O

Big (Large) data and Statistics Denmark

Christian Torp-Pedersen

September 23, 2022

Wellcome to Biostat

Denmark

- Population of 5.6 million
  - Caucasians 90%
  - One 10-digit number for all registration in the country
  - Tax – Education – housing – birth – family relations – death – work - pension
- Healthcare system
  - Goverment run
  - Tax financed
  - Free of charge
  - Equal access to everyone
  - all prescriptions registered because of reimbursement

Danish Registers

Research Environment Statistics Denmark

- Provides access to >200 nationwide registers
- Can be coupled with register from Danish National Health Authority
- Other data can be uploaded
- Data can enter Statistics Denmark, but never leave
- Only aggregate data with at least 3 people can leave
- Use needs authorisation
- Use is monitored
- Draconic penalty for violations

Statistic Denmark don't-s

- Never disclose data with 3 or less individuals
- Avoid indirect disclosure by percentages etc.
- Do not use encrypted cpr in program logic
- Avoid unauthorized data export (photo, screendump)

Statistics Denmark do-s

- Export Your programs - Science documentation
- Big data - check your consumption of memory
- gc() with R - reduce RAM consumption

Case 1 - Does Insulin cause cancer?

- Insulin is a growth factor
- Cancer could be influence by a growth factor

Step 1 - Find diabetes diagnoses

- LPR2
  - t_adm - recnum, start, end, cpr(pnr)
  - t_diag - recnum - diag
- LPR3
  - kontakt - kontakt_id, start, end, cpr
  - diagnoser

Pattype

- Pattype in LPR2 defines type of admission
  - 0=24 hour admission 2=Outpatient
- No pattype in LPR3
  - Length of contact>12 hours?
  - Type of contact acute?

More Caveats

- The unit is a course – Danish: Forløb
- The hospital is reimbursed based on the most expensive diagnosis they register or the most expensive operation.
- The diagnosis is the discharge diagnosis – implying potential severe delays for out patient diseases
- A diagnosis that dominates the clinical situation is usually registered correctly, otherwise not
- Many rare diagnoses are correctly registered in specialist unit, otherwise not
- Operations are registered correctly as surgeons use them for creating procedure lists and because of reimbursement.
- Psychiatric diagnoses MAY be among the somatical diagnoses, but only if the patient appear both as a somatic and psychiatric patient

Diabetes

- Many treated by general practitioners that do not register diagnoses, but deliveries (ydelser) such a "visit", "visit with prescription"...
- Patients are not hospitalised for diabetes, but the diseases cause by diabetes such as cardiovascular disease

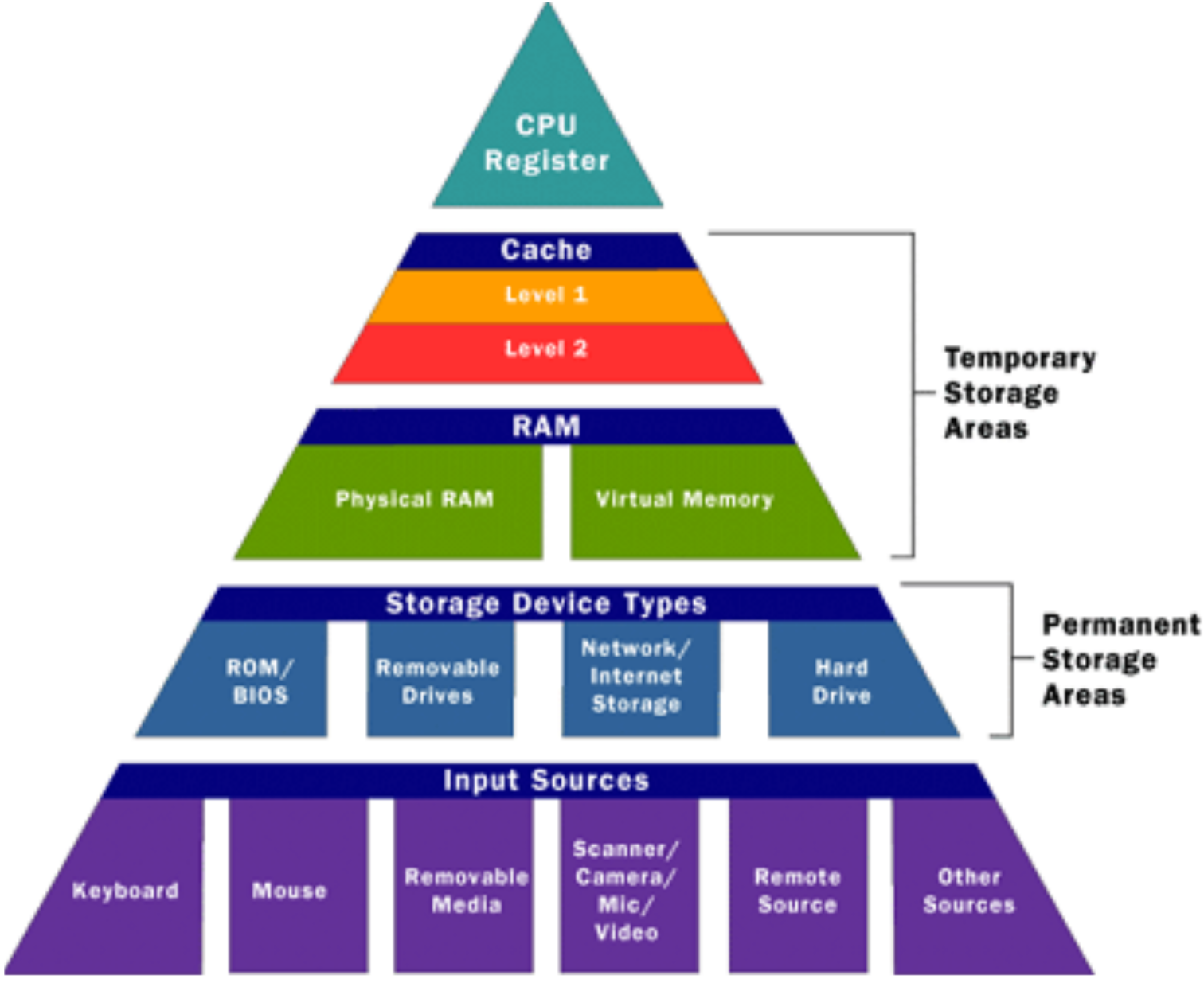
Register of Medicinal Product Statistics

Lægemiddelstatistikregisteret

- About 100 variables
- Date, ATC (drug code), amount
- About 10^9 prescriptions – roughly 80 GByte
- Both positive and negative (corrections) – if ignored zero can become two
- Diabetes has (almost) specific treatment
- Other diagnoses (eg. Psoriasis) have almost specific treatment, and some diseases (eg. Deep vein thrombosis) and be ascertained by requiring specific treatment.

Extract first diabetes prescription from SAS to R

- Use an R-package to read the SAS data directly – unbearably slow
  - library(haven)
  - myData <- read_sas("c:\\.sas7bdat")
- Use STATTRANSFER to translate the data to R – unbearably slow, but sometimes it is the easiets way to get DATES right
- Let SAS do the hard work efficiently in the background!


CPU
Register
Cache
Level 1
Level 2
RAM
Physical RAM
Virtual Memory
Temporary
Storage
Areas
Storage Device Types
ROM/
BIOS
Removable
Drives
Network/
Internet
Storage
Hard
Drive
Permanent
Storage
Areas
Input Sources
Keyboard
Mouse
Removable
Media
Scanner/
Camera/
Mic/
Video
Remote
Source
Other
Sources

SAS Datastep

```
Data work.diabetesmed; set perm.medication;
if ATC=:'A10';
keep id date ATC medication;
Run;
```

- A SAS datastep is compiled first
- It reads ONE record at a time (in reality a computer buffer)
- It records the result on temporary disk
- Finally copies the temporary file to disk
- SAS is highly efficient with modern disks

Smarter

```
Data work.diabetesmed; set perm.medication;
if ATC=:'A10';
keep id date ATC medication;
Run;
Data work.diabetesmed; set perm.medication
(keep= id date ATC medication where=(ATC=:'A10'));
Run;
```

Expand: Use ODS from sas to output variables to a csv-file to be read with fread from data.table

www.github.com/tagteam/heaven

```
diabmed <- heaven::importSAS('file-path-name',
where='atc=:"A10"',
filer=pnrobject,
keep=c(....))
```

Thrombosis with tranexamic acid