N.T.M. 2017 · 25:519–542
DOI 10.1007/s00048-017-0180-9
Published online 27 October 2017

"Mr. Database"

Jim Gray and the History of Database Technologies

Nils C. Hanwahr

„Mr. Database". Jim Gray und die Geschichte der Datenbanktechnologien

Auch wenn die verbreitete Verwendung des Begriffs Big Data vergleichsweise neu ist, geht dieser zurück auf ein Phänomen in der Entwicklung von Datenbankentechnologien mit eindeutig historischem Hintergrund. Der Informatiker Jim Gray, im Silicon Valley bekannt als „Mr. Database" bevor er 2007 auf See verschollen ging, war an vielen entscheidenden Entwicklungen seit den 1970er Jahren beteiligt, die die Basis für immer größere, schnellere und dezentralisierte Datenbanken bilden. Auf Grundlage der von Edgar F. Codd bei IBM konzipierten Prinzipien war Jim Gray an der Entwicklung von Relational Database Systemen beteiligt, und entwickelte später selbst Standards des Transaction Processing. Außerdem wirkte er mit daran, Austauschforen zwischen Wissenschaft und Industrie zu schaffen, die Funktionsstandards und Forschungsprogramme beeinflussten. Als Mitbegründer von *Microsoft Research* in San Francisco wandte sich Gray der wissenschaftlichen Anwendung von Datenbanktechnologien zu, etwa im TerraServer Projekt, einer Onlinedatenbank von Satellitenbildern. Inspiriert von Vannevar Bushs Idee des Memex entwickelte Gray seine Vision eines Personal Memex sowie eines World Memex, und postulierte letztlich ein neues Zeitalter der auf Daten basierenden wissenschaftlichen Entdeckung genannt „Fourth Paradigm Science". Dieser Artikel gibt einen Überblick über Grays Beitrag zur Entwicklung von Datenbanktechnologien sowie seiner Forschungsagenda und zeigt, dass zentrale Ideen rund um Big Data die Akteure der technologischen Entwicklung schon sehr viel länger beschäftigten als der Begriff selbst in Verwendung ist.

Schlüsselwörter: Jim Gray, Memex, Relational Database Management Systems, Transaction Processing, TerraServer, Fourth Paradigm Science

Although the widespread use of the term "Big Data" is comparatively recent, it invokes a phenomenon in the developments of database technology with distinct historical contexts. The database engineer Jim Gray, known as "Mr. Database" in Silicon Valley before his disappearance at sea in 2007, was involved in many of the crucial developments since the 1970s that constitute the foundation of exceedingly large and distributed databases. Jim Gray was involved in the development of relational database systems based on the concepts of Edgar F. Codd at IBM in the 1970s before he went on to develop principles of Transaction Processing that enable the parallel and highly distributed performance of databases today. He was also involved in creating forums for discourse between academia and industry, which influenced industry performance standards as well as database research agendas. As a co-founder of the San Francisco branch of *Microsoft Research*, Gray increasingly turned toward scientific applications of database technologies, e. g. leading the TerraServer project, an online database of satellite images. Inspired by Vannevar Bush's idea of the memex, Gray laid out his vision of a Personal Memex as well as a World Memex, eventually postulating a new era of data-based scientific discovery termed "Fourth Paradigm Science". This article gives an overview of Gray's contributions to the development of database technology as well as his research agendas and shows that central notions of Big Data have been occupying database engineers for much longer than the actual term has been in use.

Keywords: Jim Gray, Memex, Relational Database Management Systems, Transaction Processing, TerraServer, Fourth Paradigm Science

How do you know?

In January 2003, database engineer Jim Gray released a memo titled "How do you know?" to his colleagues at *Microsoft Research* in San Francisco. The memo was a meditation on what Gray's own work on database technologies had aimed to accomplish: "Wouldn't it be nice if we could just put all the books and journals in a library that would automatically organize them and start producing new answers?" (Gray 2003) Not only were databases supposed to store information in digital form, Jim Gray also wanted them to automatically generate creative ways of compiling the trove of knowledge into novel assemblages of insight. He went on to ask: "How can knowledge be represented so that algorithms can make new inferences from the knowledge base? This problem has challenged philosophers for millennia. There has been progress." (Gray 2003). While the claim that the representation of knowledge in a form that renders itself useful for computation has been an issue of philosophy for thousands of years is overstated, it has certainly been a challenge that led computer scientists to develop tools that are today assembled under the heading of Big Data. Progress has indeed been made in representing knowledge in forms accessible to algorithms, and yet this progress has a history that is closely related to the life of the author of the "How do you know?" memo, Jim Gray.

"Database researchers have labored to make it easy to define the schema, easy to add data to the database, and easy to pose questions to the database" (Gray 2003), Gray went on to write in his memo. By 2003, the issues of sorting, indexing, and organizing information had essentially been solved by deploying relational database management systems that are widely used in science and business applications to this day.[1] Jim Gray summed up the development of relational databases based on the ideas postulated by Edgar F. Codd in 1970, "the research community embraced the relational data model championed by Ted Codd. [...] After a decade of experimentation, these research ideas evolved into the SQL database language." (Gray 2003) Next, database engineers had to address the issues of how to build a database that could be spread out over various storage media, be accessed by multiple queries in parallel, and still be reliable at a level that enables one to put trust in making purchases online or carrying out financial transactions via online banking. Yet, Gray's framing of progress in database technology overlooks a more complicated history than his memo suggests.

This paper tells the story of how Jim Gray was involved in creating database technologies that allow us to sort, index, and organize information, and then went on to develop principles of transaction processing that ensure the concurrency and reliability of databases. Concluding his memo,

Jim Gray wrote: "Over the last decade, the traditional database systems have grown to include analytics (data cubes), and also data mining algorithms borrowed from the machine learning and statistics communities." (Gray 2003) Eventually, the aim of creating databases that allow for new knowledge to be gained by applying algorithms began to be realized by deploying a combination of machine learning and database technology that we often call Big Data.

In 2007, Jim Gray was lost at sea while sailing his ship *Tenacious* off the coast of San Francisco. After the US Coast Guard had to abandon the search for his sailing yacht, many of Gray's friends in the database community attempted to find him using just the distributed database systems that Gray had helped to develop. The story of the search and the methods used are aptly discussed in an article by Gray's colleague Joe Hellerstein (Hellerstein & Tennenhouse 2011). Yet, all the technology could not locate neither the man nor his ship, and Gray was eventually pronounced dead in absentia in 2012. A "Tribute to Jim Gray" was held at UC Berkeley in 2008 by Gray's family and former colleagues, whose contributions were published as a special issue of the journal SIGMOD Record. Certainly, the contributions to this tribute volume have to be regarded as the eulogies that they are, yet nevertheless, they contain valuable and highly personal information. It is difficult to find sources that are directly critical of Gray and his work, which could indicate that it is still too soon for a truly critical assessment. This paper is an attempt at such an assessment or at least a contextualization of Gray's work and positions.

Jim Gray was also actively involved in selling a narrative of linear progress in the development of database technologies that he deployed to influence research agendas, omitting the frustrations and dead-ends of research and technology development. To trace both Jim Gray's work as well as his influence on the discourse among the database technology community, this article draws on several original sources. Gray himself made available much of his personal and professional communication such as memos, technical reports, workshop presentations, and conference talks on his personal website. Furthermore, as a tireless networker and mentor of many computer scientists, he was connected to numerous people in Silicon Valley, who, in turn, make frequent references to Gray in both oral history interviews as well as interviews in newspapers and the trade press. To assess Gray's impact and influence on discourse and technology development, I draw on several of these public sources. Jim Gray believed that database technology held the promise to change the way knowledge comes into the world, an idea he called "Fourth Paradigm Science". This paper will also attempt to trace some steps in Gray's work such as his concepts of Transaction Processing, his work on the *Microsoft* TerraServer, and

his ideas of eScience to put current debates and claims about the powers and promises of Big Data, whether in commerce or science, into a broader perspective.

Mr. Database and Mr. Memex

First, we should remind ourselves of the source of the idea to create a universal library comprising automated knowledge, accessible to everyone, capable of generating new insights algorithmically, which was echoed in Jim Gray's "How do you know?" memo. Many computer scientists have been fascinated by a concept that Vannevar Bush, the Director of the Office of Scientific Research and Development during the US postwar years, had developed in an article titled "As We May Think" in the July 1945 Issue of *The Atlantic* magazine: the memex (Bush 1945).

Jim Gray included Vannevar Bush's article at the top of a recommended readings list on his personal website and frequently referred to Bush's ideas.[2] "As We May Think" addressed the swift expansion of information and information technology that has taken place during the Second World War: "Science […] has provided a record of ideas and has enabled man to manipulate and to make extracts from that record so that knowledge evolves and endures throughout the life of a race rather than that of an individual." (Bush 1945) The scientific record, of course, can also be expanded via the media of writing, books, and libraries. However, Bush envisioned such a rapid growth of information, that new technological means are necessary to store and consult the ever-expanding record of knowledge (Bush 1945).[3]

Bush was focused on analog storage media such as microphotography rather than digital storage media, and yet, his idea has inspired much of the work of Jim Gray up to his talk on Fourth Paradigm Science in 2007, which will be discussed later on (Hey et al. 2009). Remarkably, the pitfalls of Big Data are already formulated in "As We May Think", when Bush writes "we seem to be worse off than before—for we can enormously extend the record; yet even in its present bulk we can hardly consult it." (Bush 1945) This is to be achieved by the personal and associative indexing of the memex that each individual uses to trace her path through the universal record of knowledge. Despite envisioning the memex to take the form of a wooden desk-like contraption including levers, Bush had sketched out not just what drove the development of the personal computer in the 1970s, but also what could be called eScience. Notably, this appears to call for historical research as much as for scientific inquiry, thus also

foreshadowing what we have come to call Digital Humanities. Reminding us of how "As We May Think" was published just weeks before the dropping of the first nuclear bombs on Hiroshima and Nagasaki, Bush closes on a cautionary note, remarking that man "may perish in conflict before he learns to wield that record for his true good." (Bush 1945: 26).

Postwar computing in the United States, however, was dominated by the military concerns of the Cold War and focused on cryptography and cybernetic control of ballistic missiles. Thus, the power of supercomputers was taken to be a measure of progress in computing, more so than database technologies.[4] Yet, the predominance of supercomputing as the main concern of national digital infrastructure projects should also be challenged. Today, Big Data is not about the amazing speed and power of supercomputing centers, but about the amount of data in distributed systems and the kinds of novel analytics employed to mine this trove of information.

In May 2003, the National Research Council's Computer Science and Telecommunications Board met at Stanford University to listen to a presentation by Gordon Bell and Jim Gray, both of them working at *Microsoft Research* at the time. "Gordon and I have been arguing that today's supercomputer centers will become superdata centers in the future" (Markoff 2003) Jim Gray is quoted by *New York Times* technology correspondent John Markoff. While United States science policy had funded immense supercomputer infrastructure programs since the 1980s, the two IT engineers were arguing that it was no longer computing capacity but data storage capacity and ease of access that was crucial to scientific computing.[5] "Central to the Bell-Gray argument is the vast amount of data now being created by a new class of scientific instruments that integrate sensors and high-speed computers" writes Markoff. Basically, Bell and Gray argued for a reorientation of strategy for US scientific infrastructure policy, turning away from the focus on powerful supercomputers able to run intricate simulations of weather or war, toward an infrastructure that forms the foundation of computers and sensor networks by providing database technologies for the entire scientific community, reminiscent of Vannevar Bush's notion of the Memex.

Jim Gray's colleague and friend Gordon Bell was another central figure in the history of database technology and had been personally involved in establishing what became the World Wide Web through his participation at the National Science Foundation's Computing and Information Sciences and Engineering Directorate and his work on the National Research and Education Network in the late 1980s.[6] Keenly aware of technology history, he introduced one of computer science's "laws" in a 2007 paper titled "Bell's Law for the birth and death of computer classes" postulating that roughly every ten years, a new kind of computing device would come along

that rendered previous systems obsolete (Bell 2007). For example, personal desktop computers have eventually come to be replaced by various mobile and connected computing devices such as tablets and smartphones. And yet again, in 2003 Bell and Gray were announcing a new era, arguing that "data-storage technology is now significantly outpacing progress in computer processing power, […] heralding a new era where vast pools of digital data are becoming the most crucial element in scientific research." (Markoff 2003) In essence, Bell and Gray were announcing nothing less than an era of Big Data in scientific infrastructures to the National Research Council in 2003, without actually mentioning Big Data by name.

The use of the term Big Data in its current, albeit hazy, understanding is hard to pin down, although some trace it back to work by a certain John Mashey at Silicon Graphics in the late 1990s (Lohr 2013). An economist named Francis Diebold has also claimed to have coined the term and has published several versions of a paper, attempting to track the term Big Data back to about the year 2000 (Diebold 2012). And yet, it was only around the "year 2008, according to several computer scientists and industry executives, […] when the term 'Big Data' began gaining currency in tech circles." (Lohr 2012) Notably, Gordon Bell and Jim Gray, even though on the payroll of *Microsoft Research* at the time, were not speaking as marketing salesmen trying to promote a hype around Big Data; they had been working in and around Silicon Valley since the 1960s and had been involved in developing the foundation of the range of technologies that constitute Big Data today.

Jim Gray himself did not live to experience the height of the Big Data hype. He had, however, received the Turing Award, one of computer science's most prestigious awards, in 1998 for his contributions to the development of transaction processing. Transaction processing, which Grey introduced in the 1980s, has been called one of the most important algorithms of the modern world by the computer scientist and author John McCormick (Maccormick 2012: 148). The following sections will trace Jim Gray's work and career as a central figure in database technology and seek to contextualize some of the developments that lead to the assumption of a Big Data era. It is especially noteworthy how Jim Gray frequently used reflection on the historical development of database technologies to contextualize his own work and thinking in various timelines of technological breakthroughs. As a keen networker, who was well connected in the Bay Area tech community, Gray deploys the narratives of an amateur historian to locate himself within technology history and harness the focus of a research community to rally around his predictions and research agendas.

We need to be aware of how what a database is has changed crucially over time. Not just the storage hardware has been transformed from

punch-cards to magnetic tape, to hard-disks, and flash memory, but crucially the way databases were conceptualized and how one could query a database to get answers to specific questions was constantly evolving. Database technologies developed by computer scientists such as Jim Gray have enabled databases to be distributed and yet reliable, they are in ubiquitous use in the background of most digital applications, and yet the question of where a database is and what it consists of has become ever harder to pin down.

Relational Databases—Sort, Index, Organize

Born in San Francisco in 1944, Jim Gray trained as a mathematician and computer scientist, and spent practically his entire life in the San Francisco Bay Area and Silicon Valley. Following his undergraduate studies at UC Berkeley, Gray completed a PhD in computer science, allowing him to be exempt from the dreaded military draft during the ongoing Vietnam War. Following his doctorate, Michael A. Harrison, Gray's doctoral advisor at UC Berkeley encouraged him to stay in Berkeley for two more years as an IBM-affiliated post-doctoral researcher. Harrison later remarked on how spell checkers would have been a blessing for the young computer scientist, stating, "It was always surprising to me that, for someone so smart, Jim was so poor at spelling." (Harrison 2008).

Gray then went to work for IBM in 1971 at the *IBM Research Center* in San Jose, where Edgar F. Codd had just developed the concept of relational databases (Codd 1970). "Jim Gray, who we all know, knows everybody" fellow database engineer Michael Stonebraker said of him in 2007, is "is the kind of guy that just pokes his nose into everything." (Stonebraker 2007) Although competitors while Gray was involved in developing IBM's first relational database management system, called System R, and Stonebraker was building the competing INGRES database system at UC Berkeley, Jim Gray appears to have had a talent for networking and was frequently in touch with the Berkeley competitors.

Jim Gray is widely recognized to have had a significant influence on the development of database technologies since the 1970s. Following his disappearance at sea in 2007, a colleague at *Microsoft* pointed out that "Jim was one of the fathers of the database industry as we know it today. While databases were invented, per se, in the late 60's and early 70's, those early systems were not usable in most practical terms." (Vaskevitch 2008) Michael Stonebraker points out in his textbook *Readings in Database Systems* that the most influential and enduring work on IBM's System R

was Gray's contribution: "The transaction manager is probably the biggest legacy of the project, and it is clearly the work of the late Jim Gray. Much of his design endures to this day [2015] in commercial systems." (Bailis et al. 2015).

Looking back at his own work at IBM in the 1970s, Gray published a technical report at *Microsoft Research* titled "Data Management: Past, Present, and Future" in 1996, in which he places his own work in a broad historical context and traces what he believes to be six generations of data management in the history of technology:

> There have been six distinct phases in data management. Initially, data was *manually* processed. The next step used *punched-card equipment* and *electro-mechanical machines* to sort and tabulate millions of records. The third phase stored *data on magnetic tape* and used stored program computers to perform batch processing on sequential files. The fourth phase introduced the concept of a *database schema and online navigational access to the data*. The fifth step *automated access to relational databases* and added *distributed and client-server* processing. We are now in the early stages of sixth generation systems that store richer data types, notably *documents, images, voice, and video data*. These sixth generation systems are the storage engines for the emerging Internet and Intranets (Gray 1996; emphasis by author).

By manual processing, Gray means any analogue media from Sumerian clay tablets to writing and printing on paper and in books. Whether a cultural capability such as speech and writing can be reduced to information processing in a manual way is questionable, however, for scientific and commercial purposes, writing and print were used for the same ends that are today addressed by database technologies. Gray places the second era, the time of punch-cards, between Hollerith's use of them in the 1890 US census and roughly 1955. In 1951, the UNIVAC1 was delivered to the US Census Bureau and replaced thousands of punch-cards with it magnetic tape storage.[7] These databases, however, were file-oriented and used batch transaction processing, making the databases error-prone and slow to update. Online transaction processing overcame the limitations of this era to enable the use of direct access databases for applications such as stock-market trading or booking reservations by travel agents. The *Data Base Task Group* (DBTG) and *General Electric* engineer Charles Bachman developed this kind of new database, for which Bachman received the Turing Award in 1973.

Throughout the 1970s, Jim Gray had worked on developing the fifth step of his genealogy of database technologies at IBM when he was involved in constructing the major relational database management system

of the time, IBM's System R. To this day, basic relational databases use a programming language derived from the foundations of System R, the Structured Query Language, known as SQL. "In the context of the System R relational database project at IBM Research, Jim Gray developed and refined recovery techniques that ensure the reliability of the records and concurrency control methods to coordinate interactions among simultaneously executing programs accessing and modifying shared sets of records" (Lindsay 2008) Gray's former colleague Bruce Lindsay sums up his contribution.

However, IBM was unable to capitalize on the development of Jim and his colleagues. In fact, the company licensed the code for System R out to a company that is known today under the name *Oracle*, with its founder Larry Ellison. Gray comments on this technology transfer:

> Perhaps the most frustrating thing for me has been the technology transfer business. […] However, our most successful transfer has been to Relational Systems, a company which sells a System R look-alike called Oracle. Oracle entered the market this year. It is nicer than System R in many ways. Why is it that IBM, to whom we gave both the code and years of consulting, is five years behind Oracle which started in 1977 with only the System R syntax and examples? To give another example, all our ideas about distributed database are being implemented by Tandem. They credit *us* with the design. IBM is not planning to use our ideas until the late eighties (Gray 1980a).

In fact, IBM did not bring a relational database to market before 1982, naming their first commercial relational database product DB2. However, the main competitor of *Oracle's* relational database systems were not IBM's products but a group around Michael Stonebraker and Gene Wong at UC Berkeley, who developed a database system called INGRES. There had been, as was mentioned above, a spirit of collaboration between the rather academically inclined database engineers at *IBM Research* and the INGRES team, and Jim Gray frequently crossed the San Francisco Bay to meet with the INGRES developers at his Alma Mater. The competition between Michael Stonebraker's company *Relational Technology* and Larry Ellison's *Oracle*, who had licensed the technology that would become SQL from IBM, was fierce. By the early 1980s, Oracle had essentially taken over the market by aggressive marketing methods, which left Stonebraker with some resentment: "Larry Ellison had no qualms about lying to his customers", he commented in 2007 (Stonebraker 2007).

Yearning for a more dynamic and commercially oriented work environment, Jim Gray eventually quit his job at *IBM Research*: "I am resigning my position at IBM Research because it is seventy-five minutes from my

home and I am a little tired of commuting" (Gray 1980a) is how he started his resignation letter in 1980. After several comments on commuting and IBM's apparent preference to locate their research centers far away from the urban centers that Gray seemed to prefer, he goes on to lay out his personal understanding of what it means to do research: "Perhaps I should begin with a very personal statement: I aspire to be a scholar of computer science. All fields of scholarship, from religion to medicine emphasize three aspects: meditation, teaching and service." (Gray 1980a).

His frustration appears to have been long in the making, since he had circulated memos in the company before, decrying the lack of computing infrastructure and commercial product orientation at IBM. "When I left UC Berkeley to join IBM, I was surprised to find that the university provided better computing services than IBM." (Gray 1980b) Not before he entered *Microsoft Research* in 1995 would Jim Gray be able to work full time as a scholar of computer science. Yet, for the moment, Gray moved on to one of the first Silicon Valley companies that were fostering the sort of experimental work environment that so many start-ups attempt to emulate today, *Tandem Computers* in Cupertino, California.

Transaction Processing—Setting Standards

Pat Helland, an early employee of *Tandem Computers*, said about his work on fault-tolerant database systems at *Tandem*: "We read LOTS of papers but the ones that mattered were written by this fellow named Jim Gray who worked at IBM." (Helland 2008) *Tandem Computers* had been founded in 1974 and built commercial database applications that required an especially high level of reliability, such as bank transactions, cash machines, stock exchanges, and airline booking centers (Clemson 2012). Tandem's culture appears to have been the polar opposite of the corporate juggernaut IBM. As a young company, it was still run by its founders and had an "unusual [...] culture which has been adopted and adapted by many startup companies" (Nauman 2008) states former colleague John Nauman.

The *Tandem* products were supposed to process database transactions without interruptions, and were thus called NonStop. Jim Gray arrived from work on IBM's relational databases, System R and DB2, including its query language SQL, and used his experience to combine SQL for relational databases with the fault-tolerant systems developed by *Tandem* to create NonStop SQL. This was a strategic pivot for *Tandem*, since most commercial users of databases did not use SQL-based systems for their crucial distributed systems. However, Gray was able to convince *Tandem*

that an SQL-based version of their NonStop industrial product was the most cost-effective way to move forward. "NonStop SQL was developed by a relatively small team, many of whom Jim recruited from outside Tandem. He served as everything from architect to developer to cheerleader within the team while at the same time continuing to explain the benefits to *Tandem's* upper management" (Nauman 2008) John Nauman elaborates.

Gray is also credited with developing what is to this day known as the "ACID test" for database transactions. ACID is the acronym for atomicity, consistency, isolation, and durability. Atomicity postulates that one database transaction shall never be split or carried out only partly. One transaction has to be either carried out completely or it has to be rolled back in case of any faults. For example, a bank transfer has to comprise a change in both the origin and the destination account of the transfer, otherwise transferred money could either be lost or generated out of the blue. Thus, atomicity ensures the consistency of the databases involved in the transaction, although different types of databases will require appropriate conditions of consistency. Isolation is a crucial condition when a large number of transactions are processed in parallel online or in a distributed system. To ensure the efficiency of the process, "each transaction must appear to be executed as if no other transaction is executing at the same time" (Garcia-Molina et al. 2013: 9) even though in practice, many transactions are processed in parallel. Finally, durability means that it has to be ensured that after the completion of a transaction, changes in the database cannot somehow be corrupted, which would once again render the databases inconsistent.

Furthermore, Gray was involved in introducing performance benchmarks for database transactions. Moving from software engineer into a product development role at *Tandem Computers*, he was more frequently in contact with customers. "Jim kept a suit hanging on the back of his office door. If someone needed a technical spokesperson to address a customer's concerns, Jim could transform himself from a dressed-down engineer/architect to a super-product-manager" (Nauman 2008) a co-worker describes his evolving role at *Tandem*. By 1985, Gray had also published his theoretical considerations of what transaction processing benchmarks could be in his papers "One Thousand Transactions per Second" and "A Measure of Transaction Processing Power".[8] The setting of standards and measures to make performance comparable seems to have appealed to Gray as a natural networker. Also in 1985, Gray started the High Performance Transaction Systems (HPTS) Workshop. The HPTS Workshop is still held every two years on the Asilomar Conference Grounds in Pacific Grove, California, and is currently being co-organized by Gray's former colleague at *Tandem*, Pat Helland. The workshops bring

together computer science researchers from top universities with database engineers from the largest Silicon Valley companies, including *Amazon, Google*, IBM, and *Oracle*.[9]

Another yet more institutionalized forum for database hardware and software manufacturers to discuss industry standards was launched upon encouragement by Jim Gray in 1988, the Transaction Processing Performance Council (TPC) (Dewitt & Levine 2008). All of the institutions have established themselves as joint forums for database technology researchers from academia and the private sector, enabling the practitioners to exchange their experiences and collaboratively adjust the research agenda to address issues encountered in commercial applications.

The linear narrative of Big Data overlooks the importance of standards in measuring and comparing the performance of database systems. Without a common way of assessing the "size" and velocity of a database, postulates of new achievements remain vacuous.[10] In an IBM whitepaper, five "Vs" of Big Data are described to characterize the phenomenon: volume, variety, velocity, viability, and value. Especially the volume and velocity parameters of a database cannot be measured without a form of standard to compare the performance of various database and transaction systems.

In addition to networking in the commercial and academic database research community, Jim Gray also aimed to unify the field by creating common ground in the teaching of database technologies. Gray cited "meditation, teaching, and service" as his central career aims in his IBM resignation letter, yet, immersed in research and involved in a commercial company such as *Tandem Computers*, Gray did not regularly teach. Yet still, as a networker and mentor, his desire to teach had not vanished. In 1987, he wrote in a letter to his wife Donna Carnes: "I bought a Mac to write the Great American Technical Novel. I was to start March 16th, but now it is April 27th and I have yet to do anything on it. […] So in June I'll take a leave of absence from Tandem and devote myself to writing." (Carnes 2008).

Gray had taught a one-week seminar on transaction processing in Berlin in collaboration with the German academic Andreas Reuter in early 1987, and the two decided to turn the slides of their workshop presentations into a textbook. Yet, the project stalled for several years until Gray and Reuter "decided to rent a house in a small village in Tuscany named Ripa (near Carrara) and spend February through April of 1990 there." (Reuter 2008: 55) After another stint of focused writing, the textbook ended up being longer than a thousand pages and was published in 1992 under the title "Transaction Processing—Concepts and Techniques" (Gray & Reuter 1992). Usually, textbooks in computer science have a short half-life. Yet,

the textbook was well received and is still in print as one of the major texts on Transaction Processing nearly twenty-five years after its publication.

By the early 1990s, when Jim Gray left *Tandem Computers* to work for *Digital Equipment Corporation*, he had not only contributed to major developments in relational database technology and transaction processing, but had established himself as a major figure in setting standards for database performance measures as well as in teaching following generations of database engineers.

Microsoft TerraServer—a Virtual Earth

In 1995, Gordon Bell, another former employee of *Digital Equipment Corporation*, and Jim Gray were the founding directors of the *Microsoft Research Center* in the Bay Area.[11] Just after Gray had arrived at *Microsoft Research*, the company envisioned to launch a project that was supposed to impressively display to their competitors that they were capable of creating the largest online database ever conceived at the time. According to his colleague Tom Barclay, Gray was initially reluctant to work on a project that was merely a scaled-up version of an old technology, questioning the research value of such an endeavor (Barclay 2008). Yet, he appears to have been convinced by the challenge to construct an online database that exceeded one terabyte of data, postulating that the team should aim to "find both an interesting tera-byte *and* a cheap tera-byte." (Barclay 2008) Eventually, *Microsoft* chose the goal of providing images of the surface of the globe for its terabyte database ambitions and christened the project TerraServer.

Jim Gray led the TerraServer project and was able to establish a cooperation with the United States Geological Survey (USGS) to incorporate more than 2.3 Terabytes of their grayscale images. To acquire satellite images, Gray and several colleagues went on a trip to Russia, where they were able to forge a cooperation with *Sovinformsputnik*, who provided more than one terabyte of recently declassified Russian military satellite images at a resolution of about two meters. The cooperation had been established via the small firm *Aerial Images* that was attempting to capitalize on the opening up of regulation concerning the distribution of high resolution satellite images following the collapse of the Soviet Union (Barclay et al. 1999).

However, the Russians from *Sovinformsputnik* were only willing to provide the satellite images on the condition of personally meeting with the project's directors. The Russians wanted *Microsoft* to guarantee data se-

curity as well as the promise to construct and online platform for the commercial distribution of images by their US partner *Aerial Images*. Furthermore, they wanted to publicly announce the cooperation with *Microsoft* during a press conference with the Russian Space Agency. Eventually, an agreement was reached and Jim Gray participated in a press conference in Moscow announcing the cooperation between *Microsoft* and *Sovinformsputnik*. Before the Americans returned to California, the agreement was celebrated with a "nine-course meal and [we] participated in 27 vodka toasts […] We didn't sober up until we arrived back in the US two days later," (Barclay 2008) Tom Barclay recollects.

Thus, Gray and his team were able to begin constructing TerraServer in late 1996, and the online database of satellite images and aerial photographs was eventually launched on 22 June 1998. According to a *New York Times* article covering the launch of TerraServer, *Microsoft* had initially "considered creating a database for major league baseball statistics, or of every trade in the history of the New York Stock Exchange, but neither project provided enough data to suit its goals." (Richtel 1998) While *Microsoft* was dominating the operating systems market with Windows and the consumer software market with its MS Office products, the market for commercial business databases was firmly held by the old rivals IBM and *Oracle*, and not *Microsoft*'s SQL Server software. IBM spokespeople were quick to denounce *Microsoft's* claim to the largest existing database, stating "We've been at this for a while. It's good to see other companies learning to put large databases on the Internet." (Richtel 1998).

Of course, to reliably test scalability, the project would not only have to include a very large database, but would also have to attract millions of users to access the database and prove its capabilities. The TerraServer team had initially estimated a demand of about 250,000 page views per day, which was later expanded to an estimate of one million daily views. However, once TerraServer went officially online on 24 June 1998, there was a demand of more than eight million views a day, which forced the team to expand their capacity from one to ten webservers, just to be able to deliver the content at a reasonable bandwidth (Barclay et al. 1999: 5). Eventually, TerraServer was integrated into follow-up projects such as *Microsoft Virtual Earth* and *Bing Maps*, while the TerraServer website itself is no longer available.

Setting Research Agendas

In 1998, Jim Gray received the most prestigious award of the computer science community, the Turing Award. His acceptance speech was later released as a technical report at *Microsoft Research*, titled "What Next? A Dozen Information-Technology Research Goals". In his speech, Gray speaks of cyberspace as a new frontier, a "New World": "One way to think of the Information Technology revolution is to think of cyberspace as a new continent—equivalent to discovery of the Americas 500 years ago." (Gray 1999) Referring to his work as a member of the Presidential IT Advisory Committee, Gray called for a "Lewis and Clark style expedition into cyberspace." (Gray 1999: 4).

On the one hand, databases are supposed to create a representation of the world, which is supposed to render new insights into the physical world. On the other hand, Gray construes information technology as a new continent unto itself, which we are supposed to explore. This is a striking inversion of world and database, a construal that hints at Jim Gray's ideas of a new kind of epistemology associated with database interfaces that we will discuss further in the section on Gray's idea of the Fourth Paradigm.

Gray's talk also hits upon a central dilemma of past and current science policy, the question of whether the results of publicly funded research should be available for free to the public that has funded it in the first place. Furthermore, should not the public profit from the gains that are made by the commercialization of products based on such publicly funded research?[12] The unresolved problems arising from the new ubiquitous storage are the issues of privacy and intellectual property in cyberspace. "So, why isn't everything in Cyberspace? Well, the simple answer is that most information is valuable property and currently, cyberspace does not have much respect for property rights." (Gray 1999: 15) While the amount of information available online today has skyrocketed even in comparison to twenty years ago, the issues of privacy and intellectual property remain unresolved and have only become more pressing.

Gray acknowledges the issue of privacy when he posits the creation of a "Personal Memex" technology, a "box that records everything you see, hear, or read", as a research goal. Similarly, the idea of a "World Memex", Vannevar Bush's "vision of putting *all* professionally produced information into Memex" appears to Gray as a research goal within close reach since "we are getting close to the time when we can record most of what exists very inexpensively." (Gray 1999: 16, emphasis in the original) Not only is a World Memex supposed to be able to store text and other media, it is also supposed to "answer questions about the text and summarize the text as precisely and quickly as a human expert." (Gray 1999: 17) At

the time of Gray's Turing Award speech, the term Big Data was not in wide use yet, in fact Gray does not mention it at all, and yet, the Personal Memex and the World Memex as Gray construes them are a Big Data vision *avant la lettre*. An obvious model of what a Personal Memex could be are current smartphones that include various sensors and enough storage to carry around media such as pictures, music, and video files. However, the launch of a smartphone such as the iPhone in 2007 required another step in storage technology: flash storage.

While the first iPod music player, introduced in 2001, still contained a small hard disk drive, flash storage is more suitable to mobile devices, since it consumes less energy, creates no noise, and cannot be disrupted by motion. Jim Gray once again anticipated the rise of flash storage technology and summed it up in a talk given in 2006, stating "Tape is Dead, Disk is Tape, Flash is Disk, RAM Locality is King" (Gray 2006). Magnetic tape and its smaller offspring floppy disks had long since been out of use, while the market for hard disk drives had been growing and innovating relentlessly since the 1980s (Christensen 2000). By 1995, flash storage chips with a capacity of up to 16 Megabytes were available, and the capacity had risen to 16 Gigabytes by the year 2005, which is essentially what many smartphones contain. However, there were some shortcomings of the flash technology, since it was still comparatively expensive, and fairly slow at reading data.

In "Tape is Dead", Gray also proposes that one could construct an entire file system out of flash storage that would take up less energy and space and also be faster, because the separation between main memory and random access memory in computations is broken down. In 2012 SAP cofounder Hasso Plattner called Jim Gray's program for flash storage "100% true—every single word. He predicts what is happening and will happen. And we just work along." (Plattner 2012: 18) In conjunction with the power of multi-core central processing units (CPUs), this kind of memory technology is what actually enables real time "Big Data" applications, Plattner remarks "we can do things now we couldn't do before" such as "instant calculation of pricing based on the current situation in the market. Wall Street does that every single second." (Plattner 2012: 19).

Fourth Paradigm Science

Thus, by 2007, the components of what is data-intensive machine learning as it had been envisioned in Jim Gray's 1998 Turing Award lecture, were eventually coming together. According to *New York Times* author Thomas

Friedman, 2007 was the year that the era of digitalization of the 1990s entered the next level of acceleration. "In 2007, storage capacity for computing exploded thanks to the emergence that year of a company called Hadoop, making 'big data' possible for all." (Friedman 2016: 20) At the time of writing, even database giants such as IBM and *Oracle* are deploying *Hadoop* to perform analytics on unstructured data. Yet, I do not want to leap too far ahead and focus on the state of discourse around Big Data in 2007. *Wired* magazine's editor Chris Anderson caused a splash when he announced, in a brief article in 2007, the "End of Theory" (Anderson 2008). Anderson was reigning in what he thought to be no less than a new era of knowledge, no longer driven by theory and hypotheses, but data-driven and finally able to dispense with the ambiguities of discussion about whether or not researchers were asking the right questions. "With enough data, the numbers speak for themselves" (Anderson 2008) he claimed, and many have agreed, whether or not their own practice as scientists and humans actually adheres to this epistemology or not.

One practitioner in the field of database technology in particular, and what Gabriele Gramelsberger has called "eScience"[13] generally, had taken Anderson's cue even before his infamous article on the "End of Theory", Jim Gray, who made his own contribution to the business of announcing new scientific eras and "paradigms" in his speech "eScience: A Transformed Scientific Method" on 11 January 2007 at the annual convention of the Computer Science and Telecommunications Board of the US National Research Council.[14]

Gray's talk on "eScience: A Transformed Scientific Method" is instructive in laying bare the rhetorical strategies deployed in order to construct a continuity between what Gray thinks of as a new way of doing data-driven science and the history of science. His talk is also the introduction to a book published by *Microsoft Research* titled "The Fourth Paradigm". Obviously, the announcement of a fourth paradigm implies the existence of three previous paradigms that are somehow being superseded by the new method of eScience. In fact, Gray mostly focuses on the locus of calculation and hypotheses testing rather than discussing characteristics of scientific paradigms in detail. He starts out speaking about scientific paradigms, presented as largely continuous rather than incommensurable and, over and over, ends up much closer to home, discussing digital scientific infrastructures.

Crucially, it is not the sheer amount of data that Gray takes to be the central aspect of any new paradigm, it is the technology deployed in knowledge creation: digital knowledge infrastructures. Most importantly, also for Gray's work as a technologist, he is concerned with the question of where data "meets" software. The engineer of scientific infrastructures has to ad-

dress the question of whether to transport the data to the calculation or carry the calculating power to the data. Thus, the size of Big Data becomes as crucial as the speed of data transfer. In fact, the "size" of data is completely relative to the speed at which it can be transferred.

According to Jim Gray, the history of science has seen four distinct paradigms in research. Since Gray was by no means a historian of science, and probably did not aspire to be one, we should not understand his ideas as part of an historian's argument. In fact, the four paradigms may exist simultaneously or coexist as a plurality of methods within disciplines. For Gray, the first "paradigm" is empirical science that supposedly has been practiced since the time of the ancient Greeks. This is supposed to be the kind of science that describes empirical phenomena and observations. It is unclear how much quantification and hypothesizing is supposed to be involved in this kind of science, since Gray entirely disregards both philosophical origins and non-western scientific traditions. The second "paradigm" is the "theoretical branch" of science that employs generalizations and models in order to derive general knowledge about the world. Saying that this kind of science has been going on for the "last few hundred years", Gray may be thinking of the kind of mathematically driven inquiry in the natural sciences since the time of Newton and Leibniz.

The third "paradigm" according to Gray is then the use of computational simulations in science during the past few decades. Under this paradigm, complex phenomena are simulated, which requires at least some digital computational capacity. This has been feasible only since the Second World War and was not deployed on a larger scale until the expansion of scientific computing in the 1960s and 1970s. But even then, computational capacity was only accessible to a selective few, since the resources of supercomputing centers were limited and exclusively available in a few developed countries. Finally, the fourth "paradigm" according to Jim Gray is that of data exploration and eScience, which he characterizes as follows:

- Data captured by instruments or generated by simulator
- Processed by software
- Information/knowledge stored in computer
- Scientist analyzes database/files using data management and statistics (Hey et al. 2009: xviii)

Yet, one should ask how any of these characteristics constitutes a fundamental difference from the kind of research conducted under the third paradigm. Data has been captured by instruments since the development of the experimental method in science. Also, data generated by simulators is nothing exclusively used in computational sciences at the beginning of the twenty-first century. Data being processed by software also does not

seem to be anything fundamentally new, in fact, one might argue that there is no such thing as digital data that has not been processed by software. Issues with the notion of "raw data" are insightfully discussed in Lisa Gitelman's book *Raw Data is an Oxymoron*, which argues that data cannot be conceptualized independently of its infrastructure, storage hardware and database management software (Gitelman 2013).

The third point in Gray's enumeration is that knowledge and information are stored in a computer. There is no definition of what knowledge and information are in this context. Information is sometimes defined as contextualized and meaningful data, while knowledge is applied and practiced information to a specific end. Thus, one might question generally whether knowledge as such and not just data and information can be stored in a computer or database at all, independently of any knowing subject. Most importantly, however, how is the storage of information on a computer anything new in comparison to the era of computational science since the Second World War, when more and more information was stored on a variety of media? Gray fails to convince here that storage alone is a sufficient and not just a necessary characteristic of Fourth Paradigm Science.

Nevertheless, Gray concludes his talk on eScience: "The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration." (Hey et al. 2009: xix) Looking back at the ways in which Jim Gray has deployed internal memos, professional forums, extensive networking with colleagues, science policy advisory, and public appearances such as his Turing Award lecture, his 2007 talk on eScience also represents another instance of agenda setting by Gray. Gray's reputation as a prescient visionary of database technology development can in part be ascribed to the fact that he has been quite influential in shaping the course of database technology research throughout his career. Fourth Paradigm Science is still used as a marketing term by *Microsoft Research*, as well as Gray's colleague Gordon Bell, to promote their technological capabilities. And yet, Fourth Paradigm Science is absent from current discourse, while the technologies it connotes have been lumped in with Big Data.

"Don't Replace me with a Person"

This paper has traced Jim Gray's involvement in the development of relational database systems at IBM in the 1970s as well as his work on principles and standards of Transaction Processing, laying the ground-

work for the highly distributed performance of high-volume, high-velocity databases today. Platforms of discourse between academia and industry were important during the 1980s in setting industry performance standards and database research agendas. As a co-founder of the San Francisco branch of *Microsoft Research*, Gray had turned toward scientific applications of database technology in the late 1990s. His work on the *Microsoft* TerraServer was followed by further scientific collaborations constructing virtual observatories such as the Worldwide Telescope, the Sloane Digital Sky Survey, and the Ocean Observatories Initiative. Vannevar Bush's idea of the Memex, coined in 1945, informed Gray's vision of a database technology research agenda when he laid out his vision of creating a Personal Memex as well as a World Memex in his Turing Award speech in 1998.

Although Gray did not coin the term Big Data, his work, his activities in standardization and science policy, as well as his advocacy of research agendas show him to be a major trailblazer of what we are today discussing under the heading of "Big Data". Just as "the Internet" was the new promised land for entrepreneurs, which eventually failed to deliver for most but a few moguls such as Marc Zuckerberg or Seregey Brin, data itself became a new frontier for the American entrepreneurial spirit. This initiated a new space race for the data gold; the metaphor of "data mining" should actually be taken very seriously in this case.[15] Jim Gray also construed information technology as a new continent to be explored, calling for a "Lewis and Clark style expedition" into database technology research. Yet, the metaphor of exploration has two aspects to it. On the one hand, it is the technology to be explored, while, on the other hand, developments of database technology are allegedly enabling one to explore the world in an entirely new way via database interfaces and virtual observatories, as in projects such as TerraServer.

In conclusion, we have seen how Jim Gray's talk on the Fourth Paradigm as well as his previous statements on research goals incorporate two curious developments of the recent past, one epistemological and one in public culture. In epistemology, large databases, Big Data, and Fourth Paradigm Science, promise an allegedly new scientific method that will finally lend us the tools for an immediate representation of the empirical world, the plain of truth in reality that is accessed by extensive automated measurements, thus getting rid of the last subjective and soft human factors of knowledge infrastructures. This is a promise that needs to be considered with reservations. Scholars of scientific infrastructures such as Paul Edwards and Geoffrey Bowker remind us of the need for infrastructure inversion.[16] Only through a turning upside down of the scientific infrastructure are we going to be able to fully comprehend the knowledge derived from data-intensive science. This poses a challenge to many of the claims circulated in

the Big Data discourse and stresses the importance of scholarship in communicating and contextualizing the results of any sort of alleged Fourth Paradigm Science.

In public discourse, on the other hand, Big Data is connected to a movement of American popular culture that has, more or less, succeeded in announcing the next endless frontier for exploration and expansion: data. Data is the new space, both literally and figuratively, that entrepreneurs and government agencies scramble to control, sometimes with very real aims of control and surveillance, as in the case of the National Security Agency, at other times with more hazy and commercial aims such as in the most massive advertising operations history has ever witnessed, Google and Facebook.[17] The narratives supporting this "data frontier" discourse have their origins in discussions of research agendas and science policy reaching back to Vannevar Bush in 1945, and have been transported, among others, by well-connected prolific database engineers such as Jim Gray.

"Don't replace me with a person, replace me with a fully configured 4341 for the exclusive use of the R* project," is how Jim Gray concludes his resignation letter at IBM (Gray 1980a). One might equally well ask what the use of humans as people remains to be when our knowledge is automated in a World Memex and the personal memories and assembled narratives that comprise any individual are rendered digitally immortal by a Personal Memex.

Endnotes

1 A good introduction to data mining and its history is Matthew L. Jones' (2017), "Querying the Archive: Data Mining from Apriori to PageRank" in Lorraine Daston (ed.), *Science in the Archives. Pasts, Presents, Futures*.
A seminal reference on databases in science is Geoffrey C. Bowker (2005), *Memory Practices in the Sciences*. Useful overviews are e. g. Thomas J. Bergin and Thomas Haigh (2009), The Commercialization of Database Management Systems, 1969–1983 as well as Thomas Haigh (2009), How Data Got Its Base: Information Storage Software in the 1950s and 1960s, and Avi Silberschatz, Michael Stonebraker and Jeffrey D. Ullman et al. (1990), Database Systems: Achievements and Opportunities.

2 http://jimgray.azurewebsites.net/ section "Recommended Articles".

3 For further context on Vannevar Bush's influence see G. Pascal Zachary (1997), *Endless Frontier: Vannevar Bush, Engineer of the American Century.*

4 Compare Paul N. Edwards (1996), *The Closed World. Computers and the Politics of Discourse in Cold War America*. As well as Naomi Oreskes and John Krige (eds.) (2014), *Science and Technology in the Global Cold War,* and George Dyson (2012), *Turing's Cathedral. The Origins of the Digital Universe.*

5 Compare Committee on Innovations in Computing and Communications: Lessons from History, National Research Council (1999). *Funding a Revolution: Government Support for Computing Research*.

6 See Jane Abbate (1999), *Inventing the Internet*.

7 See Jon Agar (2003), *The Government Machine. A Revolutionary History of the Computer*, and George Dyson (2012), *Turing's Cathedral. The Origins of the Digital Universe.*

8 See Jim Gray et al. (1984), One Thousand Transactions Per Second. Proceedings of IEEE; Jim Gray (1985), A Measure of Transaction Processing Power.

9 See www.hpts.ws/index.html.

10 While IBM uses these five V widely, as does Gartner, there is also a discussion about how useful these are in defining Big Data, see www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters.

11 For more on Gordon Bell's work on self-archiving see Alec Wilkinson (2007), Remember This? A project to record everything we do in life.

12 See questions of Open Science in *Science as an Open Enterprise* (2012).

13 Gabriele Gramelsberger (2010), *Computerexperimente. Wandel der Wissenschaft im Zeitalter des Computers;* ibid. (ed.) (2011), *From Science to Computational Studies. Studies in the History of Computing and its Influence on Today's Sciences.*

14 See Jim Gray, "eScience: A Transformed Scientific Method" in Tony Hey et al. (2009), *The Fourth Paradigm. Data-Intensive Scientific Discovery*.

15 Compare e. g. Jaron Lanier (2013), *Who Owns The Future?*

16 See Geoffrey C. Bowker (2005), *Memory Practices in the Sciences*, Bowker, and ibid. and Susan Leigh Star (1999). *Sorting Things Out. Classification and Its Consequences*.

17 For a cultural history of these ideas see Fred Turner (2006), *From Counterculture to Cyberculture. Stwart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. For a popular critique see e. g. Evgeny Morozov (2013), *To Save Everythink, Click Here. The Folly of Technological Solutionism*.

References

Abbate, Jane 1999. *Inventing the Internet.* Cambridge, MA: MIT Press.

Agar, Jon 2003. *The Government Machine. A Revolutionary History of the Computer.* Cambridge, MA: MIT Press.

Anderson, Chris. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine* (23rd June 2008).

Bailis, Peter, Joseph M. Hellerstein and Michael Stonebraker (eds.) 2015. *Readings in Database Systems*. URL: http://www.redbook.io/all-chapters.html (26.09.2017).

Barclay, Tom 2008. TerraServer and the Russian Adventure. *SIGMOD Record* (37)**:** 59–60.

Barclay, Tom, Jim Gray and Don Slutz 1999. *Microsoft TerraServer: A Spatial Data Warehouse.* Technical Report, Microsoft Research San Francisco.

Bell, Gordon 2007. *Bell's Law for the Birth and Death of Computer Classes. A Theory of the Computer's Evolution.* Microsoft Research San Francisco.

Bergin, Thomas J. and Thomas Haigh 2009. The Commercialization of Database Management Systems, 1969–1983. *Annals of the History of Computing* (IEEE 31)**:** 26–41.

Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences.* Cambridge, MA: MIT Press.

Bowker, Geoffrey C. and Susan Leigh Star 1999. *Sorting Things Out. Classification and Its Consequences.* Cambridge, MA: MIT Press.

Bush, Vannevar. 1945. As We May Think. *The Atlantic* (July 1945).

Carnes, Donna 2008. Ode to a Sailor. *SIGMOD Record* (37)**:** 16–18.

Christensen, Clayton M. 2000. *The Innovator's Dilemma. The Revolutionary Book That Will Change the Way You Do Business.* New York: Harper Business.

Clemson, Gaye I. 2012. *Tandem Computers Unplugged. A People's History.* Campbell, CA: Fast Pencil Inc.

Codd, E.F. 1970. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM* (13)**:** 377–387.

Committee on Innovations in Computing and Communications: Lessons from History, National Research Council 1999. *Funding a Revolution: Government Support for Computing Research.* Washington, D.C.: Press, National Academy.
Daston, Lorraine (ed.) 2017. *Science in the Archives. Pasts, Presents, Futures.* Chicago: The University of Chicago Press.
Dewitt, David J. and Charles Levine 2008. Not just Correct, but Correct and Fast. A Look at Jim Gray's Contributions to Database System Performance. *SIGMOD Record* (37)**:** 45–49.
Diebold, Francis X. 2012. A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline. URL: http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf (25.09.2017).
Dyson, George 2012. *Turing's Cathedral. The Origins of the Digital Universe.* London: Penguin.
Edwards, Paul N. 1996. *The Closed World. Computers and the Politics of Discourse in Cold War America.* Cambridge, MA: MIT Press.
Friedman, Thomas L. 2016. *Thank You for Being Late. An Optimist's Guide to Thriving in the Age of Accelerations.* New York: Farrar, Straus and Giroux.
Garcia-Molina, Hector, Jeffrey D. Ullman and Jennifer Widom 2013. *Database Systems: The Complete Book.* New York: Pearson Education.
Gitelman, Lisa 2013. *"Raw Data" is an Oxymoron.* Cambridge, MA: MIT Press.
Gramelsberger, Gabriele 2010. *Computerexperimente. Wandel der Wissenschaft im Zeitalter des Computers.* Bielefeld: Transcript.
Gramelsberger, Gabriele (ed.) 2011. *From Science to Computational Studies. Studies in the History of Computing and its Influence on Today's Sciences.* Zürich: diaphanes.
Gray, Jim 1980a. *A Critique of IBM's Computer Science Research.* IBM Research, San Jose.
Gray, Jim 1980b. *MIP Envy: a Programming Complex.* IBM Research, San Jose.
Gray, Jim 1985. A Measure of Transaction Processing Power. *Datamation* (1st April 1985).
Gray, Jim 1996. Data Management: Past, Present, and Future. *IEEE Computer* (29)**:** 38–46.
Gray, Jim 1999. *What Next? A Dozen Information-Technology Research Goals.* Technical Report, Microsoft Research San Francisco.
Gray, Jim 2003. *How Do You Know?* Technical Report, Microsoft Research San Francisco.
Gray, Jim 2006. *Tape is Dead, Disk is Tape, Flash is Disk, RAM Locality is King.* Technical Report, Microsoft Research San Francisco.
Gray, Jim and Andreas Reuter 1992. *Transaction Processing: Concepts and Techniques.* San Francsico, California: Morgan Kaufmann Publishers.
Gray, Jim, Bob Good, Dieter Gawlick, Pete Homan and Harald Sammer 1984. One Thousand Transactions Per Second. Proceedings of IEEE Compcon-85. San Francisco.
Haigh, Thomas 2009. How Data Got Its Base: Information Storage Software in the 1950s and 1960s. *Annals of the History of Computing* (IEEE 31)**:** 6–25.
Harrison, Michael A. 2008. Jim Gray at Berkeley. *SIGMOD Record* (37)**:** 26–27.
Helland, Pat 2008. Knowledge and Wisdom. *SIGMOD Record* (37)**:** 28–29.
Hellerstein, Joseph M. and David L. Tennenhouse 2011. Searching for Jim Gray: A Technical Overview. *Communications of the ACM* (54)**:** 77–89.
Hey, Tony, Stewart Tansley and Kristin Tolle (eds.) 2009. *The Fourth Paradigm. Data-Intensive Scientific Discovery,* Redmont, Washington: Microsoft Research.
Lanier, Jaron 2013. *Who Owns The Future?* London: Allen Lane.
Lindsay, Bruce G. 2008. Jim Gray at IBM. The Transaction Processing Revolution. *SIGMOD Record* (37)**:** 38–40.
Lohr, Steve. 2012. How Data Became So Big. *The New York Times* (11th August 2012).
Lohr, Steve. 2013. The Origins of "Big Data": An Etymological Detective Story. *The New York Times* (1st February 2013).
Maccormick, John 2012. *9 Algorithms That Changed the Future: The Ingenious Ideas That Drive Today's Computers.* Princeton: Princeton University Press.
Markoff, John. 2003. In Computing, Weighing Sheer Power Against Vast Pools of Data. *The New York Times* (2nd June 2003).
Morozov, Evgeny 2013. *To Save Everythink, Click Here. The Folly of Technological Solutionism.* New York: Public Affairs.
Nauman, John 2008. Jim Gray's Tandem Contributions. *SIGMOD Record* (37)**:** 41–44.

Oreskes, Naomi and John Krige (eds.) 2014. *Science and Technology in the Global Cold War,* Cambridge, MA: The MIT Press.
Plattner, Hasso 2012. Oral History of Hasso Plattner (Computer History Museum, Palo Alto). URL: http://www.redbook.io/all-chapters.html (26.09.2017).
Reuter, Andreas 2008. Is There Life Outside Transactions? Writing the Transaction Processing Book. *SIGMOD Record* (37)**:** 54–58.
Richtel, Matt. 1998. Huge Microsoft Photo File Is Part of a Bigger Picture. *The New York Times* (25th June 1998).
Royal Society Science Policy Centre 2012 *Science as an Open Enterprise.* London: The Royal Society.
Silberschatz, Avi, Michael Stonebraker and Jeffrey D. Ullman 1990. Database Systems: Achievements and Opportunities. *ACM Sigmod Record* (19)**:** 6–22.
Stonebraker, Michael 2007. Oral History of Michael Stonebraker (Computer History Museum, Palo Alto). URL: http://www.computerhistory.org/collections/catalog/102635858 (26.09.2017).
Turner, Fred 2006. *From Counterculture to Cyberculture. Stwart Brand, the Whole Earth Network, and the Rise of Digital Utopianism.* Chicago: The University of Chicago Press.
Vaskevitch, David 2008. Jim Gray: His Contribution to Industry. *SIGMOD Record* (37)**:** 35–36.
Wilkinson, Alec. 2007. Remember This? A Project to Record Everything We Do in Life. *The New Yorker.* (28th May 2007).
Zachary, G. Pascal 1997. *Endless Frontier: Vannevar Bush, Engineer of the American Century.* New York: The Free Press.

Nils C. Hanwahr
Rachel Carson Center for Environment and Society
Ludwig-Maximilians-Universität München
Leopoldstraße 11a
80802 München
Germany
nils.hanwahr09@alumni.imperial.ac.uk