MORGAN & CLAYPOOL PUBLISHERS

Perspectives on Business Intelligence

Raymond T. Ng
Patricia C. Arocena
Denilson Barbosa
Giuseppe Carenini
Luiz Gomes, Jr.
Stephan Jou
Rock Anthony Leung
Evangelos Milios
Renée J. Miller
John Mylopoulos
Rachel A. Pottinger
Frank Tompa
Eric Yu

SYNTHESIS LECTURES ON DATA MANAGEMENT

M. Tamer Özsu, *Series Editor*

Perspectives on Business Intelligence

Synthesis Lectures on Data Management

Editor
M. Tamer Özsu, ***University of Waterloo***

Synthesis Lectures on Data Management is edited by Tamer Özsu of the University of Waterloo. The series will publish 50- to 125 page publications on topics pertaining to data management. The scope will largely follow the purview of premier information and computer science conferences, such as ACM SIGMOD, VLDB, ICDE, PODS, ICDT, and ACM KDD. Potential topics include, but not are limited to: query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Perspectives on Business Intelligence
Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr., Stephan Jou, Rock Anthony Leung, Evangelos Milios, Renée J. Miller, John Mylopoulos, Rachel A. Pottinger, Frank Tompa, and Eric Yu
2013

Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications
Amit Sheth and Krishnaprasad Thirunarayan
2012

Data Management in the Cloud: Challenges and Opportunities
Divyakant Agrawal, Sudipto Das, and Amr El Abbadi
2012

Query Processing over Uncertain Databases
Lei Chen and Xiang Lian
2012

Foundations of Data Quality Management
Wenfei Fan and Floris Geerts
2012

Incomplete Data and Data Dependencies in Relational Databases
Sergio Greco, Cristian Molinaro, and Francesca Spezzano
2012

Business Processes: A Database Perspective
Daniel Deutch and Tova Milo
2012

Data Protection from Insider Threats
Elisa Bertino
2012

Deep Web Query Interface Understanding and Integration
Eduard C. Dragut, Weiyi Meng, and Clement T. Yu
2012

P2P Techniques for Decentralized Applications
Esther Pacitti, Reza Akbarinia, and Manal El-Dick
2012

Query Answer Authentication
HweeHwa Pang and Kian-Lee Tan
2012

Declarative Networking
Boon Thau Loo and Wenchao Zhou
2012

Full-Text (Substring) Indexes in External Memory
Marina Barsky, Ulrike Stege, and Alex Thomo
2011

Spatial Data Management
Nikos Mamoulis
2011

Database Repairing and Consistent Query Answering
Leopoldo Bertossi
2011

Managing Event Information: Modeling, Retrieval, and Applications
Amarnath Gupta and Ramesh Jain
2011

Fundamentals of Physical Design and Query Compilation
David Toman and Grant Weddell
2011

Methods for Mining and Summarizing Text Conversations
Giuseppe Carenini, Gabriel Murray, and Raymond Ng
2011

Probabilistic Databases
Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch
2011

Peer-to-Peer Data Management
Karl Aberer
2011

Probabilistic Ranking Techniques in Relational Databases
Ihab F. Ilyas and Mohamed A. Soliman
2011

Uncertain Schema Matching
Avigdor Gal
2011

Fundamentals of Object Databases: Object-Oriented and Object-Relational Design
Suzanne W. Dietrich and Susan D. Urban
2010

Advanced Metasearch Engine Technology
Weiyi Meng and Clement T. Yu
2010

Web Page Recommendation Models: Theory and Algorithms
Sule Gündüz-Ögüdücü
2010

Multidimensional Databases and Data Warehousing
Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen
2010

Database Replication
Bettina Kemme, Ricardo Jimenez-Peris, and Marta Patino-Martinez
2010

Relational and XML Data Exchange
Marcelo Arenas, Pablo Barcelo, Leonid Libkin, and Filip Murlak
2010

User-Centered Data Management
Tiziana Catarci, Alan Dix, Stephen Kimani, and Giuseppe Santucci
2010

Data Stream Management
Lukasz Golab and M. Tamer Özsu
2010

Access Control in Data Management Systems
Elena Ferrari
2010

An Introduction to Duplicate Detection
Felix Naumann and Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview
Raymond Chi-Wing Wong and Ada Wai-Chee Fu
2010

Keyword Search in Databases
Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2009

Perspectives on Business Intelligence

Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr., Stephan Jou, Rock Anthony Leung, Evangelos Milios, Renée J. Miller, John Mylopoulos, Rachel A. Pottinger, Frank Tompa, and Eric Yu

www.morganclaypool.com

ISBN: 9781627050937 paperback
ISBN: 9781627050944 ebook

DOI 10.2200/S00491ED1V01Y201303DTM034

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #32
Series Editor: M. Tamer Özsu, *University of Waterloo*
Series ISSN
Synthesis Lectures on Data Management
Print 2153-5418 Electronic 2153-5426

Perspectives on Business Intelligence

Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr., Stephan Jou, Rock Anthony Leung, Evangelos Milios, Renée J. Miller, John Mylopoulos, Rachel A. Pottinger, Frank Tompa, and Eric Yu

SYNTHESIS LECTURES ON DATA MANAGEMENT #32

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

In the 1980s, traditional Business Intelligence (BI) systems focused on the delivery of reports that describe the state of business activities in the past, such as for questions like "How did our sales perform during the last quarter?" A decade later, there was a shift to more interactive content that presented how the business was performing at the present time, answering questions like "How are we doing right now?" Today the focus of BI users are looking into the future. "Given what I did before and how I am currently doing this quarter, how will I do next quarter?"

Furthermore, fuelled by the demands of Big Data, BI systems are going through a time of incredible change. Predictive analytics, high volume data, unstructured data, social data, mobile, consumable analytics, and data visualization are all examples of demands and capabilities that have become critical within just the past few years, and are growing at an unprecedented pace.

This book introduces research problems and solutions on various aspects central to next-generation BI systems. It begins with a chapter on an industry perspective on how BI has evolved, and discusses how game-changing trends have drastically reshaped the landscape of BI. One of the game changers is the shift toward the consumerization of BI tools. As a result, for BI tools to be successfully used by business users (rather than IT departments), the tools need a business model, rather than a data model. One chapter of the book surveys four different types of business modeling. However, even with the existence of a business model for users to express queries, the data that can meet the needs are still captured within a data model. The next chapter on vivification addresses the problem of closing the gap, which is often significant, between the business and the data models. Moreover, Big Data forces BI systems to integrate and consolidate multiple, and often wildly different, data sources. One chapter gives an overview of several integration architectures for dealing with the challenges that need to be overcome.

While the book so far focuses on the usual structured relational data, the remaining chapters turn to unstructured data, an ever-increasing and important component of Big Data. One chapter on information extraction describes methods for dealing with the extraction of relations from free text and the web. Finally, BI users need tools to visualize and interpret new and complex types of information in a way that is compelling, intuitive, but accurate. The last chapter gives an overview of information visualization for decision support and text.

KEYWORDS

business intelligence, big data, business modeling, vivification, data integration, information extraction, information visualization

Contents

CHAPTER 1

Introduction and the Changing Landscape of Business Intelligence

Stephan Jou and Raymond Ng

1.1 INTRODUCTION

A lot has changed since 1958 when IBM researcher Hans Peter Luhn first coined the term Business Intelligence (BI) as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action toward a desired goal" [Luhn, 1958]. In particular, the BI domain has seen dramatic changes in its use of temporal information, the nature of the data to analyze, cloud computing, user-centric and consumable analytics. All of these changes demand new, enabling research and technology capability that are exemplified by this book.

BI systems have traditionally focused on the delivery of reports that describe the state of business activities in the past. Questions like "How did our sales perform during the last quarter?" were answered through straightforward query generation and execution against structured and multidimensional data warehouses and delivered to end-users in a static report, such as a PDF document or simple web page.

In the 1990s, there was a shift from static reports of past performance to more interactive content that presented how the business was performing at the present time, answering questions like "How are we doing right now? This month, this day, this second?" This shift to real-time business intelligence was supported with new technologies: animated real-time dashboards, interactive filters, prompts, and multidimensional gestures augmented the classical, static content, while in-memory databases and other performance-enabling infrastructures surfaced.

Now the focus of business intelligence systems has shifted in the time domain yet again. In addition to asking questions about the past and the present, BI users are looking into the future. "Given what I did before and how I am currently doing this quarter, how will I do next quarter? How am I predicted to perform, and how does that compare to competitors who are in a similar situation? How can I tweak what I am doing right now to optimize my future?"

The addition of statistical and predictive analytical techniques—optimization, predictive modeling, simulation, and forecasting—to traditional BI methods—querying, reporting, and analysis—

has resulted in solutions that can predict and optimize the future. The development and incorporation of these future-facing technologies is one of the key events that strongly distinguishes classical business intelligence systems of the past from the new "business analytics and optimization" systems that have emerged from industry.

Dramatic changes have also occurred in the data and information that we now want to analyze. When BI systems were first adopted by industry, data warehouses were still being formed and sparsely populated, and therefore the BI technologies that emerged emphasized data collection and integration. The relational and OLAP databases held static snapshots of primarily numerical, very structured data held in well-defined schema. However, this picture has changed, and many in the industry today characterized the new data challenges with the "three V's" of volume, variety, and velocity.

These dramatic changes in data helped fuel the emergence and importance of enabling cloud computing technologies. In many ways, it was these new data demands that led to the development of our current ability to create and use virtualized computational, storage, analytical, and other infrastructural or platform services in various cloud environments.

Much of the new data of interest to businesses, particularly unstructured social data, comes from the Internet and was, therefore, "born" in the public cloud. At the same time, however, significant value continues to be extractable from data being generated or collected on-premise inside the corporate firewall, with potentially even greater value from the combination of both public and on-premise data. The large volume and high velocity nature of the data means copy operations from one location to another are impractical and sometimes impossible. As a result, it made sense to flexibly locate the systems required to analyze the data as close to its origins as possible, resulting in the development of public, on-premise (private), and hybrid cloud computing environments.

At the same time, analyzing the new varieties and large volumes of data requires significant computational power. The new predictive, text, and social analytics algorithms can sometimes require large numbers of computers—hundreds, thousands, even more—to work on the problem simultaneously in order to return a result with sufficient velocity. The rapid sequestration and orchestration of a large number of (virtual) computational and storage units for a relatively short period of time to perform these sorts of analyses cannot be practically done using traditional hardware infrastructure; a cloud computing infrastructure is the only cost-effective method of enabling this next generation of business analytics solutions.

Finally, the ability to programmatically define the deployed infrastructure in a cloud computing environment means that we can treat infrastructure as if it was software: we can algorithmically describe and tune the infrastructure to match the analysis we want to perform. This affords us tremendous flexibility and power compared to traditional BI infrastructure, allowing us to design large scale analytical solutions that were previously expensive or intractable.

Furthermore, traditional BI systems had a well-defined and controlled consumption model. IT specialists owned the data warehouses and the processes surrounding the data stored within those warehouses. Report authors created the reports and dashboards that consumed the data, and

the resultant web pages or PDF files were delivered to BI end users. BI specialists created BI object models or predictive models to add semantics, rules, and predictive capabilities to the reports.

Now, the focus has definitely shifted from those specialized groups to the individual BI end users as the primary audience for the industry's next generation software. These new end users are very familiar with new technologies such as the Internet and mobile devices, and expect a level of interactivity, performance, and ease-of-use that traditional BI software has had to evolve and aspire to. They want to be able to analyze and combine the data that sits on their desktop files or that they have discovered on the Internet, and not just the enterprise data that has been made available to them by the IT specialists. They expect results in minutes to seconds, not hours to days. And finally, these end users might not have the training of a classical statistician, but they understand their business domain area completely and want to be able to perform deep predictive analyses themselves, without involving the BI specialists.

Building systems to enable these new user-centric and consumable analytics involves more than just recognizing the primacy of user experience and human-centric design to software, although that is certainly important. The challenges to supporting such systems range from infrastructural—how can we perform such difficult calculations quickly enough to power a responsive UI?—to conceptual—how can we bridge the modeling gap between database representation to the end user's business concepts—to representational—how can we visualize and represent information that is inherently complex in a way that is both consumable and accurate, without misleading the user unintentionally? The net result is a focus on user experiences that are powerfully intuitive while also being delightful, and a focus on methods to make very complex analytics usable, comprehensible, and consumable by the ordinary business user.

1.2 THE ROLE OF RESEARCH AND THIS BOOK

The business intelligence industry is going through a time of incredible change. Predictive analytics, high volume data, unstructured data, social data, mobile, consumable analytics, and data visualization are all examples of demands and capabilities that have become critical within just the past few years, and growing at an unprecedented pace.

This book introduces research problems and solutions on various aspects central to BI. The target audience of the book includes first-year graduate students and senior undergraduate students in Computer Science, and researchers and practitioners in either Computer Science or Business Administration.

Chapter 2 provides an industry perspective on how BI has evolved. It first describes the systems in early days of BI and classic BI. It then discusses how game-changing trends have drastically reshaped the landscape of BI, with a projection into the future generation of BI tools. This sets up the rest of the book regarding ongoing research and development of BI tools.

One of the game changers is the shift toward the consumerization of BI tools. As a result, for BI tools to be successfully used by business users (rather than IT departments), the tools need to speak the language of the business user. However, a key problem with many existing BI tools is

that they speak more an IT language than a business language. In other words, a data model is not what is useful to a user; rather, it is a business model that is needed. Chapter 3 surveys four different types of business modeling.

While the needs of a user are expressed in a business model, the data that can meet the needs are captured within a data model. There is a significant gap between the two models. Chapter 4 on vivification addresses the problem of bridging the gap between the two models. It discusses the development of mappings that connect the business schema with the database schema, and outlines various strategies for dealing with incompleteness and uncertainty that arise from the bridging process.

The trend toward bigger data forces our systems to integrate and consolidate multiple, and often wildly different, data sources. As a result, it is often the case that a business query requires data to be retrieved and integrated from multiple sources. Chapter 5 describes some of the challenges that need to be overcome, including schema and semantic heterogeneity and ontology integration. It gives an overview of several integration architectures for dealing with these challenges and for efficient query answering.

While the book so far focuses on the usual structured relational data, the remaining chapters of the book turn to unstructured data, an ever-increasing and important component of the bigger data trend. Chapter 6 on information extraction describes methods for dealing with the extraction of relations from free text, which may be embedded in web pages, emails, surveys, customer call records, etc.

Finally, addressing the demanding expectations of the new users of our BI solutions requires innovation and research in new ways of interacting with users to give them an interactive discovery and guided analysis process, of visualizing new and complex types of information in a way that is consumable, compelling, and delightful, but also accurate. The last chapter of the book presents tools for visualizing text in data, as well as general visualization tools for BI users.

The topics selected for this book are aligned with the research done by collaborators within the pan-Canada Business Intelligence Network funded by the Natural Sciences and Engineering Research Council of Canada. To create an innovation platform for pre-competitive BI research within Canada, the network aims to enhance Canadian business competitiveness through the development of intelligent data management and decision-making solutions. The authors of this book are all network participants.

Note that the network includes many research projects not covered by the chapters in this book. Two notable omissions are data cleansing and cloud computing. Different data sources, particularly social data, also imply different levels of trust than the traditional "clean" data found in a data warehouse. The question of how to cleanse data from these new and complex data sources is an important research direction. Within the Synthesis Lectures on Data Management series, the reader is referred to the lecture by Bertossi [2011], which addresses some of the data cleansing issues encountered in BI systems. Within the same series, the lecture by Deutch and Milo [2012] addresses modeling and querying of business processes beyond the discussion in Chapter 3. And the lecture

by Carenini et al. [2011] considers information extraction beyond what is discussed in Chapter 6. Last but not least, the article by Armbrust et al. [2010] provides a comprehensive overview on cloud computing. We refer the interested reader to the aforementioned papers for more details on these topics.

CHAPTER 2

BI Game Changers: an Industry Viewpoint

Rock Leung, Chahab Nastar, Frederic Vanborre, Christophe Favart, Gregor Hackenbroich, Philip Taylor, and David Trastour

2.1 INTRODUCTION

To compete in today's markets, business users need to effectively use a large volume of data to make strategic and efficient decisions to help a business meet its goals. With the decreasing price of computer data storage, businesses are collecting and storing more business data at a greater detail. However, the increasingly large amounts of data are becoming more difficult to access and analyze, in part because the data are often stored in a variety of data formats in a variety of storage systems. Thus, despite the investments in storing business data, a recent survey by BusinessWeek [Hammond, 2004] found that a majority of the business users today are still going on "gut feel," and not utilizing the data from these systems to make effective business decisions. Researchers of this survey found that only one fifth of respondents say they always have the right amount of information to make an informed business decision, and over three quarters were aware of situations where managers made bad business decisions because these managers did not have sufficient information.

BI technologies, especially Business Analytics (more detailed definitions to be given later), are designed to empower business users to more efficiently make sense of vast amounts of data and make better decisions. Businesses are increasingly investing in BI software and the effective use of BI is seen as a competitive advantage. Business Analytics software is forecasted to grow faster than the overall enterprise software market [Vesset et al., 2010].

Although past and current BI solutions have helped business users, these are evolving rapidly to meet new business needs and incorporate recent technological advances. These business and technological trends present many opportunities for next generation BI technologies to better support business users. Pursuing these opportunities also raises many new research questions about how to make effective use of these new technologies. For example, how can BI systems make better use of artificial intelligence, social networks, or mobile computing devices to better support the business user?

In this chapter we discuss recent changes in Business Intelligence from our viewpoint as industrial researchers. Our team, the Business Intelligence Practice, is a group in SAP Research

that focuses on exploring the use of new technologies in next generation BI systems. As part of SAP, a market leader in enterprise application software, we are exposed to BI market trends, as well as the needs of large and small businesses. Further, we continually collaborate with academia with expertise in a variety of research areas such as text analytics, predictive analytics, visual analytics, semantic mashup, enterprise search, and collaborative decision making.

We begin this chapter by defining BI in context of business data and business users' actions. We then describe past and current BI systems and how they support the business user. The section that follows discusses many new business needs and technology trends that have contributed to the evolution of the BI. We then present our vision of the next generation of BI systems. Although we refer to businesses throughout the chapter, many of our claims also generalize to other organizations including non-profit organizations.

Through this chapter, we hope to make two contributions. First, we summarize how BI systems have evolved and share our vision on what the next generation of BI will look like. Second, we hope to help academic and industry researchers position their research work in the new view of BI and focus on research questions that are the most relevant to today's BI challenges.

2.2 DEFINING BUSINESS INTELLIGENCE

While there are varying definitions for *BI*, Forrester defines it broadly as a "set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information [that] allows business users to make informed business decisions with real-time data that can put a company ahead of its competitors" [Evelson, 2008]. In other words, the high-level goal of BI is to help a business user turn business-related data into actionable knowledge. BI traditionally focused on reports, dashboards, and answering predefined questions [Beller and Barnett, 2009]. Today BI also includes a focus on deeper, exploratory, and interactive analyses of the data using *Business Analytics* such as data mining, predictive analytics, statistical analysis, and natural language processing solutions [Evelson, 2008].

Consider the following cycle shown in Figure 2.1 involving a business user and her business's data. Starting from the *data* circle (Figure 2.1, bottom circle), the user accesses enterprise data such as financial transactions (e.g., "$19.99," "2," "21432"), which is generated and stored by the business. To begin making sense of this data, semantics are added to turn the data into more useful *information* (e.g., "2 shirts (product ID 21432) were bought at $19.99 each"). To analyze this information, a business user (e.g., store manager, another analyst) can choose, or be presented with, information that is relevant, trustworthy, and suitably presented for her purposes, in order to generate higher-level *knowledge* (e.g., whether her past actions/strategy helped her meet a monthly revenue goal).

The business user then interprets the knowledge gained from the data to determine what to do next. Specifically, the user has goals associated with her job (e.g., monthly revenue target), and one or more strategies to meet those goals (e.g., increase sales volume for popular products). The user can determine from the data whether the goals are being met and whether strategies are working. Continuing her data analysis and then weighing various options, the user *decides* what actions that

she would like to take (e.g., offer discounts, advertise product). The *actions* within the user's means are executed (e.g., discount product ID 21432), which the user hopes will have a positive impact on business but will need to confirm later by analyzing future data.

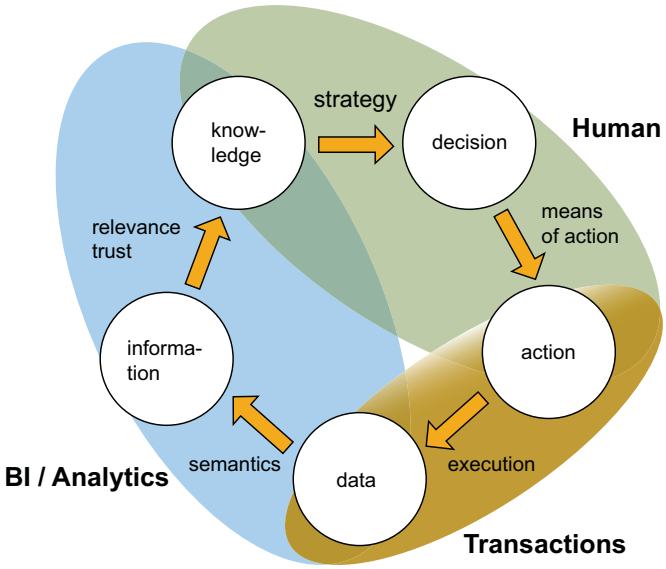


Figure 2.1: The virtuous cycle of business data

In the cycle of business data, BI (Figure 2.1, blue oval) is used to transform data to information to knowledge that the business user can then act on. In other words, BI takes a wide variety of high-dimensional, low-semantics data and refines the data into low-dimensional knowledge with high-semantics (i.e., fewer but more useful dimensions).

The development of BI systems taps several research and development areas. BI draws from work in databases to ensure, for example, that large volumes of business data are easily accessible, have minimal errors, and can be combined with different sources. BI also draws from work in data mining, text analysis, semantic analysis, and many other research areas that transform data into information. Further, BI draws from areas such as human-computer interaction, information visualization, and other areas that help the business user analyze, explore, and create knowledge from the information derived from business data. BI technologies also draw from work in networking and computer architectures.

2.3 EARLY DAYS OF BI

In the early days of BI, business data were stored in traditional databases, as shown in Figure 2.2. Data consisted of operational systems data and Enterprise Resource Planning (ERP) data. Accessing the data and processing it to a more consumable form required the IT and analytical skills of an IT

specialist. Thus, the business user needed to go through an IT specialist to access and analyze the business data. The turnaround between posing a business question to getting an answer often took weeks.

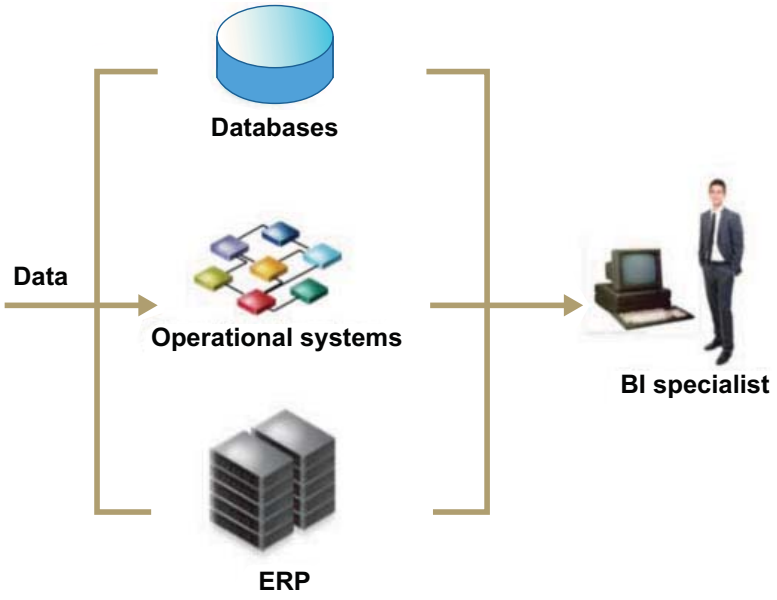


Figure 2.2: BI systems in the "early days."

2.4 CLASSIC BI

In the early 1990s, BI systems evolved into what we call Classic BI by adding layers of data "staging" to increase the accessibility of the business data to business users. Data from the operational systems and ERP were extracted, transformed into a more consumable form (e.g., column names labeled for human rather than computer consumption, errors corrected, duplication eliminated). Data from a warehouse were then loaded into OLAP cubes, as well as data marts stored in data warehouses. OLAP cubes facilitated the analysis of data over several dimensions. Data marts present a subset of the data in the warehouse, tailored to a specific line of business. Using classic BI, the business user, with the help of an IT specialist who had set up the system for her, could now more easily access and analyze the data through a BI system.

2.5 GAME-CHANGING TRENDS

A number of game-changing trends have recently emerged that we believe will significantly transform how BI is used, and affect the way product developers and researchers need to look at BI. We believe

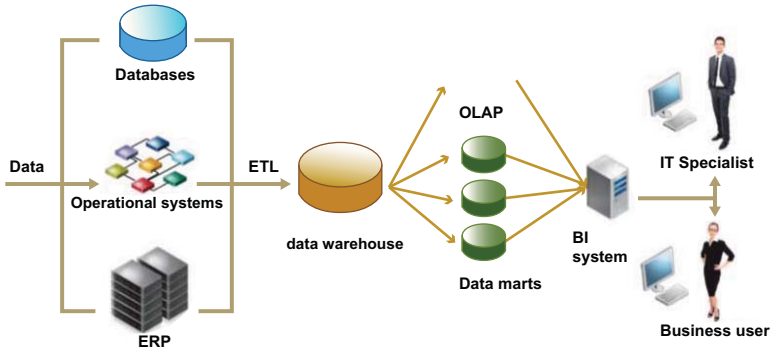


Figure 2.3: Classic BI system.

that BI, particularly Business Analytics, is at a tipping point in terms of its complexity, sophistication, and ease-of-use. These trends not only require new and advanced BI tools, but also raise new exciting research questions for industry and academic researchers to tackle.

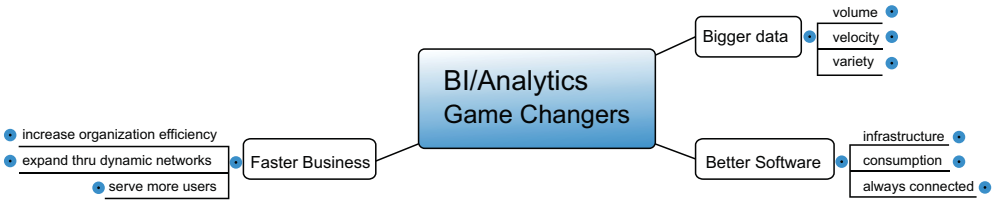


Figure 2.4: Game changing trends in Business Intelligence

2.5.1 FASTER BUSINESS

To stay competitive, businesses need to be as efficient as possible, be more innovative through dynamic networks, and serve more users.

Increasing Organization Efficiency

Businesses are empowering their employees to help them more independently make better and faster decisions. Businesses are working toward giving their employees access to the business data they need, when they need it, to help them perform their job effectively. This applies to employees at all levels, from top executives to those directly supporting customers. While business users often work at a desktop computer, there is also a need to support these users when they are in other settings such as meeting rooms, their commutes, and at the customer site. Thus, BI systems that support collaboration and mobile computing are needed. In addition, having business users access

and analyze business data by themselves, without help from IT or others helps to increase efficiency. Thus, self-service BI and other work systems can also increase the efficiency of the business.

Businesses are also delegating more tasks to computers, freeing their employees to focus on other work. For example, approximately a third of all stock trade volume on the New York Stock Exchange are performed by machines [EDN, 2012]. There is a trend in BI, which is common in other domains, to have technology (e.g., software agents) take actions automatically in "normal" cases and only involve humans in edge cases or exceptional situations.

Further, businesses are adopting new work processes to empower their employees and improve responsiveness and organizational efficiency. For example, many businesses are helping their employees better align their goals with those of the company through new performance management tools such as Key Performance Indicators and Management by Objectives. Businesses are also adopting agile product development processes (e.g., scrum, lean) in order to help employees work more efficiently together.

Innovation through Dynamic Networks

Businesses need to innovate to compete, but they often cannot rely on internal research and development to sustain the pace of innovation desired by businesses. We are seeing an increased desire by businesses to innovate with other organizations, as well as end users, through dynamic networks. Many businesses are using an open innovation model to work with other companies and academic researchers to generate and productize ideas [Chesbrough et al., 2006]. The NSERC BI Network exemplifies this open innovation model.

New business models are also emerging that require more powerful business analytics for guidance. Many businesses, particularly technology-producing ones, are adopting models like "freemium" and ad-powered models. Others are focused on selling to the "Long Tail," building a large number of products or services, each being bought at relatively small quantities but collectively total a large quantity. Thus, many businesses have a relatively larger and more diverse customer (and potential customer) base that they need to analyze and track. Understanding, serving, and selling to this customer base requires more powerful business analytics.

Serving More Users

BI systems are increasingly being used by end users and operational business users, and not only analysts. Thus, the next generation of BI systems needs to be designed for people who have less experience with analytical tools and less training on these systems.

Given the increasing number of networked computational devices (e.g., laptops, smart phones, tablets) it is increasingly more feasible for a business to reach more users. SAP, for example, wants to increase the number of SAP software users from millions of users to 1 billion users by 2015. Reaching more users requires better understanding a variety of target users, their needs, and how they use their devices.

2.5.2 BIGGER DATA

The cost of acquiring and storing data has declined significantly and thus businesses increasingly want to analyze more data (i.e., Big Data) to remain competitive. Big Data has often been characterized by increased volume, velocity, and variety [Russom, 2011]. We discuss each of these three dimensions below.

Volume

Like most digital data, the volume of business data is increasing over time. Businesses want to capture data in greater detail, in order to uncover more insights. As businesses rely more on computers for conducting business, data are being generated by more computer users, such as business employees and customers. Data are also generated by the increasing number of powerful mobile computing devices (e.g., smart phones, tablets), and connected sensors (e.g., Internet of Things). The amount of world's digital information data doubles every 18 month. What tools are needed to help business users manage and analyze huge volumes of data?

Velocity

Business data are also being collected at a greater rate. Business users also want to reduce the time it takes to answer a question, ideally in real time. In the early days of BI, analyzing enterprise data for a particular business question often took several days. In those days, reporting and analysis was often done in batches. Technology has enabled businesses to reduce the time it takes to access and process their data. Current BI systems are allowing businesses to analyze their data in "near real-time." Future BI systems will one day support real-time analysis of business data, and analysis of data streams. What tools are needed to help business users process increasingly greater rates of business data generation?

Variety

A business's enterprise data are traditionally structured, trusted, internal, and based on objective facts. However, businesses would now like to analyze a wider variety of data to discover additional insights. The variety of data that businesses are interested in analyzing can be categorized according to the following dimensions:

- *Structured/unstructured*: Enterprise data include data that are generally structured in a set of well-defined fields, and is typically stored in tables. Enterprise data can also consist of unstructured data (e.g., content of a text document, video content) that often have to be mined or modeled in order to extract meaningful information.

- *Trusted/non-certified*: Data can come from a variety of sources. Data can come from a trusted source (e.g., generated within the business). Enterprise data are generally trusted by the business, especially after the data are cleaned (e.g., data from a data warehouse). Data can also come from an external source and its accuracy, completeness, objectivity may not be certain. Online

data (e.g., databases available for free on the Internet) are generally considered non-certified, at least initially. A set of data can move along this trustedness dimension; for example, the data can be more trustworthy if the data are consistent with other trusted data or if other data from the same source are found to be trustworthy.

- *Internal/external*: Data can be generated within an organization or externally. Businesses have traditionally focused on analyzing internally generated business data. However, enterprise processes are now distributed and much of the business data will be collected outside the company's walls. For example, businesses often rely on a supply chain consisting of partners and suppliers, and an ecosystem of sellers. Some businesses use GPS truck monitoring to evaluate the provisioning chain and optimize production and/or product delivery. In addition, businesses are increasingly interested in analyzing online social networks and product review websites, generally externally hosted, to better understand consumer behaviours and market trends.

- *Facts/opinions*: Business data are generally facts (e.g., financial transactions, number of hours worked), but business data can also consist of opinions (e.g., employee satisfaction, customer ratings on a product).

Analyzing a wider variety of data requires new tools and techniques to combine, as well as differentiate, different types of data. How can tools allow users to easily analyze different types of data together? How can these tools also help users differentiate between the two types of data when analyzing them together?

2.5.3 BETTER SOFTWARE

Major technological advances in computer software and hardware have also provided opportunities to meet many needs related to faster businesses and processing Big Data. There are three advances that are particularly relevant for BI: infrastructure, data and software consumption, and increased connectedness.

Infrastructure

BI systems are incorporating cloud technologies, which are changing the way BI is deployed. Enterprise technology has historically been deployed "on premise" at the customer's site. However, BI technologies are now being offered as a service through the cloud, providing increased scalability, connectedness, and ease of deployment. Cloud technologies provide businesses more agility and flexibility in their IT systems to scale on demand. These technologies are also always on, which is crucial for online and mobile services to end users. Further, cloud technologies enable customers to use new BI systems sooner, without the need for upfront deployments of servers and software.

Advances in database technology have increased the business user's ability to quickly analyze data. In-memory technologies allow entire databases to reside in a server's memory instead of relatively slower disk storage, speeding up database accesses by orders of magnitude [Plattner, 2009,

Plattner and Zeier, 2011]. This technology has been found to be well suited for structured data and real-time processing. What other data processing and analysis can in-memory technologies help speed up?

Advances in distributed computing have increased a business's ability to store and process very large volumes of data. For example, the Apache Hadoop software framework, consisting of the Hadoop distributed File System and MapReduce distributed computing engine, can store and process petabytes of data [Apache, 2011]. Hadoop has been found to be suited for large volumes of data, unstructured data, and batch processing.

Consumption

Enterprise software has traditionally been data-centric, but this software is moving toward being more user-centric. In fact, enterprise software is lagging behind consumer software on this front, and there's a clear need to "consumerize" enterprise software [Moore, 2011].

How can enterprise software be designed to be easier for users to consume? Business users are generally consumers of other technologies (e.g., smart phones, Internet search engines, social networks, and computer games) so enterprise software may benefit from incorporating the many new features and intuitive interaction methods used on those technologies. Researchers are also exploring how "gamifying" enterprise software can help ease consumption by adding game elements to motivate engagement and help users learn to perform initial tasks and then more advanced tasks [Burke and Hiltbrand, 2011]. Tailoring the user experience for each individual user or small demographics may also make the software easier to consume. Personalizing the user experience requires measuring and making effective use of contextual data (e.g., a user's location, organizational role, current task, previous tasks). Social networks also enable users to collaboratively consume data and make decisions as a group.

Businesses also need new tools to consume Big Data. For example, new analytical tools like Visual Analytics applications are needed to explore large volumes of data. As another example, new tools are needed to process the wide variety of data a business user is interested in. Structured data have traditionally been accessed by business users through database queries. In contrast, text and multimedia content have traditionally been accessed through searching on keywords. The frontiers between structured data and content are now quickly vanishing and the business user wants to search many different types of data through one information retrieval tool such as the familiar search input text field (e.g., Google Search, Bing). Moreover, business data are not only consumed by business users but by machines as well. How can we design BI tools to make use of machine-to-machine data?

Increased Connectedness

Business users are increasingly connected through technology. For example, the latest mobile phones enable users to remotely communicate and collaborate more with others. Business users are now able to access their business's data in order to make better decisions away from their desk (e.g., customer

site, manufacturing plant, while commuting). To make better use of BI through mobile devices, researchers are exploring how to personalize the data and user experience. How can we make use of the user's position in the company, access to data, task, location, and other contextual information to personalize their experience?

Many businesses are interested in making use of social networks. Social networks enable users to engage with other individuals, groups, and communities. Using social networks, businesses can tap into both the wisdom of the crowd, as well as the network of experts. Social network can also be used to evaluate partners, suppliers, products, potential recruits, and co-workers. How can social networks and online collaboration tools be used to support decision making? How can these technologies help business users analyze data?

More devices and sensors are being connected to the Internet (e.g., Internet of Things). The Internet of Things can be thought of as an automated part of collecting insights. Data from sensors such as RFID tags for goods, GPS tracking for trucks, gate control for building's entrance, and building's temperature sensors can be used to answer business questions and provide actionable insights for a whole company, a product team, or an individual employee.

2.6 NEXT-GENERATION BI

Given the many game changes in BI that were described in the earlier section, we predict that the next generation BI software will include evolutions in both hardware infrastructure and data processing. As stated earlier, new advanced infrastructures will change how much data is stored and how quickly we can access large amounts of it (see Figure 2.5). Instead of using traditional relational databases, a business can use distributed storage systems such as Hadoop to store data warehouses and large volumes of other data that the business wants to analyze. In-memory technologies (e.g., SAP HANA, IBM solidDB, Oracle TimesTen) will allow business users to execute data queries thousands of times faster than current generation systems, greatly increasing the volume of data that can be analyzed at once as well as the interactivity of the analytical tool. We envision the distributed storage/computing system integrating the data sources and feeding the in-memory technology. The analytics will access the in-memory or distributed computing system as needed.

Next-generation BI systems will also allow users to process greater varieties of data and produce better insights from it. As mentioned earlier, the goal of BI is to refine all kinds of high-dimensional, low-semantics data and to present low-dimensional knowledge with high-semantics to the end user. BI systems traditionally only enabled the analysis of structured business data (e.g., financial transaction) but new systems will analyze many more types of data. As shown in Figure 2.6, next-generation BI systems will also process multimedia content (e.g., video, image, sound), text content, data steams (e.g., RSS, logs, device sensor data, smart items), and graphs, often together with structured data.

These various types of data are stored in a knowledge base that pushes the data to the user and enables the business user to pull data from the knowledge base (see Figure 2.6). The data can be pushed to the user through personalized (i.e., contextual, recommended) dashboards and alerts. The

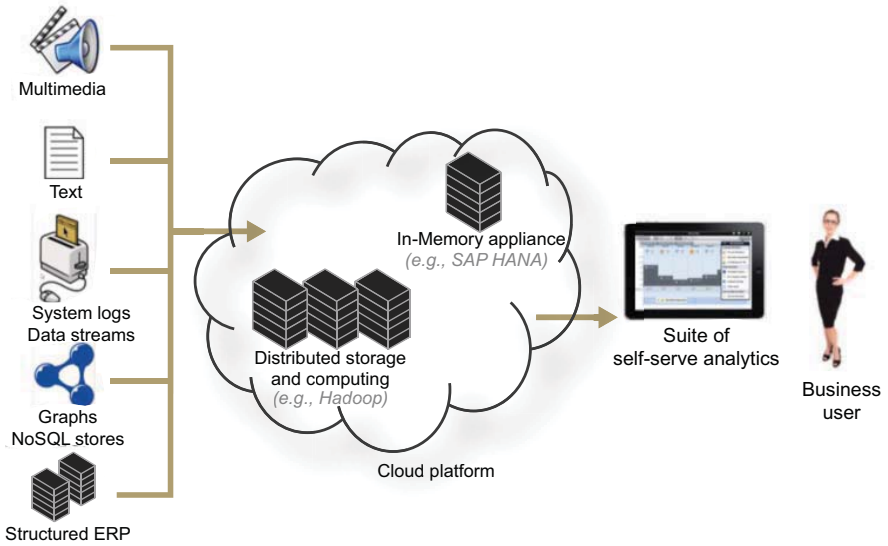


Figure 2.5: Our vision for the next-generation BI system.

user can also pull the data through reports, queries/searches, and data exploration. These systems will be designed to support self-service, enabling the business user to analyze data with little assistance from IT specialists (see Figure 2.5).

BI systems will also have a "control loop" in which data can be written-back to the data sources or knowledge database. This write-back can be done explicitly by the user, say to correct or annotate the data. The BI system can also record the user's behaviour or status (e.g., geo-location, query logs) in the data sources, to provide additional contextual data for future personalized presentation or data analysis. Furthermore, there may be a loop in which a machine makes decisions on the users' behalf (e.g., using business rules) and the business user is not directly involved.

2.7 CONCLUSIONS

Businesses are relying increasingly on Business Intelligence to remain competitive in their market. Past BI systems have allowed business users to access and analyze business data with the help of an IT specialist. Current BI systems reduce the dependency on an IT specialist, and help users make better sense of their data. We listed a number of new business needs and technology trends that both require, and help to develop, next-generation BI systems. New BI systems will allow business users to analyze a larger volume, increased velocity, and wider variety of data, with minimal involvement from IT.

This is an exciting time for BI research, as there are many opportunities to develop more powerful and easy-to-use analytical tools for business users. We have listed a few of the many

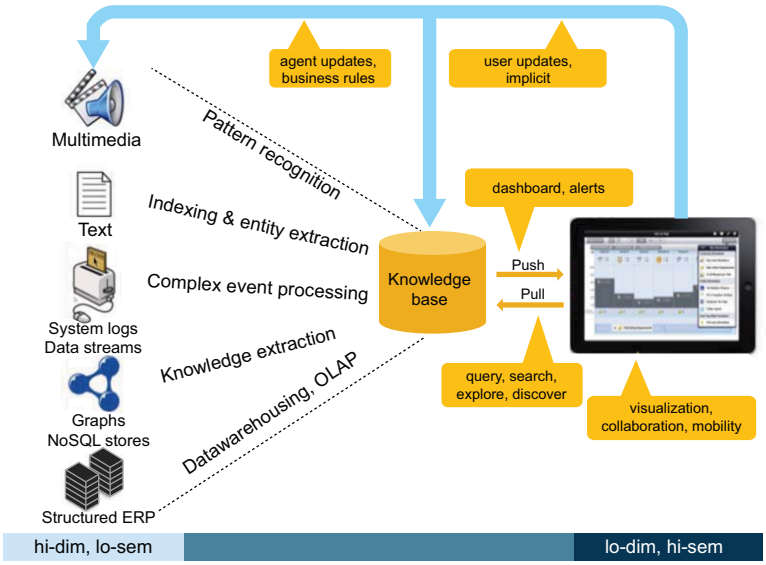


Figure 2.6: Processing new types of data.

research questions that need to be explored to continue advancing BI and realize the next-generation BI systems. Next-generation BI will enable businesses and other organizations to gain more insights from their data and make better decisions.

Authors' Biographies

RAYMOND T. NG

Dr. Raymond T. Ng is a Professor of Computer Science at the University of British Columbia. He received a Ph.D. in Computer Science from the University of Maryland in 1992. His main research area for the past two decades is on data mining, with a specific focus on health informatics and text mining. He has published over 150 peer-reviewed publications on data clustering, outlier detection, OLAP processing, health informatics, and text mining. He is the recipient of two best paper awards, from the 2001 ACM SIGKDD conference, which is the premier data mining conference worldwide, and the 2005 ACM SIGMOD conference, which is one of the top database conferences worldwide. He was a program co-chair of the 2009 International Conference on Data Engineering, and a program co-chair of the 2002 ACM SIGKDD conference. He was also one of the general co-chairs of the 2008 ACM SIGMOD conference. He was an editorial board member of the Very Large Database Journal and the IEEE Transactions on Knowledge and Data Engineering until 2008.

PATRICIA C. AROCENA

Patricia C. Arocena is a Research Assistant in Computer Science at the University of Toronto. She received her M.Eng. in Electrical and Computer Engineering in 2001 and her Ph.D. in Computer Science (expected 2013), both from the University of Toronto. Her research focuses on developing techniques to support efficient and practical use of schema mappings in information integration, and in particular, on embracing *incompleteness* in the context of data-driven decision making.

DENILSON BARBOSA

Denilson Barbosa is an Associate Professor of Computing Science at the University of Alberta. He obtained a Ph.D. in 2005 from the University of Toronto, working on Web data management. He received an IBM Faculty Award for his work on XML benchmarking, and an Alberta Ingenuity New Faculty Award for his work on extraction and integration of data from the Web. He received the Best Paper award at the 26th IEEE International Conference on Data Engineering (ICDE 2010). At the time of writing, he was a lead investigator on the NSERC Strategic Network on Business Intelligence, through which the SONEX system for large-scale relation extraction on the web is developed.

GIUSEPPE CARENINI

Giuseppe Carenini is an Associate Professor of Computer Science at the University of British Columbia. He is also an Associate member of the UBC Institute for Resources, Environment and Sustainability (IRES). Giuseppe has broad interdisciplinary interests. His work on natural language processing and information visualization to support decision making has been published in over 80 peer-reviewed papers. Dr. Carenini was the area chair for "Sentiment Analysis, Opinion Mining, and Text Classification" of ACL 2009 and the area chair for "Summarization and Generation" of NAACL 2012. He has recently co-edited an ACM-TIST Special Issue on "Intelligent Visual Interfaces for Text Analysis." In July 2011, he published a co-authored book on *Methods for Mining and Summarizing Text Conversations*. In his work, Dr. Carenini has also extensively collaborated with industrial partners, including Microsoft and IBM. Giuseppe was awarded a Google Research Award and an IBM CASCON Best Exhibit Award in 2007 and 2010 respectively.

LUIZ GOMES, JR.

Luiz Gomes, Jr., is currently a Ph.D. student at the University of Campinas. Prior to that he conducted graduate research at the University of Waterloo and spent several years gaining research-oriented experience in industry and academia. He has worked in such diverse and exciting areas as information extraction, data mining, data integration, and complex network analysis.

STEPHAN JOU

Stephan Jou is currently a Technical Architect, Research Staff Member, and Sr. Manager at IBM's Business Analytics Office of the CTO, and has over fifteen years of experience designing, building, and inventing software from inception to release, from a small start-up to one of the largest software development companies in the world. In his career at Cognos and IBM, he has architected and led the development and productization of over ten 1.0 Cognos and IBM products in the areas of cloud computing, mobile, visualization, semantic search, data mining, and neural networks. His current role at IBM focuses on translating academic and IBM research into product strategy for the Business Analytics division at IBM. Stephan holds a M.Sc. in Computational Neuroscience and Biomedical Engineering, and a dual B.Sc. in Computer Science and Human Physiology, all from the University of Toronto.

ROCK ANTHONY LEUNG

Rock Anthony Leung is a Senior Researcher at SAP and manages its Academic Research Center (ARC), which initiates and supports collaborative research projects with academia. Through ARC, Rock actively works with graduate students, professors, and SAP stakeholders to explore and validate novel solutions in business intelligence, visual analytics, and other research areas. Rock is also a Scientific Advisory Committee member of the NSERC Business Intelligence Network. Rock earned a Ph.D. in Computer Science from the University of British Columbia (UBC), specializing in Human-Computer Interaction research. His research work has been published in numerous prominent journals and conferences. He has also actively contributed to several professional development programs at UBC and has received awards for his service and leadership.

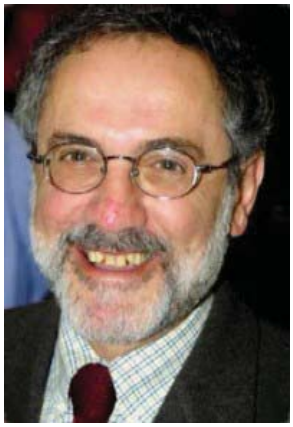
EVANGELOS MILIOS

Evangelos Milios received a diploma in Electrical Engineering from the NTUA, Athens, Greece, and Master's and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology. Since July of 1998 he has been with the Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, where he served as Director of the Graduate Program (1999-2002) and as Associate Dean–Research since 2008. He is a Senior Member of the IEEE. He was a member of the ACM Dissertation Award committee (1990-1992), a member of the AAAI/SIGART Doctoral Consortium Committee (1997-2001), and he is co-editor-in-chief of Computational Intelligence. At Dalhousie, he held a Killam Chair of Computer Science (2006-2011). He has published on the interpretation of visual and range signals for landmark-based navigation and map construction in robotics. He currently works on modeling and mining of content and link structure of Networked Information Spaces, text mining, and visual text analytics.

RENÉE J. MILLER

Renée J. Miller is Professor and the Bell Canada Chair of Information Systems at the University of Toronto. She received BS degrees in Mathematics and in Cognitive Science from the Massachusetts Institute of Technology. She received her MS and Ph.D. degrees in Computer Science from the University of Wisconsin in Madison. She received the US Presidential Early Career Award for Scientists and Engineers (PECASE), the highest honor bestowed by the United States government on outstanding scientists and engineers beginning their careers. She received the National Science Foundation Early Career Award, is a Fellow of the ACM, the President of the VLDB Endowment, and was the Program Chair for ACM SIGMOD 2011 in Athens, Greece. She and her IBM co-authors received the ICDT Test-of-Time Award for their influential 2003 paper establishing the foundations of data exchange. Her research interests are in the efficient, effective use of large volumes of complex, heterogeneous data. This interest spans data integration, data exchange, knowledge curation and data sharing. In 2011, she was elected to the Fellowship of the Royal Society of Canada (FRSC), Canada's national academy.

JOHN MYLOPOULOS



John Mylopoulos holds a distinguished professor position (chiara fama) at the University of Trento, and a professor emeritus position at the University of Toronto. He earned a Ph.D. degree from Princeton University in 1970 and joined the Department of Computer Science at the University of Toronto that year. His research interests include conceptual modeling, requirements engineering, data semantics, and knowledge management. Mylopoulos is a fellow of the Association for the Advancement of Artificial Intelligence (AAAI) and the Royal Society of Canada (Academy of Sciences). He has served as program/general chair of international conferences in Artificial Intelligence, Databases and Software Engineering, including IJCAI (1991), Requirements Engineering (1997), and VLDB (2004).

RACHEL A. POTTINGER

Rachel A. Pottinger is an associate professor in Computer Science at the University of British Columbia. She received her Ph.D. in computer science from the University of Washington in 2004. Her main research interest is data management, particularly semantic data integration, how to manage metadata (i.e., data about data), and how to manage data that is currently not well supported by databases.

FRANK TOMPA

Frank Tompa has been a faculty member in computer science at the University of Waterloo since 1974. His teaching and research interests are in the fields of data structures and databases, particularly the design of text management systems suitable for maintaining large reference texts and large, heterogeneous text collections. He has co-authored papers in the areas of database dependency theory, storage structure selection, query processing, materialized view maintenance, text matching, XML processing, structured text conversion, text classification, database integration, data retention and security, and business policy management. He has collaborated with several corporations, including Oxford University Press, Open Text, and IBM, and served as a member of the Scientific Advisory Committee for the Business Intelligence Strategic Network (BIN). In 2005, the University of Waterloo and the City of Waterloo announced the naming of the road *Frank Tompa Drive* in recognition of Professor Tompa being one of those who "epitomize the energy and enterprise that characterize the University of Waterloo." He was named a Fellow of the ACM in 2010 and awarded a Queen Elizabeth II Diamond Jubilee Medal in 2012, both for contributions in the area of text-dominated data management.

ERIC YU

Eric Yu is Professor at the Faculty of Information, University of Toronto, Canada. His research interests are in the areas of information systems modeling and design, requirements engineering, knowledge management, and software engineering. Books he has co-authored or co-edited include: *Social Modeling for Requirements Engineering* (MIT Press, 2011); *Conceptual Modeling: Foundations and Applications* (Springer, 2009); and *Non-Functional Requirements in Software Engineering* (Springer, 2000). He is an associate editor for the *Int. Journal of Information Systems Modeling and Design*, and serves on the editorial boards of the *Int. J. of Agent Oriented Software Engineering*, *IET Software,* and the *Journal of Data Semantics*. He received his Ph.D. in Computer Science from the University of Toronto.