PREPRINT

To appear in Lorraine Daston, ed. *Sciences of the Archive* (Chicago)

QUERYING THE ARCHIVE:
DATA MINING FROM APRIORI TO PAGERANK

Matthew L. Jones

mljones@columbia.edu

In 1998, amid the blossoming of large-scale corporate, government, and academic "data warehouses," Usama Fayyad was worried.

> If I were to draw on a historical analogy of where we stand today with regards to digital information manipulation, navigation, and exploitation, I find myself thinking of Ancient Egypt. We can build large impressive structures. We have demonstrated abilities at the grandest of scales in being able to capture data and construct huge data warehouses. However, our ability to navigate the digital stores and truly make use of their contents, or to understand how they can be exploited effectively is still fairly primitive. A large data store today, in practice, is not very far from being a grand, write-only, data tomb.[1]

Fayyad, then at Microsoft, previously at the Jet Propulsion Laboratory, explained that he hoped "the recent flurry in activity in data mining and KDD ["Knowledge discovery in databases"] will advance us a little towards bringing some life into our data pyramids."[2] "Big data" there was plenty. Ready access to that data and its significance, not so much. Excavating this entombed data, Fayyad argued, required richer forms of querying, especially transforming techniques drawn from statistics and the machine learning branch of artificial intelligence. In the same year, two Stanford computer science graduate students were hard at work pushing the latest in data mining techniques to apply to a far more anarchic archive, not a data pyramid of well-orgranized corporate data, but a jumble of non-curated, interlinked pages: the World Wide Web. Their end-product: the Google search engine.

This chapter stresses the centrality of the database community of academic computer science and industry, the community entrusted with figuring out how to secure digital archives, in the creation of data sciences of the 2000s. Database practitioners could never forget the scale of data, understood not as something intangible but as something

physical existing on slow hard drives, something incapable of being resident in memory, something requiring time to move from place to place and from drives to processors. Large amounts of data broke things easily, as Fayyad noted in 1998.

> A typical statistical package assumes small data sets and low dimensionality. For example, suppose you want to do some simple database segmentation by running […] a basic simple method for clustering data. Let's say you have managed to lay your hands on an implementation in some statistical library. The first operation the routine will execute is "load data." In most settings, this operation will also be the last as the process hits its memory limits and comes crashing down.[3]

Adopting statistical and artificial tools to very large databases required dramatic modifications to the tools, epistemic shifts, and transformations in the epistemic values and protocols surrounding their use.

This chapter focuses on two centers of tremendous activity in mining databases, just down the road from each other: the data mining group at Stanford and IBM's Almaden research center in San José. While the technical means for contending with the scale of the archive remained underdetermined, the creators of data mining offered powerful technological determinist narratives holding that contending with great volumes of data requires the development of new algorithms, the loosening of traditional account of statistical rigor, the creation of new epistemic virtues, and the creation of new experts.[4] The challenging materiality of corporate and government digital archives came to justify fundamental transformations in practices and values: new tools for automation that demanded heightened skills in data cleaning and 'munging' and alternative forms of algorithmic judgment.[5] The sheer scale of data was held to demand—and to justify—new forms of scientific knowledge, at times in conflict with long-held views of statistical rigor.

Materiality of the archive

The "data warehousing" guru Ralph Kimball explained in 1995 that makers of large databases in the 1980s and 1990s had prioritized perfecting storage over improving access to data.[6] Building large databases that could accurately, durably, and securely record large numbers of transactions and other forms of cleaned data had dominated database research and practical implementation. In the 1970s and 1980s, database

researchers had focused squarely upon putting stuff into databases.[7] Central to the developments of database practice was an insistence that changes to the database be "atomic": any set of connected changes to a database must be entirely completed or not completed at all. If you transfer $100 from one account to the other, the database system must either alter both accounts or alter neither. Researchers had to solve the problem for large-scale systems, where hardware, software, and power were presumed to fail with great regularity—as they do.[8]

However necessary their efforts were, Kimball explained, "we came perilously close to forgetting why we bought relational databases in the first place."[9] Seeking meaningful information from massive databases had been insufficiently prioritized. Kimball proclaimed the situation to be dire: "To be blunt, today's systems are very good at transaction processing and pretty horrible at querying." The time to create new forms of access was now.

> Fortunately, the chief executives in most companies have long memories. They remembered the promise that we would be able to "slice and dice" all of our data. These executives have noticed that we have almost succeeded in storing all the corporate data in relational databases. They also haven't forgotten that they have spent several billion dollars. From their point of view, it is now time to get all that data out.[10]

By the mid-1990s database practitioners in industry and academia alike were focusing attention upon new techniques for querying increasingly large databases.

Data mining was initially one of the many efforts to improve querying.[11] For database management researchers, data mining—or as they originally called it, "database mining"—was just such a better way of getting at the contents of databases. A primary driver of IBM's data mining effort, Rakesh Agrawal, explained in 2003: "I'm a database person, so my view of data mining has been that it is essentially a richer form of querying. We want to be able to ask richer questions than we could conveniently ask earlier."[12] In 2004, the Microsoft database researcher Jim Gray explained, "We are slowly climbing the value chain from data to information to knowledge to wisdom. Data mining is our first step into the knowledge domain."[13]

KDD

Data mining, or, as it was more formally dubbed in the 1990s, Knowledge Discovery in Databases (KDD), is the activity of creating non-trivial knowledge suitable for action from databases of vast size and dimensionality.[14] Data miners never speak of an *information* overload. Their mantra is, "We're data rich, but information poor." For them, information, taken to be largely synonymous with knowledge, comprises interesting, non-trivial patterns in data. Their task was to create such "interesting," "actionable" patterns from vast quantities of data in practically computable ways.

Underlying the practice of data mining is a *critique* of artificial reason—a recognition of the limits of the abilities of human beings and machines when faced with vast amounts of data, followed by the creation of tools more suited to limited human abilities, limited computing memories and speeds, and relatively short-term, real-world goals for investigation. Data mining concerns databases of very large size—millions or billions of records, usually with elements of high dimensionality (meaning that every record typically comprises a large number of elements). For each record in a retail database, a data mining operation might seek unexpected relationships among the item purchased, the store's zip code, the purchaser's zip code, variety of credit card, time of day, date of birth, other items purchased at the same time, even every item viewed, or the history of every previous item purchased or returned. Performing reasonably fast analyses of high-dimensional, messy real-world data is central to the identity and purpose of data mining, in contrast to its predecessor fields such as statistics and machine learning. The technical and social solutions of data mining involve converting theoretical algorithms into everyday practices and high dimensional data into actionable knowledge.

According to KDD advocates, traditional scientific approaches to data—and the traditional competencies of scientists—simply could not keep up with the volume of data and multidimensionality possible thanks to computers. Something else is needed, something less pure—because it deals with vast impurities of dynamic data, nearly always from a particular business, governmental, or scientific research goal. A now canonical programmatic piece explains, "[…] scientists can reformulate and rerun their experiments should they find that the initial design was inadequate. Database managers rarely have the luxury of redesigning their data fields and recollecting the data."[15] Establishing the legitimacy of KDD meant demonstrating that lack of luxury and

showing its techniques to be productive and meaningful in dealing with such challenging data.

The contrast with statistics and machine learning, however polemically overdrawn, is philosophical, methodological, and institutional; data mining involves a different scientific "way of life" with far different epistemological virtues from its antecedent disciplines. The contrast is also historical. As a field, data mining had statistical concerns and used numerous statistical techniques but did not emerge from a statistical culture; it emerged less from a culture self-consciously attempting to be a branch of mathematics than one uniting the practices of machine learning and of the management and use of large-scale databases for corporations and scientific researchers.

The sweet smell of positivism wafted over data mining from the start. Through data unwedded to theory, data mining promises to overcome wonted ways of dividing the world:

> With traditional statistical modeling, an analyst would pose a question such as: 'Are higher-income people prone to be more loyal to a warehouse club than those with lower income levels?' and the hypothesis would either be supported or unsupported. Data mining, on the other hand, potentially would provide more insight by pointing out other factors contributing to store loyalty that the analyst would not otherwise have been able to consider testing.[16]

Science studies practitioners are not the only ones likely to question such claims. One of the great figures of AI, John McCarthy, wrote in a presentation written after spending time in IBM's database laboratory in San José, that positivistic data mining was "BAD PHILOSOPHY and INADEQUATE COMPUTER SCIENCE."[17]

Computational constraints and database culture

The notes for a lecture course on Data Mining at Stanford in 2000 detail the various communities involved in data-mining: statistics, artificial intelligence, "Visualization researchers," and "Databases." Like the IBM researchers, the prominent Stanford database researcher Jeffrey Ullman explained,

> We'll be taking this [database] approach, of course, concentrating on the challenges that appear when the data is large and the computations complex. In a

> sense, data mining can be thought of as algorithms for executing very complex queries on non-main-memory data.[18]

These concerns, in turn, encouraged an environment in which creative database practitioners adapted analytical processes from elsewhere in computer science and computational statistics to allow them to work with much bigger realms of data.

While futurists and politicians opined endlessly about the utopian possibilities of virtual spaces, cyber-realms and dystopian visions of alternate reality matrices, database practitioners had the essential job of working within the constraints of materialized systems to enable the preservation, storage, and quick access to data. Disk access times, efficient use of memory, component failure, and distributed and parallel computing are central concerns in nearly every database paper and the maintenance of working databases.[19]

Sophisticated statistical and machine learning algorithms are typically devised for sets of data that can easily fit in memory, or that require a relatively small use of slower disk access. Adapting such algorithms to huge quantities of data that cannot be held in memory is non-trivial: different epistemic values and metrics for gauging difficulty and success are brought into play. Many of the developments of greatest significance to data mining were efforts to choose among the tradeoffs necessary to make statistical and machine learning algorithms scale.

One constraint is of the utmost importance for database miners. The database cannot be entirely stored in a computer's memory: "the number of times we need to read each datum is often the best measure of the running time of the algorithm."[20] A generation of practitioners devoted tremendous effort to reworking algorithms to reduce disk access time and the use of memory. The algorithms of big data, deeply indebted to an earlier generation of machine learning tools, matter just because they were made to scale. And whatever the increase of processor speed, successes in parallelizing processing, and the availability of memory, the size and dimensionality of data keep these values central for large data mining operations.

Concerns with computational efficiency and the constraints of disk access times were no mere surface phenomenon: they were deeply inscribed into the algorithms themselves and their implementations in real, faulty machines.

Association mining: Apriori and Quest

Suppose you run a large grocery store chain. For merchandising purposes, you'd like to know which items customers tend to purchase together, so that you could optimally place those items in your store to maximize revenue. You have an enormous database that collects all the items purchased together in every transaction. The "market-basket problem" is to find items that tend to be purchased together. More formally, the problem asks us to discern from our database a series of "association rules" where the presence of some set of items suggests a highly likelihood of another, so that $\{X_1, X_2, X_3, X_4, \ldots, X_n\} \rightarrow Y$ with a high degree of probability. Most famously, researchers working on a data set from the Osco drug store chain discovered that {diapers}->beer, that is, there's a fairly high probability that someone buying diapers is likely to buy beer. Note that the converse, {beer}->diapers, has a much lower probability.[21]

Researchers in the Quest group at IBM's Almaden research lab in San José developed an algorithm called Apriori to discover just such association rules to deal with the market-basket problem. The researchers in Quest cast their investigations in terms of how "database technology should be enhanced," notably, "classification, queries on sequences, successive query refinements, query language extensions and optimization as some of the key technical issues in this area." Already in 1991 the group had experienced some success with "an algorithm for discovering rules (patterns) in a historical database of customer transactions."[22] The latter would eventually become the Aprioi association algorithm, published in the flagship journal of database management in 1993.[23]

Soon dubbed one of the "top ten data mining algorithms," Apriori spurred a broad range of researchers in industry and academia to improve it, apply it to new domains, and to discern new uses for it. Of little interest within the more rarified domains of artificial intelligence, machine learning, and statistics, such association algorithms accounted for

a sizeable percentage of early successes of the data mining community, quickly moving from the drug store to genomics and beyond.

Association algorithms produced rules in great number, indeed in huge numbers. This turned out to be a boon. In an interview, Agrawal explained,

> When we started doing data mining, we were concerned that we were generating too many rules, but the companies we worked with said, "this is great, this is what exactly what we want!" The prevailing mode of decision making was that somebody would make a hypothesis, test if the hypothesis was correct, and repeat the process. Once they had data mining tools, the decision-making process changed. Now they could use a data mining algorithm to generate all rules, and then debate which of them were valuable.[24]

Traditional statistical notions of significance had little place here. Instead, the focus was on the value of "interestingness."[25]

Computational efficiency as productive constraint

The IBM Quest team worked from the "perspective of database mining as the confluence of machine learning techniques and the performance emphasis of database technology." The confluence demanded that both fields mutually influence one another: "Unfortunately the database systems of today offer little functionality to support such 'mining' applications. At the same time, statistical and machine learning techniques usually perform poorly when applied to very large data sets."[26] Most data mining algorithms were drawn from statistics or the branch of artificial intelligence called machine learning, typically housed within computer science departments. Even as they shared an interest in algorithms and their use on sets of data, practitioners of the three diverse fields had—and have—radically different goals and values when implementing and transforming these basic algorithms. No database practitioner would discuss an algorithm without a focus on scale. For a database practitioner, scale means reworking an algorithm to deal with large volumes of data on real computers taking non-trivial time to perform various operations.

Computational statisticians, machine learning specialists, and database engineers have produced a large number of different algorithms to perform classification, decision trees, neural nets, clustering algorithms, various forms of regression, "support vector

machines," and so forth. Algorithms requiring both data and human classification of some of that data, called "training data," are examples of "supervised learning." Even more radical, in many ways, were algorithms that simply worked on the data without requiring a "domain-specific" expert to classify some cases. Such approaches are known as "unsupervised learning," and from them a million positivistic promises to funders have sprung. No human "trains" the computer to make certain associations or classifications by giving it human produced classifications of the data.

Database researchers could not simply import algorithms as described in statistical papers or machine learning textbooks; scaling algorithms usually required radical reworking, both of code and of values at stake. An early paper on the topic explains that approaches in machine learning and statistics "do not adequately consider the case that the dataset can be too large to fit in main memory. In particular, they do not recognize that the problem must be viewed in terms of how to work with a [*sic*] limited resources (e.g. memory that is typically, much smaller than the size of the dataset) to do the clustering as accurately as possible while keeping I/O costs low."[27] Recognizing such constraints changes the problem to be solved:

> We adopt the problem definition [of clustering] used in Statistics, but with an additional database-oriented constraint: *The amount of memory available is limited (typically, much smaller than the data set size) and we want to minimize the time required for I/O.* A related point is that it is desirable to be able to take into account the amount of *time* that a user is willing to wait for the results of the clustering algorithm.[28]

Providing a solution to this problem demands a radically different kind of algorithm with a choice of statistical trade-offs that made analyzing larger sets of data tractable. Distance based approaches to clustering, for example, "assume that all data points are given in advance and can be scanned frequently. They totally or partially ignore the fact that not all data points in the dataset are equally important with respect to the clustering purpose, and that data points which are close and dense should be considered collectively instead of individually."[29] From the database perspective, not all the data points could be treated the same, and so they *should* not be. The algorithm creates something called a "CF tree," a condensation of the total data that balances the value of accounting for every data vector with the need to fit within memory and reduce scans of the disk. The authors explain the balance of the values. The algorithm is

1. fast because (a) no I/O operations are needed and (b) the problem of clustering the original data is reduced to a smaller problem of clustering the subclusters in the leaf entries;
2. accurate because (a) a lot of outliers are eliminated and (b) the remaining data is reflected with the finest granularity *that can be achieved given the available memory*.[30]

Further developing this approach to extending clustering algorithms to large datasets, the (then) Microsoft researcher Usama Fayyad and several collaborators offered a yet more demanding set of "*Data Mining Desiderata*":

1. **One scan:** The algorithm requires at most one database scan with early termination highly desirable.
2. **Anytime algorithm:** The algorithm is always able to provide a "best" answer at anytime during its computation (i.e. it exhibits "online, anytime" behavior).
3. **Interruptible and Incremental:** The algorithm is suspendable, stoppable and *resumable.* Incremental progress can be saved for continued computation later, possibly on new data.
4. **Limited RAM Requirement:** The algorithm works within the confines of a limited memory (RAM) buffer, allocated by the user, insuring good behavior when run as a server process.
5. **Forward-only cursor:** The algorithm has the ability to operate with a forward-only cursor over a view of the database.[31]

To the insistence on attention to input/output and memory constraints, these practitioners underscored the value of considering different provisional runs of an algorithm. Rather than waiting for an algorithm to complete and issue a final result, a provisional result can always be rendered. The results of nearly two decades of work now figures centrally in courses on mining massive data sets, including the descendent of the Stanford CS course described above.

In the late 1990s, IBM's Almaden Lab in San José hosted an ongoing seminar series that brought in academic researchers, industrial researchers and IBM's own employees, and more generally served as a center of sociability for the local data mining community.[32] Many of the transformative papers presented there would become standard works in scaling statistical and machine learning algorithms for use in real machines with large data sets.

One Wednesday morning in November 1997, a Stanford Computer Science graduate student came south to Almaden to speak on the topic of "Mining the Web":

> A new project at Stanford is the WebBase project. The goals are to collect a large amount of data from the Web and to make it available for research. While the project is relatively new (several months), it has already produced some interesting results.

The speaker, Sergey Brin, was the organizing force of a data mining group, MIDAS—***MI****ning* ***D****ata* ***A****t* ***S****tanford*, which had the support of several faculty members, each a pioneer in database management. At its regular meeting, the MIDAS group discussed the state of field, from algorithms to ethics: "Topics range from admistrative [sic] issues and grant proposals to conference-style presentations by students and visitors."[33] For his talk at IBM, Brin promised to range widely over their work on the web.

> I will talk about some of the things we have discovered with this data and some algorithms that have been developed including link analysis, quality filtering, searching and phrase detection.

The project would soon bear much algorithmic fruit.[34] And, before long, many billions of dollars. The webpage for the Stanford Data Mining group notes, "The most impressive and useful demo is the super search engine, called Google, built by Larry Page and Sergey Brin."[35]

The genealogy of PageRank

Numerous computer science communities were underprepared in the 1990s to contend with the vast, expanding, and decidedly non-curated World Wide Web. A generation of work dedicated to producing reliable, consistent corporate data bases and data warehouses had not created tools nearly adequate for the Web. The field of Information Retrieval, dedicated precisely to the creation of tools for search and extraction, had created numerous techniques for indexing and querying in what had previously been considered large databases, such as library catalogs or indices of journals.[36] Those search and indexing tools were designed for highly standardized, curated, centralized collections of text or other data such as a collection of journals with their metadata; they struggled with both the non-standard and unstructured quality of Web pages and their number.[37] Like many machine-learning algorithms, information retrieval algorithms did not scale easily to the number of pages in the Web. Early search engines tended to index the words in documents, which proved disappointing and was easily manipulated

simply by adding long lists of popular words to a webpage. Web-search positivism of the simplest kind had failed. By the mid- 1990s, search seemed to many an unpromising approach to the Web and the major industry players focused increasingly on curated portals, exemplified by the approach of Yahoo. Search came to dominate after 2000, with the gradual, then exponential, rise of Google. (This short chapter must skip over Jon Kleinberg's simultaneous invention of a similar approach to ranking search results using indices of authorities.)[38]

In 1998, Brin, in the Database group at Stanford, and his fellow graduate student Larry Page, in the Human Computer Interactions group, attempted to devise newer association algorithms for a generalized market basket problem. Rather than looking at items in consumers' baskets, they looked for associations within documents on the Web. Their approach, called "Dynamic Data Mining," did not "exhaustively explore the space of all possible association rules"—as the web was far too big to do so:

> . . . when standard market basket data analysis is applied to data sets other than market baskets, producing useful output in a reasonable amount of time is very difficult. For example, consider a data size with tens of millions of items and an average of 200 items per basket. . . . A traditional algorithm could not compute the large itemsets in the lifetime of the universe.[39]

Just as machine learning algorithms had to change to deal with the scale of early database mining, association mining algorithms had to change to deal with the scale of the early Web. Combining his long standing interests in mathematics with the values and concerns of a database group, Brin undertook, in his earliest publication, to modify classical statistical algorithms for "nearest neighbor search" to contend with "large metric spaces"—that is, to be able to deal with very large databases with high dimensional vectors.[40] In their adaptation of such an approach to commercial databases, Brin and Page exemplified the drive of database practitioners to minimize disk and memory usage.

> To make an entire pass over the data for every set of candidates is prohibitively expensive. . . . Instead of making mining a finite, closed-ended process which produces a well defined, complete set of items, we make it a continuous process, generating continually improving sets of itemsets . . . . the process takes advantage of intermediate counts to form estimates of itemsets' occurrence so there is no need to wait until the end of a pass to estimate an itemset's weight.[41]

They likewise recognized the value of designing algorithms capable of yielding provisional answers as they continued to scan the data.

Brin and Page, along with their collaborators, argued that the scale of the Web, which made it so challenging, simultaneously made it deeply promising:

> we take advantage of one central idea: the Web provides its own metadata through its link structure, anchor text [the visible name of the link], and partially redundant content. This is because a substantial portion *of* the Web is *about* the Web. To take full advantage of this data would require human intelligence or more. However, simple techniques that focus on a small subset of the potentially useful data can succeed due to the scale of the web.[42]

Based within a database community deeply interested in transforming existing statistical and machine learning techniques, Brin and his collaborators were prepared not just to deal with scale, but to make it into a central resource for discovery. Fundamentally, they realized that the scale of the Web included vast human effort to classify and categorize the Web in billions of piecemeal ways. Rather than creating any form of artificial intelligence capable of classifying the Web itself, they created a mechanism for leveraging human judgment at great scale.

Brin and Page's greatest breakthrough in mining the Web came in adapting an everyday academic practice into algorithmic form most fruitful at vast scales. Following an insight of Page's, they adapted the idea of counting high-quality citations to gauge the authority or value of academic work. Web pages could be "ranked" as more or less authoritative by counting citations, that is, links to pages. More authoritative pages are those that have been linked to by other authoritative pages. The total numbers of links to a page counted far less than the authority of the pages linking to that page. They called the result PageRank, and they soon made it central to a new search engine, Google.

Google search emerged from within a culture fusing database values and practice with machine learning. Much scholarship on Google search has focused tightly on the clever development and implementation of PageRank as a problem in linear algebra and computation of Markov chains.[43] Brin and Page recognized from the start the need for structuring databases capable of implementing the beautiful mathematics on fallible and limited machines.

> Google's data structures are optimized so that a large document collection can be crawled, indexed, and searched with little cost. Although, CPUs and bulk input output rates have improved dramatically over the years, a disk seek still requires about 10 ms to complete. Google is designed to avoid disk seeks whenever possible, and this has had a considerable influence on the design of the data structures.[44]

A process for leveraging human judgment at mass scale, PageRank had to be materialized in a creatively designed set of databases. PageRank and its instantiation within commodity hardware in time led to the development of new architectures for distributed databases and distributed analytic processing, called BigTable and MapReduce respectively. These developments of these approaches figure centrally in the distributed storage and processing of big data today.

In a 2009 piece, "The Unreasonable Effectiveness of Data," three Google affiliates premised the power of huge data sets on the dispersed human curation of the data available "in the wild."[45] Like the IBM database researchers happy to find lots of interesting results in the data, the Google researchers praise the discovery of the multitude of rules found, say, in understanding translation. Memorization, not first principles, have allowed the advances in machine translation, they argue:

> Instead of assuming that general patterns are more effective than memorizing specific phrases, today's translation models introduce general rules only when they improve translation over just memorizing particular phrases (for instance, in rules for dates and numbers). Similar observations have been made in every other application of machine learning to Web data . . .

The authors craft from this a conception of linguistics itself, as a new form of big data natural history:

> For those who were hoping that a small number of general rules could explain language, it is worth noting that language is inherently complex, with hundreds of thousands of vocabulary words and a vast variety of grammatical constructions. Every day, new words are coined and old usages are modified. This suggests that we can't reduce what we want to say to the free combination of a few abstract primitives.[46]

In their celebration of drawing upon vast arrays of data from the Web, the Google researchers explicitly condemned several generations of work trying to build up computational linguistics and other fields through the arduous process within Artificial

Intelligence of creating generalized "ontologies." The granularity of knowledge—of language and many other domains—becomes evident far better through leveraging the human curation of information. Data mining has thus yielded a new positivism, grounded less in the data themselves, than in a billion classifications taken together. Truly this was a new science grounded in empirical specificity of the archive, with the help of its plentiful, often contradictory, cacophonous finding aids.

Volume will never win, for Volume is our Friend

Brin and Page weren't the only people recognizing the dangers and potential of volume in the late 1990s. A 1996 interview in the highly classified house magazine of the US National Security Agency with a deputy director turned to the question of the volume of world communication to be spied upon:

> Let me add to all of that the third biggest challenge facing us, and that is volume. And I could just end the sentence there and everything is said. [Paragraph Redacted]
> That gives you some idea of the daunting challenge volume presents, forcing us to look for new technologies.

However great a challenge, volume, the deputy director underscored, would not destroy signals intelligence: "Volume will never win, the reason being that volume is not the only way the world is constructed."[47] By 2006, a top-secret email "Volume is Our Friend," suggests a newfound confidence in the Agency's ability to contend with data overload: indeed, the enabling quality central to the celebration of big data elsewhere. The bigger the volume, the better.

---

1 Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery," *Bulletin of the Technical Committee on Data Engineering* 21, no. 1 (1998): 48.

2 Ibid.

3 Usama Fayyad, "Taming the Giants and the Monsters: Mining Large Databases for Nuggets of Knowledge," *Database Programming and Design* 11, no. 3 (1998).

4 For accounts of earlier history of data, see David Sepkoski, "Towards 'A Natural History of Data': Evolving Practices and Epistemologies of Data in Paleontology, 1800–2000," *Journal of the History of Biology* 46, no. 3 (August 2013): 401–44, doi:10.1007/s10739-012-9336-6; Bruno J. Strasser, "Data-Driven Sciences: From Wonder Cabinets to Electronic Databases," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43, no. 1 (March 2012): 85–87, doi:10.1016/j.shpsc.2011.10.009; Paul Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (Cambridge: MIT Press, 2010). See Sepkoski and Strasser's essays in this volume.

5 For data preparation, see Jankovi*ć* and Strasser's essays in this volume.

[6] Ralph Kimball, "The Database Market Splits," accessed June 18, 2013, http://www.kimballgroup.com/1995/09/01/the-database-market-splits/.

[7] The main academic histories of database systems are Thomas J. Bergin and Thomas Haigh, "The Commercialization of Database Management Systems, 1969–1983," *Annals of the History of Computing, IEEE* 31, no. 4 (2009): 26–41; Thomas Haigh, "How Data Got Its Base: Information Storage Software in the 1950s and 1960s," *Annals of the History of Computing, IEEE* 31, no. 4 (2009): 6–25; Avi Silberschatz, Michael Stonebraker, and Jeffrey D. Ullman, "Database Systems: Achievements and Opportunities," *ACM Sigmod Record* 19, no. 4 (1990): 6–22; more generally, see Richard L. Nolan, "Information Technology Management Since 1960," in *Nation Transformed by Information: How Information Has Shaped the United States from Colonial Times to the Present*, ed. Alfred Dupont Chandler and James W. Cortada (New York: Oxford University Press, 2000), 217–56; and for the very recent history, see Jeff Hammerbacher, "Information Platforms and the Rise of the Data Scientist," in *Beautiful Data: The Stories Behind Elegant Data Solutions* (O'Reilly Media, 2009), 73–84. Fine studies of the profound effects of database design on scientific cultures include Hallam Stevens, *Life out of Sequence: A Data-Driven History of Bioinformatics* (Chicago: University of Chicago Press, 2013), 139–141, 161, 168–9; Sabina Leonelli and Rachel A. Ankeny, "Re-Thinking Organisms: The Impact of Databases on Model Organism Biology," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43, no. 1 (March 2012): 29–36, doi:10.1016/j.shpsc.2011.10.003. See also Geoffrey C. Bowker, *Memory Practices in the Sciences* (Cambridge, Mass.: MIT Press, 2005).

[8] Silberschatz, Stonebraker, and Ullman, "Database Systems," 11.

[9] Kimball, "The Database Market Splits."

[10] Ibid.

[11] For a good sense of the array of research projects in addition to data mining, see Jim Gray, "The Next Database Revolution," in *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, 2004, 1–4, http://dl.acm.org/citation.cfm?id=1007570.

[12] Marianne Winslett and Rakesh Agrawal, "Rakesh Agrawal Speaks Out on Where the Data Mining Field Is Going, Where It Came From, How to Choose Problems and Open Up New Fields, Our Responsibilities to Society as Technologists, What Industry Owes Academia, and More," 2003, http://www.sigmod.org/publications/interview/pdf/D15.rakesh-final-final.pdf.

[13] Gray, "The Next Database Revolution."

[14] While "Data mining" and "KDD" are usually used interchangeably, the term "KDD" was introduced to clarify that the most ideational algorithmic processes of processing data are only a small subset of wide ranges of processes needed to produce "knowledge" from large databases; see Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining* (Menlo Park, CA, USA: AAAI/MIT Press, 1996), 1–34.

[15] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus, "Knowledge Discovery in Databases: An Overview," in *Knowledge Discovery in Databases*, ed. Gregory Piatetsky-Shapiro (Cambridge, MA: AAAI/MIT Press, 1991), 8.

[16] Shawn Thelen, Sandra Mottner, and Barry Berman, "Data Mining: On the Trail to Marketing Gold," *Business Horizons* 47 (2004): 26, doi:10.1016/j.bushor.2004.09.005.

[17] John McCarthy Papers (SC0524) Department of Special Collections and University Archives, Stanford University Libraries, Stanford, Calif.; Accession 2012-055, box 11, folder 11. He published a version of this paper: John McCarthy, "Phenomenal Data Mining: From Data to Phenomena," *ACM SIGKDD Explorations Newsletter* 1, no. 2 (2000): 24–29.

[18] Jeffery D. Ullman, "CS 345 Data Mining Lecture Notes" (unpublished, 2000), 1, http://infolab.stanford.edu/~ullman/mining/allnotes.pdf.

[19] For the "materiality" of data and its significance, see especially Edwards, *A Vast Machine*.

[20] Ullman, "CS 345 Data Mining Lecture Notes," 3.

[21] Dan Power, "Origins of Beer & Diapers," *DSS News* 3, no. 23 (2002), http://www.dssresources.com/newsletters/66.php; Ronny Kohavi, "Origin of 'Diapers and Beer,'" July 6, 2000, http://www.kdnuggets.com/news/2000/n14/8i.html.

[22] Laura M. Haas and Patricia G. Selinger, "Database Research at the IBM Almaden Research Center," *SIGMOD Rec.* 20, no. 3 (September 1991): 97, doi:10.1145/126482.126493. Research on Quest enabled by Wayback machine caches of IBM Quest website from the mid to late 1990s.

[23] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, 207–16, doi:10.1145/170035.170072.

[24] Winslett and Agrawal, "Rakesh Agrawal Speaks Out."

[25] Debate around "interestingness" continues around association mining. For a mid-1990s perspective, see the important Avi Silberschatz and Alexander Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems," *IEEE Transactions on Knowledge and Data Engineering* 8, no. 6 (1996): 970–74, doi:http://doi.ieeecomputersociety.org/10.1109/69.553165.

[26] R. Agrawal, T. Imielinski, and A. Swami, "Database Mining: A Performance Perspective," *IEEE Transactions on Knowledge and Data Engineering* 5 (1993): 914.

[27] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4–6, 1996*, ed. H. V. Jagadish and Inderpal Singh Mumick (ACM Press, 1996), 104.

[28] Ibid., 103.

[29] Ibid., 104.

[30] Ibid., 107.

[31] Paul S. Bradley, Usama M. Fayyad, and Cory Reina, *Scaling EM (Expectation-Maximization) Clustering to Large Databases*, Technical Report MSR-TR-98-35 (Microsoft Research, 1998), 2.

[32] The list of seminars as of the end of 1998 are available at http://web.archive.org/web/19990116232602/http://www.almaden.ibm.com/cs/quest/seminars.html and http://web.archive.org/web/19980210042739/http://www.almaden.ibm.com/cs/quest/seminars-hist.html.

[33] The webpage for MIDAS is preserved at http://infolab.stanford.edu/midas/; a list serve of the data mining group can be found on Yahoo e-groups. See Jeffrey D. Ullman, "The MIDAS Data-Mining Project at Stanford," in *Database Engineering and Applications, 1999. IDEAS'99. International Symposium Proceedings*, 1999, 460–64, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=787298.

[34] A printed version of this material appeared as Sergey Brin, Rajeev Motwani, and Terry Winograd, "What Can You Do with a Web in Your Pocket," *Data Engineering Bulletin* 21 (1998): 37–47.

[35] http://infolab.stanford.edu/midas/

[36] For the early history of some of these tools, see Rosenberg's essay this volume.

[37] Thomas Haigh, "The Web's Missing Links: Search Engines and Portals," in *The Internet and American Business*, ed. William Aspray and Paul Ceruzzi (Cambridge, Mass.: MIT Press, 2008), 160–161. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Seventh International World-Wide Web Conference (WWW 1998)*, 1998, http://ilpubs.stanford.edu:8090/361/, §3.1. ". . . most of the research on information retrieval systems is on small well controlled homogeneous collections such as collections of scientific papers or news stories on a related topic."

[38] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the Association of Computing Machinery* 46, no. 5 (September 1999): 604–32, doi:10.1145/324133.324140.

[39] Sergey Brin and Lawrence Page, *Dynamic Data Mining: Exploring Large Rule Spaces by Sampling*, Technical Report (Stanford InfoLab, November 1999), 2, http://ilpubs.stanford.edu:8090/424/.

[40] Sergey Brin, "Near Neighbor Search in Large Metric Spaces," *21th International Conference on Very Large Data Bases (VLDB 1995)*, 1995, http://ilpubs.stanford.edu:8090/113/. The historical section of Brin's paper offers a rich multiple genealogy of such efforts.

[41] Brin and Page, *Dynamic Data Mining*, 7.

[42] Brin, Motwani, and Winograd, "What Can You Do with a Web in Your Pocket," 2.

[43] Amy N. Langville and Carl D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton: Princeton University Press, 2006).

[44] Brin and Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", §4.2.

[45] For the human transformation of archival materials, see Lemov and Taub's essays in this volume.

[46] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *Intelligent Systems, IEEE* 24, no. 2 (April 2009): 12, doi:10.1109/MIS.2009.36.

[47] Redacted, "Confronting the Intelligence Future (U) An Interview with William P. Crowell, NSA's Deputy Director (U)," *Cryptolog* 22, no. 2 (1996): 1–5.