Euclid's Elements: Introduction to "Proofs"

Euclid is famous for giving proofs, or logical arguments, for his geometric statements. We want to study his arguments to see how correct they are, or are not.

First of all, what is a "proof"? We may have heard that in mathematics, statements are proved to be either true or false, beyond any shadow of a doubt. In my opinion, this is not quite right. Rather we play a game of logical deduction, beginning from a set of assumptions that we pretend are true. Then assuming or pretending those things are true, we ask what other things would be true as well. We do not discuss whether the things we are pretending to be true really are true, but if we run across a world where they are really true, then we may be sure that anything else we deduce logically from them will also be true in that world. So there are statements we take for granted, called axioms or postulates or assumptions, and then there are statements called theorems that we deduce or "prove" from our assumptions. So we don't know that our theorems are really true, but in any world where the assumptions are true, then the theorems are also true.

In Euclidean geometry we describe a special world, a Euclidean plane. It does not really exist in the real world we live in, but we pretend it does, and we try to learn more about that perfect world. So when we "prove" a statement in Euclidean geometry, the statement is only proved to be true in a perfect or "ideal" Euclidean plane, but not on the paper we are drawing on, or the world we are living in.

It's a game like Monopoly, or dungeons and dragons, where we have a certain goal we want to achieve, but there are rules we agree to play by. We only win if we follow the rules. Sometimes the game gets boring, or too hard, and then we may just change the rules to make it easier or more fun.

The rules Euclid tried to play by are stated in his 5 postulates, and his "common notions". One fun thing about reading Euclid is trying to catch him using a rule he forgot to state. There are several places where this happens. It is also fun to see how clever he is at getting by without some of the rules we usually give ourselves to make the proofs easier. E.g. he makes his constructions not with ruler and compass, but with "straightedge" and compass. We are so used to saying "ruler" that I am going to do this sometimes, but his straightedge does not have marks on it like our ruler.

So Euclid's geometry has a different set of assumptions from the ones in most schoolbooks today, because he does not assume as much as we often do now. That makes some of his proof harder than ones in schoolbooks, because he does not give himself as much to go on. His geometry is also different from that of professional mathematicians because he forgot to state some postulates that he actually uses, and mathematicians have had fun suggesting postulates he might have or should have stated. Different people have different opinions about what the best postulates should be, so there is more than one way to do Euclidean geometry. We will look at Euclid's own version and make some choices of our own to fill in any gaps we notice.

At first we are going to try to use only postulates 1-4, as Euclid did, as well as his common notions. Those postulates are roughly as follows:
1) We can draw a finite line segment between any two different points.
2) We can extend a finite line segment as far as we want in a line.
3) We can draw a circle if we are given a center and a point on the circumference.
4) All right angles (half of a "straight angle") are equal.

Euclid also compares the size of different figures, the size of a collection of segments ia like the sum of their lengths, and the size of a plane figure is something like its area. But Euclid does not use numbers to measure either length or area, so he needs some rules to tell when two figures have the same size, or smaller or larger size. His "common notions" are really postulates for the concept of equal size. They mostly say that "equality" of (size of) figures behaves as we expect E.g. if you add figures of the same size you get new figures of the same size, and a figure that fits inside another with room to spare is not equal to it in size.

Week One:

Day One: Congruence
The first notion is that of congruence of triangles. A triangle is formed by three points not in a line, called vertices, and the three segments connecting them, called sides. Two triangles are congruent if there is a correspondence between their vertices, such that all three corresponding sides and all three corresponding angles are equal. E.g. triangle ABC is congruent to triangle DEF by the correspondence A<-->F, B<-->E, C<-->D, if corresponding sides are equal: AB = FE, BC = ED, CA = DF, and corresponding angles are equal: <ABC = < FED, <BCA = <CAB = <DFE.

Short cuts for recognizing congruence
We want to know when triangles are congruent by using less information than that. A congruence of triangles makes 6 corresponding sides and angles equal, but usually it is enough to know less. How much do you think would be enough?

Usually three pieces of data is enough, but not just any three. What do you think is enough to know? [SAS,SSS, ASA, AAS, but not SSA, or AAA. However you do learn something from AAA, and in fact SS-obtuseA is enough, i.e. two corresponding sides and a corresponding obtuse angle, or right angle.]

Many books assume one or two or even three of these, maybe all four, as postulates, but Euclid gives "proofs" for all of them. We are going to examine his arguments and see how convincing they are, and whether they really follow just from the few postulates he assumed.

Just remember it is like a game, where the other player is trying to win. Of course most of the time Euclid is going to win his own game. But to understand his strategy, we should read his argument step by step and ask for a good reason for everything he does. So go through his argument carefully and ask yourself why each step is allowed by the rules. By rules, we mean the postulates and the common notions, and any propositions that have been proved before the one we are doing at the moment. So in proving Prop. I.4, we can also use Propositions I.1, I.2, and I.3. Thus, we can construct an equilateral triangle, and can make a copy of a given segment anywhere we want. Our book contains the reasons for some arguments in the margin.

Book I: Congruence, parallels, and area

Congruence:
The first congruence result in Euclid is Proposition I.4, the famous "side - angle - side" or SAS rule. Many books today take this as a postulate, i.e. they just assume it to be true. Euclid proves it, but this is one important place where he lets himself do something he never said you could do in his postulates. See if you can guess what it is. This is not obvious and I am sure I never would have noticed it. Again we are sort of like fish and may not notice the sea we are in, because it seems so natural.

Rigid motions and SAS:

Prop. I.4: If two triangles ABC and DEF have two sides equal, AB = DE, and AC = DF, and also have the included angles equal, <BAC = <EDF, then the other corresponding sides and angles are also equal; i.e., BC = EF, and <ABC = <DEF, <ACB = <DFE.

proof of I.4: Assume given triangles ABC and DEF with sides AB and DE equal, sides AC and DF equals, and angles BAC and EDF equal. He claims that also sides BC and EF are equal, angles ABC and DEF are equal, and angles ACB and DFE are equal.

First Euclid "applies" or moves the triangle ABC onto triangle DEF. I.e. he places point A down on point D, and lines up sides AB with side DE. Then he says point B lies on point E because sides AB and DE were equal. Now we should ask ourselves what rule or postulate allows this? There is no rule saying we can move triangles, but we can copy segments, so we could always copy segment AB onto DE starting at D, and we would end at E since AB equals DE. But now he says also that the angle BAC has moved over also to angle EDF. Since we do not have a rule that we can copy angles, it is not clear to me how he gets a copy of angle BAC over on top of angle EDF. Anyway, if we grant him that he can move his triangle without changing the sides or the angles, then angle BAC would lie on top of angle EDF, and then C would end at F.

Now that B is at E and C is at F, it does seem that segment BC must equal segment EF, because they start and end at the same points. Of course postulate 1 that says we can draw a line segment between any two points, does not say only one line, so we must add that to postulate 1. The picture in Euclid's proof appears to show two such segments, and there was some argument saying this is impossible but the editor removed it as unnecessary. What do you think?

Summary of I.4: This proof uses "rigid motions" without good reason. Nowadays, some people add the postulate that one can move triangles without changing their sides or angles, so that Euclid's proof becomes ok, and many other people prefer to just take prop. I.4 as a new postulate. Which would you do? For this reason there is not just one version of postulates for "Euclidean geometry".

The proof also needs an expanded version of postulate 1, that only one segment can join the same two points.

Isosceles triangle principle, and self congruences
The next proposition "the isosceles triangle principle", is also very useful, but Euclid's own proof is one I had never seen before.

Prop.I.5: If triangle ABC is "isosceles", e.g. if sides AB and AC are equal, then the opposite angles ABC and ACB, are also equal.
proof: Euclid gives a clever but complicated proof, using Prop.I.4,. First he extends sides AB and AC to longer, still equal, segments AE and AF, then he considers the new triangles AEC and AFB. Since sides AE and AF are equal, as are sides AC and AB, and also angles EAC and FAB, the correspondence A<-->A, B<-->C, F<-->E, is a congruence of triangles ACE and ABF, by SAS (I.4). Thus angles BEC and CFB are equal, as are sides EC and FB. Since also sides BE and CF are equal, thus B<-->C,C<-->B, F<-->E is a congruence of triangles BCE and CBF. Since now angles ABF and ACE are equal, and also angles BCE and CBF, then by subtraction so are angles ABC and ACB. QED.

Discussion: Notice that points E and F can be anywhere below points B and C, as long as segments AE and AF are equal. But why not take E = B and F = C? Then we get that the correspondence A<-->A, B<>-->C, C<-->B is a congruence of triangle ABC with itself, so that angles ABC and ACB correspond. Thus these angles are equal!

This much shorter argument was found later, perhaps by Proclus. Almost all books today use this shorter proof. Why do you think Euclid did not use it? Is it confusing to think of a congruence of a triangle with itself, rather than of two different triangles? This is why I have emphasized the correspondence that gives a congruent, instead of just saying that two triangles “are congruent”. That is easier to say, but it is more precise to say that a certain correspondence is a congruence, since there may be more than one.

To make this argument clearer some books draw another copy of triangle ABC with the sides flipped over so B and C are interchanged. This may make it easier to picture because now there are two triangles instead of just one. What do you think? Which proof do you prefer? Does drawing the second copy of the triangle help?

Did Euclid use anything in this proof that he did not state?

Remark: Isosceles is a great spelling word!

[Euclid also proves the converse, but this seems unnecessary, since it also follows from Prop. I.18, larger angle / larger side theorem.
Prop.I.6: A triangle with two angles equal has the opposite sides equal as well.
proof: His proof is clever and uses I.4 and common notion 5, seemingly in reference to the area of a triangle, but he could also use I.5 and use CM5 for angles. I.e. if one side were shorter he copies that shorter side in the longer side, obtaining a triangle inside the original one but congruent to it by SAS, a contradiction.

Discussion: His proof seems to assume a triangle cannot be congruent to another triangle inside it, but there is no postulate stating that. I.e. why can a triangle not be both inside of and also congruent to another triangle? It seems more clear for angles, i.e. that an angle cannot be both greater than and equal to another angle. Or have we missed something? What does he mean by saying two triangles are “equal”? This will come up again and we will learn more about “equality” of triangles. The converse I.6 seems easier to me after we have ASA congruence. It also follows from Prop. I.18.]

Day Two:

The next proposition, “side-side-side”, is also famous and useful. It was my favorite in school, probably because it is easy to remember. Teaching this course I have learned it is also the one that lets us copy triangles by ruler and compass by copying their sides.

Line separation and “side - side - side” (SSS):
Prop.I.7-8: If ABC and DEF are two triangles and the correspondence A<-->D, B<-->E, C<-->F makes corresponding sides equal: AB = DE, AC = DF, and BC = EF, then the corresponding angles are also equal, <BAC = <EDF, <ABC = <DEF, <ACB = <DFE.
proof: Using a rigid motion, move base EF over to lie on base BC with E at B, and F at C. Flip triangle DEF if necessary to get vertex D on the same side of base BC as vertex A. Then Euclid claims that vertex D will lie on vertex A. His argument is by contradiction. There are several cases.

Try to imagine how this could fail? Say D lies on segment AC but not at A. [Since F lies at C, then side AC would not be equal to side DF. So this cannot happen.]

Now suppose D lies outside triangle ABC and off to the right (so neither triangle ABC nor DBC contains the other). Then look at triangles BAD and CAD. Since sides BA and BD are equal, by I.5 angles BAD and BDA are equal. But for the same reason, angles CAD and CDA are equal. But CDA contains BDA, which equals BAD, which contains CAD. Thus CAD is smaller than CDA by common notion 5, a contradiction.

What other cases can you think of? (D inside triangle ABC? Then what? Note that sides AB and AC divide that side of line BC into 4 parts. So D could be in any one of those parts, or it could be on one of those lines. So there seem to be 8 cases. But some of them are symmetrical wrt others and do not need to be done again. I think there are essentially three cases.)

Discussion: The cases we listed are based on the assumption that a line separates the plane into two sides. But how do we know this? This is another fact Euclid forgot to state, although he has used the language of “sides” of a line several times, e.g. in postulate #5, and proposition 7. It is related to the fact that the plane has two dimensions, which is not so easy to make precise. So we need another postulate, that every line separates the plane into two sides, and we should say that we can tell when two points are on the same side of a line, because then the segment joining them does not cross the line.

Alternate proof: Croix suggested a different looking proof, which also starts out moving base EF over to lie on base BC with E at B, and F at C. But he then flips triangle DEF if necessary to get vertex D on the opposite side of base BC from vertex A, instead of the same side. Then he also connects up segment AD, which now crosses the base BC = EF at some point X. Suppose X lies between B and C. Then he also considers the isosceles triangles ABD and ACD. By the isosceles triangle principle again, angles BAD and BDA are equal, as are angles CAD and CDA. But now we have a positive conclusion that, by adding, angles BAC and BDC are equal. Thus by SAS triangles BAC and BDC are congruent. Since DBC is congruent to our original triangle DEF.

This simple variation on Euclid’s argument has an advantage, for people who wish to avoid the law of excluded middle, that it is direct instead of indirect. There are again several cases to deal with depending on where the point X is, since it could lie outside the base BC, or could lie at B or C, and again we need the plane separation property to be sure we have all cases.

Remarks: After proving SSS, we no longer need to assume rigid motions to move a triangle, because we can copy segments. Thus by SSS we can copy triangles and thus also angles. In particular we do not need a protractor to copy angles. Some books assume SAS and that one can copy both segments and angles. By SAS, copying two segments and one angle lets us copy a triangle without assuming rigid motions. Which do you prefer? In fact I think Euclid no longer uses rigid motions from here on.

Notice also that if Euclid is assuming rigid motions, then he could use them to move or copy segments, so why did he assume he could copy segments?

Some books use straws and strings to support SSS physically, but note this only shows that the triangle does not “wobble” if the sides do not change. There could still be another different triangle with the same sides. [For instance there are usually two different triangles with the same SSA data, but they are not close to each other, so you cannot wobble one into the other. So showing that a triangle cannot wobble when certain measurements are fixed only suggest there are only a finite number of such triangles, it does not really argue there is only one.]

Exercise: Use a compass to copy a triangle ABC onto a segment EF which equals BC. [Place center of compass at E and draw circle of radius BA. Then place it at F and draw a circle of radius CA. They intersect at D so that DEF is congruent to ABC.]

Exercise: Angle - side - angle congruence (ASA)
The next congruence result is called ASA, “angle-side-angle”. It can be proved in the same way as the previous ones. **You try it.**

[Proposition I.26(a) If two triangles ABC and DEF have two angles equal, say <ABC = <DEF, and <ACB = <DFE, as well as the included sides BC = EF, then the other corresponding sides and angles are also equal, i.e. the correspondence A<-->D, B<-->E, C<-->F is a congruence.
proof: We could move triangle ABC over onto DEF so that base BC lies exactly long EF. Then we just need to show that vertex A lies on vertex D. By the equality of angles, we know side BA does lie along side ED, and we also know side FD lies along side CA. But then vertex A, which is on both sides BA and CA, must lie on sides ED and FD, i.e. A must lie at D. Notice we did not use any of the earlier propositions in this argument, but we used a rigid motion.

Euclid does not use this proof. Rather he can now copy an angle by using SSS, and he knows by SAS, that copying two sides and an angle, suffices to copy a triangle. So now that Euclid can copy sides and angles and he has SAS, he stops using rigid motions, and instead he just copies triangles. So he says he will copy side BA onto side ED, from E to X. Then by SAS, he knows triangle XEF is congruent to triangle DEF. But then angle XED should equal angle ACB, which only happens if point X = point D.
QED.

Exercise(Prop.I.6): Prove that a triangle with two equal angles has the two opposite sides equal, using ASA and the same congruence as for the isos. triangle principle.

Question: what happens when you try this proof for AAS?

Answer: You need to know that if you draw two different segments px and py from the same point p to the same line L, (with p not on L), then the two angles the segments px and py make with L are different. And we do not know how to prove this yet! So next we introduce the wonderful "exterior angle theorem" of Euclid. Prop. I.16.

Day Three: Angle - angle - side congruence (AAS) and exterior angles

Now what about the last famous congruence theorem, AAS? What happens when we try to prove that as before? Suppose in triangles ABC and DEF, that angles B and E are equal, and angles C and F are equal, and sides AC = DF. Now we can move triangle ABC over onto DEF so that side AC lies on side DF, and angle C fits onto angle F, so that base BC lies along base EF. But how do we know that vertex B is at vertex E? WE are supposed to have angle ABC equal to angle DEF, and if vertex B is closer to vertex C=F than E is, it looks as if angle ABC would be larger than angle DEF, because it is an "exterior angle" for triangle DEB. But we do not know anything about exterior angles yet. So we need to prove that result.

This exterior angle theorem of Euclid, is proved without using postulate 5, and this is a theorem I did not see in my high school class. It is much easier to prove using postulate 5, but Euclid gives a clever proof with only the other postulates, so we want to see how he did it. This will make the theorem true also in another geometry called hyperbolic geometry, where postulate 5 does not hold.

Exterior angle theorem and "vertical angles"

Exercise(I.15): If two lines cross at P they make four angles with alternate or "vertical" angles equal in pairs. [Hint: all "straight angles" are equal, by postulate___ ?]

Remark: We need to know how to bisect a segment now. (This was discussed in constructions class.)

Prop.I.16(Exterior angle theorem): If any side of a triangle be extended, the exterior angle formed is greater than either of the two remote interior angles.
Proof: Given triangle ABC, extend base BC past C to D. Claim angle ACD is greater than angle BAC. To prove this we will try to make a copy of angle BAC inside angle ACD.
Bisect side AC at X and extend BX equally to Y, so that BX = XY. Connect YC. Now claim B<-->Y, A<-->C, X<-->X is a congruence of triangles BAX and YCX, by SAS, using vertical angles and construction. Hence angle <ACY = <XCY = <XAB = <CAB, and <ACY is inside angle <ACD, hence smaller. QED.

Exercise: Why is angle ACD also larger than angle ABC? [Hint: What angle is “vertical” to <ACD?]

Exercise: Prove Prop. I.26b(AAS)

Exercise: Prove Prop. I.17: In any triangle, any two angles together are less than a straight angle.

The triangle inequality:
Next Euclid explains exactly which segments can be sides of a triangle. First he proves a useful fact related to the isosceles triangle principle. I.e. not only are the angles opposite equal sides equal, but the smaller angle is opposite the smaller side. This is another very useful and natural result I do not recall from my high school class.

Larger side / larger angle theorem:
Prop. I.18: In any triangle, the greater side is opposite the greater angle.
Proof: Assume not. Let angle <ACB be greater than <BAC and assume side AB is less than base BC. Extend side BA past A to X so that BX equals BC. Then in triangle BXC, sides BX and BC are equal so angles <BCX = <BXC. Now <BAC is exterior to <BXC hence <BAC is greater. But <BAC was assumed smaller than <ACB = <BCX = <BXC, a contradiction. QED.

Remark: Prop. I.18 also implies Prop.I.6, converse of isosceles triangle principle.

Prop.I.20.(triangle inequality): In every triangle, any two sides together are greater than the third side.
Proof: The previous proposition, may suggest trying to make a new triangle in which one side equals two sides of the original triangle, and then look at the angles opposite. So extend side BA to X, so that BX = BA + CA, i.e. AX = CA, and look at triangle BXC. Since sides AX = CA, angle BXC = <AXC equals angle <ACX, which is inside angle <BCX. Hence the side opposite <BCX is larger, i.e. BX = BA + AC > BC. QED.

Existence of triangles:(a construction)
The opposite question, of whether any three segments for which this is true can actually be sides of a constructed triangle is more subtle, and belongs to the study of what constructions are possible.

These seem to me like the most important congruence results. [Hint: learn them.]
Summary: We have proved congruence theorems SAS, SSS, ASA, and AAS. We have also proved the isosceles triangle principle, greater angle - greater side, equality of vertical angles, exterior angle theorem, and triangle inequality.

We have had to assume as new postulates: existence of rigid motions, there is only one segment joining two points, every line separates the plane into two sides. However, we have not needed to use postulate #5 so far. That changes tomorrow.

Day Four: Parallel lines

Two lines are parallel if they do not meet, no matter how far extended. In my high school class we assumed that through a point off a line there is one and only one line parallel to the given line. This is not the same as Euclid's "parallel postulate" #5. It will turn out that his postulate #5 gives a rule for two lines not to be parallel, while the exterior angle theorem gives a way to show that some lines are parallel. Here is another place where Euclid could do more with less, than some modern geometry books.

First we see how Euclid proved that parallel lines exist, with only the first 4 postulates (and our extra ones), and the propositions we have proved up to now.

[We don't seem to need the next proposition for this proof.
Prop. I.17: In any triangle, any two angles together are less than a "straight angle".
Proof: If we add to any angle of a triangle the corresponding exterior angle, we get a straight angle. The exterior angle theorem says that each of the other two angles of the triangle is less than the exterior angle. Thus adding either of them to the original angle gives less than a straight angle. QED.]

Prop.I.27: (alternate interior angle theorem, AIAT). If two lines cross a common third line, and two alternate interior angles they make are equal, then the two original lines are parallel.
Proof: (By contradiction.) Suppose they are not parallel. Then they meet say at X. Let the points where the third line meets these two lines be A and B, and consider triangle ABX. I claim it violates Prop. I.16, because one of the alternate interior angles is an exterior angle for this triangle, while the other equal one is a remote interior angle. Hence they cannot be equal by Prop.I.16. This contradiction proves the theorem. (What do you think?) QED.

Prop.I.28: If two lines cross a common third line and the interior angles they make on the same of that third line add to a straight angle, then the first two lines are parallel.
Proof: This implies that two alternate interior angles are equal so we are done by the previous result. (Do you agree?) QED.

Exercise: Two lines perpendicular to the same line are parallel to each other.

[Now we need to be able to copy angles, or construct perpendiculars.]
Prop.I.31: Through a point P off a line L, there passes at least one line parallel to L.

Proof: We know how to copy triangles, hence also angles, using SSS. So choose any point Q on L and draw line PQ. Then through point P construct a line M making an angle with PQ equal to the alternate interior angle made by PQ and L. Then M must be parallel to L. **QED.**

second construction:
Just drop a perpendicular K from P to L. Then erect a perpendicular M to K through P. Then the angles formed by L and M on the same side of K are both right, hence add to a straight angle. **QED.**

Up to now, Euclid has never used postulate #5, but the next familiar result requires it.
Exercise:Prop.I.29 Prove the Z principle (converse of the AIAT), using postulate #5. I.e. if L and M are parallel, and K meets both of them, then any two alternate interior angles are equal. (**Hint**: Show two interior angles on the same side of K add to a straight angle. Why does that help?)

Cor: Through a point P off a line L there is exactly one line parallel to L.
proof: Drop a perpendicular K from P to L. We already know that the line perpendicular to K at P is parallel to L. We will show that no other line through P is parallel to L. If M is any line through P not perpendicular to K, then on one side or the other M makes an angle less than a right angle. On that side, the two interior angles K makes with L and M add to less than a straight angle, so by postulate #5, the lines L and M meet on that side. Hence M is not parallel to L. **QED.**

Now at last we can prove a famous result of Euclidean geometry that would not be true without postulate #5.
Prop.I.32 **(i)** The angles of any triangle ABC add up to a straight angle, and
(ii) an exterior angle of a triangle is equal to the sum of the remote interior angles.
Proof: (i) Pass a line L through the vertex A parallel to the base BC. Then by the converse of the Z principle, the two angles adjacent to angle BAC are equal to the angles <ABC and <ACB. Since the three angles at the vertex A do add to a straight angle, so do the interior angles of triangle ABC. **QED.**
(ii) An exterior angle adds up to a straight angle with the adjacent angle, as do also the two remote interior angles, so the exterior angle equals those two angles. **QED.**

Remark: Although this is more precise than the earlier exterior angle theorem, we used that theorem to prove the existence of a parallel line. So books that do not present the earlier exterior angle theorem cannot prove the current result without assuming both the existence and uniqueness of parallel lines.

Day five: Parallelograms

We have studied triangles, i.e. three sided polygons, so next we look at some important four sided polygons.

(Parallelogram is another great spelling word.)

Definition: A parallelogram is a four sided plane figure such that each side is parallel to the opposite side. (For instance if the vertices are ABCD in clockwise order, then AB is parallel to CD and BC is parallel to DA.)

There are several other ways to recognize a parallelogram.
Prop.I.33 If a quadrilateral has one pair of opposite sides are both equal and parallel, then it is a parallelogram.
Proof: Draw a diagonal and use SAS and Z principle (converse of AIAT), to show the two triangles are congruent. Then AIAT shows the other two opposite sides are parallel. **QED.**

Prop.I.34 In a parallelogram all opposite sides and all opposite angles are equal, and the diagonal bisects the parallelogram into two congruent triangles.
proof: The diagonal bisects it by Z principle and ASA congruence, hence the opposite sides (and angles again) are equal. QED.

Exercise: A plane quadrilateral is a parallelogram if and only if one of these is true:
i) opposite sides are parallel
ii) opposite sides are equal
iii) opposite angles are equal
iv) the diagonals bisect each other
v) one pair of opposite sides are both equal and parallel

Day six: Area, or “content” of plane figures,

Next Euclid introduces a concept of plane figures as being “equal” to one another rather than congruent. When we say equal we usually mean two things are the same thing, so for us the word equal is even stronger than congruent. But for Euclid it is weaker. When he wants to say two figures are actually the same figure he says they “coincide”.

He has spoken before about two segments together being “equal” to another one, and of two angles together being “equal” to another one. In those cases the figures were not congruent, but merely had the same size in some sense. Here again, when Euclid says two plane figures are equal, he means they have the same size, or as we would say, the same area.

But Euclid is not using numbers to measure size, so he cannot assign a numerical area to a figure. So how does he know when two figures are “equal”, or have “equal content”, as we say today? We can find out by reading his proofs that two figures are “equal”.

Prop.I.43. If we choose a point on the diagonal of a parallelogram and pass two lines through it parallel to the two sides, dividing the original parallelogram into four smaller ones, then the two “off diagonal” parallelograms are equal.
Proof: These figures result by subtracting congruent triangles from triangles that are also congruent, all by Prop. I.34. Hence they are “equal” (in content). **QED**.

Thus Euclid considers congruent figures to be equal, and “differences” of congruent figures also to be equal. This makes sense if equality is in the sense of size or area.

Here is the fundamental way Euclid determines two figures have the same content.
Prop. I.35. Two parallelograms on the same base and in the same parallels, are equal.

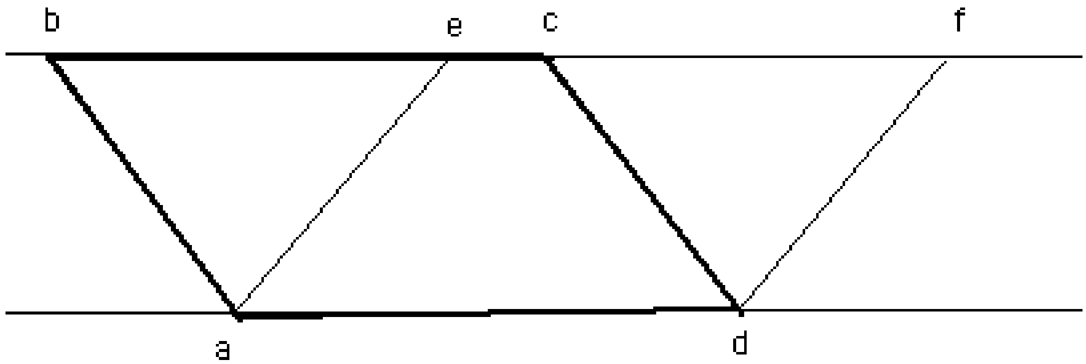
[Note: we would say the two parallelograms have the same height, but Euclid cannot say this since height is a number. See how clever he is at giving a purely geometric way of saying two parallelograms have the same height.]

Proof: He proves this by cutting up one figure into pieces that fit together to form the other figure. Or rather he does this after adding the same piece to both figures. So he is saying that two figures are equal if they can be decomposed into pieces which are congruent to each other, or if we can add congruent pieces to both of them so that afterwards the resulting enlarged figures can be decomposed into congruent pieces.

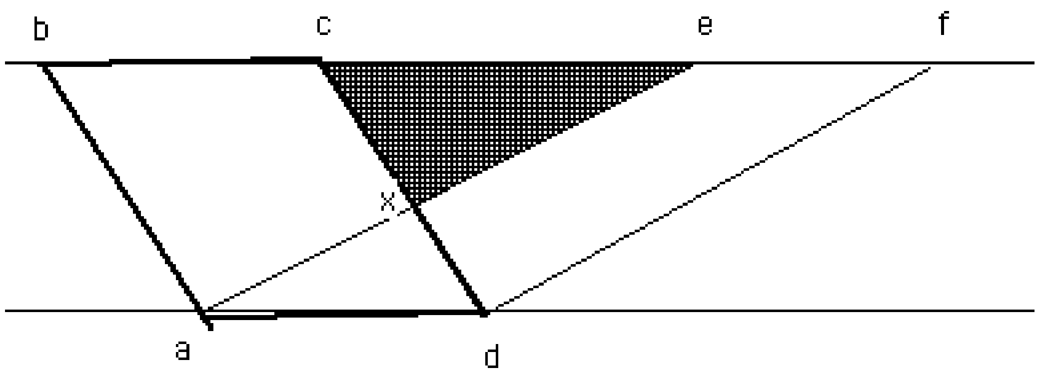
More simply, congruent figures are equal, sums of equal figures are equal, and differences of equal figures are equal.

His proof is quite nice and interesting to me. After reading it I realized I have never explained this correctly in my calculus classes. There are two cases and I only did the easier case.

When two parallelograms have the same base it is easier to cut one into pieces and reassemble it to form the other (on the same base) if the left top vertex of one is between the top vertices of the other.



Euclid does the harder case where both vertices of one parallelogram are entirely to the right of both vertices of the other. In this case it is not so easy to cut one into pieces that can be arranged to form the other, so he uses subtraction.



I.e. here triangles abe and dcf are congruent by SAS (do you see why?). Then if we add in triangle axd to both, and subtract triangle cxe from both, we get the two parallelograms. Hence those have equal content, or area.

Problem: In fact even in this harder case, you can decompose one parallelogram into pieces that form the other without subtraction. Can you figure out how? [For this solution you may assume Archimedes postulate, that any segment can be copied several times end to end, until it reaches any point on the extended line through it.]

Prop. I.36. Assume the bases are only equal (in length) rather than the same, and on the same line, and both vertices are on the same parallel line. Prove they are equal.
Proof: How do you suggest we reduce to the previous case? What clever thing does Euclid do to avoid using a rigid motion? What property of “equality” of figures does he use here? (transitivity) Can you justify that property for the relation of equal content? In particular, is “equal content” an equivalence relation?

Next we get an interesting argument for triangles that seems to be missing a step.
Prop. I. 37. Two triangles on the same base, and in the same parallels, are equal.

Proof: This time Euclid does not decompose one triangle into pieces that form the other, nor does he even do it using subtraction. Rather he doubles the triangles, to form two parallelograms that are equal. Assuming halves of equals are equal (this is the missing common notion), I.37 seems to follow from I.36. **QED.**

Discussion: Euclid has not said in common notions that halves of equals are equal. Does this follow from his other notions? If two figures are not equal, does one have to be smaller than the other? Why? If half of figure A were not just smaller than half of B, but equal to a strictly smaller part of half of figure B, would it follow that A is smaller than B? If we knew that, then we could argue that if A is equal to B then half of A could not be equal to a part of half of B, nor vice versa, so the halves must be equal. But it seems this goes beyond what has been stated about equality of figures.

I.e. even if two figures are made of the same congruent pieces, cutting each figure in half does not always cut the pieces in half, so it is not clear how one can decompose the halves also into congruent pieces. Indeed it seems easier to argue that adding equals to equals gives equals, so why did Euclid postulate (as common notion) the easier fact rather than the harder one?

An alternative argument
Hartshorne's book contains an argument (also occurring in Hilbert), reducing the case of triangles to the case of parallelograms, but without assuming halves of equal figures are equal. Rather he bisects one side of one triangle, and reassembles the pieces to make a parallelogram equal to the original triangle, not one twice as large. Then he must prove, that the same line through the midpoint of that side, and parallel to the base, also bisects the side of the other triangle. This is the first case of the theory of similar triangles, and it took me a bit of work.

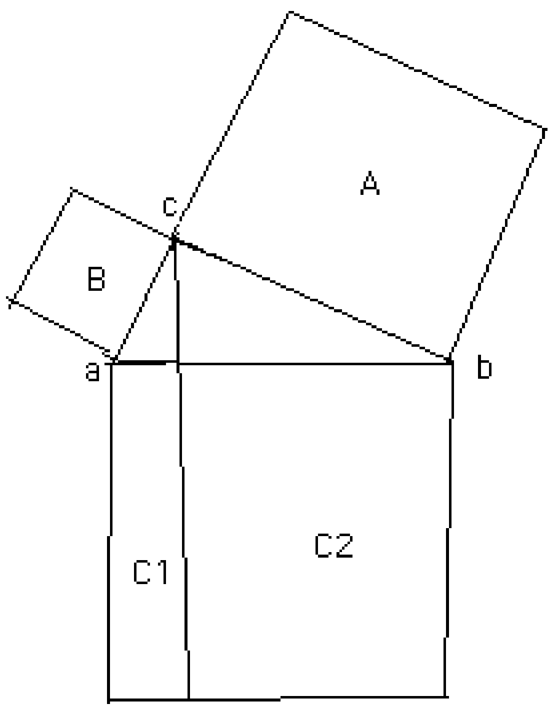
The moral is the theory of similarity always makes questions on areas easier, but that theory takes a lot of work to develop. Thus books that use similarity to make area theorems like Pythagoras look easy, are often hiding the difficulties of similarity theory. Indeed some books define area of triangles as ½ base times height, use that to establish the theory of area, and then later use the theory of area to prove the foundational facts of similarity theory.

The problem here is that a triangle has three different bases and three different heights, and the most natural way to prove that all three choices give the same area, is to use the theory of similarity. So these books are hiding the need for similarity in their development of area, and then using area to justify similarity theory. This is called "circular reasoning".

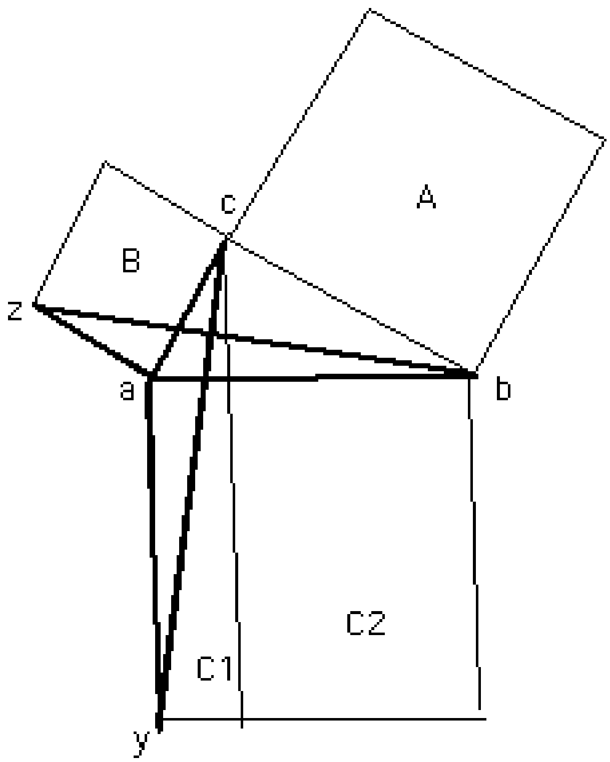
The whole point: For both parallelograms and triangles, area depends only on base and height.

Pythagoras' theorem
Now we want to prove the most famous theorem in geometry, using Euclid's own proof. There are many other proofs, one using similar triangles that is very algebraic, and others using decompositions. Euclid does not have the theory of similar triangles at this point, so he uses his theory of decomposing into triangles with the same base and in the same parallels. I had never seen this proof until a few years ago. An animation of it is available on the web at "Cut the knot".

Prop. I.47 If squares are constructed on the sides of a right triangle, the square on the side opposite the right angle is equal to the other two squares taken together.
Proof: We will give the proof in pictures. If the right angle is at c, drop the perpendicular as below, dividing the square on the hypotenuse into two rectangles, C1 and C2. Euclid will show that square B equals rectangle C1 (in area), and square A equals rectangle C2 (in area).



To do so, he will show this is true of triangles formed by bisecting the respective rectangles and squares.



The triangles <zab> and <cay> are congruent by SAS, since angles <zab and <cay both equal angle <cab plus a right angle, and the sides (ac) and (az) are sides of the same square, as are sides (ay) and (ab).

Now by Prop. I.37, triangle <zab> equals triangle <zac> in content, hence it equals half of square B. Similarly triangle <cay> is equal to half of rectangle C1.

Since doubles of figures with equal content also have equal content, C1 has equal with square B. Similarly rectangle C2 has equal content with square A. Adding, square C = C1+C2 has equal content with the sum of squares A+B. **QED.**

Week Two: Lines and angles in circles, geometric algebra, construction of a regular pentagon, and Archimedes' results on volume

This week we will discuss some topics from Books II, III, IV, and XII of Euclid. In Book II Euclid does elementary algebra purely geometrically. We will use line segments in place of numbers, and we will define multiplication and division using segments. The product of two segments is defined to be the rectangle having those segments as sides, or any other rectangle equal to it in size (i.e. with same content or area). We will learn to take square roots and solve quadratic equations geometrically, at least those equations with positive real roots.

One of our goals is to learn how Euclid constructed a regular pentagon (5 sided polygon) in a given circle, as we have done with an equilateral triangle (3 sides), a square (4 sides), and a regular hexagon (6 sides). Euclid first shows (Prop.II.11) how to solve the quadratic equation $X^2 + RX - R^2 = 0$, and then he shows how to use the solution to construct a pentagon. He proves that the length X that solves this equation is the side of a regular decagon in a circle of radius R. Then connecting every other point of the decagon gives a regular pentagon.

[We showed last Friday that the positive solution to the quadratic equation above, $X^2 + RX - R^2 = 0$, can be found by completing the square to get $(X+R/2)^2 = R^2 + (R/2)^2$. Then Euclid uses Pythagoras to find the solution of this equation. I.e. he constructs a right triangle with sides R and R/2, whose hypotenuse is thus X+R/2. Then to get X we just subtract R/2 from the hypotenuse. This solution process is Euclid's Prop. II.11. This shows how to construct the solution of this equation, but we still have to show that solution is really the side of a regular decagon in the circle of radius R. The proof of that is Prop. IV.10. The proof is usually given in modern books using trigonometry, but I like Euclid's beautiful geometric argument. **(Try to include a picture of the pentagon construction.)**

This is a long story, and we have a lot to learn before we get there. So we take it one step at a time. First we study how lines and circles meet, then the angles lines make in circles, and the area of rectangles associated to lines meeting in circles. Finally this is applied to the theory of similarity, and to construct an isosceles triangle whose vertex angle is 1/5 of a straight angle. (This is the triangle made by joining two consecutive vertices of a regular decagon in a circle, to the center of the circle.)

Week Two. Day 1) How lines and circles meet

Tangent lines to circles.

We all probably know that in Euclidean geometry a line meets a circle at most twice, and meets it just once if and only if it is tangent to the circle. Indeed it is almost universal today to define a line as tangent to a circle at a point P if and only if the line meets the circle at P and nowhere else, and even to claim that this definition is Euclid's own. After reading Euclid, it seems this usual definition differs from the one Euclid gave, although the two may be proved to be logically equivalent. We will discuss Euclid's definition and prove it is equivalent to the more commonly used one.

Euclid's Definition: A line is tangent to (or "touches") a circle at a point P, if the line meets the circle at P, but if produced further, does not "cut" the circle.
[Definition 2, Book III, page 51 of our Green Lion text.]

As often happens with Euclid's definitions, the problem here is in understanding what it means. I.e. what does "cut" mean? In my father's dictionary from 1936, it says one meaning of cut is to separate or divide into parts, also to cross, as when "lines cut one another". It also mentions examples such as "cut a corner" or take a shortcut, in all of which cases one does go inside a region by crossing a boundary. Thus it is plausible that Euclid meant the following as his definition of tangent line to a circle.

Definition: A line is tangent to (or "touches") a circle at a point P, if the line meets the circle at P, but does not cross from outside to inside the circle near that point.

Since there are several competing properties here, we will give them numbers. Suppose a line L meets a circle at P. Here are two statements that may or may not be true:

1. The line L does not meet the circle again, i.e. the line L meets the circle only once.
2. The line L does not cross the circle at P, from outside to inside (or vice versa).

Remark: Property 1 is a "global" property of the line and the circle, because it looks at every point where they meet. Property 2 is a "local" property, because it looks only at the way they meet near the point P.

Euclid immediately proves there is a connection between tangency and the total number of times a line meets a circle, i.e. between properties 1 and 2, but without doing or saying so completely explicitly.

Exercise:
a. By using the exterior angle theorem and the larger side / larger angle theorem, prove that if A is a point off a line M, then the shortest segment joining A to M, is the perpendicular from the point A to the line M.
If X is the point where the perpendicular from A meets the line M, then X is called the foot of the perpendicular.
b. Prove that as we move away from the foot of the perpendicular, the segments joining the line M to the point A get longer. I.e. if Y,Z are also points of M, and if Y is between X and Z, then AZ is longer than AY, and AY is longer than AX.
c. In particular if two points of M are the same distance from A, then the two points are on opposite sides of the foot of the perpendicular from A to M.

Here is Euclid's result.
Prop.III.2. If a line L meets a circle in two distinct points P,Q, then every point of L between P and Q lies inside the circle.
Proof: If O is the center of the circle, then since OP and OQ are the same length, the perpendicular from the center O to the segment PQ must meet the segment PQ at some

point X which lies nearer to O than either point P or Q. Thus X is inside the circle. By the large angle/larger side theorem, (or the Pythagorean theorem) as in the exercise above, the distance from O to a point of the line through PQ increases as the point moves away from X. Hence all points between X and P and between X and Q lie inside the circle, while points of L which lie outside segment PQ, also lie outside the circle. **QED.**

Cor: A line cannot meet a circle at more than two points.
Proof: If a line were to meet a circle at three distinct consecutive points P, X, Q, then point X would violate the previous result. I.e. since P and Q are on the circle, the point X must be inside the circle and not on it. Hence there cannot be more than two points where a line meets a circle. **QED.**

Note: Although Euclid did not say so, the proof of III.2 shows that points not in the interval PQ, lie wholly outside the circle.

Cor: A line that meets a circle more than once must cross the circle from outside to inside.
Proof: By the previous corollary a line that meets a circle more than once must meet it exactly twice. Then by the proof of III.2, the line crosses the circle at both points where it meets it. **QED.**

Cor: If a line meets a circle but does not cross it, then it meets the circle only once.
Proof: This is the contrapositive of the previous corollary, hence is also true. QED.

This shows that statement 2 above (Euclid's definition of tangent) implies statement 1 (the usual definition of tangent). To complete the proof of equivalence of these statements, and show that Euclid's definition of tangent line is equivalent to the usual one, we prove the inverse of the previous corollary.

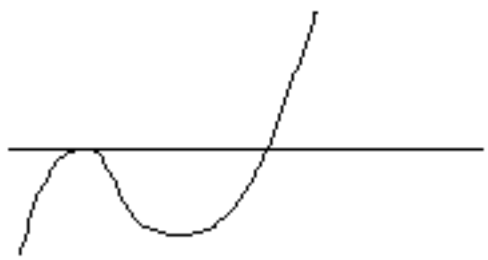
Lemma: If a line L meets a circle at P, and L crosses from the outside to the inside of the circle, then L also meets the circle again.
Proof: If the line L passes through the center O of the circle, then the point Q of L, on the other side of O, with OP = OQ, is also on the circle, so L does meet the circle twice.

Now assume L does not contain the center of the circle. If L meets the circle at P and crosses into the circle, there is a point say X of the line which lies inside the circle. That point X then is closer to the center O of the circle than is P, so P is not the foot of the perpendicular to L from O. Thus if we drop a perpendicular from O to L, meeting it at Y, then OY < OP, so Y is inside the circle.

Look at the right triangle OYP. If we flip this triangle about the line OY, or equivalently if we copy the segment YP onto L, on the other side of Y, we get segment YQ, for some point Q of L. Now triangles OYP and OYQ are congruent right triangles by SAS, so OQ equals OP. Since P is on the circle, and Q is the same distance from O as P is, Q is also on the circle, and we have proved that L meets the circle again. **QED.**

We have proved that if statement 2 above is false then statement 1 is also false, so by contraposition, we have proved that statement 1 implies statement 2. Since we had already proved that statement 2 implies statement 1, we have proved the two are equivalent. Thus it does not matter which definition you use for a tangent line, Euclid's original one, that the line meets but does not cross the circle, or the more common one that the line meets the circle only once.

Remark: Although the two definitions are equivalent for circles, Euclid's definition applies to curves much more general than circles, because Euclid's definition is "local", as a general definition of tangent must be. Look at the following picture of a curve that goes up then down, then up again.



The horizontal line meets the curve twice, hence is not tangent by the usual definition. But if we use Euclid's definition, we can see that the line is tangent at the left point, because it does not cross the curve there. It is not tangent at the second point because it does cross the curve there. Thus in this more general example, Euclid's definition still gives the right answer. Therefore if we want to use our knowledge of tangent lines in new situations, it is valuable to know Euclid's definition.

Remark: By googling "Euclid's definition of a tangent line" I found a reference in the "Century dictionary and cyclopedia", prepared by William Dwight Whitney, and Benjamin E. Smith, from 1911, where they give Euclid's definition as we have done. I.e. they say that Euclid defined a tangent as "a line that meeting a circle and not crossing it when produced". So at least 100 years ago there were people who interpreted Euclid's definition the same as we have done.

Whitney and Smith also point out that this definition fails only at points of "inflection" of a curve, [a point where it changes curvature from convex to concave]. Can you see the inflection point in the picture above, where the tangent line does cross the curve? [Hint: the tangent line at that special point only meets the curve once.] For a pdf file on tangent lines according to Descartes, which works for curves defined by polynomials, see http://www.physicsforums.com/showthread.php?p=2961525, post #6. The picture of the curve above is also reproduced in that file, if it does not appear here.

Do tangent lines exist?
It is not yet clear that tangent lines to circles exist! I.e. it still is not clear, at least not logically, that there are any lines that meet a circle once, or that do not cross a given circle. But Dr.T. has shown us where to look for them. He has proved that if a line L meets a circle with center O at P, but L is not perpendicular to radius OP, then L must meet the circle again, hence L is not tangent to the circle. Let's repeat his proof here.

Lemma: if a line L meets a circle at P, but L is not perpendicular at P to OP, then L must meet the circle again.
Proof: If L is not perpendicular to OP, then the perpendicular to L from O meets L at some other point X. Now the same construction we made in the proof above, constructing a triangle OXQ congruent to triangle OXP, shows that L meets the circle again. **QED.**

Thus if L meets the circle at P but is not perpendicular to OP at P, then L is not tangent to the circle.

If we consider the contrapositive of his statement, we get this result.
Cor: If a line L is tangent at P to a circle with center O, then L is perpendicular to radius OP.

It follows that the only possible candidate for a tangent line at a point P on a circle, is the line perpendicular at P to radius OP. In fact our arguments already show this line is a tangent, as we explain next.

Lemma: If a line L meets a circle with center O at P, and L is perpendicular to the radius OP, then L does not cross the circle at P, hence L is tangent to the circle.
Proof: We prove the contrapositive. If L is not tangent to the circle then L crosses inside the circle at P, hence the foot of the perpendicular from O to L is inside the circle and not at P. I.e. if L is not tangent to OP then L is not perpendicular to OP. Thus the contrapositive also holds: if L is perpendicular to OP at P, then L is tangent to the circle at P. **QED.**

Euclid proves the same thing in his amazing Proposition III.16, which we break up into several statements.
Prop.III.16a. If P is a point on a circle with center O, and L is a line perpendicular to OP at P, then L remains (except for the point P) wholly outside the circle.
Proof: We know by the exercise above that if L is perpendicular to OP at P, then for every other point Q on L the segment OQ is longer than OP. Thus every other point Q of L is outside the circle. **QED.**

Cor ("Porism"): The line perpendicular to radius OP at P is tangent to the circle at P.

Remark: Euclid says this corollary is obvious ("manifest"), so unfortunately he does not give a proof. If he had done so, we could see more clearly which definition of tangent he was really using. However, since the line L perpendicular to OP at P has been shown to remain wholly outside the circle except for the one point P, it follows both that it meets the circle only once, and also that it does not cross inside. So both equivalent definitions are satisfied, and it is certainly true that L is tangent there to the circle.

The other part of Euclid's Prop. III.16 is to me even more remarkable, since it contains the germ of the modern definition of a tangent line as a "limit of secants". I did not know that this idea was in Euclid, 2,000 years before the invention of the calculus.

Prop. III.16b. If P is a point on a circle with center O, and if L is a line perpendicular to radius OP at P, then no other line M through P "can be interposed" between L and the circle. I.e. if M is any other line through P, then M meets the circle again at $Q \neq P$, and the arc between P and Q lies entirely inside the angle between L and M.
Proof: If $M \neq L$, then the perpendicular to M from O meets M at $X \neq P$, and X lies inside the circle. By the previous arguments, if Q is the point on M on the other side of X from P, with PX = XQ, then OP = OQ, so Q also lies on the circle. Then by III.2, the arc between P and Q lies inside the angle between L and M. **QED.**

Remark: This result says in modern language, that the line L perpendicular to OP at P, is a "limit of the secant lines PQ, as $Q \rightarrow P$". I.e. for every angle made by lines M and N, with vertex P and containing the line L, there is an arc centered at P, such that for every point Q in that arc, the secant PQ lies within the given angle.

[**Aside:** When you study calculus, you will learn this is exactly the "epsilon / delta" proof that L is the limit of the secant lines PQ, as $Q \rightarrow P$. I.e. for every epsilon $e>0$, there is a delta $\partial > 0$, such that whenever $|Q-P| < \partial$, then $|PQ - L| < e$.
In plainer language, whenever Q is a point near P (closer than ∂), the secant PQ is near the line L (closer than e).]

This definition of tangent line, as a limit of secant lines, is presumably the one given by Newton 2,000 years after Euclid, and still used today in calculus. I have seen it stated that Newton re - read Euclid just before giving this definition. This is a reminder that even the greatest geniuses have looked for ideas in the classic works of antiquity.

Euclid also discusses how circles can meet each other. Some of the results follow easily from the proofs of earlier propositions.

Exercise: Prove two circles can meet at most once on each side of their line of centers. (Hint: Use the first proof, i.e. Euclid's own proof, of SSS congruence.)

Deduce: Cor: Prop.III.10. A circle cannot cut another circle at more than 2 points.

Euclid also defines two circles to be tangent if they meet but do not “cut” one another, and proves:
Prop.III.12-13: Two circles meet at most twice, and if tangent they meet only once.

In particular it seems clear he is not defining tangency of circles as circles intersecting only once, since then he would not have proved they cannot be tangent twice. This argues to me that the word “cut” does not mean “meet again”, but means “cross”.

Week 2. Day 2: Polygons and angles in circles

Some constructions from Book IV we can make

We quickly surveyed several nice ruler and compass constructions we can make by erecting and dropping perpendiculars, and copying segments. Everyone should be sure he/she knows how to:

Inscribe a square in a given circle
Inscribe a circle in a given square
Circumscribe a square about a given circle
Circumscribe a circle about a given square
Same 4 constructions for circles and regular pentagons,
regular hexagons, and regular octagons.

Then we discussed circumscribing a circle about any triangle. This used the basic construction of a perpendicular bisector of a segment. I.e. recall we have proved in constructions class (see Props. I.9-I.10-I.11) the following result.

Lemma: If segment AB is the base of an isosceles triangle, then the perpendicular bisector of AB passes through the vertex of that triangle. Conversely, if ABC is a triangle whose vertex C lies on the perpendicular bisector of AB, then the triangle is isosceles.

Cor: (**Prop. III.1)** The center of a circle lies on the perpendicular bisector of every secant.
Proof: The center is equidistant from the endpoints of every secant, hence if B,C are points on a circle and P is the center, the triangle PBC is isosceles with base BC. Hence the perpendicular bisector of the secant BC passes through the center P. **QED.**

Cor: To construct the center of a circle, either find the perpendicular bisector of any secant, and then find the midpoint of the resulting diameter, or find the perpendicular bisectors of two non - parallel secants, and see where they meet.

In particular, if all three vertices of a triangle lie on a circle, then the center of the circle is at the intersection of the perpendicular bisectors of any two sides. Thus the vertices of a triangle all lie on a circle if and only if the three perpendicular bisectors of the sides all meet at one point. (Then we say they “concur”, or come together.)

To show the vertices of a triangle always lie on a circle, we will prove the next result.
Prop.III.5. The point where the perpendicular bisectors of two sides of a triangle intersect, is equidistant from all three sides.
Proof: Bisect sides AB and AC in triangle ABC perpendicularly, and assume the bisectors meet at P. Since P is on both bisectors, then by the Lemma above, both triangles PAB and PAC are isosceles. I.e. sides PA = PB, and sides PA = PC. Thus all three sides are equal PA = PB = PC, and P is equidistant from all three vertices A,B,C. Hence all three vertices A,B,C lie on the circle with center at P and radius = PA. **QED.**

Cor: In any triangle, the three perpendicular bisectors of the sides are concurrent, i.e. they all meet at one point.
Proof: In the previous construction we get a circle centered at P, and all three sides AB, AC, BC of the triangle are secants of that circle. Hence the perpendicular bisectors all pass through the center of the circle. **QED.**

Prop. III.4 Given any triangle ABC, one can inscribe a circle in it, i.e. there exists a circle tangent to all three sides of the triangle.
Proof: Imagine this is true. If the center of the triangle is P, and the circle is tangent to the three sides at the points X,Y,Z, then the three radii PX, PY, PZ are equal and perpendicular to the various sides. Assume X is opposite vertex A, Y is opposite B, and Z is opposite vertex C. Connect P up to points X,Y,Z, and also to vertices A,B,C. This forms 6 right triangles all with a vertex at P.

Consider the two triangles with vertex at C. They have a common side, and each has a radius as another side. Since they are right triangles, by Pythagoras all three sides are equal, and the triangles are congruent by SSS. It follows that the segment PC bisects the angle at C, i.e. angle <ACB. Similarly all three angles of the triangle are bisected by the three segments from P. I.e. PA bisects <BAC, and PB bisects <ABC.

It follows that IF our triangle is going to have an inscribed circle, then the center of that circle is a point common to all three of the angle bisectors of the triangle. So for the proposition to be true it is necessary for the three angle bisectors to have a common point. Let’s try and prove that always happens.

Start from triangle ABC and let the angle bisectors of angles A and B meet at P. It can be shown P must be inside the triangle, using our separation postulate #8. It also looks to me as if the perpendicular from P to any side, always meets that side between the two vertices, and not outside the triangle on the extended line through the vertices. This seems assumed in the picture on page 86 of the Green Lion edition of the Elements.

Label the feet X,Y,Z of the three perpendiculars so that X is opposite angle A, Y is opposite angle B, and Z is opposite angle C. Then draw segments PX, PY, PZ, and PA, PB, PC, forming 6 right triangles as before. Now by AAS, triangles PAY and PAZ are congruent, so that PY = PZ. By the same reason triangles PZB and PXB are congruent, so that PZ = PX. Hence all three perpendiculars are equal PX=PY=PZ. Thus the circle centered at P with radius PX, has PX, PY, PZ as radii, and is tangent to all three sides of the triangle. **QED.**

In particular we have proved the following result.
Cor: In any triangle, the three angle bisectors all meet at one point.

Remark: A segment joining a vertex of a triangle to the midpoint of the opposite side is called a "median". It is true that all three medians of a triangle meet at a common point, but the proof I know for Euclidean geometry uses the theory of similarity. In fact this theorem is true also in hyperbolic geometry, and therefore also in neutral geometry. Thus there should be a proof that does not use Euclid's fifth postulate, but I have never seen one. Maybe one of you will find one and show me. Some notes on concurrence theorems is on my web page, class notes, #9 f, http://www.math.uga.edu/~roy/, but not this one.

Angles in circles
Another result I did not remember from high school is that all angles subtending the same arc of a circle are equal, no matter where their vertices are on the circumference. One nice way to show a lot of things are equal, is to show they are all equal to one special case. The only special angle subtending a given arc of a circle is the one with its vertex at the center. Fortunately this one can be used to study the angles with their vertices on the circumference, by the next result.

Prop. III.20 The angle with vertex at the center of a circle is double that of any angle with vertex on the circumference, if the two angles have the same arc as base.
Proof: Take an angle less than a straight angle, as Euclid always does.

You will need to fill the details of these proofs. If you get stuck, look at Euclid's proofs.
Case one: One side of the angle with vertex on the circumference, passes through the center. Then the result follows from the isosceles triangle principle and the Euclidean exterior angle theorem. (Exterior angle equals sum of remote interior angles, Prop. I.32, in Euclid or I.32.ii in our notes of week one, day 4.)

Case two: If the angle with vertex on the circumference encloses the angle at the center, the result follows by addition of two instances of case one.
[Draw a line from the vertex on the circumference passing through the center and extend it to a diameter. Now you have two angles, both with one side through the

center, and the original angle is their sum. Case one applies to each of these separately, and then you add the results.]

Case three: If the center is outside the angle with vertex on the circumference, the result follows by subtraction of two instances of case one.
[Again draw the diameter through the vertex on the circumference. Once more you have two angles both with one side through the center. The original angle is the difference of these angles, so apply case one again to both angles and subtract.]
QED.

Cor. Prop.III.21. Two angles in the same arc of a circle, with vertices on the circumference, are equal.
Proof: They are both equal to ½ of the angle with vertex at the center and in the same arc, so they are equal to each other, if halves of equals are equal. **QED.**

Next we want to take the limiting case where the vertex on the circumference comes to coincide with one of the endpoints of the base arc. [If we were in a calculus course, we would not need to do this as a separate case. We would just observe that the angles vary continuously as we move the vertex, but they have been constant for every position of that vertex. Hence they must remain constant as the vertex approaches an endpoint of the arc. I.e. the limiting value of a constant function is that same constant.] Notice that as in calculus, and in Euclid's prop. III.16, in this case one of the sides of the angle, being a limit of secants, has become a tangent line.

Prop. III.32. An angle in the arc AB in a circle, with vertex on the circumference, is equal to the angle between the secant AB and the tangent at B to the given arc.
Proof: It suffices by Prop. III.20 to prove the angle the tangent makes with the secant is again equal to half the angle at the center in the arc AB. If the center is O, the triangle AOB is isosceles and has base angles t and t. Since the tangent line at B is perpendicular to the radius at B, the angle s between the tangent and the secant satisfies t+s = right, hence 2t+2s = straight angle. But the angle u at the center satisfies u+2t = straight, since u,t,t are the angles of triangle AOB. Thus u+2t = 2s+2t, so u = 2s.
QED.

Day 3) Geometric algebra and Law of Cosines

Euclid Book II:

Now we want to do some geometric algebra, adding subtracting and multiplying segments, instead of numbers. We want to do algebra just using geometry. So our letters X and A, B ... represent line segments rather than numbers. Then we want to add and multiply them, still without using numbers. So if A and B are segments, we define their sum by sticking them end to end, and we say two segments are equal if they have the same length, i.e. they are congruent.

We define the product AB of two segments to be a rectangle with base A and height B. We say two products are equal if the rectangles have the same content or area, in the sense we discussed last week. I.e. AB = CD if the rectangles AB and CD can be decomposed into congruent pieces, possibly after adding some congruent pieces to both.

Then we want to know the usual properties of multiplication are true. To see that AB = BA as we expect, i.e. that "commutativity" holds, notice the rectangles AB and BA are congruent, by a rotation, hence they have the same content.

The other property we want for our multiplication is "distributivity". I.e. multiplication "distributes" or "spreads out" over addition. Thus we want **A(B+C) = AB +AC.**

Notice there are two kinds of addition in that formula. B+C is the sum of two segments, where we just lay them end to end. Addition of two rectangles means considering them together as one figure, so we lay them end to end, or if they don't line up perfectly, we just make sure they don't overlap.

Thus distributivity, i.e. A(B+C) = AB + AC, says that the rectangle with one side A, and the other side B+C, can be decomposed into two rectangles, one AxB and one AxC. This is easy to see from a picture. This is Prop. II.1. Moreover it seems to me that Props II.2 and II.3 are special cases of Prop. I.1, and thus really do not need separate proofs.

We skipped this part of Book I, but it is shown in Props. I.42-45, how to start from any polygon and find a rectangle of the same content, and having one given side. Thus to compare the size of two rectangles (or any two polygons), you can change them into two other rectangles with one congruent side, and then check whether the other sides are also congruent. I.e. given rectangles AB and CD, you can always change them into equal rectangles of form XY and XZ. Then they are equal if and only if Y and Z are congruent segments.

So it is possible to decide by ruler and compass construction whether two rectangles have the same area, without using numbers to measure that area. Notice this is a version of the cancellation property. Since XY = XZ is equivalent to X(Y-Z) = 0, this says that a product is zero only if at least one of the factors is zero. This is another property we like to have for multiplication.

We will prove Prop. II.14 below which tells us how to construct a square equal to a given rectangle. This gives us another way to compare the area of two rectangles, since once they are made square, the larger one is the one with the larger side.

Now once we have commutativity and distributivity, it seems that all Euclid's propositions II.1 through II.10 follow just by logic from those two properties. You have to be a little careful, because all letters refer to segments, and segments are always of positive length, so you can only subtract a smaller segment from a larger one. This makes a small inconvenience when we are doing algebra, because we have not introduced negative segments.

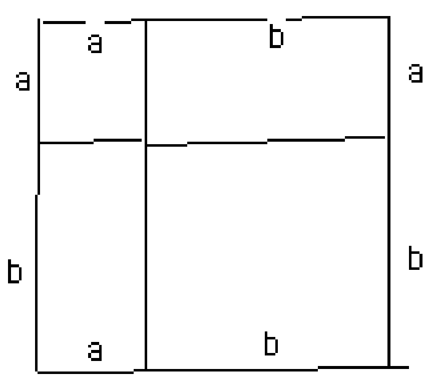
Still I feel there is basically nothing new in any of the first 10 propositions, they are just elaborate versions of AB = BA, and A(B+C) = AB + AC. It does help in some cases to look at the geometric proofs though, because some of them are so striking, that they make the algebra more obvious. This is the idea behind using aids like omnifix cubes in instruction. You can get your hands physically on the geometry more than on the algebra, and this makes it more memorable for many of us.

Propositions II.11 – II. 14 are more interesting, because they combine the algebraic rules developed in the earlier propositions with Pythagoras, to solve interesting quadratic equations. We have already discussed II.11, the equation whose solution gives the "golden ratio" used to construct a pentagon.

To be able to compute algebraically, and not have to draw the pictures every time, we want to learn a couple of useful formulas. You have seen these in algebra, and they also hold here, because they follow logically from the same rules that hold in algebra.

1) (Prop. II.4) (A+B)^2 = A^2 + 2AB + B^2.
This says the square with sides A+B, can be decomposed into four pieces, an A-square, a B-square, and two AxB rectangles.

proof: Look at this picture.



QED.

Remark: I was a senior in college when I first saw this picture, in a class with Professor Jerome Bruner, famous psychologist of learning. I was astonished it was so simple.

Proving it however is also easy without any pictures, just as a logical consequence of the two rules we have discussed. To see that, notice that commutativity applies to any product of segments. Thus $(\ ).[\{\ \} + <\ >] = (\ )\{\ \} + (\ )<\ >$, is true for any segments we put inside the various brackets. So we can put $(A+B)$, and $\{A\}$, and $<B>$. Then we have (without keeping up the fancy bracket shapes):$(A+B)(A+B) = (A+B)(A) + (A+B)(B)$, since we always expand on the right hand bracket.

Then by commutativity this equals $A(A+B) + B(A+B)$, and by distributivity again this equals $A^2 + AB + BA + B^2 = A^2 + AB +AB + B^2$, by commutativity again, and this equals $A^2 + 2AB + B^2$. **QED.**

Which proof is clearer to you, the algebraic one or the geometric one?

Here's another fact we will use a lot.

2)(Prop. II.5). $(X+Y)(X-Y) = X^2 -Y^2$.

This is not quite as obvious to me. Since we don't have negative segments, we cannot expand it algebraically as easily as the last one, because we don't know what (X) times $(–Y)$ means. Geometrically, Euclid's picture on page 40 of the Elements does not speak to me as eloquently as the picture for II.4 either.

However, it turns out to be algebraically the same as the previous one. If we assume Y is shorter than X, and call $A = Y$, and $B = X-Y$, then we have $A+B = X$.
Then $X^2 = (A+B)^2 = A^2 + 2AB + B^2 = Y^2 + 2 Y[X-Y] + [X-Y]^2$.

Now use distributivity to factor out $[X-Y]$ from the last two terms and get
$Y^2 + (2Y+[X-Y])[X-Y] = Y^2 + (X+Y)[X-Y]$.

Then since $X^2 = Y^2 + (X+Y)(X-Y)$, we get $X^2-Y^2 = (X+Y)(X-Y)$. **QED.**

To see it geometrically, make an X square and divide both sides into Y and $(X-Y)$. Then it is easy to see that within this X square we have an $X\text{x}(X-Y)$ rectangle and a $Y\text{x}(X-Y)$ rectangle, which together make an $(X+Y)\text{x}(X-Y)$ rectangle, that fills up the X square, except for a Y square. Thus $(X+Y)(X-Y) = X^2 –Y^2$.

Euclid's construction also shows that any two segments can be written as $X+Y$ and $X-Y$. I.e. he takes any segment and divides it into two pieces U,V in any way at all. Then he also divides the same original segment in half, and calls half of it X. Thus $U+V = 2X$. Then the longer piece, say U, is more than half, so $U = X+Y$, for some positive length segment Y. Now Y has been added onto X to make U, and Y has also been taken away from the other half X, so what remains is $V = X-Y$. Thus $UV = (X+Y)(X-Y)$.

This shows how to write any rectangle as a difference of two squares. It also shows how to write any odd integer as a difference of the squares of two integers. E.g. if we write 19 as 1(19), and take half of the sum of the factors 1 and 19, getting 10, we have $1 = 10-9$ and $19 = 10+9$, and thus $19 = 10^2 - 9^2 = 100-81$. (Why do we need the number to be odd?) Can you write 11 as a difference of two squares? Can you construct a right triangle with one side equal to sqrt(11)? What are the other side lengths? How about a right triangle with a leg = sqrt(39)? Or sqrt(105)? Or sqrt(171)?

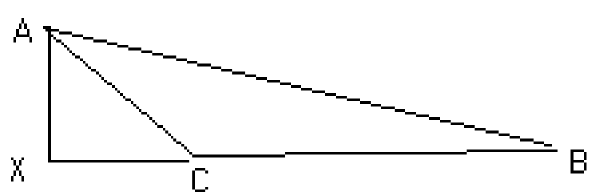
This ability to take square roots geometrically allows us to transform any rectangle into a square of equal area.
Prop. II.14. If A,B are any two segments, one can construct a segment X such that AB = X^2.
Proof: Euclid uses his trick of writing any product as a difference of two squares, and then applies Pythagoras. I.e. lay the segments A,B end to end to obtain one segment PQ of length A+B, divided at E, and bisect that segment at O. Let half of PQ be called R, and construct a semi circle of radius R on segment PQ.

Erect a perpendicular at E, the point dividing A from B on PQ, meeting the circumference of the semi circle at F, and call this perpendicular EF = X.
Denote the segment OE by C, so that A = PE = R-C, and B = EQ = R+C. Then AB = (R-)(R+C) = R^2 –C^2 = X^2, by Pythagoras applied to the right triangle OEX. **QED.**

Law of Cosines

Pythagoras says the square on the side opposite a right angle "equals" the two squares on the sides containing the angle. If the angle is acute, the square on the side opposite it is smaller than the two squares on the sides containing it, and if obtuse the square it is greater. The law of cosines tells exactly how much less or how much greater; in particular it says the discrepancy is twice the area of a certain rectangle. These are propositions 12-13, Book II of Euclid. He uses the geometric algebra we have discussed to do this. I never knew this was in Euclid either.

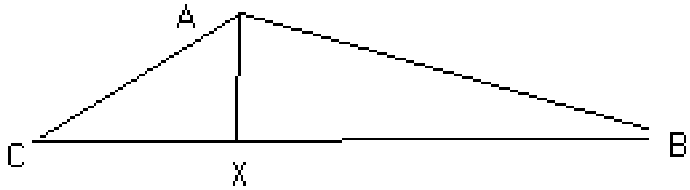
Prop. II.12 (Law of cosines, obtuse case): Let ABC be a triangle on the base BC, with obtuse angle at C, and vertex at A. Drop a perpendicular from A to the line extending base BC, meeting it at X, outside segment BC.
Then (AB)^2 = (AC)^2 + (BC)^2 + 2 (BC)(CX).



Proof: By Pythagoras applied to right triangle AXB, we have (AB)^2 = (AX)^2 + (BX)^2. From IV.4, this equals (AX)^2 + (CX)^2 + (BC)^2 + 2 (BC)(CX). By Pythagoras applied to triangle AXC, this equals (AC)^2 + (BC)^2 + 2 (BC)(CX). **QED.**

Exercise: Prove:
Prop. II.13: (Law of cosines, acute case): Let triangle ABC on base BC have an acute angle at C, and vertex A. Drop a perpendicular from A to base BC, and assume it

meets the base at X, between B and C.
Then $(AB)^2 = (AC)^2 + (BC)^2 - 2 (BC)(CX)$.



Remarks: What does this theorem have to do with cosines? If you recall the definition of the cosine of angle <C in the picture for the acute case above, $\cos(<C) = |XC|/|AC|$, the ratio of the numerical lengths of the two sides. Hence $\cos(<C).(AC) = XC$, an equality of segments. Substituting this into Euclid's formula above gives us $(AB)^2 = (AC)^2 + (BC)^2 - 2 (AC)(BC).\cos(<C)$, and this is the usual law of cosines in trigonometry. It also works for the obtuse case, since the cosine of an obtuse angle is negative, so the minus signs cancel and give us the formula in II.12 above.

Week two, Day 4) Theory of similar triangles

Now we are ready to prove the result that underlies our theory of similarity. This is sometimes called the "power of the point".

Constancy of products in secants,

Constancy of products for secants through a fixed point interior to a circle.

Prop.III.35. Given a point in a circle, if a secant be drawn through that point, the product of the two segments into which it is divided at that point is always the same.

Proof: It suffices to show every such product equals that obtained from the line through the center. So let P be the point and O be the center and assume $P \neq O$. (Exercise what if P=O?). Drop a perpendicular of length u from O to the secant dividing it at Q into two equal halves each of length s. If $|QP| = b$, then the secant is divided at P into segments of lengths $(s+b)$ and $(s-b)$.

On the other hand the diameter through P is bisected at O, and if $|OP| = a$, this diameter is divided at P into segments of lengths $(R+a)$ and $(R-a)$ where R = radius of the circle.

So we want to show that $(s+b)(s-b) = (R+a)(R-a)$. Looking at the right triangle OPQ, Pythagoras gives: $a^2 = u^2 + b^2$.

If the secant meets the circle at X and Y, then looking at right triangle OQY, Pythagoras gives us $R^2 = s^2 + u^2$.

Subtracting these gives $R^2 - a^2 = s^2-b^2$. i.e. $(R+a)(R-a) = (s+b)(s-b)$. **QED.**

Summary: In a circle of radius R and center O, if we choose a point P inside the circle, if A = segment OP, then any secant passing through the point P will be divided into two segments whose product is a rectangle equal in area to the difference of the two squares R^2-A^2.

If we were using numbers to measure lengths, then the product of the lengths of the two parts of a secant through is always equal to r^2-a^2, where r is the length of the radius and a is the distance of P from the center. This is the usual statement in books.

Application to similar triangles: We say a correspondence between two triangles is a "similarity" if corresponding angles are equal. If there is a similarity between them we call the triangles "similar".

If we draw two secants intersecting at a point P inside a circle, and connect up the 4 points A,B,X,Y, where the secants meet the circumference, we obtain two similar triangles: APX, and YPB, by Proposition III.20 on constancy of angles. For these triangles we have proved that (XP)(YP) = (AP)(BP), where this is the product of segments we have defined above. I.e. this equation means the rectangles formed by these segments have equal areas in the sense of congruent decompositions.

If we define the segment ratios (XP)/(AP) = (BP)/(YP) to be equal if and only if the rectangles (XP)(YP) = (AP)(BP) are equal, a perfectly natural definition by "cross multiplication", then we can say at least that those similar triangles formed by two intersecting secants in a circle, have corresponding sides in the same ratio.

If we were measuring side lengths and areas by numbers, then we have proved that $|XP|\,|YP| = |AP|\,|BP|$, where absolute values denotes length of a segment. If we divide by $|YP|.|AP|$, this says that $|XP|/|AP| = |BP|/|YP|$, and we get the usual numerical similarity relation of equal ratios for similar triangles.

We claim that all similar triangles can be formed by two intersecting secants in some circle. In fact if we have any two similar triangles, in the sense of having corresponding angles equal, and we arrange them so that two equal angles form vertical angles, and the other two pairs of non corresponding sides are arranged as collinear, then the circle that contains three of the vertices, will also contain the fourth vertex, or else the principle of constancy of angles would be violated. Thus all similar triangles occur in a circle like this, and satisfy the similarity relation.

In particular we can develop the theory of similarity of triangles without numbers.

[This theory of similar triangles differs from that in Euclid. I was trying to present similarity without spending as much time as Euclid does to prove Prop. VI.2, since our course is so short. Then I noticed one can derive it from Prop. III.35. So this theory was created especially for this epsilon camp class. This theory is also more general than Euclid's because it does not depend on approximating ratios by rational ratios. Indeed Euclid's theory requires another missing axiom called Archimedes axiom, which says that given three points A,B,C on a line, we can lay off copies of segment AB end to end, until some copy goes past the point C. This epsilon camp theory of similarity, unlike Euclid's own theory, works without that axiom, and proves similarity in non Archimedean Euclidean geometries.]

Remark: Now that we have similarity we could go backwards and use it to prove Pythagoras. Many people know this proof, which may have preceded Euclid's.

Similarity proof of Pythagoras
I.e. if we have a right triangle ABC with sides a,b,c, and drop a perpendicular from vertex C to side AB dividing it at E into pieces of lengths x and y, we have three similar triangles ABC, ACE, and CBE. If $|AE| = x$, and $|BE| = y$, then $x+y = c$, and $c/a = a/y$, and $c/b = b/x$. Thus $a^2 = cy$ and $b^2 = cx$, so $a^2 + b^2 = cx + cy = c(x+y) = c^2$. **QED.**

Of course this reasoning is circular, since we have used Pythagoras to prove the theory of similarity. There is also a very easy, but equally circular, argument for Prop. III.35 using similarity.

Exercise: Assume the theory of similarity for triangles and give a quick proof of III.35, the constancy of products, or the "power of the point".

Exercise: Give an easy (and non circular) proof of Prop. II.14, by showing it is a special case of Prop. III.35.

Exercise: Prove a generalized Pythagorean theorem, in which the polygons on the sides of the triangle are equilateral triangles. See Prop. VI.31 for a still more general version, where the figures on the sides are any similar figures at all.

Remarks on circularity of reasoning: Circularity occurs when something is used to prove another thing, and then the opposite is also done, with neither of them being rather proved or assumed beforehand. Circularity can thus be avoided in two ways: either prove one of them legitimately from principles already known, or simply assume one of them as a postulate.

For example, in Birkhoff's system he takes a strong version of similarity as a postulate right at the beginning. This makes his system of geometry very efficient, but he has given himself a very powerful tool from the start, which may not be very intuitive to some of us. In Harold Jacobs' book on the other hand, he assumes the theory of area as a postulate. Then he gives Euclid's proof of similarity using area. Thus in both cases there is no circularity of reasoning. Hilbert and Hartshorne, in their books listed under suggested reading on the epsilon camp student forum, develop theories of segment addition and multiplication different from ours, then base the theory of similarity on their segment arithmetic, and finally base the theory of area on their similarity theories. Therefore these approaches are also not circular, and moreover they use very few axioms.

It is only when we try to prove area makes sense using similarity, and then try later to prove similarity makes sense using area, that circularity occurs. If we use area to prove similarity makes sense without either assuming area as a postulate or proving the needed properties of area, perhaps it is more accurate to say a logical "gap" has occurred, rather than circular reasoning.

The moral is, to be clear and complete, one should always either say what one is assuming, or else show it does not need to be assumed by proving it from earlier principles which have been stated.

These remarks are directed at someone who wants to understand the logical structure of the subject and to appreciate the attempts by mathematicians beginning with Euclid, to render this structure flawless and beautiful. Preparing to compete in exams is different. Then we want the quickest way to solve every problem. In contests we usually take all of mathematics essentially as postulates, and try to use them to solve problems we have not seen before. Therefore a very clever person can do well in a contest simply by knowing the statements of theorems. A mathematician however always wants to know the proofs. You are potential mathematicians, so these notes are written for you from that perspective. It is my opinion however that they will also help you in contests, by deepening your understanding of what to use in various situations.

Exercise: Give an easy proof of Prop. II.14, by showing it is a special case of Prop. III.35.

Day 5: Euclid's proof for the construction of a regular pentagon

As we learned yesterday in Dr.T's class, the only primes p for which one can construct regular p - gons, are the Fermat primes, i.e. those primes of form 2^(2^n) + 1. Moreover there are only 5 such primes known, F0 = 3, F1 = 5, F2 = 17, F3 = 257, and F4 = 65,537. Among these, so far we only know how to construct a regular 3-gon, or equilateral triangle. We have stated how to construct a regular 5 - gon or pentagon, but have not seen why it works. We do that today. At the end we will try to say something about why only Fermat prime p-gons are constructible.

Interestingly, the proof of Prop. IV.10 justifying the construction of a regular pentagon, uses variations of the same two results we used yesterday to prove Prop. VI.2 about similar triangles. Namely we again look at constancy of angles Prop.III.32, and constancy of products Prop. III.36-37, the “power of the point”. One difference today is we have similarity theory and we can we can use it to make some proofs easier. First we do the tangential case of constancy of angles.

Prop.III.32. Let A,B be two points on a circle defining an arc less than or equal to half a circle. If C is any point on the ray tangent to the circle at A and pointing along the arc AB, and P is any point outside the arc AB, then the angles <APB and <CAB are equal.
Proof: Exercise: Hint: by Prop. III.21, it suffices to prove it when AP passes through the center of the circle. Try this case yourself before reading Euclid’s proof.

Remark: If we think in terms of limits as they do in calculus, let the point P approach the point A as a limit. Then the secant PB approaches the secant AB, and the ray PA approaches the tangent ray AC at A. Since Prop.III.21 implies the angle between the rays PA and PB remains constant as P approaches A, it seems evident (by “continuity”) that the angle is still the same in the limit, i.e. that the angle between the secant AB and the ray AC is still the same angle. [Advice: This type of thinking, i.e. the idea of things changing continuously as they move, can sometimes help you guess the answer to hard problems.]

Next we look at constancy of products, or the power of the point, when the point is outside the circle.

Exercise (Prop. III.17): Show how to construct a right triangle, given the hypotenuse and one side. Use this method to construct a tangent to a circle from a given point outside the circle.

Prop. III. 36. Let there be given a circle of radius R and a point P outside the circle. Draw a segment from P to cut the circle at A and again at B. Then draw a segment from P meeting the circle tangentially at one point C. Then $(PA)(PB) = (PC)^2$, [in terms of equality of areas of rectangles as usual, via congruent decompositions].
Proof: This is pretty much like the argument yesterday for III.35 using Pythagoras as in Euclid, but it is even easier using similarity, so we use similarity.

Look at the triangles PBC and PCA, and note that they share the angle at P, and that angles <PBC and <PCA are equal by Prop. III.32. Therefore these triangles are similar, so corresponding sides are in the same ratios. Hence $(PA)/(PC) = (PC)(PB)$, and therefore $(PA)(PB) = (PC)^2$. **QED.**

Cor: If P is a point outside a circle and we draw two segments from P to the circle, the first meeting it at A and then again at B, and the second one meeting the circle at C and then at D, then we have $(PA)(PB) = (PC)(PD)$.

Proof: By III.36, both products equal $(PE)^2$, where PE is tangent to the circle at E, hence they are equal to each other. **QED.**

The equation in III.36 gives us a new way to recognize the tangent to a circle, because the converse is also true.

Prop. III.37. Let P be a point outside a circle, and draw a segment PB from P meeting the circle first at A and then again at B. Draw a second segment from P to a point C on the circle. If $(PA)(PB) = (PC)^2$ then segment PC is tangent to the circle at C.
Proof: If not and the segment PC meets the circle again at $D \neq C$, then the right side of this equation would equal $(PC)(PD) \neq (PC)^2$. Thus the segment meets the circle only once, at C, hence is tangent there. **QED.**

Remarks: To see how this formula for the products relates to the previous case when the point was inside the circle, look at the picture as the point P moves outside the circle, and you will see that the two portions of the secant change into the whole segment and the portion of it outside the circle. So this result is a "continuous extension" of the previous case.
The two formulas can be stated in a unified way as follows. Let R be the radius of the circle and let S be the distance from P to the center, and let A,B be the two points where the segment meets the circle. Then the two theorems Props.(III.35-36) together say that for every segment, the product $(PA)(PB)$ always equals $|R^2-S^2|$. Some people prefer to focus on the number $S^2 - R^2$, without the absolute value, since then you can tell whether the point P is outside or inside the circle according to whether this number is positive or negative.

Now we can give Euclid's proof that his regular pentagon construction works. It suffices to construct a regular decagon.
Prop.IV.11: Let X be a solution of the quadratic equation $X^2 = R(R-X)$. Then X is the side of regular decagon in a circle of radius R.
Proof: If we look at a regular decagon inscribed in a circle, and draw segments connecting every vertex to the center, we get 10 congruent isosceles triangles. If we could construct just one of these triangles, we could make our decagon, by placing them together.

Since there are 10 triangles whose central angles combine to make a full circle, each central angle must equal 1/5 of a straight angle. Since they are also isosceles the base angles are equal, and since the angles of a triangle add to a straight angle, the base angles must each be 2/5 of a straight angle. Therefore it suffices to prove the following statement.

Prop.IV.10: If $X <$ (less than) R are segments such that $X^2 = R(R-X)$, then in the isosceles triangle with base X and sides R, the base angles are both equal to twice the vertex angle.

Proof: Let the triangle be ABC with vertex at A and base BC.
Copy a segment AD on side AC, equal to the base BC = X. Draw the segment BD dividing the triangle into two triangles, ABD and BDC. Euclid will prove both of these triangles are isosceles. (Do you see why that will prove the proposition?)
Let the vertex angle at A be a, let angle <DBC = b, and let angle <DBA = c. Then by the Euclidean exterior angle theorem angle <BDC = a+c, and since triangle ABC is isosceles angle <BCA = b+c.

Next Euclid makes a brilliant argument that angle angles a and b are equal as follows. Draw a circle through the three points ABD. Then C is a point outside the circle and we have two segments from C meeting the circle. One of them is CA which meets it at D and A, and the other is segment CB which meets it at least at B. Euclid observes that CB is actually tangent to the circle, because the equation $X^2 = R(R-X)$ is equivalent to the product equation $(CD)(CA) = (CB)^2$. Isn't that amazing!?

Since by Prop.III.37 segment (CB) is tangent to the circle, it follows from Prop.III.32 that angles <CDB and <BAC are equal. That is, a = b. Hence triangle BDC is isosceles, and thus side BD = BC = X.

But also side AD = X by construction, so triangle ABD is also isosceles, hence angles <BAC = a, and <ABD = c are also equal. Hence a=b=c, and the original triangle ABC has both base angles equal to 2a, twice the vertex angle. **QED.**

Remarks: While giving this argument I noticed that the cleverest part, namely using the extra circle, and propositions III.32 and III.37 to show that angles a and b are equal, can be finessed by using similarity, which Euclid did not have available in Book IV. I could not bear to omit Euclid's beautiful tour de force argument above, but now I will give the easier similarity proof, because you are more likely to be able to use this idea again some time. I.e. Euclid's proof is more like a trick that you use only once, and this one is more like a method, that as the joke goes, you should be able to use at least twice. [Mathematician's joke: "A method is a trick you use twice."]

Prop.IV.10: If X < (less than) R are segments such that $X^2 = R(R-X)$, then in the isosceles triangle with base X and sides R, the base angles are both equal to twice the vertex angle.
Alternate proof: Again draw the same triangles and label the angles as before, a,b,c, and we want to prove that a = b = c. Oops, I guess we need a slight variation of the similarity principle we proved.

SAS similarity: If two triangles have one angle equal, and if the sides adjacent to that angle are in the same proportion, then the triangles are similar with sides corresponding in that order.
This is an easy corollary and converse to our version of Prop.VI.2, as can be seen by arranging the triangles as in the picture on page 125, and let's assume it.

[Proof sketch: If two triangles have one angle equal we can move them until they have that angle in common, as with triangles ADE and ABC in Euclid page 125. Then we have proved that if the other two angles are also equal, then the sides are in the same ratio, so $(AD)/(AB) = (AE)/(AC)$. Then look at the picture and imagine what happens if we move point E further down the side toward C. Not only does the angle <AED get smaller, but the ratio of sides $(AE)/(AC)$ gets larger. So if the angles <AED and <ACB are NOT equal, then those sides AE and AC are no longer in the same ratio as sides AD and AB. The converse of this says that the triangles have angle A in common, and if those two pairs of adjacent sides are in the same ratio, i.e. if $(AD)/(AB) = (AE)/(AC)$, then the two base angles are also equal and the triangles are similar. QED.]

OK let's use it to prove Euclid's proposition IV.10.

Look at a picture of our triangles in Prop.IV.10 (with our labeling, not Euclid's). I.e. let the triangle be ABC with vertex at A and base BC, and sides AB and AC both = R, and base BC = X, where $X^2 = R(R-X)$. Copy a segment AD on side AC, equal to the base BC = X. Draw the segment BD dividing the triangle into two triangles, ABD and BDC. Euclid will prove both of these triangles are isosceles. Label the angles as <BAC = a, angle <DBC = b, and angle <ABD = c. We claim a=b=c. First we will show a=b.

Notice that triangles ABC and BDC share an angle at C. Moreover the adjacent sides are in the ratios $(R-X)/(X)$ and (X/R). But these ratios are equal by hypothesis since $X^2 = R(R-X)$, so those triangles are similar in that order.

Hence angle b = <DBC = <BAC = a. Then as before, triangle DBC is isosceles so DB = X. Then ABD is isosceles since DB = AD = X, so a = c. Hence a = b = c, and we are done. I.e. triangle ABC has angles a, 2a, and 2a. Thus 5a = straight angle, so the vertex angle a = 1/5 of a straight angle. **QED.**

Remarks on impossible constructions:

I want to say something about why the only regular p-gons that can be constructed with prime p, are for the Fermat primes $p = 2^{(2^n)} + 1$, or equivalently $p – 1 = 2^{(2^n)}$. This discussion will necessarily be very abbreviated (but longer than the one in class). First of all, note that a number of form $2^{(nm)}+1$ where m is odd is never prime, because it equals $(2^n)^m + 1$, and any number of form x^m+1 where m is odd can be factored.

E.g., you probably know the basic example of $x^3+1 = (x+1)(x^2+x+1)$. Then $x^5 +1 = (x+1)(x^4+x^3+x^2+x+1)$, and so on…. Wait a minute, why am I going to so much trouble? Probably you know the factor theorem from algebra implies that a polynomial f(x) is divisible by x+1 if and only if f(-1) = 0. Since plugging x = -1 into $x^m + 1$ does give zero when m is odd, x+1 always divides x^m+1 for m odd. So $2^n +1$ always divides $(2^n)^m + 1$ when m is odd, by taking $x = 2^n$.

Thus if a prime has form 2^n + 1, then it has form 2^(2^n) + 1. Hence it suffices to show that every constructible prime must have form 2^n + 1.

Theorem: If p is a constructible prime, then it has form p = 2^n+1.
Description of proof: Notice that every construction involves finding points by intersecting lines and circles. This is the key to understanding what constructions are possible.

Assume we are in an Archimedean geometry, i.e. Euclid's geometry plus the extra axiom of Archimedes, that says any segment can be laid off repeatedly until it reaches any other point. Then we can introduce real numbers as coordinates, by introducing a pair of perpendicular axes in the plane. Then every point of the plane can be described by a pair of real coordinates (x,y). Conversely we may assume that every pair of real coordinates (x,y) represents a point of the plane, but they need not all be constructible. We want to examine which ones among the many points of the plane are constructible by ruler and compass.

We start from only two points (0,0) and (1,0), and ask what points can be constructed from these. E.g. we can lay off as many copies of these as we wish along the x axis, so we get all points of form (n,m) where n,m are integers. Then we can also subdivide the interval between (0,0) and (1,0) into n equal parts for every natural number n, and then lay off copies of those, so we also get all rational points on the x axis of form (n/m, 0) where n,m, are integers, and m ≠ 0.

Since we can construct the perpendicular to the x axis we also get the y axis, and then we can lay off rational points on the y axis. Now we can construct perpendiculars to both x and y axes and intersect them, so we also get all "rational points" of the plane, i.e. all points of form (n/m, a/b), with a,b,n,m, integers and mb ≠ 0. What else can we get?

Well we also get points that arise from intersecting lines determined by two rational points, with other such lines, or with circles with rational centers and a rational point on the circumference, or two such circles.

Now intersecting sets defined by equations means solving the equations simultaneously, so we want to know what kind of numbers occur as simultaneous solutions of equations of lines and circles. In fact all we need to know is the degree of the equations. Solving two linear equations with rational coefficients just gives rational solutions, so intersecting such rational lines does not give any new points. Moreover intersecting two rational circles gives two points which are also obtained by intersecting one circle with a rational line (subtract the equation for the two circles to get the equation of the line). So we look at points obtained by intersecting rational lines and rational circles.

An equation for a line through two rational points looks like ax+by = c, where a,b,c, are rational numbers. An equation for a circle determined by a rational center and rational point on its circumference looks like (x-u)^2 + (y-v)^2 = w^2, where u,v,w, are rational numbers. To solve ax+by = c, and (x-u)^2 + (y-v)^2 = w^2, simultaneously, we solve the linear equation for y, getting: by = c – ax, then y = c/b – ax/b, and substitute this for y in the quadratic equation. The result is some complicated quadratic equation in x, Maybe (x-u)^2 + ([c/b – ax/b] – v)^2 = w^2. I don't care what it is exactly, as I am only interested in its degree, or the fact that it is quadratic, i.e. of degree two. We call solutions of quadratic equations with rational coefficients "quadratic" numbers.

Conversely we know how to use Pythagoras to solve quadratic equations whose coefficients are segments, or numbers, that we have already constructed. So we can construct points in the plane whose coefficients are solutions of quadratic equations with rational coefficients, and any algebraic combination of those numbers. E.g. we can construct the point (sqrt(2)-sqrt(5), 1+sqrt(7/3)). Similarly we can construct all quadratic points, i.e. points whose coefficients are quadratic numbers. These quadratic points are the points that only require one "quadratic step" to construct, i.e. one use of the compass.

What next? Well once we have those quadratic guys we can intersect more lines and circles. So now we are solving quadratic equations whose coefficients are quadratic numbers. We call these biquadratic numbers. Thus we can construct all biquadratic points I claim the solution of such an equation is also the solution of an equation of degree 4, but with rational coefficients. I.e. we claim all biquadratic numbers are also quartic numbers, or degree 4 numbers.

E.g. if we have an equation like X^2 – sqrt(3).X + 1 = 0, with quadratic coefficients, we can rewrite it as X^2 + 1 = sqrt(3)X, and square both sides, to get (X^2+1)^2 = 3X^2, or X^4 + 2XC^2 + 1 = 3X^2, which becomes X^4 –X^2 + 1 = 0. Thus our number X becomes a solution of a 4^{th} degree equation with rational coefficients. I general every biquadratic number is a quartic number. The idea is that biquadratic numbers are those that only require two quadratic steps to construct, or two uses of the compass. As in this example, they are all quartic numbers, i.e. they satisfy degree 4 equations with rational coefficients. Now what about points that require three quadratic steps to construct?

If we have X^2 – 2^(1/4)X -3 = 0, where the coefficient 2^(1/4) is biquadratic hence quartic, we get X^2 -3 = 2^(1/4)X, and raising both sides to the 4^{th} power gives (X^2-3)^4 = 2X^4, which is an equation of degree 8 with rational coefficients. So the triquadratic number X, which requires three uses of the compass, is of degree 8, or an "octic" number.

I don’t know how to make this entirely clear, but as we go on what happens is that the points we get are solutions of equations of degree 2,4,8,16,32,….., i.e. degree 2^n, with rational coefficients. And that is all we can get. More precisely, a point that can be constructed in n steps satisfies an equation with rational coefficients, and has degree dividing 2^n, (since steps not using the compass have degree one). It follows that a point (x,y) cannot be constructed unless its coefficients are solutions of an equation of some degree 2^k with rational coefficients.

Now this applies also to complex numbers x+iy corresponding to our points. I.e, if (x,y) is a constructible point, then the complex number z = x+iy must satisfy an equation of degree 2^n for some n. But the first vertex on the unit circle of a regular p - gon, after the point (1,0), is exactly “1/p^ th” of the way around the circle. Since multiplying complex numbers adds their angles and multiplies their lengths, it is a complex solution of the equation z^p – 1 = 0.

Now this equation factors as z^p – 1 = (z-1)(z^(p-1) +….+z + 1) = 0, and since z=1 is the only solution of the first factor, the complex number we want is a solution of the second factor, which has degree p-1. For that point, which gives the first vertex of the regular p-gon, to be constructible, we must have p-1 = 2^n, i.e. p = 2^n+1. Then of course since p is prime, we have seen it must have form 2^(2^k) + 1, i.e. it must be a Fermat prime.

These ideas are usually taught in a college abstract algebra course as an application of linear algebra. One possible source is the book Abstract algebra, a geometric approach, by Theodore Shifrin, or my math 4000 notes #4f, the last couple lectures, on my web page at UGA. http://www.math.uga.edu/~roy/

Day six), Cavalieri principle, Volume of pyramid, cone, sphere, Surface area of sphere

To day I want to describe a progression from Euclid to Archimedes to Newton and Barrow, in the theory of area and volume.

Fundamental principle: same base and height implies same area or volume

Recall the basic result Prop. I.35 and I.37, about plane area for parallelograms and triangles that area depends only on base and height. This result was easier to prove for parallelograms than triangles by decomposition, and then Euclid used the fact that a triangle is half a parallelogram to deduce it for triangles. Recall also that to avoid a new axiom that halves of equals are equal, Hilbert proved a special case of the principle of similarity for triangles whose side lengths have been cut in half. Generalizing the theory of area to a theory of volume will also benefit from using similarity.

Prisms, parallelepipeds and pyramids
To understand three dimensional solids it is natural to proceed by analogy with the two dimensional case. The three dimensional analog of a parallelogram is a parallelepiped whose faces are all parallelograms, or more generally a "prism", a solid whose base and top are congruent polygons lying in parallel planes, and whose side faces are parallelograms. All slices of a prism parallel to the base give polygons congruent to the base and hence congruent to each other. There are as many different types of prisms as there are polygons. A prism with triangular base is a "triangular prism", and a prism with a rectangular base is a "rectangular prism", etc…. A "right prism" also has its side edges perpendicular to the base, and the sides are then rectangles.

The three dimensional analog of a triangle is a pyramid, a figure formed from a polygonal base by joining every point of the base to a single vertex point outside the plane of the base. The pyramid is triangular if the base is triangular and so on…

The basic result about volume is again that the volumes of prisms and pyramids depend only on the base and height of the solid. Euclid first develops the theory of volume for parallelepipedal prisms using similarity, and then deduces the theory for pyramids. If you browse Euclid's Books 10 and 11 you can spot uses of similarity by looking for references to Book V or VI in the margins of the proofs.

To pass from the case of parallelepipeds to pyramids Euclid abandons his two dimensional technique of finite congruent decompositions, and adopts a limiting process. Just as one can decompose a triangle into a parallelogram occupying half its area, plus two more triangles similar to the original one, by drawing two segments joining the midpoints of the sides, Euclid shows how to decompose a pyramid into two parallelepipeds occupying more than half the volume, plus two more pyramids similar to the original one, again by joining midpoints of various sides. This construction repeated, shows how to express the volume of a pyramid as a limit of volumes of parallelepipeds. Then the fact that volume depends only on base and height for parallelepipeds, implies the same result for pyramids.

It was unknown for 2,000 years whether this limiting process is necessary, and even Gauss thought about it unsuccessfully. Finally in 1900, in response David Hilbert's famous "problem lecture" emphasizing the importance of this question, Max Dehn proved that unlike the two dimensional case for polygons of equal area, in fact one cannot always decompose two three dimensional polyhedra of equal volume into a finite number of congruent pieces. So Euclid was again justified in his approach.

Archimedes
It is easier to apply similarity if we give ourselves another advantage and appeal to the ideas of a slightly later mathematician, the great Archimedes of Syracuse. Archimedes was an amazing man who earned enormous respect in his native city by inventing and constructing many devices for the benefit and protection of the town. In Plutarch's

"Lives", the story of the siege of Syracuse by Marcellus contains a fascinating account of Archimedes' role in the defense of the city from the Roman army and navy.

It is said that Archimedes constructed huge cranes that lifted Marcellus' ships from the water and dumped the sailors into the sea, as well as great lenses to focus the rays of the sun and set the ships on fire. He made catapults that heaved missiles on the advancing land armies and adjusted their range according to the distance of the soldiers being bombarded. He built machines to release showers of arrows from slots in the walls of the city, behind which the defenders remained protected.

It was said that after a time the besieging soldiers became so afraid of Archimedes' devices that they would run if even a rope or stick were put forth over the wall of the city. Finally Marcellus gave up hope of taking the city by force as long as Archimedes defended it, and encamped to wait for the city to run out of food, which eventually succeeded.

When the city was at last overrun, the army had strict orders Archimedes not be harmed but they were not followed. A soldier encountered Archimedes considering a mathematical problem and not realizing who he was, demanded Archimedes come along as a captive. Archimedes was either too absorbed in thought to hear the order or did not care to obey and was killed. Some accounts even suggest that Archimedes ordered the soldier to stand away from his diagram so he could continue his work.

Although his mechanical devices amazed his fellow citizens, Archimedes had so little regard for them himself that he left no written works devoted to applications of his mathematics. On the other hand he was very proud of his achievements in pure mathematics as we remark below.

Volumes by "slicing"

In considering the limit method used by Euclid (possibly due originally to Eudoxus), it seems Archimedes refined it by again using limits to express areas and volumes, but in a simpler more systematic way. Much as we do today in integral calculus, he approximated areas and volumes by a finite sequence of horizontal strips or slabs, which are rectangles or right prisms, and then took the limit.

I want to emphasize this is my somewhat speculative, but educated, restatement of what he did since I have not closely studied his work. His discussion in "The Method" implies it was originally inspired by a consideration of how different solids balance each other when placed at different distances on a scale. Thus he used physical reasoning to measure volume by its relation to weight and momentum. It is instructive to google some of the wonderful illustrations of his ideas that exist on the web.

“Cavalieri’s principle”
The powerful new idea that follows directly from this technique of approximation by parallel right prisms, is that two solids whose “slice areas” are the same at every height, must have the same volume. This is known today as Cavalieri’s principle after an Italian mathematician who apparently discovered it some 1800 years after Archimedes. More precisely, imagine two solids lying between the same two parallel planes. If every plane parallel to these planes cuts plane figures of equal area on the two solids then the solids have the same volume.

We can apply this principle in two dimensions to establish again the fact that triangles on the same base and in the same parallels have equal area. The two dimensional Cavalieri’s principle says that if every line parallel to the base of the two triangles cuts both triangles in congruent segments, then the triangles have the same area. For triangles on the same base and in the same parallels, it follows from similarity that the segments cut at the same height have the same ratio to the common base, hence are equal. Thus Cavalieri implies the triangles have the same area.

In three dimensions imagine two pyramids on the same triangular base and both having vertices on the same plane parallel to the base. Each side face of a pyramid is a triangle whose base is a side of the base triangle of the pyramid. Again by similarity, horizontal slices at the same height are triangles whose sides have the same ration to the base, hence are equal. Thus by SSS the slices of the two pyramids at the same height are congruent triangles, and by Cavalieri the pyramids have equal volume. With a little more effort, the same result follows for pyramids on any polygonal base, e.g. by triangulating the base polygon.

This method, that equal slice areas imply equal volume, is so powerful that Archimedes was able to deduce the volume of a sphere inscribed in a cylinder, a result he was so proud of that he asked it be engraved as an epitaph on his tombstone, which apparently was done. We sketch that result and then another one he discovered, which was erased from his manuscript in the middle ages, and rediscovered some 100 years ago.

Volume of a pyramid
Just as it was fundamental to know that a triangle has half the area of a parallelogram with the same base and height, a fundamental result for volume is that a pyramid has 1/3 the volume of a prism with the same base and height.

In one nice case we can actually decompose a prism into three pyramids. I.e. a cube can be cut into three congruent right pyramids, and it is instructive and fun to illustrate this with cardboard models as we did in class. One can also visualize this as follows. Consider a corner of the cube and notice there are three faces adjacent to that corner. Choose one face and form a right square pyramid with that base by joining every point of that face to the opposite corner of the cube. Since there are three such faces around that first corner, there are three such right pyramids and they fill up the cube. These pyramids are all congruent by a rotation around the axis joining the two opposite corners of the cube. Thus each of these right pyramids has 1/3 the volume of the cube.

Exercise: Figure out the dimensions of the faces of these three pyramids and construct them so they fit together to give a cube.

In class Joshua pointed out it is easier to visualize decomposing a cube into 6 pyramids, each having one face of the cube as base, and vertex at the center. Each of these has the same base as the cube but only ½ the height. By similarity, if we consider pyramids with the same square base as a cube and the same height, the slice areas would double and so would the volumes. Thus it would only take 3 such pyramids to equal the volume of the cube, verifying again that a pyramid with the same base and height as a cube has 1/3 the volume. More generally, any pyramid with the same base and height of a prism has 1/3 the volume of the prism.

Now just as Joshua suggested looking at cube as a union of pyramids with faces as bases and vertices at the center, we can look at any polyhedron this way, especially symmetrical ones like the 5 regular polyhedra. E.g. Imagine an icosahedron and imagine joining each triangular face to the center of the icosahedron. Each face again forms the base of a pyramid with vertex at the center. Thus the icosahedron is a union of 20 triangular pyramids whose bases make up the surface and whose heights equal the radius of an inscribed sphere.

To get the formula for the volume of an icosahedron, we would add up the volumes of the 20 pyramids. To compute it we need the edge lengths of the triangular faces and the radius of (a sphere inscribed in) the icosahedron. In fact either of these measurements determines the other, but I do not know the formula relating them. But at least we can see that the volume equals 1/3 the radius times the surface area i.e. the area of all the faces, since the volume of each pyramid is 1/3 the radius times the area of its base. For a cube the radius is ½ the side length of a face. Thus the volume of a cube of side s equals s/6 times the surface area, i.e. s^3 = (s/6)(6s^2).

Archimedes' method of computing the volume of a sphere was to compare the volume the sphere with those of the cylinder and the cone. We discuss those two figures next.

Volume of a cone

A cone is a pyramid whose base is not a polygon but a circle. Euclid also assumed his cones are vertical, i.e. that the line joining the vertex to the center of the base is perpendicular to the base. In his famous work on spheres, Archimedes said that a sphere is a cone whose base is the surface of the sphere and whose vertex is at the center. This is analogous to the discussion above, where we considered a polyhedron as a (family of) pyramid(s) with base the surface of the polyhedron and vertices at the center of the polyhedron.

The same principle applies to a circle. When we inscribe a polygon in a circle the polygon is an approximation to the circle. The more sides the polygon has, the better it approximates the circle. If we connect each vertex of the polygon to the center we can think of the polygon as a family of triangles with vertices at the center of the circle, and with bases equal to the circumference of the polygon. The area of the polygon is thus an approximation to that of the circle, and equals ½ the total base of these triangles times their common height, i.e. equals ½ the circumference of the polygon times the radius. As the number of sides of the polygon increases, the circumference of the polygon approaches the circumference of the circle and the area of the polygon approaches the area of the circle. Thus the area of a circle equals ½ the circumference of the circle times the radius. This formula $A = ½ CR = (1/2)2\pi R.R = \pi R^2$ is the usual one for the area of a circle. Indeed the number π is defined by the formula $C = 2\pi R$, i.e. $\pi = C/2R$.

A cone is approximated by a pyramid whose base is a polygon with a large number of sides. As the number of sides increases, the base of the pyramid approaches the circular base of the cone, and the volume of the pyramid approaches that of the cone. So the volume of a cone is also 1/3 the area of the base times the height. This gives us the formula $V = (1/3)\pi R^2 H$, for the volume of a right circular cone of base radius R and height H. In particular the volume formula for a cone is simpler than for a pyramid.

Relation between volume and surface area for a sphere

Think of an icosahedron circumscribed about a sphere as an approximation to the sphere. If we join the points of each face of the icosahedron to the center of the sphere we decompose the icosahedron as above into a family of pyramids whose bases make up the surface of the icosahedron and whose common height equals the radius of the sphere. The volume of the icosahedron equals the sum of the volumes of the pyramids associated to each face. This volume equals 1/3 the sum of their base areas times their common height, i.e. 1/3 the surface area of the icosahedron times the radius of the sphere. If we circumscribe a polyhedron with more faces about our sphere, we get a better approximation. The surface area of the polyhedron approaches closer to the surface area of the sphere and the volume of the polyhedron approaches closer to that of the sphere. Thus in the limit, the volume of the sphere equals 1/3 the surface area of the sphere times its radius. This does not compute either of these quantities, but now if we can compute the volume of a sphere we will know the surface area as well.

Recall if C is the circumference of a circle and R its radius, we know the area of the circle is (½)CR. Now we have an analogous formula for a sphere. If S is the surface area of a sphere and R is its radius, then the volume of the sphere is (1/3)SR. It is tempting to conjecture that in 4 dimensions the volume of the 4 ball of radius R is R/4 times its (3 dimensional) surface area, and so on, and this is true. Nonetheless we still need to compute either the volume or the surface area.

Volume of a sphere
Archimedes computed the volume of a sphere by comparison with the volumes of a cylinder and a cone using Cavalieri's principle as follows. It is a little easier to describe using a hemisphere than a full sphere. So consider a hemisphere of radius R inscribed with equator at the bottom in a cylinder of base radius R and height R. Consider also an inverted cone, i.e. one with vertex at the bottom, inscribed in the same cylinder. This cone has (upper) base radius R and height R.

By Cavalieri's principle we can compare the volumes of these figures by comparing their slice areas at the same heights. At height zero, the cylinder has slice area πR^2 as does the hemisphere, while the cone has slice area zero, since its vertex is at the bottom. We claim the slice areas of the cone and the hemisphere add up to that of the cylinder at every height.

Consider the slice areas at height x. Each of the figures has as slice a circle at every height, so we only need to compute their radii at height x. For the cylinder every radius is R and every slice area is πR^2. The cone has base radius R and height R, so by similar triangles the circular slice at height x has radius x, hence area x^2. The hemisphere has at height x a radius r that satisfies x^2 + r^2 = R^2, by Pythagoras. Hence the radius r satisfies r^2 = R^2-x^2, and the slice area is πR^2 –πx^2. Thus indeed the slice area at height x of the hemisphere plus that of the cone, equals that of the cylinder. Thus by Cavalieri the volumes also add up.

Since the cone has volume = 1/3 base times height, and the cylinder has volume equal to base times height, the hemisphere must have volume 2/3 base times height, or 2/3 (πR^2)R = 2/3 πR^3. This agrees with what we learned in school, since we were told the volume of a full sphere is V = 4/3 πR^3. From Archimedes' comment above we also get the surface area S of the sphere, since V = (R/3)S, so S = (3/R)V = 4πR^2.

Volume of a "bicylinder"
Right after recording this result in his famous book, Archimedes states that the volume of a "bicylinder" can be done in the same way, but the solution is not given. A bicylinder is the solid made by intersecting two circular cylinders of the same radius, meeting at right angles. The reason for the missing solution is very interesting. Archimedes' works were written on parchment, which became very valuable as time went on, and there was an effort to re - use parchment for other purposes. Sometime in the middle ages when appreciation for Archimedes' work had presumably diminished, his book was washed to provide parchment for use as a prayer book. Ironically the pages that were washed and re - used can still be read today, because the washing was not totally successful, and the prayers were written perpendicularly to the mathematics. The unfortunate part is that the prayer book was shorter than Archimedes works and some unneeded pages of Archimedes writings were removed entirely. It is those lost pages that contained the solution to the problem of the volume of the bicylinder.

Still we can try to guess the solution by analogy with his solution for the sphere, since Archimedes said the two were similar. The first job is to compute the slice area of a bicylinder. This solid is rather hard to visualize but it looks like a sort of pagoda reflected in a pond, with a point at the top and bottom. The key to visualizing the slices is to realize that a horizontal slice of the intersection of two cylinders is just the three way intersection of the two cylinders and the horizontal plane. Moreover, intersection is associative, commutative and even distributive over itself, as an operation. Thus the intersection with the plane can be done to each cylinder separately and then the results can be intersected. But a horizontal plane meets a horizontal cylinder in a rectangle, and two perpendicular rectangles of the same width meet in a square. Thus the horizontal slices of the bicylinder are all squares.

Let's write down the area of one of these square slices at height x. By analogy with the case of a sphere where we looked only at a hemisphere, we look at the top half of the bicylinder. If the two cylinders had radius R, the slice at height zero is a square of side 2R. At height x, we are intersecting two rectangles of the same width and we only need to consider one rectangle to compute the width. This is obtained by slicing a semi circle perpendicularly at height x, so half the width r of the slice satisfies x^2 + r^2 = R^2. So the square slice has area (2r)^2 = 4(R^2-x^2). This looks familiar from the sphere case but with 4 in place of π.

Ok, now we have to fill in the missing pieces of Archimedes' solution. In the sphere case we had three solids, all with horizontal slices which were circles, and the radii were R, x, and sqrt(R^2-x^2). Now we have only one solid, the half bicylinder, with slice which is a square with half its side length r = sqrt(R^2-x^2). We want to come up with two more solids also having horizontal slices which are squares, presumably with half their side lengths equal to x and R. Moreover they should be analogous to a cylinder and a cone.

What is analogous to a cylinder, but has square horizontal slices of constant size? The obvious choice is a half cube, and for every slice to have half its side length equal to R, it should have side 2R and height R.

What is analogous to an inverted cone but has square horizontal slices, with side length equal to 2x at height x, i.e. with side length varying directly with the height? It seems clear it should be an inverted square pyramid of height R with (upper) base a square of side 2R.

If we compare the slice areas of these figures at height x, we get 4R^2 for the cube, 4x^2 for the square pyramid, and 4(R^2-x^2) for the bicylinder. Hence the volumes also add in the same way, and half the bicylinder has volume equal to the volume of the half cube minus the volume of the square pyramid. Since the square pyramid again has 1/3 the volume of the half cube, the volume of the half bicylider is 2/3 that of the half cube. Hence the full the bicylinder has 2/3 the volume of the cube, or (2/3)(8R^3) = (16/3)R^3.

The same relation $S = (3/R)V$ between the surface area and volume holds again, so the bicylinder has surface area $S = 16R^2$. Surface area is often quite hard to calculate by calculus, and I have not seen this surface area computation done for a bicylinder in a modern calculus book.

Newton's approach

We have seen that Archimedes knew that the volume of a figure is determined by all its slice areas. The next advance is a way to go from a formula for the slice area to a formula for the volume. This may not have occurred to the Greeks because of their preference for geometry over algebra. It may be that this advance was not possible until the rise of algebra for expressing formulas. Today a calculus student learns to do the calculation of the volume of a sphere as follows.

From the fundamental theorem of calculus, and the method of "volumes by slicing" we learn in calculus, that if the slice area formula for a solid at height x is $A(X) = x^n$ then the volume formula for the portion of the solid up to height x is $V(x) = [x^{(n+1)}]/(n+1)$. Thus to get the volume formula we raise the power of the area formula by one, and then divide by the new power. This is called anti-differentiation, or "integration" for polynomials.

Recall the slice area formula for a sphere, $A(x) = \pi R^2 - \pi x^2$. Applying the rule above gives volume formula $V(x) = \pi R^2 x - \pi(x^3/3)$, (because the power of x in the first term πR^2 was $x^0 = 1$). Setting the height x equal to R, gives the volume of the hemisphere $V(R) = (2/3)\pi R^3$, so again the volume of the sphere is $(4/3)\pi R^3$.

Exercise: Use calculus to find the volume of the bicylinder from its slice area formula.

Epilogue to epsilon camp geometry notes

Volume calculations with and without calculus

Terminology: Today we speak of the interior of a sphere as a ball, so we would not speak of finding the volume of the sphere in three space, but only its surface area. I.e. we also consider the sphere in three dimensional space to be only 2 dimensional even though it is curved and lives in three space. Thus we say the area of the 2 - sphere of radius R is $4\pi R^2$ and the volume of the 3-ball is $(4/3)\pi\hat{}R^3$.

"Surface area" versus volume: If we try to compute the volume of the 3- sphere in 4 space, or the 4 dimensional volume of its interior, the 4- ball, we have a relationship between the sphere and the ball analogous to the one found by Archimedes. Just as the volume of the 3-ball equals R/3 times the surface area of the 2-sphere, the (4 dimensional) volume of the 4- ball of radius R equals R/4 times the (3 dimensional) volume of its surface, the 3- sphere. So again we only need to find one of them.

Volumes by slicing
If we try to use calculus to do this 4 dimensional volume in the same way as for the 3-ball, using volumes by slicing, an algebraic difficulty arises. The radius of the slice is always a square root, and in odd dimensions that square root is raised to an odd power, which makes it a fractional power that is harder to anti - differentiate. I.e. the slice area of the three ball is the area of a 2-ball or disc, which was $\pi r^2 = \pi(R^2-x^2)$ by Pythagoras. This is a nice integral power of x and is easy to anti -differentiate as we saw earlier. However the slice volume of a 4 ball is the volume of a 3-ball, namely $(4/3)\pi r^3 = (4/3)\pi(R^2-x^2)^{(3/2)}$, since again $r^2 = R^2-x^2$. This formula is harder to anti differentiate. Although it is possible to do it using trig functions, and you will learn this in calculus, we will use an easier approach.

Area of a 2-disc revisited:
horizontal slices: If we look back at the area of a 2-ball or disc, it is more natural to look at it as an expanding family of circles, rather than a growing stack of straight slices. I.e. if we grow the area upwards, with the slice at height x being a segment of length 2r, where $r^2 = R^2-x^2$, then we have a slice length formula $2r = 2(R^2-x^2)^{(1/2)}$, which is hard to anti - differentiate. In fact you will recall we did this area problem by taking a limit of areas of triangles rather than by calculus.

circular slices: It is easy also by calculus if we grow the area outwards, with the leading edge of the growing area being a circle of radius x. Then the slice length is the length of this circle, which is $2\pi r = 2\pi x$, and this is easy to anti differentiate, as πx^2. Setting $x=R$ gives us πR^2 immediately as the area of the 2-disc, i.e. the area of the interior of the circle. [I am reminded of a remark long ago by a friend of mine, an Indian artist who saw me shading a disc with horizontal lines in a drawing, and said he would never do it that way, but would draw expanding circles instead. It just did not make visual sense to him my way.]

center of mass : The area calculation for a disc can also be done without calculus in a way used by the Greek mathematician Pappus, and understood by Archimedes, as follows. One knows in physics that the momentum of a body can be computed by thinking of the entire body as located at one point, its center of mass, or center of gravity. A 2-disc is generated by revolving one radius around the center of the circle. The area generated is equal to the length of the radius multiplied by the distance traveled by its center of mass, or center of area. Since the center of the radius of length R is the point at distance R/2 from the center, when the radius revolves around the center that point travels a distance of $2\pi(R/2) = \pi R$. Since the radius has length R, thus the area generated equals πR^2.

This calculation is essentially the same as the one we did earlier using limits of triangles. Recall that calculation yielded the formula $A = (1/2)CR$, where C is the circumference of the circle or the limit of the bases of the triangles. Another way to look at the area formula for a triangle is that it equals the height times the average base which is B/2, the length of a segment parallel to the base but halfway up the triangle. Then C/2 is the limiting value of the total average base of the triangles approximating the circle, so the formula for the area of the circle again equals (C/2)R. This of course equals the distance traveled by the center of the revolving radius times its length.

Volume of a 3-ball revisited:
i) horizontal slices: For the volume of the 3-ball the opposite algebraic situation occurs. I.e. if we think of the 3-ball in terms of horizontal slices, the formula x^2 +y^2 +z^2 = R^2 for the 3-ball, gives at height x, the formula y^2+w^2 = R^2-x^2 = r^2, the 2-disc of radius r = sqrt(R^2-x^2) as we know. Then the slice area is πr^2 = π(R^2-x^2), which is easy to anti-differentiate, as we did before, getting π(R^2x – x^3/3). Setting x=R of course gives (half) the volume of the 3-ball as (2π/3)R^3.

ii) cylindrical slices ("shells"): We can also look at a 3-ball as obtained by revolving the right half of a 2-disc around the y axis. If we think of the 2-disc as a union of vertical lines drawn from x = 0 to x = R, we have a line of height H, where H^2 = R^2-x^2. Thus the revolved segment sweeps out a cylindrical slice of the 3 ball having area 2πrH = 2πx.sqrt(R^2-x^2). This formula is harder to anti -differentiate than the horizontal slice formula above, but not too hard using the "chain rule" or "substitution" formula which one learns in calculus.

iii) center of mass: We can calculate the volume of a ball in principle by multiplying the area of the half disc, by the distance traveled by its center of mass. However it is not at all obvious where the center of mass is for a half disc. Since we already know the volume of 3-ball, we can use it backwards to locate that center of mass as follows. If the center of mass of the right half of the disc of radius R is located at distance r from the y axis, then the volume (4/3)πR^3 of the 3-ball, equals the distance 2πr traveled by this point times the area (1/2)πR^2 of the half disc. Thus we should have 2πr(1/2)πR^2 = (4/3)πR^3. This implies, let's see now, π^2.r.R^2 = (4/3)πR^3, so r = 4R/(3π) I hope, or a little closer than halfway to the y axis. This makes sense because the half disc is thicker near the y axis.
Since Archimedes knew the volume of 3-ball, he would have known this center of mass as well.

Volume of a 4 dimensional ball
horizontal slices: If we consider the 4-ball of radius R, with equation x^2+y^2+z^2+w^2 = R^2, the horizontal slice at height x is the 3-ball with equation y^2 + z^2 + w^2 = (R^2-x^2), of radius r = sqrt(R^2-x^2). (As usual we only consider half the 4-ball, starting with the slice at height x = 0 being the 3-dimensional "hemisphere" y^2 + z^2 + w^2 = R^2.) Here again, since 3 is an odd dimension our slice volume formula equals (4/3)πr^3 = (4/3)π(R^2-x^2)^(3/2), which is again hard to anti-differentiate without using complicated trig formulas and a new technique of "integration by parts".

"Cylindrical shells": Just as in all previous cases, we can generate a 4-ball by revolving half a 3-ball around an axis. Although is hard to visualize, we proceed exactly by analogy, and consider revolving the horizontal slices of that half 3-ball around an axis. Thus the slice at height x would be a 2 disc of radius r, where r^2 = R^2-x^2 as before, and this 2 dimensional slice would revolve around a circular path of length 2πx. Thus the full revolved 3 dimensional cylindrical slice of the 4 –ball would have volume 2πx(πr^2) = 2π^2.x.(R^2-x^2) = 2π^2R^2.x – 2π^2.x^3. This formula is easy to anti-differentiate using the familiar formula x^(n+1)/(n+1) for the anti-derivative of x^n, yielding π^2R^2.x^2 – (1/2)π^2x^4. Setting x = R, gives the volume of the 4-ball as (1/2)π^2.R^4.
Any calculus student could do this calculation but most have not seen it.

Centers of mass: Again we would not know the center of mass of half a 3-ball without using the volume formula above for a 4-ball, so cannot yet compute that volume this way, but we can use Archimedes' trick to do a calculation that Archimedes could have done. Namely he showed that the volume of half a 3-ball equals the difference of the volumes of a cylinder min us that of a cone. Now the center of mass of a cylinder is obviously half way up, and Archimedes knew that just as the center of mass of a triangle is 1/3 of the way up from the base, the center of mass of a cone is ¼ the way up from the base.

Thus we can use centers of mass and subtraction to get the volume of a 4-ball. I.e. a cylinder of height R and base radius R has center of mass at height R/2, and volume πR^2.R, so revolving it around an axis at its base gives 4 dimensional volume of 2π(R/2).πR^2.R = π^2.R^4. Now the inverted cone of height R and base radius R has center of mass at distance ¼ of the way from its base, hence distance (3R/4) from the axis, and volume (1/3)πR^2.R. Thus revolving it generates a 4 dimensional volume equal to (2π)(3R/4). (1/3)πR^2.R = (1/2)π^2.R^4. Subtracting the volume of the revolved cone from that of the revolved cylinder, gives the 4 dimensional volume of the revolved half 3-ball, i.e. the volume of the full 4-ball as π^2.R^4 - (1/2)π^2.R^4 = (1/2)π^2.R^4.

Remarks: This is another little calculation I made up just for you guys. I.e. I have not seen how to generalize Archimedes' work to 4 dimensions before, although it is probably out there somewhere in the wide wide world.
The only thing I have not justified is how Archimedes knew the center of mass of a cone. He probably discovered it using a balance beam, and then justified it later mathematically. One approach is to take the limit of an approximation by slabs as follows. Chopping an inverted cone of height R and base radius R, into n horizontal slabs and approximating those slabs by discs, gives n discs of base radii (1/n)R, (2/n)R,…..,(n/n)R, and all of height R/n. The disc of radius kR/n generates force on the balance beam proportional to its distance kR/n from the fulcrum or balance point, which causes it to try to revolve around a circle also of radius kR/n.
Thus the disc of radius kR/n, and volume (R/n)(πk^2R^2/n^2) would generate force or work or something proportional to (kR/n) (R/n)(πk^2R^2/n^2). Not being a physics guy, I prefer to multiply this by 2π and think of it as generating 4 dimensional volume of (2πkR/n) (R/n)(πk^2R^2/n^2) = 2π^2.k^3.R^4/n^3.
Adding up over all n of these slabs for k=1,…,n, gives total 4 dimensional volume of (1/n^4)[2π^2.R^4]/(1^3+2^3+….+n^3). Now we know, and maybe Archimedes did too, a formula for the sum of those cubes of form n^4/4 + lower degree terms. Thus the formula becomes (1/4)(2π^2R^4)(1/n^4)(n^4 + terms of degree 3 or less in n). As n→infinity, this approaches the limit (1/2)(π^2R^4). To get this same result by revolving a mass of the same volume as the cone and concentrated at one point at distance r from the axis we would need (2πr)(1/3)(πR^3) = (1/2)(π^2R^4). Solving for r gives r = (3/4)R (measured from the vertex of the cone), as we claimed.

Volume of the 3-sphere, i.e. "surface area of the 4-ball"
To get the 3-dimensional volume of the surface of the 4-ball we can use Archimedes' relation and just multiply the 4-dimensional volume of the ball by 4/R. This gives 2π^2R^3 as the surface "area" of the 4-ball, i.e. for the 3-dimensional volume of the 3-sphere.