Models of Scientific Discovery: What Do They Explain?

Albrecht Heeffer

albrecht.heeffer@ugent.be

Ghent University
Center for Logic and Philosophy of Science
Blandijnberg 2, B-9000
Ghent, Belgium
(tel. +32 9 2653979)

Abstract

Motivated by the renewed interest in knowledge discovery from data (KDD) by the artificial intelligence community, this paper provides a critical assessment of models of discovery in science. The most influential research program, called BACON, is using the model of data-driven induction. Two of the main claims by this research program, the descriptive and constructive power of data-driven induction, are evaluated by means of two historical cases studies: the discovery of the sine law of refraction in optics and Kepler's third law of planetary motion. I will concentrate on the first claim that the computational model used by BACON provides an explanation of the real historical cases. The explanatory power of BACON's model will be challenged by providing evidence that the data used by the program – despite the claims being made by the authors – does not correspond with the historical data available to Kepler and his contemporaries. Furthermore, it is shown that for the two cases the method by which the general law was arrived at did not involve data-driven induction. Finally, the value of the data-driven induction as a general model for scientific discovery is being questioned. An outline of an alternative model of hypothesis generation and testing will be offered for one of the case studies.

Introduction

Any rational approach to scientific discovery has to deal with ampliative reasoning. One of the most important ampliative mechanisms to expand our knowledge about nature is inductive reasoning. Since Hempel (1945) however, many philosophical problems with induction have been formulated. Some have dismissed the role of induction in science altogether: "A theory of induction is superfluous. It has no function in a logic of science" (Popper, 315). After Popper, the distinction between the context of discovery and the context of justification has been used by many to deal with the latter only and to move the difficult problem of induction outside the realm of scientific explanation. From the late seventies onwards, cognitive scientists and researches with the domain of artificial intelligence (AI) started formulating models for dealing with the context of discovery. Based on the seminal work by Herbert Simon (1972), creativity, such as scientific discovery, was approached as a rational process – using the same kind of mechanisms as one does in puzzle solving. Most of the research on scientific discovery focused on one particular kind of induction, named data-driven induction. The central tenet of data-driven induction in science is that scientists discover quantitative laws of nature by a process of inductive generalization from observational data. Using Simon's model of a goal-directed state-space search, the problem of discovery in science was reduced to finding the right heuristics in a general, generic and domain-independent model of problem solving. Thomas Nickles (1994) formulated this approach pertinently as "the neo-enlightment counterpart of universal Reason, a faculty that could in principle solve any (solvable) problem in any domain". Several systems have been build to model such processes the most influential one has been BACON, authored by Pat Langley, Gary Bradshaw, Jan Zytkow and Herbert Simon. This research program spanned a period of 1978 to 1990, involving research groups from several universities and resulting in many publications. The most representative overview is a book titled *Scientific Discovery. Computational Explorations of the Creative Processes*. (Langley e.a., 1987). In the rest of this paper I will mostly refer to this publication although my arguments also apply to other KDD approaches which endorse the model of data-driven induction.

Aims and motivation

My aim is to show that the approach taken by the BACON team is not an adequate one for explaining and modelling scientific discovery *in general*. This is exactly the claim made by the BACON team: not only is their book "concerned more with describing and explaining scientific discovery than with providing a normative theory of the process", they also call their model constructive, "it exhibits a set of processes that, when executed, actually make scientific discoveries" (Langley e.a. 1987, 7). I will present here a critical assessment of these two main presuppositions, 1) their model describes and explains historical cases of scientific discovery and 2) the BACON program is sufficient to make scientific discoveries. I will do so, not so much from the viewpoint of philosophy of science but more from a contextual historical perspective. We will focus on two "successes" of the BACON program: the rediscovery of Kepler's third law and the sine law of refraction.

Before discussing the main criticisms against these claims, I have to make a disclaimer. The arguments discussed below deal with one particular form of inductive reasoning: data-driven induction. In no way is it implied that inductive reasoning has no function in scientific discovery. Some arguments will focus on the use of observational data by the BACON program. Neither is suggested here that observational data does not play a significant role in the formulation of general quantitative laws. The possibility of inducing general laws from empirical data is not questioned either. Even the claim that BACON is able to find some laws from data fed to the program is not at issue here. Furthermore, I fully acknowledge the importance of the BACON program towards a rational explanation of the process of scientific discovery and endorse the main starting point that creativity and discovery in science can and must be explained as problem-solving processes.

One more question has to be answered. What is the relevance today of formulating methodological objections against the BACON program? While a simplified reduction of scientific discovery as data-driven induction has been mostly dismissed in currently prevailing philosophy of science, the idea is still very prominent in recent artificial intelligence research. Two new technologies have given a new impetus to data-driven induction: data mining since the early 1990's and knowledge discovery from data (KDD). In 1995 a Special Interest Group on Knowledge discovery and Data Mining (SIGKDD) was created. In 1999 this group founded a new journal on this topic. Several new conferences were created to communicate new research results within this domain: Research Issues on Data Mining and Knowledge Discovery (DMKD), the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), and The annual Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKKD). The Langley book is often quoted within these scientific communities. ISI Web of Science lists 290 citations, Google scholar 841 articles. These provide us reasons enough to reassess the philosophical and methodological foundations of the program.

The descriptive and explanatory power of the BACON program

While early research in computer chess was motivated by the study of human problems solving (as with Newell and Simon, 1972) the domain rapidly developed into a race to beat the best humans in chess. It became clear that the calculating power of computers took more benefit from the model of state-space search than humans did. In computer chess there is a tradeoff between the amount of domain-dependent knowledge you add and the search depth. Chess knowledge in computer programs is kept to a minimum in order to gain some more moves in search depth. Current programs reach up to 20 ply (halve moves) in middle-game positions, which is sufficient to beat the world champion in chess. Place this against the famous quote by the Cuban chess grandmaster Capablanca when asked how many moves he can think ahead: "Only one move ahead – the right one", he replied. The moral of this story is that computer chess works very different from the cognitive processes involved in human chess. Therefore, chess programs do not explain much about the cognitive capacities of chess masters. In other words: computer chess is no model for human problem solving in chess.

Let us now move to the question if BACON provides an adequate model for discovery in science. The first discussion point is whether the BACON researchers intended to produce a model of scientific discovery. The answer is affirmative. Not only is their intention to provide a model for scientific discovery in general, they also claim that it explains the historical cases covered in the book. It is not without some pride that Langley e.a. (1987, 340) conclude their presentation of BACON with the words "We would like to imagine that the great discoverers, the scientists whose behavior we are trying to understand, would be pleased with this interpretation of their activity." Unexpectedly, they have also claims about the amount of time scientists spend on the activities of data-driven induction, modeled by the BACON program: "Although we have no quantitative data on which to build an estimate, it is reasonable to suppose that the induction of laws from data might typically occupy *only* from 1 to 10 percent of a scientist's time" (Langley e.a., 1987, 112;

emphasis mine). Since quantitative laws of nature are discovered only so often, even the lower estimate of 1 percent seems to be a very high occurrence of data-driven induction by scientists.

In the next section I will assess to what degree two of the successful discoveries by BACON contribute to an understanding of the actual historical discoveries: the sine law and Kepler's third law. There are two reasons for selecting these two cases. The first is that both are related to Johannes Kepler. Kepler is a grateful subject for any historical study. In contrast with his contemporaries as Descartes he abundantly documented his line of reasoning including his failures as well as his successes. While Kepler did not succeed in formulating the sine law he came very close to it and we may suspect that others came to the discovery following the same lines of investigation. The second reason is given by the BACON team itself. According to Langley e.a. (1987, 224), the sine law is one of the "three instances where, in the actual history of the matter, the data – essentially the same data that were available to BACON – were interpreted erroneously before the "correct" law was discovered". The other two are Kepler's third law and Boerhaave's law. The fact that scientists fail to arrive at a quantitative law when the correct data is available to them appears to be an interesting phenomenon. If only data-driven induction is at play, such failures should not happen. Langley e.a. (1987, 224) explain it as follows:

> In all of these cases, the error arose from accepting "loose" fits of a law to data, and the later, correct formulation provided a law that fit the data much more closely. If we wished to simulate this phenomenon with BACON, we would only have to set the error allowance generously at the outset, then set stricter limits after an initial law had been found.

Such claim implies that BACON – and the model of data-driven induction – not only can explain the successes of scientific discovery but the failures as well. This will be taken into consideration in a later section.

BACON's "discoveries"

The sine law of refraction has been "rediscovered" by two versions of the BACON program. BACON.4 was given as input a table of 9 data lines which contains the sines of angles of incidence and refraction for three angles and combinations with three media (see Table 1).

Table 1: Input data for the BACON program

Medium1	Medium2	sin A_1	sin A_2	sin A_1/sin A_2
vacuum	vacuum	0.500	0.500	1.00
vacuum	vacuum	0.707	0.707	1.00
vacuum	vacuum	0.866	0.866	1.00
vacuum	water	0.500	0.665	0.75

vacuum	water	0.707	0.940	0.75
vacuum	water	0.866	1.152	0.75
vacuum	oil	0.500	0.735	0.68
vacuum	oil	0.707	1.039	0.68
vacuum	oil	0.866	1.273	0.68

The program “discovers” that the ratio between the two sines is invariant for a given combination of media. Against the possible objection – which I will raise anyway in the last section – that giving the sines as input to the program is given away the discovery, Langley e.a. write (1987, 142, note 5):

> BACON.4 does not have heuristics for considering trigonometric functions of variables directly . Thus, in the run described here we simply told the system to examine the sines. In the following chapter we will see how BACON can actually arrive at the sine term on its own in a rather subtle manner.

This subtle manner appears to be that instead of the sines of the angles BACON.5 is given the distances the hypotenuse and the opposite side of a rectangular triangle. But as we all know from secondary school, the ratio of these two parameters amounts to exactly the same.

The other discovery is Kepler’s third law, which describes a relation between the average distance of a planet from the Sun and the time of its orbit. This was discovered by Kepler shortly before finishing his *Harmonices Mundi*, published in 1619. Newton later proposed a more general form of the law applying to any two objects orbiting around a common center of mass. Langley e.a. (1987, 56) point out that in both cases the discovery involves a single step of data-driven induction: “The important point to notice here is that, in discovering the relation between acceleration and distance , the only step of induction is the inference of Kepler’s third law from the observation – a data -driven induction”. The “rediscovery” of the law was taken as a task for the first version of their program BACON.1. The input data consisted of three lines with the two relevant parameters (see Table 2). The program “discovers” that the expression D^3/P is invariant for the input data.

Table 1: Input data of BACON.1 for discovering Kepler's third law

Planet	Distance (D)	Period (P)	D/P	D^2/P	D^3/P
A	1.0	1.0	0.500	1.0	1.00
B	4.0	8.0	0.707	2.0	1.00
C	9.0	27.0	0.866	3.0	1.00

The authors admit that the input data “were contrived to fit Kepler’s law exactly” (Langley e.a., 1987, 69). The problem of early modern astronomy was that observational data did provide only approximate measures of distance and period. If BACON has to deal with real-world data which this increases the complexity dramatically. The BACON.3 program seems to work with real-world data. The” discovery” of Kepler’s third law by BACON.3 is achieved from the distance relative to the earth and a slope (see Table 3).

Table 2: Input data of BACON.3 for discovering Kepler's third law

Planet	Distance (D)	Slope (s)	Ds	D^2s	D^3s
Mercury	0.387	4.091	1.584	0.613	0.971
Venus	0.724	1.600	1.158	0.839	0.971
Earth	1.000	0.986	0.986	0.986	0.971
Mars	1.524	0.524	0.798	1.217	0.971
Jupiter	5.199	0.083	0.432	2.247	0.971
Saturn	9.539	0.033	0.319	3.044	0.971

A slope together with an intercept defines a linear relation between the given parameters *a* and *b* as in $as + i = b$. A separate table is listing the angle between the sun and the planet seen from a fixed point and the distance determines an invariant slope. As can be seen from the third line in Table 3, the result for D^3/s should be 0.986 but it still fits to 0.971. This is due to the modifiable noise margin allowed when looking for invariants.

The historical data

In order to assess the descriptive and explanatory value of BACON for the two selected historical cases, it is necessary to verify that the data given to BACON actually corresponds with what was known to Kepler and his contemporaries. As cited before, Langley e.a. claims the sine law is one of the “instances where, in the actual history of the matter, the data – essentially the same data that were available to BACON”. However this is actually not the case. For many centuries the refraction tables from Ptolemy’s *Optics* were the main authority.

Table 3: the refraction tables from Ptolemy’s *Optics*

Air/water		Air/glass		Water/glass	
Incidence	Refraction	Incidence	Refraction	Incidence	Refraction
10	8.0	10	7.0	10	9.5
20	15.5	20	13.5	20	18.5

30	22.5	30	19.5	30	27.0
40	29.0	40	25.0	40	35.0
50	35.0	50	30.0	50	42.5
60	40.5	60	34.5	60	49.5
70	45.5	70	38.5	70	56.0
80	50.0	80	42.0	80	62.0

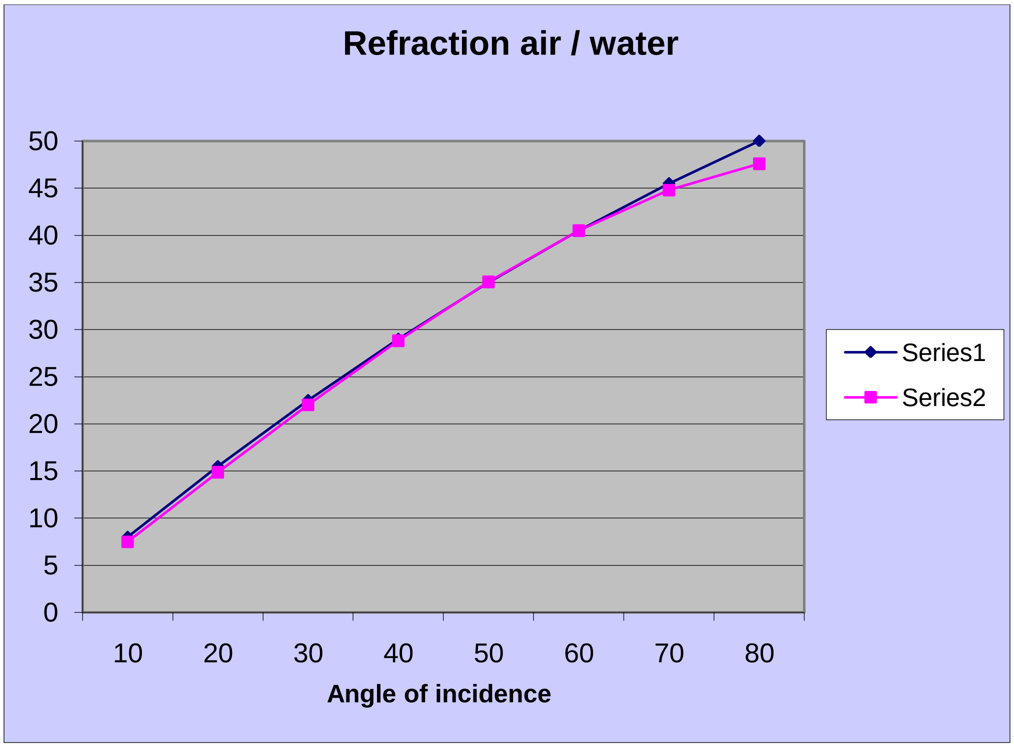
Around c.160-168 AD, Ptolemy collected data by setting up carefully contrived experiments, using a bronze instrument. He reached quite accurate measures for the angles of incidence and corresponding angle of refraction between three types of media: air / water, air / glass and water / glass (Smith, 1996; *Optics*, V, 7-11; 20-21; 31-35, see Table 4).

How do these tables compare with the data used by the BACON programs? For a start, remark that Table 1 in the Langley book lists the sines for incidence and refraction between vacuum and water. Vacuum did not exist as a concept at the time of Ptolemy and started to evolve only after the experiments by Robert Boyle and Robert Hooke after the 1650's. It was unfeasible to set up experiments with a vacuum in the first half of the seventeenth century. Apart from this historical blunder, also remark that the sines for the denser media, water or oil, are higher than those for vacuum. As was known already by Ptolemy, the angle of refraction is lower than the angle of incidence when moving from a rare to a dense medium. The refractive index of water is 1.33, defined as the sine of and the angle of incidence from vacuum, divided by the angle of refraction in water. From Table 1 we read that the refractive index for water is 0.75. BACON seems to discover the inverted sine law! A conclusion that cannot be avoided is that not Table 1 nor the data used by BACON.5 correspond in any way with the historical data.

The story gets even worse. If BACON would have used the historical data for data-driven induction, it would have arrived at a different law! To demonstrate this, let us look again at Table 4. There is something peculiar about the angles. The angles of refraction all fit within a full or halve degree. This could be explained as rounding errors caused by the measuring equipment. However, I tried to imagine what one would arrive at if one tries to induce a general quantitative law from the Ptolemy's observational data. For that purpose I used the polynomial curve fit function of the symbolic computation program *Mathematica*. To my great surprise Ptolemy's data perfectly match with three simple quadratic relations between angles of incidence (i) and angles of refraction (r): $r = \frac{(33-i)^2}{4}$ for air/water, $r = \frac{(29-i)^2}{4}$ for air/glass and $r = \frac{(39-i)^2}{4}$ for water/glass. BACON.3 looks for linear relation between the input data and would not find this relation. However Langley e.a. (1987, 171, note 1) mentions that an early version of BACON.5 used a polynomial curve fit module. On the 1981 IJCAI conference Langley, Bradshaw and Simon, (1981), reported on the discoveries of this program, including the sine law of refraction. As is shown here, the early version of BACON.5 should have found the quadratic relations.

The graph in Table 5, the data from Ptolemy (series 1) fits the sine law (series 2) very closely for smaller angles of incidence. The difference becomes apparent only for degrees of incidence higher than 70 degrees. This raises some questions. Did anyone notice this before? Apparently yes. Gilberto Govi, who published a Latin edition of Ptolemy's *Optics*, was the first to point out the existence of a quadratic relation: $r = ai - bi^2$ (Govi, 1885, pp XXII). A. Mark Smith (1996) finds the relation r = $R\text{-}(n^2 d_2\text{-}nd_2)/2$.

Table 4: comparing the quadratic relation against the sine law



A second question is why Ptolemy's data shows this remarkable regularity. Albert Lejeune who published a French translation of the *Optics* suspects that the figures may have been "adjusted" by previous translators ("Il n'est pas absolument excluque les tables aient été régularisées par un interpolateur grec our arabe", Lejeune 1956). However, a satisfactory explanation is found in Otto Neugebauer treatment of Babylonian astronomy (Neugebauer 1957, 111). Since 500 BC it was the practice to constantly diminishing increments for the construction of astronomical tables. The computation of the ephemerides was achieved by such method, which Neugebauer calls a zigzag function.

Table 5: The "computation" of Ptolemy's refraction tables

Incidence	Refraction	Arithmetical progression with decreasing increments
10	8.0	
20	15.5	8.0 + (8.0 – 0.5)

30	22.5	15.5 + (7.5 – 0.5)
40	29.0	22.5 + (7.0 – 0.5)
50	35.0	29.0 + (6.5 – 0.5)
60	40.5	35.0 + (6.0 – 0.5)
70	45.5	40.5 + (5.5 – 0.5)
80	50.0	45.5 + (5.0 – 0.5)

We can indeed reconstruct Ptolemy's table by means of such zigzag function, as shown in Table 6. Such customs raise serious objections about a simple view of scientific discovery as data-driven induction from observational data. The data used for astronomy and optics in antiquity does not consist of raw observational data but is already shaped by geometrical or arithmetical models, believed to govern the organisation of nature and the universe. Such an epistemological view is clearly present in Kepler. Kepler did not use Ptolemy's refraction tables but a slightly different one from Witelo. Witelo's *Perspectiva,* is a seminal work on optics from the thirteenth century. It was known to Kepler through the Risner edition of 1572 (for a modern edition see Unguru, 1977). Witelo's refraction tables differ from Ptolemy only for the first line and only for water to air. The refracted rays appear under an angle of 7°55' in the manuscripts and 7°45' in the printed edition. Kepler published in 1604 an extensive work on optics as a critique on Witelo, titled *Ad Vitellionem paralipomena, quibus astronomiae pars optica traditur* (for a modern edition and English translation see Donahue, 2000). Chapter 4 is completely devoted to refraction. Proposition 8 contains a passage which reveals Kepler's attitude towards observational data and Witelo's refraction tables in particular. Kepler decomposes the angle of incidence and refraction into two components and then uses an algebraic relation to calculate the composite again. He compares his results against Witelo's table (Donahue 2000):

> This tiny discrepancy should not move you; believe me: below such a degree of precision, experience does not go in this not very well-fitted business. You see that there is a large inequality in the differences of my figures and Witelo's. But my refractions progress from uniformity and in order. Therefore, the fault lies in Witelo's refractions. You will believe this all the more, if you look to the increments of the increments in Witelo. For they increase through 30 minutes. It is therefore certain that Witelo laid his hand upon his refractions gathered from experience so as to bring them into order through an equality of the second increments.

Here Kepler shows that he understands that the tables are adjusted into an order of decreasing increments. He also believes that there are some hidden relations between the angles and that observational data only approximates data arrived by calculations. His method, as revealed here, starts from a hypothetical relation for which he constructs an algebraic (as in proposition 8) or geometrical model (discussed below). From this he calculates the expected angles of refraction and checks them against observational data.

For the second case of Kepler's third law, two parameters determine the law: the average distance of planets from the sun and the periods of their revolution around the sun. Table 7 lists the data used by the BACON program, data available to Kepler from Tycho Brahe and data used by Kepler when writing the *Hamonices Mundi*.

Table 6: Data for the mean radius compared

Planet	BACON	Tycho	Kepler 1618
Mercury	0.387	0.387	0.388
Venus	0.724	0.723	0.724
Earth	1.000	1.000	1.000
Mars	1.524	1.524	1.524
Jupiter	5.199	5.202	5.200
Saturn	9.539	9.539	9.510

Given the modifiable noise margin of BACON.3 the data for the average distance of planets from the sun can be considered the same as the historical data. However, for the second parameter BACON.3 "is given the angle found by using the fixed star and the planet as the two endpoints and the primary body (the Sun) as the pivot point" (Langley 1987, 99). This data allows to calculate an invariant slope for each planet. Nowhere amongst the many tables in the *Harmonices Mundi* do we find a table like this.

In conclusion we may state that the firs condition for an adequate explanation of historical cases, the use of the same historical data, is not met for the two cases discussed here. Despite the claims made by the BACON team their program is given data which is different from the actual data used at the time of the discoveries. This seriously undermines the descriptive and explanatory ambitions of the BACON program.

The historical methods

Having established that BACON's input data does not follow the historical sources we can deal with the second, more important question: does the method of data-driven induction fit the historical discoveries? Let us start with the sine law of refraction. As is now established, the sine law was discovered independently by Thomas Harriot around 1602, by Willebrord Snellius in 1621 and by René Descartes between 1626 and 1628. Descartes was the first to publish the law in his *Dioptrique* of 1637. In some sense you can also say that Pierre Fermat, who tried to disprove

Descartes, came to the same discovery in 1662. Kepler attempted to formulate a law in his *Paralipomena* of 1604. The fact that these natural philosophers came to the same discovery independently within a very short time period is an interesting phenomenon of scientific discovery. If accurate data was available since Ptolemy and the discovery is only a matter of data-driven induction, why did I take fifteen centuries to come to the sine law? Furthermore, why did several individuals came to the discovery within a matter of a few decades? The answer is beyond the scope of the current paper but the contributions of the medieval perspectivist tradition should be mentioned here. These contributions include the decomposition of rays into orthogonal components and the idea of the conservation of one component during refraction (Heeffer 2006). By the beginning of the seventeenth century there was a consensus that there must exist some geometrical relation or an arithmetical proportion between the components of the rays of incidence and the refracted rays. This insight is very well reflected in Kepler's introduction to "the measurement of refraction" (Kepler 1604, Ch. 4, § 2; Frish, 1859 Bd II, p. 182; translation by Donahue 2000, p. 102):

> Since density is obviously a cause of refraction, and refraction itself appears to be a kind of compression of light (i.e., towards the perpendicular), it comes to mind to ask whether the ratio of the media in the case of densities is the same as the ratio of the bottom of the spaces that light has entered into and strikes, first in an empty vessel, and then one filled with water.

Table 7: Kepler's hypotheses

H01	EQ/EH
H02	FQ/FH
H03	EQ^2/EH^2
H04	a.EQ/b.EH
H05	FHEC/FQEC
H06	EQ^3/EH^3
H07	IY/IP
H08	GC/IE
H09	CE/CK
H10	FH/FX
H11	CK/FX

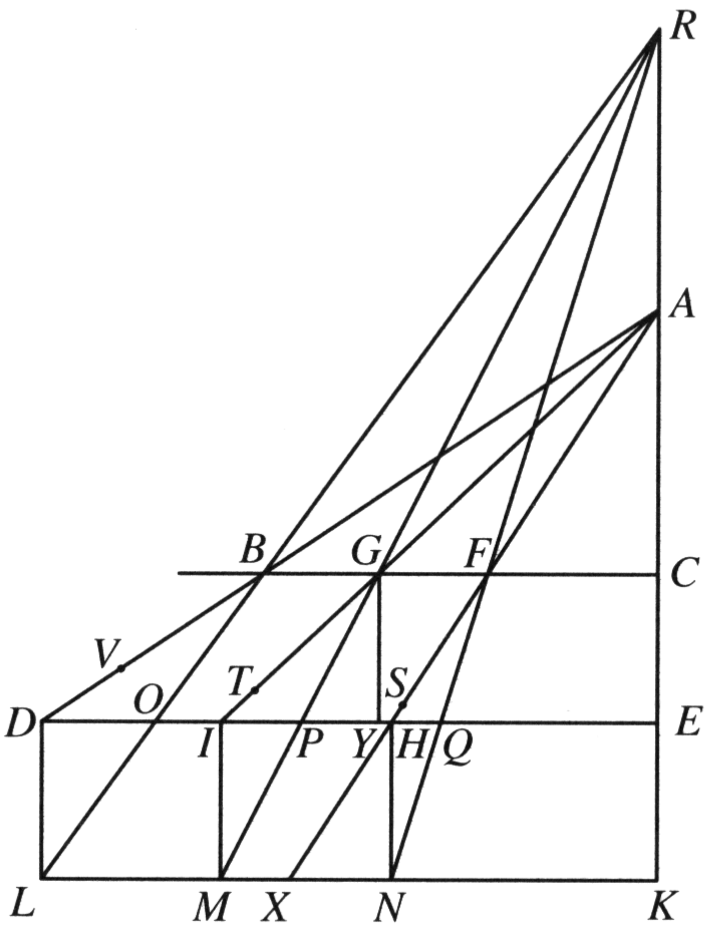


Figure 1: Kepler's geometrical model for refraction

So, Kepler starts from the general hypothesis that in a geometrical model of refraction (shown in Figure 1) some component of the ray of incidence is in a specific ratio to a corresponding component of the refracted ray. This ratio should be invariant with the angle of incidence. To able to check the invariances he draws three angles of incidence, AFC, AGC and ABC. The clever representation of Figure 1 allows Kepler to reason about refraction using geometrical knowledge of rectangular triangles. Based on the refraction tables he then draws three corresponding angles of refraction, RFC, RGC and RBC. He then considers the ratio's between components of the incidence rays and the refracted rays. The first one, which I will call H1, looks at the ratio of the distance of the refracted ray at the bottle of the vessel with the distance of the unrefracted ray, e.g. EQ/EH. The next step is to either prove or disprove the hypotheses considered. Kepler's approach is to eliminate hypotheses by either deduction or observation. In order to be invariant the EQ/EH should be the same as EP/EI and EO/ED. Kepler dismissed this first hypothesis as it "is refuted by experience". He then goes on formulating alternative hypotheses. They are summarized in Table 8. Each of the formulated hypotheses gets refuted in the end.

With some disappointment Kepler moves on to a next section writing "Hitherto, we have followed an almost blind plan of enquiry, and have called upon luck". However, his plan was not so blind, but he was unlucky. In the next section I will demonstrate that his method was the right one to discover the sine law.

Kepler's method for the sine law

We can design a model based on Kepler's general hypothesis. In fact, the hypothesis he uses is more general than in the quotation above, because more ratios are investigated than those on the bottom of the vessel. A fair reformulation of Kepler's central hypothesis would be the following:

> The ratio of the optical densities of two media is proportional to some ratio of two line segments in the geometrical representation of a light ray traversing the two media, the first line segment related to the angle of incidence, the second to the angle of refraction.

When this central hypothesis is implemented in a computational model, it would generate all of Kepler's hypotheses H01 to H11, as well as many more (Heeffer 2003). As can be noticed from Table 8, with exception of H08 and H11, in all ratios the two line segments have one point in common. If a restriction is added to the starting hypothesis, that the line segments used in the ratios, should start in the same point, the model generates only nine (linear) instances. These ratios cover all of Kepler's except the special case H07.

The next step is to either prove or disprove the generated hypotheses. Kepler's approach is to eliminate hypotheses by deduction or observation. Most hypotheses can be refuted deductively, by simple geometrical reasoning. For example, H11 states that the ratio *CK*/*FX* remains the same for varying angles of incidence. This is evidently not the case. As the angle of incidence increases, the length of *FX* will increase while *CK* remains the same. Therefore the ratio *CK*/*FX* decreases and hypothesis H11 is refuted. The tragedy is that Kepler succeeded in formulating both a suitable representation of the problem and the correct hypothesis that some geometrical proportion corresponds to the refraction index. He failed in identifying the correct ratio. Both *FR*/*FA* and *FN*/*FH* correspond to the ratio of optical densities of the two media. These ratios can be proved to be constant by geometrical reasoning. Line segment *FC* allows establishing a relation between the angles of incidence and refraction as it is the same side of the right-angled triangles *FRC* and *FAC*. This unfortunate oversight was Kepler's failure in discovering the sine law. Both Snell and Descartes have read the *Paralipomena* and undoubtedly found here their main inspiration for the sine law. Snell's formulation was based on a ratio of cosecants equaling *FH*/*FN* (Volgraff, 1918, p. 21b).

$$FA.\sin(FAC) = FR.\sin(FRC)$$

$$GA.\sin(GAC) = GR.\sin(GRC)$$

$$\frac{\sin(FAC)}{\sin(FRC)} = \frac{FR}{FA} = \frac{FN}{FH}$$

Descartes never mentioned his sources and took care not to reveal his path of discovery of the sine law. Several authors have formulated hypotheses on how Descartes came to his discovery. In an early study by Kramer, it was suggested that Descartes hit upon the sine law through his study of conic sections, in particular the problem of the anaclastic curve (Kramer 1882, p 256-8 and note 39). Others, such as William Shea believe that Descartes used the sine law for solving the anaclastic (Shea 1991). Shea argues that the demonstration of a refractometer, presented by Descartes in a letter to his lens cutter Ferrier, as the procedure leading to the sine law. Given Kepler's analysis as sketched above and the fact that Descartes called Kepler "my first teacher in optics" (in a letter from Descartes to Mersenne, 31 March 1638 and also in Descartes to Mersenne, 13 mei 1638). I consider it most likely that Descartes found in Kepler's drawing and his main hypothesis everything needed to deduce the sine law by pure geometrical reasoning (Heeffer 2006). None of the scholars who have worked on the discovery of refraction by Harriot, Snell or Descartes have reported any evidence pointing to the use of data-driven induction. Above is shown that Kepler's method is an adequate one in arriving at the sine law and that it is quite different from data-driven induction.

Kepler's method for the third law

Kepler's discovery of his third law is not as well documented as his analysis of refraction. The first and second law were published in his most important work, the *Astronomia nova* of 1609. A quantitative relation between the average distance of planets to the sun and their period eluded him for many years. It was while writing the *Harmonices mundi* several years later that he hit upon the relation of 2/3 of the powers. The law is buried within the text on page 189: "The proportion between the periodic time of any two planets is precisely sesquialter (2/3) the power of their mean distances", as if it were some side remark. But Kepler knew the significance of his discovery. He distinctively remembered the precise date of his discovery (translation by Gingrich 1975, 595):

> it was conceived on March 8 of the year 1618, but unfelicitously submitted to calculation and rejected as false, and recalled only on May 15, when by a new onset it overcame by storm the darkness of my mind with such full agreement between this idea and my labor of seventeen years on Brahe's observations that at first I believed I was dreaming and had presupposed my result in the first assumptions.

For seventeen years he possessed the necessary data to discover the law. Why did he find it in 1618 and was it reached by data-driven induction? Kepler does not tell us and neither does he give a justification for the law. It appears in the book as an auxiliary finding while dealing with the harmonies of the world. It "overcame" to him as it would be not the result of a conscious process of induction but as a simple ratio that did not occur to him before. To understand his method it is necessary to place his discovery in the right context. Kepler recorded an epiphany like this before on 19 July of 1595. He then realized that the respective radii of the five known planets measure in close correspondence with the radii of five Platonic solids nested in a specific order. This idea together with the famous depiction of the nested spheres was published in his first book, the *Mysterium Cosmographicum* or *Secret of the Cosmos*. Kepler was a deeply religious man who believed that God ordered the universe using principles of geometry. In order to understand the universe one has to look for geometrical relations which form the basis of the model that God had in mind. The *Harmonices mundi*, which contains the third law, is very much alike his first book. Here Kepler comes back to his previous ideas on the relation between platonic solids and the orbits of planets but also adds and combines this with properties of regular polygons and harmonic principles of music, hence the title, *Harmonies of the World*. From the many principles he explores, one can discern a consistent pattern. The idea of a planned universe leads Kepler to look for principles of proportion in plane and solid geometry and in music. In each instance he tries to bring proportions from music or geometry in correspondence with observational data. In other words he fits a priori data available from mathematical theories with synthetic data arrived by observations. This is the opposite of data-driven induction which formulates general laws in the language of mathematics from empirical data.

Table 8: Comparing the orbital radii calculated by harmonic proportions with observation

Planet	Harmonic	*Mysterium*
Mercury	0.385	0.360
Venus	0.738	0.719
Earth	1.000	1.000
Mars	1.505	1.520
Jupiter	4.745	5.246
Saturn	8.837	9.164

Table 9 shows and interesting example from lesser known work done by Kepler in 1599 (described by Stephenson 1994, 90-97). The first column lists the orbital radii of the planets calculated as musical proportions, starting with the ratio 3:4. These data are compared with observational data from the *Mysterium cosmographicum*. The *Harmonices Mundi* contains many such examples and the correspondence between the observed data and the calculated relations is often very close.

Concerning the third law, we do not know exactly how Kepler came to the idea. However, given the context of the book and his methodology summarized here we can assume that he fitted a mathematical proportion, “sesquialter the powers”, with observational data. As Kepler formulated sub-hypotheses using square and cubic terms as in H6 and H9 for his study of refraction, he may have proceeded along the same course for the ratio between the period and average distance of planets around the sun. In conclusion we may safely state that data-driven induction not at all fits the process that led to the discovery of Kepler’s third law.

The value of the data-driven induction model

Drawing further on the analogy with chess programming I have demonstrated that BACON is as worse in explaining the mechanisms of discovery in the two historical cases discussed as a chess program is in explaining the problem solving capacities of chess grand masters. The next question that I will address is if BACON is as good in scientific discovery as current chess programs are in beating the world champion.

Reading through the many "discoveries" by BACON in the Langley book, one cannot escape the impression that data has been contrived and procedures have been arranged ad hoc to discover what is already known. The data for discovering the sine law (Table 1) is giving away the discovery. The real discovery was made by Kepler, namely the hypothesis that some ratio between the components of the angles of incidence and refraction remains invariant with the degree of the angles. He failed to find the right line segments in his geometrical model but others such as Harriot, Descartes and Snell did, probably reasoning in the same way. Furthermore I have shown that when applying data-driven induction to the historical data this would lead to a quadratic relation between the angles of incidence and refraction which is different from the sine law. Actually, also the data used by BACON.1 for "discovering" Keplers's third law can lead to many, probably infinitely many, possible quantitative laws. The fact that BACON.1 finds precisely Kepler's third law is very suspicious as it involves a combination of a quadratic and a cubic term. The relation between the data for period and average distance in Table 2 can be expressed by one simple quadratic relation:

$$P = \frac{11}{60}D^2 + \frac{17}{12}D - \frac{6}{10}$$

It can easily be verified that for the input data D = 1, 4 and 9 this leads to P = 1, 8 and 27. Why did BACON.1 not find this simpler quadratic relation or any similar one? Because the production rules were set up to find Kepler's third law. In all the "discoveries" involving a first pass through the program for finding a slope, such as with Kepler's third law in BACON.3, the slope is already containing the relevant relation between the parameters.

So, the claims by the BACON team their program is constructive, that it did "rediscover" all these laws in physics and optics can only be met with scepticism. As data-driven induction is not adequate to describe the historical discoveries, it is neither so as a general exclusive framework for doing discovery. Data-driven induction would produce an infinite number of laws that fit the input data but are meaningless and break down for additional cases.

It is instructive to keep in mind the so-called Titus-bode law. The law was formulated during the eighteenth century as an inductive generalization of the respective distances of planets around the sun. In its modern formulation, the law expresses a measure in astronomical units a of the semi-major axis, or the longest axis of the elliptical orbit, such that

$a = 0.4 + 0.3k$ where k = 0, 1, 2, 4, 8, 16, 32, 64, 128.

Table 11 shows that the law fits the observational data well for the first four planets. Then there is a gap for $k = 8$ which could be filled up by the dwarf planet Ceres. It further continues to go well up to Uranus but the law breaks down for Neptune and Pluto.

Table 9: Data for the Titus-Bode law

Planet	k	Calculated	Observed	Error
Mercury	0	0.4	0.39	2.56%
Venus	1	0.7	0.72	2.78%
Earth	2	1.0	1.00	0.00%
Mars	4	1.6	1.52	5.26%

Many possible mathematical relations can be found by induction without being laws of science. A physical law is more than a generalization of data. A physical law is part of a broader theory and should fit with other laws in a mathematical meaningful way. A good example is the reformulation of Kepler's third law in Newtons theory of universal gravitation. Newton established that two orbiting bodies with mass m_1 and m_2 with period P having respective distances to a centre of mass d_1 and d_2 relate as:

$$(m_1 + m_2)P^2 = (d_1 + d_2)^3 = D^3$$

Kepler's third law is thus only approximately true. Because the mass of the sun is much larger than these of the planets the value of $m_1 + m_2$ approximates m_1. It is true the connection of Keplers' third law with Newton's theorey of universal gravitation that it can be understood as a law of physics, rather than an inductive generalization. The isolated context of data generalization thus raises questions on the usability of data-driven induction for physical laws.

Conclusion

I have presented an assessment of the BACON program which is very critical. There are serious methodological problems with their modelling of scientific discovery. They make claims about the explanatory power of historical cases which cannot be sustained by historical research. They present BACON as a *general* model for scientific discovery, which it is not. Data-driven induction may have some function in scientific discovery but it is certainly not the decisive creative step as presupposed by some AI approaches. However, this does not imply that scientific discovery is beyond rational explanation or models of artificial intelligence. At the contrary, the rational process of hypothesis formulation and testing – as I have proposed – does provides an alternative model for scientific discovery, and is one which can be supported by historical research. The state-space search model for discovery and problem solving, pioneered by Simon, allows for the modelling of scientific discovery without having to rely on single, simple and ad-hoc method, as is data-driven induction.

References

Donahue, William (2000) *Kepler's Optics*. Green Lion Press. Santa Fe, New Mexico.

Gingrich, Owen (1975) "The origins of Kepler's Third Law," in A. Beer and P. Beer (eds.) Kepler: Four Hunderd Years. Vistas in Astronomy, vol. 18, Pergamon, New York, pp. 595-601.

Heeffer, Albrecht (2003) "Kepler's near discovery of the sine law: A qualitative computational model", in C. Delrieux and J. Legris (eds.), *Computer Modeling of Scientific Reasoning* (Bahia Blanca, Argentinia, 2003), pp. 93-102.

Heeffer, Albrecht (2006) "The Logic of Disguise: Descartes' Discovery of the Sine Law", *Historia Scientiarum. International Journal of the History of Science Society of Japan*, 16 (2), November 2006, pp. 144-165.

Hempel, Carl G. (1945) "Studies in the logic of confirmation", *Mind* 54, 1-26, 97-121.

Govi, Gilberto (1885) *L'Ottica di Cl. Tolomeo da Eugenio*. Torino.

Kepler, Johannes (1596) *Prodromus dissertationum cosmographicarum, continens mysterium cosmographicum, de admirabili proportione orbium coelestium ... : demonstratum, per quinque regularia corpora geometrica*, Gruppenbach, Tübingen. Translated by A. M. Duncan with notes by E. J. Aiton. (1986) *The Secret of the Universe*. New York, Abaris Books.

Kepler, Johannes (1604) *Ad Vitellionem paralipomena, quibus astronomiae pars optica traditur...* Claudius Marnius & heirs of Johann Aubrius, Frankfurt, see Donahue (2000).

Kepler, Johannes (1619] The Harmony of the World. Translated into English with an Introduction and Notes by E.J. Aiton, A.M. Duncan, and J.V. Field, 1997. Philadelphia, American Philosophical Society

Kepler, Johannes (1858–1871). Joannis Kepleri Astronomi Opera Omnia. Edited by Christian Frisch. Frankfurt and Erlangen

Kramer P. (1882) Descartes und das Brechungsgesetz des Lichtes, *Abhandlungen zur Geschichte der Mathematik. Zeitschrift fur Mathematik und Physik*, t. XXVII, p 233-278.

Langley, Pat., Gary L. Bradshaw an Herbert A. Simon (1981) "BACON.5: The discovery of conservation laws". *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, 1, 121-126.

Langley, Pat, Herbert A. Simon, Gary L Bradshaw and Jan M. Zytow (1987) *Scientific Discovery. Computational Explorations of the Creative Processes*. MIT Press, Cambridge, Ma.

Lejeune, Albert (1956) *Ptolémée. L'Optique de Claude Ptolémée dans la version latine d'après l'arabe de l'émir Eugène de Sicile*. Edition critique et exégétique par Albert Lejeune. Publications universitaires de Louvain.

Nickles, Thomas (1994) "Enlightenment Versus Romantic Models of Creativity in Science - and Beyond" *Creativity Research Journal* 7, 277-314.

Neugebauer, Otto (1957) *The Exact Sciences in Antiquity*. Second Ed., reprinted by Dover 1969.

Newell, Alan and Herbert Simon, and (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Popper, Karl R. (1959) *The Logic of Scientific Discovery*, New York: Basic Book.

Risner, Friedrich (ed.) (1572) *Opticae Thesaurus*. Basel.

Shea, William R. (1991): *The magic of numbers and motion : the scientific career of René Descartes*. Science history publl. Cambridge.

Smith, A. Mark (1996) *Ptolemy's Theory of Visual Perception: An English Translation of the Optics with Introduction and Commentary*. Philadelphia: American Philosophical Society.

Stephenson, Bruce (1994) The Music of the Heavens. Kepler's Harmonic Astronomy. Princeton University Press, Princeton, NJ.

Unuguru, Sabetai (ed.) (1977) Witelo, *Witelonis Perspectiva, Liber Primus*. An English translation with introduction and commentary and Latin edition of the mathematical book of Witelo's Perspectiva, Studia Copernicana, XV, Polish academy of Sciences, Warshau.

Volgraff, J. A. (1918) *Risneri Opticam cum Annotationibus Willebrordi Snellii*. Aedibus Plantini, Gent.