Appendix D
Prediction, Error, and Shewhart's Lost Disciple, Kristo Ivanov

Purpose

This appendix summarizes Kristo Ivanov's thinking on information quality. I have referenced Ivanov in *Measuring Data Quality for Ongoing Improvement*, but his work is not well known in data quality circles. Because it has greatly influenced a range of ideas I have presented, I want to explain it in more detail.

Ivanov is a Professor of Informatics and an expert in systems theory who has written extensively on the social impact of information systems and hypersystems and the wisdom of crowds. He started his career with the question of how we define information quality. His 1972 doctoral thesis, *Quality-control of Information: On the Concept of Accuracy of Information in Data-banks and Management Information Systems*, examines and rethinks our understanding of the concepts of measurement and systems, as well as of data. This document has not been widely cited,[1] and Ivanov made his career publishing on other subjects. But this early document is remarkable in at least three ways: first in how it captures the concerns of its time (the dawn of the Information Age); second in how prescient it is, anticipating the concerns of our time; and finally, because his approach is both practical and philosophical, the document points to questions that many people have not confronted and thus he provides an additional perspective on the challenges of information quality—especially in his redefinition of workaday assumptions about the concept of "error."

Ivanov's ideas are rooted in the same place as Redman's and English's—with Walter Shewhart and the development of quality control in manufacturing. But he follows a different branch of Shewhart's legacy, that of C. West Churchman rather than W. Edwards Deming. Ivanov also draws on C. E. Shannon's information theory, though largely in order to show its limitations as a model for information quality in data banks. Writing before the rise of data warehouses, Ivanov provides a backward glance at information systems as they developed in the 1960s. He not only captures details of the problem of signal versus noise, but also points to the physical aspect of information management (the implications of the use of punch cards, for example) that many of us forget about in a world where much of the data we see never appears on paper.

Ivanov looks forward as well as backward. Despite writing at the dawn of the Information Age, Ivanov raises and explores the same set of concerns raised by data quality thought leaders beginning in the 1990s, including the need for to be cognizant of quality for potential future uses of data (p. 1.5):[2] the idea of an information chain, along which errors may be introduced (p. 2.15); the concept of a set of dimensions or facets of information quality (p. 1.2); the relationship of context to the quality of data and information; the concept of preventing errors; the risks of data misuse if data is not secured and privacy is not defined; and the need for understanding both "subjective" and

[1] Thanks to Linda Hulbert for researching citations related to Ivanov.
[2] Pagination in Ivanov's dissertation includes both a section number and a page number.

"objective" ways of looking at the data (though he does not use these terms—he rejects the idea of the objective observer as a false projection of logical positivism, while allowing for the concept of intersubjective understanding) (p. 4.42); the value of data and the cost of data errors being directly connected to the use of that data (p. 4.23).

Limitations of the Communications Model of Information Quality

In defining the problem, Ivanov quickly recognizes that "the value impact, or more specifically, the economic impact, of quality problems may rapidly increase because of the proliferation of so-called databanks and management information systems" (p. 1.5). Ivanov provides a useful summary of articles pertaining to quality problems in large data banks. Centered on the dimension of accuracy, most of these studies use C. E. Shannon's communications model of data reception and focus on the relation between message sent (input) and message received (output) and "noise" in between that reduces the amount of information transmitted (see Chapter 1 of *Measuring Data Quality for Ongoing Improvement* for a depiction of Shannon's model). They explore not only the physical structure of punch cards but also options for improving the characteristics of codes that are input into computers, as well as for reducing the human factors that contribute to errors in data banks. Ivanov rips through these studies. He points out that the authors' failure to define key terms, such as "quality" and "accuracy," along with slippery and shifting definitions of "error" and, with them, inconsistent ways of measuring errors, prevent him from being able to leverage their findings. Indeed, he points out that the most that they can "conclude" is that the problem of information quality requires more research (p. 2.12).

While the communications model of quality is adequate for describing the technical problem of transmission (the degree of similarity between input and output), Ivanov finds it otherwise inadequate to describe the challenges of information quality. Focus on the technical problem is not going to solve the accuracy problem (p. 3.6) because the accuracy problem also has a semantic component (how precisely the transmitted symbols convey the desired meaning) and an effectiveness component (whether the received meaning affects conduct in the desired way) (p. 2.17).

Ivanov has several reasons for concluding that the communications model does not go far enough to explain information quality. First, he asserts that information systems are more complex than telegraphy and that discussions using the communications analogy do not account for the added complexity. Unlike the messages being sent to a particular place and received individually, information in an information system can have multiple routes and multiple uses, some of which we cannot even anticipate. The assumption that users of such a system have a singular, constant purpose is incorrect. While he does not put it like this, in such a system, separating the message from the "noise" would be very challenging, if not impossible, since different uses of information are looking for messages in different ways—therefore, they are beset not only by different levels of "noise" but by different kinds of "noise."

Second, the communications model is focused on transmission of the message, not the content of the message. From Ivanov's perspective, content—the information itself—is the object of quality, and the content of the message may not be in a usable form to start with. The model further depends on an assumption that any information system is designed (in Ivanov's terms, is modeled) adequately to deliver the information people need, in the form they are expecting it in. Ivanov states that he began his study because he found that many errors in his firm's database turned out not to be conventional

input errors (transpositions, misread characters, etc.) but errors committed in order to keep the system going (work-arounds). This implied there was something wrong with the design of the system, which is another way of saying that the system model is inadequate for the purposes people need the system to carry out.

To work through his concerns about input accuracy and the adequacy of the system model, Ivanov discusses what he calls the relatively straightforward problem of parts inventory at a manufacturing plant. The way that an information system tracks parts for manufactured goods does not always correspond to the way that people on the plant floor track parts. If the two are not aligned—for example, if different parts are stored in the same bin or if suppliers package parts in sets unaccounted for by the system—the people responsible for entering tracking information into the system will work around the limitations of the system in order to keep work moving. Ivanov refers to this as "forcing reality to fit the model" (3.9). Is the information created by such work-arounds inaccurate because of the people who enter it or because of the system that does not allow them to account for it in any other way? He asks a question that cannot be answered outside of the context of the definitions within the system: What is the "true value"? (p. 4.23). His point is that poor system design—an inadequate model—can contribute to low-quality information as much as "human factors" do. He goes so far as to say that the quality of information expressed in error rates "may be an important indicator of the adequacy of system design or of the model. Up to now, it has been regarded as an indicator mainly of the coding [data entry] and observation process itself" (p. 3.13).

The final limitation of the communications model of information quality is that, ultimately, it requires the judgment of a person to determine whether the information received is or is not accurate; that is, an outside observer is needed to understand deviations between predictions and observations to the method of measurement (input), method of processing (model) and method of measuring (output) (p. 4.10). Therefore the communications model does not provide a very good way to measure quality—which is his goal. If it measures anything, it measures the level of "noise" or the degree to which "noise" interferes with transmission (which, indeed, was Shannon's focus).[3]

Error, Prediction, and Scientific Measurement

In order to get at better ways to measure quality, Ivanov revisits the concept of error and introduces the ideas of the prediction, detection, and prevention of errors within an information system. He recognizes that no errors exist without prediction, since logically, errors are deviations between predicted and observed values (p. 4.5). For Ivanov, Shewhart's great breakthrough was establishing scientific-statistical criteria of acceptance to limit or formalize the role of human judgment in determining quality. Shewhart's measurements formalized aspects of judgment as predictions and measurements (p. 4.28). Once established, Shewhart's concepts of accuracy and precision also serve a predictive function: to define the acceptable range of future production. In administrative functions, human judgment performs this function and often does so inconsistently or based on less than adequate criteria (p. 4.14). This predictive function limits the need for human judgment. The function

[3] Ivanov is responding to the way people use Shannon's model, rather than to the model itself in relation to the purposes for which Shannon proposed it. Thus he presents an example of the impact on thinking that results from the misapplication of a model.

of formalizing human judgment is akin to reducing uncertainty (as described by Hubbard, 2010). To the degree that aspects of "judgment" can be documented as mathematical assertions, we can use them to separate the things we are certain about from the things we are uncertain about.

When criteria for judgment are formalized, measurement becomes a means of identifying error (which Ivanov calls "disagreement") that future users of the data can be made aware of (p. 4.31). Instead of searching for accuracy in terms of "truth based on values, efficiency, or facts," Ivanov proposes the development of a criterion of measurable error (p. 4.32). In some cases, disagreement is a measure of the difference between two methods of observing (p. 4.36). In other cases, disagreement is a means of discovering hidden assumptions and thus presents the opportunity for further understanding of the system being measured (p. 4.43). "Truth" then is defined as "agreement established in the context of the strongest possible disagreement" (p. 4.4).

This assertion brings us a long way from the notion of data as "facts" when a "fact" is defined as a piece of information that is "indisputably the case …the truth about events as opposed to an interpretation (*New Oxford American Dictionary*)." It also gives us a relatively abstract notion of truth. Much of the data that many of us use seems not to require this degree of scientific rigor. If my first name is Laura and it has always been Laura, then making sure it is correctly represented does not seem to require "the context of the strongest possible disagreement." And yet …it is not always correctly represented. As you can imagine, bad things happen to "Sebastian-Coleman" all the time. And I have actually had people "correct" me when I tell them, yes, Sebastian-Coleman, both parts, is really, truly my last name and it begins with S and not C. Names are simple examples—even the strongest possible disagreement about them is not likely to be very strong. But when we get to more complex uses of what has traditionally been understood as data—numbers, measurements, calculations, aggregations—we see with greater clarity the risks that Ivanov is concerned about.

One last observation about Ivanov's discussion on quality: Ivanov saw in Shewhart several other ideas that are central to the concept of the "information product" shared by today's thought leaders. Instead of focusing solely on efficiency (output/time period), Shewhart recognized that the output is only output if it is of acceptable quality; that is, if it is produced to specification. If it is not of high quality, then "output" is simply scrap. The same idea should be applied to information. If an information system produces output that does not meet specifications, the information system and its sponsor will likely go bankrupt (p. 4.13). However, Ivanov recognizes that the manufacturing analogy goes only so far, since we do not have physical criteria with which to measure data. The quality of data is understood through activities that use the data. If data does not meet the requirements of those activities, then it is not high-quality data. Any other way of assessing the quality of data presumes a simple relationship between data and a naïve understanding of "facts" (p. 4.31). Ivanov's discussions of error and accuracy have already proven that no such relationship exists. Moreover, Ivanov observes that the same criteria needed for specifying output should also specify input—we should understand what is going into the system as well as what we expect to come out of the system (4.36).

What Do We Learn from Ivanov?

While Ivanov reaches levels of abstraction that are, at times, difficult to grasp, his insights are powerful nevertheless. To start, his review of studies on error rates is a cautionary tale about how to measure and how not to measure. While statistical measurements themselves may give us only an

approximation rather than the exact precision desired by nineteenth-century scientists, the process of measuring should be exact. It requires clear definition of terms and a defined process. People taking measurements should recognize the conditions under which they measure (especially if measurements are taken systematically). Measurement of quality should be based on practical, verifiable criteria. Internal consistency of data provides an option for measurement, but to assess internal consistency requires an understanding of the system within which items should be consistent (4.28).

Next, Ivanov changes our perspective on the concept of error as a form of disagreement, rather than simply an assessment of "correctness." To understand error, one must also understand the terms of the argument in which error is asserted and the position of the observer who concludes error is present. The most important implication of this scientific approach to quality is that quality must be built into systems. If the system and the data are seen as completely separate and separate-able, we risk misunderstanding the data. When we measure the quality of data, we are also measuring the quality of the systems in which it is created and from which it is used. This assertion is not a contradiction of what Codd said, but an example of why what Codd asserted is important. The assertion also does not mean that IT is fully responsible for data quality. Instead, it implies that system design has a direct impact on data quality and system designers need to take data quality into account in their design.

Ivanov's Concept of the System as Model

Ivanov recognizes the need for a full understanding of the information system rather than just the inputs, outputs, and noise. His caution against forcing reality to fit the model is another way of saying, don't believe the things you make up about data. To understand this idea better, it is worth taking a closer look at Ivanov's use of the word "model"—a term he uses somewhat interchangeably with the system itself.

Any system is based on a set of assumptions that can be called its "model." This model is not what we think of as the data model, but rather a paradigm of what the system is supposed to do and how it is supposed to do it. A model is a metaphor that enables us to understand the system. All models are simplifications and all contain assumptions. Each one is driven by its "theory." In sciences, a theory is a formal statement of a belief in prediction aimed at certain goals (p. 4.30). As Ivanov asserts, "every 'fact' implies a theory" (p. 4.31). In science, facts are defined not as things in themselves, but in part by the nature of the observation that collects them (p. 4.33).

In everyday life, we often do not pay attention to the "theories" behind our understanding, but they are there. They often show up directly in our language. For example, the term *data warehouse* embeds a large number of assumptions (a theory) about data and what is done to it, how it fits together, and therefore how it needs to be stored. Contrast this to the theory implied by the term *data bank* which was the common term for large data stores when Ivanov wrote his dissertation. *Banks* and *warehouses* conjure up different images and imply different priorities in relation to what they store. Storing something in a bank is different from storing something in a warehouse.

Following this line of thought, even a single question can be considered a "system" that implies a theory. For example, the following questions all ask for essentially the same piece of information, but because of the way they are phrased and the conventions most people adopt in answering them, under everyday conditions, they are most likely to result in different specific answers—which means that, as systems, they are different from each other:

When were you born?
What is your birthday?
What is your date of birth?

The first is usually looking for a year; the second, for a day and a month; the third, for a day, month, and year. As individual systems, they would look like this:

When were you born? 1776
What is your birthday? July 4
What is your date of birth? July 4, 1776

In conversation, when we get an unexpected answer to a question we simply clarify the question. In technical systems, sometimes we cannot. Systems must be designed to enable people to answer questions and abide by the conventions required to express the answers.

Since systems are designed to hold the answers to questions, it is not surprising that part of systems design comes down to asking the right questions—the questions you actually need to have answered if the system is going to do what you want it to do (requirements) and the questions about the best way to have those questions answered (system design). System design includes asking questions in such a way that they are as distinct as possible from other questions that you also need to have answered. David Loshin's assertion that systems can be mined for enterprise knowledge (particularly in the form of business rules) reflects the idea that business rules are buried within systems (Loshin, 2001). We think of discovering them through data analysis. Another aspect of knowledge mining is understanding what is buried in the design of the system itself—something we do not always think about because we assume that systems are built based on requirements and requirements do not contain "errors". And this is the part that we usually do not talk about. Our understanding of what IT is supposed to do and what the business is supposed to do gets in the way of good system design. IT expects the business to "have" requirements—predefined—ready to be "gathered." The business expects IT to "have" solutions. We stress the idea that the system is fulfilling business requirements—as if there is only one way to fill these. Based on the fact that systems are designed quite differently to meet very similar business needs in different places, there are many ways to fulfill requirements. There is a joke that asserts an elephant is a horse designed by a committee. If you have requirements for a mammal that eats grass, gives milk, and travels in herds, you could come up with a pretty wide range of "solutions."

Ivanov confronts some very large, abstract questions. Ultimately, he is arguing for better system design. It is worth noting that near the end of his dissertation, he proposes the idea of a kind of system evolution—"gradual learning and self-improvement of the information system" (p. 5.38); and in his later work, he has addressed questions related to the wisdom of crowds. What these questions mean for data quality measurement is that there are ways of approaching it scientifically, ensuring that we clearly define what we are measuring, the circumstances of the measurement, the position of the observer, and the hypothesis or prediction or expectation that we are testing. And we should not measure too many things at once. Measurements themselves should be focused and should purposefully answer particular questions. We can use them to separate what we have consensus about from what we have disagreement about.