

OUTLINE

EDA Why did Exploratory Data Analysis emerge?

What is EDA about?

EDA thinking and EDA workflow

**Technologies for
Data Exploration**

Interactive Techniques Selection and Linked Highlighting

Reordering

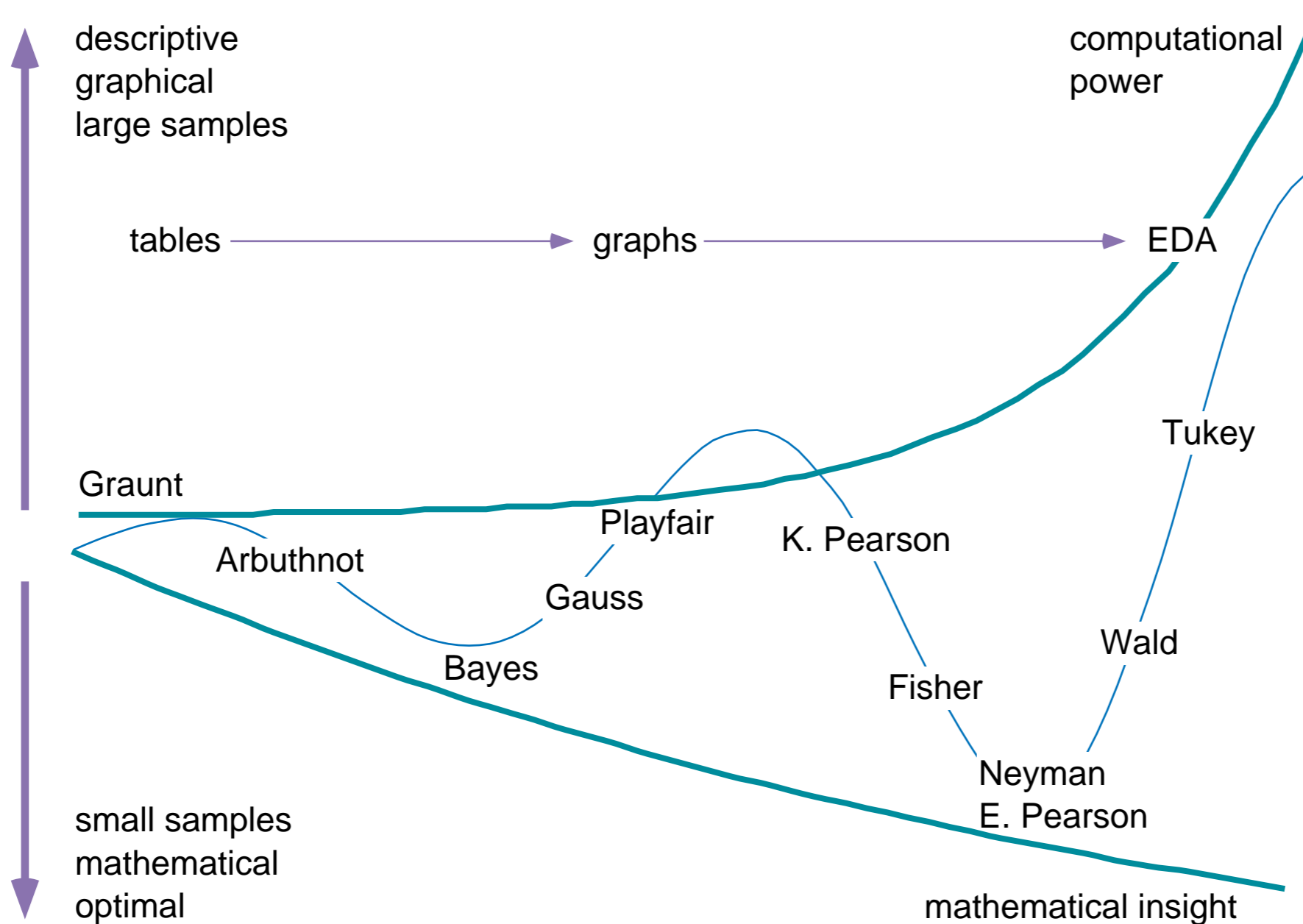
Re-Expression

Interrogation

Development Environments

Conclusions

EDA Why did Exploratory Data Analysis emerge? A historical view on Statistics



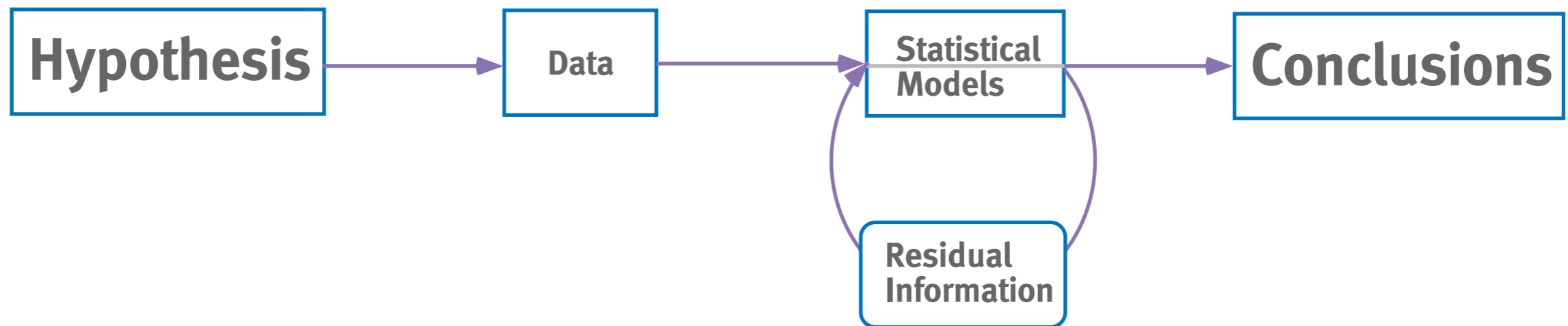
EDA What is Exploratory Data Analysis about? Looking into Tukey's 1977 book

- SCRATCHING DOWN NUMBERS (stem-and-leaf)
- SCHEMATIC SUMMARIES (pictures and numbers)
- EASY RE-EXPRESSION
- EFFECTIVE COMPARISON (including well-chosen expression)
- ...
- SHAPES OF DISTRIBUTION
- MATHEMATICAL DISTRIBUTIONS
- POSTSCRIPT

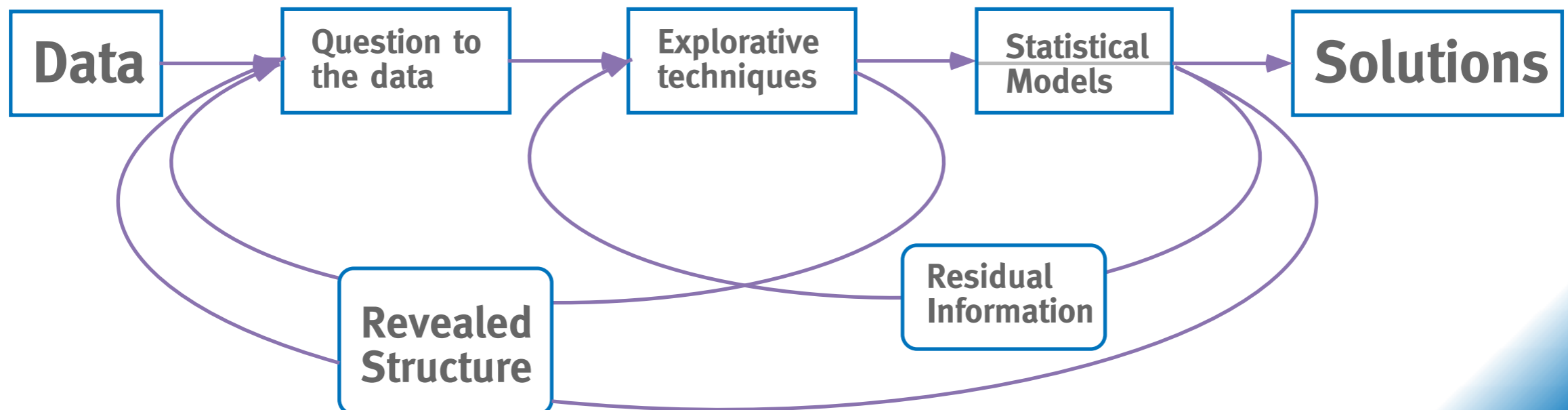
A Our relationship to the computer

EDA EDA thinking and EDA workflow About Hypothesis and Iteration

Classical



Exploratory

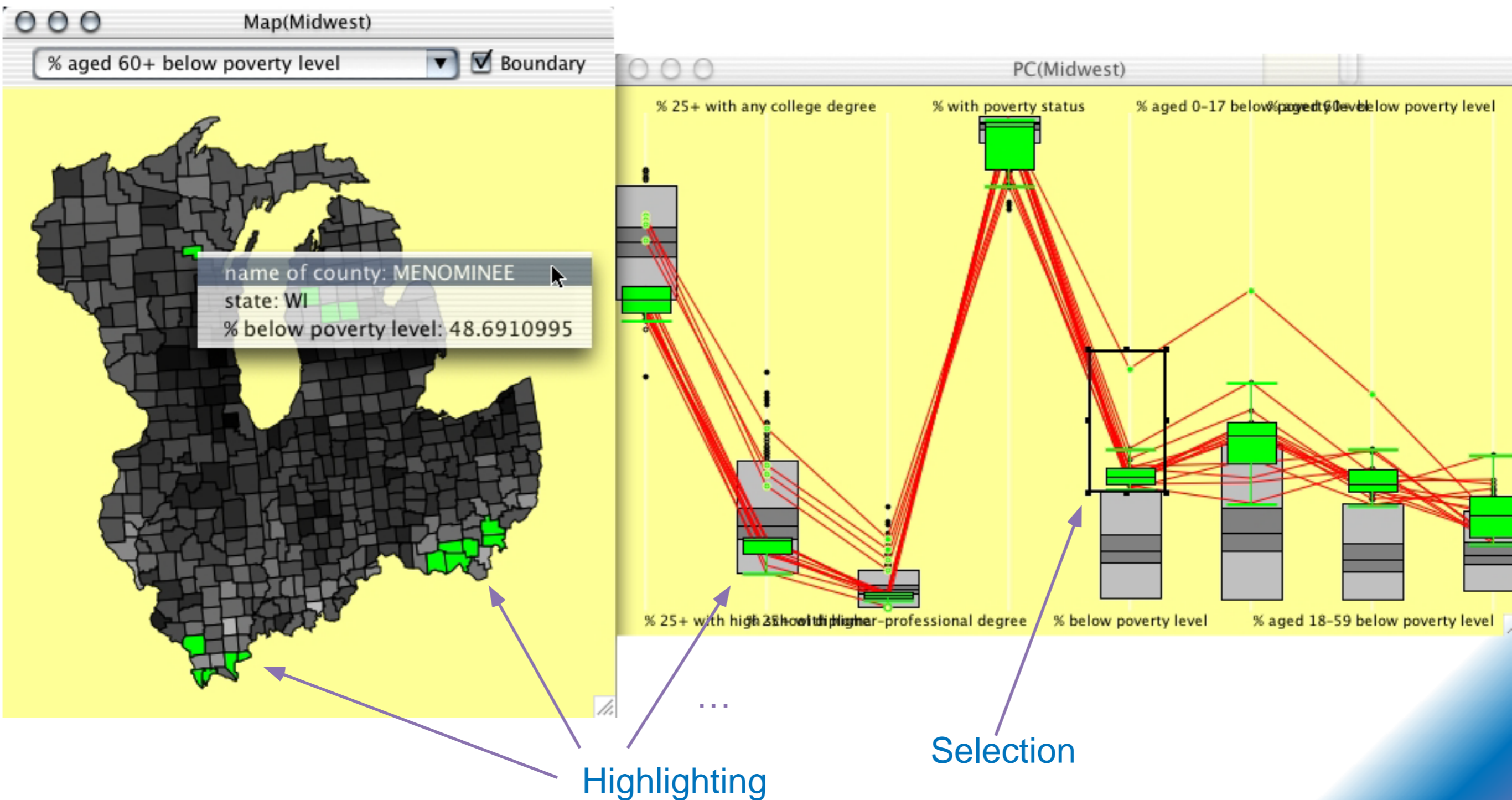


Technologies for Data Exploration

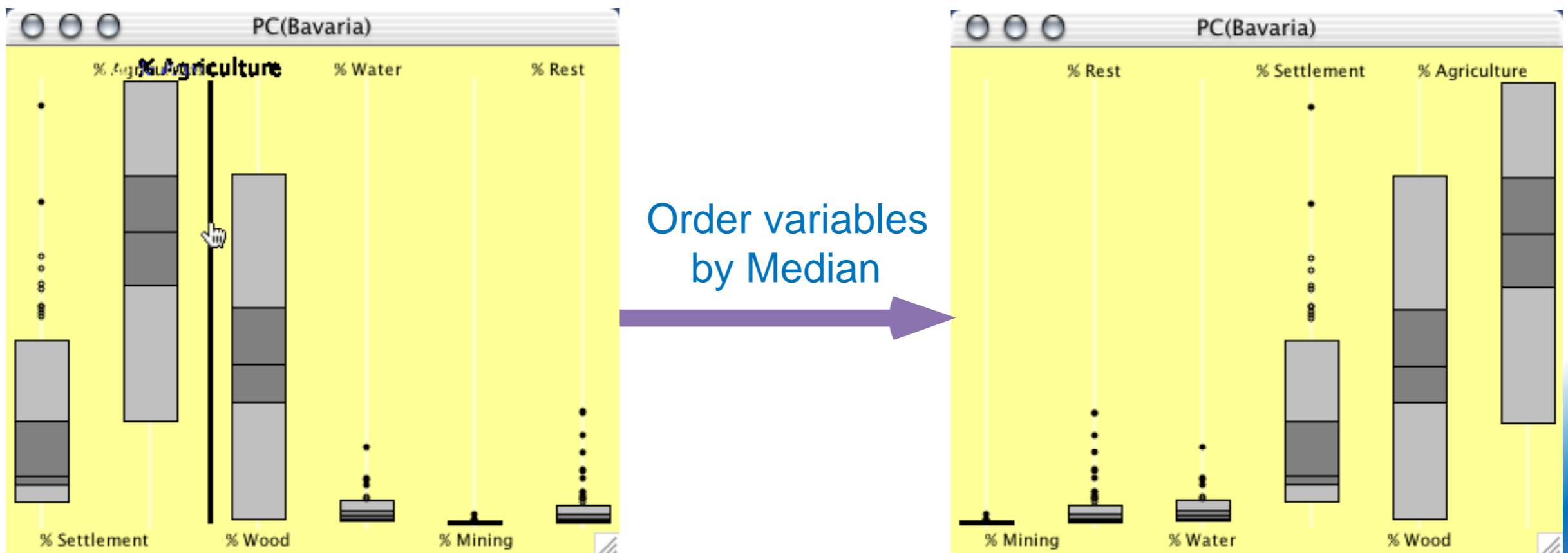
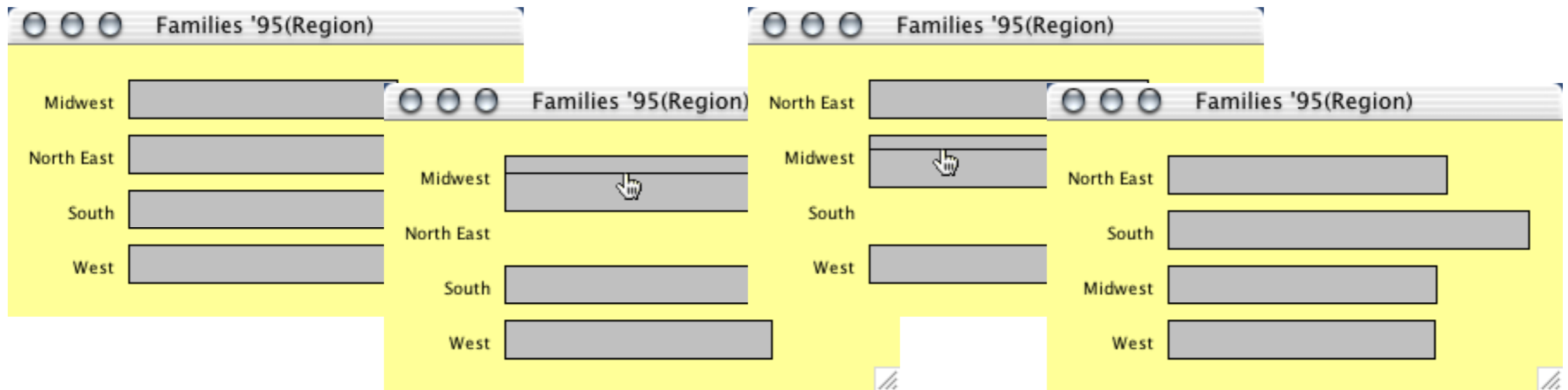
A GIS View

| | Statistics | Cartography | Data(base) Management |
|-----------------------|------------|-------------|-----------------------|
| 1. Data Input | | | ● |
| 2. Data Manipulation | | ● | ○ |
| 3. Data Management | | | ● |
| 4. Query and Analysis | ● | | ○ |
| 5. Visualization | ● | ● | |

Interactive Techniques Selection and Linked Highlighting

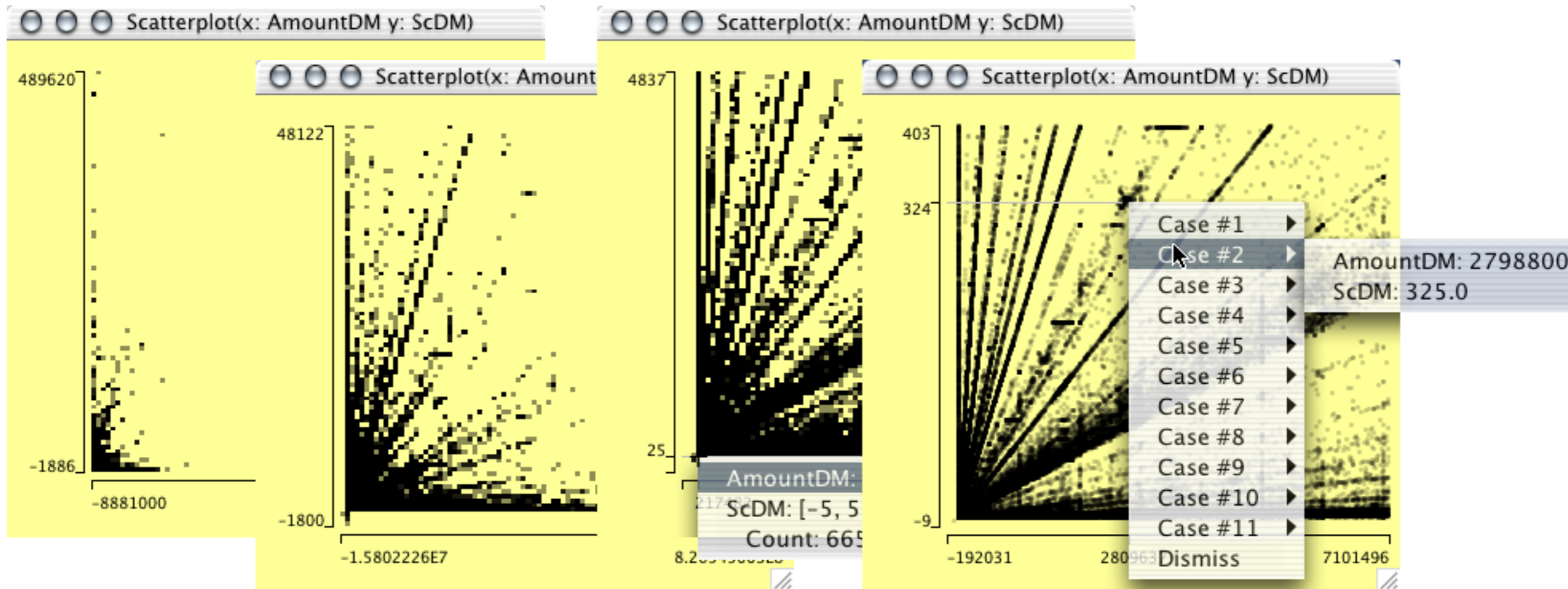


Interactive Techniques Reordering



Interactive Techniques Re-Expression

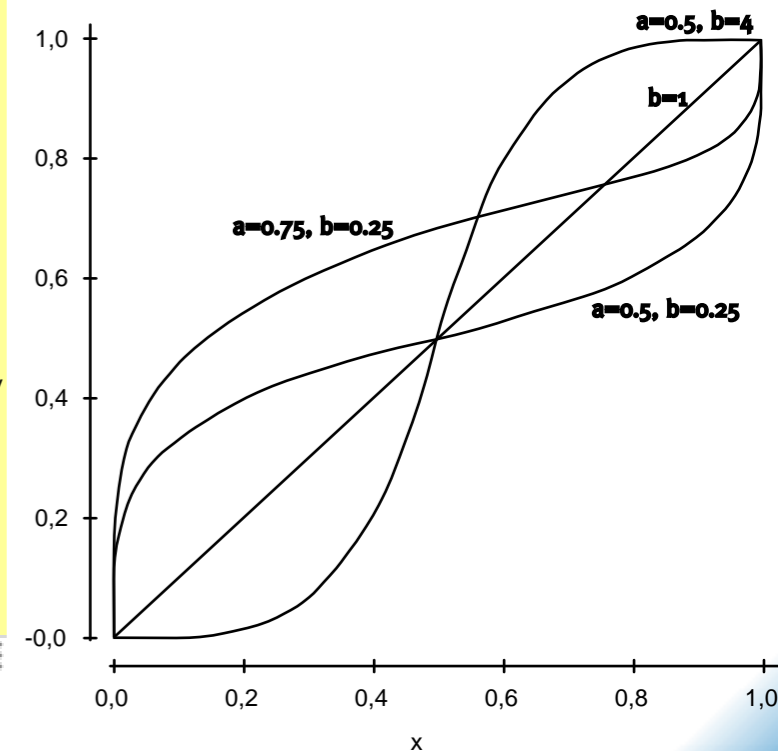
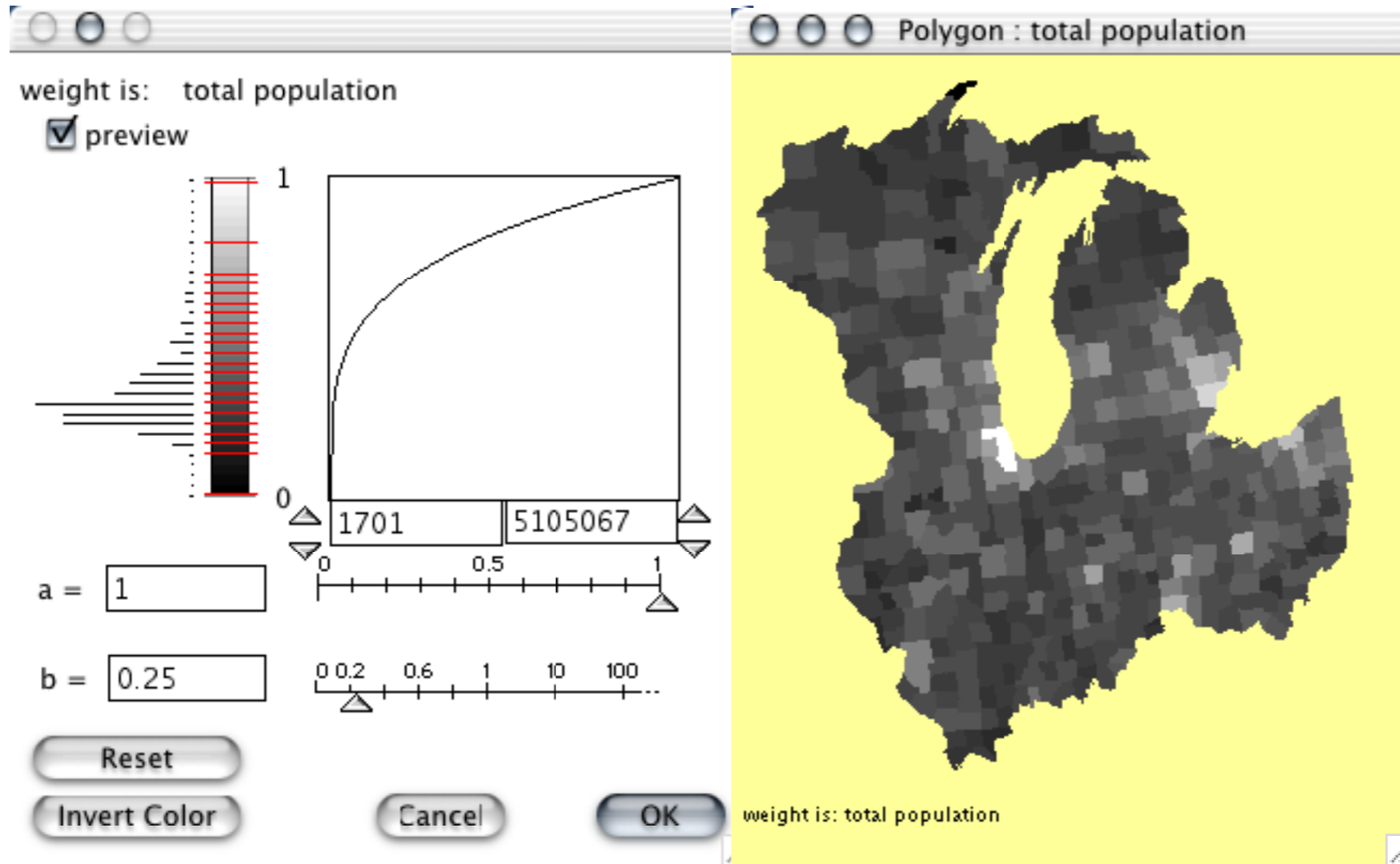
- Zooming (Example: Scatterplots)



- Logical Zooming: Display and Aggregation change
- Applicable to almost all visualizations
- Implementation is hard!

Interactive Techniques Re-Expression

- Transformation



- The formula

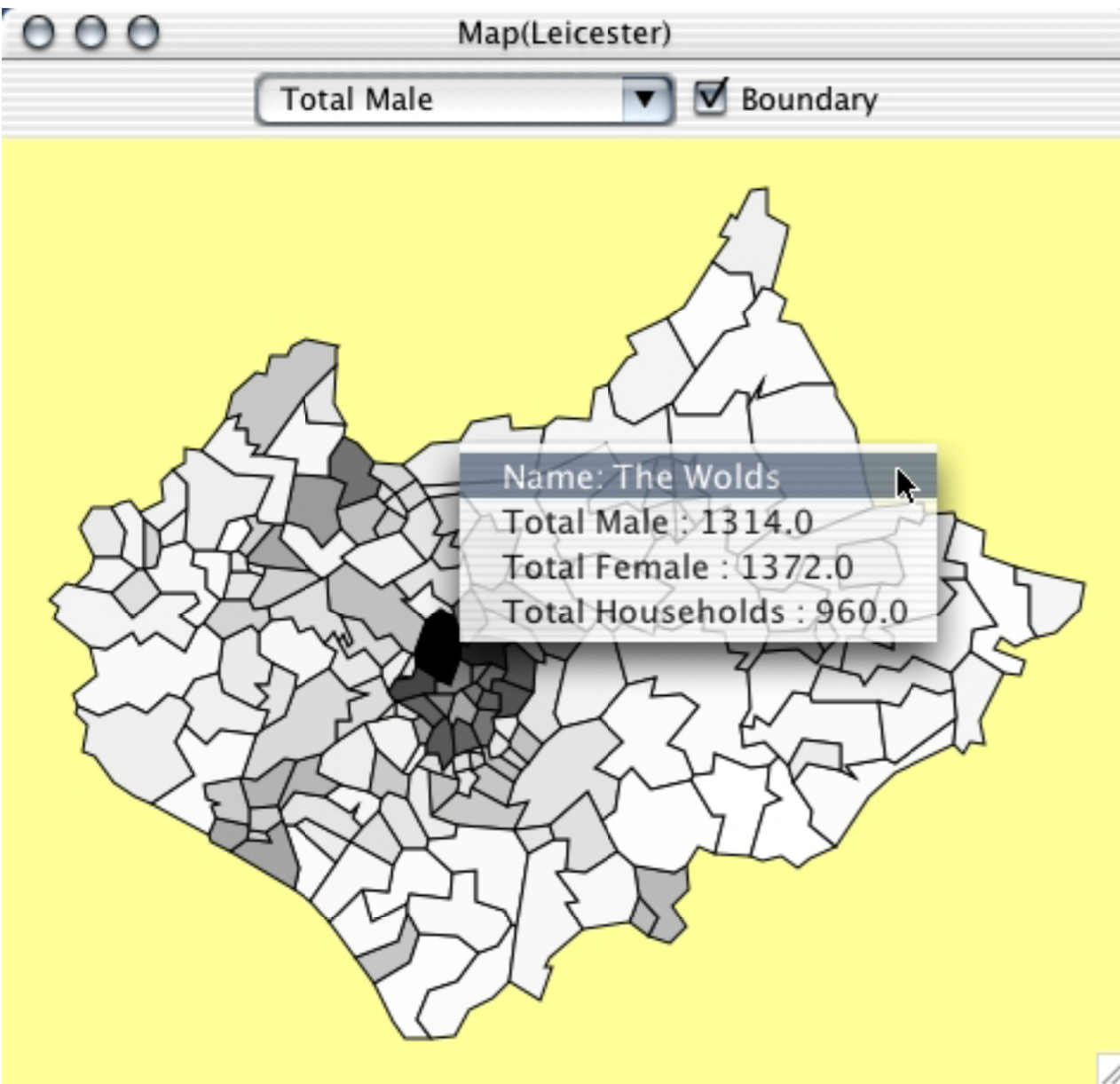
$$f(x) := \begin{cases} a \cdot \left(\frac{x}{a}\right)^b & \text{for } x \leq a \\ 1 - (1 - a) \cdot \left(\frac{1-x}{1-a}\right)^b & \text{for } x \geq a \end{cases} \quad \text{for } a \in [0, 1] \text{ and } b > 0$$

Interactive Techniques Interrogations/Queries

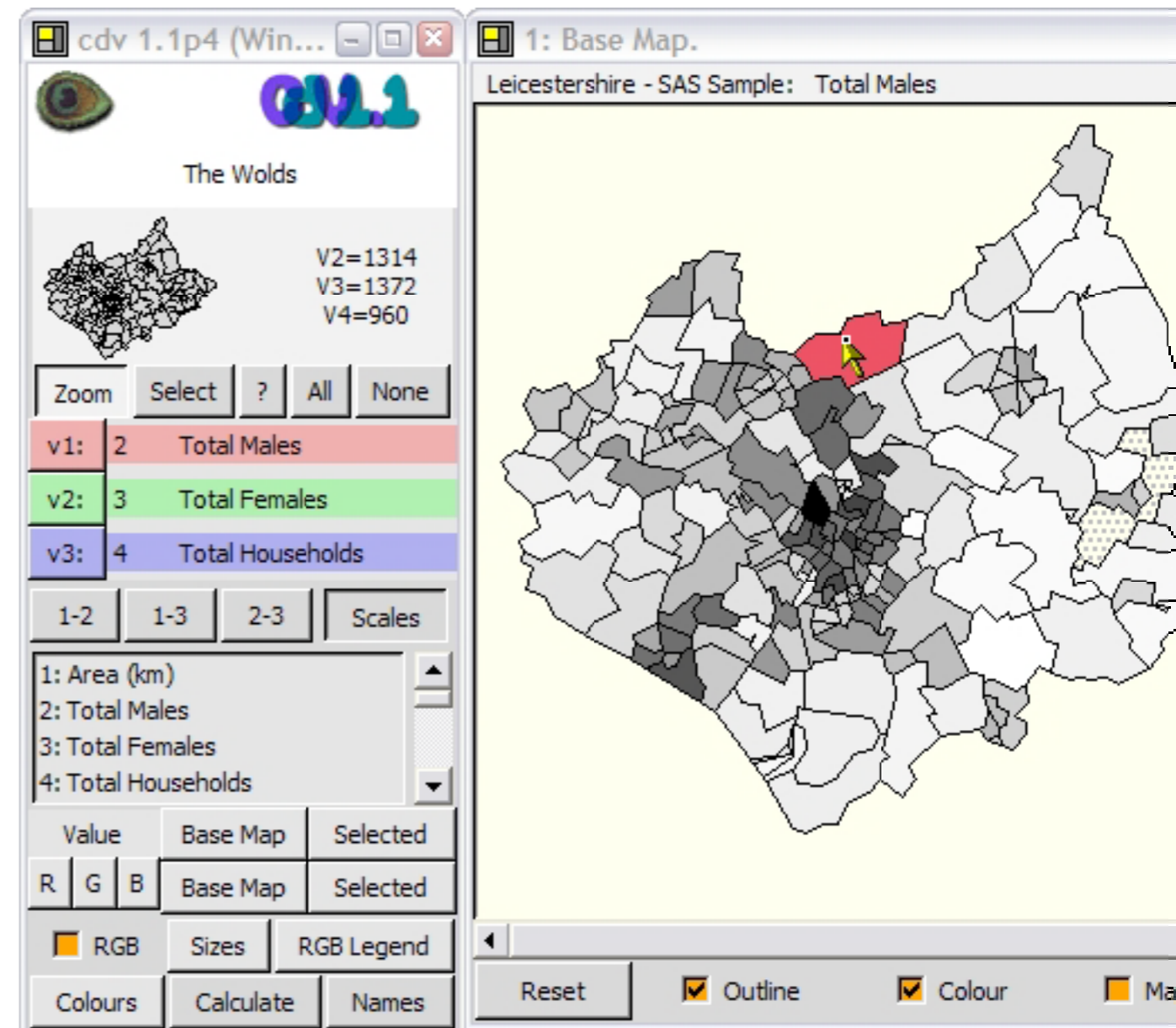
- Level of Queries
 - overview of coordinates (object independent)
 - (coordinate-) values of object that is queried
 - values of any variable of object that is queried
 - arbitrary values related to the queried object
- Closeness of Queries
 - contiguous queries:
 - Information is displayed at query position
 - distant queries:
 - Information is displayed in a separate info window

Interactive Techniques Interrogations/Queries

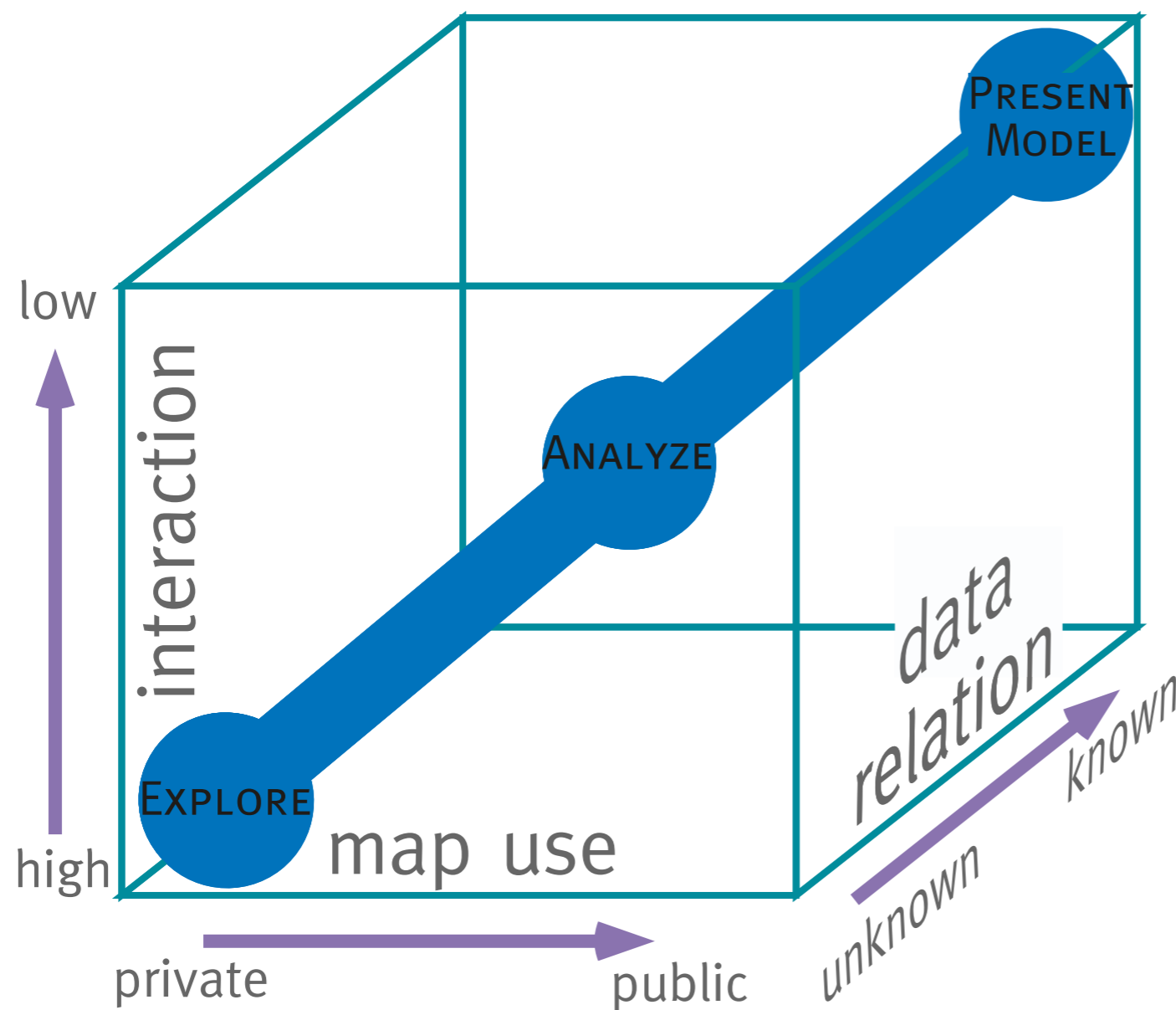
- Continuous (Mondrian)



- Distant (cdv)



Development Environment “one size fits all?”



Conclusions

- EDA is an iterative and adaptive process
- Interactive Statistical Graphics is the most powerful tool for EDA
- Interactive Statistical Graphics generalizes easily to GIS apps
- Direct Manipulation Interfaces are as important as expensive
- No single system can offer all functionality in an analysis
- Some Systems are around, but there is room for improvement
- *Thanks for you attention!*