Introduction to Data Mining

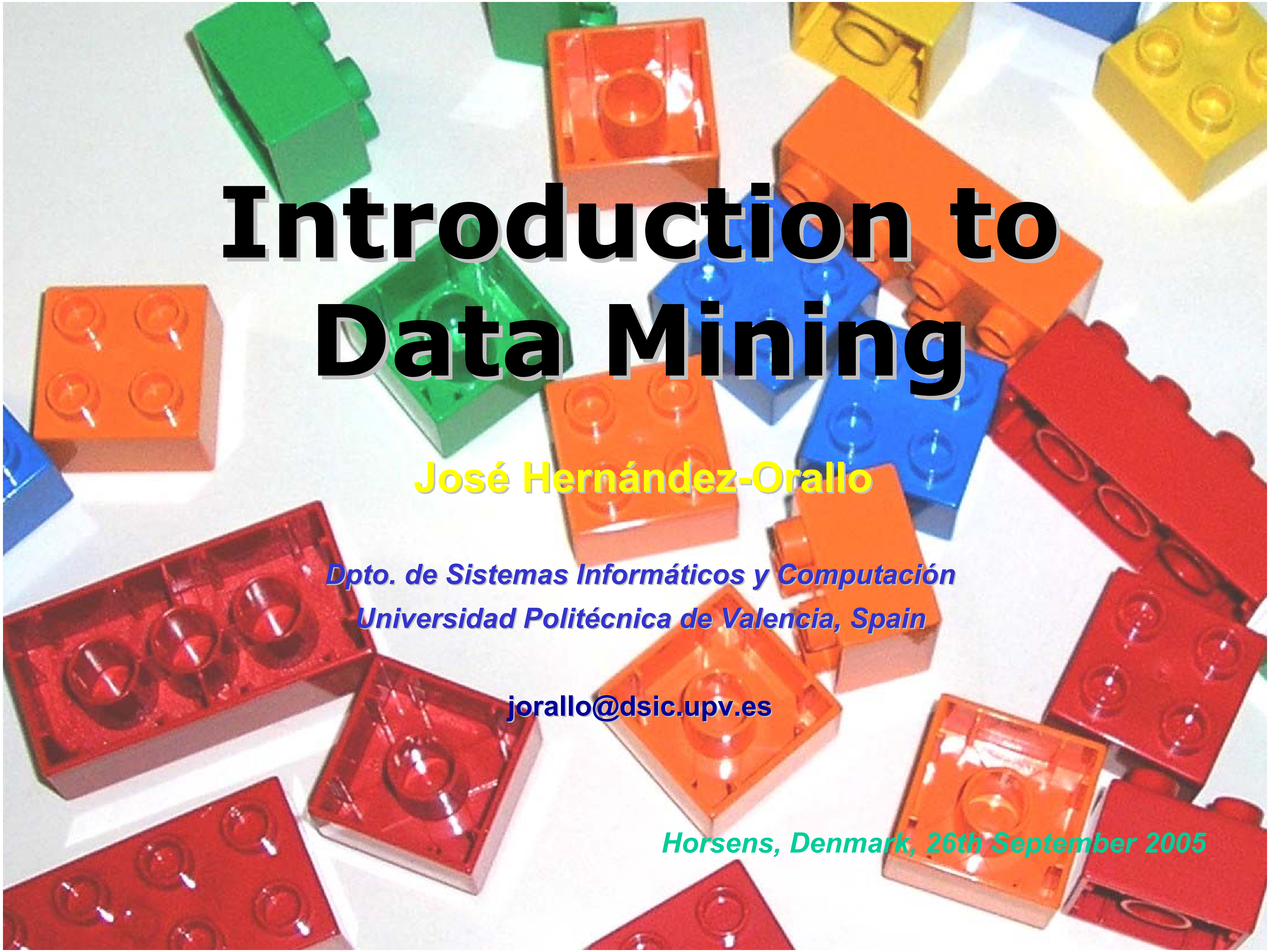
José Hernández-Orallo

Dpto. de Sistemas Informáticos y Computación

Universidad Politécnica de Valencia, Spain

jorallo@dsic.upv.es

Horsens, Denmark, 26th September 2005



Introduction to Data Mining

José Hernández-Orallo

Dpto. de Sistemas Informáticos y Computación

Universidad Politécnica de Valencia, Spain

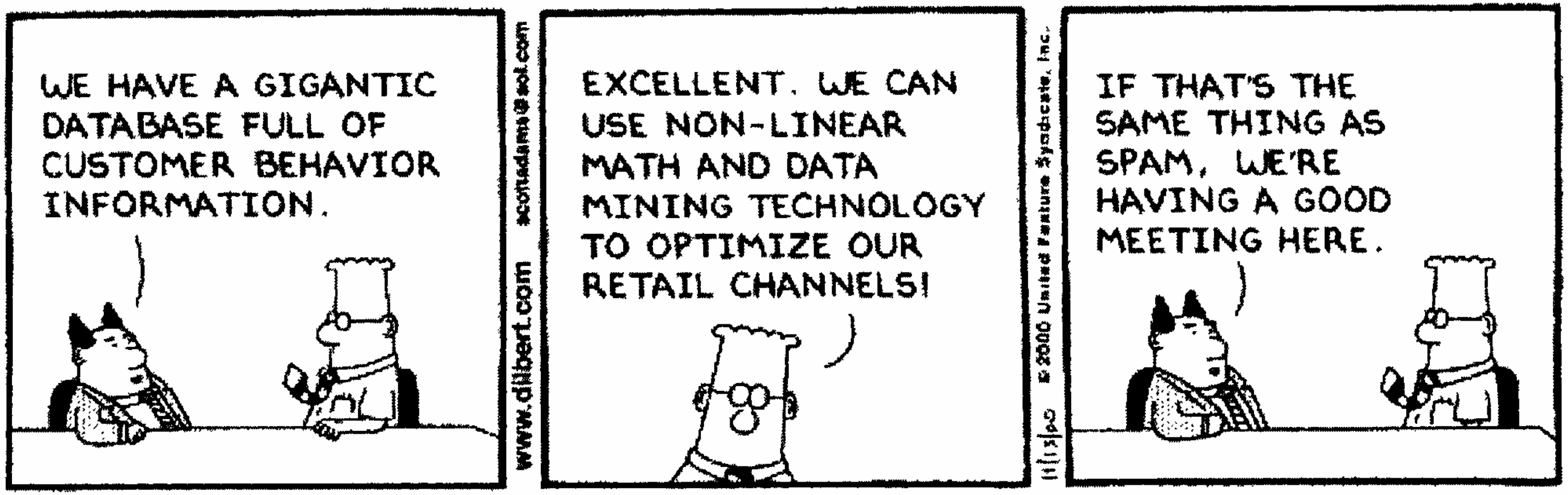
jorallo@dsic.upv.es

Horsens, Denmark, 26th September 2005

Outline

- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

• • •


WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.
EXCELLENT. WE CAN USE NON-LINEAR MATH AND DATA MINING TECHNOLOGY TO OPTIMIZE OUR RETAIL CHANNELS!
IF THAT'S THE SAME THING AS SPAM, WE'RE HAVING A GOOD MEETING HERE.
www.dilbert.com scottadams@aol.com
11/13/00 © 2000 United Feature Syndicate, Inc.

DILBERT reprinted by permission of United Feature Syndicate, Inc. [2000].

Motivation

- Old needs:
  - *Initially, the goal of information systems were to gather information about a specific domain to support decision making.*
  - *nowadays, computerised organisations with transactional software applications have specialised information systems towards a different goal: to support the basic processes in an organisation (sales, production, personnel, …).*

Motivation

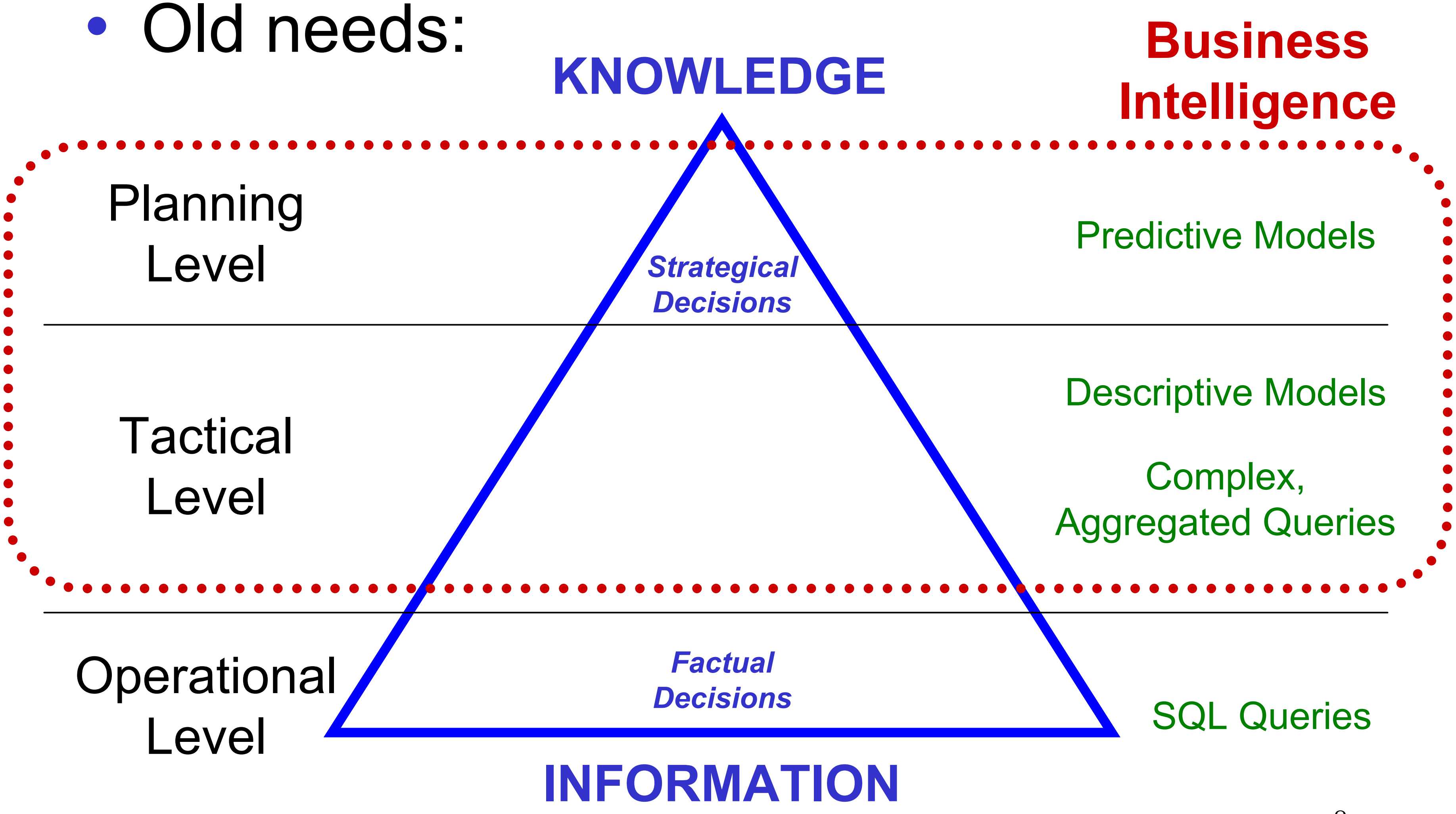
- Old needs:
  - *Information Systems nowadays are specialised for transactional work.*
    - *So are database models, database design methodologies, database management systems, etc.*
  - *These systems are not adequate for analysing the information or to convey strategic knowledge.*
    - Inappropriate organisation/architecture
    - Lack of information-to-knowledge tools.

Motivation

- Old needs:
  - Two very different kinds of data processing:
    - OLTP (*On-Line Transactional Processing*): transaction-oriented, the typical scenario in an operational database.
    - OLAP (*On-Line Analytical Processing*): oriented to analyse specific issues which are relevant for decision making.

Motivation

- Old needs:



Motivation

- Old needs:
  - *Business Intelligence?*
    - Collection of information technologies which can provide an organisation with the knowledge requirements to make strategic decisions. These comprise complex data extraction from the existing information systems and information exploitation through data analysis tools.
    - This also comprises the classical business function: “analyse the available information in the operational systems to support the executive decision making”.
  - Alternative names:
    - Business Performance Management (Meta Group).
    - CPM: Corporate Performance Management (Gartner).
    - Enterprise Performance Management

Motivation

- Old needs:
  - *Problem:*

IS’s and DB’s make it possible to convert information into information, but make it difficult to convert information into knowledge.

  - *Information Overload:*

This cannot be done manually any more.

Motivation

- Old tools:
  - Data Source:
    - Transactional Database:
      - Relational or Object-Relational Data Organisation
  - Information Delivery (Exploitation) Tools:
    - Batch Reporting Tools
    - EIS (Executive Information Systems)
    - Traditional DSS (Decision Support Systems)
    - Desktop Reporting Tools (Spreadsheet Add-ins, …)

Motivation

- New tools:
  - Data Source:
    - Data warehouses or operational data stores.
      - Other models: multidimensional
      - Other organisations: datamarts
  - Exploitation tools:
    - OLAP (On-Line Analytical Processing) Tools.
    - Data Mining
    - Simulation

Motivation

- Needs and Tools:
  - Operational decisions → operational queries:
    - Requirement Examples:
      - “Are there available seats for flight IB-2323?”.
        - » TOOL: SQL.
      - “List the products which currently have a special offer”.
        - » TOOL: SQL/Reporting tools.

Motivation

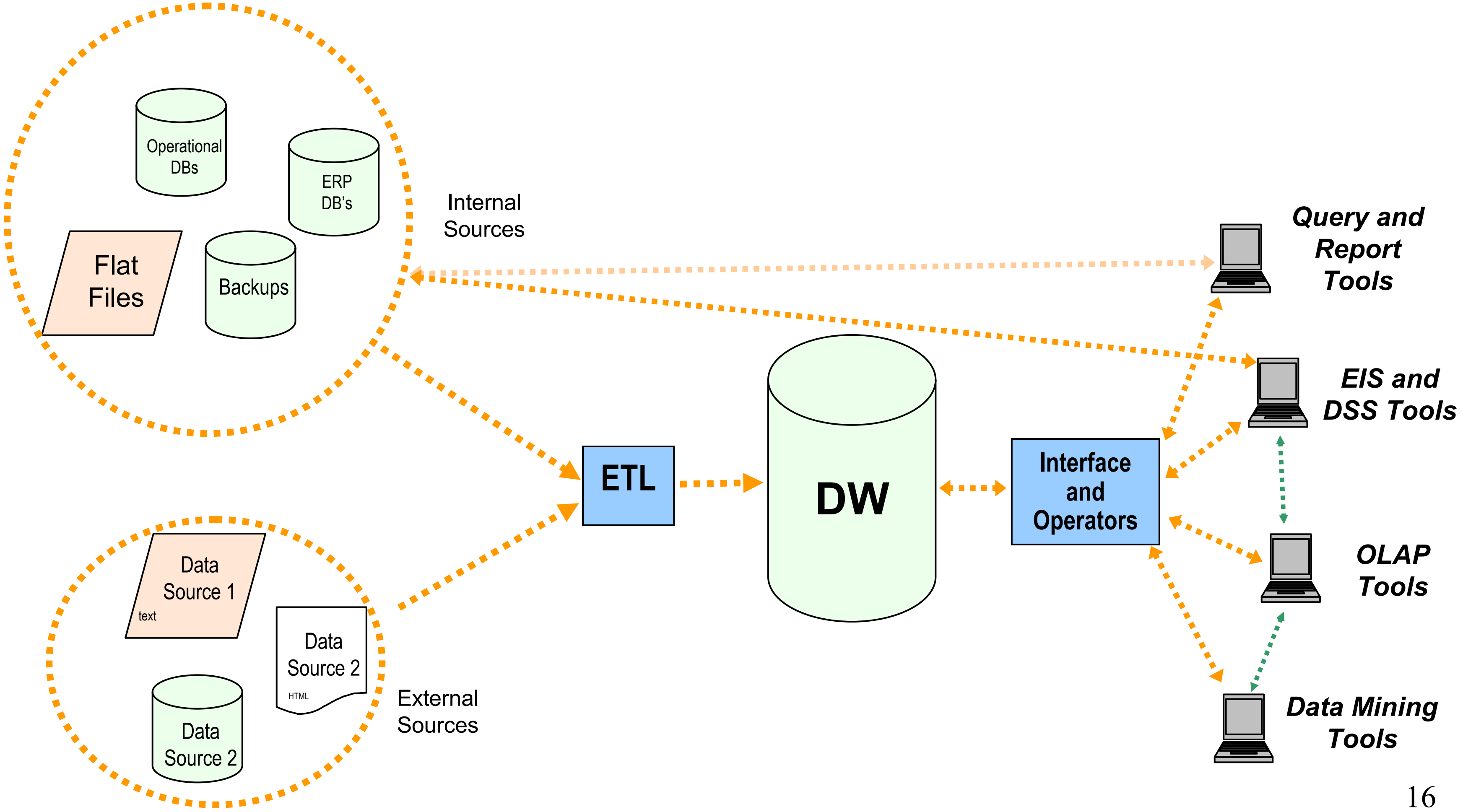
- Needs and Tools:
  - Tactical decisions → aggregated queries:
    - Requirement Example:
      - “Top ten sellers in the sales department during the last year”.
        - » TOOL: SQL, OLAP tools, data warehouses
      - “Total sales of seaside supermarkets in the second quarter of 2005 vs. total sales of interior supermarkets”.
        - » TOOL: EIS/DSS, OLAP tools, data warehouses

Motivation

- Needs and Tools:
  - Planning → descriptive or predictive models:
    - Requirement Examples:
      - “Products which are sold together frequently in seaside supermarkets”.
        - » TOOL: Descriptive Data Mining
      - “Sales expected for next quarter for all products in category “multimedia””.
        - » TOOL: Predictive Data Mining
      - “Expected penetration rates of offer 1 after offer 2 in interior supermarkets”.
        - » TOOL: Predictive Data Mining + Simulation

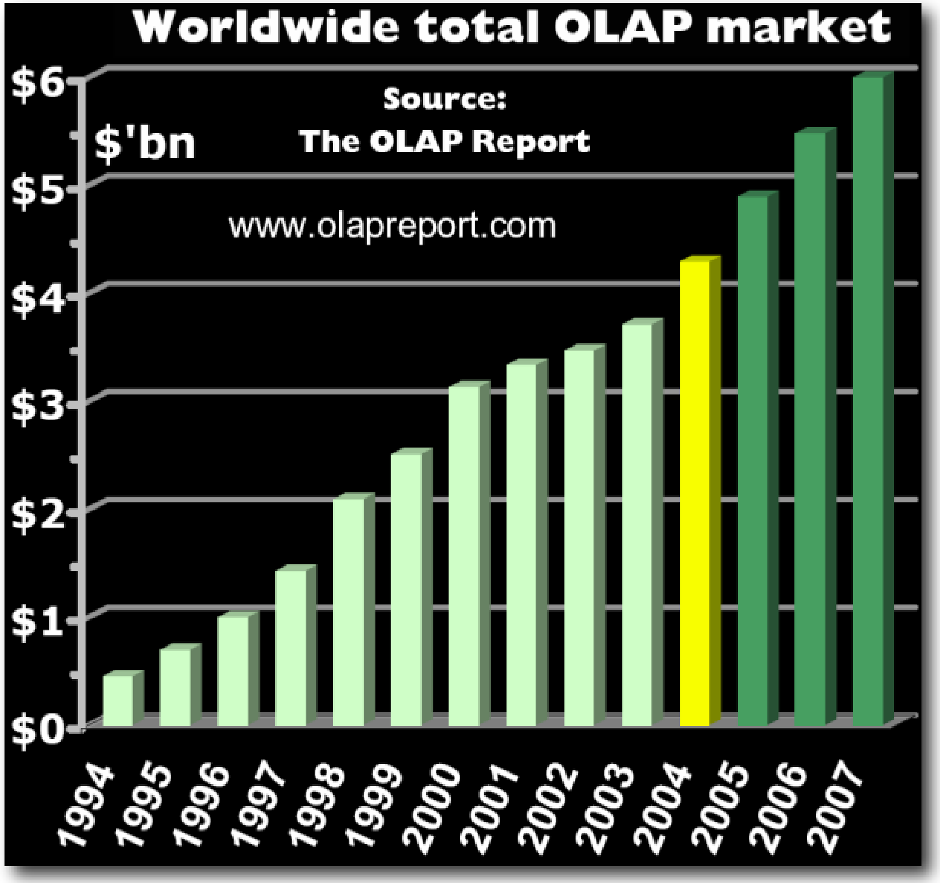
Motivation

- Architecture:



Motivation

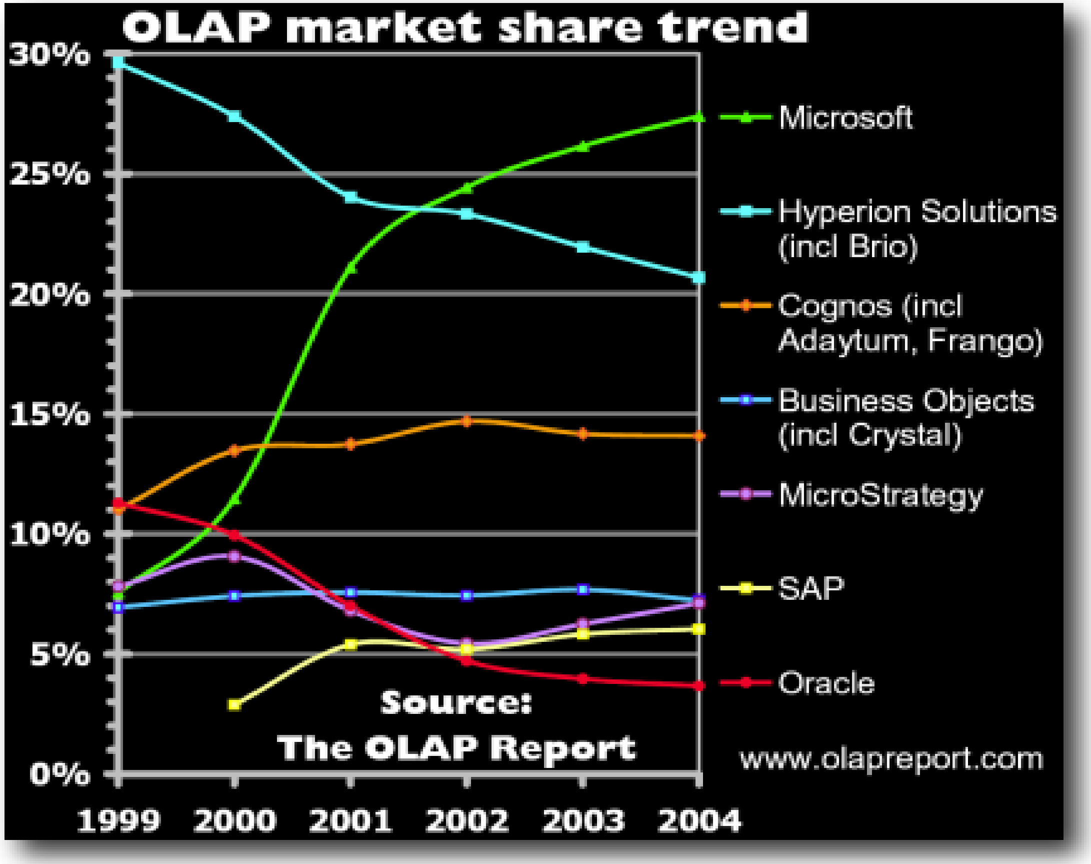
- Some economical trends:



Source: http://www.olapreport.com/market.htm

Motivation

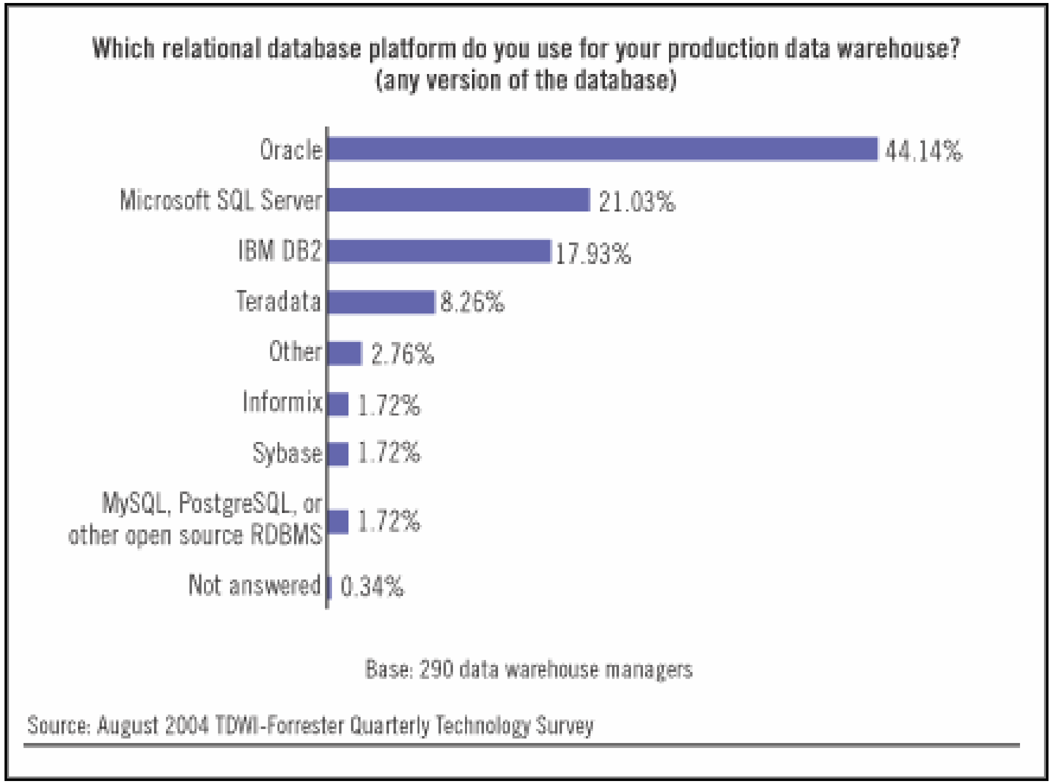
- Some tool shares:



Source: http://www.olapreport.com/market.htm

Motivation

- Share on data warehouse systems (based on RDBMS, the overwhelming majority):



Source: http://www.dmreview.com/article_sub.cfm?articleId=1023914

Motivation

- Some economical trends:

Data Mining

According to research house Gartner, the BI services market is expected to grow through 2007 at a compound annual growth rate of 9.2 percent, buoyed by an uptake in the management of BI application services.

The worldwide DW-BI services forecast represents three to four percent of the total IT services opportunity.

Source: http://www.crmbuyer.com

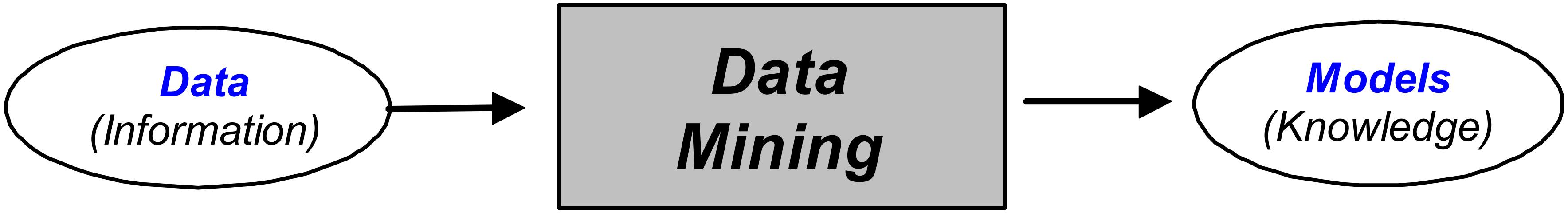
Motivation

- OLAP vs. Data Mining:

OLAP	DM
Which is the average accident rate among smokers and non-smokers?	Which are the best predictors for accidents?
Which is the average telephone bill of my current customers vs. the ex-customers who quit the company?	Will X leave the company? Which factors affect churn?
Which is the average daily purchase amount between stolen credit cards operations and legitimate ones?	Which purchase patterns are associated to credit card frauds?

Motivation

- A (simplistic) view of Data Mining:



Models are the “product” of data mining

- Strategic decisions are supported on the inferred models.

Outline

- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

Examples

- BANK AGENT:

  Must I grant a mortgage to this customer?

- SUPERMARKET MANAGER:

  When customers buy eggs, do they also buy oil?

- PERSONNEL MANAGER:

  What kind of employees do I have?

- TRADER in a RETAIL COMPANY:

  How many flat TVs do we expect to sell next month?

Examples

- **BANK AGENT:**

Must I grant a mortgage to this customer?

Historical Data:

cId	Credit-p (years)	Credit-a (euros)	Salary (euros)	Own House	Defaulter accounts	…	Returns-credit
101	15	60.000	2.200	yes	2	…	no
102	2	30.000	3.500	yes	0	…	yes
103	9	9.000	1.700	yes	1	…	no
104	15	18.000	1.900	no	0	…	yes
105	10	24.000	2.100	no	0	…	no
...	…	…	…	…	…	…	…

Data Mining

Pattern / Model:

If Defaulter-accounts > 0 **then** Returns-credit = no
If Defaulter-accounts = 0 **and** [(Salary > 2.500) **or** (credit-p > 10)] **then** Returns-credit = yes

Examples

- **SUPERMARKET MANAGER:**

When customers buy eggs, do they also buy oil?

Historical Data:

BasketId	Eggs	Oil	Nappies	Wine	Milk	Butter	Salmon	Endive	...
1	yes	yes	no	yes	no	yes	yes	yes	...
2	no	yes	no	no	yes	no	no	yes	...
3	no	no	yes	no	yes	no	no	no	...
4	no	yes	yes	no	yes	no	no	no	...
5	yes	yes	no	no	no	yes	no	yes	...
6	yes	no	no	yes	yes	yes	yes	no	...
7	no	no	no	no	no	no	no	no	...
8	yes	yes	yes	yes	yes	yes	yes	no	...
...	...	...	...	...	...	...	...	...	...

Data Mining

Pattern / Model:

Eggs → Oil : Confidence = 75%, Support = 37%

Examples

- **PERSONNEL MANAGER:**

What kind of employees do I have?

Historical Data:

Id	Salary	Married	Car	Children	Rent/ Owner	Union	Off sick/year	Work years	Gender
1	10000	yes	no	0	Rent	no	7	15	M
2	20000	no	yes	1	Rent	yes	3	3	F
3	15000	yes	yes	2	Owner	yes	5	10	M
4	30000	yes	yes	1	Rent	no	15	7	F
5	10000	yes	yes	0	Owner	yes	1	6	M
6	40000	no	yes	0	Rent	yes	3	16	F
7	25000	no	no	0	Rent	yes	0	8	M
8	20000	no	yes	0	Owner	yes	2	6	F
15	8000	no	yes	0	Rent	no	3	2	M
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Data Mining

Pattern / Model:

- **Group 1:** Without children and in a rented house. Low participation in unions. Many days off sick.
- **Group 2:** Without children and with car. High participation in unions. Few days off sick. More women and in rented houses.
- **Group 3:** With children, married and with car. More men and usually house owners. Low participation in unions.

Examples

- **TRADER in a RETAIL COMPANY:**

  How many flat TVs do we expect to sell next month?

Historical Data:

PRODUCT	Month–12	...	Month–4	Month–3	Month–2	Month–1	Month
Flat TV 30’	20	...	52	14	139	74	?
Video-dvd-recorder	11	...	43	32	26	59	?
Discman	50	…	61	14	5	28	?
Five star fridge	3	…	21	27	1	49	?
Three star fridge	14	...	27	2	25	12	?
…	…	…	…	…	…	…	...

Data Mining

Pattern / Model:

Linear Model: Flat TV Sales for Next Month:

$$V(\text{Month})_{\text{flatTV}} = 0.62 \cdot V(\text{Month-1})_{\text{flatTV}} + 0.33 \cdot V(\text{Month-2})_{\text{flatTV}} + 0.12 \cdot V(\text{Month-1})_{\text{DVD-Recorder}} - 0.05$$