Edited transcript of April 10, 2021 “Math Beyond the Horizon” presentation: Where Am I - How does your GPS Work?, by David Hathaway

Slide 1:
I'm David Hathaway, a retired electrical engineer, and I worked at IBM for 32 years working on design automation software. But I also volunteer with the Green Mountain Club and teach map and compass navigation. I found that some of my friends who are very well educated people were very confused as to how their GPS actually worked. So that was what prompted me to dig into a little bit more research and put together this talk. So I hope that the end of the today you'll have a much better idea of what goes on inside your phone or other GPS receiver, and all of the surrounding infrastructure that's needed in order to find these very precise locations.

What I'm going to be talking about today is a specific aspect of navigation which is getting a fix. This means figuring out where you are. There other parts of navigation such as figuring out which direction you should go in to get to a target destination, and things like that I'm not going to be talking about that today.

Slide 2:
I'm going to start by talking about local navigation. What do you do if you're interested in finding your location within a relatively small area. For example if you're out hiking and want to know where you are. Second, I want to talk about global navigation, and before we go to GPS I want to talk a little bit about what people did before GPS, which is primarily

celestial navigation. This is if you're out in the middle of the ocean you have no landmarks and you want to figure out where you.

Slide 3:
Before we start talking about getting fixed we have to talk about what form the answer is going to come in. You're all familiar with different coordinate systems, such as XYZ Cartesian coordinates. But the Earth is almost a sphere (and we'll talk about what that “almost” means later and and why the difference from between a sphere and “almost” matters), so it is much more convenient to use latitude longitude and altitude rather than X Y and Z. This is also called spherical coordinates. I think you're all familiar with this. Lines of latitude remain parallel and lines of longitude converge at the poles.

Slide 4:
If we’re interested in navigation over local area we can pretend that the Earth is flat and work with a two-dimensional projection of the three-dimensional Earth surface. For the projection we’ll be using we keep the latitude and longitude lines parallel to the Cartesian X Y axes. If you are looking at a region near the north or south pole, assuming longitude lines to be parallel doesn’t work so well. I found out when I was looking at GIS or geographic information systems for Canada that they tend to use a polar or conic projection because Canada is closer to the pole than the equator, and in that conic projection the latitude lines, instead of being parallel straight lines, become parts of a circle, and the longitude lines radiate from the pole. But we're going to talk about longitude and latitude as X Y Cartesian coordinates

today. These are useful over a relatively small area. As a matter of fact, in a lot of cases we're not explicitly taking into account that coordinate system, but are using the fact that the coordinate system is built into the map we are using. The key is that a straight line on the map corresponds to a straight line of sight in the field. This just a picture of Vermont where we've taken a projection we've made all of the latitude and longitude lines parallel.

Slide 5:
Okay, so how do we do navigation local navigation with map and compass? The first thing you have to have is a compass, and a compass is basically just a an instrument for finding the angle, which we call a bearing, between north and some target. This is what an orienteering compass looks like that you might take out in field. It's what I teach how to use with Green Mountain Club. It has a base plate with an arrow that you point toward the thing you're interested in, a needle that will rotate and point to North, and a rotating dial you can use to capture and read the bearing. What you're basically doing with this compass if you're trying to find your location is to find the angle or bearing (that's the red arc in the figure) between the target and north. There's a whole bunch of stuff you have to worry about with something called declination, which is the difference between magnetic north and polar north. These are not the same direction and the declination varies from place to place. If you're if you're in Vermont you have a declination or a difference between magnetic north and polar north of about 14.5°, so magnetic north is about 14.5 degrees west of polar North. In California the declination is about the same magnitude, but in the opposite direction. So you have to know where you are to use the correct

declination, but we're not going to worry about that today.

Slide 6:
So how do you actually use a map and compass to figure out where you are? This is a a little piece of a topo map for the area around Camels Hump. Let's say I'm over in here somewhere and I want to figure out where I am. I'm going to use a system called triangulation. The first thing I do is to find some landmark that I can identify. In this case I've identified Camels Hump and I use my compass and I find the bearing toward that landmark. So I've now rotated my dial until the arrow on the dial aligns with the needle pointing north, and the direction of arrow travel arrow is pointing towards Camels Hump, and I know the bearing between them. Next I draw a line back from that landmark on my map according to that bearing, so I lay down my compass on my map and align the north arrow with the longitude gridlines on the map and draw a line parallel to my target arrow using the compass side as a straight edge. I then repeat that process for second landmark. In this case it's Mount Ethan Allen. So now I've got two different landmarks and two lines from them. And I can tell now that I'm at or near that star at the intersection of the lines. In general if I have more landmarks I would repeat this process again, and because there would be some error in my measurements, I would find out that there's a little triangle or a little region formed by the lines, because these won't coincide directly at a point, and I would assume that I'm in or near the little triangle that is formed. So this is the foundation of map and compass navigation.

Slide 7:
Here is a picture of the triangle formed by the two landmarks

and my location. The distance AB between the landmarks is known. We don't have a measurement for it but, it's embedded in the map. All of the lines to N are parallel because we've assumed that locally all of our longitude lines are parallel and I can take differences between the angle bearing that I measured to a landmark and the bearing from the other landmark to that landmark and those give me the angles at vertices A and B. I don't explicitly find those angle measures, but by drawing this picture I'm essentially creating a unique angle of that size. So what I've now got are an angle at A, a side AB, and an angle at B, and if you remember some of your geometry the ASA or angle side angle theorem of triangle congruence guarantees that there is a unique triangle including line AB with these measures, because any other triangle with those measures is congruent with it has and to superimpose on it with the same AB base line. This is the basis for a map and compass navigation.

Slide 8:
If we're not talking about a local area if we instead say I'm out in the middle of the ocean I have no idea where I am. I no longer have local landmarks. The ocean in every direction is just a flat horizon. What I do have are celestial objects like stars, the moon, the sun, and planets. So I can use those objects as landmarks instead to find my location. This is an example of how we find the latitude. Latitude is fairly easy, and is something that people have been able to find for thousands of years. Being able to find your latitude without knowing the time only works with Polaris because Polaris is pretty much directly over the North Pole within a few arc minutes at all times so I don't it's not depending on the time of day.

Slide 9:
Finding latitude longitude is much harder. The problem is that every point at a given latitude sees the same stars at the same local time of every day, so every point on a latitude line will be below some fixed star at some time of the day but if you don't have an accurate clock you have no way of knowing what longitude you are at. In order to do that you need to know the time difference between your local noon and noon at a zero reference point, which is Greenwich naval observatory in London (they kind of won out about 300 years ago when Paris and and London were vying to be the standard for the prime meridian). But the earth rotates about a quarter of a degree per minute that 15 arc minutes per time minute, so if you're off by one minute of time that'll give you about a 15-mile error at the equator. Back in the 1700 this was a huge deal, since being able to have accurate naval navigation was the one of the biggest economic advantages of country could have. So the British Parliament offered this huge prize equal to about 4 million dollars today for anyone who could come up with a clock that was accurate to three seconds per day, this being what you needed to achieve about a half a degree navigational accuracy over the 6 weeks it took to sail from Great Britain to the Caribbean. At the bottom is a picture of a clock that was the first accurate navigational accurate clock by built by John Harrison in 1730. He made a number of others, there is a lot more interesting information about him, his clocks, the Longitude Prize, and more in a the book “Longitude,” by Dava Sobel.

Slide 10: So what do we do with the time once we have an accurate time? We can generalize our latitude method. We

can determine the angle from the horizon to some distant star just like we did to find the latitude using Polaris, and in this case the object that we're using replaces Polaris. Then, using the time and an accurate an almanac, we figure out where that object is directly overhead at the time we make our measurement. So I might determine that at exactly 4:23 a.m. on April 10th that particular star is over some point X out in the middle of the Atlantic at latitude 32 and longitude -30. I can then find the angle and distance from X to my location. This finds a circle on the globe of possible locations and those locations are kind of like latitude lines around that point X that I just found, where X replaces the north pole.

Slide 11:
This is a picture of that process, where I'm assuming that some object is very far away. If you tilt the center line up it'll look just like the picture that I showed for finding latitude. Here, altitude A is the angle between the line to this distant star and my local horizon, and that angle is the same as the angle between a plane through the Earth's center that is perpendicular to the to the line to that star, where that plan is analogous to the equator in the latitude method. So this is defining a circle that's kind of like the latitude was in the previous method.

Slide 12:
There's one complication if you're looking at a nearby object like the moon, which is used a lot in navigation. It's close enough that for the accuracy that you want with navigation, you can’t assume that lines to the object (e.g., the moon) from different points on earth are parallel. This deviation from parallel is called parallax, and has to be accounted for in

navigation. The teacupnavigation.net at the bottom is where I got these pictures. It has a lot of useful information on navigation.

Slide 13:
If I put these things together I get two circles, one from object one and another circle from object two and of course I also know I'm someplace on the surface of the Earth so the intersection of those two circles defines two possible points. In this case we're making the assumption that I know within maybe a thousand miles where I am so I can tell the difference between being at this upper intersection and the lower intersection. But if you completely had no idea where you were you could repeat this process for third object like maybe Polaris and find out that I'm at this particular latitude line and an intersections of these two circles and that's how I can get my fix.

Slide 14: So all of these methods, like the map and compass method, require measuring angles. Map and compass required an angle between my target and north, and celestial navigation requires an angle between some celestial object and the horizon. These are various method that have been used over the ages for doing this. A kamal is a very simple device that is just a string with some knots on it and a fixed sized object and you use the string to measure how far you have to move the object out from your eye so that one end is lines up with the object and the other end lines up with the horizon. The rest are successively more refined angle measurement devices. The backstaff is a particularly interesting one because it's used to take sightings on the sun. You can imagine that you really don't want to stare at

the sun to find its altitude because this is not very good for your eyes. The backstaff does this by having the sun at your back, and letting you line up the edge of a shadow with the horizon.

Slide 15:
Today the most accurate tool for manual navigation is a sextant. This allows very precise measurement to within a few minutes of arc by using a split image in which you line up the horizon with a reflection of the object to which you're measuring the angle. They include sun shades so you can find the altitude of the sun without damaging your eyes. When using it, you are looking at the horizon and the reflection of the object at the same time and you get them aligned by moving this bar back and forth roughly and then tuning with this micrometer drum to get the very precise angle between the object you're sighting on and the horizon.

Slide 16:
So enough of traditional navigation. You came here to find out something about GPS which is the global positioning system. This was developed starting in the 1970s by the defense department and was intended for the military purposes. The big problem they were trying to solve was not just navigation but also weapons targeting. If you look at the number of bombs were dropped in Vietnam and in World War II and the numbers that actually hit their targets it was a very small number. Not only was this better for their military objectives, it also reduced the civilian damage that you got from bombs that went off in the wrong place. Fortunately we can use the same system for a lot more peaceful purposes today. It was first operational in 1995, and because it was a

military thing they didn't want people to have access to the detailed measurements or detailed fixes that could be used against us, so there was this thing called selective availability that was used until about 2000, where they intentionally added error to the signals sent out. It was a deterministic error and there was a special cryptography key that the military could use to figure out how to translate from the erroneous data that it sent to the real date so you could get enough data to get about a 100 to 200 M fix but you couldn't get anything more, even though the signals that are coming from the satellites are accurate enough to get you about a one meter fix. GPS is similar to the LORAN (long-range navigation) system that was using similar principles from land-based stations. The GPS system has 24 satellite and seven spares in 6 orbital planes. All satellites transmit on two frequency bands of 1.5 and 1.2 gigahertz called the L1 and L2 signals.

Slide 17:
This is a picture of one of the first generation satellites that is still in use today.

Slide 18:
This is a picture of one of the latest generation satellites I think slide is about a year old so I think there's there may be more in place right now but this is kind of the evolution and they're slowly replacing some of the old ones with the the new satellites.

Slide 19:
This is just a picture of the constellation of orbits of the satellites. You can see that they're not equatorial orbits because you want to get good coverage of the whole Earth.

Slide 20:
So how does this thing work? I will go into a lot more detail, but the first thing to know is satellites don't track you. The friend that I said was confused about how it worked said he didn't really want to use a GPS because he didn't want satellites tracking him. Satellites do not track you. They have no idea you exist. They would be doing exactly the same thing if there were 10 billion GPS receivers on the ground or if there were zero GPS receivers on the ground. They are simply sending information which allows a receiver, using methods we will talk about, to compute its location from that data. Essentially what each satellite is sending is a continuous message over and over again with its exact, and I emphasize exact, location and time. The receiver figures out how far it is between the receiver and the satellite. By knowing the speed of light and the time it took the signal to travel from the satellite to the receiver, it can determine the distance, which is the speed of light times the signal delay. So if I find out that a signal took 20,000 nanoseconds, or 20 microseconds, to travel a given distance, I know that distance is around 20,000 feet.

Slide 21:
This illustrates the signals propagating from several satellites to my location at the black spot on the earth. The signals propagate in all directions, but I've drawn the yellow lines to show the path the signals take from the satellites to my receiver. The animation shows how, even if the signals are sent at precisely the same time, they arrive at the receiver at different times, because the distance from the receiver to each satellite is different.

Slide 22:

Let's talk a little bit more detail about how this thing works. It's a little more complicated, because even though the satellites have very accurate clock that have to be within three nanoseconds of the actual time at all times, your GPS receiver has a cheap clock because you're not going to get an atomic clock in a handheld GPS device. It's a cheap clock but the drift is fairly slow, so the clock error can be considered constant during fix that is taken taken over a few milliseconds. Note that even if I had a clock that's accurate to one part in a million, which is about a second of error every 11 days, that's enough clock drift put your fix off by 300 M in one second so you cannot rely on just setting your GPS clock and leaving it; your receiver has to figure out the accurate time with every single fix. So the way this works is a satellite transmits its location which is this combination xs ys zs and time ts, where ts tells you it sent this message at time ts. The signal is received at the receiver at a time the clock in the GPS receiver clock says is time trs. Because the receiver clock is not accurate there is some error te in this signal receipt time, so the real time of signal receipt is trs - te. Now we can just rely on the speed of light. The next equation just has the Pythagorean distance squared the on the left, which is sum of the squares of the differences between satellite and receiver the x coordinate squared, the y difference squared, and the z difference squared. On the right we have c squared times the difference between the signal transmission time ts and the real signal receipt time trs - te. So now I've got four different variables that I have to find I have to find. I want my spatial coordinates xr, yr, and zr, but I also have to find the time error. I said you can't set that and forget it, you have to figure that out with every fix, because your GPS clock is not

accurate enough. It can't keep accurate enough time even for a second to rely on that clock. So what I do is I set up these four equations I've got four unknowns and by solving that system of equations I can figure out where I am and I can figure out my clock time error. This is why I think I said in the in the abstract that your GPS receiver, when you have a fix, is the most accurate clock you'll ever know you'll ever own, because at the time that it gets the fix and has calculated that te, it has to be accurate to within about three nanoseconds to get a one meter fix. It will start to drift immediately, but at that precise moment it knows the time to within three nanoseconds. You can't get that a radio clocks that receive signals from the atomic clock in Colorado signal from Denver from Colorado because it doesn't account for the time of flight between that transmitter and you. The GPS receiver is explicitly taking that into account, and as a matter of fact they are so accurate that they're the foundation for a lot of financial transactions because it's so important to make sure that you order financial transactions correctly and you know what things occur between before what other things.

Slide 23:
It's very hard to visualize the time error because it's a fourth dimension we're trying to solve 4 equations 4 unknowns we got a four dimensional problem. So here is I've tried to change it into a three-dimensional problem, meaning two X, Y spatial coordinates and T. In this picture the vertical lines that you see are the timelines of the satellites in red, yellow, and cyan, and of the receiver in green. The black grid is a moment in time and that time is moving forward in a moving up. The cones represent the signals from three GPS satellites. At a particular moment in time when the signal from

the cyan satellite reaches the green line that is my location, you'll see a little piece of a black cross. The orange ball and line represent my local time error. So when the black grid representing “now” is at a particular point, the orange ball representing what my receiver clock thinks is the time is ahead by my time error, which is the length of the orange line.

Slide 24:
Note that the four equations we are solving are nonlinear, because we have squared terms for xr, yr, zr, and te. But it's possible to convert these into linear equations. If we expand all of the terms in the first line line we get the expression in the next line. Note that all of the squared terms involve only my receiver location information, and nothing specific to a particular satellite. So if I take the difference between these equations for two different satellites, all of the squared terms for the unknowns will disappear, and I'm left with a purely linear equation. If I then take differences between three other independents pairs of satellites, I would now have a system of 4 linear equations in 4 unknowns and if you know how to solve systems of equations you could use Gaussian elimination or some other method to solve these equations. You may see the term pseudorange when you read about GPS. This is the distance your receiver would calculate to a particular satellite base on its uncorrected clock

Slide 25:
Note that there no angles involved in these calculations. This is trilateration, as opposed to the triangulation used in map and compass and celestial navigation. The real calculation is not usually done using linear equations, even though I just

showed you can do so, because there's some complexities and adjustments that are involved. The biggest of these is that you usually will use more than four satellites. Since each measurement will have some error, you would use a method like least squares to minimize error. Another issue is that there is some perturbation of signals as they travel through the ionosphere and atmosphere, and those have to be accounted for to get an accurate fix. One of the reasons that the satellites send out signals on two different frequencies is because it allows them to do some calculations to estimate that error. You also have to account for relativistic effects. Einstein's theory of gravity that says that space-time is curved and if you are near a massive object like the Earth there is some time dilation or slowing of time relative to outside of the gravity well. There is also a very tiny curvature of space-time, what you may have heard of as gravitational lensing which means light is not traveling in what we would think of is a straight line so the angles between the lines that the signals take to get between you and the satellite can actually add to very slightly more than a hundred eighty degrees in large triangles around the earth.

Slide 26:
I said at the beginning we're going to give our answers in terms of latitude longitude and altitude and I've just done all this stuff with X, Y, and Z. So how do we convert back to the coordinates we want. If we are dealing with a spherical Earth it's fairly easy. The first equation is the equation for the sphere the longitude is simply the arctangent of the ratio of the Y and Z axis. My altitude is simply going to be my radius minus the Earth's radius and then the latitude I can find once I found r as the arcsin of z/r.

Slide 27:
But in reality the Earth is not a sphere, it's closer to an ellipsoid. It's not nearly as exaggerated today shown here but I've shown this so to magnify the effect. The problem is that the altitude we're interested in is perpendicular to the ellipsoid surface. Remember the ellipsoid is supposed to be kind of a local idealized sea level horizon, so I want to go straight up from that, perpendicular to it. But for an ellipsoid the radius to that point from the center of the earth is not perpendicular to the ellipsoid. So the longitude calculation is the same, but the latitude and altitude are calculated using iterative formula. One last point that I put on this slide is that sometimes a GPS receiver can only get signals from three satellites, but we said we needed four for a fix. You can throw out one of those equations and it replace it with the equation of the ellipsoid Earth, so I can essentially assume I'm someplace on this ellipsoid. If I have three satellite equations, I once again I have four equations: 3 satellite equations and one ellipsoid equation.

Slide 28:
Another complication with altitude the Earth is lumpy. This is vastly exaggerated but the altitude is supposed to be measured against sea level, and sea level does not coincide with this nice perfect geometric ellipsoid surface. It's measured against an isogravimetric surface called a geoid isogravimetric is just a fancy word for the level to which water would flow if the whole earth were covered with water, with no waves and no tides. But the earth has mass concentrations, tectonic plates, and so on, and so this picture shows a rough approximation of the geoid height, where the

blue areas are where the geoid surface, this virtual sea level, is below the ellipsoid, and red is where it's above . It can deviate from the ellipsoid surface by +75 to -100 meters.

Slide 29:
The geoid is found by using very precise accelerometers. The altitude that your GPS receiver reports is the relative to its geoid model. It is the altitude relative to the ellipsoid model minus the local geoid height. The global geoid model has a value for every square arc minute which is about five hundred million points but your your cheap commercial GPS has a rough geode bottle built-in so that's what it uses when reporting. Professional GPS can use the full model although sometimes that involves offline processing.

Slide 30:
To further improve the accuracy of a GPS fix, there are augmentation systems. The main one is WAAS, which was developed by the FAA to allow instrument landings and takeoffs. For for those it's pretty important to know exactly the altitude of a runway and whether you're a little bit above it for a little bit below it, so you needed more accuracy than you could get out of a regular GPS receiver. There are a bunch of other alphabet soup acronyms here, but they all operate in similar ways. These all work by having a bunch of ground stations whose precisely surveyed locations are known down to the centimeter level. Each station is continuously receiving GPS signals and calculating its location using the same methods that we just talked about, and it'll get some location with won't match what it knows is its location. It determines that error and, for the WAAS system, it sends it back the satellites and those satellites then transmit that to the

receivers along with their other data. The receiver can now say determine the closest grounds stations, interpolate the errors they are currently reporting, and use that to correct the location they report. Some of the other systems used for surveying don't work in real time, but can post-process detailed signal data from special professional GPS receivers. These can be used for things like surveying. People use it if they're trying to find very small motion like tectonic plate movement.

Slide 31:
The last thing that I'm going to talk about is the velocity measurements. You would think so I mean the simplest method to do this is to simply say I found location A at time A and location B at time B, and I'll take the location difference divided by the time difference and get a velocity. And you can do that, but you get the best best you can get is about 1 meter per second accuracy which is somewhere in the order of two miles an hour, so it's not a really precise.

Slide 32:
But you can do better. You can use Doppler frequency shift in the GPS signals. Here I have the formulas for Doppler shift to this says if I got a velocity relative to my satellite, and this is along the line between them so I have to find the locate my location first so I can get the direction of a vector from me to the satellite. If I want to find the component of the velocity along that line, vrel here is positive if I'm moving closer to the satellite. So if vrel is positive I will get a higher frequency observed because the wave fronts are being pushed together as the satellite is coming toward me, and if vrel is negative I'll get us a lower frequency. Again, once I've got all of this I can

set up a similar set of equations, once again I need four satellites, but here I can compute three components of velocity plus a rate of local clock drift. So just like before we found the error in the time, here I'm finding a rate of change in distance and I'm finding a rate of change in my clock error so they they kind of go together.

Slide 33:
This last slide is just some references are used in putting this together. My email is on the title slide, so let me know if you have further questions.