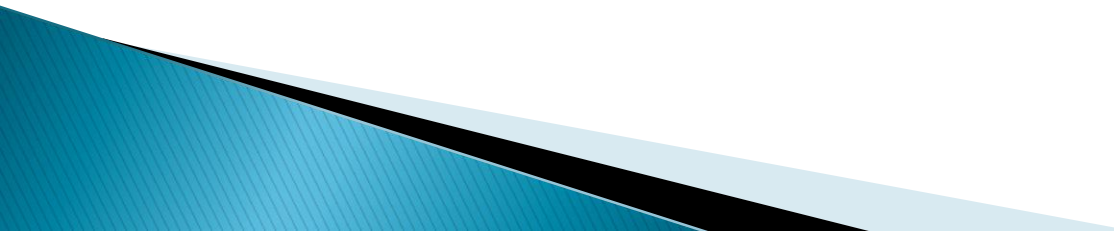


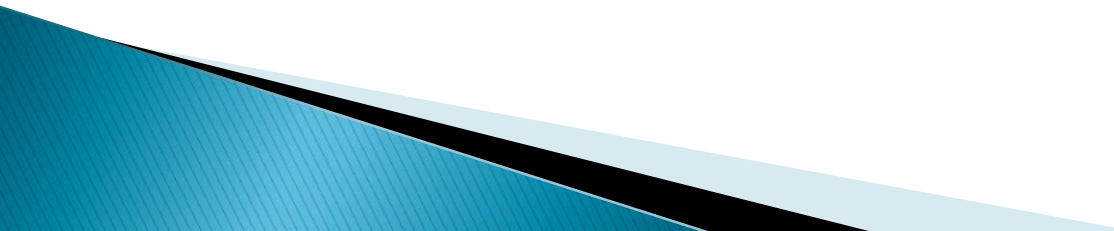
# Problems with interpretation of data mining results

David Chudán, KIZI VŠE Praha  
Seminar 14.4.2011


# Content

- ▶ Introduction into reporting
  - ▶ Business Intelligence in a few words
  - ▶ BI reporting tools
  - ▶ DM vs. BI
  - ▶ Use case
  - ▶ Presentation of the use case results
- 

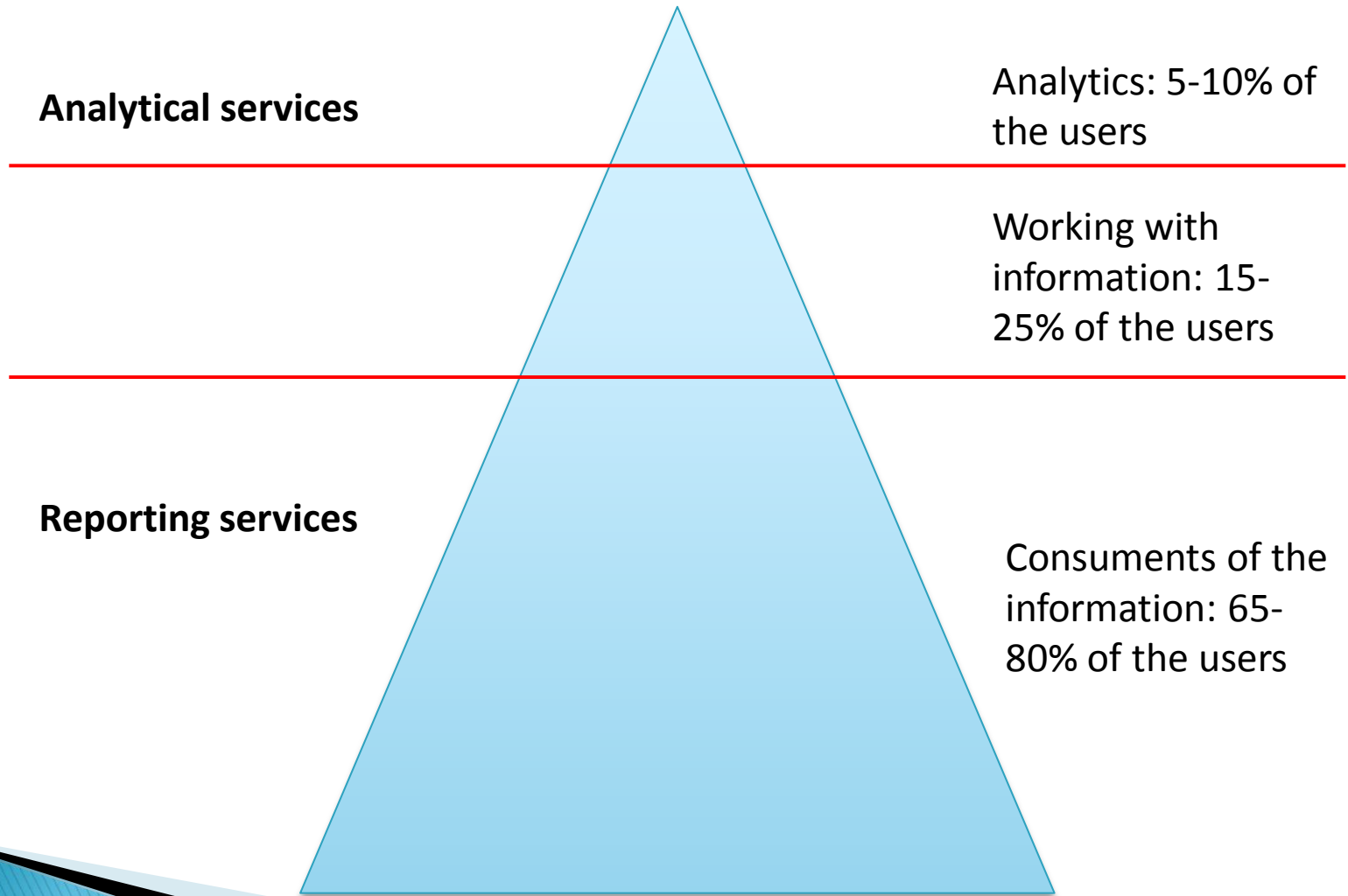
# Reporting

- ▶ Should serve as a decision support at all levels of the organizational structure
  - ▶ Is a last part of the process of gaining, saving, transforming and manipulating with data
  - ▶ Report is pre-defined, system oriented data view focused on some analytical needs
  - ▶ Report is human readable
- 

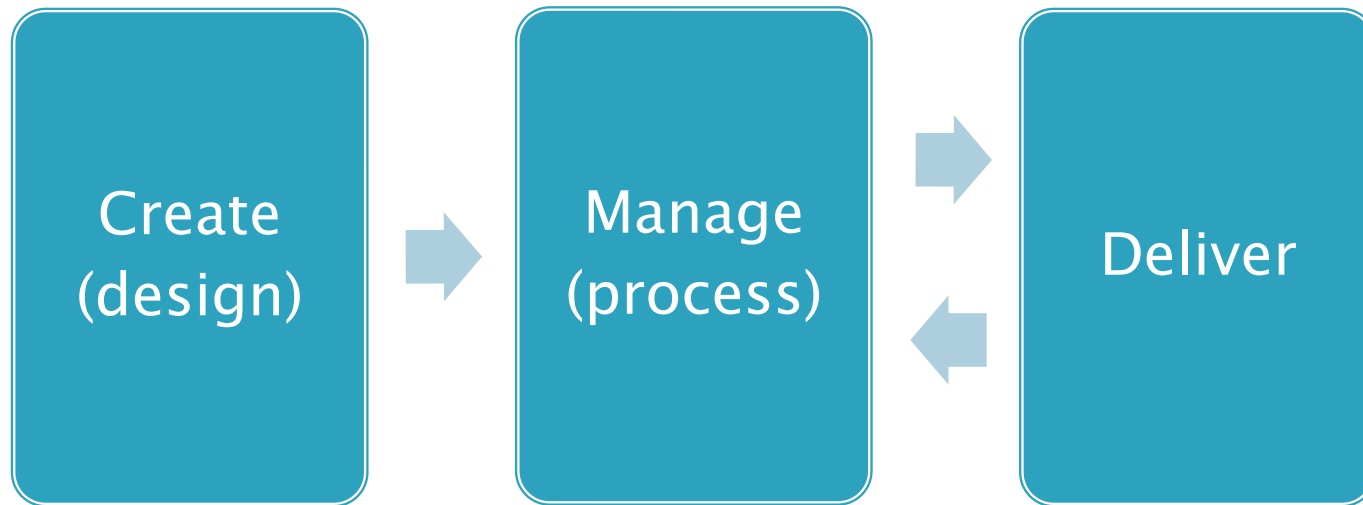
# Reports

- **Static** – similar to paper reports
  - **Interactive** – adaptable, clickable, it is possible to focus on certain area of interest
- 
- **Standard** – predefined report whose layout is not meant to be changed by the end user
  - **Ad hoc** – for less technically advanced users (without prior knowledge of DB schema or query language)
  - **Enterprise** – „in-house“ reporting from individual departments
  - **Embedded** – report generation is integral part of an application
  - **B2B** – reports for business partners
- 

# Usage of the reporting technologies



# Report lifecycle

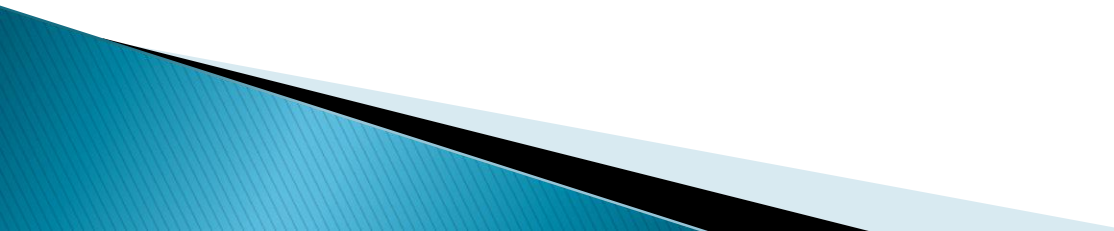


**Create** – it is likely to be possible to present a report in many forms, the need of variability, flexibility

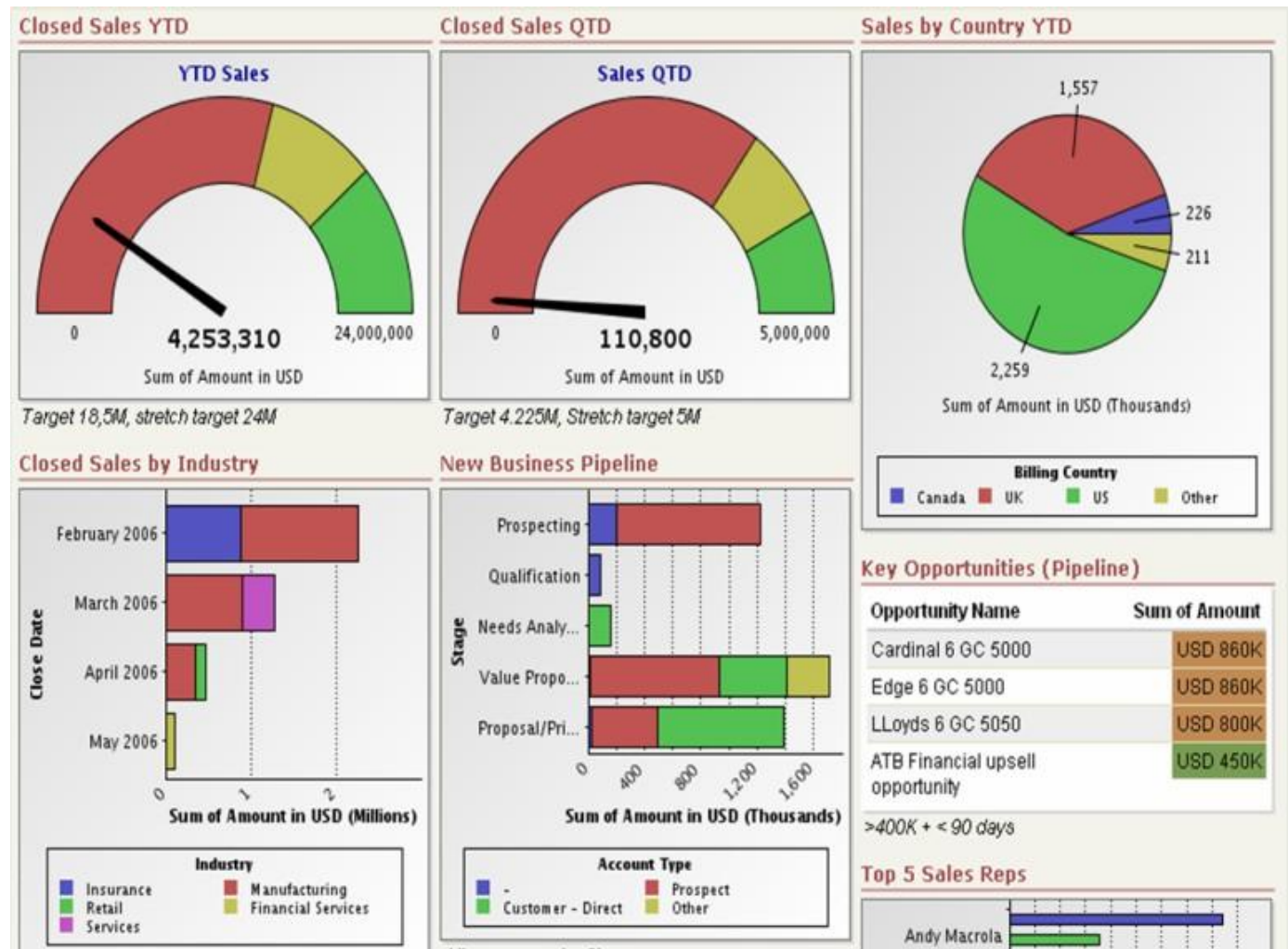
**Manage** – control and adjustments of the report proposals

**Deliver** – the method of delivery (online) and the form of delivery (e-mail)

# Dashboards

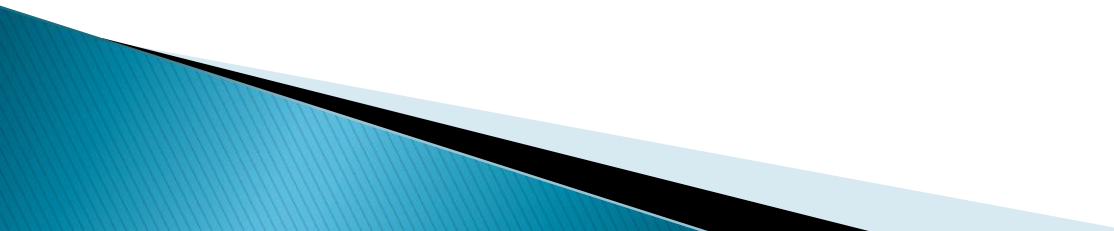
- ▶ The term is borrowed from car industry
  - ▶ Web based page where are integrated various sources from the business in real time.
  - ▶ Also application for Mac OS X operating system which is used for hosting widgets applications.
  - ▶ We can say that dashboards are interactive reports with minimum of the text.
- 

# Dashboard – example





# Manager's demand

- ▶ Managers don't have time to read long text reports (some of them don't read e-mail longer than 3 lines)
  - ▶ Managers hates anything that even remotely looks like some mathematical formulas
  - ▶ Simple, clear graphs are ideal for them
- 

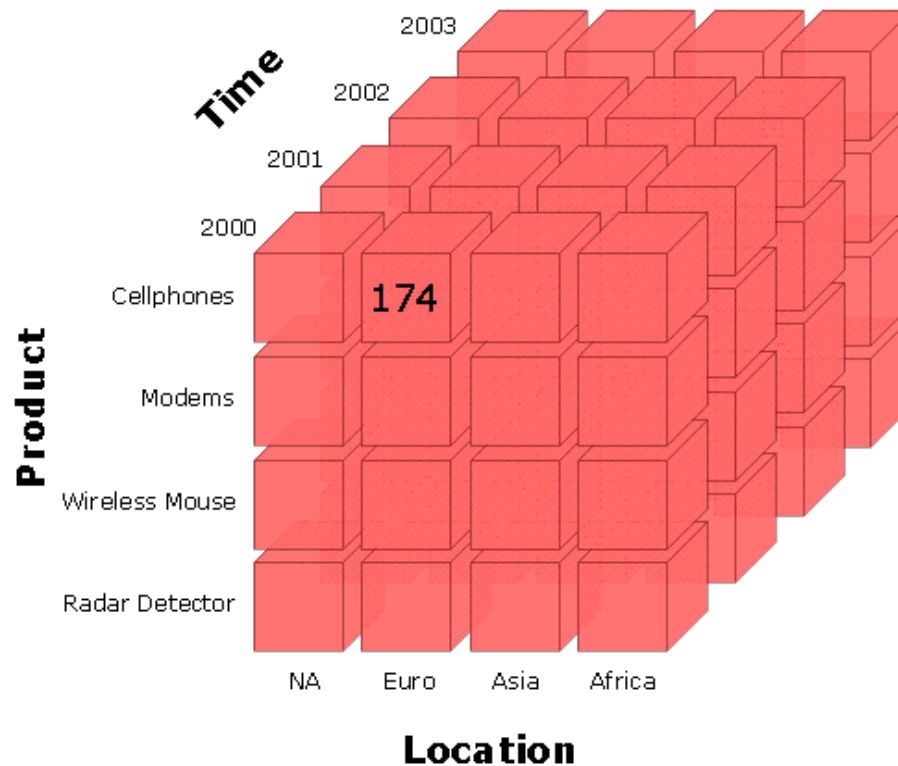
# Business Intelligence in a few words

Computer-based techniques used in spotting, digging-out, and analyzing 'hard' business data, such as sales revenue by products or departments or associated costs and incomes. Objectives of a BI exercise include understanding of a firm's internal and external strengths and weaknesses, understanding of the relationship between different data for better decision making, detection of opportunities for innovation, and cost reduction and optimal deployment of resources.

# BI basic components

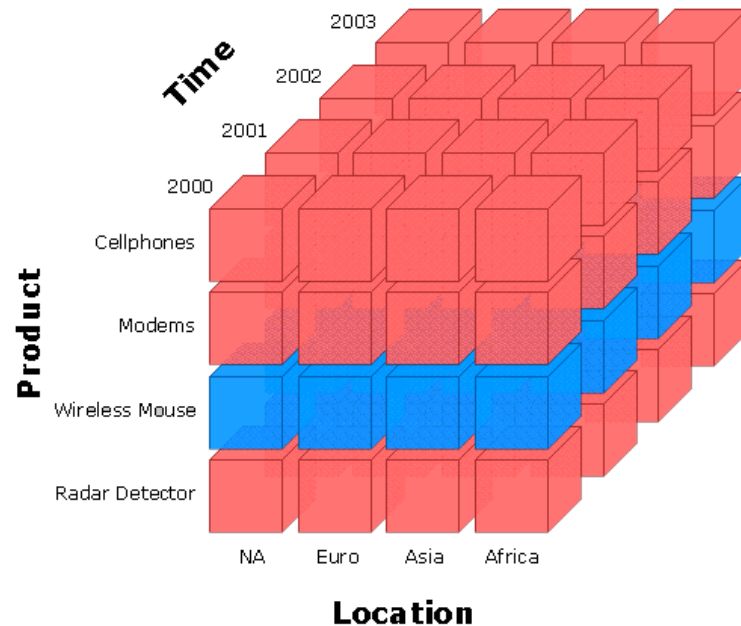
- ▶ OLTP systems
- ▶ ETL
- ▶ Multidimensional databases
- ▶ Data warehouse
- ▶ End user applications
  - Reporting
  - Executive information system

# Multidimensional databases – the OLAP cube

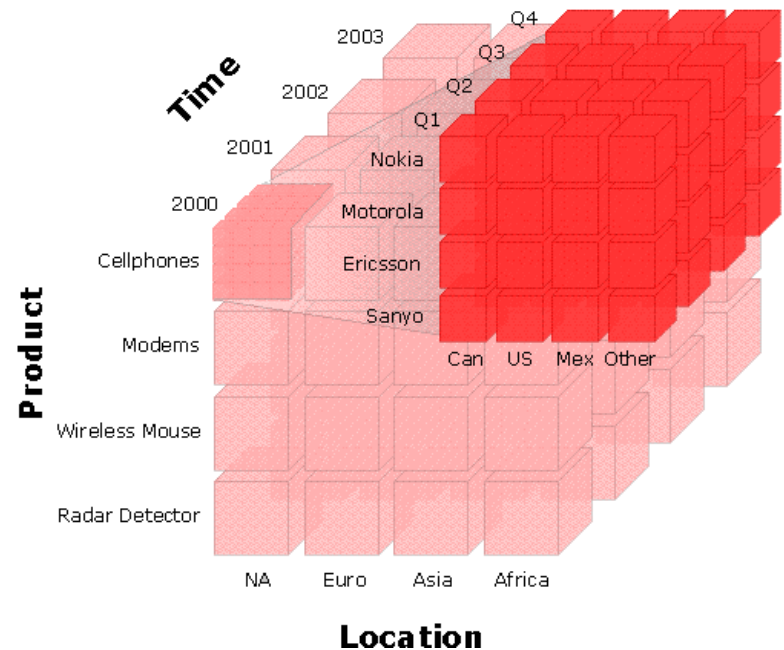


# Multidimensional databases – the OLAP cube

## Slicing

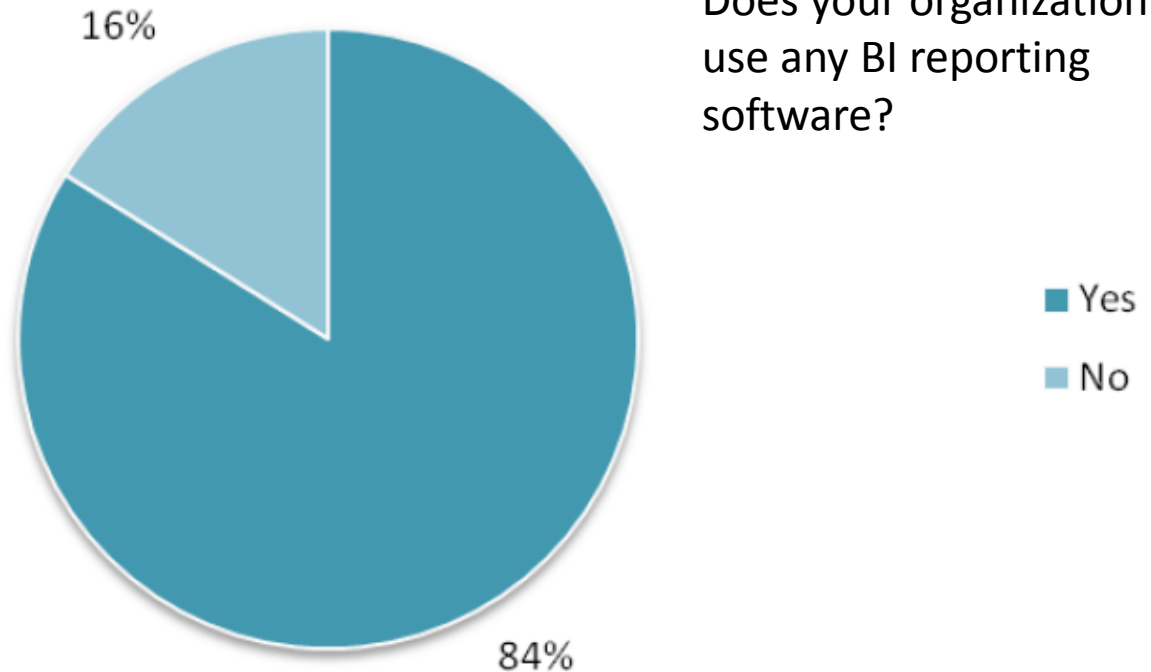


## Dicing



# Usage of BI tools and reporting

## Use of Business Intelligence and Reporting



Source: Adoption and Usage Survey: Open Source Business Intelligence and Reporting, available online at <http://www.b-eye-network.com/files/2009%20Open%20Source%20BI%20and%20Reporting%20Research%20Report.pdf>

# Some (BI) reporting tools

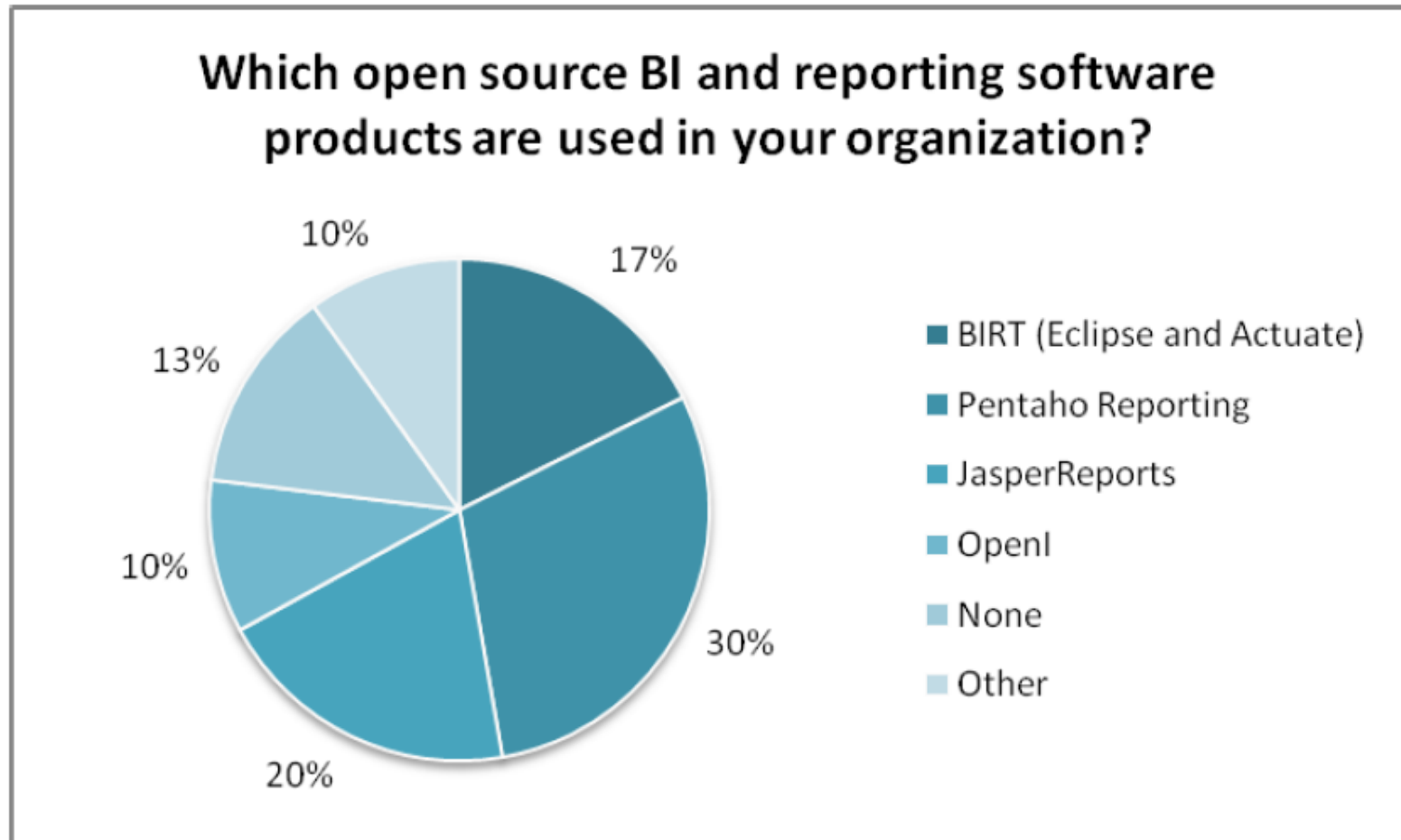
## OpenSource

- JasperReports
- Pentaho

## Commercial

- Clarity Systems Ltd
  - Crystal Reports
  - Oracle XML Publisher
  - ProClarity
  - SQL Server Reporting Services
  - Zoho Reports
- 

# Open source BI tools



Source: Adoption and Usage Survey: Open Source Business Intelligence and Reporting, available online at <http://www.b-eye-network.com/files/2009%20Open%20Source%20BI%20and%20Reporting%20Research%20Report.pdf>



# JasperReports

- ▶ Based on Java
- ▶ Enables to create almost any kind of the report imaginable including dashboards, tables, crosstabs, operational pixel-perfect print-ready layouts, and interactive web reports.
- ▶ Flexible output (many formats)

# JasperReports

## Report Viewer

Back Options Export

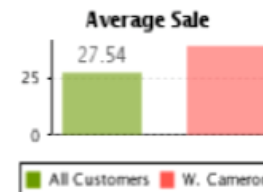
**Wildon Cameron**  
5301 Loeffler Lane  
Spokane WA 98803  
USA



Customer Card Number: 68543835878

Phone 1	891-555-9388
Phone 2	234-555-4952
Marital Status	M
Yearly Income	\$130K - \$150K
Gender	M
Total Children	4
Children at Home	4

Date Acctn Opened	19-Jun-93
Member Card	Silver
Birth Date	13-Nov-45
Education	Bachelors Degree
Occupation	Professional
Homeowner	N
Num Cars Owned	2



Purchase details for customer Wildon Cameron:

# Pentaho


- BI suite with integrated reporting, dashboard, data mining, workflow and ETL capabilities.
- Pentaho reporting
  - Broad data source support including relational, OLAP, or XML-based data sources
  - Web-based ad hoc query and reporting for business users
  - Popular output options including Adobe PDF, HTML, Microsoft Excel, Rich Text Format, or plain text

# Pentaho

- ▶ Data mining
  - Uses Weka data mining technology (clustering, segmentation, decision trees, neural networks...)
  - Output can be viewed graphically, interacted with programmatically (enabling developers to create completely custom solutions), or used data source for reports, further analysis, and other processes.
  - PMML support

# Pentaho


Pentaho Sample Report  
vFreeReport



Steel Wheels, Inc.  
Human Resources Actual vs Forecast  
Period ending June 30, 2005

Region: Central

Department	Position	Actual	Budget	Variance
<b>Executive Management</b>				
SVP Partnerships		\$367,415	\$362,100	\$24,685
SVP WW Operations		\$476,000	\$725,887	\$249,887
SVP Strategic Development		\$383,242	\$403,405	\$20,163
CEO		\$549,625	\$522,250	-\$27,375
<b>Total</b>		<b>\$1,776,282</b>	<b>\$2,043,642</b>	<b>\$267,360</b>
<b>Department Finance</b>				
Controller		\$570,373	\$577,070	\$6,697
Payroll		\$367,415	\$432,100	\$64,685
Administrative Assistant		\$827,861	\$760,990	-\$66,871
IS		\$570,759	\$577,346	\$6,587
CFO		\$770,272	\$719,855	-\$50,417
<b>Total</b>		<b>\$3,106,680</b>	<b>\$3,067,361</b>	<b>-\$39,319</b>
<b>Department Human Resource</b>				
Sexual Harassment		\$530,473	\$538,570	\$8,097
EOE		\$530,207	\$538,380	\$8,173
HR Generalists		\$856,190	\$771,225	-\$84,965
HR Training		\$397,473	\$443,570	\$46,097
Administration		\$549,625	\$552,250	\$2,625
SVP HR		\$574,895	\$570,300	-\$4,595
<b>Total</b>		<b>\$3,438,863</b>	<b>\$3,414,295</b>	<b>-\$24,568</b>
<b>Department Marketing &amp; Communication</b>				
Graphics		\$782,375	\$728,500	-\$53,875
Writer		\$405,985	\$459,650	\$53,665
Analyst Relations		\$383,375	\$443,500	\$60,125
Press Relations		\$497,296	\$524,872	\$27,576
CMO		\$827,861	\$760,990	-\$66,871
Product Marketing Mgr		\$693,531	\$665,040	-\$28,491



Steel Wheels  
500 International Speedway, Daytona, FL 32114  
(123) 455-7890 <http://www.steelwheels.com>  
Run Date: 2/2/06 1:24 PM

**TO:** Reims Collectables  
59 rue de l'Abbaye, null  
Reims, null 51100 France

**INVOICE**

Attn: Paul Henriot  
Sales Rep: 1337  
Terms: Net 30 days

Invoice #: 10121  
Account Number: 353  
Date: May 07, 2003

SKU	Product Description	Price/Unit	Qty Ordered	Total Price
S50_4713	2002 Yamaha YZR M1	\$74.85	44	\$3,293.40
S24_2360	1982 Ducati 900 Monster	\$76.88	32	\$2,460.16
S32_4485	1974 Ducati 350 Mk3 Desmo	\$86.74	25	\$2,168.50
S12_2823	2002 Suzuki XREO	\$165.68	50	\$8,284.00
S10_1678	1969 Harley Davidson Ultimate Chopper	\$81.35	34	\$2,765.90
				<b>\$18,971.96</b>

**Send Payment and Remittance Slip to:**  
Steel Wheels  
500 International Speedway  
Daytona, FL 32114

*Thank you for your business!*

Account Number: 353

Reims Collectables

Page 1 / 5

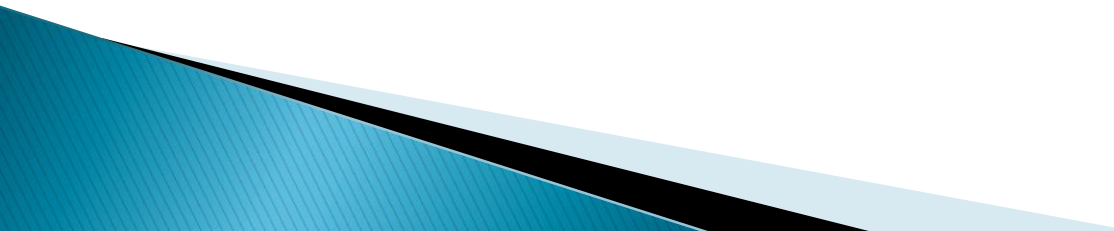
**REMITTANCE**

Reims Collectables  
59 rue de l'Abbaye, null  
Reims, null 51100 France

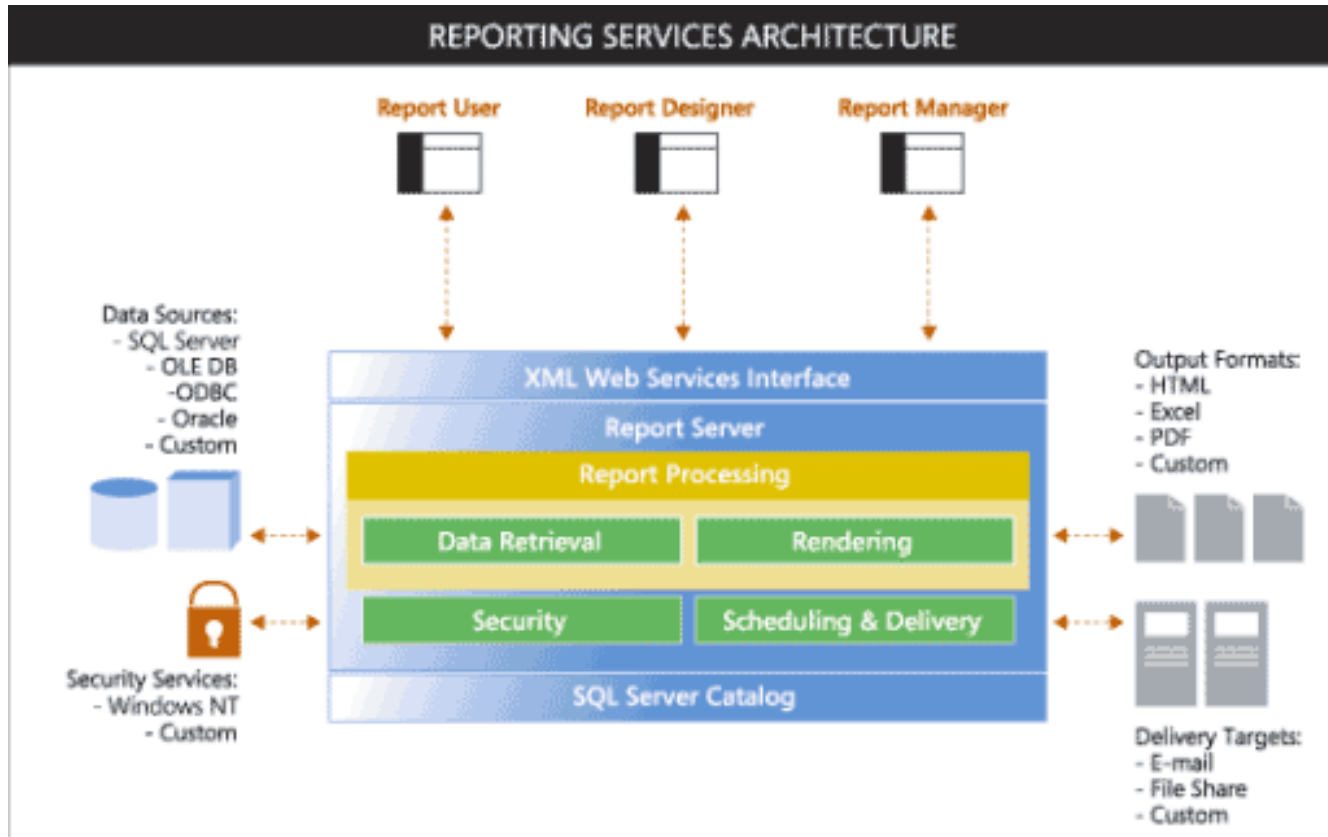
Invoice #: 10121  
Account Number: 353

AMOUNT ENCLOSED: \_\_\_\_\_

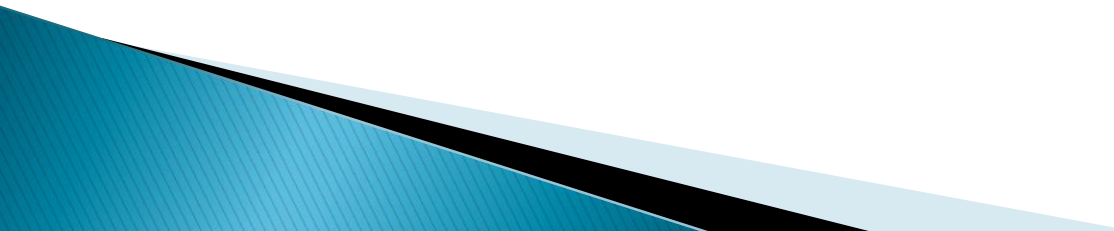
# SQL Server Reporting services

- ▶ Reports are defined in the visual environment (Report Builder, Visual Studio Report Designer...), which produces RDL language (Report Definition Language), where the reports are described.
  - ▶ Then they are managed in Report Manager
  - ▶ And delivered and presented to the end user from the web environment (or can be exported into many various formats).
- 

# SQL Server Reporting services architecture



# Business intelligence and data mining

- ▶ Data mining falls under the BI (at least in the business environment)
  - ▶ OLAP and data mining can complement each other (one supplements other), but they are used to solve different kinds of problems
  - ▶ OLAP and data mining can exist independently
- 

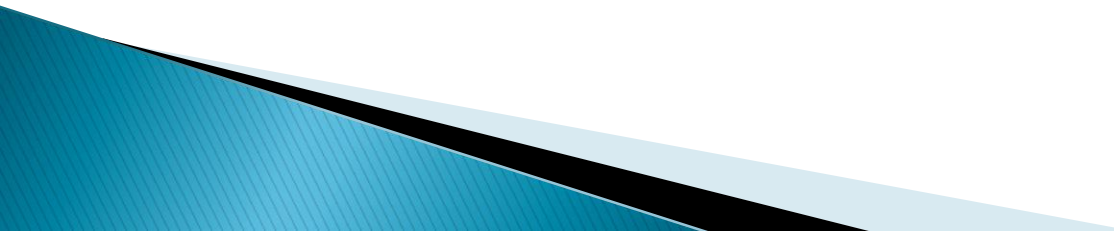


# Differences between DM and BI

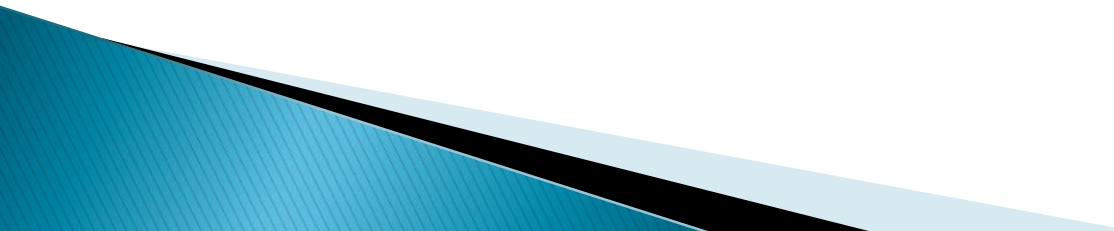
Feature	BI	DM
Usage	What is happening in the company right now?	Hidden knowledge, predictions
Technology	Summation, slicing, dicing	Many different algorithms
Data granularity	Data summation	Data at a record level
Number of attributes	Few	Many
Data size for one dimension	Small, medium	Usually huge

Modified from presentation Industry IS, available at [http://www.google.com/url?sa=t&source=web&cd=3&ved=0CCQqFjAC&url=http%3A%2F%2Fagents.felk.cvut.cz%2Fwiki%2Flib%2Fexe%2Ffetch.php%3Fid%3Dteaching%253A%26cache%3Dcache%26media%3Dteaching%3A%26apredn%3Aa0m33pis\\_komplet.pdf&rct=j&q=pr%25AFmyslov%25A9%20informa%25C4%8Dn%25C3%AD%20syst%25C3%A9my%20tom%25C3%A1%20vl%25C4%8Dek&ei=ZPmmTbK2HpHFswaAtbGRCA&usq=AFQjCNHILGGa65RuGEk49CbidPaKWarw&cad=rja](http://www.google.com/url?sa=t&source=web&cd=3&ved=0CCQqFjAC&url=http%3A%2F%2Fagents.felk.cvut.cz%2Fwiki%2Flib%2Fexe%2Ffetch.php%3Fid%3Dteaching%253A%26cache%3Dcache%26media%3Dteaching%3A%26apredn%3Aa0m33pis_komplet.pdf&rct=j&q=pr%25AFmyslov%25A9%20informa%25C4%8Dn%25C3%AD%20syst%25C3%A9my%20tom%25C3%A1%20vl%25C4%8Dek&ei=ZPmmTbK2HpHFswaAtbGRCA&usq=AFQjCNHILGGa65RuGEk49CbidPaKWarw&cad=rja)

# Use case

- Dataset from Department of Information Technologies
  - Large survey of czech IT companies
  - 600 companies, over 100 questions
  - For DM data were divided into 7 groups of attribute (companies characteristics, processes and IT management, ICT services, ICT x core business, complexity of ICT, ICT architectures, cloud computing)
- 

# Data mining

- All combinations of groups of attributes were analyzed with Founded Implication and Above average quantifiers
  - For individual combination the parameters of the DM task were set to have maximum of 40 rules
  - Total of about 1500 rules
  - **Problem** – how to determine which rules are interesting?
- 

# Interesting rules

- The rules are interesting, if they are strong and do not imply from domain knowledge (in this data set, we don't have any prepared from domain expert)
- Another problem is with similar rules, e.g.  
**Rule1:** size of company(50–249)&sector(education system) => barrier of insufficient sources  
**Rule2:** size of company(50–249)&sector(education system)&the company's characteristics(original czech company) => barrier of insufficient sources

Does the rule 2 brings anything new compares to the rule 1?

# The types of rules

- **Homogeneous rules** – the rules, in which either in antecedent or consequent prevail (over the 50%) one attribute
- **Heterogeneous rules** – the rules, in which either in antecedent or consequent do not prevail (over the 50%) one attribute
- **Dominant rules** – the rules which have confidence  $p=1$  (are valid for 100% of examples) – only for FUI quantifier

# Presentation of the rules

Vstup × Vstup	Charakteritika firmy (A)	Procesy, řízení informatiky (B)	ICT služby (C)	Priority, náklady, bariéry, efekty ICT (D)	Složitost informatiky (E)	Architektury (F1)	Cloud computing (F2)
Charakteristika firmy (A)		FUI FUI AA	FUI FUI AA AA	FUI FUI FUI AA AA	FUI FUI AA	FUI AA	FUI FUI AA
Procesy, řízení informatiky (B)			FUI FUI FUI AA	FUI FUI AA	FUI AA	FUI AA	FUI AA

- Homogeneous rules
- Heterogeneous rules
- Dominant rules

# Presentation to the end user

The form of detailed, semiautomatically created, but long analytical report.

procesům je věnována část 3. V části 4 je uvedeno zadání všech možných vztahů a kritérium pro výstup vztahu. Výstup v našem případě zahrnoval 16 vztahů, je třeba je pečlivě interpretovat, například je rozřídít na předpokládané a překvapující. Náznak takového rozřídění je v kapitole 4.1.

## 1.1 Znění jednotlivých otázek

Následuje seznam otázek, které byly v dotazníkovém šetření prezentovány firmám. Dobývání znalostí tedy vychází z odpovědí jednotlivých firem na tyto otázky.

Otázka A1: Kolik zaměstnanců má vaše firma / organizace?

Otázka A2: V jakém odvětví vaše firma / organizace působí?

Otázka A3: Jaký je charakter vaší firmy / organizace?

Otázka B2: Jak jsou ve vaší firmě / organizaci definovány podnikové procesy?

Otázka C1: Do jaké míry jsou ve vaší firmě / organizaci definovány procesy řízení informatiky?

## 2.Charakteristiky firmy

### A1 Velikost firmy

Source article: A(1,2,3) => B2C1, created: 09.01.11 04:05

Kategorie	Výčet hodnot	Frekvence
10-49	1	100
50-249	2	380
>=250	3	120

Histogram nejčastějších kategorií

Graph:  Ostatní: From:   Columns:

# Presentation to the end user

Summary for the management - From the data mining task it is clear, that 100% of large, original czech companies from the production sector has got IT department organized as one unit.

		informatiky (B)		bariéry, efekty ICT (D)	(E)		(F2)
Charakteristika firmy (A)		FUI FUI AA	FUI FUI AA AA	FUI FUI FUI AA AA	FUI FUI AA	FUI AA	FUI FUI AA

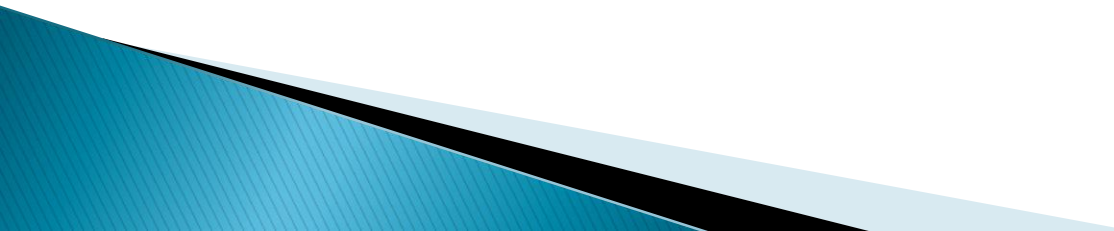
The form of short, summarizing statement.



# Presentation to the end user

- ▶ Due to management demands, short, clear statements are ideal for this.
- ▶ Possible creation:
  - Manual way
  - Automatic way
    - Conversion into SBVR
    - AR2NL System
    - Some new approach...?

# Future work, conclusion

- ▶ Work with dataset Adamek, analysis of students work in Information and Knowledge Processing course
  - ▶ Cooperation with commercial partner, the view from business perspective
  - ▶ Any other possibility to generally determine interesting rules?
- 

Thank you!

Q&A

