ARTIFICIAL INTELLIGENCE IN STATISTICS

by

Stephen M. Samuels
Purdue University

Technical Report #89-06

Department of Statistics
Purdue University

March 1989

AMS 1980 Subject Classification: Primary 62–07, Secondary 68T30
Key Words and phrases: Expert systems, statistical software, data analysis.
Research supported in part by NSF Grant DMS-8717799 at Purdue University.

◇

Artificial Intelligence in Statistics *

by
Stephen M. Samuels
Purdue University

Abstract

More intelligent statistical software is urgently needed. Of equal importance is that responsible statisticians be intimately involved in the process of developing such software and not abdicate this responsibility to, e.g., the AI community. A summary of some recent and current work is presented here.

1 Introduction

The title of this paper refers to what is commonly called "more intelligent" statistical software: in other words, statistical software which embodies features developed in the realm of Artificial Intelligence, most notably the *expert system.* The title should not be confused with "Statistics in Artificial Intelligence" which has to do with the problem of handling so-called "uncertain knowledge" in all kinds of expert systems. Both topics get equal time in Gale (1986), but only one will be discussed here.

However, there is, in Statistics in AI, an important object lesson to which I want to allude before getting back on track. AI people have apparently largely turned their backs on real probability (e.g. Bayes' formula), preferring instead such incoherencies as Belief Functions, Certainty Measures, and Fuzzy Sets. "Bayesian ideas don't work," you may hear them say, when in fact they haven't been properly tried at all. Actually, Bayesian ideas *do* work; a good example is Spiegelhalter (1984). But statisticians haven't paid enough attention to the problem, hence these other techniques have largely filled a vacuum. It is vitally important that responsible statisticians not allow the same thing to happen in the development of the statistical software of the future.

*Based on a lecture October 4, 1988 at the National Institute of Standards and Technology, Gaithersburg, MD.

2 Quotable Quotes

Here is how one particularly celebrated observer, John Tukey, assessed the magnitude of the sugject in 1985, in quotes taken from Gale (1986):

> By 1995 or so, the largest single driving force in guiding general work on data analysis and statistics is going to be the needs that have to be met to understand and improve data-analytic expert systems.
>
> ...
>
> They will ameliorate, not be cure-alls. But they are the only hope of connecting a slowly growing population of professional data analysts and statisticians to a rapidly growing population of data that requires analysis.

To many people, the "They" in Tukey's quote are some kind of expert systems, which prompts this cautious appraisal from Peter Huber (1985), also in Gale (1986):

> ... There are no real experts for data analysis, *per se*. Thus an essential ingredient for building an expert system for exploratory data analysis is lacking.
>
> ...
>
> I am afraid that the prevelant approach to expert systems in statistics—taking a circumscribed area with an established methodology (regression, time series, etc.) and aiding a naive user carrying through a highly sophisticated analysis—rather than to promote good statistics may once more promote giving the "right" answers to the wrong questions.

We'll give the last word to Tukey, although he said it in 1982:

> [the] all-too-automatic reaction by many statisticians and data analysts, the reaction that automating... is a threat to their professional position... is a serious misconception.

3 Paradigms

What kind of intelligence should statistical software have, and where should this intelligence come from? Here are two quite different approaches, the first essentially due to Huber, the second to Gale, among others.

3.1 The Lab Assistant

As Figure 3.1 suggests, a scientist has come to a real-live data analyst for help, which is being provided *in real-time!* This allows for *interaction* between them. Either one can suggest approaches and react and adapt to the results. This is, of course, vastly superior to the old-fashioned batch mode approach. But, for it to actually be feasible requires software that is more than merely "interactive." As Huber suggests, it should be responsive to the System Experts wishes in much the same way as a graduate lab assistant would be (but *much faster!!*).

"Check the correctness of these model assumptions [perhaps by doing some simulations]," we should be able to say to our software. Or "repeat the analysis we just did with this modified data;" or "modify the analysis with the same data, as follows...." Needless to say, the software would do our bidding unobtrusively, allowing us to simultaneously pursue other tasks, and politely prompt us when it's ready to show off its results. Meanwhile, our electronic lab assistant would be keeping a diary of our session—but not a passive, recording device type of diary such as currently exists in some statistical packages. This diary is to be a lot like a lab assistant's notebook. For example, the users might see on the screen a message like: "excuse me, but would you like to comment on the fit?" and the comments would become diary annotations. The diary might also contain things which are done in the background, without the users' explicit knowledge, but which the software "thinks" will be needed later.

It may be too much to ask this Lab Assistant software to actually write the final report, but the diary it produces ought to provide at least an approximation to a first draft. Furthermore, the diary should itself be "executable;" that is, we should be able to replay the session, at least the useful parts of it. (Actually this is not new; the diary command in "S" produces an executable file, which the user can edit before re-running it.)

Now what do you call an executable diary produced from a comprehensive data analysis session involving a system expert and a domain expert? Is it...could it be...an Expert System in Statistics? In other words, could it, with a little bit of tinkering, perhaps, perform some other similar analysis for some other, or the same, scientist, *automatically* without the intervention of either the live Data Analyst or the Lab Assistant software? This vision of an expert system almost literally creating itself has been a tantalizing one. Many a diary has been created with the hope of such a metamorphosis. But the results (circa 1988) have been frustrating (see the STUDENT example in the next section). To bridge the gap between executable diaries and a genuine (usually "rule-based") expert system, enter an AI guy called a Knowledge Engineer.

3.2 The Compleat Consultant

As Figure 3.2 shows, the Data Analyst now switches seats and plays a role analogous to the Scientist in the previous paradigm, providing expertise about his own specialty, Statistics. An example of this collaboration occurred at Bell Labs where Bill Gale played the role of

PARADIGM: THE LAB ASSISTANT

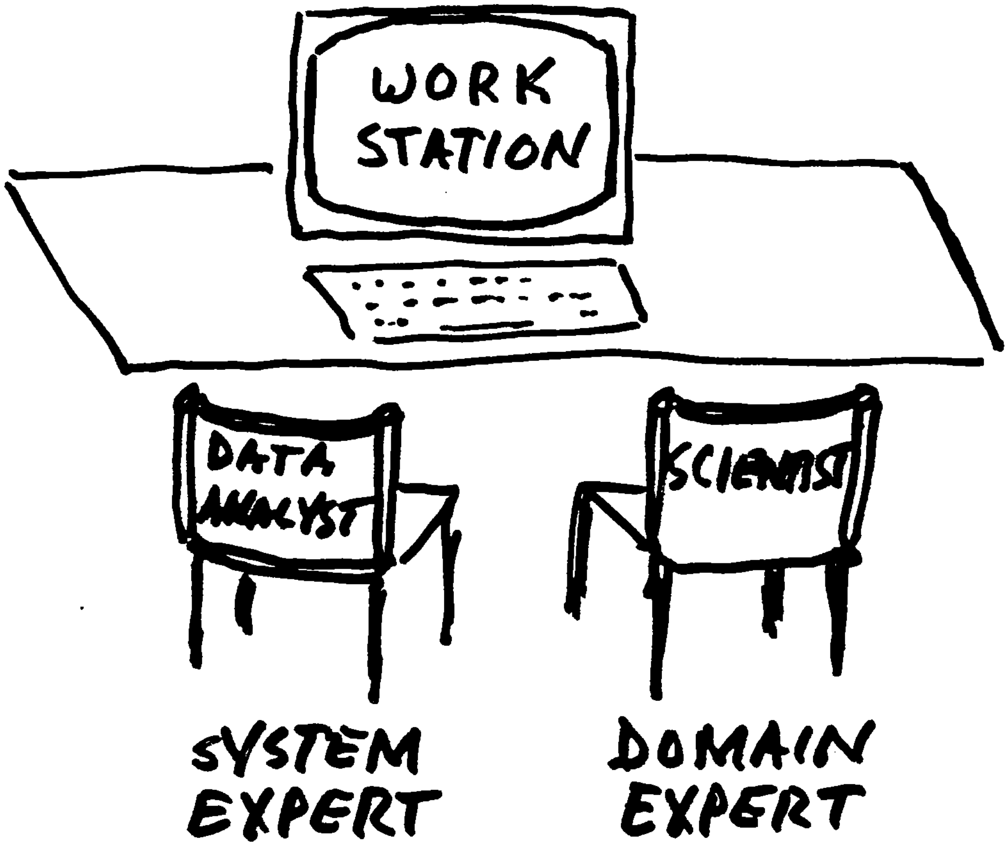


Fig. 3.1

PARADIGM: THE COMPLEAT CONSULTANT

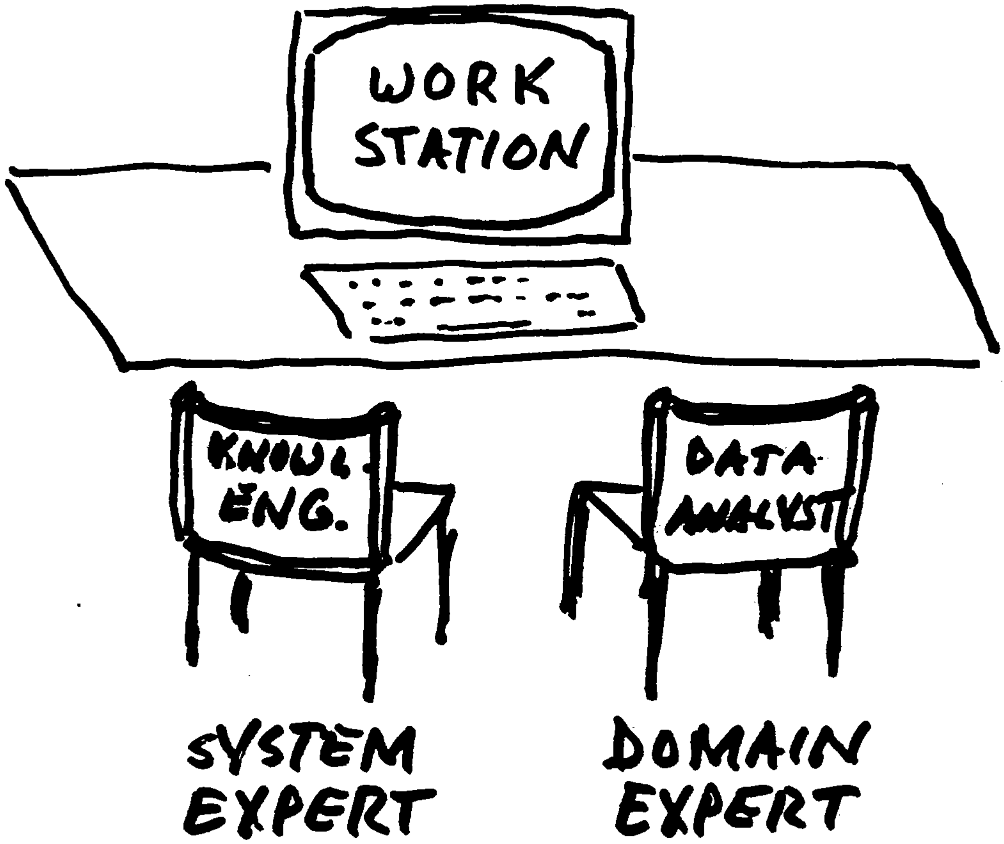


Fig. 3.2

Knowledge Engineer and Daryl Pregibon was the Data Analyst. The program they created was REX (Regression Expert); it's mentioned in the next section.

The object here is much more ambitious, to create a statistical expert system, which can then be used by statistical laypersons like the scientist, rather than by professional statisticians. How else can we do what Tukey said needs to be done?

3.3 Two goals... or one?

But notice that Tukey spoke not only of *improving* but also of *understanding* data-analytic systems. So while we are striving to automate data analysis by building expert systems, we need to be keenly aware of the need for some kind of a theory to help us understand what is really "significant" among the multitude of things which computing power allows us to see. It is interesting to note that both of these goals require us to formalize the process of data analysis, or as some have said, to try to "capture statistical strategy."

4 Prototypes

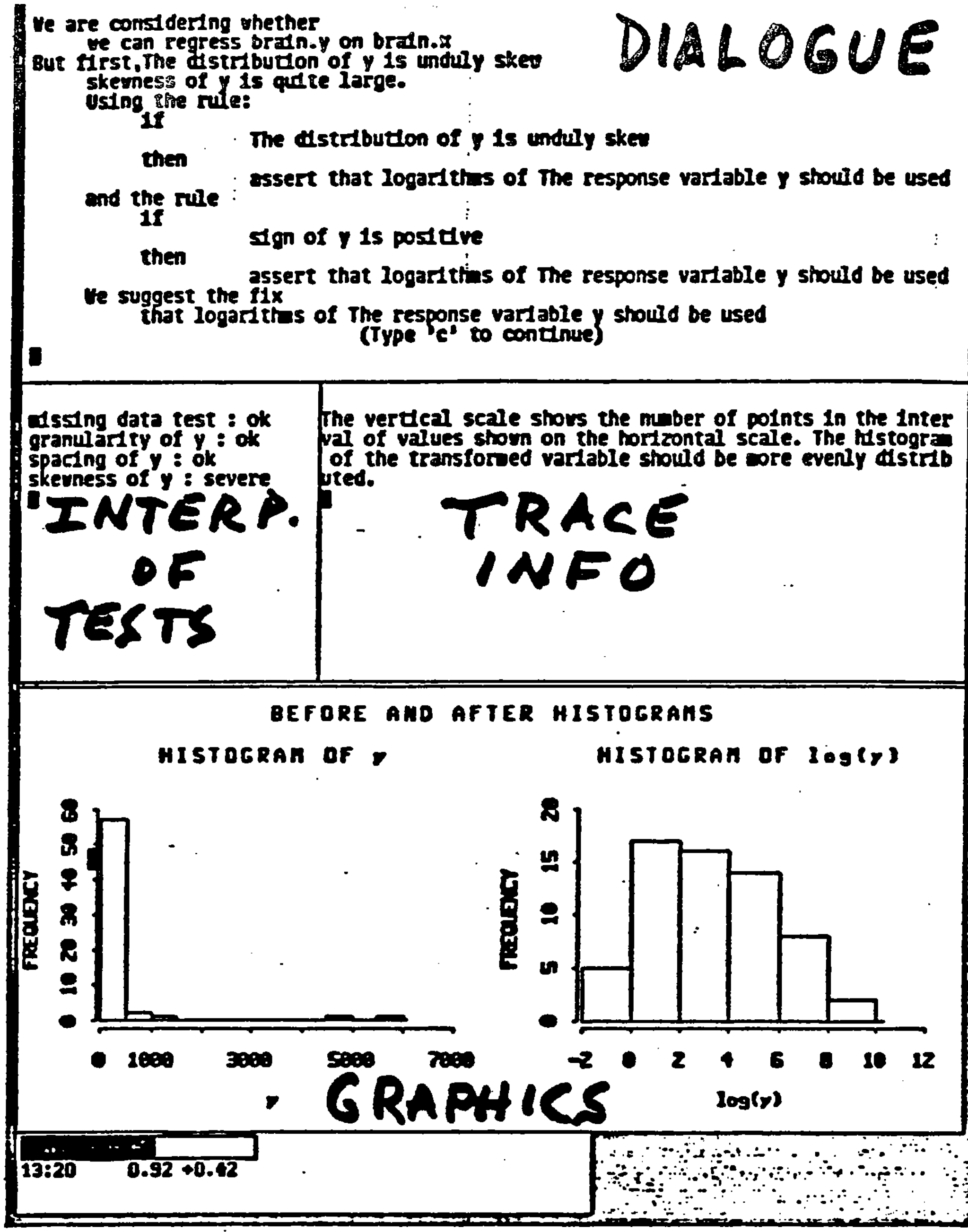
Most of the examples to follow were presented at Compstat VIII, a Symposioum held in Copenhagen in August, 1988 under the auspices of the International Association for Statistical Computing. Almost without exception, the word "prototypes" is much too optimistic; most never have and never will be put into production, or even be seriously used by anyone other than their developers.

REX Regression EXpert. Developed by Gale and Pregibon oa AT&T Bell Labs. The goal was to encode enough expert knowledge to enable a novice user to do a simple linear regression "safely." See Figures 4.1 and 4.2 for illustrations of the multi-window environment it provides the user.

STUDENT Learn how to do data analysis by studying examples, such as diaries of sessions with live experts.

TESS (Tree-based Environment for Statistical Strategies) Here Gale and Pregibon abandoned trying to model a human analyst, in part because they became convinced that only a human can supply enough of the *context.* Alternatively they now seek context-free results (note the plural). Their inspiration is the classic book of Daniel and Wood (1971, 1980), who said

> Although we usually get our notions one at a time, it does not follow that we will find the best equation(s) by accepting or rejecting each notion after examining it *once.* It will often be possible to broaden our experience by looking at each new idea in the full context of its predecessors.



4.7

Figure ~~8.10~~ We got out of the lexicon, were presented again with the choice of Figure 8.6, and asked for why. This provides a verbal answer to why the suggestion is being made in addition to the graphical answer still showing. The stylistic weakness of the response reflects the fact that REX responds by relating its *general* reasoning, not canned messages for each specific situation.

To conclude that logarithms could be used at this point, REX had to check that no negative values were present in the data. If it had found zero values, it would have asked the user whether adding a small amount to each value would change the interpretation.

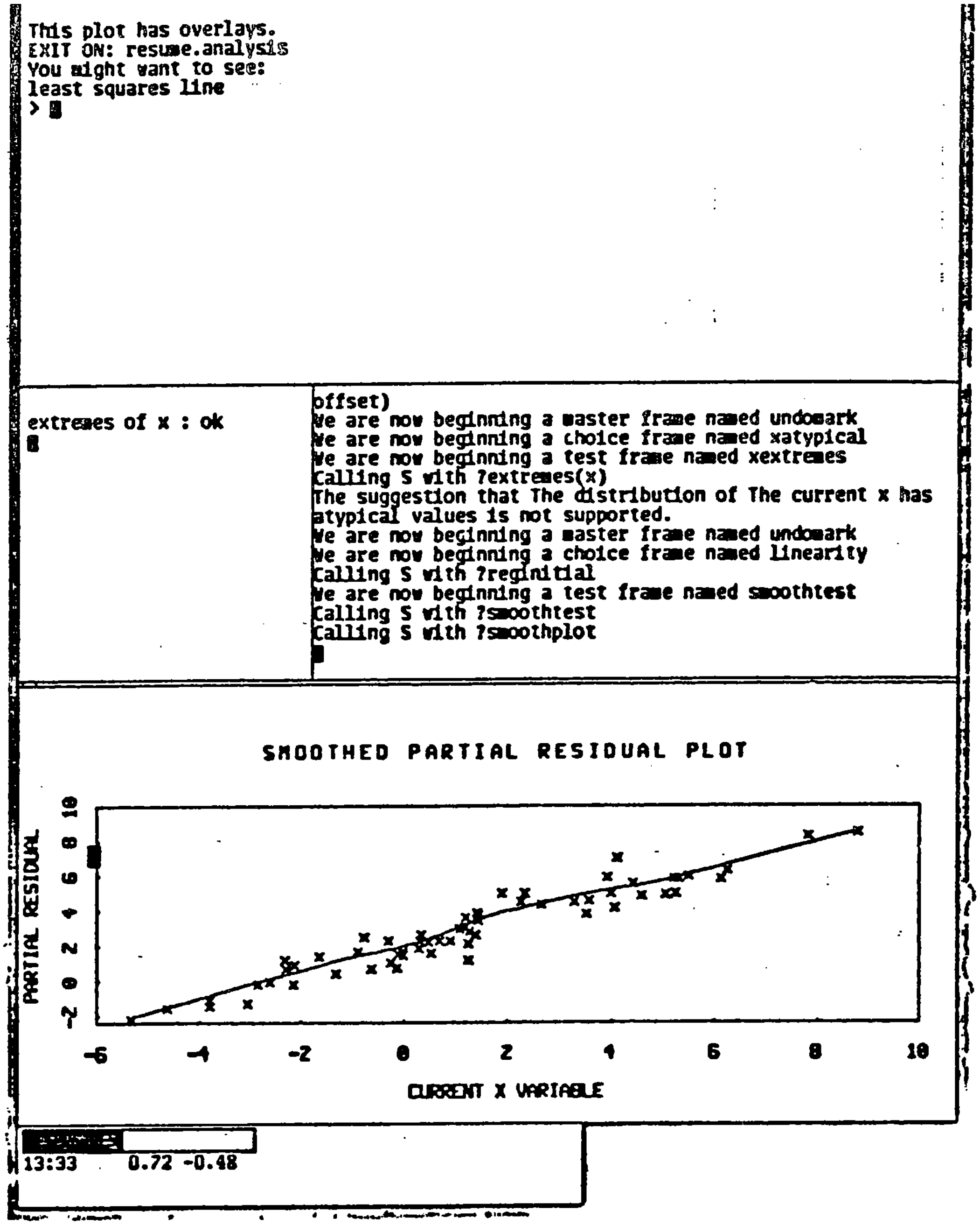


Figure ~~8.12~~ 4.2 We accepted the logarithmic transform suggested for the dependent variable. REX made a few more tests on the transformed dependent variable, then considered the independent variable. The same kind of skewness problem was found, and a logarithmic transform made there also.

The next stage of the analysis is linearity checking. The test used finds a mild nonlinearity, so REX pauses to ask if the user wants to call the problem ok. A plot is offered, which we accept. The plot show a scatter plot of the transformed variable and a smoothing curve. The curve shows a small bump in the middle. We would suggest calling this "ok." In fact, REX does not have a means of fixing this problem.

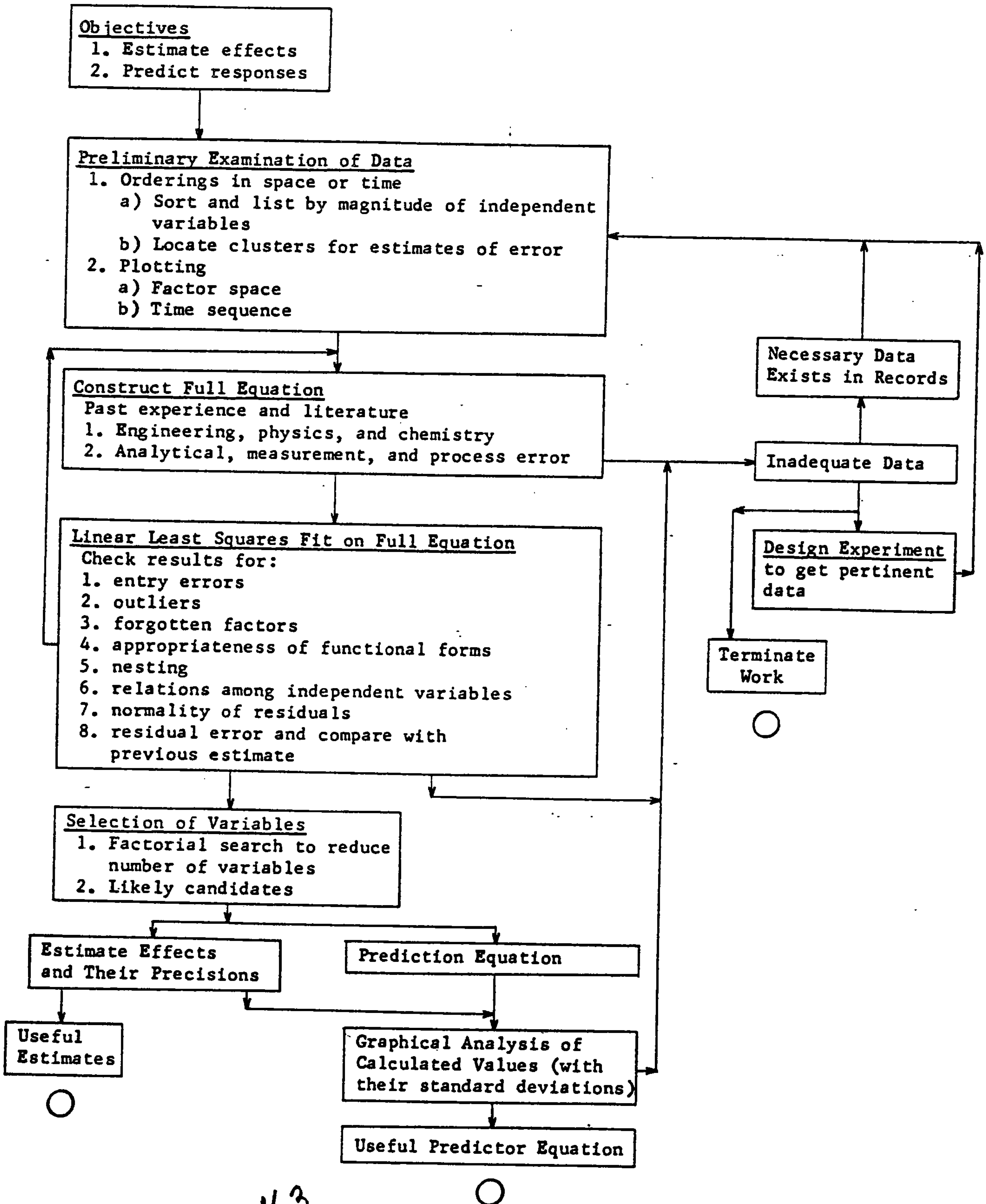


Figure ~~1.1~~ 4.3 Flow diagram for fitting equations to data.

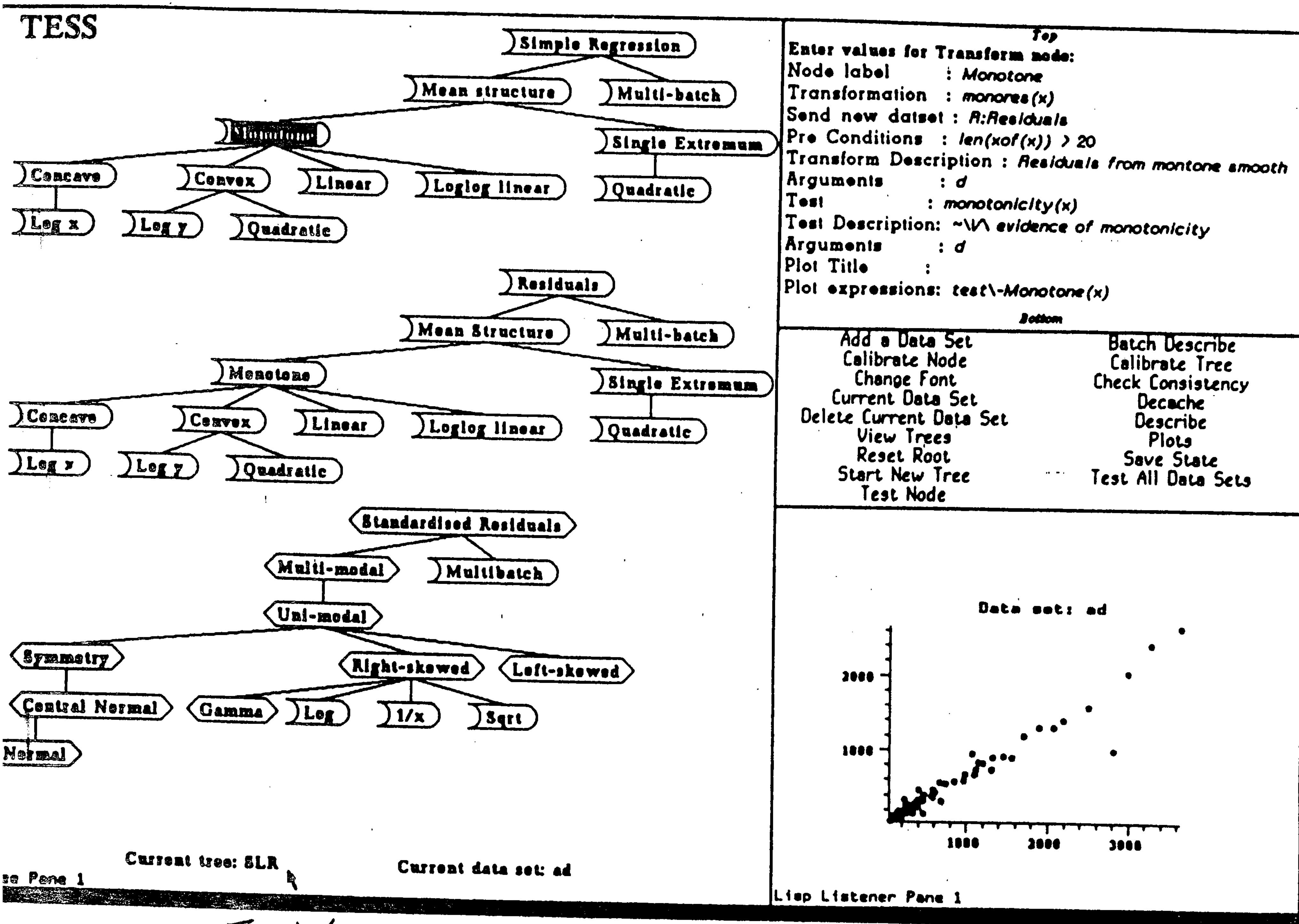


Fig 4.4
~~Fig. 2.~~ Screen image of TESS in the basic mode of interaction.

The flow chart from their book (Figure 4.3) inspired the tree structure in Figure 4.4. A central idea in TESS is the trade-off between *accuracy* and *parsimony*, the measurement of which is not yet fully worked-out.

RX Translates medical problems into statistical problems, and does regression.

MUSE (Method to USE) Selects appropriate method of analysis based on user-supplied inputs and outputs.

DINDE ("turkey" in French; don't ask why.) The work of Oldfield and Peters in the US. Object oriented: "provides a visual map of an analysis." For use by experts.

CADEMO Computer-Assisted Design of Experiments and MOdelling. A large consortium of East Germans is producing this.

SELINA another experimental design support system, developed in the UK.

PRINCE front-end for PRINCALS (PRINcipal Components AnaLysiS), developed in the Netherlands

GLIMPSE front-end for GLIM (Generalized LInear Models), developed by Nelder et al in the UK.

5 Questions and Answers

There is currently considerable debate centering on questions such as

- WHO is it for? At the present stage of development, should we primarily trying to help the professional data analyst, as in the first paradigm, or should be aiming directly at the novice? It is not too much of an oversimplification to summarize the sentiment at Compstat 88 by saying that the Americans seemed to have scaled back their ambitions from the latter toward the former, while the Europeans are concentrating—but not very ambitiously—on the latter.
- WHAT should it do? Should it do several kinds of data analysis, or just one kind; or should it be oriented toward a particular kind of application rather than toward a particular kind of Statistical technique.
- WHY do we need it?

  > Here is the moral imperative: Statistical software in its present form, made widely available by cheap computing, will precipitate much uninformed, unguided, and simply incorrect data analysis. We are obliged to do something to help.
  >
  > –Chambers (1981)

References

[1] Daniel and Wood (1980) *Fitting Equations to Data* 2nd ed. Wiley, New York.

[2] Gale, W. A., ed (1986). *Artificial Intelligence and Statistics*. Addison-Wesley, Reading MA.

[3] Hahn, G. J. (1985). More intelligent statistical software and statistical expert systems: Future directions (with discussion). *The American Statistician* **39** 1–16.

[4] Spiegelhalter, D. J., and Knill-Jones, R. P. (1984) Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroenterology (with discussion). *J. Royal Statistical Society, Series B* **147** 35–77.

STATISTICAL ANALYSIS FOR

STOCHASTIC MODELING and SIMULATION

WITH APPLICATIONS TO MANUFACTURING

ARTIFICIAL INTELLIGENCE IN STATISTICS

Stephen M. Samuels†

SMS 89-11
March 1989

PURDUE UNIVERSITY
Department of Statistics
School of Industrial Engineering

PRITSKER & ASSOCIATES, INC.

ARTIFICIAL INTELLIGENCE IN STATISTICS

Stephen M. Samuels†

SMS 89-11
March 1989

School of Industrial Engineering
Purdue University
West Lafayette, IN 47905

† Department of Statistics
Purdue University
West Lafayette, In 47907

Research supported in part by NSF Grant DMS-8717799 at Purdue University.

Also Technical Report #89-06, Department of Statistics, Purdue University.

Artificial Intelligence in Statistics *

by
Stephen M. Samuels
Purdue University

Abstract

More intelligent statistical software is urgently needed. Of equal importance is that responsible statisticians be intimately involved in the process of developing such software and not abdicate this responsibility to, e.g., the AI community. A summary of some recent and current work is presented here.

1 Introduction

The title of this paper refers to what is commonly called "more intelligent" statistical software: in other words, statistical software which embodies features developed in the realm of Artificial Intelligence, most notably the *expert system.* The title should not be confused with "Statistics in Artificial Intelligence" which has to do with the problem of handling so-called "uncertain knowledge" in all kinds of expert systems. Both topics get equal time in Gale (1986), but only one will be discussed here.

However, there is, in Statistics in AI, an important object lesson to which I want to allude before getting back on track. AI people have apparently largely turned their backs on real probability (e.g. Bayes' formula), preferring instead such incoherencies as Belief Functions, Certainty Measures, and Fuzzy Sets. "Bayesian ideas don't work," you may hear them say, when in fact they haven't been properly tried at all. Actually, Bayesian ideas *do* work; a good example is Spiegelhalter (1984). But statisticians haven't paid enough attention to the problem, hence these other techniques have largely filled a vacuum. It is vitally important that responsible statisticians not allow the same thing to happen in the development of the statistical software of the future.

*Based on a lecture October 4, 1988 at the National Institute of Standards and Technology, Gaithersburg, MD.

2 Quotable Quotes

Here is how one particularly celebrated observer, John Tukey, assessed the magnitude of the sugject in 1985, in quotes taken from Gale (1986):

> By 1995 or so, the largest single driving force in guiding general work on data analysis and statistics is going to be the needs that have to be met to understand and improve data-analytic expert systems.
>
> ...
>
> They will ameliorate, not be cure-alls. But they are the only hope of connecting a slowly growing population of professional data analysts and statisticians to a rapidly growing population of data that requires analysis.

To many people, the "They" in Tukey's quote are some kind of expert systems, which prompts this cautious appraisal from Peter Huber (1985), also in Gale (1986):

> ... There are no real experts for data analysis, *per se.* Thus an essential ingredient for building an expert system for exploratory data analysis is lacking.
>
> ...
>
> I am afraid that the prevelant approach to expert systems in statistics—taking a circumscribed area with an established methodology (regression, time series, etc.) and aiding a naive user carrying through a highly sophisticated analysis—rather than to promote good statistics may once more promote giving the "right" answers to the wrong questions.

We'll give the last word to Tukey, although he said it in 1982:

> [the] all-too-automatic reaction by many statisticians and data analysts, the reaction that automating... is a threat to their professional position... is a serious misconception.

3 Paradigms

What kind of intelligence should statistical software have, and where should this intelligence come from? Here are two quite different approaches, the first essentially due to Huber, the second to Gale, among others.

3.1 The Lab Assistant

As Figure 3.1 suggests, a scientist has come to a real-live data analyst for help, which is being provided *in real-time!* This allows for *interaction* between them. Either one can suggest approaches and react and adapt to the results. This is, of course, vastly superior to the old-fashioned batch mode approach. But, for it to actually be feasible requires software that is more than merely "interactive." As Huber suggests, it should be responsive to the System Experts wishes in much the same way as a graduate lab assistant would be (but *much faster!!*).

"Check the correctness of these model assumptions [perhaps by doing some simulations]," we should be able to say to our software. Or "repeat the analysis we just did with this modified data;" or "modify the analysis with the same data, as follows...." Needless to say, the software would do our bidding unobtrusively, allowing us to simultaneously pursue other tasks, and politely prompt us when it's ready to show off its results. Meanwhile, our electronic lab assistant would be keeping a diary of our session—but not a passive, recording device type of diary such as currently exists in some statistical packages. This diary is to be a lot like a lab assistant's notebook. For example, the users might see on the screen a message like: "excuse me, but would you like to comment on the fit?" and the comments would become diary annotations. The diary might also contain things which are done in the background, without the users' explicit knowledge, but which the software "thinks" will be needed later.

It may be too much to ask this Lab Assistant software to actually write the final report, but the diary it produces ought to provide at least an approximation to a first draft. Furthermore, the diary should itself be "executable;" that is, we should be able to replay the session, at least the useful parts of it. (Actually this is not new; the diary command in "S" produces an executable file, which the user can edit before re-running it.)

Now what do you call an executable diary produced from a comprehensive data analysis session involving a system expert and a domain expert? Is it... could it be... an Expert System in Statistics? In other words, could it, with a little bit of tinkering, perhaps, perform some other similar analysis for some other, or the same, scientist, *automatically* without the intervention of either the live Data Analyst or the Lab Assistant software? This vision of an expert system almost literally creating itself has been a tantalizing one. Many a diary has been created with the hope of such a metamorphosis. But the results (circa 1988) have been frustrating (see the STUDENT example in the next section). To bridge the gap between executable diaries and a genuine (usually "rule-based") expert system, enter an AI guy called a Knowledge Engineer.

3.2 The Compleat Consultant

As Figure 3.2 shows, the Data Analyst now switches seats and plays a role analogous to the Scientist in the previous paradigm, providing expertise about his own specialty, Statistics. An example of this collaboration occurred at Bell Labs where Bill Gale played the role of

PARADIGM: THE LAB ASSISTANT

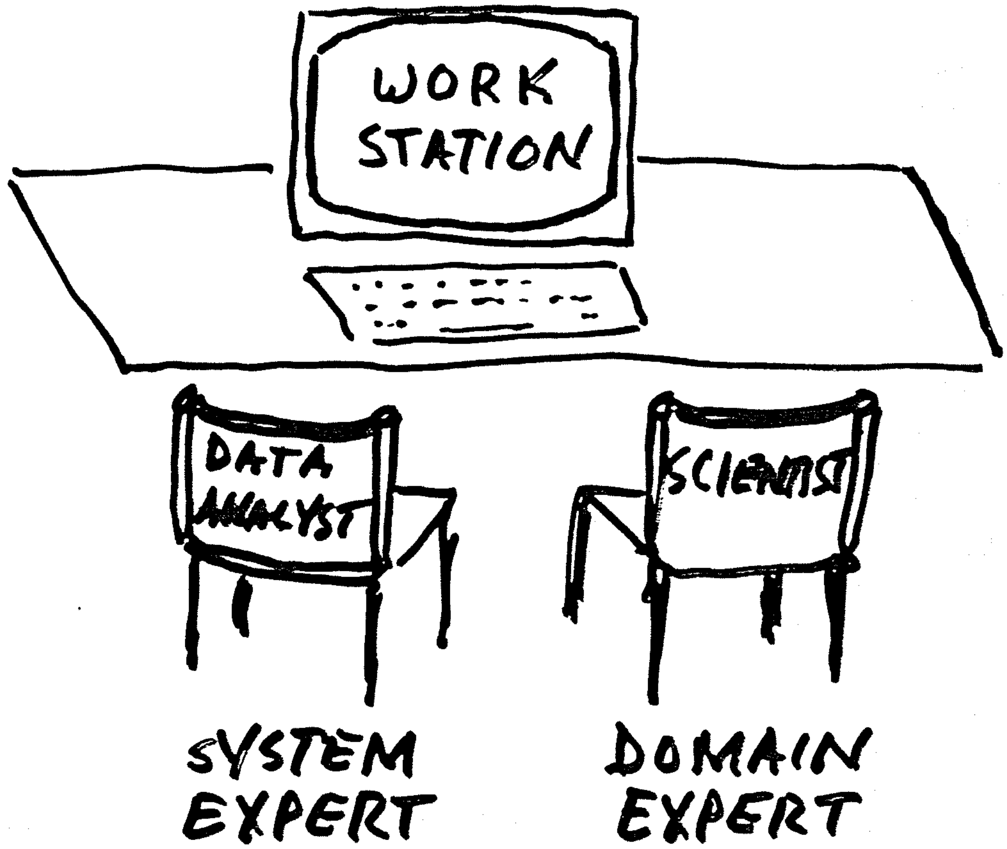


Fig. 3.1

PARADIGM: THE COMPLEAT CONSULTANT

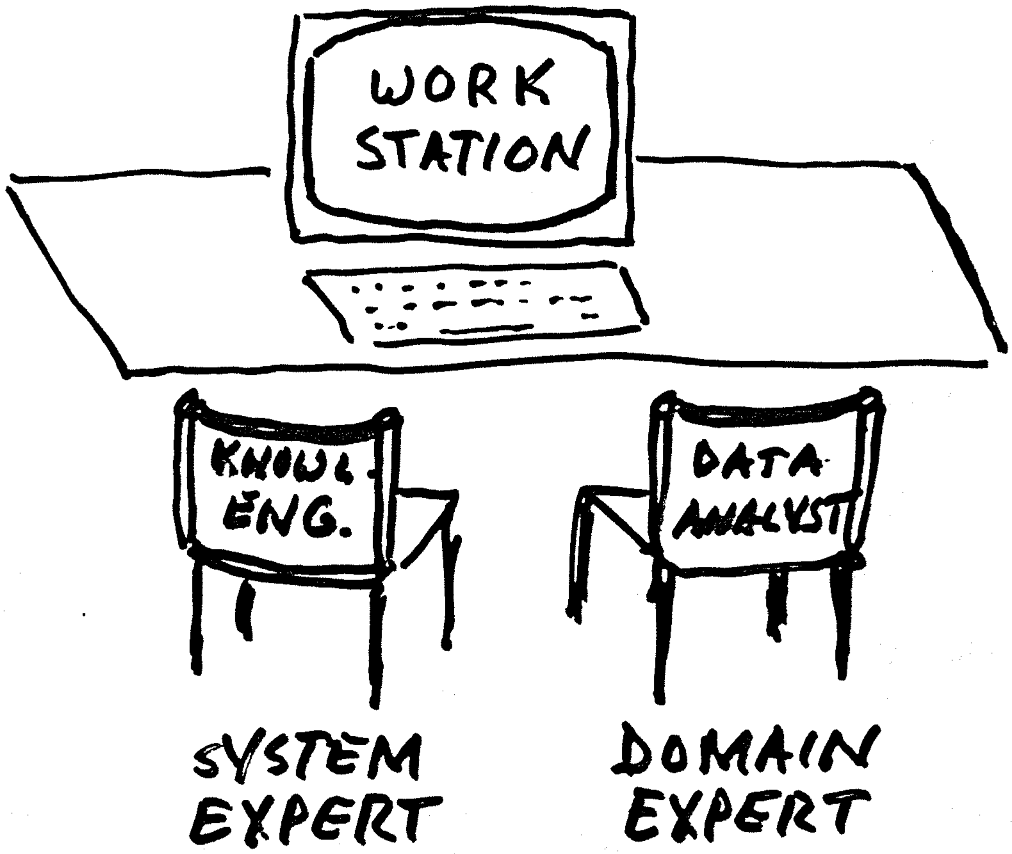


Fig. 3.2

Knowledge Engineer and Daryl Pregibon was the Data Analyst. The program they created was REX (Regression Expert); it's mentioned in the next section.

The object here is much more ambitious, to create a statistical expert system, which can then be used by statistical laypersons like the scientist, rather than by professional statisticians. How else can we do what Tukey said needs to be done?

3.3 Two goals... or one?

But notice that Tukey spoke not only of *improving* but also of *understanding* data-analytic systems. So while we are striving to automate data analysis by building expert systems, we need to be keenly aware of the need for some kind of a theory to help us understand what is really "significant" among the multitude of things which computing power allows us to see. It is interesting to note that both of these goals require us to formalize the process of data analysis, or as some have said, to try to "capture statistical strategy."

4 Prototypes

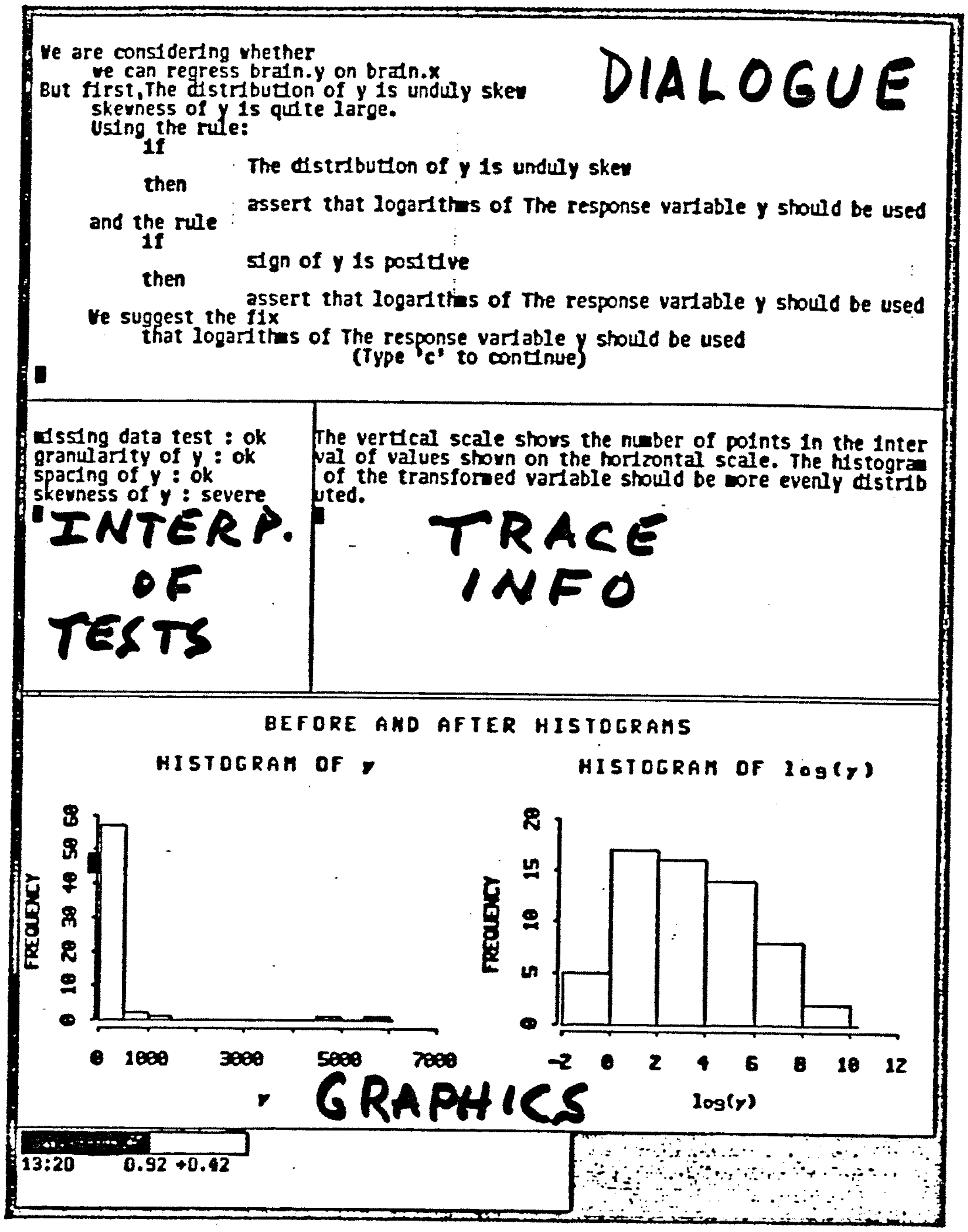
Most of the examples to follow were presented at Compstat VIII, a Symposioum held in Copenhagen in August, 1988 under the auspices of the International Association for Statistical Computing. Almost without exception, the word "prototypes" is much too optimistic; most never have and never will be put into production, or even be seriously used by anyone other than their developers.

REX Regression EXpert. Developed by Gale and Pregibon oa AT&T Bell Labs. The goal was to encode enough expert knowledge to enable a novice user to do a simple linear regression "safely." See Figures 4.1 and 4.2 for illustrations of the multi-window environment it provides the user.

STUDENT Learn how to do data analysis by studying examples, such as diaries of sessions with live experts.

TESS (Tree-based Environment for Statistical Strategies) Here Gale and Pregibon abandoned trying to model a human analyst, in part because they became convinced that only a human can supply enough of the *context.* Alternatively they now seek context-free results (note the plural). Their inspiration is the classic book of Daniel and Wood (1971, 1980), who said

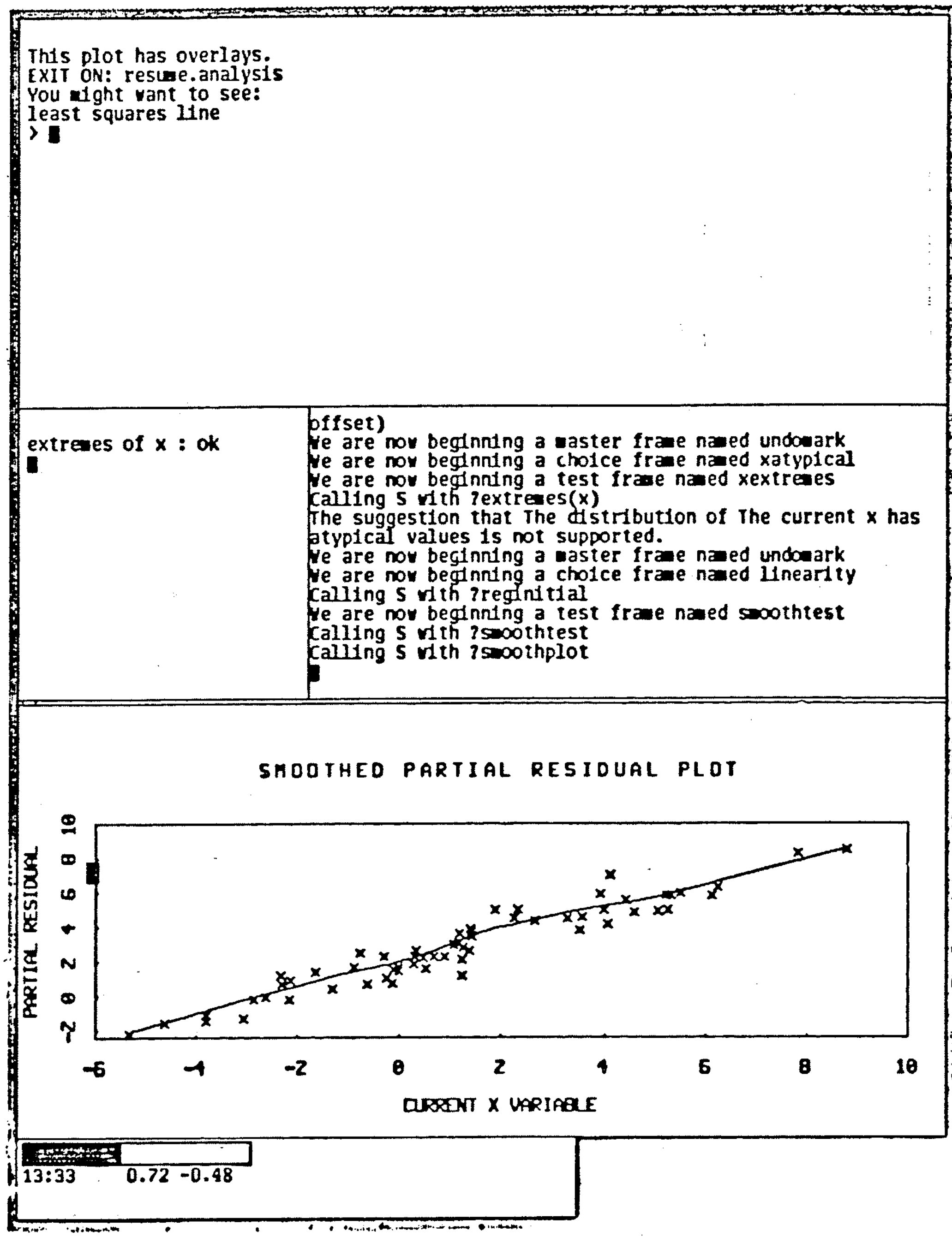
> Although we usually get our notions one at a time, it does not follow that we will find the best equation(s) by accepting or rejecting each notion after examining it *once.* It will often be possible to broaden our experience by looking at each new idea in the full context of its predecessors.



4.1

Figure ~~8.10~~ We got out of the lexicon, were presented again with the choice of Figure 8.6, and asked for why. This provides a verbal answer to why the suggestion is being made in addition to the graphical answer still showing. The stylistic weakness of the response reflects the fact that REX responds by relating its *general* reasoning, not canned messages for each specific situation.

To conclude that logarithms could be used at this point, REX had to check that no negative values were present in the data. If it had found zero values, it would have asked the user whether adding a small amount to each value would change the interpretation.



4.2

Figure ~~8.12~~ We accepted the logarithmic transform suggested for the dependent variable. REX made a few more tests on the transformed dependent variable, then considered the independent variable. The same kind of skewness problem was found, and a logarithmic transform made there also.

The next stage of the analysis is linearity checking. The test used finds a mild nonlinearity, so REX pauses to ask if the user wants to call the problem ok. A plot is offered, which we accept. The plot show a scatter plot of the transformed variable and a smoothing curve. The curve shows a small bump in the middle. We would suggest calling this “ok.” In fact, REX does not have a means of fixing this problem.

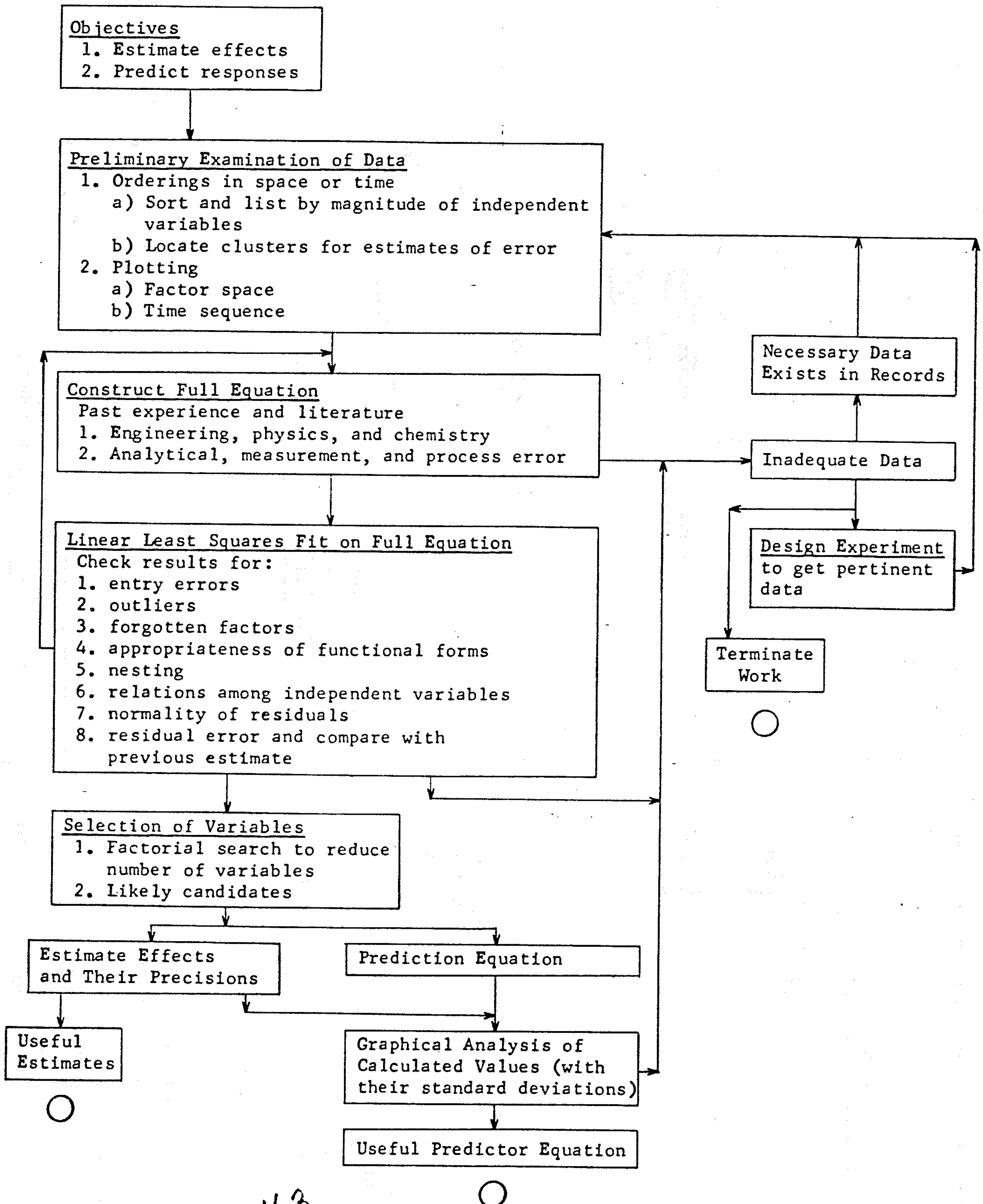


Figure ~~1.1~~ 4.3 Flow diagram for fitting equations to data.

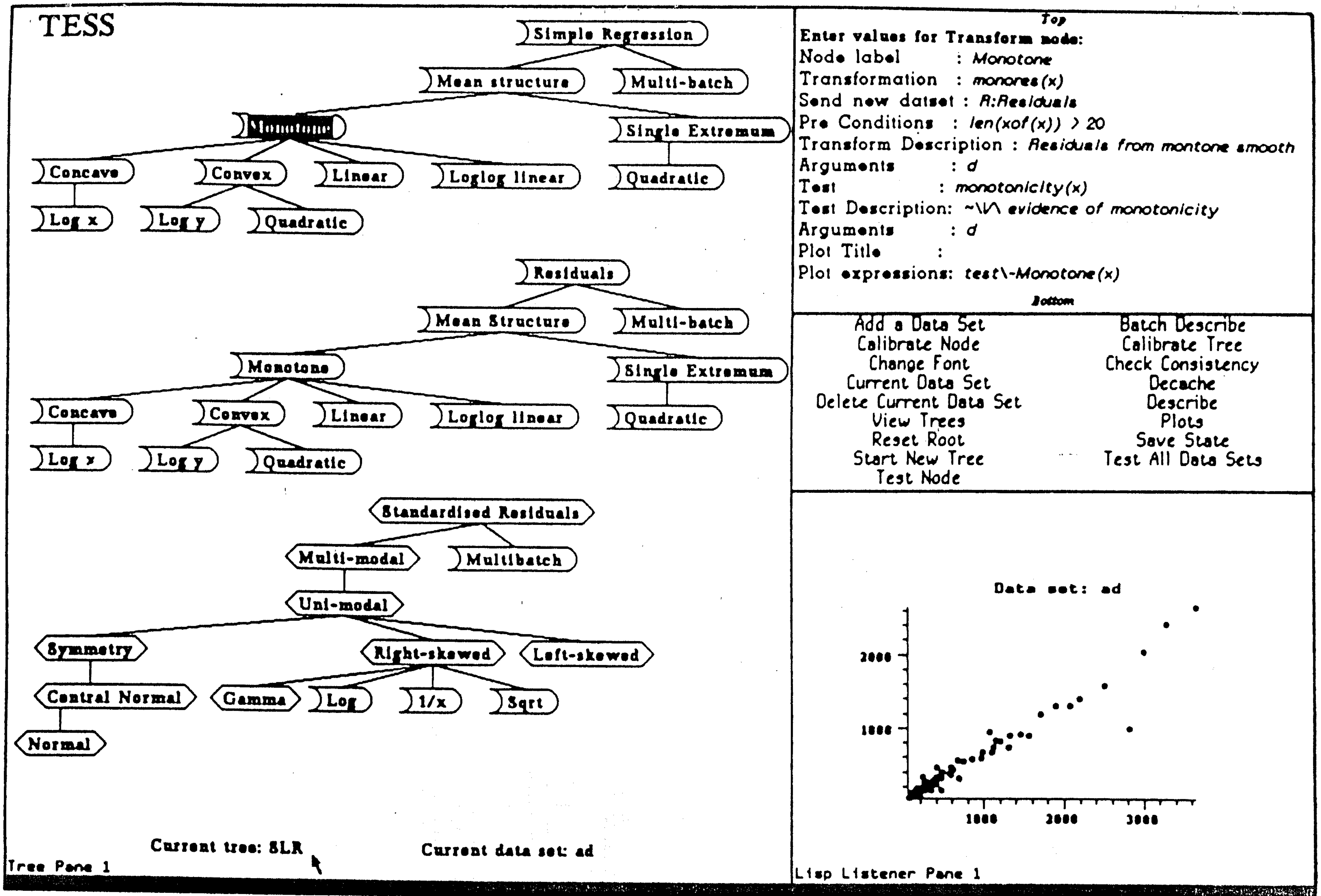


Fig 4.4

~~Fig. 2.~~ Screen image of TESS in the basic mode of interaction.

The flow chart from their book (Figure 4.3) inspired the tree structure in Figure 4.4. A central idea in TESS is the trade-off between *accuracy* and *parsimony,* the measurement of which is not yet fully worked-out.

RX Translates medical problems into statistical problems, and does regression.

MUSE (Method to USE) Selects appropriate method of analysis based on user-supplied inputs and outputs.

DINDE ("turkey" in French; don't ask why.) The work of Oldfield and Peters in the US. Object oriented: "provides a visual map of an analysis." For use by experts.

CADEMO Computer-Assisted Design of Experiments and MOdelling. A large consortium of East Germans is producing this.

SELINA another experimental design support system, developed in the UK.

PRINCE front-end for PRINCALS (PRINcipal Components AnaLysiS), developed in the Netherlands

GLIMPSE front-end for GLIM (Generalized LInear Models), developed by Nelder et al in the UK.

5 Questions and Answers

There is currently considerable debate centering on questions such as

- WHO is it for? At the present stage of development, should we primarily trying to help the professional data analyst, as in the first paradigm, or should be aiming directly at the novice? It is not too much of an oversimplification to summarize the sentiment at Compstat 88 by saying that the Americans seemed to have scaled back their ambitions from the latter toward the former, while the Europeans are concentrating—but not very ambitiously—on the latter.

- WHAT should it do? Should it do several kinds of data analysis, or just one kind; or should it be oriented toward a particular kind of application rather than toward a particular kind of Statistical technique.

- WHY do we need it?

> Here is the moral imperative: Statistical software in its present form, made widely available by cheap computing, will precipitate much uninformed, unguided, and simply incorrect data analysis. We are obliged to do something to help.
>
> –Chambers (1981)

References

[1] Daniel and Wood (1980) *Fitting Equations to Data* 2nd ed. Wiley, New York.

[2] Gale, W. A., ed (1986). *Artificial Intelligence and Statistics*. Addison-Wesley, Reading MA.

[3] Hahn, G. J. (1985). More intelligent statistical software and statistical expert systems: Future directions (with discussion). *The American Statistician* **39** 1–16.

[4] Spiegelhalter, D. J., and Knill-Jones, R. P. (1984) Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroenterology (with discussion). *J. Royal Statistical Society, Series B* **147** 35–77.

STATISTICAL ANALYSIS FOR

STOCHASTIC MODELING and SIMULATION

WITH APPLICATIONS TO MANUFACTURING