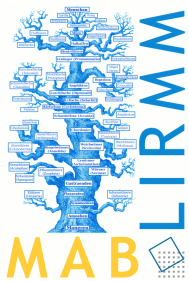
Data Mining et Big Data

Eric Rivals

LIRMM & Inst. de Biologie Computationnelle
CNRS et Univ. Montpellier

14 novembre 2015



Plan

Data Mining & Big Data

Data Mining ou « fouille de données »

Extraction de connaissances à partir de (gros volumes) de données

Big Data

- Big Data $\simeq$ Predictive Analytics $\simeq$ Data Science
- Big Analytics ou Broyage de données
- Masse de données

Domaine central d'applications

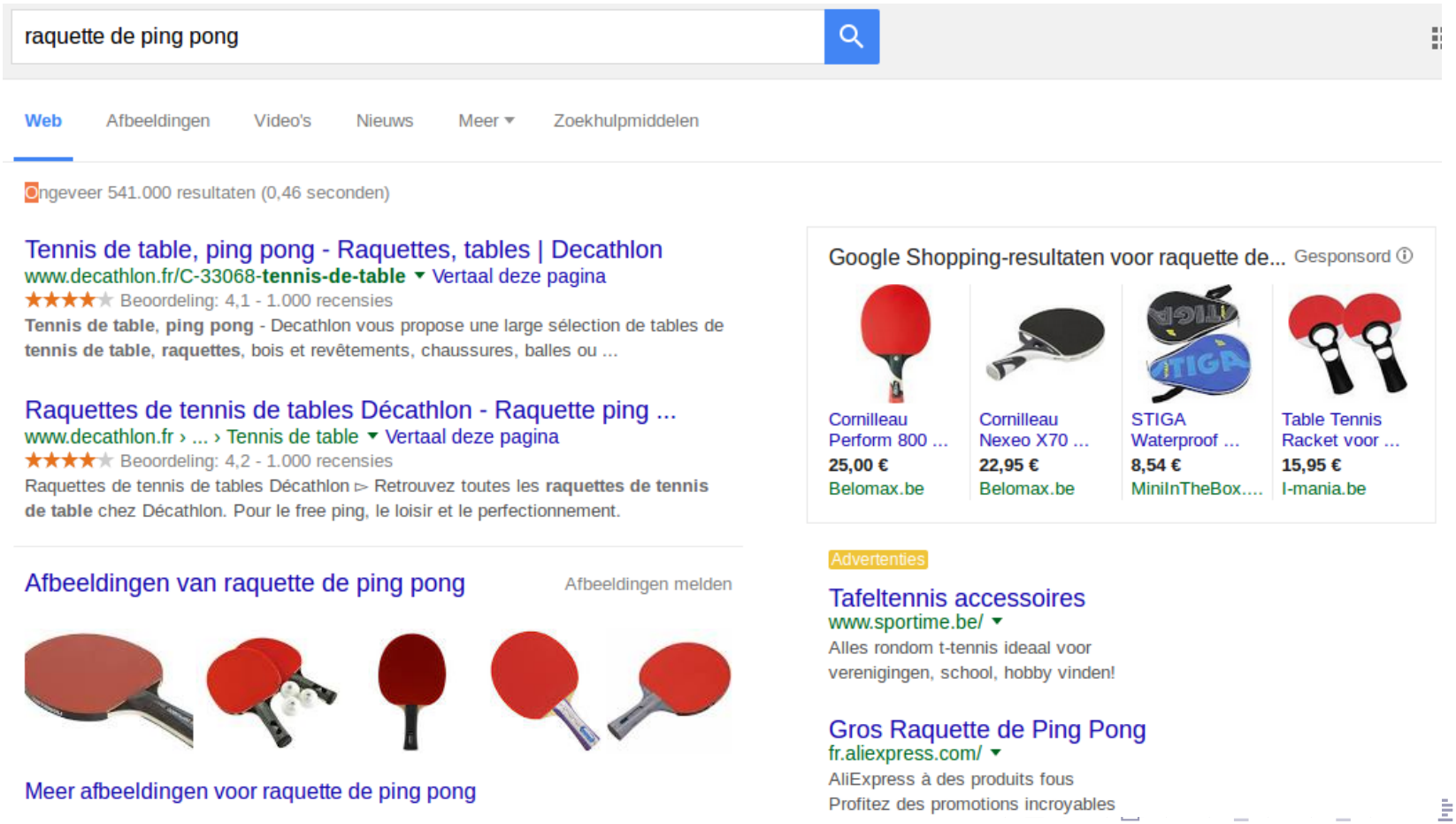
Internet, fouille des comportements, des profils

- Proposition d'annonces publicitaires
- Identification de cibles commerciales

Domaine central d'applications

Internet, fouille des comportements, des profils

- Proposition d'annonces publicitaires
- Identification de cibles commerciales



La puissance de la fouille de données ?

nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

La puissance de la fouille de données ?

nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Automatique, rapide, toujours active

Deux types d'approches

Descriptives

résumer les données (nuages de mots, statistiques, visualisation)

Deux types d'approches

Descriptives

trouver des motifs interprétables qui décrivent les données

Deux types d'approches

Descriptives

trouver des motifs interprétables qui décrivent les données

Prédictives

prédire certaines variables ou attributs en fonction d'autres variables ou attributs

Deux types d'approches

Descriptives

trouver des motifs interprétables qui décrivent les données

Prédictives

prédire certaines variables ou attributs en fonction d'autres variables ou attributs

Remarque

dans les deux cas : modélisation exploratoire

Exemples d'applications en médecine [P. Tufféry, Technip, 2010]

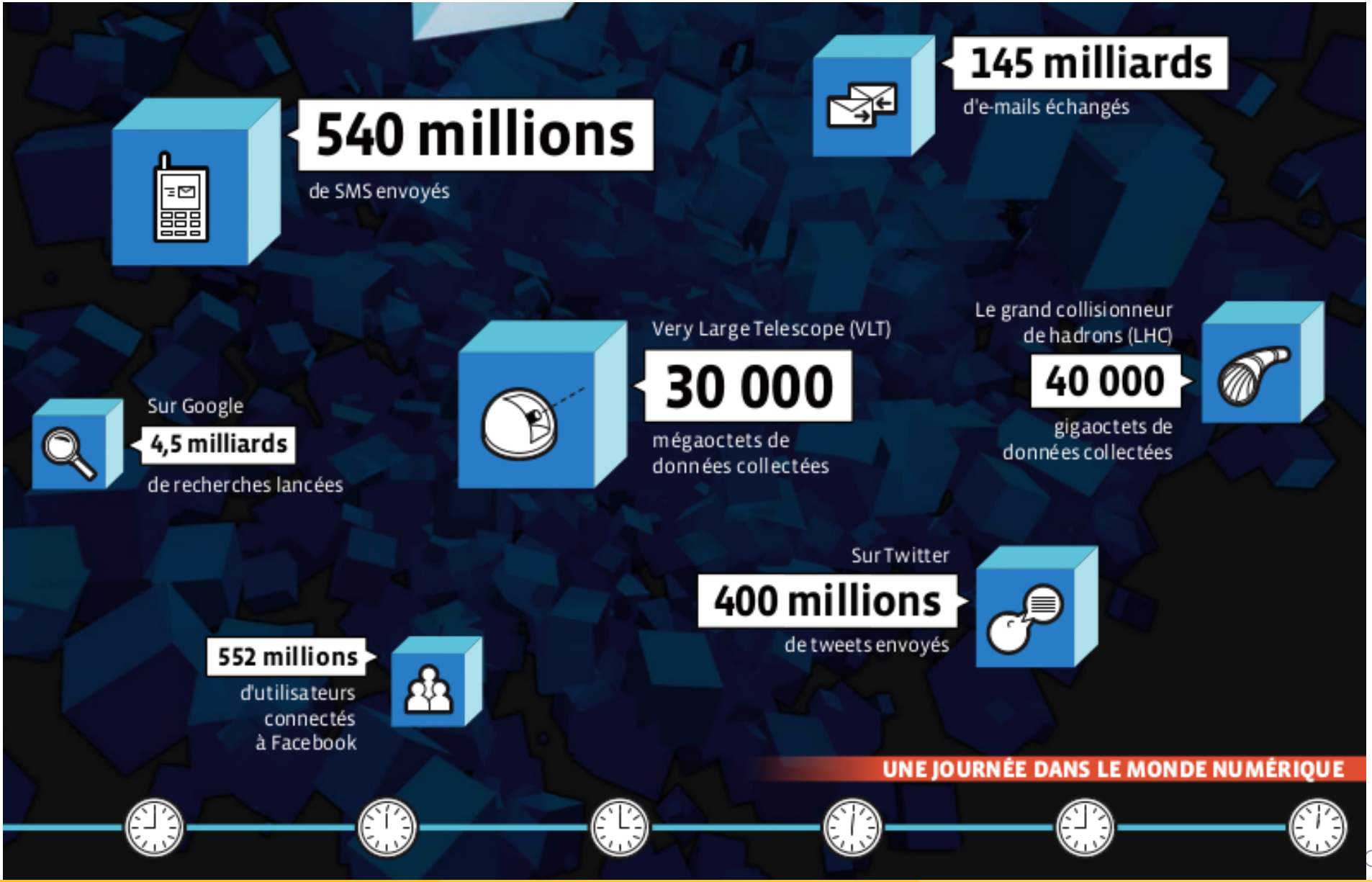
Applications descriptives

- groupes de patients réagissant de façon similaire à un traitement thérapeutique
- association de médicaments et effets secondaires

Applications prédictives

- des facteurs de décès / survie pour une pathologie à partir de données d'essais cliniques
- d'anomalies sur une échographie
- du temps de rétablissement après une opération
- fonction d'un gène / protéine par recherche de similarité entre séquences
- reconnaissance de tumeurs
- identification des erreurs de diagnostic

Raison de l'existence des Big Data [Journal CNRS n°269]



Raisons de l'existence des Big Data

- Accroissement phénoménal des volumes de données (+40 % par an)
  ex : 30 milliards de contenu sur face book chaque mois
- Internet : échanges et comportements enregistrés sous forme d'informations digitales
- Accroissement de la surveillance : capteurs, vidéo, tél. portables
- La masse d'information à explorer permet de déceler des événements ou phénomènes plus cachés
- Diminution prix stockage
- Diminution prix du calcul : distribution calcul

Les 4 V du Big Data

Volume - Variété - Vélocité - Validité

- Volume
  empêche l'utilisation d'outils classiques de calcul (ex : tableur)
  exige une répartition sur plusieurs mémoires, disques, calculateurs
- Variété
  hétérogénéité de forme, de localisation, de codage
  besoin de procédures adaptées au type de données : image, texte, valeurs
- Vélocité
  les données peuvent arriver en flux : vidéo, pages ou requêtes oueb
- Validité
  toutes les informations ne sont pas valides : erreurs, bruit, spam

Croisement ou intégration de données variées : cas médical

- Réseaux sociaux
- Données publiques ≪ open data ≫ (pollution, climat, bruit, etc.)
- Dossier patient
- Radiographies, scanner, . . .
- Rapport d'analyse
- Génome, transcriptome, épigénome

Comment combiner les informations aussi différentes ?

Data Mining est pluridisciplinaire

Mélange disciplines

- Statistiques
- Analyse de données (Statistique exploratoire)
- Base de données
- Apprentissage automatique (Machine learning), IA
- Informatique théorique, optimisation, algorithmique

Attention est portée au

- Passage à l'échelle (scalability)
- Algorithmes

  optimisation temps de calcul et mémoire
- Architectures et modèles de calcul

  distribuée : Map-Reduce, algorithmes on-line pour les flux
- Automatisation du traitement des masses de données

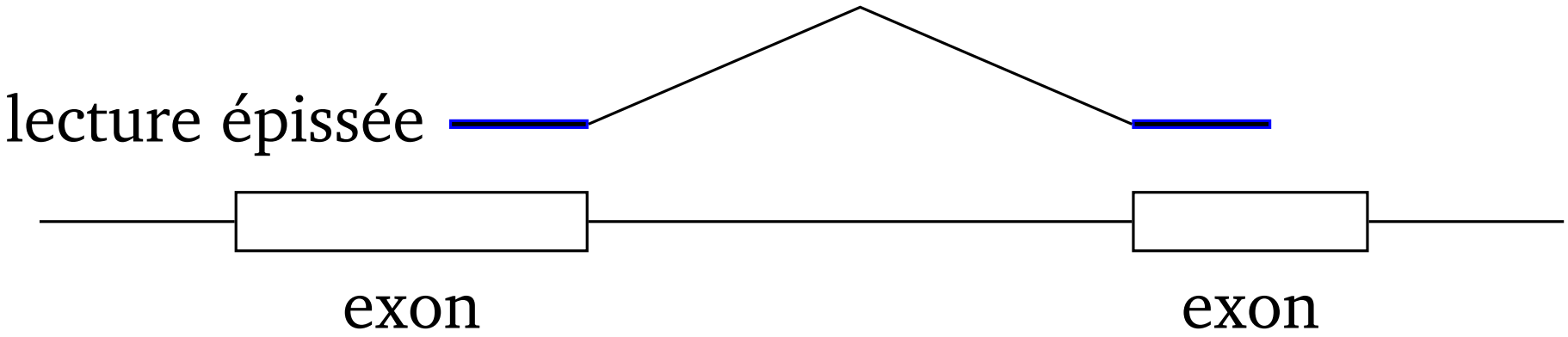
Finalités et méthodes d'analyse de données

- Description : trouver un résumé des données qui soit plus intelligible
  méthode : analyse factorielle - stat. descriptive
- Structuration : faire ressurgir des groupes « naturels » qui représentent des entités particulières
  méthode : classification
- Association : Trouver les ensembles de descripteurs qui sont les plus corrélés
  méthode : règles d'association
- Explication : Prédire les valeurs d'un attribut (endogène) à partir d'autres attributs (exogènes)
  méthode : apprentissage supervisé

Plan

Localiser des lectures d'ARN sur la séquence d'un génome

But : trouver tous les alignements des lectures sur le génome

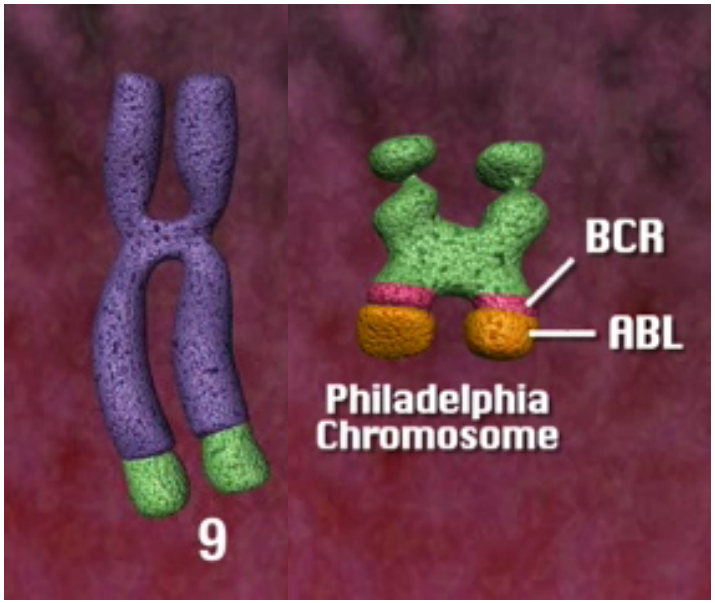


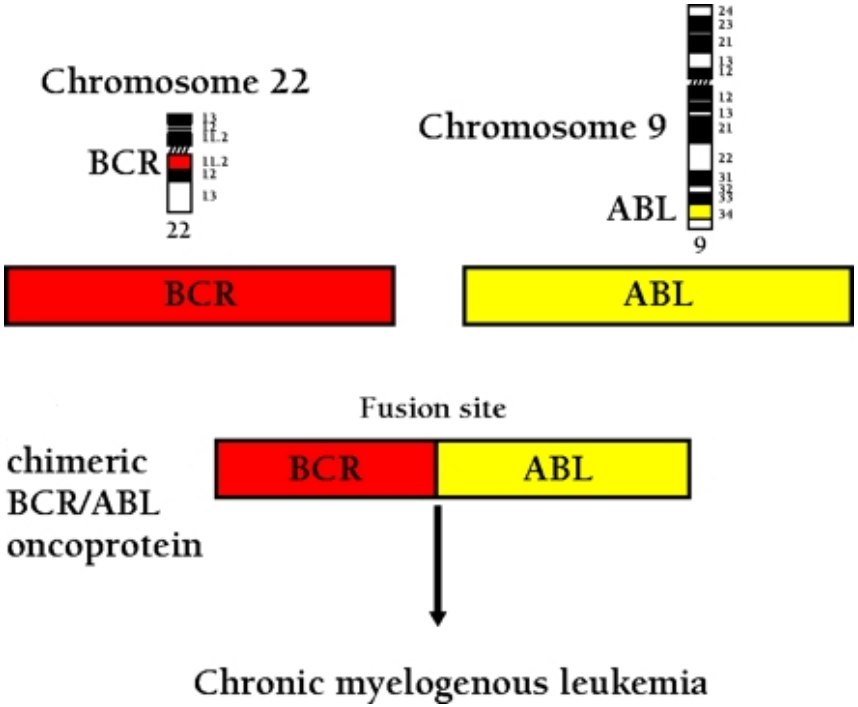
Exemple de volumes de données transcriptomiques

Transcriptome du Maïs

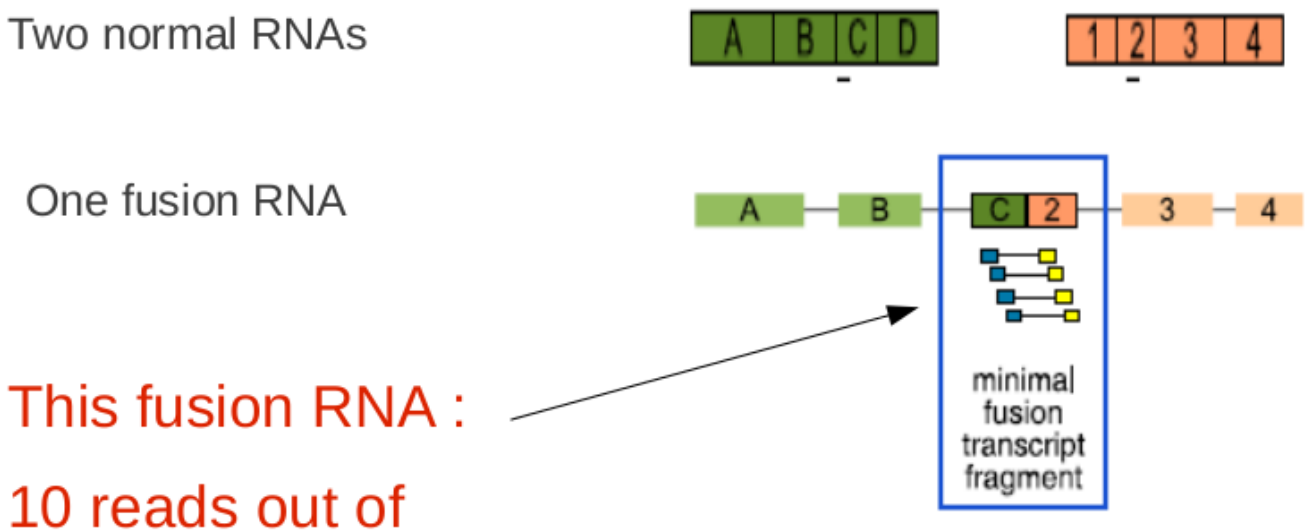
- Illumina HiSeq : 194 millions de lectures, 29 Tera - pb
- PacBio : 276000 lectures, 168 Giga - bp
- Correction d'erreurs avec LoRDEC $\simeq$ 10 heures et
- en memoire : 5 Giga octets

Détection d'ARN de fusion comme marqueurs tumoral



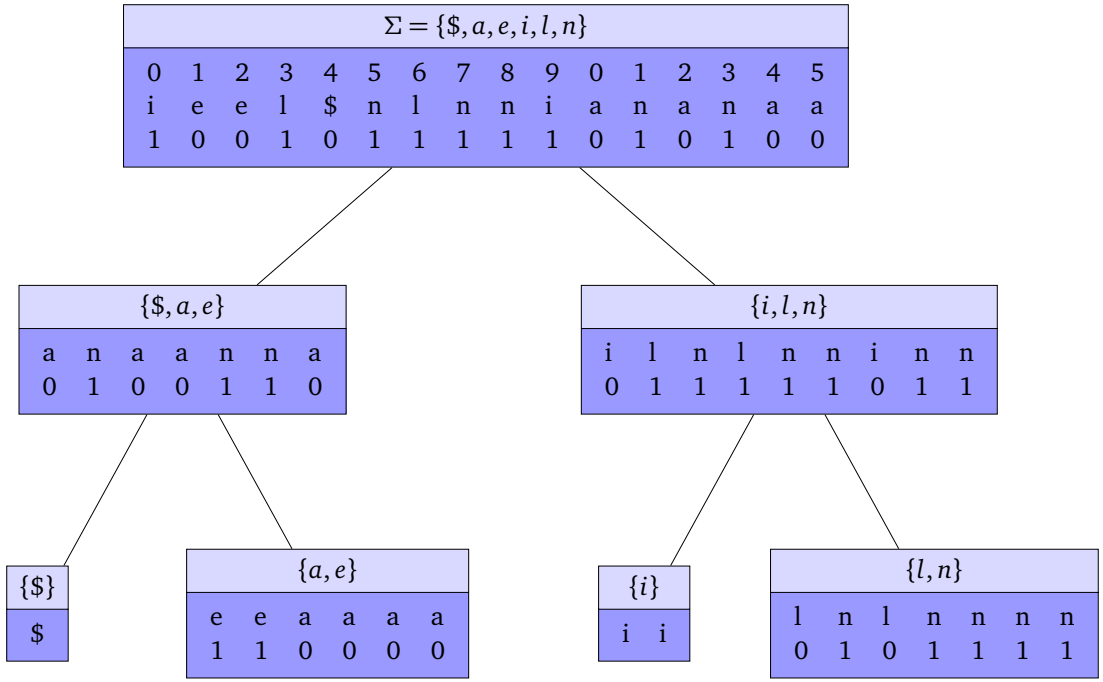


Besoin d'allier sensibilité et spécificité



Secret : structures d'indexation compressées

Index : Structure informatique en mémoire qui organise les éléments d'information d'un texte, d'une base de données pour un accès rapide.



Transformée de Burrows-Wheeler sous forme d'arbre à ondelettes

Plan

Formulation : ≪ item-sets ≫

- Chaque patient : un ensemble d'éléments dans son dossier
- éléments : traitements/médicaments et effets secondaires
- Item-set : associations potentielles entre médicaments et effets secondaires

Les régles d'association indiquent des relations fréquemment observées.

$$\{e_1, \ldots, e_i\} \Rightarrow X$$

Critères numériques

- support : nb de patients ayant ce sous-ens. d'éléments
- confiance : rapport support entre $\{e_1, \ldots, e_i\}$ et $\{e_1, \ldots, e_i, X\}$
- Intérêt de la règle :
  différence entre la confiance et la probabilité de l'élément prédit

Difficulté algorithmique

Objectif : calculer tous les item-sets avec un nb d'occurrence $>$ *seuil*

Le calcul de paires fréquentes est le goulot d'étranglement.

Observations

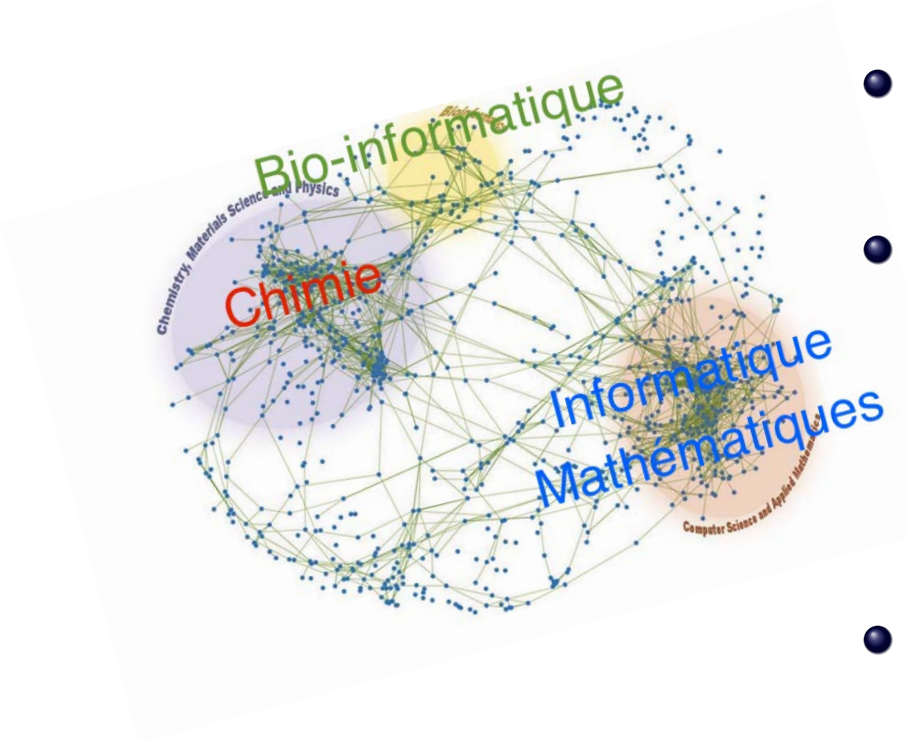
- si un ens. X est fréquent alors tout sous-ens. de X l'est aussi
- si un sous-ens. Y n'est pas fréquent, alors aucun sur-ensemble de Y ne peut l'être

Algorithmes

1. à base de fonction de hachage
   limitation mémoire par rapport au nb d'éléments
2. algorithme « A priori »
   mémoire linéaire par rapport au nb d'éléments

Plan

Recommandation : défis



[Servajean et al. 2013]

- Communauté en ligne partage des données à gde échelle
- utilise données transformées pour répondre à des requêtes pluridisciplinaires
  Ex : `modèle mathématique pour la croissance des plantes`
- Diversité des documents : prendre en compte les données de différentes disciplines (informatique, biologie, . . .)

Recommandation : incorporer confiance et diversité

- une approche probabiliste originale [Servajean et al. 2013]

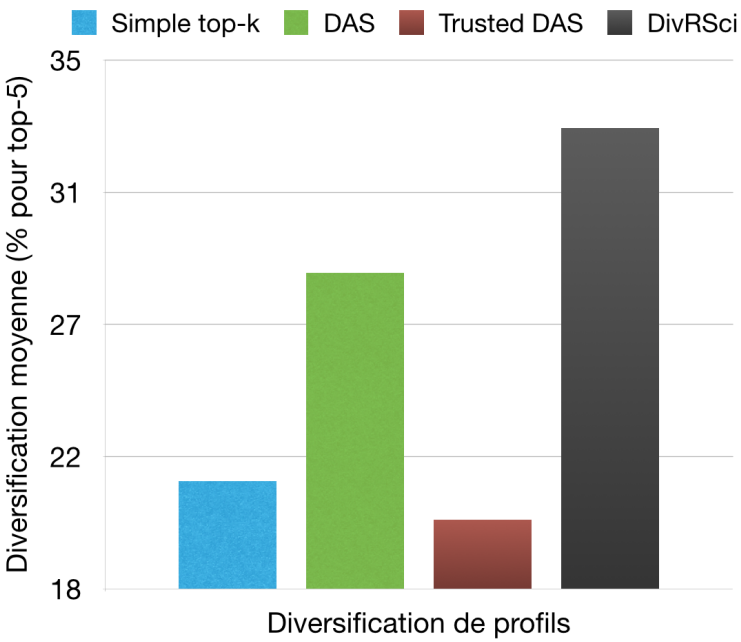
$$div_p(u_d|\{u_{d_1},...,u_{d_{i-1}}\}) = \frac{1}{N}.\sum_{v_n \in u_{d_i}} [\underbrace{rel_{trust}(v,u,q)}_{trust}.\underbrace{\prod_{v_m \in \{u_{d_1},...,u_{d_{i-1}}\}}(1-red_p(v_m|v_n)}_{Profile\ Diversification}]$$

Recommandation : incorporer confiance et diversité

- une approche probabiliste originale [Servajean et al. 2013]

$$div_p(u_d|\{u_{d_1},...,u_{d_{i-1}}\}) = \frac{1}{N}\cdot\sum_{v_n\in u_{d_i}}\Big[\underbrace{rel_{trust}(v,u,q)}_{trust}.\underbrace{\prod_{v_m\in\{u_{d_1},...,u_{d_{i-1}}\}}(1-red_p(v_m|v_n))}_{Profile\ Diversification}\Big]$$

- et des optimisations de calcul



Résultats

- montrent les gains en diversité de profils sur un jeu de données INRA
- génériques et validés par un benchmark du web
- validés avec des retours utilisateurs : 70% de satisfaction pour les requêtes interdisciplinaires

Plan

Conclusions

- ≪ Big Data ≫ viendra à la médecine

  Potentiel énorme d'applications

- ≪ Open Data ≫ à l'hôpital ?

- Aspect privé des données, sécurité et éthique

  Identification grâce aux traces des données individuelles

- Contrôle des modes de requêtes des moteurs de recherches

Références – Data Mining

- Data mining et statistique décisionnelle : l'intelligence des données
  Stéphane Tufféry, Editions TECHNIP, 3è édition

- Principles of Data Mining
  David Hand, Heikki Mannila, and Padhraic Smyth.
  MIT Press, Cambridge, MA, 2001.

- Analyse de données
  J.M. Bouroche et G. Saporta
  Collection ≪ Que sais-je ? ≫
  Presses Universitaires de France

Références – Bioinformatique et biologie

- *LoRDEC : accurate and efficient long read error correction*

  L. Salmela, E. Rivals

  Bioinformatics, doi:10.1093/bioinformatics/btu538, 30 (24) : 3506-3514, 2014.

- Detecting influenza epidemics using search engine query data

  J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant

  Nature Vol 457. 19 February 2009