there has been a lot of hype around model context protocol and I'm going to make an attempt to provide an extremely simple explanation of mCP today I will also go into technical details so that if you are building an AI application it helps you if you think about a of building a applications we are seeing a clear Evolution first we started with llm without any tools Etc then we started building this agentic Frameworks ET where llm gets help from tools and knowledge now we are entering the realm of standardized ways of interacting with these tools and knowledge so that building AI applications becomes easier let me give you an example say you are a equity research analyst working at some company let's say jeffre you want to develop a report comparing Nvidia and Tesla stock that looks something like this where you have company description at the the beginning then you have some uh Financial metrics then you have a summary of those metrics and in the end you have recent news about those companies now you are a tax heavy person so you will talk to your AI engineer friend working at Jeff and you will ask him to uh build an a application that can automatically generate this report now your AI engineer friends understands the capability of llm see they know that llm can pull the description of Nvidia and Tesla because it is part of that training data set but it cannot pull the latest stock price for example once you have somehow retrieved let's say the latest stock price llm can summarize uh that particular information okay so if you think about pure llm it has all these capabilities now you'll ask me okay if I go go to chat GPT and if I type this question it is pulling the information but look at this it is searching the web actually so chat GPT is an agent I am referring to GPT 40 model okay so now you already know the answer that you can use web search or maybe you can uh call Yahoo finance API which is called a tool to retrieve the latest information to summarize AI engineer will build this application where the heart is llm and llm is interacting with some tools which can be Yahoo finance API or web search it is also interacting with some knowledge you might have some private database in Jeff some PDF files lmm might want to pull data from it and he will write all this glue code okay glue code is basically the code that you're writing for these interactions now this can be an agentic application in which glue code is written by the agentic framework such as crew AI egno Etc or it can be a workflow application where as part of your python code you writing all this glue code this is just one application imagine jeffre is building 20 such application and think about all the companies in the world building millions of applications that is a lot of glue code it sounds like you have this old computer and you're connecting with keyboard mouse Etc through all these different wires but you will be like no today things have changed today you can connect computer using this unified interface USB C Port you can have your USB hub and you can connect all this devices well that USBC moment has arrived for AI and that is your model context protocol in this your llm will interact through model context protocol with different mCP servers so for our Jeff example think that Yahoo finance is building an mCP server or let's say Google search is building another mCP server where they're exposing all these tools and resources and prompts Etc and that way these interactions becomes easier now you're a smart programmer and you'll be like hey we are still writing gluc code right because you need to do all these interactions the answer is yes you are writing some glue code here but the AE of writing that code is increasing going back to our old diagram the problem here was not just writing the code but maintaining it because tomorrow if Yahoo finance changes their API you have to maintain this code right so all this glue code becomes a nightmare to maintain whereas with this mCP due to the standard protocol writing and maintaining the code has become easier and also since writing this code is centralized let's say Yahoo finance folks will write their mCP server so now all these 10,000 programmers in the world they don't have to write the code okay so they are kind of getting this redimed thing and they are saving time on building their AI applications let's go deeper into technical details now say you are building this chatboard for your organization which requires interacting with Google Maps API so let's say you're getting some location and you are automatically creating the to-do task using this app todoist and the technical teams at Google Maps and todoist have already built these mCP servers in mCP client which is your chatboard you will have a configuration which will tell you what kind of servers are available to this client when this chatboard starts at the beginning let's say just think about Google Maps server it will make a call

called list tools and this call will be made for Google Maps to doist all the servers which are listed Google Maps mCP server will return all its capabilities in form of this type of response so here you are saying that I can help you search places and you are providing this detailed description this description is very important because it will guide llm to uh call an appropriate tool llm has language intelligence so just by reading this description it can figure out that for whatever query if I have to search places I can use this map search places tool not only that from the query it will also extract the required parameters such as search query latitude longitude Etc let's say you are going for a hiking in lak and you'll say I'm going from lak to this place um you know show me the places is so from that natural question it will extract the parameter lak is a location and it will map it to Lang uh longitude and latitude and it will also uh determine that it needs to call this function now you will have all the tools you will have map place details and all the functionalities that Google Maps provides not only that you will have the tool description from other servers such as todoist as well so so once uh llm knows all these details now let's say you are asking this question that I'm going for hiking in lak and I need this place details what uh this application my chatboard will do is it will use this kind of a prompt so in this prompt this tool description is nothing but the combine tool description of all the tools you have available okay and it will say that choose the appropriate tool based on the user question so when you have tool description and this kind of nice prompt llm is smart enough to figure out which tool to call which parameter to extract from the user question and how to make a call get the response and how to read the response and serve to the end user here I have this uh mCP client from the python SDK that anthropic has provided when it starts it will go through all the servers remember that server configuration it will go through all the servers and each server it will ask list tools and whatever tools it is getting it will get the description of all those tools and it will put it here and look at this prompt okay so now you understand that llm is getting a question how it Maps or how it figure out an appropriate tool to call now let me show you the mCP server from Google Maps here it is listing the tools so when mCP client makes that request it will handle that request and it will list all the tools so you see this call okay and what are the tools so let's search for all the tools okay so search places tool you see search places tool geoc code tool okay so search places tool should be here So eventually it will come to this python function oh this is not python actually this is typescript so you can Implement your server in either typescript or Python and here from the user question it will derive the query location Etc and it will actually make an HTTP call to the Google Map API so it's not like you are replacing uh a rest protocol here or HTTP it's like a rapper and you are internally calling Google Map API and you are returning the response in a standardized format okay so there is a standard here so if you look at uh the input schema okay so let me search for input schema see input schema you see so for the search places tool you see here there is a standard way you will uh provide the description of the tool and also the query parameter Etc so see this input schema description Etc is part of this particular standard so you can find the this schema I'm going to provide all the links okay so this is the standard this is the schema that anybody who is building an mCP server will have to add her to so that way we have standard and some uh uniform or predictable way of communication okay so you look at this types script schema where you say input schema is this type required whatever just go through this schema and you will get an idea any mCP server will expose three capabilities tool resource and prompt their python SDK has simple examples for each of them so let's look at the tool this is a simple server with one single tool okay so if you look at the list tools function see list tool tools here it is exposing see this is an array okay so it is exposing a single tool called fatch and there is this standard description standard input schema and so on if you look at the implementation of fat it looks like this okay so you are fetching a website okay so here see you are fetching a website so whenever that fatch tool is called you call this function and you are just retrieving some information by making an HTTP call okay so this is pretty straightforward the second capability is a resource resource is um some kind of knowledge okay database files Etc and similar to list tool functions it will have list resources so when the mCP client starts it will call list tools list resources list prompts for each of

the servers so it knows the full capabilities of all the servers that it has available or has access to so in the list resources you can have a file see this is a plain file very simple example you can have a file in your uh some drive or some you know like Amazon S3 Etc you can also have a prompt so just imagine you are building mCP server for Yahoo finance as a developer you know all the prompts that AI Engineers uh might need to interact with my API okay so you will provide all those prompts VI your server so writing prompts become very easy for the mCP client okay so here you are providing one single prompt so once again you have list prompts you are providing all the prompts see this is an array and this is the simple prompt that takes context and topic and if you look at the Imp mentation it's pretty simple you have context you have topic and you are creating the prompt using that context and topic folks that's it so that is what is model context protocol I'm going to provide documentation Etc so you can read through more details there has been a lot of hype but I believe we are in early days this has a lot of potential but how this is going to evolve and how this is going to help AI Engineers solve the real problems is something that we will know as time goes okay so some people are super excited well I'm excited too but just understand that we are in early days there has been lot of hype there is some reality we'll have to see how this thing evolves we are going to come up with few more technical tutorials uh on this so I'll be building some actual servers and clients using mCP if you have have any question please post in the comment box below thank you for watching