Hey YouTube, in this video I'm going to show you how you can quickly convert any audio into text using the free open source package in Python called whisper. I'm going to show I installed it, show an example of how I ran it and compare it to an existing library. So starting off, you'll probably want to go to the whisper get hub repository that we're looking at here and they give instructions on how you can install it. Now one thing to keep in mind when you pip install just the name whisper it's not going to install the right version. We want to install from this Git repository. So just take this pip install command and run it in your environment that you're running Python. And they also mentioned here that you need FFM peg installed. There's some instructions to do it, but I already had that installed on my computer. Now that I have whisper install, let's just make some audio that I can test this on. So I'm gonna say some idioms. Idioms are usually hard for models to understand. Even though this is just speech to text. This will be kind of fun. I would love to be on Cloud 9 as a one trick pony that wouldn't hurt a fly. I'd be like a fish out of water and as fit as a fiddle to be under the weather. Let's save this off. Let's save it as a wave. They do have instructions for how we could run this just straight from the command line once it's installed. I'm gonna show you how to use the Python API, which they show here. So it's really simple. We just import whisper. Then we're gonna create our model, which is we're gonna load. model that's called base. And then just using this model object, we run transcribe on our audio file. So I named it idioms. Let's use the wave version. We want this to return the result. Now, I noticed when I ran this before, I get this error because of kuda's half tensor and float tensor. I was able to solve this. So that's something to keep in mind. If it doesn't work for you, you might need to set floating point 16 to fall. And you can see after it's run here, it detected the language already as English and then this result object has a few different Methods in them, but what we want to get inside of this is just the text and we could see that it it's looks like the result is good I would love to be on cloud nine as a one trick pony that wouldn't hurt a fly I'd be like a fish out of water and this it did mess up a little bit this fish out of water in as fit as a fiddle And maybe I didn't say it clearly enough another thing to know is when When you first run this, it's going to have to download the base model. So you might

see a progress bar going across and you'll have to download that model. And it says when you run this transcribe, it's actually taking 30 second chunks of your audio file and running predictions on it. Now there's also another approach that you can take, which is a lower level approach, where you actually create the model and then you create the audio object and pattern trim this. So you just make sure that this audio chunk is only 30 seconds. seconds or it'll pat it with 30 seconds since that's the length the model expects to have as input. Then it's making a log mouse spectrogram. It's detecting the language and we can decode here and provide a lot more options if we wanted to. If I run this cell, again get this error, which I now can set in the decoding options, FP16 equals faults. And actually this time it looks like it got everything correct. I'd be like a fish out of water. and is fit as a fiddle. So that's it for Whisper. I just want to compare it to an existing type of model. And a popular library for doing this is the speech recognition library. The way we run the speech recognition library is we import it and then create this recognizer object, which we then can load our audio file with. After that, you can take the recognizer object and there are a few different recognizing methods for that. And we're going to use the Google recognize and let's see what the result is. So it Looks like it didn't add any punctuation, and the cloud nine is different. I would love to be on cloud nine as a one trick pony that wouldn't hurt a fly. But the one thing to keep in mind is that this is actually using the Google speech recognition API. The Whisper library, you actually have the model downloaded and it's yours to use. I do also recommend you take a look at the Whisper paper, which was released with this code. They also go into detail about how the model was trained and the architecture that it's used. Whisper does work on a bunch of different languages. The performance they say varies based on the language. So you can go here on the GitHub repo where they have a plot showing which languages actually performs best for the bars here. Smaller is better and larger means it performs worse. So still pretty impressive the number of languages that this model works on.