The Universal Data Cube

Curran Kelleher

Outline

- Weave
- Problems
- The Universal Data Cube
- Solutions
- Open discussion

Weave

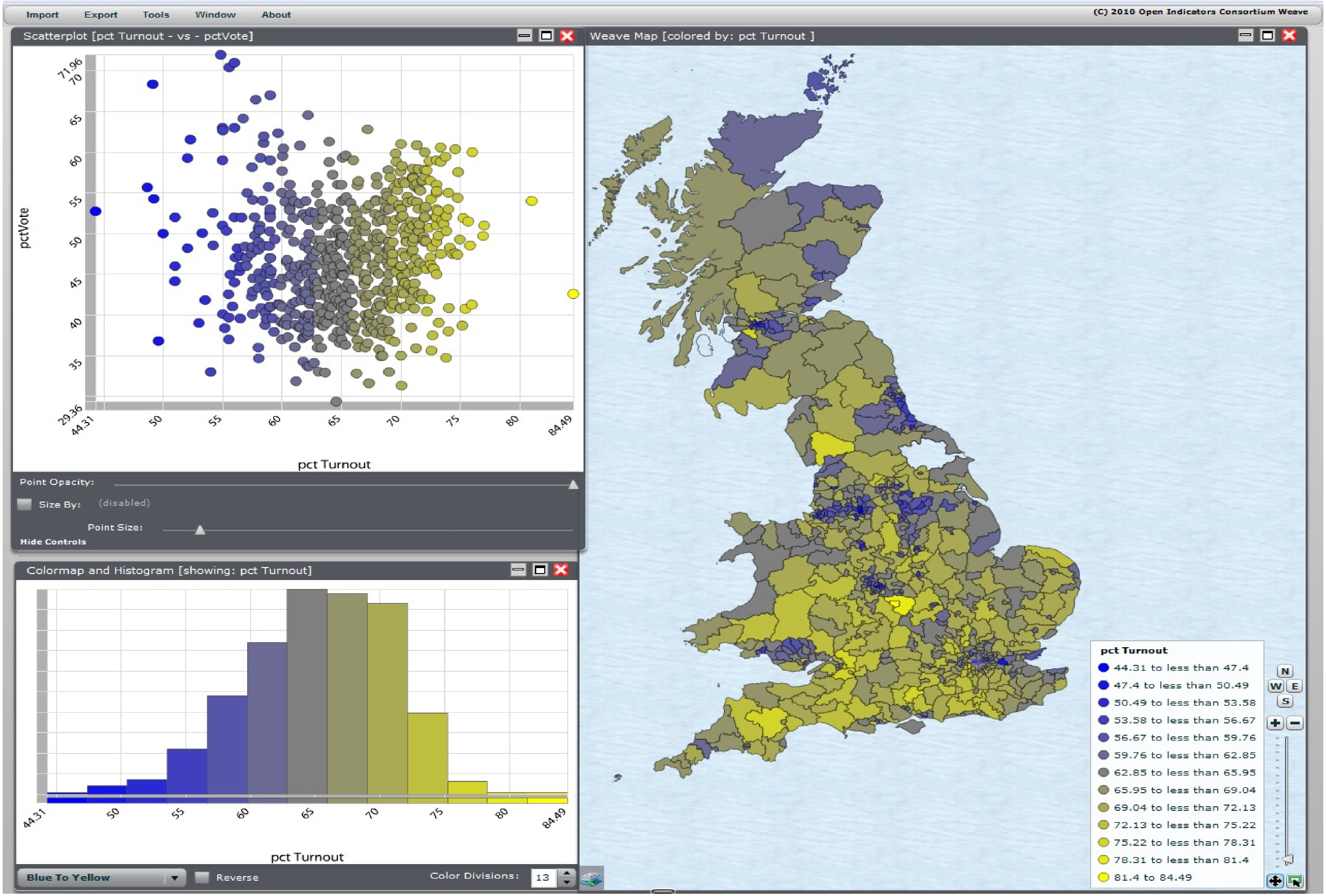
Weave

Web-based Analysis and Visualization Environment

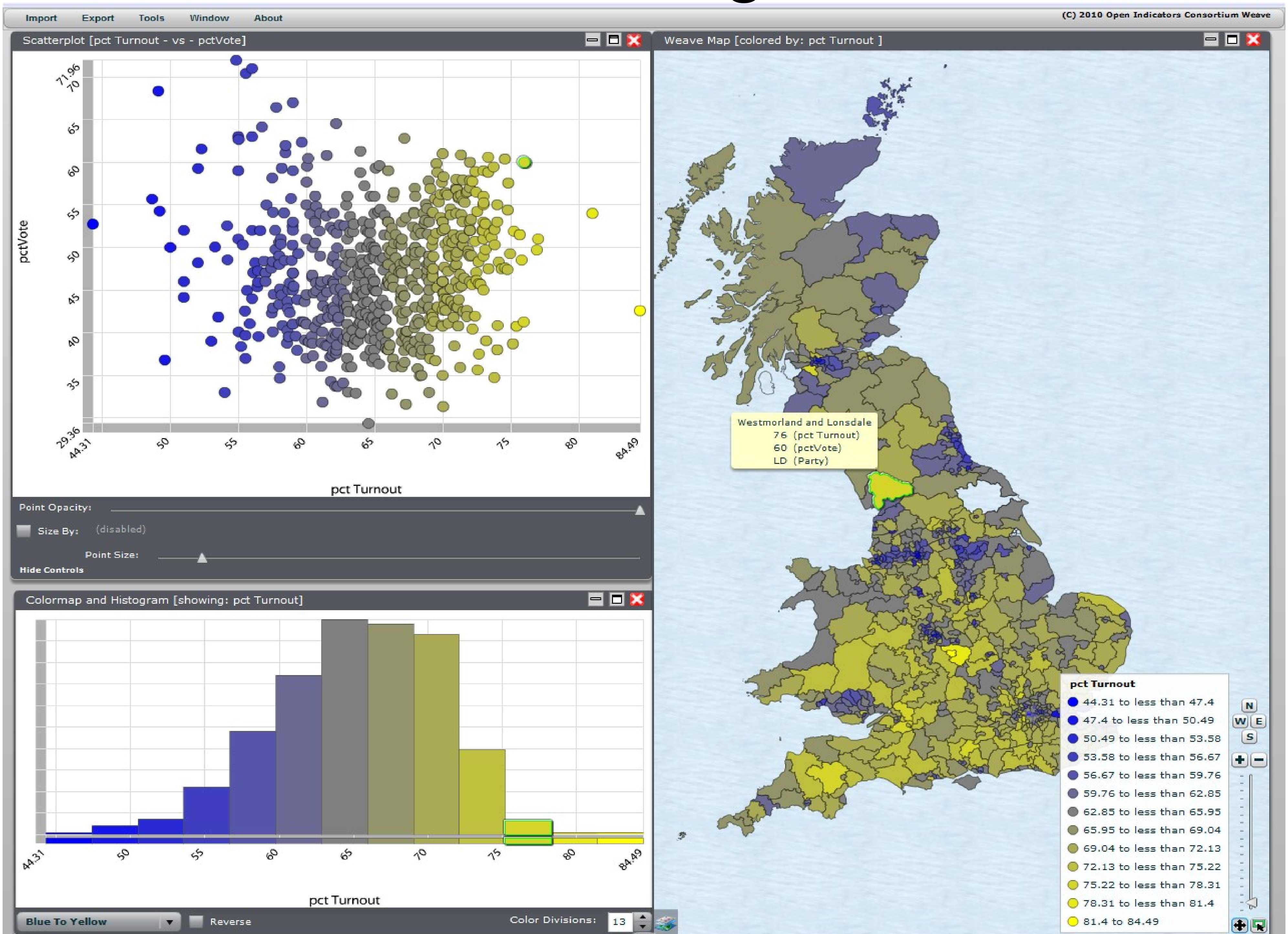
- Developed by the IVPR group
- Funded by the Open Indicators Consortium
- Client written in Adobe Flex
- Server written in Java

British Election Results in May 2010

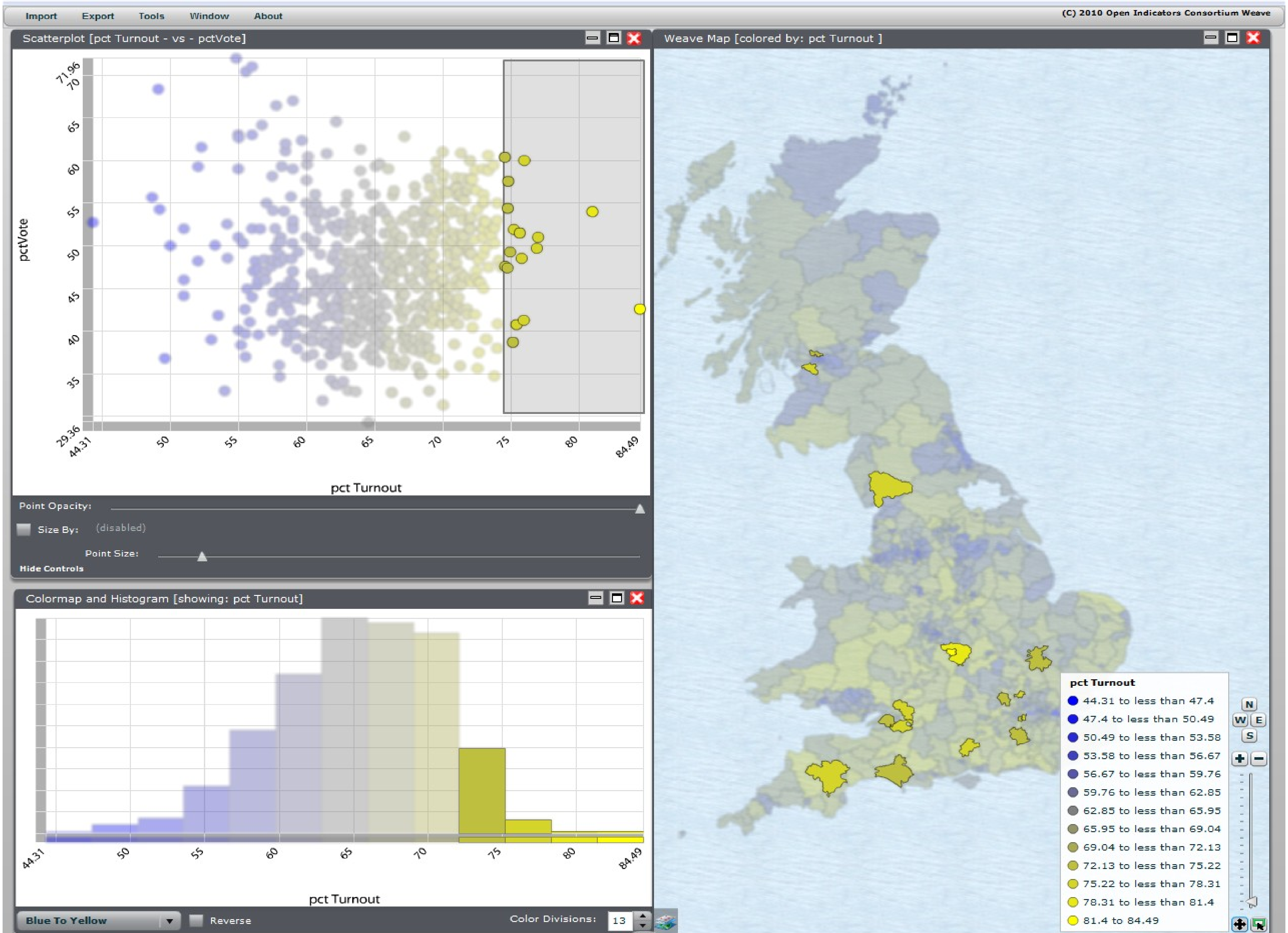
assembled by Jim Giddings



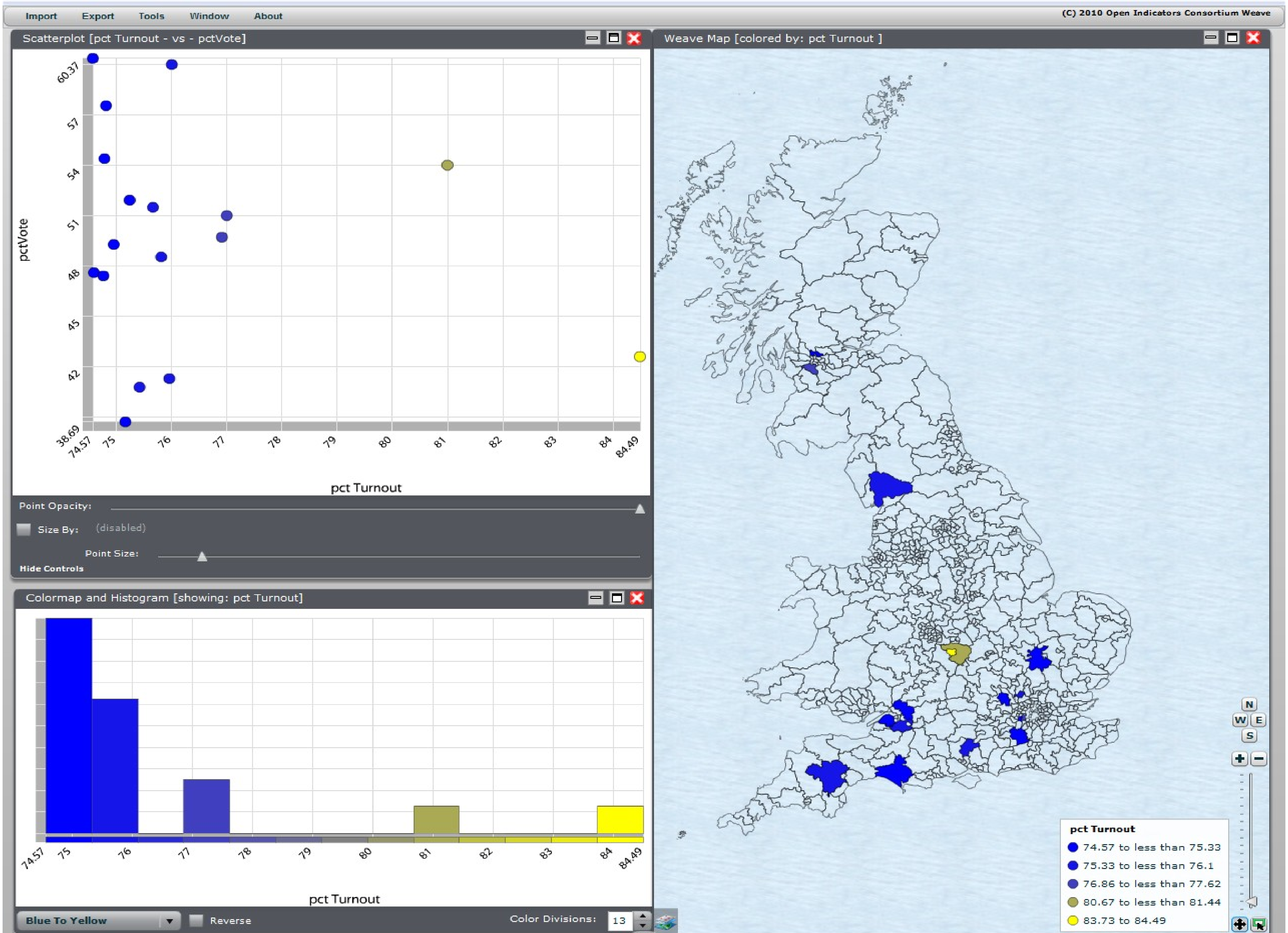
Probing


Import
Export
Tools
Window
About
(C) 2010 Open Indicators Consortium Weave
Scatterplot [pct Turnout - vs - pctVote]
pctVote
pct Turnout
Point Opacity:
Size By: (disabled)
Point Size:
Hide Controls
Weave Map [colored by: pct Turnout]
Westmorland and Lonsdale
76 (pct Turnout)
60 (pctVote)
LD (Party)
pct Turnout
44.31 to less than 47.4
47.4 to less than 50.49
50.49 to less than 53.58
53.58 to less than 56.67
56.67 to less than 59.76
59.76 to less than 62.85
62.85 to less than 65.95
65.95 to less than 69.04
69.04 to less than 72.13
72.13 to less than 75.22
75.22 to less than 78.31
78.31 to less than 81.4
81.4 to 84.49
Colormap and Histogram [showing: pct Turnout]
pct Turnout
Blue To Yellow
Reverse
Color Divisions: 13

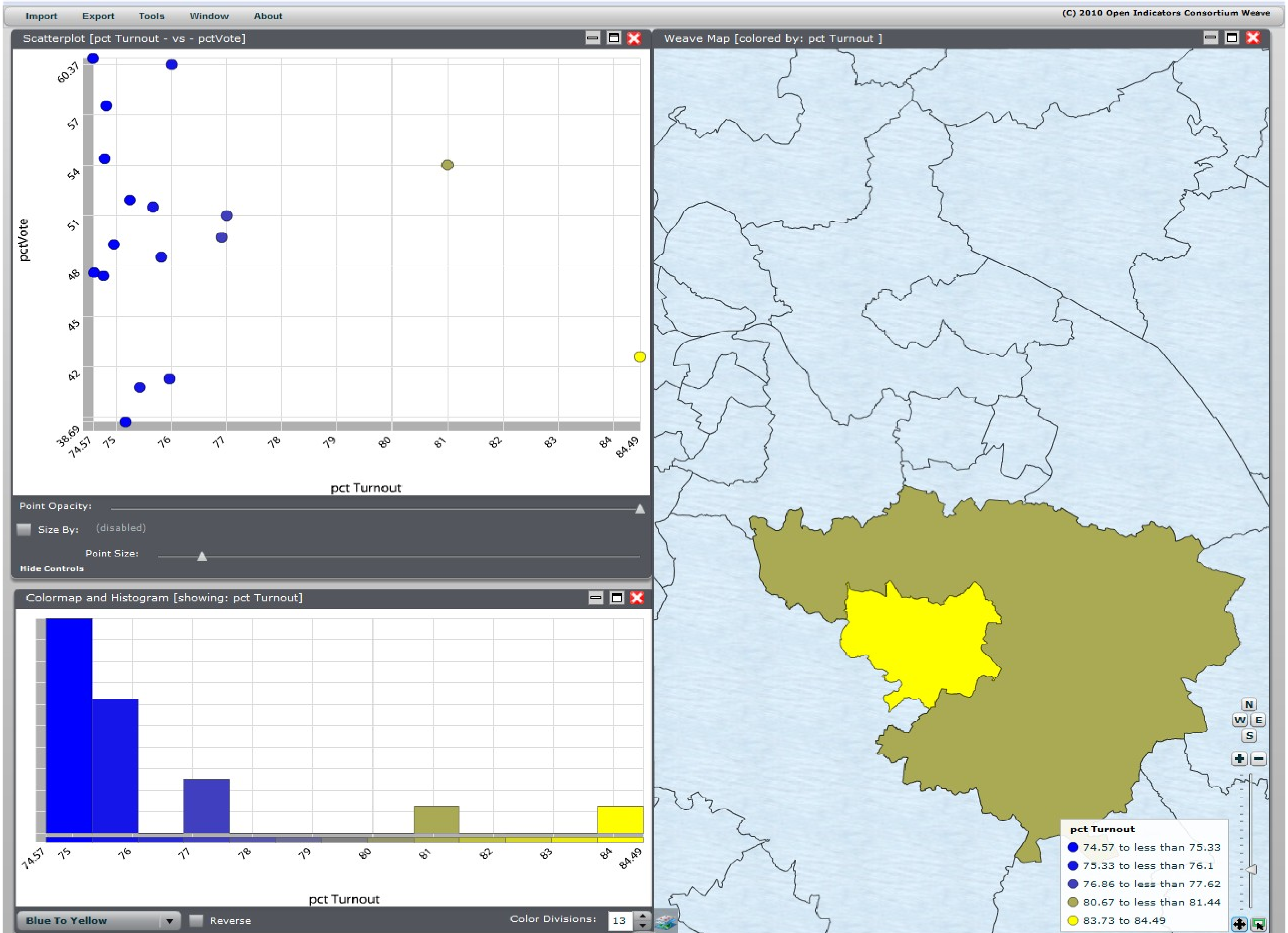
Brushed selection

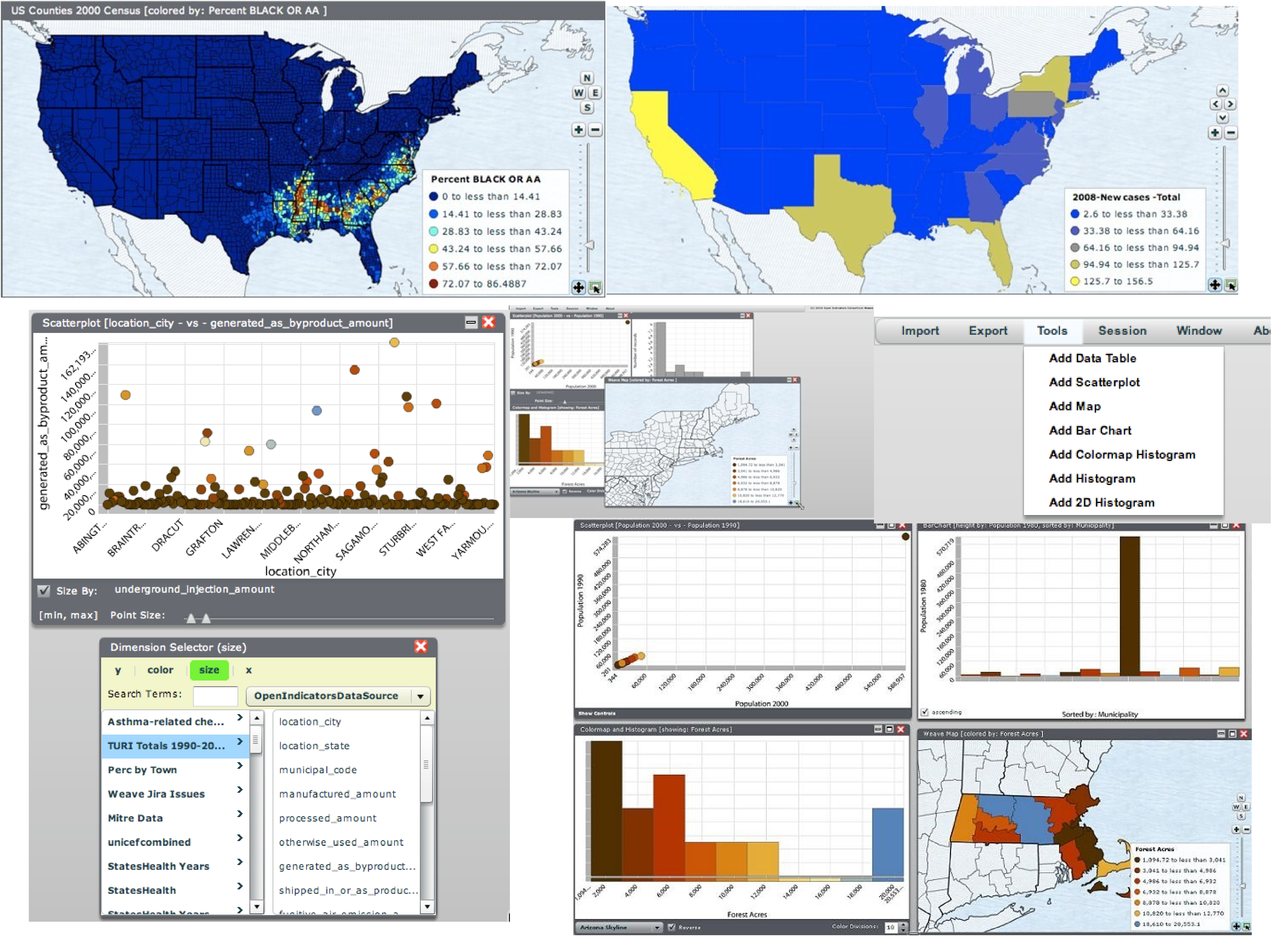


Dynamic filtering



Map navigation





US Counties 2000 Census [colored by: Percent BLACK OR AA]
Percent BLACK OR AA
0 to less than 14.41
14.41 to less than 28.83
28.83 to less than 43.24
43.24 to less than 57.66
57.66 to less than 72.07
72.07 to 86.4887
2008-New cases -Total
2.6 to less than 33.38
33.38 to less than 64.16
64.16 to less than 94.94
94.94 to less than 125.7
125.7 to 156.5
Scatterplot [location_city - vs - generated_as_byproduct_amount]
generated_as_byproduct_am...
location_city
ABINGT... BRAINTR... DRACUT GRAFTON LAWREN... MIDDLEB... NORTHAM... SAGAMO... STURBRI... WEST FA... YARMOU...
Size By: underground_injection_amount
[min, max] Point Size:
Dimension Selector (size)
y color size x
Search Terms: OpenIndicatorsDataSource
Asthma-related che...
TURI Totals 1990-20...
Perc by Town
Weave Jira Issues
Mitre Data
unicefcombined
StatesHealth Years
StatesHealth
location_city
location_state
municipal_code
manufactured_amount
processed_amount
otherwise_used_amount
generated_as_byproduct...
shipped_in_or_as_produc...
Import Export Tools Session Window
Add Data Table
Add Scatterplot
Add Map
Add Bar Chart
Add Colormap Histogram
Add Histogram
Add 2D Histogram
Scatterplot [Population 2000 - vs - Population 1990]
Population 1990
Population 2000
Show Controls
BarChart [height by: Population 1980, sorted by: Municipality]
Population 1980
ascending
Sorted by: Municipality
Colormap and Histogram [showing: Forest Acres]
Forest Acres
Arizona Skyline
Reverse
Color Divisions:
Weave Map [colored by: Forest Acres]
Forest Acres

The Weave Data Model

- Data is a collection of columns
  - Multiple simultaneous data sources are supported
- Columns are placed in a category hierarchy
- Columns have names
- Columns have associated key types
- Key type indicates what kind of thing records are
  - For example “US State FIPS code”

Weave Data Model Problems

- Hierarchical key types are not linked
  - US Counties and US States are totally independent
- Key types referring to the same things not linked
  - “US State FIPS” != “US State abbreviations”
- Columns representing the same measure with different units are not compatible
  - Population in thousands not comparable with Population in millions
- No way of resolving when two datasets provide comparable columns
  - Is column “Pop” the same thing in dataset A and B?

Data Cubes

Informative clips from the 2002 paper

Multiscale Visualization Using Data Cubes

by Chris Stolte, Diane Tang, and Pat Hanrahan

2 Related Work

Data abstraction refers to transformations applied to the data before being visually mapped, including aggregation, filtering, sampling, or statistical summarization.

Multiscale Visualization in Cartography

Data cubes are a commonly accepted method for abstracting and summarizing relational databases.

Data Abstraction: Data Cubes

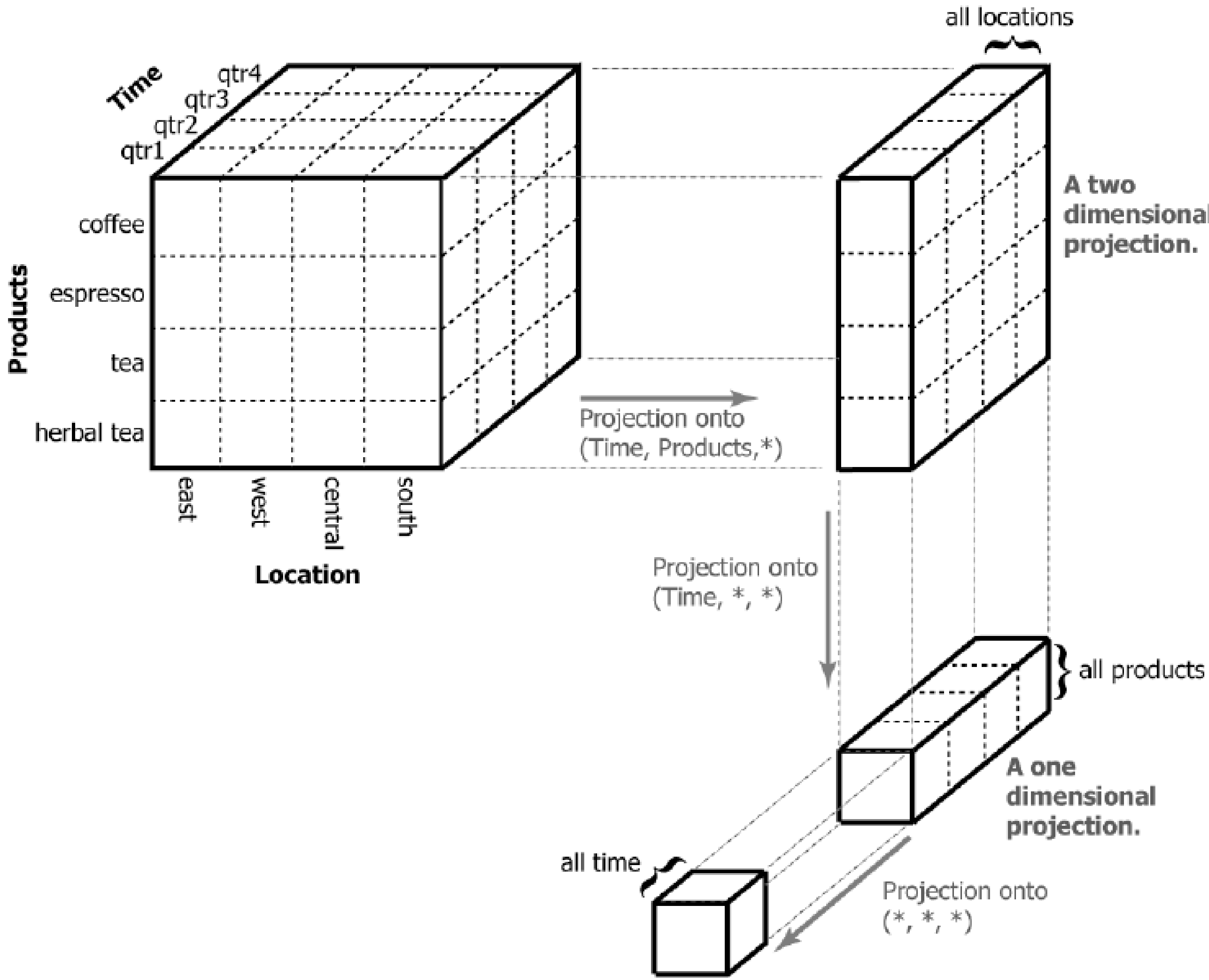
Data cubes categorize information into two classes: dimensions and measures, For example, U.S. states are a dimension, while the population of each state is a measure.

data is abstractly structured as an n-dimensional data cube. Each axis corresponds to a dimension in the data cube and consists of every possible value for that dimension. For example, an axis corresponding to states would have fifty values, one for each state.

Ev-
ery "cell" in the data cube corresponds to a unique combination of values for the dimensions.

Each cell contains one value per measure of the data cube

Projecting a three dimensional data cube



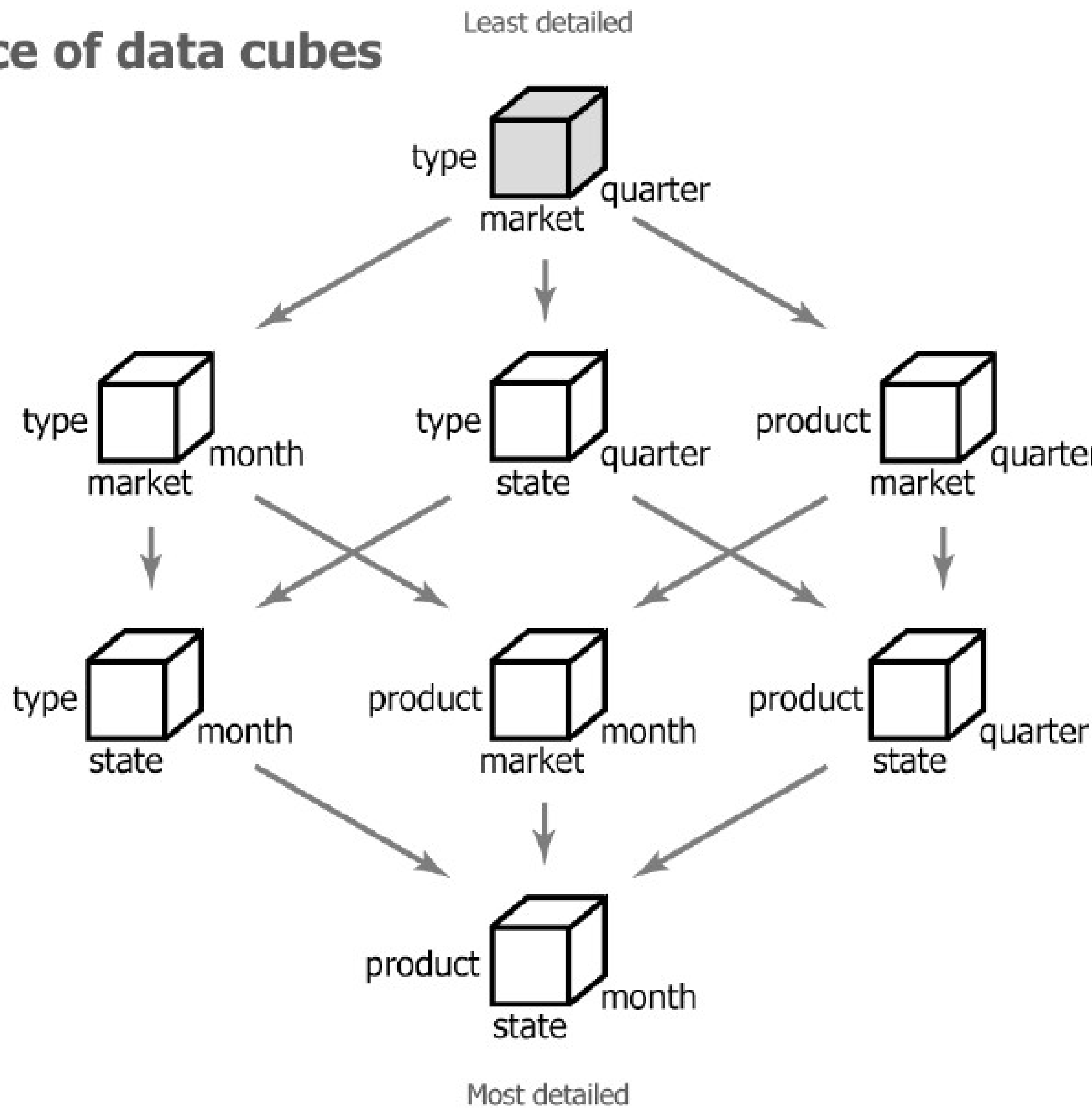
Hierarchical Data Cubes

Thus far, we have considered dimensions to be flat structures. However, most dimensions have a hierarchical structure.

For example, rather than having a single dimension "state", we may have a hierarchical dimension "location" that has levels for country, state, and county.

If each dimension has a hierarchical structure, then the data must be structured as a lattice of data cubes

The lattice of data cubes



Choosing a data abstraction corresponds to choosing a particular *projection* in this lattice of data cubes:

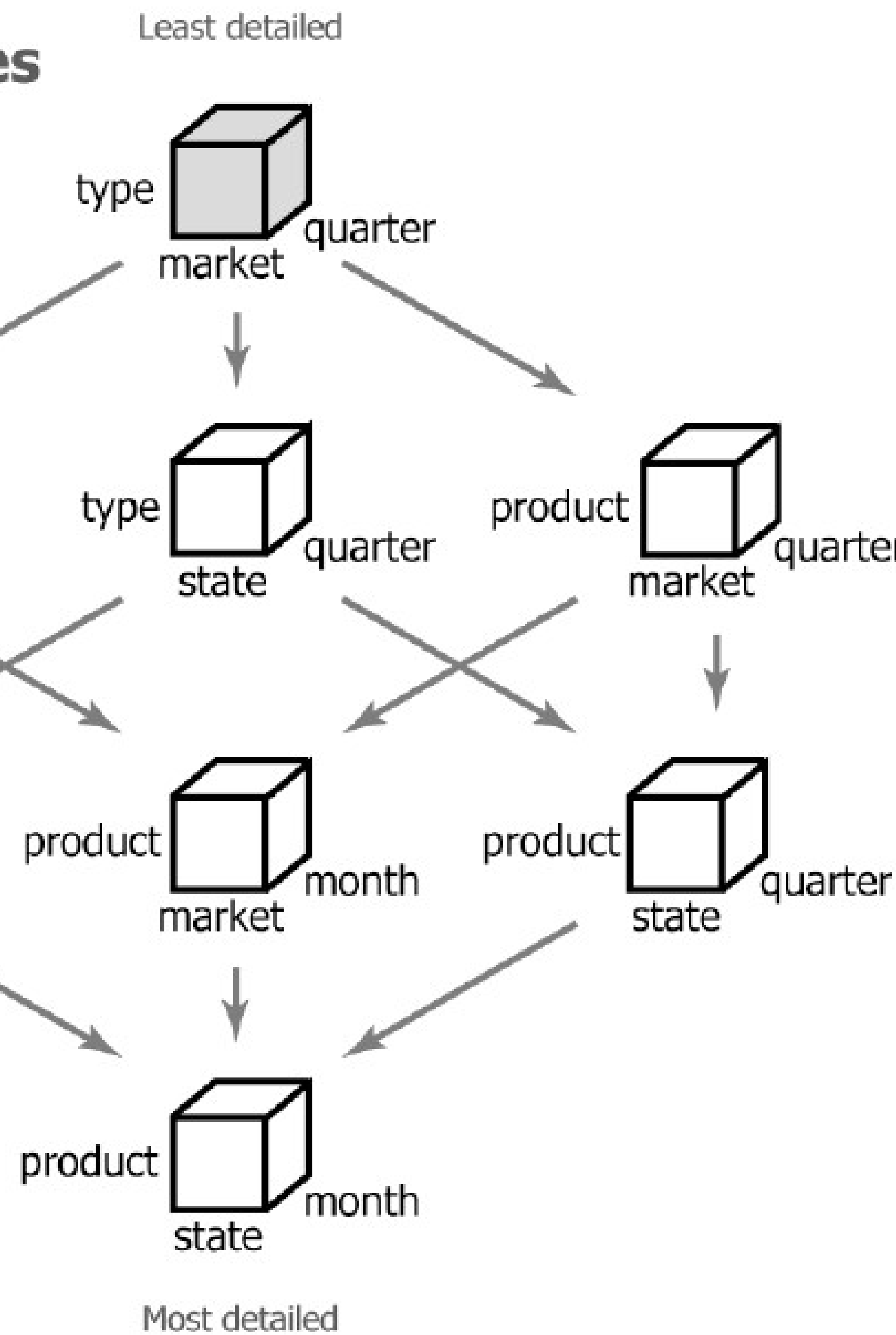
particular *projection* in this lattice of data cubes: (a) which dimensions we currently consider relevant and (b) the appropriate level of detail for each relevant dimensional hierarchy.

(b)

level of detail

identifies the cube in the lattice

The lattice of data cubes



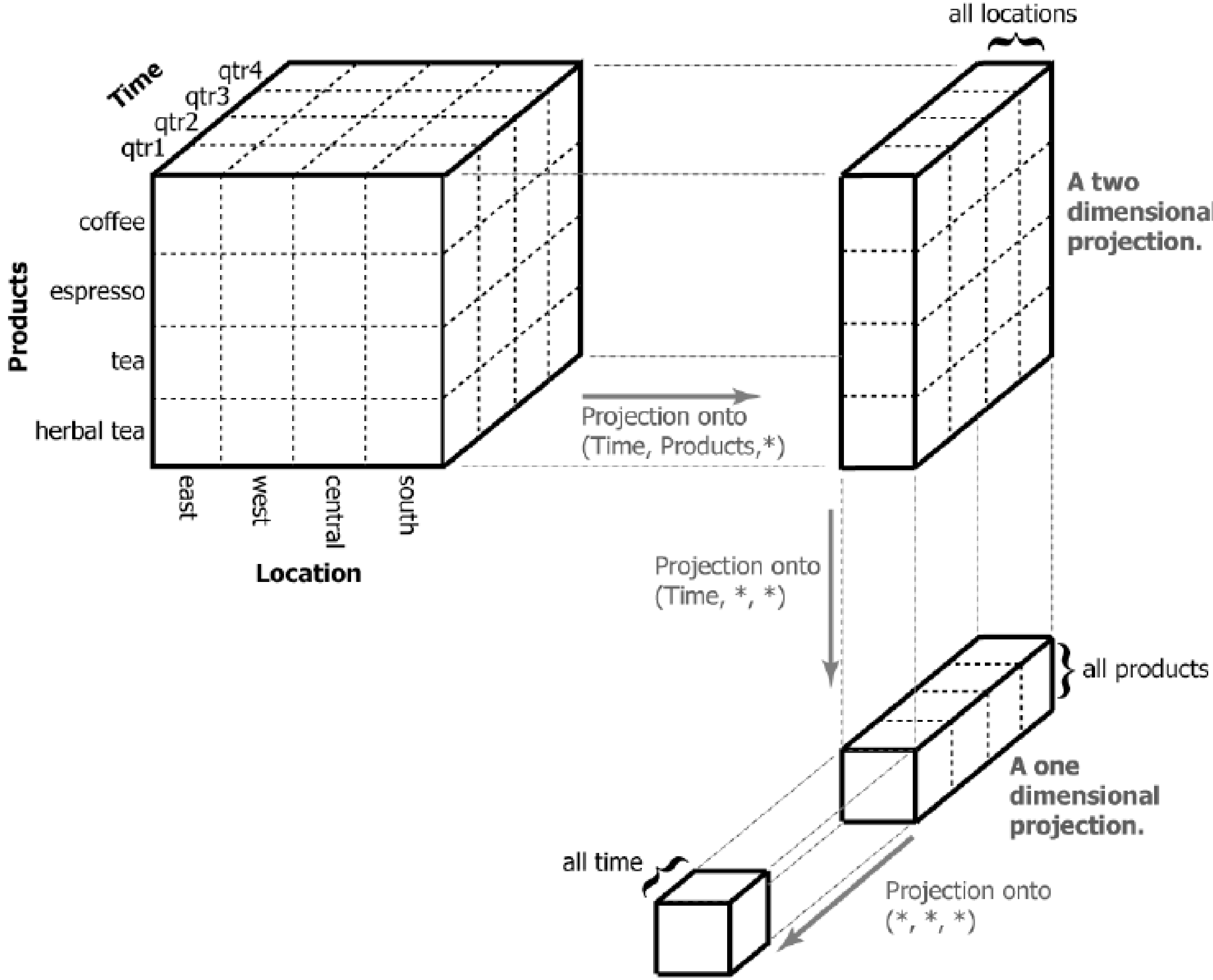
(a)

the relevant dimensions identifies which projection

of that

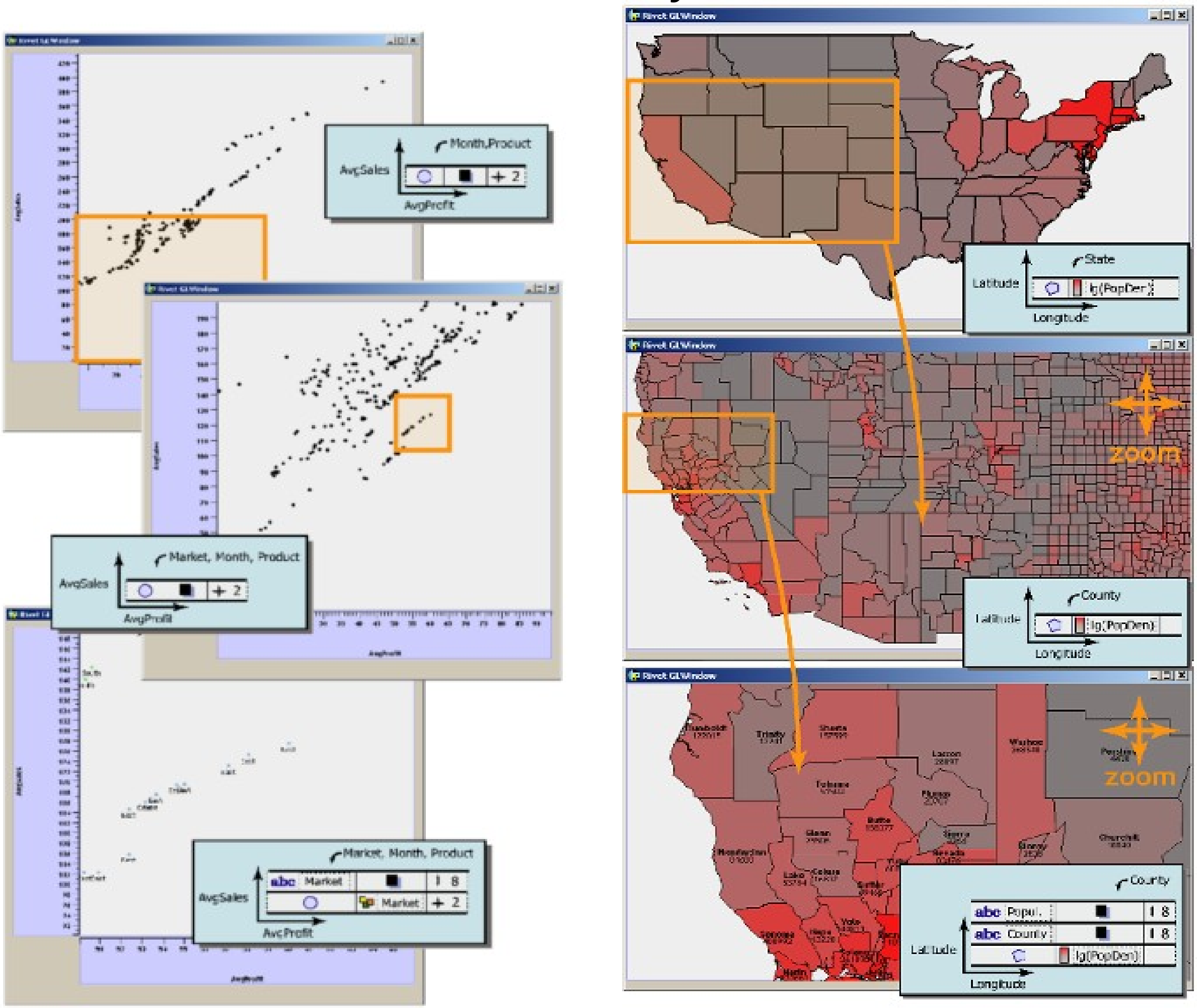
cube is needed.

Projecting a three dimensional data cube



Example: Data Cube Navigation

in the Polaris system



an example data cube: the

U.S. Bureau of Labor Statistics (BLS) Employment Dataset

The BLS Employment Dataset

as dimensions and measures

- Raw data at ftp://ftp.bls.gov/pub/special.requests/cew/
- Covers Time from 1990 to 2007
  - Data for years, quarters, and months
- Covers Space for all US States
  - Data for States and Counties
- Covers the NAICS Industry hierarchy
- Covers Ownership
  - Government (Federal, State, Local) and Private
- Contains measures employment, annual pay, total wages, and number of establishments (among others)

NAICS

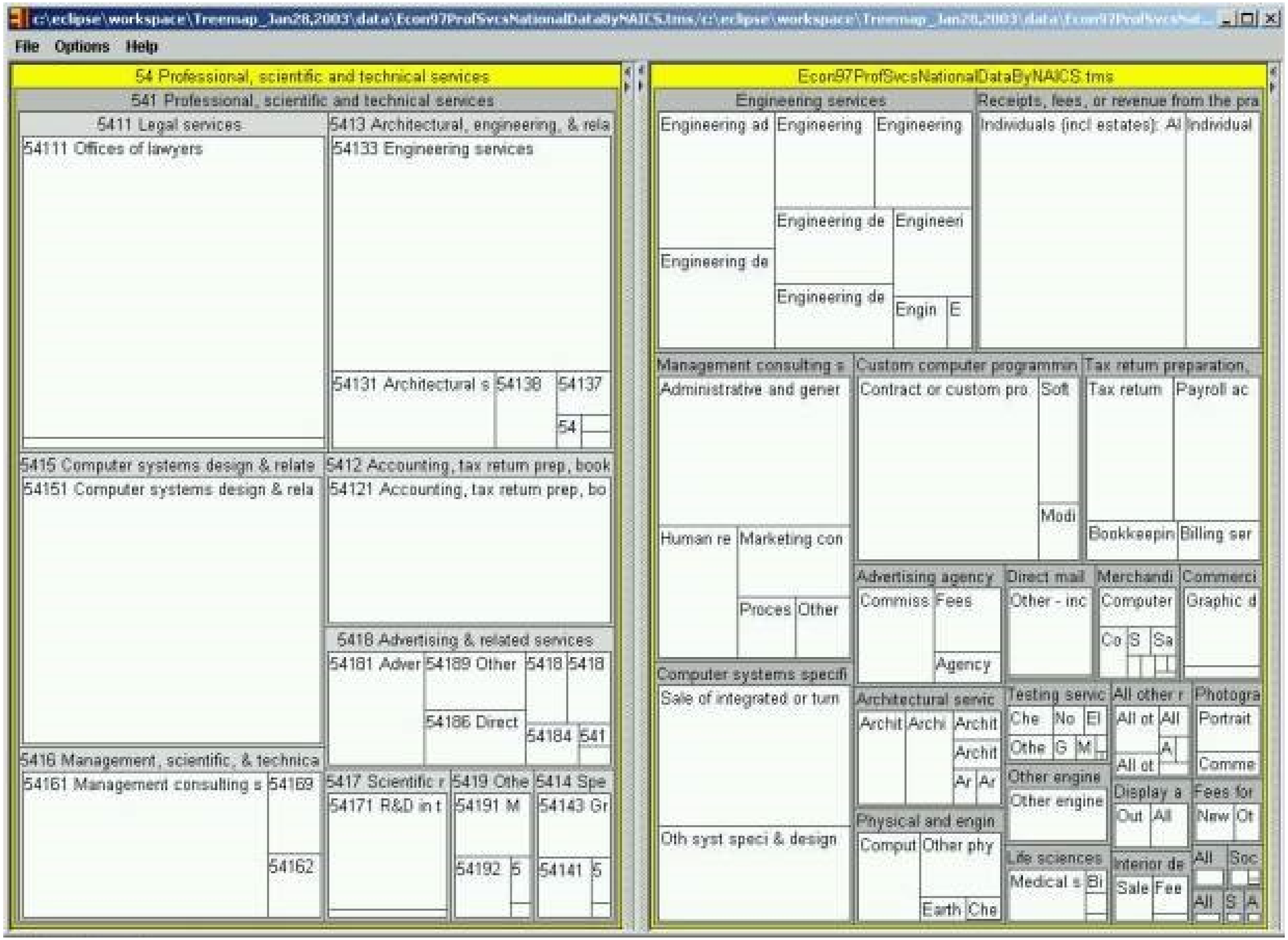
North American Industry Classification System

11	Agriculture, Forestry, Fishing and Hunting
111	Crop Production
1111	Oilseed and Grain Farming
11111	Soybean Farming
111110	Soybean Farming
11112	Oilseed (except Soybean) Farming
111120	Oilseed (except Soybean) Farming
11113	Dry Pea and Bean Farming
111130	Dry Pea and Bean Farming
11114	Wheat Farming
111140	Wheat Farming
11115	Corn Farming
111150	Corn Farming
11116	Rice Farming
111160	Rice Farming
11119	Other Grain Farming
111191	Oilseed and Grain Combination Farming
111199	All Other Grain Farming
1112	Vegetable and Melon Farming
11121	Vegetable and Melon Farming
111211	Potato Farming
111219	Other Vegetable (except Potato) and Melon Farming
1113	Fruit and Tree Nut Farming
11131	Orange Groves
111310	Orange Groves
11132	Citrus (except Orange) Groves
111320	Citrus (except Orange) Groves
11133	Noncitrus Fruit and Tree Nut Farming
111331	Apple Orchards
111332	Grape Vineyards
111333	Strawberry Farming
111334	Berry (except Strawberry) Farming
111335	Tree Nut Farming
111336	Fruit and Tree Nut Combination Farming
111339	Other Noncitrus Fruit Farming
1114	Greenhouse, Nursery, and Floriculture Production
11141	Food Crops Grown Under Cover
111411	Mushroom Production
111419	Other Food Crops Grown Under Cover
11142	Nursery and Floriculture Production
111421	Nursery and Tree Production
111422	Floriculture Production
1119	Other Crop Farming
11191	Tobacco Farming
111910	Tobacco Farming
11192	Cotton Farming
111920	Cotton Farming
11193	Sugarcane Farming
111930	Sugarcane Farming
11194	Hay Farming
111940	Hay Farming
11199	All Other Crop Farming
111991	Sugar Beet Farming
111992	Peanut Farming
111998	All Other Miscellaneous Crop Farming
112	Animal Production
1121	Cattle Ranching and Farming
11211	Beef Cattle Ranching and Farming, including Feedlots
112111	Beef Cattle Ranching and Farming
112112	Cattle Feedlots
11212	Dairy Cattle and Milk Production
112120	Dairy Cattle and Milk Production
11213	Dual-Purpose Cattle Ranching and Farming
112130	Dual-Purpose Cattle Ranching and Farming
1122	Hog and Pig Farming
11221	Hog and Pig Farming
112210	Hog and Pig Farming
1123	Poultry and Egg Production
11231	Chicken Egg Production
112310	Chicken Egg Production
11232	Broilers and Other Meat Type Chicken Production

Industry
Accommodation and food services
Administrative and waste services
Agriculture, forestry, fishing and hunting
All industries
Arts, entertainment, and recreation
Construction
Educational services
Finance and insurance
Health care and social assistance
Information
Management of companies and enterprises
Mining, quarrying, and oil and gas extraction
Other services, except public administration
Professional and technical services
Public Administration
Real estate and rental and leasing
Unclassified
Utilities
Wholesale trade

NAICS Treemap by Revenue

from University of Maryland using US Census data

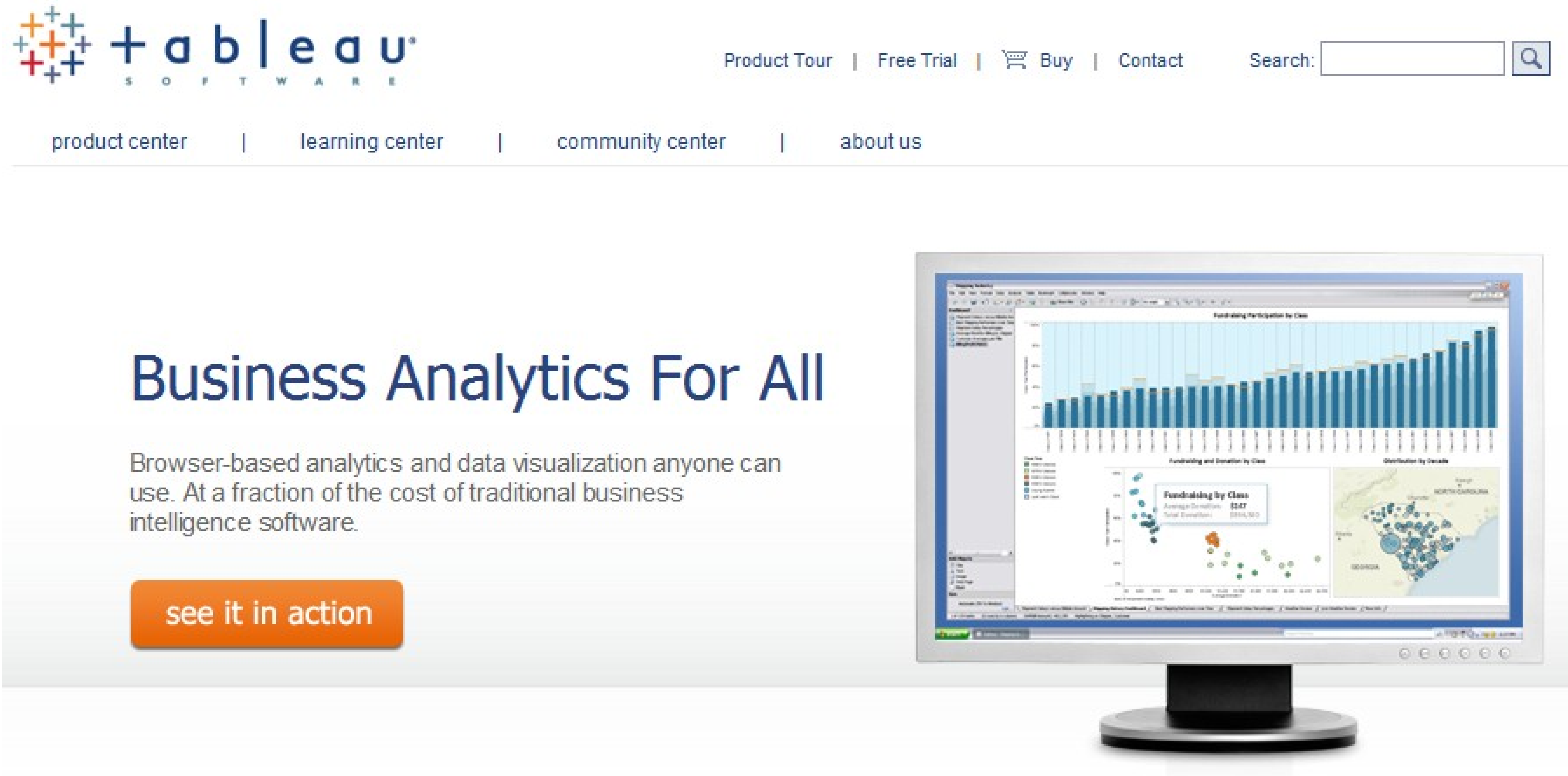


from http://hcil.cs.umd.edu/trs/2003-09/2003-09.html

Tableau

A commercial visual analysis tool

- Uses the data cube model
- From the authors of “Multiscale Visualization using Data Cubes”



The BLS Employment dataset

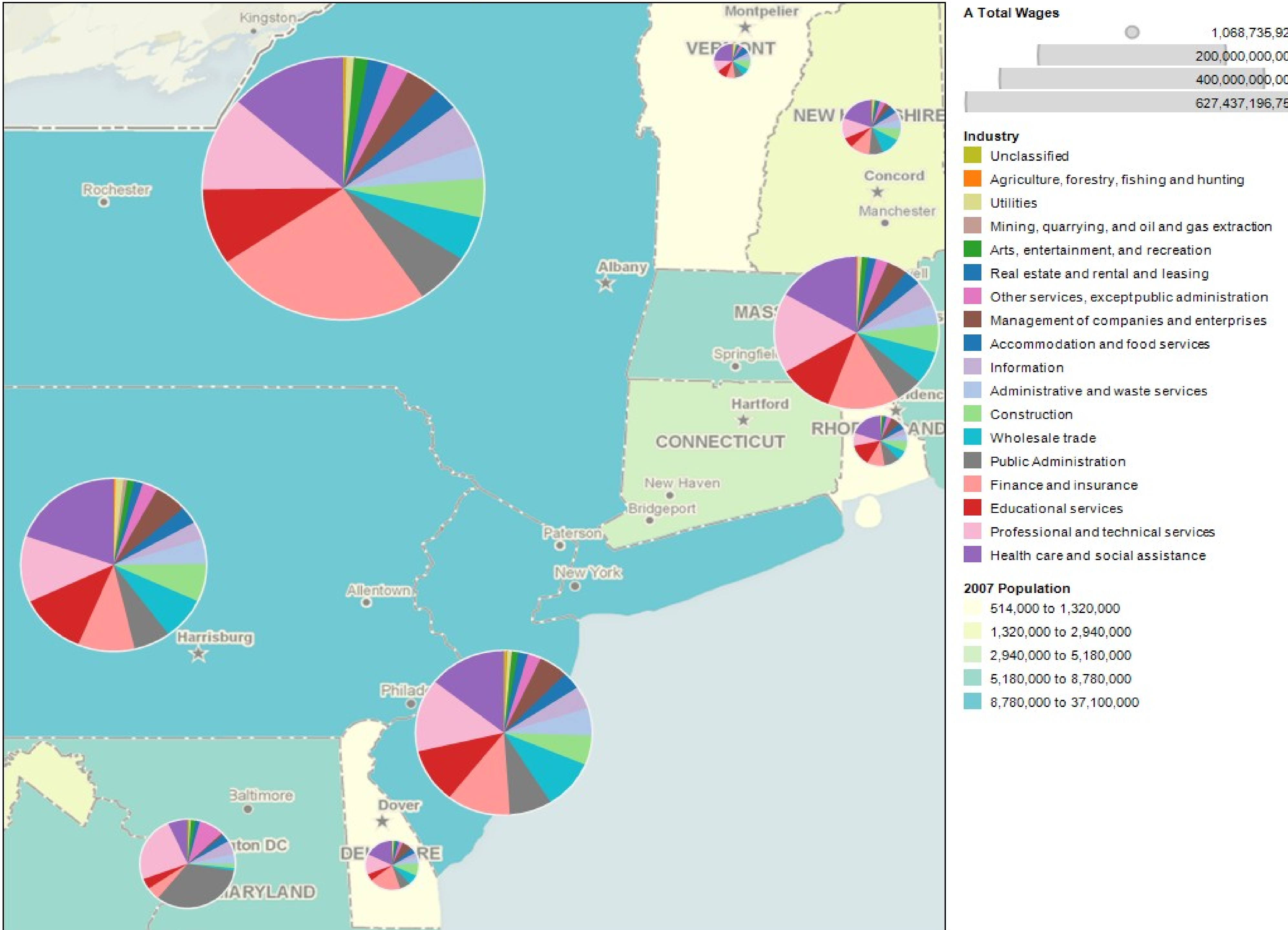
Visualized

using

Tableau

from a project by Siva Mohan and Curran Kelleher

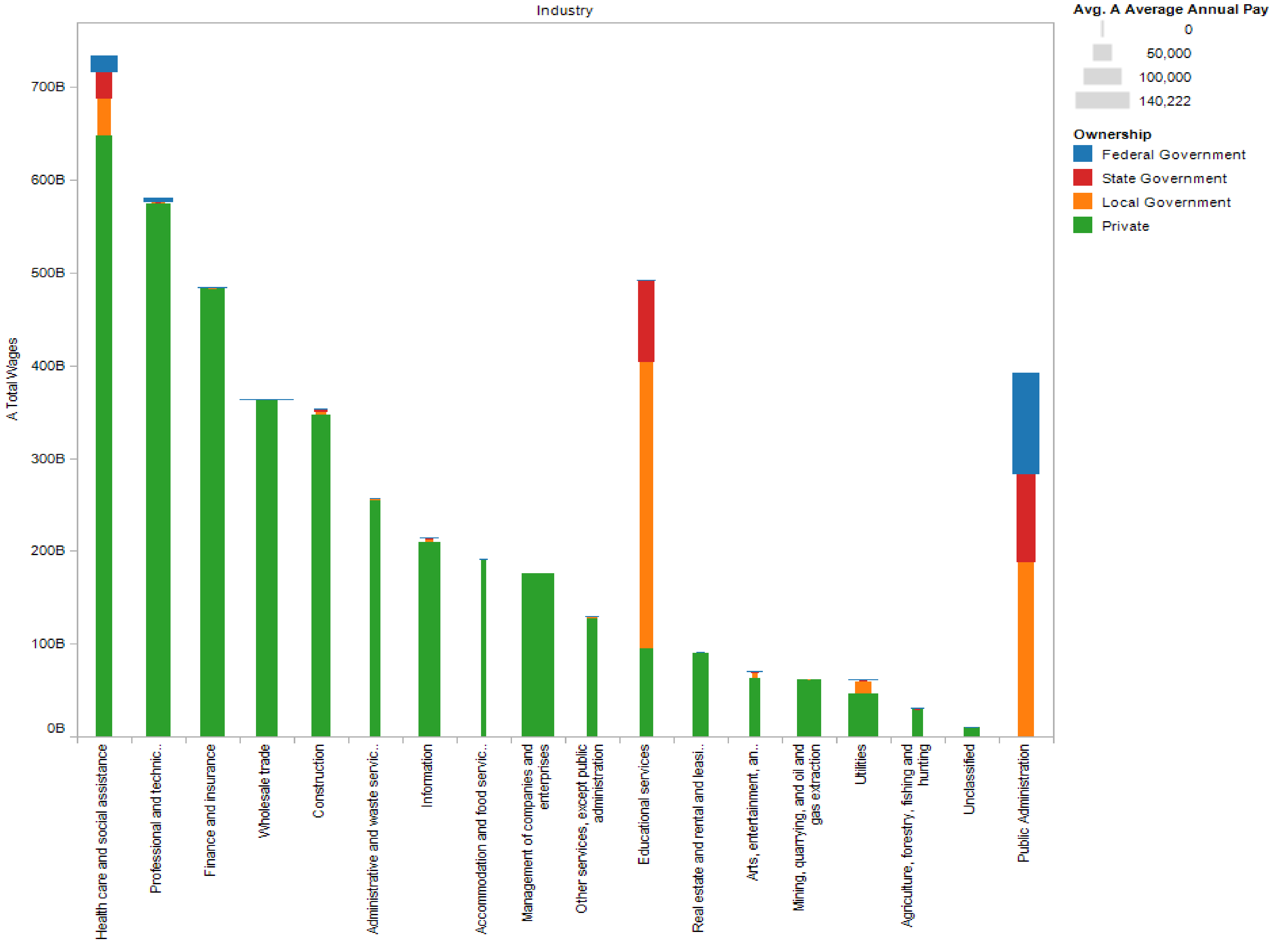
New England Pies



Map based on Longitude (generated) and Latitude (generated). Color shows details about Industry. Size shows sum of A Total Wages. Details are shown for ALPHA. The view is filtered on Industry and Exclusions (ALPHA,Industry). The Industry filter excludes All industries. The Exclusions (ALPHA,Industry) filter specifies a set.

Industry (color) and sum of A Total Wages (size) broken down by Ownership vs. NAME. The view is filtered on NAME and Industry. The NAME filter excludes 48 members. The Industry filter excludes 10 members.

Industries divided by Ownership



Sum of A Total Wages for each Industry. Color shows details about Ownership. Size shows average of A Average Annual Pay. The view is filtered on Industry, which excludes All industries.

Issues with Tableau

- No support for hierarchical data cubes
  - Only a small subset of the dataset usable: states, years, top level industries
- Dealing with Time was problematic
  - Years in different tables
  - Months in different columns
  - Tableau expects single column dimensions

The Semantic Web

Semantic Web Technologies

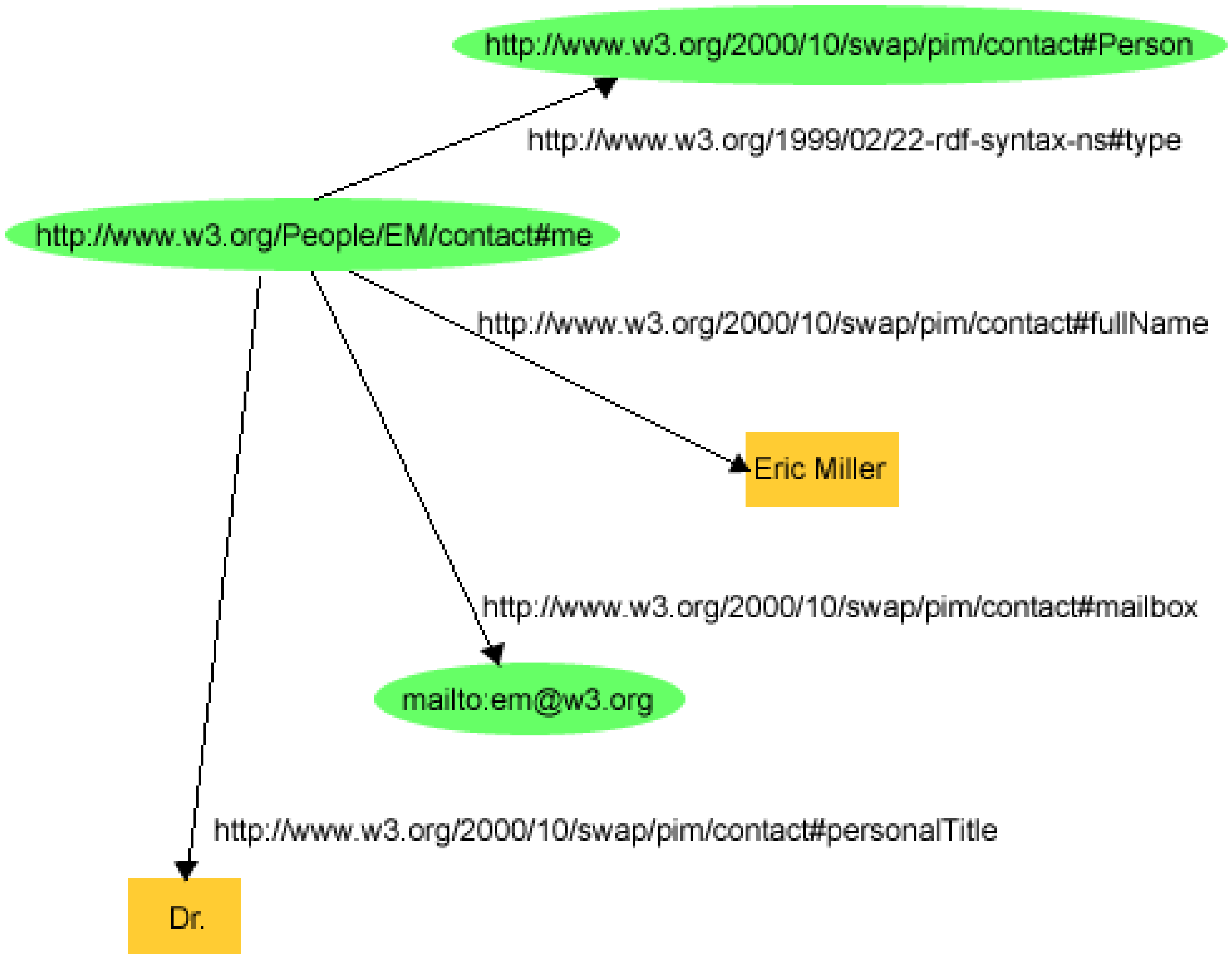
- Resource Description Framework (RDF)
  - Describes things with subject-predicate-object triples
  - Has a standard XML-RDF encoding
- Web Ontology Language (OWL)
  - Defines vocabularies for use in RDF documents
- Ontologies
  - Define classes and properties
  - Ontology design is much like object oriented design

An RDF example

```
- <rdf:RDF>
  - <foaf:Person rdf:about="http://www.w3.org/People/EM/contact#me">
      <rdf:value>Eric Miller, em@w3.org</rdf:value>
      <foaf:name>Eric Miller</foaf:name>
      <foaf:phone rdf:resource="tel:+1-(617)-258-5714"/>
      <foaf:mbox rdf:resource="mailto:em@w3.org"/>
      <foaf:nick>em</foaf:nick>
      <foaf:img rdf:resource="http://www.w3.org/People/EM/s000782.JPG"/>
      <foaf:workInfoHomepage rdf:resource="http://www.w3.org/People/EM"/>
      <foaf:workplaceHomepage rdf:resource="http://www.w3.org/"/>
    - <contact:office>
      - <contact:contactLocation>
          <rdf:value>MIT CSAIL</rdf:value>
          <contact:homePage rdf:resource="http://csail.mit.edu/"/>
        - <contact:address>
          - <contact:Address>
            - <rdf:value>
                The Stata Center, Building 32-G516, 32 Vassar Street, Cambridge MA 02139
              </rdf:value>
              <contact:city>Cambridge</contact:city>
              <contact:country>USA</contact:country>
              <contact:postalCode>02139</contact:postalCode>
            - <contact:street>
                The Stata Center, Building 32-G516, 32 Vassar Street
              </contact:street>
              <loc:coordinates>42.361860,-71.091840</loc:coordinates>
            </contact:Address>
          </contact:address>
        </contact:contactLocation>
      </contact:office>
      <foaf:knows rdf:resource="http://www.w3.org/People/Berners-Lee/card#i"/>
```

Another RDF example

from Wikipedia



Linked Data

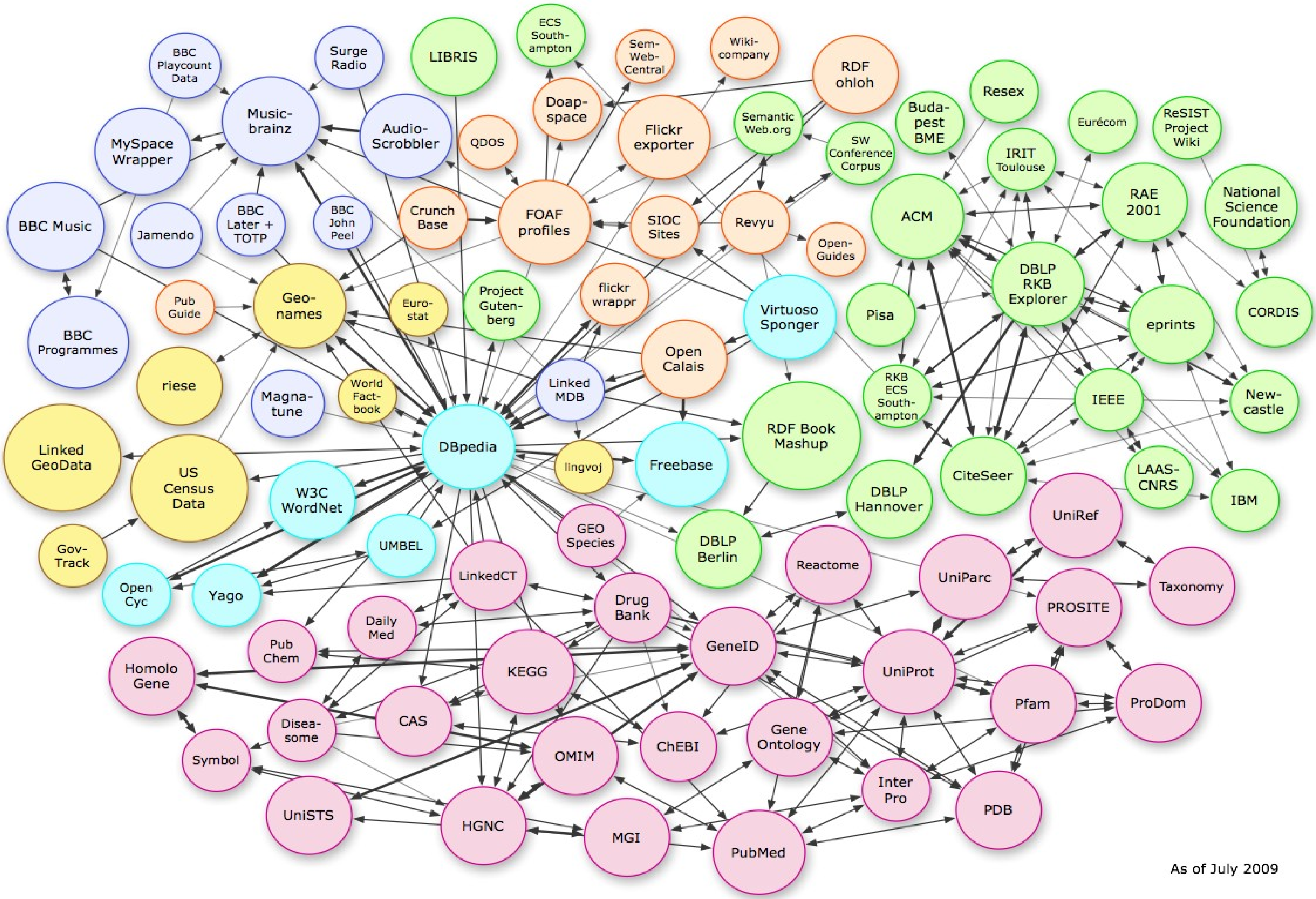
“A term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF.”
– Wikipedia

Linked Data Principles

from Tim Berners-Lee

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs. so that they can discover more things.

The Linked Data Cloud



As of July 2009

The Universal Data Cube

The Universal Data Cube System is a vision for a world wide web in which complex data sets are first class citizens, and rich web-based data visualization and analysis tools are commonplace.

Goals

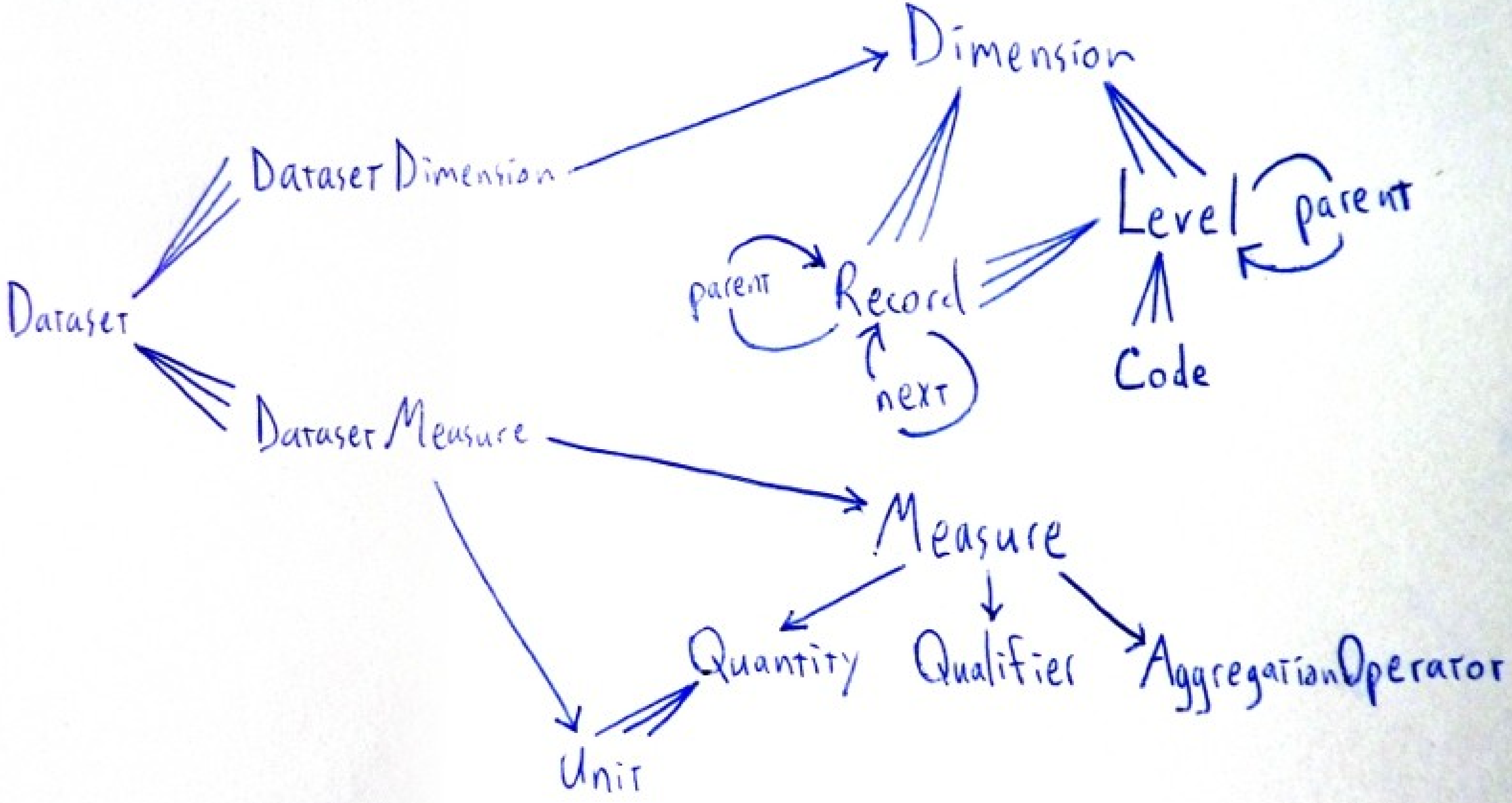
- Design an ontology for hierarchical data cubes
- Develop a system which publishes
  - Data cube metadata in the Linked Data cloud
  - A data cube query endpoint on the web
- Adapt the Weave client to use this system
- Encourage others to build more clients
- Propose it as a web standard for data publishing

The Universal Data Cube Ontology

Competency Questions

that the system must handle

- Show me the average Sepal Width for all iris classes in the Iris dataset (Barchart)
- Show me the average income for the year 2008 for the construction industry for all the US States and the counties of Texas from the BLS dataset. (choropleth map)
- Show me the total wages for top-level industries aggregated across all US States for the years 1990 to 2008 from the BLS dataset (timeseries line chart where lines are industries)


Dimension
DatasetDimension
Level
parent
parent
Record
Dataset
Code
next
DatasetMeasure
Measure
Quantity
Qualifier
AggregationOperator
Unit

Classes

Classes will follow this pattern:

[class name]

 - [property name] [property range type] (multiplicity)

 * [hidden properties internal to the server]

Multiplicity can be one of the following:

 - '1' = exactly one

 - '+' = one or more

 - '*' = zero or more

Dimension

- hasName String (1) //like "Time" or "Space"
- containsRecord Record (*)
  //e.g. "Space" containsRecord "Massachusetts"
- containsLevel Level (*)
  //e.g. "Space" containsLevel "US State"

Level

- hasName String //like "Year" or "State"
- hasNamePlural String //like "Years" or "States"
- hasParentDimension Dimension (1)
  //e.g. "US State" hasParentDimension "Space"
- containsRecord Record (*)
  //e.g. "US State" containsRecord "Massachusetts"
- hasParentLevel Level (0 or 1)
  //e.g. "US State" hasParentLevel "Country"

Record

- hasName String (1)
  //like "1990" or "Massachusetts"
- hasParentDimension Dimension (1)
- hasLevel Level (1)
- hasParentRecord (0 or 1)
- hasNextRecord (0 or 1)

Quantity

- hasName String (1)
  //like "Currency" or "Number of People"
- hasQuantityType String (1)
  //either "Magnitude" or "Multitude"
- containsUnit Unit (*)
  //e.g. "Currency" containsUnit "US Dollars"

Unit

- hasName String (1)
  //like "US Dollars" or "Persons"
- hasParentQuantity Quantity (1)
  //e.g. "US Dollars" hasParentQuantity "Currency"

AggregationOperator

- hasName String (1) //like "Sum" or "Average"

Measure

- hasName String (1)
  //like "Average Income" or "Population"
- hasQuantity Quantity (1)
  //e.g. "Average Income" hasQuantity "Currency"
- hasQualifier String (1)
  //e.g. "Teenage Girls" hasQualifier "People which are female and between age 13 and 19"
- usesAggregationOperator AggregationOperator (1)
  //e.g. "Average Income" usesAggregationOperator "Average"

DatabaseConnection

- hasName String (1) //like "BLS Database"
- containsDatabaseTable DatabaseTable (*)
- (internal) user, pass, host, and port

DatabaseTable

- hasName String (1) //like "Employment"
- hasParentDatabaseConnection DatabaseConnection (1)
- containsColumn DatabaseTableColumn (*)
- (internal) hasSQLName String

DatabaseTableColumn

- hasName String (1)
- hasParentDatabaseTable DatabaseTable (1)

Dataset

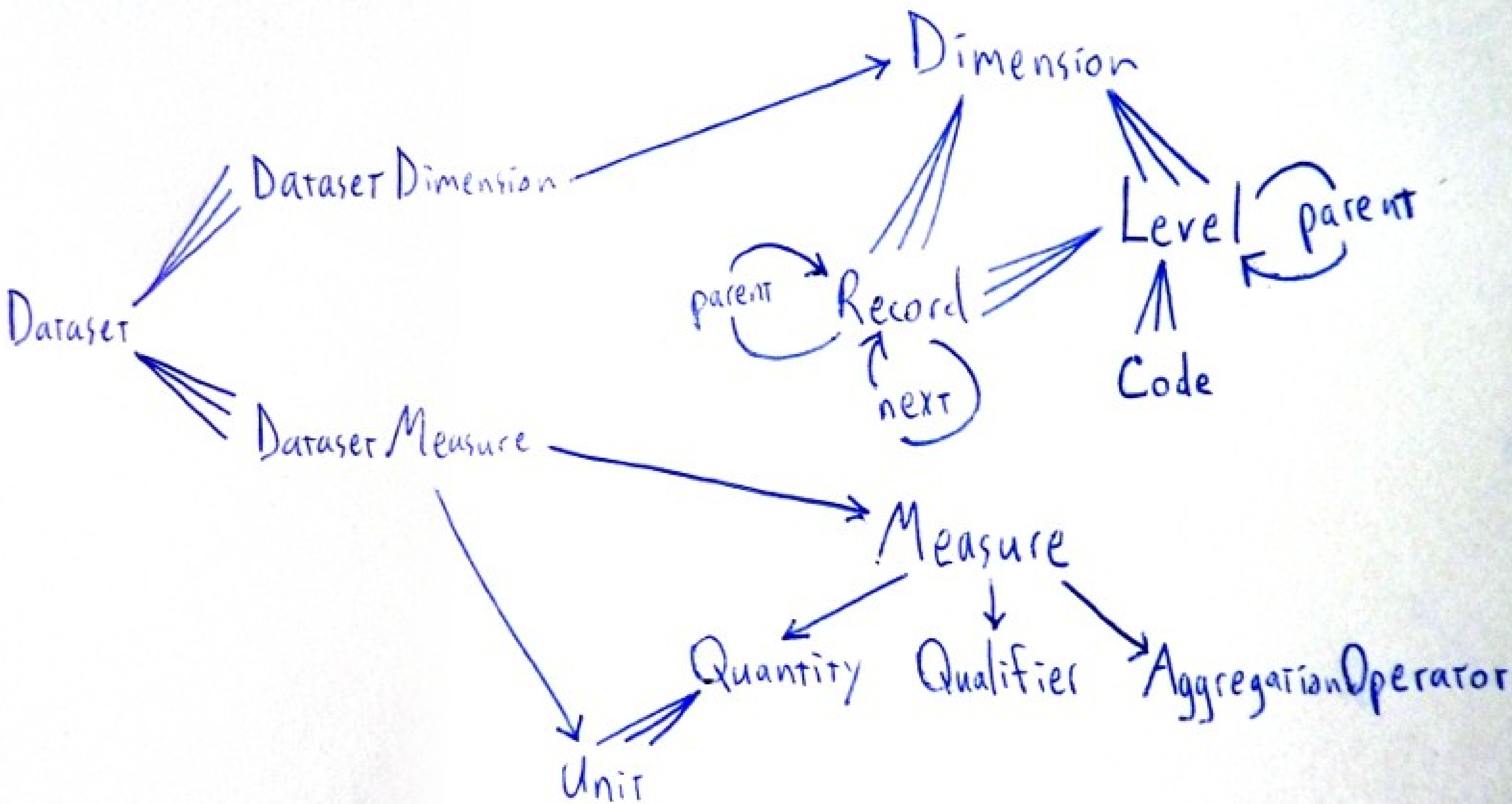
- dc:title String (1) //like "BLS Employment Dataset"
- dc:creator String (0 or 1)
- dc:subject String (0 or 1)
- dc:description String (0 or 1)
- dc:publisher String (0 or 1)
- dc:date String (0 or 1)
- dc:rights String (0 or 1)
- usesDatabaseTable DatabaseTable (*)
- usesDataCubeMapping String (1) //maps data cube metadata to relational tables
- containsDatasetDimension DatasetDimension (*)
- containsDatasetMeasure DatasetMeasure (*)

DatasetDimension

- hasParentDataset Dataset (1)
- representsDimension Dimension (1)
- containsRecord Record (*)
- containsLevel Level (*)

DatasetMeasure

- hasParentDataset Dataset (1)
- representsMeasure Measure (1)
- hasUnit Unit (1)



Dimension
Dataset Dimension
Level
parent
parent
Record
Dataset
next
Code
Dataset Measure
Measure
Quantity
Qualifier
AggregationOperator
Unit

An Example Knowledge Base

for the BLS Employment Dataset

Dimension time = new Dimension
time hasName "Time"

Dimension space = new Dimension
space hasName "Space"

Level year = new Level
year hasName "Year"
year hasNamePlural "Years"
year hasParentDimension time
time containsLevel year

Level usState = new Level
usState hasName "US State"
usState hasNamePlural "US States"
usState hasParentDimension space
space containsLevel usState

Record year1990 = new Record
year1990 hasName "1990"
year1990 hasLevel year
year1990 hasParentDimension time
time containsRecord year1990

Record ma = new Record
ma hasName "Massachusetts"
ma hasLevel usState
ma hasParentDimension space
space containsRecord ma

Quantity currency = new Quantity
currency hasName "Currency"

Quantity numPeople = new Quantity
numPeople hasName "Number of People"

Unit usDollars = new Unit
usDollars hasName "US Dollars"
usDollars hasParentQuantity currency
currency containsUnit usDollars

Unit persons = new Unit
persons hasName "Persons"
persons hasParentQuantity numPeople
numPeople containsUnit persons

Measure avgIncome = new Measure
avgIncome hasName "Average Income"
avgIncome hasQuantity currency

Measure population = new Measure
population hasName "Population
population hasQuantity numPeople

DatabaseConnection blsDatabase = new DatabaseConnection
blsDatabase hasName "Bureau of Labor Statistics Database"

DatabaseTable bls2008 = new DatabaseTable
blsTable hasName "bls2008"

blsTable hasColumn "Average Income"
blsTable hasColumn "Total Wages"
blsTable hasColumn "Employment
blsTable hasColumn "Average Income"
blsTable hasColumn "Population"

blsTable hasParentDatabaseConnection blsDatabase
blsDatabase containsDatabaseTable blsTable

Dataset blsDataset = new Dataset
blsDataset hasName "Bureau of Labor Statistics Employment Dataset"
blsDataset usesDatabaseTable blsTable

DatasetDimension blsTimeDimension = new DatasetDimension
blsTimeDimension representsDimension time
blsTimeDimension hasParentDataset blsDataset
blsDataset containsDatasetDimension blsTimeDimension

DatasetRecord bls1990 = new DatasetRecord
bls1990 representsRecord year1990
bls1990 hasParentDatasetDimension blsTimeDimension
blsTimeDimension containsDatasetRecord bls1990

DatasetRecord blsMA = new DatasetRecord
blsMA representsRecord ma
blsMA hasParentDatasetDimension blsSpaceDimension
blsSpaceDimension containsDatasetRecord blsMA

DatasetMeasure blsPopulation
blsPopulation representsMeasure population
blsPopulation hasUnit persons
blsPopulation hasParentDataset blsDataset
blsDataset containsDatasetMeasure blsPopulation

DatasetMeasure blsAvgIncome
blsAvgIncome representsMeasure avgIncome
blsAvgIncome hasUnit usDollars
blsAvgIncome hasParentDataset blsDataset
blsDataset containsDatasetMeasure blsAvgIncome

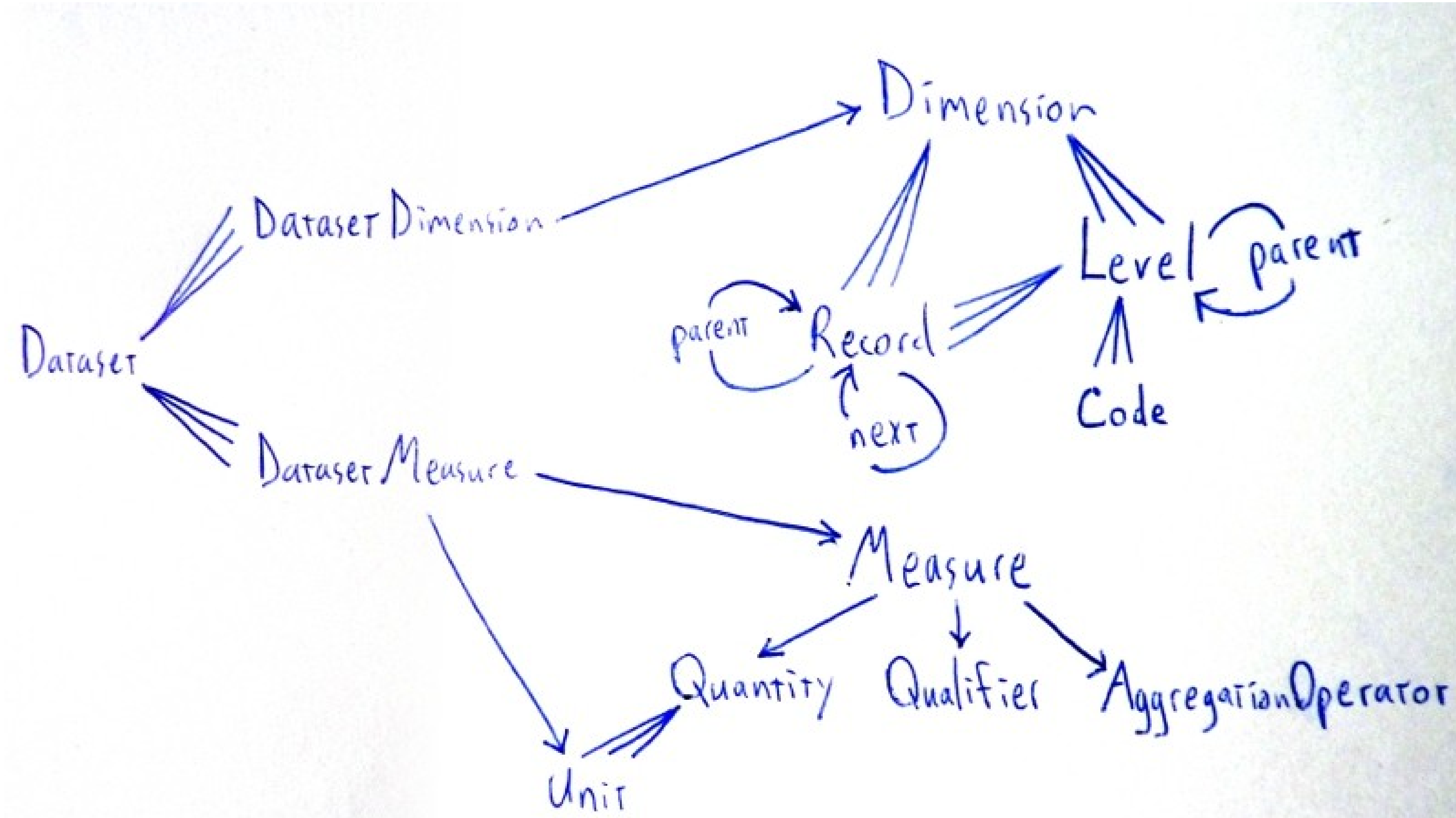
Weave Data Model Problems

- Hierarchical key types are not linked
  - US Counties and US States are totally independent
- Key types referring to the same things not linked
  - Like US State codes and US State abbreviations
- Columns representing the same measure with different units are not compatible
  - Population in thousands not comparable with Population in millions
- No way of resolving when two datasets provide comparable columns

Weave Data Model Solutions

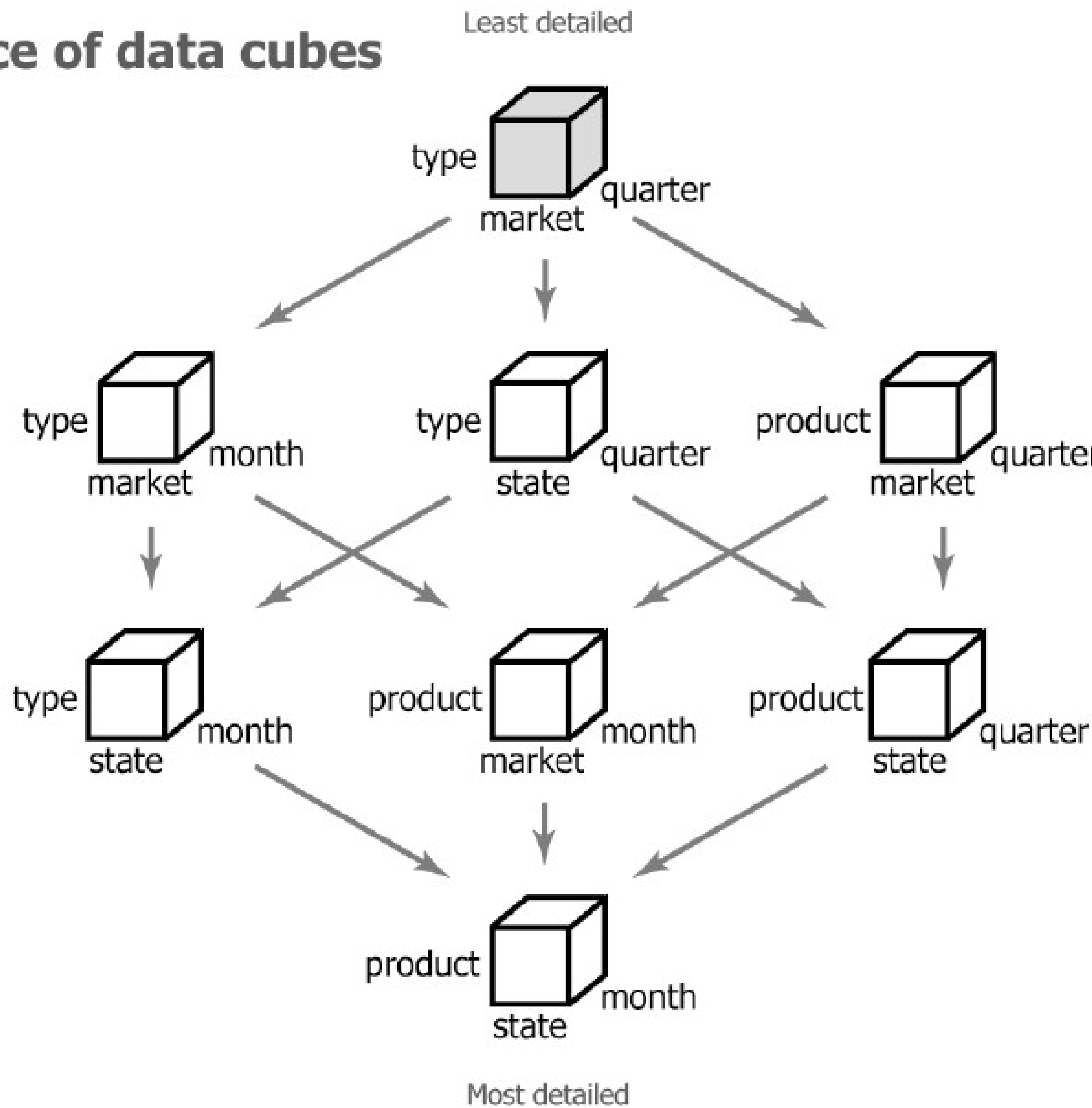
- Hierarchical key types are linked
  - Via the data cube dimension hierarchy structure
- Key types referring to the same things are linked
  - US State codes and US State abbreviations are different RecordCodes for the same record set
- Columns representing the same measure with different units are compatible
  - Population in thousands and Population in millions are two different Units within the same Quantity
- Resolving when two datasets provide comparable columns is possible
  - Because Datasets use universal Measure URIs to describe their contents

The end.



Dimension
Dataset Dimension
Level
parent
parent
Record
Dataset
next
Code
Dataset Measure
Measure
Quantity
Qualifier
AggregationOperator
Unit

The lattice of data cubes



Projecting a three dimensional data cube