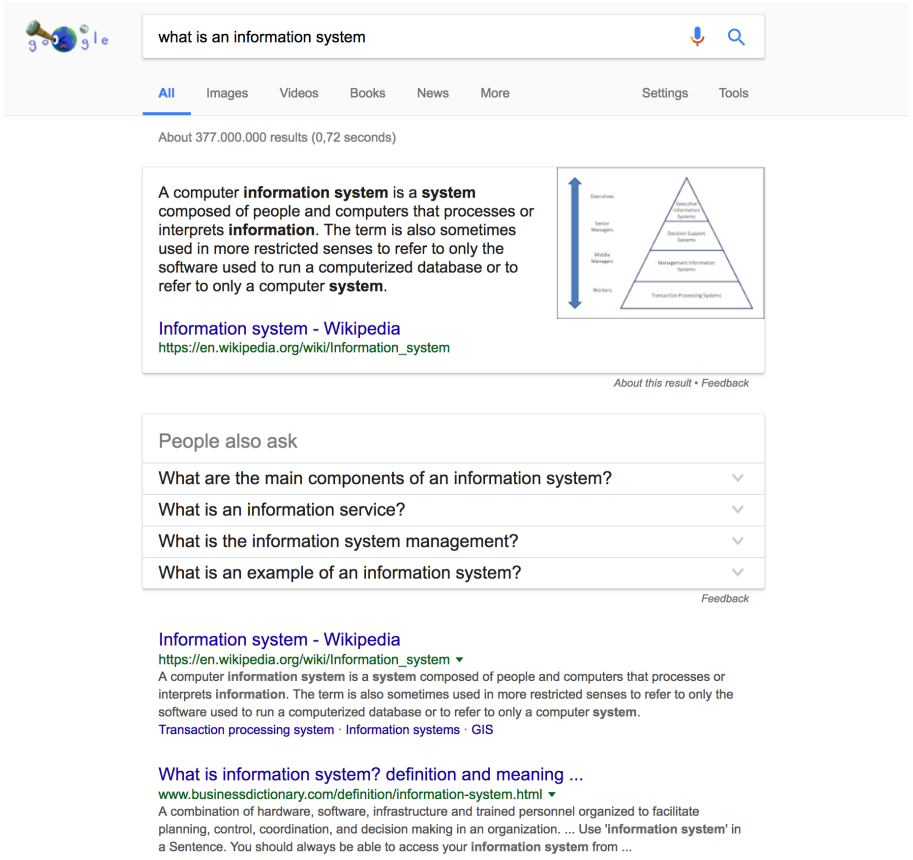
Introduction:
Distributed Information Systems
An Overview

Overview

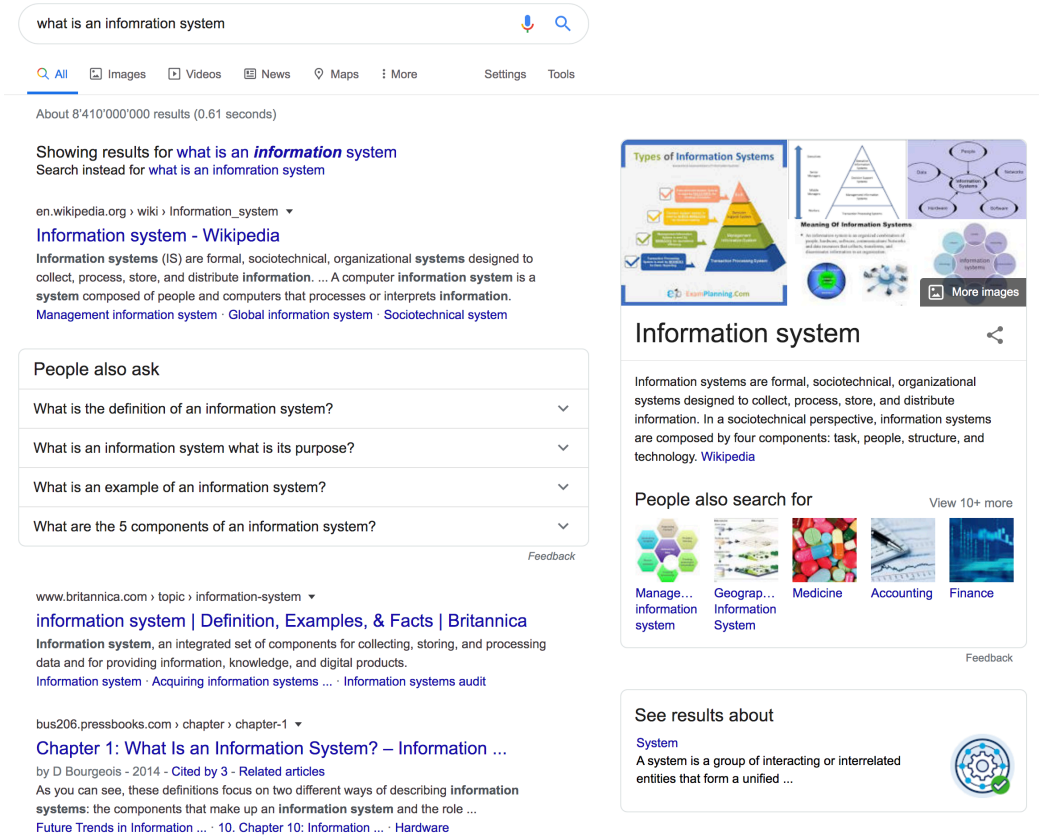
1. What is an Information System?
2. Data Management
3. Information Management
4. Distributed Information Management

1. WHAT IS AN INFORMATION SYSTEM?

Ask Google & Wikipedia



2018



2020

Information system

From Wikipedia, the free encyclopedia

> **This article has multiple issues.** Please help **improve it** or discuss these issues on the **talk page**. *(Learn how and when to remove these template messages)* [hide]
>
> - The **lead section of this article may need to be rewritten.** *(April 2017)*
> - This article **contains weasel words: vague phrasing that often accompanies biased or unverifiable information**. *(April 2017)*

Information systems (**IS**) are formal, sociotechnical, organizational systems designed to collect, process, store, and distribute information.[1] In a sociotechnical perspective, information systems are composed by four components: task, people, structure (or roles), and technology.[2]

A **computer information system** is a system composed of people and computers that processes or interprets information.[3][4][5][6] The term is also sometimes used in more restricted senses to refer to only the software used to run a computerized database or to refer to only a computer system.

Information Systems is an academic study of systems with a specific reference to information and the complementary networks of hardware and software that people and organizations use to collect, filter, process, create and also distribute data. An emphasis is placed on an information system having a definitive boundary, users, processors, storage, inputs, outputs and the aforementioned communication networks.[7]

Any specific information system aims to support operations, management and decision-making.[8][9] An information system is the information and communication technology (ICT) that an organization uses, and also the way in which people interact with this technology in support of business processes.[10]

Some authors make a clear distinction between information systems, computer systems, and business processes. Information systems typically include an ICT component but are not purely concerned with ICT, focusing instead on the end use of information technology. Information systems are also different from business processes. Information systems help to control the performance of business processes.[11]

Alter[12][13] argues for advantages of viewing an information system as a special type of work system. A work system is a system in which humans or machines perform processes and activities using resources to produce specific products or services for customers. An information system is a work system whose activities are devoted to capturing, transmitting, storing, retrieving, manipulating and displaying information.[14]

As such, information systems inter-relate with data systems on the one hand and activity systems on the other. An information system is a form of communication system in which data represent and are processed as a form of social memory. An information system can also be considered a semi-formal language which supports human decision making and action.

Information systems are the primary focus of study for organizational informatics.[15]

Information Systems

Name	Literals	Categories	Similarity	DF	Entropy	Hashtag?	Harvesting?	Wiki/Wn
big data	de: "big data" en: "big data", "bigdata"	Information Systems	0.934	3437		Yes	No	
information systems	en: "information systems"	Information Systems	0.934	1303		Yes	No	
Business Intelligence	de: "business intelligence" en: "Business Intelligence",	Information Systems; Academic	0.901	101		Yes	No	
knowledge management	en: "knowledge management"	Information Systems	0.897	435		Yes	No	
information security	en: "information security"	Information Systems	0.878	518		Yes	No	
emerging technology	en: "emerging technologies", "emerging	Information Systems	0.871	564		No	No	???
data analytics	en: "data analytics"	Information Systems	0.869	91		Yes	No	
e-business	en: "e-business"	Information Systems	0.866	159		No	No	???
change management	en: "change management"	Information Systems	0.866	256		No	No	???
cloud computing	de: "cloud computing" en: "cloud computing"	Information Systems	0.863	390		Yes	No	
data mining	en: "data mining"	Information Systems	0.863	612		No	No	???
supply chain	de: "lieferkette", "supply-chain-management"	Information Systems	0.863	951		Yes	No	
data management	en: "data management"	Information Systems	0.862	365		Yes	No	
Internet of Things	de: "internet der dinge", "iot", "yacht"	Information Systems	0.862	1151		Yes	No	
data science	de: "data science" en: "data science", "data	Information Systems	0.861	741		Yes	No	
data integration	en: "data integration"	Information Systems	0.854	149		No	No	???
analytics	de: "analytisch", "analytisches	Information Systems	0.853	2536		Yes	No	
decision support systems	en: "decision support systems"	Information Systems	0.852	113		No	No	???
information management	en: "information management"	Information Systems	0.850	519		No	No	???

Information Systems

Examples, e.g., at EPFL

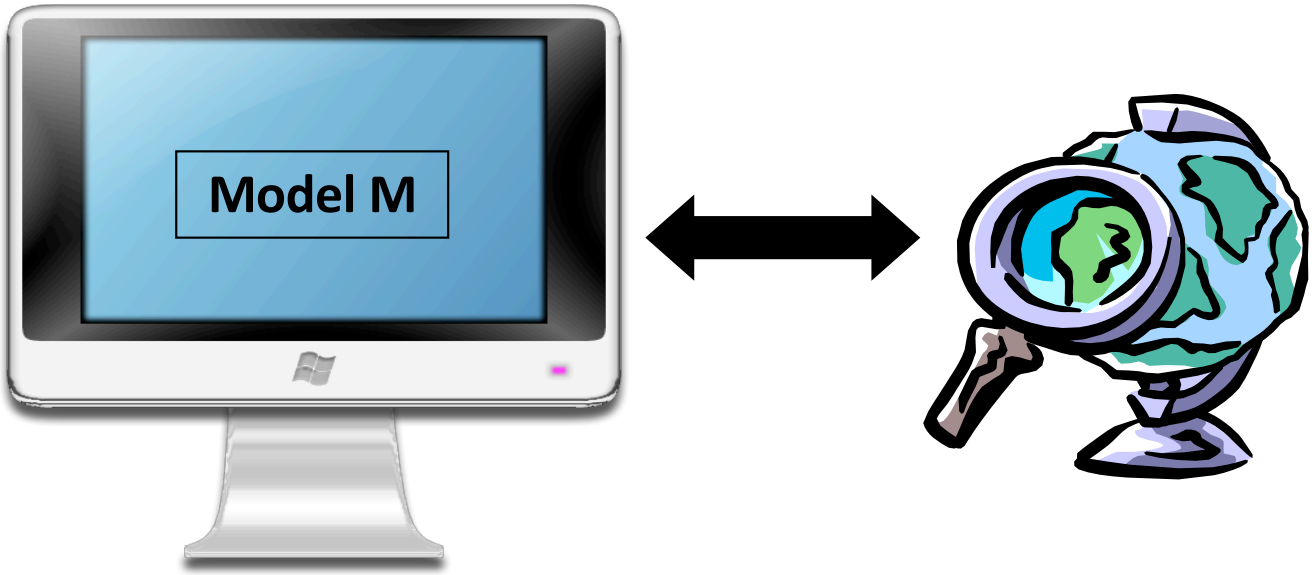
- course catalogue, accounting system, library system, news, search engine, genome information, campus map, social network, …

Computer systems that are not IS?

- IP telephony, computer game, Matlab

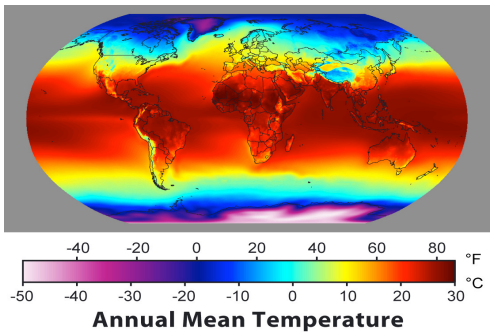
What is an Information System?

An information system is a **software** that manages a **model** of some aspect of the **real world** within a (distributed) computer system for a given **purpose**.

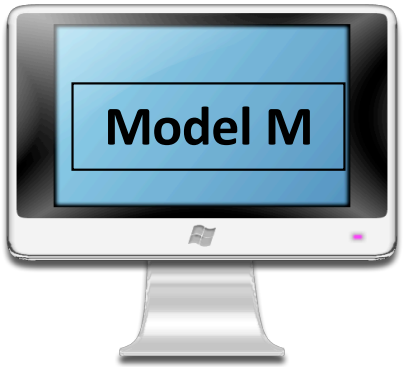


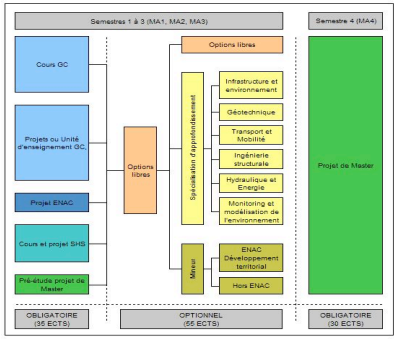
Real World Aspects

Human knowledge



Physical phenomena





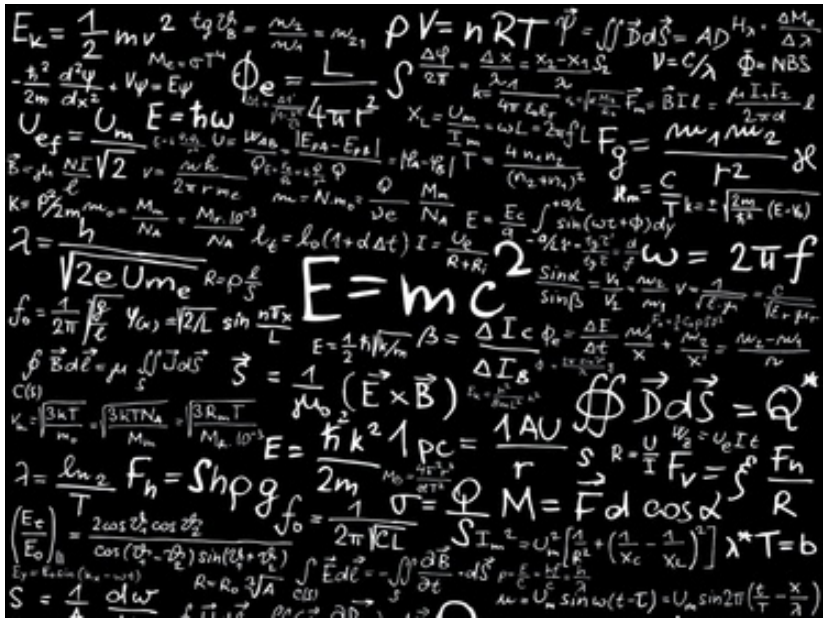
Social organization

Models

A **model** is a mathematical structure consisting of a set of

- **Constants** (or identifiers)
- **Functions** (or relations)
- **Axioms** (or constraints)

The set of constants and functions must be consistent with the axioms

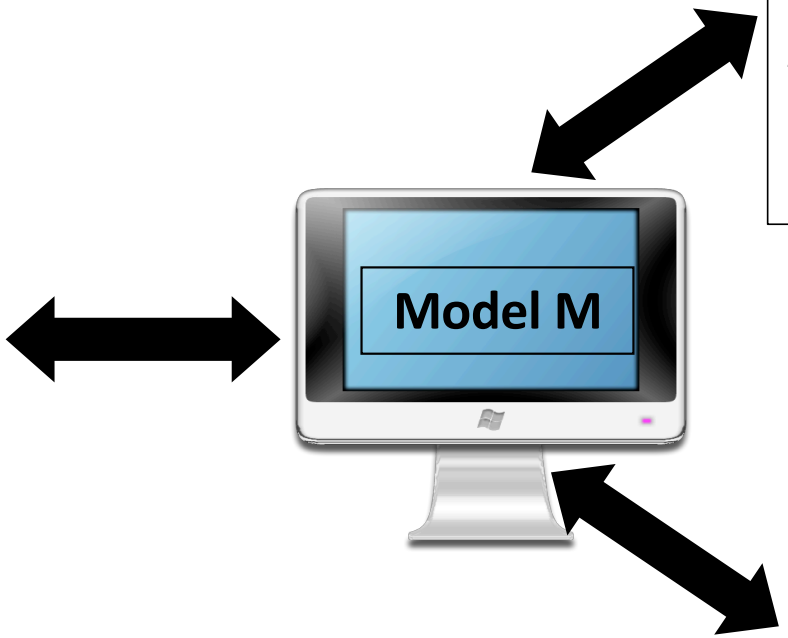


Examples of Models

Constants: coordinate values, temperature values

Function:
$temp(x, y)$
Axiom:
$-60 < temp(x, y) < 60$

Physical phenomena



Constants: document identifiers, text (sequence of words)

Function:
$similar(doc_1, doc_2)$
Axiom:
$0 \leq similar(doc_1, doc_2) \leq 1$

Human knowledge

Constants: names of people and units

Function:
$memberOf(name, unit)$
Axiom: each person belongs to at least one unit

Social organization

Representation of Functions

Functions can be represented by giving a specification or algorithm (implicitly) or by enumeration (explicitly).

We call enumerated functions **data**.

Example:

- Implicit representation: $f(x) = x^2$
- Explicit representation: $f(1) = 1, f(2) = 4, f(3) = 9$

Functions in Information Systems

Information systems strongly rely on explicit representation

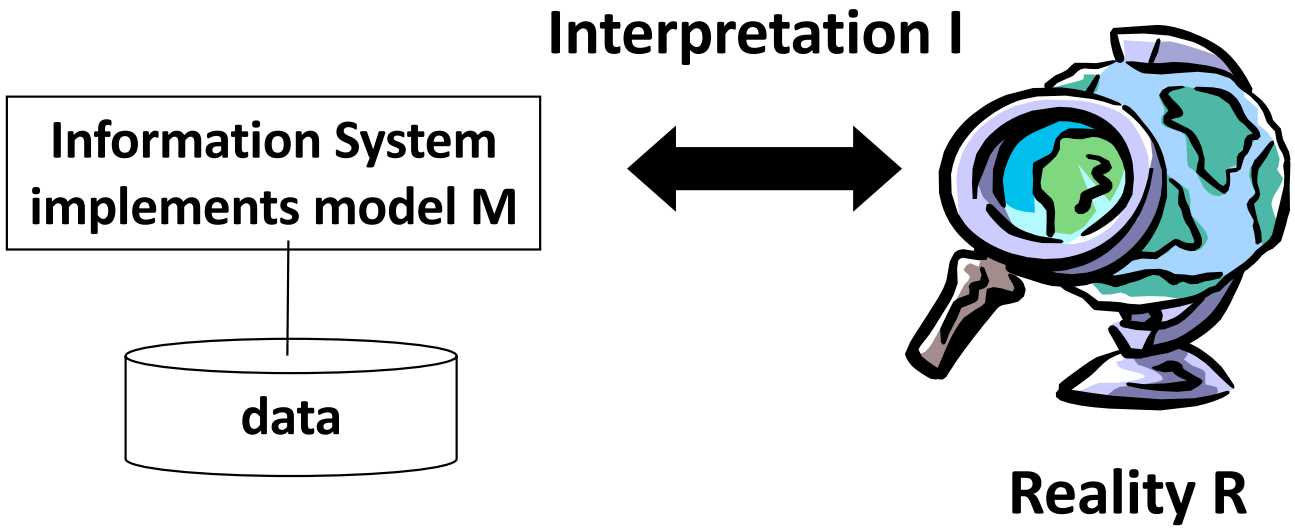
- many aspects of the world are not algorithmically defined, e.g., birthdate of a person
- difference to simulation systems

Computed functions (implicit representation) play nevertheless an important role

- queries, views, user-defined functions, etc.

How do we know that a model is a model?

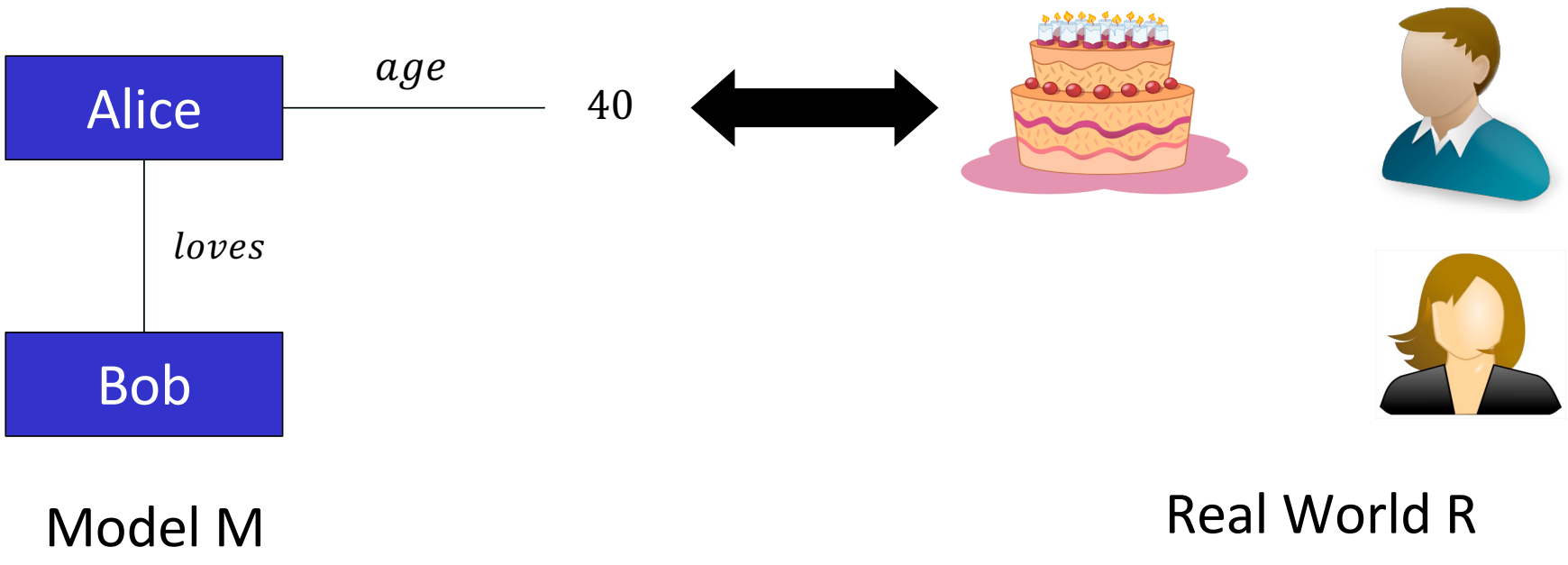
The model is linked by an interpretation (relationship) to the real world



Interpretation

The interpretation relationship is **homomorphic**

- I: M ➔ R
- preserves relationships

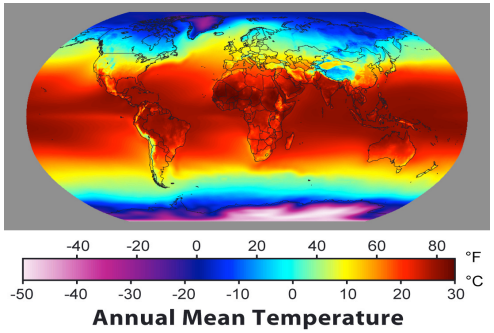


Interpretation is hard!

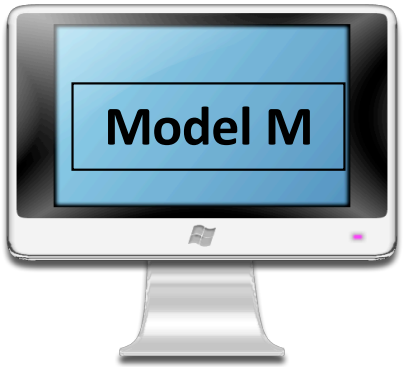
Information Retrieval

Human knowledge

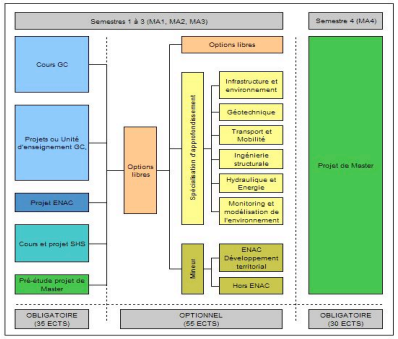
Science



Physical phenomena



Business Analysis



Social organization

Functions in models …

1. can always be computed
2. can always be represented as data
3. can sometimes be both computed and enumerated

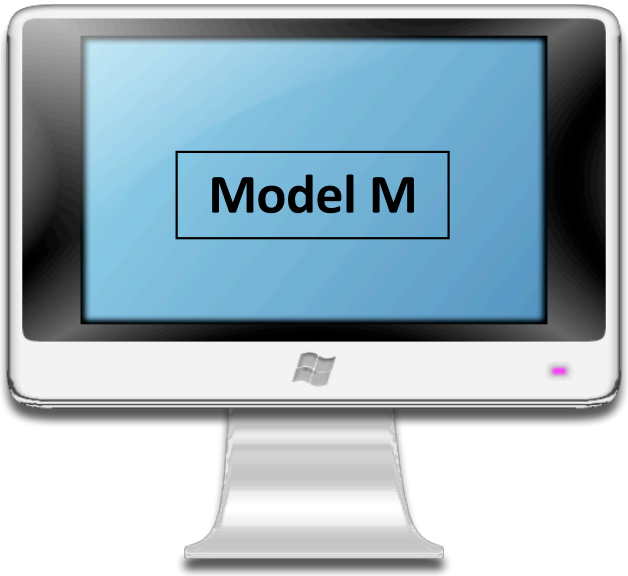
Interpretation relationships ..

1. are always computable
2. are sometimes computable
3. are uniquely defined

2. DATA MANAGEMENT

Data Models

How is a model represented in a computer system?



A model M is represented using a **data model D**.

A data model D uses **data structures and operations** for the representation of the constants, data, functions and constraints of a model M within a computer system.

Abstract Data Types

Are mathematical definitions of the properties of data structures

Example: associative array `A`

- Operations:
  `A.add(key, value), A.search(key), A.delete(key)`
- Constraint: every key occurs only once

Can represent a function $f: K \rightarrow V$

Data Structures

Implement abstract data types

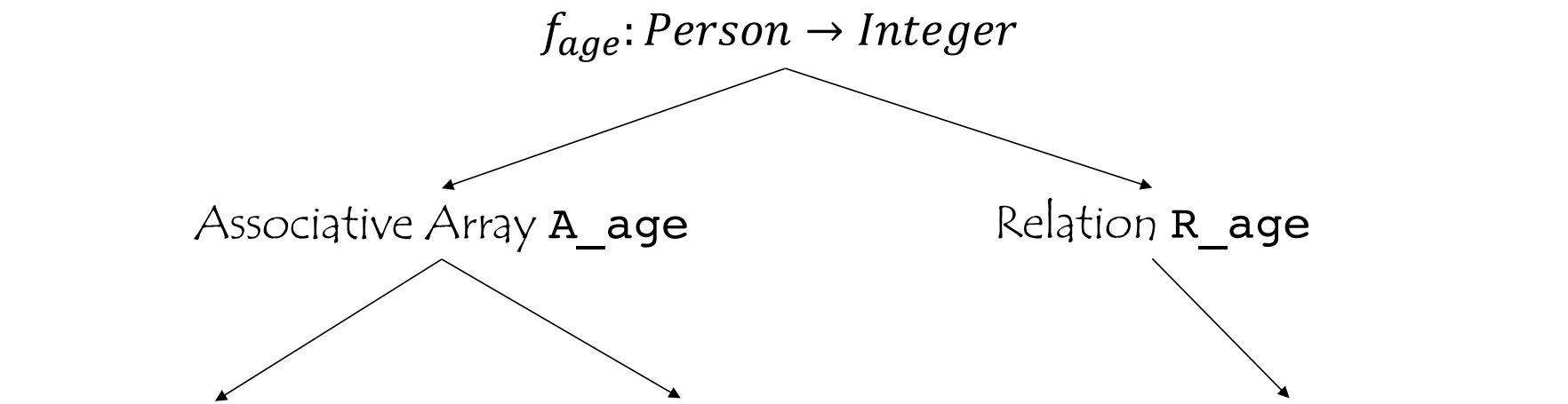
Examples of implementations of associative array:

- hash tables, binary search trees, tries

Different implementations have different performance characteristics

Relationship among Model and Data Model

The same function can be represented using different data structures




```
dict_age = {}
dict_age['Anne'] = 32
dict_age
```

```
{'Anne': 32}
```

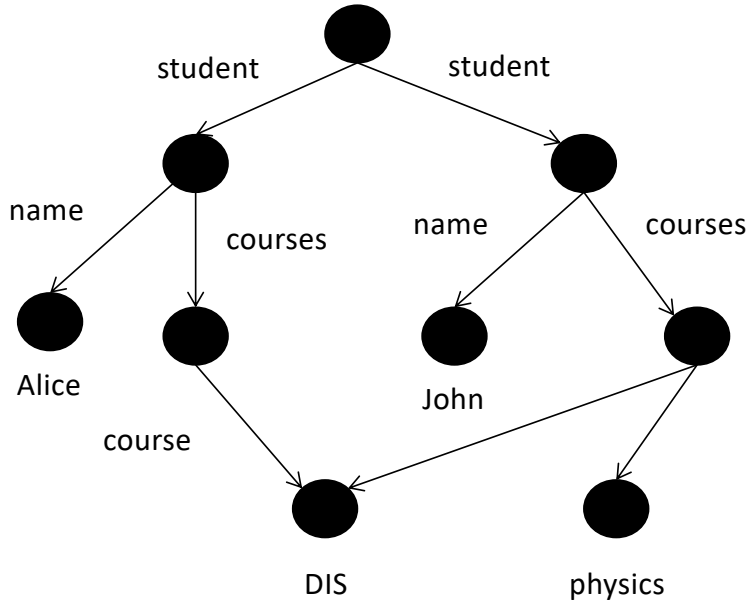
```
list_age = []
list_age.append(['Anne',32])
def get_age(list, s): [l[1] for l in list if s == l[0]]
list_age
```

```
[['Anne', 32]]
```

```
import pandas as pd
d = {'name': ['Anne'], 'age': [32]}
df = pd.DataFrame(data=d)
df
```

	age	name
0	32	Anne

Some Popular Data Structures



STUDENTID	NAME	COURSE
1234	John	DIS
3456	Bob	physics
2345	Alice	DIS

Labelled Graphs

Relational Tables

Database

The collection of data represented in a data model D is called a **database DB.**

A computer system that is designed to (generically) manage databases is called a **database management system DBMS**.

Data Modeling Languages

A data modeling language is a language used to specify data models. It consists of

1. Data Definition Language (DDL)
2. Data Manipulation Language (DML)

The specification of a data model using a DDL is called a **database schema**, or simply **schema S**.

Data Modeling Languages Components

A data modeling language specifies three main components:

- **Data structures**: a collection of data structures, which are used to represent databases.
- **Integrity constraints**: a language to express rules the data in databases has to observe.
- **Manipulation**: a collection of operators, which can be applied to the data structures, to update, transform and query the data, in a database.

Example: Relational Schema

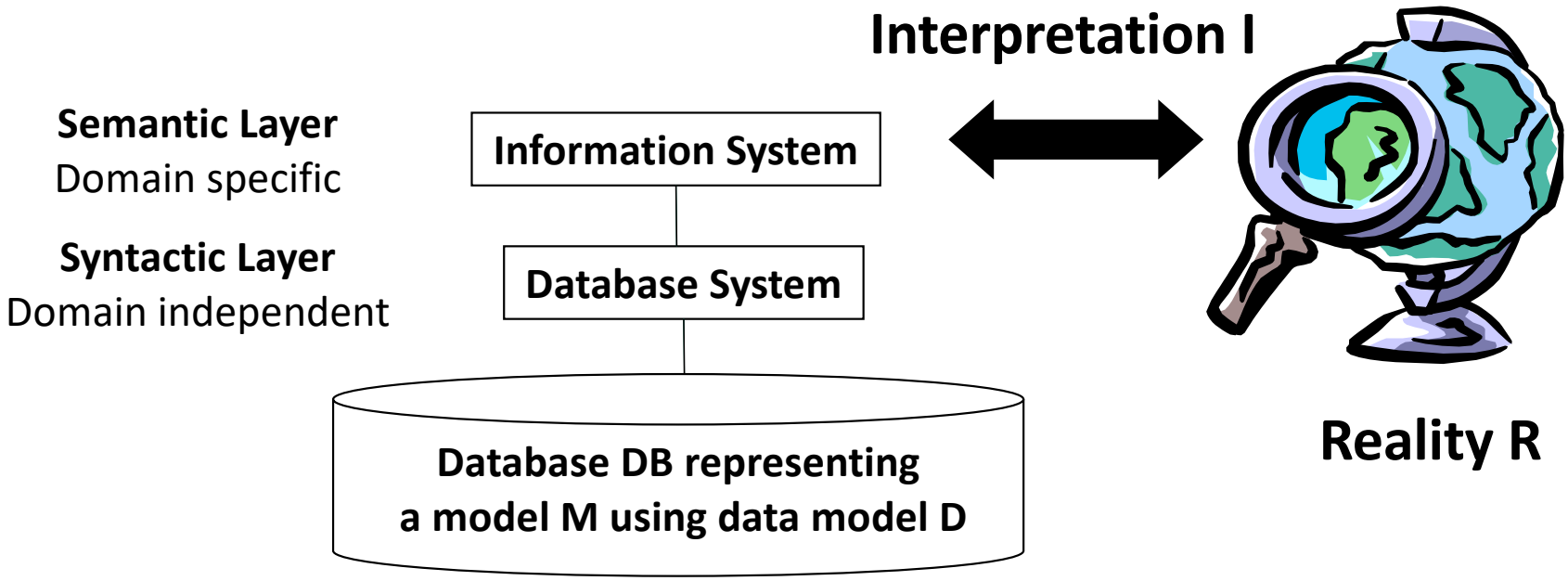
Data Definition Language

```
CREATE TABLE Student
     (Studentid NOT NULL PRIMARY KEY, Name, Course)
```

Data Manipulation Language

```
SELECT Name FROM Student WHERE Course = 'DIS'
```

Database Systems



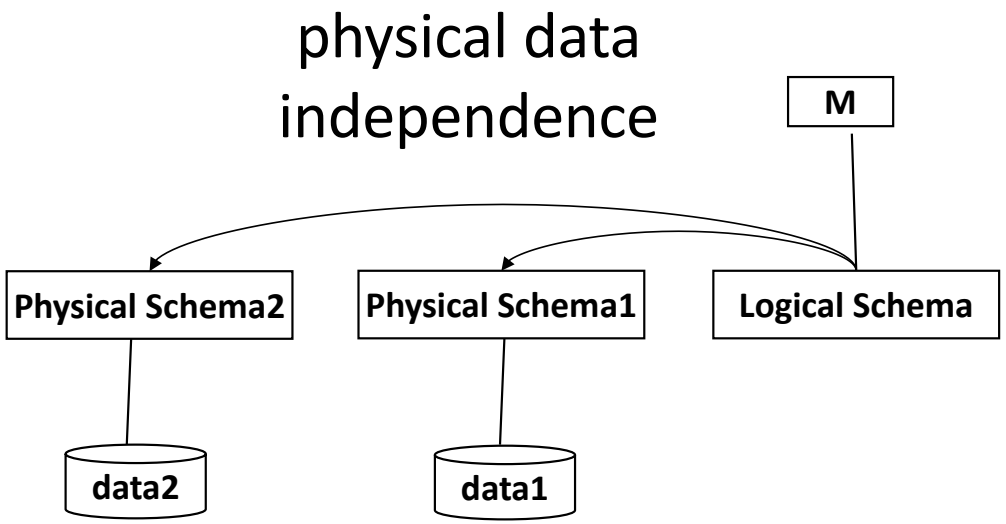
Database Management Systems

The software that implements a data modeling formalism is the **database management system**

Database management systems allow to implement many information systems. They are designed to be **application- or domain-independent**.

The data is stored using encodings of data structures exploiting available storage media (e.g. main memory or disk) and their addressing mechanisms.

Physical Data Independence



Mechanisms for supporting physical data independence
- Physical database design
- Declarative query optimization
- Transaction management

Different physical realizations of the database implementation (storage layout, query execution) do not affect the result

Key Data Management Tasks

Efficient management of large amounts of data

- efficient storage and indexing
- efficient search and aggregation

Ensuring **persistence** and **consistency** of data under updates and failures

- Persistence = data stored independent of lifetime of programs
- Consistency = data correct independent of type of failures

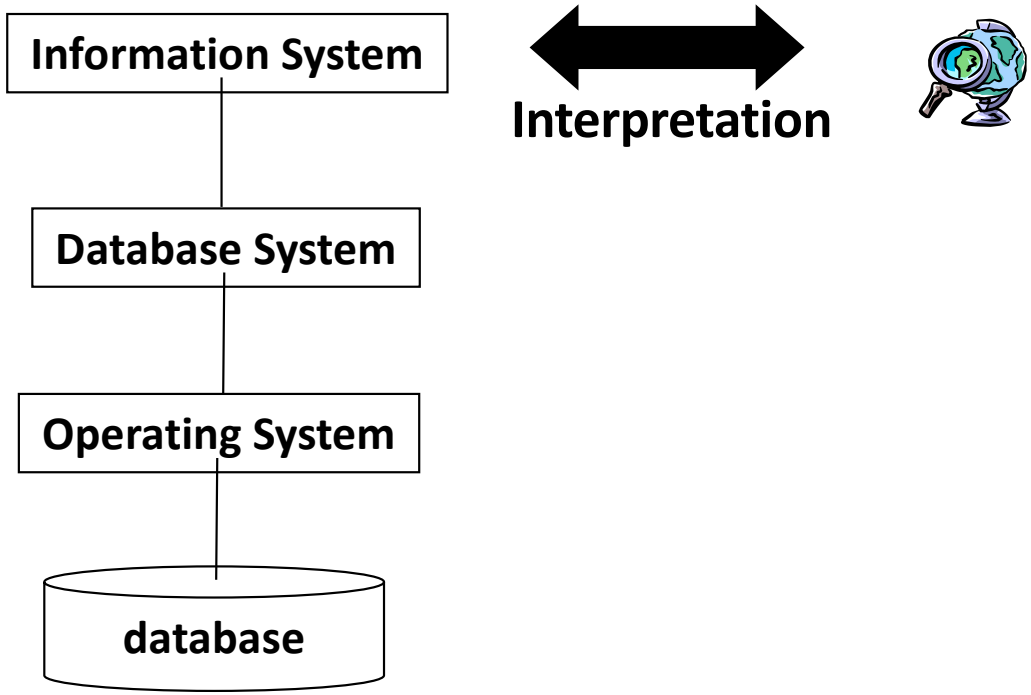
Perform both tasks by exploiting different media (e.g. memory, disk, flash, tape)

Refined View of an Information System

Semantic Layer
Application/Domain specific

Syntactic Layer
Application/Domain independent

Physical Layer
Storage



What is not specified in a data definition language?

1. The structure of a relational table
2. The query of a user
3. A constraint on a relational table
4. The definition of a view

Which is wrong? An index structure …

1. is defined as part of physical database design
2. is selected during query optimization
3. accelerates search queries
4. accelerates tuple insertion

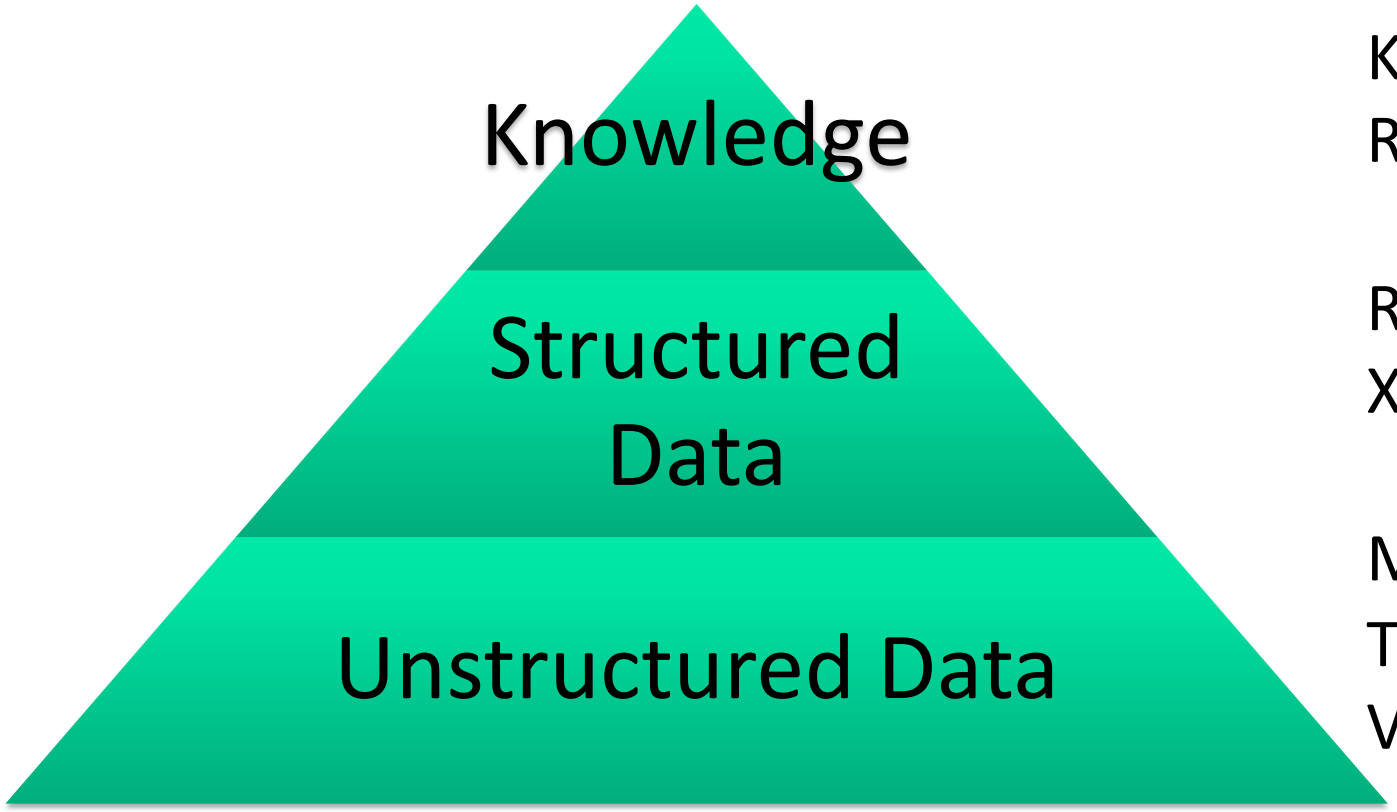
Persistence means that …

1. a change of a transaction on a database is never lost after it is completed
2. the state of a database is independent of the lifetime of a program
3. the same logical database can be stored in different ways on a storage medium

3. INFORMATION MANAGEMENT

Levels of Abstraction

Depending on the “degree of abstraction” data is classified



Knowledge Graphs
Rules

Relational data
XML and RDF data

Measurement data
Text, Images and Video

Levels of Abstraction - Characteristics

Unstructured data

- Data captured from measurements and human input
- Structure given by static data types (e.g. time series)

Structured data

- Data is structured according to a schema
- Captures relationships in the data

Knowledge

- Schema can evolve dynamically as knowledge expands
- Decisions are made using the information

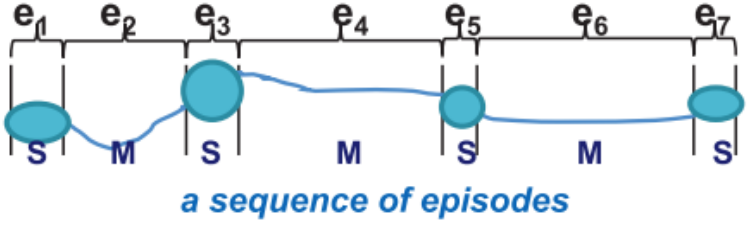
The classification is not strict!

Example

Unstructured data: a GPS trace

Bob's and Alice's GPS trace

Structured data: Road segments, Places



Places that Bob and Alice visit

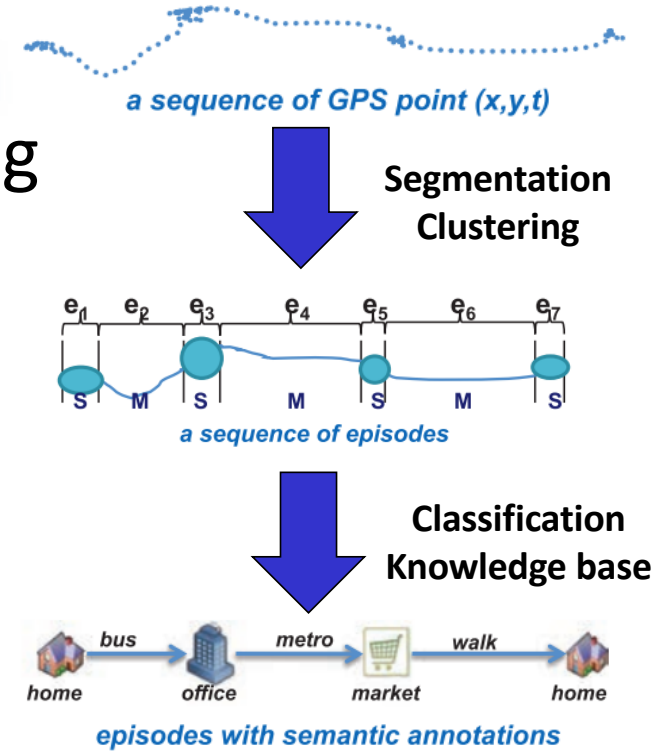
Knowledge: Concepts and Inferences

Bob and Alice are frequently together in Ouchy, thus: *Bob loves Alice*

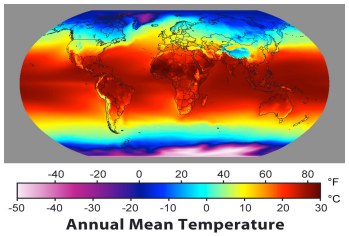
Model Building

Creating “higher level abstractions” from “lower level data”

- Using Statistical and Machine Learning methods, as well as rule-based approaches
- Typically on large datasets
- Also called: data mining, data science, data analytics etc.



Information Management Tasks



model M

Model Usage: given a model, find some specific data

Example:
Temperature model allows to find temperature at selected locations

Model Building: given data, find a model the matches the data

Example:
Given temperature measurements, find a function that interpolates for all locations

data

Model Usage

Model is used to generate data for performing various tasks

- Document retrieval
- Query answering
- Prediction

Model usage generates new data, from which more models can be built.

Model Building

Adding new functions that “interpret” the data in novel ways.

Example:

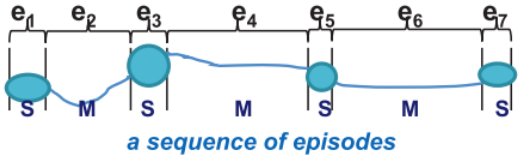
- Original model: $position(t): T \rightarrow \mathbb{R}^2$
- New model: $street(p_1, p_2): \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{B}$

The function $street$ is obtained by analyzing the velocity of position data

Methods: Data mining, Machine Learning, Data cleaning, Data integration, Inference, Rules

Representing the Models as Data

Model (Information System)



Data (Database System)

x	y	t
12	13	5:00
12	14	5:01
13	15	5:02

place	x	y	d
p111	12	13	2:00
p112	13	15	1:00

segment	xs	ys	xe	ye
s221	12	13	13	15
s222	13	15	20	27

Subject	Relation	Object
home	ISA	Place
bus	ISA	Vehicle
Bob	ISAT	home

Representing the Models as Data

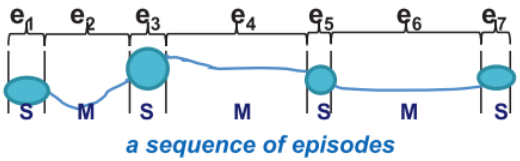
Model (Information System)

Data (Database System)

Obtain data from measurement

x	y	t
12	13	5:00
12	14	5:01
13	15	5:02

Derive model from data

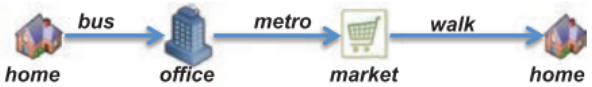


Generate data from model

place	y	d
p111	13	2:00
p112	15	1:00

segment	xs	ys	xe	ye
s221	12	13	13	15
s222	13	15	20	27

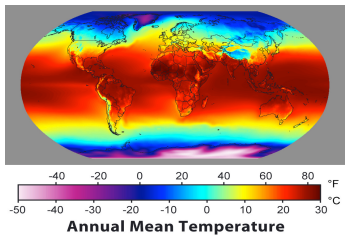
Derive model from data



Generate data from model

Subject	Relation	Object
home	ISA	Place
bus	ISA	Vehicle
Bob	ISAT	home

Information Management Tasks

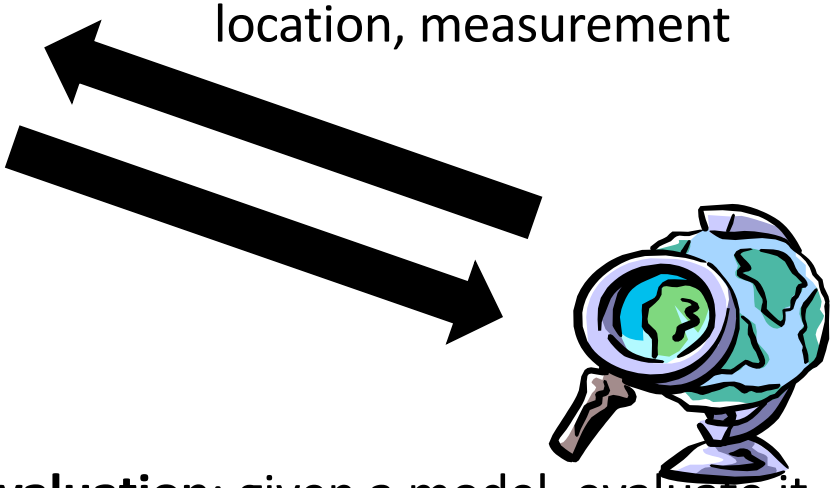


model M

data

Conceptual modeling: analyze the real world and specify a model

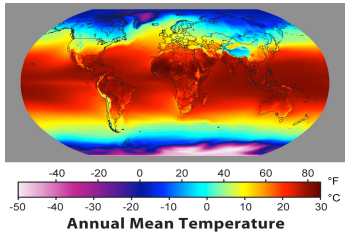
Example: define concepts temperature, location, measurement



Evaluation: given a model, evaluate it against reality

Example: compare predicted temperature to measurements

Information Management Tasks



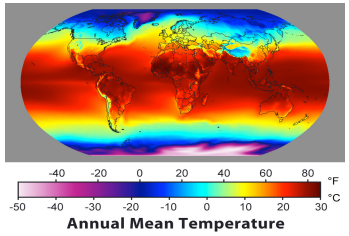
model M

data

Output: data visualization, control

Input: users, sensors

Information Management Tasks



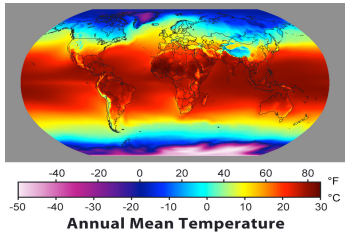
model M

data

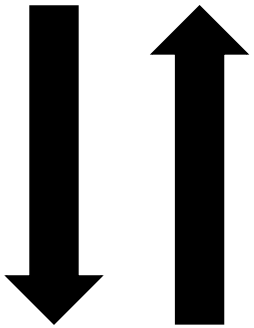
Control: adapt heating to expected temperature

Monitoring: collect data from weather stations

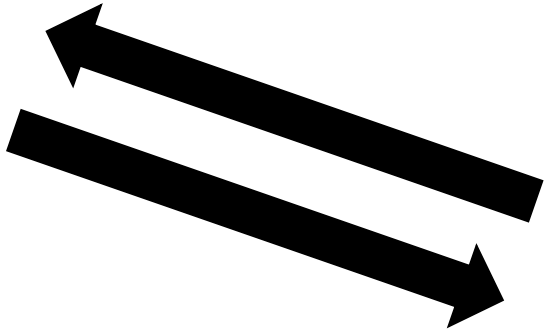
Information Management Tasks

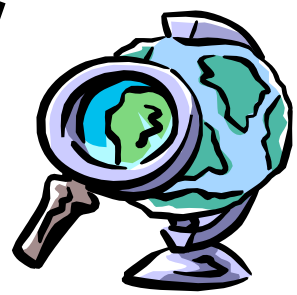


model M

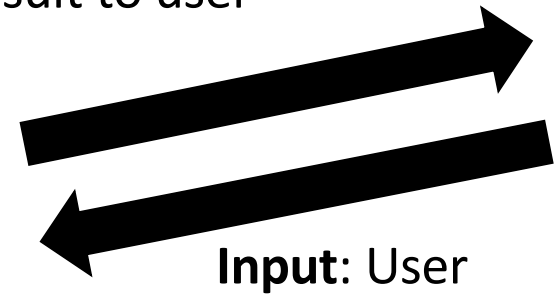


data



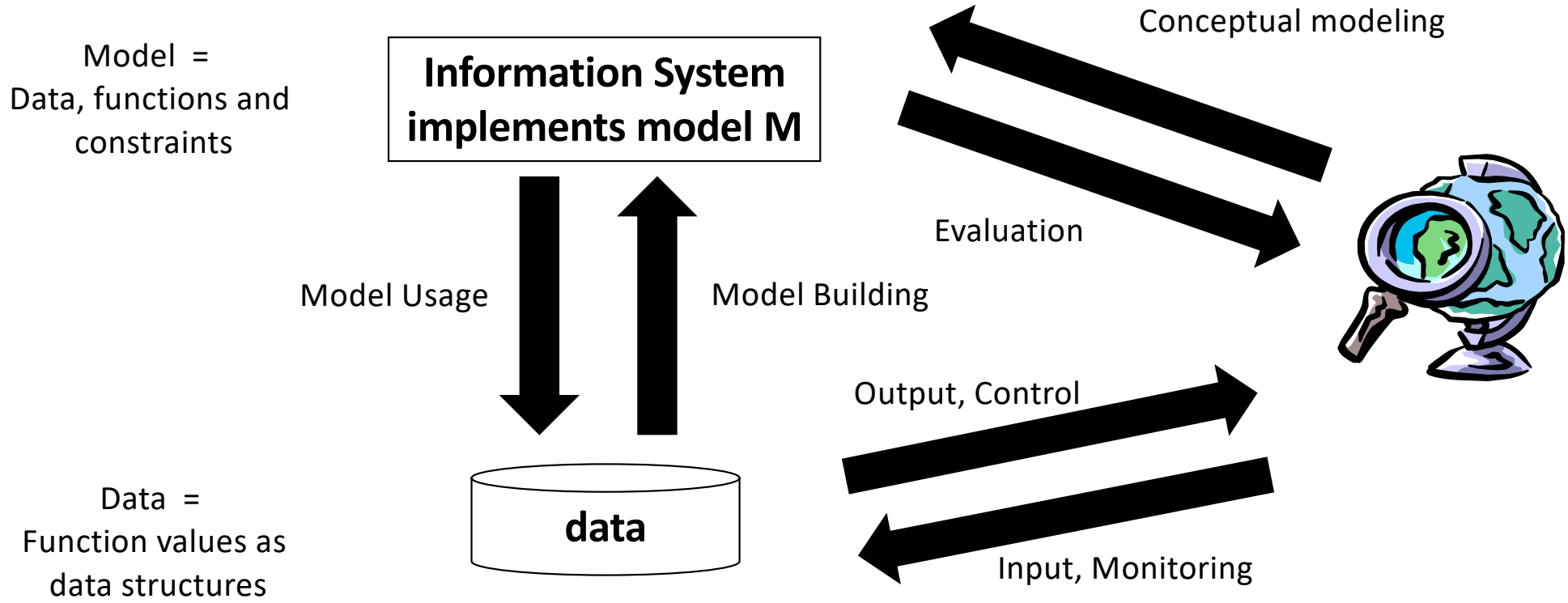


Output: show query result to user

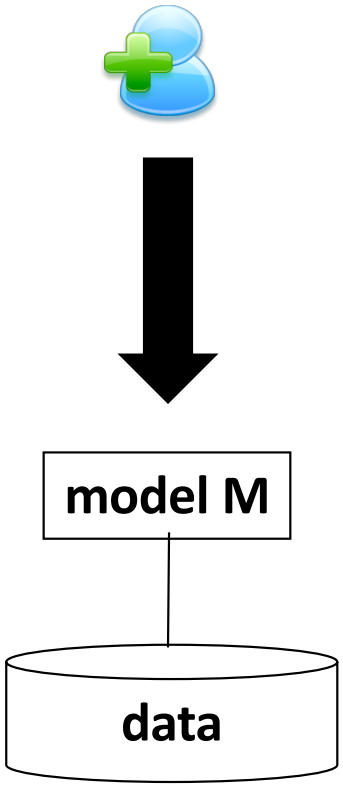


Input: User picks relevant results

Information Management Tasks



Utility



Users need information system to take **decisions**

Utility of information linked to the value achieved

Value depends on

- Importance of decision
- Quality of decision

Quality of decision depends on quality and understandability of information!

Using information systems for decision making is associated with the notion of **knowledge management**.

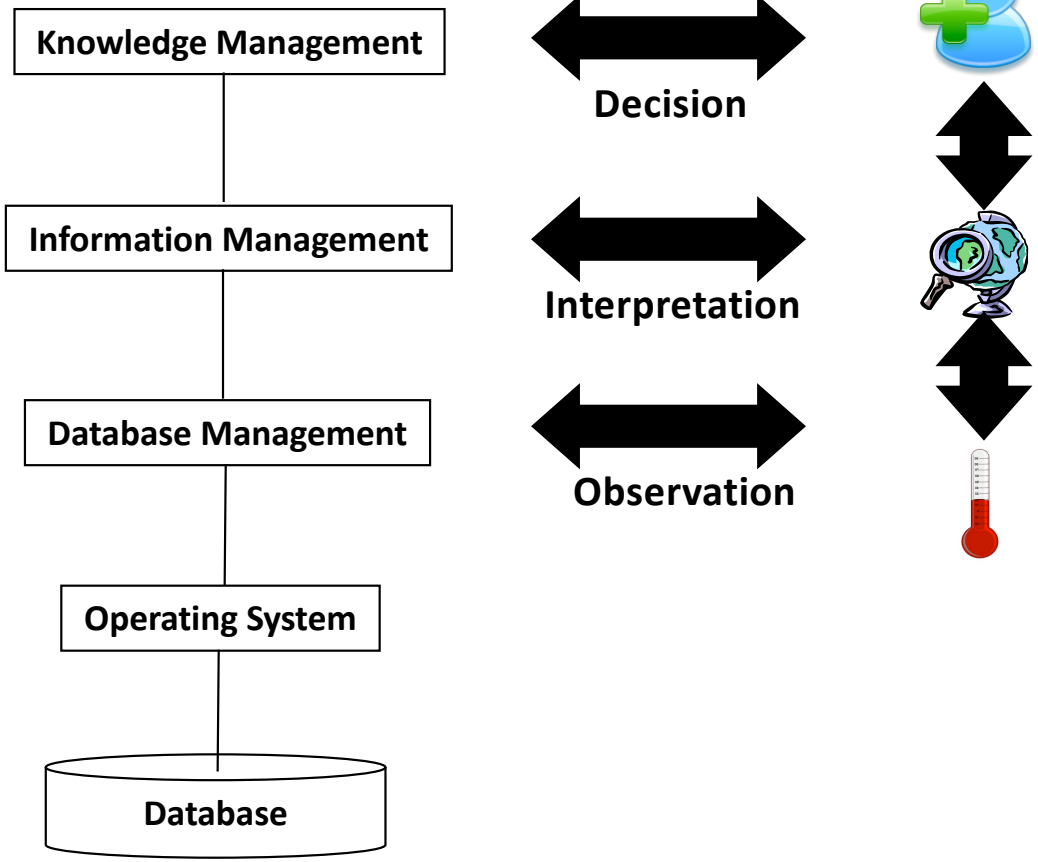
Refined View of an Information System

Pragmatic Layer
User/Community specific

Semantic Layer
Application/Domain specific

Syntactic Layer
Application/Domain independent

Physical Layer
Storage, Networks, Sensors



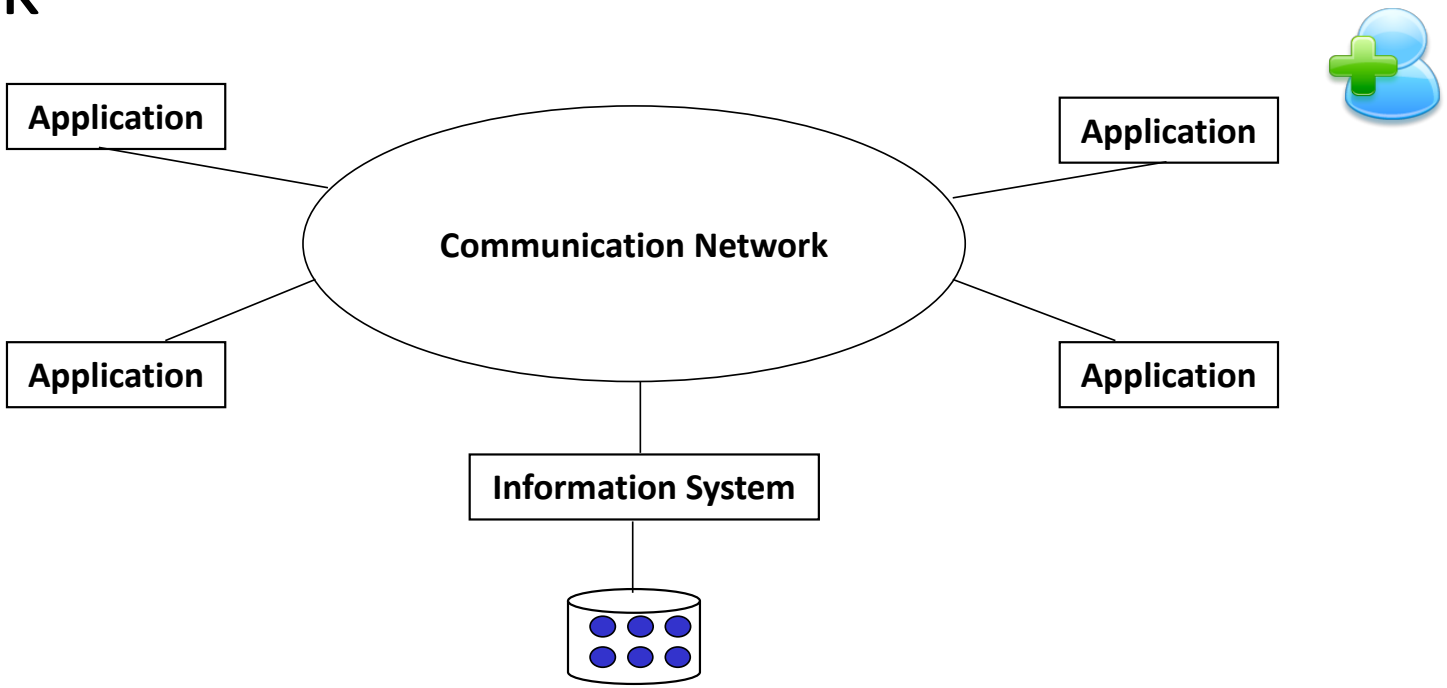
Grouping Facebook users according to their interest by analyzing the content of their posts is …

1. a retrieval task
2. a data mining task
3. an evaluation task
4. a monitoring task

4. DISTRIBUTED INFORMATION SYSTEMS

Centralized Information System

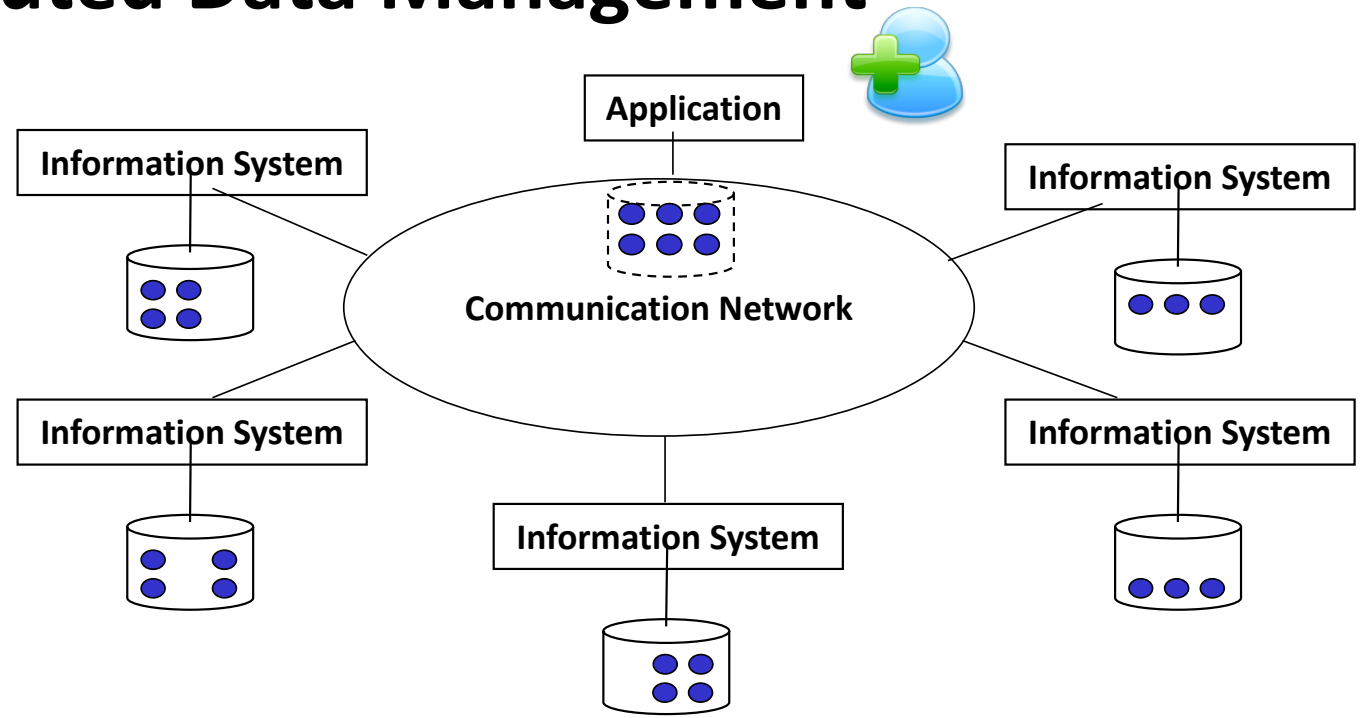
Centralized Information System on Computer Network



Physical Distribution

Use of distributed physical resources: locality of access, scalability, parallelism in the execution

Distributed Data Management

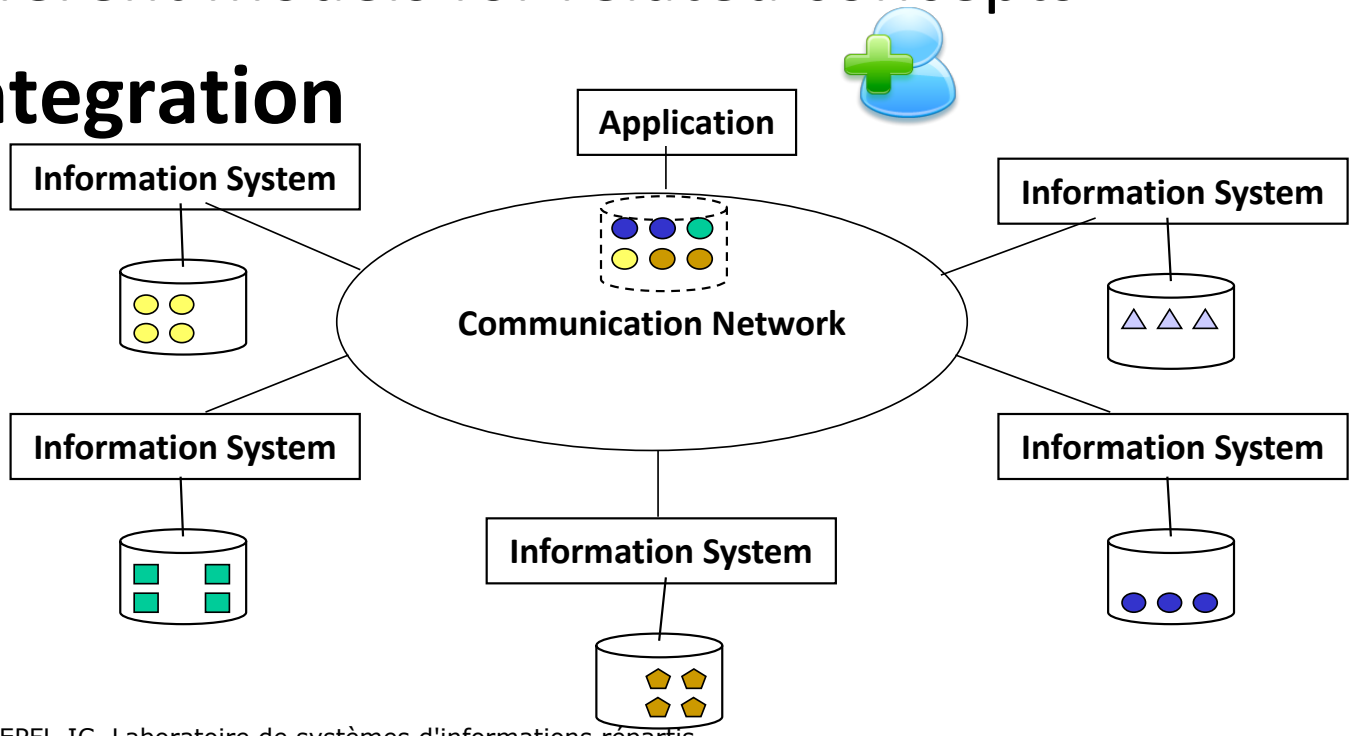


Heterogeneity – Logical Distribution

Use of different data models

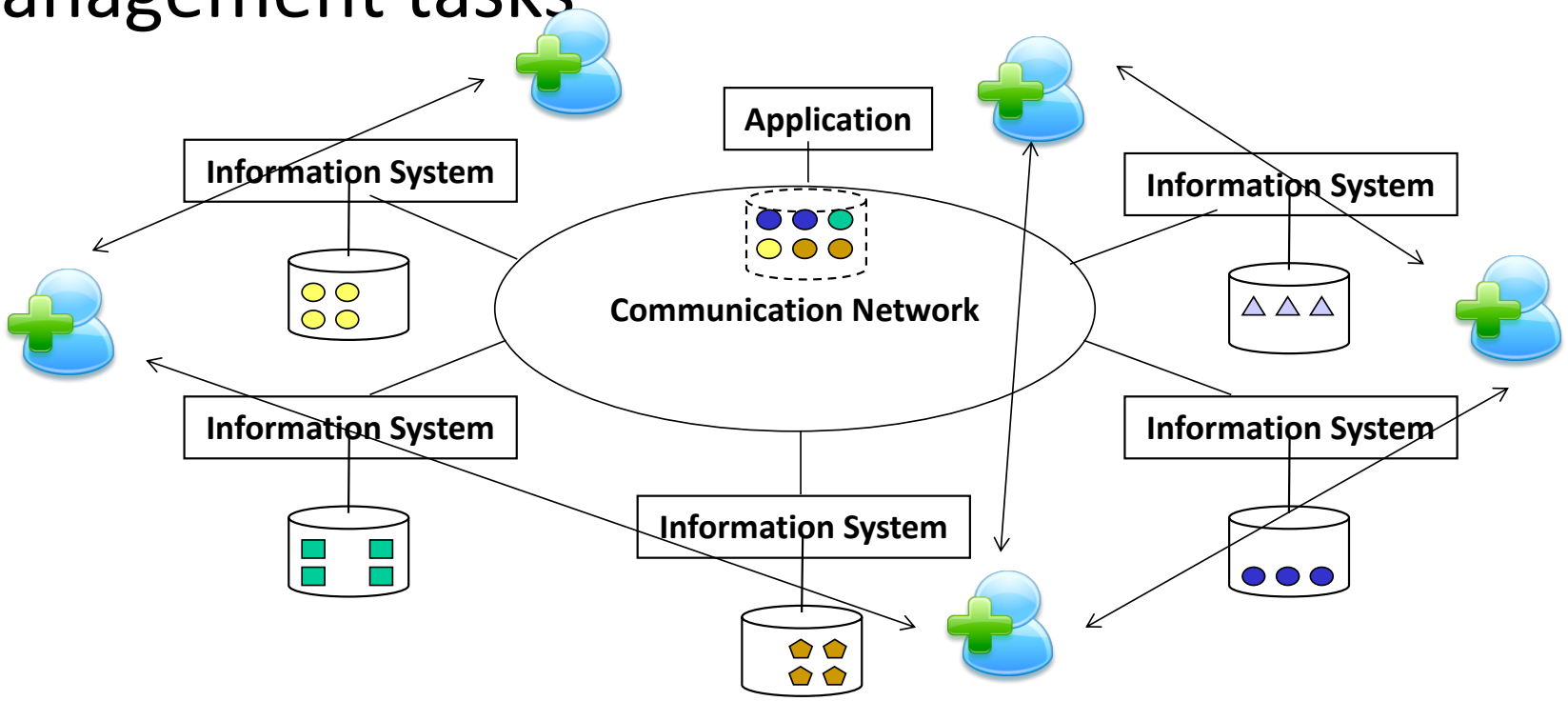
- Independently developed information systems
- Different models for related concepts

Data integration



Autonomy – Distribution of Control

Independent users have to collaborate, coordinate, negotiate, to perform information management tasks



Creating a web portal for comparing product prices is (primarily) a problem of …

1. Distributed data management
2. Heterogeneous data integration
3. Collaboration among autonomous systems

Key Issues in Distributed Data Management

Where to store data in the network?

- **Partitioning** of data
- **Replication** and **caching**
- Considering typical access patterns and data distributions

How to access data in the network?

- **Push** vs. **pull** access (query vs. filtering)
- Indexing of data in the network
- Distribution of queries and filters
- Considering the communication model

Key Issues in Heterogeneity

More Data - More Information?

Data overload

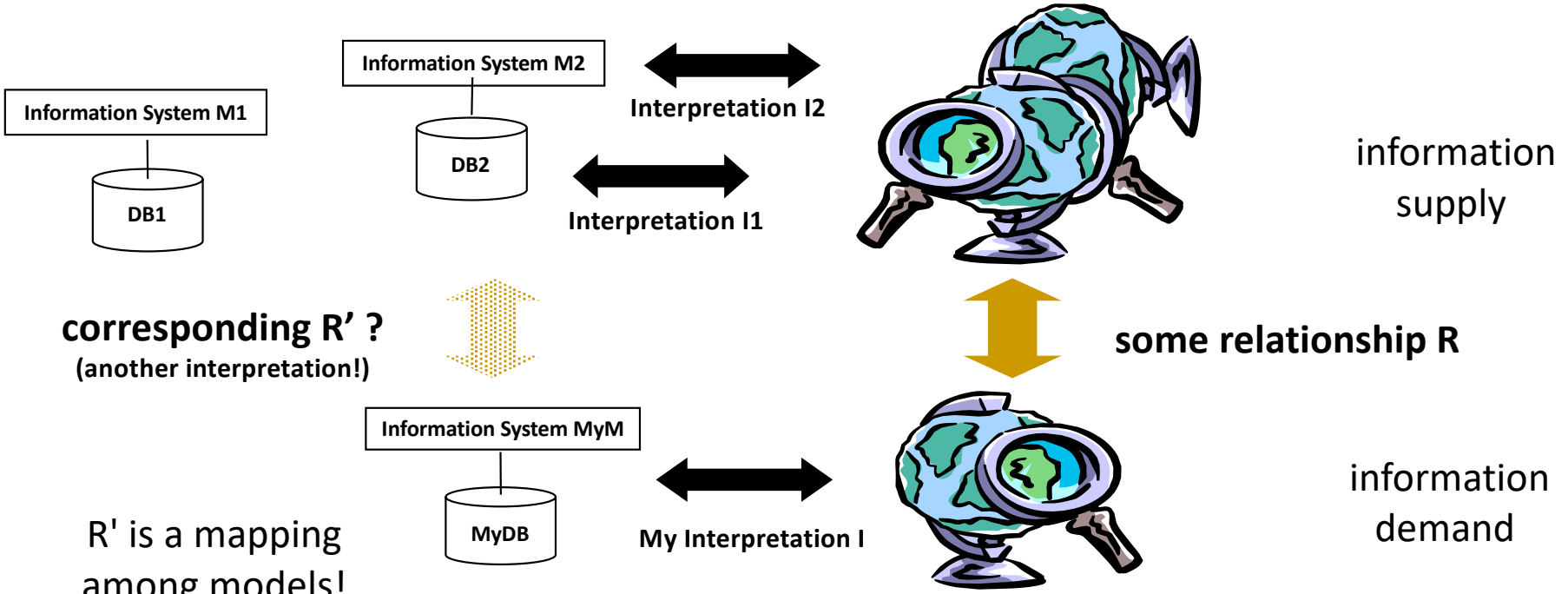
- more data, disintermediation
- More useful information ?

Information starvation

- problem: *data supply* does not match *data demand*
- models used by data provider are different from models used by data consumer

Distributed Information Management

More data! … More models!? … More useful information?



The Problem

Semantic heterogeneity

- The same real world aspect can be modelled differently
- Relating different models (and thus different interpretations) requires often human intervention; human attention is a scarce resource !

Mapping: Three Approaches

Standardization

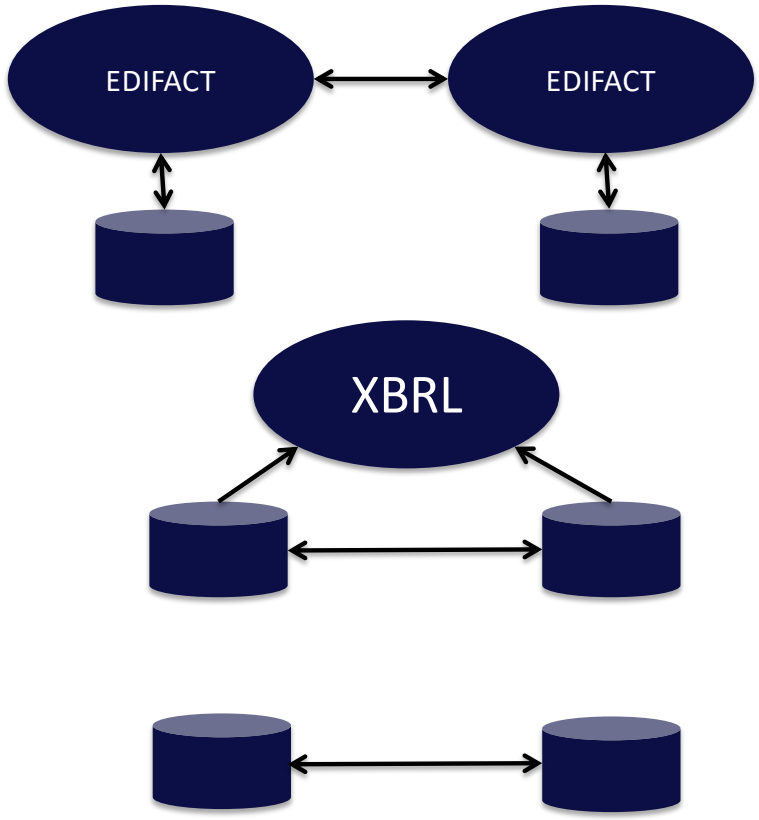
- Mapping through standards

Ontologies

- Mediated mapping

Mapping

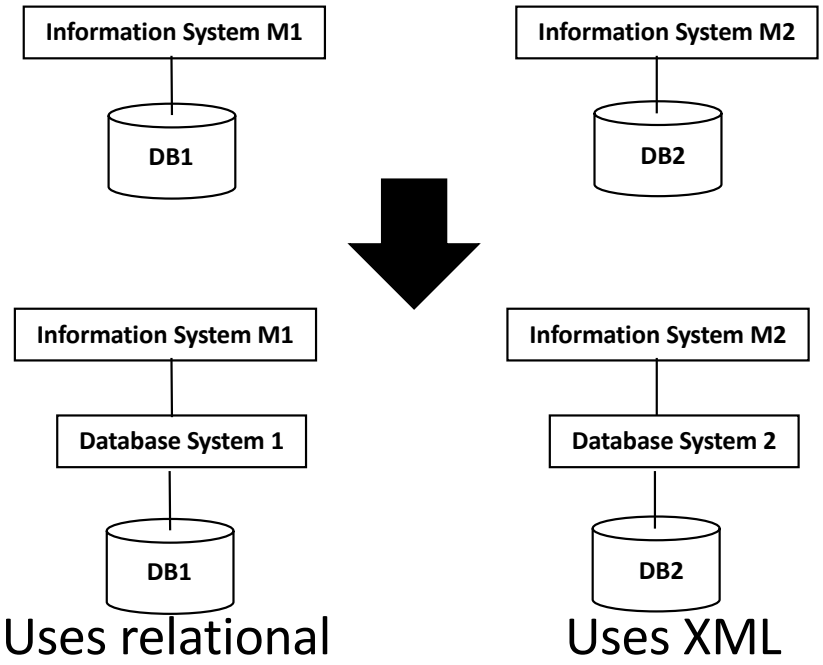
- Direct mapping



More Problems?

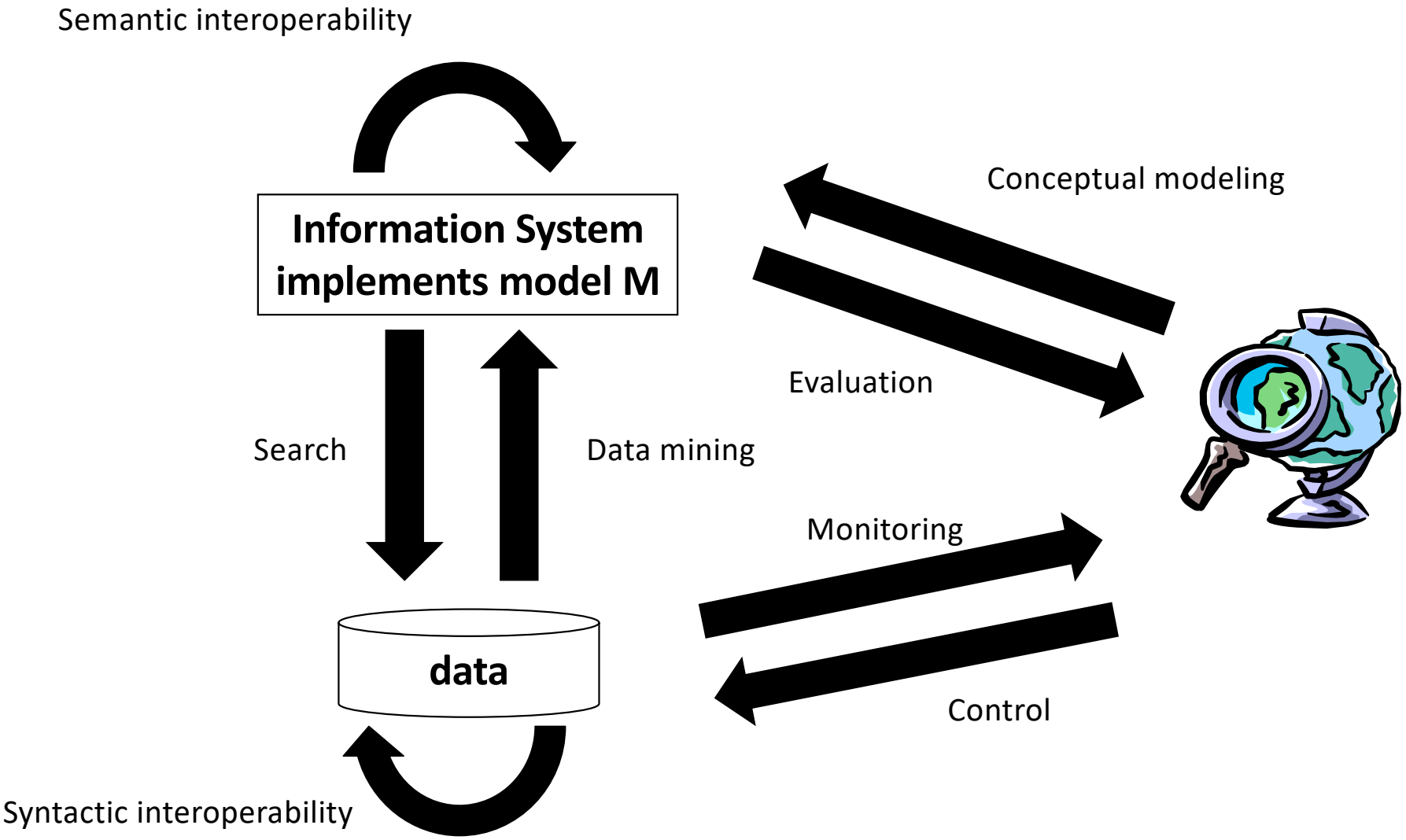
Syntactic heterogeneity

– The same data can be represented using different data models



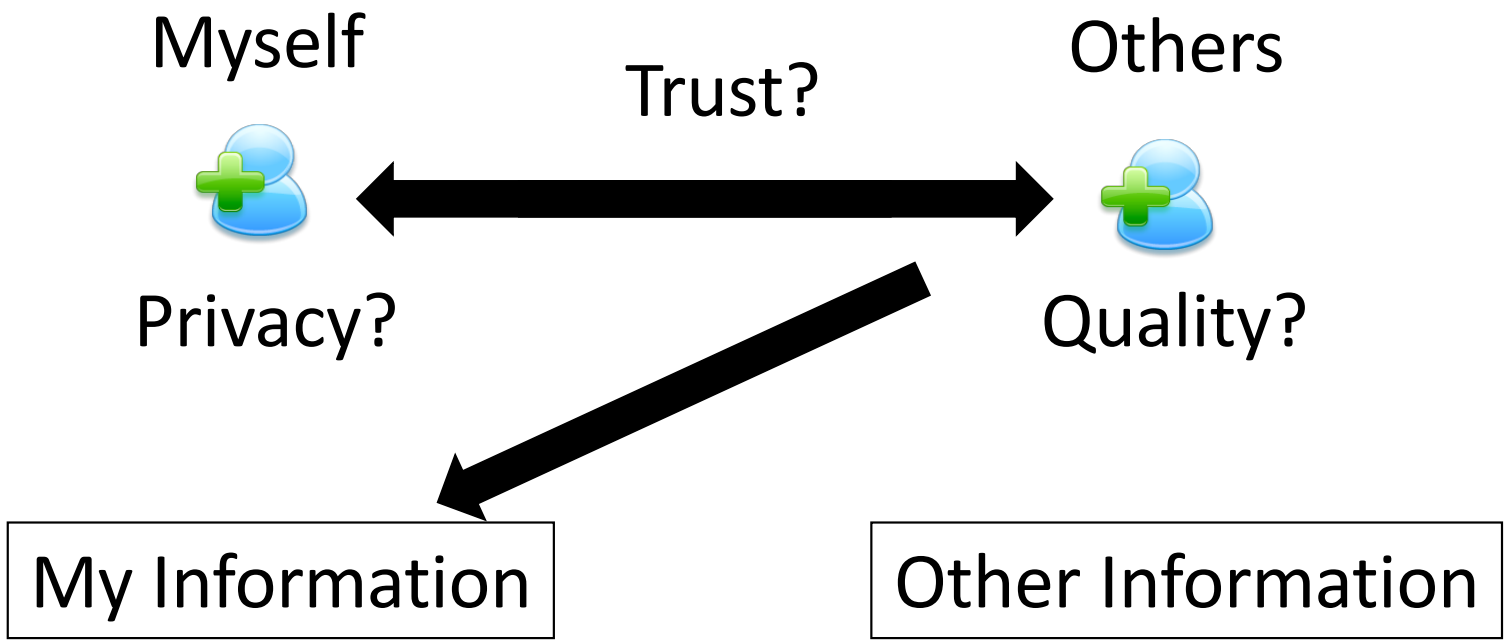
Data Models might be different!

Information Management Tasks



Key Issues in Autonomy

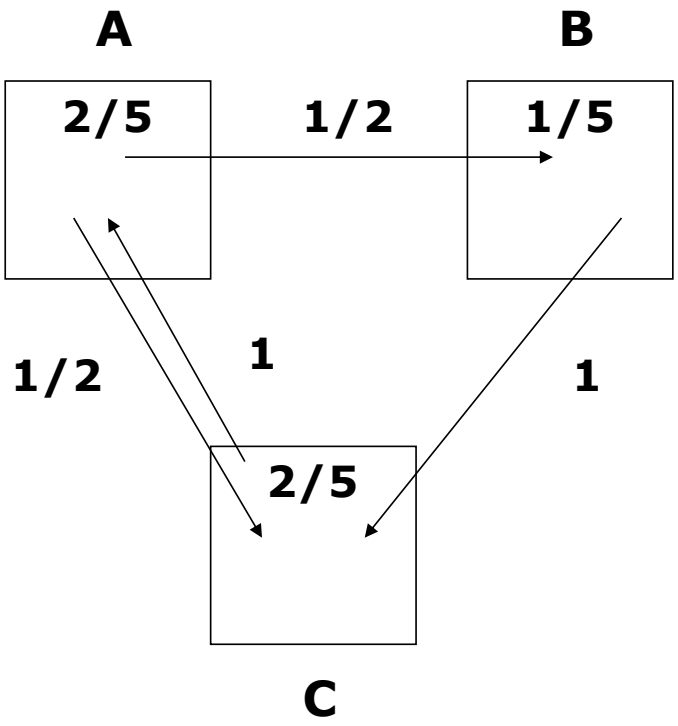
The Users Problem

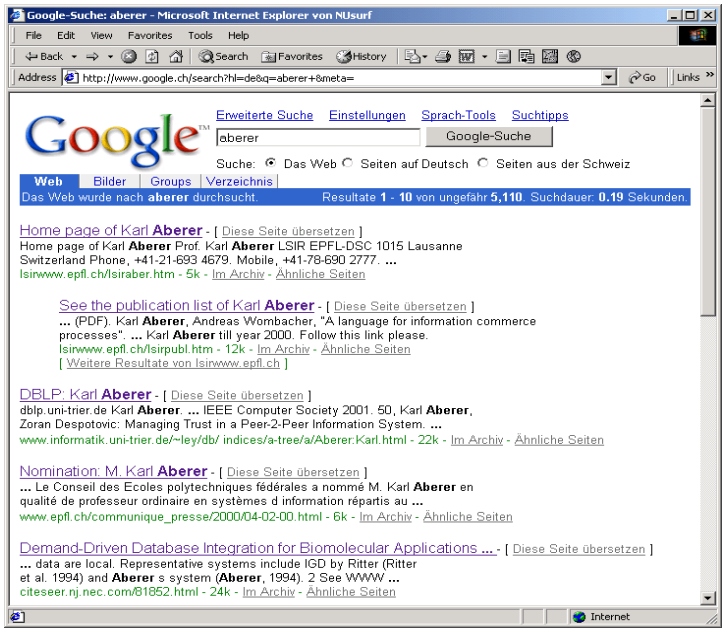


Revealing quality information increases trust,
but lowers privacy

Evaluating Quality of Information

Recommendations (e.g. Google PageRank)





Evaluating Trust

Reputation-based trust: if users behaved honestly in previous interactions, they will do so in the future

Overall profile makeup

94 positives. **91** are from unique users and count toward the final rating.

4 neutrals. **0** are from users no longer registered.

1 negatives. **1** are from unique users and count toward the final rating.

ebay ID card

seller01 (90)

Member since Tuesday, Dec 15, 1998

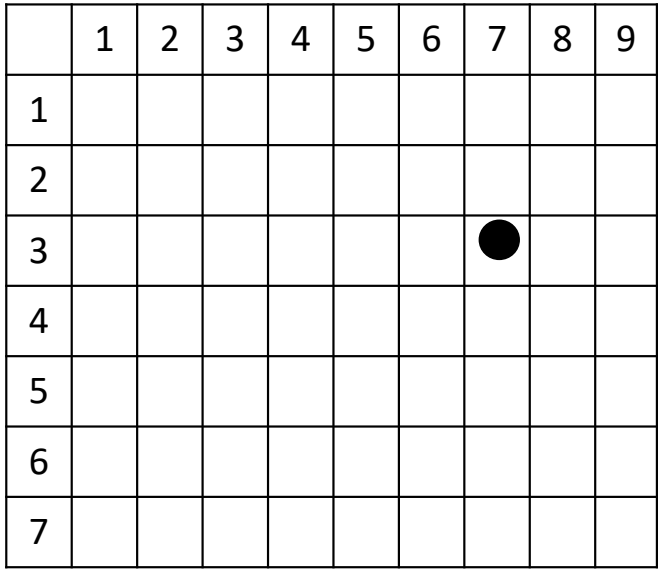
Summary of Most Recent Comments

	Past 7 days	Past month	Past 6 mo.
Positive	2	3	15
Neutral	0	0	0
Negative	0	0	0
Total	**2**	**3**	**15**
Bid Retractions	0	0	0

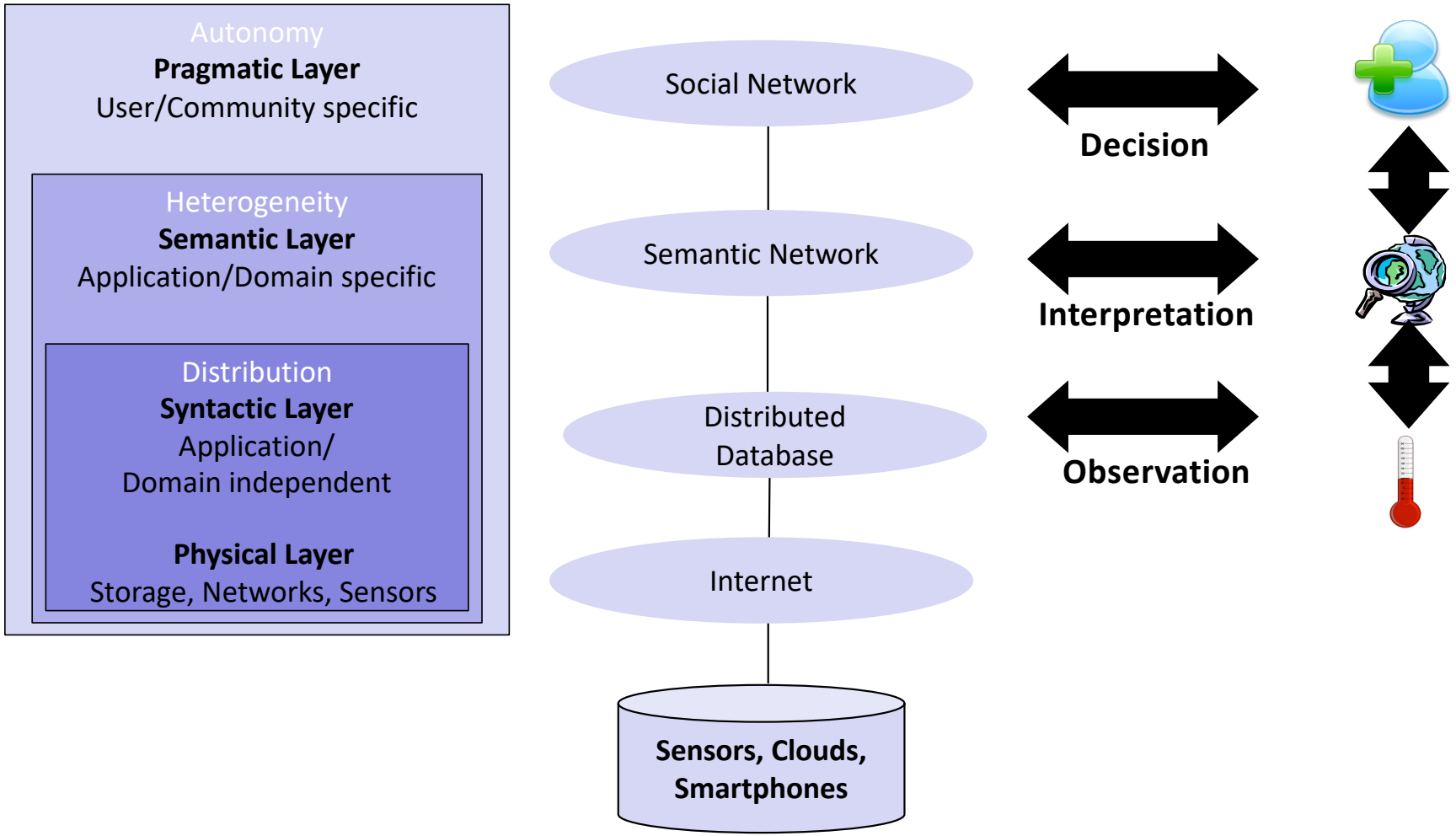
Protecting Privacy

Example: location privacy – obfuscation methods

- Perturbation: (3,7)
- Adding dummy regions: (3,5), (1,4), (6,3)
- Reducing precision: (2,5), (3,4), (3,5), (3,6), (4,5)



Refined View of a Distributed Information System



Trust is …

1. a quality of information
2. a quality of a user
3. a quality of the relationship among user and information
4. a quality of the relationship among users

Exercise

Big Data is often characterized by the four concepts of Volume, Velocity, Variety and Veracity

1. Inform yourself what is meant by those concepts
2. Identify from this lecture four problems / methods that are related to each of those four concepts