Innovations in business statistics data collection

EMOS webinar 2018
21 February 2018

Mr. Ger Snijkers, Statistics Netherlands, Heerlen, Netherlands

Acknowledgement

The research presented in this presentation is an overview of research conducted by myself as well as a number of colleagues, as listed in the reference list. The views expressed are those of the lecturer and do not necessarily reflect the policies of Statistics Netherlands.

I am grateful to:

- Daas, 2017, Statistics Netherlands
- De Broe, 2017, Statistics Netherlands
- Haraldsen and Couper, 2013, Statistics Norway and University of Michigan
- Rooijakkers, 2017, Statistics Netherlands
- Vonder, 2017, TNO Netherlands

for the use of their slides.

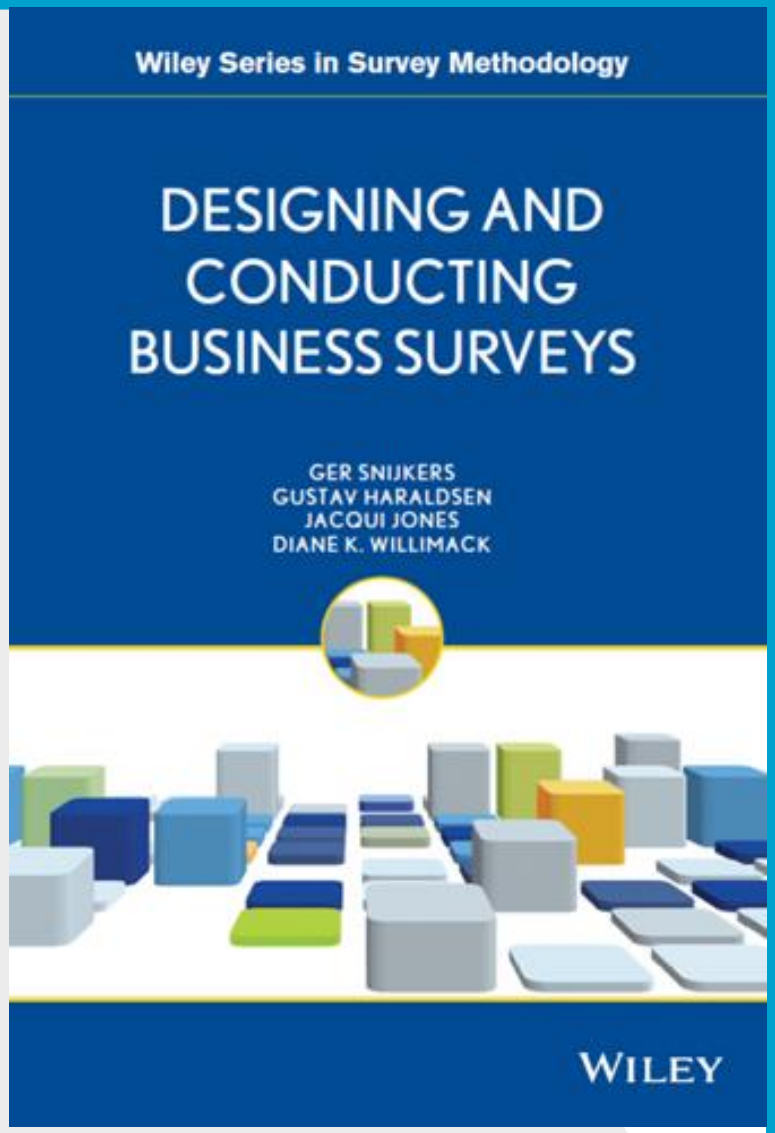
This webinar:

- I cannot discuss business surveys in full
- I will discuss:
  - Three main characteristics of business data collection
  - A brief history of business data collection
  - A general Data collection strategy
  - Directions of technological innovations
  - Methodological and organizational issues
  - A few examples of innovations

For more details:

For more details on business survey designs:

- Snijkers, Haraldsen, Jones, and Willimack, 2013, Wiley.
- ESTP course: Designing and conducting business surveys for official Statistics
  5-7 Nov 2018, Oslo
  Google: "estp eurostat 2018"
- BDCM Workshop:
  19-21 Sept 2018, Lisbon
  http://bdcmlisbon2018.ine.pt
  Deadline abstract: 20 March



Statement

Technological innovations make things possible; the applied methodology and the organisational context make it work.

E.g. introduction of :

- Web surveys (from paper to web questionnaires):
  "*We suspect that many of the survey organizations that introduce web questionnaires forget that it is not the technology in itself, but how it is utilized that determines the result.*" (Haraldsen & Couper, 2013)
- Register data
- Electronic Data Interchange: System-to-System (S2S) data communication

Overview

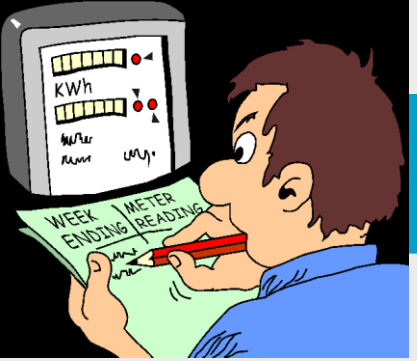
- Background on Business data collection:
  - Characteristics of business survey data collection
  - History of business data collection
  - General Data collection strategy
- Technological innovations:
  1. Computerisation of survey data collection
  2. Computerisation of the business information chain
  3. Internet as data source
  4. Internet of Things (IoT)
- Conclusions and future developments

Questions to you

- Background on Business data collection:
  - Characteristics of business survey data collection
  - History of business data collection
  - General Data
  - Logical
  - Computerisation
  - Computerisation of the business information chain
  - Internet as
  4. Intern
- Conclusio

Who of you has experience with business surveys?

What are the main issues you have to deal with?



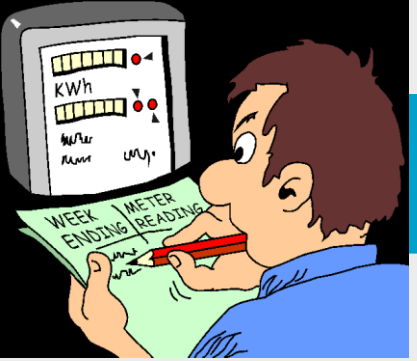
Key features of :

Business survey designs

- disproportionate samples
- self-completion Qs (web, paper)
- mainly numeric data: facts; complex Qs: matrixes
- letters and tel. contacts; large bus.: personal contacts
- post-field: re-contact to validate

Household survey designs

- equal probability samples
- mixing modes: self- and interviewer- admin. Qs
- mainly categorical data: facts and opinions
- letters and personal contacts; usage of incentives
- No post-field contacts



Key features of :

Business survey designs

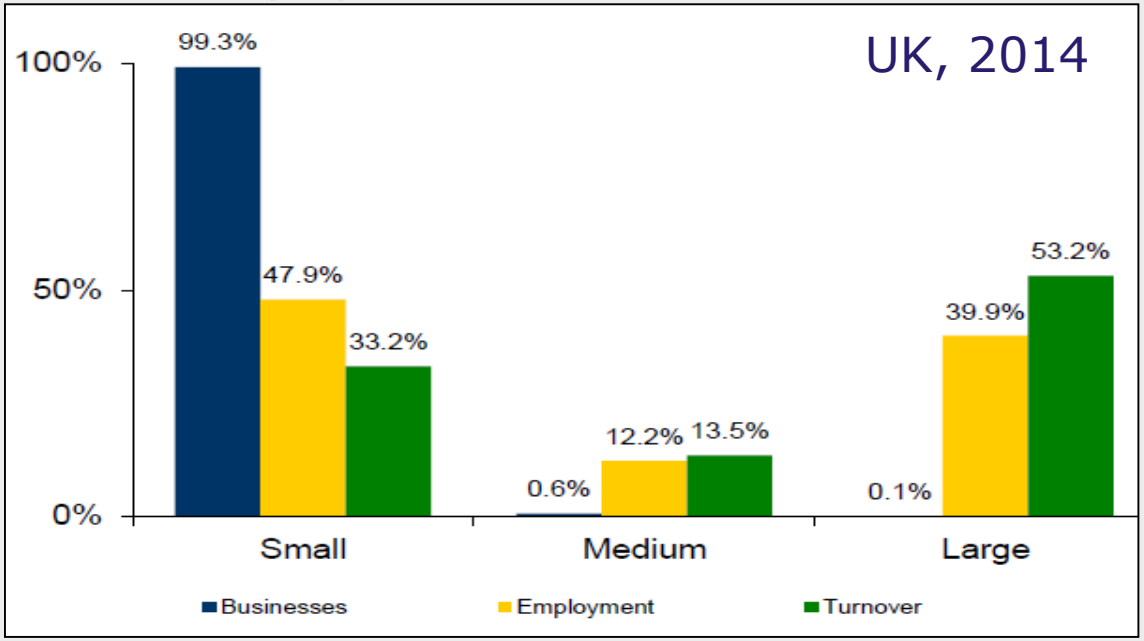
- disproportionate samples
- self-completion Qs (web, paper)
- mainly numeric data: facts; complex Qs: matrixes
- letters and tel. contacts; large bus.: personal contacts
- post-field: re-contact to validate

Why this design?

- some businesses are more important
- data need to be retrieved from various sources
- many respondents involved
- proxy reporting
- unit issues: reporting ≠ observational
- Official statistics: compulsory by law

Main characteristics of business population

1. Skewed population



2. Multi-surveyed
   - more than one
   - more than once!

   long-lasting relationship

3. Complex response process:

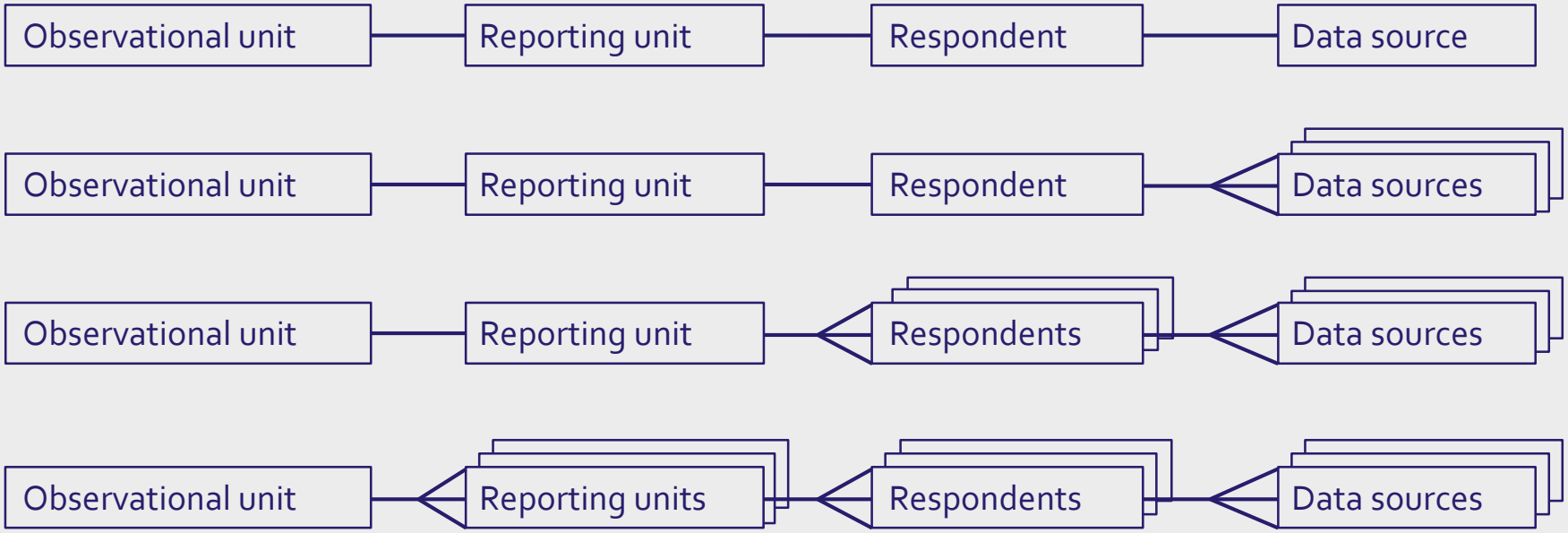
Tailoring conside-rations for business survey designs

Complex response process

- Many data sources, at various locations
- Many people, at various locations
- Many sub-units

} or a combination

- Time: when data are available, and businesses have time



Complex response process

- Many data sources, at various locations
- Many people, at various locations
- Many sub-units

} or a combination

- Time: when data are available, and businesses have time

Information distributed…

- Between people
- Between sources
- Over time

Affects measurement and unit errors

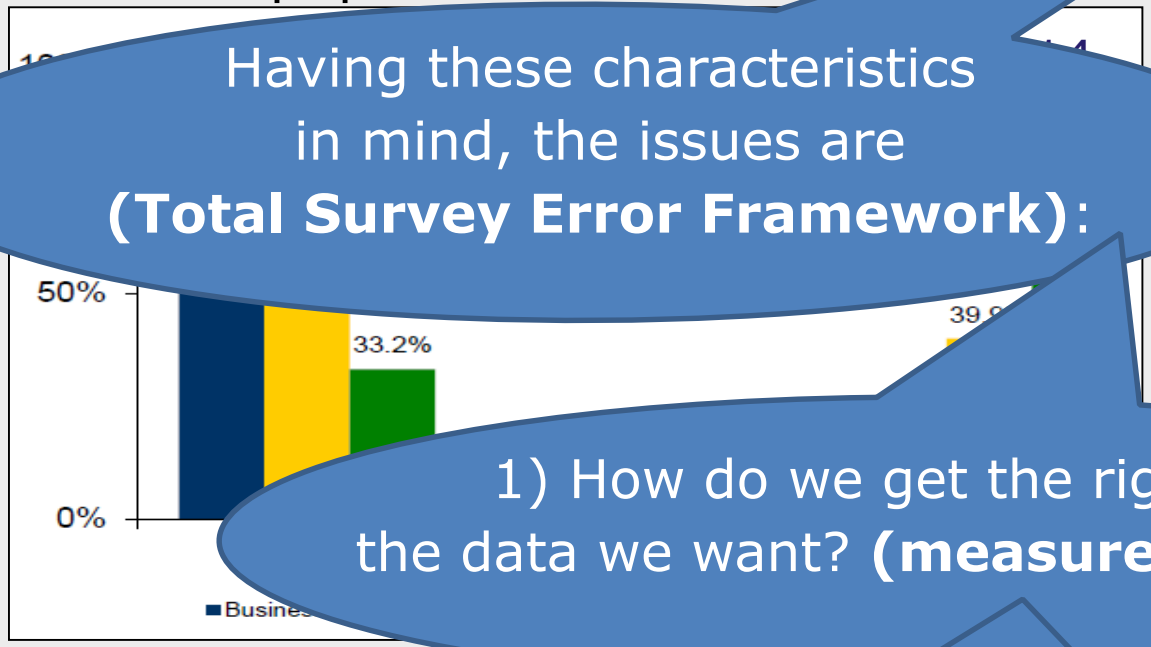
- More complicated structure
  > more steps
  > higher error risks
- Mismatch between
  > organizational structure and
  > questionnaire structure (question flow)
  higher error risks

Consequences of main features for Business Surveys Designs?

Business Survey Characteristics	Web Design Issues Particular to Business Surveys
Skewed business population	• Identify relevant subgroups (stratification) with regard to sampling, estimation, questionnaire, and communication. • Tailor to these strata (size, sector of industry, your target variable (e.g. relevance to the economy, globalisation))
Businesses are multi-surveyed	• Establish a relationship: try to get commitment from the very first contact using pre-notifications • Web portal and survey calendar to provide overview
Complex response process	• Make an effort to contact the most competent R(s), who has access to and can judge available information • Indicate the observational unit and reference period • Structure the questionnaire according to business' internal data collection process • Facilitate multiple access; print option

Main characteristics of business population

1. Skewed population



2. Multi-surveyed
   - more
   - more

3. Complex response process:

Having these characteristics in mind, the issues are **(Total Survey Error Framework)**:

1) How do we get the right data, the data we want? **(measurement error)**

2) How do we get the data from the right unit, from the pre-defined unit? **(unit error)**

Tailoring considera-

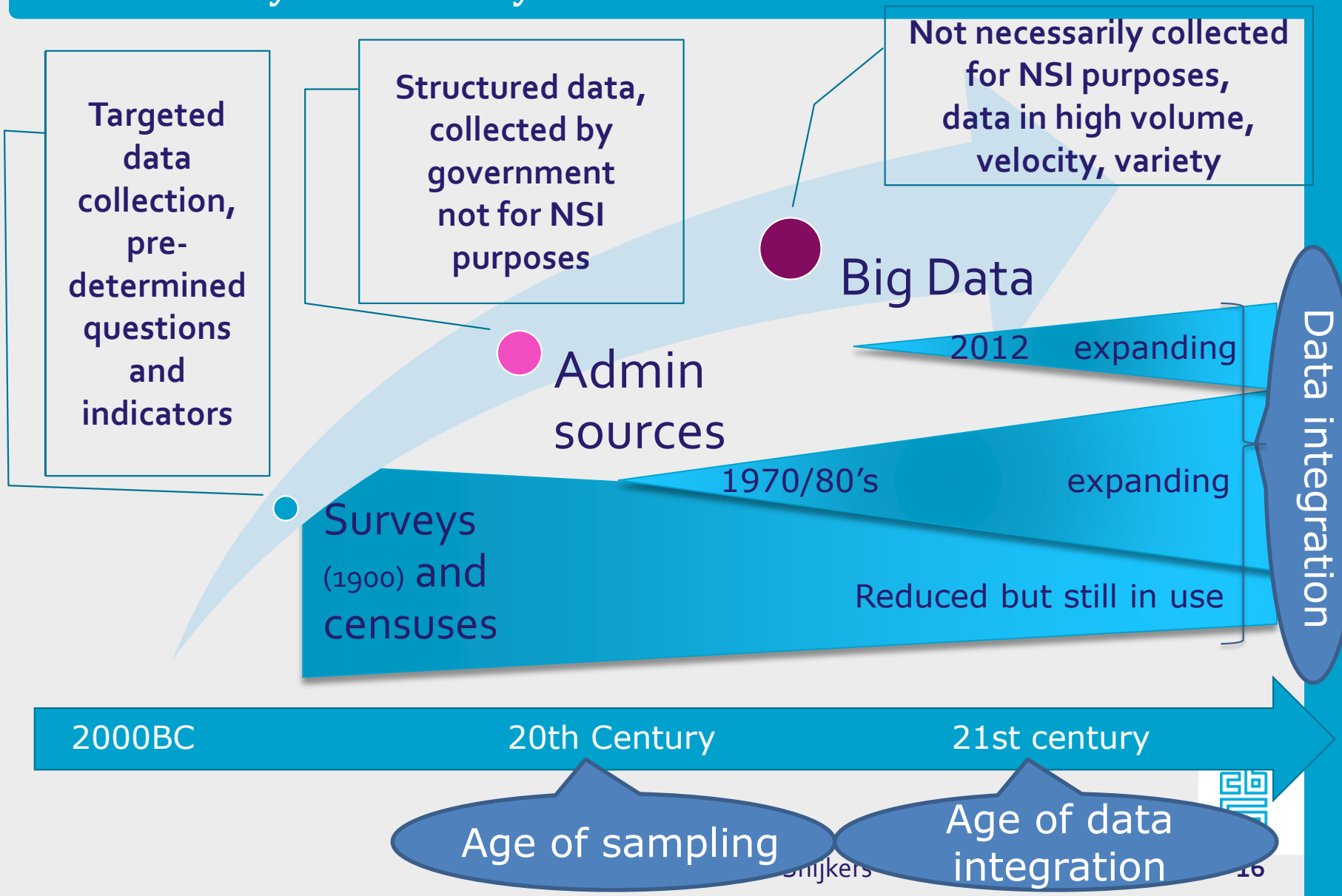
designs

Overview

- Background on Business data collection:
  - Characteristics of business survey data collection
  - History of business data collection
  - General Data collection strategy
- Technological innovations:
  1. Computerisation of survey data collection
  2. Computerisation of the business information chain
  3. Internet as data source
  4. Internet of Things (IoT)
- Conclusions and future developments

History

From single source to multi source:
from Primary to Secondary to New Data sources



History

From single source to multi source:
from Primary to Secondary to New Data sources

Targeted data collection, pre-determined questions and indicators

Structured data, collected by government not for NSI purposes

Not necessarily collected for NSI purposes, data in high volume, velocity, variety

Big Data

2012 expanding

Admin sources

1980's expanding

Surveys and censuses

Reduced but still in use

2000BC 20th Century 21st century

NSIs more and more out of control of the data collection

History: conclusions + data collection strategy

- **Mixed-mode/multi-source approach**
  - Moving from a single sources (surveys) to multiple sources
  - Moving from single-mode surveys (paper) to mixed-mode surveys, and back to a primary mode (web)
- The age of data integration
- NSIs loose control over data collection, but partner in data

General Data collection strategy:

1. Secondary sources
2. Surveys, only if necessary

➢ Reduction of questionnaires as much as possibe

➢ Many NSI have a comparable policy

Overview

- Background on Business data collection:
  - Characteristics of business survey data collection
  - History of business data collection
  - General Data collection strategy
- Technological innovations:
  1. Computerisation of survey data collection
  2. Computerisation of the business information chain
  3. Internet as data source
  4. Internet of Things (IoT)
- Conclusions and future developments

Questions to you

What innovations in business data collection is your organisation working on?

- Technological innovations:
  1. Computerisation of survey data collection
  2. Computerisation of the business information chain
  3. Internet as data source
  4. Internet of Things (IoT)
- Conclusions and future developments

Technological innovations

1. **Computerisation of survey data collection**
   a. Features of electronic questionnaires
   b. Paradata
2. **Computerisation of the business information chain**
   a. Electronic Data Interchange (EDI)
3. **Internet**
   a. Internet as data source
4. **Internet of Things (IoT):**
   Computerisation of business processes
   a. New data using EDI

Computerisation of survey data collection

Implementation of technological features, like:

- Automated routing
- Built-in edit checks
- Complex questionnaires (matrices)
- Imputation of t-1 data (historic data)
- Web log-in portals
- Paradata: collection of paradata

Both for off-line and internet Qs

Instead of simply using paper lay-outs

How to use these features?
- Methodology (Ch. 8 in Snijkers et al., 2013, Wiley)
- Organisational issues

Methodological innovations in surveys

Tailoring to the business context:

better insights in how businesses and people within these businesses think and operate

- Questionnaire *communication* instead of Q *design*:
  - Questionnaire communication design
  - Usability issues / User-interface design / interaction design
- Pre-testing of questionnaires/completion process as soon as possible in the design process
  - Feasibility studies
  - Usability + eye-tracking studies
- Apply '*influence principles*' (Cialdini) and '*nuding*' (Thaler & Sunstein) in survey communication to get response

  (Ch. 9 in Snijkers et al., 2013, Wiley)

Paradata

- Paradata = process data
  = data about the own data collection process
- Computerisation of process -> easy to collect and analyse
  - At NSI side:
    - Dates when questionnaires are received and processed
    - Log-in information
    - Cost and quality indicators
  - At Respondent side (audit trails):
    - When respondents open questionnaire
    - How they complete the questionnaire: completion process

    Example: SBS completion process (Structural Business Survey) (Snijkers & Morren, 2010)
- Provide insights in these processes

Paradata

Help to:

- Tailor the survey design to the business context:
  - Questionnaire communication design
  - Business survey communication strategy to get response
  - Efficient sampling to reduce response burden
  - Adaptive designs
- Make the survey process more efficient:
  - Applying the Deming cycle: Monitoring and improving
    (Ch. 10 in Snijkers et al., 2013, Wiley)

Technological innovations

1. **Computerisation of survey data collection**
   a. Features of electronic questionnaires: off-line, web Q
   b. Paradata
2. **Computerisation of the business information chain**
   a. Electronic Data Interchange (EDI)
3. **Internet**
   a. Internet as data source
4. **Internet of Things (IoT):**
   Computerisation of business processes
   a. New data using EDI

Innovations in business surveys

Traditional paper collection

- From Human-to-Human (H2H)
- Paper questionnaires

Transformation to digital format

- From Human-to-System (H2S)
- Web questionnaires

Unbroken digital chain

- From System-to-System (S2S)
- Electronic Data Interchange

- Before 1990s
- Paper still used up till today

- Electronic Qs
- Started in the 1990s: off-line, later on-line
- Primary mode since 2000

- Automated Data Capture
- Started in the 1990s
- Portugal
- The future

Technological innovations - 2

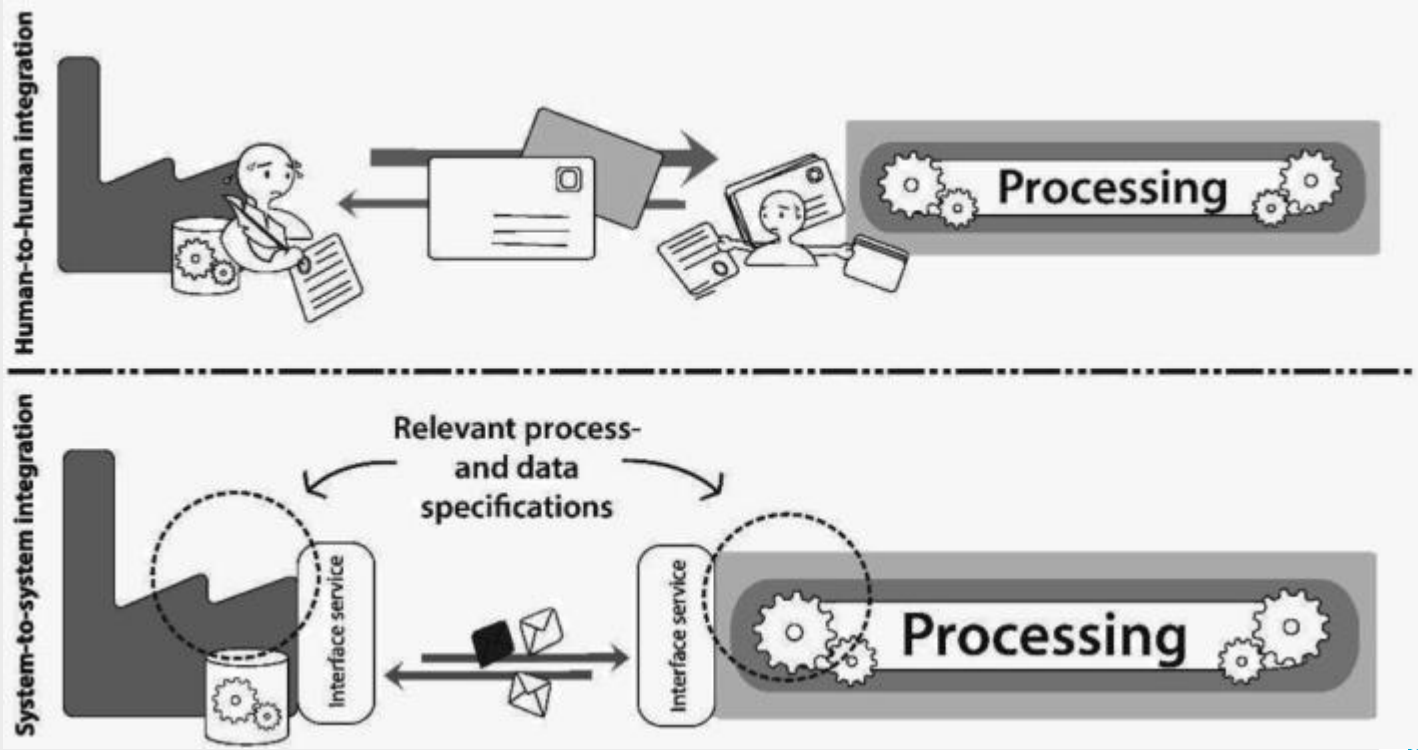
Computerisation of business information chain

Buiten, G., G. Snijkers, et al., 2018 (forthcoming), Journal of Official Statistics, ICES-V special issue

- Electronic Data Interchange (EDI) of financial data

H2H communi-cation: question-naires

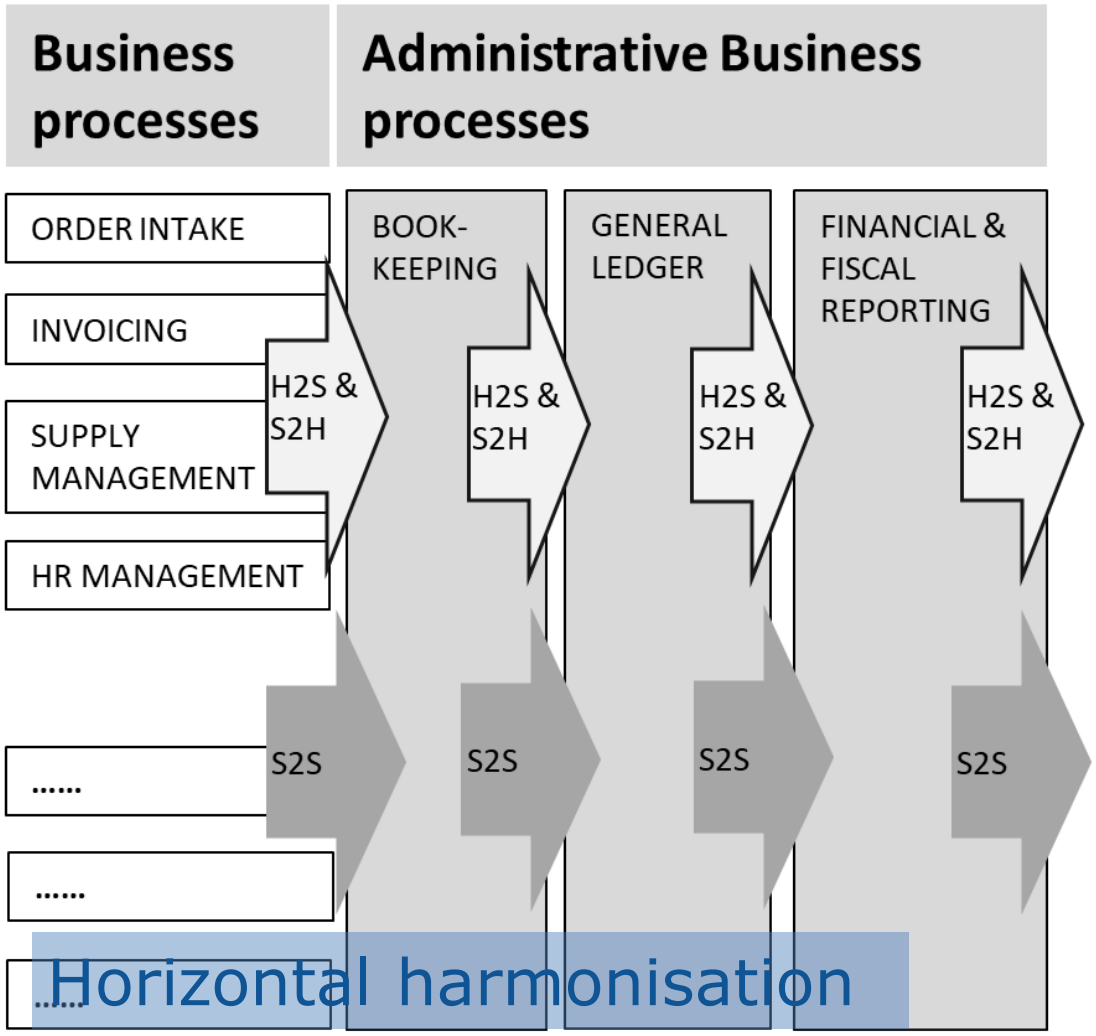
S2S communi-cation: EDI



Bharosa et al., 2015: Figure 1.3, p. 9.

Technological innovations - 2

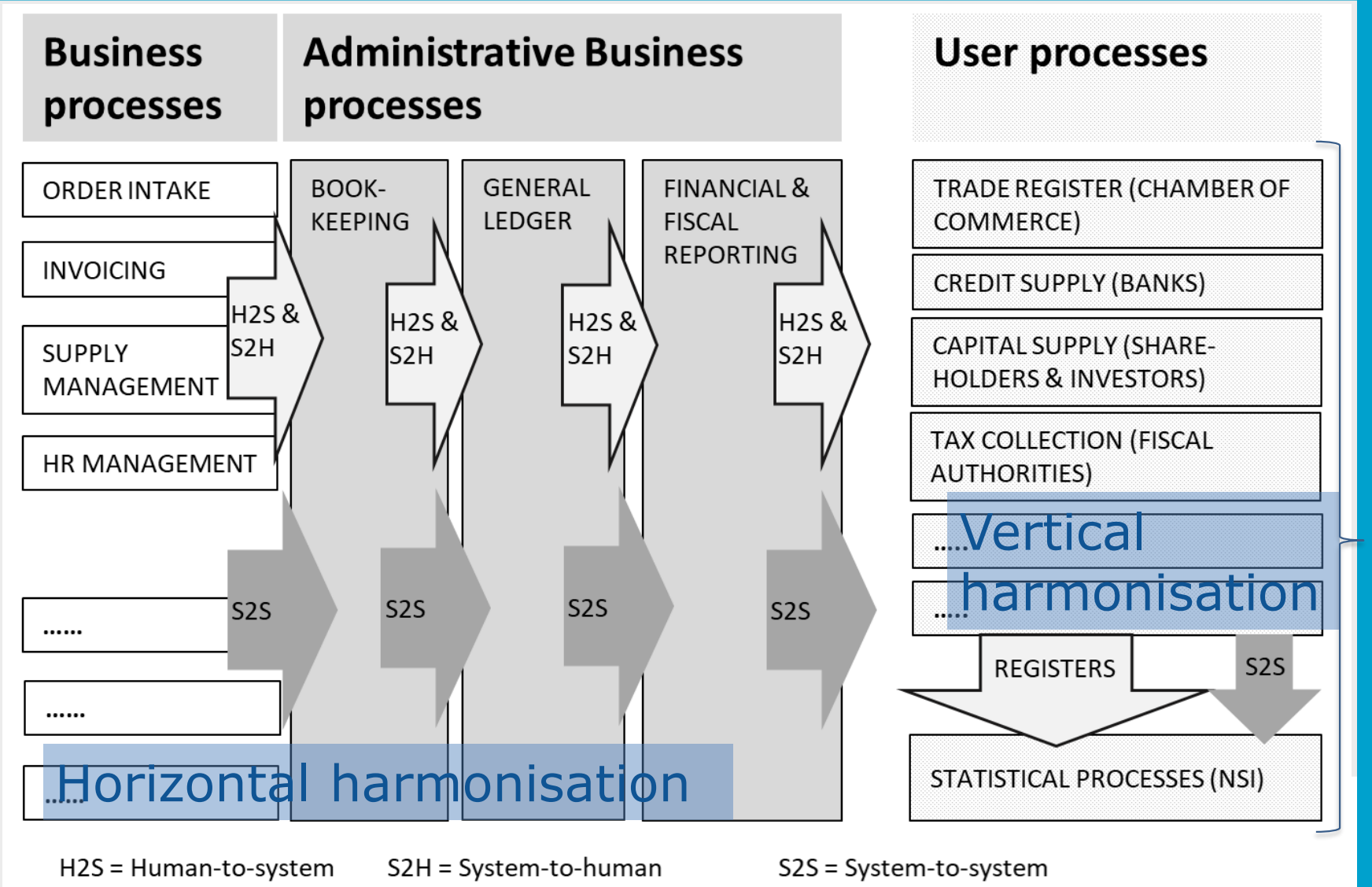
Computerisation of business information chain



H2S = Human-to-system S2H = System-to-human S2S = System-to-system

Technological innovations - 2

Computerisation of business information chain



Technological innovations - 2

Computerisation of business information chain

Investments by businesses and all other parties

EDI requirements:

- Harmonisation of financial concepts:
  - Horizontal: harmonisation of metadata within a business
  - Vertical: Standard Chart of Accounts used by all parties
  - Stability of taxonomy
- Technological standardisation
  - SBR: Standard Business Reporting - using one standard computer language (XBRL: eXtensible Business Reporting Language) on the internet
- Quality issues:
  - Measurement issues: data definition mismatch/mismapping (bookkeeping and statistical defintions), missings
  - Unit issues: is the unit correctly represented in the business records

Trust in EDI data

Technological innovati

Questions?

Computerisation of business informatic

Investments by businesses and all other parties

EDI requirements:

- Harmonisation of financial concepts:
  - Horizontal: harmonisation of metadata within a business
  - Vertical: Standard Chart of Accounts used by all parties
  - Stability of taxonomy
- Technological standardisation
  - SBR: Standard Business Reporting - using one standard computer language (XBRL: eXtensible Business Reporting Language) on the internet
- Quality issues:
  - Measurement issues: data definition mismatch/mismapping (bookkeeping and statistical defintions), missings
  - Unit issues: is the unit correctly represented in the business records

Trust in EDI data

Technological innovations

1. **Computerisation of survey data collection**
   a. Features of electronic questionnaires: off-line, web Q
   b. Paradata
2. **Computerisation of the business information chain**
   a. Electronic Data Interchange (EDI)
3. **Internet**
   a. Internet as data source
4. **Internet of Things (IoT):**
   Computerisation of business processes
   a. New data using EDI

The internet as data source

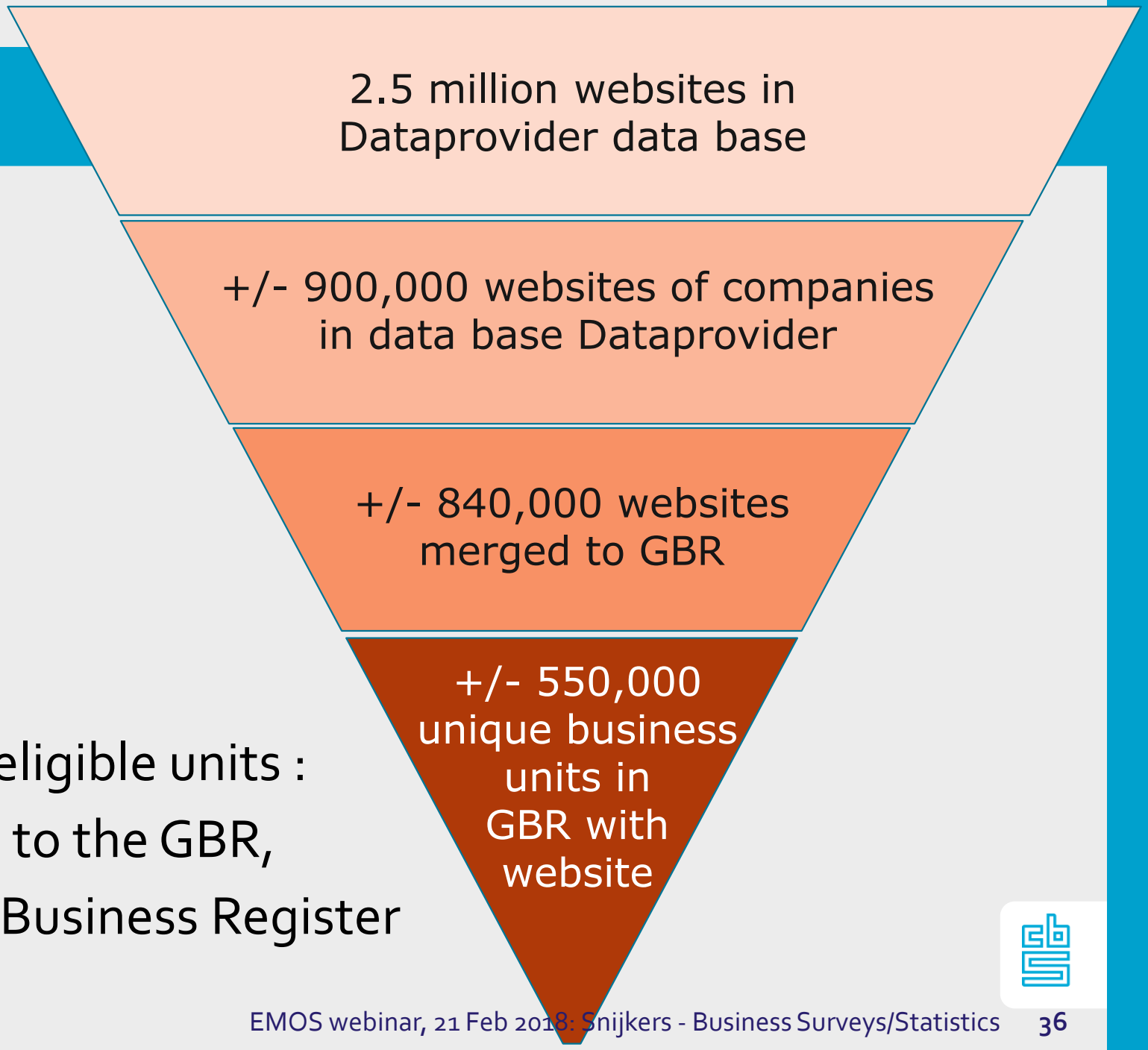
- Collecting Annual reports published on the internet and analysing the reports using text mining (chart of balance data) instead of using questionnaires
- Scraping prices on the internet
- Measuring the internet economy, instead of conducting a survey on e-commerce
- Finding information about innovative businesses
- Social media data

The internet as data source: > Measuring the internet economy

Rooijakkers, B., 2017, Measuring the internet economy with big data. Statistics Netherlands

- **Identifying businesses that do business using the internet?**
  - "Dataprovider" database: 2.5 million Dutch websites

Business information	• Country, address, company name, Chamber of Commerce number, taks number, phone number, e-mail, ...
eCommerce	• eCommerce probability, shopping cart software, delivery services, payment methods, products, prices,...
Content	• Title, description, keywords, category, language, author....
Other	• Marketing, social media, links, technical and hosting information, ...



Finding eligible units :

Merging to the GBR,

General Business Register

Applying definition of internet economy

All Dutch businesses

No website

Category A: No Income generated: businesses without a website. **A**

- Hairdresser without website
- Bakery without website
- Freelancer without website

Businesses without a website

Online presence

Category B1: Income generated indirectly through the internet (passive online presence) **B1**

- Hairdresser with website
- Shell
- DSM

Category B2: Income generated indirectly through the internet (active online presence) **B2**

- Car rental company
- Hotels
- High street store with supplementary webshop

Businesses with a website that do not belong to category C, D or E

Core of the internet economy

Category C: Income generated directly through the internet: online stores **C**

- Bol.com
- Wehkamp
- Coolblue

Online stores

Category D: Income generated directly through the internet: online services. **D**

- Relatieplanet
- Airbnb
- Marktplaats

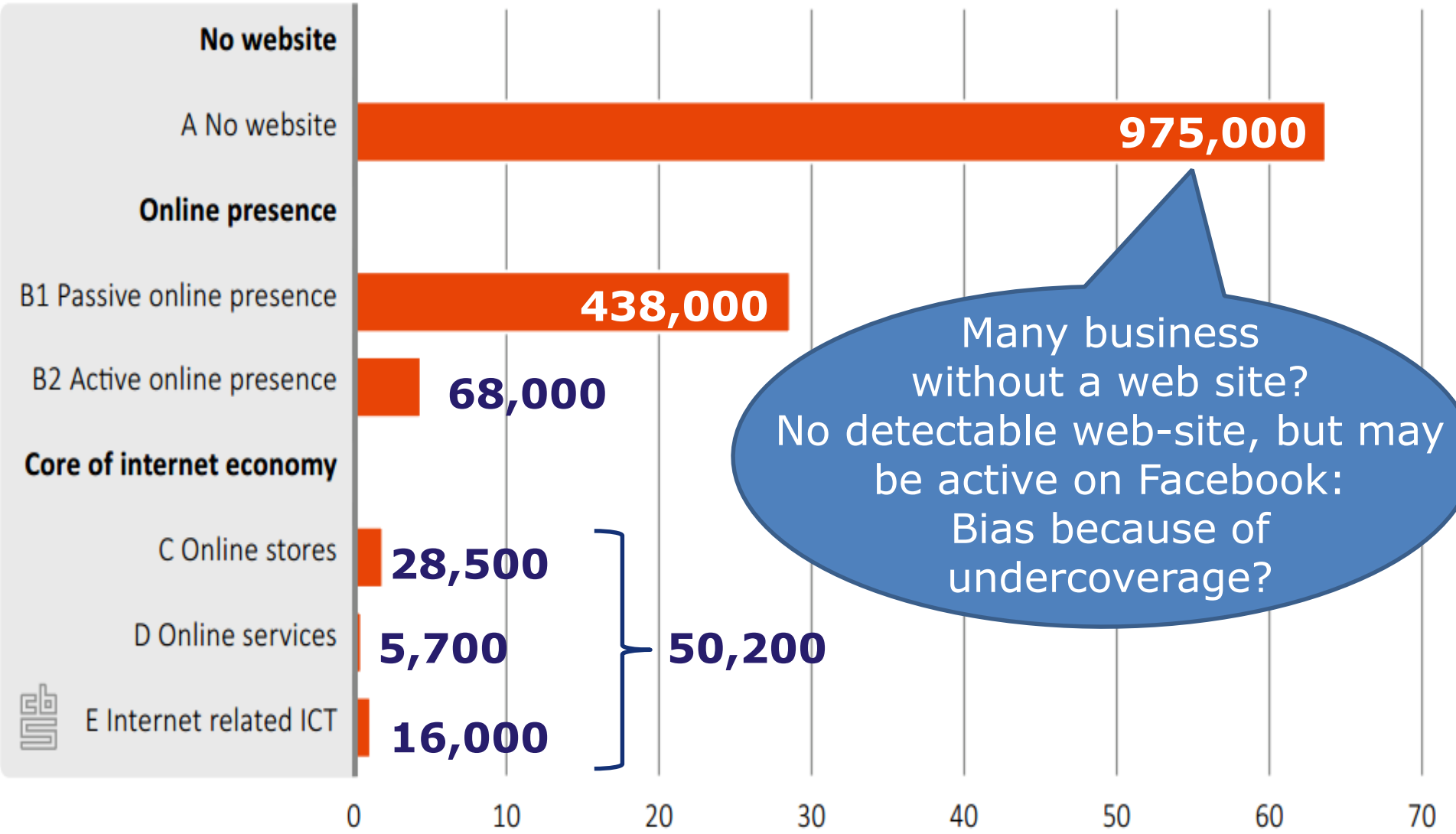
Online services

Category E: Income generated with the internet: Internet related ICT. **E**

- Web design
- Hosting
- Int. marketing

Internet related ICT

Number of business by internet category, 2015



Technological innovations

1. **Computerisation of survey data collection**
   a. Features of electronic questionnaires: off-line, web Q
   b. Paradata
2. **Computerisation of the business information chain**
   a. Electronic Data Interchange (EDI)
3. **Internet**
   a. Internet as data source
4. **Internet of Things (IoT):**
   Computerisation of business processes
   a. New data using EDI

Internet of Things (IoT)

Sensor data in businesses:

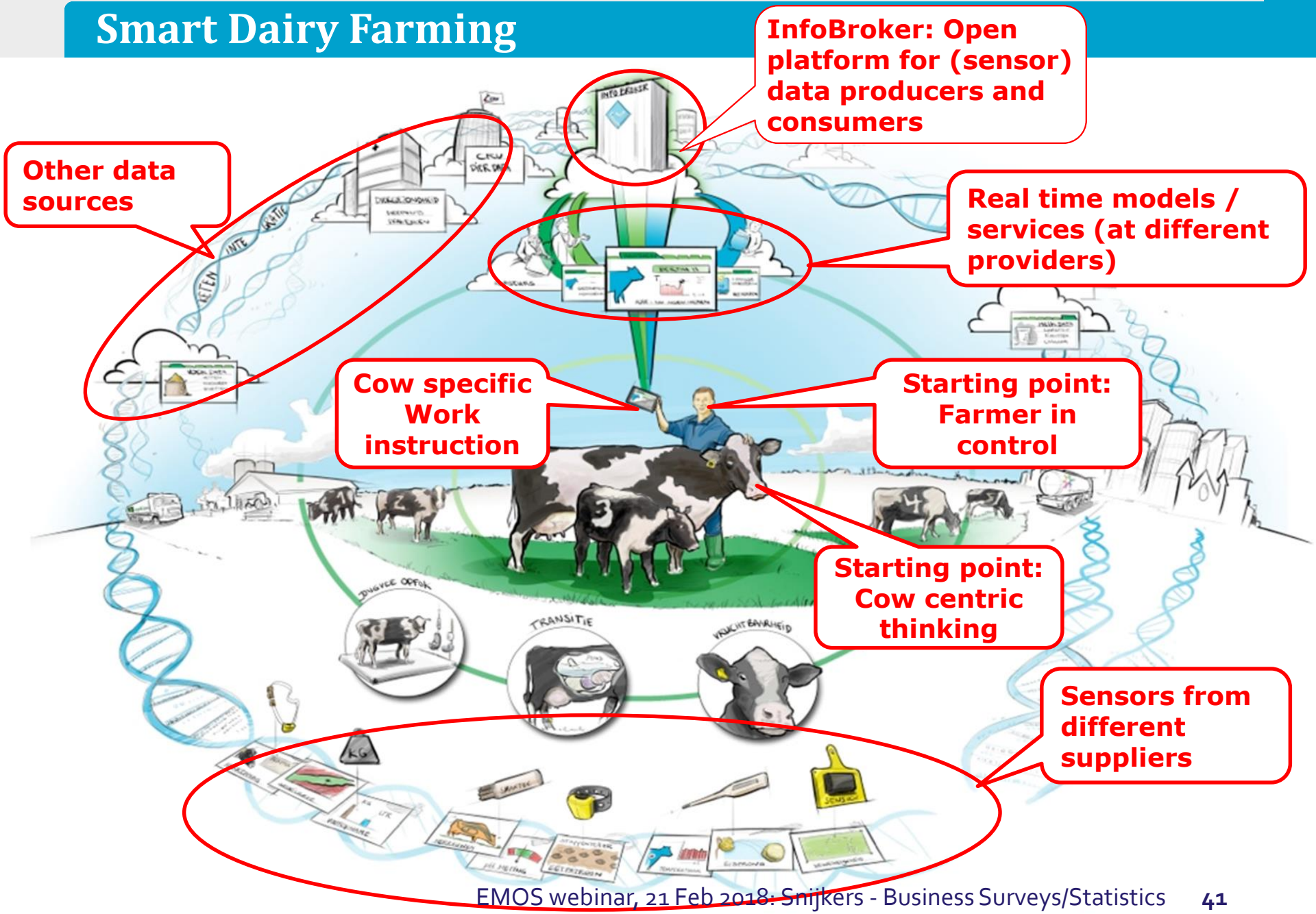
- Transportation: tracking packages
- Satellite images to estimate crop yields
- Precision or smart farming, like
  E.g. Smart Diary Farming

SmartDairyFarming

Vonder, M., 2017, Sensors going smart. Presentation at 'Big Data Matters' Seminar, Statistics Netherlands, 27 September 2017, Heerlen, Netherlands. (TNO Netherlands)

Technological innovations - 4

Smart Dairy Farming



Smart Dairy Farming

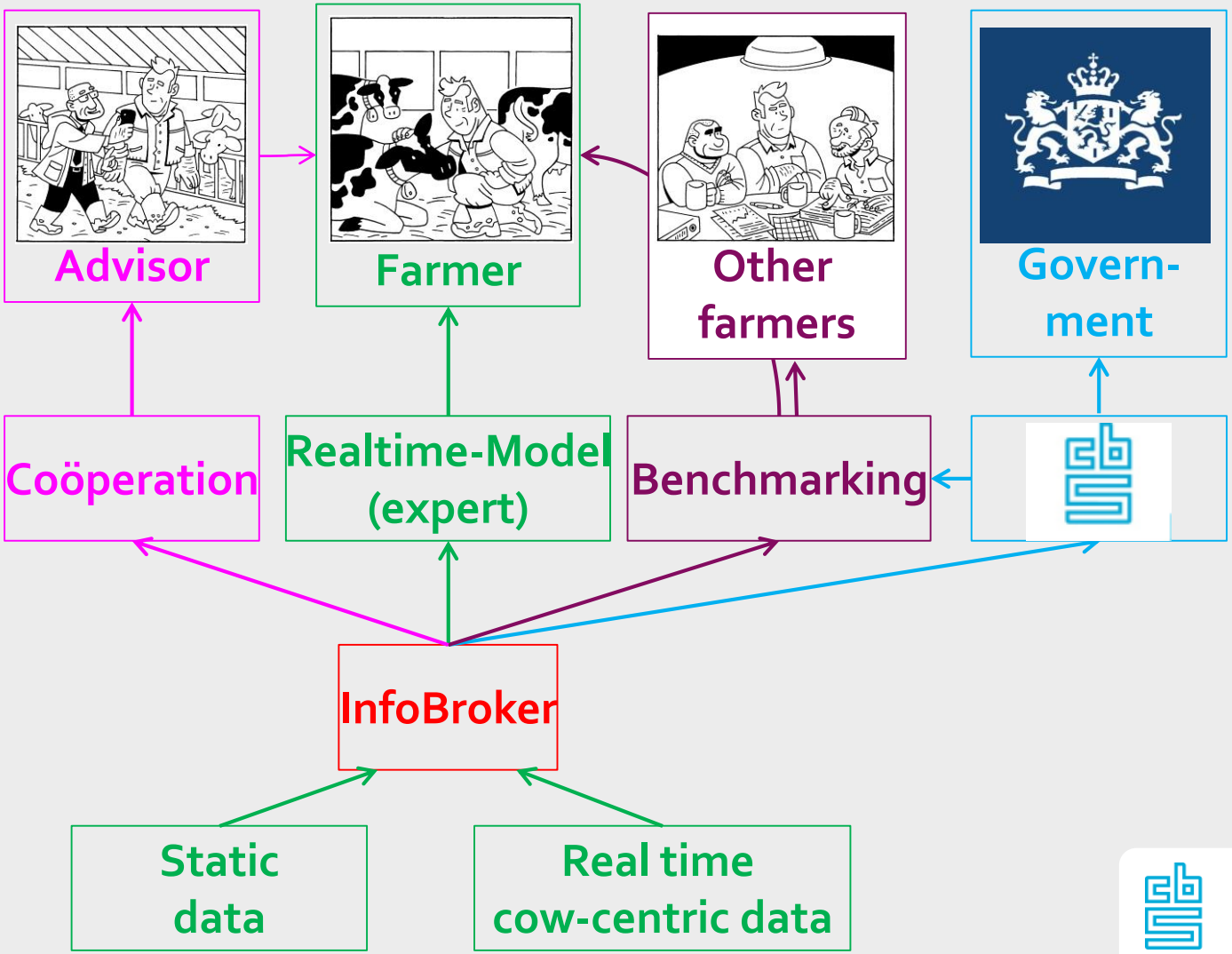
Sensor data on:

	Farm 1	Farm 2	Farm 3	Farm 4	Farm 5	Farm 6	Farm 7
# cows/calves	459	186	315	239	706	202	351
Behaviour	5x				5x		
Temperature	1x				1x		
Activity	9x	9x	3x	6x	5x	13x	9x
Milk production	16x	20x			1x	2x	19x
Feed intake	24x	24x				10x	24x
Weight	10x	6x	6x	6x	7x	6x	10x
Water intake			3x	3x			
Milk intake			7x	11x			

NB1: blue numbers are animals; not all animals are monitored for SDF (e.g. 3 and 4 only calves)
NB2: the left column gives a list of "sensor data categories" at a farm
NB3: numbers in black are the sensor fields within a category (e.g. 3 fields related to waterintake)

Smart Dairy Farming: InfoBroker

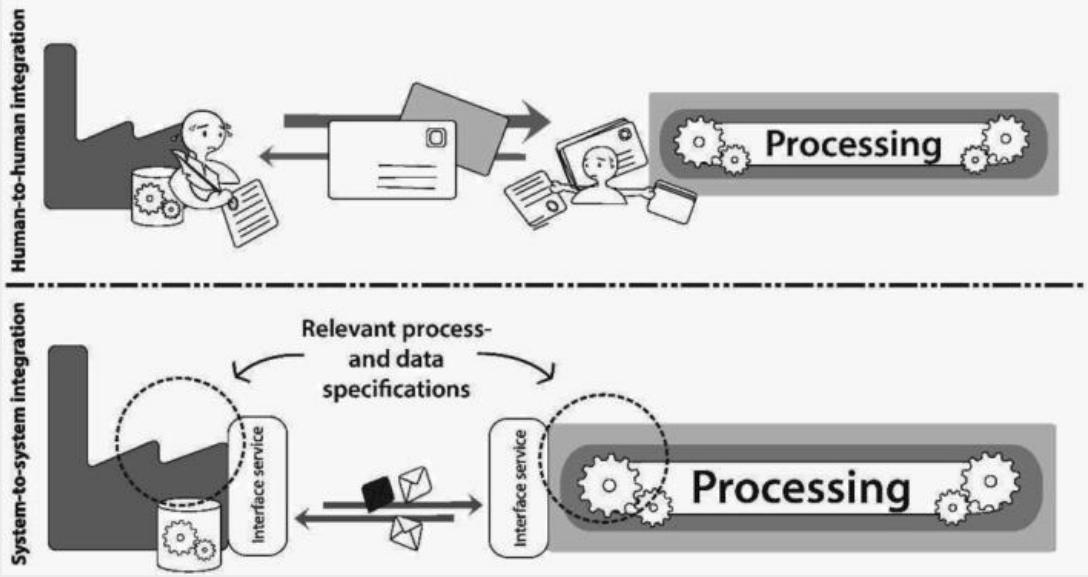
Some scenario's for using the InfoBroker:



Smart Dairy Farming

- Using EDI to collect sensor data instead of questionnaires
- Measurement and unit issues:
  - Harmonisation
  - Standardisation
  - Stability of data

  Like with EDI for financial data



- "Infobroker":
  - No central data base:
    - Location where data can be found
    - Reduces/prevents duplication of data
  - Instead of collecting data with individual businesses

Overview

- Background on Business data collection:
  - Characteristics of business survey data collection
  - History of business data collection
  - General Data collection strategy
- Technological innovations:
  1. Computerisation of survey data collection
  2. Computerisation of the business information chain
  3. Internet as data source
  4. Internet of Things (IoT)
- Conclusions and future developments

Conclusions

- **Many developments**
- **From data collection to collecting data**

  Business data collection will become more and more complex in the **age of data integration**:
  - Mixed-mode/multi-source approach
  - NSIs loose control over data collection process
  - From single-survey statistics to real-time integrated statistics
- Quality issues remain important

  (**Total Surver Error Framework**)

Methodological innovation: multi-source statistics production

- Quality evaluation of data:
  - Bias and selectivity issues: is the selection of units representative for the population as a whole?
    - Overcoverage and undercoverage issues
    - Revisiting the sample paradigm: non-probability sampling, representativeness
  - Measurement issues: do we get the data we want?
    - Definition of concepts
  - Unit issues: do we get the data from the right units?
- Linking and matching methods for all data sources: registers, survey data and big data
- Advanced statistical estimation methods combining all data sources

Organisational innovations

- From silo's to coordinated systems of data collection:
  - horizontal coordination
- Cultural shift:
  - NSIs are more and more out of control, but need to be in the forefront for new data sources
  - Trust in data from other sources
- Large-business Unit for complex and multi-surveyed enterprises:
  - Consistent communication
  - Consistent data across surveys
- Data lake:
  - Data repository of all data sources to facilitate data integration and statistics production

Future of Business Data Collection

- EDI for electronically available data, like financial data; sensor data (IoT):
  - For large, harmonised and stable data definitions
- Internet as data source
- Additional web surveys (tailored to the business context):
  - For additional data, e.g. on globalisation, out-sourcing
  - For complex businesses (unit issues) and complex data structures
  - Using feasibility studies to tailor the survey design
- Smart phone and tablets:
  - For small and simple questionnaires
  - For zero reporting

Statement

Technological innovations make things possible; technology is the enabler.
The applied methodology and the organisational context determine whether it will work.

Questions?

Thank you

References

- **Bharosa, N., R. van Wijk, N. de Winne, and M. Janssen (eds), 2015, Challenging the chain. Governing the automated exchange and processing of business information. IOS Press, Amsterdam (www.iospress.nl/book/challenging-the-chain/)**
- **Buiten, G., G. Snijkers, P. Saraiva, J. Erikson, A.-G. Erikson, and A. Born, 2018 (forthcoming), Business Data Collection: toward Electronic Data Interchange. Experiences in Portugal, Canada, Sweden, and the Netherlands with EDI. Journal of Official Statistics, ICES-V special issue.**
- Cialdini, R., 2006, Influence: the psychology of persuasion. HarperCollins Publishers.
- Cialdini, R., 2017, Pre-suasion; A Revolutionary Way to Influence and Persuade. Random House Business Publishers.
- Daas, P., S. van der Doef, and A. Hürriyetoglu, 2017, Text analysis at Statistics Netherlands. Internal presentation Statistics Netherlands, 4 October 2017, Heerlen.
- De Broe, S. 2017, Big Data Matters. Presentation at the 'Big Data Matters' Seminar, Statistics Netherlands, 27 September 2017, Heerlen, Netherlands. (Statistics Netherlands, Heerlen.)
- Eurostat, 2017, Smart Statistics. Paper discussed at the Joint Dime/ITDG plenary sessions, 14-15 February 2017. Eurostat. Luxembourg.
- Haraldsen, G., and M. Couper, 2013, How to design effective business web surveys. Short course, 14 August 2013, Bergen, Norway. (Statistics Norway, Oslo.)
- Rooijakkers, B., 2017, Measuring the internet economy with big data. Presentation at 'Big Data Matters' Seminar, Statistics Netherlands, 27 September 2017, Heerlen, Netherlands. (Statistics Netherlands, Heerlen.)
  www.cbs.nl/-/media/_pdf/2016/40/measuring-the-internet-economy.pdf?la=nl-nl

References

- **Snijkers, G. 2016, Achieving Quality in Organizational Surveys: An Holistic Approach. In Methodische Probleme in der empirischen Organisationsforschung, S. Liebig, and W. Matiaske (eds.), 33-59. Springer, Wiesbaden, Germany.**
- Snijkers, G., R. Göttgens, and H. Hermans. 2011, Data collection and data sharing at Statistics Netherlands: Yesterday, today, tomorrow. Paper presented at the 59th Plenary Session of the Conference of European Statisticians (CES): United Nations Economic Commission for Europe (UNECE), Geneva, June 14–16. Geneva: UNECE (Available at www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/20.e.pdf
- **Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D.K., 2013, Designing and Conducting Business Surveys. Hoboken, NJ: Wiley.**
- Snijkers, G., and M. Morren, 2010, Improving web and electronic questionnaires: The case of the audit trail. Paper presented at the 5th European Conference on Quality in Official Statistics, 4-5 May 2010, Helsinki, Finland.
- Thaler, R.H., and C.R. Sunstein, 2009, Nudge; Improving Decisions About Health, Wealth and Happiness. Penguin Books Ltd.
- Torres van Grinsven, V., and G. Snijkers, 2015, Sentiments and Perceptions of Business Respondents on Social Media: an Exploratory Analysis. Journal of Official Statistics Vol. 31, No. 2, pp. 283-304.
- Vonder, M, 2017, Sensors going smart. Presentation at 'Big Data Matters' Seminar, Statistics Netherlands, 27 September 2017, Heerlen, Netherlands. (TNO Netherlands)
- Zhang, L.-C., 2012, Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66(1): 41–63.