Privacy by Design
in the Age of Big Data

June 8, 2012

Ann Cavoukian, Ph.D.
Information & Privacy Commissioner
Ontario, Canada

Jeff Jonas
IBM Fellow
Chief Scientist, IBM Entity Analytics

TABLE OF CONTENTS

Foreword

I am a great lover of quotes. The 13th century Persian poet Jalal-e-din Mohammad Rumi once beautifully wrote that it is necessary to "[s]peak a new language so that the world will be a new world." If our present era is characterized as the information age, the world of Big Data is a new world in which we find ourselves. Algorithms are the lingua franca of this unchartered terrain.

However, algorithms are not the whole story — our existing algorithmic tools struggle to manage and make sense of mankind's unprecedented ability to capture and store data. In response to these new conditions a new class of algorithms designed to harness Big Data have emerged. Organizations of all sizes are now able to better leverage their hitherto trapped information assets. These Big Data developments present us with both opportunities and challenges. While organizations and consumers will benefit from more efficient operations, better customer experiences, and less fraud, waste and abuse, organizations must face new challenges if Big Data is to realize its potential without eroding cherished privacy rights and civil liberties.

One of the true visionaries leading the effort to make sense of Big Data is Jeff Jonas. Jeff Jonas is the chief scientist of the Entity Analytic Solutions group, and an IBM Fellow. In these capacities, he is responsible for shaping the overall technical strategy of next generation entity analytics and the use of these new capabilities in IBM's overall technical strategy.

Jeff Jonas applies his real world, hands-on experience in software design and development to drive innovation while at the same time delivering better privacy protections. By way of example, one breakthrough developed by Jeff Jonas involves an innovative technique enabling advanced data correlation while using only irreversible cryptographic hashes. This new technique makes it possible for organizations to discover records of common interest (e.g., identities) across systems without the transfer of any personally identifiable information. This privacy-enhancing technology, known as "anonymous resolution" significantly reduces the risk of unintended disclosure while enabling technology to contribute to critical societal interests such as clinical health care research, aviation safety, homeland security and fraud detection.

I was delighted to see Jeff Jonas present his work on analytic sensemaking over Big Data at our annual *Privacy by Design* event in Toronto, in 2011. As a technologist who really 'gets it' he presented how his latest technology incorporates a number of *Privacy by Design* principles by default — demonstrating it is possible to advance privacy protections while at the same time preserving functionality in a 'win-win,' or positive sum paradigm. This work serves as a great example that consumer privacy is not simply a compliance issue but is in fact a business imperative. Responsible innovation practices such as these are critical in order to ensure that the new world we are now creating is one where privacy and civil liberties continue to prevail.

Ann Cavoukian, Ph.D.
Commissioner

Introduction

Ninety per cent of the data in the world today was created in the last two years. It has been remarked, for example, that "[t]here was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days, and the pace is increasing."[1] Welcome to the age of Big Data. This data is being generated by sensors and humans, from practically everywhere, and at a blistering pace that surely will continue to only increase. As some refrigerators are now sold Internet-ready and prescription pill vials are now reporting on their status via the cellular network, there are big changes on the horizon.

The big change is Big Data. More specifically, how organizations will leverage Big Data analytics to maximize these growing information assets — driven by their deep interest to maximize their resources and better compete in the market.

While organizations have practical incentives to make the most of their ever-growing observation space (the data they have access to), they also have a pressing need to embed in these systems enhanced privacy protections. We outline in this paper just such an example — how an advanced Big Data sensemaking technology was, from the ground up, engineered with privacy-enhancing features. Some of these features are so critical to accuracy that the team decided they should be mandatory — so deeply baked-in they cannot be turned off.

This paper demonstrates how privacy **and** responsibility can be advanced in this new age of Big Data analytics.

1 Google CEO Eric Schmidt. Techonomy Conference in Lake Tahoe, CA. August 2010.

The Big Difference with Big Data

Big data is the next frontier for innovation, competition, and productivity. The term "Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.[2] But as technological advances improve our ability to exploit Big Data, potential privacy concerns could stir a regulatory backlash that would dampen the data economy and stifle innovation.[3] These concerns are reflected in, for example, the debate around the recently proposed European legislation that includes a 'right to be forgotten' that is aimed at helping individuals better manage data protection risks online by requiring organizations to delete their data if there are no legitimate grounds for retaining it.[4]

Organizations are developing a more complete understanding of their customers than ever before, as they better assemble the data available to them. Public health authorities, for example, have a need for more detailed information in order to better inform policy decisions related to managing their increasingly limited resources. The ability to garner insights from Big Data will without a doubt be of enormous socio-economic significance. Extracting insights from Big Data has quickly become a focus area for technologists worldwide.

The term "Big Data technologies" describes a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.[5] Today's Big Data will provide the raw material for tomorrow's innovations. Navigating this massive volume of information will require us to think about data in new and innovative ways.

While these efforts are to be welcomed, they have potential ramifications for privacy. By way of example, algorithms can now automatically infer that different digital transactions in different systems are in fact related to the activity of a single person or household. A bank that wants to better serve its customers will be eager to know if a specific customer has three relationships with the bank and has an enormous Twitter following. In the past, identifying the difference between six people each with one fact versus one person with six facts was expensive and difficult — something only the larger organizations could accomplish. Today, the advanced analytics needed to reconcile like entities over diverse data sets

2 Manyika, J., et. al. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Online: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

3 Tene, O., and Polonetsky J. (2012). Privacy in the age of big data: A time for big decisions. Stanford Law Review 64, 63.

4 Commission Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), COM (2012) 11 final (Jan. 25, 2012). Online: http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm.

5 Gantz. J., and Reinsel. D. (2011). Extracting value from chaos. IDC. Online: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

(commonly called Entity Resolution) on a Big Data scale are becoming available to organizations of all sizes.

As more data, from more sources, assembles around a single individual — despite de-identification efforts — attempts to reliably protect identity is compromised.[6] Imagine a folder that contains no references to the neighborhood you live in, the neighborhood where you work, your favorite coffee shop, and the make/model/year of your car. Without personal identifiers, could it be associated with you? As more and more individually benign facts are assembled, they collectively become strongly identifying; indeed, the right set of such data can approach your driver's license number in its ability to identify you.

This does not, however, argue against using techniques to de-identify personal data. Indeed, de-identification techniques remain crucial tools in the protection of privacy. However, we must not ignore the fact that Big Data can increase the risk of re-identification — and in some cases, inadvertently re-identify large swaths of de-identified data all at once.

Sensemaking Systems

"Sensemaking" relates to an emerging class of technology designed to help organizations make better sense of their diverse observational space. This observation space will often encompass data they have in their possession and control (*e.g.,* structured master data), as well as data they cannot control (*e.g.,* externally-generated and less structured social media).[7] Sensemaking systems will handle extremely large data sets — potentially involving tens to hundreds of billions of observations (transactions) — being generated from an ever increasing diverse range of data sources (e.g., from Twitter and OpenStreetMap to one's cyber security logs). Obviously these volumes are beyond the capacity of human review.

Sensemaking systems will be used by organizations to make better decisions, faster.[8]

From a sensemaking point of view an organization can only be as smart as the sum of its observations. These observations are collected across the various enterprise systems, such as customer enrollment systems, financial accounting systems, and payroll systems. With each new transaction an organization learns

6 Ohm, P. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA. Online: SSRN: http://ssrn.com/abstract=1450006.

7 Jonas, J. (2011). Master Data Management (MDM) vs. Sensemaking. Online: http://jeffjonas.typepad.com/jeff_jonas/2011/11/master-data-management-mdm-vs-sensemaking.html.

8 More generally in the literature, sensemaking refers to a set of meta-theoretical assumptions that lead explicitly to an overall approach to framing questions, gathering data, and conducting analyses for arriving at substantive theory. This approach has been under development, primarily through the communications research of Brenda Dervin, since 1972, but has since been guided by other disciplines. Sensemaking's core assumption is that of discontinuity. There are gaps between entities which include other people, artefacts, systems, or institutions. Information seeking is associated with these 'cognitive gaps' in our understanding. Filling the cognitive gaps in our understanding is much like asking for street directions in a foreign country.

something. When something is learned, an opportunity arises to make some sense of what this new piece of data means, and to respond appropriately.[9]

The inability of an organization to benefit from the information it has access to or has generated in the past can result in what has been referred to as 'enterprise amnesia.' Studies, for example, conducted for a major retailer found that out of every 1000 employees hired, two had been previously arrested for stealing from the same store for which they had been rehired.[10]

The challenge that organizations face in this regard is growing, because their observation space is growing too — at an unimaginable rate. Today, these observations tend to be scattered across different data sources, located in physically different places, and organized in different forms. This distribution of data makes it difficult for an organization to recognize the significance of related data points. Sensemaking seeks to integrate an organization's diverse observation space — a growing imperative if an organization is to remain competitive.

Historically, advanced analytics have been used, among other things, to analyze large data sets in order to find patterns that can help isolate key variables to build predictive models for decision-making. Companies use advanced analytics with data mining to optimize their customer relationships;[11] law enforcement agencies use advanced analytics to combat criminal activity from terrorism to tax evasion to identify theft. Naturally, these methods have their limits; for example, data mining in search of new patterns in counter-terrorism may yield little value.[12]

A new class of analytic capability is emerging that one might characterize as "general purpose sensemaking." These sensemaking techniques integrate new transactions (observations) with previous transactions — much in the same way one takes a jigsaw puzzle piece and locates its companions on the table — and use this context-accumulating process to improve understanding about what is happening right now. Crucially, this process can occur fast enough to permit the user do something about whatever is happening while it is still happening. Unlike many existing analytic methods that require users to ask questions of systems, these new systems operate on a different principle: the data finds the data, and the relevance finds the user.[13] This is represented in the figure below.

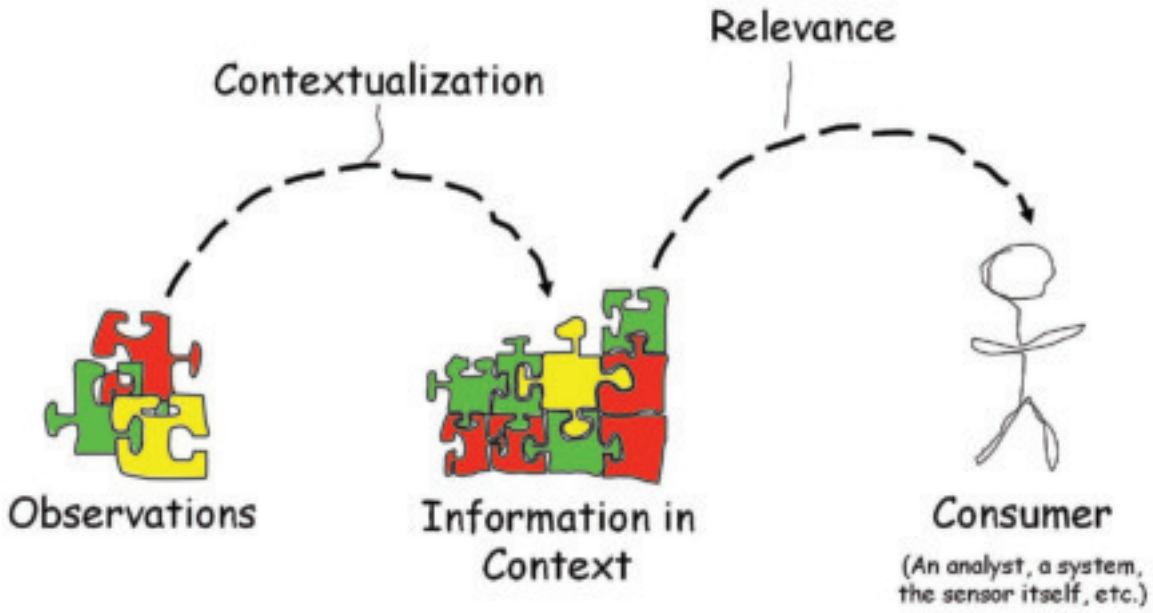
9 Jeff Jonas and Lisa Sokol (2009), "Data finds data," in Segaran, T., and Hammerbacher, J. (eds.), Beautiful Data The Stories Behind Elegant Data Solutions, O'Reilly Media. p. 105.

10 Jonas, J. (Oct 11, 2010). On how data makes corporations dumb. GigaOm. Online: http://gigaom.com/2010/10/11/jeff-jonas-big-data/.

11 Marsella, A., and Banks, M. (2005). Making customer analytics work for you! Journal of Targeting, Measurement and Analysis for Marketing. 13(4), 299-303.

12 Jonas, J., and Harper, J. (2006). Effective counterterrorism and the limited role of predictive data mining. Policy Analysis. CATO Institute, Washington, DC, 584, 1-11.

13 Jonas, J. (2009). Data finds data. Online: http://jeffjonas.typepad.com/jeff_jonas/2009/07/data-finds-data.html



When context accumulating systems are used with Big Data three surprising phenomena emerge:

1. False positives and false negatives both decrease as context reduces ambiguity. This translates directly to higher quality business decisions. Systems that are not operating on context accumulation tend to see increasing false positives and false negatives as the size of the data set grows. Context accumulation produces the opposite effect as data sizes grow.

2. In context-accumulating systems errors in the data (specifically "natural variability") are in fact helpful. Plausible variations in a name such as Ann (also spelled Anne) may be entered by the data operator and the accuracy of context-accumulating systems can be improved as a result of accumulating this variability. One example that might be familiar to many people is that when searching Google and Google responds with "Did you mean _____?" This suggestion is not coming from an internal static dictionary; rather, it has remembered everyone's error(s) in the past. If Google was not keeping this 'bad data' it would not be so smart.

3. Finally, perhaps the most counter-intuitive surprise with respect to context-accumulating systems is that integrating transactions becomes not only more accurate (point #1) but also faster, even as the data store is getting bigger. The most simplistic way to think about this is to consider why the last few pieces of a puzzle are about as easy as the first few when there is more 'data' in front of you than ever before. This phenomenon is apparently new to analytics and is apt to radically change what is possible in the Big Data era, especially in the domain of real-time, sensemaking engines.

However, in these new systems the task of ensuring data security and privacy becomes harder as more copies of information are created. Large data stores containing context-accumulated information are more useful not only to their mission holders but also to those with interests in misuse. That is, the more personally identifiable information Big Data systems contain, the greater the potential risk. This risk arises not only from potential misuse of the data by unauthorized individuals, but also from misuse of the system itself. If the analytics system is used for a purpose that goes beyond its legal mission, privacy may be at risk (for example, if unauthorized surveillance results). For this reason, organizations that want to take advantage of game-changing advances in analytics should stand back and ponder the design decisions that can enhance security and privacy.

By thinking about the privacy implications early on, technologists have a better chance of developing and baking-in privacy-enhancing features, and facilitating the deployment and adoption of these systems. Jeff Jonas has done just this. Below, we outline the privacy-enhancing features of this new technology, a "Big Data analytic sensemaking" engine.[14] This technology has been designed to make sense of new observations as they happen, fast enough to do something about it while the transaction is still happening. Because its analytic methods, capacity for Big Data and its speed are game-changing from a privacy perspective, it has been designed from the ground up with privacy protections in mind. While the result may not be perfect, it is clearly superior to one designed without reference to privacy. We hope it may inspire or guide others in the process of creating their own next-generation analytics.

Privacy by Design in the Age of Big Data

As technologies evolve, our experience and expectations of privacy also evolve. In the past, privacy was viewed as a personal good, rather than a societal one. As such, privacy was regarded as a matter of individual responsibility.[15] Jurisdictions around the world adopted data protection laws that reflected Fair Information Practices (FIPs) — universal privacy principles for the handling of personal data.[16] FIPs reflected the fundamental concepts of data management. The first, purpose specification and use limitation, required the reasons for the collection, use and disclosure of personally identifiable information needed to be identified

14 The framework was formally launched on January 28th, 2011 in Toronto at the *Privacy by Design: Time to Take Control* conference during a keynote speech by Jeff Jonas titled "Confessions of an Architect."

Jeff Jonas (2011), "Sensemaking on Streams – My G2 Skunk Works Project: Privacy by Design (PbD)" http://jeffjonas.typepad.com/jeff_jonas/2011/02/sensemaking-on-streams-my-g2-skunk-works-project-privacy-by-design-pbd.html.

15 Cavoukian, A. (2011). *Privacy by Design* in Law, Policy and Practice. Online: www.ipc.on.ca

16 FIPs was first codified in OECD (1980), OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. Online: http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html. There are a number of articulations of FIPS including The Canadian Standards Association Privacy Code, the Asia-Pacific Economic Co-operation (APEC) Privacy Framework, the U.S. Safe Harbor Principles and the Global Privacy Standard.

at or before the time of collection. Personal information should not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as authorized by law. The second concept, user participation and transparency, specified that individuals should be empowered to play a participatory role in the lifecycle of their own personal data and should be made aware of the practices associated with its use and disclosure. Lastly, FIPs highlighted the need for strong security to safeguard the confidentiality, integrity and data availability as appropriate to the sensitivity of the information.

Fair Information Practices provided an essential starting point for responsible information management practices. Over time, the task of protecting personal information was seen primarily as a "balancing act" of competing business interests and privacy requirements — a zero-sum mindset. This "balancing" approach emphasized notice and choice as the primary method for addressing personal data management. As technologies advanced, however, the possibility for individuals to meaningfully exert control over their personal information became more and more difficult. Many observers have since taken the view that FIPs were a necessary but insufficient condition for protecting privacy. Accordingly, the attention of privacy regulators has since begun to shift from compliance with FIPs to proactively embedding privacy into the design of new technologies.

An example may highlight how current privacy concerns relate to the forces of innovation, competition and the global adoption of information communications technologies. Privacy risks to data about identifiable individuals may largely be addressed with the proper use of de-identification techniques, combined with re-identification procedures. These techniques can simultaneously minimize the risk of unintended disclosure and re-identification, while maintaining a high level of data quality (a key to usability).[17] Nevertheless, complex and rapid technological change (e.g., emerging analytics) may create privacy harms as a byproduct; for example, more powerful analytics may inadvertently make it possible to re-identify individuals over large data sets. Ideally, then, privacy needs to be embedded, by default, during the architecture, design and construction of the processes. This was the central motivation for *Privacy by Design* which is aimed at reducing risks of privacy harm from arising in the first place.

PbD is based on seven (7) Foundational Principles. It emphasizes respect for user privacy and the need to embed privacy as a default condition, but preserves a commitment to functionality in a 'win-win,' or positive-sum strategy. This approach transforms consumer privacy issues from a pure policy or compliance issue into a business imperative. Since getting privacy right has become a critical success factor to any organization that deals with personal information, taking an approach that is principled and technology-neutral is now more relevant than ever. *PbD* is focused on processes rather than a singular focus directing technical outcomes. This approach reflects the reality that it is difficult in practice to favourably impact both consumer and user behaviour after the fact. Rather, privacy is best proactively interwoven into business processes and practices. To achieve this,

17 Cavoukian, A., and Emam, K.E. (2011). Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy. Online: www.ipc.on.ca.

privacy principles should be introduced early — during architecture planning, system design, and operational procedures. These principles, where possible, should be rooted into the code with defaults aligning both privacy and business imperatives.

PbD prescribes that privacy be built directly into the design and operation, not only of technology, but also how a system is operationalized (e.g., work processes, management structures, physical spaces and networked infrastructure.)[18] Today, *PbD* is widely recognized internationally as the standard for developing privacy compliant information systems.[19] As a framework for effective privacy protection, *PbD*'s focus is more about encouraging organizations to both drive and demonstrate their commitment to privacy than some strict technical compliance definition.[20]

In short, in the age of Big Data, we strongly encourage technologists engaged in the design and deployment of advanced analytics to embrace *PbD* as a way to deliver responsible innovation. In fact, we envision a future where technologists will increasingly be called upon to bake-in, from conception, more privacy-enhancing technologies directly into their products and services.

Exemplar: The Creation of a Big Data Sensemaking System Through *PbD*

In late 2008, Jeff Jonas embarked on an ambitious journey to create a sensemaking-style system. This effort started with overall architecture planning and design specifications. Over the first year of this project, while drafting and redrafting these blueprints, his team worked to embed properties that would enhance, rather than erode, the privacy and civil liberties of data subjects.

To engineer for privacy, his team weighed performance consequences, default settings, and which, if any, *PbD* features should be so hard wired into the system they literally cannot be disabled.

Over the year that spanned the preliminary and detailed design, the team created a robust suite of *PbD* features. Indeed, Jeff's team believes this sensemaking system has engineered more privacy and civil liberties-enhancing qualities into this system than any predecessor. If others differ, we welcome debate and vigorous competition as more engineers take up the challenge of *PbD*.

To this end we outline here the privacy and civil liberties-enhancing features included in Jeff's big data analytic platform for sensemaking. We share these features in hopes others will be able to draw inspiration from these features,

18 Cavoukian, A. (2010). *Privacy by Design*: the definitive workshop. A foreword by Ann Cavoukian, Ph.D. Identity in the Information Society, 3(2), 247-251.

19 On October 29, 2010, Dr. Ann Cavoukian's concept of "*Privacy by Design*" was unanimously adopted at the 32nd annual International Conference of Data Protection and Privacy Commissioners, a worldwide assembly of regulators in what has been described as a "landmark" resolution regarding *Privacy by Design*.

20 Cavoukian, A. (2011). *Privacy by Design* in Law, Policy and Practice. Online: www.ipc.on.ca.

improve upon them, build upon and extend them, and envision bigger, better and more important privacy-enhancing features. Ideally, we hope those who make important advances in this area share their good ideas as Jeff is doing.

The remainder of this section details the specific *PbD* features that Jeff Jonas and his engineering team addressed in their engineering of a next-generation sensemaking system.

1. FULL ATTRIBUTION[21]: Every observation (record) needs to know from where it came and when. There cannot be merge/purge data survivorship processing whereby some observations or fields are discarded.

Attribution refers to where the data came from. Every record contained in the database includes the metadata that points to the source of the record – this pointer consisting of a data source and a transaction ID. Full attribution means recipients of insight from our engine can trace every contributing data point back to its source. When systems use merge/purge processing it becomes difficult to correct earlier mistakes (when a different assertion should have been made) as some original data has been discarded. Full attribution also enables system-to-system reconciliation audits of the data — particularly important when dealing with large information-sharing environments.

Full attribution is so important to our sensemaking system that it cannot be turned off.

2. DATA TETHERING[22]: Adds, changes and deletes occurring in systems of record must be accounted for, in real time, in sub-seconds.

Data currency in information-sharing environments is important, especially where data is used to make important, difficult-to-reverse decisions that may affect people's freedoms or privileges. For example, if derogatory data is removed or corrected in a system of record, such corrections should appear immediately across the information-sharing ecosystem. In our sensemaking system, every reported change results in instantaneous correction.

Applying adds, changes and deletes from data-tethered systems of record cannot be turned off.

21 Jonas, J. (2006). Source attribution, don't leave home without it. Online: http://jeffjonas.typepad.com/jeff_jonas/2006/10/source_attribut.html.

22 Jonas, J. (2006). Data tethering: Managing the echo. Online: http://jeffjonas.typepad.com/jeff_jonas/2006/09/data_tethering_.html.

3. ANALYTICS ON ANONYMIZED DATA[23]: The ability to perform advanced analytics (including some fuzzy matching) over cryptographically altered data means organizations can anonymize more data before information sharing.

Every copy of data increases the risk of unintended disclosure. To reduce this risk, data should be anonymized before transfer; upon receipt, the recipient will have no choice but to anonymize it at rest (when placed into a database). And thanks to our full attribution requirement, re-identification is by design, in order to ensure accountability, reconciliation and audit. This feature permits data owners to share their information in an anonymized form that nevertheless yields materially similar results when subject to advanced analytics. Reduction of risk without a material change in analytic results makes for a very compelling case to anonymize more data, not less. We believe this privacy-protecting feature will enhance trust in information-sharing environments and will result in positive win-win outcomes.

We decided system administrators must be able to select which, if any, fields should be configured and anonymized; this feature, therefore, is at the discretion of the policy-makers.

4. TAMPER-RESISTANT AUDIT LOGS[24]: Every user search should be logged in a tamper-resistant manner — even the database administrator should not be able to alter the evidence contained in this audit log.

The question "Who will watch the watchmen?" remains as relevant today as when it was first posed in Latin two thousand years ago. People with access and privileges can, and do, occasionally look at records without a legitimate business purpose, e.g., an employee of a banking system looking up his neighbour's account. Tamper-resistant logs make it possible to audit user behavior. Implementing them may decrease violations, because where employees know such audits are possible, they may be less likely to succumb to temptation.

Following a passionate debate, we decided to include a tamper-resistant audit log subsystem as an integral and mandatory component of the sensemaking system. As such, access to a tamper-resistant audit logging mechanism is a guarantee. However, system administrators carry the responsibility of turning it on.

23 Jonas, J. (2007). To anonymize or not anonymize, that is the question. Online: http://jeffjonas.typepad.com/jeff_jonas/2007/02/to_anonymize_or.html.

24 Jonas, J. (2006). Immutable Audit Logs (IAL's). Online: http://jeffjonas.typepad.com/jeff_jonas/2006/02/immutable_audit.html.

5. FALSE NEGATIVE FAVORING METHODS: The capability to more strongly favor false negatives is of critical importance in systems that could be used to affect someone's civil liberties.

In many business scenarios, it is better to miss a few things (false negatives) than inadvertently make claims that are not true (false positives). False positives can feed into decisions that adversely affect people's lives – e.g., the police find themselves knocking down the wrong door or an innocent passenger is denied permission to board a plane. Sometimes a single data point can lead to multiple conclusions. Systems that are not false negative favoring may select the strongest conclusion and ignore the remaining conclusions. We have applied great effort to account for such conditions by creating special algorithms that favor false negatives.

This non-trivial behaviour is taking some work. We currently believe it will work as envisioned, and we hope we can make a sufficient technical, ethical, and business case to make this feature non-elective (always on).

6. SELF-CORRECTING FALSE POSITIVES[25]: With every new data point presented, prior assertions are re-evaluated to ensure they are still correct, and if no longer correct, these earlier assertions can often be repaired — in real time.

A false positive is an assertion (claim) that is made, but is not true; e.g., consider someone who cannot board a plane because he or she shares a similar name and date of birth as someone else on a watch list.

Where false positives are corrected by periodic monthly reloading, wrong decisions can persist for up to a month, even though the system had sufficient data points on hand to know beforehand. In order to prevent this, earlier assertions need to be reversed in real time and at scale, as new data points present themselves.

This happens to be the single most sophisticated technical aspect of our sensemaking system. Imagine having seen one billion records already, and now one record arrives. At this moment one must decide if this new data point can be used to correct *any* previous false positives.[26] This feature continues to get the most attention.

Once fully tested, we intend to make this feature compulsory.

25 Jonas, J. (2012). Self-correcting false positives/negatives: Exonerate the innocent. Online: http://jeffjonas.typepad.com/jeff_jonas/2012/05/self-correcting-false-positivesnegatives-exonerate-the-innocent.html.

26 Of course, false negatives can and are fixed in real time too – however these require more trivial compute.

7. INFORMATION TRANSFER ACCOUNTING[27]: Every secondary transfer of data, whether to human eyeball or a tertiary system, can be recorded to allow stakeholders (e.g., data custodians or the consumers themselves) to understand how their data is flowing.

In order to monitor information flows, information transfer accounting can be used to record both a) who inspected each record and b) where each record has been shipped off to. This log of outbound accounting (out to eyeballs or out to systems) would work much like the U.S. credit reporting system whereby at the bottom of the credit report is a log of who has pulled the file.

This increases the transparency into how systems are used. One day, it could enable a consumer, in some cases, to request an information recall.

As an added benefit, when there is a series of information leaks (e.g., an insider threat), information transfer accounting makes discovery of who accessed all records in the leaked series a trivial computational effort. This can narrow the scope of an investigation when looking for violating members within an organization.

Our information transfer accounting capability is configured at the discretion of the system administrators. We encourage adoption by having designed our underlying sensemaking data structures to support this type of usage data easily, which makes implementing this feature relatively simple.

Conclusion

Big Data has the potential to generate enormous value to society. In order to ensure that it does, opportunities to enhance privacy and civil liberties are best conceived early on. In this paper we have explored the emergence of Big Data sensemaking systems as an emerging capability with an unprecedented ability to integrate previously diversified data — and in some cases, data about people and their daily lives. The use of advanced analytics has made it possible to analyze large data sets for emerging patterns. It is increasingly apparent, however, that these techniques alone will be insufficient to manage the world of Big Data — especially given the need for organizations to be able to respond to risks and opportunities in real time. Next-generation capabilities like sensemaking offer a unique approach to gaining relevant insights from Big Data through context accumulation. While these new developments are highly welcome, building in privacy-enhancing elements, by design, can minimize the privacy harm, or even prevent the privacy harm from arising in the first place. This will in turn engender greater trust and confidence in the industries that make use of these new capabilities. The dynamic pace of technological innovation requires us to protect privacy in a proactive manner in order to better safeguard privacy within our societies. In order to achieve this

27 Jonas, J. (2007). Out-bound Record-level Accountability in Information Sharing Systems. Online: http://jeffjonas.typepad.com/jeff_jonas/2007/12/out-bound-recor.html.

goal, system designers should be encouraged to practice responsible innovation in the field of advanced analytics.

With this in mind, we strongly encourage those designing and building next-generation analytics of any kind to carry out this work while being informed by *Privacy by Design* as it relates to personally identifiable data.

Information and Privacy Commissioner,
Ontario, Canada

2 Bloor Street East
Suite 1400
Toronto, Ontario
Canada M4W 1A8

Web site: www.ipc.on.ca
Privacy by Design: www.privacybydesign.ca

June 2012

Information and
Privacy Commissioner,
Ontario, Canada