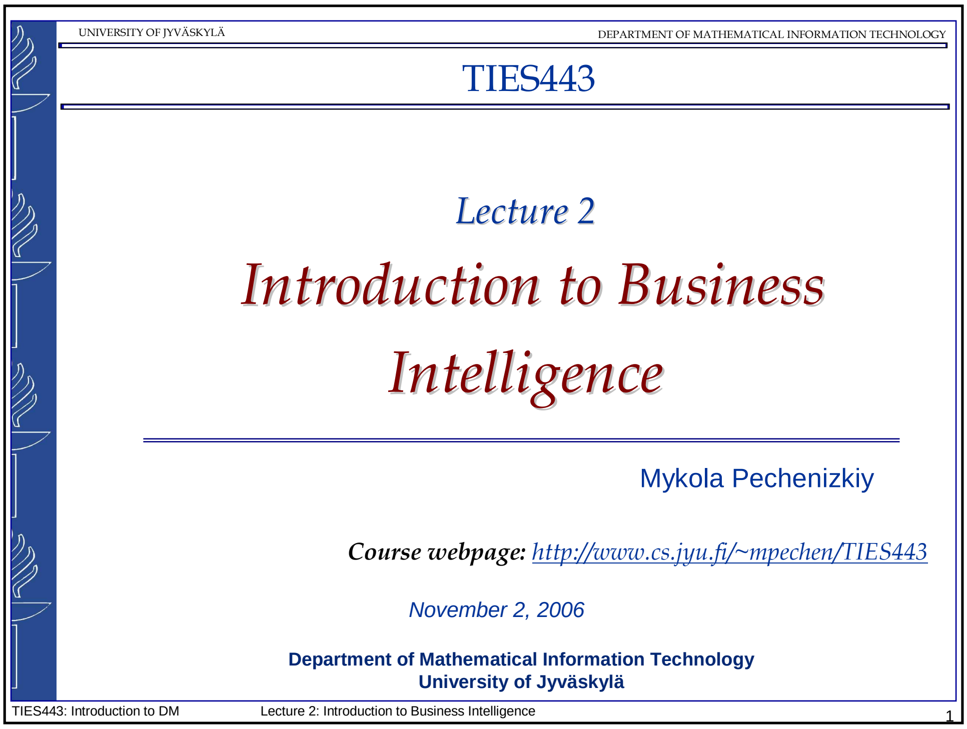


TIES443

Lecture 2

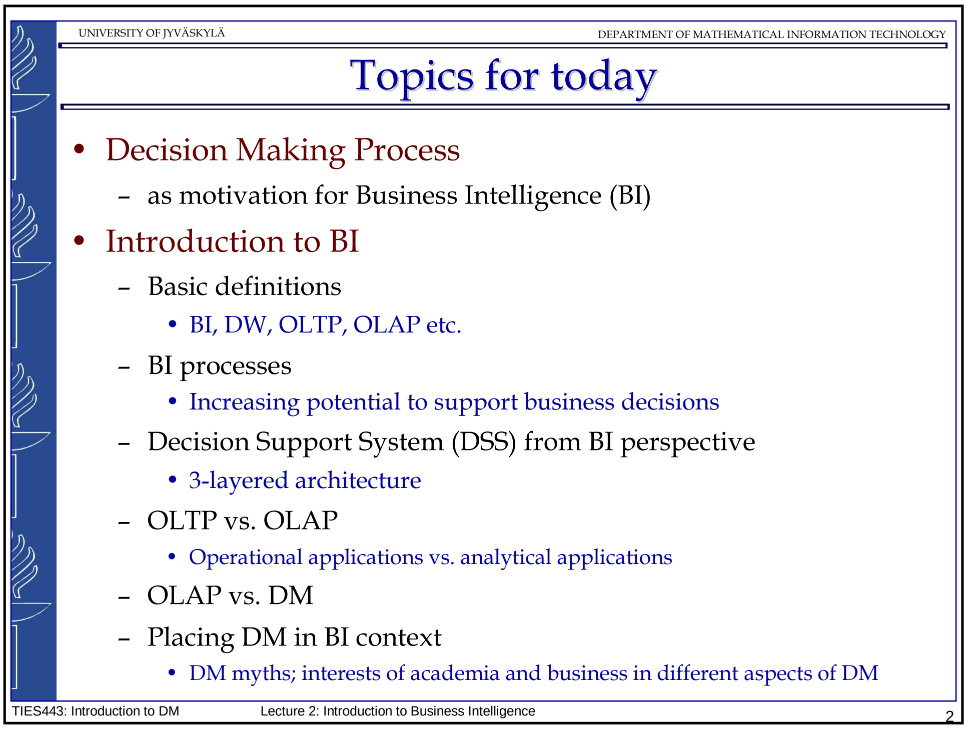
Introduction to Business Intelligence

Mykola Pechenizkiy

Course webpage: *http://www.cs.jyu.fi/~mpechen/TIES443*

November 2, 2006

Department of Mathematical Information Technology
University of Jyväskylä



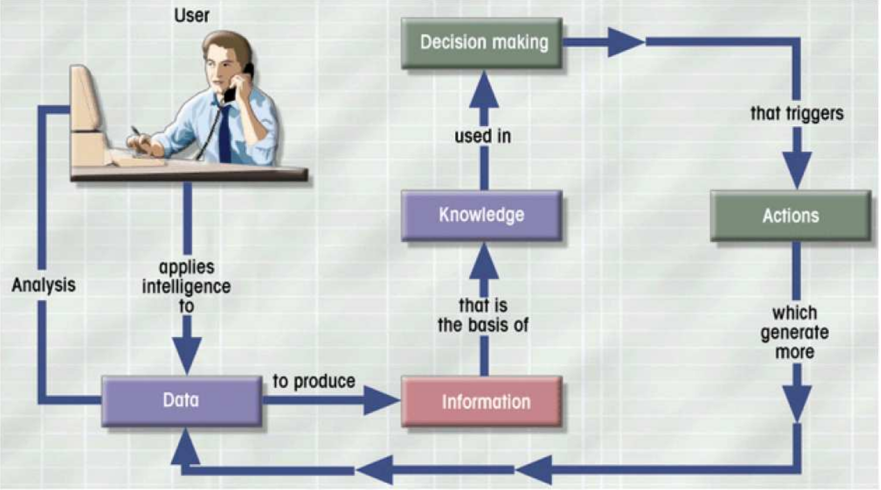
Topics for today

- Decision Making Process
  - as motivation for Business Intelligence (BI)
- Introduction to BI
  - Basic definitions
    - BI, DW, OLTP, OLAP etc.
  - BI processes
    - Increasing potential to support business decisions
  - Decision Support System (DSS) from BI perspective
    - 3-layered architecture
  - OLTP vs. OLAP
    - Operational applications vs. analytical applications
  - OLAP vs. DM
  - Placing DM in BI context
    - DM myths; interests of academia and business in different aspects of DM

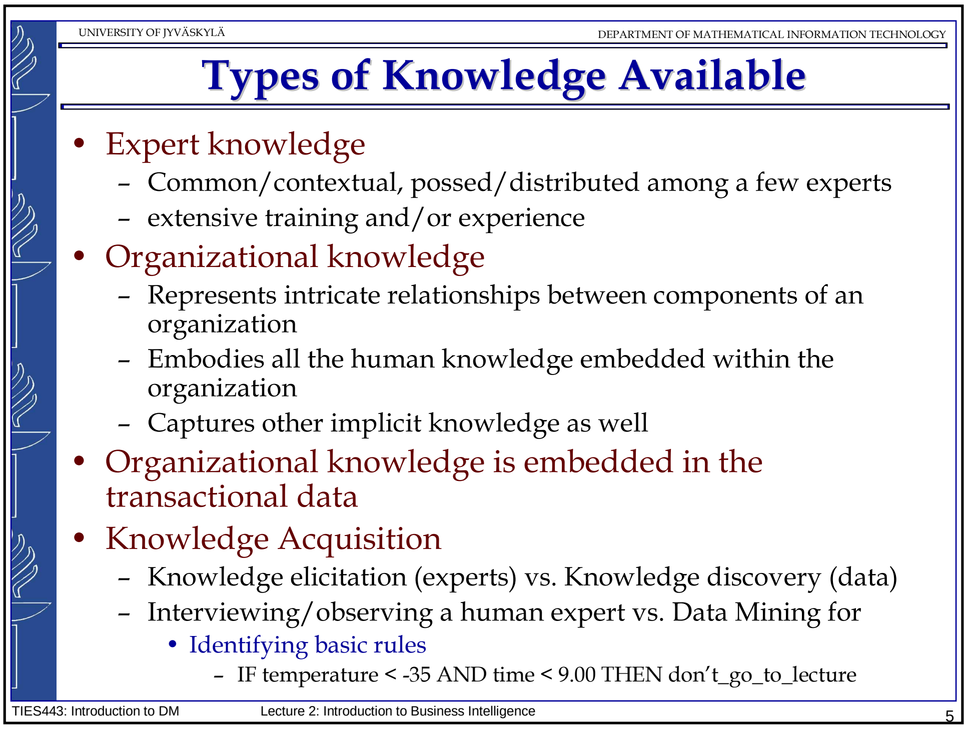
Decision Making Process

- Decision making at different levels
  - Operational
    - Related to daily activities with short-term effect
    - Structured decisions taken by lower management
  - Tactical
    - Semi-structured decisions taken by middle management
  - Strategic
    - Long-term effect
    - Unstructured decisions taken by top management
- Decision making steps include
  - Problem identification,
  - Finding alternative solutions,
  - Making a choice
- Information and knowledge form the backbone of the decision making process

Data-Information-Knowledge-Desision Making - Action cycle

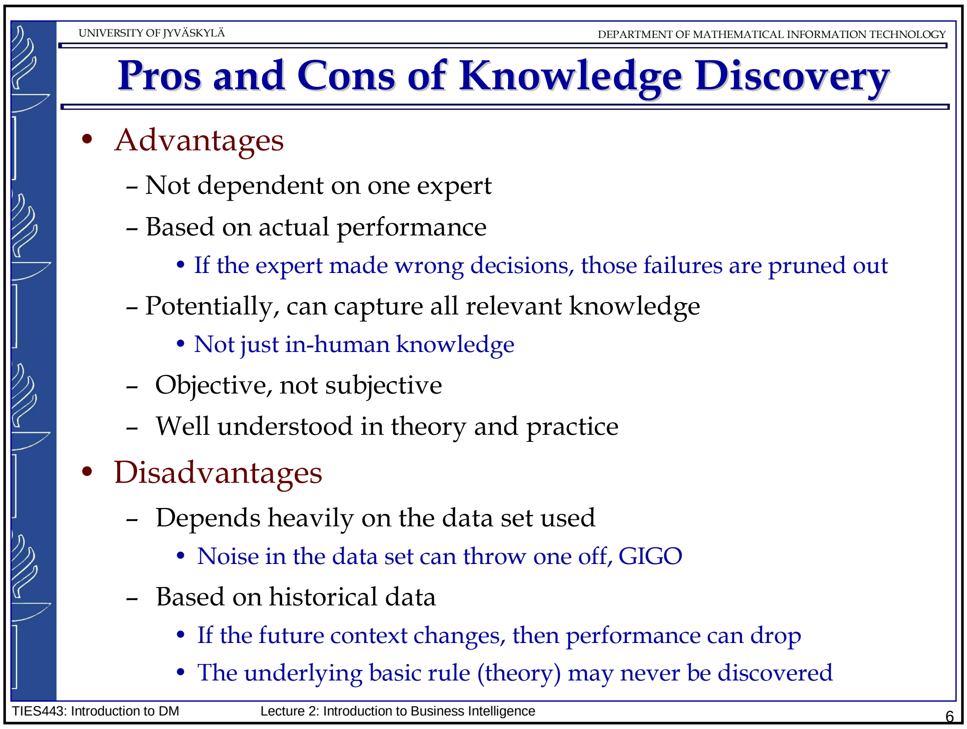


Technology is needed *"… to push information closer to the point of service to enhance decision-making, and to make the data actionable"* – SAS vision of their customers' needs



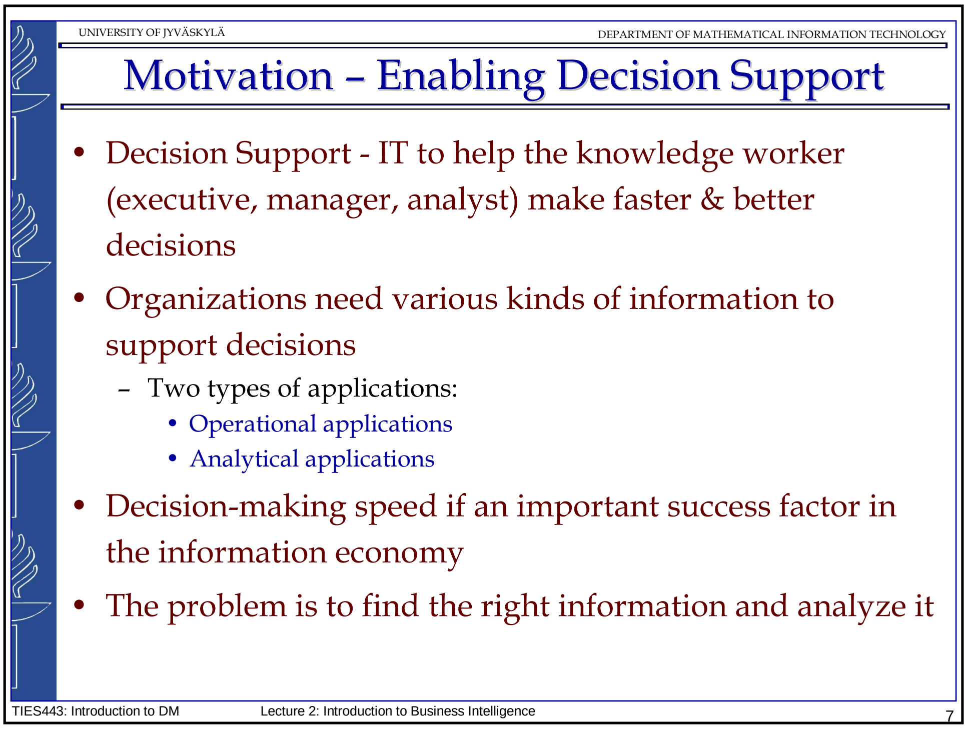
Types of Knowledge Available

- Expert knowledge
  - Common/contextual, possed/distributed among a few experts
  - extensive training and/or experience
- Organizational knowledge
  - Represents intricate relationships between components of an organization
  - Embodies all the human knowledge embedded within the organization
  - Captures other implicit knowledge as well
- Organizational knowledge is embedded in the transactional data
- Knowledge Acquisition
  - Knowledge elicitation (experts) vs. Knowledge discovery (data)
  - Interviewing/observing a human expert vs. Data Mining for
    - Identifying basic rules
      - IF temperature < -35 AND time < 9.00 THEN don't_go_to_lecture



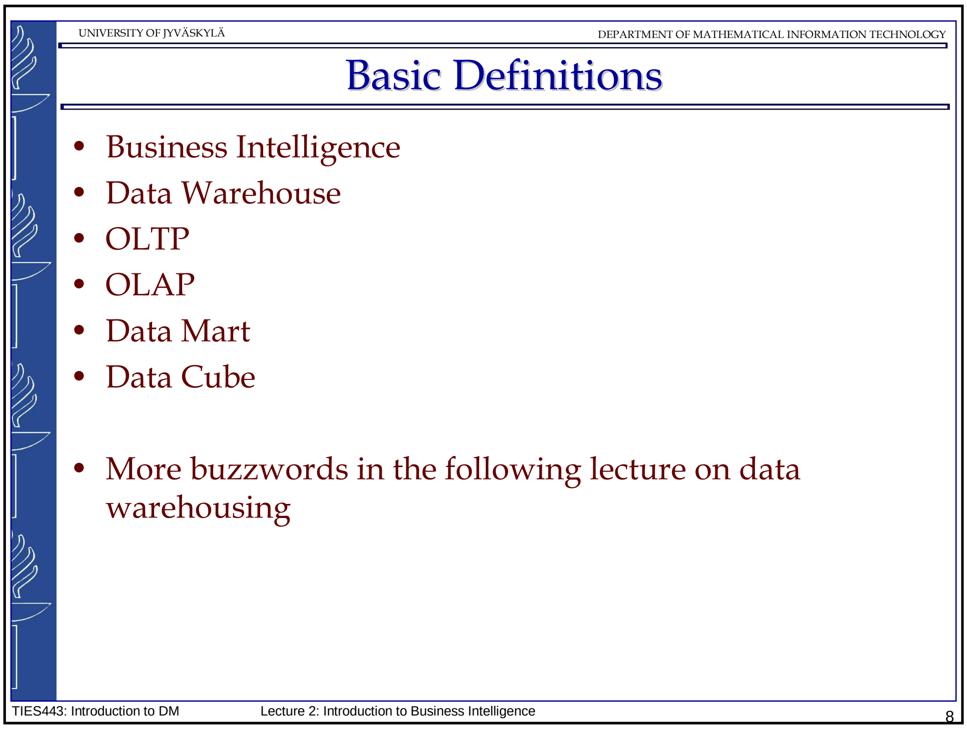
Pros and Cons of Knowledge Discovery

- Advantages
  - Not dependent on one expert
  - Based on actual performance
    - If the expert made wrong decisions, those failures are pruned out
  - Potentially, can capture all relevant knowledge
    - Not just in-human knowledge
  - Objective, not subjective
  - Well understood in theory and practice
- Disadvantages
  - Depends heavily on the data set used
    - Noise in the data set can throw one off, GIGO
  - Based on historical data
    - If the future context changes, then performance can drop
    - The underlying basic rule (theory) may never be discovered



Motivation – Enabling Decision Support

- Decision Support - IT to help the knowledge worker (executive, manager, analyst) make faster & better decisions
- Organizations need various kinds of information to support decisions
  - Two types of applications:
    - Operational applications
    - Analytical applications
- Decision-making speed if an important success factor in the information economy
- The problem is to find the right information and analyze it



Basic Definitions

- Business Intelligence
- Data Warehouse
- OLTP
- OLAP
- Data Mart
- Data Cube

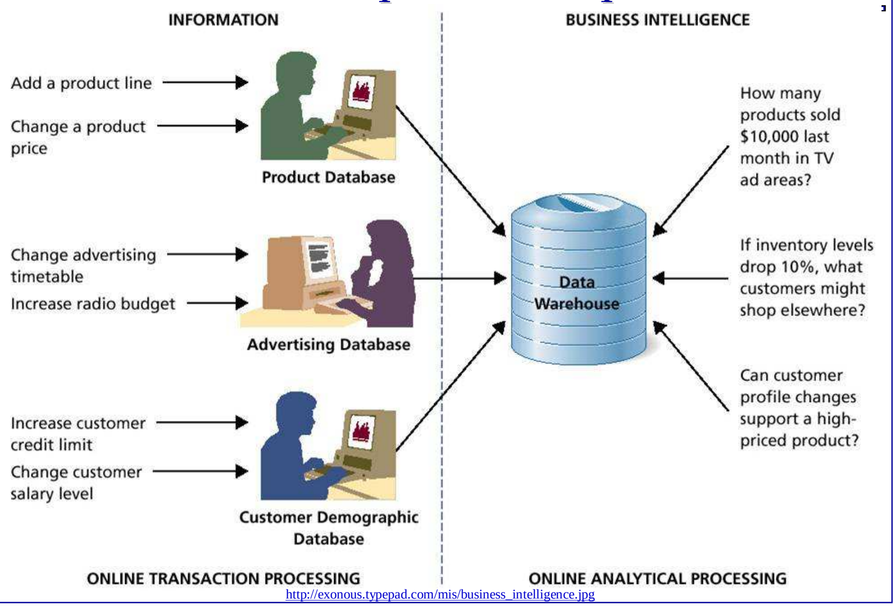
- More buzzwords in the following lecture on data warehousing

What Is Business Intelligence?

- Business Intelligence (BI) is
  - the new technology for understanding the past & predicting the future …
  - a broad category of *technologies* that allows for
    - gathering, storing, accessing & analyzing data to help business users make better decisions
    - analyzing business performance through data-driven insight
  - a broad category of *applications*, which include the activities of
    - decision support systems
    - query and reporting
    - online analytical processing (OLAP)
    - statistical analysis, forecasting, and data mining.

- BI applications can be:
  - mission-critical and integral to an enterprise's operations or occasional to meet a special requirement
  - enterprise-wide or local to one division, department, or project
  - centrally initiated or driven by user demand

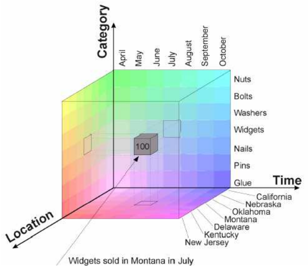
One simple BI example


INFORMATION
BUSINESS INTELLIGENCE
Add a product line
Change a product price
Product Database
Change advertising timetable
Increase radio budget
Advertising Database
Increase customer credit limit
Change customer salary level
Customer Demographic Database
Data Warehouse
How many products sold $10,000 last month in TV ad areas?
If inventory levels drop 10%, what customers might shop elsewhere?
Can customer profile changes support a high-priced product?
ONLINE TRANSACTION PROCESSING
ONLINE ANALYTICAL PROCESSING

http://exonous.typepad.com/mis/business_intelligence.jpg

What is Data Warehouse?

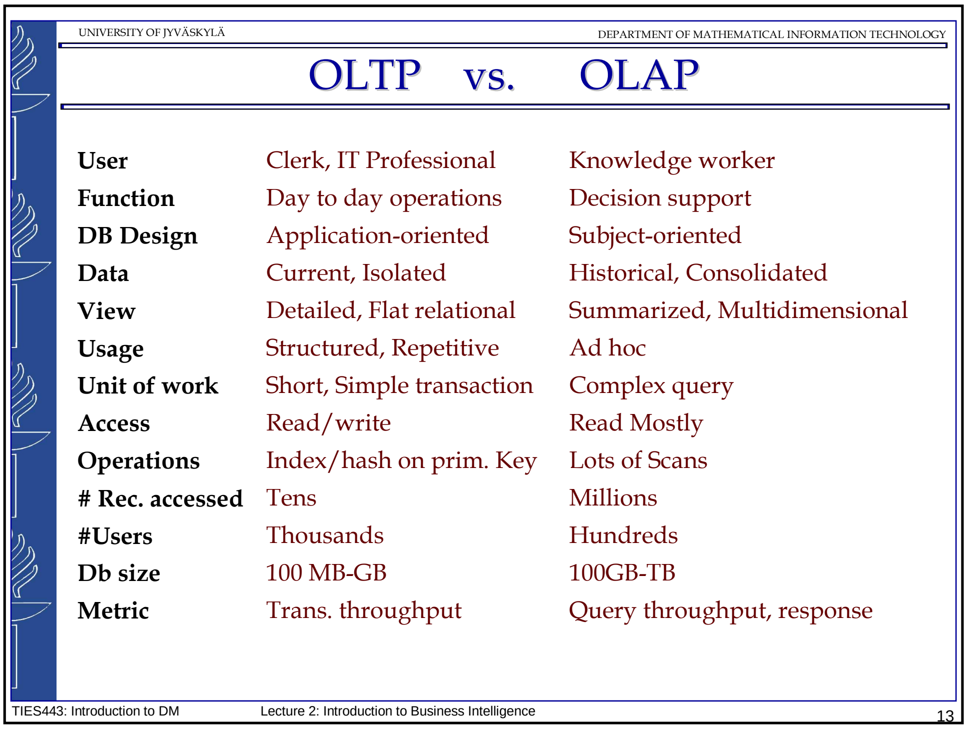
- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - A consistent database source that bring together information from multiple sources for decision support queries
  - Support information processing by providing a solid platform of consolidated, historical data for analysis
- Data warehousing:
  - The process of constructing and using data warehouses
- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- We will consider different aspect of data warehousing in the following lecture tomorrow



Widgets sold in Montana in July

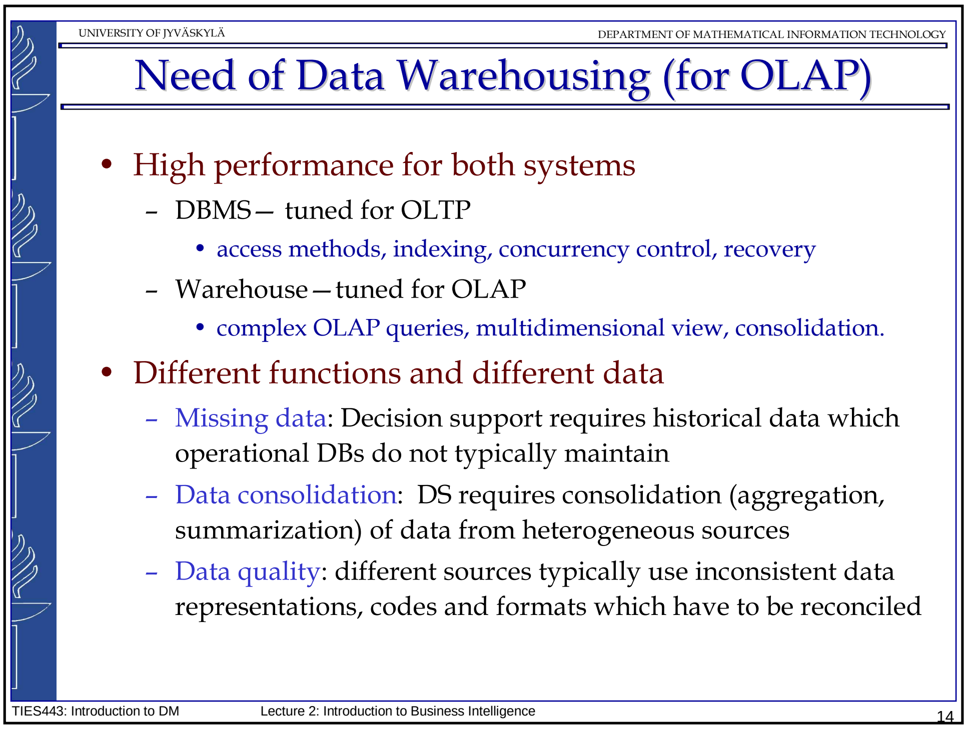
Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
  - Aims at reliable and efficient processing of a large number of transactions and ensuring data consistency
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
  - Aims at efficient multidimensional processing of large data volumes
    - *Fast, interactive answers to large aggregate queries*
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries



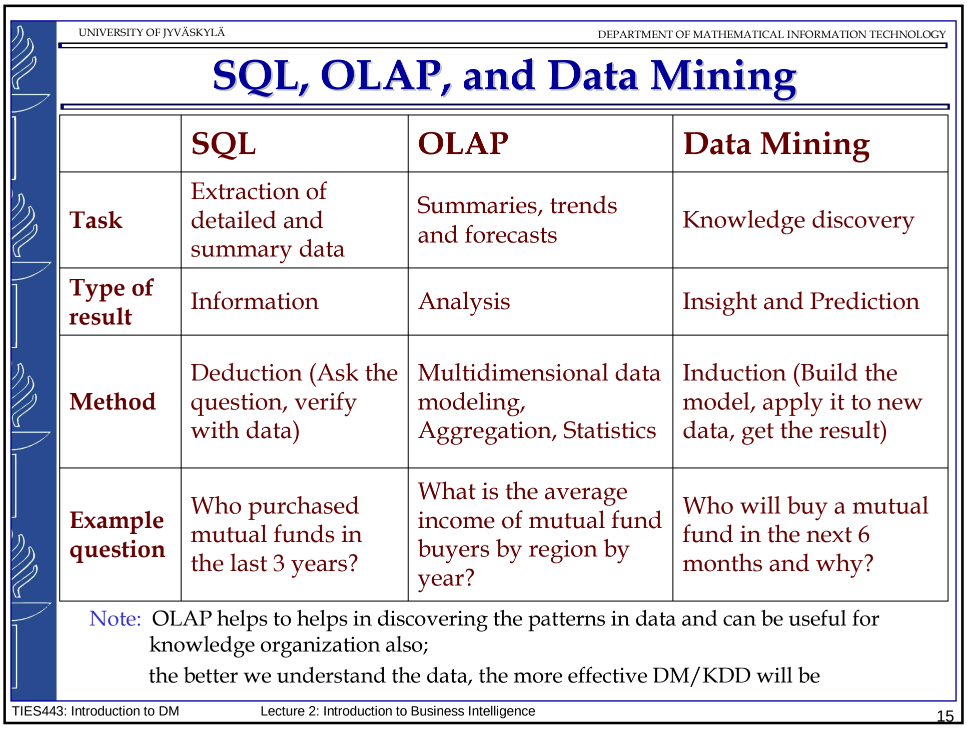
OLTP vs. OLAP

User	Clerk, IT Professional	Knowledge worker
Function	Day to day operations	Decision support
DB Design	Application-oriented	Subject-oriented
Data	Current, Isolated	Historical, Consolidated
View	Detailed, Flat relational	Summarized, Multidimensional
Usage	Structured, Repetitive	Ad hoc
Unit of work	Short, Simple transaction	Complex query
Access	Read/write	Read Mostly
Operations	Index/hash on prim. Key	Lots of Scans
# Rec. accessed	Tens	Millions
#Users	Thousands	Hundreds
Db size	100 MB-GB	100GB-TB
Metric	Trans. throughput	Query throughput, response



Need of Data Warehousing (for OLAP)

- High performance for both systems
  - DBMS— tuned for OLTP
    - access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP
    - complex OLAP queries, multidimensional view, consolidation.
- Different functions and different data
  - Missing data: Decision support requires historical data which operational DBs do not typically maintain
  - Data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - Data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled

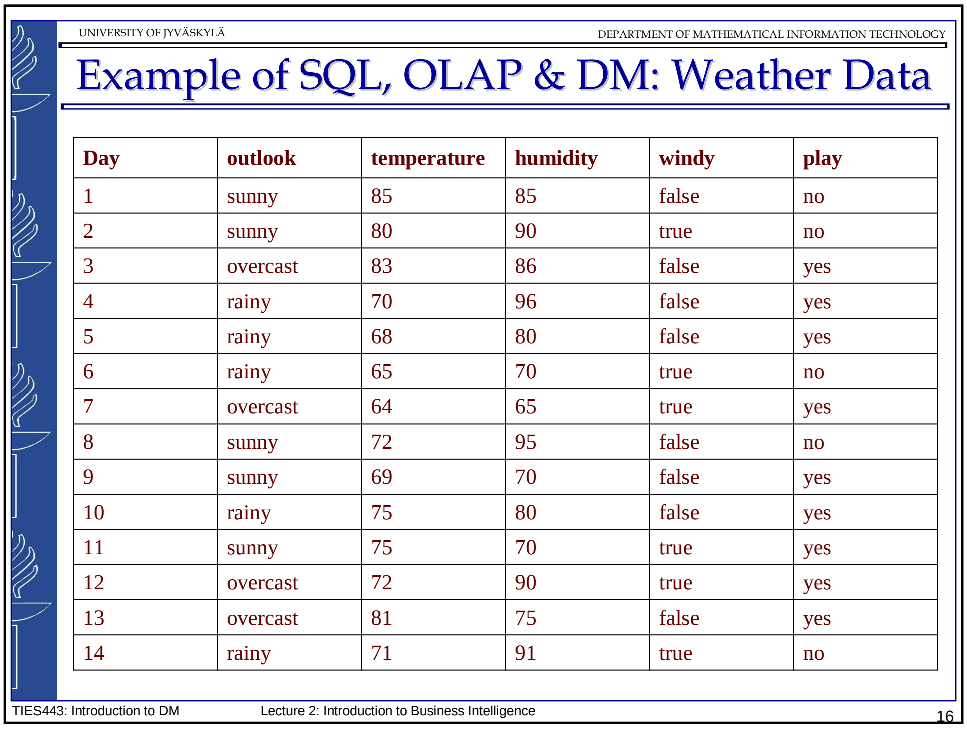


SQL, OLAP, and Data Mining

	SQL	OLAP	Data Mining
Task	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery
Type of result	Information	Analysis	Insight and Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

Note: OLAP helps to helps in discovering the patterns in data and can be useful for knowledge organization also;

the better we understand the data, the more effective DM/KDD will be



Example of SQL, OLAP & DM: Weather Data

Day	outlook	temperature	humidity	windy	play
1	sunny	85	85	false	no
2	sunny	80	90	true	no
3	overcast	83	86	false	yes
4	rainy	70	96	false	yes
5	rainy	68	80	false	yes
6	rainy	65	70	true	no
7	overcast	64	65	true	yes
8	sunny	72	95	false	no
9	sunny	69	70	false	yes
10	rainy	75	80	false	yes
11	sunny	75	70	true	yes
12	overcast	72	90	true	yes
13	overcast	81	75	false	yes
14	rainy	71	91	true	no

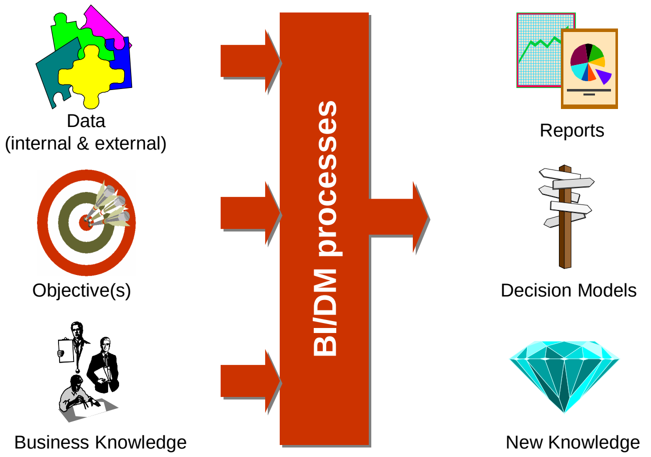
Example of SQL, OLAP & DM: Weather Data

- By querying a DBMS containing the above table we may answer questions like:
  - What was the temperature in the sunny days? {85, 80, 72, 69, 75}
  - Which days the humidity was less than 75? {6, 7, 9, 11}
  - Which days the temperature was greater than 70 and the humidity was less than 75? The intersection of the above two: {11}
- Using OLAP we can create a **Multidimensional Model** of our data (**Data Cube**).
  - E.g. using the dimensions: **time**, **outlook** and **play** we can create the following model.

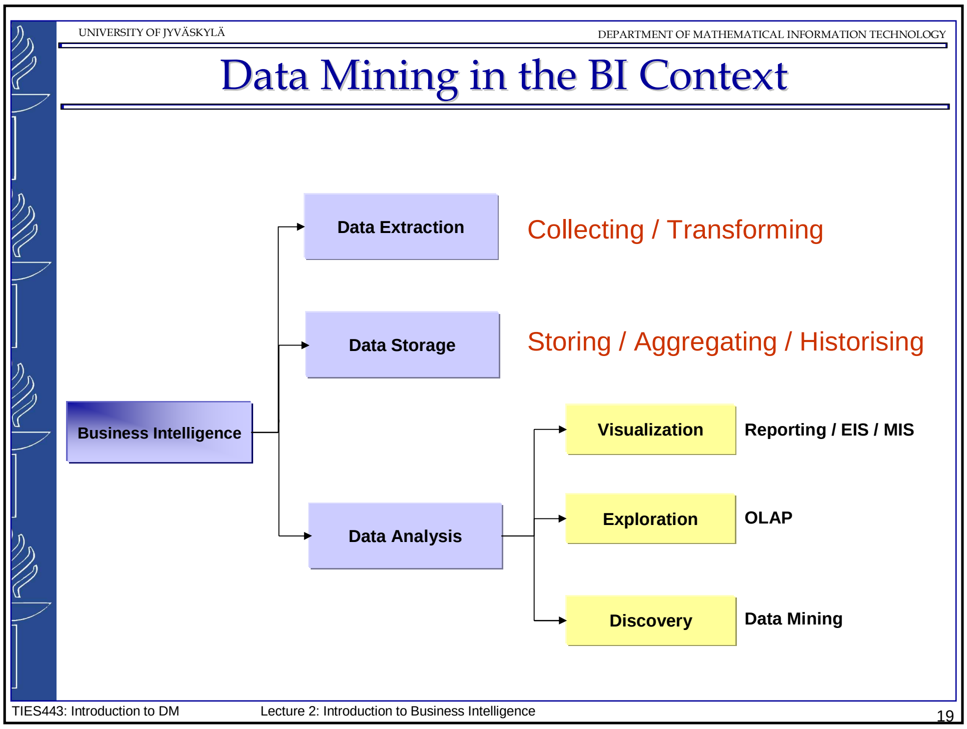
9 / 5	sunny	rainy	overcast
Week 1	0 / 2	2 / 1	2 / 0
Week 2	2 / 1	1 / 1	2 / 0

- Using the DM algorithm (e.g ID3) we can produce the following decision tree:
  - **outlook = sunny**
    - **humidity = high: no**
    - **humidity = normal: yes**
  - **outlook = overcast: yes**
  - **outlook = rainy**
    - **windy = true: no**
    - **windy = false: yes**

Input-Output View



Data Mining is a business-driven process, supported by adequate tools, aimed at the discovery and consistent use of meaningful, profitable knowledge from corporate data



Data Mining in the BI Context

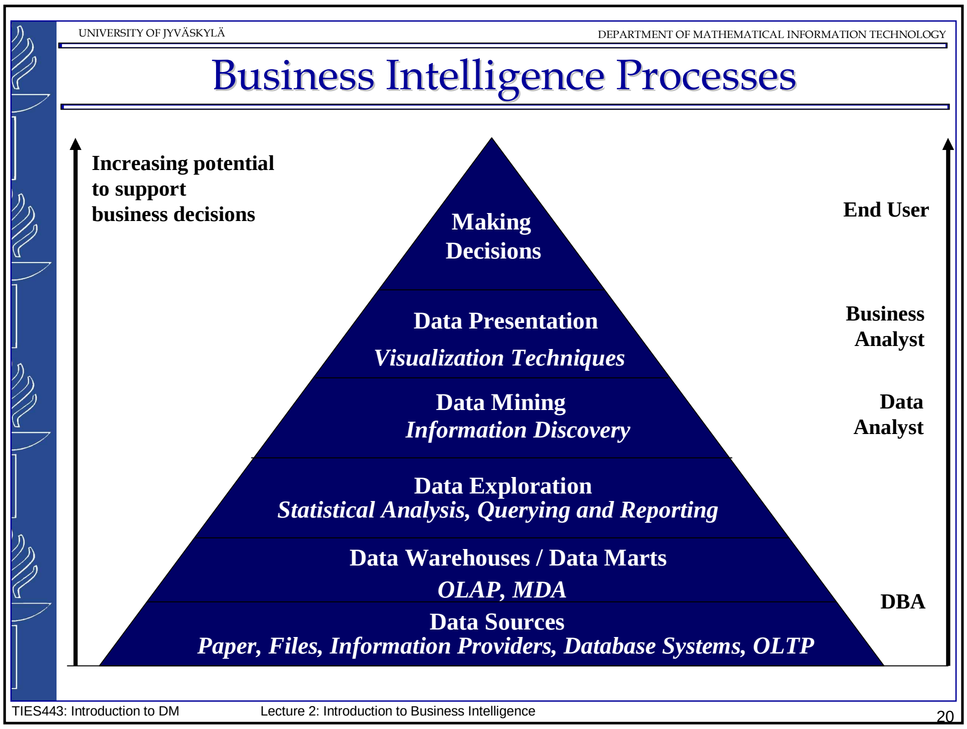
Business Intelligence

Data Extraction — Collecting / Transforming

Data Storage — Storing / Aggregating / Historising

Data Analysis

- Visualization — Reporting / EIS / MIS
- Exploration — OLAP
- Discovery — Data Mining



Business Intelligence Processes

Increasing potential to support business decisions

Making Decisions — End User

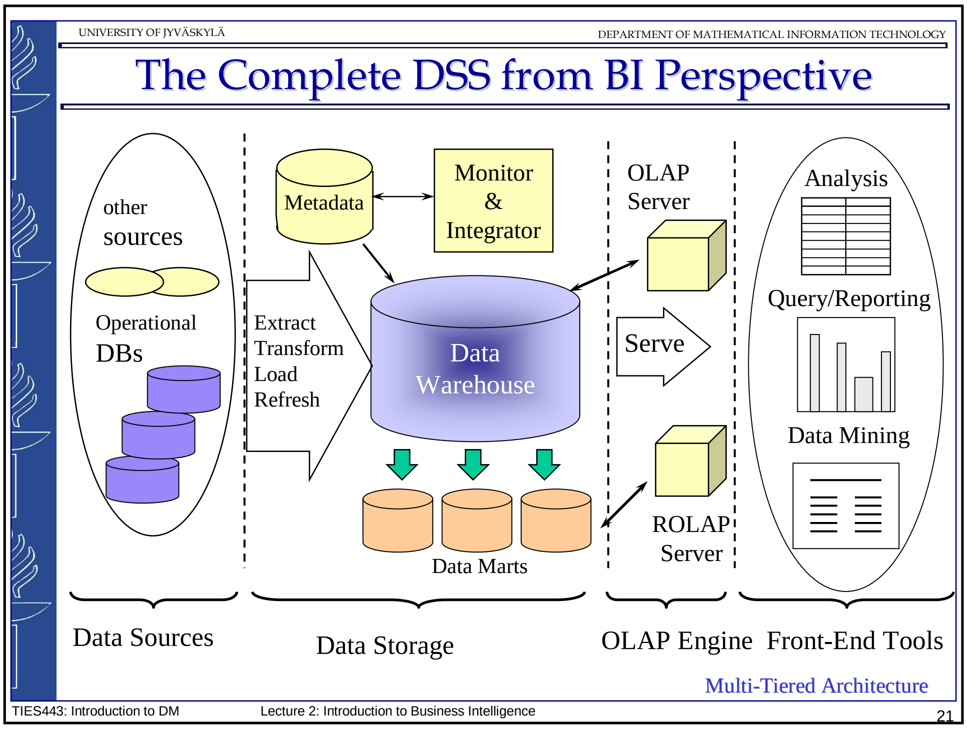
Data Presentation *Visualization Techniques* — Business Analyst

Data Mining *Information Discovery* — Data Analyst

Data Exploration *Statistical Analysis, Querying and Reporting*

Data Warehouses / Data Marts *OLAP, MDA*

Data Sources *Paper, Files, Information Providers, Database Systems, OLTP* — DBA



The Complete DSS from BI Perspective

other sources

Operational DBs

Extract Transform Load Refresh

Metadata

Monitor & Integrator

Data Warehouse

Data Marts

OLAP Server

Serve

ROLAP Server

Analysis

Query/Reporting

Data Mining

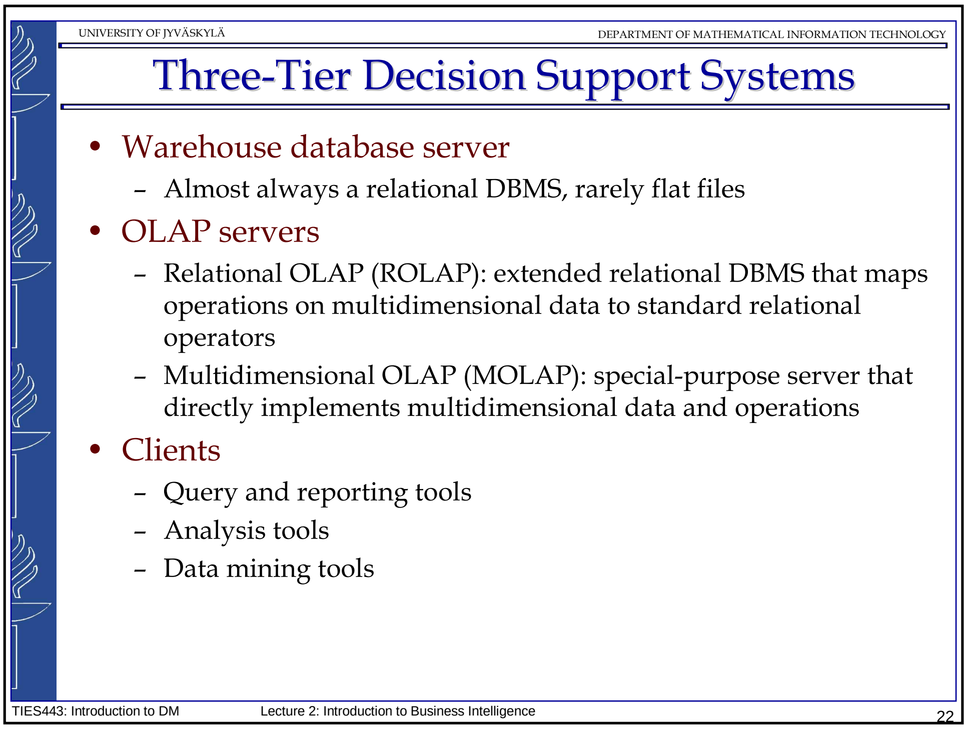
Data Sources

Data Storage

OLAP Engine

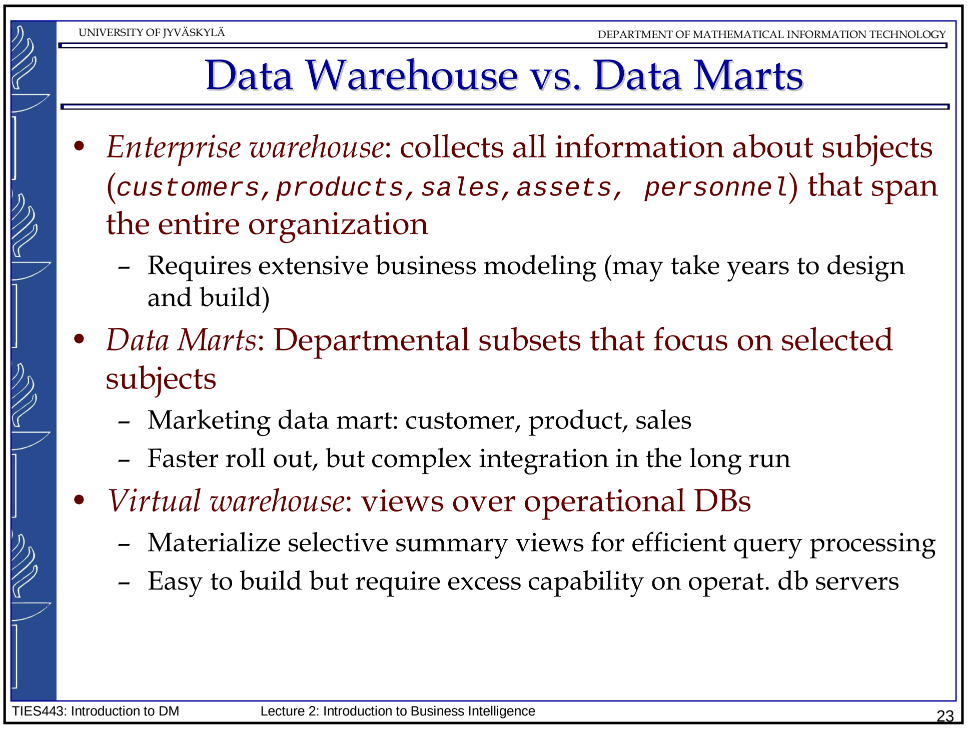
Front-End Tools

Multi-Tiered Architecture



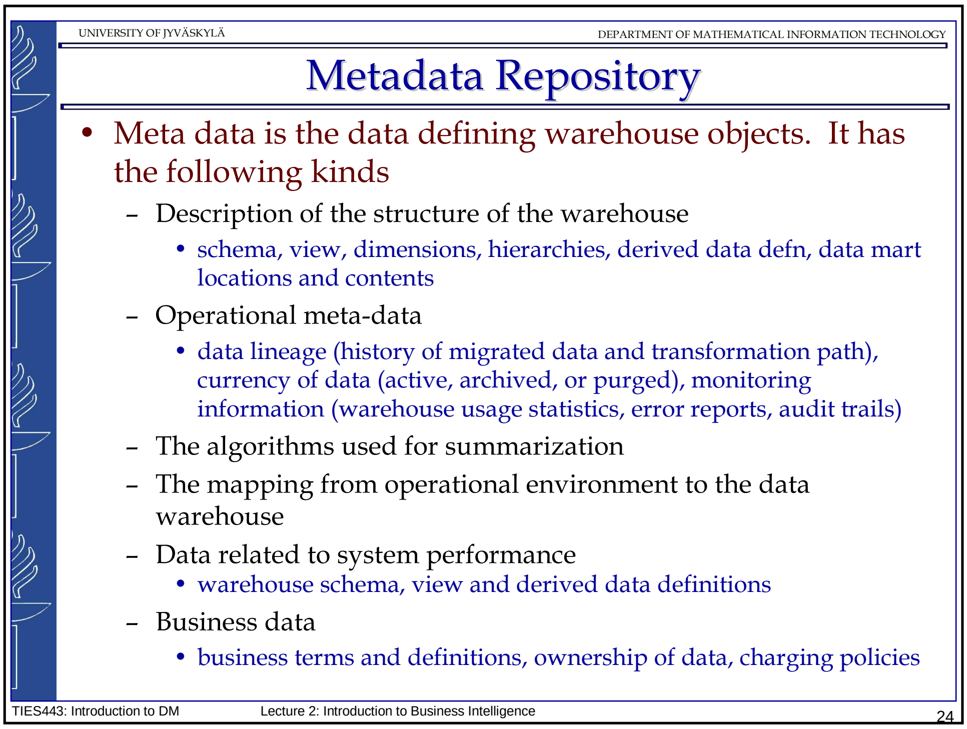
Three-Tier Decision Support Systems

- Warehouse database server
  - Almost always a relational DBMS, rarely flat files
- OLAP servers
  - Relational OLAP (ROLAP): extended relational DBMS that maps operations on multidimensional data to standard relational operators
  - Multidimensional OLAP (MOLAP): special-purpose server that directly implements multidimensional data and operations
- Clients
  - Query and reporting tools
  - Analysis tools
  - Data mining tools



Data Warehouse vs. Data Marts

- *Enterprise warehouse*: collects all information about subjects (`customers,products,sales,assets, personnel`) that span the entire organization
  - Requires extensive business modeling (may take years to design and build)
- *Data Marts*: Departmental subsets that focus on selected subjects
  - Marketing data mart: customer, product, sales
  - Faster roll out, but complex integration in the long run
- *Virtual warehouse*: views over operational DBs
  - Materialize selective summary views for efficient query processing
  - Easy to build but require excess capability on operat. db servers

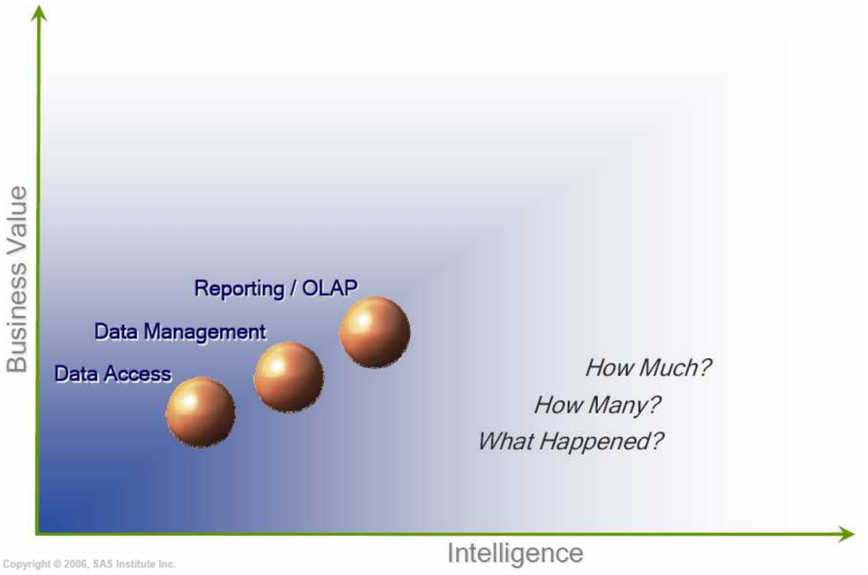


Metadata Repository

- Meta data is the data defining warehouse objects. It has the following kinds
  - Description of the structure of the warehouse
    - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
  - Operational meta-data
    - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
  - The algorithms used for summarization
  - The mapping from operational environment to the data warehouse
  - Data related to system performance
    - warehouse schema, view and derived data definitions
  - Business data
    - business terms and definitions, ownership of data, charging policies

UNIVERSITY OF JYVÄSKYLÄ DEPARTMENT OF MATHEMATICAL INFORMATION TECHNOLOGY

Business Intelligence: An old Definition

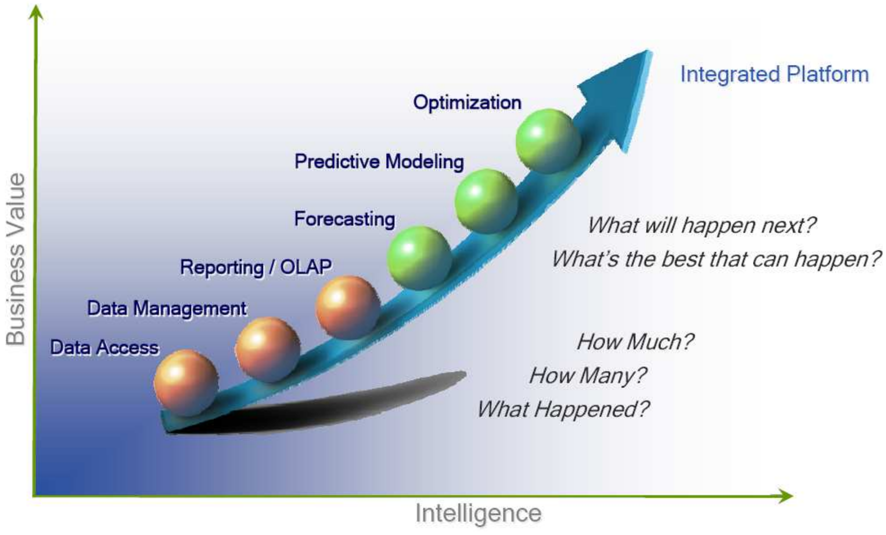


Copyright © 2006, SAS Institute Inc.

TIES443: Introduction to DM Lecture 2: Introduction to Business Intelligence 25

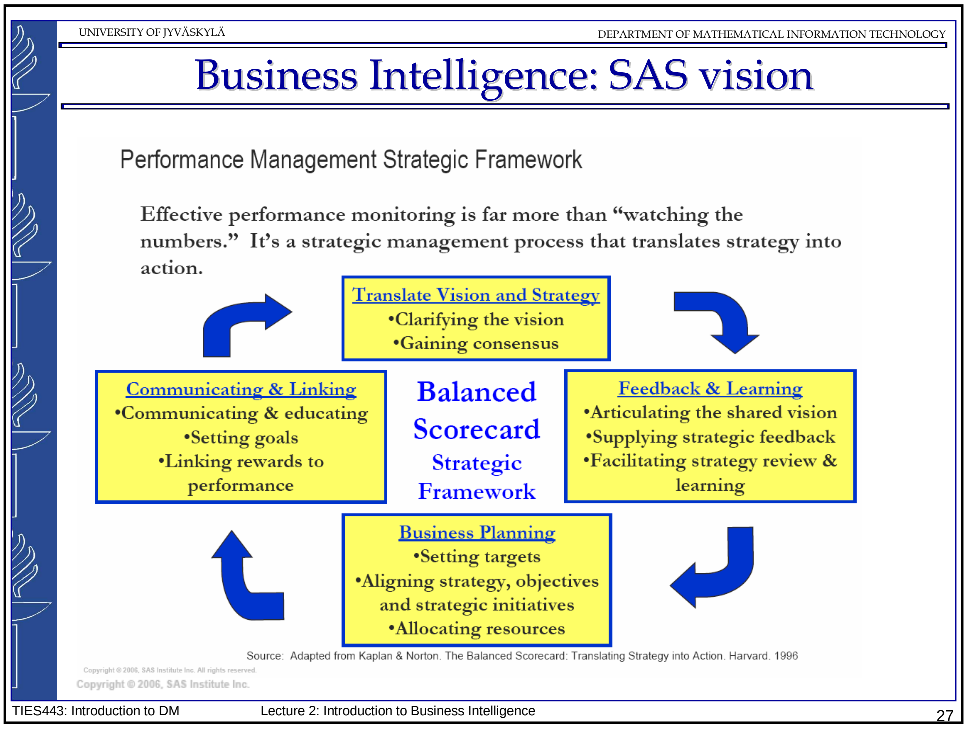
UNIVERSITY OF JYVÄSKYLÄ DEPARTMENT OF MATHEMATICAL INFORMATION TECHNOLOGY

Business Intelligence: SAS vision



Copyright © 2006, SAS Institute Inc.

TIES443: Introduction to DM Lecture 2: Introduction to Business Intelligence 26



Business Intelligence: SAS vision

Performance Management Strategic Framework

Effective performance monitoring is far more than "watching the numbers." It's a strategic management process that translates strategy into action.

Translate Vision and Strategy
- Clarifying the vision
- Gaining consensus

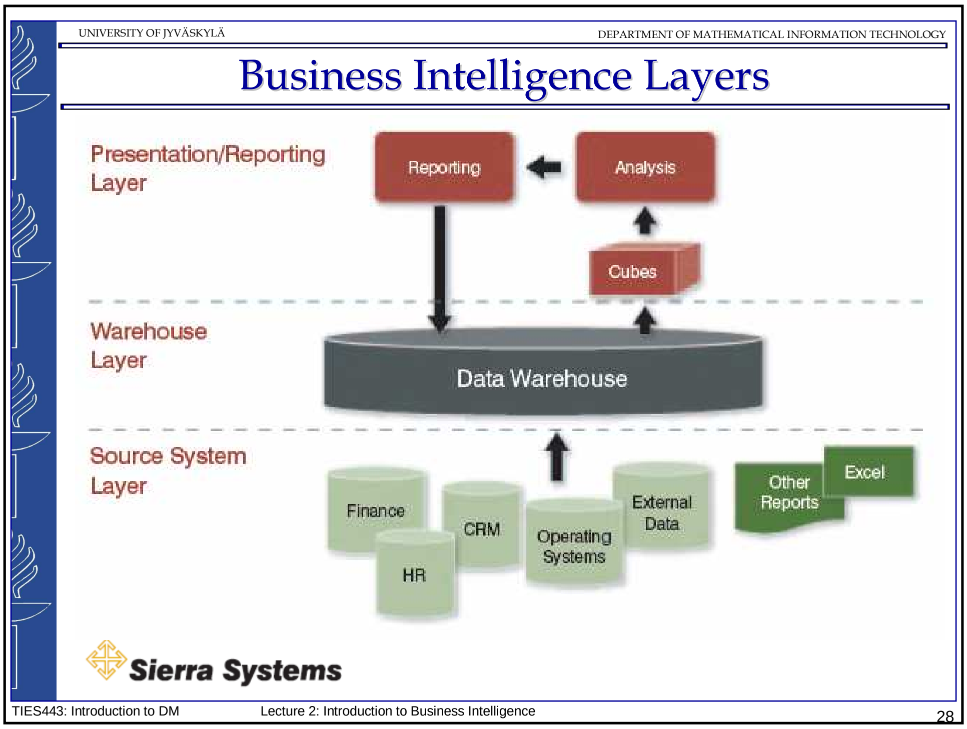
Communicating & Linking
- Communicating & educating
- Setting goals
- Linking rewards to performance

Balanced Scorecard Strategic Framework

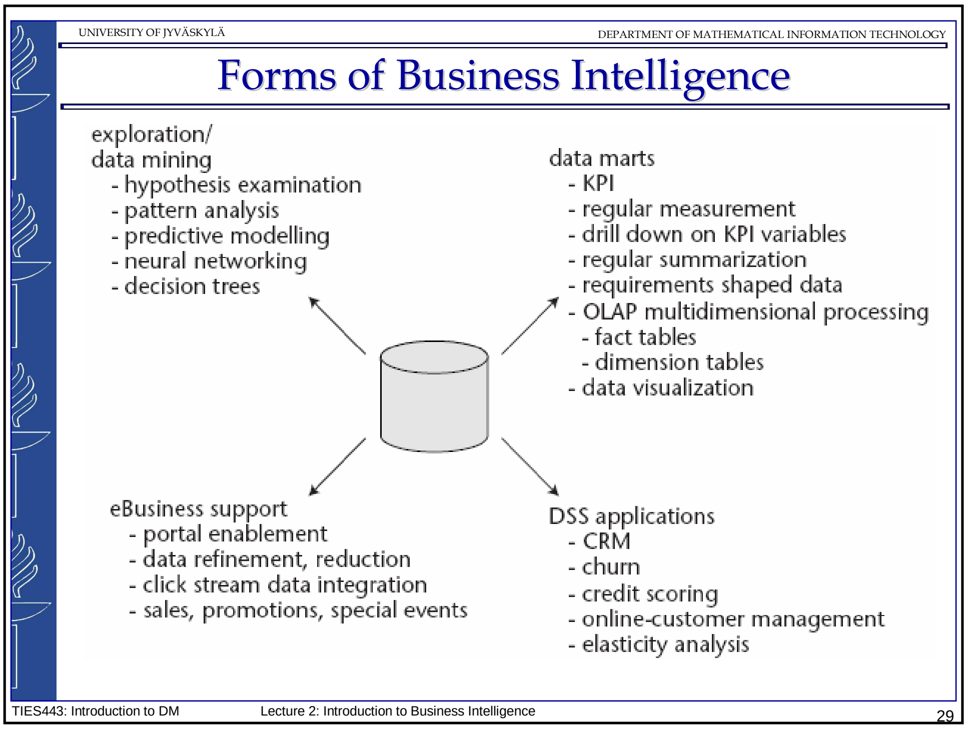
Feedback & Learning
- Articulating the shared vision
- Supplying strategic feedback
- Facilitating strategy review & learning

Business Planning
- Setting targets
- Aligning strategy, objectives and strategic initiatives
- Allocating resources

Source: Adapted from Kaplan & Norton. The Balanced Scorecard: Translating Strategy into Action. Harvard. 1996



Business Intelligence Layers



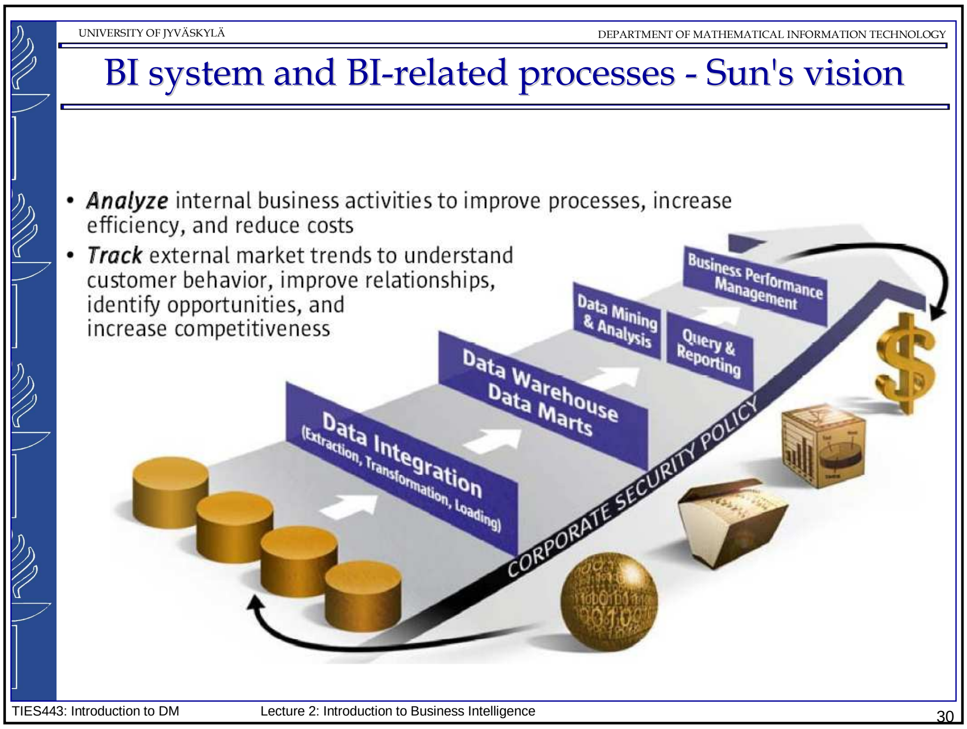
Forms of Business Intelligence

exploration/
data mining
- hypothesis examination
- pattern analysis
- predictive modelling
- neural networking
- decision trees

data marts
- KPI
- regular measurement
- drill down on KPI variables
- regular summarization
- requirements shaped data
- OLAP multidimensional processing
  - fact tables
  - dimension tables
- data visualization

eBusiness support
- portal enablement
- data refinement, reduction
- click stream data integration
- sales, promotions, special events

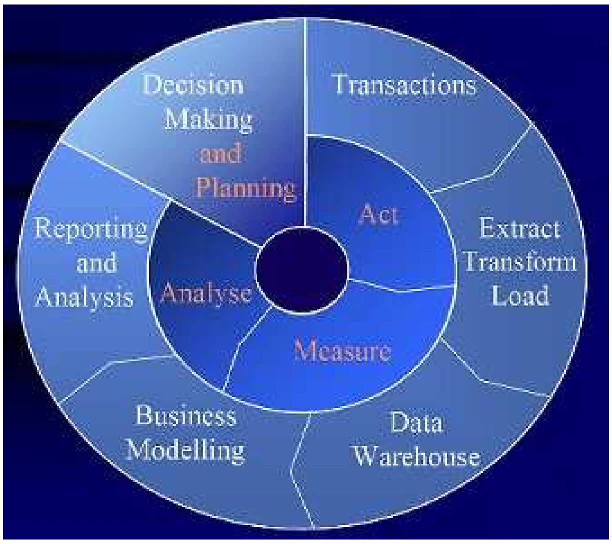
DSS applications
- CRM
- churn
- credit scoring
- online-customer management
- elasticity analysis



BI system and BI-related processes - Sun's vision

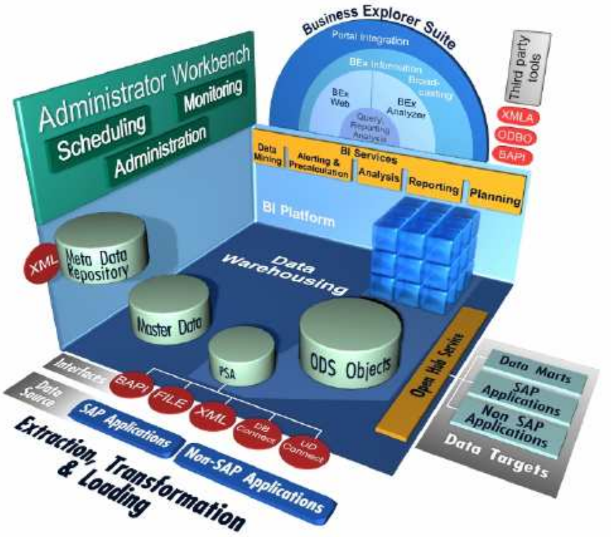
- ***Analyze*** internal business activities to improve processes, increase efficiency, and reduce costs
- ***Track*** external market trends to understand customer behavior, improve relationships, identify opportunities, and increase competitiveness

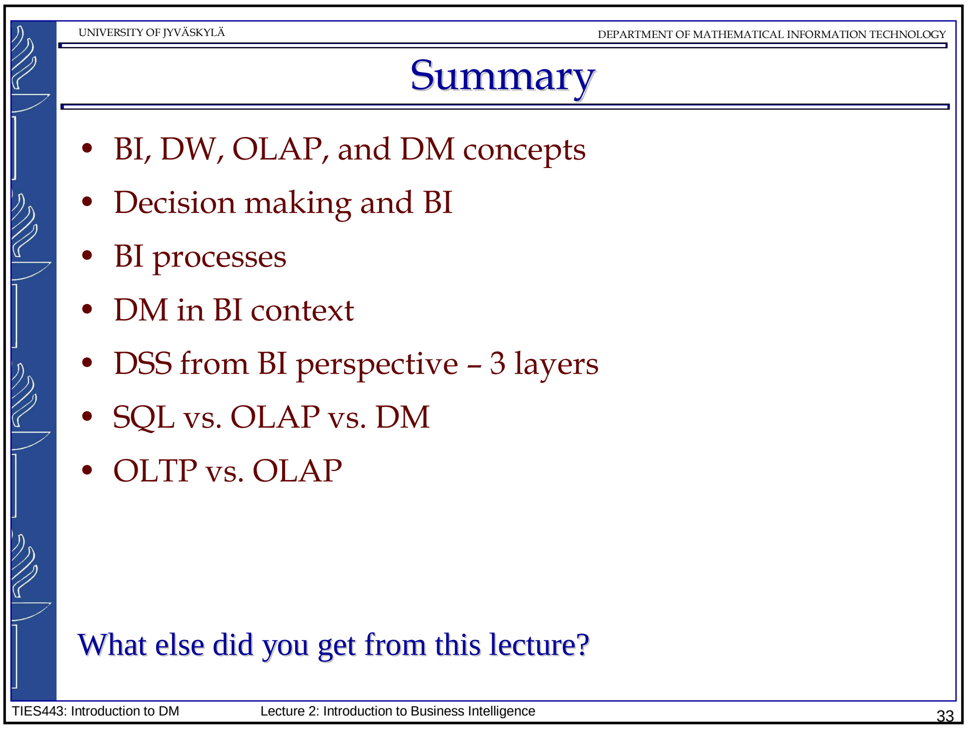
Business Intelligence Cycle



www.isa.co.uk/bi_portal.htm

Business Intelligence

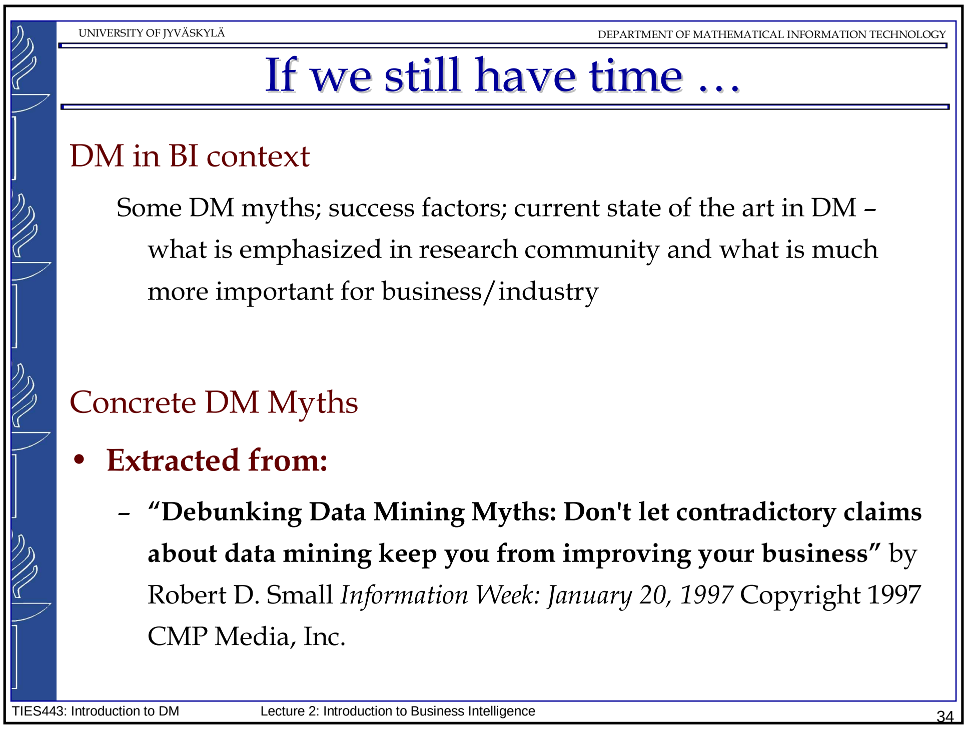




Summary

- BI, DW, OLAP, and DM concepts
- Decision making and BI
- BI processes
- DM in BI context
- DSS from BI perspective – 3 layers
- SQL vs. OLAP vs. DM
- OLTP vs. OLAP

What else did you get from this lecture?



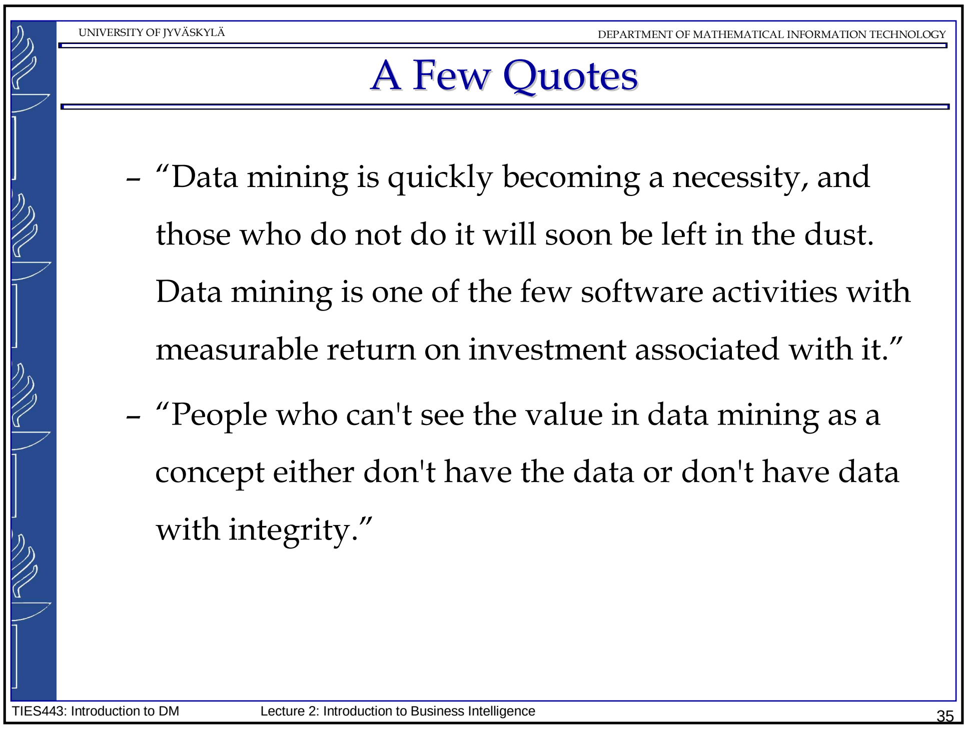
If we still have time …

DM in BI context

Some DM myths; success factors; current state of the art in DM – what is emphasized in research community and what is much more important for business/industry

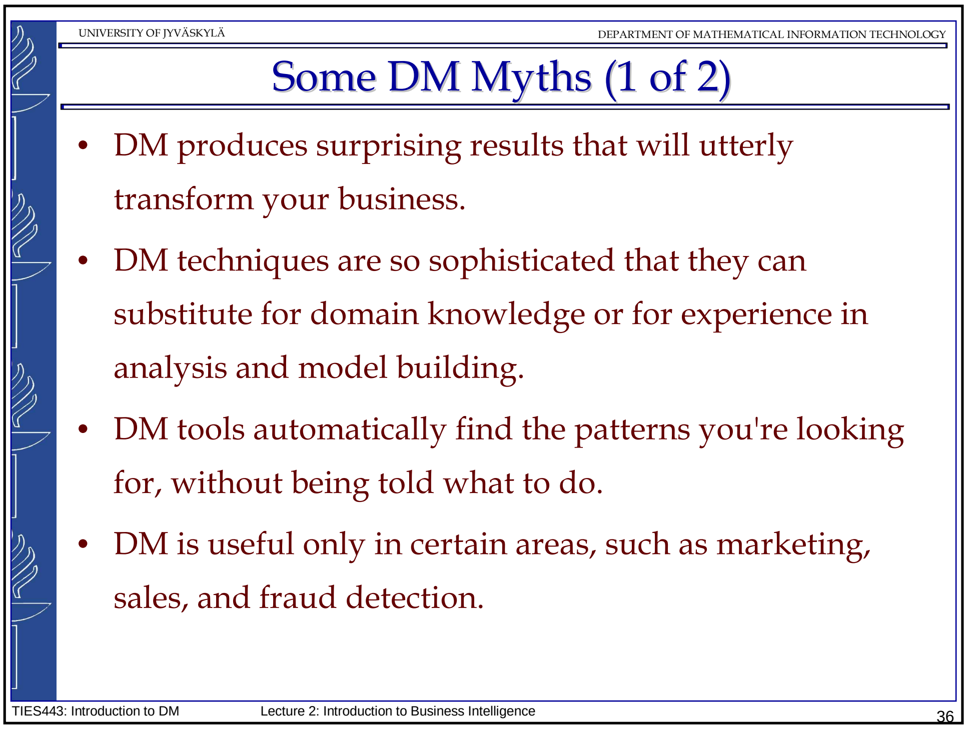
Concrete DM Myths

- **Extracted from:**
  - **“Debunking Data Mining Myths: Don't let contradictory claims about data mining keep you from improving your business”** by Robert D. Small *Information Week: January 20, 1997* Copyright 1997 CMP Media, Inc.



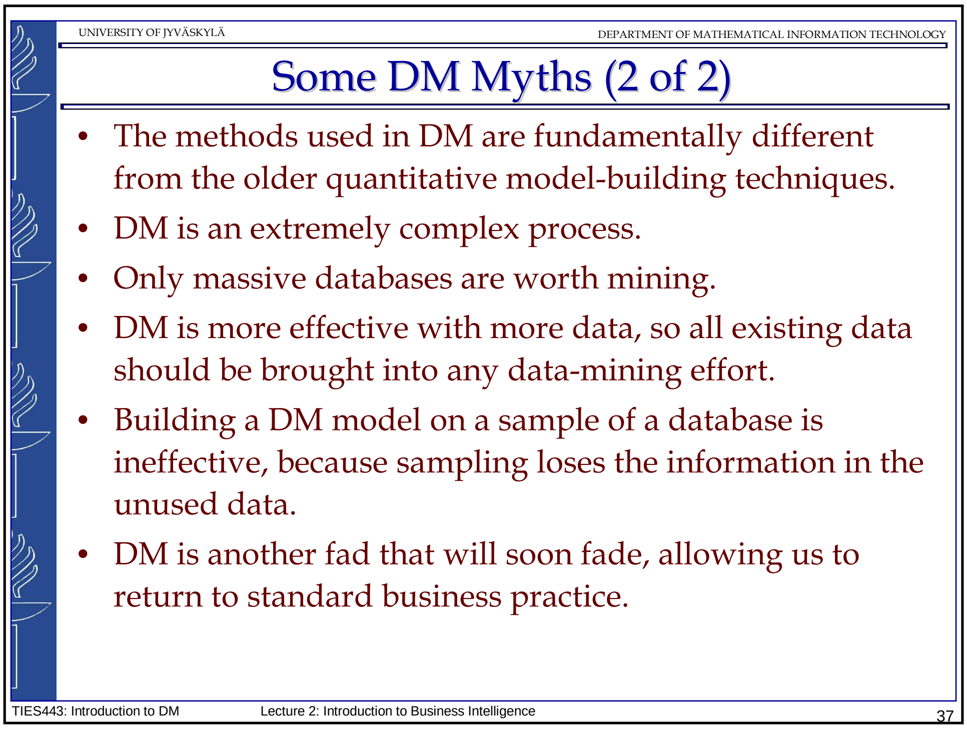
A Few Quotes

- "Data mining is quickly becoming a necessity, and those who do not do it will soon be left in the dust. Data mining is one of the few software activities with measurable return on investment associated with it."
- "People who can't see the value in data mining as a concept either don't have the data or don't have data with integrity."



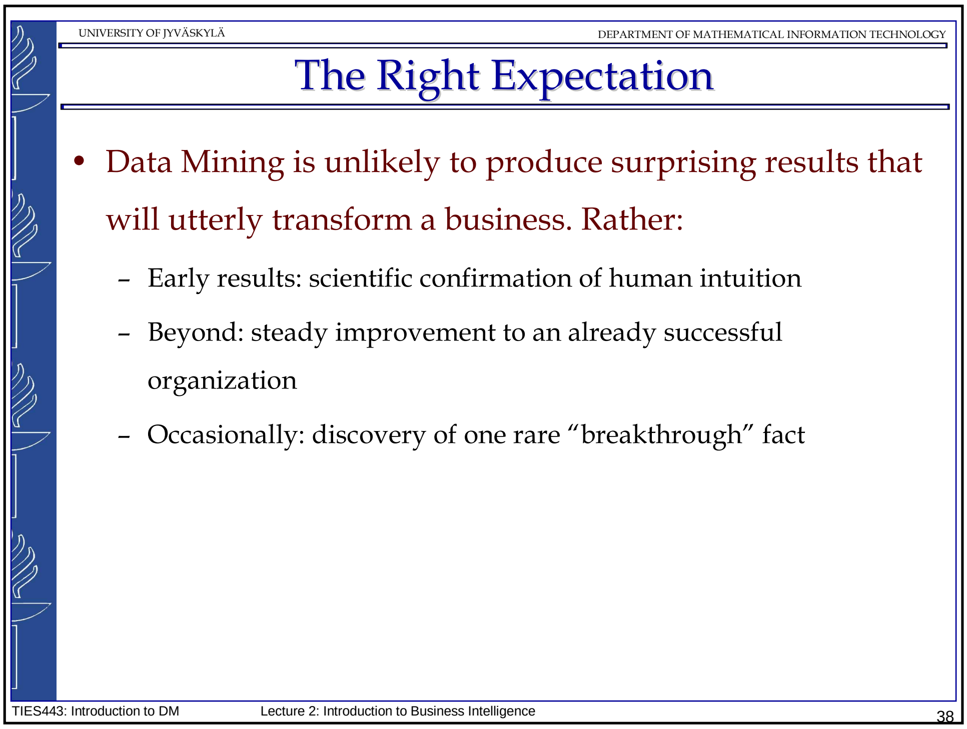
Some DM Myths (1 of 2)

- DM produces surprising results that will utterly transform your business.
- DM techniques are so sophisticated that they can substitute for domain knowledge or for experience in analysis and model building.
- DM tools automatically find the patterns you're looking for, without being told what to do.
- DM is useful only in certain areas, such as marketing, sales, and fraud detection.



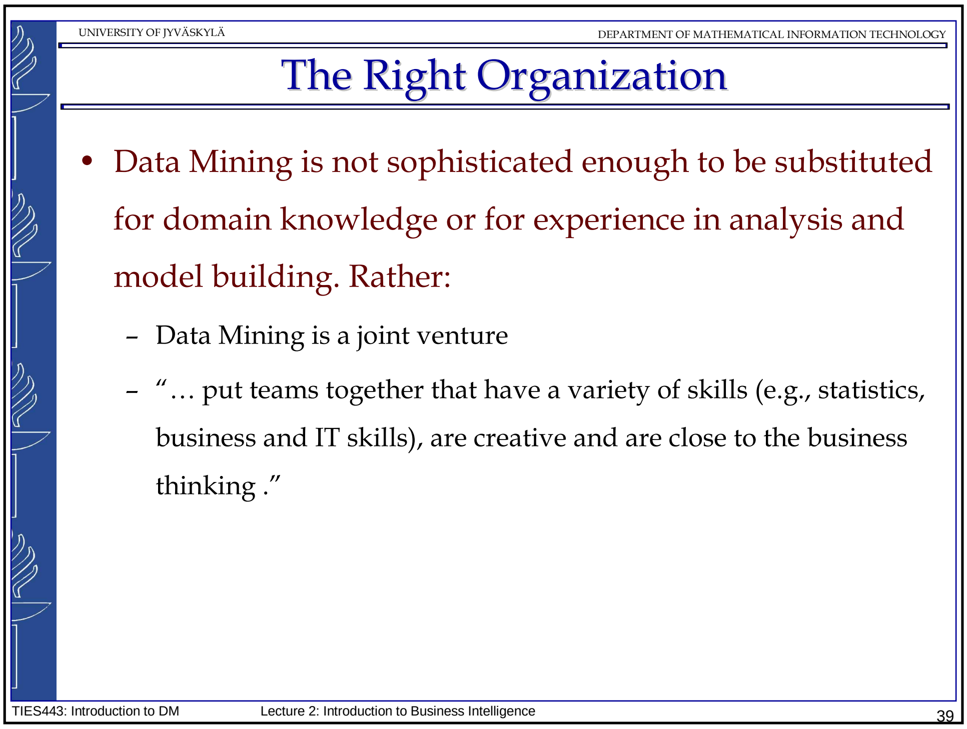
Some DM Myths (2 of 2)

- The methods used in DM are fundamentally different from the older quantitative model-building techniques.
- DM is an extremely complex process.
- Only massive databases are worth mining.
- DM is more effective with more data, so all existing data should be brought into any data-mining effort.
- Building a DM model on a sample of a database is ineffective, because sampling loses the information in the unused data.
- DM is another fad that will soon fade, allowing us to return to standard business practice.



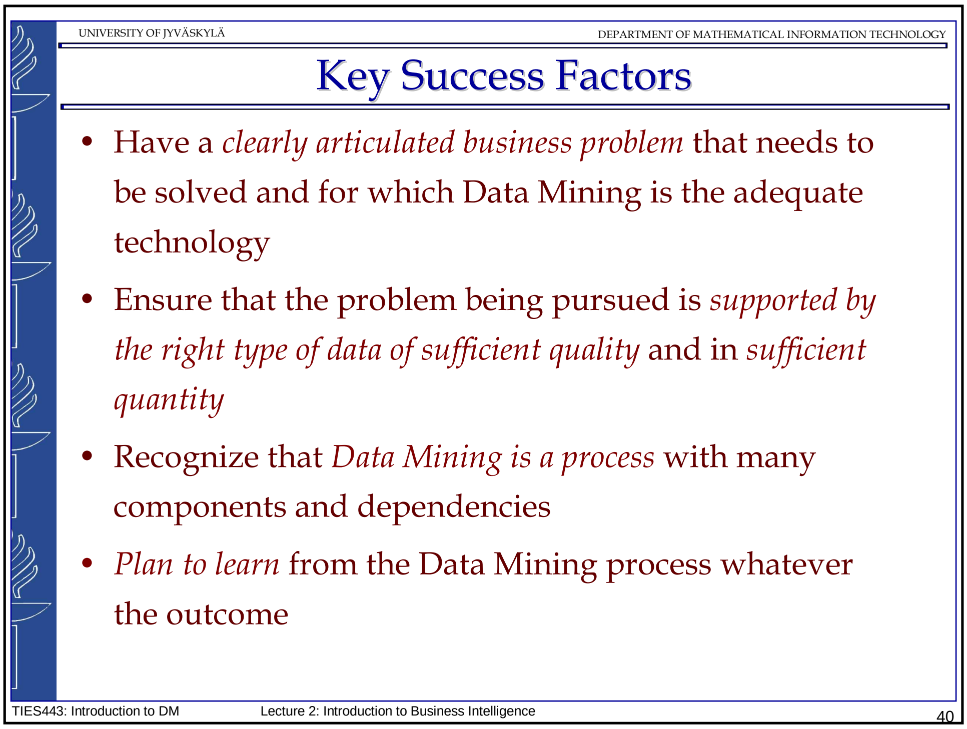
The Right Expectation

- Data Mining is unlikely to produce surprising results that will utterly transform a business. Rather:
  - Early results: scientific confirmation of human intuition
  - Beyond: steady improvement to an already successful organization
  - Occasionally: discovery of one rare “breakthrough” fact



The Right Organization

- Data Mining is not sophisticated enough to be substituted for domain knowledge or for experience in analysis and model building. Rather:
  - Data Mining is a joint venture
  - "… put teams together that have a variety of skills (e.g., statistics, business and IT skills), are creative and are close to the business thinking ."



Key Success Factors

- Have a *clearly articulated business problem* that needs to be solved and for which Data Mining is the adequate technology
- Ensure that the problem being pursued is *supported by the right type of data of sufficient quality* and in *sufficient quantity*
- Recognize that *Data Mining is a process* with many components and dependencies
- *Plan to learn* from the Data Mining process whatever the outcome

DM – state of the art

- DM is still a technology having great expectations to enable organizations to take more benefit of their huge databases.
- There exist some success stories where organizations have managed to have competitive advantage of DM.
- Still the strong focus of most DM-researchers in technology-oriented topics does not support expanding the scope in less rigorous but practically very relevant sub-areas.
- Research in the IS discipline has strong traditions to take into account human and organizational aspects of systems beside the technical ones.

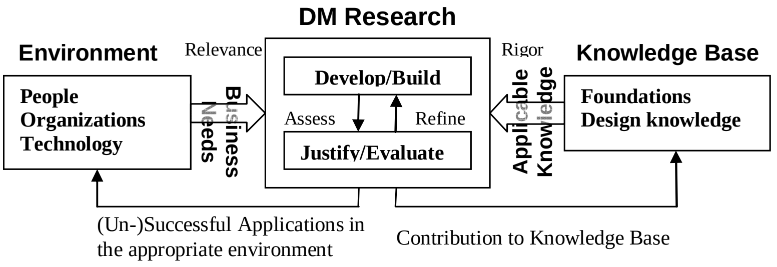
DM – state of the art (cont.)

- Currently the maturation of DM-supporting processes which would take into account human and organizational aspects is still living its childhood.
- DM community might benefit, at least from the practical point of view, looking at some other older sub-areas of IT having traditions to consider solution-driven concepts with a focus also on human and organizational aspects.
- The DM community by becoming more amenable to research results of the IS community might be able to increase its collective understanding of
  - how DM artifacts are developed – conceived, constructed, and implemented,
  - how DM artifacts are used, supported and evolved,
  - how DM artifacts impact and are impacted by the contexts in which they are embedded.

So, where are we?

- a new successful industry (as DM) can follow consecutive phases:
  1. discovering a new idea,
  2. ensuring its applicability,
  3. producing small-scale systems to test the market,
  4. better understanding of new technology and
  5. producing a fully scaled system.
- At the present moment there are several dozens of DM systems, none of which can be compared to the scale of a DBMS system.
  – This fact indicates that we are still in the 3rd phase in the DM area!

DM: Academy vs. Industry



DM Research

Environment

Relevance

Rigor

Knowledge Base

People
Organizations
Technology

Business Needs

Develop/Build

Assess

Refine

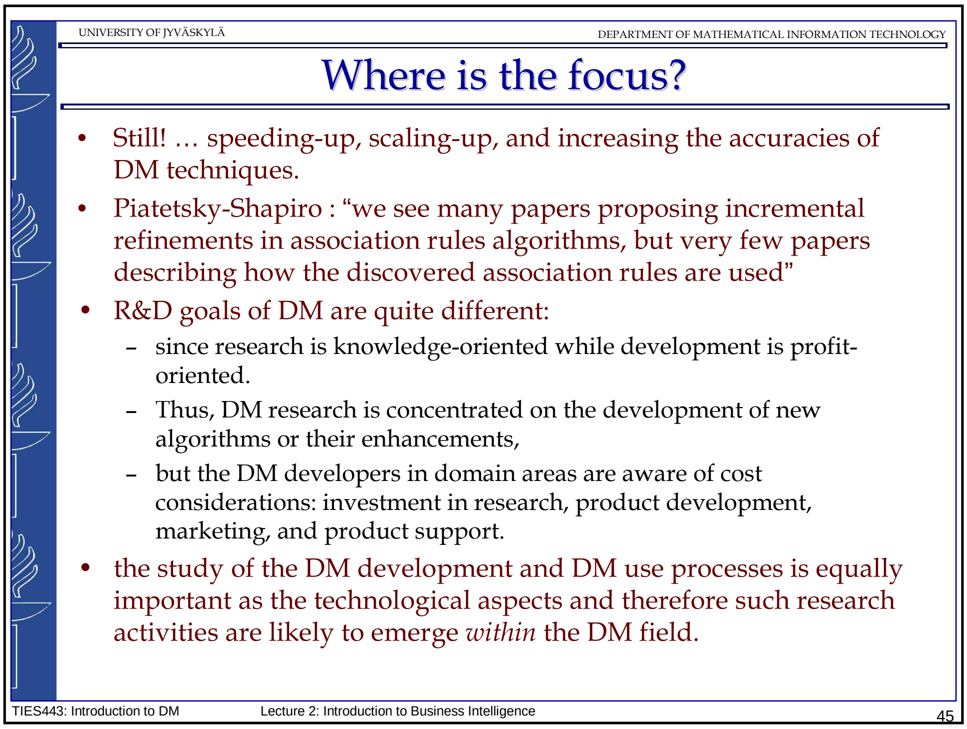
Justify/Evaluate

Applicable Knowledge

Foundations
Design knowledge

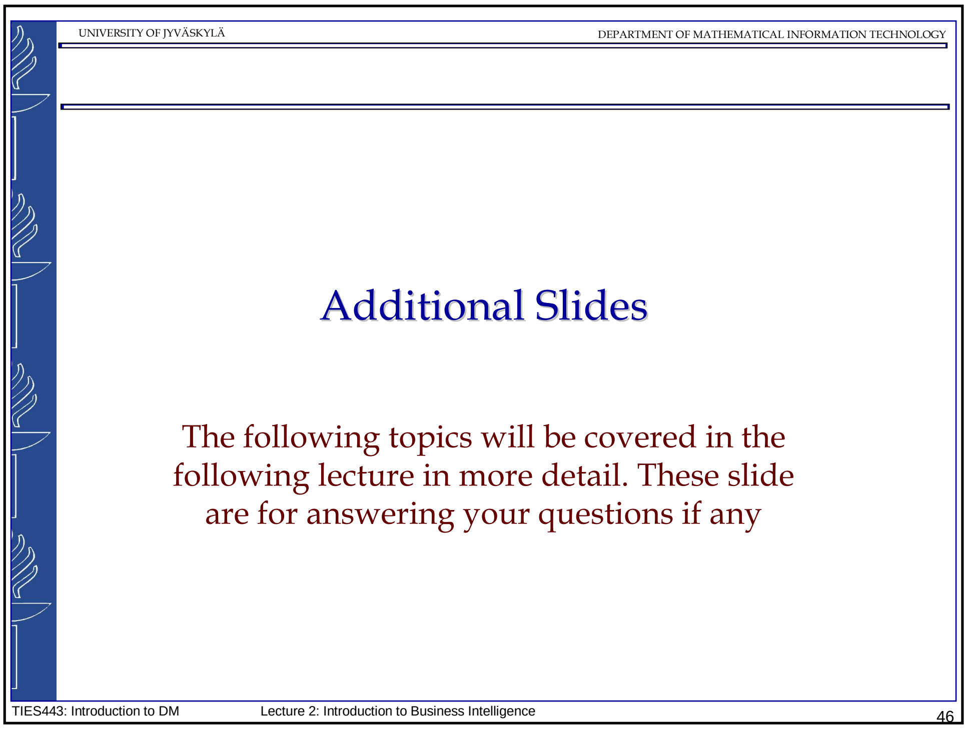
(Un-)Successful Applications in the appropriate environment

Contribution to Knowledge Base



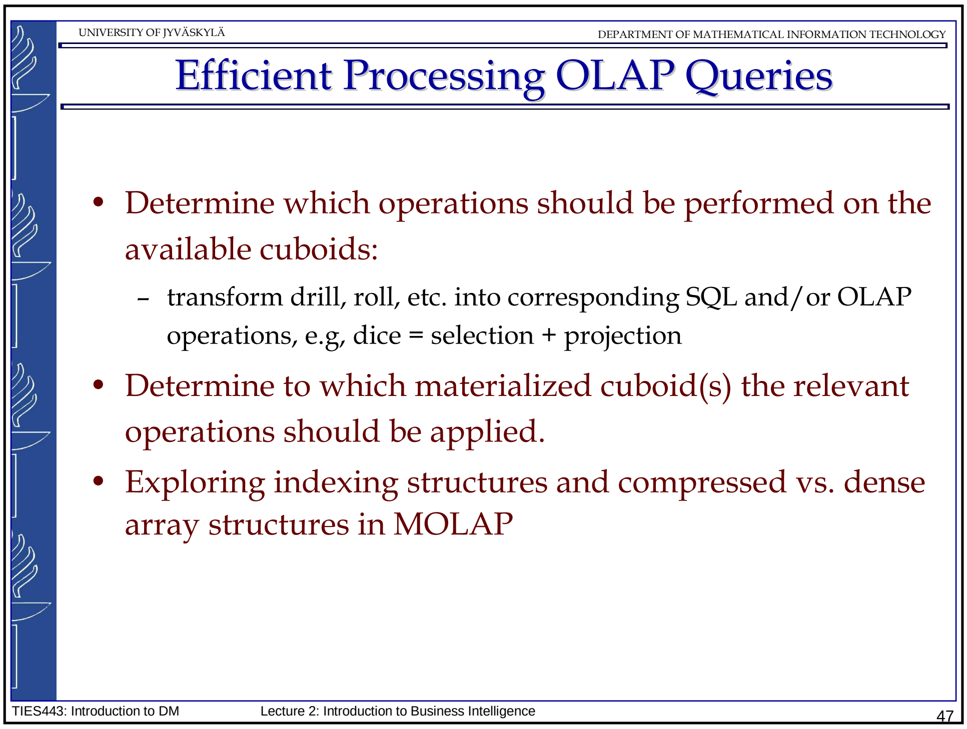
Where is the focus?

- Still! … speeding-up, scaling-up, and increasing the accuracies of DM techniques.
- Piatetsky-Shapiro : "we see many papers proposing incremental refinements in association rules algorithms, but very few papers describing how the discovered association rules are used"
- R&D goals of DM are quite different:
  - since research is knowledge-oriented while development is profit-oriented.
  - Thus, DM research is concentrated on the development of new algorithms or their enhancements,
  - but the DM developers in domain areas are aware of cost considerations: investment in research, product development, marketing, and product support.
- the study of the DM development and DM use processes is equally important as the technological aspects and therefore such research activities are likely to emerge *within* the DM field.



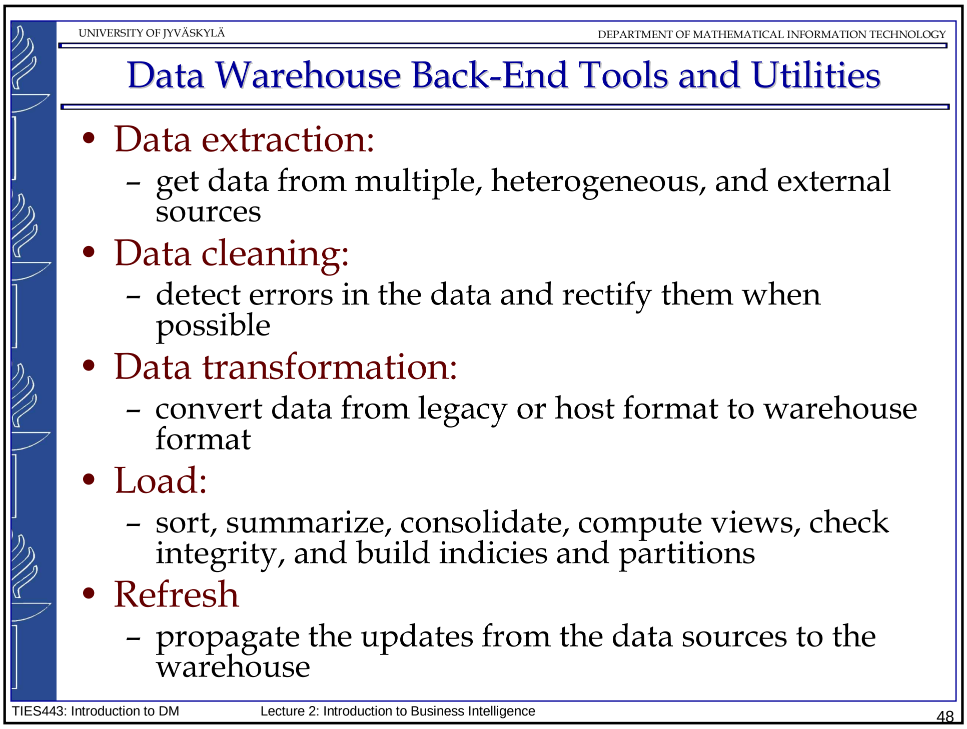
Additional Slides

The following topics will be covered in the following lecture in more detail. These slide are for answering your questions if any



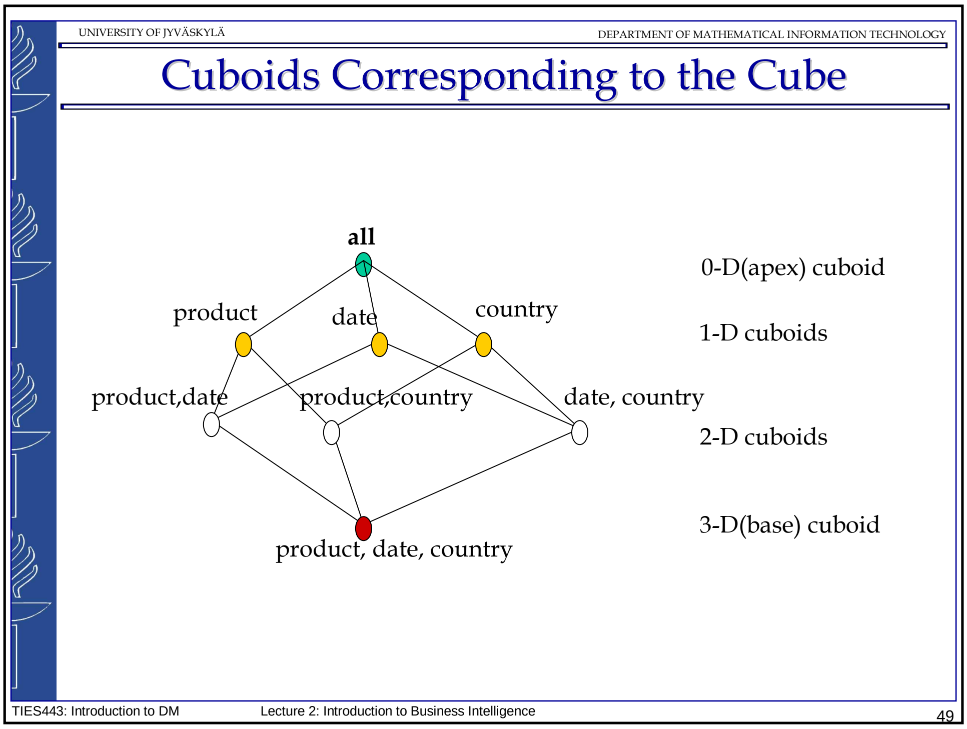
Efficient Processing OLAP Queries

- Determine which operations should be performed on the available cuboids:
  - transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g, dice = selection + projection
- Determine to which materialized cuboid(s) the relevant operations should be applied.
- Exploring indexing structures and compressed vs. dense array structures in MOLAP

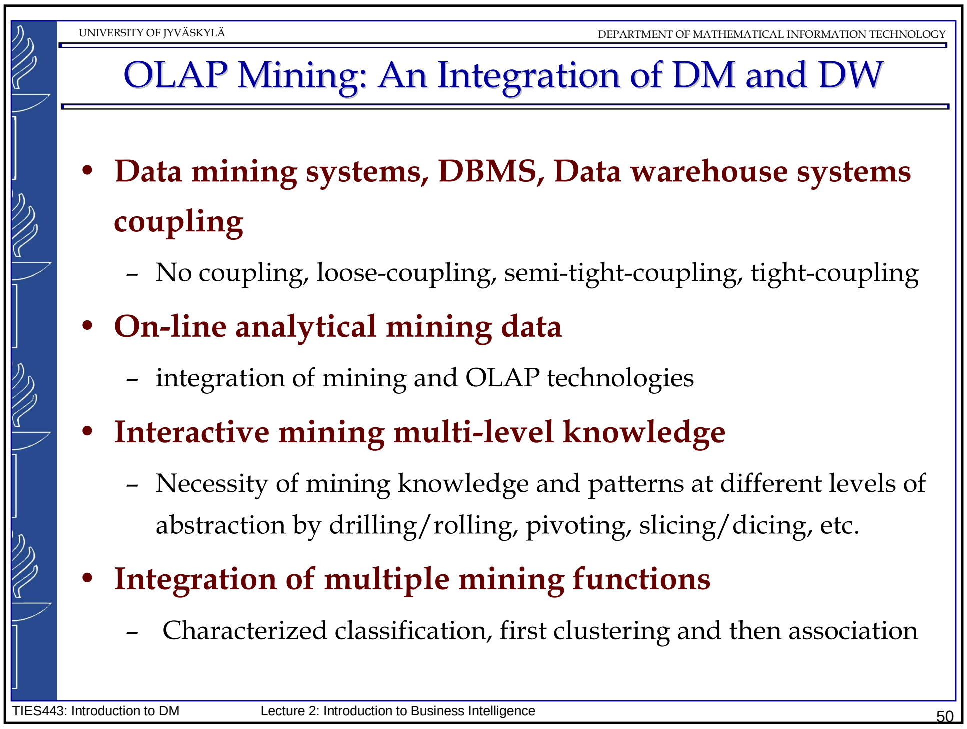


Data Warehouse Back-End Tools and Utilities

- Data extraction:
  - get data from multiple, heterogeneous, and external sources
- Data cleaning:
  - detect errors in the data and rectify them when possible
- Data transformation:
  - convert data from legacy or host format to warehouse format
- Load:
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse



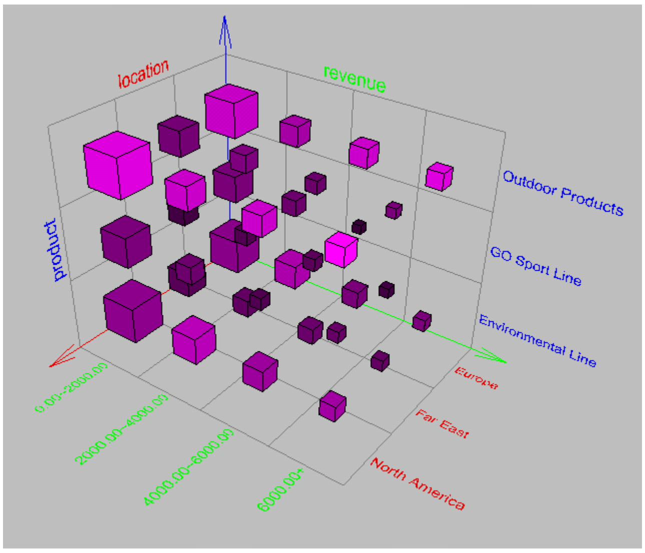
Cuboids Corresponding to the Cube



OLAP Mining: An Integration of DM and DW

- **Data mining systems, DBMS, Data warehouse systems coupling**
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- **On-line analytical mining data**
  - integration of mining and OLAP technologies
- **Interactive mining multi-level knowledge**
  - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- **Integration of multiple mining functions**
  - Characterized classification, first clustering and then association

Browsing a Data Cube



- Visualization
- OLAP capabilities
- Interactive manipulation

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice
  - *project and select*
- Pivot (rotate)
  - *reorient the cube, visualization, 3D to series of 2D planes.*
- Other operations
  - *drill across: involving (across) more than one fact table*
  - *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*