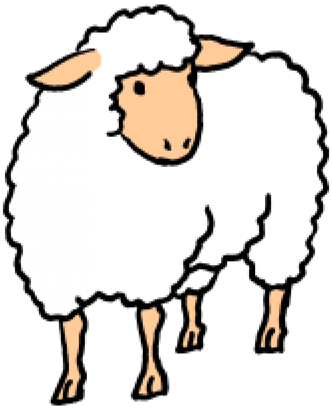
The Birth of Numbers



Steven Pemberton

The Birth of Numbers

Steven Pemberton
Amsterdam

For *u*i*n

3 4 5 6 7 8 9

Contents

- The Birth of Numbers
- The Birth of Numbers
- Contents
- Numbers
- Addition
    - Subtraction
    - Mystery Numbers
    - Zero
- Multiplication
    - One
    - More Mystery Numbers
    - The True Definition of Multiplication
- Power
    - The Complement of Power
    - Analysis
    - Yet More Mystery Numbers
- Representations
- Counting On Your Fingers
- Representing Numbers
- Why 2↓2 Isn’t a Rational Number
- Should there be a ⇑ Operator?
- Naming
- Can We Go Lower?
- Can We Go Higher?
- Examples
    - Interest
    - Computer Speeds
    - Graphing Exponentials
    - The Effect of Moore's Law
    - Networks
    - Population Growth

    - Inflation
    - Prime numbers
    - Why There Is No Perfection
    - Notes
    - Earthquakes
    - Arithmetic
    - Infectious Diseases
    - The Stock Market
    - Mortality
- Do You Really Accept Negative Numbers?
- Complete Numbers
- Generalising Numbers
    - Roof
    - Weights
    - Ball On Sloping Floor
    - Billiard Balls
- Complete Multiplication
- Angles
- What is τ?
- What About the Square Root of Minus Twenty Five?
- Multiplication is Rotation
    - Rotation
    - Understanding the Definition of Multiplication
- Conclusion
- About the Author

Numbers

Numbers are an idea, not a thing: an *abstraction* we would call them.

That is because there is no *three* that you can point to, only three *somethings*, such as three sheep, three trees, three examples. The idea of *three* is what those sheep, trees and examples have in common.

Because *three* is an abstraction, we need to have some way to make that idea real. We do this by *representing* a number in some way.

One way is by writing the name: *three*.

But we have other ways, for instance 3, as we write it, or III as the Romans wrote it. Or 三 as the Chinese write it. These are all representing the same number.

Note that all three of these representations look (a bit) like three stripes.

That’s probably because originally people started writing numbers down as stripes.

Addition

Addition is easy. You want to add numbers: you write a number of stripes for the first number, and then a number of stripes for the second, and count up how many stripes you have: // plus /// is /////. This is how children learn addition, except using fingers for stripes. This is a *procedural* definition: it tells you how to calculate the result.

We write addition as `a+b`.

There are a number of *laws* to do with addition. The most obvious one is

```
a+b = b+a
```

This is called the *commutative law*, and it says that with respect to addition it doesn't matter which way round the operands are, 2+3 will give you the same result as 3+2.

Clearly //+/// is the same as ///+//.

This is a different use of the word *law* than is usually understood in everyday life. Usually laws are things created by people to help run society. In maths it means something that is always true. It might be better to call it a *fact*.

Subtraction

The complementary operation to addition is subtraction, which we can define as

```
(a+b)-b = a.
```

This may look like an odd way to define something, but it is called a *declarative* definition: it doesn't tell you how to calculate the result (although we do learn how to do that at school), it just tells you how to recognise the right answer.

In fact we use declarative definitions a lot. Probably the first time we are taught such a definition is for square roots. We are told that

> *The square root of a number is another number that if you multiply it by itself you get the original number*

or put another way:

> *The square root of a number* n*, is a number* r *such that* $r \times r = n$*.*

This doesn't tell us how to calculate it (and we have calculators to do that for us), but it does allow us to understand what a square root is, and how to recognise one.

What does this definition for subtraction mean?

If you have a number that is the result of adding `a` and `b`, and then subtract `b` from it, we get `a` back.

So, if we want to calculate 9−3, then `(a+b)` is 9, and `b` is 3, so we have to find the value `a` such that

```
(a+3) = 9
```

which is 6.

So subtraction means “find the value such that adding the second operand to it gives us the first operand”.

You might prefer it if I say: imagine that

```
[a+b]
```

means “a number that is the result of adding a and b”, for instance 9. Then

```
[a+b]-b = a
```

is true for all possible values of `a` and `b`.

So with 9−3

```
[9]-3 = a
[a+3]-3 = a
[6+3]-3 = 6
```

Since addition is *commutative*, we also have

```
(b+a)-a = b
```

(or if you prefer

```
[b+a]-a = b
```

)

Thanks to the commutative law, this is then the same as

`[a+b]-a = b`.

In other words, if you have a number that has been formed by addition, you can get either operand of the addition by subtracting the other operand.

Subtraction on the other hand is not commutative:

`2-3 ≠ 3-2`.

This means that although we have

`[a-b]+b = a`,

we don't have

`[a-b]+a = b`.

To get `b` we can do a number of things: we can reverse the operands, and swap the operator:

`a-[a-b] = b`,

or we can negate the first operand and add:

`-[a-b]+a = b,`

or we can define a new operator, let's call it *co-plus* ⊕, whose definition is

`(x⊕y)=(-x)+y`,

and then say that

`[a-b]⊕a=b`.

Note that both of these latter two require the definition of unary −, which we will talk about shortly.

Now I can understand if you are looking at this wide-eyed, wondering what I am going on about. So let me try and explain.

If I say that two people's combined ages are 95, then if I tell you one of the ages, you can work out the other without any more information. If I say one is 50, then you know the other is 45, and vice versa. In both cases you subtract the age from the sum of the ages. This is because addition is commutative.

However, if I had said that the *difference* in age between two people is 15 years and one of them is 45, you don't know enough, because you don't know if 45 is `a` or `b`. The other one could be either 60 or 30. So you need two complements for subtraction, one to get the left operand and the other to get the right operand.

Mystery Numbers

In the act of defining subtraction there is something surprising that has happened, that modern people might not even notice, but that upset early mathematicians. Namely, it has generated new sorts of number that we didn't start off with. All of a sudden we can calculate results like "0−5" which is "the number that adding five to gives zero" (or "the number that you would get if you subtracted 5 from 0").

But the problem is, how do you write such a number down? You can't write it with stripes, like we started off with, because they are a sort of "anti-number". It must have upset early mathematicians, because you now have numbers that apparently don't exist in the real world, but they seem to obey all the rules of regular, real-world numbers.

Well, as is so often seems to have been the case with new types of numbers, they got given a very derogatory name. They could have chosen "mystical numbers" or "magical numbers" or "unreal numbers", but they chose "negative numbers". But they still had no way to write these numbers down.

Through the ages people have written negative numbers in different ways. Banks used to write them in red ink (which is why we (still) use the phrase “in the red”); annual reports of companies often use a number in brackets like “(123)” to mean a negative amount.

In fact, in mathematics we leave negative numbers uncalculated: “0−5” means “the number you would get if you subtracted 5 from 0”. Since it was always 0 that was used, it could be shortened to just “−5”; this is not strictly speaking a negative number, even if that is how we think of it nowadays: it is an operation on a positive number, an operation that we leave uncalculated, since it is good enough, or at least, the best we have.

Zero

A question that even now not all mathematicians completely agree on is whether before defining subtraction we already had zero, or whether the act of defining subtraction introduced it.

Certainly early mathematicians didn't think of zero as a number (they even had difficulties with 1): if you didn't have any sheep, how could you say that you had a number of sheep?

Either way, mathematicians now agree that zero is a number, just not whether it was there right from the start.

Zero has a special relationship with addition and subtraction. It is called the *identity value* for addition and subtraction. This is shown by the identity law for addition and subtraction:

```
a+0  =  a
a−0  =  a
```

In other words `a` remains the same (it 'keeps its identity') if you add 0 to it, or subtract 0 from it.

Multiplication

Now we’re going to go a level higher, and do almost exactly the same we just did with addition.

You can describe multiplication in terms of repeated addition:

$$a\times 3 = a+a+a$$

The complement of multiplication is division, which we can define in the same way as we did for subtraction:

$$[a\times b]\div b = a.$$

This means “find the value such that multiplying it by the second operand gives us the first operand”. Just as with addition, multiplication is commutative, and so we can also derive

$$[a\times b]\div a = b.$$

Similarly, just like subtraction, division is not commutative, since $2\div 3 \neq 3\div 2$. So again, although we have

$$[a\div b]\times b = a,$$

we do not have

$$[a\div b]\times a = b.$$

Again, we reverse the operands, and swap the operator:

$$a\div [a\div b] = b.$$

One

Multiplication and division also have an identity value, 1, since

```
a×1 = a
```

and

```
a÷1 = a.
```

We've already discussed how `-a` is a shorthand for `0-a`, "the number you would get if you subtracted `a` from 0", and mentioned that `0` is the identity value for addition and subtraction.

It is therefore notable that there is no unary operator `÷a` to represent `1÷a`. However, if you introduce it, there are some rather pleasant symmetries that arise. For instance

```
a+(-b) = a-b
a×(÷b) = a÷b
```

and

```
a-(-b) = a+b
a÷(÷b) = a×b
```

and

```
-(-a) = a
÷(÷a) = a
```

Do you see the patterns?

There is something else interesting too, which you can see from this comparison:

```
-(a-b) = b-a
÷(a÷b) = b÷a
```

What you can see from this is that unary − and ÷ are both “commute” operators. They have the effect of commuting the operands of the thing they are applied to (without having to know what the values of those operands are).

You could say that while + and × are commutative, − and ÷ are commut*able*.

Earlier we said that to extract the left-hand operand from `(a-b)` we could use `(a-b)+b`, but to extract the right-hand operand we had to either use `a-(a-b)` or `-(a-b)+a`.

This shows that from a conceptual point of view, the second of these has a certain charm: `(a-b)+b` extracts the left-hand operand `a`, and `-(a-b)+a` commutes the operands, giving `(b-a)+a`, and then extracts the left-hand operand of the result, giving `b`, the right-hand operand of the original.

So just as we defined the extraction of `b` from `a-b` as `-(a-b)+a`, we could also define the extraction of `b` from `(a÷b)` as `÷(a÷b)×a`.

Or we could define a new operator *co-times* `⊗` defined as `(x⊗y)=(÷x)×y`, which commutes and extracts the left-hand operand, and use `(a÷b)⊗a=b`.

The `÷` unary operator will be used some more shortly.

More Mystery Numbers

Just as with subtraction, division introduces a new sort of number that we didn’t start with.

We started with just positive numbers; thanks to subtraction, we got zero and negative numbers, but now thanks to division we have yet another sort of number. We can get numbers such as 2÷3, “the number that if you multiply by 3 gives you two”. Although this is often written as ⅔, we should recognise it as another example of an incompletely calculated number.

Just as with negative numbers, early mathematicians didn't consider these as numbers: they were just 'ratios'. When they finally did get accepted into the numbers club, they at least got a friendlier name: rational numbers, reflecting that they were considered ratios.

It is also worth pointing out that even though we didn’t start out with these numbers, they still follow the rules of arithmetic for addition and subtraction, although we have to add a couple of equalities:

```
a÷b + c÷d = (a×d + c×b)÷(b×d)
(a×n)÷(b×n) = a÷b
```

The first of these tells you how to express a sum of two rationals as a single rational. The second says that rationals such as `2÷4` and `3÷6` are the same as `1÷2`.

The True Definition of Multiplication

To be honest, when I defined multiplication, I did a bit of hand-waving, by just giving an example. I said

$$a\times 3 = a+a+a$$

But this is not a true definition, just an example.

You could define it declaratively as

$$a\times(b+1) = a\times b + a$$

which along with $a\times 1 = a$ which we already saw, allows you to conclude that

$$\begin{aligned} a\times 3 &= a\times(2+1) \\ &= a\times 2 + a \\ &= a\times(1+1) + a \\ &= a\times 1 + a + a \\ &= a + a + a \end{aligned}$$

but an equivalent and eventually more useful declarative definition is:

$$a\times(b+c) = a\times b + a\times c$$

What does this say? It says that you can split a multiplication up into a series of additions. (Again you might prefer the notation $a\times[b+c]$.)

For instance 3×4:

```
3×4 = 3×(2+2) = 3×2 + 3×2
3×2 = 3×(1+1) = 3×1 + 3×1
```

So therefore

```
3×4 = (3×1 + 3+1) + (3×1 + 3+1)
```

And since

```
3×1 = 3
```

we have

```
3×4 = 3 + 3 + 3 + 3.
```

You could do the same by starting out with

$$3\times 4 = 3\times(3+1)$$

(as we did above), and you would still end up with

$$3\times 4 = 3 + 3 + 3 + 3.$$

But how about

$$3\times 0$$

though?

Well, we know the identities

$$\begin{aligned} 3\times 1 &= 3 \\ 1+0 &= 1 \end{aligned}$$

So therefore we can work out that

$$3\times(1+0) = 3\times 1 + 3\times 0$$

and since $1+0 = 1$, we can also work out that

$$3\times 1 = 3\times 1 + 3\times 0$$

and so we have to conclude that

$$3\times 0 = 0.$$

The reason that

`a×(b+c) = a×b + a×c`

is a more useful definition than

`a×(b+1) = a×b + a`

is that it allows us to work out what addition and multiplication mean for fractional values as well, because there’s nothing in the `b` and `c` that requires them to be whole numbers.

As an example, what does `a×0.5` mean? Well

`a×0.5 + a×0.5 = a×(0.5+0.5) = a×1 = a`

So `a×0.5` means *a half of* `a`.

Interestingly, you could see *half* as a lower-level version of square root: the square root of a number is another number that when multiplied by itself gives the original number; the half of a number is another number that when *added* to itself gives the original number.

By the way, the rule

`a×(b+c) = a×b + a×c`

is called the *distributive* law in mathematics, but you should be aware that this is not an emergent property, but an actual definition.

Power

Now to go up yet one more level, using exactly the same patterns.

We can describe the operation of taking something to a power in terms of repeated multiplication:

$$a^3 = a \times a \times a.$$

Unfortunately, mathematics notation becomes odd at this level, and rather than using operator-style, like +, −, ×, and ÷, it starts using strange layout-style notations, as if there were something different going on.

Let's try and fix this. Here we will use a notation that is already used in some computer circles, and on some calculators, to represent taking the power, the ↑ operator, so that $a\uparrow 3$ means what classically is represented by a^3.

Of course, I've waved my hands above again, but this time I hope you can work out for yourself what the true declarative definition of power is: it uses the same pattern as multiplication.

If you do it, you will see that we can work out that just as

$$a \times 0 = 0$$

(which is the identity value for addition) we get

$$a \uparrow 0 = 1$$

(the identity value for multiplication).

The Complement of Power

Without thinking (yet) about what it means, we can also invent a complementary operation for ↑, which we will write as ↓; if ↑ is a higher-level form of multiplication, then ↓ is a higher-level form of division. Just as in the other cases, we define

`(a↑b)↓b = a`.

Unfortunately there is a difference at this level, since unlike + and ×, ↑ is not commutative: 2↑3 is 8 and 3↑2 is 9, which are clearly not equal, so we cannot just juggle operands as we did with addition and multiplication.

Instead we will invent another complementary operator to get `b`. Let's call it ⇓, so that we have

`(a↑b)⇓a = b`.

(It may help you remember which is which by noting that ↓ gives you the first operand, and ⇓ gives you the second operand of ↑)

Well, let’s now explain what these two operators mean: ↓ means “find the value that when taken to the power of the second operand gives us the first operand”. For instance

`8↓3 = 2`,

since

`2↑3 = 8`

and

`[2↑3]↓3 = 2`.

In other words, it takes a root, so that a↓b means what is classically expressed with the, frankly bizarre, notation $^{b}\surd a$.

Similarly, ⇓ means “find the value that when the second operand is raised to that power, gives the first operand”, for instance

`8⇓2 = 3`

since

`2↑3 = 8`

and

`[2↑3]⇓2 = 3`

In other words, it takes the logarithm, so that a⇓b means what is classically, and equally bizarrely, written $log_b a$.

(It can help, while learning to get used to these operators, to note a slight visual similarity between √ and ↓, and to regard ⇓ as looking a bit like a log, while not forgetting that the operands are the other way round to how they are normally written.)

The nice thing about the new notation is that it far more visually obviously expresses those relationships between power, root and log. And it uses the same methods for solving equations:

$$a+2 = 4$$
$$a = 4-2$$

$$a\times 2 = 4$$
$$a = 4\div 2$$

$$a\uparrow 2 = 4$$
$$a = 4\downarrow 2$$

$$2\uparrow a = 4$$
$$a = 4\Downarrow 2$$

But equally nice, it also suddenly exposes that literally dozens of well known identities have a similar form to identities at the two lower levels! These are listed in a table shortly, but let me just pick a couple of beauties out:

$$a\times(-b) = -(a\times b)$$
$$a\uparrow(-b) = \div(a\uparrow b)$$

This is the identity normally expressed as:

$$a^{-b} = 1/a^{b}$$

Here's another:

$$a\times(\div b) = a\div b$$
$$a\uparrow(\div b) = a\downarrow b$$

I consider this so beautiful, that ÷ just flips the operator, that even if this had been the only result, the whole exercise would have been worth it! This is the identity traditionally expressed as

$$a^{1/b} = {}^{b}\sqrt{a}$$

And now a third:

$$\div(a\div b) = b\div a$$
$$\div(a\Downarrow b) = b\Downarrow a$$

Traditionally expressed as

$$1/\log_b a = \log_a b$$

In other words ⇓ is commutable.

I may have learned this at school, but if I did, I had forgotten it. When I saw this result, I was so amazed, that even though I could prove it, I have to admit I checked it with some values as well, to make sure it was really true.

So now for completeness, to properly deal with the two new operators: how do you extract the operands from them?

For ↓ we have

`[a↓b]↑b = a`,

and

`a⇓[a↓b] = b`.

For ⇓ we have

`b↑[a⇓b] = a`

and

`a↓[a⇓b] = b`.

And to summarise the relationships, to more clearly show the patterns involved:

Operator	**a**	**b**
a+b	(a+b)−b	(a+b)−a
a−b	(a−b)+b	a−(a−b) *or* −(a−b)+a *or* (a−b)⊕a
a×b	(a×b)÷b	(a×b)÷a
a÷b	(a÷b)×b	a÷(a÷b) *or* ÷(a÷b)×a *or* (a÷b)⊗a
a↑b	(a↑b)↓b	(a↑b)⇓a
a↓b	(a↓b)↑b	a⇓(a↓b)
a⇓b	b↑(a⇓b)	a↓(a⇓b)

Just as when we defined the unary version of ÷, and we got some new equalities that had the same patterns as with unary −, with these three new operators, as I already noted, we can observe some pleasant similarities with equalities that we already know from addition and multiplication. For instance, the basic operations:

Addition	**Multiplication**	**Power**
$a+0 = a$ $a-0 = a$	$a\times 1 = a$ $a\div 1 = a$	$a\uparrow 1 = a$ $a\downarrow 1 = a$
$a-a = 0$	$a\div a = 1$	$a\downarrow a = 1$ $a\Downarrow a = 1$
	$a\times 0 = 0$ $0\div a = 0$	$a\uparrow 0 = 1$ $0\downarrow a = 0$ $1\Downarrow a = 0$
	$a\div 0 = \text{undefined}$	$a\downarrow 0 = \text{undefined}$ $a\Downarrow 0 = \text{undefined}$ $a\Downarrow 1 = \text{undefined}$

Similarly, there are several patterns with the unary operators:

Addition	**Multiplication**	**Power**
	a×(−b) = −(a×b) a÷(−b) = −(a÷b)	a↑(−b) = ÷(a↑b) a↓(−b) = ÷(a↓b)
a+(−b) = a−b a−(−b) = a+b	a×(÷b) = a÷b a÷(÷b) = a×b	a↑(÷b) = a↓b a↓(÷b) = a↑b
a+(0−a) = 0	a×(1÷a) = 1	a↑(1↓a) = 1

And finally, there are lots of patterns over the binary operators:

Addition	**Multiplication**	**Power**
	$a\times(b+c)=a\times b+a\times c$ $a\times(b-c)=a\times b-a\times c$	$a\uparrow(b+c)=a\uparrow b\times a\uparrow c$ $a\uparrow(b-c)=a\uparrow b\div a\uparrow c$
	$(a+b)\times c=a\times c+b\times c$ $(a-b)\times c=a\times c-b\times c$ $(a+b)\div c=a\div c+b\div c$ $(a-b)\div c=a\div c-b\div c$	$(a\times b)\uparrow c=a\uparrow c\times b\uparrow c$ $(a\div b)\uparrow c=a\uparrow c\div b\uparrow c$ $(a\times b)\downarrow c=a\downarrow c\times b\downarrow c$ $(a\times b)\Downarrow c=a\Downarrow c+b\Downarrow c$ $(a\div b)\downarrow c=a\downarrow c\div b\downarrow c$ $(a\div b)\Downarrow c=a\Downarrow c-b\Downarrow c$
	$(a\times b)\times c=a\times(b\times c)$ $(a\div b)\div c=a\div(b\times c)$	$(a\uparrow b)\uparrow c=a\uparrow(b\times c)$ $(a\downarrow b)\downarrow c=a\downarrow(b\times c)$
$(a+b)-c=(a-c)+b$ $(a-b)-c=(a-c)-b$	$(a\times b)\div c=(a\div c)\times b$ $(a\div b)\div c=(a\div c)\div b$	$(a\uparrow b)\downarrow c=(a\downarrow c)\uparrow b$ $(a\uparrow b)\Downarrow c=(a\Downarrow c)\times b$ $(a\downarrow b)\downarrow c=(a\downarrow c)\downarrow b$ $(a\downarrow b)\Downarrow c=(a\Downarrow c)\div b$
$a-b=(a-k)-(b-k)$	$a\div b=(a\div k)\div(b\div k)$	$a\Downarrow b=(a\Downarrow k)\div(b\Downarrow k)$
$a+(b-c)=a-(c-b)$	$a\times(b\div c)=a\div(c\div b)$	$a\uparrow(b\Downarrow c)=a\downarrow(c\Downarrow b)$
$-(b-a)=a-b$	$\div(b\div a)=a\div b$	$\div(b\Downarrow a)=a\Downarrow b$

Analysis

Looking at these similarities, there is one group that particularly sticks out:

$a+(-b) = a-b$	$a\times(\div b) = a\div b$	$a\uparrow(\div b) = a\downarrow b$
$a-(-b) = a+b$	$a\div(\div b) = a\times b$	$a\downarrow(\div b) = a\uparrow b$

This seems to suggest there should be a unary ↓ operator:

$a+(-b) = a-b$	$a\times(\div b) = a\div b$	$a\uparrow(\downarrow b) = a\downarrow b$
$a-(-b) = a+b$	$a\div(\div b) = a\times b$	$a\downarrow(\downarrow b) = a\uparrow b$

Although $\downarrow b$ means the same as $\div b$, the other unary operators are expressed in terms of their own level of operators, and so you can say that $\downarrow b$ means $b\uparrow -1$, which you can prove is the same as $b\downarrow -1$. This also has the nice property that $a\uparrow(\downarrow a) = 1$, just as $a\times(\div a)=1$ and $a+(-a)=0$.

It can also be applied to the last set of equations above:

$a-b = -(b-a)$	$a\div b = \div(b\div a)$	$a\Downarrow b = \downarrow(b\Downarrow a)$

Note its pleasant effect as a commuting operator for ⇓.

Yet More Mystery Numbers

Of course, I hope by now you are expecting us to come out of this level with more types of numbers than we began with, because you won’t be disappointed: we get *two* new types.

For several centuries it has been known that a simple expression such as 2↓2 (the square root of 2) cannot be expressed as a rational number.

This class of numbers got the derogatory name “irrational” (although the 9th century mathematician al-Khwārizmī called them *inaudible*, which is why we still call any unresolved irrational root a *surd*, Latin for *deaf*).

These numbers also still work at the lower levels of multiplication, division, addition and subtraction.

The combination of the rationals and irrationals were given the name 'real' numbers, in itself somewhat a misnomer, since 'continuous' numbers might have been a better name.

The other new type of number introduced by the complement of power is the horribly misnamed *complex numbers*, which historically emerged as solutions of expressions such as (−2)↓2. These will be the topic of a later section.

Representations

How you represent numbers can have a great effect on what you can do with them.

For instance, the Romans may well have been able to add CCCLXI and CCXXI together, but few would have been able to multiply them.

And yet nowadays, ask a schoolchild to add 361 and 221, or even multiply them, and they will be able to do it for you.

In fact even in the middle ages, multiplication was still something you learned at university. It wasn't until the 1600s that modern numbers started to be introduced and used in Europe.

Our modern numbers come from Arab mathematicians, around the year 1000, who in turn got them from Indian mathematicians, from around the year 600.

Look at this piece of Arabic text, taken from the Wikipedia article in Arabic about the Second World War. Does anything about it strike you?

الحرب العالمية الثانية: هي نزاع دولي مدمر بدأ في الأول من سبتمبر 1939 في أوروبا وانتهى في الثاني من سبتمبر 1945، شاركت فيه الغالبية العظمى من دول العالم، في حلفين رئيسيين هما: قوات الحلفاء ودول المحور. وقد وضعت الدول الرئيسية كافة قدراتها العسكرية والاقتصادية والصناعية والعلمية في خدمة المجهود الحربي، وتعد الحرب العالمية الثانية من الحروب الشمولية، وأكثرها كلفة في تاريخ البشرية لاتساع بقعة الحرب وتعدد مسارح المعارك والجبهات فيها، حيث شارك فيها أكثر من 100 مليون جندي، وتسببت بمقتل ما بين 50 إلى 85 مليون شخص ما بين مدنيين وعسكريين، أي ما يعادل 2.5% من سكان العالم في تلك الفترة.

Well, of course they use what we call Arabic numerals. But the interesting thing is that Arabic is written from right to left, and yet the numbers are written in exactly the same way as we do, 1939, and not, as you might expect 9391.

How come? Well, when Western mathematicians imported modern numerals from Arabic, they forgot to swap the order from right-to-left to left-to right.

Does this matter? A little.

Try adding two large numbers up in your head: add 1939 and 1945.

Do it now.

The answer if you got it right is 3884. But because we usually add numbers starting from the right end and working leftwards, we have to do all the addition for the whole number before saying what the result is, working from the left end rightwards. Put another way: we add numbers right to left, the same direction that Arabic text is read.

However, if numbers were the other way round, and I asked you to add the same two numbers, now written 9391 and 5491, you could speak the answer *while* doing the calculation:

> 9+5 is **4** (carry 1)
>
> 3+4 is 7, plus the carry **8**
>
> 9+9 is **8**, carry 1
>
> 1+1 is 2, plus the carry **3**

So the answer would be 4883 if they had swapped the direction on importing, and you would be able to do the calculation in your head and speak the answer while doing it. (Some computer programs on numbers would be simpler too for similar reasons.)

In fact for small numbers, we do speak our numbers that way round: we say FOURteen, SIXteen, SEVENteen. That is we say the second digit first.

Similarly, you probably know this children's rhyme:

> Sing a song of sixpence, a pocketful of rye.
> **Four and twenty** blackbirds baked in a pie.

If you look at Shakespeare's plays, you will see around two-thirds the usage of numbers is of the style "four and twenty", and the other third are of the modern style "twenty-four". This probably indicates that the style of speaking numbers was changing around Shakespeare's time.

Another interesting thing to note is that while English text such as this book is normally left-aligned, the Arabic text above is aligned on the right. But you'll notice that when we make columns of text and numbers, the text is left-aligned, and the numbers right-aligned:

```
John       300
Mary      1200
Douglas     42
```

This is because our number representation has been imported from a culture where text is right-aligned.

So the conclusion is, the style of writing numbers we use now are much easier than Roman numerals, but they could be (a little bit) easier if they had been imported the right way round.

Counting On Your Fingers

Why do we use ten as the base for our numbers? The usual answer is that it’s because we have ten fingers, and people would say that this is so clearly and undeniably obvious, it’s not worth discussing.

But to be honest this explanation has never fully satisfied me.

Firstly, not all civilisations have used base ten (for instance the Mayans used 20, the Babylonians 60, and the Egyptians 12), and even those that have used ten have not necessarily used it exclusively, like dozens for eggs, or 12 pence to a shilling, twenty shillings to a pound, or twelve months to a year, or 24 hours to a day, or seven days to a week, or multiples of two (such as 8 pints to the gallon, 16 ounces to the pound), or 1760 yards to a mile, so ten can’t be *that* obvious.

Secondly, because on the basis of the explanation of how this came about (for instance the shepherd counting sheep on her fingers until she gets to ten, and then dropping a stone in her pocket and starting again, and later counting the number of stones) it means that there are two different representations of ten: all the fingers, or no fingers and a stone; and thirdly, because I think people are more intelligent than that.

Or let me put it another way: *digit* comes from the Latin word for *finger*, but there are only 9 digits from 1 to 9, and ten fingers. What happened with that extra finger?

Now let me explain how *I* count on my fingers. I count to 4 in the ‘usual’ way, using my fingers:

1: |
2: ||
3: |||
4: ||||

but my thumb counts as 5:

5: \

and then I use my fingers again:

6: \|
7: \||
8: \|||
9: \||||

and then when I get to ten, I raise a finger on the other hand, and close all my fingers (and thumb) on the first hand.

And so twenty is two fingers on the second hand, and fifty is just the thumb on the second hand, and so on, all the way up to 99. In other words I can count to 100 on my fingers in a very natural way.

The advantages of doing it this way are that you can count to 99 on two hands, or count to ten on one hand and use the other hand for something else, like dropping stones into your pocket; and every number has a unique representation: no dual representation for ten.

Well, I’ve been counting this way since I discovered it in my youth, but recently something struck me: Roman numerals look like a picture of this. (I deliberately drew counting on my fingers above in that way to suggest it)

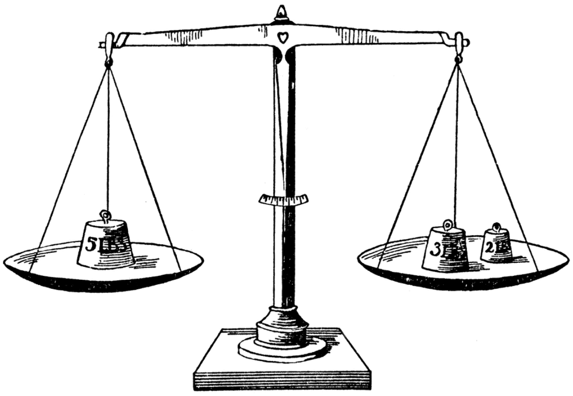
Roman numerals use letters to represent numbers. A capital I is a fairly obvious choice for the representation for one, and C is the first letter of the word *Centum*, meaning hundred, and M is the first letter of the word *Mille*, meaning thousand, but there has never been a clear explanation for me of V for five and L for fifty. But you have to admit, a V does look a bit like a thumb sticking out of a hand, as does an L. (An X for ten looks like two Vs on top of each other; the Romans didn’t have the letter W which wasn’t invented until much later, and by the way, the version of prefacing letters that are subtracted, like IX for 9 was a later development: originally it was VIIII).

This is pure speculation on my behalf, I have no hard evidence for it, but I do find it very compelling: the reason we count in tens is because the Romans did; the reason Romans counted in tens is because we have four fingers and a thumb on each hand.

Representing Numbers

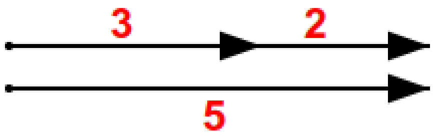
As was said at the beginning, all numbers are abstractions. Early mathematicians had problems accepting the concept of negative numbers, because they didn't seem to occur in real life. But in fact positive numbers don't 'occur' in real life either. As already pointed out, you can't point to "the number three".

However, we are not very good with abstractions in general. We like to think about them in some solid form, and so we represent numbers in some way or another in order to be able to talk about them. You can do that with three apples, you could do that with weights.



3+2=5 using weights

But since we are primarily visual, we typically represent numbers with lengths, drawing a line, and marking particular numbers on it. In mathematics, there is even the concept of “the real number line”.

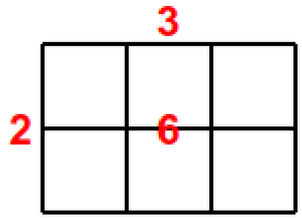


3+2 = 5 using lengths

However, there is nothing essential to this representation of numbers, and in fact a problem with concretisation of abstractions is that there is a risk of mixing the two up and thinking that the concretisation *is* the abstraction. A line is just one of many possible representations of numbers.

Sometimes we switch between representations too, often without realising it.

For instance, when explaining multiplication, we often draw a rectangle whose sides are of the length of the two numbers being multiplied, and explain that the area of the rectangle is the product of the two sides.

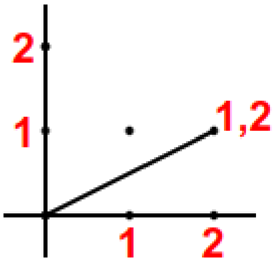


3×2 = 6

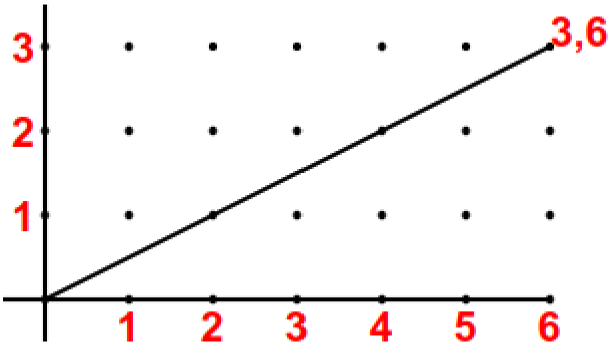
In so doing, we switch from a linear representation of numbers to an area-based representation.

Rational numbers, such as ¾, can be represented by a pair of integers (3, 4). One way of representing rational numbers is to create a two-dimensional field of dots, where in one dimension the integers are marked for the numerator, and in the other dimension for the denominator.

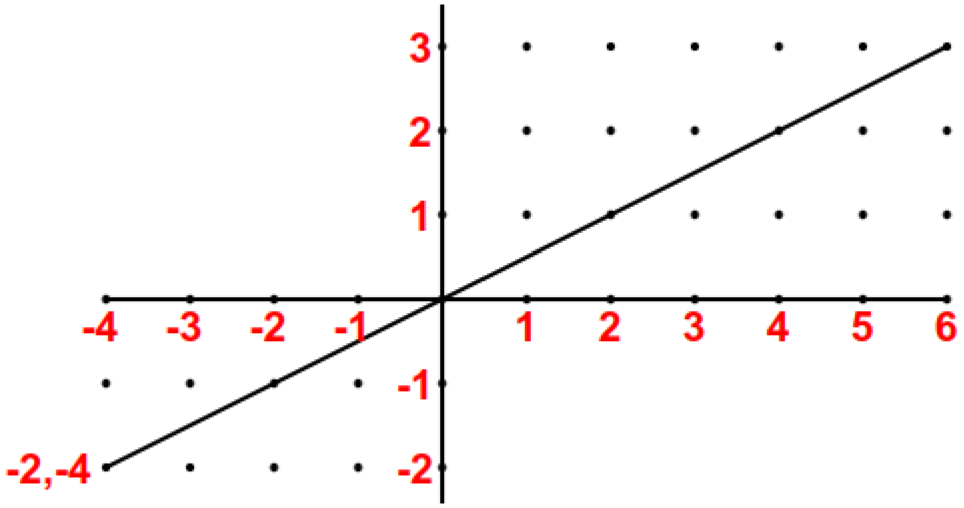
So drawing a line from the origin (0,0), through the point (1,2), represents the number ½.



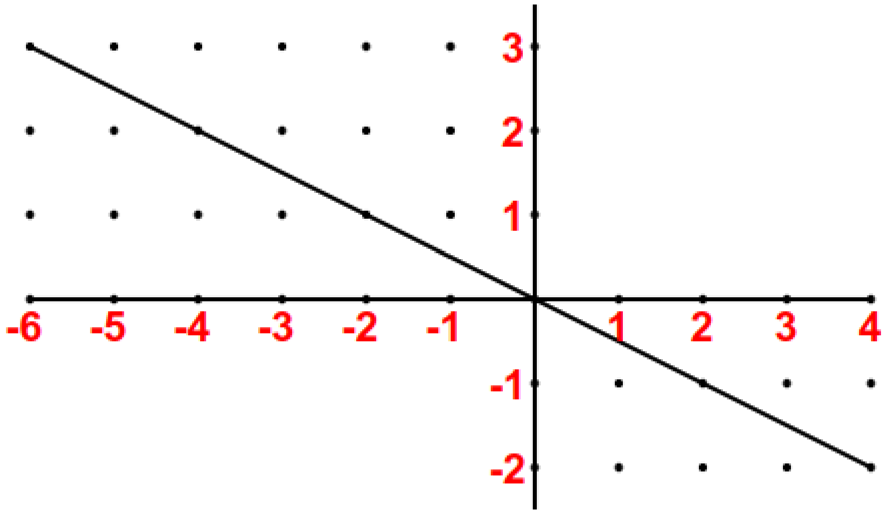
It's the line that represents the number, and if you extend it, it passes through the points (2,4) and (3, 6) as well, demonstrating that these are equivalent rational numbers.



Furthermore, it passes in the other direction through (−1, −2) and (−2, −4) and so on, also in the process demonstrating that two minuses make a plus.

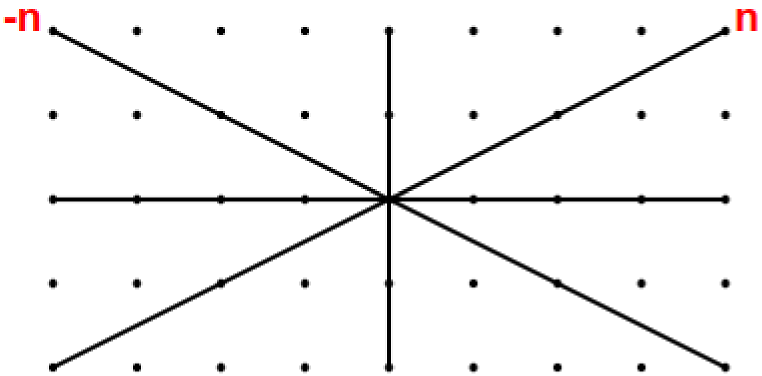


Similarly, a line from the origin to (−1, 2) represents the rational number −½, which likewise also passes through (−2, 4) as well as (1, −2) and (2, −4) , and so on.

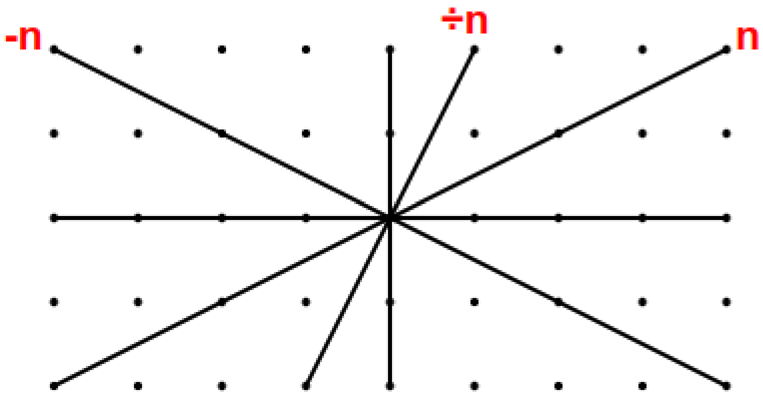


This representation has some other advantages. For instance, it also shows you that all points (0, x) represent zero, all points (x, 0) represent infinity (and actually allows you to represent infinity), and finally that (0,0) is indeterminate, since it represents all numbers, all lines passing through it.

It also has some other nice properties: the representation of the number −n is just the representation of n reflected by either the x or y axis.



Similarly, the inverse of n, ÷n, is n negated (i.e. reflected on either axis) and rotated 90 degrees (it doesn't matter in which direction), or equivalently rotated and then negated. In either case, there is always a 90° angle between −n and ÷n.



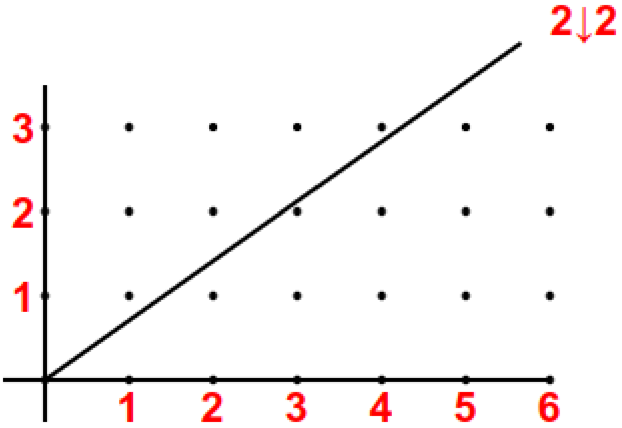
So: if the angle of the line for the number n is a, then

$$-n = 180° - a$$

and

$$÷n = 90° - a.$$

This representation can also be used to describe how the line representing a non-rational, such as 2↓2 (the square root of 2), does not pass through any integral point at all, no matter how far you extend it, because it is not a rational number.



Why 2↓2 Isn’t a Rational Number

Can we prove that the line for 2↓2 doesn't pass through any points?

Well, if it did, that is to say if 2↓2 were a rational number then it could be represented as

```
    a÷b
```

for some a and b, reduced to their lowest form.

Note that when you have a rational number like that, that while `a` and `b` can both be odd numbers, they can’t both be even, because if they are both even, you can divide them both by 2 to get a yet lower form. For instance

```
    4÷6
```

can be reduced further to

```
    2÷3
```

Reduced to their lowest form, they can both be odd, or one can be even, but they can’t both be even.

So, if it is true that the square root of 2 can be represented by `a÷b`, then we have:

`2↓2 = a÷b`

Square both sides:

`2 = (a÷b)↑2`

Expand the brackets:

`2 = a↑2 ÷ b↑2`

Multiply both sides by b↑2:

`2×b↑2 = a↑2`

This means that `a↑2` is an even number, and therefore `a` must be even too (since an odd number squared is always odd).

If `a` is even, `b` *must* be odd. Let’s try and find `b`.

Since `a` is even, then it is a multiple of some other number `c`:

`a = 2×c`

So since we had

`2×b↑2 = a↑2`

Substituting for `a` gives us

`2×b↑2 = (2×c)↑2`

Expand the brackets:

`2×b↑2 = 4×c↑2`

Divide both sides by 2:

`b↑2 = 2×c↑2`

So `b↑2` is even, which means using the same arguments as above, that `b` has to be even. So that means it is impossible to find an odd number that is the second operand of `a÷b`.

Since we can't do it, it can't exist. Therefore there are no integers `a` and `b` such that

`(a÷b)↑2 = 2`.

The square root of two is not a rational number.

In fact, it really shouldn’t be so surprising.

If we go down one level, there are rational numbers

`a÷b`

that are not integers, just as there are roots

`a↓b`

that are not rationals.

Should there be a ⇑ Operator?

Since there is a ⇓ operator, could we define the opposite of it as

`(a⇓b)⇑b = a`

just like

`(a↓b)↑b` ?

The answer is yes we could do that, but it doesn't give us any extra functionality, since to get `a` from `(a⇓b)` we can use

`b↑(a⇓b)`

just as we don't need a new operator

`(a−b)⊕a = b`

to get `b`, since we can just use

`a−(a−b) = b`.

That is to say, that if we defined ⇑ like that we would have

`a⇑b = b↑a`

i.e., just the same operator, but with its operands the other way round. However, defining it does no harm, so if you feel the need to use it, be my guest!

Naming

How should we name the new operators? Again, it would be good to look to see if we can use consistency as a guide. So how are the existing operators named:

Operation	Written	Spoken
Addition	$a+b$	a plus b
Subtraction	$a-b$	a minus b
Multiplication	$a\times b$	a times b, or a multiplied by b
Division	$a\div b$	a over b, or a divided by b
Raising to the power	a^b	a to the power (of) b
Taking the root	$\sqrt[b]{a}$	b-th root of a
Finding the logarithm	$\log_b a$	log to the base b of a

For addition and subtraction, the phrase just uses the names of the operators, neither left nor right operand playing any special role; for multiplication, a sort of adjectival form is used ("How many times have I told you?" "Four times") to describe the process of multiplication, with the right hand side being the subject, or a form where the left hand side is the subject being acted upon by the right hand side; for division either a description of the layout is used ('over'), or again the left hand side is the subject, which is acted upon by the right hand side; and for power, the left hand side is again the subject having something done to it.

So, from a consistency point of view, it’s a mess, with little to draw from. I am inclined to use ‘up’ and ‘down’ for ↑ and ↓, modelled on plus and minus, although I sometimes pronounce something like “2↓3” as “two rooted three”, when I want to emphasise the relationship to the traditional naming, using the “left hand side acted upon by the right hand side” model. That leaves us with ⇓, which I am inclined to pronounce as “logged” or “based”, again using the left hand side as subject.

Can We Go Lower?

We started with positive integers, and addition, and then showed that multiplication is a higher level based on addition, and that raising to the power is a yet higher level based on multiplication.

So could it be that addition is actually based on a lower level? And if so, what is it?

The answer is yes: the lower level is very impoverished, but addition is indeed based on it.

At this lowest level, there is but a single value, and it looks like this:

We call this blank space "zero", and when we want to represent it, in order to make it visible we can draw a circle around it, like so: 0.

There is also a single operation, called "increment". And it looks like this: /

So if we take zero, and increment it once, we get:

    /

If we increment zero twice, we get:

    //

three times:

    ///

and so on.

That’s all we have at the lowest level: zero, and increment.

As you can see, in the true tradition of these discussions, /// is just an uncomputed value of zero incremented three times.

But now, going up one level, and using the same sort of pattern we used to define multiplication in terms of addition, which was

```
a × (b+c) = a×b + b×c
a × 1 = a
```

instead of waving our hands and saying "addition is sticking two integers together", we define addition declaratively in terms of increment:

```
/a + b = a + /b
   + a = a
```

So, if I want to know the value of

//+///

I apply the first equality once, to give

//+/// = /+////

and then once more to give

/+//// = +/////

and then apply the second equality to give

+///// = /////

thus concluding that

//+/// = /////

There are several things you can see here:

1. This defines what "sticking two numbers together" means.
2. It defines zero as the identity value for addition.
3. Because "=" is commutative, so is "+".
4. It defines ‘monadic’ `+`, and shows that `+a` is the same as `0+a`, just as we claimed for monadic `-`.

What you can also see from this is that although we said earlier that there was disagreement amongst mathematicians about whether zero was there from the start or not, what we can conclude is that it really was there, but that we just didn’t see it.

What this *also* means is that we have managed to construct the integers, negative numbers, rationals, real numbers (and shortly Complete Numbers) all from the basis of a single value, zero, and a single operator, increment (well, OK, and equality).

Can We Go Higher?

There was a clever mathematician, Wilhelm Ackermann, who in the late 1920’s also saw the similarities between the definitions of × and ↑, which he defined like this:

```
a↑b = (a↑(b-1)) × a
a×b = (a×(b-1)) + a
```

and numbering the operators 3 for ↑, 2 for ×, 1 for + and 0 for increment, defined a function

```
calc(a, op, b)
```

that after dealing with some initial special cases, ended with the calculation

```
return calc(calc(a,op,b-1),op-1,a)
```

which is a generalisation of the two equalities above.

The initial special cases are:

```
if op = 0: return a+1
if (op=1 and b=0) or
   (op>1 and b=1): return a
```

The first line deals with increment, and the second with the identity values, 0 for + and 1 for the rest, thus defining all the operators (over the positive integers) only in terms of increment and the identity values. (His function differs in detail from what I’ve presented here; you will also find a similar function with only 2 parameters that came later that is wrongly identified as the Ackermann function; it should properly be called the Ackermann–Péter function to distinguish it).

Clearly as Ackermann’s function implies, there are levels above ↑. For instance with the next level above ↑, let's call it ↟, `2↟3` would represent `2↑2↑2`, and be calculated as:

`calc(2, 4, 3) = 16`

and `3↟2`, which represents `3↑3`, is calculated with

`calc(3, 4, 2) = 27`

There are difficulties at this level, because while + and × are *associative*, which means that

`(a + b) + c = a + (b + c)` and
`(a × b) × c = a × (b × c)`

this is not true for ↑; for instance:

`(2↑2)↑3 = 64` and `2↑(2↑3) = 256`

As a result, we can’t define ↟, in the same way as we did before, as

`a ↟ (b+c) = (a↟b) ↑ (b↟c)`

because it’s just not true. For instance

`2 ↟ 5 = 2 ↑ 2 ↑ 2 ↑ 2 ↑ 2`

which is 65536, but

`(2 ↟ 3) ↑ (2 ↟ 2) = 8 ↑ 4 = 4096`

and that means we can’t define what ↟ means with fractional values. We are forced to define a↟b in the Ackermann way, only over integers:

`a↟b = (a↟(b−1))↑a`

(Note that this defines 2↟3 as (2↑2)↑2 and not as 2↑(2↑2)).

So in other words anything higher than ↑ is not really useful, or interesting.

Examples

The purpose of this exercise was to develop, through consistency, a notation that is easier and more obvious to use than the classical notations. So let’s put it to the test with a few examples.

But first a word about solving equations.

If I say I bought two apples for 30 cents, almost without thinking you will know that each apple cost 15 cents. How? Well, calling the price of one apple a, we have

$$a \times 2 = 30$$

and we want to isolate a, which means we want to move everything to one side of the equals sign, except the a.

How do we do that? By applying the same operation to each side (so that the equality remains true), and then simplifying one or both sides.

In this case we divide both sides by 2:

$$(a \times 2) \div 2 = 30 \div 2.$$

Simplify, and we’re done:

$$a = 15.$$

But you should realise that there are some steps we have missed out, because they are ‘obvious’. Let me show you them, plus at each step the rule that is used:

`a×2 = 30`

`(a×2)÷2 = 30÷2` {divide both by 2}

`a×(2÷2) = 30÷2` {(a×b)÷c = a×(b÷c)}

`a×(1) = 30÷2` {a÷a = 1}

`a = 30÷2` {a×1 = a}

`a = 15` {calculate}

As you can see, doing all the steps, although straightforward, is also tedious. So in the following, when we are isolating a value, we will often leave out the obvious steps with the word ‘simplify’.

Interest

If you save an amount of money, `m`, at an interest rate of 3%, then it means that at the end of the year

`m×(3÷100)`

gets added to your account, so you then have

`m+m×0.03.`

Factoring out m, this is the same as

`m×(1+0.03)`,

which is

`m×1.03`.

In other words, each year your money gets multiplied by 1.03. So at the end of one year, you will have

`m×1.03`,

at the end of two years

`m×1.03×1.03`,

at the end of three years

`m×1.03×1.03×1.03`,

and at the end of n years,

`m×1.03↑n`.

So the general formula is:

`result = m×(1 + rate÷100)↑years.`

Let's simplify this slightly and give the expression `1 + rate÷100` the name '`r`', so that for 3% interest, `r` is 1.03. So our formula is

`result = m×r↑years`.

How much would you have to put in the bank at 3% to ensure that in five years you have 1000?

We have to isolate `m`. Take the equation,

```
result = m×r↑years
```

and divide both sides by r↑years:

```
result÷(r↑years) =
m×r↑years÷(r↑years)
```

Simplify the right-hand side:

```
result÷(r↑years) = m
```

So fill in values and calculate, and we get

```
1000÷1.03↑5 = 862.61
```

What interest rate would you have to have in order to double your money in ten years?

We have to isolate `r`. Take the initial equation

`result = m×r↑years`

and divide both sides by `m` and simplify:

`result÷m =` **`r↑years`**

We want to isolate `r`, so take a root of both sides:

`(result÷m)↓`**`years`** `= (r↑years)↓`**`years`**

replace `(r↑years)↓years` by `r`:

`(result÷m)↓years =` **`r`**

Since we want `result` to be twice `m`, and over ten years, we have to calculate:

`(2×m ÷ m)↓10`

which is

`2↓10=r`

which is approximately 1.07177 or in other words 7.177%.

How many years do you have to save at 3% to double your money?

We want to isolate `years`. Take the initial equation:

`result = m×r↑years`.

Divide both sides by `m` and simplify:

`result÷m= r↑years`

We want to isolate `years`, so take the log of both sides:

`(result÷m)⇓r = (r↑years)⇓r`

Simplify the right-hand side:

`(result÷m)⇓r = years`

`m`=1, `result`=2, r=`1.03`, which gives us:

`2⇓1.03 = years`

which gives us 23.45 years.

Computer Speeds

In 1965, Gordon Moore predicted that the density of components on integrated circuits was going to double every year at constant price, for at least ten years. Ten years later, he re-analysed the data, and increased the time to 18 months per doubling. Since then, his prophecy has held up fairly well, under the name “Moore’s Law” (even though it isn’t really a law, in any meaning of the word).

While he didn’t actually predict that computers would get twice as fast per 18 months, that has been pretty much the result.

If computers get twice as fast every 18 months, then we can represent the relative speeds by:

```
speed = 2↑(months÷18)
```

What is the annual growth?

```
2↑(12÷18)
```

which is 1.59, in other words a 59% annual growth. The monthly growth is `2↑(1÷18)`, which is 1.04, or in other words, 4%.

If only banks offered that sort of interest...

How many months before you can buy a computer that is 10 times faster?

We want to isolate `months`. Again, start with the speed equation:

speed = 2↑(months÷18)

Take the log of both sides

speed⇓**2** = (2↑(months÷18))⇓**2**

Simplify the right-hand side:

speed⇓2 = **months÷18**

Multiply both sides by 18:

speed⇓2**×18** = months

For `speed`=10 we have:

10⇓2×18 = months

which is just under 60 months, or 5 years.

For a speed gain of 100, we calculate

`100⇓2×18 = months`

which is just under 120 months, or 10 years.

But really, we didn’t need to calculate that because we already knew from the table of equivalences above that

`(a↑b)⇓c = (a⇓c)×b`

and since 100=10↑2, were were calculating

`(10↑2)⇓2`

which is the same as

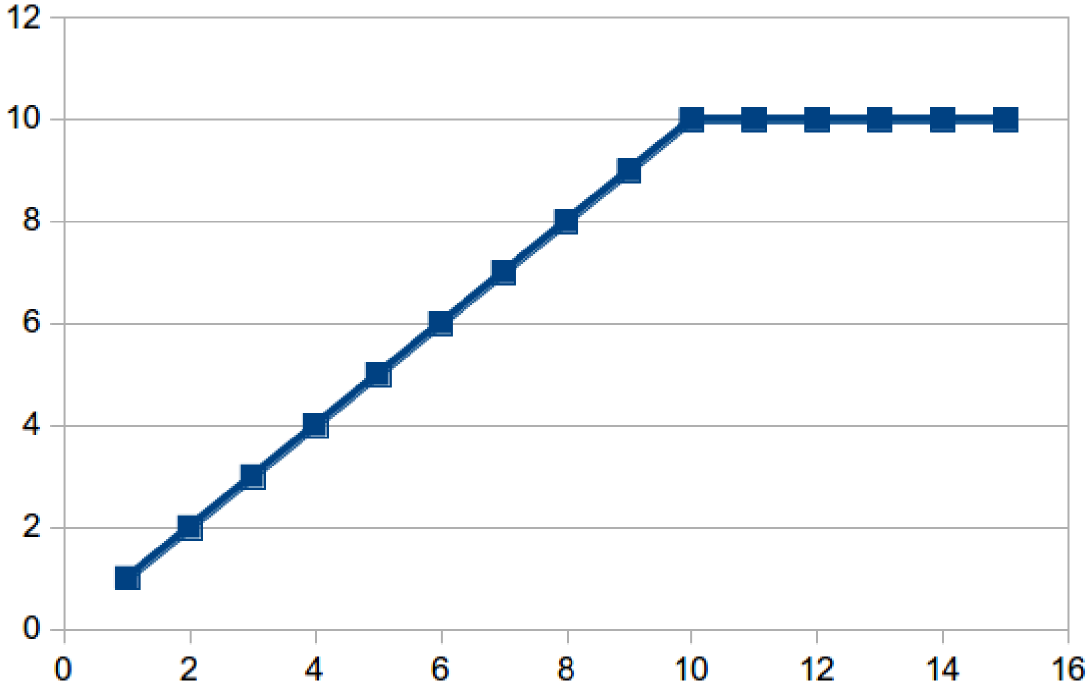
`(10⇓2)×2`

Since we had just calculated `10⇓2`, the answer had to be twice that answer.

Put another way: we knew that it takes 5 years to get a speed increase of 10; after another 5 years we would have got another speed increase of 10, and 10×10=100 times speed increase, and 5+5 = 10 years.

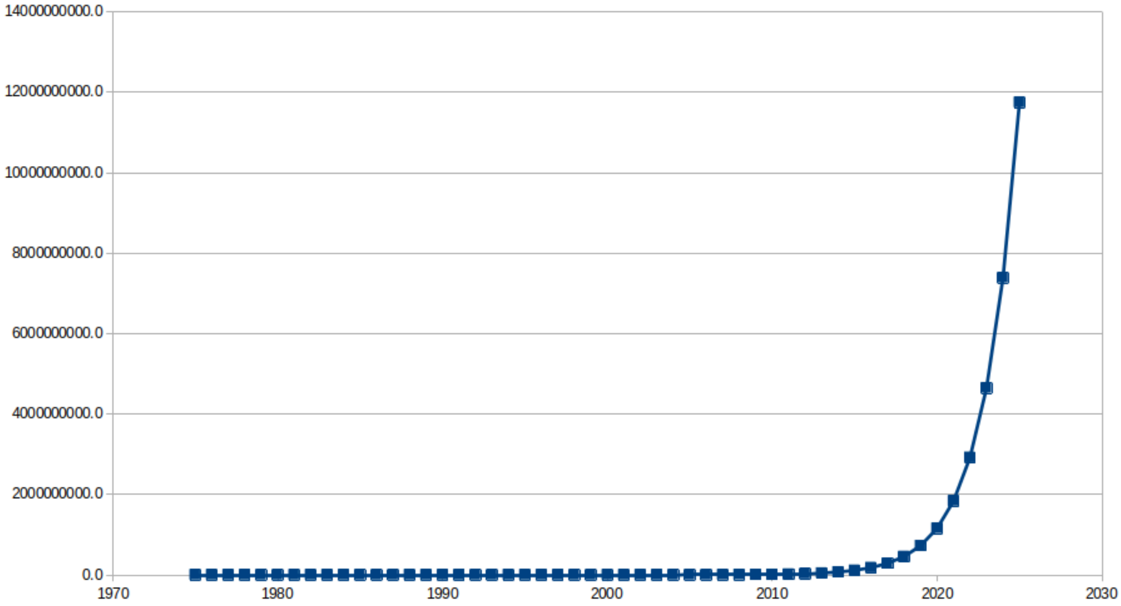
Graphing Exponentials

When you turn a tap on to fill a bath, you are *adding* a certain amount of water per minute to the bath. So if we look at the graph of the bath filling, we get something like this:



This is called a *linear* function.

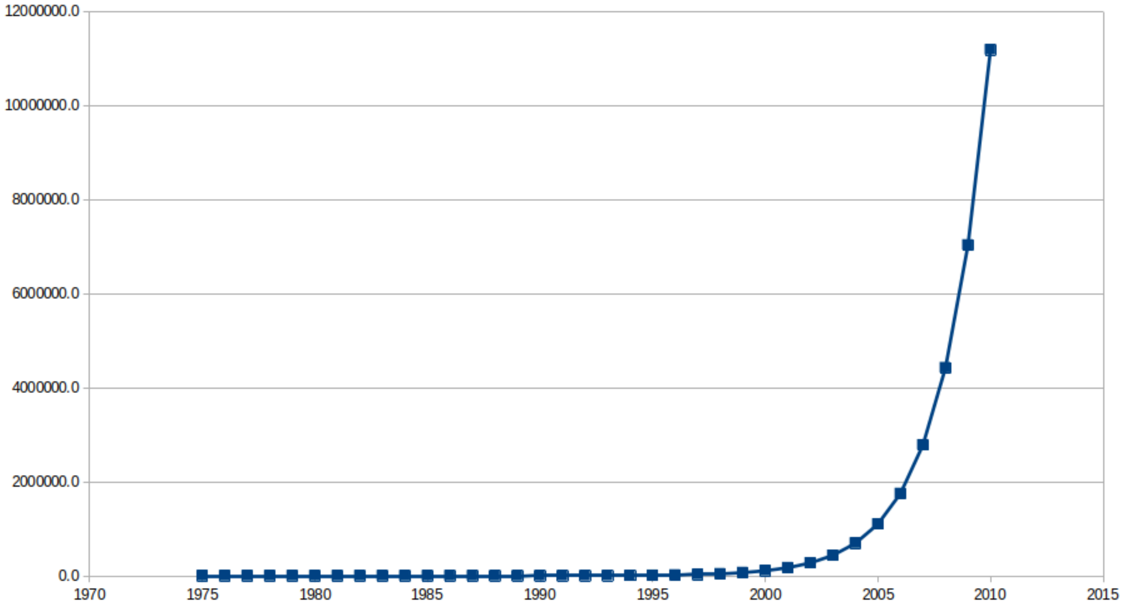
However, Moore's Law is a regular *multiplication* not addition: a doubling every 18 months. If we draw a graph of Moore's Law since 1975, it looks something like this:



In other words, a computer in 2025 would be nearly twelve thousand *million* times more powerful than one in 1975 (at the same price). This is an *exponential* function.

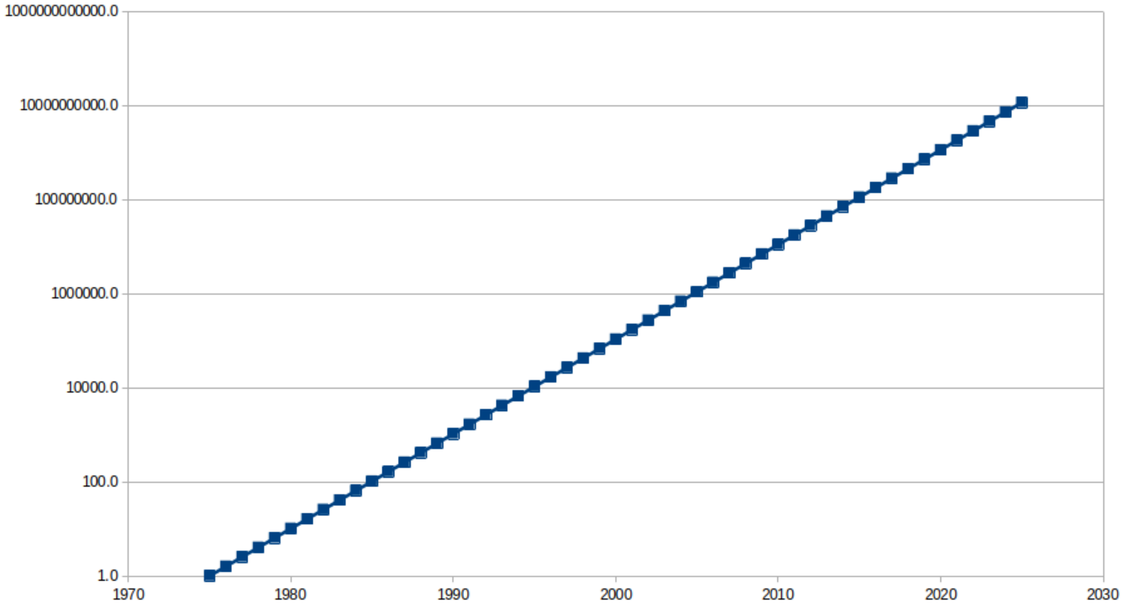
If you look at this graph, you'll see that it looks like almost nothing is happening until 2015, and then suddenly it shoots up like a rocket. Some people, when talking about exponential functions like this talk about it "having passed the knee", meaning it has finally taken off. But I have some bad news: there is no knee.

If I draw exactly the same graph, but only up to 2010, suddenly the knee appears to start around 1999. There is no knee: it is only a visual artifact of the scale you use.



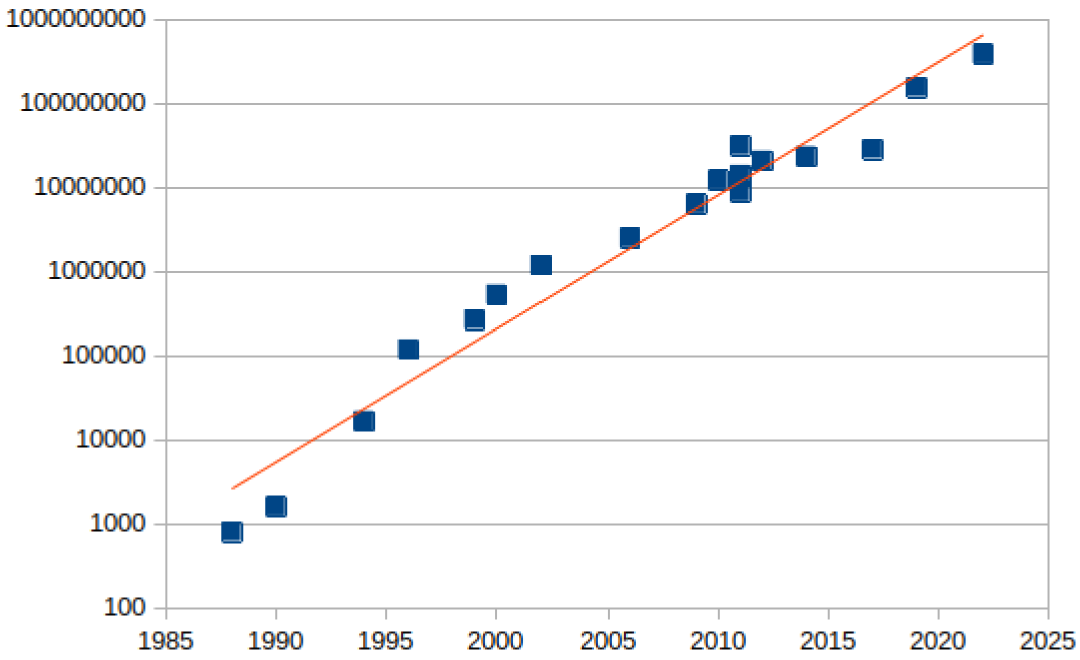
This is why it is better to graph exponential functions in a different way. On the vertical axis, rather than each step adding a certain amount, going in steps of 1, 2, 3, ... we multiply by a certain amount, for instance use steps of powers of ten 1, 10, 100, 1000, ...

Then the Moore's Law graph looks like this:



If you use a such a scale, and the graph looks like a line, then it is exponential (you can use powers of any positive number on the scale; it will still look the same).

Of course, computers don't get *exactly* twice as powerful in *exactly* 18 months. But I have been collecting data on the speed of my laptop computers since 1988, and you can see an exponential rise:



Although we are not exactly comparing like with like, in those 34 years, you see my computers going from power 800 to power 385 million. How many doublings is that? Well

```
800 × 2↑n = 385,000,000
```

so

```
2↑n = 385,000,000 ÷ 800
```

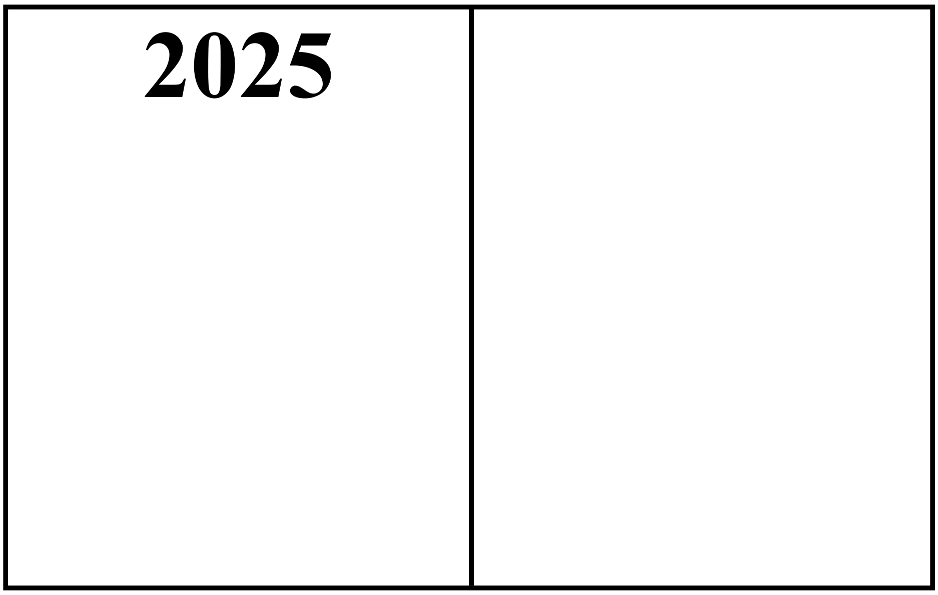
and thus

```
n = (385,000,000 ÷ 800)⇓2
```

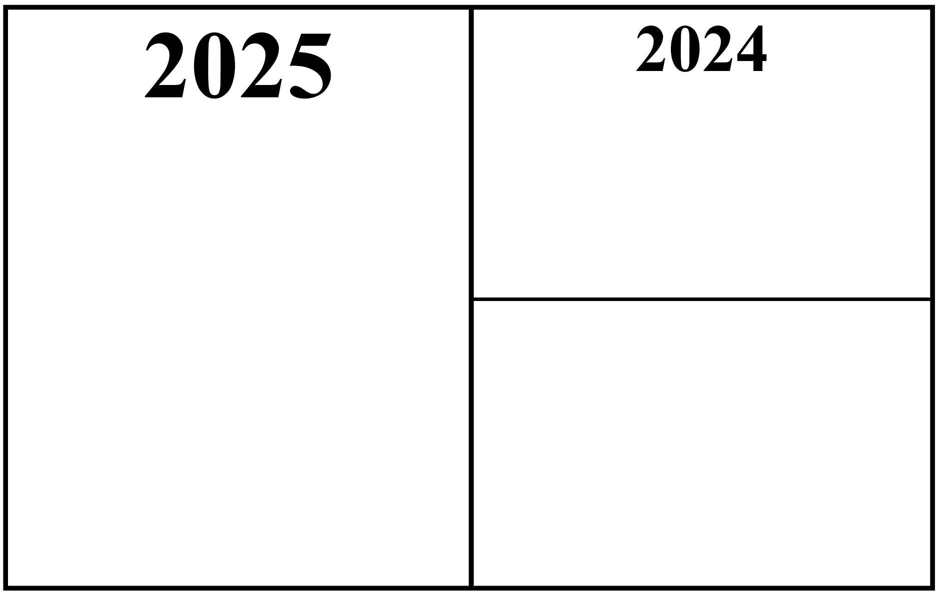
which is nearly 19 doublings in 34 years, or a doubling time of about 21½ months (34×12 ÷ 19).

The Effect of Moore's Law

OK, so we all know Moore's Law now. But often people don't understand its true effects. Take a piece of paper, divide it in two, and write this year's date in one half:

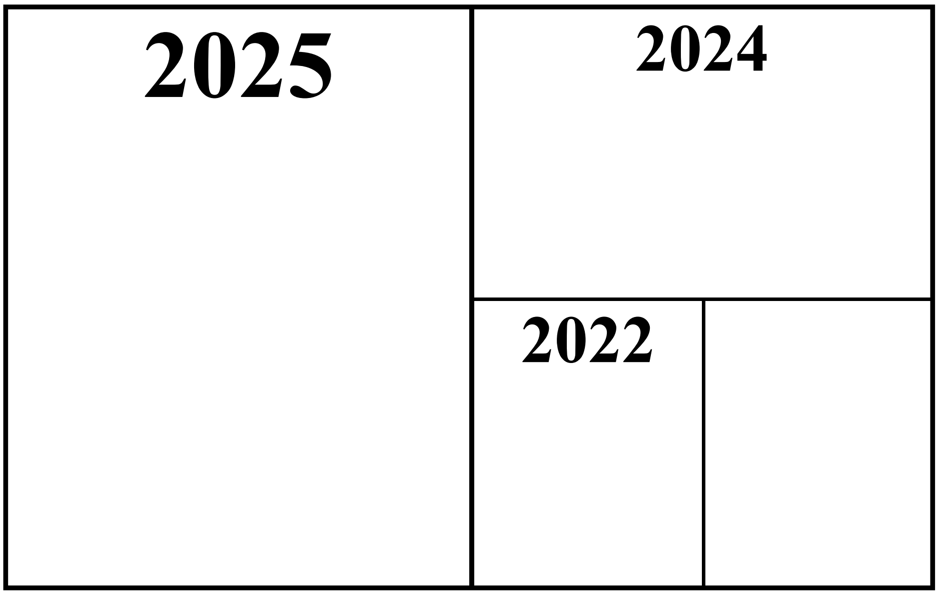


Now divide the other half in two vertically, and write the date 18 months ago in one half:

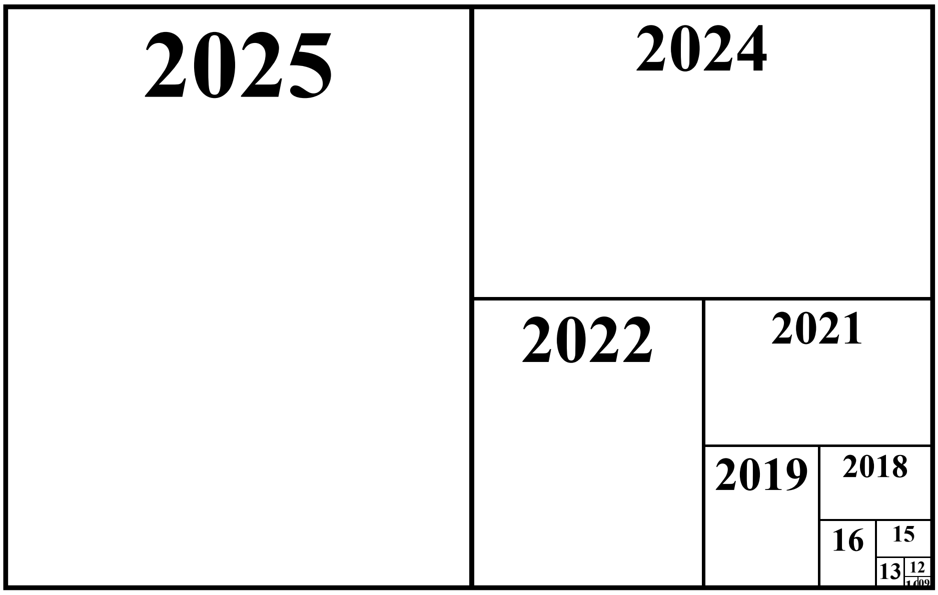


These two areas represent the speed of a computer today vs a computer bought 18 months ago at the same price.

Now divide the remaining space in half, and write the date 18 months earlier (or in other words 3 years ago):



That area represents the speed of a computer bought 3 years ago. And so on, repeat until your pen is thicker than the space you have to divide in two:



What this demonstrates is that your current computer is more powerful than all other computers you have had *put together*, assuming you don't buy a computer more often than every 18 months. (And this is true of all exponential functions by the way.)

Networks

The research centre where I work was the first internet connection in Europe on the open (non-military) internet. In 1988 the first connection from Europe to the United States was in the office next to mine, and all of Europe was connected to all of the USA at the blisteringly fast speed of 64k bits/second. (Nowadays a mobile phone is typically 1000 times faster than that). A year later the speed doubled to 128 k bits/second, and we rejoiced.

But in fact, even better than the speed increase of computers, network bandwidth doubles per year at constant price. As of this writing, where I work is the fastest internet node in the world, peaking at over 11.5T bits/second. But this is usage, not available bandwidth (which is much more). But let's calculate the annual growth in usage: in 1988 it was 64kb/s; 35 years later in 2023 it is 11.6Tb/s. What is the annual growth? Let's call it `g`:

```
64k × g↑35 = 11.6T
```

So

```
g↑35 = 11.6T ÷ 64k
```

and so

```
g = (11.6T ÷ 64k) ↓ 35
```

which is about 1.73, or a 73% increase per year averaged over the 35 years.

How many times has it doubled?

```
64k × 2↑n = 11.6T
2↑n = 11.6T ÷ 64k
n = (11.6T ÷ 64k) ⇓ 2
```

which is about 25.5 doublings. What is the doubling time?

```
35×12 ÷ 25.5
```

which is about 16.5 months.

Over 36 years, my home connection speed increased from 300 b/s to 200Mb/s, a 700,000 fold increase. So what has my annual bandwidth increase been?

```
g↑36 = 200M ÷ 300
g = (200M ÷ 300)↓36
g = 1.45
```

In other words, a 45% increase per year.

How many times has it doubled?

```
300 × 2↑n = 200M
2↑n = 200M ÷ 300
n = (200M ÷ 300) ⇓ 2
```

About 19.5 times. What is the doubling time?

```
36×12 ÷ 19.5
```

Which is about 22 months.

Assuming the same growth, how long before I could expect 1Gbps ? Well, 1G is 1024M:

```
1.45↑y = 1024M ÷ 200M
```

Do the division

```
1.45↑y = 5.12
```

We want to isolate y. Take the log of both sides:

```
(1.4↑y)⇓1.4 = 5.12⇓1.4
```

Simplify the left-hand side:

```
y = 5.12⇓1.4
```

Which is about 4.4 years.

Population Growth

Let’s assume that world population growth is exponential, which means that over some fixed period of `y` years it doubles. Let’s call the yearly growth `g`; then what this means is that

```
g↑y = 2
```

So, if we want to know what the yearly growth `g` is, we take the root of both sides:

```
(g↑y)↓y = 2↓y
```

and simplify:

```
g = 2↓y
```

which we can also write as `2↑÷y`.

So, if each year the growth is `2↑÷y`, then over a period of `n` years the growth is `2↑(n÷y)`. The world grew from 6 billion to 7 billion in the 13 years up to 2012, so we have

```
6×2↑(13÷y) = 7
```

What then is `y`, the doubling period? Divide both sides by 6:

```
2↑(13÷y) = 7÷6
```

Take the log of both sides:

```
(2↑(13÷y))⇓2 = (7÷6)⇓2
```

Simplify:

```
13÷y = (7÷6)⇓2
```

Divide both sides by 13:

```
÷y = ((7÷6)⇓2)÷13
```

Invert both sides:

`÷÷y = ÷(((7÷6)⇓2)÷13)`

Simplify

`y = 13÷(7÷6)⇓2`

Which you may also transform to this if you wish:

`y = 13×2⇓(7÷6)`

Which gives a doubling time of 58.45 years. So all other things being equal, we could expect a population of 14 billion in around 2070.

When might the population enter double figures?

`7×2↑(x÷y) = 10`

Divide by 7

`2↑(x÷y) = 10÷7`

Take the log:

`x÷y = (10÷7)⇓2`

Multiply by y

`x = y×((10÷7)⇓2)`

Which is about 30 years after 2012.

Inflation

The newspaper says that inflation last month was 10%. But by that they mean the yearly inflation. So how much did prices go up last month for them to come to that conclusion?

Well, what they are saying is that the monthly inflation rate applied 12 times gives a yearly rate of 10%, so what costs 100 now will cost 110 in a year:

```
    100 × (m↑12) = 110
```

Divide both sides by 100:

```
    m↑12 = 1.1
```

and take the root:

```
    m = 1.1↓12
```

which is about 1.008. In other words, last month's inflation was 0.8%

Prime numbers

At the time of writing, a new largest prime number has been discovered `(2↑136,279,841)-1`. A news article about it says that it is more than 16 million decimal digits longer than the previous largest prime. But how many decimal digits is the new number long?

Well, the length of the integer part of a number `10↑d` is the next integer higher than `d`. For instance:

`10↑3 = 1000`, and 1000 is length 4

`10↑4.5 = 31622.7766`, and 31622 is length 5.

So if we have `[10↑d]`, how do we get `d`? Well, by definition

`(10↑d)⇓10 = d`

so we have

`2↑136,279,841 = 10↑d`

and therefore

`(2↑136,279,841)⇓10 = (10↑d)⇓10`

`(2↑136,279,841)⇓10 = d`

However, we don't want to calculate `2↑136,279,841`, because it's a very big number. But if you look back at the table of equivalences, you'll see this one:

```
(a↑b)⇓c = (a⇓c)×b
```

in other words:

```
(2↑136,279,841)⇓10 =
(2⇓10)×136,279,841
```

and that's an easier number to calculate:

```
(2⇓10)×136,279,841 = 41024319.945
```

In other words, the new prime number is 41,024,320 decimal digits long.

Why There Is No Perfection

Imagine you want to buy a folding bike. What do you want from such a bike? Well, for instance, that it is light, strong, cheap – or at least reasonably priced, easy to fold, quick to fold, small when it is folded, comfortable to ride... there are probably a dozen such properties that you want a folding bike to fulfil.

Well, let’s just take three of them for now: light, strong, and cheap. Unfortunately, search as you may, you won’t be able to find a folding bike that matches all three: you can find strong and light, but not cheap; you can find cheap and strong, but not light; you can find cheap and light, but not strong. Just those three constraints are not satisfiable. You *could* find strong, cheap and light, but then it wouldn’t be a bike: being a bike and being able to fold it are two constraints that are non-negotiable – it *must* be a bike, and it *must* be foldable.

So in other words, you will have to relax at least one of your negotiable constraints. You could decide not to go on holiday next year and use that extra money to buy a non-cheap bike. Or you could decide to put up with a heavy bike, or you could decide to be careful, and go with a not-strong bike. In any case, you have to put up with non-perfection.

And this is why you shouldn't expect to find a perfect partner either. What might you look for in a partner? Good looking, healthy, fit, financially secure, amusing, good conversationalist, good natured, good at cooking, musical, good in bed, ... There are lots of constraints, and they will be different for different people.

Then there are probably a few that you wouldn't think of mentioning, maybe because they are non-negotiable, like: of the opposite sex/same sex depending on your preference, in an age group not too different from your own, speaks a language that you also speak, interested in you, and so on, and so on.

The problem is, as you add each constraint, the pool of potential perfect partners gets smaller and smaller, and then you have to add the constraint that it is someone you will somehow actually get to meet in a social context... This is why you should prepare yourself for not meeting the *perfect* partner: you will need to relax some of your constraints to find someone who is at least satisfactory.

Which brings us, funnily enough, to musical tuning.

You probably know that a note in music is caused by air vibrating, at a different frequency for different notes. For instance, internationally it has been agreed that the note A above middle C has a frequency of 440Hz. Not for any particularly good reason (and a slightly better choice would be 430.5Hz, but not so much better to make it worth retuning all the instruments in the world) but 440Hz is good enough.It turns out that our ears like combinations of frequencies that are related in certain simple ways, as small ratios of frequencies. For instance, a note played an octave higher is just twice the frequency, that is to say they have a ratio of 2:1. So the A one octave higher would be 880Hz. We hear a note an octave higher as almost the same note, only higher (and incidentally, it is also just by convention that we call it ‘higher’, a longer organ pipe plays a deeper note, so we could have called them the other way round).

The next-most simple ratio of frequencies is 3:2. This is the relationship between the notes called a “perfect fifth”, between the first note on the major scale, and the fifth (which are seven semitones apart). So for instance, a fifth from A is the note E, so E would have a frequency of 440 × 3÷2, which is 660Hz. And the combination of an A and an E sounds nice to the human ear.

Of course, we want to be able to play a perfect fifth from any note, not just A, so the perfect 5th from E, which is B, would have a frequency of 660 × 3÷2, which is 990 Hz. And so we can build up a whole octave based on the premise that from every note we can also play the perfect fifth.

Since 990 is above 880, we can halve the number to take it down into the octave we are building up, between 440Hz and 880Hz. So the B is at 495Hz. The fifth above that is F# which will have a frequency of 742.5, and so we can step through C#, G#, D#, A#, F, C, G, up to D, which has a frequency of 594.67Hz. The fifth above that takes us back to A, which will have a frequency of 594.67 × 3÷2, which is ... 892 Hz? But didn’t we say that A would be 880Hz?

Yes we did: in other words, it is *impossible* to make an octave of notes where you can play a perfect 5th from every other note. We have to relax one of our constraints.

Well, how about if we take the next most simple ratio, the perfect fourth, which has a ration of 4:3? Can we build an octave out of that, starting with A=440, and its perfect fourth D being 440×4÷3 = 586.7 and so on? I’m afraid that the answer to that is also no: you end up with an octave A that has a frequency of 868.1Hz, instead of the required 880Hz.

So what is to be done?

It is clear hopefully from the above that the octaves are non-negotiable. Any solution has to have the octave of any note as twice the frequency, otherwise you would get awful dissonance.

So what we can try to do is divide the octave up into 12 equal steps, equal in the sense that each semitone has the same frequency ratio with its neighbour. But what is that ratio?

Let’s call the ratio r.

Starting from A, the calculation 440×r should give us A#. Then 440×r×r should give us B, and 440×r×r×r should give us C, and so on all the way up to the next A:

```
440×r×r×r×r×r×r×r×r×r×r×r×r = 880
```

(that's twelve r's). Writing this another way:

```
440×r↑12 = 880
```

or

```
r↑12 = 880÷440
```

or

```
r↑12 = 2
```

or in other words

```
r = 2↓12
```

Well, we know how to calculate that: r is just under 1.06.

So if we calculate the resultant octave, it looks like this:

A	A#	B	C	C#	D	D#	E	F	F#	G	G#	A
440	466.2	493.9	523.3	554.4	587.3	622.3	659.3	698.5	740	784	830.6	880
					586.7		660					

Underneath E I have shown what we would ideally have for a perfect fifth from A, and under D, a perfect fourth. What you can see is that the difference is very small, less than 1Hz in both cases. So small in fact that since we *want* to hear the right tuning, we think it is properly tuned (a lack of dissonance due to cognitive dissonance).

So, the conclusion is, don't expect perfection, but if you relax some of your requirements you might just find something that so nearly matches that you can't tell the difference.

Notes

So, the note A above middle C has a frequency of 440Hz, and each of the 12 semitones in an octave are separated by a factor of 2↑÷12. So what is the frequency of middle C?

C is 9 semitones lower, so the frequency is

```
c = 440×2↑(-9÷12)
```

which we know we can also represent as

```
c = 440÷2↑(9÷12)
```

If we want, we can reduce it further to

```
c = 440÷2↑(3÷4)
```

but either way, it calculates to 261.626Hz

Which note is closest to 512Hz?

```
512 = 440×2↑(n÷12)
```

Divide by 440

```
512÷440 = 2↑(n÷12)
```

Take the log

```
(512÷440)⇓2 = (2↑(n÷12))⇓2
```

Simplify

```
(512÷440)⇓2 = n÷12
```

Multiply by 12

```
((512÷440)⇓2)×12 = n
```

Which is 2.62 (semitones higher than A). That means that the nearest note to 512Hz is C, which has a frequency of

```
440×2↑(3÷12)
```

which is 523.25Hz (unsurprisingly twice that of middle C).

If C were set at 512Hz, what frequency would A be?

```
512÷2↑(3÷12)
```

which can also be written:

```
512÷2↓(12÷3)
```

which is

```
512÷2↓4
```

or

```
512×2↓−4
```

any of which give 430.54Hz.

A frequency of 512 is 2↑9, and since octaves are constant doublings, you can now see why 430.54Hz could be seen as a good frequency for A rather than 440. It is said that Mozart used to tune his piano to A430.

Earthquakes

The Richter scale of magnitudes of earthquakes represents the amplitude (height) of the seismic waves. For a magnitude of m:

```
amplitude = 10↑m
```

The energy released on the other hand is the 1.5th power of the amplitude:

```
energy = (10↑m)↑1.5
```

which, because of the rule

```
(a↑b)↑c = a↑(b×c)
```

can also be written

```
energy = 10↑(m×1.5).
```

This grows really fast: an earthquake of magnitude 4 releases 1 million units of energy; magnitude 6 releases 1000 times that.

(Actually, the Richter scale itself is no longer used, but a new scale measured in a different way is used, but has been tuned to match the Richter scale as closely as possible.)

So what is the difference in magnitude between two earthquakes, where the second releases twice as much energy as the first?

$$e_1 = 10\uparrow(m_1 \times 1.5)$$

$$e_2 = 10\uparrow(m_2 \times 1.5)$$

$$e_2 \div e_1 = 2$$

Substitute in:

$$10\uparrow(m_2 \times 1.5) \div 10\uparrow(m_1 \times 1.5) = 2$$

Use the rule $a\uparrow b \div a\uparrow c = a\uparrow(b-c)$:

$$10\uparrow(m_2 \times 1.5 - m_1 \times 1.5) = 2$$

Take the 10 log of both sides:

$$(10\uparrow(m_2 \times 1.5 - m_1 \times 1.5))\Downarrow 10 = 2\Downarrow 10$$

Simplify:

$$m_2 \times 1.5 - m_1 \times 1.5 = 2\Downarrow 10$$

Factor out the 1.5:

$$(m_2 - m_1) \times 1.5 = 2\Downarrow 10$$

Divide both sides by 1.5:

$$m_2 - m_1 = 2\Downarrow 10 \div 1.5$$

Which is marginally above 0.2. In other words, an earthquake of magnitude 7.2 is about twice as powerful as an earthquake of magnitude 7.

Personally I was amazed by this. Until I worked it out, I assumed, as I am sure many people do, that magnitude 7 and 7.2 earthquakes were quite similar, when in fact they aren’t at all. The 7.2 is twice as destructive!

How could we fix this?

Well, a simple fix would be to introduce a new scale by just multiplying the magnitudes by 5. Then a magnitude 7.2 earthquake would become a scale 36 earthquake and a magnitude 7.4 earthquake would become a scale 37 earthquake. People could then understand better that the next number in the scale is twice as powerful, while still having a close enough relationship to the magnitude that it would be easy to convert from one to the other.

However, as I said above, twice as powerful is "marginally above 0.2", not *exactly* 0.2, and within a few points on the new scale the rule would diverge. For instance, the difference between a scale 36 earthquake and a scale 41 earthquake would actually not be 32 times as powerful as you might expect, but only 31.6. This is because the magnitude scale is based on powers of ten, and 31.6 is 1000↓2.

So what other possibilities are there?

Well, we could define a scale purely based on powers of 2. We could start at the same point, so that magnitude 0 was scale 0, and then *define* the new scale so that each point higher was *exactly* twice as powerful. What would that look like?

Well, as I said earlier, the magnitudes are the measure of amplitude, and the energy released is

```
e = 10↑(m×1.5)
```

The new scale would start at the same value for 0, which would be energy

```
10↑(0×1.5) = 1
```

and then double at each step:

```
e = 2↑s
```

So putting these together, we have

```
10↑(m×1.5) = 2↑s
```

Solving for m, we get:

```
m×1.5 = (2↑s)⇓10
m = ((2↑s)⇓10)÷1.5
```

and solving for s we get

```
10↑(m×1.5) = 2↑s
(10↑(m×1.5))⇓2 = s
```

Now tabulating these, we would get the following:

Magnitude	**Scale**
1	4.98
2	9.97
3	14.95
4	19.93
5	24.91
6	29.90
7	34.88
8	39.86
9	44.85
10	49.82

(The largest recorded earthquake ever was magnitude 9.5)

Scale	Magnitude
5	1.00
10	2.01
15	3.01
20	4.01
25	5.02
30	6.02
35	7.02
40	8.03
45	9.03

As you can see, the difference between this and the rule “multiply the old scale by 5” is very small, so maybe it’s better to just stick with that.

You might actually like to compare the two equations for m and s, and admire their symmetry:

```
m = ((2↑s)⇓10)÷1.5
s = (10↑(m×1.5))⇓2
```

Arithmetic

For hundreds of years schoolchildren have been tortured with the mathematics of powers, roots and logarithms as if they were a completely different branch of thinking. However, if we change the notation to match that of simple arithmetic, we can do no more than conclude that they are actually just one more part of simple arithmetic, with similar rules and productions.

To illustrate this point, the Wikipedia article on logarithms ends the section on bases with the statement

> *“Given a number x and its logarithm* $\log_b(x)$ *to an unknown base b, the base is given by:* $b = x^{1/\log_b(x)}$*”*

Translating this into our notation we have:

`b=x↑÷(x⇓b)`

which we know to transform to

`b=x↓(x⇓b)`

To which we reply “Ha! That’s the very definition of ⇓!”.

As another example, the Wikipedia article on roots says

> "*Simplifying radical expressions involving nested radicals can be quite difficult. It is not immediately obvious for instance that √(3+2√2) = 1 + 2√2*"

So let's prove it, and show how simple it now is! Convert to our notation:

`(3+2×2↓2)↓2 = 1 + 2↓2`

Square the right-hand side:

`= (1 + 2↓2)↑2`

Expand `(a+b)↑2` to `a↑2 + 2×a×b + b↑2`:

= **1↑2 + 2×2↓2+ (2↓2)↑2**

Replace 1↑2 with 1, and (2↓2)↑2 with 2:

= **1** + 2×2↓2+ **2**

Reorder:

= **3**+2×2↓2

Take the square root:

= (3+2×2↓2)↓**2**

QED.

Infectious Diseases

Infectious diseases are exponential as well. Different diseases are infectious at different rates, but the sneaky thing about such diseases is that a person having got the disease is infectious for a while before feeling ill: this is called the incubation period. That means that there are infectious people walking about who don't (yet) know it.

Suppose for instance there is a disease that has a very short 1 day incubation period, so you are infectious for one day without knowing it, and that in that one day while you are walking about your regular life that you manage to infect 2 others before you fall ill and retire to bed, and that those two people themselves similarly infect two people before falling ill, and so on.

Then within ten days, you have more than 1000 infected people (2↑10 = 1024). Within 20 days more than a million! (2↑20 = 1048576). Within a month, the whole country more or less is ill. What can be done to mitigate this?

If the disease is particularly serious, so that people need to be hospitalised, the main problem is that there is a limit to hospital space. There is some margin, but just wheeling more beds in is not sufficient, because training intensive care nurses takes several years.

You can't stop people getting infected (unless they are *completely* isolated), so you have to try and reduce the infection rate, so that the disease spreads more slowly. If you can reduce the infection rate, for instance reducing the doubling time to 2 days, it will take 20 days to get 1000 infected, and so on.

That is why during the Covid pandemic, we were asked to isolate, and wear masks. Not so we wouldn't get the disease, but to slow the rate of spreading, so that hospitals could at least deal with the numbers. If more are ill at any one time than the hospitals can cope with, some people will die just due to lack of care. If the numbers can be kept low enough, then at least the hospitals can cope, and the number of deaths will be minimised.

In other words, wearing a mask is not principally to help the person wearing it (though it does have a moderate effect in that way), but to reduce the amount of infectious material that the wearer can spread, which masks are very good at. Surgeons and dentists don't wear masks to protect themselves: they wear them to protect their patients.

As a real-life example, in mid January 2020, China had 278 recorded cases of Covid. A month and 4 days later, 77,000 cases were reported. By then, Italy had 215 cases, and a month and 6 days after that it had 101,000 cases. Let's call the infection rate r: r_c for China and r_i for Italy:

$$278 \times r_c{\uparrow}34 = 77000$$
$$215 \times r_i{\uparrow}35 = 101000$$

So

$$r_c{\uparrow}34 = 77000 \div 278$$
$$r_i{\uparrow}35 = 101000 \div 215$$

thus

$$r_c = (77000 \div 278){\downarrow}34$$
$$r_i = (101000 \div 215){\downarrow}35$$

Which gives as the average number of people one person infects per day as

$$r_c = 1.1799 \text{ for China}$$
$$r_i = 1.1921 \text{ for Italy}$$

Both fairly similar.

What is the doubling time (the number of days on average it takes for one person to infect 2 people)?:

$$1.1921{\uparrow}d = 2$$
$$d = 2 \Downarrow 1.1921$$
$$d = 3.945$$

in other words about 4 days (4.19 for China).

For the USA, the figures were 14 cases on February 12, 29 days later on March 12 1,645 cases, giving an r of 1.1786, and 31 days later 556,044 cases, giving an r of 1.2066.

So as you can see, a disease doesn't have to be very infectious to quickly spread over huge numbers of people, although, if you look back to the interest example, you'll recognise that 1.18 represents an 18% interest rate, *per day!*

We can also work out when the first infection in China could have happened, since

```
1 × 1.1799↑d = 278
d = 278⇓1.1799
d = 34.02
```

So this would suggest the first infection occurred a month and 4 days earlier, sometime in December.

In the Netherlands in late June 2021, the number of infected people had dropped to 500, and so the government decided to remove all restrictions. Was that a good idea? Well, 14 days later, there were 10,000 infections, so the government had to reintroduce the restrictions. What was the infection rate?

```
500 × r↑14 = 10000
r↑14 = 10000÷500
r↑14 = 20
r = 20↓14
r = 1.24
```

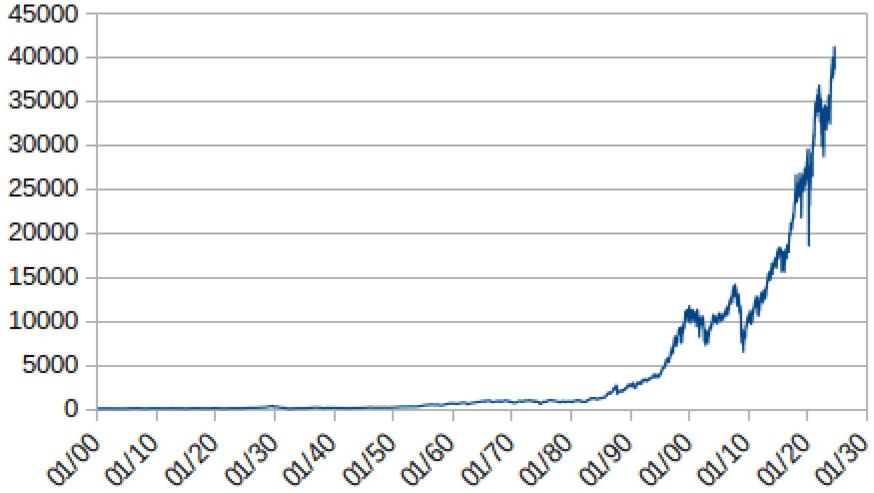
so we can conclude that people went wild, and mingled a lot in those 2 weeks (I can report that many of my teenage sons' friends went clubbing in those two weeks, and about half of them got sick).

What had the doubling period become?

```
1.24↑d = 2
d = 2⇓1.24
d = 3.22 days
```

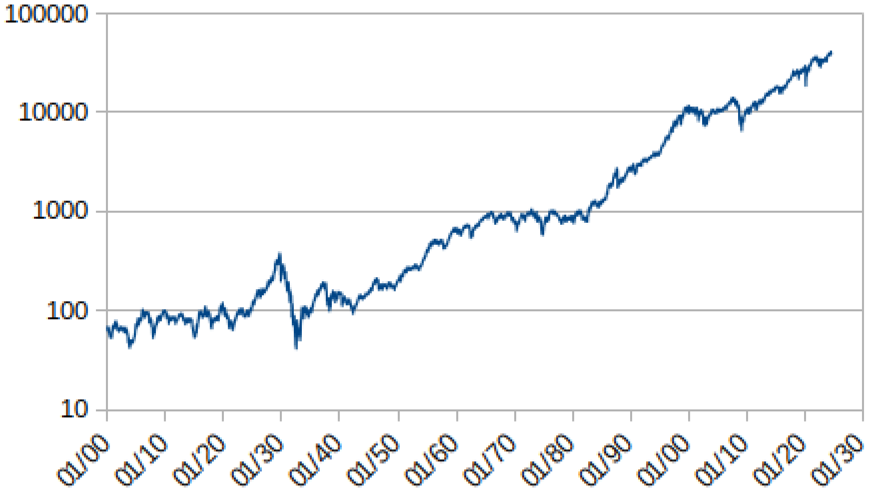
The Stock Market

If you look at a graph of a stock market (this is Dow Jones since 1900), it looks suspiciously exponential:



It shouldn't really be surprising that stock markets are exponential, since the aim is to increase their current value, not the principal value, by some percentage each year.

Plotted above on a linear scale, you can see the 2008 crisis clearly, but because of the linear scale, you can't see any detail before 1980. However, plotted on a logarithmic scale, the approximation to a straight line becomes clearer, and moreover, the detail from before 1980 becomes visible:



What this also shows is that while the crisis of 2008 was bad, and looks really bad on the linear scale (the index fell from 14093 in October 2007 to 6547 in March 2009, so more than halving in value), the great depression of the 1930s was much much worse, dropping from 381 in September 1929 to 41 in July 1932, losing 90% of its value, and not recovering until the mid 1950s.

Doing the calculations, the market has had on average about a 5.25% yearly increase since 1900, a doubling period of about 13½ years, or since 1950 a 24% yearly rise, a doubling period of about 3 years and 2 months.

So if you had invested 100 in 1950, and just left it, by 2025 it would have been worth

```
100 × 1.24↑75
```

which is a little over 1000 million!

How long ago would you have had to invest 100 to have a million now?

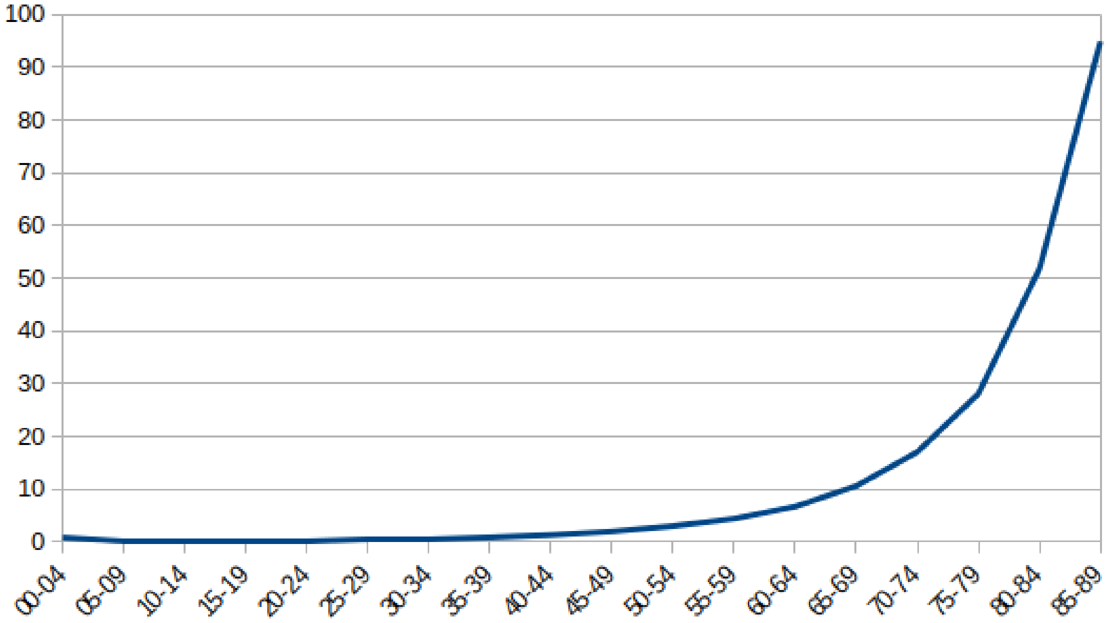
```
100 × 1.24↑y = 1000000
1.24↑y = 1000000÷100
1.24↑y = 10000
(1.24↑y)⇓1.24 = 10000⇓1.24
y = 10000⇓1.24
```

which is about 43 years ago.

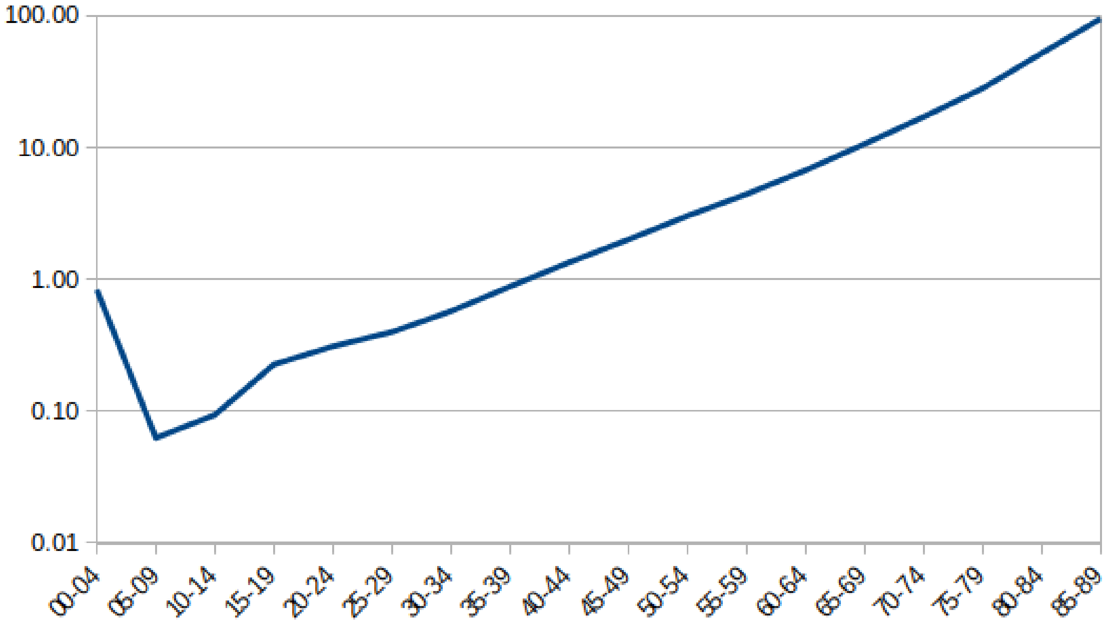
Mortality

It is obvious to say that the older you get, the more likely you are to die. But how much more likely?

We can work it out by taking the population of a country (in this case the UK in 2023), and splitting it up into 5 year age groups (0-4, 5-9, 10-14, and so on), and then divide the number of people who have died in each age group that year by the size of that group. This will then tell us the risk of dying in that age band, averaged over the population. Of course, this hides a lot of detail: smokers are more likely to die early (by about 14 years), males die earlier than women (by about 10 years on average), but it is still interesting information. Here is the (linear) graph of the risk of death in each age band under age 90, in deaths per thousand:



Again, this looks suspiciously exponential, so lets graph it logarithmically:



When I first saw this, I was amazed. What is revealed is that the linear version hides some important detail at the younger ages: the first few years are relatively quite dangerous (this is almost entirely due to newborn babies dying); the least risk is at around age 10, and after that the risk grows more or less exponentially: being about 0.03% at age 20, 0.06% at 30, 0.13% at 40, 0.3% at 50, 0.6% at 60, 1.7% at 70, 5% at 80, and 9.5% at 90 (in other words around 1 in 10 people aged 85-89 will die in a year).

Mortality is exponential!

So what is the yearly increase in risk? Taking the values from age 20 to age 89, we have an age range of 69 years:

```
0.031 × r↑69 = 9.496
r↑69 = 9.496÷0.031
r = (9.496÷0.031)↓69
r = 1.087
```

So the risk increases annually by about 9%

What is the doubling time?

```
1 × r↑y = 2
y = 2⇓r
y = 8.355
```

in other words, every 8 years and 4 months your risk of dying doubles, throughout your adult life.

Do You Really Accept Negative Numbers?

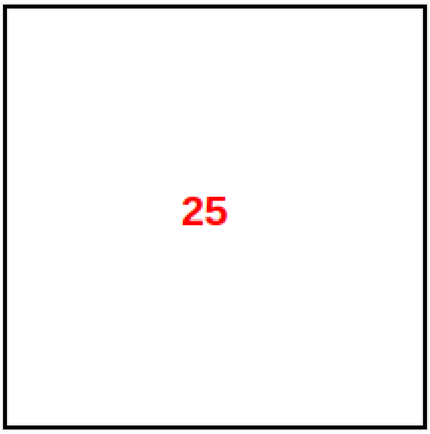
I'll say it again: all numbers are abstractions. (There, I've said it three times, it must be true.)

However, it has taken human society a long time to realise this. For instance, for hundreds of years even mathematicians didn't accept negative numbers, because they didn't seem to correspond to anything in "the real world". They were 'fictitious' or 'false' values, and solutions that gave negative numbers were ignored. You could count three sheep, but what could minus three sheep possibly mean? And addition was about making the result larger: 'plus' means 'increase', so how could `a+b` possibly be smaller than `a`, which would be the case with negative `b`? Pythagoras had no negative numbers for instance. Only recently has my bank been reporting debits as negative numbers, rather than positive amounts of debit.

Of course, we moderns accept negative numbers as full partners of the positive numbers, don't we? We accept that a car travelling backwards can be regarded as travelling at a negative velocity. We accept that you can have a negative balance in your bank account. We know how to deal with the idea of negative temperatures.

So, now that that is out of the way, let me ask you a question.

A farmer has a square field, of area 25. How long are the sides of the field?



Well, I expect you answered 5, which is the right answer.

Well, at least, one of the right answers, because −5 would be a correct answer too.

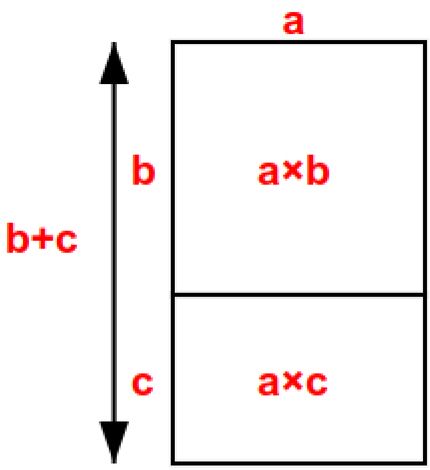
“But how can a field have a side of length −5?” I hear you ask. Now you know how early mathematicians felt about negative numbers.

So let me explain how a field can have a negative side.

Algebra and geometry often go hand in hand. We defined multiplication using

```
a×(b+c)  = a×b + a×c
```

and there is a really good way to illustrate this equality, that shows the area of the whole rectangle is made up of the sum of the two small rectangles:



$a\times(b+c) = a\times b + a\times c$

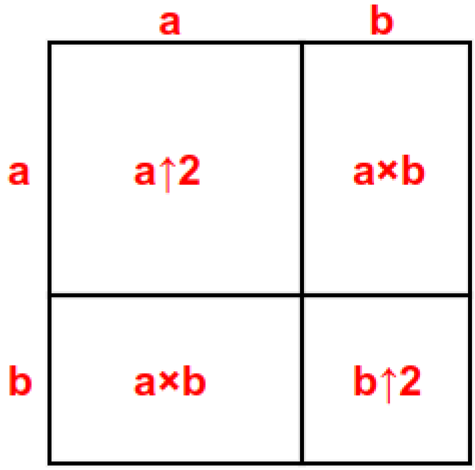
Another well-known equality (which can be derived from the one above) is

```
    (a+b)↑2 = a↑2 + 2×a×b + b↑2
```

Let’s derive it:

```
    (a+b)×(a+b)
   = a×(a+b) + b×(a+b)
   = a×a + a×b + b×a + b×b
   = a↑2 + a×b + a×b + b↑2
   = a↑2 + 2×a×b + b↑2
```

which likewise can be illustrated geometrically:



(a+b)↑2 = a↑2 + 2×a×b + b↑2

But an important aspect of such an equality is that *a*, *b*, and *c* can be *any* number. So while you can have *a*=3 and *b*=2:

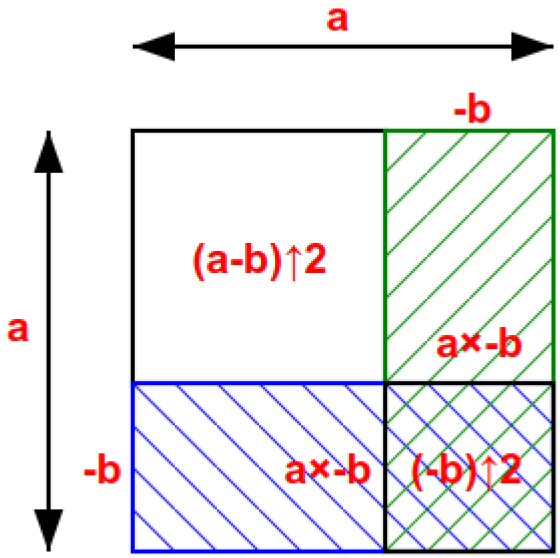
```
(3+2)↑2
= 3↑2 + 2×3×2 + 2↑2
= 9 + 12 + 4
= 25
```

you can also have *a*=3 and *b*=−2:

```
(3+(−2))↑2
= 3↑2 + 2×3×(−2) + (−2)↑2
= 9 + (−12) + 4
= 1
```

and the equality still holds.

But how does this version look in our geometric equivalent?

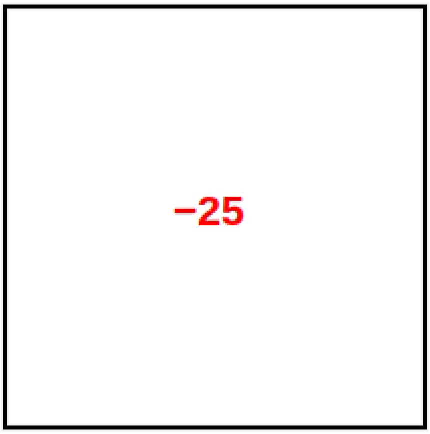


So the little square at the bottom right has sides of negative length −b, but a positive area (can you see why?) Even more interesting is that the two (overlapping) rectangles have negative area (a × −b).

So you can not only have shapes that have sides of negative length, but also shapes of negative area.

So let me ask another question to mirror the question asked at the beginning of this section.

Since we have now shown that it is possible to have shapes with negative area: what is the length of the sides of a square with area −25?

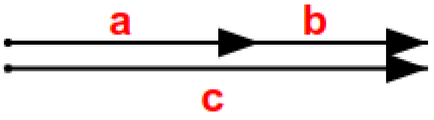


This is a question that we will come back to later.

Complete Numbers

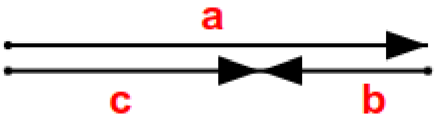
Even though society has for hundreds of years had problems with the concept of negative numbers, nowadays, we are happy to talk about negative temperatures, about negative growth, negative bank balances, and we accept that a negative speed indicates that the thing is travelling backwards.

One way of visualising addition of two numbers is to draw from zero a line of length the value of the first number, and then a line from the head of that line, of the length of the second number. The result is the length of the line from zero to the head of the second number:



a+b=c

If any number is negative, then the arrow goes in the other direction:



a+b=c with b negative.

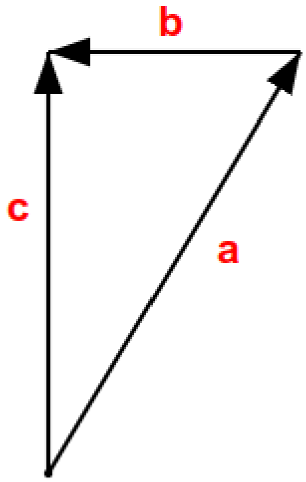
An observable application of this is someone swimming in a river. If someone is swimming at a fast 4 m/s in the same direction as the current of 3 m/s, then from the viewpoint of someone standing on the bank, they will be travelling at 7 m/s. Similarly, if they are swimming *against* the current, then from the viewpoint of the person on the bank, they will be travelling at 1 m/s.

Swimming against the current

Generalising Numbers

But now we are going to free these numbers from the bounds of the number line, and let them point in *any* direction. A number is now its length *and* its direction. For reasons I will later explain, I will call these new, unrestricted, numbers *complete* numbers.

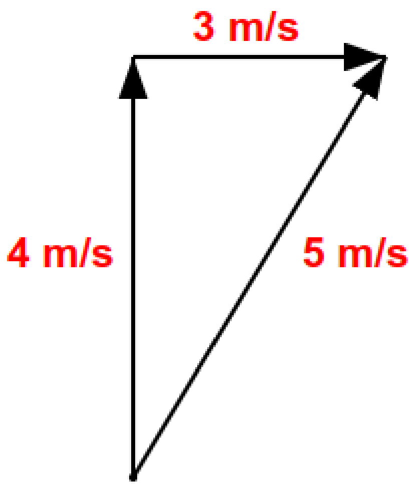
Addition is done in exactly the same way. You draw from zero a line of the length of the first number, pointing in the direction of the first number; from the head of that line you draw another line of the length of the second number, in the direction of the second number. Then the result is a line from zero to the head of the second number, and its direction is the resulting direction:



a+b=c with Complete Numbers.

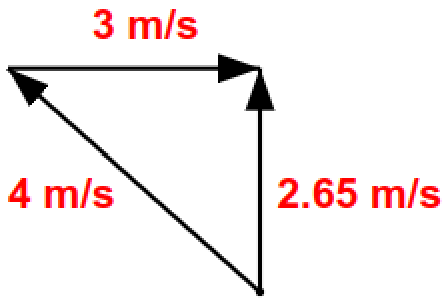
An immediate application of Complete Numbers is that you may now swim *across* the river, as well as up and down it.

So if you swim straight across the river at the speed of 4 m/s, and the river is flowing at 3 m/s, from the viewpoint of the person on the bank, you will end up swimming diagonally, at a speed of 5 m/s:



It's the same addition rule, just more general. (4@90° + 3@0° = 5@53.1°).

Similarly, if you want to swim across the river so that from the bank it looks like you are swimming straight across, you have to swim slightly into the current:

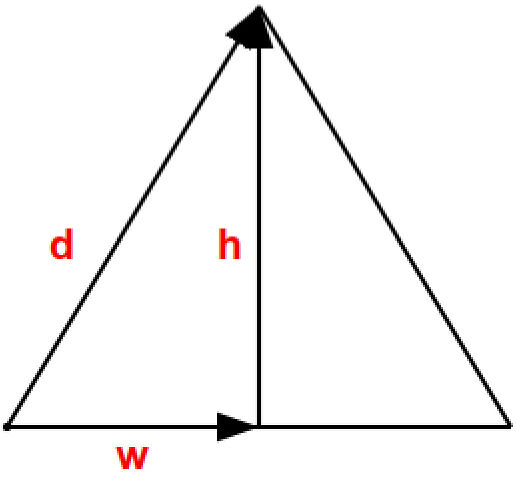


4@138.6°+3@0°=2.65@90°

Roof

But there are other applications of Complete addition.

If you know the height and width of a roof, you can easily calculate the length and angle of the roof covering:

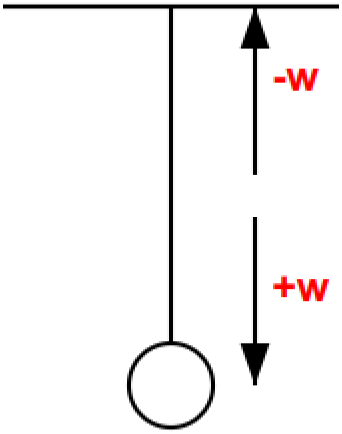


$$w+h = d$$

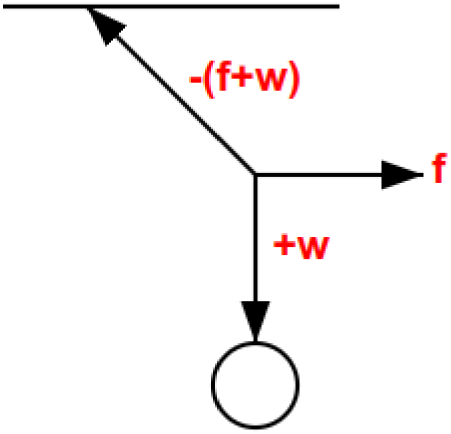
While we are looking at this example, it is worth mentioning that mathematicians would call the width a ‘real’ number, the height an ‘imaginary’ number, and the diagonal a ‘complex’ number (complex because it is both real and imaginary at the same time). I hope you can see why I think this is a misleading naming that should not be used, and all of them just be called ‘complete’ numbers.

Weights

If a weight hangs on a string from the ceiling, then the force in the string is just the same as the weight of the object, but in the other direction, so that they balance out:



However, if you attach another string to the middle of the first string, and pull on it, the force in the top half of the string is just the negative of the Complete sum of the two other forces:

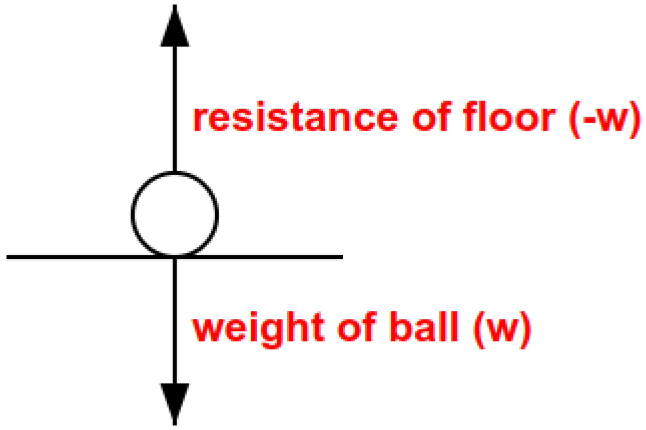


$$c = -(f+w)$$

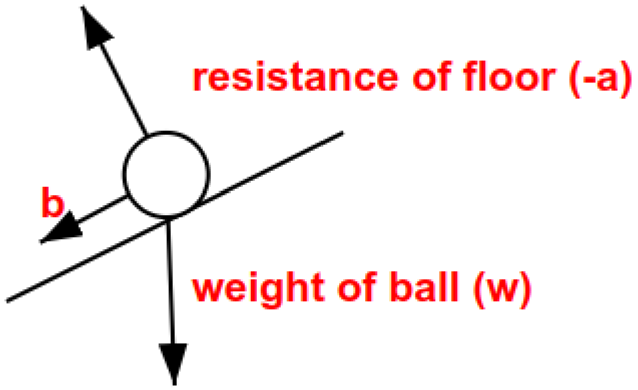
('Negative' for Complete Numbers means 'a half turn (180°) in the opposite direction'.)

Ball On Sloping Floor

If a ball is on a floor, then the floor offers the ball resistance exactly opposite to the force of the weight of the ball (otherwise the ball would sink, or crash, through the floor).



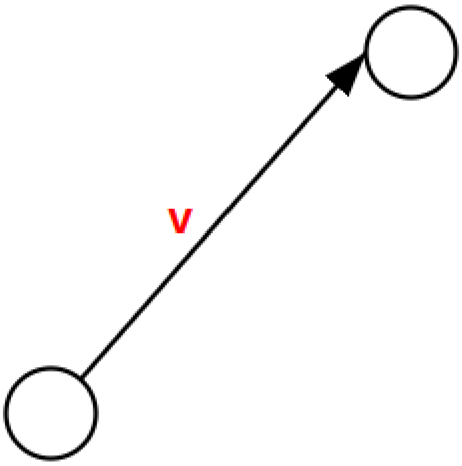
However, if the floor is sloping, the weight of the ball gets split into two forces, one which is at 90° to the floor, and one which is a force parallel to the floor, which pushes the ball down the slope. The complete sum of these two forces equals the force caused by the weight of the ball:



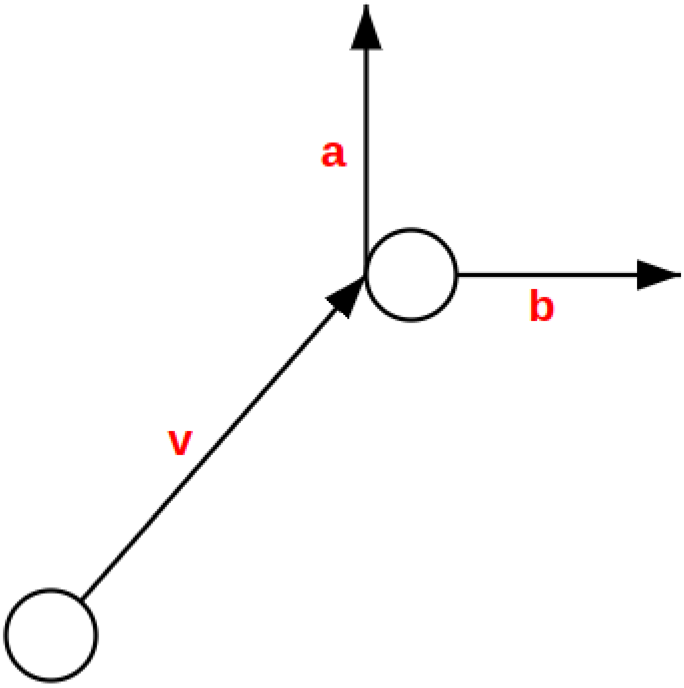
$$a+b=w$$

Billiard Balls

If a billiard ball hits another, stationary, ball of the same weight, then the Complete sum of the velocities of the two balls after the collision is the velocity of the original ball:



Before



After: $a+b=v$

In fact, even if the original ball misses the other ball, it's still true, because a+0 = a is still true with Complete Numbers.

Complete Multiplication

OK, so all those examples use addition, but what does it mean to multiply two Complete Numbers together?

Well, as was already explained, multiplication has to obey the rule:

$$(a+b)\times c = a\times c + b\times c$$

and while I shan't prove it here, the definition of multiplication that follows this rule is:

$$a@b \times c@d = (a\times c)@(b+d)$$

in other words, to multiply two Complete Numbers, you multiply the magnitudes, and add the angles.

A corollary of this is:

$$(a@b)\uparrow 2 = (a\uparrow 2)@(b\times 2)$$

and therefore

$$(a@b)\downarrow 2 = (a\downarrow 2)@(b\div 2)$$

and in general:

$$(a@b)\uparrow c = (a\uparrow c)@(b\times c)$$

and

$$(a@b)\downarrow c = (a\downarrow c)@(b\div c)$$

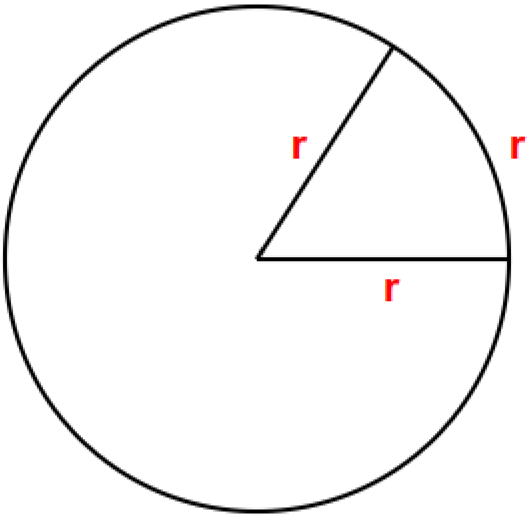
Angles

In the above examples, the angles of the Complete Numbers have been expressed in degrees, since that is what people are most at home with.

Why do we use 360° as the number of degrees in a circle? Possibly because it's close to the number of days in the year; in any case it is a nice number to split up into pieces, because it is divisible by 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 18, and 20. Only 7, 11, 13, 14, 16, 17, and 19 are missing in that row.

However, rather than degrees, mathematicians prefer to use *radians* for measuring angles (because it makes some laws simpler).

One radian is the angle such that the length of the arc drawn out by the angle is the same as the radius doing the drawing out; a sort of equilateral triangle, but with one curved side.



An equilateral triangle has three angles of 60°, and since one side of the above segment of the circle is curved, a radian is going to be slightly less than 60°. To be exact, since the circumference of a circle is $2\times\pi\times r$, there are $2\times\pi$ radians in a circle (that is $360° = 2\times\pi$ radians). Which means that one radian is equal to $360 \div (2\times\pi) = 57.3°$ (approximately).

However, for Complete Numbers it is actually handier to talk in fractions of a turn τ, where 1τ is one turn, which is the same as 360° or 2π radians. The advantage is that it is easier to take the modulo, and more understandable to talk in terms of fractions of a turn.

For instance the square of

```
2@270°
```

is

```
(2×2)@(270°+270°)
= 4@540° = 4@(540°-360°)
= 4@180°.
```

However, doing this in *turns*, we have

```
2@0.75τ × 2@0.75τ
= (2×2)@(0.75τ+0.75τ)
= 4@1.5τ
= 4@0.5τ,
```

since one and a half turns has the same effect as a half turn.

There is also the pleasant fact that if your take τ = 2×π, you can still have your radians. Then the circumference of a circle is `τ×r`.

What is τ?

Just about everyone know what π (pi) is. If you measure the diameter of a circle, you will find that the circumference of the circle is π times longer. This is true whether the circle is the size of an orange, or the size of a planet.

If a wheel is 1 metre in diameter, then its circumference is π metres.

Almost everybody knows the approximate value of π too: somewhere near 3.14 (a little bit more, 3.141592...).

How can you use it?

Well, suppose you have a bicycle. A typical bike wheel is 68cm in diameter. That means that each time it revolves, you travel 68×π cm, about 214 cm.

When you turn your pedals, it drives that big cog wheel at the front. That cog is connected to the chain that connects to a similar, smaller, cog at the back. If the big cog has 52 teeth, then each complete turn of the pedals will push 52 links of the chain forward. If the cog at the back has 26 teeth, then the 52 links in the chain will drive the back cog (and the back wheel with it) 52 ÷ 26 = 2 times round, which will cause the bike to travel 2×68× π cm, which is 4.27 metres.

Presumably the diameter was chosen as the basis for π because it is easier to measure the diameter of a cylinder or wheel than the radius.

It wasn't until much later that mathematicians realised that it is better to use the radius than the diameter when talking about circles (again, it makes some laws simpler).

Unfortunately, when they made the change, they forgot to change π to go with it, which means that mathematics is now full of formulas that include 2×π in them, such as the circumference (2×π×r).

But if we instead of using π, use τ (tau), with the value twice that of pi (in other words 6.283184…), then the circumference of a circle becomes τ×r, and the circumference of a quarter of a circle (90°) becomes 0.25×τ×r, and there are now τ radians in a whole circle, so that 180°, a half circle, becomes an angle of 0.5×τ (or τ÷2).

What About the Square Root of Minus Twenty Five?

You may know that complex numbers (as Complete Numbers are traditionally called in mathematics) were born out of a need to take the square root of negative numbers.

So to go back to the question that was asked earlier, what is the length of the sides of a field of area −25. Or to put it another way, what is the square root of −25?

Well, as we have now seen, a number like −25 is just another way of writing 25@180°, or rather, 25@0.5τ.

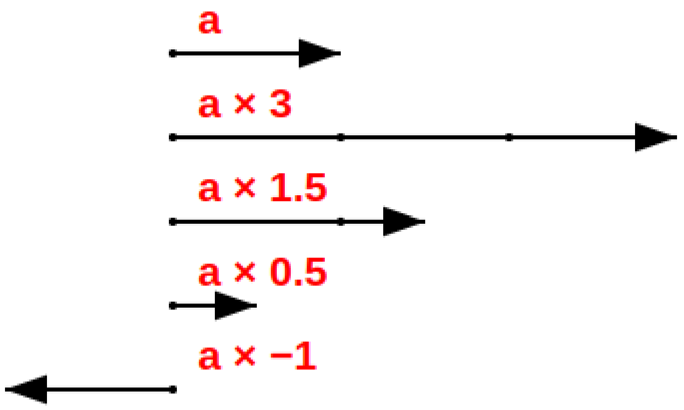
And, we know how to take the root of a Complete Number:

```
 (25@0.5τ)↓2
= 25↓2@(0.5τ÷2)
= 5@0.25τ
```

Nothing particularly special about that.

Multiplication is Rotation

Traditionally we know that multiplication is about making a number larger or smaller:



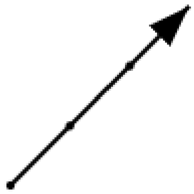
With Complete Numbers, multiplication also involves scaling like this, but with an added dimension of rotation.

But first things first. 'Regular' numbers, like `3`, are just shorthand for `3@0`. What this means is that multiplication works as normal with them: `a × 3@0` works the same, since the angle is zero:

```
a@0 × 3@0
= (a×3)@(0+0)
= (a×3)@0
```

Even if the number `a` has an angle other than zero, `a` just enlarges or contracts in the same way, at the same angle:

```
a@n × 3@0
= (a×3)@(n + 0)
= (a×3)@n
```



A number like `-3` is just shorthand for `3@0.5τ`, or if you prefer `3@÷2τ`, in other words, `3@0` flipped in the other direction. (Remember that `÷2` is just a new way of writing ½, and a half turn is 180°)

and so `3 × -1` is just

> `3@0 × 1@÷2τ`
> `= (3 × 1)@(0 + ÷2)τ`
> `= 3@÷2τ` (i.e. −3)

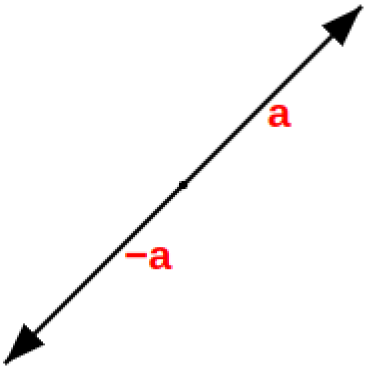
and similarly

> `-3 × -1`
> `= 3@÷2τ × 1@÷2τ`
> `= (3 × 1)@(÷2 + ÷2)τ`
> `= 3@1τ` (since a half plus a half is 1)
> `= 3@0τ` (since one whole turn is the same as no turn).

So we still have "two minuses make a plus".

By the way the operator "−" flips the number round the other way (i.e. rotates it a half turn) whichever way it is pointing. So

`−(3@÷8τ)`
= `3@(÷8 + ÷2)τ`
= `3@(5÷8)τ`

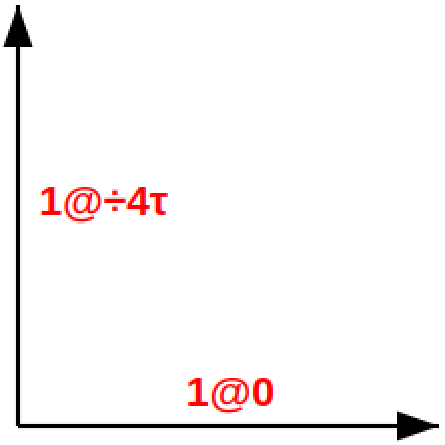


so that `a + −a` continues to be zero whichever way `a` is pointing.

Rotation

The place where most people have experience with rotation of lines is with clocks, where the lines rotate clockwise (well, there's a surprise), and the zero position is pointing upwards. However, in mathematics it's different: lines rotate anti-clockwise, and the zero position is at what you might call the three o'clock position.

This means that if we want to represent a minute hand at the twelve o'clock position, we should use `1@÷4τ`, a line of length one, rotated a quarter of a turn.

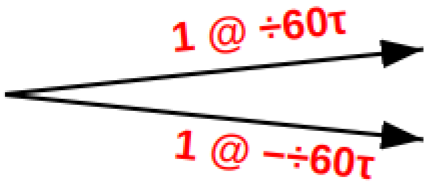


Every minute, the minute hand rotates clockwise a sixtieth of a turn (that is to say 6 degrees). Since multiplication is rotation, that is the same as multiplying it by `1@−÷60τ`

```
m ← m × 1@−÷60τ
```

But on the other hand, division is rotation in the clockwise direction, so we can also say

```
m ← m ÷ 1@÷60τ
```



The hour hand is on the other hand shorter, let's say three-quarters of the length: `0.75@÷4τ`.

The hour hand rotates much slower: every hour it moves one twelfth of a turn, which means that each minute it moves one sixtieth of one twelfth of a turn (a half degree in other words):

```
h ← h ÷ 1@÷(60×12)τ
```

or alternatively

```
h ← h ÷ (1@÷12τ)↑1@(÷60τ)
```

or alternatively

```
h ← h ÷ (1@÷12τ)↓1@60τ
```

This is because multiplying two Complete Numbers you add the angles, so to multiply the angles, you raise to the power. To summarise:

```
a@b × c@d = (a×c)@(b+d)
a@b ↑ c@d = (a↑c)@(b×d)
a@b ÷ c@d = (a÷c)@(b−d)
a@b ↓ c@d = (a↓c)@(b÷d)
```

Understanding the Definition of Multiplication

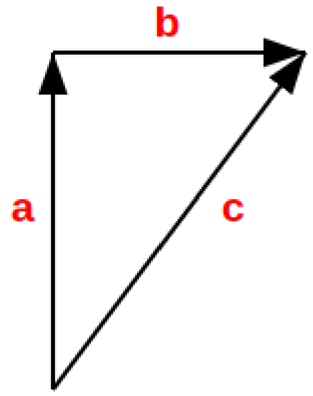
As was said before, the definition of multiplication is

`(a+b) × m = a×m + b×m`

and we said that the definition of multiplication for Complete Numbers that satisfies that definition is

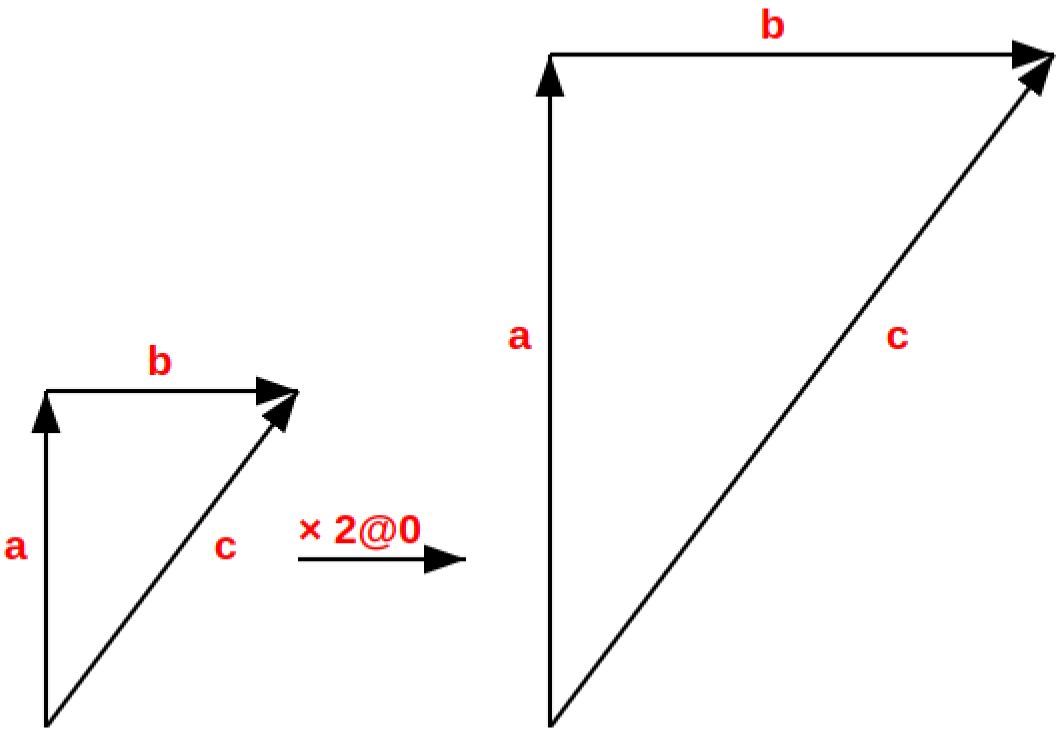
`a@b × c@d = (a×c)@(b+d)`

To give you a gut-feeling of why this works, let's take a diagram that we've seen before, showing `a+b=c`:



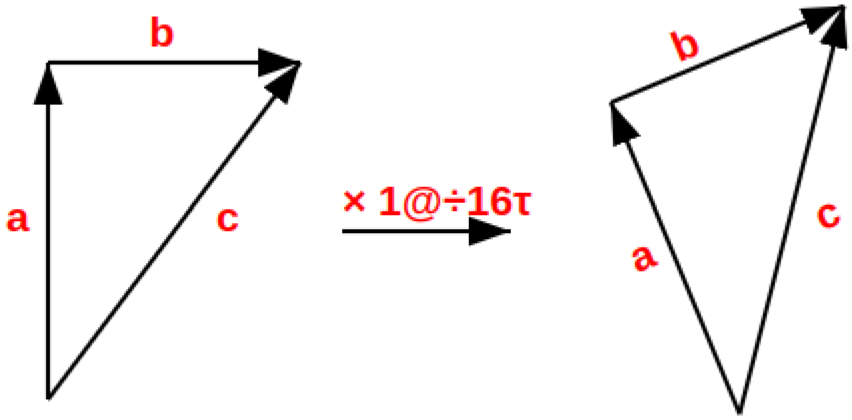
What we have to illustrate is that multiplying `a` and `b` by some number and adding them gives the same result as multiplying `c` by the same number (since that's what the definition of multiplication is saying).

So let's take a simple case first, multiplying by 2, or rather, by `2@0`:



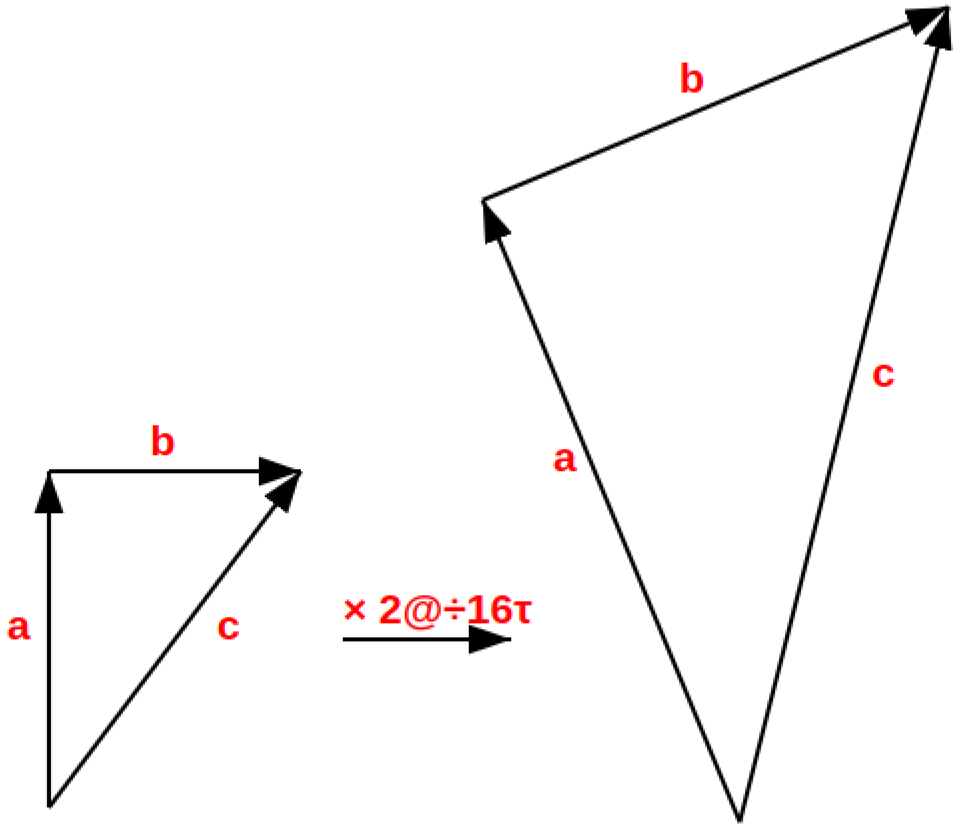
As you can see, enlarging `a` and `b` by 2 has the result of just enlarging `c` by 2, with all the angles the same.

Now let's take another easy case, multiplying by `1@÷16τ`, that is with rotation but no change in size:



So rotating `a` and `b` by the same amount causes `c` to also rotate by the same amount.

And now to combine the two, multiplying by `2@÷16τ`, in other words both enlarging and rotating:



This is not strictly speaking a proof, but I hope that it is convincing enough to help you understand why the definition of multiplication for Complete Numbers works as required, and if you are so inclined, to help you use it as the basis for creating a proof.

Conclusion

This book arose from the combination of three different thought-streams: the first was while I was looking for good examples of mathematical functions that have two different inverses, such as ↑; the second was trying to understand complex numbers properly, and why they are treated so mystically in most maths books. I studied complex numbers at university, and got all the answers right, so I thought I understood them. Only after writing this book, did I realise that I didn't *really* understand them, and only now do. (But that makes me worry that perhaps there is still yet another level of understanding to come...) The third thought-stream came from reviewing a son’s homework, and wondering “Why do they make something so simple seem so difficult?”

One of my conclusions was that mathematics teaching is influenced by the way that concepts were discovered historically, and that no one has taken the time to do a synthesis, and work out what order we ought to have discovered them in.

You should also understand that I wrote the book for my teenage sons, for their birthdays. Consequently I used a rhetorical style that I hoped would be appropriate for them (so that they would want to read it). I modelled the style on a book that I noticed one of my sons devoured, "10 Billion" by Stephen Emmott; basically one thought per page.

Hopefully, reading my text, it will all seem rather clear and straightforward. However, working out the new notations presented here consumed *reams* of paper as I tried different options, and compared different patterns of formulas. The result may seem obvious, but it went through many variations as I experimented with different options: as a colleague of mine who tries to design straightforward user interfaces once remarked: if you make it as easy to use as a coffee machine, they think of you as a plumber.

Although Complete Numbers are traditionally called ‘Complex Numbers’, and are made up of a combination of ‘Imaginary’ numbers and ‘Real’ numbers, they are, as I hope I have convinced you, neither imaginary nor complex, at least, no more imaginary than negative numbers, and no more complex than rational numbers. For that reason I've called them something reminiscent of their old name, while trying to be fairer to them.

It could be argued that the old names have been used for so long now we should just accept them as they are, and note that they don’t describe what they refer to, and move on. However, the names are both off-putting for newcomers, and misleading for laypeople, so I believe it is better to rename them to something more representative.

The traditional way of handling imaginary numbers is to reduce them to real numbers multiplied by the unit of imaginariness, i, where $i = \sqrt{-1}$. Thus is $i \times i = -1$, and a complex number might look like $3+4i$.

If early mathematicians had adopted the same approach to their fictitious negative numbers, they would have observed that the unit of fictitiousness is f, where $f = 0-1$, and noted that all fictitious numbers can be written as positive numbers times f, such as $5f$. All that you would have to remember is that $f \times f = 1$ to return us to the realms of the real world, so that for instance $5 \times 5f = 25f$, but $5f \times 5f = 25f^2 = 25$.

In other words, negative numbers would not have been given equal status with positive numbers, but treated as a poor sibling of them. So is it with the traditional treatment of imaginary numbers.

Another point worth mentioning, looking at that comparison between f and i, is that i is really just another sign like + and −: if $+3$ is a line to the right, and -3 is a line to the left, then $3i$ is just a line upwards, and $-3i$ is a line downwards, so if you are using classical representations of complex numbers, it would be more consistent to write something like $\oplus 3$ and $\ominus 3$ to emphasise the similarity, and then represent a complex number like $3+4i$ as $3\oplus 4$.

About the Author

Steven Pemberton is an Anglo-Dutch researcher affiliated with the CWI, the Dutch national research institute for mathematics and computer science, in Amsterdam, the Netherlands. He went to the same school as Stephen Hawking, and at university was tutored by Dick Grimsdale, the person who built the world's first transistorised computer, who was himself a tutee of Alan Turing.

After university, Pemberton worked coincidentally in Turing's old department in Manchester, writing software for the fifth computer in the line of computers Turing originally worked on.

Following that he was a lecturer in computing in Brighton, before moving to the CWI in Amsterdam, where he co-designed ABC, the programming language that formed the basis of Python. He was the first user of the open internet in Europe, when the first European internet node was started at the CWI in 1988. He was involved with the Web from the beginning, organising workshops at the first Web conference, and went on to co-design HTML, CSS, XHTML, RDFa, XForms, and a number of other Web technologies.

He currently researches internet notations, and declarative techniques.