So why is the game of Jeopardy hard for a computer?

We said earlier that Watson had 3,000 processors and a database of 200 million pages of information.

So shouldn't it be easy for Watson to play Jeopardy?

Unfortunately, Jeopardy has a wide variety of categories which are purposely made cryptic.

While computers can easily answer precise questions, like computing the square root of a complicated number, understanding natural language is hard for computers.

As an example, suppose we ask Watson to answer the question: Where was Albert Einstein born?

Stored in its database, Watson might have the following information: "One day, from his city views of Ulm, Otto chose a watercolor to send to Albert Einstein as a remembrance of his birthplace." As a human, we can probably figure out that the name of the city is Ulm, but this is a hard sentence to parse.

How would you tell a computer that Albert Einstein was born in Ulm using just this sentence?

OK.

So how about if we just store answers to all possible questions that could be asked on Jeopardy?

Unfortunately, this would be impossible.

An analysis of 200,000 previous Jeopardy questions yielded over 2,500 different categories, and new questions are created on Jeopardy all the time.

Well, OK, then let's just search Google for the answer to the question.

Unfortunately, no links to the outside world are permitted on Jeopardy, and this rule applied to Watson as well.

And even if Watson could search the internet for the answer to a question, it can take considerable skill to find the right web page with the right information.

So instead, Watson used analytics to answer the Jeopardy questions.

Watson received each question in text form.

Normally, the players see and hear the questions, but Watson couldn't hear anything, so they decided to feed him the questions in text instead.

With the question in text form, IBM was able to use text analytics and other analytical methods to make Watson a competitive player.

Overall, they used 100 different techniques for analyzing natural language, finding hypotheses for the questions, and ranking these hypotheses to pick an answer.

Watson had a huge database of sources and several basic tools to help understand language.

The database consisted of a massive number of data sources, like encyclopedias, texts, manuals, magazines, and downloaded pages of Wikipedia.

One of the tools Watson had was a lexicon, which describes the relationship between different words.

For example, the lexicon could tell Watson that water is a clear liquid, but not all clear liquids are water.

Another tool Watson had was a part of speech tagger and parser.

This would identify functions of words in text.

For example, it would know that the word "race" can be used as a verb or a noun.

"The students had to race to catch the bus" uses race as a verb, while "Please indicate your race" uses race as a noun.

Then, using these data sources and tools, Watson would answer a question by going through four major steps.

The first step is question analysis, where Watson tries to figure out what the question is looking for.

The second step is hypothesis generation, where Watson searches the information sources for possible answers.

After this, Watson moves on to step three, when the different hypotheses are scored.

This means that a confidence level has to be computed for each answer.

The final step is ranking the hypotheses to look for a highly-supported answer.

In the next two videos, we'll go through how each of these steps work.