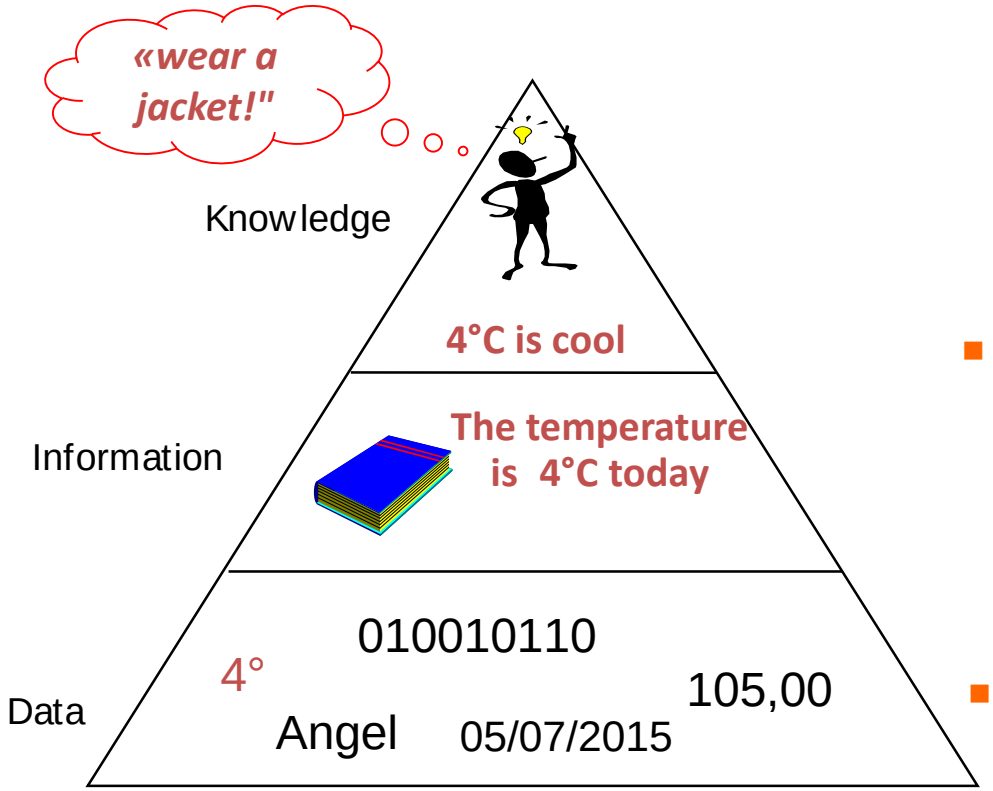
Business Intelligence and Data Warehouse

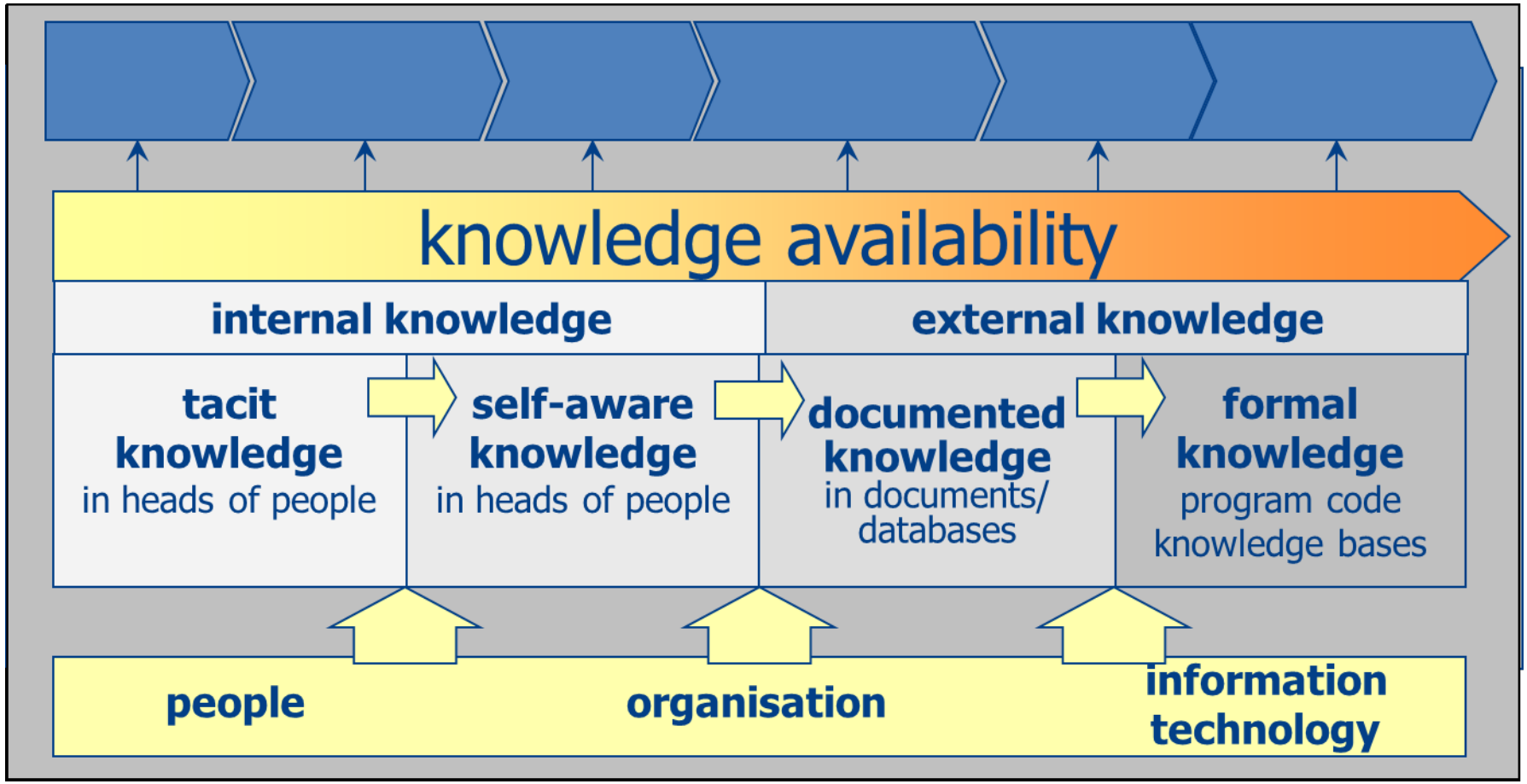
Knut Hinkelmann

Data, Information and Knowledge

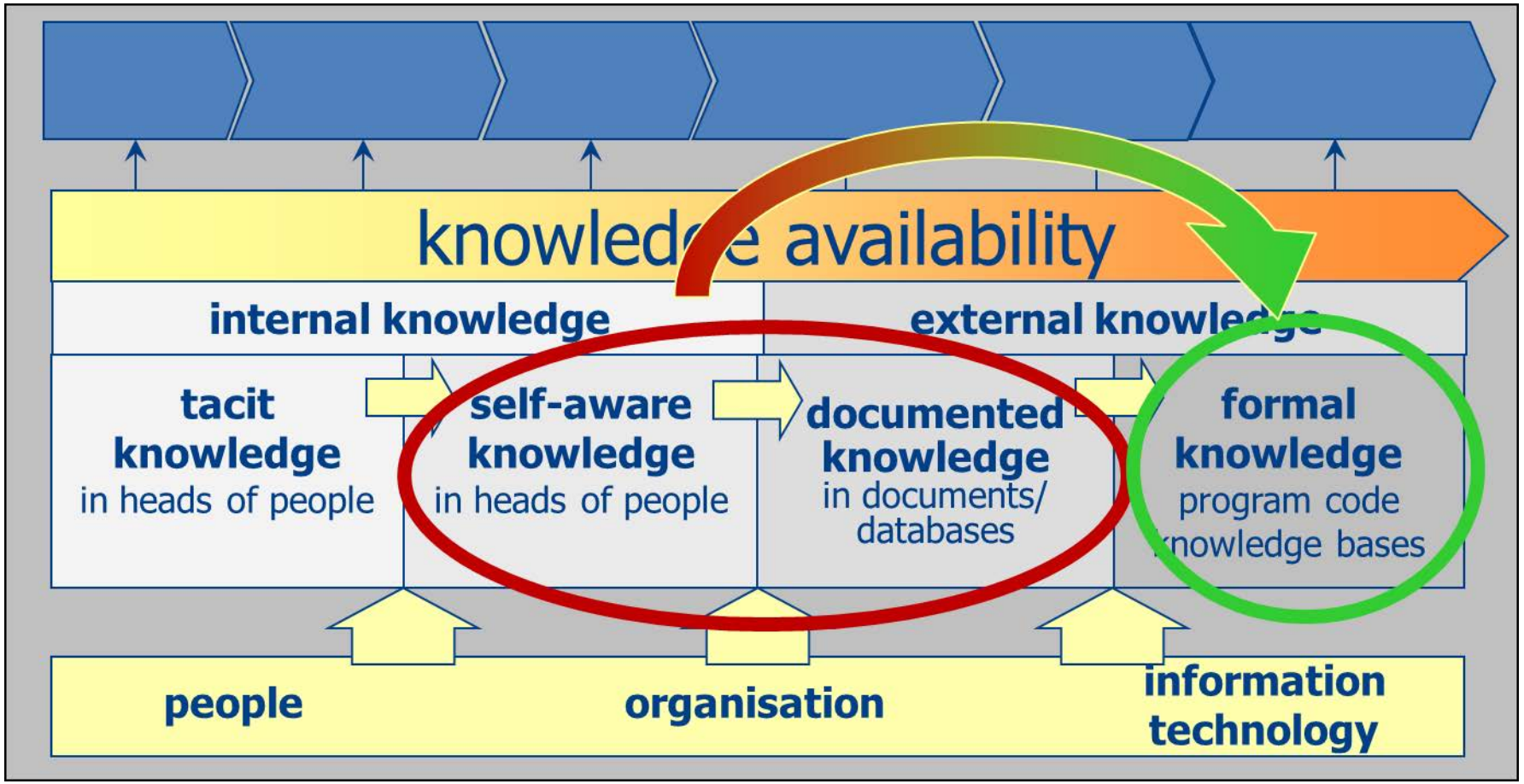


- **Knowledge** enables decisions and actions
  - originates from messages (information), experience, insight
  - is embedded into the beliefs and opinions of its owner
- **Information** is an interpretation of data, often assembled in messages
  - influences the judgment and behaviour of the recipient
  - has a significance (relevance, purpose)
- **Data** is a set of facts and/or signals
  - do not have meaning by itself
  - to understand data you need an interpretation

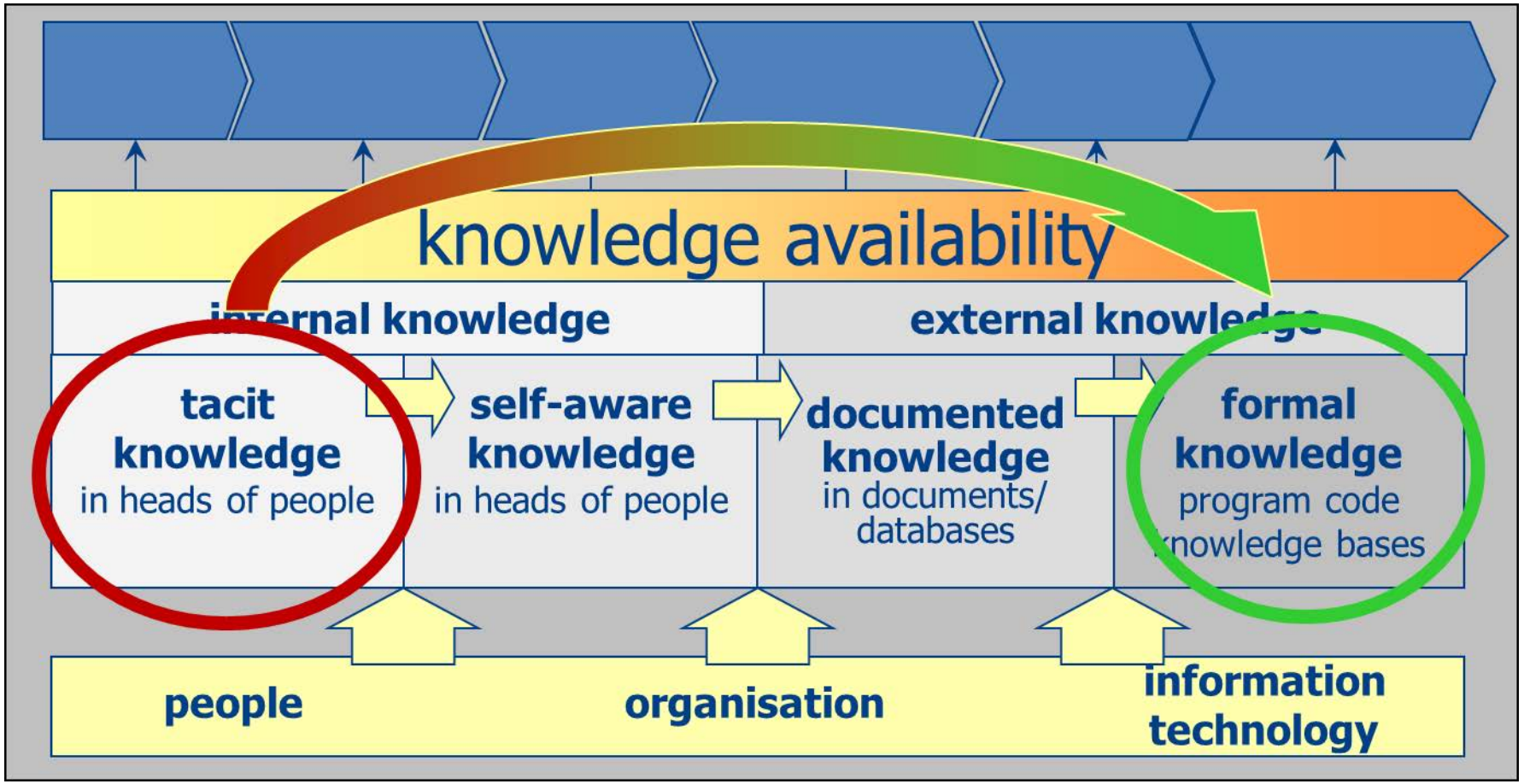
Knowledge


knowledge availability
internal knowledge
external knowledge
tacit knowledge
in heads of people
self-aware knowledge
in heads of people
documented knowledge
in documents/ databases
formal knowledge
program code
knowledge bases
people
organisation
information technology

Knowledge Engineering



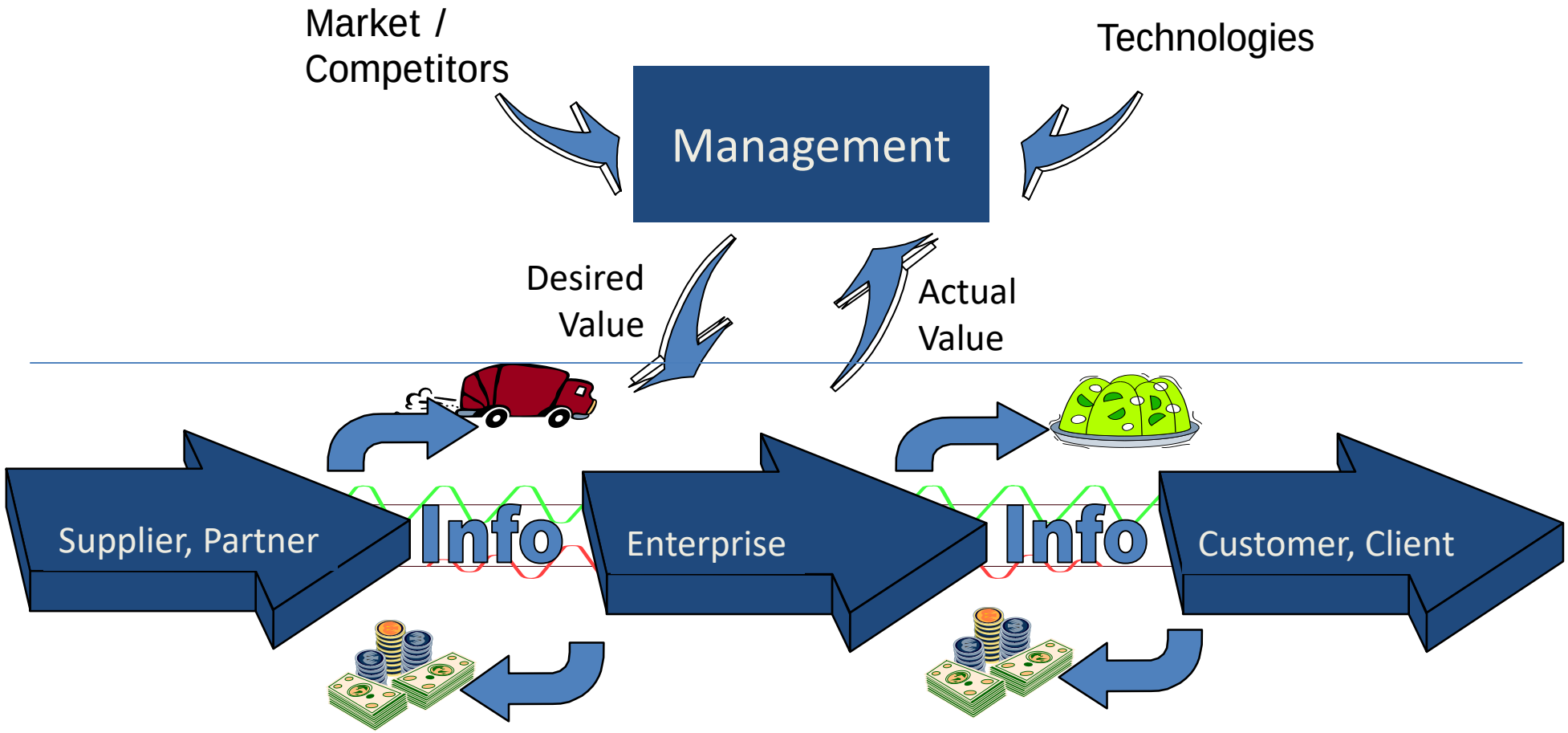
Machine Learning:
Make Knowledge explicit with the Use of Data



Business Intelligence – Definition(s)

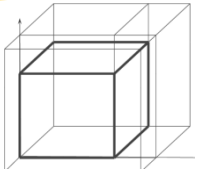
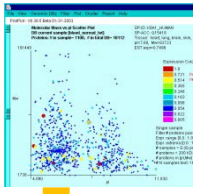
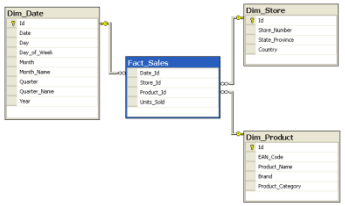
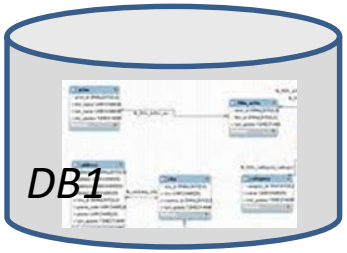
- *Sabherwal (2011):* «We define BI as providing decision makers with valuable information and knowledge by leveraging a variety of sources of data as well as structured and unstructured information. […] The key intellectual output of BI is **knowledge that enables decision making** with **information and data being the inputs**.»
- *Howson (2007):* Business Intelligence allows people at all levels of an organisation to **access, interact with and analyse data** to **manage the business, improve performance, discover opportunities, and operate efficiently.**

An Enterprise and is Context



Management = Information Processing = Decision Making

BI overview

	strategic	*operative*
Questions	• What are our goals? • Are we reaching our goals? • If not, where is the problem?	• Which credit applications should be accepted? • Who are potential csutomers for the new product?
Analyses	 *measure, aggregate, visualise*	 *Ad hoc queries, OLAP*
		 *find patterns (data mining)*
	dimensional modelling	
	 **Data Warehouse**	
	ETL ***ETL*** ***IE***	
raw data	 *DB1* *DB2*  	

Perspectives on BI – pain points

MARKETING

*For targeted campaigns, we would urgently need data about customers and their buying behavor […] ideally on an **integrated platform** where we can communicate with sales.*

MANAGEMENT

*I wanted to retrieve some numbers myself from my laptop. I then got **access to various (!) systems** […] I finally gave up*

SALES

*In most review meetings, we spend half the time discussing **which sales data are the right ones** because everyone brings their own reporting.*

ADVISORY BOARD

*Why weren't you able to **preview that trend**? All our competitors seem to have reacted long before we did!*

Why introduce BI? – primary motivations

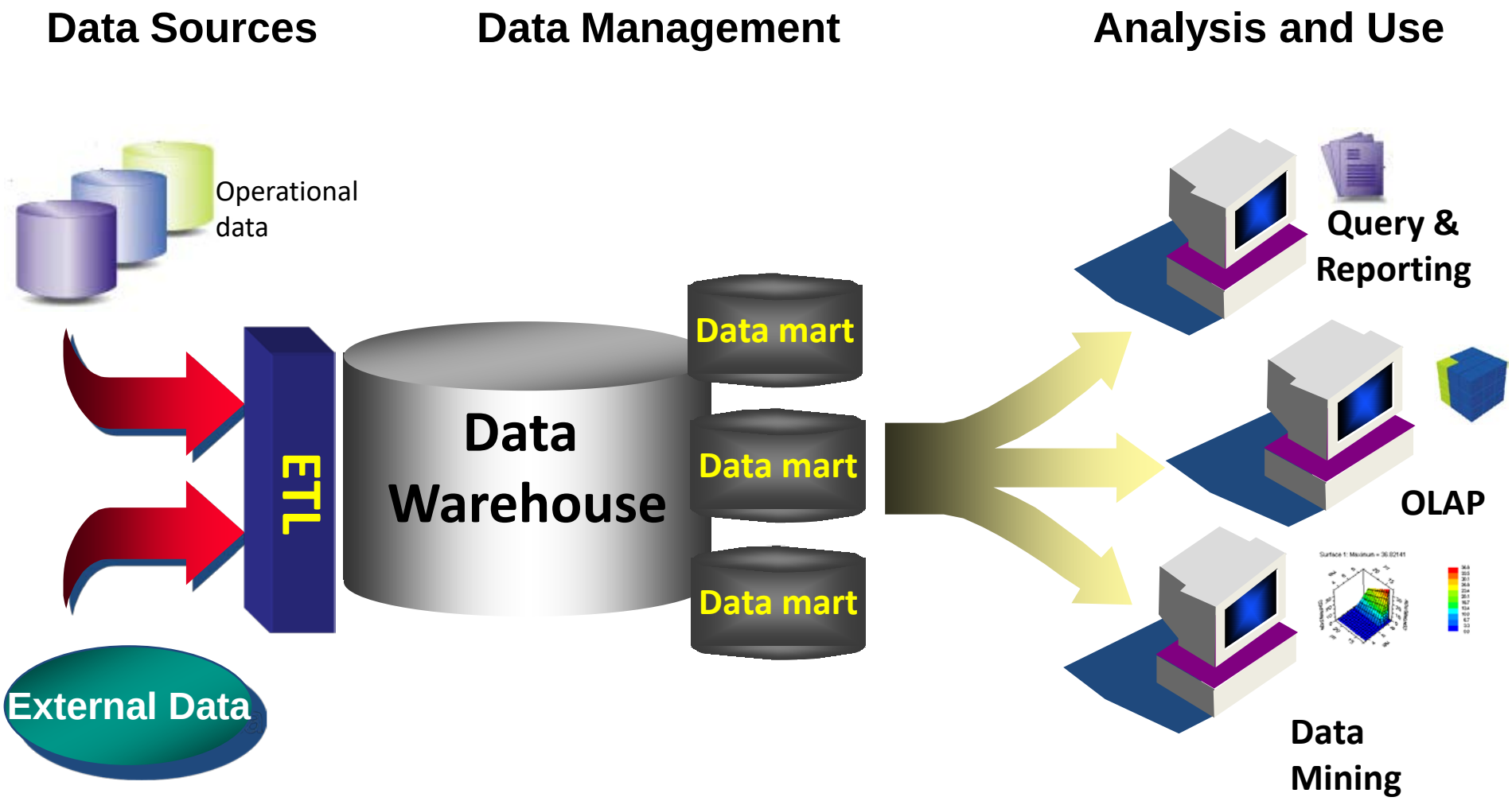
- **Drive company strategy**
  - being able to connect planning to measuring of impact (do not manage «blindly»)
- **Growth and competitiveness:**
  - anticipate market trends and adapt R&D accordingly
  - better customer relationships through better-targeted offers
  - better leverage of customer potential (cross-/up-selling)
  - optimise business processes

business drivers

- **Single point of truth**
  - no by-pass reporting, consistent data
- **Cost reduction**
  - faster access to information
  - automation of reports, self-service BI

technical drivers

Business Intelligence



Making Informed Decisions

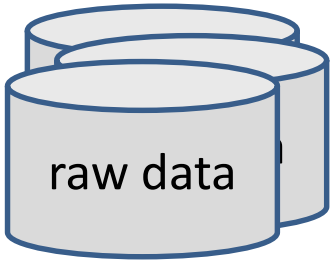
BI and fact-based decision making

- Fact-based decisions are based on information
- BI supports decision making by providing that information, usually in the following way:
  - the human decision maker (HDM) formulates the decision problem
  - the HDM identifies which information is needed to make an informed decision
  - the HDM consults a BI tool to get the answers, usually by querying or browsing (e.g. OLAP)
  - the HDM uses the answers to take an informed decision

Decision making

- **Decision making =** *The action of selecting among alternatives to achieve a goal*
  - ♦ each alternative leads to a different future
  - ♦ what is needed is the ability to predict the futures
- **Options:**
  1. Decide based on gut feeling
     - cheap in the first place
     - risk of low-quality decisions
  2. Experiment with real system (try out)
     - risky
     - time-consuming
  3. Decide based on the past:
     - data collection is time-consuming
     - difficult to determine when to stop and make a decision

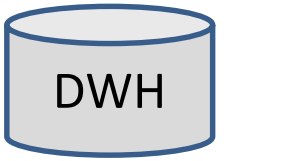
Data-driven vs. business-driven BI



We have data. What can we do with it?

ETL

Consolidate and integrate data

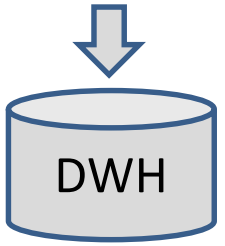


Analyze data

We have questions. How do we get answers?

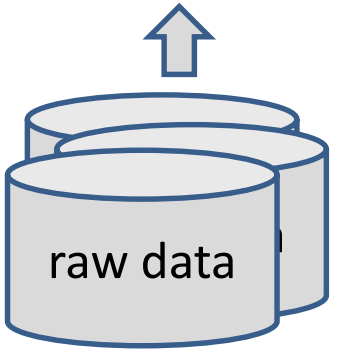
?

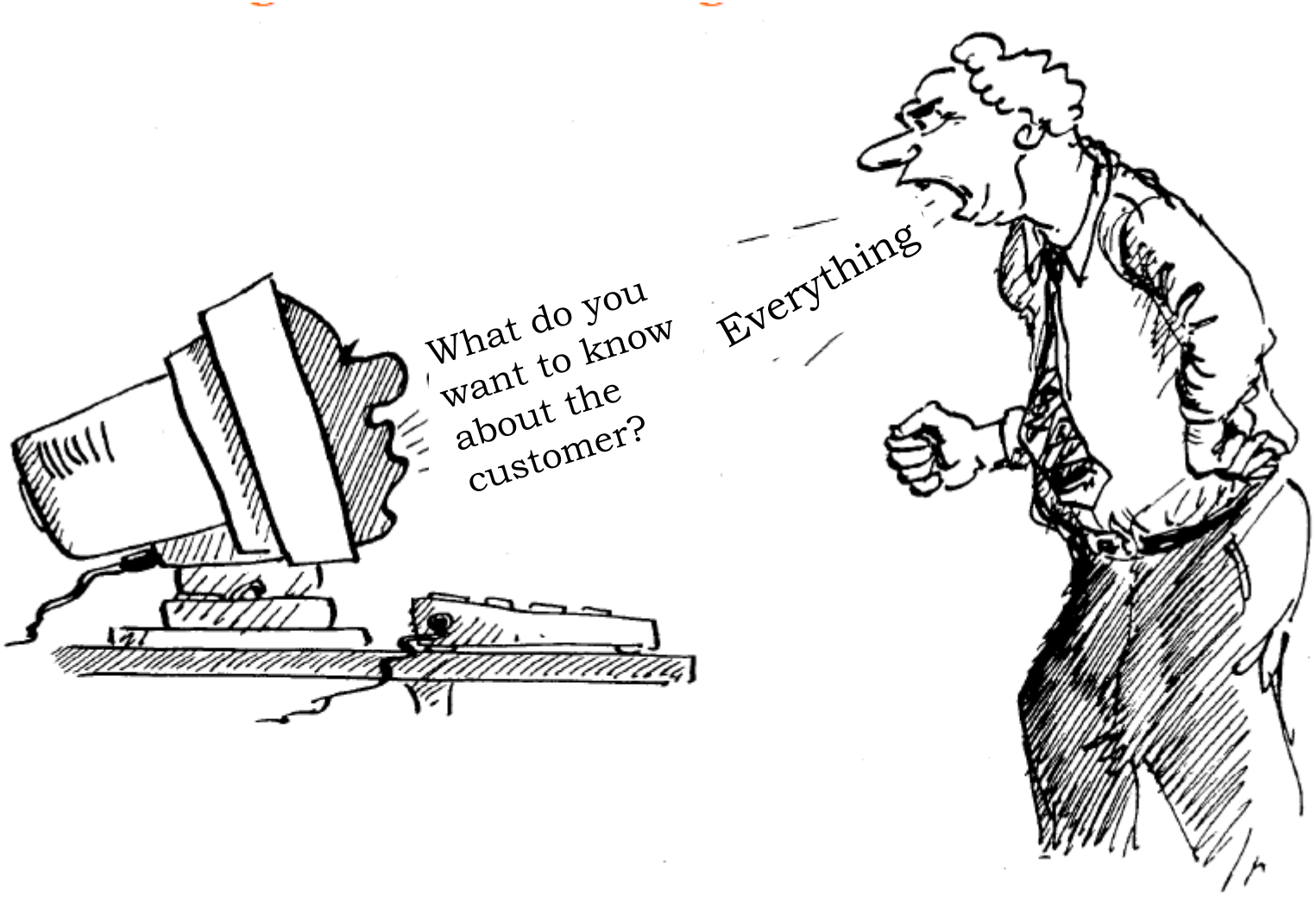
What data do we need to answer the questions?



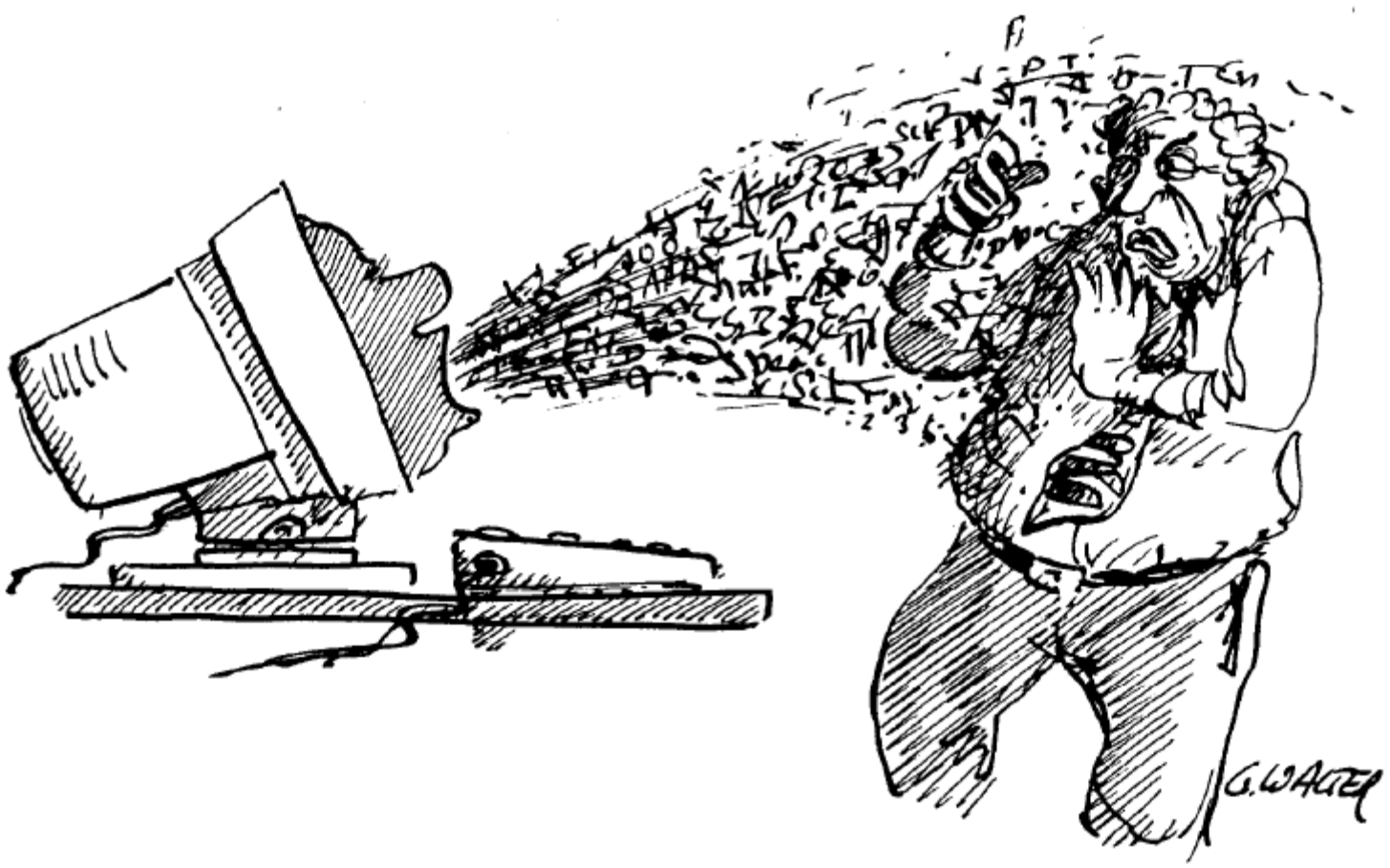
ETL

Collect and consolidate data





adapted from slides by Dani Schnider

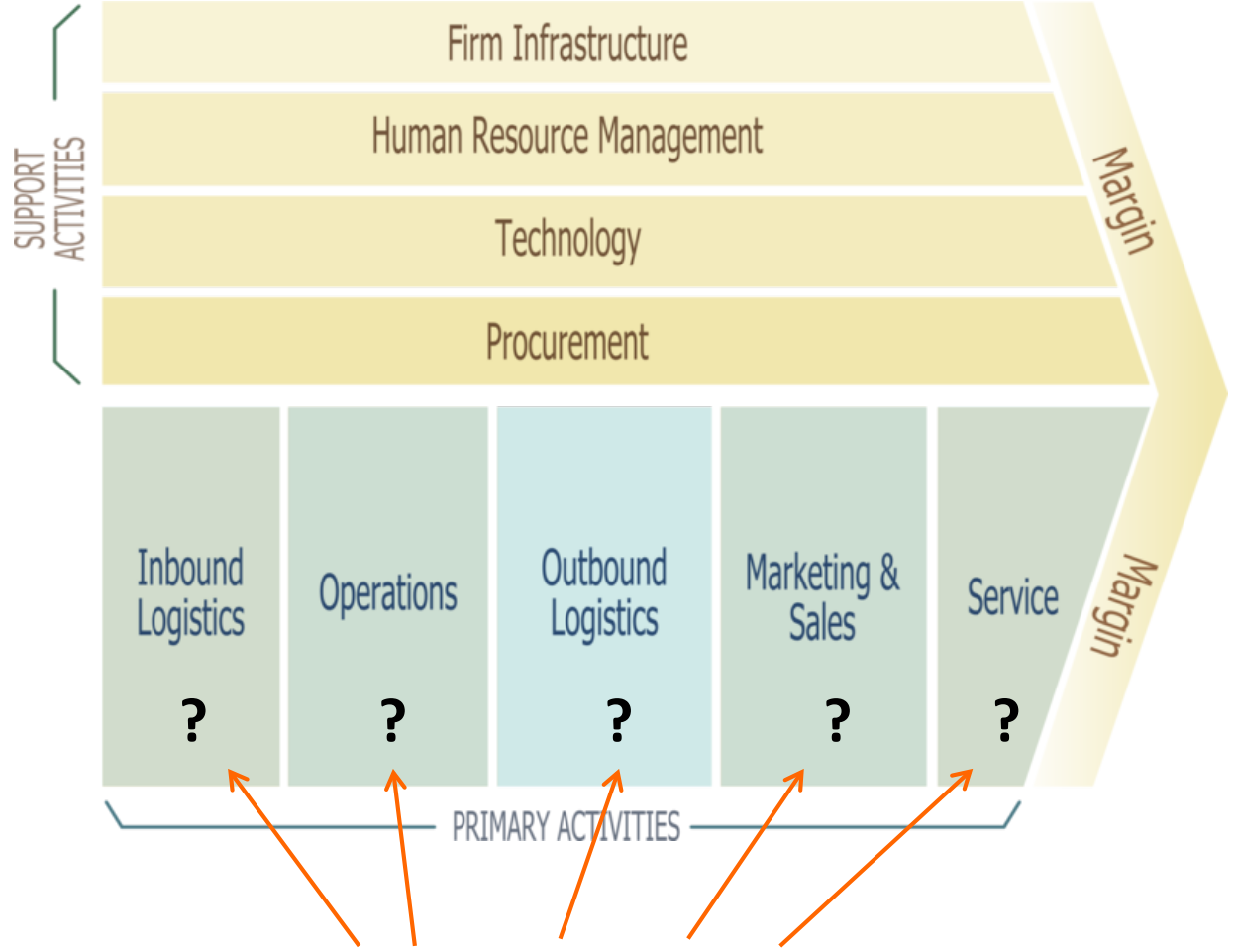


adapted from slides by Dani Schnider

Strategic decisions…

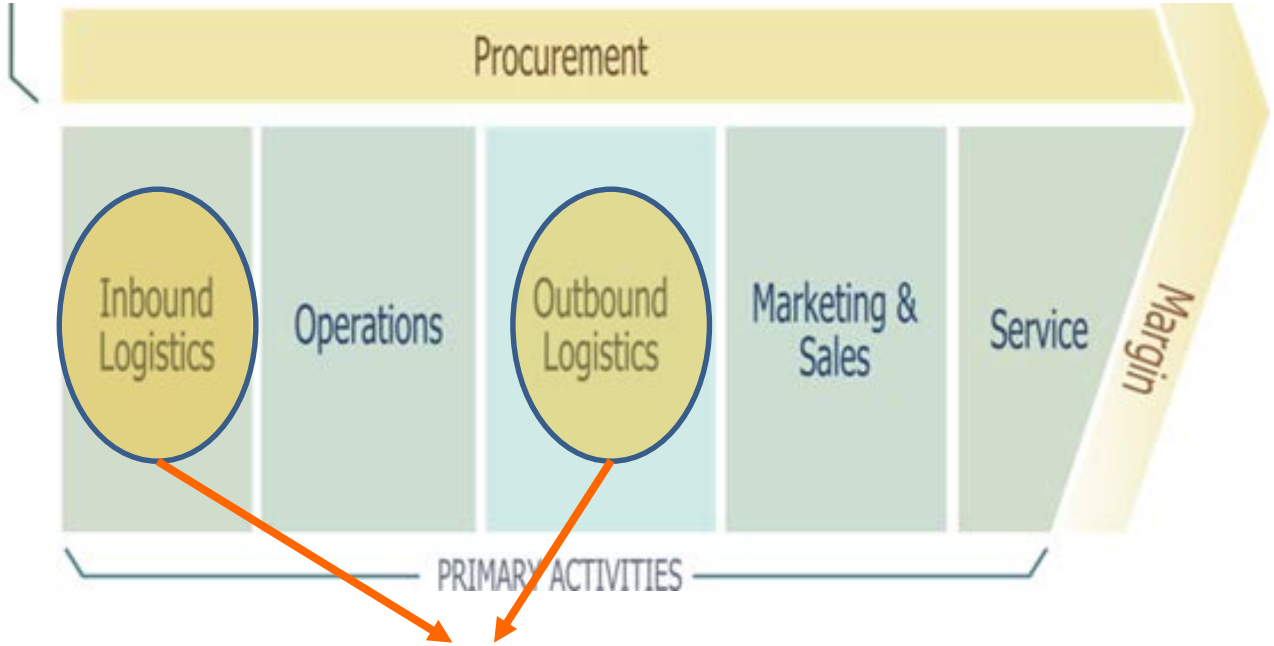
- Business Performance Management:
  - “how to perform better as a company?”
  - BI helps to achieve that by enabling measurement of achievement of strategic goals via Key Performance Indicatore (KPIs)
    1. Define **strategy**
    2. Define **goals**
       - e.g., identify key business processes to be improved, derive (concrete) strategic goals
       - for each goal, define KPIs and target values
    3. **Measure**
       - current values of KPIs (dashboard/cockpit)
       - analyse / compare current to targeted values
    4. **Decide**…
       - understand the (possible) deviation of KPI values from target!

Operative decisions: where BI creates value…



decisions to be taken in corresponding business processes?

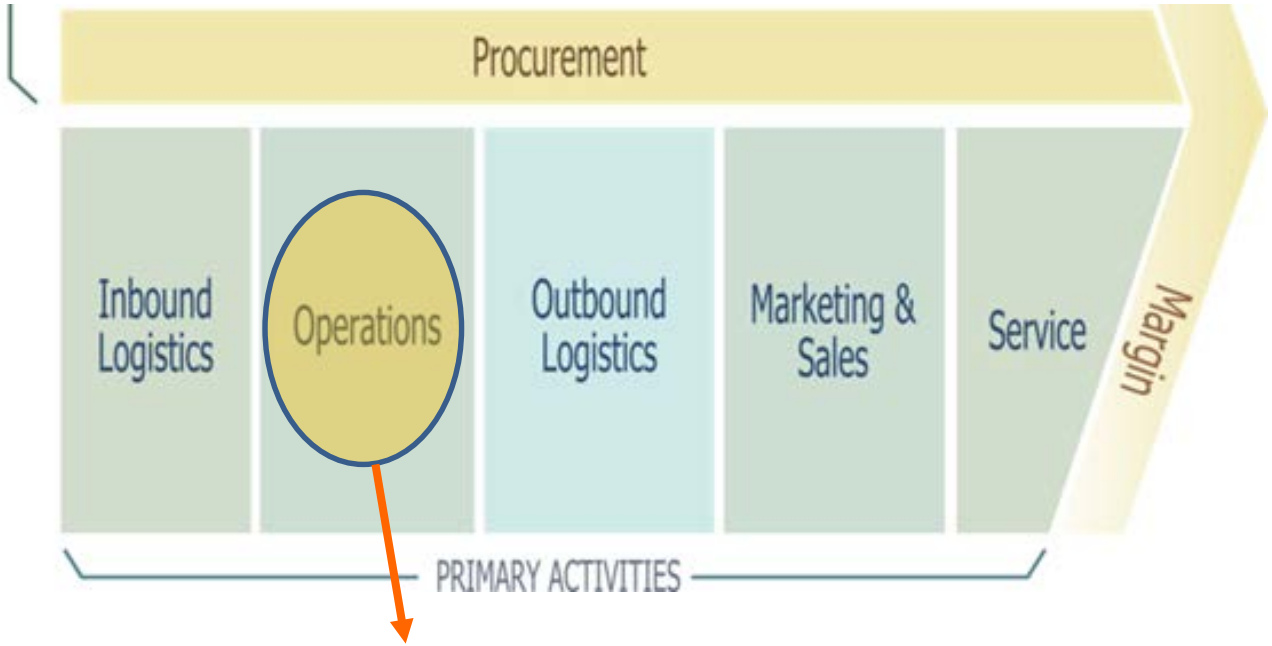
Operative decisions: where BI creates value…



Logistics:
the process of planning, implementing and controlling the efficient, effective flow and storage of goods, services and related information from the point of origin to the point of consumption for the purpose of conforming to customer requirements

- ***how to best use resources (inbound)?***
  - which parts to order, in which quantity, at what time, from which supplier?
- ***how to optimise processes (outbound)?***
  - which route/channel to use, how to schedule deliveries?

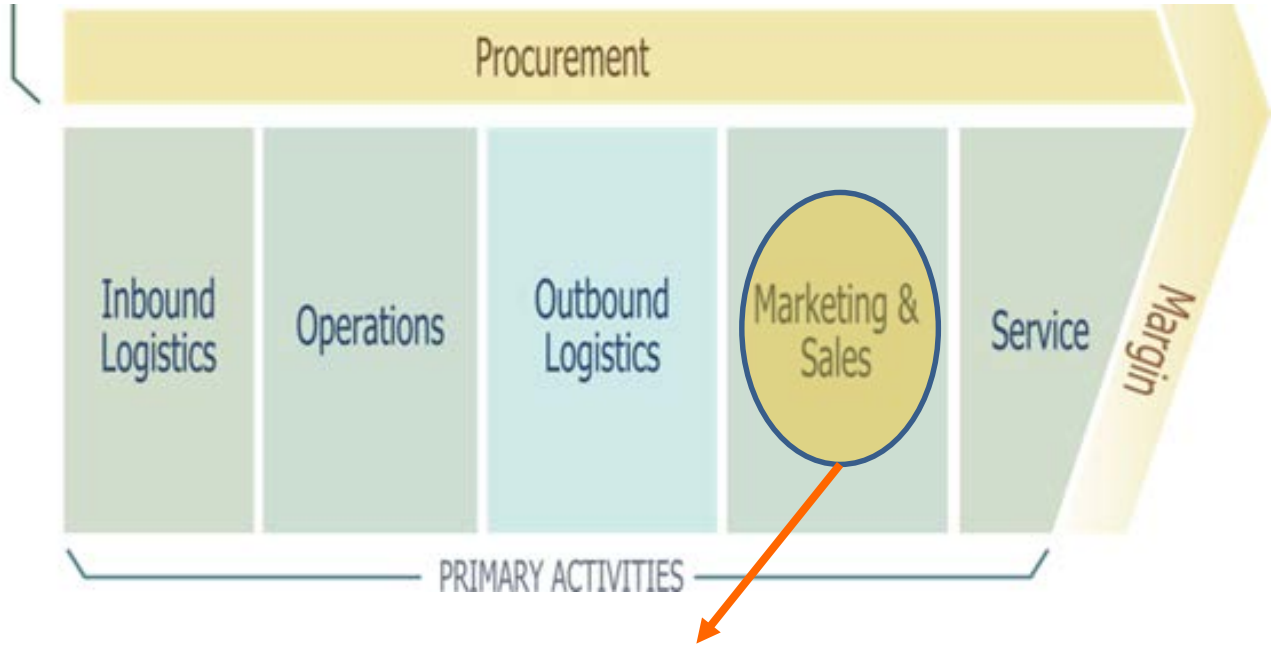
Operative decisions: where BI creates value…



Operations:
activities associated with the functions of transforming inputs into the final product form, such as machining, packaging, assembly, equipment maintenance, testing, printing, and facility operations.

- ***how to improve efficiency and effectiveness of processes?***
  - which resources to allocate, in which quantity, …

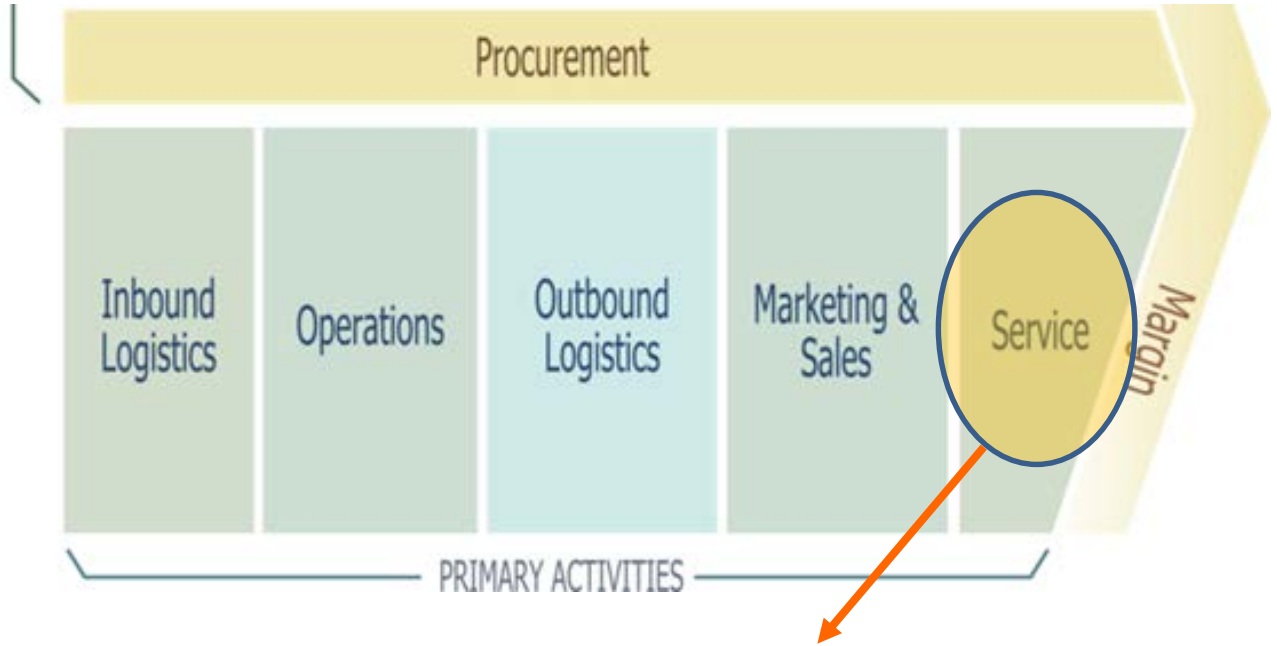
Operative decisions: where BI creates value…



Marketing/Sales:
activities associated with the functions of providing the means by which buyers can purchase the product and inducing them to do so, such as advertising, promotion, quoting, pricing, channel and sales force management.

- ***how to understand and best address the market?***
  - which customers to approach with a campaign?
  - cross-selling: which offers to make?
  - where to place products in stores?
  - client profitability: which customers to treat with special care?
  - pricing decisions

Operative decisions: where BI creates value…



Service:
activities associated with the functions of providing service to enhance or maintain the value of the product, such as installation, repair, training, parts supply, and product adjustment.

- ***how to meet customer requirements and anticipate problems?***
  - which distribution channels to use for service delivery?
  - which quality problems to address first?
  - Attrition prediction: which customers to retain with special offers?

Question types – summary

- Types of questions identified:
  - **query** for particular numbers or facts
    - *e.g. list of all policies that have been lost, list of all complaints, list of treatments that have been billed twice, list of high-value customers…*
  - **compute a measure or KPI by aggregating numbers**
    - *e.g. cost, margin, turnover, profitability*
  - **analyse KPIs / facts in different ways**
    - *e.g. sales/bookings by product/customer/sales rep/time*
    - *e.g. receipts/failures/stock by part/supplier*
    - *e.g. number of clicks/purchases by buyer/seller/page*
  - **predict**
    - *e.g. predict fraudulent transactions/claims*
    - *e.g. predict if a customer will buy a product*
    - *e.g. detect types of customers or types of complaints*

Where questions come from

- Generally speaking, companies need information to
  - monitor and improve **performance**
  - recognize and mitigate **risks**
  - recognize and seize **opportunities**

- All this can happen both on a strategic and an operative level

Monitor and improve performance

- **Strategic level**: be able to measure if strategic goals are achieved
  - e.g. be able to measure the satisfaction of our customers over the last year
    → so that we can decide to change our customer service model

- **Operative level**: monitor performance within certain business processes, in small time intervals
  - e.g. find out that/why (individual) customers are not satisfied today
    → so that we can decide to call them and find a solution

Recognise and mitigate risks

- **Strategic level**: be able to recognise general threats to our business
  - e.g. become aware that sales in certain product category are dropping dramatically (which is threatening our whole business)
    → so that we can revise our product portfolio

- **Operative level**: be able to recognise risks related to individual processes, customers, suppliers, employees, …
  - e.g. in telecommunications, be able to predict if a customer is going to cancel (or not renew) her contract
    → so that we can decide to make a special offer to that customer

Recognise and seize opportunities

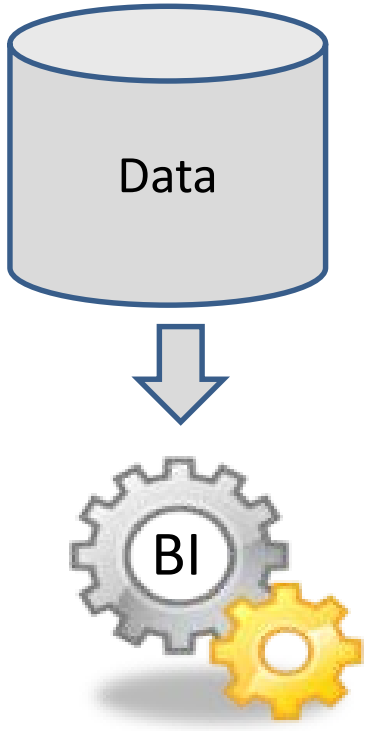
- **Strategic level**: be able to recognise general opportunities for our business
  - e.g. become aware that (potential) customers are asking for a certain kind of product or product feature in social media
    → so that we can decide to develop such a product

- **Operative level**: be able to recognise opportunities related to individual process instances, customers, suppliers, employees…
  - e.g. recognise that we can cross-sell a certain product to an existing customer
    → so that we can decide to make the customer aware of that product

Data Warehouse – BI Backend

Remember…

- … transform **raw data** into meaningful and useful **information**…
- **Raw data** is the starting point!



Where the data comes from

- Internal data sources:
  - (Transactional) standard business applications: sales data, accounting, SCM, ERP, CRM, …
  - Legacy databases, spreadsheets
  - Web data: clickstreams from server logs, application logs
  - textual documents (from DMS, CMS, intranet, email,…)

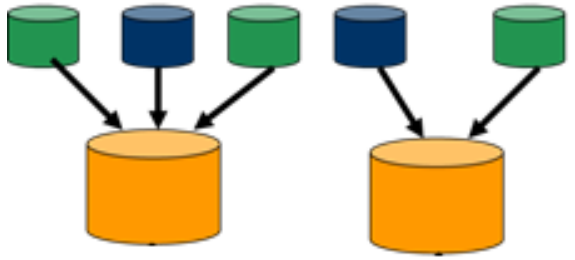
structured-ness

- External data sources:
  - Web and web 2.0

CRM – Customer Relationship Management
SCM – Suppy Chain Management
ERP – Enterprise Resource Planning

BI tools – backend

- Observations:
  - ♦ many questions involve multiple (types of) data
  - ♦ sometimes the data can be expected to originate from more than one source system
  - ♦ for answering the questions, data from various sources needs to be connected
    - example: «Which is the best way to distribute product XYZ to customers?" → involves information about customers (e.g. profitability, behaviour) as well as about channels (e.g. cost of each channel)



Data warehouse

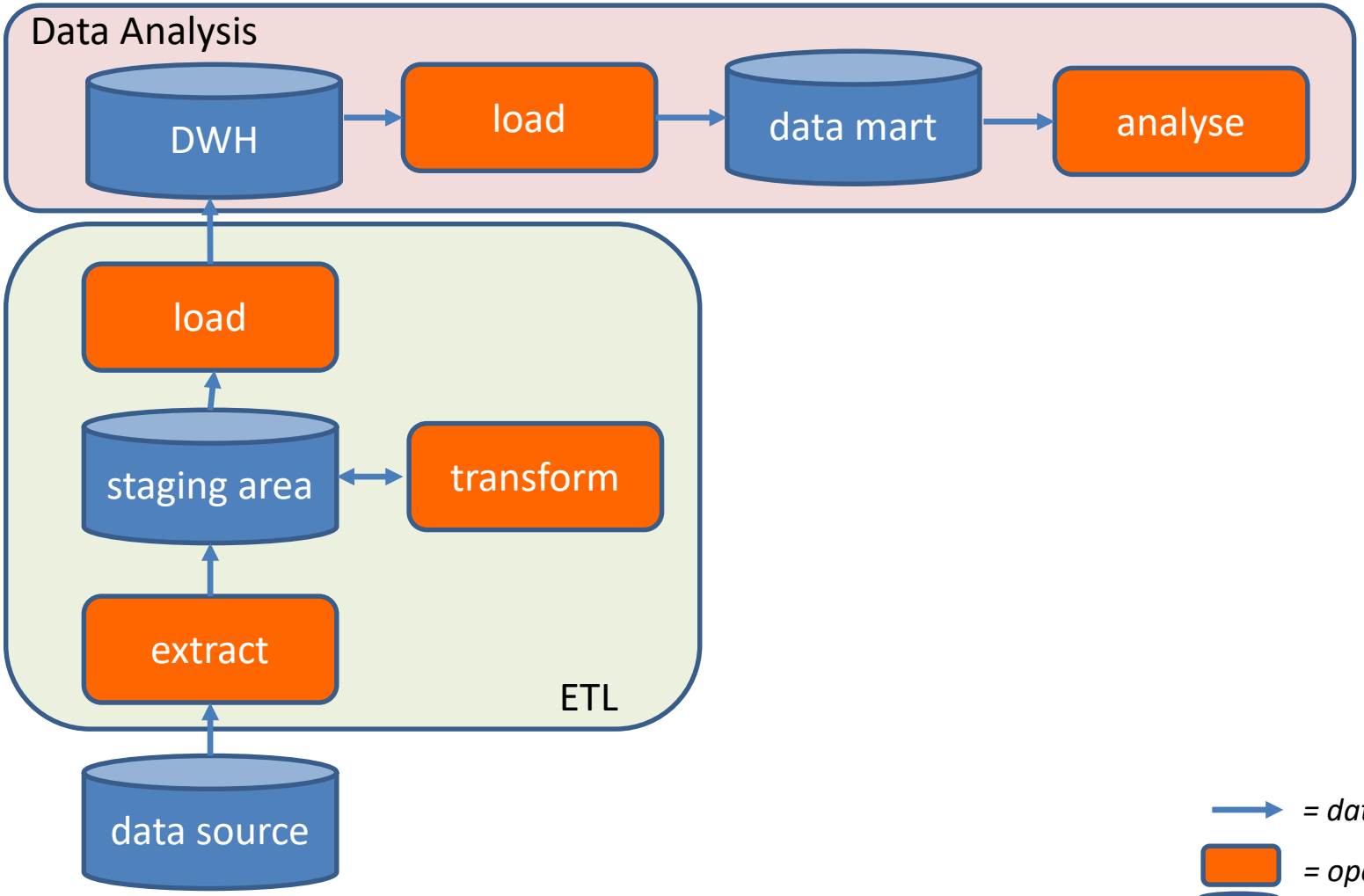
- *A data warehouse is*
  - *"a **copy** of transaction data specifically structured **for querying and reporting" (Kimball et al. 2008)***

  or

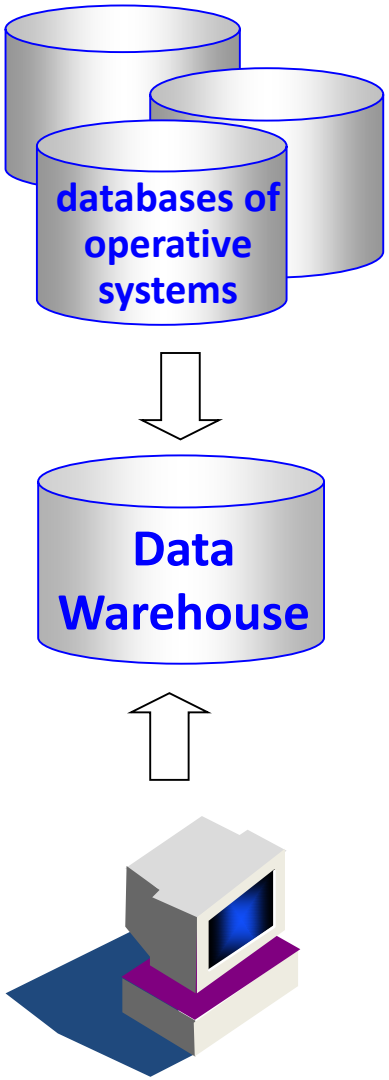
  - *"an environment […] comprising a data store and […] tools for data extraction, loading, storage, access, query and reporting […] to **support decision-oriented management queries**" (Bashein/Markus, 2000)*

- *Data warehousing is the entire process of data **extraction, transformation, and loading** of data to the warehouse and the **analysing** the data by users and applications.*

Reference Architecture – Overview



Data Warehouse



- A Data Warehouse is a database that supports strategic decisions by providing
  - high-volume and regular excerpts from operative databases
  - often aggregated[1]
  - also for ad hoc[2] analysis

- Essential characteristics (Inmon 2005):
  - Subject oriented
  - Integrated
  - Time variant
  - Nonvolatile

1) combined, consolidated (e.g. als sum, average, indicators)
2) without preparation, in contrast to standardized analysis

Integration, Time variance

- **Integration:** provide a «single version of the truth»
  - remove redundancy, inconsistency, semantic contradictions (see ETL processes later)
- **Time variance:** DWH maintains historic data, data is collected over a long time period with information when it was valid

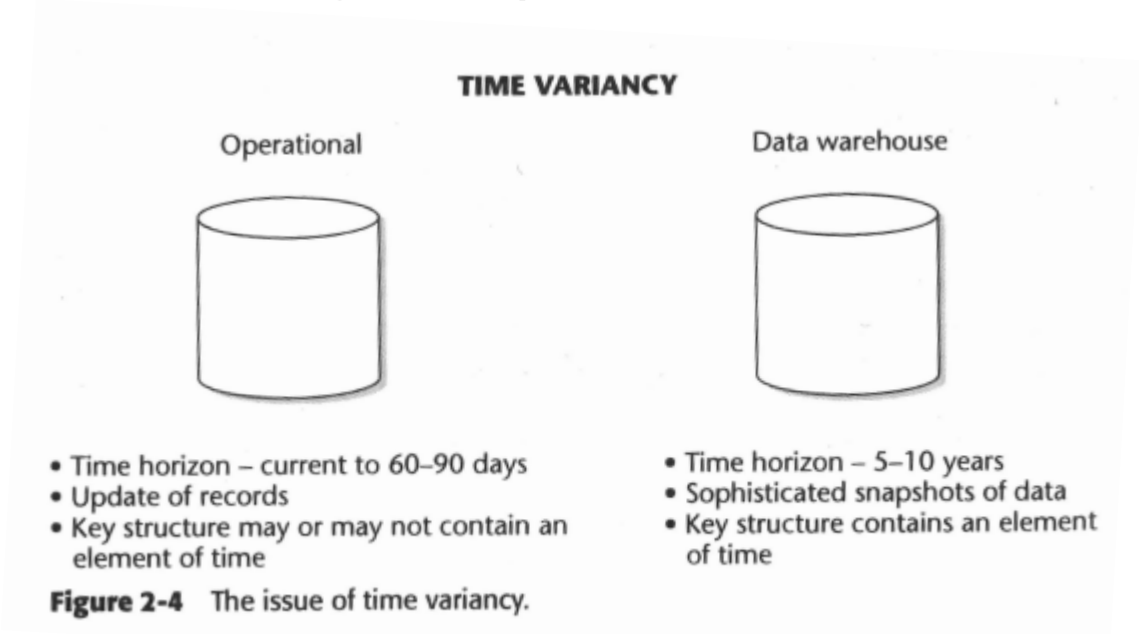


Figure 2-4 The issue of time variancy.

figure taken from B. Inmon: Building the data warehouse.

Non-volatility

- **Non-volatility:** data is not updated by end users on a regular basis
  - bulk loading, «read-only» access

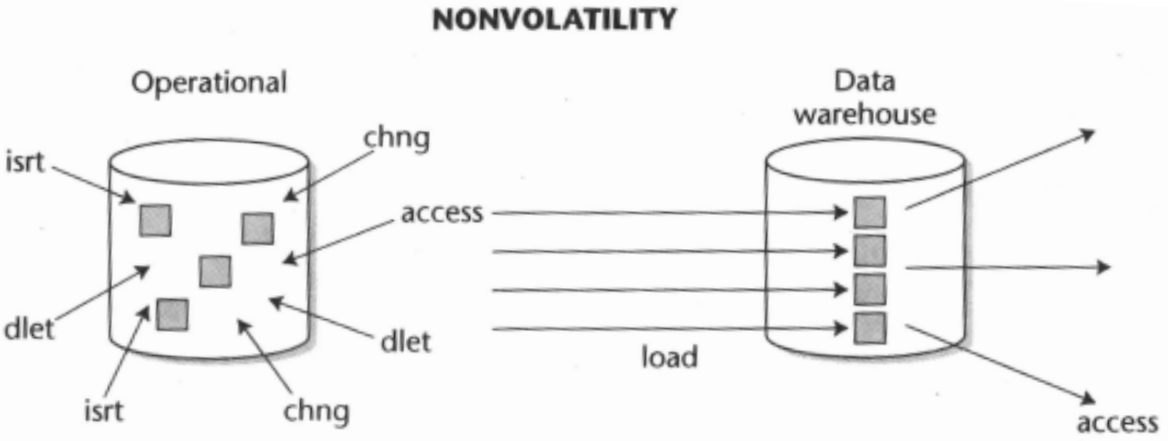


Figure 2-3 The issue of nonvolatility.

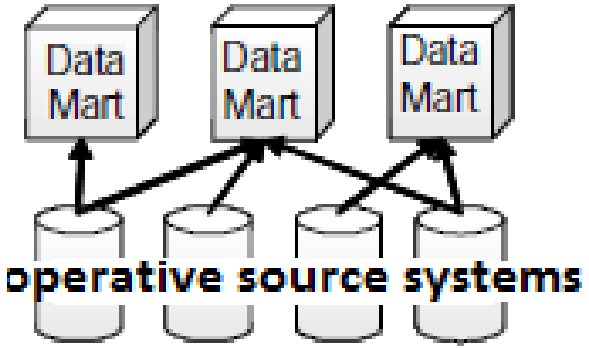
figure taken from B. Inmon: Building the data warehouse.

Data Marts

A **data mart** stores data for a *limited* number of subject areas. It is used to support *specific* applications.

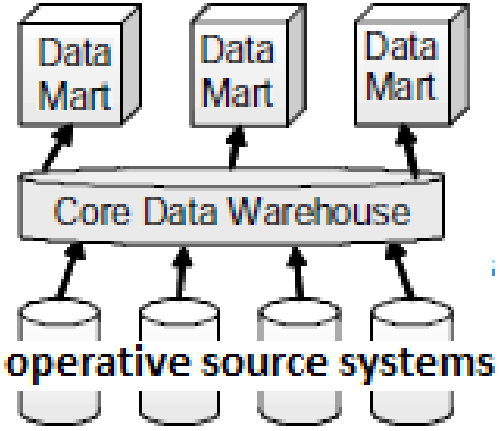
Independent data mart

- created directly from source systems
- Possibly joined into a data warehouse later



Dependent data mart

- Source data are aggregated into a data warehouse
- data marts are created as subsets (e.g. for efficiency reasons)



following Kemper et al. fig. 2.4

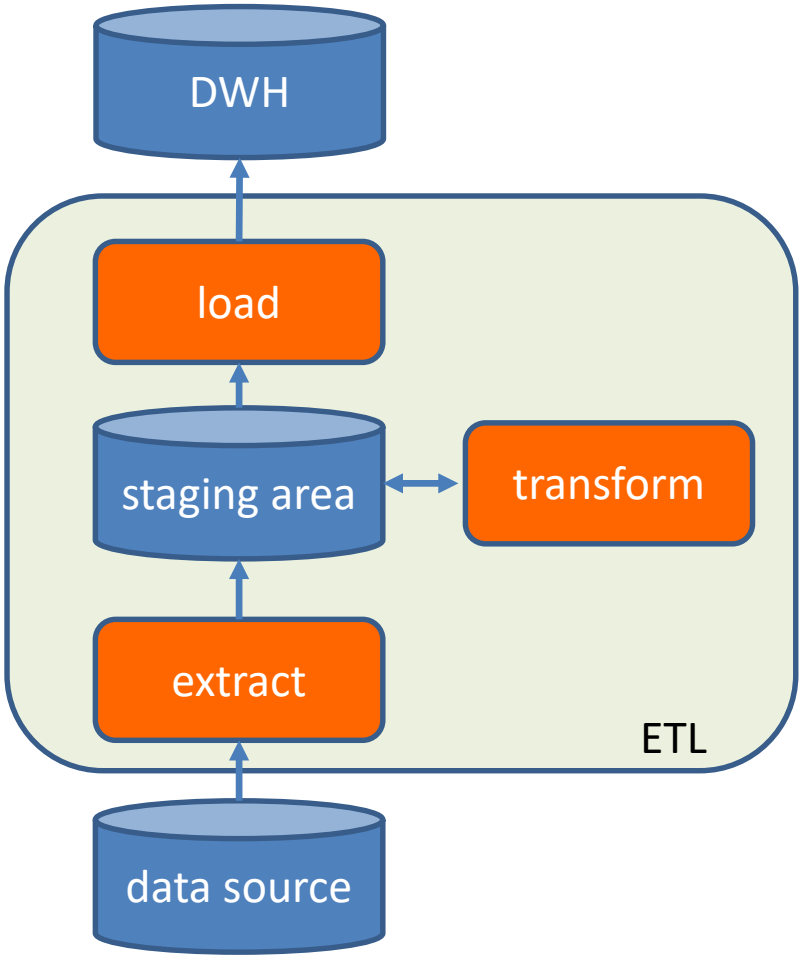
Data Marts: Departmental vs. Enterprise?

- Question: should data marts be enterprise-wide or departmental?

  - Answer 1: Data marts should be organised around business processes (orders, invoices,…), not department boundaries!
  - Answer 2: … but they don't necessarily have to be enterprise-wide (depends on the business process)!

ETL – Extract,Transform, Load

ETL process



- The process of
  - **extracting** relevant data from source systems
  - **transforming** the data into the target format defined for the DWH or data mart
  - **loading** the data into the DWH

ETL process – Extraction

- **Extraction =** Collecting information to be added into DWH
  - Read source data into staging area
  - Control the selection of data that should be copied
- Extractions can happen…
  - … periodically
  - … on human request
  - … event-based (e.g. when a certain number of changes has occurred)
  - … upon each change

ETL process – Transformation

- **Transformation =** adapting data, data quality and schemas to the requirements of users
  - **Filtering:** remove syntactic and semantic defects of data
  - **Harmonisation:** map source schemas to the target schema of the DWH
    - syntactic harmonisation: schema integration + data integration
    - business harmonisation
  - **Aggregation:** aggregate data along dimension hierarchies (e.g. «customer», «customer segment», «all»)
  - **Enrichment:** pre-compute values of frequent interest and store as new attributes
    - on the basis of harmonised/aggregated data

ETL- Filtering: Error Classes

	1. class: Automatic identification	2. class: (Semi-) automatic identification
Syntactic	- Known formatting variants (abbreviations, date formatting etc.) - encoding problems	- Spelling variants/errors
Semantic	- Missing values (incompleteness) - redundancy (duplicates) - non-unique identifiers - missing referential integrity	- Incorrectness (e.g. outliers) - inconsistencies (violating business rules or contradictions) - dummy values

adapted from Kemper et al. fig 2.8

ETL- Filtering: Correction Measures

- Correction measures
  - **1st class:**
    - *incompleteness:* define rules to fill in missing values (e.g. fill sales values with ones from previous month or planned ones)
    - *duplicate detection:* often there is a combination of values that unambiguously identifies a record => if these are the same, match!
    - *formatting/encoding/non-unique id issues*: simple scripting
  - **2nd class:**
    - *spelling variants/errors:* use string similarity, thesauri (extend as you go along)
    - *general incorrectness: hard to spot automatically, can define automatic sanity checks…*
    - *outliers:* statistic analyses
    - *inconsistencies:* checks based on business rules

ETL - Harmonization

- These are parts of tables that should be integrated in a DWH. What harmonisation tasks/problems do you see?

CustomerID	Name	City
11	Peter	Rom
15	Paul	Camerino
18	Mary	Olten
25	Joe	Bern

PurchaseID	CustomerID	Date	ProductID
1002	11	5 May 2015	SE4256
1003	18	5 May 2015	EA4516
1004	11	6 May 2015	EA4516
1005	25	6 May 2015	RG3452

ComplaintID	Complaint	Person
36536	Return	George
44363	Failure	Paul
46344	Failure	John

ETL – Harmonisation: Schema integration

Problem	characteristics	Example: data source 1	Example: data source 2	Solution
Synonyms	Attributes with different names have identical meaning	Attribute «employee» contains employee name	Attribute «staff» contains employee name	Choose an attribute name
Homonyms	Same attribute name refers to attributes with different meaning	Attribute «partner» refers to name of customer	Attribute «partner» refers to name of supplier	Choose different attribute names

ETL – Harmonisation: data integration (1)

Problem	characteristics	Example: data source 1	Example: data source 2	Solution
Deviating primary keys (synonyms)	Same entity has different id in different operational DBs	Customer «Smith» has id 376_ACC in accounting application	Customer «Smith» has id 7843_CC in call center application	Record linkage: identify identical entities via overlapping attribute values; use mapping table

- How to detect entity identity?

ETL – Harmonisation: data integration (2)

- Mapping tables: allow to map updates in sources to DWH records

AD_SYS	...	customer	LOADTIME
AD-FX8257		Müller	31DEC2009:23:03:08
AD-FH2454		Meier	31DEC2009:23:03:08
AD-FX7059		Schulz	31DEC2009:23:03:08
AD-FT2567		Schmitz	31DEC2009:23:03:08
...	...	...	...

AC_SYS	customer	customerStatus
3857_ACC	Müller	A
3525_ACC	Meier	A
3635_ACC	Schulz	A
3566_ACC	Schmitz	B
...	...	...

CC_SYS	cust_grp	customer	LOADTIME
59235395	retail	Müller	31DEC2009:23:03:08
08485356	industry	Meier	31DEC2009:23:03:08
08555698	industry	Schulz	31DEC2009:23:03:08
85385386	retail	Schmitz	31DEC2009:23:03:08
...	...	...	...

AD=customer service
CC = call center
AC = accounting

Kunde_ID	cust_id	...	AD_SYS	CC_SYS	AC_SYS	...	LOADTIME
0001	Müller		AD-FX8257	59235395	3857_ACC		31DEC2009:23:03:08
0002	Meier		AD-FH2454	08485356	3525_ACC		31DEC2009:23:03:08
0003	Schulz		AD-FX7059	08555698	3635_ACC		31DEC2009:23:03:08
0004	Schmitz		AD-FT2567	85385386	3566_ACC		31DEC2009:23:03:08
...	...	...	...	...	...	...	...

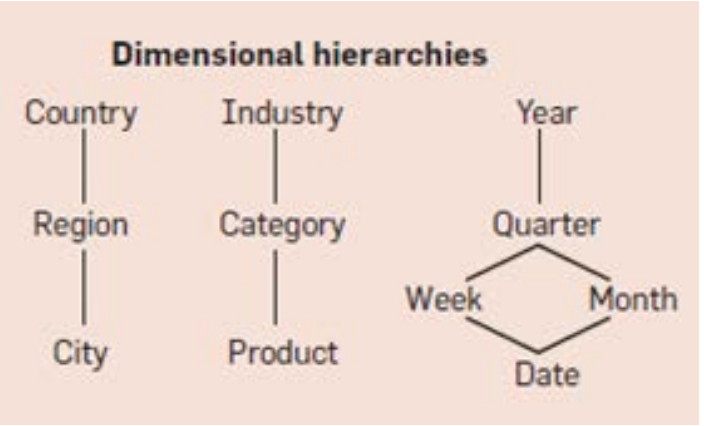
adapted from Kemper et al.

ETL – business harmonisation

- **adjust figures/values**
  - consolidate figures from various databases based on their (business) meaning, e.g. apply rules to map location- or department-specific value deviations
  - convert currencies and units (e.g. inch → cm)
- **adjust granularity**
  - decide for a level of granularity (e.g. monthly or quarterly)
  - harmonise according to period (source systems may have differing granularity, e.g. quarters vs. years)
  - aggregate all values on that level (e.g. sum all records/receipts of one day together)

ETL - Aggregation

- Aggregate data based on dimensional hierarchy
  - usually, aggregates are pre-computed for performance reasons
  - introduces «controlled redundancy»
  - aggregates become invalid when hierarchies and/or source data change…

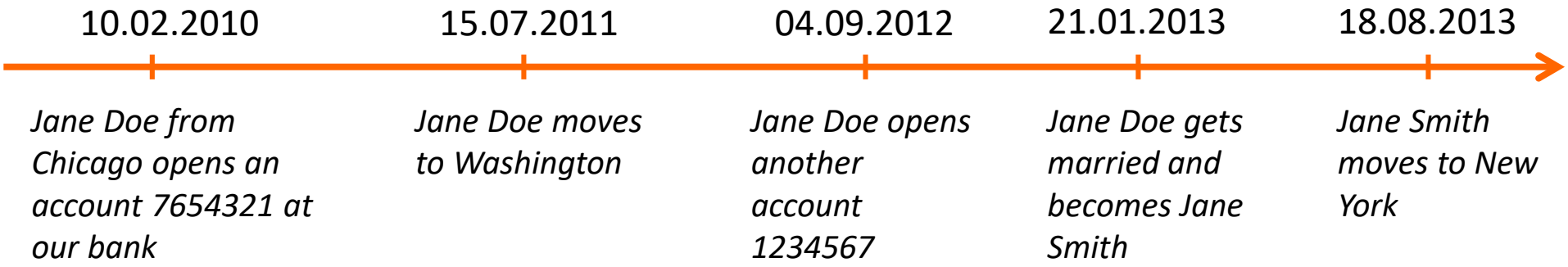


ETL - Enrichment

- add new attributes that are functions of existing data; compute these functions and store the result
  - sums, averages or more complicated computations (e.g. profitability)
  - based on harmonised and/or aggregated data
  - same motivation as aggregation: performance
  - introduces another «controlled redundancy»

Slowly Changing Dimensions

Example: customer dimension change



- who's the owner of the bank account 1234567?
  - as of today: Jane Smith from New York
  - as of 31.12.2012: Jane Doe from Washington
  - as of 31.12.2011: there is no such bank account

Type I: no history

Cust_id	Cust_name	Cust_city	…
1	John Allan	Chicago	…
2	Chris Lee	Boston	…
3	Jane Doe	Chicago	…

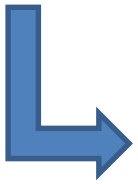
Cust_id	Cust_name	Cust_city	…
1	John Allan	Chicago	…
2	Chris Lee	Boston	…
3	Jane Doe	Washington	…

15.07.2011

old value is simple overwritten with new value

Type II: full history

Cust_id	Cust_name	Cust_city	Valid from	Valid to
1	John Allan	Chicago	10.03.2008	
2	Chris Lee	Boston	02.06.2010	
3	Jane Doe	Chicago	10.02.2010	



Cust_id	Cust_name	Cust_city	Valid from	Valid to
1	John Allan	Chicago	…	
2	Chris Lee	Boston	…	
3	Jane Doe	Chicago	10.02.2010	14.07.2011
3	Jane Doe	Washington	15.07.2011	20.01.2013
3	Jane Smith	Washington	21.01.2013	17.08.2013
3	Jane Smith	New York	18.08.2013	

18.08.2013

- every intermediate state is documented, validity range of values is signalled via «valid from», «valid to» attributes
- «valid from» becomes part of primary key

Type III: limited history

Cust_id	Previous Cust_name	Current Cust_name	Effective date cust_name	Previous Cust_city	Current Cust_city	Effective date cust_city
1		John Allan			Chicago	10.03.2008
2		Chris Lee			Boston	02.06.2010
3		Jane Doe			Chicago	10.02.2010

18.08.2013

Cust_id	Previous Cust_name	Current Cust_name	Effective date cust_name	Previous Cust_city	Current Cust_city	Effective date cust_city
1		John Allan			Chicago	10.03.2008
2		Chris Lee			Boston	02.06.2010
3	Jane Doe	Jane Smith	21.01.2013	Washington	New York	18.08.2013

- keeps the n previous values, each in a separate new column (in the example: n=1)
- effective date column(s) show(s) when the change occured