THE SWISS CENTER FOR BIOMETRICS RESEARCH AND TESTING

- Improving Information Quality enhances Business Performance
  PAGE 03

- Data Storytelling, next step for data visualization and BI tools
  PAGE 06

- Introduction into the legal aspects of Big Data (part #2: Ownership)
  PAGE 19

Sponsors:
WE ARE BACK!

I am very happy to introduce to you the second issue of the Swiss Analytics Magazine! When working on the first issue, the future of the magazine was not that clear: it all depended on its success. Since you are reading this second issue, you understand that the success was present. We had very positive feedback. In addition to the online version, 700 printed copies were distributed around Switzerland. We were thus very motivated to produce the current issue.

The most time consuming part for the first issue was certainly the layout, which was done completely manually. For this second issue, we have collaborated with Gasser Media for the magazine layout and printing. We are also here again thanks to our 3 sponsors: INFORMS, Itecor and Trivadis. You will find more information about them inside the magazine.

In this second issue, you will find original content about several data-related topics. Sébastien Marcel presents biometric projects at IDIAP (cover story). Philippe Nieuwbourg introduces the concept of data storytelling. Frank Block describes concepts behind data quality. Polly Mitchell-Guthrie explains INFORMS analytics certification. Dario Mangano illustrates the database leaf schema. Christian Laux continues to discuss the law aspects of Big Data. Sandro Saitta reviews the book *R in Action*. Finally, Christian Gügi presents his association, the Swiss Big Data user group.

In conclusion, we are back! And this is thanks to Gasser Media, our 3 sponsors and all readers and analytics passionate that spread the word about our magazine.

In the name of the committee
Sandro Saitta,
President of the Swiss Association for Analytics
IMPROVING INFORMATION QUALITY ENHANCES BUSINESS PERFORMANCE

Crucial information that drive important decision-making processes is often of poor quality and costs companies a fortune. We explain how Information Quality can be enhanced and how it contributes to company success in a measurable way.

Data and Data Quality – what is it? Why is it important? What makes it unique?

Data today is everywhere! There is so much of it that we now even call it BIG DATA! In this article we will not distinguish between information and data, which is a rather lengthy and theoretical discussion. The fact is that what for some people will be considered as information, for others will be considered as nothing more than data.

Data has become the new oil of our economy and is now the main ingredient to our information society. The rise of the Information Age has come hand in hand with the rise of IT capabilities and reduction of costs in terms of storage, speed, and transmission of data. The global connectivity and information exchange has further contributed to accelerate this development. Nearly every aspect of today’s private and corporate life is captured through digital data, which is processed and exchanged coming from credit cards, banks, stock exchanges, E-commerce sites, mobile telephones, internet behavior, etc.

A very important and new way of looking at data has emerged in the 90’s when Richard Wang (MIT) introduced the concept of the “data product”. Just like any

AGENDA OF EVENTS

Upcoming data-related events in Switzerland

SEPTEMBER

30th
Züri Machine Learning Meetup #8
place to be defined,
www.meetup.com/Zurich-Machine-Learning/events/197154132

OCTOBER

8th - 10th
Swiss Days of Official Statistics 2014
Yverdon-les-Bains,
www.statoo.ch/jss14

21st - 22nd
Jazoon’14
Zürich, www.jazoon.com/conference

NOVEMBER

7th (afternoon)
Analytics & Statistics for CRM
Hotel Mövenpick Lausanne,
http://event.swiss-analytics.ch

JUNE 2015

12th
2nd Swiss Workshop on Data Science (SDS|2015)
Eulachpassage, Winterthur,
www.zhaw.ch/de/zhaw/institutezentren/uebergreifende-institutezentren/dlab/sds2015.html

Want your future event to be listed here? Send an email to info@swiss-analytics.ch
other manufactured product, data products have to be designed, produced and their quality controlled.

The following image shows the analogy between traditional and data products. (fig.1)

This changed the view from looking at data as a simple by-product of business/manufacturing processes to being the main objective of a specific process. Thereby, data came to the center of focus and attention. Similar to traditional processing where Quality Assurance and Management have since long been part of the “culture”, Data Quality is only recently becoming part of data processing.

**Poor Information Quality hurts**

Many companies experience the presence of poor quality information for their daily operations, forcing them either to come to decisions based on “educated guessing” or putting them into a state of decision paralysis where “making no decision” is just as good. Even worse, some companies are largely optimistic about the quality of their data damaging even more seriously the company’s health.

According to “data guru” Tom Redman “The costs of poor quality are enormous. Some costs, such as added expense and lost customers are relatively easy to spot ... these costs are roughly 10% of revenue for a typical organization” (Tom Redman, Data Quality: The field guide, 2001).

Domains which are typically heavily impacted by Data Quality issues are the following:

**Financial Risk Management:** Recent years have been a showcase for the devastating impact poor data has on correctly assessing the financial risks behind people, companies and investment instruments. Regulations such as Basel II for financial markets have pushed the complexity of risk assessment models very far, making them extremely hungry for a lot of good quality data. Risk measures are often wrong as a consequence of incorrect and incomplete data.

**Customer Relationship Management (CRM):** CRM focuses on acquiring new clients, developing existing client relationships and maximizing customer lifetime value. To this end, detailed and high-quality customer information is required. The presence of poor customer information leads to the inaccurate estimation of present and future customer value, making assignment of the adequate sales and marketing budget on a customer level an impossible task. As a result, “bad customers” will be treated as good ones and “good customers” as bad ones, finally driving customer satisfaction down and customer churn up.

**Business Intelligence (BI):** Today, very powerful software packages enable organizations to analyze and distribute results in terms of reports or dashboards. Regrettably, poor quality information is often fed into these systems, resulting in inconsistent information and analysis results spreading like a virus across the organization, hindering decision making instead of enabling it. User confidence and adoption quickly drops putting previous massive technology investments at risk. Employees prefer to return to their “old data sources” and will spend most of their valuable time gathering, transforming and validating data.

**Curing the disease – Measuring and improving information quality**

In order to manage IQ we choose a simple definition: Information is considered to be good quality if it is fit for all its purposes. What does this imply? Knowledge workers (i.e. those who use data and information in their daily life to carry out their work) must define the purposes and the meaning of fitness; they define what information quality means.
This makes “IQ Management” a business topic.

Several IQ management approaches exist and have been successfully applied to large organizations i.e. in the banking and telecommunication sectors. They are often based on a multi-dimensional hierarchical IQ models focused on quality dimensions such as correctness, completeness, accessibility, ease of manipulation, interpretability, relevancy, security, confidentiality, and timeliness of information. Each of these approaches usually consists of several more or less complex steps. Simplifying, in general four steps need to be carried out:

Common phases in Data Quality Management methodologies

Measure Information Quality: To start, the IQ scope is defined in terms of the concerned information systems (e.g. data warehouses, CRM systems, risk management systems), which are then assessed in two ways. Firstly, a qualitative assessment of IQ is carried out based on the perception of the stakeholders, i.e. those who on a daily basis are confronted with data and their quality problems. To this end specially developed questionnaires and structured interviews are carried out with the stakeholders. Secondly, a quantitative assessment is carried out by means of data profiling techniques to detect suspect data patterns such as wrongly formatted phone numbers, future birthdates or missing values. As a result, the organization’s IQ landscape can be obtained to help understand where the major issues occur and where the business impact is the greatest. This will help focus on especially critical and poorly performing information systems and processes.

Analyze IQ Cause-Effect Chains: In this step data defects are linked with the organization’s key performance indicators (KPI). Root-cause-analyses are now carried out to understand the origins of previously identified IQ issues. Powerful impact model can then be derived to quantify which information quality indicators impact key performance indicators and how strongly. This ideally gives the cost of poor data quality in terms of economic and risk. Cost can be direct cost caused by the presence of data defects but also opportunity costs arising from the reduction of business opportunities as a consequence of data defects. For example a poorly populated birth date customer data attribute will worsen the targeting capabilities of marketing campaigns and will reduce their returns.

Build IQ Roadmap – Prepare for Data Governance: Now, a detailed list of IQ improvement actions quantifying the
required investments and estimated benefits (KPI) is produced. Based on this, business teams assign a priority for each action. This leads to the IQ Management Roadmap, which is now totally driven by business impact and RoI considerations.

**Execute and Monitor IQ Roadmap – Information becomes a key asset:** Executing the IQ Roadmap should be seen as “standard project work”, ideally under the umbrella of the company’s Data Governance organization. Reiterating the whole process guarantees continuous improvement of IQ.

**Enhancing Information Quality generates tangible benefits**

Measuring and managing Information Quality as presented here assures that information is managed, as it should: as one of the company’s main assets. It maximizes the leverage effect of information and its contribution to the company success in a measurable way. Its unique and innovative approach includes the information consumer (knowledge worker, employee) as a key player and quantifies very quickly the business impact of IQ issues.

As a result many benefits are obtained such as enhanced risk management (reduced legal and reputational risk, enhanced fraud detection), cost reductions (less costs for detection and correction of information problems, avoiding penalty payments due to non-compliance with regulations, better productivity due to lower idle times), revenue assurance (less lost collections, reduced customer churn), and higher employee satisfaction (reinforced organizational trust, fast and accurate decision-making, powerful forecasting, consistent management reporting).

As a positive side-effect, the presented approach is ideal for giving Data Governance a sense right from the very start, by arming it with highly relevant insights and a ready-to-execute action plan to have an enterprise impact in a minimal timeframe.

---

**DATA STORYTELLING, NEXT STEP FOR DATA VISUALIZATION AND BI TOOLS**

The next evolution of Business Intelligence and Performance Management tools is to become data storytelling tools. Tomorrow, your company will probably hire a data storyteller. Let me tell you this story.

A long time ago in a secret boardroom far, far away... John Doe, the promising young CFO of USA Inc. gained a promotion, telling his CEO the story he wanted to hear, about the quarterly results of his company, and the strategic plan he proposed to enhance corporate performance. After years in a famous business school, John Doe learned data visualization, and added a special skill called data storytelling or how to tell stories based on corporate data.
Data Storytelling: mixing knowledge on data visualization and communication

For some people, the concept of storytelling refers to a “dirty” practice. Something related to politicians, communication “artists”… people with more style and less content than “real” professional people. But, stories or narratives have been shared in every culture as a means of entertainment, education, cultural preservation, and instilling moral values.

As available data increase, as Big Data seems to be a key concept for business models evolution, CFOs needs to embrace this topic and learn how data can bring new value to their job. The 3 “V” defines Big Data: Volume, Velocity, and Variety. But, I often add another one, a V for Visualization, as the more data you have, the more complex it is to summarize them and bring them to your audience. And I end my own formula by a fifth V about Value, as every Big Data project should clearly be funded only after having identified the value it will bring to your company. (Volume + Velocity + Variety) x Visualization = Value, is my magic formula for the data century we have entered in. To explain the value, you need technics and methods. And these methods are the same than used for years in marketing and communication, but applied to finance and performance management.

Why CFOs needs to tell stories

CFOs are not sad people locked up in their offices, frowning while creating PDF reports full of insane figures! They are smart people loving to create comprehensive charts and to share their results and strategy with others. They have to! As a parent, are you telling the same story to a kid, a teenager, or a young adult? No!

In 2014, CFOs have to speak in front of board, executive committee, financial analysts, press, and shareholders... You can’t tell the same story to all these people.

As a movie, the story of your corporate data has to be organized to keep the audience focused. As Steve Jobs said in his Stanford speech in 2005: creating a story is “connecting the dots” between your data. A good story begins with a pitch. Before working on your presentation, you will write, just for your eyes, a sentence of four lines to explain “what?” “How?” and “What for?” The next step is the narrative plan: an initial situation, a troublemaker, events, a solution and a settlement, it’s known as the Freitag’s pyramid. During your story, you will alternate between “what is” phases and “what could be” phases. “What is” phase helps you to stay connected to your audience. They will accept the next “What could be” phase. They trust you because you tell them things they know and agree (“What is” phases).

Let’s take an example! Here is the pitch of a movie you will certainly recognize: “It’s the story of a scientist who builds a resort in which old species of dinosaurs are recreated from their old DNA, to make them live again and be the heroes of a this entertainment theme park.” First part of the pitch explains us “what” the movie is about; second part “how” the dinosaurs will be recreated; third part, the “what for”, the objective of the park founder. Of course, it’s the pitch of Jurassic Park. Applied to a professional situation, like the meeting to highlight last quarter results to the board, the pitch could be “Our Company publish strong results this quarter, thanks to a large customer we won in China for our top selling brand, and these results will help us to reimburse our bank loan quicker than expected”.

The narrative plan will alternate different phases:

• You will begin with the summary of your last meeting, acted decisions... what people knows already
You will give the good news, the strong results you have to share with your board. It’s a high intensity moment.

Perhaps you will continue with a lower intensity list of your top selling products.

And announce the huge effect of the Chinese customer on your quarterly results.

After that, you will come back to things your audience know, the amount of debt your company have…

And finally propose to use part of the quarterly results to reimburse the most expensive loan you have with one bank.

In front of financial analysts, you will maybe keep the same pitch, but will change your narrative plan replacing the confidential list of your top selling products by an overview of your competition; and the details on your debt, by a comparison between your debt and the cash dividend you serve to your shareholders since 5 years.

Five rules of Data Storytelling

Even if it can be difficult to change your habits, you should:

• Not try to be the star of your presentation: your audience is!

• Turn off your computer: That’s not when you use leftovers that you can create an original recipe! Don’t open your slide’s fridge to create your new presentation.

• Enjoy! You have to love your ideas and to love your audience!

• Don’t tell everything at the beginning. You have to keep a piece of mystery… for the end

• To be or not to be? That is NOT the question! Occupy the space, the room, and the scene… Occupy audience eyes, ears… You have to BE here, with them!

New tools are coming

As CFOs more depends on software for collecting, storing, and analyzing data, it’s normal that this new idea of telling stories around data, find a declination in the software industry. CFOs are Excel and PowerPoint addicts and no business intelligence tool succeeded in replacing Excel. But, presenting results could really be improved. Today, CFOs take data from Excel or BI tools, create graphics, or report, and copy-paste into PowerPoint. Conclusion: no re-use is possible from a presentation to another, no interaction with data, no automatic update. All the process needs to run again for every presentation. Yellowfin was the first BI tool to include a presentation function that allows you to add slides inside reporting. Data are automatically updated, and the presenter can interact with them and answer questions without leaving the presentation.

Tableau Software invested a lot in Data Storytelling and version 8.2 added data points. A data point is a view of data, which can be stored, like a dynamic and interactive screenshot. Data points can be combined into a storyboard to create the all story. In Microsoft PowerView, the same concept of storyboard is used to copy-paste moments of your analytic process and create a story. A French BI tool, Coheris Liberty, just launched a new version including similar combined views to create stories. And Qlik just released Qlik Sense, which includes a data storytelling module, close to Tableau’s. Domo is another business intelligence tool to follow as they enter the data storytelling space.

In the Corporate Performance Management (CPM) market the most advanced tool seems to be Tidemark. In Tidemark you can create storylines that
are infographics dedicated to financial KPIs. And storylines can be gathered into Playbooks that help you to create an automated and interactive corporate report, based on the story you want to tell to your audience.

CPM and BI tools will all include functions to create stories. You will be able to combine narration and interaction, with tools that will include audio and video recording. In many occasions, they will replace PowerPoint. If you want to follow this emerging market, you can find an updated cartography of data storytelling tools, available for free at www.datastoryteller.com/carto/. (fig.1)

...a couple of years later, our former young CFO, John Doe, became communication professional, adding storytelling knowledge and tools on his desk. And professors in business schools added a data storytelling for finance course to their program.

Philippe Nieuwbourg is an independent analyst focusing on Business Intelligence and Big Data. He publishes the blog www.datastoryteller.com about Data Storytelling.

Follow him on Twitter @nieuwbouрг
Big Data & Data Privacy with Itecor

Big Data for a small world

IT leaders from various companies ranging from local entities to the public sector to large multinational all trust Itecor in the Big Data field.

What drives them is their need for a Swiss Cloud partner. They found us searching for an agile, reliable and local partner.

Do you need cutting-edge technologies solutions and data scientists to implement them?

Strengthen your BI to leverage better your business opportunities using distributed computing.

Our Big Data service? Providing solutions in:

- Data analysis in a multidimensional world
- Best practices to evolve faster
- Distributed computing development
- Infrastructure provisioning and running

Our services are developed around four practices put in place in different industries having different challenges. Each dimension is represented by one face of the Itecor cube.

www.itecor.com
THE SWISS CENTER FOR BIOMETRICS RESEARCH AND TESTING

Among traditional methods of identification, biometrics stand out as a favourable solution today for identity management problems of large populations with increasing mobility. In a well-designed system, biometric security ensures that only authorized persons can access to the protected facility or information, because it assesses a person’s most unique physical and behavioural features that can be practically sensed by devices and interpreted by computers.

Biometric security also provides convenience to the users as their obligation to carry a badge or remember a PIN is eliminated. In fact, frequent users of the internet and mobile devices are required to create and remember multiple passwords. Truly secure passwords are long, complex and ideally changed on regular basis. However in reality, most users have few passwords that they use repeatedly for various applications and long periods of time.

On the other hand, biometric systems offer an attractive alternative – improving security while reducing demands on the user. Consequently, security solutions that depend on biometric patterns of users, such as fingerprint, facial structure, voice or gait are becoming increasingly widespread.

Driven by the biometrics passports that are currently in use in many countries, more and more biometric-enabled applications are employed in daily life, mainly for login to personal electronic devices via fingerprint (iPhone 5S Touch ID, Fujitsu Lifebook) or face (Samsung Galaxy 4 Face Unlock, HP Face Recognition) recognition. (fig.1)

Although the market for biometrics based identity management is dominated by some key players (e.g. Safran Morpho, Cognitec) that focus mainly on high security applications in business-to-business market, new exploitation routes for business-to-consumer market are currently explored by small and medium sized enterprises (e.g. KeyLemon, Visidon, Mobbeel, Facebanx, BioID) that address to computer login, universal login for social networks or cloud computing solutions. Moreover, these routes are also undertaken by innovative companies as demonstrated by their strategic moves, such as the acquisition of Face.com by Facebook in 2012, the acquisition of PolarRose by Apple in 2010 and the acquisition of NevenVision, PittPatt, and Viewdle by Google respectively in 2006, 2011 and 2012.

Despite this stimulating and rapidly growing market, a crucial security issue is still to be considered by most companies: vulnerability to attacks. It has been shown recently that conventional biometric techniques, such as fingerprint or face recognition are prone to one of the most potent and damaging threats involving personal data – identity fraud, mostly known as spoofing.

1. Face recognition on an iPhone 4
Spoofing

Spoofing, also referred to as presentation attack, is a direct attack performed at the sensor level outside the digital limits of the system. Therefore, no digital protection mechanisms can be employed against it. In an attempt to spoof a biometric system, an intruder tries to masquerade as a valid user by forging a fake biometric sample and presenting it to the sensor to be captured. Anti-spoofing refers to the measures to detect and avert these attempts. Unfortunately, it is unclear that anti-spoofing modules are present in commercial biometric authentication products today, possibly placing personal security at high risk. This vulnerability is currently raising a lot of attention at the international standardization level. (fig.2)

Trusted Biometrics under Spoofing Attacks

In the face of all these developments and in the hope to provide clear and concise information regarding the vulnerabilities of a range of biometric systems to spoofing attacks, a large consortium of academic and industrial partners launched a research project named Trusted Biometrics under Spoofing Attacks (TABULA RASA), funded by the European Commission under the Framework Programme Seven (FP7).

The goal of this project is to research, develop, evaluate and transfer anti-spoofing solutions. For this purpose, an inventory of spoofing attacks is created against various existing systems and their vulnerabilities are evaluated.

Finally, appropriate countermeasures are developed and transferred to companies to be integrated into their biometric systems.

This work depends primarily on designing and collecting databases for the analysis of spoofing attacks and measuring the success of the designed countermeasures. For instance, as one of the academic partners, Idiap Research Institute has collected face spoofing databases using printed pictures, photos on mobile phone/tablets and 3D masks and utilized those databases to develop countermeasures based on texture analysis and motion estimation.

Biometrics Evaluation and Testing

With the collaborative efforts of its partners, BEAT aims to create one unified document detailing the future
evaluation of biometrics, as well as to enable the realisation of accurate and repeatable comparisons between different biometric technologies by providing an online platform (the BEAT platform) and associated toolbox.

The BEAT platform will need to be a hub for transparent, free and accessible information as a prerequisite to employ BEAT as a bridge between research community and industry partners and an academic tool for the facilitation of data swapping, result verification and education. (fig. 3)

3. Main page of the online BEAT platform

The data provided by TABULA RASA regarding existing and potential spoofing attacks and the international consensus being fostered under the BEAT project represent a significant blow to the ambitions of data criminals. However, the plans and ability of potential attackers remain largely unknown. New forms of attacks may arise for any of the modalities investigated and it is still unclear how the developed countermeasures will generalize. Hence, it is clear that preventative efforts must continue – ideally, in a collaborative and international environment.

In view of this and to serve as the legacy of the EU projects TABULA RASA and BEAT, the Swiss Centre for Biometrics Research and Testing is created with the support of the Swiss State of Valais and the Idiap research institute that will be hosting the Center.

The Swiss Centre for Biometrics Research and Testing will aim to facilitate collaboration between academia via its members.
Its core missions are:
• to carry out high quality research
• to train talented engineers and researchers
• to foster technology transfer from academia to industry
• to propose evaluation and testing services in the domain of biometric security.

Being split between partners (active researchers) and affiliates (industry partners), the center ensures an environment of mutual benefit. While the members create research proposals for funding consideration according to the affiliates’ needs, the affiliates fund and drive the research carried out among the members. This structure helps industry partners to gain unparalleled access to cutting-edge researchers and their expertise and supports researchers to acquire the opportunity to bid for funding from driven and relevant organisations.

**Researchers – Become a partner!**

The establishment and maintenance of a thriving research environment is an impractical task without these three pillars of support for research funding, technology transfer and career development. The Swiss Center for Biometrics Research and Testing brings these three pillars under its roof and helps to bridge the gap between fundamental research and commercial applications by providing a hub for researchers and companies to interact and collaborate.

Becoming a partner to the Center offers unparalleled benefits and is easily possible by referring to two companies for affiliateship. Individual researcher participants from academic institutions or non-for-profit research institutions that are associated with the Center by means of the partnership agreement have the right to propose research projects to affiliates for funding in a time and cost efficient manner with much less risk. Additionally, they gain a huge opportunity to put their research findings into real-world practice.

**Companies – Become an affiliate!**

In today’s fast changing industrial environment, design and development of new technologies and commodities is crucial. Firms must keep pace with the continuous evolution and development of technology. To this end, they have to invest in research and development activities to discover new scientific or technological knowledge for advancement of their processes and services.

Unfortunately, these activities mostly do not yield immediate profit and involve higher uncertainty with respect to other company activities. This is truer for industrial fields driven by cutting edge technology, such as biometrics business.

In order to overcome such adversity, the Swiss Center for Biometrics Research and Testing offers biometrics and security companies the means to build strong ties with talented researchers and engineers in the field and to make recommendations and influence the course of their research projects by becoming an affiliate.

Affiliates have right to delegate representatives to the Affiliate Advisory Board which is designated to meet on annual basis and make recommendations on the research projects to be carried out within the Center and the allocation of resources to research proposals. Additionally, they are granted the right to obtain worldwide, royalty-free, non-exclusive, perpetual licenses to the scientific results obtained within the projects funded by the Center.
BOOK REVIEW: R IN ACTION

R has recently been ranked as the 9th most popular programming language by the IEEE. It is widely used due to several reasons: latest machine learning algorithms are implemented, it is easy to prototype applications and it is free. I discovered the author of R in Action – Data Analysis and Graphics with R, Robert Kabacoff, through his famous website, Quick-R (www.statmethods.net). If you are programming in R, you will inevitably reach his blog when searching for information on the Internet. I found the content of his blog clear and detailed. These two words summarize very well his book.

R in Action covers several topics such as R programming, regression, missing values, ANOVA, missing data, graphics and bootstrapping. Kabacoff’s book is neither a tutorial, nor a reference manual. The author takes time to explain statistical concepts and then present the way to solve them in R. The book is full of advices. Functions are also well explained using schema (for example for the melt/case functions). The code is clearly documented. At the end of the book, you will be surprised to have learned not only R, but also statistics.

One point to be noted about R in Action (although it is clear in the title): this is not a book about data mining, nor machine learning. It won’t go into discussing how to solve a data mining business problem or how to implement a machine learning algorithm. However, if you’re serious about R, you should really consider Kabacoff’s book. Also to be highlighted that the author has made efforts to solve the same problem in various manners (with different packages). R in Action is, according to me, the reference book for R. Don’t hesitate, you can go for it!

SWISS BIGDATA USER GROUP

The Swiss Big Data user group (SBDUG) is the largest user group for Big Data related topics in Switzerland. Founded in April 2012, the group aims to bring together people from all industries who are interested in dealing with large amounts of data, to offer interesting talks, hands-on advice and a forum for exchange and networking. Topics include but are not limited to: Apache Hadoop, NoSQL storage, search technologies or any other Big Data related technology - presented from both a technical and a business perspective.

It is a non-commercial, private initiative, founded and led by Christian Gügi with currently 430+ members. The SBDUG is sponsored by YMC AG and meets on a regular basis, usually every 2-3 months.

So far, most of the events took place at the ETH in Zurich. Nevertheless, event locations are flexible and events can take place throughout Switzerland and can also be provided by companies. The audience at these informal events is quite diverse, ranging from IT professionals, data analysts/scientists to business people. On average around 70 people from numerous renowned companies as well as from startups come to a meeting.

For more information about the group or past talks please visit www.bigdata-usergroup.ch or www.slideshare.net/SwissHUG/presentations

Christian works as a Big Data Solutions Architect at YMC.

If you are interested in giving a talk or becoming the sponsor of a future meeting please contact: christian.guegi@gmail.com.
BENEFITS OF CERTIFICATION
- Advances your career potential by setting you apart from the competition
- Drives personal satisfaction of accomplishing a key career milestone
- Helps improve your overall job performance by stressing continuing professional development
- Recognizes that you have invested in your analytics career by pursuing this rigorous credential
- Boosts your salary potential by being viewed as experienced analytics professional
- Shows competence in the principles and practices of analytics

APPLICATIONS
- Prepare to apply by reviewing Candidate Handbook & Study Guide Draft
- Arrange now to secure academic transcript and confirmation of “soft skills” to send to INFORMS

COMPUTER-BASED TESTING
It is now more convenient than ever to schedule your CAP exam in more than 700 Kryterion test centers in more than 100+ countries. To find the location closest to you, check this site: www.kryteriononline.com/host_locations/

QUESTIONS? certification@mail.informs.org

**DOMAINS OF ANALYTICS PRACTICE**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
<th>Weight*</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Business Problem (Question) Framing</td>
<td>15%</td>
</tr>
<tr>
<td>II</td>
<td>Analytics Problem Framing</td>
<td>17%</td>
</tr>
<tr>
<td>III</td>
<td>Data</td>
<td>22%</td>
</tr>
<tr>
<td>IV</td>
<td>Methodology (Approach) Selection</td>
<td>15%</td>
</tr>
<tr>
<td>V</td>
<td>Model Building</td>
<td>16%</td>
</tr>
<tr>
<td>VI</td>
<td>Deployment</td>
<td>9%</td>
</tr>
<tr>
<td>VII</td>
<td>Life Cycle Management</td>
<td>6%</td>
</tr>
</tbody>
</table>

*Percentage of questions in exam 100%
THE LEAF SCHEMA

A Star is Dead

I wanted to start this article by bumptiously saying that the Data Warehousing fathers stopped their thoughts too early, and finally I decided to tell you not less bumptiously that I invented a new concept that will replace definitely the star schemas...

Of course I am not serious! I will only try here to explain that there is better model to host your presentation layer than a traditional Star Schema. But let’s start with some reminders.

Dimensional Modeling

Remember, a Star Schema is made of a central Table containing the facts and surrounding table called Dimension.

Loading a Star Schema is not very difficult, but a certain sequence must be followed.

A Leaf Is Born

Remember the Data Vault Modeling Concepts... Only three types of tables: Hubs, Links and Satellites... Hubs and links containing only the keys. The satellites are split by rate of change.

Let’s take a look to a very simple example: The relation between a customer and a product.

The Customer could look to something like this:

And if we decided to change a little the way we designed our satellites in our Business Data Vault Layer?

This is what I propose:

Let's put all the information destined to be part of a future dimension in dedicated satellites, and let’s create a special satellite with the measures destined to go to a fact table.

Now that we have all the future facts in separates satellites. Imagine the same process for the Products Satellites...

And here is the new concept: try to imagine a dimension made of small dimensions, with the same key.

And yes, it is working!
Your final Data Mart will be a collection of Leafs as shown in the figure below.

Now we can see that the process to transform satellites into Leafs and Hub and dedicated satellites into Facts is very simple and most important it is automatable.

**Virtualization**

If you are looking for performance, you will probably start to think about hosting your presentation layer on an In Memory appliance.

Most of Business Intelligence providers are offering solutions to put your Data Mart in memory and to use virtualization techniques in order to gain in development time and performance.

SAP Hana, Oracle Exadata, IBM Netezza are among the most important players in this domain.

With the Leaf Schema concept, you can imagine to use a view on your RAW Data Vault Model to build virtual Data Marts. It is up to you to decide if you prefer to have a physical Data Mart or a virtual one.

**Dario Mangano** just started a new challenge as CIO of the biggest event management company in Switzerland: Palexpo. He previously was head of the Knowledge Management Department of Nestlé Nespresso, based at Nestlé Nespresso Headquarters in Lausanne Switzerland. In this role, he lead the Business Intelligence, Enterprise Content Management and Global Data Services departments. Driven by the motto “From Data, To Information, To Knowledge, To Action!” he was responsible for defining and implementing the strategy to transform internal structured and unstructured data into meaningful information and knowledge to support strategic decision-making.

Dario Mangano holds a PhD degree in computer sciences and a license in group dynamics. He is also the founder of the Swiss Network of Business Intelligence Professionals, Owner of the largest Swiss BI Event: The Swiss BI Days and BI Awards and is the author of the book: The Integrated Data Hub™. He is the inventor of the Leaf Schema concept.

Twitter: @DarioMangano
Linkedin: ch.linkedin.com/in/dariomangano
Book: www.createspace.com/4066552
INTRODUCTION INTO THE LEGAL ASPECTS OF BIG DATA
(PART #2: OWNERSHIP)

Introduction
In part #1 of this introduction², we have laid out some thoughts to the definition of Big Data. While the definition of what is Big Data may not be supported by everyone, everyone will agree that businesses can gain huge benefits from controlling and/or processing data. So, they turn to commercial data providers to purchase data respectively to have it processed, that gives them those insights. The problem is: The law defines processing of data but has no direct rule available to determine what „purchase“ means, in this situation. Purchasing a car, a loaf of bread, or a book – those transactions will give the purchaser ownership in the car, bread or book. How is it with data? This question touches on the very grounds of what is data law. We begin with some definitions and then tackle ownership in data.

Definitions
Data, Information and Knowledge
The law indeed has a number of definitions for information or data⁴, but is lacking a uniform definition for information or data. The definitions we use here are as follows:
- **Data**: In common language, the term “data” is used in two ways. First, data can be described to mean information that is stored in a format that can be processed by a machine. And second, the term is generally used to refer to a single entry of a “data collection”. This is the meaning used in Swiss criminal law, for example.
- **Information**: Information is the meaning we aim to extract from something. There is hardly any better definition than the following: Information is “the it from the bit⁶”⁷. When the Swiss Data Protection Act (DPA) uses the term “data” it refers to the content (i.e. to information) and not to the format of an entry (“data”).
- **Knowledge**: There is not a lot of unanimous theory in what knowledge would be. We understand knowledge as a higher level of combined information, or better: Information put in context.

Or, to put those terms in context: “Data as an abstract concept can be viewed as the lowest level of abstraction, from which information and then knowledge are derived.”⁵

Categories of Data
In the context of the Big Data conversation, not only information in spreadsheets is data. The term “data” includes:
- **structured data**, i.e. data that are organized in a well-defined manner, e.g. in databases or other files (examples are data collections organized in spreadsheets or tables, or maps, i.e. guiding information organized alongside certain concepts);
- **semi-structured data**, e.g. markup languages (e.g. XML). Markup languages contain self-describing structures, i.e. fields to separate semantic elements and enforce hierarchies of records and fields within a data set. In standard database applications, a basic principle is the distinction between the schema (that describes the structure of the database) and data (the database instance).

1. This contribution is the second 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. Now, we want to illustrate this when discussing what ownership in data is. Following this first overview, the legal landscape is analysed in this issue in more detail. Later, specific fields of special regulation should be identified. Dealing data quality aspects can then be discussed, same as aspects of how one can be liable for the first step that is relevant for Big Data, the disclosure of data sets to the public.
2. Laux, Introduction Into The Legal Aspects of Big Data, Swiss Analytics Magazine 2014/01, 15 et seq.
3. Swiss law has a data definition in criminal law in order to determine that an outsider can be punished if he accesses information stored on a automated processing system by circumventing specific protection measures. The Swiss Data Protection Act (DPA) has a (different) definition of data to determine the rights an individual may have in information describing him or her.
Semi-structured data one can note that the distinction between the schema and the data itself gets blurred;
• unstructured data, e.g. documents containing text or graphic.
These categories may become relevant to understand in more detail how the law treats information, and data.

Intellectual Property Laws as a Point of Reference for Ownership respectively Control of Information or Data

Generally, ownership means having “title” in a thing. An owner has the ability to determine how the thing is being used. Ownership thus translates into control.

The limitations of ownership operate well in a brick-and-mortar world. In the law, ownership is defined for things like bread, a book, a car, a plough or a cow. The thing itself defines the boundaries of the control. Now, we live in the information age. The problem is that ownership in information or data is not defined in the law. With respect to ideas (or more precisely: “information”), intellectual property laws confer a property-like control over certain forms of information:
• a patent means control over a technical teaching that is new, useful and not obvious (in view of prior art). The technical teaching (i.e., the inventive idea) is information, described in a patent application that needs to be submitted to a competent authority (registration). Unlike in copyright (control over a defined embodiment), a patent controls the information as such (limitations in time and jurisdiction apply). Free flow of information kicks in outside the technical teaching, after lapse of the protection period (commonly, 20 years) or outside the jurisdiction for which a patent has been granted. Control is the result of a formal disclosure (a registration) made with respect to specific information (a novel and useful “technical teaching”).

The policy rationale, i.e. the reason why control is given, is novelty6.
• a copyright means control over information that has been embodied in a concrete and creative creation (e.g. in a book, in a movie, in music, in painting, in a statue, in photographs and in software). Control is limited to reproductions (or equivalents) of that very precise creative embodiment. Outside the creative embodiment, ideas are not restricted (free flow of information). Control is the result of a “creative embodiment” having been added to information. The policy rationale behind copyright law is creativity7.
• an industrial design means control over the form of the appearance or of the style of an industrial object (e.g. spare parts, furniture or textiles). Similarly as with copyright, control is the result of a specific embodiment8. The reason why control is given is novelty.
• a mask work means control over a series of related images used to produce an integrated circuit (also called a topography). The images represent the predetermined three-dimensional pattern of the layers constituting the integrated circuit. Similarly as with copyright, control is the result of a specific embodiment (a series of images)9. The reason why control is given is novelty.
• a trademark means control over a distinctive sign10 used to distinguish the products or services of one business from those of another business. If the sign is not used in commerce, control once obtained will be lost. Control is the result of distinctive information used in commerce following registration. The reason why control is given is distinctiveness11.

After all, one can summarize that control in information is given to the holder of an

---

6. The policy rationale is best visible in the United States constitution, that describes why the United States Congress has powers to legislate with respect to patents and copyrights: “To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries” (Article I, Section 8, Clause 8).

7. Creativity can be identified as a special form of novelty.

8. It being noted that the requirements are quite different in design law, one of them being the necessity of a registration with a competent authority.

9. It being noted that the requirements are quite different in topography law, one of them being the necessity of a registration with a competent authority.

10. Of course, a sign is information as it means something (see above).

11. Distinctiveness can be identified as a special form of novelty.
intellectual property right if “something new” has been added to the information in question (novelty, creativity, distinctiveness), either through a specific embodiment (copyrights, designs, mask works), a specific disclosure (patent law) or through use in commerce following registration (trademarks). Without such a qualification, information is not protected. In such an event, the fallback applies, being free flow of information.

Two further categories of intellectual property exist, database rights (solely in the EU) and rights in trade secrets. Those are specific as the means and effects of protection are limited in a number of ways compared to the “classical” intellectual property rights introduced above. The database right given to the producer of a database is most relevant in the present context, though. It prohibits to extract substantial portions of content from a protected database. Therefore, the database right can influence the way how one can use data.

To sum up: We have discussed the above to understand who has what rights in big data. The rule is that big data (information) per se is not protected except if someone else can claim a right in a piece of the big data set. The baseline is “free flow of information”. Free flow of information translates into “nobody owns but anyone may use”, i.e. “public domain”.

The following is a visualization of this rule:

In the chart, the blue waves can be read to represent what we have referred to above as free flow of information. Information, once available, can be reused by anyone who has access to it. “The author tells a story, it then is in the reader’s head, and the author cannot take it away from there.” That is what free flow of information means. Intellectual property law protection is in the black section (in the field of black pluses). “Pluses” because additional factors must be met so there is control over information to which access has been given. These additional factors have been described above: embodiment, disclosure, use in commerce, each with respect to something that is novel (as described above).

One observation can be added: In Switzerland, structured data (i.e. databases) are very often less protected than unstructured data such as texts, movies, music, etc. The latter are covered by copyright, thus there is a good level of protection for such types of data. Contrary to copyrighted products, according to constant case law of the Swiss Federal Court, databases that have been made publicly accessible can be spidered to a large extent. It would not be permitted to display verbatim copies of such databases, but inserting the spidered data in new products in own products is largely permitted.

Now: What Ownership in Information Means

Freedom to operate or control, both can be a motivation to ask the question of who owns information in big data. In this article, we focus on a freedom to operate perspective.

Freedom to Operate

If information contained in a big data collection is subject to intellectual property rights someone can control reuse of the information in question. In the context of big data: Aggregations of data must therefore be analyzed in light of whether third parties can claim intellectual property rights. Also it must be

---

12. This would be different under EU law, where the database rights given to a producer of a database would prohibit such copying.
checked whether other prohibitions or restrictions apply, like personality rights or contracts. From a freedom to operate perspective: Reuse of a big data collection must be cleared with the third party rights owners before the big data collection can be freely exploited, provided, of course, that exploitation falls within the scope of what the owner of the intellectual property right can prohibit. The following activities seem to be typical in the context of Big Data:

- **Extracting Meaning** ("analytics"): Extracting meaning from a data collection rarely involves the copying of data (what could be prohibited by copyright law), and **reading alone is not an infringement** from a copyright perspective. So, usually no infringement of copyright law. It is diligent to ask whether the use (or reuse) of information for a specific purpose is actually covered by patent law. However, the fact that information that is being analyzed has been disclosed in a patent application does not, per se, mean that an analysis of the overall data sets will be infringing on that patent. It is more likely, if at all, that the method applied to extract meaning touches on a method patent (which, in turn, is not depending from the information that is being analyzed). As a consequence, there is not much risk exposure for him or her who limits himself or herself on the extracting of meaning from available data sources.

- **Reproduction**: However, reproducing data that are covered by third party rights almost always is critical. He or she who intends to aggregate data for reuse purposes may easily violate copyright laws or patent laws (prohibited reproduction).

- **Public display**: Displaying aggregated data collections in public (e.g. on the world wide web) will almost always violate third parties if such are present.

From a freedom to operate perspective it is not enough, however, to ask whether a third party has an intellectual property right in information. Any third party influence on the data set in question should be tracked. In order to identify those third party interests, the following should be taken into account:

- **Personally identifiable data**: Does the data set in question contain personally identifiable data? If so, a data subject could claim rights it has under the data protection act. Should the data set consist of personal information, then it may be necessary to notify the data subject (i.e. the individual described by such data set) of the analytical activities with respect to such data set.

- **Regulations**: Is there specific regulation to be complied with? Specific regulations will need to be complied with, if applicable. Regulation can relate to the persons described by the data (banking law, attorney client relationship) or the subject matter described by the data (geographical information, meteorological information, all as made available by public authorities). Further, national security interests may have an impact on reuse of data.

- **Contractual limitations**: Are there contractual rights that may hinder the exploitation of the data set in question? If the data scientist has received the information under specific contractual limitations (e.g. a non-use obligation that may be contained in a non disclosure agreement) he or she must obey to those in order to avoid sanctions from the data provider.

To be clear: The above describes the legal analysis in case a data scientist accesses information that in fact is available. If the information the data scientist wishes to use is not accessible, he or she first needs to get access. There are a number of legal remedies under which a data scientist can claim access to certain information (e.g. by

---

13. Data protection laws will be discussed in the next issue of this series.
means of a “Freedom of Information Request”, the data scientist can obtain a vast variety of information from public authorities. If data remains locked, however, it is a criminal act, for example, to circumvent access protection measures.

Control Perspective

If someone wants to be the owner of information (e.g. a platform operator aggregating user data may wish to state his ownership) then it should be asked whether he can claim any of the intellectual property rights outlined above. Where none of the qualifications (embodiment, disclosure, use in commerce, each with respect to something that is novel as described above) apply, ownership in a data set (or aggregations of data) must be determined otherwise. In such a scenario, and on the basis of the example given (platform operator) the following is relevant:

• Access control: Can the platform operator access to information? As a rule, the answer is yes: Data is stored on the operator’s servers and not accessible to the public.

• Reuse control: Can the platform operator control reuse? Yes. For so long that no one else can access the data it is the platform operator alone who can reuse the data. If a third party is given access to the information the platform operator should do so only under restrictive agreements stating clearly what obligations the recipient has with respect to the information in question.

• Deletion control: In such agreements, the platform operator will need to impose clear obligations on the recipient at what conditions, and by when, the recipient must delete the information.

• Compliance: If the platform operator can ensure compliance in all his dealings with such data (usage, forwarding) it can make sure that no third party actually has a say in what it does. This means: No third party intellectual property rights in the data; no data subject having a data protection claim or any other personality right; full compliance with potential regulations. If no third party and no authority can raise concerns then the information is free from any lien towards a third party.

All those steps are relevant to determine whether the holder of a data set can be declared to have ownership in that data set.

Outlook

In the next issue, we will dig deeper to analyse how the laws of personal data/data protection laws may impact big data.

Dr. Christian Laux, LL.M., is 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 officers 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.
Bedeutet Big Data auch Big Kosten?


www.trivadis.com | info@trivadis.com
Certification has roots in medieval craft guilds, which established standards, set ethical guidelines for practice, and functioned as professional associations. The Institute for Operations Research and Management Science (INFORMS) is the largest professional society in the world for analytics, operations research, and management science. Our Certified Analytics Professional (CAP®) Program is a new effort to establish standards and set guidelines. And certainly the practice of analytics is a craft, requiring special skills in its execution. Let me explain, and maybe by the end you’ll be interested in getting certified yourself.

**INFORMS, a modern-day guild?**

INFORMS is certainly not a medieval guild, but it is a global organization of 11,000 members, of whom 50% are academics, 30% practitioners, and 20% students. Members are highly educated – 96% have or are earning an advanced degree, and 51% have PhDs. We have taken on the championing of analytics in many ways, including establishing continuing education programs; an analytics maturity model; engagement with industry, government, and academia; and certification.

Our decision to establish a certification program shares goals with the guilds: to advance the craft by introducing standards of quality, identify individuals with breadth of knowledge, and encourage continued competency. While guilds were quasi-governmental in requiring membership to practice that craft (akin to governments requiring licenses for certain professions today), certification is a voluntary credential granted by a non-governmental body. We conducted a Job Task Analysis (JTA) study by a panel
of subject matter experts to define the knowledge, skills, and abilities necessary to effectively practice analytics. They identified seven domains: business problem framing, analytics problem framing, data, methodology selection, model building, deployment, and model lifecycle management.

What is required for certification?

There are five “E’s” of the Certified Analytics Professional (CAP®) program: education, experience, exam, effectiveness, and ethics. The education and analytics-related work experience combination required is three years for those with a master’s degree (any guess where that term originated?) in a related discipline (e.g. statistics, mathematics, computer science, etc.), five years with a bachelor’s degree in a related discipline, and seven years with a degree in an unrelated discipline (waivers of the educational eligibility requirements will be considered on a case by case basis). The exam tests skills and knowledge in the seven JTA domains with 100 multiple choice questions that must be completed in three hours or less. Computer-based testing is available at 700+ locations worldwide (including five in Switzerland).

Analytics requires more than technical skills and knowledge, so a confirmation of effectiveness in the area of soft skills by an employer or client is required. And, CAP® requires agreement with the code of ethics. For more details on applying see the CAP® Candidate Handbook and the Complete CAP® Study Guide.

Why get certified?

The target group is early to mid-career professionals, or “apprentices” seeking to become “journeymen,” although thankfully today apprentices are paid! Those advanced in their careers may have a resume whose mastery stands on its own, but early career professionals may find the differentiation offered by a portable, independent, third-party validation of knowledge of certain standards helpful. Not unlike the journeyman certificate or letter, in this era of increasing demand for analytics professionals certification attests to knowledge and skills beyond a person’s own word. It demonstrates commitment to career and to craft and adherence to accepted standards. It can increase earnings potential and provide the personal satisfaction of achieving a milestone. INFORMS and CAP® may be less “explosive” than Sechseläuten festivities, but these benefits will endure for years to come.

Polly Mitchell-Guthrie is the Senior Manager of the Advanced Analytics Customer Liaison Group in SAS’ Research and Development Division, where her team serves as a bridge between R&D and external customers and internal SAS divisions. Polly serves as Vice Chair of the INFORMS Analytics Certification Board. She has an MBA from the Kenan-Flagler Business School at the University of North Carolina at Chapel Hill, where she also received her BA in Political Science.
SWISS ANALYTICS EVENTS

By Sandro Saitta, President of the Swiss Association for Analytics.

On June 19th took place the 4th Swiss Analytics Event. The topic was “Analytics & Forecasting”. It was held in Hotel Mövenpick Lausanne, with more than 30 people (who preferred analytics to the World Cup). We had 3 excellent speakers. Lionel Moret (Forecaster at MeteoSwiss), Georg Kropat (PhD student at UNIL) and Ilkka Huotelin (Independent Consultant). We would like to thank our sponsor, PwC, who allowed us to organize this event.

Our next event will be a joint effort of the Swiss Association for Analytics (SAA) and the Swiss Statistical Society (SSS). This free afternoon event will take place November 7th and the theme is “Analytics & Statistics for CRM”. Our keynote speaker is Dean Abbott, Chief Data Scientist, Smarter Remarketer, Inc. We will also have speakers from Nestlé, Aduno Gruppe, SAS, DiaSys and SandSIV. The event is sponsored by PwC, SAS and IMSD. More information and free subscription at event.swiss-analytics.ch

The Swiss Association for Analytics will also take this opportunity to organize its constitutive General Assembly on November 7th. It is free and accessible to anybody interested in analytics. Schedule and further details will be communicated through the LinkedIn group (www.swiss-analytics.ch)

Looking forward to meeting you there.

The Committee

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 IDIAP (www.idiap.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 September 2014, Le Locle, Switzerland.