

The Intelligent Data Analysis System for Social Science

**- Incorporating Object-oriented
and Knowledge-based approaches**

- Alex Liu, Ph.D.
- Director
- Research Methods Institute
- Los Angeles, CA, USA
- in <http://www.researchmethods.org/ida.pdf>
- © 2004 Research Methods Institute

Purposes

- Go beyond “cookbook fallacy”
- Assess the impacts of computer technology on data analysis in social science
- Review technologies of modern data analysis - “Intelligent Data Analysis IDA”
- Plan for an IDA system for social science

About Data Analysis

- Data analysis is a highly complex activity.
- Data analysis is a repetitive, cyclical search for understanding data.
- (iterative process)

Computer's Impacts (1)

- More datasets to analyze
- More information to process
- Sizes & types of data changed
- For example, besides survey data, news reports, and information collected from eGovernment, online voting, ...

Computer's Impacts (2)

- A lot more new tools available
- -- expert systems, bootstrap, simulation, genetic algorithms, neural nets

Computer's Impacts (3)

- “computers do not crunch numbers, they manipulate symbols --- Margaret Boden
- Boden, M. A. 1977 *Artificial Intelligence and Natural Man*. Hassocks: Harvester Press
- Data can be symbols, graphs, ...
- Anything in electronic format can be analyzed.

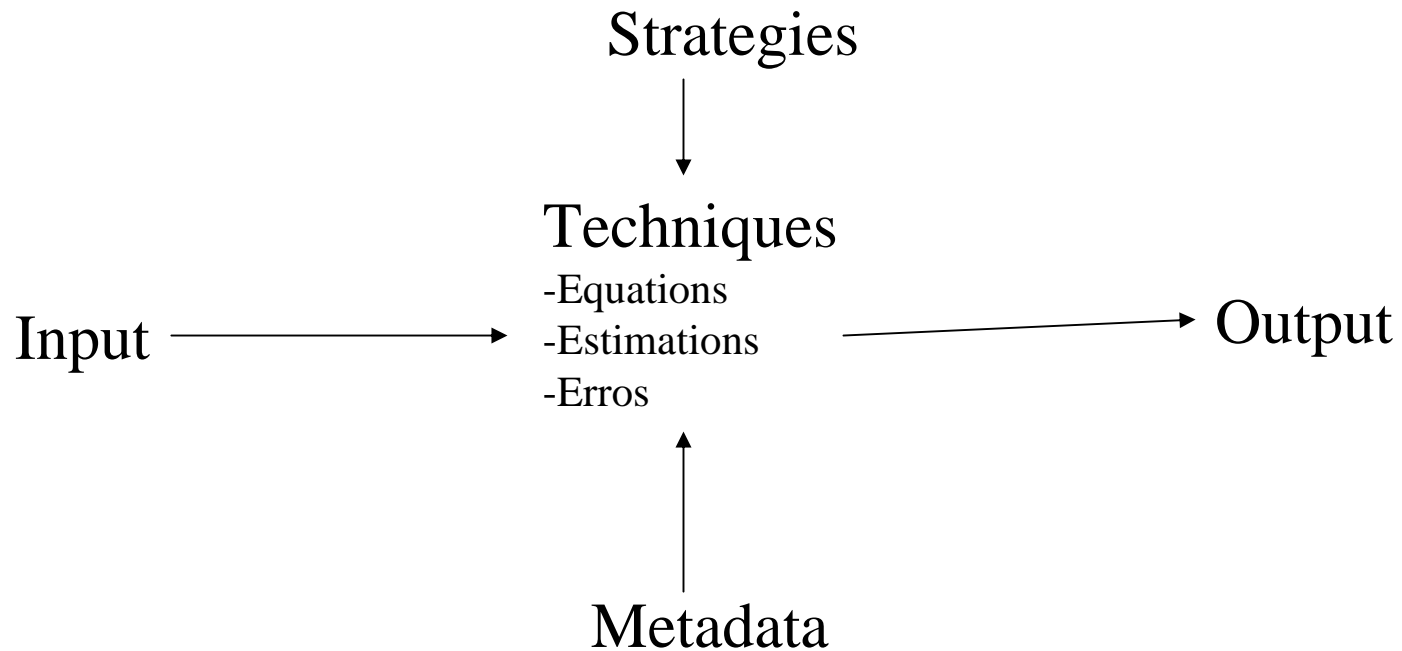
Key Impacts

- Computer frees us from the need to focus on the manipulative details of data analysis and allow us to examine higher level issues of problem formulation.
- Most statistics users already take an object-oriented view.

What is happening already?

- With SPSS, SAS, STATA, S-plus
- Users do not necessarily understand the numeric and algebraic details.
- Users describe their research in terms of this SPSS language
- INPUTS – **BOX** - OUTPUTS

The Working Structure



Higher Level Thinking

- Research Goals (understanding, prediction, ...)
- Types of data – designs, sample sizes, populations, error structures
- Data quality control
- Methods selection
- Model selection

- Strategies

Higher Level Issues (1)

- Statistical Strategies
- - to summarize the data analysis process
- - steps, decisions and actions taken during the process of analyzing data
- “good strategy” as the hallmark of IDA

Higher Level Issues (II)

- Problem formulation \leftrightarrow problem-solving model
- Key application factors identified in the formulation of problem
- General characteristics of the model under consideration
- MAPPING model to problem

Higher Level Issues (III)

- Methods selection
- So many methods to select
- Possible to define an “optimal” or near-optimal one?

Higher Level Issues (IV)

- What are the important factors in choosing a method for problem solving?
- - prediction accuracy, computational efficiency, interpretability
- - quality of model, engineering considerations, available resources, logistical constraints
- MAPPING methods with problems

Higher Level Issues (V)

- Data quality
- Accuracy, completeness, consistency, timeliness
- To handle outliers

Needs from analysts

- Ways to incorporate various datasets
- Ways to incorporate various methods
- Ways to work on higher level issues

Future Work for IDA

- 1) New Data Analysis Training Methods
- 2) Knowledge-based IDA Systems
- 3) Further Research on IDA

1) New Training

- Emphasize on strategies rather than on techniques
- Cookbook fallacy
- Cognitive Research on the Impacts of IDA
- (F. Young 2003)

2) Developing knowledge systems

- Systems to handle strategies
- Systems to map models with problems
- Systems to map methods with problems
- Systems on data quality control
- and
- A system to integrate everything

Formal representation of statistical strategies

- 1) data-analysis procedures to accomplish a specified data analysis task,
- 2) data-analysis actions (choices, decisions, etc.)
- 3) the interactions between procedures and actions to accomplish the data-analysis task

Mapping Systems

- Map models to problems
- Map methods to problems

Data Quality Control Systems

- Outliers

the RM Integrated System

- Users will focus on high level thinking
- Multi-methods
- Multi-datasets
- Knowledge-based
- Multi-output



Data
Object

Method
Object

Model
Object

Prototype - components

- Knowledge representation
- Inference engine
- Data mining tools
- Multi data inputs
- Multi outputs

Modern
User Interface

More on the system

- Users decide on higher level issues and supply datasets (data object)
- The system produces the model object

