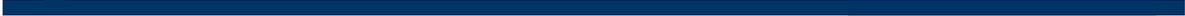


Magical thinking of data mining : lessons
from CoIL challenge 2000
par Charles Elkan



Présenté par Mirija RAKOTONANAHARY

9 mars 2006



Contexte



- ◆ *CoIL - Computational Intelligence and Learning - challenge 2000*
- ◆ *Concours d'apprentissage supervisé pour le data mining*
- ◆ *Analyse les articles lors du concours*

Overfitting (surapprentissage)

- ◆ *formule la variance ($n-1$ au lieu de n)*
- ◆ *Source du problème*
 - *dans un espace très grand*
 - *travailler sur un modèle simple*



Sélection d'attribut

- ◆ *Heuristique*
- ◆ *Interaction entre attributs*
- ◆ *Nouveaux attributs (produit croisé)*

Comparaison des méthodes

- ◆ *Null hypothesis (2 premiers)*
- ◆ *Hypothèse de McNemar $((|n_{10}-n_{01}|-1)^2/n_{10}+n_{01})$*



Raisonnement magique



- ◆ *Raisonnement non probabiliste*
- ◆ *Méthode idéale*
- ◆ *Méthode mal préparée*



Choix de méthodes

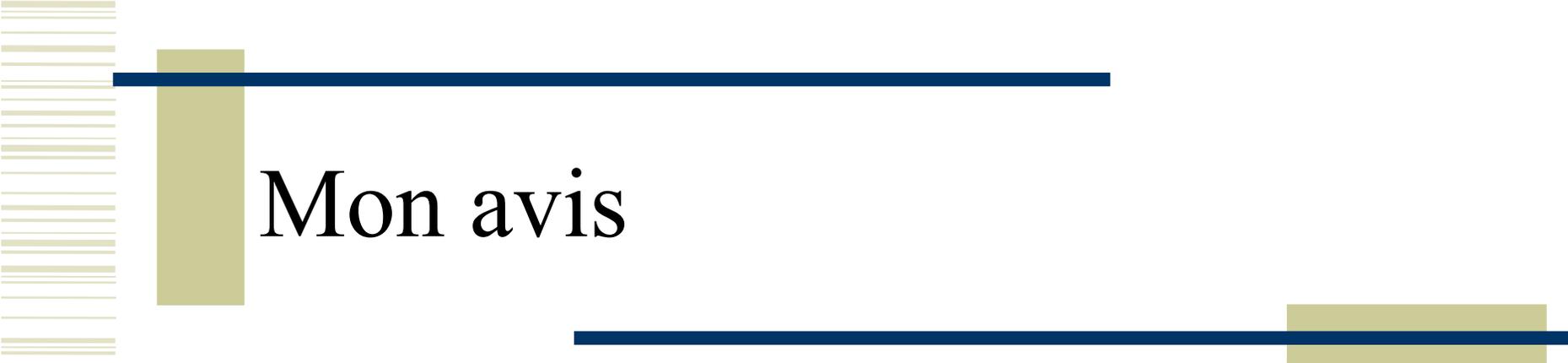


- ◆ *Ne pas oublier les contraintes (temps, flexibilité)*
- ◆ *Connaître les méthodes*
- ◆ *Méthode mal préparée*



Reformulation

- ◆ *Poser le concours autrement*
- ◆ *CRISP-DM (Cross Industry Standard Process for Data Mining)*
- ◆ *Méthode existante*
- ◆ *Interaction entre attributs + boosting => amélioration*
- ◆ *Oubli des significations statistiques*



Mon avis

- ◆ *Les méthodes les plus simple => bon résultat*
- ◆ *« Meta article »*