Key challenges in the management of Big Data

for the institution and the individual

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE

Barteld Braaksma

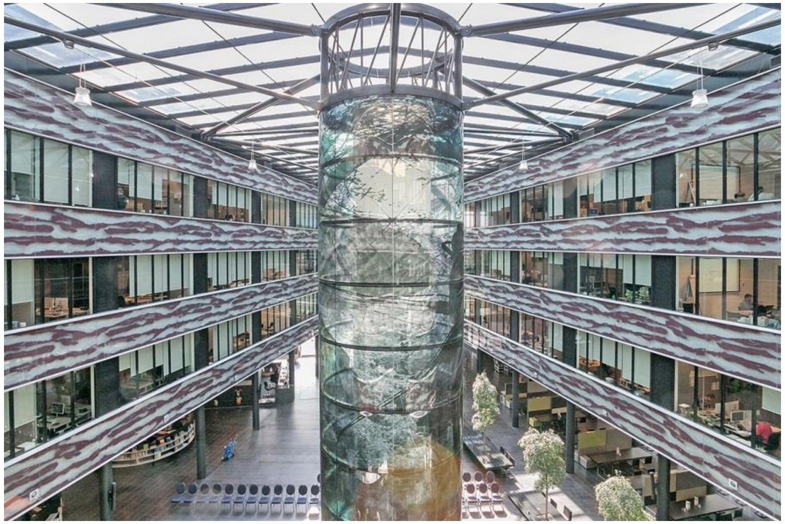
25 October 2018

Centraal Bureau voor de Statistiek (CBS)

- Established 1899, 5 employees at Binnenhof
- 1949-1973: 23 locations in The Hague
- 1973-2008: Voorburg
- 2008-now : Leidschenveen
- 1978-now : Heerlen

- Since 10-10-10: Bonaire

Two (three) offices- or sixteen?



- 1982: 3600 employees
- 2018: 1800 employees

Most employees now with higher (HBO/WO) education

Urban Data Center

Departmental data center (2x)

Rural data center

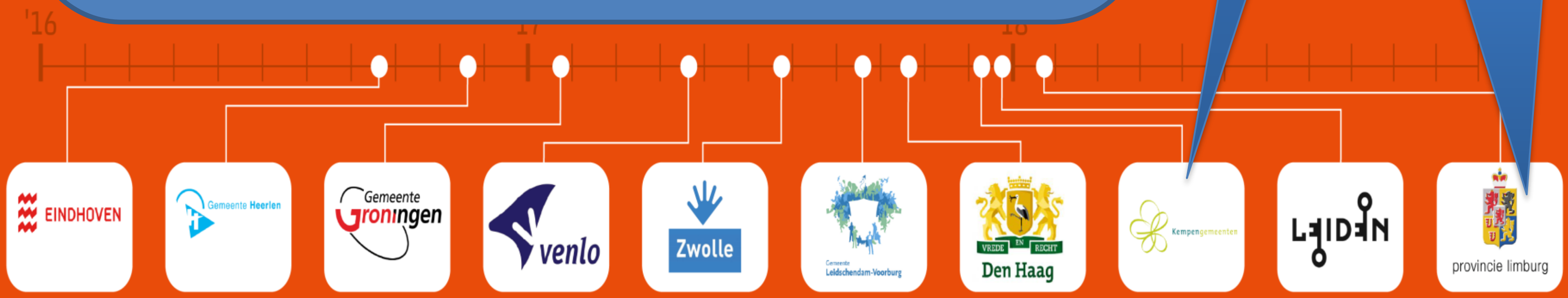
Academic data center

Provincial data center

CBS has started to create Urban Data Centers

To connect CBS data and CBS data-expertise to cities leading to:

- ***a better understanding of a city***
- ***better informed city decisions***
- ***better city finances***
- ***harmonized and standardized data (local – regional – national – international)***



A statistical treasure chest

Statistics Netherlands (CBS) enables people to have debates on social issues on the basis of reliable statistical information

Statistical results next to data on individual persons, companies, etcetera

Confidentiality protected by law (GDPR and CBS Act)

CBS products and services

Traditional

- ✓ Regular official statistics (annual, quarterly, monthly, …)
- ✓ Press releases and other publications
- ✓ Dissemination to EU, IMF, OECD, UN, …
- ✓ Machine-readable open data
- ✓ Microdata access for scientific purposes
- ✓ Commissioned statistics
- ✓ Applied research
- ✓ Statistical software
- ✓ Statistical training

CBS was elected best government organisation of the year 2015

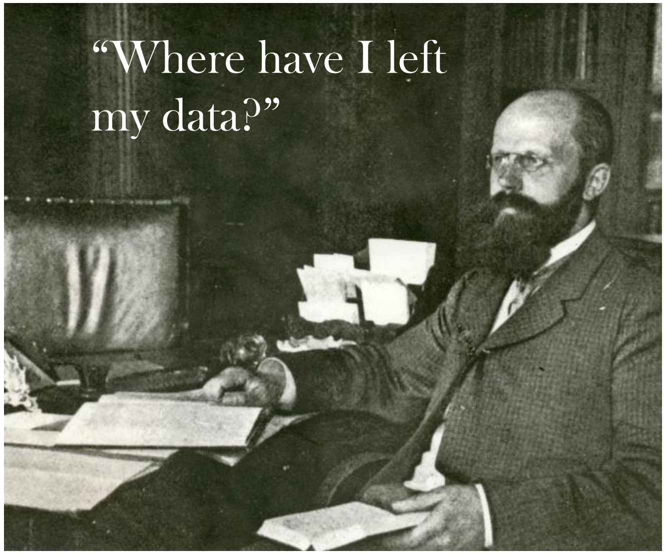
New and under construction

- ✓ Urban Data Centers
- ✓ Statistical consultancy
- ✓ Third party data services
- ✓ Experimental statistics
- ✓ Data camps and hackathons
- ✓ CBS Academy
- ✓ Information dialogue

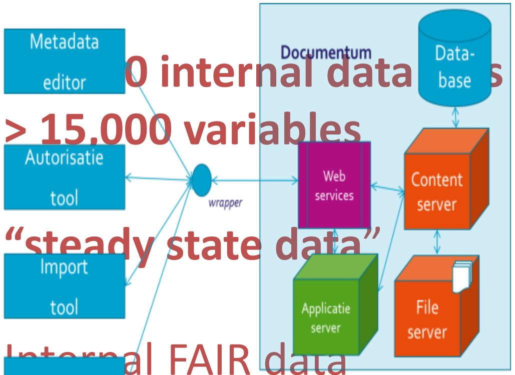
BESTE 2015
OVERHEIDS
ORGANISATIE
VAN HET JAAR

Data Service Centre (DSC)

for internal data management



“Where have I left my data?”

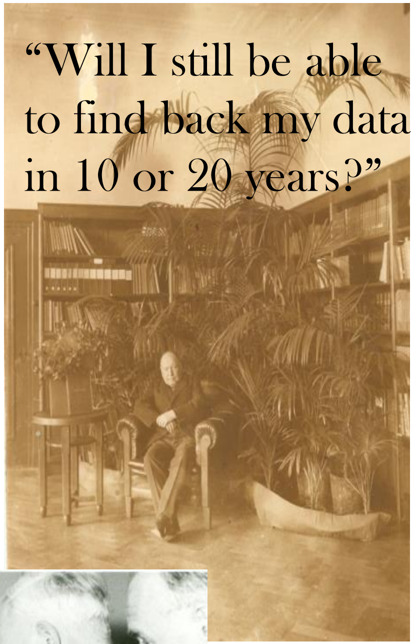


[illegible]0 internal data[illegible]s

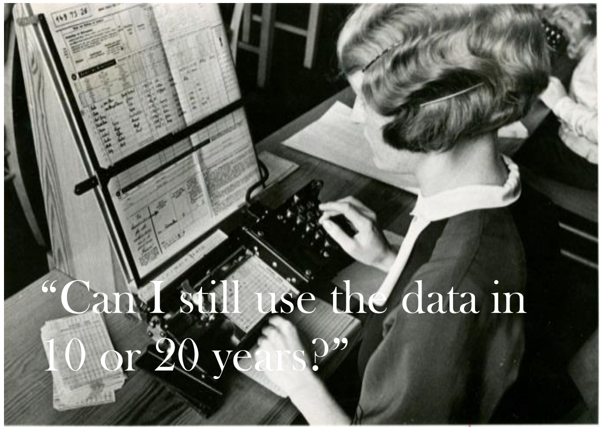
> 15,000 variables

“steady state data”

Internal FAIR data *avant la lettre*



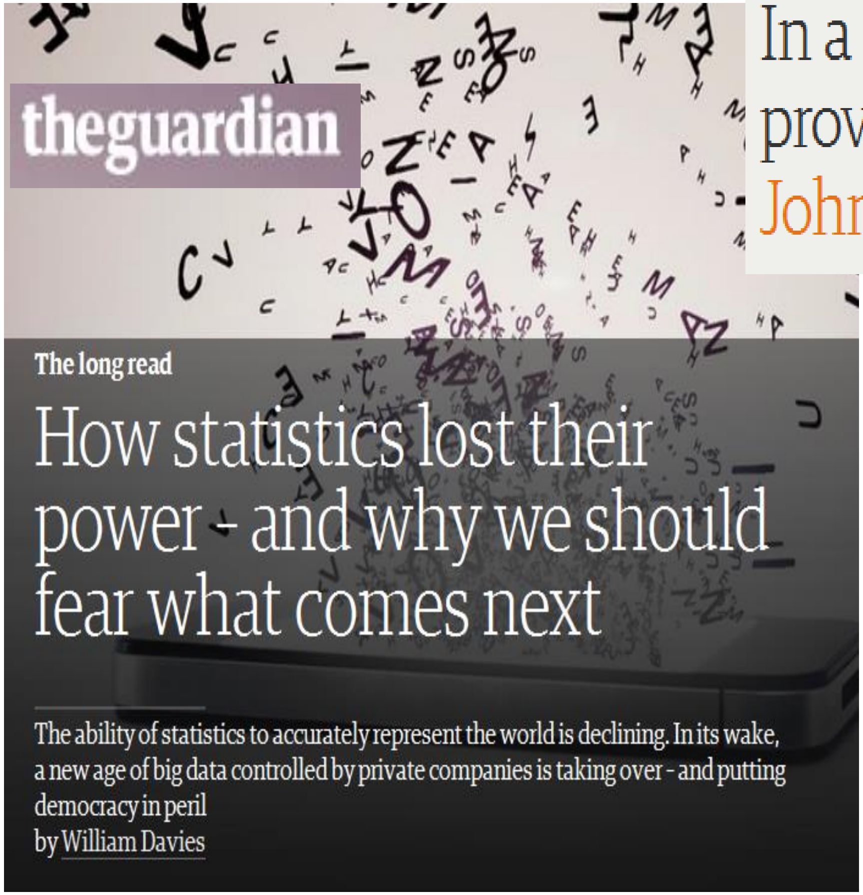
“Will I still be able to find back my data in 10 or 20 years?”



“Can I still use the data in 10 or 20 years?”

“Can ik control and manage access to the data?”

But: is the time of statistics over?



theguardian

The long read

How statistics lost their power - and why we should fear what comes next

The ability of statistics to accurately represent the world is declining. In its wake, a new age of big data controlled by private companies is taking over - and putting democracy in peril

by William Davies

In a post-truth world, statistics could provide an essential public service

John Pullinger

(National Statistician UK)

By combining the best of both worlds

Statistics Netherlands

https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

Data, data everywhere

Information has gone from scarce to superabundant.

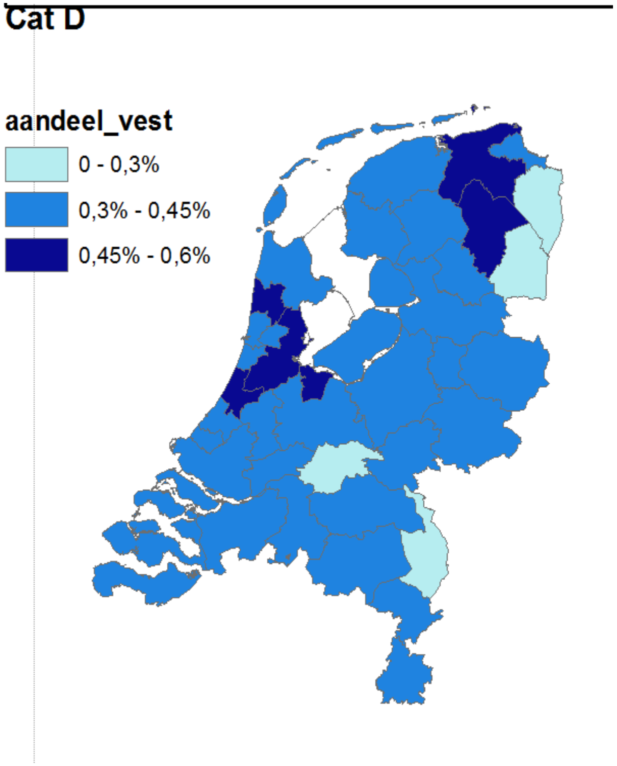
The Economist

CBS answer: Center for Big Data Statistics

Center for
Big Data Statistics

- Academic, public, private partners
- Both national and international
- Various relationships

Measuring the internet economy
in The Netherlands



Main research question:

“What is the importance of the internet to the Dutch economy?”

The aim of the research project was fourfold:

1. Determine a pragmatic definition of “the internet economy”
2. Show the importance and size of the internet economy in NL
3. Show the possibilities of new measurement methods
4. Explain differences from regular statistics/concepts

Center for
Big Data Statistics

Google

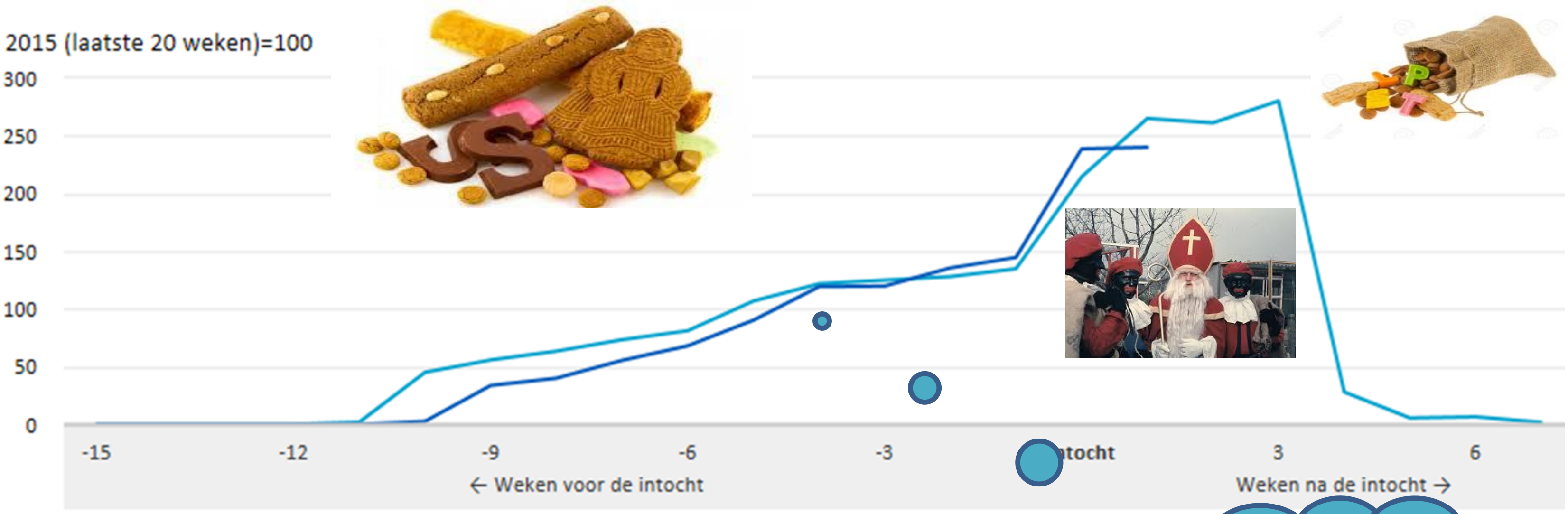
dataprovider

2.5 million Dutch websites
linked to business register

The Peppernut Index

seasonal cookies and candy

Omzetindex van sinterklaassnoepgoed in supermarkten

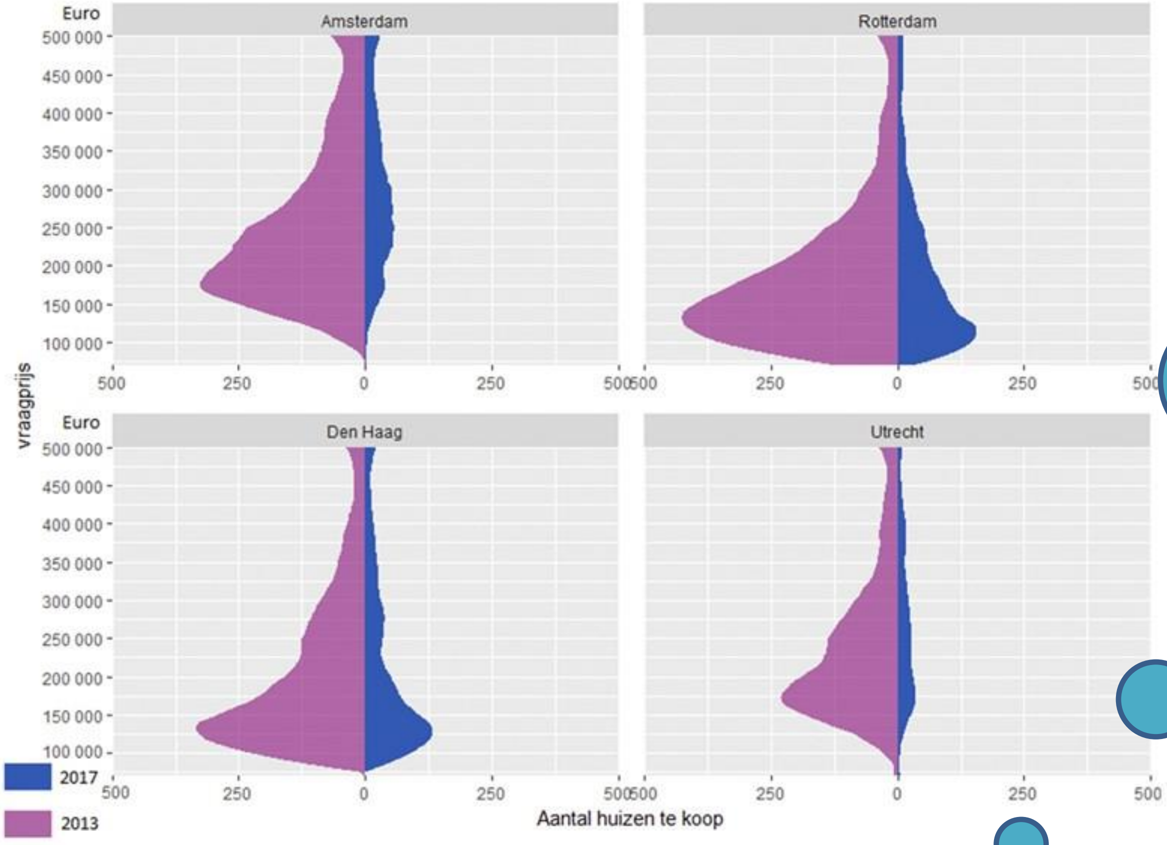


Could we say something about regional food patterns?

1234567890

Center for
Big Data Statistics

The Dutch online housing market



How to integrate such an API in the CBS environment?

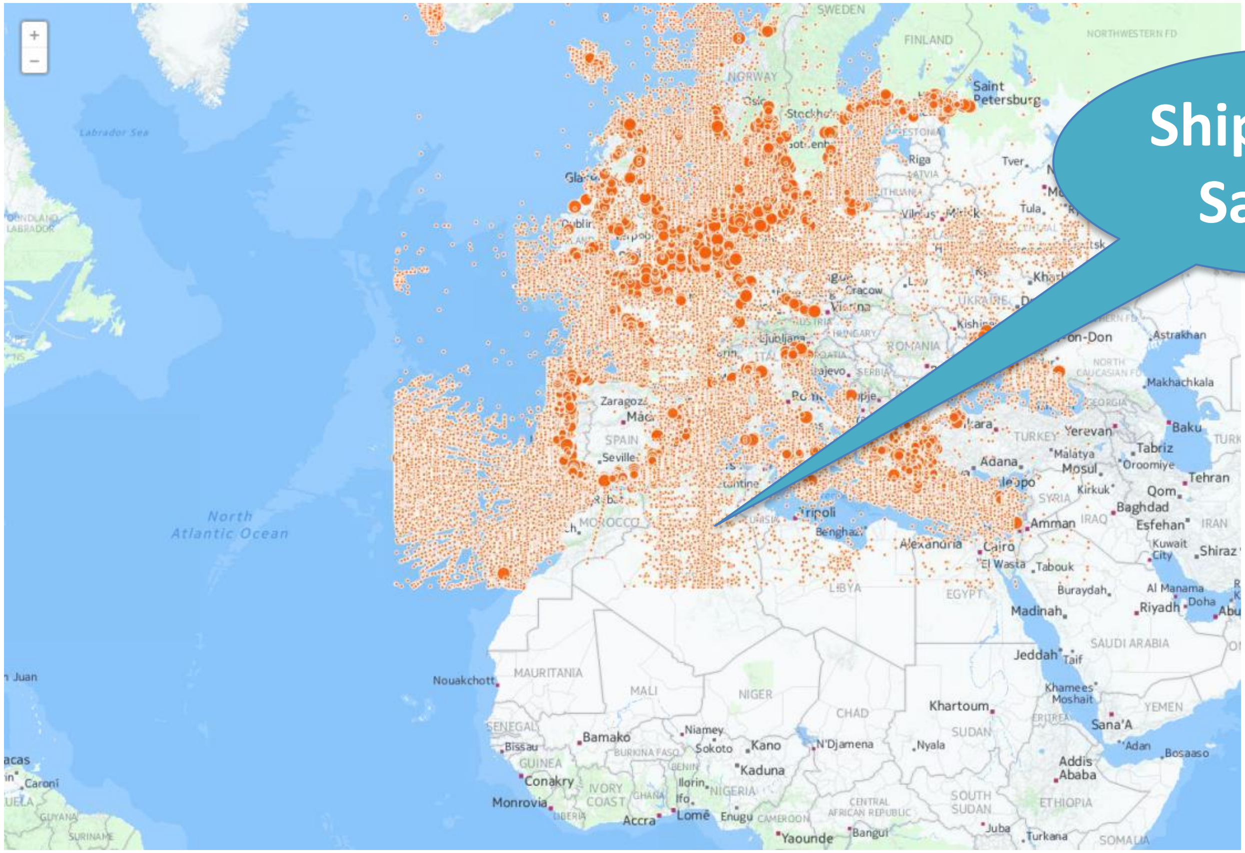
Number of houses for sale online, 2013+2017

Source (API): JAAP.NL ALLES OVER HUIZEN |data services|

Center for Big Data Statistics

What about quality of big data?

AIS data from ships

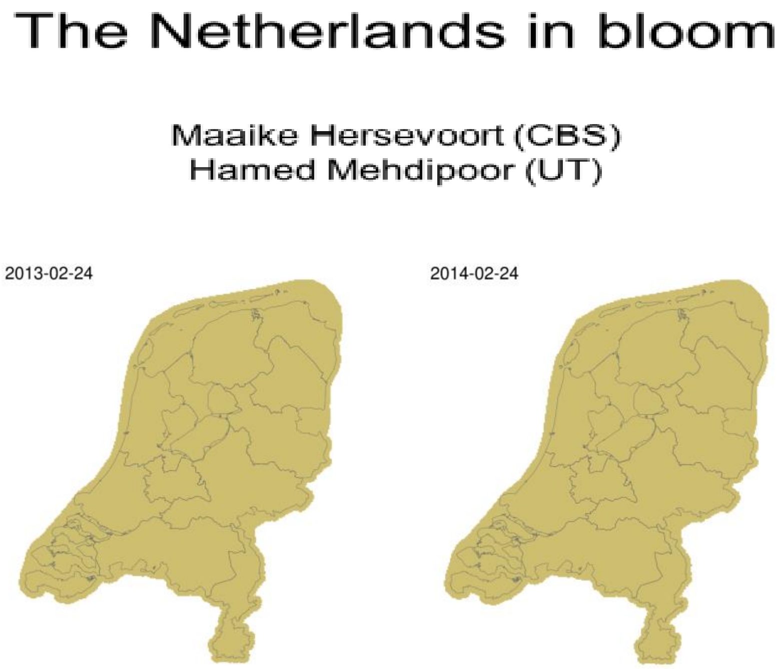


Ships in the Sahara?

Automatic identification system
Worldwide standard using GPS-like sensors
For most ships obligatory

Innovative collaboration models:

CBS-UTwente Data Camp 2015

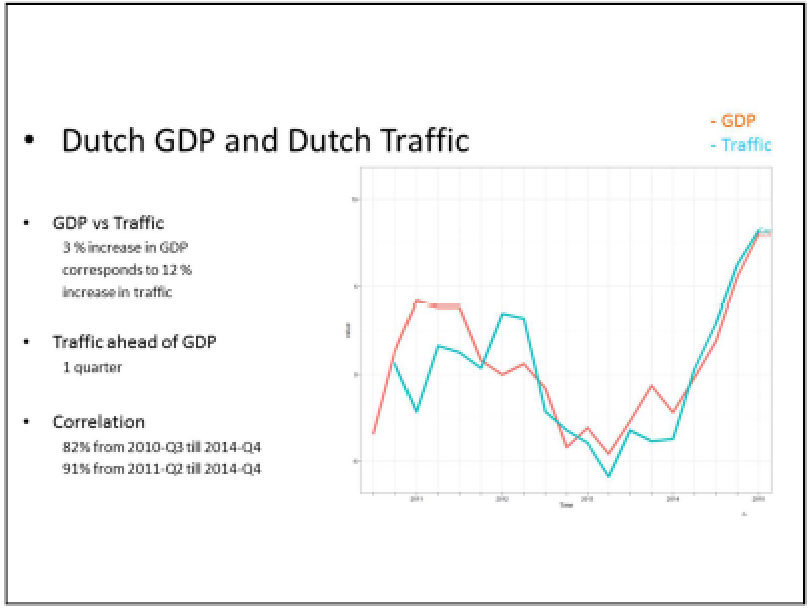


The Netherlands in bloom

Maaike Hersevoort (CBS)
Hamed Mehdipoor (UT)

Do cars drive the economy?

Dan Ionita (UT) and Ronald van der stegen (CBS)



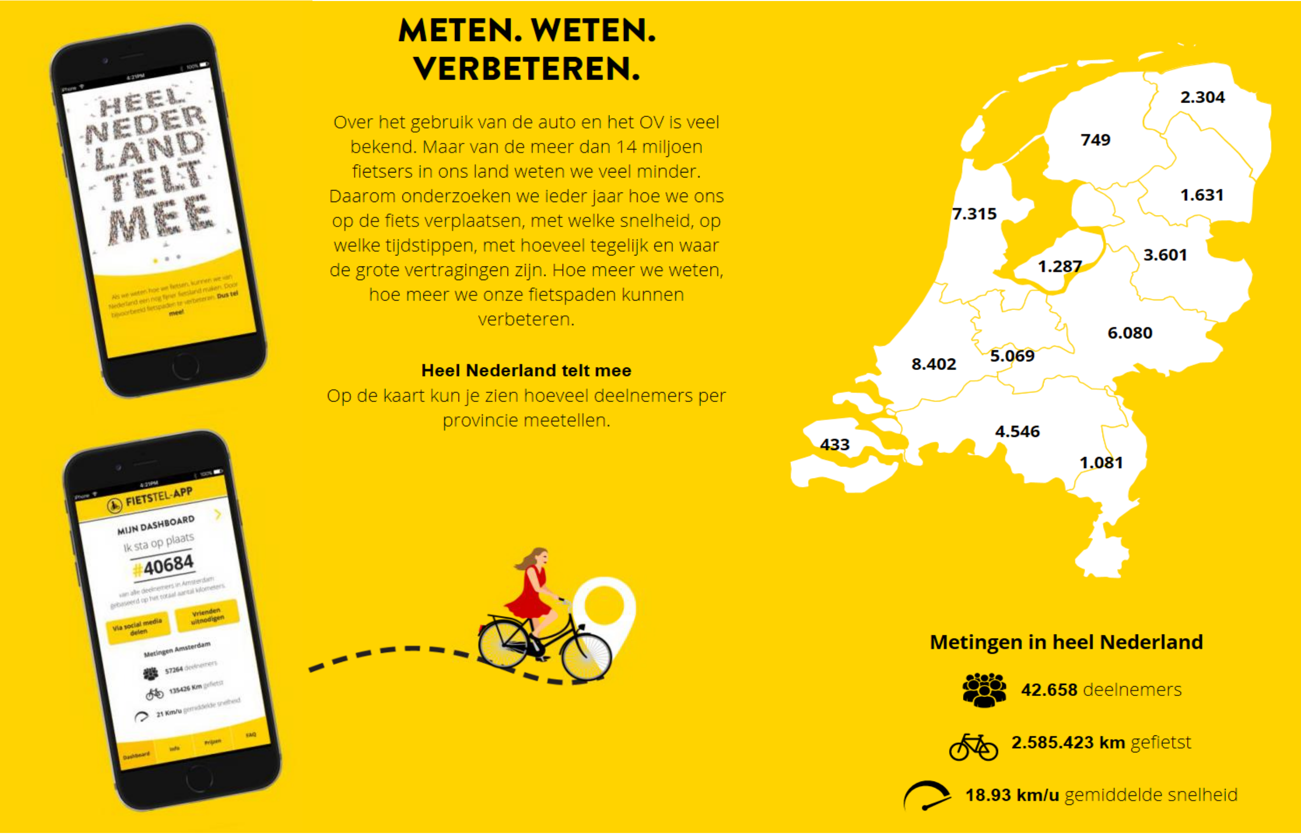
- Dutch GDP and Dutch Traffic

- GDP vs Traffic
  3 % increase in GDP corresponds to 12 % increase in traffic
- Traffic ahead of GDP
  1 quarter
- Correlation
  82% from 2010-Q3 till 2014-Q4
  91% from 2011-Q2 till 2014-Q4

Center for
Big Data Statistics

National bicycle counting week

Using the crowd with the help of students



HEEL NEDERLAND TELT MEE

Als we weten hoe we fietsen, kunnen we van Nederland een nog fijner fietsland maken. Door bijvoorbeeld fietspaden te verbeteren. **Dus tel mee!**

FIETSTEL-APP

MIJN DASHBOARD

Ik sta op plaats

#40684

van alle deelnemers in Amsterdam gebaseerd op het totaal aantal kilometers.

Via social media delen | Vrienden uitnodigen

Metingen Amsterdam

57264 deelnemers

135426 Km gefietst

21 Km/u gemiddelde snelheid

Dashboard | Info | Prijzen | FAQ

METEN. WETEN. VERBETEREN.

Over het gebruik van de auto en het OV is veel bekend. Maar van de meer dan 14 miljoen fietsers in ons land weten we veel minder. Daarom onderzoeken we ieder jaar hoe we ons op de fiets verplaatsen, met welke snelheid, op welke tijdstippen, met hoeveel tegelijk en waar de grote vertragingen zijn. Hoe meer we weten, hoe meer we onze fietspaden kunnen verbeteren.

Heel Nederland telt mee
Op de kaart kun je zien hoeveel deelnemers per provincie meetellen.

Metingen in heel Nederland

42.658 deelnemers

2.585.423 km gefietst

18.93 km/u gemiddelde snelheid

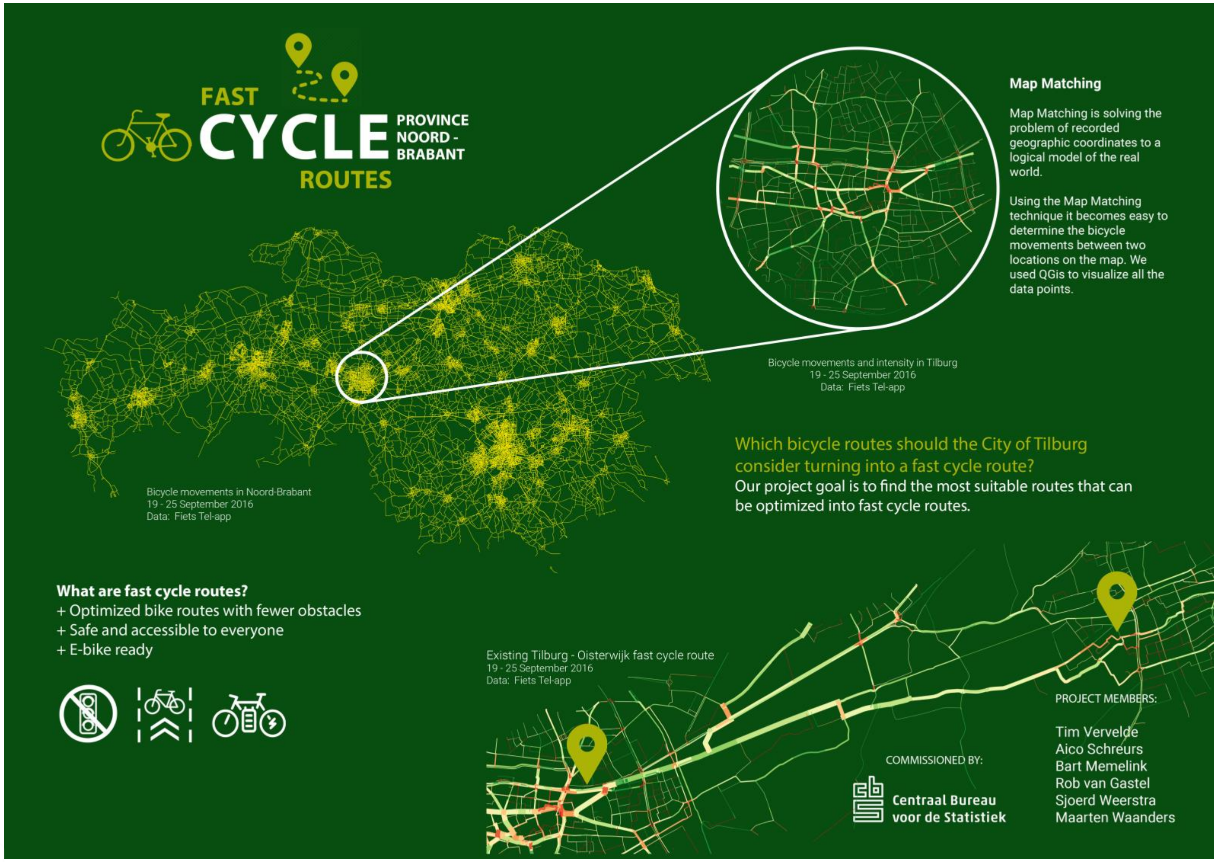
Center for
Big Data Statistics

Fontys

Can we turn this into statistics?

National bicycle counting week

Using the crowd with the help of students



FAST CYCLE ROUTES PROVINCE NOORD - BRABANT

Map Matching

Map Matching is solving the problem of recorded geographic coordinates to a logical model of the real world.

Using the Map Matching technique it becomes easy to determine the bicycle movements between two locations on the map. We used QGis to visualize all the data points.

Bicycle movements and intensity in Tilburg
19 - 25 September 2016
Data: Fiets Tel-app

Bicycle movements in Noord-Brabant
19 - 25 September 2016
Data: Fiets Tel-app

Which bicycle routes should the City of Tilburg consider turning into a fast cycle route?

Our project goal is to find the most suitable routes that can be optimized into fast cycle routes.

What are fast cycle routes?

+ Optimized bike routes with fewer obstacles
+ Safe and accessible to everyone
+ E-bike ready

Existing Tilburg - Oisterwijk fast cycle route
19 - 25 September 2016
Data: Fiets Tel-app

COMMISSIONED BY:

Centraal Bureau voor de Statistiek

PROJECT MEMBERS:

Tim Vervelde
Aico Schreurs
Bart Memelink
Rob van Gastel
Sjoerd Weerstra
Maarten Waanders

Center for Big Data Statistics

Fontys

Yes we can! See also the site

Data dilemmas

Users are impatient
and create own solutions

Can we use third-party HPC facilities?

How to give external users secure and easy access to microdata?

How to make data delivery more flexible?

Analysts spend too much time on data preparation

How can we process (near) real-time information?

How to keep data quality and governance under control?

Adding new information is time consuming and costly

How to secure anonymization and pseudonymization?

How can I use cloud-based solutions fast and safe. Is it legal?

I want a single version of the truth!

I want to implement a Data Management Platform

How to use the potential of big data without copying complete data sets?

How can I support self service?

Better data access for scientific research

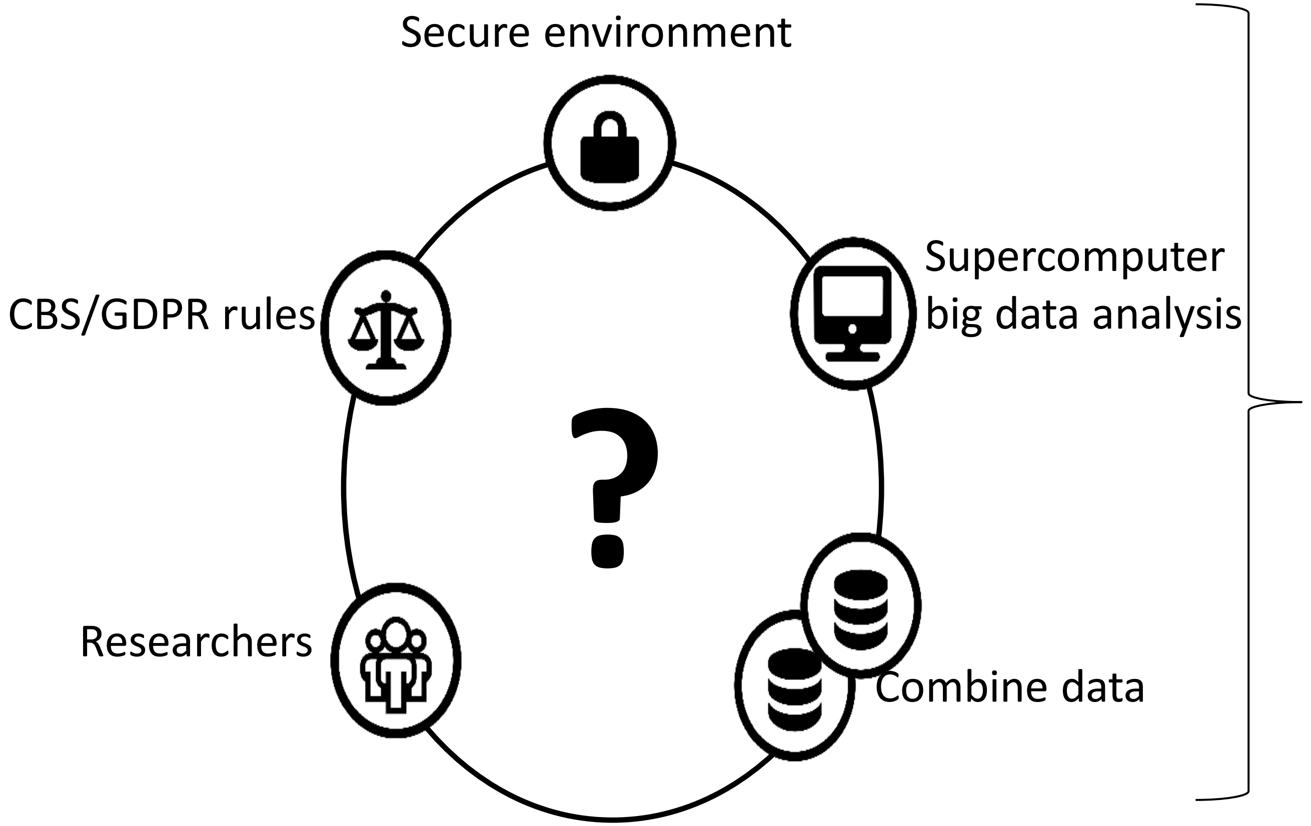
From the ODISSEI website:

(Open Data Infrastructure for Social Science and Economic Innovations)

ODISSEI works to develop a **sustainable research infrastructure** for the social sciences in the Netherlands. Through ODISSEI, researchers within the social sciences will have **access to large-scale, longitudinal data collections** connected to registrations from Statistics Netherlands (CBS). (…)

(…) The creation of a national data platform has been **included in the National Roadmap** for Large-Scale Scientific Infrastructure 2016-2020.

Proof of Concept ODF- ODISSEI Data Facility


Secure environment

Supercomputer big data analysis

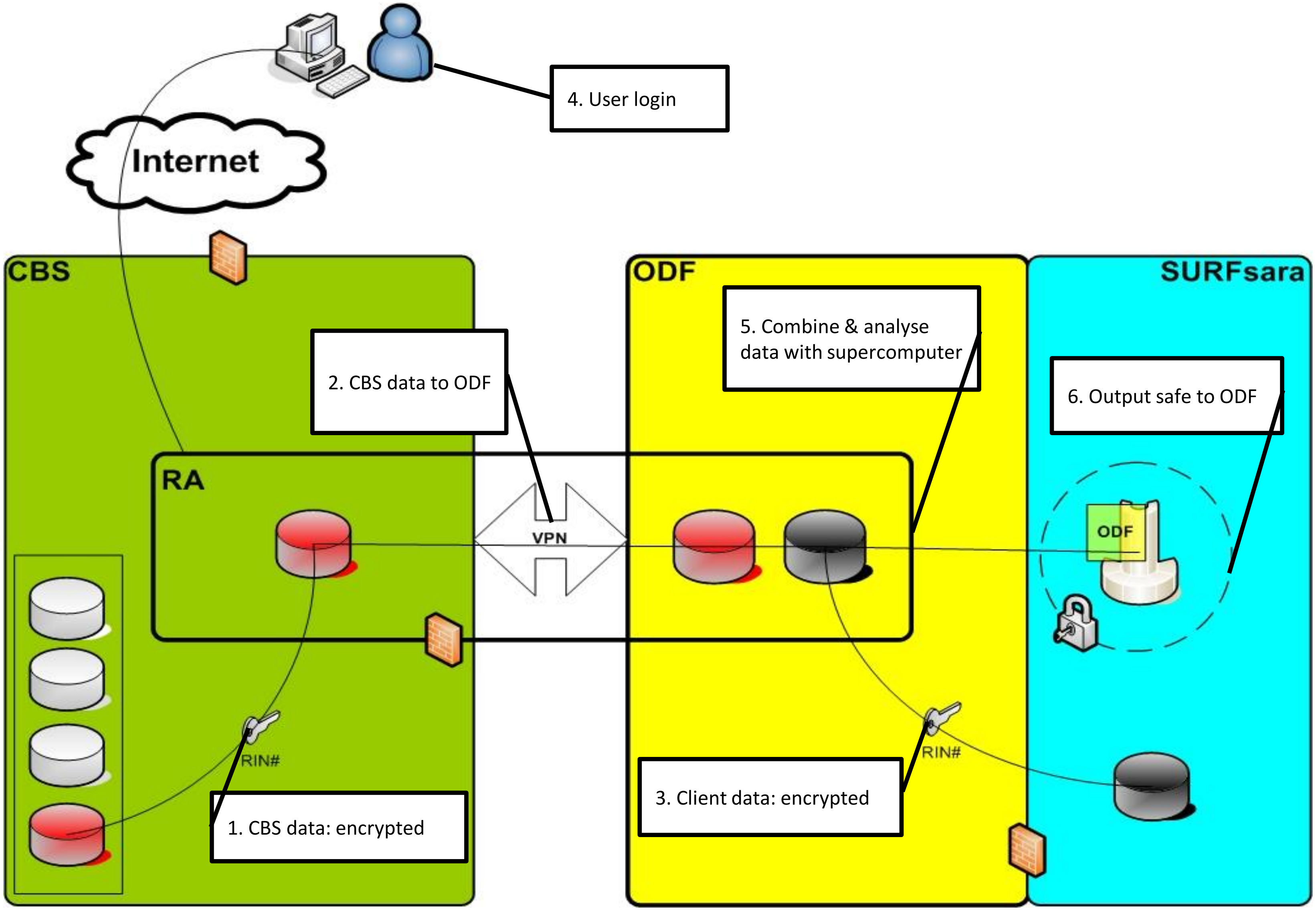
Combine data

Researchers

CBS/GDPR rules

Proof of Concept challenges

ODF implementation



4. User login

5. Combine & analyse data with supercomputer

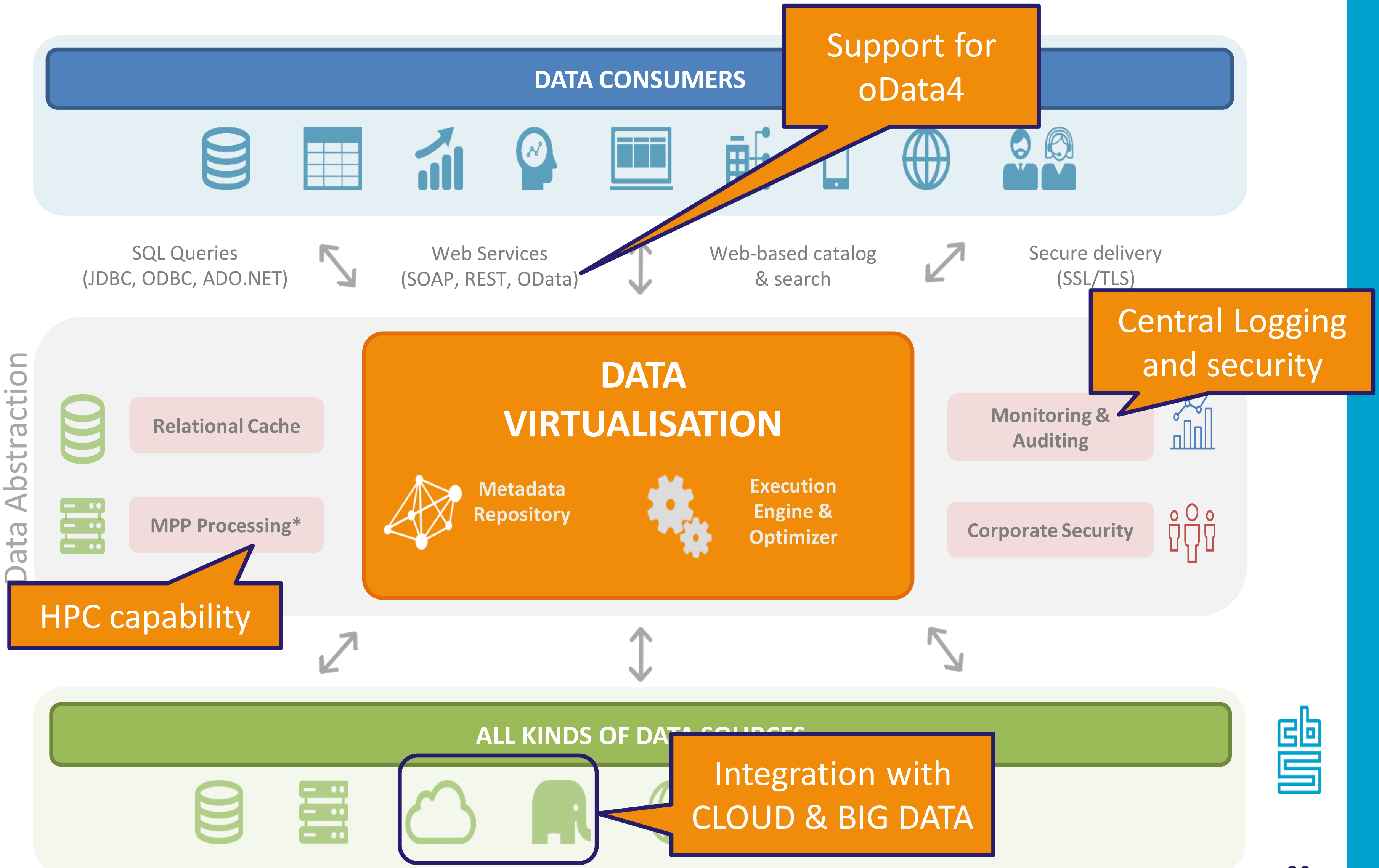
2. CBS data to ODF

6. Output safe to ODF

1. CBS data: encrypted

3. Client data: encrypted

Towards a Data Virtualisation Architecture



Support for oData4

DATA CONSUMERS

SQL Queries (JDBC, ODBC, ADO.NET)

Web Services (SOAP, REST, OData)

Web-based catalog & search

Secure delivery (SSL/TLS)

Central Logging and security

Data Abstraction

Relational Cache

MPP Processing*

HPC capability

DATA VIRTUALISATION

Metadata Repository

Execution Engine & Optimizer

Monitoring & Auditing

Corporate Security

ALL KINDS OF DATA SOURCES

Integration with CLOUD & BIG DATA

* Massive Parallel Processing

Hal Varian ‘On workers and managers’

The McKinsey Quarterly, Januari 2009:

I keep saying the sexy job in the next ten years will be statisticians (…)

Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

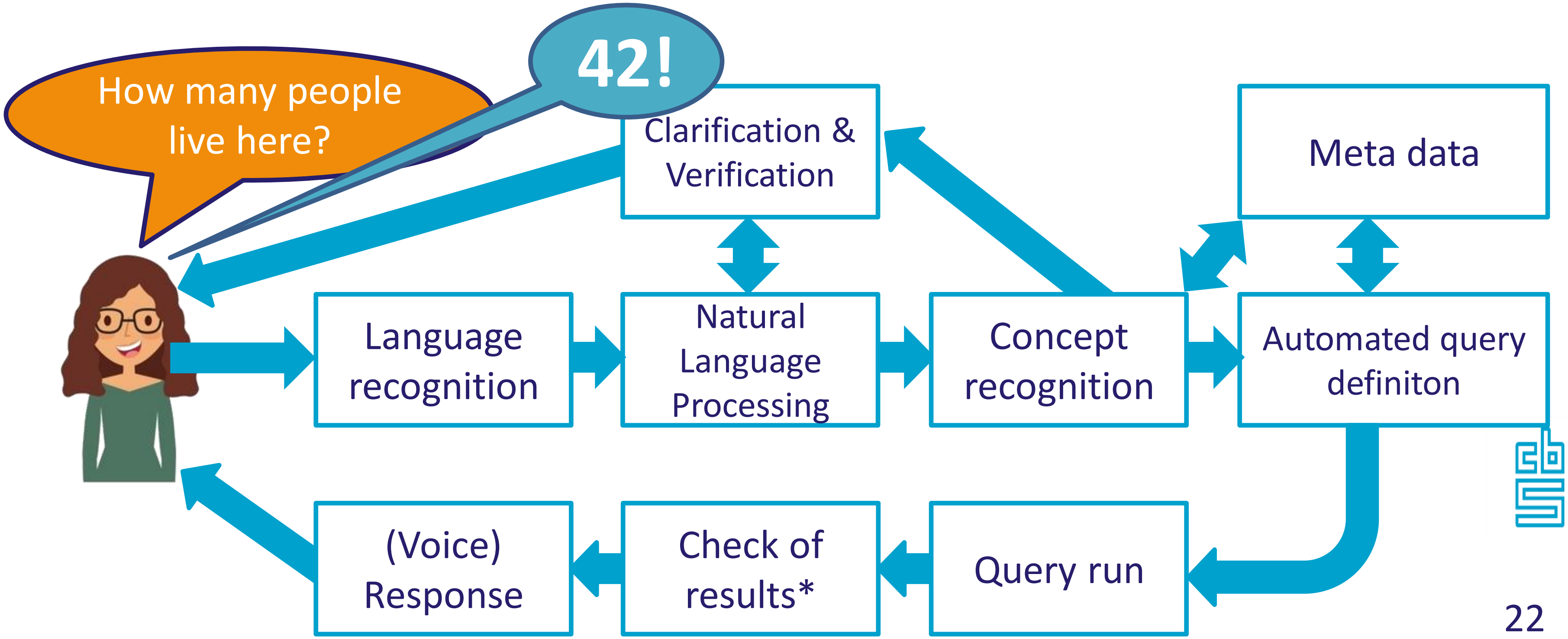
I think statisticians are part of it. (…) Managers need to be able to access and understand the data themselves.

- Chief economist at Google
- Emeritus professor at Berkeley

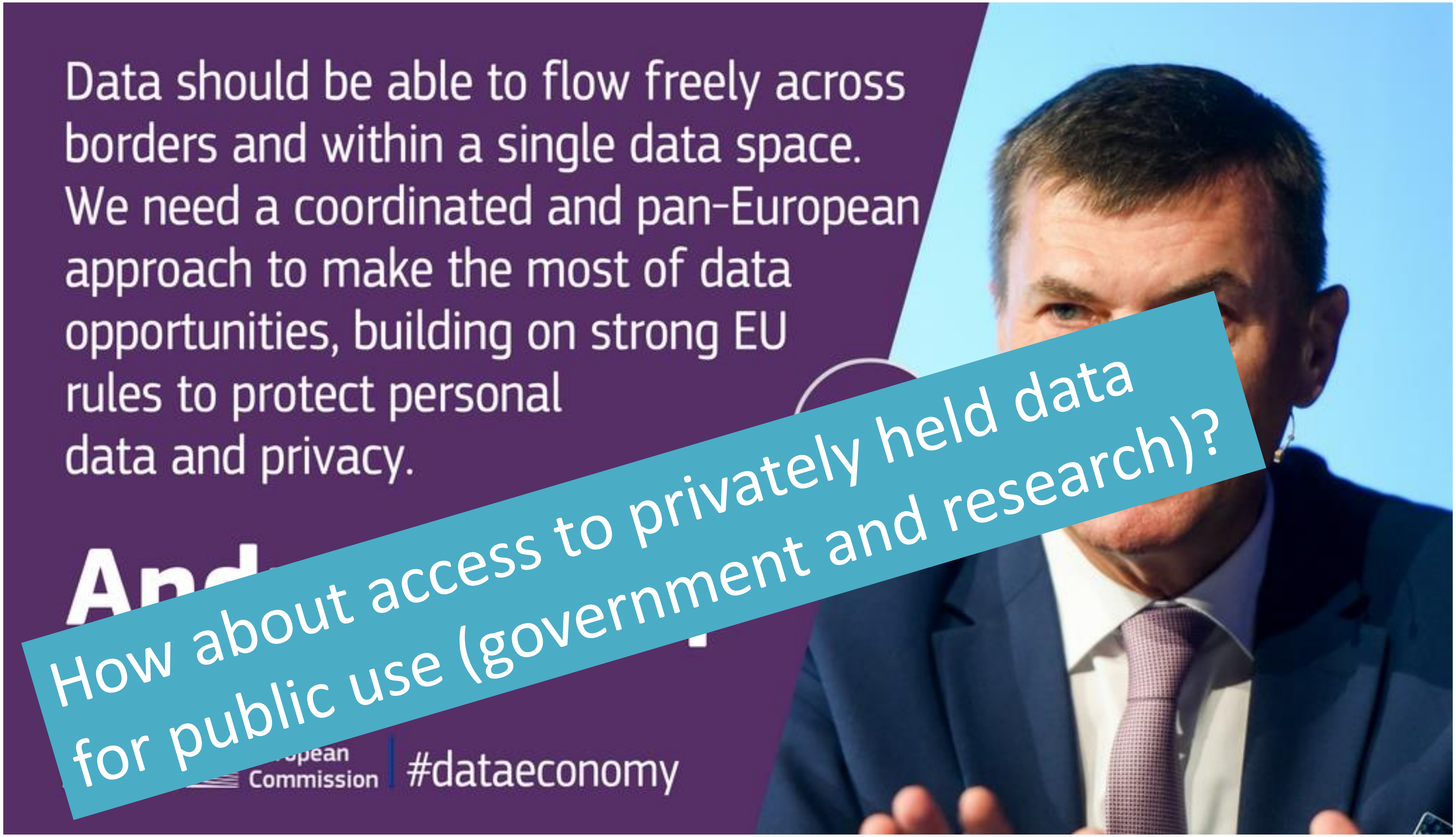
The Information Dialogue

“An **arbitrary user** can ask CBS any **arbitrary question** at any **arbitrary moment** via any **arbitrary platform** (desktop, tablet, mobile device).

Next, CBS **clarifies the question** in a partly or fully **automated dialogue** and, based on available content (text, images, data, audio, visuals and data visuals) **a single complete answer is given in a format demanded by the user**.”


42!
How many people live here?
Clarification & Verification
Meta data
Language recognition
Natural Language Processing
Concept recognition
Automated query definiton
(Voice) Response
Check of results*
Query run

Data is an EU priority



Data should be able to flow freely across borders and within a single data space. We need a coordinated and pan-European approach to make the most of data opportunities, building on strong EU rules to protect personal data and privacy.

How about access to privately held data for public use (government and research)?

Takeaway messages (long versions)

1) The importance (and volume, variety and velocity) of data is rapidly growing. This calls for new methods and architectures of data management.

2) New relations between societal partners are emerging. Collaboration and partnerships are crucial in moving forward. Data sharing is not just an option, it is a must!

3) A combination of top-down (frameworks, guidelines) and bottom-up (experiments, concrete cases) approaches works best.

4) The importance of soft skills (communications, negotiation, sensitivity to the environment) for researchers is increasing. The ivory tower disappears.

5) Whatever you do, support at top management level is extremely important- create the conditions for change and innovation.

Takeaway messages (short versions)

1) New methods and architectures of data management are needed
2) Collaboration and partnerships are crucial in moving forward
3) Combine top-down and bottom-up approaches
4) The importance of soft skills is increasing
5) Create the conditions for change and innovation at top level

AFTER EIGHT MONTHS, SENIOR MANAGEMENT FINALLY APPROVED YOUR PROJECT PLAN.
Dilbert.com DilbertCartoonist@gmail.com
IT'S TOO LATE. ALL OF THE TECHNOLOGY HAS CHANGED AND OUR COMPETITORS HAVE LEAPFROGGED US.
12-8-09 © 2009 Scott Adams, Inc./Dist. by UFS, Inc.
MAYBE YOU COULD WRITE A NEW PLAN.
OR WE COULD GET THE SAME RESULT BY RESUB-MITTING THIS ONE.

Facts that matter