The program of the event was the following:

13h00 | General Assembly of the Swiss Association for Analytics (SAA)

14h00 | Start of the event – introduction from the Committee

1st Half | Customer Intelligence

14h10 | Keynote Speech – Dean Abbott – The Revolution in Retail Customer Intelligence (50 + 10 mins)

15h10 | Valérie Ameel and Raquel Fandos Marin – Recommendation Systems in CRM (20 + 5 mins)

15h35 | 15 mins break

2nd Half | CRM

15h50 | Sponsor Presentation from SAS (10 mins)

16h00 | Dr. Jukka Hekanaho – Incorporate Voice of Customer into CRM Analytics (20 + 5 mins)

16h25 | Werner Zürcher – Using survival analysis to identify profitable members (20 + 5 mins)

16h50 | Olivier Gosset – How to marry CRM & big data and make it (more) real? (20 + 5 mins)

17h15 | Networking apéro

Our next event will take place on June 18th 2015 at 6 pm in Lausanne.

INTRODUCTION INTO THE LEGAL ASPECTS OF BIG DATA[1] (PART #3: PRIVACY ASPECTS – BASICS)

Introduction

This is the third article in a row to give an introduction to the legal aspects of Big Data. In part #1 of this introduction[2], we have laid out some thoughts to the definition of Big Data, and we have complemented this view in a part #2 by defining ownership in information. Big Data is disputed because of privacy issues: We went too far if our neighbor knows more about us than we do ourselves. Instead of "neighbor" use any other role you can think of: businesses, governments, etc. And this is why we should take a closer look onto the privacy implications of Big Data.

While this present article focuses more on the basics of what data protection is and means, a follow-up article in this magazine will tackle specific, and more practical, topics, responding to what actually a researcher or a company can do when it comes to personal data.

Starting Point

If personal data are being used in big data collections numerous concerns can be raised. We discuss some of them hereinafter. Personal data sets can be injected into such data collections as follows[3]:

- Personal data may be submitted by individuals at their own initiative, either to an open group of recipients (e.g., by posting personal information on online social networks) or to the receiving organization, only, as a requirement of a service (e.g. web forms, etc.).

1. This contribution is the third article in a series of articles to discuss legal aspects of data and data analytics. In the 2014/01 issue of this magazine, we have started this series on legal aspects of big data with a definition of "Big Data". Two key words to remind what we have focused on: a kaleidoscope approach, and contracts. Then, we discussed what ownership in data is.

2. Christian Laux, Introduction Into The Legal Aspects of Big Data, Swiss Analytics Magazine 2014/01, 15 et seq.

3. Examples taken from International Working Group on Data Protection in Telecommunications (n° 675.48.12), Working Paper on Big Data and Privacy, Privacy principles under pressure in the age of Big Data analytics, 4, para. 12.

- Personal data may be collected automatically, in connection with the use of specific services (e.g., toll booth transaction data and location data). Such collection of data may also be carried out unknowingly.
- Personal data (e.g., detailed customer records) may come from a partnering organization that is sharing its information with the receiving organization.

Sometimes, such data collection occurs on the basis of legal requirements. The receiving organization would gather such data and use it, and potentially combine it with own records (e.g. customer records, etc.). The received information about persons can be used to enrich existing databases. As a result, conclusions inferred from the processing and analysis of data collected for previous and other purposes often are new information. And derivatives drawn from various sets of information may be personal data, even if the sources appeared to be anonymous at the start.

Data Protection and Privacy

Data protection[4], as a legal discipline, does not exactly do what it says on the tin. "Data Protection" seems to imply that *data* are protected. But Data Protection in fact is a concept to protect *individuals*[5]. Contrary, the concepts to protect *data* have been discussed in our second contribution.[6]

Individuals protected by *data protection* or *privacy* are called "data subjects" under European legislations, including the Swiss legislation.

Data protection regimes lean to either of two separate methods of protection:

- European style data protection laws use the term "personal information" or "personal data" to describe information that can be linked to a person. European style data protection laws resemble to Intellectual Property Rights (IPR), giving the data subject a right to object to using personal data related to their person.
- US style privacy laws are different. They apply a *transactional kind of protection* (instead of an IPR based approach). He or she who can access information can reuse it, except if restricted by an agreement with the source of information[7]. This is how privacy policies work in the US, they provide for the notice (e.g. by the owner of a website) and the consent (of the user), resulting in said transactional method of protection. The operator of a website would engage in unfair competition and possibly be subject to statutory sanctions if it were in breach of the website's privacy policy. And the user has a contractual claim to enforce the privacy policy. This method of protection is backed by sector specific regulation. In the U.S., sector specific regulation is based on the understanding that certain categories of data imply a much broader risk to individuals than other data[8]. Such protected data items are referred to as personally identifiable information (PII)[9].

Data Protection and Privacy from a Society Perspective

Is Data Protection Important? And if so: Why?

Many data subjects do not experience an important fear if they share specific information. "After all, we have nothing to hide". The truth is that privacy matters even if one does not have anything to hide[10]. Data protection as a discipline is more structural in nature. The argument of a German court was that data protection is important in order to avoid chilling effects: If people cannot trust their private sphere is protected they must assume it is not – which in turn would eventually keep them from freely expressing themselves. Freedom of speech, however, is at the core of democracy. Silence due to the fear of being subject to observation would

4. In this article, we refer to "data protection" as a synonym to "privacy", noting, however, that "privacy" is known in the US jurisdiction, and data protection in continental jurisdictions. Further, it should be mentioned that the US do not have an established body of law protecting one's personality in the same way as it is being implemented in continental European jurisdictions.

5. Bruno Baeriswyl, Big Data zwischen Anonymisierung und Re-Individualisierung, in: Weber/Thouvenin: Big Data und Datenschutz - Gegenseitige Herausforderungen, Zurich 2014, 57.

6. Christian Laux, Introduction Into The Legal Aspects of Big Data, Swiss Analytics Magazine 2014/02, 19 et seq.

7. Typically, but not necessarily, source of information is the data subject.

8. E.g. telephone numbers, fax numbers, e-mail addresses, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers, certificate/license numbers, device identifiers and serial numbers, URLs, IP addresses, biometric identifiers, full face photos in the context of health related information under the Health Insurance Portability and Accountability Act of 1996 (HIPAA), 45 C.F.R. § 164.514(b)(1).

9. It is common to state that there is a legal difference between PII (US model of protection) and personal information (EU model of protection). However, mere nomenclature should never be emphasized too much to compare legal regimes. The single most important question is what protection / legal consequences either regime provides for.

10. Daniel J. Solove, http://chronicle.com/article/Why-Privacy-Matters-Even-if/127461/.

dramatically affect democracy and much more. Professor Daniel J. Solove summarizes potential damages to a data subject's privacy as follows: "They involve less the overt insult or reputational harm to a person and more the creation of the risk that a person might be harmed in the future."[11]

Potential damages in case of a Privacy Breach

_____ But yes, privacy risks are difficult to measure and understand[12]. Insult and reputational harm in fact *are* risks that can materialize if one's privacy is invaded. But, there is a more compelling way to describe the risk inherent to a privacy breach: Ohm suggests referring to this risk as the *database of ruin*.

The database of ruin exists only as a hypothesis: "It is the worldwide collection of all of the facts held by third parties that can be used to cause privacy-related harm to almost every member of society."[13] The term describes what happens if the hypothesis turned true, and if it were possible to "join the data from all of the databases in the world together into one, giant, database-in-the-sky – an irresistible target for the malevolent."[14] Assuming such a *database of ruin* would exist: Anyone accessing it could do tremendous harm. This is the privacy issue in big data. Ohm claims that "Regulators should care about the threat of harm from re-identification because this database-in-the-sky contains information about all of us."[15]

PII is an ever-expanding category

_____ "The trouble is that PII is an ever-expanding category"[16]. Often cited research has shown that it is possible to re-identify individuals on the basis of an anonymized data set if only the adversary had some additional knowledge[17]. These days, many have disclosed a great deal of information on social media sites, so the possibility of having useful outside information available to re-identify a dataset has increased significantly and the question is asked whether anonymization still is an appropriate remedy to protect privacy. As anonymization is so important, we should add some comments about anonymization.

Anonymization

Anonymity is rewarded by the law

_____ It is relevant to note that the Swiss data protection act (DPA) exempts anonymized use from its scope. Obviously, if a data set has no implications on a person's individual situation then there is also no need to make use of that data set subject to specific rules. Accordingly, as anonymization is rewarded by law in most jurisdictions, anonymization becomes a key discipline.

Definition

_____ "Anonymization is a process by which information in a database is manipulated to make it difficult to identify data subjects"[18]. But anonymization must be made properly. "It is a well-known fact that the removal of direct identifiers alone is generally insufficient to properly de-identify datasets"[19]. On a high level, the following approaches are known to anonymize data:

- **K Anonymity:** Researchers suggest a range of methods (e.g. suppression of data fields; or generalization) to make several entries in a data record less distinguishable. A table satisfies k-anonymity if every record in the table is indistinguishable from at least {k – 1} other records with respect to every set of quasi-identifier attributes the table contains (ZIP code, age, etc.). Such a table is called a k-anonymous table.
- **L Diversity:** L diversity is a step forward to improve anonymity compared to what K anonymity brings. The problem in K anonymity is that values outside the given quasi-identifiers (on the basis of which K anonymity has been

11. Daniel J. Solove, A Taxonomy of Privacy, 154 U. PA. L. REV. 477, (2006) nn. 487-488.

12. Jane Yakowitz, Tragedy of the Data Commons, 25 Harvard Journal of Law & Technology Review 1 (Fall 2011), 39.

13. Paul Ohm, Broken Promises of Privacy: Responding to Surprising Failure of Anonymity, 57 UCLA LAW REVIEW 1701 (2010), 1746.

14. Ohm (Fn. 13) 1748.

15. Ohm (Fn. 13) 1748.

16. Ohm (Fn. 13) 1742.

17. Latanya Sweeney: Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000; Arvind Narayanan / Shmatikov Vitali: Robust De-anonymization of Large Sparse Datasets, The University of Texas at Austin, 2008; Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: Identifying personal genomes by surname inference., Science (New York, N.Y.) [2013, 339(6117):321-324]. See Bruno Baeriswyl, "Big Data" ohne Datenschutzleitplanken, in digma 2013.1, 14 et seq.

18. Ohm (Fn. 13) 1707; Thomas Probst, Die Verknüpfung von Personendaten und deren rechtliche Tragweite, in: Astrid/Probst/Gammenthaler (Hg.): Datenverknüpfung, Problematik und rechtlicher Rahmen, Zürich 2011, 13 ff.

19. Ann Cavoukian / Daniel Castro, Big Data and Innovation, Setting the Record Straight: De-identification Does Work, 3 (http://www2.itif.org/2014-big-data-deidentification.pdf; 2 February 2015), citing Article 29 Working Party, "Opinion 05/2014 on Anonymisation Techniques," April 10, 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, p. 9; "Guidelines Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," U.S. Department of Health & Human Services, http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/ guidance.html.

implemented) may still remain distinguishable. The method is a form of group based anonymization by reducing the granularity of a data representation. The L diversity model adds intra-group diversity for sensitive values to the anonymization mechanism.

- **T Closeness:** The T closeness model requires that the values of certain data elements in a class shall not differ from their overall distribution (or not by more than by a given threshold "t"). The value of t constrains the additional information an adversary gains after seeing a single data set[20].

Due to the fact that PII is "an ever-expanding category" anonymization is no longer a solution to protect an individual's, or the society's risk exposure related to PII. Legal scholars even go so far to unmask privacy protection approaches that rely on anonymity as a "privacy theatre".[21] These are harsh words. But they somewhat need to be balanced against statements of relevant authority suggesting that it is by far less expensive to hack a computer than to de-anonymize a data set.[22] The latter statements imply that the grave concerns with anonymization are somewhat overstated. In any event, the "idea that we can single out fields of information that are more linkable to identity than others has lost its scientific basis and must be abandoned"[23]. It is a very well possible scenario that legislators will abandon the safe harbor for anonymized data in future legislations.

How do Courts handle the Modern Challenges to Anonymization?

As laid out, there are voices warning that anonymization does not bring the benefits so far anticipated by lawmakers. But if law is what the courts will decide in fact[24] it is useful to understand whether and how the courts absorb the concerns regarding the potential of anonymization summarized above.

Interestingly, courts do not seem to work with the statement that there is no such thing as anonymization. Courts wish to put singular successes of re-identification in context and ask what these successes mean in general:

When considering the latter, we should always keep an eye on what isolated successes of re-identification might mean:

"[T]he fact that one expert in data anonymity can manipulate the data to determine identity does not necessarily mean, without more, that a threat exists that other individuals will be able to do so as well, nor does it in any way define the magnitude of such a threat or whether that threat, if it in fact even exists, renders the release of the data an act that reasonably tends to lead to the identity of specific persons."[25]

It was an Illinois Appellate Court that went on to receive an expert witness statement from one of the privacy researchers who is known to have re-identified an individual governor of the state of Massachusetts, Dr. Sweeney. Based on Dr. Sweeney's statements, the Appellate Court confirmed the lower court's findings that disclosure of a certain data set would not reasonably lead to de-identification. The courts applied the following line of argument:

"We find it difficult to believe that an individual with less knowledge, education, and experience than Dr. Sweeney has would have been able to navigate the six-step process as adeptly as she did. Clearly, Dr. Sweeney's methodology required knowledge and analytical skills beyond that of the average person. The circuit court even engaged Dr. Sweeney in an extended discussion of her methodology. The trial judge stated, 'I want to go through the project with you step by step as if I was a computer literate person attempting to recreate what you did.' Significantly, although Dr. Sweeney's responses to this line of questioning by the court indicate in great detail how she

20. Jianmeng Cao / Panagiotis Karras / Panos Kalnis / Kian-Lee Tan: SABRE: A Sensitive Attribute Bucketization and REdistribution framework for t-closeness, VDLB Journal (2011) 20:59-81, 60.

21. "Anonymization has become 'privacy theater'; it should no longer be considered to provide meaningful guarantees of privacy." (Ohm (Fn. 13) 1743, quoting Paul M. Schwartz, Reviving Telecommunications Surveillance Law, 75 U. CHI. L. REV. 287, 310–315 (2008)).

22. Yakowitz, 41: "A malfeasor with no specific target in mind is still better off using hacking techniques rather than de-anonymization algorithms." Yakowitz also points out that the unwanted disclosure of private records (e.g. by employees, etc.; Yakowitz calls these events "data spills", i.e. mishandling of unencrypted data) is by far the most important reason for a data breach.

23. Ohm (Fn. 13) 1732.

24. "The prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law", Oliver Wendell Holmes, Jr., The Path of the Law, 10 Harvard Law Review 460-461 (1897).

25. Southern Illinoisan v. Department of Public Health, No. 5-02-0836, Appellate Court of Illinois, Fifth District, June 9, 2004, http://www.state.il.us/court/Opinions/AppellateCourt/2004/5thDistrict/June/Html/5020836.htm (2 February 2015).

knew what to do, her responses lack concreteness and specificity regarding the extent to which others would be able to do the same. Nor did the defendants present any other evidence on this point."[26]

Conclusion

In short, big data is a shift. And it is one of the key challenges of our society to find responses how to best deal with the ever increasing availability of data around us and about us. Hopefully, anonymization techniques still improve. Alternatively, businesses could come up with approaches how to improve the current situation for the user. Quite some initiatives are on their way these days with a focus to either improve identity management techniques, to empower users to retain "their" data, or to build privacy-friendly technical alternatives to the not-so-privacy-friendly mass products out on the market. Then, we do see opportunities how the laws could be improved. Together with governments and other researchers this is a topic to elaborate on in more detail, and separately. – On a more practical level, the fourth contribution in the series about legal aspects of Big Data (in this magazine) will respond to specific privacy-related questions that can come up in practice.

26. http://www.state.il.us/court/Opinions/SupremeCourt/2006/February/Opinions/Html/98712.htm (2 February 2015).

INTRODUCTION TO THE LEGAL ASPECTS OF BIG DATA[1] (PART #4: TEN PRACTICAL PRIVACY QUESTIONS)

Introduction

This is the fourth article in a row to give an introduction to the legal aspects of Big Data. While the contribution #3 (also in this magazine) outlines the theoretical basis for what data protection is and can achieve, this fourth article intends to answer specific questions asked in conversations about data protection and big data. In this article, we want to apply a more practical focus and discuss how data protection aspects can be complied with. In the following, we summarize questions that are regularly raised, and possible responses.

Is it OK to Access Publicly available Data?

Data can only be lawfully used if it has been procured lawfully. Data purchased from dubious sources may cause difficulties to companies. Accordingly, when building data bases to work on, companies should focus on lawful methods of procurement. Now, from a Swiss law perspective, it is generally acceptable to consult data that has been made available to the public by data subjects. Statutory law explicitly states: *"As a rule, there is no breach of privacy if the data subject has made the data generally accessible and has not expressly prohibited its processing."* The law implicitly suggests that exceptions to the otherwise very broad wording *("there is no breach of privacy")* can be made. One exception is an explicit declaration alongside the data made available ("do not reuse") – such declarations are not commonly made, though. Another exception is data that obviously has not been published by the data subject itself (e.g. data from a "leaking site").

1. This contribution is the third article in a series of articles to discuss legal aspects of data and data analytics. In the 2014/01 issue of this magazine, we have started this series on legal aspects of big data with a definition of "Big Data". Two key words to remind what we have focused on: a kaleidoscope approach, and contracts. Then, we discussed what ownership in data is.