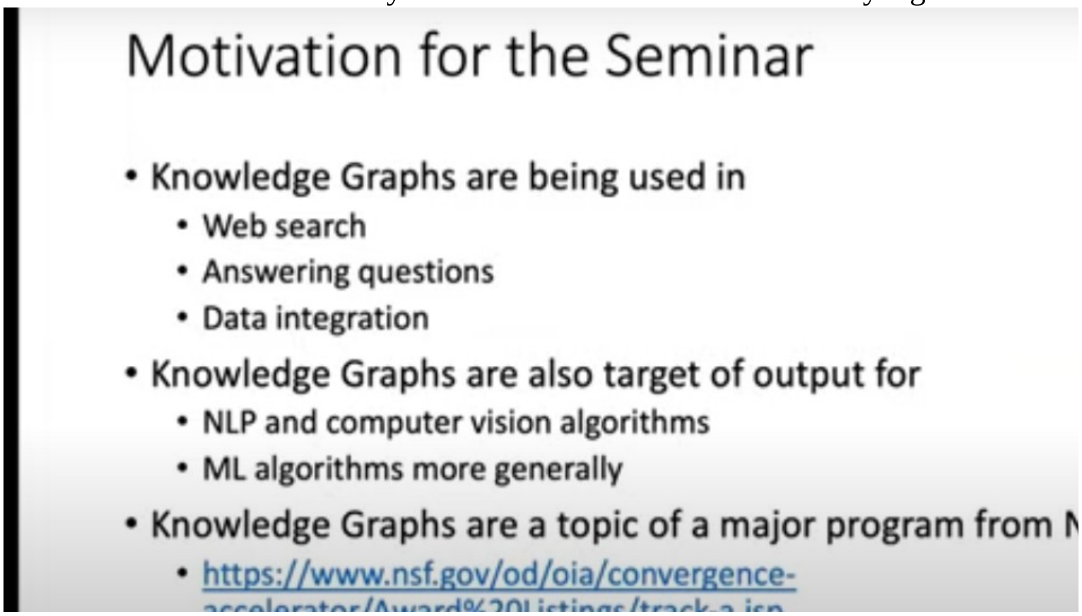
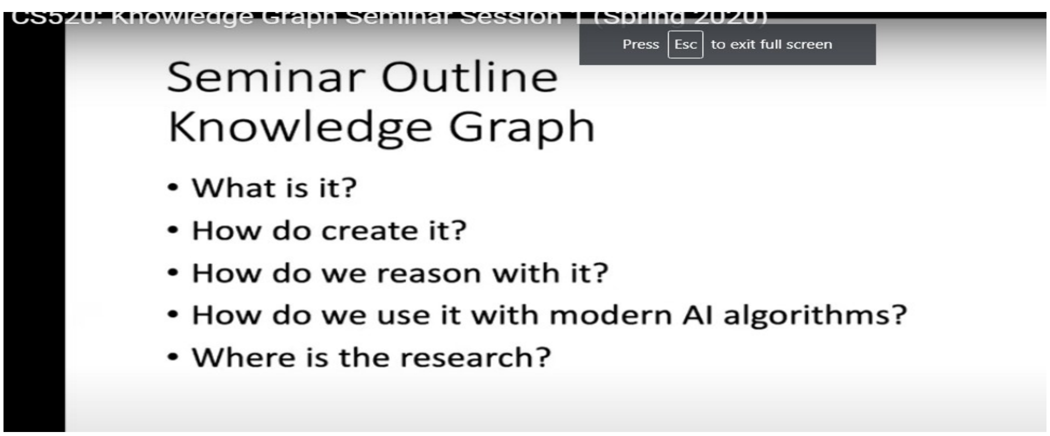
Notes from the CS520: Knowledge Graph Seminar (Spring 2020)
Knowledge 520 course Graphs:How should AI explicitly represent knowledge?
Gary Berg-Cross

Focus and Motivation: What is a knowledge graph (KG)?

Motivation – KGs now widely used. Used to store info extracted by algorithms.



Outline:



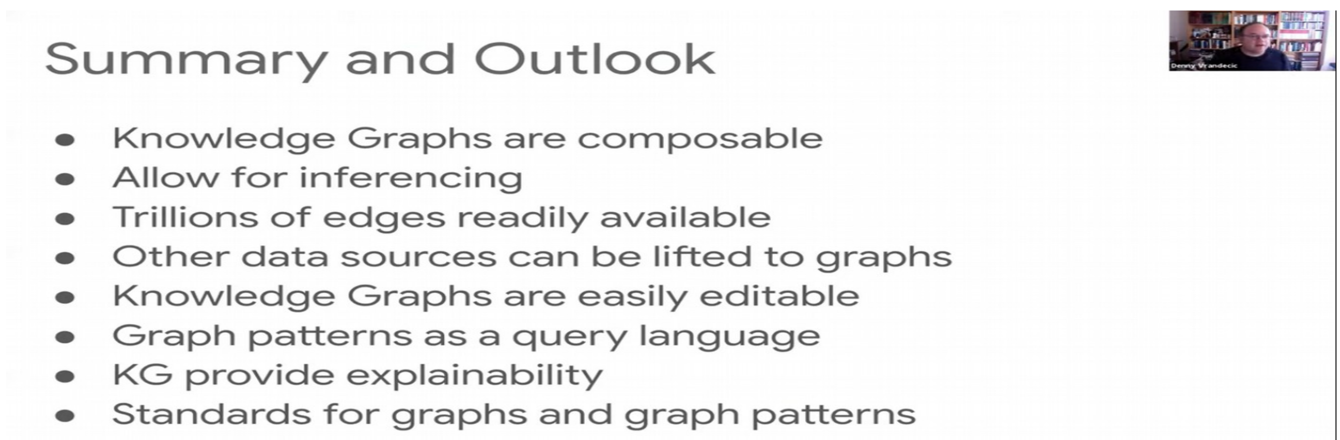
CS520: Knowledge Graph Seminar Session 1 **What is a knowledge graph?**

Denny Vrandečić Knowledge Graphs have found wide usage in many applications in various industries, for diverse research tasks, and by increasingly also by hobbyists and

student developers. In this talk we will informally introduce the ideas behind knowledge graphs, show use cases and applications, and how they have proven to be useful
We build a KG because: It is easy for knowledge to be hidden, get lost in natural language, if you have lots of text.
Need to data information that is actionable.

Intro to Wikidata which supports Wikipedia.
Shows connectivity of the graph say for cities of the world.
8 billion nodes, connected using RDF standard.
LoC also has a large KG using RDF.
More than 4800 catalogs connected. 12 million publishers of RDF. Using schema.org
So what can we do with this? Can as to give every node X connected by relation Y
Graph patterns.
With a simple query you get the geo-coordinates of every x connected by y.

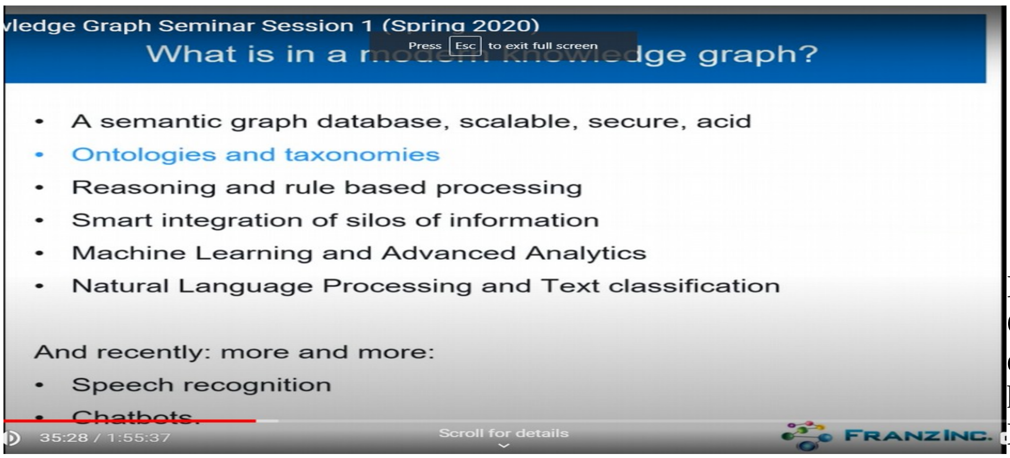
Note, what we know of the world is inconsistent, so Wikidata allows this which amounts to alternate models of reality.



Jans Aasman CEO Allegro-graph

We will discuss a Knowledge Graph technology approach that encapsulates a novel Entity-Event Model, natively integrated with domain ontologies and metadata, and dynamic ways of setting the analytics focus on all entities in the system (patient, person, devices, transactions, events, operations, etc.) as prime objects that can be the focus of an analytic (AI, ML, DL) process.

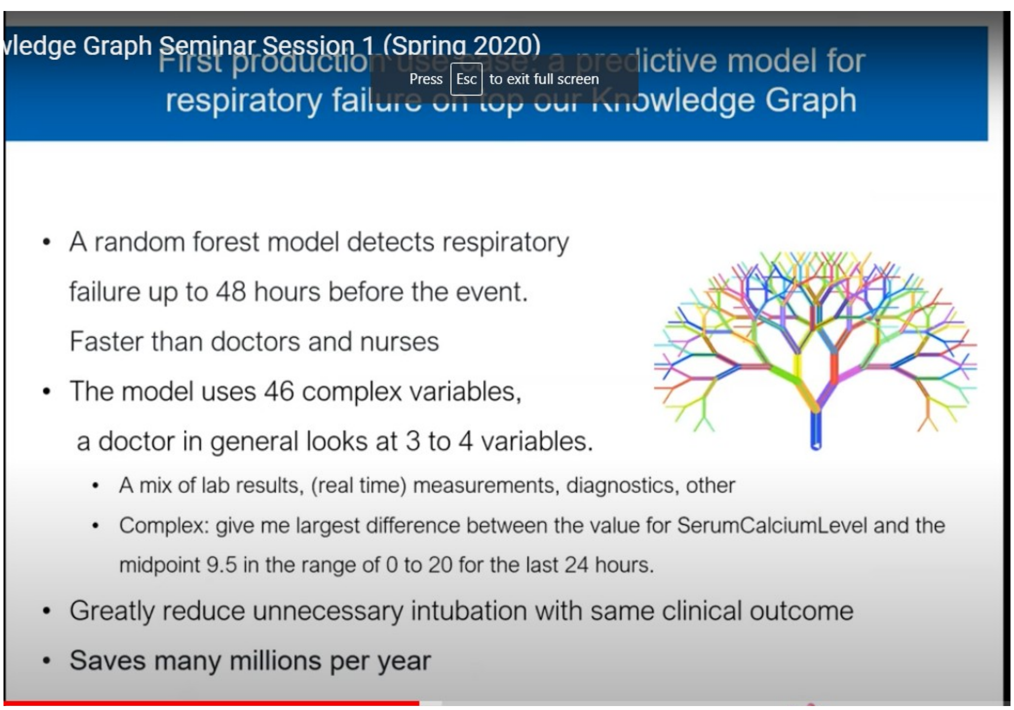
We will demonstrate core elements of a knowledge graphs --- its knowledge representation and its core principles, example of one or two knowledge graphs that show how modern knowledge graphs are a mix of graph database technology and NLP and ML techniques.

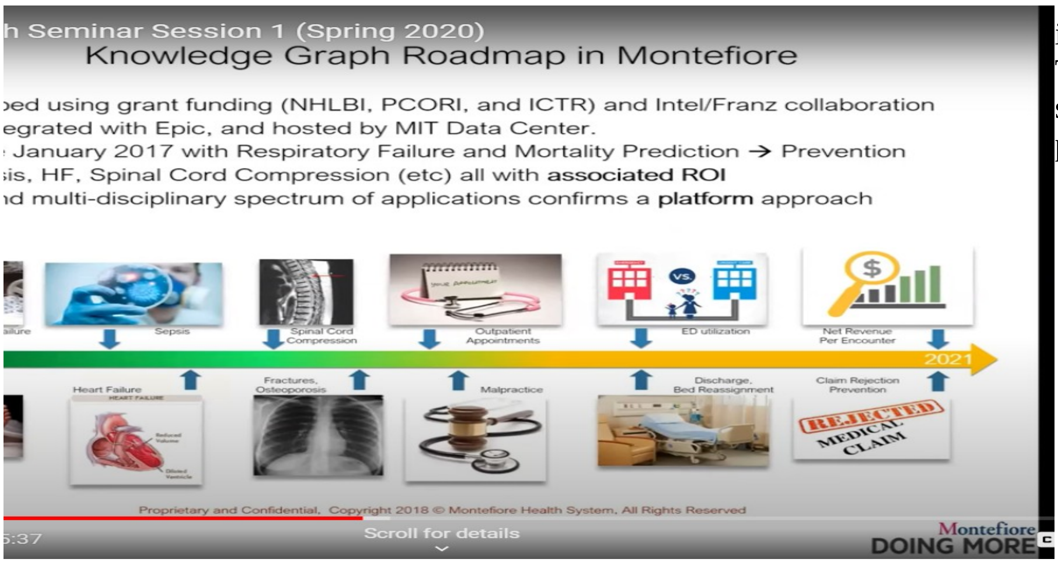


Example of a Chomsky KG demonstrates how to build a KG from text on Chomsky's publications.

Has Gruff graph DB

Large HC an events based KB.





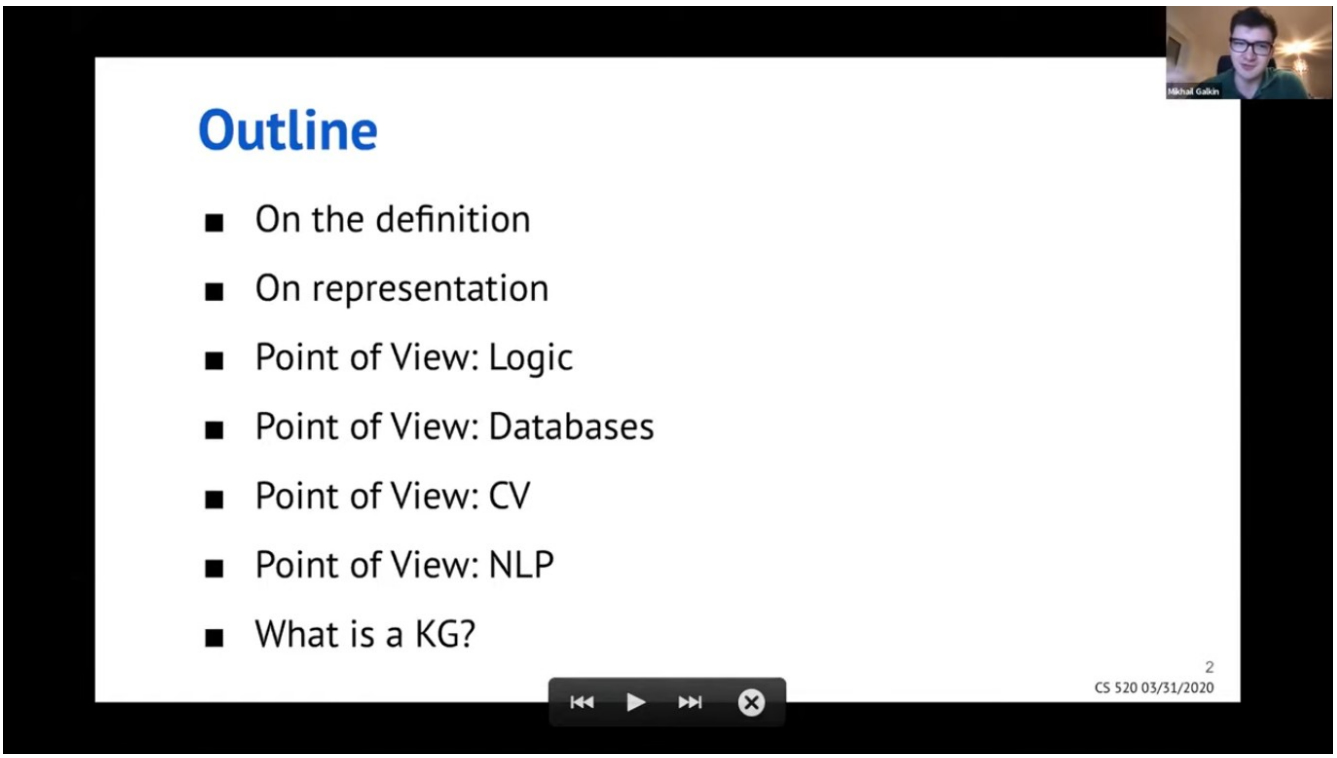
Turn everything into an event. They have the same data shape pattern.

Mikhail Galkin (Dresden) Outline: In this talk we will discuss various perspectives and points of view on understanding knowledge graphs as seen from the recent ML and AI conferences [1-4]. Finally, we will identify common attributes that pertain to knowledge graphs.

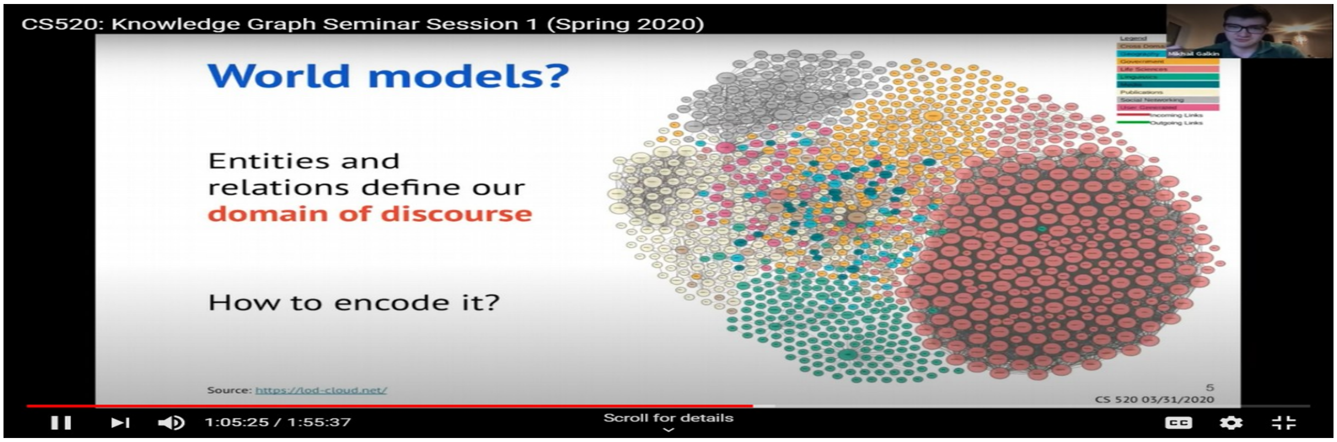
The slides for the presentation are available here.

1. Knowledge Graphs at AAAI 2020
2. Machine Learning on Graphs at NeurIPS 2019
3. Knowledge Graphs at EMNLP 2019
4. Knowledge Graphs at ACL 2019

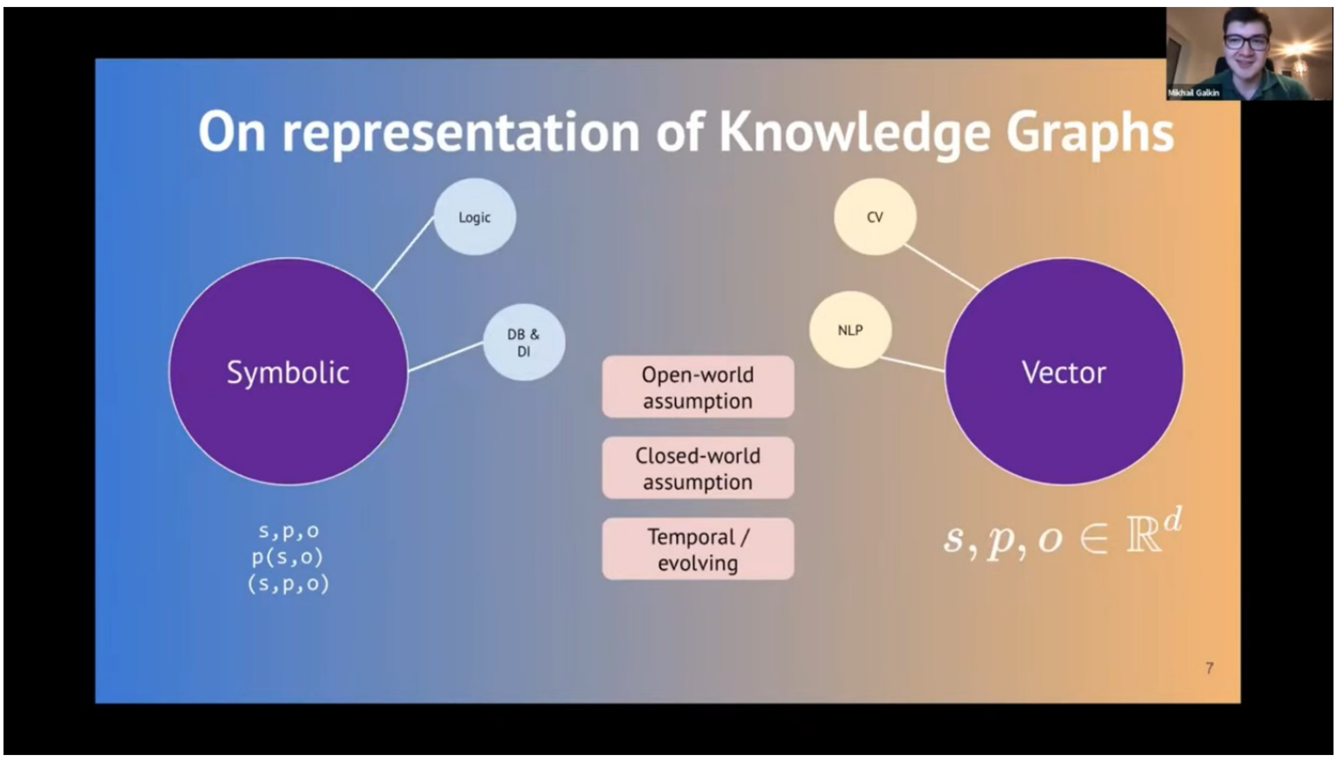
What do we need to reach a shared understanding of KG definition" Some are mathematical all are correct.



World models some with shared vocabularies.
How do we encode these? Symbolic (tiples) or vector (embedding in high dim space).



Might be open or closed world.



There are different points of view like logic using logic programming. Consistent set of facts.
Or the logic of DL
Ontology is at the T box level while KG is at the A-Box level.

Dbs get integrated into a KB, gives them power.

More AI (CHH + RPNPOV: Scene graphs build related objects identified in an image or series of images.
Kgs can be built from text.

Named entity recognition (apple as a suface form label has many meanings)

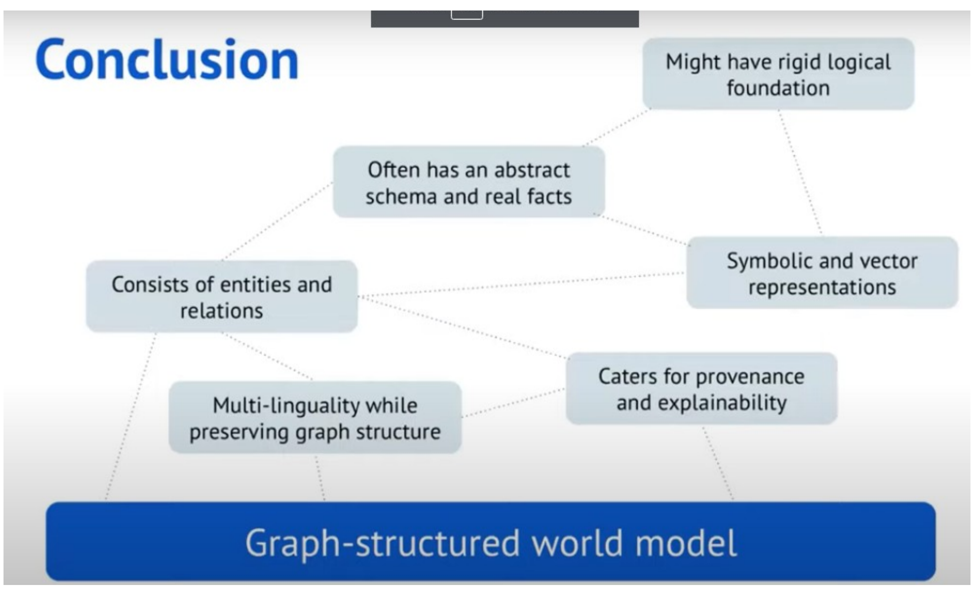
Another task in KB building is relation linking.

All of these are used in NL question answering

Language models are used to predict a word in a sentence – often trained on large textbookgood benchmark results

Qs
are ontologies and Kgs the same? How related?
Yes technically but not practically. You can get by w/o a real ontology like Chomsky, but for a banking application you need a strict ontology for restrictions on what you include in the KG. It forces you to be consistent.
A little top-level ontology helps finds errors in a KG.

Ontology defines the set of labels used in a KG.



Do we us the ACID properties for Kgs? Not unless we are using the KG for transactions.

Limitations – rdf has limited expressiveness compared to natural language.

Time series are not well expressed in KGs.

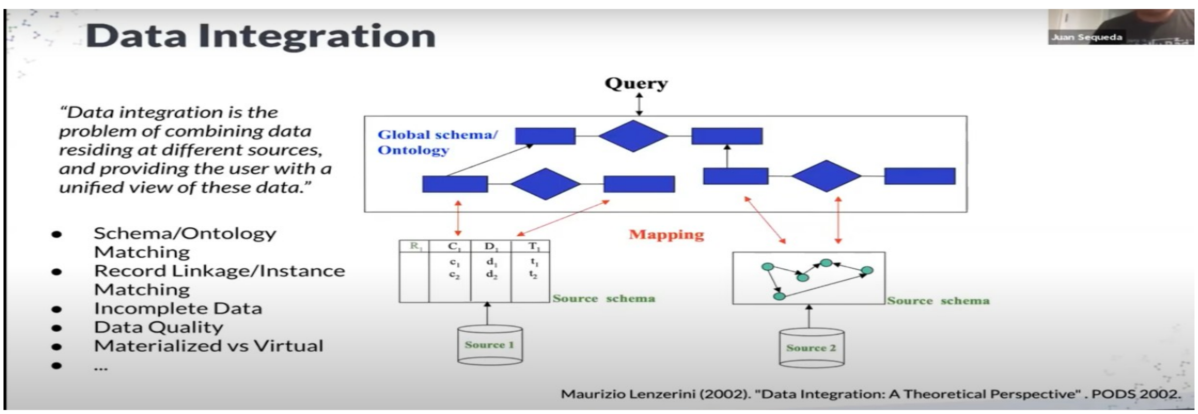
CS 520: Knowledge Graph Seminar Session 2 **How to create a knowledge graph?**

Juan Sequeda (Data.world) Data Integration has been an active area of computer science research for over two decades. A modern manifestations is as Knowledge Graphs which integrates not just data but also knowledge at scale. Tasks such as schema and ontology matching are fundamental in the data integration process. Research focus has been on studying this phenomena from a technical point of view (algorithms and systems) with the ultimate goal of automating this task.

In the process of applying scientific results to real world enterprise data integration scenarios to design and build Knowledge Graphs from enterprise databases, we have experienced numerous obstacles. In this talk, I will share insights about these obstacles. I will argue that we need to think outside of a technical box and further study the phenomena of data integration with a human-centric lens: from a socio-technical point of view.

Enterprise scenario are RDBs from enterprises to support business questions.

Integration issue.



Map data to the ontology schema. Known problems such as data quality.

Too many tables with impossible labels. Hard to join the tables since data was modeled for application and not for integration.

Large scale mapping is hard in practice, since there are many names and what do they mean?

Kbs may enable us to bridge a data meaning gap. There is a socio-tech solution with people, methods and tools. It more than linked data it is a linking of knowledge and data. Need a Knowledge Scientist role. (formerly KE)

Tools include GRA.FO.

Take away *Technology Fallacy* Kane et al.

Chris Ré Theory and systems for weak supervision If you want to build a high-quality machine learning product, build a large, high-quality training set. At first glance, this seems as useful as the statement “if you want to be rich, get a lot of money.” However, a key idea driving our work is that new theoretical and systems concepts including weak supervision, automatic data augmentation policies, and more, can enable engineers to build training sets more quickly and cost effectively.

Along with state-of-the-art results on benchmarks, these concepts have allowed our group and collaborators to build a range of state-of-the-art applications including patient-care monitoring on electronic health records, automatic triage systems for radiologists, and enabling cardiologists to spot rare abnormalities in video MRI—along with widely used products from Apple and Google. This talk describes the theoretical and systems challenges that such applications create.

Much of this work is open source and available at http://snorkel.org or my website.

ML Applications are Models plus (training) data plus hardware

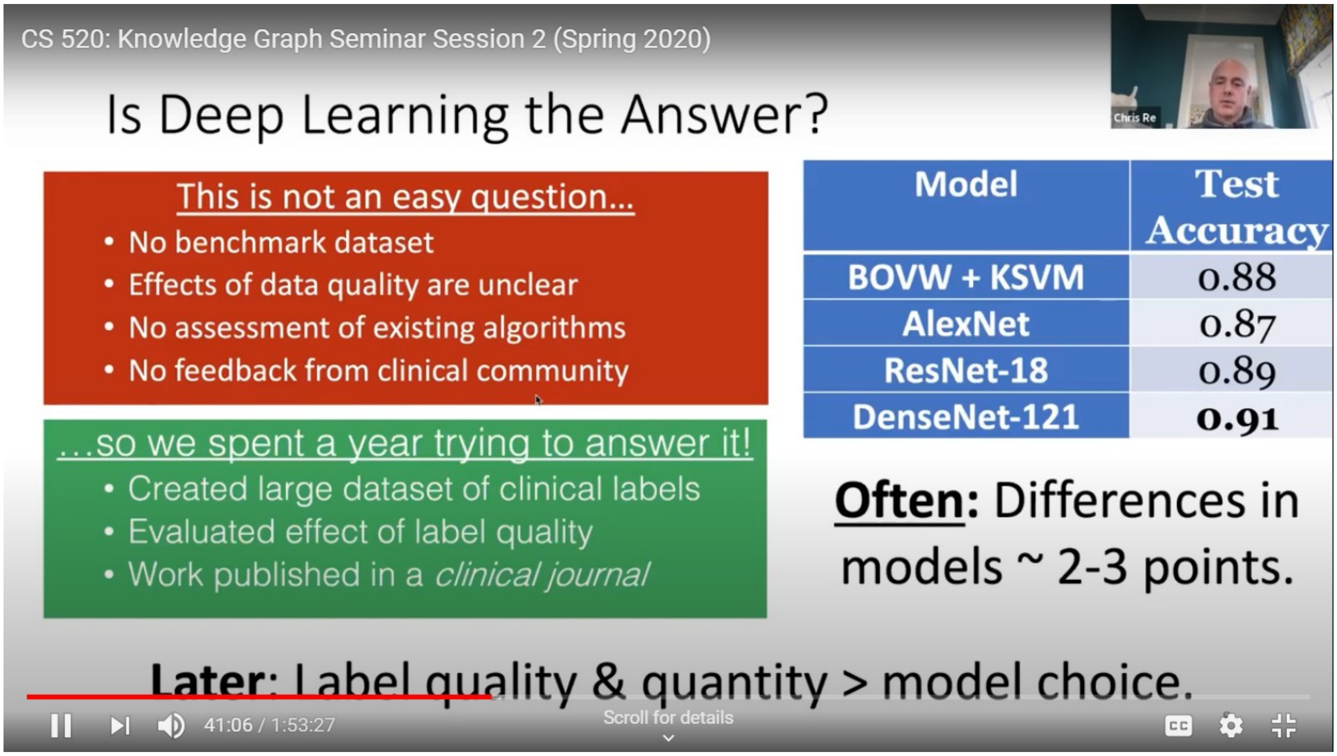
Not enough data. Usually it comes from a messy, noisy process.

Can we build a mathematical approach to clean this up?

Supervised ML is where the action is. Example of reading lung X-ray images.

Needed to create a large image DB. Takeaway prglem with image Dbs and training bases are critical (dense-net etc.)

Q on Open sources tools for inference – pellet is one.



Forefront need data augmentation of accuracy. Quality of labels more important than exact models

Created datasets is slow, expensive and static which is a problematic.

So they have looked at faster dynamic Ml approaches that are also cheap.

But the labels will be lower quality due to simple labeling functions.

Image classifier. They blend KG data and ML processes like looking for triples in graphs.

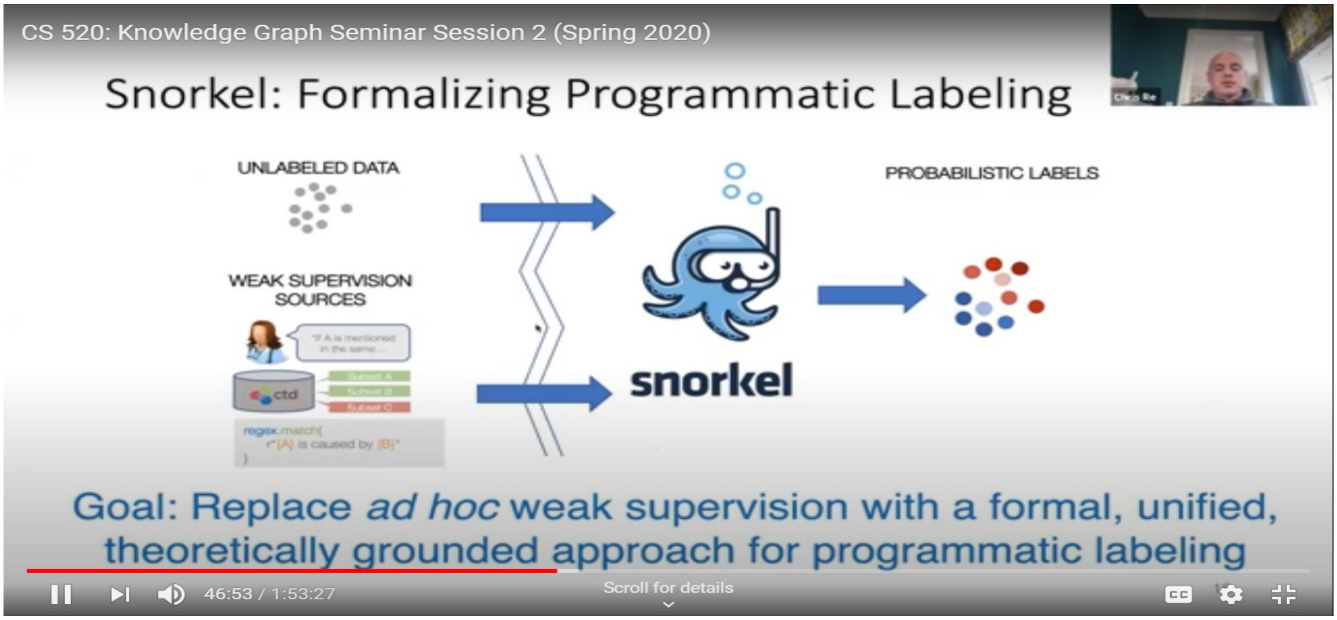
Snorkel uses labeling functions for extraction from text reports and they try to use deep learning model the noisy data.

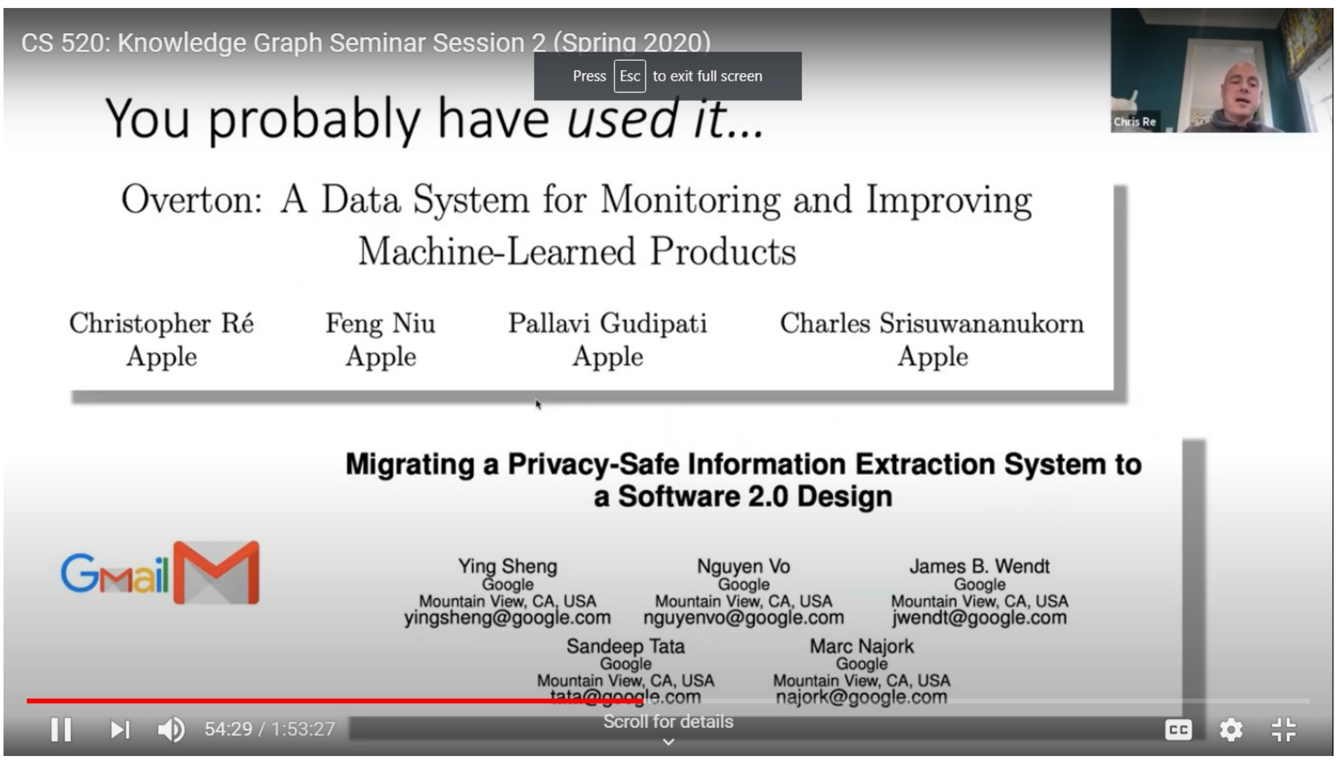
DL discovers things that your function ignored. It matches results obtained over years.

Example of a use – Google's Snorkel Drybell,

Shows how SW is changing. Data problem becomes coding. You have supervision stack to build code.

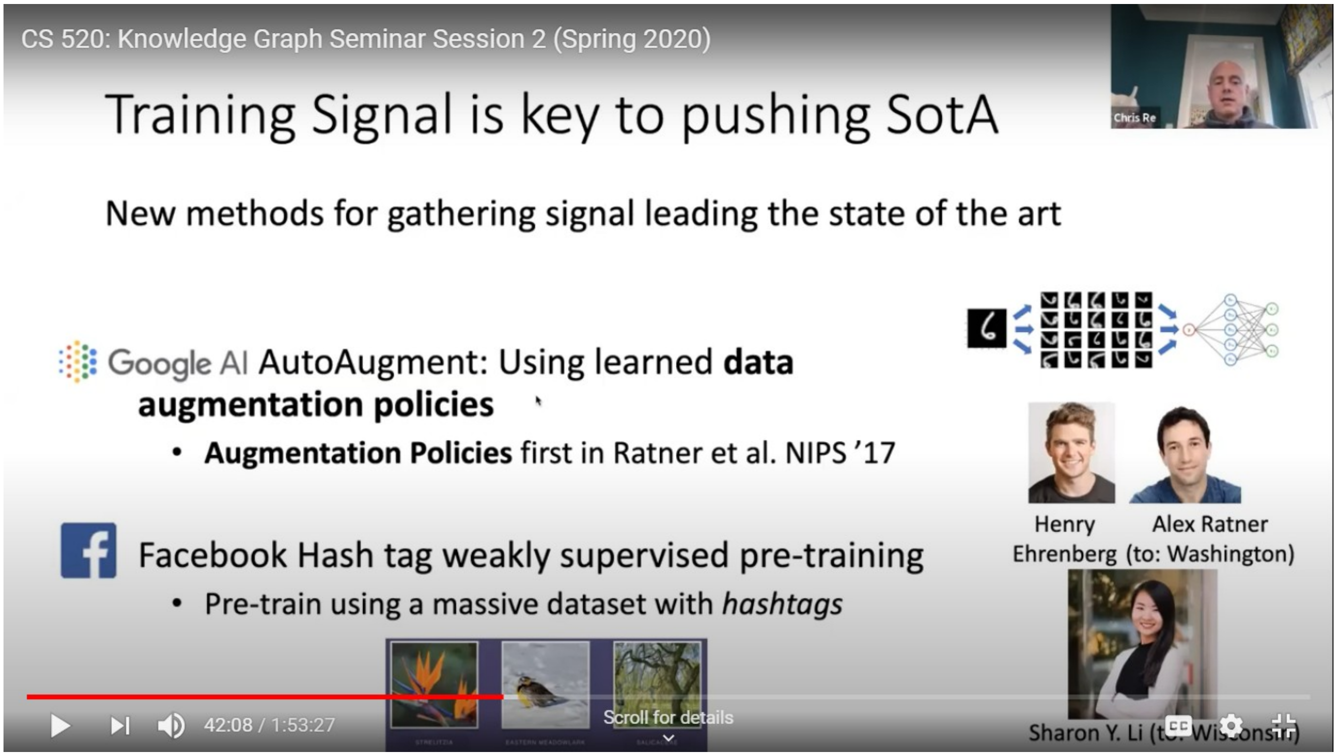
Another use: Overton

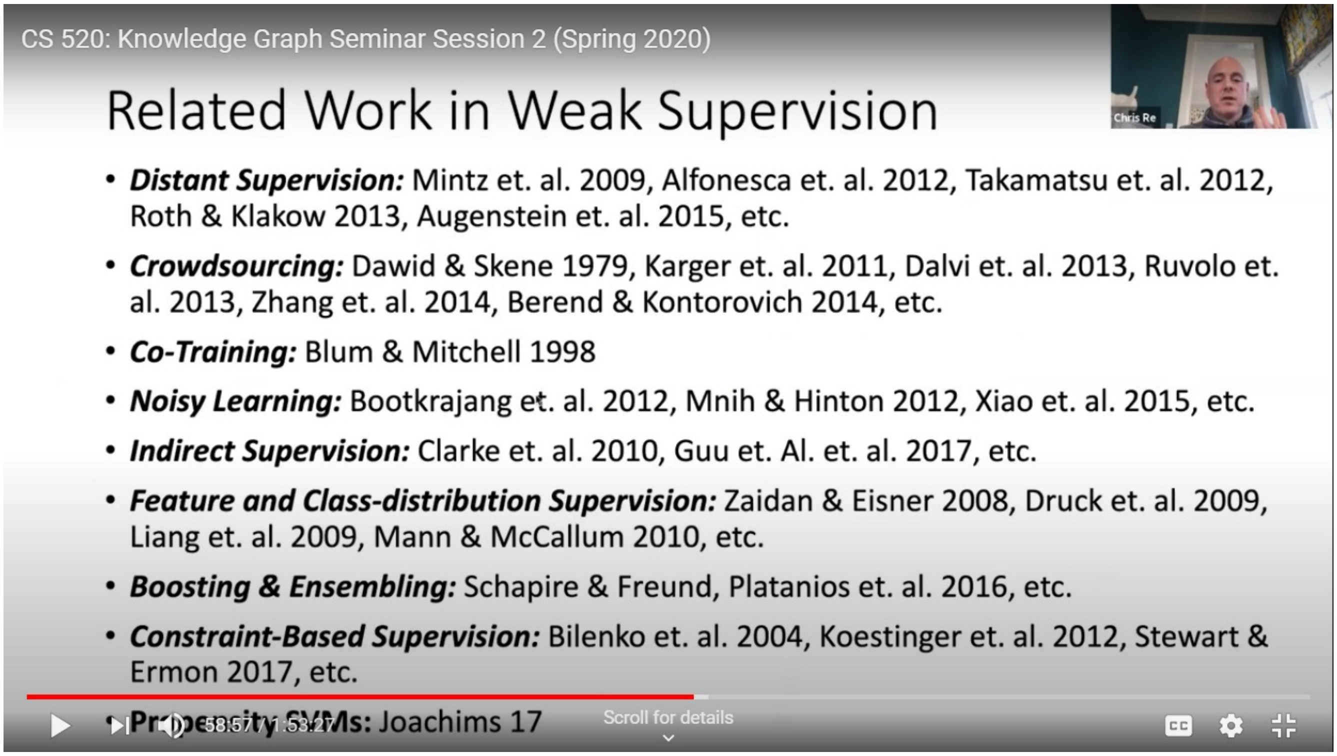


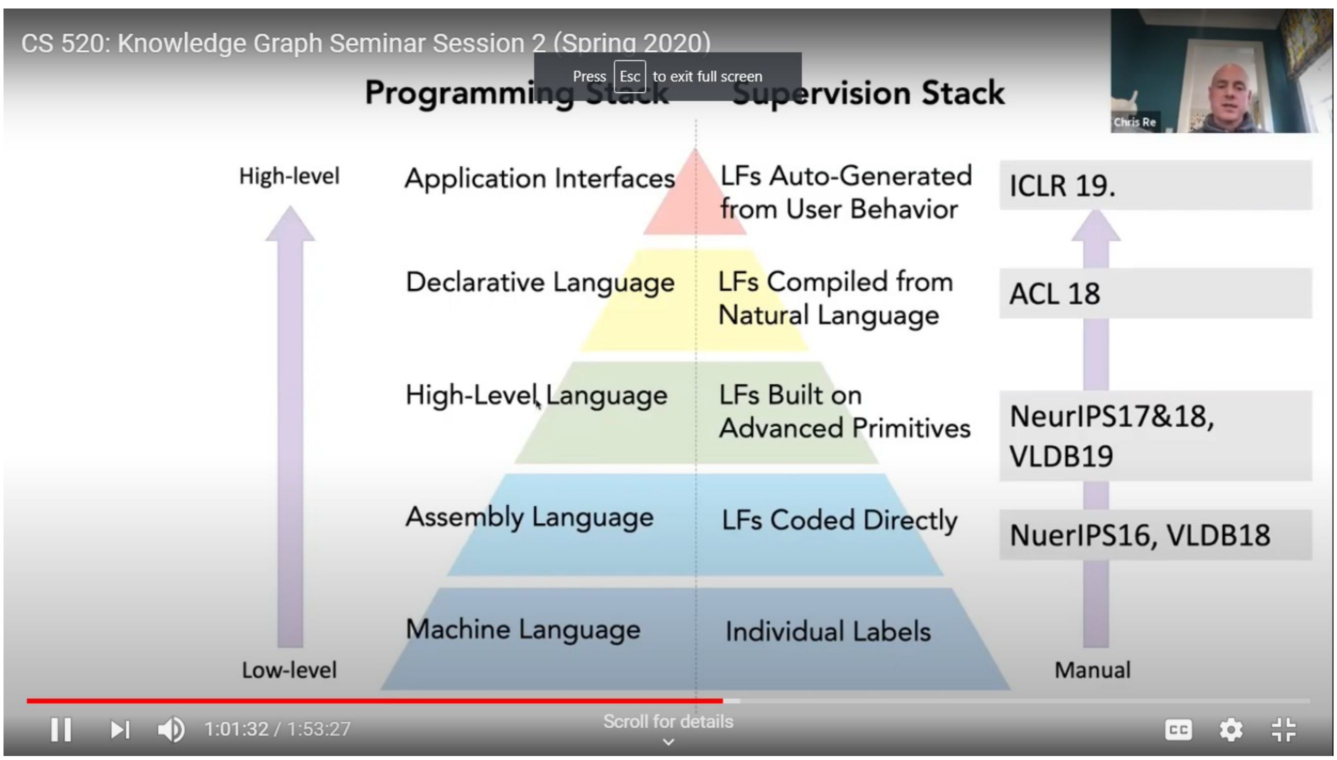


Jason Fries' recent work noted.

Sharon Li's blob post on text augmentation


CS 520: Knowledge Graph Seminar Session 2 (Spring 2020)
Chris Re
Training Signal is key to pushing SotA
New methods for gathering signal leading the state of the art
Google AI AutoAugment: Using learned data augmentation policies
• Augmentation Policies first in Ratner et al. NIPS '17
Facebook Hash tag weakly supervised pre-training
• Pre-train using a massive dataset with hashtags
Henry Ehrenberg
Alex Ratner (to: Washington)
Sharon Y. Li (to: Wisconsin)
Scroll for details
42:08 / 1:53:27


CS 520: Knowledge Graph Seminar Session 2 (Spring 2020)
Chris Re
Related Work in Weak Supervision
• Distant Supervision: Mintz et. al. 2009, Alfonesca et. al. 2012, Takamatsu et. al. 2012, Roth & Klakow 2013, Augenstein et. al. 2015, etc.
• Crowdsourcing: Dawid & Skene 1979, Karger et. al. 2011, Dalvi et. al. 2013, Ruvolo et. al. 2013, Zhang et. al. 2014, Berend & Kontorovich 2014, etc.
• Co-Training: Blum & Mitchell 1998
• Noisy Learning: Bootkrajang et. al. 2012, Mnih & Hinton 2012, Xiao et. al. 2015, etc.
• Indirect Supervision: Clarke et. al. 2010, Guu et. Al. et. al. 2017, etc.
• Feature and Class-distribution Supervision: Zaidan & Eisner 2008, Druck et. al. 2009, Liang et. al. 2009, Mann & McCallum 2010, etc.
• Boosting & Ensembling: Schapire & Freund, Platanios et. al. 2016, etc.
• Constraint-Based Supervision: Bilenko et. al. 2004, Koestinger et. al. 2012, Stewart & Ermon 2017, etc.
• Propensity SVMs: Joachims 17
58:57 / 1:53:27
Scroll for details


CS 520: Knowledge Graph Seminar Session 2 (Spring 2020)
Press Esc to exit full screen
Chris Re
Programming Stack
Supervision Stack
High-level
Low-level
Application Interfaces
Declarative Language
High-Level Language
Assembly Language
Machine Language
LFs Auto-Generated from User Behavior
LFs Compiled from Natural Language
LFs Built on Advanced Primitives
LFs Coded Directly
Individual Labels
ICLR 19.
ACL 18
NeurIPS17&18, VLDB19
NuerIPS16, VLDB18
Manual
1:01:32 / 1:53:27
Scroll for details

See Snorkel.org

Xiao Ling K based construction at Apple Siri one of the goals Siri aims at is to be a *know-it-all* question answering system, capable of answering questions from hundreds of millions of users about nearly anything. The question answering system is backed by a knowledge graph that was automatically constructed from a vast number of data sources including natural language text, HTML tables, and many others.

In this talk, we will give a brief overview of automatic knowledge base construction (AKBC) at Siri and discuss two concrete problems, Wikipedia infobox extraction and entity resolution. Wikipedia infobox extraction has been somewhat considered solved in the literature. We will show why this is still a challenging problem in the AKBC context and present our work published in NAACL 2019. We will also discuss in-progress work in tackling entity resolution by embedding entities to vector spaces for matching and promising results in our preliminary experiments.

Want to answer open queries with Siri. But how is the knowledge graph to do this built?

4 sources turning into canonical triples – text, semi-structured (DBPedia infobox info), structured data and curated. These 4 triples are then fused and inference is run to add new concluded knowledge.

Problems – DBPedia has know problems like wrong anchor links.

Built Robust InfoBox Extractor (RIBE).

RIBE components – 1 Relation extraction (M R N) is an important strategy. Ex. Membership relation for Beatles.

2. Is entity linking (uses classifier).

Entity Resolution, a key problem. Need resolution of entities in many data sets around. By hand can take 6 months.

So work on identifying equivalence of 2 entities, maybe in 2 sources combined in the KG. Names are often not reliable.

Approach embed all entities in real-value vectors and then match entities in that space. Use a NN to function approximate the hand coded heuristics used to identify entities.

Entity Linking is the other key activity.

Questions

Q on open Qs...There are many. Ontology alignment is one. Needed for messy data. Is there an automated way to do this.

Still and open issue on the best way to do data augmentation.

Open Q on how to automate parts of the current methodologies to build Kgs.

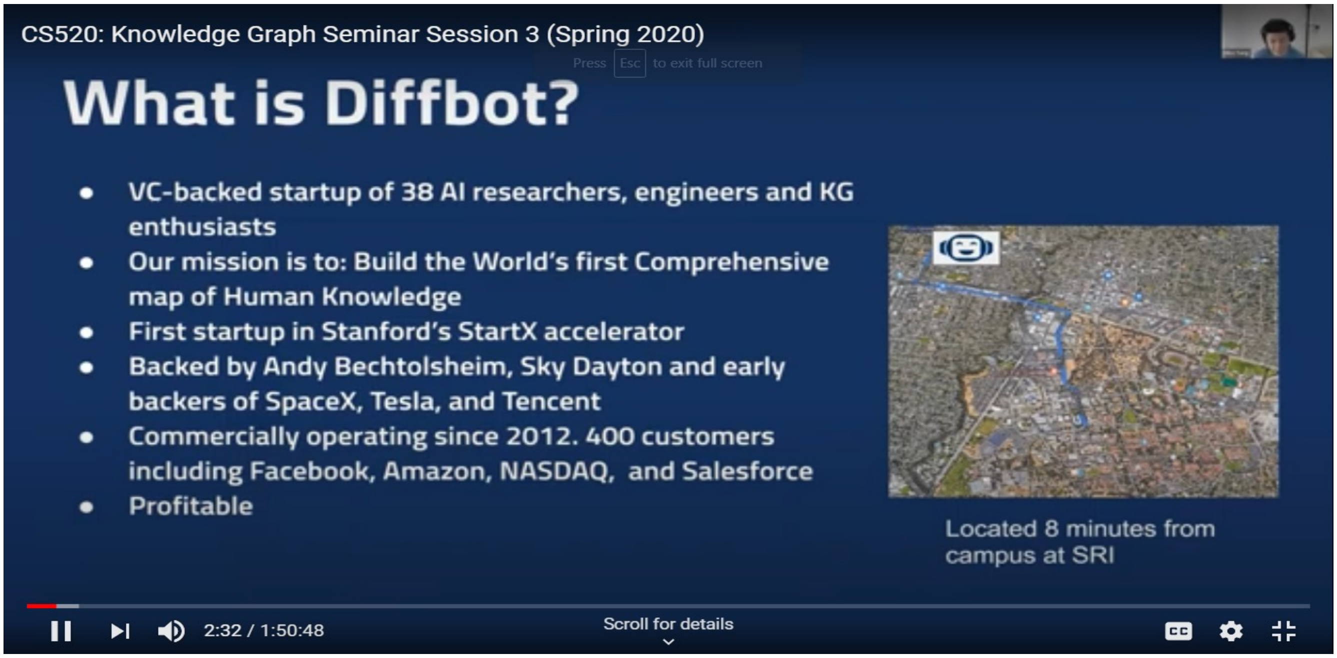
Knowledge Graph Seminar Session 3

What are some advanced knowledge graphs?

Mike Tung Diffbot KB trying to be a comprehensive map of human knowledge-based. Computer scientists have long dreamed of building a "semantic web" - a machine-readable network of knowledge that can be used by intelligent systems to perform useful services. However, current approaches to building knowledge graphs that involve a human element are either narrow and vertical applications, or broad but shallow, limiting the knowledge to the popular entities. We argue that the only way to build applications powered by real-world knowledge is with fully-automated approaches that are able to acquire knowledge on demand.

Mike Tung will give an overview of the architecture behind the Diffbot Knowledge Graph, an AI-generated, production knowledge graph built from crawling the entire web. He will describe the component technologies including web-scale crawling, rendering and classification, automated visual extraction, natural language extraction of facts from text and computer vision extraction of facts from images, record-linking, knowledge fusion, and search.

Uses ML



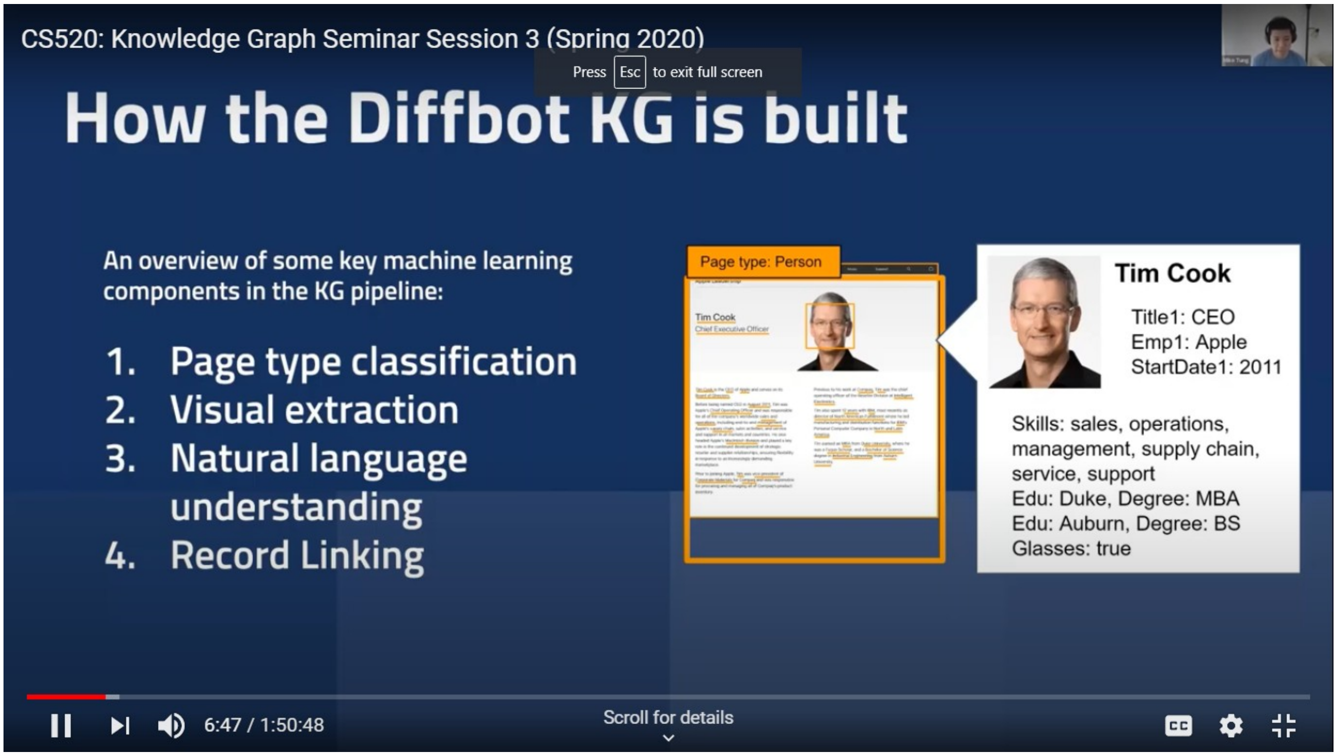
How it is built: First crawl the web, but comprehensively. They do a render of the page as if it is like a video game accessing each pixel, font types, colors, locations. It is like what a human experiences.

This data is serializes as input to their ML. Then it is classified (is it a product page or article?)

Used by their customers to mo monitor what is happening on the web.

Good precision.

Then do visual extraction using visual features. Pick out names of product, related products, headline. You can earch in English but find articles on that topic in another language.



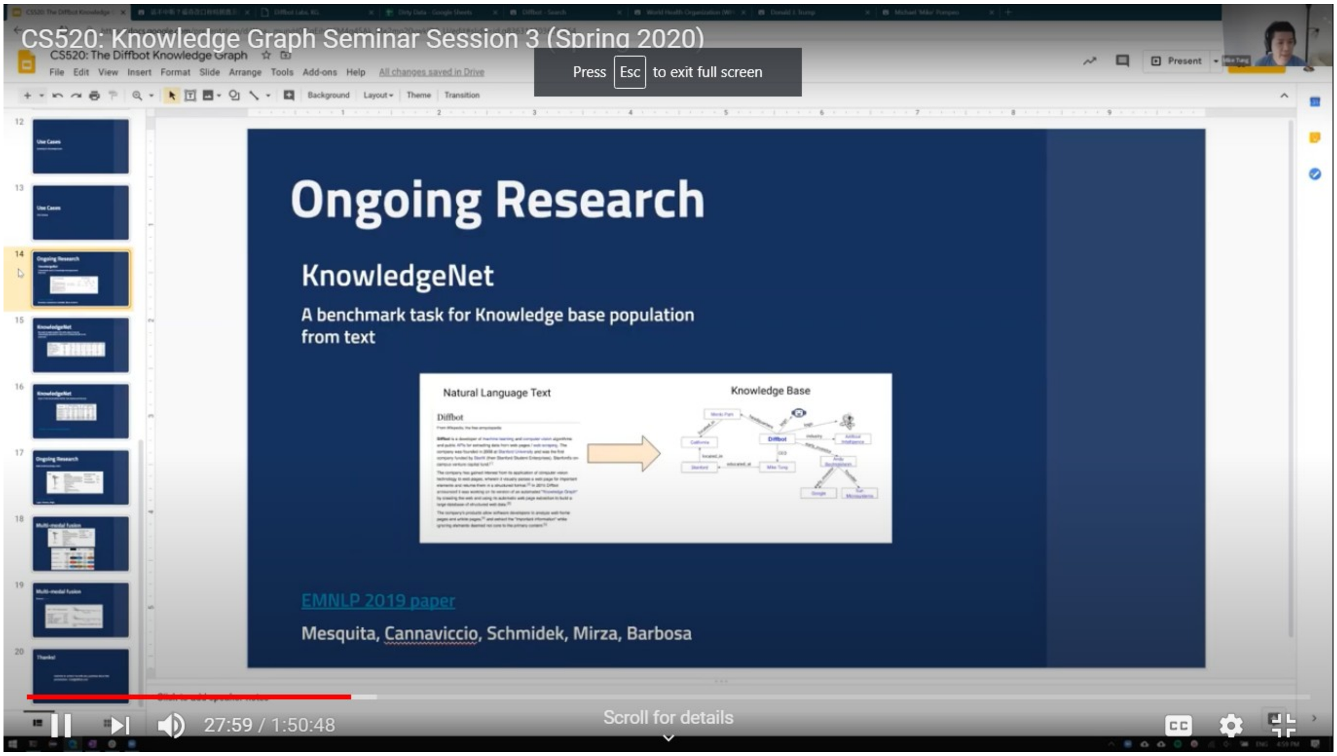
NLP uses Open (world) relation extraction uncovers new relations not explicitly expressed in texty on the web.

Image extraction works similarly and ML is then used to integrate image and text extracrted knowledge.

One example used concerned knowledge about the CDC.

Created tools to clean up data from tables. For example, typos and different addresses. They enhance the data directly by linking it to their KG. Items get resolved and incomplete data gets filled in. Info from one language source can be placed in another.

There is a paper on how to populate Kgs using NLP. Free access for research groups.

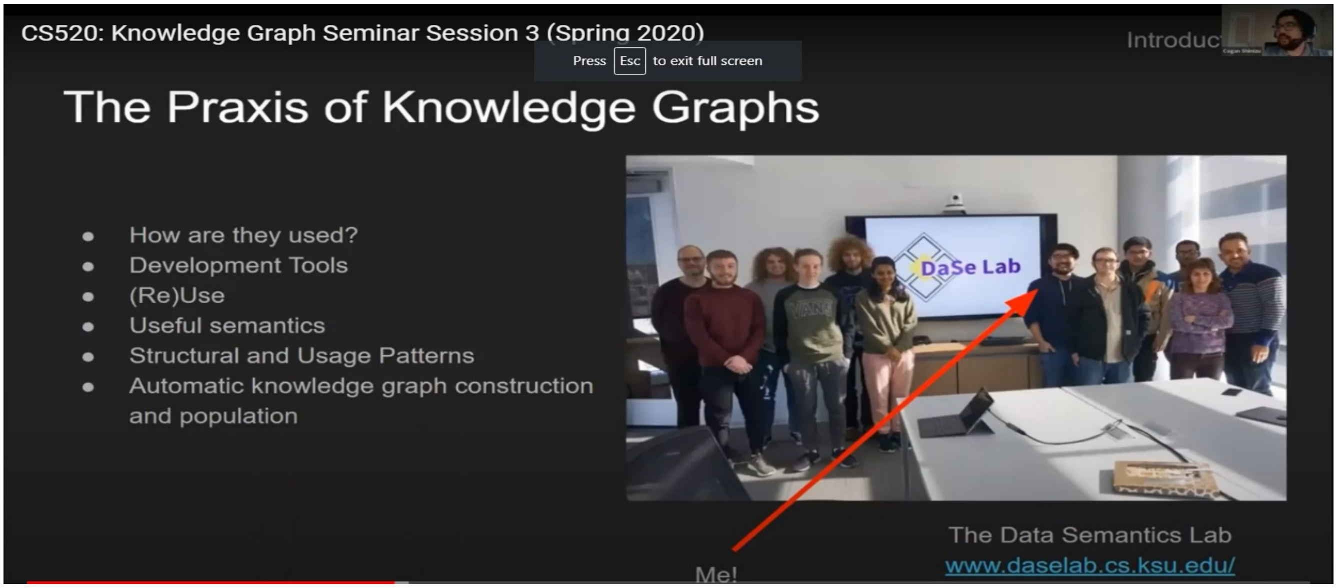


Other examples – multi-modal knowledge fusion of text and images released for researchers. Comes out with a probability that a fact is true.

Has to hand spam using knowledge trustworthiness (from a blog vs Wikipedia)

Cogan Shimizu modular ontology engineering. Knowledge graphs are poised to be significant disruptors in both private and public industry. However, as with any complex entity, developing a knowledge graph requires significant expertise and time, as well as maintenance. In order to reduce the upfront and sustained costs, it is necessary to design the schema of a knowledge graph (also known as an ontology), in a way that supports interoperability, reuse, and maintainability. This presentation will put forward the notion of pattern-based modular ontology engineering and its supporting methodology and infrastructure.

One area is praxis of Kgs.



How can we develop a highly useful KG? What infrastructure is useful for this?

Modular ontology engineering is one thing. Kgs often are hard to reuse. They may lack a clear schema. Schemas are important to understand the structure in the KG.

But what constitutes a good schema? Illustrated using a University example from a Wikipedia data and what seems to be the underlying schema.

A good schema will allow adding contextual info such as the time that enrollment is true. So we need to add blank nodes for this type of context. Makes the knowledge clearer.

We also want schemas that are low cost for maintenance and adaptability Doesn't require as much expertise or time.

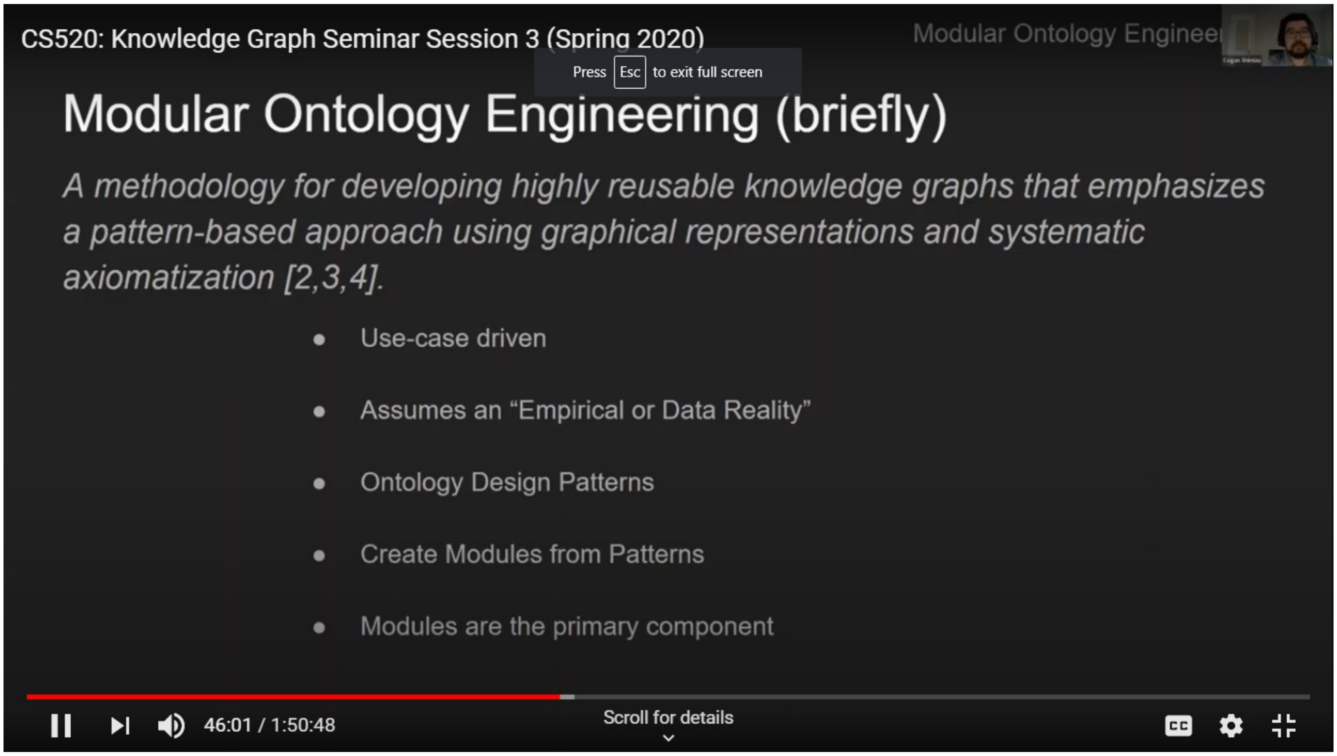
Should be Human readable.

Design decisions should be clear.

Methods to achieve these include use of logic to express the schema with precision. For example OWL DL. Promotes the FAIR data practices.

For maintenance and adaptability we use modularity. It allows Kgs to evolve. Modules act as a bridge between human conceptualization and data. They allow a plug and play.

Methods of modular ontology engineering.



They use schema diagrams to informally depict relations of classes. There are ambiguities in these which have to be overcome subsequently. What does a particular edge represent?

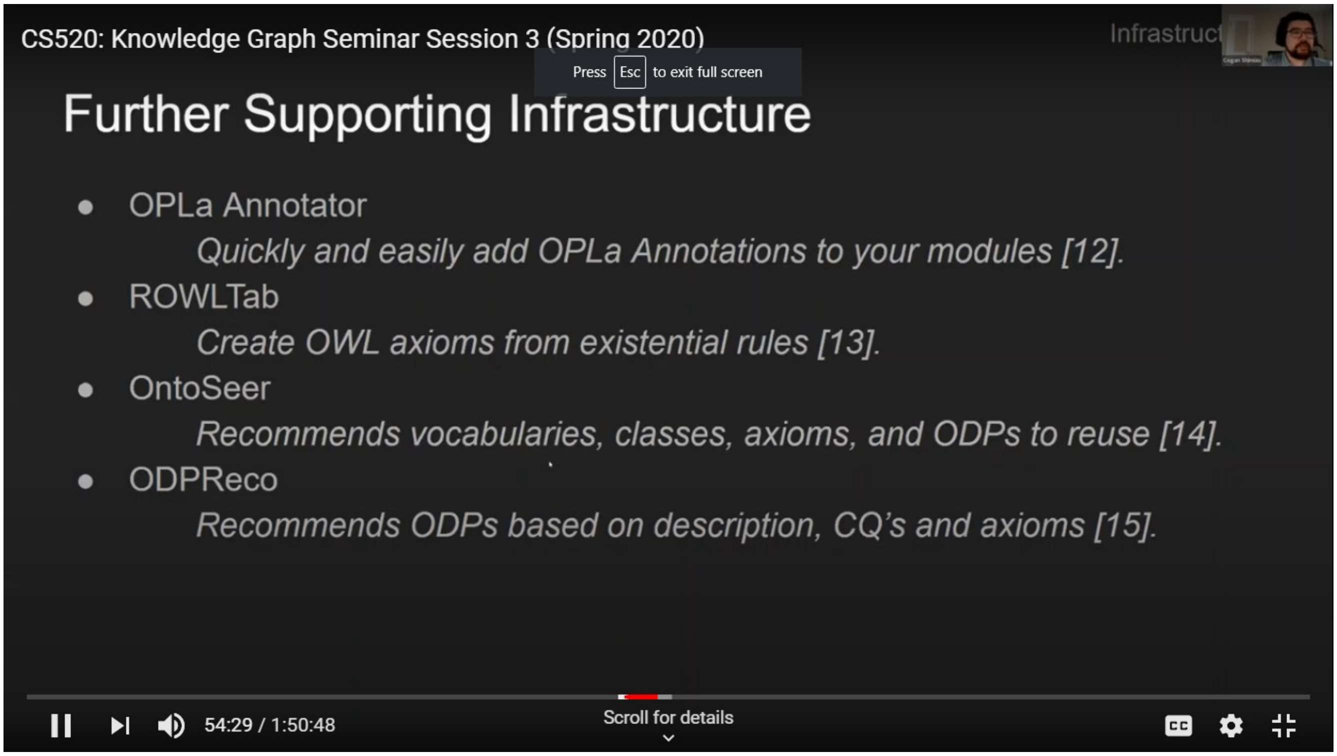
They use a set of 17 or so axioms (such as temporal extent) to detail the diagrams.

For example we may add a temporal role to the diagram.

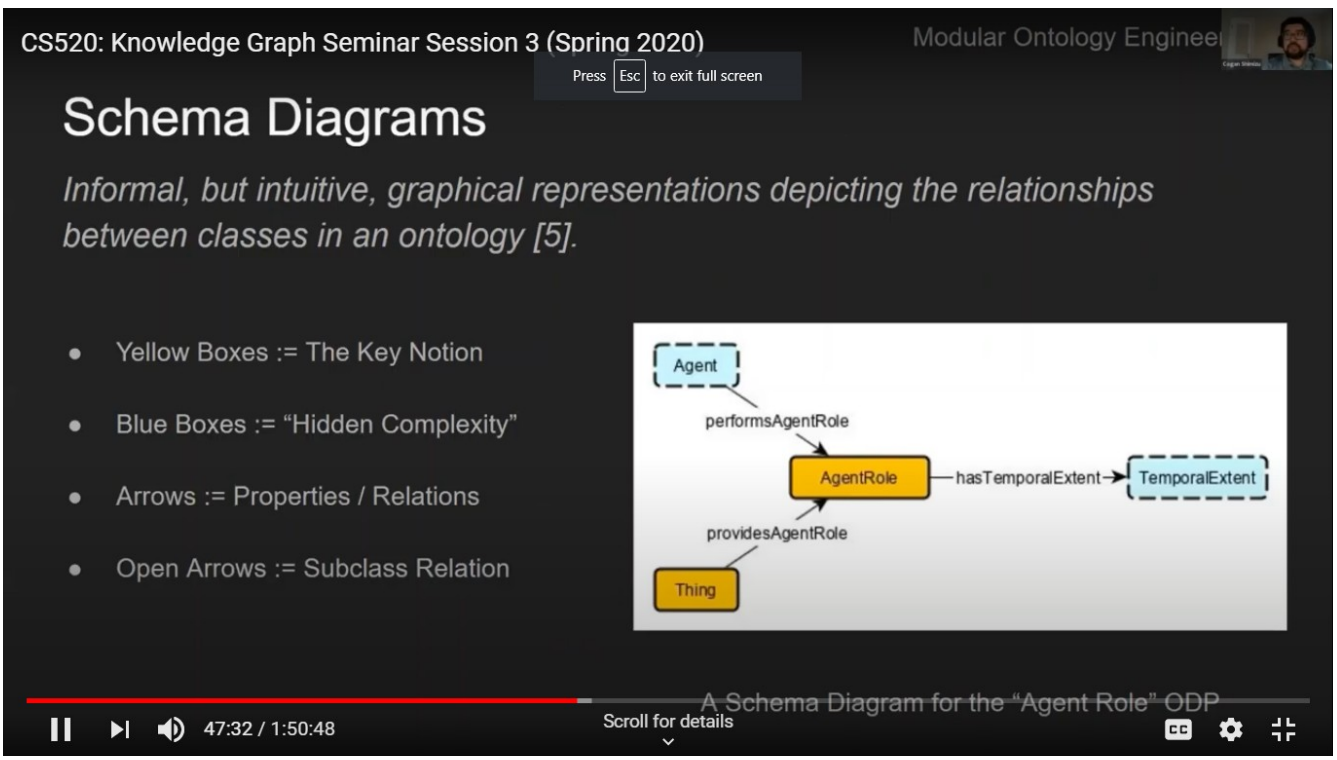
List of steps such as scoping since you can't model everything.

They have come up with some tooling that plugs into Protege : CoModiDE.

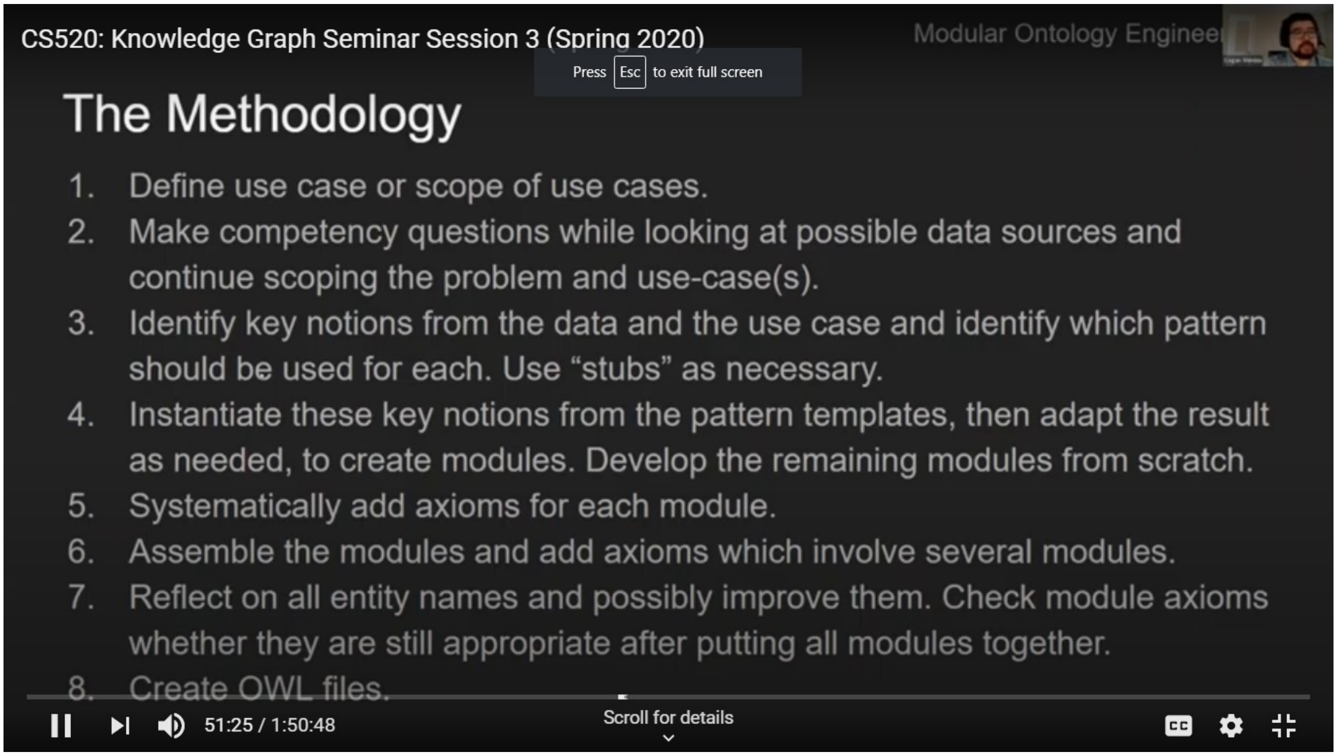
See CoModiDE.com



CS520: Knowledge Graph Seminar Session 3 (Spring 2020)
Infrastruct
Press Esc to exit full screen
Further Supporting Infrastructure
OPLa Annotator
Quickly and easily add OPLa Annotations to your modules [12].
ROWLTab
Create OWL axioms from existential rules [13].
OntoSeer
Recommends vocabularies, classes, axioms, and ODPs to reuse [14].
ODPReco
Recommends ODPs based on description, CQ's and axioms [15].
54:29 / 1:50:48
Scroll for details



CS520: Knowledge Graph Seminar Session 3 (Spring 2020)
Modular Ontology Enginee
Press Esc to exit full screen
Schema Diagrams
Informal, but intuitive, graphical representations depicting the relationships between classes in an ontology [5].
Yellow Boxes := The Key Notion
Blue Boxes := "Hidden Complexity"
Arrows := Properties / Relations
Open Arrows := Subclass Relation
Agent
performsAgentRole
AgentRole
hasTemporalExtent
TemporalExtent
providesAgentRole
Thing
A Schema Diagram for the "Agent Role" ODP
47:32 / 1:50:48
Scroll for details



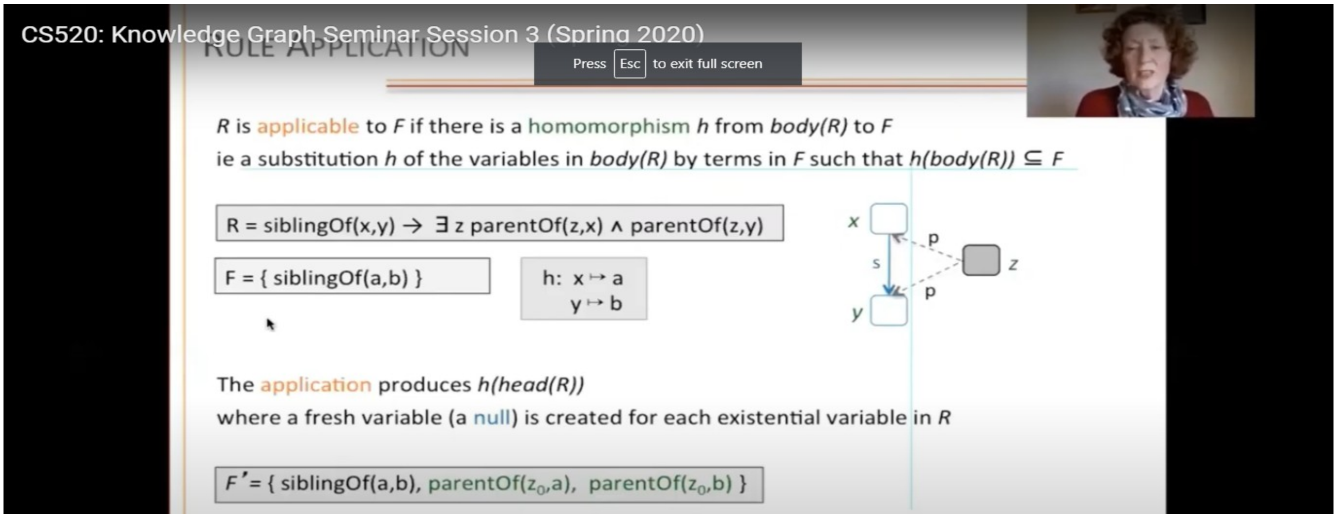
[Marie-Laure Mugnier]Reasoning on data with Existential Tools

Existential rules are a knowledge representation language dedicated to reasoning on data. This syntactically simple yet powerful language overcomes limitations of both Datalog, the language of deductive databases, and prominent ontological languages like desription logics. After introducing existential rules and their salient properties, we will present the main paradigms to query knowledge graphs in this framework, whether these graphs are materialized or virtually defined from multiple data sources.

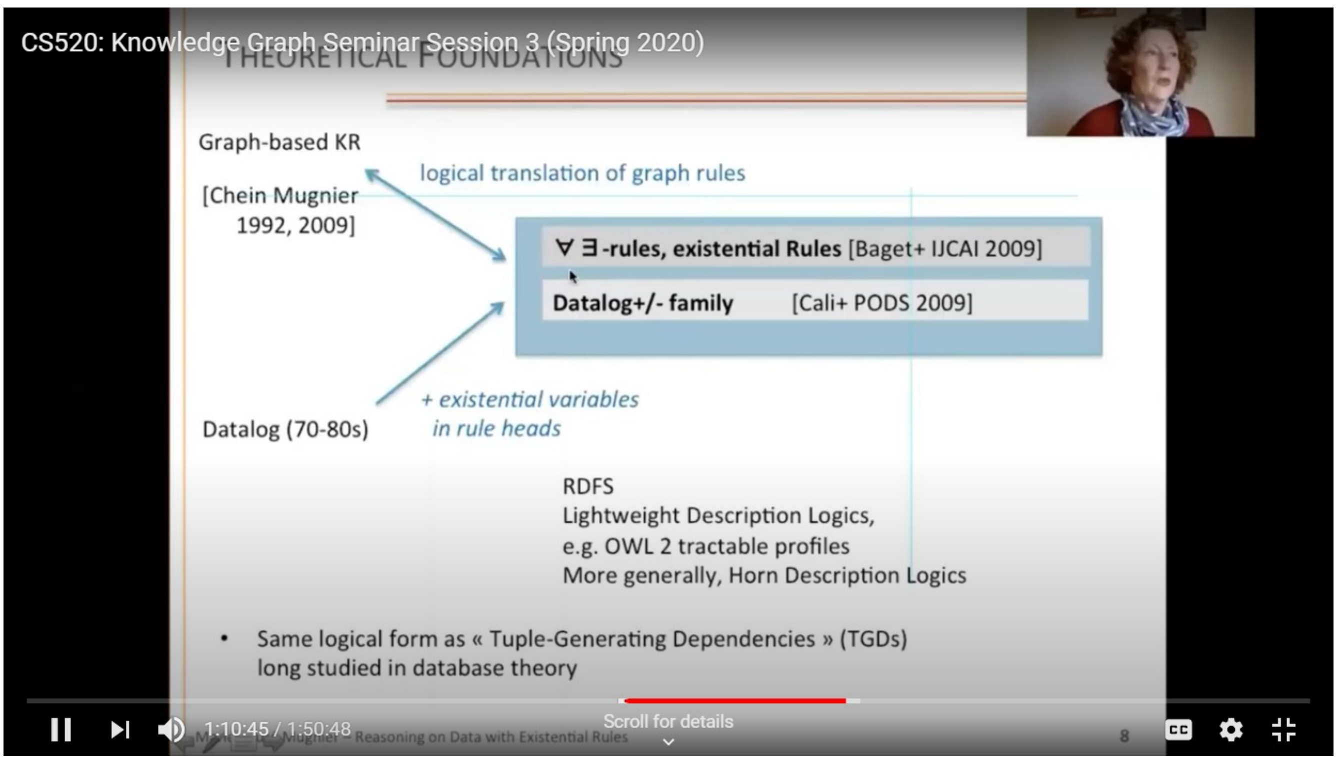
Use logic to represent Kgs Defined the idea of an existential rule to identify an unknown node in a set of facts.

Used a conflict of interest situation to discuss what types of knowledge is needed to answers conflict questions.

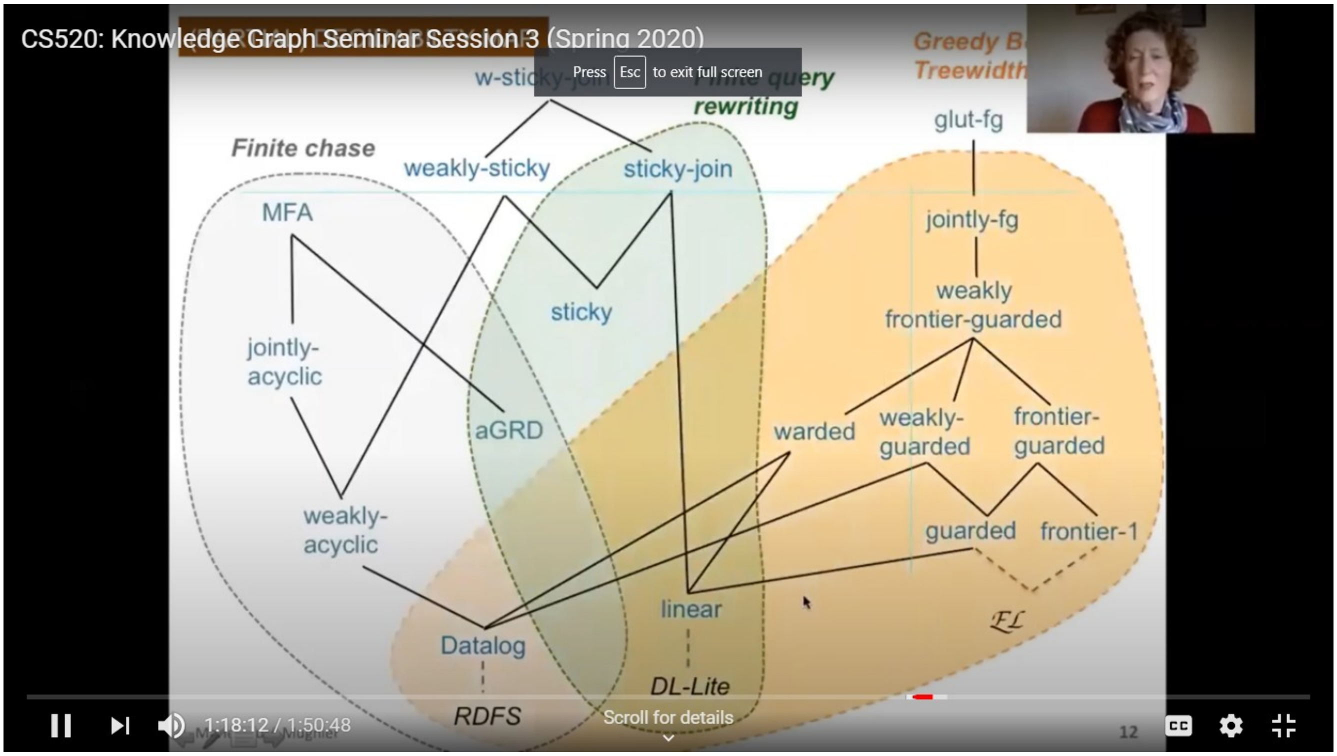
Conflict of interest rule defined with complex relations of objects – beyond unary and binary relations. We reify such n-ary relations. Allows flexible descriptions with first class objects.



Research lines from investigating graph-based rules. Find all the answers to query in a KG.

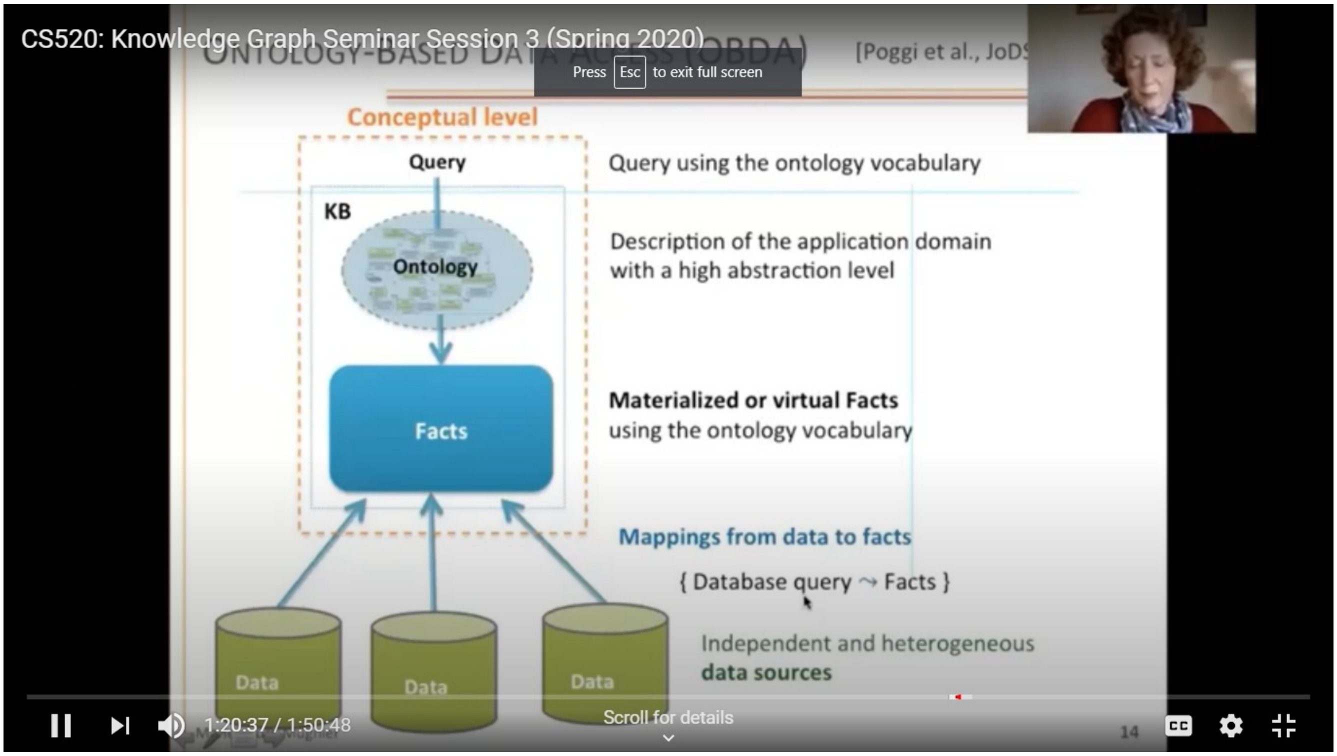


To make queries tactible they have identified some pattern of rules that are decidable.
So we have families of queries.
Forward chaining called Chase to rewrite (materialize) the rule.
Backward chaining or rewriting to a first order query is not also applicable.

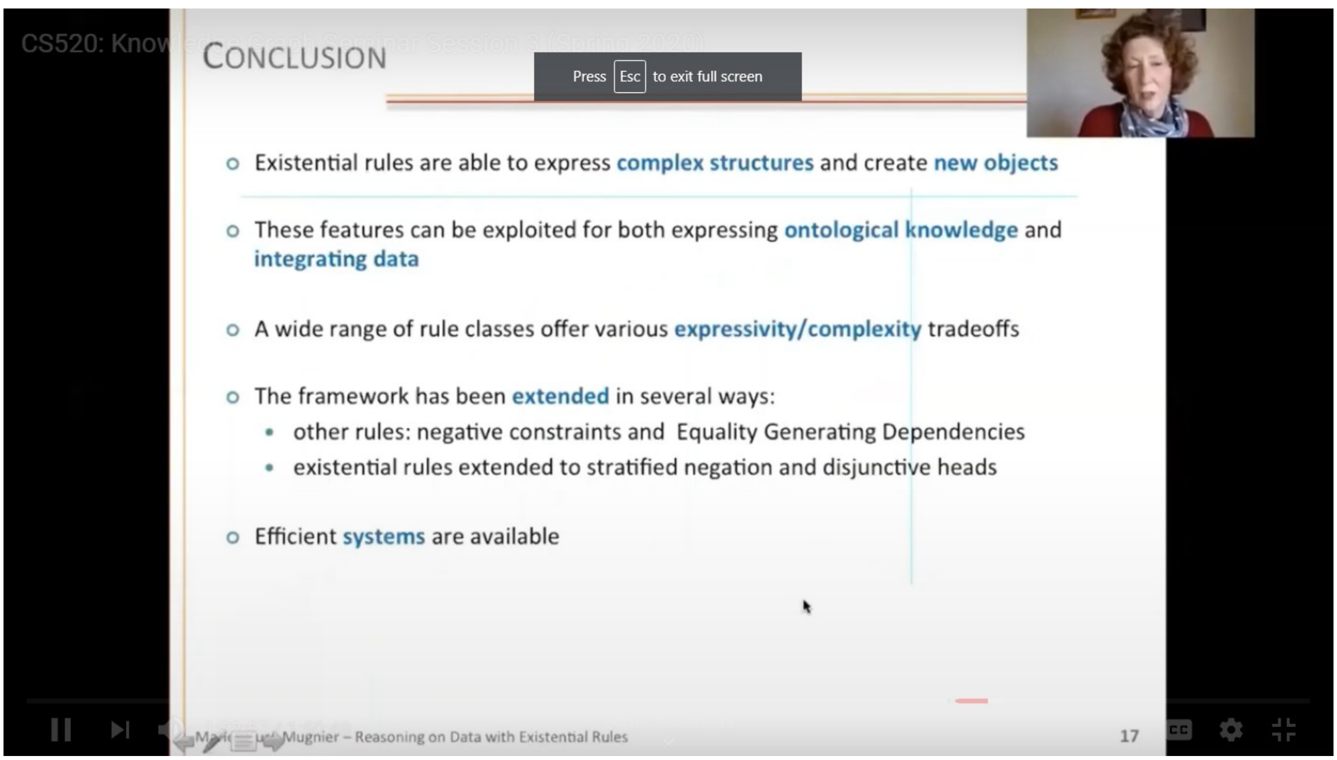


There is a corresponding view of complexity – such as polynomial.

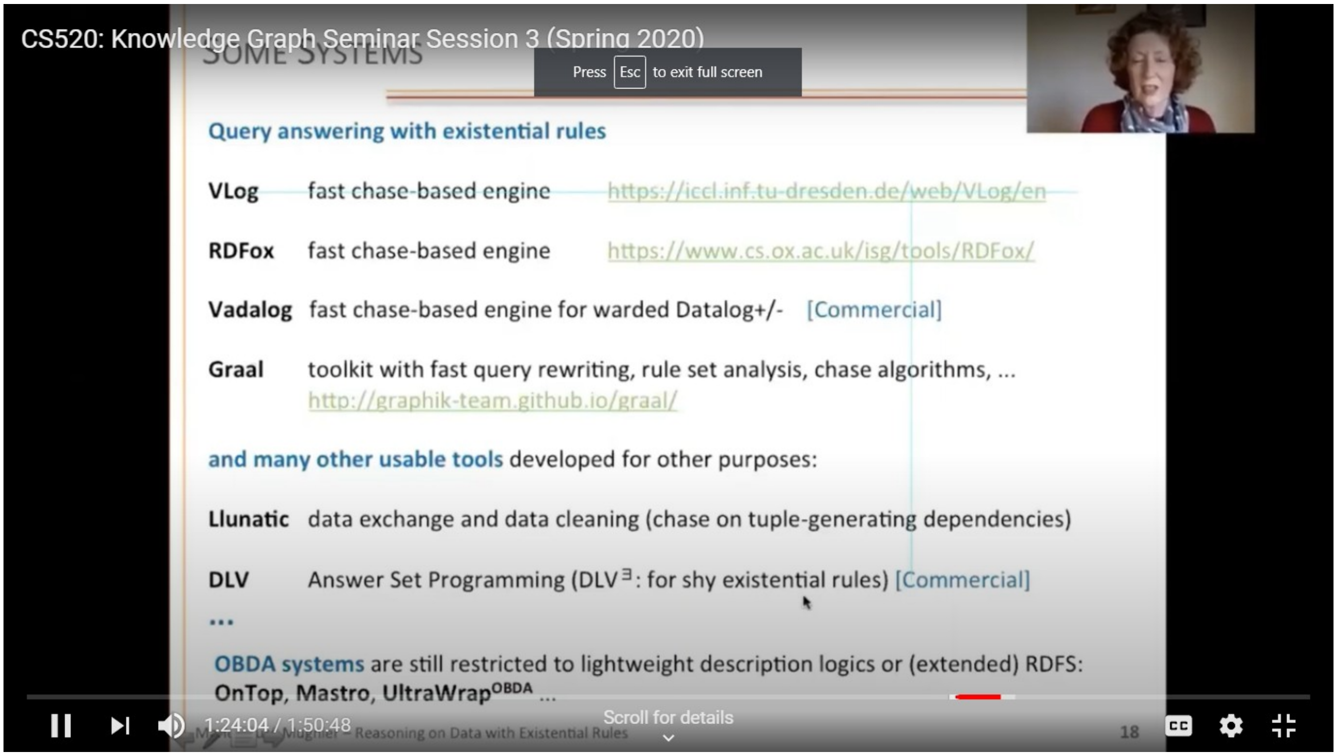
Ontology Based Data Access framewrok. Mappings allow us to select data for our application.
The framework is being extended using existential rules which allow powerful mappings that in turn find new knowledge such as missing values.



Summary



Applications of Existential rules:



Questions and Answers

Q for Mike on DIFFBOT Do they use a modular approach and ontologies?
An open answer in ML field. They can find the ontologies and may use an upper level depending on the application. The lower level needs experts.
Q Do you have a fact that people have one mother?
A This is extracted from the data.

Q on developing pattern libraries
A Being done as part of OKN funded research – emphasis on geo-information.

Q on CoVID work using Kgs.
A Mike - some work crawling to see what effect social distancing makes. Or can Kgs reveal interactions between drugs.
Some work on logistics also being done by Krzysztof Janowicz,

Q Does DIFFBOT use schema.org?
It has access and can use it as a feature. But it is often just used as markup to enhance presence on the Web.

Knowledge Graph Seminar Session 4

What are some knowledge graph inference algorithms?

An Hai Doan Fundamental Operation for Building Kgs (Magellan Project and *Cloudmatcher*at UW)

Recent years have witnessed the emergent success of graph neural networks (GNNs) for modeling structured data. However, most GNNs are designed for homogeneous networks, in which all nodes or links have the same feature space and representation distribution, making them infeasible for representing real-world evolving heterogeneous graphs, such as knowledge graphs. In this talk, I will introduce GNN architectures that can model billion-scale heterogeneous graphs with dynamics. The focus will be on how we design the graph attention and temporal encoding mechanisms to capture the heterogeneous and dynamic natures of real-world graphs. With this, I will further discuss the strategies of pre-training such GNNs for general graph mining tasks. Finally, to handle Web-scale data, I will introduce the heterogeneous mini-batch graph sampling algorithm for efficient and scalable training. Extensive experiments show the promise of GNN pre-training for billion-scale (knowledge) graphs in practice.

Example of a customer DB. Entity **matching** (EM) is a key. Algorithms have gotten very complex.

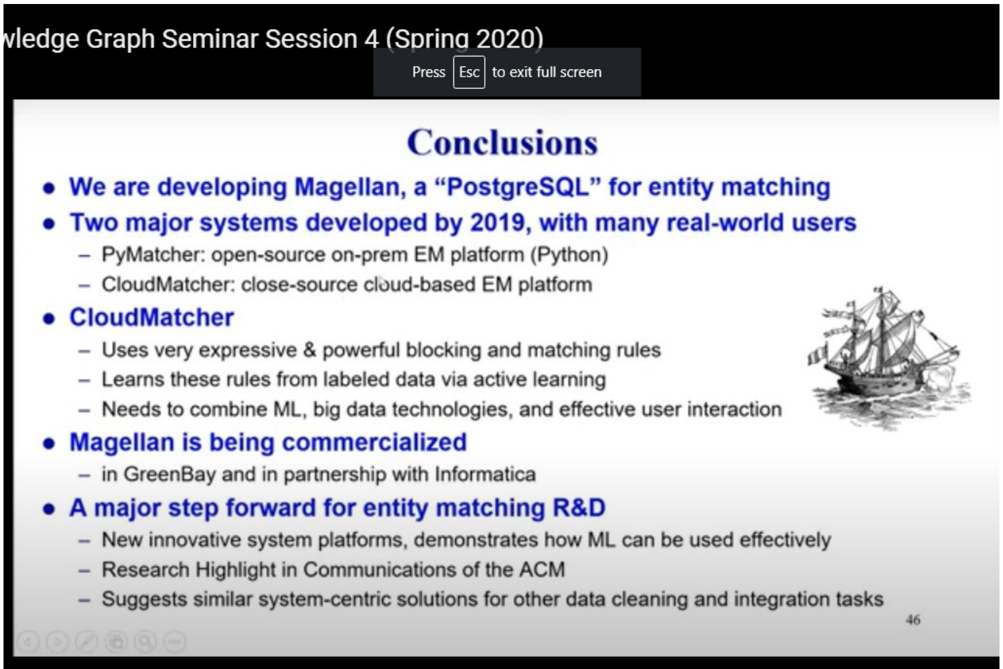
(Magellan Projec took an end-to-end approach, PyMatcher a PostgreSQL world.

Cloudmatcher allow non-expert users to do matching by answering questions (labeling for active learning)

Two main steps to reduce the complexity are

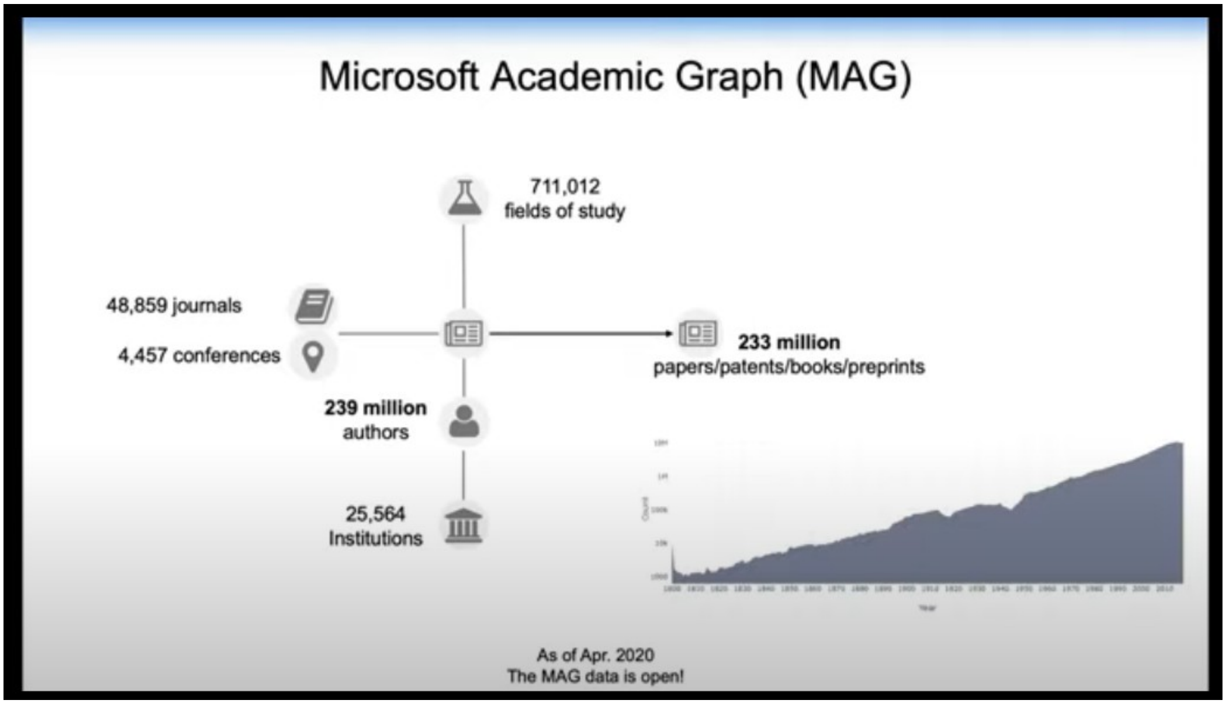
1. blocking (with a variety of similarity measures & rules on feature vectors from tokenized atrtributes) to only look at likely cases and

2. matching say using random forest approach that encodes matching rules again on feature vectors.



Yuxiao Dong Learning with Academic (Domain) Knowledge Graphs

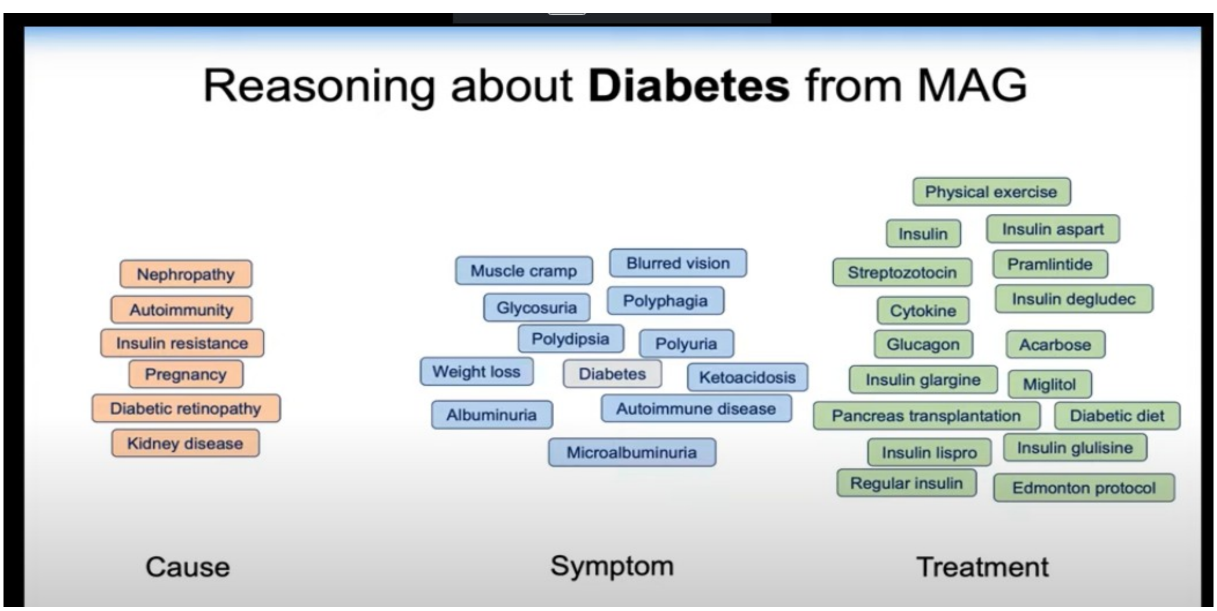
Compared to say Google Scholar they are more interested in the semantics than terms. And they are much larger.



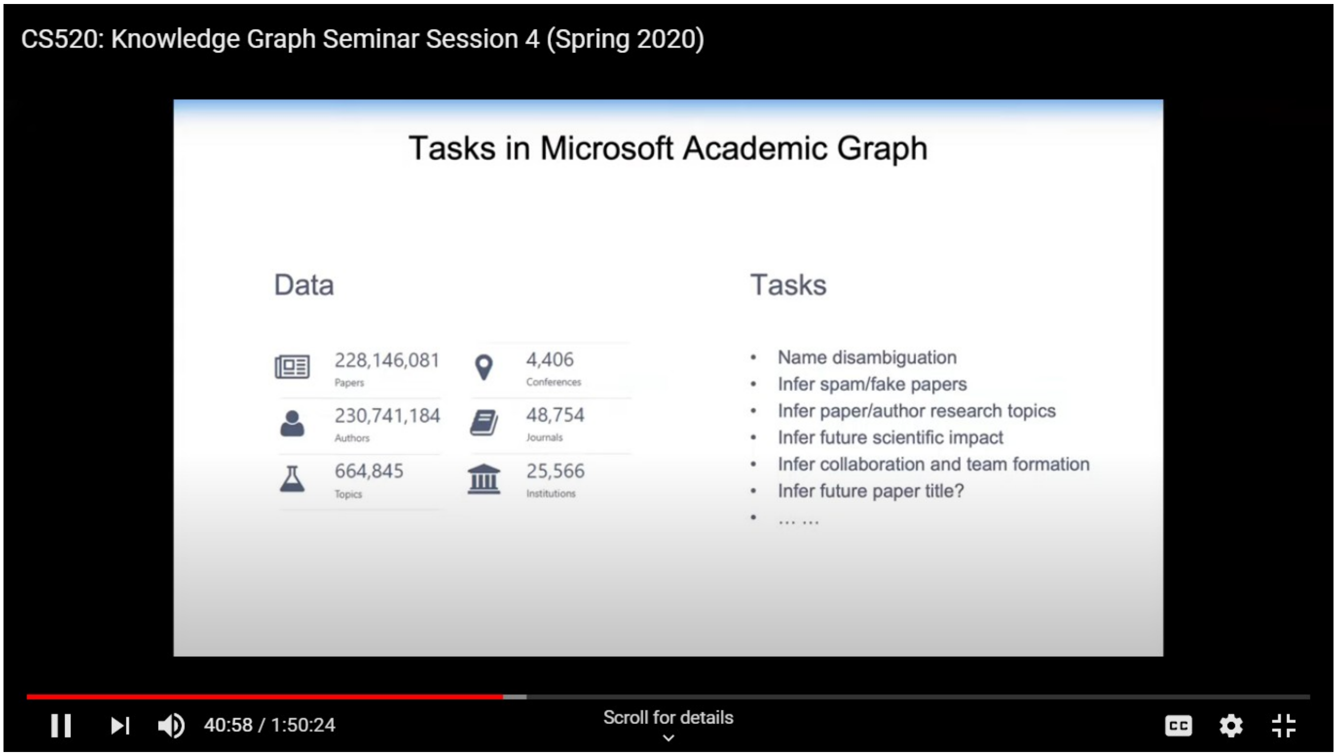
Now we have heterogeneous Kgs (BING Academic KG etc) of entities and relations. So we need a way to identify disambiguate entities (say journal Nature or Yang Yang).

What are the topics of an article is very difficult.

They are trying to reason about Diabetes and Covid-19 with MAG.



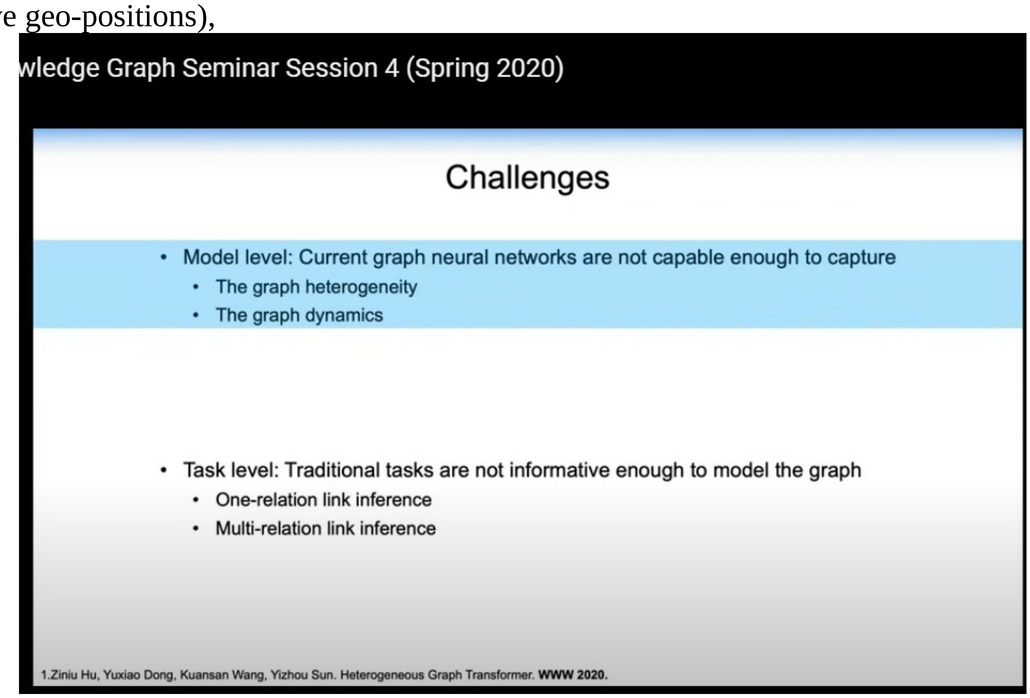
Research - Do we need a tool for each task? Trying to pre-train the language models in NLP.



Training is in 2 parts – at the model level and the task level.

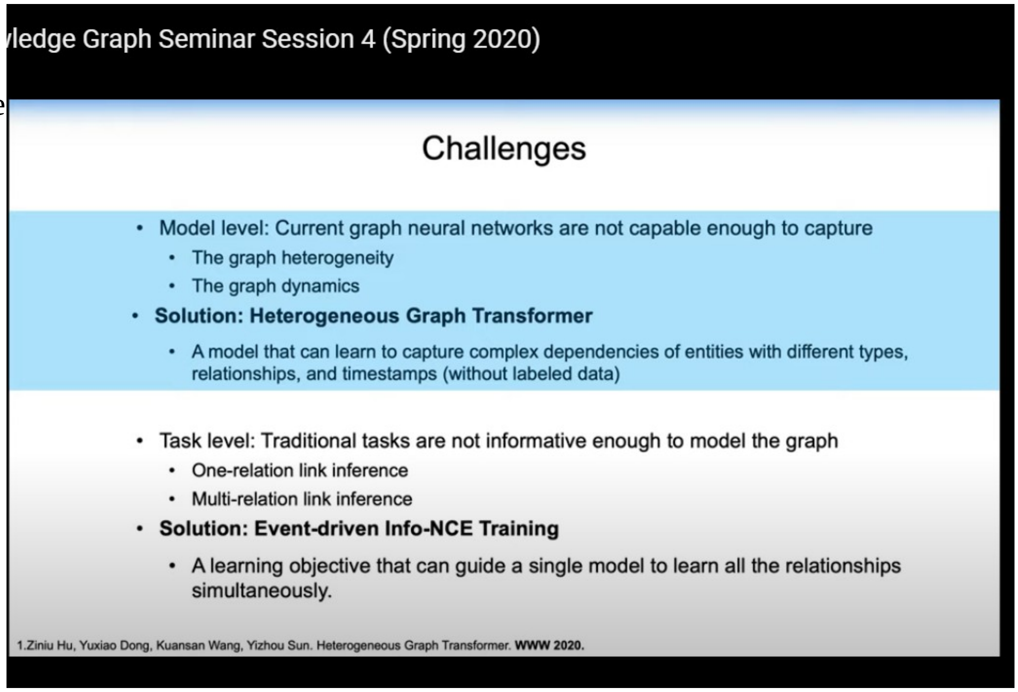
At the model level they introduce more complex relations – heterogeneous graph transformer approach

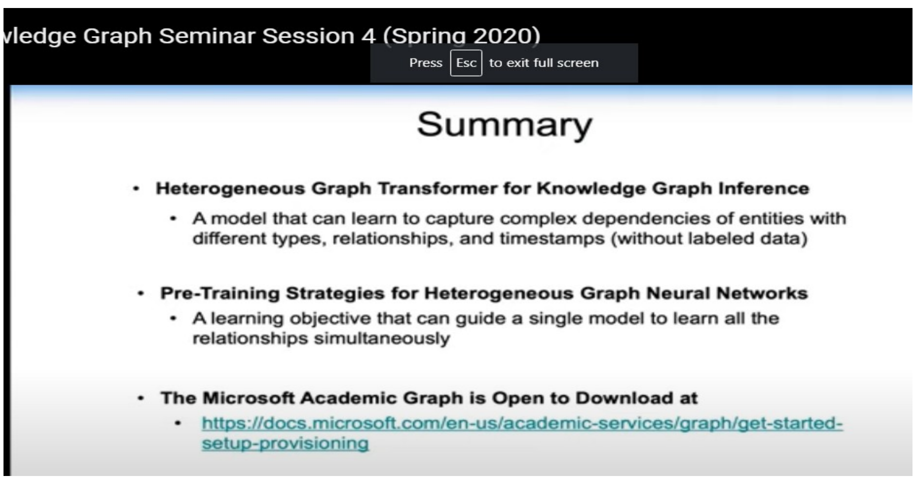
For the dynamic part they use relative temporal encoding. So different timestamps cam affect each other (like relative geo-positions),



Summary

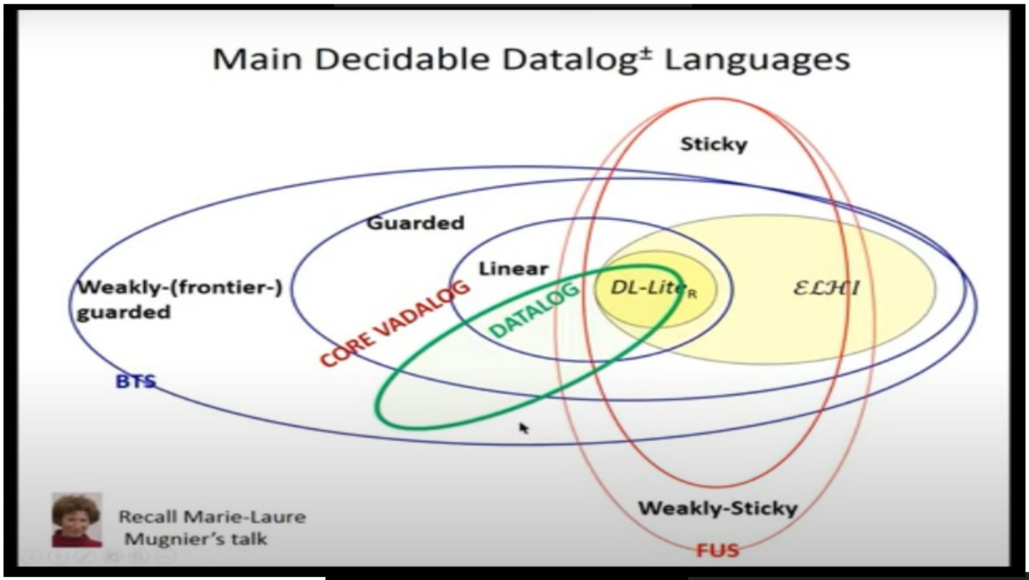
As part of their Info-NCE training they introduce fake or adversarial papers (events) so that it learns to discriminate these.

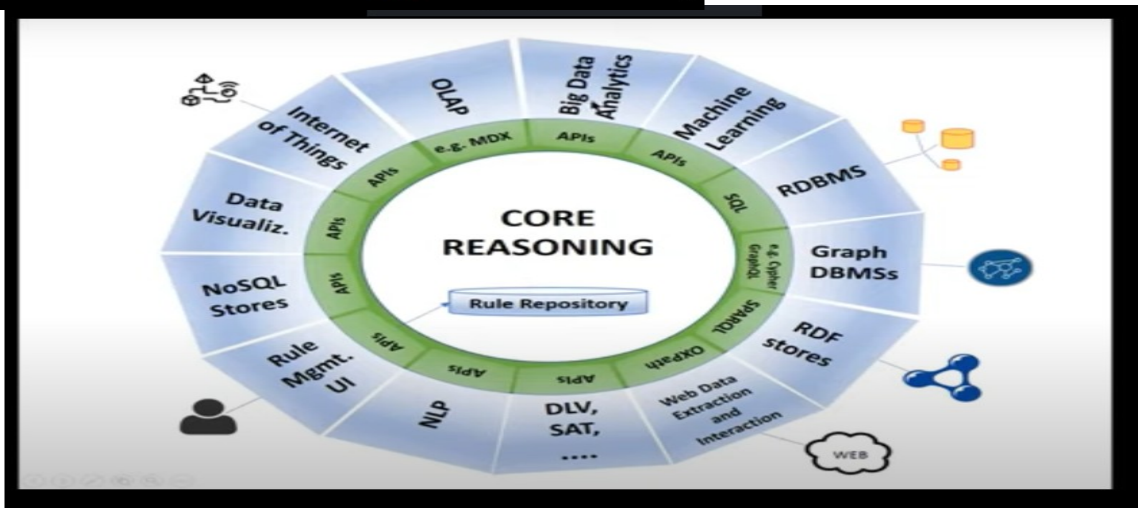




Georg Gottlob VADALOG a language and system like a full Datalog with restrictred use of existential quantifiers and strategfied negation – it extends Datalog in a decidable way

the VADALOG Flower showing interfaces.





Note a graph DB does not suffice. There may be no reasoning there. DL does not allow us to express somethings we need that are in Dbs such as “married from x to y” knowledge in these rules

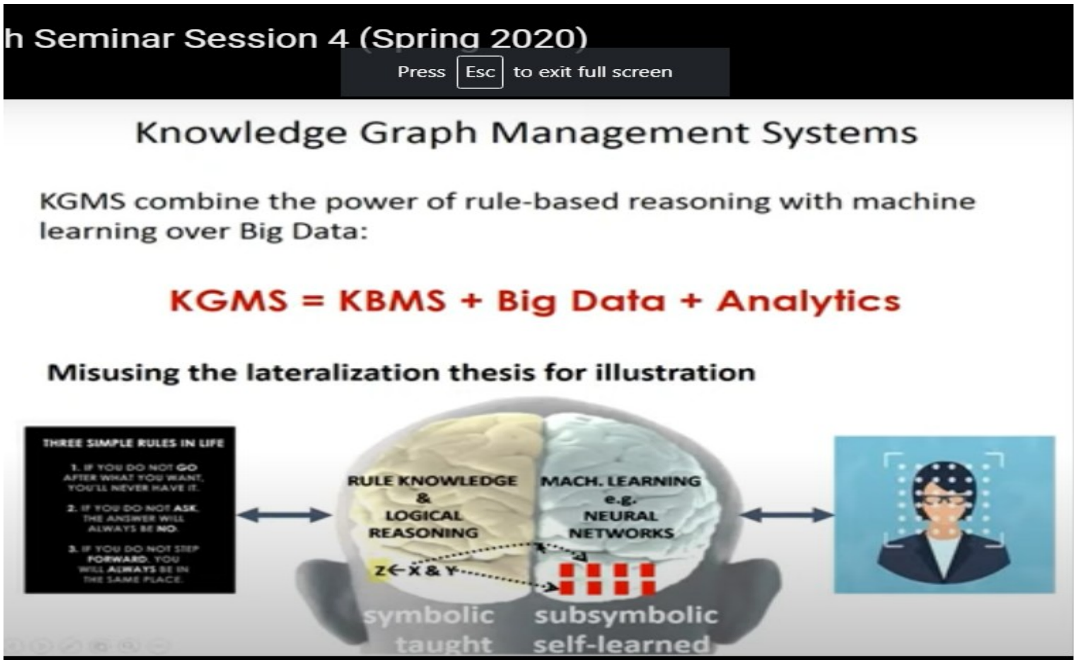
Hard to express some things like aggregate functions and recursion - X controls y if

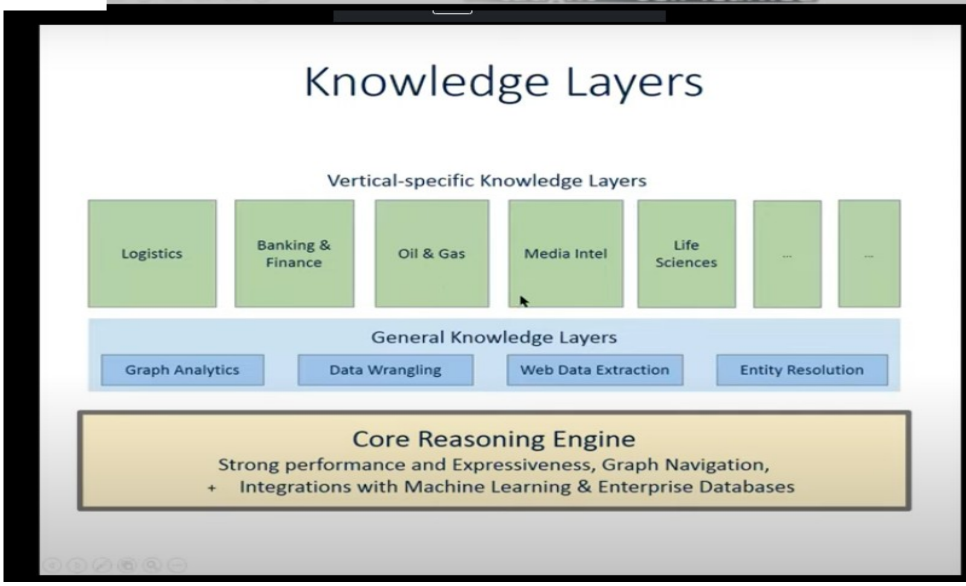
- x directly hold over 50% of y or’
- x controls a set of companies that jointly hold over 50% of y

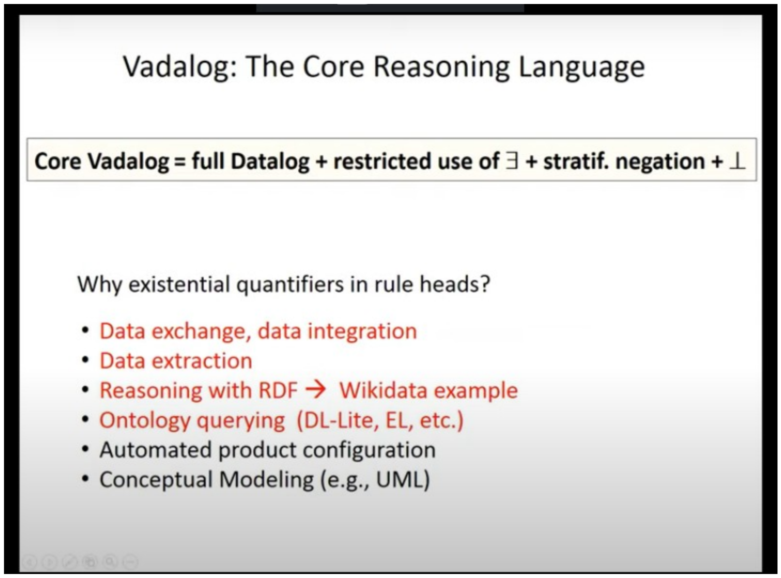
Clumsy to express in SQL

Other examples involve commonsense – creditworthiness if you live in multiple places.

Knowledge layers

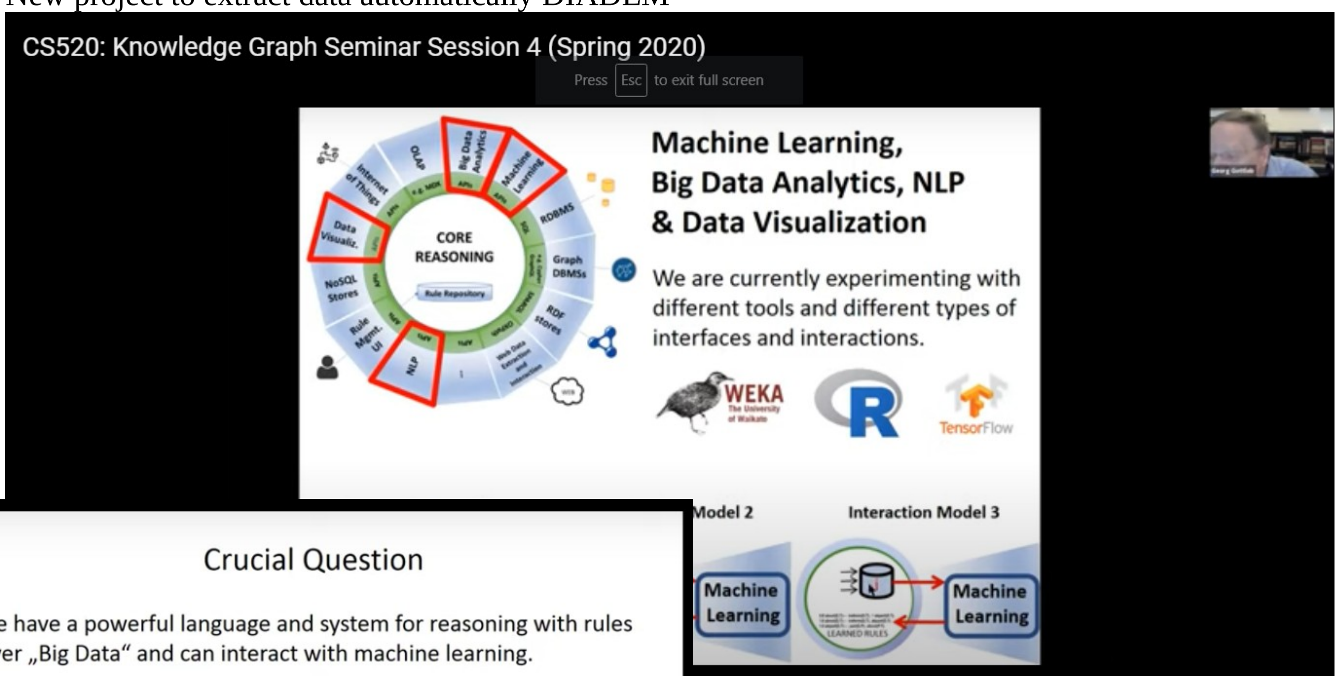


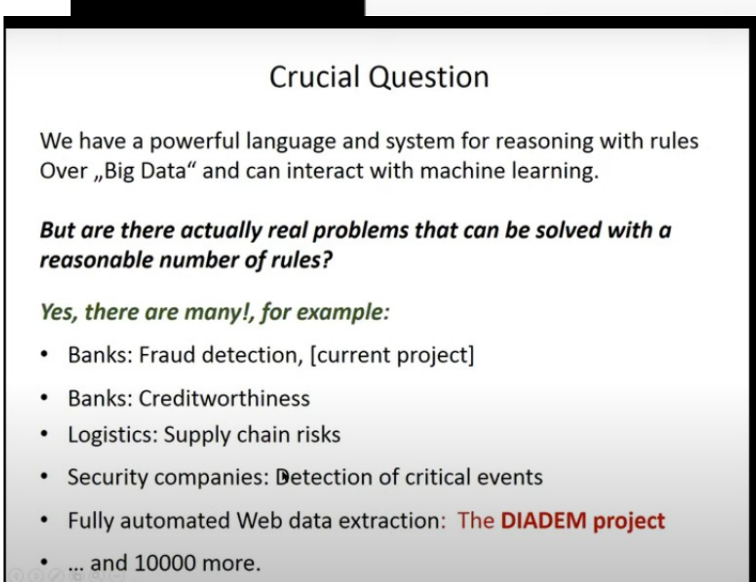




Research on good algorithms to do useful thing with ML.

New project to extract data automatically DIADEM





Questions:

For Georg have you tried using NL to express the rules? Yes, thought of this. Can express rules in SQL.

For MAG have you ever tried to couple this with rules to supplement the ML.

A. Yes, there are rules employed say in name disambiguation for ORCID IDs. Or a URL from an academic institution gives high value.

Knowledge Graph Seminar Session 5 **How to evolve a knowledge graph?**

Héctor Pérez-Urbina (google graph work)   KG Evolution and Maintenance

We are usually modeling a dynamic world so things changes and you need to adapt.

Here are some things we found:

- Vocaloids –fictional characters that have albums forced them to change their assumption on what could be an artist.

- The book domain is involved. From books you may want to talk about other types of “publications”. but need to have a general concept. But the applications on top of the KB may already use a property for a book. So generalizations can cause problems.

- CEOs are modeled with a type with start and end date.

- Music albums are covered in multiple sources. The schemas differ so integration is hard.

- Inference can be hard when based on wrong assumptions. Not only movies are played in a cinema. Operas got classified as movies since shown there

But modifying a KG can be easier than modifying a DB. Adding a column is harder than adding a triple. Factors you need to take into account:

- What is changing? Might be the scope. Adding some model,

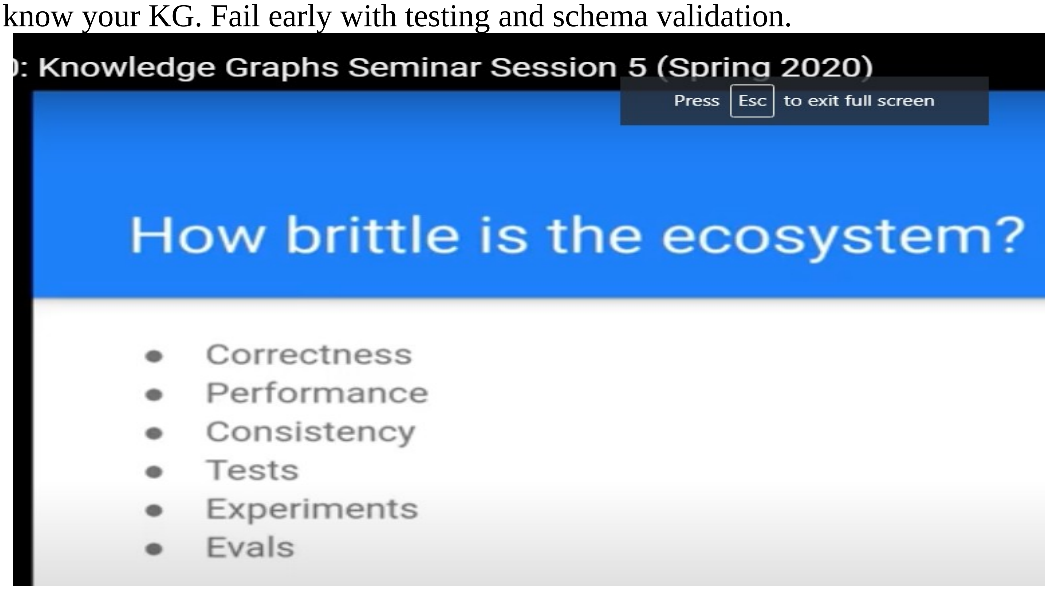
- Is there data already? It might be noisy, incomplete and need to be integrated. Is it a one time migration or continually synchronized. How do you deal with retractions?

- Who is using the data? To some users people and animal are disjoint. Not to biologists. Assumptions of a model can change. Groups may be switching to a new version.

- How complex or brittle is the ecosystem? You may have many SPARQL endpoints, many models integrated and many sources. How quickly are changes to the schema made? What is done for truth maintenance? Are constraints mutable? Not allowed if it breaks the schema.

Brittleness – if billions of triples need to change to make something correct you have to take it into account.

So know your KG. Fail early with testing and schema validation.



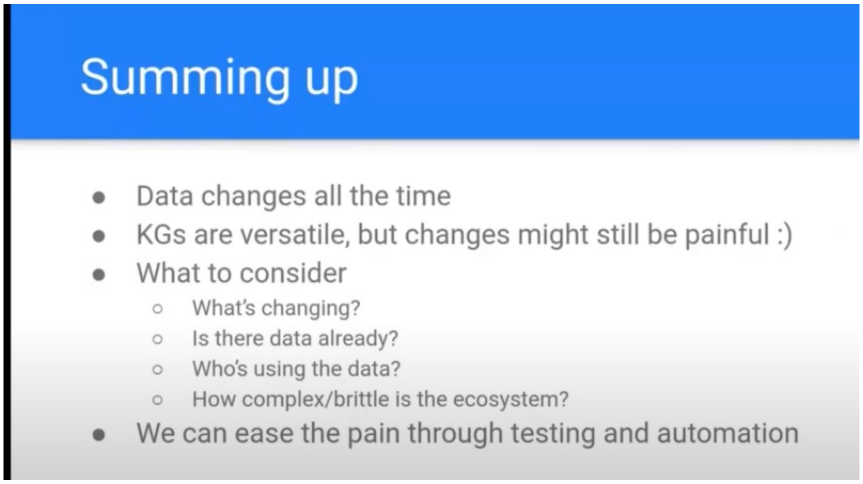
Get support for experiments and do comprehensive evals.

You need a governance and policy to manage the KG construction and maintenance.

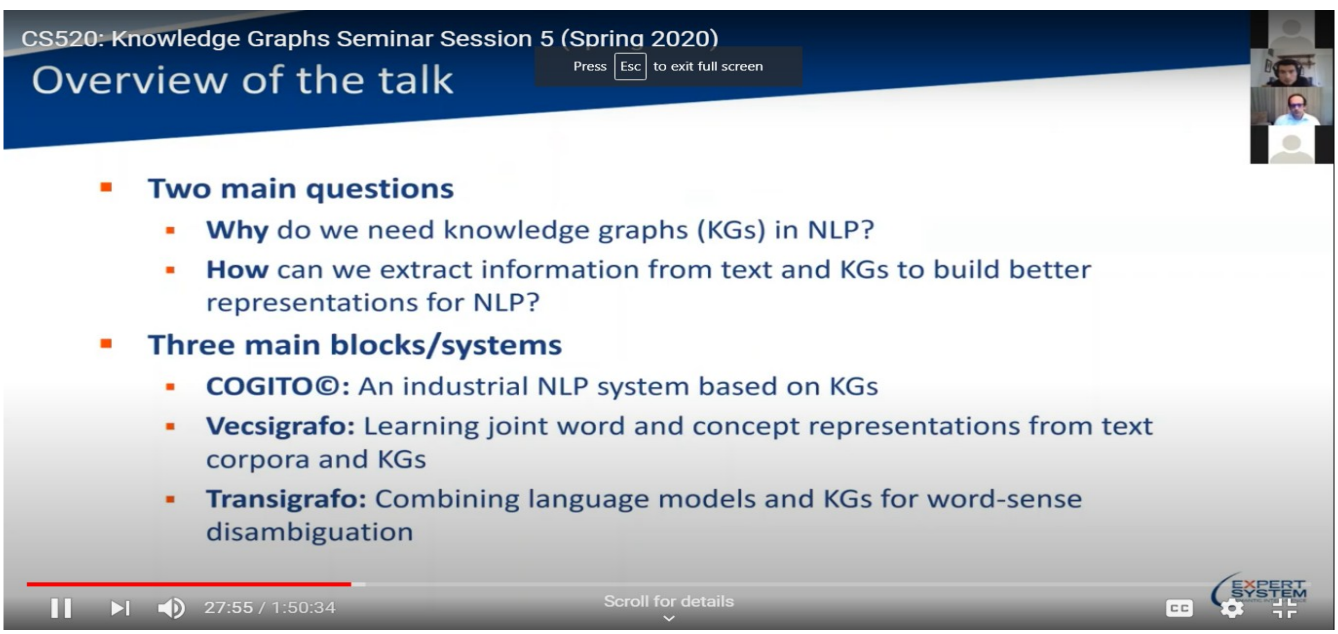
Automation is your friend here: pipelines, automated orchestrate schema release, help with rollbacks and emergency fixes.

Example of a perfect storm situation with corona virus where things were changing rapidly. The schema evolved quickly as different sources of data popped up. Parking lots now became medical testing areas.

They could keep up with a KG and doubt they could with Dbs.



José Manuel Gómez-Pérez KG(s) for NLP



Constrast of 2 approachhes – KB and Neural

Demo of Cogito to understand text on the 2016 US election.

Extracts details on people you can drill down on and visually shows places

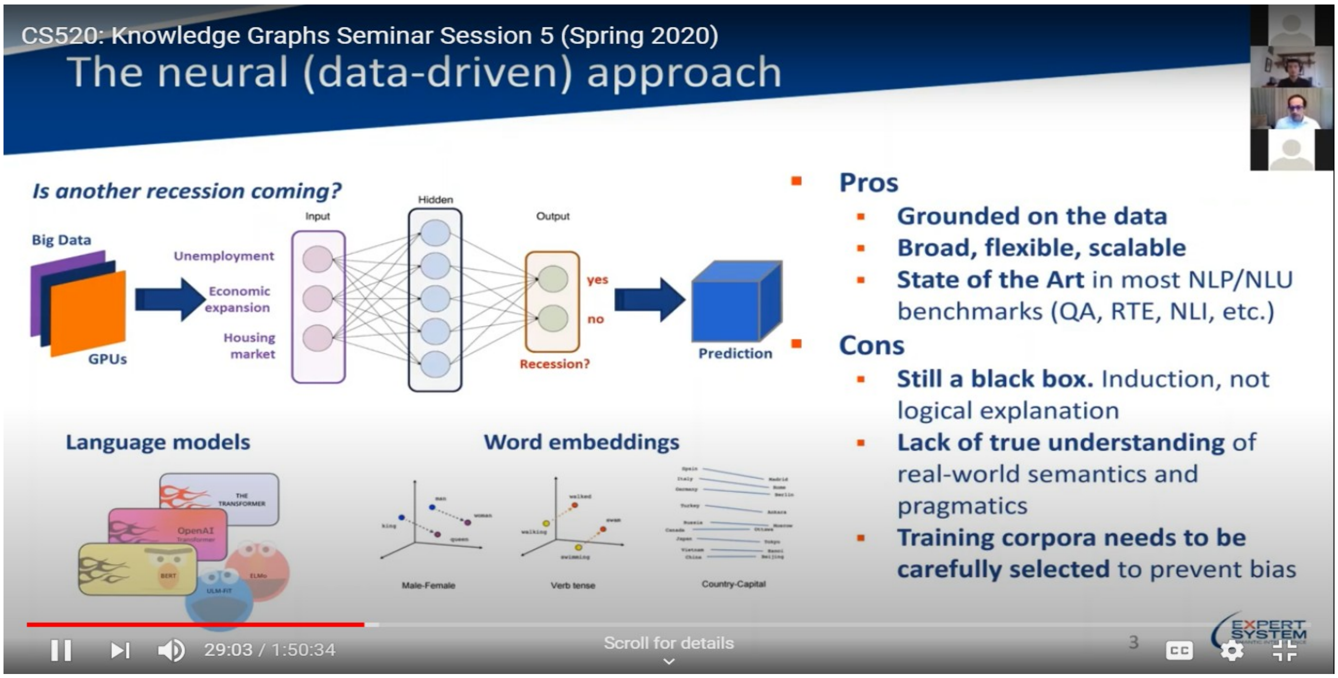
Demo of Vecsigrafo learns words and concept embeddings in a shared vector space.

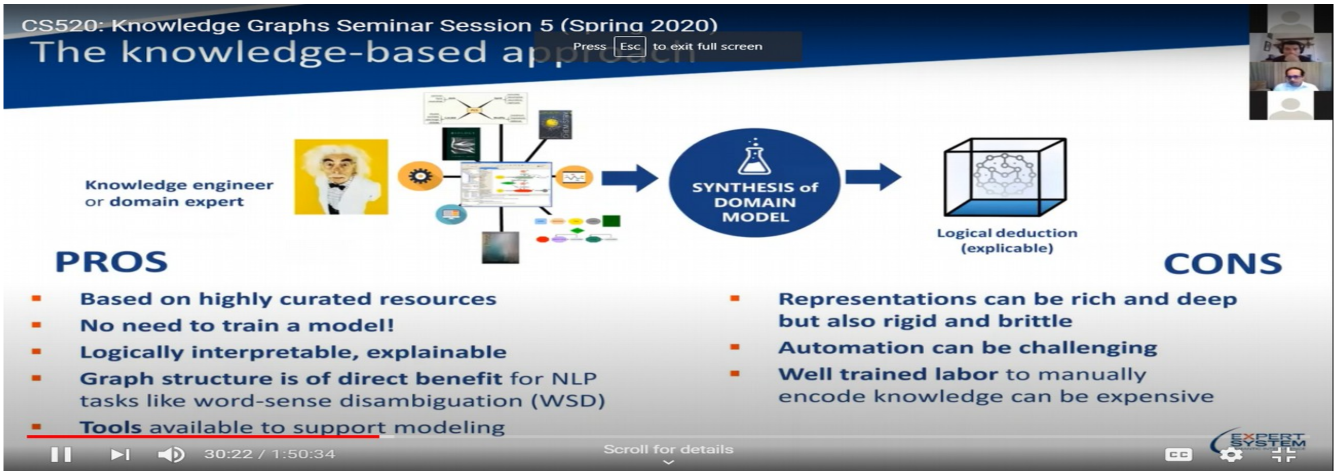
Can be used to detect bot messages rather than human.

Some progress on cross lingual mapping between languages but KG(s) differ in size between 2 languages.

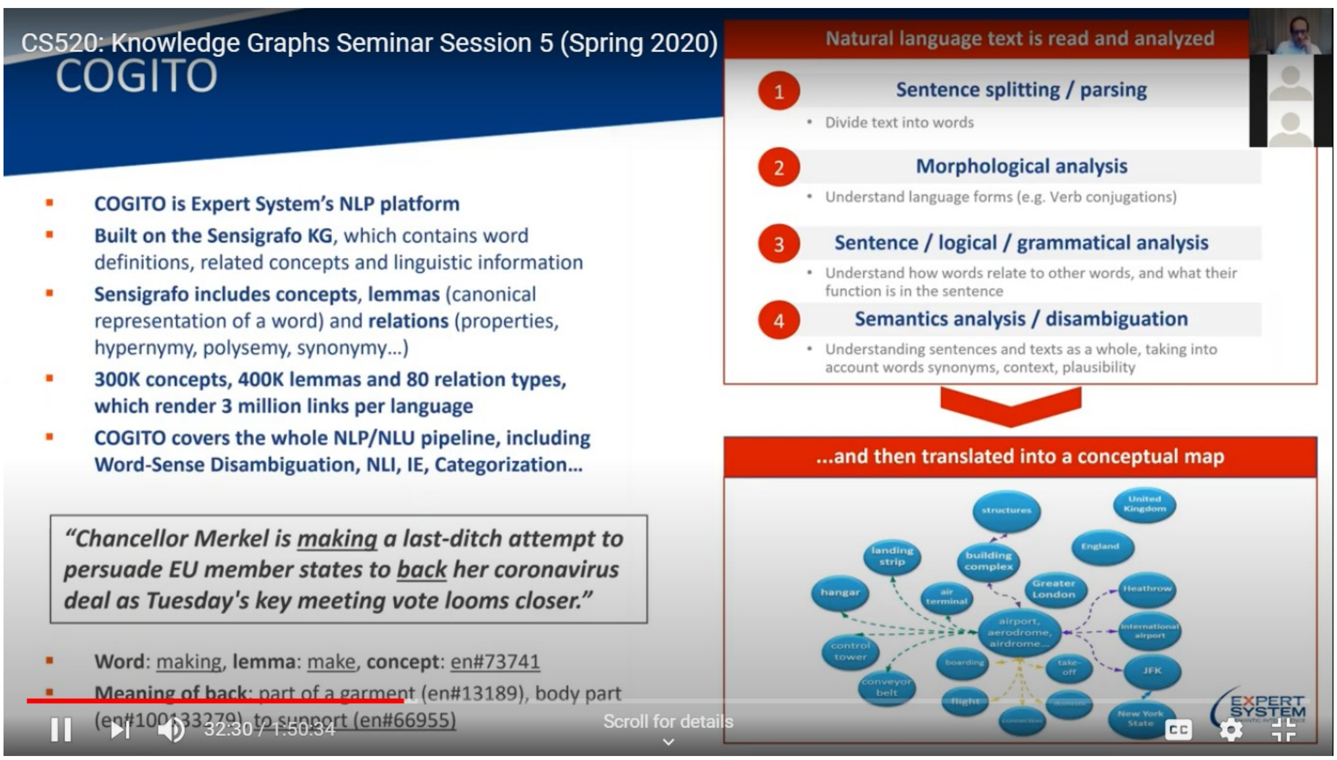
Also there are different strategic decisions made in the 2 different languages.

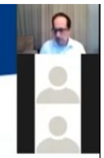
Still challenging to co-evolution tasks.





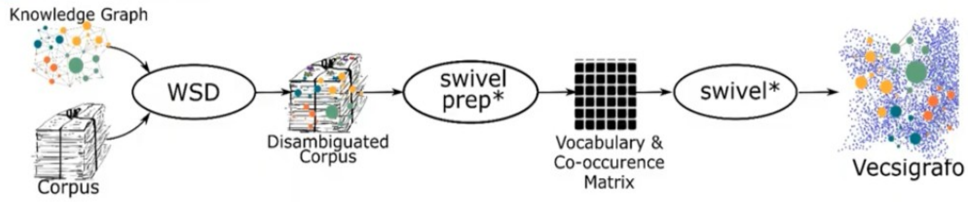
Co-evolution example = concepts like financing work





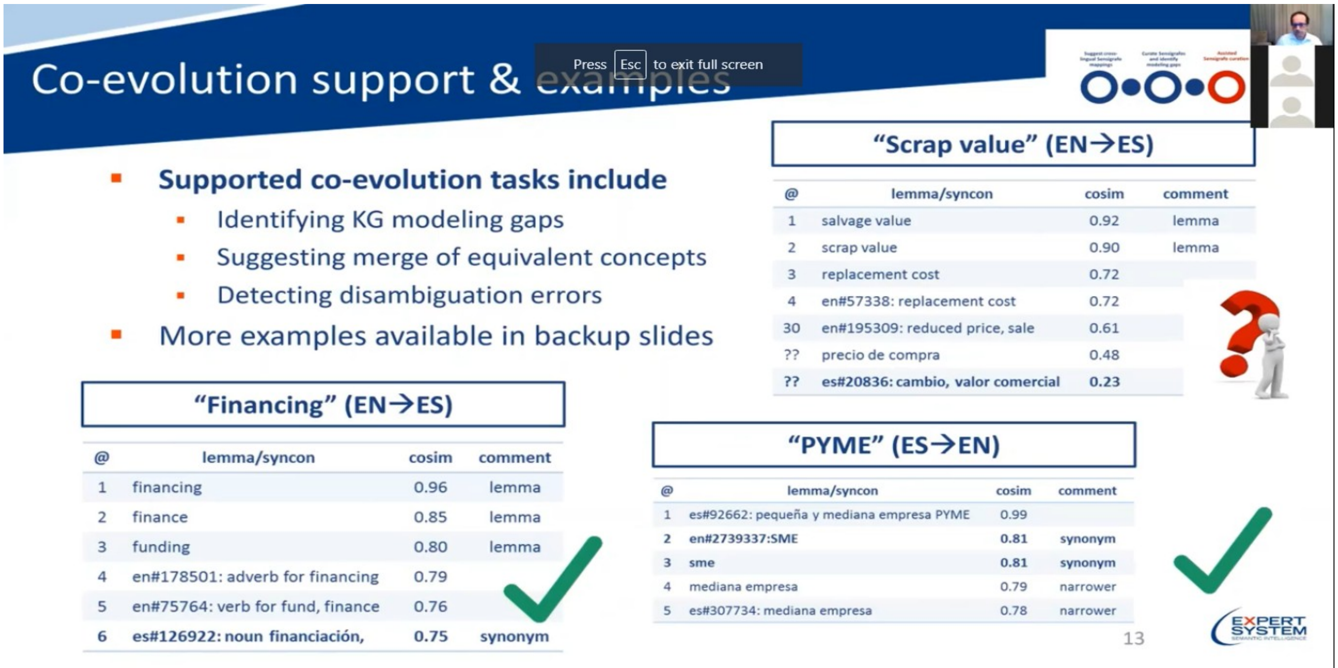
Press Esc to exit full screen

Vecsigrafo: Crossing the NLP chasm

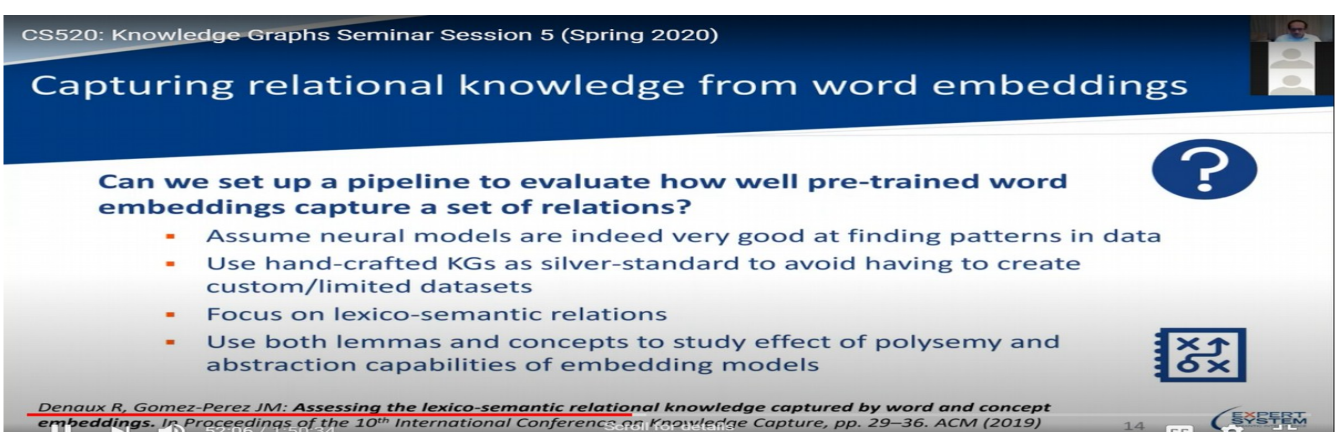


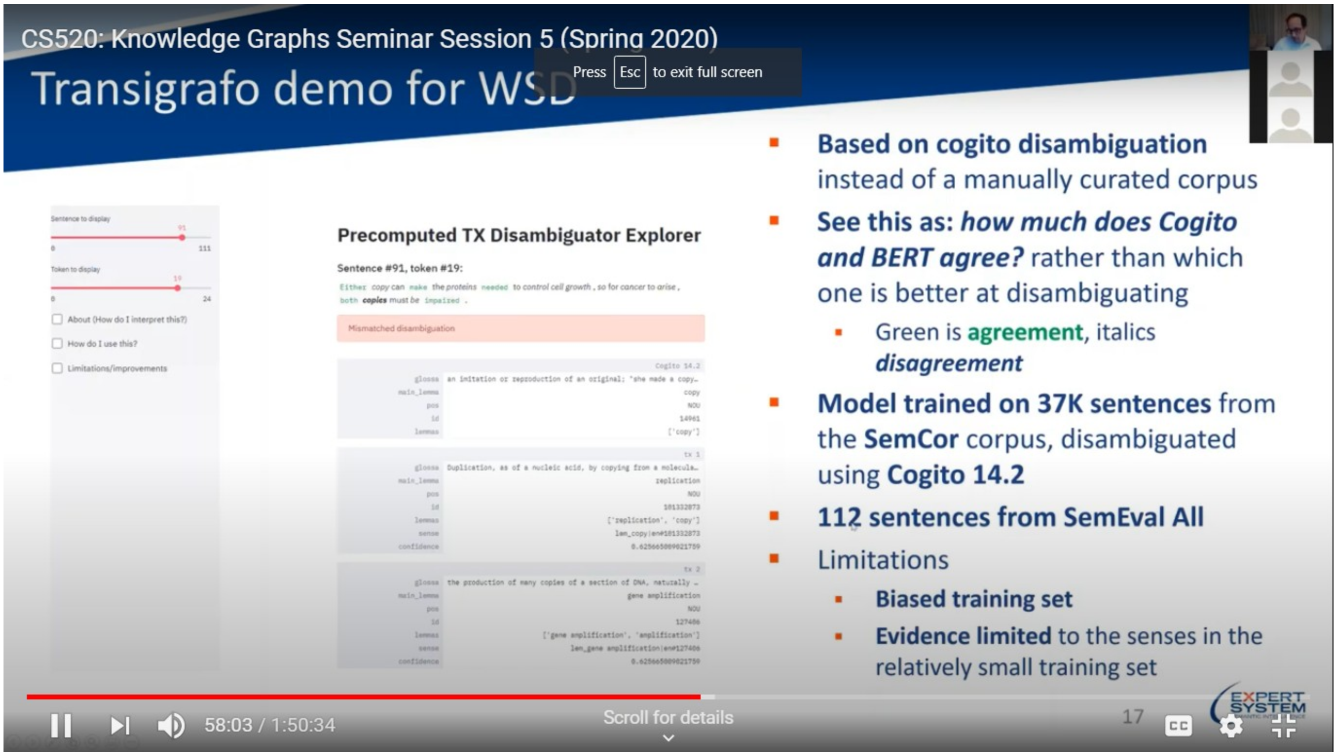
- **Vecsigrafo jointly learns word and concept embeddings** in a **shared vector space**
- Unlike knowledge graph embeddings, Vecsigrafo combines **corpus-based and graph-based** approaches **to build enriched word representations**
- **Extends (*) the Swivel algorithm** with both lexical and semantic entries as part of the vocabulary
  - **The corpus is lemmatized following different tokenization strategies, disambiguated and expanded**
  - **Word and semantic (KG lemma and concept) embeddings are then jointly learnt**

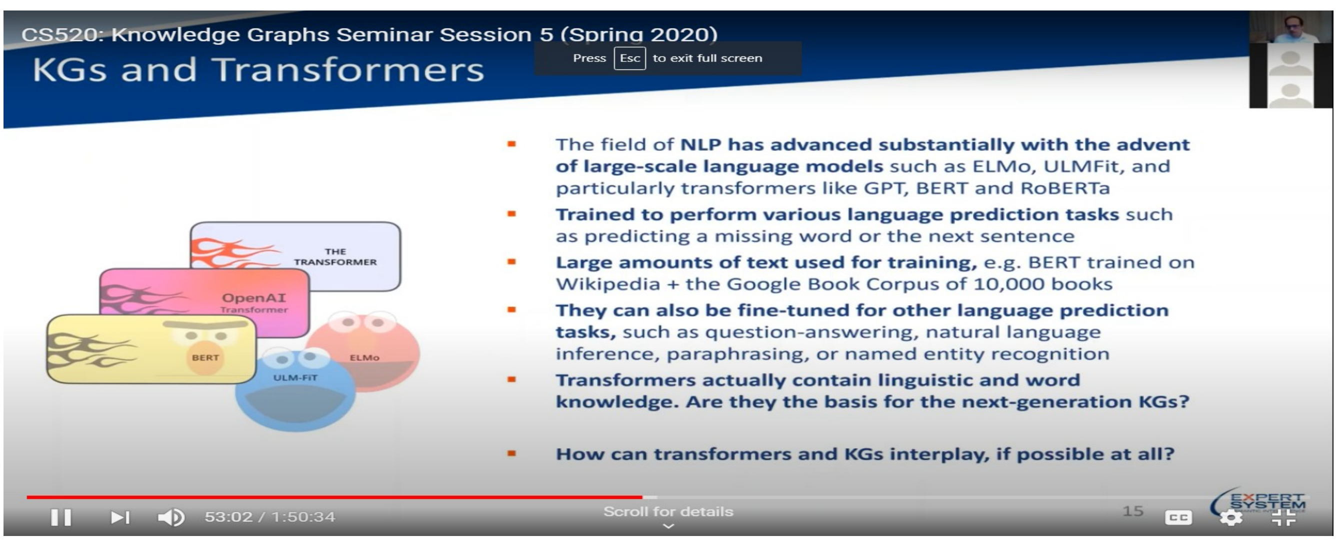
Denaux R and Gomez-Perez JM. 2019. ***Vecsigrafo: Corpus-based Word-Concept Embeddings.*** *Semantic Web Journal (2019), 68–96. https://doi.org/10.3233/SW-190361*



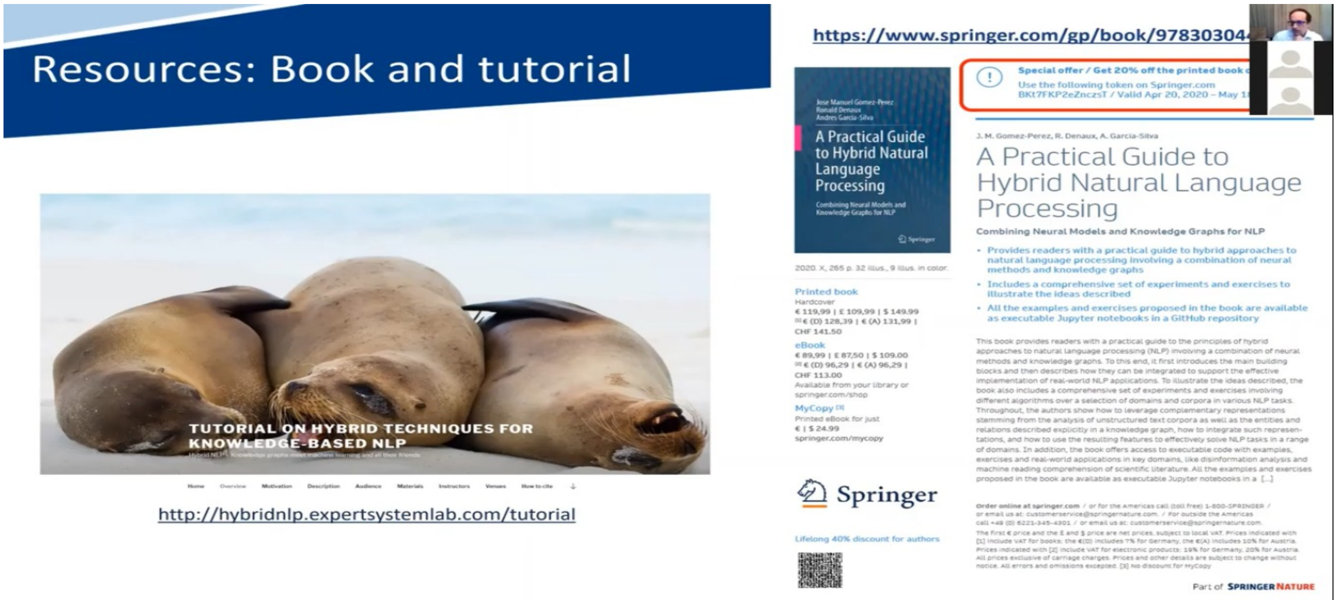
Transigrafo Decoupling linguistic a(how sentience are built and KR (concept model) allows advance







Summing up sources



Mike Uschold (briefed us) Obliterate Silos with KG

Silos are a big problem in any enterprise. They arise from limitations in relational database technology including the lack of explicit semantics. We describe how to avoid silos using ontologies and knowledge graphs. We show how it works in practice and illustrate with case studies. We warn against the use of these newer technologies to gain a local advantage in an organization but ultimately recreating silos across the wider enterprise. The use of an enterprise ontology as a schema to populate an RDF-based knowledge graph opens the door to removing silos and never creating them again. The technology is mature and ready for prime time.

The slides for the presentation are available here.

Knowledge Graphs Seminar Session 6: **How do users interact with knowledge graphs?**

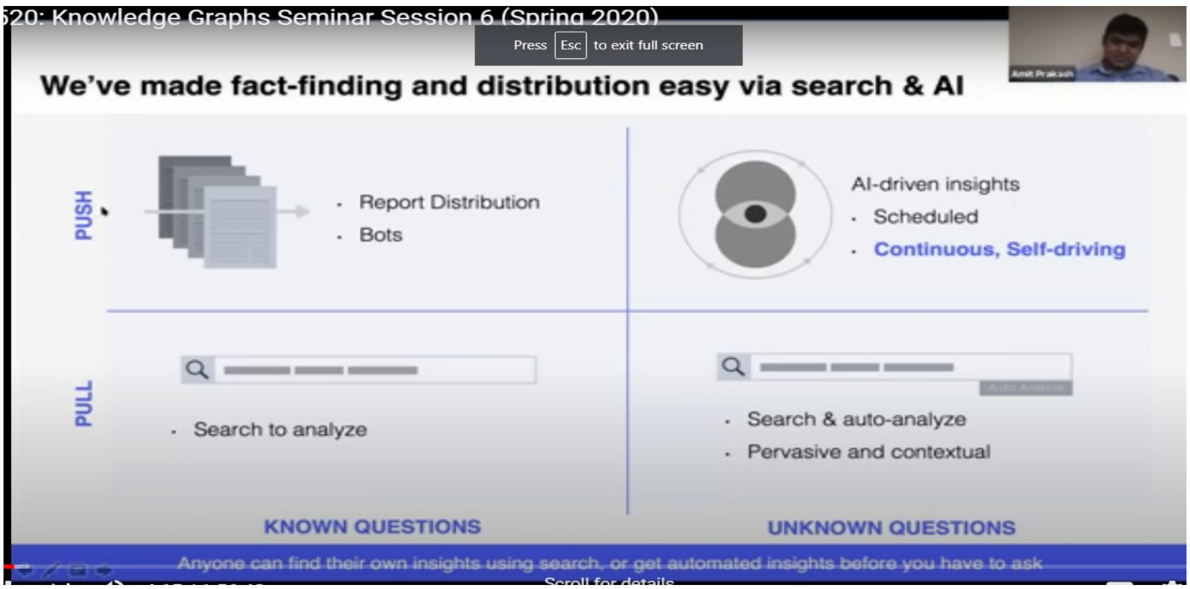
Amit Prakash (early BING work) Search and AI Driven Analysis (not really knowledge graph oriented

System called ThoughtSpot to help people get to the facts. Most systems are static, data driven

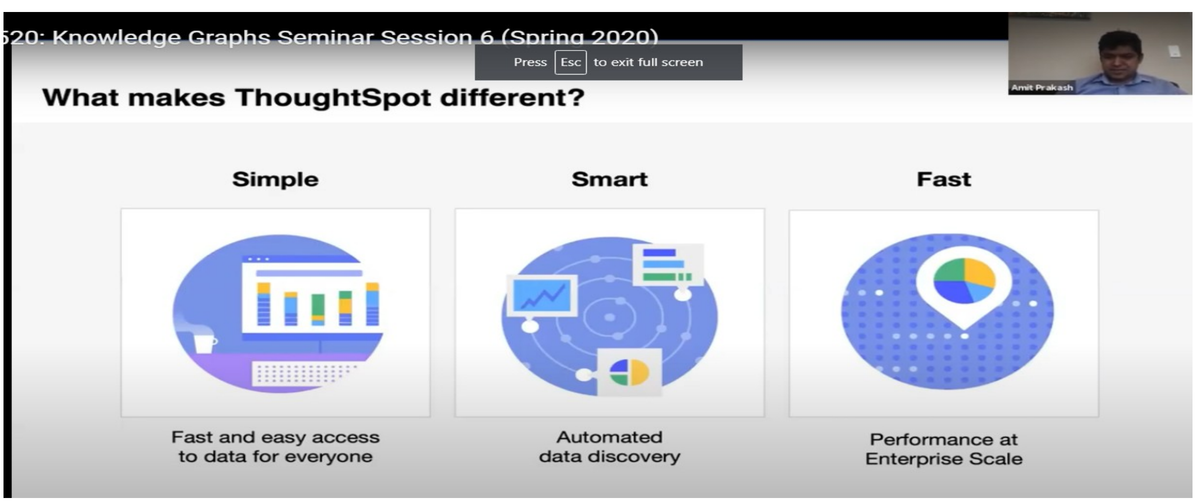
and broken.

Not good for answering a sequence of questions. And people aren't knowledge experts and don't know how to ask and design a questions to system.

Most systems can't handle Big Data.



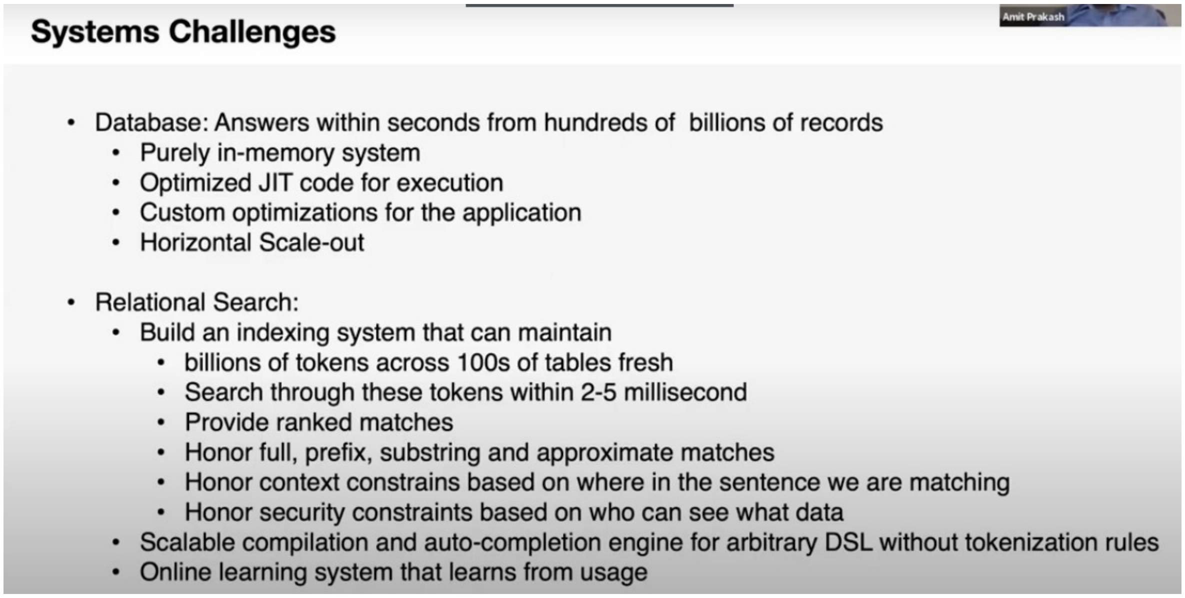
How it is different:



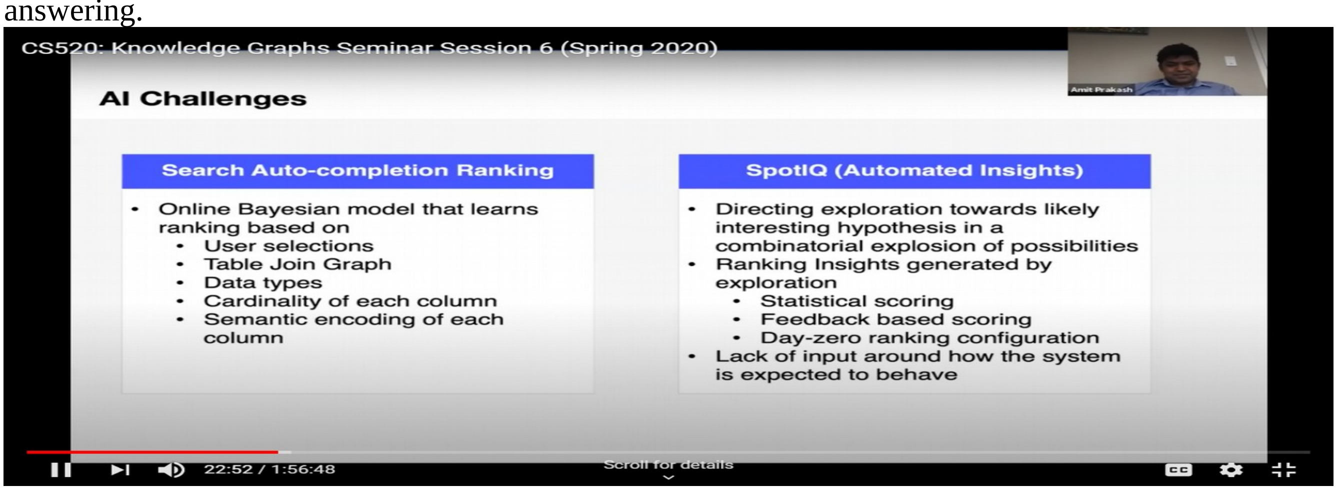
A quick Demo

Search IQ is a product in development that might use a KG.

You can teach the system as it goes along so it gets smarter.



Search IQ involves NL (using a NN, parse tree and pattern matching) and real world knowledge. How to bootstrap it with knowledge and how can it fail gracefully letting the user know it has trouble answering.

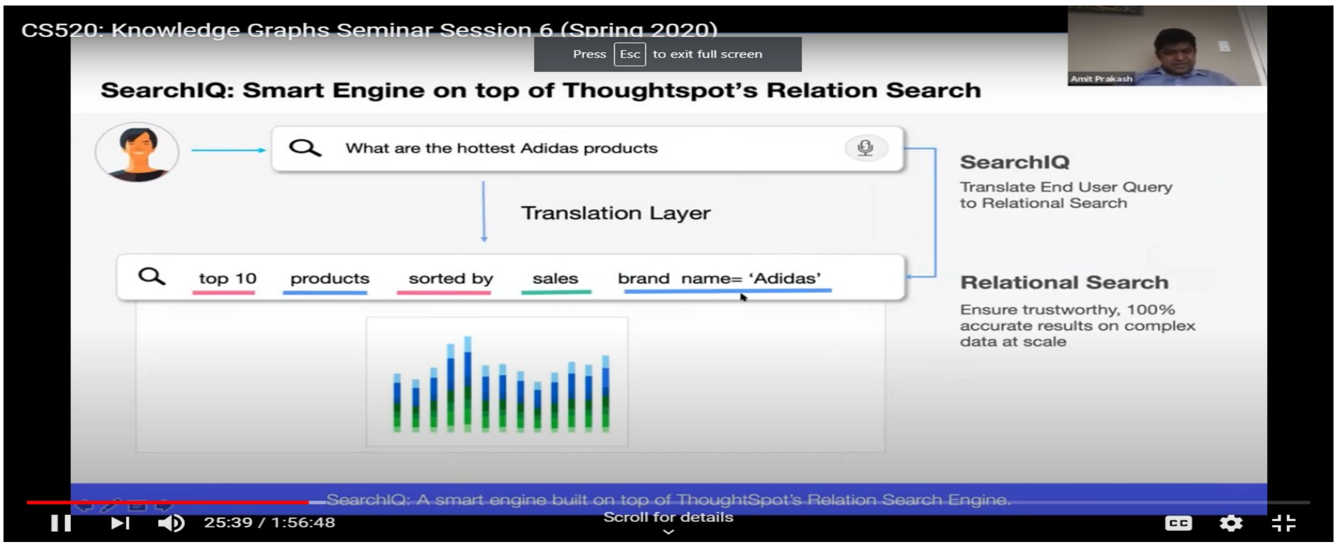


This provides an opportunity to teach the system with a user feedback “, vs “what’s the longest movie name?”

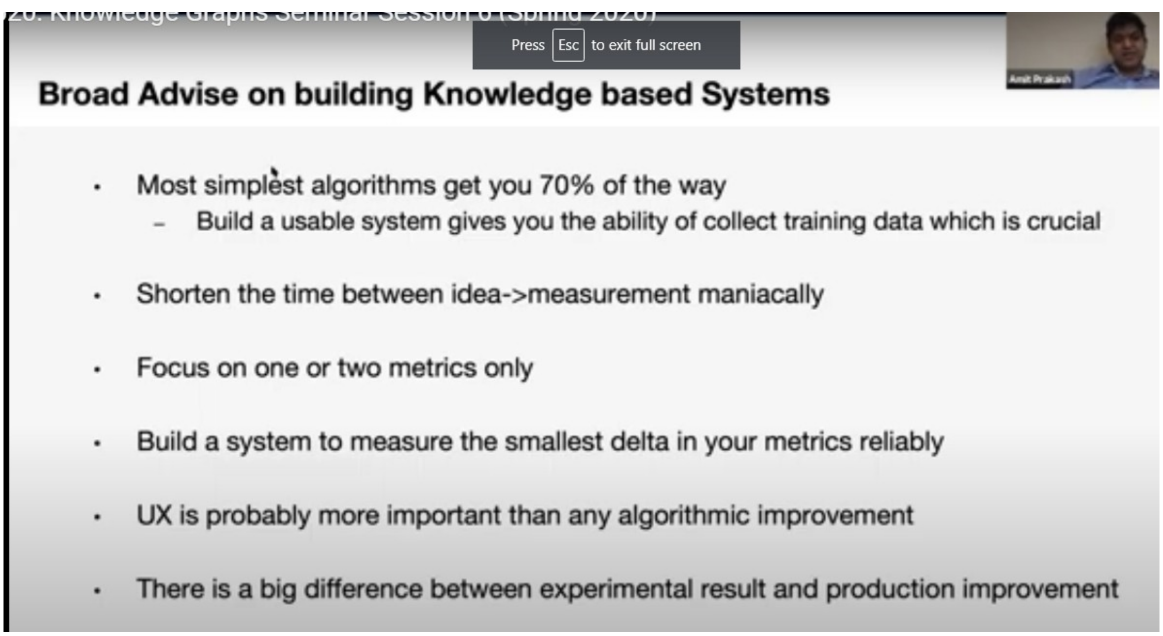
Another example is Qs with best and worse. How to sort these for users? Worst should be last?

There work includes:

- mining entities from text
- context and
- how to support transfer learning with ML technology.



Take aways:

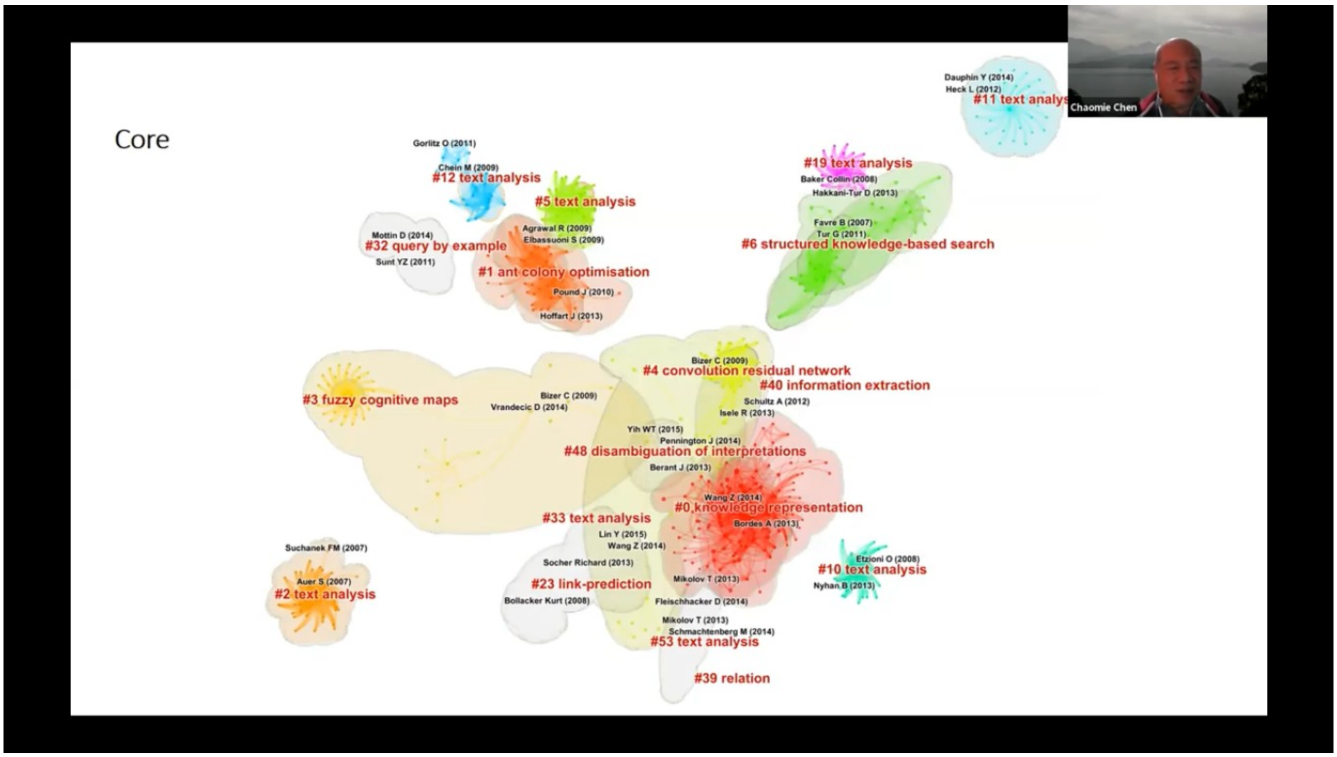


Chaomei Chen Making sense of a field of research?

What has been done in KG research? Graphical view when you search the field

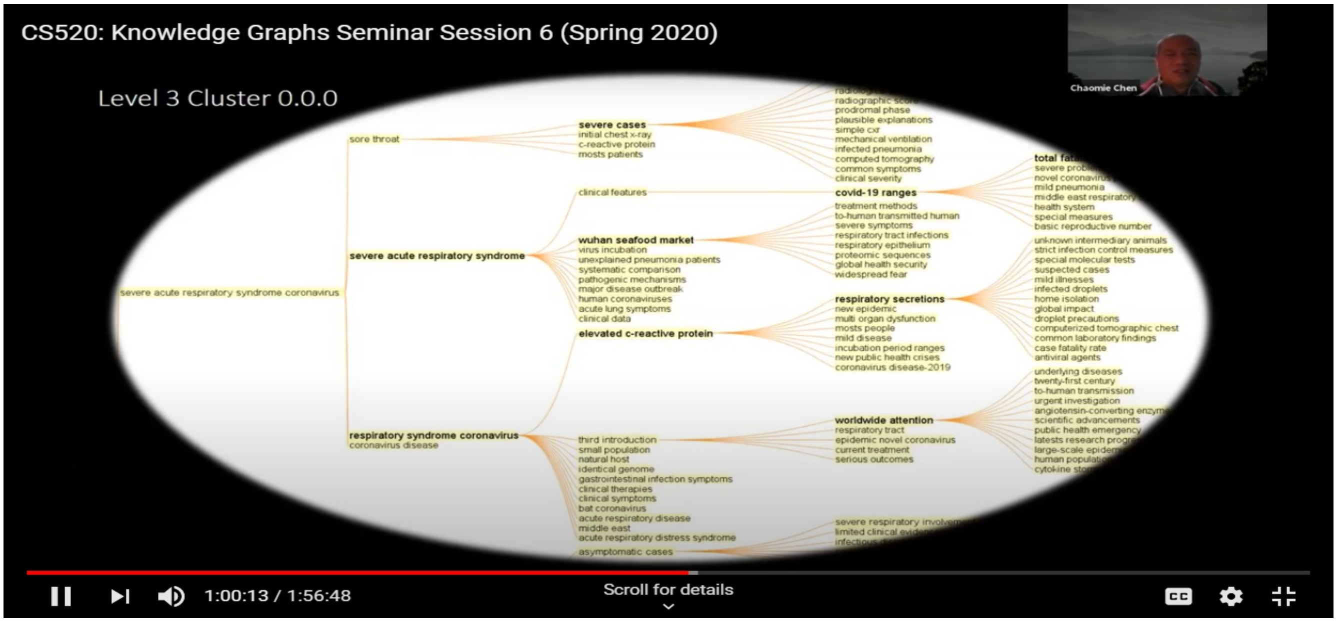
Lens.org and CiteSpace -We can see labels on the clouds of areas that involve KG(s). A core area is knowledge representation.

We see some of the neighborhoods -like CRN and text analysis.



You can also look at how the core has changed over time. KG and corona virus used as example to see clusters of concepts and articles on these topics such as vaccine development.

There are 2 or 3 levels of detail that you can go to and get there interactively.


CS520: Knowledge Graphs Seminar Session 6 (Spring 2020)
Chaomie Chen
Level 3 Cluster 0.0.0
severe acute respiratory syndrome coronavirus
sore throat
severe acute respiratory syndrome
respiratory syndrome coronavirus
coronavirus disease
severe cases
initial chest x-ray
c-reactive protein
mosts patients
clinical features
wuhan seafood market
virus incubation
unexplained pneumonia patients
systematic comparison
pathogenic mechanisms
major disease outbreak
human coronaviruses
acute lung symptoms
clinical data
elevated c-reactive protein
third introduction
small population
natural host
identical genome
gastrointestinal infection symptoms
clinical therapies
clinical symptoms
bat coronavirus
acute respiratory disease
middle east
acute respiratory distress syndrome
asymptomatic cases
radiographic
prodromal phase
plausible explanations
simple cxr
mechanical ventilation
infected pneumonia
computed tomography
common symptoms
clinical severity
covid-19 ranges
treatment methods
to-human transmitted human
severe symptoms
respiratory tract infections
respiratory epithelium
proteomic sequences
global health security
widespread fear
respiratory secretions
new epidemic
multi organ dysfunction
mosts people
mild disease
incubation period ranges
new public health crises
coronavirus disease-2019
worldwide attention
respiratory tract
epidemic novel coronavirus
current treatment
serious outcomes
novel coronavirus
mild pneumonia
middle east respiratory
health system
special measures
basic reproductive number
unknown intermediary animals
strict infection control measures
special molecular tests
suspected cases
mild illnesses
infected droplets
home isolation
global impact
droplet precautions
computerized tomographic chest
common laboratory findings
case fatality rate
antiviral agents
underlying diseases
twenty-first century
to-human transmission
urgent investigation
scientific advancements
public health emergency
1:00:13 / 1:56:48
Scroll for details

References


520: Knowledge Graphs Seminar Session 6 (Spring 2020)
Press Esc to exit full screen
Chaomie Chen
Related Topics
Cascading Citation Expansion
• Chen, C., Song, M. (2019) Visualizing a Field of Research: A Methodology of Systematic Scientometric Reviews. PLoS One 14 (10), e0223994.
Epistemic and Other Types of Uncertainties
• Chen, C., Song, M., Heo, G. E. (2018) A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty. Journal of Informetrics 12 (1), 158-180.
• Chen, C., Song, M. (2017) Representing Scientific Knowledge: The Role of Uncertainty. Springer.

Leilani Gilpin topic Explaining Explanations (XAI)

Did an large scale literature review.

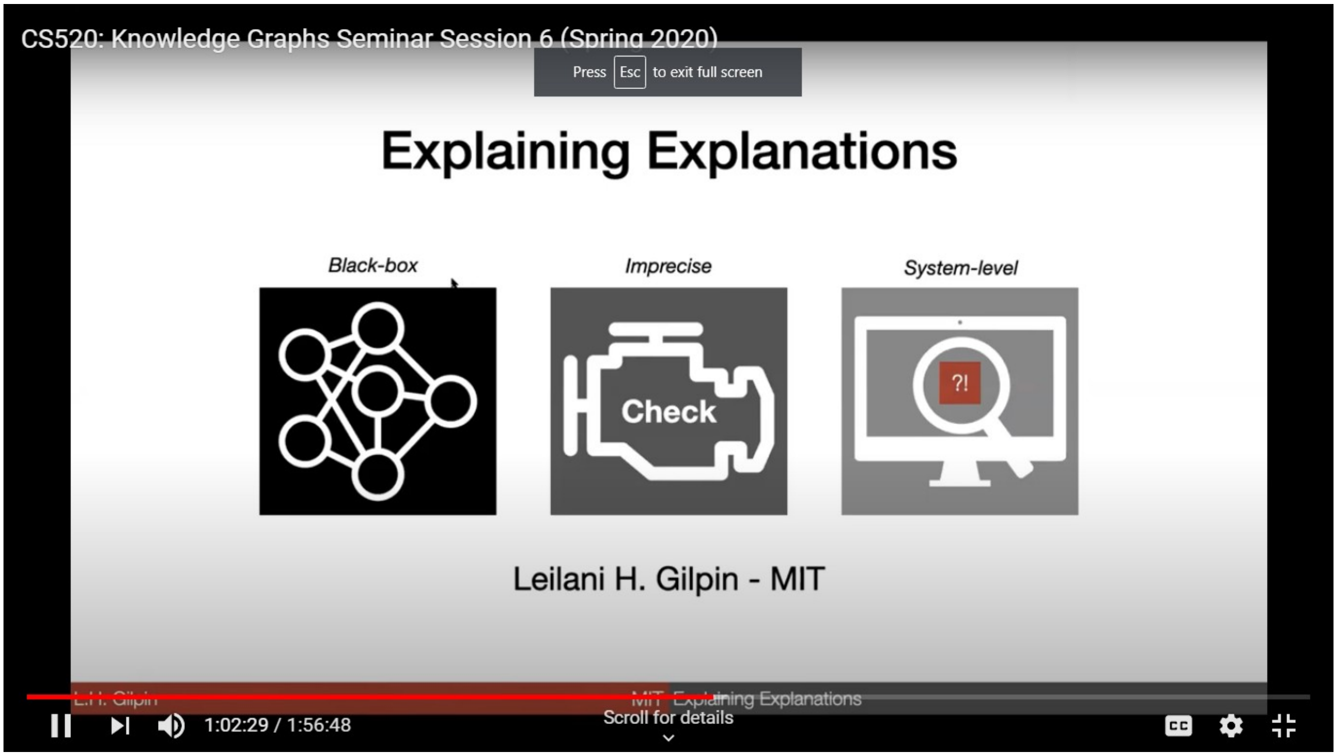
Imprecise explanations where and error occurs.

Why we need explanations? When systems break why and where did this happen.

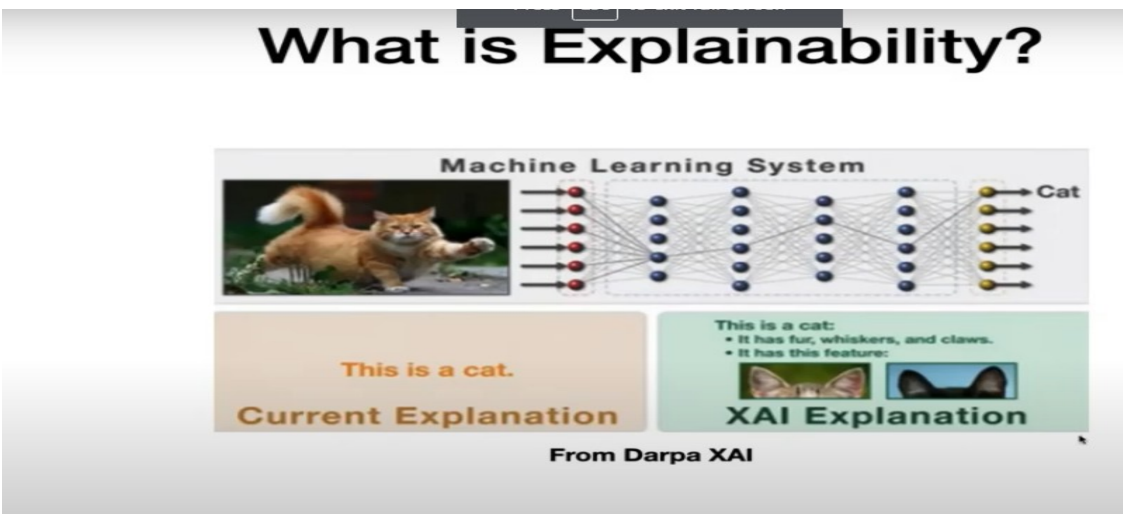
Once we have autonomy on a complex system we can have problems like update OS doesn't allow us to access files.

And the systems lack commonsense. Need for better decisions.

Explanations in AI has been around, but now a big field with deep nets.

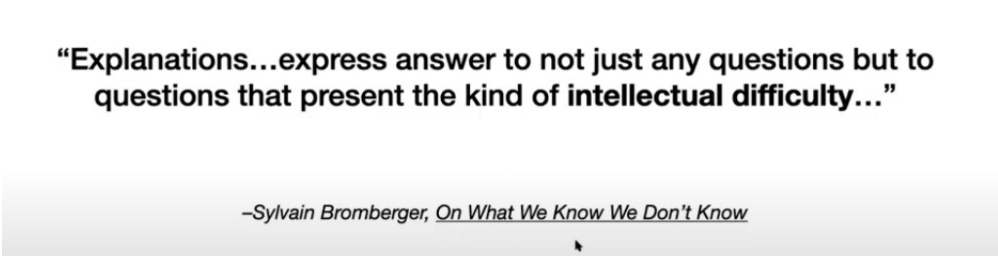


Can't talk about features.



Explanations have to be deeper and give new insights

Philosopher's view:

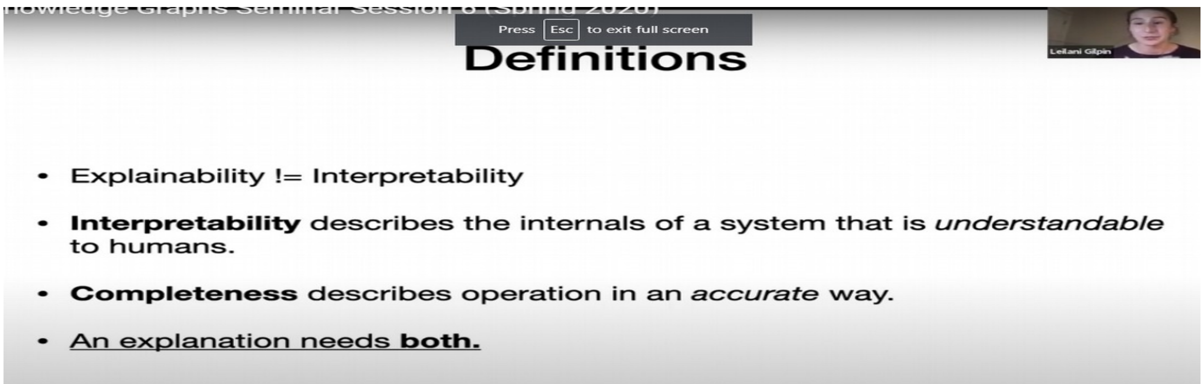


Deep NNs are everywhere now including self driving cars, Go and classification in medical tasks.

So we are used to them but they aren't very understandable.

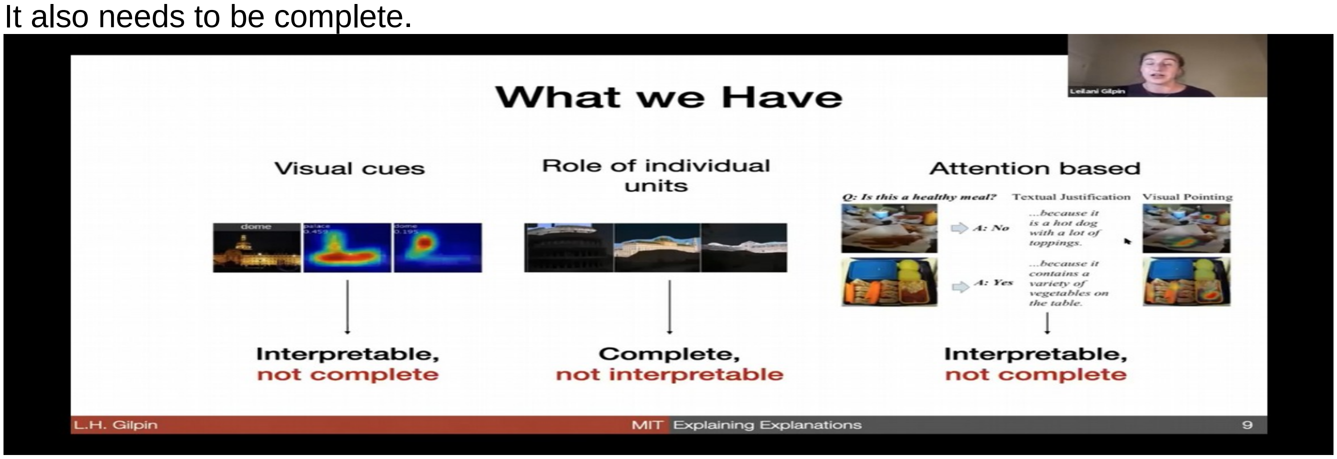
One problem is that there isn't good theory on what goes on or should in the middle layers of a DNN. So when they fail we don't understand why.

A definition of explanation:



it is not the same as interpretability which is about a model like a decision tree.

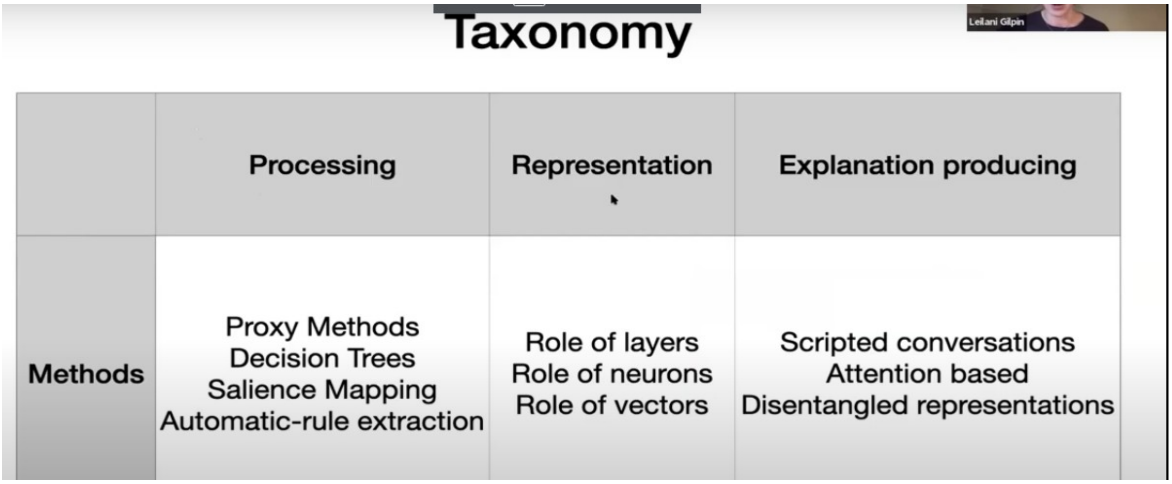
It also needs to be complete.



We don't have good explanation (what is being explained) and this matters say in a fraud detection or errors in cancer classifier – turns out it was largely based on image accuracy.

Visual cues are explaining a different thing than role of individual units (representation).

A



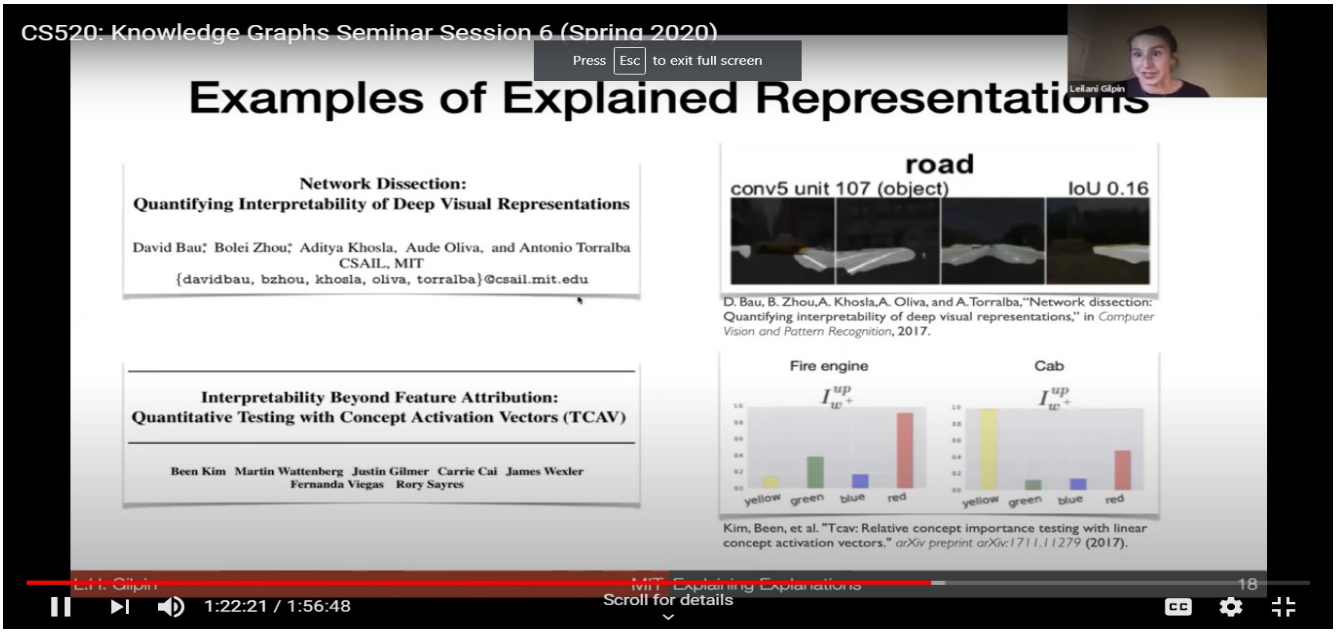
Taxonomy

	Processing	Representation	Explanation producing
Methods	Proxy Methods Decision Trees Salience Mapping Automatic-rule extraction	Role of layers Role of neurons Role of vectors	Scripted conversations Attention based Disentangled representations

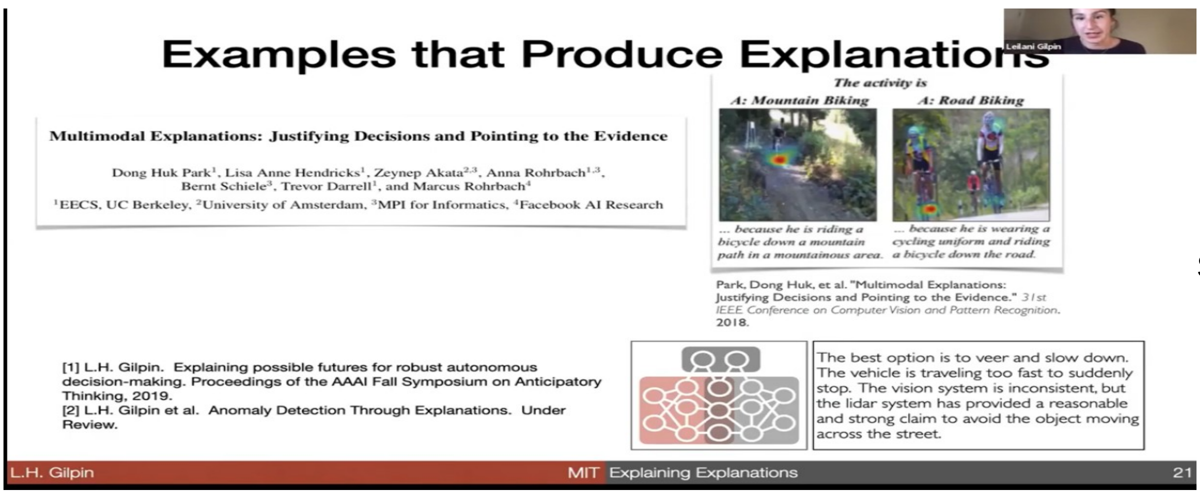
Taxonomy from 87 papers on explanations

Example of methods and what they do well and badly.

DeepRed as an example, a slaiency map looks like a decision tree.

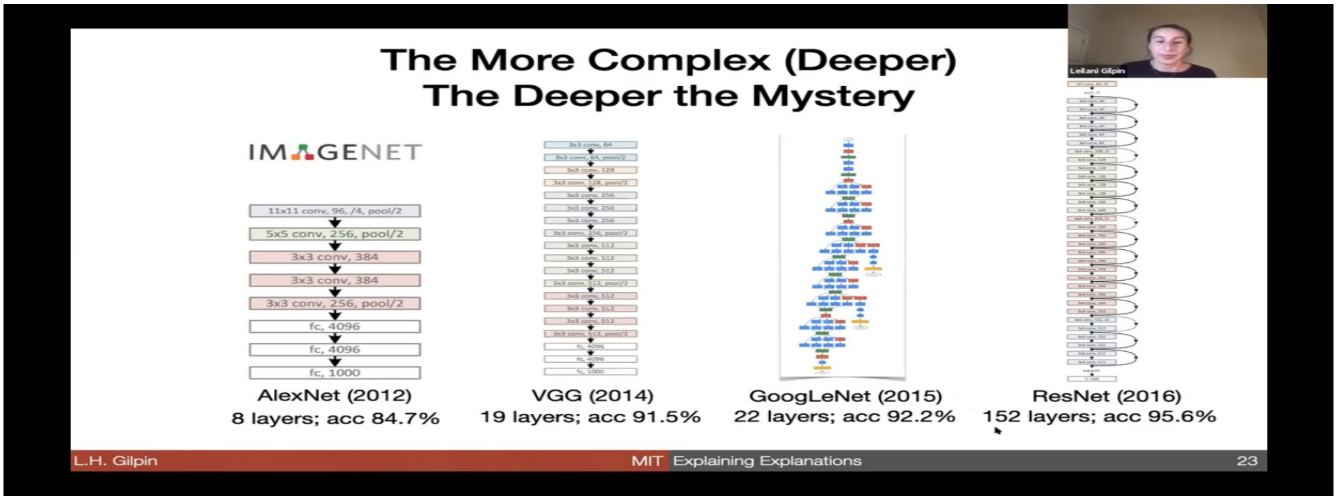


VQA and explanation by default.



So

These approaches are good, but we need to understand it better. Say how we come up with justifications.

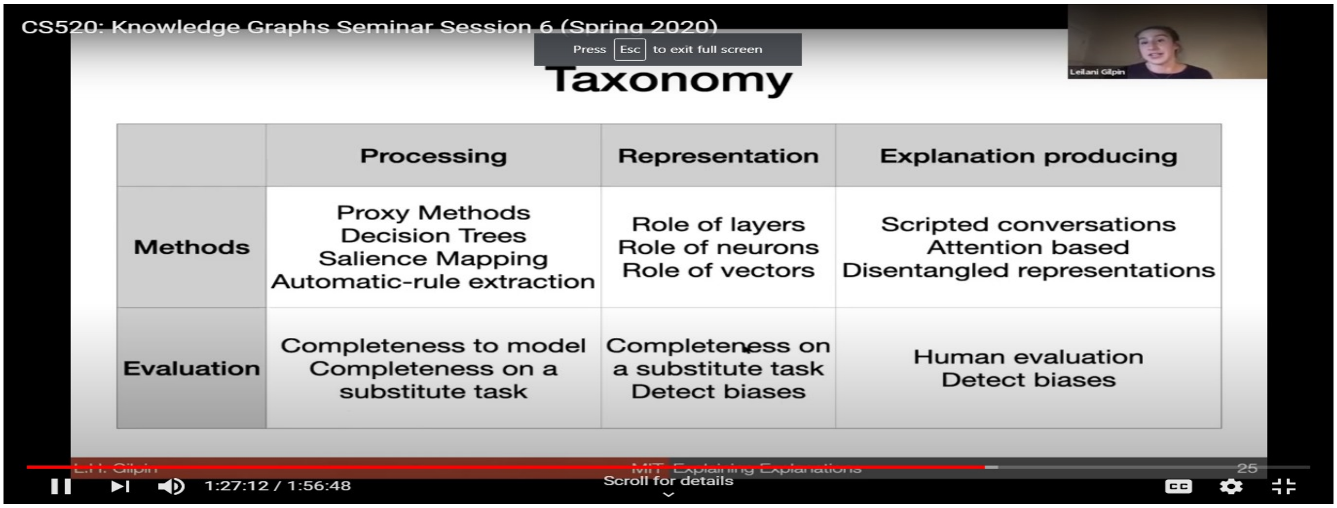


For visual cues you ask how complete it is.

For Individual units it is completion on other tasks

For attention based you need to work with users.

Recommendations – need to understand what is being explained.

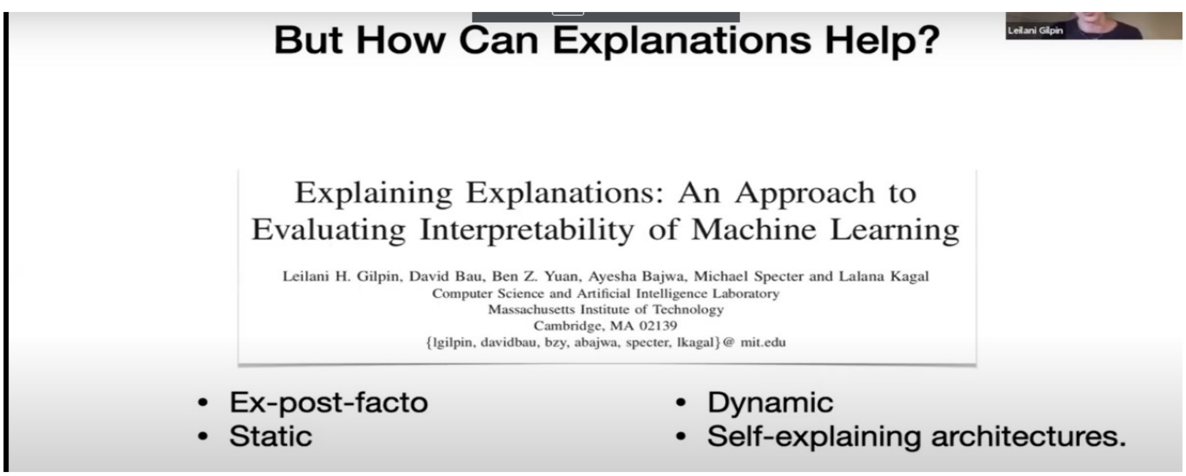


How explanations can help

Challenges with metrics and standards – what makes a good explanation.

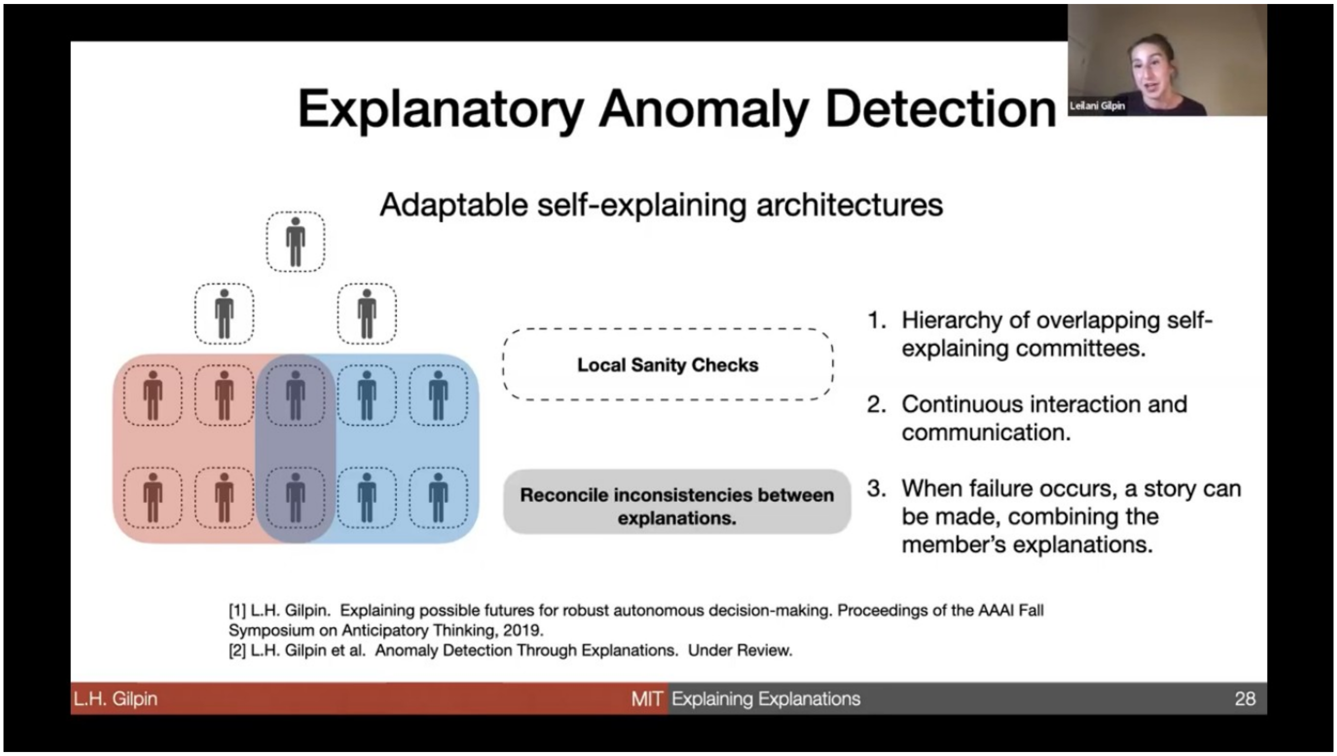
We don't have a fixed data set to test on and gauge metrics. Metrics are usually fuzzy and based on information feedback from evaluations.

Need benchmarks for mission critical apps. How to design scenarios



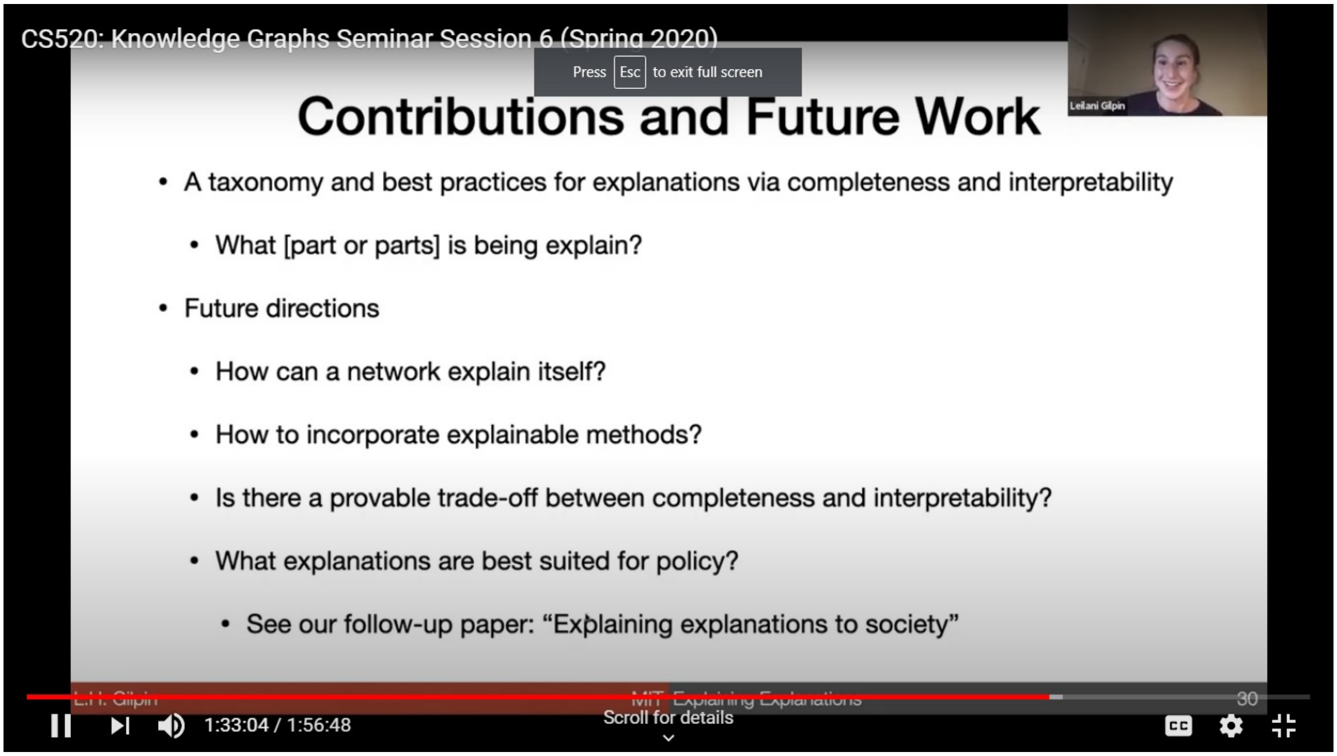
Current work on explanatory anomaly detection.

Thinking about dynamic explanations and architectures to support adaptive learning.



Local sanity test to constantly explain what it is doing and learn from mistakes.

We need to make a NN explain itself.



<u>Questions</u>

Are human explanations complete or even accurate?

Can’t peer into brain to check. But even if we don;’t know how we arrived at an explanation we can supplement these with commonsense. I feel hot and similar to how I felt when sick.

Can we construct a complete model of a research domain?

A. Globally this is hard and maybe philosophical. We get local views.

Q. A system that reaches a human interpretable explanation...are we losing somethings?

A. There is a tradeoff but it is open we don’t know and don’t have a base to test.

Q Do we look at counterfactual explanations? I didn’t make a left turn. What would have happened?

A. Yes we are looking at that. Example,“You were denied because you didn’t make a payment in 2012.”

Q. Do Q and A systems handle counterfactuals?

A. Not yet.

Q what is an explanation is a lie? Say it reads a false billboard.

A. We based some on physics. But we need user feedback.

Q. What level of explanation is useful?

A. we have very verbose ones, but it is user sensitive. Some need terse answers.

Q. Is Ontological precision/validity used?

A, Mostly not. We have a traditional KB, haven't looked at faults. We do some as part of reasoning, some generalization from embeddings of words.

Q. Different approaches to handle open vs. open situations

CS520: Knowledge Graphs Seminar Session 7

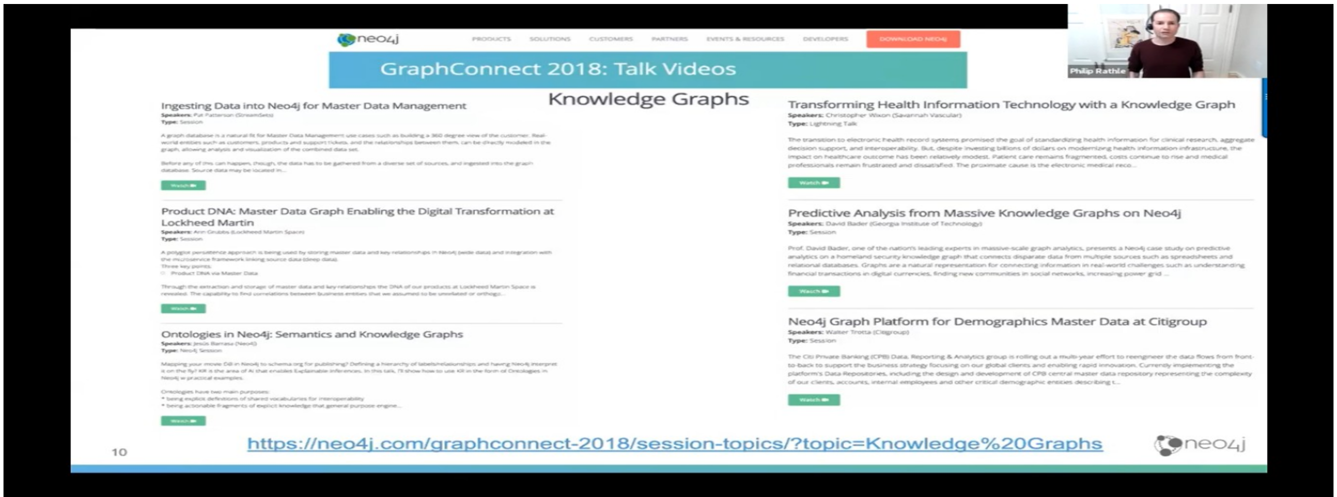
Implementing Knowledge Graphs in Products

What are some prelevant graph engines in industry?

Philip Rathle (Neo4J) Property Graphs (Neo4J browser)and Graph Algorithms

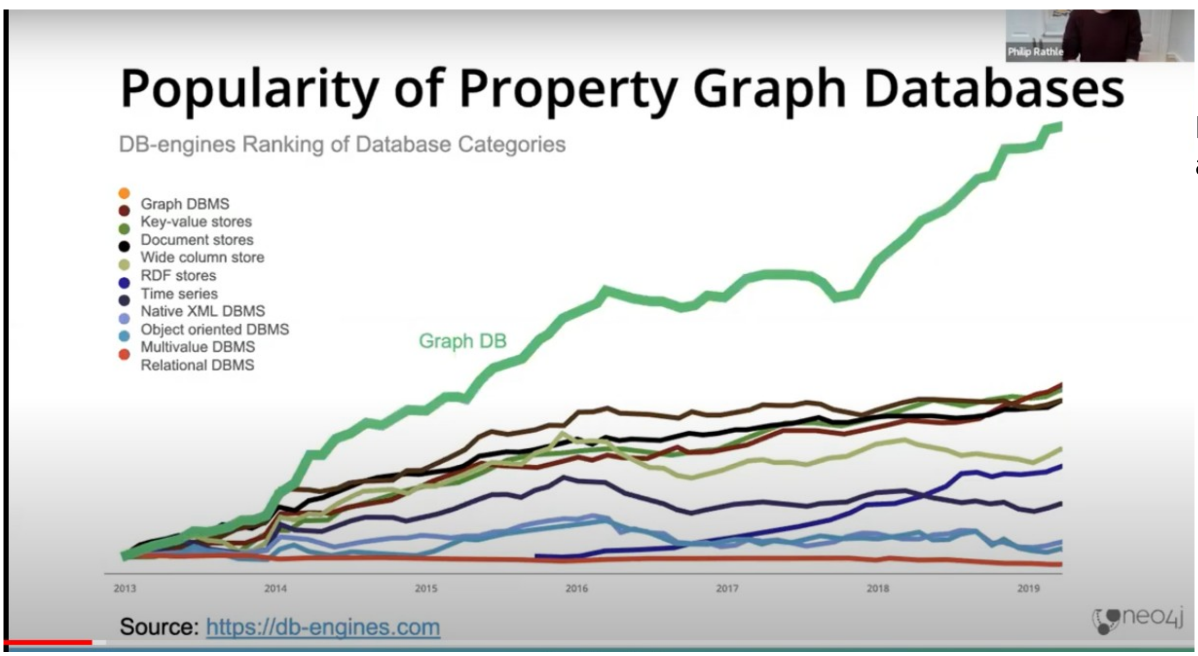
NASA and Ebay Conversational systems, or German Center for Diabetes Research (covidgraph.org), ICIJ on Panama papers of 2.6 TB of 11.5 million documents curated into a KG.

People consider many types of graphs to be a KG



Evidence for property graph popularity – motivated by data storage & management interests, also querying and interest of developers and applications. This is a different vision than RDF,

PG nodes and relationships have internal structure unlike with RDF.

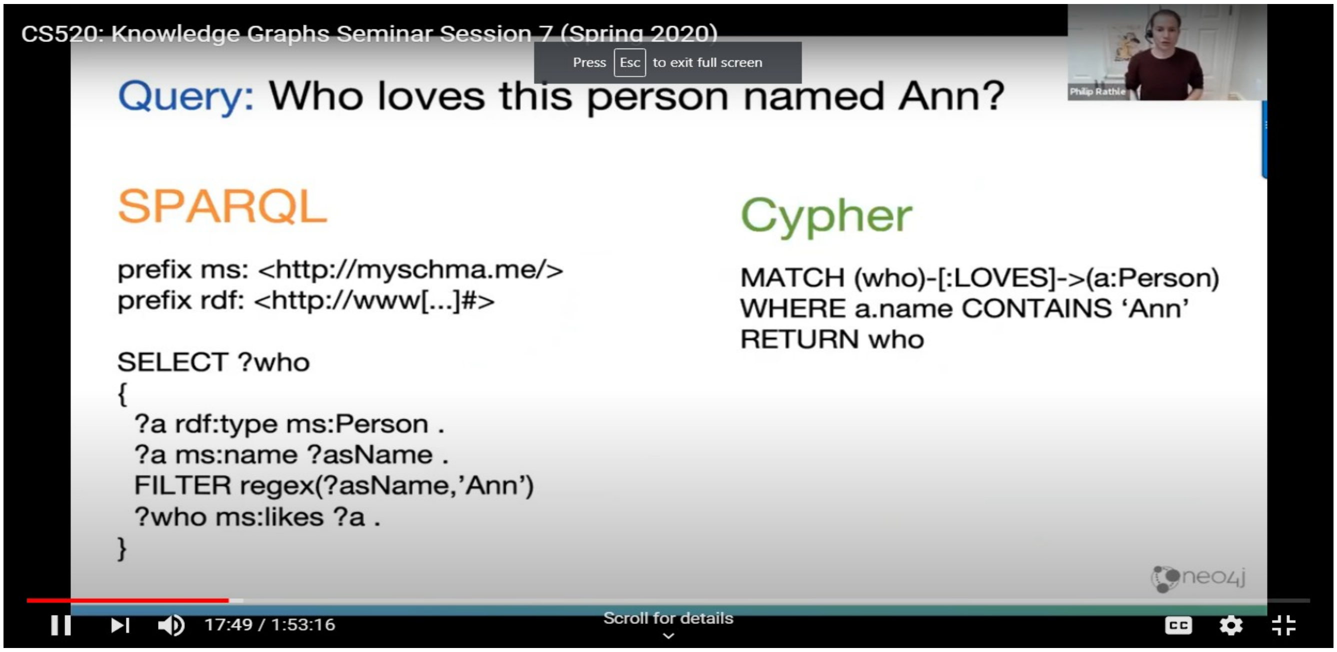


Nodes and

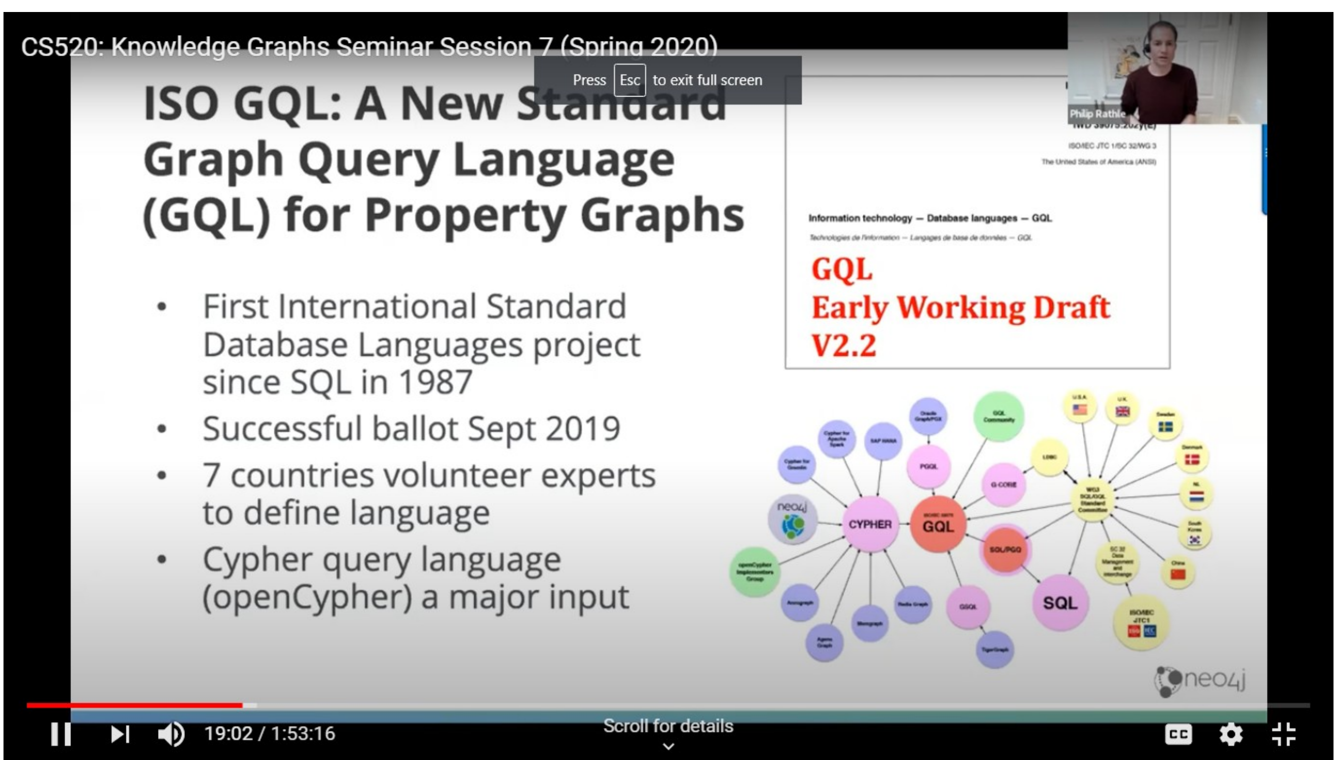
Relations are synonymous with vertex and edge.

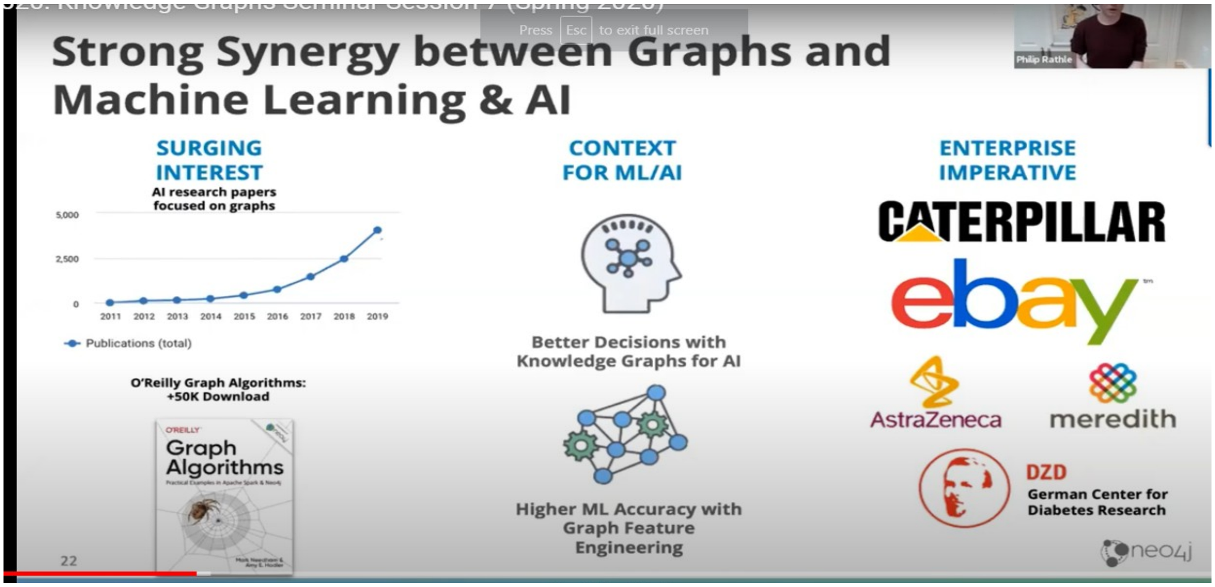
In PG RDB rows become nodes and join tables become a relationship.

An open language based on patterns is Cypher.

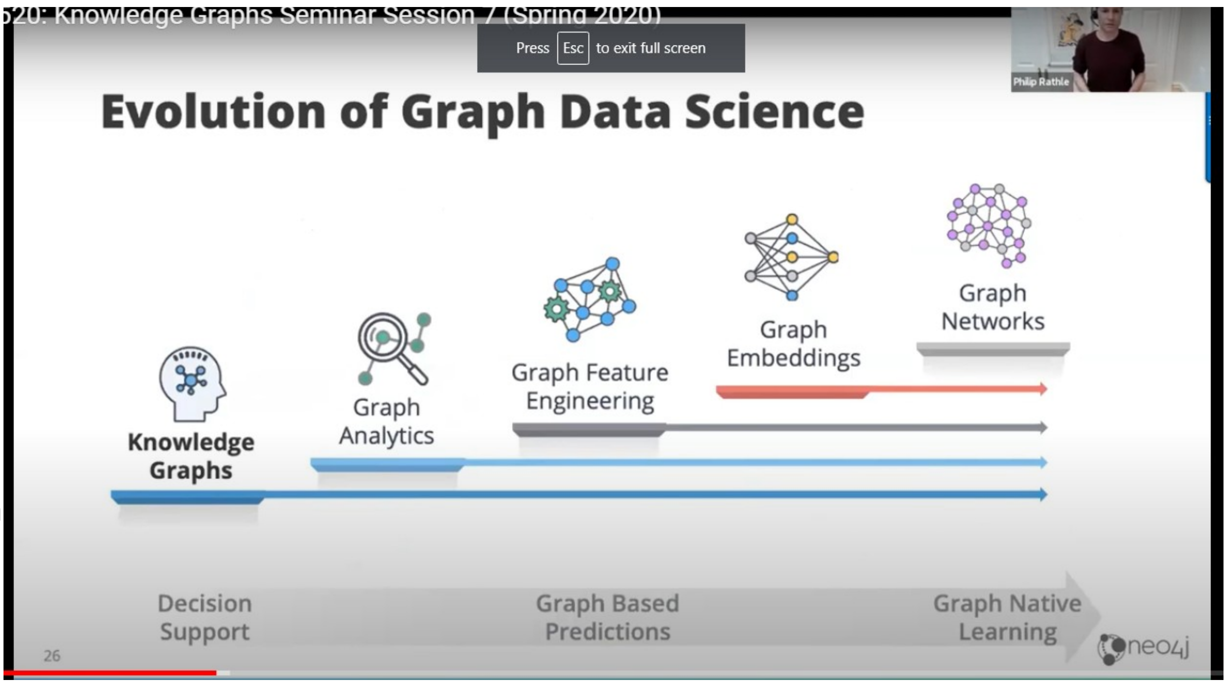


Big space for this work between AI and KG – graph data scienc. We learn about a node in a network by the nodes around it.



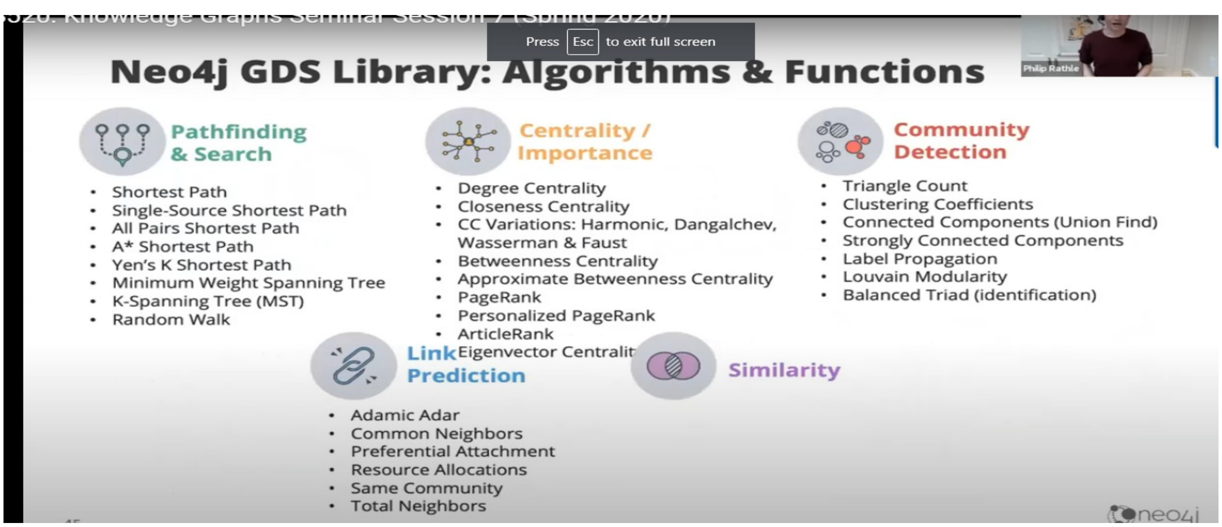


We
see
evolution in areas like graph feature engineering.



The
graph

library with lots of algorithm and functions. There is an ecosystem see neo4j.com



A variety of customer use cases discussed with benchmark results for queries.

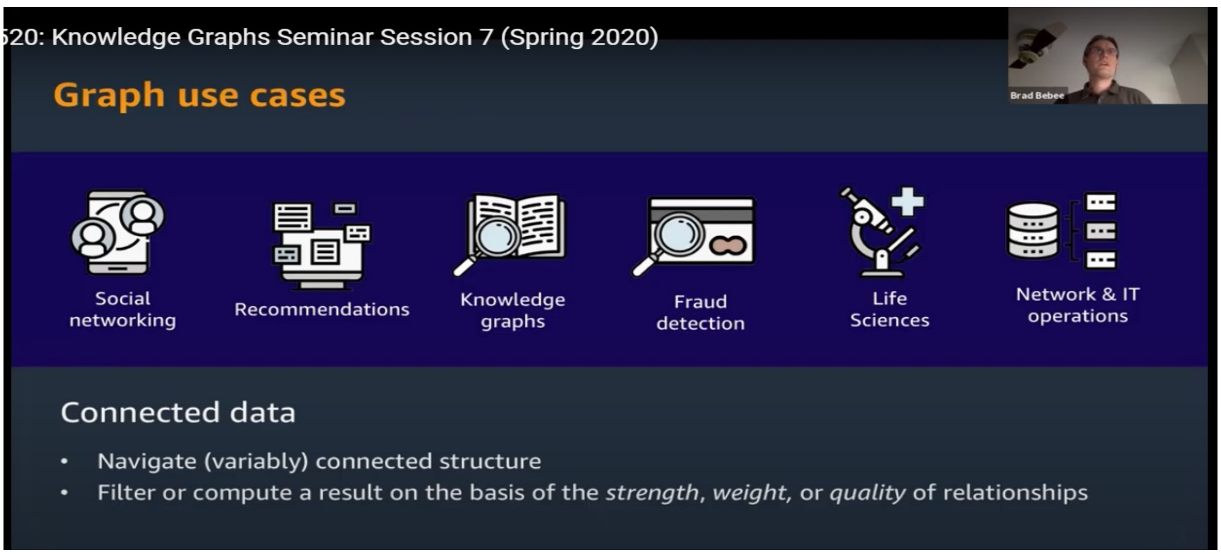
Brad Bebee Putting data into context with Amazon Neptune

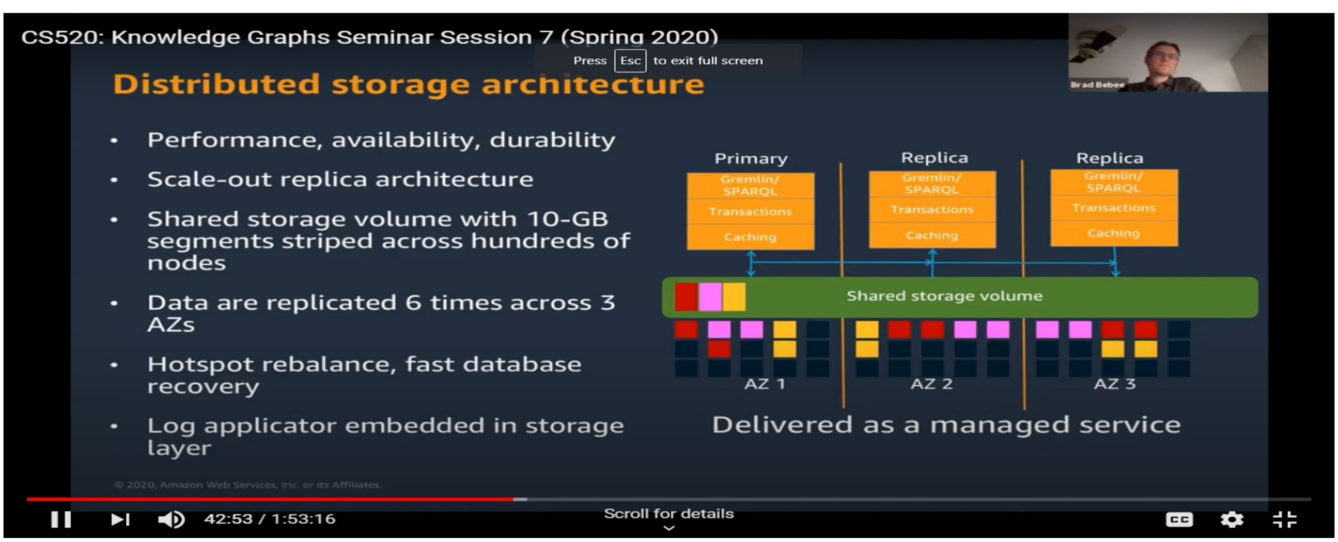
It's just a graph so just make it so.

Background included 2014 Blazegraph which was picked by Wikipedia like other customers like linking things and for applications. You can make new relations easily.

Graphs make sense many places.

The Q becomes what type of system should I use to manage the data. SQL had problems expressing some things and for joins. RDBs aren't optimized for graphs. You often have to make big changes to add relations.



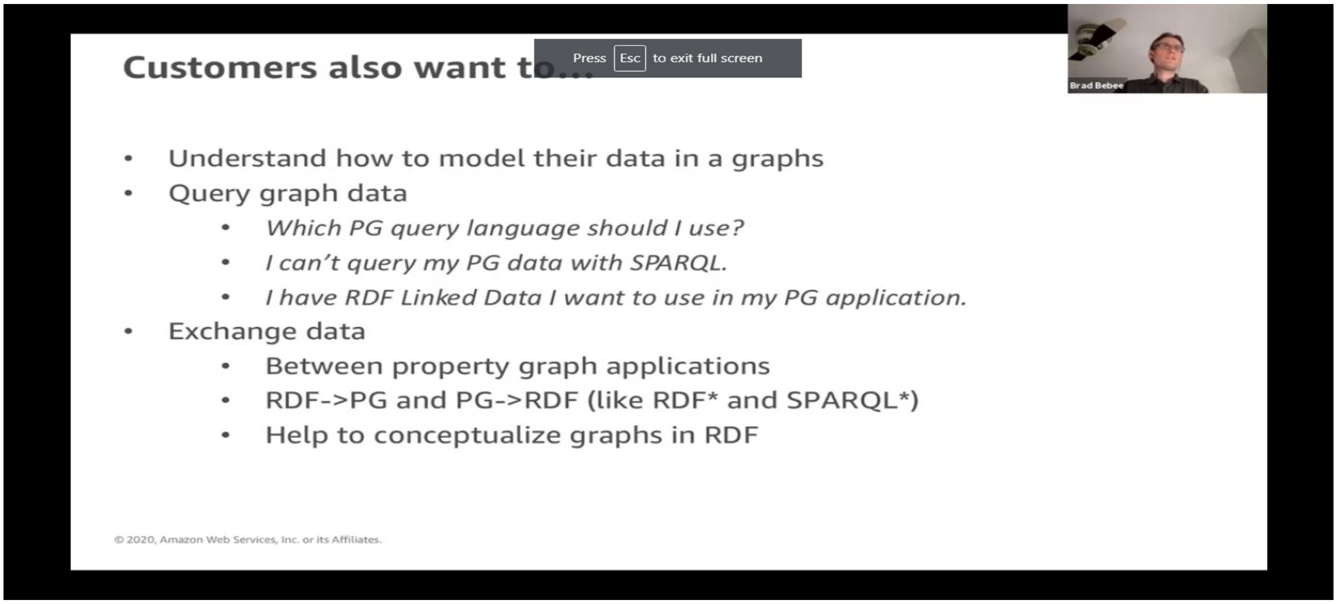


But users are not typically thinking of different graph models like RDF or property graphs.

They want us to make this easier and not have to worry about the model (may affect tools?)

So they provide both with their architecture- they use a quad format (S-P-O, graph) to store data, They use several indices such as S-P-O, graph or S-P-graph-O to make retrieval fast.

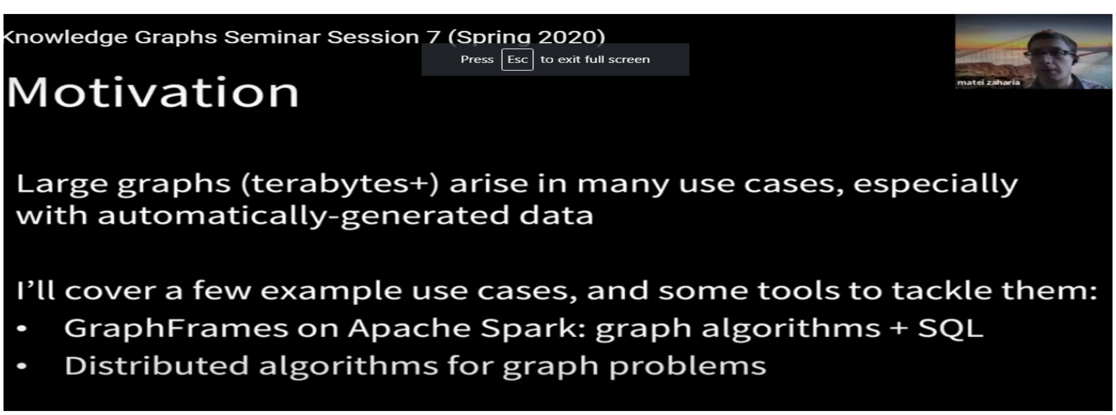
What customers want:



Question – what is the different PG or RDA and how to chose?

A. Not for customers but scale issues with billions of nodes. If building a graph standardization efforts will move on and build support for one language.

Matei Zaharia (open source databricks -platform on top of clouds) Large scale analytics with Apache Spark.

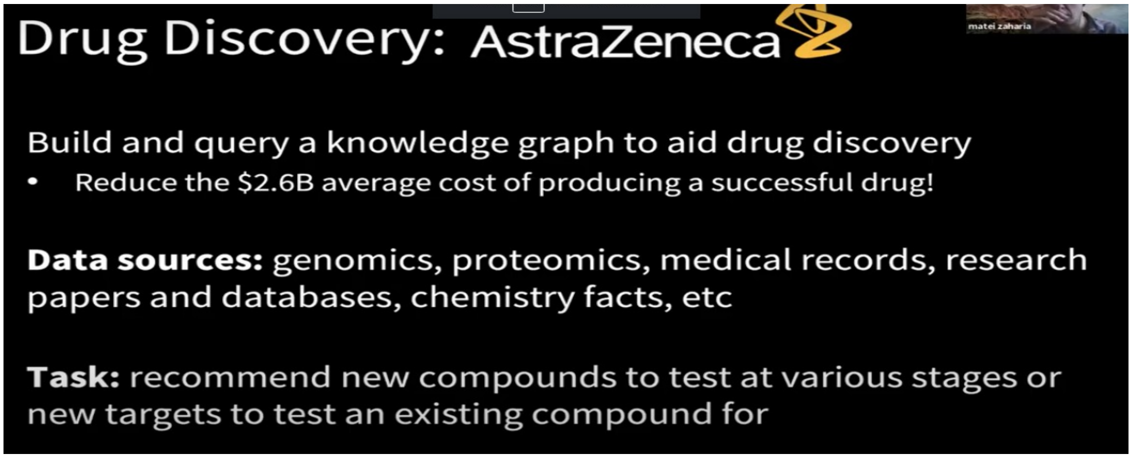


Cool applications with analytic automated generated data.

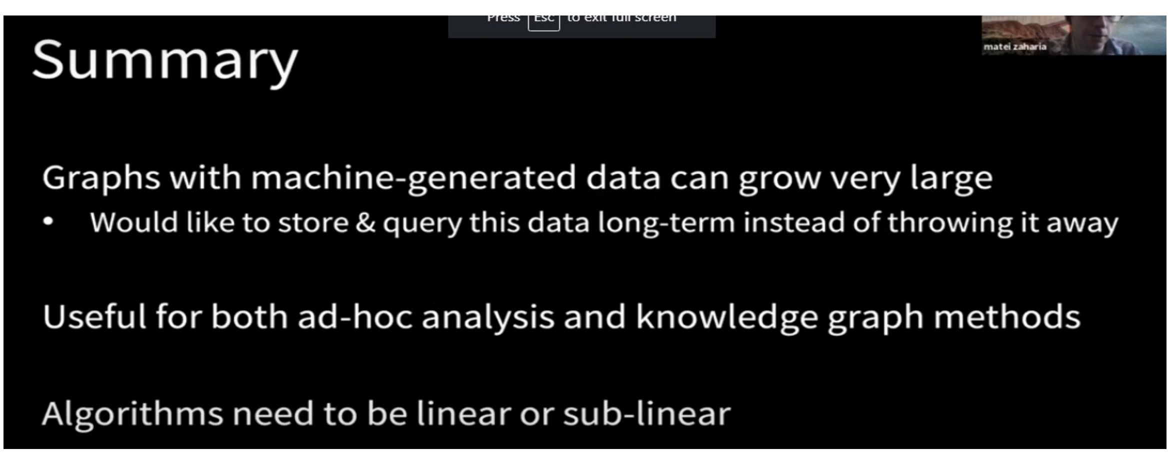
Use cases: FINRA regulatory organization to detect illegal automated trades.

Large volume, saved data and look for patterns.

Drug discovery is another example:



Use case summary:



Tools & APIs – used GraphX now using GraphFrames package now. It implements graph computations on top of a SQL engine.

Table Fields provide into about a vertex of edge/relation (like a PG) Allows you to do SQL query.

Motif finding also allows Nei4J pattern matching using search for structures based on something similar to the Cypher language. Use together with filtering in airport flight example,

Questions – what is the role of ML in these use cases?

A. Often part of the workflow as needed by the client. Often this is part of the pattern match. A graph output can be sent to a ML application.

Neptune may use the entire KG to train a ML application.

Q of suitability of Dbs vs Kgs. Some differences on scaling maintenance and use of schema free approaches. Varies by application too differences in handling time-series where RDBs are good.

Q. Do you worry at the ontological level?

A Neptune sees a need for ontologies. They support triples from ontology for truth maintenance. It is a bit like hot sauce. A little is good but you can have too much that unbalances things.

Q Q what rules operate?

Neo4J stores an RDF version sedately. Often the rules are hard to maintain.

Q. what types of visualization is used to represent a KG.

A. It is used to understand the KG. With no schema this (neo4j Browser) is helpful.

Visualizing certain workflows is helpful. Customers want to see the KG in the context of their analytic tool.

Q. What are the next set of features they are working on?

Neo4J just released a suite has reactive programming and security and works with Bloom.

Neptune wants to bring in AI and making it easier to use.

Apache SPARK wants low cost data versioning and better checking of data quality. Putting lots of effort into performance.

What is the role of knowledge graphs in machine learning?

Jure Leskovec Reasoning in KG using Embeddings

knowledge graphs are heterogeneous graphs where facts are represented as triples (h,.r.t) but are often imcomplete.

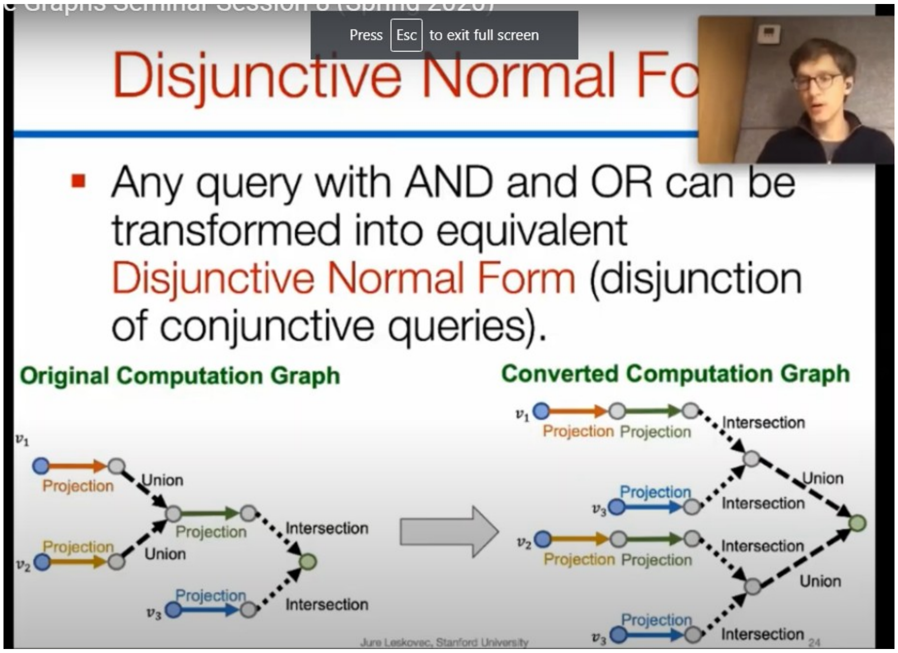
Can we use ML to complete them,say with linked predictions (but that takes time since matching dpesn't scale) or multi-hop, logical queries?

But this is hard because a KG is large, noisey and incomplete, often with uncertainties and people want fast query times.

Solution to map the KG to an embedding space. Fore queries we translate to a spacial concept and do spatial search, spatial operators like overlap for a space close the query space.

They represent the query as a box and look for intersections using a projection operation.

Challenge handling arbitrary and/or queries- so it scales.

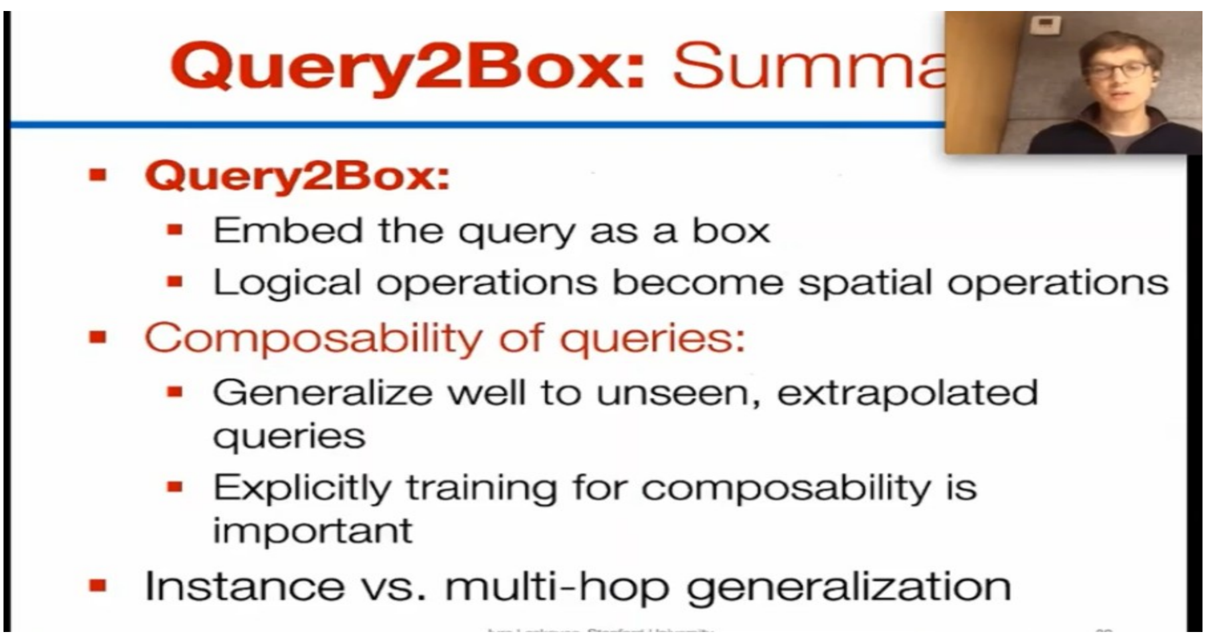


Training approach positive with answers in the KB and negative with random items.

Does the method generalize to hand new queries on missing info?

Train with at least one missing edge (incomplete knowledge) relation.

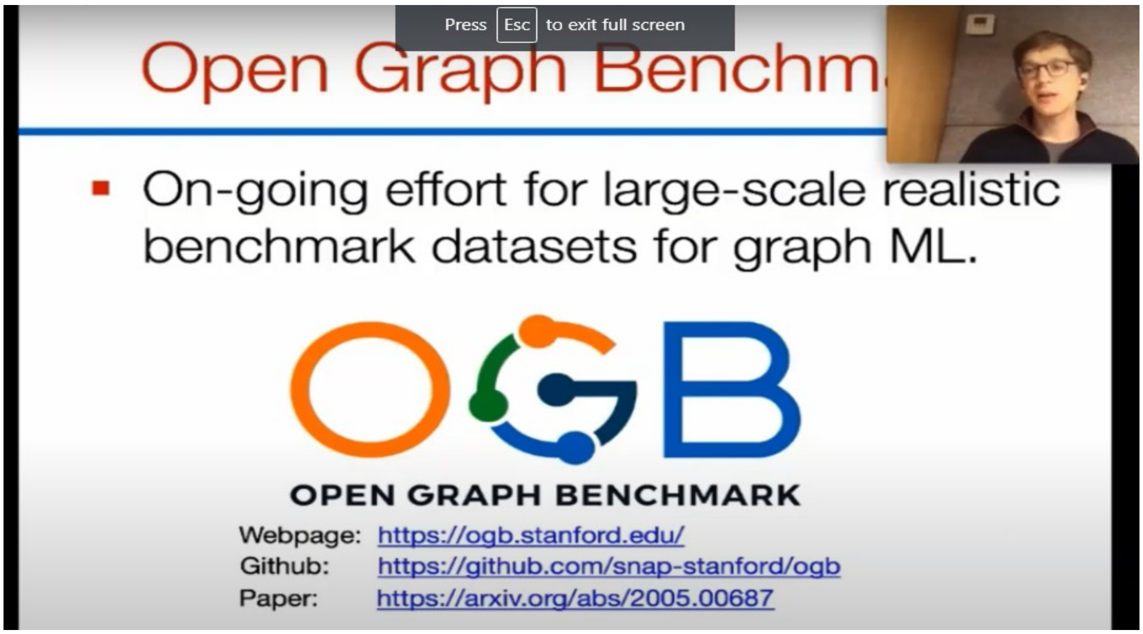
Test on Freebase



Future work is aimed at handling negation and uncertainty.

Also interested in developing a open graph benchmark DB to test ML with ML.

Interest is 3 sizes in science domain.



Luna Dong (Amazon Product graph team) Is there a natural synergy of ML and KG?

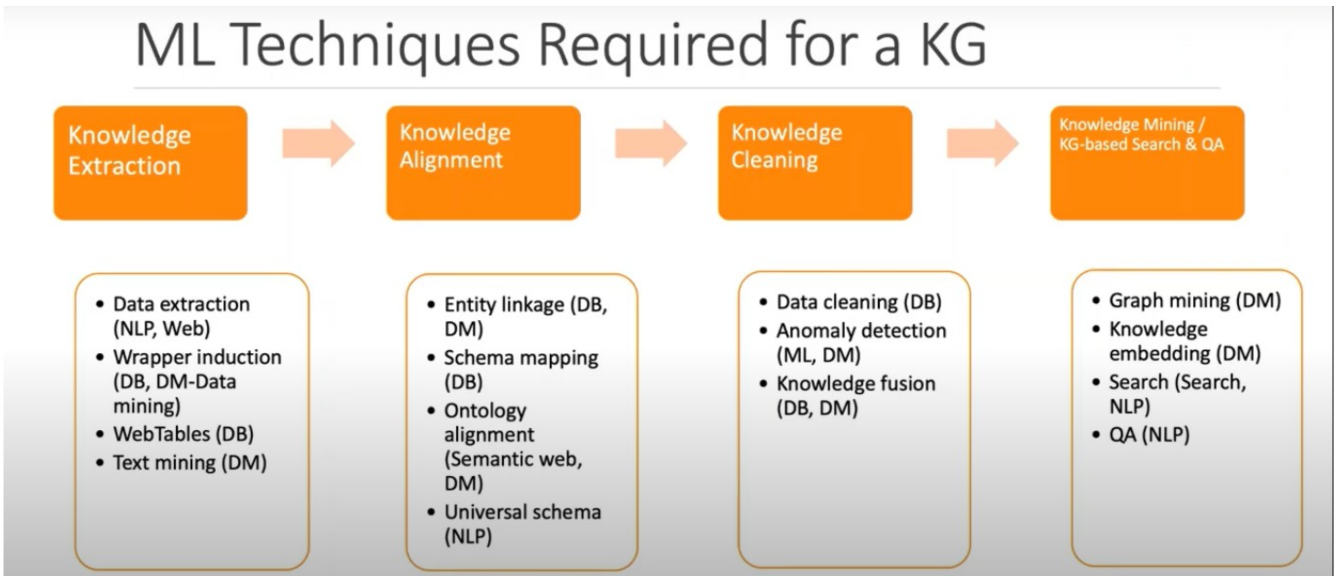
What is the Role of KG in ML

KG is a gold mine for ML but also a testbed

The more knowledgeable the more intelligent seems obvious and visa versa.

So knowledgeable returns better search results, can be an assistant, makes better suggestions,

So need to absorb K from sources (extraction), align them, clean up and mine them



10 challenges

1. Vast Unknown Space Size means you may not know all the types and relations beforehand, they change and how to organize them.

They use mining to help organize the taxonomy

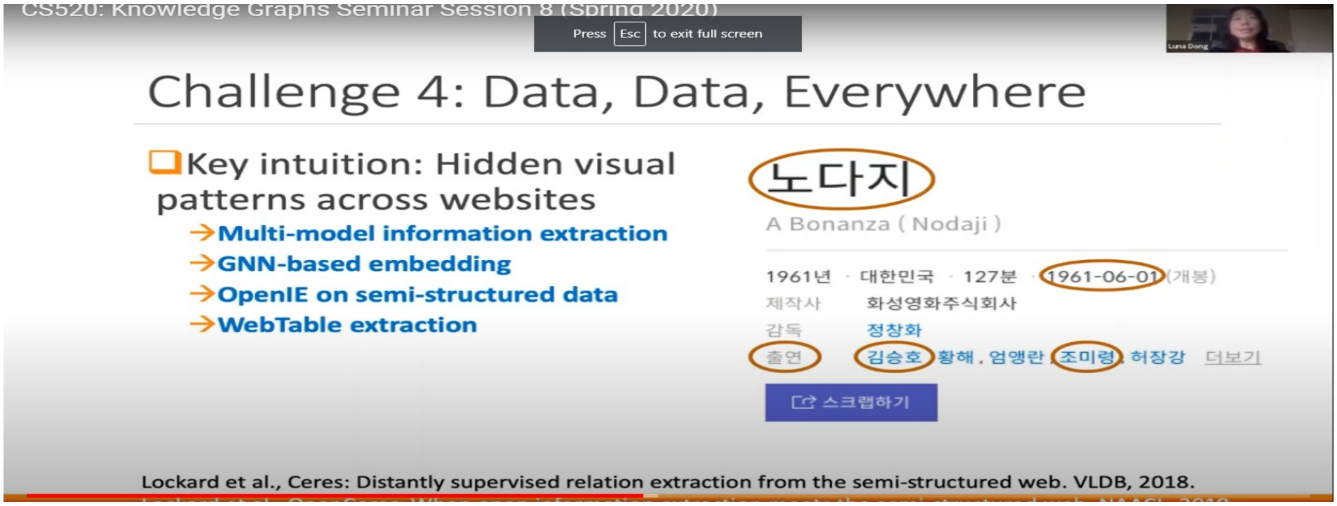
2. Data is not well structured ,some is in titles.

Important info also in images. Do NER or data extraction.

3. Different ways that entities are identified heterogeneity. Need to align the data by exploring the neighborhood. Might do knowledge embedding and mapping,

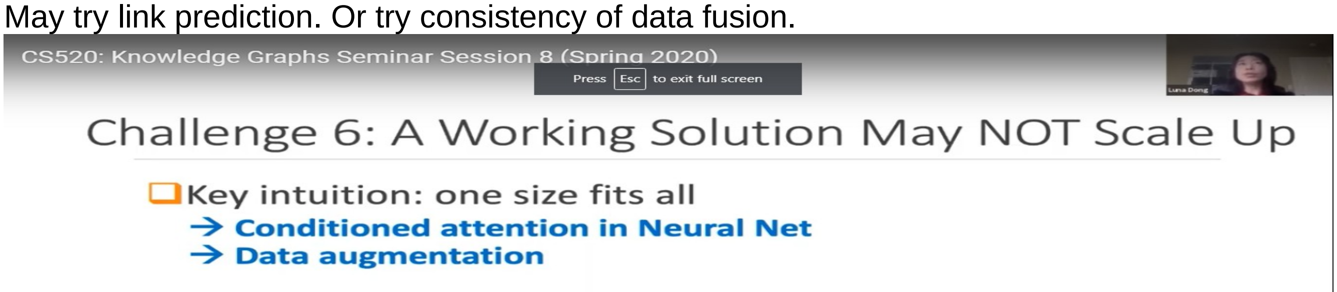
4. So much data with different layouts– not just text,some semi-structured from Dbs.

How to extraction? Often hidden patterns such as headings, where relation is,attributes and dates,



5. What data is trustworthy? For cleanup look at reduncency, in same embedding space?

May try link prediction. Or try consistency of data fusion.



6. What if it doesn’t scale, because each category uses a different model with different vocabularies.

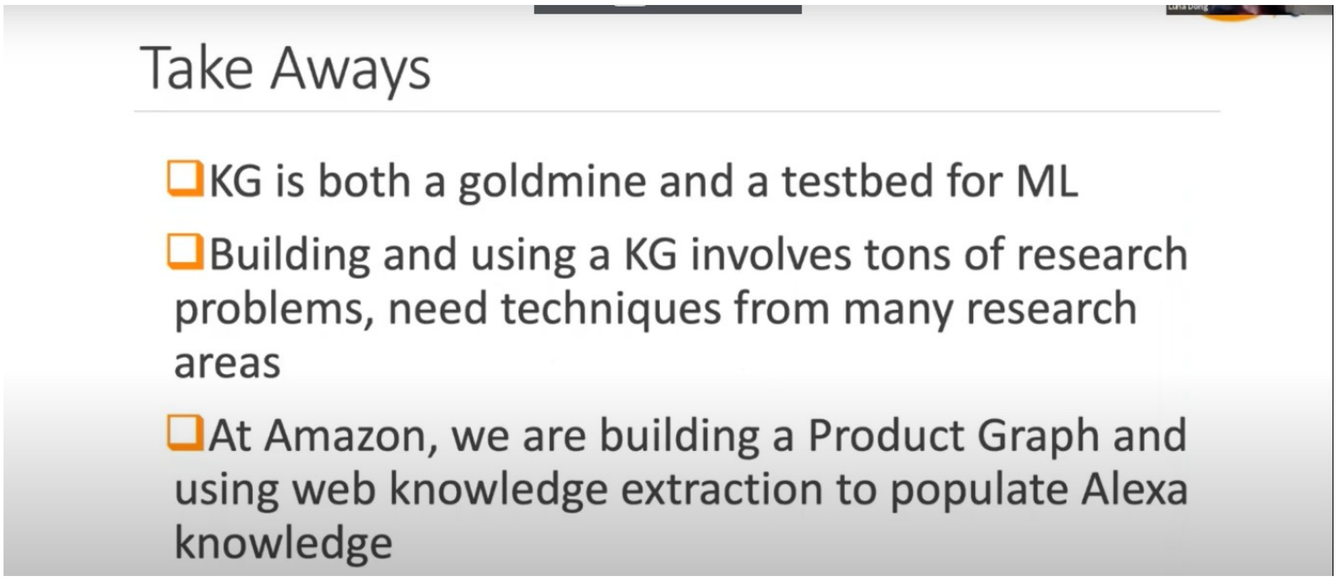
7. training data for so many cases. A variety of learning approaches on labels

8. How to use a KG – find the fact in a KG? Entities may not be searched by name.

Can you find related items or do query refinement or involve user feedback.

9, Gaining insight from the graph facts. Do random walks

10. can you explain it? Interpretability. Still working on this



Robert Hoffman (IHMC) challenges for XAI

Emphasis on Psychological challenges of XAI.

State of the art??? Really challenges because we borrow terms from ordinary language or science to apply elsewhere - these may be orthogonal. It misleads many. - cross disciplinary confusion. (neural nets are not like neurons or brains

Interpretations get used as explanations.

There is little deep investigations into explanations.

It requires more than a good interface.

The traps of Goofball science.

Explanation is a process not just a matter of diagrams or a statement.

The user needs to explore the solution and explanatory space. Why does it get things right or wrong?

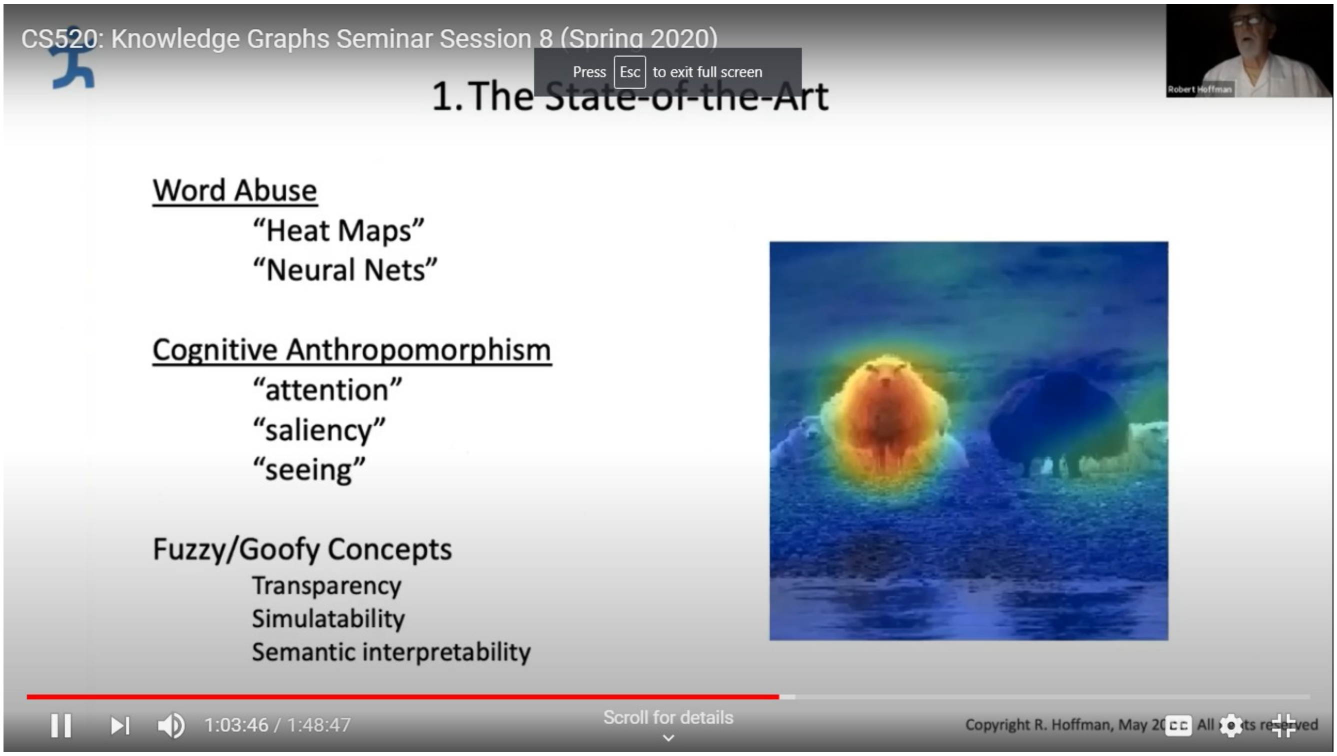
AI systems make errors that humans would never make..due to CSK. They often don ‘t

 have human concepts. And users often have different views than a designer.

Example of foot recognition and old wine in new bottles. KG is an example along with concept maps and intelligent tutoring systems. Need dialog more than just text.

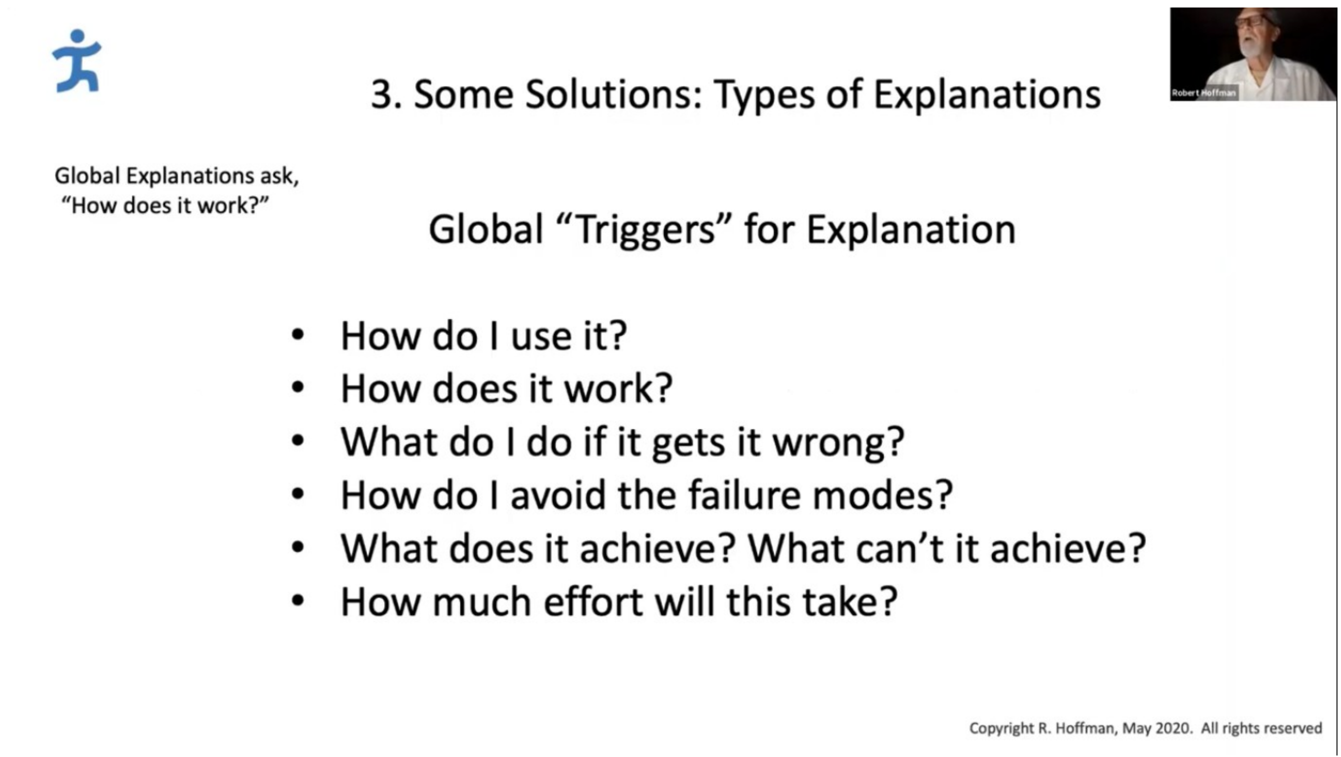
Where is the knowledge in a bubble diagram?

Users have views:



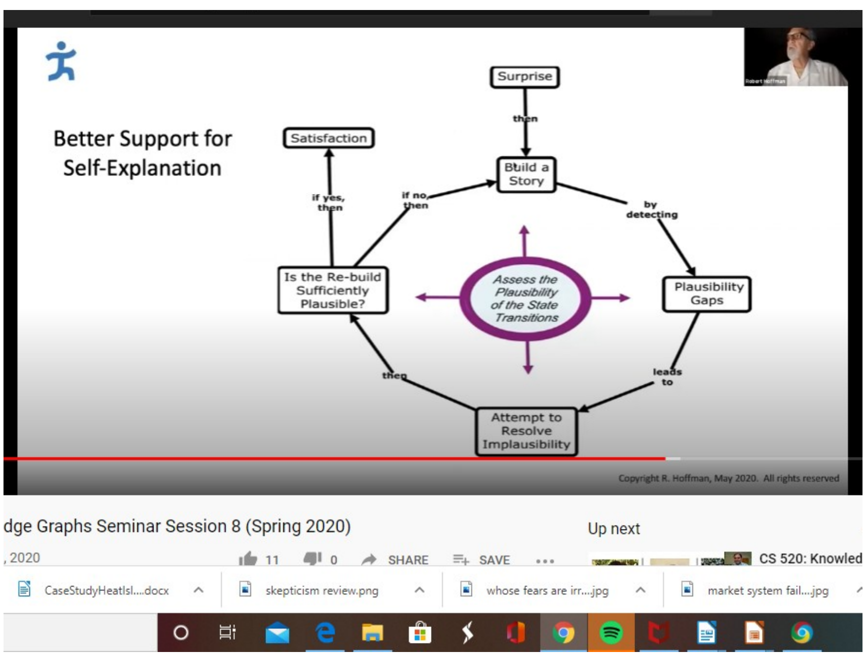
Triggers.

How does it work, what can it achieve?

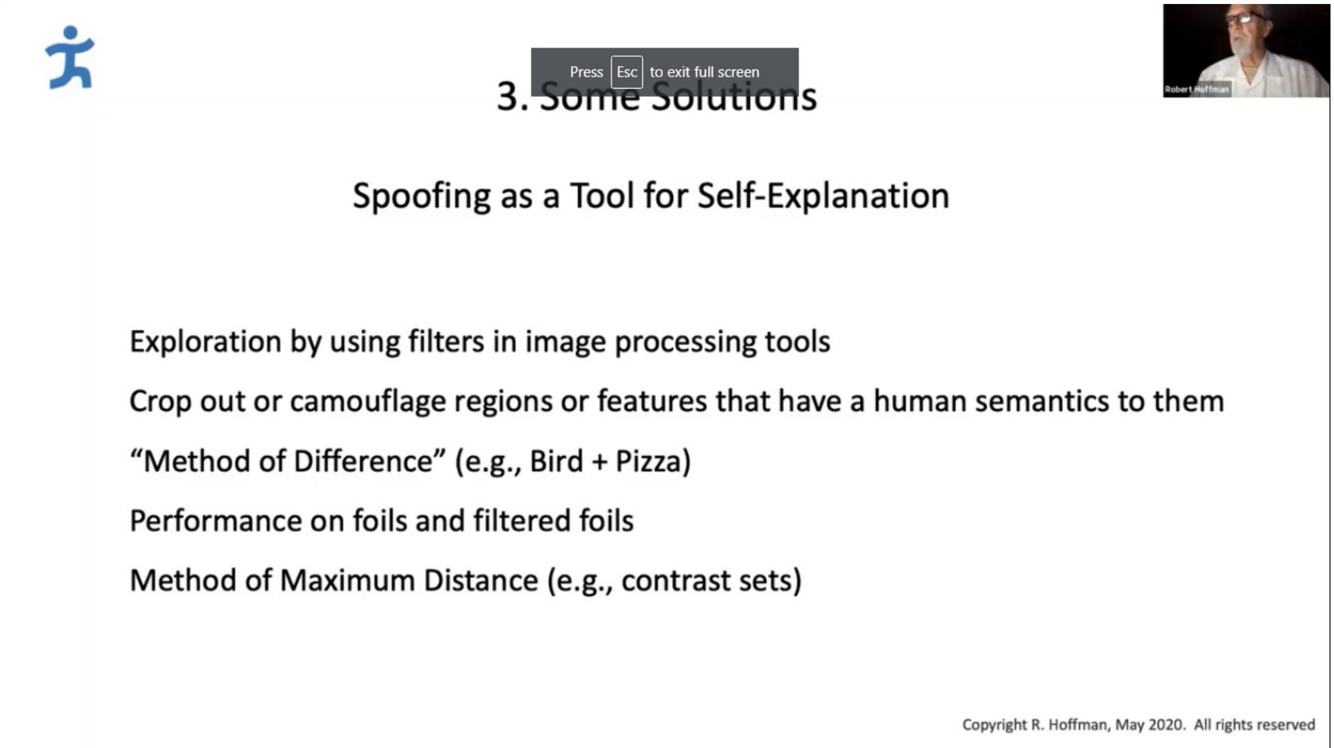


Are there solutions? Better support for self explanation using what ML provides as a mere clue or e idea of stories and plausibility for complex system. What an ML knows may be a metaphor-like thing.

List of solutions that really are just help -graphics etc.



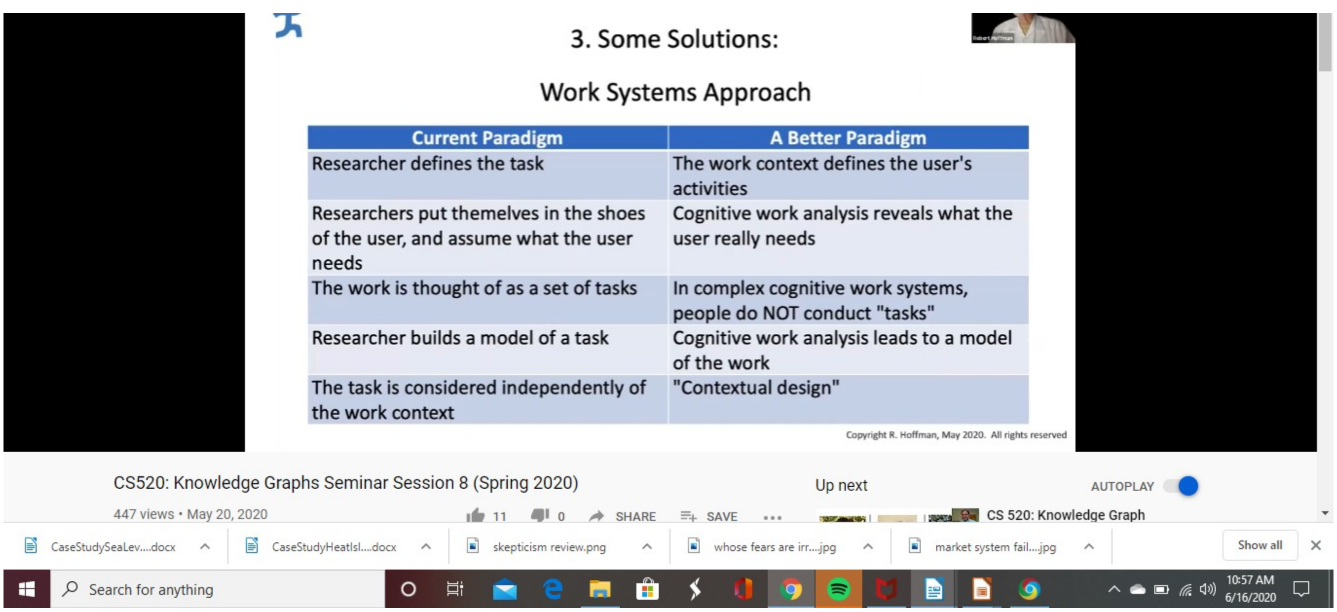
Do we just need more data? Spoofing needed too.



He likes shifting the paradigm as a solution

Work context defines the task. (Cognitive work analysis)

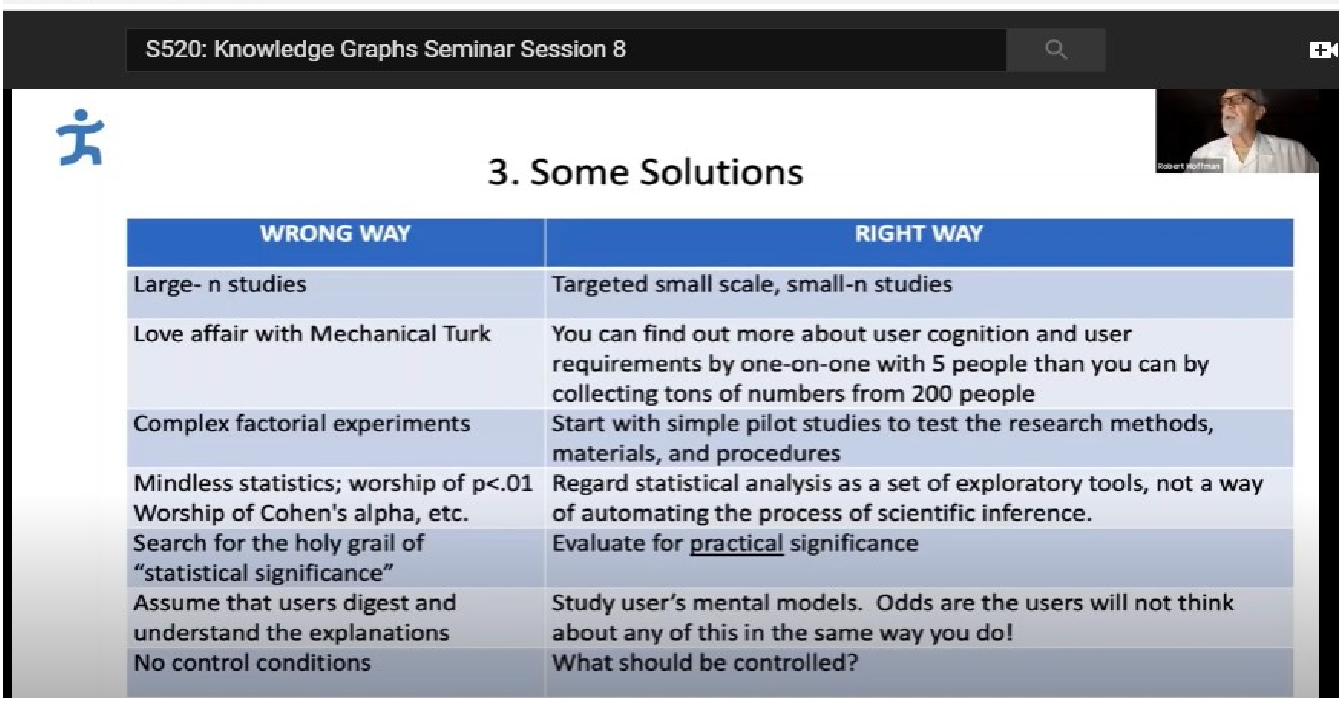
An we need to involve the end user earlier in the process:



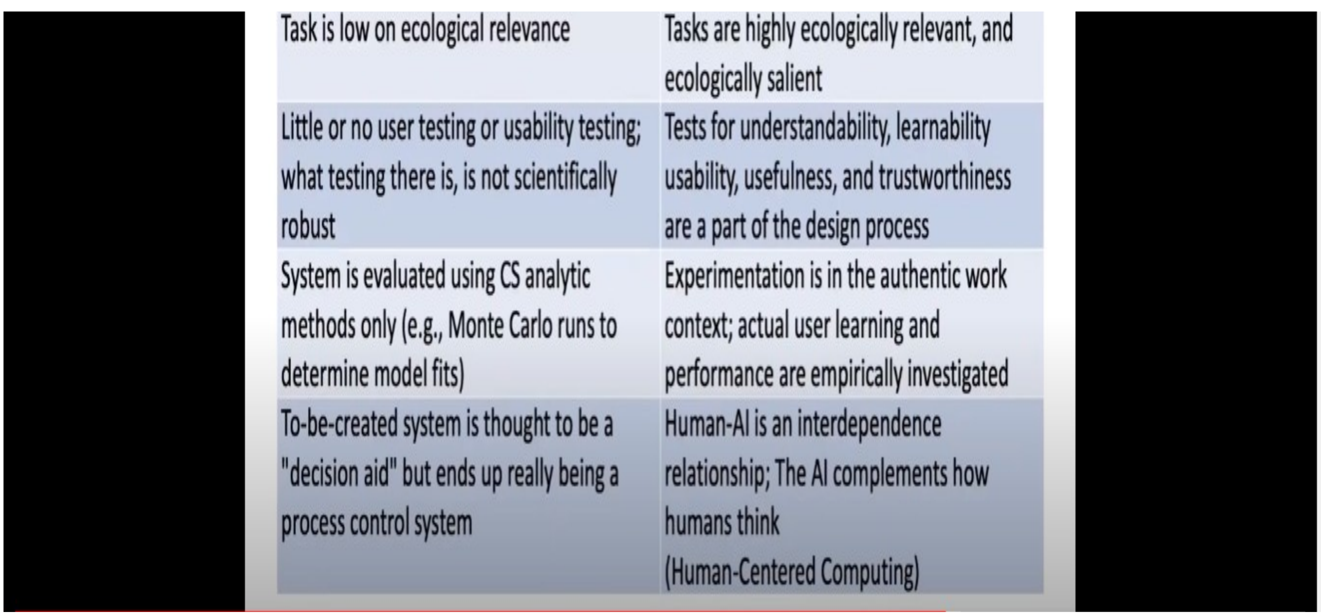
Need a more integrated approach and not a process control but one of interdependent computing.

Need more integration with experimental psychology.

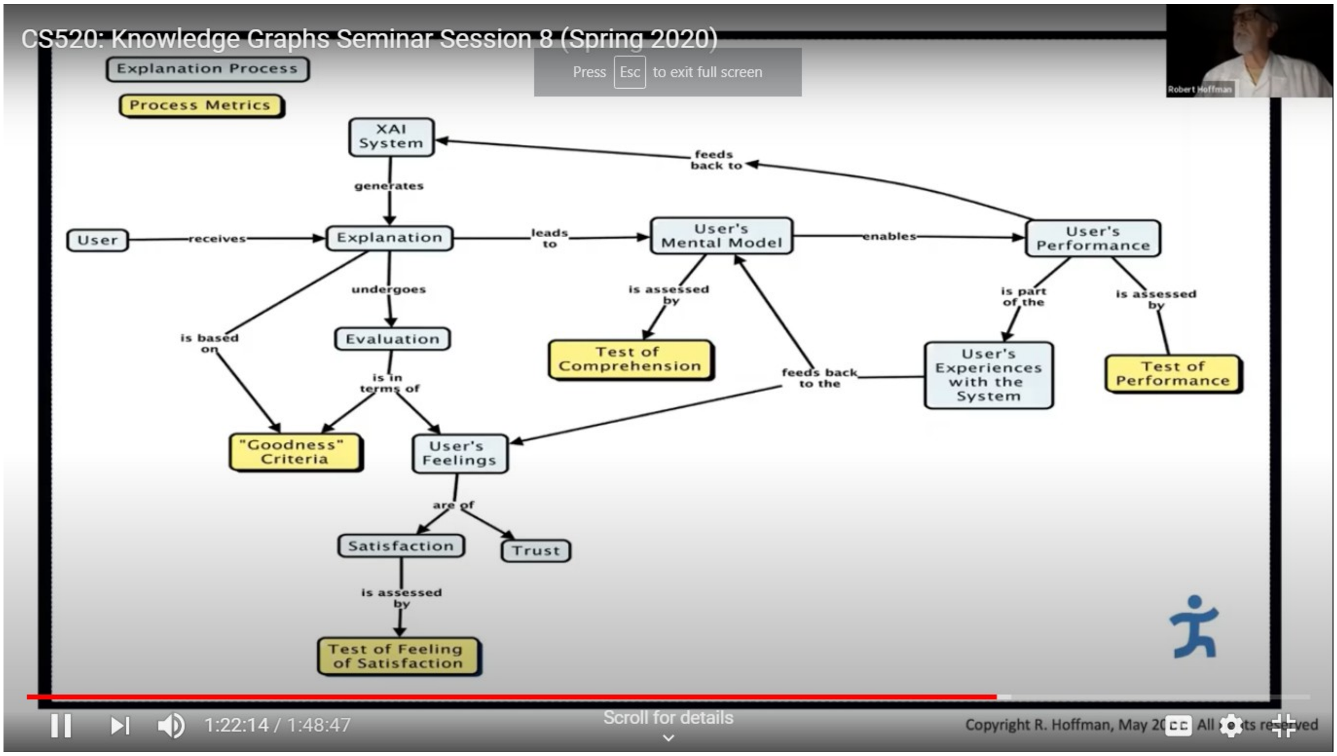
The right vs wrong way to get Psych involved:



WRONG WAY	RIGHT WAY
Large- n studies	Targeted small scale, small-n studies
Love affair with Mechanical Turk	You can find out more about user cognition and user requirements by one-on-one with 5 people than you can by collecting tons of numbers from 200 people
Complex factorial experiments	Start with simple pilot studies to test the research methods, materials, and procedures
Mindless statistics; worship of p<.01 Worship of Cohen's alpha, etc.	Regard statistical analysis as a set of exploratory tools, not a way of automating the process of scientific inference.
Search for the holy grail of "statistical significance"	Evaluate for practical significance
Assume that users digest and understand the explanations	Study user's mental models. Odds are the users will not think about any of this in the same way you do!
No control conditions	What should be controlled?



Task is low on ecological relevance	Tasks are highly ecologically relevant, and ecologically salient
Little or no user testing or usability testing; what testing there is, is not scientifically robust	Tests for understandability, learnability usability, usefulness, and trustworthiness are a part of the design process
System is evaluated using CS analytic methods only (e.g., Monte Carlo runs to determine model fits)	Experimentation is in the authentic work context; actual user learning and performance are empirically investigated
To-be-created system is thought to be a "decision aid" but ends up really being a process control system	Human-AI is an interdependence relationship; The AI complements how humans think (Human-Centered Computing)

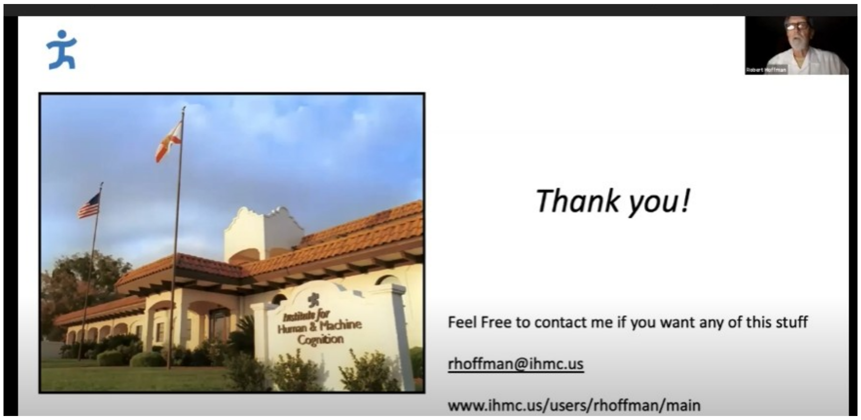


DARPA XAI approach.

Questions

We are trying to build a conceptual map, but can't do it an edge at the time…

A. What we are talking about is more like John Sowa's concept diagrams/maps.



Q for Luna – trying to extract from things loosely designed. Can you ask for more cooperation and get more structure?

A, Yes, but we would get more abuse to make products stand out. They may provide small misinformation. We do want to get sellers and buyers engaged to describe better products.

Q How do you decide what is a new product?

A. We have entity linkage system and use matching even images. We ask sellers to look first at product types.(self driving curation too._

Q. What is the role of explanation in Amazon system?
A. It can improve recommendations, relevant but perhaps different.

Q What do you think about realist ontologies like BFO? Do the high-level distinctions play a role
A. In Amazon they thought it an easy part as a backbone, but much more semantics to it.

Q	Jure- What do you do when an answer doesn't make sense to the user?
A. The model isn't perfect. It is statistical.
Q for Robert, as a Psychologist, what problem can we help solve...we have been doing this for a while.
A. It depends on what the project defines as a problem. How do you measure making explanations better? It might be help with user terminology which is often mentalistic.

Q. Will your KG ever be done?
A. Some think it a 100 year project...it is hard, but is why they are working on a self driving approach to add new types by itself...includes a human in the loop.
The DARPA rapic knowledge project ran into problems with self guided approaches. Cyc may also be considered a 100 year project that ran into hard problems.

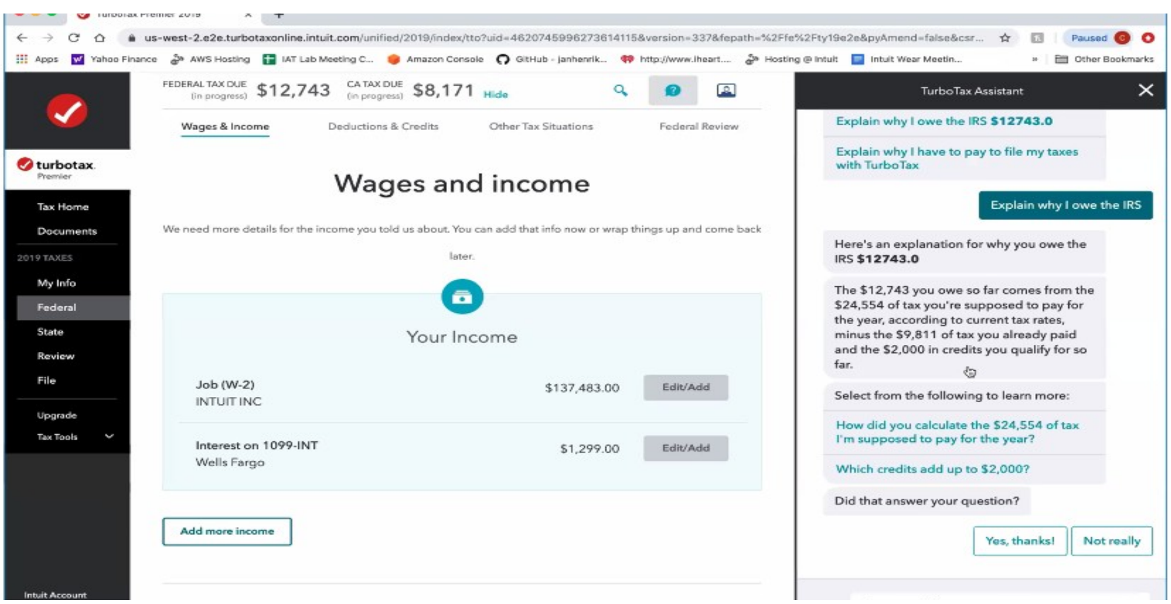
S520: Knowledge Graphs Seminar Session 9

What are some high value use cases of knowledge graphs?

<u>Jay Yu</u> Intuit KG Uses case

Intuit has an AI driven platform to provide good financial advice and explanation.

Intuit explains why it is engaging intelligent assistant how it came up with a return using KG for reasoning

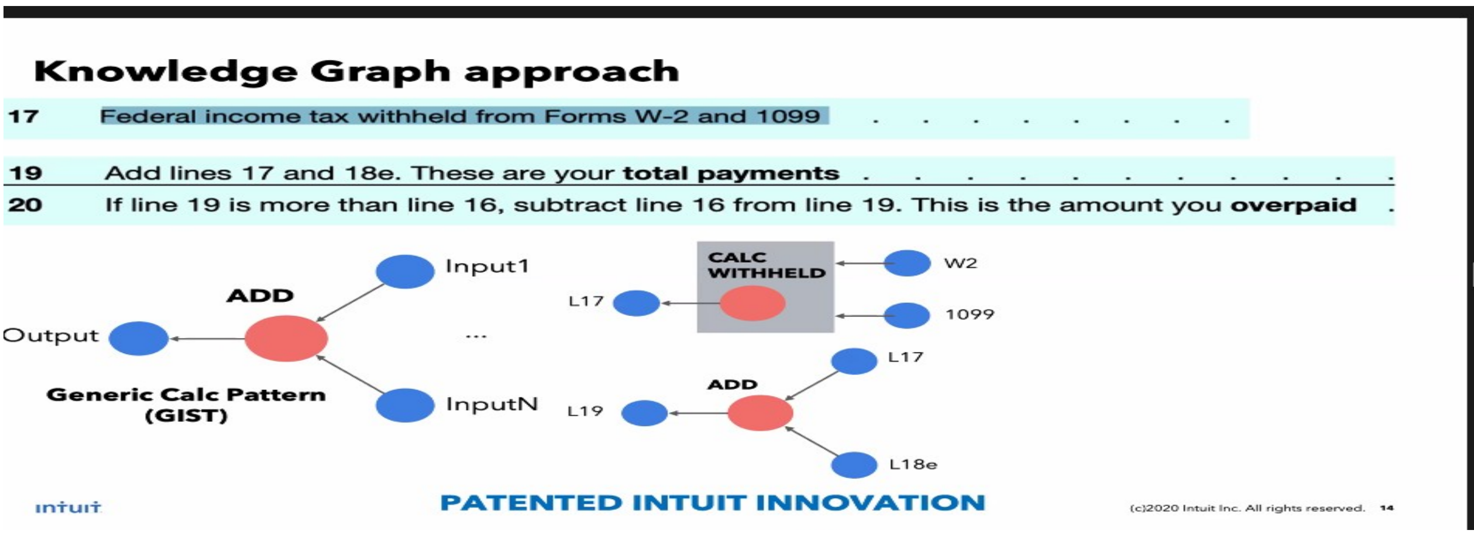


AI and KG have their own role and KG is the foundation of their data.

How they use a KG...the US tax code if difficult so there is lots to encode supporting thousands of calculations. An example is for eligibility.
Limits of procedural approach -be explicit so can provide explanations.

Be declarative and you can compose things into a larger graph.

Develop patterns for the numerous calculations. These are generic patterns for calculations that can be stitched together. An analysis of the graphs allows you to explain the calculation. You can reason over the graph and identify if information is missing.

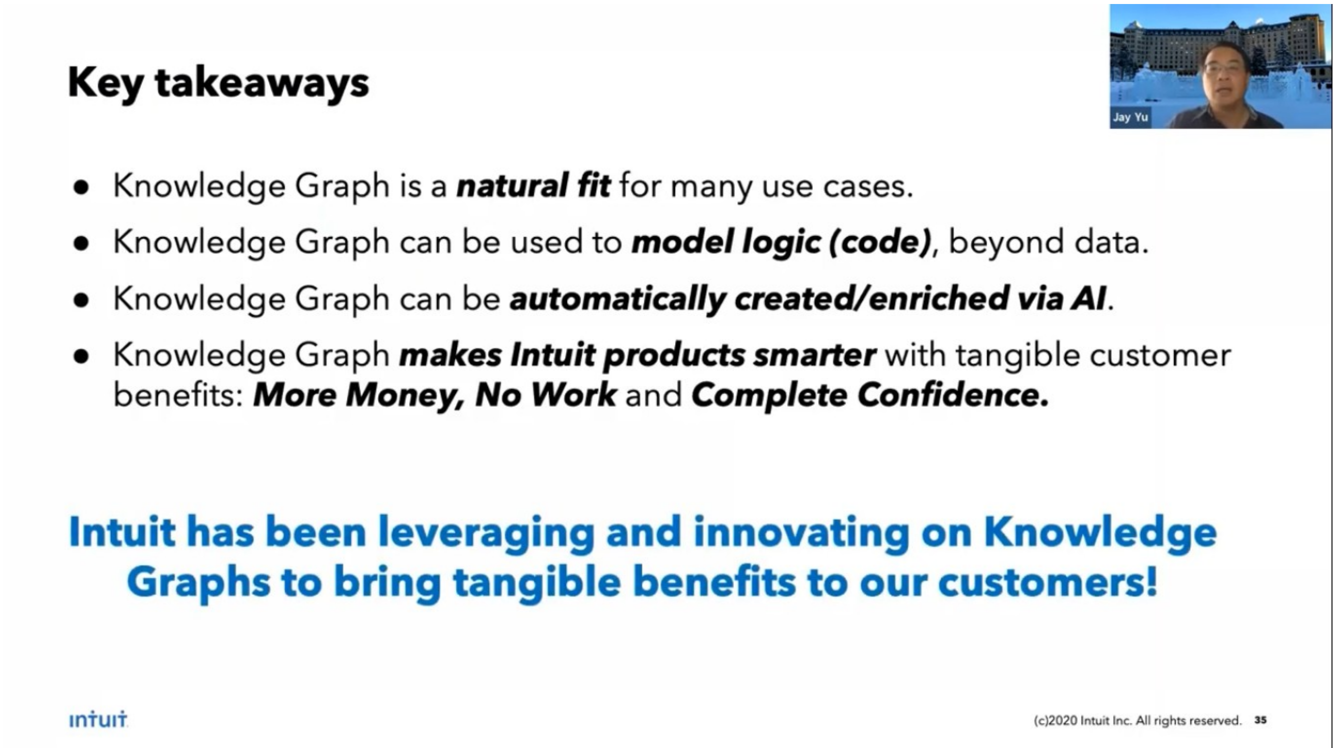


Several examples of how they build their KG and why it is better than a procedural approach.
As discussed by others they use various approaches to merge data and support entity resolution which might need a new relation.
The allow users to provide feedback.

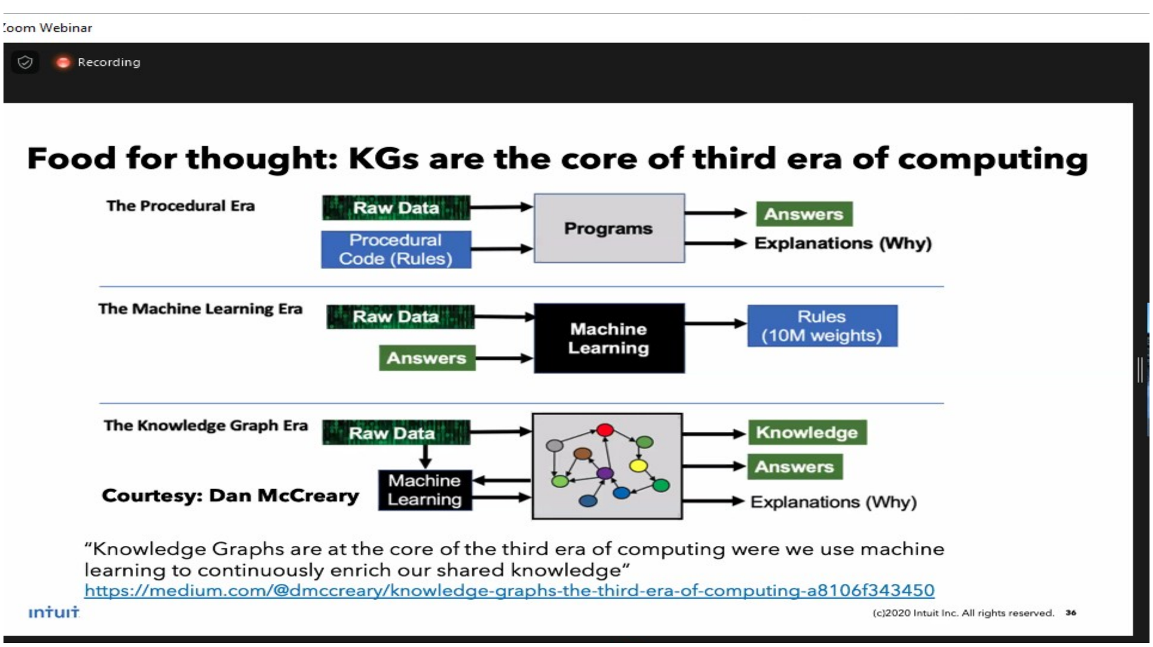
Comparison between two approaches

	PROCEDURAL	KNOWLEDGE GRAPH
Logic	Procedural	Declarative
Execution	Sequential	Dependency-driven
Modularity	By best practice	By design
Explainability	Blackbox	Built-in
Personalization	Hard-coded, limited	Dynamic, automatic
Automation	Not manageable	Manageable
Testability	Harder	Easier
Optimized for	Exceptional complex logic	Most tax logic
Target Developer	Engineers	Domain experts

intuit



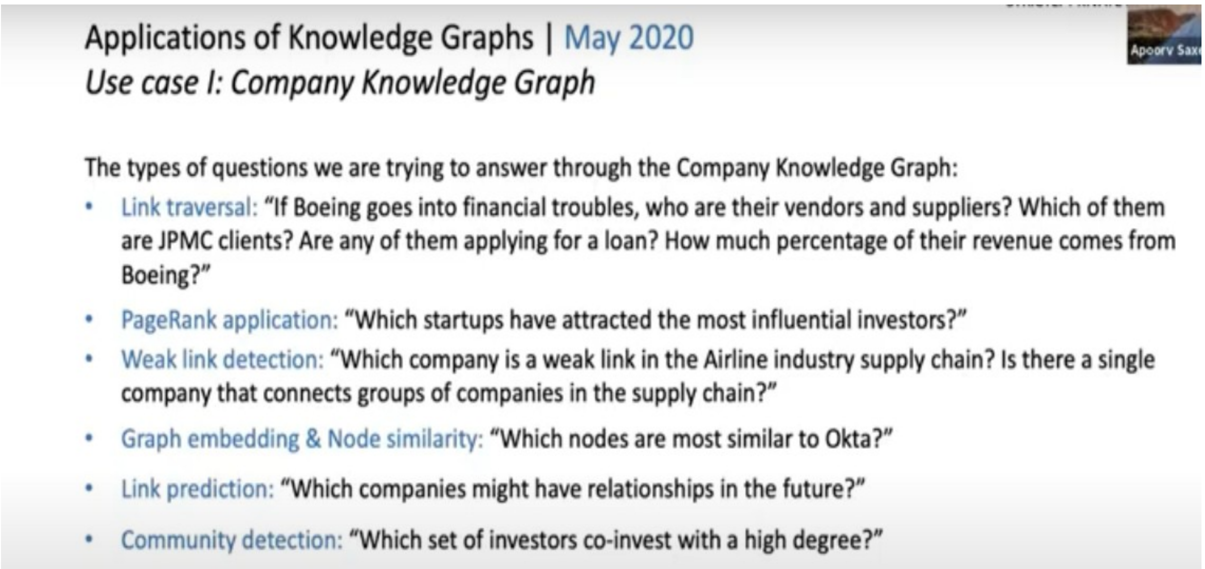
How to think about Knowledge Graphs.



Apoorv Saxena(JP Morgan) Application of Knowledge Graphs

A few use cases and challenges. Finance and Banking. KG used for fraud and risk. The KG allows them to see how fraud clusters grow over time.

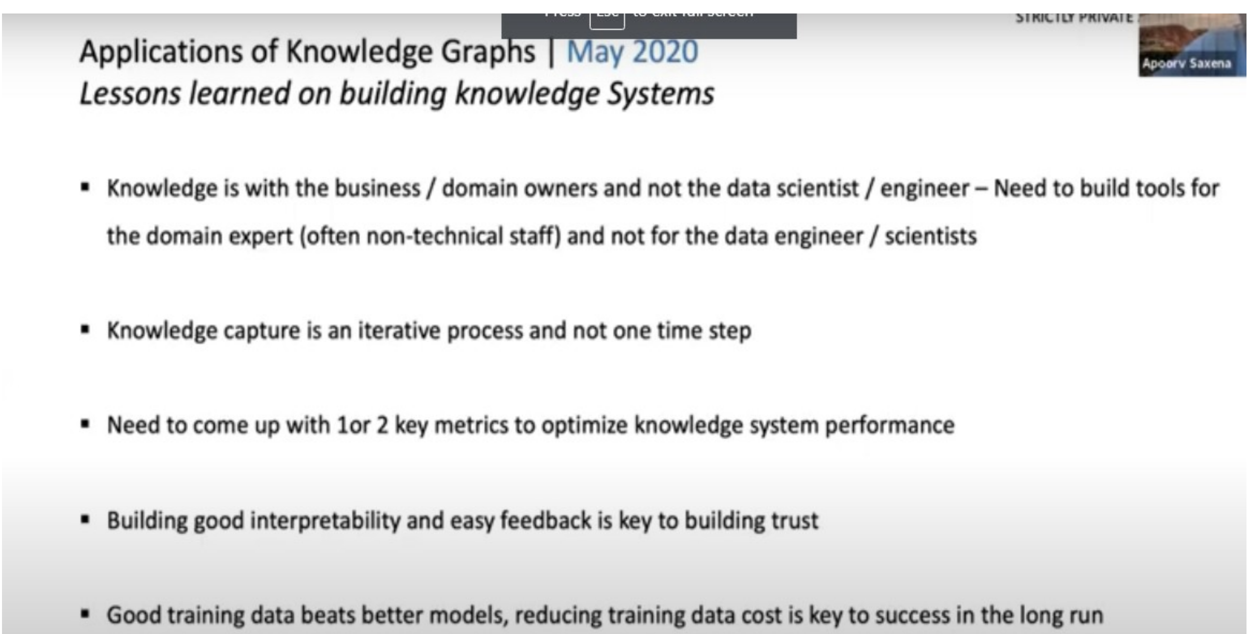
We see how a KG is used to answer questions in this Table. Includes recommendation systems.



Problems are similar to previously described – integrating various sources of data and the problem of unique names, NER and extracting information. Dynamic data caused challenges of temporal nature, authoritative sources and keeping up to date.

There are regulations on access and who can see what.

Many business users are not sophisticated so they have trouble with queries with Cipher. They use NL queries that get converted using BERT.



David Newman Wells Fargo High Value Use Cases (also FIBO)

Talking more conceptually than operationally.

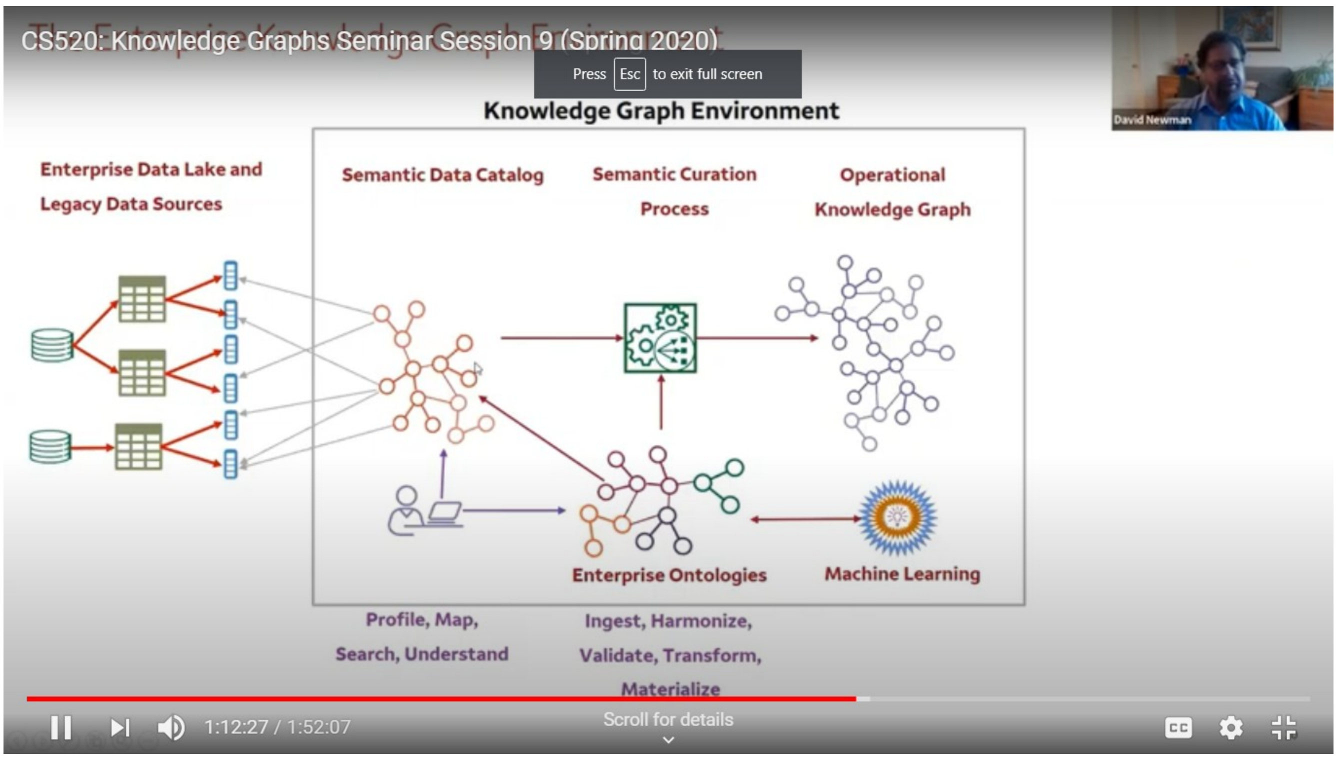
KG as a parading shift. How to improve our metadata and reduce data wrangling.

How to handle data silos? The data paradigm is old and we need to move on from the high costs and risks. KG provides a strategic solutions with a common semantic data

model understandable by humans and machines.

There is KG trajectory from data to semantic web now to KG 3.0 with machine understanding using ML.

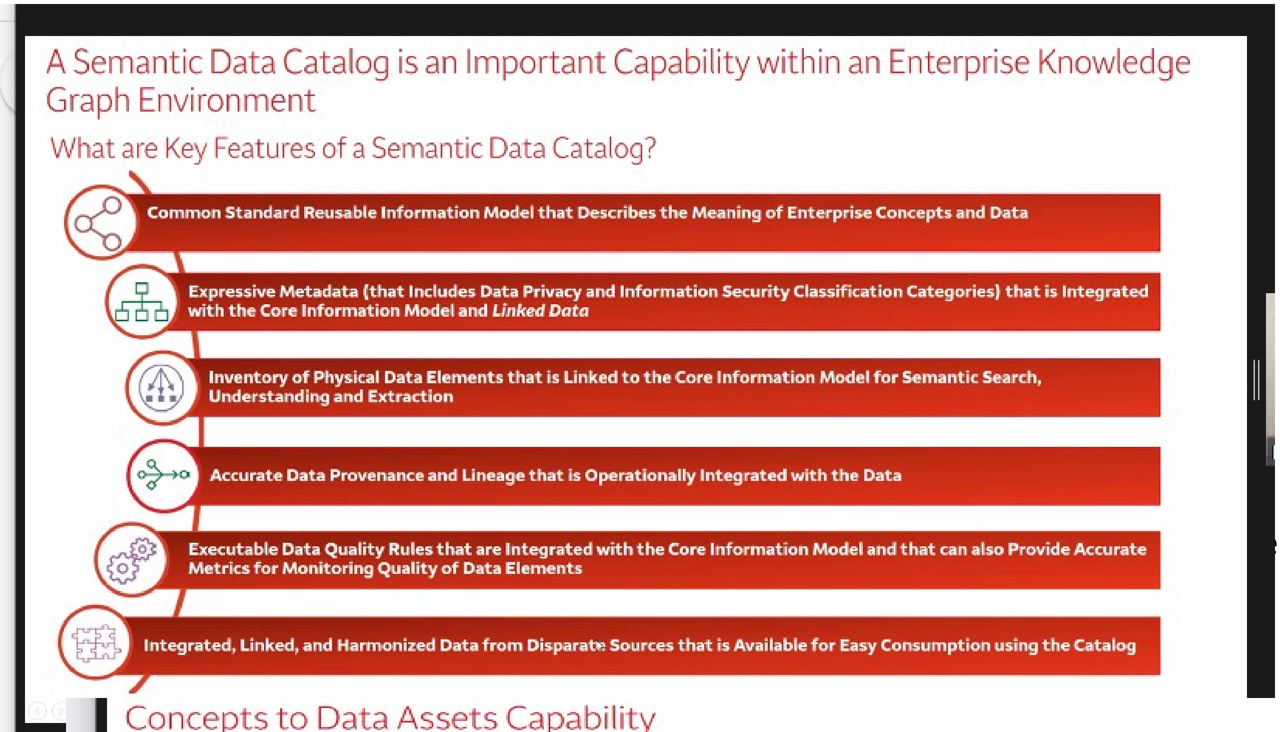
Groups like EDM provide a start on data for open, reusable enterprise KGs – also usable as a

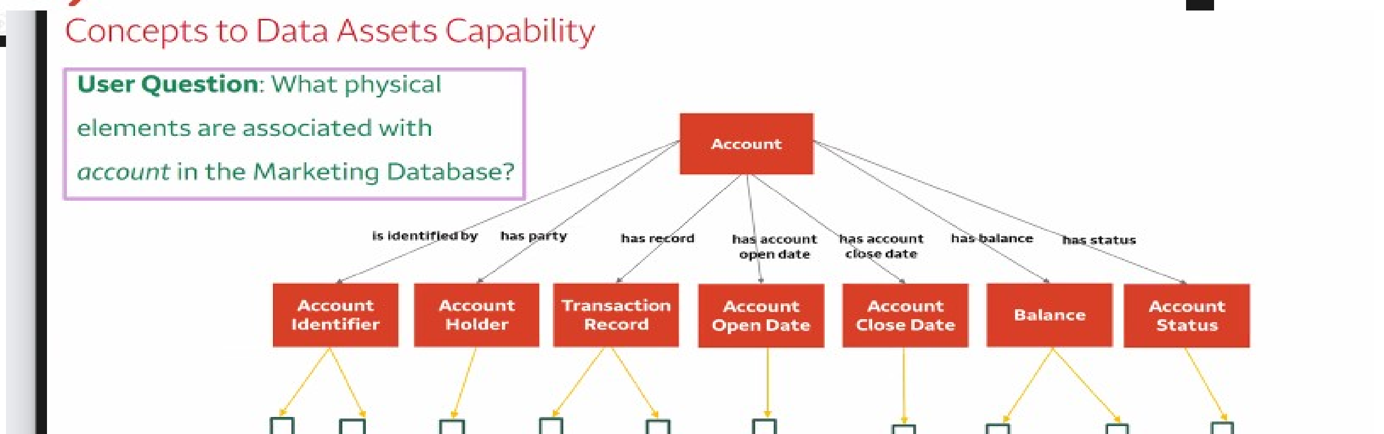


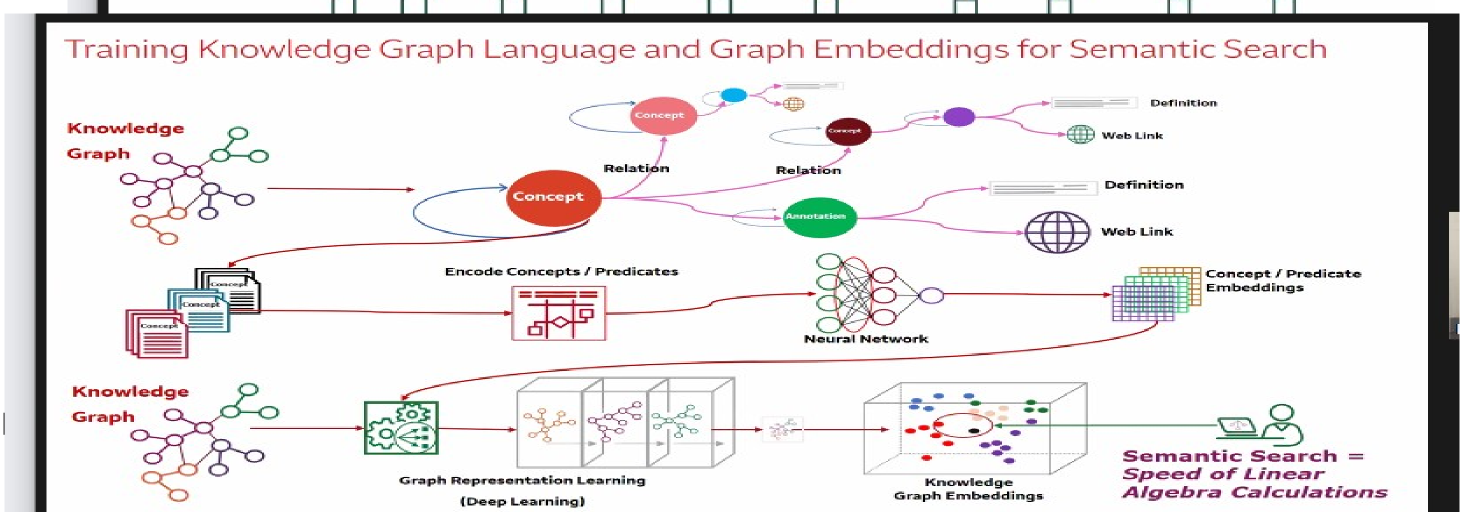
reference model since they are based on an enterprise ontology.

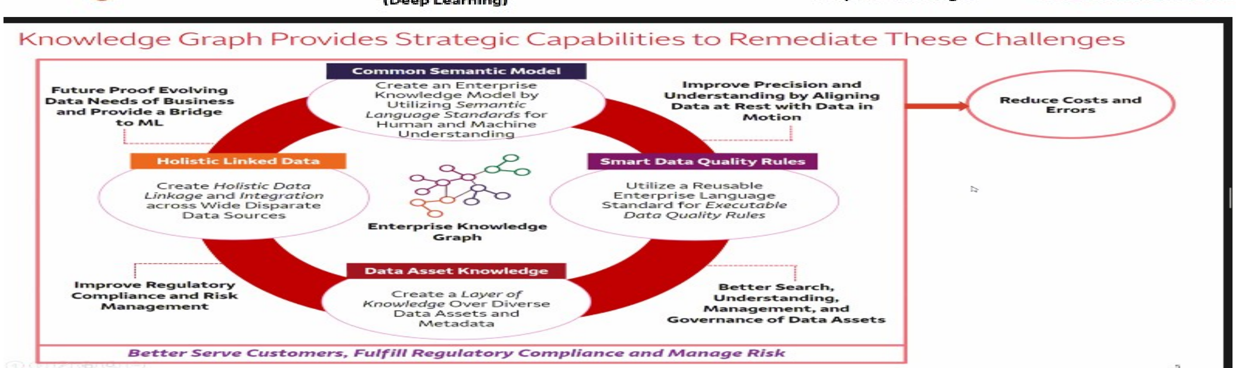
The idea of a semantic data catalog to help semantic search and queries for understandings.

There is a methodology that goes along with this using ML for link prediction, resolution, embedding ….


A Semantic Data Catalog is an Important Capability within an Enterprise Knowledge Graph Environment
What are Key Features of a Semantic Data Catalog?
Common Standard Reusable Information Model that Describes the Meaning of Enterprise Concepts and Data
Expressive Metadata (that Includes Data Privacy and Information Security Classification Categories) that is Integrated with the Core Information Model and *Linked Data*
Inventory of Physical Data Elements that is Linked to the Core Information Model for Semantic Search, Understanding and Extraction
Accurate Data Provenance and Lineage that is Operationally Integrated with the Data
Executable Data Quality Rules that are Integrated with the Core Information Model and that can also Provide Accurate Metrics for Monitoring Quality of Data Elements
Integrated, Linked, and Harmonized Data from Disparate Sources that is Available for Easy Consumption using the Catalog


Concepts to Data Assets Capability
User Question: What physical elements are associated with *account* in the Marketing Database?
Account
is identified by
has party
has record
has account open date
has account close date
has balance
has status
Account Identifier
Account Holder
Transaction Record
Account Open Date
Account Close Date
Balance
Account Status


Training Knowledge Graph Language and Graph Embeddings for Semantic Search
Knowledge Graph
Concept
Relation
Concept
Relation
Definition
Web Link
Annotation
Definition
Web Link
Encode Concepts / Predicates
Neural Network
Concept / Predicate Embeddings
Knowledge Graph
Graph Representation Learning
(Deep Learning)
Knowledge Graph Embeddings
Semantic Search = Speed of Linear Algebra Calculations
21


Knowledge Graph Provides Strategic Capabilities to Remediate These Challenges
Future Proof Evolving Data Needs of Business and Provide a Bridge to ML
Common Semantic Model
Create an Enterprise Knowledge Model by Utilizing *Semantic Language Standards* for Human and Machine Understanding
Improve Precision and Understanding by Aligning Data at Rest with Data in Motion
Reduce Costs and Errors
Holistic Linked Data
Create *Holistic Data Linkage and Integration* across Wide Disparate Data Sources
Smart Data Quality Rules
Utilize a Reusable Enterprise Language Standard for *Executable Data Quality Rules*
Enterprise Knowledge Graph
Improve Regulatory Compliance and Risk Management
Data Asset Knowledge
Create a *Layer of Knowledge* Over Diverse Data Assets and Metadata
Better Search, Understanding, Management, and Governance of Data Assets
Better Serve Customers, Fulfill Regulatory Compliance and Manage Risk

Questions:

How are you three groups sharing data?

A. it is the common data model such as the EDM council is discussing with others.

There are standards for sharing some of this. Some governments provide the standard such as what “money” means. KG will be the next step. We need education for users to understand a KG more.

Embeddings can also be used to guide this sharing.

Q On the idea of a computation graph. The level is at the same as the computation itself. The explanation of a computation can be so complex as to not be understandable.

A. Right. We use simple examples. The IRS publishes regulations for things in XML schema, but eventually the IRS will have to define this by say a graph.

Q The future?

A. FIBO will be crowdsourced supported by a community process.

A Jay, the future of KB is good together with ML. We need to make these open and understandable.

Apoorv Kbs will provide a base for ML.

CS 520: Knowledge Graphs Seminar Session 10

What are some open research questions on knowledge graphs (discussed during chat at June 10th session

Richard Socher Multi-Hop Knowledge Graph Reasoning with Deep Reinforcement Learning (1 part of the future)

Slides

Real world KG are important. But in the long term, how much can they capture? They are

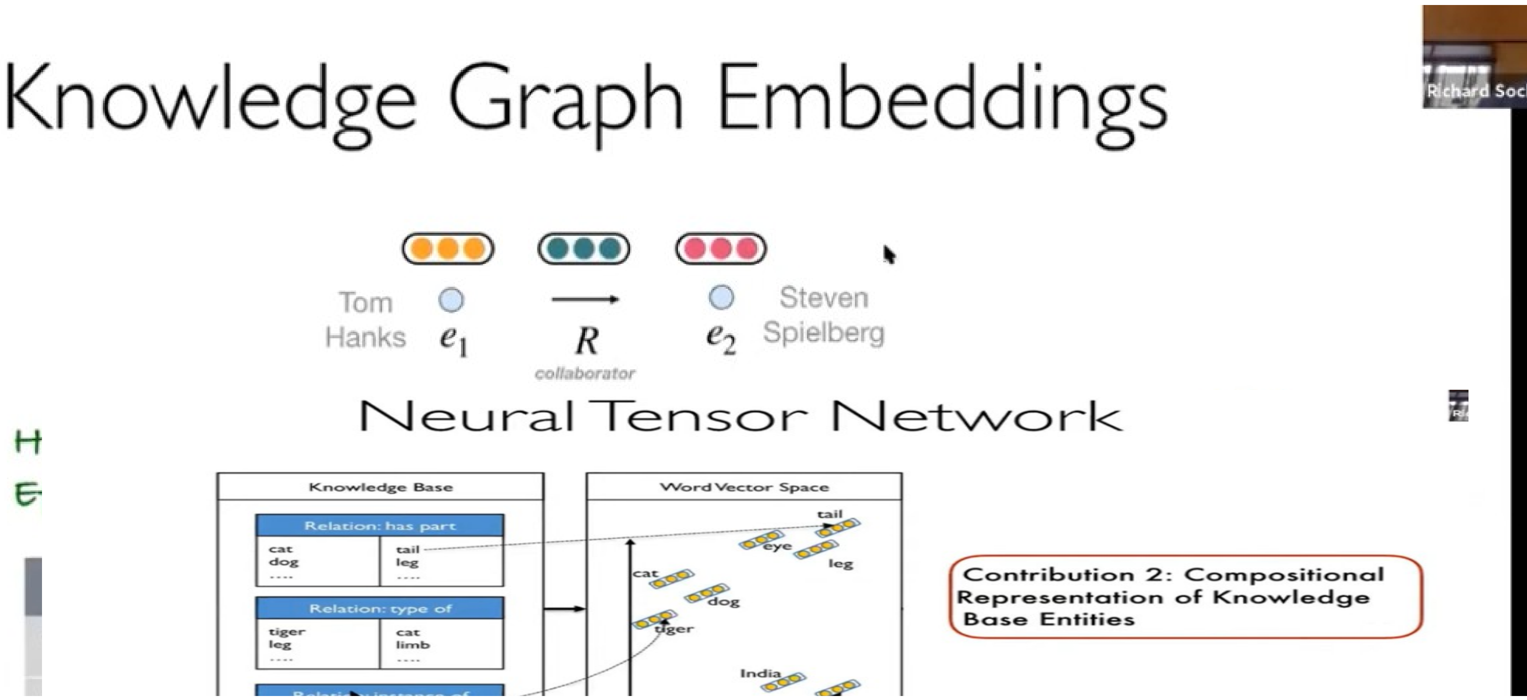
incomplete (who is a leader? Changes over time.) What about hypothetical?

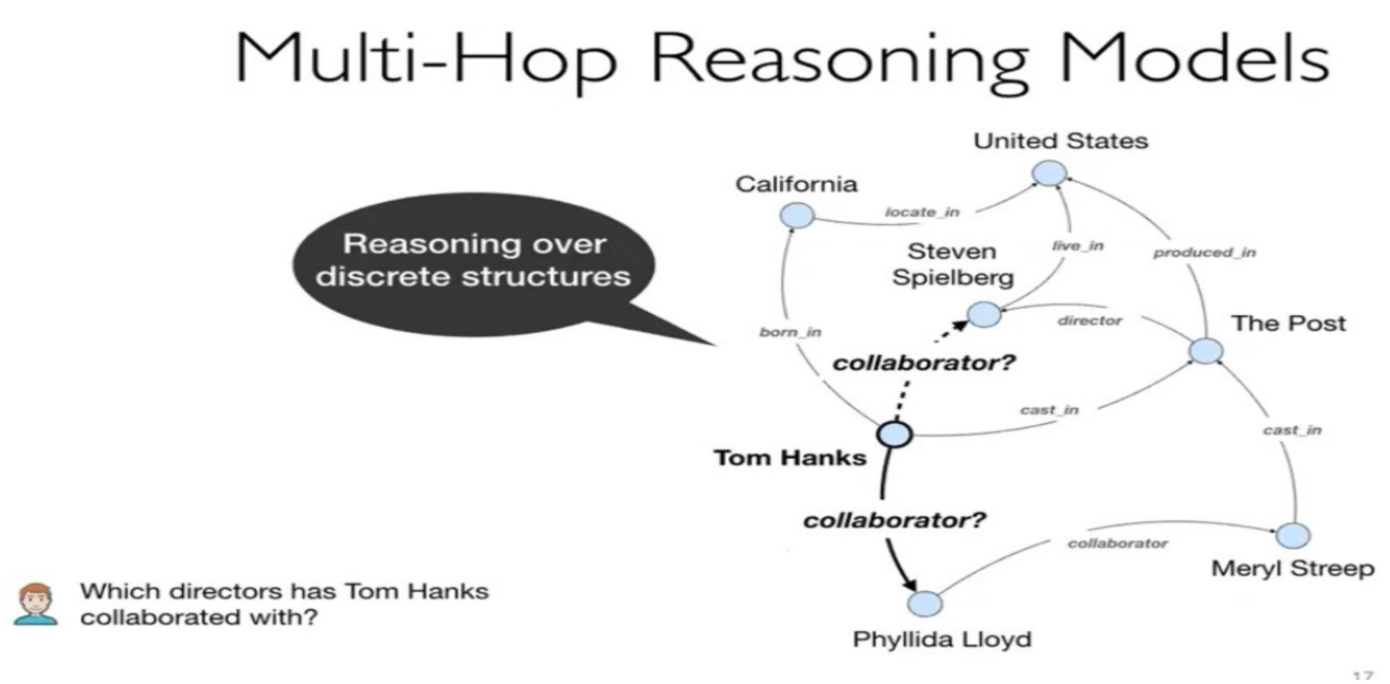
Can be useful for Covid-19 and products, orders.

We need to be more complete in filling in connections and reinforcement learning can help.

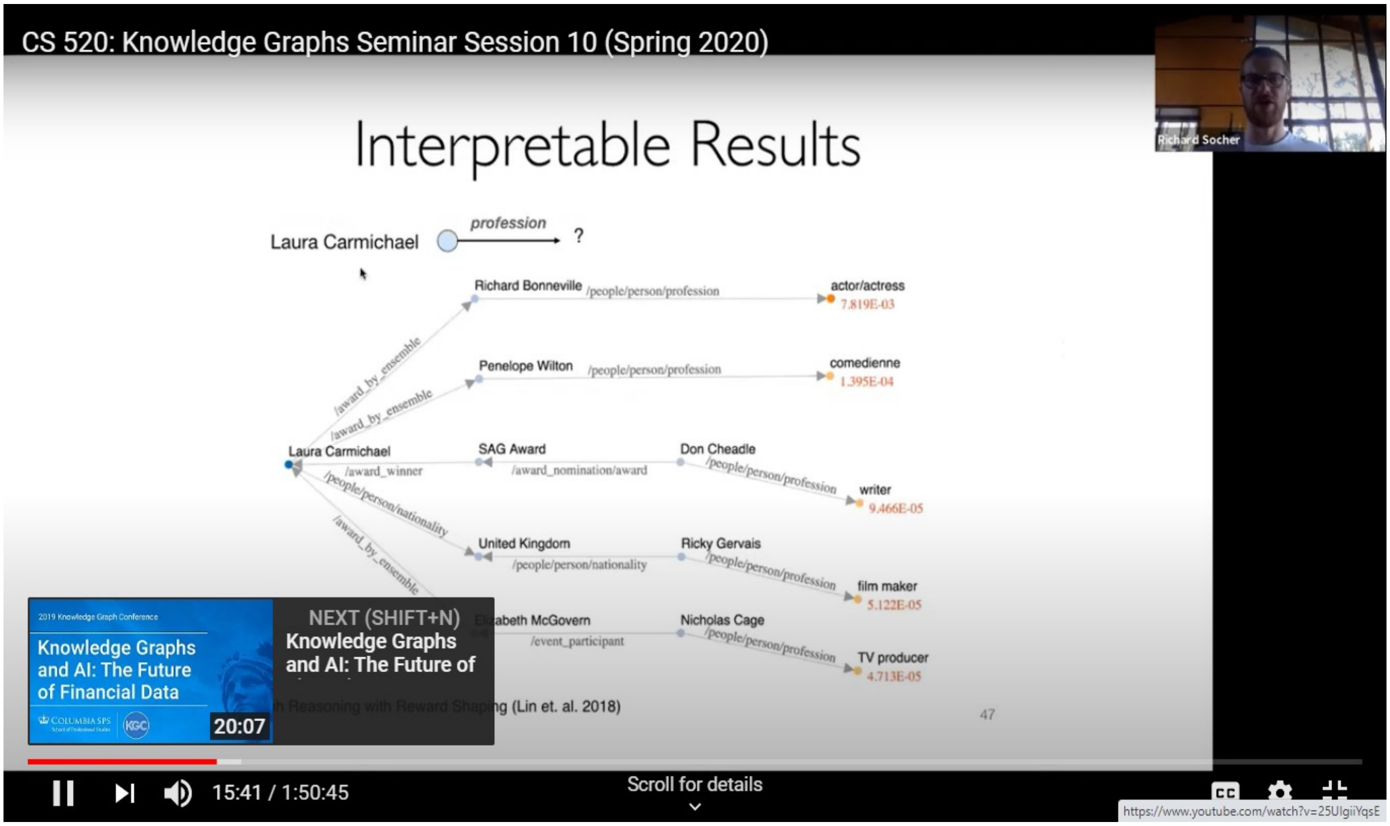
We need interpretably and complete, accurate models.

Simple and efficient Approach – embed entities as vectors into a NN architecture. You train on these.

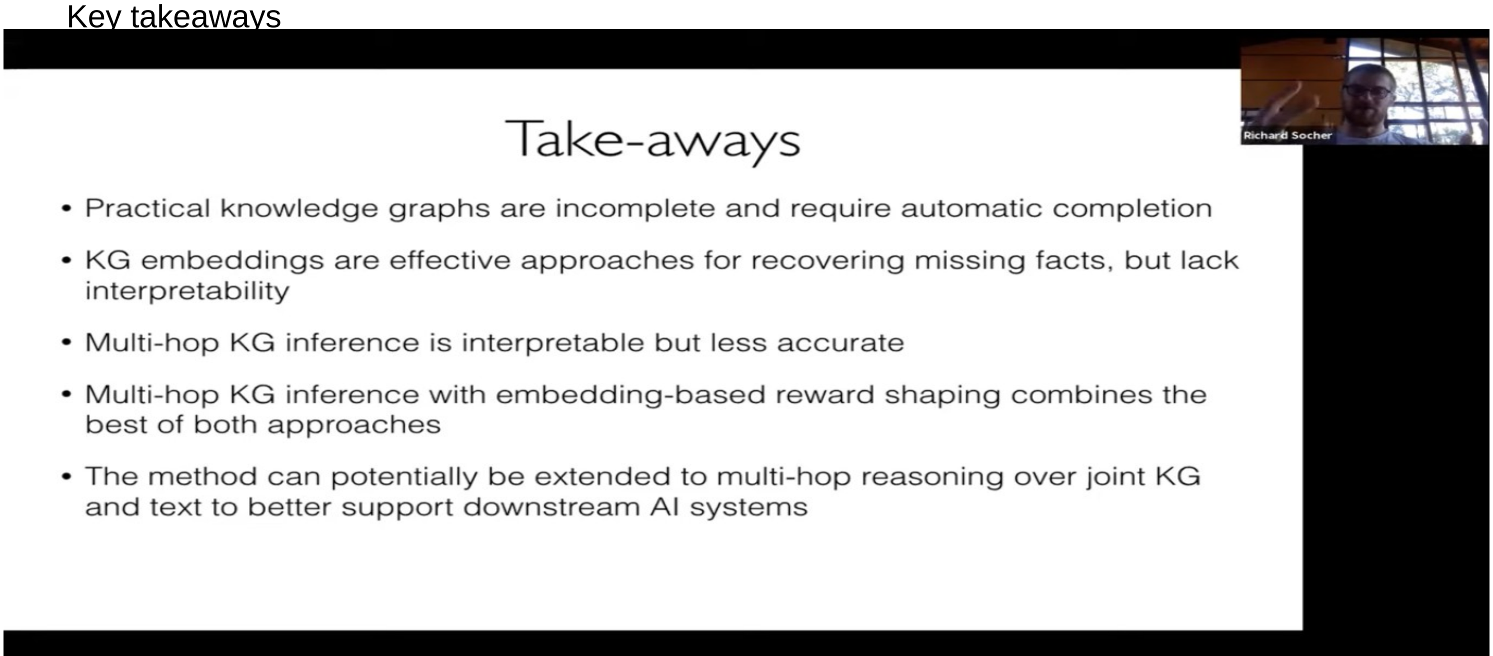




Interpretable with inference w/o lots of feature analysis



Key takeaways



Mark Musen (Stanford) What Do Knowledge Graphs Really Know?

Slides

Since the earliest days of AI, there has been a struggle to define what we really mean by knowledge and how to determine when a computational agent displays intelligence. Allen Newell had considerable influence in asking the AI community to think about knowledge in behavioral terms. He argued that knowledge representations do not manifest intelligence until some process comes along and operates on them. In that perspective, knowledge graphs are not intrinsically knowledgeable at all. We need agents to do something with our graphs to cause intelligent actions. With all the excitement about knowledge graphs, we've been focussing intently on the graphs. In the next round of research, we need to be thinking more about the knowledge.

Knowledge graphs give us the ability to represent unimaginable numbers of entities in the world and the relationships among them.

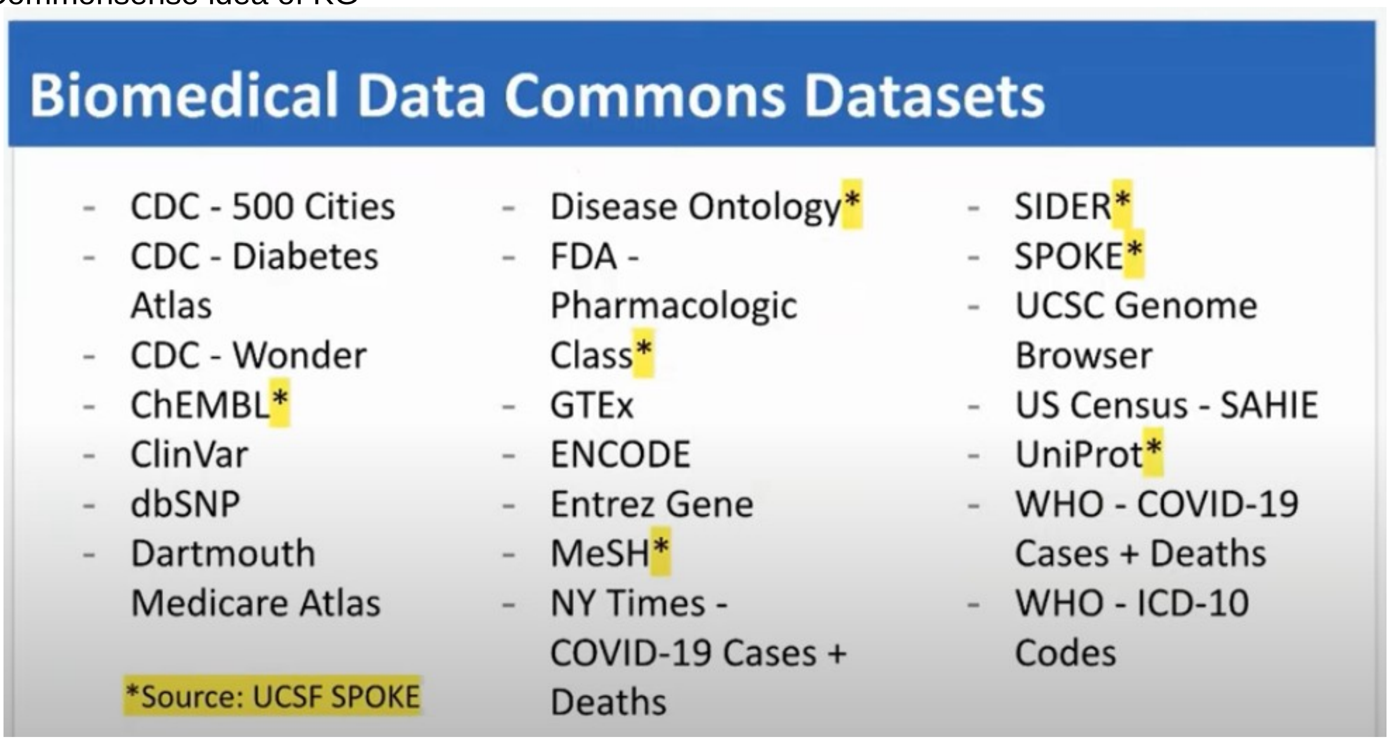
It's time to apply the same kind of principled thinking that has led us to the development of knowledge graphs to the construction of intelligent agents that can demonstrate, through their behaviors, what our graphs really know.

RV Guha Title: DataCommons

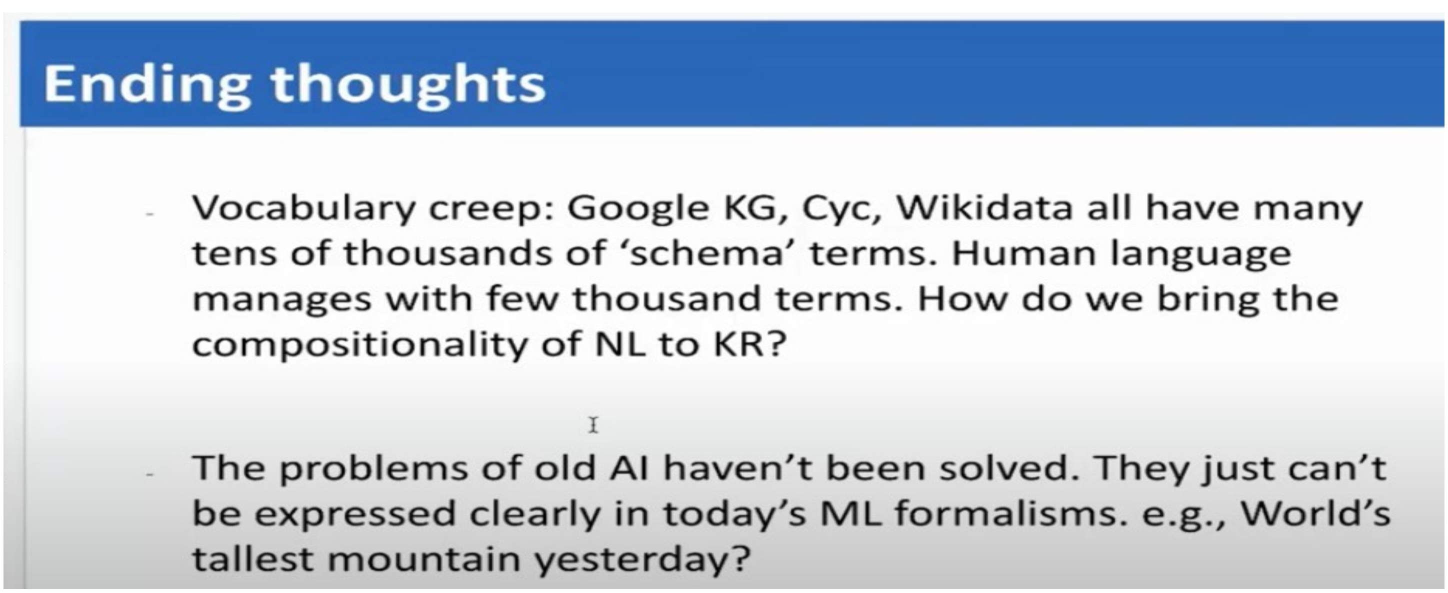
Slides

Publicly available data from open sources are a vital resource for students and researchers in a variety of disciplines. Unfortunately, processing these datasets to make them useful --- scraping, cleaning, normalizing, joining --- is tedious, error prone and has to repeated by every group. DataCommons attempts to alleviate some of this pain by synthesizing a single Knowledge Graph from many different data sources. It links references to the same entities (such as cities, counties, organizations, etc.) across different datasets to nodes on the graph, so that users can access data about a particular entity aggregated from different sources. Like the Web, the DataCommons graph is open - any user can contribute data or build applications powered by the graph. We are jumpstarting the graph with data from publicly available sources such as CDC, Census, BLS, FBI, etc. and are looking to engage with the academic community to take it further.

Commonsense idea of KG



Ending Thoughts



Questions and discussion
Language models can generate ahead.

ML and NLP can capture commonsense and generalize beyond what is read.