

Longitudinal analysis using warehouse techniques

M.H.J. Vucsan

J. Kardaun

Statistics Netherlands



Contents

The problem

The population register

New storage solutions

- Warehouse explained
- OLAP tools explained

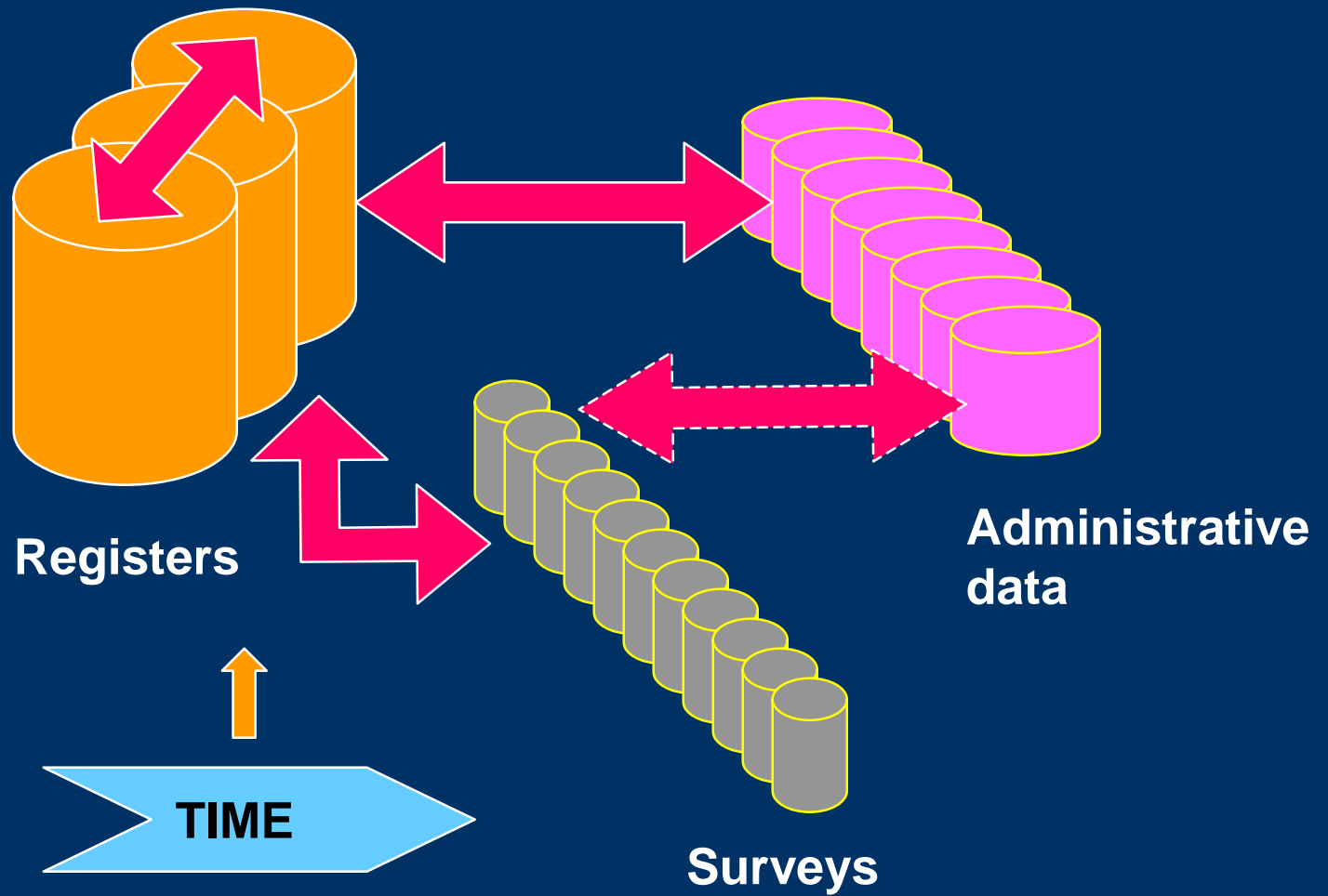
Longitudinal use of the warehouse

- Longitudinal data mart
- Simple longitudinal analysis

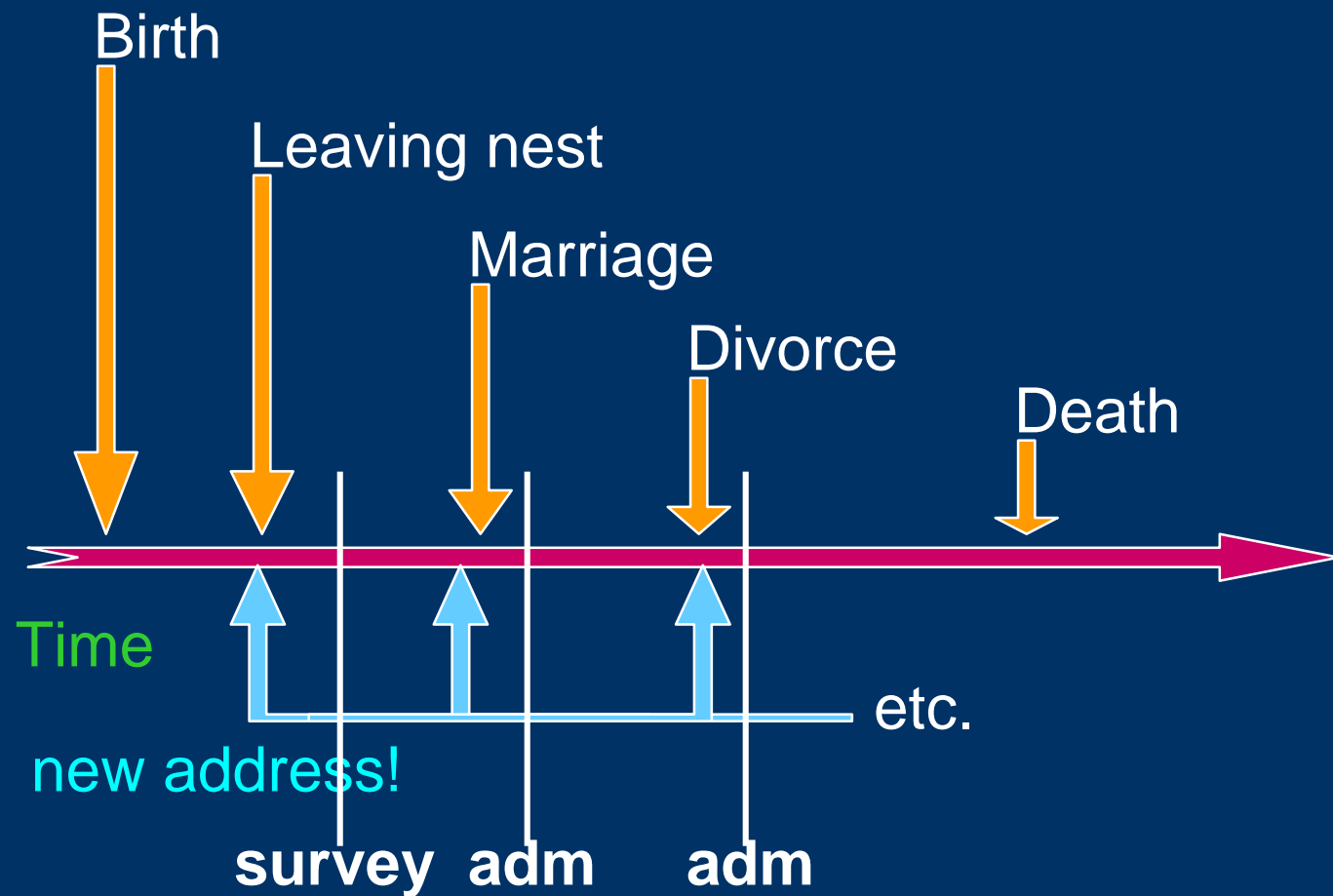
Conclusion



The Problem (1)



The Problem (2)



The Population register

- **Yearly snapshot**
 - All municipalities submit snapshot
 - Dates vary; January to February
- **Continuous updates**
 - Automatic reception
 - Coupled to inter municipal network
- **Statistics Netherlands**
 - Keeps obsolete copy (few months)
 - Integrates snapshots and updates



(New) Storage Solutions

Data Warehouse



The rise of BI (0)



The rise of BI (1)

- Analysis of point-of-sale data
- Analysis of real time sales
 - Television
 - web stores
- Analysts not proficient with programming tools
- Explorative type of analysis, output unknown to programmers



The rise of BI (2)

- **FILE: (200 mil x 4k = 800Gb)**

- Sold articles to person at moment in time

- **Attributes**

- Date, customer-id, name, number, address, amount, salesperson, invoice#, article#, article name etc. (4kb per record)

This file tells you all there is to know about the sales side of this business



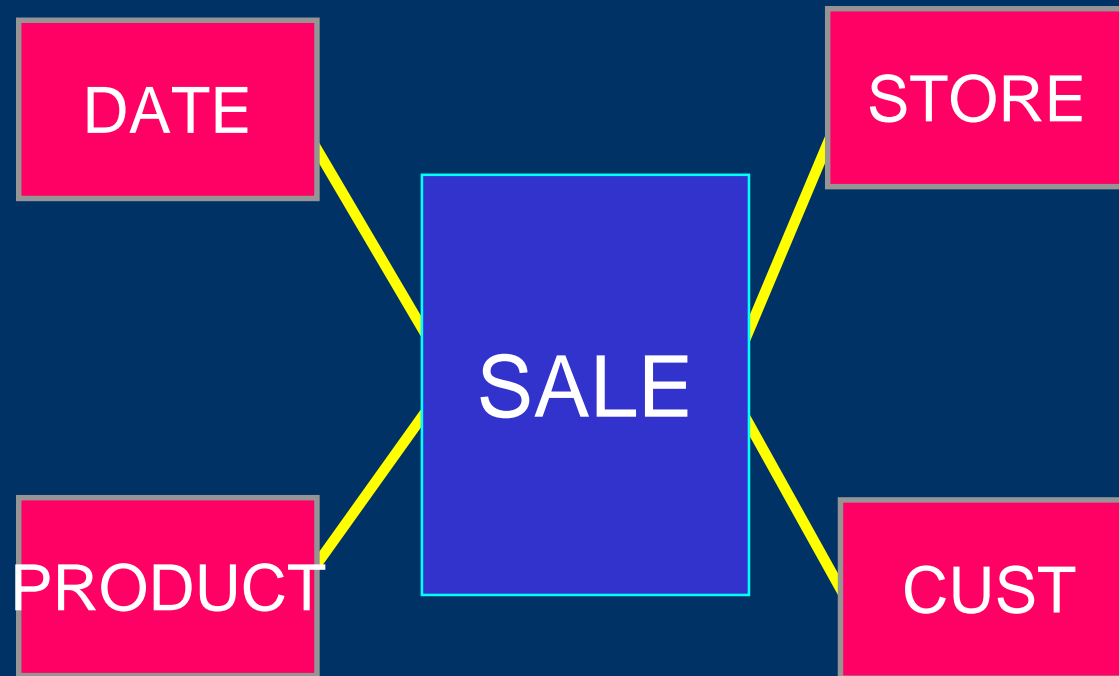
The rise of BI (3)

- **dissociate the file:**

- Customer attributes ($20.000 \times 1k = 20 \text{ Mb}$)
- Salesperson attributes ($100 \times 1k = 100 \text{ Kb}$)
- Location attributes ($6 \text{ mil} \times 1k = 6 \text{ Gb}$)
- Date attributes ($1000 \times 1k = 1\text{Mb}$)
- Above keys + amount ($200 \text{ mil} \times 25b = 5\text{Gb}$)



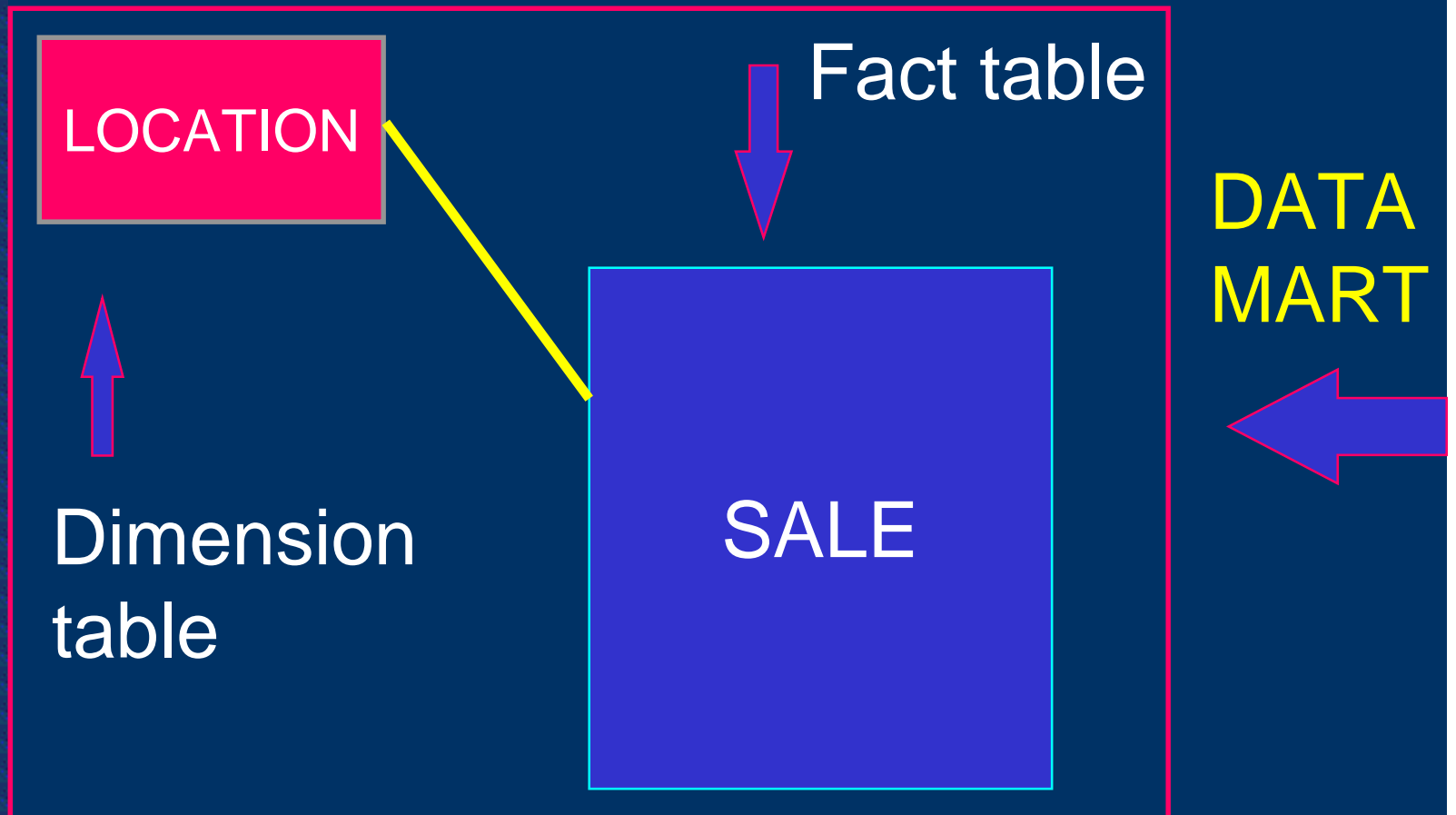
The rise of BI (5)



Data Mart



The Dimensional Model



(New) Storage Solutions

OLAP Engines



Olap Engines (1)

- **Main functions:**
 - Performance increase
 - Access control
 - Presentation layer
- **Most important subsystem:**
 - Pre aggregation engine



OLAP Engines (2)

- **How Aggregates work:**

- **Reduce cardinality in the dimensions and use hierarchic structure:**
 - Create an aggregate without person-ID
 - Create an aggregate without municipality
- **Direct:** summing over province uses the aggregate on province and the combined persons
- **Indirect:** summing over country uses aggregate on province etc. (automatic)



OLAP engines (3)

	12	600	1000
<i>key</i>	<i>province</i>	<i>municip.</i>	<i>hamlet</i>
00234	limburg	sittard	ondermergel
00233	limburg	sittard	urmond
00288	limburg	geleen	munstergeleen

Reduction of records to be summed: 1000x max



Automatic aggregates

Data Storage and Aggregation Wizard

Set storage size and performance options

You can set options for your query performance or aggregation storage. Setting a high performance gain improves query speeds by building more aggregations, but requires increased storage.

Add aggregations until:

- ☐ Estimated storage reaches GB
- ☒ Performance gain reaches %
- ☐ I click Stop

For more information on these options, click Help.

Click Start to design the aggregations.

Performance vs. Size

Size (MB)	Performance (%)	Description
0	0	Start
1.5	35	Aggregations answer few queries
5.5	85	Aggregations answer most queries



Longitudinal Analysis

NOT looking at events in a certain period

BUT

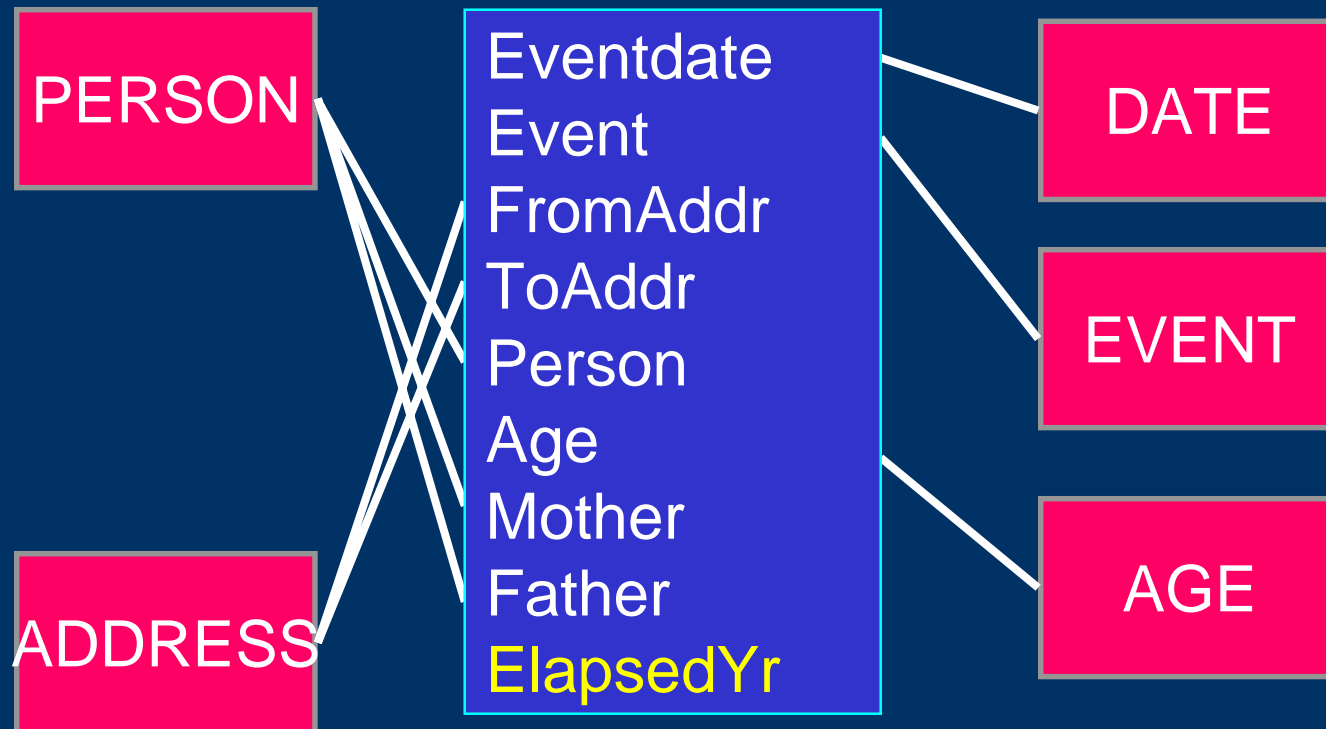
- looking at a sequence of events (at least 2)
- preferably over a long, long period
- preferably considering the “population” from which the events arise
- choosing a relevant time origin and scale

SO:

- not trends of averages/ distributions
- but averages/distributions of trends



Data mart MIGRATION



simplified



Near-Longitudinal Analysis

Mean value of duration, compare different groups:

- For all movers to Amsterdam last year; time at previous address, grouped by age etc.
- For all movers in the Netherlands last year; time at previous address

Mean value of duration, compare different eras

- For all movers to Amsterdam last year; time at previous address, grouped by age etc.
- For all movers to Amsterdam 5 years ago; time at previous address, grouped by age etc.



Time \neq Time in L.A.

Time measure as

- Calender date (Period)
- Age
- Time since event (marriage, migration, childbirth, widowing)

Central limitation

- Age, Period, Cohort are dependent



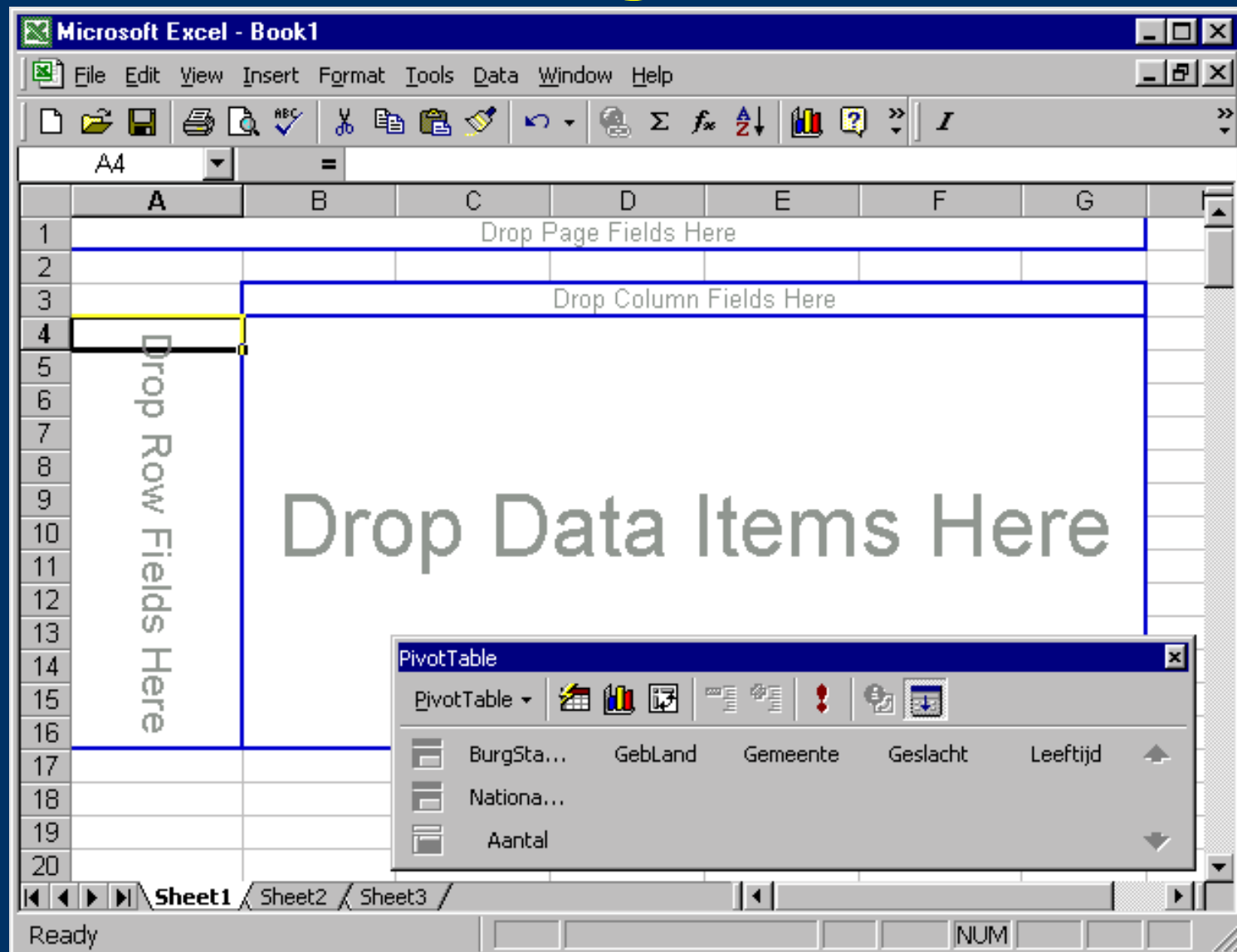
Longitudinal analysis (2)

Pitfalls:

- **Need an accurate description of the population at risk**
 - Example: left/right handedness
 - time of residence before migration \neq time of residence of movers
- **Beware of censoring problems**
 - Example: The graveyard fallacy
- **And other periods of non-observation**
 - Example: Giving birth while expatriate



Interactive usage



Interactive usage

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

A4 = Werelddeel

	A	B	C	D	E	F	G
1	Drop Page Fields Here						
2							
3	Aantal	Bsgroep1					
4	Werelddeel	Gehuwd	Gescheiden	Ongehuwd	Weduwstaat	Grand Total *	
5	Afrika	119180	20622	102342	2380	244524	
6	Azie	280237	41059	114971	26270	462537	
7	Europa	6493339	694622	6483732	841110	14512803	
8	Midden-Oosten	18149	2200	31578	482	52409	
9	Noord-Centraal-Amerika	12939	2565	19091	468	35063	
10	Oceanie	5303	866	5764	91	12024	
11	Zuid-Amerika	87517	48483	132305	6237	274542	
12	Grand Total *	7016664	810417	6889783	877038	15593902	
13							
14							
15							
16							
17							
18							
19							
20							

PivotTable

PivotTable

BurgSta... GebLand Gemeente Geslacht Leeftijd

Nationa...

Aantal

Sheet1 Sheet2 Sheet3

Ready

NUM



Conclusions (1)

Use of OLAP and Warehouse techniques will speed up analysis considerably

- **Quicker than sequential files**
- **Variables all available and linked**
 - Especially over time
- **Integrated administration and meta**
 - This comes with the warehouse as consequence



Conclusions (2)

Rapid exploration will enable analysts to build up a much better and completer intuition about the statistic population

Population Registers can open a wealth of longitudinal insights, provided that the history of the records is maintained



Thank You !!

