Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup

A. Yeh, L. Hirschman, A. Morgan, *BioInformatics*, **19**, 2003

Marie Guégan

Cours : Fouille de données et apprentissage

M2R Informatique, mars 2006.

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

Introduction

- Les bases de données biologiques
  - Intérêt : réunir et traiter automatiquement les données
  - Origine : articles du domaine
  - Tri des articles et de l'information : **manuel**

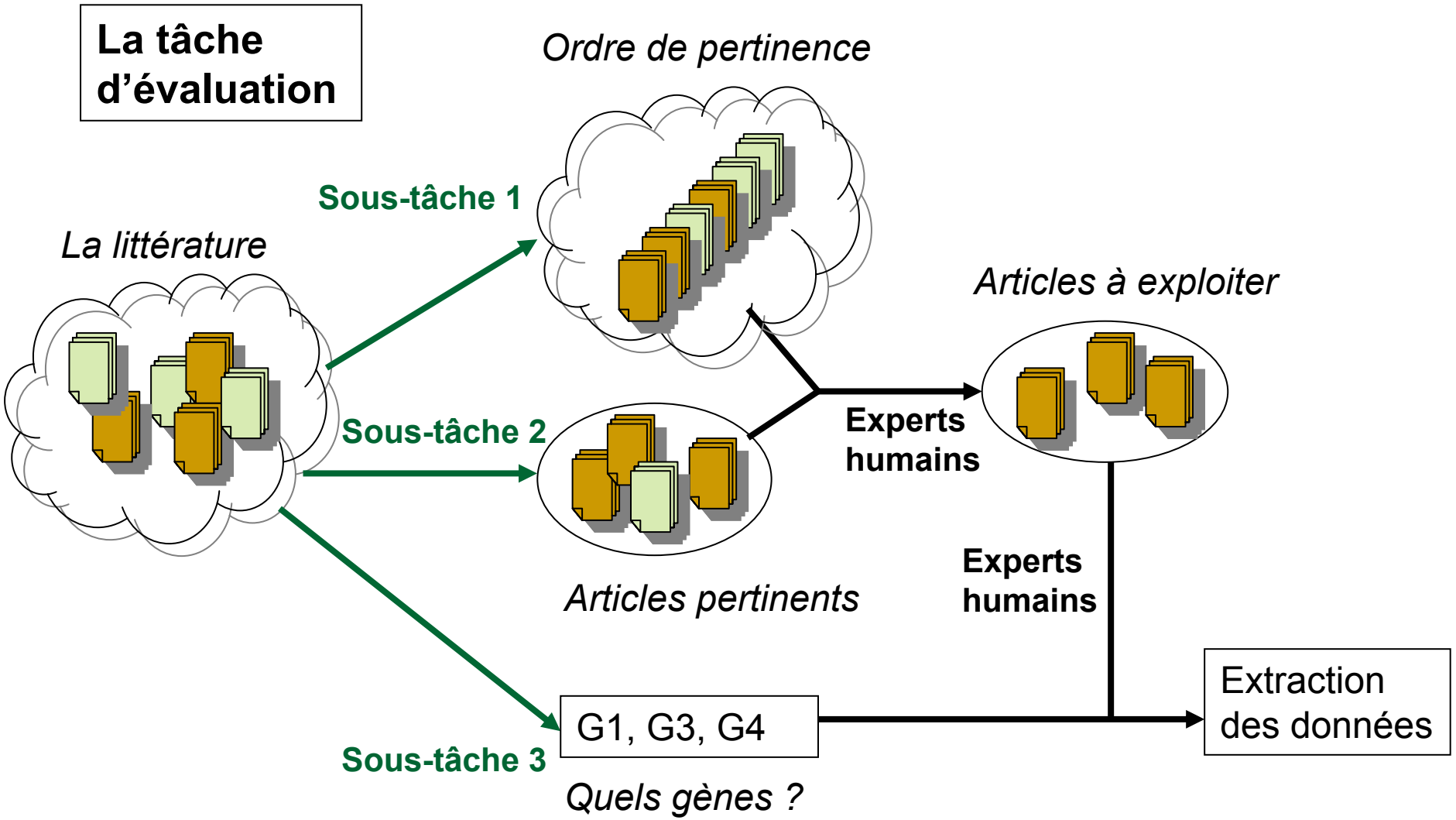
- Le défi des auteurs
  - Surabondance de la littérature ➔ Surcharge manuelle
  - Aides automatiques inspirées de la fouille de données
  - Mais aucun cadre pour les évaluer

La campagne d'évaluation KDD Challenge Cup 2002, Tâche 1

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

1. Élaboration de la campagne



1. Élaboration de la campagne

- Les données
  - Articles fournis par l'équipe de FlyBase, Harvard

- Difficultés rencontrées
  - Gènes : polysémie / synonymie dans le corpus
  - Bruit : variance dans les données fournies par les experts
  - Manque de justification de leur choix par les experts

→ Une tâche finalement simplifiée mais d'utilité indiscutable

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

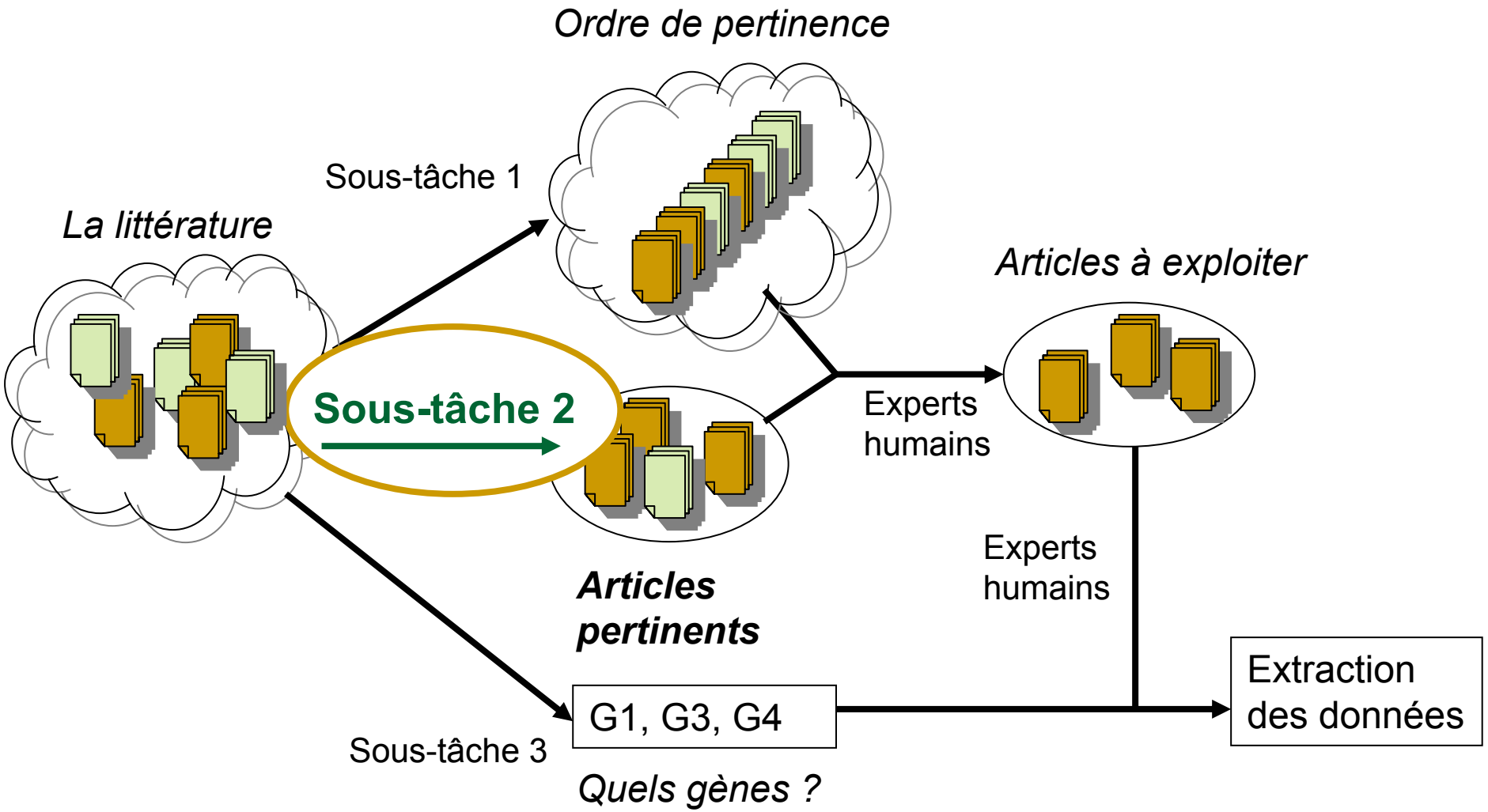
2. Résultats de la campagne

- Evaluation :
  - Courbe ROC
  - F-mesure

- Equipe gagnante (18 équipes) : [Regev et al., 2003]
  - Règles construites manuellement pour reconnaître des motifs
  - Importance des légendes des figures

- Points communs
  - approche « **sac de mots** » écartée
  - des équipes de **statisticiens**, non de biologistes
  - aide par des experts **biologistes**

2. Résultats de la campagne



2. Résultats de la campagne

- Retour sur les données et la sous-tâche 2

  Sous-tâche 2 : cet article est-il intéressant ?

  - Classification oui/non (positifs/négatifs)
  - En grande majorité chez les participants :

    % marqués positifs < % positifs effectifs

    → Des participants très **prudents :**
    tendance à décider en faveur des négatifs

  - Conséquence : négatifs mieux retrouvés que les positifs

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

3. Leçons à tirer de la campagne

1. Préparation de la campagne

- Difficultés à constituer les données
  - Version complètes rarement en libre accès
  - Traitement automatique → choix d'articles HTML (PDF!)
  - Formatage des textes : figures éliminées, parfois avec les légendes
  - …
- Lien trop implicite entre la donnée dans la BD et le passage du texte d'où elle est tirée
  - Nécessité d'experts biologistes
  - Nécessité d'experts de la BD en question

3. Leçons à tirer de la campagne

2. Evaluation de la campagne : objectifs atteints

- Répétition et coût :
  - Données standard constituées en un temps raisonnable et réutilisables
  - Mesure d'évaluation stable et compréhensible
- Attractivité :
  - Pb de biologistes << Equipes de biologistes + statisticiens
- Un facteur de progrès :
  - Tâche à effectuer réaliste
  - Mais qui relève des défis

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

Conclusion des auteurs

- Une campagne réussie
- D'autres tâches en vue
- Souhait : + grande participation des biologistes

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

Discussion et conclusion

- Qualité
  - Un article bien écrit, assez complet
  - Quelques informations redondantes

- Thème
  - Très centré sur la préparation de la campagne
  - Pas d'analyse détaillée des idées des participants
    → cf. [Elkan, 2001] : leçons du CoIL Challenge 2000.

Discussion et conclusion

- Les articles du type « leçons tirées »
  - Prise de conscience récente de leur importance
  - Aujourd'hui, appels à ce type d'articles : *Machine Learning*, **57**, 2004
  - Catégorisation : [Lavrac et al., 2004]
    - Problèmes de l'application de méthodes à des données réelles
    - Apport à la compréhension générale de problèmes irrésolus ← [Elkan, 2001]
    - Etudes de cas bien détaillées ← **[Yeh et al., 2003]**

Plan de la présentation

- **Introduction**
- **1. Élaboration de la campagne**
  - La tâche d'évaluation
  - Données et difficultés des organisateurs
- **2. Résultats de la campagne**
  - Mesure d'évaluation et approches des participants
  - Un retour sur les données
- **3. Les leçons à tirer**
  - Préparation de la campagne
  - Évaluation de la campagne elle-même
- **Conclusion des auteurs**
- **Discussion et conclusion**

Bibliographie

- La campagne
  - FlyBase Consortium (2002)
    The flybase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.,* 30, 106–108.
  - *SIGKDD Exploration newsletter, KDD Cup 2002 (task 1) :*
    - Yeh,A., Hirschman,L. and Morgan,A. (2003)
      Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles
    - Ghanem,M.M., Guo,Y., Lodhi,H. and Zhang,Y. (2003)
      Automatic scientific text classification using local patterns
    - Regev,Y., Finkelstein-Landau,M. and Feldman,R. (2003)
      Rulebased extraction of experimental evidence in the biomedical domain
    - Shi,M., Edwin,D.S., Menon,R., Shen,L., Lim,J.Y.K. and Loh,H.T. (2003)
      A machine learning approach for the curation of biomedical literature

Bibliographie

- Articles complémentaires

  - Hirschman,L., Park,J.C., Tsujii,J., Wong,L. and Wu,C.H. (2002)
    Accomplishments and challenges in literature data mining for biology.
    Bioinformatics, 18, 1553–1561.

  - Elkan, C. (2001).
    Magical thinking in data mining: Lessons from CoIL Challenge 2000.
    In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 426–431.

  - Nada Lavrac, Hiroshi Motoda, Tom Fawcett (2004)
    Editorial: Data Mining Lessons Learned
    Machine Learning, 57, 5-11

  - R. Wirth and J. Hipp. (2000)
    CRISP-DM: Towards a standard process model for data mining. In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining
    (PADD'00), pages 29–39.