#### Prof. Dr. Katharina Morik

## Head of the Collaborative Research Center SFB 876, TU Dortmund University

#### How do you define Big Data?

Big Data is defined by high volume, high velocity, and high variety data sets that challenge methods of storage, search, and analytics. The notion of "big" is seen with respect to the resources available for computation. Given a very small computer, even a medium-sized data set is considered "big."

#### How will Big Data transform society in the future?

Big Data often stems from cyber-physical systems, which produce data in a stream. Data stem from sensors in streets, cars, medical devices, telescopes, production (factories), logistics, social networks, search engines, etc.

First, very large data collections are well utilized to optimize diverse processes. For instance, traffic flow can be optimized to save energy. Factories can be optimized through adaptive quality control so that processing stops or is changed as soon as the predicted quality is no longer sufficient. Distributed analysis optimizes the logistics routing on the basis of sensors that are integrated into each container and communicate with each other.

Second, science increasingly relies on Big Data analytics. For instance, astrophysical entities, such as compact regions in active galaxies (active galactic nucleus (AGN)), can be determined from the data produced by physical experiments. Medical diagnoses and therapy make good use of novel sensors, which deliver data from, e.g., one's breath, bodily fluids, or from inside one's body through minimally invasive cameras. The very high-dimensional genomic data is medical Big Data, which offers opportunities to better understand diseases, such as cancer.

Third, Big Data analytics enables new business models. Personalized advertisements, insurance contracts, and even demand-driven movie productions have already become a reality.

It is important that society and its citizens discuss how to derive value from data to address societal needs. The following questions require answers for all data sets and their applications:

- Who gathers the data?
- Who stores the data?
- Who has access to the data?
- Who develops and offers services based on the data?
- Who pays for the data?
- Who benefits from the data?

Our interest in privacy and transparency does not need to conflict. Individuals' privacy and open data that are available to the public can be combined. However, there are still many topics that urgently require broader public discourse.

Let me also name a particular concern of mine: climate change. On the one hand, computation consumes energy and requires cooling. We have to improve this situation by having better systems and processes. On the other hand, computation may reduce energy consumption if processes and products become more efficient and sustainable thanks to Big Data analytics. There is a special feature on data mining for sustainability in the *Data Mining and Knowledge Discovery Journal* (2012) and a book on the topic is also scheduled to come out later in 2016. Both show real-world applications of Big Data for sustainability.

# What are the key focus areas of your Collaborative Research Center SFB 876?

The Collaborative Research Center 876 brings together Big Data and cyber-physical systems. We work on novel frameworks, algorithms for streaming data, highly parallel algorithms, and algorithms for distributed data. The approximation of methods for constrained devices brings analytics closer to cyber-physical systems. On the other hand, cyber-physical systems are enhanced to offer adaptive platforms, measure energy consumption, and open their operating systems.

### Please describe a few exciting projects that you are currently working on.

Basic research on analyzing smartphone data by probabilistic graphical models also made practical applications possible. We may run the learning algorithm on a smartphone so that its operating system can optimize its processes, e.g., prefetching files based on predictions. We have developed an approximation algorithm that learns probabilistic graphical models, which only compute on integer numbers, which hence is very energy-efficient and runs on extremely restricted devices. As the principal investigator, I am working with my Ph.D. student Nico Piatkowski on this topic for project A1 of SFB 876.

We are conducting research on smart factories, in which predictions are based on distributed sensors that predict the quality of the end-result every second of the

process. Applying this to real factories saves considerable energy and materials. I am working on this with my Ph.D. students Marco Stolpe and Hendrik Blom as part of project B3 of SFB 876.

Using Big Data analytics in astrophysics is also a necessity. In this field, the data are so huge, that without smart data analytics, signals cannot be separated from the noise. We are working on data from two experiments that aim to understand astrophysical processes by finding active galactic nuclei. The data from the project IceCube are gathered at the South Pole and our goal is to detect neutrinos. Data from the projects FACT and MAGIC are gathered at La Palma; our goal here is to detect gamma rays. I hope that we can even feed back analysis results to the telescopes; as soon as one telescope detects an event, it should be able to communicate it in such a way that other telescopes turn in the respective direction. Together with my colleague, Wolfgang Rhode, from the Department of Physics, I am leading project C3 of SFB 876.

## In your opinion, what will be the most significant changes in Big Data over the next decade?

Prognosis is based on similar observations in the past. We do not have enough data to predict developments in Big Data.