Puzzles and Peculiarities:

How Scientists Attend to and Process Anomalies During Data Analysis

Susan B. Trickett

Christian D. Schunn

J. Gregory Trafton

Abstract

We used an "in vivo" methodology to investigate how scientists working alone or with a single colleague in informal data analysis sessions deal with anomalous data. We focused on the extent to which several scientists paid attention to anomalies, both immediately and over time, and the processes by which they addressed them. We found that the scientists paid more immediate attention to anomalies than to expected phenomena and also that they continued to devote more attention to anomalies, in some cases over considerable periods of time. We further identified a pattern of response to anomalies that differed from the response to expected phenomena. This pattern included identifying specific features of the phenomenon, proposing a hypothesis to account for it, elaborating that hypothesis by reference to the visual display, and considering the anomaly in the context of the surrounding dataset. We consider how our results mesh with those of previous work and propose some possible explanations for the differences.

Introduction

Carl Linnaeus published the first edition of his classification of living things, the Systema Naturae, in 1735. Shortly thereafter, while having lunch with a colleague at the University of Leiden, he was in the middle of explaining the nature of his classification system, when the colleague stopped him in mid-explanation. A beetle had crawled onto the table, and the colleague wanted to know where this particular type of beetle fit into the classification system. Linnaeus examined the bug carefully, and frowned. Then he squished the bug with a thumb, flicked it from the table, and asked, "What beetle?"

This amusing anecdote, while very unlikely to have actually happened, illustrates one view of scientists: that they are very fond of their own theories and frameworks and not particularly tolerant of data that contradicts the theory. By contrast, philosophers of science have argued that anomalies are crucial to the advancement of science, both in terms of refining existing theories and in terms of developing whole new theories and frameworks (Kuhn, 1970; Lakatos, 1976). Moreover, when asked, scientists themselves believe anomalies are important (Knorr, 1980).

However, the general question for the psychology of science on this topic is how scientists *actually* deal with anomalies, rather than how science *ought* to proceed or how scientists perceive themselves as operating. Psychology has had two general approaches to investigating how scientists deal with anomalies. Researchers using the first approach asked how people respond to negative evidence in a concept identification task. In particular, they asked what sort of evidence they seek for their hypotheses and what they do when the evidence contradicts a working hypothesis (Mynatt, Doherty, & Tweney, 1977; Wason, 1960). Results generally showed (or were interpreted as showing) that scientists are susceptible to confirmation

bias, in that they appear to seek confirming evidence and ignore negative "anomalous" evidence (Mahoney & DeMonbreun, 1977). Additional evidence for this apparent confirmation bias came from sociological studies based on interviews of practicing scientists (Mitroff, 1974).

Researchers using the second approach have focused on practicing scientists performing analyses of authentic scientific data (as opposed to abstract hypothesis-testing tasks), using either historical case studies or "in vivo" observations of practicing scientists as they work. The historical case studies have found mixed results (Chinn & Brewer, 1992; Gorman, 1995; Kulkarni & Simon, 1988; Nersessian, 1999; Tweney, 1989; Tweney & Yachanin, 1985). Kulkarni and Simon found some famous scientists used an "attend to surprising results" heuristic as the impetus for important discoveries (Kulkarni & Simon, 1988; Kulkarni & Simon, 1990), but Chinn and Brewer (1992) suggest there are a range of responses that scientists use, from focusing on the anomaly and changing one's theory all the way to ignoring the anomaly entirely. Tweney (1989) argues that Faraday profitably used a "confirm early, disconfirm late" strategy that might also relate to how scientists respond to anomalous data when it occurs.

Kevin Dunbar pioneered the "in vivo" approach to analyzing the cognitive activities of practicing scientists as they work (Dunbar, 1995; Dunbar, 1997). To date, his emphasis has been on molecular biology labs and the activities that take place in regular lab group meetings. In an analysis of how scientists respond to anomalous data, Dunbar (1995, 1997) found that scientists do indeed discard hypotheses that are inconsistent with evidence, and that they devote considerable attention to anomalies when they occur.

It is important to note that Dunbar's studies deal with reasoning in larger groups working in semiformal settings.[1] The question still remains how individual scientists or pairs of scientists in less formal settings (e.g., as they analyze data in their offices or labs) might respond to anomalous data. Indeed, Dama & Dunbar (1996) showed that important reasoning occurs across several individuals in these lab group settings, such that inductions, for example, are not generated entirely by just a single individual, but rather collaboratively across individuals.

Alberdi , Sleeman, & Korpi (2000) combined the in vitro and in vivo approaches, and conducted a lab study of botanists performing a categorization task. In particular, they presented the botanists with "rogue" items (anomalies to an apparent categorization scheme) and recorded the scientists' responses. They found that scientists did indeed attend to anomalies. Alberdi et al. also identified several strategies by which scientists tried to resolve anomalous data. "Instantiate" was a key strategy. In using the instantiate strategy, subjects activated a schema they had not hitherto considered, which directed them to investigate new features in the data. Thus, they were essentially exploring their theoretical domain knowledge in an attempt to fit the rogue and non-rogue items into a category. They generated new hypotheses about the rogue items that would assimilate the anomalies into their current theoretical understanding of the domain.

The goal of the current chapter is to build on this past work by developing a methodology by which expert scientists can be studied as they conduct their own research in more individual settings. We then investigate scientists' responses to anomalies, both immediately and over time, and determine whether their response to anomalies is different from their response to expected phenomena. This will enable us to ask (and answer) three specific questions: 1) do scientists commonly notice anomalies during data analysis? 2) do scientists focus on anomalies when they

[1] Dunbar also examined lab notebooks and grant proposals, and interviewed scientists in more

are noticed (as opposed to attending to them but not substantially diverting their activities in response to them)? and 3) when scientists examine anomalies, by what processes do they do so?

Method

In order to investigate the issues discussed above, we have adapted Dunbar's in vivo methodology (Dunbar, 1997; Trickett, Fu, Schunn, & Trafton, 2000). This approach offers several advantages. It allows observation of experts, who can use their domain knowledge to guide their strategy selection. It also allows the collection of "on-line" measures of thinking, so that the scientists' thought processes can be examined as they occur. Finally, the tasks the scientists perform are fully authentic.

Participants

Our participants were one individual and one pair of scientists, working on data analysis in their own offices. These scientists were videotaped while conducting their own research. All the scientists were experts, having earned their Ph.D.s more than six years previously. In the first dataset, two astronomers, one a tenured professor at a university, the other a fellow at a research institute, worked collaboratively to investigate computer-generated visual representations of a new set of observational data. At the time of this study, one astronomer had approximately 20 publications in this general area, and the other approximately 10. The astronomers had been collaborating for some years, although they did not frequently work at the same computer screen and at the same time to examine data.

In the second dataset, a physicist with expertise in computational fluid dynamics worked alone to inspect the results of a computational model he had built and run. He was working as a

---

individual settings to make sure focusing on lab groups lost no important activities.

research scientist at a major U.S. scientific research facility and had earned his Ph.D. 23 years beforehand. He had inspected the data earlier but had made some adjustments to the physics parameters underlying the model and was therefore revisiting the data.

All participants were instructed to carry out their work as though no camera were present and without explanation to the experimenter. The individual scientist was trained to give a talk-aloud verbal protocol (Ericsson & Simon, 1993). We recorded the two astronomers' conversation as they engaged in scientific discussion about their data. It is important to emphasize that all participants were performing their usual tasks in the manner in which they typically did so. At the beginning of the session, the participants gave the experimenter an explanatory overview of the data and the questions to be resolved, and after the session, the experimenter interviewed the participants to gain clarification about any uncertainties. During the analysis session, however, the experimenter did not interrupt the participants, and the interactions between participant and experimenter were not included in our own analyses. The relevant part of the astronomy session lasted about 53 minutes, and the physics session, 15 minutes.

All utterances were later transcribed and segmented. A segment consisted of either a short sentence or a clause, if the sentence was complex. All segments were coded by two coders as on-task (pertaining to data analysis) or off-task (e.g., jokes, phone interruptions, etc.). Inter-rater reliability for this coding was more than 95%. Off-task segments were excluded from further analysis. On-task segments (N = 649 for astronomy and N = 173 for physics) were then grouped into episodes, based on visualizations (N = 11 for astronomy and N = 8 for physics). A new episode began when the scientists opened a new data visualization. This grouping of the protocol into episodes allowed us to focus on the more immediate reaction to anomalies.

The Tasks and the Data

Astronomy The data under analysis were optical and radio data of a ring galaxy. The astronomers' high-level goal was to understand its evolution and structure by understanding the flow of gas in the galaxy. In order to understand the gas flow, the astronomers must make inferences about the velocity field, represented by contour lines on the two-dimensional display.

The astronomers' task was made difficult by two characteristics of their data. First, the data were one- or at best two-dimensional, whereas the structure they were attempting to understand was three-dimensional. Second, the data were noisy, with no easy way to separate noise from real phenomena. Figure 1 shows a screen snapshot of the type of data the astronomers were examining. In order to make their inferences, the astronomers used different types of images, representing different phenomena (e.g., different forms of gas), which contain different information about the structure and dynamics of the galaxy. In addition, they could choose from images created by different processing algorithms, each with advantages and disadvantages (e.g., more or less resolution). Finally, they could adjust some features of the display, such as contrast or false color.

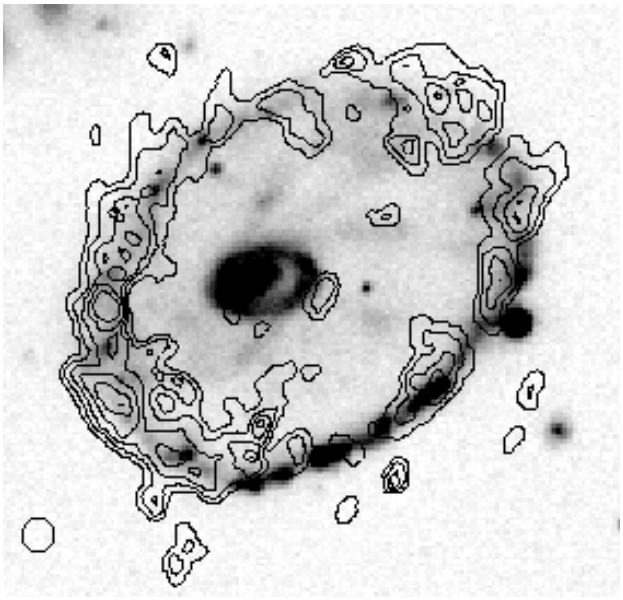

Figure 1: Example of data examined by astronomers. Radio data (contour lines) are laid over optical data.

Physics The physicist was working to evaluate how deep into a pellet a laser light will go before being reflected. His high-level goal was to understand the fundamental physics underlying the reaction, an understanding that hinged on comprehending the relative importance and growth rates of different parts of the pellet and the relationship of that growth to the location of the laser. The physicist had built a computer model of the reaction; other scientists had independently conducted experiments in which lasers were fired at pellets and the reactions recorded. A close match between model and empirical data would indicate a good understanding of the underlying phenomenon. Although the physicist had been in conversation with the experimentalist, he had not viewed the empirical data, and in this session he was investigating only the results of his computational model. However, he believed the model to be correct (i.e., he had strong expectations about what he would see), and in this sense, this session may be considered confirmatory.

The physicist's data consisted of two different kinds of representation of the different modes, shown over time (in nanoseconds). The physicist was able to view either a Fourier decomposition of the modes or a representation of the "raw" data. Figure 2 shows an example of the physicist's data. He could choose from black-and-white or a variety of color representations, and could adjust the scales of the displayed image, as well as some other features. He was able to open numerous views simultaneously. A large part of his task was comparing images, both different types of representation of the same data and different time slices represented in the same way.

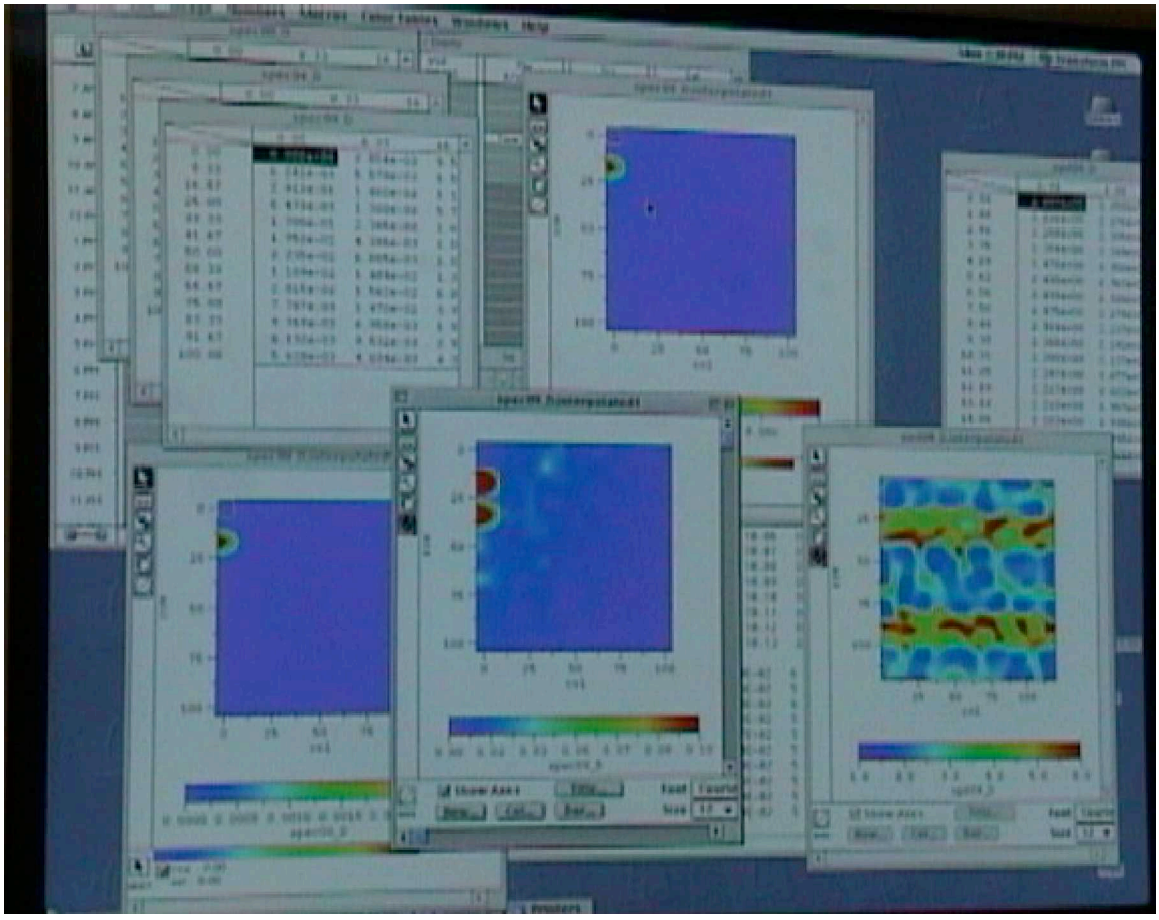

Figure 2: Example of data examined by physicist: Fourier modes (left) and raw data (right)

Coding Scheme

Our goal in this research was to investigate scientists' responses to anomalous data. First, we wanted to establish whether and to what extent the scientists noticed and attended to anomalies. Second, we wanted to investigate the processes by which they responded to anomalous data.

Both protocols were coded independently by two different coders. Initial inter-rater reliability for each code was greater than 85%. Disagreements were resolved by discussion. Any coding disagreements that could not be resolved were excluded from further analysis.

Noticings In order to establish which phenomena—anomalous or not—the scientists attended to,

we first coded for the scientists' *noticing* phenomena in the data. A noticing could involve merely some surface feature of the display, such as a line, shape, or color, or it could involve some interpretation, for example, identifying an area of star formation or the implied presence of a mode. Only the first reference to a phenomenon was coded as a noticing; coding of subsequent references to the same phenomenon is discussed below.

Because our investigation focused on the extent to which the scientists attended to anomalies in the data, we further coded these noticings as either "anomalous" or "expected," according to one or more of the following criteria: a) in some cases the scientist made explicit verbal reference to the fact that something was anomalous or expected; b) if there was no explicit reference, domain knowledge was used to determine whether a noticing was anomalous or not;[2] c) a phenomenon might be associated with (i.e., identified as like) another phenomenon that had already been established as anomalous or not; d) a phenomenon might be contrasted with (i.e., identified as unlike) a phenomenon that had already been established as anomalous or not; e) a scientist might question a feature, thus implying that it is unexpected. Table 1 illustrates these codes.

[2] The coders' domain knowledge came from textbooks and interviews with the scientists.

CRITERION	CODE	EXAMPLE
Explicit	Anomalous	What's *that funky thing*…That's odd Look at that! Ooh! That's interesting. That's very interesting. Okay, what's interesting about this is that *this mode here*…
Domain Knowledge	Expected	You can see that *all the H1* is concentrated in the ring… OK, that was the *initial 2D spike*, all right, the 2D perturbation that we put on the surface
Association	Anomalous	You see *similar kinds of intrusions* along here
Contrast	Expected	That's odd…As opposed to *these things*, which are just the lower contours down here
Question	Anomalous	I still wonder why *we don't see any H1* up here in this sort of northern ring segment? So, I should be seeing *another peak* somewhere in here. And I don't see that there.

Table 1: Noticings (in italics): anomalous or expected

Subsequent References One of our questions was the extent to which the scientists attended to anomalies. The coding of noticings captured only the first reference to a phenomenon of interest; we needed to establish how frequently they made subsequent reference to each noticing. Consequently, all subsequent references were also identified and coded.[3] Not all subsequent references immediately followed a noticing; frequently, the scientists returned to a phenomenon after investigating other features of the data. Subsequent references were identified both within

[3] In the astronomy dataset, frequently the first interaction between them after a noticing was to make sure the scientists were both looking at the same thing. Subsequent references that served purely to establish identity were *not* included in the analyses.

the episode in which the noticing had occurred and across later episodes.

The rest of the coding scheme addresses *how* the scientists responded to the anomalies, in particular immediately after they noticed the anomalies. To investigate the scientists' immediate response to their anomalous findings, we coded ten utterances following each noticing, whether anomalous or expected (excluding utterances in the astronomy dataset that merely established which phenomenon was under discussion). We coded the presence or absence of each type of utterance identified below. We anticipated that scientists would attempt to produce hypotheses for the anomalies, and that some of these hypotheses would be discussed further. Based on the results reported by Alberdi, et al. (2000), we investigated the extent to which elaboration of hypotheses was grounded in theory or in the visual display of the data. We also anticipated the use of additional strategies and inspected the data to identify strategies that emerged, as discussed below.

Identify Features Perhaps because the data the scientists were analyzing was visual, we found that when the scientists noticed a phenomenon, they often elaborated on specific features of the phenomenon that either had attracted their attention or were characteristic of the phenomenon.

Hypotheses Statements that attempted to provide a possible explanation for the data were coded as hypotheses. All hypotheses were further coded as *elaborated* or *unelaborated*. Elaboration consisted of one or more statements that either supported or opposed the hypothesis. Hypotheses that were not discussed further after they were proposed were coded as unelaborated.

When a hypothesis was elaborated, we coded whether the elaboration was *theoretical* or *visual*. When evidence for or against a hypothesis was grounded in theoretical domain knowledge, elaboration was coded as theoretical; when evidence came from the display, it was coded as visual.

Place in context A strategy that emerged from our examination of the data was considering the noticed phenomenon in relation to other data. Thus we coded whether or not the scientist placed the noticing in context, and whether that context was another part of the dataset (*local*) or the scientist's own theoretical knowledge (*global*).

Results

Noticing Anomalies

Our first question was whether the scientists commonly noticed anomalies. Recall that a "noticing" is a first-time reference to a phenomenon. Table 2 presents the total number of noticings for each dataset and the percentages of anomalous and expected phenomena ("not coded" refers to those noticings the coders did not agree on). As Table 2 shows, at least one-third of the phenomena the astronomers commented upon and almost one-half the physicist noticed were coded as anomalous. Thus, out of all the phenomena that caught the scientists' attention, a notable proportion were unusual in some way.

	Total Noticings	Anomalous	Expected	Not Coded
Astronomy	27	33%	48%	19%
Physics	9	44%	44%	12%

Table 2: Frequency of anomalous and expected noticings

Another measure of how commonly the scientists noticed anomalies in the data is the rate at which they commented on an aspect of the data that was anomalous. In the astronomy dataset, the 9 anomalies were identified over the course of 53 minutes, or 1 anomaly for approximately every 6 minutes of data analysis. In the physics dataset, the 4 anomalies were identified over the course of 15 minutes, or 1 anomaly for approximately every 3.75 minutes of analysis. It appears then, that noticing anomalies in the data was a relatively common occurrence for these scientists.

Attending to Anomalies

Once the scientists had identified something unusual in the data, what did they do with this observation? There are several possible reactions, including immediately attempting to account for the anomaly, temporarily disregarding it and returning to it later, or moving on to investigate another, perhaps better understood aspect of the data. One way to investigate this issue is to determine whether their response to anomalies was different from their response to expected phenomena.

We investigated this issue by counting how often the scientists made subsequent reference to a noticing immediately upon identifying it. If anomalies and expected phenomena are of equal interest, we would expect the scientists to make a similar number of references to both the anomalous and expected patterns. However, if anomalies play a more important role in their efforts to understand the data, we would expect them to pay more attention (measured by the number of subsequent references) to anomalies than to expected observations.

As Table 3 shows, for both the astronomy and physics datasets, scientists paid more attention to anomalies than expected phenomena, $t(28)= 3.22$, $p<.01$. In the case of astronomy, the anomalies received over four times as many subsequent references within the same episode as the expected phenomena. The physics dataset follows a similar pattern, with more than three times as many references to anomalies as expected phenomena. The results indicate that the scientists were more interested in phenomena that did not match their expectations, and are thus in stark contrast to the findings of the confirmation bias literature.

Recall that the studies of confirmation bias, even those that involved scientists, involved abstract tasks that required no theoretical or domain knowledge. These scientists we observed, however, were working in their own domain, and, more importantly, on their own data. Thus

they were likely to have strong expectations grounded in deeply-held beliefs that had been built up over years of practice. Insofar as anomalies had the potential to challenge those beliefs, the scientists were likely to have a "real-life" investment in understanding and eventually resolving anomalous data.

	Anomalies	Expected
Astronomy	7.4	1.7
Physics	5.0	1.2

Table 3: Mean number of subsequent references per noticed object for anomalies and expected phenomena *within* the same visualization episode.

Another measure of the attention the scientists paid to anomalies is the number of subsequent references they made to phenomena across episodes. Recall that a new episode was coded when the scientists switched to a new visualization. A new visualization presented the scientists with the opportunity to direct their attention to new phenomena that may have been newly visible (or more salient) in the new visualization. Thus, refocusing their attention upon a phenomenon they had already investigated indicates a relatively high measure of interest on the scientists' part. In both datasets, the scientists continued to pay more attention to the anomalies than the expected phenomena, $t(27) = 2.08$, $p < .05$. As Table 4 shows, both sets of scientists increased their attention to the anomalies over visualizations, more so than to the expected phenomena. This pattern was especially strong in the astronomy dataset.

	Anomalies	Expected
Astronomy	8.3	0.6
Physics	7.7	5.2

Table 4: Mean number of subsequent references per noticed object for anomalies and expected phenomena *across* visualization episodes

Examples of Attending to Anomalies

In both datasets, some anomalies in particular seemed to take hold of and keep the scientists' attention. In these cases, the scientists revisited the anomalies multiple times, and over many different intervening visualizations and noticings. For example, one of the astronomers noticed an area in the galaxy with "a substantial amount of gas" but where no star formation was apparently occurring. The other astronomer proposed two hypotheses that might account for the lack of star formation and then adjusted the contrast of the image in order to allow a better view of the phenomenon. This view allowed the astronomers to confirm the lack of star formation in the relevant area. They then created a new visualization, specifically to allow a closer inspection of the phenomenon. Unsatisfied with what they learned from this visualization, they decided to create yet another visualization ("Let's return to good old black and white") and compared the anomalous area with surrounding phenomena in the galaxy. This comparison led to one astronomer's proposing a new hypothesis, which was rejected by the other. The astronomers then seemed to concede that this anomaly, while potentially important, could not be accounted for at this time:

Astronomer 2: "I mean, it's, it's a strong feature. Can't ignore it."

Astronomer 1: "Can't ignore it. Yeah, well…Gloss over it."

The astronomers then directed their attention to other phenomena in the galaxy.

However, despite both their apparent agreement to table further discussion of this anomaly and their intervening focus on several other phenomena, the astronomers continued to

return to this anomaly several more times ("Getting back to the blob…"; "Note that this, that's the only sort of major blob of H1 that we don't see associated with, you know, really dense pile of H1 we don't see associated with star formation")[4]. Furthermore, they proposed four additional hypotheses about this particular anomaly, considered whether it would be cost-effective to collect more data to investigate it further, and proposed it as the focus of some computational modeling one astronomer was to engage in. The anomaly, originally noticed in the fifth visualization episode, continued to hold their attention into the ninth visualization episode—approximately 20 minutes later.

A similar situation occurred in the physics dataset. The physicist identified an anomalous growth pattern amongst several modes: "Okay, what's interesting about this is that this mode here, while being very large on this plot, at roughly a nanosecond later, is not very big and that while that original seed dominates, we have to go two modes up to actually see a result." He determined that this would be of interest to the experimentalist (recall that the physicist's data derived from computational models, which he hoped would predict the experimental data collected by another scientist). He then opened a new visualization, and noted that the same pattern occurred in this new visualization and proposed an explanation. He concluded, "That's basically what the lesson is here," suggesting that the explanation satisfactorily accounted for the phenomenon. He viewed one more visualization in order to examine a different phenomenon before apparently concluding that the session was complete ("All right, at this point I got to, ah, show these things to our experimentalist and get his experimental data and see how well we agree.") At this point, the scientist began to discuss procedural issues with the experimenter, explaining, for example, what his next steps would be. However, despite having acknowledged

[4] H1 refers to hydrogen gas.

that the data exploration session was over ("Yeah, I think I'm pretty much done"), the physicist abruptly interrupted himself to consider in more detail the anomalous growth pattern he had identified two visualizations earlier. Although he appeared earlier to have been satisfied with his explanation for the anomaly, he expressed renewed interest and continued perplexity: "Was outrun by the next one down. And I don't know, just don't know. I'll have to get someone else's interpretation of that. I don't understand that. The high modes are supposed to take off, they're supposed to run faster." He then proposed and explored a new hypothesis for this discrepant growth rate, but was, in the end, unable to conclusively account for it: " It'll be nice when I, um, show this to the actual experimentalist because they have a better feel for this stuff than I do." Thus, the anomaly continued to hold the physicist's attention even after he had proposed an apparently satisfactory explanation for it. He continued to devote attention to this anomaly across two visualization episodes beyond his original noticing of it and after attending to other phenomena in the data, as well as discussing his future intentions for this dataset.

Immediate Response to Anomalies

In addition to this prolonged attention to several anomalies, we have shown that in general, whenever the scientists noticed an anomaly, they paid immediate attention to it. However, we have not analyzed the content of that immediate attention to anomalies. In order to understand how the scientists dealt with the anomalies upon noticing them, we now turn to the results of the second part of our coding scheme, which was applied to the ten utterances that immediately followed the initial noticing of anomalies and expected phenomena.

Identify Features In approximately half the instances when they noticed a phenomenon, the scientists elaborated on that phenomenon by identifying one or more of its features. However, as Figure 3 shows, the scientists were only slightly (and non-significantly) more likely to identify

specific features of the anomalies as the expected noticings, and this pattern held for both domains.

Although identifying the features of a phenomenon occurred equally for both types of noticing, it seems likely that this strategy served a different purpose for the expected noticings from for the anomalies. For the expected noticings, the scientists seemed to elaborate on the details almost as a kind of mental checklist of the expected characteristics, as though affirming the match between what would be expected and the phenomenon itself. For example, the physicist began his session by plotting a visualization of something he acknowledged that he understood well (a "baseline") and confirmed, "OK, that was the initial 2D spike." He then identified more specific features of this spike, by noting its origin location ("All right, the 2D perturbation that we put on the surface") and finally its nature ("That's basically the sixty micron, uh, the sixty micron perturbation"). This additional information provided a more complete, confirmatory description of the phenomenon itself.

In contrast, when the scientists identified features of anomalous noticings, they seemed to focus on aspects of the phenomenon that were unusual. Early in the session, for example, one of the astronomers noticed an anomalous pattern in the velocity contours ("Look at the little sort of, er, sort of intrusion of the velocity field here…"). He then identified specific features of the "intrusion" by describing it further: "The velocity, the velocity contour goes 'woop.' It sort of dips under, sort of does a very non-circular motion thing. I mean, notice that the, the contours don't sort of cross the ring, they sort of go…." These utterances all serve to elaborate on specific features of the phenomenon that specify how it *deviates* from what the scientist expected.

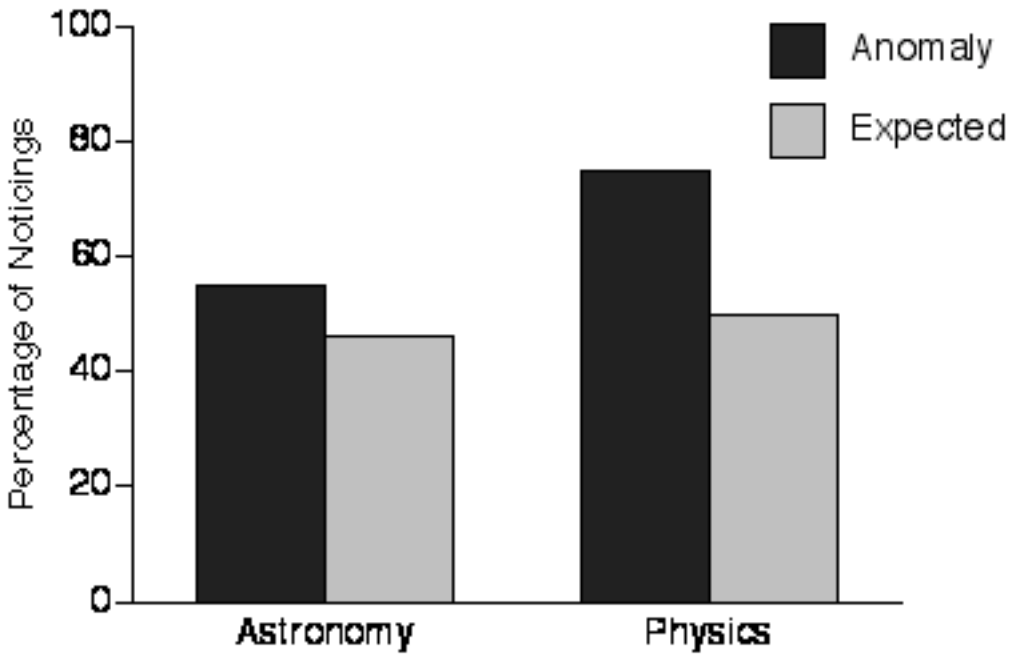

Figure 3: Percentage of noticings for which scientists identified features

Propose Hypothesis When the scientists encountered an anomaly, in most cases they proposed a hypothesis or explanation that would account for it, whereas they rarely did so for the expected phenomena (see Figure 4). This difference was significant, $\chi^2(1) = 7.5$, $p<.05$, and the pattern was very strong in both domains.

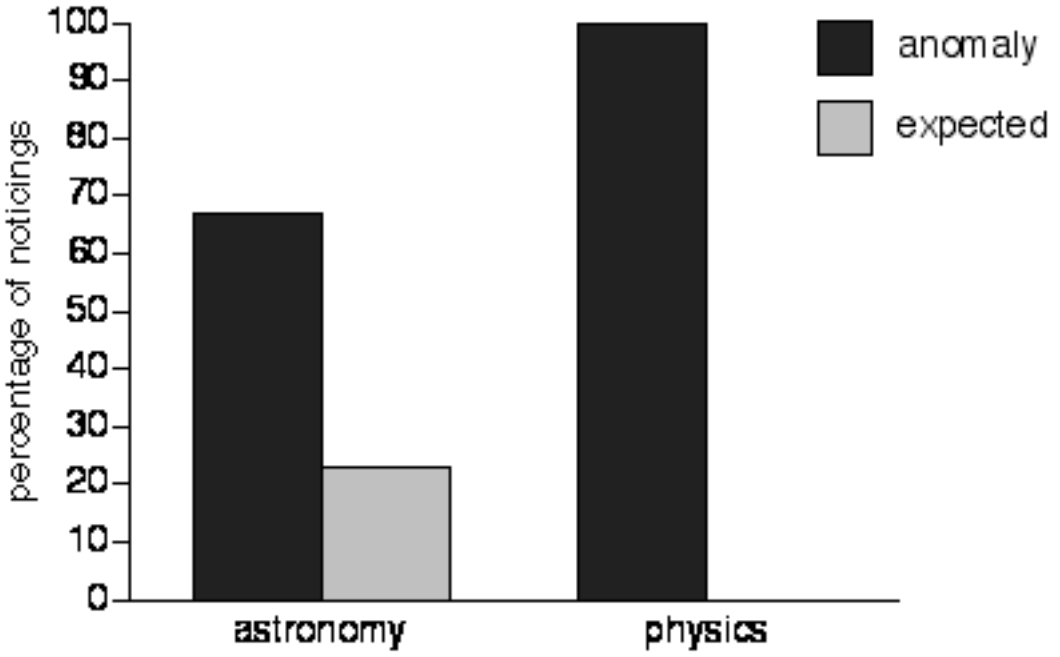

Figure 4: Percentage of noticings for which hypotheses were proposed

As Figure 4 shows, most anomalies were followed within the next ten utterances (that is, almost immediately) by some effort to explain what might have caused them, whether based in the underlying theories of the relevant science or some characteristic of the data collection process or the visualization itself. For example, for the anomalous "intrusion of the velocity field" mentioned above, one of the astronomers asked if this pattern were "significant." His colleague immediately proposed a theoretically-based cause for the "dip" in the velocity contour: "Oh yeah, that's sort of like, um, what can it mean? A streaming motion." Somewhat later, one of the astronomers noticed an anomalous lack of hydrogen gas, where he expected to find it: "I still wonder why we don't see any H1 up here in this sort of northern ring segment." He then suggested, "Ahh. I'm beginning to wonder if we didn't have enough velocity range all of a sudden." In other words, he proposed that the observation itself was not sensitive enough to

detect the hydrogen gas, a hypothesis grounded in the data collection process that would explain why the gas did not appear on the image.

The physicist was also likely to propose hypotheses about anomalous phenomena. After he noticed the discrepant growth pattern among the modes discussed above, he asked, "So what happened? What happened?" He then suggested, "That's the non-linear feature of this, because I added this mode with that mode and got him. The higher coupling coefficient, and as a result he peaks up sooner. And once this guy starts off, he goes very quickly." He thus proposed a hypothesis that pertained to the physics underlying the reaction. In another instance, he failed to detect an expected peak: "So I should be seeing another peak somewhere in here. And I don't see that here." As with the astronomy example above, he proposed an explanation for this grounded in a characteristic of the display, i.e., that the scale of the image was simply too small for the expected peak to be visible: "So let's re-scale. Let's see if I can see it. Because this amplitude may not be very big over here."

The scientists' treatment of the expected phenomena was quite different. For example, one astronomer noted, "Star formation is concentrated in the ring as well" (an expected phenomenon). He elaborated slightly on this observation by indicating the grounds for his comment ("Which is what this color map shows, you've got all the star formation is really concentrated inside the ring"), but made no further comment about this phenomenon. The physicist showed the same tendency not to account for expected phenomena: "Now, clearly we got a fundamental growth right here. Which is the…uh, second harmonic." Again, he identified a feature of the growth ("the second harmonic") but did not propose a hypothesis to explain it.

Hypothesis Elaboration Once the scientists had proposed a hypothesis (primarily about the anomalies), in most cases they elaborated on that hypothesis. Elaboration might consist of more detailed specification of the hypothesis itself or of evidence for or against the hypothesis. Figure 5 presents the proportion of hypotheses that were elaborated within each domain for expected and anomalous noticings. In most cases, scientists attempted to elaborate the hypotheses, for both expected and anomalous noticings (note that there were no hypotheses to elaborate in the expected physics case). As Figure 5 shows, the scientists rarely proposed a hypothesis without considering it further; however, as our next analysis shows, the source of the elaboration was quite different for anomalies than for expected phenomena.

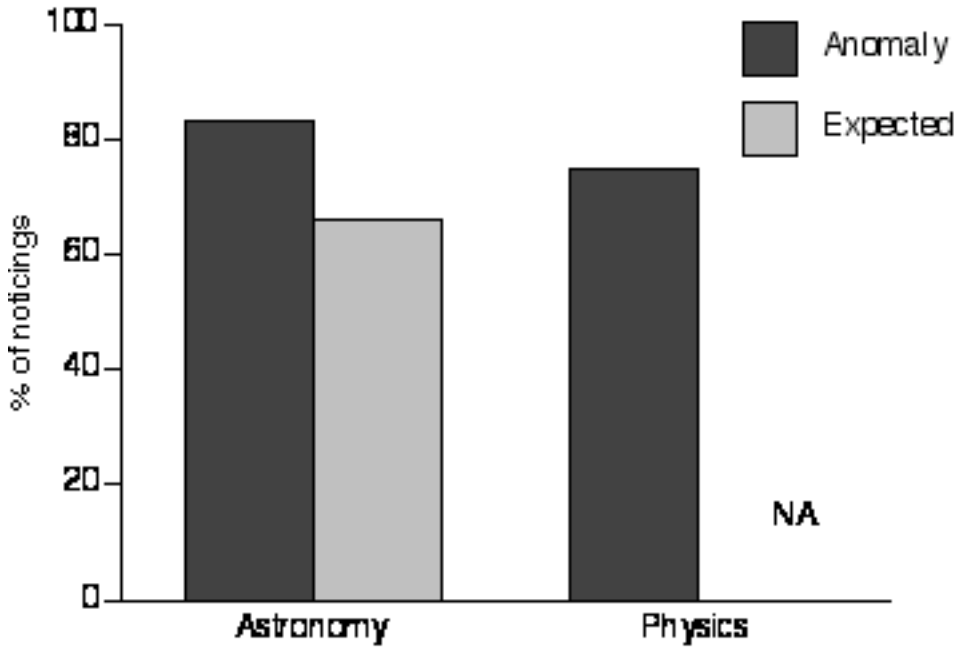

Figure 5: Percentage of hypotheses that were elaborated

Source of Elaboration For the physics dataset, there were not enough elaborated hypotheses to analyze further. However, the astronomy dataset showed a strong contrast in the type of

elaboration the scientists used to further consider their hypotheses, and this difference is illustrated in Figure 6. For the anomalies, evidence pertaining to four of the five hypotheses came from the visual display, whereas the two hypotheses about expected noticings were developed theoretically.

To illustrate the difference in the type of elaboration, consider the following exchange as the astronomers attempted to understand an anomalous pattern of hydrogen gas in the galaxy:

Astronomer 2: "I mean, OK, let's go with er, cosmic conspiracy, think there's another galaxy in the background?

Astronomer 1: "No."

Astronomer 2: "OK. The velocity contours, would they be any different there?"

Astronomer 1: "Noo. Remember the, um, let's see, that corresponds to right about—here."

Astronomer 2: "Well, the velocity contours are doing something there."

Astronomer 2 proposes the hypothesis that another (unobserved) galaxy could account for the pattern of gas and considers whether, if this were the case, the representation of velocity contours in the display would appear different in a specific location (*there*). He looks strictly to the visual pattern to investigate the implications of the hidden galaxy hypothesis. Astronomer 1 rejects the hypothesis (*No*) and similarly focuses on the visual display (*that corresponds to right about—here*). Astronomer 2 continues to focus on the unusual pattern of movement shown by the displayed contours (*the velocity contours are doing something there*). The focus of the whole exchange is the visual display itself, rather than any theoretical issues that underlie the representation.

In contrast, the following exchange occurs as the astronomers discuss an area of the galaxy about which they have agreed, "Yeah, exactly, I mean that's not too surprising,"—in other words, the phenomenon is expected.

Astronomer 1: "That might just be gas blowing out from the star-forming regions"

Astronomer 2: "But that's not a star-forming region, though…that one"

Astronomer 1: "No, absolutely not."

In this case, first the astronomers agree that the phenomenon represents nothing unusual. Astronomer 1 proposes a hypothesis to account for it *(That might just be gas blowing out from the star-forming regions)*, an explanation that Astronomer 2 rejects. He rejects it on the theoretical grounds that the area in question is not a star-forming region, to which Astronomer 1 readily agrees. The astronomers are focused on theoretical domain knowledge rather than on the visual patterns represented on the display.

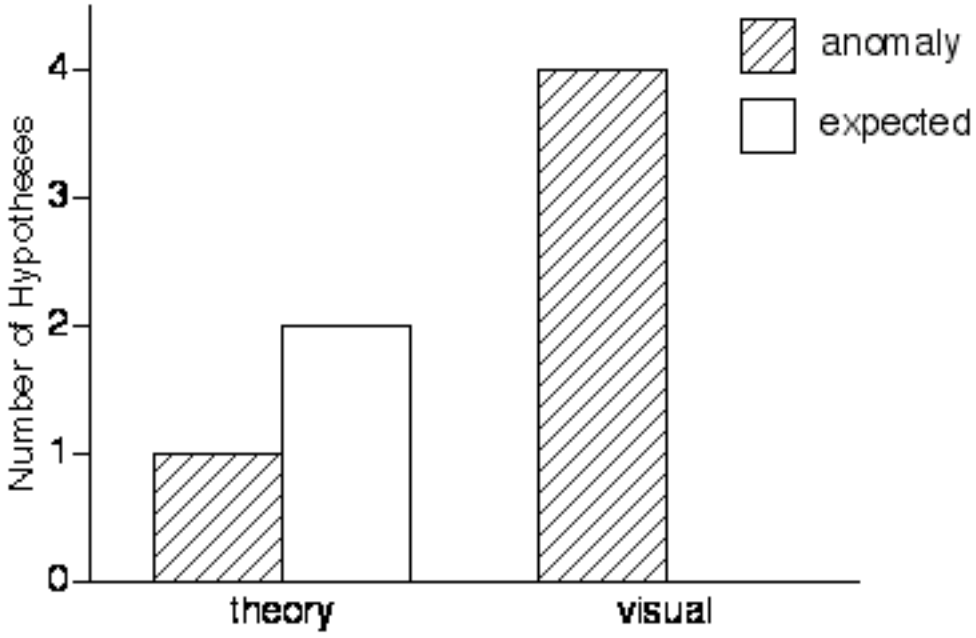

Figure 6: Elaboration type for hypotheses (astronomy dataset only)

Place in Context In addition to (or instead of) developing hypotheses about the noticings, the scientists also might consider the noticing in relation to other information, either theoretical information in memory (global context) or information about the current dataset (local context), or they might not place it in either context. In fact, none of the noticings was considered directly in the context of the scientists' theoretical knowledge (global). However, the scientists considered the noticings in the context of the current dataset (local), and this sequence occurred more frequently for the anomalies than for the expected phenomena, especially in the astronomy dataset (see Figure 7), $\chi^2(1) = 9.21$, $p<.01$.

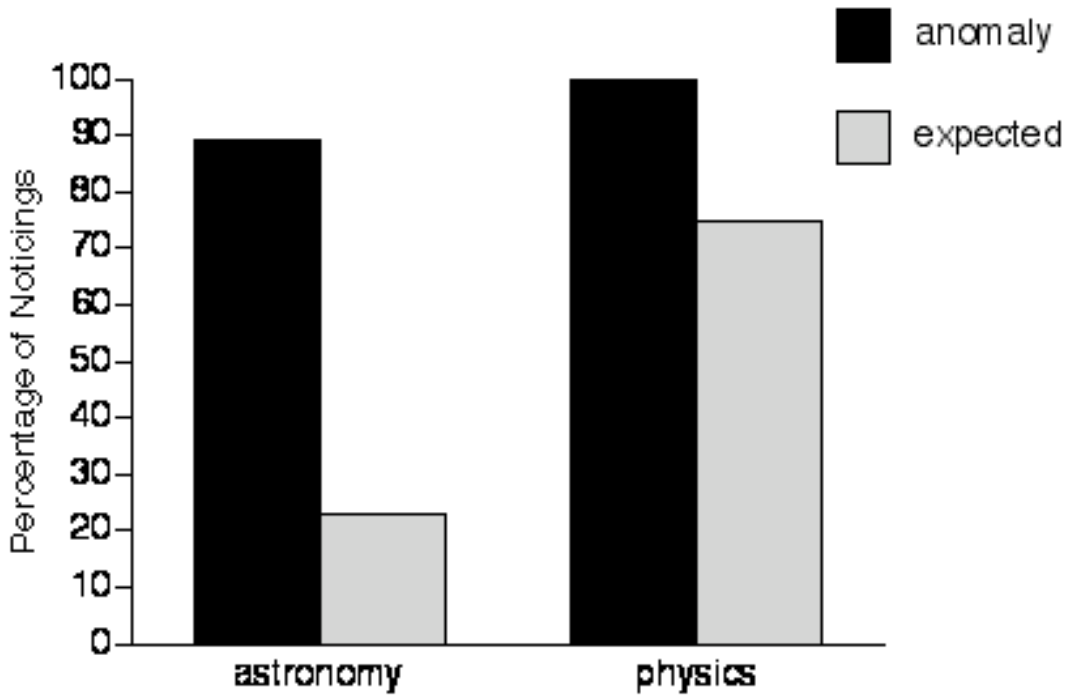

Figure 7: Percentage of noticings put in local context

How did the scientists consider phenomena in the context of information about the current dataset? Again, referring to the anomalous "intrusion of the velocity field" noticed by one of the astronomers, after identifying the features and proposing the "streaming motion" hypothesis discussed above, the discussion continued:

Astronomer 1: "You sort of see that all through this region. You see velocity contours kind of, sort of cutting across the outer ring."

Astronomer 2: "Yeah, it's kinda odd, it seems like the outer ring's…"

Astronomer 1: "And it's only in this region."

Thus, the astronomers made an explicit comparison between the anomaly and the immediately surrounding area of the galaxy, apparently in an effort to fit the phenomenon into its surrounding context and further understand its possible significance.

Although there was less contrast in his response to the anomalous and expected noticings, the physicist also considered the noticings in the context of the surrounding data between the phenomenon and its context. For example, the physicist noticed that the data implied the existence of an unexpected mode: "That means there's somebody in between him. Oh man, oh man." He then considered this implication in the context of the visible modes: "Does that mean that guy's supposed to be…?" before rebuilding the entire visualization and then carefully counting the placement of these contextual modes: "Okay, now. Where is that? [Counting along the visualization axis] That's roughly about thirty. Thirty, right. And this one is roughly about…" In both cases, the scientists appear to be trying to understand the anomaly in terms of the larger local context, that is, the area immediately surrounding it.

General Discussion and Conclusions

We have examined the behavior of scientists at work, analyzing their own data. Our results show that these scientists not only commonly notice anomalies in the data, but also attend to them, both immediately and over multiple visualization episodes. These results are in keeping with the findings of Dunbar (1997) and Alberdi et al. (2000), although they contrast with the results of the confirmation bias literature(e.g., Mahoney & DeMonbreun, 1977).

Beyond corroborating the finding that scientists notice and attend to anomalies, our results also show that these scientists' response followed a particular pattern. Furthermore, their investigation of anomalies was quite different from their treatment of expected phenomena. When they noticed an expected phenomenon, after identifying or describing its features, the scientists were likely to engage in no further elaboration of the phenomenon. On the rare occasions when they did attempt to account for it by proposing a hypothesis, they sought evidence in their own theoretical knowledge, rather than in the visually displayed data. By contrast, however, for anomalous noticings, the scientists attempted to account for the anomaly by proposing a hypothesis. They then elaborated the hypothesis, primarily by seeking evidence in the visual display, and finally considered how the anomaly related to neighboring phenomena within the same display.

One question that arises concerning our results is the nature of the anomalies detected by these scientists. Just as there is a range of possible responses to anomalies, the anomalies themselves also fall along a continuum, from those that are potentially important to those that are ultimately minor. Thus, some anomalies represent major deviations from what current theory predicts or can explain. Such anomalies would require re-thinking the fundamental assumptions

of the science; they have the potential to cause the kind of paradigm shift described by Kuhn (1970) and thus to revolutionize a field. Presumably, however, such anomalies are quite rare. At the other extreme (and much more likely) are less significant irregularities that can be explained within the current theoretical framework. Some anomalies may actually aid the scientist in gaining insight about data or developing theories (see Trickett, Trafton, Schunn & Harrison, 2001, for a discussion of these "framework" anomalies).

In some cases, the scientists in our study were able to resolve an anomaly within the session, and they seemed satisfied with their explanation. Nonetheless, these anomalies were initially puzzling for the scientists, and may have been revisited more than once. In other cases, the scientists could not resolve the anomaly in the session we observed. In one instance, for example, the physicist declared that he was going to have to consult with a colleague to get his interpretation of an especially baffling phenomenon, and in another, the astronomers determined that they would conduct further investigations by constructing some models of a particular anomaly.

Perhaps what is most important to note here, in attempting to generalize our findings to other scientists, is that there was nothing to indicate that the scientists themselves knew in advance whether an irregularity would be accounted for relatively quickly or not. Rather, it was through a process of hypothesizing and contextual exploration that the scientists initially attempted to resolve anomalous results, resorting to other techniques (such as consultation, modeling, or further data collection) only when they were unable to account for an anomaly by these means.

A related issue is the extent to which the scientists questioned the soundness of their data, when they were confronted with an anomaly. Gorman, for example, has found that confirmation

is a useful heuristic when the data may be in error (Gorman, 1986). Penner & Klahr (1996) have also explored the issue of noise, or error, in the data. Penner and Klahr found that people who were warned that some of the data might be incorrect ran replications of suspect experiments, in order to try to evaluate their validity. Performing replications would be costly and time-consuming for both sets of scientists in our study. Nonetheless, one might expect that in general scientists would take anomalies seriously only when they are sure that such anomalies are genuine irregularities in the data, rather than errors or some other form of noise. Given the extent to which these scientists attended to and focused on anomalies, this would suggest that these scientists strongly believed their data to be extremely reliable.

However, this was not necessarily the case. The astronomers made a number of comments about noise in the data. These comments included complaints about the quality of data to be obtained from certain telescopes (in their frustration, they even referred to one as "that piece of garbage"), expressions of awareness about the potential for interference, and doubts about the relative positions of the phenomena they were observing (the problem of representing three-dimensional objects in two dimensions). In addition, in subsequent interviews with the experimenters, they acknowledged that noise in the data is a constant source of difficulty. In contrast, in general, the physicist seemed to believe firmly in his data. Recall that he had constructed a theoretically grounded computational model, and his data were the results of running that model. Thus he had strong reasons to believe in the quality of the data. However, in one instance, he did not detect a mode and, because he was convinced that it should be there, he re-scaled the visual representation of the data until it became visible. In other words, he showed some mistrust of the data when it contradicted his strong expectation. Thus, the scientists were

certainly aware of the possibility of error in the data; however, it appears that they investigated anomalies regardless of their beliefs about the quality of the data.

A final question concerns the issue of working collaboratively versus working individually, and its relationship to noticing and attending to anomalies. It has been shown both computationally (e.g., Clearwater, Huberman, & Hogg, 1991) and empirically (e.g., Okada & Simon, 1997) that working collaboratively can lead to more efficient problem-solving and to more discovery than does working alone, with the appropriate collaboration strategies. One might therefore conjecture that scientists working collaboratively would be more likely to notice anomalies than scientists working individually. It certainly makes sense intuitively to imagine that "two heads are better than one" in this situation, simply because two individuals operate with two slightly different knowledge bases. However, the question cannot be answered by the present dataset. Because each analysis session focussed on a different domain, and hence different data, it is impossible to compare the performance of the two groups. This issue remains one to be addressed by future experimental research.

Our results mesh in part with those of other researchers in that they provide further evidence for the important role played by anomalies as scientists analyze and reason about data. However, our results differ from those of Alberdi et al. (2000) in some significant ways. When the botanists in their study encountered an anomaly, they were most likely to use a strategy of theory-driven search for an explanation. The scientists in our study, however, sought support for hypotheses in the visually displayed data, and attempted to place the anomaly in the local context of neighboring phenomena. Only hypotheses about expected phenomena were developed at a theoretical level.

Thus we found an anomaly with respect to previous work in our own data, and here we attend to it, providing some theoretical discussion of its origins. There are several possible explanations for this difference. Situational differences in the tasks performed by the participants in these two studies might affect their strategy. For the botanists, categorization was the goal *per se*. Although the astronomers and physicist were performing some categorization tasks, this was done in service of understanding the data as a whole, in order to build a mechanistic theory. For example, the galaxy the astronomers were investigating had already been classified as a ring galaxy. Their task was thus not to fit the galaxy itself into an existing classification scheme. Rather, their goal was to construct an understanding of the galaxy's structure and its evolution by interpreting the evidence in the display, a task which itself involved, but was not confined to, categorization. The difference in their goals might account for the different strategies they used. Another possibility is that the botanists were getting immediate feedback on their hypotheses, whereas the other scientists had to generate their own feedback. In this sense, the botanists' task is similar to a supervised learning task, whereas the astronomers and physicist were in a situation where learning was unsupervised (Hertz, 1991). It is plausible that the uncertainties inherent in this situation can account for the fact that these scientists sought feedback in the empirical data in the display rather than jumping immediately to their theoretical domain knowledge. One might expect that later in the research process, the scientists would shift to more theoretical explorations of the anomalies.

Future work, historical, in vivo, or in vitro, should explore this theoretical explanation in further detail. We have elaborated further the account of how scientists attend to anomalous data, providing some insights into the strategies that scientists use for elaborating and resolving these

anomalies. But much work remains, and we expect that the in vivo method will continue to be fruitful in continuing this work.

Acknowledgments

This research was supported in part by grant number 55-7850-00 to Greg Trafton from the Office of Naval Research and by grant number DASWO1-00-K-0017 to Christian Schunn from the Army Research Institute and grant number NOOO14-01-0-0321 to Christian Schunn from the Office of Naval Research. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U. S. Navy.

References

Alberdi, E., Sleeman, D. H., Korpi, M. (2000). Accommodating surprise in taxonomic tasks: The role of expertise. Cognitive Science, 24(1), 53-91.

Chinn, C. A., & Brewer, W. F. (1992). Psychological responses to anomalous data. Paper presented at the 14th Annual Meeting of the Cognitive Science Society, Bloomington, IN.

Clearwater, S., Huberman, B. & Hogg, T. (1991). Problem-solving by committee. Science, 253, 1181-1183.

Dama, M., & Dunbar, K. (1996). Distributed reasoning: An analysis of where social and cognitive worlds fuse. In G. W. Cottrell (Ed.), Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society (pp. 166-170). Mahwah, NJ: Lawrence Erlbaum Associates.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), The nature of insight (pp. 365-395). Cambridge, MA: MIT Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward & S. M. Smith (Eds.), Creative thought: An investigation of conceptual

structures and processes (pp. 461-493). Washington, DC, USA: American Psychological Association.

Gorman, M. E. (1986). How the possibility of error affects falsification on a task that models scientific problem solving. British Journal of Psychology, 77, 85-96.

Gorman, M. E. (1995). Hypothesis testing. In S. E. Newstead & J. St. B. T. Evans (Eds.), Perspectives on thinking and reasoning: Essays in honour of Peter Wason (pp. 217-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). Introduction to the theory of neural computation. Reading, MA: Addison-Wesley.

Knorr, K. D. (1980). Manufacture of knowledge: an essay on the constructivist and contextual nature of science. Oxford, New York: Pergamon Press.

Kuhn, T. S. (1970). The structure of scientific revolutions, 2nd edition. Chicago: University of Chigago Press.

Kulkarni, D., & Simon, H. A. (1988). The process of scientific discovery: The strategy of experimentation. Cognitive Science, 12, 139-176.

Kulkarni, D., & Simon, H. A. (1990). Experimentation in machine discovery. In J. Shrager & P. Langley (Eds.), Computational models of scientific discovery and theory formation . San Mateo, CA: Morgan Kaufmann.

Lakatos, I. (1976). Proofs and refutations. Cambridge, UK: Cambridge University Press.

Mahoney, M. J., & DeMonbreun, B. G. (1977). Psychology of the scientist: An analysis of problem-solving bias. Cognitive Therapy and Research, 1, 229-238.

Mitroff, I. I. (1974). The subjective side of science. New York: Elsevier.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmational bias in a simulated research environment: An experimental study of scientific inference. Quarterly Journal of Experimental Psychology, 29, 85-95.

Nersessian, N. J. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), Model-based reasoning in scientific discovery (pp. 5 - 22). New York: Kluwer Academic/Plenum Publishers.

Okada, T. & Simon, H. A. (1997). Collaborative discovery in a scientific domain. Cognitive Science, 21(2), 109-146.

Penner, D. E., & Klahr, D. (1996). When to trust the data: Further explorations of system error in a scientific reasoning task. Memory & Cognition, 24(5), 665-668.

Trickett, S. B., Fu, W.-T., Schunn, C. D., & Trafton, J. G. (2000). From dipsy-doodles to streaming motions: Changes in representation in the analysis of visual scientific data., Proceedings of the 22nd Annual Conference of the Cognitive Science Society . Mahwah, NJ: Erlbaum.

Trickett, S. B., Trafton, J. G., Schunn, C. D., & Harrison, A. (2001). "That's odd!" How scientists respond to anomalous data. In Stenning, K. & Moore, J. (Eds.),Proceedings of the 23rd Annual Meeting of the Cognitive Science Society. Mahwah,NJ: Erlbaum.

Tweney, R. D. (1989). A framework for the cognitive psychology of science. In B. Gholson, W. R. Shadish, Jr., R. A. Neimeyer, & A. C. Houts, (Eds.), Psychology of science: Contributions to metascience (pp. 342-366). Cambridge, England UK: Cambridge University Press.

Tweney, R. D., & Yachanin, S. A. (1985). Can scientists rationally assess conditional inferences? Social Studies of Science, 15, 155-173.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12, 129-140.