POTSDAM INSTITUTE FOR CLIMATE IMPACT RESEARCH

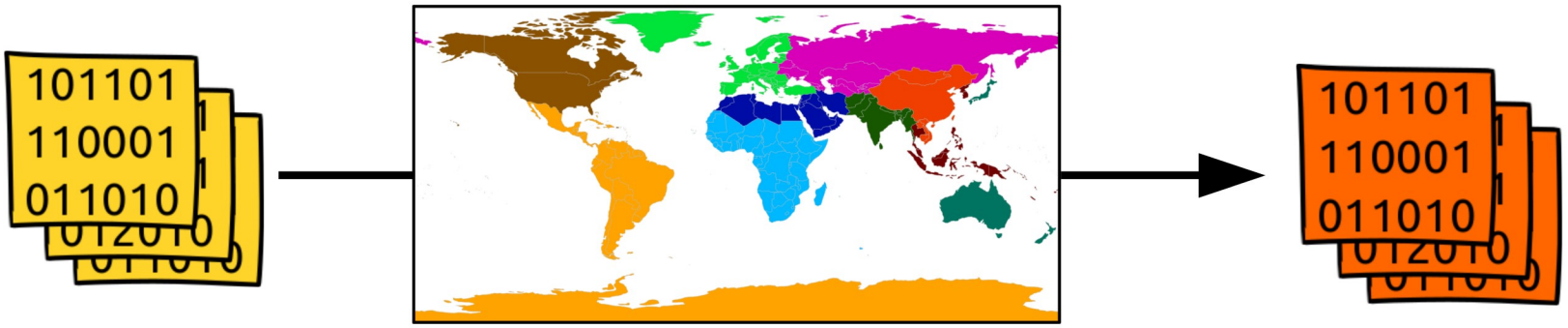
Bringing structure into data processing work-flows in R

Jan Philipp Dietrich, Lavinia Baumstark, Anastasis Giannousakis, Benjamin Leon Bodirsky, David Klein

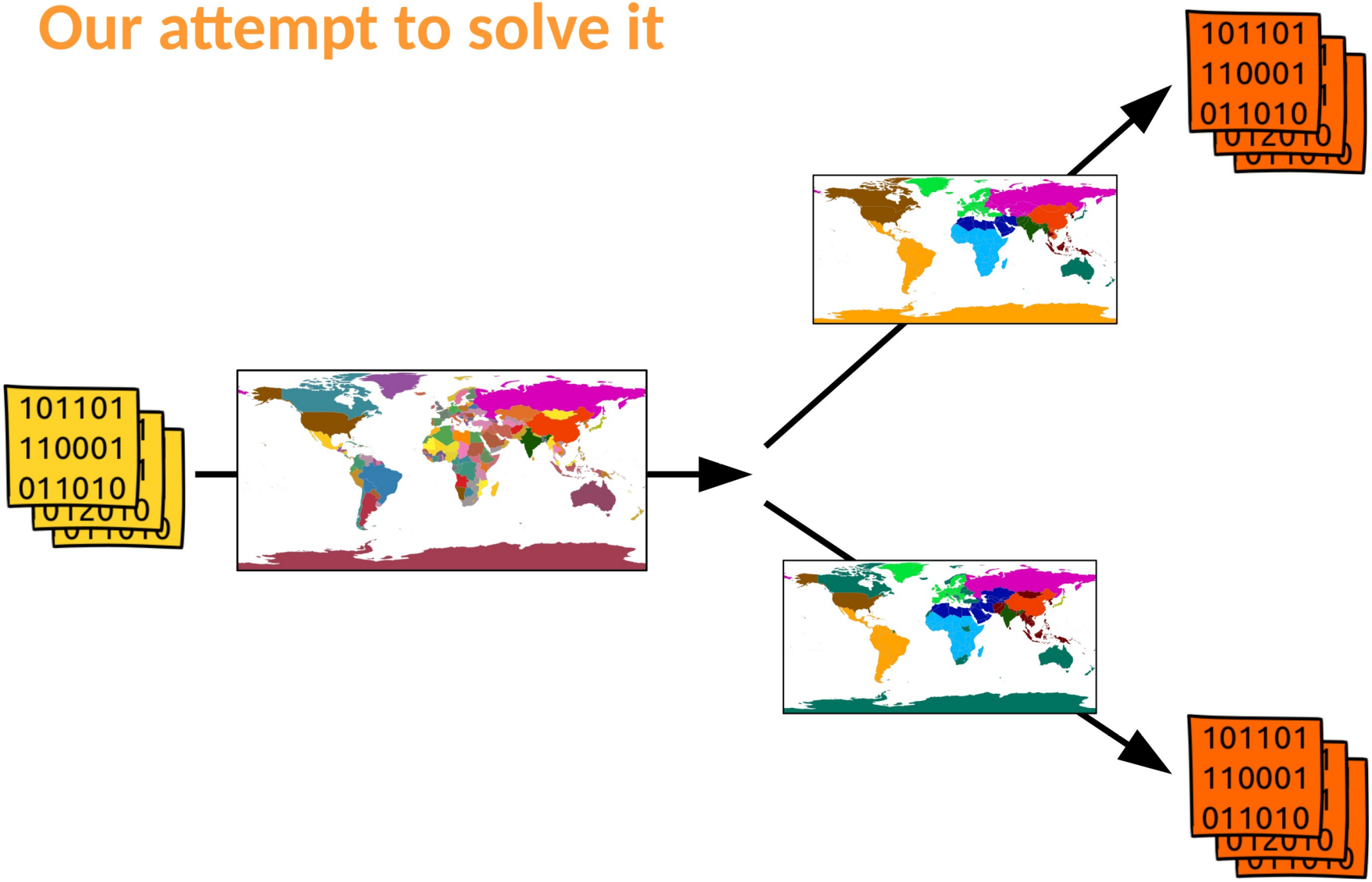
8 September 2017

Leibniz Association

The problem

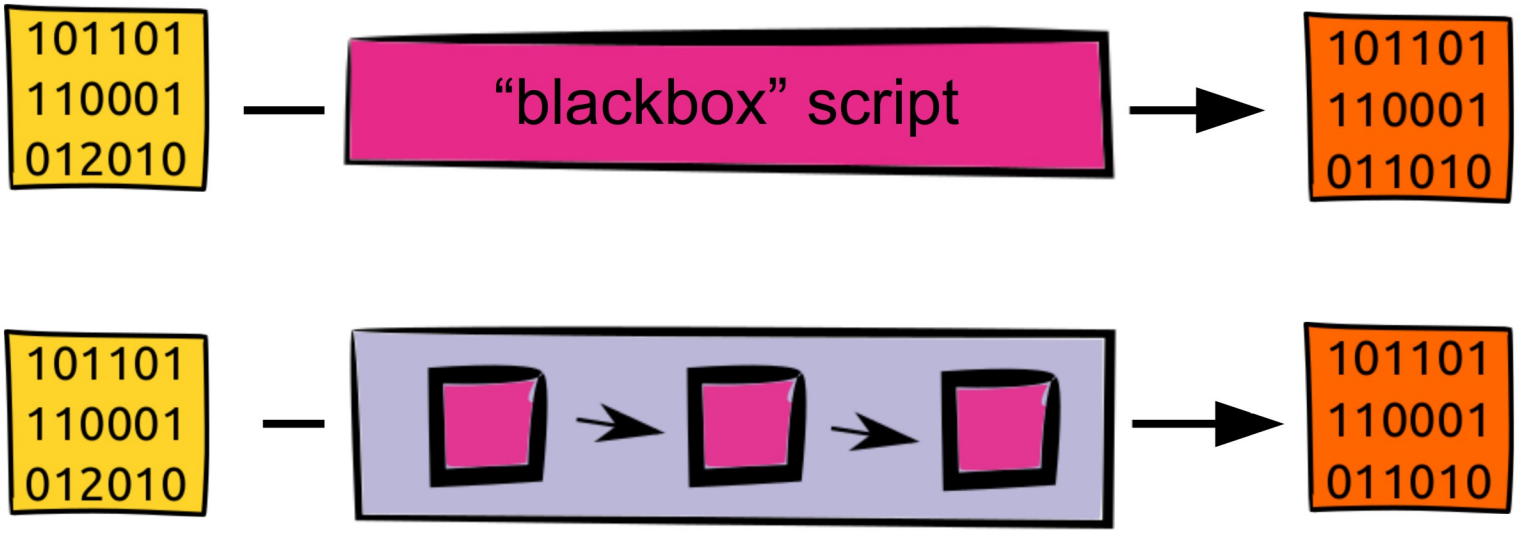


Our attempt to solve it

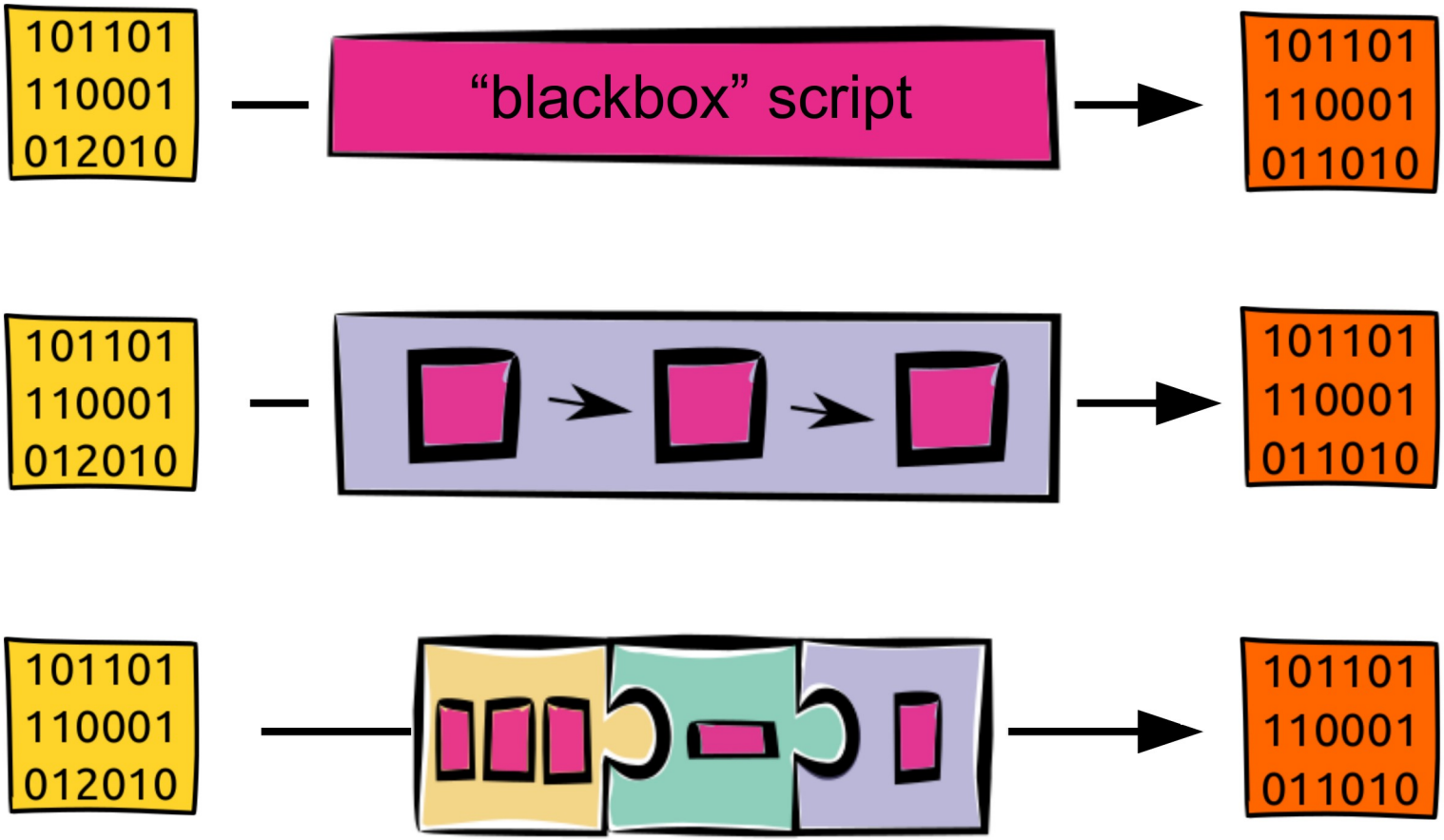


Our attempt to solve it

Our attempt to solve it



Our attempt to solve it



The derived framework

readSource

calcOutput

retrieveData

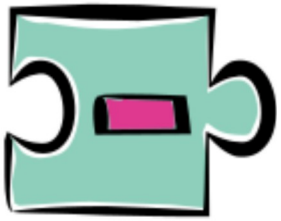
1) Download data
2) Read data and convert to standardized data format
3) Bring data to country-resolution

The derived framework

readSource

calcOutput

retrieveData



- Calculate required data
  - Filtering of data
  - Merging of data different data sources
  - Data harmonization
- Provide spatial aggregation (e.g. weight)

The derived framework

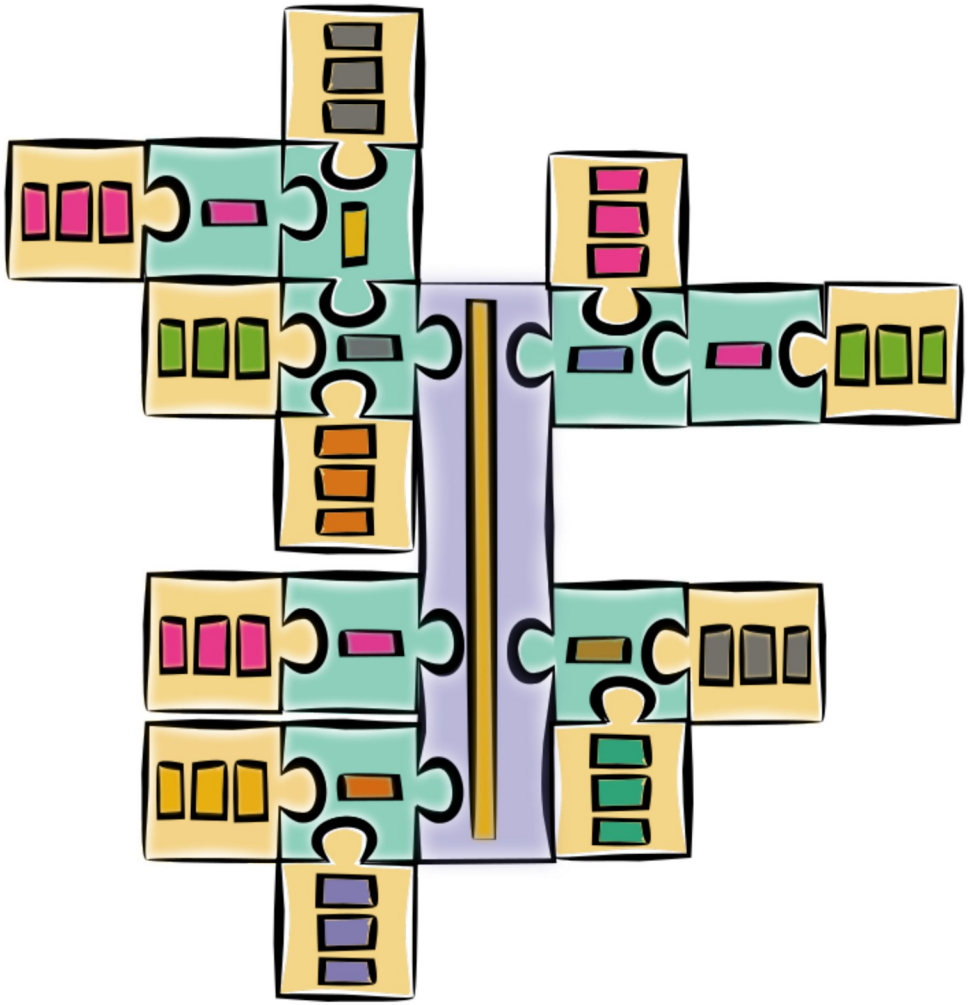
readSource

calcOutput

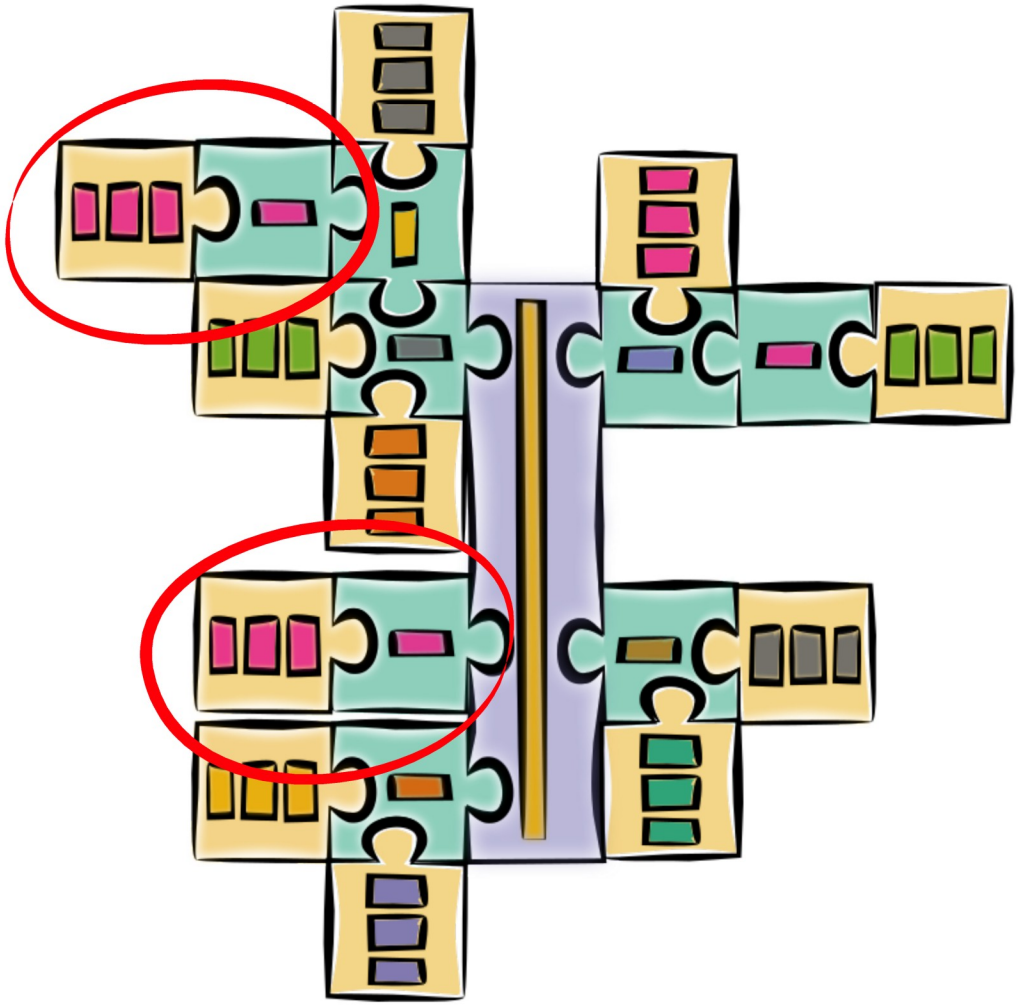
retrieveData

- Collecting data sets
- Coordinate packaging of aggregated data

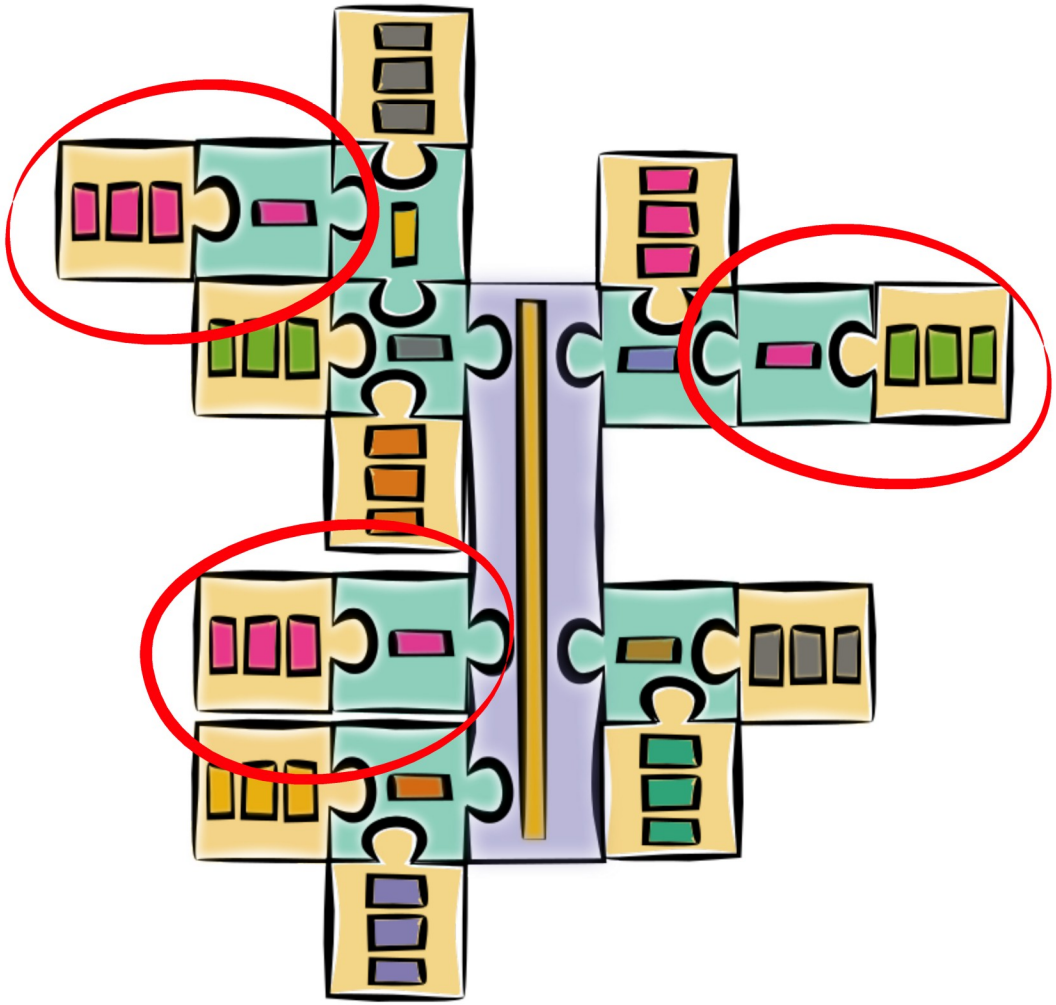
The derived framework



The derived framework

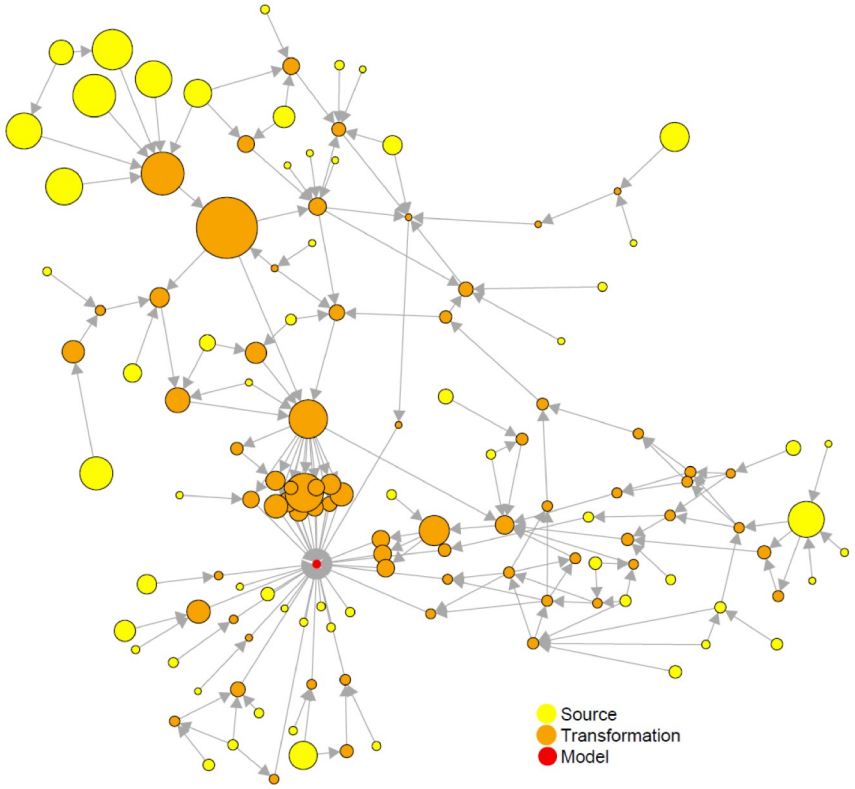


The derived framework



Unanticipated side effects

- A lot of low hanging fruits
  - Meta-data generation
  - Sanity checks
  - Data processing network
  - Data caching
  - Structured log file



- User report faster development
- Broader usage than planned

- Change in focus:

  Spatial aggregation → reproducibility & transparency

Additional Information

- **The R package:**
  - **Name: MADRaT**
    “May All Data be Reproducible and Transparent”
  - **License: BSD2**
  - **Git: https://github.com/pik-piam/madrat**
  - **CRAN: https://CRAN.R-project.org/package=madrat**

- **License of this presentation: CC-BY-4.0**
  https://creativecommons.org/licenses/by/4.0/

- **Contact: dietrich@pik-potsdam.de**

Backup Slides

wrapper functions

```
calcOutput("ours")
```

```
readSource("yours")
```

user functions

```
calcOurs <- function() {
  a <- readSource("yours")
  #do some fancy calculations
  return(list(x=x,weight=weight,unit="-",
        description="Some example calculations"))
}
```

```
readYours() {
  x <- read.csv("example.csv")
  return(as.magpie(x))
}

convertYours(x) {
  y <- toolAggregate(x,"mapping.csv")
  return(y)
}

downloadYours() {
  download.file("http://exam.ple/data.zip"
                , destfile = "data.zip")
  unzip("data.zip")
  unlink("data.zip")
}
```

Backup Slides

wrapper functions

```
retrieveData("example", rev=1.2,
             modelfoler="example",
             regionmapping="example.csv")
```

user functions

```
fullEXAMPLE <- function(rev=0) {
  if(rev>=1) {
    calcOutput("ours", round=2, file="ours.cs4",
    destination="testfolder")
  } else {
   stop("No calculations for rev<1 available!")
  }
}
```

PIK