4000/6000, Day 21 Polynomials

We want to study the ring of polynomials in one "variable" X, i.e. linear combinations of powers of a variable X of form $a_0+a_1X+....+a_nX^n$, and see how it is similar to the ring of ordinary decimal integers, $a_0+a_1(10)+....+a_n(10)^n$, which are certain linear combinations of powers of the integer 10. We will review how to add, multiply and divide polynomials, stressing this analogy with ordinary integers, and show how to define a "size" function that allows us to imitate the proofs of the most important properties of integers, namely the linear combination property, (relatively) prime divisibility property, and unique factorization. For reasons discussed briefly below, we will usually discuss polynomials with coefficients chosen from a field F, or at least a domain such as Z.

Later, by imitating the equivalence relations used to define modular integers, we will see that "modular polynomials" provide a universal model for studying all "root field" extensions of the rationals Q. The smallest field containing the square root of 2, i.e. elements of form a + b(sqrt(2)) may be represented by polynomials over Q, if we agree that multiples of X^2-2 are equal to zero. Gaussian integers, i.e. expressions of form a + b (sqrt(-1)) = a+bi, are represented by polynomials over Z, with X^2+1 set equal to zero. The smallest field containing the cube root of 2, i.e. expressions of form a + b(cuberoot(3)) + c (cuberoot(3^2)) are just polynomials over Q subject to the condition that X^3-2 and its multiples, are all zero.

As a consequence, we will analyze the "vector space dimension" of these fields, and eventually deduce as a corollary, that some angles cannot be trisected using only straightedge and compass, hence there is no such construction that will trisect all angles. Recall that the Pythagoreans only reluctantly granted the fact that the rationals Q do not equal the reals R, whereas today we know the reals are an infinite dimensional vector space over Q. The solution to the trisection problem requires a comparison between intermediate fields of different dimensions, lying between the rationals and the reals. (The dimension of a constructible field is always a power of 2, while the solution to some trisection problems lies in a field of dimension 3.) This is probably why the Greeks could not solve this problem.

The ancient problem of "squaring the circle", i.e. of constructing a line segment of length π (or sqrt(π)), is also known to be impossible, but the proofs known seem much more difficult, and we will not take time to present one. The difference is that the field Q(π) is infinite dimensional over Q, which makes it even further from being constructible. (As an example of how difficult questions involving numbers as abstruse as π can be, I believe it is still unknown whether the sum (e +π) is actually rational, where e is the familiar base of natural logarithms, which we showed earlier to be irrational. I.e. while it seems very unlikely the sum is rational, apparently no one knows how to prove it is not. To put it another way, if you prove this, you will have one of the most famous PhD theses in mathematics ever produced at the University of Georgia.)

Polynomials.

Let A be any (commutative) ring, and consider the ring of all polynomials with coefficients in A. This ring is denoted A[X], and consists of elements of form $a_0+a_1X+......+a_nX^n$, where the coefficients a_i are elements of A. Notice that when we write a polynomial as $a_0+a_1X+......+a_nX^n$, it is understood that all coefficients of powers of X higher than n are zero, but it is also possible that some or all of the coefficients a_j with $j \leq n$, for example a_n, may be zero as well. We agree that two polynomials are equal if and only if all their coefficients are equal. For example $1 + X + 0(X^2) = 1+X$, since in both cases all coefficients of powers X^n with $n \geq 2$, are zero, as well as the coefficients of X and X^0 being equal.

Although explicit computations are easier with integer, modular integer, or rational coefficients, it is important for theoretical purposes to consider also polynomials with coefficients in any field, such as R, C, Q(i), Q(sqrt(2)), as well as Q, and Z_p where p is a prime integer, and in a domain such as Z, and sometimes even a more general ring. Polynomials with coefficients in a ring A, are often called polynomials "over A".

Adding polynomials

Recall how to add integers, by adding numbers in the same columns and then "regrouping". We add 365 to 879 in the usual way. I.e. 9+5 = 4 ones and 1 ten. then 1+6+7 = 4 tens and 1 hundred, and finally 1+3+8 = 2 hundreds and 1 thousand.

```
     11
     365
     879
     ___
 1244
```

We could rewrite this as $3(10^2) + 6(10) + 5$

plus $8(10^2) + 7(10) + 9$

equals

$$11(10^2) + 13(10) + 14.$$

Then we have to "regroup".

I.e. $14 = 4 + 1(10)$,

giving $11(10^2) + 14(10) + 4$,

and now $14(10) = 1(10^2)+4$,

so we get $12(10^2) + 4(10) + 4$.

Then $12(10^2) = 1(10^3) + 2(10)$,

which gives $1(10^3) + 2(10^2) + 4(10) + 4 = 1{,}244$.

Addition of polynomials is similar only easier, since we do not have to regroup.
I.e. consider adding the polynomials $3X^2 + 6X + 5$ and $8X^2 + 7X + 9$, with the same coefficients

as the integers 365 and 879. The sum is obtained by adding the coefficients of the same powers of X, as follows.
It is useful to add them in columns as we do with integers.

$$\begin{array}{l} 3X^2 + 6X + 5 \\ 8X^2 + 7X + 9 \\ \hline 11X^2 + 13X + 14. \end{array}$$

And that is all there is to it, no regrouping needed, since no constant, no matter how large, can ever become a multiple of X, and no constant multiple of X can ever become a multiple of X^2.

It is important to replace missing coefficients by zeroes when adding like this, just as we do in positional notation for integers. I.e. just as 307 means $3(10^2)+0(10)+7$, the polynomial $3X^2+7$ equals $3X^2+0X+7$. Thus to add $3X^2+7$ and $X^3-X + 8$, using columns, requires arranging them as follows, with zeroes in place of the missing coefficients. It also seems helpful to me to write the 1's explicitly in front of variables with "no" coefficient, which means the coefficient is 1.

$$\begin{array}{l} 0X^3 + 3X^2 + 0X + 7 \\ 1\,X^3 + 0X^2 -1X \ + 8 \\ \hline 1X^3 \ + 3X^2 \ -1X + 15 \ = \end{array}$$

$$X^3 + 3X^2 - X + 15.$$

Notice that we have rewritten the polynomial $3X^2+7$ which has many unwritten zero coefficients, as
$0X^3 + 3X^2 + 0X + 7$, to make it have coefficients in as many places as the polynomial $X^3-X + 8$.
I.e. any polynomial $a_0 + a_1X + a_2X^2 + \ldots\ldots + a_nX^n$ of degree n, can be rewritten as
$a_0 + a_1X + a_2X^2 + \ldots\ldots + a_nX^n + a_{n+1}X^{n+1} + \ldots\ldots + a_mX^m$, where $m \geq n$, if we set the coefficients $a_{n+1} = \ldots\ldots = a_m = 0$.

The presence of zero coefficients makes it tricky to write a general formula for addition, but we will try.

I.e. to add two polynomials of form
$a_0 + a_1X + a_2X^2 + \ldots\ldots + a_nX^n$ and
$b_0 + b_1X + b_2X^2 + \ldots\ldots + b_mX^m$, where $m \geq n$, set
$a_{n+1} = \ldots\ldots = a_m = 0$, and rewrite the first polynomial as

$a_0 + a_1X + a_2X^2 + \ldots\ldots + a_mX^m$. Then the sum is
$(a_0+b_0) + (a_1+b_1)X + (a_2+b_2)X^2 + \ldots\ldots + (a_m+b_m)X^m$,

where $a_{n+1} = ... = a_m = 0$.

Multiplying polynomials

The same story holds again for multiplication of polynomials, i.e. you multiply as for integers but there is no need to regroup at the end. To illustrate we multiply 365 by 879 as follows. First multiply the usual way, mentally carrying.

$$\begin{array}{r}
365 \\
879 \\
\hline
3285 \\
2555 \\
2920 \\
\hline
320{,}835
\end{array}$$

Written out in more detail as powers of 10, we get the following.

$$\begin{array}{r}
3(10^2) + 6(10) + 5 \\
8(10^2) + 7(10) + 9 \\
\hline
\end{array}$$

We multiply by one term at a time.

$$\begin{array}{rrrrr}
 & & 27(10^2) & + 54(10) & + 45 \\
 & 21(10^3) & + 42(10^2) & + 35(10) & \\
24(10^4) & + 48(10^3) & + 40(10^2) & &
\end{array}$$

Then add like powers of 10.

$$\begin{array}{rrrrr}
\hline
24(10^4) & + 69(10^3) & + 109(10^2) & + 89(10) & + 45
\end{array}$$

Finally we regroup.

$$= 24(10^4) + 69(10^3) + 109(10^2) + 93(10) + 5$$

$$= 24(10^4) + 69(10^3) + 118(10^2) + 3(10) + 5$$

$$= 24(10^4) + 80(10^3) + 8(10^2) + 3(10) + 5$$

$$= 32(10^4) + 0(10^3) + 8(10^2) + 3(10) + 5$$

$$= 3(10^5) + 2(10^4) + 0(10^3) + 8(10^2) + 3(10) + 5$$

$$= 320{,}835.$$

Now we multiply the analogous polynomials, with the same coefficients, $3X^2+6X+5$, by $8X^2+7X+9$. We line them up in the same way by columns, according to powers of X.

$$\begin{array}{rrrr} 3X^2 & +6X & +5 & \\ & 8X^2 & +7X & +9 \\ \hline \end{array}$$

We multiply again by one term at a time.

$$\begin{array}{rrrrr} & & 27X^2 & +54X & +45 \\ & 21X^3 & +42X^2 & +35X & \\ 24X^4 & +48X^3 & +40X^2 & & \end{array}$$

Then add like powers of X.

$$\begin{array}{c} \hline 24X^4 + 69X^3 + 109X^2 + 89X + 45 \end{array}$$

That is it, no regrouping needed.

Theoretically we describe multiplication as follows.
The product of $f = a_0+a_1X+......+a_nX^n$, by
$g = b_0+b_1X+......+b_mX^m$, is the sum of all the terms $a_ib_jX^iX^j = a_ib_jX^{i+j}$.

Using sigma notation,

$$\Big(\sum_{0\le i\le n} a_iX^i\Big)\Big(\sum_{0\le j\le m} b_jX^j\Big) = \sum_{\substack{0\le i\le n\\ 0\le j\le m}} a_ib_jX^{i+j}.$$

The main point to remember is you have to multiply each term in f by every term in g, and then add all the products together. For example to multiply a cubic by a quartic, you will get 20 products you have to add up.

Finally, since the expression $\sum_{\substack{0\le i\le n\\ 0\le j\le m}} a_ib_jX^{i+j}$ will have many terms of the same degree, we should combine terms of like degree in X, as follows.

$$\sum_{\substack{0\leq i\leq n\\ 0\leq j\leq m}} a_ib_jX^{i+j} = \sum_{0\leq k\leq(n+m)} \left(\sum_{\substack{i+j=k\\ i,j\geq 0}} a_ib_j\right) X^k.$$

If n = m = 2, we get $(a_0+a_1X+a_2X^2)(b_0+b_1X+b_2X^2)$

$= a_0b_0 + (a_0b_1+a_1b_0)X + (a_0b_2+a_1b_1+a_2b_0)X^2$

$+ (a_1b_2+a_2b_1)X^3 + (a_2b_2)X^4.$

For example, we can compute $(1+X+X^2)(2-X+3X^2)$ as

$2 + (-1+2)X + (3-1+2)X^2 + (3-1)X^3 + 3X^4$

$= 2+X+4X^2+2X^3+3X^4.$

In the example above, note that there is only one contribution to the term of highest degree, (and at most one contribution to the term of lowest degree as well). I.e.

$(a_0+a_1X+a_2X^2)(b_0+b_1X+b_2X^2)$

$= a_0b_0$
$+ (a_0b_1+a_1b_0)X$
$+ (a_0b_2+a_1b_1+a_2b_0)X^2$
$+ (a_1b_2+a_2b_1)X^3$
$+ (a_2b_2)X^4.$

If we multiply a quadratic by a cubic we get a similar pattern,

$(a_0+a_1X+a_2X^2)(b_0+b_1X+b_2X^2+b_3X^3)$

$= a_0b_0$
$+ (a_0b_1+a_1b_0)X$
$+ (a_0b_2+a_1b_1+a_2b_0)X^2$
$+ (a_0b_3+a_1b_2+a_2b_1)X^3$
$+ (a_1b_3+a_2b_2)X^4$
$+ (a_2b_3)X^5.$

In each case there is only one contribution to the terms of lowest and of highest degree, and usually more possibilities for terms of intermediate degree. There can however be many terms with the same number of contributions. For example if we multiply by a constant a_0, every term has only one contribution.

Then we get for instance,

$(a_0)(b_0+b_1X+b_2X^2+b_3X^3)$
$= (a_0b_0)$
$+(a_0b_1)X$
$+(a_0b_2)X^2$
$+(a_0b_3)X^3$.

Multiplication by a linear factor only allows at most two contributions to each term, but they still form a symmetrical pattern. I.e.

$(a_0+a_1X)(b_0+b_1X+b_2X^2+b_3X^3+b_4X^4)$
$= a_0b_0$
$+ (a_0b_1+a_1b_0)X$
$+ (a_0b_2+a_1b_1)X^2$
$+ (a_0b_3+a_1b_2)X^3$
$+ (a_0b_4+a_1b_3))X^4$
$+ (a_1b_4)X^5$.

The number of contributions starts out at 1, with a_0b_0, then goes up by one with each increase in the degree of the term, until the maximum possible number of n+1 is reached at the term of degree n. Then it stays there until the term of degree m, after which it starts going back down, one at a time, until it reaches one again at the top term of degree n+m, with a_nb_m.

Degree of a non zero polynomial.

If a polynomial has form $f(X) = a_0+a_1X+......+a_nX^n$, where $a_n \neq 0$, we call the term a_nX^n, the "leading term", and a_n the "leading coefficient", and n the "degree" of f. Thus the degree of f is the highest power of X occurring in f with a non zero coefficient.

We do not assign a degree to the zero polynomial, since no term in the zero polynomial has a non zero coefficient, (although some people like to say the zero polynomial has degree equal to minus infinity). Thus the degree is a function deg:A[X]-{0}--->N U {0}, where N U {0} denotes the non negative integers. A polynomial of degree zero is called a "constant".

Behavior of the degree function.

If we multiply $f = a_0+a_1X+......+a_nX^n$, by $g = b_0+b_1X+......+b_mX^m$, we get the sum of all the terms $a_ib_jX^iX^j = a_ib_jX^{i+j}$. As noted above, different products can give a term of the same degree. But there is only one way to get a term of highest possible degree. I.e. the only term of highest degree in the product $fg = (a_0+a_1X+......+a_nX^n)(b_0+b_1X+......+b_mX^m)$, is the term $a_nb_mX^nX^m = a_nb_mX^{n+m}$, of degree n+m.

If f has degree n, and g has degree m, i.e. if the coefficients a_n and b_m are non zero, and if the ring A of coefficients is a domain, then the product a_nb_m is also non zero. Since this is the coefficient of the highest power of X occurring in the product, i.e. of X^{n+m}, it follows that whenever the coefficients are chosen from a domain, the degree of the product of two polynomials is the sum of their degrees. I.e.

Lemma: If the ring A of coefficients is a domain, for example a field, then the degree function obeys the law deg(fg) = deg(f) + deg(g), for all non zero polynomials f,g in A[X].

Remark: This result is not true for rings which are not domains. For example if A = Z_9, then $(1+3x)(1+3x) = 1+6X + 9X^2 = 1+6X$, where the coefficients are meant to be integers mod 9. Thus the degree of this product of two polynomials of degree one, has degree one, instead of degree two. This is one big reason we prefer to use coefficient rings which are domains.

Units in a polynomial ring over a coefficient domain.

If the coefficient ring is a domain, then the degree of a product is the sum of the degrees of the factors, and since all constants have degree zero, then we cannot get 1 as a product unless both factors have degree zero. Thus the only possible units have degree zero. I.e. the only possible units in the polynomial ring are non zero constants.

Thus the units of a polynomial ring A[X] where A is a domain, are exactly the units of the ring A. In particular, the units in a polynomial ring with coefficients in a field are just the non zero constants. The units are thus exactly the polynomials of degree zero (since the zero polynomial has no degree).

If the ring of coefficients is not a domain, determining the units is much harder. For example, if A = Z_9 as above, then $(1+3X)(1-3X) = 1-9X^2 = 1$, so both 1-3X, and 1+3X are units in $Z_9[X]$.

This complication in recognizing units is another big reason we prefer coefficient rings which are domains.

Irreducible polynomials

The analog of a prime integer is called an “irreducible” polynomial. As usual a polynomial f in A[X] is irreducible if and only if f is not zero and not a unit and whenever f = gh, where g,h are polynomials in A[X], then one of g or h must be a unit. Again we see that we cannot recognize irreducible polynomials unless we can recognize units, another reason to restrict attention to coefficient rings which are domains. If A is contained in anmother ring B, then a polynomial f which is irreducible in A[X] may be come reducible in B[X]. We say then that f is irreducible "over A", but reducible "over B".

Assuming the coefficient ring is a field F, the constants are precisely zero and the units, so we can describe irreducible polynomials simply as follows. A polynomial f (over a field) is irreducible if and only if f is non constant and whenever f factors as f = gh, then either g or h is a constant.

For example, all polynomials of degree one over a field are irreducible, such as X-1, or 6X-7, or $(\pi^2)X+e$. Whether or not a polynomial of higher degree is irreducible depends on the field. for example over the real field, the polynomial X^2+1 is irreducible, but the same polynomial is reducible over the complex field, since it factors as (X+i)(X-i). Indeed over the complex field, the only irreducible polynomials are those of degree one, while over the real field, all irreducible polynomials have degree 1 or 2.

Over the reals, all polynomials of degree one are irreducible, and a polynomial of degree 2 of form aX^2+bX+c, with $a \neq 0$, is irreducible if and only if the discriminant b^2-4ac is negative. Over the rational numbers there are irreducible polynomials of every degree. Indeed if p is a prime integer then for every $n \geq 1$, the polynomial $X^n - p$ is irreducible over Q. It is an interesting and important fact, proved by Gauss, that a polynomial in Z[X] which is irreducible over Z, cannot become reducible over Q.

Divisibility theory and factorization of polynomials.

We want to carry out all the same basic results for polynomials that we did for integers, in terms of studying their divisibility and factorization properties. Fortunately the whole shooting match goes through in perfect analogy with the results we have been learning for Z and Z[i]. The basic tool is the size function, in this case the degree, and the first step is to show it satisfies a nice division theorem. As often seems to happen, this one key proof is the ugliest proof of all, so I will do this one. But you should be able to state it, by analogy with our other division theorems. (I.e. you can always divide by any non zero polynomial, and the remainder should have smaller degree than the divisor.) Actually if you have done a little long division with polynomials, this proof is just a matter of saying that the procedure you have done will always work, which you probably already believe anyway.

Review long division of polynomials

The same thing is true for division as for multiplication, i.e. division of polynomials is "easier" than division of integers, since there are no complications arising from regrouping or "carrying". For instance suppose we want to divide the integer 274 into the integer 8312. Since 2 goes into 8 four times, we might be tempted to begin the division with a 4, but when we multiply back we get a number which is too large.

$$\begin{array}{r} 4 \\ 284\,\overline{)8312} \\ 1136 \end{array}$$

So we might try a 3, but this is also too large.

$$\begin{array}{r} 3 \\ 284\,\overline{)8312} \\ 852 \end{array}$$

Finally we try a 2, and get lucky.

$$\begin{array}{r} 2 \\ 284\,\overline{)8312} \end{array}$$

$$\begin{array}{r}
568\\
\hline
263
\end{array}$$

then we bring down the 2 as usual, and continue.

$$\begin{array}{r}
2\\
284\overline{)8312}\\
568\\
\hline
2632
\end{array}$$

Now we have to divide 284 into 2632. Since 2 goes into 26, gee 13 times, we could guess that, but we know the answer is always at most 9, so we guess 9. Then multiplying back gives 2556, which works. I.e.

$$\begin{array}{r}
2\\
284\overline{)8312}\\
568\\
\hline
2632\\
2556\\
\hline
76
\end{array}$$

Thus the remainder is 76, so we get the equation 8312 = (2)(284) + 76.

We claim on the other hand, that dividing polynomials takes all the guessing out of what the answer should be. Let's assume we are dividing by a monic polynomial, such as X^2+8X+4, and dividing into any other polynomial, say $6X^4-X^2+20X+1$.

When we arrange this for division we must remember that the columns of the problem represent powers of X, and we must put in zeroes to represent missing powers of X. Thus our dividend $6X^4-X^2+20X+1$, must be written as $6X^4 + 0X^3 - X^2 + 20X + 1$. So we write the problem as follows.

$$X^2+8X+4\overline{)6X^4+0X^3-X^2+20X+1}$$

Now this time the difference is that we only need to look at the leading power of X in our two polynomials. that determines our quotient exactly. I.e. since X^2 goes into $6X^4$ exactly $6X^2$ times, that is the first term of the quotient, no guessing, no trial and error. I.e. we have,

$$\begin{array}{r} 6X^2 \\ X^2+8X+4 \,\big)\overline{6X^4+0X^3-X^2+20X+1} \end{array}$$

Then multiply back, change signs, and add (or subtract if you prefer).

$$\begin{array}{l} -6X^4 - 48X^3 - 24X^2 \\ \hline \quad -48X^3-25X^2 \end{array}$$

and then bring down the remaining terms

$$\begin{array}{l} \hline -48X^3-25X^2+20X+1 \end{array}$$

Now this is our new division problem, and the next term of the quotient is obtained by dividing X^2 into the new leading term, which gives -48X, and we multiply back, and change signs again, and add.

$$\begin{array}{r} -48X^3-25X^2 \quad +20X \quad +1 \\ +48X^3+384X^2+192X \qquad\; \\ \hline 359X^2+212X \quad + 1, \end{array}$$

do it again

$$\begin{array}{r} -359X^2-2872X-1436 \\ \hline -2660X-1435. \end{array}$$

Hence $6X^4-X^2+20X+1$
$= (X^2+8X+4)(6X^2-48X+359) -2660X-1435.$

This actually checks. Written out in columns:

$$\begin{array}{r} 6X^2-48X+359 \\ X^2+8X+4 \,\big)\overline{6X^4+0X^3-X^2+20X+1} \\ -6X^4 \;\; -48X^3-24X^2 \qquad\qquad\quad \\ \hline -48X^3 \;\; -25X^2 \; +20X \;\; +1 \\ +48X^3+384X^2+192X \qquad\; \\ \hline 359X^2+212X \;\; + 1 \\ -359X^2-2872X-1436 \\ \hline \end{array}$$

$$-2660X-1435.$$

Although the dividing part is easier than dividing integers, obviously we have to multiply and subtract a lot of integers to carry this out.

Let's try an easier one, like one in the book. Let's divide X^3-8 by X^2-X-2, writing the answer on the right.

$$\begin{array}{l}
X^2-X-2 \overline{\big)\, X^3 + 0X^2 + 0X - 8}\,;\ X+1 \\
\qquad\quad -X^3 + X^2 + 2X \\
\hline
\qquad\qquad\quad X^2 + 2X - 8 \\
\qquad\qquad\quad -X^2 + X + 2 \\
\hline
\qquad\qquad\qquad 3X - 6
\end{array}$$

Thus X^3-8 = $(X+1)(X^2$-X-2) + (3X-6).

We can prove by the first step of this division process, and the well ordering principle, that division by a monic polynomial is always possible, although it seems obvious this process must work. I.e. as long as your divisor has degree no bigger than the polynomial you get from subtracting, you can continue to divide and lower the degree every time. Thus eventually the degree of the remainder becomes lower than that of the divisor, (or the remainder becomes zero).

Here is the proof.

Division theorem for polynomials over a field: Given two polynomials f,h, over a field, where $f \neq 0$, there exist polynomials, g, r, such that
1) h = fg+r, and
2) either r = 0, or deg(r) < deg(f).

Proof: Since we can always divide through by a non zero constant, we may as well assume that the leading coefficient of f is 1, to simplify the notation in the proof. Ok,
$f = X^n+a_{n-1}X^{n-1}+....+a_1X+a_0$, and
$h = b_mX^m+b_{m-1}X^{m-1}+....+b_1X+b_0$, and we want to find a polynomial g such that h -fg = r has lower degree than f. We just choose our g so that fg has the same leading coefficient as h, and then when we subtract fg from h we get something of lower degree. We keep doing this until the degree is lower than deg(f). Maybe we can do this by induction (or well ordering). I.e. suppose deg(h) < deg(f). That is the really easy case. Then just take g = 0 and r = h. I.e. then h = 0(f) + h, and h has lower degree than f. done.

Now assume deg(h) ≥ deg(f), and assume there is some h for which this is impossible. Then there is such an h of lowest possible degree, so take that one. I.e. assume that deg(h) ≥ deg(f) and that the theorem is true for all polynomials of degree lower than h. Now we just want to multiply f up until it has the same leading term as this h, and then subtract it off. So consider the monomial b_mX^{m-n}. If we multiply this by f we get
$(b_mX^{m-n})f = (b_mX^{m-n})(X^n+a_{n-1}X^{n-1}+....+a_1X+a_0) =$

$b_mX^m+ (b_mX^{m-n})(a_{n-1}X^{n-1}+....+a_1X+a_0) = b_mX^m+ k(X)$, where k is a polynomial of degree

less than deg(h), i.e. less than m. Note also that this polynomial $(b_mX^{m-n})f = b_mX^m + k(X)$, has the same leading term as h does. Thus when we subtract, we get
$h - (b_mX^{m-n})f =$

$b_mX^m+b_{m-1}X^{m-1}+....+b_1X+b_0 - (b_mX^m+ k(X)) =$

$b_{m-1}X^{m-1}+....+b_1X+b_0 - k(X) = s(X)$,

is some polynomial of degree $\leq$ m-1. Thus we can divide f into s. I.e. there exist polynomials t,r such that

$s(X) = ft + r$ and $\deg(r) < \deg(f)$. But then
$h - (b_mX^{m-n})f = s(X)$,

so $h = (b_mX^{m-n})f + s(X) =$

$(b_mX^{m-n})f + ft + r =$

$f(b_mX^{m-n} + t) + r$, where deg(r) is still less than deg(f). **QED.**

Remark: A polynomial whose leading coefficient is 1, is called "monic". Note that the only place in this proof where we assumed we were working over a field was when we divided through by the leading coefficient of f to get a monic polynomial to divide by. Thus in fact the division theorem is true when dividing by monic polynomials over any ring at all. I.e.

Corollary (of proof): **Division theorem for monic polynomials over a ring:** Given two polynomials f,h, over a ring, where $f \neq 0$ is monic, there exist polynomials, g, r, such that
1) h = fg+r, and
2) either r = 0, or deg(r) < deg(f).

Definition: A "root" of a polynomial f(X) is an element r of some ring containing the coefficient ring of f, such that f(r) = 0.

Corollary (Root - factor theorem): If f(X) is a polynomial over a ring A, and r is an element of A, then r is a root of f(X), i.e. f(r) = 0, if and only if X-r is a factor of f(X).
proof: If X-r divides f(X), then there exists a g(X) such that f(X) = (X-r)g(X). Then setting X=r shows that f(r) = (r-r)g(r) = 0. Conversely, by the monic division theorem, given f(X) and r, there exists a polynomial g(X) in A[X] and a constant c in A, such that f(X) = g(X)(X-r) + c. Substituting X = r on both sides we see that f(r) = c. Hence X-r divides f(X) if f(r) = c = 0. QED.

(**Remark**: By repeating this process in the same way, we can write f(X) = (h(X)(X-r)+d)(X-r) + c = $h(X)(X-r)^2$ + d(X-r) + c, where d is the "derivative of f at r". This allows us to define the derivative of any polynomial, without limits.)

Corollary: If f(X) is a non constant polynomial of degree 2 or 3 over a field F, then f is reducible over F if and only if f has a root in F.
Proof: If r is a root of f in F then f(X) = (X-r)g(X) where deg(g) = 1 or 2, by the previous corollary. In the other direction, if f is reducible, then f(X) = g(X)h(X) where both g and h have positive degree and deg(g)+deg)h) $\leq$ 3. Hence at least one factor has degree 1. If say deg(g) = 1, then g(X) = aX+b for some a$\neq$0, b in F. Then r = -b/a is a root of g, hence also a root of f in F.
QED.

Take home problems:
Assuming these results, prove the following:
(Assume all polynomials have coefficients in a field.)

Due Wednesday:
1) Given f,g, not both zero, if d = fh+gk is a $\neq$ 0 linear combination of lowest possible degree, then d divides both f and g.

2) If f,g are relatively prime polynomials, i.e. if their only common divisors are non zero constants, then there exist polynomials h,k, such that fh+gk = 1.

3) If f,g are relatively prime polynomials and f divides gh, then f divides h.

4) If f is irreducible, and f divides gh, then f divides g or f divides h.

Due Friday.
5) If f is any $\neq$ 0, non unit polynomial, then f can be written as a product of (one or more) irreducible polynomials.

6) If $f_1,.....,f_n$, $g_1,....,g_m$ are irreducible polynomials, and if their products are equal, i.e. if $(f_1)(.....)(f_n) = (g_1)(....)(g_m)$, then n = m, and after possibly renumbering, we have $f_1 = c_1g_1$, $f_2 = c_2g_2$, ... , $f_n = c_ng_n$, where each c_i is a non zero constant.

Greatest common divisors.

Given two elements a,b of a ring A, we call another element d, a gcd of a,b provided two properties hold:
1) d divides a and b, and
2) any common divisor of a and b divides d.

Problem 1 above allows us as usual to prove the existence of a gcd for two polynomials f,g, not both zero, over a field.
I.e. Given f,g, not both zero, if d = fh+gk is a $\neq$ 0 linear combination of lowest possible degree, then d divides both f and g. We claim any such d is a gcd of f,g. We must check the two properties above. First, d divides f and g by problem 1. Then by the three term principle, since d is a linear combination of f and g, any common divisor of f and g also divides d. In a general

ring, it might be better to call these gcd's rather "universal common divisors" than greatest common divisors, although in the case of polynomials, any common divisor of f,g of largest degree, is a gcd.

These gcd's are not unique, since any non zero constant multiple of one gcd is another one. However we can divide any gcd through by its leading coefficient and get a unique monic gcd. In more general rings, like the Gaussian integers, there is no way to choose a special gcd, and then any two elements a,b have as many gcd's as there are units in the ring.

Recall the Euclidean algorithm for calculating a gcd, from the division algorithm. Let's use it to calculate the gcd of X^3-8 and X^2-X-2 in Q(X). Recall we begin by dividing the smaller into the larger, and finding the remainder.

Above we found that X^3-8 = (X+1)(X^2-X-2) + (3X-6). Then we divide the remainder into the old divisor. Since we are in a field, we divide 3X-6 by 3 to get a monic polynomial X-2. We can change this back later.

$X-2\,\overline{)\,X^2 - X - 2}$; X+1

$-X^2+2X$

$\overline{\qquad\qquad\qquad}$

$X-2$

$X-2$

$\overline{\qquad\qquad\qquad}$

Thus X^2-X-2 = (X-2)(X+1), so dividing by the original polynomial 3(X-2) gives us X^2-X-2 = [3(X-2)][(1/3)(X+1)]. Notice that although the quotient changes, the remainder stays the same. Now since this last remainder is zero, the previous remainder is our gcd, i.e. gcd(X^3-8, X^2-X-2) = 3X-6.

We can work backwards to actually write this gcd in terms of the original two polynomials. Since there was only one division step, this is trivial.
I.e. X^3-8 = (X+1)(X^2-X-2) + (3X-6), so X^3-8 - (X+1)(X^2-X-2) = (3X-6), expresses (3X-6) as a linear combination of (X^3-8) and (X^2-X-2). If we want the unique monic gcd, i.e. X-2, we divide through by 3, getting
(1/3)(X^3-8) - (1/3)(X+1)(X^2-X-2) = (X-2).

Exercise: Over Q, since X^2+1 is irreducible, and does not divide X^3+2X+1, their gcd should be 1. Find an expression of form f(X)(X^2+1) +g(X)(X^3+2X+1) = 1, with f,g in Q(X).

4000/6000 Modular polynomial rings

We have shown many parallels between the ring of integers and the ring of polynomials over a field. there is one construction we have made for integers that we have not yet made for polynomials, the parallel of the modular integers. I.e. recall that given any integer n we defined an equivalence relation on Z such that two integers r,s were equivalent mod n, if and only if n divides r-s, i.e. if and only if r = s+ nk for some integer k.

Then we checked that this relation was indeed an equivalence relation, and that it respected the arithmetic operations, i.e. sums of equivalent integers were equivalent, as were products. In our present situation if F is any field and F[X] the polynomial ring over F, if h is any non constant polynomial in F[X], define an equivalence relation on F[X] where f,g are equivalent mod h, if and only if f-g is divisible by h, i.e. if and only if f = g + kh for some polynomial k.

Then, as with modular integers, when we add and multiply equivalent polynomials and get equivalent answers. I.e.
Claim: If g, g_1 are equivalent mod f, and if also h , h_1 are equivalent mod f, then gh is equivalent to g_1h_1, and g+h is equivalent to $g_1 + h_1$.
proof: This is easy, but we must be careful, as always. We are assuming that $g = g_1 +fr$, and $h = h_1 +fs$, for some polynomials r,s. Then just add them and get $g+h = g_1 +fr + h_1 +fs = (g_1 + h_1) +fr +fs =$
$(g_1 + h_1) +f(r +s)$. Thus g+h is equivalent to $g_1 + h_1$.

Also if we multiply we get $gh = (g_1 +fr)(h_1 +fs) = g_1h_1 +g_1 fs +fr h_1 +frfs$
$= g_1h_1 +f(g_1 s +r h_1 +rfs)$, hence gh is equivalent to g_1h_1. **QED**

This says the set of all equivalence classes of elements of F[X] forms a ring, which we write as F[X]/(f). I.e. to define the sum or product of two equivalence classes, just pick one representative of each class, add or multiply those, and then take the equivalence class of the answer. The claim says all choices will give the same equivalence class as an answer. I.e. if we write [g] for the equivalence class of g, then we define [g]+[h] to be [g+h]. And this works because the claim shows that if $[g] = [g_1]$, and $[h] = [h_1]$, then also $[g+h] = [g_1+h_1]$. These equivalence classes are like polynomials, except all multiples of h are set equal to zero.

Lemma: If h is irreducible, then the ring F[X]/(h) is a field containing a copy of the constant field F.
proof: Recall Karl's' proof that Z_p is a field if p is a prime integer. Imitate it. I.e. if h is irreducible and f is not divisible by h, then there are polynomials g,k such that fg + hk = 1. then when we set h = 0, we get
fg = 1 mod h, i.e. [f][g] = [1], mod h. So [g] is the inverse of [f], mod h.

To show this new ring contains F, check that no two elements of F become equal, mod h. I.e. if c,d are constants in F, and [c] = [d], then h must divide c-d, but h has positive degree and c,d have degree zero. So for every $c \neq d$ in F, their equivalence classes [c], [d] are different in F[X]/(h). Thus the field F is injected into the field F[X]/(h). **QED.**

Recall also that in Z/(n), (another notation for Z_n), all equivalence classes were represented by integers in the set {0,1,2......,n-1}, since you could always divide a larger integer by n and throw away a multiple of n, leaving just the remainder, which is smaller than n. In the same way we can divide any polynomial by h(X) and throw away a multiple of h, leaving only the remainder,which has lower degree than h. More precisely.

Lemma: For any polynomial h, irreducible or not, every equivalence class in the ring F[X]/(h(X)) is represented (by zero or) a polynomial of degree less than deg(h).
Proof: Given any polynomial f(X) we can write f(X) = q(X)h(X) + r(X), where r=0 or deg(r) < deg(h), and r is equivalent to f (mod h). **QED.**

Thus if deg(h) = n, every element of the ring F[X]/(h(X)) is equivalent to a polynomial of form $a_0+a_1X+.....+a_{n-1}X^{n-1}$, where the a's are in F.

For example, in Q[X]/(X^2-2), the elements are all equivalent to ones of form a+bX, with a,b, rational numbers.

Similarly, the elements of Q[X]/(X^2-3) are all equivalent to ones of form a+bX where a,b, are rational numbers.

Note however, although these last two rings look the same from the standpoint of addition, they are very different in terms of multiplication, since in the first we set $X^2 = 2$, and in the second we set $X^2 = 3$.

Similarly, Q[X]/(X^2+1) is also made up of elements of form a+bX, with a,b, rational, but here we set $X^2 = -1$, when we multiply.

Thus the first field is essentially the same as Q(sqrt(2)), the second as Q(sqrt(3)), and the third as Q(i), where $i^2 = -1$.

These similarities allow us to calculate multiplicative inverses in all these fields, and in more general ones too.

Example: Let F = Q, the rationals.
Let's try the field Q[X]/(X^2+1). I.e. we claim the inverse of a+b[X] here should be (a-b[X])/(a^2+b^2).

I.e. look for polynomials f,g such that f(X)(X^2+1) + g(X)(a+bX) = 1. Try it first with just a+X.

Divide X+a $\overline{)X^2+0X+1}$; X -a

$-X^2 - aX$

$$\begin{array}{r} -aX + 1 \\ +aX + a^2 \\ \hline 1+a^2 \end{array}$$

So we get $X^2+1 = (X+a)(X-a) + 1+a^2$,

so $1+a^2 = (X^2+1) - (X+a)(X-a)$,

so $1 = (X^2+1)/(1+a^2) - (X+a)(X-a)/(1+a^2)$.

So when we set $(X^2+1) = 0$, we get $1 = (a+X)\ [(a-X)/(1+a^2)]$.

Now to invert a+bX. we have the inverse of $(a/b + X)$ as

$[(a/b - X) /(1+(a/b)^2)] = [(ab-b^2X)/(a^2+b^2)]$.

I.e. $[(a/b) + X][(ab-b^2X)/(a^2+b^2)] = 1$,

so also, shifting one b to the left factor, gives

$[(a) + bX][(a-bX)/(a^2+b^2)] = 1$,

so the inverse of $a+b[X]$, mod X^2+1, is $(a-b[X])/(a^2+b^2)$.

Does this look familiar from working with complex numbers? That is because this field $Q[X]/(X^2+1)$ is essentially the same as the field Q(i).

Example 2: Compute an inverse in $Q[X]/(X^2-2)$.
Let $h(X) = X^2-2$, and F = Q, the rational numbers, and compute the inverse of $a+[X]$, where a is in Q.
solution: We want to solve the equation $(X^2-2)g(X) + (a+X)(k(X)) = 1$, for some polynomials g,k. Then k will be the inverse of a+X, mod(X^2-2). We use the Euclidean algorithm again.

I.e. divide $X^{2}2$ by X+a, getting.

$$\begin{array}{l} X+a \ \overline{)X^2+0X-2} \ ; \quad X - a \\ \qquad -X^2-aX \\ \hline \qquad\quad -aX-2 \end{array}$$

$+aX+a^2$

---

a^2-2 So the remainder is a^2-2 a unit.

I.e. we have $X^2 - 2 =(X-a)(X+a) + (a^2-2)$. Hence
$a^2 -2 = X^2-2 - (X-a)(X+a)$,

so $1 = (X^2- 2)/(a^2) - (X-a)(X+a)/(a^2-2)$
$= (X^2 - 2)/(a^2) + (X+a)[(a-X)/(a^2-2)]$.

Now setting $X^2-2 = 0$, gives $(a-X)/(a^2-2)$ as the inverse of $a+X$, mod X^2-2.

If we want the inverse of $a+bX$, we divide by b, find the inverse of $a/b + X$ as $[(a/b) - X]/[(a^2/b^2) - 2] = [ab-b^2X]/[a^2-2b^2]$, so that

$[(a/b) + X] [ab-b^2X]/[a^2-2b^2] = 1$, hence multiplying and dividing by b on the left, gives

$[a + bX] [a-bX]/[a^2-2b^2] = 1$.

This should look familiar, from our experience with the field Q(sqrt(2)) where the inverse of a+bsqrt(2) was $(a\text{-bsqrt}(2))/(a^2-2b^2)$.
Claim: In fact these new modular rings are not new rings at all, but merely abstract version of the subfields Q(i) and Q(sqrt(2)) of R, generated by roots of the irreducible polynomials X^2+1, and $X^2 - 2$.

Example 3: In the same way we claim that the modular polynomial ring $Q[X]/(X^3-2)$ is "isomorphic to" the field $Q(\sqrt[3]{2})$. This lets us figure out how to find inverses in that field by using the Euclidean algorithm.

I.e. every element of the field $Q[X]/(X^3-2)$ is represented by a polynomial f of degree ≤ 2. If $f(X)$ is any non zero polynomial of degree ≤ 2, then f is relatively prime to (X^3-2), so there is an equation of form

$f(X)g(X) + h(X)(X^3-2) = 1$.

Setting $X = \sqrt[3]{2}$ in this equation gives

$f(\sqrt[3]{2}).\ g(\sqrt[3]{2}) = 1$, so $g(\sqrt[3]{2})$ is the inverse of $f(\sqrt[3]{2})$.

Use this method to find the inverse of $\sqrt[3]{2}$, and of $1+\boxed{\sqrt[3]{2}}$.

Since $\sqrt[3]{4} = (\sqrt[3]{2}\,)^2$, the field $Q(\sqrt[3]{2}\,)$ consists of elements of form $a + b\sqrt[3]{2} + c\sqrt[3]{4}$, where a,b,c, are rational numbers.

To invert $\sqrt[3]{2}$, we want to invert the element represented by the polynomial X, so divide X into X^3-2, as follows:

$$\begin{array}{l}
X\,\overline{\big)X^3+0X^2+0X-2} \;\;;\;\; X^2 \\
\;-X^3 \\
\hline
\qquad\qquad\quad -2
\end{array}$$

so we get $X^3-2 = X(X^2) - 2$. Thus $2 = X(X^2) - (X^3-2)$, so setting $X = \sqrt[3]{2}$,

gives $2 = X(X^2)$, so $1 = X[(X^2)/2]$. Thus the inverse of X, mod X^3-2, is $X^2/2$. I.e. the inverse of $\sqrt[3]{2}$, is $(\sqrt[3]{4}\,)/2$.

To invert $1+\sqrt[3]{2}$, do the same procedure for 1+X. I.e. divide 1+X into X^3-2.

$$\begin{array}{l}
X+1\,\overline{\big)X^3+0X^2+0X-2} \;\;;\;\; X^2-X+1 \\
\quad -X^3-X^2 \\
\hline
\qquad\quad -X^2 \qquad -2 \\
\qquad\quad +X^2 \;\; +X \\
\hline
\qquad\qquad\qquad X-2 \\
\qquad\qquad\quad\; -X-1 \\
\hline
\qquad\qquad\qquad\; -3
\end{array}$$

Thus $X^3-2 = (X+1)(X^2-X+1) -3$, so $3 = (X+1)(X^2-X+1) - (X^3-2)$. Now set $X = \sqrt[3]{2}$, and get $3 = (1+\sqrt[3]{2}\,)(\sqrt[3]{4} - \sqrt[3]{2} + 1)$. So the inverse of $1+\sqrt[3]{2}$, is

$(\sqrt[3]{4} - \sqrt[3]{2} + 1)/3$. Let's check it.

I.e. multiply back: $(1+\sqrt[3]{2}\,)(\sqrt[3]{4} - \sqrt[3]{2} + 1) = (\sqrt[3]{4} - \sqrt[3]{2} + 1) + (2 - \sqrt[3]{4} + \sqrt[3]{2}\,)$

$= 3$. Hence 1/3 of that is indeed 1.

Now you try it for $1+\boxed{\sqrt[3]{4}}$. Recall $\sqrt[3]{4} = (\sqrt[3]{2}\,)^2$ so $1+\boxed{\sqrt[3]{4}}$ corresponds to $1+X^2$ mod X^3-2.

Thus we divide X^2+1 into X^3-2, getting $X^3-2 = X(X^2+1) -(X+2)$. Then $X+2 = X(X^2+1) -(X^3-2)$, so now we divide $X+2$ into X^2+1, and get $X^2+1 = (X+2)(X-2) +5$, so $5 = X^2+1 - (X+2)(X-2)$
$= X^2+1 - (X+2)(X-2) = (X^2+1) - (X(X^2+1) -(X^3-2))(X-2)$
$= (X^2+1)[1-X(X-2)] +(X-2)(X^3-2) = (X^2+1)[1+2X-X^2] + (X-2)(X^3-2)$.

Now set $X = \sqrt[3]{2}$, and get $5 = (\boxed{\sqrt[3]{4}}+1)[1+2\sqrt[3]{2} -\boxed{\sqrt[3]{4}}]$, Check it.
$(\boxed{\sqrt[3]{4}}+1)[1+2\sqrt[3]{2} -\boxed{\sqrt[3]{4}}] = (\boxed{\sqrt[3]{4}} + 4 - 2\sqrt[3]{2}) + (1 + 2\sqrt[3]{2} -\boxed{\sqrt[3]{4}}) = 5$.

So $(\boxed{\sqrt[3]{4}}+1)^{-1} = [1+2\sqrt[3]{2} -\boxed{\sqrt[3]{4}}]/5$.

Here is an explanation of why this trick works.
Theorem: The modular polynomial ring Q[X]/(X^2-2), is "isomorphic to" the field Q(sqrt(2)), i.e. there is a correspondence between their elements which is 1 to 1 and onto, and which preserves the ring operations of addition and multiplication.

Proof: Recall the elements of the ring Q[X]/(X^2-2) are equivalence classes of polynomials in Q[X], where two polynomials are equivalent if and only if their difference is divisible by (X^2-2). We claim this is the same equivalence relation induced by evaluating the polynomials at X = sqrt(2). I.e. intuitively, setting (X^2-2) = 0, is the same as setting X^2 = 2, i.e. as aetting X equal to a square root of 2.

So define a map Q[X]-->Q(sqrt(2)), which sends f(X) to f(sqrt(2)). I.e. just set X equal to sqrt(2) in f(X). Since evaluating the sum f+g is th4 same as evaluating f and g separately and then adding the values, and the same holds true for multiplying, this map preserves addition and multiplication. I.e. setting X = sqrt(2) in f(X) and also in g(X), and then multiplying the values, gives the same answer as first multiplying out the polynomials f and g, and then setting X equal to sqrt(2).

We claim that f(sqrt(2)) = g(sqrt(2)) if and only if f-g is divisible by X^2-2. Certainly if f-g = (X^2-2)h(X), then f(X) = g(X) + (X^2-2)h(X), so setting X=sqrt(2) on both sides gives f(sqrt(2)) = g(sqrt(2)) + 0.

In the other direction suppose f(sqrt(2)) = g(sqrt(2)). Then we claim that f-g is divisible by X^2-2. By the division algorithm we can try to divide it, and we get an equation of form f(X)-g(X) = (X^2-2)h(X) + r(X), where r(X) has degree < 2. But then setting X = sqrt(2) on both sides gives 0 = r(sqrt(2). But there is not polynomial of degree ≤ 1 with rational coefficients satisfied by sqrt(2) except the zero polynomial, so r=0. I.e. f-g is divisible by X^2-2, as claimed. **QED.**

4000/6000 Day 23 Modular polynomial rings part 2

If F is any field and F[X] the polynomial ring over F, if f is any non constant polynomial in F[X], define an equivalence relation on F[X] where g,h are equivalent mod f, if and only if g-h is divisible by f, i.e. if and only if g = h + kf for some polynomial k.

Then, as with modular integers, when we add and multiply equivalent polynomials and get equivalent answers. I.e.
Claim: If g, g_1 are equivalent mod f, and if also h , h_1 are equivalent mod f, then gh is equivalent to g_1h_1, and g+h is equivalent to $g_1 + h_1$.
proof: This is easy, but I must be careful, as always. We are assuming that g = g_1 +fr, and h = h_1 +fs, for some polynomials r,s. Then just add them and get g+h = g_1 +fr + h_1 +fs = ($g_1 + h_1$) +fr +fs =
($g_1 + h_1$) +f(r +s). Thus g+h is equivalent to $g_1 + h_1$.

Also if we multiply we get gh $= (g_1 + fr)(h_1 + fs) = g_1h_1 + g_1fs + fr\,h_1 + frfs$ $= g_1h_1 + f(g_1s + r\,h_1 + rfs)$, hence gh is equivalent to g_1h_1. **QED**

This says the set of all equivalence classes of elements of F[X] forms a ring, which we write as F[X]/(f). I.e. to define the sum or product of two equivalence classes, just pick one representative of each class, add or multiply those, and then take the equivalence class of the answer. The claim says all choices will give the same equivalence class as an answer. I.e. if we write [g] for the equivalence class of g, then we define [g]+[h] to be [g+h]. And this works because the claim shows that if $[g] = [g_1]$, and $[h] = [h_1]$, then also $[g+h] = [g_1+h_1]$.

Lemma: If f is irreducible, then the ring F[X]/(f) is a field containing a copy of the constant field F.
proof: (We proved this in the previous notes.)

Since f was an irreducible polynomial, if it had degree ≥ 2 then by the root factor theorem it had no root in F. Hence he most important property of the new field F[X]/(f) is the next one.

Theorem: The polynomial f(X) has a root in the new field F[X]/(f).
proof: It is confusing to have so many X's running around, so it is better if we pick a new letter, like T, and reconstruct the field as F[T]/(f(T)). Since I cannot make a bar over a T, denote the class of T by t. Then f(t) = the class of f(T) = the class of 0. So t = class of T, is a root of f(X) in the field F[T]/(f(T)). QED.

In the language of the book, "Xbar" is a root of f in the field F[X]/(f). I.e. f("Xbar") = class of f = "0bar."

Example: Let F = Q, the rationals.
Let's try the field $Q[X]/(X^2+1)$. I.e. we claim the inverse of a+b[X] here should be $(a-b[X])/(a^2+b^2)$.

I.e. look for polynomials f,g such that $f(X)(X^2+1) + g(X)(a+bX) = 1$. Try it first with just a+X.

Divide X+a $\overline{)X^2 + 0X + 1}$; X -a

$-X^2 - aX$

$-aX + 1$

$+aX + a^2$

$1+a^2$

So we get $X^2+1 = (X+a)(X-a) + 1+a^2$,

so $1+a^2 = (X^2+1) - (X+a)(X-a)$,

so $1 = (X^2+1)/(1+a^2) - (X+a)(X-a)/(1+a^2)$.

So when we set $(X^2+1) = 0$, we get $1 = (a+X)\ [(a-X)/(1+a^2)]$.

Now to invert $a+bX$. we have the inverse of $(a/b + X)$ as

$[(a/b - X)/(1+(a/b)^2)] = [(ab-b^2X)/(a^2+b^2)]$.

I.e. $[(a/b) + X][(ab-b^2X)/(a^2+b^2)] = 1$,

so also, shifting one b to the left factor, gives

$[(a) + bX][(a-bX)/(a^2+b^2)] = 1$,

so the inverse of $a+b[X]$, mod X^2+1, is $(a-b[X])/(a^2+b^2)$.

Does this look familiar from working with complex numbers? We claim this field $Q[X]/(X^2+1)$ is essentially the same as the field $Q(i)$.

Example 2: Compute an inverse in $Q[X]/(X^2-2)$.

Let $h(X) = X^2-2$, and $F = Q$, the rational numbers, and compute the inverse of $a+[X]$, where a is in Q.

solution: We want to solve the equation $(X^2-2)g(X) + (aX+b)(k(X)) = 1$, for some polynomials g,k. We use the Euclidean algorithm.

I.e. divide X^22 by $X+a$, getting.

$X+a\ \overline{)X^2+0X-2}$; $X-a$

$-X^2-aX$

$-aX-2$

$+aX+a^2$

---

a^2-2 So the remainder is a^2-2 a unit.

I.e. we have $X^2 - 2 =(X-a)(X+a) + (a^2-2)$. Hence
$a^2 -2 = X^2-2 - (X-a)(X+a)$,

so $1 = (X^2 - 2)/(a^2)$ - $(X-a)(X+a)/(a^2-2)$. Now setting $X^2-2 = 0$ gives that $(a-X)/(a^2-2)$ is the inverse of $a+X$, mod X^2-2.

If we want the inverse of $a+bX$, we divide by b, find the inverse of $a/b + X$ as $[(a/b) - X]/[(a^2/b^2) - 2] = [ab-b^2X]/[a^2-2b^2]$, so that

$[(a/b) + X] [ab-b^2X]/[a^2-2b^2] = 1$, hence multiplying and dividing by b on the left, gives

$[a + bX] [a-bX]/[a^2-2b^2] = 1$.

This should look familiar, from our experience with the field Q(sqrt(2)) where the inverse of a+bsqrt(2) was (a-bsqrt(2))/(a^2-2b^2).

Day 23: 4000/6000 Reducibility of polynomials over the rational field (Corrected version)

The ring Q[X] has "unique" factorization of non constant polynomials into irreducible ones (i.e. unique up to constant multiples, and reordering of factors), but it is hard even to recognize an irreducible polynomial in Q[X]. We need some results to help us do this. Here are five basic ones, that still do not go very far.
There is a close connection between the concepts of irreducibility in Z[X] and irreducibility in Q[X]. They are not quite the same since prime integers in Z are irreducible factors in Z[X] and units in Q[X], which means that some reducible "polynomials" in Z[X], like 2X, or 6, become irreducible or even units in Q[X], but other than worrying about integer factors, there is little difference. The main result (due to Gauss) is that in the other direction things are ok, i.e. irreducible polynomials in Z[X] remain irreducible in Q[X].

NOTE: In general, a polynomial in Z[X] which is irreducible in Q[X] and whose coefficients have no common prime factor in Z, is irreducible also in Z[X]. (This is elementary to prove.)

I) If f(X) in Q[X] has degree 2 or 3, then f is reducible in Q[X] if and only if f has a root in Q, (and we know how to find all possible rational roots).

II)(Gauss) If f(X) cannot be factored into factors of degree ≥ 1 in Z[X], then it cannot be factored into factors of degree ≥ 1 in Q[X], i.e. f is irreducible in Q[X]. Thus a polynomial f in Z[X] of degree ≥ 1 is irreducible in Q[X] if and only if f cannot be factored in Z[X] into two factors both of degree ≥ 1.

III) If f has degree $n \geq 1$ in Z[X], and there is some prime p in Z such that the element {f} represented by f in the ring $Z_p[X]$ still has degree n, and {f} is irreducible in $Z_p[X]$, then f is also irreducible in Q[X]. If the coefficients of f have no common prime factor in Z, then f is even irreducible in Z[X].

IV)(Eisenstein). If f is of degree $n \geq 1$ in Z[X], if $f = a_0 + a_1X + a_2X^2 + ...+a_nX^n$, and if there is some prime p in Z such that p divides the coefficients $a_0, a_1, ..., a_{n-1}$ of f, but p does not divide a_n, and p^2 does not divide a_0, then f is irreducible in Q[X].

V. If f has degree 4 or 5 in Z[X], and the element {f} it represents in $Z_2[X]$ has the same degree, has no root in Z_2, and is not divisible by X^2+X+1, then f is irreducible in Q[X]. (Similarly, if f has degree 4 or 5 in Z[X], and the element {f} it represents in $Z_3[X]$ has the same degree, no root in Z_3, and is not divisible by any of X^2+1, X^2+X+2, or X^2+2X+2, then f is irreducible in Q[X].

We have proved I. already.

proof of II: Let f(X) = g(X)h(X) where g,h, are of degree ≥ 1 in Q[X]. Then taking common denominators of the coefficients, there is some n in Z and m in Z such that $ng = g_1$ has integer coefficients and also $mh = h_1$ has integer coefficients. Then we have $nmf = g_1h_1$, as an equation in Z[X]. We claim we can cancel the factors n,m on both sides, without introducing fractions,

and hence leaving a factorization of f by integer polynomials. Let p be any prime factor of nm. Then p divides the left side so it also divides the right side, i.e. p divides the product g_1h_1 in the ring Z[X]. Now we know nothing about this ring but in fact we claim it has a sort of prime divisibility property, at least for prime integers, and in particular
Claim: Since p divides the product g_1h_1 in Z[X], then either p divides g_1 or p divides h_1. (Notice that for p to divide a polynomial means that it divides every coefficient).

We will use the trick of "mapping" a polynomial g in the ring Z[X] to a corresponding polynomial {g} in the ring Z_p[X]. We do this simply by replacing each integer coefficient a of g by the element {a} in Z_p which it represents. This mapping from Z[X] to Z_p[X] preserves all the operations on polynomials. I.e. sums and products in Z[X] correspond to sums and products in Z_p[X]. I.e. if g,h are elements of Z[X], then the element {gh} represented by their product gh in Z[X] equals the product {g}{h} in Z_p[X]. We take this fact for granted without proof (as does the book). We have learned this behavior can be risky, but I assure you that in this case it is absolutely all right.

To prove the Claim, consider the polynomials $\{g_1\}$, {h}, and $\{g_1h_1\}$ represented by g_1, h_1, and g_1h_1, in the ring Z_p[X]. Our hypothesis says that the p divides g_1h_1, so the element $\{g_1h_1\}$ is zero in Z_p[X]. Since Z_p is a field, the polynomial ring Z_p[X] is a domain, so since by my assurances to you we also have $\{g_1\}\{h_1\} = \{g_1\}\{h_1\}$, either $\{g_1\}$ or $\{h_1\}$ must be zero in Z_p[X]. To be zero measn all coefficients are zero in Z_p. So p divides all coefficients of either g_1 or h_1, as claimed. **QED for Claim.**

Now in the equation nmf = g_1h_1, if we factor mn into prime integers, we have something like $(p_1)(p_2)(........)(p_r)f = g_1h_1$. And we have just proved that p_1 divides all coefficients of say g_1. Then we can cancel p_1 on both sides getting say $(p_2)(........)(p_r)f = g_2h_1$, where all polynomials f,g_2,h_1, still have integer coefficients. Then the same proof says we can do the same with p_2, then with p_3,......, and eventually with p_n. After these cancellations we have an equation like f = g_sh_t where g, h have the same degrees as g,h, but have integer coefficients. Thus from a factorization of the integer polynomial f into factors of degree ≥ 1 with rational coefficients, we have constructed a factorization of f into factors of the same degrees, but with integer coefficients. **QED.**

proof of III: We will show if f has degree ≥ 1 in Z[X] and is reducible in Q[X], and if {f} in Z_p[X] has the same degree as f, then {f} is also reducible in Z_p[X]. I.e. if f is reducible in Q[X], then by part II we may assume that f = gh, where g,h are non constant in Z[X], and then we have also {f} = {g}{h} in Z_p[X]. To show {f} reducible in Z_p[X] we only have to show that {g} and {h} are non constant.

This is not obvious since we could have say g = 3X + 1, so when we reduce mod 3, this becomes the constant {1}. But here we use our hypothesis on the degree. Since Z is a domain we have deg(f) = deg(g) + deg(h), and since p is prime, hence Z_p is also a domain, we have deg{f} = deg{g} + deg{h}. Since also deg({g}) ≤ deg(g), and deg({h}) ≤ deg(h), and by hypothesis we have deg({f}) = deg(f), then also we must have deg({g}) = deg(g), and deg({h}) = deg(h). Thus {g} and {h} are also non constant, so {f} is reducible. The contrapositive of this implies that if {f }has the same degree as f, and if {f} is irreducible in $Z_p[X]$, then f is also irreducible in Q[X].

Under the hypotheses of III, since f is irreducible in Q[X], then by II, f cannot be factored into factors of degree ≥ 1 in Z[X], hence the only way f can be reducible in Z[X] is if f has a non unit factor of degree 0, i.e. a non unit integer factor, i.e. an integer other than 1 or -1, which is a factor of every coefficient. Thus if in addition to the hypotheses of III, the coefficients of f have gcd 1, then f is actually irreducible in Z[X]. **QED.**

proof of IV: Let $f = a_0 + a_1X + a_2X^2 + ...+a_nX^n$ have degree n in Z[X], and assume that f is reducible in Q[X]. Then by II, f = gh, where $g = b_0 + b_1X + b_2X^2 + ...+b_rX^r$, and $h = c_0 + c_1X + c_2X^2 + ...+c_sX^s$, are polynomials in Z[X]of degrees $r,s \geq 1$, where r+s = n. Then since by hypothesis p divides $a_0 = b_0c_0$, but p^2 does not , then p divides one of b_0 or c_0 but not both. Say p divides b_0 but not c_0.

Since by hypothesis p does not divide $a_n = b_rc_s$, then p does not divide either of b_r or c_s. Thus p divides b_o and maybe some more b's but not all of them. Pick the one of smallest index that it fails to divide. I.e. say p divides $b_0,b_1,....,b_{t-1}$, but p does not divide b_t, where $0 < t \leq r < n = r+s$.

Then look at the coefficient a_t of f. This coefficient is $a_t = b_tc_0 + b_{t-1}c_1 + b_{t-2}c_2 + ...+b_0c_t$. This is true even if some of these c's are zero, e.g. if deg(h) < t. Since p divides all coefficients of f below the one of degree n, and since $n > r \geq t$, then p divides a_t. But this is a problem, since p also divides all the b's below b_t, but not b_t and not c_0. So p does not in fact divide a_t by the generalized 3 term principle, a contradiction. Thus f is not reducible in Q[X]. **QED.**

Proof of V, for Z_2. By III it suffices to check that {f} is irreducible in $Z_2[X]$. If not {f} has an irreducible factor of either degree 1 or 2. If it has a factor of degree 1 it would have a root, and the only irreducible quadratic polynomial over Z_2 is X^2+X+1 (exercise 6, p.111). Similarly the only irreducible polynomials, according to the book's list, are the ones given. (Exercise.)

Example: If p is any prime integer, and n > 0, then $X^n - p$ is irreducible in Z[X], and also in Q[X].

Example: $X^3 +79X + 99$ becomes X^3+X+1 mod 2, where it has the same degree ≤ 3, but no roots, hence is irreducible in $Z_2[X]$. Thus $X^3 +79X + 99$ is irreducible in Q[X], and since it is monic, it is also irreducible in Z[X].

Look at problems in section 3.3 on determining reducibility over Q of polynomials. (They are not all so easy.)

a. X^3+4X^2-3X+5,
b. $4X^4-6X^2+6X -12$,
c. $X^3 + X^2 +X + 1$,

d. X^4 - 180,
e. $X^4 + X^2$ -6,
f. X^4 -2X^3 +X^2 + 1,
g. X^3 +17X +36,
h. X^4 +X + 1,
i. X^5 +X^3 + X^2 + 1,
j. X^5 +X^3 +X + 1.

Second proof of Claim needed for Gauss Lemma.

In the proof above of the Claim, used in the proof of Gauss's result, part II, of the class notes, we gave the proof from the book, which used the concept of reducing a polynomial with integer coefficients to one with modular integer coefficients.

This proof uses an abstract idea, modular integers, but has the advantage of replacing multiples of p by zeroes, thus rendering the argument easier. I.e. it is easier to understand an expression when there are a lot of zeroes in it.

Here is a more elementary proof, more complicated, but with no abstract stuff, and nothing to take on faith. It may seem easier for that reason. (I will not use the same letters as above.)

Claim: Assume f,g,h, are in Z[X], f = gh, and p is a prime integer that divides every coefficient of f. Then either p divides every coefficient of g, or p divides every coefficient of h.

Proof: We will prove the contrapositive. I.e. we will show that if there is some coefficient of g that p does not divide, and also some coefficient of h that p does not divide, then there must be a coefficient of f that p does not divide.

It may help to do an example. Suppose g = 3 + 6X + 2X^2 +3X^3 and h = 12 + 4X +2X^2. Then we claim 3 cannot divide all coefficients of gh. Notice 3 divides all coefficients of g up until the X^2 coefficient, and all coefficients of h up until the X^1 coefficient. Adding those degrees gives 2+1 = 3, so look at the X^3 coefficient of f = gh.

The X^3 coefficient of gh is a sum of products,
(6X)(2X^2) + (2X^2)(4X) + (3X^3)(12). Each of these products involves a factor from one of the coefficients of p did divide in either g or h, except for (2X^2)(4X), obtained from multiplying the two lowest coefficients of g and h that p did not divide. Since p divides two of these terms but not the third, by the three term principle 3 cannot divide their sum.

Here is the general proof:

Assume that g = $b_0 + b_1X + ...+b_nX^n$, and h = $c_0 + c_1X +.....+c_mX^m$. Let b_r be the lowest degree coefficient of g that p does not divide, and let c_s be the lowest degree coefficient of h that p does not divide. I.e. p divides every coefficient in g of form $b_{(r-i)}$ for i > 0, and every coefficient in h of form $c_{(s-j)}$ for j > 0.

Then we claim that p does not divide the coefficient of the $X^{(r+s)}$ term in gh. Look at the coefficient of the $X^{(r+s)}$ term in gh.

It equals a sum of terms of form $b_{(r-i)}c_{(s+i)}$ and terms of form $b_{(r+j)}c_{(s-j)}$, with $i > 0$ and $j > 0$, as well as the term b_rc_s.

Now by hypothesis, p divides all the coefficients of form $b_{(r-i)}$ and all the coefficients of form $c_{(s-j)}$. Hence p divides every term of form $b_{(r-i)}c_{(s+i)}$ and every term of form $b_{(r+j)}c_{(s-j)}$, but not the one term b_rc_s. Hence by the generalized three term principle, p cannot divide their sum, i.e. p cannot divide the coefficient of the $X^{(r+s)}$ term in $f = gh$. **QED**

Notice that to follow a proof like this, or any proof of this nature, you must have done enough examples that you can visualize in your mind's eye what the terms look like.

Remark: This is the same proof as before, but without the modular integers. I.e. after reducing these polynomials mod p, the terms of g below degree r become zero, so the lowest non zero coefficient of {g} is $\{b_r\}X^r$, and similarly the lowest non zero coefficient of {h} is $\{c_s\}X^s$. Then the lowest coefficient of {gh} is $\{b_rc_s\}X^{(r+s)}$, which is also non zero in Z_p, since Z_p is a domain. But this coefficient of {gh} equals the coefficient of $X^{(r+s)}$ in gh, only taken mod p. Hence p does not divide the corresponding coefficient of gh.

So the simplification here was to reduce the coefficients of g and h mod p separately, and then multiply, instead of multiplying first and then reducing mod p. That these give the same result was exactly the step we (and the book) omitted proving, hence the reason the book's proof is easier is that it left out some of the steps. (Of course the book had done this proof in the case of modular integers Z_p, but not in the related case of polynomials with modular integer coefficients $Z_p[X]$.)

4000/6000 Fundamental homomorphism theorem

We have omitted four topics from the course that are contained in the first five chapters of the book, (i) the quadratic and cubic formulas, (ii) the isometries of the plane, (iii) the study of finite fields more complicated than the rings Z_p, (iv) the general concept of an ideal I in a ring R, the associated construction of a modular ring R/I, and the theorem that a surjective "homomorphism" f:R-->S induces an isomorphism of R/I with S, where I is the ideal {x in R: f(x) = 0}.

The quadratic formula is familiar from high school. If you make a clever substitution, you can transform any quadratic equation into one that has the simpler form $X^2 = d$. Then the answer can at least be denoted as X = sqrt(d), although computing sqrt(d), or approximating it, remains a problem. There is an even cleverer substitution for cubics, that transforms them essentially into the form $X^6 + bX^3 + c$. This is a quadratic in disguise, namely Y^2+bY+c, where the variable is $Y = X^3$, so one can use the quadratic formula to write Y in terms of a square root, and then X is the cube root of that expression for Y. Thus the solution to a cubic can always be written using only square roots and cube roots. This is nice and can even be done for 4th degree or quartic equations, but the procedure is unenlightening, since it is not easy to understand how anyone thought of these tricks. The really insightful investigations were made later, which reveal why these tricks run out of steam with 5th degree equations, quintics, and why there are quintic equations whose solutions cannot be written using formulas containing only coefficients of the equations, and various roots of expressions in them. Thus the time to look at the cubic formula may be when studying the later material on Galois theory.

The history of this topic is also exciting. Intellectuals used to challenge each other to algebra duels, tantalizing each other with problems to solve, and the guy with the cubic formula always came out on top. Since this situation was intolerable to his rivals, some of them also doped out the formula I believe, maybe with some secret help from someone who had been told the trick and promised to secrecy. Finally Girolamo Cardano published the formula in a book, not the man who found it first, and ever since, history has placed called it "Cardano's formula", even though everyone knows he did not discover it. This enraged the true discoverer, even though Cardano gave him proper credit in the book.

Many of us also know of the tragic history, centuries later, of young Evariste Galois, a brilliant but rash French student and political activist, whose insinuated threats against the king probably led to the duel in which he was killed at the age of 20 or 21. Feverishly, the night before the duel, he wrote up the results of his research, hoping someone would make use of it afterwards. It took some time for the great mathematician Liouville to actually understand what Galois had written that night, and reveal to the world the beautiful and complete answer to the problem of solution formulas he had discovered. After his tragic death, or assassination, he was at last regarded as a genius, whereas in his life he had even been denied access to the best schools. Recall in analyzing constructibility problems, that we have studied only the vector dimension of root fields, ignoring the multiplication rule in a field of form Q[X]/(f), and using only the degree of f. This is sufficient to get some handle on fields obtained by adjoining only square roots, i.e. constructible fields. To understand whether a root field can be obtained by successively adjoining higher order roots, i.e. whether its elements can be expressed using "radicals", Galois analyzed the symmetries of the field, the family of self isomorphisms of the field, called its "Galois group". This is covered later in our book.

The isometries of the plane are nice too, but we cover them in another course on transformation geometry, and they are in there as a predecessor to the study of groups, in chapter 6. Those of you continuing into 4010/6010 will benefit from reading those sections before (or while) studying chapter 6 on groups. The finite fields containing more than p elements, for example there is one containing p^2 elements, and one containing p^n elements for every $n > 0$, must be interesting, but I guess I don't know much about why that is. They do not greatly interest me - just ignorance I guess. Maybe a number theorist could enlighten me on their value, or someone knowledgable about codes. I confess though, I have always become interested in almost anything in mathematics, after learning enough about it to appreciate it.

The last topic above, ideals and modular rings and homomorphsms, is undeniably important, and we have already been using them in several fundamental examples. The general statements in chapter 4 are probably less important than the examples we have already studied, but it is helpful to see how the important examples we have studied give rise to the more general idea in chapter 4, so I will say a few things here summarizing sections 4.1 and 4.2.

Basically they generalize the construction of the modular rings Z_n and F[X]/(g(X)). I.e. in those cases we took a ring, and an element of that ring, and we set all multiples of that element equal to zero, and set two elements equal to each other if they differ by a multiple of that one element. The new idea is to use more than one element of the ring. I.e. suppose R is a ring, and we take two elements x,y of R, and set equal to zero any multiple of x, and also any multiple of y. Since the sum of two zeroes is also zero, then we have to set equal to zero also any sum of a multiple of x and a multiple of y. Thus we set all linear combinations of form ax+by, equal to zero, where a,b are in R. Then to get an equivalence relation in our ring, we set two elements of R equal to each other iff they differ by an element of form ax+by. We get a modular ring of these equivalence classes which we write as R/(x,y). Thus in this new ring R/(x,y), elements like [ax+by] become equal to [0], and hence two elements like [z] and [z+ax+by] become equal.

Remark: In the case of most of the rings we have been studying this would not give us anything new, because in both Z and F[X], the set of linear combinations of two things is just the same as the multiples of the gcd of the two things. This is true in any ring with a division algorithm. This is exactly whjat the linear combination propperty says. I.e. in any ring with the division theorem, given any two elements, not both zero, the smallest non zero linear combination of them, their gcd, divides both of them. Hence the set of linear combinations of the two original elements is exactly the same as the set of multiples of the gcd. But this is not true in the ring Z[X], and not in rings of polynomials with more variables, like Z[X,Y], Q[X,Y], or Q[X,Y,S,T]. E.g. there is no one polynomial in Z[X,Y] whose multiples are the same as linear combinations of the two elements X,Y. For example, in Z[X,Y], both X and Y are linear combinations of X,Y, but the only polynomials dividing both X and Y are 1 and -1. Thus 1 is the gcd of X,Y. Yet 1 is not a linear combination of X,Y. (Why not? A linear combination of X,Y equals zero when we set X = 0 = Y, but 1 never equals zero.) Thus the multiples of 1 are not the same as the linear combinations of X,Y.

In those rings we need the more general concept of ideals consisting of linear combinations of more than one element. In particular we can do this construction for more than two elements. I.e. let R be a ring and $x_1,....,x_n$ any finite set of elements in R. Then we can set equal to zero any linear combination of form $a_1x_1+....+a_nx_n$, with $a_1,....,a_n$ elements of R. Again we set two elements of R as equivalent if their difference is a linear combination of this type. We get another modular ring which we denote by $R/(x_1,....,x_n)$. The set of all these linear combinations $a_1x_1+....+a_nx_n$ is called the "ideal" generated by the set $x_1,....,x_n$. Since the sum of two linear combinations is another linear combination, and a multiple of a linear combination is also one, it follows that any linear combination of linear combinations is also a linear combination. Thus an ideal is closed under taking linear combinations. For this reason, it is common to define an ideal in R more abstractly as any non empty subset I of R which is closed under taking linear combinations of its elements with coefficients from R. Then we get a corresponding modular ring called R/I, whose elements are equivalence classes of elements of R, under the equivalence relation that x and y in R are equivalent (mod I) iff x-y belongs to the ideal I.

The reason for this construction is just to give a way of forming a lot of different rings out of a few standard rings. I.e. all the fields Z_p can be constructed in this way from the ordinary integers Z. More useful to us, every subfield F of C, which is finite dimensional as a Q vector space, can be constructed in this way from the one ring Q[X]. Let us prove at least part of that fact, as an illustration of the "fundamental homomorphism theorem" from chapter 4. We take for granted the more difficult result that every finite dimensional field extension of Q has the form Q[r] where r is a root of an irreducible polynomial in Q[X]. (In general a finite dimensional field extension of Q has the form $Q[r_1,....,r_n]$ where the r_i are roots of different polynomials $f_1,....,f_n$, but a nice theorem says that we can always find a suitable element r such that $Q[r] = Q[r_1,....,r_n]$. Remember for instance we showed that Q[sqrt(2), i] = Q[sqrt(2)+i].) Then we prove that every such field Q[r] can be represented by the modular construction above, applied to Q[X]. First we introduce a language for comparing different rings.

Definition: Given two rings R,S, a homomorphism from R to S is a function with domain R and range S, i.e. it is a mapping f:R-->S that takes each element x of R and transforms it into an element f(x) of S. It is also required that sums go to sums and products go to products, i.e. that for all x,y in R, we have f(x+y) = f(x) + f(y), and f(xy) = f(x)f(y). Finally we also require that f(1) = 1.

Examples: The map from Z to Z_n taking x to [x] is a homomorphism. The map from Q[X] to Q[i] taking X to i and more generally taking f(X) to f(i), is a homomorphism. If g is an element of Q[X], the map taking X to {X} in Q[X]/(g), and more generally taking f(X) to {f(X)}, is a homomorphism.

Definition: A homomorphism which is both one to one and onto) is called an isomorphism. If there is an isomorphism from R to S we say that R and S are isomorphic. (There may be no isomorphism, or more than one, between two given rings.)

In some sense isomorphic rings are almost the same. I.e. they may consist of different elements, but their elements have the same arithmetic relations to each other. There is a way to make elements of one ring correspond to elements of the other, so that corresponding pairs of elements have corresponding sums and products. Anything you understand about one ring, you also understand about every ring isomorphic to it. Isomorphic rings are not quite the same, because there may be more than one isomorphism betwen them, i.e. there may be more than one way to match up their elements still preserving all the ring operations. Thus if there are several isomorphisms between two rings and we have not chosen any particular isomorphism, then given an element of one ring there is still no way to say which element of the other ring it corresponds to. Once you choose an isomorphism, then the two rings may be regarded as the same, by means of that isomorphism, i.e. then you know for each element of one ring, which element of the other it corresponds to, by the given isomorphism.

If a ring has some symmetry, it may be isomorphic to itself in an interesting way, i.e. by some correspondence other than the trivial one taking each element to itself. For instance the complex numbers C have one non trivial isomorphism taking a+bi to its conjugate a-bi, since we know the product of conjugates is the conjugate of the product, and also the conjugate of a sum is the sum of the conjugates. The secret of Galois' theory which reveals why some equations have solution formulas and some do not, is to go beyond our study of just the dimension of a field extension, and explore also the symmetries of the field, i.e. the isomorphisms it has. Since every root field extending Q is isomorphic to a modular ring constructed from Q[X], understanding isomorphisms of root fields will be easier if we study first how to construct homomorphisms of modular rings. That is the content of the "fundamental homomorphism theorem".

First we illustrate it in our most important example.

Theorem: Let r be any complex number which satisfies an irreducible polynomial g in Q[X]. Then the two fields Q[r] and Q[X]/(g) are isomorphic, under the mapping taking {X} to r, and more generally taking {f(X)} to f(r).

Proof: First we show the mapping Q[X]/(g)-->Q[r] taking {f} to f(r) is well defined. I.e. in the ring Q[X]/(g), the element {f} is equal to the element {f + gh} for every polynomial k in Q[X]. Thus the rule defining our mapping must tell us to send {f} to the same element of Q[r] as we send {f + gh} to. But we send {f} to f(r) and we send {f+gh} to (f+gh)(r) = f(r) + g(r)h(r) = f(r), since g(r) = 0. So this is all right and we really have defined a function or mapping from Q[X]/(g) to Q[r].

Next we ask if this mapping preserves addition. I.e. does {f+h} map to the sum of what {f} and {h} map to? Well {f+h} maps to (f+h)(r) = f(r) + h(r), so yes. The same goes for multiplication, since (fh)(r) = f(r)h(r). What about {1}? Does it map to 1? Yes, since evaluating the constant polynomial 1 at r, gives the number 1.

Next we check one to one and onto. Ontoness is easy since every element of Q[r] has the form f(r) for some polynomial f in Q[X]. This is the image under our map of the element {f}, so the map is onto. For one to oneness, we ask when two elements {f} and {h} have the same image in Q[r], i.e. when does f(r) = h(r)? Well this happens exactly when (f-h)(r) = 0. Here is where we use the hypothesis that g is irreducible. I.e. we know that r is root of g, and g is irreducible, so by a proof we have given before, any polynomial in Q[X] satisfied by r must be a multiple of g. I.e. suppose k is a polynomial in Q[X] which is a multiple of g. Then say k = gm, for some polynomial m, so k(r) = g(r)m(r) =m 0, since g(r) = 0. Conversely if k is not a multiple of g, then since g is irreducible, g and k are relatively prime, and thus there exists a linear combination of form gA + kB = 1. Then since r is not a root of the RHS, it is not a root of both g and k. I.e. if k is not a multiple of g, then r is not a root of k.

Thus we have proved that if {f} and {h} have the same image in Q[r] under our map, i.e. if f(r) = h(r), then (f-h)(r) = 0, so f-h is a multiple of g. But if f-h = gk, then f = h+gk, so then by definition of equality in our ring Q[X]/(g), we must have {f} = {h}. I.e. if two elements {f} and {h} in the ring Q[X]/(g) map to the same element of Q[r] under our map, then those two elements are equal in Q[X]/(g). Thus our map from Q[X]/(g) to Q[r] is one to one. Thus our map taking {f} to f(r) is an isomorphism from Q[X]/(g) to Q[r]. **QED.**

The general result is less interesting, but will come in handy in future explicit situations, so we give the definitions.
Definition: Let R be any ring and $\{x_i\}_A$ any indexed collection of elements of R, possibly infinite. Let $I = I(\{x_i\}_A)$ = the set of all finite linear combinations $b_{i1}x_{i1}+....+b_{in}x_{in}$ of elements x_i in the collection $\{x_i\}_A$ with coefficients b_i in R. This set is called the ideal generated by the set $\{x_i\}_A$.

Lemma: (a) The ideal I generated a non empty subset of a ring R, is always non empty and is closed under forming linear combinations. I.e. given any finite set of elements $y_1,....,y_n$ of the ideal I, and any set of coefficients $b_1,....,b_n$ of R, the element $b_1y_1+....+b_ny_n$ is also in I.

(b) Conversely any subset I of R which is both non empty and closed under forming linear combinations is the ideal generated by some non empty subset of R, for example we could take the whole set I as a generating set.
Proof: This sort of thing is just done by bashing it out step by step. Certainly an ideal I generated by a non empty set is itself non empty since it contains 0, obtained from any finite linear combination with all coefficients equal to zero. Then we can check that a sum of linear combinations is a linear combination, and more generally that a linear combination of linear combinations is again a linear combination. Thus an ideal is non empty and closed under forming linear combinations.

Now if I is any non empty set which is closed under forming linear combinations, consider the set of all linear combinations of elements of I. This is by definition the ideal generated by the set I. We claim I itself is the ideal generated by I. We certainly do get each element in I as a linear combination of elements of I. Namely if a is in I, then a = 1a, is a linear combination with one non zero coefficient equal to 1. On the other hand since we assumed I is closed under forming linear combinations, we cannot get anything else except elements of I as linear combinations, so we get exactly I as the ideal generated by I. Well that wasn't so bad, just boring. **QED.**

This next result is more interesting, the connection between ideals and homomorphisms.
Lemma: Let R,S be rings and f:R-->S a homomorphism from R to S. If we define ker(f) = kernel(f) as the set of those elements x of R such that f(x) = 0, then ker(f) is an ideal of R.
Remark: I have no idea why this set is called the "kernel of f". In linear algebra a more plausible term is used such as "null space of f", since "null" is roughly a synonym for zero. "Kernel" is probably a translation of a German word from one of the early papers on the topic. Remember the first person to use a concept gets to name it whatever they want.
Proof: It suffices by the previous result to show that ker(f) is non empty and closed under linear combinations. First we show closure under (finite) linear combinations. Let $x_1,....,x_n$ be any finite set of elements of ker(f), i.e. assume they all map to zero under f, and let $a_1,....,a_n$ be any elements of R. We must show that $a_1x_1+....+a_nx_n$ also maps to zero. Since f is a homomorphism, hence preserves sums and products, $f(a_1x_1+....+a_nx_n) = a_1f(x_1)+....+a_nf(x_n)$ which equals zero, since for each i = 1,....,n, we assumed $f(x_i) = 0$.

Next we show ker(f) contains 0, hence is non empty. To show f(0) = 0, let's see, surely I can do this after all these years, but right now it is not coming to me. Oh, I remember now, since 0 = 0+0, then f(0) = f(0+0) = f(0) + f(0), so subtracting f(0) from both sides gives 0 = f(0). OK, I admit sometimes it helps to have some stuff memorized. **QED.**

The reason ideals are important is in the following result, the converse of the previous one, that every ideal is a kernel of some homomorphism. Well. almost all. Actually in any ring R, technically R itself is an ideal, and that ideal, called the unit ideal, cannot be the kernel of a homomorphism because we said a homomorphism had to send 1 to 1. If R were the kernel os sone homomorphism then all of R, including 1, would go to zero. So we prove every other ideal except R itself, is the kernel of some homomorphism.

lemma: In a ring R, let I be any ideal except R, and define a modular ring R/I as equivalence classes of elements of R, where x,y in T are equivalent in R/I if and only if x-y belongs to I. Then the mapping f:R-->R/I taking an element a to its equivalence class {a} is a surjective (i.e. onto) homomorphism with kernel I.

Proof: The definition of addition and multiplication in R/I is as follows. We define {a}+{b} to be {a+b}. For this to make sense, we must check that if {a} = {c} and {b} = {d} then {a+b} = {c+d}. But {a} = {c} means that a-c is in I, and {b} = [d} means that b-d is in I. Then since I is closed under addition, also (a-c) + (b-d) = (a+b -(c+d)) is in I. Hence {a+b} = {c+d}, as desired.

For multiplication, assume again that {a} = {c} and {b} = {d} and ask whether {ab} = {cd}. Well, {a} = {c} so a-c is in I, and {b} = {d} so b-d is in I, so their product is also in I, so we want to show that also ab-cd is in I. I think I remember how to do this. Let's try this: a-c = x in I, so a = c+x, and c-d = y is in I, so c = d+y, so ab = (c+x)(d+y) = cd + dx + cy + xy. Now since an ideal is closed under linear combinations, the fact that x,y, are in I implies that also dx + cy + xy is in I. So ab-cd = dx + cy + xy is in I. Hence {ab} = {cd} as desired. Thus the map taking a to {a} is a homomorphism.

That f is onto is trivial, since if {a} is any element of R/I then a maps to {a}, so f is onto. For the kernel, this is also trivial since if a goes to zero, then {a} = {0}. But by definition {a} = {0} iff a-0 = a belongs to I. **QED.**

Finally we want to show a partial converse to the last result.

Proposition: If f:R-->S is a surjective homomorphism with kernel I, then S is isomorphic to R/I, via the map F taking {r} to f(r).

Proof: First well definedness of F. I.e. we want to map {r} to f(r). So we must check that if {r} = {s} then F({r}) = f(r) = f(s) = F({s}). But by definition, if {r} = {s} then r-s is in I = ker(f), so f(r-s) = 0, hence f(r) = f(s), as desired. As to the homomorphism property for F, we claim that F({ax+by}) = F({a})F({x})+F({b})F({y}). But this is trivial since F({ax+by}) = f(ax+by), by definition of F, = f(a)f(x) + f(b)f(y), since f is a homomorphism = F({a})F({x})+F({b})F({y}), by definition of F. Thus once F has been shown to be well defined, the homomorphism property for F follows immediately from that for f.

Now to show F is an isomorphism, we must show it is both one to one and onto. Since f was onto, if y is any element of S, then there exists some element x of R such that f(x) = y, and then F({x}) = f(x) = y, so F is also onto.

For one to oneness, we must show if two elements {a},{b} of R map to the same element of S under F, then {a} = {b}. But if F({a}) = F({b}), then f(a) = f(b), so f(a-b = 0, so a-b belongs to I, so {a} = {b}, as desired. **QED.**

The fact that there is nothing new in these constructions for our familiar rings in which the division theorem is true, follows from the same ideas as always, but we spell it out explicitly as follows.

Definition: An ideal I in R is called "principal" if there is some one element a of that ideal, such that I consists precisely of all multiples of a by elements of R. A ring is called a "principal ideal domain" (or p.i.d., or just pid) if it is a domain, and all ideals in R are principal.

Proposition: If R is a domain having a notion of size for which the division theorem is true, then R is a principal ideal domain, and if I is any ideal of R, and a is a non zero element of I of smallest size, then I consists of all multiples of a.

Recall: Our hypothesis means that there is a size function d:R-{0}-->N assigning a natural number to every non zero element of R such that for any two elements a,b of R with $a \neq 0$, we can divide b by a, either exactly or with a remainder smaller than a, in the following sense. There exist elements q,r such that b = aq + r, and either r= 0, or d(r) < d(a).

proof of proposition: The zero ideal of R is generated by the element 0, so it is principal. Hence assume I contains some non zero elements, and that a is a non zero element of smallest possible size. We will show every element of I is a multiple of a, so let b be any elemnt of I. Then by hypothesis there exist elements q,r of R such that b = qa + r, and either r = 0, or d(r) < d(a). We will show first that r belongs to I. This is trivial since r = b-qa, and both a and b are in I, and since an ideal is closed under linear combinations, the linear combination b - qa of b and a is also in I. Thus if r were non zero the fact that d(r) < d(a) and that a has smallest size of all non zero elements of I would be a contradiction. Hence r = 0, so b = qa, hence b is indeed a multiple of a, and we are done. **QED.**

The property of all ideals being principal lies behind most of the results we obtained earlier from the division theorem. For example it is easy to prove the linear combination property, and the relatively prime divisibility property from the principal ideal property as follows.

Proposition: if R is any principal ideal domain, and a,b are relatively prime, then there exist elements x,y in r such that ax+by = 1.
proof: Consider the ideal (a,b) generated by a and b, i.e. consisting of all linear combinations of a and b. Since R is a principal ideal domain, this ideal must be principal, so there exists an element u in R such that (a,b) = (u) = all multiples of u. Since (a,b) = (u), we can write u as a linear combination of a and b, so there exist elements z,w such that za + wb = u.

But since both a and b are in (a,b) = (u), then both a and b are multiples of u, i.e. u is a common factor of a and b. But the assumption that a and b are relatively prime means then u must be a unit. Since u is a unit, if v is its inverse we then have vza+vwb = vu = 1. Thus if x = vz and y = vw, we get xa + yb = 1 as desired. **QED.**

Proposition: If a,b, are relatively prime elements of a principal ideal domain R, and a divides bz, for some z in R, then a divides z.
proof: This follows by the same argument as before since we can use the linear combiantion property. I.e. there exist x,y such that ax+by = 1. Then axz + byz = z, so since a divides a and bz, a also divides axz and byz, i.e. a divides the left side of the equation axz + byz = z, so a

divides the right side too. **QED.**

The prime divisibility property follows as usual, as does the uniqueness of prime factorization, (up to order and units). We must work a little harder to prove existence of prime factorization in a principal ideal domain, (without using the division theorem) but it is true. I will sketch the argument that a non unit element of a pid has at least one irreducible factor. If R is a p.i.d., and a is a non unit in R, supopose a has no irreducible factors. Then we can write a = a_1a_2 where (both factors are non units and) neither factor is irreducible. Hence we have a proper containment of ideals (a) contained in (a_2). Since a_2 is not irreducible we can write $a_2 = a_3a_4$, where neither a_3 nor a_4 is irreducible. We have another proper containment of ideals (a) contained in (a_2) contained in (a_4). If we never find an irreducible factor of a, this process keeps up forever and we get an infinite sequence of ideals properly containing those before: (a) contained in (a_2) contained in (a_4) contained in (a_6),........, etc. This gives a contradiction as follows. The union of an infinite family of increasingly larger ideals is easily shown to be closed under linear combinations, hence to be an ideal, hence to be equal to (b) for some b in R. Then b must be contained in one of the ideals (a_{2n}). But if b is in (a_{2n}), so are all multiples of b, i.e. so is the whole union of all the ideals (a_{2k}). This says that all the supposedly larger ideals (a_{2k+2}), (a_{2k+4}),...etc, are actually no larger than (a_{2k}), a contradiction to our assumption.

In a similar way we can prove that any non unit element a in a pid, factors completely into irreducible elements. I.e. suppose a is a non unit but cannot be written as a product of irreducible elements. Then a is not itself irreducible so a = a_1a_2 where (both factors are non units and) neither factor is irreducible. Moreover if both factors can be factored into irreducible factors, then combinaing those factors would give an irreducible factorization of a, so at least one factor say a_2 cannot be written as a product of irreducibles. Then a_2 is not irreducible, so $a_2 = a_3a_4$, where neither a_3 nor a_4 is irreducible, and again at least one factor say a_4, cannot be written as a product of irreducibles. This gives again an infinite sequence of non unit elements a, a_2, a_4,...., each one divisible by the next one, hence an infinite increasing sequence of principal ideals, hence a contradiction as before.

Thus every domain having the division theorem is a principal ideal domain, and every principal ideal domain is a unique factorization domain (or ufd). There are principal ideal domains which do not have the division property, but I do not happen to know any examples, and there are many unique factorization domains which are not principal ideal domains. Indeed we could easily prove that the domain Z[X] (in which the ideal (2,X) is not principal) is a unique factorization domain, using Gauss's lemma to deduce it from the unique factorization property for Q[X]. A similar method shows also that R[X] is a ufd whenever R is a ufd, and hence by induction that k[X_1,....,X_n] is a ufd for any field k.