3 Data Management

3.1 Motivation

One of the most exciting opportunities of the emerging Information Age is to extract useful findings from the immense wealth of data and information acquired, computed, and stored by modern information systems. This is witnessed by both professionals and single users that every day extract valuable pieces of information from very different kinds of data sources, e.g., files and emails on their laptops, data coming from their company databases, or data available on the Internet.

The big opportunity of the Information Age

Unfortunately, as described in Chapter 1, there are many obstacles, which impede the effective exploitation of such an opportunity: users and analysts may get overwhelmed by irrelevant, or inappropriately processed or presented information – the information overload problem.

Obstacles come from the fact that datasets are often very large and growing incrementally, data sources are heterogeneous and are typically distributed. As a consequence, it is necessary to take this into account when studying, assessing, and giving recommendations about techniques for managing data. This is particularly challenging and the following issues need to be considered:

Many obstacles need to be overcome

- **Heterogeneity of data sources.** In a number of applications, it is necessary to integrate and query data coming from diverse data sources; this is inherently difficult and not especially well researched. Logic based systems, balancing expressive power and computational cost, represent the state of the art solutions; however such approaches are neither well understood nor easy to use.
- **Different data types.** Data comes in a variety of types and with different structures. It is challenging to analyse, in an integrated fashion, numeric and non-numeric data, together with images, videos, models, and data presenting particular entities as found in geographic and temporal data (as discussed in more detail in Chapter 5).
- **Data streams.** In many application areas, the data is in the form of streams, that is, data from a source that frequently produces new pieces of information (sensor data, stock market data, news data, etc.). Further investigation is required to deal with the conceptual and technical issues.
- **Working under pressure.** In some applications, such as emergency management, the analytical process must be performed as quickly as possible in order to make timely critical decisions. In such cases, 'classical' data management flow methods, involving data experts are not appropriate and 'traditional' data activities like data querying, cleaning, integration, etc. need to be accelerated.

- **Time consuming activities.** Managing different data formats or measurement units, null values, column names, etc. can be a complex and time consuming activity, even with small and simple datasets.

In the last decades, significant research effort has been directed towards managing and exploring large amounts of data, and two robust disciplines have emerged: data management and visual analytics.

Data management ensures data consistency and standards

Data management is a well understood field, researched over the past 30 years, and provides methods for effectively dealing with large datasets. The techniques aim to ensure data consistency, avoiding duplication and handling data transactions in a formal way. They rely on a common and well understood model, the relational model, useful to exchange and integrate data, and they exploit a highly optimised and standardised data access interface, which is called the SQL query language.

Figure 3.1: Visual analytics: a visionary scenario. Excerpt from the VisMaster Video, http://videotheque.inria.fr/videotheque/doc/635

Visual analytics is interactive and allows for exploratory analysis

Visual analytics has emerged only recently compared to the related topics of information visualisation and data mining. The advantages of visual analytics are that it deeply involves the user in the analysis loop, exploiting his perceptive and cognitive capabilities. It can be employed in a dynamic manner, with quick visual interaction and switching of analysis paradigms, and it is intended for exploratory analysis, especially when the goals are not clearly defined.

However, in spite of the strong advances in these two synergetic fields, a big gap exists between them, which obstructs the integration of these two disciplines. The main issues are:

- **Dynamicity.** Classical data management activities rely on the relational model and on the SQL query language and are highly optimised for a simple and inherently static two step interaction: query formulation and collecting results. With large datasets (billions of items), this approach is

unlikely to provide the response (approximately 100msec) necessary for good interaction[99].
- **Standards**. While data management is based on well-known and accepted standards (i.e., the relational model and the SQL query language) visual analytics applications tend to access and handle data in a proprietary way, lacking a shared, proven, and efficient solution.
- **User interaction life-cycle.** From the end user's point of view, who is only interested in finding information, data management interactions are essentially single user, quick, and one shot: the user expresses a query against the data, collects the results and analyses it. In contrast to this, visual analytics activities last a long time and may involve several users. Thus, assistance is required for long-term activities and collaborative work that are currently poorly supported by classical data management techniques.

Integration of data management and visual analytics is important, as illustrated by this scenario

The following scenario illustrates a tight integration of data management and visual analytics capabilities. It describes the research activities of several doctors working in different hospitals across Europe.

Doctors are coordinating their efforts to achieve a better understanding of several new allergy cases that have been reported in different European cities. The new allergy mainly affects the hands of 5-9 year old children and while it is very irritating it is not serious: it resolves itself spontaneously in about two weeks, or in a few days if treated with a common antihistamine. What puzzles the doctors is that the disease appeared at the same time in different locations across Europe and that a reasonable explanation is not available.

Integration and analysis of different data sources

A smart integration engine allows for seamless integration of data coming from different sources and in different formats, including patients' location and personal data, pictures about the allergy, notes from doctors, and case histories. Several interactive visualisations are available in the system, tightly integrated with automatic analytical tools. Exploring the data structure and content, helps doctors in choosing the most appropriate ones.

Collaboration among users

Data and findings are shared among the doctors, and the system allows for collaborative work and for saving and reopening the analytical processes. Using such a system, the doctors are able to select all the cases that belong to the new allergy, discarding similar but not related situations. After that, they start to investigate the environment in which the children live, searching for some common patterns (alimentation, dressing, pollution, climatic situation, etc.). Again, this requires new complex integration activities and analysis tools. After two weeks of research they conclude that there are not relevant similar patterns.

New analysis directions

One doctor starts to compare the temporal evolution of the allergy and its response to medicines by linking to a large medical dataset describing allergy cases. He discovers a strong pattern similarity with some relatively rare contact allergies generated by a kind of rigid plastic largely used for toys and food containers; this allergy usually manifests itself after prolonged contact with the substance. The doctor shares these findings through the system, but some research on toys and food containers fail to find that substance. Another doctor points out a fact that is rather obvious but has previously gone unnoticed: while

the allergy affects both right and left hands, most cases involve the right hand. A quick analysis reveals that the less frequent left hand cases correspond to left-handed children. The analysis moves again to the alimentation of the children focusing, this time, not on the food components but on the plastic associated with the food (i.e., boxes, bags, etc.) and on the probability of touching plastic parts.

The cause is discovered

Eventually a doctor discovers that a European company is marketing a new brand of lollipop, quite popular among children, and that the lollipop's plastic stick contains the allergenic component.

To summarise, nowadays, analysts and end users have the opportunity of extracting useful pieces of information from a wealth of data. However, several obstacles stand in the way and we have seen how data management and visual analytics need to address different, and sometimes complementary, facets of the problem. In order to effectively exploit this challenging situation, an integration between these two approaches is required, reducing the gap that exists between them. Such a process requires the solution of several theoretical and practical issues that, if not adequately addressed, could seriously compromise the opportunity that the new Information Age offers.

3.2 State of the Art

3.2.1 Data Management

This section focuses on the main research fields active in the context of data management, emphasising activities and results that are particularly relevant for visual analytics; aspects associated with visualisation issues will be discussed in Section 3.2.2.

Relational Technology

The relational technology[44] is based on research from the 1970s: Ted Codd's visionary paper introduces the relational model and the System R research project at IBM's San Jose Research Lab, in which the SQL query language appeared. In the relational data model, data is represented in tables that are connected to each other by attribute values, without any explicit navigational link in the data. The flexibility offered by this feature and SQL meant that the relational model rapidly replaced the now largely obsolete network and hierarchical data models.

Relational DBMSs dominate the market

Nowadays, relational systems dominate the market and rely on a very mature computer science technology. Modern RDBMSs (Relational Database Management Systems) allow for accessing the data in a controlled and managed fashion. They present a clear separation between data structure and content, and incorporate robust means of handling security and data consistency that is ensured by arranging data management in Atomic, Consistent, Isolated,

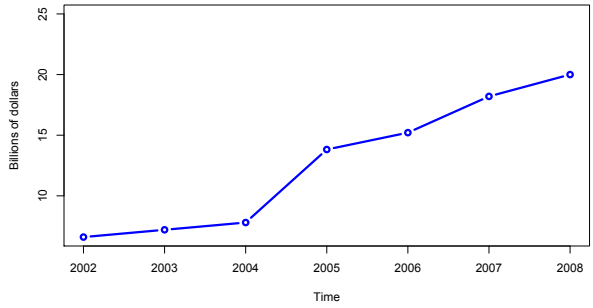

Figure 3.2: Purchases of relational database licenses in the last years (in billions of $)

and Durable transactions (so called transactions' ACID property). This permits seamless concurrent data access and data recovery in a collection of databases that is physically distributed across sites in a computer network (Distributed RDBMS), hiding the distribution details from the users that access the data through a common interface, using the widely accepted SQL query language. A coherent set of theorems and practical research on query optimisation and data indexing allows relational systems to deal with very large datasets.

The market of RDBMS is still growing: the worldwide sales of new licenses of relational database management systems (RDBMS) totalled about $20 billion dollars in 2008, increasing about three times the 2002 revenue of $6.6 billions, according to Gartner, Inc. as shown in Figure 3.2.

Data Integration

Providing unified and transparent access to a set of heterogeneous sources

Data integration is the problem of providing unified and transparent access to a set of autonomous and heterogeneous sources, in order to allow for expressing queries that could not be supported by the individual data sources alone. There is a big and still growing need for systems and techniques that support such a process, and very likely it is one of the major challenges for the future of IT. The problem is ubiquitous in modern software systems, and comes in different forms: data sources are characterised by a high degree of heterogeneity (e.g., different data models, different data types, different nomenclature, different data units, etc.), raising many challenges, and a number of methodologies, architectures, and systems have been developed to support it.

Data integration can be centralised, that is being performed within the same organisation (e.g., Enterprise Information Integration) or can be decentralised, involving two or more organisations, usually based on a peer-to-peer architecture. The latter assumes a data-centric coordination among the autonomous

organisations to dynamically expose a view of their data using an agreed data schema.

The integration can be virtual or materialised. In the first case, the data does not move from the original source and the integration is performed at query time; in the second case chunks of data are physically exchanged before the query process and collected in a single place (e.g., data warehousing).

The most relevant approach for visual analytics is the centralised, virtual information integration that represents an evolution of ideas dating back to the 80s. A collection of theoretical results is available, but a robust and definitive solution is still far from being reached. The available solutions foresee several tools for data source wrapping and database federation (e.g., DB2 Information Integrator), providing a common model for exchanging heterogeneous data and allowing physical transparency (i.e., masking from the user the physical characteristics of the sources), handling heterogeneity (federating highly diverse types of sources), preserving the autonomy of the data sources, and ensuring scalability (distributed query optimisation).

Semantic integration

However, these tools do not provide conceptual data transparency, i.e., they present the data as it is stored within the sources, leaving the heterogeneity arising from different naming, data representation, etc., unsolved. The most promising solution to this problem is called semantic integration[23] and is based on the idea of computing queries using a logic based engine that exploits a conceptual view of the application domain (i.e., an ontology), rather than a flat description of the data sources. Such a description, called a global schema, is independent from the sources that are mapped through a logic language into concepts of the global schema. A solution that is being adopted more often is to use, as a logic language the so called 'description logics' that are a subset of the first order logic and balance expressive power and computational cost.

Data Warehousing, OLAP and Data Mining

Data warehousing, OLAP (On-Line Analytical Processing), and data mining share many of the goals of visual analytics: they are intended for supporting, without the explicit use of visualisations, strategic analysis and decision-supporting processes.

Data warehousing for decision making

A data warehouse[62] is an integrated repository of data that can be easily understood, interpreted, and analysed by the people who need to use it to make decisions. It is different from a classical database for the following reasons: it is designed around the major entities of interests of an organisation (e.g., customers, sales, etc.), it includes some related external data not produced by the organisation and it is incremental, meaning that data, once added, is not deleted, allowing for analysing temporal trends, patterns, correlations etc. Moreover it is optimised for complex decision-support queries (vs. relational transactions). The different goals and data models of data warehousing

stimulated research on techniques, methodologies and methods, which differ from those used for relational DBMS.

The term OLAP[31] refers to end-user applications for interactive exploration of large multidimensional datasets. OLAP applications rely on a multidimensional data model created to explore the data from different points of view through so called data cubes (or data hypercubes), i.e., measures arranged through a set of descriptive categories, called dimensions (e.g., sales for city, department, and week). Hierarchies are defined on dimensions, (e.g., week ... month ... year) to enable additional aggregation levels. A data cube may hold millions of entries characterised by tens of dimensions and one of the challenges is to devise methods that ensure a high degree of interactivity. One solution is to pre-compute and store aggregated values for different levels of the hierarchies and reduce the size of the data (see below), thus sacrificing precision for speed. Another consideration is system usability. The user can only explore a small number of dimensions at any one time (i.e. the hypercube needs to be projected onto two-dimensional or three-dimensional spaces) and hence to gain insights into high dimensional data may require long and sometimes frustrating explorations.

Mining insights from large datasets

Data mining is the process of discovering knowledge or patterns from massive amounts of data through ad hoc algorithms. Data mining can be categorised based on the kinds of data to be analysed, such as relational data, text, stream, Web data, multimedia (e.g., image, video), etc. Its relationship with visualisations became more prevalent in the 90s when the term 'visual data mining' emerged, denoting techniques for making sense of data mining algorithms through different visualisations, built on both the mined data and on the results produced by the algorithms. The topic of data mining is further discussed in Chapter 4.

Data Reduction and Abstraction

In the context of data management, data reduction techniques have been used to obtain summary statistics, mainly for estimating costs (time and storage) of query plans in a query optimiser. The precision is usually adequate for the query optimiser and is much cheaper than a full evaluation of the query.

Data reduction can improve query optimisation and interaction

More recently the focus has moved onto data reduction techniques to improve the interactivity for OLAP applications operating on large amounts of data stored in the organisation data warehouse. Due to the analytical and exploratory nature of the queries, approximate answers are usually acceptable.

In summary, the purpose of data reduction in the context of data management is to save computational or disk access costs in query processing or to increase the systems responsiveness during interactive analysis. Data reduction relies on various techniques, like histograms, clustering, singular value decomposition, discrete wavelet transforms, etc. A comprehensive summary of data reduction techniques for databases can be found in the New Jersey data reduction

report[11]. Data reduction techniques can be usefully exploited in the context of visual analytics by reducing the number of dimensions and/or the complexity of relationships.

Data Quality

Databases often have to deal with data coming from multiple sources of varying quality - data could be incomplete, inconsistent, or contain measurement errors. To date, several research lines and commercial solutions have been proposed to deal with these kinds of data errors, in order to improve data quality.

Linking different views of the same data

Data conflicts have been studied by statisticians that needed to resolve discrepancies rising from large statistical surveys. One of the first problems of this kind was the presence of duplicated records of a person[43], and the devised practical and theoretical solution, called record linkage, allowed the collection and linkage of all the related data records, producing a unique and consistent view of the person. It was quickly understood that record linkage was only one of a larger set of problems, such as wrong, missing, inaccurate, and contradicting data, and in the late 1980's, researchers started to investigate all problems related to data quality. This line of research was advanced by both the increasing number of scientific applications based on large, numerical datasets and by the need to integrate data from heterogeneous sources for business decision making.

Restoring missing data

The problem of missing data was initially studied in the context of scientific/numerical datasets, relying on curative methods and algorithms able to align scientific data. More recently, the focus has moved on to non-numerical data and in particular, dealing with inherently low quality datasets such as information extracted from Web and sensor networks. MystiQ[19] is an example of research into building general purpose tools to management uncertain data.

Polishing the data

Dealing with missing data and duplicate records is only part of the overall process of data cleansing. We also need to identify and either correct or reject data that is incorrect or inaccurate, possibly through the use of aggregate data statistics. Additionally, the data many need to be standardised by, for example, adjusting the data format and measurement units.

3.2.2 Data Management and Information Visualisation

The data management research field acknowledges the key role that information visualisation can play in enhancing data management activities through ad hoc visualisation. In the following section, we describe some examples, which show the synergy that exists between data management and information visualisation.

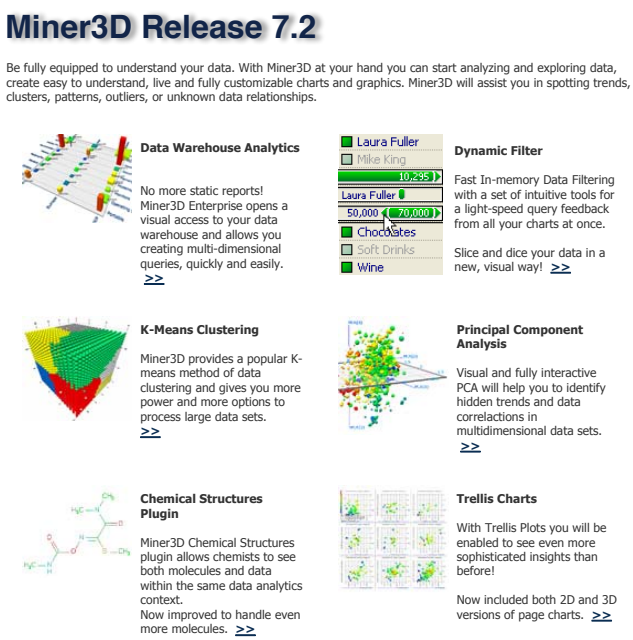

Figure 3.3: The commercial system Miner3D

Visual Data Mining

Using visualisation to enhance data mining

The inherent difficulties associated with data mining algorithms together with the need to justify the data mining results, has stimulated the development of integrated environments in which suitable visualisations are used as a complementary technique to support data mining. The combination of visualisation and data mining is known as 'visual data mining'.

This new field presents strong correlation with several synergic fields, i.e., knowledge discovery in databases, information visualisation and human-computer interaction. Common elements have been recognised that need to be specified when developing methodologies for visual data mining. These include: a) the initial assumptions posed by the respective visual representations; b) the set of interactive operations over the respective visual representations and guidelines for their application and interpretation, and c) the range of applicability and limitations of the visual data mining methodology. Several research projects have dealt with this new challenging research field, like the 3D Visual Data Mining (3DVDM) project in Aalborg University representing the dataset in various stereoscopic virtual reality systems. Commercial products such as VMiner3D[35] (see Figure 3.3), demonstrate the usefulness of the combination of data mining and visualisation.

Visual OLAP

A clear trend in business visualisation software exists, showing a progression from basic and well-understood data visualisations to advanced ones, and this is the case of OLAP applications that present emerging techniques for advanced interaction and visual querying.

To date, the commonly used OLAP interfaces enhance the traditional way of arranging data on a spreadsheet through automatic aggregation and sort functions that store the result in a second table (called a pivot table). This allows the end user to explore data cubes through traditional visualisation techniques such as time series plots, scatterplots, maps, treemaps, cartograms, matrices etc., as well as more specialised visualisations such as decomposition trees and fractal maps. Some applications also integrate advanced visualisation techniques developed by information visualisation researchers.

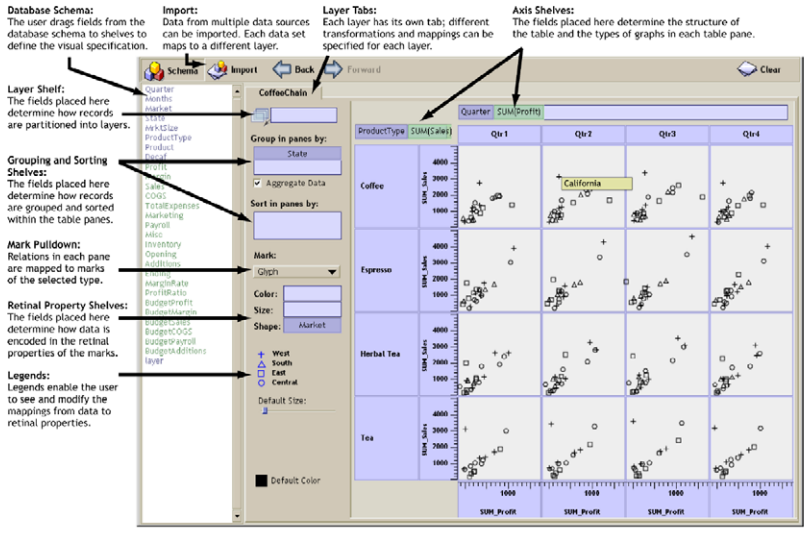

Figure 1: *The Polaris user interface. Analysts construct table-based displays of relational data by dragging fields from the database schema onto shelves throughout the display. A given configuration of fields on shelves is called a visual specification. The specification unambiguously defines the analysis and visualization operations to be performed by the system to generate the display.*

Figure 3.4: The Polaris interface as presented in the original paper[107] © 2002 IEEE

Two pioneering systems in visual online analytical processing

Polaris and ADVIZOR are among the first attempts in such a direction. Polaris is a visual tool for multidimensional analysis developed at Stanford University[107] and inherits the basic idea of the classical pivot table interface, using embedded graphical marks rather than textual numbers in the table cells. The types of supported graphics are arranged into a taxonomy, comprising of rectangle, circle, glyph, text, Gantt bar, line, polygon and image layouts (See Figure 3.4). Currently, Tableau Software commercialises the pioneering Polaris work. ADVIZOR represents the commercialisation of 10 years of research in Bell Labs on interactive data visualisation and in-memory data management[39].

It arranges multidimensional data from multiple tables onto a series of pages, each one containing several linked charts (from 15 chart types), which facilitates ad hoc exploration of the data.

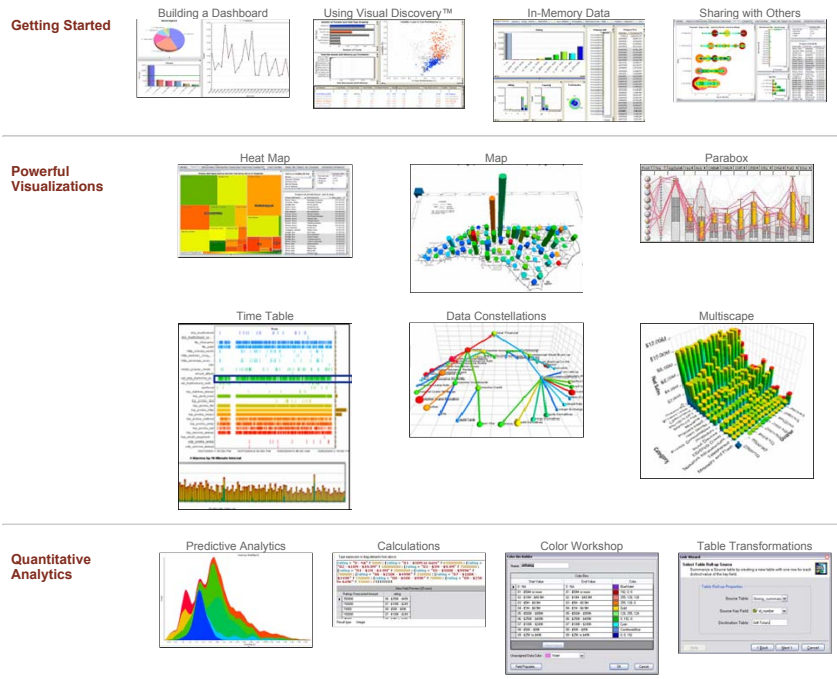

Figure 3.5: The rich set of linked visualisations provided by ADVISOR include barcharts, treemaps, dashboards, linked tables and time tables

Visual Data Reduction

Scaling visual presentations to high dimensional data or to billions of records is one of the challenges that visual analytics has to deal with, and requires a tight collaboration between data management and visual analytics. Visualising large datasets, often produces cluttered images, and hence various clutter reduction techniques[41] have been developed by the information visualisation community to ameliorate this. However, as mentioned previously, visual analysis is limited if the system does not allow for quick interaction while exploring the data. This requires new scalable data structures and algorithms for data reduction and/or innovative hierarchical data management. While several proposals are available, e.g., sampling, density-based hierarchical data aggregation and multi-resolution representation, a common understanding on how to interactively visualise vast datasets does not exist.

Visualising large datasets often produces cluttered images

Some clutter reduction techniques are fairly complex (e.g., dimension reduction), while other ones are relatively straightforward (e.g., uniform sampling).

However, disregarding the technical details and the computational aspects, all of them share a set of common problems:

- When is the technique needed?
- Which technique is the most appropriate for a given dataset and visualisation?
- To which extent do we have to apply it (e.g., how much do we have to sample)?
- How can we evaluate the quality of the reduced dataset?

Measuring the visualisation quality with quality metrics

Addressing these questions highlights some complicated issues, involving the data structure and cardinality, the user task, the chosen visualisation as well as perceptual and cognitive aspects. Several proposals addressed these issues using objectives *quality metrics* that capture some relevant data reduction measure. For example, the number of data items being displayed on a scatterplot can be used as a trigger to decide *when* to apply a reduction technique (e.g., sampling), *how much* to sample, and *to compare* the final result with another technique (e.g., data clustering). Obviously, several, non trivial parameters affect this process, like threshold values, data distribution, image size, etc. However, the matter still deserves further study as several issues are still far from being solved, such as how to assess the quality of a visualisation produced from the application of a data reduction technique.

Figure 3.6: Visual data reduction preserving density differences through visual quality metrics. The image shows a curative sampling algorithm applied to the topmost scatterplot producing a more comprehensible image (below). The measurements on the right allow the user to formally assess the performance of the algorithm

Data reduction techniques belong to two different categories:

- those that use quality metrics to optimise non visual aspects, e.g., time, space, tree balancing, etc;

- those that use quality metrics to optimise the *visualisation* of some data aspects relevant to the analytical process. We call this activity *visual data reduction*.

Visual data reduction perfectly fits the visual analytics philosophy: a) an automated analysis is performed on the data, discovering and measuring different relevant data aspects (e.g., strong correlation, outliers, density differences) and b) such measures are used as quality metrics to drive and evaluate the data reduction activity, presenting the end user with the visualisation that best conveys these interesting data aspects. But we must ensure that these facets are effectively *presented* to the end user as image degradation or perceptual issues could hide the precious insights highlighted by the automatic analytical process.

Visual quality metrics: size, visual effectiveness and feature preservation

Quality metrics adopted in the visual data reduction process must encompass data issues and visualisation issues in a strongly integrated fashion. In an example of this[17] the authors categorise these *visual quality metrics* in three classes: size metrics, visual effectiveness metrics, and feature preservation metrics. Size metrics have a descriptive character (e.g., number of data points) and are the basis for all further calculations. Visual effectiveness metrics are used to measure the image degradation (e.g., collisions, occlusions, etc.) or outliers. Examples are data density, collisions, and the data-ink-ratio. Feature preservation metrics are the core of visual quality metrics and are intended for measuring how correctly an image represents some data characteristics, taking into account data issues, visual aspects and perceptual aspects. Figure 3.6 illustrates such a feature preserving data reduction technique[16] (non uniform sampling) for 2D scatterplots, driven and evaluated by visual quality metrics that use data, visualisation and perceptual parameters, with the main goal of discovering and preserving data density differences.

Visualisation for the Masses

Visualising data on the Web has improved considerably in the last few years

Tools supporting search, browsing, and visualisation have dramatically improved in the past decade, so that it is possible to design and implement new Web based system integrating data management technologies and interactive information visualisation. An example of these new opportunities comes from the Swedish non-profit company Gapminder[1], acquired by Google that designed a system to make world census data available to a wider audience. The relational data of the census can be explored with two linked visualisations, a geographical map plus a two-dimensional scatterplot that uses colour and size to visualise two additional attribute values (see Figure 3.7). The system is easy to use and by allowing the user to explore the change of the variables over time, its effectiveness is enhanced considerably.

In summary, data management and visual analytics are two disciplines that, together, are able to exploit the opportunities coming from the Information Age. This survey of the state of the art shows, however, that while strong results have

[1] `http://graphs.gapminder.org/world/`

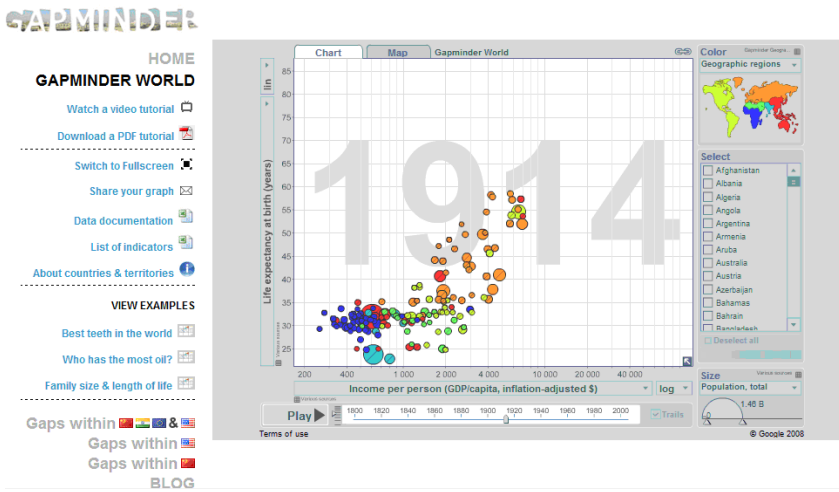

Figure 3.7: The GapMinder visualisation clearly shows the correspondence that exists between life expectancy at birth and income per person. Moreover, it shows the dramatic differences that exist between India, North Europe and North America

been reached within these two fields there are still some unexplored issues and some requirements that have not been addressed that impede the integration of fields. In particular, we need better logic based integration systems, a tighter integration between visualisation and data, more precise quality indicators, visual oriented data reduction techniques, and new interaction paradigms for large sets of users. In order to progress, it is mandatory to make these two disciplines more aware of each other's results and needs, addressing both common problems and specific issues.

3.3 Challenges and Opportunities

Symbiotic dependency of visual analytics and data management

In this section, we highlight the most important challenges that data management and visual analytics have to deal with to better exploit the opportunity of the Information Age. As, on the one hand, data management is intrinsic to visual analytics, solving some of data management's open problems will enhance visual analytics applications. On the other hand, specific visual analytics issues will pose new challenges for the data management community. Therefore, it is important to reflect upon the different perspectives and the inherent relationships when considering the mutual roles played by data management and visual analytics.

Moreover, as a general consideration, it is worth noting that some critical visual analytics issues are the subject of research effort in different scientific areas, often with different names. For example, the activity of data sampling is performed with different goals in both data management and information

visualisation research activities, but a shared view of problems and techniques between these two research fields does not exist. Therefore, it is essential to find and encourage such potential synergies.

Uncertainty

How to visualise missing data?

Solving issues resulting from incomplete, inconsistent, or erroneous data is crucial for both visual analytics and data management. Therefore, both robust and agreed methodologies are required. However, visual analytics looks at these issues in a different way and the straightforward adoption of the solutions proposed in the data management field could be either a valid solution or an obstacle to the analysis process. For example, assume that we are dealing with a missing or erroneous value. The data management techniques may use some curative algorithms, providing an alternative (e.g., interpolated or statistically computed) value for the bad data, but this solution can hide important facts; perhaps the value is empty because a person omitted to enter a value on the form to evade paying tax or an out of range value indicates a faulty sensor?

Data visualisation also has methods of dealing with missing data and so it has to be decided whether data management or the visualisation has responsibility for managing this. Whatever subsystem takes charge, it is necessary to remember the decisions made during the cleaning activities so that the user can be made aware of any uncertainties in the data.

Data Integration

Need for new integration systems

The integration of heterogeneous data is a core data management activity and its importance and use are increasing. Logic based systems, balancing expressive power and computational cost represent state of the art solutions. Visual analytics can greatly benefit from such an approach and does not raise particular issues in such a context, apart from situations that require quick decision making (e.g., emergency management) or upon performing data integration without expert support. In such cases, the integration engine should present an interface intended for non expert users and be able to make decisions with incomplete information, e.g., while integrating data coming from the Web. This is a new and challenging requirement, not addressed by traditional data management research.

Semantics Management

Making semantics a first class citizen

Associated with data integration activities, is the need for managing all the data semantics in a centralised way, for example, by adding a virtual logic layer on the top of the data itself. For example, data semantics could be used to describe synonyms such as 'is-a' relationships (e.g., a student is-a person and, as a consequence, everything holds for person holds for a student as well)

and constraints (e.g., you must be at least 18 years old to hold an Italian car driving license). This is useful not only for data integration, but also for dealing with all the semantic issues involved in the analytical process, like metadata management, abstraction levels, hierarchical structures and externalisation. Visual analytics applications should manage all the available semantics at one point, under the responsibility of the database management system. That includes also the semantics that are discovered during analytical (manual and automatic) activities – once discovered it should be added to the top virtual logic layer.

Such a challenging kind of semantic integration has been not researched in both visual analytics and data management fields and could represent an interesting starting point for cooperation between the two disciplines. This also represents a strong opportunity: semantics discovered during the analytical process could be usefully exploited for improving database activities and database performances, e.g., improving data cleaning and the query optimisation process.

Data Provenance and Integrity of Results

Where does the data come from?

While performing visual analytics and data management activities, the end user may need to inspect and understand the path associated with some data items. That includes, at least, a) the physical source hosting the original data, b) the reason why the data belongs to the observed dataset (that is useful when the query process is performed using some kind of logical deduction), and c) a way for better understanding the results of an automatic analysis step (e.g., a data mining classification or a record linkage). Moreover, while performing long and complex visual analytics activities, it could be useful to view the series of actions that generated the current display, i.e., what data and what transformations have been used to generate the actual visualisation?

Data Streaming

Visual analytics applications sometimes have to deal with dynamic data (i.e., new data is received on a regular basis) whilst the analysis process is running. For instance, a visual analysis of a social network, based on a live feed of their data. The analysis has to gracefully adjust to the updates; stopping the process and triggering a total re-computation would not be appropriate.

Continuous flows of data require special study

The following three aspects of data streams require further study, at a conceptual and technical level, in order to address the visual analytics and data management objectives:

- **Building data stream management systems.** That implies studying architectures and prototypes, stream-oriented query languages and operators, stream processing and efficient algorithms to keep an up-to-date online connection to the data sources.

- **Designing efficient algorithms for stream analysis.** In particular, we need algorithms that are able to proceed in an incremental way, mining information from the stream and capturing both trends and overall insights.
- **Change detection analysis.** Sometimes the analysis looks for relevant changes that happen within the stream, allowing for the fast detection of new or unexpected behaviours.

Time Consuming Low Level Activities

Managing diverse data types can be time consuming for the analyst

Data management and visual analytics problems are not always due to the large size of the dataset. Dealing with small details such as data heterogeneity, data formats and data transformation can be a time consuming and hence an unwelcome burden on the analyst. In these cases, new consistency checking languages could offer assistance, relieving the analyst of coding in SQL. In general, there needs to be a better comprehension of the role of low-level data management activities in the visual analytics process.

Further time consuming activities include selecting the appropriate view on the data, joining relevant tables and data sources, selecting the best visualisation for exploring the dataset, and associating data values to visual attributes. These call for some form of automation, which is able to assist the analyst in speeding up the overall analysis process. This issue is strongly connected with heterogeneous data integration and semantics management, as mentioned earlier, and researchers should address logic based solutions, capturing both predefined and discovered semantics in order to drive automatic or semi-automatic algorithms.

Interactive Visualisation of Large Databases

Visualising billions of items

Whilst the storage and retrieval of data from very large datasets is well understood, supporting effective and efficient data visualisations with, say billions of items and/or hundreds of dimensions, is still a challenging research activity. In particular, we need to provide the user with rapid feedback while exploring the data. Promising solutions come from different techniques of (visual) data reduction, able to scale on both data cardinality and data dimensions. Additionally, there are proposals to pre-compute metadata, e.g., indexing or aggregating data for visualisation purposes. However, the field is still a challenging one, and more formal approaches are needed, e.g., using formal definition of quality and visual quality metrics.

Researching this topic is crucial for visual analytics and its importance is also being acknowledged in the data management area. This suggests the pursuit of joint research efforts in areas such as new scalable data structures, novel algorithms for data reduction, innovative hierarchical data management and supporting visual analytics applications to adopt the data models of data management.

Distributed and Collaborative Visual Analytics

Visual analytics activities are more complex and longer than issuing a single query

Visual analytics activities are longer and more complex than issuing a single query against a dataset and exploring the result; moreover, they often involve several users at different sites. The process itself requires intermediate steps, saving and presenting partial results, annotating data and insights, resuming previous analysis, as well as sharing findings with the different users. Also, it is beneficial to be able to automatically reapplying the same visual analytics steps on different datasets or on a changed one, as with streaming data.

Long term and collaborative activities are poorly supported by classical data management techniques, and in order to reduce this gap, new research directions should be instigated, exploring collaborative systems explicitly designed to help the visual analytical process.

Visual Analytics for the Masses

Managing personal data is increasingly prevalent

The volume of personal digital data (i.e., emails, photos, files, music, movies, etc.) is increasing rapidly and with the availability of new Web based systems integrating data management technologies and information visualisation, this opens up new opportunities for visual analytics applications and new challenges. The home user becomes a naive analyst and this requires different interaction modalities for non-expert people and raises heterogeneity (data source and devices), scalability, and user acceptance issues.

Summary

Challenges for both the visual analytics and data management communities

Many challenges and opportunities associated with data management and visual analytics exist. They are related to solving basic data management problems that will help visual analytics activities, or to addressing problems arising from specific requirements of visual analytics. On the other hand, data management could fruitfully exploit by some results coming from visual analytics research. However, in order to make progress in the visual analytics field, we need to address some critical issues such as uncertainty problems, semantic data integration and semantics management, data provenance, data streaming, interactive visualisation of huge datasets, solving process intensive activities, and designing visual analytics systems intended for the general public. Dealing with these issues is a challenge that both communities have to take up, in order to take advantage of the increasing information opportunities available today.

3.4 Next Steps

By examining the state of the art in data management and the requirements of visual analytics, it is clear that the two displines would mutually benefit

from increased cooperative research and have subsequently identified particular challenges faced by these communities. We now describe what we would consider to be useful next steps towards stimulating future developments in data management and visual analytics research that will eventually enable new solutions that exploit the strengths of modern data management technology in the context of advanced visual analytics scenarios.

We recommend the activation of research projects bringing together the interdisciplinary competencies of visual analytics and data management, in order to progress and to gain a better understanding of the problems associated with the challenges described in the previous section. In particular, we point out three main areas that should to be addressed:

- The development of a new generation of data integration systems, based on the most advanced data management results and targeted to the specific, compelling visual analytics requirements, like high heterogeneity of data sources and data types, critical time constraints, and methods for effectively managing inconsistent and missing data.
- The development of new data reduction and analysis techniques for dealing with the modern complex data, such as high dimensional data and data streams.
- The development of new algorithms, data structures and visual data reduction techniques to facilitate the interactive visualisation of extremely large datasets.

Effort is required to distribute the ideas discussed in this chapter to potentially interested colleagues. This involves the dissemination of information about the potential advantages of enabling visual analytics in the data management research field, as well as the dissemination of information about the state of the art in data management and its according promises to visual analytics researchers. In addition, we suggest that steps are taken to improved literacy in each of the two fields in respective of other field, and as a longer term goal, update the educational curriculum to reflect the interdisciplinary nature of this topic.