Beyond the limits of CBS RA

Efficient programming and the ODISSEI Secure Supercomputer

Erik-Jan van Kesteren

Before we start…

About me

- Assistant professor at **Methodology & Statistics**, Utrecht University
- Team lead for the **ODISSEI Social Data Science (SoDa) team**
  advancing data- & computation-intensive research in social science
- Strong advocate for **open science**

About today

- We will **not solve all your computation problems** in less than an hour ☺
- I will show **R code**, but the principles hold for other environments too
- If you work at an **ODISSEI member organisation**: you can get help from the ODISSEI Social Data Science team!
- https://github.com/sodascience/cbs_microdata_computing

Today

- The CBS RA technical
- The trinity of trouble
  - Storage
  - Memory
  - Compute
- Tips & question time!

The CBS RA: technical

CBS RA technical

The CBS RA server

- One big physical computer
- Virtual machines: you log in via Citrix and a “windows computer” is instantiated for you
- 4 virtual CPUs, 48Gb RAM & 100Gb storage
- Heavy option: 128Gb RAM

Resources are limited and (FOR NOW) shared across users

- Disk access is shared
- Memory is shared
- Computation is shared

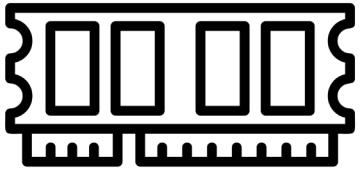
Top tip #1

Run your heavy tasks during low - intensity hours on the RA environment

Tackling the trinity of trouble

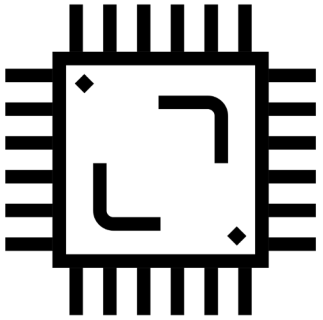
storage

Hard Disk by Creative Stall
from NounProject.com



memory

CPU by Liberus from
NounProject.com

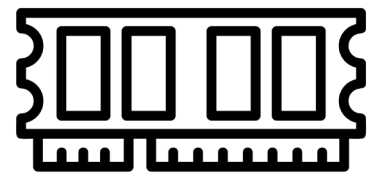


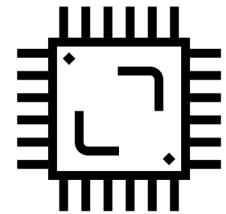
compute

CPU by DinosoftLab from
NounProject.com

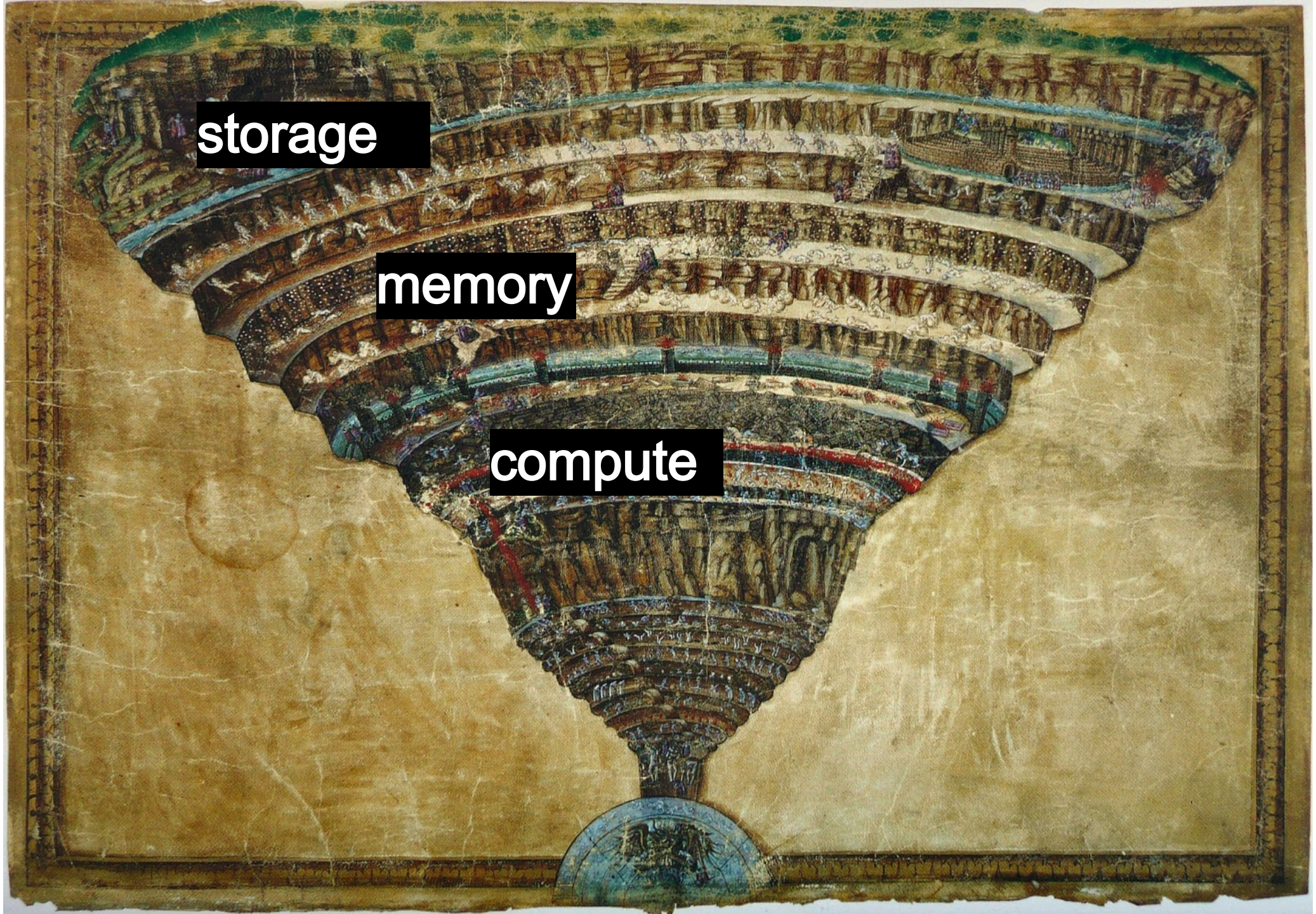
loading
importing

saving
storing





processing


storage
memory
compute


data do not fit on disk
cannot allocate vector of size 223.1 Gb
?? Omg this will take forever!

Storage

Geachte relatie,

Uit een meting op maandag 4 april 2022 blijkt dat project 0000, Titel van het project, een ruimtebeslag kent van **133** GB. De limiet voor het project is **100** GB.

Als u de extra capaciteit daadwerkelijk nodig heeft, dan kunt u een verzoek indienen om extra capaciteit bij te kopen. De kosten hiervoor bedragen 25 euro per 50 GB per maand.

Met vriendelijke groet,

Firstname Lastname
DBD Team Dataservices

CBS | Henri Faasdreef 312 | Postbus 24500 | 2490 HA Den Haag
Email: microdata@cbs.nl
Volg statistiekcbs op twitter | facebook | instagram

Top tip #2

If you can afford it, just buy extra storage space for your project ☺

Efficient project folder structure

```
my_project/
├── raw_data/
│   ├── questionnaire_data.csv
├── processed_data/
│   ├── questionnaire_processed.rds
│   ├── analysis_object.rds
├── img/
│   ├── plot.png
├── 01_load_and_process_data.R
├── 02_create_visualisations.R
├── 03_main_analysis.R
├── 04_output_results.R
├── my_project.Rproj
├── readme.md
```

Efficient project folder structure

```
my_project/
├── raw_data/
│   └── questionnaire_data.csv
├── processed_data/
│   ├── questionnaire_processed.rds
│   └── analysis_object.rds
├── img/
│   └── plot.png
├── 01_load_and_process_data.R
├── 02_create_visualisations.R
├── 03_main_analysis.R
├── 04_output_results.R
├── my_project.Rproj
└── readme.md
```

At CBS, this is on a different disk!
Does not count towards your 100GB quote

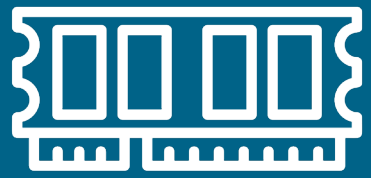
Make these objects efficiently stored
Depends on your application

Top tip #3 (small open science digression)

Create a clear code folder, export your code from the RA, and publish it!

Efficiently storing large R datasets

Live coding 1



Memory

Top tip #4

Read your program's error messages!
They give a lot of diagnostic info

In R:

Error: cannot allocate vector of size 745.1 Gb

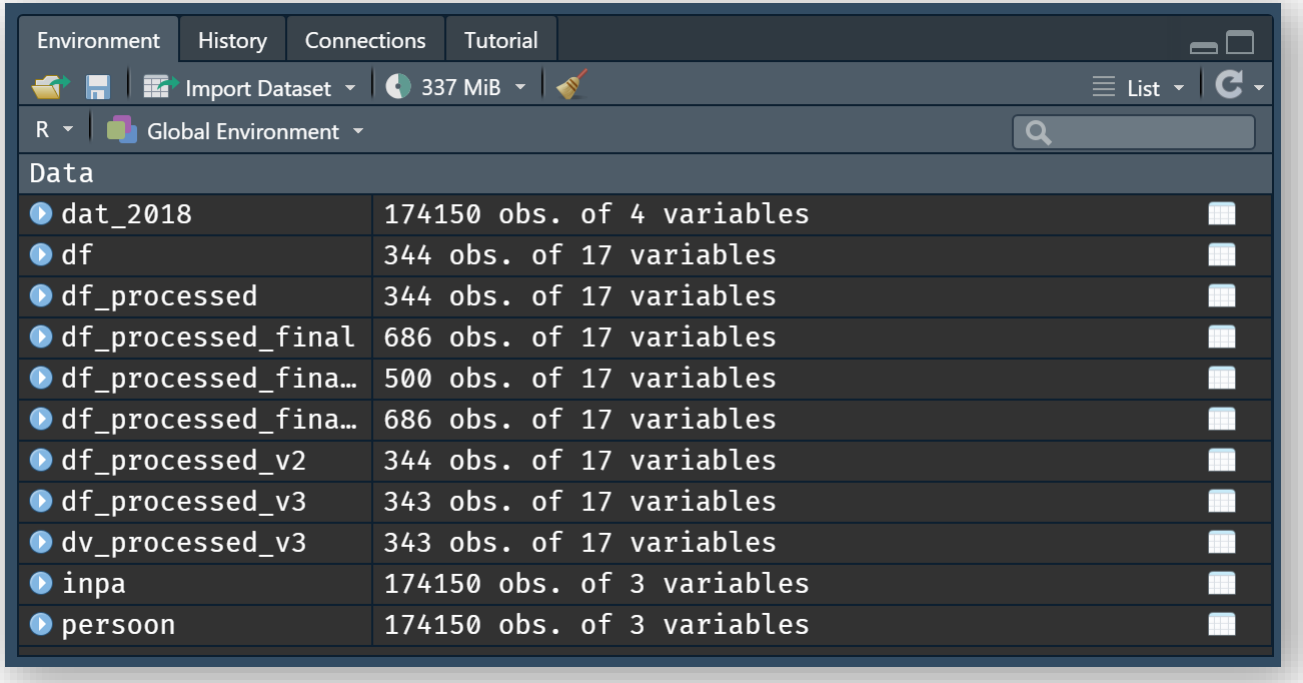
In Python (numpy)

numpy.core._exceptions._ArrayMemoryError: Unable to allocate 745. GiB for an array with shape (100000000000,) and data type float64

In Stata

(no clue, I really don’t use Stata??)

Clean your session / environment



Efficiently processing large datasets

Live coding 2

Larger-than-memory data

- Sometimes, your data really is larger-than-memory
- It is possible to do analyses on datasets which are on-disk

Two options:

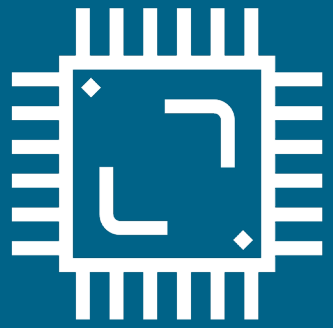
- Create chunked data objects
- Create a proper database

Top tip #5

Investigate whether the “heavy” RA machine will solve your memory issues

Visualisation & regression with larger -than- memory data

Live coding 3



Compute

Compute-heavy applications

- Large simulations, e.g.,
  - agent-based models
  - computational models
  - complex systems stuff
  - statistical simulations (large power analyses)
- Many different conditions
  - Perform some computation for each neighbourhood in NL
- Bayesian estimation with large models (many parameters, many posterior samples)

Speeding up a function with C++

Live coding 4

Embarrassingly parallel

Many independent computations, l ittle or no effort is needed to separate the problem into a number of parallel tasks

- Simulations
- Applying a function to many conditions
- Running a piece of code with many different settings
- Bootstrapping

Top tip #6

Is your problem parallelizable? Look into the ODISSEI Secure Supercomputer

Training

Supercomputing for Social Scientists with R

Would you like to understand how to work with a supercomputer and translate your R workflow from a graphical-user-interface (GUI) on your desktop to a scripting/automated workflow that leverages the resources of a supercomputer?

Sign up

24 May 2022

9.00-17.00

Online

Sign up

Event type	Prerequisite knowledge	Costs
Training	No prior knowledge required	Free

Top tips, collected

- Run your heavy tasks during low-intensity hours on the RA environment
- If you can afford it, just buy extra storage space for your project ☺
- Create a clear code folder, export your code from the RA, and publish it!
- Read your program's error messages! They give a lot of diagnostic info
- Investigate whether the "heavy" RA machine will solve your memory issues
- Is your problem parallelizable? Look into the ODISSEI Secure Supercomputer
- Want to know more? Join the workshop.

Thank you!

https://odissei - data.nl

https://www.surf.nl/en/agenda/supercomputing - for - social - scientists - with - r

https://github.com/sodascience/cbs_microdata_computing

@SoDa_NL

Questions?

Default light slide

Default subheading

This is the body of the text

Default subheading

Note that the text is not black, but “black, text 1, lighter 25%”

Default subheading

This makes things easier on the eyes

Default subheading

This is the body of the text

Default dark slide

Default subheading

The dark slide brings some variation

Default subheading

It can highlight important aspects of the presentation.

Default subheading

This is the body of the text

Default subheading

This is the body of the text

Is this an impact slide?

Here is an impactful slide with a sentence on it.

Here is a topic related to the aforementioned question.