Synthetic Data

IOM/Microsoft collaboration

Prepared by the Counter-Trafficking Data Collaborative (CTDC)

2023-03-06

IOM UN MIGRATION

CTDC

What are synthetic data?

- Synthetic data are **artificial data** generated from real data
- It cannot be linked back to the original data (including cases), so we can ensure data confidentiality and privacy.
- The statistical properties and relationships from the original data are preserved in synthetic data.
- The advantages of synthetic data over other forms of data are, e.g.,
  - **Aggregate data**: Researchers can no longer make connections between traits
    - Example: The Global Estimates of Modern Slavery
  - **K-anonymized data**: The redaction of outliers can lead to significant data loss
    - Example: CTDC data pre-2021

Example: Raw Data vs. K-anonymized Data

Raw Data

Gender	Age	isForcedLabour
Female	19	1
Male	18	1
Male	20	1
Male	37	
Female	35	
Female	31	

Anonymization

K-anonymized Data (k = 2)

Gender	AgeBroad	isForcedLabour	*k*
Female	18–20	1	1
Male	18–20	1	2
Male	18–20	1	2
Male	30–38		1
Female	30–38		2
Female	30–38		2

Example: Synthetic Data vs. K-anonymized Data

Synthetic Data ?

Gender	AgeBroad	isForcedLabour
Female		
	18–20	1
Male		
	30-38	
		1
Male	18–20	1
Male	18–20	1
Female	30–38	
Female	30–38	

K-anonymized Data (*k* = 2)

Gender	AgeBroad	isForcedLabour	*k*
~~Female~~	~~18–20~~	~~1~~	~~1~~
Male	18–20	1	2
Male	18–20	1	2
~~Male~~	~~30–38~~		~~1~~
Female	30–38		2
Female	30–38		2

**Both datasets are generated from the same raw data. For CTDC, k = 10.*

Overcoming challenges with synthetic data

Problem – Then

- CTDC's previous solution was **labour-intensive and partner reliant**
- K-anonymization results in the **loss** of potentially useful data
- Risk of **re-identification and reprisals**
- Market providers were **expensive**

Solution – Now

- Microsoft Research's solution **automatically** prevents the publication of rare attributes
- **Preserves** the statistical properties and relationships in the original data
- Differential privacy **guarantees against any privacy attacks**
- **Open-source**

Synthetic Data
Timeline and achievements

2017
CTDC launch

2019
TAT Accelerator program

2020
Published a paper with Microsoft Research on the privacy-preserving data solution

2021
Released the Global Synthetic Dataset

2022
Released the Global Victim-Perpetrator Synthetic Dataset

IOM UN MIGRATION

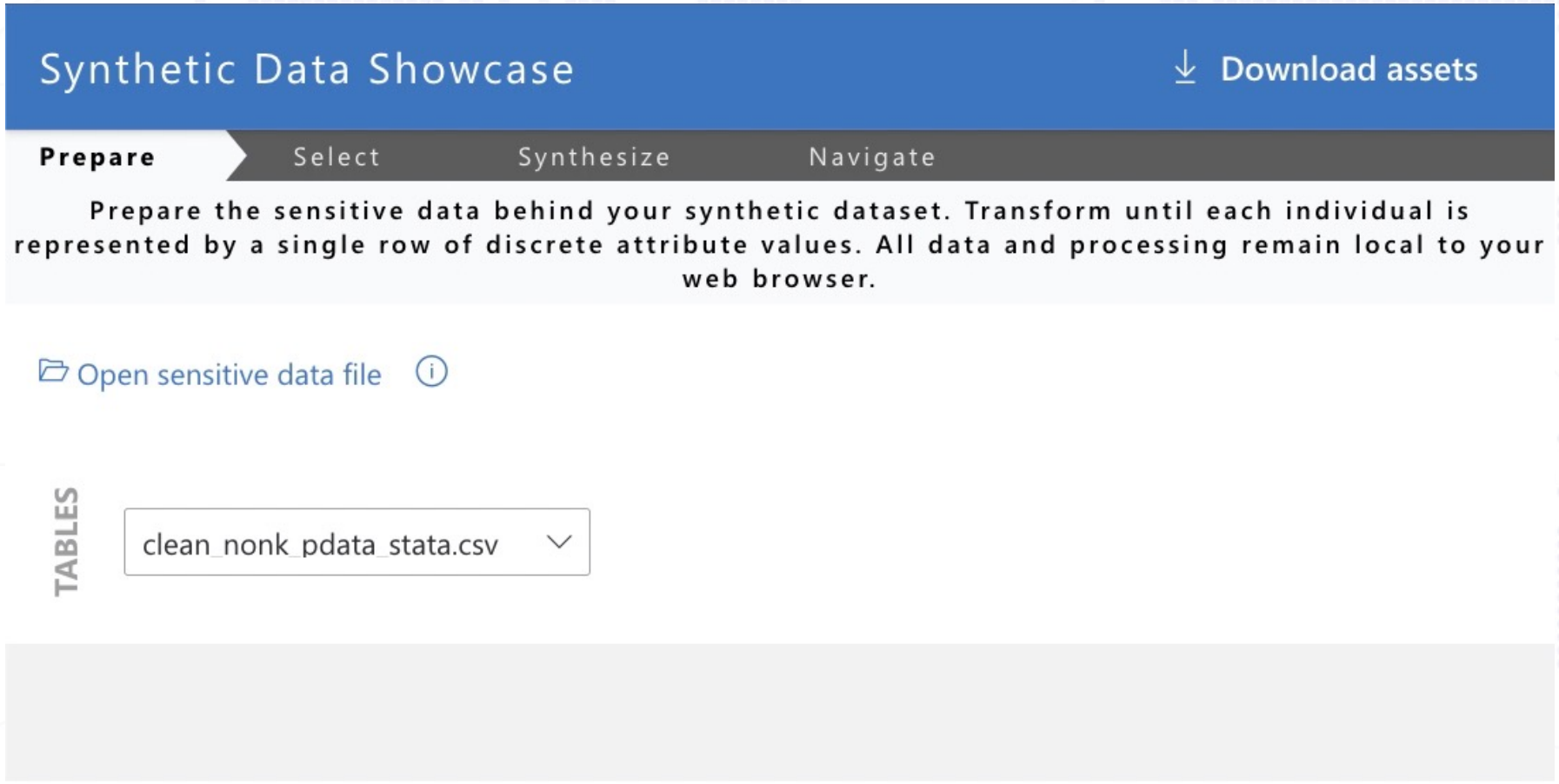
Synthetic Data
Tech Against Trafficking Accelerator Program

- Tech Against Trafficking invited IOM's CTDC to participate the 2019 Accelerator Program.
- The partnership focused on 3 workstreams:
  1. **Privacy-preserving mechanism**:
     Develop a solution for analyzing case data while protecting victim privacy.
  2. **Data standards:**
     Address data standards/consistency related to victim case management.
  3. **Stakeholder engagement:**
     Maximize utility and impact of the CTDC platform.
- IOM has benefitted from substantial in-kind support from Microsoft Research to support CTDC.

Microsoft Research's Synthetic Data Showcase

- Synthetic Data Showcase can automatically generate 3 elements: i) a synthetic data, ii) aggregate data, and iii) data dashboards.
- The tool provides 2 approaches to create anonymous datasets: i) differential privacy and ii) k-anonymity.
- The tools are available with **command-line** options in Python/Rust or a locally run **web application** using Javascript and Web Assembly.

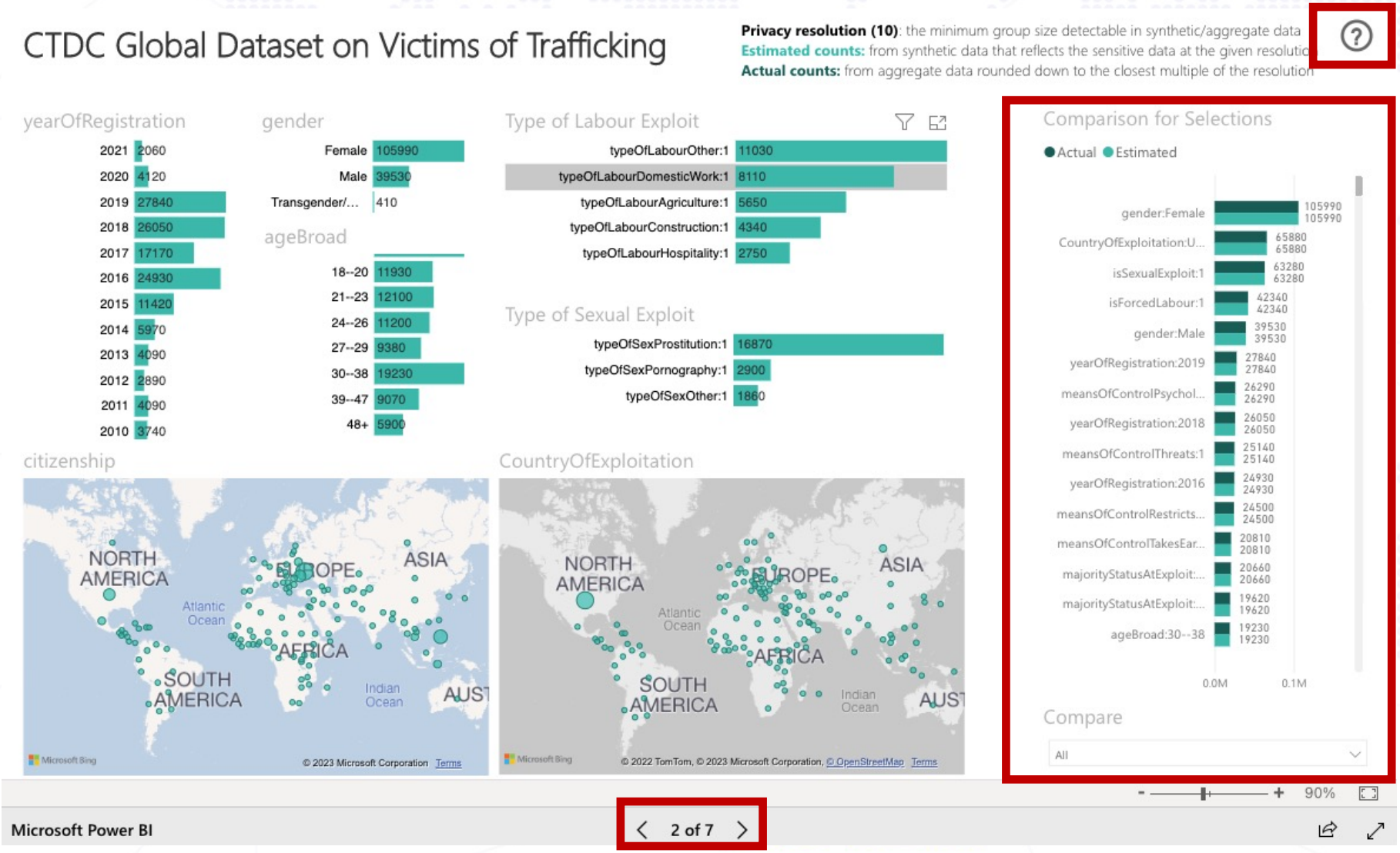
Synthetic Data Showcase
Free-to-use web application



IOM UN MIGRATION

Global Synthetic Dataset

You can download data, codebook, and data dictionary

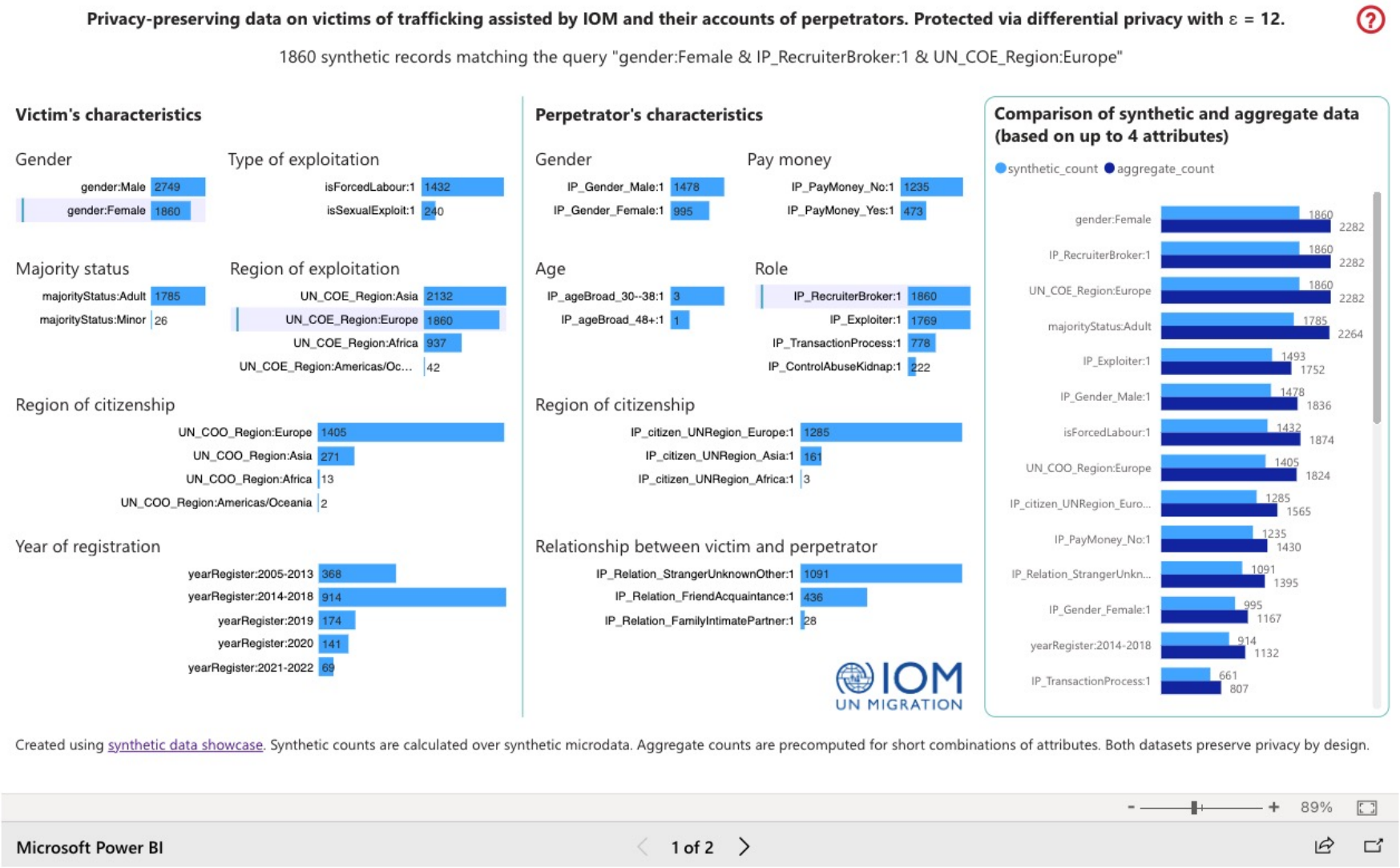


Victim's characteristics, e.g.,

- Country of exploitation
- Citizenship
- Type of trafficking, labour exploit, sexual exploit
- Means of control
- Trafficking duration
- and more…

Global Victim-Perpetrator Synthetic Dataset

You can download data, codebook, and data dictionary



Perpetrator' characteristics, e.g.,

- Role
- Region of citizenship
- Victim-perpetrator relationship
- Pay money

Victim's characteristics, e.g.,

- Type of exploitation
- Region of exploitation and citizenship

Synthetic Data
Relevance to stakeholders

- More data (that are published safely) can enable more effective research and scalable responses.
- Victim-perpetrator relationships and victims' characteristics can help advance the understanding of risk factors for vulnerability.
- The technology can be used by any stakeholder who wants to collect and publish sensitive data while protecting individual privacy.
- Synthetic data, if well used, can strengthen the evidence base on human trafficking and help address this grave human rights violation.

Want to know more?

- Take this free and self-paced e-learning course on "Standardized Human Trafficking Survivor Data Management"
- Visit the CTDC website
- 🚨 Read/consult this forthcoming joint IOM-UNODC report, "Leveraging Administrative Data to Strengthen the Response to TIP (ICSTIP)" and guidance

Appendix: Different forms of data

Type	Definition
Raw data	Data collected on *data subjects* and **not processed**.
Partially de-identified data	Data **modified only marginally** by removing direct identifiers (e.g., name, ID number, IP address).
Aggregate data	Data combined and presented in a **summarized format** in the form of statistics, etc. (e.g., Global Estimates 2022).
K-anonymized data	Data modified by removing direct identifiers, **reducing details** (e.g., 23 becomes 18–24), and **redacted outliers** (e.g., k = 10).
Synthetic data	Data that are **artificially created** rather than obtained through direct measurement, but the statistical properties and relationships from the original data are preserved.

Appendix: Synthesizing data with full k-anonymity (Microsoft)

Step 1

	ATTRIBUTES/COLUMNS			
row ID	A	B	C	D
1	a1	b1	c1	d1
2	a2	b2	c1	d1
3	a1	b2	c2	d1
4	a2	b1	c1	d2
5	a2	b2	c3	d1
6	a1	b1	c3	d1
7	a1	b2	c1	d2

value of attribute	freq.	rel. freq.	cumu. freq.	RAND [0,1]
a1	4	0.25	0.25	0.1
b1	3	0.19	0.50	
c1	4	0.25	0.69	
d1	5	0.31	0.94	
total	16			

new1	a1			

Step 2

	ATTRIBUTES/COLUMNS			
ID	A	B	C	D
1	a1	b1	c1	d1
3	a1	b2	c2	d1
6	a1	b1	c3	d1
7	a1	b2	c1	d2

value of attribute	freq.	rel. freq.	cum. freq.	RAND [0,1]
b1	2	0.29	0.29	
c1	2	0.29	0.57	0.6
d1	3	0.43	1	
total	7			

new1	a1		c1	

Step 3

	ATTRIBUTES/COLUMNS			
ID	A	B	C	D
1	a1	b1	c1	d1
7	a1	b2	c1	d2

value of attribute	freq.	rel. freq.	cumu. freq.	RAND [0,1]
b1	1	0.5	0.5	
d1	1	0.5	1	0.75
total	2			

new1	a1		c1	d1

Step 4

	ATTRIBUTES/COLUMNS			
ID	A	B	C	D
1	a1	b1	c1	d1

We could not include attribute B

new1	a1	NULL	c1	d1
new2	NULL	b1	c1	d1

* "freq." stands for frequency. "rel." stands for relative. "cumu." stands for cumulative.