TALL SOM FORTELLER

{dapla}

Or our migration to the cloud and open source software

MORTEN WIIG LETNES

Statistisk sentralbyrå
Statistics Norway

Agenda

- The data platform in general
- GIS in the cloud
- Open source GIS
- Positives and negatives

What is Dapla

- A **da**ta **pla**tform: a place for working on data with code (in the cloud)

MS stock photos

Statistisk sentralbyrå
Statistics Norway

Old platform

New platform

Kongsvinger

Do-It-Yourself

Google Cloud

Finland

sas

R

jupyter

python™

ORACLE

GitHub

git

ArcGIS Pro

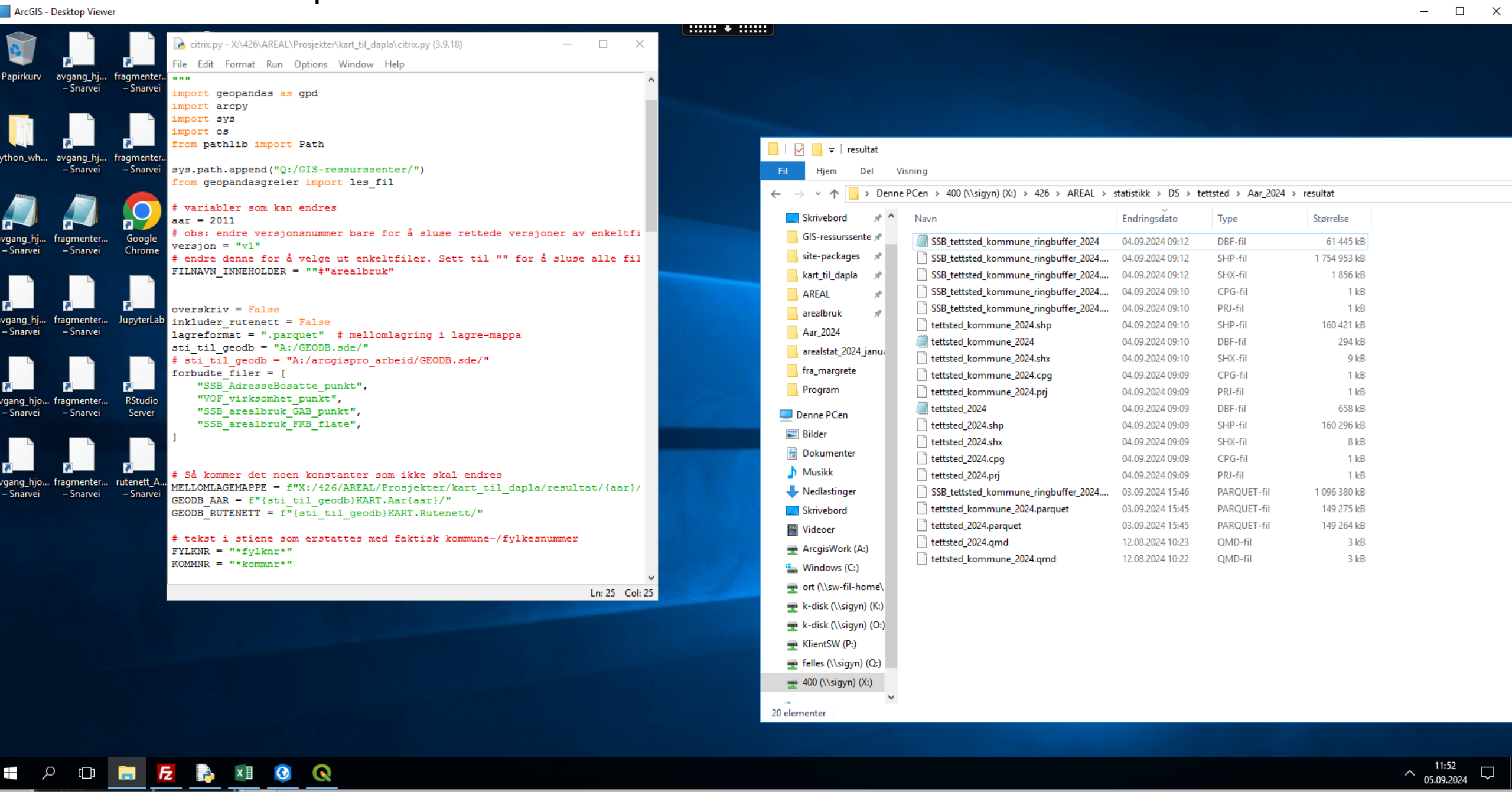
Statistisk sentralbyrå
Statistics Norway

- So instead of paying for licences to software, we pay Google for storing our data and providing hardware
- Will it be cheaper? Difficult to say.

Main differences to the old platform

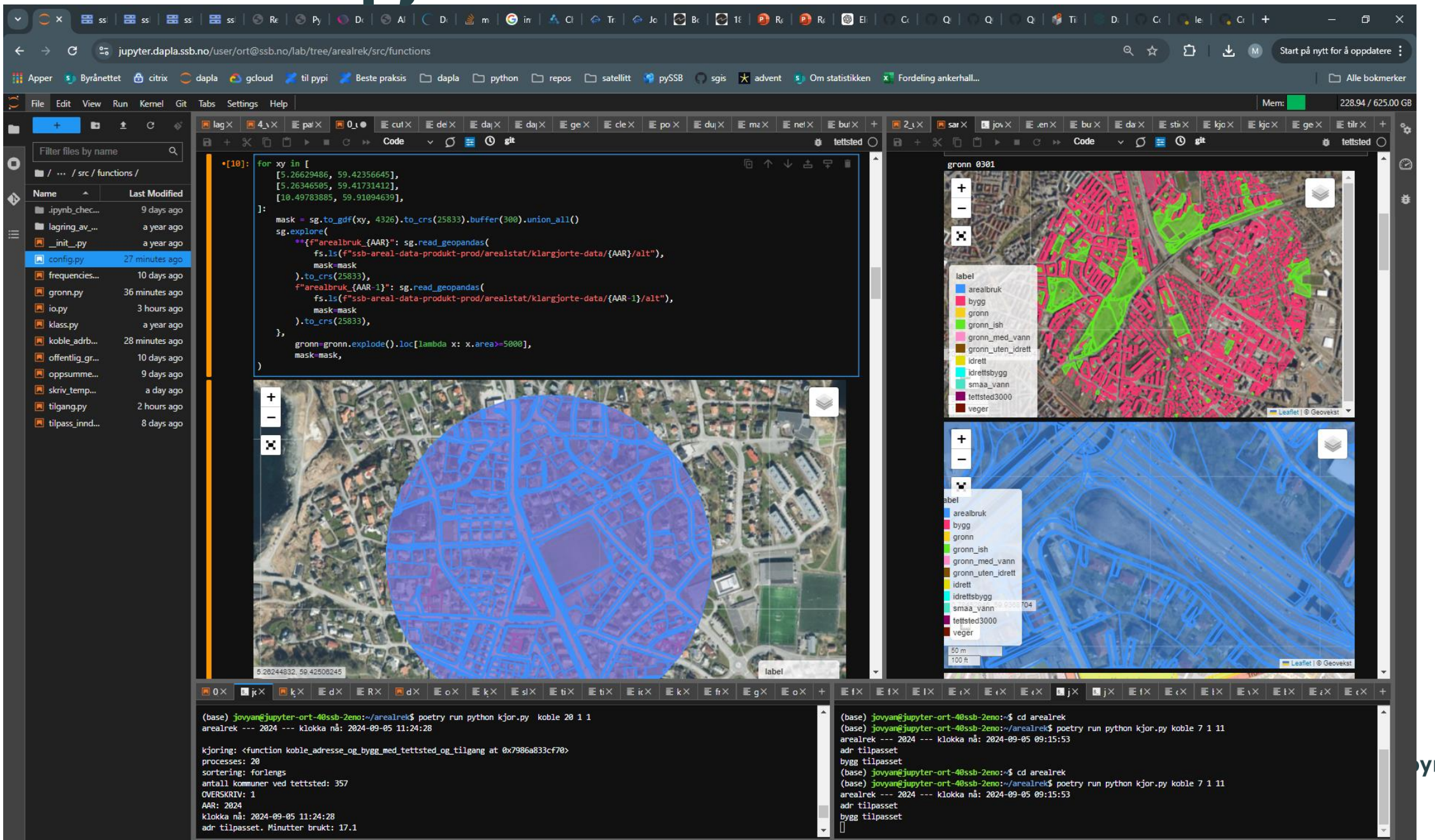
- Data stored in the «cloud»
- Mostly open souce software
- More code based – less «model builder», less manual editing

Old platform

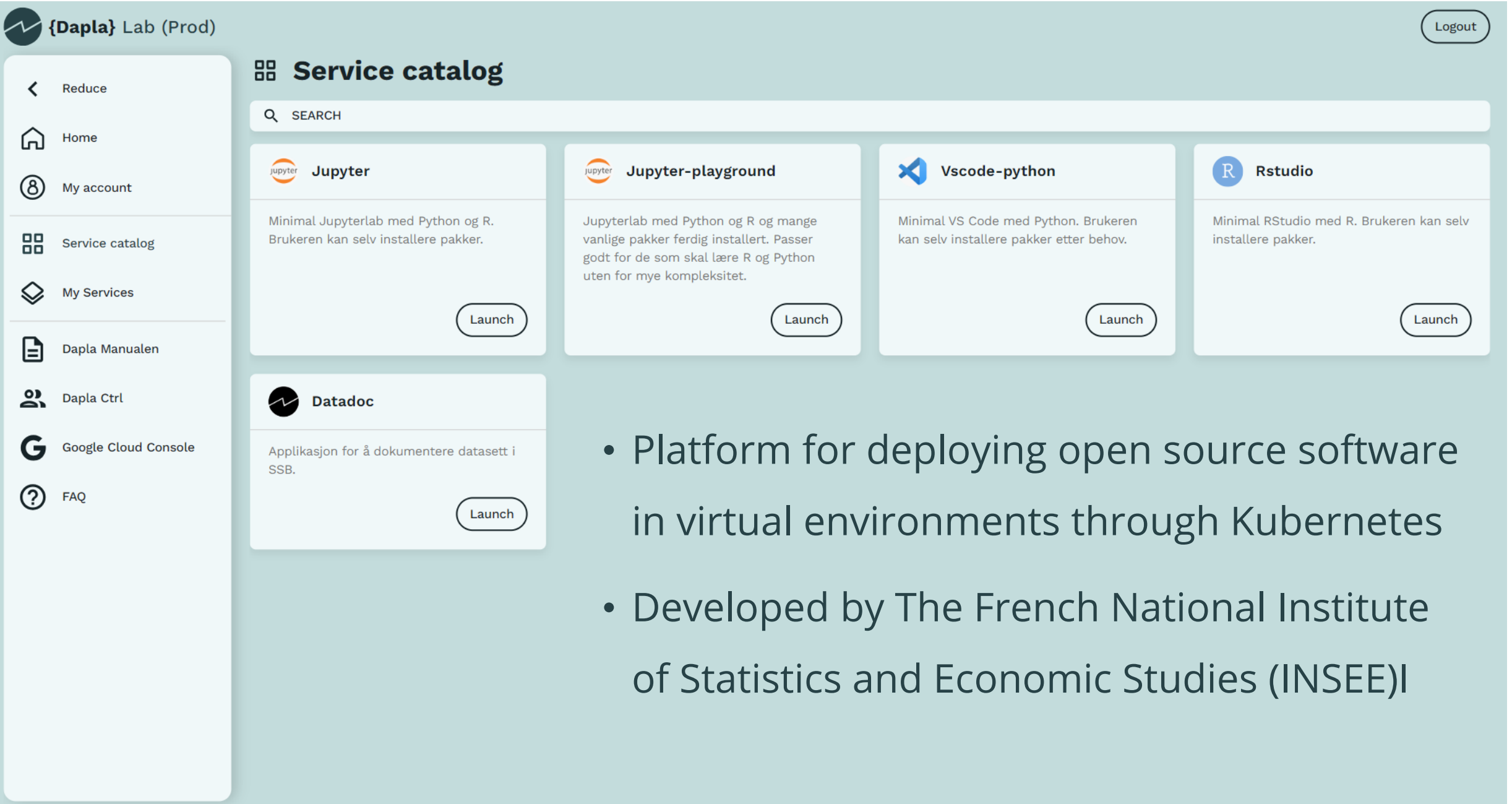


New platform

Production: Jupyter Notebook or terminal



New: Onyxia, more software for production



- Platform for deploying open source software in virtual environments through Kubernetes
- Developed by The French National Institute of Statistics and Economic Studies (INSEE)I

Coming soon to Onyxia: QGIS

Code: GitHub

statisticsnorway / tettsted

Type / to search

Code | Issues | Pull requests | Actions | Projects | Wiki | Security 30 | Insights | Settings

tettsted

Internal

generated from statisticsnorway/ssb-minimal-template

Watch 2 | Fork 0 | Star 0

main | 1 Branch | 2 Tags | Go to file t | Add file | Code

mortewle Merge pull request #14 from statisticsnorway/kjor2024 — 16fbe4b · last week — 95 Commits

merknader	2024	3 weeks ago
src	tettsted 2024 ferdig	last week
tests	json-fil med nye navn og nummer	last year
.cruft.json	merge	last year
.gitattributes	Initial commit	last year
.gitignore	nesten ferdig, hvis alt er riktig	3 weeks ago
.pre-commit-config.yaml	pre-commit	last year
LICENSE	merge	last year
README.md	Update README.md	last year
SECURITY.md	Initial commit	last year
SSB-PROJECT-INSTRUKSJONER.md	merge	last year
_BACKUP_4897.gitignore	merge	last year
_BASE_4897.gitignore	merge	last year
_LOCAL_4897.gitignore	merge	last year

About

No description, website, or topics provided.

Readme
View license
Security policy
Activity
Custom properties
0 stars
2 watching
0 forks

Releases 1

Endelig versjon 2023 Latest
on Aug 23, 2023

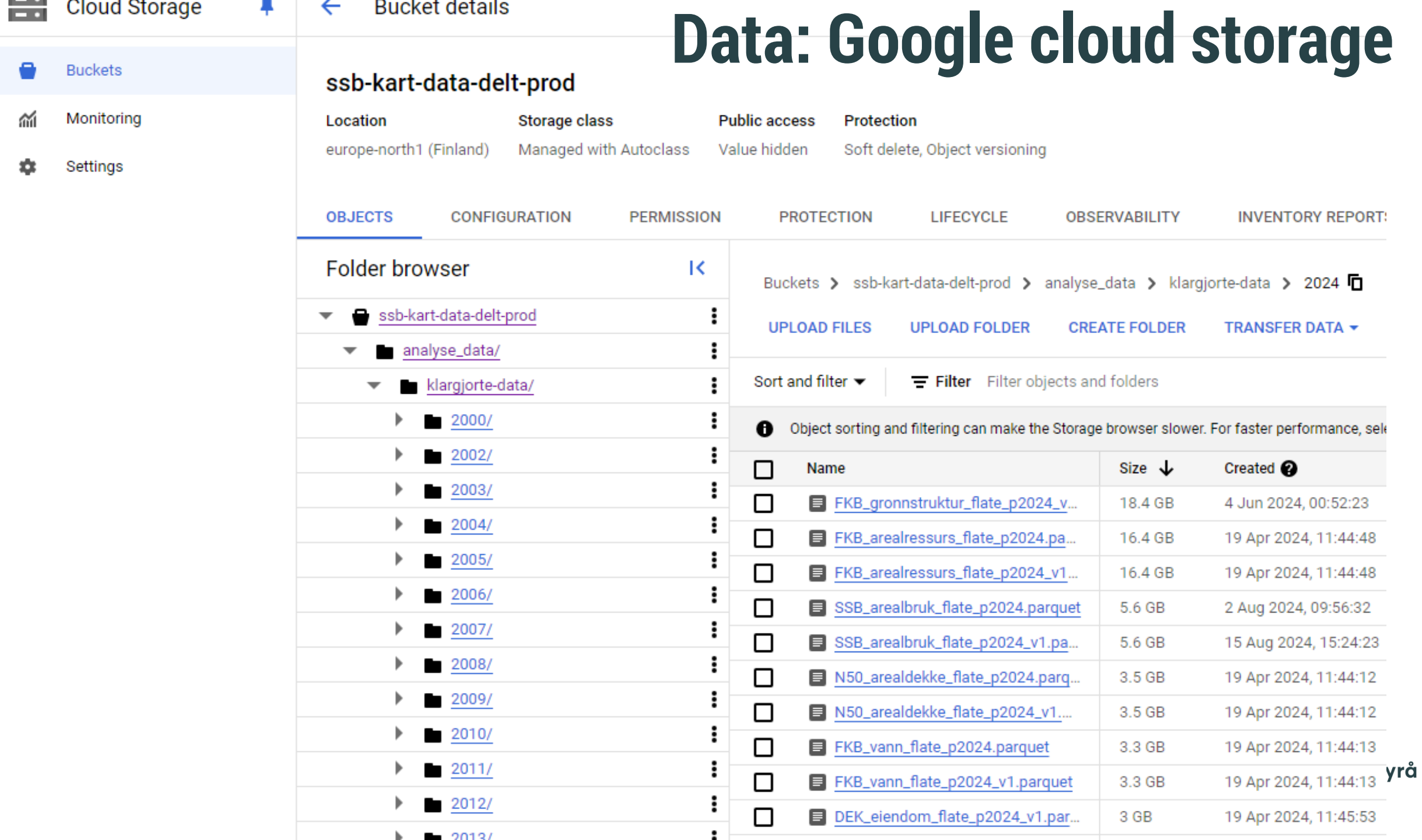
Packages

No packages published
Publish your first package

Languages

Python 76.9% Jupyter Notebook 23.1%

Data: Google cloud storage



Dapla team

- Each group has their own team, which initially only has access to their own data.
- Then the team can apply to get access to data from other parts of the organisation
- For security reasons

Data format: (geo)parquet

- Ridicilously fast to read/write
- Takes up a lot less space
- Open source

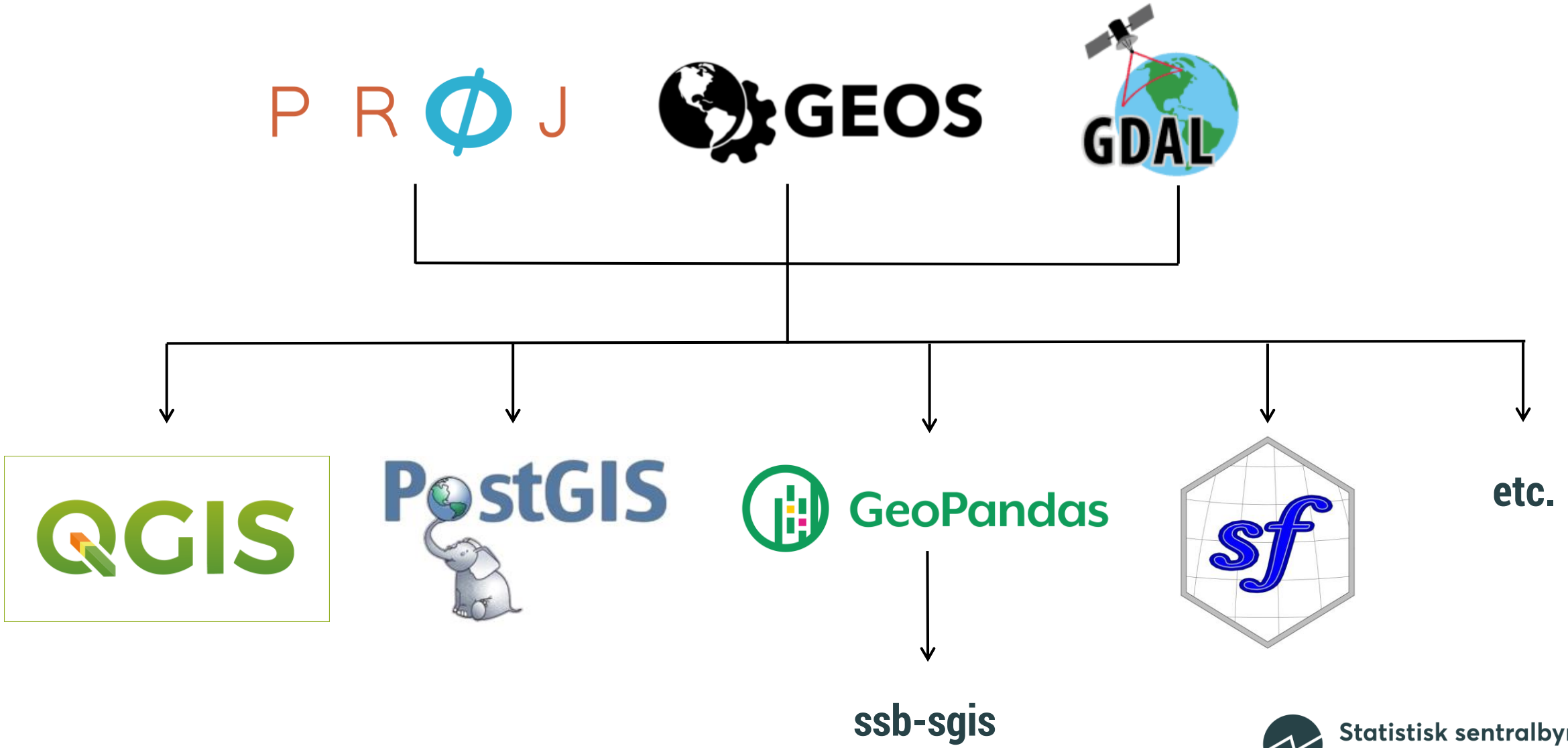
Why more code based?

- Reproductible: Makes a production more automated, less chance for human errors
- Transparent: people can check your code for errors
  - Without having software like ArcGIS
  - The plan is that all code will be publicly available (for journalists etc. to dig)

GIS in the cloud

- Currently, we use:
  - Geopandas in Python
  - sf in R
- Coming soon: QGIS

Open source GIS



GeoPandas (Python)

- Built on top of Pandas, a widely used package for non-GIS data science
- Features the most basic GIS tools (buffer, dissolve, overlay, spatial join)
- But missing some key elements of the ArcGIS toolboxes

GeoPandas: simple syntax

```python
def excecute_some_gis_stuff(muni_number: str) -> None:
    """Some example code to be used in this presentation only."""
    construction = sg.read_geopandas(f"{TEMP}/inndata/anlegg/{muni_number}.parquet")
    land_use = sg.read_geopandas(f"{TEMP}/inndata/arealbruk/{muni_number}.parquet")
    buildings = sg.read_geopandas(f"{TEMP}/inndata/ssb_bygg/{muni_number}.parquet")

    buildings.geometry = buildings.buffer(25)

    spatial_joined = land_use.sjoin(buildings)

    intersected = spatial_joined.overlay(construction, how="intersection")

    erased = spatial_joined.overlay(construction, how="difference")

    intersected_and_erased = pd.concat([intersected, erased])

    sg.write_geopandas(
        intersected_and_erased, f"{TEMP}/dont_use_this_data/{muni_number}.parquet"
    )

```

byrå

ssb-sgis

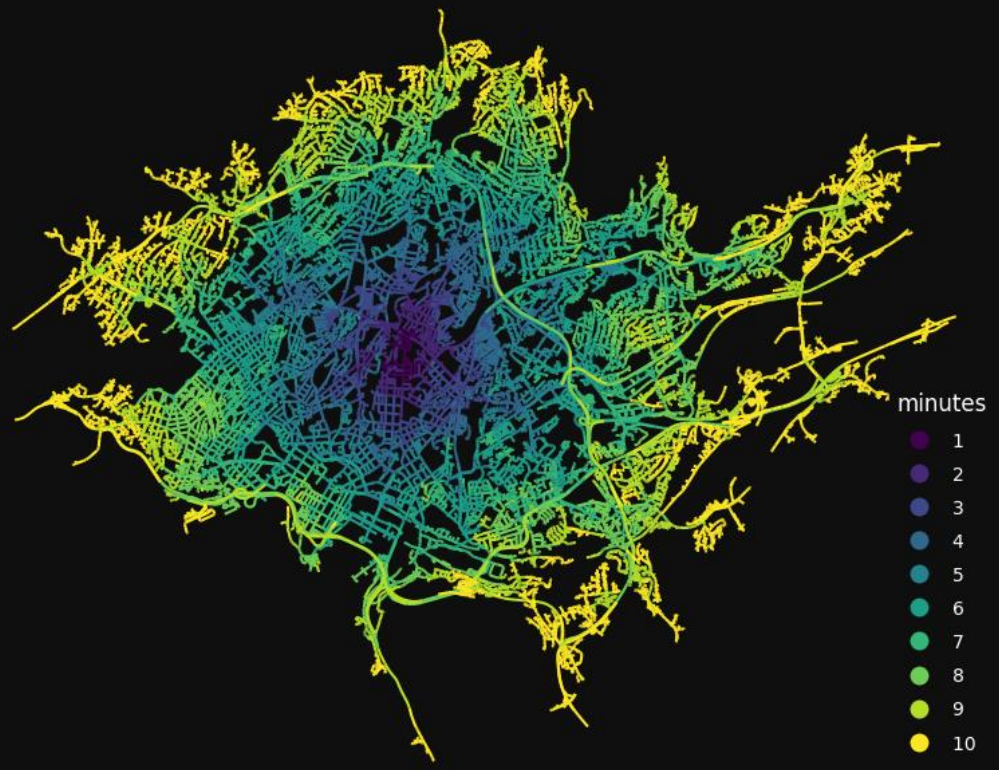
- Python package built on top of geopandas
- Developed by Stats Norway
- Publicly available (https://github.com/statisticsnorway/ssb-sgis)
- Gives the user more GIS functionality when using GeoPandas
  - Map making
  - Network analysis
  - Raster data
  - Eliminating sliver polygons, closing polygon holes etc.

sgis – network analysis

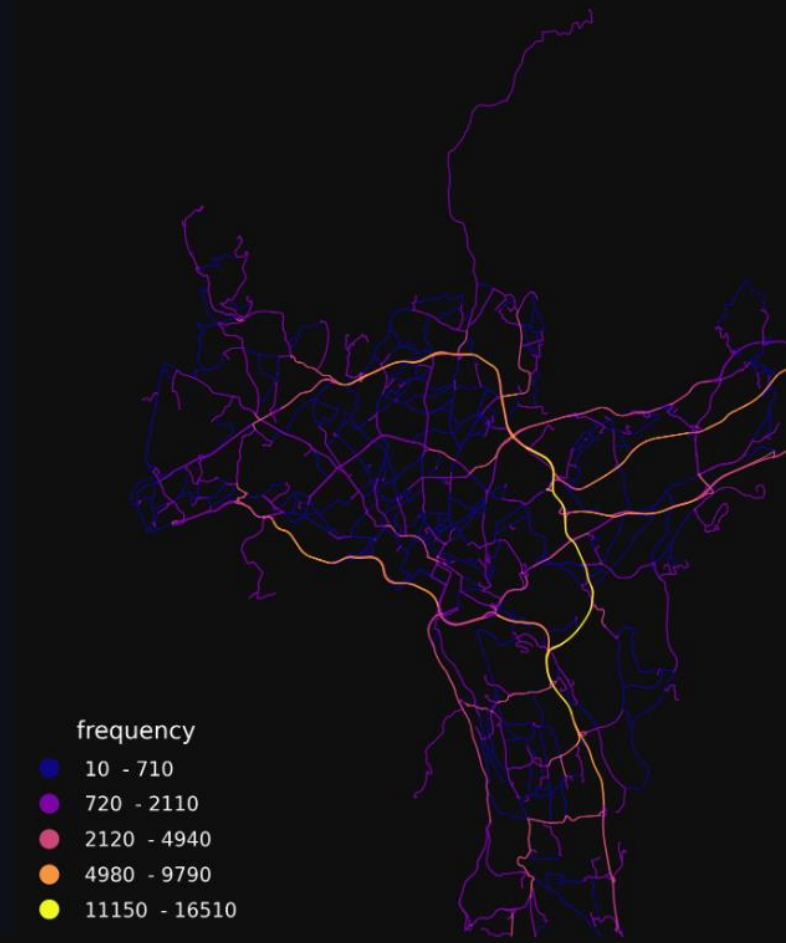
	origin	destination	minutes
0	0	0	0.000000
1	0	1	13.039830
2	0	2	10.902453
3	0	3	8.297021
4	0	4	14.742294
...	...	...	...
999995	999	995	11.038673
999996	999	996	17.820664
999997	999	997	10.288465
999998	999	998	14.798257
999999	999	999	0.000000

[1000000 rows x 3 columns]

Roads that can be reached within 1 to 10 minutes

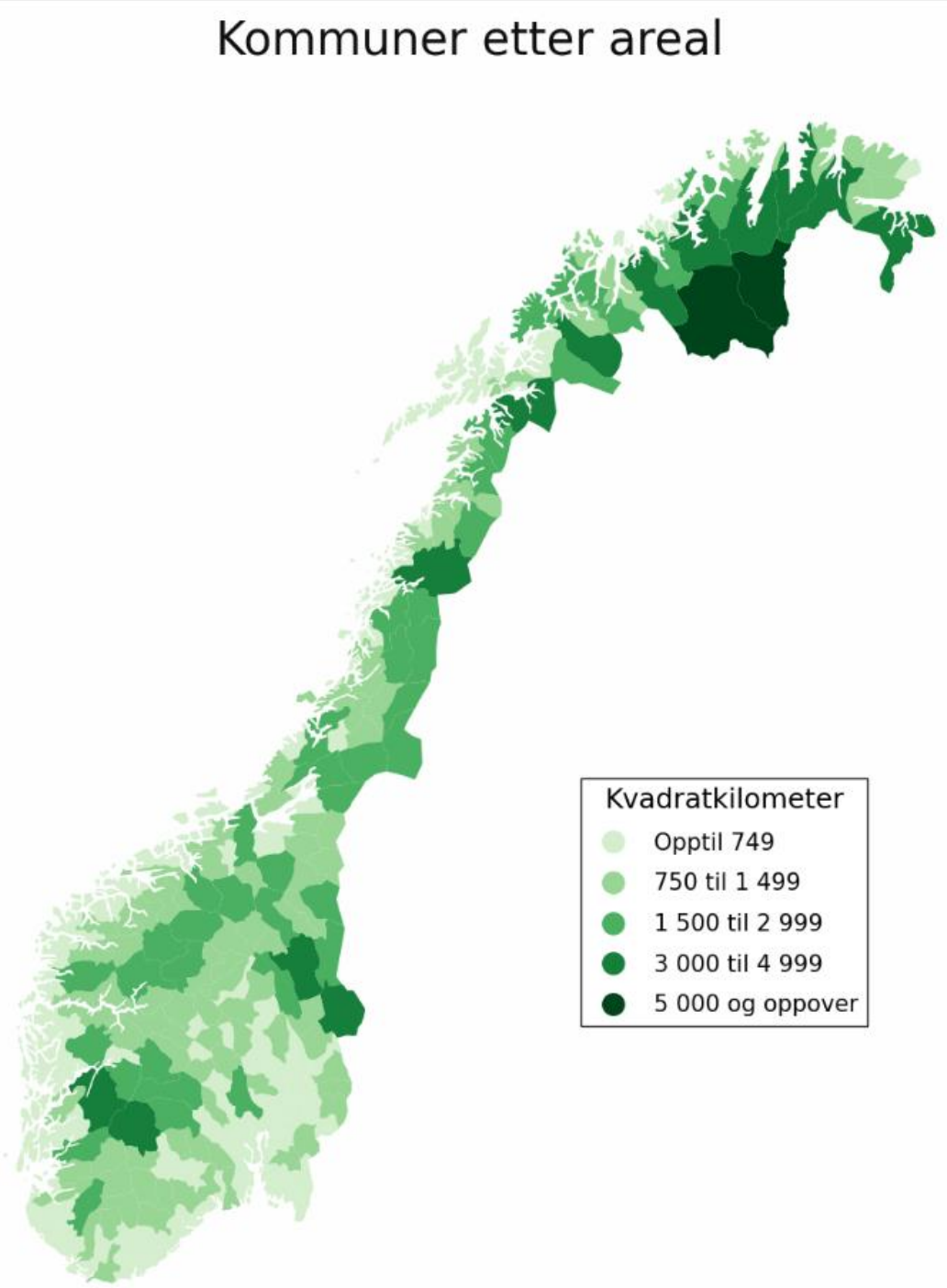


Number of times each road was weighted * 10

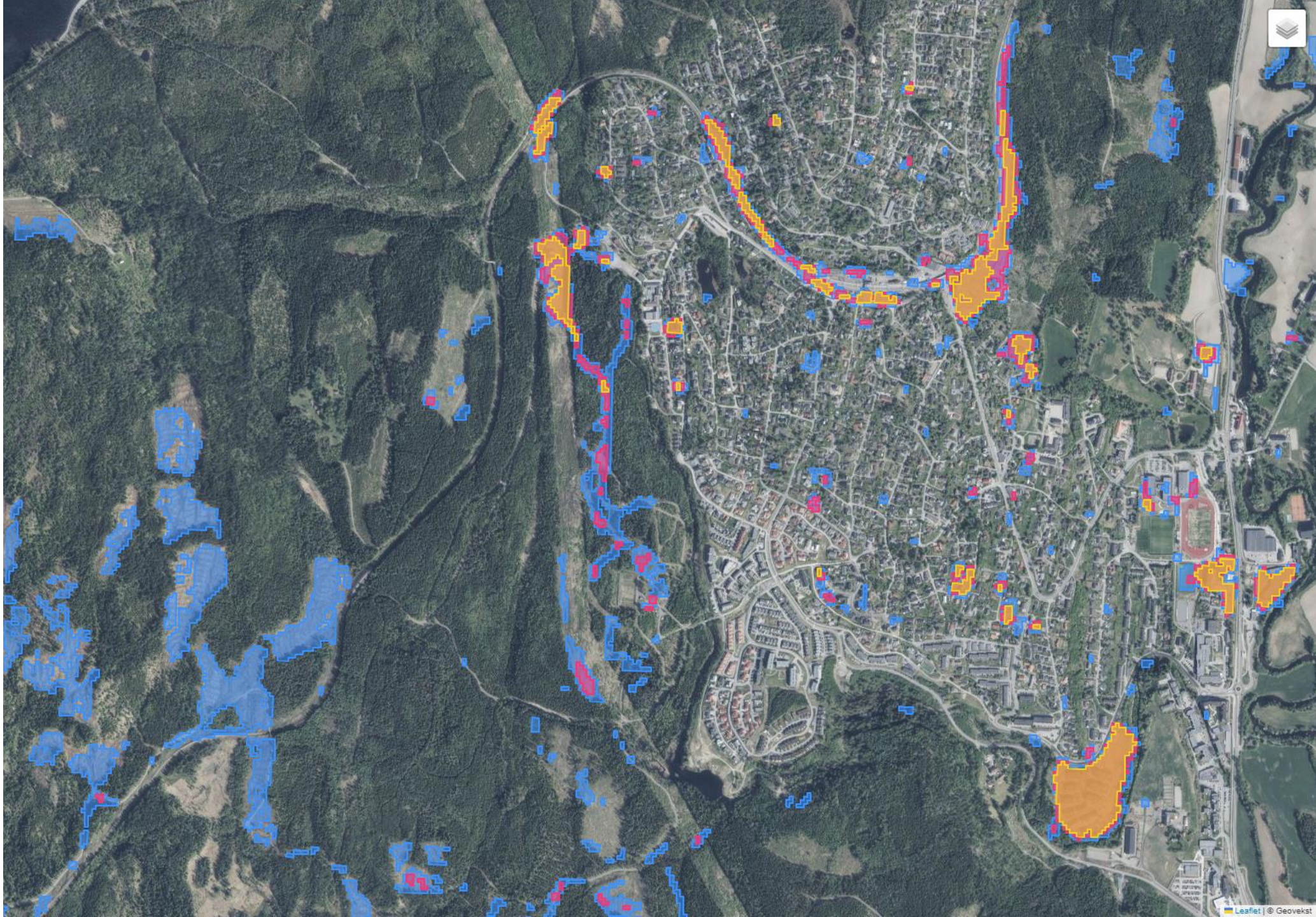


Map making

```
m = sg.ThematicMap(kommuner, column="km2", size=15)
m.title = "Kommuner etter areal"
m.legend.title = "Kvadratkilometer"
m.legend.position = (0.8, 0.35)
m.cmap = "Greens"
m.cmap_start = 50

m.bins = [749, 1499, 2999, 4999]

m.legend.labels = [
    f"Opptil 749",
    "750 til 1 499",
    "1 500 til 2 999",
    "3 000 til 4 999",
    f"5 000 og oppover",
]

m.plot()
```



Raster



Is cloud migration and open source a good idea?

The good – open source

- Cheap/free
  - And we avoid being dependent on the pricing models of large companies (except Google)
- Transparent: gives Stats Norway more control of what's going on under the hoop
- More freedom: it is often easier to combine multiple open source frameworks than combining licence-based software

The good – Python/geopandas

- Fast.
  - Faster algorithms (buffer, dissolve, overlay) than ArcGIS
  - Easy to parallelise – no locked files or «scratch work environments»
  - Productions that took 2-3 weeks in ArcGIS, can take 2-3 hours in the cloud
- Short, easy syntax
- Endless possibilities for customising functionality to your needs

The good – cloud migration

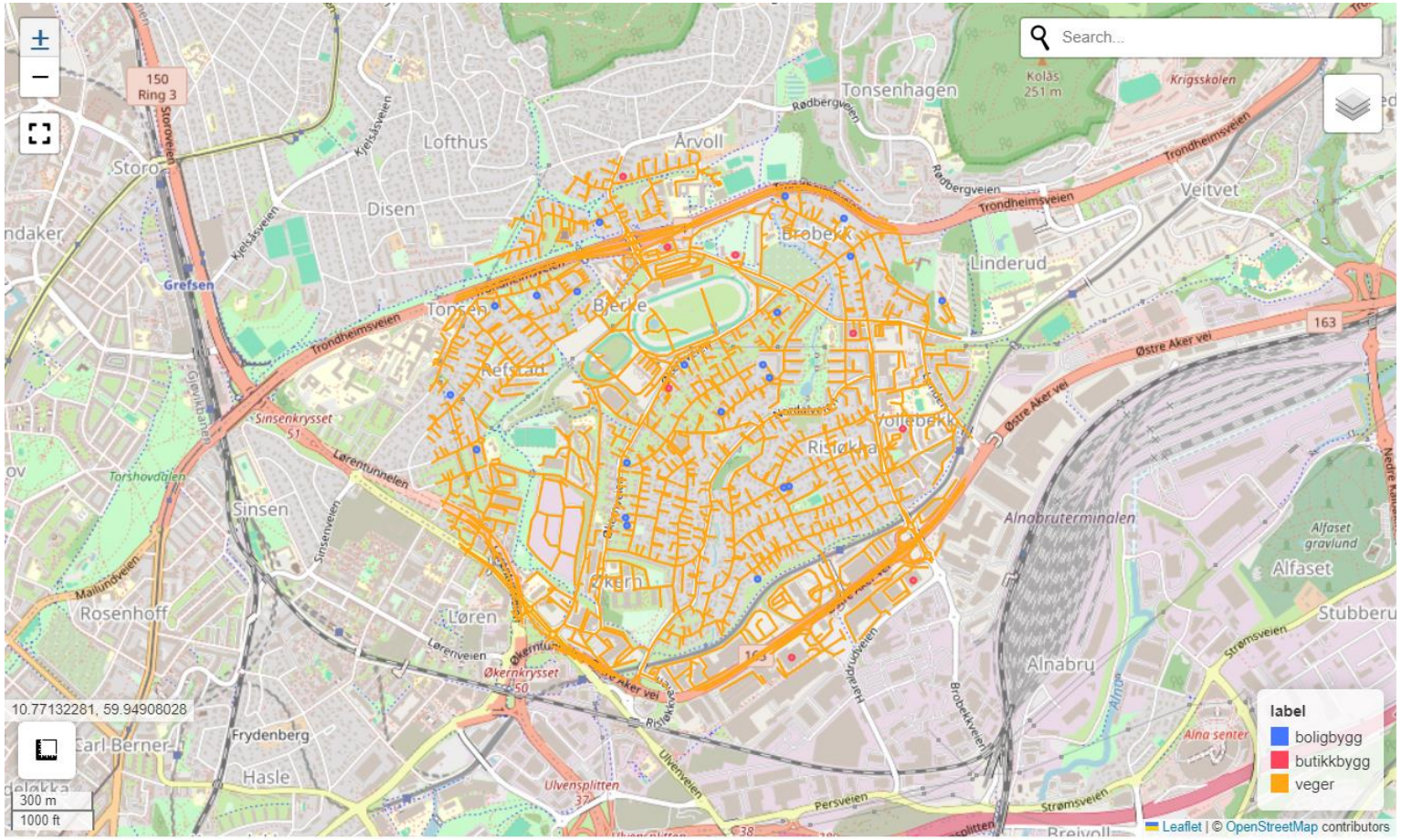
- «Endless» data storage
  - Makes it possible to do jobs we couldn’t before (big data, machine learning)
- No need for maintaining an in-house data storage
- More secure

The bad – cloud migration

- Complex pricing model: difficult to calculate total costs
- The economy of the organisation is at the mercy of Google
- Added complexity for statisticians
  - Dapla teams/data access
  - Data transfer from/to cloud

The bad – geopandas/Python

- Difficult to explore large amounts of geodata



Statistisk sentralbyrå
Statistics Norway

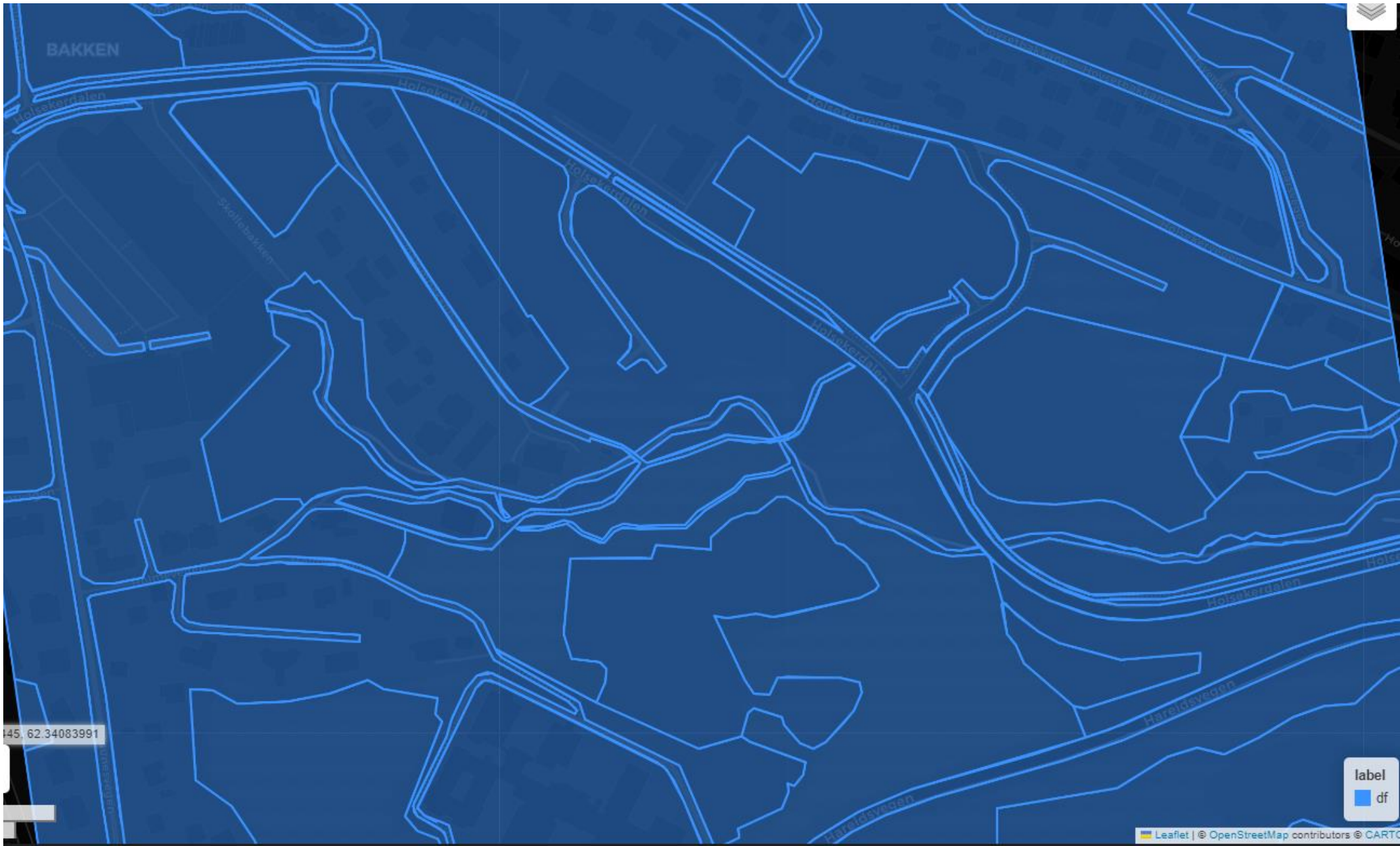
The bad – geopandas/Python

- Difficult to explore large amounts of geodata
- No good editing possibilities
- Difficult for non-programmers
  - At least at first
  - But sometimes always

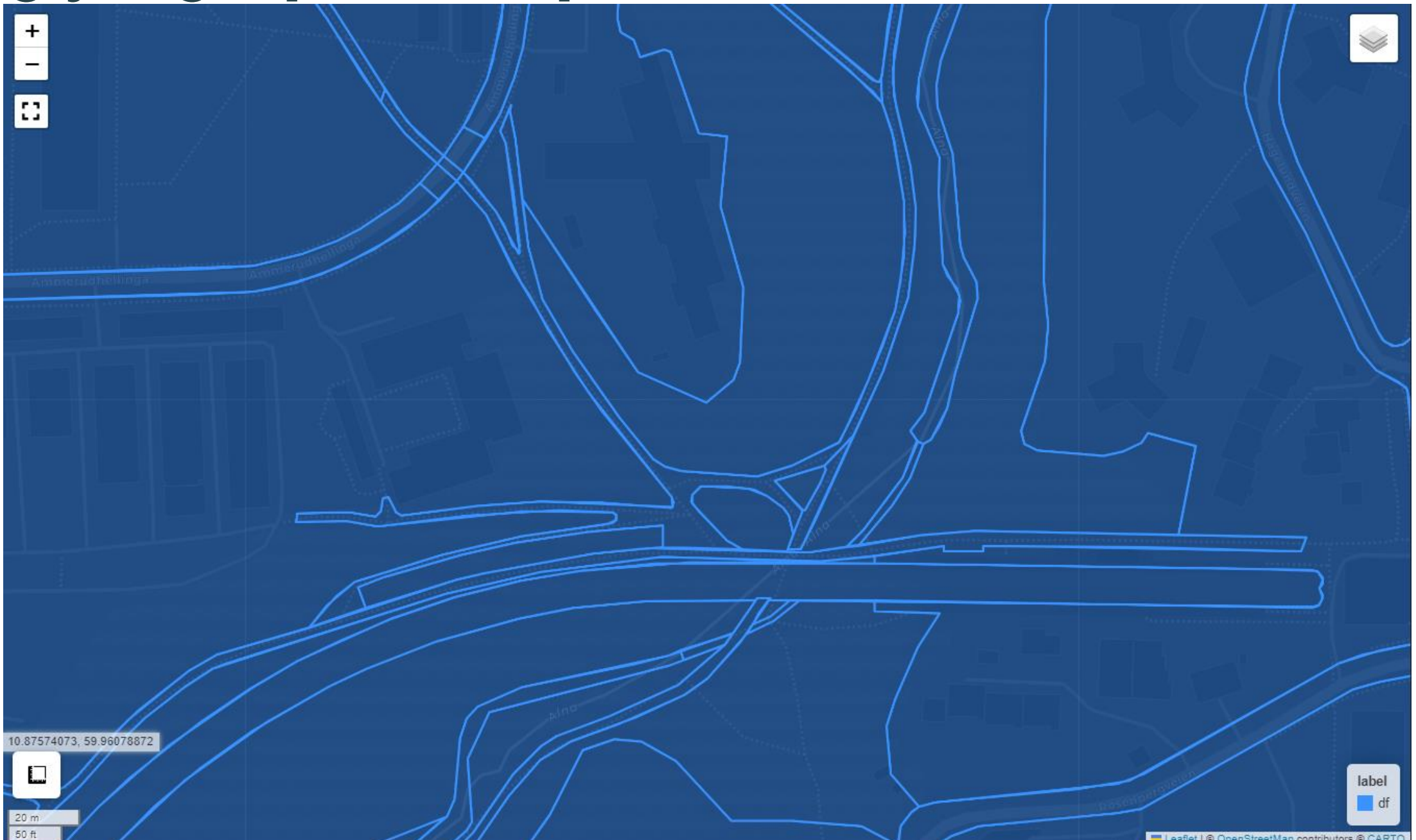
The ugly – geopandas/open source GIS

- Bugs that stop the program:
  - Error messages originating in the underlying C++ source code (GEOS)
  - Happens during complicated overlay and dissolve operations
- And worse: silent bugs without error messages
  - Gives erronous results
- Only applies to one production land use classification

The ugly – geopandas/open source GIS



The ugly – geopandas/open source GIS



Is there a solution?

- I have contacted the developers. They haven’t found time to fix this yet. But hopefully they will eventually.
- Snapping/cleaning function: it should be possible to recreate the magic happening under ArcGIS’s hood.
  - But I haven’t been able to create a perfect algorithm yet.

So should you do cloud and open source?

- If you need more data storage: Yes
- Probably easier to not be a test pilot, but instead learn from our mistakes
- It is possible to just do open source and not cloud
  - Will make a cloud transition easier and cheaper in the future
- Or the opposite: switch to cloud and keep your existing codebase
  - Esri has its own cloud service

TALL SOM FORTELLER

Thank you

Statistisk sentralbyrå
Statistics Norway