KSH HUNGARIAN CENTRAL STATISTICAL OFFICE

# Safe Centre at the Statistical Office: Hungarian Experience

**_Dr. Zsolt Németh Ph.D._**
*chief adviser*
*zsolt.nemeth@ksh.hu*

**CERGE-EI**
***March 22, 2018***

# General Framework

- We live in the most complex society of human history.
- The speed of changes has never been experienced before
  - The most rapid changes are in technologies – the unbelievable expansion of
    - storage capacities
    - processing speed

↓

**This is Data Revolution**

# Data Revolution

- The „new oil" comes in three main forms:
  - **Big Data**
    - featured by 3-x Vs
      - volume, variety, velocity, variability, veracity + complexity etc.
  - **Open Data**
    - Free availability of data for reusing, analyzing, republishing, sharing etc.
  - **Administrative data**
    - Data collected for non-statistical purposes
    - Full coverage is an objective
    - Method of data collection and data processing is determined by the public administration

- The „new oil" presents itself as researchable data for social scientists

# Problems of Big Data*

- Information Society entered into a new phase: Algorithmic Society
- Algorithms and AI are the machines; Big Data is the fuel that makes the machines run.
- The Algorithmic Society features the collection of vast amounts of data on individuals and facilitates new forms of surveillance, control, discrimination and manipulation, both by **governments** and by **private companies**.
- In the Algorithmic Society, surveillance and data collection are now widely distributed, but there is no guarantee that they will be democratically controlled. Data about many people are collected in many places, but a relatively small number of people have the resources and the practical ability to collect, analyze, and use these data.
- The state, while always remaining a threat to free expression, also needs to serve as a necessary counterweight to developing technologies of private control and surveillance.

*Jack M. Balkin: *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation* Yale Law School, Public Law Research Paper No. 615

# Dataism

- "Dataism declares that the universe consists of data flows, and the value of any phenomenon or entity is determined by its contribution to data processing„*
- The problem is the same as that of Big Data: the more capable data mining systems become, the more businesses and government agencies can use these systems to spy on consumers or citizens in many different ways.
- Dataism and Big Data jeopardize privacy.
- Adequate institutional and structural responses are not in place yet

*Harari, Yuval Noah (2017). Homo Deus: A Brief History of Tomorrow. UK: Vintage Penguin Random House. p. 428*

# Statisticians on the battlefield

- Who are statisticians?
  A. Harmless plant eaters and/or lazy bureaucrats – a very popular view
  B. To be a statistician is a professional calling to tell the public, **what the world looks like ☺**

- Why do we produce statistical information?
  - „Independent and high quality statistics are essential for a democratic society." (Tim Holt)
  - „Good statistics are much cheaper than bad decisions." (Janez Potocnik, ex EU commissioner, 2011)
  - Official statistics are essential part of the public good

# Official statistics in the age of AS

- Official statistical institutions are responsible for producing high quality statistics on society, economy and environment
- **Challenges**
  - Never experienced rapid changes in the world
  - Consequently rapid changes in users' needs
  - There are new and powerful competitors in information production
- **Opportunities**
  - Rapid growth in storage capacities and processing speed
  - New data sources: Big Data, Open Data, administrative sources
  - Extensively decreasing production costs
  - The activity of big international organizations (UN, OECD, Eurostat)
  - New legal regulation on standards of quality assurance, data processing and behavior
  - The popularity of evidence-based decision making

# A big step forward

- Statistical offices have to find a very narrow path for how to
  - protect individual data of data providers;
  - provide access to microdata.
- Official statistics count as absolutely trust based business.

**All the collected statistical data can be used exclusively for purposes of statistical analysis.**

**A new type of co-operation between researchers and statistical offices: access to microdata**

- Different forms of access to microdata:
  - Access to anonymized microdata sets
  - Remote execution
  - Remote access
  - Safe Centre access
    - Inside the HCSO infrastructure
    - Outside the HCSO infrastructure

# Jan Nepomucky and data protection

# Access to anonymized microdata sets

- HCSO sends microdata sets to researchers requesting the dataset, **based on a contract**.
- Microdata sets are datasets that contain information on observation units. Anonymized microdata sets are the form of microdata that have been modified by statistical disclosure control methods in order to reduce to an acceptable level, in accordance with current best practices, the disclosure risk of statistical units to which they relate.
- Anonymized microdata sets are always released with **accompanying methodological documentation,** describing the methodology used during anonymization and their effects on the dataset.
- Release of anonymized microdata sets is available only for **scientific purposes** and access is granted only for approved research projects that meet all researcher accreditation criteria.
- In the context of release of anonymized microdata sets **all datasets** managed by the HCSO can be requested.
- **Contract and confidentiality commitments** are to be signed for data requests successfully approved after researcher accreditation and evaluation of the request form from professional and data protection point of view.
- In case the requested anonymized microdata set is not available in the form and with the content as requested, the HCSO might **charge for the production** of the requested dataset. For the production of such anonymized microdata set a fee has to be paid to be paid (expert day equivalent).

# Remote execution and remote access

## Remote execution

- Datasets are managed by the HCSO (use of external datasets is also possible)
- Remote execution environment is based on the syntax files and/or specifications provided by the researcher.
- HCSO produces the research outputs as indicated in the specifications.
- Output checking.
- It doesn’t really work properly
- Only 2 outputs in 2017

## Remote access

- Remote access provides more or less the same services for the researchers as the Safe Centre.
  - Service is still inside the HCSO infrastructure but out of Budapest. (In the Szeged office of the HCSO.)
  - The rules and procedures are the same as in the SC in Budapest
  - On-line access to the SC server in Budapest
  - 2 work stations, low utilization

# Safe Centre access

# General rules of SC access

Access in the Safe Centre is provided to de-identified micro data sets only for **scientific purposes**, respecting the protection of individual statistical data and the data protection regulations. In the Safe Centre, researchers access datasets

- prepared for research in safe environment
- with a CCTV (closed-circuit television) surveillance system in place.

Safe Centre access is available only for approved research projects that meet all researcher accreditation criteria.

- **Process of accreditation**:
  - Initiation of Safe Centre access
  - Evaluation of Safe Centre access requests
  - Signing contract and confidentiality commitment
  - Providing Safe Centre access

# Datasets in the Safe Centre

- **Standard datasets prepared for research**
  The list of such datasets is available and kept up-to-date on the HCSO website.
- **Datasets prepared based on specific data requests**
  If no suitable dataset is available on the list of standard datasets prepared for research for the given research purpose, preparation of additional datasets can be requested ***for a fee***.
- **Linked datasets**
  During the preparation of the requested datasets, linkage of different datasets or preparation of such datasets for linkage is possible. In the latter case the HCSO assigns such technical IDs to the microdata sets that allow no direct identification but the linkage of the different datasets.
- **External datasets**
  External datasets can also be used for producing research outputs, in addition to the datasets prepared by the HCSO for the relevant research project. These external datasets must be listed on the data request form for Safe Centre access.

# Standard datasets in the Safe Centre

- **The list of standard datasets *free of charge*, prepared for research:**
  - **Population and Housing Census microdata sets**
    - 10% sample of the Population and Housing Census 2001
    - Micro census of 2005
    - 10% sample of the Population and Housing Census 2011

  - **Labour Force Survey (LFS) microdata sets, Quarterly datasets**
    2003 – 2016 + Q2 2017
  - **Household Budget Survey (HBS) and EU-SILC microdata sets**
    2005 – 2016 + Well Being microdata dataset 2013, Material deprivation
  - **Farm Structure Survey - EUROFARM microdata sets**
    2000, 2003, 2005, 2007, 2010
  - **European Health Interview Survey (EHIS) microdata sets**
    2009, 2014
  - **Time Use Survey microdata sets** 2010

# Other datasets in the Safe Centre

- Specific datasets can be requested **for a fee**. The HCSO provides access to such specific datasets after the process of granting access to the Safe Centre.
    - Datasets from HCSO's regular data collections
    - Datasets from administrative sources – if the data owner assigns to use it for scientific purposes
    - Other surveys, like:
        - ECB's Household Finance and Consumption Survey (HFCS) 2014, 2017
        - Large-sample Hungarian dwelling survey 2015
        - OECD's Programme for Assessment of Adult Competencies (PIAAC) – fieldwork in progress
        - Time Use Survey (TUS) – to be collected at the end of this decade
- Linked datasets - HCSO assigns such technical IDs to the microdata sets that allow no direct identification but the linkage of the different datasets.
- External datasets
        - External datasets can also be used for producing research outputs, in addition to the datasets prepared by the HCSO for the relevant research project – input checking!!!

# Available firm-level datasets

- External trade (1992-2015)
- Business Register information (2012-2016)
- Innovation (2006-2015)
- Industrial prices (1998-2014; 2015)
- Sale of industrial products (1995-2015)
- R+D (2004-2015)
- Community Innovation Survey (2006-2014)
- Balance Sheet data (1992-2015)
- FDI (2000-2013)
- VAT statistics (2015-2016)

# Metadata and other information

- Guideline for researchers.
- Methodological guidelines, information on data collections, questionnaires (including the names of the variables).
- HCSO publications on related topics.
- List of variables in the files, indicating the exact name, measure, number of the question on the questionnaire to which the variable contains the answer.
- The description of classifications, nomenclatures that are used in the files (in effect at the given time and at the reference time of the data).
- The labels of the variables in a separate file or built into the data file.

# What can and what cannot be used in the SC

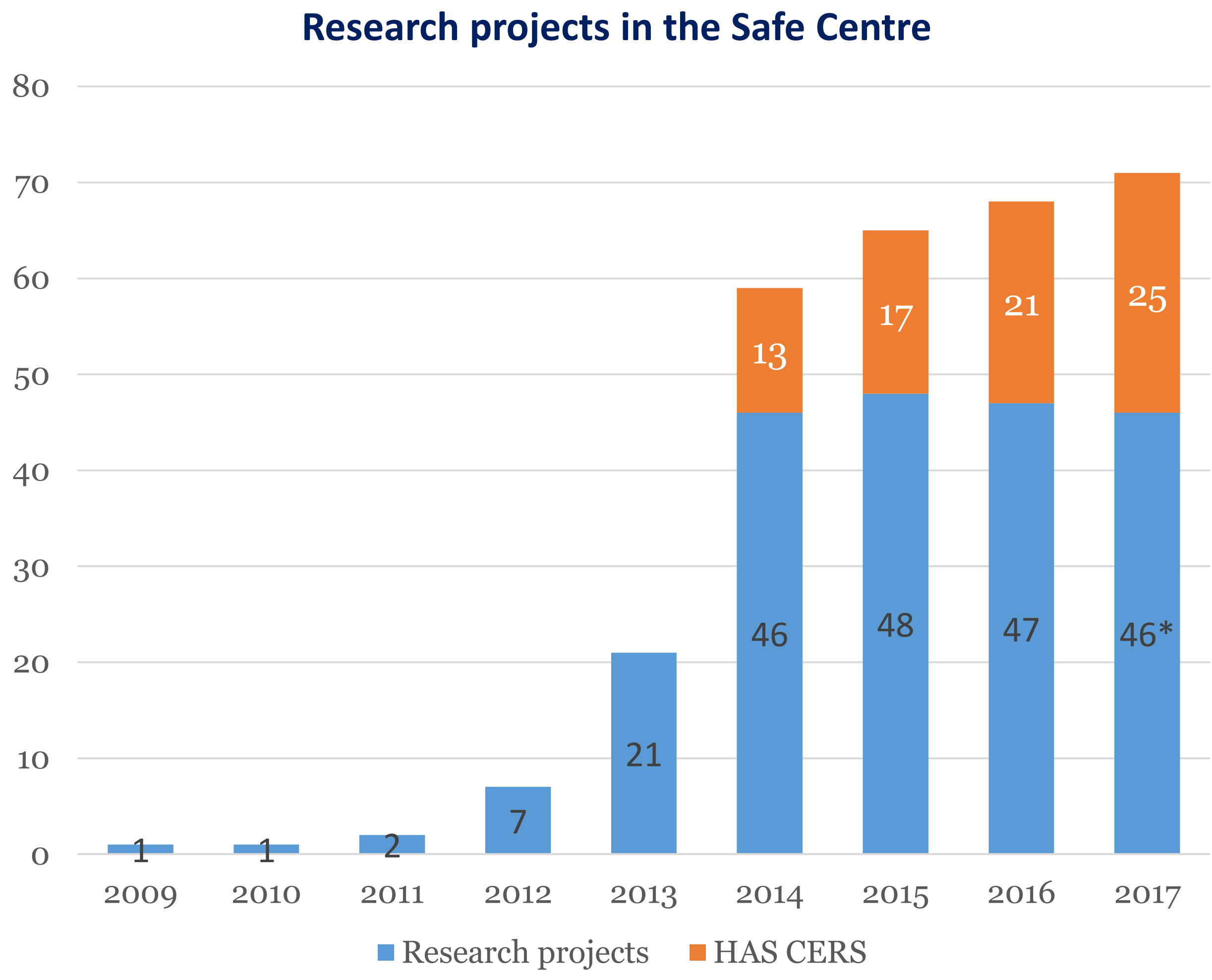
- **Available IT tools:**
  - STATA 12.0 SE
  - SPSS 22.0 BASE
  - SAS 9.4, SAS Enterprise Guide 7.1
  - Microsoft Office 2013
  - Stat/Transfer v12
- **Forbidden:**
  - Printing of documents.
  - Copying data to external data storage.
  - Copying the data used for research onto the hard drive of the local client PC.
  - Connecting any instrument to the client PC.
  - Entering the Safe Centre with laptop, phone or any other instrument capable of mobile communication and recording.
  - Use of internet and e-mail.
  - Taking notes prepared in a non-electronic form from the Safe Centre.
  - Changing the system settings.

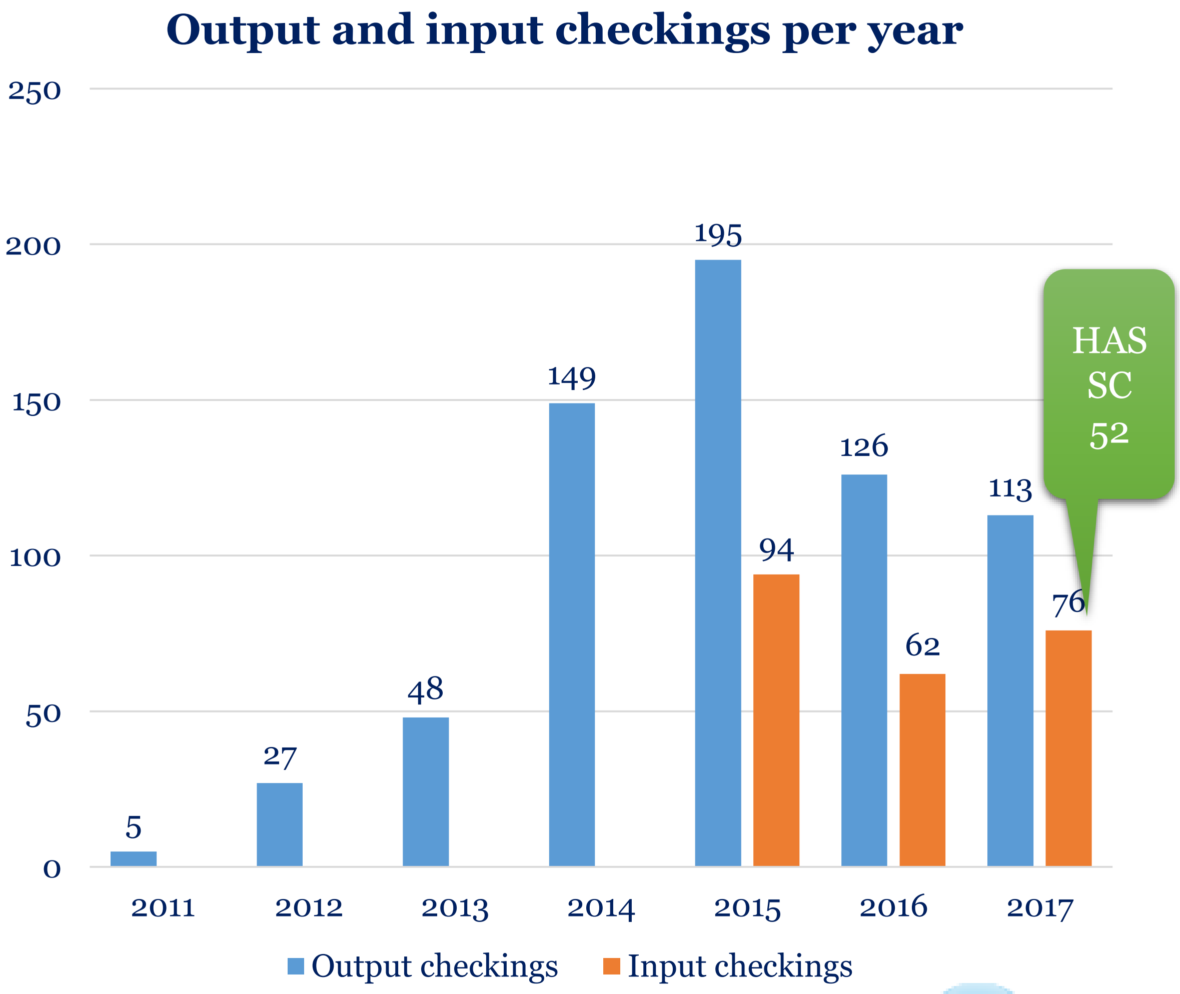
# The Safe Centre in Budapest

# Output checking

- The HCSO fully examines the research outputs produced in the Safe Centre from statistical confidentiality point of view, prior to their release to the researcher. The HCSO provides access to the researcher only to those research outputs that are checked and approved against statistical confidentiality.
- The purpose of the output checking procedure is to check the produced research outputs against disclosure; to ensure that the research outputs **do not allow identification of and disclosure of information on statistical units.**
- In order to support the output checking procedure, all research outputs produced in the Safe Centre must be documented. Researchers have to document their research outputs using the standard documentation form of the research outputs.
- Outputs are sent to researchers by e-mail.
- Output checking is recently free of charge in the case of datasets in the same category and it is charged in all other cases. → This policy should be reconsidered.

# Safe Centre in figures





* 6 remote accesses

# Pricing of Output Checking

| One type of research result | | | |
|---|---|---|---|
| **Types of Output** | **Denomination** | **Number of research results** | **Expenditures in shifts** |
| 1 | diagrams, graphs | max. 25 db pieces | ¼ day |
| 2 | statistical model results | max. 50 executed program | ¼ day |
| 3 | mean, deviation and one dimensional tables | max. 25 tables | ¼ day |

| Several types of research results | | | |
|---|---|---|---|
| **Types and size of outputs** | **Pieces of research results** | **Total size of outputs** | **Expenditures in shifts** |
| 4 | max. 5 files | max. 1 MB | ¼ day |
| 5 | max.10 files | max. 3 MB | ½ day |
| 6 | max. 25 files | max. 5 MB | 2/3 day |
| 7 | max. 50 files | max. 8 MB | 1 day |
| 8 | max 100 files | max. 10 MB | 2 days |
| 9 | 101 files and more | - | based on quantity of outputs |

| | **2017** | | |
|---|---|---|---|
| **Research results by type** | **HCSO SC** | **HAC CERS** | **Total** |
| | **pieces** | | |
| 1 | 0 | 2 | 2 |
| 2 | 2 | 13 | 15 |
| 3 | 7 | 4 | 11 |
| 4 | 42 | 14 | 56 |
| 5 | 2 | 8 | 10 |
| 6 | 1 | 3 | 4 |
| 7 | 2 | 0 | 2 |
| 8 | 1 | 1 | 2 |
| 9 | 1 | 0 | 1 |
| Others | 4 | 6 | 10 |
| **Total** | **62** | **51** | **113** |

# HCSO strives to meet users' requirements

- Since 2015 HCSO has operated de facto 2 SCs
  - The basic one
  - One for HAS CERS
- CERS invested in HCSO's SC
  - They bought their own servers for their research projects
  - HCSO integrated them into its IT infrastructure
    - CERS paid a fee for operation, surveillance, service, trouble-shooting etc.

# A jump to a new terrain

- HAS invested into a new science building
  - CERS and Centre for Social Sciences are located in the new building
  - CSS conducts basic research in political science, sociology, minority studies and law.
- CERS initiated negotiations with HCSO about a Safe Centre in the new science building

# Safe Centre outside the HCSO infrastructure

- After long negotiations HCSO and CERS signed a contract on a SC outside the HCSO infrastructure.
  - The duration of the contract is for 5 years
- The contract provides a prudent regulation on data accessibility and data protection.
- Data protection has three chief areas:
  - legal
  - physical
  - IT

  regulations
- Contract contains a very detailed regulation of duties of the partners.

# The SC at HAS CERS

- HCSO and HAS CERS operate **jointly** the SC
- HAS CSS is also authorized to use SC
- A special review committee at CERS evaluates new research initiatives and forwards it to HCSO

## IT solutions

- VMware Horizont virtual platform
- 2 servers, 12 workstations
- VPN (Virtual Private Network)
- Public, encrypted Internet connection (cost efficient)

# Duties of HCSO

- Ensures the necessary storage capacity and performs a night save on jobs completed during the day
- Submits ID and password for researchers
- Ensures a shared directory where researchers can share and store files – except for micro data
- Provides the necessary softwares
- If it is required, HCSO prepares test files, which represent well the dataset, but they are inappropriate for research
    - Test files can be used outside the SC
- Does the Input Checking of external datasets
- Monitors CCTV screens
- Executes output checking in 7 shifts

# Duties of CERS

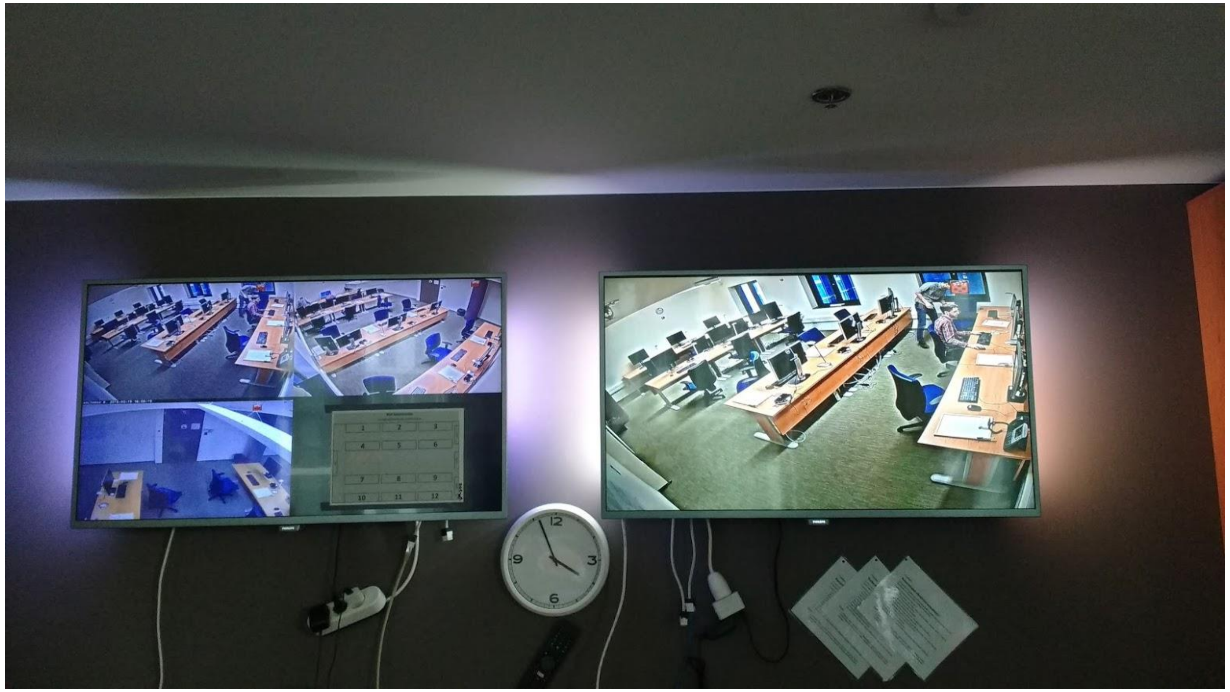
- Informs researcher about the regulation of SC
- Ensures that non-authorized persons do not enter the SC
- Organizes and controls the daily operation of SC – regularly informs HCSO
- Monitors CCTV screens, observes the safety rules and supervises the work in SC
- In case of a security incident, they immediately inform HCSO

# Annual costs

Cca. 30.000 EUR / year which covers the following:

- Updating datasets
  - If the annual fee is exceeded, CERS pays 130 EUR/expert day
- IT operation costs
- Supervision of security
- Output checking

# Safe Centre at CERS and its’ CCTV surveillance



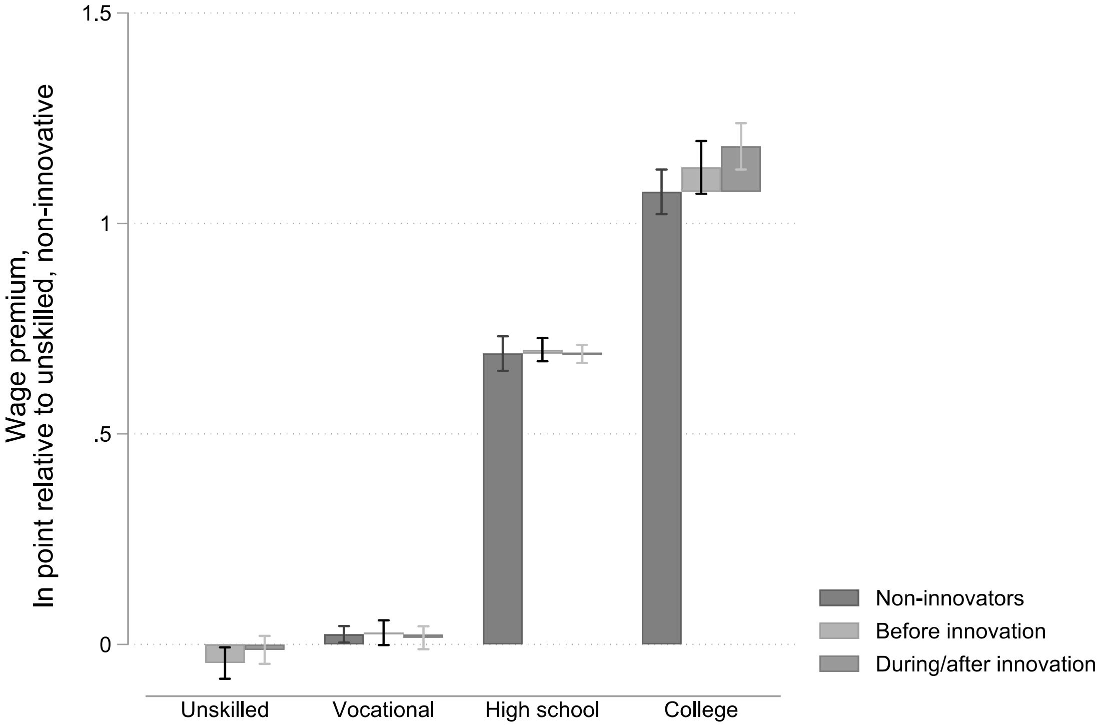
# Research results in the Safe Centre

# Innovation and within-firm wage inequality

- **Attila Lindner, Balázs Muraközy, Balázs Reizer**
- **Research question**
  - Innovation is often considered as skill-biased
    - A key driver of increasing inequality
  - Little is known about
    - What type of innovation causes an increase in inequality
      - Technological or also management?
      - R&D-based innovation or simply the adoption of existing technologies
    - How important is the within-firm component?
- Data
  - Linking the Structure of Earning Survey to the Community Innovation Survey and Balance Sheet data
- Method
  - Running worker-level (Mincer-type) regressions with
    - Wages as dependent variable
    - Firm-level innovation status and its interaction with education level as explanatory variables
- Results
  - Innovative firms pay higher wage premia for college educated workers even before the innovation takes place
  - Innovation, especially management innovation leads to increased college wage premium

# Innovation and within-firm wage inequality

Figure 2: Results of baseline regression



*Notes:* This figure shows the wage advantage, in log point terms, of workers with different skill levels, relative to unskilled workers working in non-innovative firms, estimated from Mincer-type equations with skill-innovation variable interactions and firm- and skill level-year fixed effects. The spikes/caps show 95 percent confidence intervals, where standard errors are clustered at the firm level.

# Example: GEO

GEO is a data set of 45,500 census tracts (CT) and a matrix of commuting times and costs from one CT to another

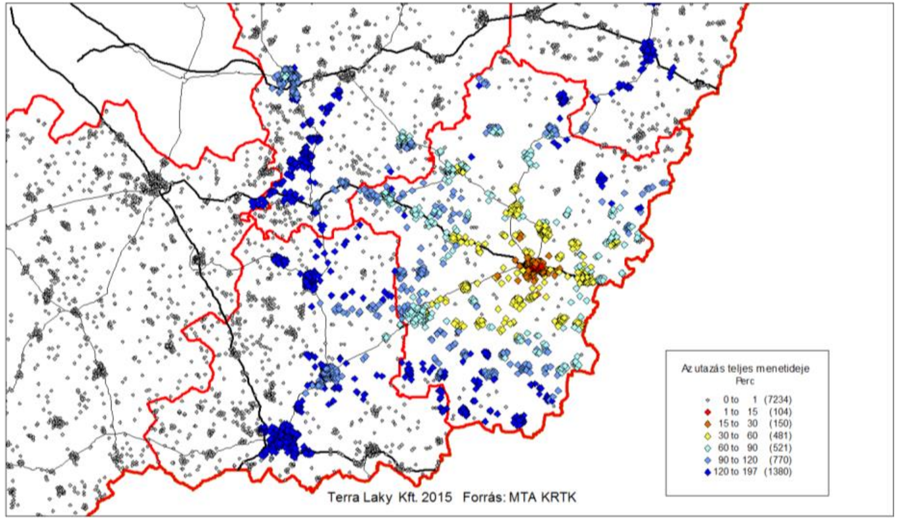
GEO was built by a joint effort of the HCSO, the Academy of Sciences, CEU and 3 business firms, and was financed by the HAS.

Census-based data on each CT's population + firm-level data on (nearly) all employers in the CT + educational institutions + health care providers

Notes: Reference year 2011. Commuting by public transport and car are both considered. Firm's financial data are available.

**When the above data are merged with the 2011 Census and the 2016 Micro census (in the Safe Centre): one can draw the relevant geographical environment of any person or firm.**

## Commuting times by public transport from the centre of Békéscsaba (South-East Hungary)



Terra Laky Kft. 2015 Forrás: MTA KRTK

## Examples of research questions

How do neighborhood characteristics and the accessibility of schools and jobs affect NEET* among 15-24 year old youth?

The effects of cutting the mandatory school age (from 18 to 16) in ethnic ghettos versus other neighborhoods

Where to open an industrial zone?

*NEET – Not in Education, Employment or Training

# Health Differences at birth between Roma and Non-Roma Children in Hungary*

- Linkage of two big datasets:
  - Birth records – as administrative data
    - It contains all live births since 1970
    - The content of data is completely harmonized across time starting in 1981
    - Information on the date of birth, place of residence at birth, gender, birth weight, and age of gestation of the newborn babies
    - the date of birth, level of education, employment, and residence of both the mother and the father
    - Birth records do not contain ethnic markers
  - Population census of 2011
    - The census of 2011 identifies the ethnic identity of all adult respondents who choose to declare their ethnic background

*Tamás Hajdu – Gábor Kertesi – Gábor Kézdi

http://www.mtakti.hu/wp-content/uploads/2017/11/BWP1712.pdf

# Continuation 1

- Difficulties
  - Neither birth records nor census records have personal identifiers in Hungary, such as social security numbers
  - Names are permanently erased from the census records and are not recorded for birth records
- Solution
  - Birth records contain:
    - the gender and the exact date of birth of the newborn child and the mother/father
    - as well as the city, town or village of residence at birth
  - Census records contain :
    - the gender,
    - the place of residence at birth,
    - the exact date of birth of the individuals and the birth year and month of their children, but not their parents.
  - It means that identifying the mother/father-child pairs (or in other words, the date of birth of the parents) in the census records was possible for children living with their parents during the census of 2011

# Continuation 2

- The most important variables used for the linkage
  - the exact date of birth of the child and the mother,
  - the gender of the child,
  - the residence of the mother at the time of the birth of the child
  - dates of previous live births to the mother were available in the birth records, which helped linking siblings
- The percentage of birth records successfully linked:
  - 90% of live births after 1995 are successfully linked
  - the success rate declines continuously as we consider earlier births, to below 60% in 1981

# Continuation 3 – Research results

*Figure 1*

**Trends in average birth weight**

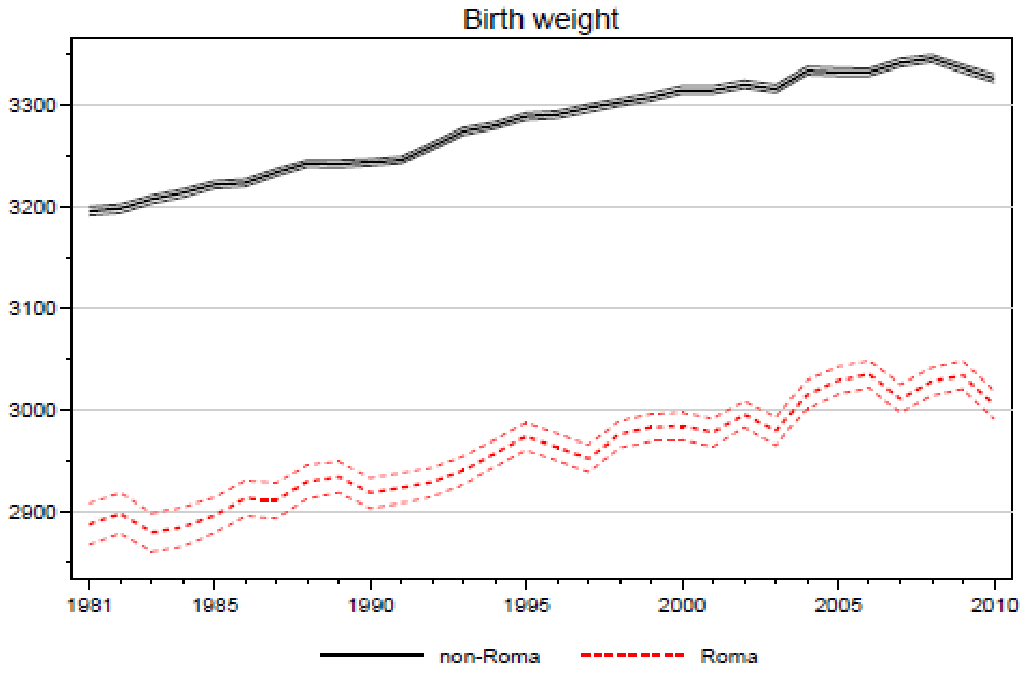
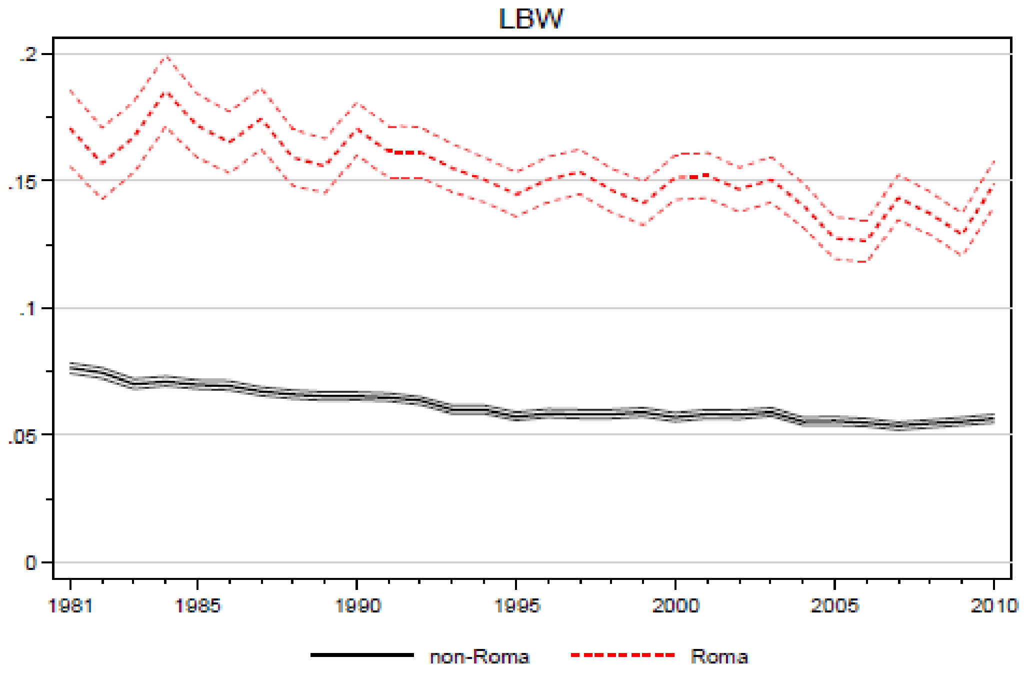


*Figure 2*

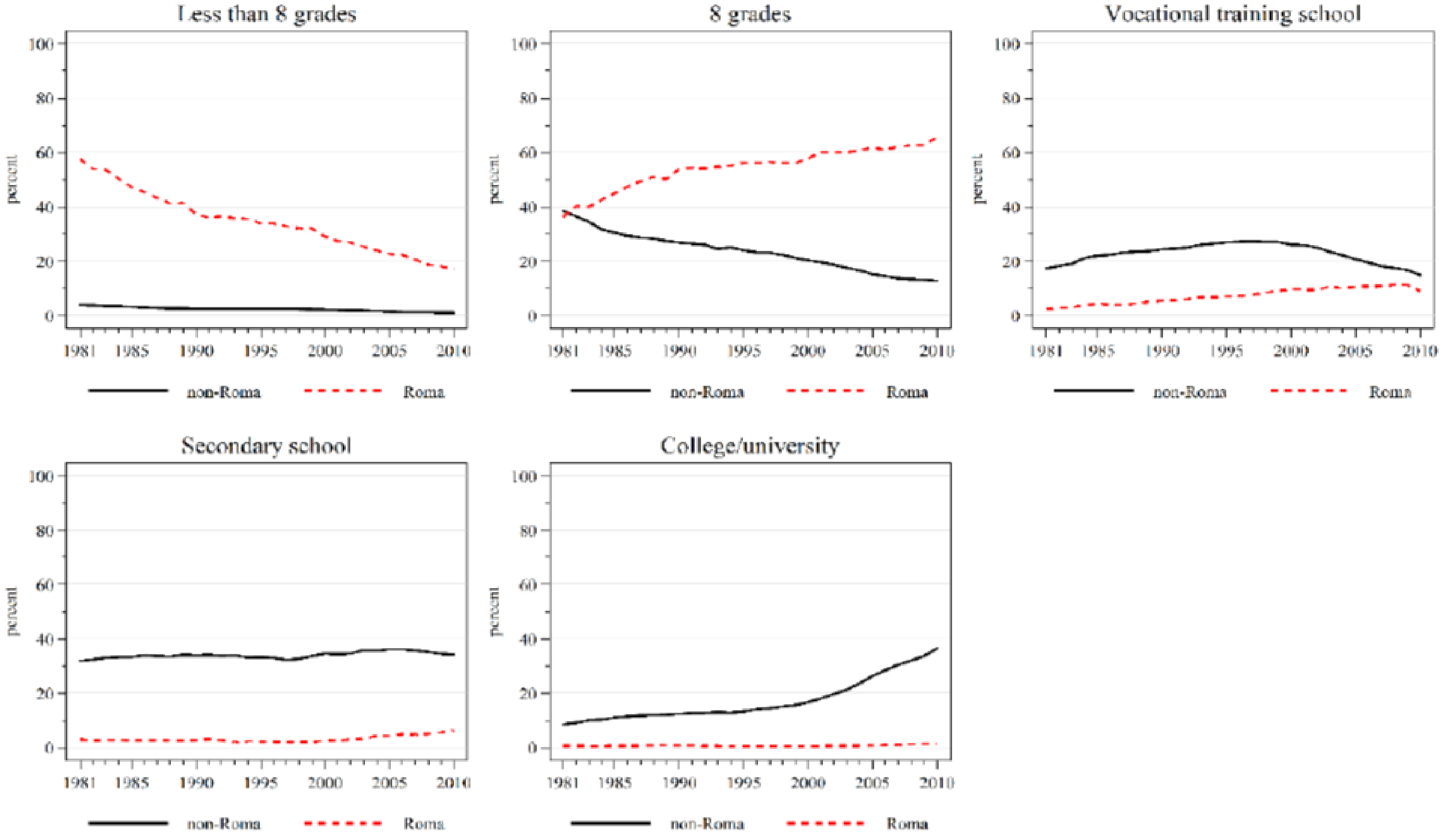
**Trends in the fraction of low birth weights**



# Continuation 4 – Research results

*Figure 9.*

**Trends in educational attainment of Roma (dashed line) and non-Roma (continuous line) mothers**

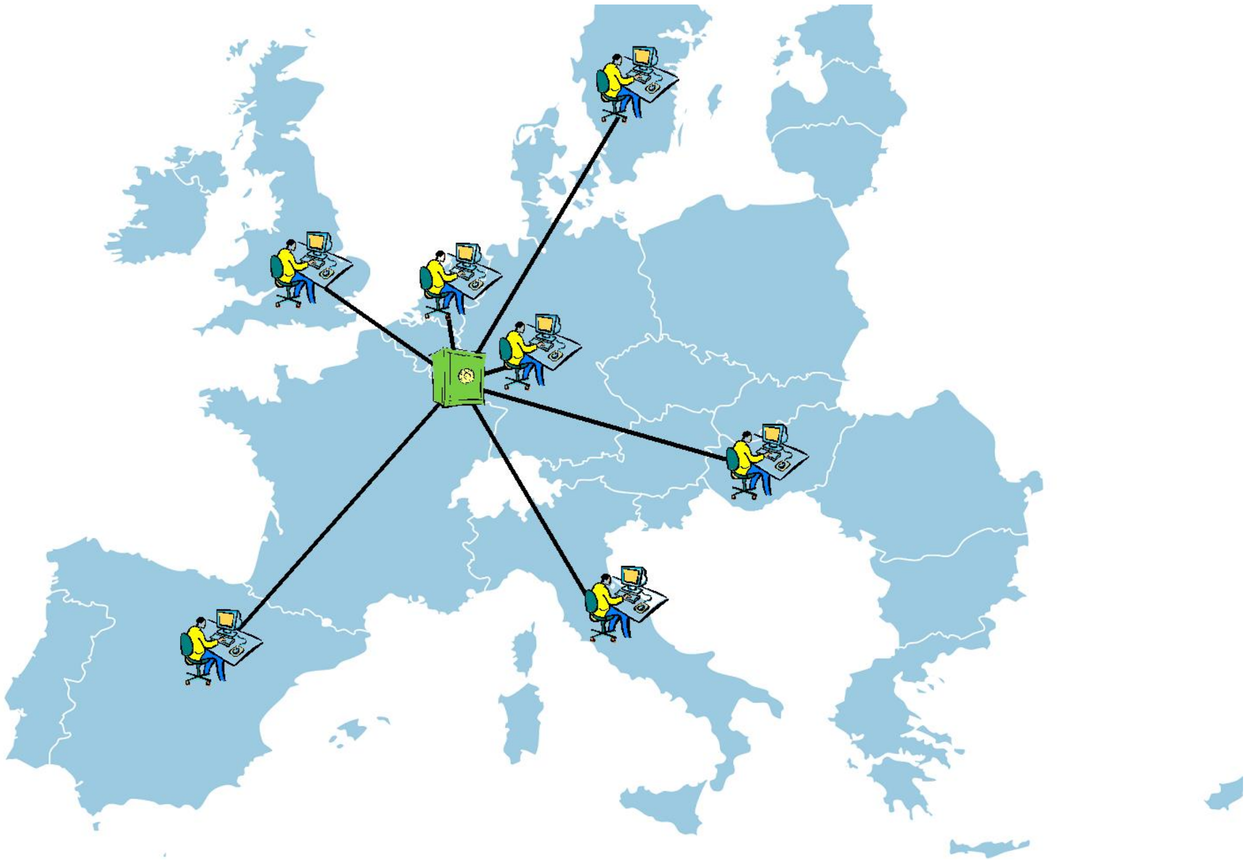


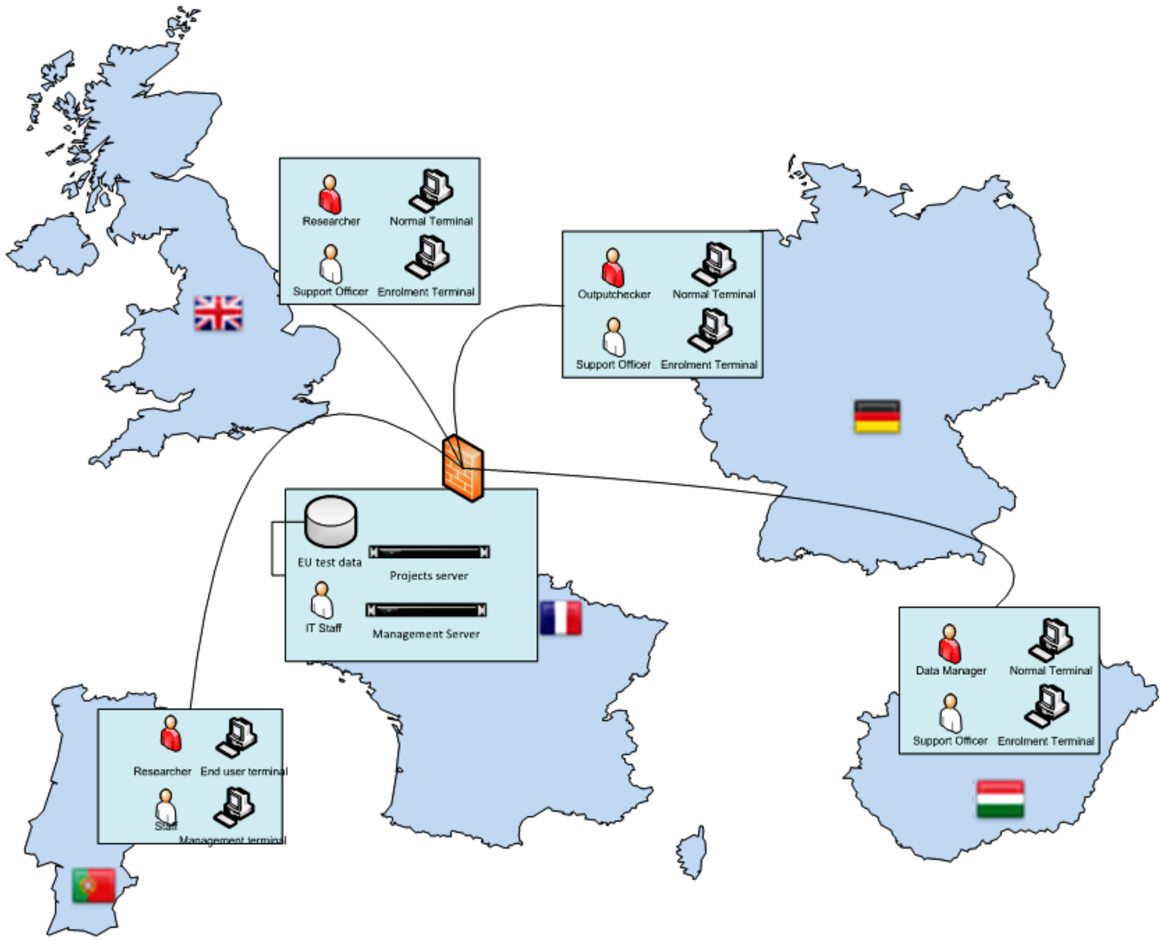
# The Future?

DARA project

(Decentralized And Remote Access to Confidential Data in the ESS)

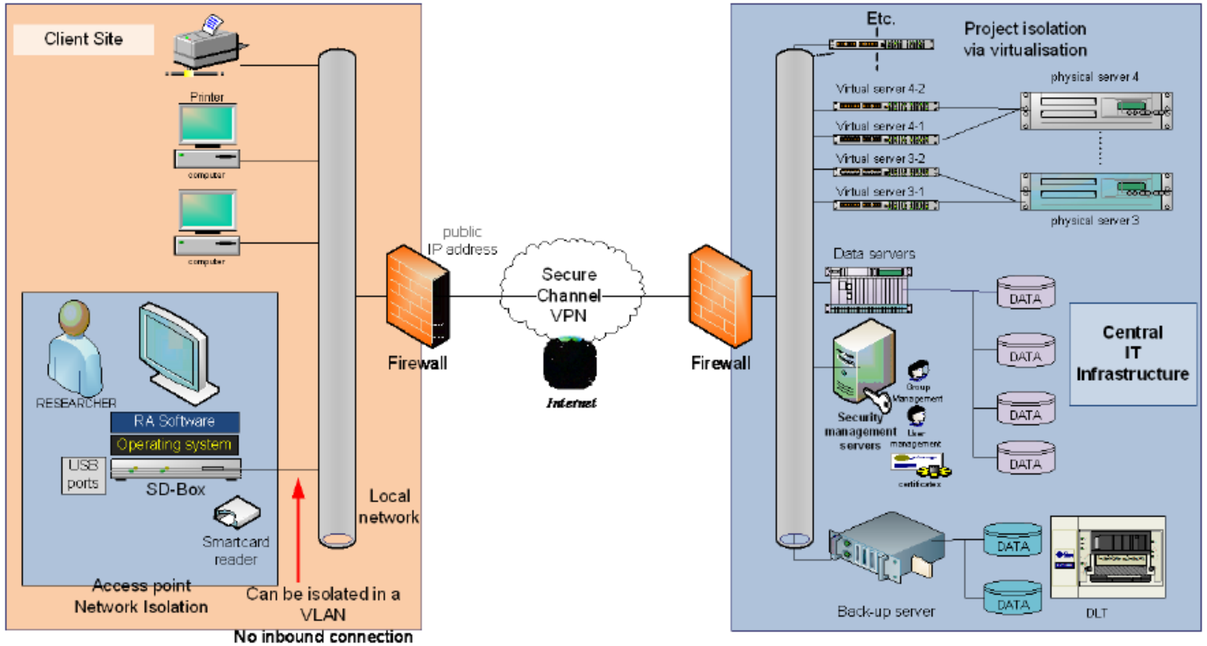
# Planned and realized DARA system



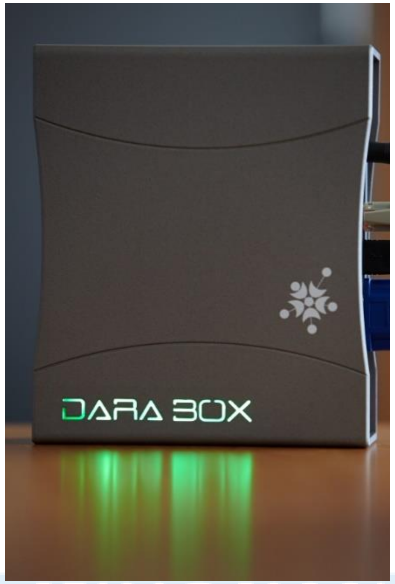


# Safe communication in the DARA – DARA box

- Biometric ID
- Prompt access to data at Eurostat from the SC of any member state



No inbound connection
Only port 443 open to a specific server



# Closing remarks

# Benefits of using micro data

- By using micro data the world appears in a completely new perspective
  - This breakthrough for social scientists is like the microscope for biologists or the radio telescope for astronomists
- Legal duty
  - European Statistics Code of Practice
  - National Statistics Code of Practice
  - The mission of the European Statistical System*: “We provide the European Union, the world and the public with independent high quality information on the economy and society on European, national and regional levels and make the information available to everyone for decision-making purposes, research and debate.”*
  - Regulation on European Statistics: „In order to align concepts and methodologies in statistics, an adequate interdisciplinary cooperation with academic institutions should be developed.” (REGULATION (EC) No 223/2009 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 March 2009)
  - Hungarian Act on Statistics
- Running a Safe Centre is profitable for statistical offices
- Micro data and Safe Centres are relevant tools for policy evaluations

# Difficulties

- Do statisticians like accessibility to micro data and Safe Centres?
  - ☹ ☹ ☹ vs. ☺
- A certain cultural change would be desirable in how statisticians perceive the importance of their profession
  - innovative approach
  - proactive behavior
  - co-operation and even more co-operation
  - commitment to producing useful, relevant and high quality statistics

# In conclusion

**A transparent and properly regulated use of micro data**
**in Safe Centers**
**should become best practice and**
**a relevant answer to the challenges of**
**Data Revolution and Algorithmic Society.**