Safeguard Customer Privacy Without Sacrificing Analytical Capability with Diffprivlib

Team 4
Laura Cattaneo, Tzu-Hsuan Lin, Harshitha Kuriminisetty, Casey Easterday

University of Minnesota
Driven to Discover℠

Anonymization is not sufficient for customer privacy

Almost 90 percent of the US population has a unique combination of 5-digit zip code, gender, and date of birth*

New York taxi details can be extracted from anonymised data, researchers say

FoI request reveals data on 173m individual trips in US city - but could yield more details, such as drivers' addresses and income

The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now

Researchers reverse Netflix anonymization

Robert Lemos, SecurityFocus 2007-12-04

The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

*https://dataprivacylab.org/projects/identifiability/paper1.pdf

Storing data imposes financial, reputation, and regulation risks

$8.64M

Average cost of a data breach in US‡

280

Average days to identify and address a breach‡

$$$?

Company reputation is not quantifiable

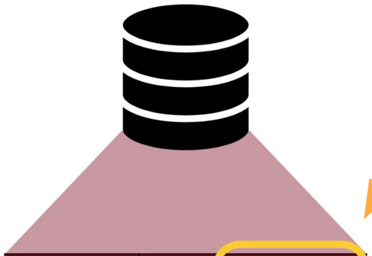
29

States enacted, enforced, or considered privacy legislation in 2019*

*https://gcn.com/articles/2019/07/31/state-privacy-laws.aspx

‡https://www.ibm.com/security/data-breach

Diffprivlib mitigates risks of using data for competitive advantage



Name	Age	Age (Privatized)
John Smith	46	48
Steve Roberts	35	31
…	…	…
Mary Williams	72	74
Average	**51**	**51**

Diffprivlib **adds noise** to confidential data while **maintaining statistical integrity**

Users can conduct **accurate analyses** and build **valid models** with privatized data

Gain population or customer **insights** while **safeguarding privacy**

Diffprivlib is the product of many years of differential privacy research

Differential privacy has been **researched and evolving since early 2000**

Offers **mathematically provable guarantee of privacy** against: differencing attack, linkage attack, reconstruction attack*

*Dwork & Roth, 2014

Why it Works

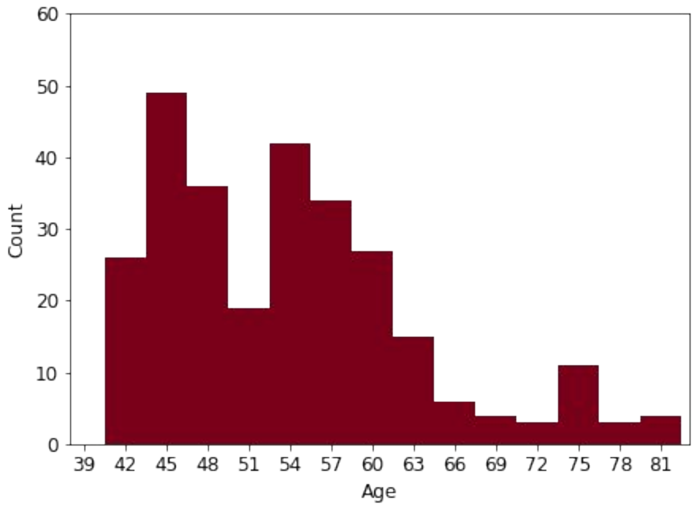
Noise added to each value is **different for each record**

Name	Age	Age with noise(Privatized)
Guy Gilbert	49	52
Kevin Brown	44	43
....	.....	.....
Rachel Valdez	55	55
Lynn Tsoflias	60	61
Average Age	**54**	**54**

Non-random noise **creates combinations of customer data** that do not exist and are **untraceable**

Despite added noise, analytical capability is preserved

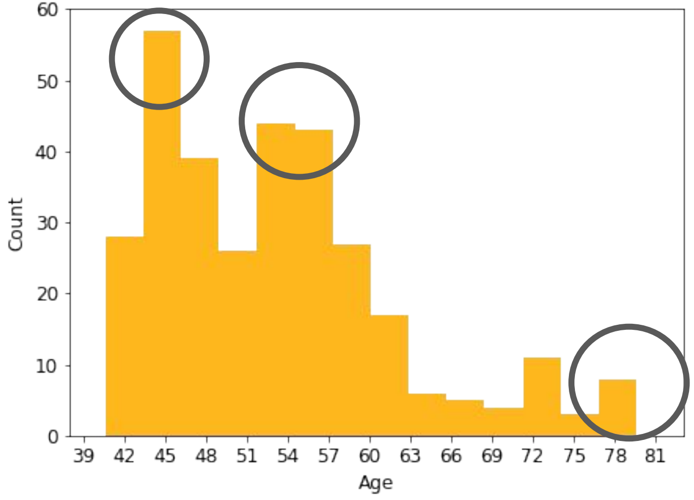
Age Distribution **(Original Dataset)**



Mean: 52.2981

Standard deviation: 8.9533

Age Distribution **(Masked Dataset)**

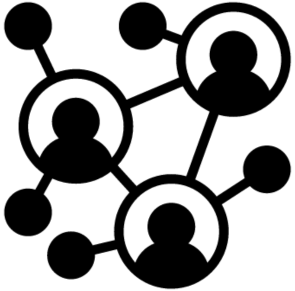


Mean: 52.4317

Standard deviation: 9.0723

Diffprivlib can help firms acquire competitive edge

Reduce risk of sharing sensitive data externally

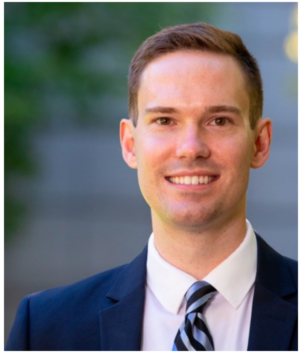


Expand pool of potential **partners**

kaggle NUMERAI

Crowd-source analytics solutions

Questions? Feel free to contact us

Laura Cattaneo
catta008@umn.edu

Casey Easterday
easte060@umn.edu

Harshitha Kuriminisetty
kurim006@umn.edu

Tzu-Hsuan Lin
lin00491@umn.edu

For more technical details, check out our [GitHub]