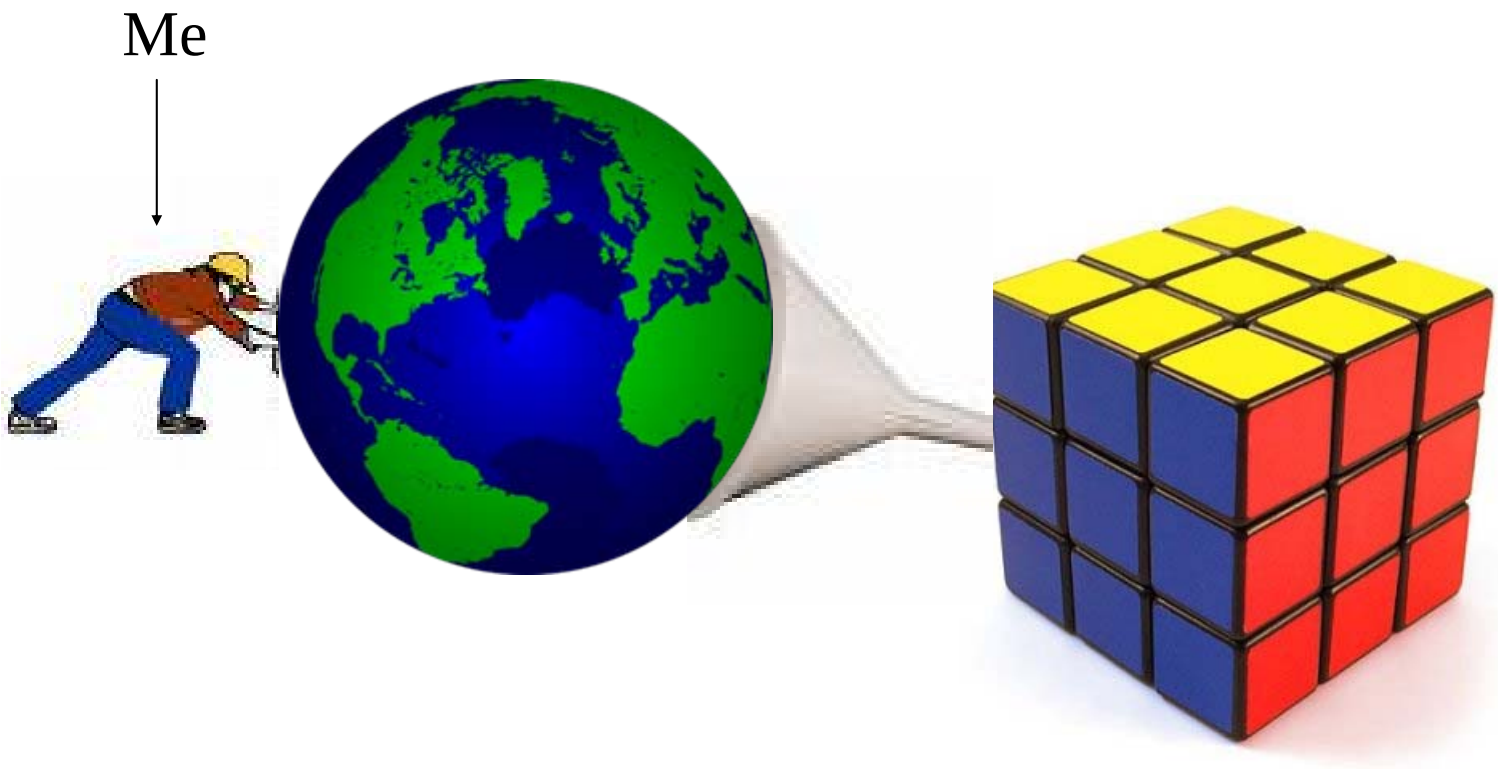
AALBORG UNIVERSITY

Warehousing The World: Challenges From New Types of Data

Professor Torben Bach Pedersen
Department of Computer Science
Aalborg University, Denmark

Speaker Intro

- Well, I try to squeeze the world into cubes…

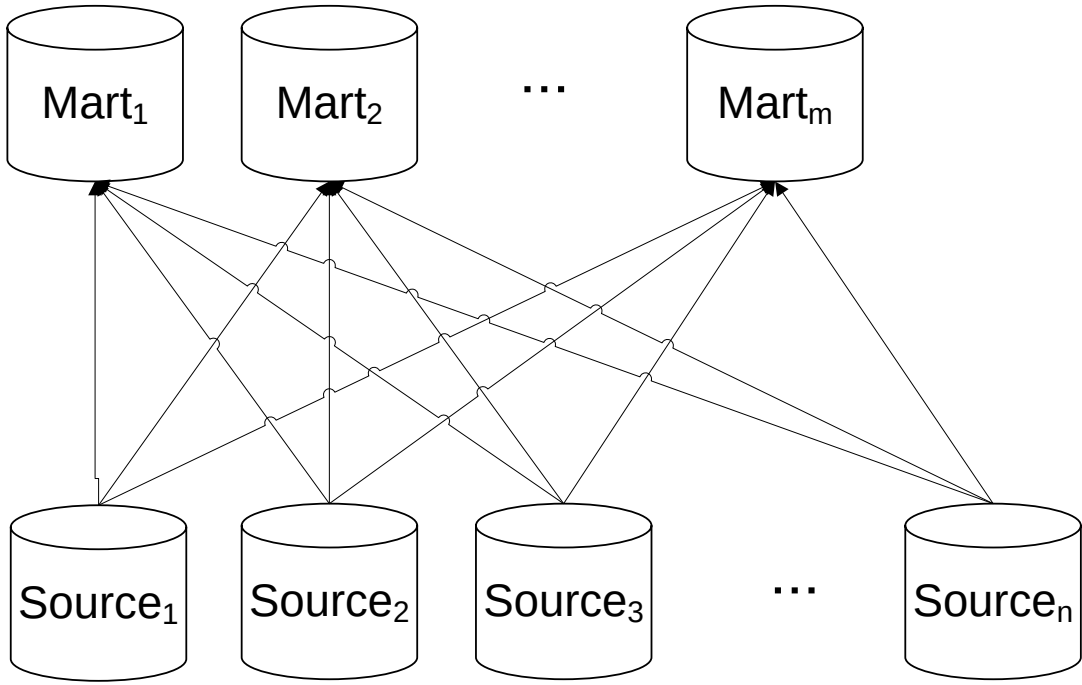


Talk Overview

- Data warehousing and business intelligence
  - Current status
- What is missing?
  - Support for new types of data
  - Associated challenges
  - Partial solutions
- The World Warehouse
  - An integrated solution
  - Challenges for the World Warehouse
- Conclusion and future work

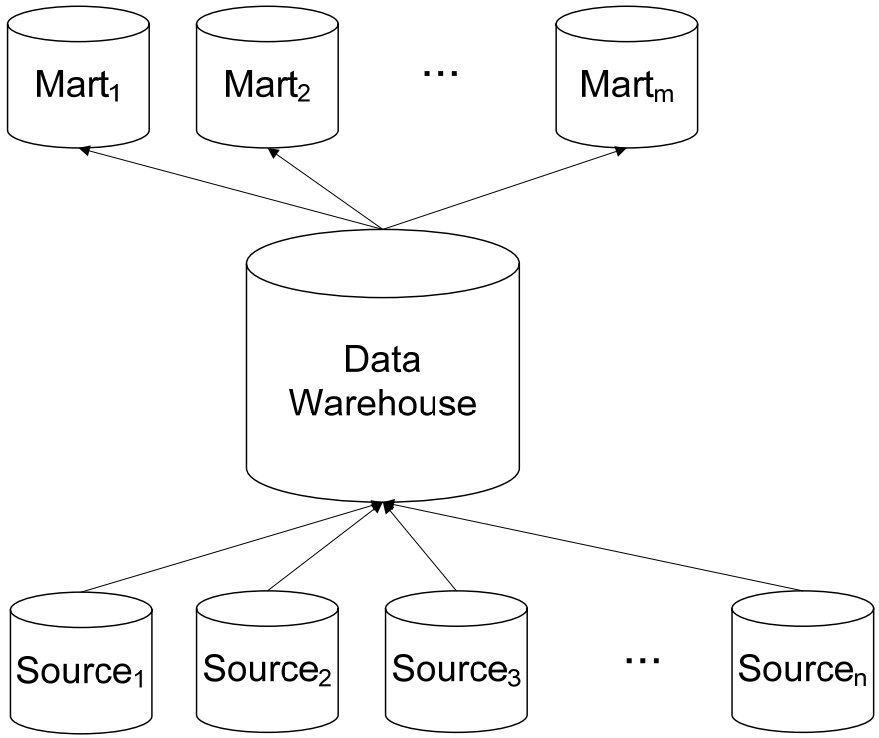
Data Warehouse Refresher

- Why was it that data warehouses were smart?
- In the old days, systems looked like this:
- And that was not so smart…
- n*m connections must be coded/maintained
- Even worse, **different** views on the world



Data Warehouse Refresher 2

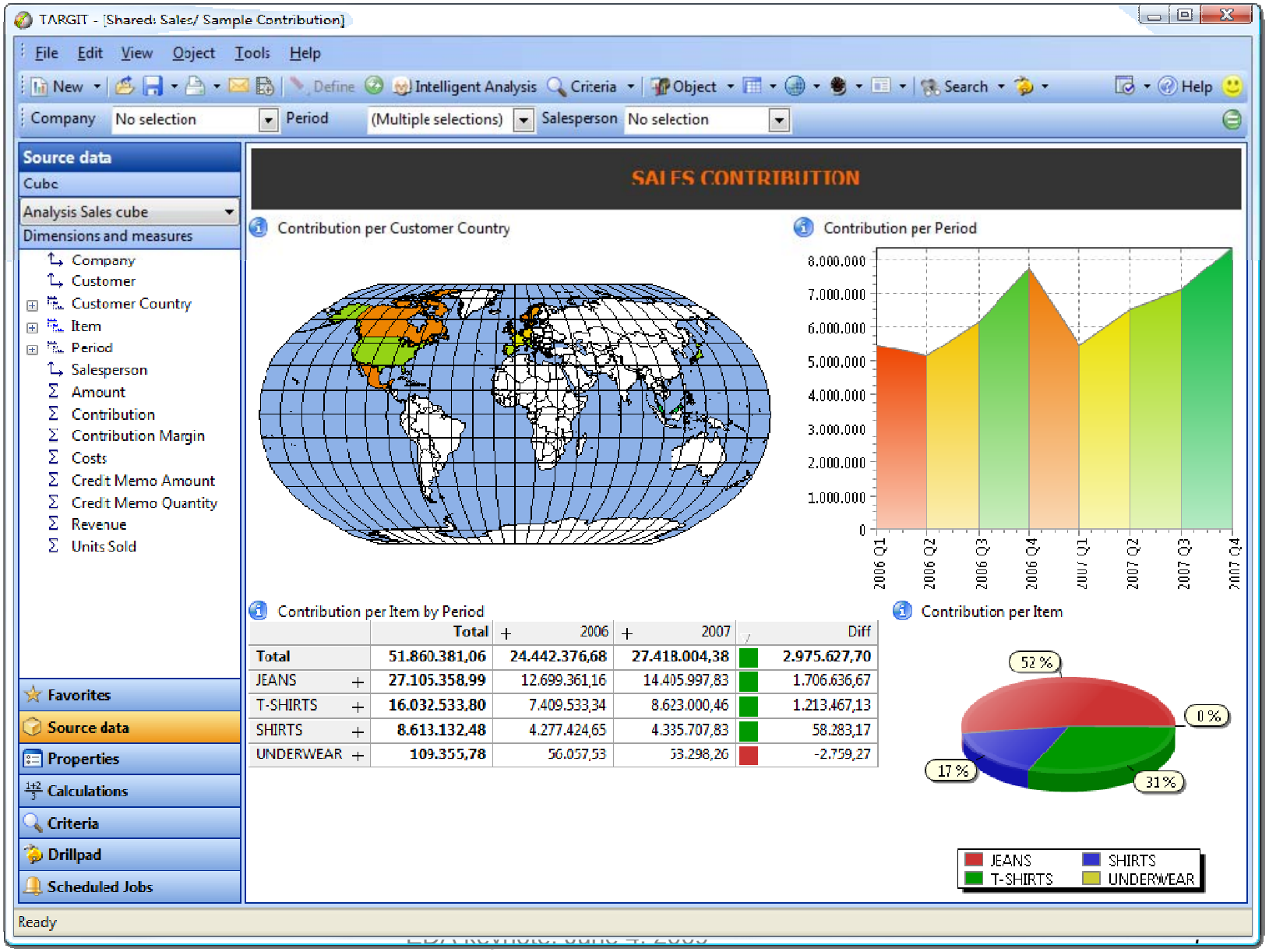
- A data warehouse looks like this:
- And that is much smarter:
- Only n+m connections must be coded/maintained
- Even better, **common** view of the world



Multidimensional Data

- DWs are based on a multidimensional data model
- Important business events, e.g., sales, modeled as *facts*
- Facts characterized by hierarchical *dimensions*, e.g., time and product
- Associated numerical *measures*, e.g., sales price
- The multidimensional model is unique in providing a framework that is
  - Intuitive
  - Effective
  - Allowing data to be viewed/analyzed at desired level of detail
  - Supports excellent performance
- Note: MD data is about “as structured as you can get!”

Example BI tool: TARGIT BI Suite



Status 2009

- Almost all (large) organizations have some kind of data warehouse
- With a business intelligence (BI) solution on top
- (Pretty good) control of finance data, sales data, etc.

- Are we then ”done” with DW+BI?

- Absolutely not! ☺

What is missing?

- Traditional DWs work well for traditional, *structured* data
- But DW data only cover *very little* of an organization's data
- So, DWs only solve *a small part* of the real integration and analysis needs of most organization

- So, what is missing is:
  - …the rest of the world!

New Types of Data

- Structured data is quite well supported
  - Relational data + multidimensional data in DWs
  - But other types of data are not:
- Text data is found everywhere
  - Documents, emails, web pages
- Semi-structured/XML data
  - Electronic catalogs, semantic web data
- Mobile, pervasive and ubiquitous computing:
  - Large quantities of geo-related data
  - Data from a large amount of sensors
- Analytical models of data developed through data mining
  - Used, e.g., to predict the future
- All this must be integrated and used for BI

Problems with New Types of Data

- Problem with current technologies:
  - All these different types of data/models cannot be integrated and analyzed in a *coherent* fashion
- Instead, applications must develop *separate* ad-hoc solutions for integration and analysis
  - Typically for each pair of data types
  - For example, combining relational and text data

Integration of New Types of Data

- Trend 1: Integration of semi-structured/XML data
- Trend 2: Integration of text data
- Trend 3: Integration of geo-data
- Trend 4: Integration of sensor data
- Trend 5: Data streams
- Trend 6: Integration of analytical models
- Trend 7: Privacy

- Some solutions/systems offer partial support, but there is still a long way to go…
- I will discuss the issues, and show a few partial solutions

Integration of XML/Semi-structured

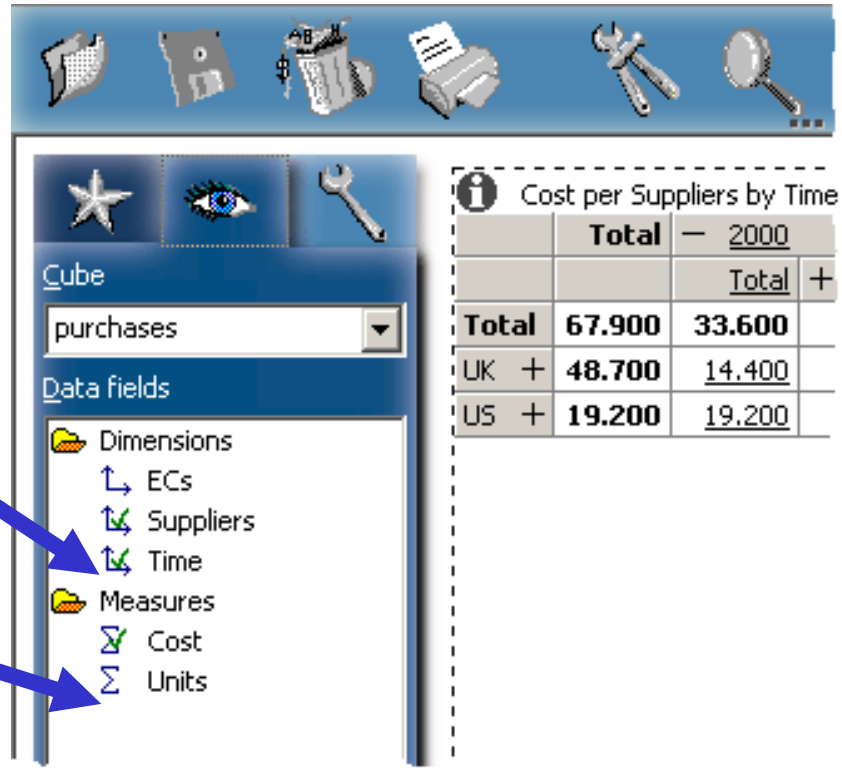
- XML data is everywhere
- XML data is ”semi-structured”
  - Simple example: emails in XML format
  - Some data is quite structured, e.g., email To/From/CC fields
  - Some data is un-structured, e.g., the email text
- A lot of valuable information only found in XML data
- Problems related to BI
  - BI systems (generally) only handle nice, structured data ☹
- Benefits of XML integration
  - New types of analyses
  - ”compare number of emails to/from our customers to their share of sales, are we using to much time on some of them?”
  - ”who is (not) communicating with who in our company?”

Example: Integration of XML Data

- XML data as **logical** dimensions/measures
- Prototype with TARGIT

```
<?xml version="1.0"?>
 <SupplierCities>
  <Supplier>
   <SupplierName>A.A. Corp.</SupplierName>
   <City>Aalborg</City>
  </Supplier>
       ⋮
</SupplierCities>
```

```
<?xml version="1.0"?>
<OverheadExpenses>
 <Supplier Name="A.A.">
  <EC Name="EC1">200</EC>
  <EC Name="EC4">300</EC>
 </Supplier>
       ⋮
</OverheadExpenses>
```



Semantic Web Data

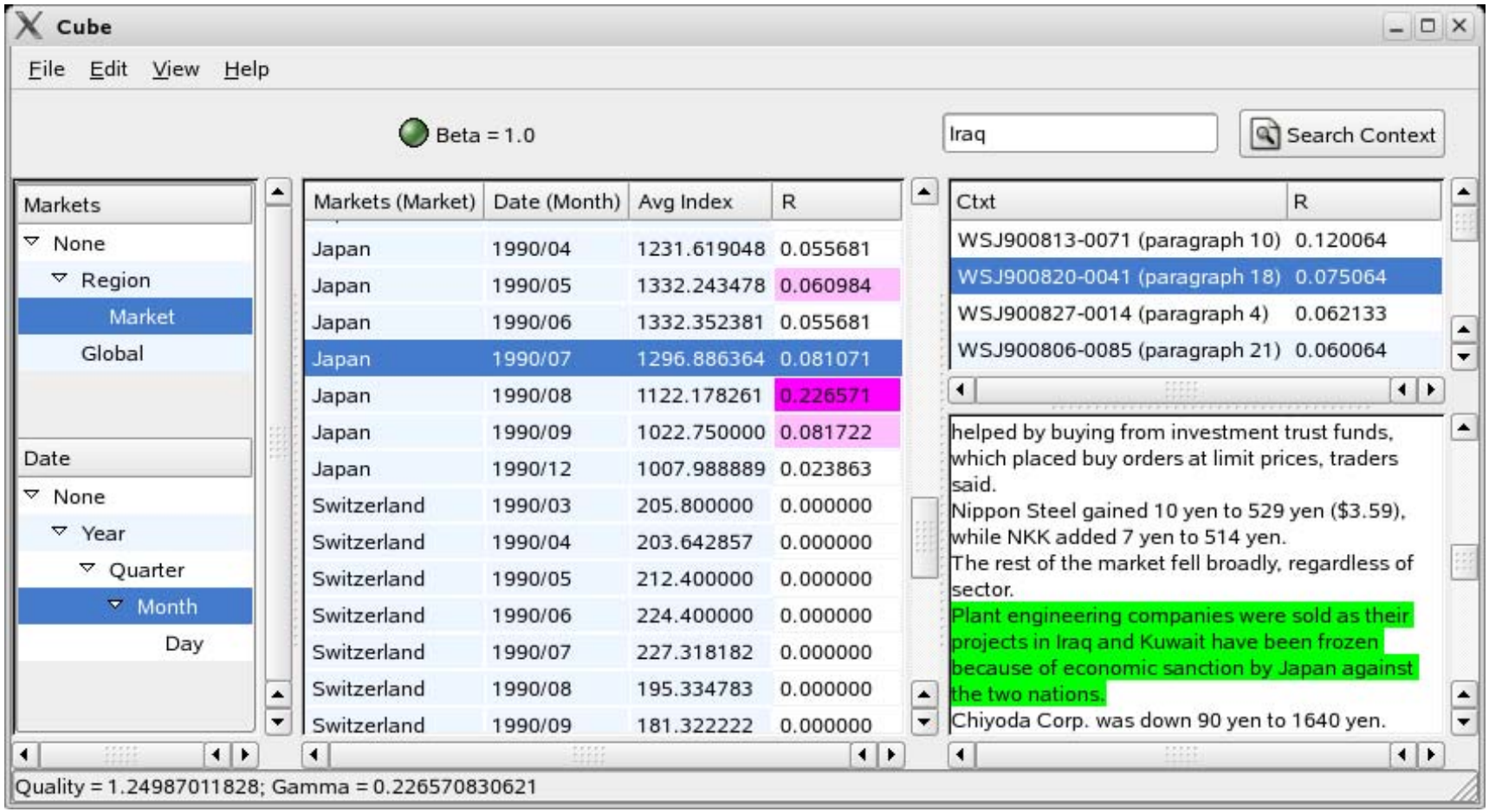
- A very interesting new development
- Semantic web data
  - RDF
  - OWL
  - Used to specify ontologies
- Often used for capturing semantics of existing (web) data
  - Open world assumption, new knowledge can be added later
- Wide range of “structuredness”
  - From very structured data (ontologies)
  - To quite unstructured data (“scattered” (s,p,o) triples)
- Reasoning capabilities: a new thing for most data models

Integration of Text Data

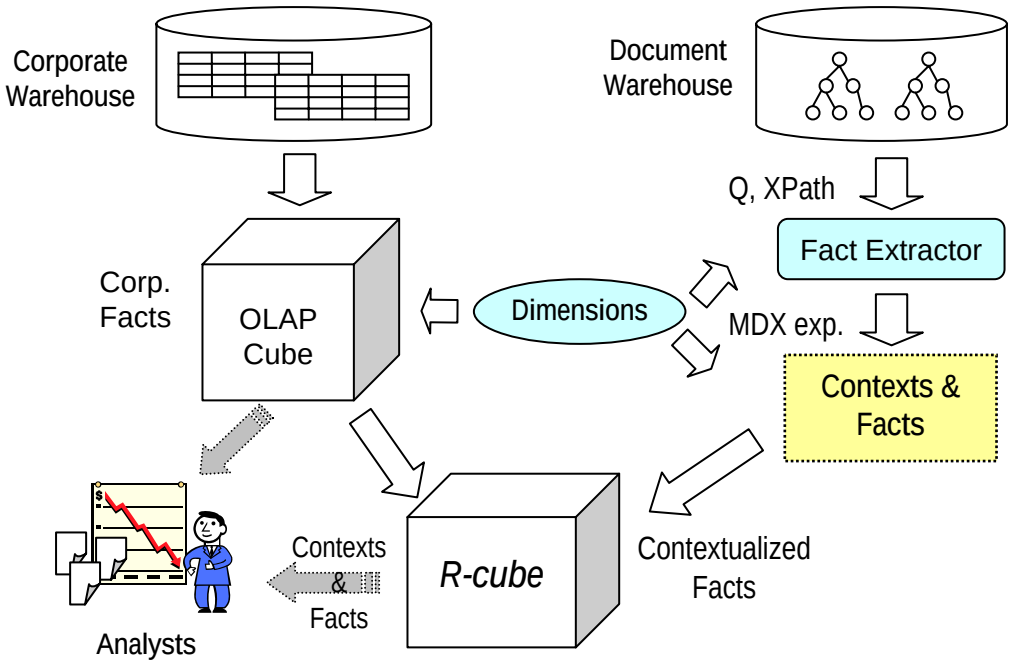
- Text data is everywhere
  - Web, news, market analyses..
  - A lot of valuable information **only** found in text data
- Problem: BI systems cannot handle text ”in a smart way”
  - Cannot ”link text and numbers”
- Benefit: new types of analyses
- Early 2003: stock analyst thinks ”the US will soon invade Iraq, how does that affect my portfolio?”
- ”Hmm, what happened during the Gulf War? Search on ‘Iraq’ ”

Example: Relevance Cubes

- Linking to cube shows that *Japanese* stocks were hit particularly hard during the Gulf War

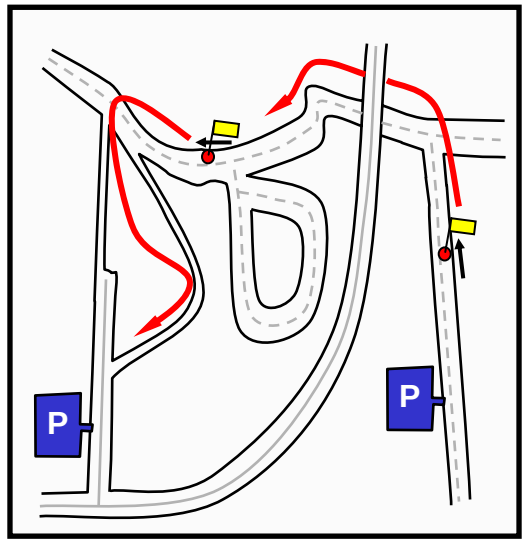


Example: Relevance Cubes



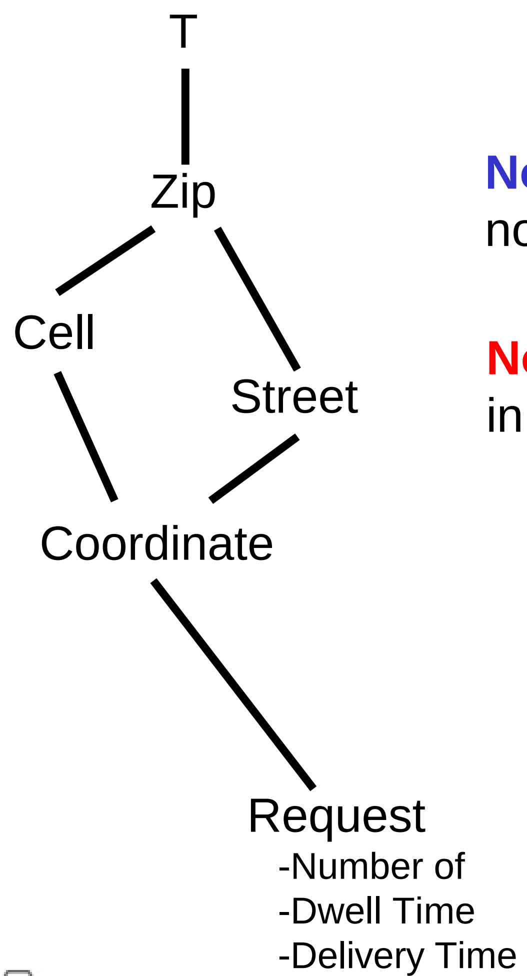
Integration of Geo Data

- Geo data everywhere
  - GPS, Google Maps, buildings, roads, infrastructure….
- Problem: BI systems cannot handle geo data
  - How to turn a road network into a dimension???
  - How to analyze driver routes in MS Analysis???
- Benefit: new types of analyses
  - Ride-share: which drivers often drive on (part of) the same route at the same time?
  - When are the most potential customers coming by my shop?



Non-Standard Dimensions

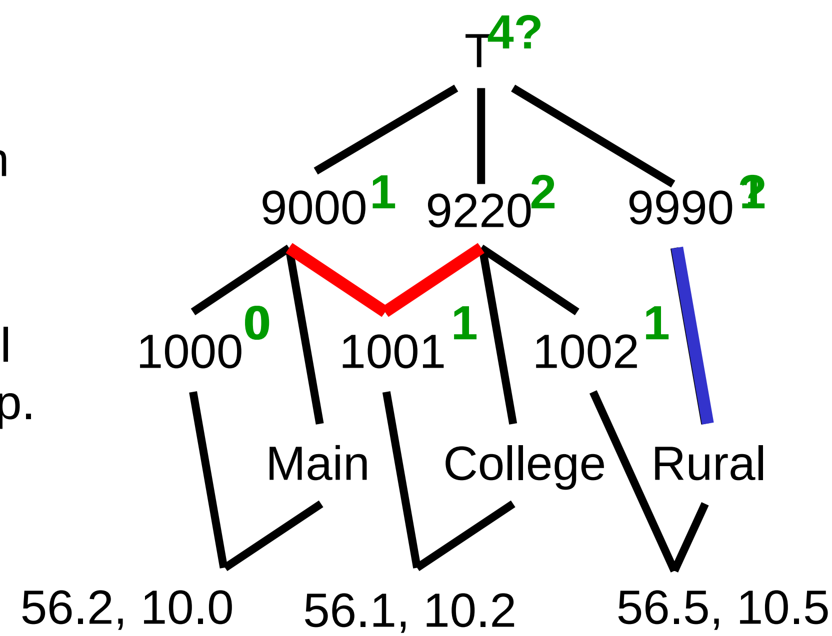
Location dimension



Non-onto: Zips with no cell.

Non-strict: One cell in more than one Zip.

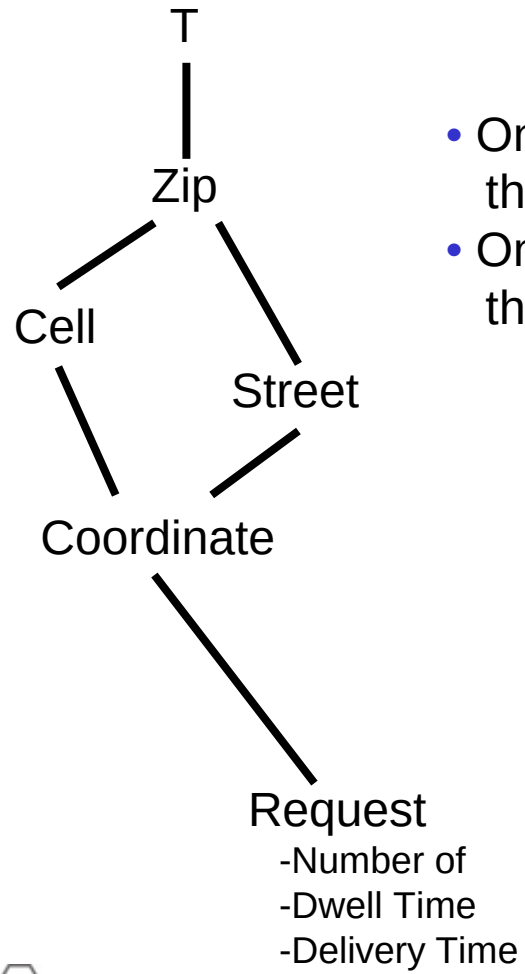
Location dimension instance



- Cannot be captured using standard multi-dimensional data models.
- Double counting possible.
- One solution is **hierarchy normalization**

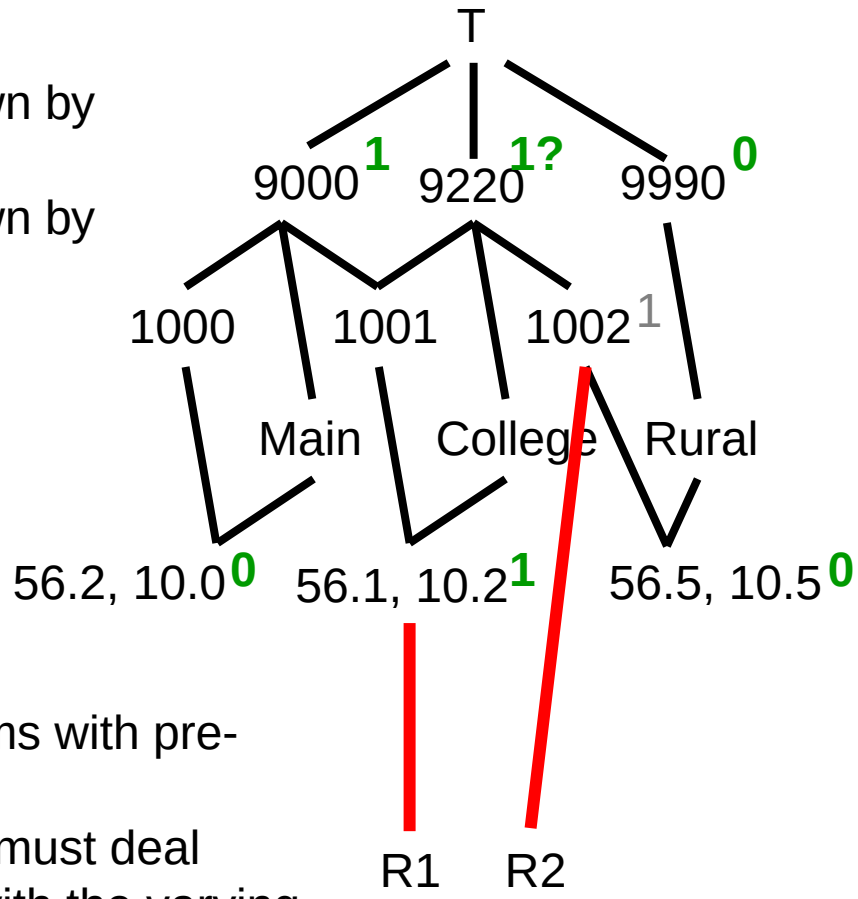
Imprecision and Varying Precision

Location dimension



- One location is known by the coordinate.
- One location is known by the cell.

Location dimension instance

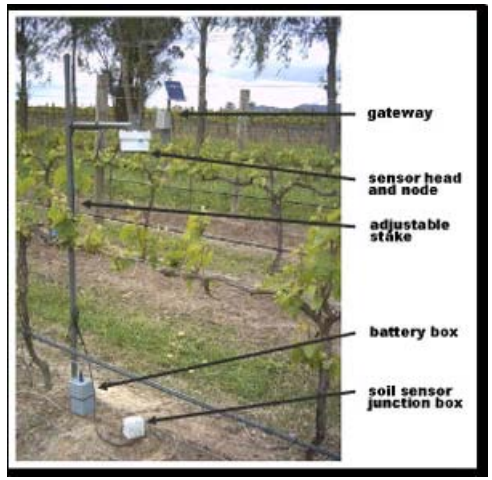


- Again problems with pre-aggregration.
- The systems must deal intelligently with the varying precisions.

Integration of Sensor Data

- Sensors appearing everywhere in our surroundings
  - Temperature, moist, soil, RFID, GPS….
  - Passive or active
  - Mote: sensor/CPU/RAM/transm
- Organized in wireless sensor networks, see right
- Problem: BI systems do not handle sensor data well
  - Data streaming in every second, no connection, sensors don't work, imprecise data, wrong values, central computation not possible….
- Benefit: new types of analyses
  - Connection between temp., soil and yield?

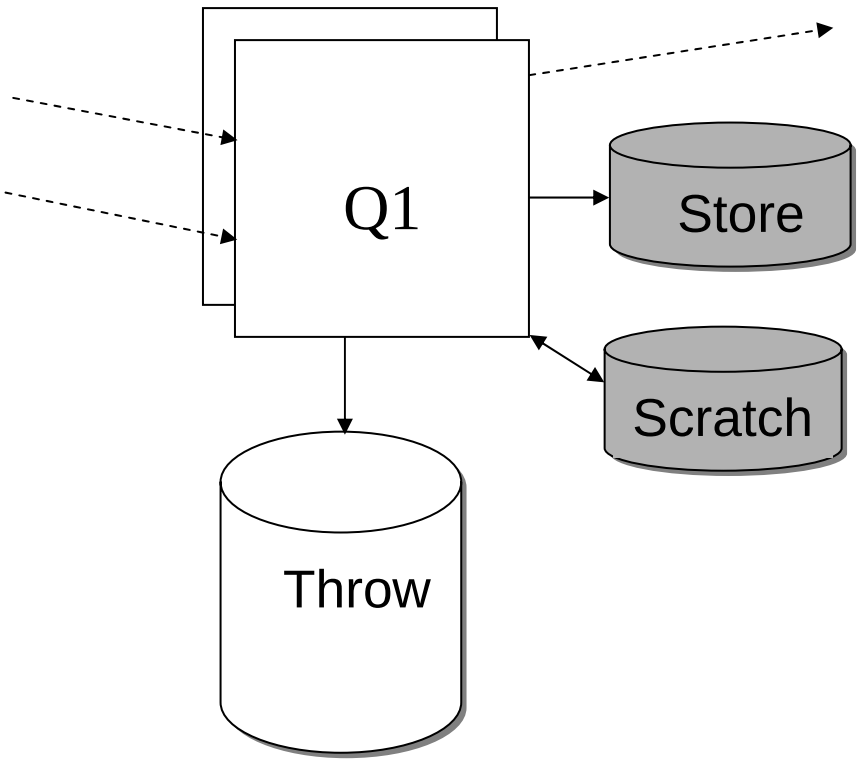
[Levis et al. CACM 51(7)]



[Crossbow, 2007]

Data Streams

- Often **too much data** for traditional ”save in DW and analyse”
  - AT&T: Internet backbone
  - Sensor network data
  - Detail data ”not interesting”
- New ”paradigm”: data streams
- ”Analyse and throw away”
  - **Continuous** queries
  - Data in/out in streams
  - **Some** data put in **Store**
  - Temp data put in **Scratch** (RAM)
  - Unneccasry data discarded
- New type of sw: DSMS
- Not handled by current BI



Integration of Analytical Models

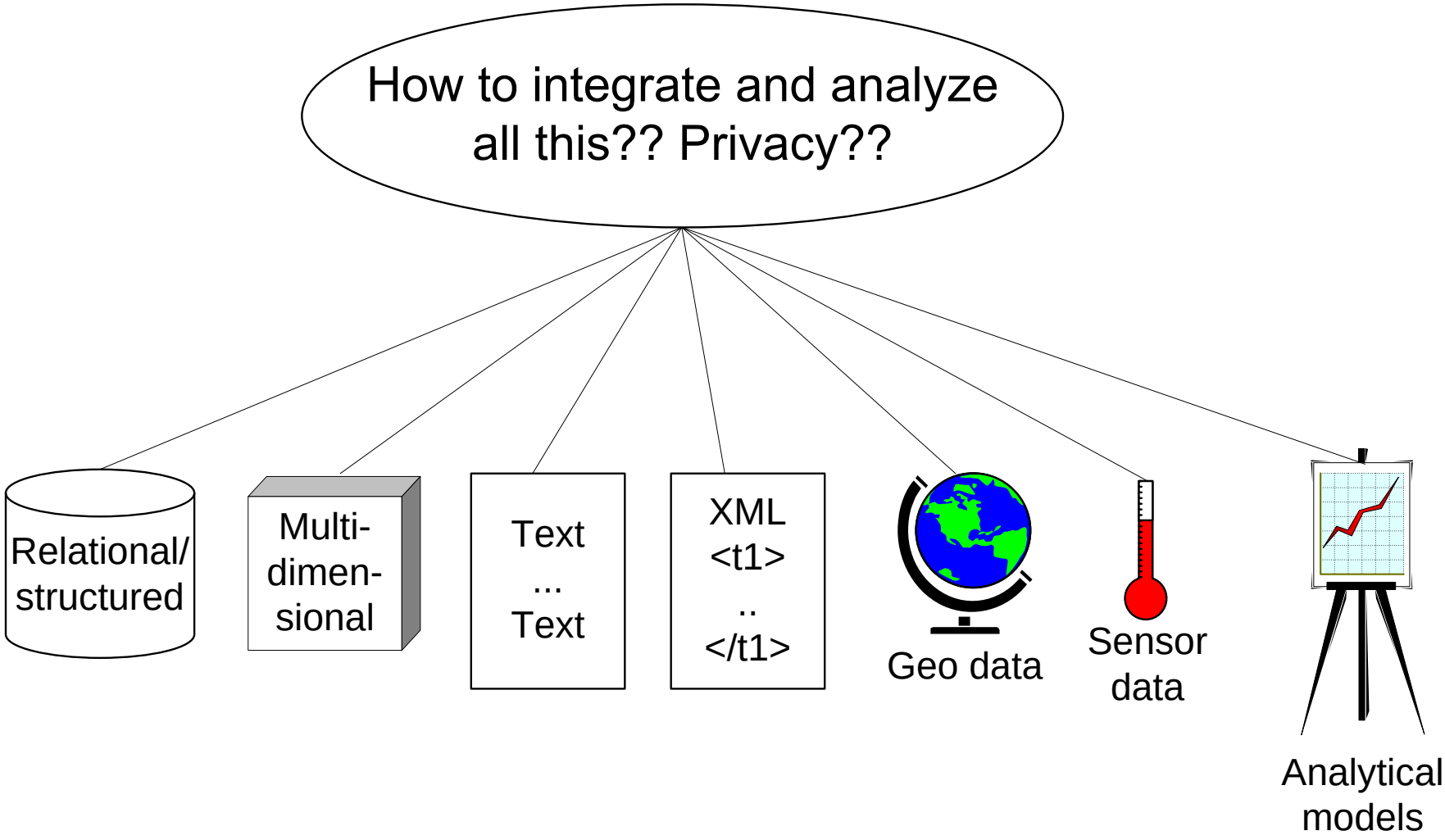
- We have data about past, present, and future
- Past: databases
- Present: data streams
- Future: forecasting/prediction models
- But this is not integrated!
- Problem: 3 different systems for handling this data
  - Not integrated
- Data should be managed in the same system
  - Only difference: ”future” data is more ”imprecise”
- Benefit: new types of analyses
  - ”Where were/are the traffic jams yesterday, right now, and in 20 minutes?”
  - ”Show traffic on our web site for 2007-10”

daisy

Privacy

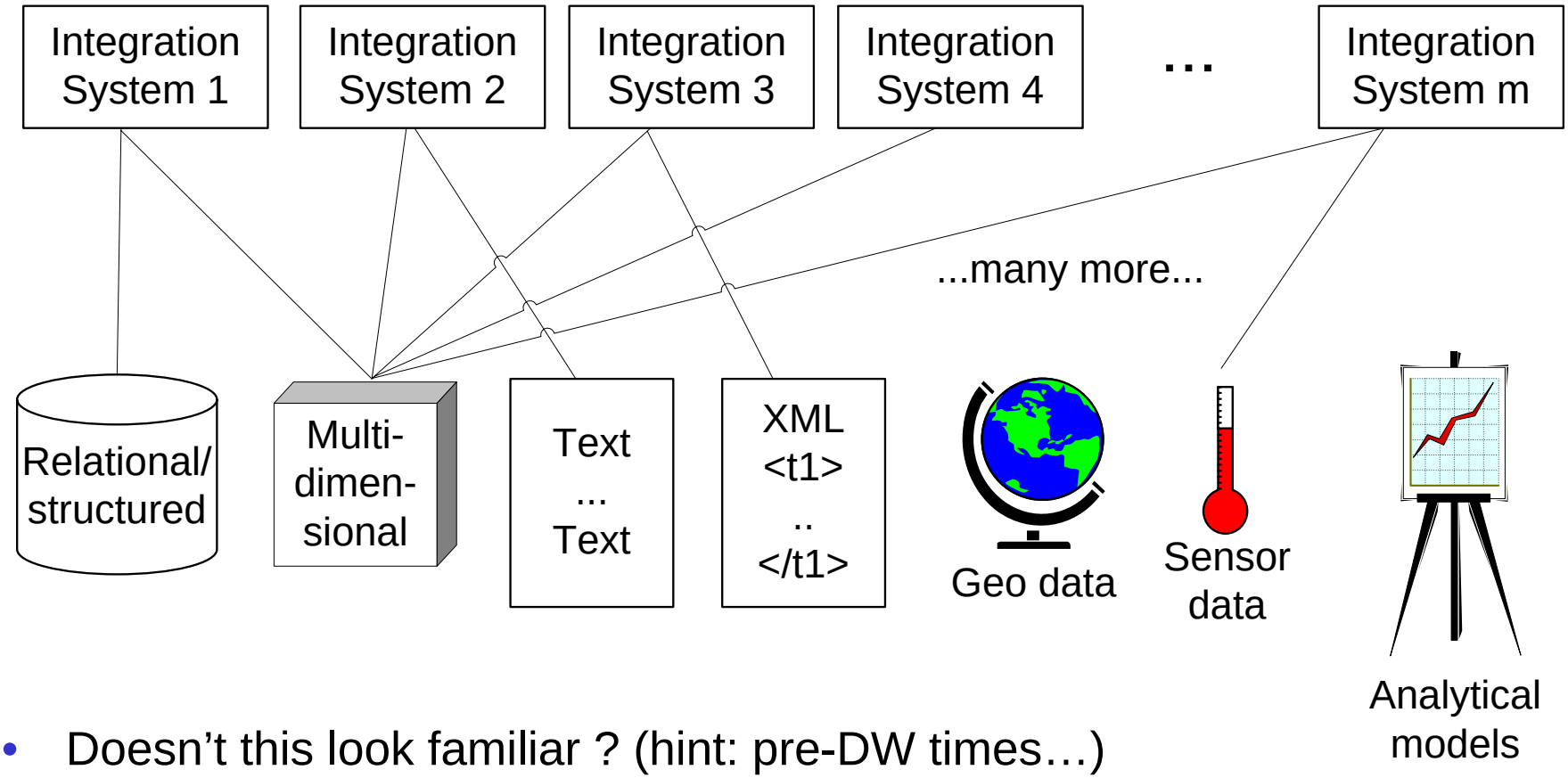
- Privacy becoming an increasing concern
- Data about individuals accumulated like never before
  - Web surfing, web sites like Facebook, GPS…
  - Plus data in ERP, public systems, etc.
- ”Joining” theses sources expose ”sensitive” knowledge
  - Holiday pic on Facebook->White Pages->burglary
  - The driver is only with customers in half of the working hours
- Often, detailed data is not revealing
  - A single GPS position doesn’t say so much
- But trends can be revealing
  - ”Every Thursday he is at a certain hotel from 13-14”
- Problem: BI systems don’t know what is ”sensitive”
  - And if they do, they only know at the detailed data level
- Benefit: find valuable trends, **without** upsetting people

Taking a Step Back…



Existing Solutions

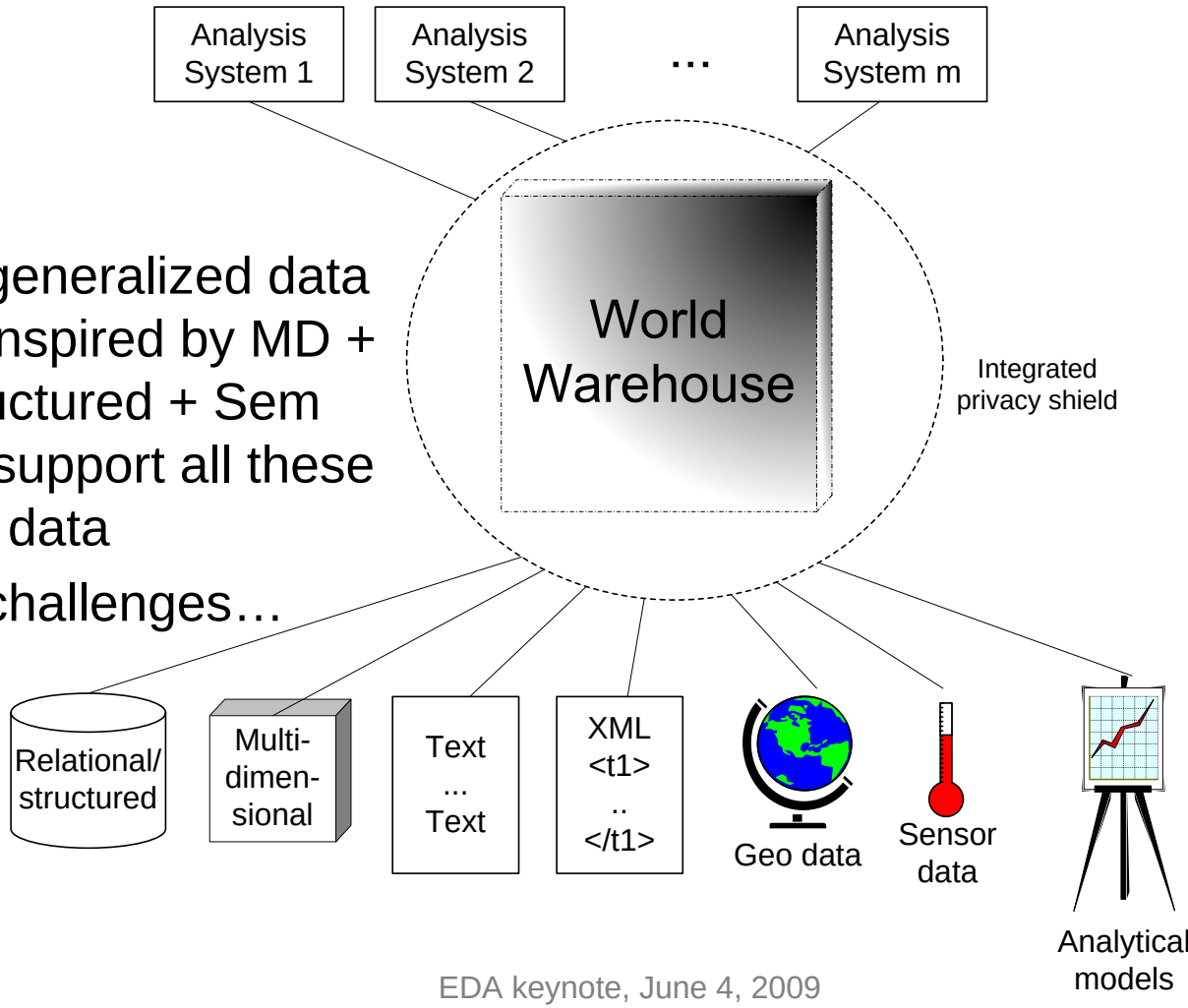
- ”Pair-wise” integration
- Many! different systems…my own work included…



- Doesn’t this look familiar ? (hint: pre-DW times…)

The World Warehouse

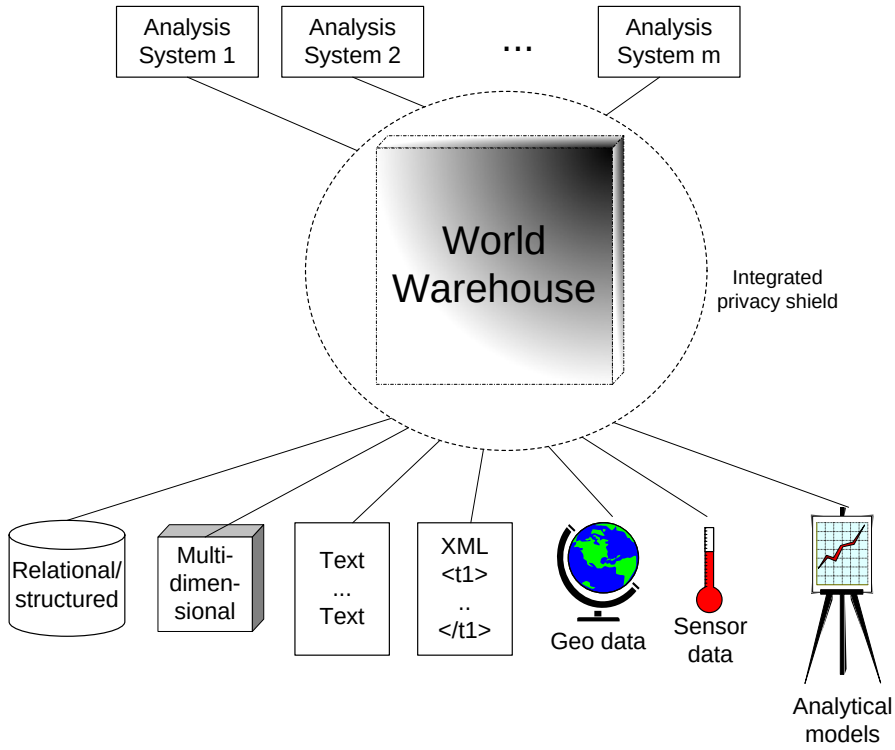
- Generalize the DW success!



- Create generalized data model, inspired by MD + semistructured + Sem Web to support all these types of data
- Lots of challenges…

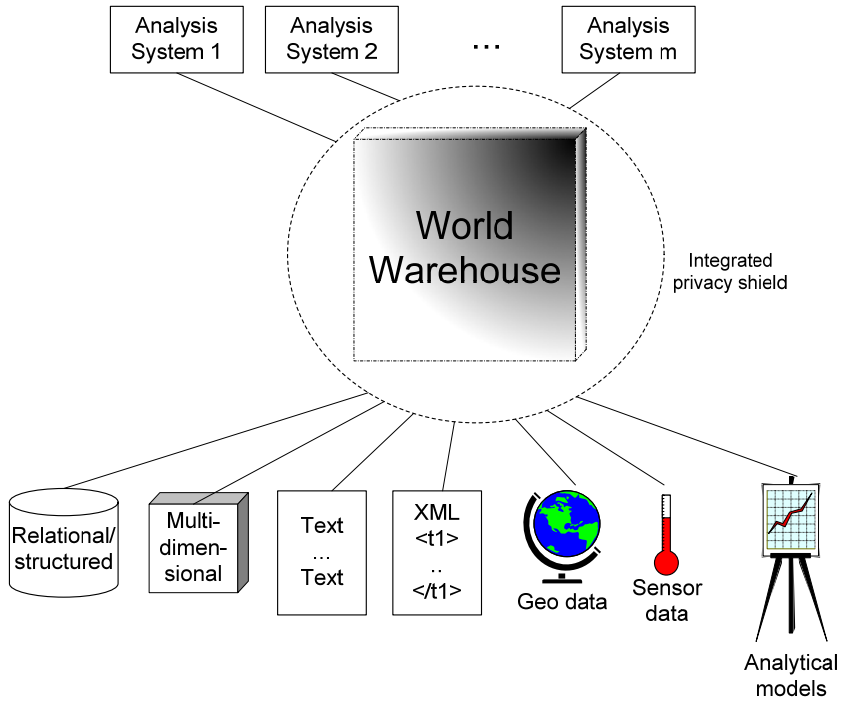
The World Warehouse

- Overall idea: repeat the “data warehouse success” for integrating different types of data
- Data of a particular type should only need to be “integrated” once
- Integrated results put into common, “harmonized” data store (WW)
- WW handles all these types of data (or *derivations*) for data analysis
- The WW is a cube, meaning, i.e., based on MD principles
- WW content has different “shades,” data is “not just black and white.”
- All WW data has a built-in notion of “perfection” (precision/certainty)
- Data may be very precise and totally certain (like ord. DW data)
- Or imprecise and uncertain (sampling errors, data from analytical models)



The World Warehouse

- Sources (different types) connected through only ***one*** "connection"
- Difficult task of integrating particular type of data handled once-and-for-all, by mapping into WW data model (+algs/tools)
- Analysis systems have only one "connection" each to the WW
- Take advantage of all functionality and data available in WW
- No need to perform integration themselves (as the systems mentioned earlier)
- WW has "integrated privacy shield."
- When data comes from sources, shield analyzes data + performs modifications (aggregation, swapping,...) before storing
- When data is requested from an analysis system, WW may perform further modifications of results



The World Warehouse

- The WW approach means that the “complexity” of the integration of all the different types of data for:
  - *n* types of data
  - *m* analysis systems
- Drops to *n*+*m* (from *n***m*)!
- The “hard” tasks
  - Integrating a new type of data
  - Protecting privacy
- Are generally handled only once
  - By the WW rather than in the analysis systems
  - Great relief for the development of the analysis systems.

A New Data Model

- Basis for the WW will be a novel kind of data model
  - Should encompass the best of several worlds
  - Multidimensional modeling concepts (superior for analysis)
  - Flexibility and generality from semi-structured data models
  - Borrow useful Semantic Web concepts
- Support a much wider range of data
  - Geo-related data (geo models, etc)
  - Sensor data
    - High speed data streams, missing or incorrect values, etc.
  - Semi-structured and unstructured data
    - Enabling analysis across structured, semi-structured, and unstructured data
  - Imperfect (imprecise, uncertain, etc.) data
- Support for privacy management

daisy

Research Topics

- Develop **complete** “infrastructure”
- Query languages
- Query processing/optimization techniques
- Data integration techniques
- Techniques for integrating databases, sensors, and analytical/predictive models of data
- Integrate contributions into a common prototype system
  - Open source project
- Integrated system enables solutions to be evaluated experimentally using large volumes of real-world data

daisy

Benefits and Challenges

- The same benefits to all the described data types as is currently available in traditional DWs for structured data
- WW enables the integration and analysis of all types of data using the developed data model and query language
- Distinguishing feature: all-encompassing ”privacy shield”
  - All queries to the DW pass through/approved by shield
- Five challenges
  - Warehousing data about the physical world
  - Integrating structured, semi-structured, and unstructured data in DWs
  - Integrating the past, the present, and the future
  - Warehousing imperfect data
  - Ensuring privacy in DWs
- Novel to consider the challenges in combination

Data About The Physical World

- Data stemming from the physical world have unique characteristics
- Geo-related data
  - GPS readings, maps, transportation networks
- Data from sensors in the environment
  - Temperature, humidity,…
- Issues include
  - Handling various geo models
  - Managing high speed data streams
  - Missing or incorrect values, etc.

Structured, semi-struct., un-struct.

- The WW needs to be able to effectively integrate semi-structured and unstructured data for analysis purposes
- For enabling analysis across structured, semi-structured, and unstructured data.
- How to overcome the issue that:
  - Multidimensional data are usually very homogeneous and structured
  - While semi-structured and unstructured data is, by nature, very heterogeneous (and obviously not very structured)?
- Idea: store *derivations* of data, rather than data itself
  - Store the fact that a particular sentence in a particular document is related to the sale of vegetable oil in the Japanese market
  - Rather than storing the sentence itself

Past, Present, and Future

- The WW has to support the seamless and integrated querying of
  - Past data (as current DW data)
  - Current data (continuously streaming in from sensors),
  - Future data (predicted using analytical models).
- It should be possible to say:
  - “SELECT sales FROM cube WHERE month=<next month>”
  - just as easily as selecting data from the last month.
- Idea: break down traditional distinction between
  - “**real**” data values and functions/models that **describe** data
- These two aspects should be seen as a **duality** of the same thing
  - Like the duality of particles and waves in nuclear physics.
- The conversion between the two aspects can be achieved by
  - “folding” data into models
  - “unfolding” models into data
- Unfolding mechanism means that models/functions can be used in queries just as ”real” data values.
- This unified view enables easy integration of past data (DWs), present data (sensors), and predicted future data (models)

Warehousing Imperfect Data

- In the WW all data values have an attached
  - Uncertainty
  - Imprecision
- Both ”**real**” (historical) and ”**fake**” (future, predicted) data
- Always having notion of ”imperfection” makes it natural to compress/aggregate data into patterns/models
  - Wavelets, probabilistic models, …
- Models can be “unfolded” to (re-)provide original data.
- One particular challenge is how to balance:
  - Complexity of managing data imperfection
  - Requirements for high performance analysis

daisy

Ensuring Privacy in DWs

- Privacy is hard to realize effectively…
- The idea of folding/unfolding can actually aid in privacy protection
  - Privacy can be protected by folding (aggregating/ compressing/...) actual data values into patterns (which is just one kind of function/model describing the data)
  - This creates some imprecision, but this is also captured natively in the WW
  - Current privacy protection approaches (generalization, condensation, randomization, cloaking,…) are actually all special cases of this general mechanism, so the benefits of a more general approach can be significant
- Idea for integrated privacy management ”shield”:
  - Enforcement mechanism based on **certification**
  - Privacy requirements for a particular data item are built into the data item itself using a special “**privacy dimension**”.
  - Any query accessing data item (typically an aggregation function) must provide a **certificate** stating how the query preserves privacy
  - Certificate issued by a trusted external party
  - Certificate matched against the privacy requirements
  - If requirements are met, the data item releases the desired value, otherwise it will refuse to release the value or provide a properly anonymized value instead.

daisy

Conclusion

- DWs work very well for structured data
  - Multidimensional data model, …
- But fail to support many new types of data
  - Text, semistructured, geo data, sensor data, data streams, analytical models, …
  - Privacy an increasing issue
- Current solutions provide ”point-to-point” integration
  - Does not scale as new types of data arrive
- Solution: The World Warehouse
  - ”Repeat DW success”
  - Develop new, powerful data model and computing infrastructure
  - A number of challenges must be addressed

Future Work

- Well, most of it…☺
- Thanks a lot Maguelonne Teisseire for inviting me
  - And to the whole TATOO team for hosting me
- Entrepôts de Données est la future !
- Questions?