Does Metadata leak Privacy?

Danning Zhan, Rihan Hai

TU Delft

Motivation

01

Data Sources

Tabular data
Distributed sources

02

Data Augmentation

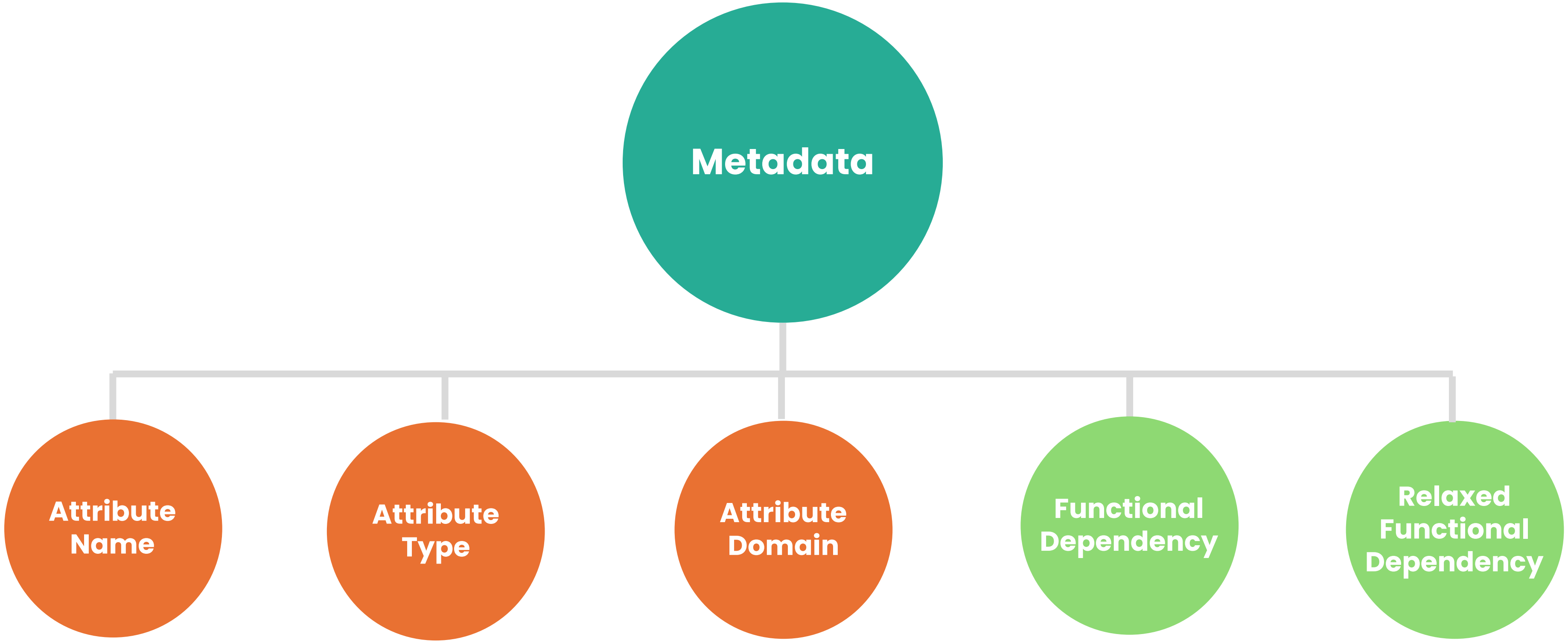
Sources will augment data attributes
Validate the data augmentation

03

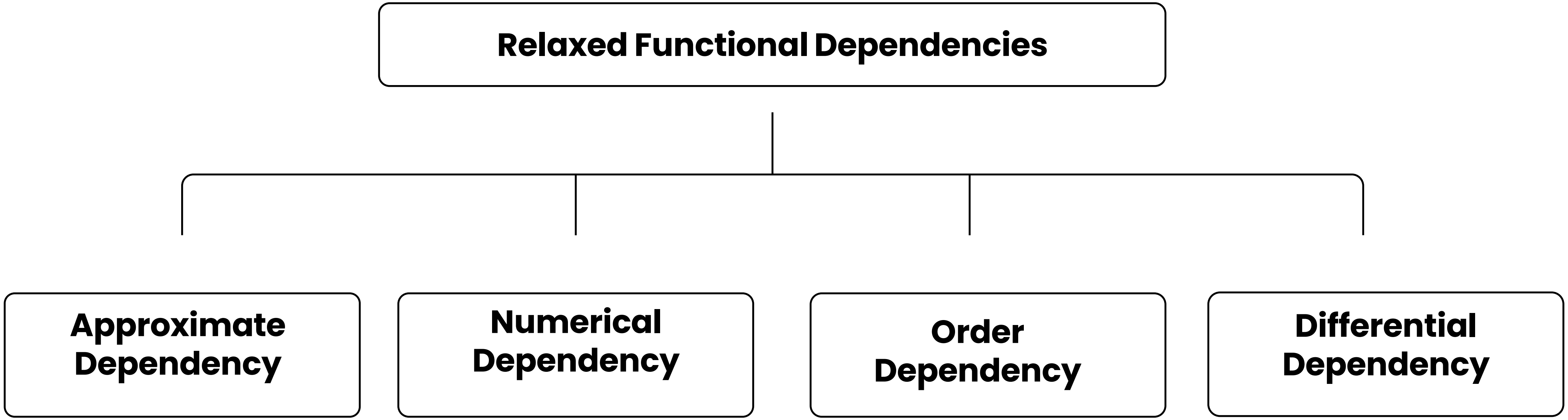
Federated Learning

Downstream Analytical model.

Tabular Data Metadata



Tabular Data Dependencies



Relaxed Functional Dependencies

Approximate Dependency

Numerical Dependency

Order Dependency

Differential Dependency

Privacy

Motivation

GDPR

- Recent Regulations such as GDPR
    - Data storage and data processing

Formal Definition

Privacy Leakage definition

- Data point identifiability
- Data value Inference

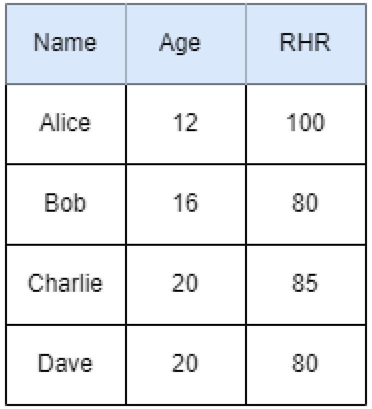
Data Types

Data Types

- Categorical
    - Exact Matching
- Continuous
    - Distance Metric

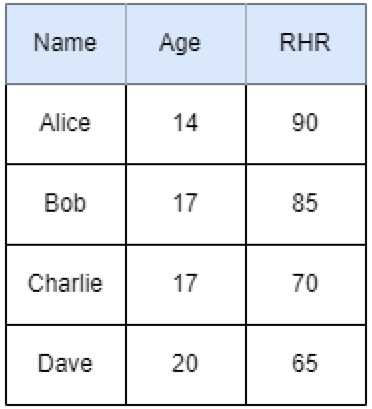
Privacy Analysis

- From Discovery - Annotations
- Both tables satisfy the same metadata
  - Attribute names
  - Attribute domains
  - Functional Dependencies
  - Relaxed Functional Dependencies
- Probabilistic
  - The dependencies
  - The values



Name	Age	RHR
Alice	12	100
Bob	16	80
Charlie	20	85
Dave	20	80

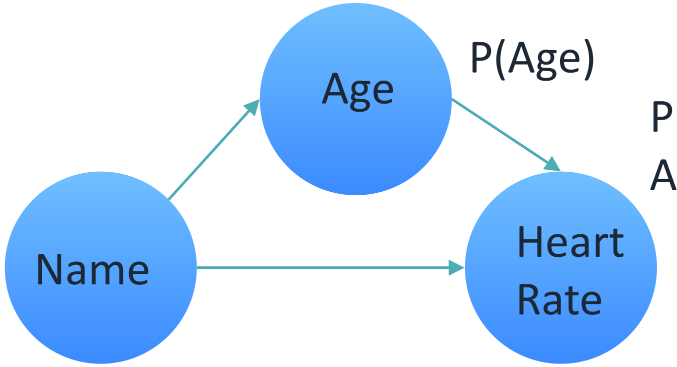
Original



Name	Age	RHR
Alice	14	90
Bob	17	85
Charlie	17	70
Dave	20	65

Copy

{ Name → Age , Age → Heart Rate, Name → Heart Rate}



P(Age)

P(Mapping) = Age → Heart Rate

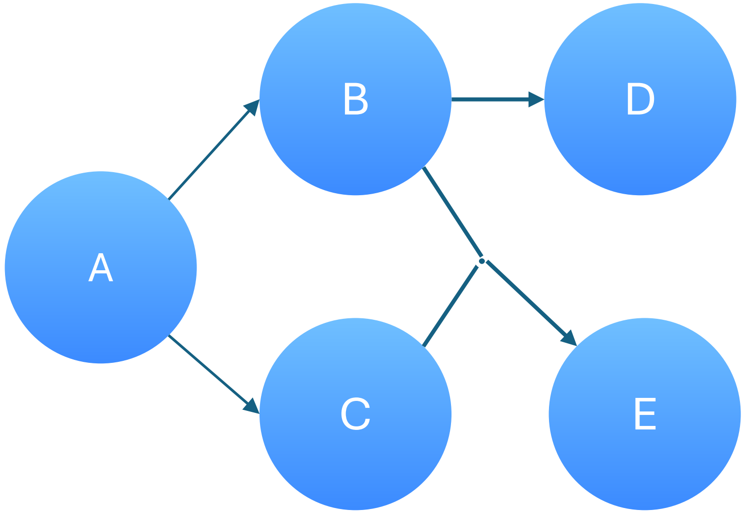
P(Heart Rate) = P(Age) * P(Mapping)

Privacy Analysis

Approximate Functional Dependency – ε

- Dependencies:
  - Directed Graph Traversal
  - Generated from the graph
    - Mapping: A ➔ B, A ➔ C, B➔D, (B, C) ➔ E
- Values
  - Generated from assumed distribution
  - Consistency with Dependencies
  - Matching follows a Binomial Distribution

{ A ➔ B , A ➔ C, B ➔ D, (C,B) ➔ E}



Experiment

Number of Exact Matches for Categorical Attributes

Dep	Attr 1	Attr 3	Attr 11	Attr 12
Rand Gen	44	44	33	44
Func Dep	44.082	43.954	32.815	NA
Ord Dep	44	32	29	47
Num Dep	56	NA	NA	NA

Metric Evaluation of Continuous Variables

Dep	Attr 0	Attr 2	Attr 4	Attr 5	Attr 6	Attr 7	Attr 8	Attr 9
Rand Gen	580.49	1169.96	0.43	114.17	10.14	138.69	1.71	0.93
Func Dep	580.25	1172.4	0.43	114	10.11	138.6	1.71	NA
Ord Dep	581.43	1383.86	0.24	17.33	9.63	139.44	1	1.41
Num Dep	708.58	NA	NA	NA	NA	NA	NA	NA

Conclusion and Takeaway

Privacy

Non-zero probability of leakage

Agnostic to assumed distribution

Metadata

Properties

Non-unique

Implicit to data

Metadata Dependency

Dependency Between metadata

Data Source → Attribute Names

Attribute names → Attribute Domain

Attribute Domain → Functional Dependencies

Future Works

Other metadata

Will communicating more metadata lead to more privacy leakage?

Validate Data

Can we validate the data without communicating metadata?

TUNE IN TO THE REST OF OUR TEAM

Aditya:
SiloFuse: Cross-Silo Synthetic Data Generation with Latent Tabular Diffusion Models.

Tuesday, May 14th 16:21 – 18:00
Theatre 12