Data Privacy Detective Podcast – Episode 57: Protecting Data Privacy Within Databases

Dec 04, 2020

Categories:

Blogs
Podcast
Technology Blog

Authors:

Joseph J. Dehner

Respecting individual privacy while achieving reliable databases.

We all value privacy – at least to some extent. But some of us want to be famous, and all of us want to connect with friends and acquaintances. We like the convenience from technology that requires our personal information to operate. So we share our personal details in many ways, and our data flows like water down a stream into lakes and oceans, some of which we'd prefer to avoid. And our information becomes a piece of society's knowledge base. Databases like the U.S. Census have essential purposes, but they're only reliable and complete if we are comfortable sharing our data. How to respect individual privacy and achieve reliable databases? That's a challenge!

In this podcast, Alex Watson, co-founder and CEO of Gretel.ai, explains two essential phrases to understand how this can be done. Alex founded a security startup called Harvest.ai, which was acquired by Amazon Web Services in 2016 when he became AWS General Manager and it launched its first customer-facing security offering. Gretel.ai is an early-stage startup that offers tools to help developers safely share and collaborate with sensitive data in real-time.

Alex explains that privacy is a problem rooted in code, not in compliance. By **auto-anonymization**, the personal data of an individual is separated from the underlying data so that the database where the information is needed comes to it without identifying the individual. The essential information is shared without allowing someone to know which individual's information it is. While nothing is hack-proof, auto-anonymization eliminates the link between an individual and data about that individual as it moves to another user. Personal privacy is preserved in the transmission and further use.

The other key phrase to understand is **differentially private synthetic data**. Data Privacy Detective Podcast 55 offers an introduction to the topic. This phrase means that information within a database has been changed to eliminate the ability to trace back the data to a particular individual. The information is private and individual to a person, but as pieces of data are shared for a purpose, they are not traceable to a specific person. The database user only needs the provided information, not the identity of individuals who contributed to each piece.

Let's consider two use cases.

Uber has scooters in some cities available for short-term use. The user provides information that lets Uber bill for the usage (Uber obviously needs to know who is using and paying). When this service was launched, it allowed third parties to determine who was using each scooter and where the person was or had been. By using AI to separate the information about a scooter's location from a particular user's identity, Uber could share the information with a city for traffic and other urban planning purposes, while assuring users that their personal identity was not being shared with third parties that had no reason (or business) knowing the user's personal identity. Protocol story about Uber example

A second example concerns health data. A medical database was used to develop recommendations about heart disease. But when the database was analyzed to ensure that it was reliably useful, it was found that 68% of the individuals within the database were male. This wrongly skewed the database against females, who have different heart disease risks and profiles from men. A database weighted two thirds to one sex was not properly useful to provide gender-neutral conclusions. In this case, an individual's identity was not essential to compile a database of equally representative individuals by gender. Gretel post about UCI heart disease example on Kaggle

Listen to episode 57 below.

There is great public benefit in encouraging people to share sensitive data – e.g., public health databases, sociological research,

Census Bureau studies. But people will share their private data only if they are comfortable knowing it will not be misused. Database users should ensure that they do not acquire personal data that identifies individuals without the need to have that information.

Auto-anonymization and differentially private synthetic data – two phrases one should know. Their proper usage can achieve privacy by design. This will be an important contribution to creating reliable databases humankind needs to advance public health and other social good.

If you have ideas for more interviews or stories, please email info@thedataprivacydetective.com.

FBT Gibbons