Lecture 18: What Comes First? Data or Theory?

Should science be based only on experience?

Science requires data. Without data, science becomes pure speculation. But should science be based on nothing but experience? Some philosophers argue that this is not possible. Still, the empiricist ideal holds a strong position among scientists and many contemporary philosophers.

One scientific ideal, called radical naturalism, is that science should be a purely empirical matter. This means that anything that goes beyond the data should be considered metaphysical speculations (ontology), and thus not scientific. This would fit with Hume's strict empiricism.

In this lecture, we will look at some problems with assuming that science can be purely empirical. It seems that at least some non-empirical assumptions must be held also in science.

What is the role of data in science?

The type of observation or experience that science is concerned with is called *data*. What role do data play in science? Why do we collect data and how do they relate to scientific theories? There are various views on this:

Data describe or uncover facts. Sometimes data are used to map an aspect of reality: that married men live longer than unmarried men or that low birthweight is linked to adult unemployment. Such statistical data count as scientific results in some disciplines.

Data indicate scientific hypotheses. Once we have the data, this might generate a number of hypotheses. If married men live longer than unmarried men, perhaps it indicates that men are chosen as mates by women because of their genetic advantage. Or perhaps being in a caring relationship is good for one's health. Without the data, we might not even think of asking these questions.

Data are used to test scientific hypotheses. Sometimes we start with a question or hypothesis that we try to test against data. To find out whether genetically modified food is harmful for the environment, one could test the hypotheses against data.

Data provide evidence for hypotheses. Data are required for scientific proof or evidence. The type of data will depend on the scientific discipline and their preferred methods.

Data offer scientific (causal?) explanation. Scientific knowledge is often motivated by questions of *why* and *how*. The data might then count as explanations, showing us what actually happens (e.g. lung cancer from smoking and asbestos).

Data offer a basis for prediction. If something happens regularly and over various contexts, we might use these data to predict future events. We then expect the data set to have application beyond the set itself (external validity). Can we know this without a theory?

Can data determine scientific theories?

We saw that Francis Bacon, with his inductive method, thought that data could generate scientific theories. Given a set of data, only one theory could be the correct one. This might be a necessary assumption for a strict empiricist, since knowledge is on this view thought to come from experience only.

A problem with this view is the problem of *underdetermination*: data do not determine which theory is correct. This means that, given any set of data, various theories could fit and even explain the data equally well.

One then needs *abduction* in addition: inference to the best explanation. But what is the best explanation? That depends on what else we know.

Susan Haack uses the analogy of a crossword puzzle to explain the structure of knowledge. Data has to fit into a coherent system of foundational beliefs. But these might be wrong.

The ideal of raw data

Another empiricist ideal is that our data should be neutral or objective, not influenced by our own expectations. Bacon's inductive method specified this. Scientific observations and data should ideally be free of theories, dogmas and presuppositions.

Note that this is a reason why scientists have to state whether they have a conflict of interest when the results are published.

In his book, *Patterns of Discovery* (ch. 1), Norwood Russell Hanson (1924-1967) argues that the ideal of raw, neutral and theory-independent observation data is an illusion. To observe is theory-dependent, influenced by what we are taught to see.

Example: an expert looking into a microscope has learned what to look for and will understand and interpret it accordingly. Both a child and a layperson can see: they are not blind. But they cannot see what the physicist sees; they are blind to what she sees. Hanson argues that the only way in which an expert and a layperson "see the same" is the retinal response. But this is not seeing: "There is more to seeing than meets the eyeball… The eyeball, like a camera lens, is blind." Seeing is an activity.

In her book, *Data-Centric Biology*, Sabina Leonelli agrees that data are not mind-independent or neutral. 'Data' means 'what is given', but despite of this, she says, data are clearly *made*. "They are results of complex interactions between the researcher and the world". Any evidential value that data might have will depend on the context of scientific inquiry. One should be cautious of using large databases for various scientific purposes without critical consideration of this.

Is everything in science unobservable?

Science often deals with unobservable entities: electrons, genes and photons, are inferred from measurements on macro size objects. And what about abstract entities, such as *history*, *culture*, *ideals*, *time* or *causation*? To gather relevant data, we must first define what it is that we are looking for.

Calling something a protein, RNA, gene or molecule, is already heavily dependent on a theoretical foundation. Without this theory, these are just meaningless terms. How could data about these things then remain *un*influenced by theory?

The strict empiricist ideal of science ignores the fact that a large part of science is non-empirical. This includes:

- Concepts: probability, causation, genes, proteins
- Tools: mathematics, statistics
- Methods: quantitative, qualitative, experimental
- Ontology: reductionism, dualism, determinism

Discussion questions

What is the empiricist ideal of science?

What are the problems with this ideal?

What is problem of underdetermination?

What is the idea of 'raw data'?

What is theory-laden or theory-dependent data?

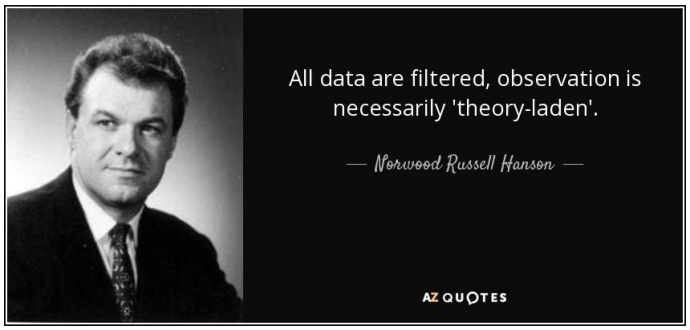
Do you think there could be some raw data, or theory-independent research, even if some data are theory-dependent?

If what we see is dependent on what we expect to see, how could this influence our scientific results?

If Hanson and Leonelli are right in their criticism of raw data, would this mean that science cannot be objective?

What do you think Haack's crossword puzzle view of knowledge means?

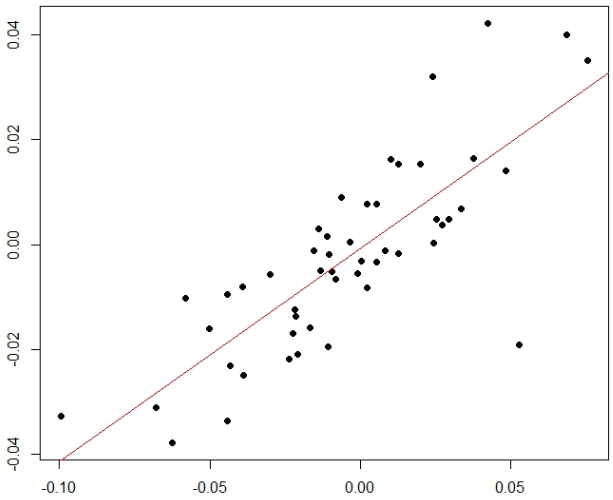
Do you think that it is important for scientists to reflect upon the relationship between data and theory? Why, or why not?



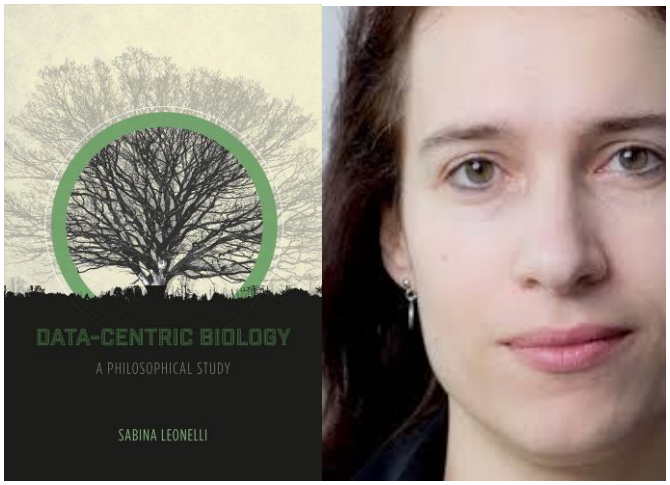
Norwood Russel Hanson was a philosopher of science, and the first to introduce the idea of theory-laden observations.

Susan Haack (1945-) is a philosopher of science and logic.

Hanson argues that seeing as an expert and as a layperson are very different observations. Few experts could find out what was wrong with this lung scan. Can you?



Sometimes our statistical models influence how we interpret the data. Why do these data form a straight line rather than a zigzag pattern?



Sabina Leonelli works on the research processes, scientific outputs and social embedding of Open Science, Open Data and Big Data.