Statistical databases in theory and practice
Part I: Concepts and definitions

Bo Sundgren

2010

Databases and statistical databases

- ”A database is a well organised collection of data. It should be easy to process and update data in the database, and to add new data to the database. It should also be easy to retrieve data from the database, both planned and unplanned retrievals.” (Sundgren, 1981)
- A statistical database is a well organised collection of statistical data. It should be easy to process and update statistical data in the database, and to add new statistical data to the database. It should also be easy to retrieve statistical data from the database, both planned and unplanned retrievals.

Data warehouses

- A data warehouse, or corporate data warehouse, is a database, or a well integrated collection of databases (including metadata), meant for use by an organisation as a whole (and/or its customers)

Basic definition of a register
(register in a strict and narrow sense)

- A register is an authorised, up-to-date list of all objects belonging to a certain population
- The objects are uniquely identified by an authorised identifier, such as person number for persons, organisation number for enterprises and other organisations, etc
- In addition to the identifier, a register may contain additional basic and up-to-date information about the objects, such as name (not necessarily unique) and location and other contact information, e.g. address and telephone number

Data and statistical data

- Data may be the result of a measurement or a direct observation performed by a human being, or by a measurement instrument designed by a human being.
- Data may also be the result of a mental and physical process, where a human being tries to represent and communicate information by means of data.
- Data may be tranformed into other data by means of processes, designed by human beings and executed by human beings, machines (e.g. computers), and/or human beings and machines in cooperation.
- Statistical data are data used and/or produced for statistical purposes.

Statistical data

- Microdata – data about individual objects
- Macrodata – summarised (aggregated) data about collectives of objects, estimated values of statistical characteristics, ”statistics”
- Metadata – ”data about data”:
  - exploratory metadata, e.g. metadata to be used by search engines
  - explanatory metadata, e.g. definitions, quality declarations
  - technical metadata, e.g. formats and data types

What do we mean by ”statistics”?

- ”figures”, numeric data, presented in tables and graphs
- summarising data about collectives of objects
- summarising data obtained from individual observations by means of some kind of aggregation process

Examples

- the number of companies in Sweden are ...
- the average age of persons living in Sweden is ...
- the number of road accidents in Sweden last year was...
- consumer prices in Sweden have risen by ... percent last month
- the GNP per capita of Sweden is ...

Statistical characteristic

- a statistical measure (m) applied on
- the (true) values of a variable (V); V may be a vector
- for the objects in a population (O)

- O.V.m = statistical characteristic
- O.V = object characteristic
- V.m = parameter

Examples of statistical characteristics

- number of persons living in Sweden at the end of 2001
- average income of persons living in Sweden at the end of 2001
- correlation between sex and income for persons living in Sweden at the end of 2001

Statistical measures

- enumeration measures
- summation measures
- central measures
- variation measures
- covariation measures

Examples of statistical characteristics and their conceptual components

STATISTICAL CHARACTERISTIC	POPULATION	VARIABLE	MEASURE
Number of persons living in Sweden at the end of year 2001	persons living in Sweden at the end of year 2001	Φ (none)[1]	enumeration (frequency count)
Average income of persons living in Sweden at the end of year 2001	persons living in Sweden at the end of year 2001	income, in SEK, during year 2001	average (arithmetic mean)
Correlation between sex and income of persons living in Sweden at the end of year 2001	persons living in Sweden at the end of year 2001	(sex at the end of 2001, income during 2001)	correlation
Volumes of iron, in tons, produced by iron mines in Sweden during 2001	iron mines in Sweden existing at some time during 2001	volume, in tons, during year 2001	sum
Number of road accident in Sweden during 2001	road accident that took place in Sweden during the year 2001		enumeration (count)

[1] Enumeration, or frequency counting, may be seen as a function (statistical measure) with no argument. Alternatively, it may be regarded as a function with one argument, where the argument is a variable, which takes the value “1” for all objects to be counted; the statistical measure would then be “sum”.

Statistic

- an estimator (e) applied on
- *observed* values of an *observed* variable (V');
- for a set of *observed* objects (O') *allegedly* belonging to a population (O)

Ideally the value of a statistic O'.V'.e should be "close to the true value of the statistical characteristic O.V.m that it aims at estimating

Examples

- the estimated number of persons living in Sweden at the end of 2001
- the estimated average income of persons living in Sweden at the end of 2001
- the estimated correlation between sex and income for persons living in Sweden at the end of 2001

Complex statistical characteristic

A generalised statistical characteristic where

- the single *population* of a simple statistical characteristic is generalised into a *structured set of related populations*; the structuring is typically a *cross-classification* of a main population into a hiearchy of *subpopulations*

- the single *parameter* of a simple statistical characteristic is generalised into a *vector of parameters*; all the parameters in the vector are supposed to be relevant for all (or at least an important subset of) the subpopulations

- the *reference time* is treated as a separate dimension, which may assume a vector of values on a time scale, typically a series of years and/or months, etc

Statistics	Number of foreign citizens living in Sweden and their average yearly incomes by citizenship, region, sex, and age. Years 1996-2003.
Population	foreign citizens living in Sweden at the end of year y
Population-defining properties	1. country of residence = Sweden 2. citizenship = non-Swedish
Counted object	person
Related objects	
Subpopulations	crossclassification of population objects
Classification variables and value sets	1. citizenship (country code) 2. region (county.municipality) 3. sex (male, female) 4. age (five year age groups)
Parameters	1. number (of objects in the population) 2. average (income)
Reference times	year (1996, ..., 2003)

Example 1: Analysis of the contents and structure of a statistical message

Statistics	Number of working persons in Sweden, 16 years of age and older, living in the region (night population) by region of dwelling, region of work, sex, occupation, socio-economic status, income class, activity class of working place. Year 1990.
Population	working persons, 16+ years old, living in Sweden at the time of the reference time, t
Population-defining properties	1. country of residence = Sweden 2. working status = working 3. age > 15 years
Counted object	person
Related objects	dwelling (where person lives) establishments (where person work)
Subpopulations	crossclassification of population objects
Classification variables and value sets	1. region of dwelling (country.municipality) 2. region of work (county.municipality) 3. sex (male, female) 4. occupation (ISCO) 5. socio-economic status (SEI82) 6. activity class of working place (NACE)
Parameters	number (of objects in the population)
Reference times	time of the population census 1990

Example 2: Analysis of the contents and structure of a statistical message

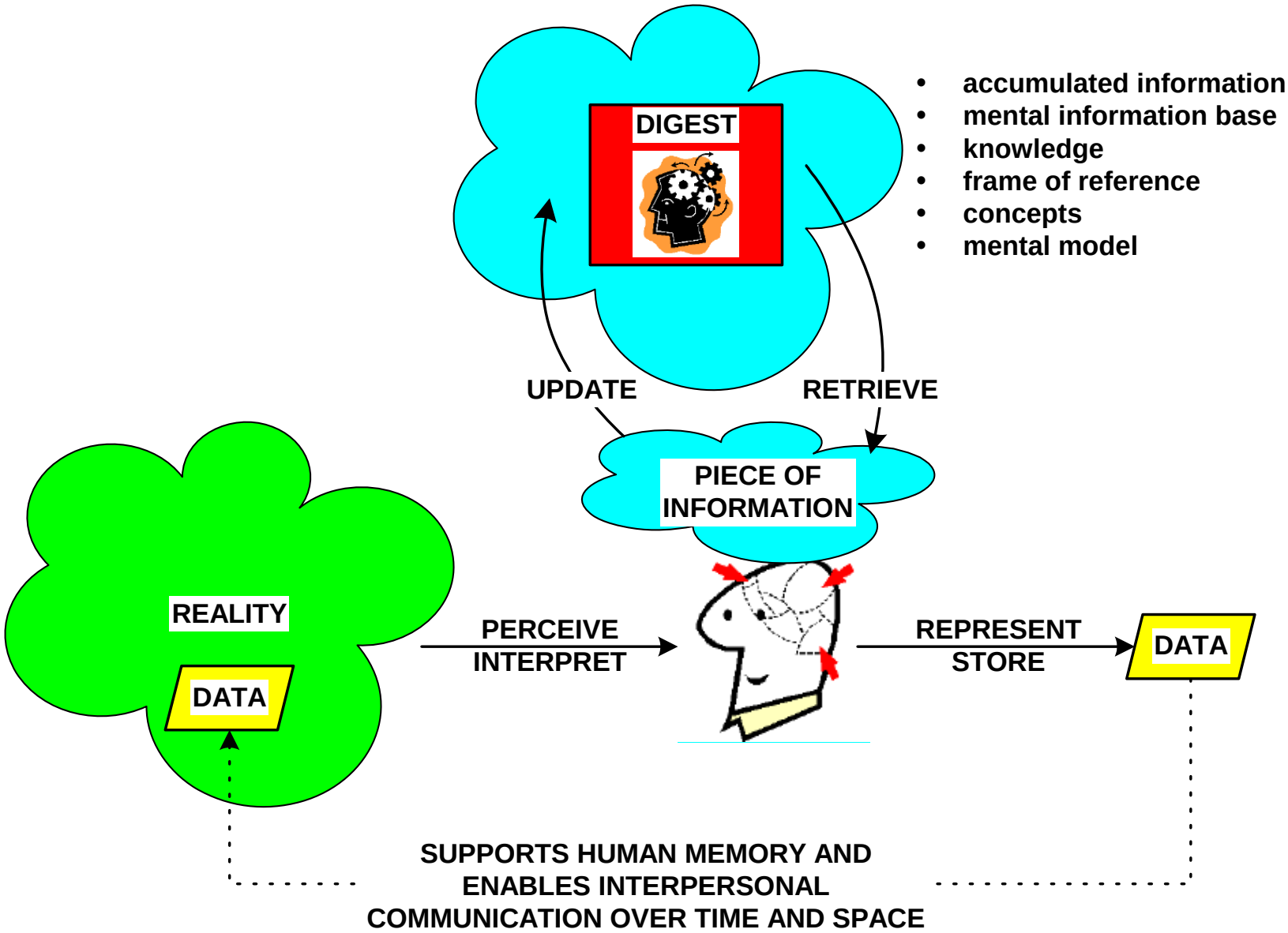
Statistics	Number of migrations in Sweden by sex, age, from_region and to_region. Years 1998-2003.
Population	migration events concerning persons living in Sweden (before and/or after the event) that have taken place during the reference year, y
Population-defining properties	1. country of residence of person = Sweden
Counted object	migration event
Related objects	person who migrates, dwelling from which the person migrates, dwelling to which the person migrates
Subpopulations	crossclassification of all migration events by sex of the migrating person, region of the dwelling from which the person migrates, and region
Classification variables and value sets	1. sex of the migrating person 2. age of the migrating person 3. region of dwelling from which the person migrates 4. region of dwelling to which the person migrates
Parameters	number (of objects in the population/subpopulation)
Reference times	years 1998-2003

Example 3: Analysis of the contents and structure of a statistical message

Information

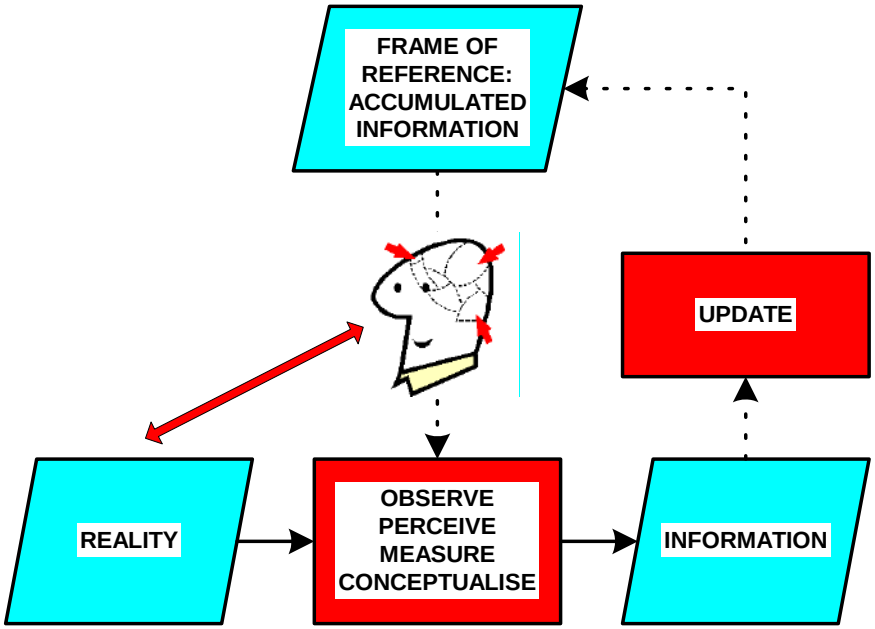
- Information may be the result of a human being's attempt to interpret data, an interpretation process.
- Information may also be the result of a perception process, where a human being perceives the external world by means of her senses.

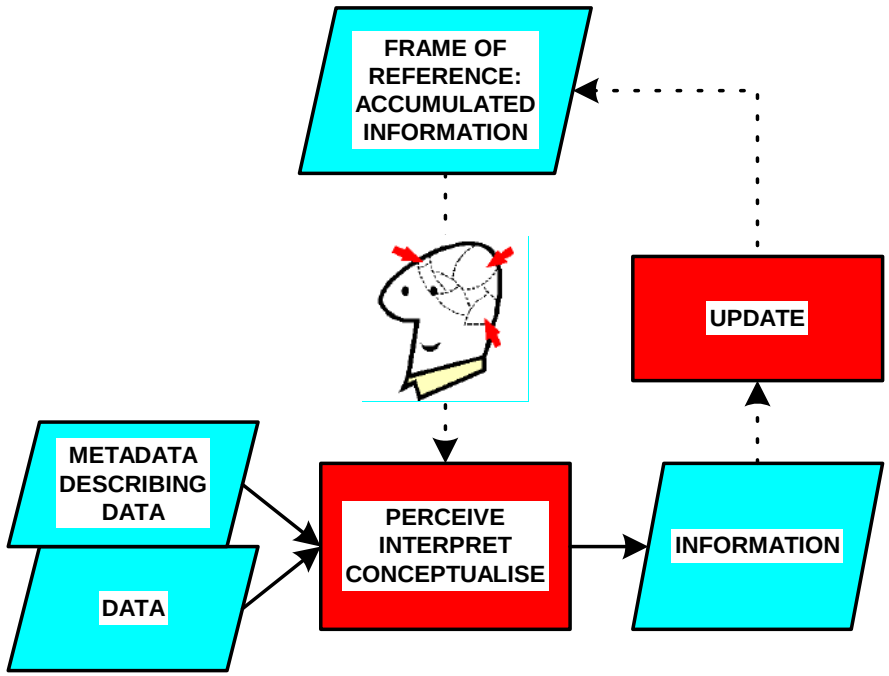
Reality – information – data



- accumulated information
- mental information base
- knowledge
- frame of reference
- concepts
- mental model

Frame of reference - Metadata





The infological equation (Langefors)

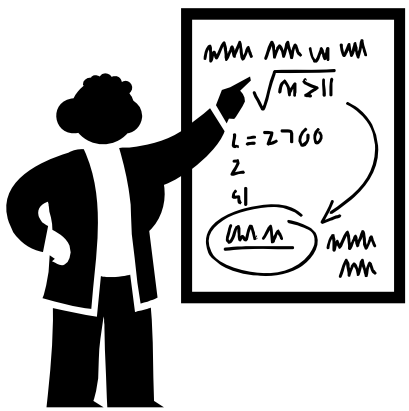
- I = i(D, S, t)

where

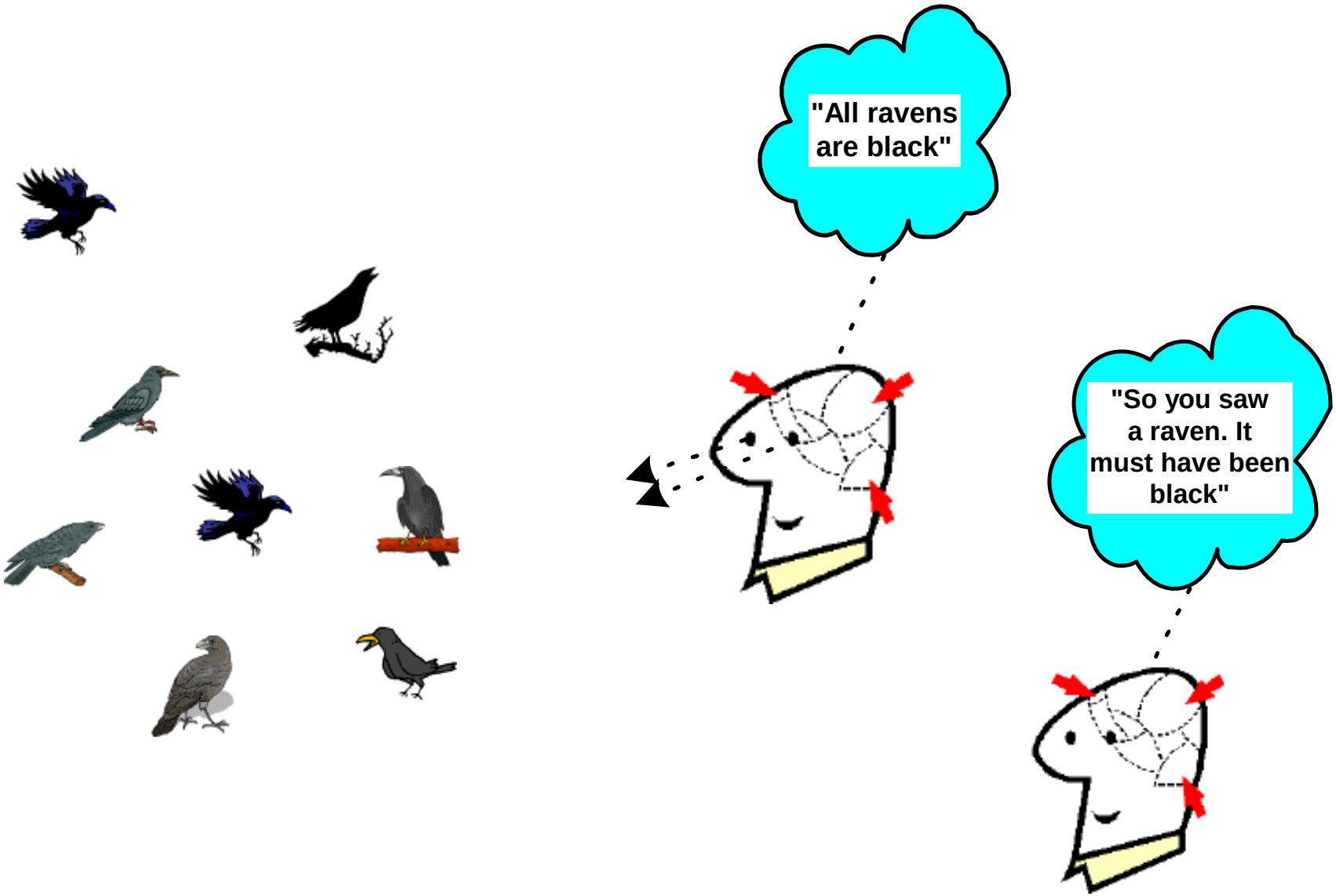
- I is the information contents obtained by a human being
- i is the process of interpretation and creation of meaning
- D is the received data
- S is the frame of reference, or accumulated knowledge, used by the interpretor
- t is the time used for interpretation

Information and knowledge

- Facts and rules (laws)
- Specific information and general information
- Information → Knowledge → Wisdom



Knowledge formation: induction and deduction

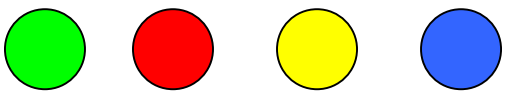


What is a concept and how is it formed?

- Categories like object types, e.g. “person”, “dog”, and property types, e.g. “age”, “colour”



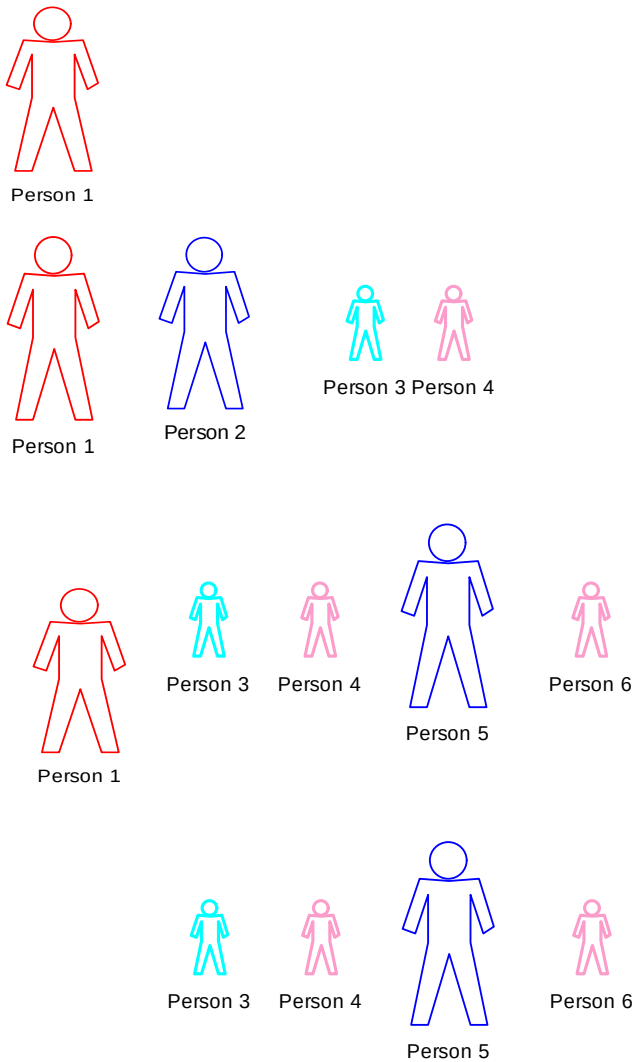
- We recognise dogs as dogs when we see them. Why?
- Dogs have many things in common: head with nose, two ears, two eyes, mouth, body with four legs, tail; a dog barks now and then, ...
- But what if a dog has only three legs? Could be an exception because of accident or disease, ...



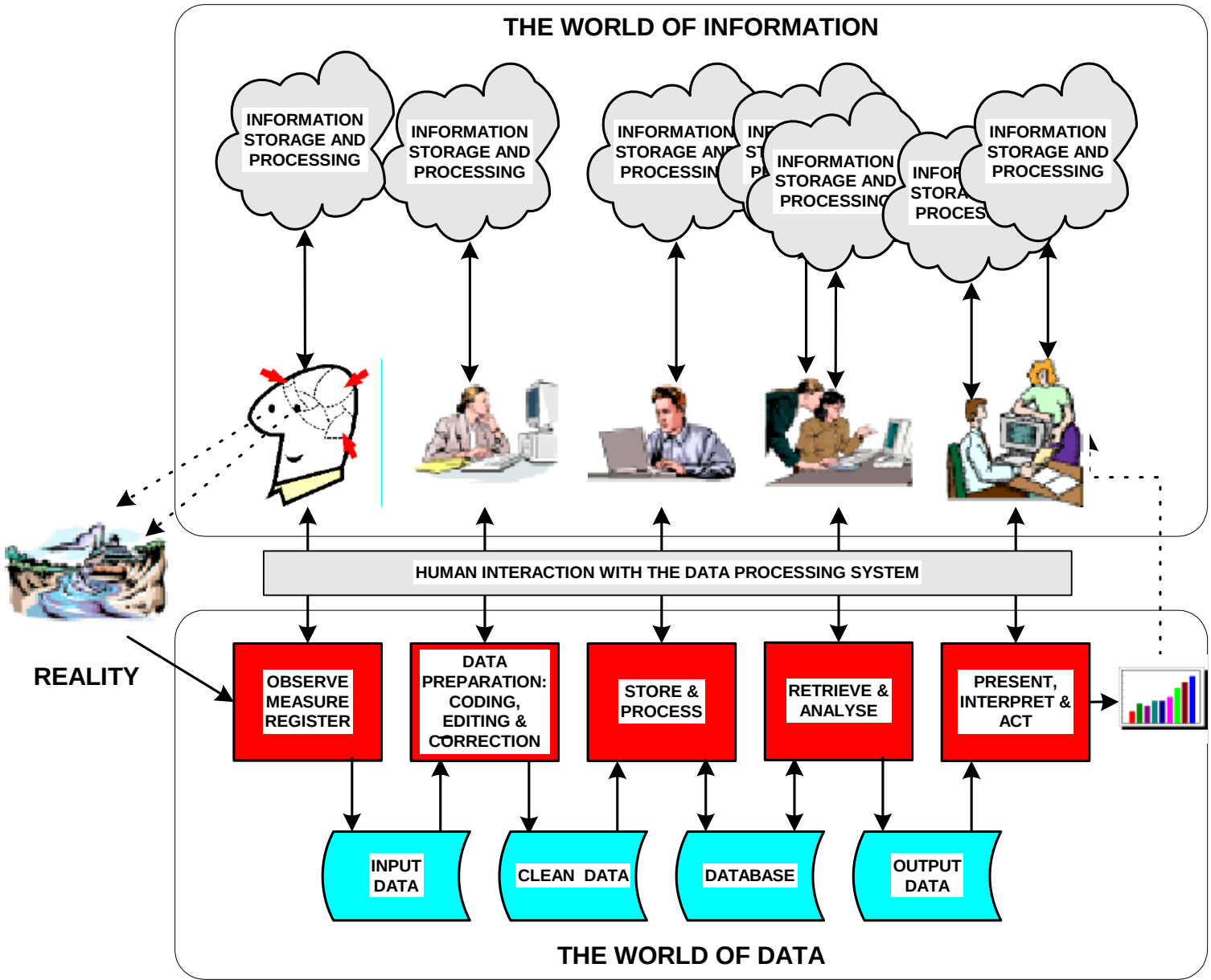



- We recognise colours when we see them.
- Concepts may be combined into new concepts:
- The colour of a car.

Defining in space and time: What is a household?



Information & data processing system



Information systems and statistical information systems

Information systems

- capture data intended to represent, directly or indirectly, information stored in human minds
- store and process data with or without help of computers or other technical tools
- produce and communicate data that are interpreted into information by human minds

Statistical information systems

- capture, store, process, and communicate statistical data

Interpretation of data and communication of information

We can never be sure that

- different people interpret the same data in the same way
- a receiver of data interprets the data as intended by the sender

Operative and directive/analytical information systems: typical tasks

OPERATIVE INFORMATION SYSTEMS	*DIRECTIVE (ANALYTICAL) INFORMATION SYSTEMS*
Automating or supporting manual processes; often repetitive	Supporting planning and control processes; often non-repetitive
Supporting repetitive processes within a business function	Supporting decision-making ad hoc
Taking note of regular events (transactions, operative decisions); routine	Supporting strategic decisions; non-routine
Supporting a business process initiated by a customer until it is completed	Supporting analytical activities, e.g. research and development

Operative and directive information systems: typical properties

OPERATIVE INFORMATION SYSTEMS	*DIRECTIVE INFORMATION SYSTEMS*
Users and usages known at systems development time	Users and usages partially unknown
Information necessary for operative processes; must be provided despite costs	Information improving the quality of directive processes; there is a trade-off between value and cost
Repetitive usage	Ad hoc usage
Data collection well planned and an integral part of the system	Combine available data from different sources
Strong connections between collection and use of data	Data are used for new purposes
Users know the meaning and quality of data relatively well	Metainformation has an important role

Part I: Extra material

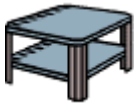
Bo Sundgren

2010

Concept formation: Examples

- **HOUSEHOLD**
  - What is a household?
  - When is a household born, and how does it die?
  - Which changes can a household undergo and still remain the same household?
- **COMPANY**
  - What is a company?
  - Which changes can a company undergo without ceasing to be the same company?
- **CUSTOMER**
  - What is a (potential, actual, former) customer?
  - When does a person become a customer, and when does he or she cease to be a customer?

Concept formation: ”What is a table?”

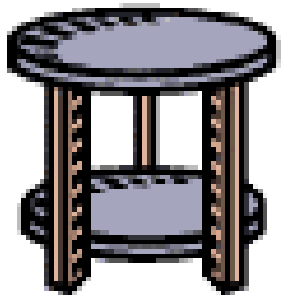
Dialogue between Robinson and Friday when they meet in Robinson’s home in England.

- ***Friday***: What is a table?
- ***Robinson***: A table is a sqare board with four legs.
- ***Friday*** (*pointing to a chair*): I understand, that is a table.
- ***Robinson***: No, that is not a table. A table does not have a back.

”What is a table?” (2)

- ***Friday*** (*pointing to a chest of drawers*): Well, but then that is a table, because it has a board, four legs, and no back.
- ***Robinson***: No, it is a chest of drawers. Tables do not have drawers under them.
- ***Friday***: I understand. A table is a sqare board with four legs, without a back, and without drawers under it.

”What is a table?” (3)



- ***Friday***: Then (*pointing to Robinson’s round dining-room table with three legs*) obviously that cannot be a table, since it is round and has only three legs.
- ***Robinson***: Actually it is. The board does not have to be square, and the number of legs does not have to be exactly four.
- ***Friday***: How many legs are necessary?
- ***Robinson***: There must be at least three legs, because otherwize the table cannot stand. But there is no upper limit. It depends on how big the table is. In practice I don’t think there are tables with more than ten legs.

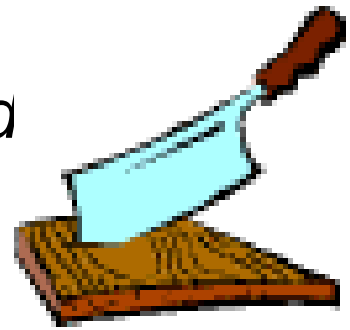
”What is a table?” (4)

- ***Friday*** (*pointing to a table in Robinson’s garden that is made from a round millstone, resting on a thick piece of cement in the middle*): Thus that cannot be a table, because it has only one leg.
- ***Robinson***: Yes, it is a table! I was wrong, when I said that a table has have at least three legs. But it cannot have less than one leg.

- ***Friday*** (*getting sight of a drop-leaf table sitting on the wall*): Then clearly that is not a table, since it has no legs at all.
- ***Robinson***: Indeed that is a table, too. It seems that a table does not have to have any legs. A board is sufficient.

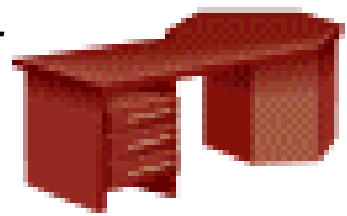
”What is a table?” (5)

Robinson brings forward a cutting-board in ord to cut some pieces of cake for the coffee.

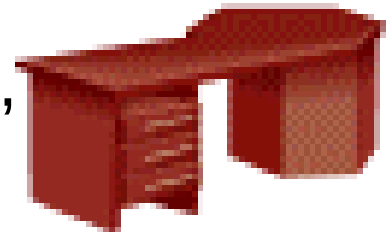


- ***Friday*** *(watching the cutting-board):* Is that a table?

Robinson becomes uncertain and starts muttering something. He quickly brings Friday to the sleeping-room, where he knows for sure that there are no strange tables hanging on the wall or having no legs. But he has forgotten his writing-table that he has inherited from his father a heavy piece of furniture with a cabinet under it on one side, and a set of drawers on the other.

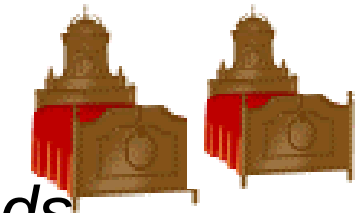


”What is a table?” (6)

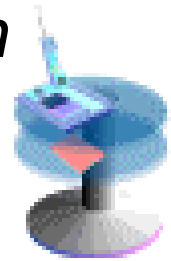


- ***Friday*** *(smiling):* That is certainly not a table, because it has drawers under it.
- ***Robinson****:* Hm, it is a table. It is called a writing-table, and I use it when I sit working. It can be distinguished from a chest of drawers by the fact that it has a free space in the middle.
- **Friday**: Then, if I have got it right, a table is a board, without a back, and with free space below.
- ***Robinson*** *(with a sigh of relief):* Exactly, I could not have put it better myself!

”What is a table?” (7)

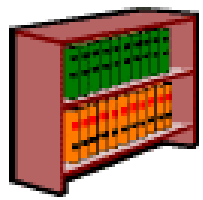


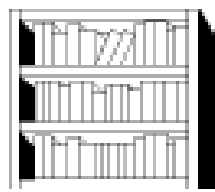
Robinson suddenly comes to think of the beds that have some free space under them, and brings Friday quickly out of the bedroom, before he starts asking about them... They enter the library, where Friday discovers a round table with two small boards below each other.



- ***Friday**:* That cannot be a table, because it has two boards. If it is a table, all those things must also be tables.

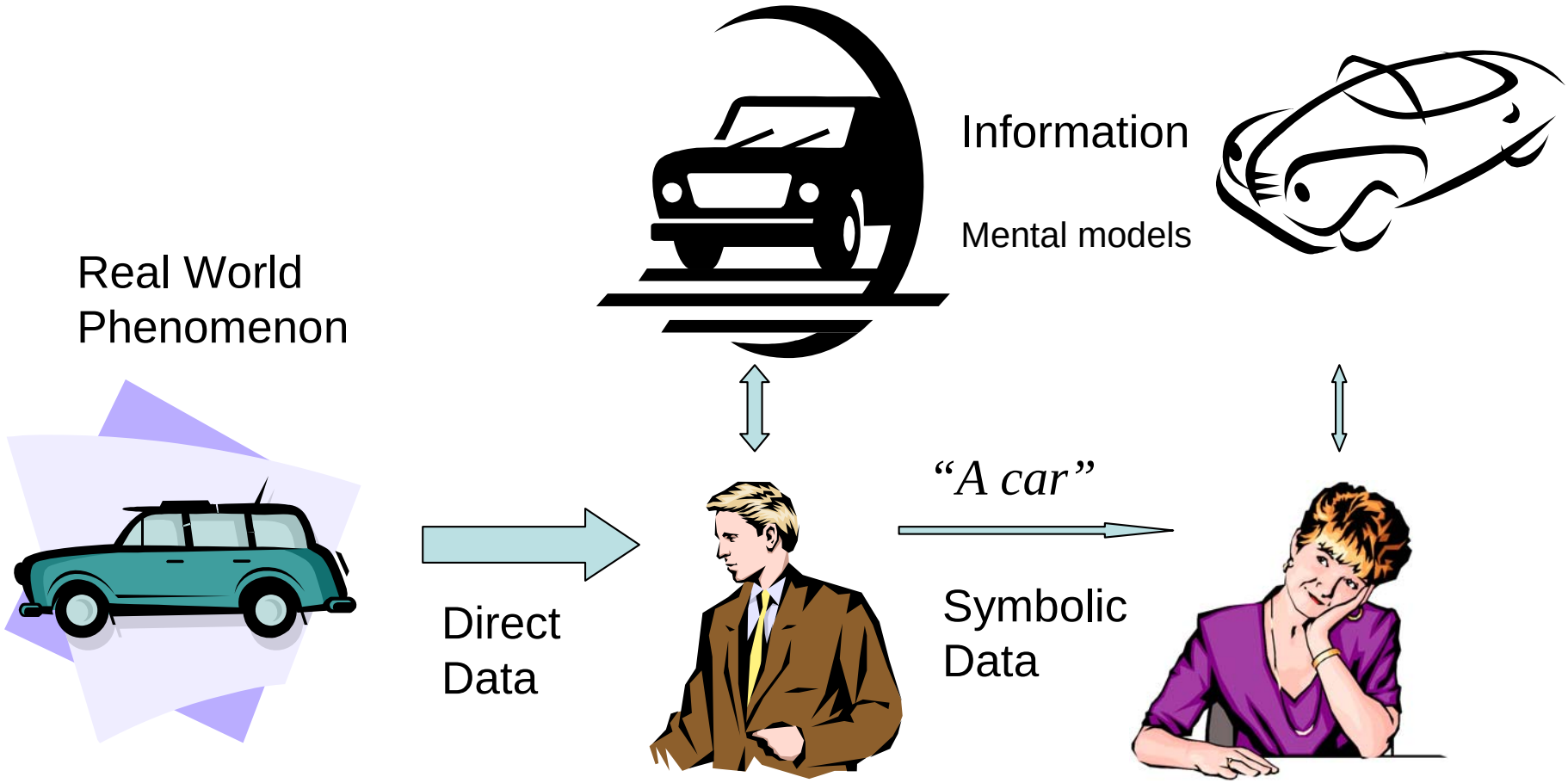
  He says, pointing to all the book-shelves. Robinson, who has so far had an angle’s patience, now loses all this, and cries out something that is not suitable to reproduce.





(Adapted from Per Flensburg: “Personal computing”, doctoral thesis, Lund 1986.)

Data and information



Direct and direct communication