

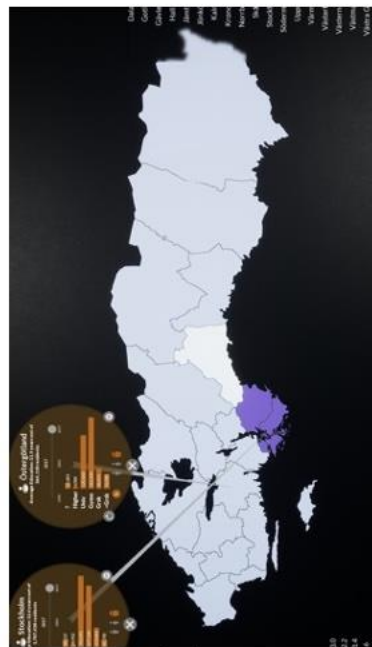
INTERSECTIONALITY & SYNTHETIC DATA

Ericka Johnson

Professor of Gender & Society, Linköping University

Graduate School Director for the Wallenberg AI, Autonomous Systems & Software Program, Humanity & Society

Swedish region with the largest proportion of people with university degrees



Real data



Synthetic data

Intersectional hallucinations

SYNTHETIC DATA

Images



Text



Research

Safety

ChatGPT

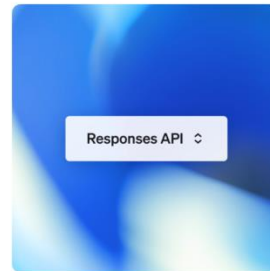
Sora

API Platform

For Business

Stories

Company



Data in numbers

GenAI & Data imbalances

You can use AI to
generate data

Women and heart attacks

Different skin colours for
facial recognition algorithms

Balance data

Privacy

Avoid regulation

”portable data”

Amplification

Adjust the income levels of a
certain group to train a more
equitable mortgage allocation
algorithm.

But...

SYNTHETIC (fake) DATA

Edge cases...

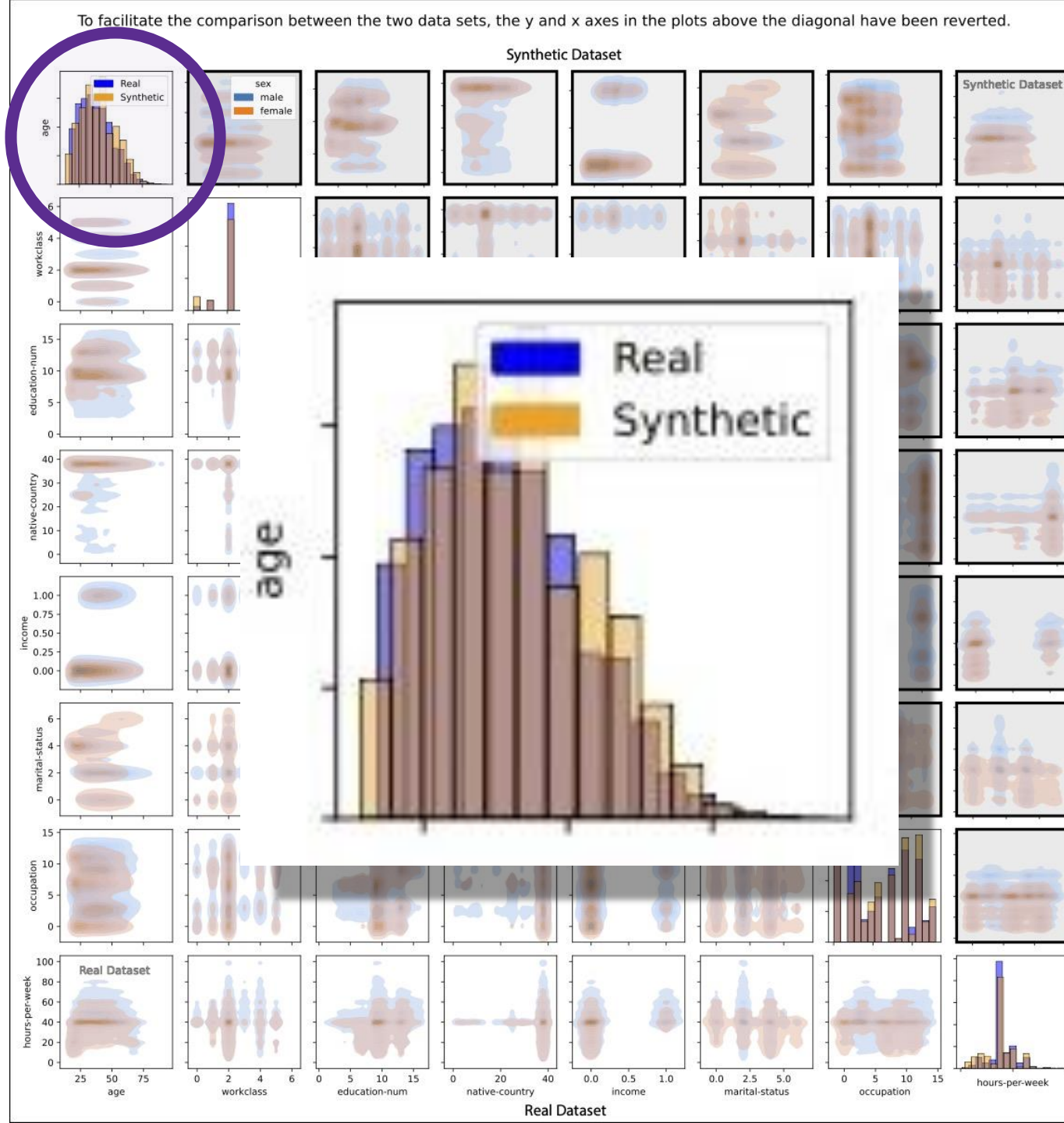
...go missing

'Cambodia', 'Canada', 'China',
'Columbia', 'Cuba', 'Dominican-
Republic', 'Ecuador', 'El-Salvador',
'England', 'France', 'Germany',
'Greece', 'Guatemala', 'Haiti',
'Honduras', **'Hong Kong'**,
'Hungary', 'India', 'Iran', **'Ireland'**,
'Italy', 'Jamaica', 'Japan', **'Laos'**,
'Mexico', 'Nicaragua', 'Outlying-
US(Guam-USVI-etc)', 'Peru',
'Philippines', 'Poland', 'Portugal',
'Puerto-Rico', **'Scotland'**, 'South',
'Taiwan', **'Thailand'**,
'Trinidad&Tobago', 'United-
States', 'Vietnam', **'Yugoslavia'**

And...

- Complicated relationships between attributes.
- Not always known relationships.

Age in the real
& the synthetic data

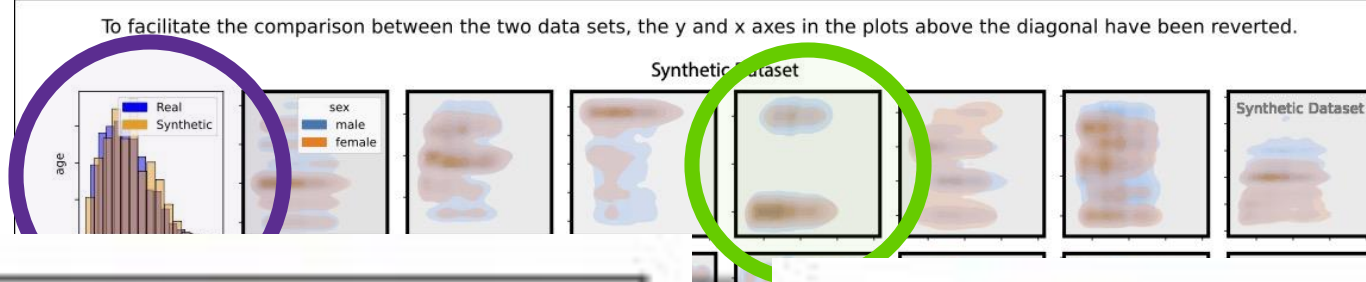


Synthetic USA
census data, 1990

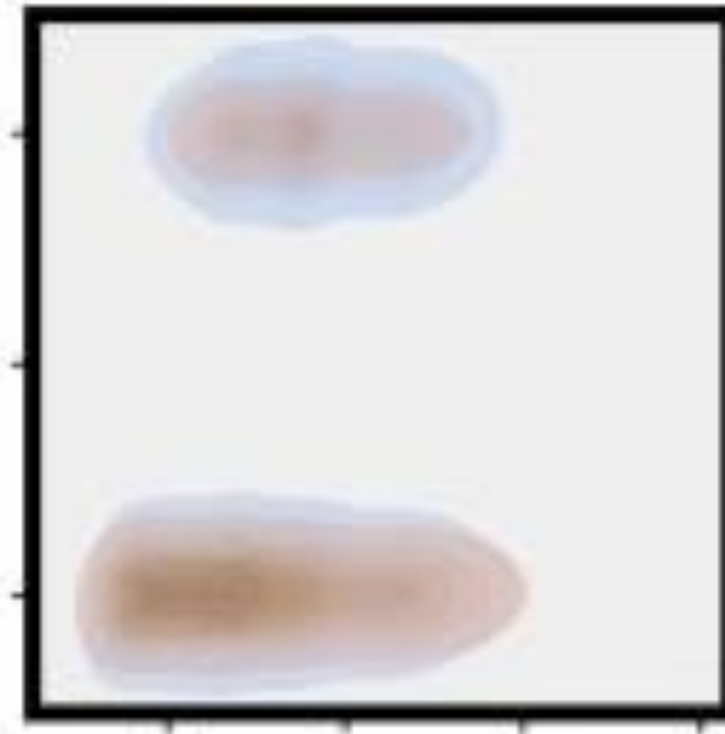
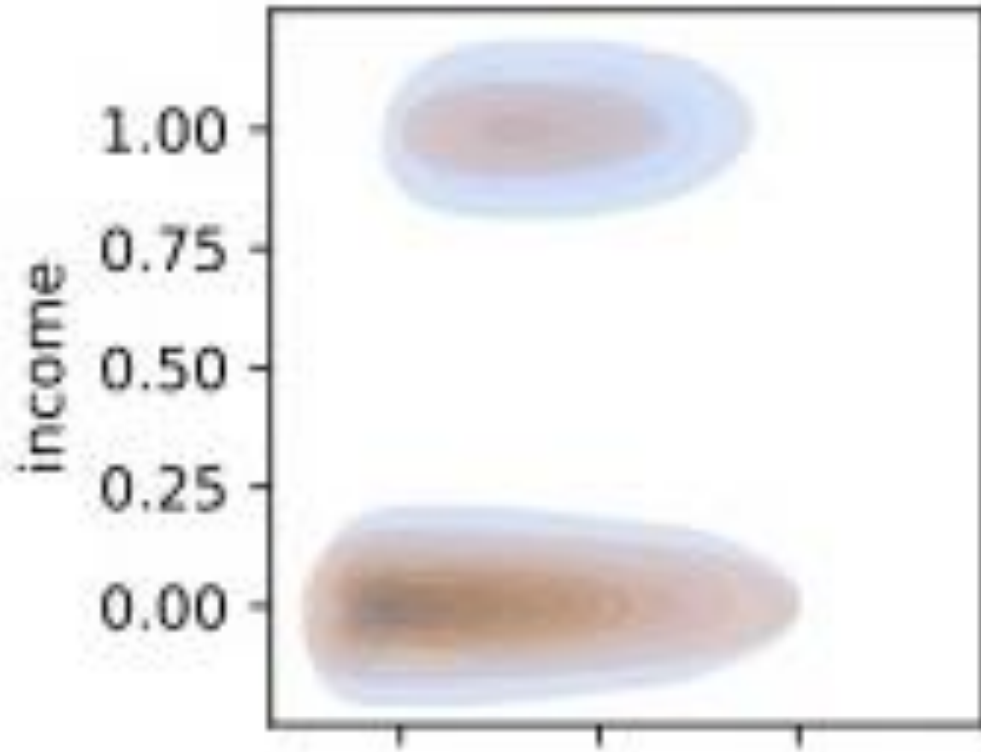
USA Census
data, 1990

*Synthetic data generated and
visualized by Dr. Saghi Hajisharif*

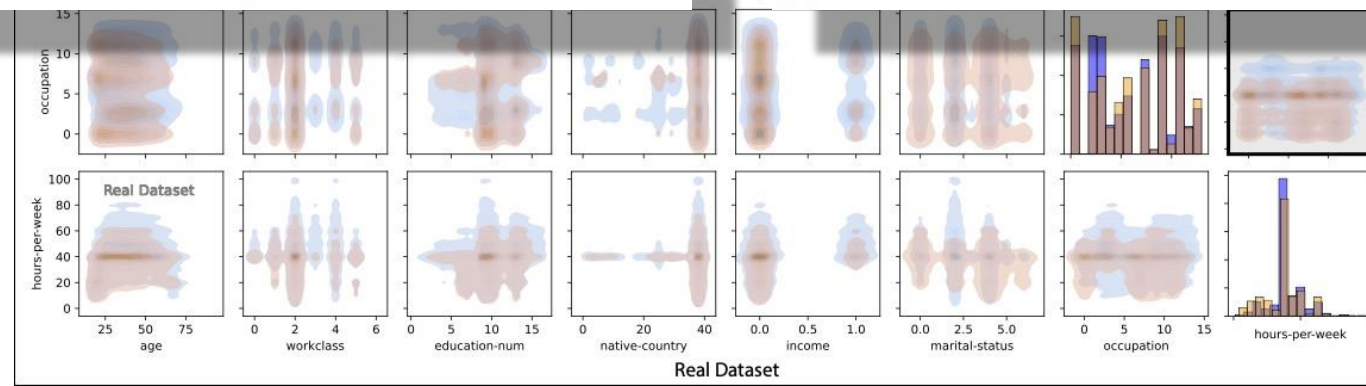
Age in the real
& the synthetic data



Synthetic USA
census data, 1990

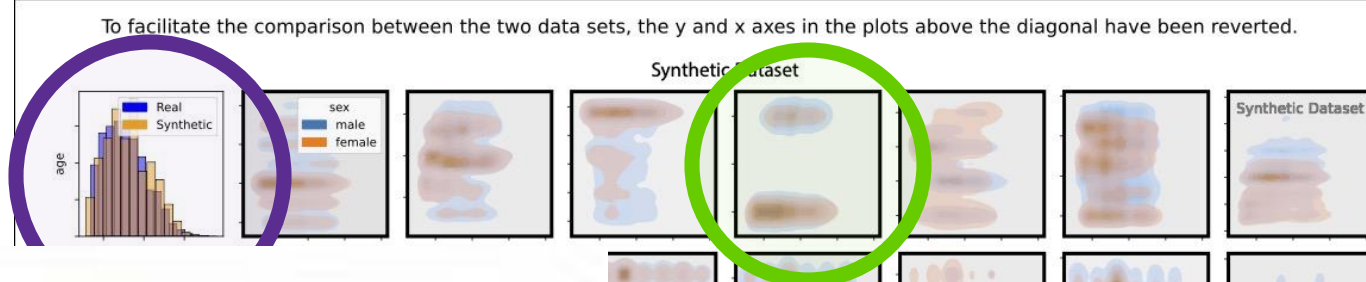


USA Census
data, 1990



*Synthetic data generated and
visualized by Dr. Saghi Hajisharif*

Age in the real
& the synthetic data

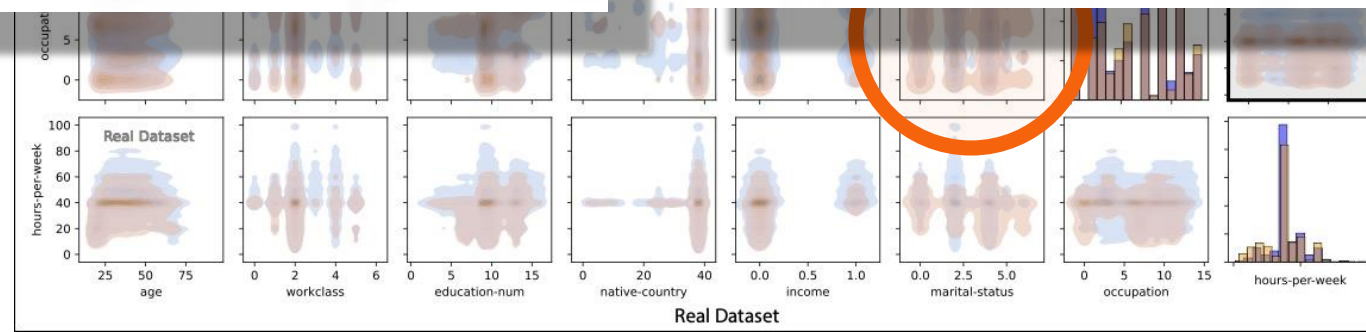


Synthetic USA
census data, 1990

1 female husband

259 female husbands

USA Census
data, 1990



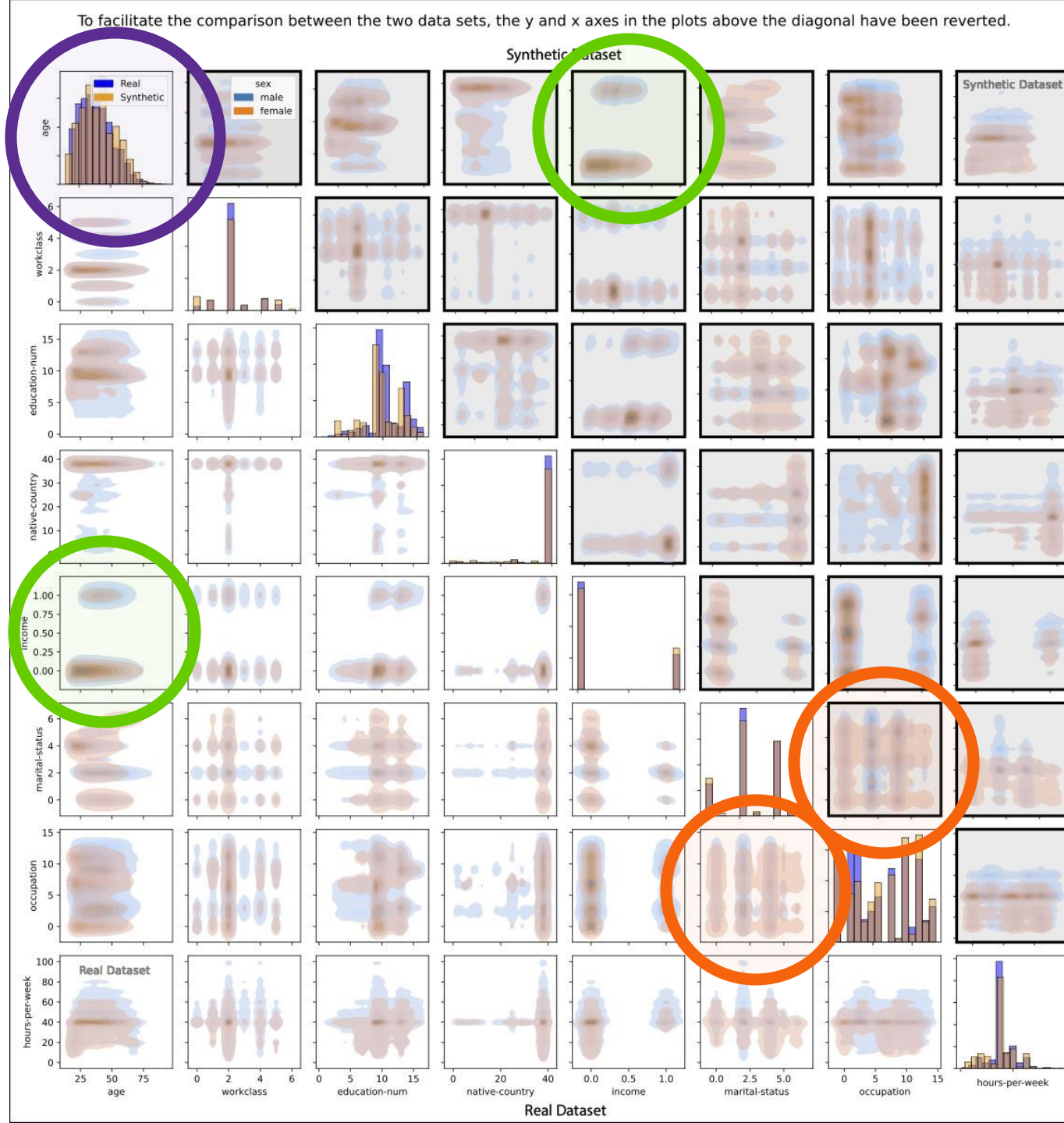
Synthetic data generated and
visualized by Dr. Saghi Hajisharif

Age in the real
& the synthetic data

Intersectional
fidelity

Age, income
& gender

USA Census
data, 1990



Synthetic USA
census data, 1990

333 points =
husband/wife
+ single

Occupation,
marital status
& gender

Intersectional
hallucination

*Synthetic data generated and
visualized by Dr. Saghi Hajisharif*

The figure displays a 10x10 grid of plots comparing Real and Synthetic datasets across 10 variables: age, workclass, education-num, native-country, income, marital-status, occupation, and hours-per-week. The diagonal plots show histograms for Real (blue) and Synthetic (orange) data. The off-diagonal plots show density distributions for each variable pair. Three circles highlight specific areas: a purple circle around the top-left (age, workclass), a green circle around the top-right (age, income), and an orange circle around the bottom-right (occupation, hours-per-week).

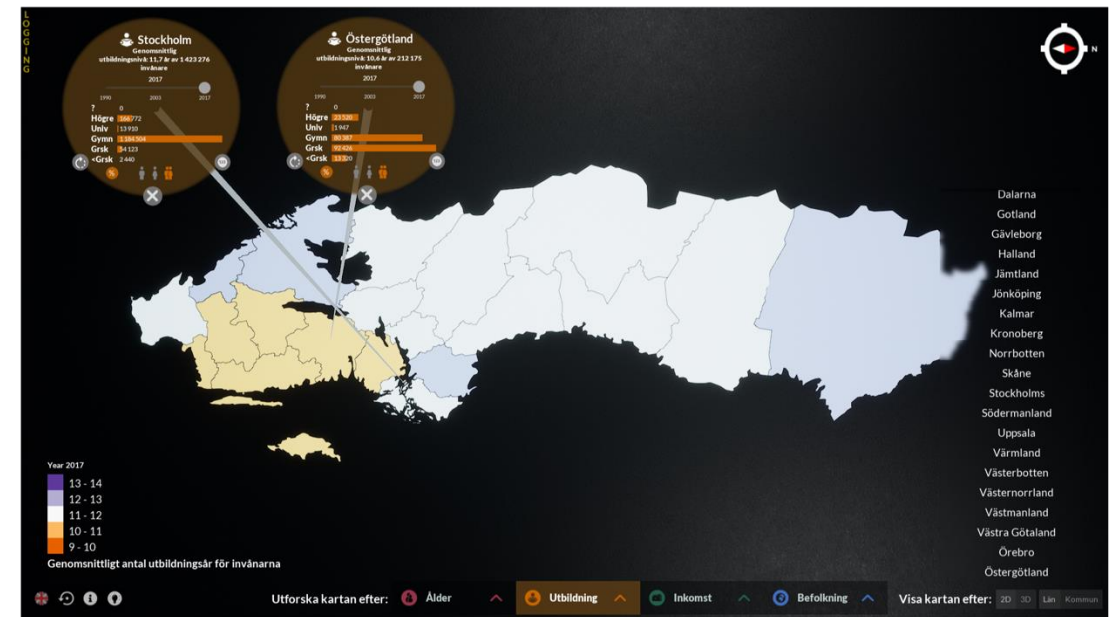
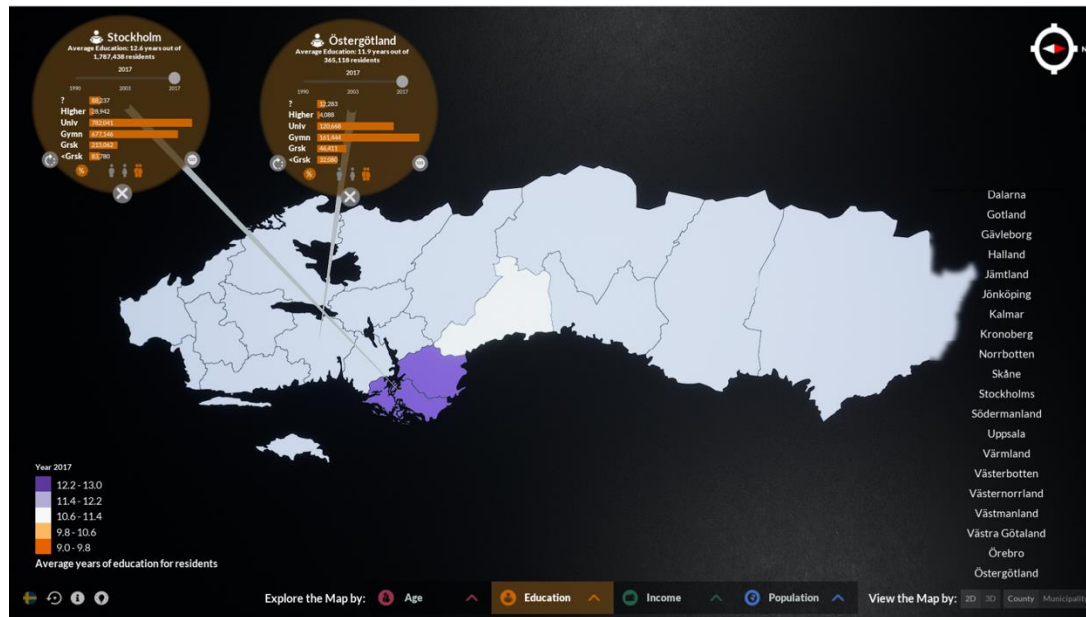
Data generated by:
Zhang, Hengrui, et al. "Mixed-type tabular
data synthesis with score-based diffusion in
latent space." 2024.

68 points =
husband/wife
+ single

*Synthetic data generated and
visualized by Dr. Saghi Hajjsharif*

SYNTHETIC SWEDEN IN NUMBERS

Q2: Which county has the largest proportion of residents with university degrees?



GUARDRAILS for transparency and reproducibility

BUILDING GAURDRAILS



AI Policy Lab
at Umeå University

HOME ABOUT RESEARCH AIPEX NEWS AND EVENTS

METADATA/README ELEMENTS FOR SYNTHETIC STRUCTURED DATA REPOSITORIES TO ENCOURAGE TRANSPARENT, REPRODUCIBLE AND RESPONSIBLE DATA SHARING



Swedish National Data Service

Published February 18, 2025 | Version 1

Managing and publishing synthetic research data

Swedish National Data Service 

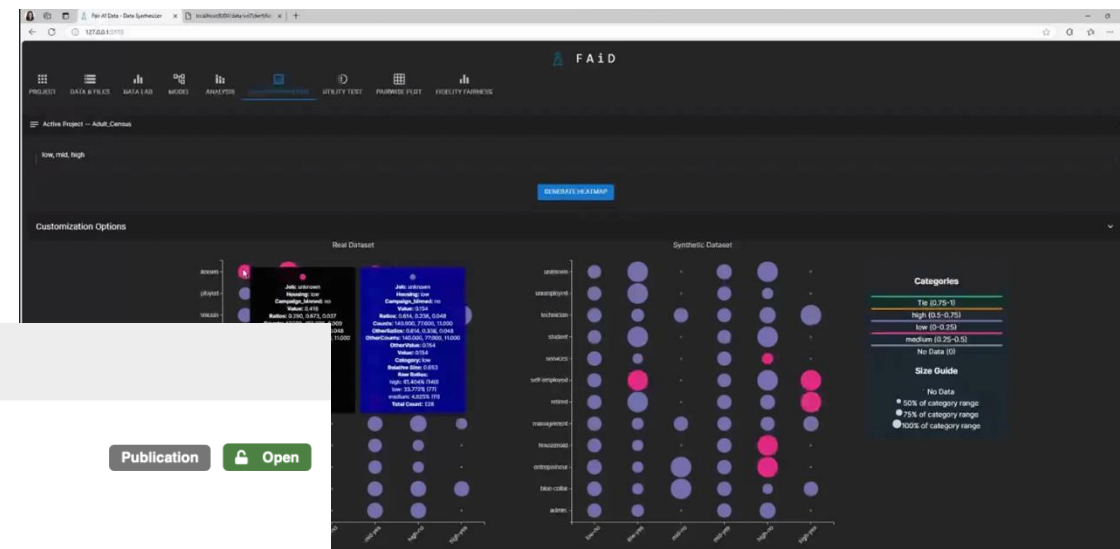
Contributors

Producer: Swedish National Data Service 

Researchers: Johnson, Ericka^{1,2} ; Kasmire, J.^{3,4} ; Hennetier, Victor² 

Show affiliations

This document provides guidance on organizing and documenting datasets that contain synthetic data to simplify publication in a research data repository. Unlike datasets collected from the "real world", synthetic data often require additional details to facilitate reproduction and reuse. This document summarizes the essential information that you should provide when sharing synthetic data in a research data repository to ensure that the data can be easily understood and efficiently reused by others.



<https://zenodo.org/records/14887525>

<https://aipolicylab.se/2025/03/20/metadata-readme-elements-for-synthetic-structured-data-made-with-genai-recommendations-to-data-repositories-to-encourage-transparent-reproducible-and-responsible-data-sharing/#ai-metadata>

<https://fairaidata.com/synthetic-data-for-open-science>

THANKS



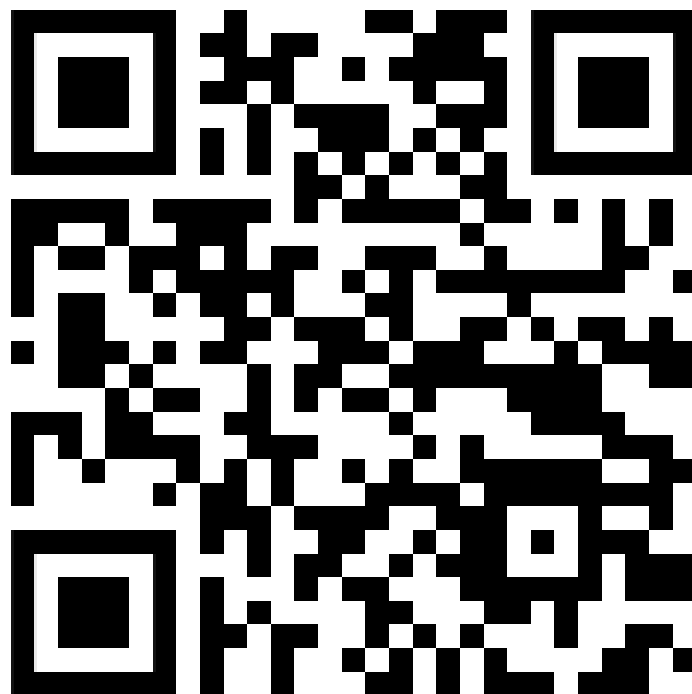
Gruppen för Syntetisk Data:

Dr. Saghi Hajisharif

Yin He

Rickard Falk

Måns Gezelius



Tisdag 12 november 2024

IVAs 100-lista 2024:
Sveriges mest lovande
forskningsprojekt

