

Web sémantique et intelligence artificielle

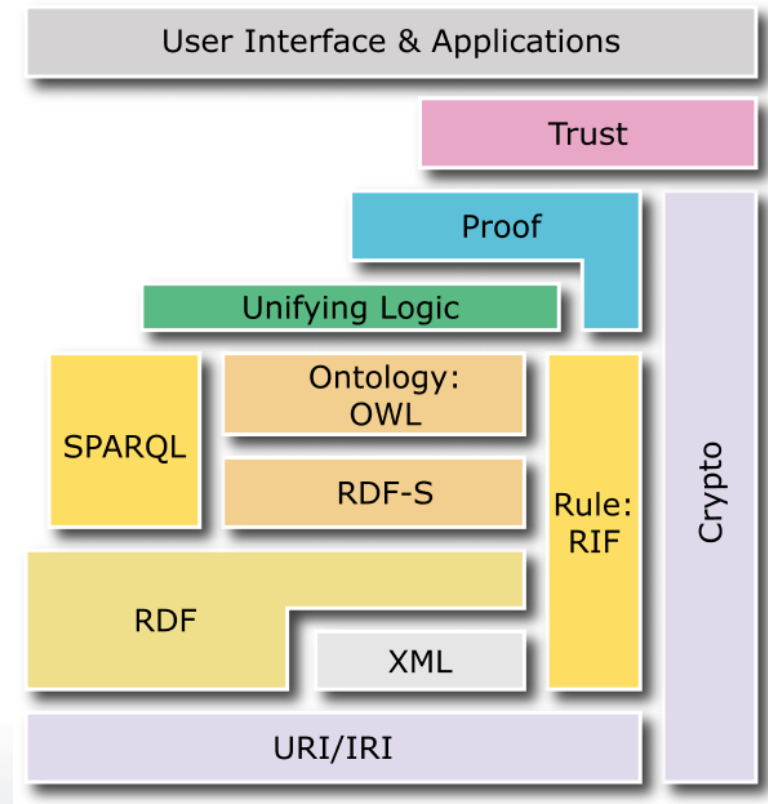
- **Comment le Web Sémantique peut propulser certaines avancées en intelligence artificielle?**

- 1. Il intervient dans 77% des tâches reliées à la création de modèles d'apprentissage automatique**

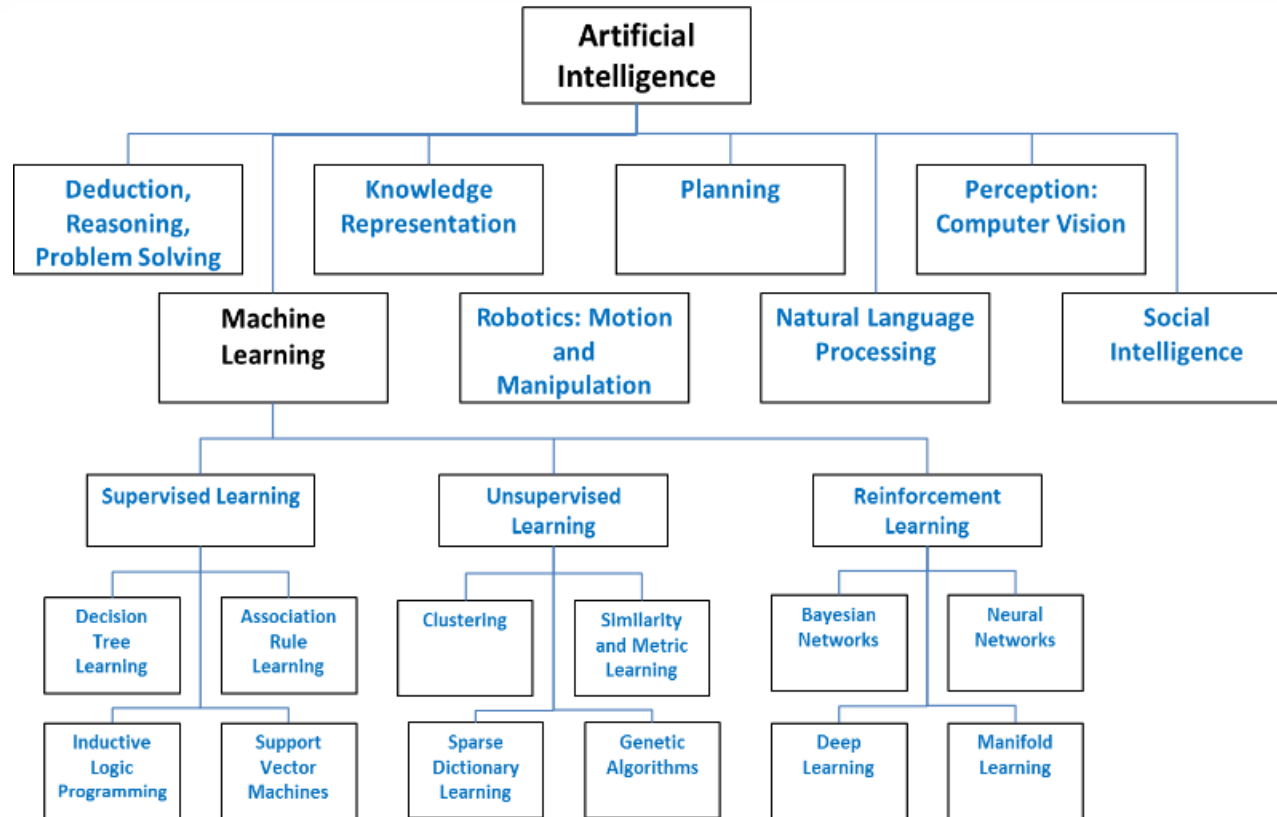
- 2. Il fait partie intégrante d'un modèle mathématique**

Qu'est-ce que le Web Sémantique?

- Un ensemble de technologies d'aide à la gestion et à la valorisation de l'information



Intelligence Artificielle



Algorithmes d'apprentissages

« gives computers the ability to learn without being explicitly programmed » - Arthur Samuel, 1959

- **Apprentissage supervisé**
 - ✓ Exemples d'entrées et de sorties
- **Apprentissage non-supervisé**
 - ✓ Trouver la structure des données par elle-même
- **Apprentissage par renforcement**
 - ✓ L'action de l'algorithme sur l'environnement produit une valeur de retour qui guide l'algorithme d'apprentissage

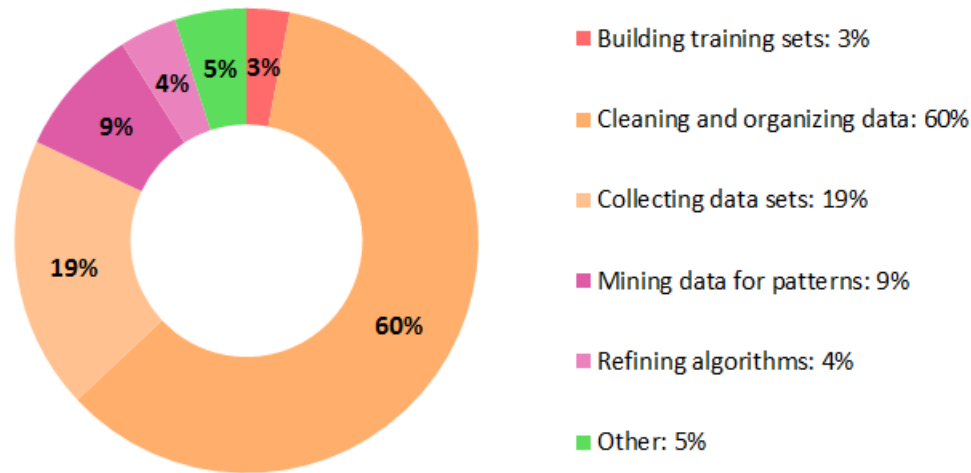
L'information: l'essence de l'apprentissage automatique

- **Consensus: la disponibilité d'une quantité importante de données numériques et structurées est l'un des facteurs déterminants ayant contribué aux avancées récentes de l'intelligence artificielle, particulièrement dans le domaine de l'apprentissage automatique**

« The most important new development is that today we can provide these algorithms with the resources they need to succeed. Figure 1.8 shows how the size of benchmark datasets has expanded remarkably over time. »

- Deep Learning, Ian Goodfellow, Yoshua Benbgio, and Aaron Courville

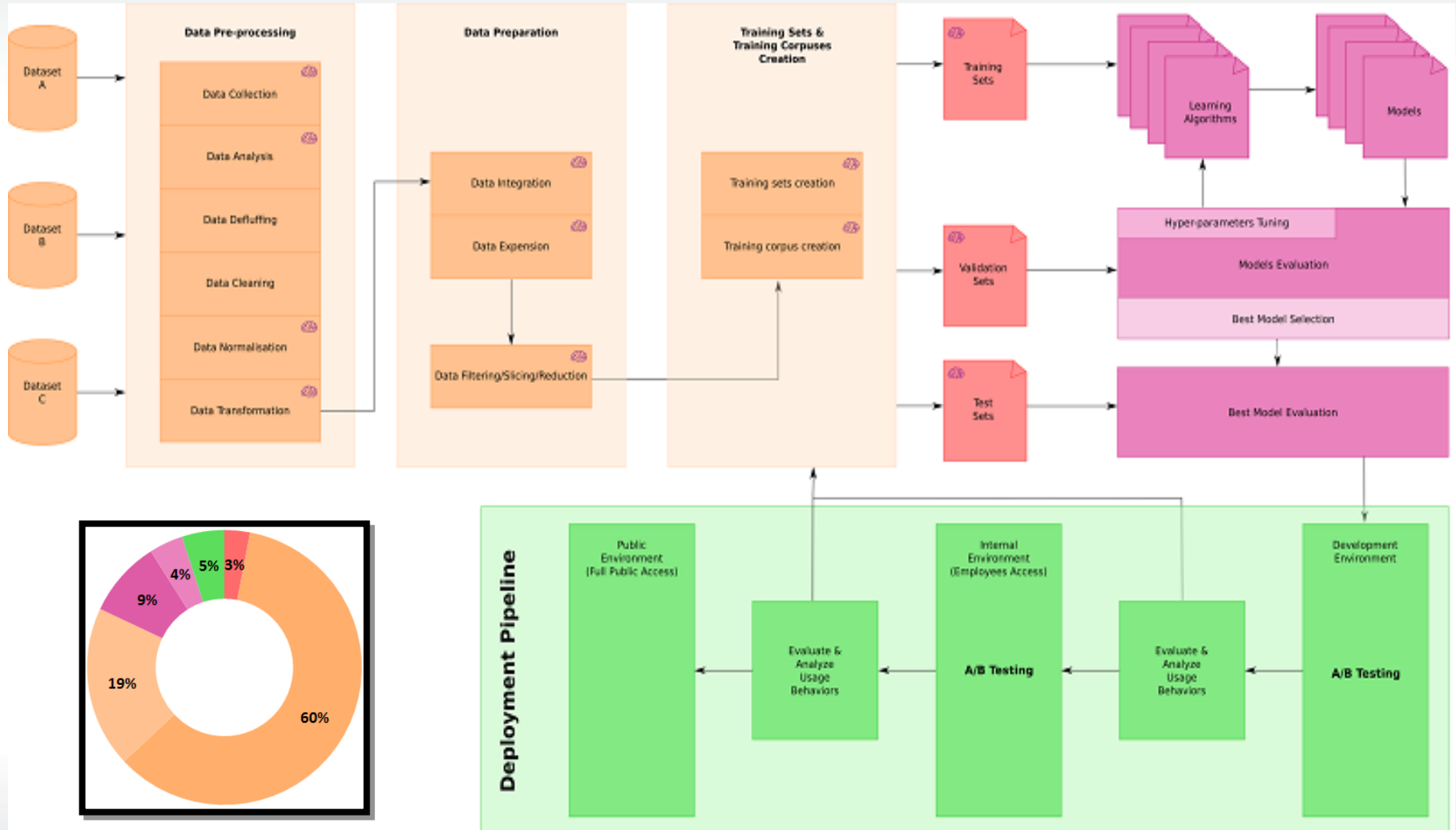
“What data scientists spend the most time doing?”



- Enquête de 80 Data Scientist par CrowdFlower en 2016

- 77% du temps des scientifiques de la donnée dépensé à effectuer des tâches connexes

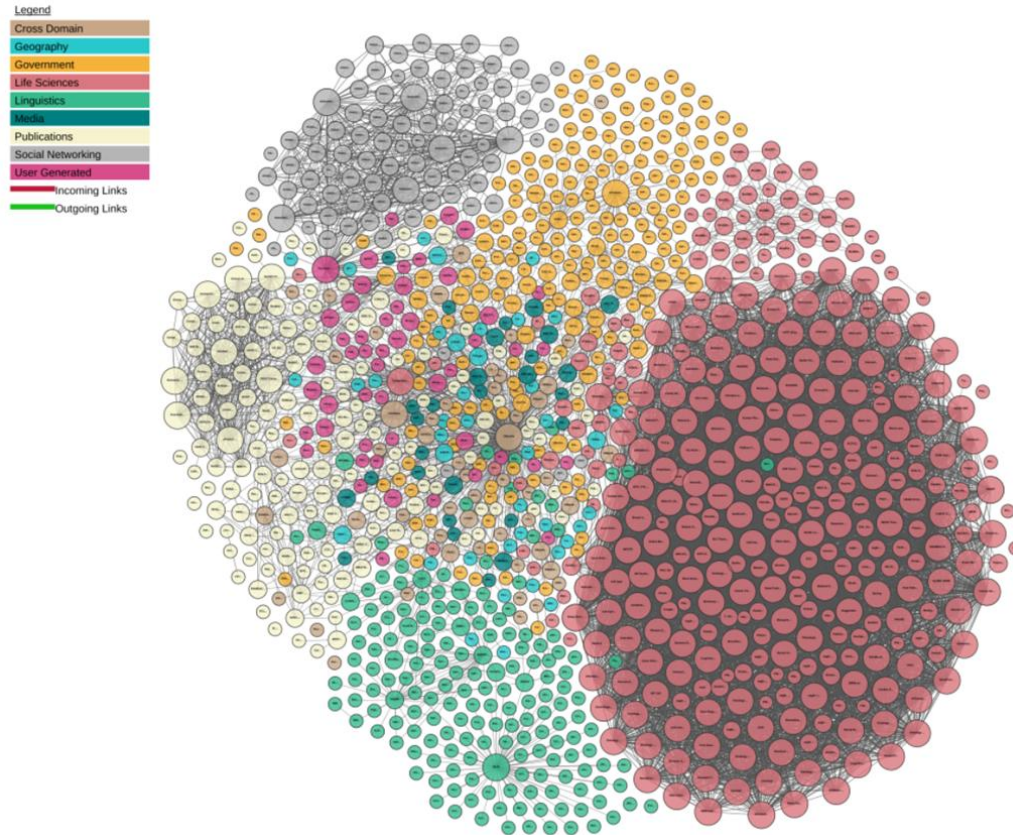
Quels sont ces tâches?



Cas d'utilisations

- **Cas: Acquisition de jeux de données**
 - ✓ Linked Open Data: Jeux de données intégrés
- **Création d'un corpus de textes spécialisés**
 - ✓ Création de jeux de données d'apprentissages
 - ✓ Nettoyer et organiser l'information
 - ✓ Acquisition de jeux de données
- **Cas: Classification automatique de documents**
 - ✓ Création de jeux de données d'apprentissages
 - ✓ Nettoyer et organiser l'information
 - ✓ Acquisition de jeux de données

Cas 1: Acquisition de jeux de données

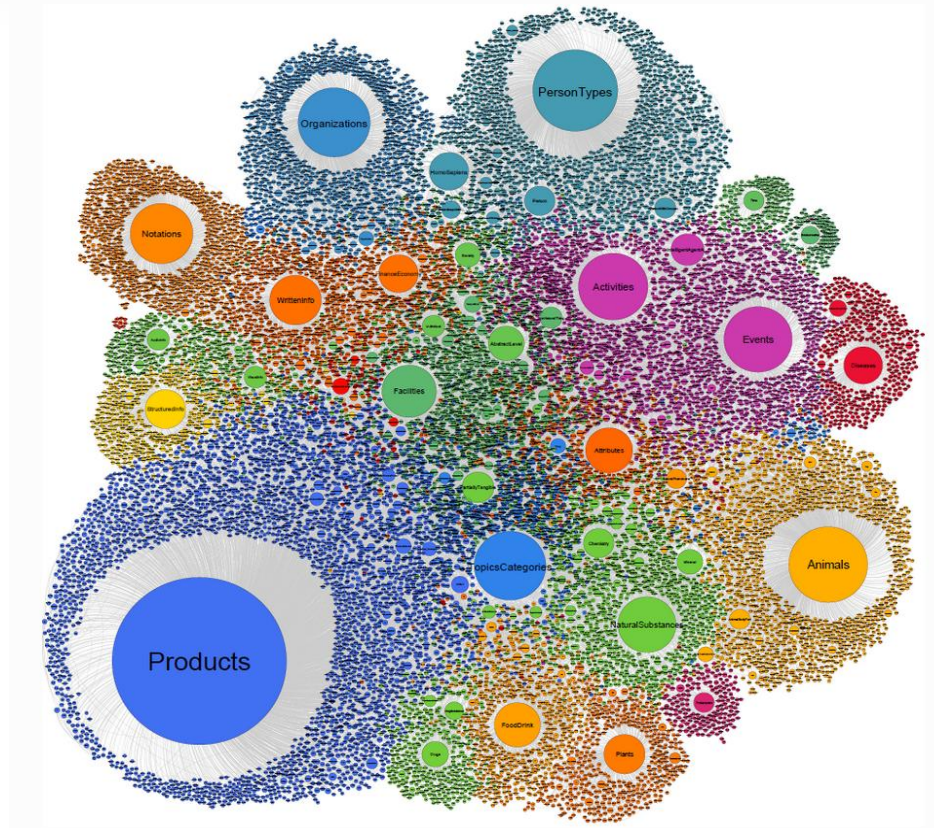


Linked Open Data

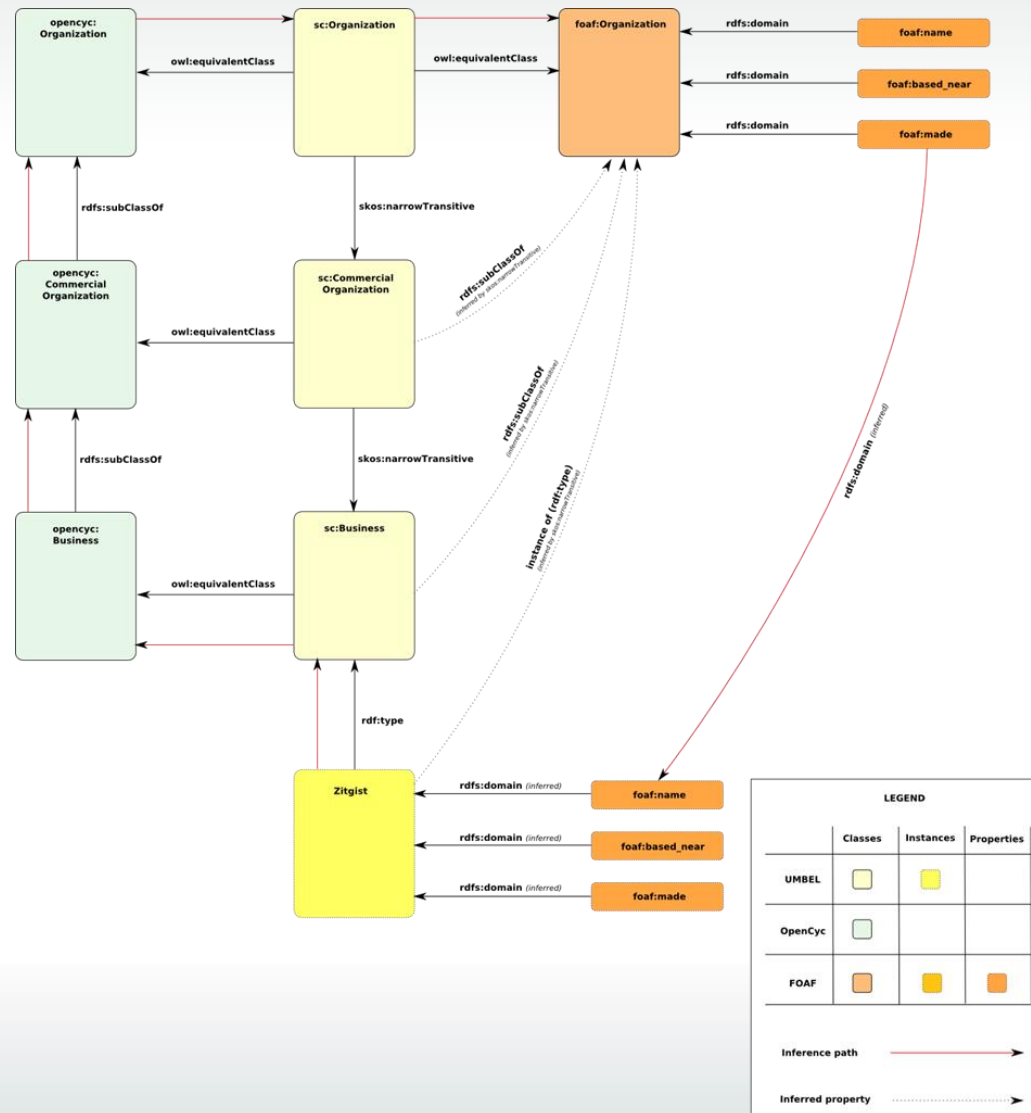
- **2973 jeux de données intégrées**
- **Accessible via SPARQL**
 - 149,423,660,620 (149 milliards) triplets
- **Accessible via jeux de données sérialisées**
 - 192,230,648 (192 millions) triplets
 - 54,225,309 entités

- *LODStats* (stats.lod2.eu)

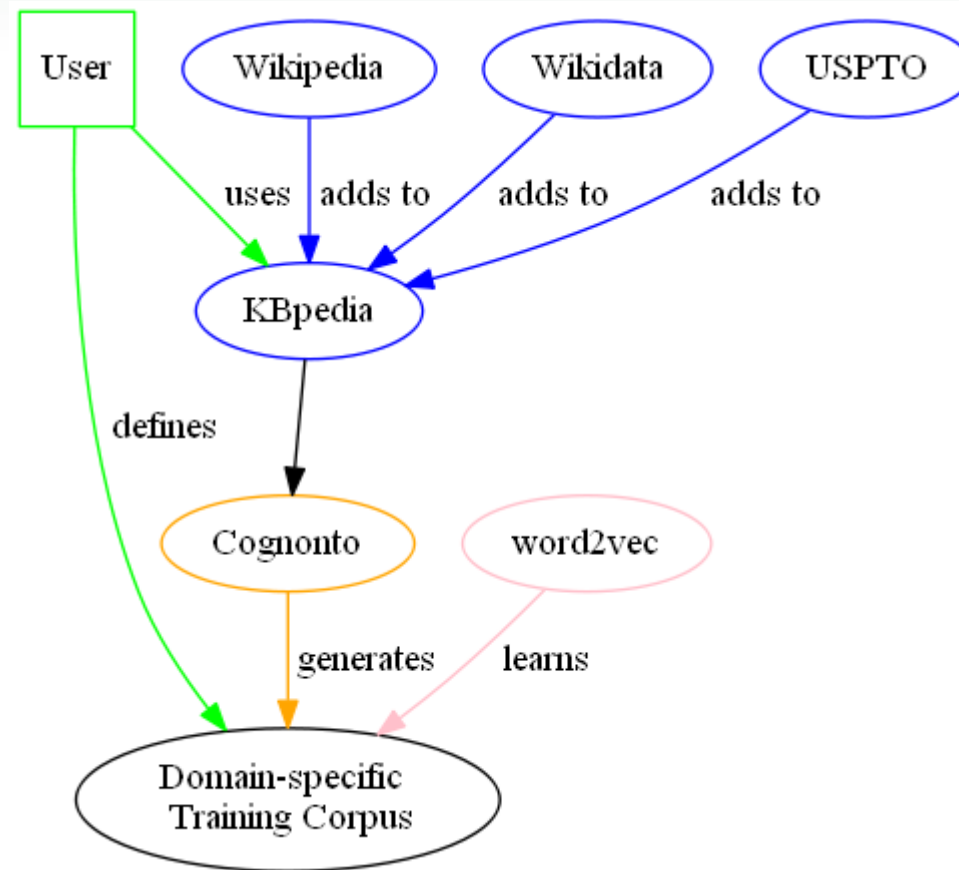
Utilisation d'une base de connaissance



Liens entre les classes des ontologies



Cas 2: Création d'un corpus de textes spécialisés

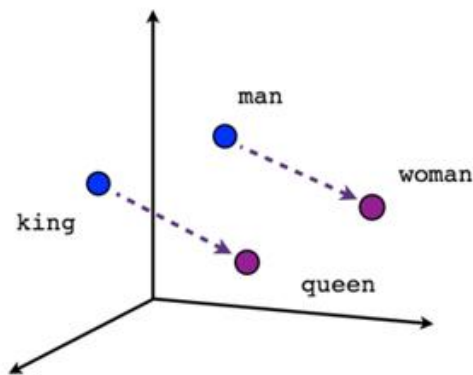


Corpus: musiciens

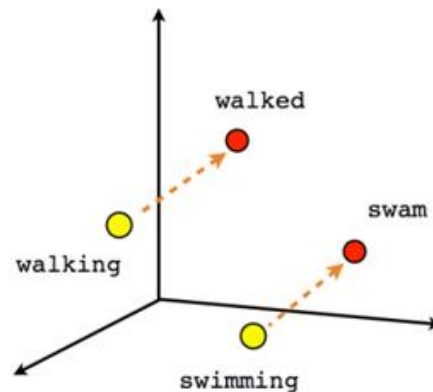
- 47,263 documents relatifs aux musiciens de Wikipedia
- Comparaison avec le « *Google News model* » comprenant 3 milliards de mots

Algorithme d'apprentissage: word2vec

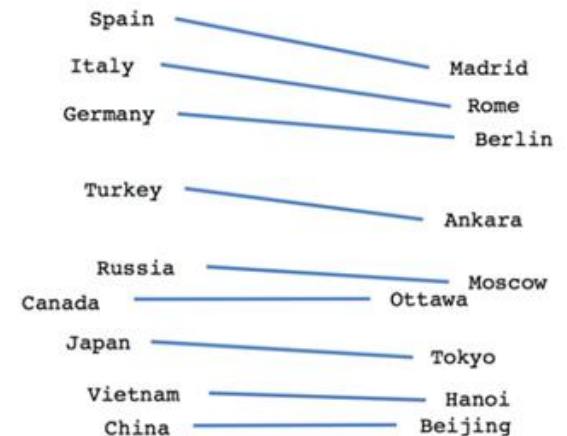
- *Word2vec est un algorithme qui transforme les mots en vecteurs, de sorte que les mots ayant une signification similaire finissent par se rapprocher les uns des autres.*
- *King – man + woman = queen*



Male-Female



Verb tense



Country-Capital

Metal + Death = ?

Metal + Death = ?

```
(.wordsNearest musicians-model ["metal" "death"] [] 10)
```

- death
- deathcore
- mathcore
- metal
- metalcore
- melodic
- gothic
- thrash
- grindcore
- ~~present_labels_insideout~~

```
(.wordsNearest google-news-model ["metal" "death"] [] 10)
```

- death
- ~~stool~~
- ~~untimely_death~~
- metal
- Death
- ~~thallium_toxic~~
- ~~blackened_shards~~
- ~~Tunstall_bled~~
- ~~compressional_asphyxia~~
- metal_grindcore

Metal - Death + Smooth = ?

Metal - Death + Smooth = ?

```
(.wordsNearest musicians-model ["metal" "smooth"] ["death"] 5)
```

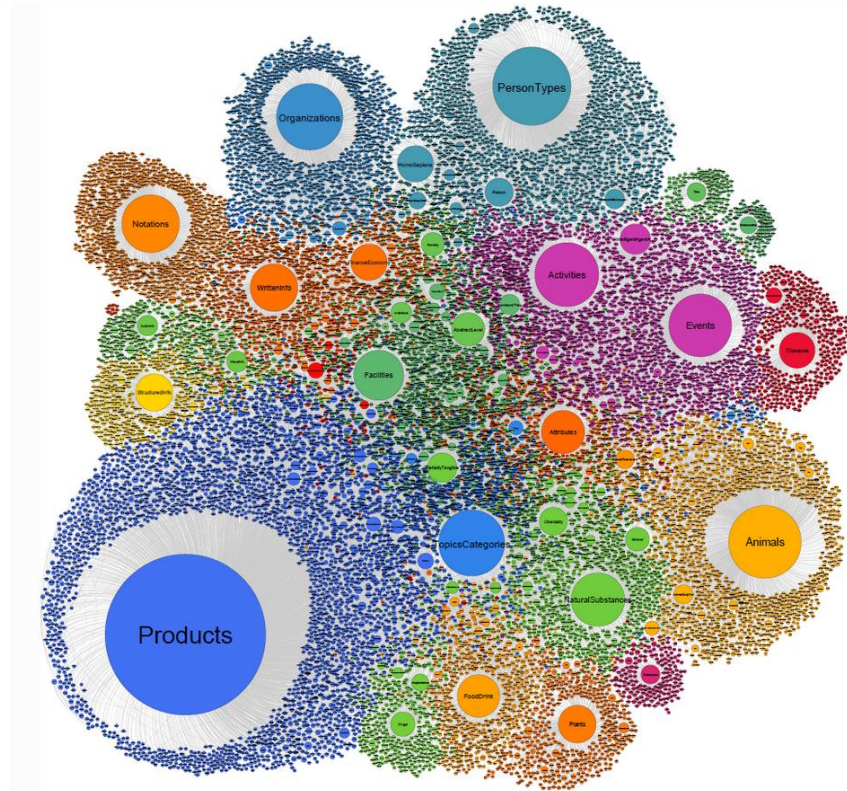
- smooth
- funk
- soul
- disco
- daniel_håkansson_genres

```
(.wordsNearest google-news-model ["metal" "smooth"] ["death"] 5)
```

- smooth
- metal
- ~~• shredding_solos~~
- ~~• chromed_steel~~
- ~~• metallic~~

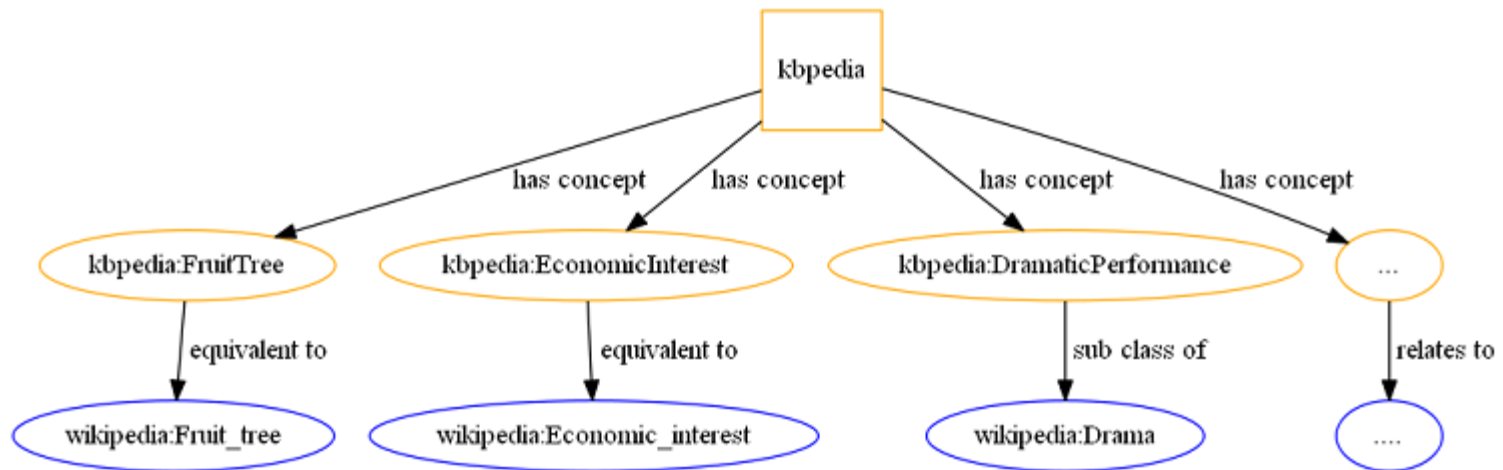
Cas: Classification automatique de documents

- Utilisation d'une base de connaissance



Cas 3: Classification automatique de documents

- Création d'un corpus de texte général



Cas: Classification automatique de documents

- Création d'un corpus de texte relatif au domaine de la musique

```
(define-domain-corpus ["http://kbpedia.org/kko/rc/Music"  
    "http://kbpedia.org/kko/rc/Musician"  
    "http://kbpedia.org/kko/rc/MusicPerformanceOrganization"  
    "http://kbpedia.org/kko/rc/MusicalInstrument"  
    "http://kbpedia.org/kko/rc/Album-CW"  
    "http://kbpedia.org/kko/rc/Album-IBO"  
    "http://kbpedia.org/kko/rc/MusicalComposition"  
    "http://kbpedia.org/kko/rc/MusicalText"  
    "http://kbpedia.org/kko/rc/PropositionalConceptualWork-MusicalGenre"  
    "http://kbpedia.org/kko/rc/MusicalPerformer"]  
    "resources/kbpedia_reference_concepts_linkage.n3"  
    "resources/domain-corpus-dictionary.csv")
```

< >

Cas: Classification automatique de documents

- **Création d'un corpus d'entraînement général**
 - ✓ 24,374 paged Wikipedia
- **Création d'un corpus d'entraînement spécialisé sur la musique**
 - 62 pages Wikipedia

Cas: Classification automatique de documents

- **Création de l'interpréteur sémantique**

- Utilisation de l'algorithme ESA (*Explicit Semantic Analysis*) pour la création des vecteur numériques
- Utilisation de l'algorithme SVM (*Support Vector Machine*) pour la classification des documents

Cas: Classification automatique de documents

■ Création d'un jeu de référence

```
(defn create-gold-standard-from-feeds
  [name]
  (let [feeds ["http://rss.cbc.ca/lineup/topstories.xml"
               "http://rss.cbc.ca/lineup/world.xml"
               "http://rss.cbc.ca/lineup/canada.xml"
               "http://rss.cbc.ca/lineup/politics.xml"
               "http://rss.cbc.ca/lineup/business.xml"
               "http://rss.cbc.ca/lineup/health.xml"
               "http://rss.cbc.ca/lineup/arts.xml"
               "http://rss.cbc.ca/lineup/technology.xml"
               "http://rss.cbc.ca/lineup/offbeat.xml"
               "http://www.cbc.ca/cmlink/rss-cbcaboriginal"
               "http://rss.cbc.ca/lineup/sports.xml"
               "http://rss.cbc.ca/lineup/canada-britishcolumbia.xml"
               "http://rss.cbc.ca/lineup/canada-calgary.xml"
               "http://rss.cbc.ca/lineup/canada-montreal.xml"
               "http://rss.cbc.ca/lineup/canada-pe1.xml"
               "http://rss.cbc.ca/lineup/canada-ottawa.xml"
               "http://rss.cbc.ca/lineup/canada-toronto.xml"
               "http://rss.cbc.ca/lineup/canada-north.xml"
               "http://rss.cbc.ca/lineup/canada-manitoba.xml"
               "http://feeds.reuters.com/news/artsCulture"
               "http://feeds.reuters.com/reuters/businessNews"
               "http://feeds.reuters.com/reuters/entertainment"
               "http://feeds.reuters.com/reuters/companyNews"
               "http://feeds.reuters.com/reuters/lifestyle"
               "http://feeds.reuters.com/reuters/healthNews"
               "http://feeds.reuters.com/reuters/MostRead"
               "http://feeds.reuters.com/reuters/peopleNews"
               "http://feeds.reuters.com/reuters/scienceNews"
               "http://feeds.reuters.com/reuters/technologyNews"
               "http://feeds.reuters.com/Reuters/domesticNews"
               "http://feeds.reuters.com/Reuters/worldNews"
               "http://feeds.reuters.com/reuters/USmediaDiversifiedNews"]]

    (with-open [out-file (io/writer (str "resources/" name ".csv"))]
      (csv/write-csv out-file [{"class" "title" "url"}])
      (doseq [feed-url feeds]
        (doseq [item (:entries (feed/parse-feed feed-url))]
          (csv/write-csv out-file "" (:title item) (:link item) :append true))))))
```

Cas: Classification automatique de documents

- Évaluation de différents modèles

```
(evaluate-model "w200" "resources/gold-standard-2.csv")
```

```
True positive: 23  
False positive: 6  
True negative: 314  
False negative: 2  
  
Precision: 0.79310346  
Recall: 0.92  
Accuracy: 0.9768116  
F1: 0.8518519
```


Cas: Classification automatique de documents

- **Avantages**

- Création rapide de différents jeux de données d'apprentissages ("*slice and dice*")
- Évaluation rapide de la performance de chacun des modèles

Autres cas d'application

- Résoudre les ambiguïtés entre les concepts automatiquement balisés dans un texte
- Identifier les entités dupliquées entre deux jeux de données
- Identifier et ajouter des concepts manquants d'une ontologie en provenance d'une autre structure

Merci! Des questions?

- **Frédérick Giasson**

- ✓ Courriel: fred@fgiasson.com
- ✓ Cognonto: <http://cognonto.com>
- ✓ Site personel: <http://fgiasson.com>
- ✓ Blog: <http://fgiasson.com/blog/>