

Графы знаний

Лекция 1 - Введение

М. Галкин, Д. Муромцев

Содержание

- Информация по курсу
- Введение
- Почему графы?
- История развития
- Сценарии использования
- Открытые научные проблемы
- Демо

Зачем?

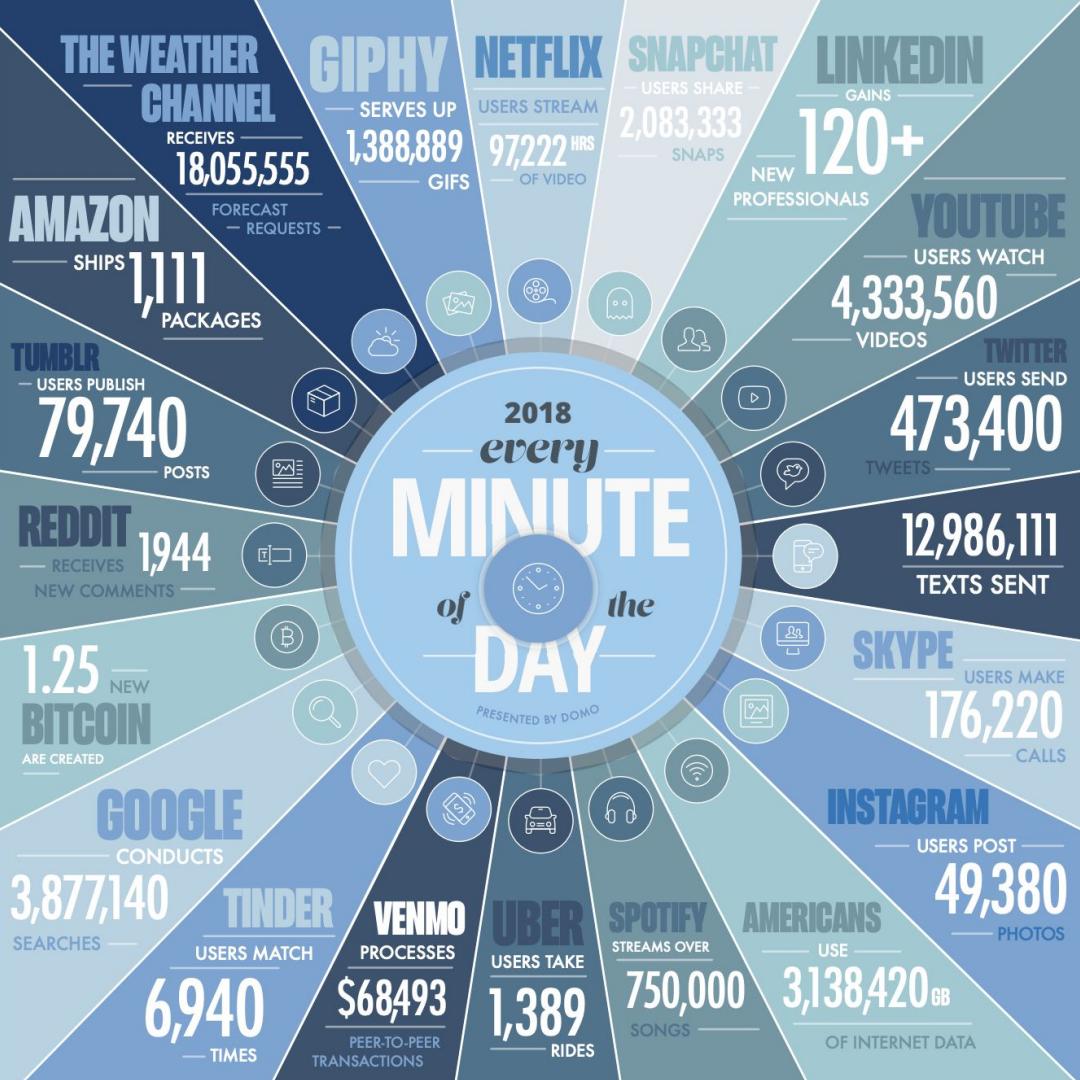
1. Графы знаний и семтех в целом очень быстро растут в применениях
2. Крайне мало информации и курсов на русском языке

План курса

1. Введение
2. Представление знаний в графах - RDF & RDFS & OWL
3. Хранение знаний в графах - SPARQL & Graph Databases
4. Однородность знаний - Reification & RDF* & SHACL & ShEx
5. Интеграция данных в графы знаний - Semantic Data Integration
6. Векторные представления графов - Knowledge Graph Embeddings
7. Введение в теорию графов - Graph Theory Intro
8. Машинное обучение на графах - Graph Neural Networks & KGs
9. Некоторые применения - Question Answering & Query Embedding

Введение

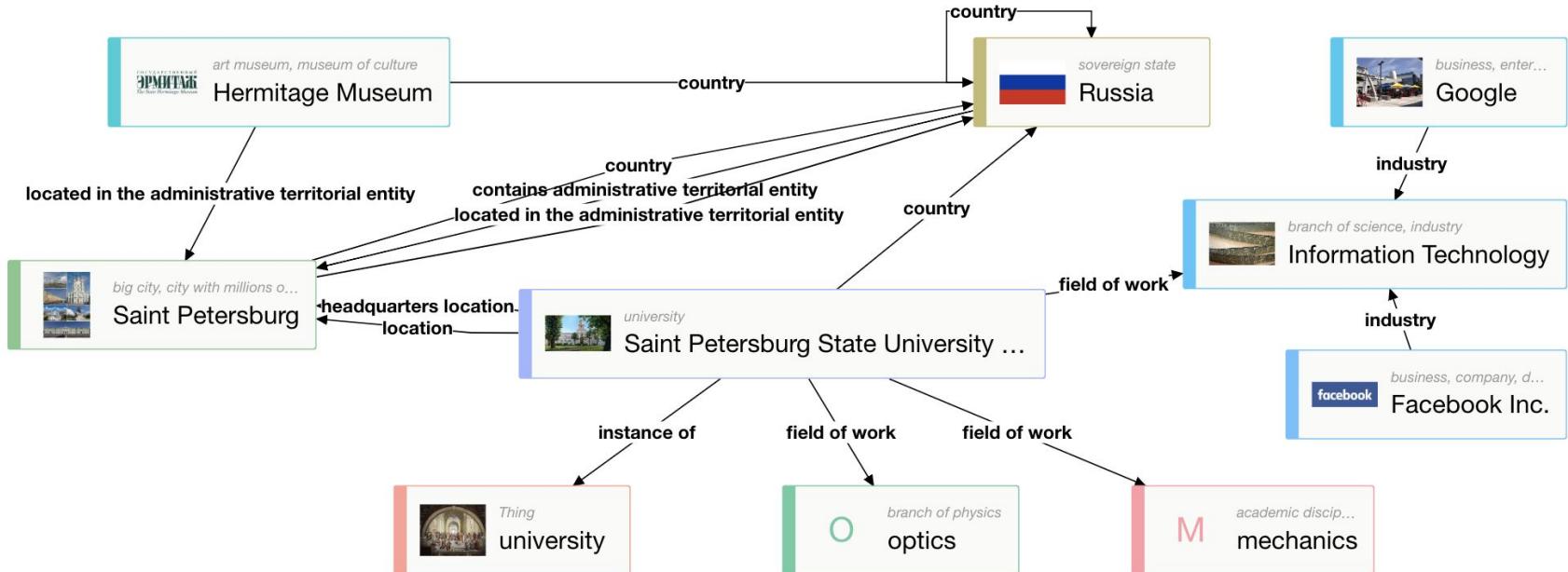
- Внушительные объемы генерируемой информации
- Проблема:
Связать разрозненные данные в коллекцию знаний, понимаемую машинами



https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf

Введение - графы

- Граф - современная форма представления знаний в сети, состоящая из уникальных сущностей (узлов) и связей между ними (ребер)



Почему графы?

- Окружающий мир довольно хорошо структурирован - есть **сущности и связи** между ними
- **Связи** часто имеют еще большее значение, чем сущности
- Графы - естественная абстракция над сетевыми структурами (карты, деревья решений, визуализации)
- Графы хорошо изучены математически

Почему графы?



Yuri Gagarin

Soviet cosmonaut

Yuri Alekseyevich Gagarin was a Soviet pilot and cosmonaut. He became the first human to journey into outer space when his Vostok spacecraft completed one orbit of the Earth on 12 April 1961. [Wikipedia](#)

Born: March 9, 1934, Klushino

Died: March 27, 1968, Kirzhach

Space missions: Vostok 1

Nationality: Soviet

Height: 1.57 m

Movies: First Orbit

People also search for

[View 15+ more](#)



Neil
Armstrong



Valentina
Tereshkova



Alexei
Leonov

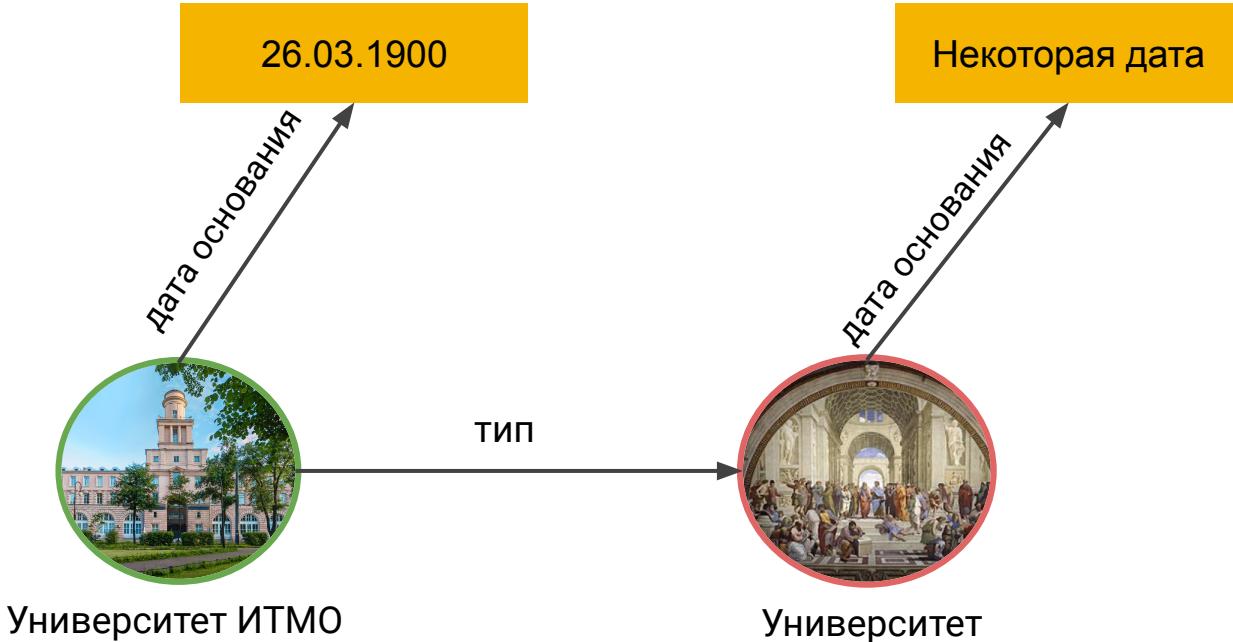


Sergei
Korolev

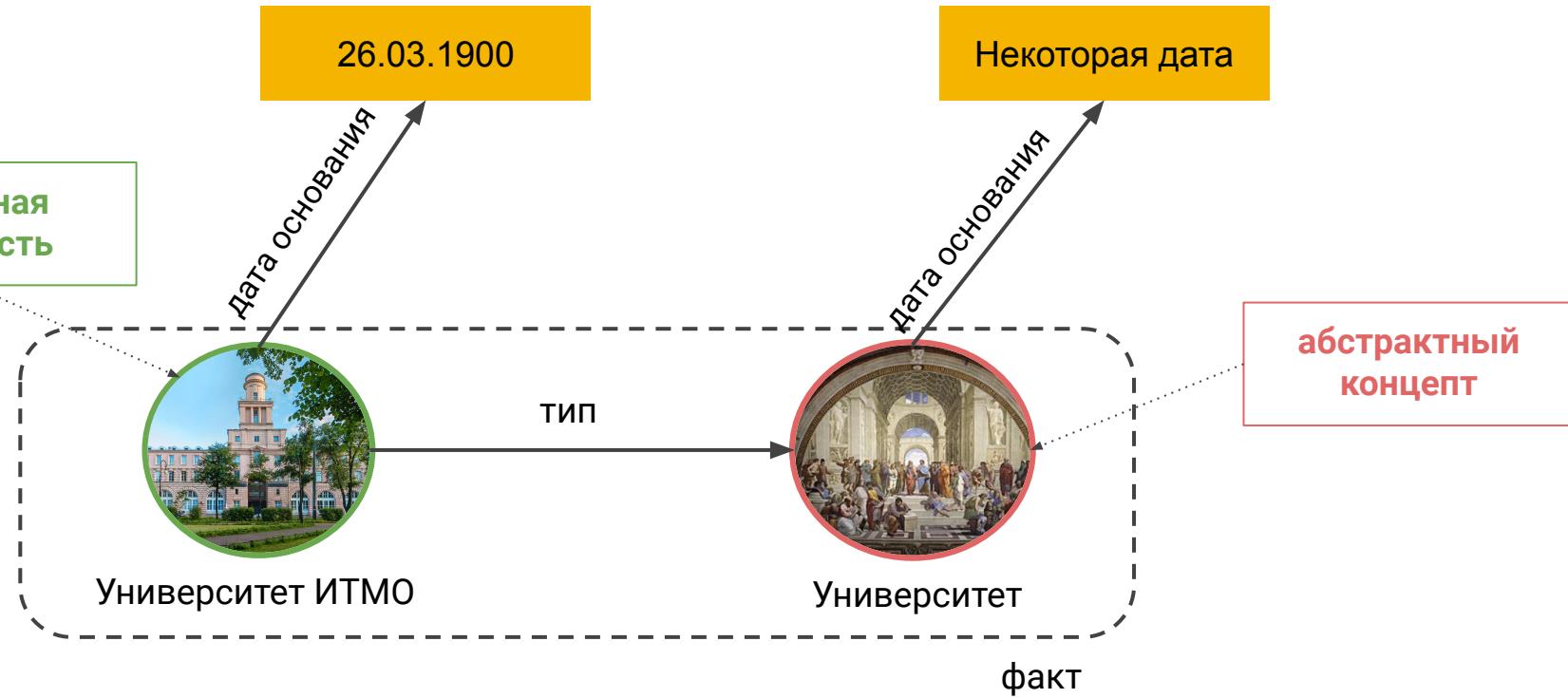
Сущность (Entity) - узел графа, обозначающий абстрактные или материальные вещи (люди, места, концепции)

Предикат (Predicate) - ребро графа, обозначающее некоторую связь между сущностями, например “родился в”

Почему графы?



Почему графы?



Почему графы?

- Можно моделировать абстрактные концепты (**Университет**) и логические высказывания
“У университетов как правило есть дата основания”
- Можно моделировать и конкретные сущности (**Университет ИТМО**)
“Университет ИТМО был основан 26 марта 1900 года”

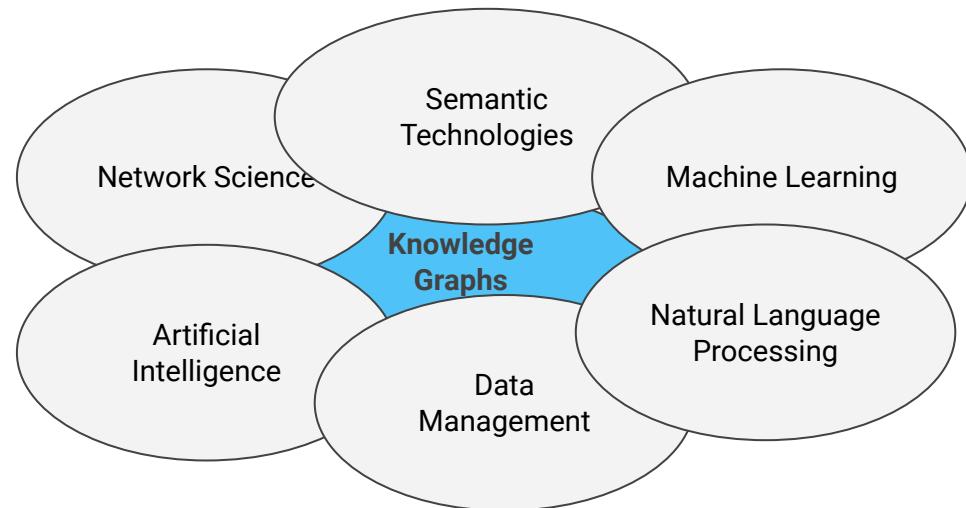
Почему графы?

- Можно моделировать абстрактные концепты (**Университет**) и логические высказывания
“У университетов как правило есть дата основания”
- Можно моделировать и конкретные сущности (**Университет ИТМО**)
“Университет ИТМО был основан 26 марта 1900 года”
- Можно производить автоматический вывод новых знаний и объяснять результаты
- А может реляционная СУБД?
 - Нет - графы как правило полуструктурированы
 - Нет - мощный логический аппарат

Исследования графов знаний

На стыке:

- Семантических технологий
- Машинного обучения
- Обработки естественных языков
- Управления и хранения данными
- Наука о сетях
- Искусственного интеллекта в целом



Представление графов знаний



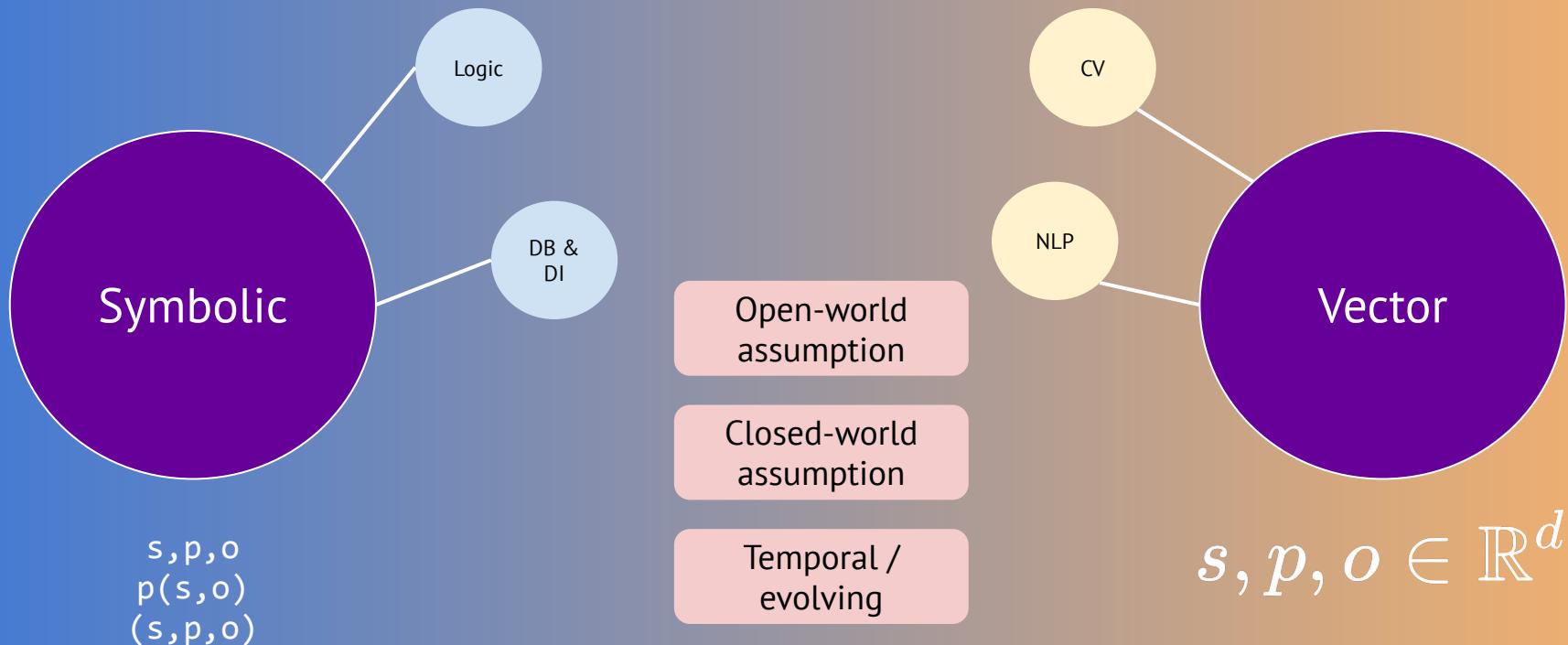
Symbolic

s, p, o
 $p(s, o)$
 (s, p, o)

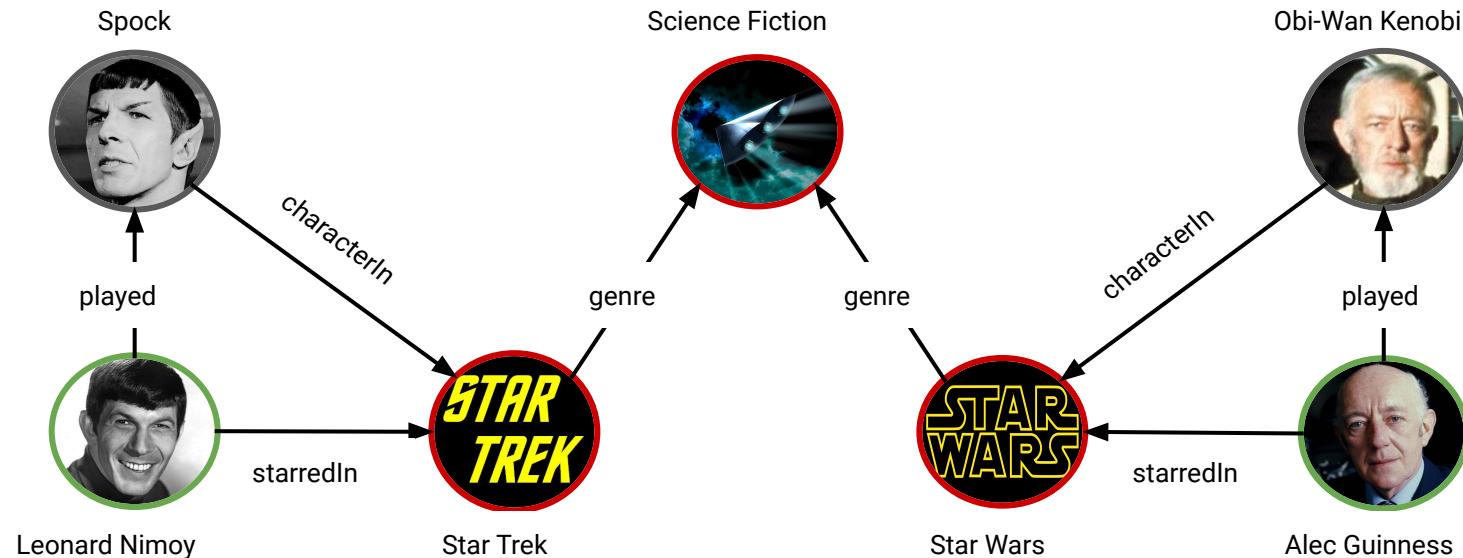
Vector

$s, p, o \in \mathbb{R}^d$

Представление графов знаний



Symbolic: Триплеты



Leonard Nimoy played

Leonard Nimoy starredIn

Spock

Star Trek

Spock

Star Trek

characterIn

Star Trek

genre

Alec Guinness

Alec Guinness

Obi-Wan Kenobi

Star Wars

starredIn

played

characterIn

genre

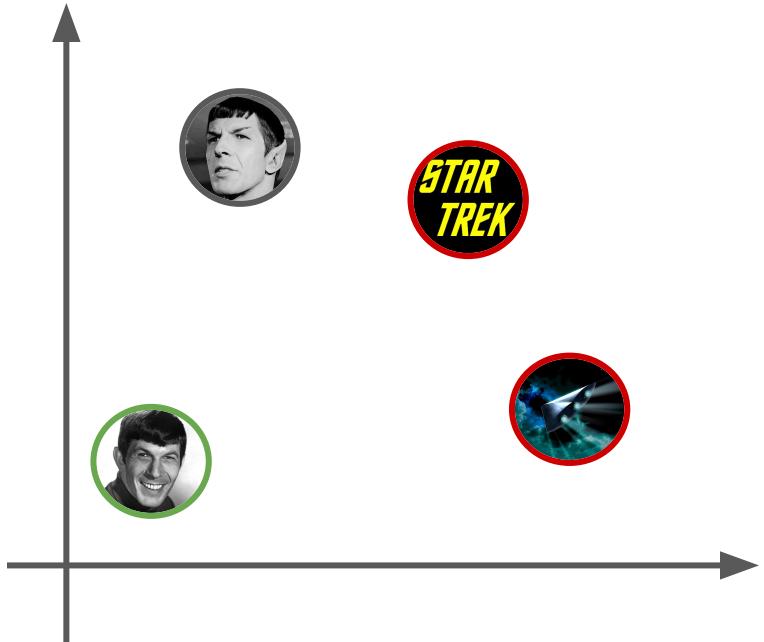
Star Wars

Obi-Wan Kenobi

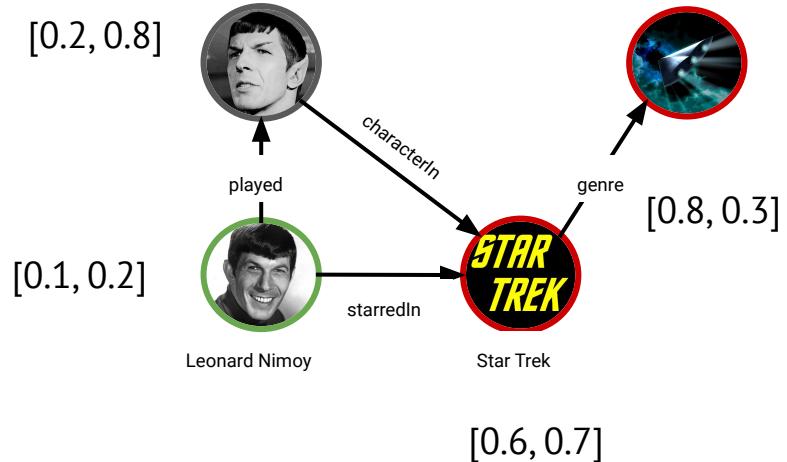
Star Wars

Science Fiction

Vector: Embeddings



$$E \in \mathbb{R}^{N_e \times d}$$
$$R \in \mathbb{R}^{N_r \times d}$$



История развития



2007



- Получена парсингом инфобоксов Википедии на всех языках
- Более 6 миллиардов фактов
- Первый де-факто стандарт создания и публикации графов знаний

История развития



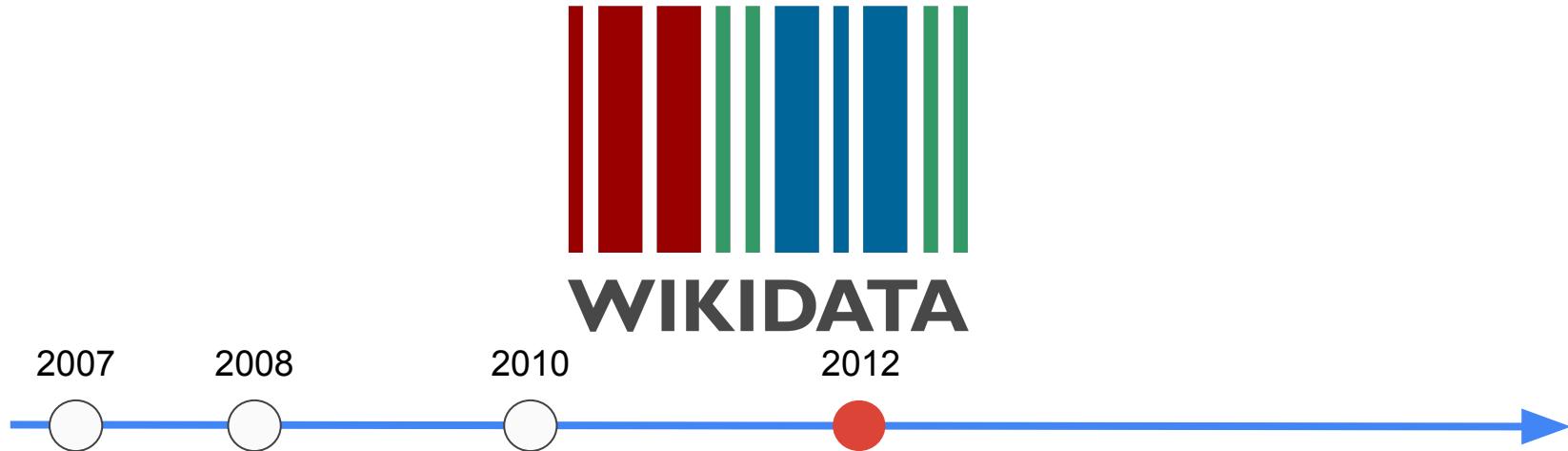
- Получена обработкой Википедии и семантического тезауруса WordNet
- Более 120 миллионов фактов

История развития



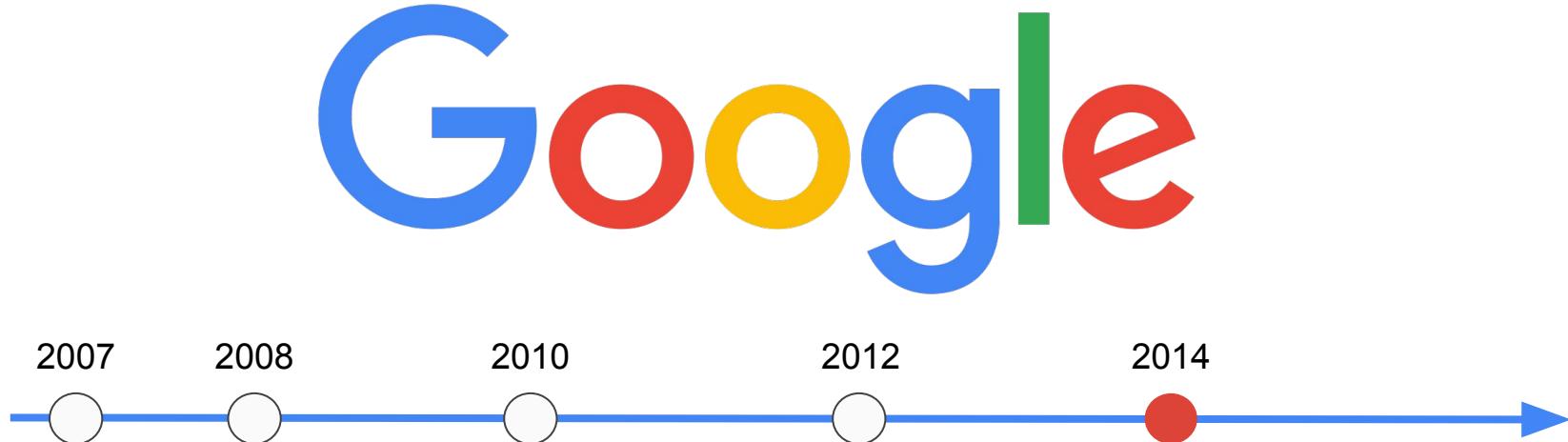
- NELL - Never Ending Language Learner
- Читает веб-страницы и ссылки, выделяя факты в граф знаний
- Более 14.5 миллиардов фактов

История развития



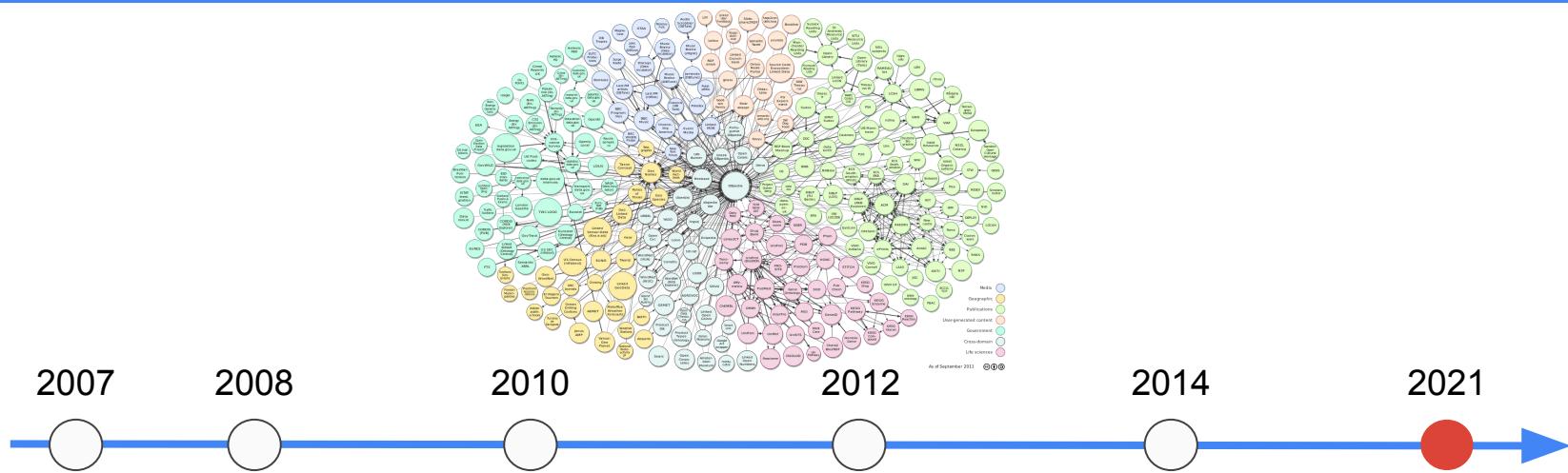
- В отличие от DBpedia, создан чтобы служить источником фактов для инфобоксов Википедии
- Гибкая схема моделирования и представления знаний
- Более 7 миллиардов фактов о 90 миллионах сущностей
- Google и Facebook загружают свои датасеты именно сюда

История развития



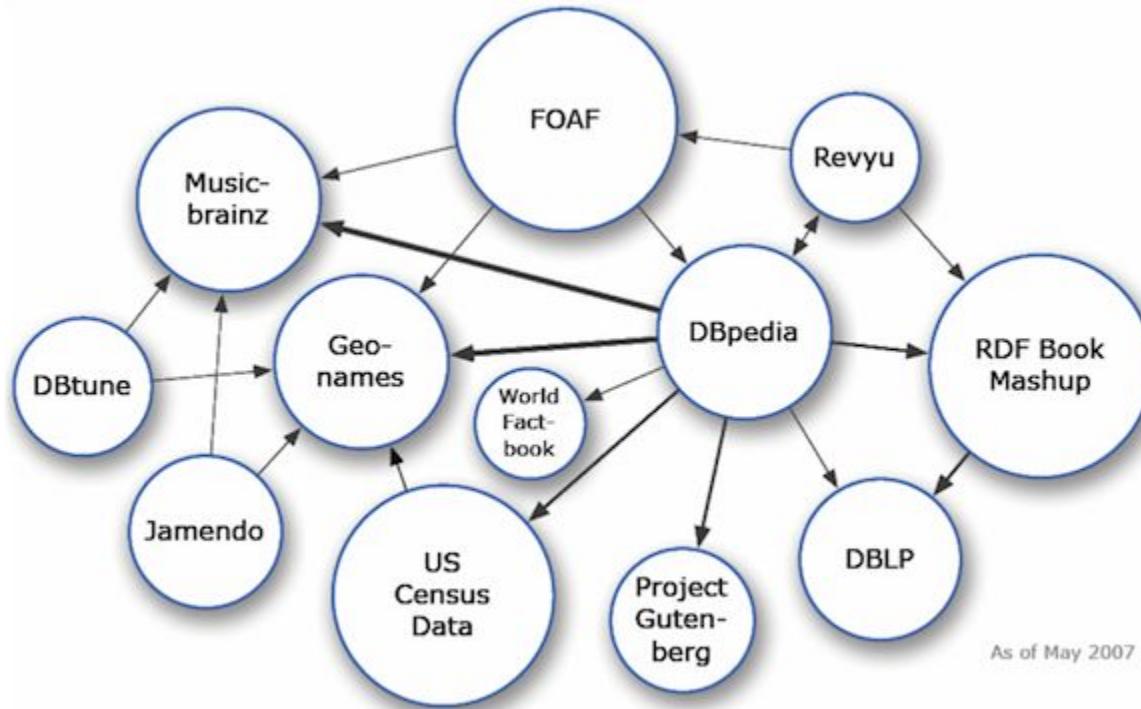
- Google Knowledge Graph - развитие графа Freebase (2007)
- Интегрирует общеизвестные факты и пользовательские данные
- После анонса остальные компании захотели “такое же”

История развития

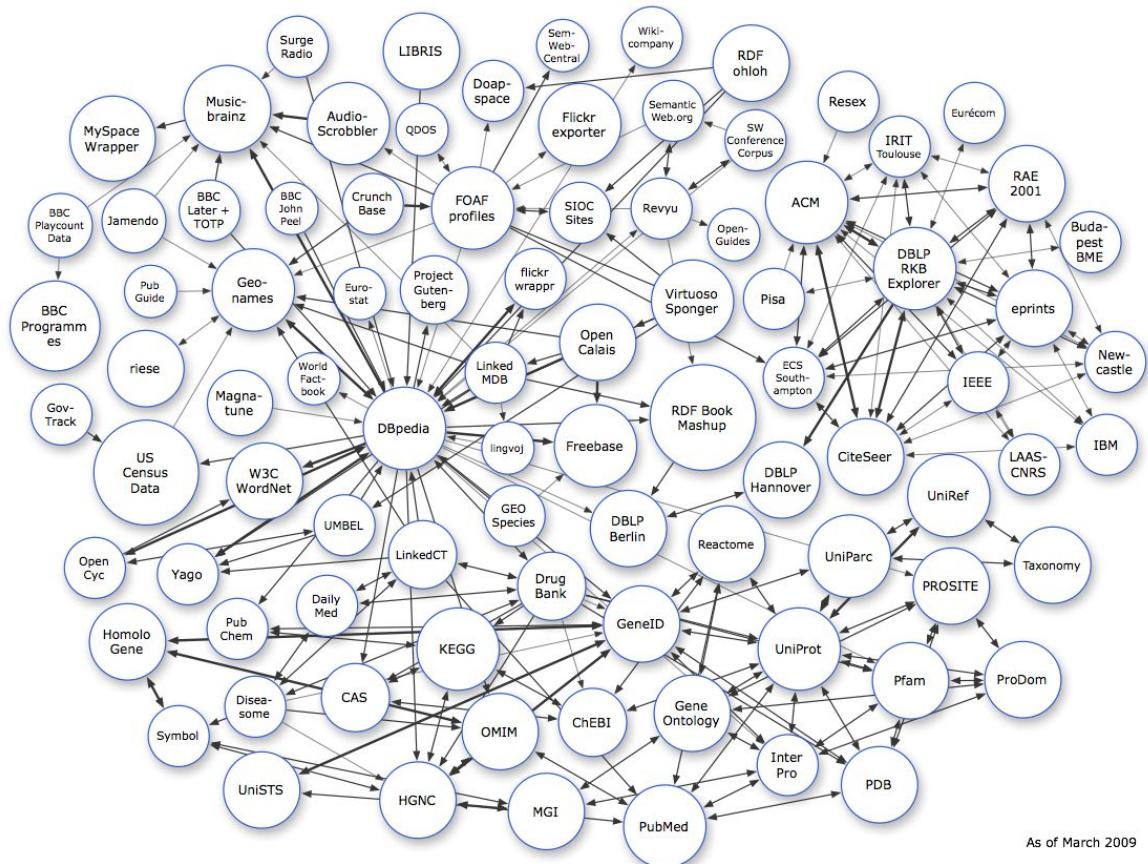


- Открытые графы знаний
- Специализированные (domain-specific) графы знаний
- Персональные синтезируемые в процессе обработки графы

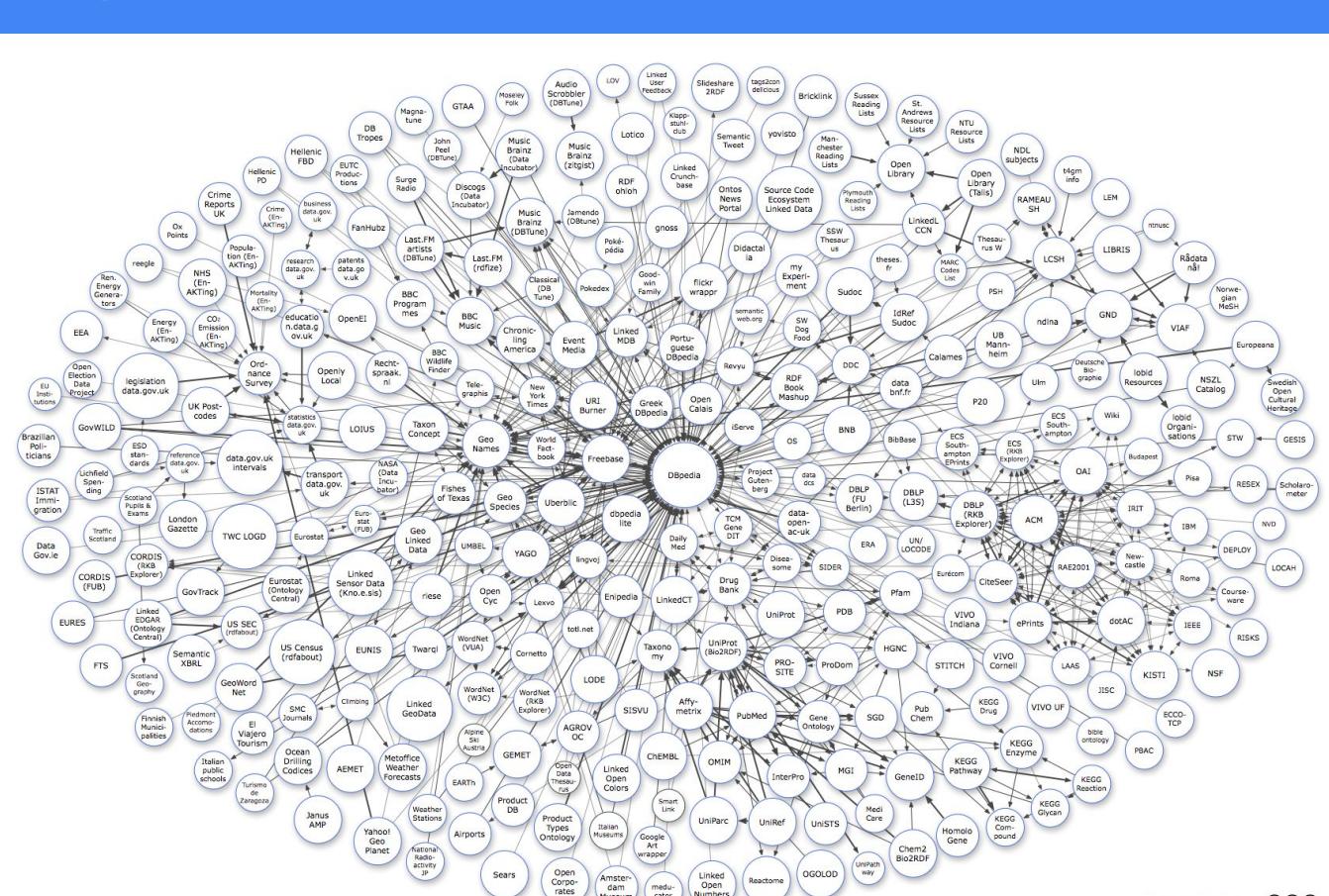
Linked Open Data - 2007



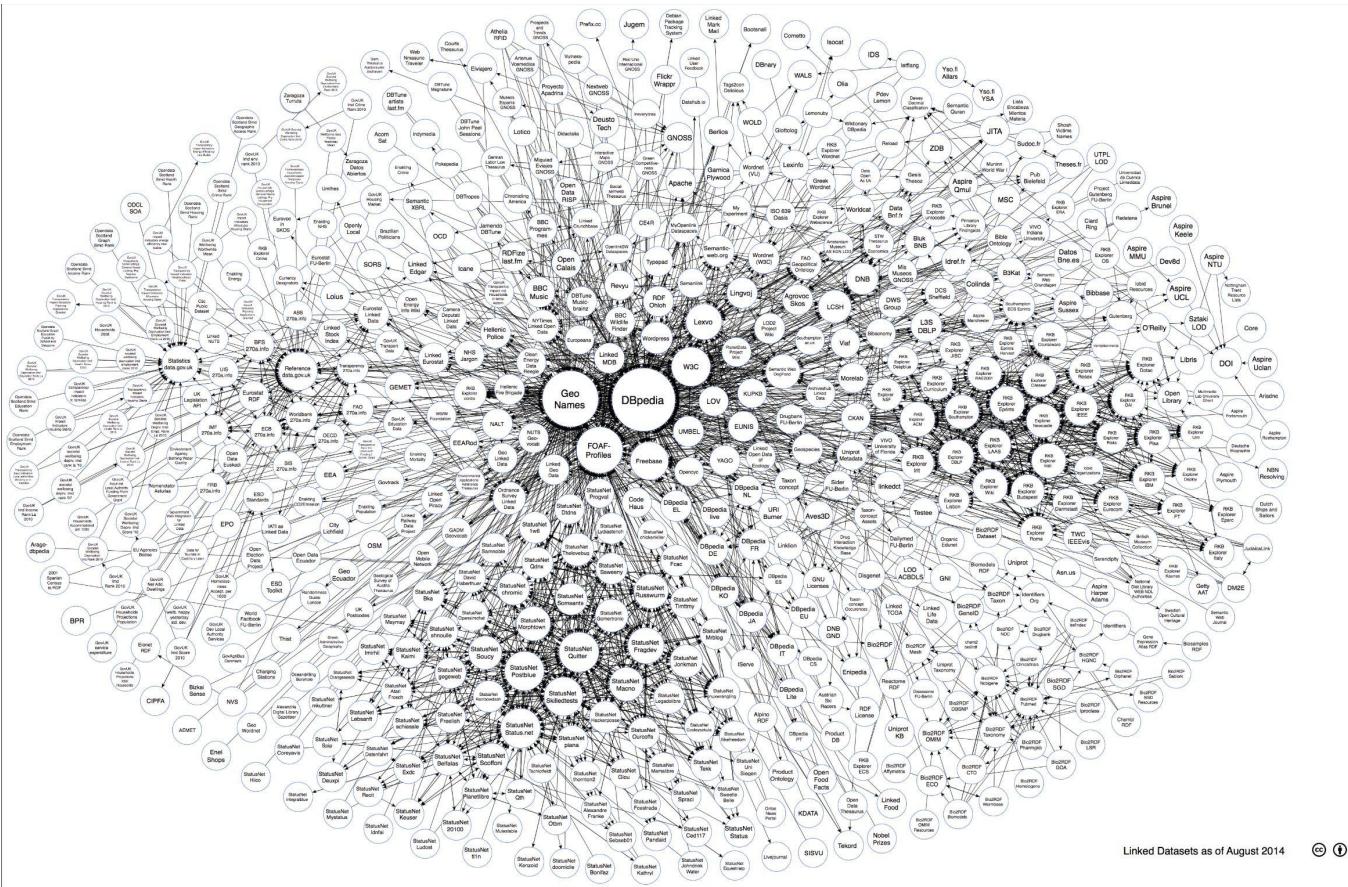
Linked Open Data - 2009



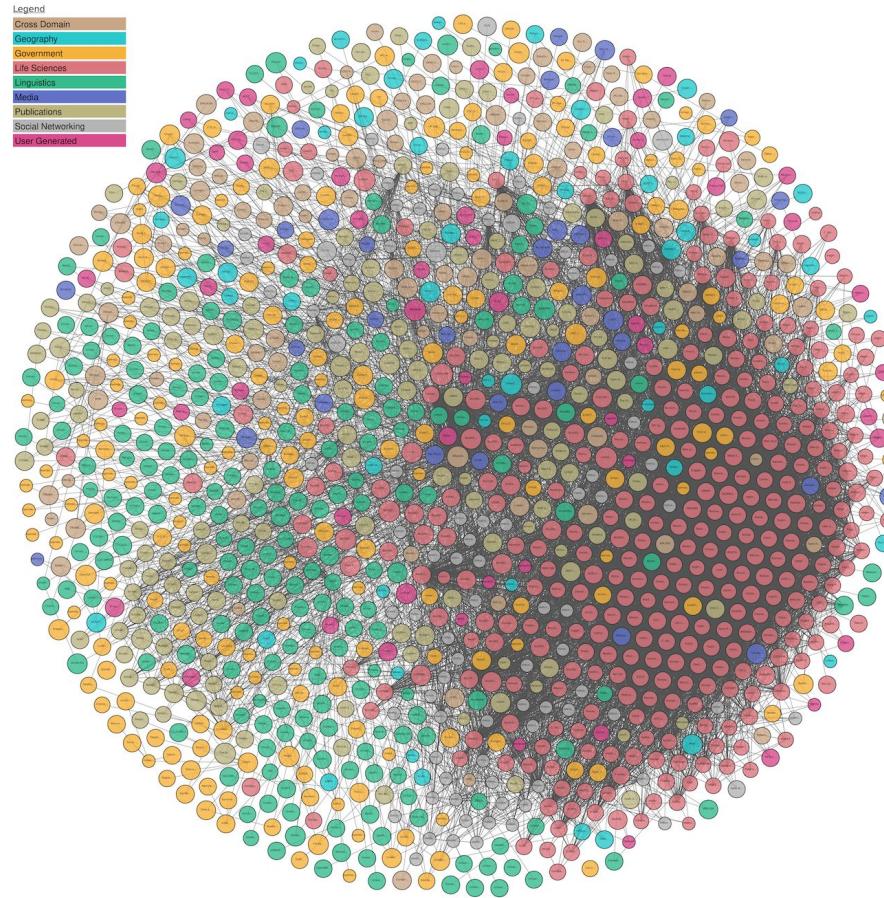
Linked Open Data - 2011



Linked Open Data - 2014



Linked Open Data - 2020

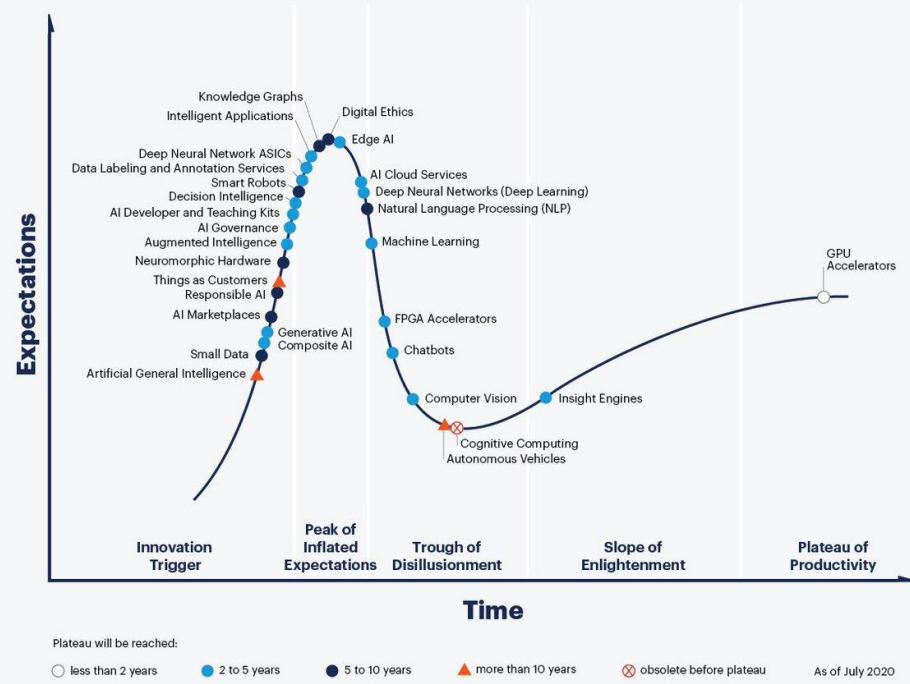


Сценарии использования графов знаний

Анализ рынка:

- Графы знаний на подъеме!
- Технологии, использующие графы знаний, на подъеме!
- В среднесрочной перспективе спрос со стороны бизнеса будет расти

Hype Cycle for Artificial Intelligence, 2020



gartner.com/SmarterWithGartner

Source: Gartner
© 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S.

Gartner

Сценарии использования графов знаний

Largest change in market cap by company (2009 to 31 March 2018)

	Company name	Location	Industry	Change in market cap 2009-2018 (\$bn)	Market cap 2018 (\$bn)
1	Apple	United States	Technology	757	851
2	Amazon.Com	United States	Consumer Services	670	701
3	Alphabet	United States	Technology	609	719
4	Microsoft Corp	United States	Technology	540	703
5	Tencent Holdings	China	Technology	483	496
6	Facebook	United States	Technology	383(1)	464
7	Berkshire Hathaway	United States	Financial	358	492
8	Alibaba	China	Consumer Services	302(1)	470
9	JPMorgan Chase	United States	Financials	275	375
10	Bank of America	United States	Financials	263	307

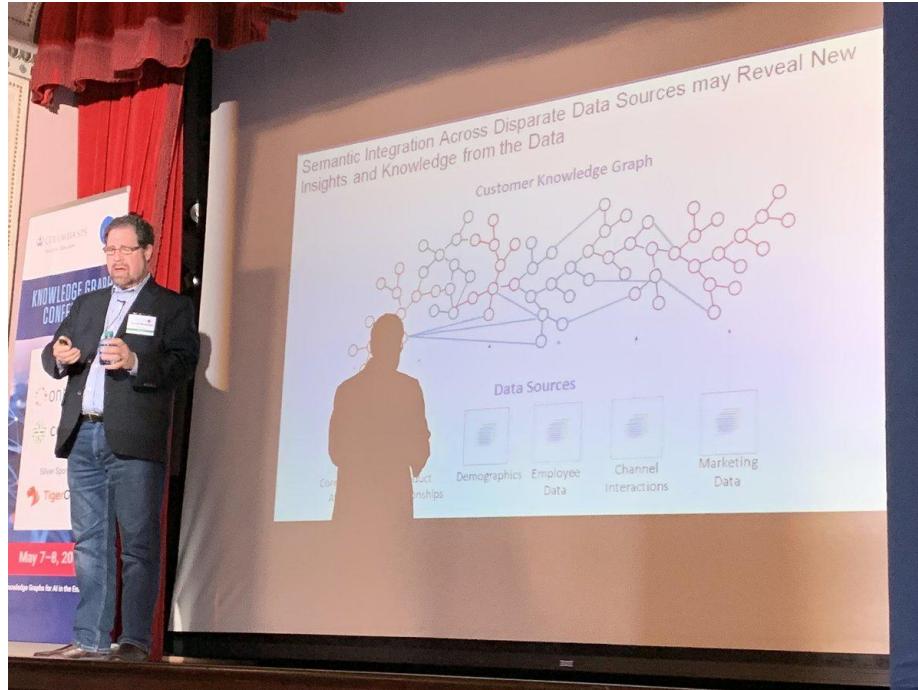
(1) Change in market cap from IPO date

(2) Market cap at IPO date

Sources: Bloomberg and PwC analysis

Сценарии использования графов знаний (KGC 2019)

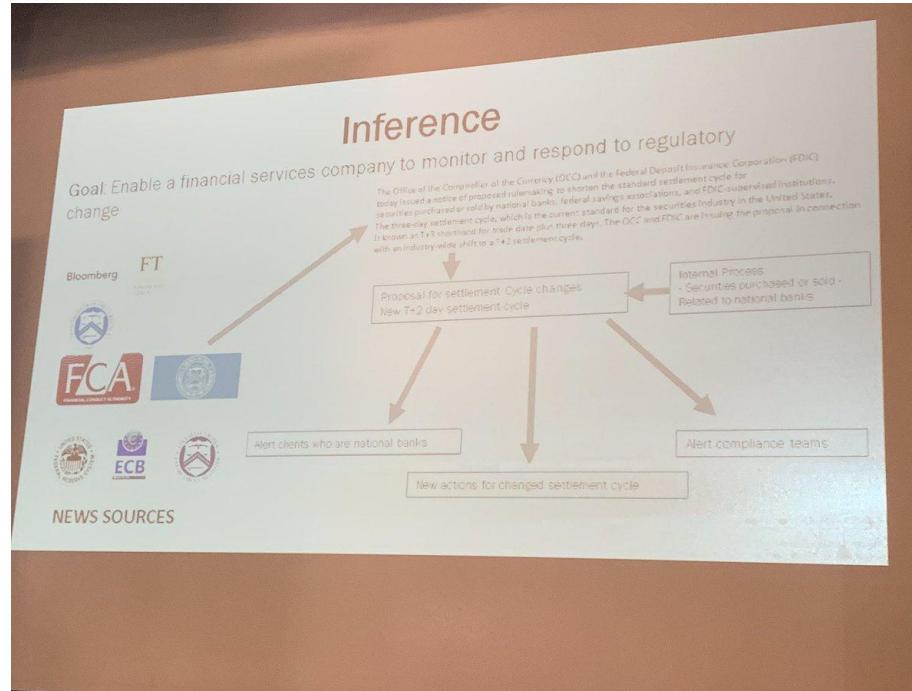
- Банки: анализ транзакций, fraud detection



Goldman Sachs, Wells Fargo

Сценарии использования графов знаний (KGC 2019)

- Банки: анализ транзакций, fraud detection
- Консалтинг: графы на основе юридических документов



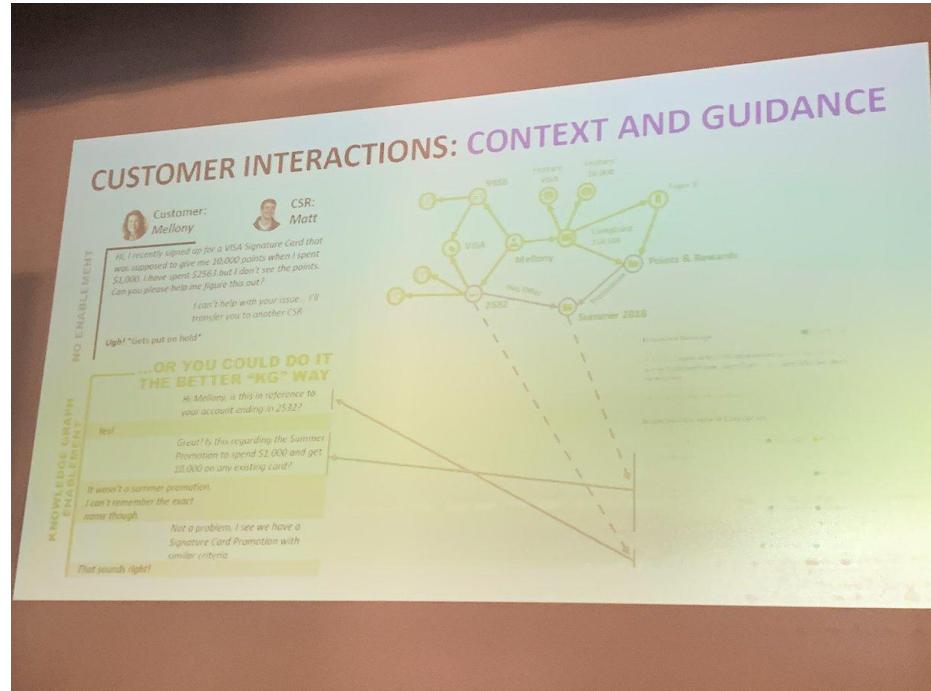
Сценарии использования графов знаний (KGC 2019)

- Банки: анализ транзакций, fraud detection
- Консалтинг: графы на основе юридических документов
- Индустрия: supply chain management



Сценарии использования графов знаний (KGC 2019)

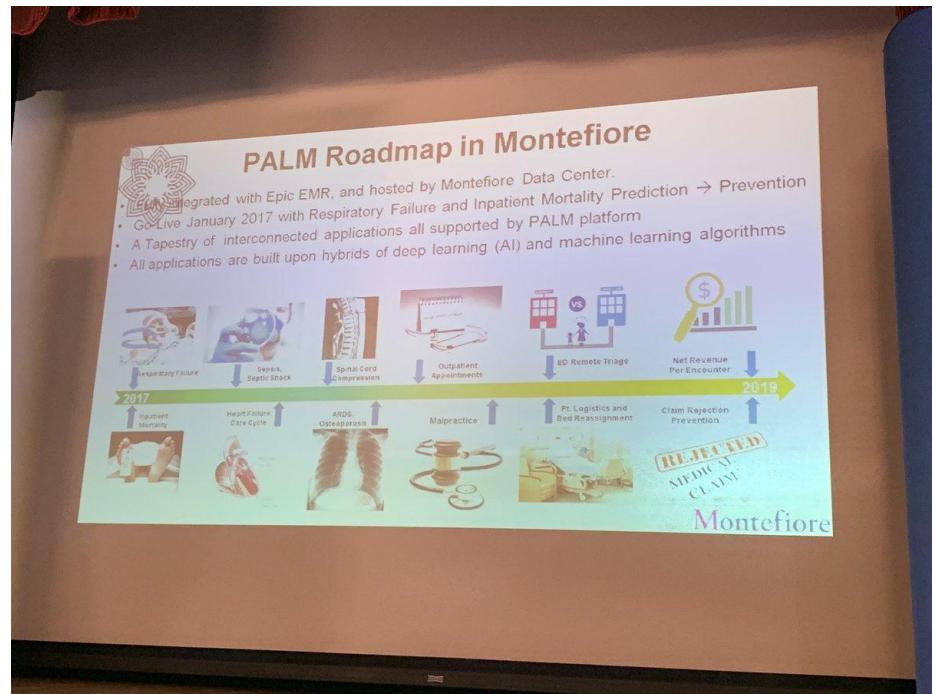
- Банки: анализ транзакций, fraud detection
- Консалтинг: графы на основе юридических документов
- Индустрия: supply chain management
- Сфера услуг: чат-боты



DiffBot, AirBnB

Сценарии использования графов знаний (KGC 2019)

- Банки: анализ транзакций, fraud detection
- Консалтинг: графы на основе юридических документов
- Индустрия: supply chain management
- Сфера услуг: чат-боты
- Здравоохранение: EHR (electronic health records)



Сценарии использования графов знаний

❖ Обогащение результатов поиска и ответы на простые вопросы

когда родился альберт эйнштейн?

Все Картинки Новости Видео Карты Ещё Настройки Инструменты

Результатов: примерно 363 000 (0,63 сек.)

Альберт Эйнштейн / Дата рождения

14 марта 1879 г.

Похожие запросы

Стiven Уильям Хокинг 8 января 1942 г. Исаак Ньютоn 4 января 1643 г. Эдуард Эйнштейн 28 июля 1910 г.

Оставить отзыв

Эйнштейн, Альберт — Википедия
https://ru.wikipedia.org/wiki/%D0%95%D0%B9%D0%BD%D1%8C%D1%82%D0%B5%D0%BD%D0%B8%D1%8F,_%D0%A1%D0%BB%D0%BE%D0%BC

Альберт Эйнштейн (нем. Albert Einstein, МФА [ˈalbɛt ˈajnʃtaɪn]; 14 марта 1879 ... править код]. Альберт Эйнштейн родился 14 марта 1879 года в южно-германском городе Ульме, в небогатой еврейской семье. Мозг Альberta Эйнштейна · Marić, Mileva · Эйнштейн, Эльза · Планк, Макс

Биография Альбера Эйнштейна - РИА Новости, 14.03.2014
<https://ria.ru/20140314/999156297.html>

Альберт Эйнштейн

Физик-теоретик

Альберт Эйнштейн — физик-теоретик, один из основателей современной теоретической физики, лауреат Нобелевской премии по физике 1921 года, общественный деятель-гуманист. Жил в Германии, Швейцарии и США. [Википедия](#)

Родился: 14 марта 1879 г., Ульм

Умер: 18 апреля 1955 г., Princeton Medical Center at Plainsboro, Нью-Джерси, США

Супруга: Эльза Эйнштейн (в браке с 1919 г. до 1936 г.), Милева Marić (в браке с 1903 г. до 1919 г.)

Образование: Цюрихский университет (1905 г.), ЕШЭ

Дети: Эдуард Эйнштейн, Ганс Альберт Эйнштейн, Лизэрл Эйнштейн

КНИГИ

Относите... специаль... и общая... 1916 г.

The Quotable Einstein

Мир, каким я его вижу 1949 г.

Идеи и мнения 1954 г.

The Meaning of Relativity 1922 г.

Ещё 35+

Оставить отзыв

Сценарии использования графов знаний

- ❖ Ответы на сложные вопросы, заданные на естественном языке



Royal Society

The President, Council, and Fellows of the Royal Society of London for Improving Natural Knowledge, commonly known as the Royal Society, is a learned society for science and is possibly the oldest such society still in existence. Founded in November 1660, it was granted a royal charter by King Charles II as "The Royal Society". The Society is the United Kingdom's and Commonwealth of Nations' Academy of Sciences and fulfills a number of roles; promoting science and its benefits, recognising excellence in science, supporting outstanding science, providing scientific advice for policy, fostering international and global cooperation, education and public engagement. The society is governed by its Council, which is chaired by the Society's President, according to a set of statutes and standing orders. The members of Council and the President are elected from and by its Fellows, the basic members of the society, who are themselves elected by existing Fellows. As of 2016, there are about 1,600 fellows, allowed to use the postnominal title FRS (Fellow of the Royal

Max Planck Medal

The Max Planck medal is the highest award of the Deutsche Physikalische Gesellschaft (English: German Physical Society); the world's largest organization of physicists; for extraordinary achievements in theoretical physics. It is awarded annually. The winner is awarded with gold medal and hand-written parchment.

Show detailed

Which award did Albert Einstein receive?

Nobel Prize in Physics

The Nobel Prize in Physics (Swedish: Nobelpriiset i fysik) is a yearly award given by the Royal Swedish Academy of Sciences for those who conferred the most outstanding contributions for mankind in the field of physics. It is one of the five Nobel Prizes established by the will of Alfred Nobel in 1869 and awarded since 1901; the others being the Nobel Prize in Chemistry, Nobel Prize in Literature, Nobel Peace Prize, and Nobel Prize in Physiology or Medicine. The first Nobel Prize in Physics was awarded to a German physicist Wilhelm Röntgen in recognition of the extraordinary services he has rendered by the discovery of the remarkable rays (or x-rays). This award is administered by the Nobel Foundation and widely regarded as the most prestigious award that a scientist can receive in physics. It is presented in Stockholm at an annual ceremony on December 10, the anniversary of Nobel's death. Through 2016, a total of 203 individuals have been awarded the prize.

Copley Medal

The Copley Medal is a scientific award given by the Royal Society, London, for "outstanding achievements in research in any branch of science." It alternates between the physical and the biological sciences. Given every year, the medal is the oldest Royal Society medal still awarded, and probably the oldest surviving scientific award in the world, having first been given in 1731 to Stephen Gray, for "his new Electrical Experiments: – as an encouragement to him for the readiness he has always shown in obliging the Society with his discoveries and improvements in this part of Natural Knowledge".

Show detailed

Matteucci Medal

The Matteucci Medal is an Italian award for physicists, named after Carlo Matteucci. It was established to award physicists for their fundamental contributions. Under an Italian Royal Decree dated July 10, 1870, the Italian Society of Sciences was authorized to receive a donation from Carlo Matteucci for the establishment of the Prize. Matteucci Medalists Source: Italian Society of Sciences)* 1868 Hermann Helmholtz * 1875 Henri Victor Regnault * 1876 William Thomson, 1st Baron Kelvin * 1877 Gustav Kirchhoff * 1878 Gustav Wiedemann * 1879 Wilhelm Eduard Weber * 1880 Antonio Pacinotti * 1881 Emilio Viviani * 1882 Augusto Righi * 1887 Thomas Edison * 1888 Heinrich Rudolph Hertz * 1894 John Strutt, 3rd Baron Rayleigh * 1895 Henry Augustus Rowland * 1896 Wilhelm Conrad Röntgen and Philipp Lenard * 1901 Guglielmo Marconi * 1903 Albert Abraham Michelson * 1904 Marie Curie and Pierre Curie * 1905 Henri Poincaré * 1906 James Dewar * 1907 William Ramsay * 1908 Antonio Garibaldi * 1909 Orso Mario Corbino * 1910 Heike

Time 100: The Most Important People of the Century

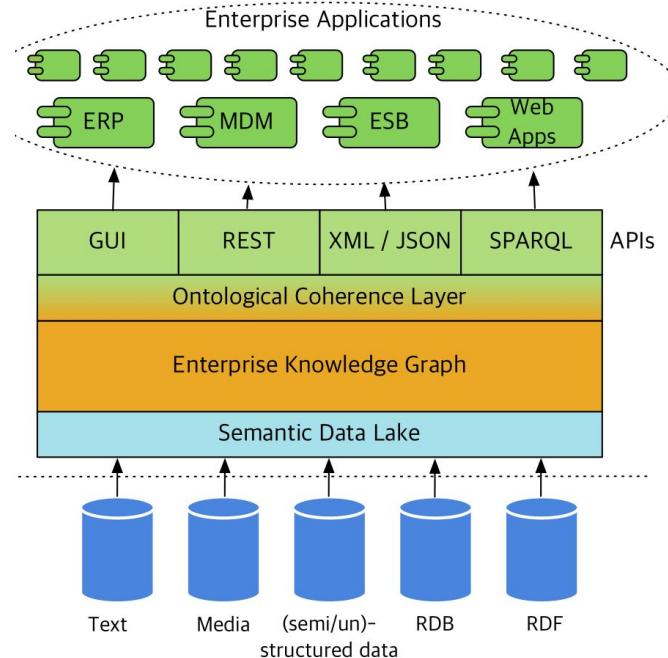
Time 100: The Most Important People of the Century is a compilation of the 20th century's 100 most influential people, published in Time magazine in 1999. The idea for such a list started on February 1, 1998, with a debate at a symposium in Hanoi, Vietnam. The panel participants were former CBS Evening News anchor Dan Rather, historian Doris Kearns Goodwin, former New York governor Mario Cuomo, then-Stanford Provost Condoleezza Rice, publisher Irving Kristol, and Time managing editor Walter Isaacson. In a separate issue on December 31, 1999, Time recognized Einstein as the Person of the Century.

Show detailed

<http://asknow.sda.tech>

Сценарии использования графов знаний

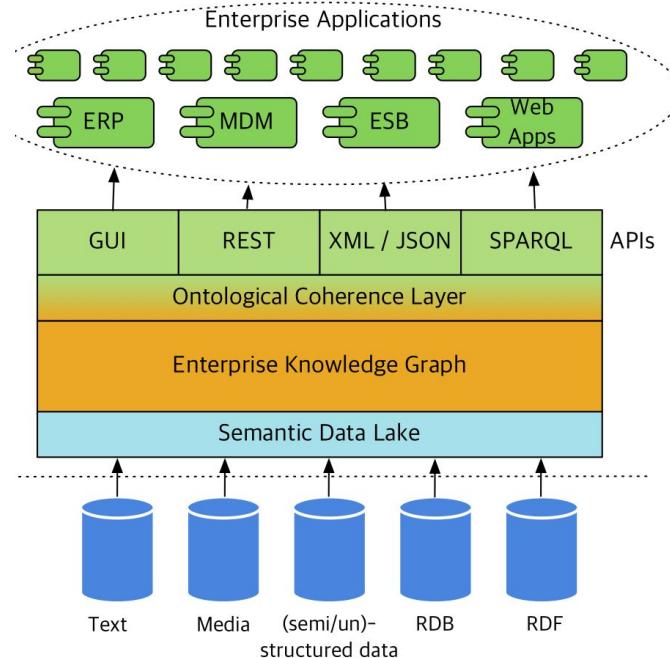
- ❖ Интеграция данных из гетерогенных источников в один физический или виртуальный граф знаний



Сценарии использования графов знаний

- ❖ Интеграция данных из гетерогенных источников в один физический или виртуальный граф знаний

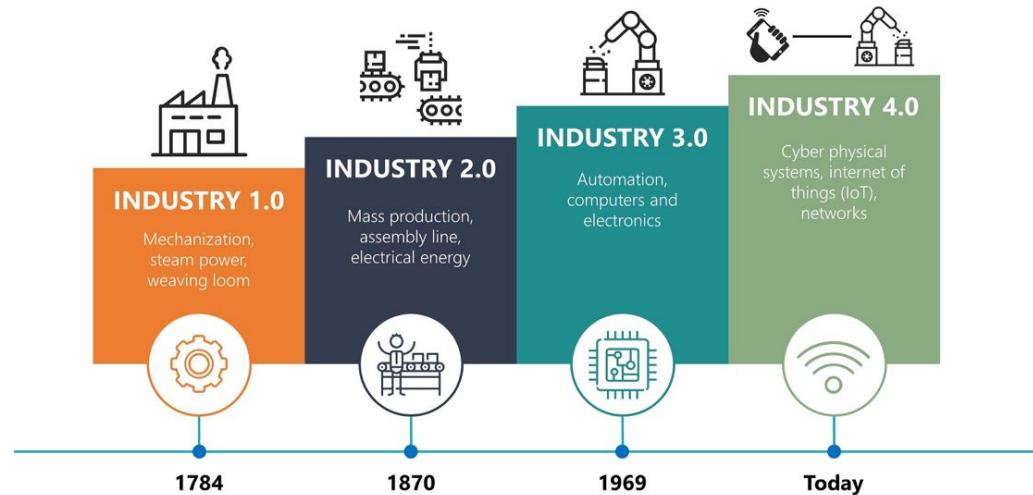
- Множество датасетов интегрируются с графом знаний
- Один запрос к графу пересыпается нужным источникам



Сценарии использования графов знаний

❖ Индустрия 4.0

- Кибер-физические системы (КФС)
- Взаимодействие КФС между собой без участия человека



Сценарии использования графов знаний

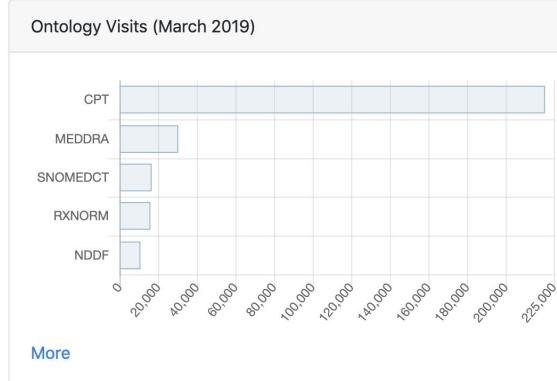
❖ Биотехнологии, медицина

➤ BioPortal - крупнейший репозиторий биомед-графов, > 140 миллиардов фактов

Metrics ?

Classes	347,358
Individuals	0
Properties	200
Maximum depth	28
Maximum number of children	2,355
Average number of children	5
Classes with a single child	36,680
Classes with more than 25 children	2,690
Classes with no definition	343,630

Visits ⬇️



BioPortal Statistics

Ontologies	768
Classes	9,410,439
Resources Indexed	48
Indexed Records	39,537,360
Direct Annotations	95,468,433,792
Direct Plus Expanded Annotations	144,789,582,932

Сценарии использования графов знаний

❖ Биотехнологии, медицина

- BioPortal - крупнейший репозиторий биомед-графов, > 140 миллиардов фактов
- Bio2RDF - интегрирует 35 датасетов в граф > 11 миллиардов фактов



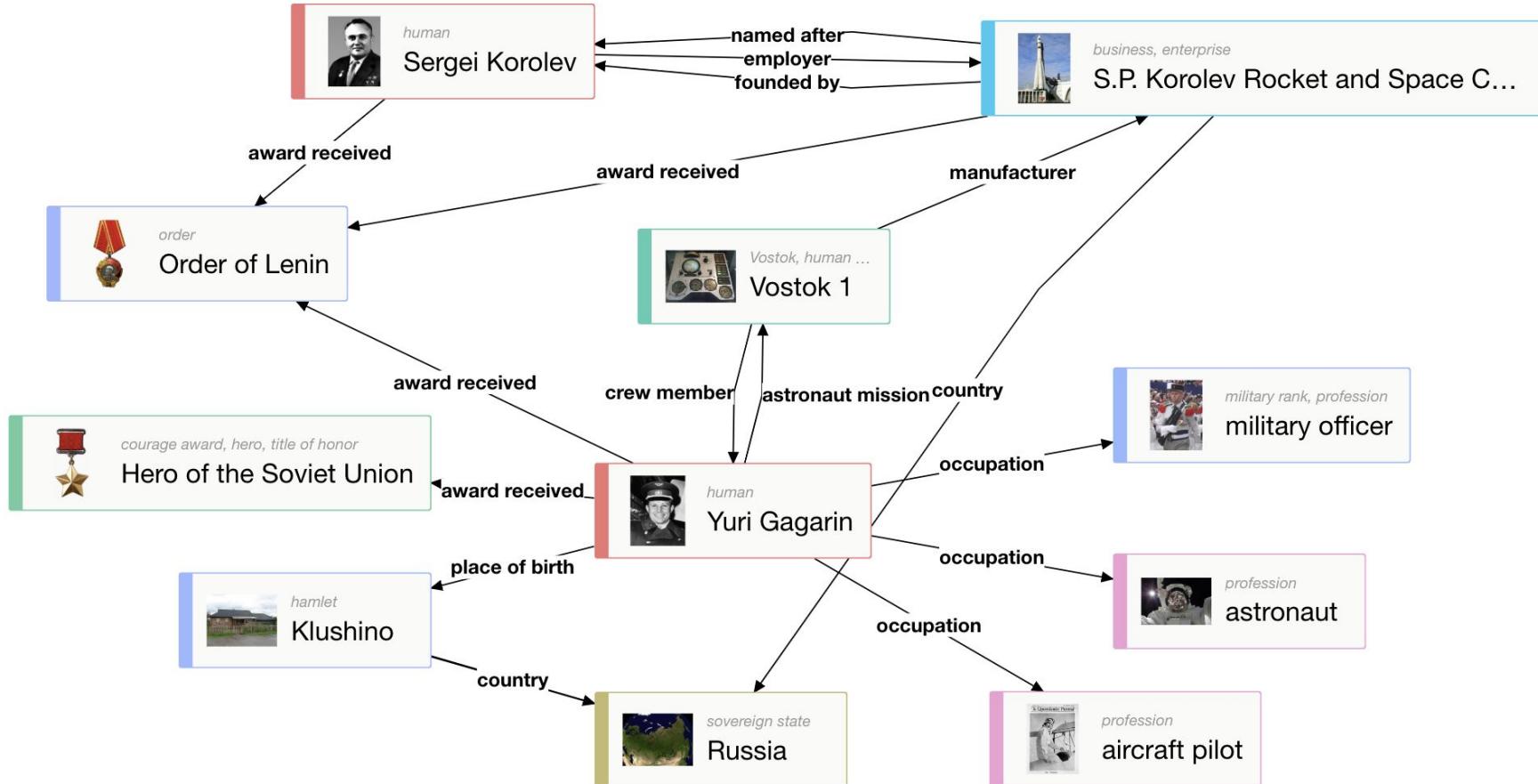
Сценарии использования графов знаний

❖ Биотехнологии, медицина

- BioPortal - крупнейший репозиторий биомед-графов, > 140 миллиардов фактов
- Bio2RDF - интегрирует 35 датасетов в граф > 11 миллиардов фактов
- PubMed - собрание аннотаций медицинских и биологических статей с 1951 года



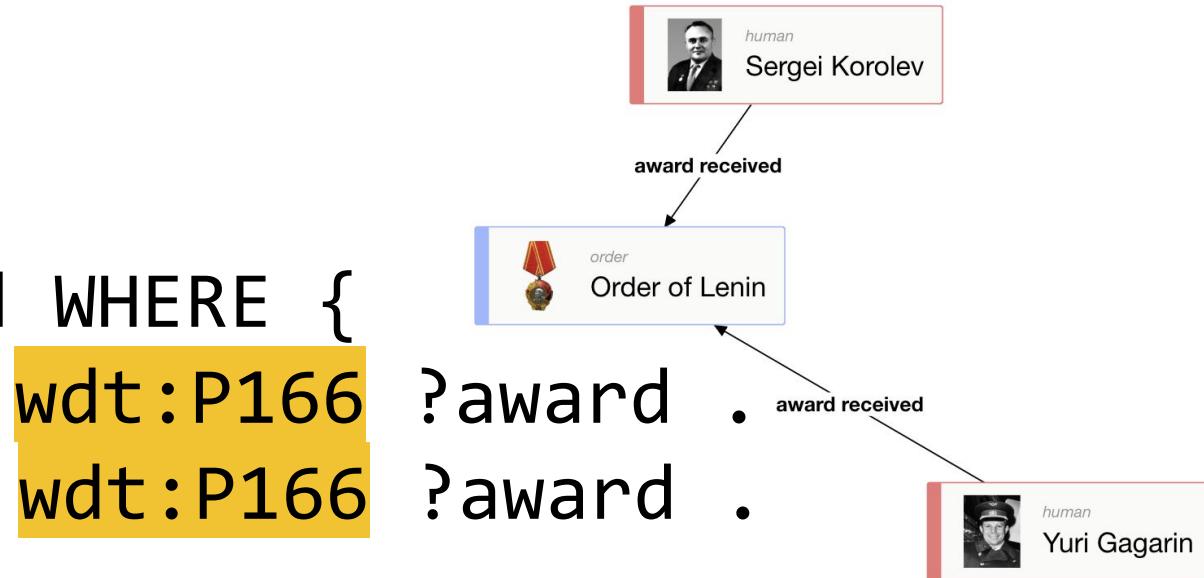
Открытые научные проблемы - NER & RL



Открытые научные проблемы - NER & RL

Обладателями каких наград являлись и
Юрий Гагарин, и Сергей Королев?

```
SELECT ?award WHERE {  
    wd:Q7327 wdt:P166 ?award .  
    wd:Q170842 wdt:P166 ?award .  
}
```



Named Entity Recognition & Linking

Задача - обнаружить части предложения, которые являются сущностями в имеющемся графе

Обладателями каких наград являлись и
Юрий Гагарин, и **Сергей Королев?**



wd:Q7327



wd:Q170842

Relation Linking

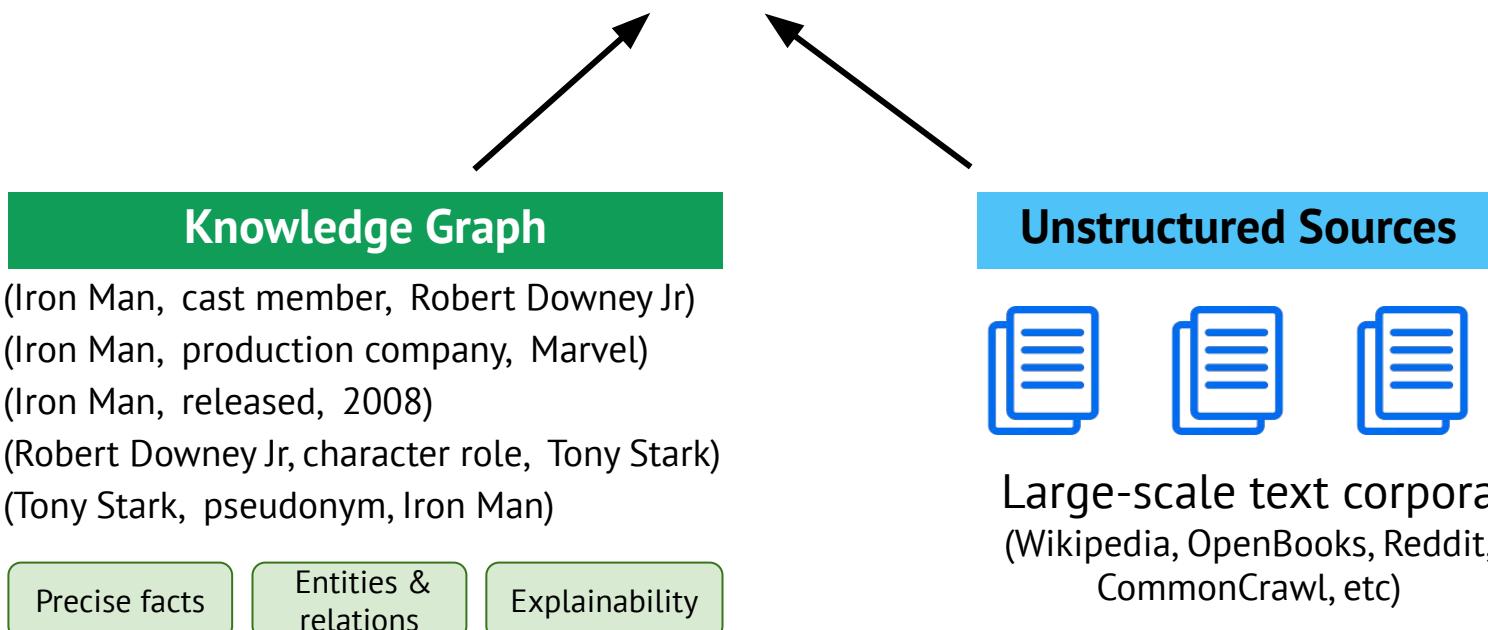
Задача - обнаружить части предложения, которые являются связями в имеющемся графе

Обладателями каких наград являлись и
Юрий Гагарин, и Сергей Королев?



Language Modeling

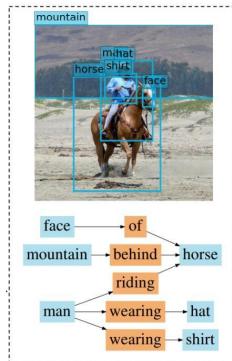
Robert Downey Jr. portrayed [MASK] in the Marvel movie in 2008.



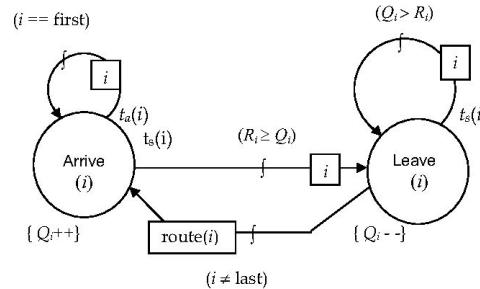
KGs in Graph ML



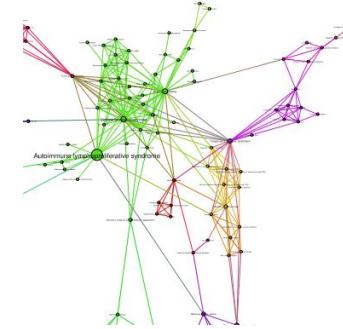
Cell similarity networks



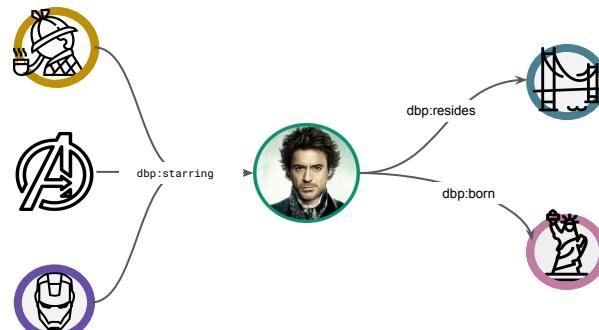
Scene Graphs



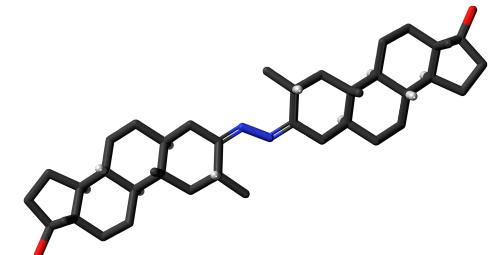
Event graphs



Disease pathways

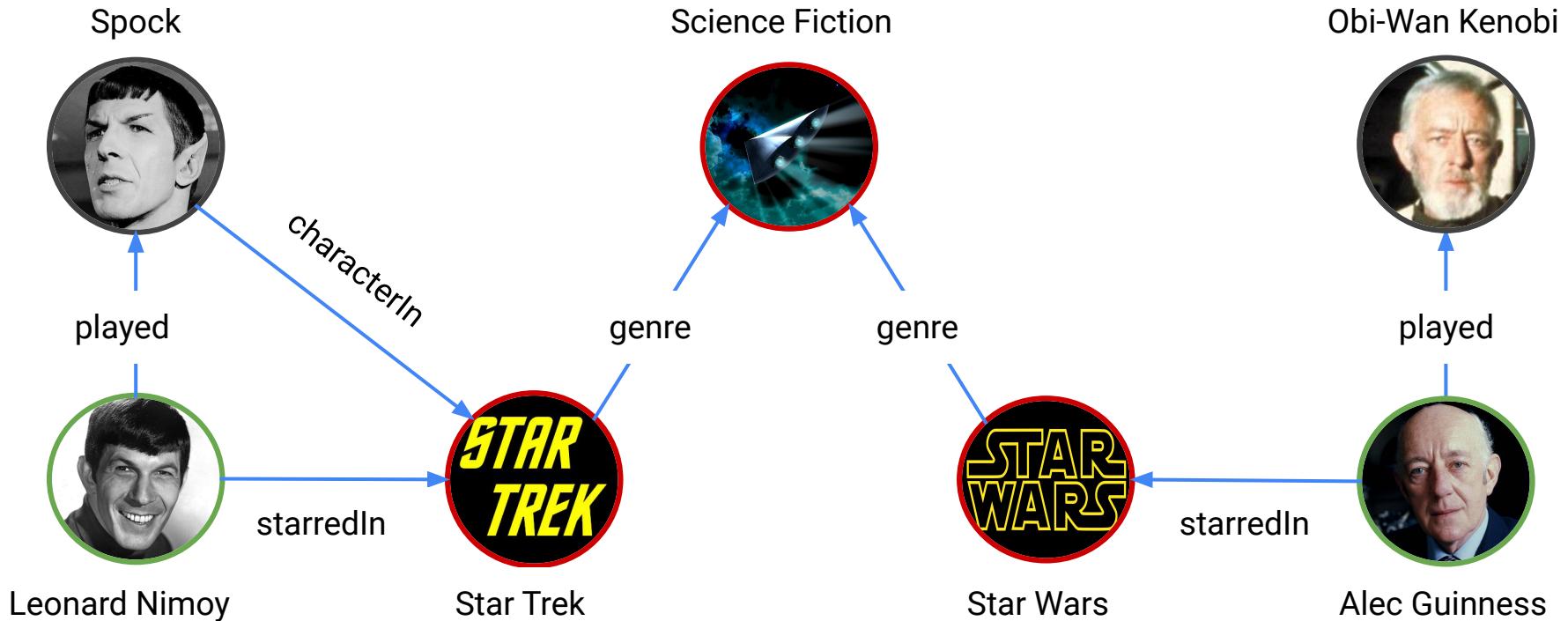


Knowledge Graphs

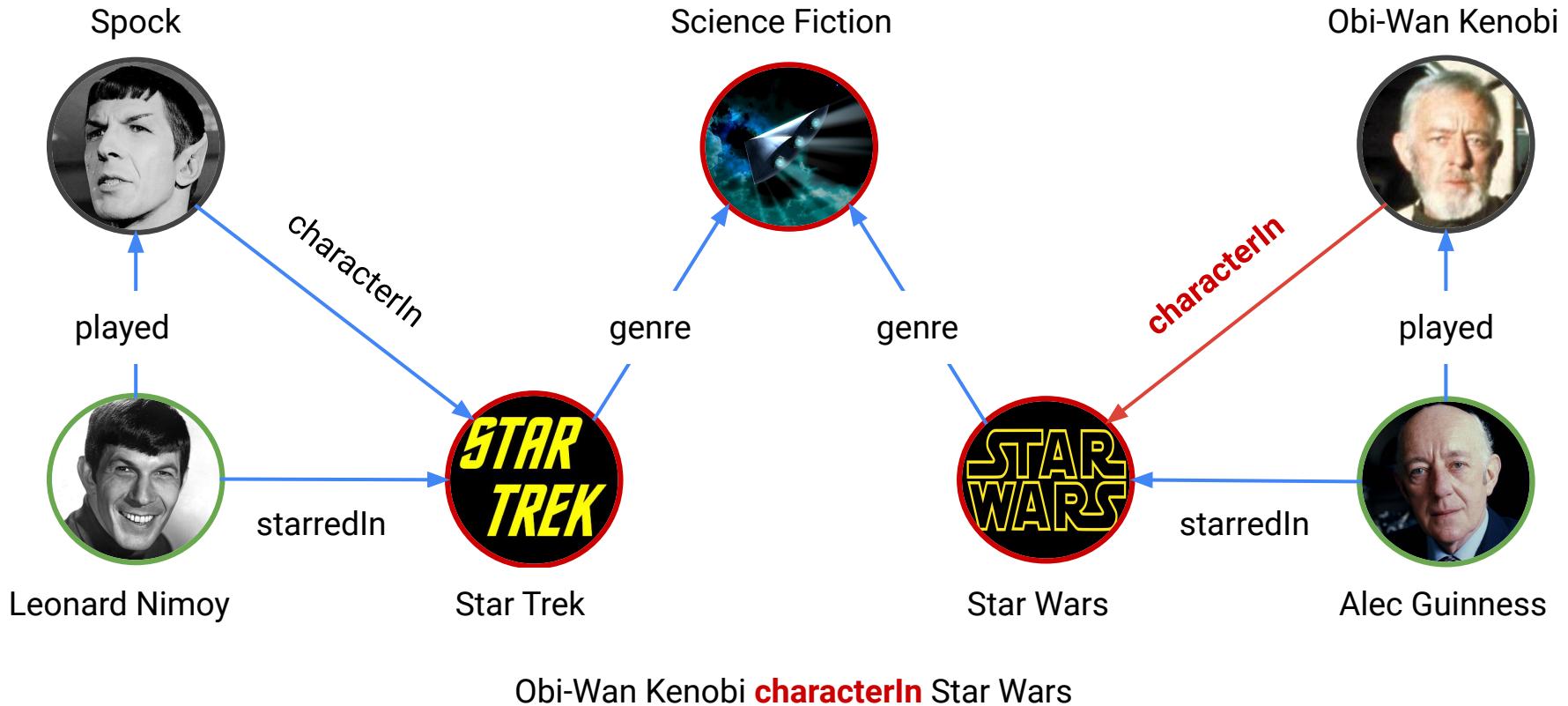


Molecules

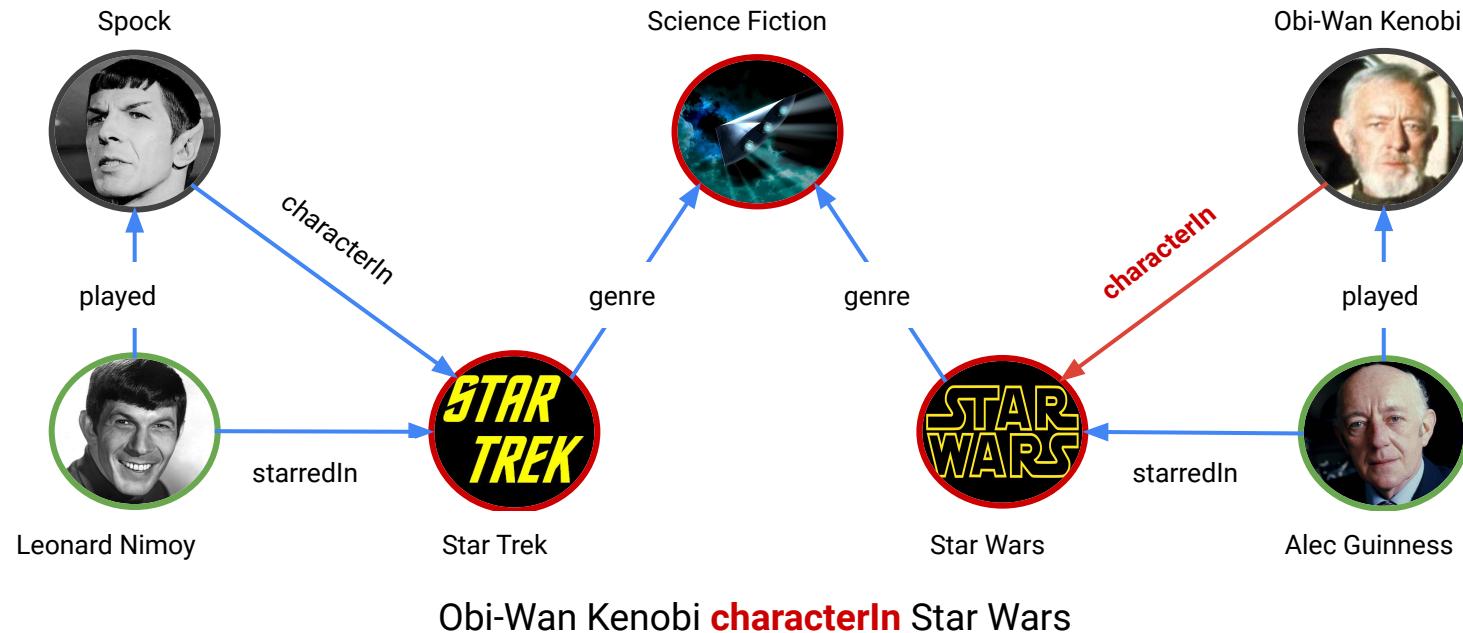
Открытые научные проблемы - Knowledge Graph Completion



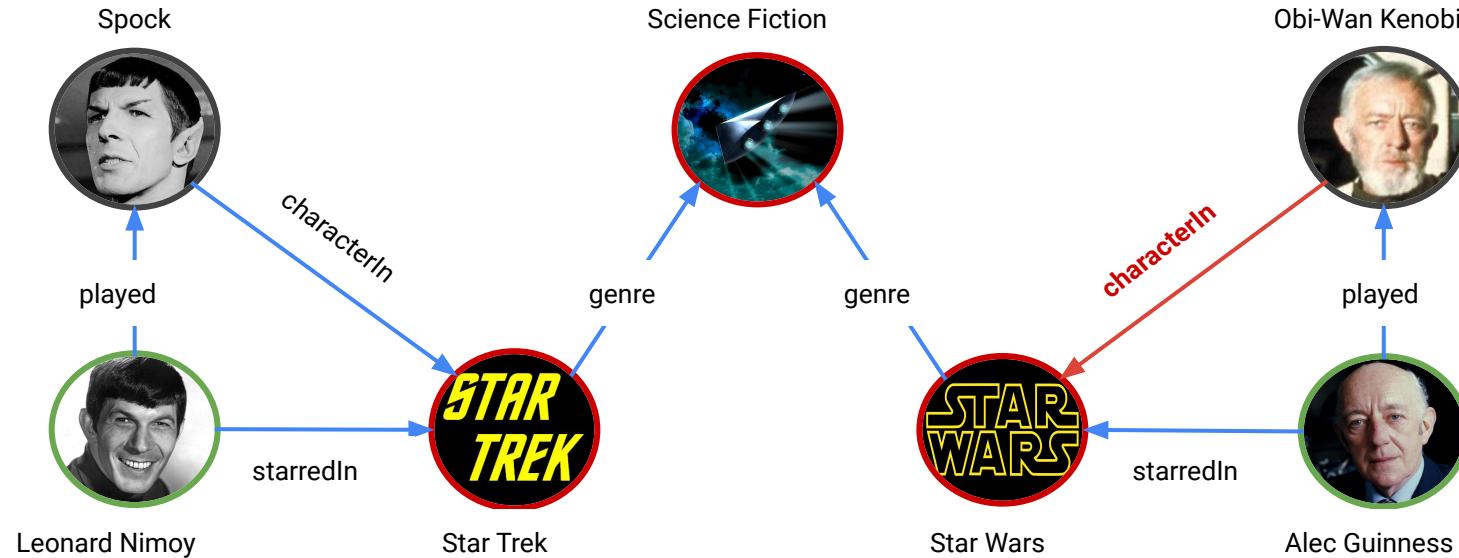
Открытые научные проблемы - Knowledge Graph Completion



Knowledge Graph Completion - Символический вывод


$$\text{starredIn}(X,Y) \wedge \text{played}(X,Z) \rightarrow \text{characterIn}(Z,Y)$$
$$\text{starredIn}(\text{Alec Guinness}, \text{Star Wars}) \wedge \text{played}(\text{Alec Guinness}, \text{Obi-Wan Kenobi}) \rightarrow \text{characterIn}(\text{Obi-Wan Kenobi}, \text{Star Wars})$$

Knowledge Graph Completion - Статистический вывод



Spock = [0.1, 0.2, 0.3]

Leonard Nimoy = [0.4, 0.8, 0.1]

Star Trek = [0.22, 0.34, 0.87]

characterIn = [0.1, 0.1, 0.6]

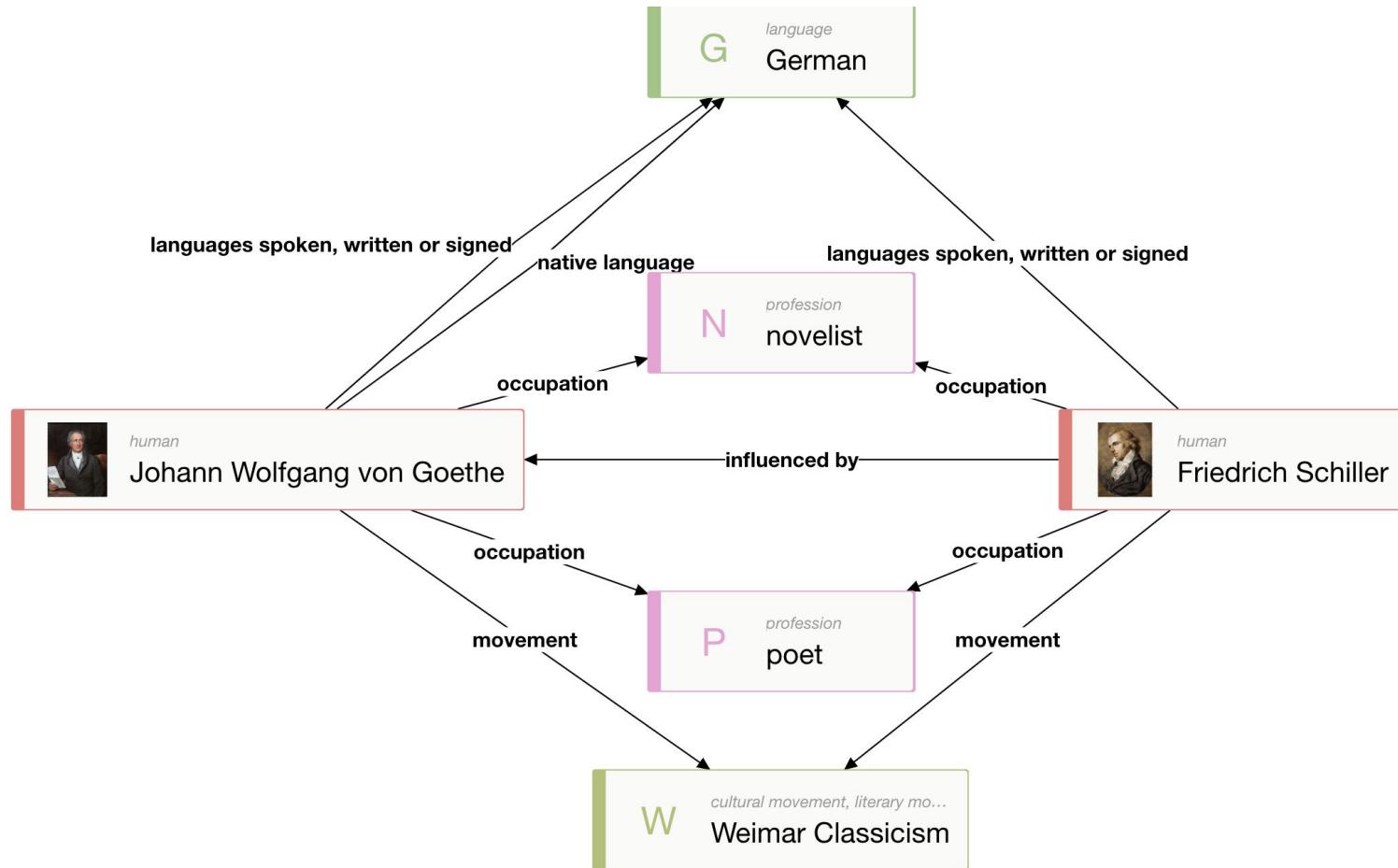
Obi-Wan = [0.05, 0.25, 0.37]

Alec Guinness = [0.33, 0.5, 0.3]

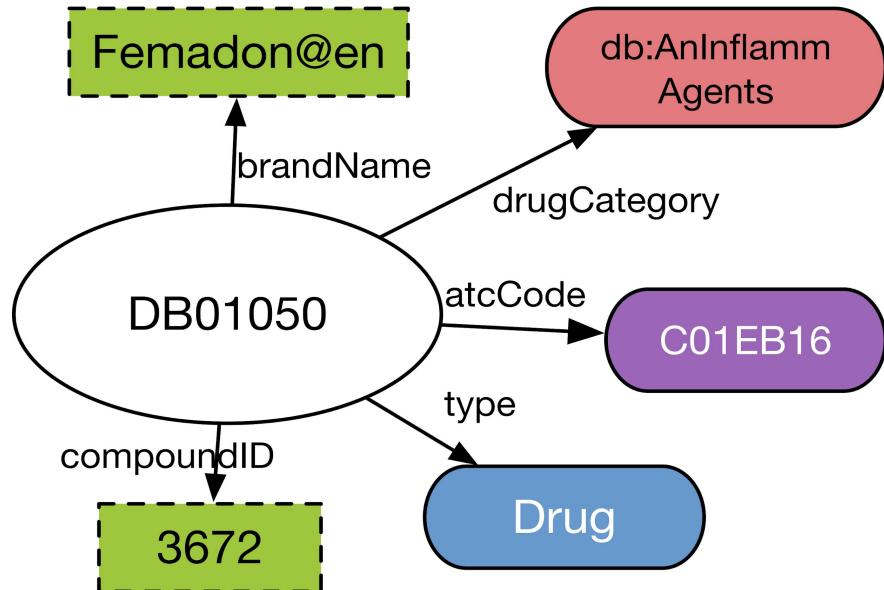
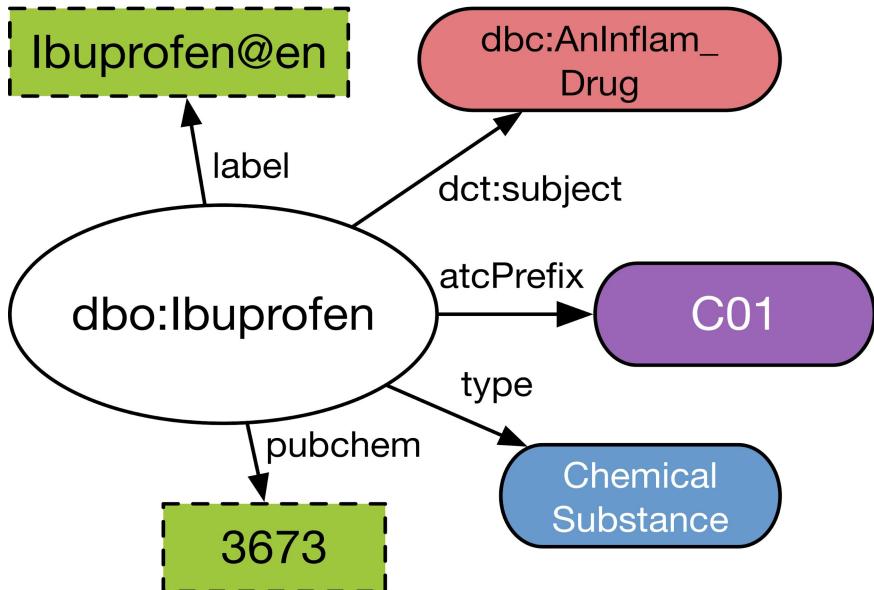
Star Wars = [0.18, 0.4, 0.9]

[0.13, 0.15, 0.53]

Открытые научные проблемы - Семантическая близость



Открытые научные проблемы - Семантическая близость



Открытые научные проблемы - Проверка утверждений

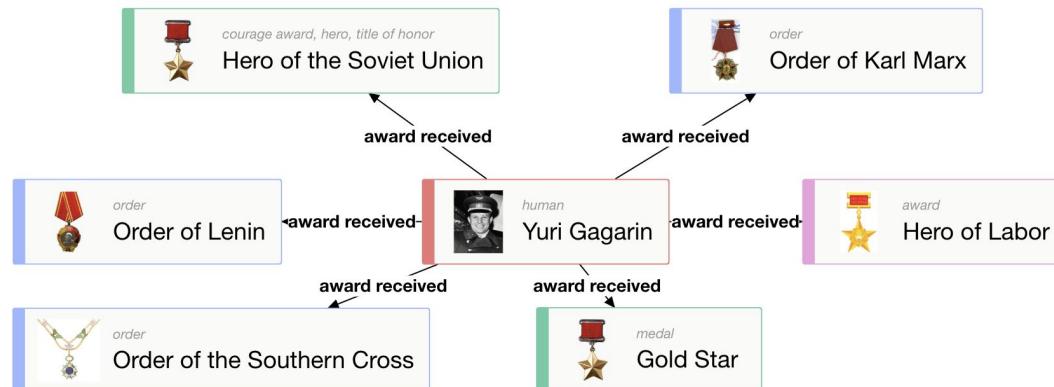
Задача - обнаружить факты в графе, подтверждающие или опровергающие данное утверждение

Юрий Гагарин обладал множеством
государственных наград

Открытые научные проблемы - Проверка утверждений

Задача - обнаружить факты в графе, подтверждающие или опровергающие данное утверждение

Юрий Гагарин обладал множеством государственных наград



wd:Q7327

Конференции

- The Web Conf (WWW)
- International Semantic Web Conference (ISWC)
- Neural Information Processing Systems (NeurIPS)
- International Conference on Learning Representations (ICLR)
- AAAI Conference on Artificial Intelligence (AAAI)
- International Conference on Machine Learning (ICML)
- International Joint Conference on Artificial Intelligence (IJCAI)
- Association for Computational Linguistics (ACL/ NAACL)
- Empirical Methods in NLP (EMNLP)

- Enterprise Data World
- Knowledge Graph Conference

В следующей серии

1. Introduction
2. **Представление знаний в графах - RDF & RDFS & OWL**
3. Хранение знаний в графах - SPARQL & Graph Databases
4. Однородность знаний - Reification & RDF* & SHACL & ShEx
5. Интеграция данных в графы знаний - Semantic Data Integration
6. Векторные представления графов - Knowledge Graph Embeddings
7. Введение в теорию графов - Graph Theory Intro
8. Машинное обучение на графах - Graph Neural Networks & KGs
9. Некоторые применения - Question Answering & Query Embedding

Демо

<https://wikidata.metaphacts.com>