Teaching Data Management: Key Competencies and Opportunities

Andreas Grillenberger, Ralf Romeike
Computing Education Research Group
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstraße 3
D-91058 Erlangen
{andreas.grillenberger, ralf.romeike}@fau.de

Abstract: Data management is a central topic in computer science as well as in computer science education. Within the last years, this topic is changing tremendously, as its impact on daily life becomes increasingly visible. Nowadays, everyone not only needs to manage data of various kinds, but also continuously generates large amounts of data. In addition, Big Data and data analysis are intensively discussed in public dialogue because of their influences on society. For the understanding of such discussions and for being able to participate in them, fundamental knowledge on data management is necessary. Especially, being aware of the threats accompanying the ability to analyze large amounts of data in nearly real-time becomes increasingly important. This raises the question, which key competencies are necessary for daily dealings with data and data management.

In this paper, we will first point out the importance of data management and of Big Data in daily life. On this basis, we will analyze which are the key competencies everyone needs concerning data management to be able to handle data in a proper way in daily life. Afterwards, we will discuss the impact of these changes in data management on computer science education and in particular database education.

Keywords: Data Management, Key Competencies, Big Data, NoSQL, Databases, Data Privacy, Data Analysis, Challenges

1 Introduction

Nowadays, data take a key position in nearly everyone's daily life. Enormous amounts of data are generated and organized every day, for example when storing documents or music, when using social media, but also in a less obvious way during activities like using public transport (by using electronic tickets), when consulting a doctor (by using electronic health insurance cards), and so on. Therefore, handling data in daily life has many facets: data may be captured wittingly or unwittingly, it may be stored locally or in online (cloud) stores, it may have different structures, and so on. As these data are captured almost everywhere, data analysis enables the reconstruction of large parts of the daily routine with high statistical relevance. For example, this is the case when taking together data coming from public transportation with the payments with credit cards and the data captured by a smart meter used in the private household. With these information, the daily routine like working hours, shopping habits but also which devices are used at home may be reconstructed by analyzing the times one uses public transportation, where someone goes shopping and by analyzing the power consumption information captured by the smart meter (Molina-Markham, Shenoy, Fu, Cecchet, Irwin, 2010). While this example deals with data collected about a person by third parties, in rarer cases people also collect data on themselves: Participants in the "Quantified Self" movement (also known as "life logging") actively gather and analyze data on their own life for different purposes, sometimes for improving health, for improving well-being, or for improving own productivity, in other cases just out of curiosity.

Today, data have a clear influence on daily life and this influence is continuously increasing. A central task is handling these data in a responsible way, as storing, modifying, deleting and using data are typical aspects concerning everyone's life. Storing large amounts of data not only includes selecting an appropriate data store, but also structuring and organizing this data, deciding whether to create backups (and which backup methods to use), synchronizing data between multiple devices (and perhaps users), and so on. This also involves protecting own data and such about other persons from being manipulated, lost or from being used abusively, but also methods for guaranteeing the authenticity of data. For recognizing the necessity for doing so, people also need to recognize the value of data. This is especially possible when analyzing data on their own. Nowadays, such possibilities are opened up for everyone because various comfortable analysis tools are available for free, but also because of the Open Data movement, which targets on publishing as many data sets as

possible in order to allow many-sided usage. These data analysis methods can also enable people to make data captured in daily life valuable for personal use and to extract new information out of known data.

In computer science, the aspects mentioned before are often summarized under the term "Big Data". This is a topical subject in various contexts, as it not only affects people's life as well as CS, but also has strong impact on economy, politics and security. Typically, the term Big Data is described as handling large amounts of data with varying structures and high velocity (Laney, 2001). This includes continuously generating data but also fast processing of such in nearly real-time. Various topics that are being discussed in public dialogue today are strongly affected by Big Data and particularly by data analysis, e. g. early data retention or surveillance programs of intelligence agencies. These topics are often hard to understand, as data management and data analysis are complex topics with rising importance, while the knowledge needed for understanding them is not part of current (CS) education.

In the context of these current developments, the relevance of data management for people changes fundamentally: While until now, data management and specifically databases were topics that were mainly considered as relevant for educational and professional use, these topics are now affecting the whole life of everyone. Despite this changing impact, current teaching considers databases as central topic of data management education (cf. e.g. Brinda, Puhlmann, Schulte, 2009; Seehorn et al., 2011). In the future, other topics like data safety and data privacy will gain importance in everyday life but also in education. Thereby, the purpose of data management education changes tremendously. While the current focus of teaching is set on concepts of and knowledge on databases, in future this emphasis needs to be changed to fostering competencies needed for everyone. Hence, current database education needs to be revised and adapted to these new requirements in order to teach sustainable concepts and aspects of data management that are not only relevant in computer science and computer science education, but also in daily life. So, for being able to deal with the new possibilities and threats, new key competencies coming from data management have to be fostered in class.

In order to point out these new and reappraised key competencies, in this paper we will first describe main aspects of data management and Big Data that are relevant for teaching but also for the students' life. On this basis, we will point out major key competencies, which people need for being able to deal with the new possibilities and threats evolving in context of data management. Finally, we will outline the consequences for computer science educati-

on by discussing main challenges computer science education will have to deal with in the future.

2 Data Management in the Context of Big Data

Handling data is an important task today. This includes planning, organizing and utilizing data, methods which are typically summarized by the term "data management" (Bodendorf, 2005). In addition to these aspects, data management also comprises for example evaluating the quality of data, acquiring data, securing access to data as well as data backup and recovery (DAMA International, Mosley, Brackett, Earley, 2009). These aspects are clearly concerning data management in daily life, as described before.

The topic data management changes tremendously under the influence of current developments like Big Data: Although data management has been an important task in daily life for years, people often only consider it as topic in computer science (education), because the influences of data on daily life were only hardly recognizable so far. Nowadays this influence becomes increasingly obvious: While in 2012 approximately a total of 2.7 Zetabytes (10^{21} Byte) of data existed, about 2.5 Exabyte (10^{18} Byte) of additional data were generated per day (IBM, 2012). However, Big Data is not only characterized by the large amount of data and this high rate of data generation, but also by the variety of data: About 80 % of these data are unstructured or of varying structure, as they are especially coming from social media. For example, in 2012 about 100 Terabytes of data were uploaded daily on Facebook and 230 Million tweets were sent on Twitter per day (IBM, 2012). Therefore, Big Data is often characterized by (at least) three Vs: "volume", "variety" and "velocity" (Laney, 2001), other Vs like the "veracity" of data are added in some descriptions. When dealing with Big Data, new challenges are evolving: usually, the most commonly used relational database management systems are optimized for consistent, durable and reliable storage of data. In contrast, newly emerging databases (often summarized under the term NoSQL[1] databases) set the focus on fast processing of distributed stored data and thus accept limitations, such as of consistency. In addition, other aspects of data management, like data safety and privacy are strongly affected by these new developments.

1 Nowadays, the term NoSQL is interpreted as "Not Only SQL" (Edlich, n.d.), while originally it was used as name for a database management system not supporting SQL at all (Strozzi, 1998).

3 Key Competencies in Data Management

With the increasing impact of data management, CS education needs to analyze which knowledge and skills people need in order to deal with this topic in their life. As data management is a complex demand with strong influence on different fields of daily life, skills necessary for successfully coping with this topic are described as key competencies according to the definition by Rychen, Salganik (2001). These will be derived from all parts of data management – including its main aspects structuring, organizing and utilizing data (Bodendorf, 2005) – but also from discussing the consequences of data management and its influence on data privacy. Therefore, we will hereafter point out the most important key competencies related to the topic data management. These key competencies will be illustrated by examples describing their relevance for people's life.

3.1 Storing data

Nowadays, everybody stores and manages enormous amounts of data every day, e.g. text files, videos, music, e-mails and so on. For storing data, various possibilities exist that strongly differ especially in the following aspects:

- storing data offline or in the cloud,
- storing data as files in a file system or as entries in a database,
- using specialized stores like media stores or not,
- storing data in a structured or unstructured way.

In most cases, data are stored as files in file systems, locally or in the cloud, e.g. when dealing with documents, photos, music and so on. In contrast, there are also data stores that are specialized on only a few data types. For example, e-mail clients are data stores for e-mails but also for contacts and in some cases calendar entries, while music could instead be stored in media libraries, often together with videos and pictures. By using an application specialized for concrete use cases, dealing with specific types of data will be distinctly simplified. As each time when storing data the requirements differ, it is not possible to decide in general if to use specialized stores or not, as these stores also have disadvantages: for example, they often use proprietary file formats and thus comprise the threat of a "vendor lock-in"[2]. Also, all other aspects mentioned

2 The term "vendor lock-in" describes the dependence on a single vendor of products. This is for example the case, when data are stored in a proprietary file format, so that using them in another vendor's product is hardly possible.

above have to be decided from case to case. These decisions are summoned by the question: *"Which data should be stored – and where and how?"* Therefore, handling data implicitly involves knowledge on the different possibilities for storing data.

The decision which data store with which functionalities should be used in a concrete use case has to be made as the case arises. So, for being able to deal with data, and especially large amounts of such, in a proper way, it is necessary that students *understand and apply different ways for storing data.*

3.2 Handling metadata

All these large amounts of data that are generated every day bring additional information with them, the so-called metadata. These are for example visible as attributes of a file (e.g. creation as well as last-modified date, author/creator…), as log files (e.g. in cloud storage services: "file 'example.txt' was deleted on 2014-02-01 09:07 by 'user1'") or as tracked changes in a document. Although most metadata are actually accessible by the user, they are often disregarded: While final versions of a document are typically cleaned from e.g. comments and notes, often metadata are not revised. These data may contain information not supposed to be contained in the finalized document, or they may have been generated from old content, like old or wrong keywords. In addition, information that may be confidential, like the concrete author of a document or the time span between creation and last modification is usually included. There are various examples, where confidential data were disclosed by metadata included in published documents. For example, in 2005, the United Nations published a report on Syria's involvement in a murder; this document not only contained the visible information, but also tracked changes. These annotations were only hidden and contained, i. a. names of persons involved in this plot that were not supposed to be disclosed (Zeller, 2005). Another example for the hidden generation of metadata is the automated geotagging of photos taken by smartphones: typically, such devices add the current GPS position as metadata to all photos taken with this device. When sharing such photos, the user then probably shares more information than intended without noticing it. Therefore, when handling data in daily life, people need the key competency to *note that additional data may be included in data sets as metadata.*

On the other side, these metadata may simplify dealing with data: by adding additional information as metadata, locating data is simplified and accelerated. In particular, metadata are necessary when searching for information by substantial criteria, as most file contents cannot be interpreted directly, so sear-

ching by content might only be possible for pure text files. For example, when dealing with photos, adding metadata that describe the place the photo was taken at, or by marking persons who are visible on it, locating this photo afterwards will obviously be simplified in comparison to searching without such information. As metadata are typically considered by search engines of operating systems, but also within most of the currently used database management systems, being able to deal with metadata simplifies daily data management a lot. However, also the disadvantages of using metadata should be kept in mind: not only the effort of assigning and maintaining them may be relevant for the decision whether such information should be added, because even the usefulness of them strongly depends on the concrete use case. Additionally, while the existence of metadata strongly benefits reading data from the data store, typically writing operations are slowed down, because not only the data but also the metadata need to be updated in order to ensure consistency. Therefore, another key competency in data management is to *understand the purpose of metadata and use them in a proper way.*

3.3 Handling redundancy and consistency

When managing data, people will always have to deal with redundancies and inconsistencies: for example, for files related to multiple topics it often seems reasonable to store copies of the file in multiple folders of the file system. This leads to inconsistent data when one file is being updated while at least one copy is (accidentally) left untouched. Summarizing, data stored redundantly tend to become eventually inconsistent. Since this problem, needing a duplicate copy of a file in another location/folder, is not unusual when saving data, students need to understand the consequences of storing data redundantly. In addition, they have to deal with this requirement, e.g. by creating a link to the data at the second location instead of saving a real copy of it. Therefore, to *understand the consequences of storing data in a redundant way* as well as to *save data in a proper way in order to prevent inconsistencies* are key competencies of data management.

Today, another common cause for inconsistencies is the synchronization of data between multiple devices. Nowadays, one person carries in average 2.9 (mobile) devices including laptops, smartphones and tablets (Truong, 2013), and the overall number of devices used by one person may be even higher. Data are often synchronized between two or more of these devices, and not only read but also modified on these. This leads to inconsistencies when modifying data that was earlier changed in another location, but not successfully

synchronized to the other devices yet. This leads to different possible consequences, dependent on the application used for synchronization and on the type of synchronized data: while only in special cases (such as pure text files) conflicts may be automatically resolved, in most cases duplicate data will come up or in the worst case data will be lost. So, another key competency in this topic, which is needed to be able to understand threats when synchronizing data, is to *understand the consequences of synchronizing data and deal with synchronization conflicts.*

While commonly redundancies and inconsistencies should be avoided, there are also use cases where both concepts are used intentionally: for example, backups are redundant copies of the original files and will become inconsistent as soon as the original files are modified again. However, in this case redundancy occurs by design, because backups serve as fallback copies, especially for the case that data are accidentally deleted or changes must be reverted. So, they need to be redundant to the original file (for restoring) and need to become inconsistent when the original file is being modified (for reversing changes). Today, as in most operating systems different backup functionalities exist, people also need to be aware of the different ways for creating backups: continuous backup vs. backup at discrete points in time, incremental vs. complete backup, hot vs. cold backup. These possibilities clearly differ in used hard disk space, in the speed of the backup and restore processes, and in the typical frequency of backups. For each use case, it is necessary to decide, which aspects are required – a decision which has to be done in context of the value of the concrete data. Therefore, another key competency in this field is to *create redundant data sets for backup/data safety in a proper way.*

3.4 Data safety and encryption

Nowadays, a great part of one's personal life is captured as data. With smartphones and other mobile devices, an increasing amount of moments is immediately captured, for example in form of posts in social media, photos, but data is also in background, e.g. as position data, log files of sensors and so on. Often, the data everyone manages and generates are not only stored on stationary desktop PCs, but also on mobile devices as well as portable USB drives without any security measures, and they are often transferred via insecure communication channels. This results in privacy issues, but also enables even more problematic threats like identity theft. In professional context, financial losses may occur. Storing data on mobile devices or storage media enlarges the risks of unauthorized usage of these data, of manipulations and of data theft.

Consequently, it is an important task to store and transfer private or confidential data in a secure way. This may be reached by different ways. A typical method for securing data in case of theft of the device, on which the data are stored, is to restrict access to the device. This is often accomplished by using password protection or similar authentication methods. Although this will increase data safety, as accessing data becomes more difficult, data safety cannot be guaranteed by this method, because data may yet be accessible stored on the device. A simple approach for overcoming such authentication methods is reading the data store (for example the hard drive) using another device that does not enforce the authentication. Users need to be aware, that usually typical authentication methods cannot suffice to secure their data, as they are only a hurdle for accessing these. Therefore, people must differentiate between restricting the access to devices and to the data itself. To (relative strictly) ensure data safety, it must be prevented that the meaning of data is recognizable without the required permissions. This is the goal of data encryption: While encrypted data might still be read from the hard disk, they have no value for anyone until being decrypted using the right key. So, another key competency when managing data is to *understand the difference between restricting access to a device or service and protecting the data stored on it* as well as to *encrypt data and communication* in order to prevent unwanted access to these data.

Another aspect concerning data safety is to decide whether to confide specific data or not. As for example, the attribute *author* of files, e-mails and so on is typically not protected against changes nor is the content itself; these data carry the risk of being manipulated. As in various use cases it is necessary to be able to trust data, everyone needs to be aware of methods for checking the authenticity of data. For example when reading e-mails, today most people keep in mind that these messages may contain non-authentic content, as they may be somewhat junk or phishing mails. However, with an increasing quality of such messages, it would be harder to figure out if an e-mail is authentic or not. Other data than e-mail are often less questioned, because threats are less obvious and less discussed in public dialogue. Therefore, methods for proving the authenticity of data will become more relevant in the future, especially because an increasing amount of legally relevant tasks is done via electronic communication methods. One technique for guaranteeing the authenticity of the sender information as well as the validity of the content is by digitally signing the data. This enables the recipient to check if data were manipulated. Therefore, it is necessary that people *know methods for guaranteeing the authenticity of data and use them in a proper way.*

3.5 Using methods of Data Analysis

Today, various sets of information and data are available and accessible for free. However, only few people are able to use these data in another way than by just looking at them or analyzing them manually. For example, by combining data from various sources, interesting new use cases may be found as well as new information may be extracted. Today, this is possible for everyone: different simple tools for analyzing data are provided for free by the large Big Data companies like Google or IBM. In addition, there are simple tools for creating mash ups, a form of integrating multiple data and especially media. An example for using open data sets is evaluating whether to book a hotel in a concrete neighborhood in another way than by reading the opinions of former visitors. For example, the City of New York offers many data they capture daily as open data sets.[3] This includes calls to the service number 311, which are concerning complaints on noise, street or sidewalk conditions and so on.[4] While the direct results of analyzing these data are relatively obvious, they can also be combined with other data, such as restaurant inspection results[5], in order to gain prediction factors, for example if the noise conditions in a borough and the ratings of the restaurants in this part of a town correlate or not. Doing such data analysis is possible with simple techniques, for example included in spreadsheet applications or available as online tools. Typical data analysis methods are grouping of data ("clustering", in this case by neighborhood), categorizing them by different characteristics ("classification", in this case e.g. the types of service calls but also the restaurant grades) or by determining interdependencies (if-then-relations) between data ("association", e.g. the described analysis if the restaurant grades and noise conditions correlate). So, another key competency in data management is to *use, find and combine data in order to gather new information*.

By analyzing data themselves, people learn to recognize the threats for data privacy coming from data analysis. With the ability to combine data from different sources, it is only a small step into discovering that the same methods may be used when analyzing personal data. Even data strongly anonymized or pseudonymized according to data privacy acts may be deanonymized – so data privacy acts would be bypassed. This was for example the case, when AOL released a set of search data, which included a user's search terms as well

3 NYC Open Data: https://data.cityofnewyork.us.

4 Analyzation of New York City 311 Service requests: http://opendatabits.com/new-york-city-311-servicerequests-open-data.

5 Meshup of NYC 311 calls together with restaurant inspection results: http://opendatabits.com/nyc-restaurantinspections-results-open-data.

as a unique person ID related to the user, but without revealing personal data of this person. By analyzing these data, some data analysts were rapidly able to recognize some persons out of these data and discovered their real name, contact data, as well as their search habits at AOL's search engine (Barbaro, Zeller Jr., 2006).

Therefore, another key competency of data management, which involves not only data analysis but also data privacy, is to *know the threats for data privacy and analyze data with keeping ethical demands in mind.*

3.6 Being aware of Data Traces and Data Privacy Issues

With the possibility to store and analyze huge amounts of data, different threats for data privacy are evolving. As mentioned before, metadata may be harmful if the user does not know about them, or when handling them in an inappropriate way. In addition, a lack of data safety and encryption strongly affects data privacy. This threat is even intensified when dealing with modern devices, applications and services, as various types of data on this usage and on the user are captured continuously. While the main aim of some services is capturing data in a relatively obvious way, such as in social networks, in other cases they are generated in a hidden way besides the intended use, for example as log files. Additionally, applications supposed to generate data, like the mentioned social networks, tend to store more data than actually needed for the service to work. While in some cases, the user is aware of this data generation and actively decided for capturing these data, such as when participating in the "Quantified Self" movement, this is usually not the case. However, by combining different sources of such data, large parts of daily life may be reconstructed. As nowadays, everyone uses different kinds of data stores, but also applications using these data, one leaves digital data traces everywhere. Hence, the question if one trusts an application/service or not becomes increasingly important for data privacy, as when using an application there is usually no possibility to decide whether it is allowed to capture data.

Examples are typical chat applications that offer the possibility to display "last online" times to contacts: by having a look on these times (and perhaps comparing them to the ones of other persons) other people can digitally trace persons with simplest methods. Depending on the concrete application, this tracking is even possible without prior contact to a person, only by adding them on the contact list without a need for approval of this request by the person added. Therefore, the decision to disclose own data, like these online

times, but also the uploaded photo and the status message, is implicitly made when registering with this service and using it.

So, raising students' awareness on such abilities and threats is an important aspect when discussing data privacy. As such methods for collecting data are typically hard to discover and in most cases cannot be prevented, users need to be aware of these possibilities in order to be able to recognize hints on such issues, e.g. the "last online" times in chat applications mentioned before. This is especially, when confiding large amounts of data to one provider, e.g. when storing data on cloud storage services.

Therefore these aspects of data privacy are summoned by the questions *"Who stores which data on me? Who has which information on me? Who can I confide data about me?"* So, to *note own data traces* but also to *know the possible threats of using data storage services* are important key competencies, which are hard to foster, because tracking such traces is mostly done in an invisible way, as well as threats when using data storage services are typically hard to discover.

3.7 Overview

As described before, several key competencies in data management could be identified. These will be summarized in order to get a complete overview:

People…

- understand and apply different ways for storing data
- note that additional data may be included in data sets as metadata
- understand the purpose of metadata and use them in a proper way
- understand the consequences of storing data in a redundant way
- save data in a proper way in order to prevent inconsistencies
- understand the consequences of synchronizing data and deal with synchronization conflicts
- create redundant data sets for backup/data safety in a proper way
- understand the difference between restricting access to a device or service and protecting the data stored on it
- encrypt data and communication
- know methods for guaranteeing the authenticity of data and use them in a proper way
- use, find and combine data in order to gather new information
- know the threats for data privacy and analyze data with keeping ethical demands in mind

- note own data traces
- know the possible threats of using data storage services

It is clearly visible, that these competencies face different aspects of data management. But at the same time, they are all strongly related to daily life. Additionally, most of these key competencies have one central aspect in common: they represent newly occurring decisions, which are necessary in order to deal with data management in a proper way. This mirrors the current developments in computer science: while until the last years, mainly one database system was leading and used for most use cases, there is nowadays a great variety of such systems which evolved since the development of the NoSQL databases. This makes it necessary to choose the database depending on the use case.

4 Challenges for Computer Science Education

In contrast to their relevance for everyday life, the key competencies described before do not yet receive sufficient attention in current data management education. With the increasing relevance of data management, current curricula for computer science education need to be revised with keeping the new requirements and possibilities in mind. By considering these aspects in class, computer science education changes tremendously: Especially, while current data management education mainly focuses on databases, in the future the relevance of various additional topics will increase clearly, while other current topics may then be less important. The key competencies developed above, need to be considered in computer science education, since no other subject in general educational schools can foster these, because more than basic knowledge on these topics is required.

In contrast to these new requirements, current computer science education mainly focuses on relational database management. Since the topic databases was intensively discussed in the context of computer science education in between the years 1986–1998, only occasionally papers and articles on this topic were published. While in the earlier of these years the relevance of (mainly relational) databases in class was the main topic of research, in the later of these years and now on, the focus of publications is set on supporting the teaching of databases. Therefore, various tools were discussed, especially for teaching SQL, e.g. by Grillenberger, Brinda (2012) and by Sadiq et al. (2004). Since 1998, only few research results on this topic were published, in particular there are no such publications on current developments like Big Data or the increasing relevance of data management in daily life, yet. In addition, currently

no compilation of key competencies concerning this field exists. Only different educational standards, like the K-12 Computer Science Standards by the Computer Science Teachers Association (Seehorn et al., 2011) or the German Educational Standards for Computer Science in Lower Secondary Education (Brinda, Puhlmann, Schulte, 2009), mention some competencies on this topic.

In the following, we will outline some of these competencies for comparing them with the key competencies in data management described above. By having a look on the educational standards, important competencies on data management/database education are found particularly in the topics "structuring data", "data safety" and "data privacy". The former includes aspects of creating and using data structures, e. g. students "know principles for structuring documents and use them in an appropriate way"[6], they "know and use tree structures, for example directory trees"[6] or they "navigate in directory trees and manipulate directory trees in a proper way"[6] (Puhlmann et al., 2008). The latter ones – "data safety" and "data privacy" – need to be distinguished, even though they are strongly related to each other. Data safety focuses on the technical aspects, like preventing prohibited access to confidential data, encrypting such data and so on, while data privacy focuses on using data (and especially personal data) in a proper way. So, competencies needed concerning data safety are for example to *"explain the principles of security by examining encryption, cryptography, and authentication techniques." (*Seehorn et al., 2011), while a typical competency on data privacy is to *"evaluate situations in which private data should be shared"*[6] (Puhlmann et al., 2008). Also, an important competency considering data management is to *"use data analysis to enhance understanding of complex natural and human systems."* (Seehorn et al. (2011)).

By comparing these competencies currently considered as important in computer science education with the ones described before, a clear difference is visible: Although most competencies of current database education will remain relevant in future, various additional ones are supplemented. The competencies fostered today are strongly related to computer science, while in future key competencies of data management will especially face handling data in daily life.

In addition, the topic databases will change clearly in context of Big Data and the newly emerging NoSQL databases. In order to meet the main requirements of storing Big Data, the management of large amounts of data with high variety and high velocity, new types of databases arose. These non-relational

6 Original in German, translation by authors.

databases are typically summarized under the term NoSQL[7] Various concepts of databases that were so far assumed as being fundamental to this topic are dropped by these databases, in order to speed up access to distributed stored data. For example, while consistency is a main requirement of relational databases and part of the ACID[8] criteria, this concept is dropped in NoSQL databases, because they are only "eventually consistent" according to their main requirements summarized as BASE[9]. Therefore, in order to teach sustainable concepts and aspects of data management and databases, the concepts fundamental to databases – and not only for relational database management systems or for NoSQL databases – need to be analyzed.

5 Conclusions

As described in this paper, by considering the new aspects coming from current developments like Big Data and because of the increasing importance of data management for daily life, data management education changes tremendously. By discussing the new aspects coming from these topics in class, various key competencies of data management will be fostered. Although these new key competencies are becoming increasingly relevant, they are mostly not yet fostered in current education on the topics data management or databases. Especially aspects coming from data privacy and data analysis will increase in importance in future data management education, but also all the other key competencies described before need to be fostered, as key competencies are "of prime importance for a successful life and effective participation in different fields of life" (Rychen, Salganik, 2001).

This will also ensure a better fit between education and daily life, because today most people use applications involving Big Data analysis multiple times daily, but without being able to notice the collection of their data or its analysis – and often without even knowing about possible consequences. Additionally, as everyone handles and generates large amounts of data continuously, also the importance of data management in daily life increases continuously: such as when storing data in different data stores (like the file system, media libraries and so on), when synchronizing data between multiple applications and/or devices or when creating backups of data.

7 NoSQL nowadays is interpreted as "not only SQL" (Fowler & Sadalage, 2012). In original, by Carlo Strozzi (1998), this term was used as name for a database not supporting SQL.

8 ACID is the abbreviation for Atomicity, Consistency, Isolation and Durability, the four main requirements on relational databases (Elmasri & Navathe, 2011).

9 BASE is the abbreviation for Basically Available, Soft-State, Eventually Consistent, the three main requirements on NoSQL databases (Edlich, n.d.).

In addition, Big Data has strong impact on current and newly emerging professions. Especially, various professions are changing when considering aspects of Big Data as well as of data analysis. Also, at the moment new professions are evolving, like the data scientist (Davenport, Patil, 2012). In this profession, aspects coming from informatics, like data analysis, are combined with mathematical ones, especially coming from statistics. Therefore, knowledge on fundamental concepts and methods of handling Big Data will have sustainable influence.

References

All electronic sources were retrieved on 17th April 2014.

Barbaro, M., Zeller Jr., T. (2006, August). A Face Is Exposed for AOL Searcher No. 4417749.

Bodendorf, F. (2005). *Daten- und Wissensmanagement [Data and Knowledge Management]*. Springer.

Brinda, T., Puhlmann, H., Schulte, C. (2009). Bridging ICT and CS: Educational Standards for Computer Science in Lower Secondary Education. In *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education* (pp. 288–292). New York, NY, USA: ACM.

DAMA International, Mosley, M., Brackett, M. H., Earley, S. (2009). *The Dama Guide to the Data Management Body of Knowledge Enterprise Server Edition*. Technics Publications Llc.

Davenport, T. H., Patil, D. J. (2012). Data scientist: the sexiest job of the 21st century. *Harvard Business Review, 90*(10), pp. 70–77.

Edlich, S. (n.d.). NOSQL Databases. Retrieved from http://nosql-database.org

Elmasri, R. A., Navathe, S. B. (2011). *Fundamentals of Database Systems*. ADDISON WESLEY Publishing Company Incorporated.

Fowler, M., Sadalage, P. J. (2012). *NoSQL Distilled – A Brief Guide to the Emerging World of Polyglot Persistence* (1. ed.). Amsterdam: Addison-Wesley.

Grillenberger, A., Brinda, T. (2012). eledSQL: A New Web-based Learning Environment for Teaching Databases and SQL at Secondary School Level. In *Proceedings of the 7th Workshop in Primary and Secondary Computing Education* (pp. 101–104). New York, NY, USA: ACM.

IBM. (2012). The Flood of Big Data. *Infographic*. Retrieved from http://ibmdatamag.com/2012/04/managing-the-big-flood-of-big-data-in-digital-marketing

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E., Irwin, D. (2010). Private Memoirs of a Smart Meter. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building* (pp. 61–66). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1878431.1878446

Puhlmann, H., Brinda, T., Fothe, M., Friedrich, S., Koerber, B., Röhner, G., Schulte, C. (2008). Grundsätze und Standards für die Informatik in der Schule: Bildungsstandards Informatik für die Sekundarstufe I [Principles and standards for computer science in schools: educational standards for computer science lower secondary]. *Supplement to LOG IN, 150/151*.

Rychen, D. S., Salganik, L. H. E. (2001). *Defining and selecting key competencies*. Hogrefe & Huber Publishers.

Sadiq, S., Orlowska, M., Sadiq, W., Lin, J. (2004). SQLator: An Online SQL Learning Workbench. In *Proceedings of the 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education* (pp. 223–227). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1007996.1008055

Seehorn, D., Carey, S., Fuschetto, B., Lee, I., Moix, D., O'Grady-Cunniff, D., … Verno, A. (2011). K–12 *Computer Science Standards*. Computer Science Teachers Association, Association for Computing Machinery.

Strozzi, C. (1998). NoSQL: a non-SQL RDBMS. Retrieved from http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/nosql/Home Page

Truong, K. (2013). INFOGRAPHIC: Users weighed down by multiple gadgets – survey reveals the most carried devices. *Sophos Naked Security*. Retrieved from http://nakedsecurity.sophos.com/2013/03/14/devices-wozniak-infographic

Zeller, T. J. (2005, November 7). Beware Your Trail of Digital Fingerprints. *The New York Times*. Retrieved from http://www.nytimes.com/2005/11/07/business/07link.html

Biographies

Andreas Grillenberger is research associate at the Faculty of Engineering and doctoral candidate at the Computing Education Research Group of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany.

Ralf Romeike is the head of the Computing Education Research Group at the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany.