Introduction to Large Language Models (LLMs)

Prof. Tanmoy Chakraborty, Prof. Soumen Chakraborti
Department of Computer Science & Engineering
Indian Institute of Technology, Delhi

Lecture 35

Knowledge and Retrieval: Temporal Knowledge Graphs

Next, we come to the important issue of time in knowledge graphs. In other words, temporal knowledge graphs. Everything changes. Honolulu is a city of USA, but even this is not an eternal truth. It couldn't have happened before the USA came into existence, for example.

Barack Obama was born in Honolulu at a certain point of time. Barack Obama was president of USA through a particular time range. During a disjoint other time range, George W. Bush was president of the USA. And Barack Obama has been spouse of Michelle Obama since 1992.

As with many other things in the universe, all of this pass. Everything changes. Change is the only constant. So far, the representations of knowledge graphs that we have discussed have no notion of time. It's as if every fact holds throughout eternity.

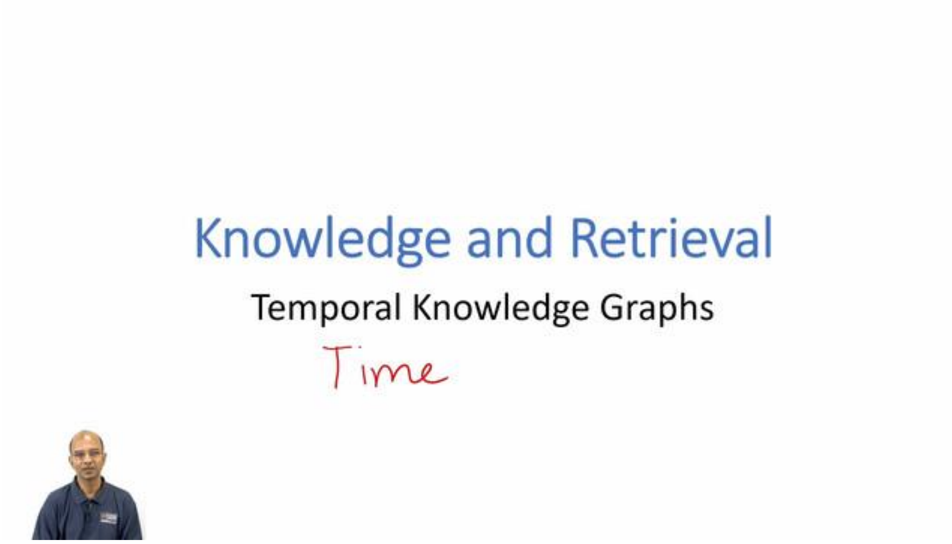
In this part of this module, we are going to discuss various ways in which temporal information can be incorporated in knowledge graph representations and inference. Here are the tasks which we might want to solve if we are given temporal knowledge graphs. For simplicity, we will limit our representation of time to either a single point of time, a single interval of which a point is a degenerate case, or recurrent instance through time. That's an important type of thing. For example, an Olympic game happens every four years, et cetera, et cetera.

So a temporal tuple or a temporal fact has subject relation and object as before, but it also has a time interval given by a beginning and an end. If the beginning and end are exactly the same time, then it's like an instant. At what granularity are we going to measure time?

Milliseconds or millennia? That depends on the application. Milliseconds may be more appropriate for Olympic swimming records, whereas millennia may be more appropriate for the rise and fall of civilizations. For example, Barack Obama was president of US starting in 2009 and finishing in 2017.

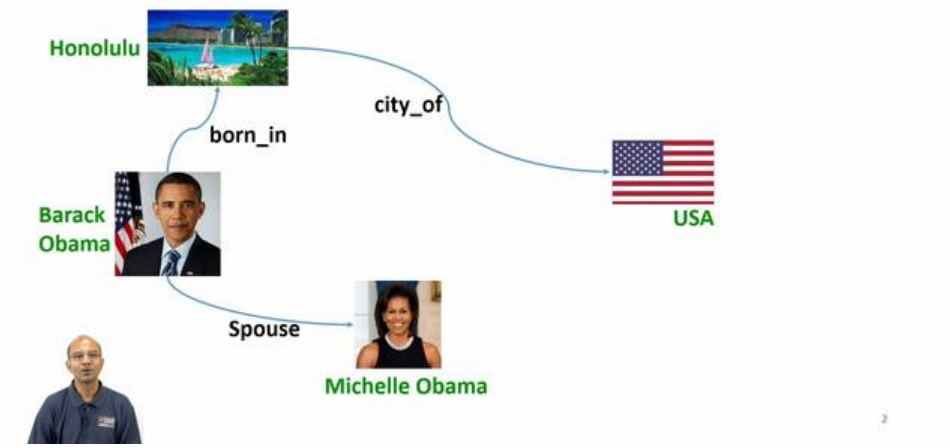
There are various link prediction tasks. For example, just like in standard knowledge graph completion, I could leave the subject unknown or the object unknown. In each of these cases, I may provide positive and negative examples. For example, if I say Barack Obama, president of between 2009 and 2017, but I mask off the USA or I provide USA as the fact, that's a positive fact. Whereas if I provide India in place of the object in that same time interval, or for any time interval, that would be a negative example.

Similarly, we could in fact keep the times the same, but one of them has the correct precedent, the other has an incorrect precedent. So now link prediction gets much more complicated because we are providing the temporal information as well. There's also now the new time prediction task, where I might give the subject relation and object. All three of them are given, but I don't know during what time this relationship holds. I could say Barack Obama, President of USA, and I could ask for the interval and 2009 to 2017 would be correct, 2001 to 2009 would be wrong.

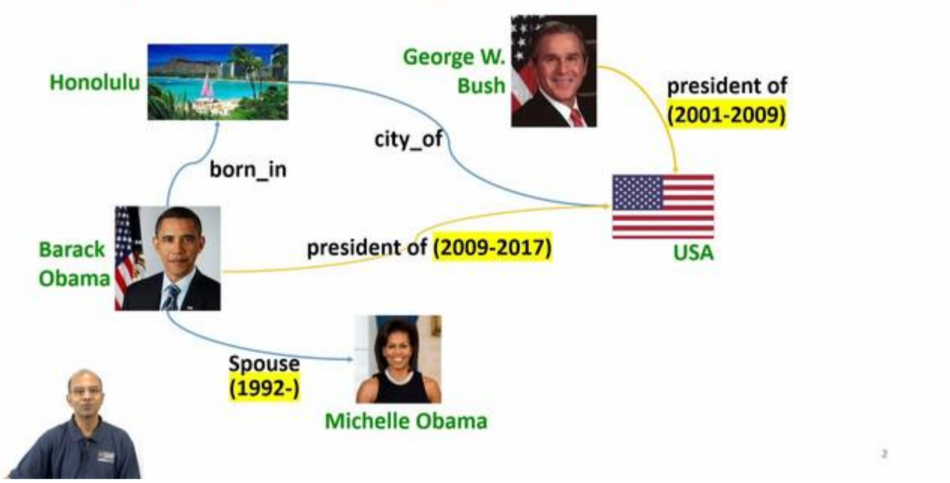


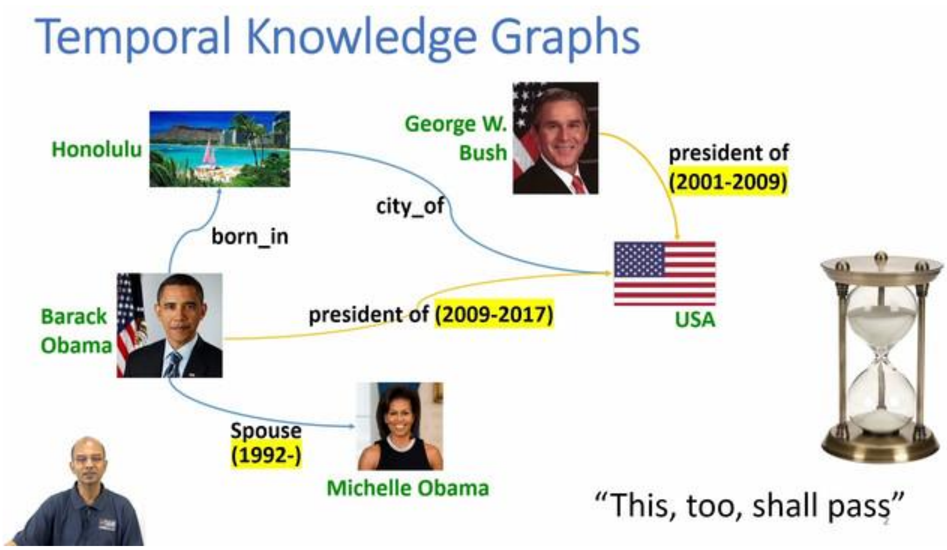
So these are now the new added kinds of tasks in case of temporal knowledge graphs. How do we extend knowledge graph completion to temporal knowledge graph completion? Recall that earlier subject, relation and object or their representations were inputs to my scoring function f and I got as output a belief in the fact. Now, I also have start and end times as potential additional arguments. Remember that any of these five could be missing as inputs and then you have to maximize the belief with regard to choices of the missing input. But how do these start and end times fit in? How should F be redesigned or extended? Where should time start interfering with the workings of F? Should it modify the inputs, the subject relation and object? Should those get time varying representations? Or should subject relation and objects have time independent representation and only F starts recognizing the start and end times? These are very difficult questions and these are not merely of academic interest because even fairly sophisticated initial attempts at designing a temporal f had very serious shortcomings.

Temporal Knowledge Graphs



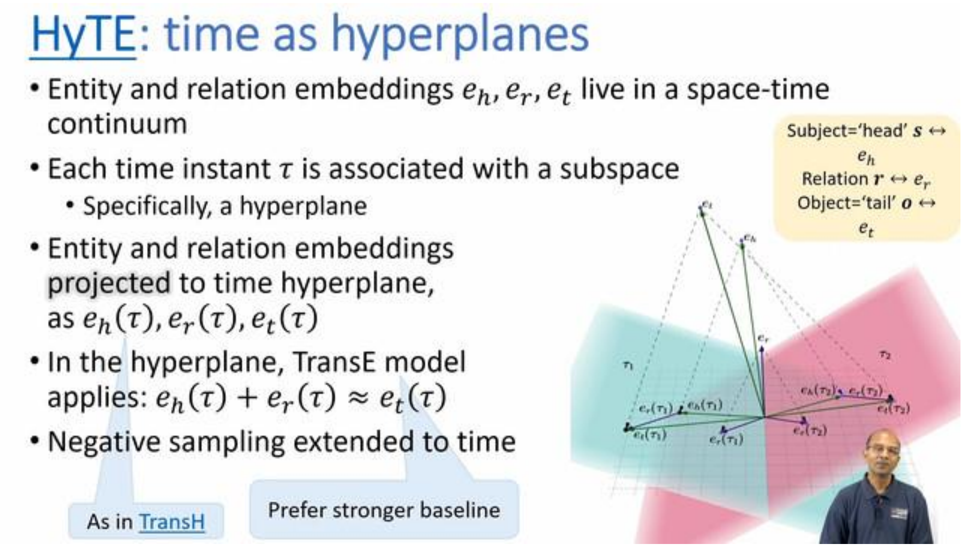
Temporal Knowledge Graphs



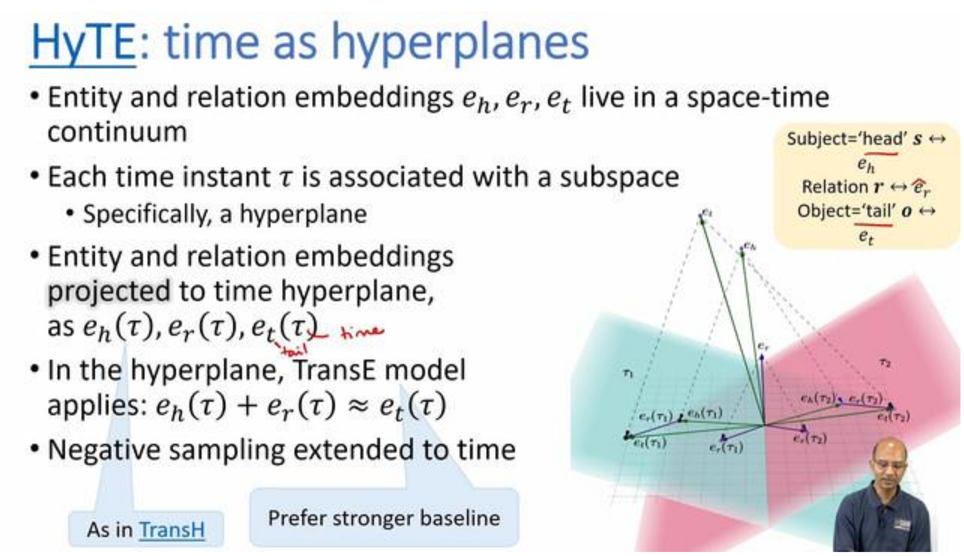


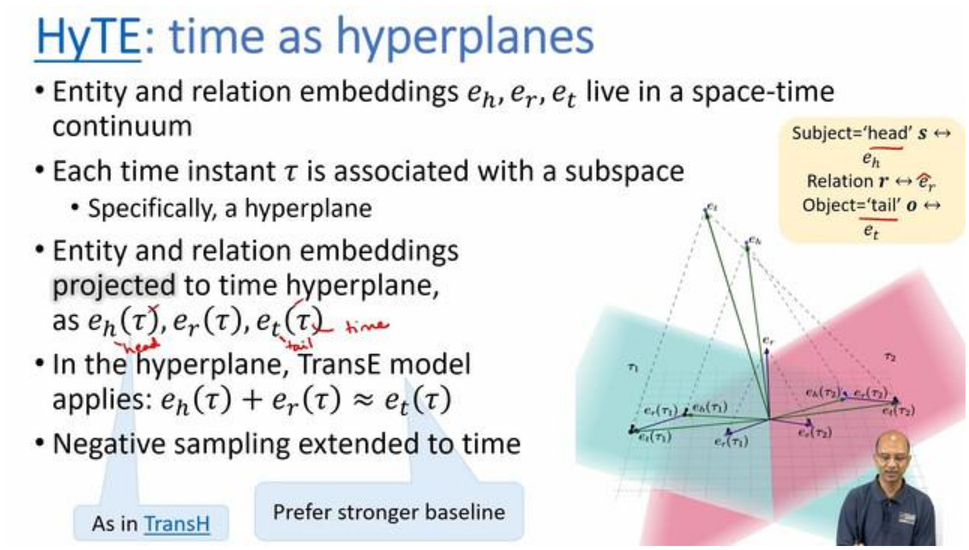
For example, in the training fold we might have a triple or now a quintuple which is Shin Ae Ra is married to Cha In Pyo, these are all celebrities in Korea, from 1995 onwards. That's already known in the train fold. In the test fold, there is a query, Shin Ae Ra was born in which year? A recent temporal knowledge graph completion method responds with 2013. So this celebrity was born in 2013 but got married in 1995. That's a bit anti-causal which happens in movies but it's a real difficult to accept in real life.

A very early attempt to mitigate these kinds of nonsense outputs or inferences was HyTE. HyTE, as the name suggests, used hyperplanes to model time. So every instant of time was modeled using a hyperplane as is shown in this diagram. Entity and relation embeddings would be projected to the time hyperplane. Now here in this slide alone, because I'm reusing diagrams from the paper, the subject is also called the head and the object is also called the tail.

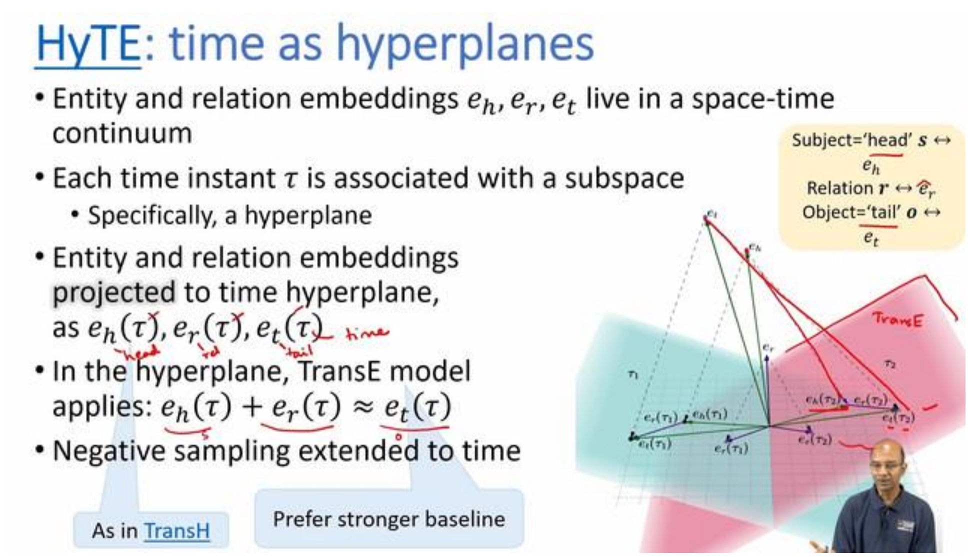


The time is represented by a unit vector ET. So tau is the time, right? And ET is the representation of the tail entity, EH is the representation of the head entity, all at this time. And similarly, the relation has a representation for the given time, tau. Pictorially, consider tau two as a time. Here is my head entity.



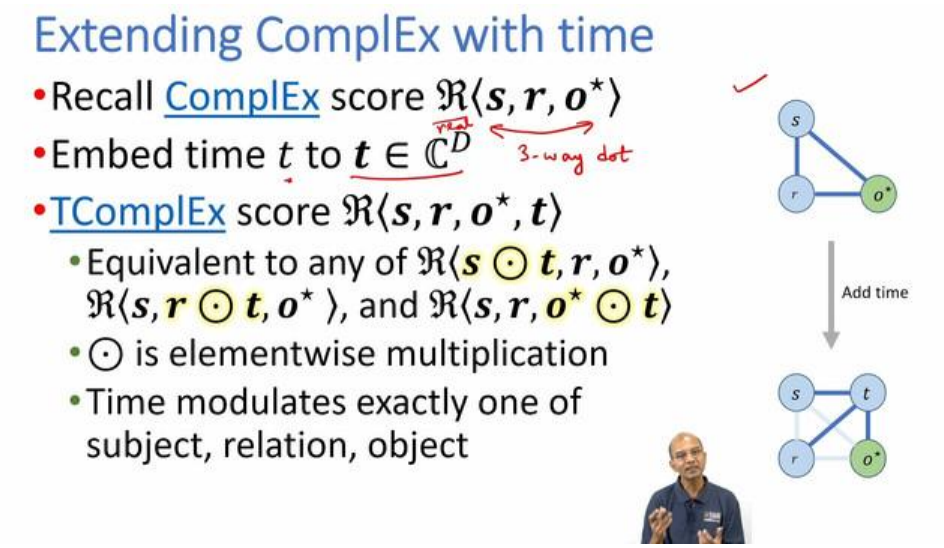


It gets projected down to the head entity's representation at the time tau 2. Similarly, the tail gets projected to the tail's representation at time 2. And on this hyperplane, the pink hyperplane, I expect trans E to take effect. Recall that in trans E, the head or subject plus the relation should approximate the tail or the object. But now that happens not in space in general, but only on the plane of the time point that I'm interested.



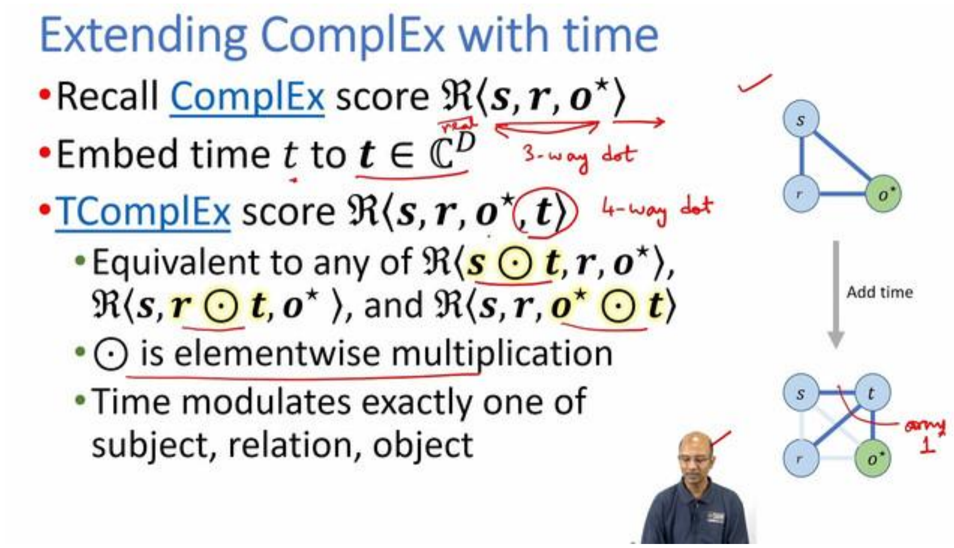
At a different time point tau one, these relations might be very different. A different relation may hold between the head and the tail, okay? Now, this is all very nice, but transy was a very weak model, if you might remember, that it couldn't support many-to-many relationships. And so we prefer a stronger baseline. So let's now jump over to something like complex and see how such things can be enhanced with time. Recall that in complex, the scoring function was the real part of the three-way inner product between the subject, the relation, and the conjugate of the object vector.

This is represented by this upper diagram, which says that there is all-to-all relationship between the three of them. There is a three-way inner product. And that's represented by this triangle on the right. Now that we have time like high t but generalizing it, we are going to embed time to another complex vector in the same dimensional place c to the d. How can we design a score which now takes complex vectors bold s, bold r, bold o and bold t and puts them together? T-complex takes the very first step.



It says let me just do whatever was done earlier, but I'm going to extend it to a four-way dot product, which also takes in times complex vector as a fourth argument. As a simple homework, you can show that this four-way dot product is actually equivalent to any of these three forms. In one, time interacts with only the subject as a elementwise multiplication. In the other, time interacts only with relation or time interacts only with object conjugate. And this is shown in the lower picture where we show that the triangle between S, R, and O star is made more faint, but we now have time possibly interacting with any one of them.

So, any one. and that's what T-complex does. In words, time modulates exactly one of subject, relation or object. However, this gives the impression that all relations are subject to temporal changes and that might not be the case. Some relations could indeed be timeless or time independent.



TNT-complex implements this extension. it mixes the two parts. So either it does what T-complex was doing, where we are sensitive to the time, and in the second part, we do standard complex, where there is no participation of time. RT is a representation of the relation in a time-sensitive manner and R is a time-independent representation of the same relationship. Again, one can show as a simple exercise that this formula is equivalent to that formula where we first take S, we combine RT with T using an elementwise product, but we add on the timeless or eternal aspect of R, and then we take a three-way inner product or dot product like we used to do before.

So T acts in a very minimal way on only the R part, not the S or the O part. Training reveals what kind of relation R is. If R is strongly time varying, then RT will turn out to be the informative part of R's representation, and R will dwindle to nothing or a very low noisy norm vector. Whereas if R is actually timeless, then R will carry the brunt of the representation burden, whereas RT will become either meaningless or low norm. TNT complex is generally better than high TE and this is not surprising because high TE depends on trans E as a baseline whereas TNT complex depends on complex which is much better than trans E.

However, we can go further than this and improve time sensitivity of complex even more. Earlier, you could think of R as a relation with judges whether S and O can be related by it, independent of time. If you have taken a course on graphical models, you might recognize the style of this diagram. You might call this blue square a factor, and these are variables. In this case, of course, they are variable vectors.

So this factor is represented as can S participate with O involving relation R. And so this factor gives you the earlier score as per complex. But now we add on many more factors. Can S participate in a relation R at time t at all irrespective of O? Can anyone be president of any organization at age 10? Unlikely. So who judges that? This second factor.

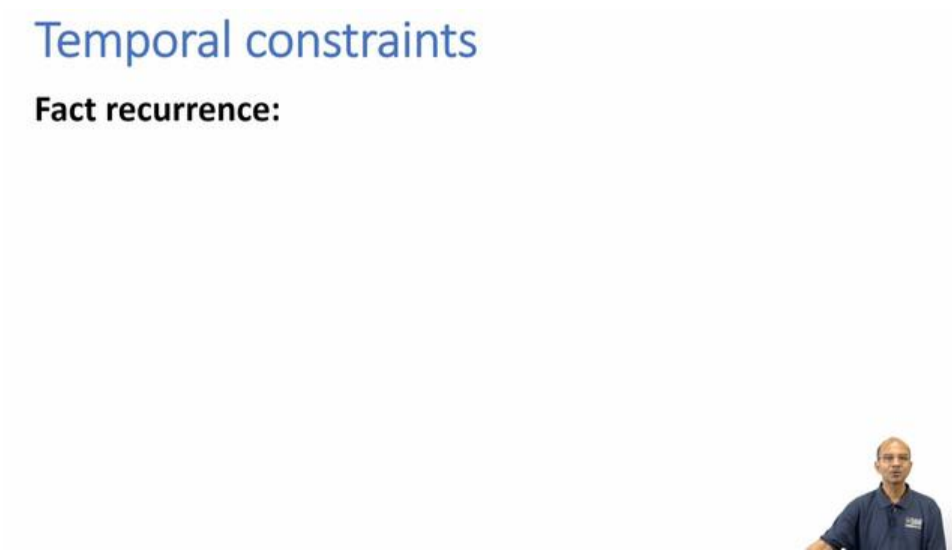
So RST, a second representation of the relation after RSO, judges if S can participate in that relation at time T, independent of O. There is no connection to O here. Similarly, can O participate in a relation R at time T? And that's judged by a representation ROT. Can anyone have any relationship to the company Google in 1980, where it was not even born? No. That is what this factor judges.

Finally, can S and O interact via any relation at time T? That is this last factor. For example, could Mozart and Elvis Presley interact in 2005? No, because both of them are dead by that time. So that's the four factors. And these four come together to make a judgment of the confidence in the overall fact. We reuse these three factor scoring functions from complex.

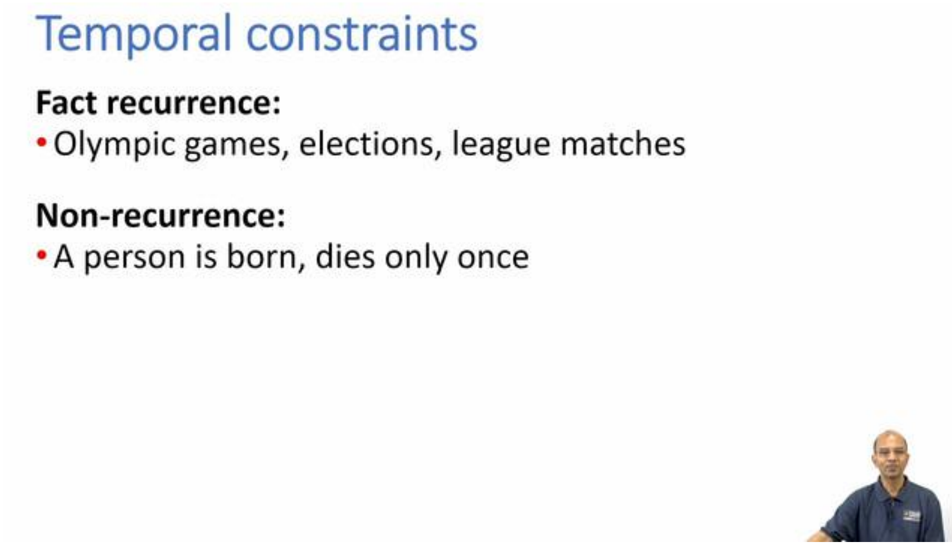
So each of them looks very similar to complex, except so all of these have the real parts. but these are weighted by these hyperparameters, the weight hyperparameters, and these have to be balanced through cost validation. This model is already better than both T-complex and TNT-complex, but there are actually two more signals still to exploit. First, it's often said that history may not repeat, but it rhymes. If you look at all the years in which Winter Olympics are held, they all differ by four years.

And this should give us enormous confidence that no Winter Olympics was held in, say, 2015. I don't need to know anything else about the world. But there are softer versions of this. From births to getting married, Obama spent 31 years. From birth to his first marriage, Einstein spent 24 years.

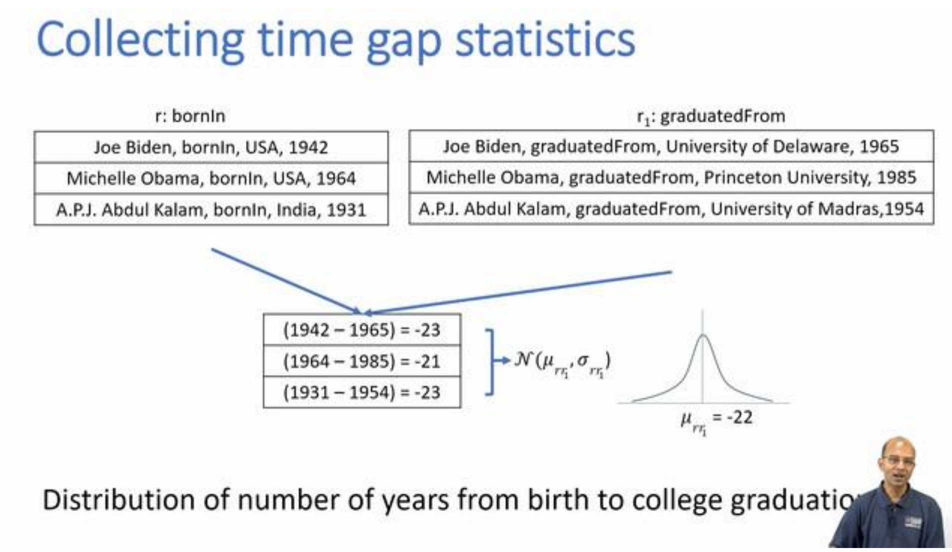
From starting higher education to getting their highest distinction in life, Barack Obama took 20 years and Einstein took 28 years. Given the nature of human life or even the nature of organizations, of events, there are strong regularities between gaps like these. Periodic events have almost exact gaps, but others may be subject to person-to-person or organization-to-organization variations, but there are strong regularities in this. And we should be able to use that to avoid those kinds of blunders that people are married before they are born. So what kind of temporal constraints are we talking about? There could be recurrence, as we mentioned, Olympic games, elections, league matches, championships.



or non-recurrence. A person is born and dies only once. There are, of course, important fictitious exceptions to this. And then there are orderings between relations. You have to be born before you marry or die.



And then there are distributions over time gaps between the relations. Some person died in minus the person was born in is typically 70 years, but this depends on which country which century of human civilization, et cetera. So there is variability and our model should be robust to that. I won't describe the model in detail, but I'll give the gist and the intuition behind it. Suppose we collect two tables.

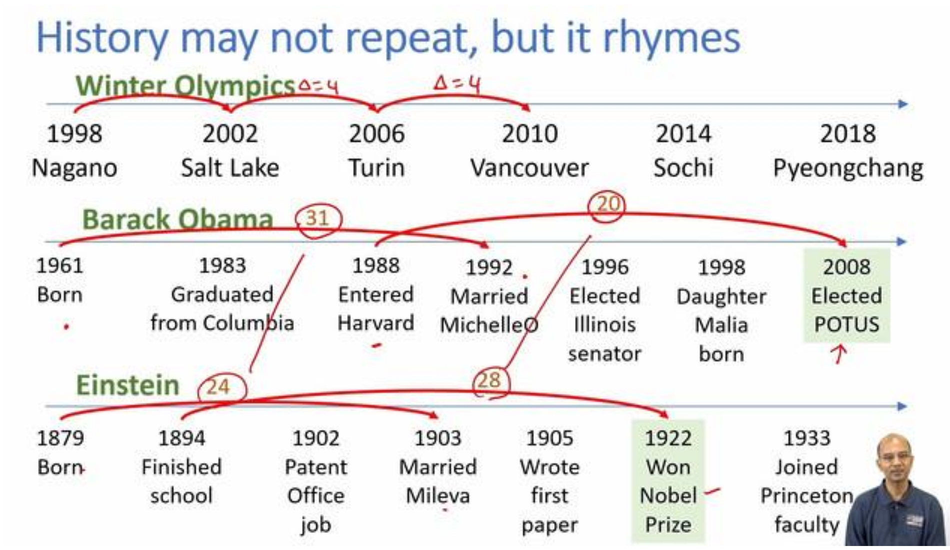


One table pertains to the born in relationship. The other is the from the graduated from relationship. So we know that Joe Biden was born in USA in 1942, APJ Abdul Kalam was born in India in 1931. Meanwhile, from the graduated from table, we see that Joe Biden graduated from the University of Delaware in 1965, whereas Abdul Kalam graduated from the University of Madras in 1954. Now we construct a table of the differences between these years.

So for Joe Biden, we compute 1942 minus 1965, and we get minus 23. For Abdul Kalam, we get 1931 minus 1954, also minus 23. And we model this for simplicity. You can do much better, I'm sure, but as a first exercise using a normal distribution. If we do this for real tables, we might get a mean of minus 22 and some standard deviation around it.

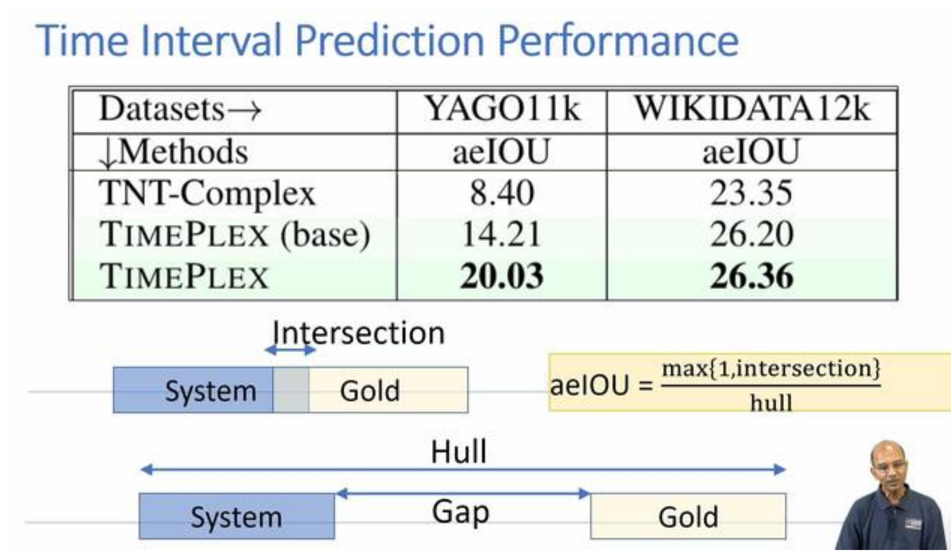
So this is a distribution of the number of years that elapse from birth to college graduation of the people in my knowledge graph. And I can do this in principle for pretty much all pairs of temporal tables in my knowledge graph. Given this kind of data set and given this kind of distribution, if I'm asked, how likely is it that Obama was born in 1961? Suppose I want to give a likelihood for this. Well, a human would say, tell me something more about Barack Obama and times of those events. When did he graduate from college? When did he become president? Now I'm going to use my distributional knowledge from other people to decide how likely it is that the gap between a proposed birth year and known graduation from and precedent years fits my pre-trained distributions, those normal variables, normal distributions.

And I can put them together through some suitable network and come up with a confidence that Barack Obama was indeed born in 1961. You can find the details in the papers. Similarly, we can also model recurrence. In case of recurrence, we know that there is this significant blip in likelihood, almost like an ECG trace. And now we can judge, if a proposal is made that an election was held in the US in 2016, well, am I getting blips at 2020? Am I getting a blip at 2008? If so, 2016 becomes more likely because there are multiples of four years.

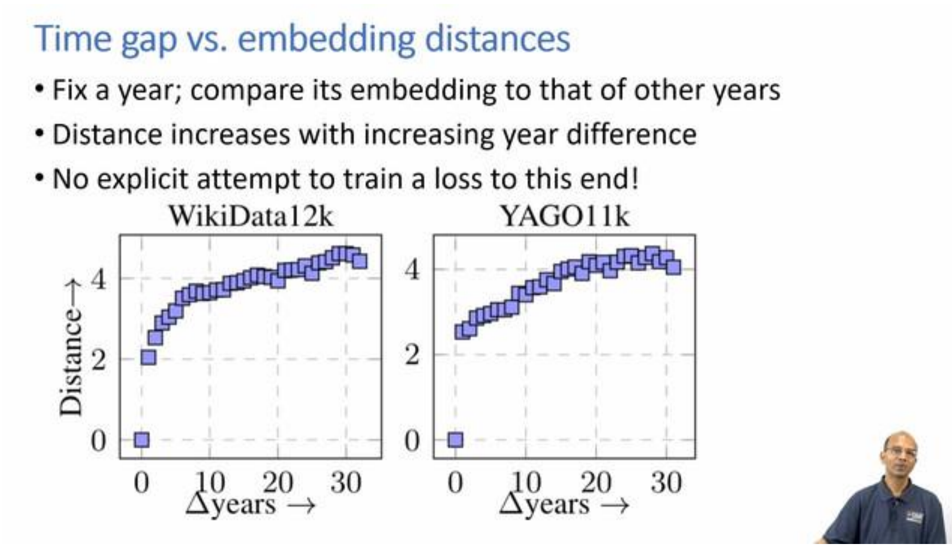


We can model all of this and combine these three kinds of signals. We can have the timeplex baseline, remember our molecule like diagram, and then we can have plausibility from the recurrence model of a single relation and we can have plausibility from relation pairs R and R prime all over the knowledge graph. Various pruning heuristics will be required to keep our complexity in check. Overall, we can train it from gold, positive and negative facts and those can back prop into all the three branches and train our older models as well as the mean and standard deviation of our temporal distributions. And this really works magic.

Even the timeplex base which is the molecule type diagram does better than TNT complex and complex of course but the full model with these temporal constraints works out far better than things we have seen before. The evaluation measure now becomes non-trivial and has to be designed carefully. I won't have time to get into that. Also, in case of time interval prediction, performance has to be measured much more carefully than just MRR. MRR is now a bit more meaningless because the system output and the gold output in case of time are intervals in general.



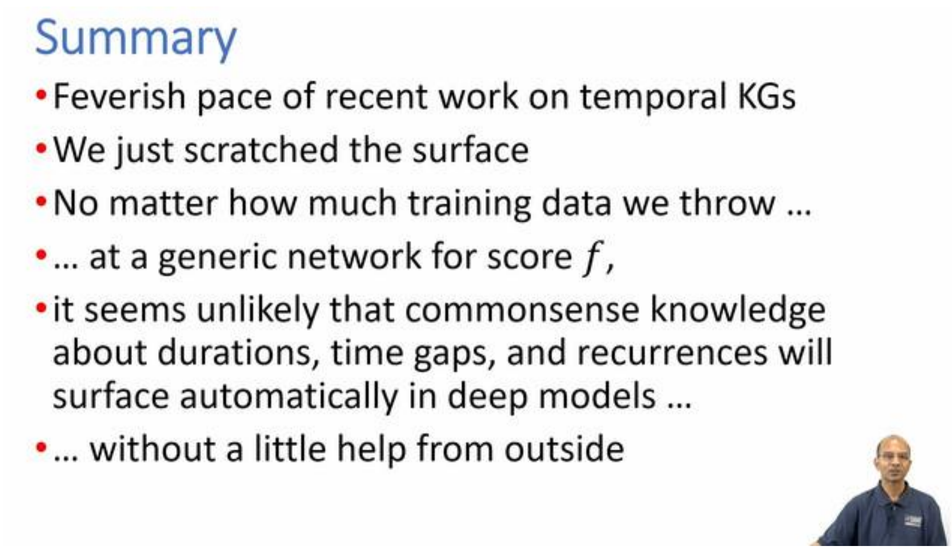
So what is the intersection between the system interval and the gold interval? We might be tempted to deal with it similar to our box or rectangle embeddings, but that actually doesn't work because again here, we would like to penalize as the gap increases the score of performance of the system. And there's a new measure called AEIOU, or Area Enhanced Intersection Over Union, that you can find out in the paper. But the more interesting experiments are like this. Suppose we fix a year and find its embedding, complex embedding. and we compare its embeddings to other years.



But remember, these years have numeric years attached to it. Maybe this is 1960, and that is 1980, which is 20 years off, and here is 2000, which is 40 years off. What happens as the gap in years increases with regard to the vector distance between these? Let's say L2 distance. between these vectors. As we can see here, quite remarkably, although no explicit attempt was made to actually foster this kind of behavior, as the gap in years increases, the L2 distance between the vectors also increases.

And this is quite remarkable. Next, we come to what use all this is. If we now ask, in the training, knowledge base, there is already a labeled triple or quadruple, saying Shin-era is married to Cha In-pyo from 1995 onwards, and now in the test fold, we ask what was the birth year of Shin-era in South Korea, but I'm looking for the birth year. If you look at standard TNT complex, despite trying to model time explicitly, its distribution is very flat. Whereas, because of our much more detailed modeling of how time goes into the complex knowledge graph model, we see a clear peak of likelihood where Shin-era might have been born.

And in fact, our prediction will be much earlier. Approximately 1965 or 1967, whose AEIOU score is much better than TNT complexes, absurd prediction of being born after getting married. So the summary so far is that there has been a feverish pace of recent work on temporal knowledge graphs. Here we just scratched the surface. No matter how much training data we throw at a generic network like complex for scoring F, it seems unlikely that common sense knowledge about durations, time gaps, recurrences, and other temporal constraints will surface automatically in deep models without a little help from outside.



And I have given a vignette of a few models which try to incorporate time representations in a much more delicate way in the earlier champion knowledge graph completion models such as complex.