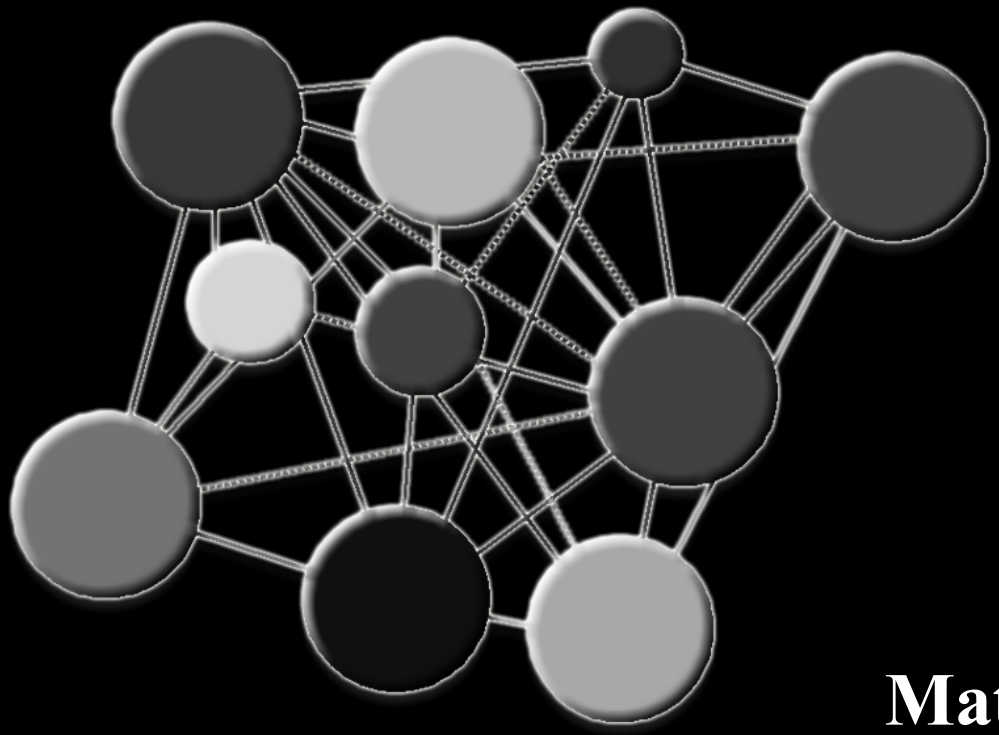
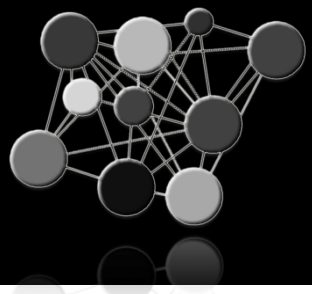


Mathieu Lafourcade
Manel Zarrouk
Alain Joubert

Inférence de règles déductives par abduction

Plan



- Où ça ? réseau lexico-sémantique
- Quoi ? règles
- Comment ? savoir compter
- Combien ? bien ou pas bien ?

Contexte ~ acquisition en sémantique lexicale

A quoi de telles ressources peuvent servir ?

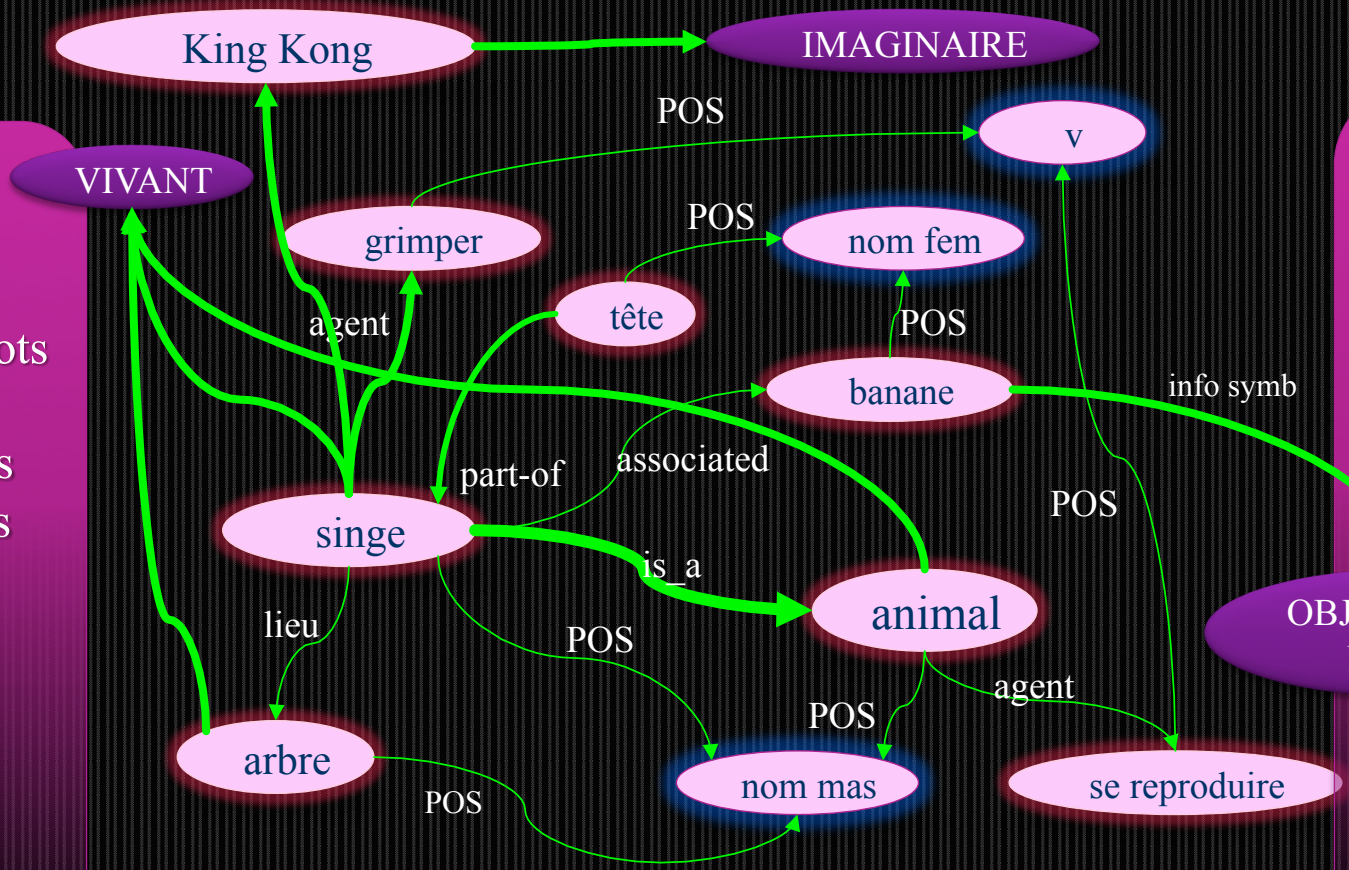
- Applications automatiques nécessitant des K du monde, linguistiques et terminologiques : analyse de texte, RI, TA …
- Assistance lexicale

Comment les construire ?

- Manuellement (par des experts ou des linguistes)
  Long, coûteux, statique, normatif
- Automatiquement (par extraction depuis des corpus)
  Difficile + connaissances implicites absentes des textes
- Crowdsourcing
  Efficace, mais gros silences si contributif

Contexte ~ réseau lexico-sémantique (RLS)

une représentation en sémantique lexicale



Nœuds

Termes
Usage de mots
Concepts
Informations linguistiques

Symboles

Relations

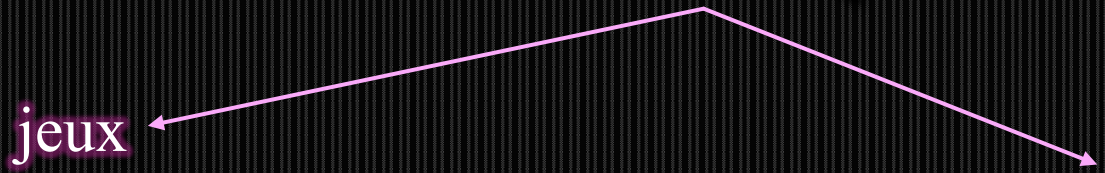
Ontologiques
Sémantiques
Lexicales
Symboliques

Contexte ~ approches par les foules

Considérations : coût, qualité, durée

Peu cher – efficace – rapide - populaire

Externalisation ouverte (crowdsourcing)



Sur des données (JDM)
Sur des connaissances (Règles)

Frontière poreuse

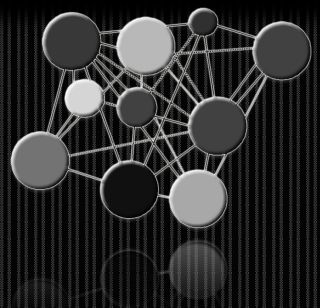
Contexte ~ approches par les foules

RLS JDM
construit par externalisation
ouverte

JeuxDeMots

Jeu en ligne
Contributions **non-négociées**

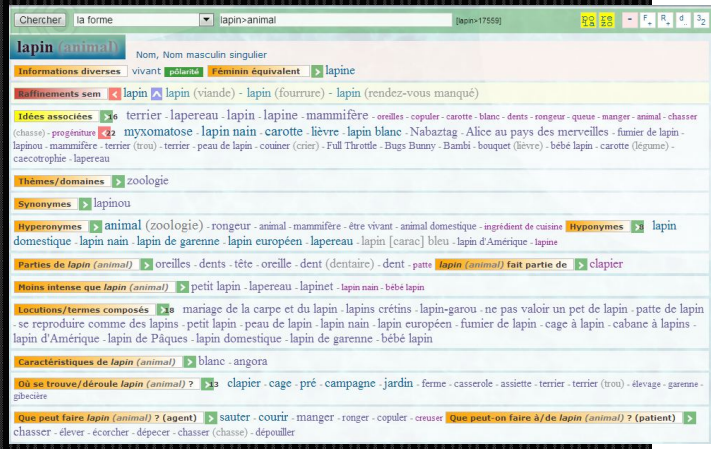
Jeux de mots .org



259 984 termes
2 154 207 relations
60 types de relations

Diko

outil contributif
Contributions **négociées**



Contexte ~ Application au RLS JDM

SILENCE + BRUIT

Relations triviales ignorées par les utilisateurs mais primordiales pour la bonne qualité de la ressource et à son utilisation par les applications du TAL

Surtout issue des jeux: termes ou consignes difficiles contrainte de temps

Enrichir et consolider le réseau par le remplissage des silences

Déduire des relations manquantes à partir de celles déjà existantes

Approche uniquement endogène

Que peut-on déduire de façon (quasi) certaine ?

Nous avions des schémas

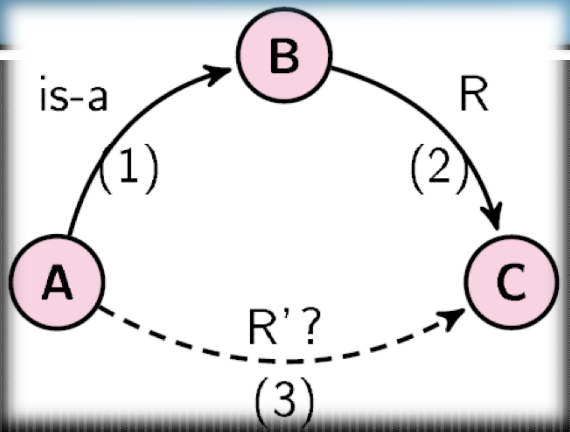


Schéma descendant fondé sur la transitivité de la relation ontologique *is-a*.

générique => spécifique

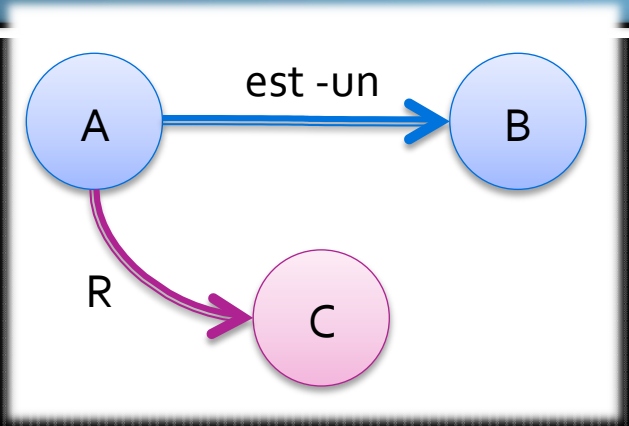
(A est un type de B) et *(B a une relation R avec C)*
alors *(A possède la relation R avec C)*

Les relations (1) et (2) sont les prémisses
La relation (3) est la conclusion proposée et devant être validée

$$chien \xrightarrow{is-a} canidé \wedge canidé \xrightarrow{has-part} croc \Rightarrow chien \xrightarrow{has-part} croc$$

Information en extension

Nous avons des règles



Si A est un type de B
alors
A possède la relation R avec C

Avec B et C instanciés

H : Valide si toujours vraie

(bon ok, … presque toujours vraie)

règles fondées
sur la relation ontologique *est-un*.

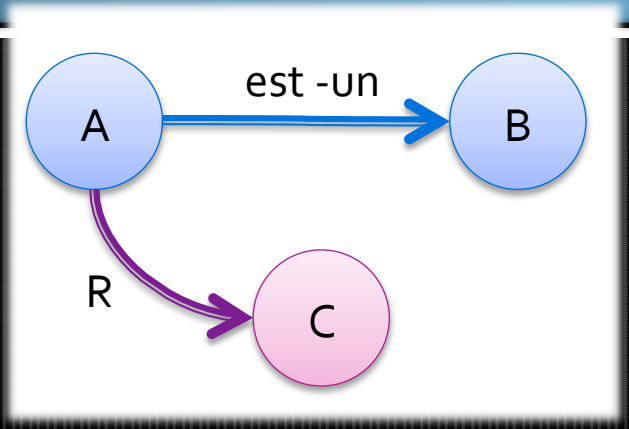
est-un;gastéropode => est-un;animal
est-un;invertébré => est-un;animal

est-un vertébré => est-un animal

est-un;oiseau => a-comme-partie;aile>oiseau
est-un;oiseau => a-comme-partie;bec>oiseau

est-un;vertébré => a-comme-partie;
système nerveux
est-un;vertébré => a-comme-partie;nerf

Nous avons des règles ~ pourquoi ?



Vers de l'extraction de connaissance

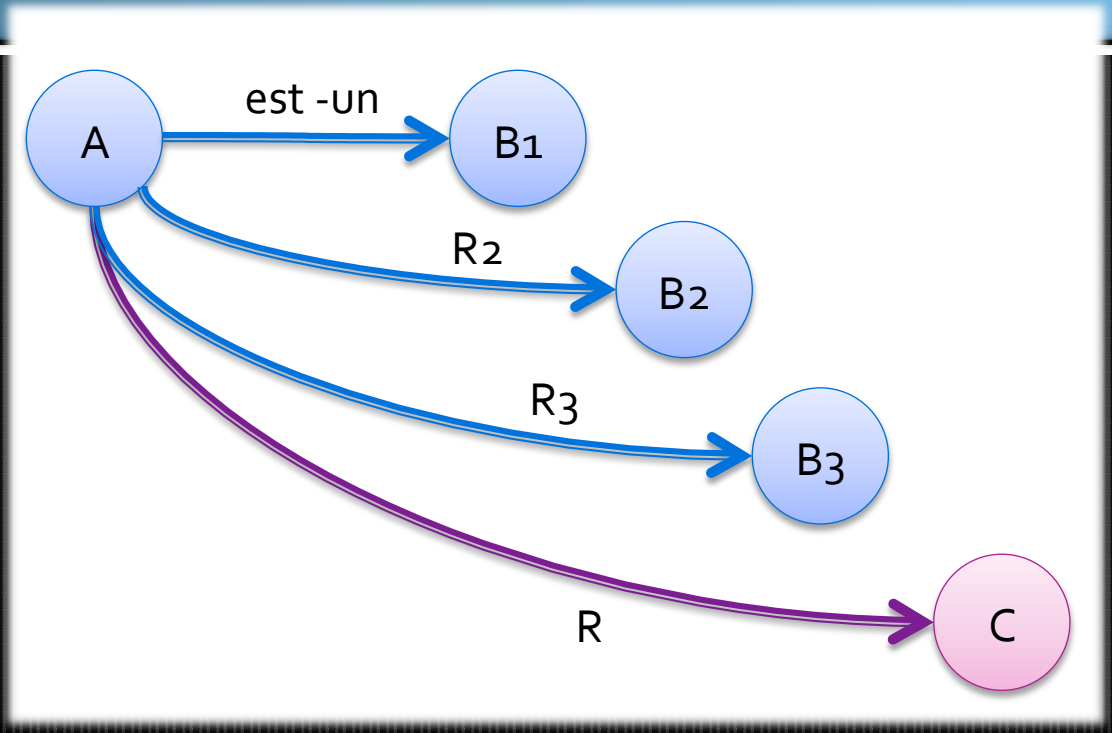
utilisable lors d'une analyse sémantique
(et pas uniquement pour consolider un RLS)

comptables, comparables, fusionnables
Directement compréhensibles

Approche moins générique que les schémas
mais règle qui si validée s'applique toujours
Information en intension

Pour l'instant, pas de validation automatique des règles

Nous avons des règles



est-un;film
& domaine ; cinéma
=> est-un;film > oeuvre

est-un;avion
& contient; passager
& contient ; hôtesse de l'air
=> est-un; avion de ligne

est-un;film > oeuvre
& attribut-deforme;~majusculisé
=> info-sem; entité nommé

Vers de l'extraction de connaissance
avec plusieurs prémisses
(en plus de la relation ontologique)

Comment ? ~ du comptage …

Par abduction sur le réseau JeuxDeMots

=> inférer une règle à partir des exemples d'occurrences existant dans le réseau => exploration massive permanente

```
Pour chaque terme A
    pour chaque B hyperonyme de A
        pour chaque A =R=> C
            +1 pour la règle
                x est-un B alors x =R=> C
```

Pas si couteux que cela, même avec 2 millions de relations et 21 741 hyperonymes – 5 minutes sur ce PC – règles potentielles générées quotidiennement

Comment ? ~ du comptage …

Comptage d'occurrences de relation dans le réseau

- nombre d'occurrences (termes hyponymes pour lesquels la relation existe et est de poids > 0)
- nombre de silences (termes hyponymes pour lesquels la relation n'existe pas)
- nombre de cas faux (relation existante avec poids < 0)
  ~ exceptions (cas faux annotés comme tels)

Comment ? ~ du comptage …

Filtrage sur le poids des relations…
Blocage ontologique (pseudo unification)

- soit *x* =*est-un*=> *A alors x* =*R*=> *C* une règle candidate

- si il existe la règle *x* =*est-un*=> *B alors x* =*R*=> *C*
  et
  A a pour hyperonyme *B* (*A* =*est-un*=> *B*)
  alors
  on ne propose pas la règle

- pour le moment pas de nettoyage
  on souhaite garder la mémoire des règles validées

Combien ?

Produire tout ce qui peut l'être d'un coup... au 21 juin 2013 :

Seuil	Nb de règles	
1	360341	
2	13067	
3	5657	
4	3522	
5	2629	
7	1529	← seuil flottant retenu (pour la production)
10	776	
20	316	

Loi de puissance encore une fois

Valider ou invalider ces règles (les plus productive en premier)

Pas ou pas bien ?

Pour l'instant :

- 1400 règles acceptés
- 470 règles rejetées
- Sur ces 1870 règles, 4 jugées absurdes (99% de règles légitimes)

Validation manuelle par crowdsourcing (niveau meta / JDM)

Choix par le votant/validateur selon l'intérêt pressenti de la règle

- productivité potentielle (nb d'exemples applicables)
- pertinence (nb d'exemples trouvés)
- nb d'exceptions (réelles ou supposées)

Pas ou pas bien ? ~ quelques règles acceptées

métal;6;_INFO-SEM-SUBST;36
matière>substance;6;_INFO-SEM-SUBST;36

bateau;6;coque;9;
bateau;6;coque>bateau;9
navire;6;coque;9;
navire;6;coque>bateau;9

avion;6;aviation;3
avion;6;aéronautique;3

avion;6;prendre l'air>décoller;24
avion>véhicule;6;prendre l'air>décoller;24
avion;6;voler>déplacement aérien;24
avion>véhicule;6;voler>déplacement aérien;24

avion>véhicule;6;pilote;28

Rapidement

Pas ou pas bien ? ~ règles sur des informations symboliques à deux prémisses

personnage;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
groupe industriel;6;_INFO-SEM-ORGA;36;CAPITALIZE-P
groupe industriel;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
institution;6;_INFO-SEM-ORGA;36;CAPITALIZE-P
institution;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P

maison d'édition;6;_INFO-SEM-ORGA;36;CAPITALIZE-P
maison d'édition;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P

ville;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
pays;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
département>territoire;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
cours d'eau;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
montagne;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
province;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P

film>217196;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
oeuvre;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
roman>littérature;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P
bande dessinée;6;_INFO-SEM-NAMED-ENTITY;36;CAPITALIZE-P

Rapidement

Pas ou pas bien ? ~ quelques règles rejetées

être vivant;6;yeux;9
métier>103465;6;yeux;9
métier>103465;6;parler;24
métier;6;parler;24
métier;6;yeux;9
animal;6;patte palmée;9
ingrédient de cuisine;6;mourir;24
métier>profession;6;tête;9
véhicule;6;ciel;15
véhicule>moyen de transport;6;coque;9
animal>zoologie;6;pondre;24
animal;6;comestible;17
animal>zoologie;6;blanc;17

Rapidement

Conclusion et perspectives

Cadre de la recherche de stratégie de consolidation

Construction de connaissances

- du statistique vers le symbolique / procédural
- utilisable lors d'une analyse sémantique de texte
- applicable aux algorithmes de propagation

(règle représentable dans un RLS)

les règles rejetées sont gardées

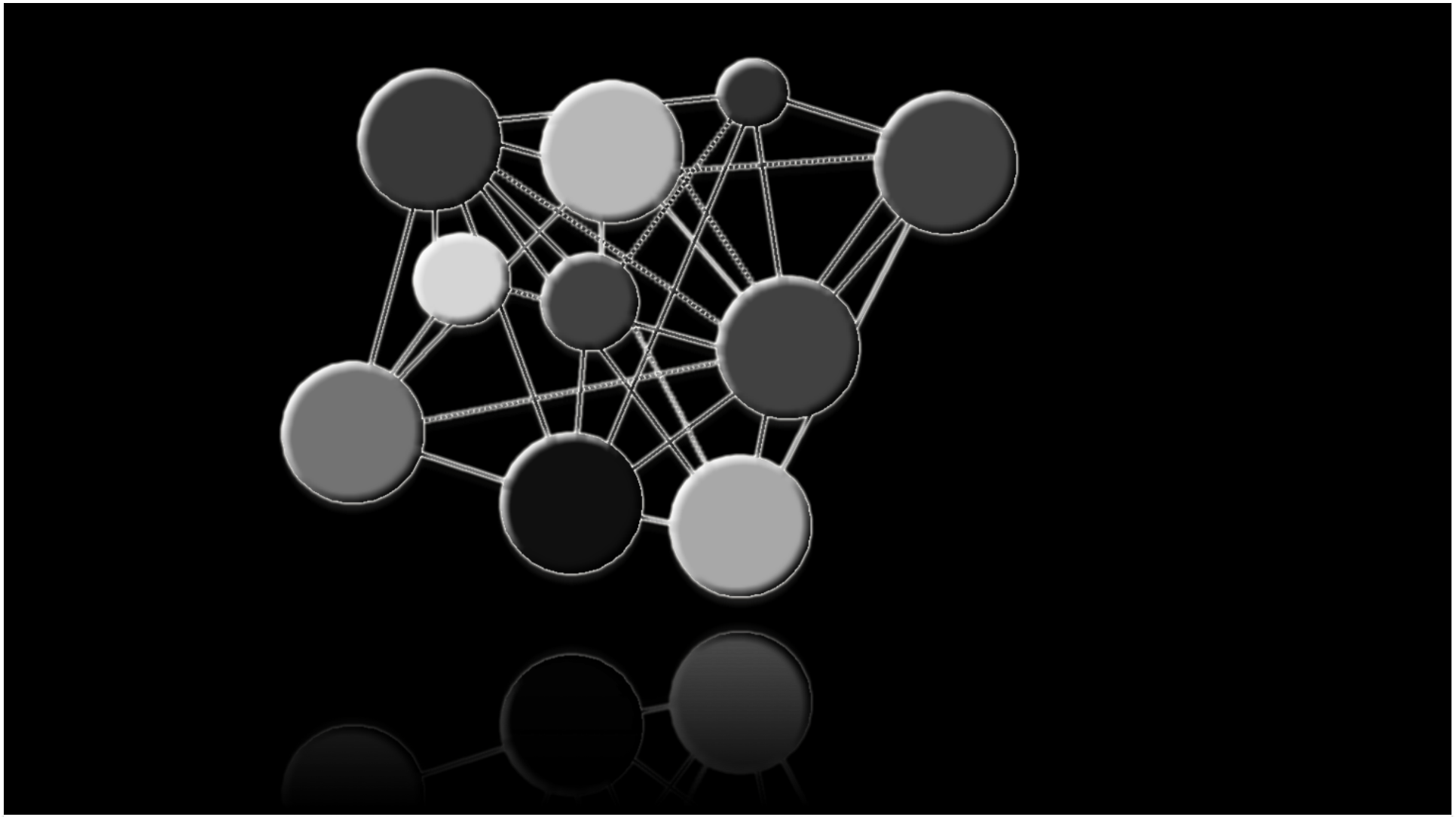
elles ont autant (voire plus) de valeur que les règles validées

Vers la validation automatique de règles ?

en fonction des rapports entre

les exemples, les silences (?) et les cas faux/exceptions

Apprentissage (SVM) sur les règles validées par crowdsourcing



Merci