RAIRD

Remote Access Infrastructure for Register Data

Johan Heldal[*], Elin Monstad[**], Terje Risberg[*] and Ørnulf Risnes[**],

[*]Statistics Norway (SN), [**]Norwegian Social Science Data Services (NSD)
<johan.heldal@ssb.no / ornulf.risnes@nsd.uib.no>

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality
Helsinki, Finland, 5 to 7 October 2015)

Statistisk sentralbyrå
Statistics Norway

NSD

The Research Council of Norway

The Situation in Norway

- Norway has a large number of registers of individuals established for administrative or statistical purposes.
- Examples of such registers:
  - The Central Population Register (CPR, administrative)
  - National Database of Educations (NUDB, statistical)
  - Registers for administration of the social security, welfare system, pensions and retirements, taxation etc.
  - Many others
- These registers are being merged by a unique person identification number to ***event history data bases*** that are updated every year.
- Demand for such data is **high**. Researchers want safe access without time consuming bureaucracy.

Research access today

- A few highly trusted researchers get access to some of these registers at their own premises (a weak point).
- Other researchers in approved institutions can apply for access to selected data for specific projects.
- The application procedure is cumbersome and time consuming.
- Permissions from the Norwegian Data Protection Authority and Statistics Norway, in some cases other register owners, are needed
- Research based on register data has shown a high pay off. We wish to extend their use to a wider range of researchers than those who have access today.

RAIRD

1. Remote Analysis server with event history data
2. Micro data are invisible (Anonymous User interface)
   - Only statistical output should be visible.
   - Metadata should be public domain

Metadata

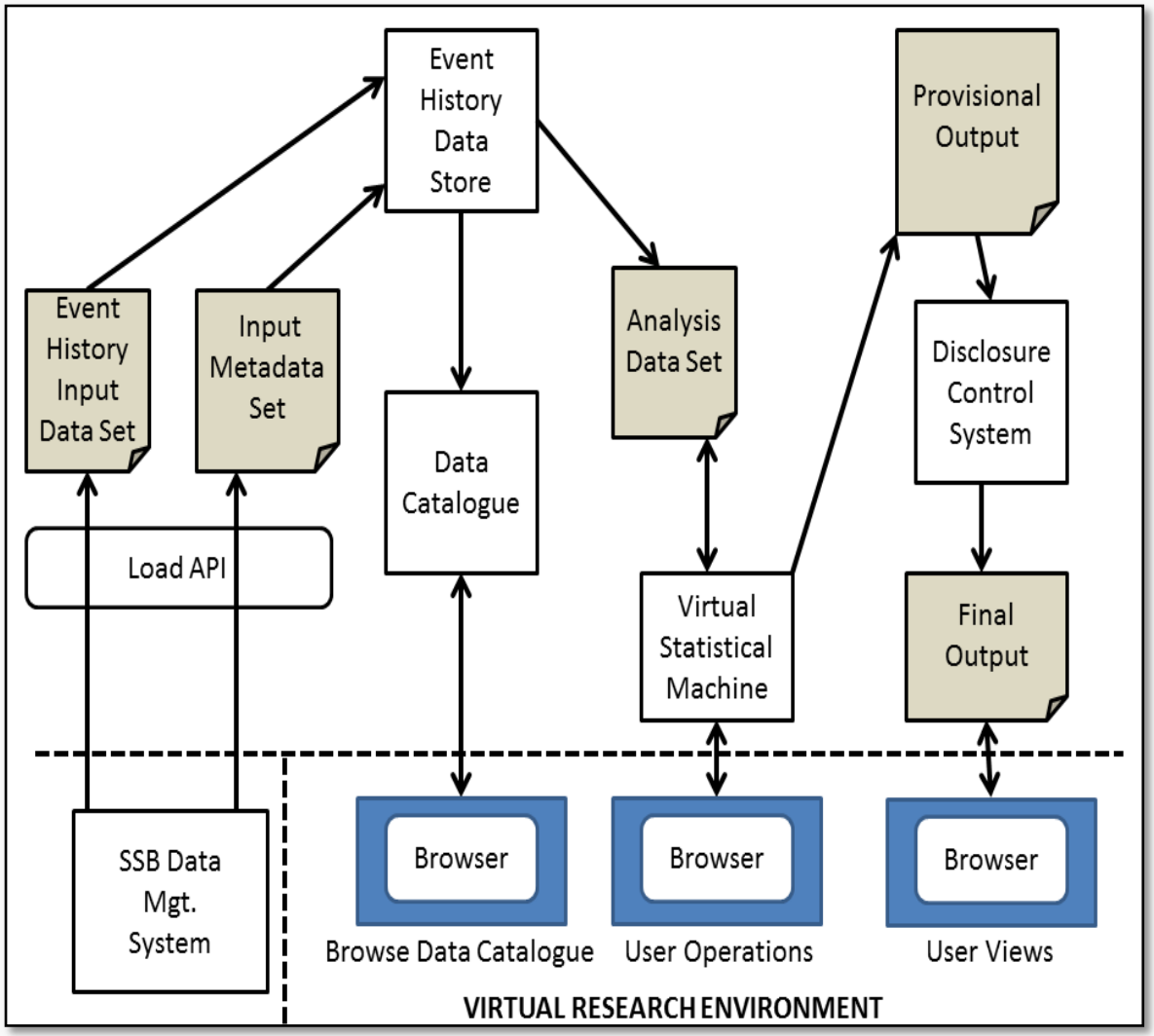
3. A way to get around legal procedures for access.
   - Simpler and cheaper(?) access for more researchers.
4. Requirement: Safe (enough) statistical output.

Comparison to neighbour countries

- Sweden offers similar microdata through the LISA database and the MONA-system (Microdata Online Access).
- Statistics Denmark offers microdata for researchers through its Research Services.
- Both countries provide access from researchers desks.
- In both countries microdata are visible on the screen.
- In Norway: To make register data accessible with RA without the extensive application process, microdata must be invisible (AUI).
- AUI shifts the DC focus from microdata record disclosure to the statistical output.

The RAIRD information model

All operations go through the **V**irtual **S**tatistical **M**achine. **VSM** allows controlled exposure of methods and Managed Execution Environment.



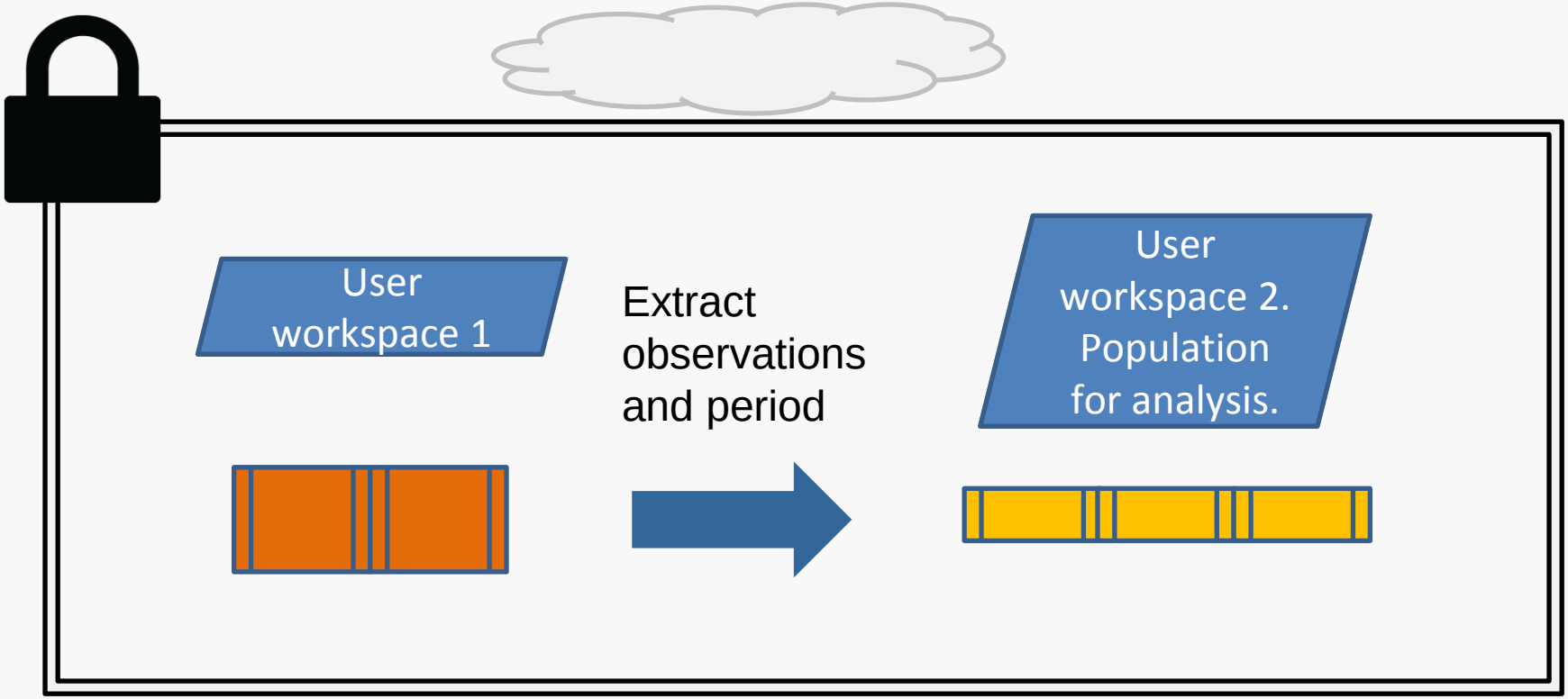
Extract variables from Data Store

VRE

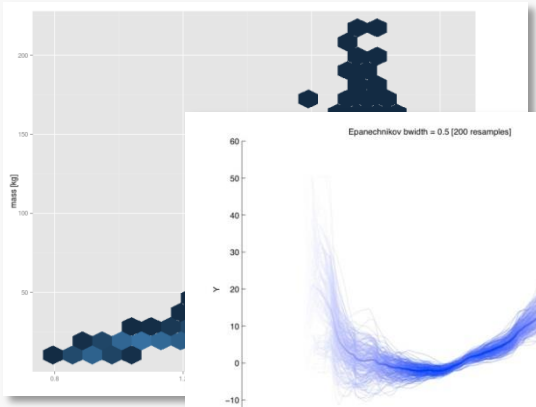
Data-Store

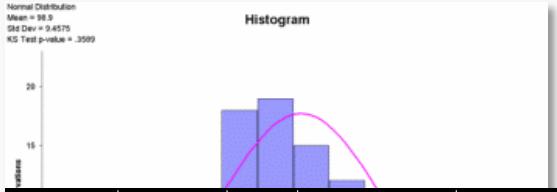
User workspace

Extract observations



Analyze data

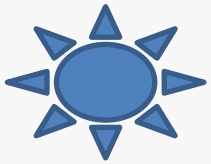





```
. regress  gdppcap xgrowth mgrowth consump popgrowth

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  4,    34) =   73.35
       Model |  139.845267     4  34.9613167            Prob > F      =  0.0000
    Residual |  16.2059352    34  .476645151            R-squared     =  0.8961
-------------+------------------------------           Adj R-squared =  0.8839
       Total |  156.051202    38  4.10661058            Root MSE      =  .69039

------------------------------------------------------------------------------
    gdppcap |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
    xgrowth |   .0966019   .0210758     4.58   0.000     .0537706    .1394331
    mgrowth |   .0867174   .0269233     3.22   0.003     .0320027    .1414321
    consump |   .8316915   .1225229     6.79   0.000     .5826951    1.080688
  popgrowth |  -.9844089   .8412343    -1.17   0.250    -2.694003    .7251848
      _cons |  -.7584808   1.009879    -0.75   0.458    -2.810803    1.293841
------------------------------------------------------------------------------
```



User workspace 2

Statistical Analysis

Safe results

Target Groups

- The system should serve the medium and less sophisticated researchers (low/medium access level)
- PhD and master students should be allowed access.
- The ambition is to develop the system to satisfy more and more advanced users.
- We want to provide access directly to users desks, *not* through Safe Centres.

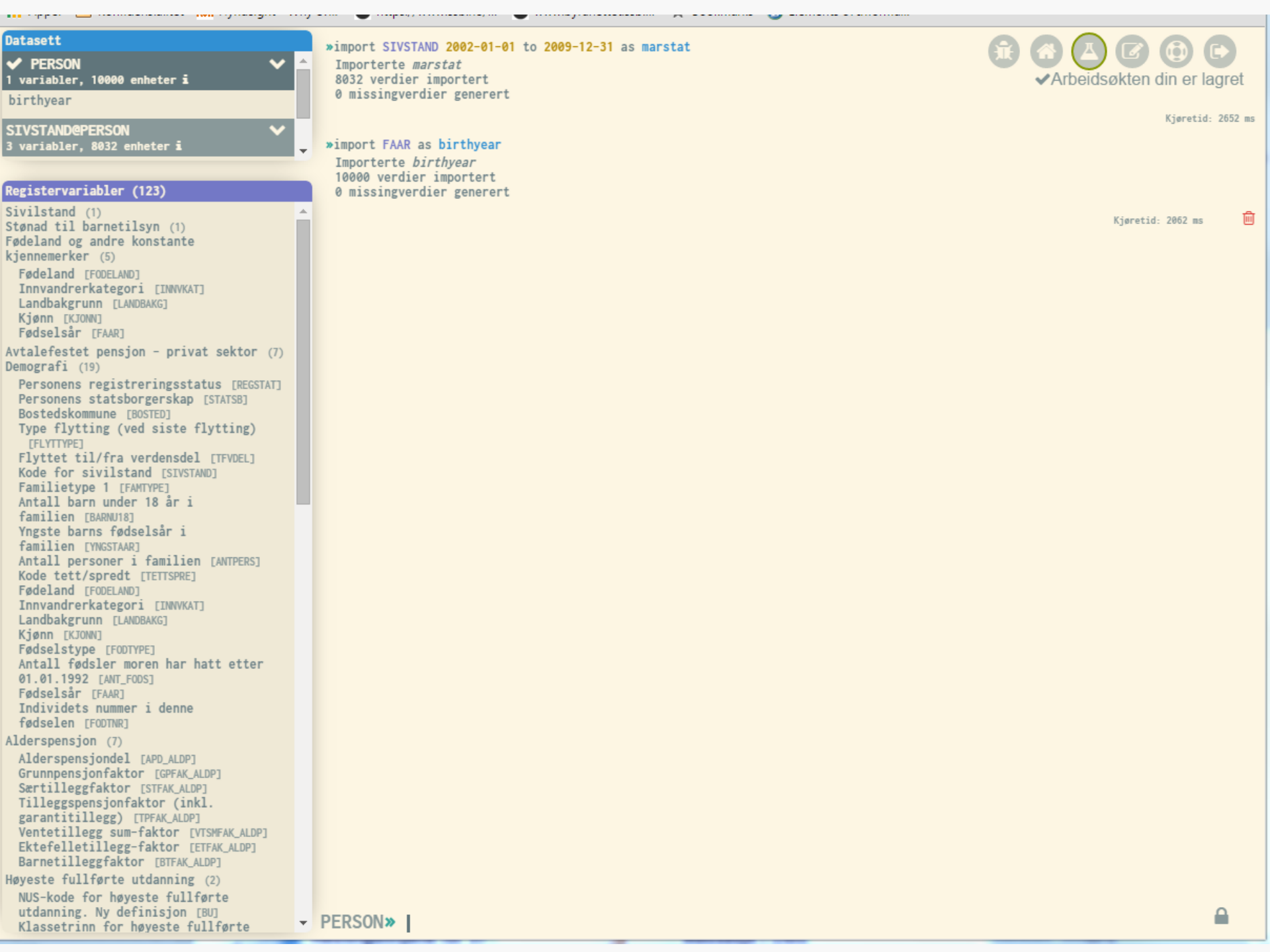
Prototype

- NSD is working on a prototype for an AUI solution.
- With AUI we cannot use any standard software.
- Basic software: Python programming environment (SciPy, NumPy, Pandas, Statsmodels)
- We establish a STATA-like command interface.

```
>> import HOVED 2002-01-01 to 2010-12-31 as jobseeker
  Imported jobseeker
  8368 values imported
  0 missing generated
>>drop if jobseeker !=1
  5627 units were removed from the data set.
>>
```

Risks associated with RAIRD - 1

- With AUI disclosure risk will be associated with the statistical output only, not with micro data themselves.
- **In general:** We consider disclosure attacks that can be considered *realistic scenarios.*
- Tabular output: Zeroes, Small counts, group disclosure.
- Aggregations: Dominance
- *Differencing* is a challenge for tabular output.
- Graphical output
  - Scatter plots, residual plots, boxplots
  - Can be fixed with existing software.


Datasett
PERSON
1 variabler, 10000 enheter
birthyear
SIVSTAND@PERSON
3 variabler, 8032 enheter
Registervariabler (123)
Sivilstand (1)
Stønad til barnetilsyn (1)
Fødeland og andre konstante kjennemerker (5)
Fødeland [FODELAND]
Innvandrerkategori [INNVKAT]
Landbakgrunn [LANDBAKG]
Kjønn [KJONN]
Fødselsår [FAAR]
Avtalefestet pensjon - privat sektor (7)
Demografi (19)
Personens registreringsstatus [REGSTAT]
Personens statsborgerskap [STATSB]
Bostedskommune [BOSTED]
Type flytting (ved siste flytting) [FLYTTYPE]
Flyttet til/fra verdensdel [TFVDEL]
Kode for sivilstand [SIVSTAND]
Familietype 1 [FAMTYPE]
Antall barn under 18 år i familien [BARNU18]
Yngste barns fødselsår i familien [YNGSTAAR]
Antall personer i familien [ANTPERS]
Kode tett/spredt [TETTSPRE]
Fødeland [FODELAND]
Innvandrerkategori [INNVKAT]
Landbakgrunn [LANDBAKG]
Kjønn [KJONN]
Fødselstype [FODTYPE]
Antall fødsler moren har hatt etter 01.01.1992 [ANT_FODS]
Fødselsår [FAAR]
Individets nummer i denne fødselen [FODTNR]
Alderspensjon (7)
Alderspensjondel [APD_ALDP]
Grunnpensjonfaktor [GPFAK_ALDP]
Særtilleggfaktor [STFAK_ALDP]
Tilleggspensjonfaktor (inkl. garantitillegg) [TPFAK_ALDP]
Ventetillegg sum-faktor [VTSMFAK_ALDP]
Ektefelletillegg-faktor [ETFAK_ALDP]
Barnetilleggfaktor [BTFAK_ALDP]
Høyeste fullførte utdanning (2)
NUS-kode for høyeste fullførte utdanning. Ny definisjon [BU]
Klassetrinn for høyeste fullførte
»import SIVSTAND 2002-01-01 to 2009-12-31 as marstat
Importerte marstat
8032 verdier importert
0 missingverdier generert
Kjøretid: 2652 ms
»import FAAR as birthyear
Importerte birthyear
10000 verdier importert
0 missingverdier generert
Kjøretid: 2062 ms
Arbeidsøkten din er lagret
PERSON»

Risks associated with RAIRD - 2

- Regressions:
  - High leverage points.
  - Cross product matrices
- We miss studies of disclosure risks for event history settings, but believe that scenarios for cross sectional data still apply.
- Inferential attacks are not seen as very likely.

We want solutions that

- do not distort the microdata (no inconsistencies)
- provide useful analysis and statistical output
- avoid costly manual output control.
- embrace as many statistical methods as possible.
  - We have to start with the most basic.

Dealing with risk 1

- User authorisation
  - Different access levels for different researchers.
- Strict access control.
- Logging of all activity on the system.
  - All activity can be reproduced
  - Store logs as data that can be routinely analysed with RAIRD itself.
- Censoring tables or analysis with
  - too few observations (e.g. 20)
  - too few obs. per cell or estimated parameter in average (e.g.10)
  - too sparse tables

Dealing with risk 2

- Drop observations at random to imped differensiation (Sparks et al. (2008))
- Automatic output control.
  - Add noise to *results* on the fly.
  - Noise *fixed* for repeated runs of the same output.
  - We are inspired by the Australian Data Analyser.
- Focus on robust statistical methods.
  - Reduces the effect of high leverages.
- Hexbin plots, modified boxplots
- Possibly: Adding noise to estimating equations
  - Chipperfield & O'Keefe (2014)

Compared to the present situation

- RA with AUI will be a way to get around legal obstacles for access to register micro data in Norway.
  - Easier access for more researchers
- The practice of giving access at own premises should seize.
- RAIRD will take micro data safety to a new level in Norway.

Researchers attitudes

- Positive to remote access as such
  - Need not have the responsibility for storing confidential data safely at their own premises.
  - As far as they can do what they have been able to do until now.
  - Appreciate less bureaucracy for access.
- Reluctant to Anonymous User Interface
  - They are used to be able to see microdata.
  - We mean that is a question of habit and culture.
- They have their spikes out when they hear the word “noise”.
  - They believe we mean noise on micro data.

Conclusion – Short run

- We must build RAIRD *bottom-up, not top-down*.
- Start with access to the less advanced researchers (master and PhD students) with safe output for the more basic methods.
- Implement higher access levels with more advanced methods with less disclosure control for the more advanced and trusted researchers.

Conclusion – Long run

- Implement disclosure control methods with safe output for more advanced statistical methods.
- Win advanced researchers' confidence in the system and its disclosure control methods.
- We hope the RAIRD technology once will be established as the standard for all access to statistical microdata in Norway.