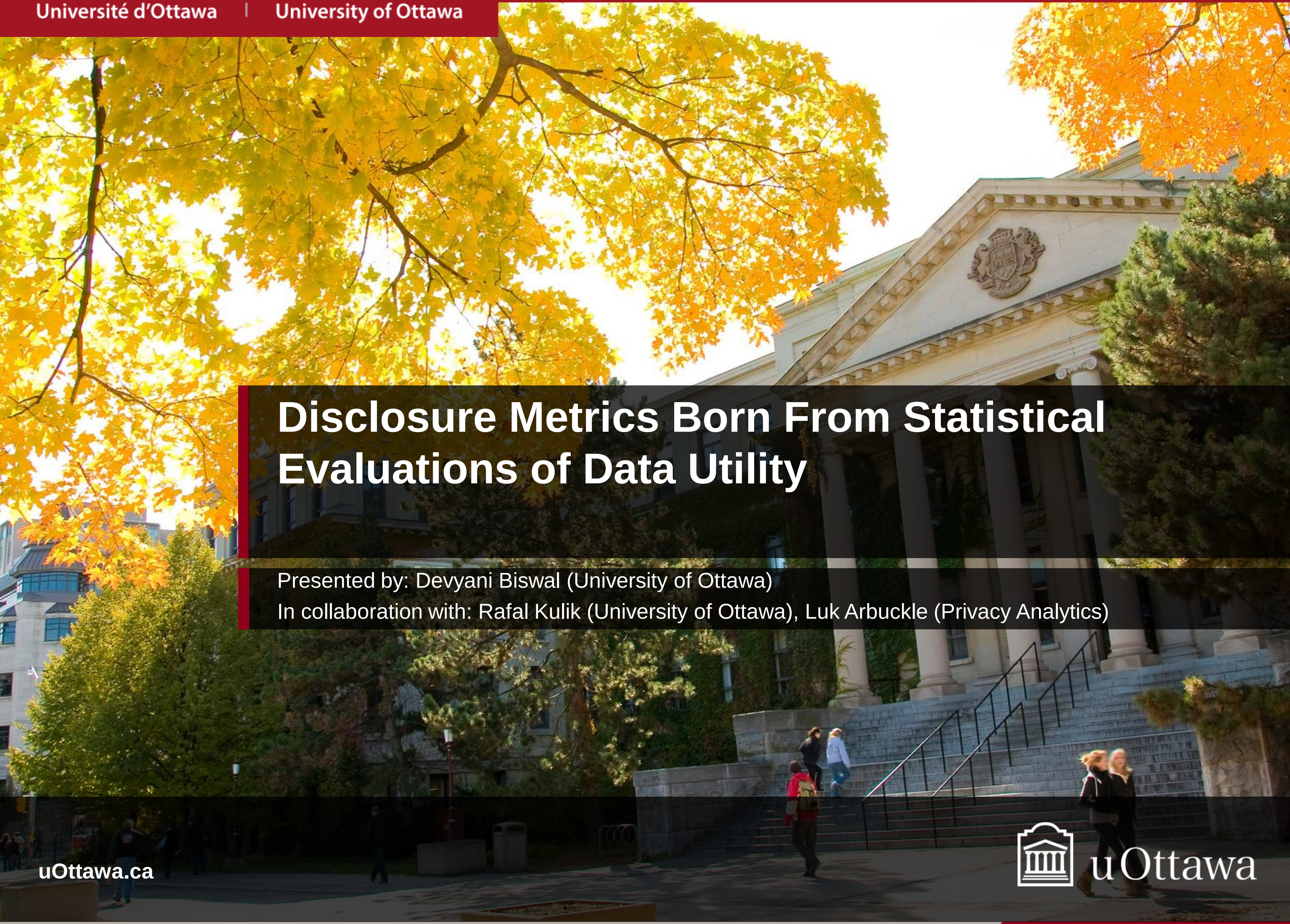
Université d'Ottawa | University of Ottawa



Disclosure Metrics Born From Statistical Evaluations of Data Utility

Presented by: Devyani Biswal (University of Ottawa)
In collaboration with: Rafal Kulik (University of Ottawa), Luk Arbuckle (Privacy Analytics)

uOttawa.ca

uOttawa

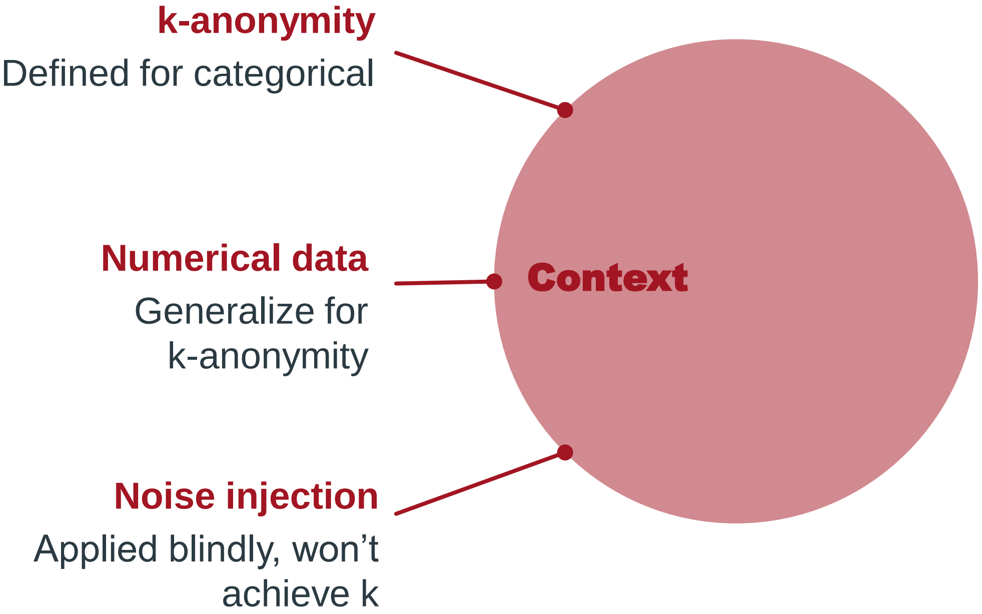
Agenda

- Background
- Privacy Models
- Empirical Results
- Ongoing Research

Context and Motivation

BACKGROUND

Context



k-anonymity

Defined for categorical

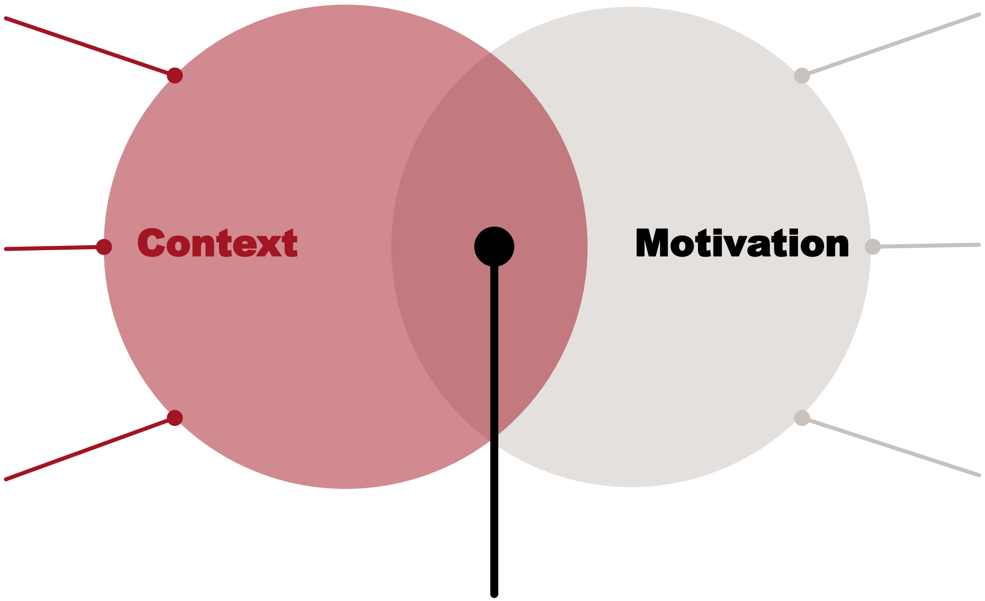
Numerical data

Generalize for k-anonymity

Noise injection

Applied blindly, won't achieve k

Context and Motivation



k-anonymity
Defined for categorical

Numerical data
Generalize for k-anonymity

Noise injection
Applied blindly, won't achieve k

Data utility
Improve over existing

Privacy bound
Use the same k

Recipient
Produce meaningful statistics

Research

Theoretical

PRIVACY MODELS

Categorical Data

k-PRAM

Input: $\underline{X} = (X_1, \dots, X_n)$

Output: $\underline{X}^r = (X_1^r, \dots, X_n^r)$

$$\left.\begin{array}{c} X_{(1)} \\ \vdots \\ X_{(n)} \end{array}\right\} \rightarrow \left.\begin{array}{l} G_1 = [X_{(1)}, X_{(i)}] \rightarrow G_1^r = [X_{(1)}^r, X_{(i)}^r] \\ \vdots \\ G_m = [X_{(j)}, X_{(n)}] \rightarrow G_m^r = [X_{(j)}^r, X_{(n)}^r] \end{array}\right\} \rightarrow \begin{array}{c} X_1^r \\ \vdots \\ X_n^r \end{array}$$

$X_{(i)} \longrightarrow X_{(i)}^r$ with uniform probability.

Numerical Data

k-Noise

Input: $\underline{X} = (X_1, \dots, X_n)$

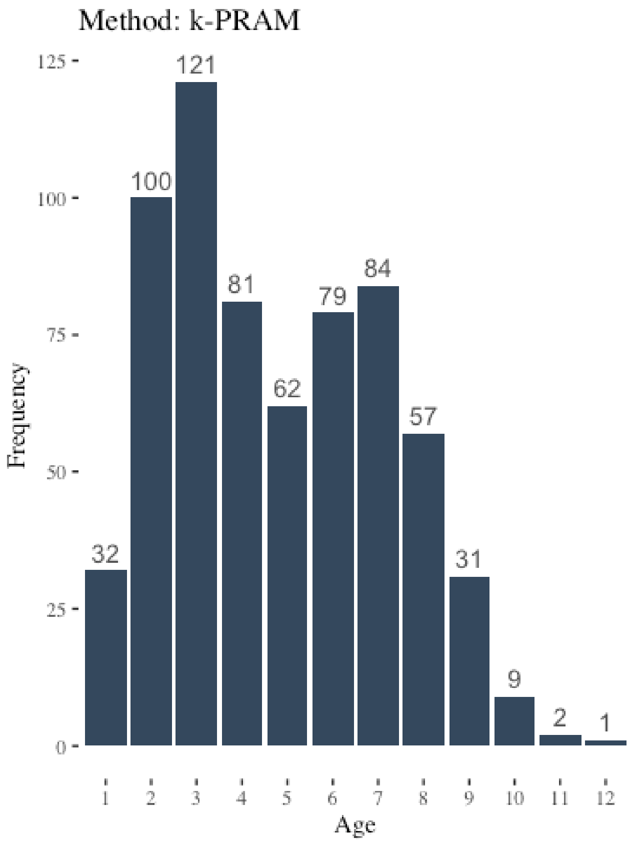
Output: $\underline{X}^r = (X_1^r, \dots, X_n^r)$

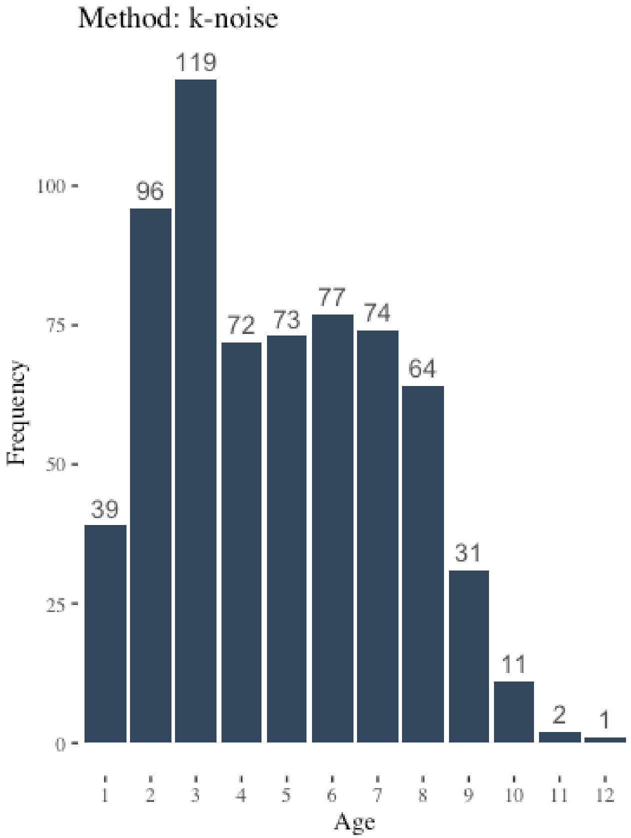
$$\begin{array}{ccccc}
\hline
X_1 & & X_1 + \mathrm{Uni}(-a,a) & & X_1^r \\
\vdots & \longrightarrow & \vdots & \longrightarrow & \vdots \\
X_n & & X_n + \mathrm{Uni}(-a,a) & & X_n^r
\end{array}$$

Experiments

EMPIRICAL RESULTS

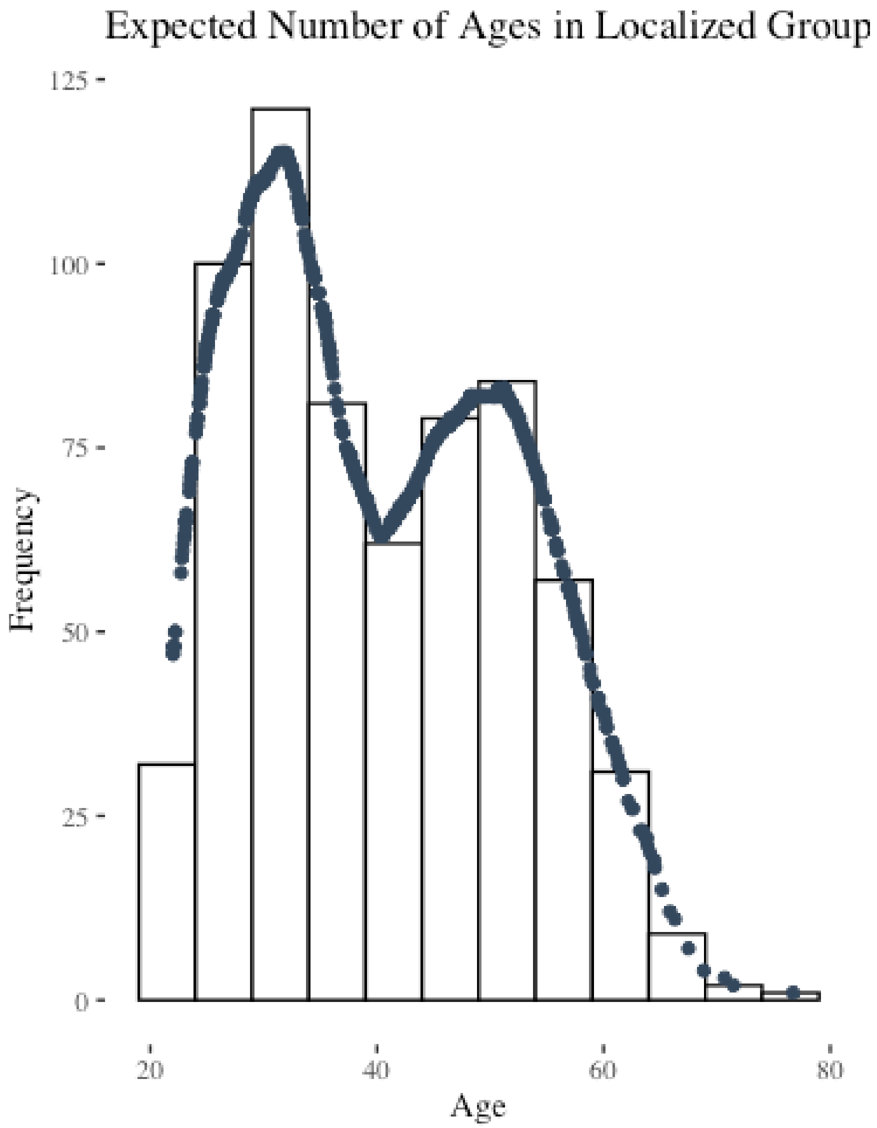
Baseline Level of Privacy



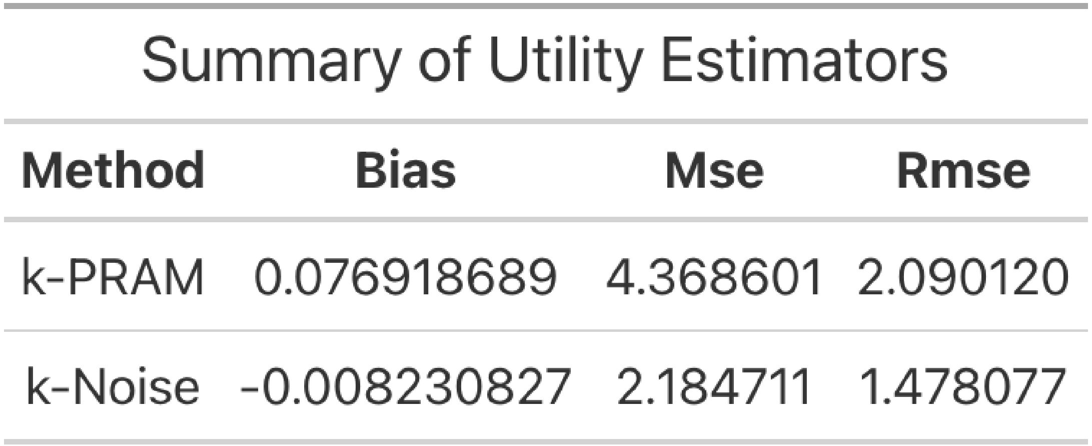


Baseline Level of Privacy

Expected number or records within a neighbourhood of [-2.5,2.5] years of each randomized record in X.



Data Utility Measures



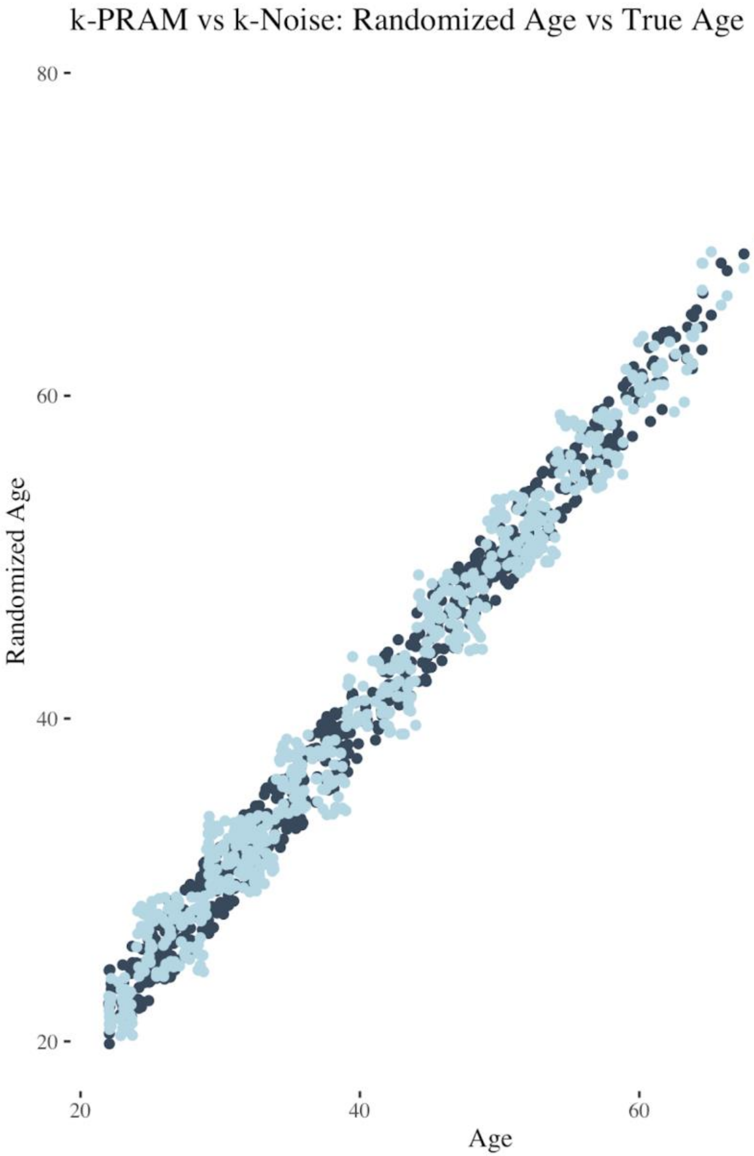
Summary of Utility Estimators

Method	Bias	Mse	Rmse
k-PRAM	0.076918689	4.368601	2.090120
k-Noise	-0.008230827	2.184711	1.478077

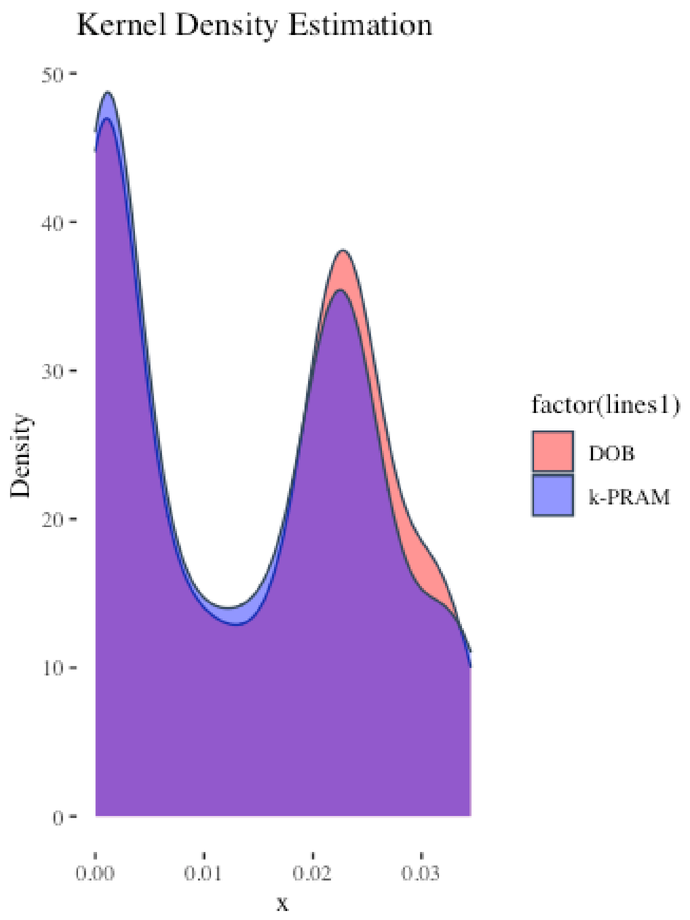
Data Utility Visual

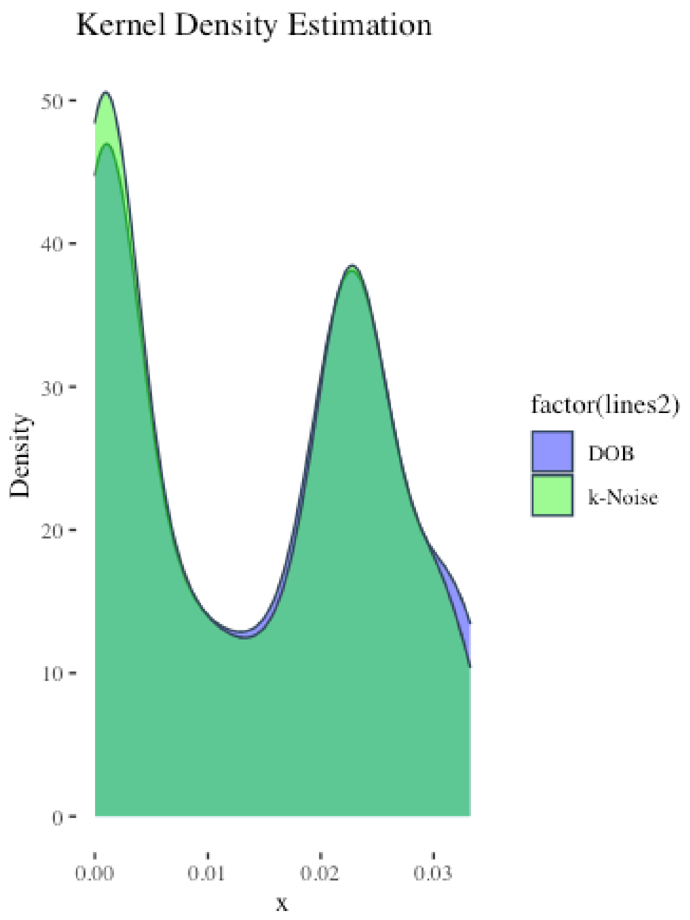
The clustered scatter plot represents k-PRAM randomized individuals.

The bias calculated is visibly noticeable when comparing the two methods against the true value.

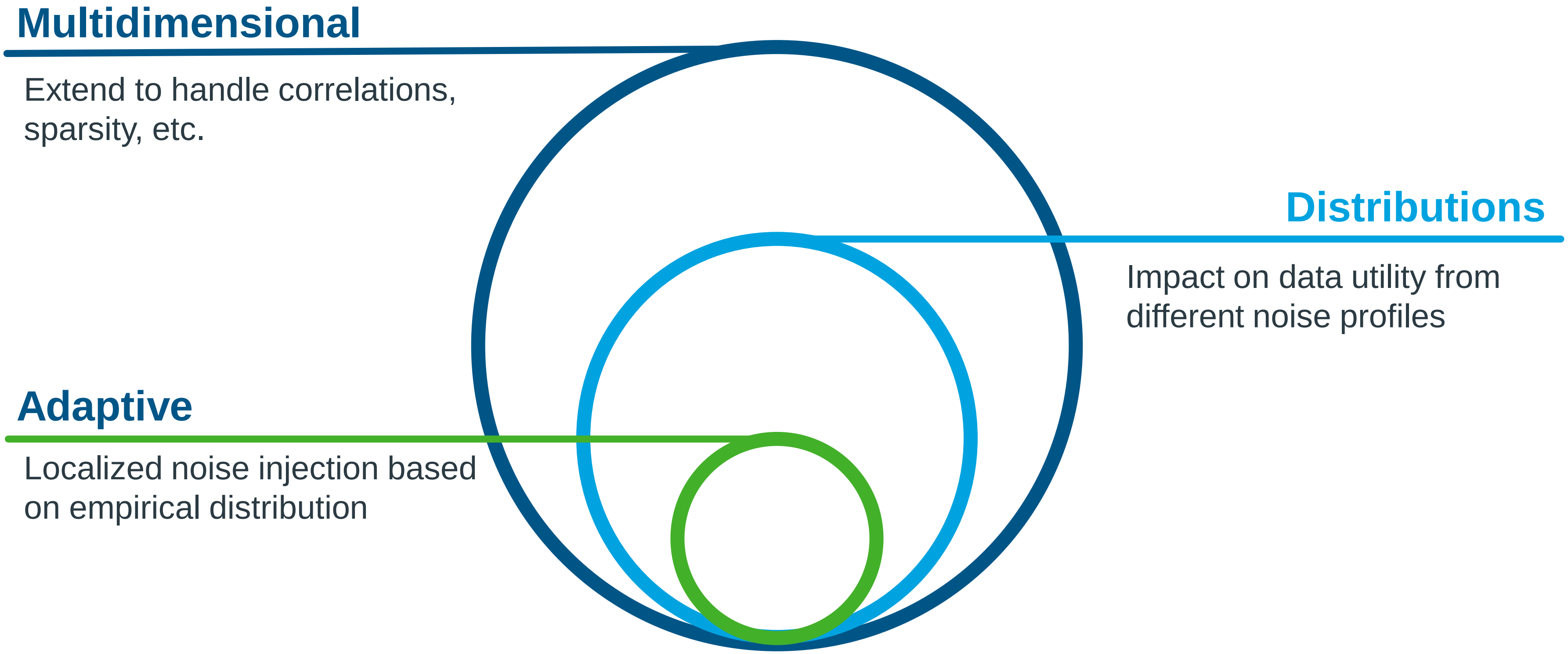

k-PRAM vs k-Noise: Randomized Age vs True Age

Data Utility Densities





Ongoing research



Multidimensional

Extend to handle correlations, sparsity, etc.

Distributions

Impact on data utility from different noise profiles

Adaptive

Localized noise injection based on empirical distribution