Data Warehousing

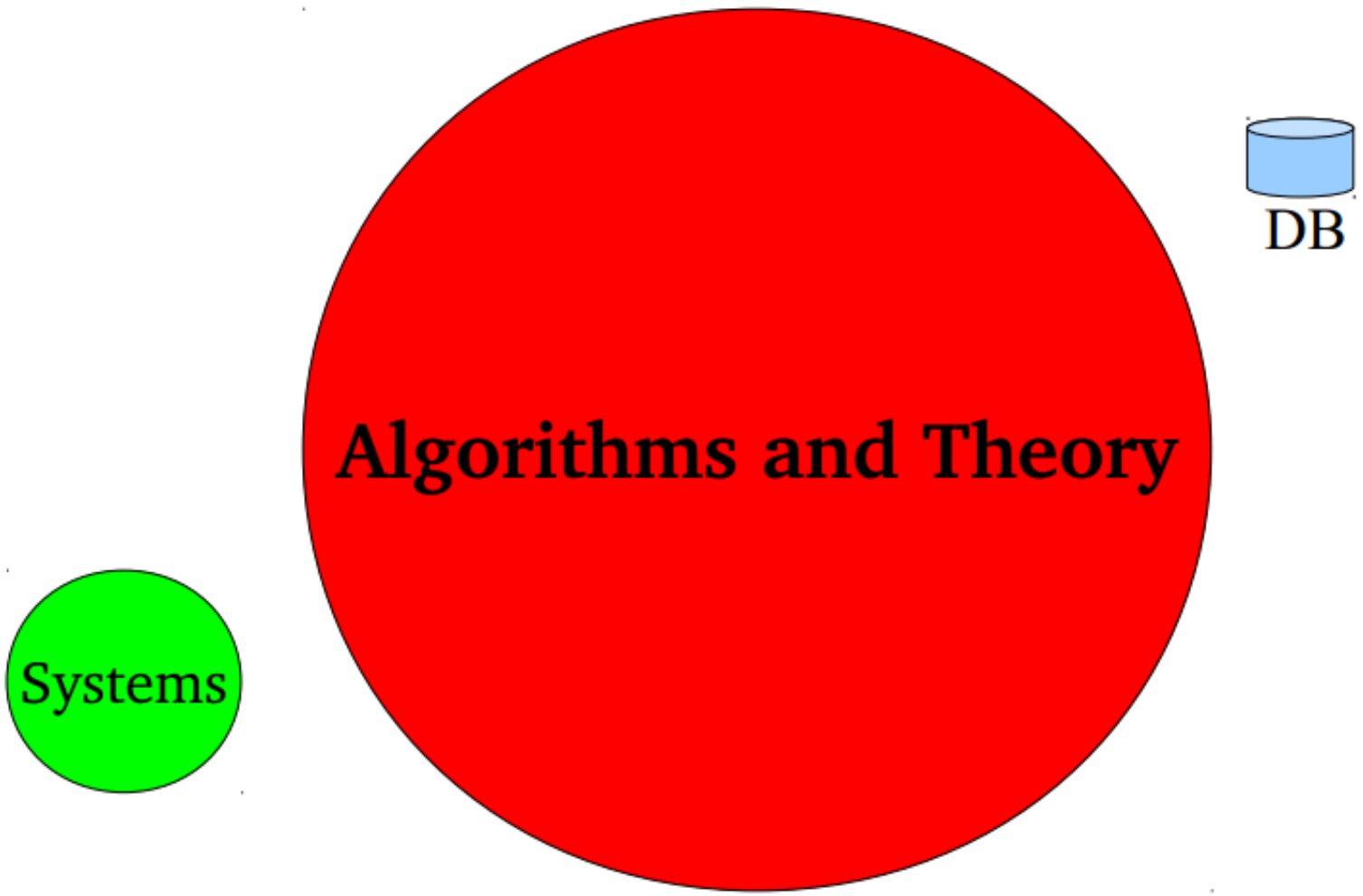
Business Intelligence

Outline

- **Introduction**
- **Business intelligence**
- **OLTP vs. OLAP**
- **Data integration**
- **Methodological framework**
- **DW definition**

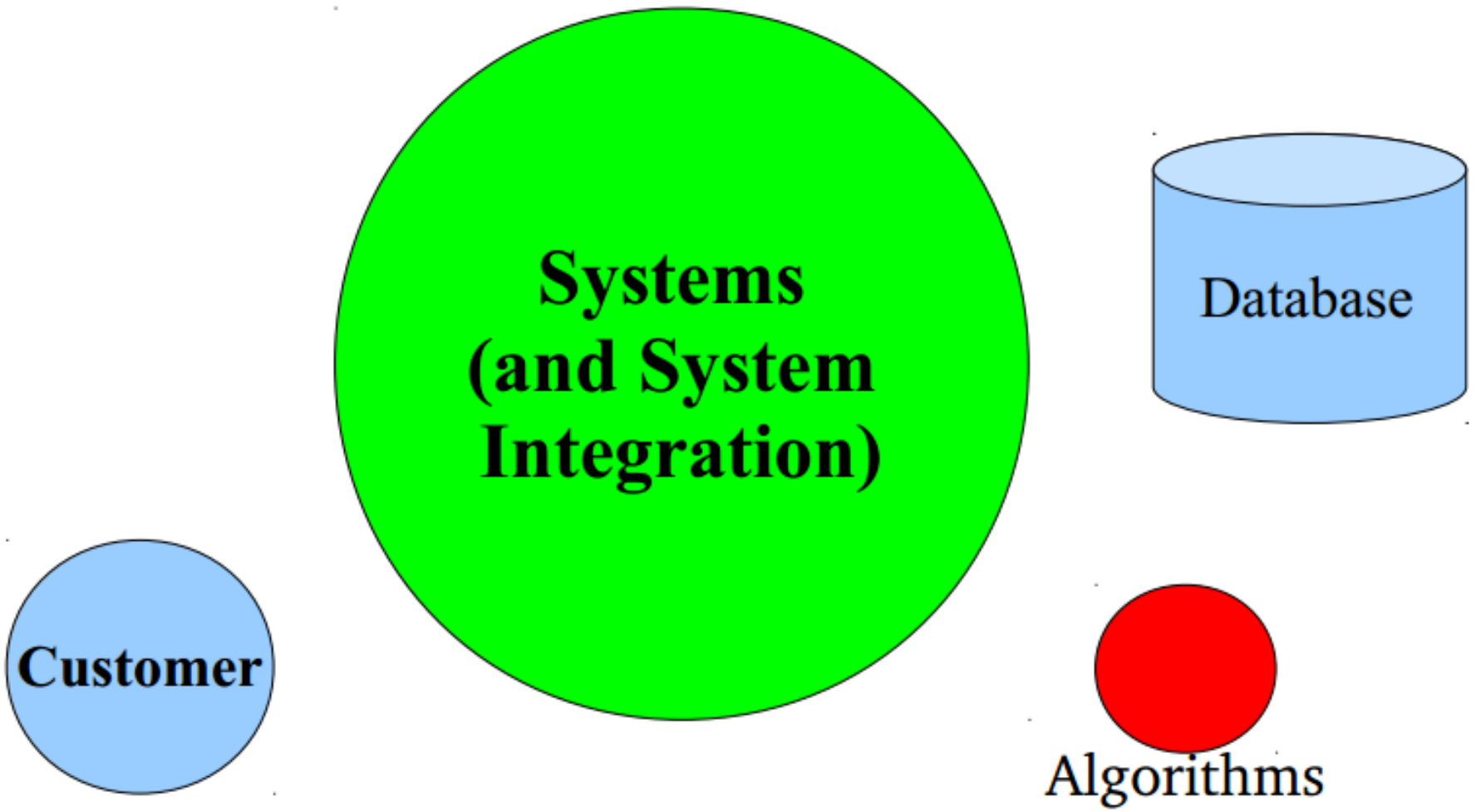
The Big Picture of DW/1

- What's important for researchers:



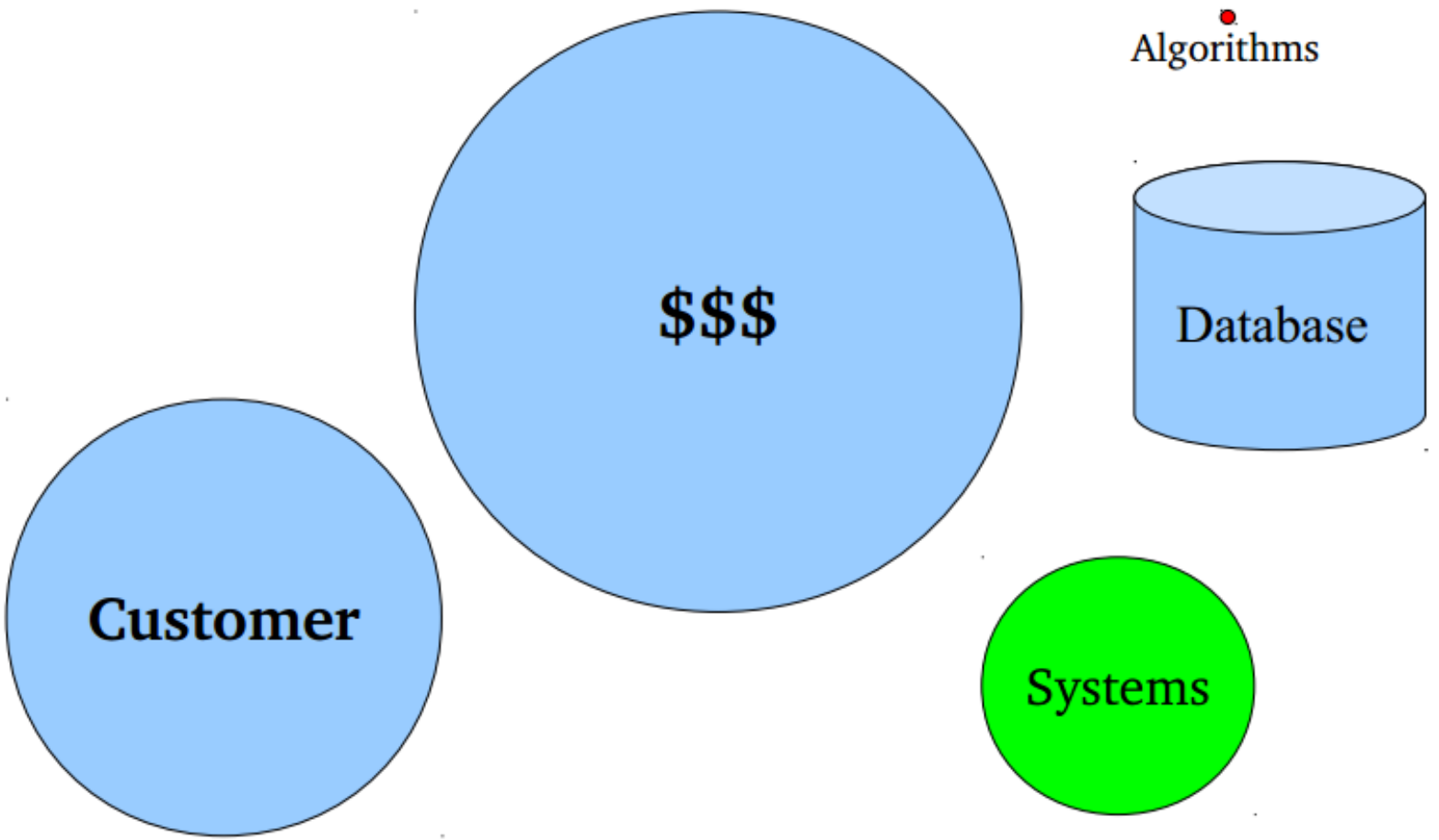
The Big Picture of DW/2

- What's important for real world applications:



The Big Picture of DW/3

- What's important for businesses:



Computer Science and Decision Making

- An **exponential increase** in operational data has made computers the only tools suitable for providing data for decision-making performed by business managers.

- The massive use of techniques for analyzing enterprise data made information systems a **key factor to achieve business goals.**

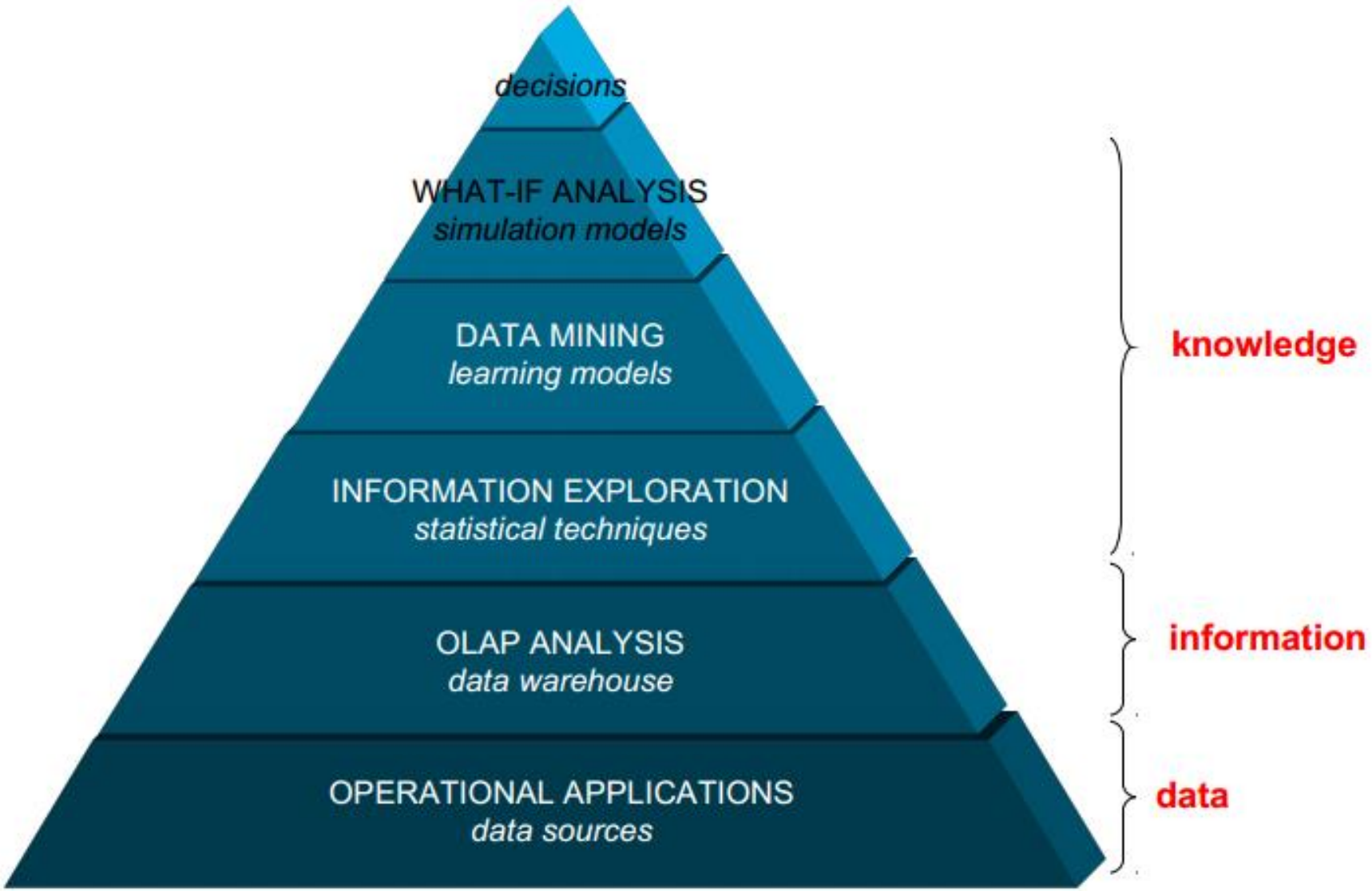
What is Business Intelligence?/1

- BI is a set of processes, tools, and technologies transform business data into timely and accurate **information to support decisional processes**
  - **Data Warehousing** (DW)
  - On-Line Analytical Processing (OLAP)
  - Data Mining (DM) and Data Visualization (VIS)
  - Decision Analysis (what-if)
- BI systems are used by decision makers to **get a comprehensive knowledge** of the business and to define and support their business strategies.
- The goal is to enable data-based decisions aimed at gaining competitive advantage, improving operative performance, responding more quickly to changes, increasing profitability, and, in general, creating added value for the company.

What is Business Intelligence?/2

- BI is the "opposite" of Artificial Intelligence (AI)
  - AI systems **make** decisions **for** the users
  - BI systems **help** users make the **right** decisions, based on the available data
  - Many BI techniques have roots in AI, though.

The BI Pyramid



Example BI Queries

- **Ql**: On October 11, 2010, find the 5 top-selling products for each product subcategory that contributes more than 20% of the sales within its product category.

- **Q2**: What was the percent change in market share for a grouping of my top 20% of products for the current three-month period versus same period year ago for accounts that grew by more than 20 percent in revenue?.

- Regular database models and systems are not suitable for this type of queries.

BI is Crucial and Growing/1

- The Web made BI more necessary:
  - Customers do not appear ”physically” in the store
  - Customers can change to other stores more easily
- Thus
  - You have to know your customers using data and BI.
  - Web logs make is possible to analyze customer behavior in more detailed than before (what was **not** bought?)
  - Combine web data with traditional customer data
- Wireless Internet adds further to this:
  - Customers are always ”online”
  - Customer’s position is known
  - Combine position and knowledge about customer => very valuable

BI is Crucial and Growing/2

1) **Complex and unusable models**
   - Many DB models are difficult to understand
   - DB models do not focus on a single clear business purpose
2) **Same data found in many different systems**
   - Example: customer data in many different systems
   - The same concept is defined differently
3) **Data is suited for operational systems**
   - Accounting, billing, etc.
   - Do not support analysis across business functions
4) **Data quality is bad**
   - Missing data, imprecise data, different use of systems
5) **Data are "volatile"**
   - Data deleted in operational systems (6 months)
   - Data change over time - no historical information

BI: Solution

- A **new analysis environment** with a **data warehouse** at the core, where data is
  - Integrated (logically and physically)
  - Subject oriented (versus function oriented)
  - Supporting management decisions (different organization)
  - Stable (data is not deleted, several versions)
  - Time variant (data can always be related to time)

Definition of a Data Warehouse/1

- Barry Devlin, IBM Consultant

> A data warehouse is simply a **single, complete,** and **consistent** store of data obtained from a **variety** of sources and made available to end users in a way they can **understand** and **use** it in a **business context.**

Definition of a Data Warehouse/2

- W. H. Inmon, Building the Data Warehouse

A data warehouse is a

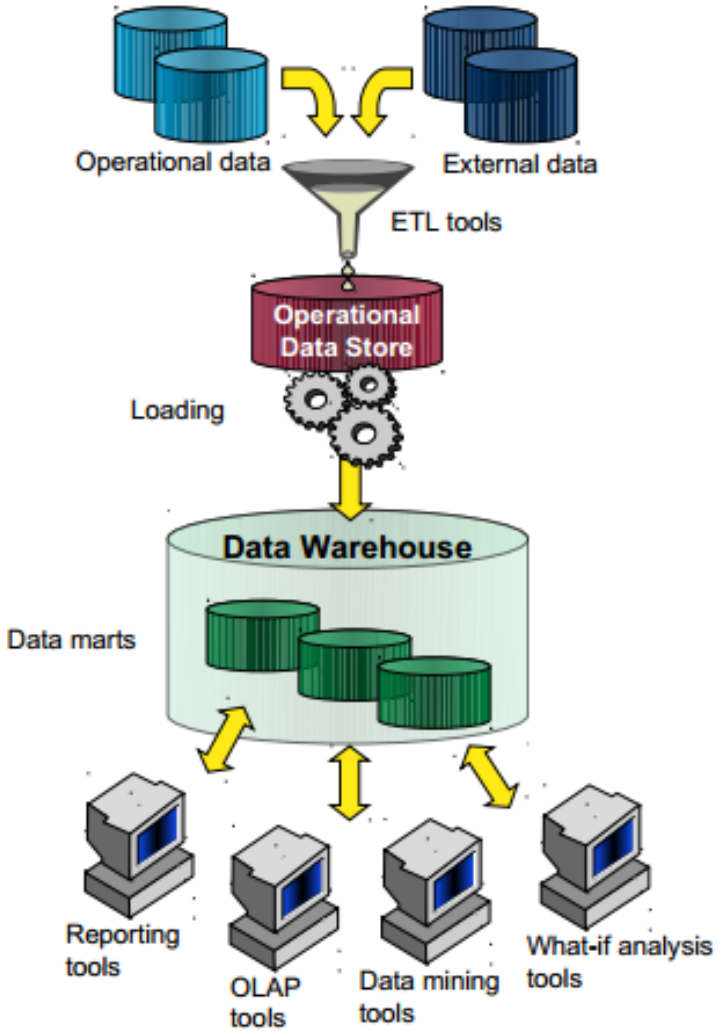
- ✓ **subject-oriented,**
- ✓ **integrated,**
- ✓ **time-varying,**
- ✓ **non-volatile**

collection of data that is used primarily in

- ✓ **organizational decision making.**

DW Architecture

- Basic elements of a Data Warehouse environment



Source layer

Data staging

Reconciled layer

Data warehouse layer

Analysis

EXTRACTION, TRANSFORMATION, AND LOADING:

ETL processes extract data from sources, transform and clean them, and finally load them in the ODS and in the data warehouse

OPERATIONAL DATA STORE:

Operational data obtained after integrating and cleansing source data. As a result, those data are integrated, consistent, appropriate, current, and detailed

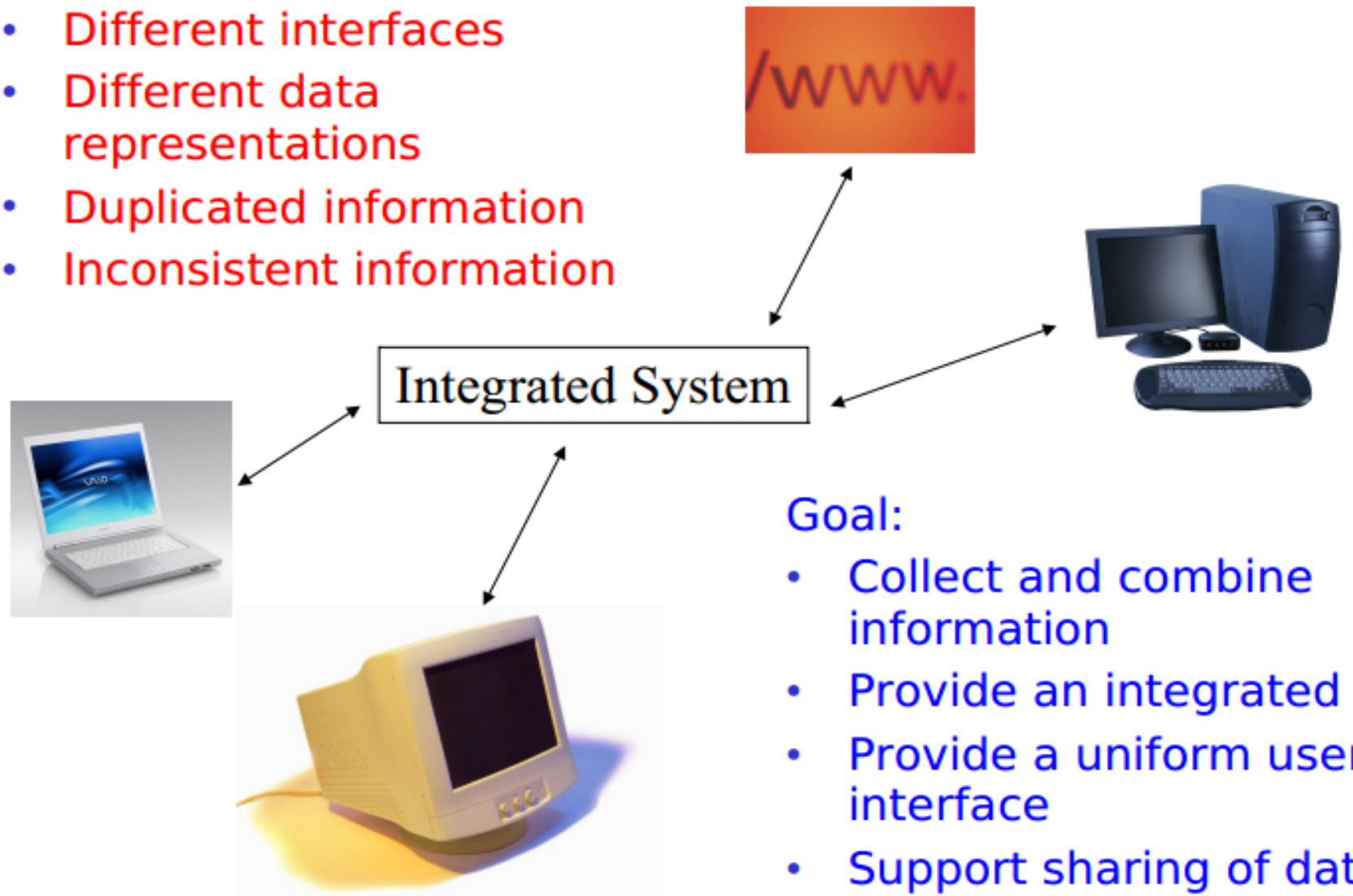
DATA MART:

A subset or an aggregation of the data stored to a primary data warehouse. It includes a set of information pieces relevant to a specific business area, corporate department, or category of users

Data Integration

Problem:

- Different interfaces
- Different data representations
- Duplicated information
- Inconsistent information



Goal:

- Collect and combine information
- Provide an integrated view
- Provide a uniform user interface
- Support sharing of data

Query-Driven Data Integration

- **Data is integrated on demand (lazy)**
- **PROS**
  - Access to most up-to-date data (all source data directly available)
  - No duplication of data
- **CONS**
  - Delay in query processing
    - Slow (or currently unavailable) information sources
    - Complex filtering and integration
  - Inefficient and expensive for frequent queries
  - Competes with local processing at sources
  - Data loss at the sources (e.g., historical data) cannot be recovered
- **Has not caught on in industry**

Warehouse-Driven Data Integration

- **Data is integrated in advance (eager)**
- **Data is stored in DW for querying and analysis**
- **PROS**
  - High query performance
  - Does not interfere with local processing at sources
    - Assumes that data warehouse update is possible during downtime of local processing
    - Complex queries are run at the data warehouse
    - OLTP queries are run at the source systems
- **CONS**
  - Duplication of data
  - The most current source data is not available
- **Has caught on in industry**

OLTP versus OLAP/1

- **On-Line Transaction Processing (OLTP)**
  - Many "small" queries on a small number of tuples from many tables that need to be joined
  - Frequent updates
  - The system is always available for both updates and reads
  - Smaller data volume (few historical data)
  - Complex data model (normalized)

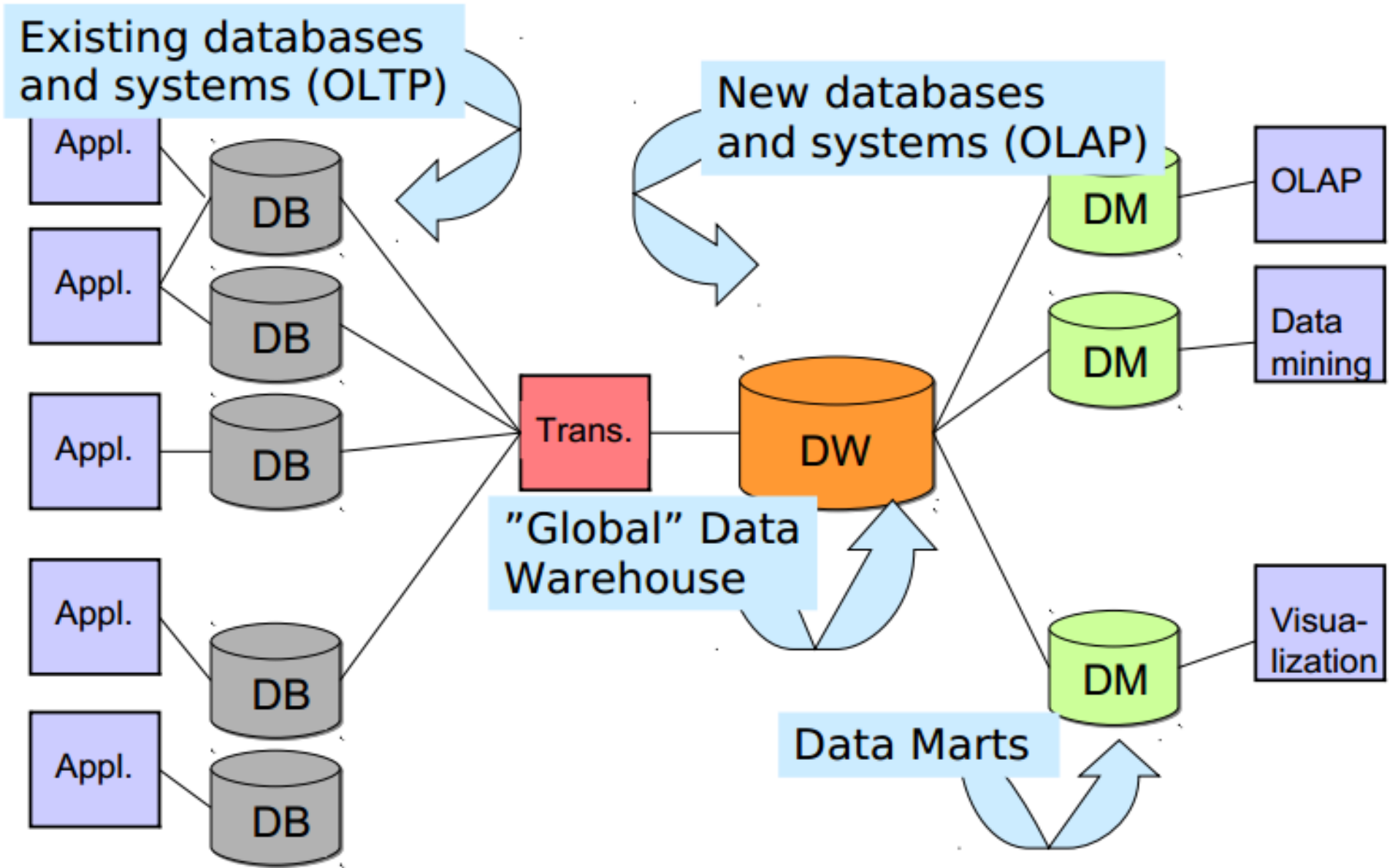
- **On-Line Analytical Processing (OLAP)**
  - Fewer, but "bigger" queries that typically need to scan a huge amount of records and doing some aggregation
  - Frequent reads, in-frequent updates(daily, weekly)
  - 2-phase operation: either reading or updating
  - Larger data volumes (collection of historical data)
  - Simple data model (multidimensional/de-normalized)

OLTP versus OLAP/2

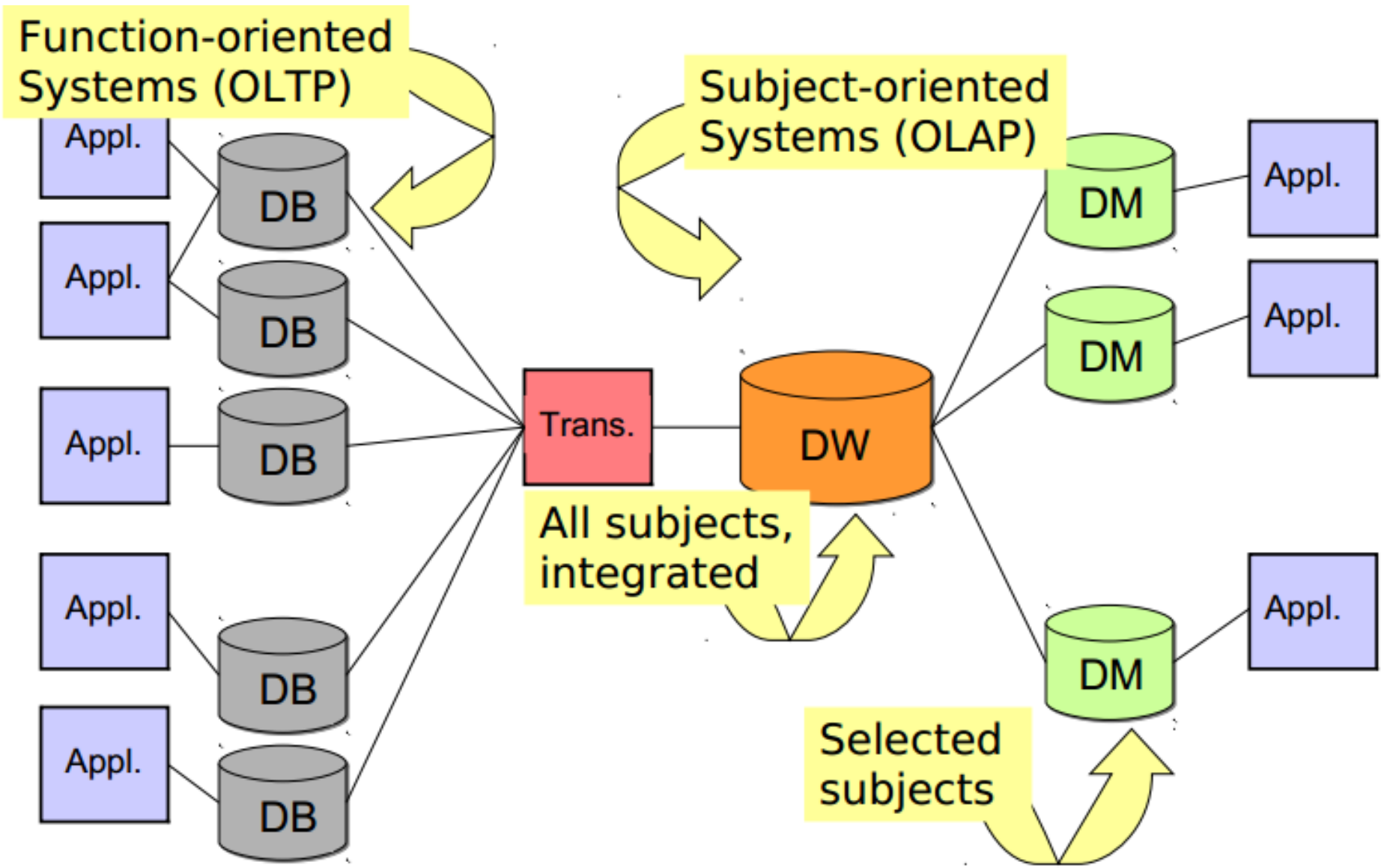
- A mix of analytical queries (OLAP) with transactional routine queries (OLTP) inevitably slows down the system, and this does not meet the needs of users of both types of queries.

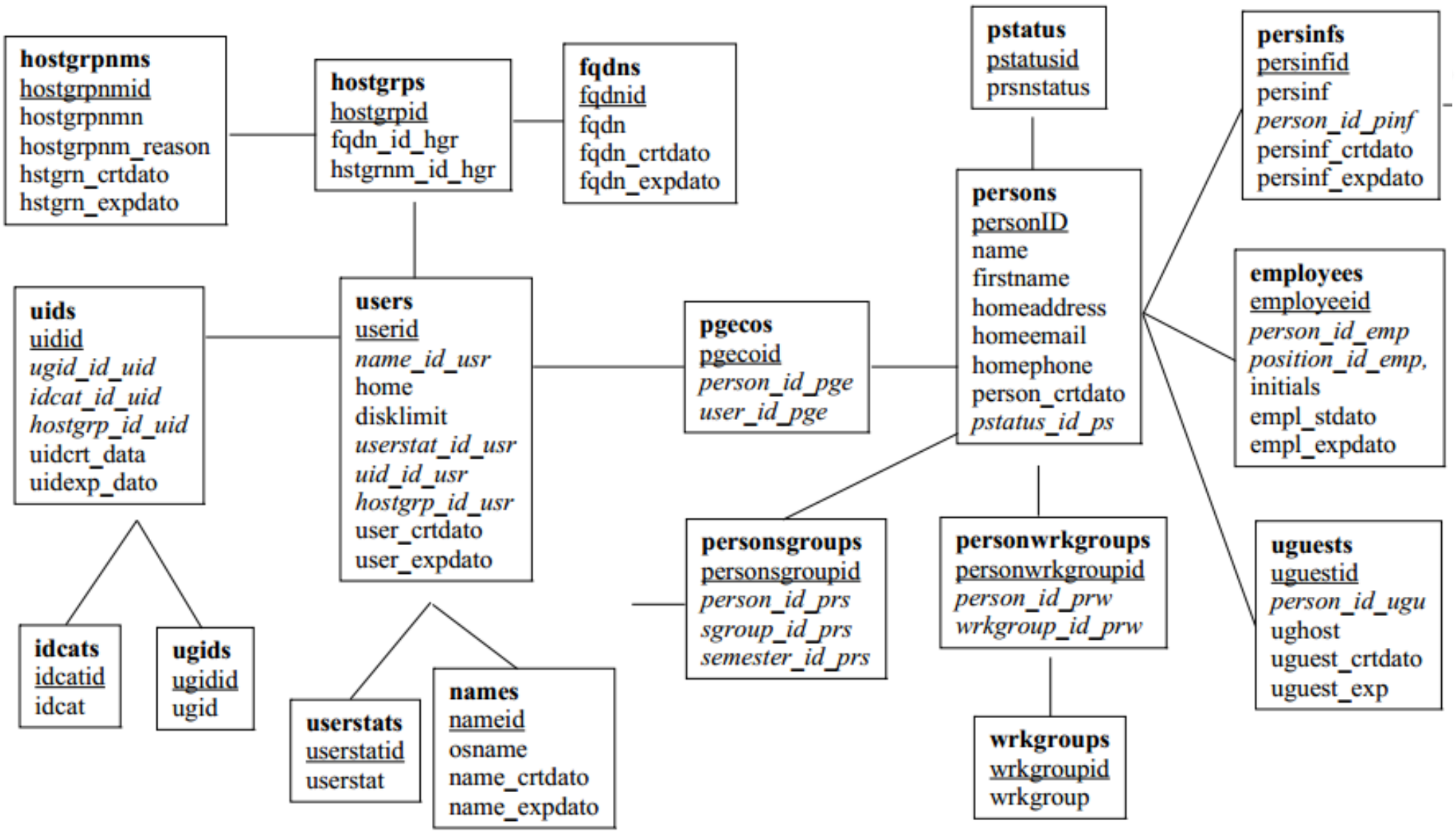
- Separate ***OLAP*** from ***OLTP*** by creating a new repository that integrates data from various sources and then makes data available for analysis and evaluation aimed at decision-making processes

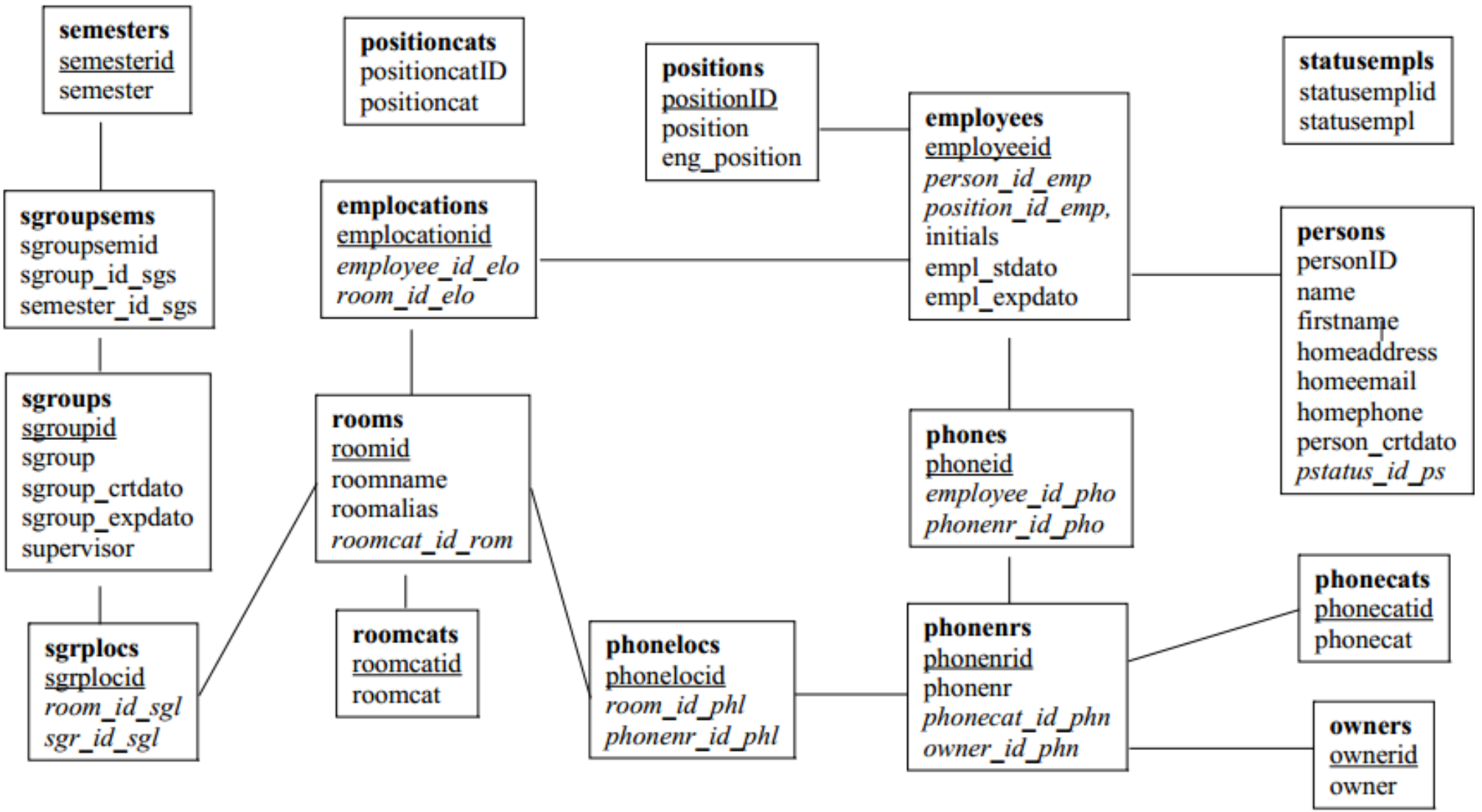
OLTP versus OLAP/3



OLTP versus OLAP/4




hostgrpnms
hostgrpnmid
hostgrpnmn
hostgrpnm_reason
hstgrn_crtdato
hstgrn_expdato
hostgrps
hostgrpid
fqdn_id_hgr
hstgrnm_id_hgr
fqdns
fqdnid
fqdn
fqdn_crtdato
fqdn_expdato
pstatus
pstatusid
prsnstatus
persinfs
persinfid
persinf
person_id_pinf
persinf_crtdato
persinf_expdato
uids
uidid
ugid_id_uid
idcat_id_uid
hostgrp_id_uid
uidcrt_data
uidexp_dato
users
userid
name_id_usr
home
disklimit
userstat_id_usr
uid_id_usr
hostgrp_id_usr
user_crtdato
user_expdato
pgecos
pgecoid
person_id_pge
user_id_pge
persons
personID
name
firstname
homeaddress
homeemail
homephone
person_crtdato
pstatus_id_ps
employees
employeeid
person_id_emp
position_id_emp,
initials
empl_stdato
empl_expdato
idcats
idcatid
idcat
ugids
ugidid
ugid
userstats
userstatid
userstat
names
nameid
osname
name_crtdato
name_expdato
personsgroups
personsgroupid
person_id_prs
sgroup_id_prs
semester_id_prs
personwrkgroups
personwrkgroupid
person_id_prw
wrkgroup_id_prw
uguests
uguestid
person_id_ugu
ughost
uguest_crtdato
uguest_exp
wrkgroups
wrkgroupid
wrkgroup


semesters
semesterid
semester
positioncats
positioncatID
positioncat
positions
positionID
position
eng_position
statusempls
statusemplid
statusempl
sgroupsems
sgroupsemid
sgroup_id_sgs
semester_id_sgs
emplocations
emplocationid
employee_id_elo
room_id_elo
employees
employeeid
person_id_emp
position_id_emp,
initials
empl_stdato
empl_expdato
persons
personID
name
firstname
homeaddress
homeemail
homephone
person_crtdato
pstatus_id_ps
sgroups
sgroupid
sgroup
sgroup_crtdato
sgroup_expdato
supervisor
rooms
roomid
roomname
roomalias
roomcat_id_rom
phones
phoneid
employee_id_pho
phonenr_id_pho
phonecats
phonecatid
phonecat
sgrplocs
sgrplocid
room_id_sgl
sgr_id_sgl
roomcats
roomcatid
roomcat
phonelocs
phonelocid
room_id_phl
phonenr_id_phl
phonenrs
phonenrid
phonenr
phonecat_id_phn
owner_id_phn
owners
ownerid
owner

OLTP Example: OncoNet

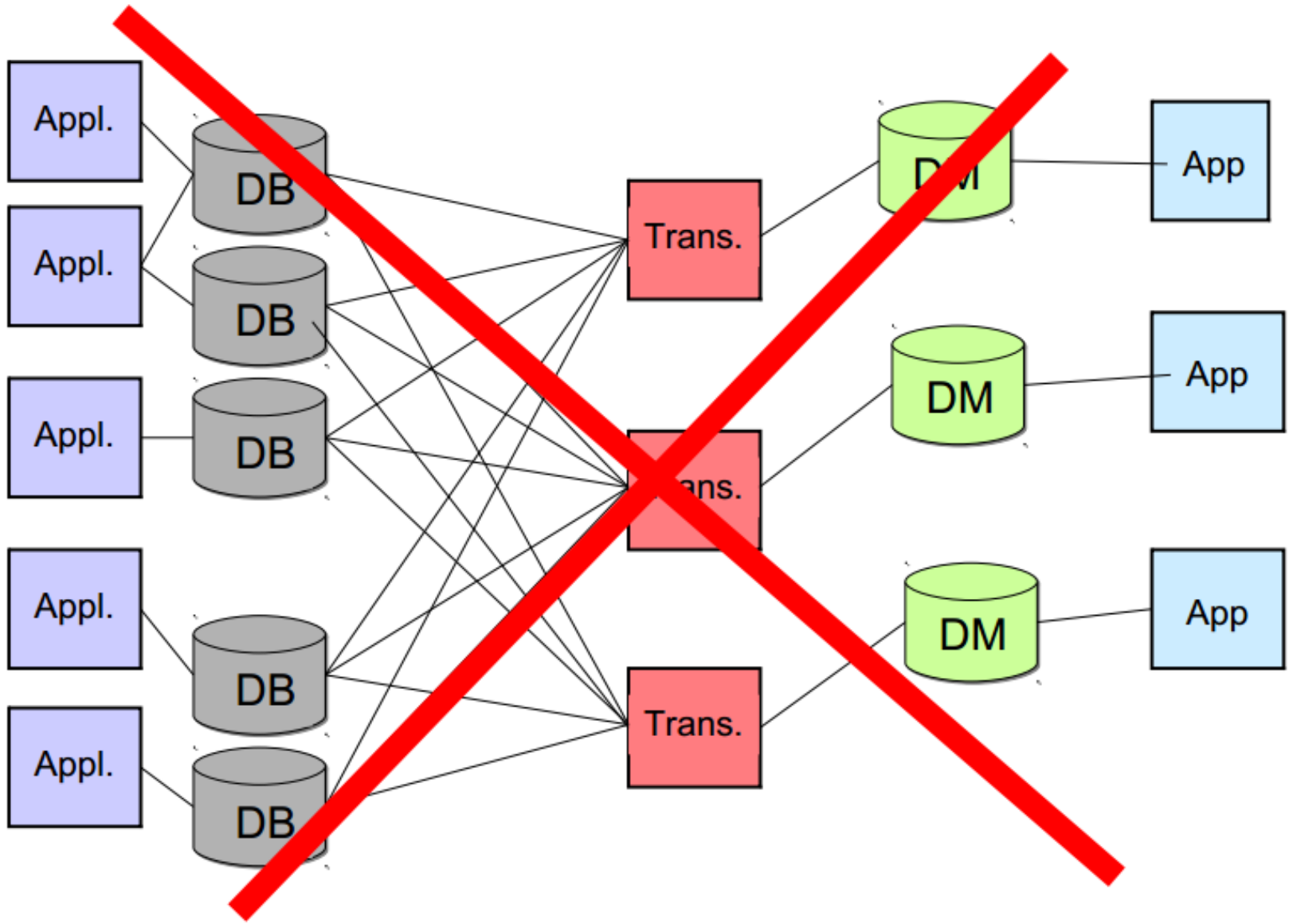
- OncoNet is a system for the management of patients undergoing a cancer therapy
  - > 200 tables
- Well-suited for daily management of patients
- **But**: statistical analysis are expensive
  - takes up to 12 hours
  - tables are locked for that time
  - run queries over weekend

- A DW approach reduced the runtime of the same queries to a few seconds.

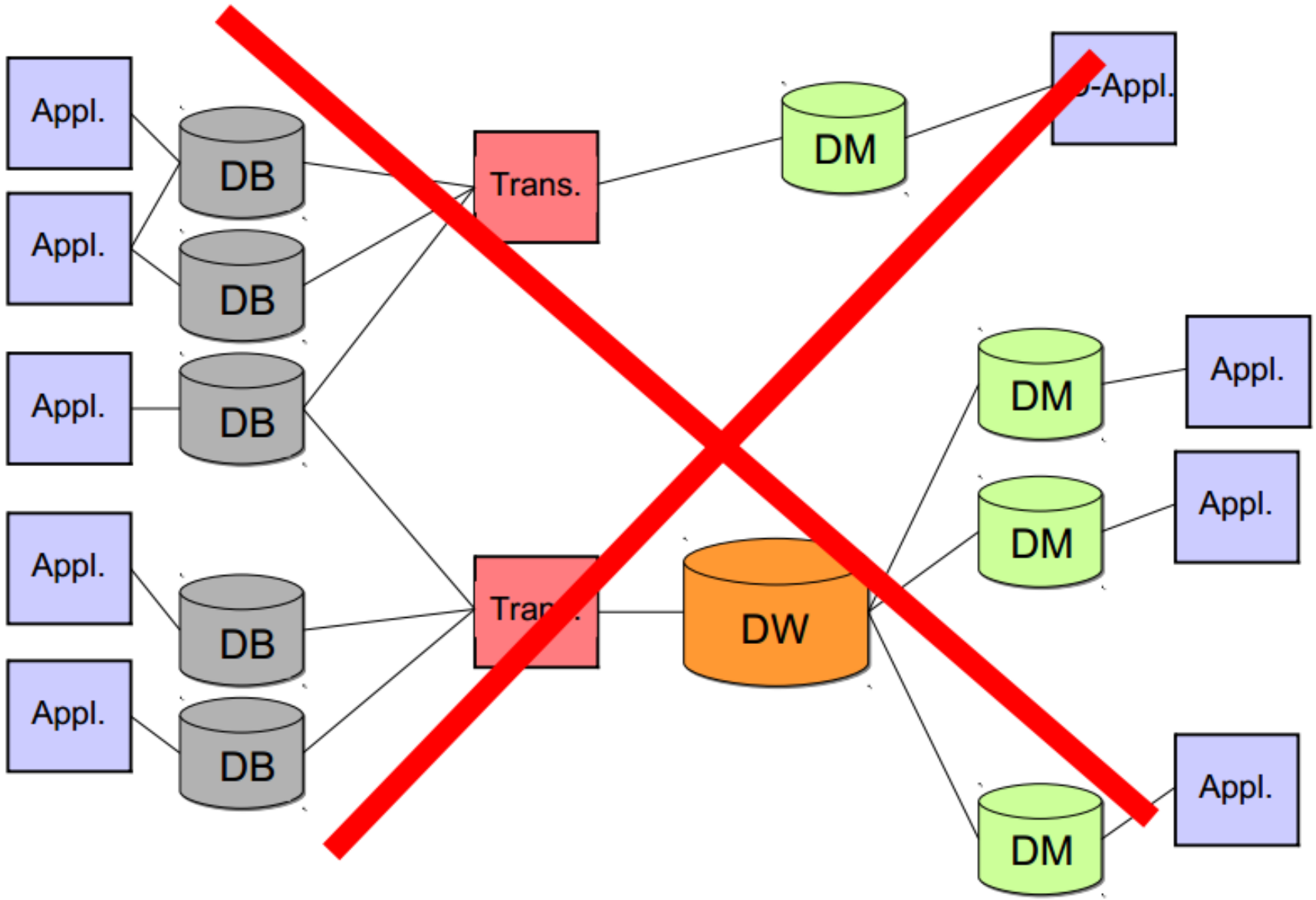
Methodological/Design Framework

- Building a DW is a **very complex task**, which requires an **accurate planning** aimed at devising satisfactory answers to organizational and architectural questions.
- A large number of organizations **lack experience and skills** that are required to meet the challenges involved in DW projects.
- The reports of DW project failures state that a major cause lies in the **absence of a global view** of the design process: in other terms, in the **absence of a design methodology**.
- Methodologies are created by closely studying similar experiences and **minimizing the risks for failure** by basing new approaches on a constructive analysis of the mistakes made previously.

Many Ways not to Do/1



Many Ways not to Do/2



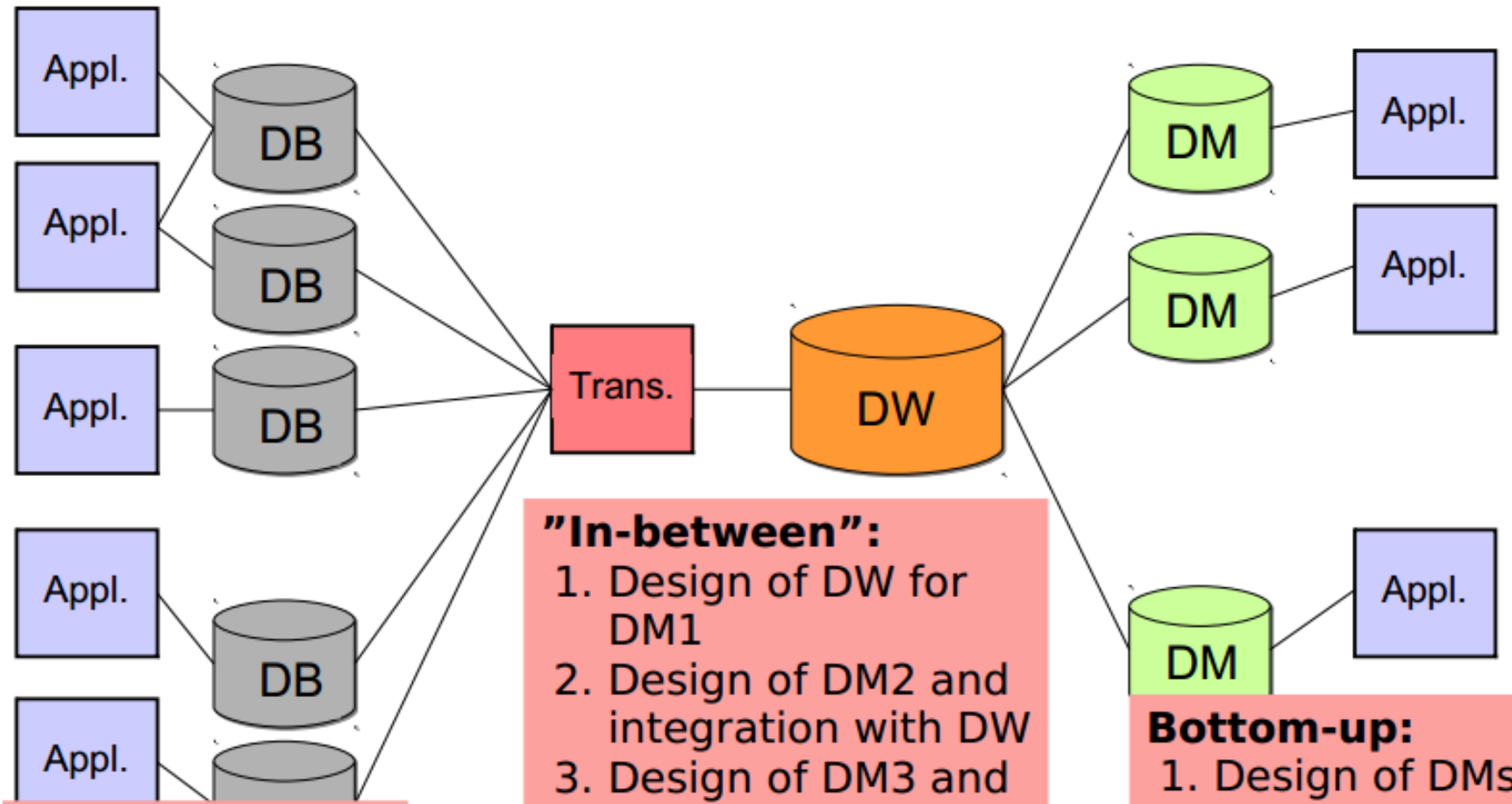
Top-down Approach

- Analyze global business needs, plan how to develop a data warehouse, design it, and implement it as a whole
  - This procedure is promising: it is **based on a global picture** of the goal to achieve, and in principle it ensures consistent, well integrated data warehouses.
  - High-cost estimates with **long-term implementations discourage** company managers from embarking on these kind of projects.
  - Analyzing and integrating all relevant sources at the same time is **a very difficult task**, even because it is not very likely that they are all available and stable at the same time.
  - It is **extremely difficult to forecast** the specific needs of every department involved in a project, which can result in the analysis process coming to a standstill طريق مسدود.
  - Since **no working system is going to be delivered in the short term**, users cannot check for this project to be useful, so they lose trust and interest in it.

Bottom-up Approach

- DWs are incrementally built and several data marts are iteratively created. Each data mart is based on a set of facts that are linked to a specific department and that can be interesting for a user group
  - Leads to concrete **results in a short time**
  - Does **not require huge investments**
  - Enables designers to investigate **one area at a time**
  - Gives managers a **quick feedback** about the actual benefits of the system being built
  - Keeps the interest for the project constantly high
  - May determine a partial vision of the business domain

Top-down vs. Bottom-up Approach



Summary

- BI is well-recognized and is a combination of a number of techniques to support decision making.
- DW is at the core of BI that
  - provides a complete, consistent, subject-oriented and time-varying collection of the data;
  - allows to separte OLTP from OLAP.
- Applications that **use** the DW include OLAP, data mining, visualization
- BI can provide many advantages to an organization
  - Creates added value by transforming data into information
  - Provides comprehensive knowledge about your business
  - A good DW is a prerequisite for BI
  - But, a DW is a **means** rather than a **goal** … it is only a success if it is heavily used
- Following a clear design methodology is important.