
Università degli Studi di Milano
Master Degree in Computer Science

Information Management course

Teacher: Alberto Ceselli

Lecture 01 : 06/10/2015

Practical informations:

- Teacher: Alberto Ceselli
(alberto.ceselli@unimi.it)
- Course weekly schedule:
 - Tuesday 11.00 - 13.00 (3 nord)
 - Wednesday 11.00 - 13.00 (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

$A \text{ and } B \rightarrow C$

search for rules like

$A \text{ or } B \rightarrow C$

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

Why Information Management?

A set of consolidated (still improving) techniques in data management:

Data Mining

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

©2012 Han, Kamber & Pei. All rights reserved.

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



Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
 - What items are frequently purchased together?
- Association, correlation vs. causality
 - A typical association rule
 - Diaper \rightarrow 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, ...)

Why Information Management?

Two arising issues in data management:

Big Data
+
Analytics

Bottom Issue: Big Data

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005



It's estimated that **2.5 QUINTILLION BYTES**
[2.3 TRILLION GIGABYTES]
of data are created each day



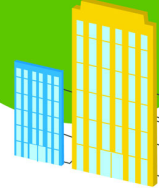
Volume SCALE OF DATA

6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES**
[100,000 GIGABYTES]
of data stored



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

Variety DIFFERENT FORMS OF DATA

30 BILLION PIECES OF CONTENT are shared on Facebook every month



4 BILLION+ HOURS OF VIDEO are watched on YouTube each month



400 MILLION TWEETS are sent per day by about 200 million monthly active users



The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



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

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



Poor data quality costs the US economy around **\$3.1 TRILLION A YEAR**



Veracity UNCERTAINTY OF DATA

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

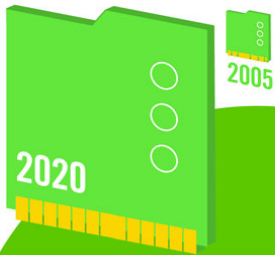
in one survey were unsure of how much of their data was inaccurate

Big Data: Volume

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

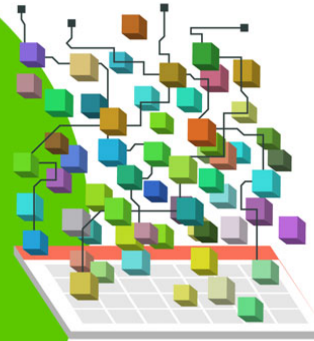


It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



Volume
SCALE OF DATA



6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION



Most companies in the U.S. have at least

100 TERABYTES

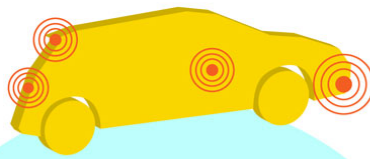
[100,000 GIGABYTES]

of data stored

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

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

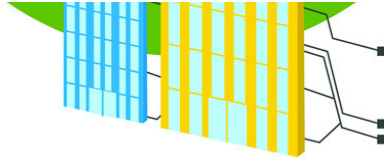
As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and

Big Data: Velocity

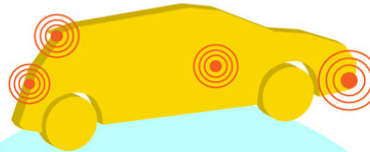


WORLD POPULATION: 7 BILLION



Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



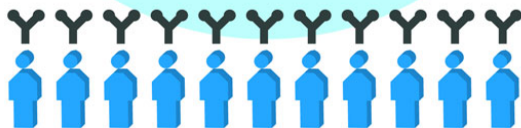
Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

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



and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States



Big Data: Variety

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



Variety
DIFFERENT
FORMS OF DATA

**30 BILLION
PIECES OF CONTENT**

are shared on Facebook every month



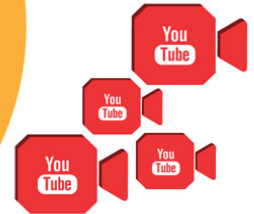
By 2014, it's anticipated there will be

**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**



**4 BILLION+
HOURS OF VIDEO**

are watched on YouTube each month



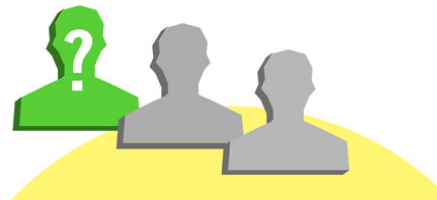
400 MILLION TWEETS

are sent per day by about 200 million monthly active users



**1 IN 3 BUSINESS
LEADERS**

don't trust the information they use to make decisions



Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



Big Data: Veracity

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States



FIELDS OF CONTENT

are shared on Facebook every month



400 MILLION TWEETS

are sent per day by about 200 million monthly active users

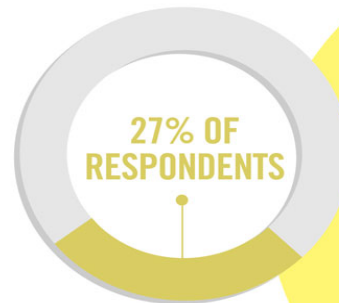
1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around

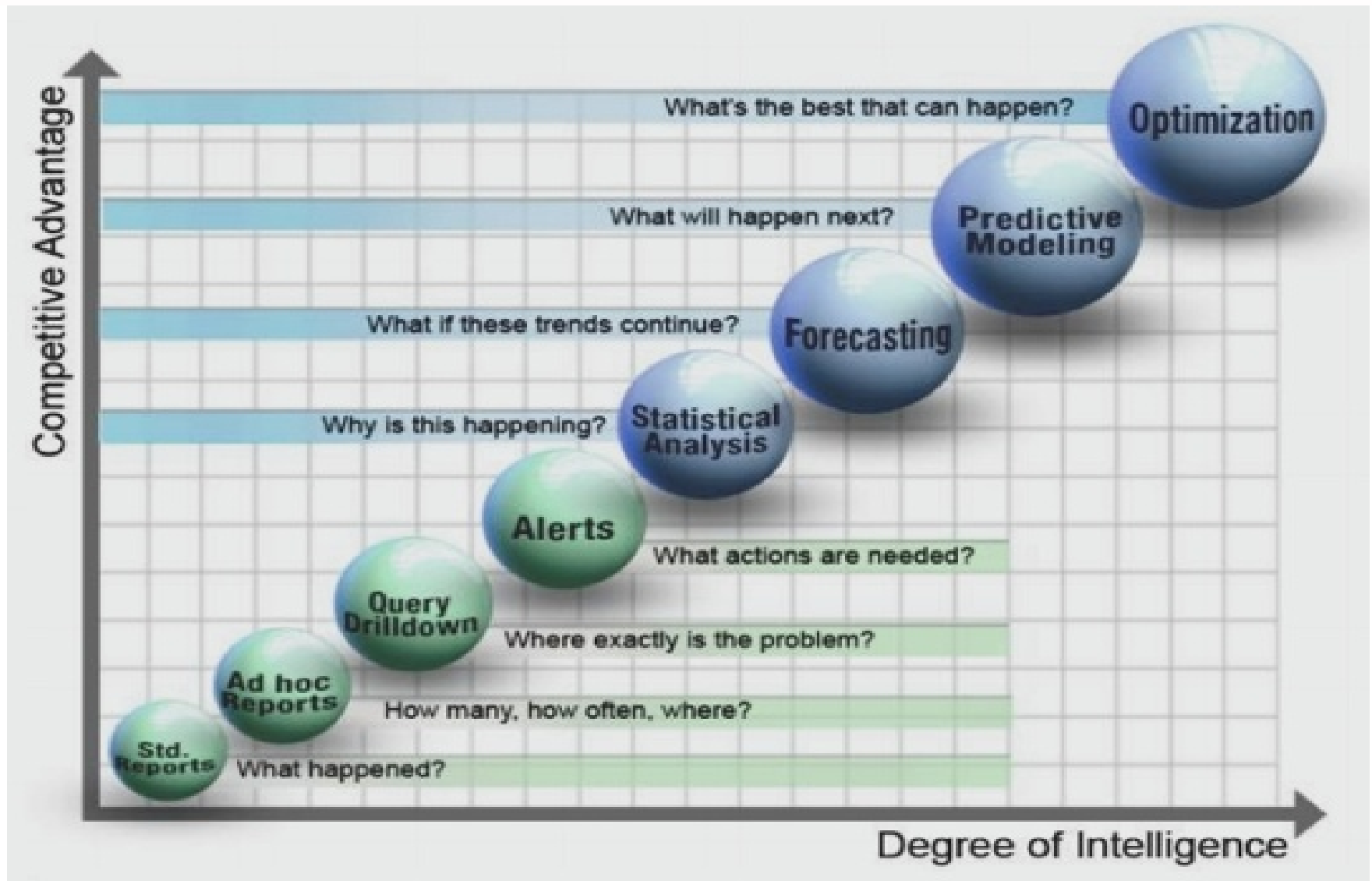
\$3.1 TRILLION A YEAR



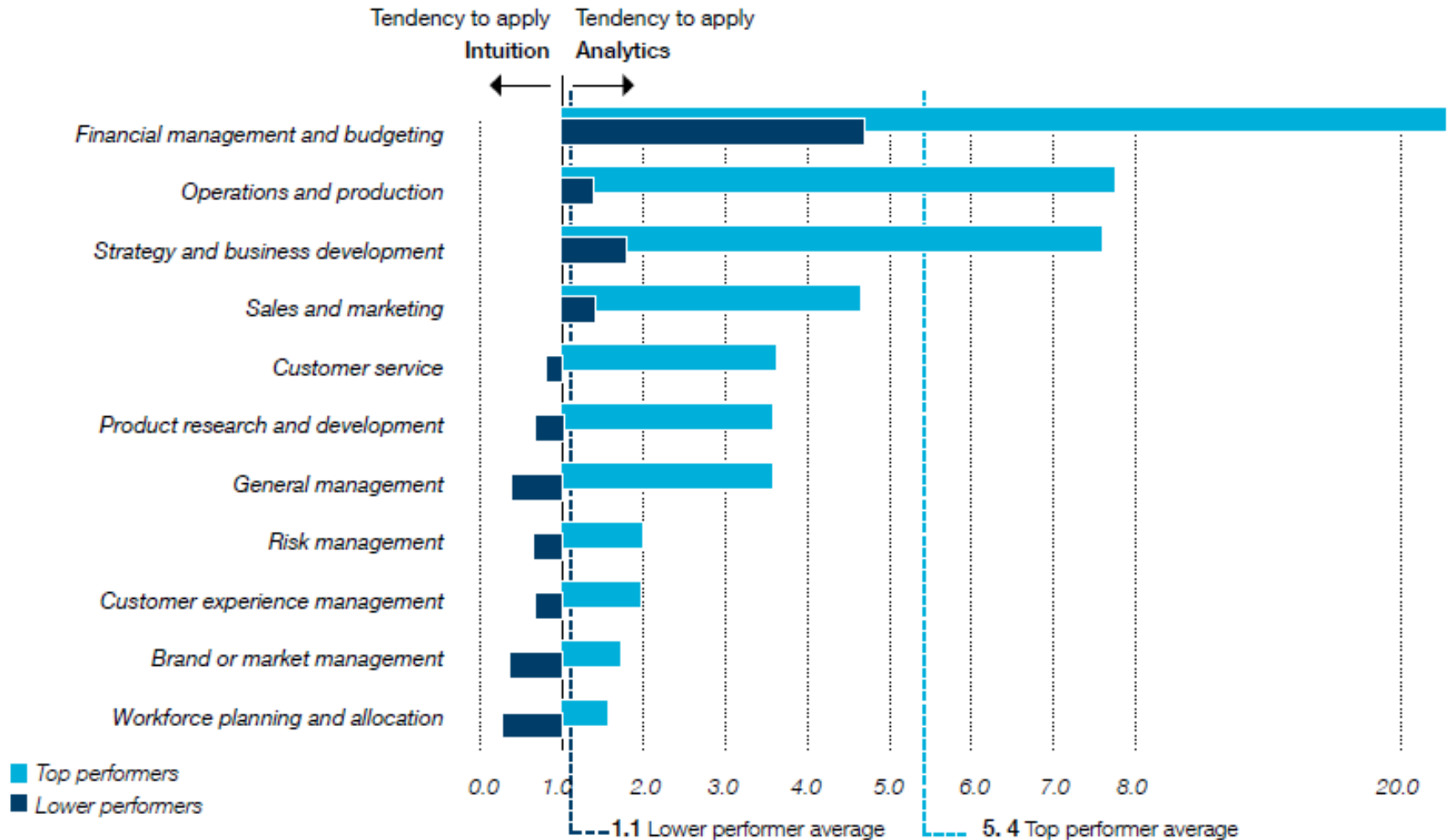
in one survey were unsure of how much of their data was inaccurate

Veracity
UNCERTAINTY OF DATA

Top Issue: Analytics (SAS)



Top Issue: Analytics (IBM)

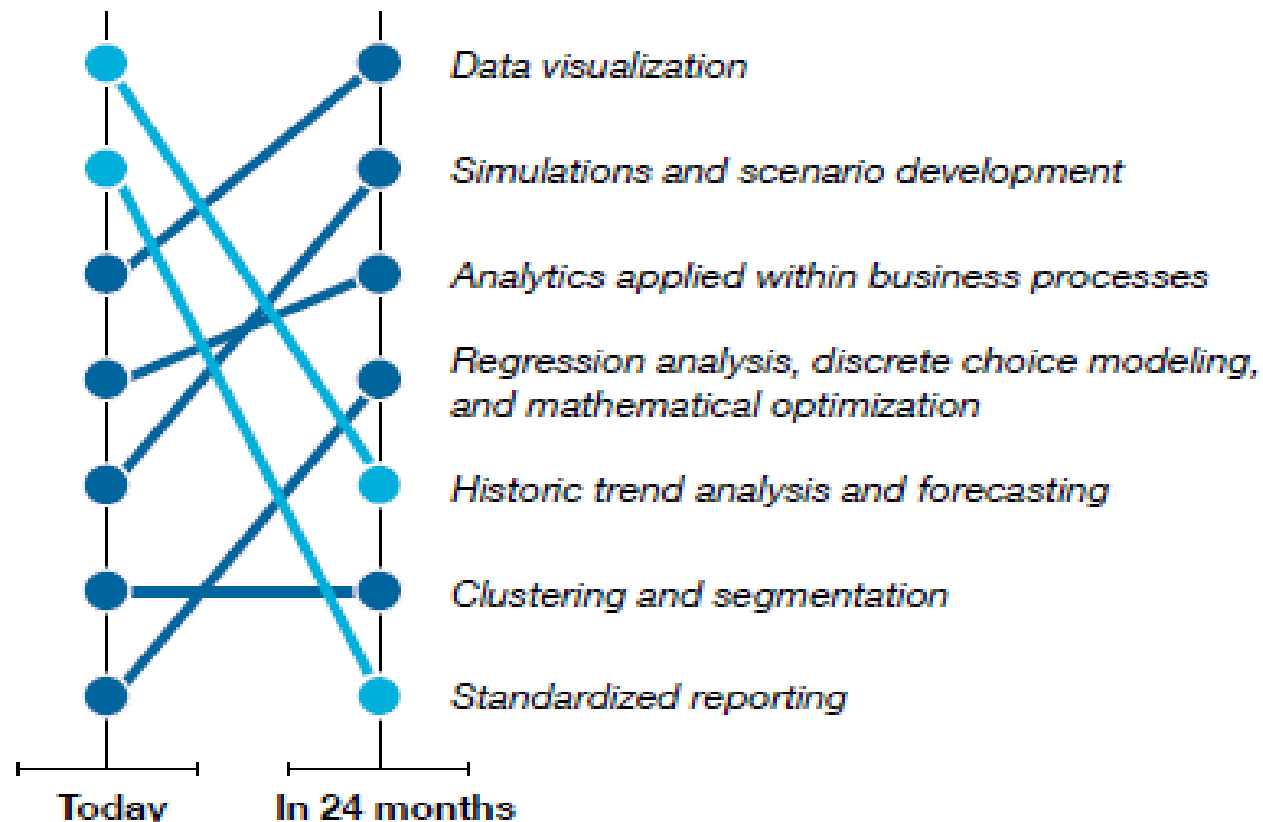


Note: Respondents were asked about their organization's application of analytics to the activities listed above. A score of 1.0 indicates an equal likelihood of applying either analytics or non-analytic methods, while a score of 0.0 indicates a tendency to use non-analytic methods.

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)

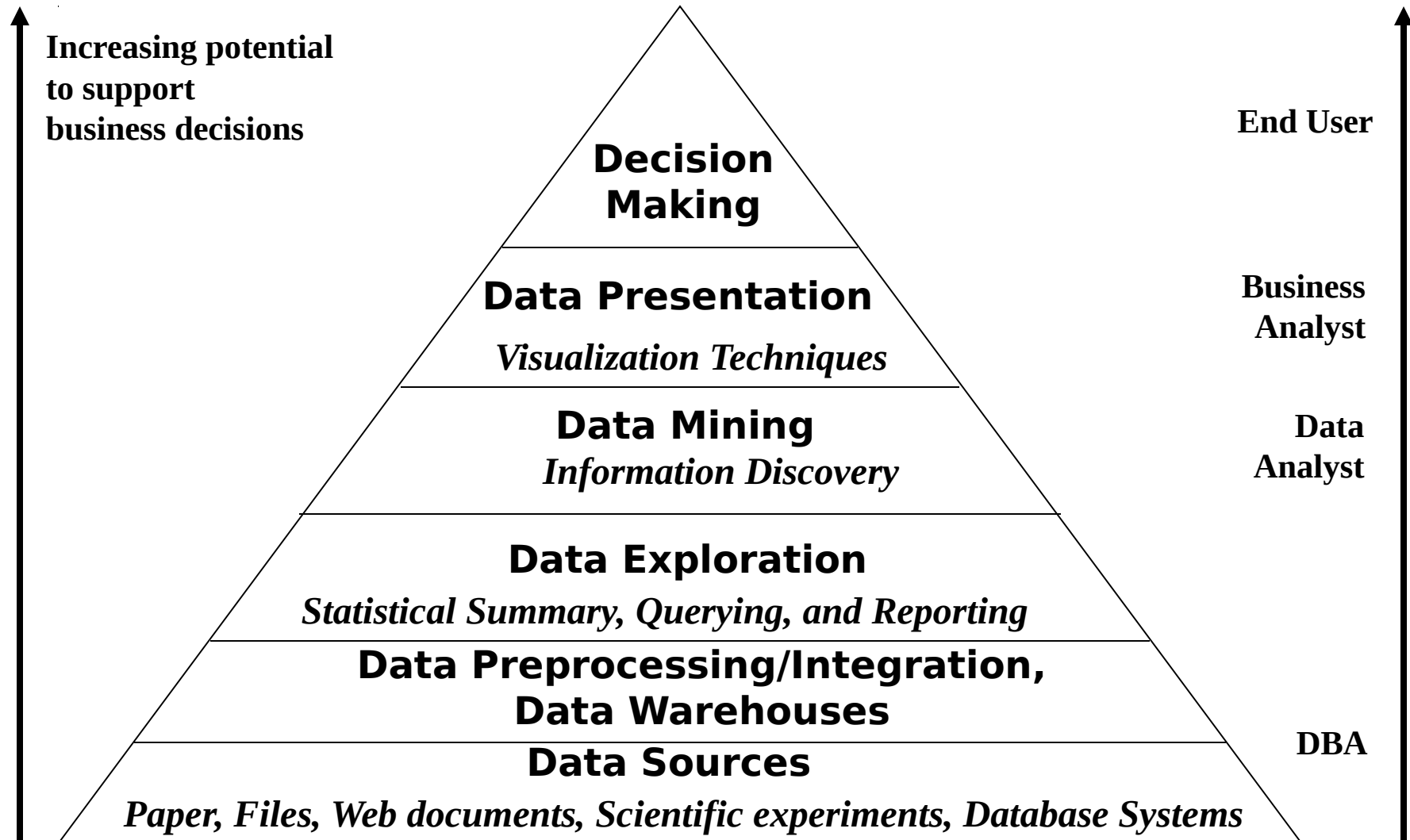
Relative value of tools



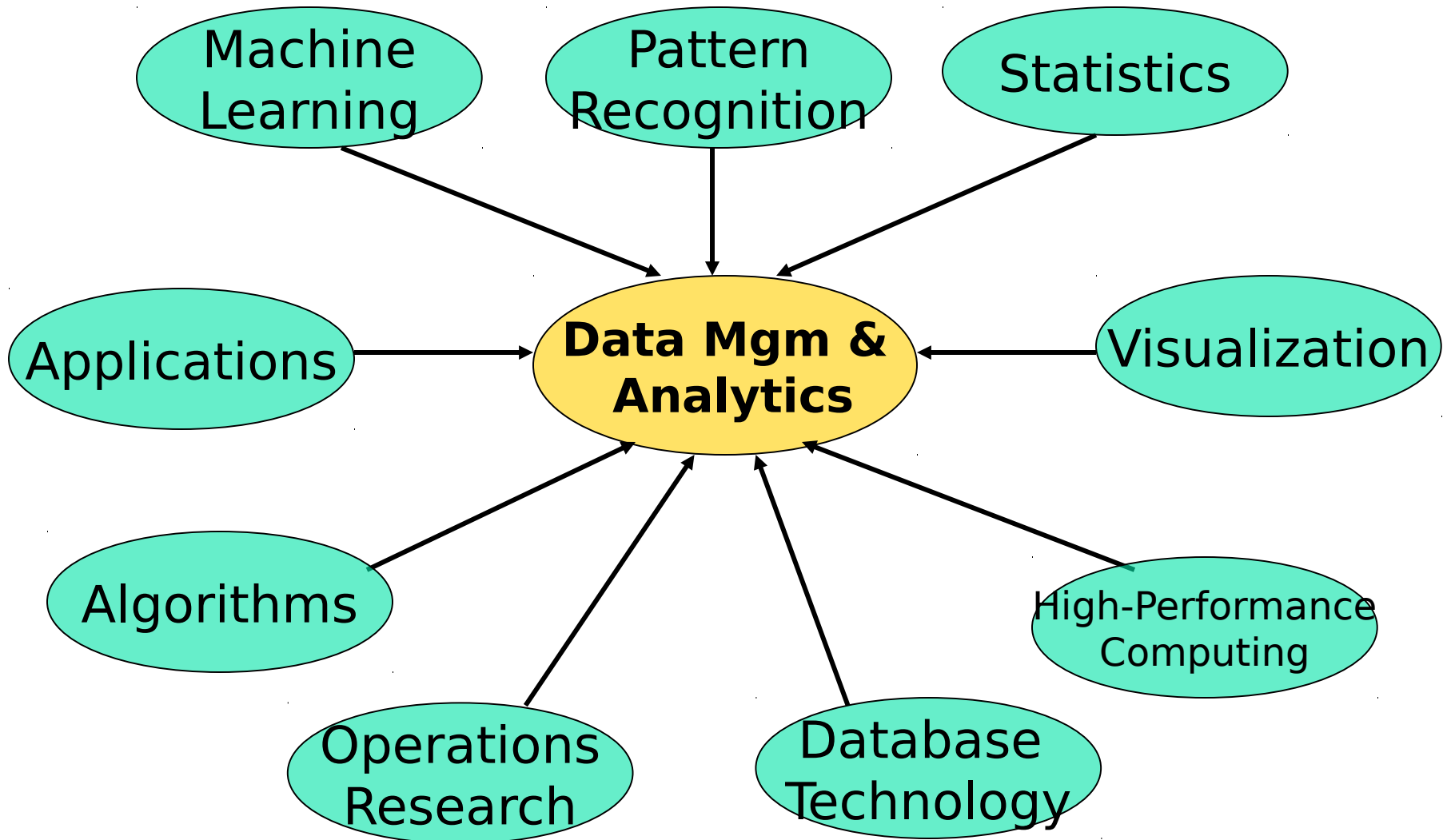
Note: Respondents were asked to "Select the type of analytics creating the most value in your organization today, and which types you believe will create the greatest value 24 months from now? (Select up to three in each timeframe.)"

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.

Data Mining in Business Intelligence



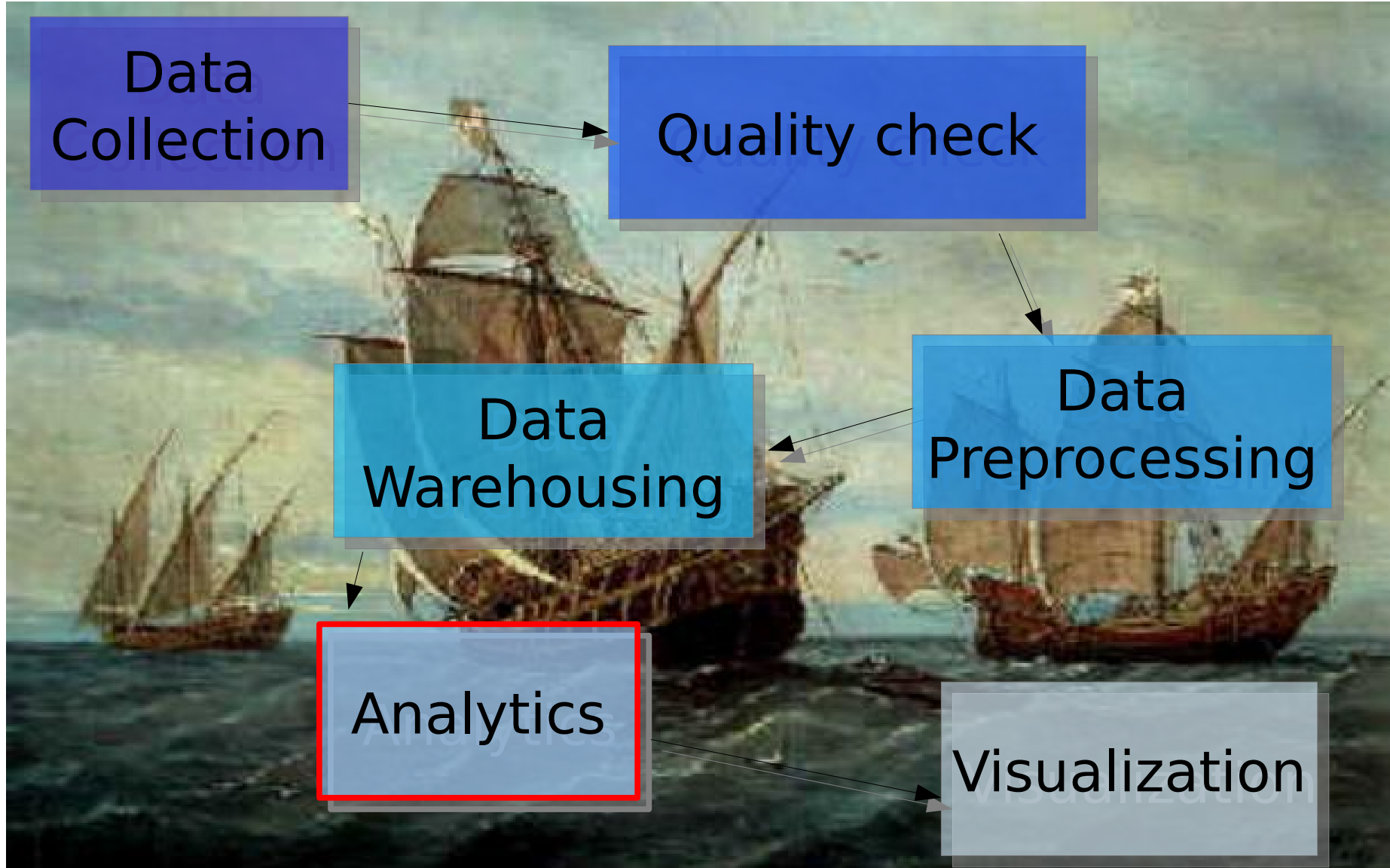
Data analytics: Confluence of Multiple Disciplines



Summarizing: why Confluence of Multiple Disciplines?

- Tremendous amount of data
 - Algorithms must be highly scalable to handle such as terabytes 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

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)*