INTERNET OF THINGS AND ANALYTICS

Value added Business Processes through Predictive Analytics

Asymov, Psychohistory and Predictive Analytics

Meta-mining: a new meta-learning framework

Sponsor:

Sas
### TABLE OF CONTENTS

| STORY & EXTRAS : CONCRETE APPROACH TO ADD VALUES TO BUSINESS PROCESSES THROUGH PREDICTIVE ANALYTICS | 3 |
| AGENDA OF FUTURE EVENTS | 3 |
| ASIMOV, PSYCHOHISTORY AND PREDICTIVE ANALYTICS | 5 |
| ANALYTICS AT THE EDGE : EXAMPLES OF OPPORTUNITY IN THE INTERNET OF THINGS | 7 |
| META-MINING : A NOVEL META-LEARNING FRAMEWORK FOR AUTOMATIC DATA MINING WORKFLOW PLANNING AND OPTIMIZATION | 10 |
| SWISSTEXT 2016 : AUTOMATIC TEXT UNDERSTANDING IN SWITZERLAND | 16 |
| BOOK REVIEW: “DATA SCIENCE FOR DUMMIES”, BY LILLIAN PIERSON” | 17 |
| MEMBERSHIP | 19 |

---

**SPRING RELEASE**

This year started out strong with already two events and our second General Assembly. This assembly held the first change of Presidency since its creation, where Sandro Saitta, staying in the committee, passed on the role to Amine Mansour who was elected by the committee.

It is now already the fourth year of existence for the Swiss Association for Analytics and more than 10 events and 5 magazines, and this success is thanks to you! Analytics is at the heart of multiple key challenges of our society and its application is always very diverse. Therefore it is the perfect timing for another issue of our magazine. Another set of articles dedicated to diverse subjects, in addition to many posts on our LinkedIn group. We thank you all for being active and for your continuous support.

In the following pages you’ll find an article from Davis Wu on a Concrete Approach to Add Values to Business Processes through Predictive Analytics. Followed by another from Sandro Saitta on Asimov, Psychohistory and Predictive Analytics, SAS and Fiona McNell provide us with a view on the opportunities of the Internet of Things and finally Phong Nguyen writes on Meta-Mining a novel meta-learning framework to automate part of the data mining process!

Common columns include the agenda for a few upcoming events, membership information and a book review by Amine Mansour: “Data Science for Dummies”, by Lillian E. Pierson. Enjoy your reading!

A friendly reminder to those who wish to support the association. Please join us as an official member of the Swiss Association for Analytics. You’ll find necessary information on our website:

www.swiss-analytics.com/membership

You’ll receive our magazine at home and have access to event presentations.

---

Amine Mansour
President of the Swiss Association for Analytics
Senior Consultant and Big Data Subject Matter Leader at Itecor

www.swiss-analytics.com
STORY & EXTRAS: CONCRETE APPROACH TO ADD VALUES TO BUSINESS PROCESSES THROUGH PREDICTIVE ANALYTICS

Author: Dr. Davis Wu, Global Lead of Demand Planning, Nestlé S.A.

From Frankfurt to São Paulo, from Toronto to Kuala Lumpur and Bangkok, in 2015 I have visited many Nestlé companies in different countries. My main mission was to support them in developing predictive analytics capabilities, mainly the use of SAS Forecasting Solutions, to improve their demand planning results, as well as deeper collaboration between Supply Chain, Sales, Marketing and Finance. The program is under the name of Future of Demand Planning (FDP) in Nestlé.

Two questions often came up in the discussions, workshops and trainings during the visits:

Demand Planner: ‘how can I make my views accepted in the Monthly Business Planning discussions?’

Business Executive Manager (head of business unit): ‘what is in it (FDP) for me?’

The first question is about the credibility and quality of work Demand Planners should bring to the Monthly Business Planning discussions.

The second question is about the added value that Future of Demand Planning (FDP) can bring to cross-functional collaboration.

The notion that ‘FDP can improve efficiency in Demand Planning, and Demand Plan Accuracy & Bias’ is insufficient in answering these questions.

Something more concrete and tangible is needed for the demand planning community to visualize the benefits of FDP program and take practical steps to materialize them.

Here is how and we focus on two simple but powerful key words: STORY and EXTRAS.

STORY:

Story is about what’s behind the numbers that demand planners are presenting in Monthly Business Planning meetings. Much more than the numbers themselves. Demand Planners must be able to articulate what makes up a forecast. The facts, the quality and rigor of analysis that they have put into it. This will help address the first question ‘how can I make my views accepted in Monthly Business Planning discussions?’

The Story comes with 3 key elements:

1. Historical Analysis

The trend of base demand (everyday sales), hidden pattern of seasonality, historical performance of promotions are some of the factual results that Future of Demand Planning program (FDP) can bring out through advanced statistical modelling.
2. Activity Plans

The timing and mechanics of future activity plans, incremental estimates based on past similar events, are among those that can be statistically modelled in the Forecasting Solution, based on the inputs typically communicated from Key Account Managers, Customer Managers and etc.

3. Assumptions

New Product launches, customer or distribution gains or losses, cannibalization effects are examples of elements that must be factored in as part of the total forecast. These are often not statistically modelled but provided by Marketing and Sales managers as assumptions.

The forecast should be presented, often in a series of plots, to suit the needs and audiences in Monthly Business Planning discussions, e.g. in both volume and value, appropriate level, such as Category, Key Customers or Strategic SKUs.

EXTRAS:

Extras is about what additional value Demand Planners can bring through Future of Demand Planning program (FDP). Value in the eyes of the Business Executive Manager, the Business. Something that demand planners have not previously been able to provide in our current capacity. This links to the second question asked by Business Executive Managers ‘What is in it (FDP) for me?’

The below diagram captures the 3 key areas that are now made possible as a result of the use of advanced analytics in Demand Planning.

1. Confidence Corridor

All forecasts presented must come with a range within a certain confidence level. No forecast is a set number. It is an estimate within a range. This is now standard with forecasts generated from the Forecasting Solution. The Confidence Corridor is represented as a +/- band.

For low level forecasts (e.g. SKU/ Customer) the Confidence Corridor is relatively wide. For more aggregated forecasts (e.g. Category, sub-Category) the confidence corridor will be much tighter.

The concept helps the business to avoid setting a forecast to a specific target. It also helps Demand Planners justify/protect their numbers.

2. Insights

There are rich insights waiting to be untapped as a result of advanced modelling in the Forecasting Solution. This is especially so when causal modelling is used for predicting promotions/event incrementals or impacts of any cyclical activities.

The Forecasting Solution Nestlé uses has functions, such as Parameter Estimate Table, that provide sources of information about the relativity of promotion mechanics, seasonal impacts, media contribution to volume, impacts of month-end or quarter-end push. All of them provide Demand Planners invaluable insights that have not been available to them before. These insights add tremendous value to cross-functional collaboration and help sustain the credibility of Demand Planners.

3. ‘What-if’ Scenarios

Visualize that Demand Planners sit with Key Account Managers to interactively
model scenarios such as ‘What would be the volume impact if the mechanics or timing of certain events change?’ This is now entirely possible with advanced analytical capability in the ‘Scenario Planning’ function in the Forecasting Solution. This gives business planning a revolutionary capability to assess new possibilities on how to close a gap or create new competitive advantage.

These three ‘Extras’ give the Demand Planning community unprecedented capabilities to collaborate with cross-functional teams.

Although the amount of ‘extras’ could vary widely depending on the granularity of planning levels and availability of data to support advanced modelling, this new capability is entirely possible in both modern trade and distributor environments.

Future of Demand Planning program (FDP) is a journey in Nestlé to fundamentally improve the way of forecasting, collaborating with cross-functional teams, adding values to Monthly Business Planning processes. Benefits are already being shown in many Nestlé companies around the world. More to expect in coming months and years.

ASIMOV, PSYCHOHISTORY AND PREDICTIVE ANALYTICS

By Sandro Saitta, Manager, Data Science, Expedia

“You don't need to predict the future. Just choose a future -- a good future, a useful future -- and make the kind of prediction that will alter human emotions and reactions in such a way that the future you predicted will be brought about. Better to make a good future than predict a bad one.”

Isaac Asimov, Prelude to Foundation

If you like hard science fiction with stories that evolve over thousands of years and detailed characters, then you should read Asimov1. In particular, the Foundation series. How is this related to Predictive Analytics? The Foundation series is based on the concept of Psychohistory, which is the study of the future, based on maths, at the level of an entire population2.

Hari Seldon, the main character at the beginning of the series, is a mathematician specializing on Psychohistory.

He can predict the future but only at a large scale. He will use his knowledge to prepare Humanity to recover from an unavoidable war which will destroy nearly everything (I will let you read the series to know more about the story).

---

1 http://en.wikipedia.org/wiki/Isaac_Asimov
2 http://en.wikipedia.org/wiki/Psychohistory_(fictional)
The first question is: was Asimov a precursor in Predictive Analytics?

There is a major difference between Psychohistory and Predictive Analytics: the target scale. Whereas Psychohistory focuses on predicting behaviour at a population level, Predictive Analytics is applied at an individual level. Obviously, objectives are not the same. Predictive Analytics is driven by applications such as marketing and fraud detection. Psychohistory is used to study the evolution of an entire civilization.

The next question is: which of the two disciplines is the toughest? According to Pedro Domingo, Psychohistory is harder than Predictive Analytics. In his book The Master Algorithm, he writes:

“In Isaac Asimov Foundation, the scientist Hari Seldon manages to mathematically predict the future of humanity and thereby save it from decadence. […] According to Seldon people are like molecules in a gas, and the law of large numbers ensures that even if individuals are unpredictable, whole societies aren’t. Relation learning reveals why this is not the case. If people were independent each making decisions in isolation, societies would indeed be predictable, because all those random decisions would add up to a fairly constant average. But when people interact, larger assemblies can be less predictable than smaller ones, not more.”

This is certainly the reason why Predictive Analytics is currently used, while Psychohistory remains science fiction as of today. Asimov proposed two main axioms for Psychohistory, and their relations to Predictive Analytics are worth discussing:

Axiom 1 – “The population whose behaviour was modeled should be sufficiently large”

Axiom 2 – “The population should remain in ignorance of the results of the application of psychohistorical analyses”

On the first axiom, Daniel Zeng, writes in a recent editorial of IEEE Intelligent Systems that it is becoming “irrelevant” in a world of Big Data. The second axiom is also valid for specific Predictive Analytics applications. In fraud detection, for example, if the fraudster knows the model used to discover him, he will change his behaviour accordingly.

In conclusion, although Psychohistory is fictional, it shares common aspects with Predictive Analytics. Psychohistory, although still science fiction, is definitely a key research area and actors from Wall Street and Governments would pay a fortune to predict the future of an entire country, over a long period of time. Wouldn’t you?

---

1 Professor and author of The Master Algorithm (http://tinyurl.com/zxqqfow)


4 Foundation and Empire, the second book of the Foundation Series.
ANALYTICS AT THE EDGE:
EXAMPLES OF OPPORTUNITY IN THE INTERNET
OF THINGS

By Fiona McNeill, Global Product Marketing Manager, SAS

The data streaming in and out of organizations from electrical and mechanical sensors, RFID tags, smart meters, scanners, mobile communications, live social media, and more results in staggering volumes of information. When all these sources are networked to communicate with each other — without human intervention — the Internet of Things (IoT) is born.

The IoT market is estimated to include nearly 26 billion devices, with a “global economic value-add” of $1.9 trillion by 20201 and nearly $9 trillion in annual sales by 20202. By all accounts, IoT is a new type of industrial revolution.

But to derive useful knowledge from the tide of streaming source data — and participate in this new economy — you must have analytics.

In traditional analysis, data is stored and then analyzed. But with streaming data, analytics must occur in real time, as the data passes through. This allows you to identify and examine patterns of interest as the data is being created. The result is instant insight and immediate action.

So before the data is stored, in the cloud or in any high-performance repository, the event stream is automatically processed. And using analytics to decipher streaming data as close to the device as possible creates a new realm of knowledge for many industries. Let’s look at a few examples.

The Internet of Things in health care

In health care, analyzing IoT data can result in increased uptime for machines that treat cancer, which means that patients are treated when they are scheduled. If a treatment time is missed, it can be up to 40 percent less effective, so reducing service interruptions is critical. By monitoring hundreds of sensors, identifying issues early and proactively correcting them, service personnel armed with the necessary information and parts arrive together.3 Elekta, a Swedish company that provides equipment and clinical management to help treat cancer and brain disorders, has cited a 30 percent reduction in site visits because of such monitoring.4

With rising global populations and corresponding increases in disease and health care costs, the remote patient monitoring market doubled from 2007 to 2011 and is projected to double again by 2016.5

And what if these scenarios went beyond monitoring device status or patient conditions — to predicting machine reliability in advance of parts beginning to malfunction? Servicing would then move from being proactive to being optimized across each supplier’s landscape of devices. And foreseeing patient problems

3 Todd DeSisto, Opening Session, Axeda Connexion Conference, Boston, MA, May 6, 2014.
before they even experience symptoms could avert adverse events altogether.

Event streams that know more than just existing conditions, and which evaluate future scenarios using advanced analytics, are now within the realm of possibility.

How do you apply predictive capabilities to IoT data? High-performance analytics environments are designed to examine complex questions and produce models. These algorithms are then coded into the data streams, along with any data normalization and business rules to detect patterns associated with the defined future scenarios. So in addition to monitoring conditions and thresholds, you can use the data stream to assess likely future events.

The Internet of Things in manufacturing

The automobile industry is stepping up development detection systems for imminent collisions to determine when to take evasive action. Based on radar and other types of remote technologies, driving conditions are monitored to assess – and ultimately avoid – collisions. These collision avoidance systems assess the likelihood of a collision event and automatically prescribe mechanical changes to the vehicle if the driver doesn’t respond – including deceleration and external lighting changes. Potential accident reduction from wide deployment could surpass $100 billion annually in savings.

In fact, the “Industrial Internet” – which combines physical machinery, networked sensors and software – has extensive use and promise in manufacturing, including production optimization, product development and aftermarket servicing. GE predicts $1 trillion in opportunity per year by improving how assets and resources are used and how operations and maintenance are performed within industrial industries.

The Internet of Things in energy

A detailed view of energy consumption patterns is needed to understand energy usage, daily spikes and workload dependencies. And beyond just manufacturing – lighting alone takes 19 percent of the world’s electricity. Optimized alternate energy sources can have a significant impact on all sectors.

---

7 A term coined by GE.
For example, a single blade on a gas turbine can generate 500GB of data per day.\textsuperscript{10} Wind turbines constantly identify the best angles to catch the wind, and turbine-to-turbine communications allow turbine farms to align and operate as a single, maximized unit.

Historically, the only way to know what was happening with a turbine – even if it was on and working – was to climb 330 feet and see. Remote monitoring provides new eyes on the status of these energy generators.

What if the same data could be used to forecast? Efficiency in green energy means we can store more energy for use when the wind is low. Predicting when excess energy is available can help determine when to charge batteries, for example, further extending the efficiencies of alternate energy sources.

Of course, the energy market provides one of the most well-known examples of IoT technology altering the customer landscape. With dynamic smart meter billing, customers have new choices, which lead energy companies to adopt a more customer-centric approach. The utility smart grid transformation is expected to almost double the customer information system market, from $2.5 billion in 2013 to $5.5 billion in 2020.\textsuperscript{11}

**The Internet of Things in retail**

Customers are also at the center of IoT analytics in retail, where some companies are studying ways to gather and process data from thousands of shoppers as they journey through stores. This “in-store geography” informed by sensor readings and videos considers how long shoppers linger at individual displays, recording what they ultimately buy.

With the goal of optimizing store layout, these data points can also be tied to smart-device Wi-Fi networks. In addition to appropriately targeting shoppers for promotions in-store, retailers can ask customer opinions – using IoT data to initiate an interaction, customizing the shopping experience and enhancing loyalty.

**Taking action with IoT data**

Event streams monitor patterns of interest. Sensors and devices generate lots of data that describes existing conditions. Analysis of conditions informs what actions are necessary – either immediately as an alert notification, or with pre-planning from predictive and other advanced analysis methods.

Of course, analysis always leads to more questions – directing what additional sensors (and data) can be collected to measure new aspects of the conditions, elements of the event or more detail in the scenario to understand different patterns.

IoT data, by itself, isn’t the value. Just as with traditional data sources, it’s the ability to take the insights and then act on them that provides value. To know what to do in the moment, use analytics at the edge.

For more information on the Internet of Things and Event Stream Processing go to www.sas.com/iot

---

\textsuperscript{10} GE Software, “System 1™ and Proficy Smartsignal™ Keep Watch.”

\textsuperscript{11} Navigant Research, “Electric Utility Billing and Customer Information Systems.”
META-MINING: A NOVEL META-LEARNING FRAMEWORK FOR AUTOMATIC DATA MINING WORKFLOW PLANNING AND OPTIMIZATION

Data Mining (DM) or Knowledge Discovery in Databases (KDD) refers to the computational process in which low-level data are analyzed in order to extract high-level knowledge (Fayyad et al., 1996). This process is carried out through the specification of a DM workflow, i.e. the assembly of individual data transformations and analysis steps, implemented by DM operators, which composes the DM process with which a data analyst chooses to address her/his DM task. Standard workflow models such as the CRISP-DM model (Chapman et al., 2000) decomposes the life cycle of this process into five principal steps: selection, pre-processing, transformation, learning or modeling, and post-processing. Each step can be further decomposed into lower steps. At each workflow step, data objects are consumed by the respective operators which either transform them or produce new data objects that flow to the next step following the control flow defined by the DM workflow. The process is repeated until relevant knowledge is created, see Figure 1.

Despite the recent efforts to standardize the DM workflow modeling process with a workflow model such as CRISP-DM, the (meta-)analysis of DM workflows is becoming increasingly challenging with the growing number and complexity of available operators (Gil et al., 2007). Today’s second generation knowledge discovery support systems (KDSS) allow complex modeling of workflows and contain several hundreds of operators; the Rapid-Miner\(^1\) platform, in its extended version with Weka\(^2\) and R\(^3\), proposes actually more than 500 operators, some of which can have very complex data and control flows, e.g. bagging or boosting operators, in which several sub-workflows are interleaved.

As a consequence, the possible number of workflows that can be modeled within these systems is on the order of several millions, ranging from simple to more elaborated workflows with several hundred operators. Therefore the data analyst has to carefully select among those operators the ones that can be meaningfully combined to address her knowledge.

---

\(^1\) https://rapidminer.com/
\(^2\) http://www.cs.waikato.ac.nz/ml/weka/
\(^3\) https://cran.r-project.org/
discovery problem. However, even the most sophisticated data miner can be overwhelmed by the complexity of such modeling, having to rely on his/her experience and biases as well as on thorough experimentation in the hope of finding out the best operator combination.

With the advance of new generation KDSS that provide even more advanced functionalities, it becomes important to provide automated support to the user in the workflow modelling process, an issue that has been identified as one of the top-ten challenges in data mining (Yang and Wu, 2006). During the last decade, a rather limited number of systems have been proposed to address this challenge. In the two next sections, we will review the two most important research avenues: the first one takes a planning approach based on an ontology of DM operators to automatically design DM workflows, while the second one, referred as meta-learning, makes use of learning methods to address, among other tasks, the task of algorithm selection for a given dataset.

**Ontology-based Planning of DM Workflows**

Bernstein, Provost, and Hill (2005) propose an ontology-based Intelligent Discovery Assistant (IDA) that plans valid DM workflows – valid in the sense that they can be executed without any failure – according to basic descriptions of the input dataset such as attribute types, presence of missing values, number of classes, etc. By describing into a DM ontology the input conditions and output effects of DM operators, according to the three main steps of the DM process, pre-processing, modeling and post-processing, see Figure 2, IDA systematically enumerates with a workflow planner all possible valid operator combinations, workflows, that fulfill the data input request. A ranking of the workflows is then computed according to user defined criteria such as speed or memory consumption which are measured from past experiments.

Zakova, Kremen, Zelezny, and Lavrac (2011) propose the KD ontology to support automatic design of DM workflows for relational DM. In this ontology, DM relational algorithms and datasets are modeled with the semantic web language OWL-DL, providing thereby semantic reasoning and inference to query over a DM workflow repository. Similarly to IDA, the ontology characterizes DM algorithms with their data input/output specifications to address DM workflow planning. The authors have developed a translator from their ontology representation to the Planning Domain Definition Language (PDDL), with which they can produce abstract directed-acyclic graph workflows using a FF-style planning algorithm. They demonstrate their approach on genomic and product engineering data mining.
(CAD) use-cases where complex workflows are produced which can make use of relational data structure and background knowledge.

More recently, the e-LICO\(^4\) project featured another IDA built upon a planner which constructs DM plans following a hierarchical task networks (HTN) planning approach. The specification of the HTN is given in the Data Mining Workflow (DMWF) ontology, (Kietz, Serban, Bernstein, and Fischer, 2009). As its predecessors the e-LICO IDA has been designed to identify operators which preconditions are met at a given planning step in order to plan valid DM workflows and does an exhaustive search in the space of possible DM plans.

However, none of the three DM support systems described here consider the eventual performance of the workflows they plan with respect to the DM task that they are supposed to address. Moreover, they tend to deliver an extremely large number of plans, DM workflows, which are typically ranked with simple heuristics, such as workflow complexity or expected execution time, leaving the user at a loss as to which is the best workflow in terms of the expected performance in the DM task that she needs to address. To address this problem, we have proposed a novel meta-learning approach, aptly called meta-mining, which we will describe in the following section.

**Meta-learning and Meta-Mining**

There has been considerable work that tries to support the user in view of performance maximization for a very specific part of the DM process, that of modeling or learning. A number of approaches have been proposed, collectively identified as meta-learning or learning-to-learn (Brazdil et al., 2008; Kalousis, 2002; Hilario, 2002; Soares & Brazdil, 2000). The main idea in meta-learning is that given an unseen dataset the system should be able to select or rank a pool of learning algorithms with respect to their expected performance on this dataset; this is referred as the algorithm selection task (Smith-Miles, 2008). To do so, one builds a meta-learning model from the analysis of past learning experiments, searching for associations between algorithm’s performances and dataset characteristics.

In the work of Hilario, Nguyen, Do, Woznica, and Kalousis (2011), we proposed a novel meta-learning framework that we called meta-mining or process-oriented meta-learning applied on the complete DM process. We associate workflow descriptors and dataset descriptors, applying decision tree algorithms on past experiments, in order to learn which couplings of workflows and datasets will lead to high predictive performance. The workflow descriptors were extracted using frequent pattern mining accommodating also background knowledge, given by the Data Mining Optimization (DMOP) ontology, on DM tasks, operators, workflows, performance measures and their relationships. However the predictive performance of the system was rather low, due to the limited capacity of decision trees to capture relations between dataset and workflow characteristics that were essential for performance prediction.

To address the above limitation, we presented in the work of Nguyen, Wang, Hilario, and Kalousis (2012b) an approach that learns heterogeneous similarity measures, associating dataset and workflow characteristics. These similarity measures reflect respectively: the similarity of the datasets as it is given by the similarity of the relative workflow performance vectors of the workflows that were applied on them; the similarity of the workflows given by their performance based similarity on different datasets; the dataset-workflow similarity based on the expected performance of the latter applied on the former.

\(^4\) http://www.e-lico.eu
However the two meta-mining methods described here were limited to select from, or rank, a set of given workflows according to their expected performance, i.e. they cannot plan new workflows given an input data set.

Retrospectively, we presented in the work of Nguyen, Kalousis, and Hilario (2011) an initial blueprint of an approach that does DM workflow planning in view of workflow performance optimization. There we suggested that the planner should be guided by a meta-mining model that ranks partial candidate workflows at each planning step. We also gave a preliminary evaluation of the proposed approach with interesting results (Nguyen, Kalousis & Hilario, 2012a). However, the meta-mining module was rather trivial, it does a simple nearest-neighbor search over the dataset descriptors to identify the most similar datasets to the dataset for which we want to plan the workflows. Within that neighborhood, it ranks the partial workflows using the support of the workflow patterns on the workflows that perform best on the datasets of the neighborhood. The pattern-based ranking of the workflows was cumbersome and heuristic; the system was not learning associations of dataset and workflow characteristics which explicitly optimize the expected workflow performance, which is what must guide the workflow planning. In the next section, we will present our meta-mining system which overcomes all the limitations described so far to address the task of planning DM workflows with respect to a performance measure.

**A Meta-Mining System to support DM workflow planning and optimization**

In Nguyen, Kalousis, and Hilario (2014), we follow the line of work we first sketched in the work of Nguyen et al. (2011). We couple tightly together a workflow planning and a meta-mining module to develop a DM workflow planning system that given an input dataset designs workflows that are expected to optimize the performance on the given dataset. The system with its components is illustrated in Figure 4. As first component, we have the Data Mining Experiment Repository (DMER) which contains all past DM experiments and meta-data (MD) from which we learn heterogeneous similarity measures, associations between dataset and workflow descriptors that lead to optimal performance, according to our method described in Nguyen et al. (2012b). Then we have as second component the Meta-Miner that exploits the learned associations to guide the AI-planner in the workflow construction during the planning of DM workflows for the input dataset. This last is provided by the User Interface, our third component, together with a DM goal such as classification.
The Meta-Miner and the AI-planner then tightly interact together to form our IDA (fourth component) where at each planning step the planner submits a set of partial candidate workflows that are ranked by the Meta-Miner in order to prioritize those workflows that are expected to deliver the best performance on the input dataset. Finally in the fifth component the system delivers after planning a set of optimal plans ranked by order of their expected performance from which the user can select which one to apply on her dataset. We evaluate the system on a number of real world datasets and show that the workflows it plans are significantly better than the workflows delivered by a number of baseline methods.

To the best of our knowledge, our meta-mining system is the first system of its kind, i.e. being able to design DM workflows that are specifically tailored to the characteristics of the input dataset in view of optimizing a DM task performance measure. We believe that it is a promising approach to support DM practitioners and industries in their daily tasks. Future work includes refining the DMOP ontology to support more DM algorithms as well as combining meta-mining with reinforcement learning to update our DM knowledge by shifting new DM problems in our system.

About the Author
Phong Nguyen holds a PhD in machine learning and data mining from the University of Geneva. He is currently Senior Data Scientist in Expedia working on sort algorithms and recommender systems. This article is the first one of two articles describing the meta-mining system proposed in the author’s PhD thesis. The second article will follow in the next issue of the Swiss Analytics Magazine and will describe in more details the proposed system.

Bibliography
Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. AI magazine, 17(3):37, 1996.


SWISSTEXT 2016: AUTOMATIC TEXT UNDERSTANDING IN SWITZERLAND

SwissText is a one-day conference on automatic text analytics, which takes place June 8, 2016 in Zurich. SwissText will bring together practitioners and researchers and give an overview of existing solutions and technologies in automatic text understanding/natural language processing/computational linguistics.

The conference will cover the most interesting topics that are relevant for real-world applications: for instance how to detect positive or negative messages on Twitter, what is state-of-the-art in machine translation, how to extract entities such as company and person names from arbitrary texts, or which types of texts can be generated automatically.

Broad Support by Swiss Community

SwissText has broad support in the analytics community: it is organized by the Datalab of Zurich University of Applied Sciences (ZHAW), more than 10 Swiss universities and research associations are official partners of the conference, and it is supported and funded by CTI, the Swiss Federal Commission for Technology and Innovation, and by SpinningBytes, a private company for data analytics.

Keynotes will be given by three distinguished experts in the area: Katja Filippova, NLP expert at Google Zurich; Prof. Paolo Rosso, head of the Natural Language Engineering Lab in València; and Jürg Attinger, innovation mentor of CTI, who will give hands-on tips how to get federal funding for innovative products and services.

Presentations for the Audience

The conference will see three types of presentations: Surveys will give a broad overview of state-of-the-art technologies and solutions in text understanding; Showcases will present successful projects; and Open Problem Presentations, where practitioners from the industry present “their” open problem.

Get Involved

Participation fee for SwissText 2016 is CHF 80.- and includes admission to all sessions, lunch and after-conference apero. Registration is now open until end of May 2016.

For more information on SwissText 2016, visit the conference website at www.swisstext.org.

About the author:
Mark Cieliebak
Conference Chair of SwissText 2016
(Other roles: Lecturer at Zurich University of Applied Sciences (ZHAW) CEO of SpinningBytes AG)
BOOK REVIEW: “DATA SCIENCE FOR DUMMIES”, LILLIAN PIERSON

By Amine Mansour

Some of you must know Lillian Pierson due to her involvement in the Data Science and Big Data domain, being a data scientist and data science journalist/author. In this sense, it was a pleasure for me to review this book. I am always looking for new resources to introduce people to the “buzzy” domain of Data Science and share a little bit of what I do with them, without hassling them with too much of the technical side. How to do it best than with a “dummies series” book?

Right at the start of the book, expectations are set: you will see a lot of content in a rather concise book. However, quality is there and most important aspects of Data Science are present: Probability and Statistics, Clustering and Classification, Visualization and business concepts and more. Furthermore, when people go through the learning curve of Data Science and have no prior experience in similar fields, it is a capital mistake to start with pure Data Mining/Data Science technical books, but it is also a mistake to start with pure business oriented literature. What is precious in this book is that it takes you (of course not through extreme details) to the basics of essential skills such as Probability and Statistics.

However, as a corollary, with the objective to cover as many technical aspects and segments of Data Science as possible, a slight weakness can be noted: Not focusing enough on more thoroughly developed data science business cases. Of course it is present in the book and as an example Part 5 shows four real-world applications, from Data Science in Journalism to Predicting Crime-Rates, but many of today’s most discussed cases could be shown to target a broader audience. Also the author does assume that readers must have prior technical background, and focuses heavily on programming which is a bit contradictory to the purpose of making Data Science accessible to a broader audience.

In the end, this is still a very good book that I recommend to those who wish to learn more about data science and more particularly those who are eager to see some technical concepts behind theory, but this is only an introductory book and if you like the concepts, you should add to it a pure business oriented book and a pure technically oriented one.
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 at home, twice a year (available for addresses in Switzerland 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 01.01.2016, your membership is valid for 2016.

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:

<table>
<thead>
<tr>
<th>Name</th>
<th>Swiss Association for Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td>Chemin des Fontannins 12, 1066 Epalinges</td>
</tr>
<tr>
<td>Bank</td>
<td>PostFinance</td>
</tr>
<tr>
<td>Account</td>
<td>85-122140-9</td>
</tr>
<tr>
<td>IBAN</td>
<td>CH65 0900 0000 8512 2140 9</td>
</tr>
<tr>
<td>BIC</td>
<td>POFICHBEXXX</td>
</tr>
</tbody>
</table>

Step 2

Send an email to member@swiss-analytics.com 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 shipping by mail
- Event or magazine expenses when no sponsor is found
LE MEILLEUR DE VOS DONNÉES !
COLLECTER, ORGANISER, DIFFUSER.

ÉDITION
GESTION DE PROJET
DEPUIS 2001

IMPRESSION
PRODUCTION À LA DEMANDE
DEPUIS 1947

PUBLICATION
WEB, APP, CROSS-MEDIA
DEPUIS 2004

ADRESSE
Gasser Media SA
Jambe-Ducommun 6a
CH-2400 Le Locle

CONTACT
Tél. +41 32 933 00 33
info@gasser-media.ch
www.gasser-media.ch

Pour suivre nos actualités :
www.gasser-media.ch/blog