8.3 DATA VS. INFORMATION: A SYSTEM PARADIGM

Fred C. Billingsley
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, California 91109

This is a paper about thinking.
It may not seem to be, but it is.
As such, it won't give any answers.
But it should give you some ideas.

In the justification stage of any new system, the proponent is usually asked to provide some sort of Benefit/Cost analysis. Because, for a data system, there is no value in the data per se, justification must be found in the benefits in the use of the data. If, as is usually the case, the instrument or system designer is not a "user", his recourse is to survey the user community to obtain some sort of consensus on the utility. The generally unsatisfactory nature of the resuls is reflected in the large number of times the users are surveyed, resurveyed, and re-resurveyed. Something must be missing, or the answers would have been found.

The thrust here is not the justification of the system itself, but rather the justification of the selection of the various technical parameters which the system must meet. This justification (i.e., optimum parameter tradeoff) must be done in relation to the ability of the user to turn the cold, impersonal data into a live, personal decision or piece of information.

Therein, of course, lies the sleeper: the data system designer requires <u>data</u> parameters, and is dependent on the user to convert his <u>information</u> needs to these data parameters. This conversion will be done with more or less accuracy, beginning a chain of inaccuracies which propogate through the system, and which, in the end, may prevent the user from converting the data which he receives into the information he requires. The concept to be pursued will be that errors will occur in various parts of the system, and, having occurred, will propogate to the end. Modeling of the system may allow an estimation of the effects at any point and the final accumulated effect, and may provide a method of allocating an error budget among the system components.

Inaccuracies will be considered to be of two types, which may be stated in terms of transfer functions for each of the system components considered: 1) Calibration--the difference between the stated transfer function and reality; 2) Uncertainty--the error bars around each stated function and measurement.

This paper presents the results of one phase of research performed by the Jet Propulsion Laboratory, California Institute of Technology, sponsored by the National Aeronautics and Space Administration under Contract NAS7-100.

We begin by modeling an information system as shown in Figure 1. The forward model is required to convert units of information to units of required data, and answers the question "What set of measurements will best carry (i.e., allow the best derivation of) the information?" This box provides the set of measurement "requirements" to the measuring system, which responds with a set of real measurements which will hopefully be somewhat near to the desired set. However, the data system may have an inaccurate or uncertain transfer function, so that the set of apparent measurements presented to the information model deviate further from reality. It is with this set that the user attempts to derive his information, using the Information Model.

Evaluation of the system takes place in two levels as suggested in Figure 2. Note that the evaluation (and, therefore, the design) of a total information system is the joint responsibility of the user and the data system designer, as model boxes under the cognizance of each are involved. The data system designer cannot be held for the inadequacies/uncertainties in either the forward or the information models, although he is deeply interested in the validity of each.

Care must be taken in designing the models, and the systems which they represent. Figure 3 applies to both models and systems. At the low end of complexity, the system may only provide a nominal solution to the information problem Ⓐ, and so the potential errors due to the design may be quite large. But at least, Ⓑ, the data can be obtained. At the other extreme, a complex modelⒹ can produce the desired results quite precisely, if only the data required for the solution could be obtained Ⓒ. If it can be identified, the saddle point, Ⓔ, is the optimum complexity to design to. In the case of registration of Landsat, for example, the saddle point may be found to be at the 0.5-1.5 pixel level, fairly broad, with the moderate gains obtained with very complex processing being very costly or the requisite complex data (e.g., world-wide GCPs) being unobtainable, or at the low complexity end, simple processing producing only moderate registration accuracy.

SYSTEM DESIGN

We will be concerned primarily with the data system design. This includes the choice of the parameters (e.g., spectral bands, resolution, etc.), the exactness with which they must be maintained, the calibration process including the availabilty of required ancillary data, data latency, and the uncertainties associated with each of these items. This must be done in the context of the complete information system. The data system block is diagrammed in Figure 4.

Two approaches may be taken to the data system design: 1) Optimize the data system by minimizing the summation of the deviations of the delivered products from the desired measurements; 2) Optimize the total information system by minimizing the decreases in obtainable information (by the users) due to deviations in the desired measurements from the requested set. One of these approaches is used implicitly, if not explicitly, in any system design. They do not necessarily lead to the same choice of parameters.

Thus, the data system design model for 1) is diagrammed in Figure 5. It can be seen that this is a linear programming problem. The loss function to be minimized is the (weighted, according to the importance of the various disciplines) sum of the deviations in the data delivered from each discipline request. The parameters available to the designer are the sensor bogie parameters, anticipated interference factors (such as sensor vibrations, ground altitude relief displacements, orbit uncertainties, etc.), the ability to measure these, the calibration forward model (i.e., how do we plan to remove the errors?), data system procedures, availability/accuracy of calibration references, and the procedures used to rectify (apply the calibrations). To properly choose between the parameters, coefficients pertaining to the sensitivity of results to variations in each parameter and to the importance of the various parameters are required. (For example, how important to Discipline A is the difference between prompt registration to 1 pixel vs. delayed registration to 0.3 pixel; how important are these relative to overlay matching or to absolute geodetic location, and how important is Discipline A in the total scheme of things?) These coefficients, if available at all, will gen rally be only poorly known. Note, however, that if they are not explicitly stated, they will be assumed by the data system designer with or without affirmation by the discipline users. A caveat to the users!

In approach 2), the user and his forward and information models are explicitly treated, as demonstrated in Figure 6. In this case, the information system design must take into account the effect of the real data on the information conversion in the information models, recognizing that it will be different from the desired data and will be accompanied by the accumulated uncertainties. In addition to the set of data system parameters, available also are potential changes in the forward and information models (e.g., the user may have to do things differently than first planned if the anticipated real data is too divergent from the data desired or if it will be accompanied by too large errors.) In an informaton-driven system, the information losses allowed will place tolerances on the real data. This requires that the information model be accompanied with a sensitivity analysis. The information/forward model linear programming optimization will allow the user to trade off the various desired parameters requested, allowing for the anticipation of data realities and the influences of the other disciplines on the total information system outcome.

Again, a caveat to the users—this procedure, usually implicit, requires the choosing of sensitivity coefficients, also usually implicit. The user with particularly sensitive requirements had best make his needs known!

Just as the various errors propagate to the end ("downstream"), in an information-driven system the tolerances will propagate upstream. The implication is that, in contrast to the normal single-thread system which requires ever-tighter tolerances in the earlier stages, it may be possible for certain users to pick off data earlier in the data stream before errors have had a chance to accumulate, and for them to do their own processing. This may relieve error tolerances on the remainder of the system.

In addition to the sensitivity coefficients for a single parameter, cross coefficients may be important in analyzing the tradeoffs. Four examples of interdependency of variables are sketched in Figure 7. For example, with Variable A being spatial resolution and Variable B being data rate: Case I (upper left), a user may have lots of computer capability, so that data quantity is no problem, but increasing resolution improves things up to a point after which (say) scatter in the data decreases his performance. In Case IV (lower right) a smaller user finds the same type of resolution optimization, but total data quantity hurts, so that at some point the increase of data with resolution becomes the limiting factor.

As a second example, let the variables be ability to register (in pixels) vs. the pixel size, and let us consider three cases: 1) user doesn't care about registration at all, because he is only looking at a single image; 2) desired geodetic location (in, say, meters) is constant regardless of resolution because the user must register to GCP at the same location accuracy independent of resolution; this user requires that the per pixel registration get better as the pixels get larger; 3) for overlay purposes, the same fractional pixel accuracy is required regardless of the pixel size. These are sketched in Figure 8, together with a hypothetical system peformance. The heavy line in Figure 8 indicates the ridge of optimum performance from the user point of view. The intersection of the anticipated system performance with the user ridge indicates the design optimum.

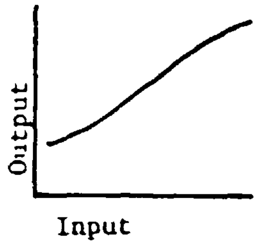
It is realized that the bogie parameters, sensitivity coefficients, and cross-sensitivity coefficients required to do a quantitative system optimization will generally not be available. Nevertheless, these are implicitly defined in the system designer's mind. He will make a mental evaluation of the user forward and information models, and try to decide which parameters are important and which can be slighted, and then proceed to the data system design. Much fundamental research remains to be done to define the forward and information user models and to obtain the sensitivity factors, to allow these to be used in a quantitative total information system design.

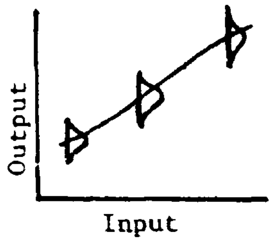
SYSTEM ANALYSIS

Somehow the system design is arrived at, the sensor built and data delivered. The user must then work with whatever data is now available, together with its errors. At that point he has only the information model to vary--that is, he will do whatever necessary to derive the desired information. We will leave him to his troubles, and consider the data system itself. The task at this point is to evaluate the system performance (see Figure 9).

The normal desire in designing a system is that each function be a 1:1 translation, with a change in dimensions only, until finally the "correct measurement" is the same as the "desired measurement". We will therefore model each function as "somewhat linear, with a bias" (figure left), but let the output be produced with some uncertainty (figure right). For a linear system with the input having any possible value with equal probabilty, the probability that the output has a value in a certain range is found by convolving the probability density function of the error with that of the

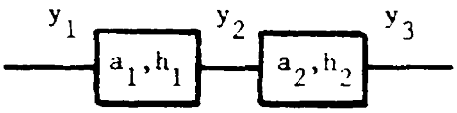
Transfer
Without
Error
Propagation





Transfer
With
Error
Propagation

signal, and integrating between limits representing the range of interest. (Note that the probability distribution function of the error is equivalent, in one dimension, to the point spread function of the image case. The symbol h will therefore be used.) Define the "gain" of each stage as Output/Input = a, so that for a two stage system,



$$y_2 = a_1 y_1 * h_1$$

$$y_3 = a_2 y_2 * h_2 = (a_1 y_1 * h_1)\, a_2 * h_2$$

For constant total system gain $a_1 a_2$, minimum error occurs when $a_1 >> a_2$. This leads to the engineers' old rule of thumb: put as much gain ahead of any noise sources as possible.

If the error sources are Gaussian, the convolutions become root-mean-square additions. However, in the registration case, it is not yet clear whether the error sources are Gaussian, so the RMS addition must be used wih caution. It is also not clear whether an RMS statement of the errors is the one most useful to the user in evaluating the system performance. For example, it may be more important for the user to know where the displacement errors occur (worse in areas of high relief and predictable in direction) than it is for him to know an RMS value (which in itself may be suspect).

It should be noted that a statement of the errors occurring in various parts of the system is of marginal use by itself unless, perhaps, one or more is glaringly bad. Not until the system model is built (implicitly or explicitly) can the error propagation be estimated. During system design, the propagation estimate is used to establish tolerances on the components, and during evaluation it will be used with the actual expected errors to check performance and to identify critical error contributors. After the contributing error sources and their interactions are identified, the following questions may be asked of each source and of the system as a whole:

- What is the intended component performance?
- What is the component actual expected performance?
- How may the performance be verified?
- What correction methods are available for system use?
- What correction methods are available for user use?
- How well can the correction methods potentially work?
- How well may the correction methods actually work?

* How do inaccuracies propagate through the system?
* How do uncertainties propagate through the system?
* Where are the major inaccuracy or uncertainty sources?

Finally, it is to be expected that there may be breakdowns during operation, or that there may be operational problems in performing the component functions. Considering probability of correct operation as a system criterion, the following questions are pertinent:

* What slack is there in the design to allow for problems?
* If a problem occurs, will the system fail catastrophically or gracefully?
* What is the probability that the system will remain up (within specs) for X% of the time?
* How hard does the system seem to be to operate?
* What potential for operator errors are present?
* Are work-arounds for various envisioned errors defined?
* How friendly are the system interfaces to the users?
* Where are the operational bottlenecks?
* Are there any serious single-point failure points?

FINAL POINTS

It can be seen that the various "User Requirements Surveys" have not asked the right questions, or at least have not asked the questions within a milieu to allow the user to respond with the coefficients required by the system designer. The most recent system survey by GSFC has taken a step in the right direction by presenting to the users several potential systems among which the users were to indicate the relative usefulness. But the necessary grossness of the differences prevents any fine tuning of the parameters.

It is not clear that this fine tuning is even possible, given the diversity of users within each discipline, let alone among the disciplines. No plateaus of, say, registration accuracy, have been found beyond which there is a marked loss of utility of the data. The loss of utility with poorer performance has not, perhaps cannot, be stated for the various disciplines. And the aggregation of the losses will produce a loss curve with a gradual slope, with no cliffs.

In the long run, it may well be found that all of the potentially obtainable information is already in the user surveys which are available and that users really cannot define their coefficients, much less anticipate the coefficients of others. This is the "low complexity" end of the spectrum. In this case, the advances in system performance will be more technology driven, and the users must make of it what they will. (In any event, once a system is designed, this is the situation.) It then remains to the system personnel to define what data quality results at various points in the system, and hopefully to allow the users to obtain data of various quality to suit their needs. The system error model will be used to do the evaluation, identify and remove successive error predominant sources, to provide ever-better data to the users, and to serve as a source of information for subsequent systems.

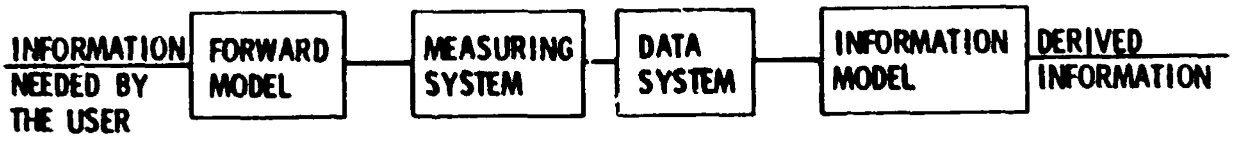


Figure 1. Information System Overall Block Diagram

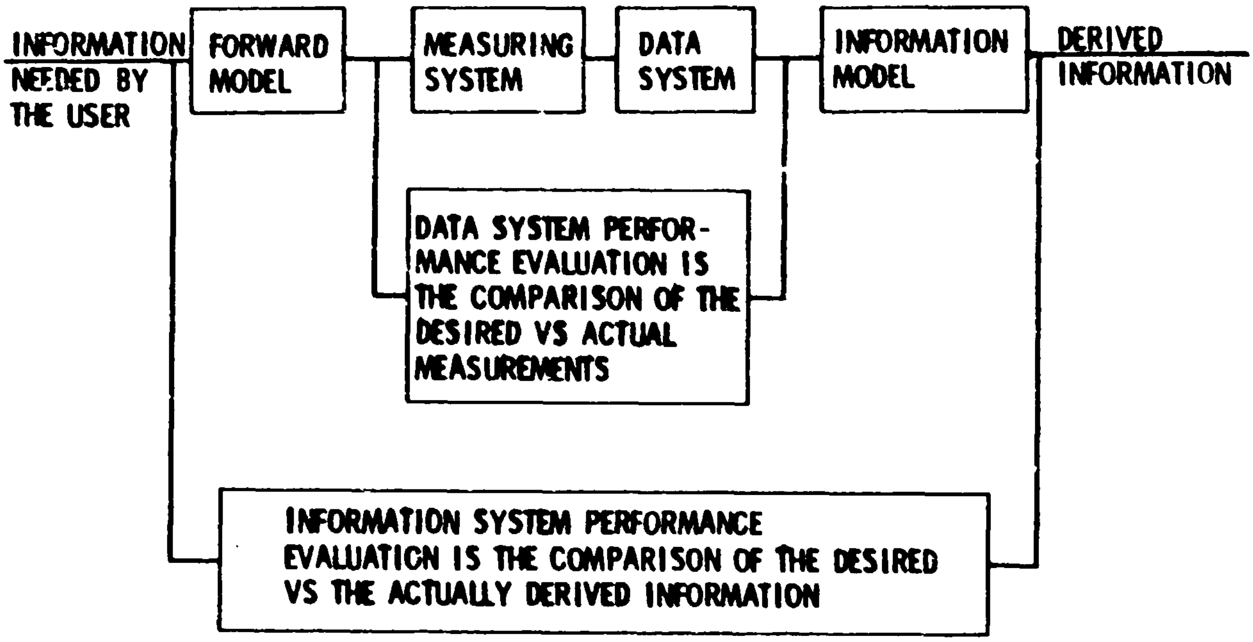

INFORMATION NEEDED BY THE USER
FORWARD MODEL
MEASURING SYSTEM
DATA SYSTEM
INFORMATION MODEL
DERIVED INFORMATION
DATA SYSTEM PERFORMANCE EVALUATION IS THE COMPARISON OF THE DESIRED VS ACTUAL MEASUREMENTS
INFORMATION SYSTEM PERFORMANCE EVALUATION IS THE COMPARISON OF THE DESIRED VS THE ACTUALLY DERIVED INFORMATION

Figure 2. Performance Evaluation at Data System or Information System Level

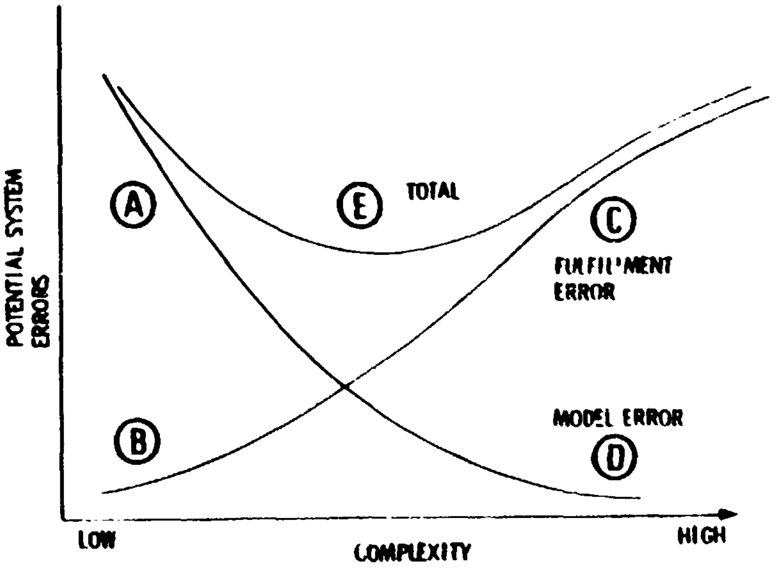


Figure 3. Model and System Complexity Optimization

ORIGINAL PAGE IS OF POOR QUALITY

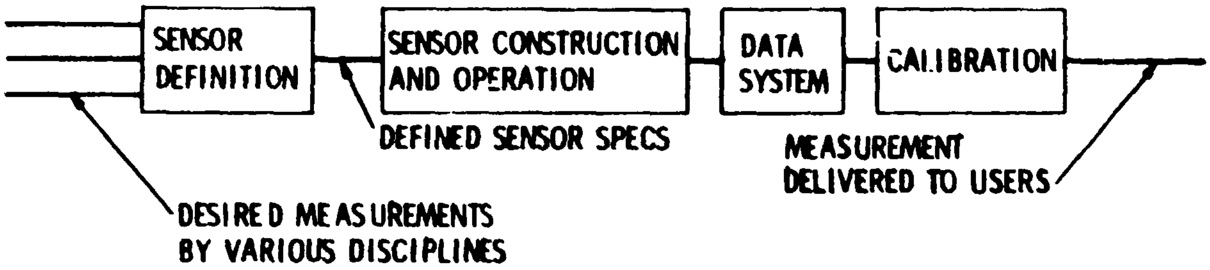

SENSOR DEFINITION
SENSOR CONSTRUCTION AND OPERATION
DATA SYSTEM
CALIBRATION
DEFINED SENSOR SPECS
MEASUREMENT DELIVERED TO USERS
DESIRED MEASUREMENTS BY VARIOUS DISCIPLINES

Figure 4. Data System Block Diagram

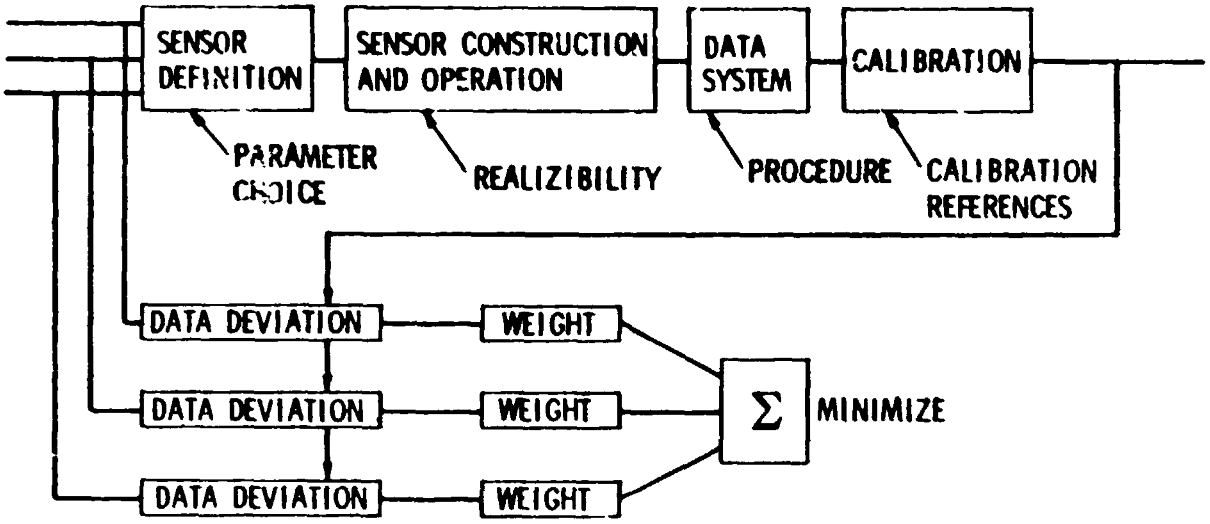


Figure 5. Design Model for Data System Optimization

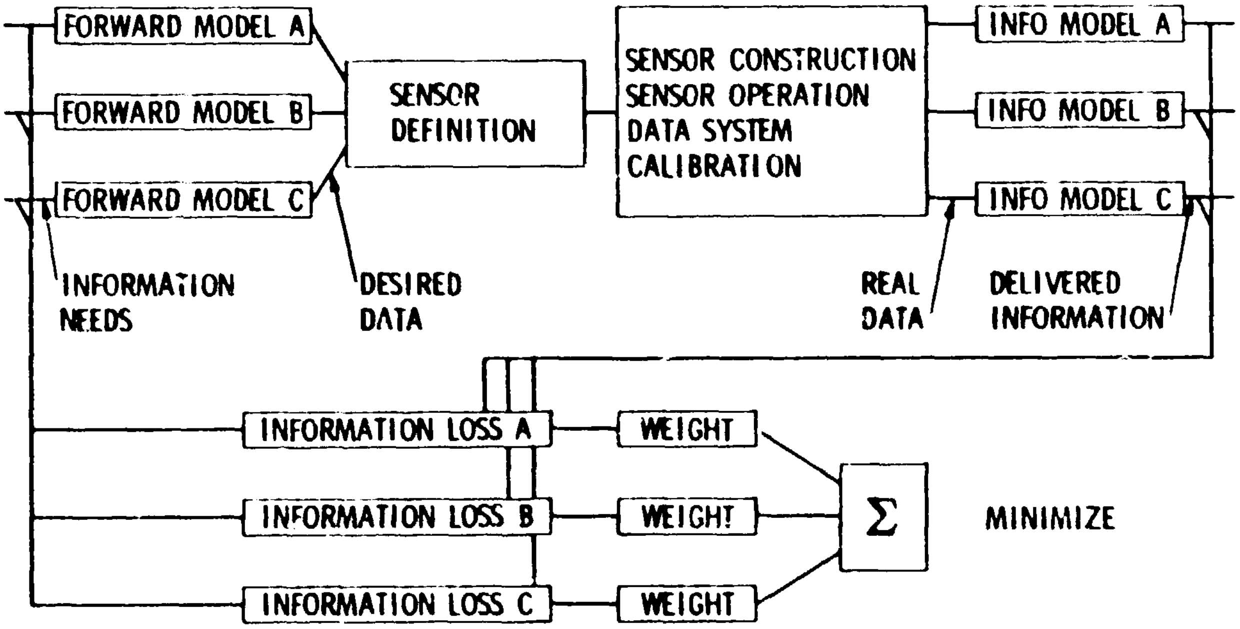
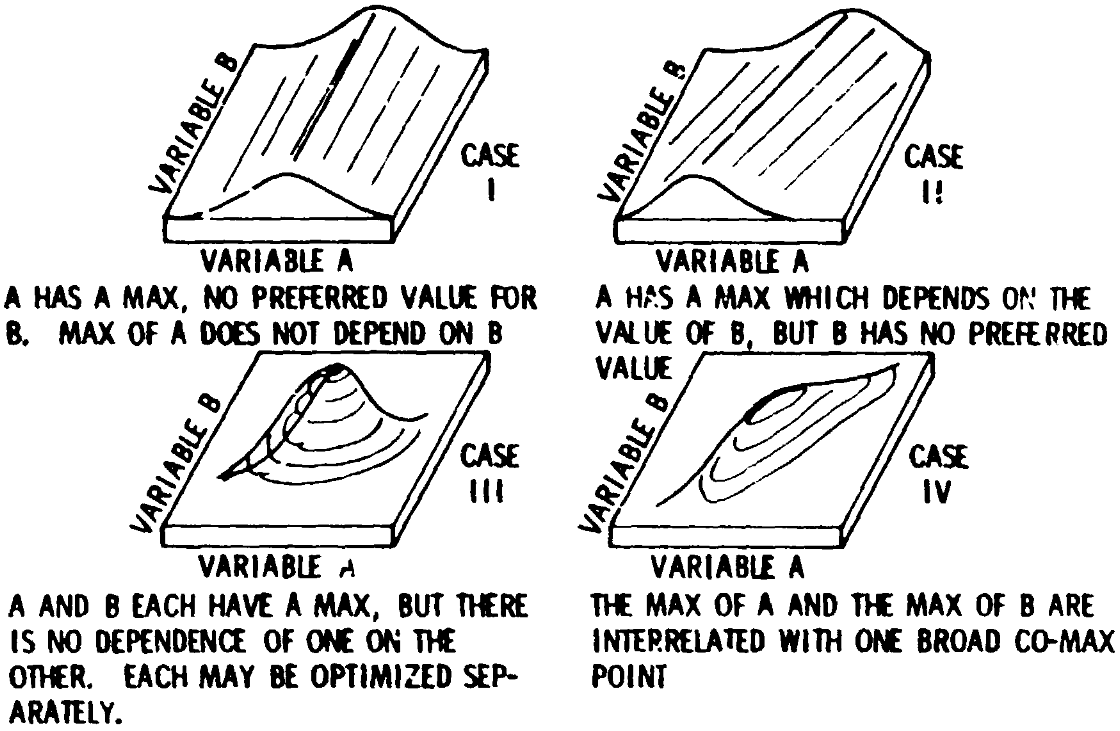


Figure 6. Design Model for Information System Optimization

ORIGINAL PAGE IS
OF POOR QUALITY



A HAS A MAX, NO PREFERRED VALUE FOR B. MAX OF A DOES NOT DEPEND ON B

A HAS A MAX WHICH DEPENDS ON THE VALUE OF B, BUT B HAS NO PREFERRED VALUE

A AND B EACH HAVE A MAX, BUT THERE IS NO DEPENDENCE OF ONE ON THE OTHER. EACH MAY BE OPTIMIZED SEPARATELY.

THE MAX OF A AND THE MAX OF B ARE INTERRELATED WITH ONE BROAD CO-MAX POINT

Figure 7. Interaction of Sensitivity Coefficients

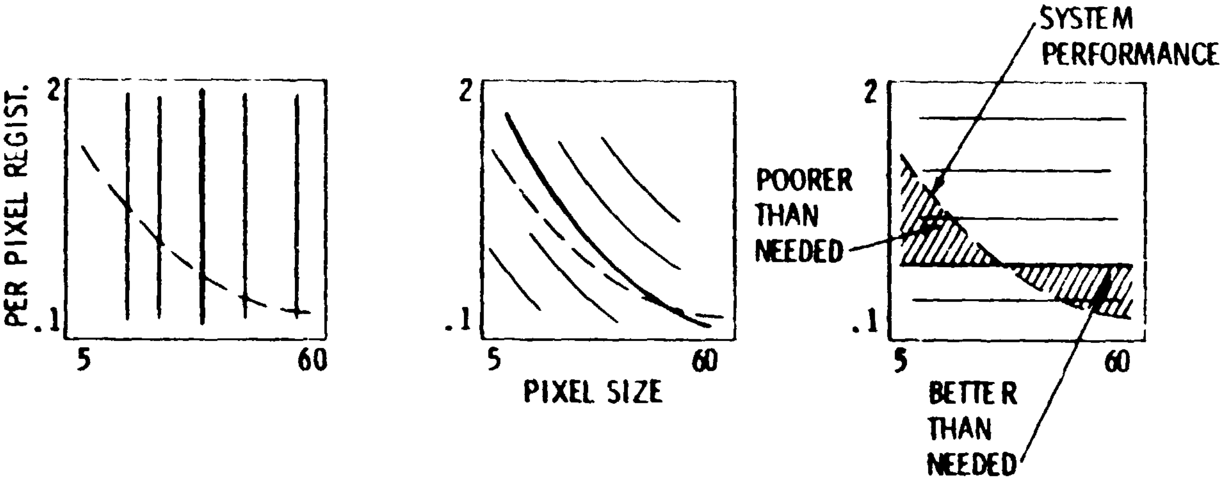


Figure 8. System Operating Point Selection

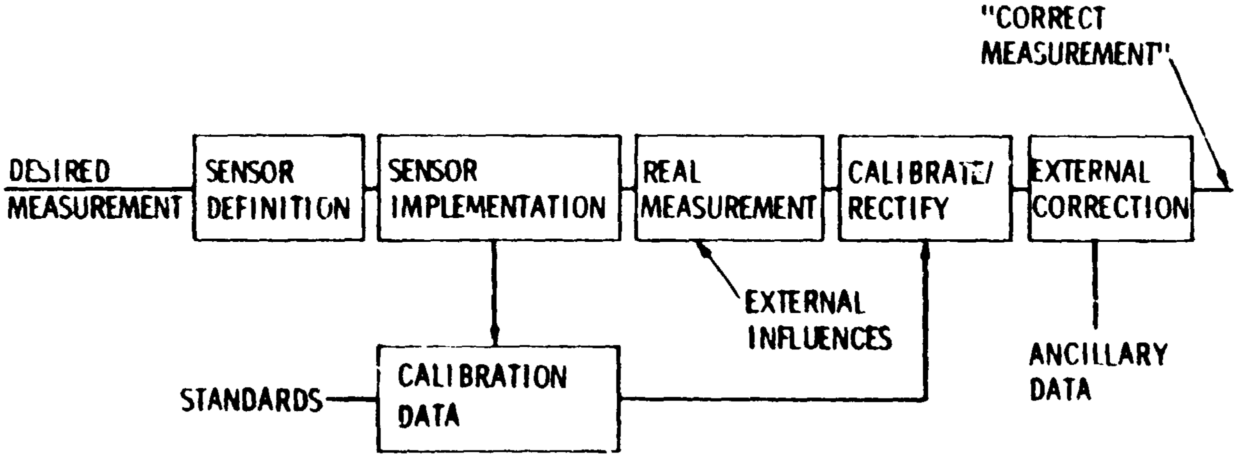


Figure 9. System Error Budget Model