

# **A Cognitive Model of Data Analysis and its Implementation as a Human-Computer Interface**

**Forrest W. Young**

**L.L. Thurstone Psychometric Laboratory  
University of North Carolina  
Chapel Hill, North Carolina**

Overheads for an invited presentation at the annual meeting of the American Psychological Association, held in Toronto, Canada, Aug 9-13, 1996.

These slides, and the ViSta system are available at no cost from: <http://forrest.psych.unc.edu/>

## **Outline**

- 1. Statistical Strategy**
- 2. Cognitive Modes**
- 3. Software Modes**
- 4. ViSta - The Visual Statistics System**  
**WorkMaps, GuideMaps, SpreadPlots**
- 5. Individual Differences and Data Analysis**
- 6. Proposed Empirical Usability Study**
- 7. Conclusion**

## 1: Statistical Strategy

- [Forrest Young, John Smith & David Lubinsky](#) have proposed a cognitive model for data analysis. It is discussed in:  
 Young, F.W. & Smith, J.B. (1991) *Towards a Structured Data Analysis Environment: A Cognition-Based Design*. In: Buja, A. & Tukey, P.A. (Eds.) *Computing and Graphics in Statistics*. 36, 253-279. New York: Springer-Verlag.
- Young, F.W. & Lubinsky, D.J. (1995) *Guiding Data Analysts with Visual Statistical Strategies*. J. Comp. and Graphical Statistics. 4: pp. 229-250.
- We present a definition of [statistical strategy](#): A statistical strategy is a formal representation of an expert statistician's conceptual structuring of:
  1. the [data analysis procedures](#) needed to accomplish a specified data analysis task;
  2. the [data analyst's actions](#) (choices, decisions, etc.) that are possible with the data analysis procedures;
  3. and the [relationships](#) between the procedures and actions needed to accomplish the data analysis task.
  4. Note that "the [data analysis task](#)" is to understand a specified data analysis object (data set or data model).

## 2: Cognitive Modes

- [Cognitive Modes](#): Our model involves cognitive modes that differ in terms of the statistical strategies involved.
- The Cognitive Modes are
  1. [Structural](#) — the cognitive mode that is active when the analyst is constructing, maintaining or revising the data analysis.
  2. [Exploratory](#) — the cognitive mode that is active when the analyst is exploring the data and generating hypotheses.
  3. [Confirmatory](#) — the cognitive mode that is active when the analyst is confirming hypotheses.
- Note that
  1. The analyst's cognitive mode [changes](#) according to what the analyst is doing during the data analysis.
  2. The [profile](#) of cognitive modes depends on the analyst's level of expertise.

### 3: Software Modes

**Software Modes:** The data-analysis environment should have visible software modes (windows) to support the different cognitive modes.

- Each software mode is tailored to the requirements of the cognitive mode.
- A key aspect of our work is that the data-analysis environment should change to reflect the analyst's changing [cognitive modes](#).

Software Modes and Cognitive Modes:

1. [WorkMaps](#) & [GuideMaps](#) support the [structure](#) cognitive mode.
2. [SpreadPlots](#) support the [exploratory](#) cognitive mode.
3. [Command Lines](#) and [alphanumeric reports](#) support the [confirmatory](#) cognitive mode.

### 4: ViSta: The Visual Statistics System

#### [Implementation:](#)

Our cognitive model of data analysis is implemented in [ViSta](#):

Young, F.W. (1966) *ViSta, The Visual Statistics System*. L.L. Thurstone Psychometric Laboratory Research Memorandum 94-1(b), Univ. N. Carolina, Chapel Hill, NC.

#### [ViSta runs under](#)

- MS-Windows (3.1 & 95)
- MacOS (68040, PowerPC)
- Unix with X11

#### [Availability:](#)

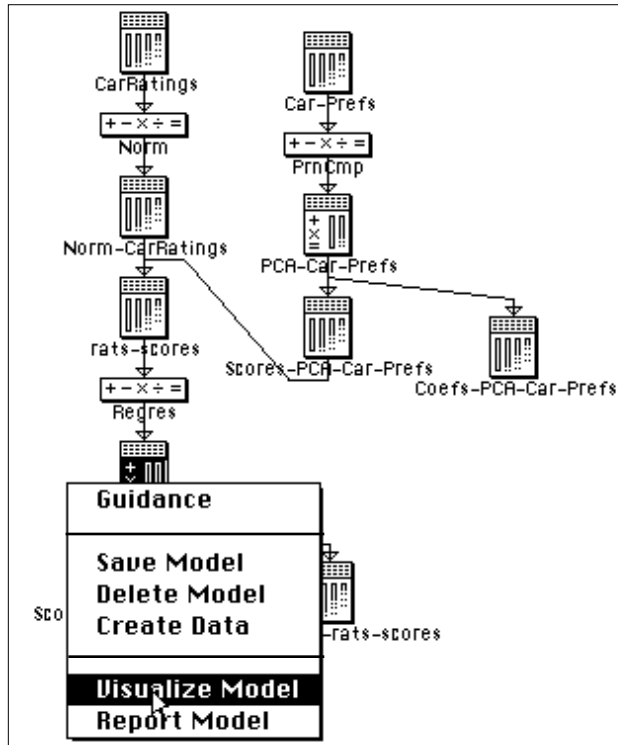
ViSta is [free](#) from <http://forrest.psych.unc.edu/>.

- [Code](#) may be freely copied and redistributed, with certain restrictions.
- [Documentation](#) is also available for free from the above site.

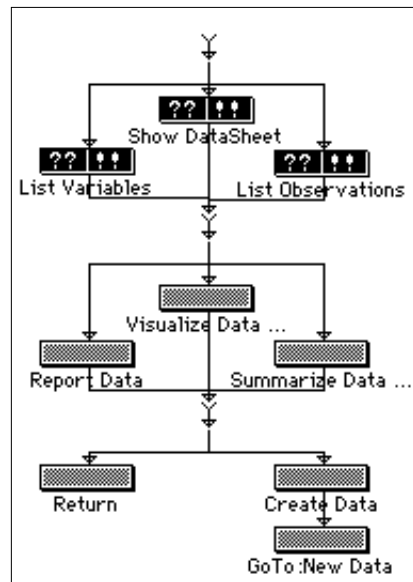
#### [Examples](#)

ViSta's WorkMaps, GuideMaps and SpreadPlots are shown on the next pages.

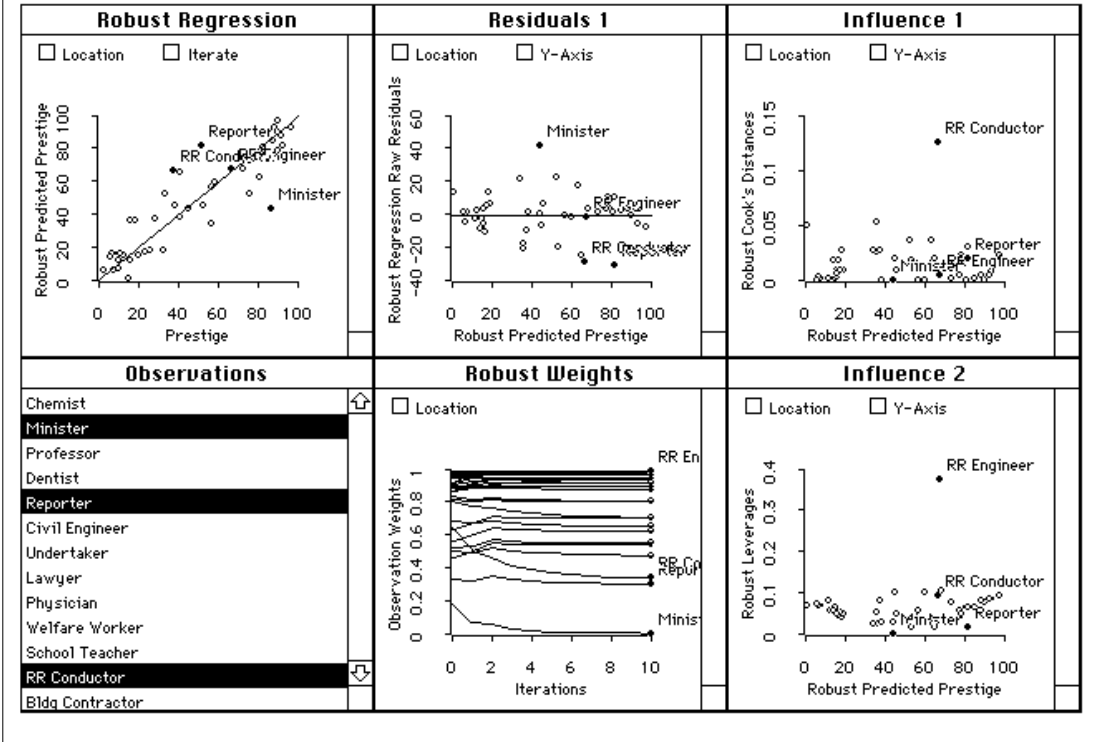
## ViSta's WorkMaps - For Structure Mode



## ViSta's GuideMaps - For Structure Mode



## ViSta's SpreadPlots - For Exploratory Mode



## ViSta's Text Windows — For Confirmatory Mode

**ViDAL:** ViSta's Data Analysis Language (command lines and scripts):

```

> (load-data "animals")
; loading animals
; finished loading "animals"
*Object: 7304556, prototype = TABLE-DATA-OBJECT-PROTO>
> (analysis-of-variance)
*Object: 7665180, prototype = ANOVA-MODEL-OBJECT-PROTO>
> (report-model)
    
```

**Alphanumeric Reports** of Statistical Analyses (produced by the code above):

SUMMARY OF FIT:					
R Squared:	0.96				
Adjusted R Squared:	0.85				
Sigma hat (RMS error):	0.95				
Number of cases:	24				
Degrees of freedom:	6				
ANALYSIS OF VARIANCE: MODEL TEST					
Source	Sum-of-Squares	df	Mean-Square	F-Ratio	P-Value
Model	132.54	17	7.80	8.64	0.00685
Error	5.42	6	0.90		
Total	137.96	23			
ANALYSIS OF VARIANCE: EFFECTS TESTS					
Source	Sum-of-Squares	df	Mean-Square	F-Ratio	P-Value
Species	51.04	1	51.04	56.54	0.00029
Subject	14.58	2	7.29	8.08	0.01987
Season	47.46	3	15.82	17.52	0.00226
Species*Subject	2.58	2	1.29	1.43	0.3104
Species*Season	7.46	3	2.49	2.75	0.13454
Subject*Season	9.42	6	1.57	1.74	0.25918
Error	5.42	6	0.90		
Total	137.96	23			

## 5: Individual Differences & Data Analysis

**Individual Differences:** We argue that [data analysts vary](#) in data analysis sophistication.

- We identify four types of analysts:  
[novice](#), [competent](#), [sophisticated](#), and [expert](#).
- We believe that the sophistication level alters the [profile](#) of cognitive modes.  
For example, we believe that novices would spend the most time in structure mode, while sophisticates and experts would spend the least time in it.

We also believe that [different computing environments](#) are needed for the [different levels of sophistication](#).

With [ViSta](#) these different computing environments are provided by the combinations of software modes that are needed by the user:

- [Novice](#) analysts need an environment that provides structured guidance. These users will use [GuideMaps & WorkMaps](#).
- [Competent](#) analysts need an un-guided but structured environment. These users will use [WorkMaps](#), but not [GuideMaps](#).
- [Sophisticated](#) analysts need menus and command lines. These users will use [Menus](#), but will use [neither](#) [GuideMaps](#) nor [WorkMaps](#).
- [Experts](#) will use command lines and will dispense with [Menus](#) as well as [GuideMaps](#) and [WorkMaps](#).

## 6: Proposed Empirical Usability Study

### [Our Fundamental Hypothesis:](#)

The naive data analyst analyzing data in an environment that visually guides and structures the analysis will be more productive, accurate, satisfied, etc., than when the same analyst performs the same analysis without visual guidance and structure.

### [A Secondary Hypothesis:](#)

Data analysts with more experience will spend less time using the visual aids for structure and guidance than those with less experience.

### [Subjects:](#)

Undergraduates in an introductory Psychology Statistics course.

### [Design:](#)

1. [Guided and Structured](#) Visual Interface: [GuideMaps](#) and [WorkMaps](#)
2. [Unguided, Structured](#) Visual Interface: [WorkMaps](#) but no [GuideMaps](#)
3. [Unguided, Unstructured](#) Visual Interface: [Menus](#), but no [GuideMaps](#) and [WorkMaps](#)
4. [Non-Visual](#) Interface: [Command lines](#), no [Menus](#), [GuideMaps](#) or [WorkMaps](#).

## Proposed Empirical Usability Study

### Response Measures:

These measures are yet to be defined. They will tap the user's

1. data analysis productivity
2. data analysis accuracy
3. satisfaction with the data analysis

### Hypothesis:

Since these are naive subjects, we predict that the subjects with the most visual aids will perform best. That is, group 1 will perform best, group 2 next best, group 3 worse than 1 and 2, and group 4 the worst.

### Stay Tuned:

We will perform this usability study during the coming academic year.

## 7: Conclusion

### Current trends:

Developments in hardware and software will continue to increase the widespread use of data analysis software.

- Thus, the need to understand the data analysis process and to improve data analysis systems will continue to increase.

Our cognitive model of data analysis is a way to understand the data analysis process. We believe it will prove to be a useful way.

- ViSta, our Visual Statistics System is firmly based on our cognitive model and on current developments in computer science.
- Our usability study, is designed to show whether ViSta, and the cognitive model it is based on, are
  1. useful in improving the data analysis process,
  2. useful in improving our understanding of that process.

We hope that we can soon report that empirical results support the usefulness of our cognitive model of data analysis, and its' software implementation.

*PostScript error (--nostringval--, get)*