- **Lesson 2: How collecting data? (the process of data collection/management)**
  - Type of data and type of informative system, Batini - Scannapieco (pagg 7 - 19)
  - Data flows, from on paper docs to electronic charts
  - Real-time database drawn from an electronic health record for a thoracic surgery unit: high-quality clinical data saving time and human resources, Michele Salati

Data, in general, describe real world objects in a format that can be stored, retrieved, and processed by a software procedure.

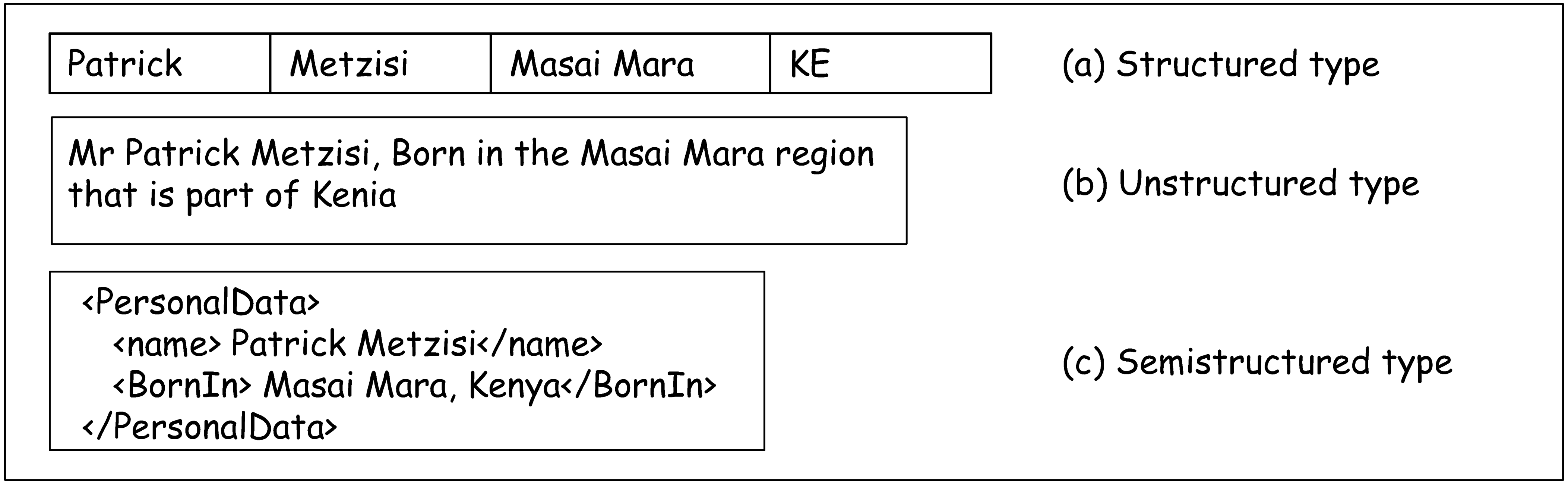
- **Lesson 2: How collecting data? (the process of data collection/management)**

TYPE OF DATA

(1) *Structured data*, is aggregations or generalizations of items described by elementary attributes defined within a domain. Domains represent the range of values that can be assigned to attributes and usually correspond to elementary data types of programming languages, such as numeric values or text strings. Relational tables and statistical data represent the most common type of structured data.

(2) *Semistructured data*, is data that have a structure which has some degree of flexibility. Semistructured data are also referred to as schemaless or self-describing [Abiteboul et al. 2000; Buneman 1997; Calvanese et al. 1999]. XML is the markup language commonly used to represent semistructured data. Some common characteristics are: (1) data can contain fields not known at design time; for instance, an XML file does not have an associated XML schema file; (2) the same kind of data may be represented in multiple ways; for example, a date might be represented by one field or by multiple fields, even within a single data set; and (3) among the fields known at design time, many fields will not have values.

(3) *Unstructured data*, is a generic sequence of symbols, typically coded in natural language. Typical examples of unstructured data are a questionnaire containing free text answering open questions or the body of an e-mail.



Patrick	Metzisi	Masai Mara	KE

Mr Patrick Metzisi, Born in the Masai Mara region that is part of Kenia

```
<PersonalData>
  <name> Patrick Metzisi</name>
  <BornIn> Masai Mara, Kenya</BornIn>
</PersonalData>
```

Fig. 2. Different representations of the same real-world object.

The same quality dimension will have different metrics according to the type of data. For instance, syntactic accuracy is measured as described in Lesson 3 in the case of structured data. With semistructured data, the distance function should consider a global distance related to the shape of the XML tree in addition to the local distance of fields.

The large majority of research contributions in the data quality literature focuses on structured and semistructured data

OTHER CLASSIFICATIONS OF DATA

(1) *Elementary data*, represents atomic informations of the real word (i.e. age, sex).

(2) *Aggregated data*, is obtained by a collection of elementary data submitted to an aggregation function (i.e. mean hemoglobin level of patients submitted to coronary stent).

This classification could influence the rigor for measuring parameters (metrics) of data quality

---

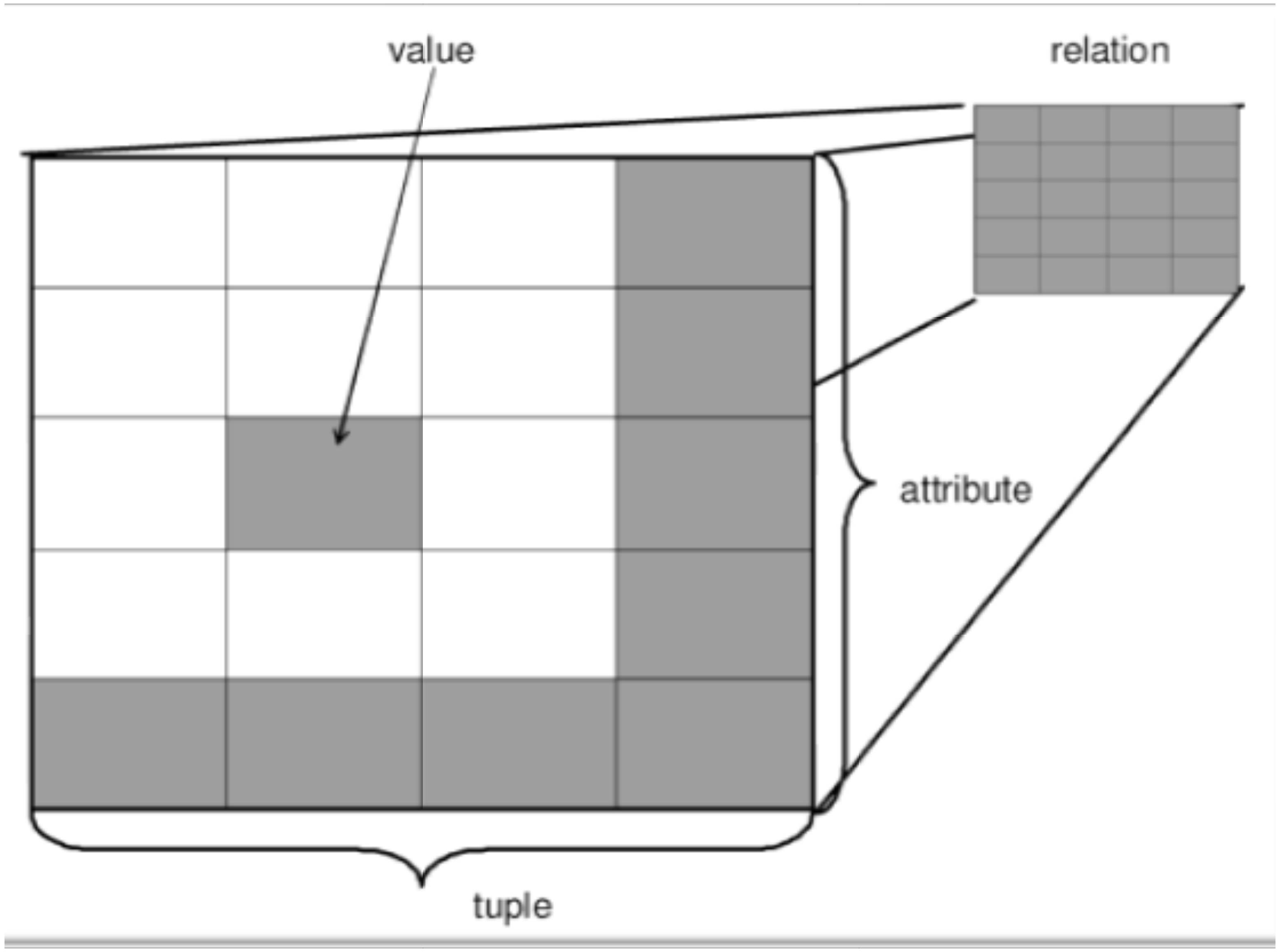
Taking into account a temporal dimension, we could classify:

(1) *Stable data*, its change is extremely unlikely.

(2) *Long term variation data*, its rate of update is low (i.e. follow up data)

(3) *Frequently changeable data*, its rate of update is high (i.e. daily blood pressure)

The assessment of data quality increases the level of complexity with the rise of data update


value
relation
attribute
tuple

TYPES OF INFORMATION SYSTEMS

Different IS architectures or, simply, types of information systems are distinguished on the basis of the degree of data, process and management integration supported by a technical system. As the degree of integration of data, process, and management decreases, the data quality assessment and improvement techniques that can be applied become more sophisticated.

At the same time, data quality assessment and improvement is more challenging.

TYPES OF INFORMATION SYSTEMS

(1) In a *monolithic information system*, applications are single-tier and do not provide data access services. Although data are usually stored in a database that can be queried, separate applications do not share data. This can cause data duplication, possibly affecting all quality dimensions.

(2) A *data warehouse (DW)* is a centralized collection of data retrieved from multiple databases. Data warehouses are periodically refreshed with updated data from the original databases by procedures automatically extracting and aligning data. Data are physically integrated, since they are reformatted according to the data warehouse schema, merged, and finally stored, in the data warehouse.

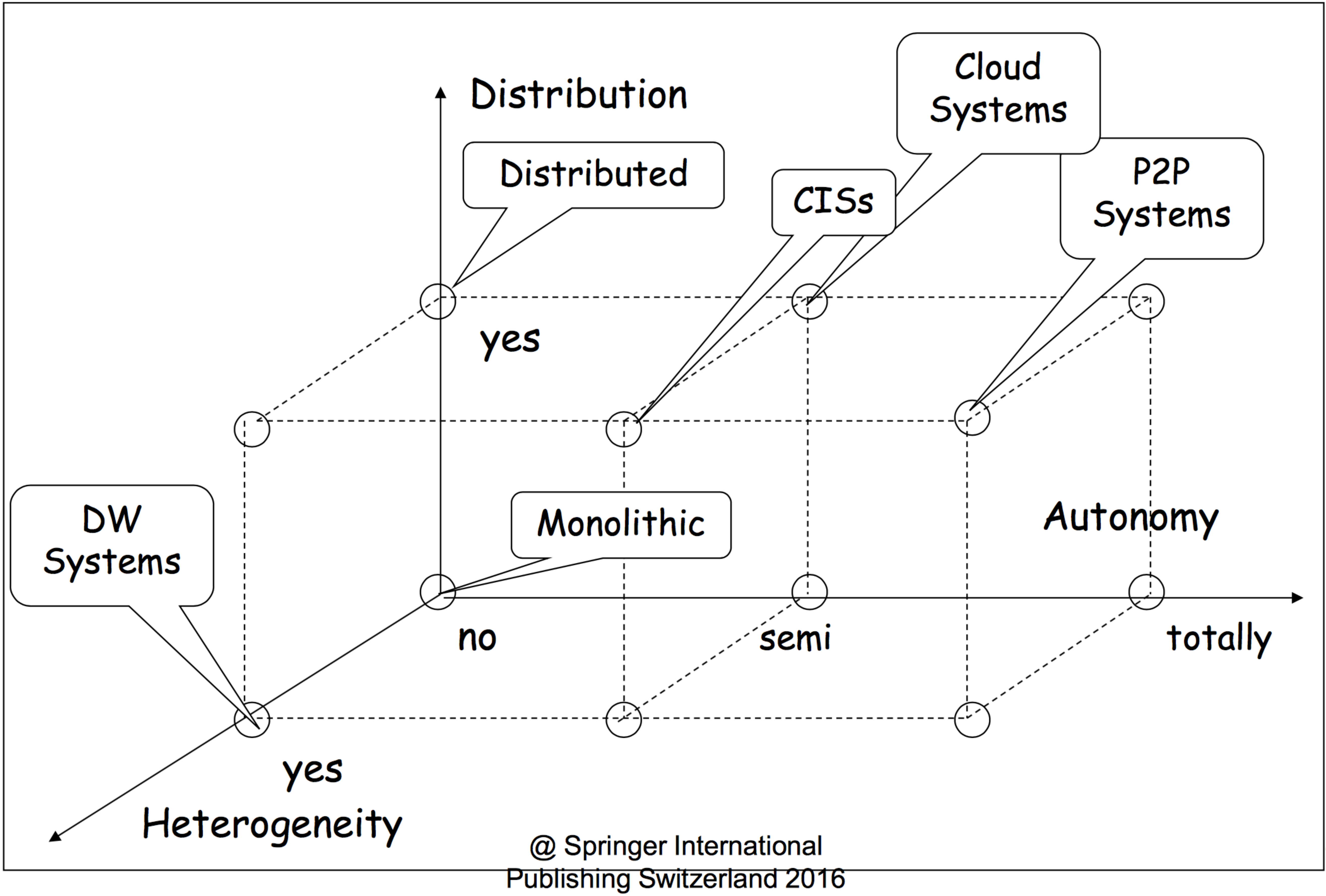
(3) A *distributed information system* is a collection of application modules coordinated by a workflow. Applications are typically divided in tiers, such as presentation, application logic, and data management, and export data access functionalities at different tiers. Data can be stored in different databases, but interoperability is guaranteed by the logical integration of their schemas.

(4) A *cooperative information system* (CIS) can be defined as a large-scale information system that interconnects multiple systems of different and autonomous organizations sharing common objectives [De Michelis et al. 1997]. Cooperation with other information systems requires the ability to exchange information. In CISs, data are not logically integrated, since they are stored in separate databases according to different schemas. However, applications incorporate data transformation and exchange procedures that allow interoperability and cooperation among common processes. In other words, integration is realized at a process level.

(5) In the literature, the term *Web Information System* (WIS) [Isakowitz et al. 1998] is used to indicate any type of information adopting Web technologies. From a technical perspective a WIS is a client/server application. Such systems typically use structured, semi structured, and unstructured data, and are supported by development and management tools based on techniques specific to each type of data.

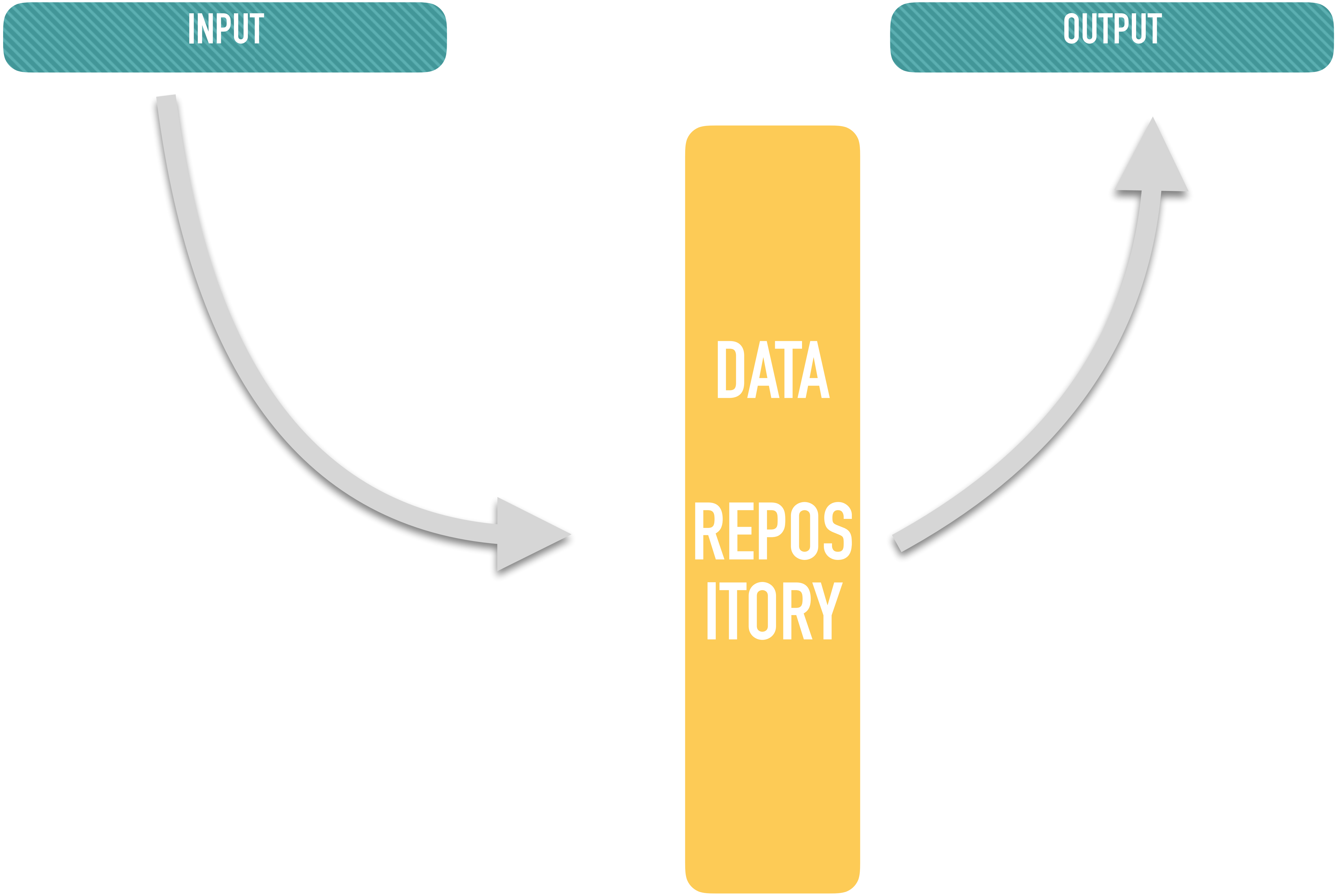
(6) In a *peer-to-peer information system* (P2P), there is no distinction between clients and servers. The system is constituted by a set of identical nodes that share data and application services in order to satisfy given user requirements collectively. P2P systems are characterized by a number of properties: no central coordination, no central database, no peer has a global view of the system, Peers are autonomous and can dynamically connect or disconnect from the system. However, peers typically share common management procedures.

Types of information systems

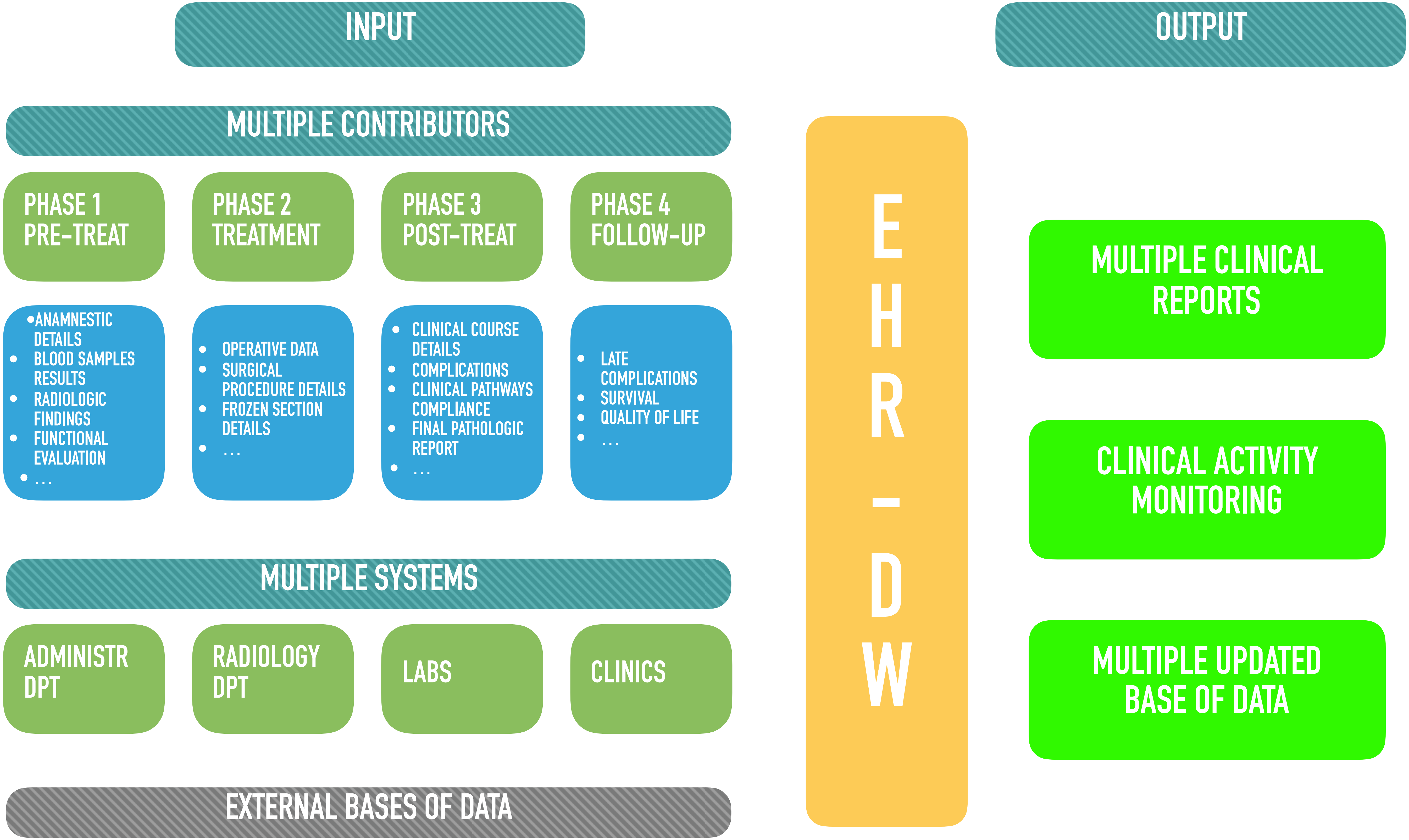


@ Springer International Publishing Switzerland 2016

DATA FLOW



DATA FLOW



INPUT

OUTPUT

MULTIPLE CONTRIBUTORS

PHASE 1
PRE-TREAT

PHASE 2
TREATMENT

PHASE 3
POST-TREAT

PHASE 4
FOLLOW-UP

- ANAMNESTIC DETAILS
- BLOOD SAMPLES RESULTS
- RADIOLOGIC FINDINGS
- FUNCTIONAL EVALUATION
- …

- OPERATIVE DATA
- SURGICAL PROCEDURE DETAILS
- FROZEN SECTION DETAILS
- …

- CLINICAL COURSE DETAILS
- COMPLICATIONS
- CLINICAL PATHWAYS COMPLIANCE
- FINAL PATHOLOGIC REPORT
- …

- LATE COMPLICATIONS
- SURVIVAL
- QUALITY OF LIFE
- …

MULTIPLE SYSTEMS

ADMINISTR
DPT

RADIOLOGY
DPT

LABS

CLINICS

EXTERNAL BASES OF DATA

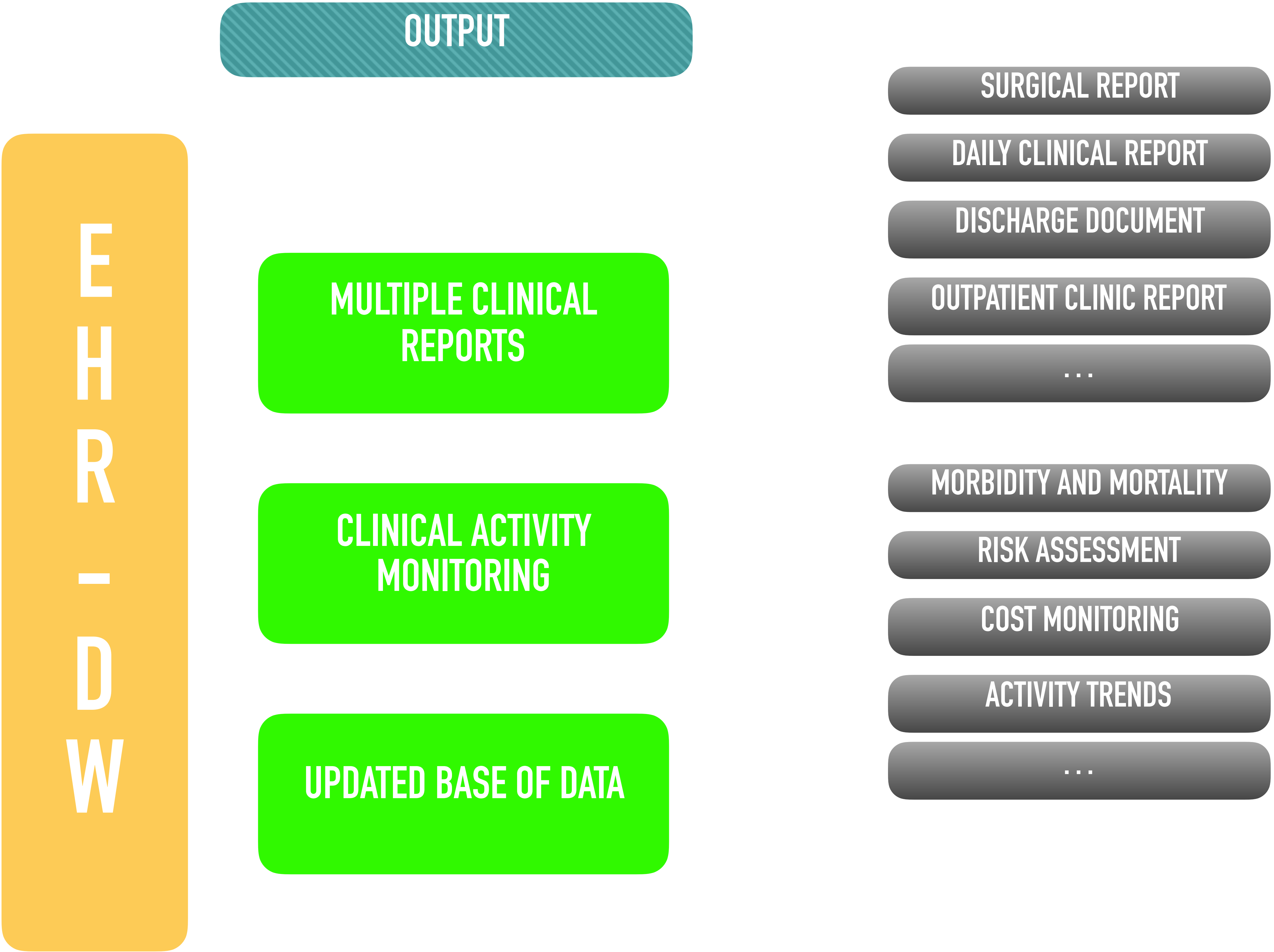
EHR-DW

MULTIPLE CLINICAL REPORTS

CLINICAL ACTIVITY MONITORING

MULTIPLE UPDATED BASE OF DATA

DATA FLOW, IMPACT ON CLINICAL CARE



OUTPUT

EHR - DW

MULTIPLE CLINICAL REPORTS

- SURGICAL REPORT
- DAILY CLINICAL REPORT
- DISCHARGE DOCUMENT
- OUTPATIENT CLINIC REPORT
- . . .

CLINICAL ACTIVITY MONITORING

- MORBIDITY AND MORTALITY
- RISK ASSESSMENT
- COST MONITORING
- ACTIVITY TRENDS
- . . .

UPDATED BASE OF DATA

DATA FLOW, IMPACT ON CLINICAL CARE

OUTPUT

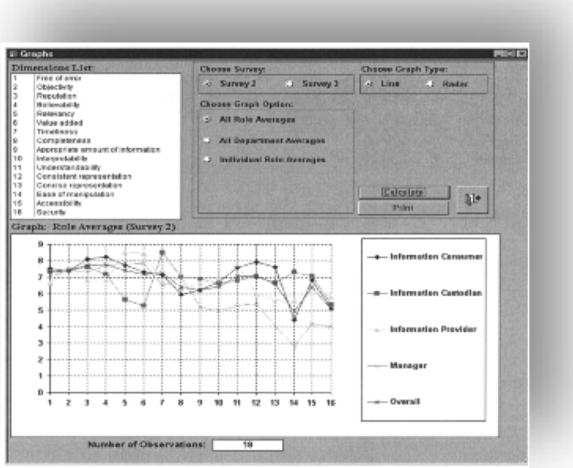
EHR-DW

MULTIPL
R

CLINICAL
MONITORI

UPDATED BASE OF DATA

AUTOMATICALLY CALCULATED
PERIODICALLY CHECKED
DATA QUALITY CONTROLS



Graphs

Dimensions List:

1 Free of error
2 Objectivity
3 Reputation
4 Believability
5 Relevancy
6 Value added
7 Timeliness
8 Completeness
9 Appropriate amount of information
10 Interpretability
11 Understandability
12 Consistent representation
13 Concise representation
14 Ease of manipulation
15 Accessibility
16 Security

Choose Survey: ● Survey 2 ○ Survey 3

Choose Graph Type: ● Line ○ Radar

Choose Graph Option: ● All Role Averages ○ All Department Averages ○ Individual Role Averages

Graph: Role Averages (Survey 2)

Number of Observations: 18

DATA QUALITY ASSESSMENT (twice per year)

QUALITY IMPROVEMENT STRATEGIES

SCIENTIFIC ACTIVITY WITH THE TQM APPROACH

DATA FLOW, IMPACT ON CLINICAL CARE

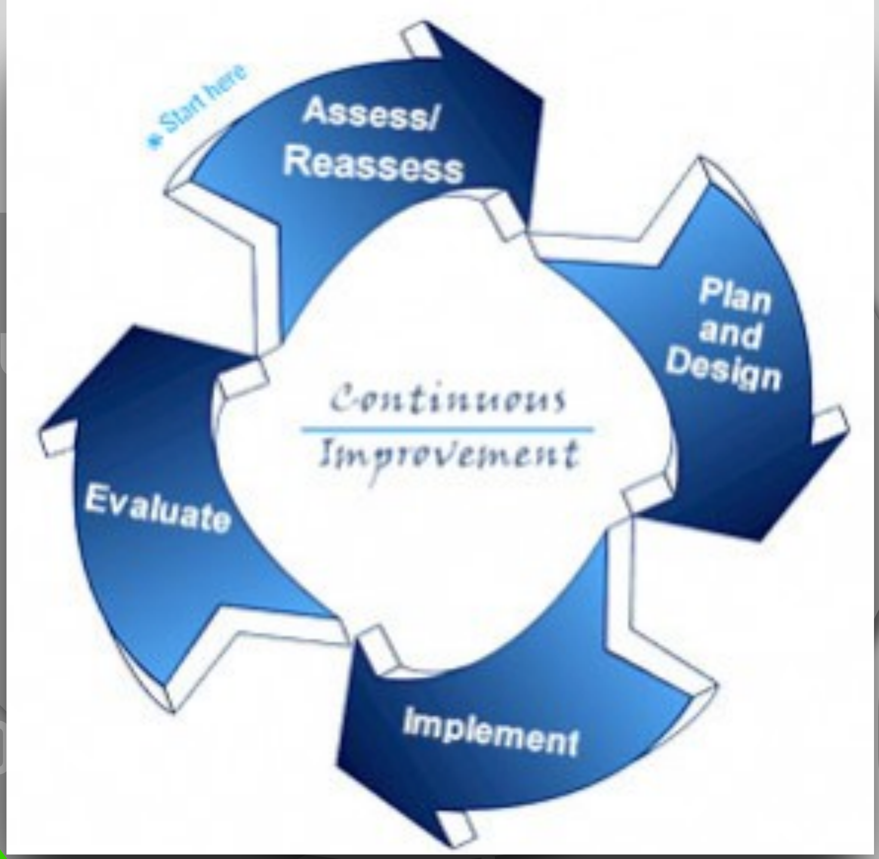
CURRENT PRACTICE ANALYSIS

TPUT

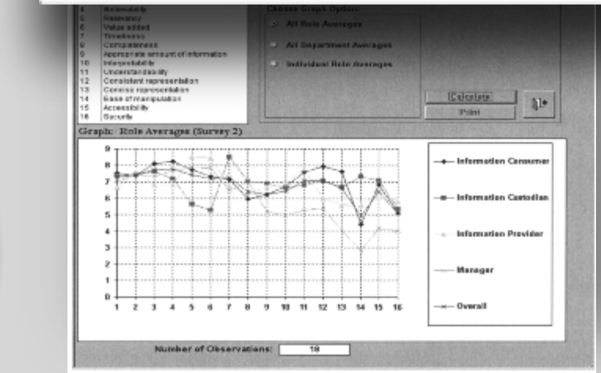
E
H
D
W

MULTIPL
R A

CAL
MONITO



NEW MANAGEMENT STRATEGY DESIGN



4	Retrievability
5	Relevancy
6	Value added
7	Timeliness
8	Completeness
9	Appropriate amount of information
10	Interpretability
11	Understandability
12	Consistent representation
13	Concise representation
14	Ease of manipulation
15	Accessibility
16	Security

Choose Graph Option:

○ All Role Averages

○ All Department Averages

○ Individual Role Averages

Graph: Role Averages (Survey 2)

Number of Observations: 18

IMPACT ON QUALITY INDICTORS EVALUATION

DATA QUALITY ASSESSMENT (twice per year)

EMENT STRATEGIES

UPDATED BASE OF DATA

DAILY PRACTICE IMPLEMENTATION

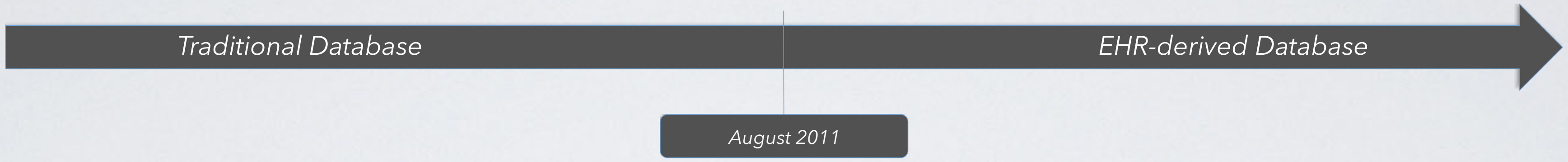
SCIENTIFIC ACTIVITY WITH THE TQM APPROACH

REAL-TIME DATABASE DRAWN FROM ELECTRONIC HEALTH RECORD FOR A THORACIC SURGERY UNIT: HIGH QUALITY CLINICAL DATA SAVING TIME AND HUMAN RESOURCES

Michele Salati, Cecilia Pompili, Majed Refai, Francesco Xiumè, Armando Sabbatini, Alessandro Brunelli - Ospedali Riuniti, Ancona, Italy

OBJECTIVES: In times of cost restraints, clinical data represent the centerpiece of nearly every initiative designed to bend the health care quality and cost curves and to promote research.

The aim of the present study was to verify if the implementation of an Electronic-Health-Record (EHR) in our thoracic surgery unit allowed to create a high quality clinical database (eDB) saving time and costs, in comparison to the traditional database (tDB).



Traditional Database

EHR-derived Database

August 2011

t-DB

- ✓ DATA RETRIVED FROM CLINICAL ON PAPER DOCS
- ✓ A DATA MANAGER IMPUTED SELECTED DATA INTO THE DB

DATABASE COMPARISON FOR:
COMPLETENESS
ACCURACY

e-DB

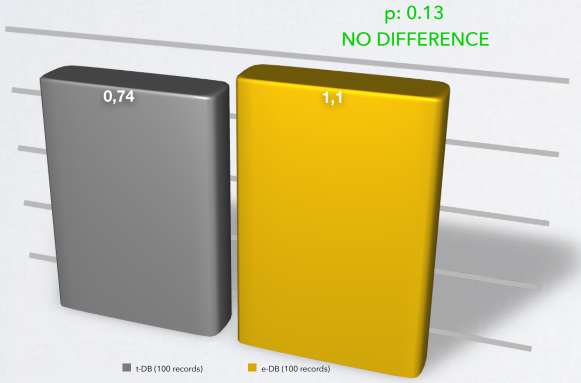
- ✓ MULTIPLE PHYSICIANS COMPILE EHR AS CLINICAL ROUTINARY PRACTICE
- ✓ EHR AUTOMATICALLY GENERATES THE DB (NO DATA-MANAGER ENTRY)

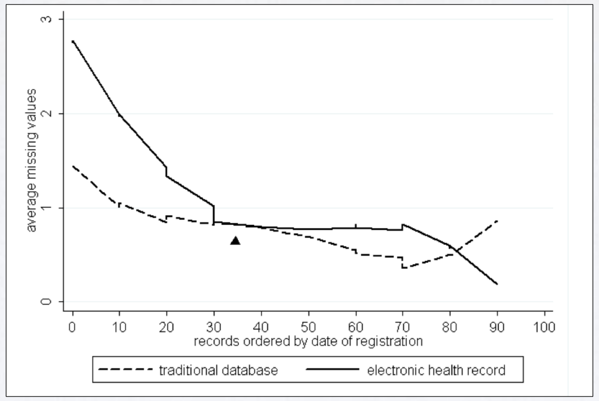
REVISION LAST 100 RECORDS

REVISION FIRST 100 RECORDS

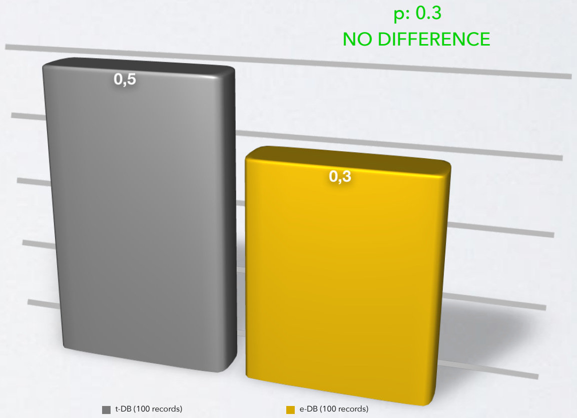
MISSING VALUE RATE									**INACCURATE VALUE RATE**		
t-DB (100 records)	*0.74 ± 1.9*	*p: 0.13*	**t-DB (100 records)**	*0.74 ± 1.9*	*p: 0.03*	**t-DB (100 records)**	*0.74 ± 1.9*	*p: 0.6*	**t-DB (100 records)**	*0.5 ± 1*	*p: 0.3*
e-DB (100 records)	*1.1 ± 2*		**e-DB (first 35 records)**	*1.9 ± 2.7*		**e-DB (last 65 records)**	*1.9 ± 2.7*		**e-DB (100 records)**	*0.3 ± 0.7*	

MISSING VALUE RATE





INACCURATE VALUE RATE



Database	Phases	Mean time required (min)	Total single record (min)	Total entire database (100 pts) (hours)
tDB	on paper clinical chart compilation	27	49	81.7
	clinical chart completion and closure	8		
	data entering in the electronic database	14		
eDB	HER compilation during preoperative evaluation	25	40	66.7
	entering surgical data	3		
	entering outcomes	4		
	entering staging and final check	8		

CONCLUSIONS:

1. EHR allowed to obtain a base of clinical data with an high quality level (completeness and accuracy rates above 99%), comparable to that of a traditional database
2. At the same time, the possibility of automatically generate a real-time database reduces the time and the human resources costs involved in data collection processes.