



Exploiting Data for Risk Evaluation

Niel Hens

Center for Statistics

Hasselt University

Sci Com Workshop 2007



Overview

- Introduction
- Database management systems
 - Design
 - Normalization
 - Queries
- Using different databases in risk evaluation
 - Merging databases
 - Using different databases in analyses
- From databases to analyses
 - Issues
 - Modern statistical methods
- Discussion



Risk Evaluation

1. Hazard identification
2. Dose-response assessment
3. Exposure assessment
4. Risk characterization



Introduction

- Diversity
 - Information
 - Techniques
- Interdisciplinary
 - Chemistry
 - Epidemiology
 - Management
 - Medicine
 - Microbiology
 - Sociology
 - Statistics
 - Toxicology
 - ...



Introduction

- Data from studies stored in databases
 - Large databases
 - Numerous variables
- Data base management system (DBMS)
- Data exploitation
- Analyses
 - Estimating parameters
 - Mathematical models
 - Deterministic
 - Stochastic



DBMS-advantages

- Minimal data redundancy.
- Data consistency.
- Integration of data.
- Sharing of data.
- Enforcement of standards.
- Ease of application development.
- Uniform security, privacy and integrity.
- Data independence.



DBMS-examples

- Oracle
- Ingres
- Informix (Unix)
- DB2, SQL/DS (IBM)
- Access (Microsoft)
- SQL Server (Microsoft +)
- Many older (Focus, IMS, ...)
- Many limited PC (dBASE, Paradox, ...)

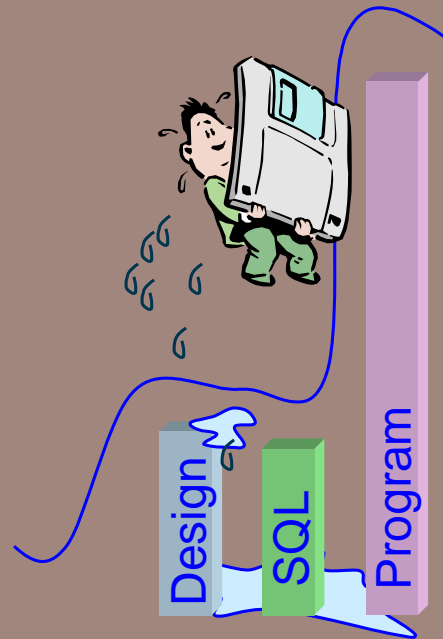


DBMS-structures

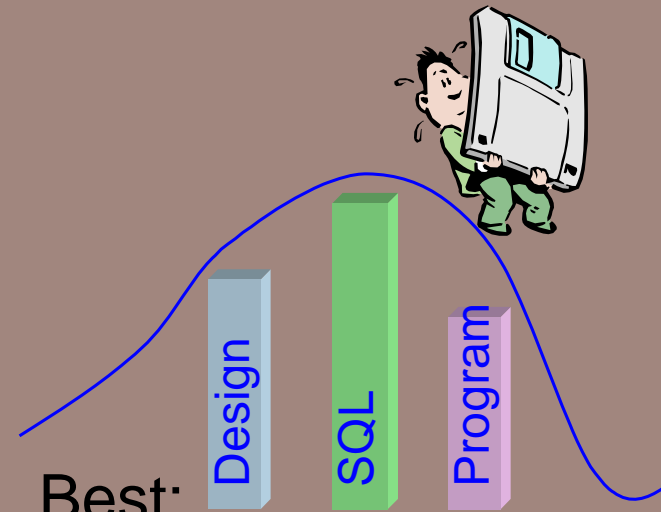
- Hierarchical Databases
- Network Databases
- Relational Databases
- Object-Oriented Databases



DBMS-Design



Worst:
Compensate for poor design
and limited SQL with programming.



Best:
Spend your time on
design and SQL



DBMS-Design

1. Identify the exact goals of the system.
2. Talk with the users to identify the basic forms and reports.
3. Identify the data items to be stored.
4. Design the classes (tables) and relationships.
5. Identify any 'business' constraints.
6. Verify the design matches the 'business' rules.



DBMS-Normalization

Ensures minimal data redundancy

1. Start from the overall data coming from different sources
2. Normalize the table according to specific rules
 1. 1st – 2nd – 3rd – 4th normal form
 2. Hidden Dependencies (Boyce Codd – etc)
3. Referential integrity



DBMS-Queries

Normalization: splitting databases in tables

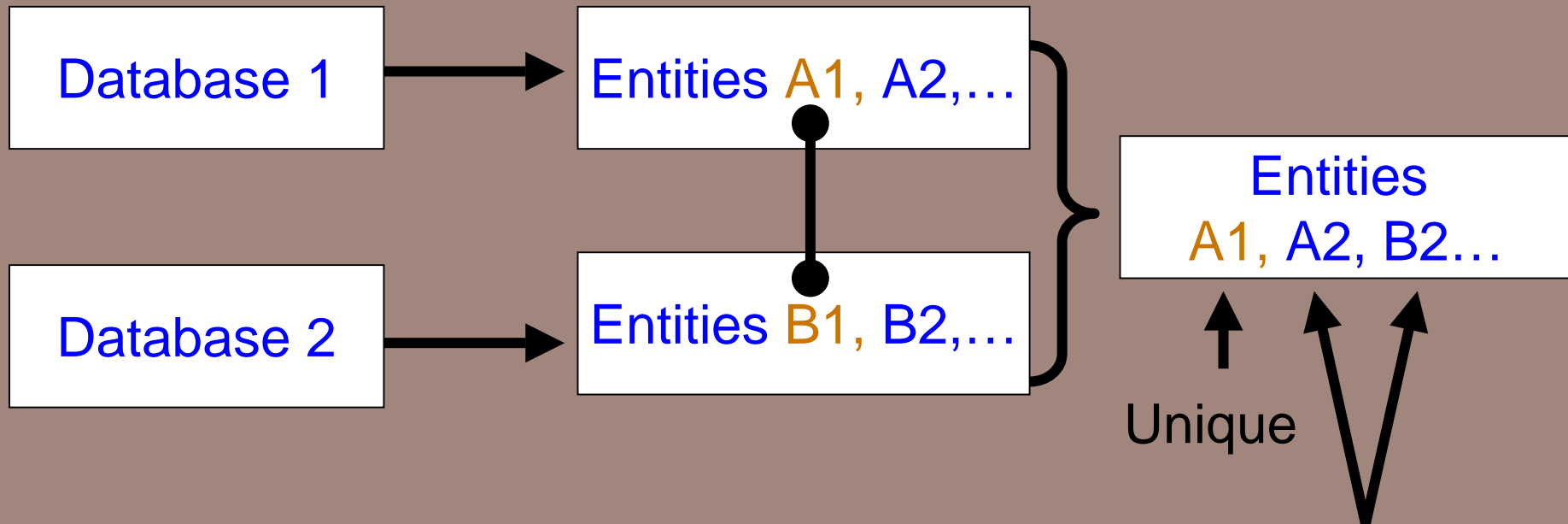


Queries: merging tables to answer specific questions
e.g. SQL, QBE, ...



Merging Databases

Similar & Different entities



Need for data dictionaries !!!
Need for database documentation !!!



Merging Databases

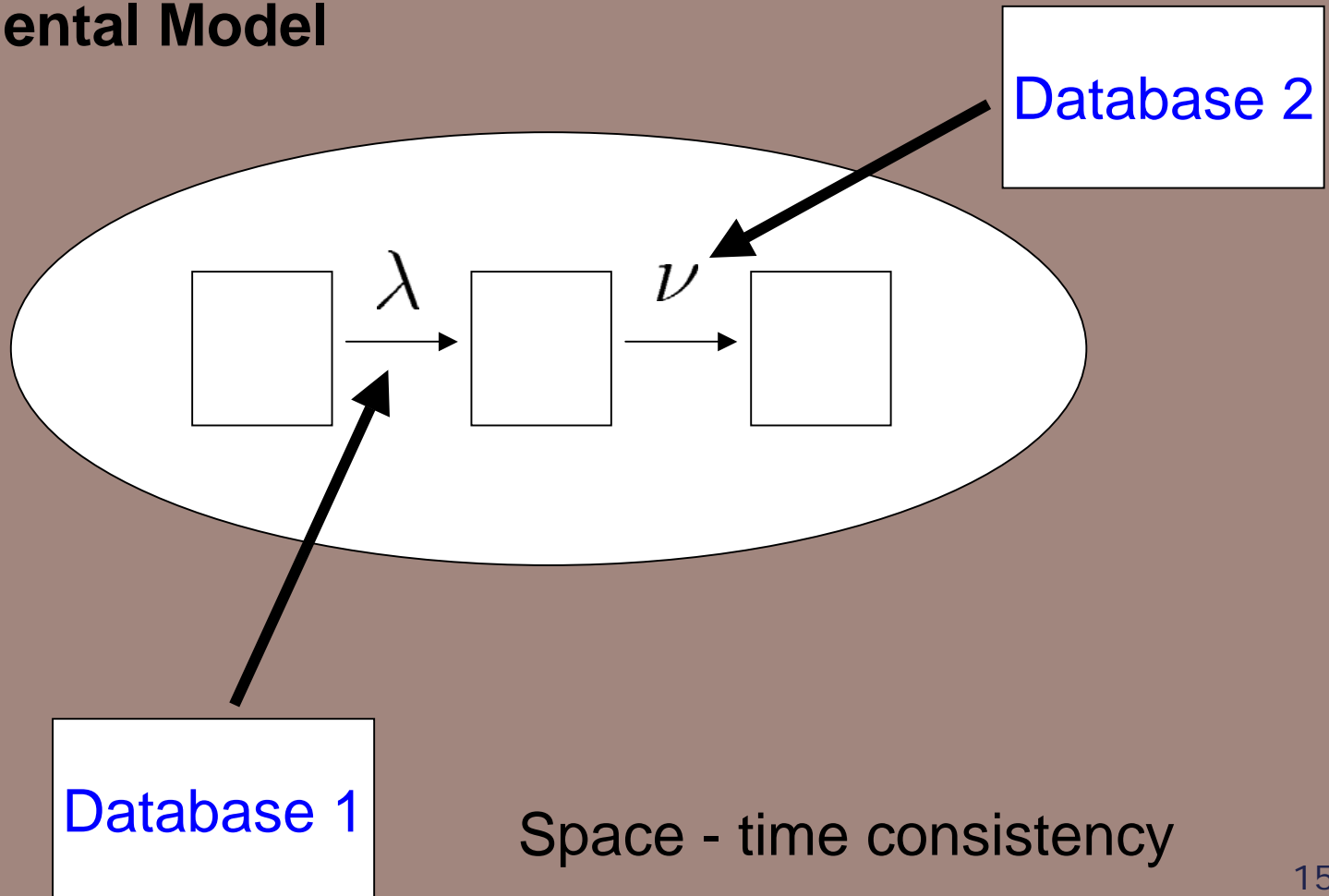
- Extended coding
- Renormalization
- Re-evaluation of queries
- Consistency checks
- ...

Common example: date (formats)



Consistency across database usage

Example: Compartmental Model





De-linked databases

- Lost link?
 - pragmatic solutions
 - increased uncertainty
- Back tracing



From Databases to Statistics



- Fragmented information
- Need queries to obtain information
- Storage possibilities
- ...

- Rectangular Data Format
- Limited Storage
- ...

Compatible ?



From Databases to Statistics

Therefore:

- Transition ideally done only once
- Spend time on developing queries
- Assess the information needed in the analysis
- Limit data manipulation in the statistical program
- Document all steps

Common problems

- Identification of risk factors
- Handling of missing data
- ...



Statistics to the rescue ?

YES:

- Data Mining
- Incomplete Data
- Diagnostic Uncertainty
- Detection Limits
- Bayesian Framework – MCMC
- ...

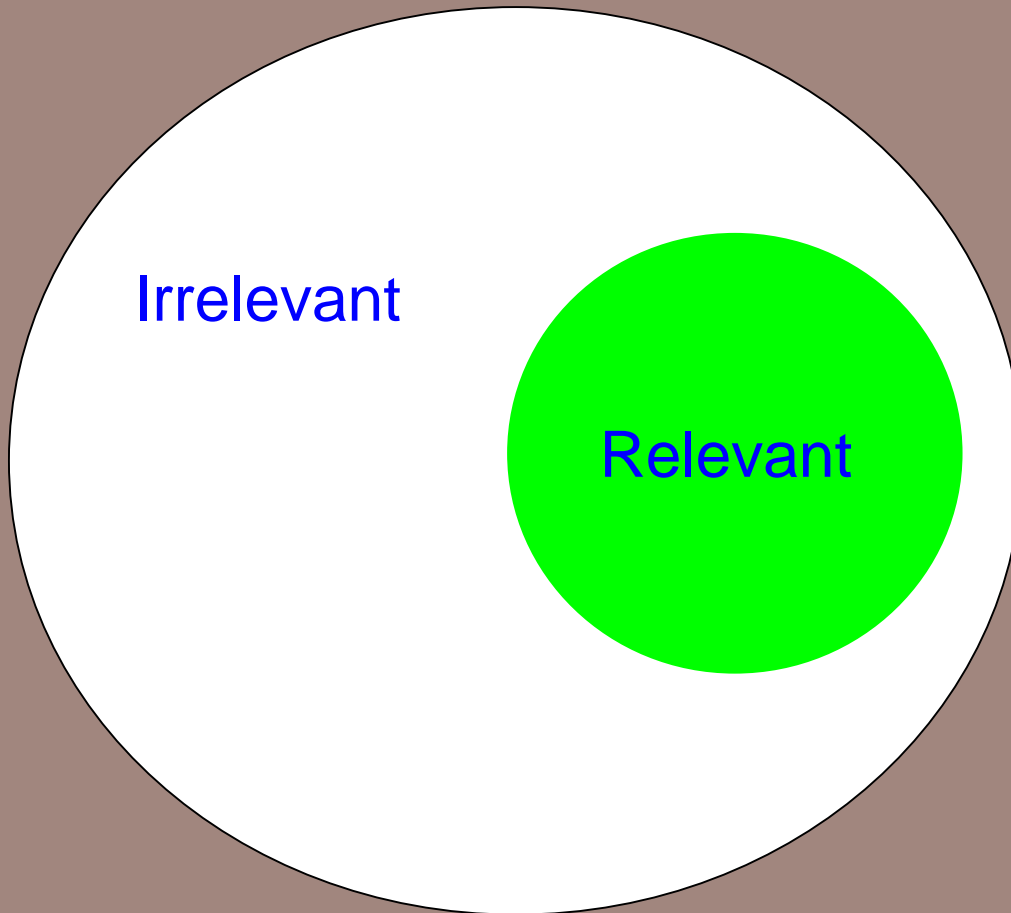
NO: Limitations

- Assumptions
- No general solutions



Data Mining

Identifying risk factors



Data mining is the principle of sorting through **large amounts of data** and picking out **relevant information**.

It is **increasingly used** in the **sciences** to extract information from the enormous data sets generated by modern experimental and observational methods.



Incomplete Data

- Coding
- Different software – different coding
 - 'NA' – 'NULL' – '.' – '9' – '99' - ...
 - Combination of coding
 - DANGEROUS !!!
 - Continuous variables: misinterpretation
 - Discrete variables:
 - Nominal: new category
 - Ordinal: no ordering possible
- Investigate missingness distribution prior to analyses



Incomplete Data

- Incomplete data
 - Mostly completely lost
 - Inevitably loss of efficiency
 - Bias? → not always
- Modeling incomplete data
 - Rubin (1978)
 - MCAR, MAR and MNAR
 - MAR a reasonable assumption?
Several analyses possible!
 - Sensitivity analyses



Detection limits

- Truncated distribution
- Ignorance \rightarrow bias
- Censoring
 - True response Y

– Observed response

$$Y_c = \begin{cases} Y & \text{if } Y > c \\ c & \text{if } Y \leq c \end{cases}$$

– Account for detection limit



Diagnostic Uncertainty

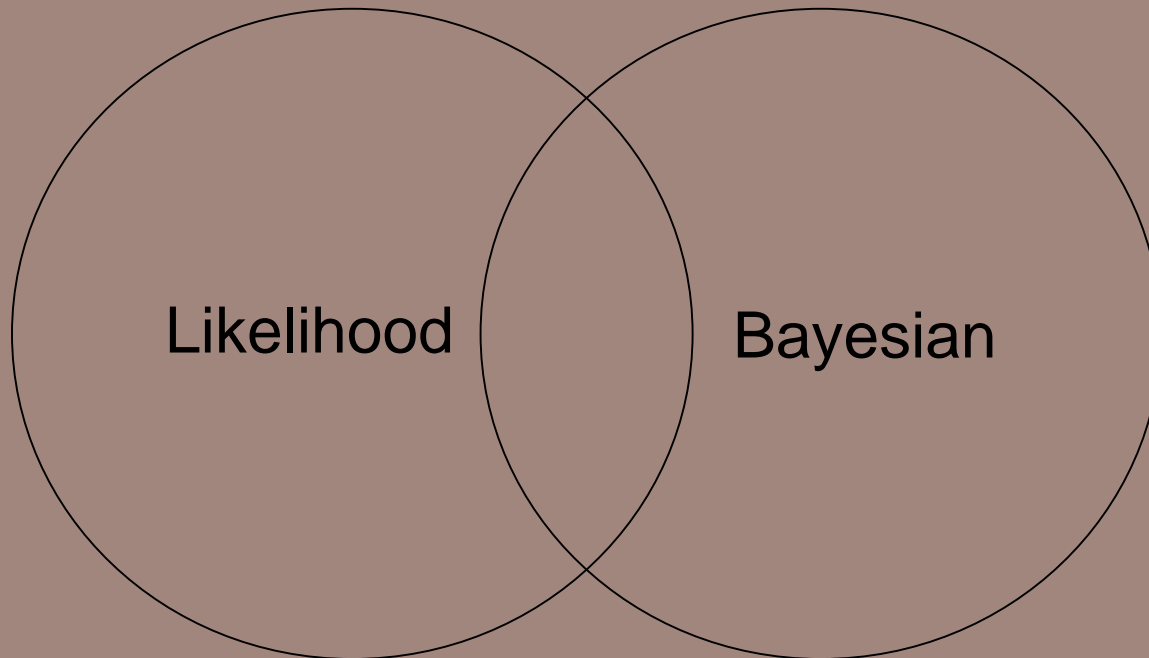
1. True \leftrightarrow Apparent prevalence
Correction possible:
using sensitivity and specificity

2. Dichotomization
Depends on cut-off values
Mixture modeling, no cut-off needed



The Bayesian paradigm

Two opposite worlds?



Two complementary worlds!



The Bayesian paradigm

- Likelihood: data
- Bayesian:
 - prior x likelihood = posterior
 - Allows to use prior information
 - Sensitivity analysis on prior
 - Link to
 - Multilevel modeling
 - MCMC methods



Discussion

- Exploiting databases
 - Available data
 - Well-structured databases
- Merging information
 - If foreseen – relatively easy
 - Not foreseen – increased uncertainty



Discussion

- Ignorance = invalid results
- Modern statistical methods
 - do exist
 - don't solve everything



“Know where to find the information
and how to use it.
That's the secret of success.”

Albert Einstein