## CELEBRATING OUR 1’000TH MEMBER!

In February, we accepted our 1’000th LinkedIn member. Although our goal is to be a local association, it’s a pleasure to have achieved that amount of members in Switzerland and around the world. Thanks to all of you for that!

If you want to support the association, receive the present magazine at home and have access to event presentations, please join us as an official member of the Swiss Association for Analytics. Information is available in the present magazine and on the association website: [www.swiss-analytics.ch/membership](http://www.swiss-analytics.ch/membership)

The current issue contains plenty of analytics articles. Amine Mansour interviewed Alex Monnier from Itecor about Big Data consulting. Christian Laux proposes the third and fourth parts of his article about the legal aspects of Big Data. Manuel Martin Marquez writes about information discovery at CERN. Dean Abbott gives us an excerpt from his last book *Applied Predictive Analytics*. Martin Jaggi uncovers the Zurich Machine Learning and Data Science Meetup. Common columns include agenda of upcoming events, description of the Swiss Analytics Event of November 2014 and a book review. Enjoy the reading!

---

Sandro Saitta  
President of the Swiss Association for Analytics  
Data Scientist at Expedia

---

**TABLE OF CONTENTS**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEBRATING OUR 1’000TH MEMBER!</td>
<td>2</td>
</tr>
<tr>
<td>PRINCIPLES IN FIXING MISSING DATA</td>
<td>3</td>
</tr>
<tr>
<td>AGENDA OF EVENTS</td>
<td>3</td>
</tr>
<tr>
<td>BOOK REVIEW: FORECASTING</td>
<td>5</td>
</tr>
<tr>
<td>INFORMATION DISCOVERY: EXPLORING DATA-DRIVEN DECISION-MAKING</td>
<td>6</td>
</tr>
<tr>
<td>A VIEW ON BIG DATA CONSULTING</td>
<td>9</td>
</tr>
<tr>
<td>ANALYTICS AND STATISTICS FOR CRM: EVENT FEEDBACK</td>
<td>12</td>
</tr>
<tr>
<td>INTRODUCTION INTO THE LEGAL ASPECTS OF BIG DATA (PART #3: PRIVACY ASPECTS — BASICS)</td>
<td>13</td>
</tr>
<tr>
<td>INTRODUCTION TO THE LEGAL ASPECTS OF BIG DATA (PART #4: TEN PRACTICAL PRIVACY QUESTIONS)</td>
<td>17</td>
</tr>
<tr>
<td>ZURICH MACHINE LEARNING AND DATA SCIENCE MEETUP</td>
<td>22</td>
</tr>
<tr>
<td>MEMBERSHIP</td>
<td>23</td>
</tr>
</tbody>
</table>

---

**Due to a maintenance, please use the following address to access our website:**

**PRINCIPLES IN FIXING MISSING DATA**

Excerpted from Chapter 4 of Mr. Abbott’s book

*Applied Predictive Analytics, Wiley 2014*

---

Missing value correction is perhaps the most time-consuming of the variable cleaning steps needed in data preparation. Whenever possible, imputing missing values is the most desirable action. Missing value imputation means changing values of missing data to a value that represents a plausible or expected value in the variable if it were actually known. These are the most commonly used methods for fixing problems associated with missing values.

### Listwise and Column Deletion

Many datasets have missing values in most if not all records, especially in domains related to the behavior of people like customer acquisition, customer retention and survey analysis. One may begin with more than 1 million records, but after removing all records with any missing values at all — called listwise deletion — one may have only 10,000 remaining. Clearly, this isn’t desirable and is rarely done in customer analytics.

Sometimes, you find that predictive modeling software performs listwise deletion by default without alerting the user that this is what is happening. You only discover this default when you examine reports on the models that have been built and can see that very few records were actually used in building the models.

An alternative to listwise deletion is *column deletion*: removing any variable that has any missing values at all, leaving only variables that are fully populated. This approach solves the listwise deletion problem of removing too many records, but may still be too restrictive. If only a few of the values are missing in a column, removing the entire column is a severe action. Nevertheless, both listwise deletion and column deletion are practiced, especially when the number of missing values is particularly large or when the timeline to complete the modeling is particularly short.

### Imputation with a Constant

This option is almost always available in predictive analytics software. For categorical variables, this can be as simple as filling missing values with a “U” or another appropriate string to indicate missing. For continuous variables, this is most often a 0. For some continuous variables, filling missing values with a 0 makes perfect sense. Bank account balances for account types that an individual does not have can be recoded with a 0, conveying information about the account that makes sense (no dollars on account). Sometimes, imputing with a 0 causes significant problems, however. Age, for example, could have a range of values between 18 and 90. Imputing missing values with a 0 clearly doesn’t convey a value that is possible, let alone likely, for donors in a dataset.

### Mean and Median Imputation for Continuous Variables

The next level of sophistication is imputing with a constant value that is not predefined by the modeler. Mean imputation is perhaps the most commonly used method for several reasons. First, it’s easy; mean values are easy to compute and often available without extra computational steps. Second, the
idea behind mean imputation is that the value that is imputed should do the least amount of harm possible; on average, one expects the values that are missing to approach the mean value. However, mean imputation has problems as well. The more missing values imputed with the mean, the more values of the variable fall exactly at the mean, so a spike emerges in the distribution. The problem is primarily with algorithms that compute summary statistics; the more values imputed with the mean, the smaller the standard deviation becomes when you compare the standard deviation of the variable before and after imputation.

For example, in a particular data set, a variable AGE has a mean of 61.64 and standard deviation with 16.62. After imputing missing values for AGE when AGE has 25 percent, 40 percent, and 62 percent missing, the standard deviation shrank from 16.6 to 14.4, 12.9, and 10.2 respectively because of the spike of imputed values falling at the mean. This can mislead some algorithms as the build model parameters.

Mean imputation is by far the most common method, but in some circumstances, if the mean and median are different from one another, imputing with the median may be better because the median will represent better the most typical value of the variable. However, median imputation can be more computationally expensive, especially if the number of records in the data is large.

**Imputing with Distributions**

When large percentages of values are missing, the summary statistics are affected by mean imputation. An alternative to this is, rather than imputing with a constant value, to impute randomly from a known distribution. For the variable AGE, if instead of imputing with the mean (61.6) you impute using a random number generated from a normal distribution with mean 61.6 and standard deviation 16.6. These imputed values will retain the same shape as the original shape of AGE, though there will be small differences because AGE was not shaped exactly like a normal distribution originally. If the variable appears to be more uniformly distributed, you can use a random draw from a uniform distribution rather than a normal distribution for imputation.

Under most circumstances, imputing missing values from a distribution rather than from the mean is preferred numerically, though sometimes the simplicity of mean imputation can be the deciding factor in its use.

**Random Imputation from Own Distributions**

A similar procedure is called “hot deck” imputation, where the “deck” referred originally to the days when the Census Bureau had cards for individuals. If someone did not respond, one could take another card from the deck at random and use that information as a substitute.

In predictive modeling terms, this is random imputation, but instead of using a random number generator to pick a number, a random actual value of the variable of the non-missing values is selected. Predictive modeling software rarely provides this option, but it is easy to do. One way to do it is as follows:

1. Copy the variable to impute (variable “X”) to a new column.
2. Remove missing values and randomly scramble that column.
3. Replicate missing values and randomly scramble that column.
4. Join this column to the original data.
5. Impute missing values of variable X with the value in the scrambled column.

The advantage of random imputation from the same variable is that the distribution of imputed values matches the populated data even when the shape of the data is not uniform or normal.

---

Mr. Abbott is the author of Applied Predictive Analytics (Wiley, 2014) and co-author of IBM SPSS Modeler Cookbook (Packt Publishing, 2013). He is a highly-regarded and popular speaker at Predictive Analytics and Data Mining conferences and meetups, is on the Advisory Boards for the UC/Irvine Predictive Analytics Certificate as well as the UCSD Data Mining Certificate programs, and has served multiples times on the program committee for the industrial track of KDD conferences. He has a B.S. in Mathematics of Computation from Rensselaer (1985) and a Master of Applied Mathematics from the University of Virginia (1987).

---

Dean Abbott is Co-Founder and Chief Data Scientist of Smarter Remarketer, Inc., and President of Abbott Analytics, Inc. in San Diego, California. Mr. Abbott is an internationally recognized data mining and predictive analytics expert with over two decades experience applying advanced data mining algorithms, data preparation techniques, and data visualization methods to real-world problems, including fraud detection, risk modeling, text mining, personality assessment, response modeling, survey analysis, planned giving, and predictive toxicology.

---

**Imputing with Distributions**

When large percentages of values are missing, the summary statistics are affected by mean imputation. An alternative to this is, rather than imputing with a constant value, to impute randomly from a known distribution. For the variable AGE, if instead of imputing with the mean (61.6) you impute using a random number generated from a normal distribution with mean 61.6 and standard deviation 16.6. These imputed values will retain the same shape as the original shape of AGE, though there will be small differences because AGE was not shaped exactly like a normal distribution originally. If the variable appears to be more uniformly distributed, you can use a random draw from a uniform distribution rather than a normal distribution for imputation.

Under most circumstances, imputing missing values from a distribution rather than from the mean is preferred numerically, though sometimes the simplicity of mean imputation can be the deciding factor in its use.

**Random Imputation from Own Distributions**

A similar procedure is called “hot deck” imputation, where the “deck” referred originally to the days when the Census Bureau had cards for individuals. If someone did not respond, one could take another card from the deck at random and use that information as a substitute.

In predictive modeling terms, this is random imputation, but instead of using a random number generator to pick a number, a random actual value of the variable of the non-missing values is selected. Predictive modeling software rarely provides this option, but it is easy to do. One way to do it is as follows:

1. Copy the variable to impute (variable “X”) to a new column.
2. Remove missing values and randomly scramble that column.
3. Replicate missing values and randomly scramble that column.
4. Join this column to the original data.
5. Impute missing values of variable X with the value in the scrambled column.

The advantage of random imputation from the same variable is that the distribution of imputed values matches the populated data even when the shape of the data is not uniform or normal.
Imputing Missing Values from a Model

Imputing missing values as a constant (the mean or median) or with a random value are quick, easy, and often sufficient to solve the problem of missing data. However, better imputation can be achieved for non-MCAR missing values by other means.

The model approach to missing value imputation begins with changing the role of the input variable with missing values to now be a target variable. The inputs to the new model are other input variables that may predict this new target variable well. The training data should be large enough, and all of the inputs must be populated; listwise deletion is an appropriate way to remove records with any missing values. Keep in mind that even a moderately accurate model can still produce good imputations; the alternatives are either random or constant imputation methods.

Modelers have two problems with using models to impute missing values. First, if the software doesn’t have methods to impute values from models automatically, this kind of imputation can take considerable time and effort; one must build as many models as there are variables in the data.

Second, even if the imputation methods are done efficiently and effectively, they have to be done again when the models are deployed, any missing values in data to be scored must first be run through a model, adding complexity to any deployment process.

Nevertheless, the benefits may outweigh the problems. It seems that the modeling algorithms most often included in predictive analytics software to automatically impute missing values are decision trees and k-nearest neighbor, both of which can work well with imputing continuous and categorical values.

BOOK REVIEW: FORECASTING

While working on forecasting (understand “time series analysis”) I found several interesting and state of the art articles from Rob J. Hyndman. He is the co-author, with George Athanasopoulos of Forecasting: Principles and Practice. This is an excellent concise and comprehensive text explaining concepts behind forecasting, common algorithms and how to implement them in R (for a business view of forecasting, I advise Future Ready).

The book presents key concepts of forecasting. From judgmental forecasting (which can be useful when you have no or few data) to simple/multiple regression, time series decomposition, exponential smoothing (ETS), ARIMA and a few more advanced topics such as Neural Networks. I would suggest to the author to add Support Vector Regression (SVR) and ensemble learning for the next edition of the book. Each concept of the book is covered through examples with real data. What is most appreciable about the book is how concise and readable it is. Each sentence is useful to understand the described concept, nothing “superflu”.

The book contains a good overview and schema about each technique and how to set their meta-parameters. The R codes are well presented, concise and easy to implement and test. The book can easily be used to teach forecasting since each chapter contains exercises. In conclusion, Forecasting: Principles and Practice is THE book to learn time series analysis algorithms and how to implement them.

Review by Sandro Saitta, Data Scientist at Expedia
**1. CERN’s accelerator complex.**
The LHC is the last ring in a complex chain of particle accelerators. The smaller machines are used in a chain to help boost the particles to their final energies and to provide beams to a whole set of smaller experiments, which also aim to uncover the mysteries of the Universe.

At CERN, the European Organization for Nuclear Research, fundamental research is carried out to shed light on the most essential questions physicists around the world are currently addressing. Why do particles have mass? Why is no antimatter left in the Universe? Does supersymmetry really exist? What was the state of matter right after the big bang? Of what is 96% of the Universe made? What of dark matter and dark energy? These are just a few examples.

The instruments set up at CERN to answer these questions are purpose-built particle accelerators and detectors. Accelerators boost beams of particles and make them collide. Detectors observe and record the results of these collisions. CERN’s accelerator complex is made up of a series of accelerators that work together to push particles to over 99% of the speed of light. Each machine boosts the energy of a beam of particles, before injecting it into the next machine in the

---

---
sequence — until the beam finally arrives in the Large Hadron Collider (LHC). The LHC is the world’s largest scientific instrument (contained within a 27-kilometer circular tunnel underground) and the most powerful particle accelerator ever built. Beams of protons travel around the LHC over 11000 times per second. They are guided around the accelerator ring by more than 6000 superconducting magnets, which are chilled to a temperature colder than outer space.

The design and construction of the LHC has pushed the boundaries of technology in many fields. It has required the development of novel and innovative technologies, many of which are now present in our own daily lives. The World Wide Web is one example, but many others are detailed in the Global Science Forum (GSF) report published by the Organization for Economic Co-operation and Development (OECD), “The Impacts of Large Research Infrastructures on Economic Innovation and on Society: Case studies at CERN”.

Inline with the aforementioned need to accelerate technological development in order to overcome the unprecedented challenges of the LHC, CERN openlab was founded in 2001. CERN openlab is a public-private partnership between CERN and leading ICT companies, which offers a neutral ground for carrying out advanced R&D in a unique and demanding environment. Within the CERN openlab framework we have targeted the exploration of data analytics technologies as a core task. Our objective is to further develop the data-driven control and monitoring systems of CERN’s accelerator complex.

Over recent decades, CERN has been very successful in developing custom data-driven control and monitoring systems to ensure the safe and efficient running of its scientific instruments. This has required significant investment in data storage and management. Fully exploiting the control and monitoring data helps us to obtain a deeper understanding of the factors that affect the performance and availability of the accelerator complex. This is vital for achieving maximum operational efficiency and in ensuring informed decision-making.

In spring 2015 the second run of the LHC will begin, following a two-year period of maintenance and upgrade work. During this second run, the LHC will significantly increase its luminosity and the energy per beam will almost double. This is an exciting moment for the scientific community, but this also leads to important challenges for all the subsystems that comprise CERN’s accelerator complex and the various engineering teams responsible for them, thus making detailed analysis of the control and monitoring data even more important.

CERN’s particle accelerator infrastructure is comprehensively heterogeneous. A number of critical subsystems, which represent cutting-edge technology in several engineering fields, need to be considered: cryogenics, power converters, magnet protection, etc. The historical monitoring and control data derived from these systems has persisted mainly using Oracle database technologies, but also other sorts of data formats such as JSON, XML and plain text files. All of these must be integrated and combined in order to provide a full picture of the overall status of the accelerator complex. Therefore, a key challenge is to facilitate easy access to, flexible interaction with, and dynamic visualization of heterogeneous data from different sources and domains.

A technology that can play an important role in tackling these challenges is “Information discovery”. “Information discovery” is clearly about much more than just representing data; it must also unlock understanding, facilitate exploration, and spot hidden insights, correlations and patterns. This is indispensable in enabling more intelligent decision-making at all levels and in our case for improving the efficiency of operations at CERN.

Demo of the solution implemented at CERN, based on Oracle Endeca Information Discovery
In order to test the capabilities of this technology, we carried out an extensive review of potential solutions — commercial, open source and custom — on the basis of covering the following key general requirements:

**Speed of implementation:** Prototyping, validating assumptions, and delivering results as quickly as possible from different perspectives is a must in our environment. Thousands of CERN users, accelerator operators, infrastructure controllers and equipment experts, with a wide range of expertise, dig into the data from different perspectives. The technology is pushed to the limits in order to handle the large amount of data generated — in the order of terabytes — and the different data perspectives.

**Volume and diversity of users:** CERN is an international organization where scientists, engineers and specific domain specialists collaborate to deepen our understanding of the Universe. Together with its open nature, this makes CERN a unique environment, with people representing more than 110 nationalities. In principle, the main users of this data are the operation and control teams, but any CERN member may potentially have a unique and valuable point of view on the data that needs to be captured.

In addition, the “information discovery” technology must enable rapid integration of different data sources and exploitation the data in a simple, efficient and user-driven manner. We need to be able to study the control and monitoring data in order to improve the operation of the accelerator complex. Specifically, we use it to find patterns in complex accelerator events, to discover hidden insights and relations between domains, and to track faults and predict potential problems.

**Finding patterns in the accelerator events:** The operations groups, controls teams, and systems continuously generate a large number of events. These events are important to determine the status of the accelerator complex at any given moment in time. The core event information is text-based, meaning advanced text analytics mechanisms are indispensable. The extracted knowledge from the event description is used to spot areas in need of investigation and it helps us find subtle correlations between different subsystems.

**Discovering hidden insights and relations between domains:** If data is to be analysed successfully, it is essential that its nature — including the processes through which it was generated — is properly understood. One of the technology areas that makes this possible is so-called visual analytics. Information discovery techniques support us in going beyond ordinary visualisation of the data and enable us to find hidden patterns and insights, thus revealing previously unknown links between domains.

**Tracking faults, operational issues, and predicting potential problems:** Our goal is to improve the performance and availability of CERN’s accelerators. This work is closely related to the study of faults and potentially related operational issues within the various subsystems (source of the faults, root cause analysis, potential implications on other subsystems, actual and estimated duration, accelerator down-time, etc.). Improved fault analysis capabilities can help us to predict future problems. The capability of the technology to correlate information from a wide range of data sources plays an important role in this.

In summary, data is everywhere and extracting value from it is essential. “Information discovery” is a technology that can help your organization to discover insight into its data. You’re probably not running the largest scientific instrument ever built, but you still want to go beyond data visualization to unlock hidden knowledge. At CERN openlab we have evaluated various solutions. Of them, Oracle Endeca...
Information Discovery has demonstrated potential through a variety of proof-of-concept tests designed using real control and monitoring data. Endeca has emerged as one of the most promising approaches for helping us to explore future possibilities for enabling better decision-making — as well as more efficient operational control — based on insights extracted from the data.

A VIEW ON BIG DATA CONSULTING

Interview with Alex Monnier, by Amine Mansour

In your past 15 years in the IT consultancy world, have you noticed an important paradigm shift which may explain the success of Advanced Analytics and Big Data in industry nowadays?

Indeed, company Data warehouses (DWH) matured, new tools arrived at all levels, and now business and IT are blending because of the pressure for results.

On the IT DWH side, enterprises with mature Business Intelligence (BI) are facing issues of change management. Changes on structured information in RDBMS or cube systems are slow, painful and expensive. The cost of the data was too high with too few ways to reduce it. Therefore, the need of placing data according to factors such as value and business criticality has arisen.

Now, new distributed computing techniques around Hadoop for example, allow for business value extraction from larger datasets in any form respecting the correct tradeoff between cost and speed. This will allow companies to follow the fast pace of business evolution and...
#1 Business Analytics

✓ Leader in Business Intelligence
✓ Leader in Analytic Applications
✓ Leader in Data Warehousing

oracle.com/bi
or call 1.800.ORACLE.1
constant changes in business requirements. Business decisions are now based upon different data types such as video, images and IoT generated data.

In the same time, Advanced Analytics is being successful in many business cases partly due to current technology that allows the use of more complex algorithms and the fact that more people are able to understand the underlying concepts.

New DWH models and repository types enable this distributed computing and enhancement of BI systems to integrate everything, from unstructured to structured, fuzzy to precise facts, into a unified enterprise system.

What is the usual way to undertake a consultancy project in this field?

For a chance of success, the need of change has to be recognized at two levels simultaneously: business and technical. Some try to integrate these processes only in the technical point of view without having convinced business leaders of the risk of making decisions solely or primarily on gut-feeling. Understand me well, gut-feeling is good but it is based on human perception and experience. This is still needed for all problems where there is previously not enough data. The issue is that experience is slow to build and fades away with business changes. With Big Data, we can now expand our natural abilities feeding our perception with new sources of large amounts of data that are synthetized using machine detected models.

On the other side, initiatives driven only by businesses can have negative impacts from continuously changing goals, and lack of understanding with respect to impacts of technical implementations. Selecting the right project size with technically integrable and business actionable results is the most important step. Depending on the company’s data maturity, it is essential to start small to prove actionable results and increase confidence gradually.

Many people believe challenges lie within the choice of tools, and the complexity of the technical details. Is that your experience?

For the time being, there are enough tools to solve anything, however only a subset of them are ready in terms of maturity and stability for production systems. Companies don’t want to update and put at risk their working processes.

The key takeaway is that a tool cannot solve all problems by itself. Software are like a medical devices, as sophisticated as they may be, they will be of no use without a qualified doctor and the right diagnostic.

How does one detect a business pain and what is the difficulty in being able to relieve this pain through new technologies?

Business pains are mostly known but people tend to get used to living with them. Often we hear: “we have always worked in this way”. One needs now to reevaluate the assumptions behind each issue such as lack of quality or loss of services. One way to address it is through the use of criticisms that are posted on forums and shared widely. With technics like Natural Language Processing (NLP), we are now able to organize and leverage these data sources. The biggest issue is still to select the problems to be solved that will deliver actionable results and value.

Being able to fail fast using methodologies such as Agile is now the key point in this field. Iterative processes of business and data understanding with Data Scientists enable the creation of new ideas. As an example we use with some of our clients RapidMiner which fits in this new paradigm of bridging technology, Advanced Analytics and Business Models, allowing both parties to work alongside in an efficient way.
If there is a key take-away among all those experiences? What is the common factor across different industry you noticed the most?

The project selection process is central. Key factors are data quality, existing and potential sources of data, data privacy, internal politics, cost, return, and innovation strength. As our customers are realizing the importance of all the data they currently have and are paying for, they try using it to be innovative and find ways to stay aligned with their business environment.

You mention innovation, and a common concern detected is the large coverage around innovation, without seeing the direct impact. In your sense, what are the purposes of innovation within the Big Data field for business deciders and the added benefits?

Big Data is a buzz word. But fact based decision making was and will always be central. Innovation in domains where you already have data will produce quick wins but gathering new data based on business vision is business exploration, which can be seen like a scientific and experimental setup. In the end, it is to be able to augment our capabilities with data.

What are some examples of questions that you would ask at the beginning?

Do you want to be innovative? Do you want to create new products? Do you want to address the need of different customers? Are you currently looking for the right price for the right customer? Where is the right place for your activity? What is the impact of your promotion? Big Data allows now for a new set of answers to these questions.

In our day to day business, we help our customer fill the gap between tools, technologies and results.

**ANALYTICS AND STATISTICS FOR CRM: EVENT FEEDBACK**

On Friday 7th took place the 5th event of the Swiss Association for Analytics in Lausanne. It was a joint event with the Swiss Statistical Society. About 80 analytics professionals attended the event. I would like to thank our two gold sponsors, PwC and SAS, as well as our partner sponsor IMSD.

It was a great pleasure to have the world-known predictive analytics expert Dean Abbott as a keynote speaker.
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.


• 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, 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. Contrary, the concepts to protect data have been discussed in our second contribution.

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. 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. Such protected data items are referred to as personally identifiable information (PII).

**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. 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
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.”

Potential damages in case of a Privacy Breach

But yes, privacy risks are difficult to measure and understand. 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 risks to 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.

PII is an ever-expanding category

“The trouble is that PII is an ever-expanding category.” 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. 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 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.” 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.” 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 applied) 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.”

14. Ohm (Fn. 13) 1748.
15. Ohm (Fn. 13) 1748.
16. Ohm (Fn. 13) 1742.
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.

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”.

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 decide in fact, 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 anonymization 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 in fact even exists, renders the release of the data an act that reasonably tends to lead to the identity of specific persons.”

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
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.


INTRODUCTION TO THE LEGAL ASPECTS OF BIG DATA¹
(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.
Dr. Christian Laux, LL.M., is a Zurich based attorney for technology, advertising and e-commerce matters, combining both inside and outside counsel experience with a passion for technology.

In his daily practice Dr. Laux advises on all transactional aspects related to software, infrastructure and data distribution agreements, on matters of Intellectual Property, Telecommunications, Privacy, Corporate and Commercial Law. Litigation, in particular commercial litigation, domain name disputes and trademark litigation.

Dr. Laux has extended experience in data related issues, acts as Data protection officer for companies, and is board member of Opendata.ch, an association dealing with data that are lawfully accessible to the public for the benefit of transparency, economical efficiency and more.

Dr. Laux completed his legal studies at universities in Zurich (Switzerland), Paris (France) and Stanford (CA, USA). His Ph.D. thesis discusses relevant issues about agreements covering copyrighted assets.

How to otherwise procure data?

Data that has been procured from a public space can be reused without notifying the data subject (see here before, “there is no breach of privacy”). This is true at least if it was the data subject who has made his or her data accessible there. Data that has been procured from other sources, namely directly from the data subject or from third parties can be used only under more restrictive conditions:

- **data procured directly from a data subject:** the law requires the company to make transparent what it intends doing with the data. In the context of big data collections, companies should make sure that they publish adapted terms of use in order to express that data they use will be run by analytics tools so to permit the company to better learn about its customer base. Companies need to have accountability mechanisms in place (internal process when to release, when to reuse data, how to de-identify data, including by keeping up with technological developments). Companies maintaining big data collections should also publicly commit not to re-identify the data (e.g.: “It is our policy to treat data lawfully. Within those boundaries we may wish to better understand who interacts with us, in what way. We reserve the right to analyze data we have accessed, or that has been submitted to us. We apply methods to anonymize the data, and have process to ensure that, within our sphere of influence, data are not being re-identified”). Such information might be combined with commitments to de-identification and minimum security standards.

- **data procured from a third party** (without any restrictive obligations): data can be procured from a third party. If the data received is sensitive information then the procuring company must inform the data subject about it. The party disclosing the data to the procuring company most likely will also have an obligation to make the disclosure transparent to the data subject.

- **data accessed in the context of an outsourcing relationship:** an outsourcer is not permitted to exploit data received by its customer under an outsourcing regime. Outsourcers must process data on behalf of the outsourcing customer, only. Reuse in the aggregate and after thorough anonymization is not prohibited, though.

Is it OK for an Organization to combine various data sets?

Combination of data is a form of data processing that is not explicitly prohibited by Swiss data protection law. Accordingly, combining information is permitted, except if the processing is totally disconnected to the reasonable and lawful business purpose of the company analyzing the data. Namely if the data processing aims at identifying individuals, or worse, to pursue an unlawful purpose, then processing would be unlawful and could be stopped by anyone affected by the data processing. Unlawful purposes can include denigration of an individual, whether in the public or a closed space (criminal law), analysis of a customer’s purchasing preferences in order to exclude that individual from benefits (such as rebates) other customers would naturally get (unfair business practice to increase revenues), etc. If combination of data sources does not have any other purpose than identifying an individual then such practice is detrimental to the core rules of data protection and most likely will result in unlawful processing of data. If combination of data sources accidentally results in de- anonymization this might not make the entire data processing setup unlawful. The company using the data should, however, establish methods to avoid accidental de-anonymization of external records. Otherwise an investigation against the
Welcome to the new world of Management Intelligence

Welcome to the new world of BOARD.
Today there is a business decision environment that is very different from all others. An extremely advanced environment, yet simple and intuitive, where you can finally have all the benefits of Business Intelligence, Predictive Analytics and Corporate Performance Management within a single software solution.

The secret is called the “toolkit”, a revolutionary approach that provides you with your custom tailored applications: from reporting to analysis, from dashboards to scorecards, from planning to simulation. You can not only create your own solutions, but also modify them in a simple, fast and secure way. Enter the World of BOARD. Discover how easy it is to embrace this ALL IN ONE environment, where everything works perfectly and where the most advanced technology will provide you with the solutions you need.

Make the right decision with BOARD.

The new world of BI & CPM

www.board.com
data processor could easily come to the conclusion that the setup as such was targeted to build personalized records.

The above rules stem from the principle of proportionality. This principle is the cornerstone of data protection law. Big data analysis practices that go beyond reasonableness would be lawful only with the consent of individuals concerned. Procuring such consent may work in some circumstances (where the user will need to click through a certain form) but not in other circumstances. Asking for the data subjects’ consent in the abstract is not feasible, be it that not all of the potentially concerned individuals are known, be it that other individuals would not feel comfortable being contacted by companies on a very frequent basis.

Is it OK for an Organization to forward data to third parties for processing?

A company wishing to perform big data analysis is permitted to use subcontractors for analyzing purposes. Under the outsourcing friendly framework available under Swiss law such outsourcing setups can be implemented with relatively little administrative burden provided the outsourcing provider (the subcontractor) does not use the data made accessible to it for purposes other than as instructed by the customer. The subcontractor must accept audit rights and must commit to minimum security standards as required by data protection law. The company that uses an outsourcing provider should also, in enforceable agreements, require the subcontractors to commit not to re-identify the data.2

In the scenario outlined in this section, a notification to the user is not necessary. If data is shared with third parties and if the third parties are not restricted to reuse the data then a notification to the user is necessary.3

In light of the risk of re-identification:

Is it OK for an Organization to publish anonymized data?

Data protection law, from a Swiss law perspective, controls acts and omissions of individuals and companies that can be attributed to them. The mere release of an anonymized data set is not resulting in a breach of personality of anyone else. This is true even if the data set can later be de-anonymized by someone else. If, down the road, de-anonymization occurs then the publisher of the data set is not the malfeasor. Malfeasor is who has de-anonymized the data set. The publisher will not be looked at as a contributor to the breach, except of course in the event of conspiracy or gross negligence. To avoid negligence the publisher of data should make sure that data sets are anonymized according to the state of the art. Appropriate procedures should be applied.

Is it necessary to provide transparency to a user about what is being done with the data?

As discussed above, there may be instances where a user must be notified about data processing practices. Namely, if sensitive information is processed; if the privacy practices of the company do not appropriately identify the possibility of big data analysis (and if, as a result, it is not transparent to the data subject what is being done with its data); or if data are being disclosed to third parties. In such instances it is necessary to provide the user with more transparency. Otherwise, reaching out to individual users to explain the company’s privacy practices does not seem to be practical.

Is it necessary to obtain an individual’s consent? And if so: Whose consent?

As shown above, consent requirements are triggered if storage (longer than needed) and usage (no restrictions to the purpose identified when

---


5. Thouvenin (Fn. 3) 65.
collecting the data\textsuperscript{6}) are excessive. The problem is that consent cannot practicably be obtained. This is why big data collections in fact do risk being unlawful if they are the basis for too far reaching analytics.

Now, the problem cannot be solved by seeking consent. Or, to put it differently:

A setup relying on consent is hardly a very practicable setup. Given that the consent requirement is a consequence of overbroad analytics practices it would be more efficient to limit the analytics practices to what is still proportionate\textsuperscript{7}. In other words: Big Data analytics procedures should be made proportionate in all respects.

Correctness of data

Some writers discussing legal boundaries of big data argue that data sets easily turn inaccurate when combined with other data sets. This is correct. However, correctness of data is not something that could be responded to in the abstract. Data cannot be considered incorrect by default, not even in the context of Big Data. A data set must, however, be considered incorrect if a data subject raises an objection and claims a certain data set to be wrong. Then, the entity managing the database must correct such data or, if requested, delete the respective data set.

Secure Infrastructure

The DPA requires the manager of the database(s) in question to provide for appropriate data security measures. As big data collections are relevant from a personal data protection perspective each company engaging in big data analysis practices should make sure it has secure infrastructure in place that meets the minimum security level established by the law. This contribution is not the place to discuss those requirements in more detail. Entry point to the conversation about security measures is article 7 of the Swiss DPA (Data Security) which reads as follows: “Personal data must be protected against unauthorised processing through adequate technical and organisational measures. The Federal Council issues detailed provisions on the minimum standards for data security”.

Is it relevant what Value a Company will derive from the personalized data set?

For some companies it will be important to perform big data analysis in order to gain internal insights and to improve their business practices. Others will be able to commercialize, not the data, but the insights they have won on the basis of the big data analysis. Some claim that it would be unfair if individuals could not have a fair share of the gains a company won based on data of others. That may be true but we should also state that the Swiss DPA has not anticipated such argument. The Swiss DPA pursues the goal to protect an individual’s personality, not the individual’s commercial interests. This is why, absent unfair market practices, it is not relevant whether the company exploiting its big data know how makes money and draws other benefits, or not. This is, at least, the response from a data protection perspective.

Summary

Data protection law can find answers to questions raised when it comes to Big Data and data protection. Based on the law as it is currently in force, companies wishing to engage in big data practices can successfully proceed. A company engaging in big data practices will need to ensure the big data analysis remains proportionate in all respects. Whether the law appropriately protects the society is another question. We do see opportunities how the laws could be improved. But this does not mean that big data “does not work”, today. This translates as follows: Big Data analytics need diligent review but can be made compliant.

\textsuperscript{6} Thouvenin is not convinced the purpose limitation lawfully restricts reuse of data, Thouvenin, in Weber/Thouvenin, 77.

\textsuperscript{7} It is understood that “overbroad” and “proportionate” are open legal terms and, thus, unclear. But legal practitioners have learned how to build to fill these blanket terms with life. It would, however, go too far to explain in this article by what means these blanket norms can be filled, in practice.
ZURICH MACHINE LEARNING AND DATA SCIENCE MEETUP

A pub crawl in Paris marked my first encounter with a Meetup group. Such open events of interest groups have recently formed around the globe for various topics, covering knitting, language exchange over volleyball and various technology meetups.

In the area of machine learning and data science, I was inspired by the meetups in the San Francisco bay area, where the free events regularly draw an attendance of around 300 people. Why not do the same in Switzerland?

The goal of the Zurich Machine Learning and Data Science Meetup is to bring together practitioners, academics and just anybody interested in similar topics in a relaxed atmosphere. Our monthly events usually consist of two or three presentations, which could be about a cool application, new trends, some interesting research result, or even a tutorial on using open source tools. Last but not least we often conclude the events with an apéro for networking and follow-up discussions. Data science is inherently a very diverse and inter-disciplinary topic, split between several disciplines in universities, and also very fragmented in many small and large companies in industry.

Therefore, I think that the possibility of connecting similarly interested people can be particularly valuable in this area. As today marks the first year anniversary of our Zurich based meetup, a look back might be in order. Surprisingly, despite our relatively narrow focus of topics, we have now already become the largest technology meetup group in Switzerland. Our current 725 members are even out-numbering the Swiss Bitcoin enthusiasts, which were established earlier.

The highlight of our first year in the series was a simultaneous event jointly organized with our “sister” groups in Europe: the machine learning meetup groups in Paris, Berlin and London.

The main talk was given by one of the superstars of the field, Andrew Ng, co-founder of Coursera, former director of the Stanford AI Lab and now Chief Scientist at Baidu.

Over video-conference, we listened to a nice introduction to deep learning, and enjoyed a lively Q&A session afterwards. Apart from the possibility to attract more prominent speakers, such joint events also offer interesting opportunities to connect across the borders.

Dr. Martin Jaggi is a post-doctoral researcher in machine learning at ETH Zurich. Before that, he was a research fellow at the Simons Institute in Berkeley, US, working on the theory of big data analysis, and a postdoctoral researcher at Ecole Polytechnique in Paris, France. He has earned his PhD in Machine Learning and Optimization from ETH Zurich in 2011, and a MSc in Mathematics also from ETH Zurich, interrupted with several shorter stints in industry (Google, Netbreeze, Avaloq). He is broadly interested in methods for the analysis of large datasets, distributed training algorithms, open source software and machine learning applications for example in medicine, computer vision or text analysis.

Linkedin: ch.linkedin.com/in/martinjaggi

LINKS:
website: www.meetup.com/Zurich-Machine-Learning
youtube channel: www.youtube.com/user/zurichML/videos
While most of our events so far took place at ETH, we also explored other venues and were hosted by Google Zurich, by IBM Altstetten, and also at the University of Zurich. Another event which I found very enjoyable was the Data Science Connect evening in July, where three meetup groups teamed up for a nice evening hosted by IBM, gathering the Zurich Big Data Developers group, the Swiss Big Data User Group and our group.

We try to offer the recordings of all presentations freely on our YouTube channel.

With the entire organization being non-profit and fully volunteer based, I am always looking out for speakers, apero sponsors, suggestions about possible topics, or help with processing the videos for example. Please get in touch, and help us to improve the meetup!

MEMBERSHIP

Support the Swiss Association for Analytics and become a member!

The Swiss Association for Analytics is a non-profit organization. In addition to sponsors covering event and magazine expenses, we propose a paid membership of 30 CHF/year to support the association.

Membership benefits

Benefits of the Swiss Association for Analytics membership are the following:

– Become an official supporter of the association
– Receive the printed magazine home, twice a year (Swiss addresses only)
– Access presentations and pictures from association events

Membership is valid for a full calendar year (from 01.01 to 31.12). If you have paid before 01.01.2015, your membership is valid for 2015.

How to become a member of the Swiss Association for Analytics?

Step 1

Send the amount of 30 CHF, if possible through e-banking, to the following bank account:

Name: Swiss Association for Analytics
Address: Chemin des Fontannins 12, 1066 Epalinges
Bank: PostFinance
Account: 85-122140-9
IBAN: CH65 0900 0000 8512 2140 9
BIC: POFICHBEXXX

Step 2

Send an email to member@swiss-analytics.ch with the following information:

– Name (as in the above payment)
– Address where you would like the magazine to be sent (magazine is sent to Swiss addresses only)
– Email (needed to send you the password to access the member area on the website)

How will the money be used?

The membership fees will be used by the association to cover costs related to:

– Website domain name and hosting
– Magazine sending by mail
– Event or magazine expenses when no sponsor is found

IMPRESSUM

The Swiss Analytics Magazine is proposed by the Swiss Association for Analytics (www.swiss-analytics.ch). The magazine is printed by Gasser Media (www.gasser-media.ch). Cover image: courtesy of CERN (http://home.web.cern.ch). Articles are written by experts in analytics. The views expressed in this magazine do not necessarily represent the views of the Swiss Association for Analytics. It is not allowed to reproduce, duplicate, copy, sell, resell or exploit any portion of the magazine without agreement of the Swiss Association for Analytics (info@swiss-analytics.ch).

Printed in April 2015, Le Locle, Switzerland.