Searching for Dependencies at Multiple Abstraction Levels

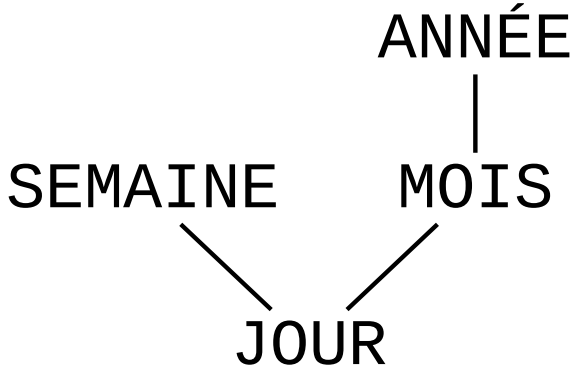
Toon Calders, Raymond T. Ng et Jef Wijsen

Matthieu Objois et Pierre Senellart

Cadre : hiérarchies de généralisation

- Bases de données relationnelles
- Hiérarchie sur les valeurs des attributs (*niveaux*)

EXEMPLE



Dépendance roll-up (DRU)

- Extension de la notion de dépendance fonctionnelle
- EXEMPLE

$$(Ville, \texttt{RÉGION}) \rightarrow (Taux\_Natalité, \texttt{POURMILLE})$$

Cette DRU est vérifiée

⇕

Au sein d'une même région, toutes les villes ont le même taux de natalité arrondi au $1‰$ *le plus proche.*

Plan de l'exposé

1. Introduction

2. Motivations
   - ✓ OLAP
   - ✓ Élimination de redondance
   - ✓ Prédiction de données

3. RUDMINE
   - ✓ Support
   - ✓ Confiance
   - ✓ Gen-schéma

4. Algorithme
   - ✓ Principe
   - ✓ Simulation
   - ✓ Performances

5. Conclusion

OnLine Analytic Processing

- Déterminer la granularité intéressante pour OLAP
- EXEMPLE
  - $(Ville, \texttt{RÉGION}) \rightarrow (Taux\_Natalité, \texttt{POURMILLE})$ *vérifiée*
  - $(Ville, \texttt{PAYS}) \rightarrow (Taux\_Natalité, \texttt{POURMILLE})$ *non vérifiée*

$\Longrightarrow$ données intéressantes : taux de natalité par *région*

Élimination de redondance

- Réduction du nombre de données

  EXEMPLE

  $(Horaire, \texttt{HEURE}), (Station, \texttt{RÉGION}) \rightarrow (Temp°, \texttt{ENTIER})$ *vérifiée*

  $\Longrightarrow$ on peut ne garder que les relevés par *heure* et par *région*

- Aide à la conception de schémas

Prédiction de données

- Découverte de DRU $\Rightarrow$ possibilité de :
  - compléter des données manquantes
  - prévoir de nouvelles données
- Analogue aux techniques classiques en Fouille de Données (règles d'association. . .)

Support d'une DRU

- Caractérise le nombre de cas dans lesquels la règle peut s'appliquer

- EXEMPLE

$$(Ville, \texttt{RÉGION}) \rightarrow (Taux\_Natalité, \texttt{POURMILLE})$$

support de cette DRU : proportion de paires de n-uplets de la base ayant des villes de la même région

Confiance d'une DRU

- Caractérise le nombre de cas dans lesquels la règle est vérifée

- EXEMPLE

$(Ville, \texttt{RÉGION}) \rightarrow (Taux\_Natalité, \texttt{POURMILLE})$

confiance de cette DRU : proportion de paires de n-uplets de la base ayant le même taux de natalité arrondi au $1‰$, parmi ceux ayant des villes de la même région

Schéma de généralisation (Gen-schéma)

- Un gen-schéma d'une relation est un ensemble de couples $(attribut, niveau)$ non redondant.

- On note $G \trianglelefteq H$ si le niveau de chacun des attributs dans H est plus grand que celui dans G.

- EXEMPLE

  2 gen-schémas de $\{(Date, \texttt{JOUR}), (Produit, \texttt{PRODUIT})\}$:

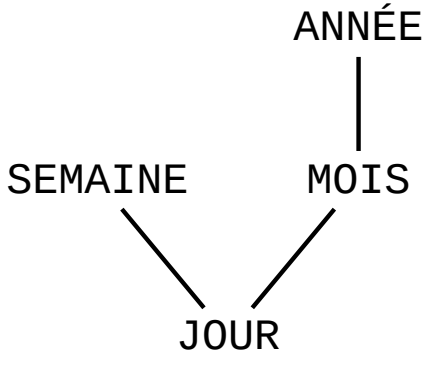
  $$\{(Date, \texttt{MOIS}), (Produit, \texttt{PRODUIT})\} \trianglelefteq \{(Date, \texttt{ANNÉE})\}$$

Principe de l'algorithme

- Parcours de l'ensemble des gen-schémas (qui forment potentiellement la partie gauche de la DRU), du plus général au plus spécifique
- À chaque étape (*strate*), génération des spécialisations immédiates des gen-schémas actuels
- Test du support et de la confiance des candidats, pour retenir ceux qui dépassent les seuils fixés
- Élagage possible

Simulation de l'algorithme

DRU de la forme : $? \rightarrow (Vente, \text{k€})$ $(s_{seuil} = 0{,}25\,;\, c_{seuil} = 0{,}7)$



1. Introduction
2. Motivations
3. Problème RUDMINE
4. Algorithme
   ✓ Simulation
5. Conclusion

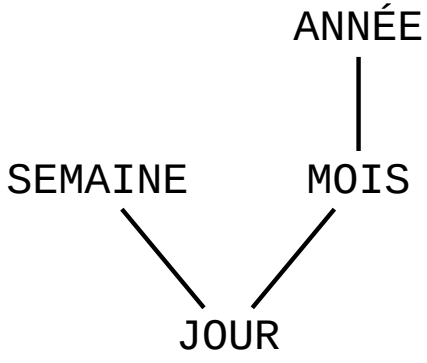
Simulation de l'algorithme

DRU de la forme : $? \rightarrow (Vente, \text{k€})$ $(s_{seuil} = 0{,}25\,;\, c_{seuil} = 0{,}7)$

$s = 1$

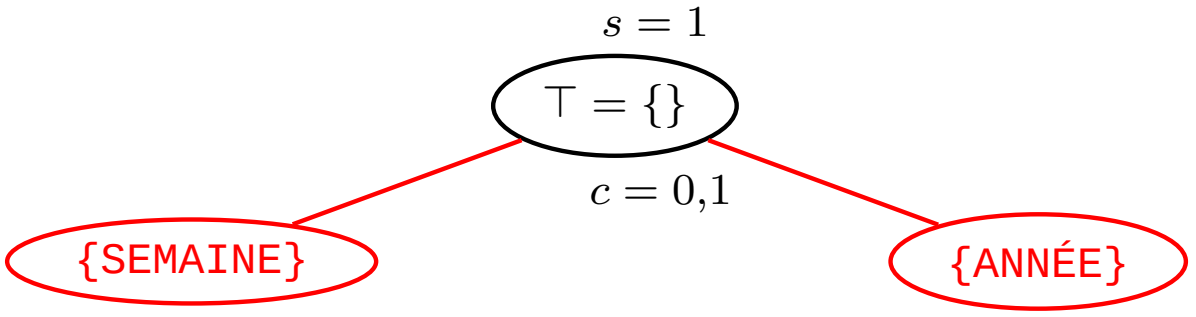
$\top = \{\}$

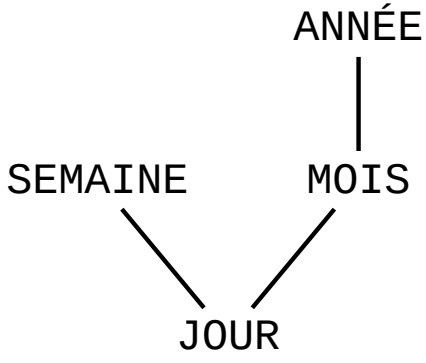
$c = 0{,}1$



Simulation de l'algorithme

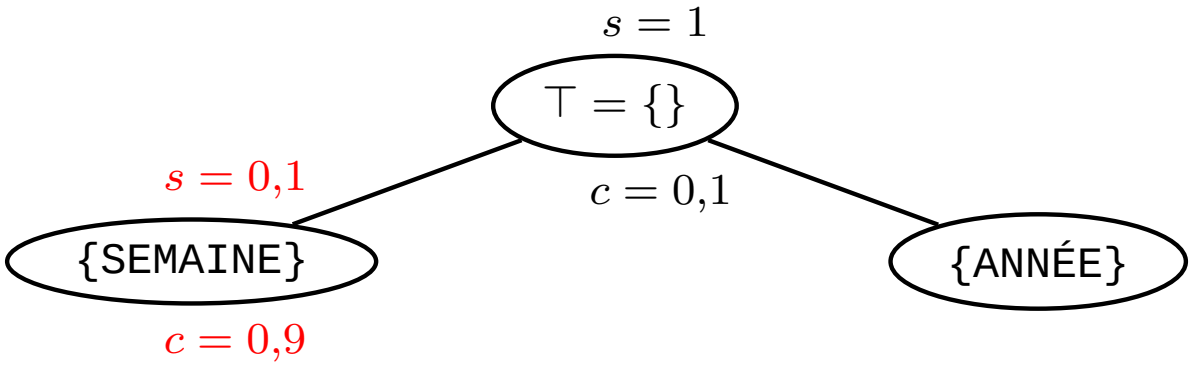
DRU de la forme : $? \rightarrow (Vente, \text{k€})$ ($s_{seuil} = 0{,}25$; $c_{seuil} = 0{,}7$)

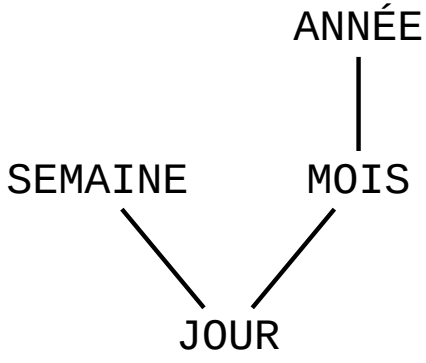




Simulation de l'algorithme

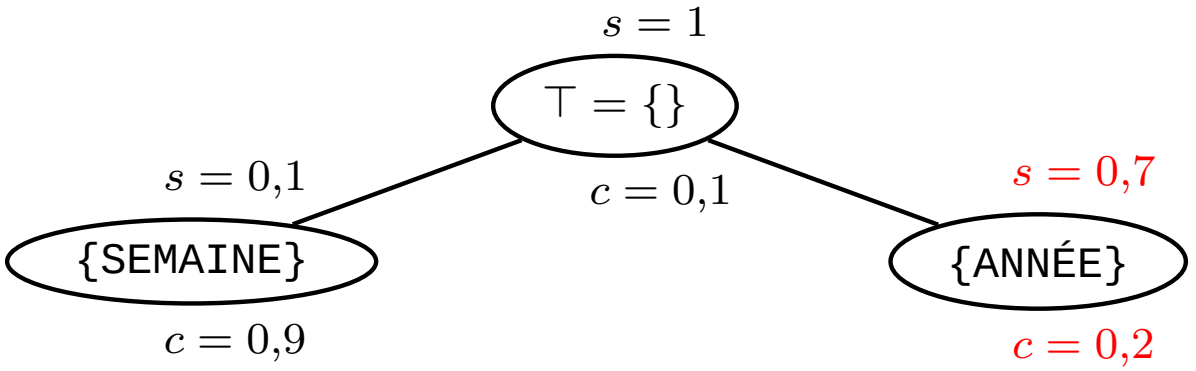
DRU de la forme : $? \rightarrow (Vente, \text{k€})$ $(s_{seuil} = 0{,}25\,;\, c_{seuil} = 0{,}7)$

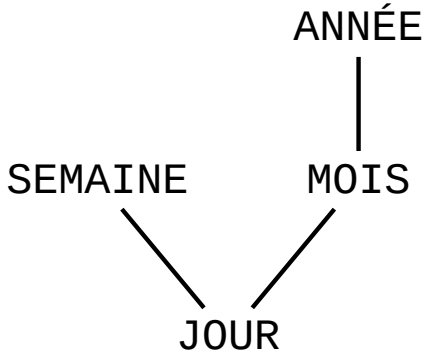




Simulation de l'algorithme

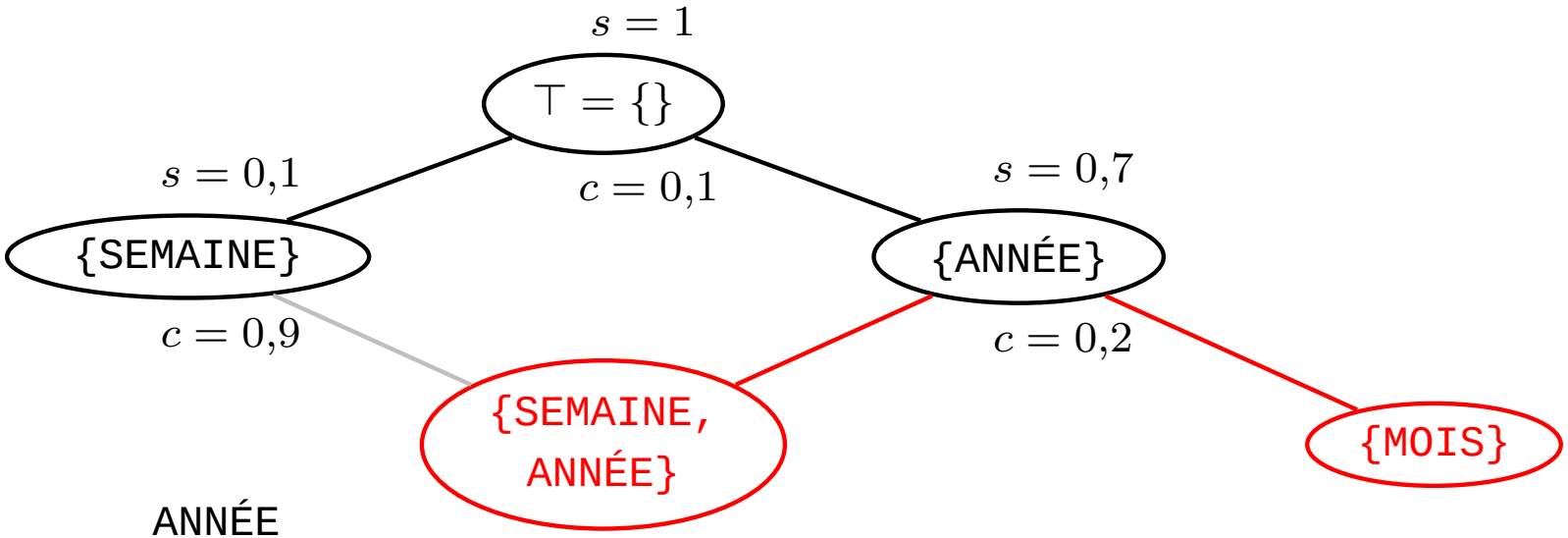
DRU de la forme : $? \rightarrow (Vente, \text{k€})$ $(s_{seuil} = 0{,}25\,;\, c_{seuil} = 0{,}7)$

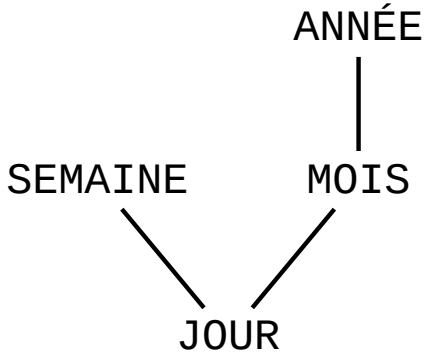




Simulation de l'algorithme

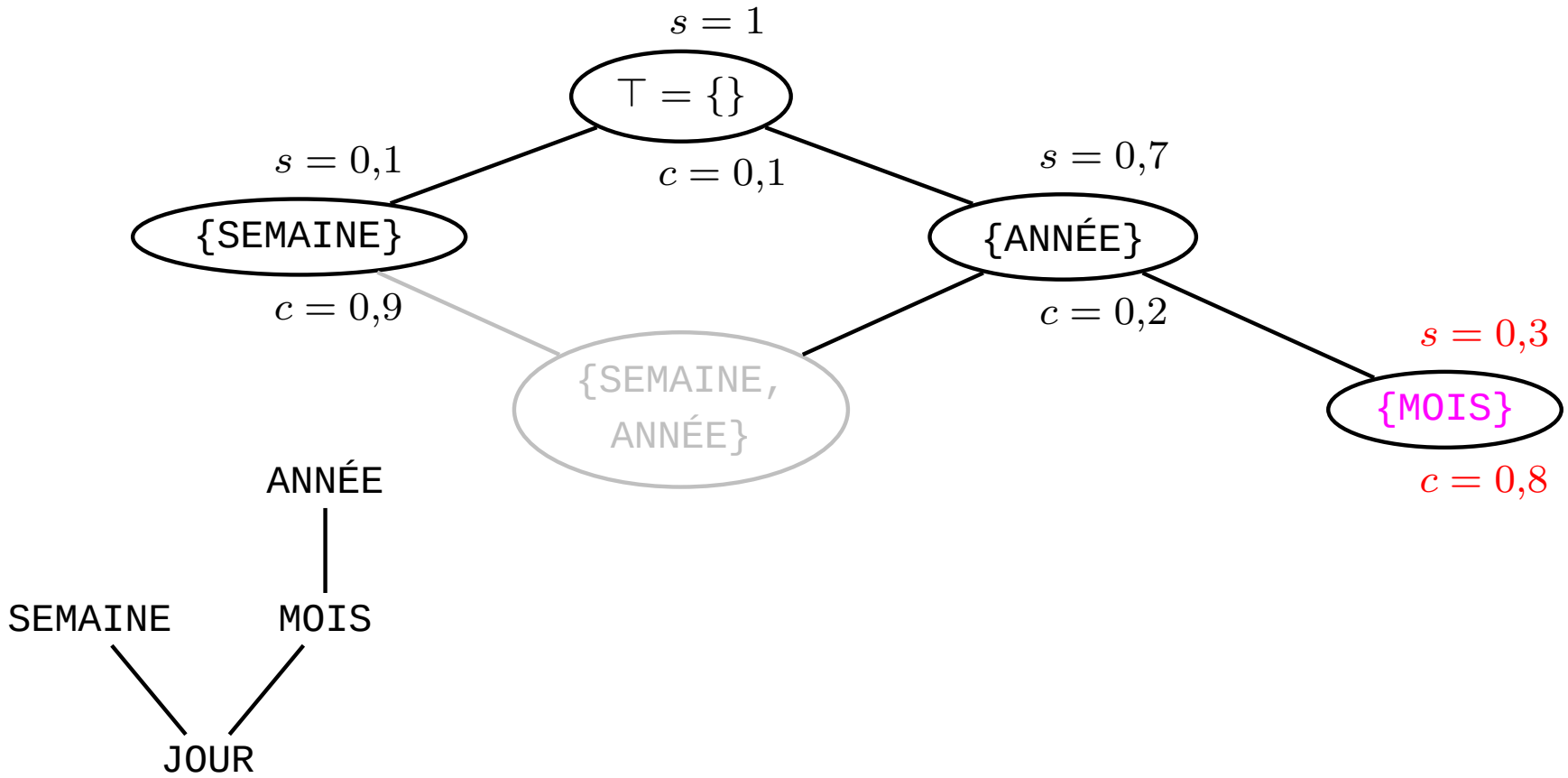
DRU de la forme : $? \rightarrow (Vente, \text{k€})$ $(s_{seuil} = 0{,}25\,;\, c_{seuil} = 0{,}7)$





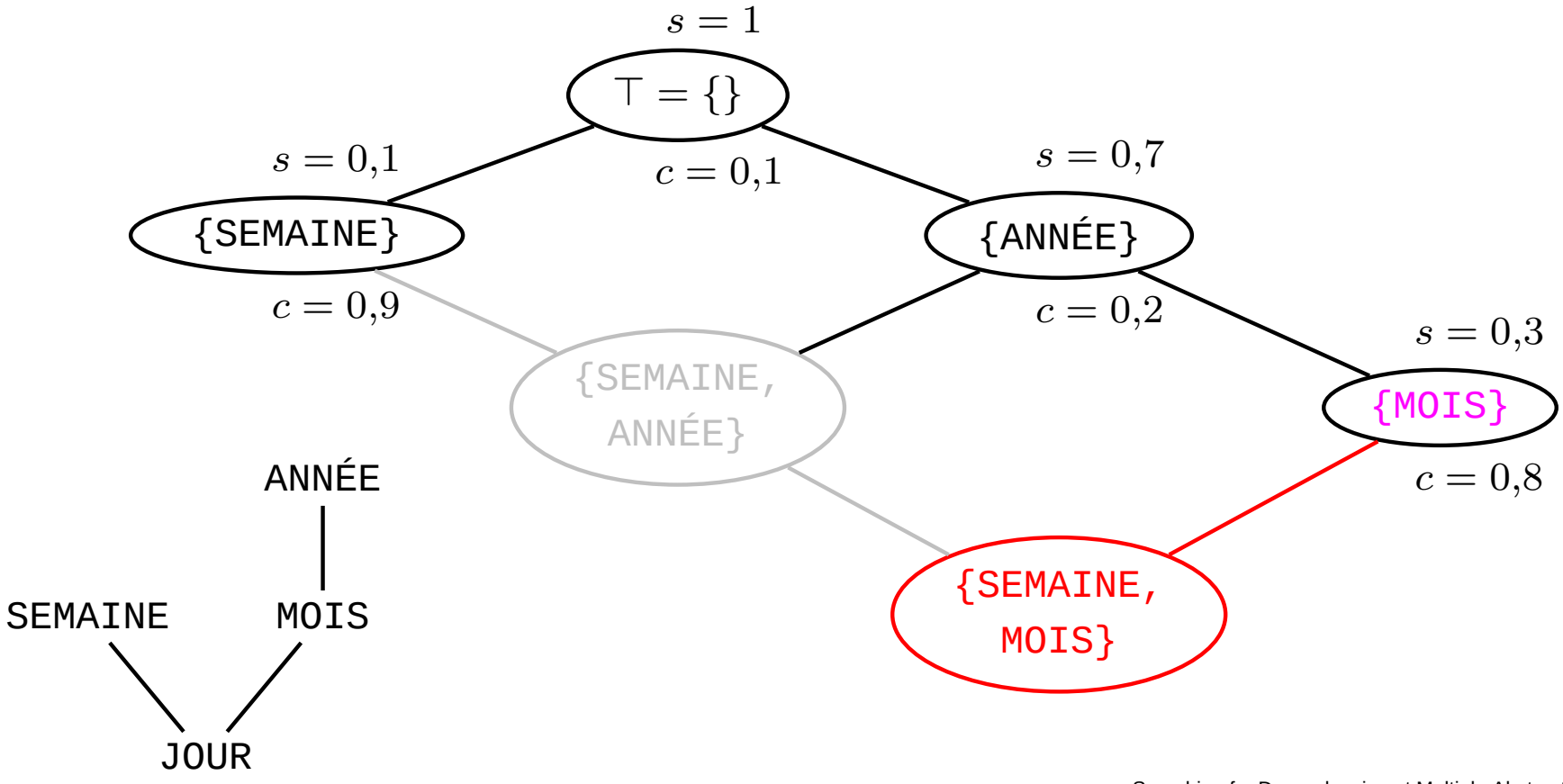
Simulation de l'algorithme

DRU de la forme : $? \rightarrow (Vente, \text{k€})$ $(s_{seuil} = 0{,}25\,;\, c_{seuil} = 0{,}7)$



Simulation de l'algorithme

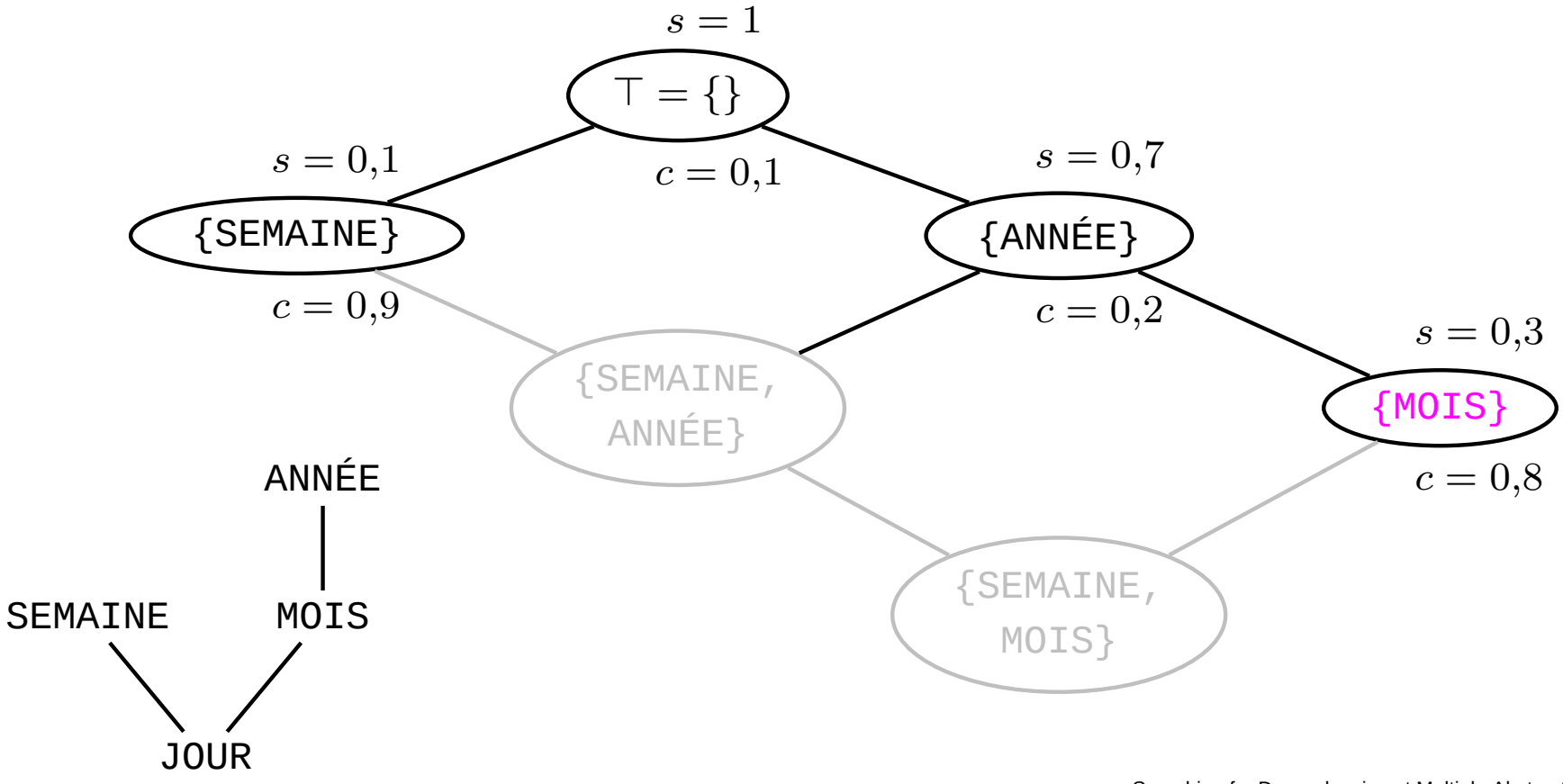
DRU de la forme : $? \rightarrow (Vente, \text{k€})$ $(s_{seuil} = 0{,}25\,;\, c_{seuil} = 0{,}7)$



1. Introduction
2. Motivations
3. Problème RUDMINE
4. Algorithme
   ✓ Simulation
5. Conclusion

Simulation de l'algorithme

$(Date, \texttt{MOIS}) \rightarrow (Vente, \texttt{k€})$



Performances

- Résultats théoriques
  - NP-complet en le nombre d'attributs
  - Linéaire en le nombre de n-uplets
- Performances en pratique
  - Exponentiel en le nombre d'attributs (légère surlinéarité en le nombre de gen-schémas produit)
  - Linéaire en le nombre de n-uplets

En résumé…

- Dépendances roll-ups : extension de la notion de dépendance fonctionnelle
- Utilisation des hiérarchies de valeurs d'attributs pour l'extraction d'informations
- Algorithme efficace

Perspectives

- Généralisation de la notion de DRU (distance entre valeurs)
- Conception de méthode de mise en forme normale d'un schéma de bases de données (extension de l'algorithme de Boyce-Codd)
- Test de l'efficacité des résultats en prédiction