AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

Data Science to Patient Value (D2V)
UNIVERSITY OF COLORADO **ANSCHUTZ MEDICAL CAMPUS**

Kung Faux Pandas: Simplifying privacy protection

Oral Presentations - Privacy and Bias in Data Science

S45

Seth Russell

University of Colorado Anschutz Medical Campus

Twitter: @magic__lantern

GitHub: magic-lantern;

Project URL: https://github.com/CUD2V/kungfauxpandas

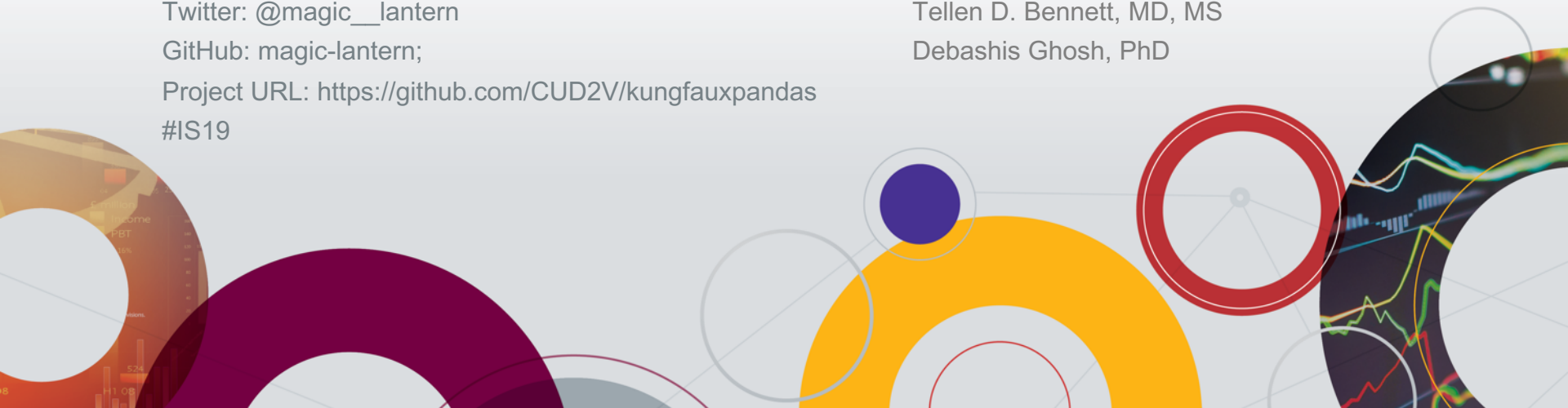
#IS19

paper co-authors:

James King MIDS

Tellen D. Bennett, MD, MS

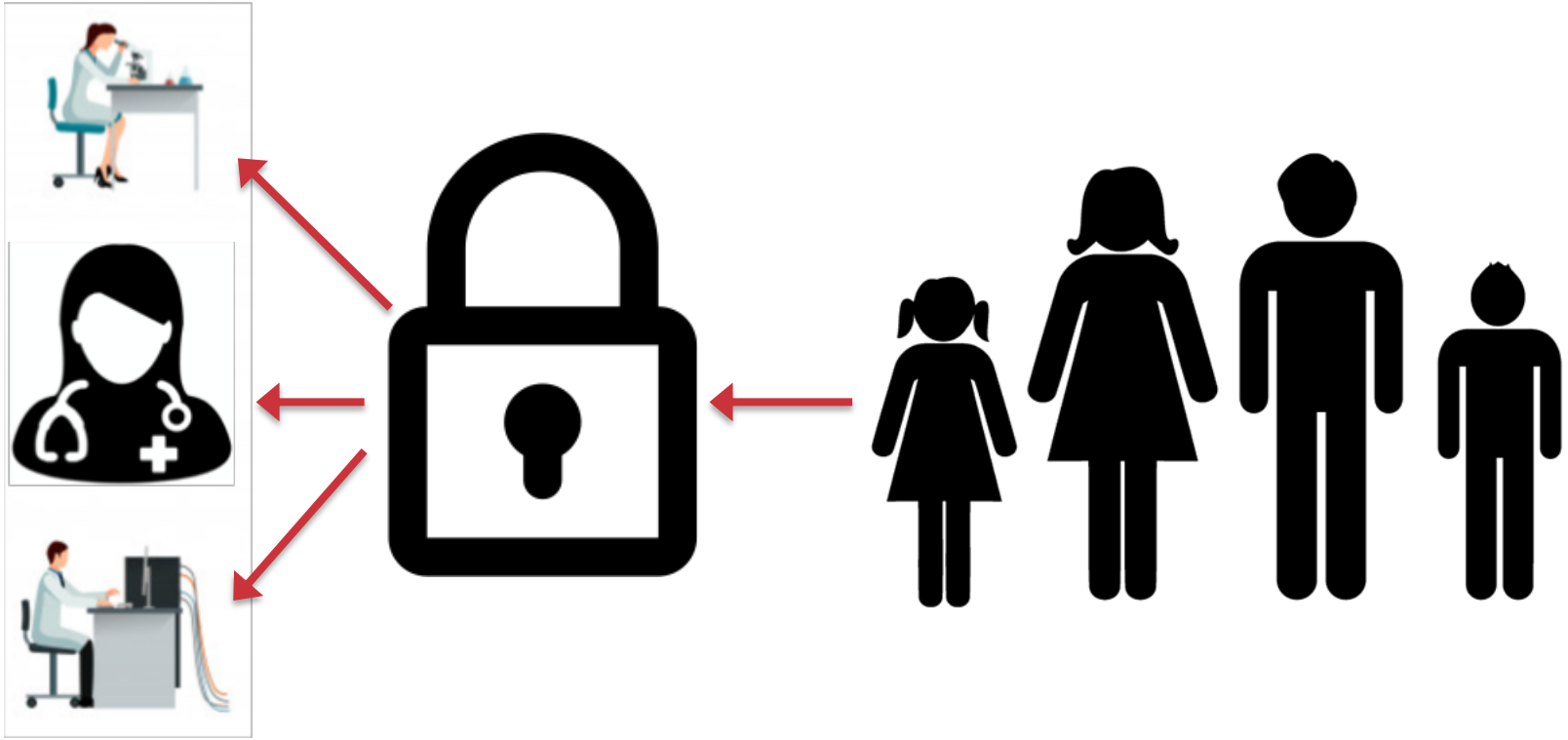
Debashis Ghosh, PhD



Disclosure

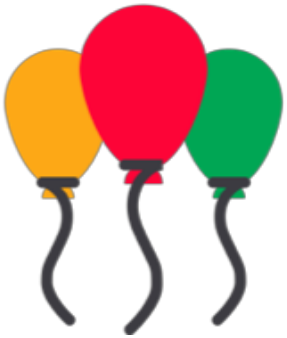
I and my co-authors have no relevant relationships with commercial interests to disclose.

Why Kung Faux Pandas?

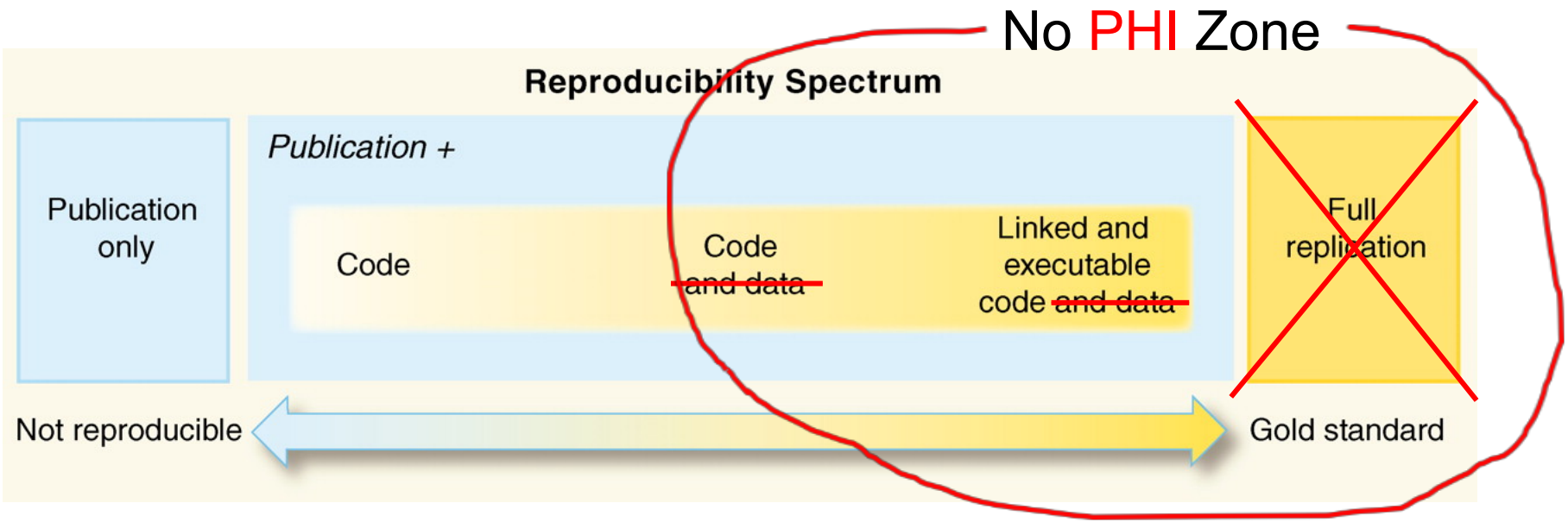


https://github.com/CUD2V/kungfauxpandas

Why Kung Faux Pandas?

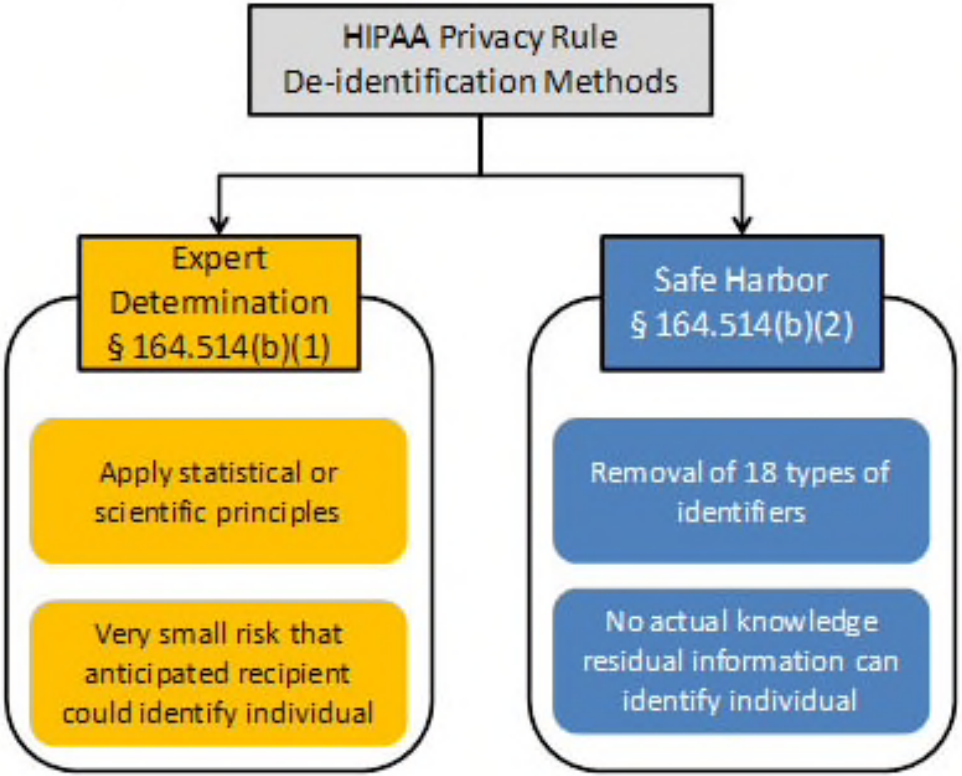


Reproducibility and Replicability



Peng RD. Reproducible research in computational science. Science (New York, Ny). 2011 Dec;334(6060):1226– 1227. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383002/

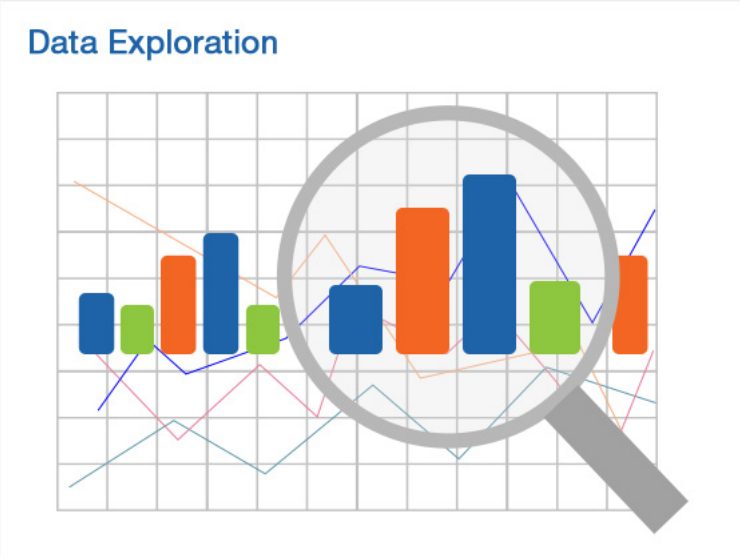
Reproducibility and Replicability



https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Why Kung Faux Pandas?

Privacy protection through data synthesis



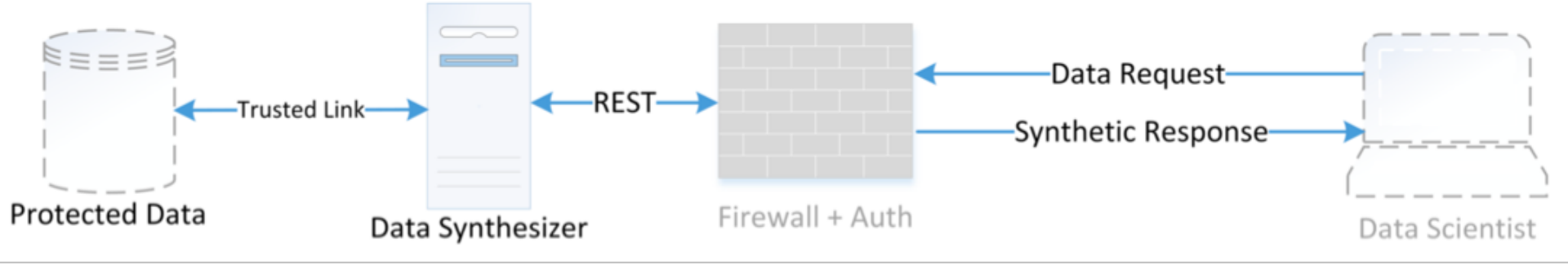
https://github.com/CUD2V/kungfauxpandas

Why Kung Faux Pandas?

Privacy protection through data synthesis

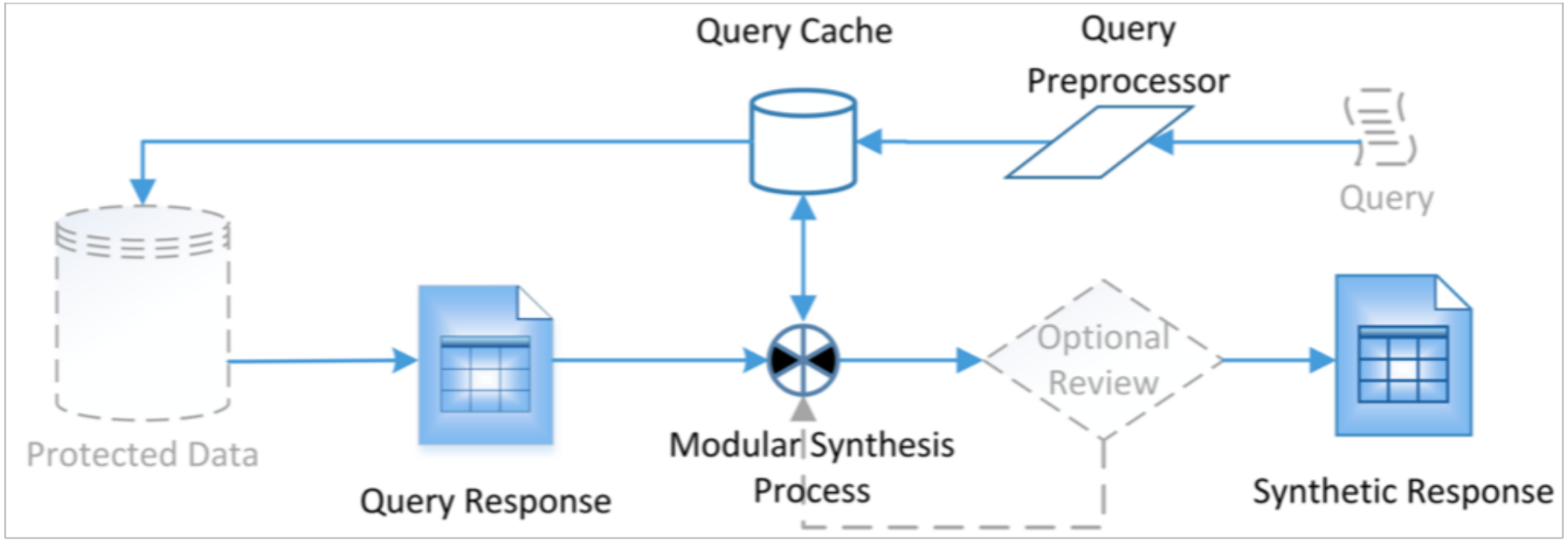
https://github.com/CUD2V/kungfauxpandas

KFP System Architecture Part 1



https://github.com/CUD2V/kungfauxpandas

KFP System Architecture Part 2



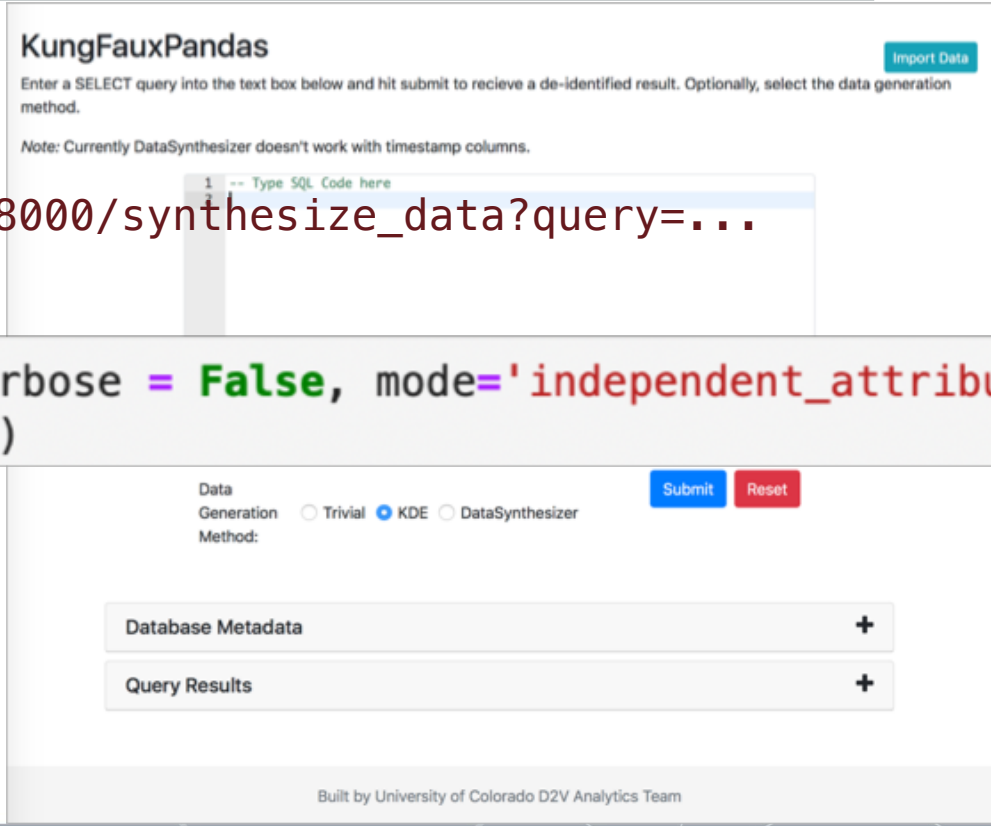
Provided Interfaces

- HTML/JS UI
- REST API

```
HTTP GET http://localhost:8000/synthesize_data?query=...
```

- Python API

```
kfpd.plugin = KDEPlugin(verbose = False, mode='independent_attribu
fdf=kfpd.read_sql(sql,conn)
```



https://github.com/CUD2V/kungfauxpandas

Included Synthesis Methods

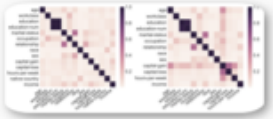
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

SciPy.org

scipy.stats.gaussian_kde

DataSynthesizer: Privacy-preserving synthetic datasets



To facilitate collaboration over sensitive data, we present DataSynthesizer, a tool that takes a sensitive dataset as input and generates a structurally and statistically similar synthetic dataset with strong privacy guarantees.

Haoyue Ping, Julia Stoyanovich and Bill Howe

Proceedings of SSDBM, 2017

Included Synthesis Methods

Independent Attribute Mode

SubjectId	EncounterId	Source	StartDate	Code	Type
47644390	69	Encounter	2017-10-05 16:35:35	S52283B	ICD-10-CM
51058881	269	Patient History	2017-06-15 11:11:41	S80242S	ICD-10-CM
99487391	172	Patient History	2017-09-19 13:59:50	H65111	ICD-10-CM
43700149	830	Encounter	2017-06-21 11:22:53	O00112	ICD-10-CM
76049031	843	Billing	2017-09-19 06:54:25	S60412D	ICD-10-CM

Included Synthesis Methods

Correlated Attribute Mode

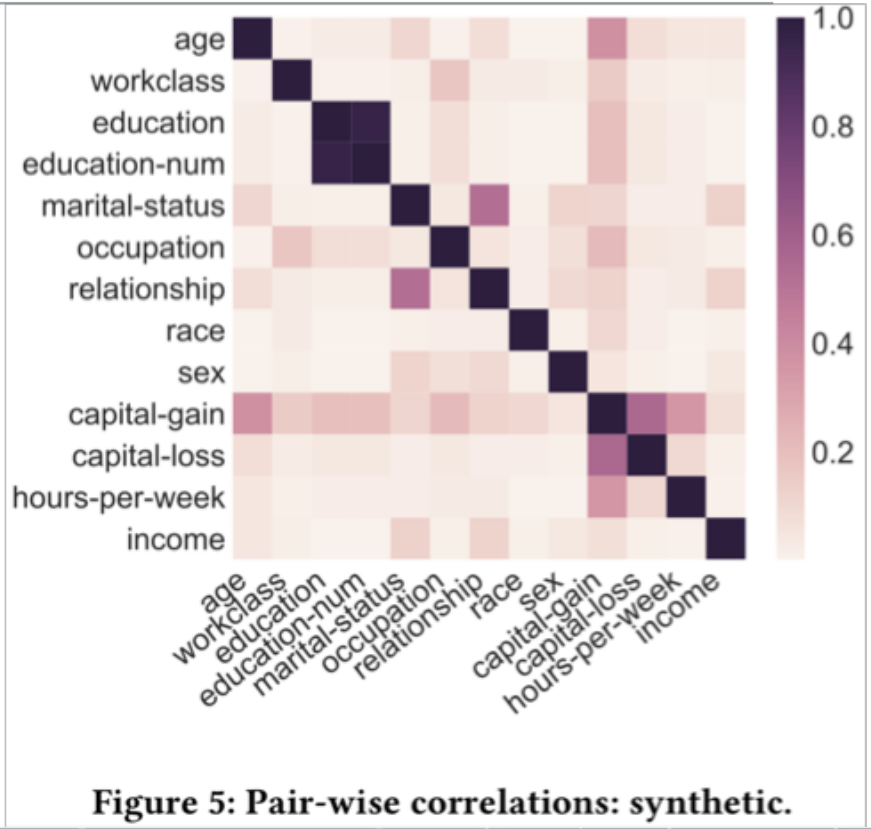


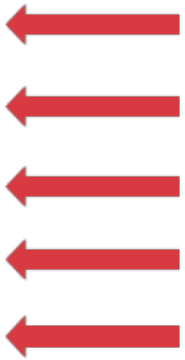
Figure 5: Pair-wise correlations: synthetic.

"DataSynthesizer: Privacy-Preserving Synthetic Datasets"
https://dl.acm.org/citation.cfm?doid=3085504.3091117

Performance of Synthesis Methods

Time to analyze and synthesize 100,000 rows

Method	Seconds
Kernel Density Estimator Independent	0.1
Kernel Density Estimator Correlated	120.0
DataSynthesizer Independent	2.6
DataSynthesizer Correlated	160.0
DataSynthesizer Correlated/custom config	2124.0



sourcecode/python/notebooks/performance_tests.ipynb

Python Pandas compatible API

KFP Setup (sourcecode/python/notebooks/presentation_example.ipynb)

```
1  import sqlite3
2  import pandas as pd
3  import numpy as np
4
5  conn = sqlite3.connect("../../data/sample_data.db")
6  cursor = conn.cursor()
```

```
1  from kungfauxpandas import KungFauxPandas, TrivialPlugin, DataSynthesizerPlugin, KDEPlugin
2  kfpd = KungFauxPandas()
```

```
1  sql = '''
2  SELECT d."SubjectId", d."EncounterId", d."Source", d."Code", d."Type", MAX("FlowsheetValue")
3   FROM diagnoses d LEFT JOIN flowsheet f ON d."EncounterId" = f."EncounterId"
4   GROUP BY d."SubjectId", d."EncounterId", d."Source", d."Code", d."Type"
5   ORDER BY NumLoggedScores DESC
6   limit 1000
7  '''
```

Python Pandas compatible API

	SubjectId	EncounterId	Source	Code	Type	MaxScore	MinScore	NumLoggedScores
0	40552133	288	Encounter	A4152	ICD-10-CM	100.0	0.0	294
1	83299697	625	Encounter	A414	ICD-10-CM	100.0	0.0	286
2	96360391	985	Encounter	A400	ICD-10-CM	100.0	0.0	278
3	43551783	984	Encounter	A392	ICD-10-CM	98.0	0.0	272
4	92110570	934	Encounter	A4151	ICD-10-CM	99.0	0.0	256

https://github.com/CUD2V/kungfauxpandas

Python Pandas compatible API

```
1  kfpd.plugin = KDEPlugin(verbose = False, mode='independent_attribute_mode')
2  fdf=kfpd.read_sql(sql,conn)
3  fdf.head()
```

	SubjectId	EncounterId	Source	Code	Type	MaxScore	MinScore	NumLoggedScores
0	-10308664	141	Billing	N950	ICD-10-CM	100.0	1.0	135
1	31175541	241	Problem List	A5486	ICD-10-CM	100.0	0.0	132
2	67298643	838	Encounter	A327	ICD-10-CM	99.0	0.0	137
3	49070319	807	Patient History	S45001D	ICD-10-CM	97.0	0.0	127
4	71361333	205	Patient History	Y629	ICD-10-CM	99.0	0.0	141

https://github.com/CUD2V/kungfauxpandas

Python Pandas compatible API

```
1  kfpd.plugin = DataSynthesizerPlugin(mode='independent_attribute_mode', verbose=False)
2  fdf=kfpd.read_sql(sql,conn)
3  fdf.head()
```

	SubjectId	EncounterId	Source	Code	Type	MaxScore	MinScore	NumLoggedScores
0	59091572.0	649.0	Problem List	uzfdu	ICD-10-CM	99.729218	0.0	140.0
1	68624515.0	98.0	Encounter	ujrtry	ICD-10-CM	99.036495	0.0	129.0
2	66943722.0	393.0	Encounter	lclylb	ICD-10-CM	95.935947	0.0	135.0
3	56174670.0	738.0	Billing	wzrbbu	ICD-10-CM	99.738615	0.0	129.0
4	52450025.0	117.0	Encounter	zuhamr	ICD-10-CM	99.080197	0.0	133.0

https://github.com/CUD2V/kungfauxpandas

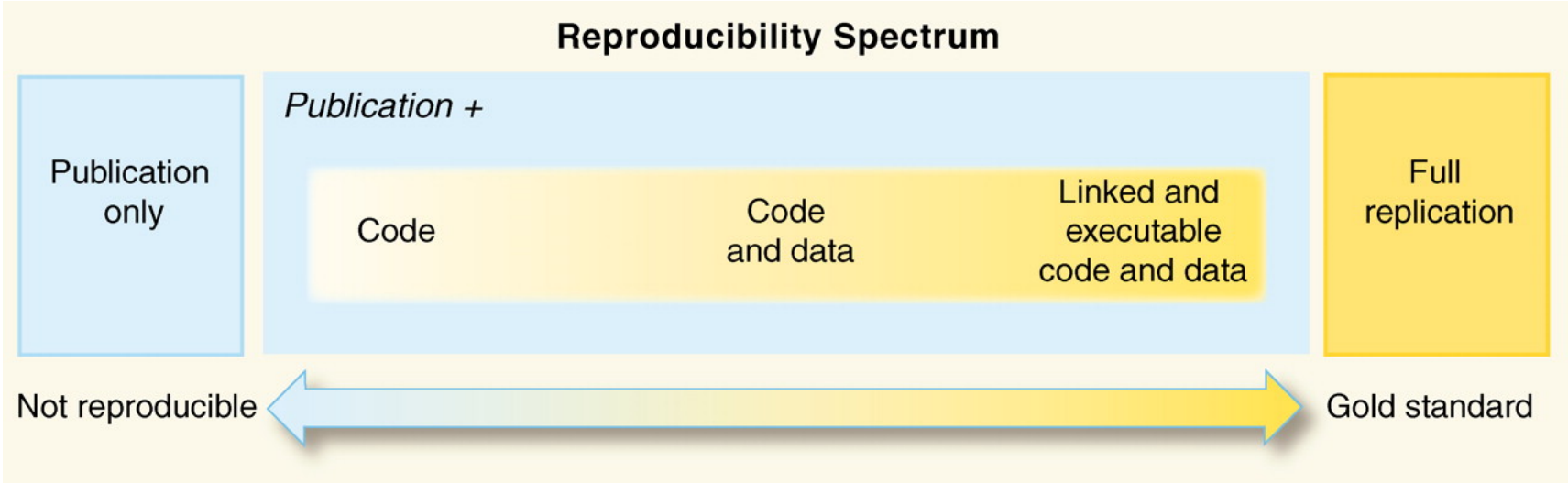
Python Pandas compatible API

```
kfpd.plugin = DataSynthesizerPlugin(mode='correlated_attribute_mode',
                                    candidate_keys = {'SubjectId': True, 'EncounterId': True},
                                    categorical_attributes = {'Source': True,
                                                              'Code': True,
                                                              'Type': True,
                                                              'MaxScore': False,
                                                              'MinScore': False,
                                                              'NumLoggedScores': False}
                                    )
fdf=kfpd.read_sql(sql,conn)
fdf.head()
```

	SubjectId	EncounterId	Source	Code	Type	MaxScore	MinScore	NumLoggedScores
0	0	0	Encounter	S61203D	ICD-10-CM	95.453327	0.0	211.0
1	1	1	Billing	S89311P	ICD-10-CM	99.027386	0.0	267.0
2	2	2	Encounter	S66291A	ICD-10-CM	98.926171	0.0	194.0

https://github.com/CUD2V/kungfauxpandas

Reproducibility and Replicability



Peng RD. Reproducible research in computational science. Science (New York, Ny). 2011 Dec;334(6060):1226– 1227. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383002/

Reproducibility and Replicability

What have we done to promote reuse and reusability?

- Software, documentation, paper, this presentation, etc. available at: https://github.com/CUD2V/kungfauxpandas
- Docker Container: docker pull blackspot/synthesis
- Unit tested software using https://pytest.org

Limitations

- Generates synthetic data only when given a specific SQL query or pandas.DataFrame
- Performance
- Get legal advice before releasing any data generated by KFP
- Provided web application does not include SSL certificates, so data is not encrypted

Areas for Further Research

- Additional plugins.
- Approval and review of results before returning to end user.
- More specific controls – for example minimum input and output sizes
- Develop metrics for privacy guarantees (differential privacy, plausible deniability, etc.)
- Publish code as python package
- Others based on feedback from users

Thank you!

Image credits

- https://www.behance.net/gallery/5408155/Family-Icon
- https://www.freepik.com/free-vector/scientist-decorative-icons-set_2871430.htm
- https://www.vectorstock.com/royalty-free-vector/doctor-icon-medical-consultation-female-physician-vector-18769418
- https://www.flaticon.com/free-icon/lock-icon_26053
- https://cdn3.iconfinder.com/data/icons/wedding-4/512/clink-512.png
- https://cdn.iconscout.com/icon/free/png-256/balloon-decoration-christmas-xmas-celebration-party-30627.png
- https://cdn2.iconfinder.com/data/icons/communication-2/512/Pencil_And_Paper-512.png
- https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/07/Data-Exploration.jpg

Items to include?

Independent reproduction and replication of results are critical components of scientific research. Barriers to data access and data sharing are some of the most important impediments to reproducibility in health data science research. Health care organizations are often reluctant to share data even when it has been de-identified. Through data synthesis, fears about data sharing can be reduced. Through providing ad-hoc synthetic data, health data scientists can have lower barriers to accessing private data as well as reduced barriers to sharing artificial data derived from protected data

Reproducibility: the "ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline,"

Replicability: "the chance that an independent experiment targeting the same scientific question will produce a consistent result."

[Your Presentation]

[Your presentation on this and next slides]