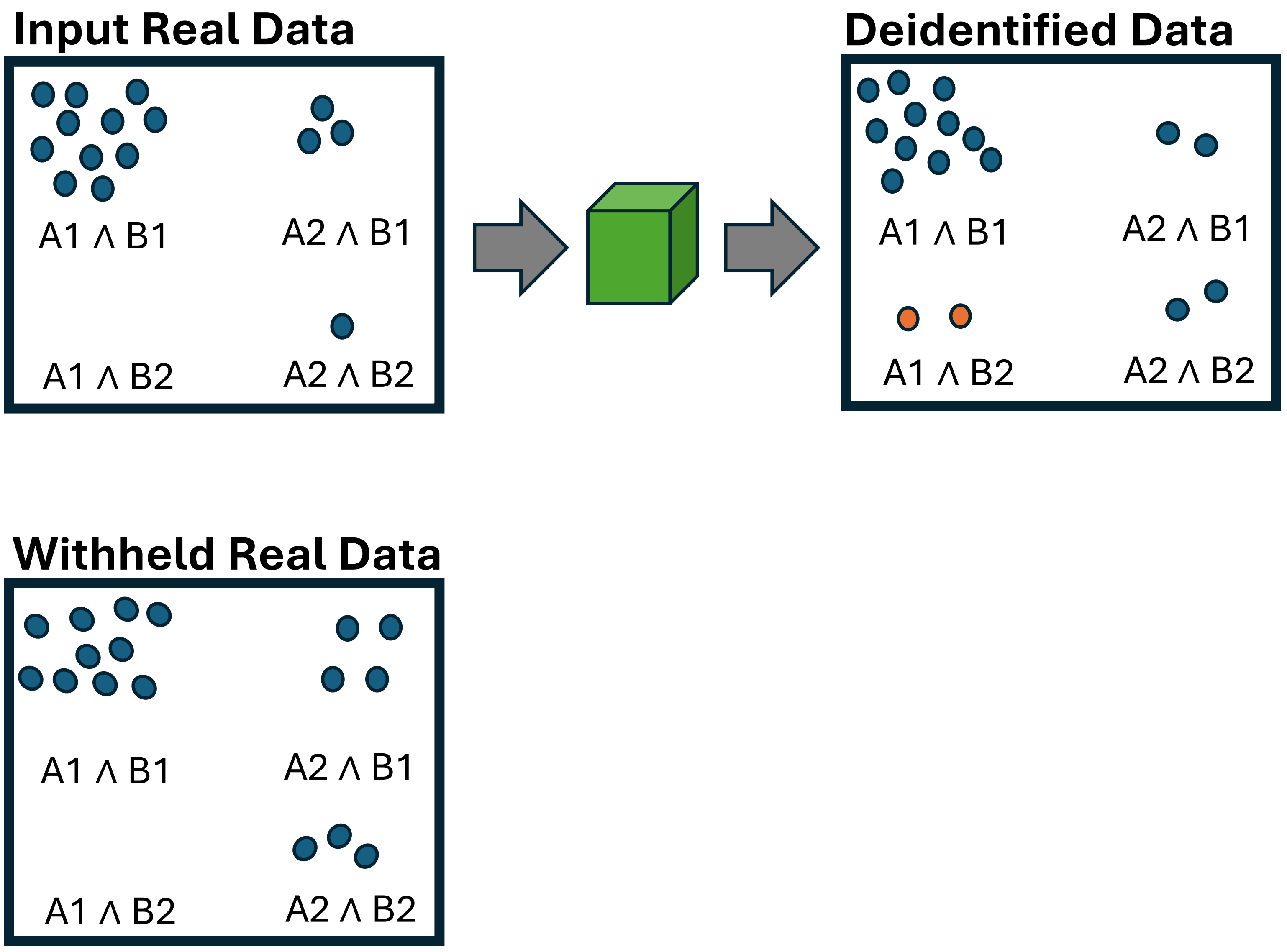
Measuring Privacy for Deidentified Data:



The pieces of our problem

Measuring Privacy for Deidentified Data:

Could anyone use the deidentified data to do something bad that would make us regret having released it?



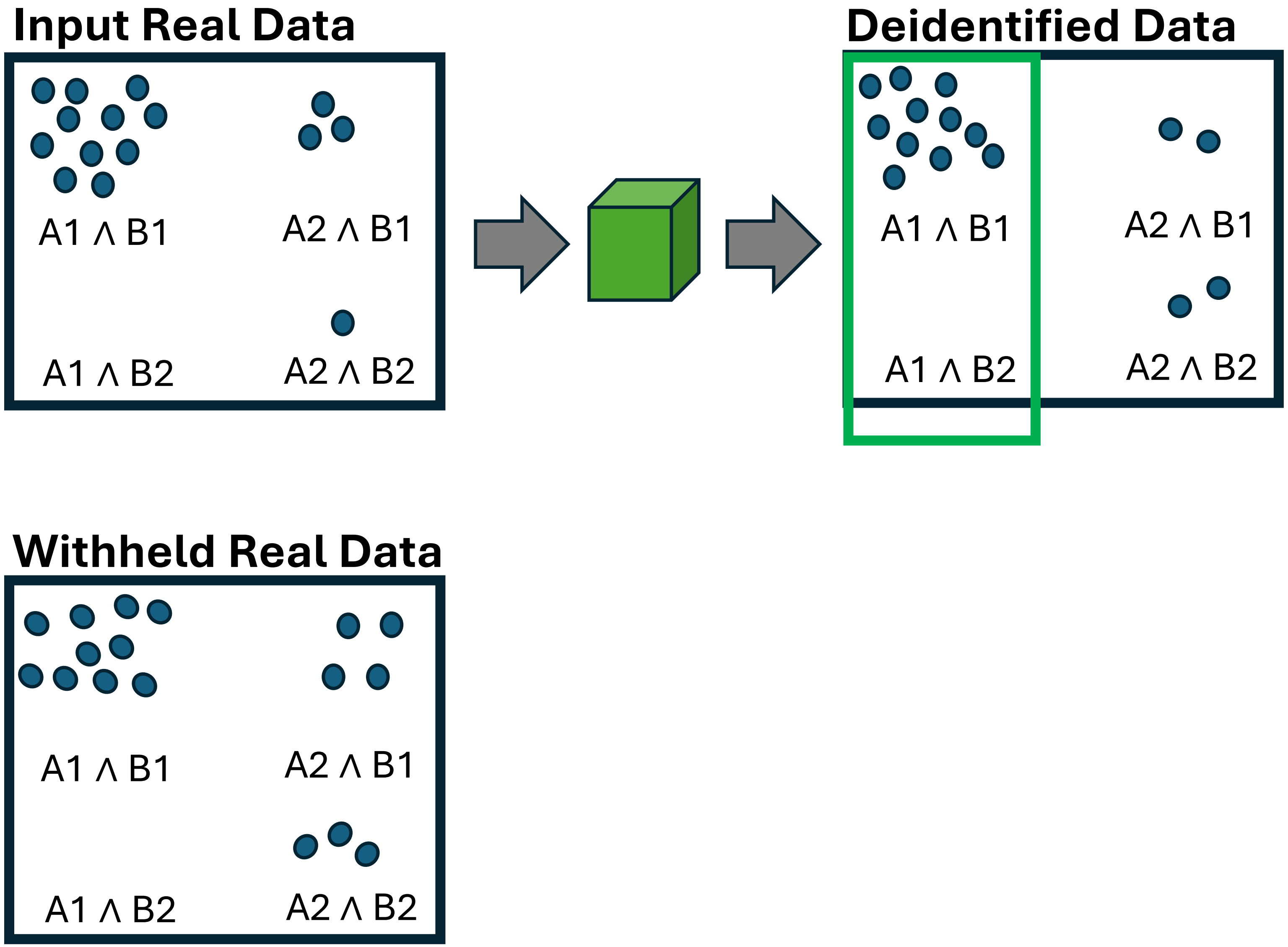
Can I use the released data to learn something invasive about you?

Attribute Privacy:

If I know someone's identifying features A, could I use the deidentified data to infer invasive features B

(1) better than I could using the withheld data? [Anonymeter: Inference]

(2) worse than I could using the input real data? [Synthpop: DiSCO]



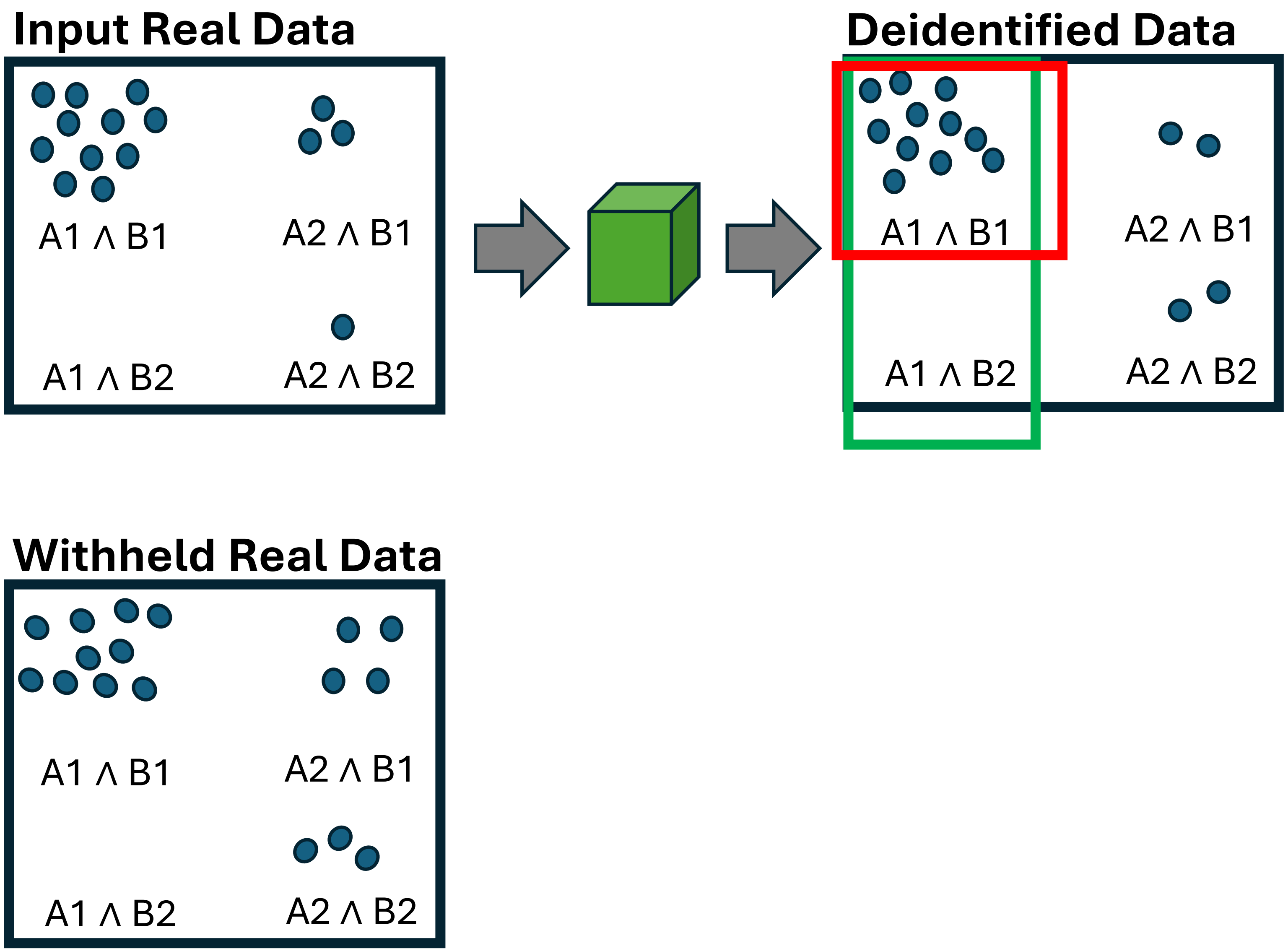
Can I use the released data to learn something invasive about you?

Attribute Privacy:

If I know someone's identifying features A, could I use the deidentified data to infer invasive features B

(1) better than I could using the withheld data? [Anonymeter: Inference]

(2) worse than I could using the input real data? [Synthpop: DiSCO]



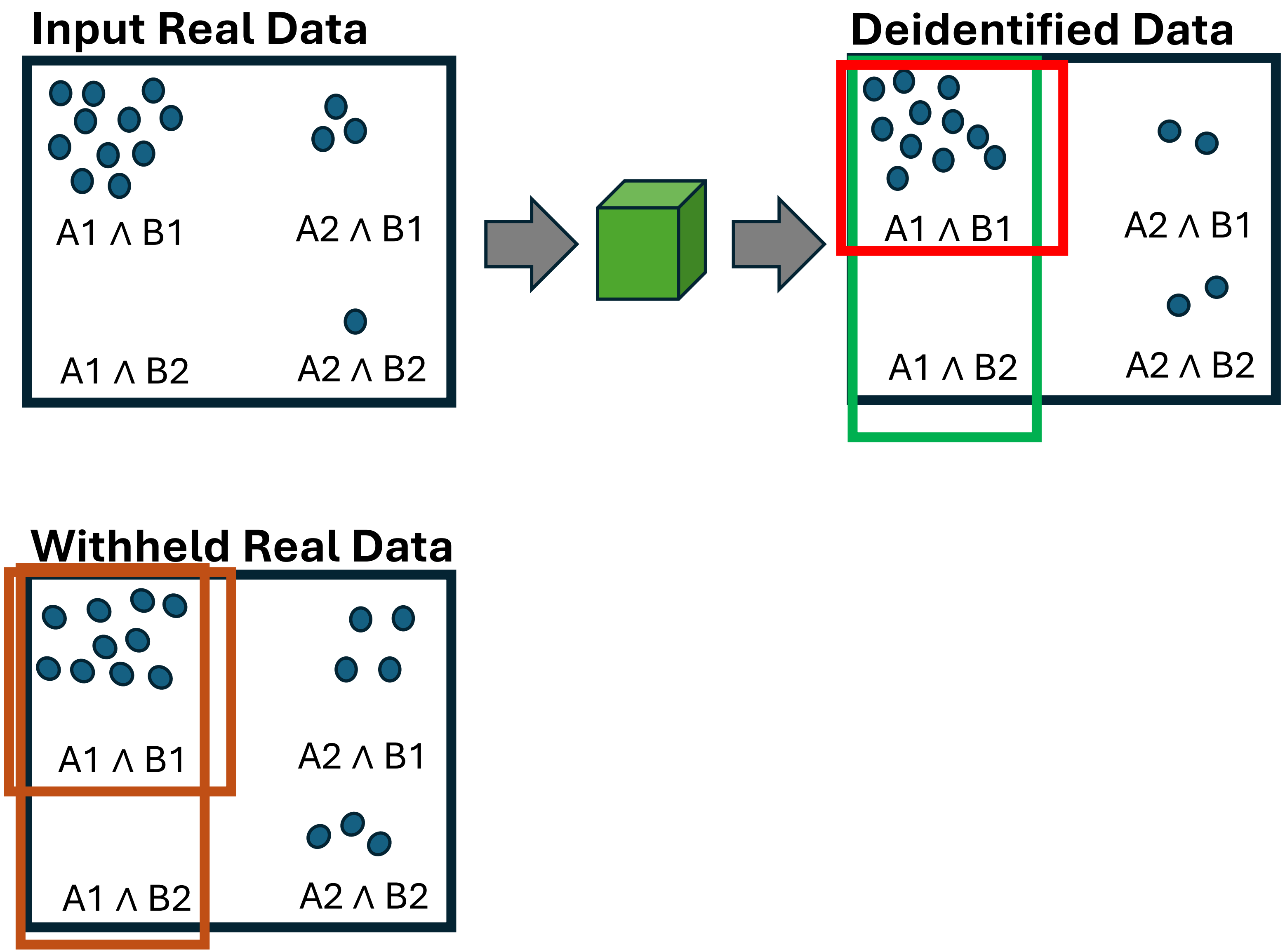
Can I use the released data to learn something invasive about you?

Attribute Privacy:

If I know someone's identifying features A, could I use the deidentified data to infer invasive features B

(1) better than I could using the withheld data? [Anonymeter: Inference]

(2) worse than I could using the input real data? [Synthpop: DiSCO]



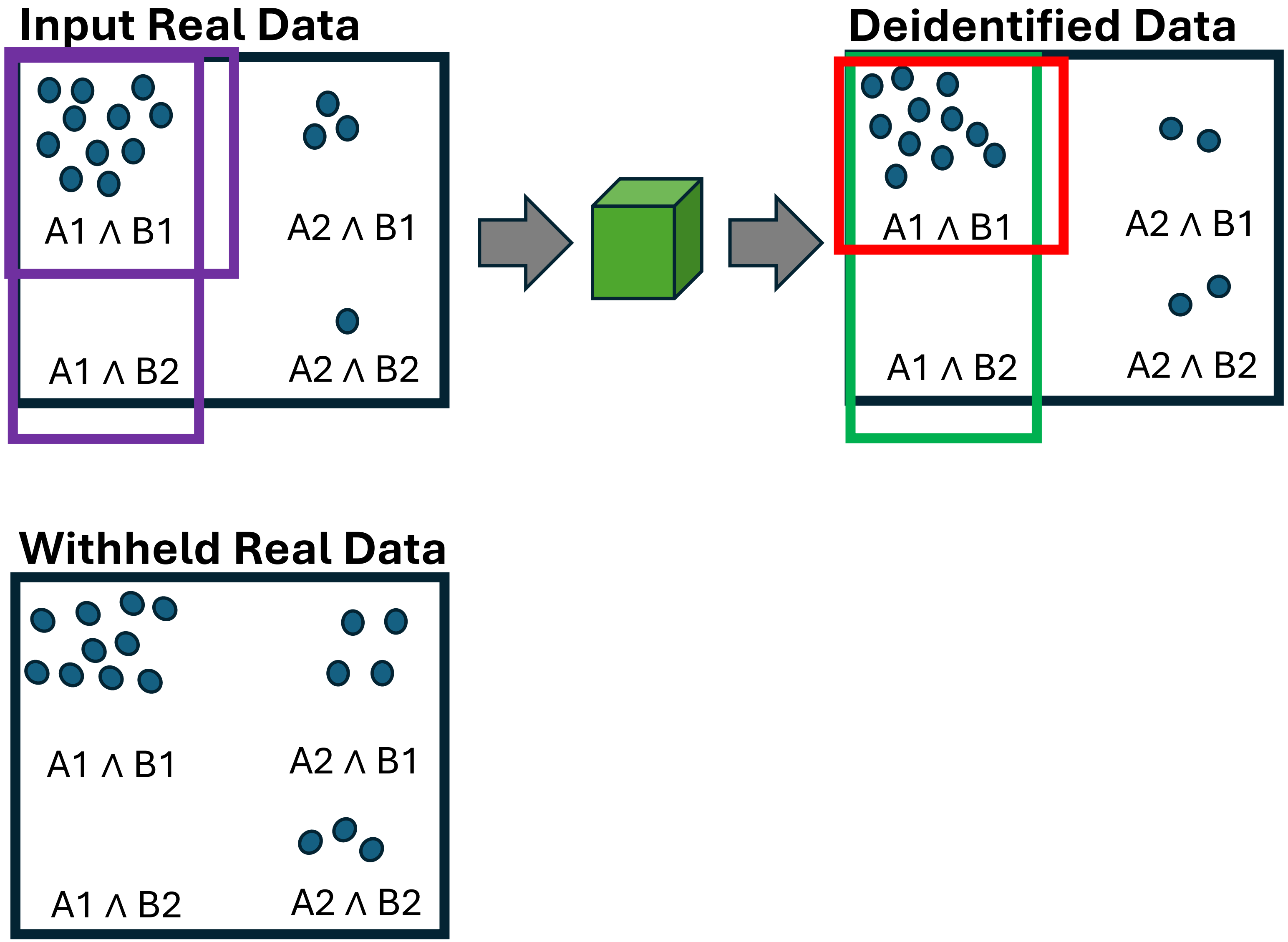
Can I use the released data to learn something invasive about you?

Attribute Privacy:

If I know someone's identifying features A, could I use the deidentified data to infer invasive features B

(1) better than I could using the withheld data? [Anonymeter: Inference]

(2) worse than I could using the input real data? [Synthpop: DiSCO]



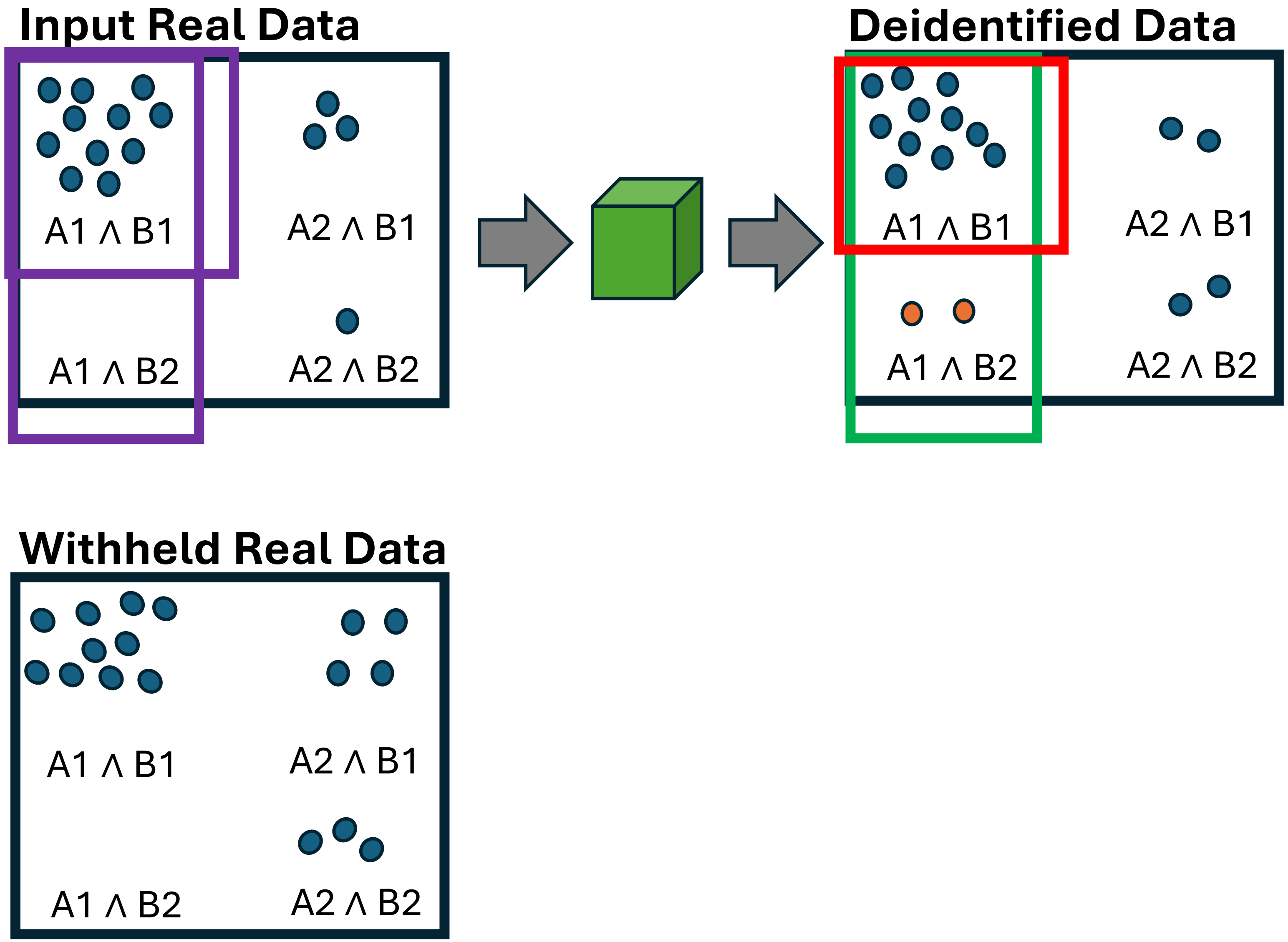
Can I use the released data to learn something invasive about you?

Attribute Privacy:

If I know someone's identifying features A, could I use the deidentified data to infer invasive features B

(1) better than I could using the withheld data? [Anonymeter: Inference]

(2) worse than I could using the input real data? [Synthpop: DiSCO]

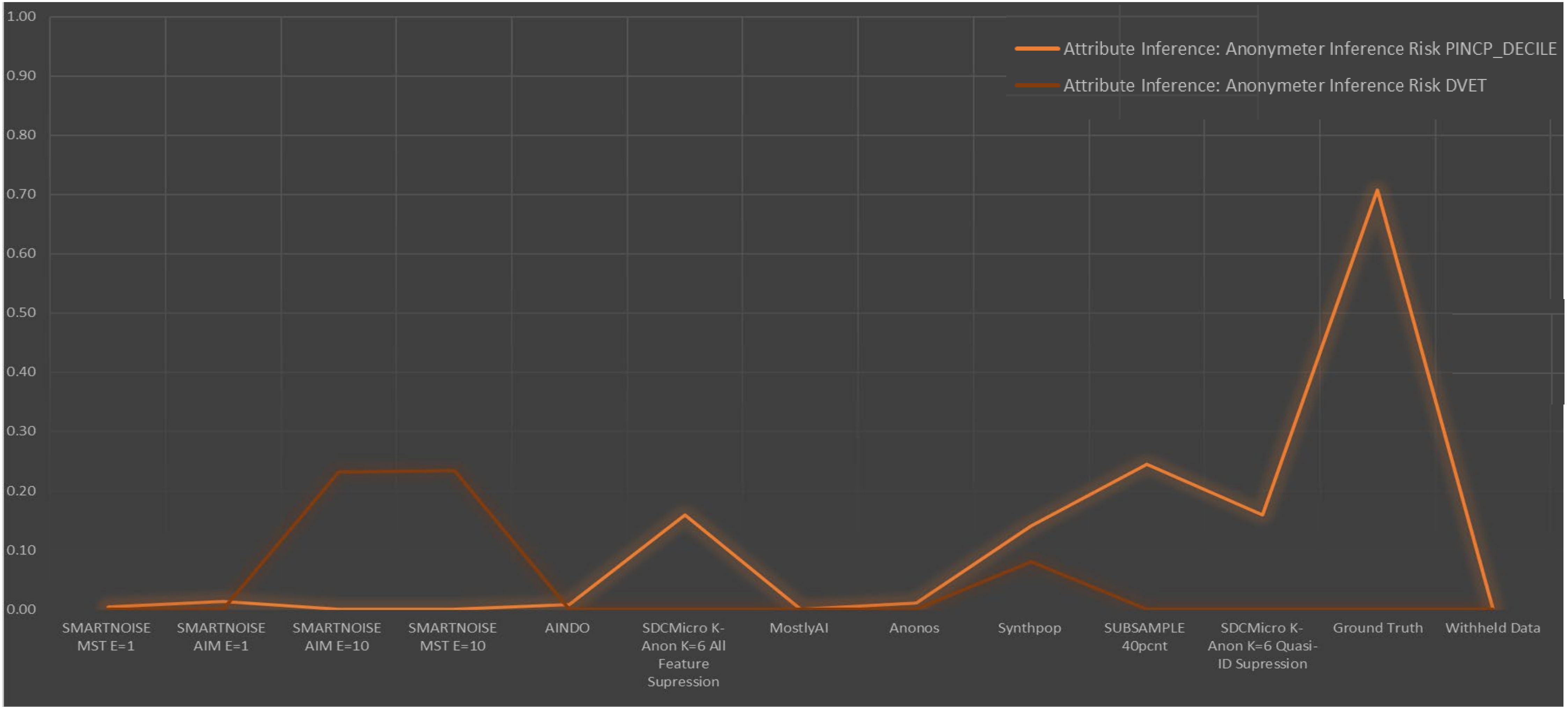


Can I use the released data to learn something invasive about you?

Attribute Privacy: If I know someone's identifying features A, could I use the deidentified data to infer invasive features B

(1) **better than I could using the withheld data? [Anonymeter: Inference]**

(2) better than I could using the input real data? [Synthpop: DiSCO]

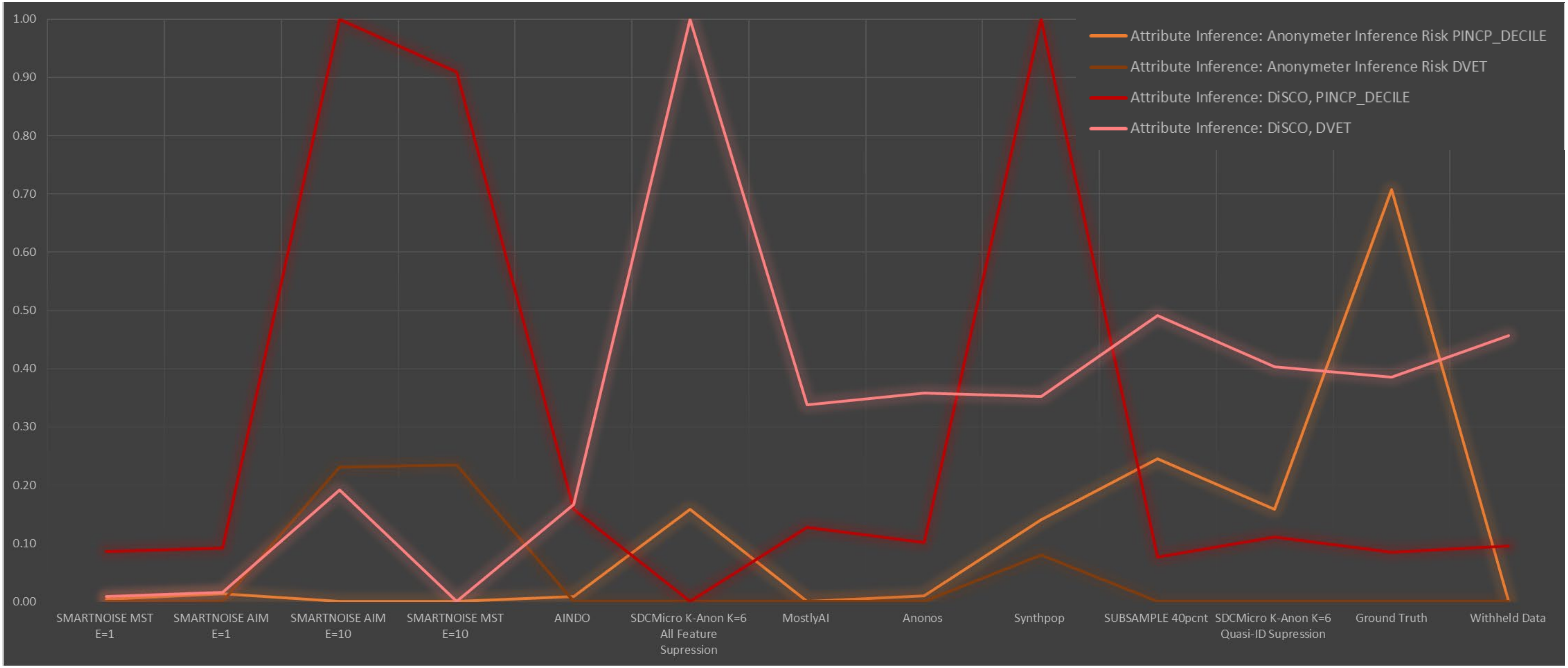


Can I use the released data to learn something invasive about you?

Attribute Privacy: If I know someone's identifying features A, could I use the deidentified data to infer invasive features B

(1) **better than I could using the withheld data? [Anonymeter: Inference]**

(2) **better than I could using the input real data? [Synthpop: DiSCO]**



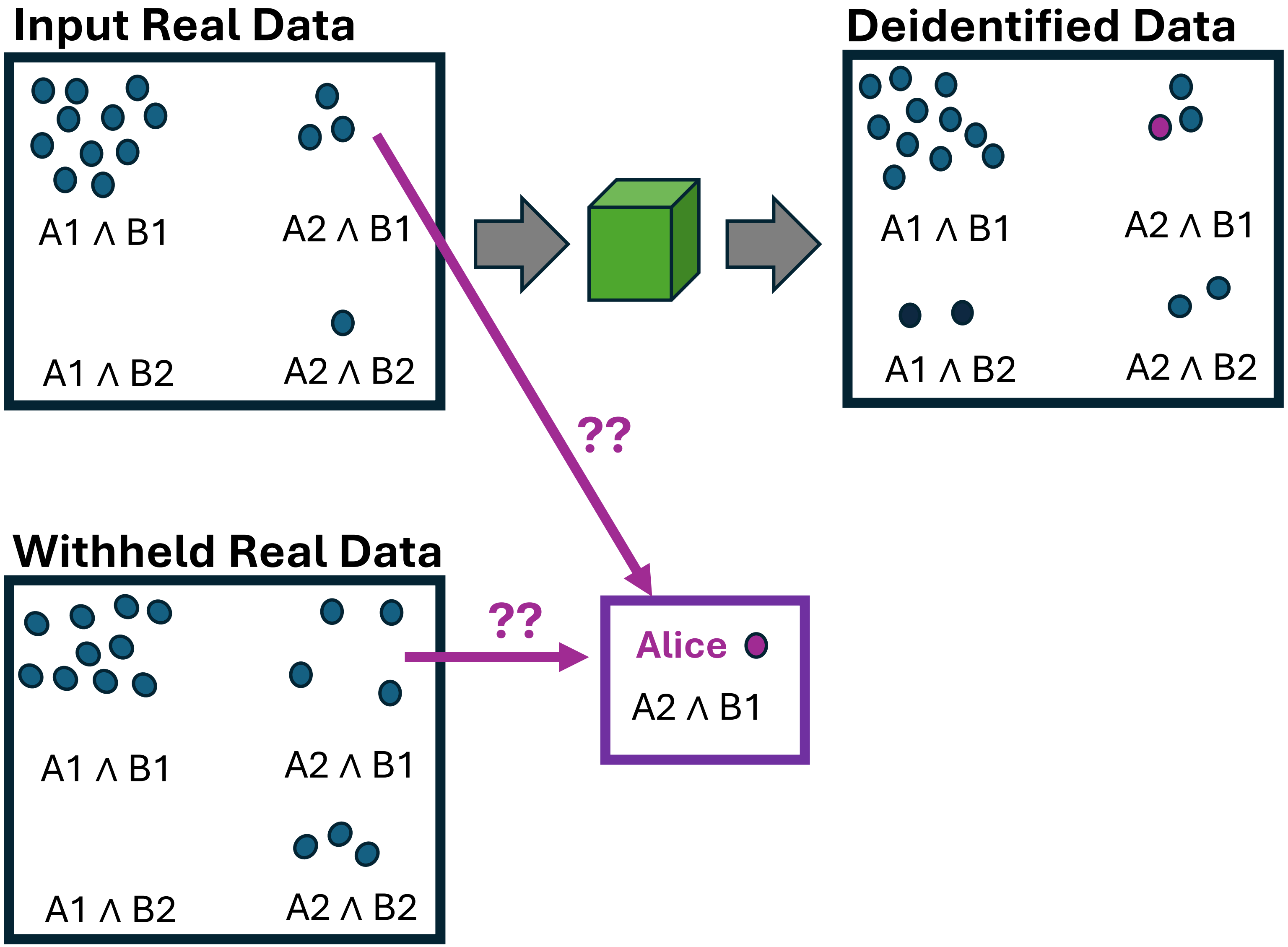
Can I use the released data to learn whether you were in the input data?

Membership Privacy:

If I know someone’s record, can I tell if they were in the (problematic) input data set?

(1) by checking if vulnerable outlier records look very similar to deidentified records? [Tumult: Empirical Leakage]

(2) by checking if deidentified data has dense spots where the withheld data doesn’t [SynthCity: DoMIAS Prior]



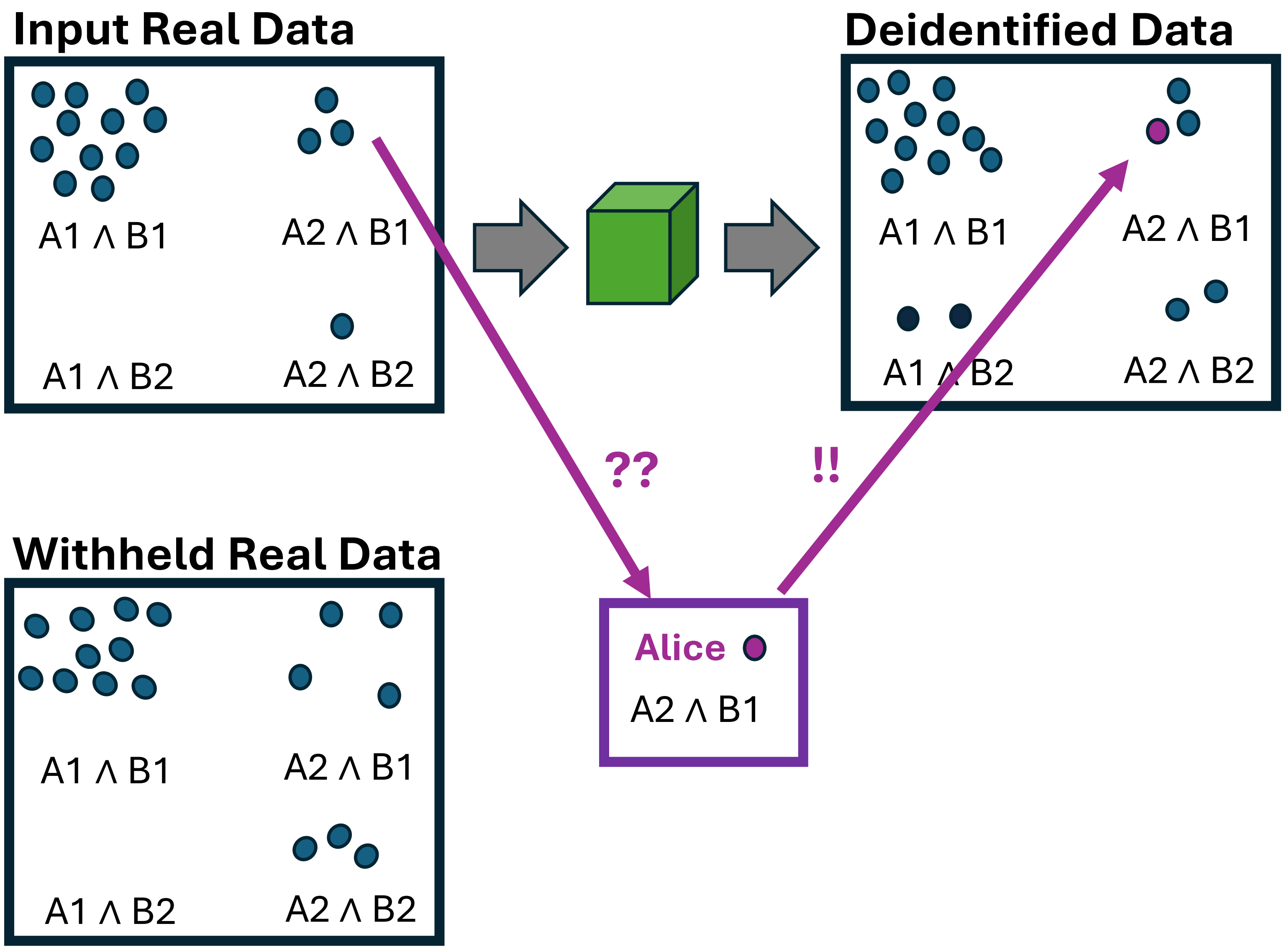
Can I use the released data to learn whether you were in the input data?

Membership Privacy:

If I know someone's record, can I tell if they were in the (problematic) input data set?

(1) by checking if vulnerable outlier records **look very similar** to deidentified records? [Tumult: Empirical Leakage]

(2) by checking if deidentified data **has dense spots** where the withheld data doesn't [SynthCity: DoMIAS Prior]

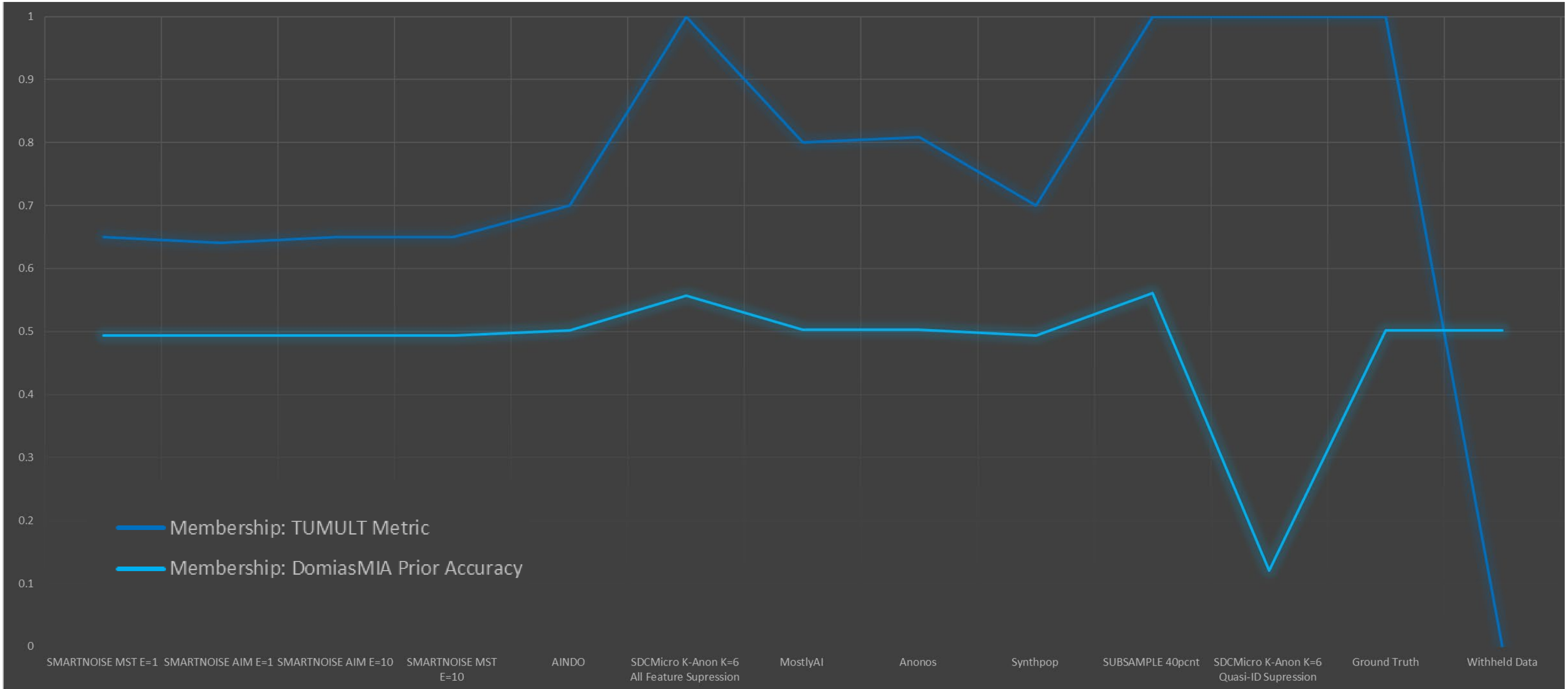


Can I use the released data to learn whether you were in the input data?

Membership Privacy: If I know someone's record, can I tell that they were in the input data set

(1) **by checking if vulnerable outlier records look similar to deidentified records? [Empirical Leakage]**

(2) **by checking if deidentified data has dense spots where the withheld data doesn't [DoMIAS Prior]**

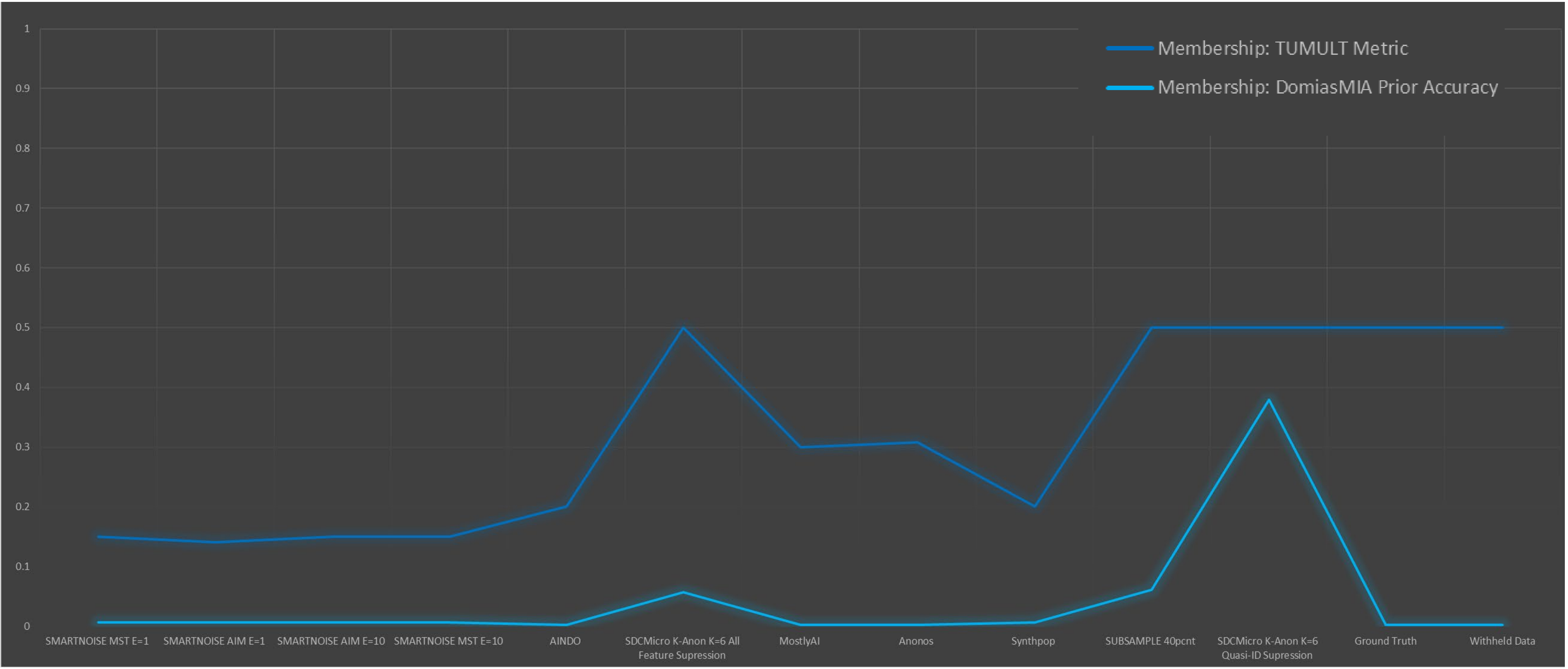


Can I use the released data to learn whether you were in the input data?

Membership Privacy: If I know someone's record, can I tell that they were in the input data set

(1) **by checking if vulnerable outlier records look similar to deidentified records? [Empirical Leakage]**

(2) **by checking if deidentified data has dense spots where the withheld data doesn't [DoMIAS Prior]**



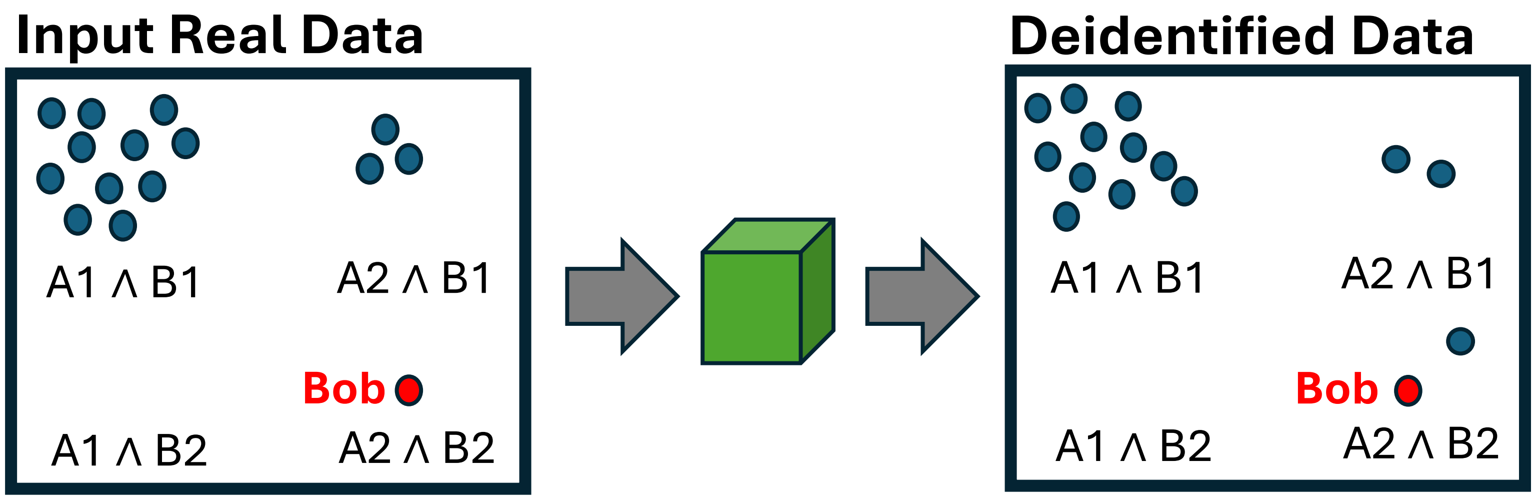
Can I use the released data to.... *find you*?

Reidentification Risk:

An anonymization concept: Can the deidentified data help me find your record because...

(1) your unique outlier record survived deidentification? [CRC: Unique Exact Match]

(2) the released deidentified data forms a bridge between two real data sets with features A and features B? [Anonymeter: Linkability]



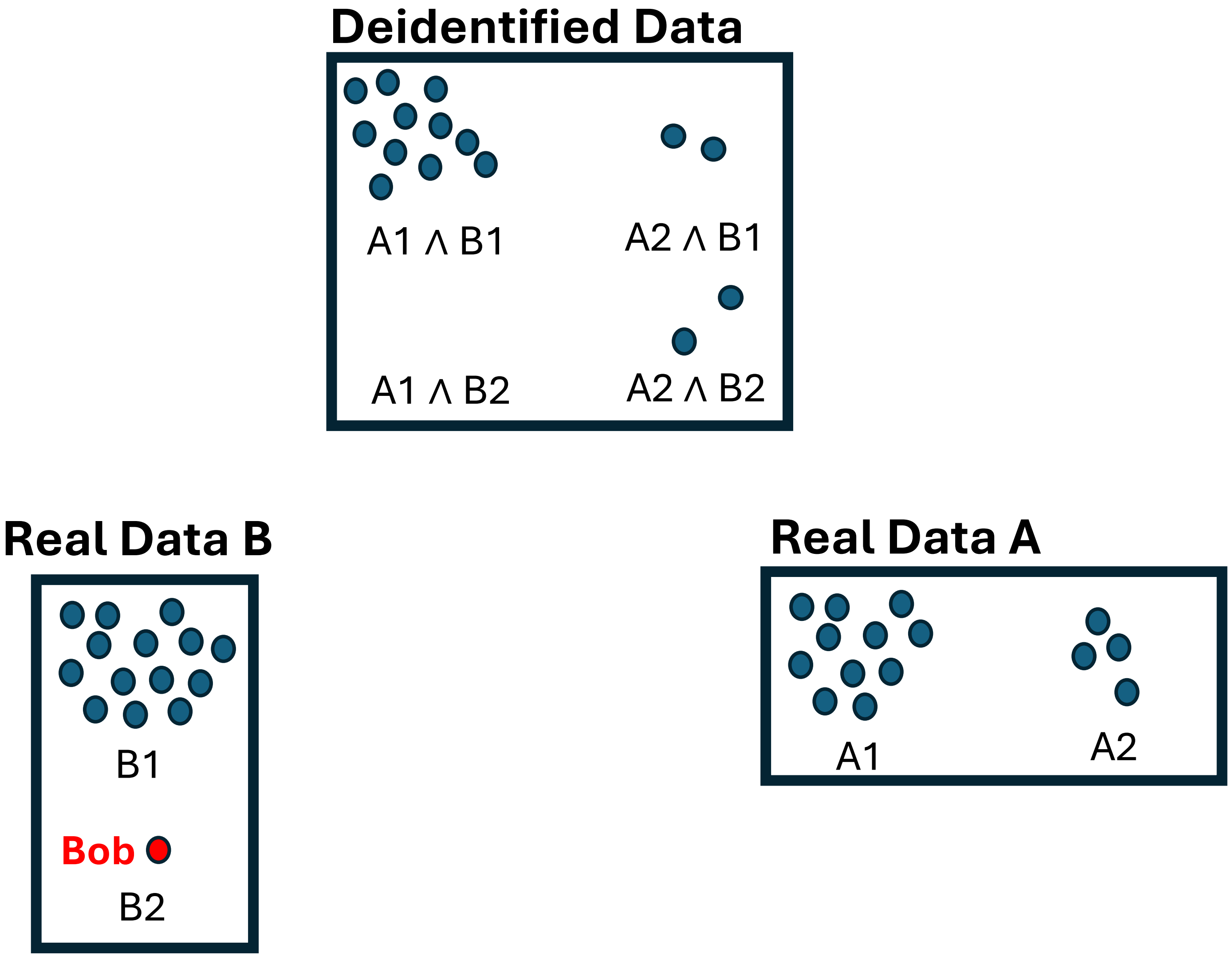
Can I use the released data to…. *find* you?

Reidentification Risk:

An anonymization concept: Can the deidentified data help me find your record because…

(1) your unique outlier record survived deidentification? [CRC: Unique Exact Match]

(2) the released deidentified data forms a bridge between two real data sets with features A and features B? [Anonymeter: Linkability]



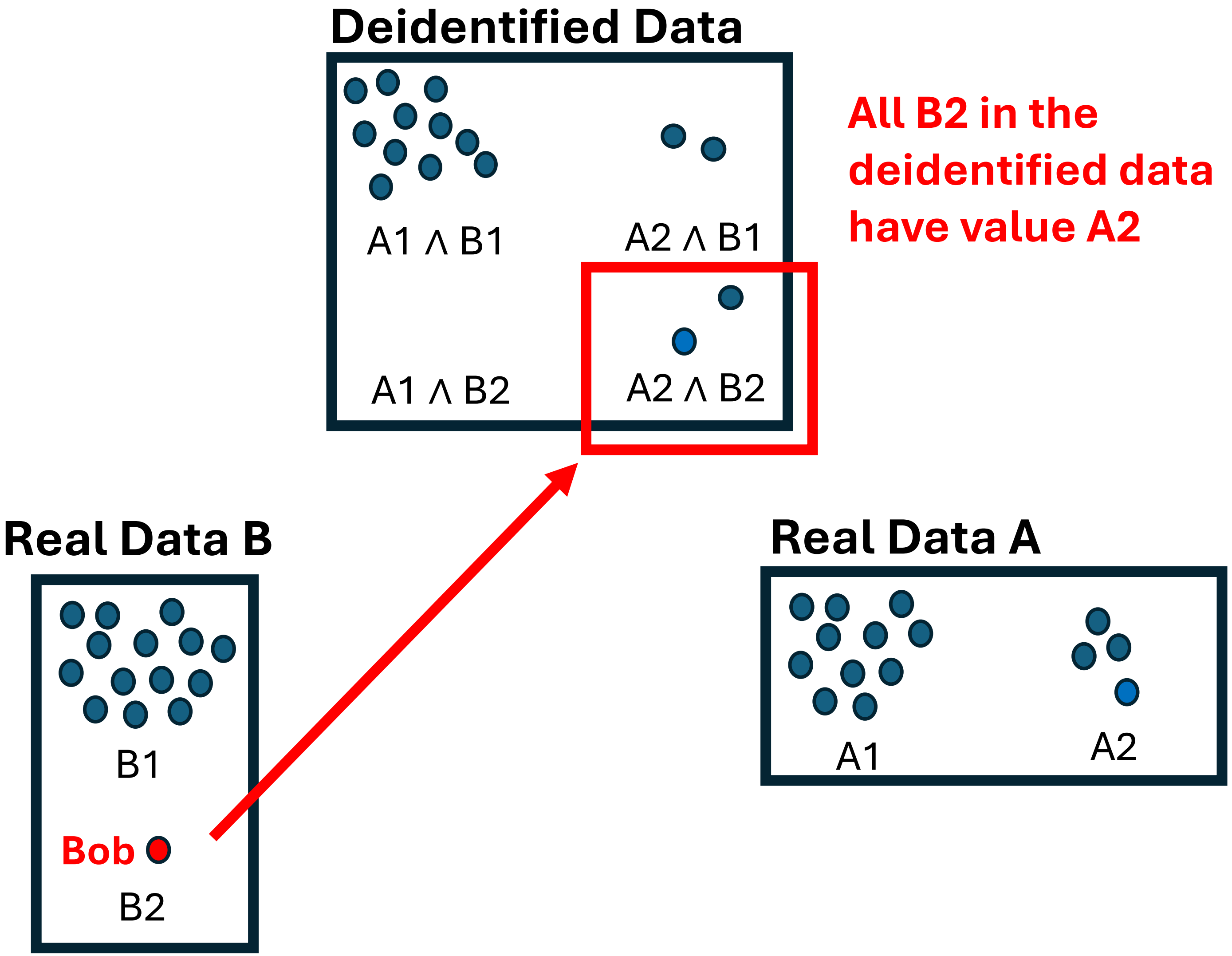
Can I use the released data to.... *find* you?

Reidentification Risk:

An anonymization concept: Can the deidentified data help me find your record because...

(1) your unique outlier record survived deidentification? [CRC: Unique Exact Match]

(2) the released deidentified data forms a bridge between two real data sets with features A and features B? [Anonymeter: Linkability]



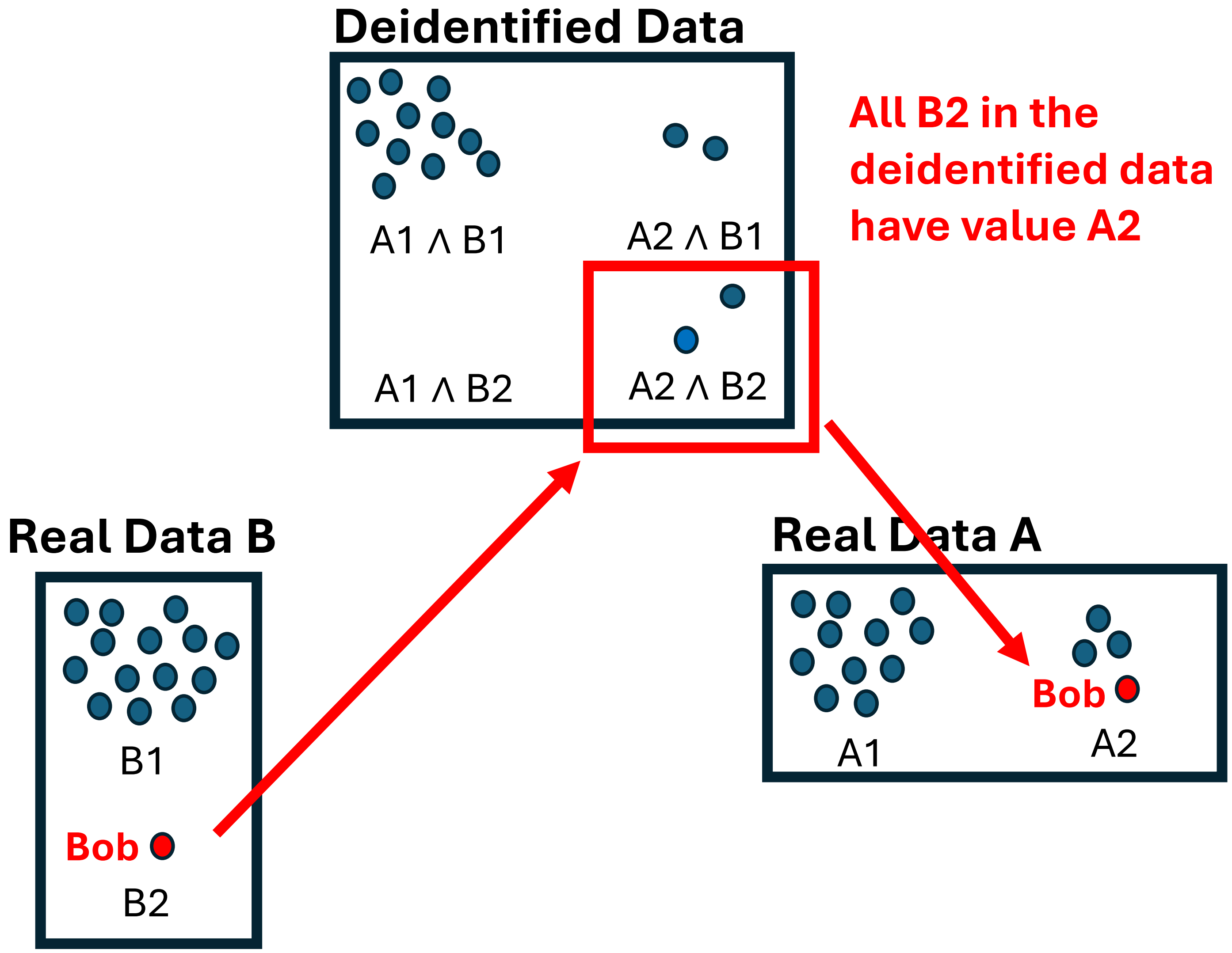
Can I use the released data to.... *find you*?

Reidentification Risk:

An anonymization concept:
Can the deidentified data help me find your record because...

(1) your unique outlier record survived deidentification?
[CRC: Unique Exact Match]

(2) the released deidentified data forms a bridge between two real data sets with features A and features B?
[Anonymeter: Linkability]

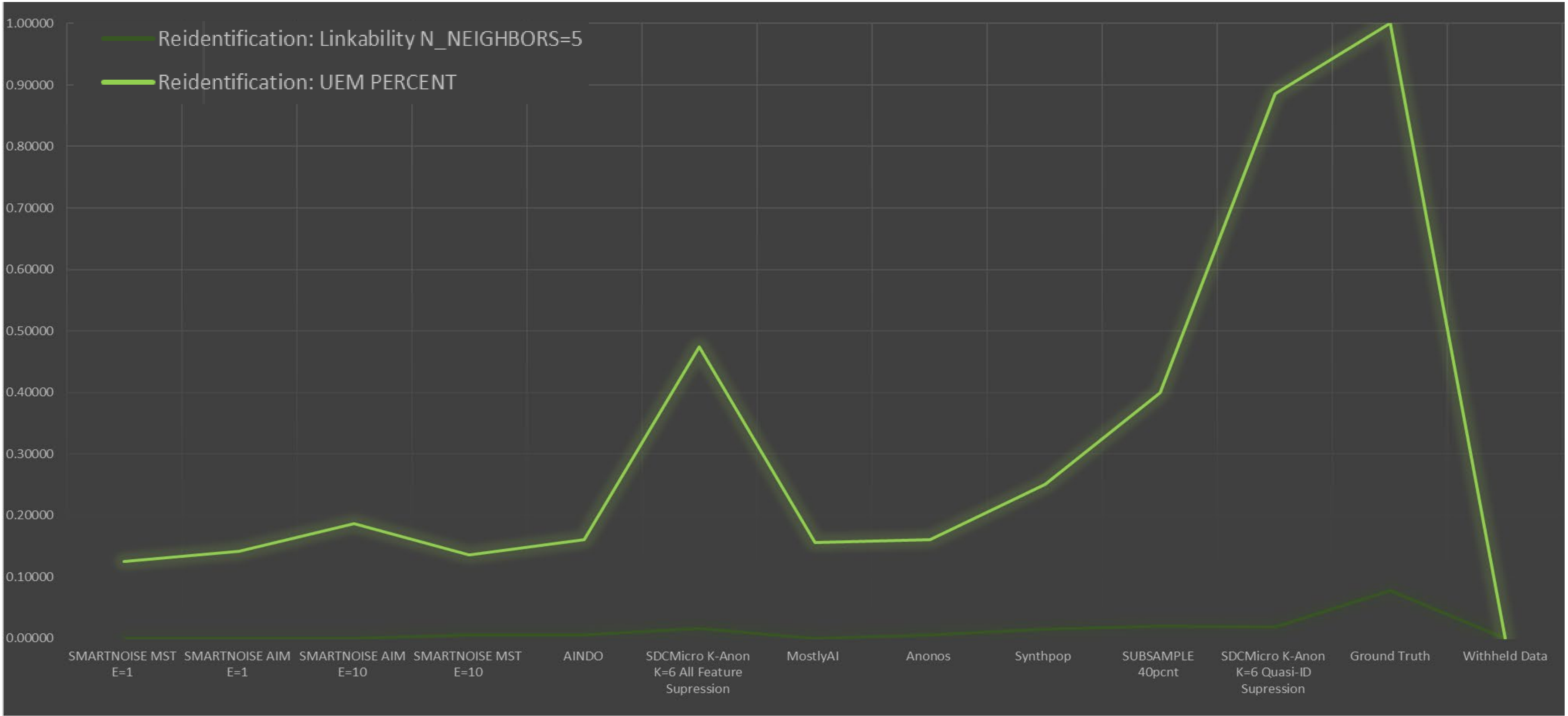


Can I use the released data to.... *find you*?

Reidentification Risk: Can the deidentified data help me find your record because...

(1) **your unique outlier record survived deidentification (as % of unique records)? [Unique Exact Match]**

(2) **the released data is a bridge between two real data sets with features A and features B? [Linkability]**



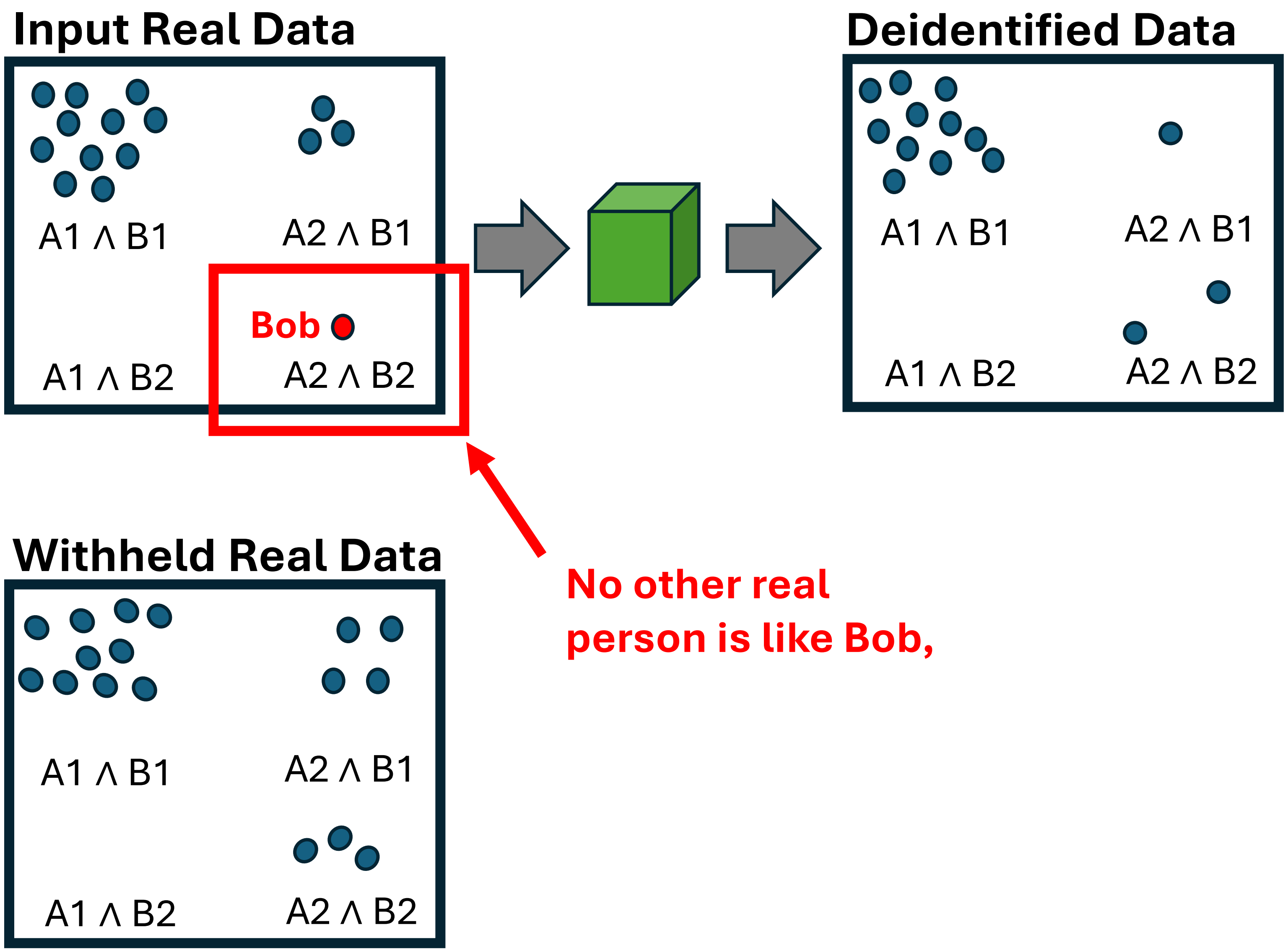
Can I use the released data to imagine that I've found you?

Singling Out:

Are there records in the deidentified data that look like they're distinct (unusual) real people? This could happen if the deidentified record is:

(1) **Closer to a real outlier person**, say Bob, than any other real person is close to Bob. [Synthcity: Identifiability]

(2) An outlier (unique feature combo) and there's a matching outlier in the real data. [Anonymeter: Singling Out]



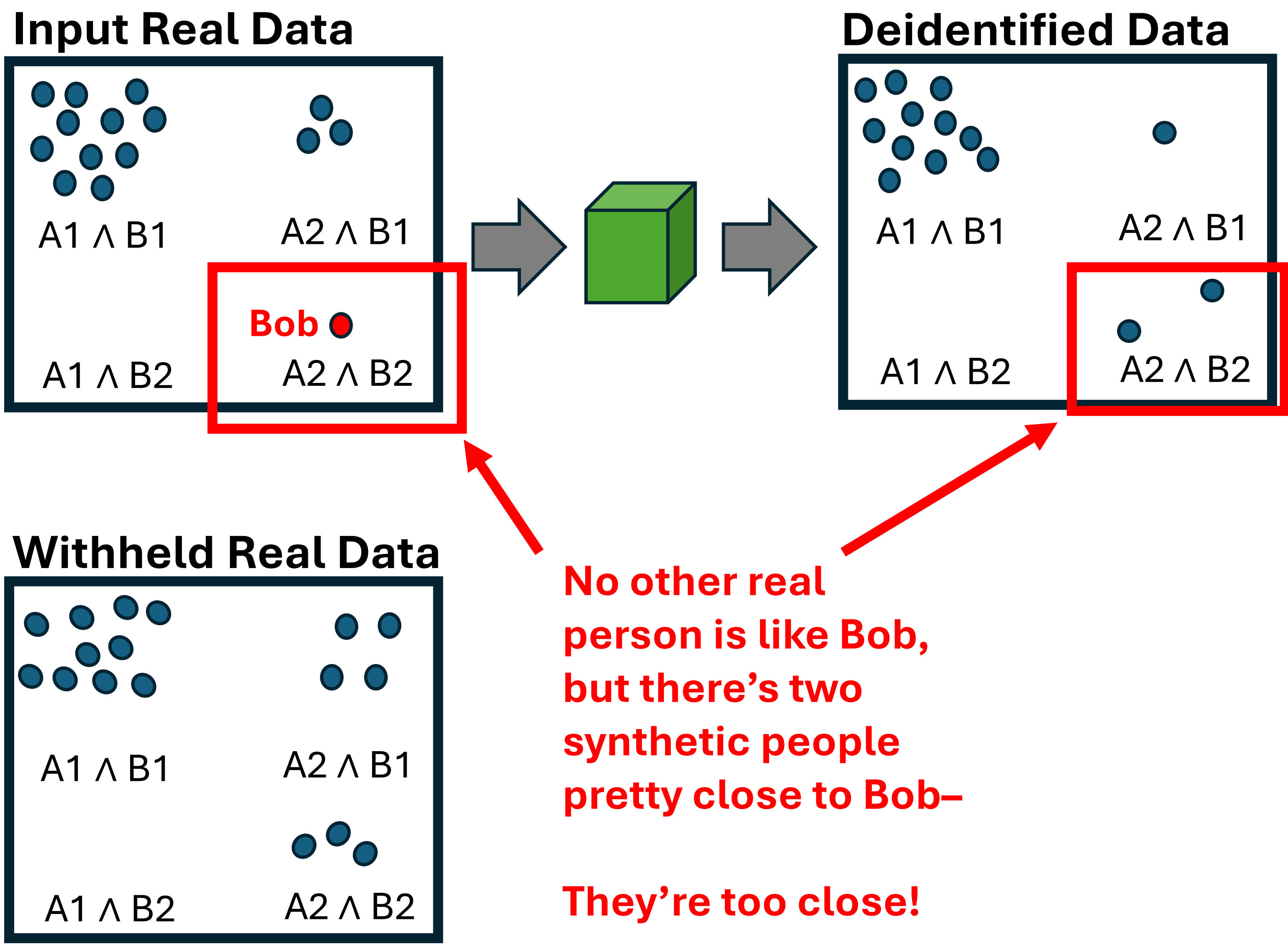
Can I use the released data to imagine that I've found you?

Singling Out:

Are there records in the deidentified data that look like they're distinct (unusual) real people? This could happen if the deidentified record is:

(1) **Closer to a real outlier person**, say Bob, than any other real person is close to Bob. [Synthcity: Identifiability]

(2) An outlier (unique feature combo) and there's a matching outlier in the real data. [Anonymeter: Singling Out]



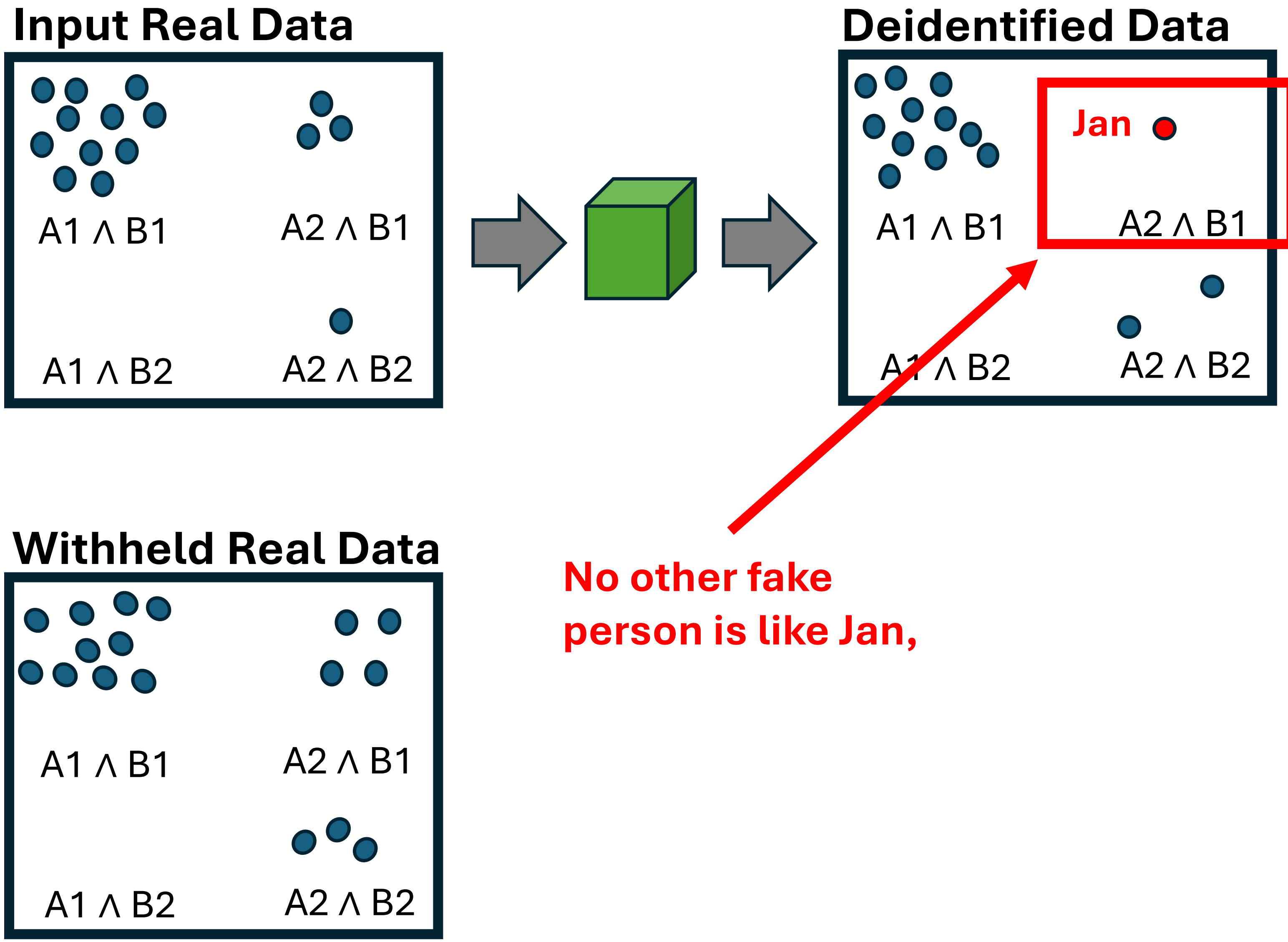
Can I use the released data to imagine that I've found you?

Singling Out:

Are there records in the deidentified data that look like they're distinct (unusual) real people? This could happen if the deidentified record is:

(1) Closer to a real outlier person, say Bob, than any other real person is close to Bob. [Synthcity: Identifiability]

(2) A **synthetic outlier (unique feature combo)** and there's also a matching outlier in the real data. [Anonymeter: Singling Out]



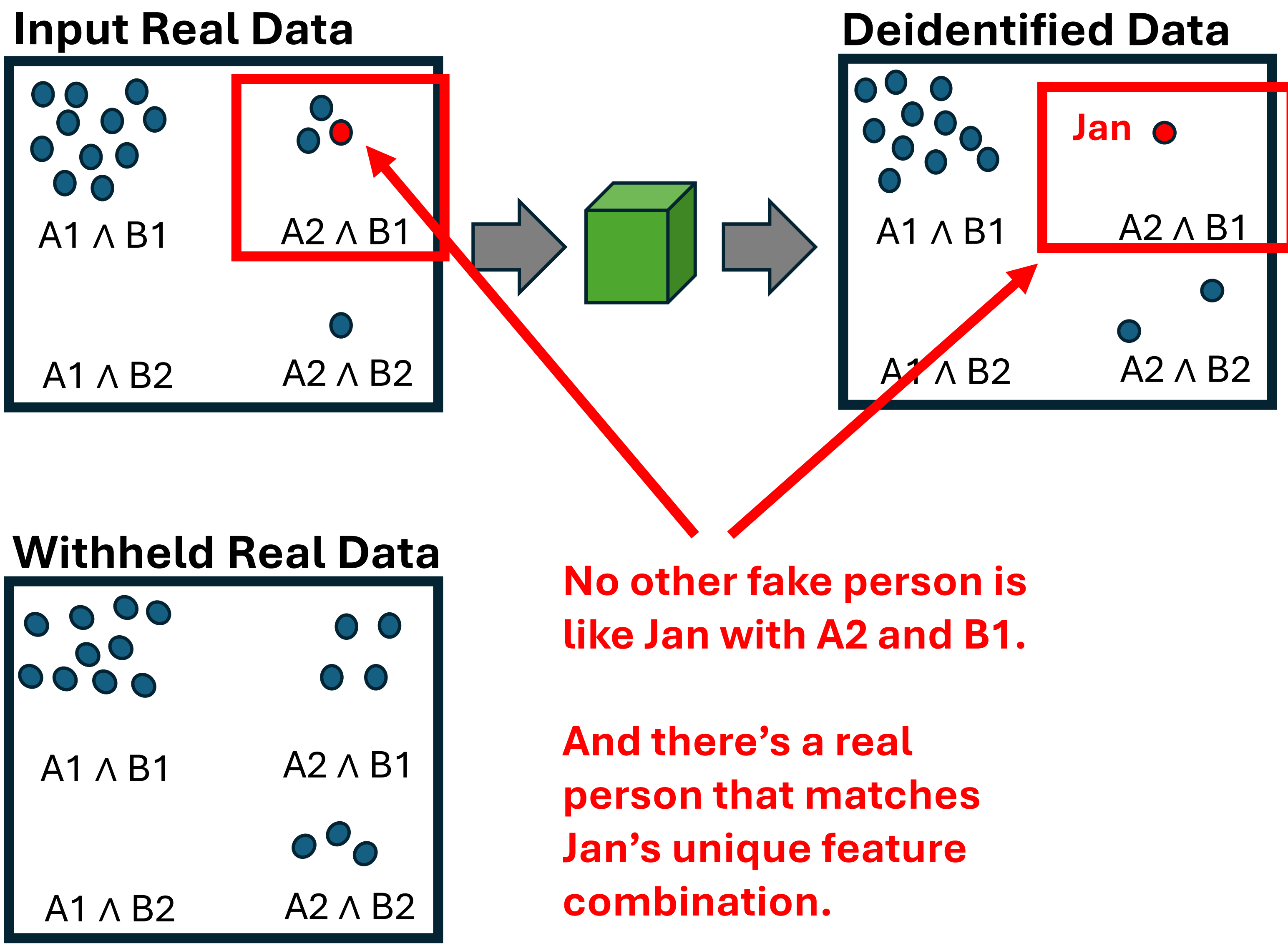
Can I use the released data to imagine that I've found you?

Singling Out:

Are there records in the deidentified data that look like they're distinct (unusual) real people? This could happen if the deidentified record is:

(1) Closer to a real outlier person, say Bob, than any other real person is close to Bob. [Synthcity: Identifiability]

(2) A **synthetic outlier (unique feature combo)** and there's also a matching outlier in the real data. [Anonymeter: Singling Out]

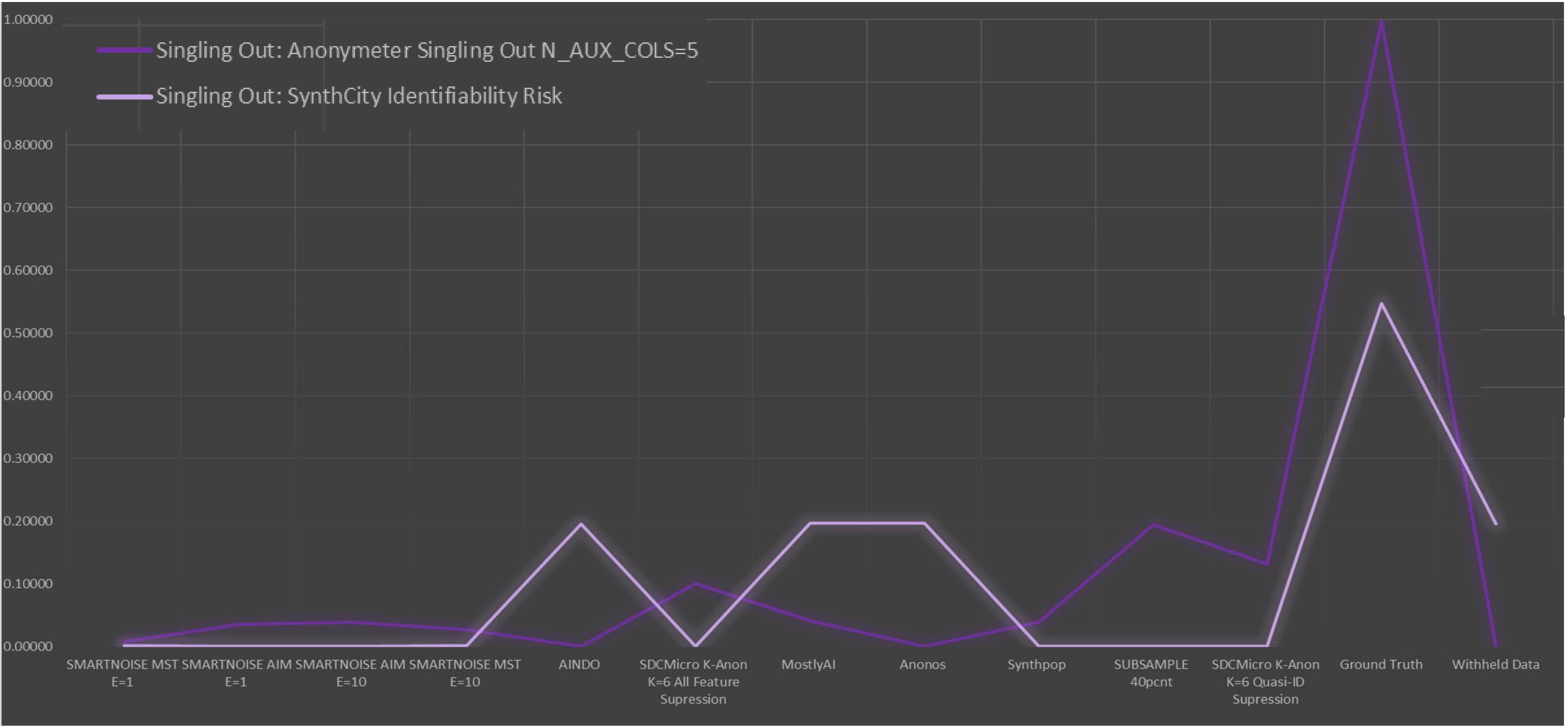


Can I use the released data to imagine that I’ve found you?

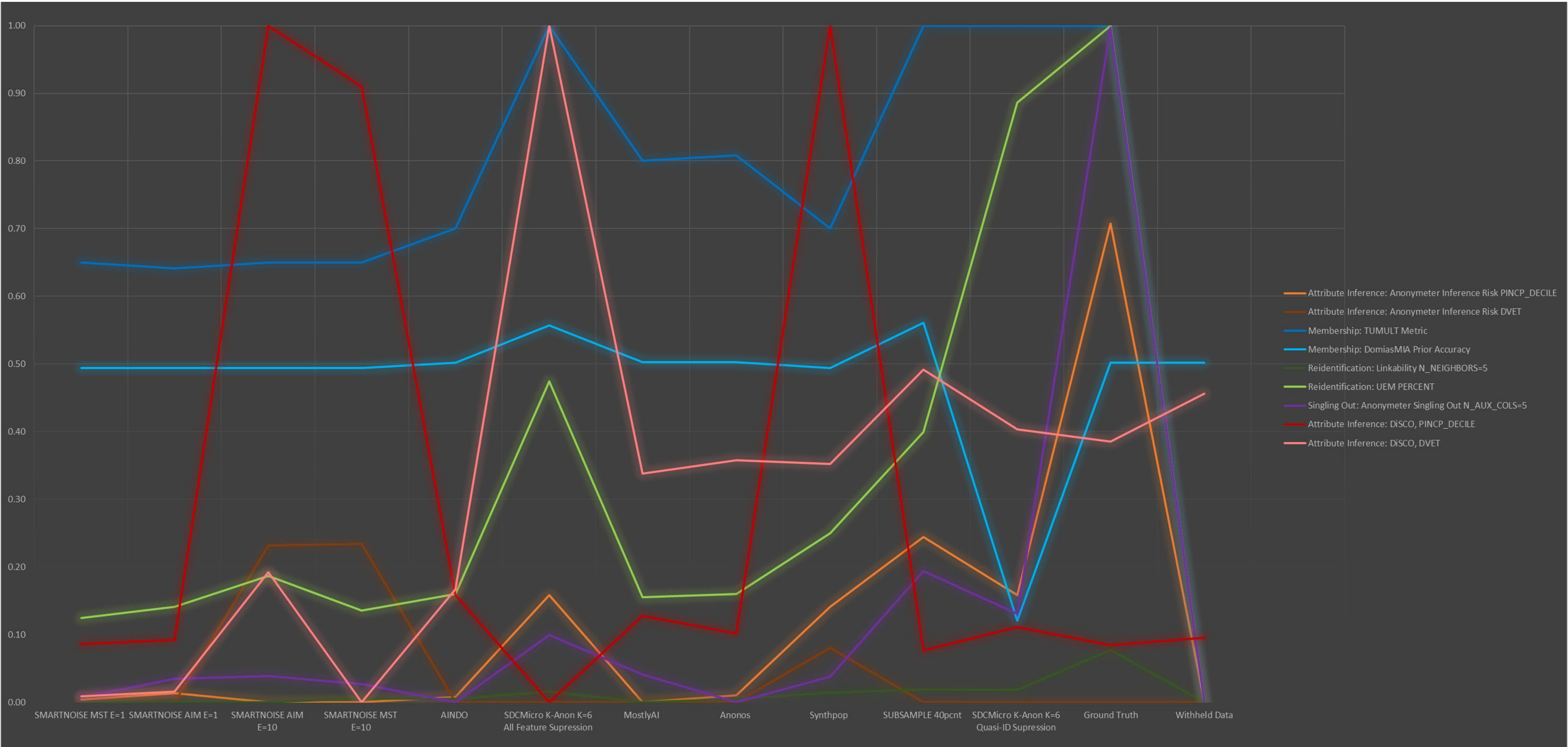
Singling Out: Are there records in the deidentified data that look like they might be distinct real people?

(1) **Fake records closer to a real outlier person than any other real person is. [Synthcity: Identifiability]**

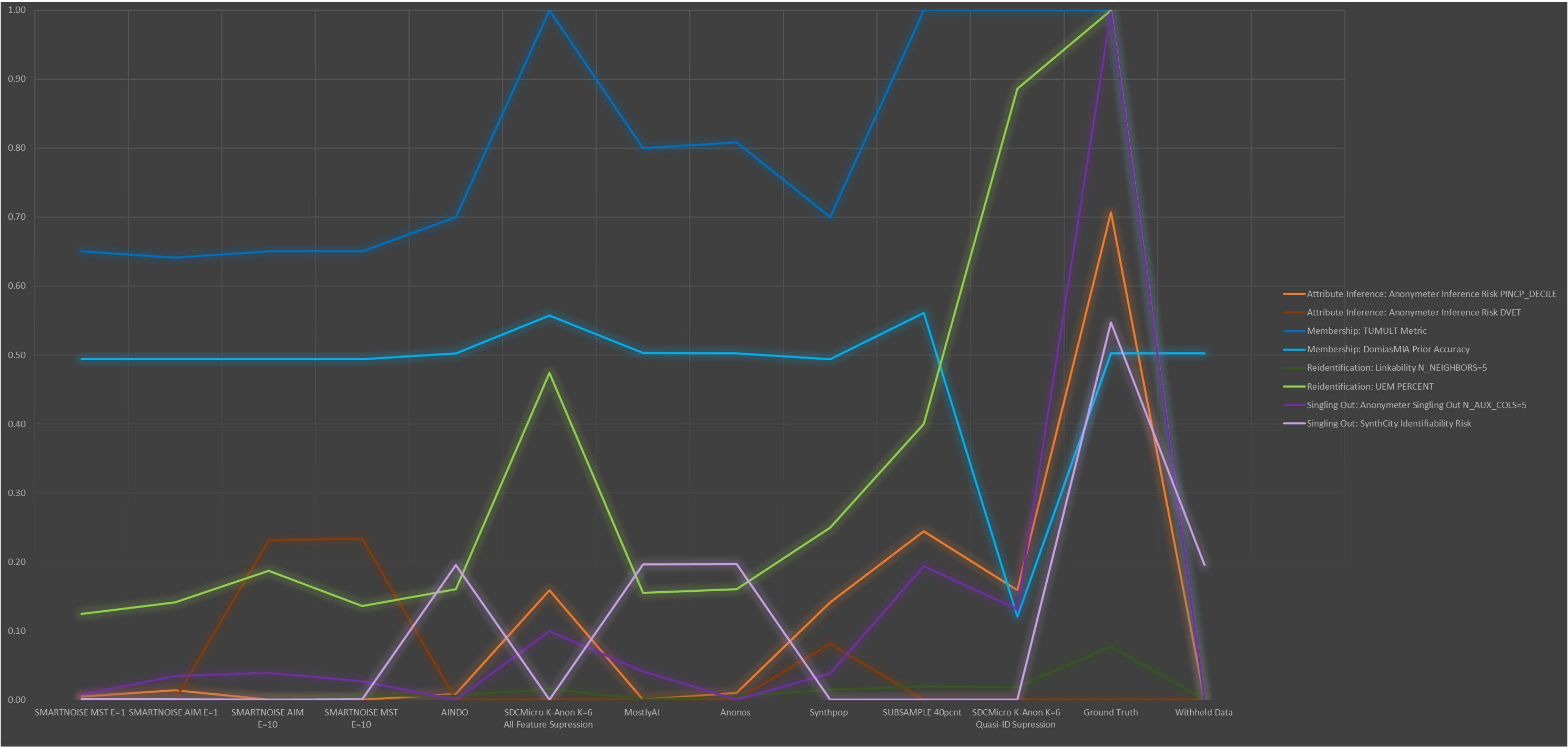
(2) **Fake outlier records (unique feature combo) with matching real records [Anonymeter: Singling Out]**



So what's private? Is there consensus?



Is there consensus? ...what if we only look at author-tuned metrics?

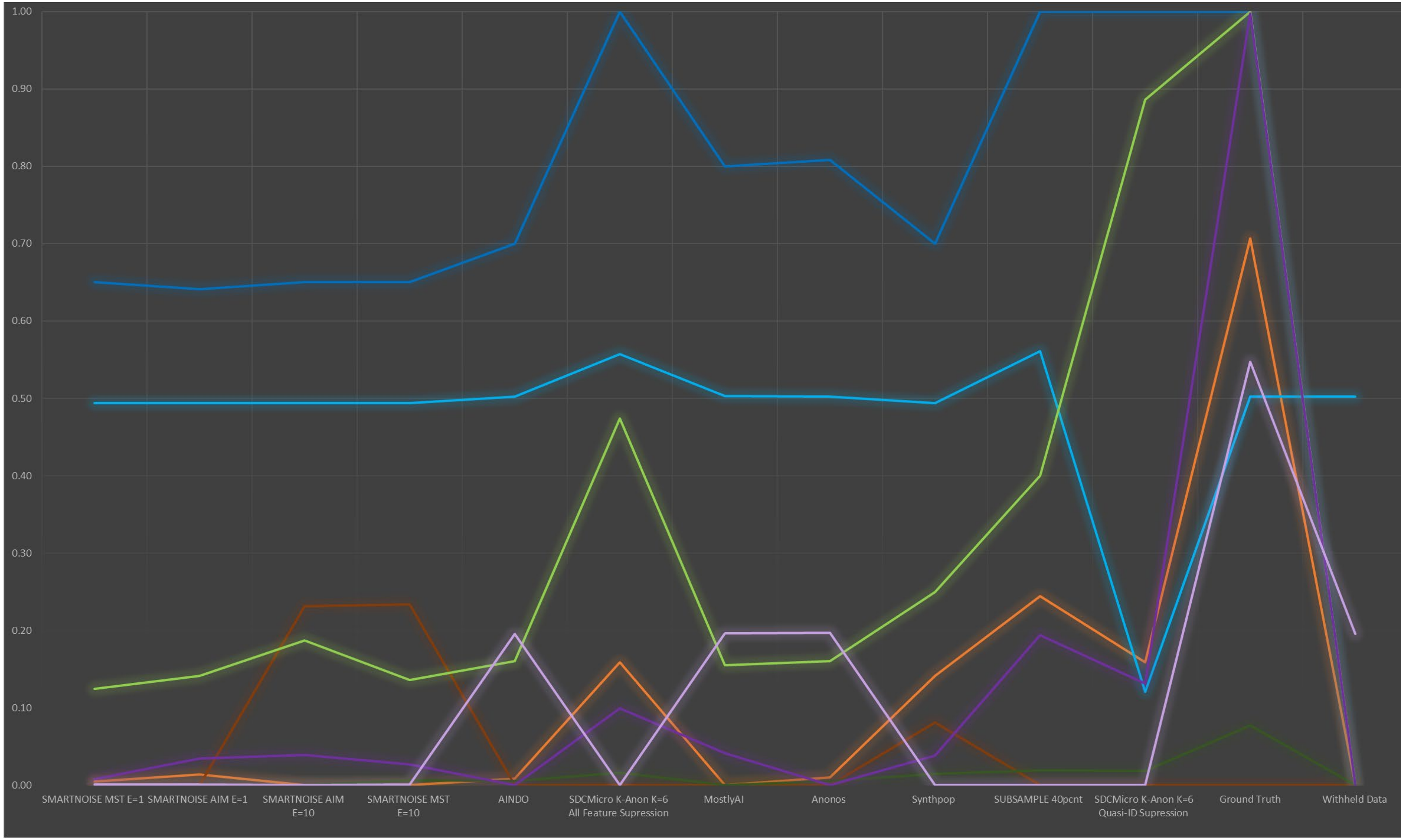


Tuning and configuring privacy metrics (open questions):

Measuring Privacy for Deidentified Data:

Possibly we need to check how we're configuring the metrics:

- Similarity between records with mixed numerical and categorical features
- Quasi-identifiers
- KNN pool size (k)
- etc...



Collaborative Research Cycle: *Red-Teaming* Deidentification Algorithms

Measuring Privacy for Deidentified Data:

Possibly we need to check how we're configuring the metrics.

...or maybe we should just see if any of you can ***hack it?***

Collaborative Research Cycle: *Red-Teaming* Deidentification Algorithms

Measuring Privacy for Deidentified Data:

Possibly we need to check how we're configuring the metrics.

...or maybe we should just see if any of you can ***hack it?***

Next up:

We will be running a community Red Team Exercise in Spring of 2025

***B*etween now and then, we'll be hosting Privacy Metric Discussion Hours, where you can learn the details of the metrics we just flew through, and meet metric designers.**
Sign up on our site to keep in touch!

Thank you!

Questions?

gary.howarth@nist.gov

christine.task@knexusresearch.com