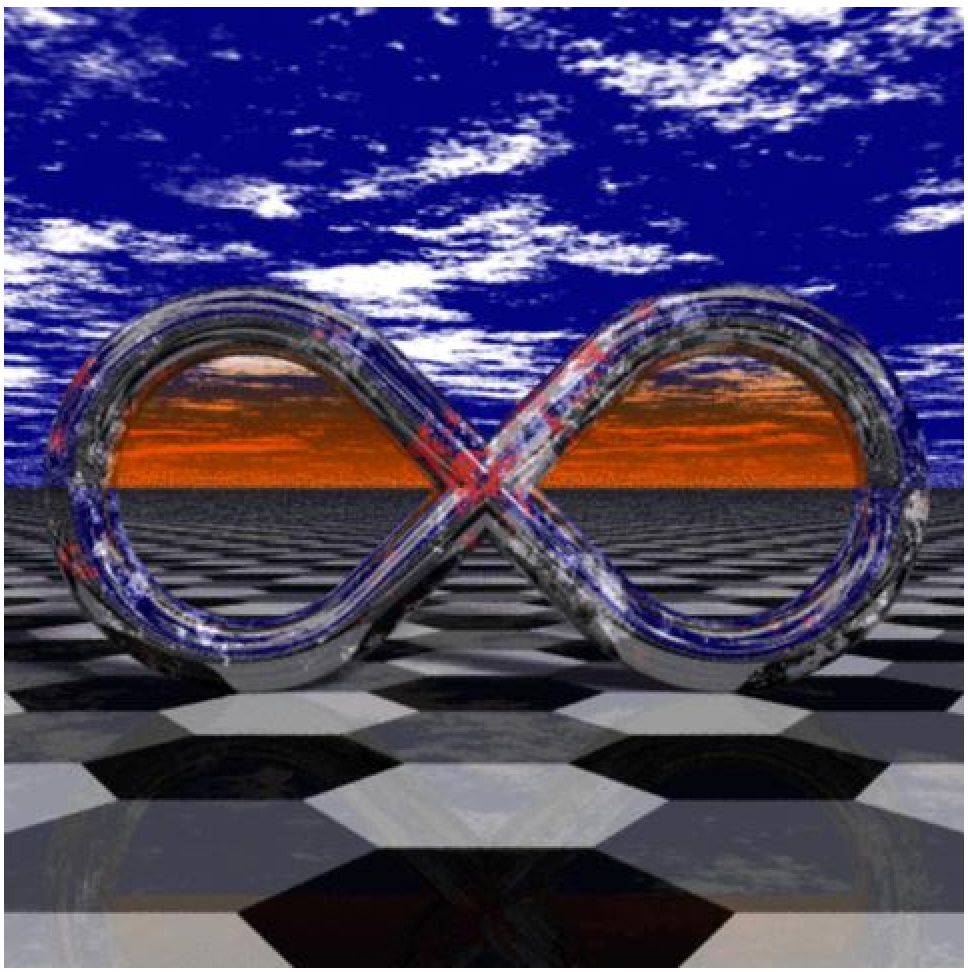
The Two Faces of Infinity

Dr. Bob Gardner

Great Ideas in Science (BIOL 3018)



From the webpage of Timithy Kohl, Boston University

INTRODUCTION

Note. We will consider infinity from two different perspectives: (1) as the world's largest number (of course, it is neither a number nor of this world!), and (2) as the size of a set. You have probably met infinity in the first setting while taking calculus. You are likely familiar with the idea of infinite sets, but the details of such sets may surprise you and, unless you are a math major, you probably have *not* explored these details (which are only a little over 100 years old).

ZENO'S PARADOX

Zeno of Elea (circa 490–435 BCE)

Note. The most famous arguments concerning infinity are probably those of Zeno of Elea. "Zeno's Paradox" deals with both the ideas of infinity and continuity. It can be explained like this: Imagine that I am standing 10 feet from an open door. In order to walk through the door, I must first walk half the distance to the door. Next, I must walk half of the remaining distance (1/4 of the original distance). Next, I must walk half the remaining distance (1/8 of the original distance), and so forth. At each stage, I must walk half the remaining distance (at stage n, the distance to go is $1/2^n$ of the original distance). Since this requires an infinite number of steps, I can never reach the door.

Note. The problem with Zeno's Paradox is that it assumes the spacial dimension can be cut into *arbitrarily small* pieces, whereas the temporal (time) dimension cannot be cut into such pieces. This is the resolution of the "paradox." All I need do is cover each of the new distances in half the time of the distance in the previous stage. However, this still requires that we deal with an infinite sum:

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots = \sum_{n=1}^{\infty} \frac{1}{2^n}.$$

Note. In order to make sense out of an infinite sum, we must pursue, in analogy with the "arbitrarily small" idea, an "arbitrarily close" idea. This approach to infinite sums (and more generally, limits) was not formalized until the early 1800's when Augustine Cauchy gave us he definitions we use today.

CAUCHY AND LIMITS

Augustine Cauchy (1789–1857)

Definition. An infinite sum $\sum_{i=1}^{\infty} a_i$, where each a_i us a real number, is called a *series*. The *nth partial sum* of a series is the sum of the first n terms: $S_n = \sum_{i=1}^{n} a_i$. These partial sums then define a *sequence* of real numbers $S_1, S_2, S_3, \ldots = \langle S_n \rangle$.

Note. We now want to see if the sequence of partial sums approaches some limit.

Definition. The *limit* of a sequence $\langle S_n \rangle$ is the number L if:

$$\text{for all } \epsilon > 0, \text{ there exists a natural number } N \text{ such that}$$
$$\text{if } n > N \text{ then } |S_n - L| < \epsilon.$$

This is denoted $\lim_{n \to \infty} S_n = L$. If sequence $\langle S_n \rangle$ has a limit, then it is said to *converge*. If a sequence does not converge, then it is said to *diverge*. A series converges with *sum* L if the limit of its sequence of partial sums is L.

Note. The idea of convergence of a sequence is that the terms of the sequence can be made *arbitrarily close* to L by going *sufficiently far* out the sequence. Notice that the limit concept DOES NOT say anything along the lines of "getting closer and closer, but never gets there." The important (informal) idea is that the terms "get close to L and stay close to L."

Definition. A sequence is *geometric* if consecutive terms are in a constant ratio. That is, $\langle a_n \rangle$ is a geometric sequence if $a_n = ar^{n-1}$ for all $n \in \mathbb{N}$.

Theorem A. The nth partial sum of a geometric sequence $\langle a_n \rangle = \langle ar^{n-1} \rangle$, where $a \neq 0$ and $r \neq 1$, is

$$S_n = \frac{a(1 - r^{n+1})}{1 - r}.$$

Theorem B. If $|r| < 1$ then the geometric sequence $\langle r^n \rangle$ has a limit of 0. That is, $\lim_{n\to\infty} r^n = 0$.

Theorem C. If $|r| < 1$, then the geometric series $\sum_{i=1}^{\infty} ar^{i-1}$ has sum $\frac{a}{1-r}$: $\sum_{i=1}^{\infty} ar^{i-1} = \frac{a}{1-r}$.

Proof. By Theorem A, the nth partial sum of the series is

$$S_n = \frac{a(1 - r^{n+1})}{1 - r} = \frac{a}{1-r} - \frac{ar^{n+1}}{1-r}.$$

Therefore the sum of the series is

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \left(\frac{a}{1-r} - \frac{r^{n+1}}{1-r} \right) = \frac{a}{1-r} - \frac{\lim_{n\to\infty}(r^{n+1})}{1-r} = \frac{a}{1-r} - 0$$

by Theorem B. ∎

Note. Theorem C now allows us to resolve Zeno's Paradox. The total distance traveled (when we start with the initial distance 10 feet, as described above) is

$$\sum_{n=1}^{\infty} \frac{10}{2^n} = \frac{10/2}{1 - 1/2} = 10 \text{ feet.}$$

If I move at a steady rate, then the amount of time this takes is (proportional to) $\displaystyle\sum_{i=1}^{\infty} \frac{1}{2^i} = \frac{1/2}{1 - 1/2} = 1$. So, for example, if I go 10 ft.sec, then I can reach the door in 1 second (well, DUH!).

Note. Whenever we discuss limits, we are assuming a certain property of the real numbers called "completeness." Completeness can best be described in terms of an airplane which takes off from the ground.

COMPLETENESS AND AIRPLANES

We use an intuitive situation to illustrate some properties of the continuum.

Imagine that an airplane taxis down a runway and takes to the air. Once in the air, the plane remains in the air (instead of, say, the wheels bouncing on the runway before the plane gains altitude). Also, the plane is either in the air or on the ground (as long as a wheel is on the ground, we say that the plane is on the ground and that it is not in the air until all parts of the plane are off the ground — technically, we are interested in the height of the plane which is either zero or positive, once positive stays positive, and changes in a continuous way). We ask the question: "Is there

(a) a first point (in time) at which the plane is in the air,

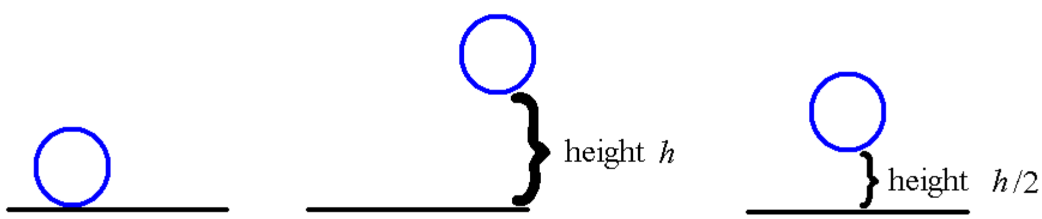
(b) a last point at which the plane is on the ground,

(c) both (a) and (b), or

(d) neither (a) nor (b)?"

By exploring these four possible cases, we will get a sound idea of what the continuum is.

Consider case (a). *If* there is a first point in time at which the plane is in the air, then we get an easy contradiction. Suppose the point in time at which this occurs is called time t. Since the plane is in the air at time t, then it must have some height, call it h. However, before reaching height h, the plane must have been at height $h/2$ (that is, half the height h). Since h is a positive number, then $h/2$ is a positive number and the plane is off of the ground with a height of $h/2$ at some time *before* time t.[1] But then time t is *not* the first point in time at which the plane is in the air. Since the assumption of case (a) leads to a contradiction, then case (a) cannot hold, and one of the other cases (b), (c), or (d) must hold.



Our airplane (symbolically reduced to a single wheel) on the ground. The airplane at time t with height h. Before reaching height h, the airplane had to be a height $h/2$ (indicating that, in fact, t was not the first point in time when the plane was in the air).

Let's skip case (b) for now and concentrate on case (c). Now when we talk about the continuum, we assume that between any two points there is another point. The same holds for the real numbers (which we relate to a line by the number line correspondence). For example, between any two

[1]We have assumed that the plane does not take a quantum leap and jump into the air to a height of h without crossing all of the space from height 0 to height h. This is an assumption that height is a continuous function of time and an application of the Intermediate Value Theorem from calculus.

numbers a and b is the average of these two numbers $(a+b)/2$. *Suppose* case (c) holds and that there is both a first point in time when the plane is in the air and a last point in time when the point is on the ground. Since a plane cannot be on the ground and in the air at the same time, then these must be different times, call them times a and b. Since time $(a+b)/2$ is between time a and time b, we ask the question "where is the plane at time $(a+b)/2$?" If the plane is on the ground, then time a was not, in fact, the last time when the plane was on the ground since $(a+b)/2$ is later in time than time a. If the plane is in the air, then time b is not, in fact, the first time when the plane was in the air since $(a+b)/2$ is earlier in time than time b. Since the plane must be *somewhere*, these two contradictions show us that case (c) is impossible.[2]

So we are now left with cases (b) and (d). It is much harder to eliminate one of these. In fact, the elimination of one of these is part of the definition of the continuum. Let's explore in more detail. We consider two sets of time values: the set G of times when the plane is on the ground and the set A of times when the plane is in the air. Together, these two sets include all time values. That is, the union of sets G and A make up the whole number line. Now, every time value from set G is less than

[2] A quicker way to dispatch with case (c) is to observe that we already know there is not a first point in time when the plane is in the air (that is, case (a) cannot hold), so it cannot be the case that *both* (a) and (b) hold.

every time value from set A (and, of course, every time value from set A is greater than every time value from set G). Case (b) then translates into the situation that set G has a largest element. Case (d) translates into the situations that set G does not have a largest element and set A does not have a smallest element. Now the idea that a set of numbers which is bounded below may not have a smallest element is not actually a problem. Consider, for example, the set of all positive numbers (that is, all numbers greater than 0). The set is bounded below (by 0, say), but there is no smallest positive number. To see this, we suppose that there is some number q which is the smallest positive number. But then $q/2$ is a positive number and is smaller than q, indicating that any candidate smallest number fails (similar to the argument above that there is not a first time when the plane is in the air). We can similarly show that a set of numbers can be bounded above but not have a largest element (for example, consider the negative numbers). So it is quite possible for sets of numbers to not have extremal (i.e., maximum or minimum) elements. In fact, the set N of negative numbers and P of positive numbers satisfy this property. The problem generated by case (d) is that we cannot have two sets with this property which together give the whole number line. For example, sets N and P have the property but, when unioned together, they do not give the whole number line since they omit 0. Now our sets G and A must "butt up against one another" in a sense similar to the way the negative and positive number sets do. That is, the sets cannot have any distance between them. Now *if* case (d) holds, then even though there is no distance between the two sets, there might be a single point between the sets (like the situation where 0 is between sets N and P). In our story about the airplane, such a hypothetical point p must actually lie in one or the other set since the airplane has to either be on the ground or in the air at this point in time. If p is in set A, then it is the smallest element of set A and case (a) holds. However, we know that case (a) cannot hold. If p is in set G, then it is the largest element of set G and it is the last point in time when the plane is on the ground. This in fact *is* the case! However, we still have only hypothesized the existence of this point p. We have not yet given an argument as to why the sets G and A *must* have this point p. We can observe that if such a point does not exist, then there is a hole in the number line between the two sets. We cannot prove this since, it turns out, this is part of the definition of the continuum! The idea that there are no holes in the number line continuum is called (quite appropriately) the "Axiom of Completeness." One way to state this axiom is: When the number line is cut into two nonoverlapping connected pieces, then either the piece on the left has a largest element or the piece on the right has a smallest element (but not both). This particular statement of the Axiom of Completeness is due to Richard Dedekind in 1872.

Richard Dedekind (1831–1916)

An alternate statement of the Axiom of Completeness is that every set of numbers with an upper bound has a least upper bound (or equivalently, every set of numbers with a lower bound has a greatest lower bound). Since set G is bounded above (recall that any element of A is greater than all elements of G) and set A is bounded below (by any element of G), then set G has a least upper bound and set A has a greatest lower bound. Since sets G and A do not overlap and together give the whole number line, the least upper bound of G must be the same as the greatest lower bound of A. In fact, the point p described above is this common point.

INFINITY AND CARDINALITY

Note. Historically, it took over 2 millenia for the idea of completeness to become understood. The problem is really caused by the irrational numbers.

Definition. The set of *natural numbers* is $\mathbb{N} = \{1, 2, 3, \ldots\}$. The set of *integers* is $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. The set of *rational numbers* is $\mathbb{Q} = \{p/q \mid p, q \in \mathbb{Z}, q \neq 0\}$. The *real numbers* are denoted $\mathbb{R}$. The set of *irrational numbers* is denoted $\mathbb{R} \setminus \mathbb{Q}$.

Note. Now, we explore the sizes of sets. If two sets are finite, then it is easy to explore their sizes — they are the same size if they have the same number of elements. However, this does not work for infinite sets. Consider $A = \{0, 1, 2, \ldots\}$ and $B = \{1, 2, 3, \ldots\}$. To deal with infinite sets, we must use ideas due to George Cantor.

Georg Cantor (1845–1918)

Definition. Two sets A and B are of the *same cardinality* ("size"), if their elements can be matched up pairwise. That is, if there is a one-to-one and onto function from set A to set B.

Note. The sets $A = \{0, 1, 2, \ldots\}$ and $B = \{1, 2, 3, \ldots\}$ are of the same cardinality since $f(n) = n + 1$ is a one-to-one and onto function from A to B. The sets $\mathbb{N}$ and $\mathbb{Z}$ are of the same cardinality since we can make the mapping:

$$
\begin{aligned}
1 &\rightarrow 0 \\
2 &\rightarrow 1 \\
3 &\rightarrow -1 \\
4 &\rightarrow 2 \\
5 &\rightarrow -2 \\
&\vdots
\end{aligned}
$$

That is, $f(2n) = n/2$, $f(2n+1) = -n/2$ maps $\mathbb{N}$ to $\mathbb{Z}$ in the desired fashion. These examples illustrate one funny property of infinite sets: they can be the same cardinality as a proper subset.

Definition. A set is *countable* if it is of the same cardinality as a subset of natural numbers. Therefore, set A is countable if there is a one-to-one function from set A into $\mathbb{N}$.

Note. Informally, a countable set is one for which a complete list can be made: first element, second element, third element, ...

Theorem. A union of two countable sets is countable.

Proof. Let $A = \{a_1, a_2, a_3, \ldots\}$ and $B = \{b_1, b_2, b_3, \ldots\}$. Then the elements of $A \cup B$ are:

$$a_1, a_2, a_3, \ldots$$

$$b_1, b_2, b_3, \ldots$$

We just need a way to list all of the elements above... ∎

Theorem D. A countable union of countable sets is countable.

Proof. Let the sets be $A_1 = \{a_{11}, a_{12}, a_{13}, \ldots\}$, $A_2 = \{a_{21}, a_{22}, a_{23}, \ldots\}$, $A_3 = \{a_{31}, a_{32}, a_{33}, \ldots\}$, ... Then the elements of the union $\bigcup_{n=1}^{\infty} A_n$ are:

$$\begin{array}{ccccc} a_{11} & a_{12} & a_{13} & a_{14} & \cdots \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots \\ a_{31} & a_{32} & a_{33} & a_{34} & \cdots \\ a_{41} & a_{42} & a_{43} & a_{44} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

As in the previous theorem, we just need a way to list all of the elements above... ∎

Note. Consider how the rational numbers are distributed: between any two integers there are an infinite number of rational numbers. In fact, between *any* two real numbers there are an infinite number of rational numbers. It follows, then, that for any real number r, there is not a "first" rational number larger than r. Therefore, the distribution of rational numbers along the real line is much different from the distribution of the integers. So is the set of rational numbers larger than the set of integers? Surprisingly, NO!

Theorem. $\mathbb{Q}$ is a countable set.

Proof. Effectively, we need to make a list of the rational numbers. This seems tricky, since there is no "first" rational number. Consider the following representation of all positive rational numbers:

$$\begin{matrix} 1/1 & 1/2 & 1/3 & 1/4 & \cdots \\ 2/1 & 2/2 & 2/3 & 2/4 & \cdots \\ 3/1 & 3/2 & 3/3 & 3/4 & \cdots \\ 4/1 & 4/2 & 4/3 & 4/4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{matrix}$$

These numbers can be counted as illustrated previously, and so the positive rational numbers $\mathbb{Q}^+$ is countable. Similarly, $\mathbb{Q}^-$ can be counted. So by Theorem D, $\mathbb{Q}$ is countable. ∎

Note. You are probably familiar with the property of rational numbers which says: a real number is rational if and only if its decimal expansion either terminates or has a repeating pattern. Therefore, irrational numbers have decimal expansions which neither terminate nor repeat. This means that any real number can be given an infinite decimal expansion (which is not all 0's after some point). This can be accomplished for rational numbers with terminating decimal expansions by reducing the last nonzero digit by 1 and then adding an infinite number of 9's after it. We use this idea to show that the numbers between 0 and 1 are not countable.

Note. As in the case for rationals, we can show that between any two real numbers there is an irrational number. So there seems to be a similar distribution along the real line for both rationals and irrationals. This makes it all the more surprising that there are *more* irrationals than rationals! (Well, it is probably surprising that there exists *any* uncountable sets!!!)

Theorem. The numbers in the interval $(0, 1)$ are uncountable.

Proof. As discussed above, every number in this interval can be written (uniquely) with an infinite decimal expansion. We now give a proof by contradiction. Suppose, to the contrary of the claim, that there exists a complete listing of these numbers:

$$\begin{array}{ccccccc} a_1 & = & 0. & x_{11} & x_{12} & x_{13} & \cdots \\ a_2 & = & 0. & x_{21} & x_{22} & x_{23} & \cdots \\ a_2 & = & 0. & x_{31} & x_{32} & x_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

We now proceed to create a number $(0, 1)$ which is not in the list, hence showing that any such listing is not complete and hence that the set $(0, 1)$ is not countable. Define $a \in (0, 1)$ where $= 0.x_1 x_2 x_3 \cdots$ by defining the nth decimal of a to be different from the nth decimal of a_n:

$$x_n = \begin{cases} 1 & \text{if} \quad x_{nn} \neq 1 \\ 2 & \text{if} \quad x_{nn} = 1. \end{cases}$$

Since a differs from a_n in the nth digit, then $a \neq a_n$ for all n, and hence it is not in the list. This method of proof is called Cantor's Diagonalization Argument. ∎

Note. The above result shows that the irrationals in $(0,1)$ are uncountable (we already know that the rationals in $(0,1)$ are countable, so if the irrationals were also countable then we would have $(0,1)$ countable by Theorem C). Therefore we see that they are legitimately *more* irrationals than rationals, even though there are an infinite number of both. This means that some infinite sets are bigger than other infinite sets! In fact, there are lots of different kinds of infinities!

Definition. The *power set* of a set X, denoted $\mathcal{P}(X)$, is the set of all subsets of set X.

Example. If $X = \{1,2,3\}$ then $\mathcal{P}(X) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$.

Theorem. If the cardinality of set X is n, denoted $|X| = n$, then the cardinality of the power set of X is 2^n: $|\mathcal{P}(X)| = 2^n$.

Note. Of course, the power set of a finite set is larger than the set itself. Surprisingly, this is also true for infinite sets.

Theorem. *Cantor's Theorem.* For any set X, its power set $\mathcal{P}(X)$ has a greater cardinality than that of X: $|\mathcal{P}(X)| > |X|$.

Note. Cantor's Theorem implies that there can be no largest set — the power set of any candidate largest set is always strictly bigger! This means that there is a whole chain of infinities:

$$|\mathbb{N}| < |\mathcal{P}(\mathbb{N})| < |\mathcal{P}(\mathcal{P}(\mathbb{N}))| < |\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))| < \cdots$$

We have already seen that $|\mathbb{N}| < |\mathbb{R}|$, so we might wonder if $|\mathcal{P}(\mathbb{N})| = |\mathbb{R}|$ — this is, in fact, the case. We can also show that any infinite set is at least as large as $\mathbb{N}$. Therefore, the cardinality of $\mathbb{N}$ is the smallest type of infinity! By convention, this level of infinity is denoted as aleph-naught: $\aleph_0$.

Note. So it is reasonable to wonder if there is a set larger than $\mathbb{N}$ and smaller than $\mathbb{R}$. This very famous question has a surprising answer.

Note. The *Continuum Hypothesis* states that there is no set with cardinality strictly larger than $|\mathbb{N}| = \aleph_0$ and strictly less than $|\mathbb{R}| = c$ (the cardinality of the continuum). Therefore if we accept the Continuum Hypothesis, then it is reasonable to denote $|\mathbb{R}| = \aleph_1$. Another notation (in analogy to the finite case) is to say $|\mathbb{R}| = \aleph_1 = 2^{\aleph_0}$. We can then denote the chain of infinities as

$$|\mathbb{N}| = \aleph_0 < |\mathcal{P}(\mathbb{N})| = |\mathbb{R}| = 2^{\aleph_0} = \aleph_1 < |\mathcal{P}(\mathcal{P}(\mathbb{N}))| = |\mathcal{P}(\mathbb{R})| = 2^{\aleph_1} = \aleph_2 < \cdots$$

However, no one has actually given a proof of the Continuum Hypothesis. In fact, in 1939 Kurt Gödel showed that the Continuum Hypothesis does not contradict the axioms of set theory (i.e., it is consistent with them), and in 1964 Paul Cohen proved that the Continuum Hypothesis does not follow from the axioms of set theory (i.e., it is independent of them). This means that the Continuum Hypothesis is neither true nor false — it is *undecidable* (shades of Gödel's incompleteness here)!!!

Kurt Godel (1906–1978) Paul Cohen (1934–)

Note. A mathematical Platonist is someone who believes that objects of math have some kind of existence independent of the human mind. Such a person would find the undecidability of the Continuum Hypothesis unpleasant! They would likely think that it is *really* either true or false, only that we do not know which. A nonPlatonist (which I am), sees the objects of math as simply symbols which are manipulated according to certain rules — these symbols don't represent anything having any sort of existence independent of the existence given to them by the axiomatic system. A nonPlatonist would consider the Continuum Hypothesis intrinsically neither true nor false. Though if it is helpful to assume its validity as a new axiom, then that could be done and the symbol manipulation taken up from there...

REFERENCES

All images of mathematicians are from MacTutor History of Mathematics archive (`http://www-history.mcs.st-and.ac.uk/`). Most of the information comes out of the head of Dr. Bob Gardner, following the teaching of these ideas for several years, but the following references contain much of the information given here.

1. Kirkwood, J. *An Introduction to Analysis*, 2nd Edition, PWS Publishing, 1995.

2. Wallace, D. *Everything and More: A Compact History of* ∞, W. W. Norton, Great Discoveries series, 2003.