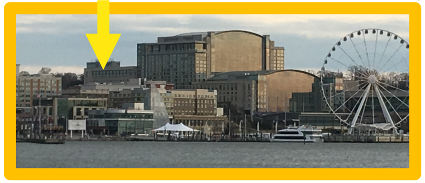
Synthetic Data Quality Metrics: Relative vs. Absolute

Christine Task, PhD
Senior Computer Scientist
Knexus Research Corporation

The views and opinions expressed in this talk are those of the authors and not the U.S Census Bureau.

The views and opinions expressed in this talk are those of the authors and not the National Institute of Standards and Technology.

Knexus Research

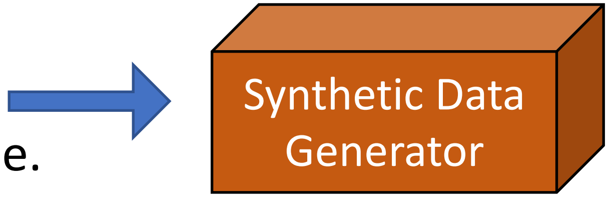


Knexus Research is a small R&D company located in the **DC area** at National Harbor, MD.

We have two active projects supporting Privacy-preserving Synthetic Data Generation:

- For the **US Census Bureau**, the Knexus CenSyn team is providing evaluation, research, engineering and production software development support for Census privacy efforts.
- As technical lead for the **NIST Differentially Private Synthetic Data Challenge**, Knexus provided technical guidance for the first national challenge in Differential Privacy.

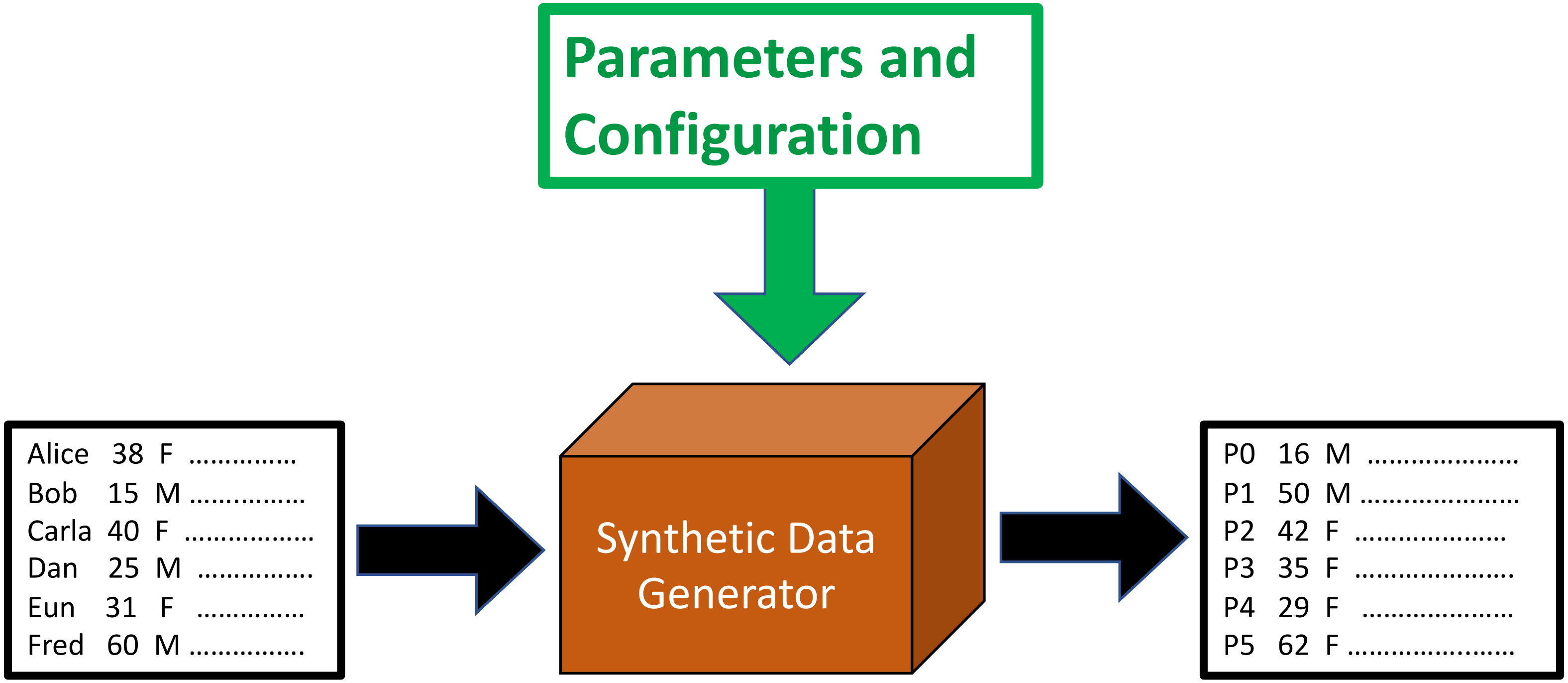
Over the past year, we've thought very carefully about what it means to make a good one of these.



Synthetic Data Generator

This talk will provide a quick orientation to synthetic data quality evaluation

Synthetic Data Generation



Parameters and Configuration

Synthetic Data Generator

Alice	38	F	....
Bob	15	M	....
Carla	40	F	....
Dan	25	M	....
Eun	31	F	....
Fred	60	M	....

P0	16	M	....
P1	50	M	....
P2	42	F	....
P3	35	F	....
P4	29	F	....
P5	62	F	....

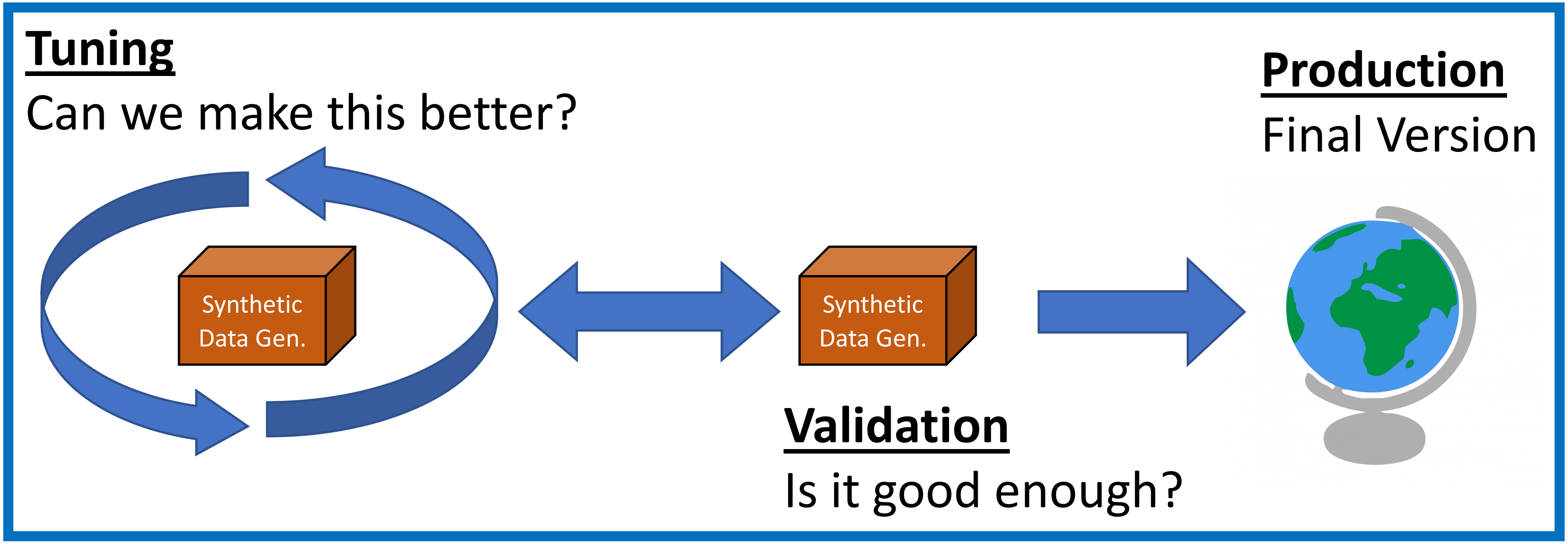
Ground Truth Data

Data on ***real*** individuals

Synthetic Data

Data on ***sanitized*** individuals

Evaluation Process for Synthetic Data Generators



Tuning

Can we make this better?

Validation

Is it good enough?

Production

Final Version

Synthetic Data Generation is essentially a task of fitting a generative model to a data-set. The basic evaluation process is familiar from data analytics.

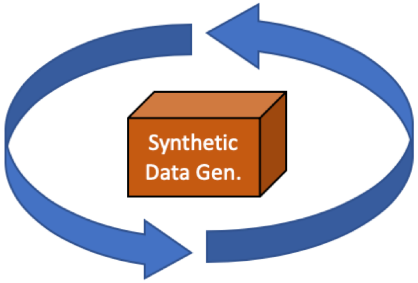
However, these models output complex, high dimensional, potentially sparse data, which will be used in turn to train models in downstream analytics, with accuracy degrading at each step in the chain. Careful attention to evaluation is vital.

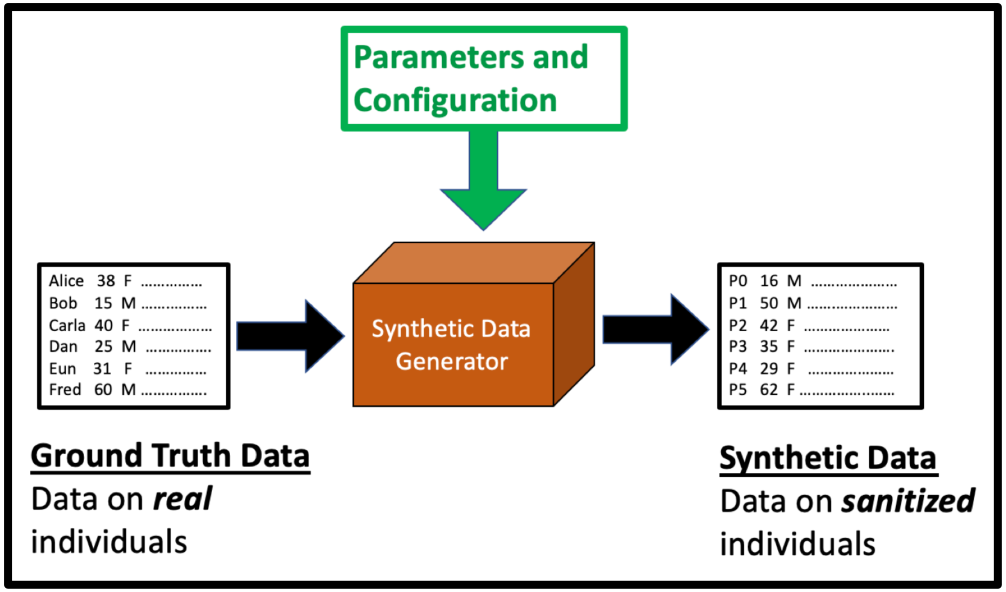
Snapshot Metrics for Tuning:

- Synthesizer output quality depends on: Data Encoding/Pre-processing, Model Choice, Training Process, Post-processing… and parameter choices for all of the above.
- To compare different options we need ***relative*** metrics--quick snapshots that allow us to study the quality distributions of synthesizer output.

Tuning

Can we make this better?



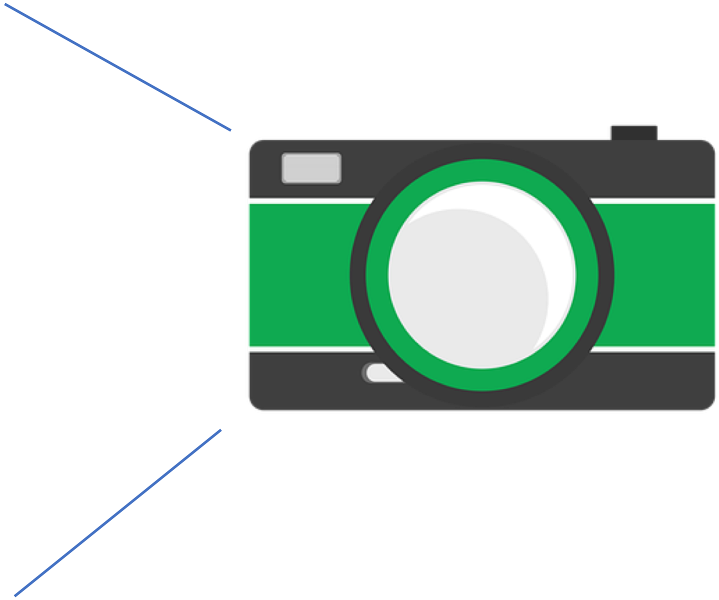


Alice	38	F	....
Bob	15	M	....
Carla	40	F	....
Dan	25	M	....
Eun	31	F	....
Fred	60	M	....

P0	16	M	....
P1	50	M	....
P2	42	F	....
P3	35	F	....
P4	29	F	....
P5	62	F	....

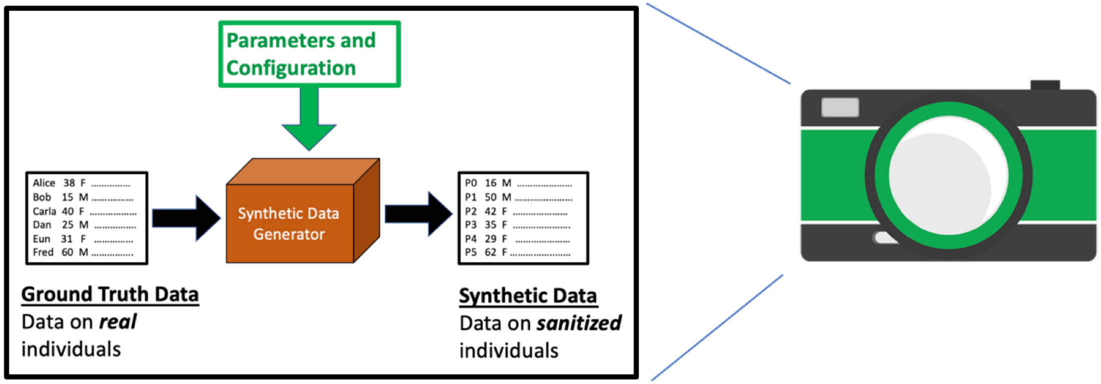
Ground Truth Data
Data on ***real*** individuals

Synthetic Data
Data on ***sanitized*** individuals



Snapshot Metrics for Tuning:

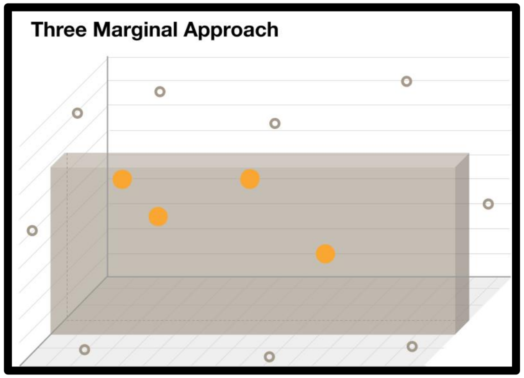
To compare different options we need ***relative*** metrics--quick snapshots that allow us to study the quality distributions of synthesizer output.



Below are a few example classes of these metrics:

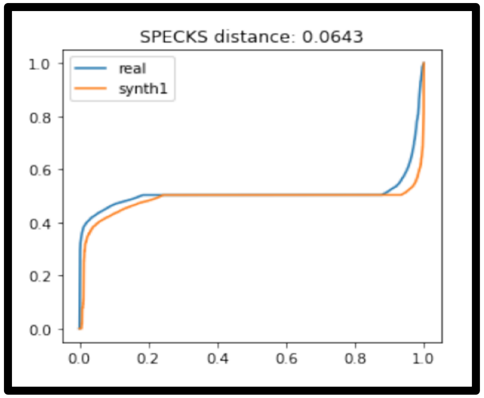
Distance Based Metrics

These compute absolute deviation under norms (L1, L2)



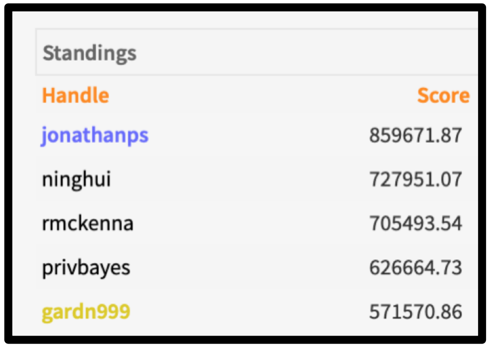
Propensity Based Metrics

These rely on classifiers distinguishing the real and synthetic data (SPECKS, pMSE)



Randomized Heuristics

These efficiently capture distributional similarity (randomized 3-marginal, row pool)

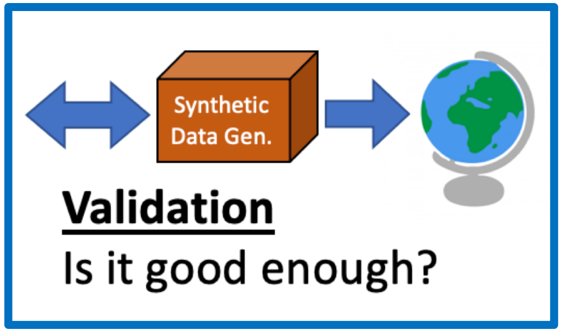


Standings

Handle	Score
jonathanps	859671.87
ninghui	727951.07
rmckenna	705493.54
privbayes	626664.73
gardn999	571570.86

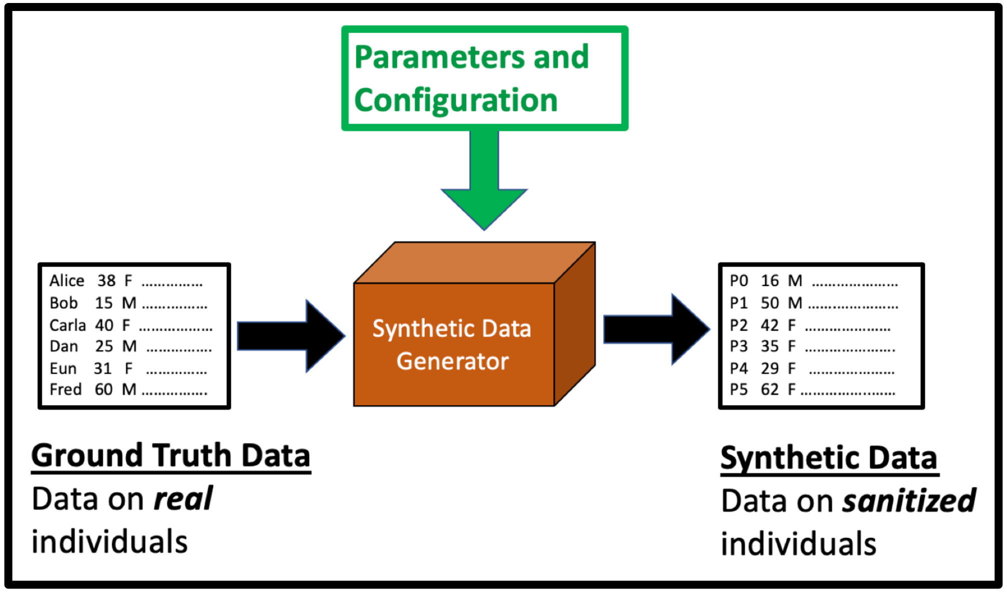
Deep-dive Metrics for Validation:

- Snapshots are convenient, but have problematic shortcomings: blind spots, bias… they don’t tell us how the synthetic data will work in practice.
- To validate synthesizers, we need ***absolute*** metrics--deep dive tools that help identify, understand, and measure the impact of distributional discrepancies between the ground truth and synthesizer output.



Validation

Is it good enough?



Alice	38	F	……………
Bob	15	M	……………
Carla	40	F	………………
Dan	25	M	……………
Eun	31	F	……………
Fred	60	M	……………

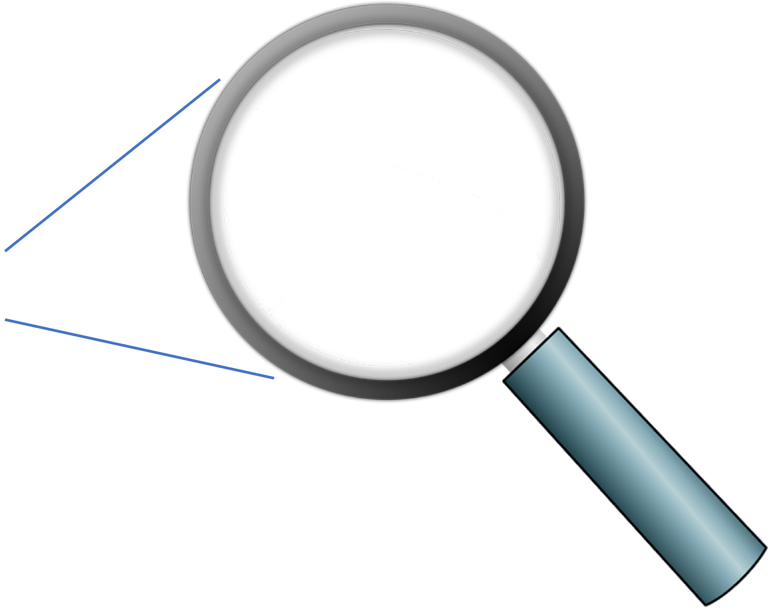
Ground Truth Data

Data on ***real*** individuals

P0	16	M	…………………
P1	50	M	…………………
P2	42	F	…………………
P3	35	F	…………………
P4	29	F	…………………
P5	62	F	…………………

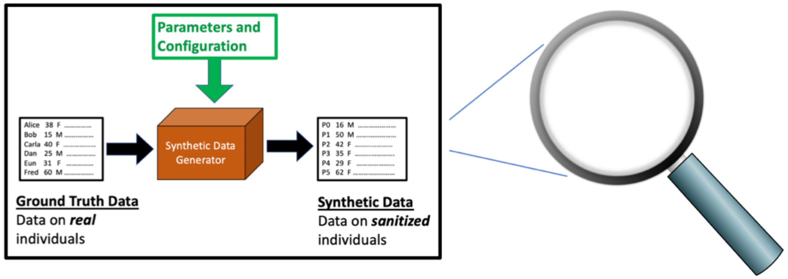
Synthetic Data

Data on ***sanitized*** individuals



Deep-dive Metrics for Validation:

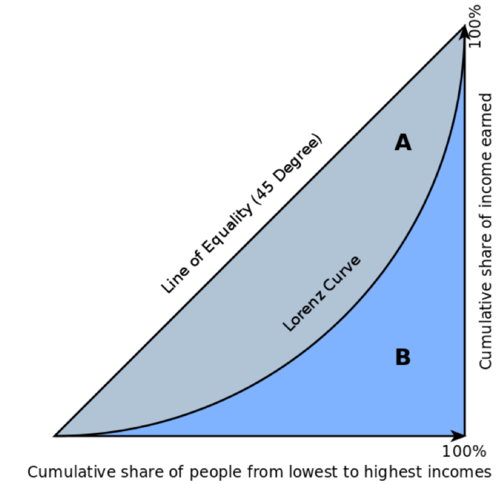
To validate synthesizers, we need ***absolute*** metrics--deep dive tools that help identify, understand, and measure the impact of distributional discrepancies between the ground truth and synthesizer output.



Below are a few example classes of these metrics:

Analytics and Use Cases

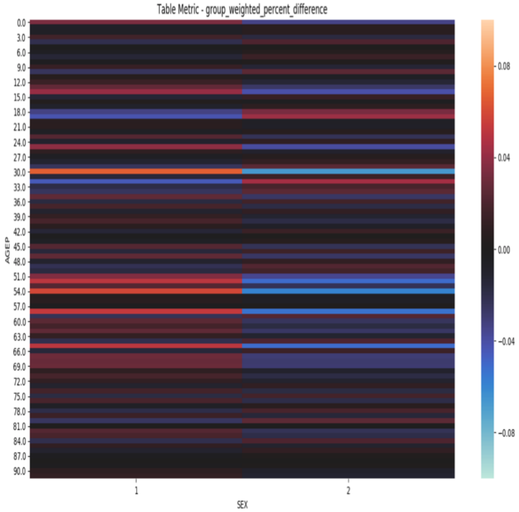
Example analytics can check challenging cases such as long tails, differences-of-differences



Graphical representation of the Gini coefficient

Heatmap Tools

Table deviation heatmaps can identify fine grained regions and patterns of problems



Frequent Itemset Analysis

Post-processing on distance analysis can identify variables most to blame for deviations

CAD_NUMBER
DISPATCH_DATETIME
HOSTPITAL_DATETIME
WATCH_DATE
ALS_UNIT
UNIT_SEQUENCE

Questions?

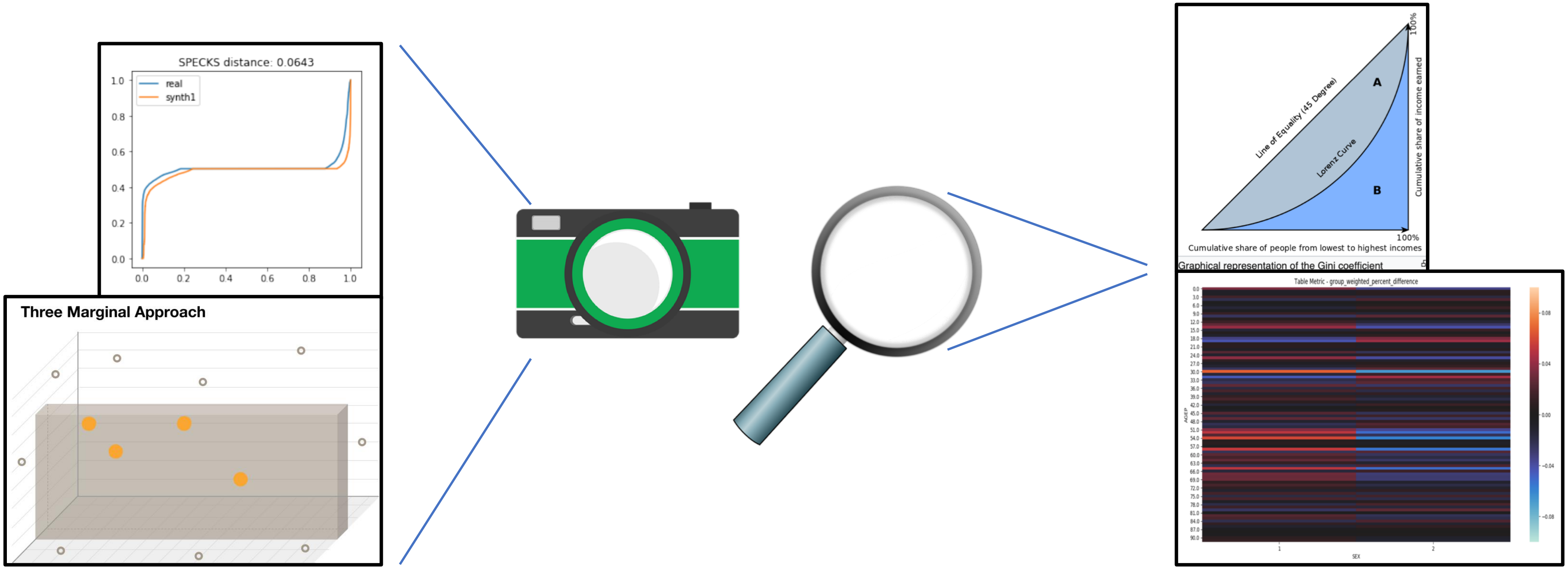
Lightning talks rarely have much time for questions—Come talk to me afterwards or send me an email if this was interesting, or if you have insight/ideas to share!

Contact Details

Talk Topic: Synthetic Data Quality Metrics: Relative vs. Absolute

Affiliation: **Knexus Research Corporation**

Contact Email: Christine.Task@knexusresearch.com


SPECKS distance: 0.0643
real
synth1
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Three Marginal Approach
Line of Equality (45 Degree)
Lorenz Curve
A
B
100%
100%
Cumulative share of income earned
Cumulative share of people from lowest to highest incomes
Graphical representation of the Gini coefficient
Table Metric - group_weighted_percent_difference
AGEP
SEX