DYCCEUS

Ateliers du groupe FMR
28th of March, 2018

data.cube: An R Library for the Exploration of Multidimensional Data

Robin Lamarche-Perrin

CNRS / ISC-PIF / LIP6

Mail: Robin.Lamarche-Perrin@lip6.com
Sources: https://github.com/Lamarche-Perrin/data.cube
Webpage: https://www-complexnetworks.lip6.fr/~lamarche/

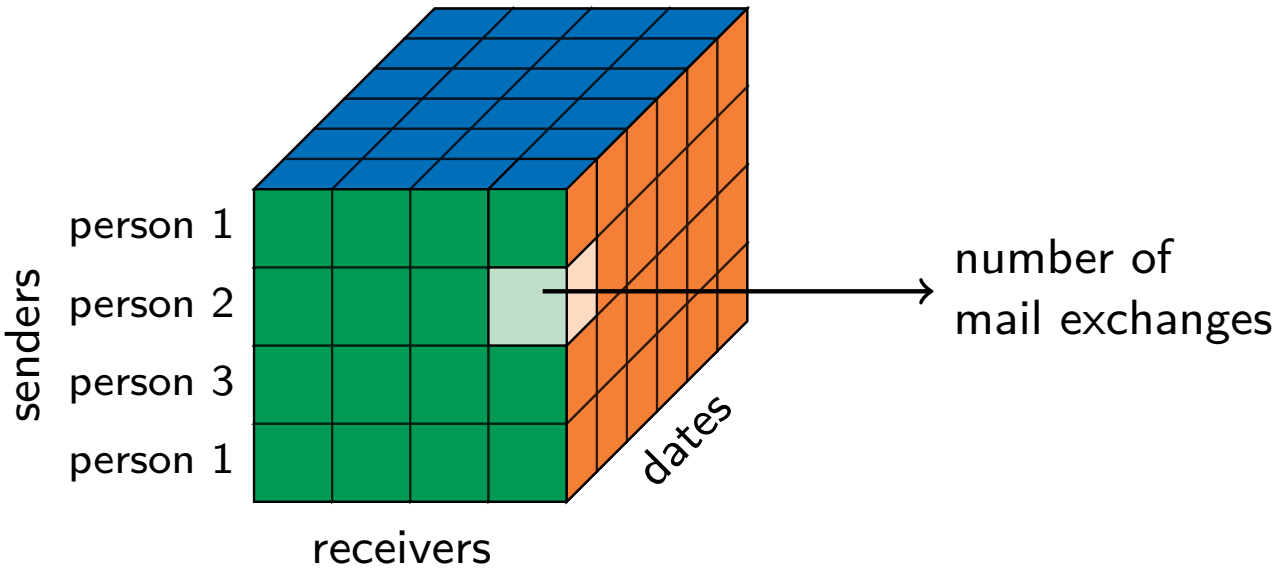
Explore multidimensional data...

...to discover statistical outliers

...to find macroscopic patterns

...to measure diversity, *etc.*

datacube of 3 dimensions
and 1 variable:



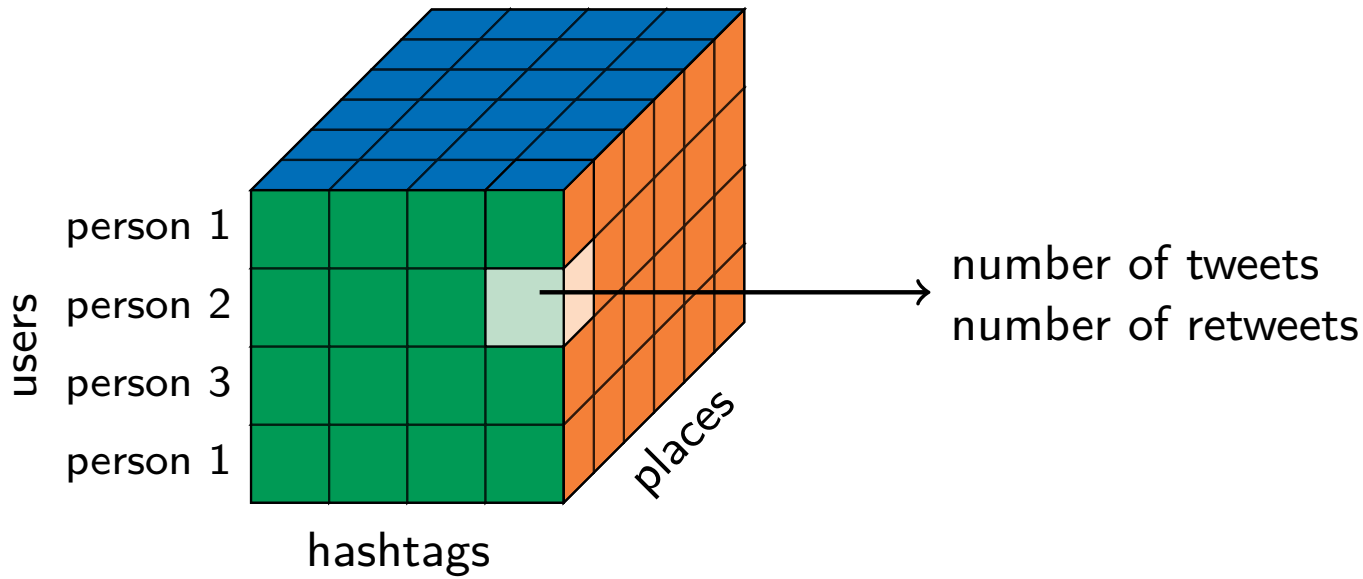
Explore multidimensional data...

...to discover statistical outliers

...to find macroscopic patterns

...to measure diversity, *etc.*

datacube of 3 dimensions
and 2 variables:



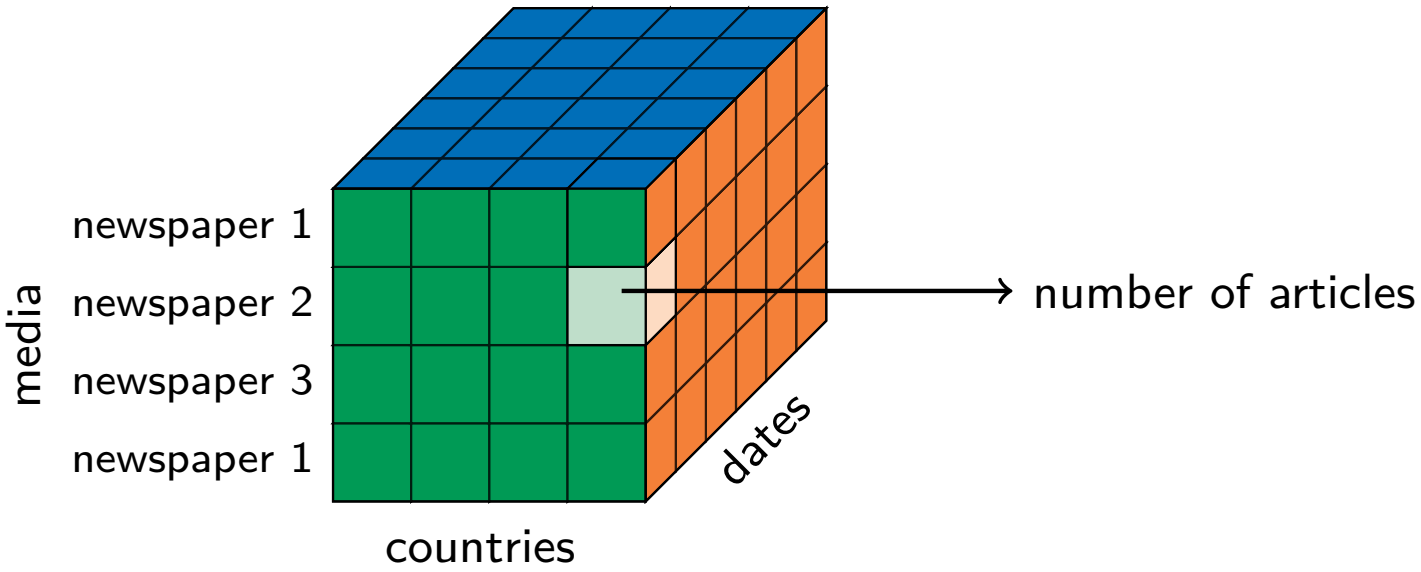
Explore multidimensional data...

...to discover statistical outliers

...to find macroscopic patterns

...to measure diversity, *etc.*

datacube of 3 dimensions and 1 variable:



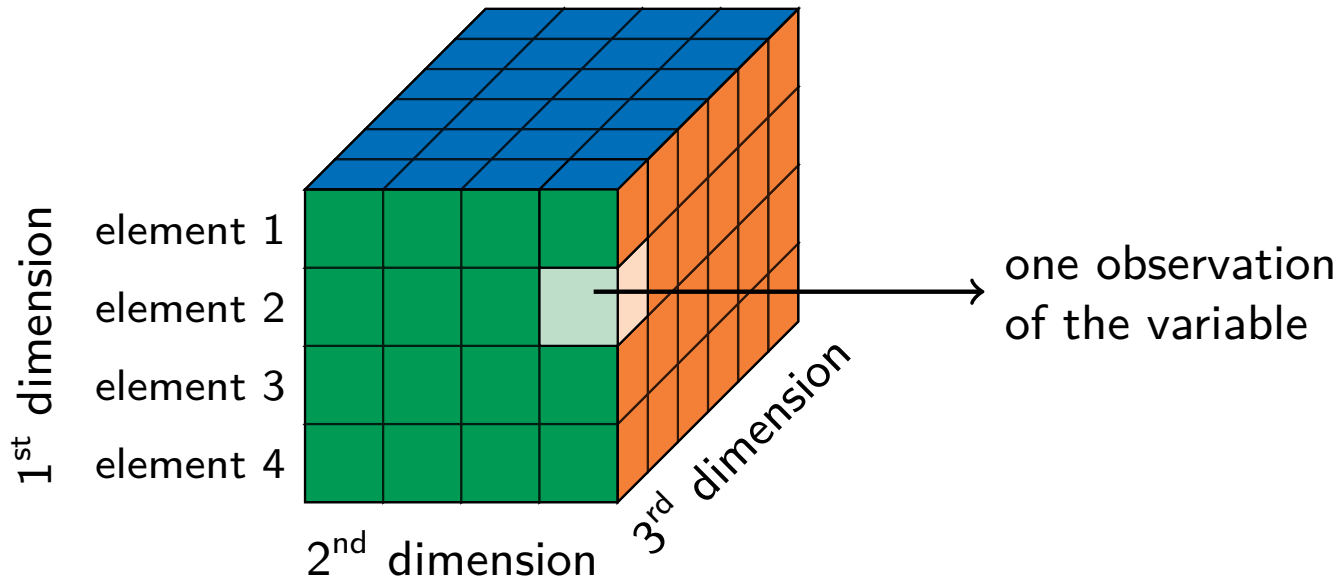
Explore multidimensional data...

...to discover statistical outliers

...to find macroscopic patterns

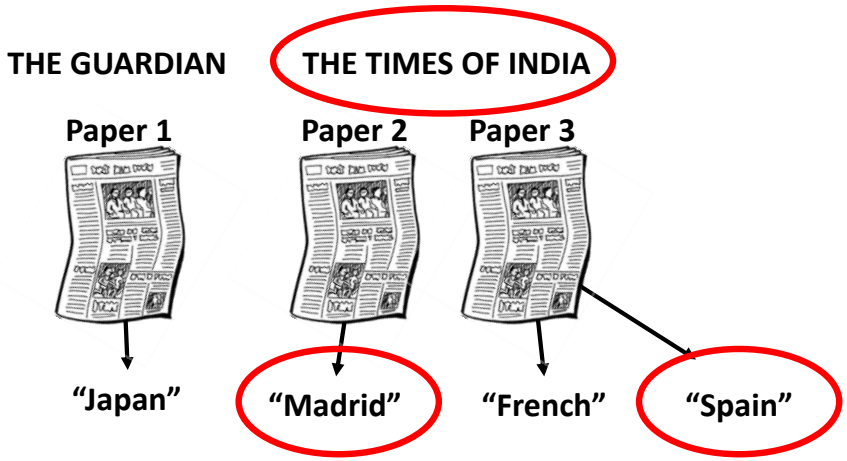
...to measure diversity, *etc.*

datacube of 3 dimensions and 1 variable:

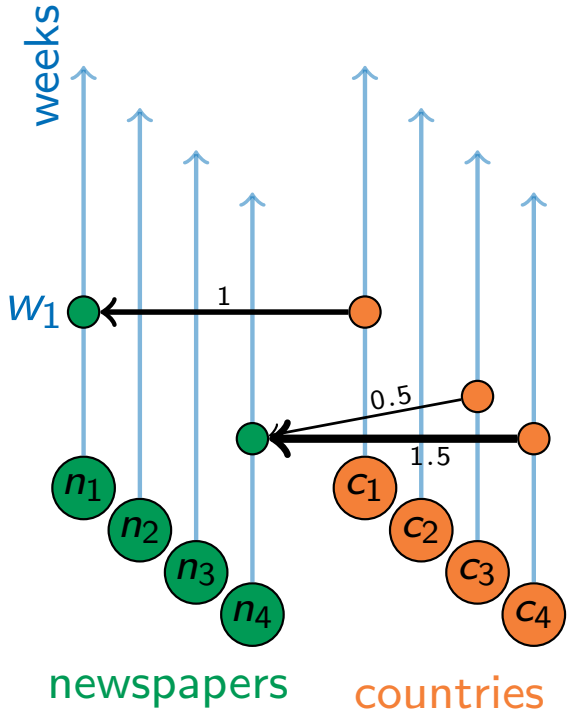


The Geomedia Data

Corpus of newspaper articles



Weighted temporal bipartite graph

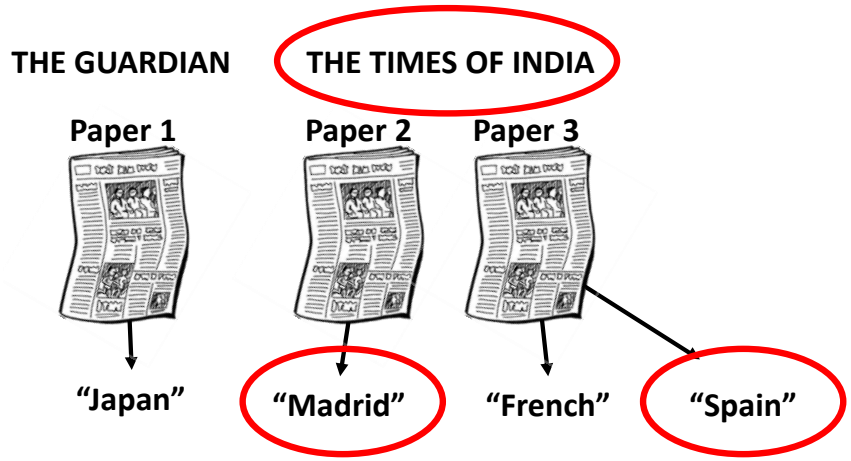


Example:

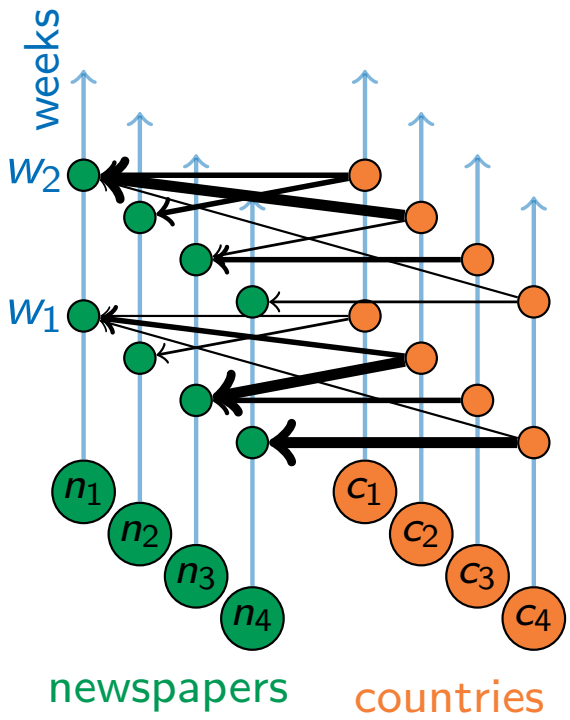
1 article published by media n_1 during week w_1 and citing country c_1
1 article published by media n_4 during week w_1 and citing country c_4
1 article published by media n_4 during week w_1 and citing countries c_3 and c_4

The Geomedia Data

Corpus of newspaper articles



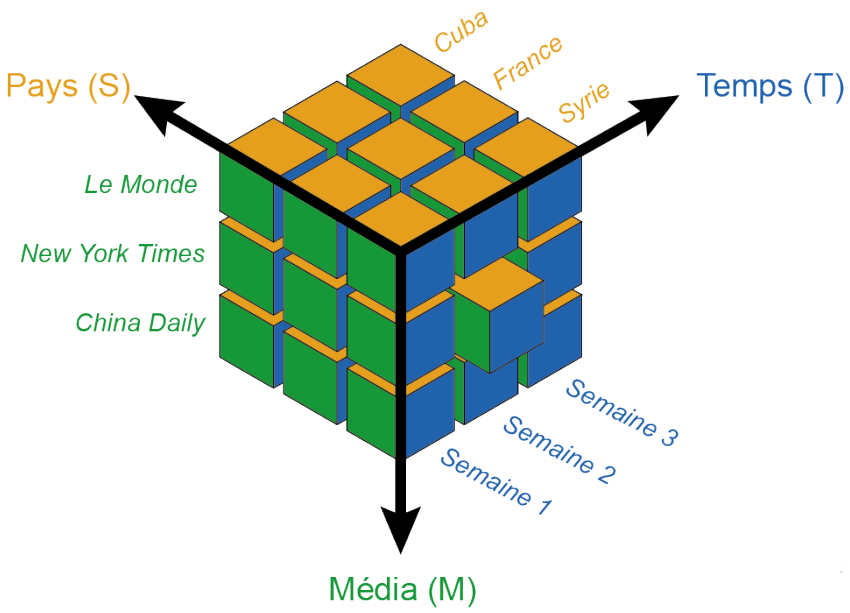
Weighted temporal bipartite graph



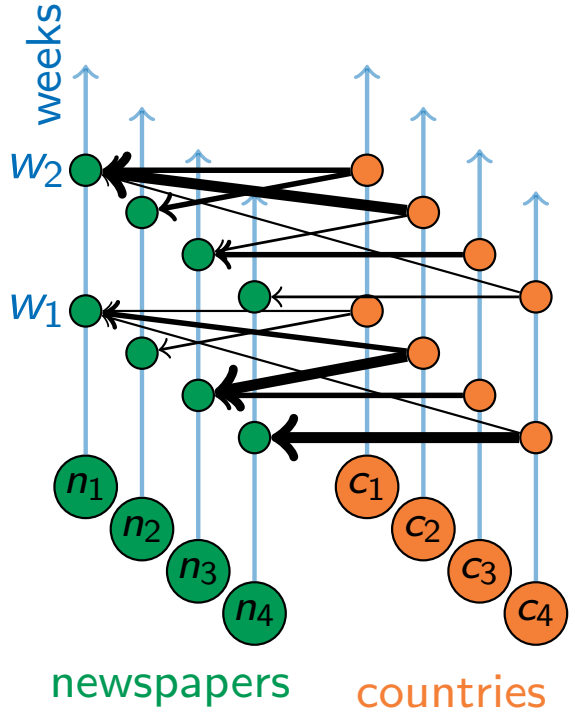
Corpus: 407 364 articles
published by **36 newspapers** (in 23 different states)
during **79 weeks** (from 01/01/2014 to 30/06/2015)
and citing **205 countries** (recognised by the UN)

The Geomedia Data

Geomedia Datacube
(newspapers × countries × weeks)

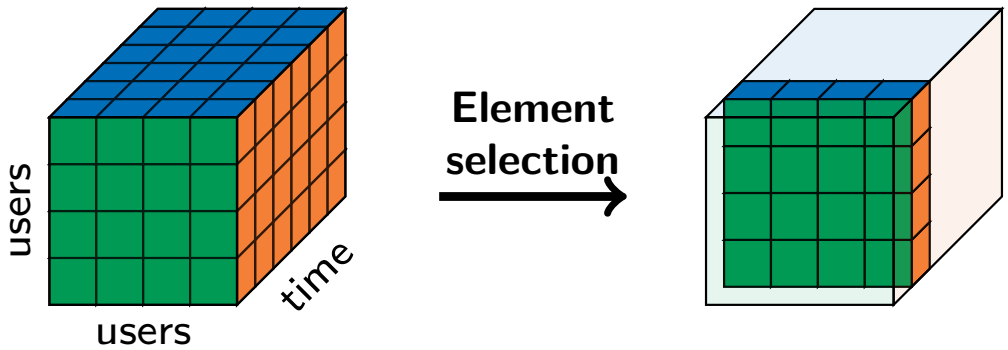


Weighted temporal bipartite graph

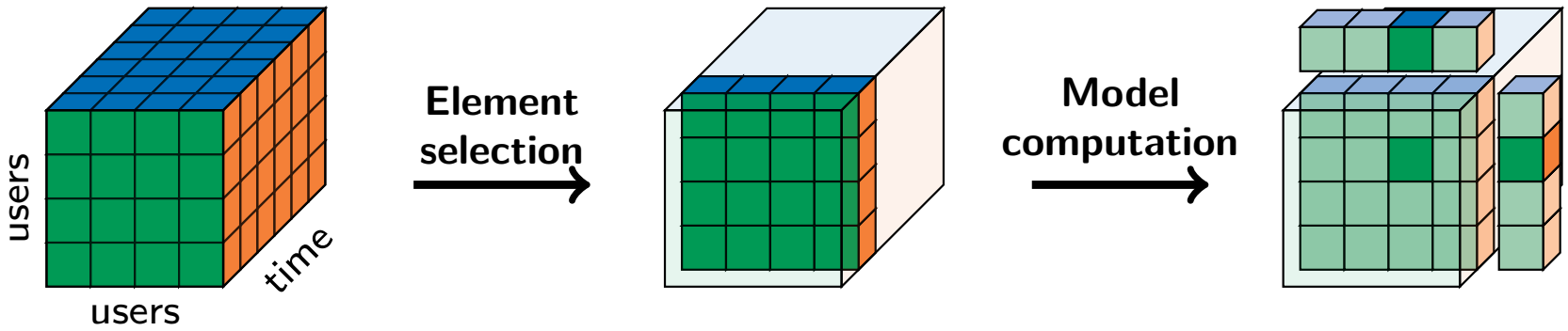


Corpus: 407 364 articles
published by **36 newspapers** (in 23 different states)
during **79 weeks** (from 01/01/2014 to 30/06/2015)
and citing **205 countries** (recognised by the UN)

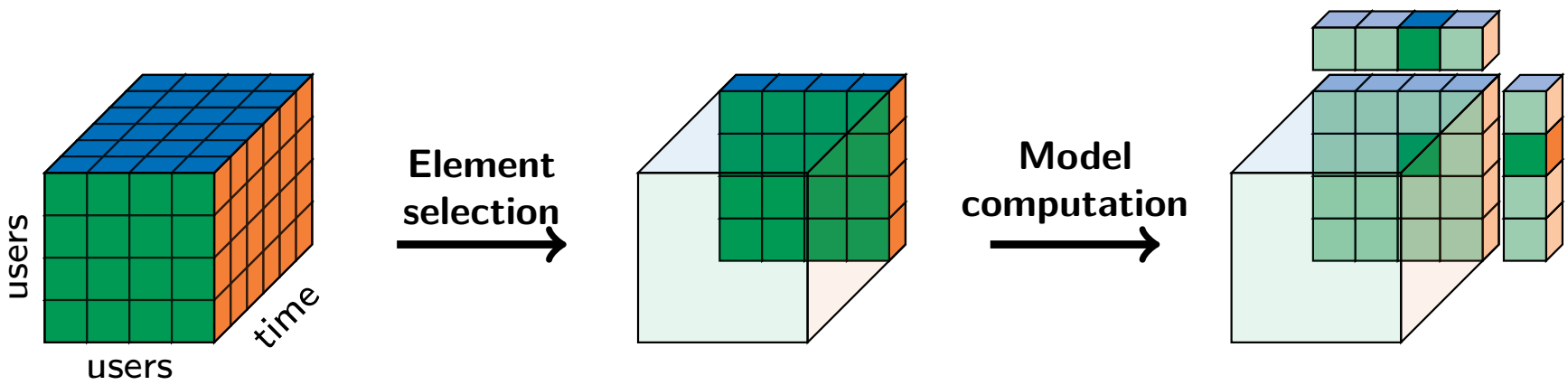
Basic operations on datacubes



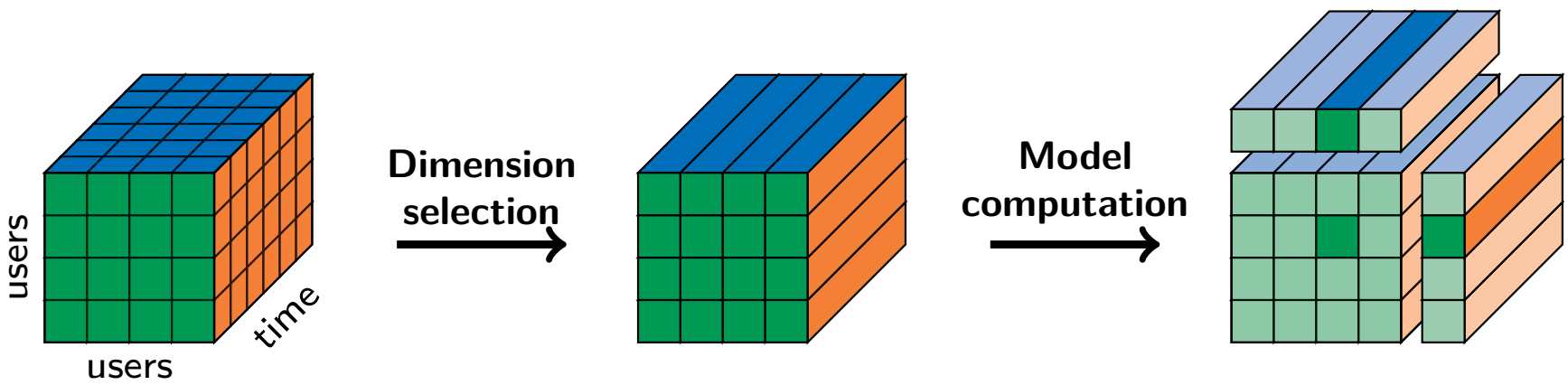
Basic operations on datacubes



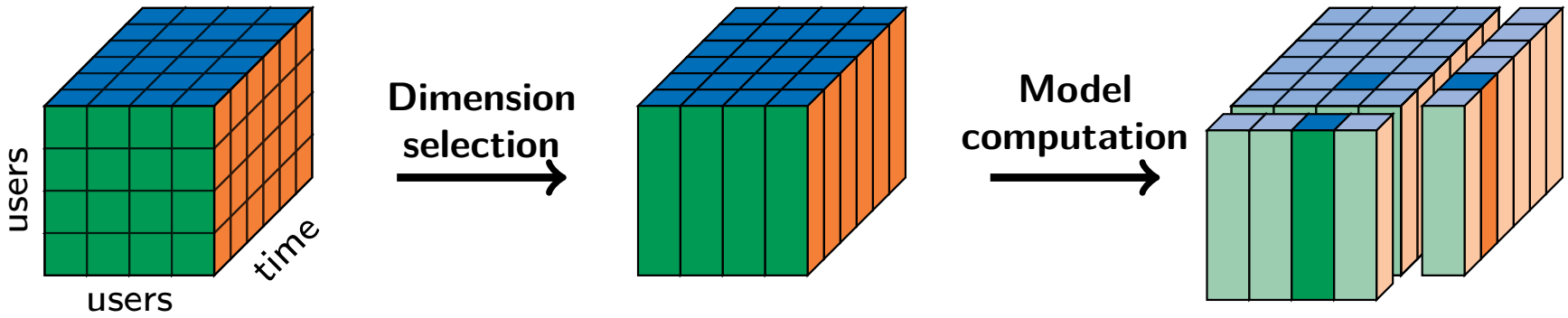
Basic operations on datacubes



Basic operations on datacubes



Basic operations on datacubes



Functions of the library (1/3)

From dataframes to datacubes:

`read.csv (filename)`	→	Load CSV file into dataframe
`as.data.cube (df)`	→	Transform dataframe into datacube
`as.data.frame (dc)`	→	Transform datacube into dataframe

Datacube properties:

`summary (dc)`	→	Get general information
`dim.nb (dc)`	→	Get number of dimensions
`dim.names (dc)`	→	Get names of dimensions
`elm.nb (dc, [dim])`	→	Get number of elements
`elm.names (dc, [dim])`	→	Get names of elements
`src (dc)`	→	Print data structure

Functions of the library (2/3)

Dimension manipulation:

`select.dim (dim1, dim2, ...)`	→	Select dimensions (and aggregate others)

Elements manipulation:

`arrange.elm (dim)`	→	Reorder elements by names
`arrange.elm (dim, var)`	→	Reorder elements by values
`remove.elm (dim, elm.array)`	→	Remove elements in a given dimension
`select.elm (dim, elm.array)`	→	Select elements (and remove others)
`select.elm (dim, top.nb, var)`	→	Select top elements
`select.elm (dim, bot.nb, var)`	→	Select bottom elements
`select.elm (dim, filter)`	→	Filter and keep elements wrt condition

Observations manipulation:

`arrange.obs (var)`	→	Reorder observations by values
`select.obs (filter)`	→	Filter and keep observations wrt condition

Functions of the library (3/3)

Data visualisation:

`plot.obs (var, type="bar")`	→	Plot observations (bar plot)
`plot.obs (var, type="line")`	→	Plot observations (line plot)
`plot.obs (var, sep.dim)`	→	Plot observations (superposed plots)
`biplot.obs (var, x.dim, y.dim)`	→	Plot observations (2D plot)

Find outliers:

`compute.model (dim1, dim2, ..., [deviation.type])`
→ Compute a marginal model and find statistical deviations
→ Give access to new variables: `model`, `ratio`, `deviation`, and `outlier`

`plot.outliers ()` → Plot outliers after having computed a model