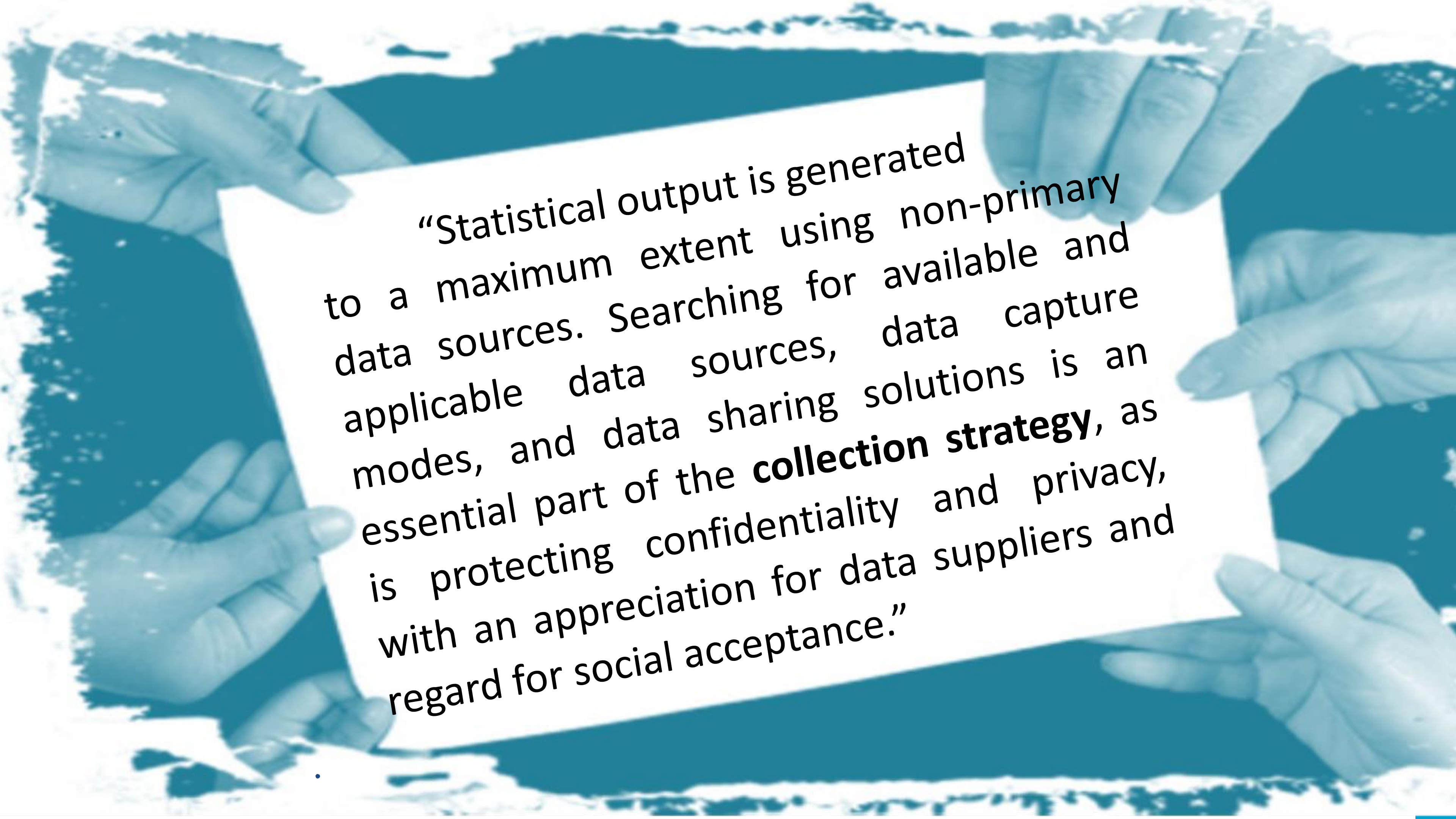
Future Advanced Data Collection

The Future is NOW

Irene Salemink
NPSO Innovatie dag 26 November 2019



“Statistical output is generated to a maximum extent using non-primary data sources. Searching for available and applicable data sources, data capture modes, and data sharing solutions is an essential part of the **collection strategy**, as is protecting confidentiality and privacy, with an appreciation for data suppliers and regard for social acceptance.”

Future data collection

- Strategic assignments for NSO’s
- Shaping the future by being smart
- Trends
- Challenges and constraints
- Future data collection
- ACTION
- Conclusions

Strategic assignments NSO’s

Strategic assignments NSO’s

- Demand driven and user centric
  - Broad range statistics/information
  - Fit-for-purpose
  - Accurate and timely
  - Policymakers
    - Private sector
      - Society

- Fact based policy making
  - Actionable Intelligence

Strategic assignments NSO’s

- Dealing with complexity
    - Complex societal and economic phenomena
    - Real time statistics
    - Tailored to all aggregation levels
    - More detail, regional, local
    - Fit –for-purpose quality

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

New information products

Discussion Paper

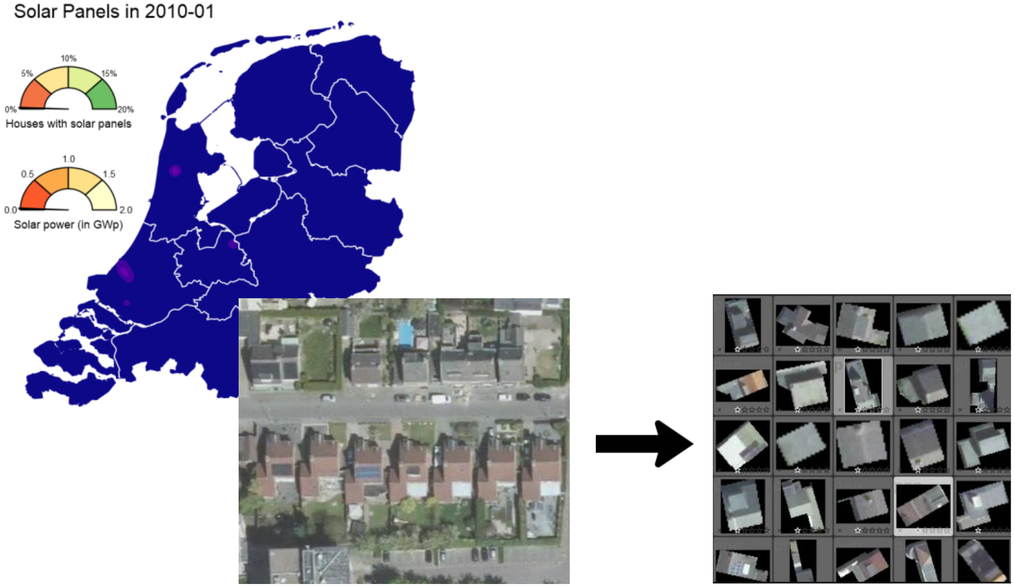
Measuring the internet economy in The Netherlands: a big data analysis

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2016 | 14

Lotte Oostrom
Adam N. Walker
Bart Staats
Magda Slootbeek-Van Laar
Shirley Ortega Azurduy
Bastiaan Rooijakkers

Complex phenomena

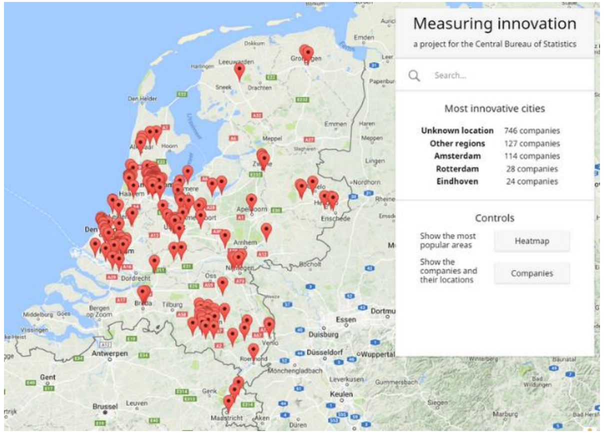


Energy transition

Support local government with insights to support sustainable development goals

Production Installation Register (solar panels)
VAT registers from Tax authority
Basic register Addresses and buildings

Complex phenomena



Measuring innovation

a project for the Central Bureau of Statistics

Search...

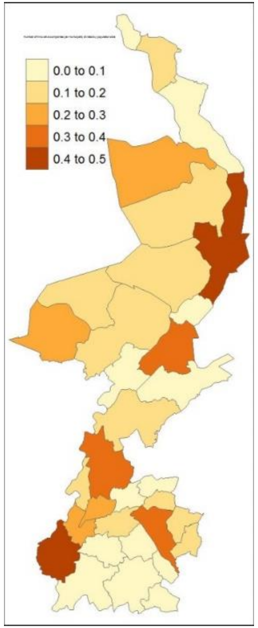
Most innovative cities

Unknown location	746 companies
Other regions	127 companies
Amsterdam	114 companies
Rotterdam	28 companies
Eindhoven	24 companies

Controls

Show the most popular areas — Heatmap

Show the companies and their locations — Companies



Innovative enterprises

Web scraping and text mining to identify small innovative enterprises

Classification
Linking to background characteristics

Shaping the future by being smart

Shaping the future by being smart

- Official statistics become “smart statistics”
  - Smart technologies
  - Smart data

- Guaranteed confidentiality

- Privacy by design

Trusted Smart Statistics

Trusted Smart Statistics

- Facts for evidence based policy making
- Quantitative monitoring of development and progress of policy
- Society oriented
- Reliable and innovative
- Protect confidentiality and privacy

Nature of data collection is bound to change

Trends

Trends

- Datafication

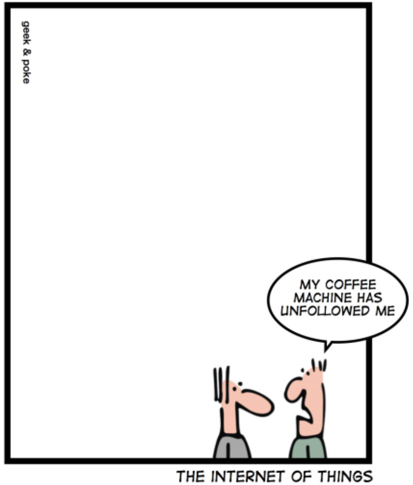
- Data storage and access

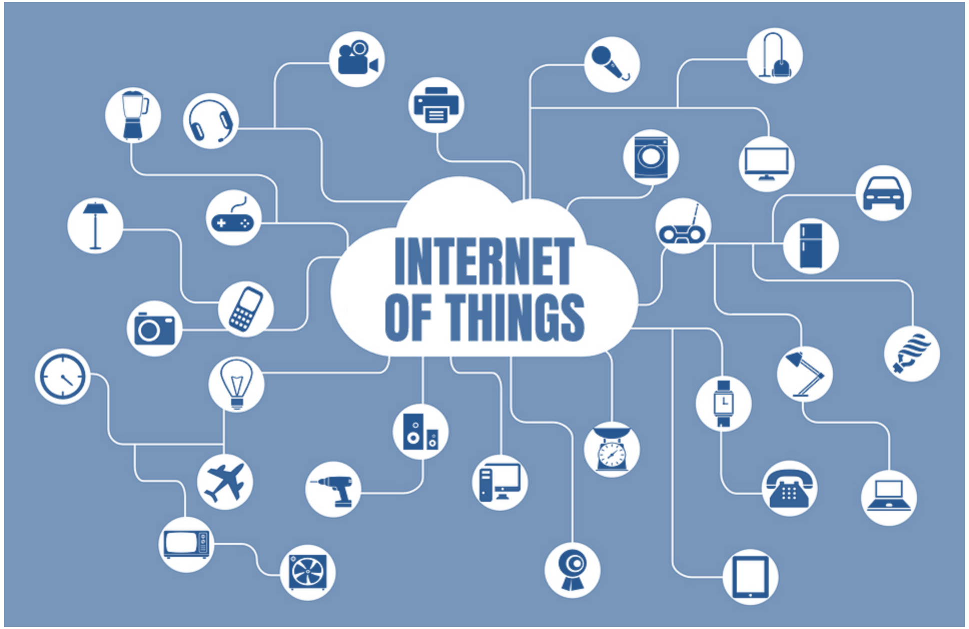
- Computing power and analytical capabilities

Administrative data use at CBS

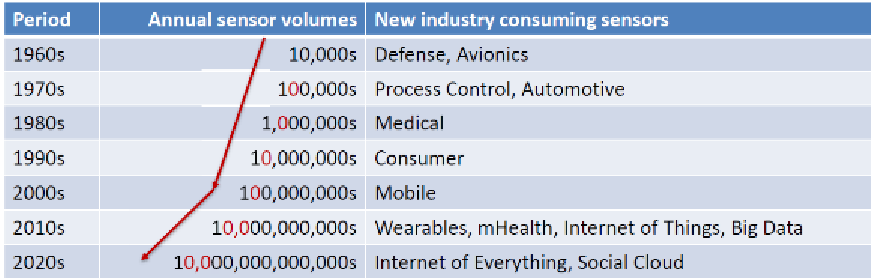
200+
registries

Sensors and IoT





We Live in Exponential Sensor Times



Period	Annual sensor volumes	New industry consuming sensors
1960s	10,000s	Defense, Avionics
1970s	100,000s	Process Control, Automotive
1980s	1,000,000s	Medical
1990s	10,000,000s	Consumer
2000s	100,000,000s	Mobile
2010s	10,000,000,000s	Wearables, mHealth, Internet of Things, Big Data
2020s	10,000,000,000,000s	Internet of Everything, Social Cloud

Mobile market accelerated
the sensor growth by an
order of magnitude

Data storage and access

- Need for :

  Fast, easy, and free access to all relevant data

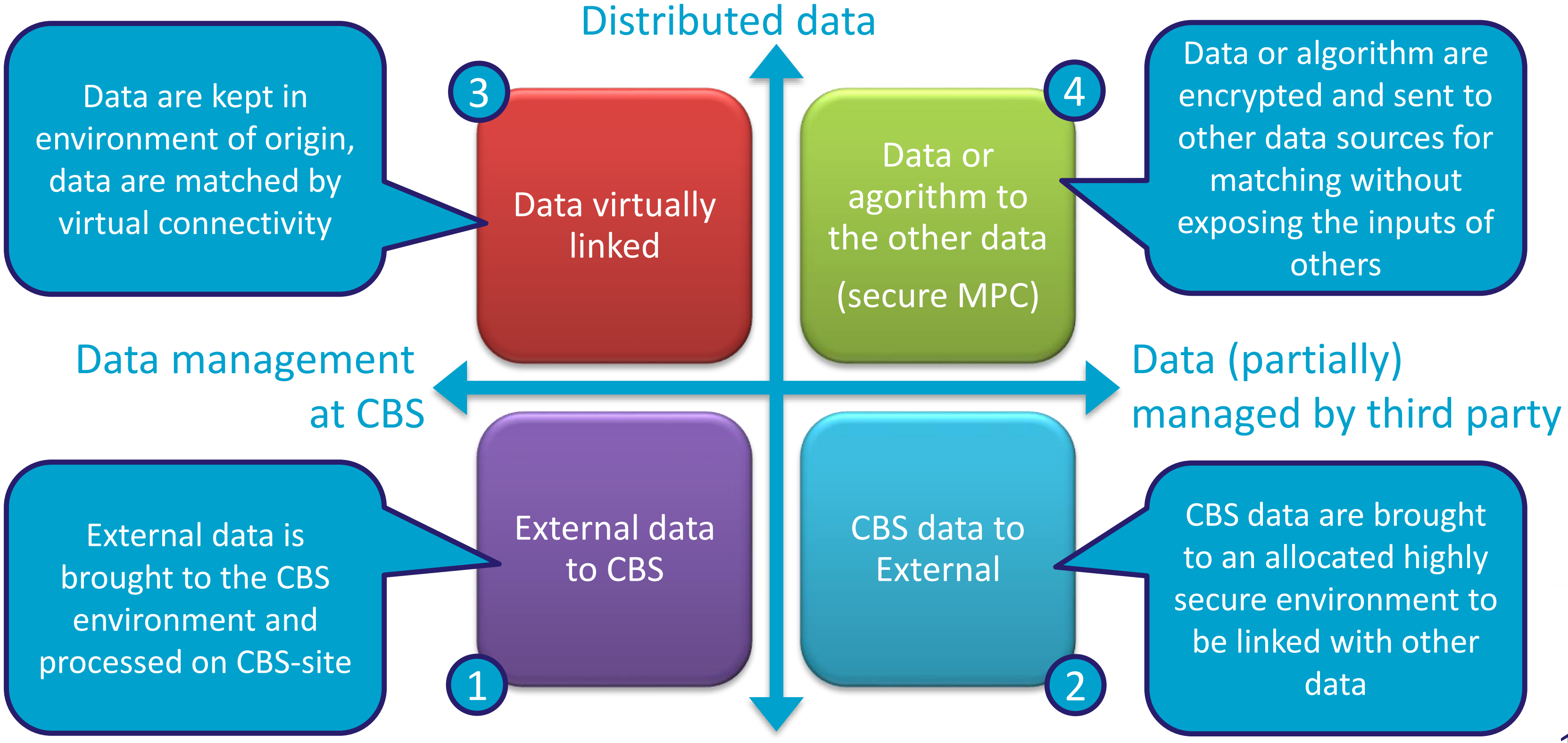
- Reality:

  Datasets; too big to copy, not allowed legally to “leave the building”, need matching between multiple (different) sources, require knowledge, only the proportion that is needed can be accessed…

- Solution:

  Dependent on type of data sharing

Data architecture patterns



Distributed data

Data are kept in environment of origin, data are matched by virtual connectivity

3

Data virtually linked

4

Data or agorithm to the other data

(secure MPC)

Data or algorithm are encrypted and sent to other data sources for matching without exposing the inputs of others

Data management at CBS

Data (partially) managed by third party

External data is brought to the CBS environment and processed on CBS-site

1

External data to CBS

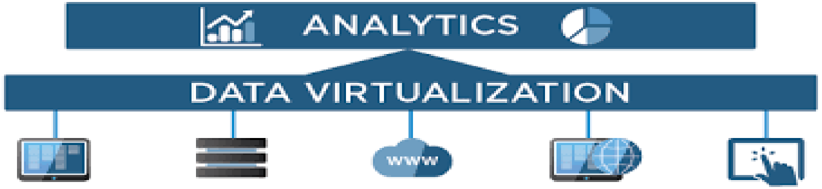
2

CBS data to External

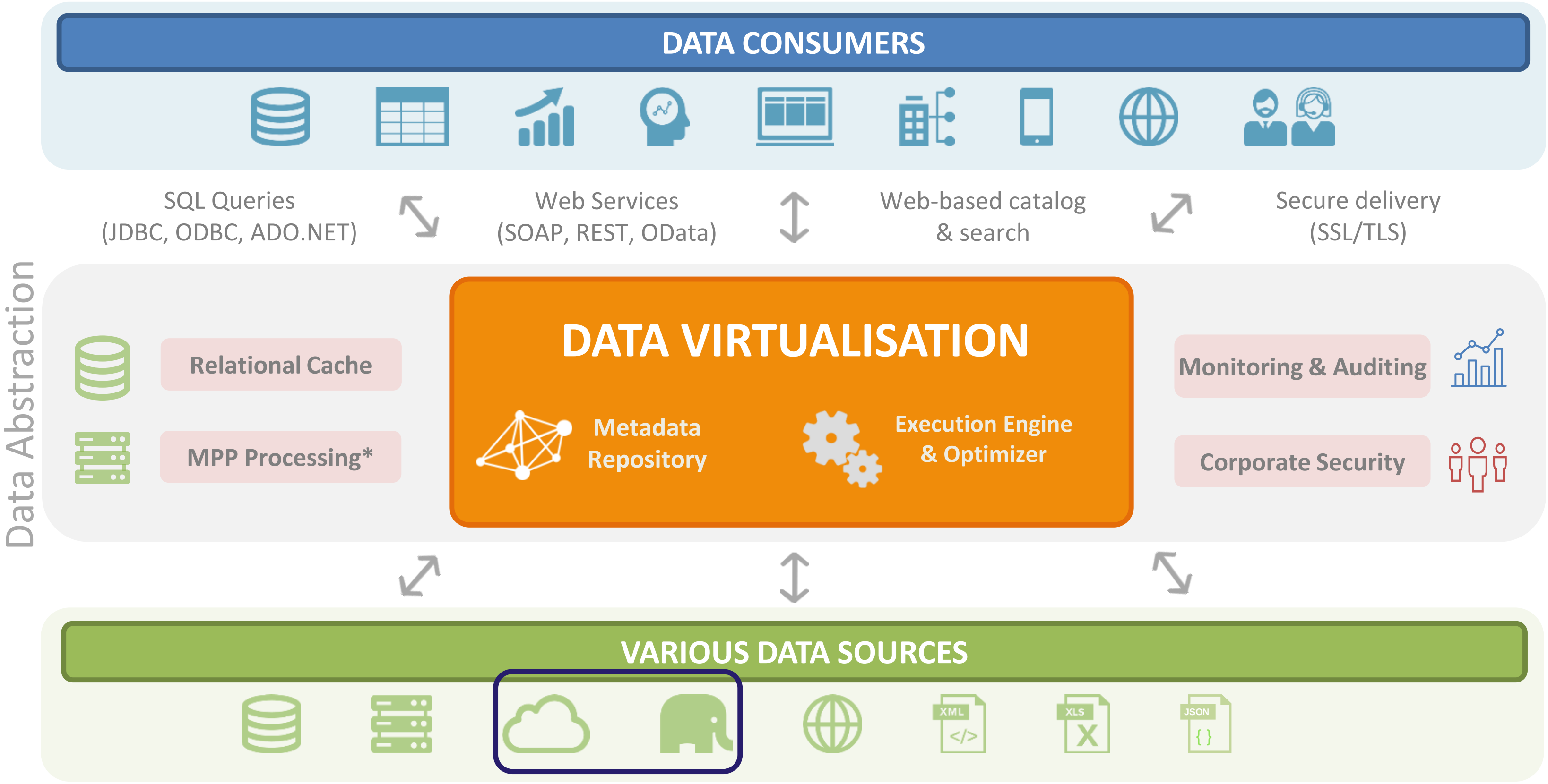
CBS data are brought to an allocated highly secure environment to be linked with other data

Data storage and access

- These 4 patterns come with capabilities that need further investigation
    - Privacy preserving analytic techniques
        - Secure multi party computation
    - Data virtualisation and data abstraction
    - Metadata management



Data virtualisation architecture



DATA CONSUMERS

SQL Queries
(JDBC, ODBC, ADO.NET)

Web Services
(SOAP, REST, OData)

Web-based catalog
& search

Secure delivery
(SSL/TLS)

Data Abstraction

Relational Cache

MPP Processing*

DATA VIRTUALISATION

Metadata
Repository

Execution Engine
& Optimizer

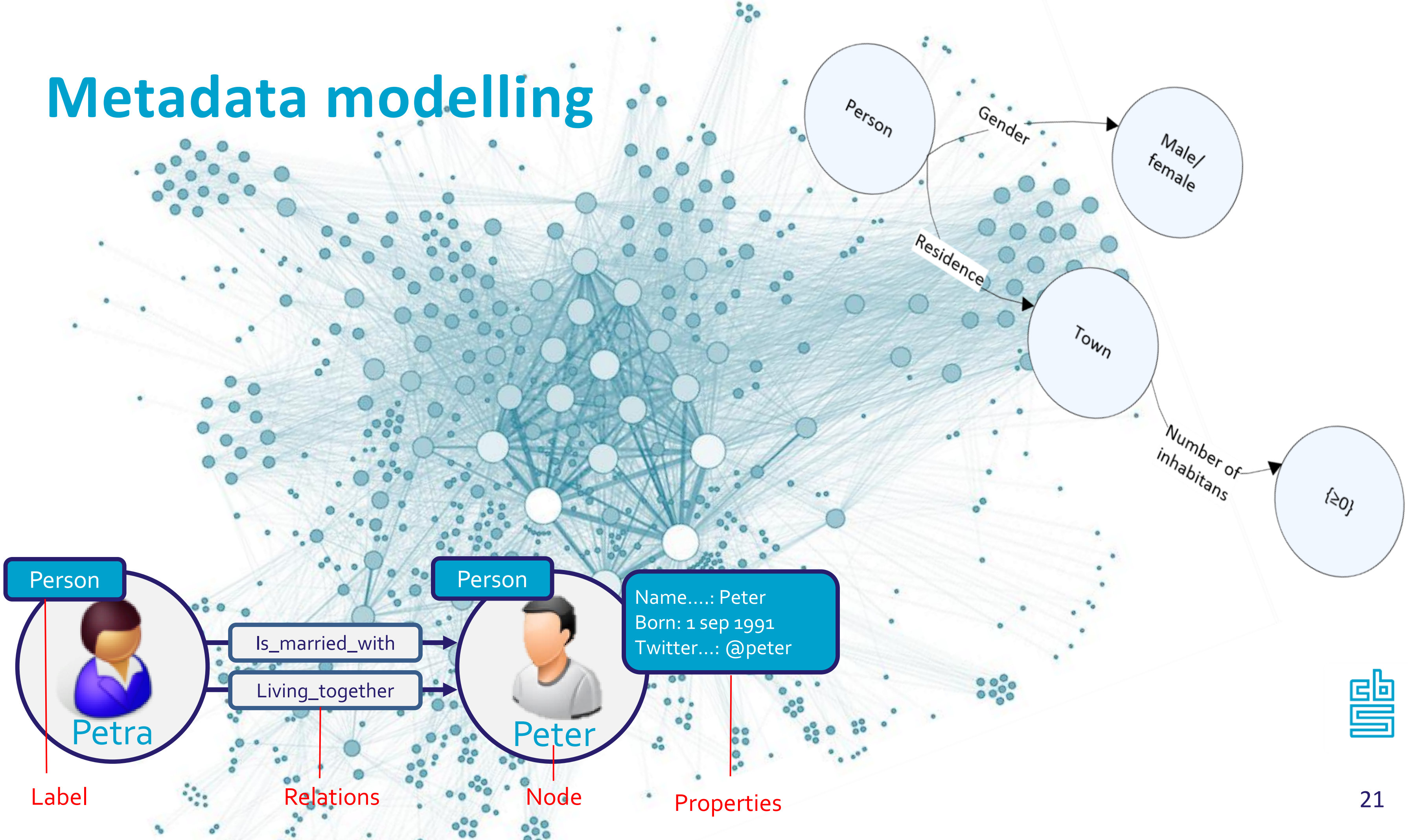
Monitoring & Auditing

Corporate Security

VARIOUS DATA SOURCES

* Massive Parallel Processing

Metadata modelling



Name....: Peter
Born: 1 sep 1991
Twitter...: @peter

Computing power and analytics

- Explore and Confirm
  - Deductive; data analysis to explain, check or validate ideas
  - Inductive; data analysis to generate new ideas

- Edge analytics
  - Analysis and data quality framework are brought to the data gathering devices instead of moving the data to the (centralized) analytics and quality frameworks

Challenges and constraints

Challenges

- Information Gap due to a Data gap

- Burden to society and response rates
- Privacy protection and difficult acces tot data tot data

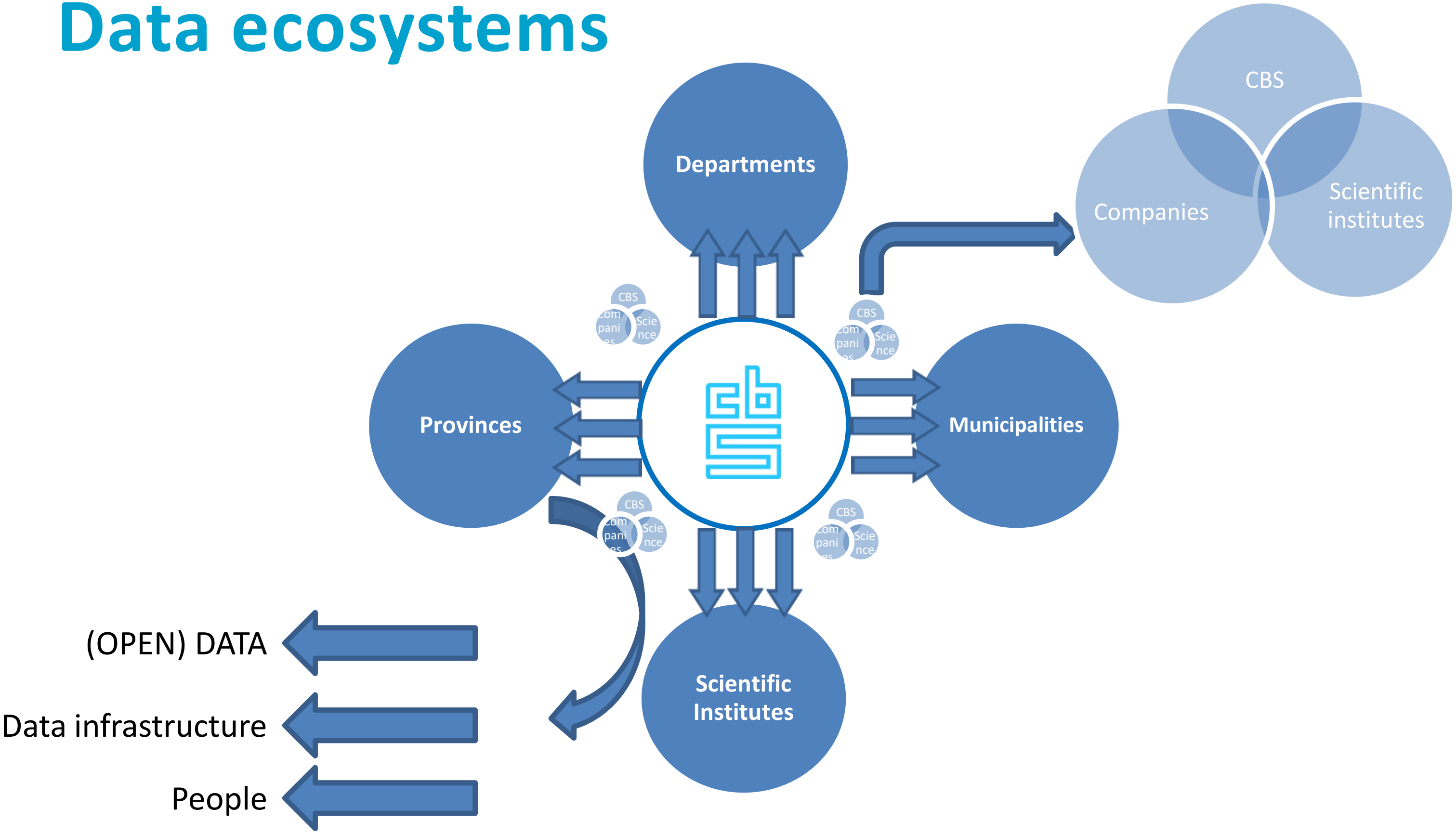
- Methodology

Future data collection

CBS in relation to data owners & end users

- Signalling trends and potential future needs
- Meet the public demand for information
- Data hub
- CBS fulfils role of a platform
- Enabling broad cooperation between
    - Governmental bodies
    - Municipalities
    - Companies
    - Scientific institutes

Data ecosystems



Departments

Provinces

Municipalities

Scientific Institutes

CBS

Companies

Scientific institutes

(OPEN) DATA

Data infrastructure

People

Tapping into and unlocking new data sources

- Datascouting
  - Awareness of data sources being there
  - Usability and availability of the data source
  - Organization and facilitation of the acquisition, tapping into and unlocking of new data sources

- Data scout en community
  - Link between internal and external stakeholders

Surveys – primary data collection

- CAWI hybrid mode
- Multi mode and multi source
- Custom-fit data collection*
- Experimenting
  - Crowdsourcing
  - Interactive Voice Response
- Validation of alternative sources / data

Call to ACTION

Call to action

- Methodology
- Collect – Connect – Link
- Metadata
- Proprietary sensor networks
- Legal frameworks and social acceptability

Methodology on combining data

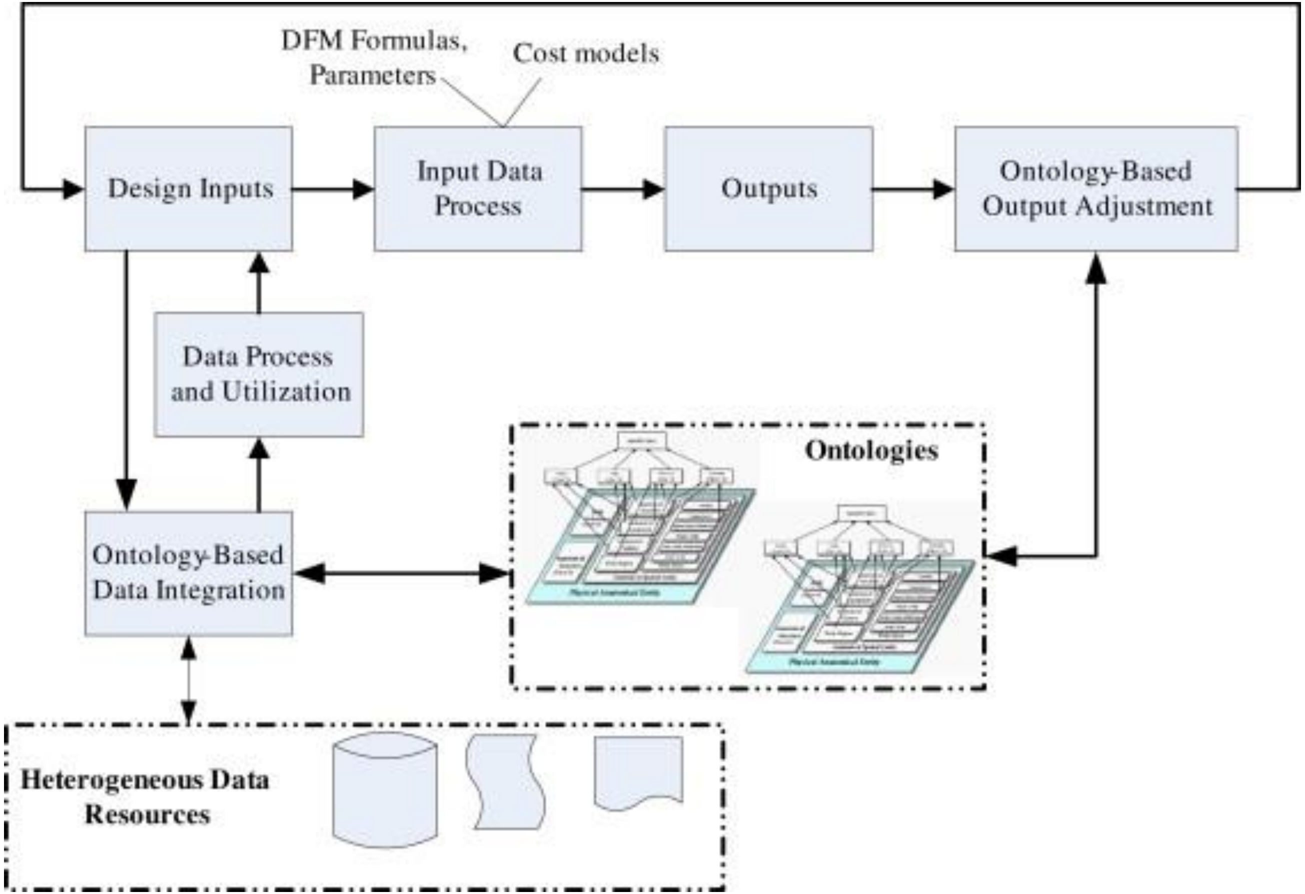
- Multiple collection modes and different data sources
    - units to be matched do not equal source units
    - sources do not contain overlapping units
    - matching errors
    - variables in multiple sources with different measurement errors

- Complex mixed mode designs

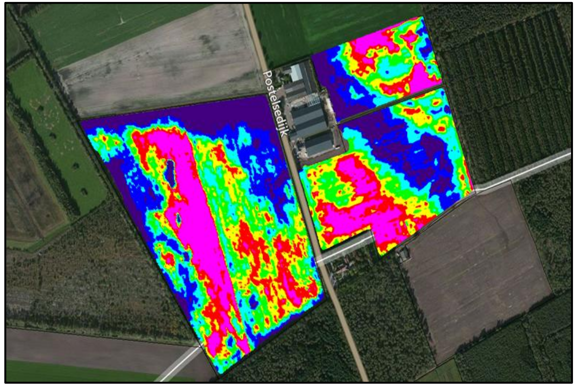
- Extent, combine and/or renew existing techniques
    - Probabilistic matching
    - Matching with supervised machine learning
    - Synthetic matching

- Integration by design

Integration by design

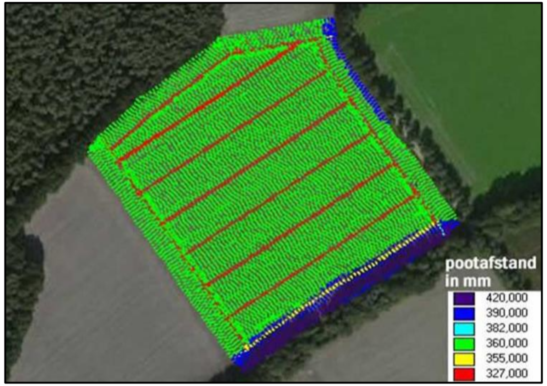

DFM Formulas,
Parameters
Cost models
Design Inputs
Input Data Process
Outputs
Ontology-Based Output Adjustment
Data Process and Utilization
Ontologies
Ontology-Based Data Integration
Heterogeneous Data Resources

Sensor networks - precision agriculture cycle



Winter

- Draw parcels
- Yield potential
- Tractor lanes

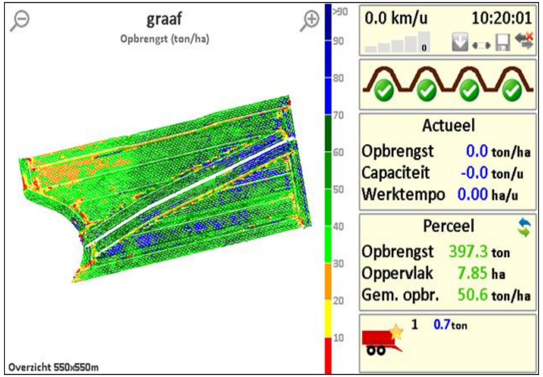


Spring

- Fertilization
- Variable planting

Summer

- Additional fertilization, pesticides & water
- Based on sensordata



0.0 km/u 10:20:01

Actueel
Opbrengst 0.0 ton/ha
Capaciteit -0.0 ton/u
Werktempo 0.00 ha/u

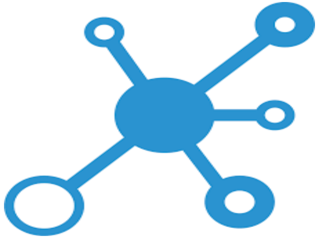
Perceel
Opbrengst 397.3 ton
Oppervlak 7.85 ha
Gem. opbr. 50.6 ton/ha

1 0.7 ton

Autumn

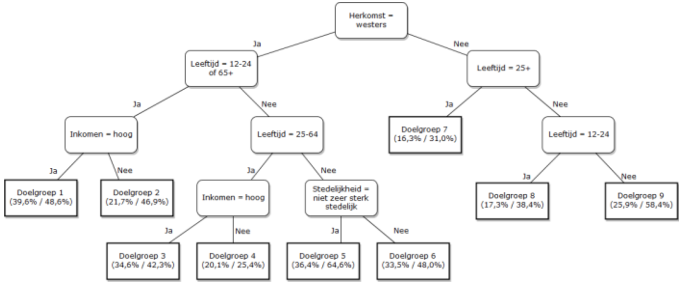
- Harvest
- Storage

…and also….



BIG
DATA
ANALYTICS
STORAGE

□□□□□


Herkomst = westers
Ja
Nee
Leeftijd = 12-24 of 65+
Leeftijd = 25+
Ja
Nee
Ja
Nee
Inkomen = hoog
Leeftijd = 25-64
Doelgroep 7 (16,3% / 31,0%)
Leeftijd = 12-24
Ja
Nee
Ja
Nee
Ja
Nee
Doelgroep 1 (39,6% / 48,6%)
Doelgroep 2 (21,7% / 46,9%)
Inkomen = hoog
Stedelijkheid = niet zeer sterk stedelijk
Doelgroep 8 (17,3% / 38,4%)
Doelgroep 9 (25,9% / 58,4%)
Ja
Nee
Ja
Nee
Doelgroep 3 (34,6% / 42,3%)
Doelgroep 4 (20,1% / 25,4%)
Doelgroep 5 (36,4% / 64,6%)
Doelgroep 6 (33,5% / 48,0%)

Conclusions

- Influence of NSI on content of captured data diminishes
- Availability and technology greatly determine the data to be collected
- New sources increase complexity
    - Validation, measuring errors & biases, mixed modes and multiple sources are a challenge
- AND possibilities
    - New – More – Cheaper – Faster

Conclusions

- Data collection becomes Advanced
- Data collection through data connection
- Statistical production process has to follow
- Integration by design

Vision paper can be found at

https://www.cbs.nl/en-gb/uitgelicht/statistics-netherlands-at-isi-wsc-2019

Facts that matter