Data processing in observation

Vincent Israel-Jost
IHPST, 17, rue du Four, 75005 Paris France
Extended version of a communication presented at SPSP, Minneapolis, June 18-20 2009
contact : vincent_israel_jost@yahoo.fr

July 31, 2009

Although it has become extremely common for scientists to make use of sophisticated instruments in combination with computational treatments of data in their observational practices, only the former has caused a debate among philosophers. Whether it is appropriate to make use of such instruments as telescopes, microscopes, PET scans, MRI devices and so forth is a question that has received various answers, depending on how the concept of observation is to be understood. The classical empiricist view, that sees unaided perception as the only way that human beings can really learn about facts of the world, has notably been followed by van Fraassen [1] while many other philosophers have defended the idea that instruments can legitimately be used in observation, so long as the processes involved in data acquisition meet some requirements. Hence, Maxwell [2], Shapere [3], Hacking [4] and others have opposed not only to the observational/theoretical distinction that shapes the empiricist view, most strikingly in its logical form, but also to the observable/unobservable distinction. Indeed, to defend that instruments play a role in observation is to acknowledge that what is observable at a given time is a function of our technology and therefore, that no fixed distinction can be established in the long term regarding what is observable and what is not.

Among the defenders of a concept of observation that is not exclusively associated to human perception, it is unfortunate that none has tried to propose a systematic account of the different ways that the data acquired with an instrument are processed before they are actually used to observe some entity or phenomenon. Their analysis

remains of great value, especially in the realism/empiricism debate, as well as in some simpler situations, but it often proves inapplicable when trying to evaluate observation claims made by scientists. Examples of such claims include astronomical images, for which data have been processed to correct for a known defect of the telescope, or medical images, the great success of which relies on their offering a three-dimensional view of body structures. In the latter case, 3D images can only be produced from what is actually recorded by instruments through mathematical algorithms, which transform a set of 2D radiography-like projection images into a 3D volume. Other treatments aim to facilitate data interpretation, by detecting relevant objects in an intricate image for example. Those are commonly used by scientists as datasets tend to be larger and larger. It will be our goal in this paper to present the different kinds of data processing and to evaluate their compatibility with the desiderata of observation.

A minimal notion of observation

In this undertaking, however, it is a most unwelcomed fact that no agreement has been reached concerning what observation is in the first place. From the impossibility concluded by the works of the Vienna Circles members to concile truth and objectivity of observation sentences (see Carnap [5], Neurath [6], Schlick [7]), not much can positively be said, except that we cannot renounce objectivity, since it alone permits to enforce the condition of communicability. As Daston and Galison [8] put it, "if even godlike knowledge of the nature of things fail[s] the test of communicability, it [cannot] be science." In contrast, the truth desiderata of observation sentences is left open and we can either follow Feyerabend [9], positing that observation statements are characterized by mere widespread decidability, or add some milder truth requirements, for example that observation statements should at least be more likely to be true than other statements.

Now, since it is not the place here for debating at length over this and other problems linked to the proper definition of observation, (its association to human perception only, its appropriate language and the forth), we need to find a way to escape these difficulties. To this end, I will formulate the problem discussed in the rest of

paper as follows: *if we assume* that one observes some entity or some phenomenon by making use of an instrument that produces *raw data*, that is, data recorded by the instrument, with no post-processing, what are the typical mathematical treatments that are used by scientists to transform the raw data? And are some of those treatments compatible with observation? Can one still claim to (supposedly better) observe from the processed data the same entity or phenomenon that was already claimed to be observed (supposedly not as clearly) from the raw data?

Asking this question implies that we work with a fuzzy concept of observation, that adapts to one's preferences regarding the problems raised by this very concept. Relativizing the concept will prove quite uninteresting in some cases. The strong perception-centered empiricist thesis, that only accepts sense-data as permitting to observe, will result in no positive cases of observation from the raw data, so long as an instrument has been used, and therefore no possibility for one to observe from the processed data either. So even if the present work will not conflict with this view, the problem that we raise is simply not interesting for its advocates. However, for the philosophers who are sympathetic to the positions expressed by Maxwell, Shapere or Hacking, our work can be thought of as a plug-in to their views, that aims to make their positions applicable in many more situations than in their present state.

We could not go very far however, without giving at least some conditions for an experience to count as observational. If we were to accept that any sort of data can serve in observation, we would reach our conclusions very quickly and any type of data processing, applied to any data could serve in observation. Therefore, we need to give at least a minimal notion of what observation is, that will give us some constraints concerning what is and what is not acceptable in practice. In the rest of the paper, observation will be understood as follows: to observe is to give an interpretation of some data in terms of a real entity or phenomenon. We talk about observing, rather than interpreting, when we can only come up with one interpretation, shared by any qualified observer, which is neither ambiguous, nor dubious. In particular, this interpretation must be independant from the various theoretical views that different observers can endorse, to at least enforce the objectivity of observation. This is where, it seems, there is a general agreement among

philosophers, as we have previously noticed. When data meet these requirements and can be used to observe something, I will say that they are "observational". So if the raw data can be used for the observation of some phenomenon (i.e. are observational), can we still interpret the processed data without contaminating the observation with inferences, theories or beliefs? Are some kinds of processed data observational as well?

Raw data are not pure data

In our formulation of the problem, we contrast the raw data with the *processed data*. By "processed data", we mean the data that the scientists actually use to observe something, and that are created from the raw data, by use of some mathematical transformation. The distinction that I have in mind here, is that between what has been recorded by the instrument (a measurement or a spatially or temporally organized set of measurements) and post-processed data, that is, a new set of numbers computed from the previous one. A very basic example would be to take someones temperature with a thermometer. The number that reads on the instrument is the raw datum, just one measurement. Now if we read carefully the instructions to use the instrument properly, we might find indications to apply corrections, according to the various ways to take the temperature. For example, if the thermometer was introduced in the patients mouth, one might have to add 0.5°C to the raw datum because the oral temperature is known to underestimate the central temperature by approximately this margin. Of course, we could already claim to observe the patients temperature from the raw datum, but we should better claim that we observe the temperature of his mouth in this case, the local temperature. The correction suggested by the instructions is aimed to better approximate the central temperature of the body.

But this example shows that the notion of raw data is not very clear. We can take it to be what is given to us by the instrument, the most basic, least processed data that is available to the observer. If one used a little bit more sophisticated thermometer, that adds 0.5°C when we press a button corresponding to the oral temperature, we would have both the raw datum and the processed datum. But if one knows from the beginning that we will always use the

thermometer in the same way, he can buy one that just displays the corrected temperature and nothing else. Since there is no other data available, this is the raw datum.

We could of course say that even when the unprocessed data are not accessible to the observer, because they have not been recorded, they still exist and have been used to compute some processed data, but it seems unnecessary to argue so. Among scientists, the meaning of "raw data" seems to roughly correspond to what we suggested, namely, *the least processed data available*. We should therefore abandon the idea that the raw data be "pure" in any sense of the term. So the distinction of interest here, is not between "pure" (raw) data and "contaminated" (processed) data, but rather between two sets of data, which are both very likely to be processed in some way, but for which one set of data has been created from the other, by applying an extra mathematical transformation.

The methodology of this paper might then seem puzzling. If we can demonstrate that some processed data are observational because they have been computed from observational raw data with a mathematical treatment that is compatible with observation, the whole reasoning then relies on the assumption that the raw data are observational in the first place. But this appears to be wishful thinking, since the raw data are not even supposed to be free of mathematical processing. So our analysis only applies to a relative difference of mathematical treatment loading between two datasets. As such, it provides a basis for a justification of the observational status of data by recurrence. The question that we are addressing has the form of a recurrence hypothesis: if we assume that a dataset is observational and has has been processed n times, how can we guarantee that the dataset of rank $(n+1)$, that has been processed one more time, is observational? If we can find the conditions for this to happend, then we will just have to assume that some rank of dataset is observational. It is in this sense that this work can be interpreted as a plug-in to the various justifications that have been given by philosophers, that the data acquired with such or such instrument are observational. They provide a justification for the observational status of some rank of data. Let us now take a basic look at how data processing is achieved and its link to simulations.

Data processing and simulations

Computers play a fundamental role in todays practices of observation. It is because most instruments are linked to computer systems that much of the data recorded by instruments have a digital form (a list of numbers). This means that any image that is displayed on a computer screen is actually coded as an array of numbers[1], that can easily be considered as a matrix or a vector. If we call d the vector of raw data, the processed data p are obtained by applying some transform T to d according to the equation:

$$p = Td \tag{1}$$

The operator T is based on a mathematical model that aims to describe something that is relevant to the scientists task. More specifically, the model relies on assumptions which can be associated with three different levels: the object of investigation (that that is being observed), the acquisition of data and the perception of data by an observer. But before we go further into this distinction, it is worth noticing the similarity between data processing and computer simulations. According to Hartmann [10], "simulations are closely related to dynamic models. More concretely, a simulation results when the equations of the underlying model are solved." How then is data processing different from simulations? They are not entirely different, since they cover a number of identical situations, but they can be contrasted by several aspects. First, while it is a requirement for simulations to make use of dynamical models, that is, models that essentially describe a dynamical process, one makes use of any sorts of models (dynamic or static) in data processing. Second, in simulations, one investigates a given phenomenon, and the model is a model of this phenomenon. This is not necessarily the case in data processing, since, as we have already stated, the model can either deal with the investigated phenomenon, or with the processes of data acquisition or of human perception of the data. So from these two aspects, simulations appear like a restricted case of data processing. But there is a third aspect that needs to be stressed,

[1] A distinction must be made here between the data that are recorded by an instrument, stored as list of numbers and the *displayed data*, which are often presented as an image, created from the list of numbers. Several parameters are involved in the transformation of a list of numbers into an image: the colours associated to each number, a scale factor, possibly some rotation parameter to define the most natural orientation, etc. Other types of data display are graphs or the list of number itself.

namely, that any type of data can be put in the dynamic equation of simulations to serve as initial conditions. In data processing, the data we are talking about are characterized by a real connection to their object, they are necessary indicial in Peirce's sense. So while in simulations, we could see how an imaginary population evolves and spreads out on an imaginary planet, data processing is characterized by the use of data that refer to the actual world in a causal way. Therefore, from this third aspect, data processing is a restricted case of simulations.

The fact that non-observational data often serve as initial conditions in simulations is, I believe, essential to the understanding of why, in spite of the interest that simulations have raised among philosophers in the past decade, not much has been said about the observational status of simulations. For there is another general agreement about observation, namely that any observational data must somehow relate, more or less directly, to the real world, and not be entirely computed. In other words, the nature of observational data must be, at least partly, indicial. Hence, the questions raised in the literature dedicated to simulations have dealt with the experimental status of simulations, but not with their observational status. I argue however that it makes sense to study the observational status of processed data, because, although these data can sometimes be seen as the result of simulations, it is a special case of simulations, in which indicial data are plugged into the equation (we have actually assumed that the data d of equation (1) are not only indicial, but even observational). We will now turn to the three different situations that we have identified for data processing, starting with models that describe the object that is investigated.

Models of the observed phenomenon

From the previous remarks, we deduce that this case is the only one in which data processing can also be considered as simulations, but only so long as the underlying model of equation (1) is dynamic. It is therefore natural for us to explore both dynamic and static models of the observed phenomenon separately.

Data that have been processed with a dynamic model of the observed phenomenon are not good candidates to count as observational. The dynamic aspect results in these data serving essen-

tially for predictions, as in the case of eclipses, weather forecast or continental drift for example. Each of these examples has the same structure: we record some data that qualify as observational (planet and sun positions and velocity; temperature, pressure and clouds position; landmasses position), we then apply mathematical transforms to these data, that are based respectively on Newtons dynamic laws in the first example and a highly complex mixture of empirical and statistical models in the second and third examples. But even the data obtained by applying a deterministic law cannot qualify as observable, for we could never say, when looking at the result of a simulation that shows a solar eclipse that will take place in decades, that we are "observing" this eclipse. This is also true for phenomena that took place in the past; it would sound equally awkward and wrong to say that we are observing an eclipse that supposedly took place several centuries ago and that was not recorded by any means, just from looking at simulated data, even if these were created by transformation of real data.

Data processed with a transform based on a static model have much better potential with respect to observation. If we get back to the temperature of the human body, as measured with a thermometer, the fact that we process the datum by adding 0.5°C to what is given by the thermomter (the raw datum) does not seem to work against the idea that we are still observing the body temperature from the processed datum. This correction is based on an empirical law that states that "the temperature in the mouth is (approximately) 0.5°C lower than the central temperature". Since this law is widely accepted among observers, the processed data cannot be said to be attached to some subjective content. It is statistical in nature of course, because the correction would better be 0.4°C for some people and 0.6°C for others, but the important thing is that the correction always improve the observational status compared to the raw data. From this example, I want to suggest that transforming observational data with a static model based on a deterministic law (theoretical or empirical) could be compatible with observation, or at least with our working definition of observation. We are surely making use of a piece of theory here, but if it is completely uncontroversial (to ensure objectivity), deterministic (to ensure some form of certainty) and static (to avoid the possibility to observe what has not yet happened), it seems to be in accordance with the scientific

usage of the word "observation".

Models of data acquisition

This type of data processing is aimed to improve the readability of the raw data, by incorporating some knowledge about how they were recorded and the factors that could contribute to some loss of information. When scientists deal with blurry images, that have been recorded with a microscope for example, post-processing can be applied to correct for this effect. Like any recording device, microscopes cannot record a signal with infinite precision. A compromise has to be found between sensitivity and resolution, since one works against the other. This phenomenon has long been described in the field of signal processing: the resulting blurry image can be seen as a perfectly detailed image, convolved with the response of the detector. So instead of having a point of the object represented by a single point in the image produced by the microscope, a point is represented by a disk, the diameter of which is a characteristic of the imaging device. The relationship that stands between the response of the detector and the blurriness of the image is simple: the larger the diameter of this response, the more blurry the resulting image. If scientists have a model of (raw) data acquisition, they can process these data with the inverse of the blur model. In our case, since the raw data are convolved with some known function, they can be deconvolved with the same function, in order to obtain a sharper image, that shows microscopic features of the object with much more details. So, again, although they can already claim to observe the sample from the blurry image (the displayed raw data), the resulting processed data permit to better observe the sample. But can these processed data qualify as observational, even though one has made use of assumptions in producing them? I believe so, because if we are to consider that the blurry data are observational, then we should ask ourselves how we would interpret them, and I dont see that anyone would read them as sharp images of blurry entities. If everyone interpret them correctly, that is, as blurry images of sharp entities, it means that the knowledge that is incorporated in the deconvolution is actually not new, since we already use it when interpreting the raw data. So my view is that, in most cases, when data processing relies on a model of data acquisition (describing the

transmission of information or the instrument), we are just implementing explicitly pieces of knowledge that we already use implicitly when interpreting the raw data. This is why, I think, reconstruction methods (data processing with a model of data acquisition, which includes deconvolution ("deblurring") or tomographic reconstruction for example) yield data that qualify as observational.

Models of perception

Data processing based on models of perception is currently leading to many developments, with dozens of new papers being published every month in many different journals. This is due to the fact that, while reconstruction methods (data processing based on a model of data acquisition) apply indistinctly to any object that the scientists want to observe, since it is the same thing to deconvolve an image of cells and an image of bacteria for example, data processing based on models of perception adapts to every different problem. This type of mathematical treatments, that we can call interpretive treatments, is aimed to partly automatize the interpretation of the data, for example by running an algorithm that detects relevant objects. This is a set of very precious tools for scientists who deal with extremely large datasets, whether they are physicians, looking for small breast tumors in a stack of radiological images or astrophysicist, trying to detect new stellar objects that could cover only a few pixels in gigabytes of data. But these methods are not only about detection. Very often, the object of interest is very visible but scientists need to measure some of its geometric parameters (diameter, surface, volume, or simply their shape) to classify it. The physician often knows that his patient has a brain tumor, but he needs to check if it has increased or decreased in size after some treatment. So he must draw the contours of what he thinks is the tumor on an MRI scan, which is a very tedious task on three-dimensional data. To accomplish this task, one uses visual perception, detecting objects either by identifying their edges, when the contrast between them and the background is high and sharp, or by seeing them as relatively homogeneous areas characterized by a specific color. Models of perception usually mimick one of these two processes, or sometimes combine them for greater robustness. When implemented within an algorithm, they permit the observer to transform his imaging device - a property detector

- into an entity detector. A CAT-scan device for instance, which detects a specific physical property of matter - that of attenuating X-rays - can be turned into a tumor detector, or a bone detector etc. This not only saves a lot of time, it also generates agreement among observers and permits to obtain reproductible results, which proves impossible even when a single observer manually draws the contour of an object several times. In addition, this type of data processing relies on quite simple and universal models of perception, and should therefore not contaminate the processed data with undesired biases. Yet, I argue that such processed data are generally not observational and are not considered observational by scientists. The reason is that, although their fundamental principle is to implement universal models of perception, they are task-oriented and can therefore only perform well in normal situations, when everything turns as expected. Parameters are set in order to detect a specific class of objects, but these algorithms wont adapt to objects outside of this class, as would the expert that really looks at the image. They are widely used, however, but only in combination with the raw data, so the expert can check that he would actually have seen the same thing as the algorithm. But no one blindly accepts the results given by interpretive treatments. No physician would prescribe surgery because the algorithm, and the algorithm only, has found that the tumor has increased in size since the last scan. This is to be contrasted with the previous type of data processing: scientists do not necessarily look at the blurry image when they have the deconvolved one. We see then that interpretive treatments are reminiscent of our first type of data processing, that makes use of hypotheses concerning what is being observed. We saw that such treatments could be compatible with observation, provided that they enforce some condition of objectivity and of certainty or high probability. This is what interpretive treatments lack: because they are designed to help the observer accomplish a very specific task, they do not adapt well to unexpected situations and are prone to errors.

Conclusion

I have surveyed three types of data processing, characterized by the domain of application of the underlying models, that aim to describe respectively the object of investigation, processes of data acquisition

and the perception of the data. My conclusions are roughly that only reconstruction methods (data processing of the second type) are widely accepted as observational. This means that the resulting data can serve in observation as well as the raw data, if not better. Among the other cases, that of simulations (data processing of the first type that make use of a dynamic model) seems at odd with even a very liberal notion of observation, because it essentially serves to make predictions (or "retrodictions") by making use of a dynamic model. As for interpretive treatments (third type), they need to be highly task-specific to perform well enough and be of any use for scientists, but this make them highly fallible when dealing with unexpected situations, too much indeed to be accepted in observational practices. Finally, I suggested that data processing based on static models of the observed object could be observational, but this requires further investigation, and probably a more detailed analysis than that I was able to provide here. Besides, this part of the analysis is perhaps the most sensitive to the concept of observation that we endorse, so it might prove necessary to give a much more precise definition of this concept before we can move on this point. A final word concerning this paper: I have chosen to limit my enquiry to data recorded by instruments and processed with computers. Although it would have been too ambitious to do so here, I am sure that extending the analysis to perception would be extremely fruitful.

References

[1] B. van Fraassen, *The Scientific Image*, Oxford University Press, 1980.

[2] G. Maxwell, "On the Ontological Status of Theoretical Entities," in *Scientific Explanation, Space, and Time*, H. Feigl and G. Maxwell, Eds., pp. 3–27. Minnesota Studies in the Philosophy of Science, 1962.

[3] D. Shapere, "The concept of observation in science and philosophy," *Philosophy of Science*, vol. 49, no. 4, pp. 485–525, 1982.

[4] I. Hacking, *Representing and Intervening*, Cambridge University, 1983.

[5] R. Carnap, *The logical syntax of language*, K. Paul, Trench, Trubner and co., 1937.

[6] O. Neurath, "Protokollsätze," *Erkenntnis*, vol. 3, pp. 204–214, 1933.

[7] M. Schlick, "Über das Fundament der Erkenntnis," *Erkenntnis*, vol. 4, pp. 77–99, 1934.

[8] Lorraine Daston and Peter Galison, *Objectivity*, Zone Books, 2007.

[9] P. Feyerabend, "An Attempt at a Realistic Interpretation of Experience," in *Realism, Rationalism, and Scientific Method (Philosophical Papers I)*, pp. 17–36. Cambridge University Press, 1985, 1996.

[10] S. Hartmann, "The World as a Process: Simulations in the Natural and Social Sciences," in *Modelling and simulation in the social sciences from a philosophy of science point of view*, U. Mueller R. Hegselmann and K. G. Troitzsch, Eds., pp. 77–100. Kluwer, 1996.