Università degli Studi di Milano
Master Degree in Computer Science

# Information Management course

Teacher: Alberto Ceselli

Lecture 01 : 02/10/2014

# Practical informations:

- Teacher: Alberto Ceselli
  (alberto.ceselli@unimi.it)
- Course weekly schedule:
  - Thursday 11.00 – 13.00 (3 nord)
  - Thursday 14.00 – 17.00 (3 nord)
- Tutoring:
  - Anytime after the lectures
  - Contact me by email
  - whenever you can find me in my office
- Homepage:
  - homes.di.unimi.it/ceselli/IM

# Practical Informations (2)

- Reference book:
  - J. Han, M. Kamber (J. Pei),
    “Data Mining: concepts and Techniques”,
    2nd (3rd) edition, 2006 (2011)

- Exam:
  - Development of a project +
  - project discussion
  - (+ general check on theory)

# Why Information Management?

Let's start from a user perspective:

- I have a lot of data (queries on Google)
- I'm interested in making a decision about my future:
  - weather to follow or not the “Information Management”course
- Can I extract knowledge from data?

- http://www.google.com/trends/
  - DMBS
  - Data warehouse
  - Big data
  - Data analytics

And that's “only” statistical evaluation … we're interested in “analytics”!

# Advanced analysis example: disjunctive mapping


Instead of searching for rules like

A and B → C

search for rules like

A or B → C

# Evolution of Sciences: New Data Science Era

- Before 1600: **Empirical science**
- 1600-1950s: **Theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s: **Computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (i.e. empirical, theoretical, and computational branches)
  - Computational Science: what can I do if I am not able to find closed-form solutions for complex mathematical models?
- 1990-now: **Data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - Computing grids that make all these archives accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes
  - Data analytics is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

# A Brief History of Data Mining

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

# Why Information Management?

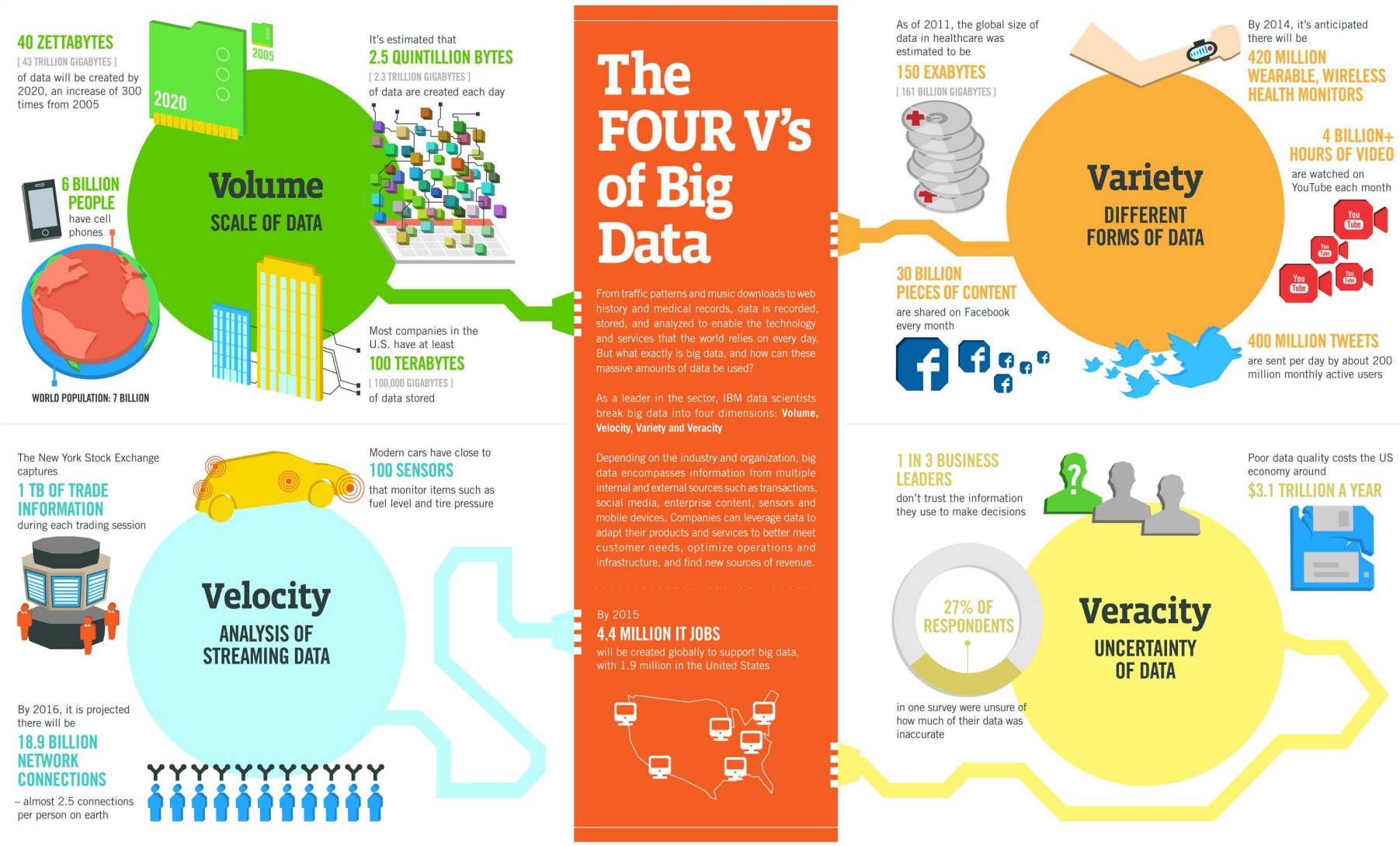
Two arising issues in data management:

Big Data

+

Analytics

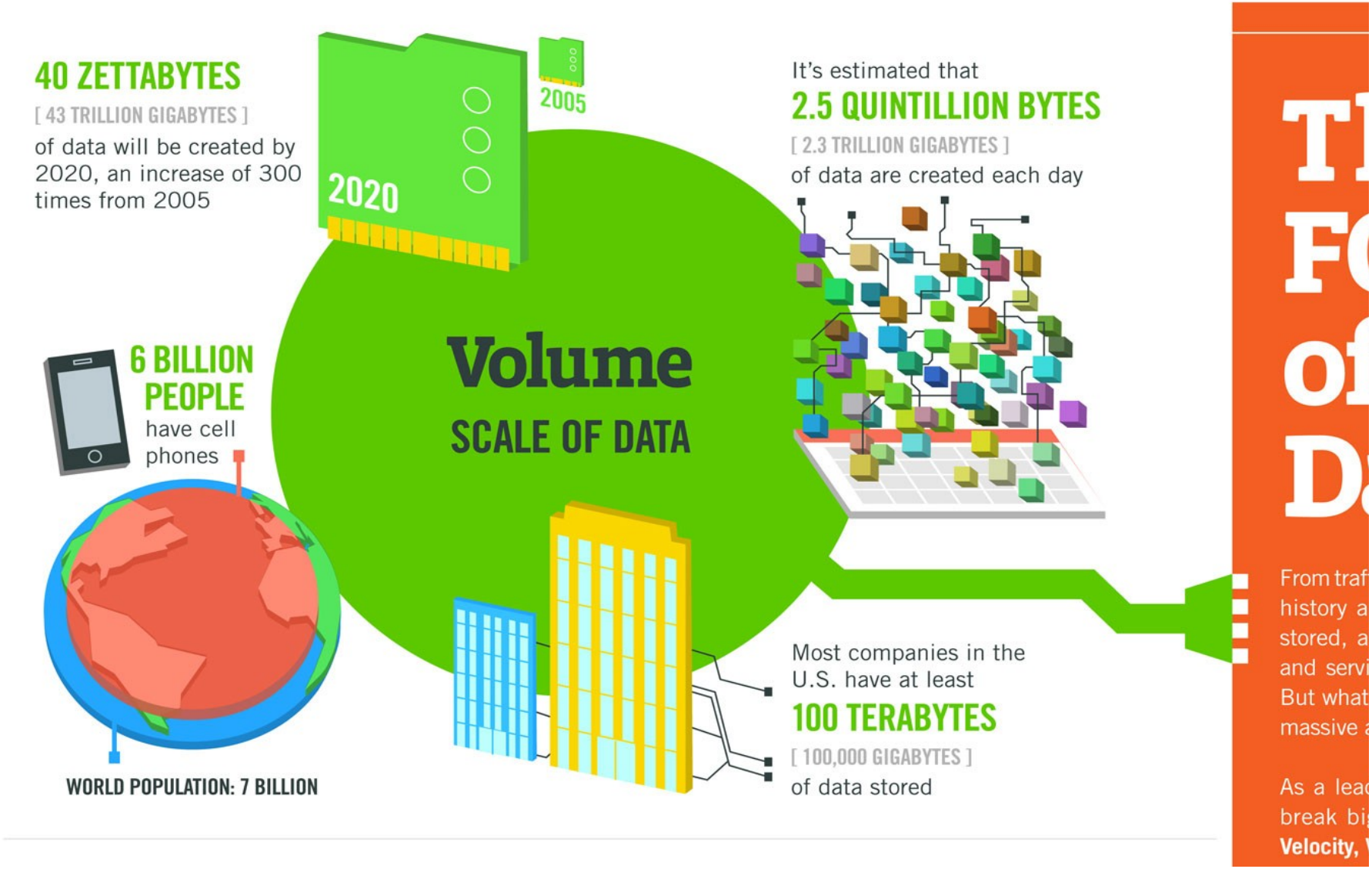
# Bottom Issue: Big Data




**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

# Big Data: Volume

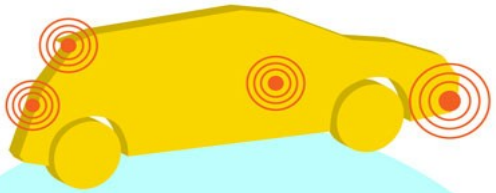


# Big Data: Velocity

The New York Stock Exchange captures

**1 TB OF TRADE INFORMATION**

during each trading session



Modern cars have close to

**100 SENSORS**

that monitor items such as fuel level and tire pressure

## Velocity

### ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

**18.9 BILLION NETWORK CONNECTIONS**

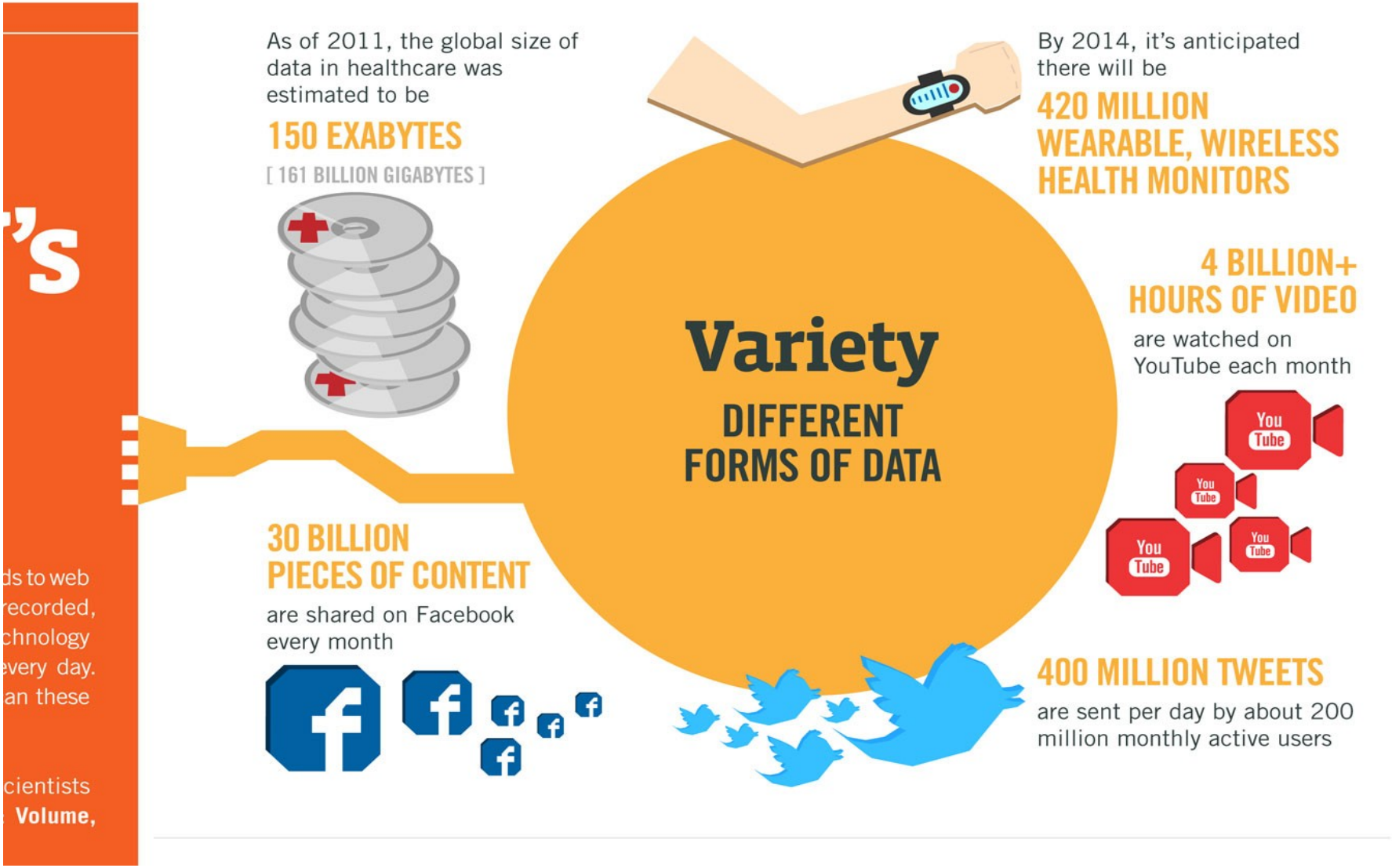
– almost 2.5 connections per person on earth

Dependi
data enc
internal a
social m
mobile d
adapt the
custome
infrastruc

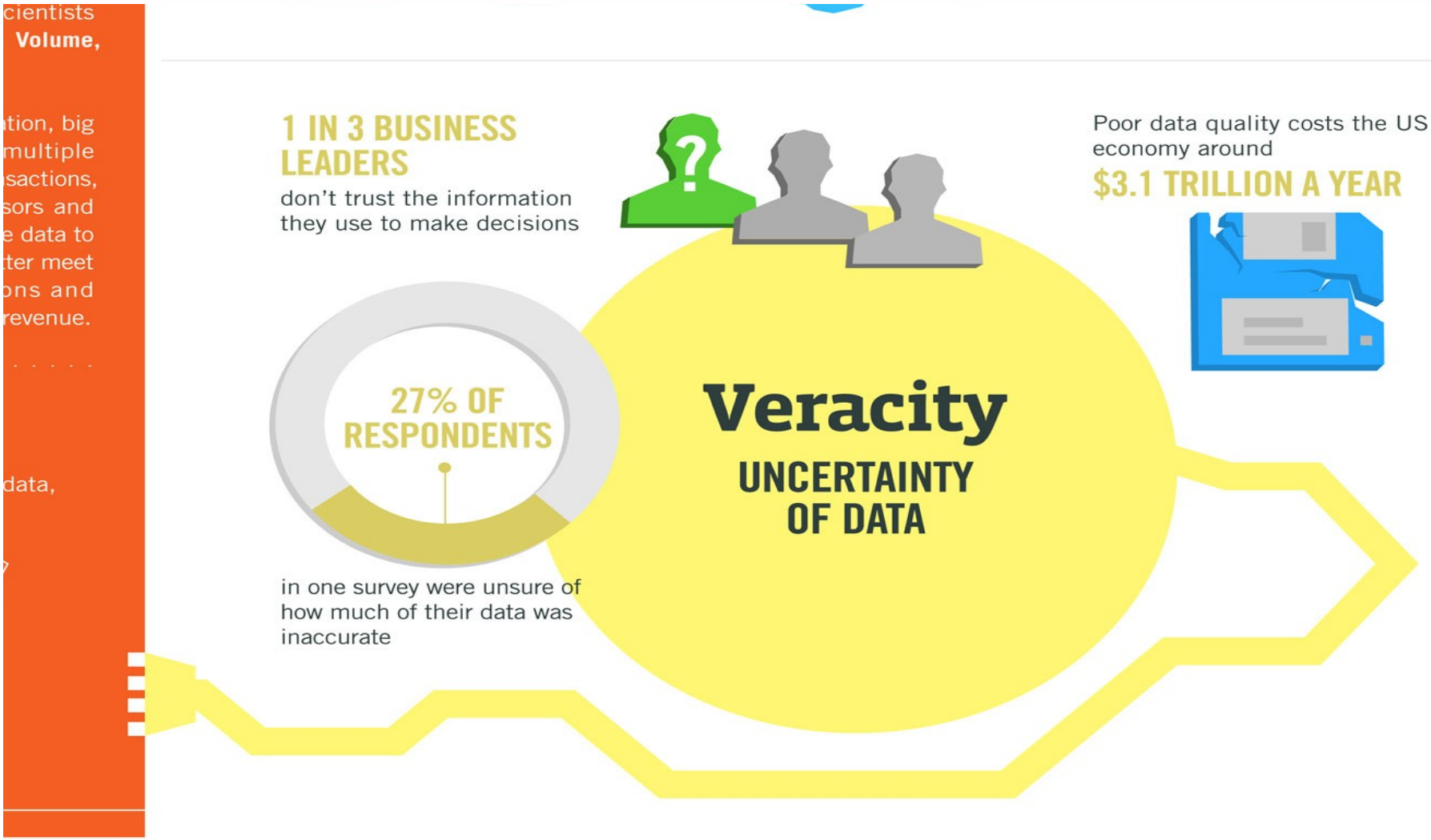
By 2015
**4.4 MI**
will be c
with 1.9

**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

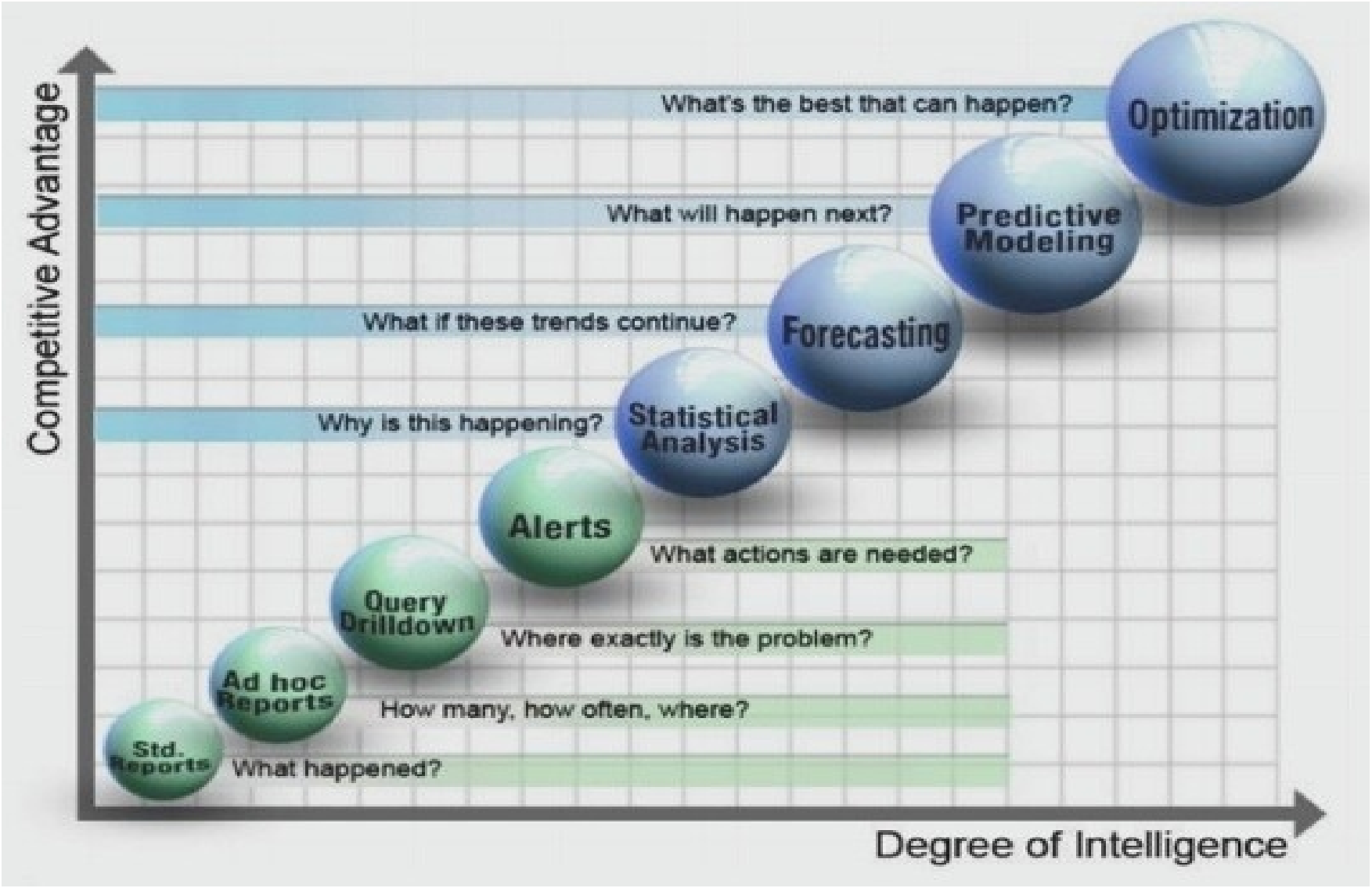
# Big Data: Variety



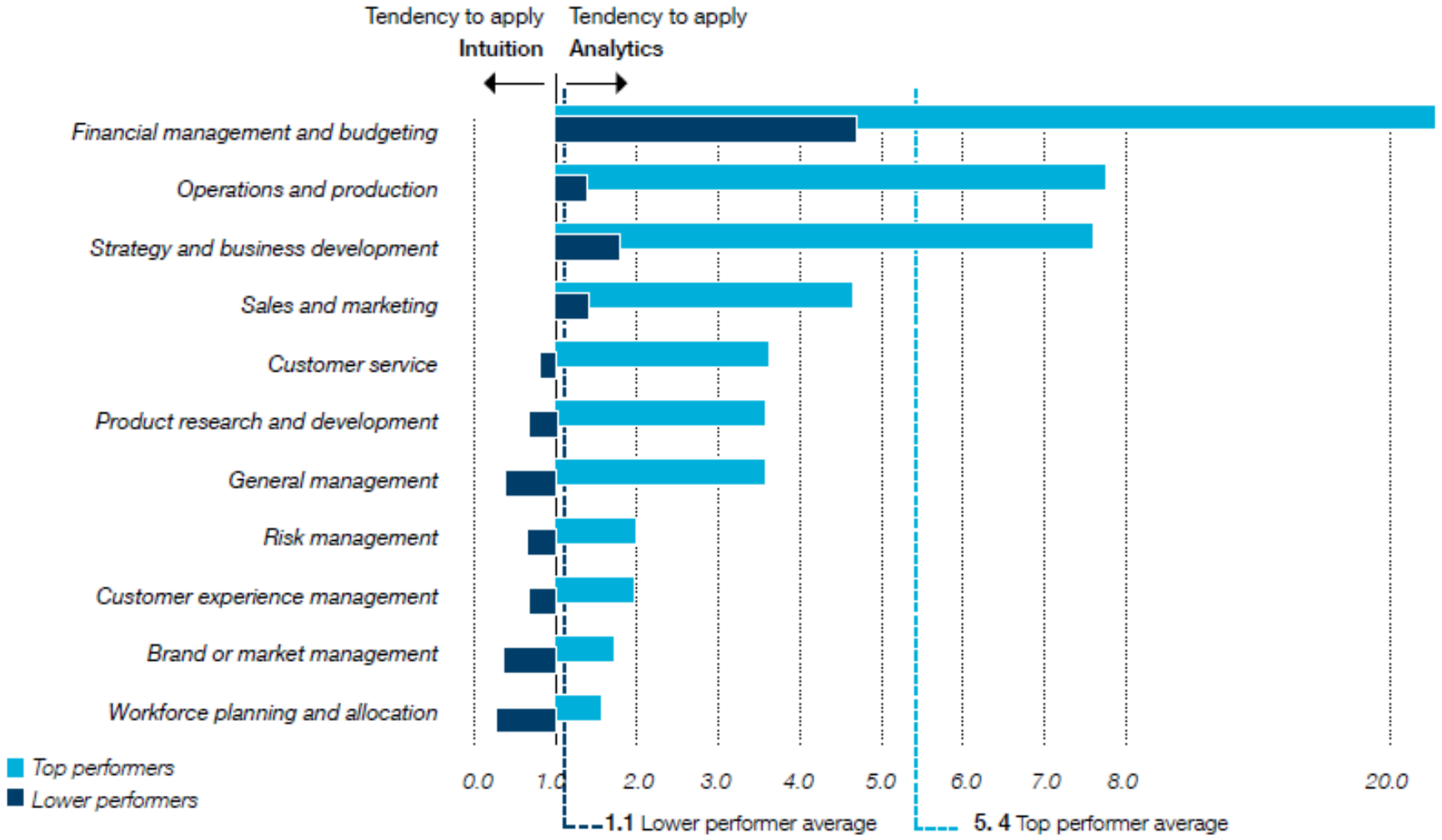
# Big Data: Veracity



# Top Issue: Analytics (SAS)



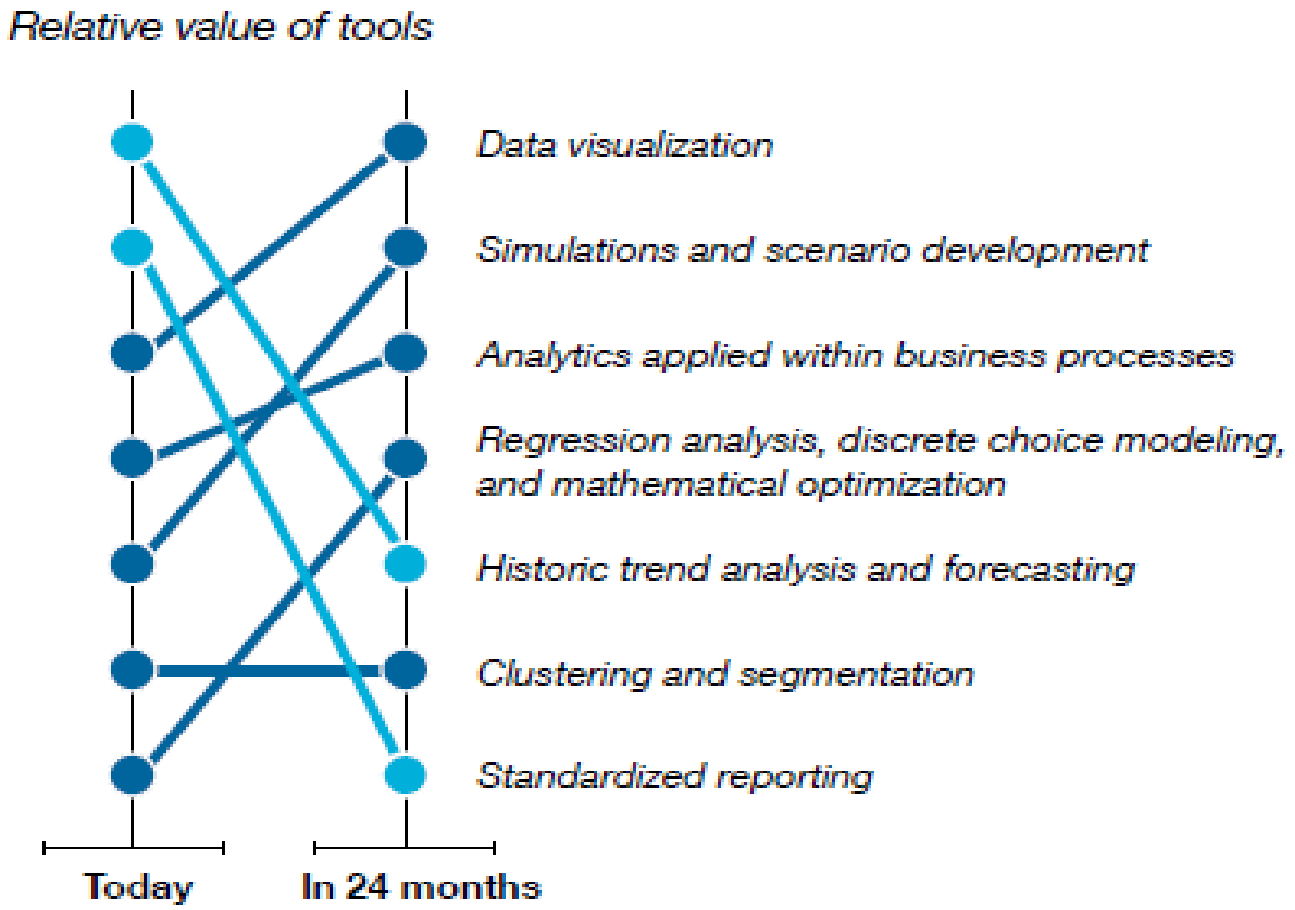
# Top Issue: Analytics (IBM)




*Note: Respondents were asked about their organization's application of analytics to the activities listed above. A score of 1.0 indicates an equal likelihood of applying either analytics or non-analytic methods, while a score of 0.0 indicates a tendency to use non-analytic methods.*

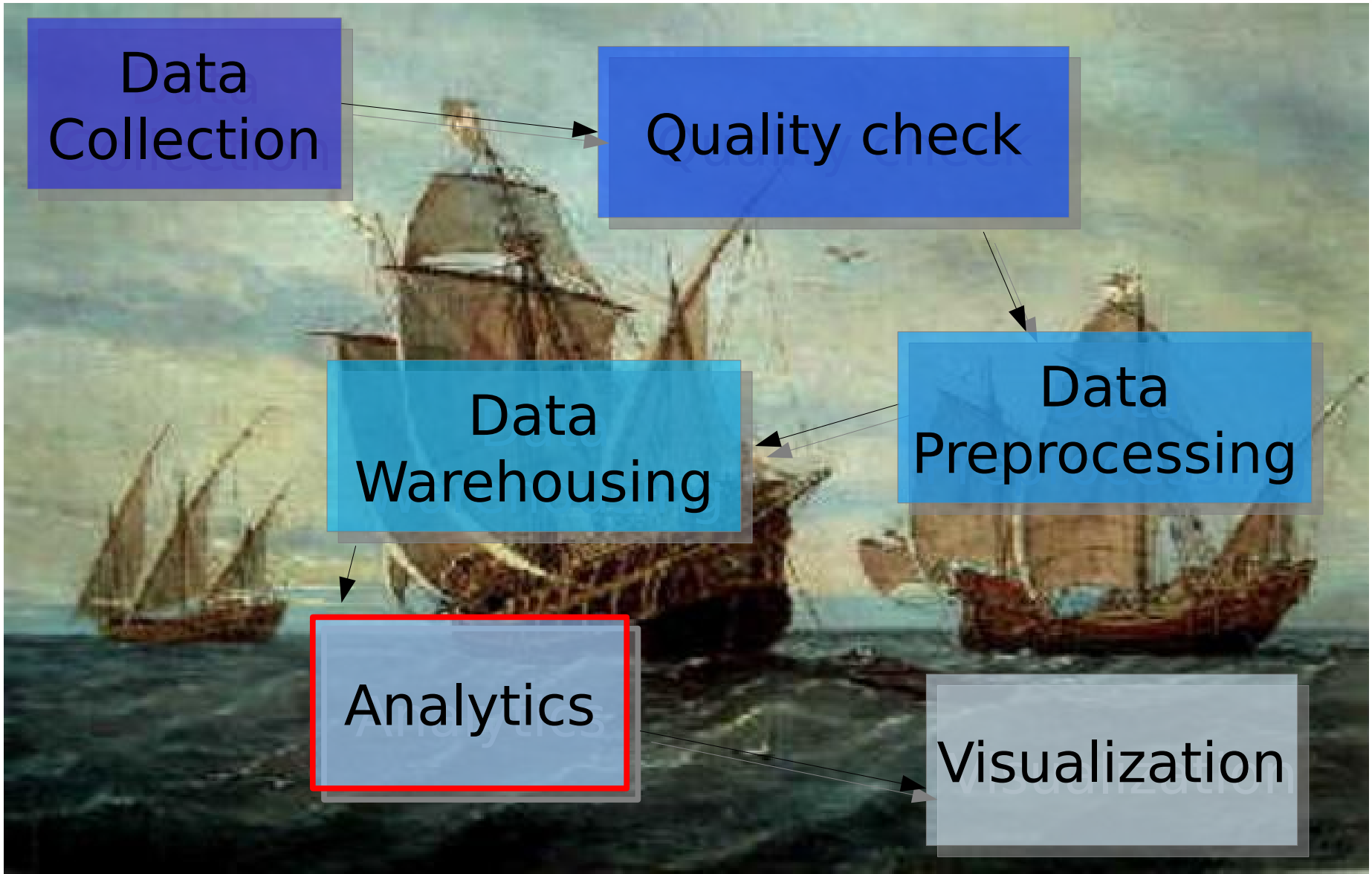
*Source: Analytics: The New Path to Value, a joint MIT Sloan Management Review and IBM Institute for Business Value study. Copyright © Massachusetts Institute of Technology 2010.*

# Top Issue: Analytics (IBM)




*Note: Respondents were asked to "Select the type of analytics creating the most value in your organization today, and which types you believe will create the greatest value 24 months from now? (Select up to three in each timeframe.)"*
*Source: Analytics: The New Path to Value, a joint MIT Sloan Management Review and IBM Institute for Business Value study. Copyright © Massachusetts Institute of Technology 2010.*

# The Data Journey



# The Data Journey

- A new professional profile: the data steward

# Information Management course syllabus

- Know your data: data objects and attribute types, basic statistical descriptions of data, measuring data proximity.
- Data preprocessing: the quality of data; data cleaning, integration; samples reduction.
- Dimensionality reduction: Principal Component Analysis; Feature selection algorithms
- On Line Analytical Processing and Data Warehousing
- Mining frequent patterns, ideas, algorithms and pattern evaluation measures
- Classification:
  - basic concepts and ideas;
  - decision tree induction models and algorithms.
  - Bayesian classification; Bayesian Belief Networks.
  - Support Vector Machines
- Clustering: partitioning, hierarchical and density-based methods; evaluation of clustering methods
- Time series analysis
- *(Data mining in networks & graphs / fraud detection)*

# Data Mining: Concepts and Techniques

## (3rd ed.)

## — Chapter 1 —

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign &
Simon Fraser University

# Chapter 1. Introduction

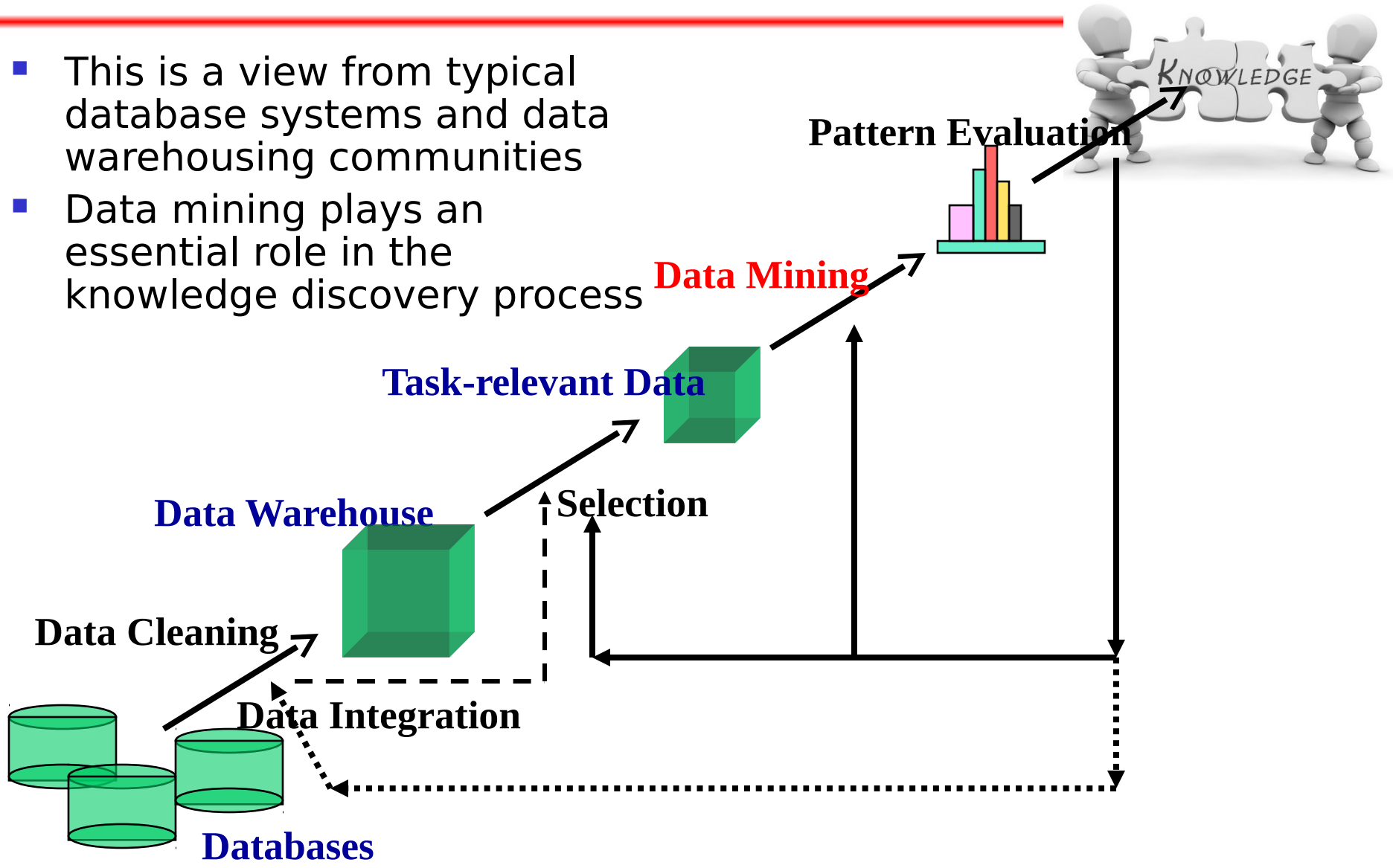
- Why Data Mining?
- What Is Data Mining? 
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

# What Is Data Mining?

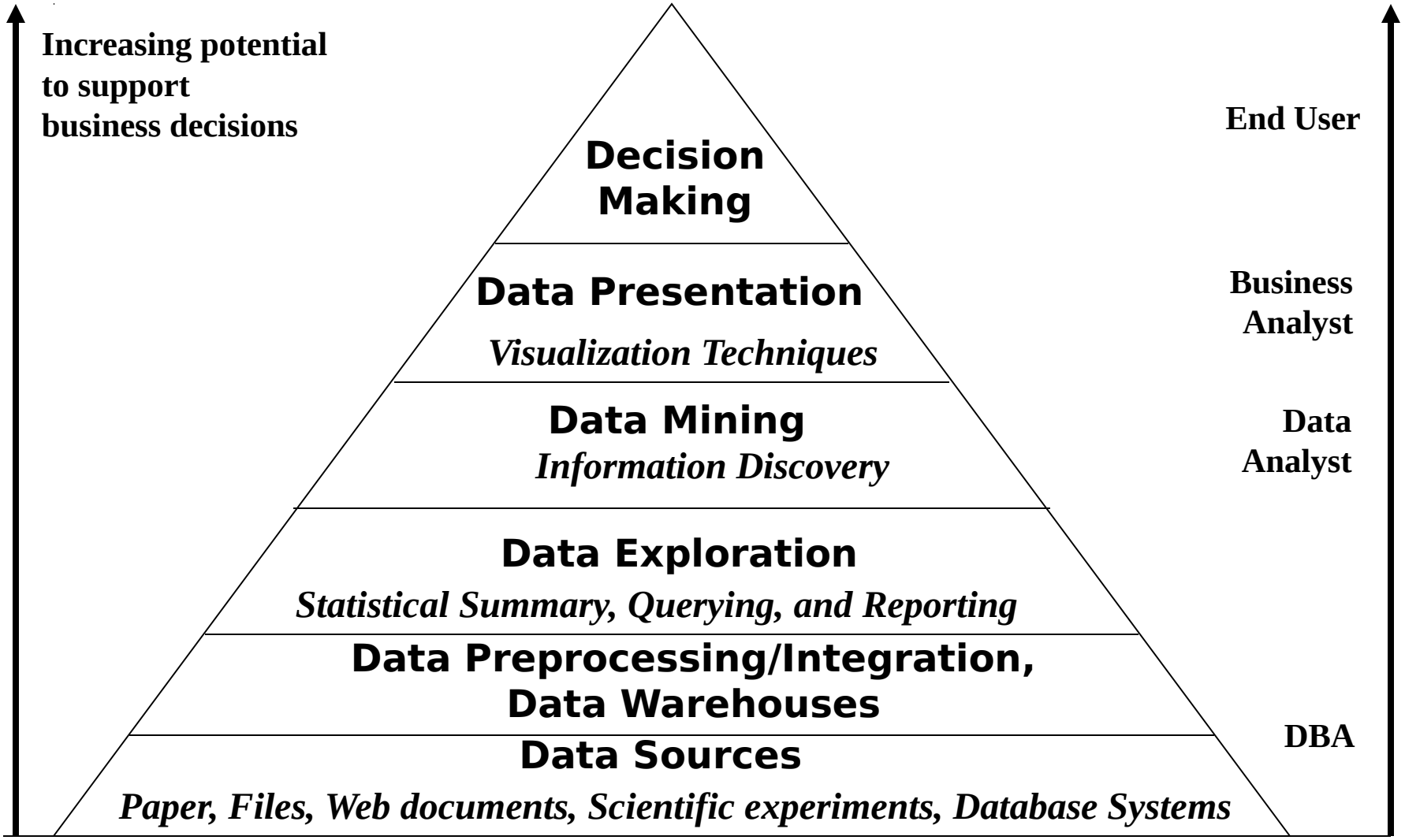
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, big data analytics, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

# Knowledge Discovery (KDD) Process

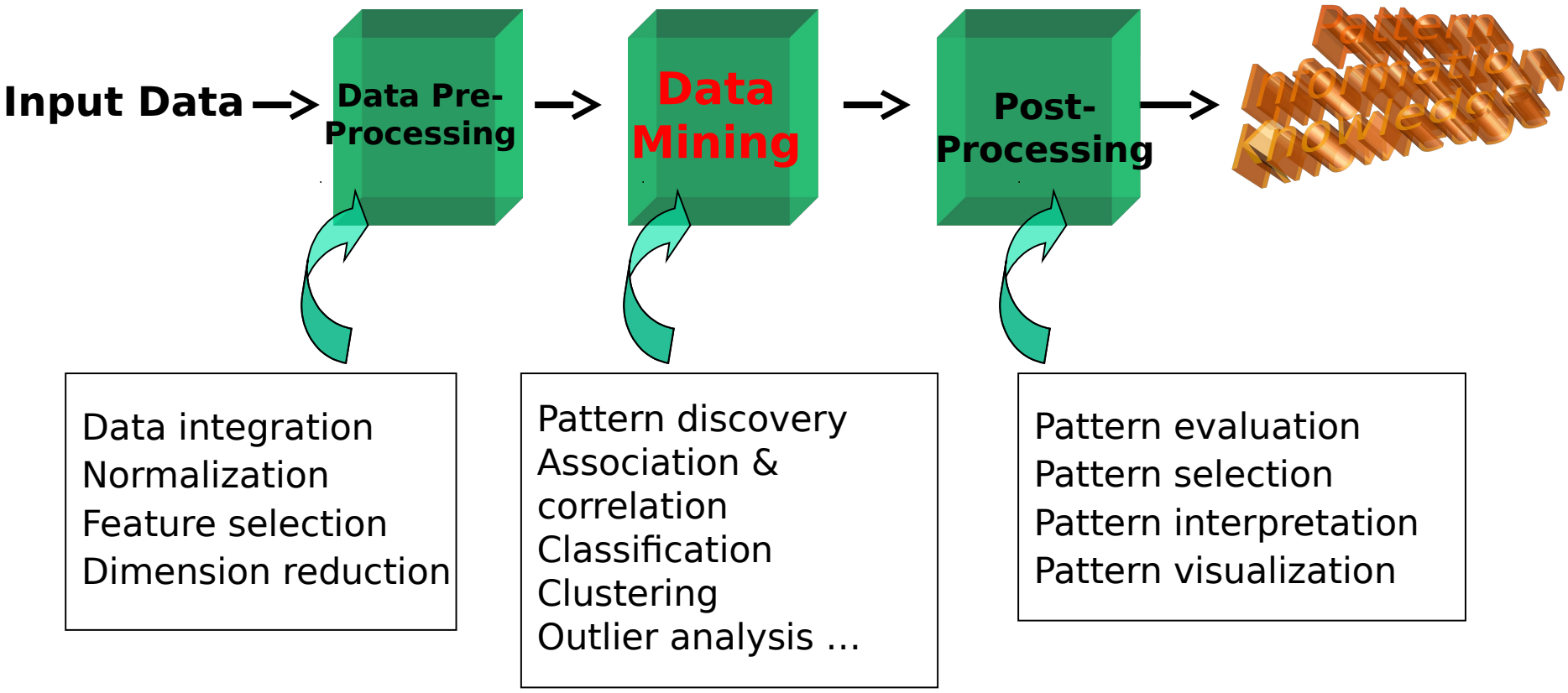
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



# Data Mining in Business Intelligence

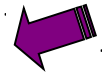


# Data Mining in Machine Learning and Statistics




- This is a view from typical machine learning and statistics communities

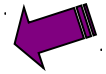
# Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining 
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

# Multi-Dimensional View of Data Mining

- **Data to be mined:** Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications:** Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

# Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined? 
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

# Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

# Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined? 
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

# Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%]  (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

# Classification

- Classification and label prediction
  - Construct models based on some training examples
  - Describe and distinguish classes or concepts for prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases,  web-pages, …

# Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

# Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

# Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

# Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks (Web community discovery, opinion mining, usage mining, …)

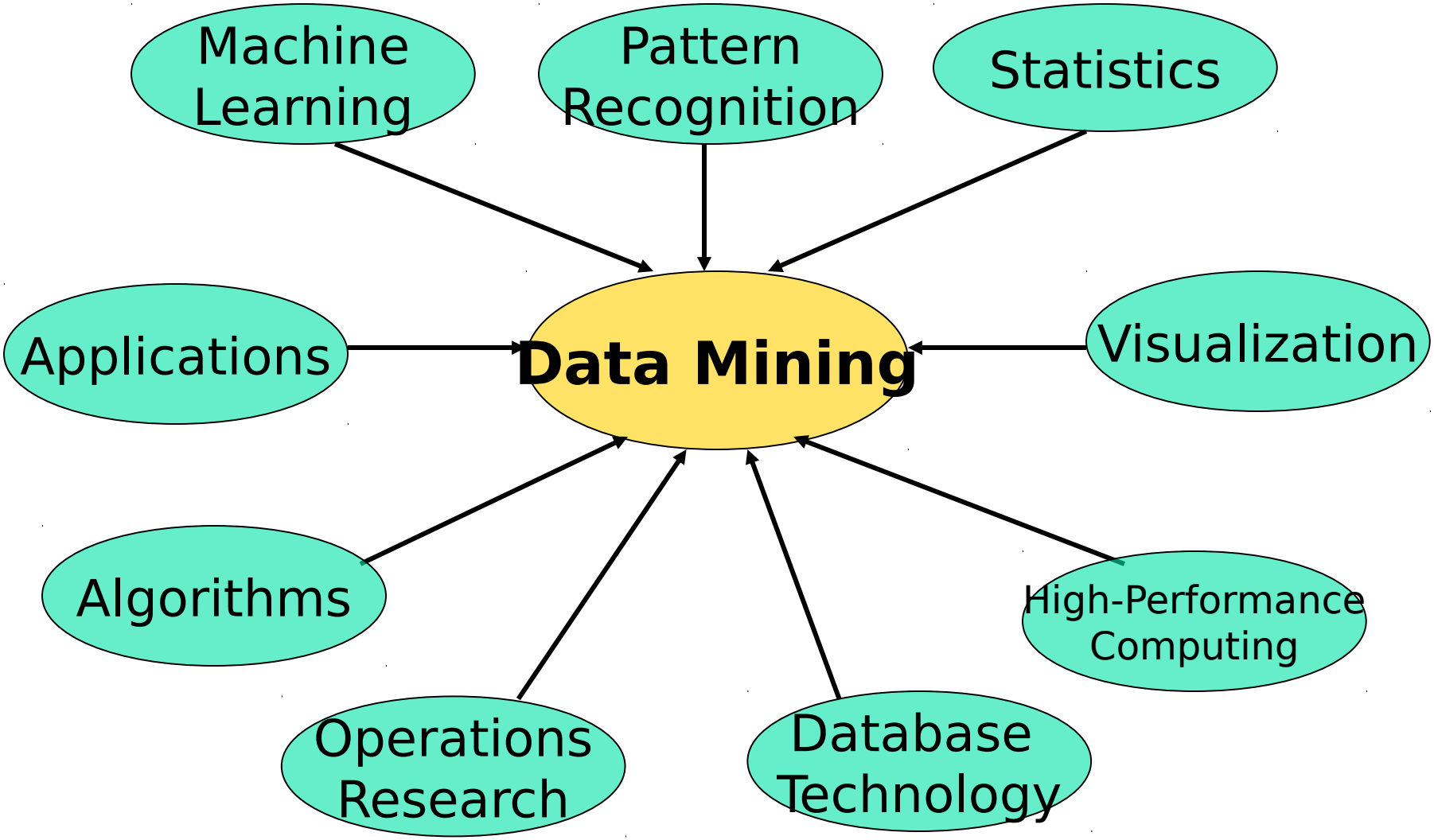
# Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of “patterns” and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness

# Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used? 
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

# Data Mining: Confluence of Multiple Disciplines



# Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

# Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

# Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

# Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining 
- Summary

# Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

# Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

# Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary 

# Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications
- Major issues in data mining