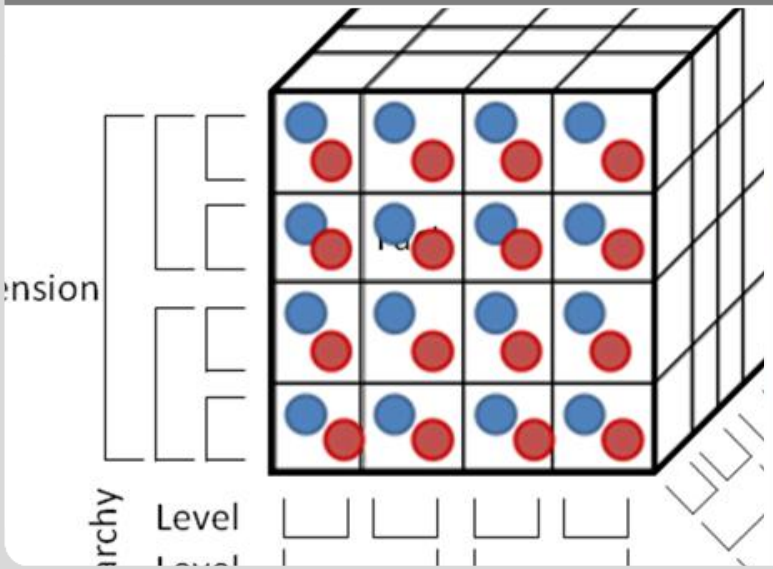
Karlsruhe Institute of Technology

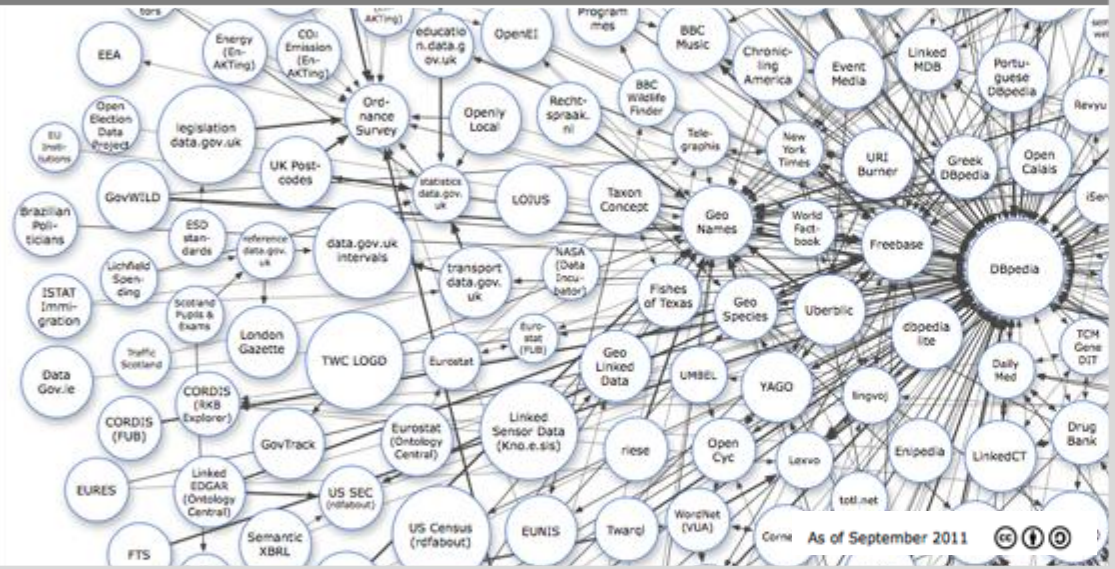
Definition and Materialisation of the Global Cube from Multidimensional Datasets on the Web

Benedikt Kämpgen, Steffen Stadtmüller, Andreas Harth

MUGS

Institute of Applied Informatics and Formal Description Methods (AIFB)

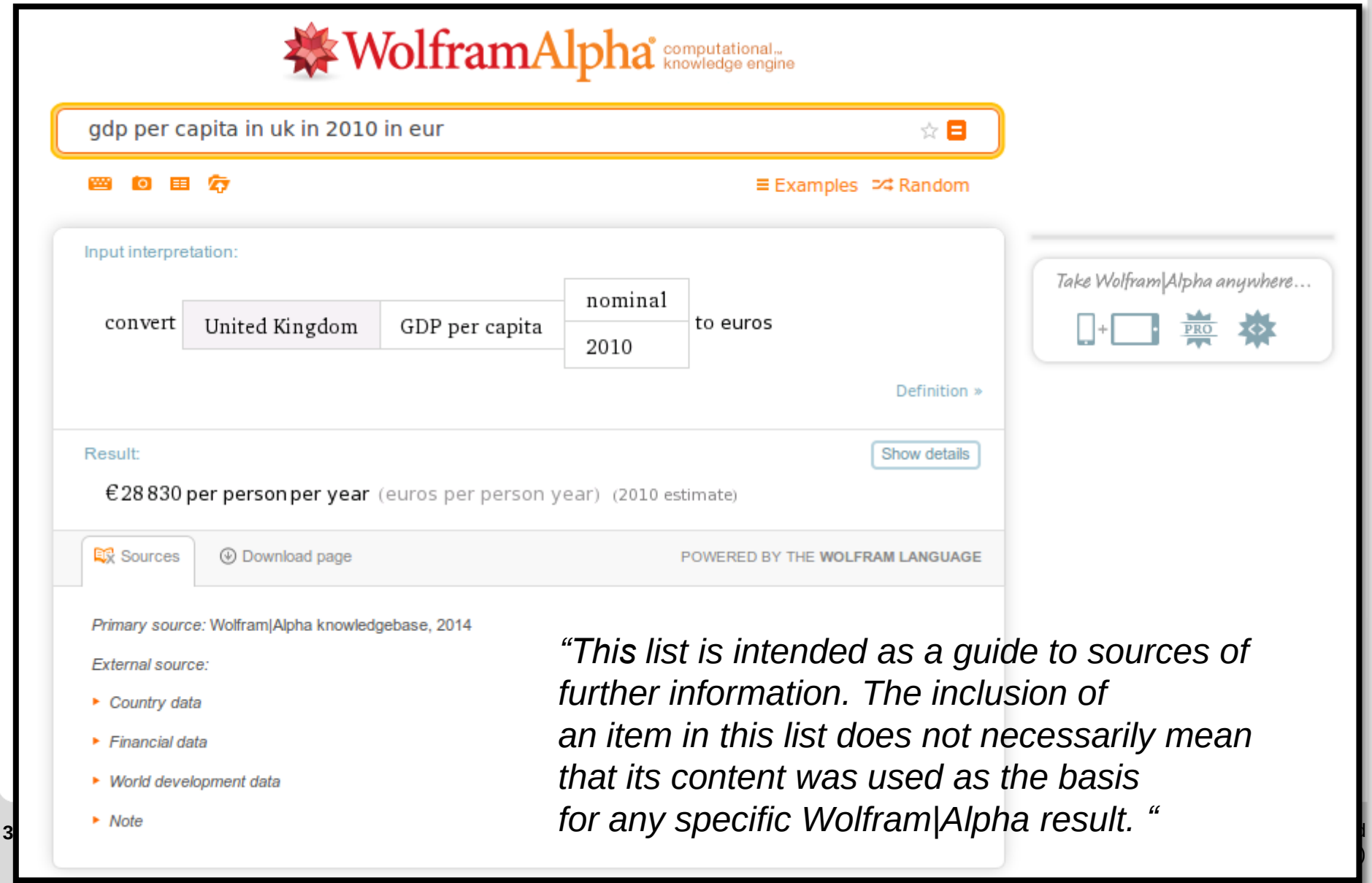




Motivation


WolframAlpha computational... knowledge engine
gdp per capita in uk in 2010 in eur
Examples
Random
Input interpretation:
convert
United Kingdom
GDP per capita
nominal
2010
to euros
Definition »
Result:
Show details
€28 830 per person per year (euros per person year) (2010 estimate)
Sources
Download page
POWERED BY THE WOLFRAM LANGUAGE
Take Wolfram|Alpha anywhere...
PRO

Motivation



"This list is intended as a guide to sources of further information. The inclusion of an item in this list does not necessarily mean that its content was used as the basis for any specific Wolfram|Alpha result. "

Query in Terms of the Global Cube

```
datacube(globalcube).
dimension(globalcube, geo).
dimension(globalcube, unit).
dimension(globalcube, date).
dimension(globalcube, indicna).
dimension(globalcube, sex). ...
measure(globalcube, obsvalue).
globalcube(Geo, Unit, Date, Indicna, ...,
  Obsvalue).

?- globalcube(uk, eur_hab, 2010, ngdph, all,
  ..., Obsvalue).
```

Outline

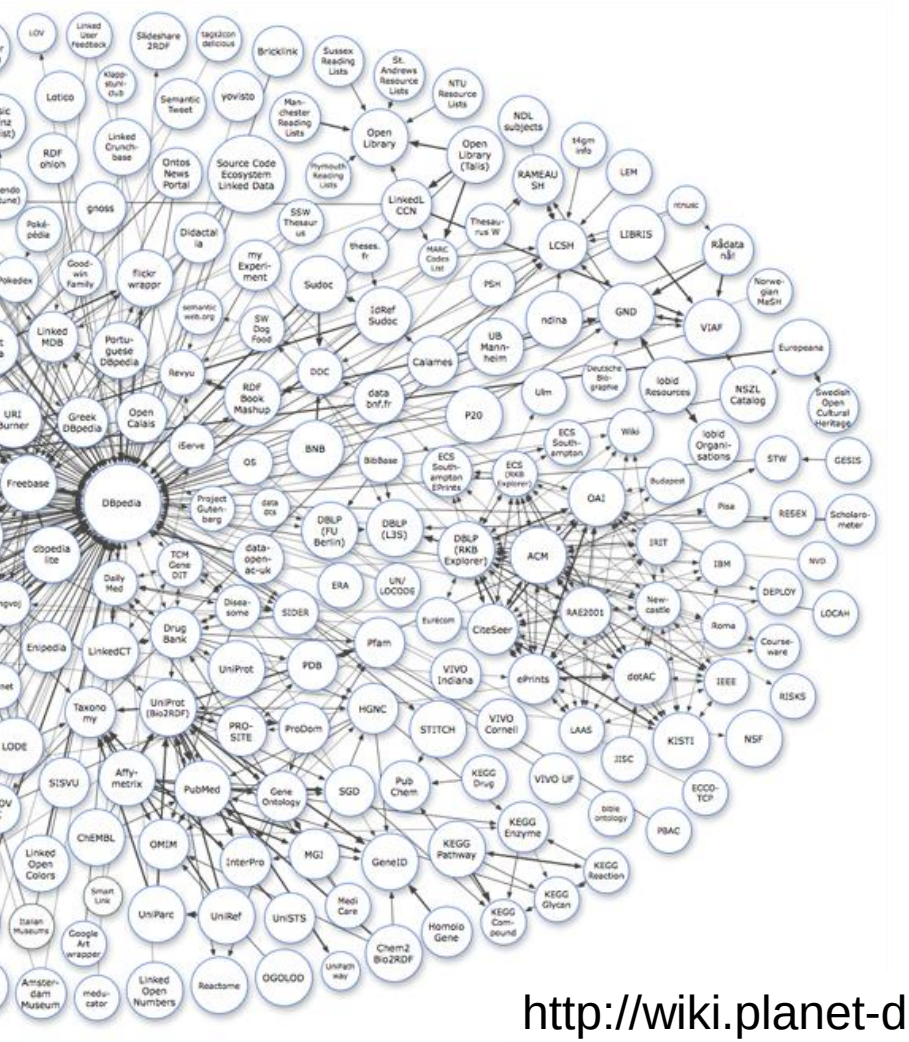
- Global Cube
- Heterogeneity Problems
- Analysis of Size of Global Cube

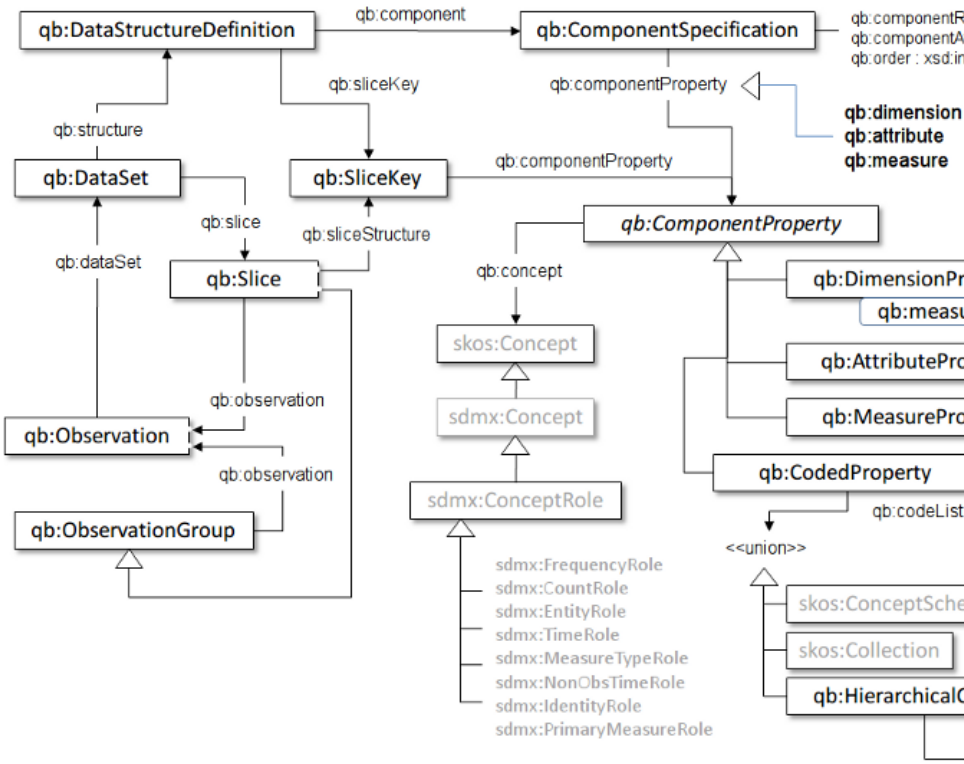
Outline

- **Global Cube**
- Heterogeneity Problems
- Analysis of Size of Global Cube

Institute of Applied Informatics and Formal Description Methods (AIFB)

Assumption: Linked Data and The RDF Data Cube Vocabulary





http://wiki.planet-data.eu/web/Datasets

Multidimensional Dataset

```
datacube(eurostat:id/nama_gdp_c#ds).
dimension(eurostat:id/nama_gdp_c#ds, estatwrap:geo).
dimension(eurostat:id/nama_gdp_c#ds, estatwrap:unit).
dimension(eurostat:id/nama_gdp_c#ds, dcterms:date).
dimension(eurostat:id/nama_gdp_c#ds, estatwrap:indic_na).
measure(eurostat:id/nama_gdp_c#ds, sdmx-measure:obsValue).

eurostat:id/nama_gdp_c#ds(Estatwrap:geo, Estatwrap:unit,
  Dcterms:date, Estatwrap:indic_na, Sdmx-measure:obsValue).

eurostat:id/nama_gdp_c#ds(uk, mioeur, 2010, ngdp, 1731809).
```

Linked Data URIs serving RDF, e.g.,
http://estatwrap.ontologycentral.com/id/nama_gdp_c#ds

Multidimensional Dataset (readable)

```
datacube(gdpcomponents).
dimension(gdpcomponents, geo).
dimension(gdpcomponents, unit).
dimension(gdpcomponents, date).
dimension(gdpcomponents, indicna).
measure(gdpcomponents, obsvalue).

gdpcomponents(Geo, Unit, Date, Indicna, Obsvalue).


gdpcomponents(uk, mioeur, 2010, ngdp, 1731809).
```

Global Cube Definition: Unempl. Fear Example (UNEMPLOY)

```
datacube(gdpgrowth).
datacube(unemployfear).

dimension(gdpgrowth, geo).
dimension(gdpgrwoth, unit).
dimension(gdpgrwoth, date).

dimension(unemployfear, geo).
dimension(unemployfear, date).
dimension(unemployfear, variable).
```

Global Cube Definition: UNEMPLOY (2)

```
dimension(globalcube, X) :-
  dimension(Y, X).


globalcube(Geo, Unit, Date, Variable,
  Obsvalue).
globalcube(Geo, Unit, Date, all,
  ObsValue) :- gdpgrowth(Geo, Unit,
  Date, Obsvalue).
globalcube(Geo, all, Date, Variable,
  ObsValue) :- unemployfear(Geo, Date,
  Variable, Obsvalue).
```

Requirements

- Correct GDP per Capita – confirmed by as many data sources as possible.

- In case there are differences between data sources – we want to know about it.

Outline

- Global Cube
- **Heterogeneity Problems**
- Analysis of Size of Global Cube

Heterogeneity Problems

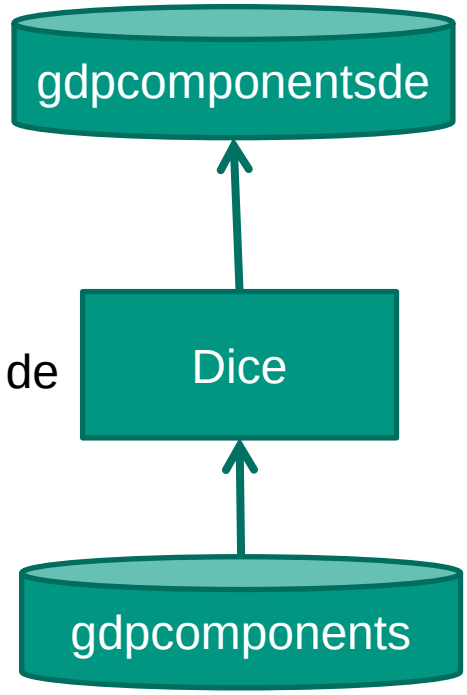
- different dimensions, e.g., sex, age...
- different names for dimensions, e.g., “geo” – “loc”
- ... or for values, e.g., “UK“ vs. “United Kingdom“
- units of measurements, e.g., “mio eur” vs “eur“
- compound indicators, e.g., GDP per Capita
- ...

Heterogeneity Problems

- different dimensions, e.g., sex, age...
- different names for dimensions, e.g., “geo” – “loc”
- ... or for values, e.g., “UK“ vs. “United Kingdom“
- units of measurements, e.g., “mio eur” vs “eur“
- compound indicators, e.g., GDP per Capita
- ...

OLAP Operations (Projection, Dice, Slice, Roll-Up, Drill-Across)

```
Dice(gdpcomponents, {(Geo = de)})
```



Usually represented in SQL over Star Schema

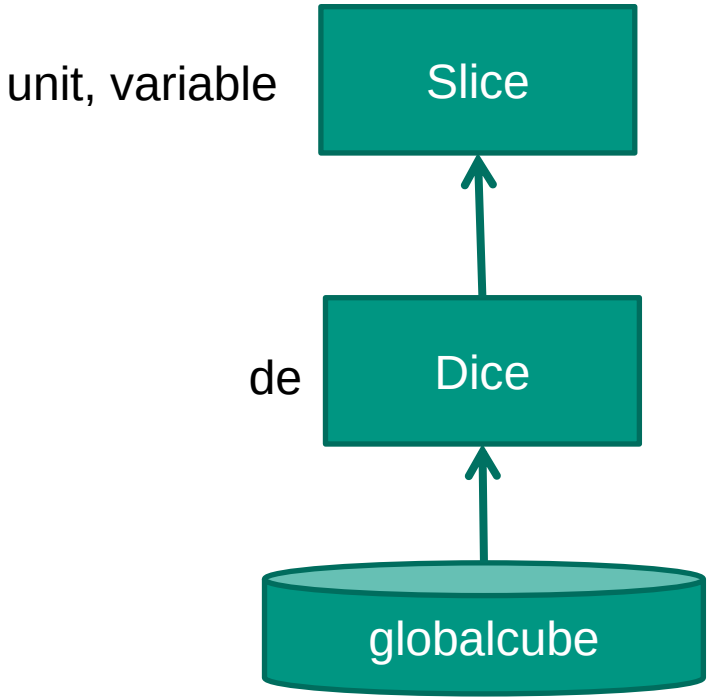
Also possible in Datalog:

```
gdpcomponentsde(Geo, Unit, Date,
  Indicna, Obsvalue) :-
  gdpcomponents(de, Unit, Date,
  Indicna, Obsvalue).
```

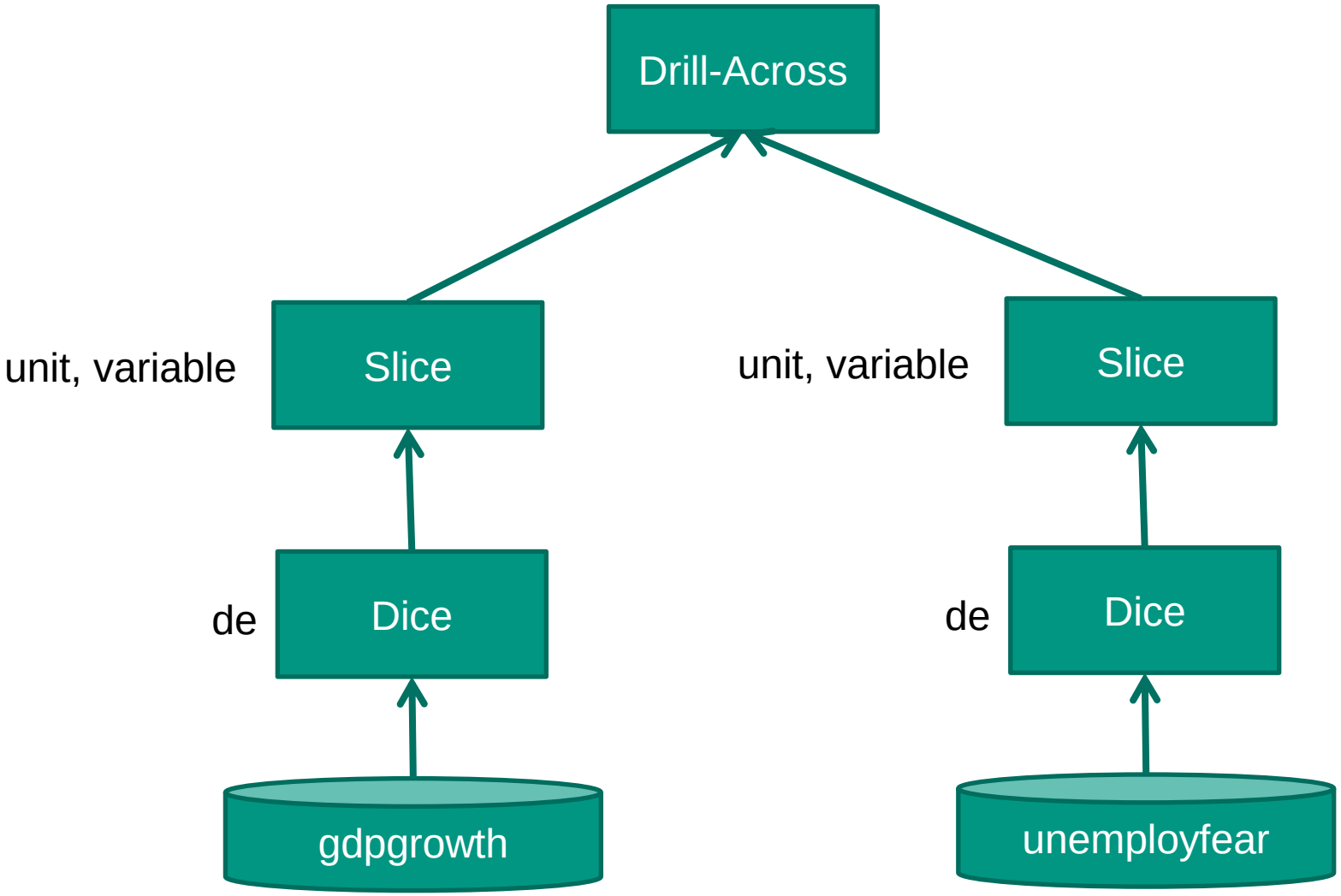
Slice / Roll-Up: Require aggregation (Abhijeet)

Query over the Global Cube (UNEMPLOY)

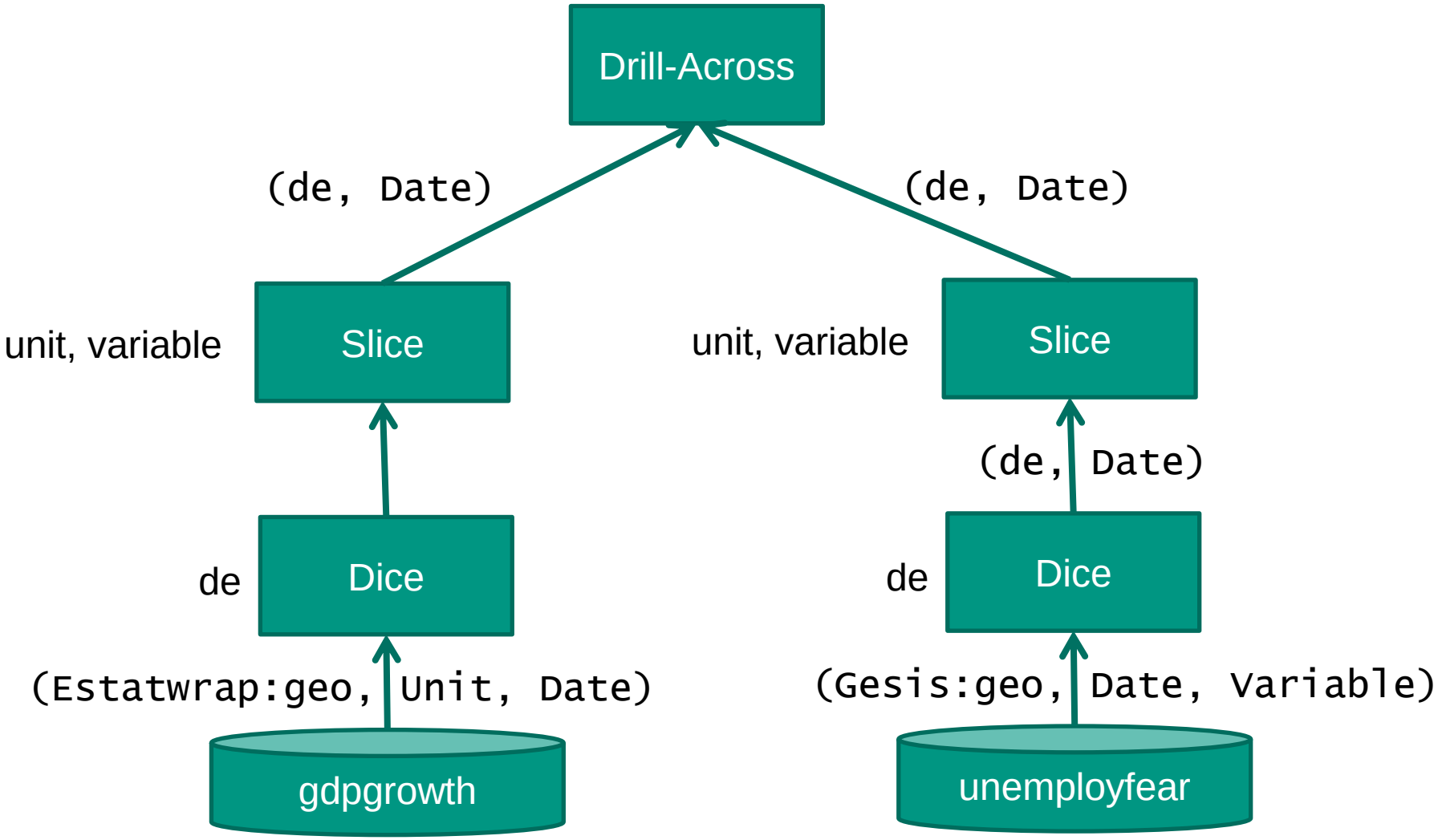
```
Slice(Dice(Projection (globalcube, { obsvalue avg }),
   {Geo = de}), {unit, variable})
```



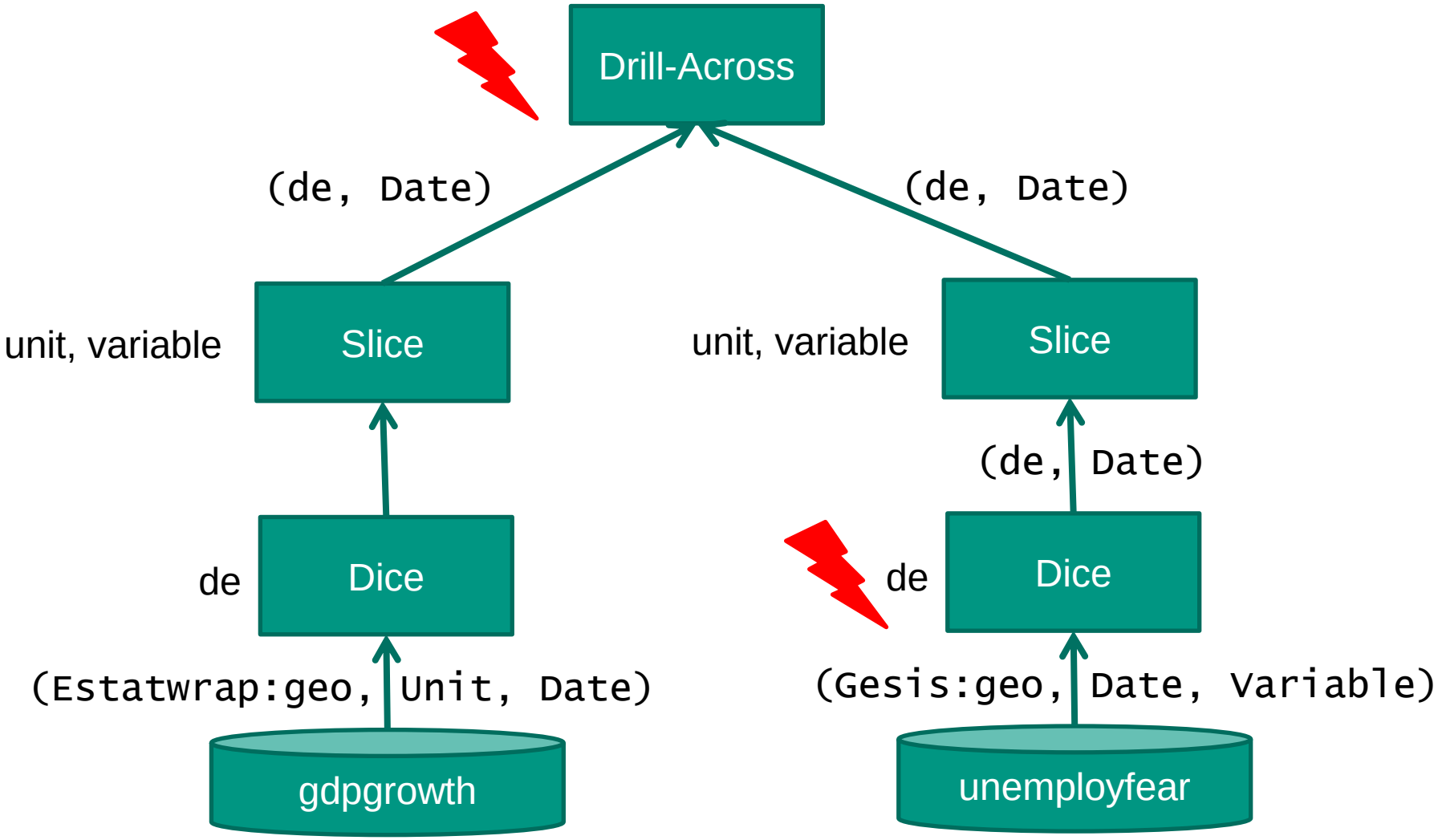
Translated Query over the Global Cube (UNEMPLOY)



Different Names Problem (UNEMPLOY)



Different Names Problem (UNEMPLOY)



Different Names Problem – Solution

OWL equivalence statements in RDF

```
owl:sameAs(estatwrap:geo, gesis:geo).
eurostat-geo:DE owl:sameAs gesis-
  geo:00 .
```

Duplication interpretation of OWL semantics

```
gdpgrowth(Estatwrap:geo, Gesis:geo,
  Unit, Date, Obsvalue)
unemployfear(Estatwrap:geo, Gesis:geo,
  Date, Variable, Obsvalue)
```

Outline

- Global Cube
- **Heterogeneity Problems**
- Analysis of Size of Global Cube

Heterogeneity Problems

- different dimensions, e.g., sex, age...
- different names for dimensions, e.g., “geo” – “loc”
- ... or for values, e.g., “UK“ vs. “United Kingdom“
- units of measurements, e.g., “mio eur” vs “eur“
- compound indicators, e.g., GDP per Capita
- ...

Units and Compounds Problem

```
gdppercapita(uk, eur_hab, 2010, ngdph,
  27800).
gdpcomponents(uk, mioeur, 2010, ngdp,
  1731809).
population(uk, hab, 2010, t, total,
  62510197).


? :- globalcube(uk, eur_hab, 2010,
  ngdph, all, all, Obsvalue).
```

Units and Compounds Problem – Solution

Conversion and Merging Correspondences (declarative)

```
MIO2EUR =
({(unit, :MIO_EUR)}, {(unit, :EUR)}, f (x) = 1, 000, 000 · x)

COMP_GDP_CAP = (
{(indic_na, :NGDP), (unit, :EUR)}, {(sex, :T), (age, :TOTAL)},
{(indic na, :NGDPH), (unit, :EUR HAB)},
f (x 1 , x 2 ) = x 1 /x 2 )
```

Convert-Cube and Merge-Cubes operations

```
datacube(MIO2EUR(X)) :- datacube(X).

datacube(COMP_GDP_CAP(X,Y)) :- datacube(X), datacube(Y).
```

Can be executed using Datalog/Prolog

Architecture of Integration System

- GDP Per Capita (GDP_CAP)

  Looked-up: 27,800

  vs.

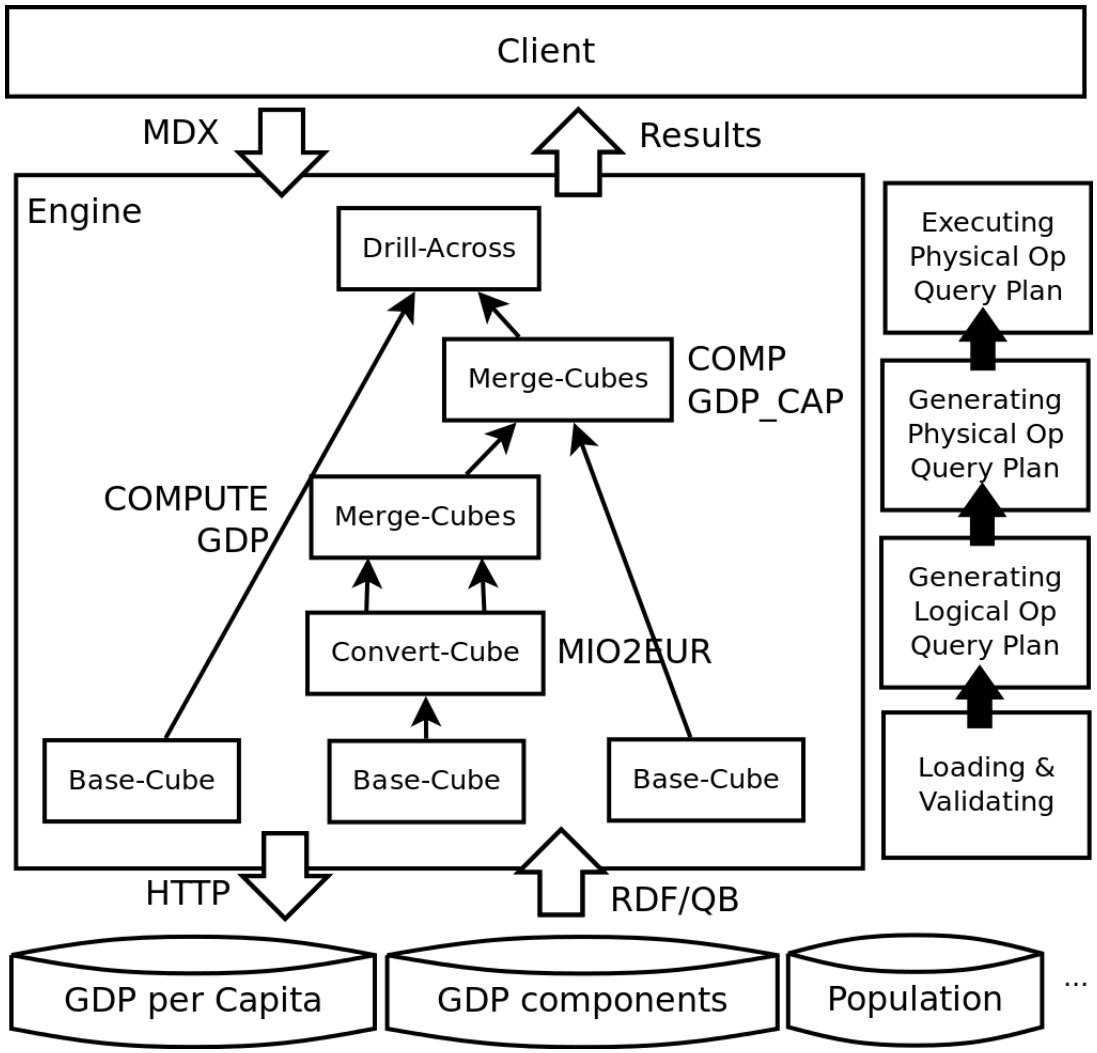
  Computed: 27,704

- Execution

  Top-Down

  vs.

  Bottom-Up



Outline

- Global Cube
- Heterogeneity Problems
- **Analysis of Size of Global Cube**

Analysis of Size of Global Cube

Depending on number of derived data cubes from conversion and merging operations:

1. No restrictions
2. No cycles
3. Matching of dimension-member combinations

How to restrict the possible number of conversion and merging operations?

1) No restrictions

```
datacube(gdppercapita).
datacube(mio2eur(X)) :- datacube(X).
```

Unlimited

How to restrict the possible number of conversion and merging operations? (2)

2) No repeated applications / cycles

```
datacube(mio2eur(X)) :- datacube(X), \+
  derivedby(X, mio2eur).
derivedby(mio2eur(X), mio2eur) :-
  datacube(X), \+ derivedby(X, mio2eur).
```

noderiveds(ds, mc) a lot, XSB Prolog would not be able to compute them in GDP per Capita scenario (1x Convert-Cube, 2x Merge-Cubes)

How to restrict the possible number of conversion and merging operations? (3)

3) Considering matching dimension-member combinations

```
/* The generation of the dataset */
datacube(mio2eur(X)) :- datacube(X),
  dimension(X,unit), ( \+
  dimensionmember(X,unit,Z);
  dimensionmember(X,unit,mioeur) ).
/* Copying over of dimensions */
dimension(mio2eur(X), Y) :- datacube(X),
  dimension(X,unit), ( \+
  dimensionmember(X,unit,Z);
  dimensionmember(X,unit,mioeur) ),
  dimension(X,Y).
```

How to restrict the possible number of conversion and merging operations? (4)

```
/* Copying over of dimension members */
dimensionmember(mio2eur(X), Y, V) :-
  datacube(X), dimension(X,unit), ( \+
  dimensionmember(X,unit,Z);
  dimensionmember(X,unit,mioeur) ),
  dimensionmember(X,Y,V), \+
  dimensionmember(X, unit, mioeur).
/* Setting the new dimension member */
dimensionmember(mio2eur(X), unit, eur) :-
  datacube(X), dimension(X,unit), ( \+
  dimensionmember(X,unit,Z);
  dimensionmember(X,unit,mioeur) ).
```

In practice few (in GDP per Capita scenario, only 54)

Conclusions

- Many different multidimensional datasets available as **Linked Data**
- No easy querying due to **heterogeneities**
- **Global Cube** as a unified schema of all such datasets
- Pay-as-you-go integration with **two mapping** approaches

- WolframAlpha and flexibly adding data sources if
  - 1) we are able to materialise the Global Cube offline (execute queries efficiently over RDF + recognise if datasets change),
  - 2) we resolve the inherent semantic conflicts in Open Government Data

A different topic...

- Looked-up vs Computed
- 27,800 27,704

Thanks!

- Looked-up vs Computed
- 27,800 vs 27,704

eurostat

Backup: Evaluation

- Implementation in OLAP4LD [1]
  - Drill-Across (loading and query processing grow linearly)

Exper.	#DS	#T	#O	L&V	QP	T
UNEMPLOY	2	3,897	362	11	11	22
EU2020a	4	19,714	2,212	18	21	39
EU2020b	8	38,069	3,992	47	56	103

  - Convert-Cube Queries (without full materialisation)

Exper.	#DS	#T	#O	L&V	QP	T
GDP_CAP	3	1,015,044	126,351	119	127	246

[1] http://olap4ld.googlecode.com/

Backup: Two smaller problems

- Integrity constraint violations
- Open World assumption

Integrity constraint violations

```
"integrity constraint violation" :-
  ds(D1, ..., Dn, M1), ds(D1, ..., Dn,
  M2), M1 != M2.
```

Solution

Drill-Across(ds1, ds2) = ds3 with

1) If dimensions(ds1) != dimensions(ds2) then:
   dimension(ds3) = dimension(ds1) UNION dimension(ds2)
   and ds3(D1, ..., Dn, M) empty.

2) Else: ds3(D1, ..., Dn, M) :- ds1(D1, ..., Dn, M1),
   ds2(D1, ..., Dn, M2), M = f(M1, M2).

f identifying "integrity constraint violations".

Open World assumption

In Open World assumption not possible

```
ds3(D1, ..., Dn, M) :- ds1(D1, ...,
  Dn, M1), ds2(D1, ..., Dn, M2), M =
  f(M1, M2).
```

Solution

We have to assume that the dimensions are known and the relation of the multidimensional dataset fixed.