

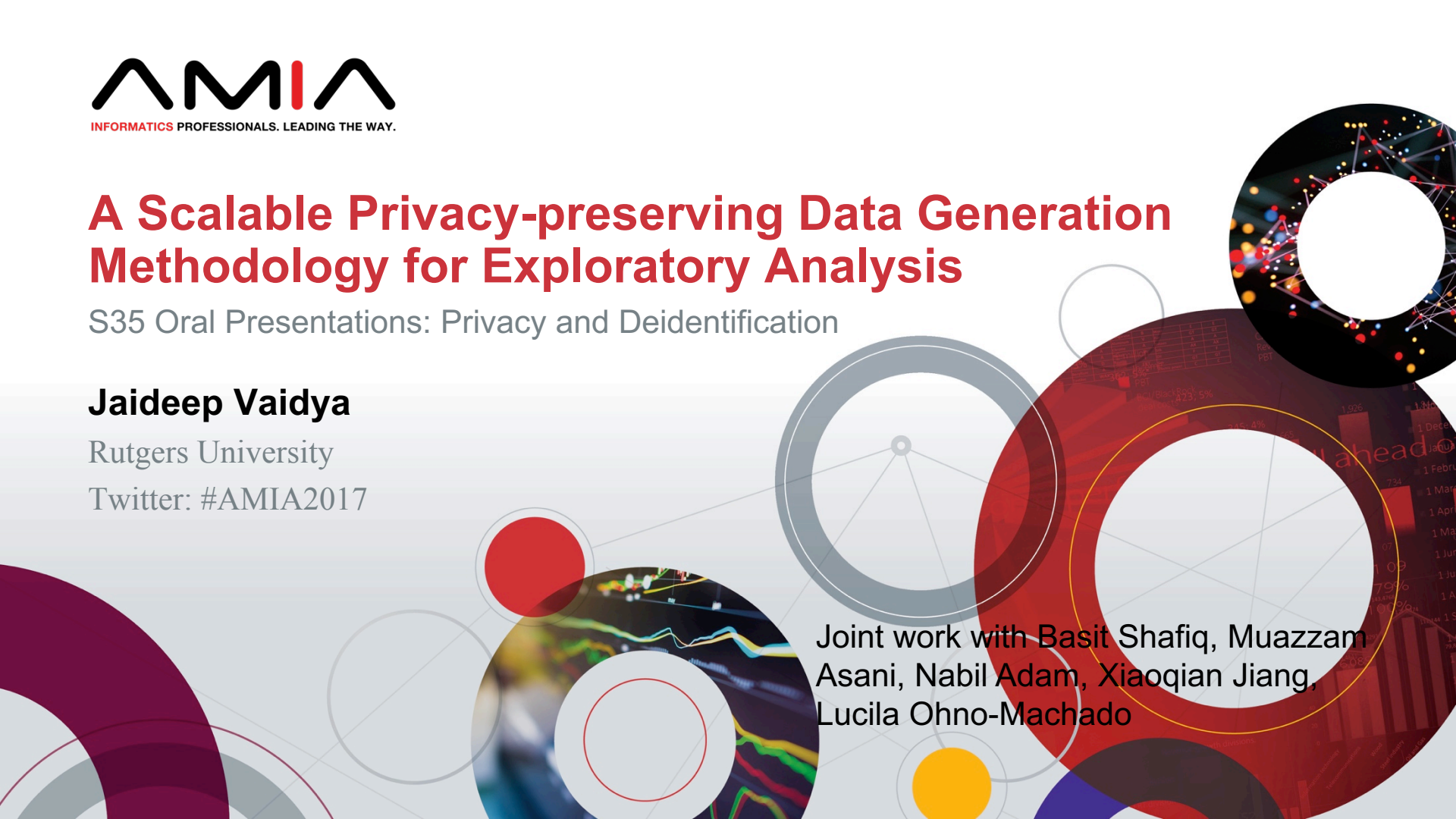
A Scalable Privacy-preserving Data Generation Methodology for Exploratory Analysis

S35 Oral Presentations: Privacy and Deidentification

Jaideep Vaidya

Rutgers University

Twitter: #AMIA2017



Joint work with Basit Shafiq, Muazzam Asani, Nabil Adam, Xiaoqian Jiang, Lucila Ohno-Machado

Disclosure and Acknowledgments

I and my spouse/partner have no relevant relationships with commercial interests to disclose.

This work was supported by the National Institutes of Health under awards U54HL108460, R01GM118574, R01HG008802, R01GM114612, R21LM012060, R01GM118609, U01EB023685, by the National Science Foundation under awards CNS-1422501 and CNS-1564034 and by the Patient-Centered Outcomes Research Institute (PCORI) under Contract CDRN-1306-04819. The work of Shafiq is supported by Pakistan's Higher Education Commission's NRPU grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research.

Learning Objectives

After participating in this session the learner should be better able to:

- Formulate an approach to generating synthetic data for exploratory analysis of biomedical data in a privacy-preserving manner.

Information is the basis of healthcare

Vast amounts of medical information available

- Genomic
- Transcriptomic
- Clinical
- Behavioral
- Social

In different modalities and forms

- EHR data
- Text
- Images
- Audio

Proper analysis can lead to breakthroughs in healthcare

The promise of big data

Due to decreased storage costs, and decreased computation costs

- Process data to realize value

Big Data has the potential to revolutionize

- Innovation, Competition, and Productivity

In the context of Healthcare

- Could create more than \$300 billion in value every year.
 - reducing healthcare expenditure by about 8 percent
- Improve healthcare efficacy
 - Enable comparative effectiveness research, reducing undertreatment and overtreatment
 - Better clinical trial design
 - Personalized medicine

The peril of big data

Volume, Velocity, Variety, and Veracity

Threat to Privacy and Security

- Individual privacy
- Competitive advantage
- Violating regulatory or confidentiality laws

The fear of big data

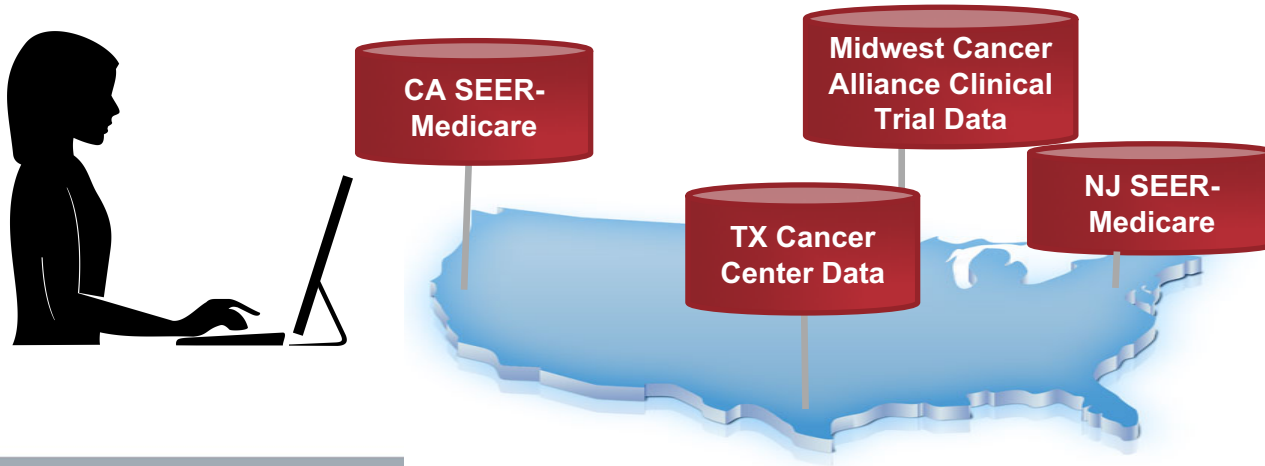
- No privacy within 5 years!

Result: Data is strictly controlled (as it should be) and often is stuck in silos (as it should not be)

Research Problem Addressed

Exploratory analysis?

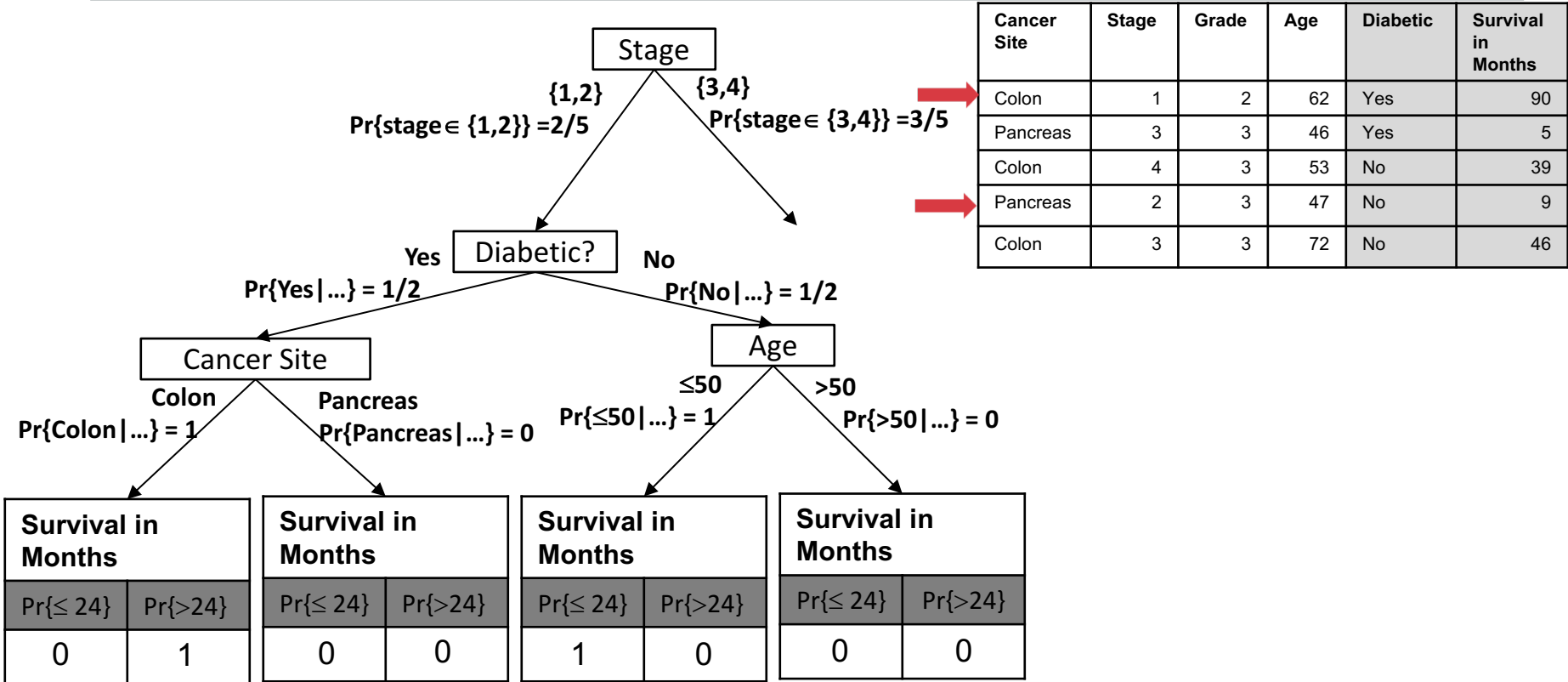
- A medical researcher Alice is interested in exploring the relationship between Vitamin D levels and cancer diagnosis at a national level
 - Specific regions where the strength of such relationship is higher/lower than national average



Research Objective

How can we generate sample datasets that preserve the structure and semantics of the original data, but not exact values, thus preserving its privacy, while providing a first-order approximation of utility

Random Decision Trees



Why Random Decision Trees?

RDTs have been shown to be an efficient implementation of the Bayes Optimal Classifier

- Empirically also shown to have good utility on various datasets

RDTs are also extremely well suited from the privacy perspective

- Randomness in structure is effective in protecting the base data
- Can be easily adapted to the differential privacy model

RDT outperforms other models in terms of computational speed

Generate multiple parameterized RDTs using the original dataset.

- Original dataset (D) can contain both nominal and numeric attributes
- Numeric attributes are discretized appropriately
- K RDTs are built in the standard way (choosing structure randomly)
- Compute conditional probability of visiting any node n , assuming that we are currently visiting the node's parent

$$pr(n) = \frac{\#instances\ in\ D\ reaching\ node\ n}{\#instances\ in\ D\ reaching\ node\ parent(n)}$$

Perform random walk over the RDTs to recreate instances

- randomly choose a RDT and perform a random walk from its root to fill in the instance value until the instance is completely generated.

Experimental Evaluation

Three real datasets used from the UCI Machine Learning Repository

- the Breast Cancer Wisconsin (Original) Data Set
- the Parkinsons Telemonitoring Data Set
- the Diabetes 130-US hospitals for years 1999-2008 Data Set

Name	# instances	# Attributes	RDT Depth	# Response Variables	Task
Breast Cancer	683	10	5 [10/2]	1	Classification
Parkinson's Telemonitoring	3178	20	5 [20/4]	2	Regression
Diabetes	98042	37	6 [37/7]	1	Classification

Note: 10-fold cross validation carried out (i.e., synthetic data built from data in 9 folds and accuracy computed for data in 10th fold. Accuracy reported is the average accuracy over all 10 iterations.)

Results for Classification

Data Set	# Classes	AUC with Original Data	AUC with Synthetic Data (no oversampling)	AUC with Synthetic Data (10-fold oversampling)	AUC with 100,000 synthetic instances
Breast Cancer	2	0.9945	0.9937	0.9933	0.9938
Diabetes	3	0.6534	0.6134	-	-

In general, the model generated from synthetic data achieves almost the same accuracy as the model generated from the original data in terms of the AUC.

Increasing the degree of oversampling tends to improve the results.

Though the accuracy for Diabetes is low with synthetic data, it is also quite low for the original data.

Results for Regression

Data Set	Response Variable	RMSE with Original Data	RMSE with Synthetic Data (no oversampling)	RMSE with Synthetic Data (10-fold oversampling)	RMSE with 100,000 synthetic instances
Parkinsons	motor_UPDRS	6.51	7.05	7.04	7.03
Parkinsons	total_UPDRS	8.14	8.93	8.86	8.87

Performance of linear regression is slightly worse with the synthetic data but is still very comparable

Significant overlap in the variables identified as significant in the regression model built from the training data and that built from the synthetic data, though the p-values varied.

Key Takeaway

From the perspective of exploratory analysis, the models generated from synthetic data do give a similar view of the data as compared to the models generated from the original data.

The synthetic data generation process is extremely efficient – generating 100,000 instances takes only a few minutes, though the process for building the RDTs is memory intensive.

We have developed an approach for generating synthetic data using RDTs that can be used for exploratory analysis

Demonstrated that it can provide a first order approximation of utility on some datasets

Future work

- Provide more accurate estimates of utility for specific data analysis tasks
- Extending the approach to longitudinal data
- Working on using it with real data in real clinical applications
- Integrating the approach into REDCap

Question 1

Consider a researcher Alice who is interested in exploring the relationship between Vitamin D levels and cancer diagnosis. Alice is aware of four different datasets which have been collected at institutions in different geographic regions in the country. In order to establish the relevance of each dataset to her study, which of the following options should Alice follow?

Answer Option

- A. Alice should wait till she has access to the entire real data from each of the four different sources to evaluate whether each dataset will be useful or not.
- B. Alice should use synthetic data generation techniques to generate a privacy-preserving variant of the required dataset by herself, which she can then use
- C. If synthetic data has been appropriately generated by the respective sites, then Alice should use the synthetic data to estimate the utility of the real data and then go through the process to obtain the real data if required.
- D. If synthetic data has been appropriately generated by the respective sites, then Alice should use the synthetic data as a proxy for the real data to carry out her study.

- A. Alice should wait till she has access to the entire real data from each of the four different sources to evaluate whether each dataset will be useful or not.
 - Incurs significant time and cost overhead with no guarantee of relevance
- B. Alice should use synthetic data generation techniques to generate a privacy-preserving variant of the required dataset by herself, which she can then use
 - She can't do this by herself without access to the data
- C. If synthetic data has been appropriately generated by the respective sites, then Alice should use the synthetic data to estimate the utility of the real data and then go through the process to obtain the real data if required.
 - If appropriate synthetic data has been created by the respective sites, it can indeed be used by Alice to get an estimate of the utility with respect to her study.
- D. If synthetic data has been appropriately generated by the respective sites, then Alice should use the synthetic data as a proxy for the real data to carry out her study.
 - Since the synthetic data is only an approximation of the real data, therefore it should not be directly used as a proxy for the real data to carry out her study.


Question 2

Consider an institution that collects/owns two different datasets A and B. A is longitudinal data (i.e., patient records with the time of patient visit, which enables finding relationships across time). For example, consider data similar to that collected by Medicare which includes multiple records for each patient, along with the time of service/visit. On the other hand, B is not longitudinal and contains only a single record for each patient (for example, SEER data), so relationships across visits cannot be found. The institution is considering making both datasets accessible to other researchers in some form that protects privacy. The CIO learns of the Random Decision Tree based synthetic data generation approach and is interested in potentially using it to enable other researchers to perform exploratory data analytics. After going through the approach, which of the following is the most likely action the CIO takes?

Answer

- A. Since no Associations are preserved, the generated synthetic data is useless. Therefore, the institution does not use the random decision tree based synthetic data generation approach for either dataset.
 - Associations across records are maintained to some extent
- B. Since associations are preserved to some extent for each attribute across time, the institution can use the random decision tree based synthetic data generation approach for the longitudinal data.
 - RDT approach does not maintain associations across records for the same patient over different time instants, so it cannot be used for longitudinal data
- C. Since associations are preserved to some extent across attributes, though not across time, the institution can use the random decision tree based synthetic data generation approach for the non-longitudinal data.
 - Associations are preserved to some extent across attributes, therefore it can be used for the non-longitudinal data for exploratory analytics.
- D. Since all associations are preserved perfectly, both across attributes and time, the institution uses the random decision tree based synthetic data generation approach for both datasets.
 - RDT approach does not maintain associations across records for the same patient over different time instants, so it cannot be used for longitudinal data

Questions?



AMIA is the professional home for more than 5,400 informatics professionals, representing frontline clinicians, researchers, public health experts and educators who bring meaning to data, manage information and generate new knowledge across the research and healthcare enterprise.



f @AMIAInformatics

🐦 @AMIAinformatics

in Official Group of AMIA

📺 @AMIAInformatics

#WhyInformatics

Thank you!

Email me at:

jsvaidya@business.rutgers.edu

