1. The algebra of two-by-two matrices. Matrices with entries in any field F of "scalars" can be added and "scaled" in the most obvious manner. However, the product of two matrices is defined by:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & u \\ y & v \end{bmatrix} = \begin{bmatrix} ax+by & au+bv \\ cx+dy & cu+dv \end{bmatrix}. \tag{$*$}$$

This operation obeys most of the usual laws of multiplication, such as $A(B+C) = AB+AC$ and $(A+B)C = AC+BC$ ("distributivity") as well as $A(BC) = (AB)C$ ("associativity"), but it does not allow you to interchange the order of factors: $AB \neq BA$ in general.

A hard look at the definition of the product should convince you that the two distributive laws do in fact hold. Associativity is not quite as obvious. One way to see it is to imagine every matrix expressed in terms of the so-called matrix units, which are matrices having a single entry $= 1$ and all the others $= 0$. The matrix unit having the 1 in row i and column j (and 0 elsewhere) is denoted by E_{ij}. It is a good exercise to check that $E_{ij}E_{kl}$ equals E_{il} if $j = k$, and otherwise yields the zero-matrix.

Multiplying once more, we see that both $E_{ij}(E_{kl}E_{mn})$ and $(E_{ij}E_{kl})E_{mn}$ give the same result (what is it?). To verify the associative law $A(BC) = (AB)C$, we imagine all three matrices expressed as linear combinations of matrix units. Repeated use of distributivity then reduces both sides of this equation to the special cases already verified.

The obvious equation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} ad-bc & 0 \\ 0 & ad-bc \end{bmatrix} \tag{$\dagger$}$$

turns out to be a real gold mine of information. The second factor appearing in it is visibly concocted from the first factor A, and is known as its *adjoint*, denoted $A^\star$. The scalar $(ad-bc)$ on the right hand side is called the *determinant*, $\det A$. Our equation (†) can thus be abbreviated as $AA^\star = (\det A)I$. You should check that $A^\star A = (\det A)I$, as well. Here is what we can deduce:

(1) If $\det A \neq 0$, the matrix A has an *inverse*

$$A^{-1} = \frac{1}{\det A} A^\star, \tag{$\ddagger$}$$

which is a matrix with the property that $AA^{-1} = A^{-1}A = I$.

(2) If $\det A = 0$, then A is *singular*, which means that there is a $B \neq 0$ with $AB = 0$.

(3) Clearly, a singular A *cannot* have an inverse (multiplying $AB = 0$ by A^{-1} would force $B = 0$). *Conclusion:* every A is *either* invertible *or* singular, and the determinant says which.

(4) Substituting the right hand side of $A^\star = (a+d)I - A$ in (†) yields the *Cayley-Hamilton relation*

$$A^2 - (a+d)A + (ad-bc)I = 0,$$

which involves another interesting scalar $(a+d)$, called the *trace* of A and denoted by $\operatorname{tr} A$.

The following item not directly related to (†) is also noteworthy.

(5) Looking back to the product AB, as given in ($*$) above, and computing its determinant, we obtain $(ax+by)(cu+dv) - (cx+dy)(au+bv) = adxv + bcyu - bcxv - adyu = (ad-bc)(xv-yu)$, i.e. the miraculous formula:

$$\det(AB) = (\det A)(\det B). \tag{$**$}$$

2. Quadratic Field Extensions. In mathematical evolution, number systems have successively expanded in order to overcome arithmetic restrictions. For instance, the set $\mathbf{N}$ of *natural* numbers $1, 2, 3, 4 \ldots$ does not allow subtraction $a - b$ unless $b < a$, and so is augmented by 0 and the negatives to form the *integers* $\mathbf{Z}$. In order to permit unrestricted division (except by 0), the latter is then enlarged to the *rationals* $\mathbf{Q}$, which encompass all fractions or "ratios" of integers.

$\mathbf{Q}$ is a *field* — which means, it allows the four operations of elementary arithmetic: addition, subtraction, multiplication, and division by anything but 0 — but it cannot measure the diagonal of a square or the circumference of a circle. Therefore it is extended to the *reals* $\mathbf{R}$, a larger field which include such "irrationals" as $\sqrt{2}$ or π. Real numbers are often described as "decimal expansions that go on indefinitely." While this is not exactly a crisp definition, most students are happy to believe in it.

For algebraic purposes, it is usually unnecessary to extend from $\mathbf{Q}$ all the way to $\mathbf{R}$. One might, for instance, only need to consider numbers of the form

$$\alpha = a + b\tau\,, \quad \text{with} \quad a, b \in \mathbf{Q}, \tag{1}$$

where $\tau = \sqrt{t}$, for some rational number $t > 0$ which is not already a square in $\mathbf{Q}$ (for instance $t = 2$). The set E of all such numbers is again a field, as it is clearly closed under $+, -, \times$, and moreover allows unrestricted division by any $\alpha \neq 0$ on account of the formula

$$\frac{1}{a + b\tau} = \frac{a}{a^2 - tb^2} - \frac{b}{a^2 - tb^2}\,\tau. \tag{2}$$

This works for every non-zero pair a, b because $a^2 - tb^2 = 0$ is impossible, since it would mean either $t = (a/b)^2$ (which was ruled out) or $b = 0$ (which would imply that $a = 0$ as well). The field so obtained is usually denoted by $E = \mathbf{Q}[\tau]$ or $E = \mathbf{Q}[\sqrt{t}]$. It is bigger than $\mathbf{Q}$ but considerably smaller than $\mathbf{R}$.

If you look closely at this construction, you notice that the actual rationality of the coefficients a, b or of the number $t = \tau^2$ was not used: these items could just as well lie in some larger subfield K of $\mathbf{R}$ not containing $\sqrt{t}$, and then the result of the construction would be labelled $E = K[\sqrt{t}]$.

Looking yet more closely, you realize that the role of $\mathbf{R}$ is just to provide an environment in which the formulas (1) and (2) make sense. But it cannot always be used, since it does not (for instance) contain finite fields or square roots of negative numbers. Suppose, for example, you wanted to start from the finite field $K = \mathbf{F}_5$ and construct $E = K[\sqrt{2}]$. Since 2 is not a square in $\mathbf{F}_5$ the arguments surrounding (1) and (2) are formally correct (the proviso $t > 0$ was used only to make a square root available in $\mathbf{R}$) — but what kind of object is the formula $\alpha = a + b\tau$ referring to ? In particular, what is τ ?

A good way out of this quandary is to say

$$\alpha = \begin{bmatrix} a & bt \\ b & a \end{bmatrix} = a \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + b \begin{bmatrix} 0 & t \\ 1 & 0 \end{bmatrix} \tag{3}$$

and that E is simply the set of all such matrices, with a and b in the ground field K. Writing (3) more compactly as $\alpha = aI + bT$, and checking that $T^2 = tI$, you easily see that this set of matrices is closed under the operations $+, -, \times$. Moreover, $\det \alpha = a^2 - tb^2 \neq 0$ for every non-zero α (which is therefore invertible). A copy of the ground field K is contained in E as the set of all matrices of the form aI. This may seem awkward at first, but it is typical for the extension game: remember how integers had to be reinterpretd to qualify as rationals, or how fractions had to be viewed as infinite sequences to fit into $\mathbf{R}$.

The advantage of the matrix gambit that it works in every scenario, and requires no mumbo-jumbo about infinities. If, however, the ground field is $\mathbf{R}$ itself and t equals -1, it yields the field $\mathbf{C} = \mathbf{R}[\sqrt{-1}]$ of "complex numbers".

3. The Complex Plane. In the particular case of complex numbers, it is customary to write $a + bi$ instead of $a+b\tau$, i.e., to denote $\sqrt{-1}$ by the letter i (for "imaginary") — unless you are an electrical engineer, in which case you use the letter j (because i is reserved for "current"). You can safely do your complex arithmetic by formally manipulating these expressions, without always thinking of matrices.

However, there is a geometric aspect to $\mathbf{C}$, for which it is handy to remember the matrix interpretation: the complex number $w = u + vi$ is, after all, just a short-hand label for the matrix $M(w) = uI + vJ =$

$$u \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + v \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} u & -v \\ v & u \end{bmatrix} = \rho \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{1}$$

$= \rho\, R(\theta)$, where ρ, θ are polar coordinates for the planar point (u, v). Note that $J = R(90°)$ is a square root of $-I$ in a very natural way. The arithmetic of complex numbers is entirely determined by the interactions of these 2×2 matrices with entries from $\mathbf{R}$, that is:

$$M(w + w') = M(w) + M(w') \qquad \text{and} \qquad M(w \cdot w') = M(w) \cdot M(w'). \tag{2}$$

You might say that the difference between $w = u \cdot 1 + v \cdot i$ and $M(w) = u \cdot I + v \cdot J$ resides mainly in the "spelling" and that the complex number actually *is* the matrix. Someone else might say that a complex number is neither a matrix nor a formal expression $u + vi$, but quite simply a pair (u, v) of reals, or (if you like) a point in the plane. It's all in the eye of the beholder: if such pairs are just being added together, we consider them "vectors"; if they are also being multiplied as indicated, we elevate them to the status of "complex numbers".

We know all about vector addition — but what is the geometric effect of complex multiplication? The answer comes from (2): after all, the vector (u', v') is just the first column of $M(w')$, which gets transformed into the first column of $M(w \cdot w') = M(w) \cdot M(w')$ via the action of $M(w) = \rho R(\theta)$. Hence: *complex multiplication by w means matrix multiplication by $M(w)$, i.e. rotating by θ and scaling by ρ.*

In the complex context, the norm or length $\rho = |w| = \sqrt{u^2 + v^2}$ is traditionally referred to as the *absolute value* of w. Its relations with addition and multiplication are governed by the rules

$$|w + w'| \le |w| + |w'| \qquad \text{and} \qquad |w \cdot w'| = |w| \cdot |w'|, \tag{3}$$

respectively. The first of these is easy: if you draw the vectors $|w|$, $|w'|$, and $|w + w'|$ as arrows, you will understand why it is known as the "Triangle Inequality". The second rule in (3) can either be deduced from the geometric description of complex multiplication just given, or it can be inferred from the fact that $|w|^2 = \det M(w)$.

In $\mathbf{C}$ it is easy to find any n-th root, i.e. solve the equation $z^n = w$ for any given $w \in \mathbf{C}$ and $n \in \mathbf{N}$. Because of (1) and (2), you just have to choose z such that $M(z)^n = M(w)$, for instance by putting $z = \sigma(\cos\phi + i \sin\phi)$ where $\sigma^n = \rho$ and $n \cdot \phi = \theta$, so that $M(z) = \sigma R(\phi)$. Since angles obey the special rule $360° = 0°$, we actually get n different solutions by setting $\phi_k = (\theta + k \cdot 360°)/n$ for $k = 0, 1, \ldots, (n - 1)$.

On the next page we shall soon prove that this goes much further: *any polynomial equation $f(z) = 0$ can be solved in $\mathbf{C}$.*

4. Complex Polynomials. Over the complex numbers any polynomial equation $f(z) = 0$ has a solution. This *"Fundamental Theorem of Algebra"* is often taken on faith, although it is not exactly plausible. Why should $\sqrt{-1}$ open the gate for *all* polynomials? Our answer will focus on the continuous real-valued function $|f(z)|$ and use the fact that, restricted to any closed planar disc, such a function must reach a minimum. Intuitively this result is more accessible; in fact, it might seem obvious. Just imagine the graph of the function stretched over the disc in question like a circus tent. If there are no rips or fuzzy edges, there must be a point of lowest height.

For the rest of this dicussion, we consider a polynomial

$$f(z) = c_0 + c_l z^l + \cdots + c_h z^h, \tag{1}$$

with non-zero complex coefficients $c_0, c_l, \ldots, c_h$ and the powers of z arranged in ascending order. Our argument will be based on the following two properties of its absolute value $|f(z)|$:

(i) $|f(z)| > |f(0)|$ for all sufficiently large values of z, and
(ii) $|f(z)| < |f(0)|$ for certain small values of z.

Let us see how each of these two statements contributes toward the proof of the theorem. Note that (ii) depends vitally on the assumption that $f(0) \neq 0$ (without which there would be nothing to prove, anyway).

1. The absolute value $|f(z)|$ *attains a minimum.* Like every continuous function, $|f(z)|$ takes on a minimum value $m(r) \leq |f(z)|$, for z ranging over a closed disc of radius r, i.e. $|z| \leq r$. In particular, $m(r) \leq |f(0)|$. Hence, if r is chosen large enough, (i) guarantees that $m(r) < |f(z)|$ for z outside that disc, i.e. $|z| > r$, as well. In other words, for sufficiently large r, the minimal value $m = m(r)$ is a global minimum. (This is quite unusual for complex functions; even such a decent one as e^z does not achieve it.)

2. This minimum cannot be positive. Let $a \in \mathbf{C}$ be the place where the minimum is attained, i.e. $|f(a)| = m$. Now form the polynomial $g(z) = f(a - z)$, whose range of values is obviously the same as that of $f(z)$. In particular, the minimum of $|g(z)|$ is also equal to $m = |f(a)| = |g(0)|$. However, if m were not zero, (ii) could be applied to $g(z)$ and would say that $|g(0)|$ is *not* minimal. Since this cannot be, we conclude that $g(0) = f(a) = 0$, as was to be shown.

To establish the properties (i) and (ii), we consider $f(z)$ written as

$$f(z) = \varphi(z) + c_h z^h \qquad \text{or} \qquad f(z) = c_0 + c_l z^l + \psi(z), \tag{2}$$

and note that $\varphi(z)$ and $\psi(z)$ are negligible for very large z and very small z, respectively. In fact, if K is a real number greater than the sum of the absolute values $|c_0| + |c_l| + \cdots + |c_h|$, you can easily check that $|\varphi(z)| < K|z|^{h-1}$ for $|z| \geq 1$, and $|\psi(z)| < K|z|^{l+1}$ for $|z| \leq 1$ (see?).

The reason behind (i) is that, for very large values of z, the order of magnitude of $f(z)$ is dictated entirely by that of the term $c_h z^h$, which overpowers $\varphi(z)$. On the other hand, (ii) is due to the fact that, for very small values of z, the behaviour of $f(z)$ resembles that of $c_0 + c_l z^l$, because $\psi(z)$ counts for so little. To see how this implies (ii), let us first look at the special case $f(0) = c_0 = 1$, and try to find a particular $z = s$ such that $|f(s)| < 1$.

Start with a complex number w such that $c_l w^l = -1$, and scale it down by a real factor $0 < \varepsilon < 1$, putting $s = \varepsilon w$. By adjusting ε we can make s so small that $\psi(s)$ is less than half the size of $c_l s^l = -\varepsilon^l$ in absolute value. Then $f(s) = 1 - \varepsilon^l + \psi(s)$ with $|\psi(s)| < \frac{1}{2}\varepsilon^l$, and hence

$$|f(s)| \leq |1 - \varepsilon^l| + |\psi(s)| < |1 - \varepsilon^l| + \frac{\varepsilon^l}{2} < 1. \tag{3}$$

If $f(0) = c_0 \neq 0$ is not equal to 1, we simply apply this argument to the polynomial $c_0^{-1} f(z)$. It then produces an s such that $|c_0^{-1} f(s)| < 1$, whence $|f(s)| < |c_0|$ as required.