Spatial data science with R

ifgi
Institute for Geoinformatics
University of Münster

Edzer Pebesma

Digital Earth Colloquium Series, U Göttingen, Nov 14, 2018

edzer.pebesma@r-project.org
https://r-spatial.org/
@edzerpebesma

overview

- Why am I here?
- What is data science?
- How can we publish reproducibly? (O2R)
- The challenge of large data sets (openEO)

Why am I here?

- I develop and distribute open source geostatistical software since 1997
- I am an active S user since '94, R user since 2000
- I am an active R contributor since 2003:
  - infrastructure packages: sp, spacetime, trajectories, sf, stars
  - method packages: gstat, rgeos, rgdal, ...
- more than 50% of citations to my work cite software contributions
- as editor and scientist, I try to actively engage in (computational) reproducibility

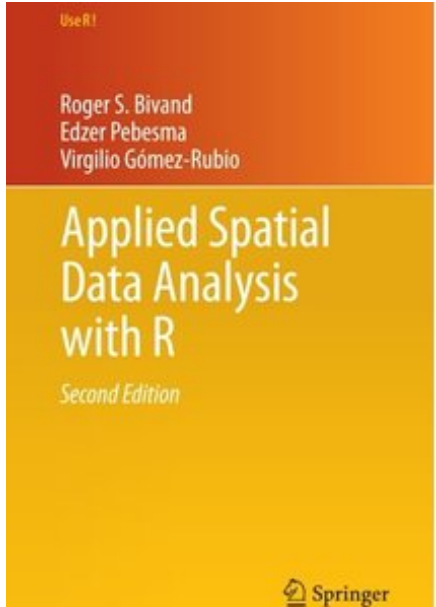
... and I am intrigued by *data science*

Why am I here?

- I develop and distribute open source geostatistical software since 1997
- I am an active S user since '94, R user since 2000
- I am an active R contributor since 2003:
  - infrastructure packages: sp, spacetime, trajectories, sf, stars
  - method packages: gstat, rgeos, rgdal, ...
- more than 50% of citations to my work cite software contributions
- as editor and scientist, I try to actively engage in (computational) reproducibility

... and I am intrigued by *data science*

2008, 2013: https://asdar-book.org/



Harvard Business Review, 2012[1]

[1] "Data scientists' most basic, universal skill is the ability to write code."

What is data science?

- old wine in new bottles?
- a way of statistics (or CS?) departments to popularize their program?
- a recipe to make money out of large amounts of chaotic data?

Whatever it is, data science is

- interdisciplinary
- augmenting applied statistics with computation and software engineering to fight data volumes and data chaos
- popular under young people
- happening on social media
- not searching for self-justification!

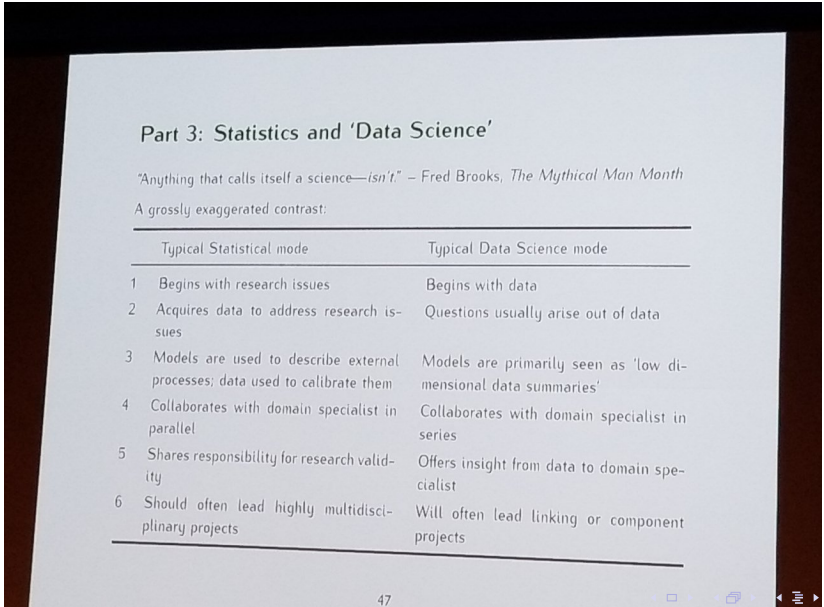
What is data science?

- old wine in new bottles?
- a way of statistics (or CS?) departments to popularize their program?
- a recipe to make money out of large amounts of chaotic data?

Whatever it is, data science is

- interdisciplinary
- augmenting applied statistics with computation and software engineering to fight data volumes and data chaos
- popular under young people
- happening on social media
- not searching for self-justification!

a Bill Venables slide @ UseR! 2018



Part 3: Statistics and 'Data Science'

"Anything that calls itself a science—*isn't*." – Fred Brooks, *The Mythical Man Month*

A grossly exaggerated contrast:

	Typical Statistical mode	Typical Data Science mode
1	Begins with research issues	Begins with data
2	Acquires data to address research issues	Questions usually arise out of data
3	Models are used to describe external processes; data used to calibrate them	Models are primarily seen as 'low dimensional data summaries'
4	Collaborates with domain specialist in parallel	Collaborates with domain specialist in series
5	Shares responsibility for research validity	Offers insight from data to domain specialist
6	Should often lead highly multidisciplinary projects	Will often lead linking or component projects

Why R?

- Because "R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS." (`https://www.r-project.org/`)
- for many: because it gets the job done in a small amount of time.
- for me: because it allows for *reproducible research*
- But so do Python and Julia? Yes, too, and it is good that there is variety.
- It allows for an incredibly fast turnaround of ideas, in an **executable** and **reusable** way
- R is a language that
  - many understand,
  - allows for strong abstractions, and
  - interfaces complex/large systems (db's, cloud-based analytics, object storage, ...)
- The R community has a friendly, "just do it" atmosphere.

What is reproducible research?

In addition to text, tables and figures of the classical paper, publishing reproducibly means that

- we also share data,
- and share the code needed to generate the results from the data
- we do that in an executable form, such
- that reproduction is as easy as downloading a PDF
- ... and with reasonable effort, data and code can be understood in relation to the paper.

How should we do that?[2]

- that does not really matter as long as it is simple and understandable, and
- with a minimum amount of metadata (DOI, author IDs, refs to publication, etc)

[2] do we ask ourselves this when writing a paper??

What is reproducible research?

In addition to text, tables and figures of the classical paper, publishing reproducibly means that

- we also share data,
- and share the code needed to generate the results from the data
- we do that in an executable form, such
- that reproduction is as easy as downloading a PDF
- ... and with reasonable effort, data and code can be understood in relation to the paper.

How should we do that?[2]

- that does not really matter as long as it is simple and understandable, and
- with a minimum amount of metadata (DOI, author IDs, refs to publication, etc)

[2] do we ask ourselves this when writing a paper??

Reproducible research: challenges

- For how long? DFG: >= 10 years
- in 10 years, everything has changed: R, python, Julia interpretiers, operating systems, compilers, packages, ...
- idea:
  - wrap the entire runtime environment in an executable form (VM, docker image)
  - store this with a recipy how to re-run
  - this encapsulates all possible dependencies
  - ... but assumes we can still run VMs or docker images 10 years from now
- DFG project O2R ("Opening Reproducible Research", `http://o2r.info/`) 2017-2018:
  - defined the "executable research compendium" (text, data, code, metadata)
  - built a system that does this,
  - aligned workflows with workflows of the library,
  - investigated re-use of ERC's, and interaction with them
- DFG project O2R2 (2019-20?): bring this in practice with SI's with Copernicus, Elsevier and develop a plug-in in OJS

Results

Please find a project description and information about the team and partners on the about page, and a complete list of publications and presentations on the publications page.

Specifications & documentation

o2r is an open project, so all our components are openly developed on GitHub. The project's findings manifest themselves in the following core specifications and documents, all of which are under development.

- **ERC specification** (source) formally defines the Executable Research Compendium and provides some background.
- **Architecture** (source) describes multiple levels of architecture, from the relation of our reprocibility service with other platforms down to internal microservices.
- **Web API** (source) defines a RESTful API for our reproducibility service, also used by our platform client.

Implementation & demo

We develop a reference implementation of the mentioned specification as Open Source software on GitHub: **https://github.com/o2r-project**

Try the online demo at https://o2r.uni-muenster.de and if you are a developer find the web API endpoint at `https://o2r.uni-muenster.de/api/v1/` .

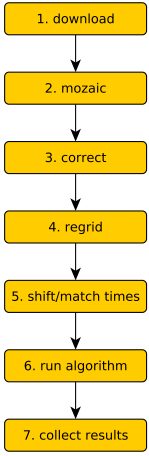
Try it out on your own machine with the reference-implementation (only Docker required!):

```
git clone https://github.com/o2r-project/reference-implementation
```

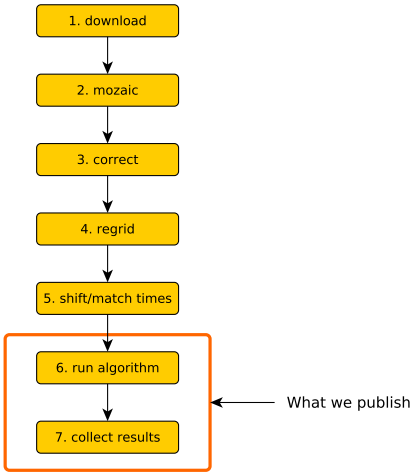
Doing this with large datasets: impossible?

- We can no longer move the data around!
- But we can try to make their organisation and processing more transparent and structured, and better understood.
- The hope is that archives
  - will remain static (only grow),
  - if reprocessed, that this is done versioned, gets documented, and will not have large implications!

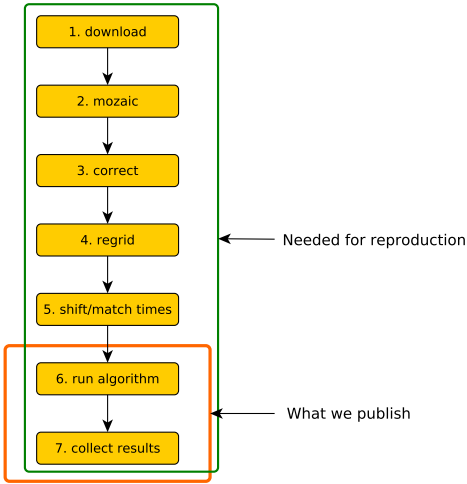
Traditional Earth Observation Research:



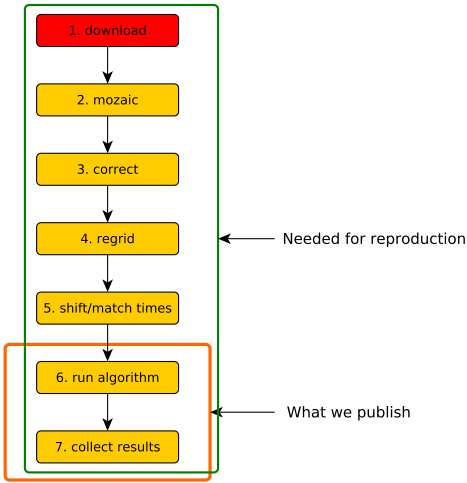
Traditional Earth Observation Research:



Traditional Earth Observation Research:



Traditional Earth Observation Research:



Current problems in Big EO data analysis

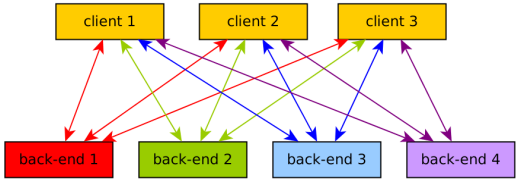
- Google Earth Engine does almost everything, for free, but not transparently
- Other cloud-based platforms are in a *much* less mature state
- All cloud-based platforms use a different interface
- Who will (even attempt to) validate results from platform A against those obtained by platform B?
- With openEO, we can. Easily.

Current problems in Big EO data analysis

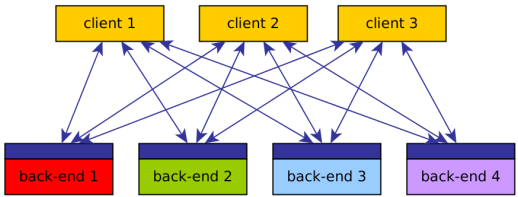
- Google Earth Engine does almost everything, for free, but not transparently
- Other cloud-based platforms are in a *much* less mature state
- All cloud-based platforms use a different interface
- Who will (even attempt to) validate results from platform A against those obtained by platform B?
- With `openEO`, we can. Easily.

openEO

without openEO:

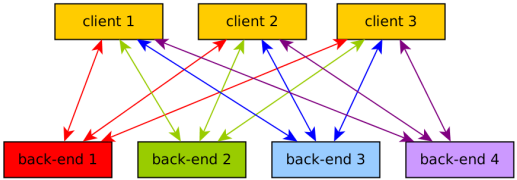


with openEO:

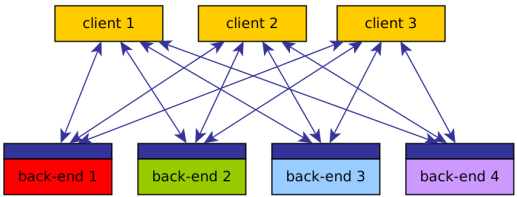


openEO

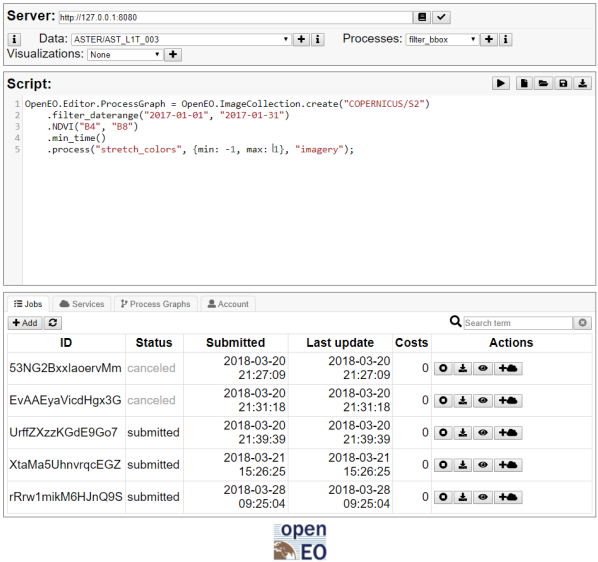
without openEO:

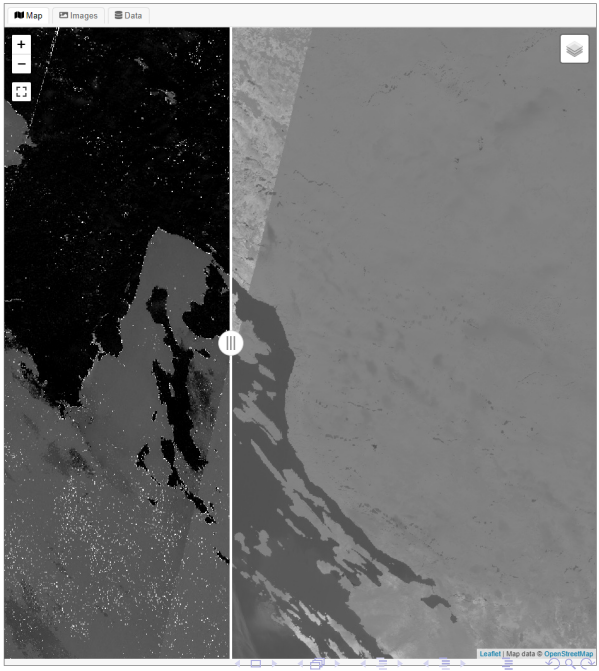


with openEO:



What is an API? Why is it a big thing?


Server: http://127.0.0.1:8080
Data: ASTER/AST_L1T_003
Processes: filter_bbox
Visualizations: None
Script:
1 OpenEO.Editor.ProcessGraph = OpenEO.ImageCollection.create("COPERNICUS/S2")
2 .filter_daterange("2017-01-01", "2017-01-31")
3 .NDVI("B4", "B8")
4 .min_time()
5 .process("stretch_colors", {min: -1, max: 1}, "imagery");
Jobs
Services
Process Graphs
Account
Add
Search term
ID
Status
Submitted
Last update
Costs
Actions
53NG2BxxIaoervMm canceled 2018-03-20 21:27:09 2018-03-20 21:27:09 0
EvAAEyaVicdHgx3G canceled 2018-03-20 21:31:18 2018-03-20 21:31:18 0
UrffZXzzKGdE9Go7 submitted 2018-03-20 21:39:39 2018-03-20 21:39:39 0
XtaMa5UhnvrqcEGZ submitted 2018-03-21 15:26:25 2018-03-21 15:26:25 0
rRrw1mikM6HJnQ9S submitted 2018-03-28 09:25:04 2018-03-28 09:25:04 0
open EO


Map
Images
Data
Leaflet | Map data © OpenStreetMap

```
{ "process_id":"min_time",
  "args":{
    "imagery":{
      "process_id":"/user/custom_ndvi",
      "args":{
        "imagery":{
          "process_id":"filter_daterange",
          "args":{
            "imagery":{
              "process_id":"filter_bbox",
              "args":{
                "imagery":{
                  "product_id":"S2_L2A_T32TPS_20M"
                },
                "left":652000,
                "right":672000,
                "top":5161000,
                "bottom":5181000,
                "srs":"EPSG:32632"
              }
            },
            "from":"2017-01-01",
            "to":"2017-01-31"
          }
        },
        "red":"B04",
        "nir":"B8A"
      }
    }
  }
}
```

Discussion

1. Which criteria do we use when we hire data scientists as faculty? E.g., when
   - candidate A has 20 scientific publications cited more than 20 times (i.e., h = 20)
   - candidate B has 30 R packages on CRAN, of which 10 have been downloaded more than a million times each this year, and has h = 5
2. Why don't we publish reproducibly? Really: it is NOT happening already!!
3. How can we do research sustainably and transparently with Petabyte-scale datasets?
4. How can we develop curricula to make 2 and 3 happen?