

Selecting among Categories

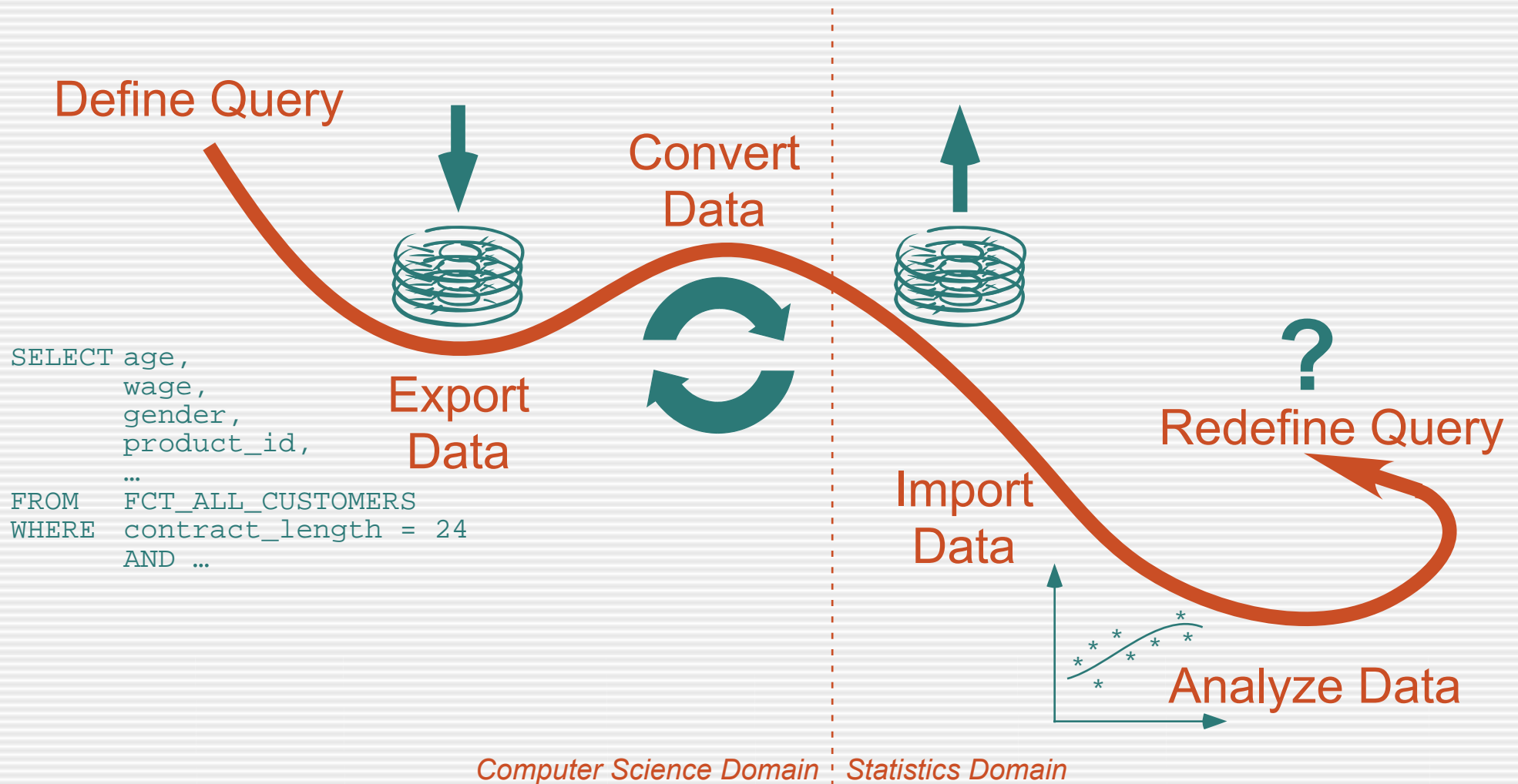
**Interactive Statistical Graphics
working on Databases**

martin.theus@math.uni-augsburg.de

- 1. The Curse of Flat Files**
- 2. Isn't the World categorical anyway?**
- 3. Talking to Databases**
- 4. Making Graphs work on categorical data**
- 5. Two Level Data Access**
- 6. Implementation**

The Curse of Flat Files

Statistics software usually does not feel comfortable with databases, but data does ...



Isn't the World categorical anyway?

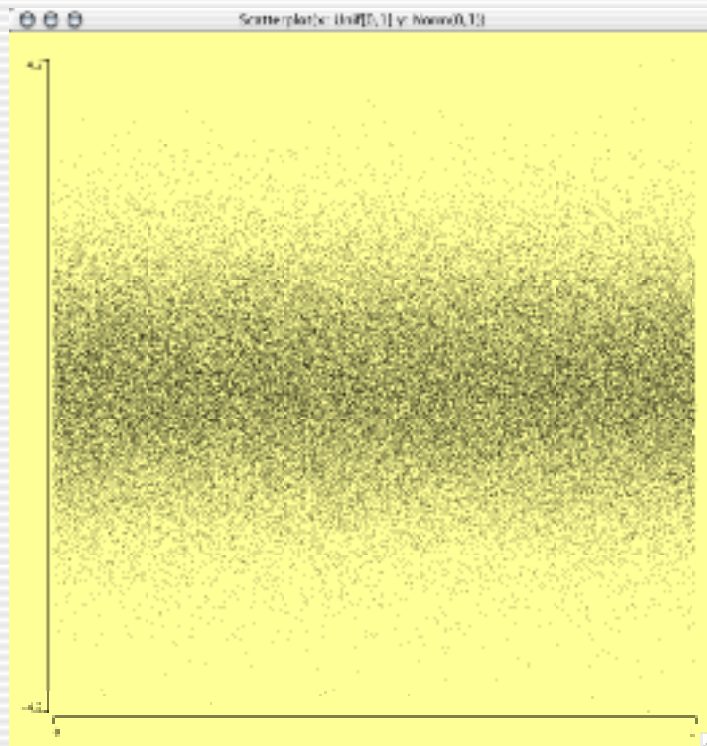
- **Most variables are measured at only a very limited resolution (e.g. Age, ...)**
- **Census data are mostly categorical by nature**
- **Attribute data are categorical by definition**
- **Classical Star Model in database design assumes categorical data as well**
- **US Census Data from UCI: 23 Vars, 18 categorical**
- **But: beware of careless discretizations**

- **Database connectivity is not very wide spread among statistics software**
- **“Getting the data out” is usually not enough**
→ **efficiency problems persist**
- **For efficiency reasons: “Get the data out you need”**
- **For programmers only:**
ODBC, JDBC are fast enough
- **Consistency problems arise:**
“Has the data changed since I performed the analysis?”

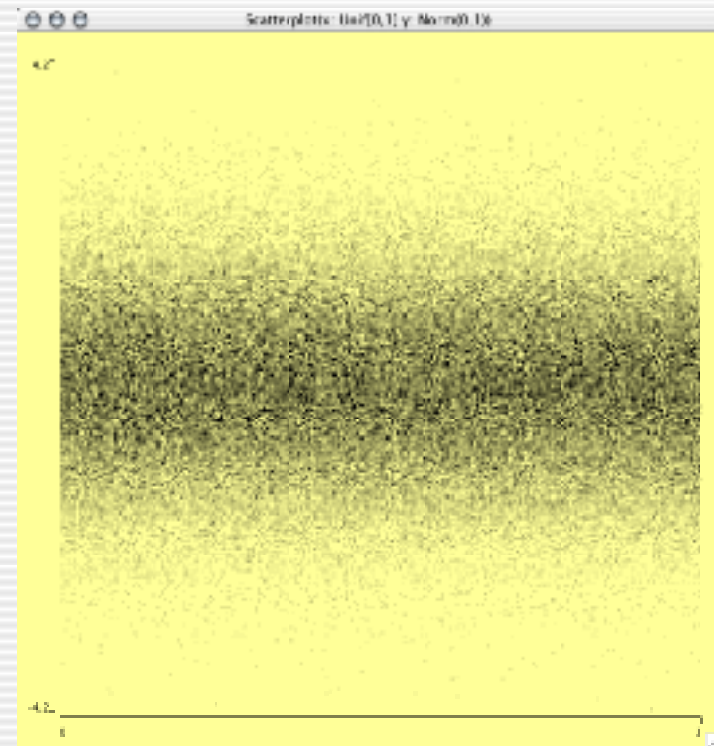
- **Graphs for categorical data work fine with DBs**
- **Only summaries of the data are needed to plot:**
 - **Barcharts / Spineplots**
 - **Mosaic Plots**
 - **Histograms / Spinograms**
 - **(Boxplots)**
- **What about graphs, which plot a single glyph for each observation?**

Making Graphs Categorical

- **Example: Scatterplot vs. Binning Plot**
100.000 $U(0,1)$ vs. $N(0,1)$ random numbers



raw scatterplot



binned data 256x256

- **Idea: Only access summaries as long as we look at the complete data set**
- **Use categorical displays to select subsets (Barcharts, Mosaics, Histograms, Binned Plots)**
- **If a subset is “small”, any plot can hold the data of the subset. (“Hot Selection Sets“)**
- **The definition of “small” depends strongly on the**
 - **speed of the database system**
 - **speed of the graphics system**
 - **the plot displayed**

Implementation in Mondrian

- **Database connection dialog:**

DB Connection

Driver: org.gjt.mm.mysql.Driver

URL: jdbc:mysql://jetta.math.uni-augsburg.de:3306/mysql

User: theusm Pwd: *****

DB: datasets

Table: families

- **Information needed:**
 - database server
 - user
 - password
 - DB – instance
 - table

- **Statistics software must not ignore data in databases**
- **Simply exchanging the flat file with the database is not efficient**
- **Efficiency problems can be rolled out to the DB**
- **Categorical data can easily be handled in databases**
- **Categorical displays work fine on databases**
- **Implementation in MONDRIAN available**
- **“Thanks for your attention!”**