TELECOM ParisTech

INSTITUT Mines-Télécom

Big Data : enjeux et usages

Talel.Abdessalem@telecom-paristech.fr

université PARIS-SACLAY

Chaire Big Data & Market Insights

- **Chaire de recherche**
- **Partenaires académiques**
  - Telecom ParisTech
  - Telecom Ecole de Management
- **Sponsors :**

GROUPE ROCHER GROUPE BPCE SNCF Deloitte.

- **Démarrage : 2014**
- **Durée du projet : 5 ans**

Principaux axes de recherche

- **Les données massives**
  - Crawl et extraction de données du Web, analyse de réseaux sociaux, optimisation…
- **L'analyse prédictive**
  - Les systèmes de recommandation, détection d'intrusion et de fraude, prédiction de risque …
- **Marketing Digital**
  - Segmentation de clients et évaluation des actions de marketing

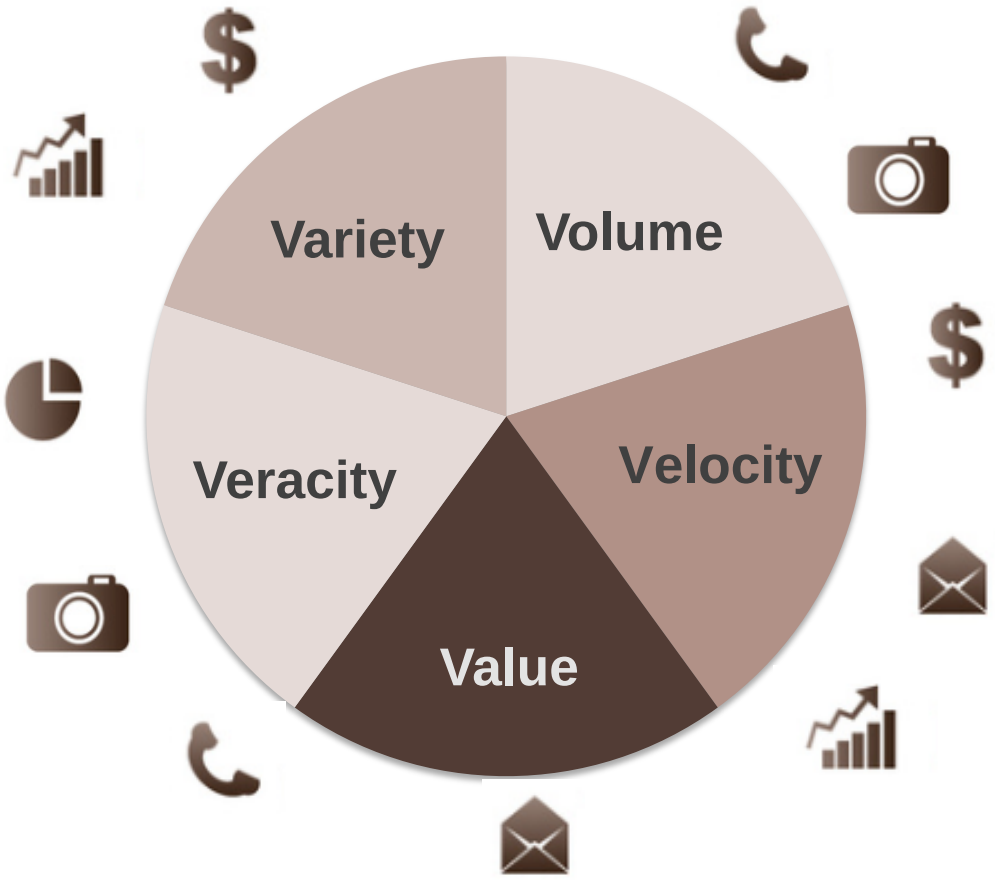
Plan

- **Qu’est ce que le Big Data ?**
- **Importance et défis pour l’entreprise**
- **Exemples de travaux et résultats**

Qu’est ce que le Big Data et quels sont les principaux usages ?

Big Data ?

Ensemble de défis souvent décrits par les 4/5Vs



Importance et défis pour les entreprises ?

Importance du Big Data

- **Facteurs déterminants**
  - Rupture technologique initiée par les géants du Web
  - Disponibilité des données
  - Prise de conscience de la valeur des données
  - Marchés de plus en plus compétitifs

- **Phénomène de mode ?**
  - Peut-être pour certaines niches sans concurrence
  - Certainement pas pour les autres…

Big Data & Transformation Digitale

- **Transformation digitale**
  - Elle est déjà là, dans notre société, dans notre vie quotidienne
  - Les entreprises doivent suivre cette transformation
    - Pour satisfaire les exigences des clients (services et qualité)
    - Pour au moins garder son avantage concurrentiel

- **Le Big Data peut servir de moteur pour la transformation digitale des entreprises**
    - Cas du groupe Accor, de SNCF…

Défis pour les entreprises

- **Qualité des données**
  - Stratégie de collecte et d'enrichissement des données
  - La qualité des prédictions dépend beaucoup de la qualité des données collectées
- **Mise en commun des compétences**
  - Experts métier
  - Experts techniques (ceux qui gèrent la data en interne)
  - Experts Big Data (nouvelles technologies, R&D)
- **Conduite de projets Big Data**
  - Tâche complexe pour les grandes entreprises
  - … à la hauteur des enjeux

Exemples de travaux et résultats

Système de Recommandation

- **S'appuie sur les notations (ratings) des produits par les utilisateurs**
- **Prédit (et recommande) les produits qui sont susceptibles d'intéresser un utilisateur**
  - Netflix : 2/3 des films loués
  - Amazon : 35% des ventes

Système de Recommandation

- **S'appuie sur les notations (ratings) des produits par les utilisateurs**
- **Prédit (et recommande) les produits qui sont susceptibles d'intéresser un utilisateur**
  - Netflix : 2/3 des films loués
  - Amazon : 35% des ventes
- **Matrice d'Utilité**

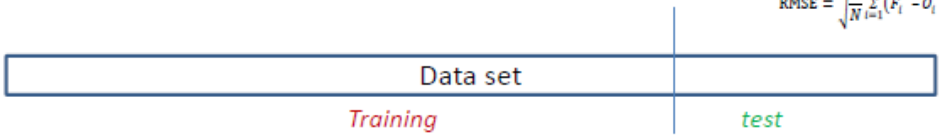
	King Kong	LOTR	Matrix	Nacho Libre
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

Système de Recommandation

- **S'appuie sur les notations (ratings) des produits par les utilisateurs**
- **Prédit (et recommande) les produits qui sont susceptibles d'intéresser un utilisateur**
  - Netflix : 2/3 des films loués
  - Amazon : 35% des ventes
- **Matrice d'Utilité**

	King Kong	LOTR	Matrix	Nacho Libre
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$$



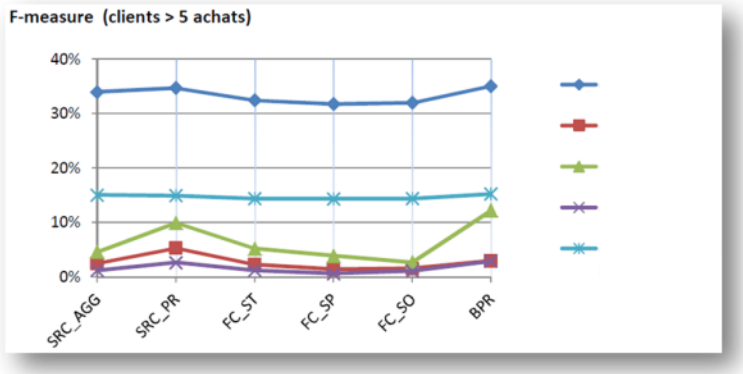
- **Evaluation**
  - Construction du modèle
  - Comparaison avec l'échantillon de test
    - Root-mean-square error (RMSE), F1-mesure …

Système de Recommandation

- **Deux approches principales**
  - Basée sur le contenu (content based)
    - Idée : recommander des produits similaires à ceux appréciés par l'utilisateur
      - Ex. Films : mêmes acteurs, même réalisateur, même genre …
  - Filtrage collaboratif (collaboratif filtering)
    - Idée : recommander des produits appréciés par les utilisateurs qui apprécient les mêmes choses que moi (notations similaires)
- **Complication**
  - La plus part des entreprises ne disposent pas de notations et n'ont que l'historique des achats
- **Défis**
  - Précision des prédictions
  - Capacité d'adaptation (évolution rapide des données, démarrage à froid, *Serendipity …)*

Recommandation d'hôtels

- **Expérimentations et tests**
  - Implantation : 6 algo. de recommandation (>7k lignes de code)
  - Echantillon : 200K clients, 38K hôtels, 345K transactions



- **Suite des travaux**
  - Recommandation de destinations
  - Modélisation Uplift pour l'évaluation des algo. de recommandation

Détection d'intrusion

- **Techniques basé sur les signatures**

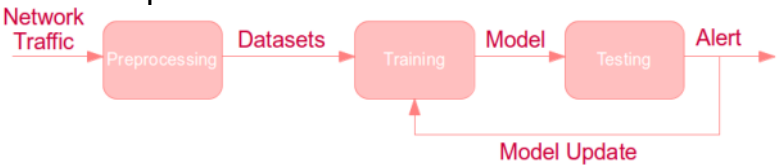
  - Base de données de signatures
  - Efficace pour les attaques connues
  - Pb : nouvelles attaques et variations autour des attaques connues

- **Détection d'anomalies**

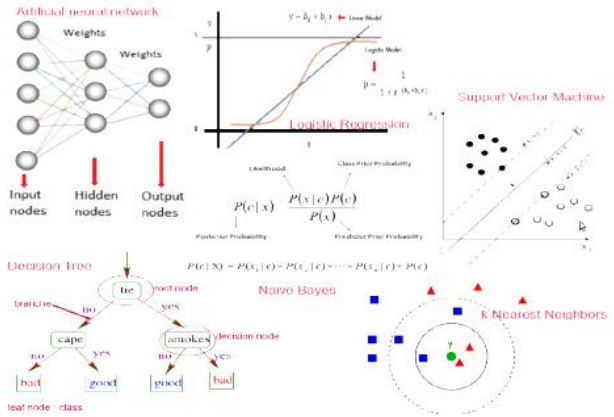
  - Détection des déviations
  - Apprend des logs (passé) et décide pour les actions à venir
  - La qualité dépend du modèle développé
  - Pb : faux positifs et scalabilité

Détection d'intrusion

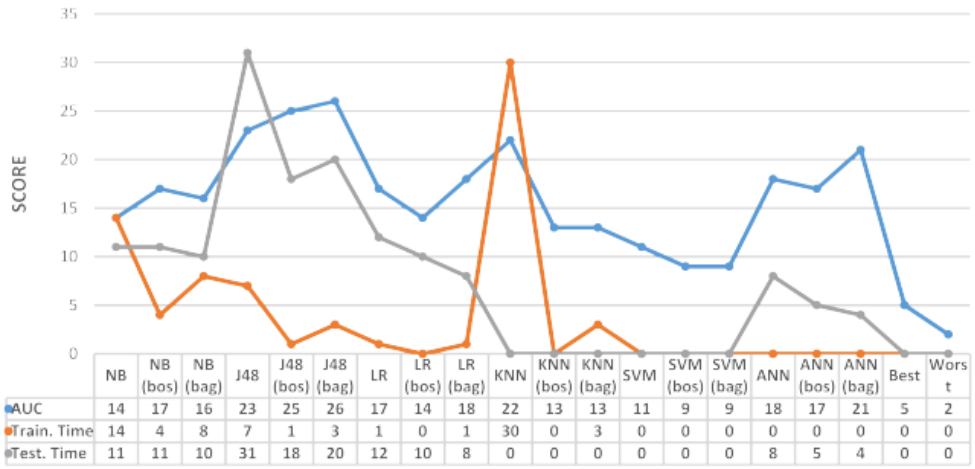
- **Utilisation des techniques de ML**
  - Décision sur la base d'un modèle appris à partir de l'historique
  - Distribution et calcul parallèle sur Spark



- **Expérimentations**
  - 9 jeux de données réels (7 publics + TPT + Deloitte)
  - 6 techniques de classification testées : régression logistique, arbres de décision, réseaux de neurones, KNN, SVM, classification naïve bayésienne.



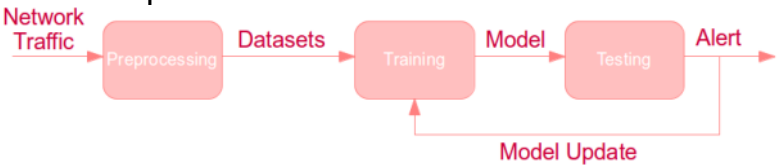
- **Résultats**



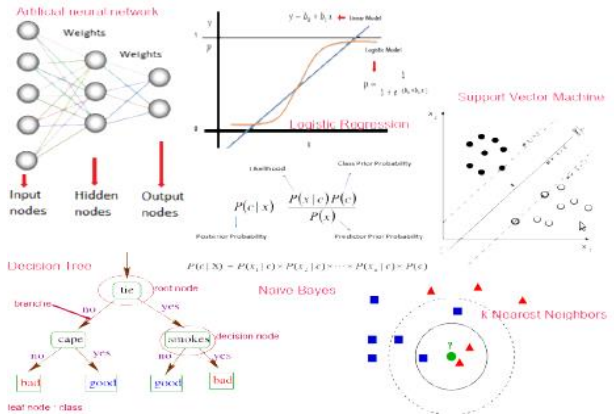
	NB	NB (bos)	NB (bag)	J48	J48 (bos)	J48 (bag)	LR	LR (bos)	LR (bag)	KNN	KNN (bos)	KNN (bag)	SVM	SVM (bos)	SVM (bag)	ANN	ANN (bos)	ANN (bag)	Best	Worst
AUC	14	17	16	23	25	26	17	14	18	22	13	13	11	9	9	18	17	21	5	2
Train. Time	14	4	8	7	1	3	1	0	1	30	0	3	0	0	0	0	0	0	0	0
Test. Time	11	11	10	31	18	20	12	10	8	0	0	0	0	0	0	8	5	4	0	0

Détection d'intrusion

- **Utilisation des techniques de ML**
  - Décision sur la base d'un modèle appris à partir de l'historique
  - Distribution et calcul parallèle sur Spark

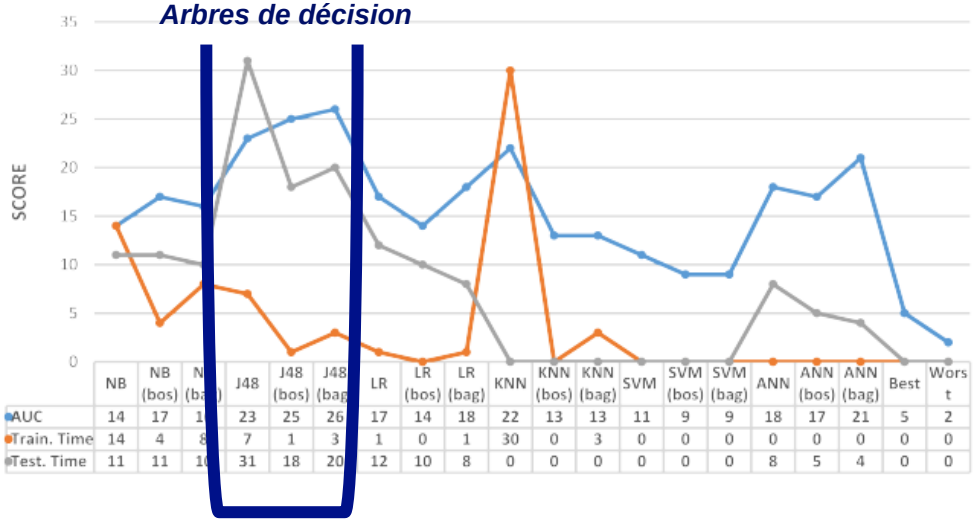


- **Expérimentations**
  - 9 jeux de données réels (7 publics + TPT + Deloitte)
  - 6 techniques de classification testées : régression logistique, arbres de décision, réseaux de neurones, KNN, SVM, classification naïve bayésienne.



- **Résultats**

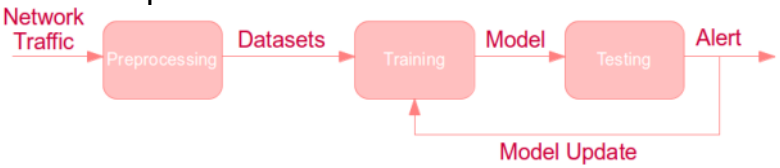
Arbres de décision



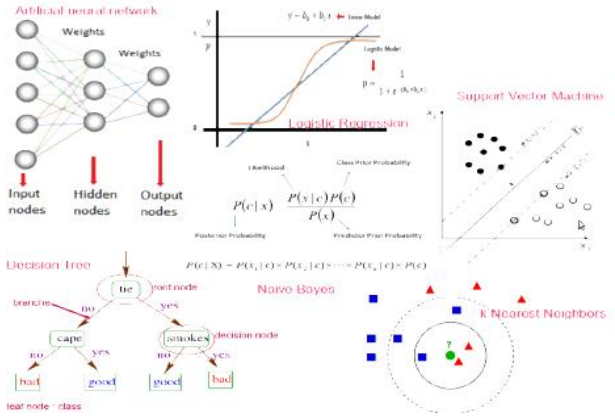
	NB	NB (bos)	N (ba)	J48	J48 (bos)	J48 (bag	LR	LR (bos)	LR (bag)	KNN	KNN (bos)	KNN (bag)	SVM	SVM (bos)	SVM (bag)	ANN	ANN (bos)	ANN (bag)	Best	Worst
AUC	14	17	1	23	25	26	17	14	18	22	13	13	11	9	9	18	17	21	5	2
Train. Time	14	4	8	7	1	3	1	0	1	30	0	3	0	0	0	0	0	0	0	0
Test. Time	11	11	1	31	18	20	12	10	8	0	0	0	0	0	0	8	5	4	0	0

Détection d'intrusion

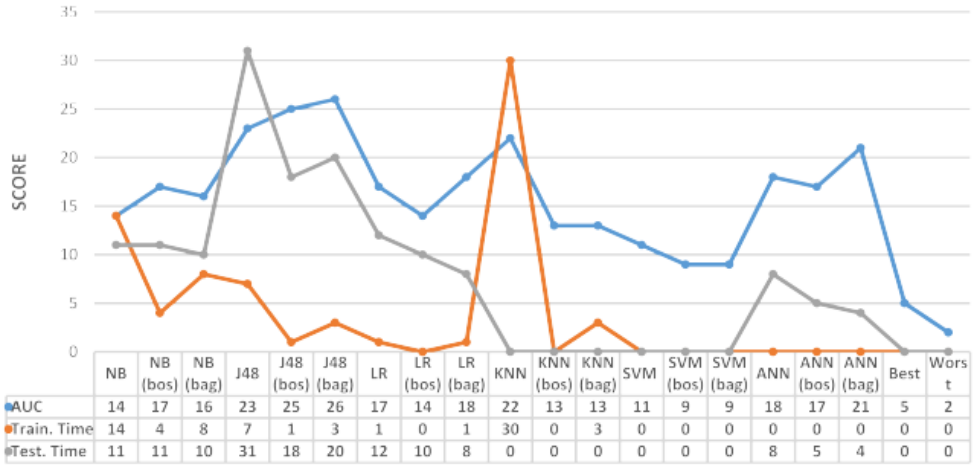
- **Utilisation des techniques de ML**
  - Décision sur la base d'un modèle appris à partir de l'historique
  - Distribution et calcul parallèle sur Spark



- **Expérimentations**
  - 9 jeux de données réels (7 publics + TPT + Deloitte)
  - 6 techniques de classification testées : régression logistique, arbres de décision, réseaux de neurones, KNN, SVM, classification naïve bayésienne.



- **Résultats**



	NB	NB (bos)	NB (bag)	J48	J48 (bos)	J48 (bag)	LR	LR (bos)	LR (bag)	KNN	KNN (bos)	KNN (bag)	SVM	SVM (bos)	SVM (bag)	ANN	ANN (bos)	ANN (bag)	Best	Worst
AUC	14	17	16	23	25	26	17	14	18	22	13	13	11	9	9	18	17	21	5	2
Train. Time	14	4	8	7	1	3	1	0	1	30	0	3	0	0	0	0	0	0	0	0
Test. Time	11	11	10	31	18	20	12	10	8	0	0	0	0	0	0	8	5	4	0	0

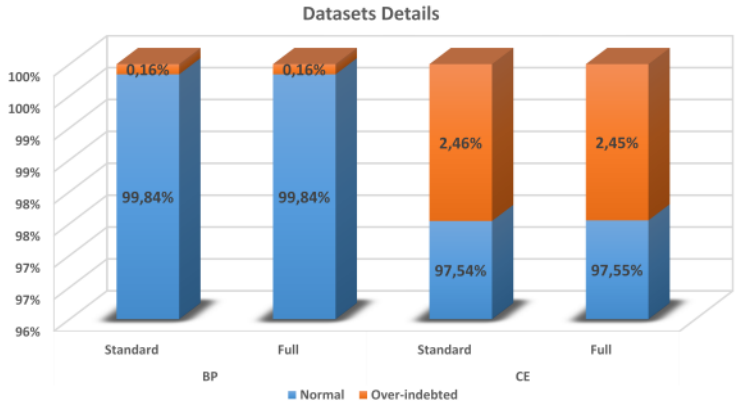
- **Suite des travaux**
  - Scalabilité, traitement à la volé (flux)

Surendettement

- **Contexte et motivation**
  - Charte d'inclusion bancaire et de prévention du surendettement
- **Objectif du travail**
  - Détection à 6 mois du risque de surendettement
  - Réduire le temps de calcul du modèle
  - Améliorer le taux de prédiction

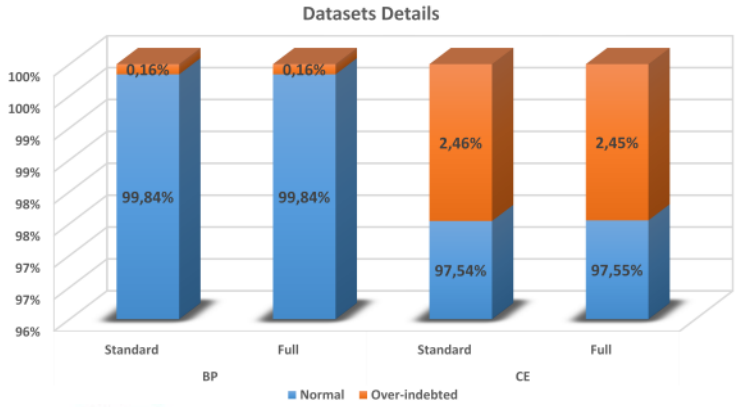
Surendettement

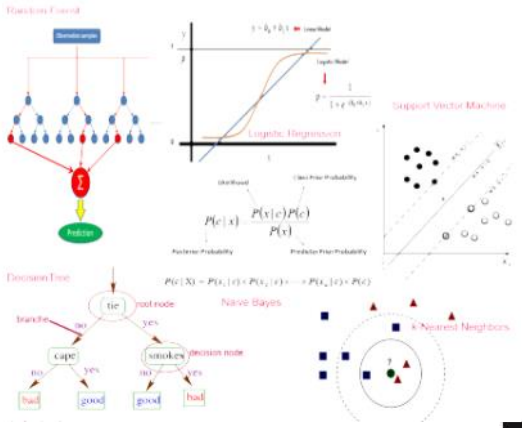
- **Contexte et motivation**
  - Charte d'inclusion bancaire et de prévention du surendettement
- **Objectif du travail**
  - Détection à 6 mois du risque de surendettement
  - Réduire le temps de calcul du moc
  - Améliorer le taux de prédiction
- **Expérimentations**
  - Jeu de données :
    - 700K clients BP
    - >1M clients CE
    - Historique : 2ans



Surendettement

- **Contexte et motivation**
  - Charte d'inclusion bancaire et de prévention du surendettement
- **Objectif du travail**
  - Détection à 6 mois du risque de surendettement
  - Réduire le temps de calcul du moc
  - Améliorer le taux de prédiction
- **Expérimentations**
  - Jeu de données :
    - 700K clients BP
    - >1M clients CE
    - Historique : 2ans
  - Six algorithmes testés :
    - Regression logistique,
    - Arbres de décision,
    - Random Forest,
    - KNN, SVM,
    - classification naïve bayésienne

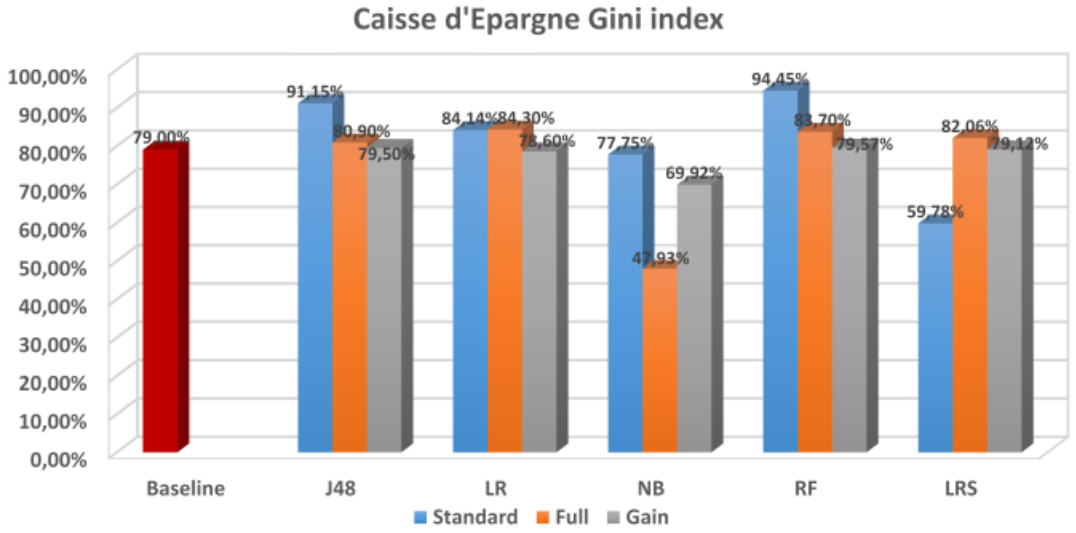




Surendettement

- **Outils utilisés**
  - Spark, MLLib,
  - Weka (réduction de dimensions)

- **Résultats**
  - Calcul ~ quelques minutes,
  - Gini ~ 0.94 (random forest)



Analyse des réseaux sociaux

- **Sources de données**
  - Facebook et Twitter,
    - Posts, commentaires, likes, shares, …
    - Petit Bateau, YR et les principales marques de cosmétiques (L'Oréal, Sephora, Kiko …)

- **Pour FaceBook**
  - GraphAPI v2.4
  - Netvizz.
  - API Blender

- **Pour Twitter**
  - TwitterAPI v1.1
  - API Blender

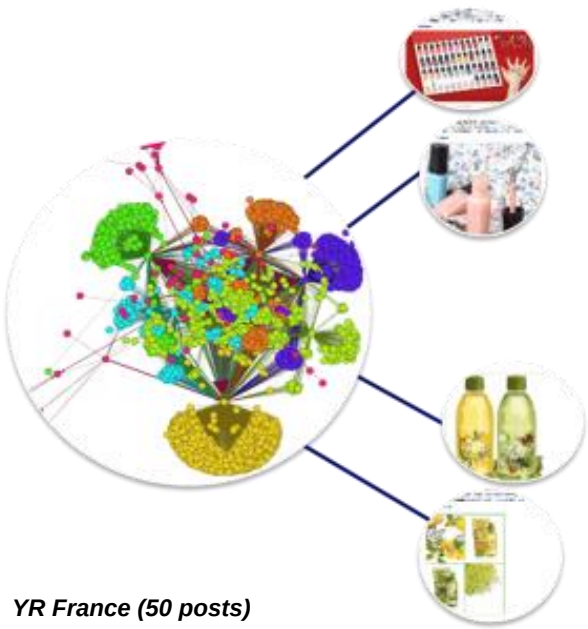
- **Stockage**
  - MongoDB pour les données de Twitter
  - Fichiers gdf pour les données Facebook

Analyse des réseaux sociaux

- **Etude comparative**
  - Taux d'engament, mesure d'influence, détection de communautés, analyse de sentiment

Analyse des réseaux sociaux

- **Etude comparative**
  - Taux d'engament, mesure d'influence, détection de communautés, analyse de sentiment
- **Réseau :** User → Post

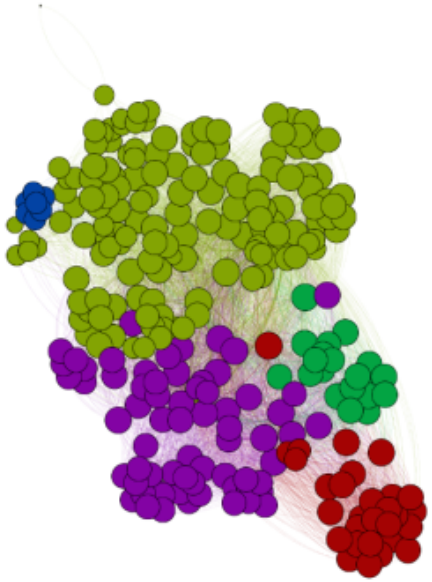


YR France (50 posts)

YR Saudi Arabia (50 posts)

Analyse des réseaux sociaux

- **Etude comparative**
  - Taux d'engament, mesure d'influence, détection de communautés, analyse de sentiment
- **Réseau :** User → User



YR UK, User-User Graph

Most Influential User (Sorted by their PageRank)

Label	Likes	Comments all	Engagement	Weighted Degree	Page Rank	Eigenvector Centrality	Closeness Centrality	Betweenness Centrality
user_101520241 19366587	39	1	40	1090	0.02618	1	1.121324	13531.25
user_594361904 034853	31	0	31	914	0.022761	0.940166	1.216912	8875.202
user_102068310 58404764	31	2	33	710	0.019701	0.649259	1.400735	6769.166
user_856098484 477340	27	0	27	726	0.019672	0.766233	1.367647	6360.272
user_102068376 57048019	14	0	14	564	0.017755	0.821832	1.371324	4742.592
user_102040675 92146279	21	0	21	624	0.017309	0.759434	1.411765	4894.957
user_156863128 0068378	18	0	18	428	0.013997	0.531197	1.569853	3305.304
user_885519094 859697	12	2	14	422	0.012931	0.612361	1.555147	1981.91
user_841564932 603698	9	0	9	374	0.012269	0.603155	1.5625	1884.302

Selected publications

- **O. D. Balalau, F. Bonchi, T-H. Chan, F. Gullo et M. Sozio. Finding Subgraphs with Maximum Total Density and Limited Overlap, WSDM 2015.**

- **M. Gueye, T. Abdessalem et H. Naacke. Dynamic recommender system: using cluster-based biases to improve the accuracy of the predictions. In Advances in Knowledge Discovery and Management, Mars 2015.**

- **M. Gueye, T. Abdessalem et H. Naacke. A Social and Popularity-based Tag Recommender. SocialCom, decembre 2014, Sydney, Australia.**

- **J.B. Griesner, T. Abdessalem et H. Naacke. POI Recommendation: Towards Fused Matrix Factorization with Geographical and Temporal Influences. In RecSys, septembre 2015.**

- **S. Lei, S. Maniu, L. Mo, R. Cheng et P. Senellart, « Online Influence Maximization ». Proc. KDD, Sydney, Australie, août 2015.**

- **C. Meng, R. Cheng, S. Maniu, P. Senellart et W. Zhang, « Discovering Meta-Paths in Large Heterogeneous Information Networks ». Proc. WWW, p. 754-764, Florence, Italie, mai 2015.**

Merci pour votre attention !