INFOCOMP 2017 International Expert Panel:

From Advanced Applications to Optimisation and Energy Efficiency: How Will Data Science Benefit to Communication and Computing in Science and Society?

June 27, 2017, Venice, Italy

The Seventh International Conference on
Advanced Communications and Computation (INFOCOMP 2017)
The Thirteenth Advanced International Conference on
Telecommunications (AICT 2017)

INFOCOMP/AICT
June 25–29, 2017 - Venice, Italy

INFOCOMP Expert Panel: . . . How Will Data Science Benefit to . . .

Panelists

- *Claus-Peter Rückemann* (Moderator),
  Westfälische Wilhelms-Universität Münster (WWU) /
  Leibniz Universität Hannover /
  North-German Supercomputing Alliance (HLRN), Germany
- *Isabel Schwerdtfeger,*
  IBM, Germany
- *Lutz Schubert,*
  University of Ulm, Germany
- *Zlatinka Kovacheva,*
  Middle East College, Oman
- *Claus-Peter Rückemann,*
  WWU Münster / Leibniz Universität Hannover / HLRN, Germany

INFOCOMP 2017: `http://www.iaria.org/conferences2017/INFOCOMP17.html`
Program: `http://www.iaria.org/conferences2017/ProgramINFOCOMP17.html`

INFOCOMP Expert Panel: . . . How Will Data Science Benefit to . . .

Pre-Discussion-Wrapup / Panel Statements:

- **Practical experiences:** Data, information, knowledge and their relation to Data Science and Data-driven Science are becoming increasingly important.
- **Methodologies and best practice:** Data/knowledge-centric approaches are best for long-term task and scenarios.
- **Data transformation:** Hare-brained methods are needed.
- **Quantity/quality:** We need to balance between quantity and quality. understanding the value of Big Data and analytics.
- **Statistical analysis vs. machine learning approach:** Statistics fail with large data bases. New algorithms can be used to search for hypotheses. Neural networks can be applied for data mining.
- **Technical solutions:** Modular, standardised, exchangable components should be created, for long-term.
- **Optimisation, energy efficiency, data locality:** View of disciplines should be fostered.

INFOCOMP Expert Panel: Post-Panel-Discussion Summary

Post-Panel-Discussion Summary (2017-06-28):

- **Data and data related methods, algorithms, and architectures are critical for computing**.
- **High End Computing and education should foster understanding of knowledge and data-centric approaches**.
- **Approaches and solutions** for undisturbed knowledge and analysis are **already emerging** (e.g., IBM Watson analysis).
- **Data is not just information**, different **views need to be integrated**, context and extrapolation cannot always be created later oder separately.
- **Statistics** is important but it **is neither a major nor the sole solution** with Big Data.
- **Practical experiences:** Data, information, knowledge and their relation to Data Science and Data-driven Science are becoming increasingly important.
- **Models:** Better models are required, integrating methodologies. Consider the insight of great thinkers of mankind. (Contrib.: Audience, Alexander Troussov).
- **Best practice:** Data-centric approaches are best for long-term task and scenarios.
- Data-centric, computing-centric, . . .: **Purpose** is important. Data-centric approaches can foster the perception and value of data. Quality, organisation, and management of data can benefit from data-centric approaches.
- **Efficiency, energy efficiency, optimisation, and economic issues should be seen from more holistic perspectives**.

INFOCOMP Expert Panel: Table of Presentations, Attached

Panelist Presentations: (presentation sort order, following pages)

- **Advancing Computing Does Mean Advancing Data Science and Long-term Resources** (*Rückemann*)
- **IBM Watson Analytics – Unbiased analytics with cognitive capabilities** (*Schwerdtfeger*)
- **Data is Not Information – Can we Close the Gap Between Science and Humanities?** (*Schubert*)
- **Challenges and opportunities in analytics of Big Data** (*Kovacheva*)

INFOCOMP 2017 International Expert Panel:
From Advanced Applications to Optimisation and Energy Efficiency:
How Will Data Science Benefit to Communication and Computing in Science and Society?

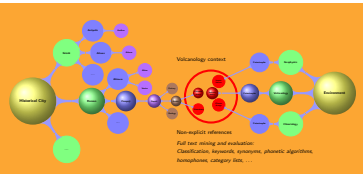
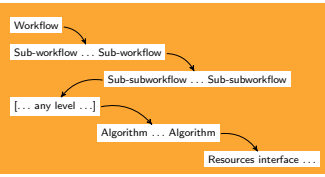
Advancing Computing Does Mean Advancing Data Science and Long-term Resources

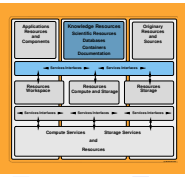
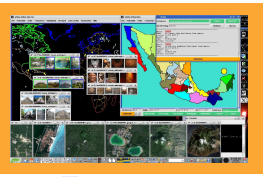
The Seventh International Conference on Advanced Communications and Computation
The Thirteenth Advanced International Conference on Telecommunications
(INFOCOMP 2017 & AICT 2017)
June 27, 2017, Venice, Italy

Dr. rer. nat. Claus-Peter Rückemann[1,2,3]

[1] Westfälische Wilhelms-Universität Münster (WWU), Münster, Germany
[2] Leibniz Universität Hannover, Hannover, Germany
[3] North-German Supercomputing Alliance (HLRN), Germany

ruckema(at)uni-muenster.de

Computing and Data Science

- **Computer:** A device (an implementation of a mathematical machine) applicable for universal automatic manipulation and processing of data.
- **Computing** is "implemented methodology".
- **Efficiency** of computing is depending on implementations (data, algorithms, technical aspects, data locality, ...).
- **Portability** is depending on implementation (data, algorithms, technical, ...).
- **Results** are depending on complement work of data, algorithms, ...
- **Data science** is focussed on data analysis, related methods, and statistics. Data science is known as data-driven science, too.
- **Data use:** Computing is (mostly) done based on (already) organised data.
- **Data organisation:** In most cases data is organised following available tools' features.
- **Data and computers:** The tools should suit best for the purpose.

Common data-centric solutions, knowledge resources, ...

- **Data organisation** is becoming increasingly important.
- **Data-centricity and data science** are becoming central issues.
- **Long-term:** Most data are required for long-term.
- **Elaborateness/quality of knowledge** is not a matter of structure or tools only. Examples of proposed matrices:
  High-End-Quality/High-End-Structure ... Low-End-Quality/Low-End-Structure.
- **Data and computers:** We should not make mistakes from past, e.g., too specialised architectural solutions.
- **General and standard:** Have modular components and multi-purpose and multi- and inter-disciplinary scenarios and fields in mind.
- **Focus:** Scientific methods, processes, and systems, e.g., to extract knowledge or insights from data in various forms.
- **Complexity and context** are carrying knowledge and information.
- **Use complexity and context**, which are intrinsic to most data.

Conclusions: Data-centric long-term solutions

- **Data and data related methods, algorithms, and architectures are critical for computing.**
- **Create data-centric, data science supported solutions ...**
- **Improve data organisation, long-term data, structures, means.**
- **Create standards/systematics/methodologies with content.**
- **High End Computing and education should foster understanding of knowledge and data-centric approaches.**
- **Efficiency, energy efficiency, optimisation, and economic issues should be seen from more holistic perspectives.**

Future ...

- **Focus:** Where is the / our / others' common focus?
- **Communication:** Specifics regarding communication / technology?
- **Computing:** Specifics regarding computing / technology?
- **Recommendations:** In which direction should data science develop?

Panel

From Advanced Applications to Optimisation and Energy Efficiency: How will Data Science Benefit to Communication and Computing in Science and Society?

Isabel Schwerdtfeger
HPC& HPSS Services Sales Leader, IBM Market DACH
Tel. +49/170 635 72 51
June 29, 2017

IBM

IBM Watson Analytics

Unbiased analytics
with cognitive capabilities

Analytics with expanding versatility

Watson Analytics continues to evolve, presenting users with new ways to address their challenges with greater efficacy and confidence:

- 1M registered for Watson Analytics in the first year
- 1.8M registered users across 39,000 organizations in the first 18 months
- Over 500 academic institutions are using Watson Analytics since July of 2015
- 2015 saw the launch of Watson Analytics for Social Media
- Use cases that span customer service, sales, marketing, production, HR/employee and IT/support
- New relationships with:
  - Datawatch for intuitive data prep and extended reach into new data sources
  - Mapbox for business focused geospatial and mapping visualizations of data

1M
YEAR1 REGISTRATIONS

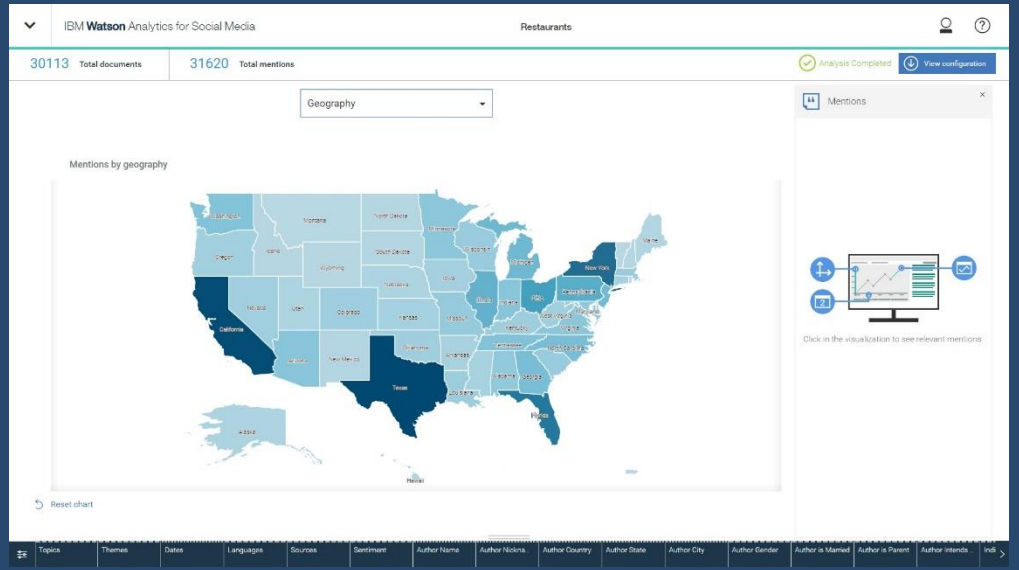
1.8M
REGISTERED

500
ACADEMIC INSTITUTIONS

Watson Analytics for Social Media:

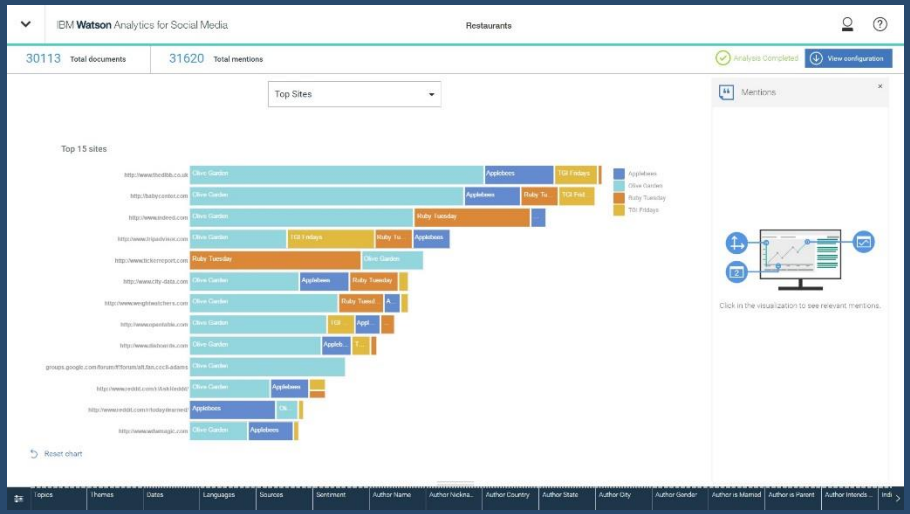
- Watson Analytics for Social Media extends the cognitive platform of Watson Analytics to the social web

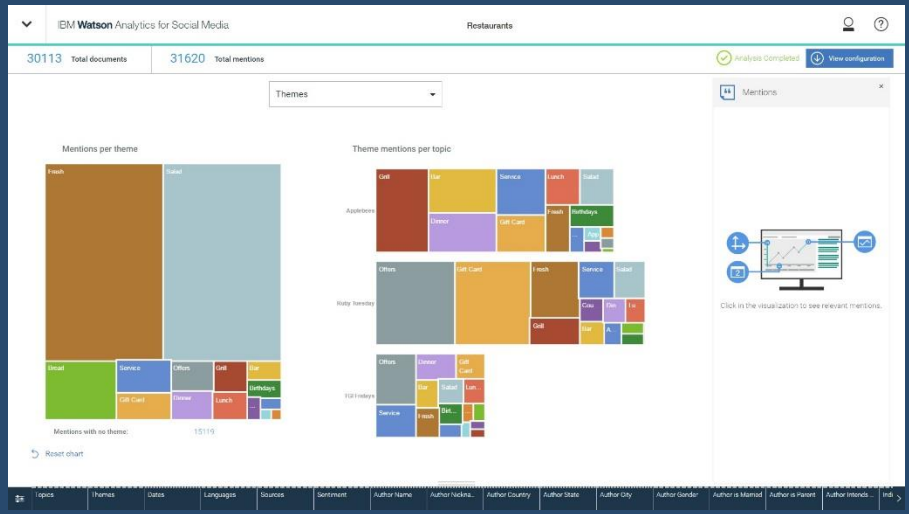
- Access, understand, reason and learn from the highly valuable, unstructured information streaming across social channels.



Identify the pulse of your audience:

- Reveal insights in conversations across all social channels
- Providing pre-built visualizations to highlight the most important insights






What can IBM Watson Analytics do for you?
See how IBM® Watson™ Analytics can unlock your full potential.
www.watsonanalytics.com
Sales
Operations
Watson Analytics
Finance
Human Resources
IT
Marketing

INFOCOMP 2017: Special Track – Tuesday, 27.06.17

18:50 - 20:20	INFOCOMP 2 / LSEEDCC			

- Large Scale Energy Efficient Data Center Concepts – LSEEDCC
- Special Guests:

Supriyo Bhattacharya
Innovo Cloud GmbH

Professor Lindenstruth
University of Frankfurt, Germany

Thank you

Isabel Schwerdtfeger

Leading Sales Professional
HPC & HPSS Sales Leader DACH
IBM Germany

Tel. +49/ 170- 635 72 51
Email: schwerdtfeger@de.ibm.com

Data is NOT Information

can we close the gap between
science and humanities?

Lutz Schubert

(lutz.schubert@uni-ulm.de)

- Senior researcher, lecturer and deputy director at the University of Ulm
- Member of the board of directors of the German chapter of the CAA (computational algorithms and quantitative methods in archaeology)
- External expert and moderator for the European Commission

Humanities is about information

- Humanities centers on the human:
- What does a certain environment condition mean to us?
- What is the meaning of language?
- How can we live in certain political conditions?

Science is about data

- Data has no meaning: it represents a physical condition
- Language is
  - sound waves
  - series of letters
  - syntax and grammar
- Politics is
  - a time series of events and conditions
- Climate is
  - temperature
  - air pressure
  - humidity
  - …

Data vs Information

The problem:

- Information is not static
- It can be put down in letters, but may change meaning over time (read Latin now)
- It cannot be computed, as it is subjective

Science is not really about data

However we perform science on data in the hope to generate information:

- CFD simulations e.g. try to identify whether a car is fast and how it can become faster
- Weather simulations are not about pressure zones, but whether we need to wear warm clothes, whether the crop grows, whether it’s safe to go out

➢ We just never developed a different way

Information based Science?

With computer science complexity coming to an end we need to rethink our way to compute, to write algorithm, to store data – centered around information

I3

- Information
- Incentive
- Intention

Read the report at:
http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=6775

CHALLENGES AND OPPORTUNITIES IN ANALYTICS OF BIG DATA

According to Asigra, a Cloud Backup company since 1986, a staggering 90% of the data in the world today have been created during only two years. And, it is predicted that the worldwide number of Internet Protocol (IP) addresses will quadruple very soon. Moreover, it is forecasted three billion people will be online creating close to eight zettabytes of data for two years only .

We are living now in unprecedented era of innovative technologies that create colossal volumes of both structured and unstructured data. We need to balance between quantity and quality.

Big Data analytics

We have to focus more and more at the understanding the value of Big Data and analytics. **Recognizing, understanding, and using Big Data in terms of scientific research** are necessary at this time in a world of ever increasing data.

Statistical analysis vs. machine learning approach

For most types of experiments, sampling data is sufficient to build an effective picture of the entire dataset and, **statistically**, we can give high levels of accuracy to predictions based on relatively small **samples**. Data collected in this way is often of very high quality. To ensure the sample is representative and accurate, the data is collected and 'cleaned' with great care. This extra care is often very expensive, however, and over the last few decades we have seen the costs of running large randomized control trials spiral upwards.

Instead of researchers creating a hypothesis and collecting data from samples, **machine-learning algorithms** plow through large data sets searching for **hypotheses**.

Main drawbacks of the classical statistical techniques:

- They impose restrictions on the number of input data: one is limited to a few inputs among dozens or hundreds available, imposing a priori variable selection, with all the inherent pitfalls ;
- Regressions are performed using simple dependency functions (linear, logarithmic) that are not very realistic ;
- The hypothesis is made that there is only one dependency function over the whole data set, instead of many distinct niches ;
- Other hypotheses imposed by their underlying theories (normal distributions, equiprobabilities, uncorrelated variables) known to be violated, but that are necessary for their good operation ;
- The need to use hare-brained methods to transform data.

More fundamentally, to quote Professor Peter D.M. MacDonald, of McMaster University, Ontario, Canada: **Traditional approaches to statistical inference fail with large databases**, however, because with thousands or millions of cases and hundreds or thousands of variables there will be a high level of redundancy among the variables, there will be spurious relationships, and even the weakest relationships will be highly significant by any statistical test.

Data Mining

Data Mining is an analytic process designed to explore data (usually large amounts of data) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

In general, the process of data mining consists of **three stages**:

- initial exploration,
- model building or pattern identification with validation/verification,
- deployment (the application of the model to new data in order to generate predictions).

Main DM methods:

- **Associations rules** - discovery of interesting relations in large databases that concern the co-occurrences of different elements;
- **Classification** - assigns items in a collection to target categories or classes;
- **Regression** - predicts numeric values along a continuum;
- **Clustering** - finds clusters of data objects that are similar in some sense to one another.

NEURAL NETWORKS

Neural Networks

Neural Networks is one of the Data Mining techniques. They are analytic techniques modelled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data. As the prediction is usually an activity of the human brain, to automate this process, it is necessary to understand “How the human brain learns?”

Main advantages of neural networks

We can outline the following advantages of neural networks comparing with other data mining techniques :

- Ability to account for any functional dependency. The network discovers (learns, models) the nature of the dependency without needing to be prompted. No need to postulate a model, etc.;
- One goes straight from the data to the model without intermediary, without recoding, without binning, without simplification or questionable interpretation;
- Insensitivity to « moderate » noise or unreliability in the data.
- No conditions on the predicted variable: it can be a Yes/No output, a continuous value, one or more classes among n, etc.;
- Ease of handling, much less human work than traditional analytical methods;

Main advantages of neural networks

- No need to manually detect collinearities;
- In segmentation, the net determines by itself how many clusters there are in each class;
- Speed of use: 10 microseconds when hardwired, a few milliseconds on a 1 GHz computer;
- Spatial relations (geomarketing etc.) are easily analysed and modelled;
- The final model is continuous and derivable and lends itself easily to further work;
- The neural networks model associations and not causes;
- The neural model is validated using a number of examples that were excluded from the learning set, called the «test set». Expected and predicted values are compared.

كلية الشرق الأوسط
Middle East College

THANK YOU!