

SMALL  
**BIG DATA**  
CONGRESS



**TNO** innovation  
for life

5 OKTOBER 2017  
SEATS2MEET UTRECHT CS

# CONTENTS

JOHAN BAAKE – IS CONTROL-M ALSO THE SECRET SAUCE TO YOUR BIG DATA SUCCESS?	5	LAMIN KINTEH – PREDICTIVE ANALYTICS IN HR, FROM TALKING ABOUT IT TO MAKING IT HAPPEN!	18
TOM BAKKER – UNBOXING THE BIG DATA BLOCK BOX: FACT AND FAIR DATA SCIENCE	6	DAVID LANGLEY – SIMULATING THE IOT: COMBINING NK AND ABM MODELLING	19
HARRIE BASTIAANSEN – DESIGN-FOR-AUDITABILITY (D4A): IN CONTROL OVER (BIG) DATA EXCHANGE	7	ELENA LAZOVIK – DATA LOCALITY ENGINE - LARGE SCALE DATA ANALYSIS CLOSE TO DATA SOURCE	20
MICHAEL VAN BEKKUM – GET OUT OF YOUR BUBBLE: DECISION SUPPORT ON UNCERTAINTIES WITH SUDS	8	WILMA MULDER – TRACK AND TRACE THE LATEST INSIGHTS	21
INGE BOUMA – PROCESS MINING + BUSINESS INTELLIGENCE CREATES VALUE	9	AD POPPER – THE BMAP TECHNOLOGY	22
ERROL VAN ENGELEN – INNOVATIVE BIG DATA PLAN	10	ROEL SCHIPHORST – PRIVACY-AWARE CROWD ANALYTICS DURING FESTIVALS	23
EDWIN HARMSMA – TRACEABILITY FOR BIG DATA SERVICES ACROSS ORGANIZATIONS	11	PAULIEN VAN SLINGERLAND – SHARE ANALYSES, NOT DATA	24
VINCENT VAN HEES – MCFLY: DEEP LEARNING FOR TIME SERIES MADE EASY	12	DAVE VAN DER SLUIS – YOU’LL UNDERSTAND IT, WHEN YOU SEE IT	25
FIEKE HILLERSTROM – HOW MUCH DO YOU TRUST YOUR BIG DATA?	13	RON SNIJDERS – INTERACTIVE INSIGHT IN BIG DATA	26
MARTIJN HOOGELAND – BIG DATA TO SMALL BITES, MONITORING SHIP’S HEALTH	14	MICHIEL STORNEBRINK – FIT WITH ONTOLOGIES	27
HENNIE HUIJGENS – STRONG AGILE METRICS	15	RON STUKKER – SMART INFRASTRUCTURE MANAGEMENT - WHAT DOES IT TAKE?	28
JOHAN DE JONG – BUSINESS OPPORTUNITIES WITH (BIG) DATA.	16	JOK TANG – CONNECTING THE DOTS	29
PETRIT KABASHI – ENGAGE & RESONATE THROUGH RELEVANT CONTENT	17	WARD VENROOIJ – SHOWING UNCERTAINTY CORRECTLY INSTEAD OF JUST ADDING SOME MORE	30
		JEROEN VLEK – A REACTIVE DATA MODEL INSTEAD OF PREDICTIVE	31
		BRAM VAN DER WAAIJ – DATA QUALITY ASSURANCE	32

# INTRODUCTION

TNO, in collaboration with the Big Data Value Center, presents the fourth Small Big Data Congress! Our congress aims at providing an overview of practical and innovative applications based on big data. Do you want to know what is happening in applied research with big data? And what can already be achieved through big data? This booklet rapidly takes you through the latest insights and advancements about big data in the Netherlands.

By providing a platform for both usual and unusual suspects, the Small Big Data Congress aims at creating unexpected combinations between industry, government and research. The three-minute pitches that are presented are about new results with or new applications of big data. We have invited SMEs, larger businesses, governments and research institutes to share their latest innovations. In this booklet you find a summary of all pitches that are presented during our 2017 version of the congress.

Best regards,

Jessica Doorn  
Business Consultant Data Science at TNO  
jessica.doorn@tno.nl

TNO brings together expertise in structured and unstructured data, data infrastructure and data governance, with a particular emphasis on data analytics and semantic technologies. In a world of connected people and devices, data, both complex data and Big Data, is becoming of ever greater importance. Our core mission is to make sense of this data, to help companies and authorities to manage their data and to enable them to make better decisions from both a societal and a business perspective.

# IS CONTROL-M ALSO THE SECRET SAUCE TO YOUR BIG DATA SUCCESS?

With our solution Control-M we automate, accelerate and optimize Hadoop and supporting technology ecosystems to enable agility, operational excellence, and a competitive edge.

Control-M is not a big data solution as such, but focus on how the big data environment can be embedded into the organization faster and more secure. This is done by automating batches in and around the big data environment from one single point of control, giving developers the tools to scale and integrate with other applications faster, resulting in faster ingestion with less service interruptions and less resources.

Customers like Navistar report benefits like 5x faster creation of actionable data with 20% savings in work time. GoPro testifies that one Data scientist can do the work of many because of Control-M. Is Control-M also the secret sauce to your Big Data success?

## CONTACT

Johan Baake  
BMC

E [johan\\_baake@bmc.com](mailto:johan_baake@bmc.com)



# UNBOXING THE BIG DATA BLOCK BOX: FACT AND FAIR DATA SCIENCE

Description: Big data can result in unintended, negative or unfair consequences for citizens, society and businesses. To address a number of key technological and societal challenges, the National Science Agenda has funded the 2-year 'Startimpuls' program "Verantwoorde Waardecreatie met Big Data". Key topics include:

1. Fairness: How to prevent or limit the negative impact of big data technologies? What is the danger of analytics and personalized news recommendation systems for democracy and access to a balanced set of information?
2. Accuracy: Which level of accuracy can be guaranteed when using big data to answer research questions?
3. Confidentiality: What potential does polymorphic encryption and pseudonymisation hold for sharing and using data in a safe and confidential way, while preserving accessibility and accuracy?
4. Transparency: How to warrant responsible collection and analysis of personal data in law enforcement
5. FAIR: How can big, sensitive and complex data be made Findable, Accessible, Interoperable and Re-usable to maximize value creation?

The multidisciplinary research program is coordinated by NWO, TNO and the Netherlands eScience Center, runs in 2018-2019 and unites more than 20 universities, HBO's and other institutes across the Netherlands. The 8 subprojects seek to address fundamental computer and data science questions as well as specific questions from the areas of news media, information services, health and security. The program will actively seek to distribute results to and enable collaborations with external stakeholders for validation, improvement and adoption.

## CONTACT

Tom Bakker  
Netherlands eScience Center  
E t.bakker@esciencecenter.nl

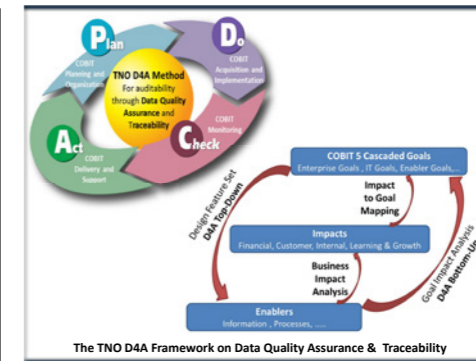


# DESIGN-FOR-AUDITABILITY (D4A): IN CONTROL OVER (BIG) DATA EXCHANGE

Current society heavily relies on the correct functioning of its 'big data multi-stakeholder' ICT-chains (and its individual links). With increasing vitality to society, the importance of being in control over the data being exchanged becomes crucial. This holds for both: – a prospective view, i.e. managing data reliability through forward-looking 'Data Quality Assurance' mechanisms, and – a retrospective view, i.e. taking accountability through backward-looking 'Traceability' mechanisms.

In 'big data multi-stakeholder' ICT-chains manual processing of data becomes unfeasible. Automated control mechanisms for both the prospective Data Quality Assurance and the retrospective Data Traceability capabilities are required. A method for embedding such control mechanisms in an audit framework is still lacking.

Therefore, TNO has developed the Design-for-Auditability (D4A) method to include the data quality assurance and traceability capabilities in the (COBIT) audit framework. It provides both a goal impact analysis method for 'bottom-up' requirement-identification and a capability feature set for 'top-down' design support. These give IT-managers the necessary guidance for process and technology improvement to (re-)gain control over their (big) data exchange processes in multi-stakeholder ICT-chains.



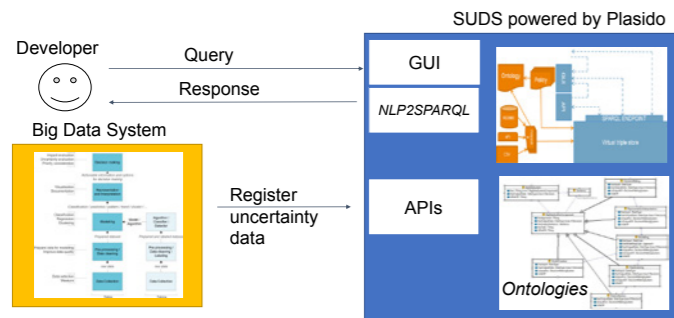
Message to the audience: Being in control requires the designed balance and alignment of goals, risks, management processes and technical measures.

The D4A framework builds upon the global COBIT framework for IT management, auditing and control. It extends COBITs well-established audit framework with TNO's technology expertise.

## CONTACT

Harrie Bastiaansen  
TNO  
harrie.bastiaansen@tno.nl

# GET OUT OF YOUR BUBBLE: DECISION SUPPORT ON UNCERTAINTIES WITH SUDS



Big data pipelines entail all sorts of uncertainties in their system components: data processing, model choices and interpretation lead to various types of uncertainties in data that are partly based on a.o. bias, assumptions, choices and ambiguity. Users require understanding of these uncertainties in order to make informed decisions. Decision support that allows users to gain insights into the types and quantities of uncertainties as well as relations between uncertainties and the effects of system components on uncertainty is therefore paramount. Most information support systems however either do not quantify the final uncertainty or present uncertainty as a quantification of a single (or aggregated)

entity. In both cases, they do not allow the user to drill down and get deeper understanding of its causes and effects.

TNO implements SUDS, an uncertainty prototype product that can be applied to any data system. SUDS (1) allows a user to query for and retrieve extensive uncertainty information on a particular decision and (2) allows uncertainty analysts to register uncertainties and aggregate them. In order to do so, SUDS deploys formal semantics and logic reasoning to infer new knowledge on decisions and uncertainties in big data systems. Users benefit from SUDS by gaining insight into influences of uncertainty in their system and in the end make more informed decisions in their use case.

## CONTACT

Michael van Bekkum  
TNO  
E [michael.vanbekkum@tno.nl](mailto:michael.vanbekkum@tno.nl)

# PROCESS MINING + BUSINESS INTELLIGENCE CREATES VALUE

The recent marriage between Process Mining - a real Netherlands' innovation - and Business Intelligence has resulted in technology, packed as Enterprise Platforms, that really drives process improvement in large organizations. Process Mining provides deep insight in processes and actionable possibilities for improving these. The combination with Business Intelligence puts these in the hands of any case and process owner: easy-to-use and directly related to the business processes of users. Insurance company Aegon, accountant EY as well as national authorities as well as the national police will serve as examples.

## CONTACT

Inge Bouma  
ProcessGold ([www.processgold.com](http://www.processgold.com))  
E [inge.bouma@processgold.com](mailto:inge.bouma@processgold.com)



# INNOVATIVE BIG DATA PLAN



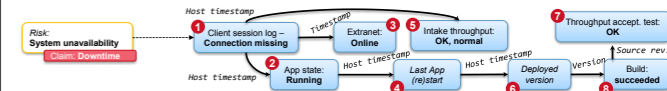
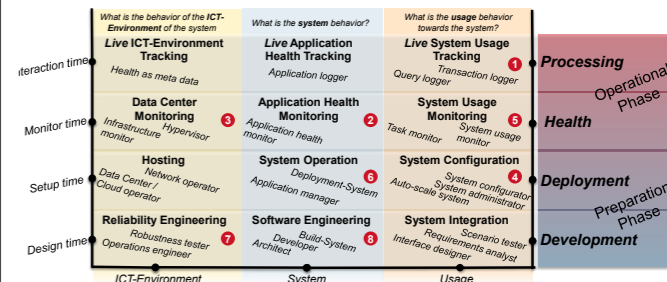
## CONTACT

Errol S. van Engelen  
 Bizmaxx  
 E errol.vanengelen@yahoo.com

# TRACEABILITY FOR BIG DATA SERVICES ACROSS ORGANIZATIONS

Trustworthy decision making in cross organizational information chains require organizations to be transparent and to be able to react on liability claims. Traceability is the technology which constructs traces through the historically captured behavior, which is the basis for transparency and helps to provide evidence for liability claims. In modern large scale Big Data Services, it is rarely seen that such complex traces can be reconstructed after an incident happened.

A matrix containing a categorization of log files is developed that helps in the process to see which traces can and cannot be made through the historical information. The vertical axis defines the period in time with respect to the Application Lifecycle: Design time, Setup time, Monitor time, and Interaction time. The horizontal axis defines the Perspectives: the usage of the system, the system itself, and the ICT-environment. Trace-technology creates trace-blueprints at design time, which ensure that the right correlations between all log files along the trace-path can be made at runtime. This way, a transparent discussion can be held to make the right Audit or Liability conclusions.



## CONTACT

Edwin Harmsma  
 TNO  
 E edwin.harmsma@tno.nl

# MCFLY: DEEP LEARNING FOR TIME SERIES MADE EASY



In a variety of scientific fields researchers face the challenge of time series classification. For example, to classify activity types from wrist-worn accelerometer data or to classify epilepsy from electroencephalogram (EEG) data. For researchers who are novel in the field of deep learning, the barrier can be high to start using deep learning. In contrast to computer vision use cases, where there are tools such as caffe that provide pre-defined models to apply on new data, it takes some knowledge to choose an architecture and hyperparameters for your model when working

with time series data. We developed mcfly (<https://github.com/Nlesc/mcfly>), an open source python library to make time series classification with deep learning easy. It is a wrapper around Keras, a popular python library for deep learning. Mcfly provides a set of suitable architectures to start with, and performs a search over possible hyper-parameters to propose a most suitable model for the classification task provided. We will demonstrate mcfly with excerpts from (multi-channel) time series data from movement sensors that are associated with a class label.

## CONTACT

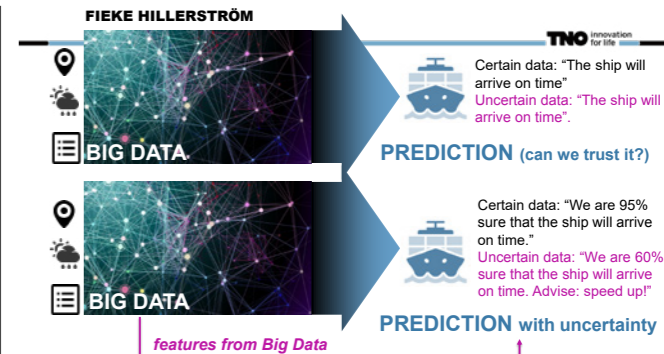
Vincent van Hees  
Netherlands eScience Center, Amsterdam  
E [v.vanhees@esciencecenter.nl](mailto:v.vanhees@esciencecenter.nl)

# HOW MUCH DO YOU TRUST YOUR BIG DATA?

## PROPAGATION AND COMMUNICATION OF OVERALL UNCERTAINTY IN BIG DATA PIPELINES

Big data pipelines are trained to make a prediction about the world and can help to perform tasks more efficiently and effectively than humans. However, one of the differences between us humans and standard computers is that we can give notice if we are not certain of the outcome of a prediction. Standard computers cannot do this yet: big data systems provide a prediction, but do not provide a degree of its certainty. This extra information would be highly valuable for the user. It would allow for better informed decisions and could build more trust in Big Data.

The lack of certainty measures mainly follows from the complexity of big data systems. In fact, it is often not clear how to attribute a degree of final "certainty" to big data results. Starting from data selection and data collection, datasets are intrinsically uncertain. Models trained on these data will also not be perfect, and the choice of a specific analytic model might introduce new sources of uncertainties. As there are many more sources of uncertainty across the full chain of big data processing, and the generated uncertainties propagate through the system, talking about the uncertainty is difficult.



We present a general methodology to analyze, model and propagate the uncertainty of a big data pipeline through the whole system and a way to communicate the derived final certainty in an interpretable manner. We implemented our methods on a computer vision system detecting interesting events in video data.

## CONTACT

Fieke Hillerström  
TNO  
E [fieke.hillerstrom@tno.nl](mailto:fieke.hillerstrom@tno.nl)

# BIG DATA TO SMALL BITES, MONITORING SHIP'S HEALTH

Inspections can help to increase the safety level of ships. Knowing where and when to inspect, can make inspection more cost effective. Just inspecting five yearly may be a waste of money. The European project Safepec developed risk models for ship's safety. Bayesian networks and available historical data is used to define relations and populate probability tables. Obviously, new ship structures do not look like those of the historical data.

Predicting the health of a specific ship from historical data can be improved by adding new information, preferable from a monitoring system on board the ship in question. The 'new' data is fed into the existing risk models and predicts the deterioration of a specific ship. Predictions are verified by 5-yearly inspections. So, risk model validation goes slowly.

## AIM

How can we accelerate the learning process? To make the 'tool' available for the market, we need validation. But that can't take 5 years. A project is needed that accelerates the learning. Industry will benefit from dedicated inspections, reduce maintenance cost and downtime from unexpected repairs.

## PLAN

- Increase datamining
- Develop methods to 'explode' inspection results to increase the knowledge gained

## CONTACT

Martijn Hoogeland

TNO

E [martijn.hoogeland@tno.nl](mailto:martijn.hoogeland@tno.nl)

# STRONG AGILE METRICS

## MINING LOG DATA TO DETERMINE PREDICTIVE POWER OF SOFTWARE METRICS FOR CONTINUOUS DELIVERY TEAMS

Omschrijving: ING Bank implemented a fully automated continuous delivery pipe-line for its software engineering activities in more than 300 teams, that perform more than 2500 deployments to production each month on more than 750 different applications. Practitioners from ING and researchers from TU Delft examined how strong metrics for agile (Scrum) DevOps teams can be set in an iterative fashion. We performed an exploratory case study that focused on the classification based on predictive power of software metrics, in which we analyzed log data derived from two initial sources within this pipeline. We analyzed a subset of 16 metrics from 59 squads. We identified two lagging metrics and assessed four leading metrics to be strong.

Read the paper: <http://swrl.tudelft.nl/twiki/pub/Main/TechnicalReports/TUD-SERG-2017-010.pdf>

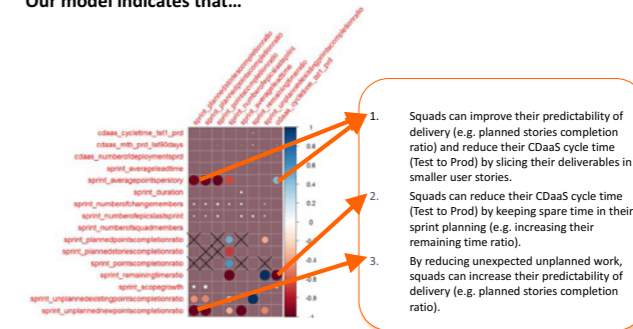
## CONTACT

Hennie Huijgens

TU Delft

E [hennie@goversdon.nl](mailto:hennie@goversdon.nl)

Our model indicates that...





# BUSINESS OPPORTUNITIES WITH (BIG) DATA.



A great part of the small and medium sized firms are still not be conscious of the possible business opportunities by using data. That conclusion could be drawn by a survey that was carried out by JADS and the Chamber of Commerce at the beginning of 2017. So it is still necessary to advise those entrepreneurs. Big data is an abstract phenomenon and not easy to implement. Although the entrepreneurs see the opportunities the step to data analysis and using the results is difficult. That is a reason for the Chamber to organize two awareness and information events.

On Friday 27th of October and Wednesday 1 November we will organize an event for small and medium sized businesses. The main topic will be the focus on business opportunities by gathering and using the available data in the organization. This event is mainly interesting for companies which are at the beginning of using big data. They notice that (big) data might become very important for the future but they do not know where to start and which further steps have to be taken. To create inspiration a couple of companies will show how they use data in their organization and, moreover, we will present a hands on plan which can be used to develop and implement data in your own company. Also during this event we will work in little groups to set the first steps in your own plan. After the event the visitors will have the opportunity to work on their own plan under the supervision of TNO specialists. This will take place in 2 sessions. For more information of the program see [www.kvk.nl/evenementen/446880/](http://www.kvk.nl/evenementen/446880/) and of course you can ask my colleagues or myself!

## CONTACT

Johan de Jong,  
Chamber of Commerce  
E [johan.de.jong@kvk.nl](mailto:johan.de.jong@kvk.nl)

# ENGAGE & RESONATE THROUGH RELEVANT CONTENT

Our vision is to melt the boundaries between the offline and online world by using innovative mobile tagging technologies. In order to achieve our goal we must convince the consumers that the activity of scanning with their smartphone always results in fulfilment of their expectations. To deliver on this promise we track key activities through our App for the purpose of delivering relevant content which is tailored and context specific.

In each interaction with the real world through our App we offer consumers relevant content based on their location, time, gender, age group, language, preferences, interaction history, behaviour and context. In doing this we make sure to respect the privacy laws so that all collected data is anonymous and our system is GDPR proof.

## CONTACT

Petrit Kabashi  
Endeavour Media  
E [petrit@endeavourmedia.nl](mailto:petrit@endeavourmedia.nl)



# PREDICTIVE ANALYTICS IN HR, FROM TALKING ABOUT IT TO MAKING IT HAPPEN!



Zyvo is reshaping the future of HR. With our award-winning and highly advanced personality and neuro-assessments, we predict employee performance with unprecedented precision.

The 'Zyvo HR Analytics Box' combines psychometrics, neuro-science and machine learning with your companies' data to make predictions about revenue, retention, customer satisfaction and other business outcomes.

Predictive analytics in HR, everybody talks about it, everybody wants it, but only few can deliver it. We will show you in 4 steps how easy it actually is to start with predictive HR analytics and how to directly impact your bottom line.

If you like to learn more about how predictive analytics can help your organization realize its ambitions and/or if you would like to partner with us, do attend our pitch!

## CONTACT

Lamin Kinteh  
Zyvo  
E [lamin@zyvo.nl](mailto:lamin@zyvo.nl)

# SIMULATING THE IOT: COMBINING NK AND ABM MODELLING

Despite industry analysts expectations of an unprecedented economic impact from big data solutions, such as emerging from the internet-of-things, many organizations are struggling to implement the technology and realize its potential. A key factor is the difficulty of understanding how to foresee what an optimal extent of data integration is, at both the product and the organizational levels. The University of Groningen, in collaboration with TNO, large corporate and SME firms, is developing new methods to explore the potential designs of connected products and combine this with the potential organizational value networks that become relevant. NK modelling, from evolutionary biology, is an approach suited to explore a phase space of data stream combinations, such as those made possible with smart, connected products. Agent-based modelling, from the social sciences, can be used to take an organizational perspective and study the relations, data flows and business models brought about by networks of firms offering such products. By linking these methods, it becomes possible to explore optimum designs for data combinations, and at the same time explore their implications for business ecosystems and strategies to make big data work.

A question for the audience: Do you believe you have found the optimum design of your big data solution?

## CONTACT

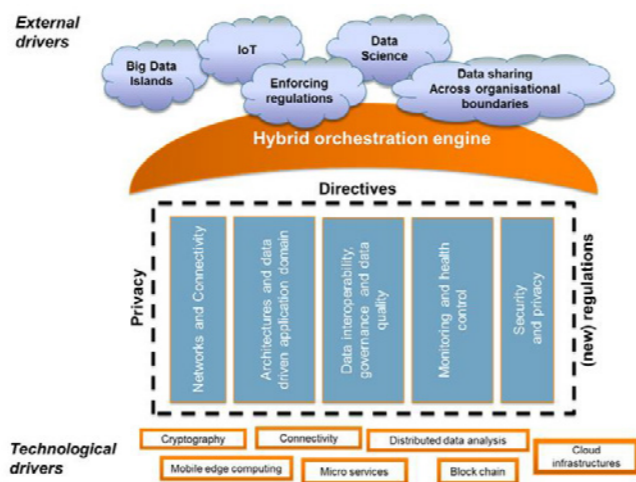
David Langley  
TNO and University of Groningen  
E [david.langley@tno.nl](mailto:david.langley@tno.nl)

# DATA LOCALITY ENGINE - LARGE SCALE DATA ANALYSIS CLOSE TO DATA SOURCE

The current method for large-scale data processing is that all data is copied to a centralized distributed data analysis platform (e.g., Apache Spark, Hadoop) and processed there. With the increase of available data and its size, the challenge of getting meaningful information from combined data sources is getting bigger. What to do if the data needed for the analysis belongs to different organizations, and the organizations cannot give you a copy of their whole Big Data islands?

Making a copy of all data to a large-scale data analysis platform is a huge challenge: for large volumes of data it is time-consuming and puts special requirements on networks, for real-time data this requires to create an entire copy of data acquisition infrastructure, and sometimes data might not be copied because of security, privacy or confidentiality reasons.

The principle of 'data locality' is a route to solve this challenge of large scale data analysis over distributed sources with possible access limitations by making the calculations where data resides. Our goal is to develop a distributed Big Data analysis engine which would bring data analysis algorithms to data where it is needed, and afterwards merge the analysis results from different organizations centrally.



## CONTACT

Elena Lazovik  
TNO  
E elena.lazovik@tno.nl

# TRACK AND TRACE THE LATEST INSIGHTS

The industry is in full swing. 3D printing, IOT and not to forget Big Data playing a major role. KxA Data Solutions, derived from Astronomy, has completed successful Big Data projects for 10 years. The applications of Big Data in the industry are numerous: Failure analyzes, historical storage for efficiency improvements, predictive maintenance operator app, etc. Using the latest technologies, we developed for Philips Drachten a complete track and trace for a production line, including all the means of production therein, such as robot, shaker bins, injection molding plants, carriers and modeled and analyzed the interrelationships between them. For the first time, it has proved possible to follow an individual product through the entire production and to determine exactly where delay and failure occur. Based on these outcomes, the line can be more effectively optimized with the ultimate goal to speed up the time for launching a new product.

If you want to get started with Big Data, we'd like to talk to you about the possibilities.

## CONTACT

Wilma Mulder  
KxA Data Solutions  
E mulder@kxa.nl

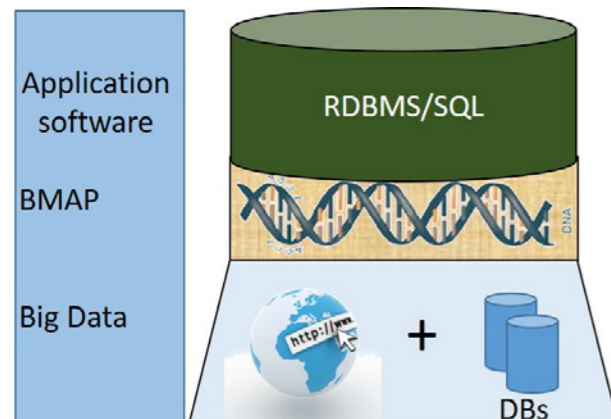


# THE BMAP TECHNOLOGY

## ENABLES BIG DATA DESIGN

Big Data is too large to manage the relations between the data with relational database modelling techniques. Instead of modelling, data scientists reconstruct relations with statistical methods and – techniques, filter out the knowledge, and explain the commercial significances.

The availability of experienced data scientists is a worldwide problem. Beside it, Big Data applications appears too expensive for large scale utilization.



The RDI challenge is to find a solution, to design and manage Big Data relations.

Inspired how the living nature rules his infinite complexity by DNA architecture, Xilion develops the BMAP-technology (Biological Modelling And Processing). On top of BMAP, we design market research applications, presented on TNO small Big Data congress last year.

The efficiency benefits of the DNA architecture approach, shows that the marginal cost price of the application can be reduced by a factor 1.000! Moreover, increasing the complexity has virtual no effect on that cost price. These results are presented on de European Research and Innovation Event at Düsseldorf last March. With a call to develop this BMAP-technology to general applicability. BMAP is a break-through to design all sorts of affordable as innovative Big Data applications. At this moment, the international R&D consortium is in formation. We need cases and reviewers.

### CONTACT

Ad Popper  
Xilion  
E [ad.popper@xilion.nl](mailto:ad.popper@xilion.nl)

# PRIVACY-AWARE CROWD ANALYTICS DURING FESTIVALS

In today's world everyone is 24/7 online. Crowd behaviour can be analyzed and optimized for specific goals. For instance optimizing the conversion of visitors of a webshop. In the offline world, such as music festivals, this is a different story. There are limited amount techniques available to monitor crowd behaviour. The result is that most crowd managers rely on common sense and experience instead of metrics. Furthermore in the offline world privacy is a very important issue, because online people accept the general terms of the website (cookies) first. In this pitch we will address how crowd behaviour during a music festival can be analyzed and at the same time protect the privacy of the visitors. As an example we will use the TT festival where we used WiFi scanners to monitor crowd movements. In the pitch we focus both on privacy issues (and how to solve them), but also metrics results of the crowd behaviour. The message to the audience would be that you can use WiFi tracking and protect the privacy of the visitors. We envisage it as an open question where the audience can comment and provide other solutions.



### CONTACT

dr. ir. Roel Schiphorst  
BlueMark Innovations BV  
E [roel@bluemark.io](mailto:roel@bluemark.io)

## SHARE ANALYSES, NOT DATA

Innovating with sensitive/personal data is challenging. On the one hand, data-analysis offers interesting opportunities for improving decision-making, research, customer profiling, personalization of products and services, etc. On the other hand, if you want to set up a new data service based on sensitive data, you need to ensure compliance with privacy legislation (new laws are coming up), avoid reputational damage and churn (remember the ING incident?), and avoid unwanted insights in your organization (e.g. derivation of the pricing strategy).

Now, imagine the data is not sensitive. Stakeholders like your customers, the government, and your own organization, would feel no need to protect the data. You could create valuable insights based on your data combined with others' data, legislation would no longer be an additional investment risk, and you could share information on your own organization for collaboration purposes and new services.

Techniques exist that remove the sensitivity of data. In our pitch we want to highlight the features of these techniques that TNO is working on. The techniques enable you to compute with data, while that data stays encrypted at all time. This removes the sensitivity, as encrypted data cannot be interpreted. As a result, you can integrate data from different sources and share data analyses with

others, without sharing the (sensitive) meaning of the underlying data.

Are you interested to explore what new possibilities these techniques can offer you? We are looking forward to further investigate the opportunities with you. So if you want to learn more, do not hesitate to contact us.

### CONTACT

Paulien van Slingerland  
TNO  
E [paulien.vanslingerland@tno.nl](mailto:paulien.vanslingerland@tno.nl)

## YOU'LL UNDERSTAND IT, WHEN YOU SEE IT

Our pitch is about how people react when they see something. How active people become when they face information. Especially when it comes to information they are dealing with. Information that allows people to make changes, creates new insights and provides growth.

We want to activate people to search for themselves and not wait for other people to show it.

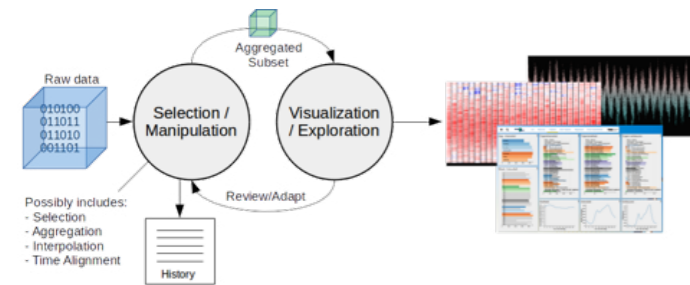
### CONTACT

Dave van der Sluis  
Quant Base  
E [dave.vandersluis@quantbase.nl](mailto:dave.vandersluis@quantbase.nl)

# INTERACTIVE INSIGHT IN BIG DATA

Dataset preparation forms a significant part in the creation of a suitable model in Big Data applications. Especially in situations where the availability of sensor data is unreliable, going through the raw data manually in an effort to select and understand the data becomes increasingly more difficult as the data grows in size and complexity.

In our work we aim to tackle this problem by allowing model makers to explore vast amounts of data visually and select subsets of the data in an interactive way, providing insights in the availability and quality of the desired data.



Using distributed technologies such as HDFS, Hive, YARN and Spark for our back-end, our solution can scale horizontally as the size of the data grows. Interactive tools implemented using Zeppelin and WebGL allows the user to explore the data and interact with the aggregated results with ease, but they can also use familiar programming languages like Python, R, Java and Scala. We use large amounts of noisy sensor data from the dairy farming industry as a use case for our project TKI Young Stock Rearing.

Question to audience: We are interested to hear the struggles the audience may have in their own efforts to explore and select vast amounts of data and how our work could ease their early stages of model selection and preparation.

## CONTACT

Ron Snijders  
TNO  
E [ron.snijders@tno.nl](mailto:ron.snijders@tno.nl)

# FIT WITH ONTOLOGIES

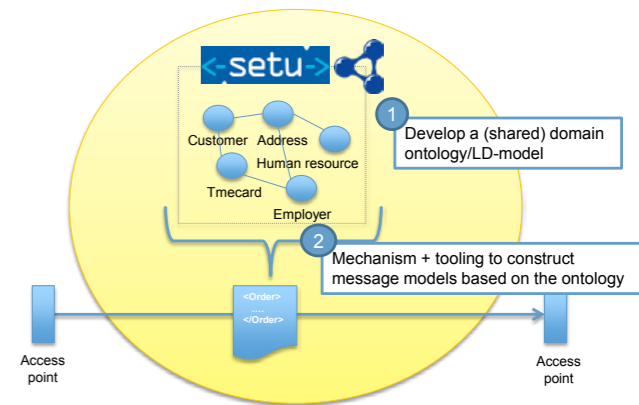
We will give a demonstration of the Semantic Treehouse application by which you can use a (shared) domain ontology/LD model to compose hierarchical messages (xml/json). With this approach you can easily develop REST APIs to provide services that are and stay in line with your domain language.

Currently B2B information exchange (orders, invoices, and all other business documents) occurs using the traditional way of sending electronic documents (xml/json). The future however lies with Linked Data (LD) supported with ontologies and semantic web technologies. These two worlds/paradigms however are miles apart! Solutions to bridge this gap are needed.

Furthermore, standardization of information now focuses on semantics and syntax of messages (e.g. EN 16931 on Electronic invoicing). The academic world looks at providing semantic meaning by means of (formal) ontologies. The latter one will make automatic translations and IT interface configuration possible. Since a big-bang introduction of semantic web technologies is not likely, we connect the 'old' and 'new' paradigms.

## CONTACT

Michiel Stornebrink  
TNO  
E [michiel.stornebrink@tno.nl](mailto:michiel.stornebrink@tno.nl)



# SMART INFRASTRUCTURE MANAGEMENT - WHAT DOES IT TAKE?

Reconciling different source data can add value to maintenance of infrastructure and make processes smarter. Cooperation between knowledge institutions and market parties is essential to tackle the real problem of aging urban infrastructure and other networks.

HAL24K, TNO and Rijkswaterstaat have therefore begun to cooperate to gain more insight into the possibilities, with old bridges as a case. Last year, we have seen with the Merwedeburg how much bother, traffic jams and above all costs are involved the “failure” of such infrastructures. We started with bringing together different data sources from and around the van Brienoordbrug in Rotterdam. When we can get insights into the material degradation of the bridge we can use this to predict the life of the bridge. The ultimate goal is to go to a proactive way of maintenance. For this we use many data sources within TNO and RWS and HAL24K's data science expertise. How do you tackle such problems and what are you up to in such a big project? We are constantly learning about the approach, machine learning models and the possible business cases. We are looking forward to find synergy with the Big Data Value Center.



## CONTACT

René Stukker  
HAL24K  
E [rene.stukker@hal24k.com](mailto:rene.stukker@hal24k.com)

# CONNECTING THE DOTS

## HOW NETWORK ANALYSIS LEADS TO SURPRISING INSIGHTS

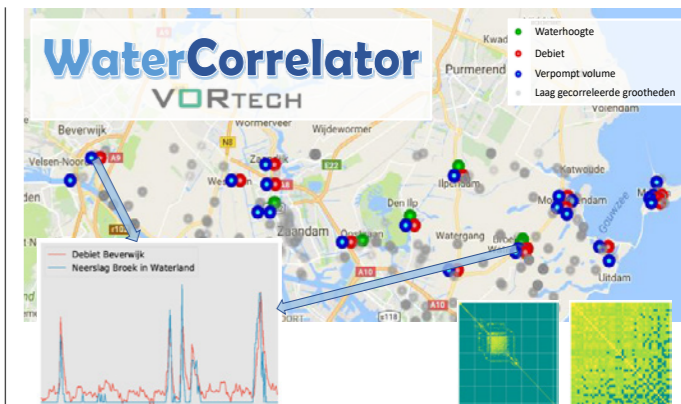
Did you know that when the water pumps in Monnickendam are working hard, you can tell that the water level will be high in Oostzaan, a place that is some 15 km away? And did you know that the temperature in the water around the equator is hardly connected to the water temperatures away from the equator?

Surprising connections like these can be made from network analysis. Such an analysis starts by finding the correlations between events in different locations and then studying the interaction network that is formed by events that are strongly correlated.

VORtech has developed software to construct and study such networks for huge data sets. An example of this is a data set of sea temperatures at a high resolution. This allows researchers to understand how climate subsystems are connected to each other. Similar software is used in our Watercorrelator, which finds relevant connections in data from water management authorities in the Netherlands.

## CONTACT

Jok Tang  
VORtech Data Science  
E [jok.tang@vortech.nl](mailto:jok.tang@vortech.nl)

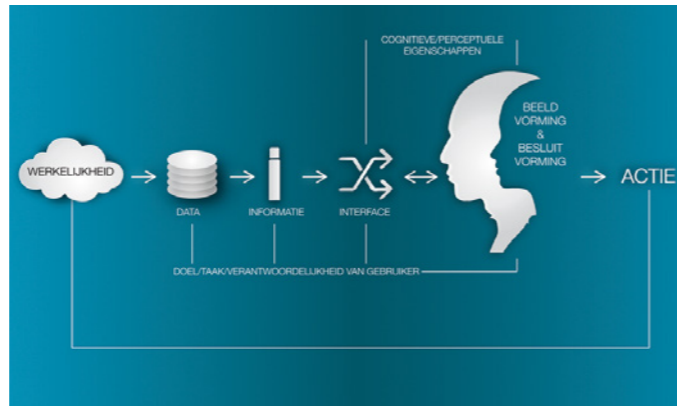


## SHOWING UNCERTAINTY CORRECTLY INSTEAD OF JUST ADDING SOME MORE

Every big data system has a certain degree and types of uncertainty originating from different locations in the process. Decision makers require a certain understanding of this uncertainty to be able to make correct decisions. Therefore, we have to make sure that these system uncertainties are communicated effectively and efficiently to the user. However, research shows that the user interpreting results or accompanying uncertainties from an interface may cause a great deal of uncertainty in itself. We see that data visualizations often fall short of their potential and fail to communicate the correct message. At TNO we develop innovative measures and methods to quickly assess the performance of various (uncertainty) visualizations in order to reduce uncertainty caused during interpretation to a minimum.

### CONTACT

Ward Venrooij  
TNO  
E [ward.venrooij@tno.nl](mailto:ward.venrooij@tno.nl)



## A REACTIVE DATA MODEL INSTEAD OF PREDICTIVE

Customer engagement is driven by what is happening around them. The majority of our time is spent watching screens, whether it is our television, smartphone or laptop. Digital advertising agencies are taking advantage of online searches by consumers with reasonable success. The ability to improve the reaction speed on unexpected developments is a new competitive differentiator for our client Mporium.

A large part of any online advertising campaign is determining how much a click and its conversion is worth. Business decisions in a Google AdWords account are based on the return-on-investment. While bidding for advertising space is typically a blind auction with special sets of rules, optimizing your bid is difficult.

External influences like accidents, weather conditions, news, PR activities and events, tv subtitles or tv guides, all have impact on online search behaviour. We use a Reinforcement Learning and Neural Networks based approach, using these external influences, to model which effect different bids have on profits of customers. We expect a significant decrease in costs for any online advertising campaign and a smarter allocation of available resources.



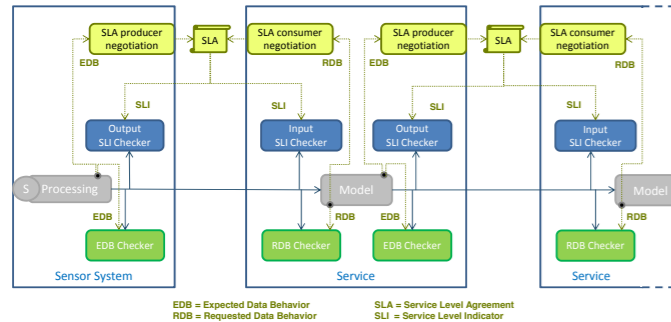
Instead of making changes to active keywords and ads manually, an additional automation model makes it possible to react faster and better.

### CONTACT

Jeroen Vlek (CTO)  
Anchormen  
E [j.vlek@anchormen.nl](mailto:j.vlek@anchormen.nl)



# DATA QUALITY ASSURANCE



Data quality assurance is essential to keep an operational data system robust and reliable through its cross organizational collaboration environment. The big question is: "does reality still meet the expectations of your model, which were defined during the model development?" Over time sensors can degrade, underlying relationships can drift, etc.

Up to a certain extent, handling changes in data input will be part of the (pre)processing of the model. Beyond that, the model will probably still produce results but might be incorrect. Data quality assurance is about monitoring data and compare that to the expectations and obligations (possibly converted into contracts).

This can be verified by implementing one or more of the following checks:

- Assure the quality of data that my system is producing:
  - a. Does the incoming data still fit my model requirements? (RDB Checker)
  - b. Does the data my system produces still follow my expectations? (EDB checker)
- Assure the agreements made with involved parties:
  - c. Does the incoming data still meet the agreement with the producer? (Input SLI checker)
  - d. Does the data(my system produces still follow the agreement made with my consumer? (output SLI checker)

Data quality assurance is part of Design-for-Auditability.

## CONTACT

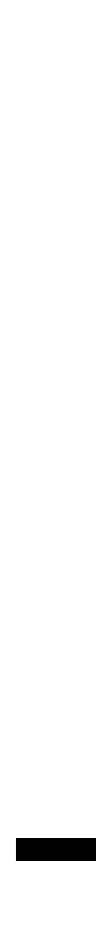
Bram van der Waaij  
TNO  
E [bram.vanderwaaij@TNO.nl](mailto:bram.vanderwaaij@TNO.nl)



34



35



**BIG DATA**  
VALUE CENTER

**TNO**

**TNO.NL**