FMS och Surveyföreningens höstmöte i Södertälje 22 november 2011

Anders & Britt Wallgren
SCB och Örebro Universitet

Vad innebär administrativa register för surveymetodiken?

1. **Survey = ?**
2. **Registerbaserad undersökning = ?**
3. **Administrativa data = ?**
4. **Samhällets administrativa register**
5. **Företagens administrativa register**
6. **Kvalitet = ?**
7. **Vilken är den stora skillnaden?**

1. **Survey = ?**

Statistics Canada Quality Guidelines

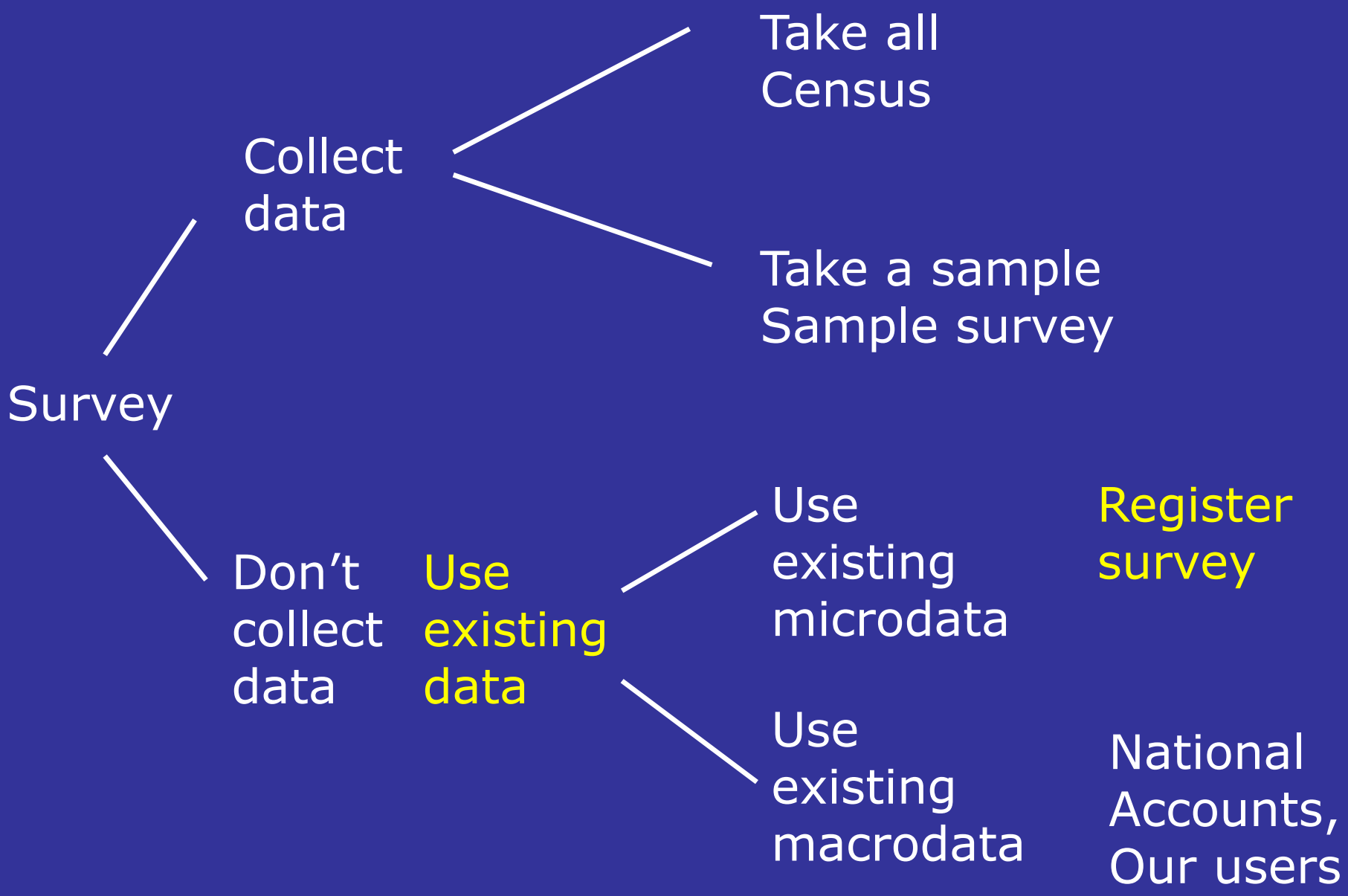
This document brings together guidelines and checklists of issues to be considered in the pursuit of quality objectives in the execution of statistical activities. It draws on the collective experience of many Statistics Canada employees. It should be useful to staff engaged in the planning and design of surveys as well as those who evaluate and analyze the results.

Preface by Ivan Fellegi

Survey

The term *survey* is used here generically to cover any activity that collects or acquires statistical data. Included are:

- a *census*, which attempts to collect data from all members of a population;
- a *sample survey*, in which data are collected from a (usually random) sample of population members;
- collection of data from *administrative records*, in which data are derived from records originally kept for non-statistical purposes;
- a *derived statistical activity*, in which data are estimated, modeled, or otherwise derived from existing statistical data sources.

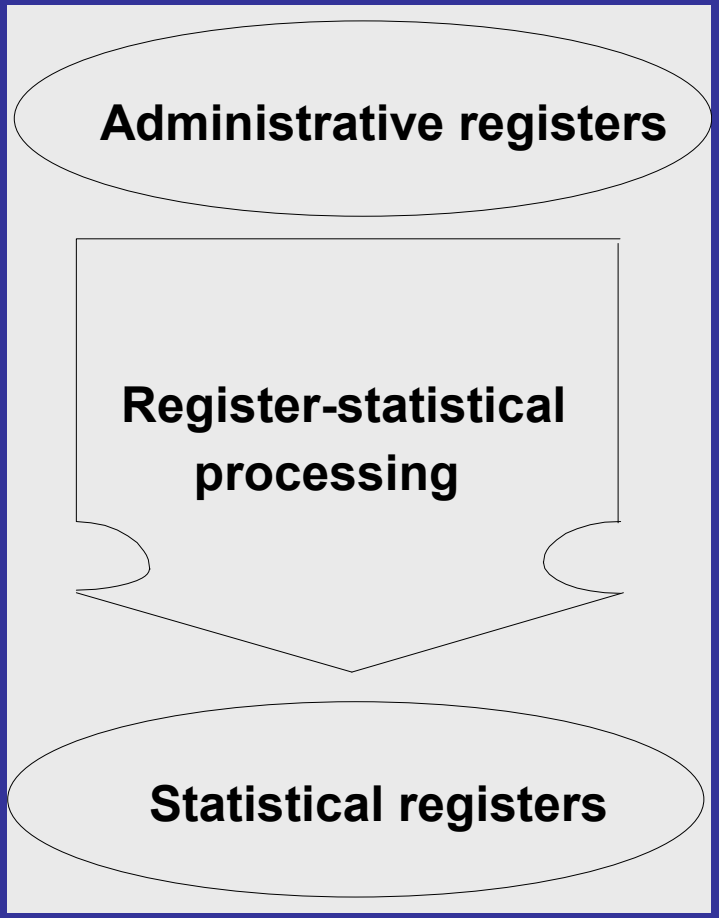

Take all
Census
Collect
data
Take a sample
Sample survey
Survey
Don’t
collect
data
Use
existing
data
Use
existing
microdata
Register
survey
Use
existing
macrodata
National
Accounts,
Our users

2. Registerbaserad undersökning = ?

Chart 1.1 Four principles on how to use administrative data

2. These administrative registers should be transformed into statistical registers.

   Many sources should be used and compared during this transformation.



Administrative object set
Administrative units
Administrative variables

Estimatorer?

Statistical population
Statistical units
Statistical variables

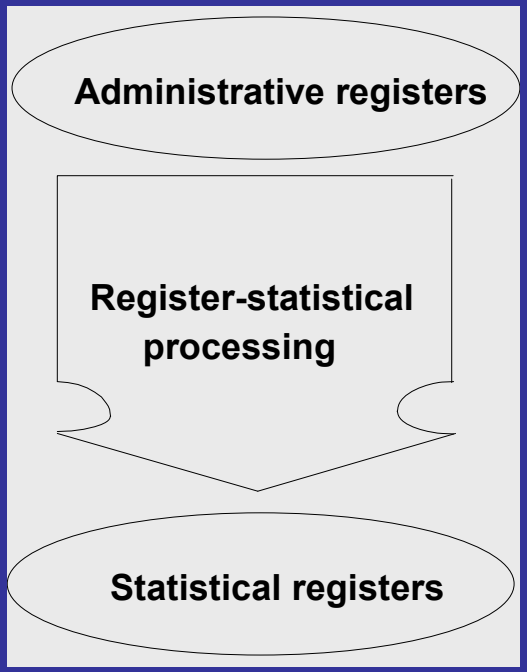
Estimatorer?

$$\hat{Y} = \sum_{i=1}^{r} d_i g_i y_i = \sum_{i=1}^{r} w_i y_i \qquad (1)$$

where r is the number of objects in the *sample* that responded in a particular cell

$$\hat{Y} = \sum_{i=1}^{R} y_i \qquad (2)$$

where R is the number of objects in the *register* in a particular table cell



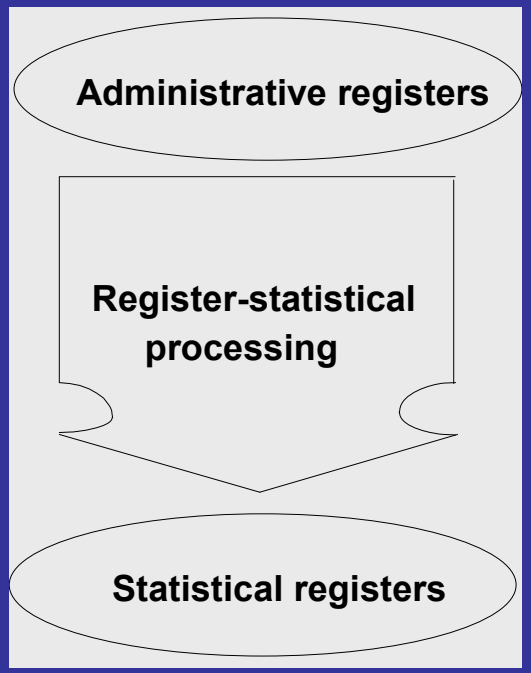
Estimatorer?

$$\hat{Y} = \sum_{i=1}^{r} d_i g_i y_i = \sum_{i=1}^{r} w_i y_i \quad (1)$$

where r is the number of objects in the *sample* that responded in a particular cell

$$\hat{Y} = \sum_{i=1}^{R} y_i \quad (2)$$

where R is the number of objects in the *register* in a particular table cell



$$\hat{Y} = \sum_{i=1}^{r} d_i g_i y_i$$

$$\hat{Y} = \sum_{i=1}^{R} y_i$$

3. Administrativa data = ?

3. Administrativa data = ?

”In principle, there is no difference between error structures found in administrative data compared to data collected via other modes”

Biemer, Lyberg: Introduction to Survey Quality, Wiley 2003

3. Administrativa data = ?

”In principle, there is no difference between error structures found in administrative data compared to data collected via other modes”

Chart 10.5 Measurement errors – comparison of data collection methods

Collecting data in sample surveys	Collecting data in admin. systems
Underlying structure of question: *Will you please try to understand our questions and try to remember?* *It is not necessary that you answer, and it does not matter what you answer, we will not do you any harm*	Underlying structure of question: *1. Report last month’s turnover before the 12th this month!* *2. Pay 25% of reported turnover before the 12th this month!* *3. If you don’t report and pay, you will have to pay extra!*

The Nature of Administrative Data

Data from admin. authorities can be of different nature:

1. Almost as statistical data - If the authority use it for their own statistics *(Statistical data are NOT legally important)*
2. Other kinds of data are legally important:

   Legal obligations are registered – a baby is born, the identities of mother and father are registered

   Ownership is registered – who owns a specific property or vehicle and has to pay tax for it

   Information from tax payers – wrong information can result in punishment

3. Data on decisions made by the authority:

The tax agency decides on taxable income and amount of tax to be paid

A court decides that a person is guilty of crime against a certain law and should get a specific punishment

Social authorities decide that a person/family should get some kind of benefit and how much money

Chart 5: Statistical questionnaire data, tax form data and administrative decision data

1. Questionnaire from NSO:	
Turnover	100
Costs	90
Profit	10

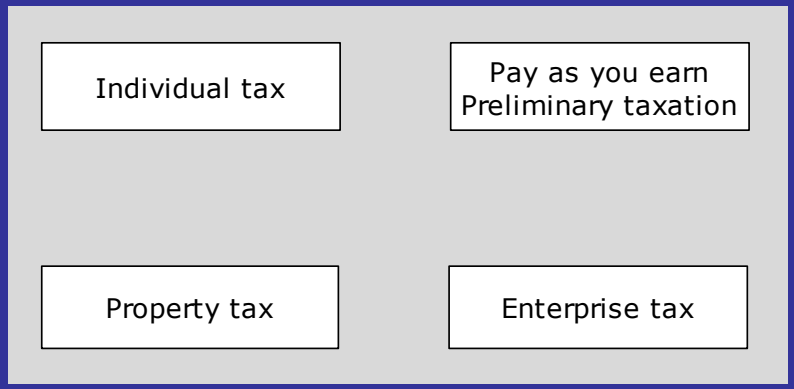
2. Tax form:	
Turnover	100
Costs	90
Profit	10

3. Decision by Tax Board:	
Turnover	100
Costs	70
Profit	30
Tax	9

4. Samhällets administrativa register

Identities of objects and Links between objects

The four main sources:



The four Base Registers with links:

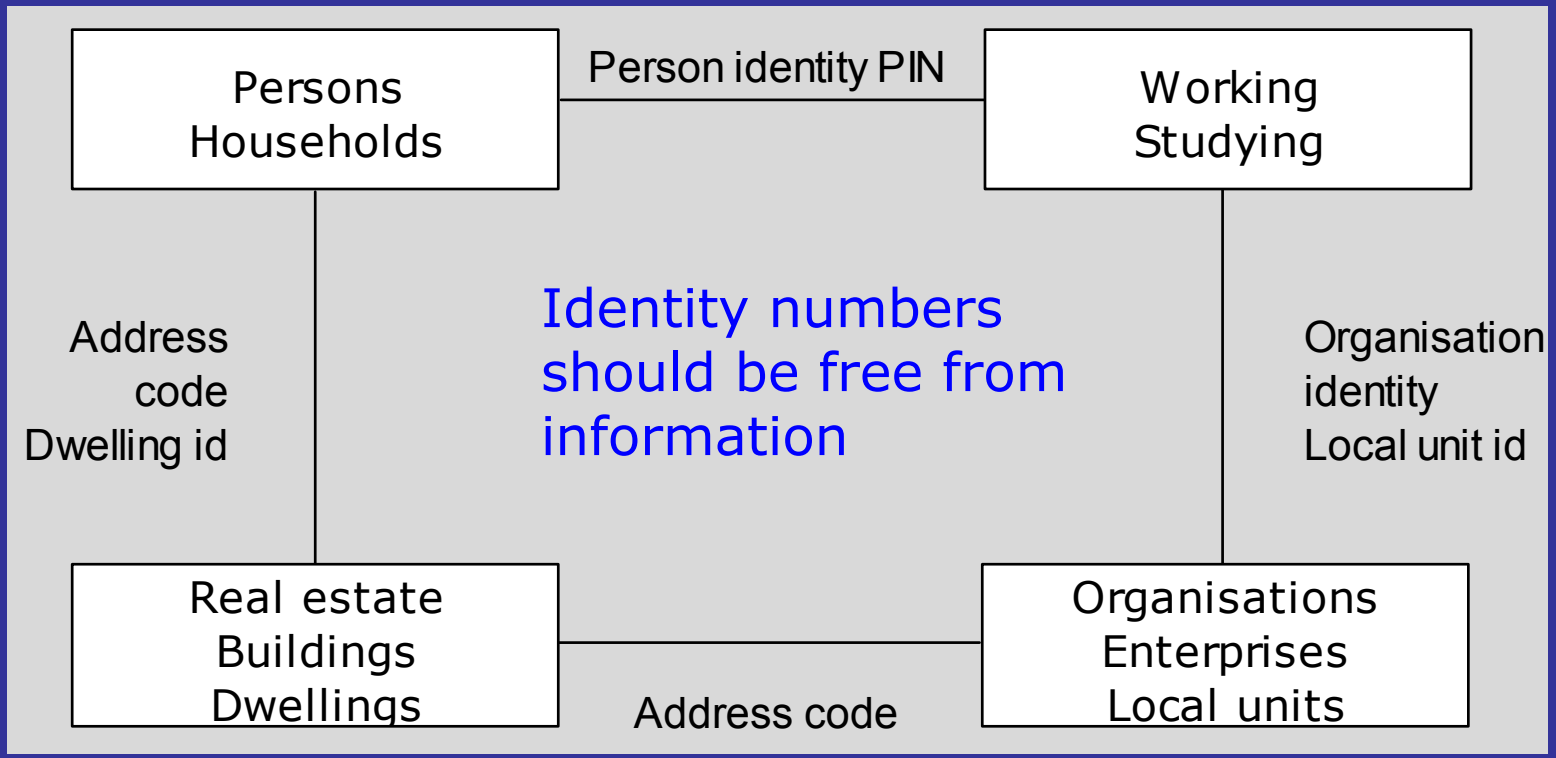
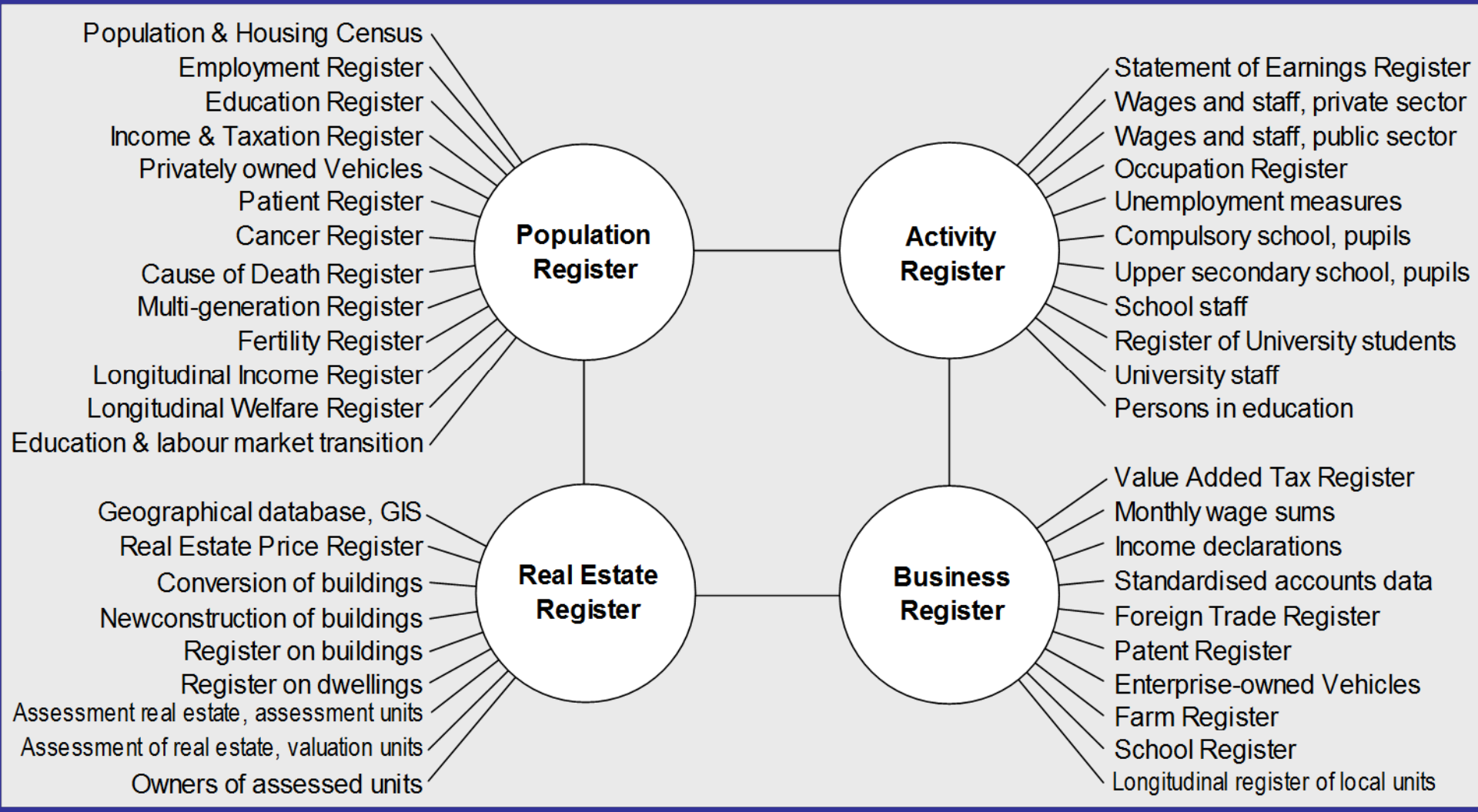


Chart 2.10 A system of statistical registers – registers by object type and subject field

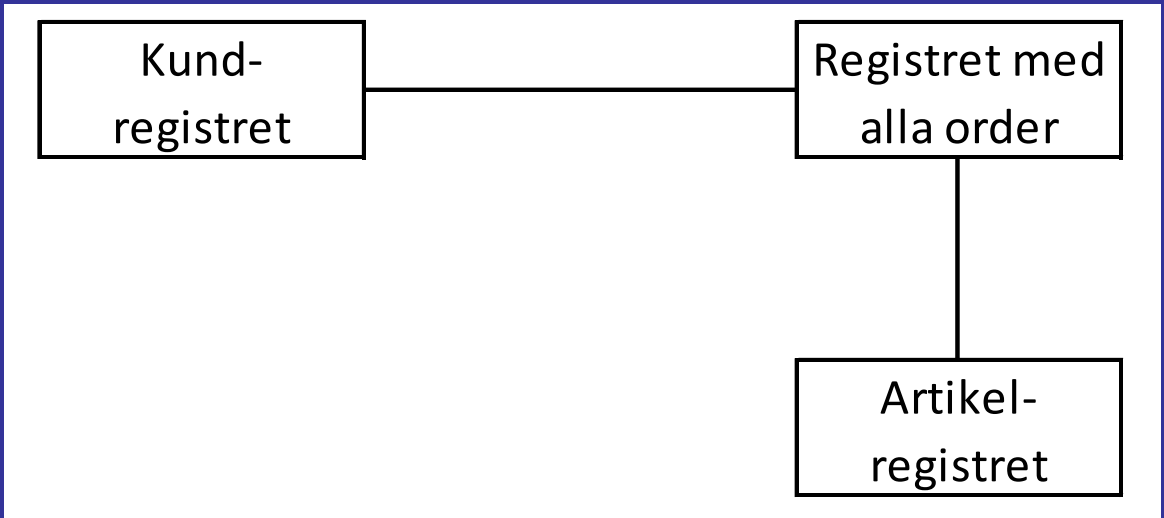


5. Företagens administrativa register

Chart 6: Administrative data based on decisions, without data collection or measurement

A customer phones and asks if enterprise X can deliver a certain quantity of a certain commodity. How much will it cost and when can it be delivered? After negotiations the following administrative data has been created:

Customer identity:	xxxx
Article number:	yyyy
Quantity:	qqqq
Price:	pppp
Delivery date:	dddd

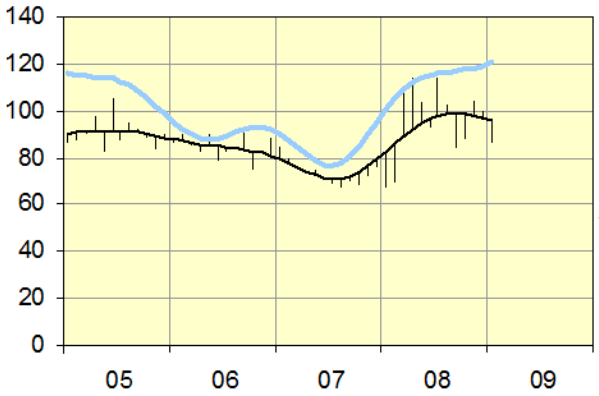


Statistiskt register med månadens fakturering skapas:

	Artikelregistret		Kundregistret				Orderregistret		Art.reg	
	Orderreg		Orderreg							Härledd var
Date	**Article-number**	**Article-group**	**Customer number**	**District**	**Industry**	**Country**	**Quantity**	**Value SEK**	**Cost SEK**	**Price SEK/Q**
2005-01-04	14	12	4	4	10	FI	210	17011	20025	81.00
2005-01-04	48	10	4	4	10	FI	375	14723	17411	39.26
2005-01-04	51	10	19	13	2	SE	163	16221	12732	99.52
2005-01-04	9	12	25	13	1	SE	70	8728	7283	124.69
2005-01-04	26	12	25	13	1	SE	245	8426	8223	34.39
2005-01-04	15	12	26	13	1	SE	189	35058	23154	185.49
2005-01-04	58	10	26	13	2	SE	2	317	216	158.50
2005-01-04	1	12	28	13	1	SE	7	575	408	82.14
2005-01-04	37	12	28	13	3	SE	16	1029	755	64.31
2005-01-04	73	11	28	13	1	SE	1275	23027	14129	18.06
2005-01-04	89	10	28	13	1	SE	125	7664	5482	61.31
2005-01-04	102	10	28	13	3	SE	249	5184	5110	20.82
2005-01-04	106	11	28	13	1	SE	850	11977	11784	14.09
2005-01-04	108	11	28	13	3	SE	1275	24722	21666	19.39
2005-01-04	106	11	69	16	1	SE	5100	59772	70702	11.72
2005-01-04	78	10	203	14	1	SE	403	12419	10094	30.82
2005-01-04	101	11	203	14	1	SE	680	19972	18877	29.37

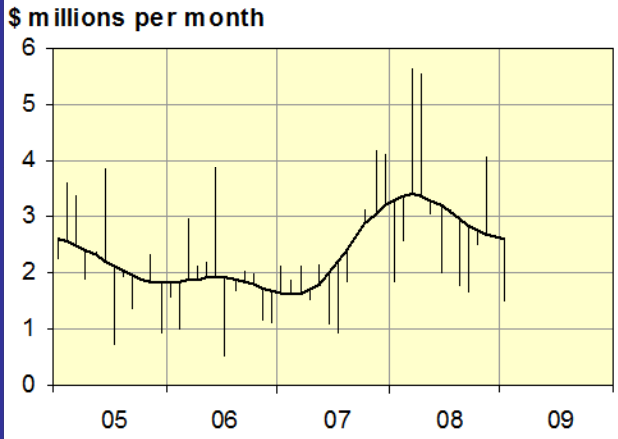
Customer 1 Data including 2009 period 1

1. Price index, 2008 = 100



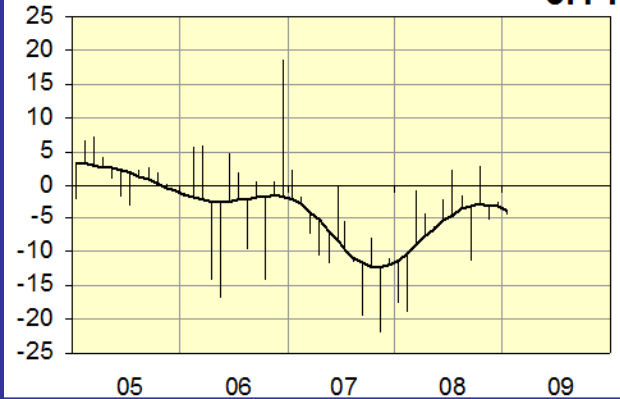
	Prices before discount	Prices after discount	Prices - raw material cost
2005	91.6	94.5	97.6
2006	84.8	88.3	80.7
2007	73.7	76.6	70.7
2008	100.0	100.0	100.0
Up to now:			
2009	86.3	86.4	77.5
Forecast			
2009	86.3		

2. Volume of sales, 2008 years prices



	$ millions per year
2005	26.4
2006	22.2
2007	27.8
2008	36.3
ARIMA-extrapolation	
2009	31.5
Forecast	
2009	31.5

3. Profit margin



	Mkr per år			Procent
	Sales-bonus	Profit	Costs	Profit margin
2005	24.1	0.5	23.6	2.0
2006	18.7	-0.4	19.1	-2.1
2007	20.5	-2.1	22.6	-10.3
2008	36.3	-1.4	37.7	-3.9
Up to now :				
2009	1.3	-0.1	1.3	-4.5
Forecast:				
2009				-4.5

Storytelling:

Relevant charts and tables together in a meaningful combination

Two “clics” to get this detailed information based on some of the 1 600 series

6. Kvalitet = ?

Main quality issues of:

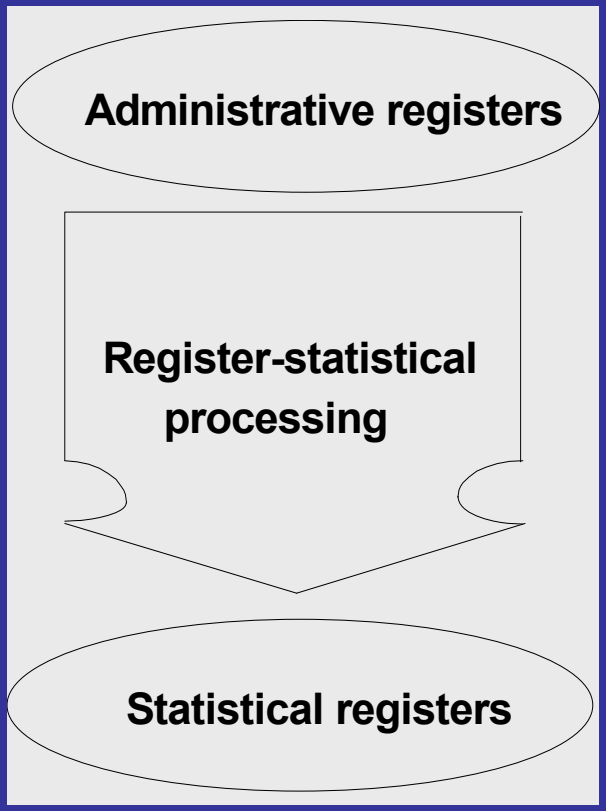
1. Sample surveys	2. Censuses	3. Register-based surveys
Errors, bad quality?		
Frame errors	Frame errors	Relevance errors ***Administrative data***
Nonresponse errors	Nonresponse errors	
Measurement errors	Measurement errors	Integration errors ***Data integration***
Sampling errors		
Data collection	***Data collection***	

The data collection process is not perfect

Relevance errors

1. Transform administrative data into statistical data

Relevance errors: <u>The transformation is not perfect</u>

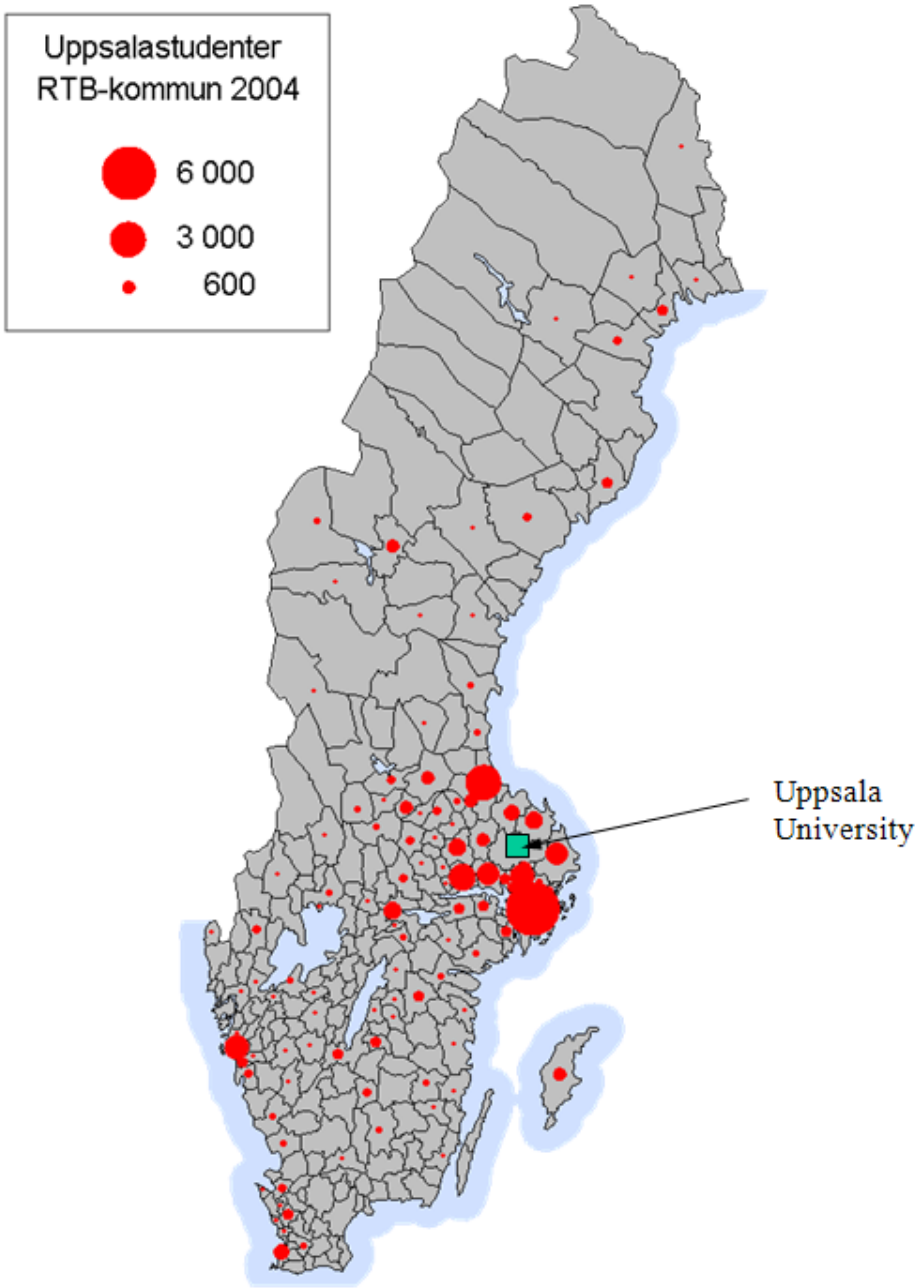


Administrative object set
Administrative units
Administrative variables

Statistical population
Statistical units
Statistical variables

Relevant variable/units?

Where do students at Uppsala University live?



According to the Population Register:

	Age	Address	Municipality
Father	45	1	1
Mother	44	1	1
Child 1	22	1	1
Child 2	20	1	1

=> One household with 4 members

In reality:

	Age	Address	Municipality
Father	45	1	1
Mother	44	1	1
Child 1	22	2	2
Child 2	20	3	3

=> Three households with 2, 1and 1 member

In the University Register there are information on present addresses of university students

"Use all relevant sources!"

Integration errors

Integration errors: The integration is not perfect

- **The populations do not agree**
- **The units do not agree**
- **The variables do not agree**

Different sources are from different points in time

Two sources are integrated:

Source 1		Source 2	
Id	Variable 1	Id	Variable 2
1	100	1	160
2	200	*	*
*	*	3	80

* = missing value

We imput missing values:

Source 1		Source 2	
Id	Variable 1	Id	Variable 2
1	100	1	160
2	200	2	150
3	100	3	80
Sums:	400		390

But, the truth is:

Source 1		Source 2	
Id	Variable 1	Id	Variable 2
1	100	3	80
2	200	1	160

Id 1 in Source 1 = Id 3 in Source 2
Id 2 in Source 1 = Id 1 in Source 2

Sums:	300	240

If units with the same identities in the sources are not the same, this will give integration errors

Two sources are integrated:

Matching always gives mismatch
How should this be interpreted?

Source 1:	Source 2:
Id 1 Id 2 Id 3	mismatch
Id 4 Id 5 Id 6 Id 7 Id 8 Id 9 Id 10 Id 11	Id 4 Id 5 Id 6 Id 7 Id 8 Id 9 Id 10 Id 11
mismatch	Id 12 Id 13 Id 14

Interpretation 1:
We trust Source 1

Source 1:	Source 2:	
Id 1 Id 2 Id 3	Id 1 Id 2 Id 3	Impute Impute Impute
Id 4 Id 5 Id 6 Id 7 Id 8 Id 9 Id 10 Id 11	Id 4 Id 5 Id 6 Id 7 Id 8 Id 9 Id 10 Id 11	
	Discard these	

If this interpretation wrong:
=> Integration error

Interpretation 2:
We trust both Source 1 and 2

Source 1:		Source 2:	
Id 1 Id 2 Id 3		Id 1 Id 2 Id 3	Impute Impute Impute
Id 4 Id 5 Id 6 Id 7 Id 8 Id 9 Id 10 Id 11		Id 4 Id 5 Id 6 Id 7 Id 8 Id 9 Id 10 Id 11	
Id 12 Id 13 Id 14	Impute Impute Impute	Id 12 Id 13 Id 14	

If this interpretation wrong:
=> Integration error

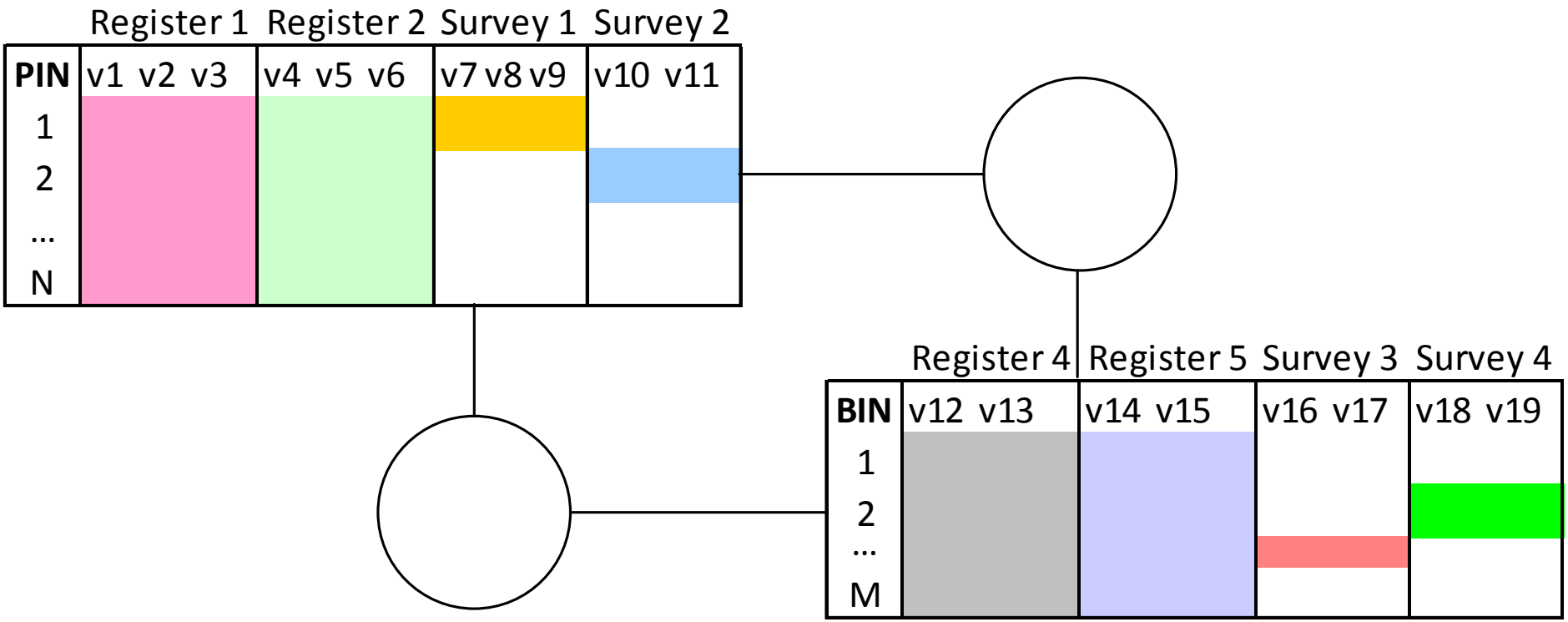
Wrong handling of mismatch will give integration errors

7. Vilken är den stora skillnaden?

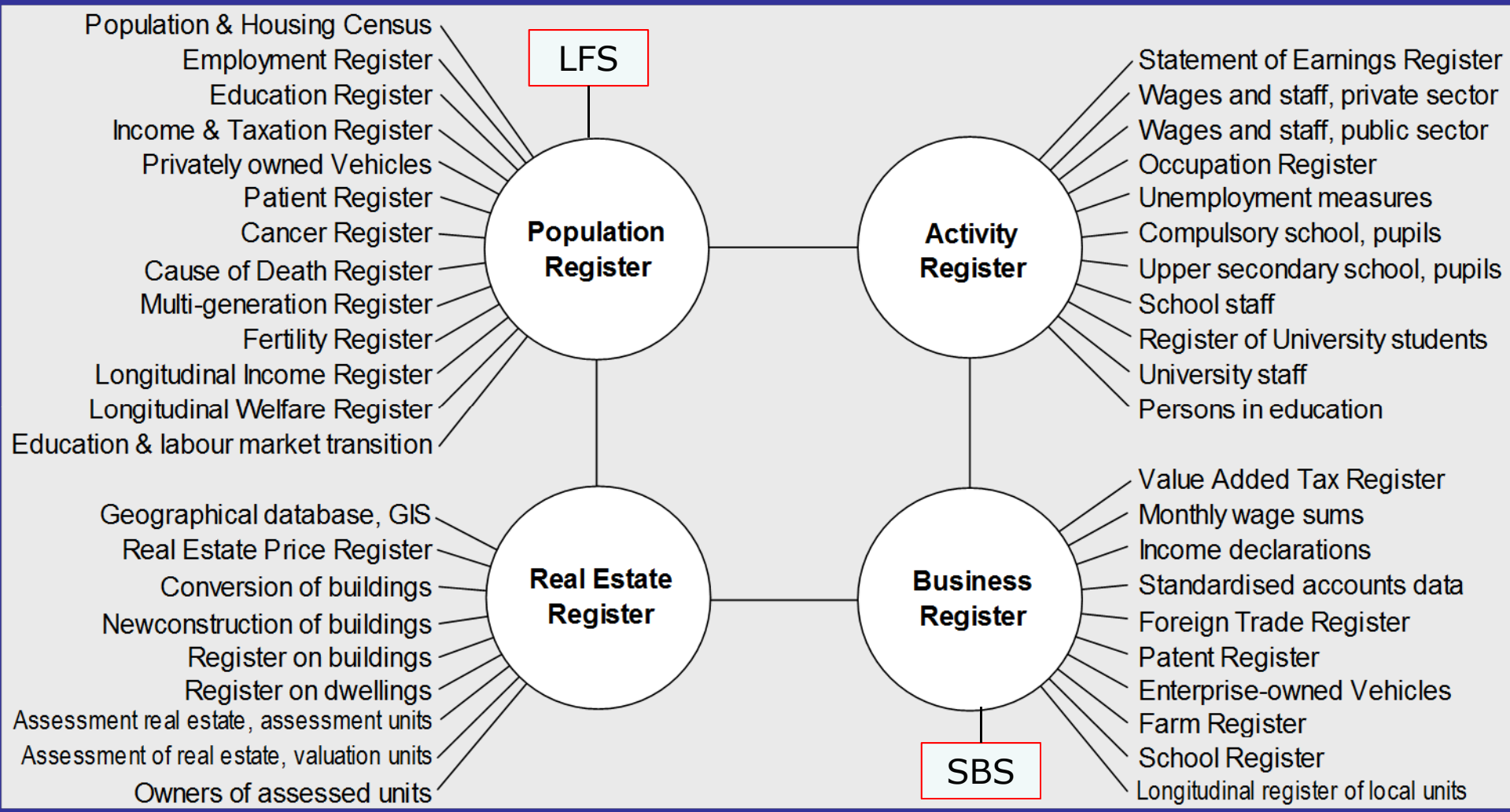
All registers and sample surveys on persons can be combined by PIN = Person Identity Numbers

All registers and sample surveys on enterprises can be combined by BIN = Business Identity Numbers

We do this in Register-based Census and National Accounts



All registers and censuses in the system can be combined
All sample surveys can be combined with all registers



No sample surveys can be combined with other sample surveys
One survey at a time thinking was OK before the system

- All registers and censuses in the system can be combined

- All sample surveys can be combined with all registers

- No sample surveys can be combined with other sample surveys. One survey at a time thinking was OK before the system but not now

- Combining sources, why:

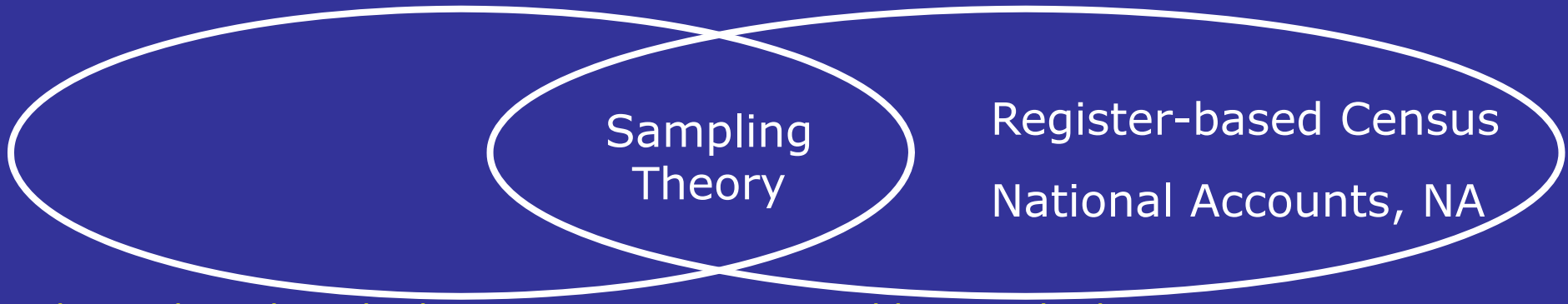
  - Subject matter reason: Much richer content

  - Methodological reason:
    Can find and correct errors:
    coverage errors, consistence, coherence

- New paradigm, new theory:
  Theory of statistical systems

Statistical science at Universities

Statistical methods at a National Statistical Office



Theory-based methods, common terms

Ad hoc methods, no common terms

Paradigm:

Probability theory
Inference theory

Census:

System of registers

NA: System of registers and sample surveys

Register-statistics in the Nordic countries was not developed with the help of methodologists

Statistical science at Universities

Statistical methods at a National Statistical Office

Theory-based methods, common terms

Paradigm:

- Probability theory
- Inference theory

New paradigm:

- New theory
- New methods
- Statistical systems

Theory-based methods, common terms

Census:

- System of registers

NA:

- System of registers and sample surveys