LEAPYEAR

BLOG REPRINT

Differential Privacy from Theory to Practice

INTRODUCTION

Emerging technologies are creating unprecedented value from data. The demand for accessing and sharing information is growing, as cloud computing, AI, and machine learning enable new value to be derived from enterprise data.

However, privacy, security, and regulatory concerns are severely restricting access to sensitive information, thus limiting the full potential impact of data. The same methods from machine learning that create value from data also pose risks to privacy—as sensitive datasets are exposed to data scientists, third parties, and models, the surface area for data exfiltration is rapidly growing.

Traditional methods to protect data such as data masking, aggregation, and redaction have time and time again failed to solve this problem. These are heuristic-based approaches—they provide no provable guarantees—and have been repeatedly reverse engineered to compromise highly sensitive personal data including medical and financial information as well as confidential business data and IP.

Restricting access to data outright is not the solution either—similar to the approaches above, strict controls limit data value, restrict the pace of innovation, and do not provide meaningful defense against privacy attacks.

The massive value from mining and sharing data sets, paired with the sophistication of modern privacy attacks, has created an imminent problem for data scientists, security officers, IT organizations, data engineers, compliance officers, business stakeholders, and regulators. There is an urgent need for a new paradigm to safely unlock value from sensitive information for the enterprise.

DIFFERENTIAL PRIVACY

For half a century, the field of cryptography has facilitated the secure sharing of electronic information—credit card transactions, file transfers, and communications—protected by cryptographic protocols. This has been possible through rigorous, mathematically proven guarantees. The assurances provided by the protocols to both individuals and businesses are backed by unassailable mathematical proof and the protocols have therefore stood the test of time.

The field of data privacy, however, has historically been devoid of such foundations. The only known approaches for exposing data for analysis while still attempting to maintain confidentiality have been heuristic based: in other words, guesswork. These approaches involved redacting certain fields that were considered especially sensitive, such as PII, or rules, such as “reveal the age of a person but not their birthdate” or “only show a result if there are more than 20 people in the sample.”

It is a well-known fact that these techniques do not work; countless studies and real-world breaches have demonstrated that these approaches can be reverse engineered. Every time an approach is breached,

privacy practitioners propose a new, slightly more sophisticated technique, such as adding noise to release "synthetic data," performing aggregations such as "k-anonymity," or instituting a broad class of statistical techniques called "data masking." Ultimately, each technique has been broken, resulting in a compromise of highly sensitive information and harming both individuals and institutions. In the absence of an alternative, these approaches continue to be used in the enterprise today, perpetuating the related inefficiencies, loss of data value, and vulnerabilities.

Differential privacy emerged 15 years ago as a solution to this problem. Differential privacy is a rigorous, mathematical definition of data privacy. When a statistic, algorithm, or analytical procedure meets the standard of differential privacy, it means that no individual record significantly impacts the output of that analysis—the mathematics behind differential privacy ensures that the output does not contain information that can be used to draw conclusions about any record in the underlying data set. With differential privacy, there is a mathematical bound on the amount of information released by the system, also known as an information theoretic guarantee. The beauty of the approach is that the sophistication level of an adversary, the amount of computational resources they use, or how much outside information they have access to does not matter; when differential privacy is satisfied, the information required to draw conclusions about individuals' private information is just not sufficient.

IMPLEMENTING DIFFERENTIAL PRIVACY

Despite the power of differential privacy as a solution to the global privacy problem, differential privacy has not been implemented for the enterprise. While there are specialized implementations by Apple and Google for targeted use cases as well as several academic research projects, no commercial solutions emerged.

One reason a larger implementation has not emerged is that differential privacy is not an algorithm or technique—it is a mathematical definition of privacy. The definition pertains to analyses: an analysis is or is not differentially private; intuitively, if the outputs of the algorithm are insensitive to individual records, then the algorithm is differentially private, and if the outputs are sensitive to individual records, then the algorithm is not differentially private. The standard set of analyses in a data scientist's toolkit—counts, means, regressions, etc. are not inherently differentially private—has to be re-designed in a way that satisfies the definition, which is a hard mathematical problem.

The process of re-designing the algorithms usually involves introducing precisely calibrated variability into the computation itself to hide the contribution of any individual records datapoint with randomization. The challenge is to introduce variability in a way that satisfies the standard of differential privacy without compromising the analytical utility of the result. Over the past 15 years, thousands of papers have been written by theoretical computer scientists, statisticians, and mathematicians on the topic of differential privacy.

Most of these papers attempt to address how, for various analytical procedures, the definition of differential privacy can be satisfied; unfortunately, the academic literature is of mixed quality and implementability.

Some papers contain errors (for instance, they are not truly differentially private), and others just cannot be practically implemented (they are impossible to translate into working code or introduce too much error when run on real world data sets). In summary, differential privacy is a standard and is a very hard standard to meet.

To meet the standard of differential privacy for the enterprise, several components need to be built, each of which contains deep technical challenges in its own right.

The following are a few examples:

- A differentially private platform needs to support the complete range of analytical functions used by enterprise analysts and data scientists. These range from aggregates, statistical functions, and data operations to machine learning algorithms. Ensuring that the system contains correct and accurate implementations of these differentially private algorithms requires understanding the literature, identifying the best theoretical results, and mapping them to working, production software. This process contains many challenges in statistical analysis and machine learning.
- For a platform to be differentially private, every computation in the platform needs to be differentially private. In other words, there needs to be no way an analyst can run a computation that is not differentially private. This is a hard problem in secure system design—to ensure that there is no way to subvert the differentially private computations to exfiltrate the data.
- The platform needs to rigorously track composition. It is not sufficient for individual computations to be differentially private—the entire set of programs that are ever executed needs to be differentially private. Tracking composition both correctly and in a way that effectively allows for continued value from the data is a hard problem in information theory.
- To be successfully deployed in the real world, a differentially private platform needs to scale, in many cases, to petabyte size data sets. This requires solving hard problems in scalable and distributed computing.
- These are just a few examples of the broad set of challenges in implementing a commercial differential platform in a way that is both secure and usable. Developing a differential privacy platform for the enterprise requires technical talent across a diverse set of technical fields. Our focus at LeapYear over the years has been to not only assemble outstanding engineers in these relatively disparate areas but also align their efforts and innovations toward a unified platform.

The scope and breadth of differential privacy poses an additional challenge, which is creating the mapping from an abstract mathematical standard to delivering solutions to real-world business problems. Assembling the technical team and developing the core technology are only steps in the journey; implementing differential privacy requires extensive experience in vertical-specific enterprise data challenges.

It involves understanding the security concerns, regulatory frameworks, data sets, analytical workflows, and business objectives and then mapping the technology, particularly the mathematical nuances of differential

privacy, to deliver a broader solution. Without the data and business context, differential privacy may be incorrectly applied, thus compromising analytical utility and potentially voiding any and all privacy guarantees.

LeapYear has a dedicated group of solutions architects who have vertical expertise, a foundational understanding of differential privacy, and experience in implementing differential privacy in the context of specific business problems across healthcare, technology, financial services, and government.

IMPACT FOR THE ENTERPRISE

Prior to differential privacy, access to information was effectively synonymous with access to data—to obtain insights from data, one had to have access to data. However, the reality is that these are distinct ideas, and the value of differential privacy is that it provides a rigorous framework for drawing a hard line between them—for the first time, differential privacy enables individuals who previously could not get access to data sets to still derive valuable insights from the information contained in the datasets.

From the experiences of LeapYear's customers, the impact of what initially seems to be a nuance is in fact far-reaching and unprecedented—with LeapYear's platform, enterprises can leverage highly sensitive data sets in ways that they could not even imagine before.

Below are a few examples of how institutions are leveraging LeapYear's platform to use sensitive data sets in new ways:

- Healthcare. Multiple top-5 U.S. health insurance companies making medical data on 100 M+ individuals available to third parties for understanding the effectiveness of therapies, without exposing any private information about individual patients.
- Retail banking. Global banks sharing insights from customer data across borders (countries with strict data residency requirements) with partners (such as co-brand card partners) and across lines of business (such as with investment research), without exposing any customer data across these boundaries.
- Capital markets. Multiple top-10 brokers analyzing data across their institutional clients' holdings and trades to develop information products for clients to better understand financial markets while ensuring that one client can never see or infer proprietary information about another client.
- Technology. Several global technology firms making user data available for research and business partnerships while ensuring that no information about individual user activity on their platforms can be viewed, exfiltrated, or reconstructed.

CONCLUSION

The progress of many major technological innovations can be traced back to a single idea. This idea tends to serve as a foundational pillar on which the future is built: for instance, the explosion of modern computing

can be traced back to the transistor, the internet to networking protocols, and information security to public key cryptography. Differential privacy has the right properties to be a pillar for privacy; it is intuitive, generalizable, and broadly applicable. The academic community has already settled on differential privacy as the de-facto standard for privacy research.

As with many technologies, there is a gap to be bridged between theory and practice for differential privacy. However, in the case of differential privacy, this gap is particularly significant—15 years of theoretical research did not produce a viable commercial solution. At LeapYear, we are a team of researchers, engineers, and business strategists with a singular focus of bridging this gap and bringing differential privacy into practice. In partnership with some of the largest stewards of data, we have demonstrated that with the right combination of resources, talent, direction, and effort, this gap is very much surmountable.

The vision at LeapYear is for every industry to have a pillar on which to build novel data-driven applications from sensitive information and unlock value in ways previously unimaginable, all while protecting the privacy of individuals and institutions to a degree never thought possible before.