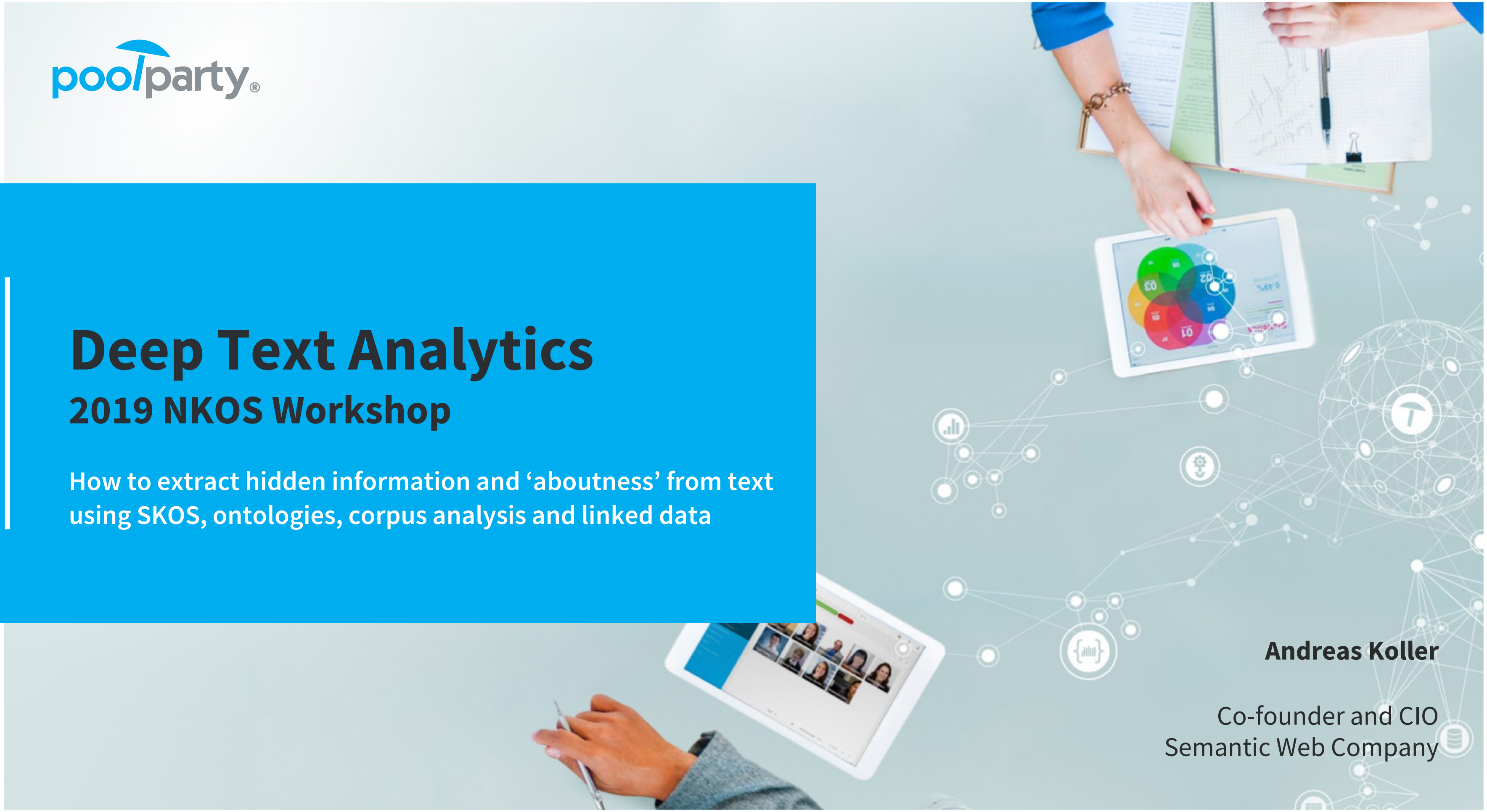
poolparty®



# Deep Text Analytics

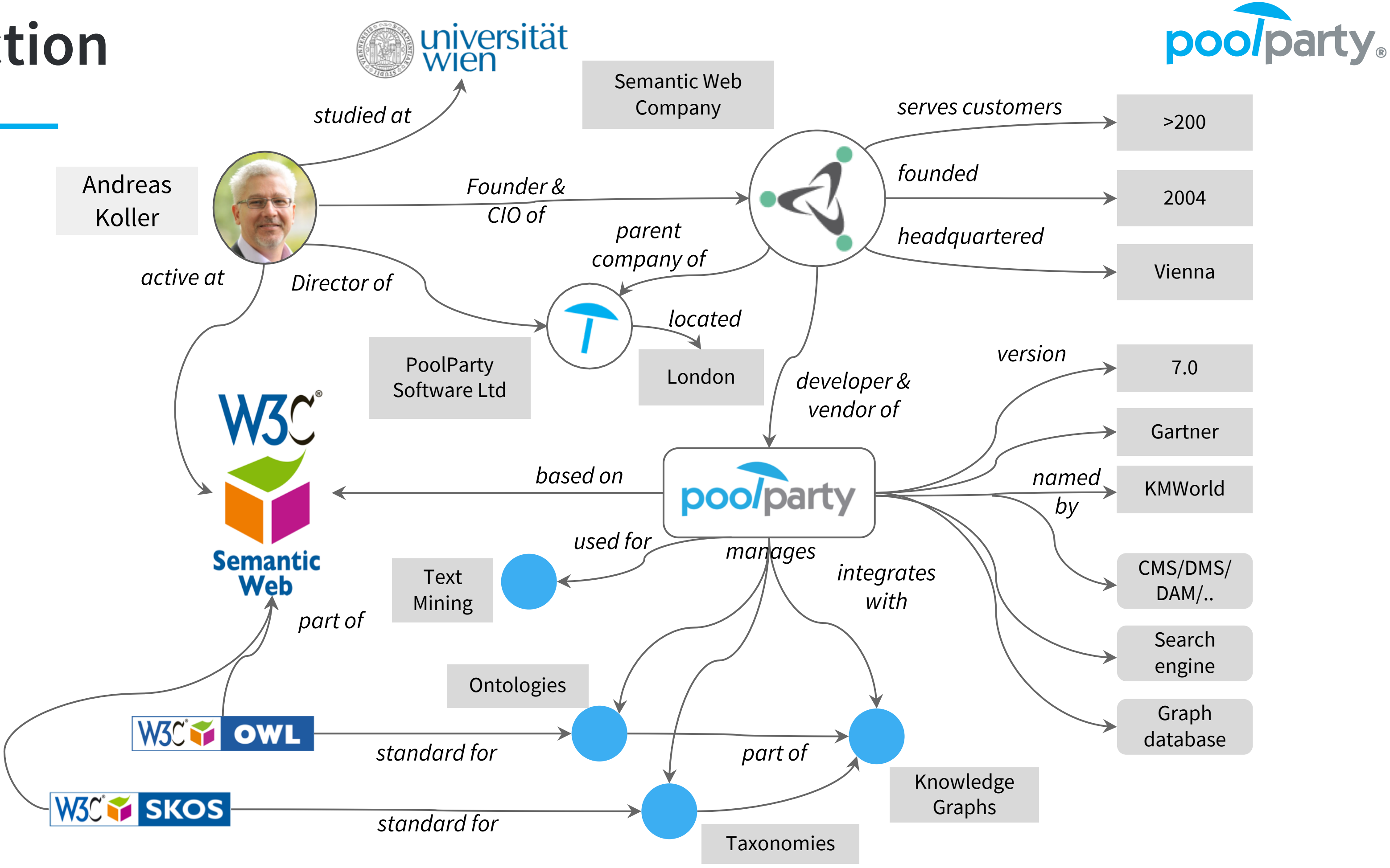
## 2019 NKOS Workshop

How to extract hidden information and ‘aboutness’ from text using SKOS, ontologies, corpus analysis and linked data

**Andreas Koller**

Co-founder and CIO
Semantic Web Company

# Introduction



# At a glance

## PoolParty Semantic Suite

- PoolParty is the most complete and most secure **semantic middleware** on the global market.
- PoolParty integrates with leading **graph database** technologies, NoSQL stores, and search engines.
- PoolParty was launched in **2009**.

## Semantic Web Company

- SWC is **developer, licensor and vendor** of PoolParty Semantic Suite.
- Established in 2004, SWC has pioneered the areas of **Enterprise Semantics & Semantic AI**.
- SWC is dedicated to use **W3C standards** like SKOS, OWL, SPARQL, etc.

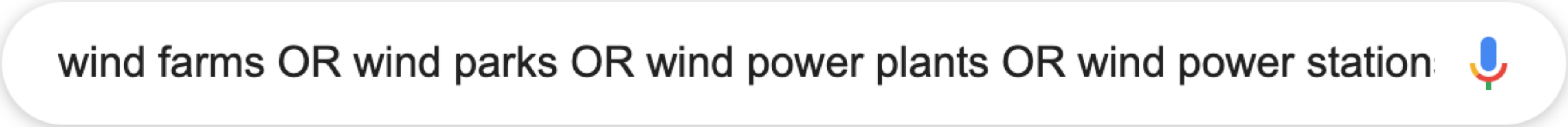
## Why Knowledge Graphs?

# WHAT’S THE PROBLEM?

Many Challenges in (Enterprise) Information Management & AI

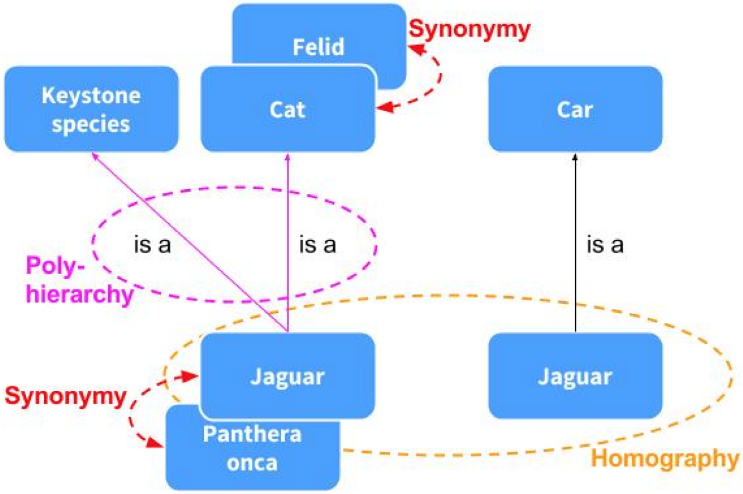
# “Search” is still about documents only

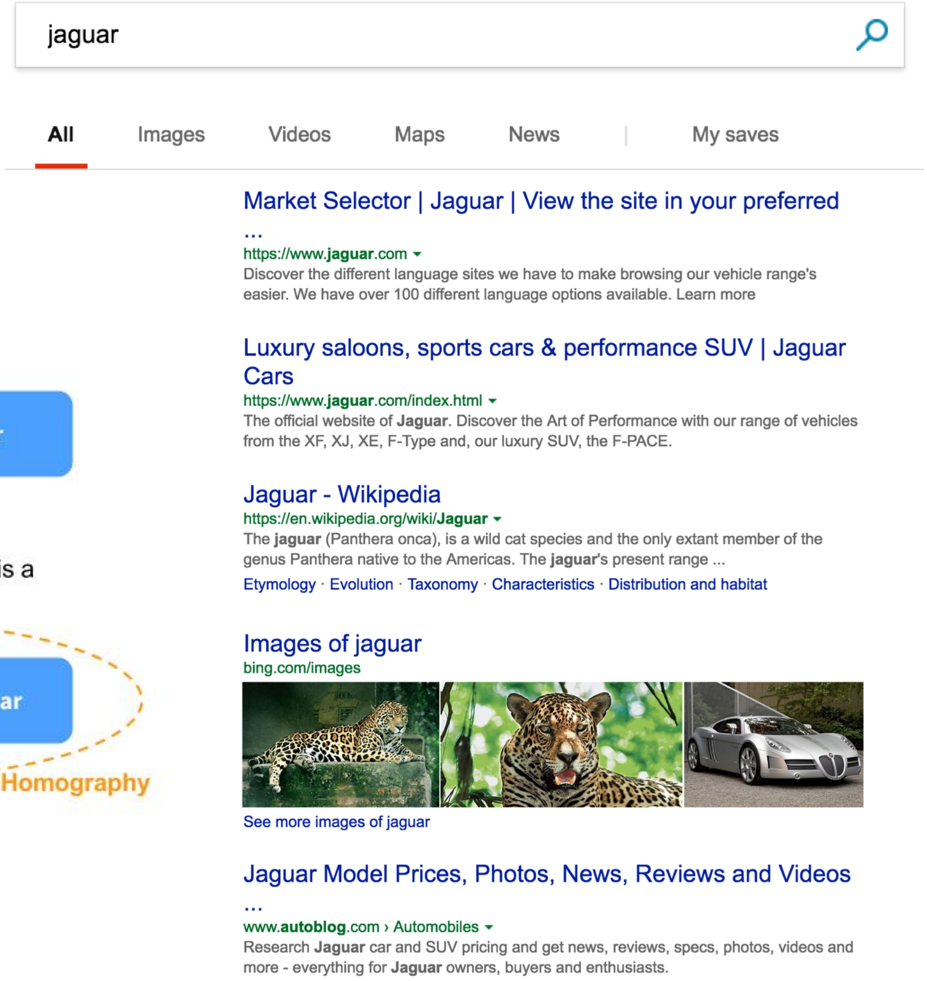
… and in enterprises it’s painful



wind farms OR wind parks OR wind power plants OR wind power station

# Missing Context





jaguar

[Market Selector | Jaguar | View the site in your preferred ...]
https://www.jaguar.com
Discover the different language sites we have to make browsing our vehicle range's easier. We have over 100 different language options available. Learn more

[Luxury saloons, sports cars & performance SUV | Jaguar Cars]
https://www.jaguar.com/index.html
The official website of **Jaguar**. Discover the Art of Performance with our range of vehicles from the XF, XJ, XE, F-Type and, our luxury SUV, the F-PACE.

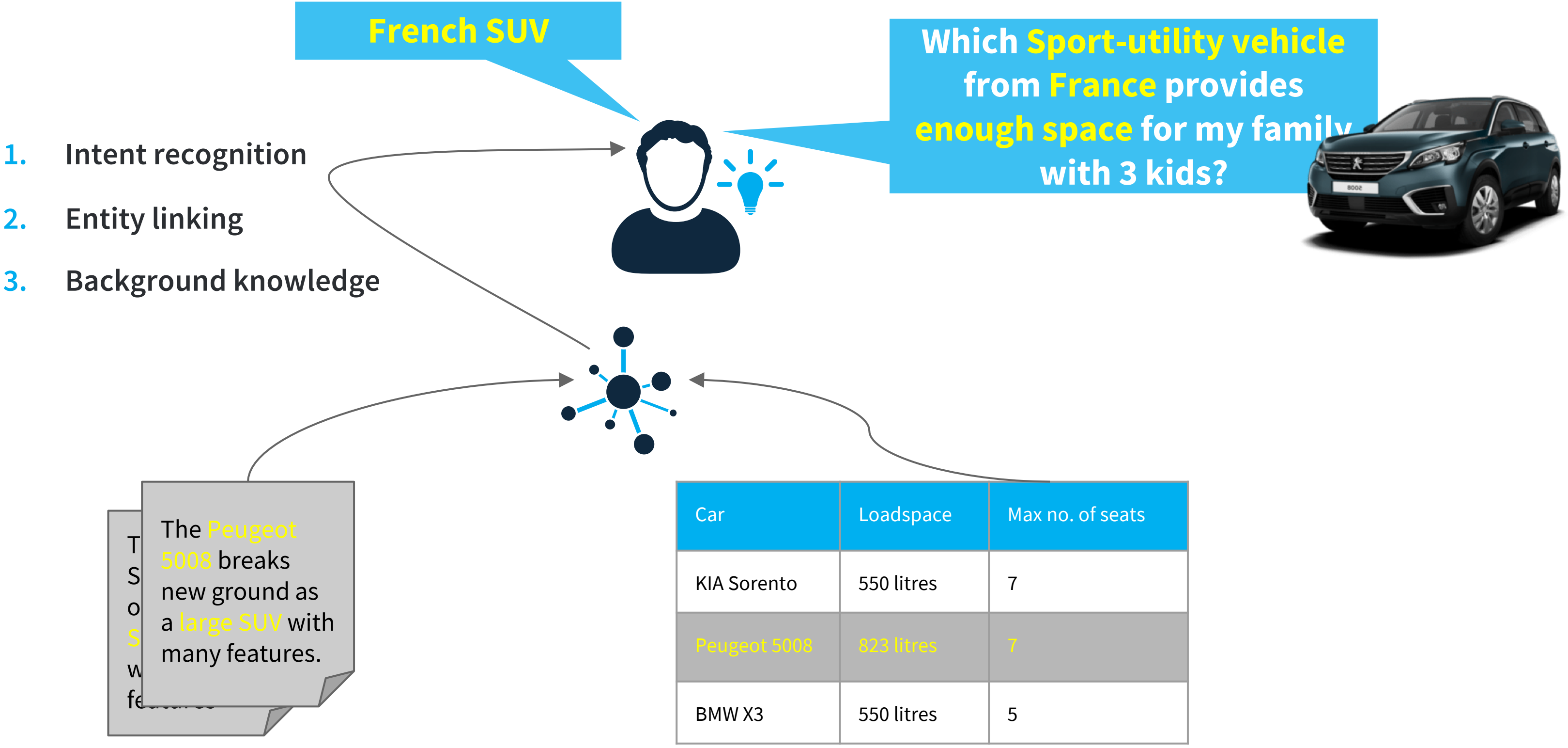
[Jaguar - Wikipedia]
https://en.wikipedia.org/wiki/**Jaguar**
The **jaguar** (Panthera onca), is a wild cat species and the only extant member of the genus Panthera native to the Americas. The **jaguar**'s present range ...
Etymology · Evolution · Taxonomy · Characteristics · Distribution and habitat

[Images of jaguar]
bing.com/images
See more images of jaguar

[Jaguar Model Prices, Photos, News, Reviews and Videos ...]
www.**autoblog**.com › Automobiles
Research **Jaguar** car and SUV pricing and get news, reviews, specs, photos, videos and more - everything for **Jaguar** owners, buyers and enthusiasts.

# Context-free data models and User-agnostic

Do machines understand user intent? Do they have enough context?



French SUV

Which Sport-utility vehicle from France provides enough space for my family with 3 kids?

1. Intent recognition
2. Entity linking
3. Background knowledge

The Peugeot 5008 breaks new ground as a large SUV with many features.

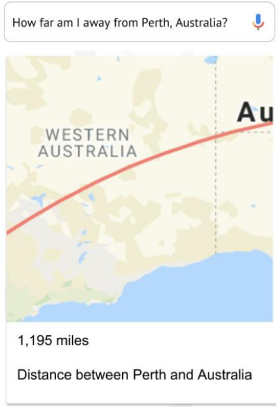
| Car | Loadspace | Max no. of seats |
|---|---|---|
| KIA Sorento | 550 litres | 7 |
| Peugeot 5008 | 823 litres | 7 |
| BMW X3 | 550 litres | 5 |

# Background knowledge (context) is key

## Support complex Q&A:

Which cities located in the Commonwealth of Nations have a population of more than 2 mio. people?

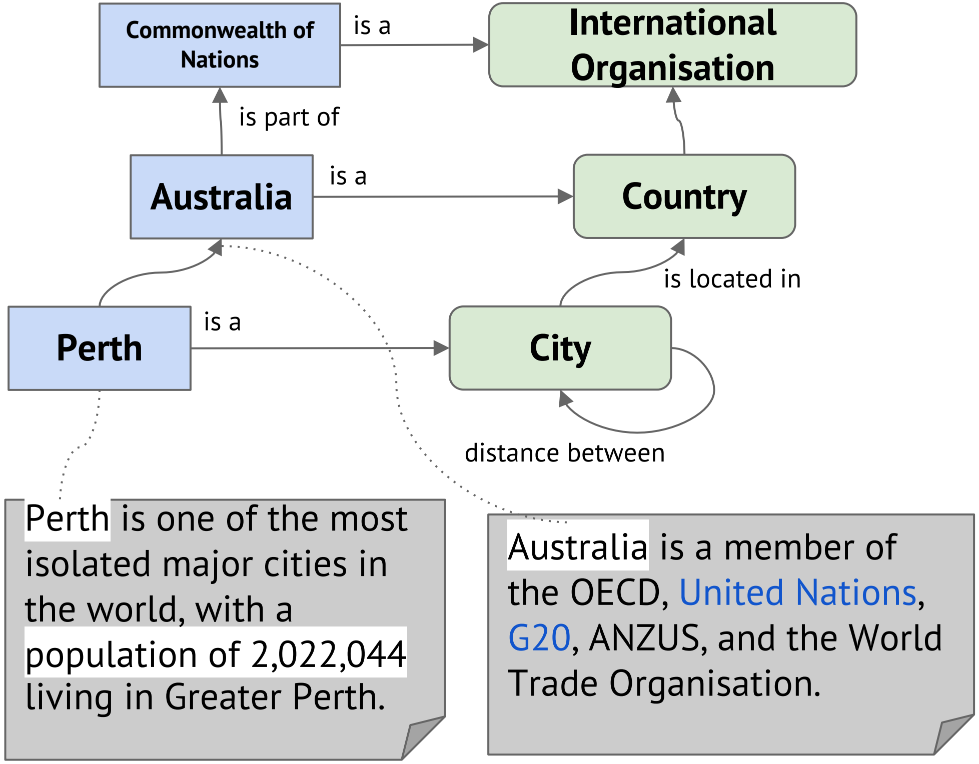
## Avoid illogical answers:



How far am I away from Perth, Australia?

1,195 miles

Distance between Perth and Australia



Perth is one of the most isolated major cities in the world, with a population of 2,022,044 living in Greater Perth.

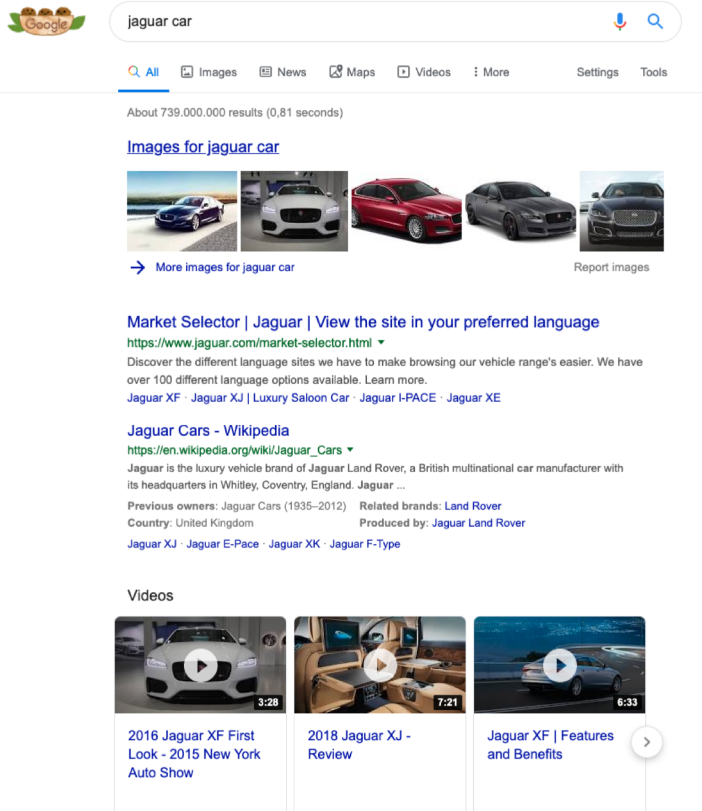
Australia is a member of the OECD, United Nations, G20, ANZUS, and the World Trade Organisation.

# How does it work?

# KNOWLEDGE GRAPHS

Bring your metadata to the next level!

# Google Knowledge Graph



jaguar car

About 739.000.000 results (0,81 seconds)

Images for jaguar car

More images for jaguar car

Market Selector | Jaguar | View the site in your preferred language
https://www.jaguar.com/market-selector.html
Discover the different language sites we have to make browsing our vehicle range's easier. We have over 100 different language options available. Learn more.
Jaguar XF · Jaguar XJ | Luxury Saloon Car · Jaguar I-PACE · Jaguar XE

Jaguar Cars - Wikipedia
https://en.wikipedia.org/wiki/Jaguar_Cars
**Jaguar** is the luxury vehicle brand of **Jaguar** Land Rover, a British multinational **car** manufacturer with its headquarters in Whitley, Coventry, England. **Jaguar** ...

Previous owners: Jaguar Cars (1935–2012)
Country: United Kingdom
Related brands: Land Rover
Produced by: Jaguar Land Rover

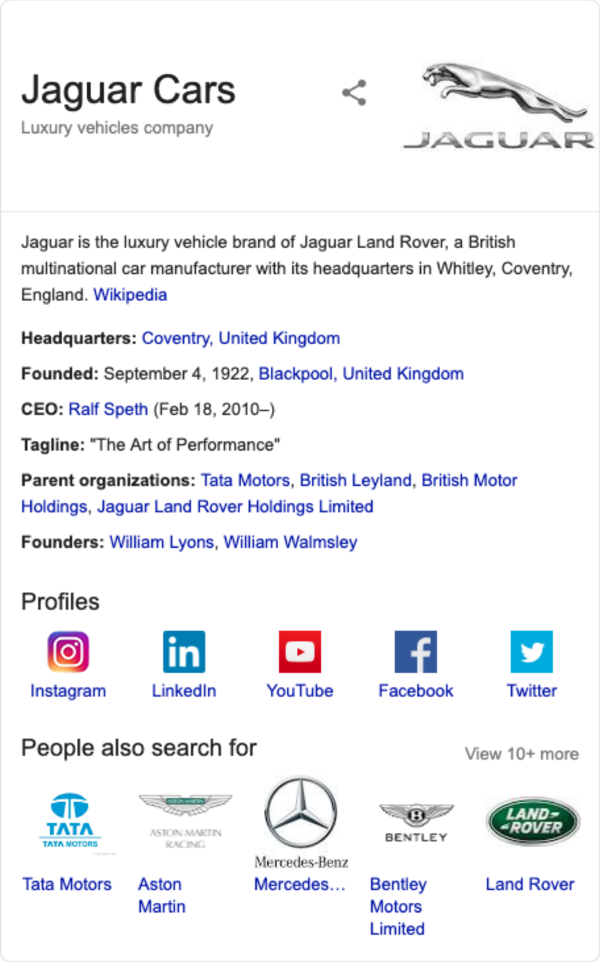
Jaguar XJ · Jaguar E-Pace · Jaguar XK · Jaguar F-Type

Videos

2016 Jaguar XF First Look - 2015 New York Auto Show

2018 Jaguar XJ - Review

Jaguar XF | Features and Benefits



Jaguar Cars
Luxury vehicles company

Jaguar is the luxury vehicle brand of Jaguar Land Rover, a British multinational car manufacturer with its headquarters in Whitley, Coventry, England. Wikipedia

**Headquarters:** Coventry, United Kingdom

**Founded:** September 4, 1922, Blackpool, United Kingdom

**CEO:** Ralf Speth (Feb 18, 2010–)

**Tagline:** "The Art of Performance"

**Parent organizations:** Tata Motors, British Leyland, British Motor Holdings, Jaguar Land Rover Holdings Limited

**Founders:** William Lyons, William Walmsley

Profiles

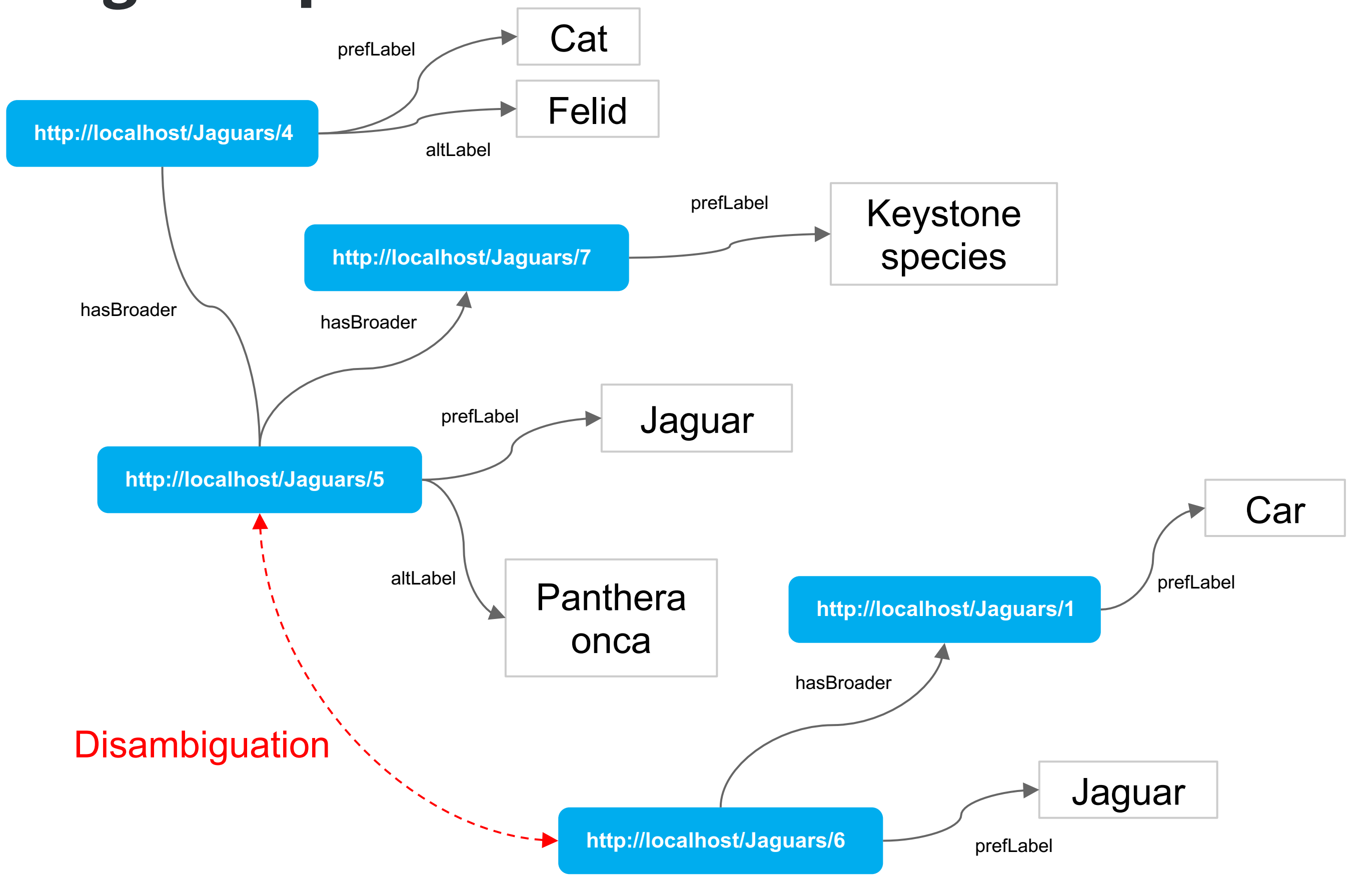
Instagram · LinkedIn · YouTube · Facebook · Twitter

People also search for — View 10+ more

Tata Motors · Aston Martin · Mercedes... · Bentley Motors Limited · Land Rover

# Example Knowledge Graph

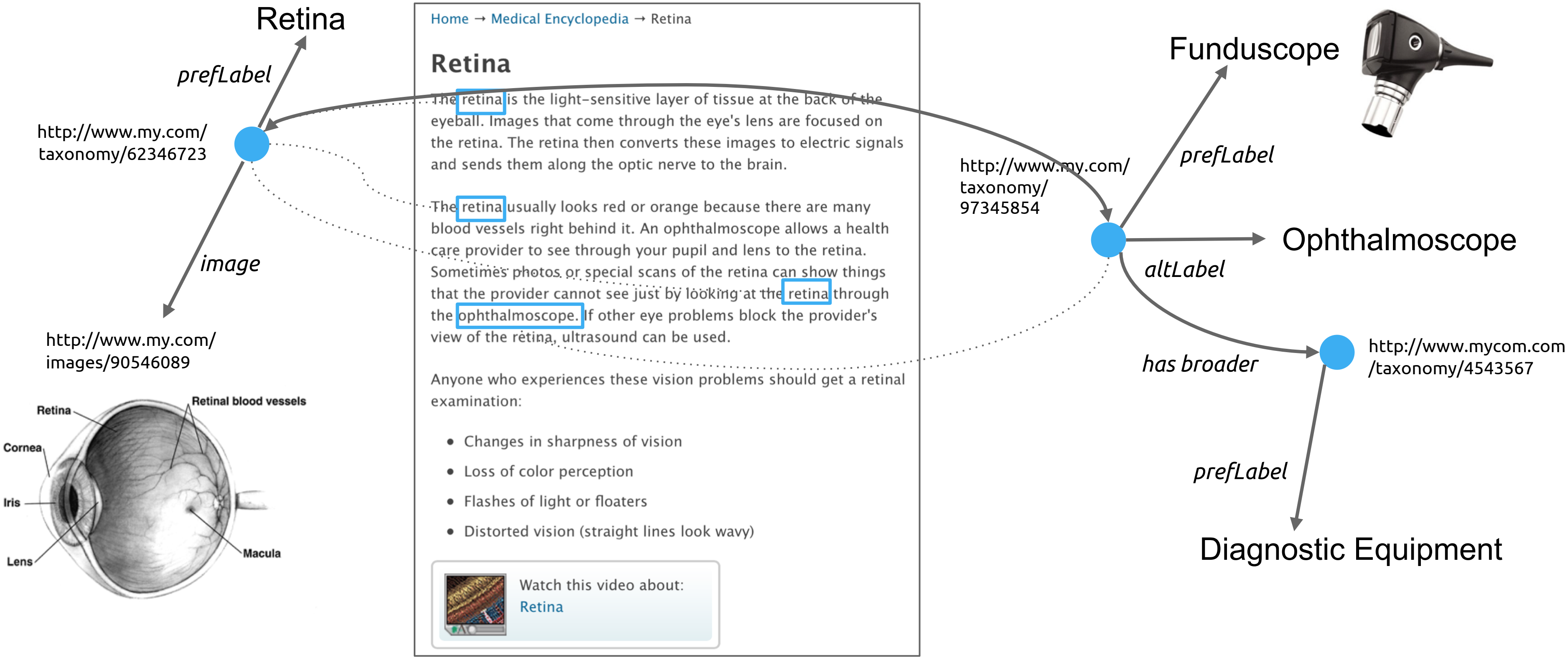
- URIs
- Concepts
- Labels
- Relations



Demo

# ‘Things’ but no strings:

Using a ‘Semantic Knowledge Graph’



Home → Medical Encyclopedia → Retina

## Retina

The retina is the light-sensitive layer of tissue at the back of the eyeball. Images that come through the eye's lens are focused on the retina. The retina then converts these images to electric signals and sends them along the optic nerve to the brain.

The retina usually looks red or orange because there are many blood vessels right behind it. An ophthalmoscope allows a health care provider to see through your pupil and lens to the retina. Sometimes photos or special scans of the retina can show things that the provider cannot see just by looking at the retina through the ophthalmoscope. If other eye problems block the provider's view of the retina, ultrasound can be used.

Anyone who experiences these vision problems should get a retinal examination:

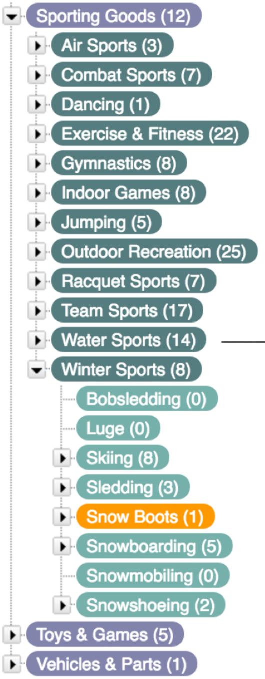
- Changes in sharpness of vision
- Loss of color perception
- Flashes of light or floaters
- Distorted vision (straight lines look wavy)

Watch this video about: Retina

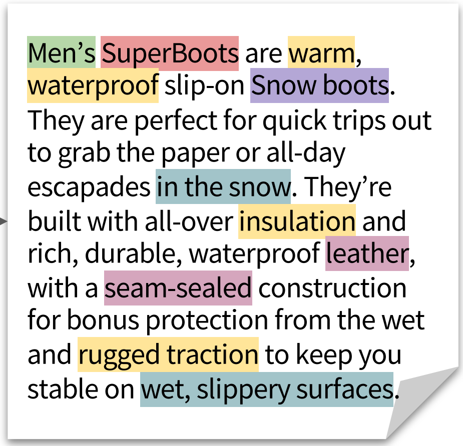
# Semantic Enrichment

- Graph-based annotation → **Entity linking**
- Machine-learning-based annotation → **Named entity recognition**
- Machine-learning based classification → **Document Classification**
- Annotation based on **Regular expressions**

## Taxonomy & Ontology Server



## Entity Extraction & Text Mining

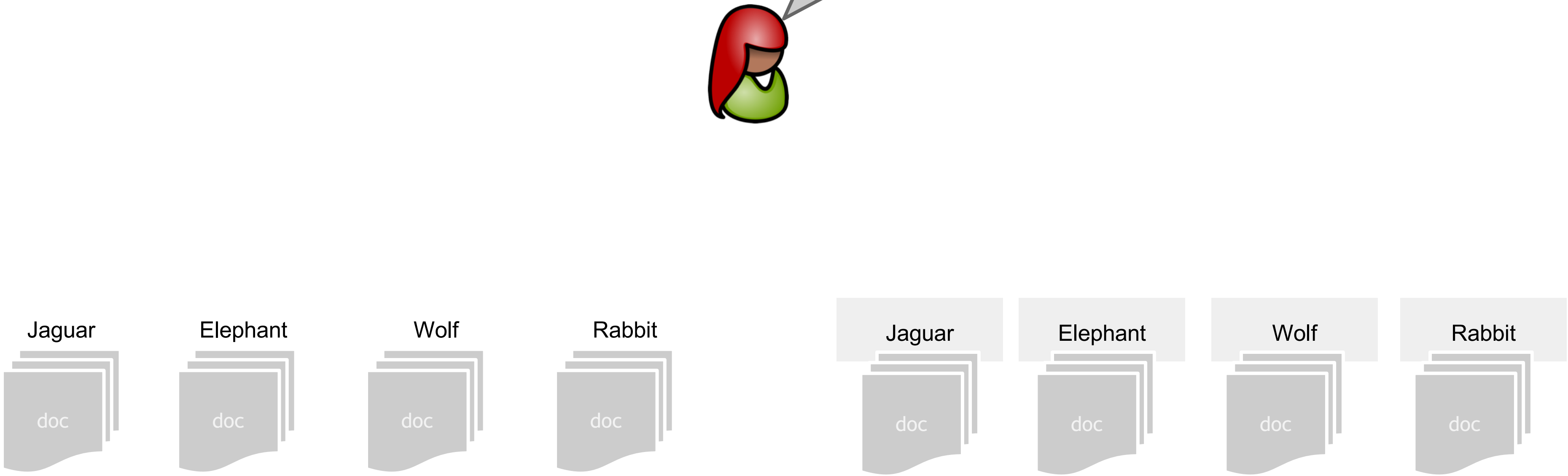

Men’s SuperBoots are warm, waterproof slip-on Snow boots. They are perfect for quick trips out to grab the paper or all-day escapades in the snow. They’re built with all-over insulation and rich, durable, waterproof leather, with a seam-sealed construction for bonus protection from the wet and rugged traction to keep you stable on wet, slippery surfaces.

# Traditional vs. Graph-based Metadata Management

Traditional approach

Show me all documents about ***Carnivores***

Graph-based approach



Jaguar

Elephant

Wolf

Rabbit

Jaguar

Elephant

Wolf

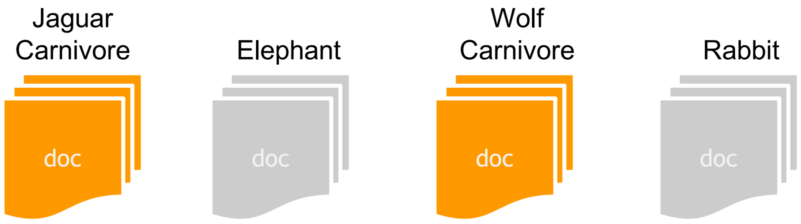
Rabbit

# Traditional vs. Graph-based Metadata Management

## Traditional approach

Show me all documents about ***Carnivores***

## Graph-based approach



Jaguar Carnivore

doc

Elephant

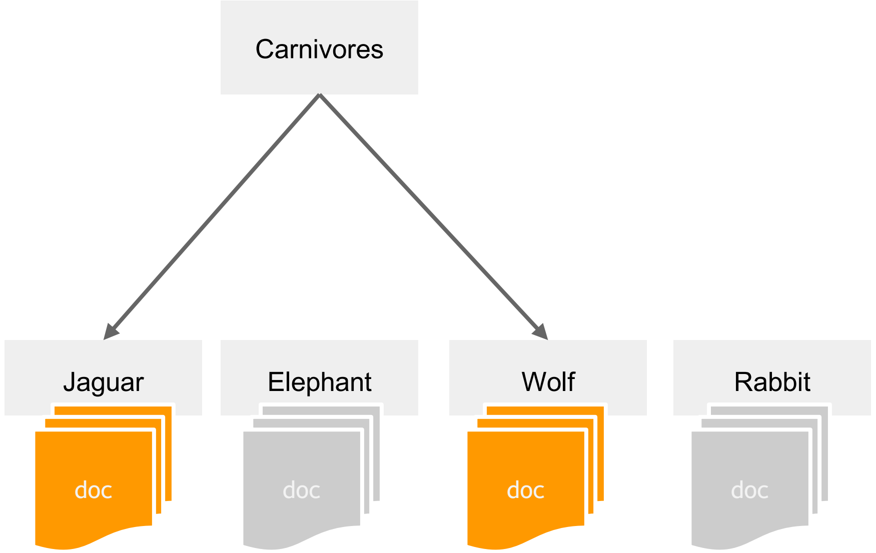
doc

Wolf Carnivore

doc

Rabbit

doc



Carnivores

Jaguar

doc

Elephant

doc

Wolf

doc

Rabbit

doc

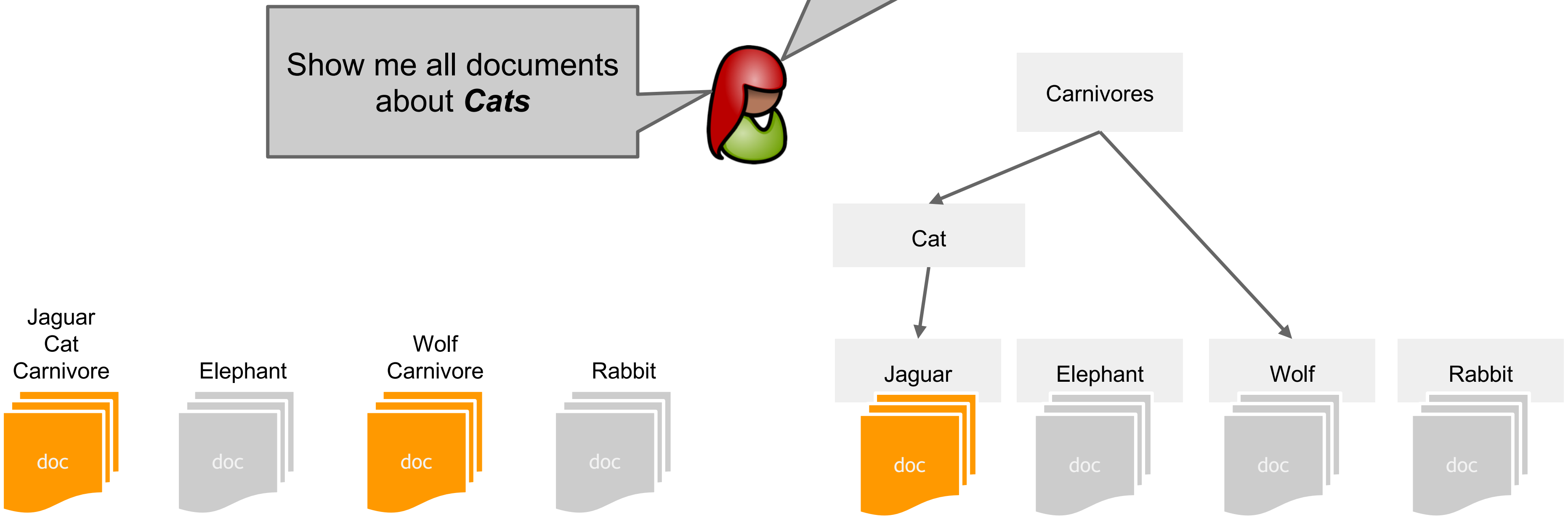
# Traditional vs. Graph-based Metadata Management

poolparty®

## Traditional approach

Show me all documents about ***Carnivores***

## Graph-based approach



Show me all documents about ***Cats***

Carnivores

Cat

Jaguar

Elephant

Wolf

Rabbit

doc

Jaguar
Cat
Carnivore

Elephant

Wolf
Carnivore

Rabbit

doc

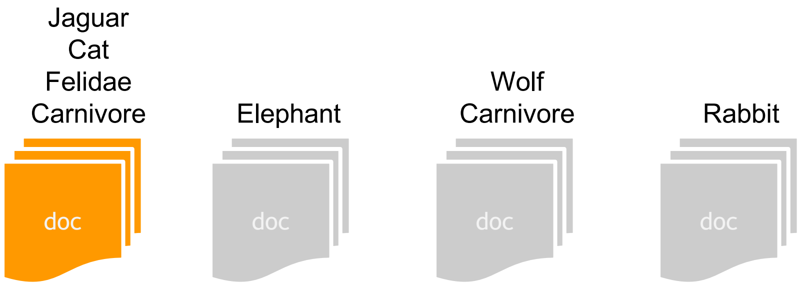
# Traditional vs. Graph-based Metadata Management

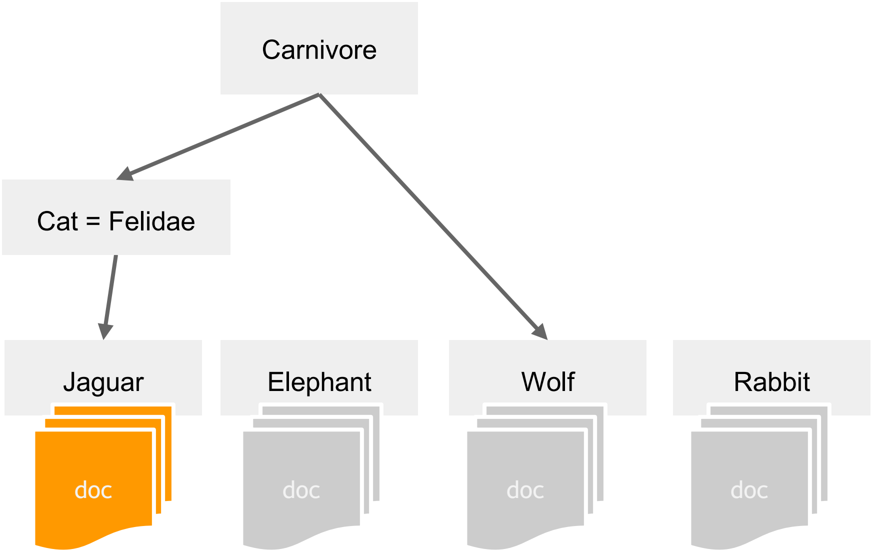
## Traditional approach

Show me all documents about ***Carnivores***

## Graph-based approach

Show me all documents about ***Felidae***

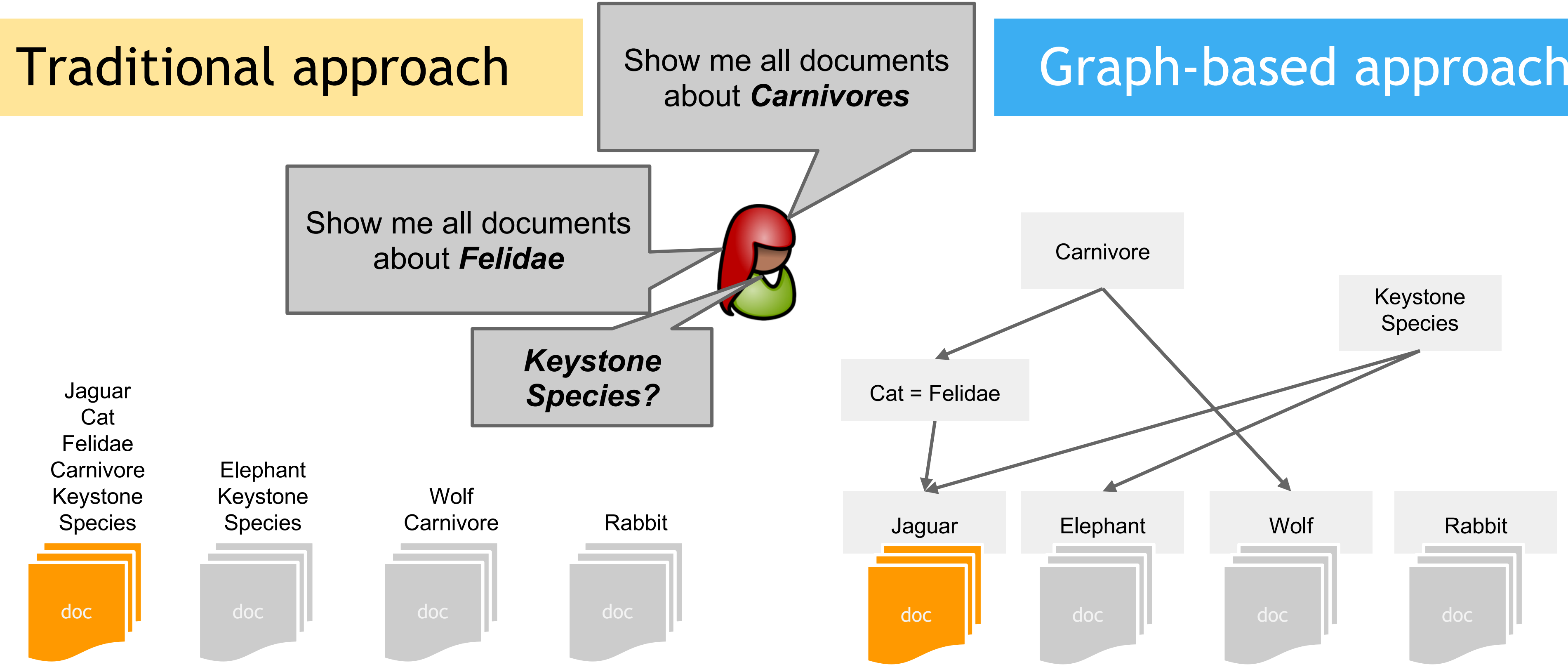




# Traditional vs. Graph-based Metadata Management

## Traditional approach

## Graph-based approach



Show me all documents about ***Carnivores***

Show me all documents about ***Felidae***

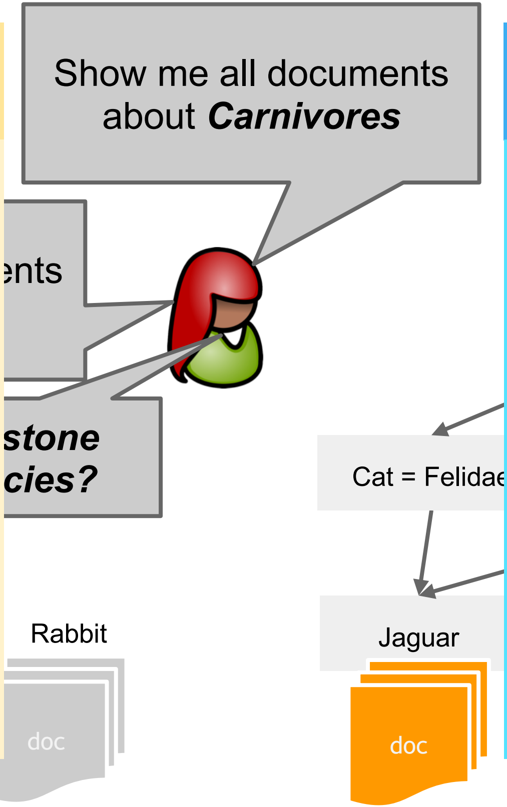
***Keystone Species?***

# Traditional vs. Graph-based Metadata Management

## Traditional approach

### Metadata per document

1. No or little network effects
2. No reuse of metadata
3. Metadata resides in silos
4. Data quality hard to measure
5. Not machine-readable

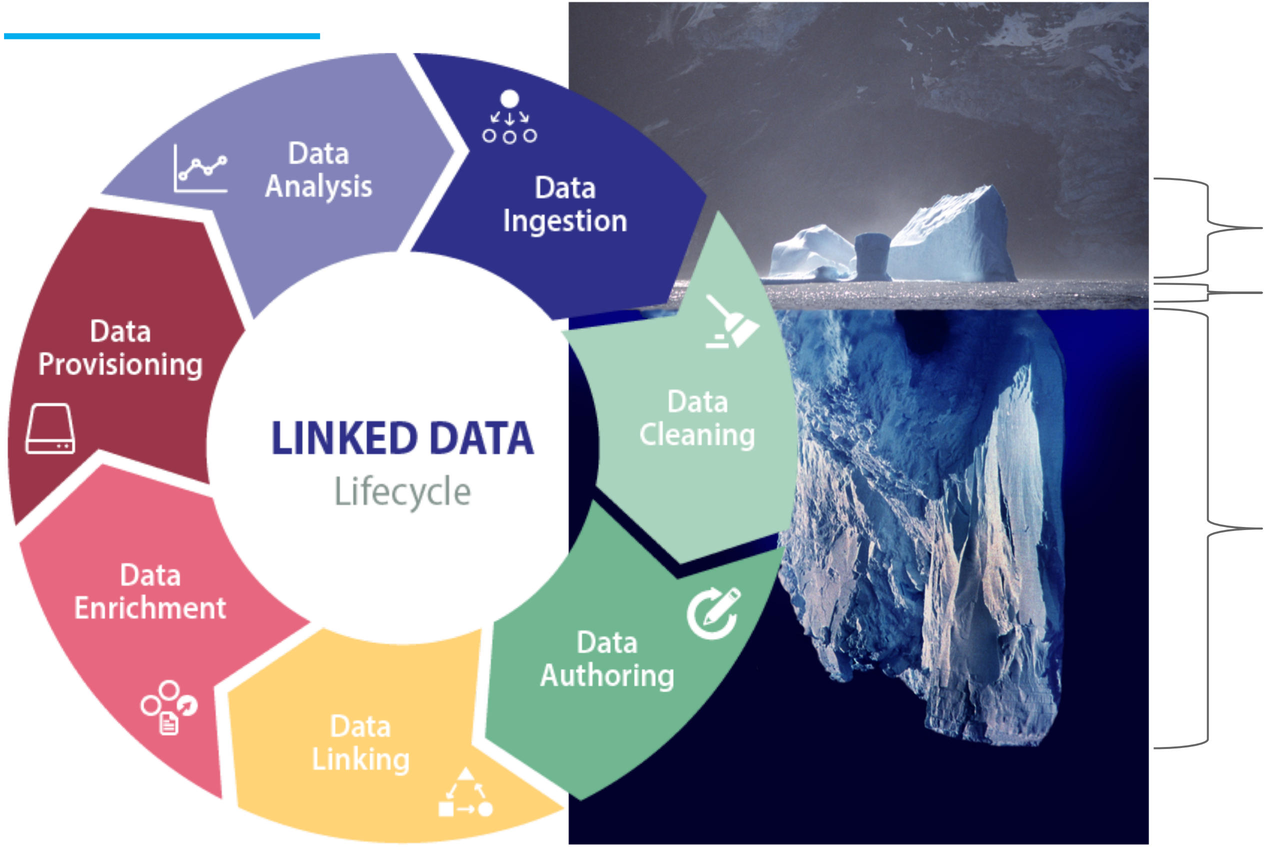


## Graph-based approach

### Knowledge about metadata

1. Explicit knowledge models
2. Reusable and measurable
3. Metadata is machine-processable
4. Standards-based metadata
5. Linkable metadata opens silos

# Knowledge Graphs as a result of human-machine cooperation



Manually created parts of graph

Supervised learning

Automatically created parts of graph (corpus analysis, RDF transformation, machine learning, ….)

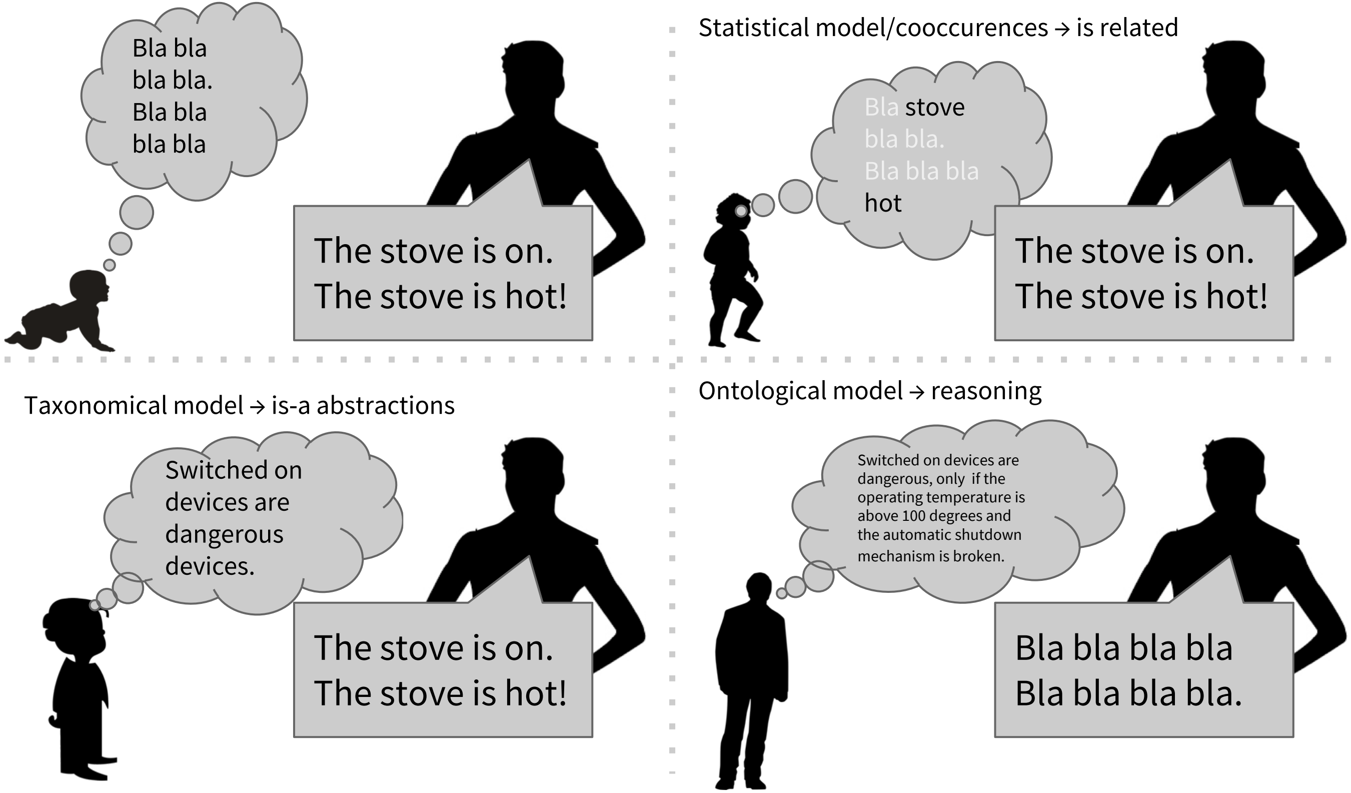
# How does it work?

# CORPUS ANALYSIS

Semantic Information Retrieval

# Bionics

## How do we learn from a lot of text?



Bla bla
bla bla.
Bla bla
bla bla

The stove is on.
The stove is hot!

Statistical model/cooccurences → is related

Bla stove
bla bla.
Bla bla bla
hot

The stove is on.
The stove is hot!

Taxonomical model → is-a abstractions

Switched on
devices are
dangerous
devices.

The stove is on.
The stove is hot!

Ontological model → reasoning

Switched on devices are dangerous, only if the operating temperature is above 100 degrees and the automatic shutdown mechanism is broken.

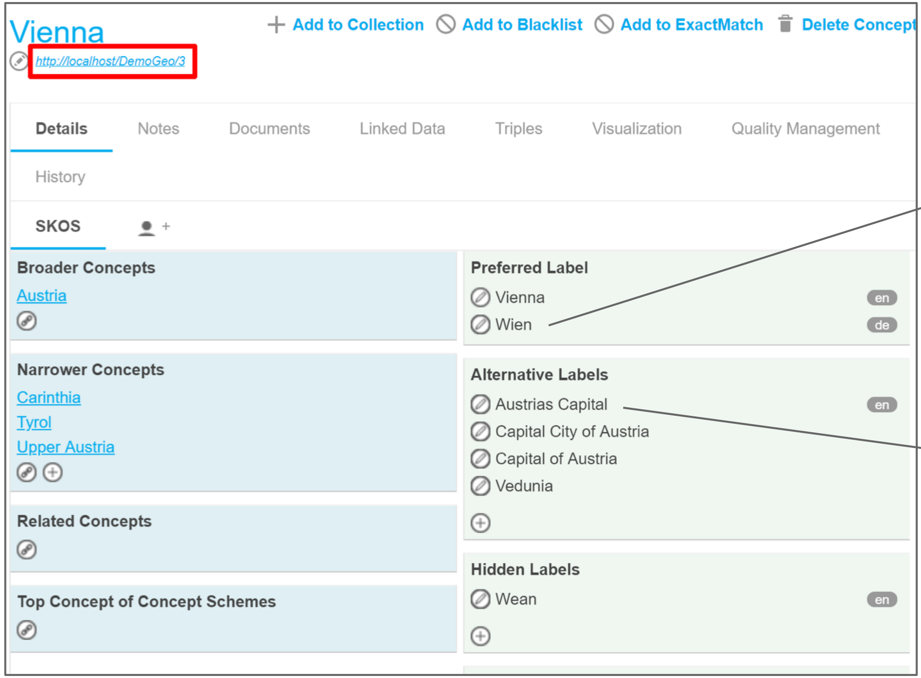
Bla bla bla bla
Bla bla bla bla.

# Extraction of the non-obvious

Complexity

| Mechanism | How does it work? |
|---|---|
| Entity extraction based on concepts instead of simple term-based extraction | Make use of synonyms in a taxonomy; disambiguation when necessary |
| Extraction of super classes and broader concepts | Make use of hierarchical structures in a knowledge graph |
| Extraction of related terms and concepts | Co-occurrence models and word embeddings |
| Extraction of 'Shadow Concepts' | Combining co-occurrence models and knowledge graphs |
| Semantic Classifier | Enrichment of training documents with metadata from a knowledge graph |
| Deep Text Analytics | Derivation of new metadata and classifications based on mechanisms as described above combined by rules |

# Entity extraction based on concepts instead of simple term-based extraction



Vienna

http://localhost/DemoGeo/3

Broader Concepts
Austria

Narrower Concepts
Carinthia
Tyrol
Upper Austria

Related Concepts

Top Concept of Concept Schemes

Preferred Label
Vienna en
Wien de

Alternative Labels
Austrias Capital en
Capital City of Austria
Capital of Austria
Vedunia

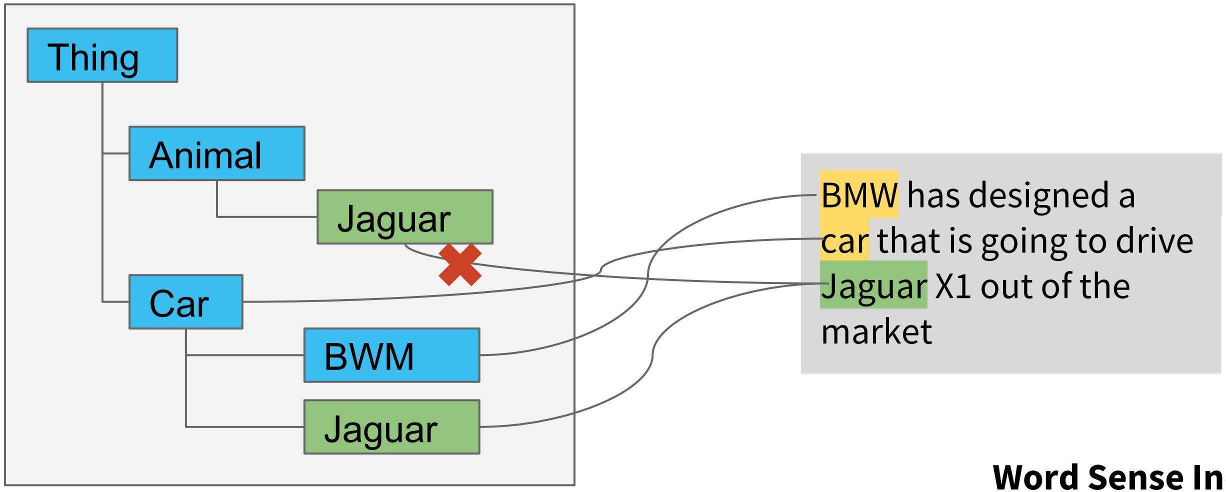
Hidden Labels
Wean en

Schaut man vom Kahlenberg auf die Donau hinunter, kann man Wien mit allen Sinnen spüren. Weinberge sind da zu sehen, dahinter glänzt das bauliche Erbe der mitteleuropäischen Metropole. Ein halbes Jahrtausend wurde hier Weltgeschichte geschrieben. Kunstgeschichte sowieso.

Austria’s capital, lies in the country’s east on the Danube River. Its artistic and intellectual legacy was shaped by residents including Mozart, Beethoven and Sigmund Freud. The city is also known for its Imperial palaces, including Schönbrunn, the Habsburgs’ summer residence. In the MuseumsQuartier district, historic and contemporary buildings display works by Egon Schiele, Gustav Klimt and other artists.

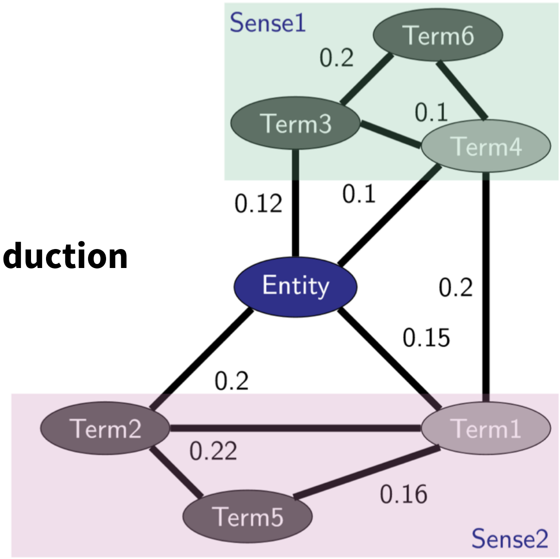
# Extraction of super classes and broader concepts

Make use of hierarchical structures in a knowledge graph



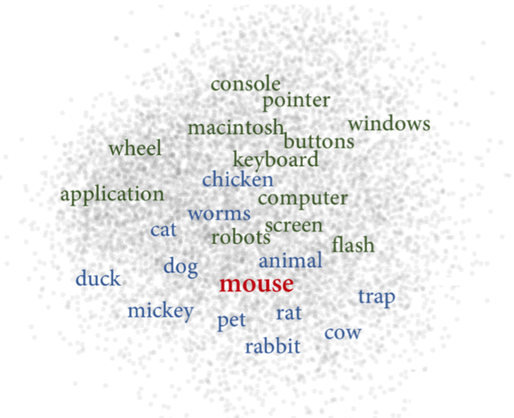
BMW has designed a car that is going to drive Jaguar X1 out of the market

**Word Sense Induction**



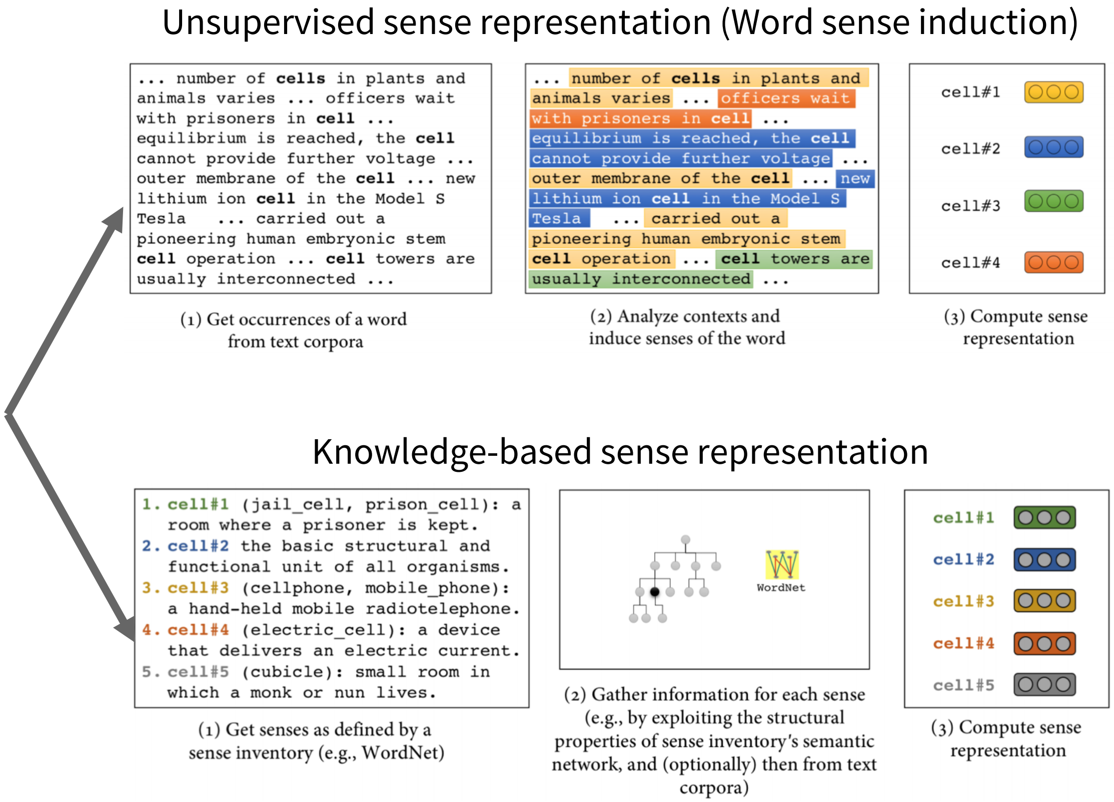
# From Word to Sense Embeddings

From Word to Sense Embeddings:
A Survey on Vector Representations of Meaning
Jose Camacho-Collados, Mohammad Taher Pilehvar
https://arxiv.org/abs/1805.04032



## The main problem:

Meaning conflation deficiency
(Word sense ambiguity)



## Unsupervised sense representation (Word sense induction)

```
... number of cells in plants and
animals varies ... officers wait
with prisoners in cell ...
equilibrium is reached, the cell
cannot provide further voltage ...
outer membrane of the cell ... new
lithium ion cell in the Model S
Tesla   ... carried out a
pioneering human embryonic stem
cell operation ... cell towers are
usually interconnected ...
```

(1) Get occurrences of a word from text corpora

```
... number of cells in plants and
animals varies ... officers wait
with prisoners in cell ...
equilibrium is reached, the cell
cannot provide further voltage ...
outer membrane of the cell ... new
lithium ion cell in the Model S
Tesla   ... carried out a
pioneering human embryonic stem
cell operation ... cell towers are
usually interconnected ...
```

(2) Analyze contexts and induce senses of the word

cell#1 ○○○
cell#2 ○○○
cell#3 ○○○
cell#4 ○○○

(3) Compute sense representation

## Knowledge-based sense representation

1. cell#1 (jail_cell, prison_cell): a room where a prisoner is kept.
2. cell#2 the basic structural and functional unit of all organisms.
3. cell#3 (cellphone, mobile_phone): a hand-held mobile radiotelephone.
4. cell#4 (electric_cell): a device that delivers an electric current.
5. cell#5 (cubicle): small room in which a monk or nun lives.

(1) Get senses as defined by a sense inventory (e.g., WordNet)

(2) Gather information for each sense (e.g., by exploiting the structural properties of sense inventory's semantic network, and (optionally) then from text corpora)

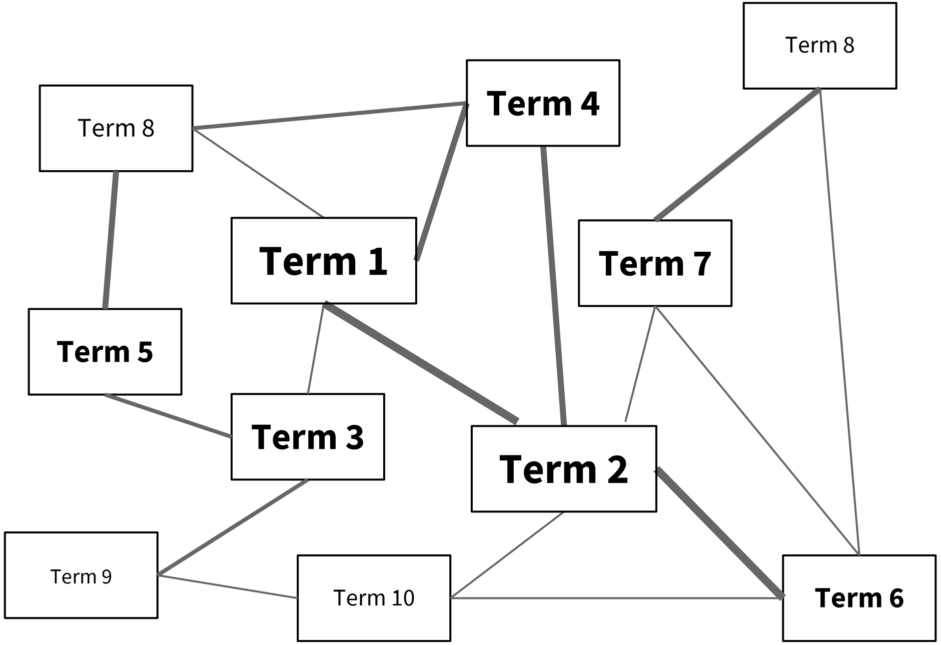
cell#1 ○○○
cell#2 ○○○
cell#3 ○○○
cell#4 ○○○
cell#5 ○○○

(3) Compute sense representation

# Extraction of related terms and concepts

Document Corpus

- Websites
- PDF, Word, …
- Abstracts from DBpedia
- RSS Feeds

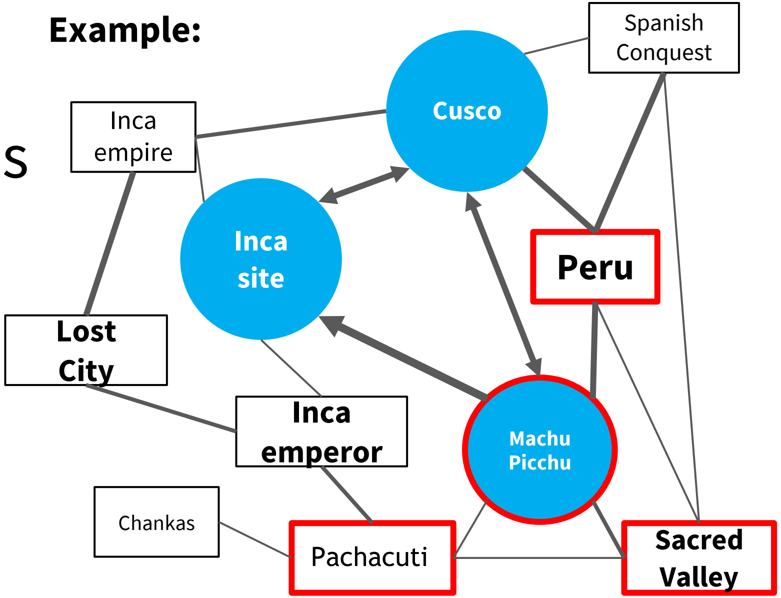


- Relevant terms and phrases
- Relevancy of terms
- co-occurrence between terms and terms

# Extraction of 'Shadow Concepts'

Use co-occurrences between concepts and terms to extract 'shadow concepts'

**Example:**



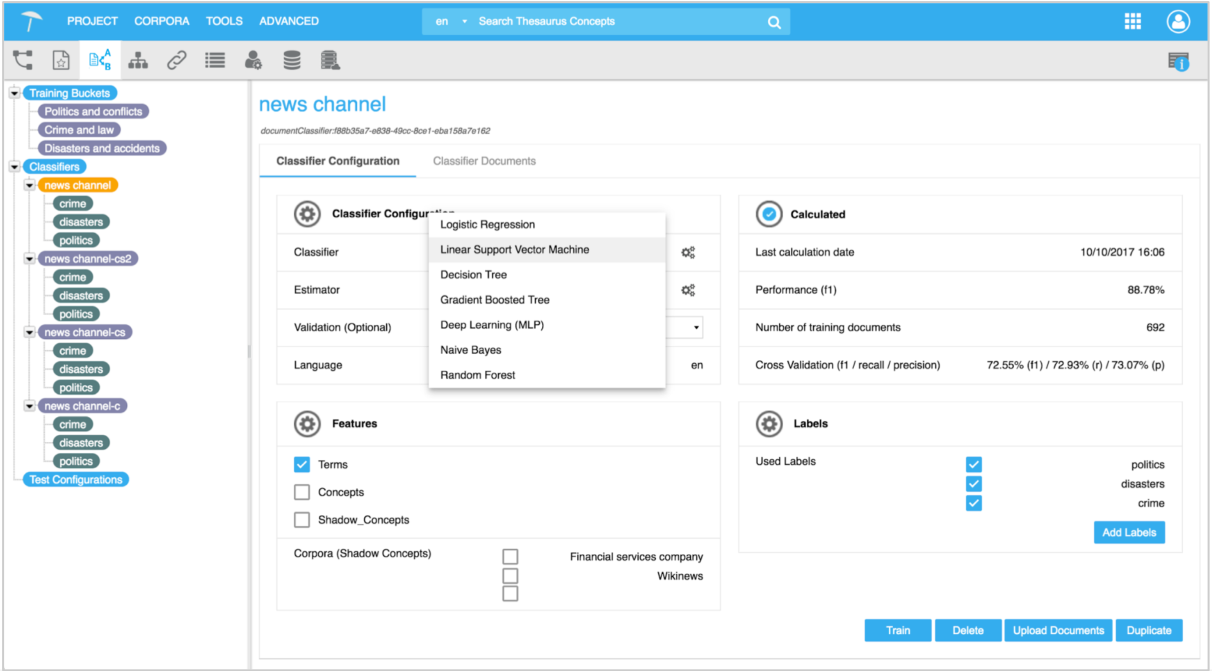
This site is a 15th-century Inca site located 2,430 metres above sea level. It is located in Cusco, **Peru**. It is situated on a mountain ridge above the **Sacred Valley** through which the Urubamba River flows. Most archaeologists believe that it was built as an estate for the Inca emperor **Pachacuti**. Often mistakenly referred to as the "Lost City of the Incas", it is the most familiar icon of Inca civilization. The Incas built the estate around 1450, but abandoned it a century later at the time of the Spanish Conquest.

In addition to explicitly used concepts and terms, ***Machu Picchu*** is extracted from the article as a *Shadow Concept*. As a prerequisite, one has to provide and analyze a representative text corpus first.

# Semantic Classifier

## Semantic Classifier

ML based on semantically enriched training data



news channel

*documentClassifier:f88b35a7-e838-49cc-8ce1-eba158a7e162*

Classifier Configuration | Classifier Documents

**Classifier Configur**

- Classifier
- Estimator
- Validation (Optional)
- Language: en

Logistic Regression
Linear Support Vector Machine
Decision Tree
Gradient Boosted Tree
Deep Learning (MLP)
Naive Bayes
Random Forest

**Calculated**

| | |
|---|---|
| Last calculation date | 10/10/2017 16:06 |
| Performance (f1) | 88.78% |
| Number of training documents | 692 |
| Cross Validation (f1 / recall / precision) | 72.55% (f1) / 72.93% (r) / 73.07% (p) |

**Features**

☑ Terms
☐ Concepts
☐ Shadow_Concepts

Corpora (Shadow Concepts)
☐ Financial services company
☐ Wikinews
☐

**Labels**

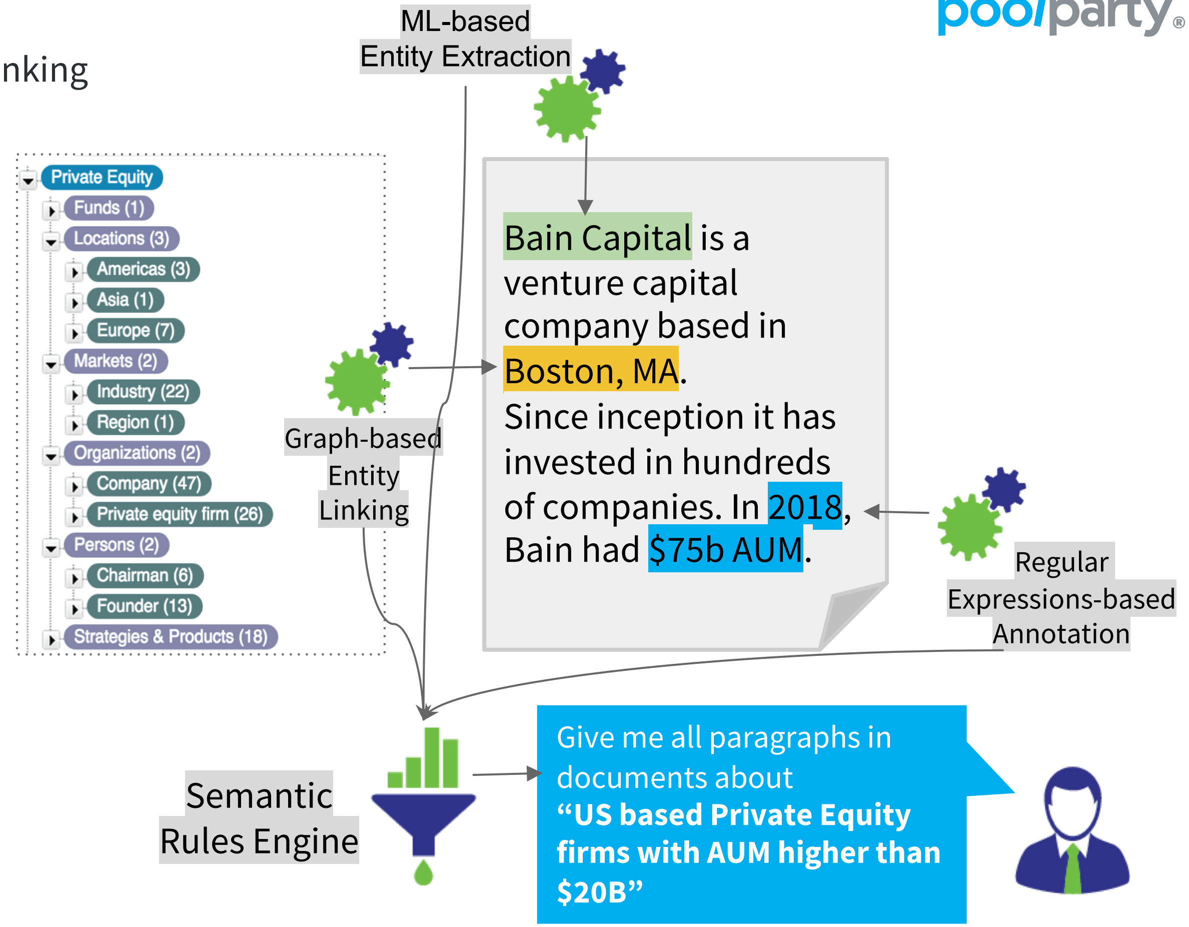
Used Labels
☑ politics
☑ disasters
☑ crime

PoolParty Semantic Classifier combines machine learning algorithms (SVM, Deep Learning, Naive Bayes, etc.) with Semantic Knowledge Graphs.

# Deep Text Analysis

## Annotation, Extraction, Classification, Linking

- Corpus statistics / Word embeddings
  → **Keyphrase extraction**
- Graph-based annotation
  → **Entity/Concept linking**
- Corpus Statistics embedded in graphs
  → **Shadow Concepts**
- Machine-learning-based annotation
  → **Named entity recognition (NER)**
- Machine-learning based classification
  → **Document Classification**
- Annotation based on rules
  → **Regular expressions**



- Private Equity
  - Funds (1)
  - Locations (3)
    - Americas (3)
    - Asia (1)
    - Europe (7)
  - Markets (2)
    - Industry (22)
    - Region (1)
  - Organizations (2)
    - Company (47)
    - Private equity firm (26)
  - Persons (2)
    - Chairman (6)
    - Founder (13)
  - Strategies & Products (18)

ML-based Entity Extraction

Graph-based Entity Linking

Regular Expressions-based Annotation

Bain Capital is a venture capital company based in Boston, MA. Since inception it has invested in hundreds of companies. In 2018, Bain had $75b AUM.

Semantic Rules Engine

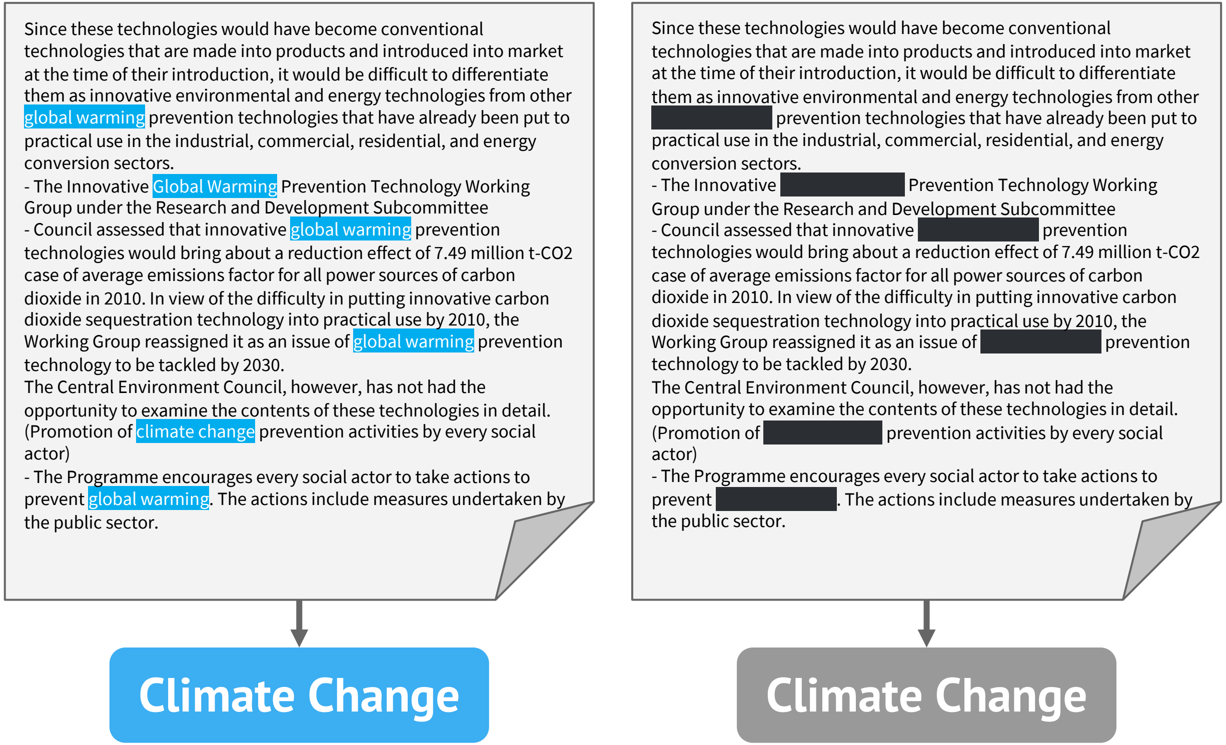
Give me all paragraphs in documents about **“US based Private Equity firms with AUM higher than $20B”**

# USE CASES

See how it can be used in your environment!

# Extract concepts from text even if not used explicitly

Some domains use text that doesn’t always call a spade a spade. With ‘shadow concept extraction’ those ‘masked’ concepts still can be surfaced.



Since these technologies would have become conventional technologies that are made into products and introduced into market at the time of their introduction, it would be difficult to differentiate them as innovative environmental and energy technologies from other global warming prevention technologies that have already been put to practical use in the industrial, commercial, residential, and energy conversion sectors.

- The Innovative Global Warming Prevention Technology Working Group under the Research and Development Subcommittee
- Council assessed that innovative global warming prevention technologies would bring about a reduction effect of 7.49 million t-CO2 case of average emissions factor for all power sources of carbon dioxide in 2010. In view of the difficulty in putting innovative carbon dioxide sequestration technology into practical use by 2010, the Working Group reassigned it as an issue of global warming prevention technology to be tackled by 2030.

The Central Environment Council, however, has not had the opportunity to examine the contents of these technologies in detail.
(Promotion of climate change prevention activities by every social actor)

- The Programme encourages every social actor to take actions to prevent global warming. The actions include measures undertaken by the public sector.

Climate Change

Since these technologies would have become conventional technologies that are made into products and introduced into market at the time of their introduction, it would be difficult to differentiate them as innovative environmental and energy technologies from other [redacted] prevention technologies that have already been put to practical use in the industrial, commercial, residential, and energy conversion sectors.

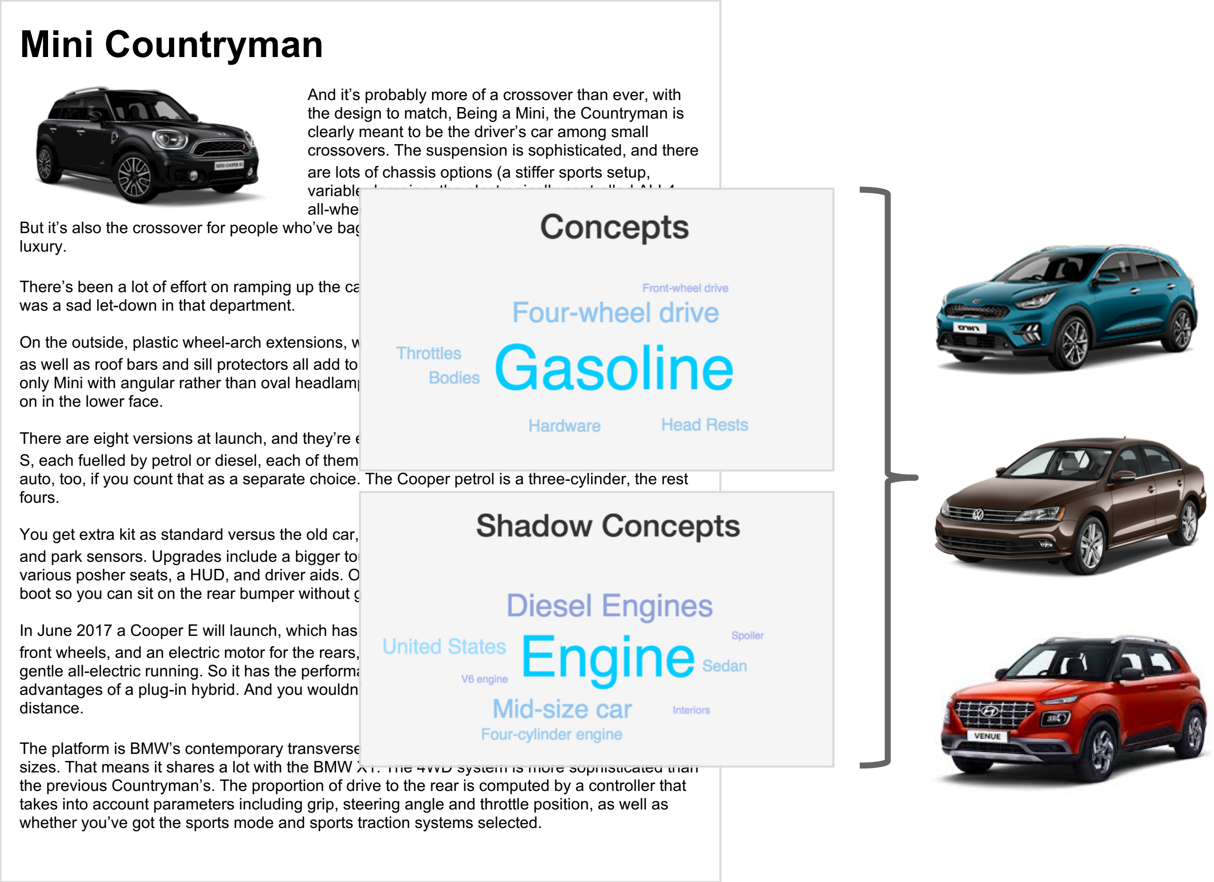
- The Innovative [redacted] Prevention Technology Working Group under the Research and Development Subcommittee
- Council assessed that innovative [redacted] prevention technologies would bring about a reduction effect of 7.49 million t-CO2 case of average emissions factor for all power sources of carbon dioxide in 2010. In view of the difficulty in putting innovative carbon dioxide sequestration technology into practical use by 2010, the Working Group reassigned it as an issue of [redacted] prevention technology to be tackled by 2030.

The Central Environment Council, however, has not had the opportunity to examine the contents of these technologies in detail.
(Promotion of [redacted] prevention activities by every social actor)

- The Programme encourages every social actor to take actions to prevent [redacted]. The actions include measures undertaken by the public sector.

Climate Change

# Use Shadow Concepts to improve Recommender Systems



## Mini Countryman

And it's probably more of a crossover than ever, with the design to match, Being a Mini, the Countryman is clearly meant to be the driver's car among small crossovers. The suspension is sophisticated, and there are lots of chassis options (a stiffer sports setup, variable

all-whe

But it's also the crossover for people who've bag

luxury.

There's been a lot of effort on ramping up the ca

was a sad let-down in that department.

On the outside, plastic wheel-arch extensions, w

as well as roof bars and sill protectors all add to

only Mini with angular rather than oval headlam

on in the lower face.

There are eight versions at launch, and they're e

S, each fuelled by petrol or diesel, each of them

auto, too, if you count that as a separate choice. The Cooper petrol is a three-cylinder, the rest fours.

You get extra kit as standard versus the old car,

and park sensors. Upgrades include a bigger to

various posher seats, a HUD, and driver aids. O

boot so you can sit on the rear bumper without g

In June 2017 a Cooper E will launch, which has

front wheels, and an electric motor for the rears,

gentle all-electric running. So it has the performa

advantages of a plug-in hybrid. And you wouldn

distance.

The platform is BMW's contemporary transverse

sizes. That means it shares a lot with the BMW X

the previous Countryman's. The proportion of drive to the rear is computed by a controller that takes into account parameters including grip, steering angle and throttle position, as well as whether you've got the sports mode and sports traction systems selected.

### Concepts

Front-wheel drive

Four-wheel drive

Throttles

Bodies

Gasoline

Hardware

Head Rests

### Shadow Concepts

Diesel Engines

Spoiler

United States

Engine

Sedan

V6 engine

Mid-size car

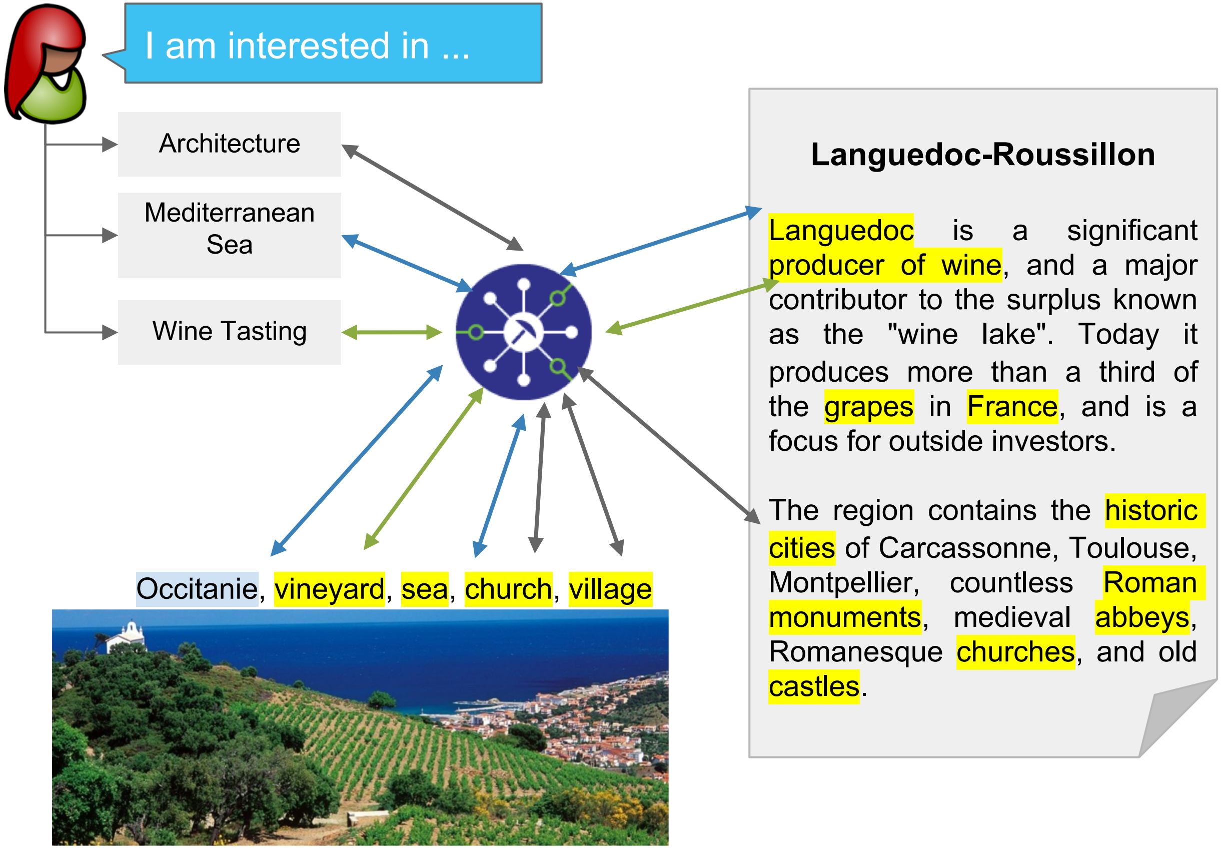
Interiors

Four-cylinder engine

# Use Case Recommender System

Connecting

- content to content
- people to content
- people to people



**Languedoc-Roussillon**

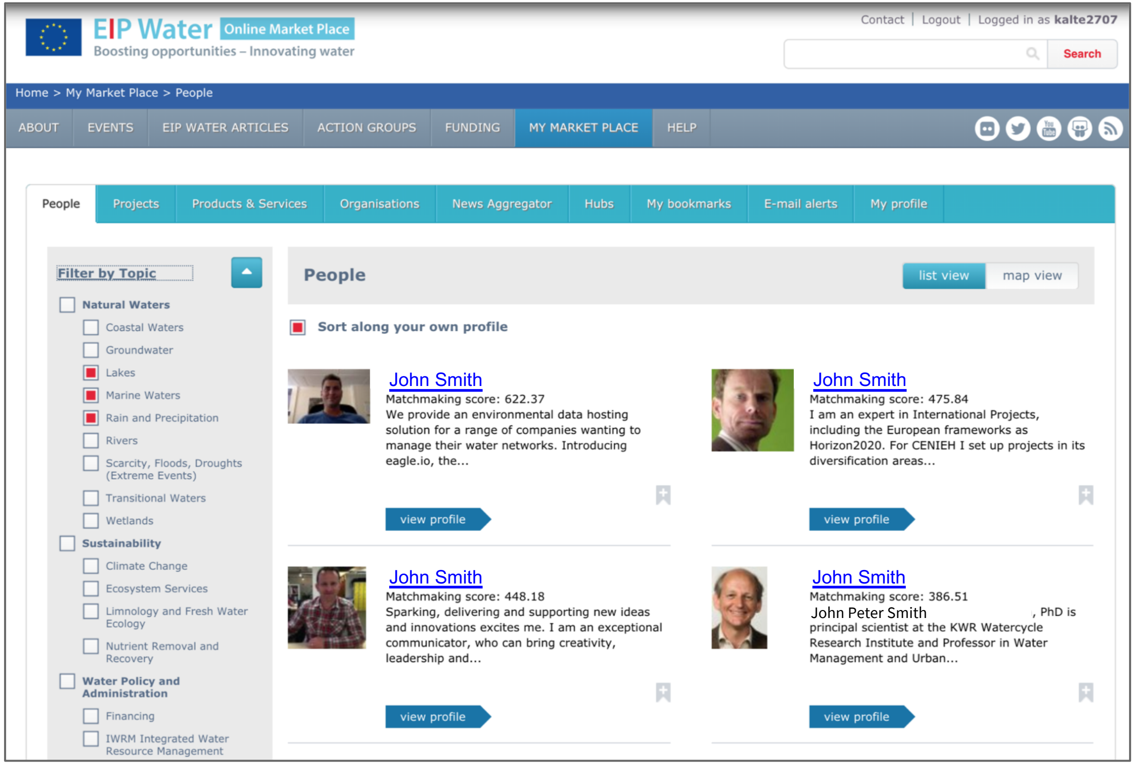
Languedoc is a significant producer of wine, and a major contributor to the surplus known as the "wine lake". Today it produces more than a third of the grapes in France, and is a focus for outside investors.

The region contains the historic cities of Carcassonne, Toulouse, Montpellier, countless Roman monuments, medieval abbeys, Romanesque churches, and old castles.

Demo

# Example EIP on Water

At the center of the marketplace is the matchmaking function



Filter by Topic

- □ Natural Waters
  - □ Coastal Waters
  - □ Groundwater
  - ☒ Lakes
  - ☒ Marine Waters
  - ☒ Rain and Precipitation
  - □ Rivers
  - □ Scarcity, Floods, Droughts (Extreme Events)
  - □ Transitional Waters
  - □ Wetlands
- □ Sustainability
  - □ Climate Change
  - □ Ecosystem Services
  - □ Limnology and Fresh Water Ecology
  - □ Nutrient Removal and Recovery
- □ Water Policy and Administration
  - □ Financing
  - □ IWRM Integrated Water Resource Management

People

☒ Sort along your own profile

John Smith
Matchmaking score: 622.37
We provide an environmental data hosting solution for a range of companies wanting to manage their water networks. Introducing eagle.io, the...

John Smith
Matchmaking score: 475.84
I am an expert in International Projects, including the European frameworks as Horizon2020. For CENIEH I set up projects in its diversification areas...

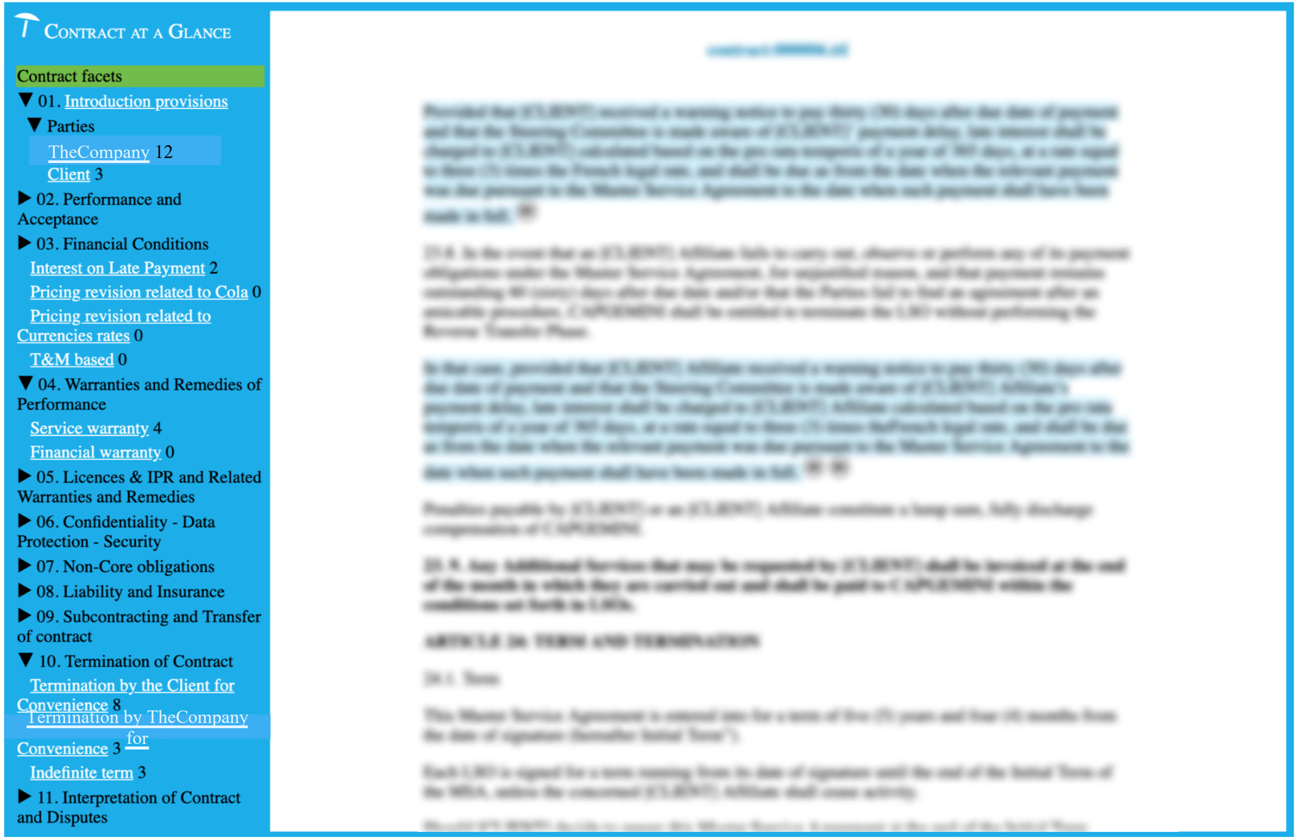
John Smith
Matchmaking score: 448.18
Sparking, delivering and supporting new ideas and innovations excites me. I am an exceptional communicator, who can bring creativity, leadership and...

John Smith
Matchmaking score: 386.51
John Peter Smith , PhD is principal scientist at the KWR Watercycle Research Institute and Professor in Water Management and Urban...

https://www.eip-water.eu/

# Contract Intelligence

Semantic analysis and pre-selection of relevant clauses in contracts



CONTRACT AT A GLANCE

Contract facets

- ▼ 01. Introduction provisions
  - ▼ Parties
    - TheCompany 12
    - Client 3
- ▶ 02. Performance and Acceptance
- ▶ 03. Financial Conditions
  - Interest on Late Payment 2
  - Pricing revision related to Cola 0
  - Pricing revision related to Currencies rates 0
  - T&M based 0
- ▼ 04. Warranties and Remedies of Performance
  - Service warranty 4
  - Financial warranty 0
- ▶ 05. Licences & IPR and Related Warranties and Remedies
- ▶ 06. Confidentiality - Data Protection - Security
- ▶ 07. Non-Core obligations
- ▶ 08. Liability and Insurance
- ▶ 09. Subcontracting and Transfer of contract
- ▼ 10. Termination of Contract
  - Termination by the Client for Convenience 8
  - Termination by TheCompany for Convenience 3
  - Indefinite term 3
- ▶ 11. Interpretation of Contract and Disputes

# Connect

## Andreas Koller

CIO & Managing Partner, Semantic Web Company

- andreas.koller@semantic-web.com
- https://www.linkedin.com/in/ankoller/
- https://twitter.com/semwebcompany

?

# Q & A