Conférence ISN
Informatique et Sciences du Numérique

Introduction à la fouille de données et au "Big Data"

Germain Forestier, Université de Haute-Alsace

02/12/2014

Parcours :

- Germain FORESTIER
- né à Strasbourg en 1984
- BAC S au Lycée Louis Pasteur
- DUT en Informatique à l'IUT d'Illkirch
- Master en Informatique à l'ULP en 2007
- Docteur en Informatique de l'Unistra en 2010
- Maître de conférences depuis 2011 à l'Université de Haute-Alsace

⇝ http://germain-forestier.info/

La fouille de données :

- La fouille de données a pour objet l'extraction **d'un savoir ou d'une connaissance** à partir de **grandes quantités de données**, par des méthodes automatiques ou semi-automatiques.
- exploration de données, fouille de données, *knowledge discovery*, *business intelligence*, *data mining*, *data analytics*, big data, etc.

Quels types de données ?

- données numériques/symboliques (capteurs, images, etc.)
- données textuelles (articles, sites web, livres, etc.)
- données complexes (vidéos, séquences, etc.)

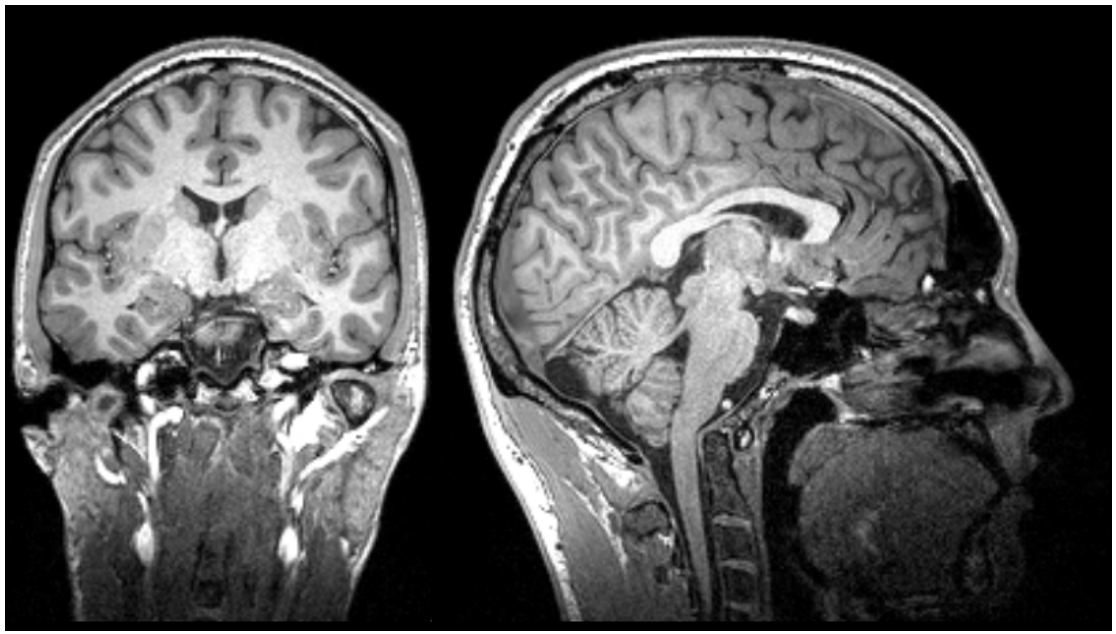
	A	B	C	D	E	F	G	H
1	age	sexe	pression	cholester	sucre	electro	taux_max	angine
2	70	masculin	130	322	A	C	109	non
3	67	feminin	115	564	A	C	160	non
4	57	masculin	124	261	A	A	141	non
5	64	masculin	128	263	A	A	105	oui
6	74	feminin	120	269	A	C	121	oui
7	65	masculin	120	177	A	A	140	non
8	56	masculin	130	256	B	C	142	oui
9	59	masculin	110	239	A	C	142	oui
10	60	masculin	140	293	A	C	170	non
11	63	feminin	150	407	A	C	154	non

données de patients

source : `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

Quels types de données ?

- données numériques/symboliques (capteurs, images, etc.)
- données textuelles (articles, sites web, livres, etc.)
- données complexes (vidéos, séquences, etc.)



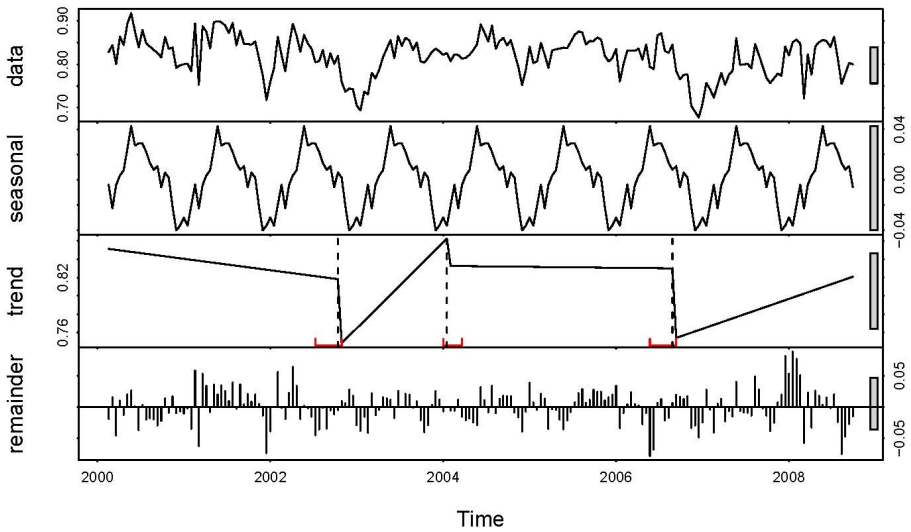
IRM de mon cerveau

Quels types de données ?

- données numériques/symboliques (capteurs, images, etc.)
- données textuelles (articles, sites web, livres, etc.)
- données complexes (vidéos, séquences, etc.)

Quels types de données ?

- données numériques/symboliques (capteurs, images, etc.)
- données textuelles (articles, sites web, livres, etc.)
- données complexes (vidéos, séquences, etc.)



informations spectrales d'une plantation de pins

source : `http://bfast.r-forge.r-project.org/`

Un exemple :

- enregistrer les symptômes d'un ensemble de patients
- enregistrer la présence d'un problème cardiaque
- créer un modèle prédictif pour les futurs patients

	A	B	C	D	E	F	G	H
1	age	sexe	pression	cholester	sucre	electro	taux_max	angine
2	70	masculin	130	322	A	C	109	non
3	67	feminin	115	564	A	C	160	non
4	57	masculin	124	261	A	A	141	non
5	64	masculin	128	263	A	A	105	oui
6	74	feminin	120	269	A	C	121	oui
7	65	masculin	120	177	A	A	140	non
8	56	masculin	130	256	B	C	142	oui
9	59	masculin	110	239	A	C	142	oui
10	60	masculin	140	293	A	C	170	non
11	63	feminin	150	407	A	C	154	non

source : `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

Un exemple :

- enregistrer les symptômes d'un ensemble de patients
- enregistrer la présence d'un problème cardiaque
- créer un modèle prédictif pour les futurs patients

	A	B	C	D	E	F	G	H
1	age	sexe	pression	cholester	sucre	electro	taux_max	angine
2	70	masculin	130	322	A	C	109	non
3	67	feminin	115	564	A	C	160	non
4	57	masculin	124	261	A	A	141	non
5	64	masculin	128	263	A	A	105	oui
6	74	feminin	120	269	A	C	121	oui
7	65	masculin	120	177	A	A	140	non
8	56	masculin	130	256	B	C	142	oui
9	59	masculin	110	239	A	C	142	oui
10	60	masculin	140	293	A	C	170	non
11	63	feminin	150	407	A	C	154	non

source : `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

Un exemple :

- enregistrer les symptômes d'un ensemble de patients
- enregistrer la présence d'un problème cardiaque
- créer un modèle prédictif pour les futurs patients

	A	B	C	D	E	F	G	H
1	age	sexe	pression	cholester	sucre	electro	taux_max	angine
2	70	masculin	130	322	A	C	109	non
3	67	feminin	115	564	A	C	160	non
4	57	masculin	124	261	A	A	141	non
5	64	masculin	128	263	A	A	105	oui
6	74	feminin	120	269	A	C	121	oui
7	65	masculin	120	177	A	A	140	non
8	56	masculin	130	256	B	C	142	oui
9	59	masculin	110	239	A	C	142	oui
10	60	masculin	140	293	A	C	170	non
11	63	feminin	150	407	A	C	154	non

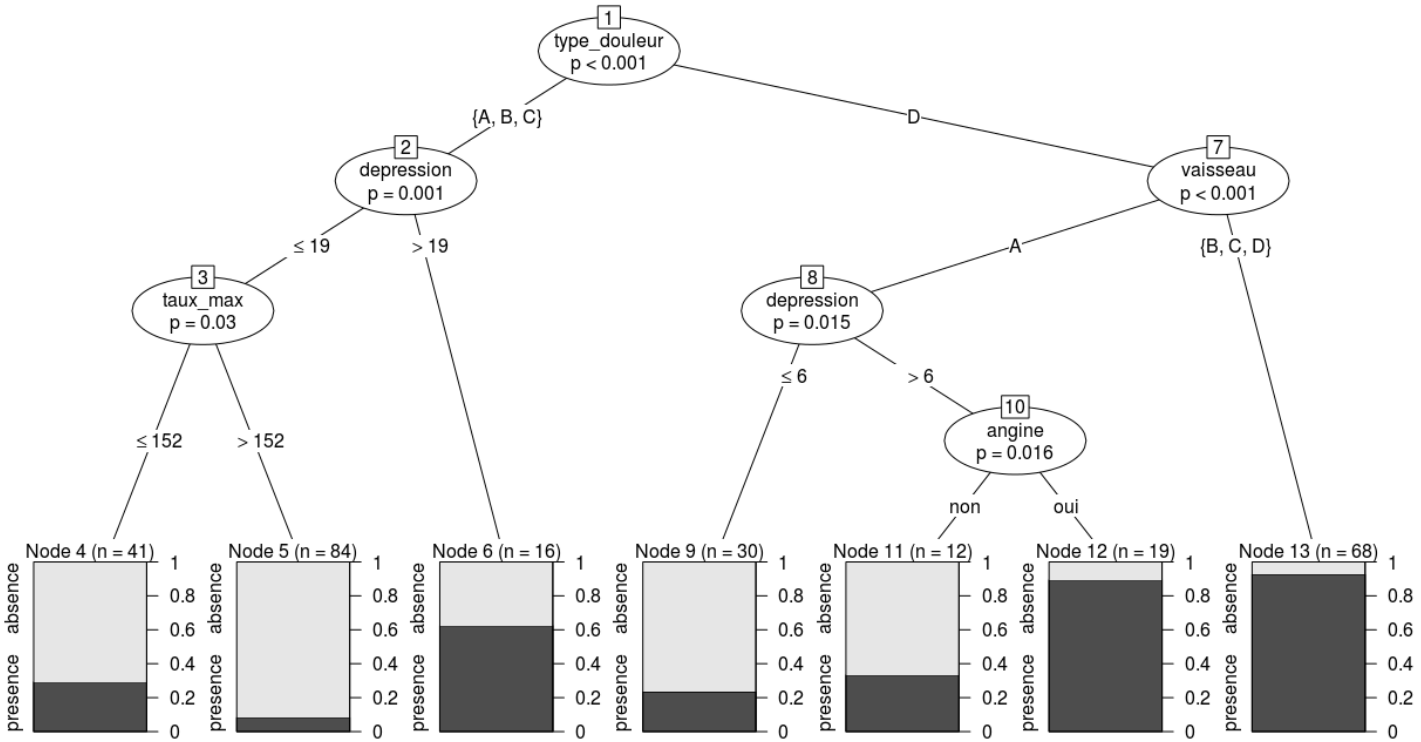
L
coeur
presence
absence
presence
absence
absence
absence
presence
presence
presence

?

source : `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

Un modèle prédictif

- arbre de décision (ID3)



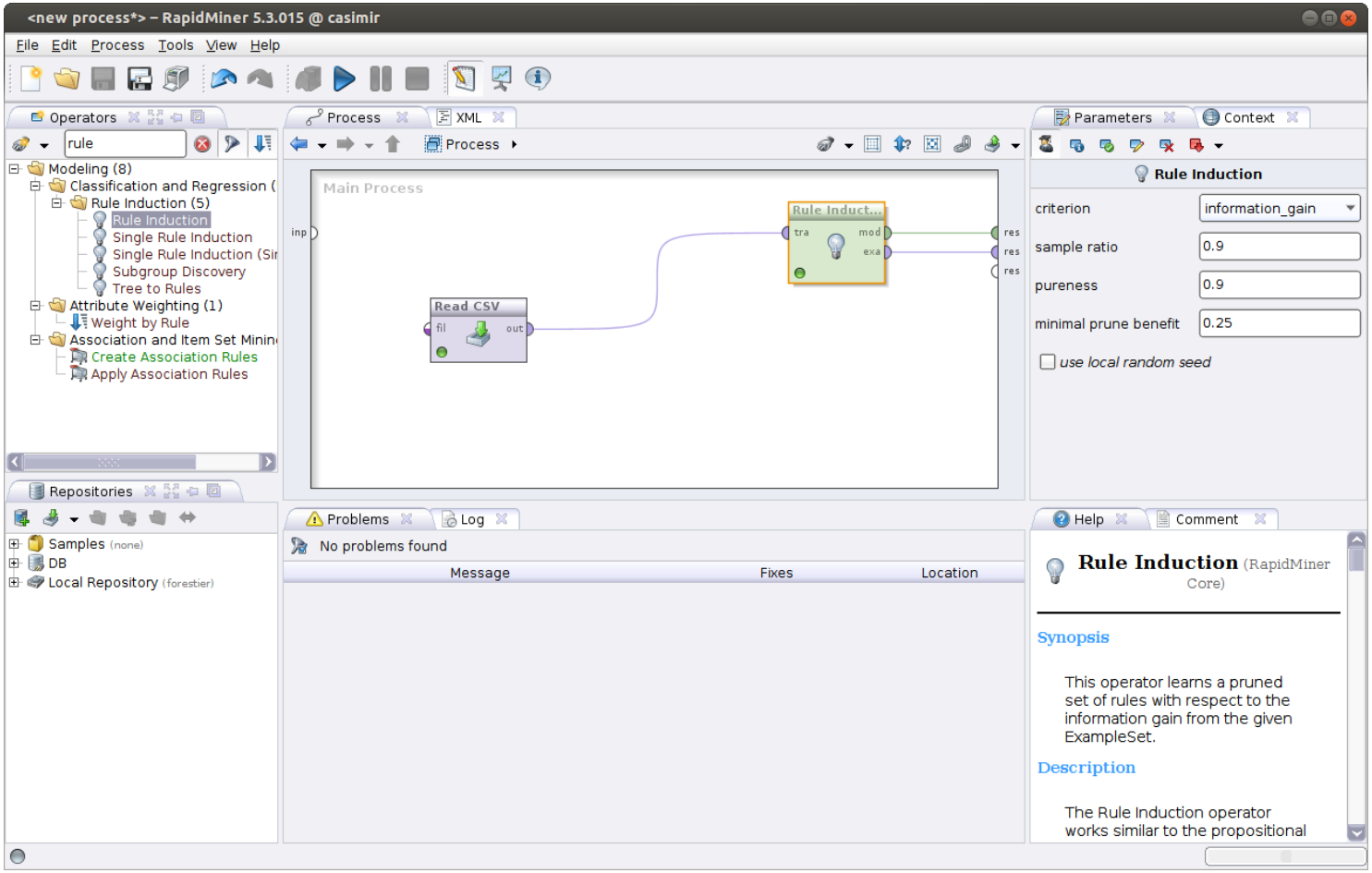
Un modèle prédictif

- apprentissage de règles (Apriori)

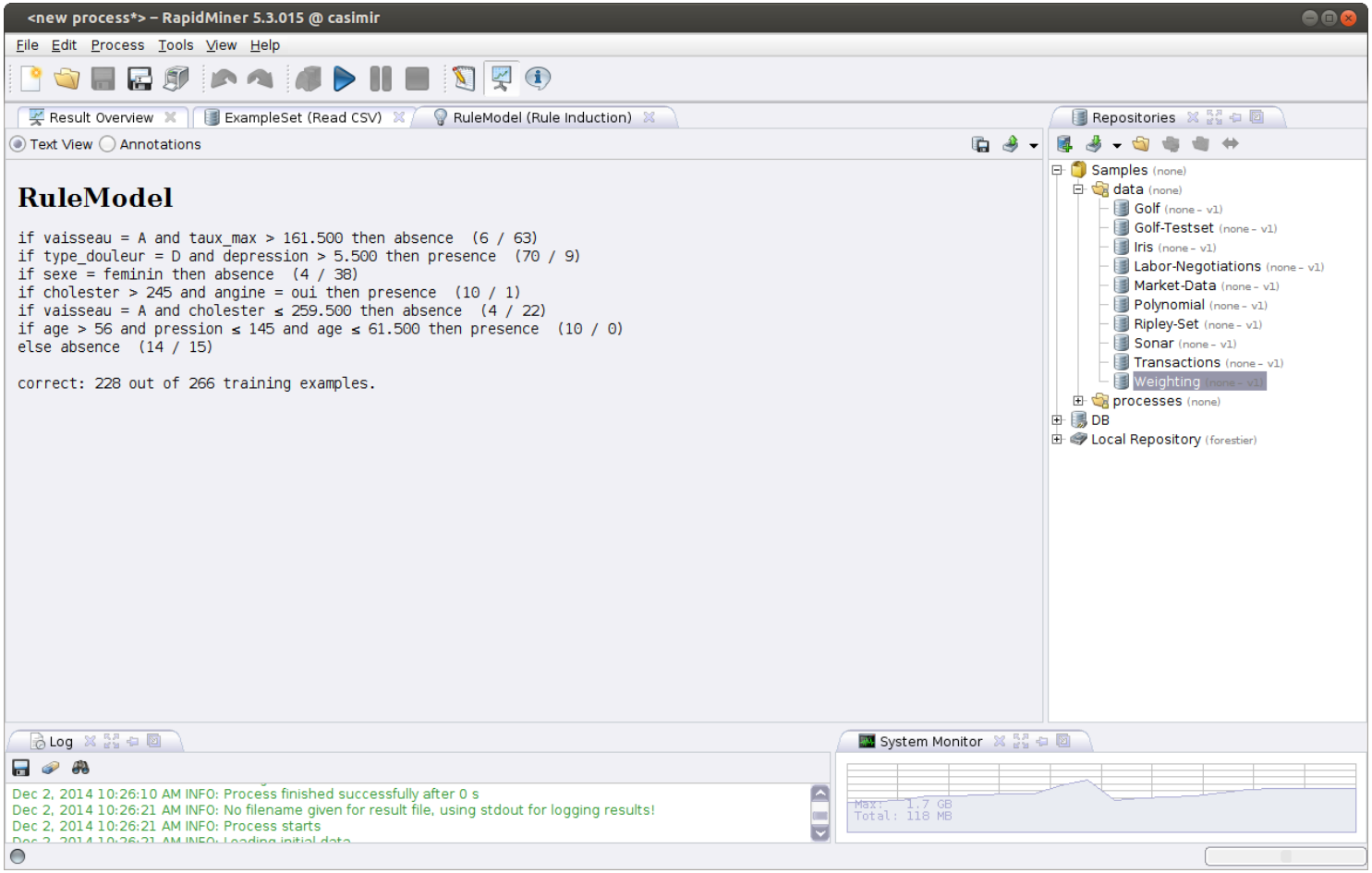
```
if vaisseau = A and taux_max > 161.500 then absence  (6 / 63)
if type_douleur = D and depression > 5.500 then presence  (70 / 9)
if sexe = feminin then absence  (4 / 38)
if cholester > 245 and angine = oui then presence  (10 / 1)
if vaisseau = A and cholester <= 259.500 then absence  (4 / 22)
if age > 56 and pression <= 145 and age <= 61.500 then presence  (10 / 0)
else absence  (14 / 15)

correct: 228 out of 266 training examples.
```

- le logiciel Rapid Miner `https://rapidminer.com/`



- le logiciel Rapid Miner `https://rapidminer.com/`



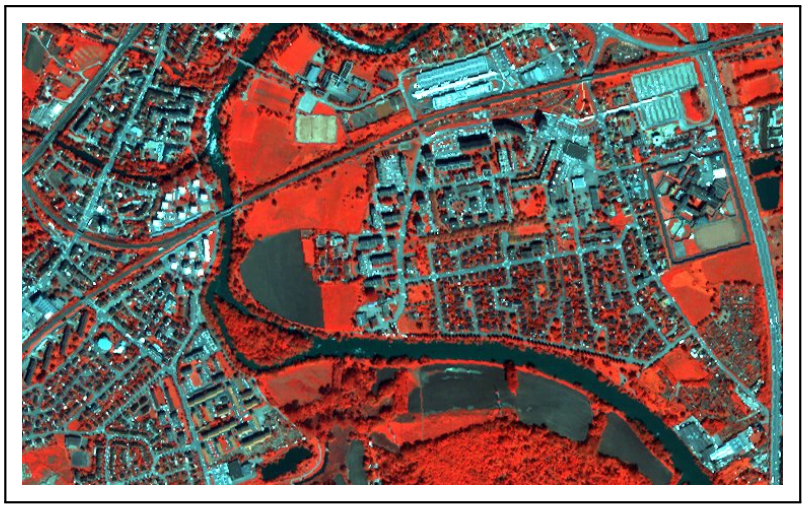
Et les "Big Data" ?!

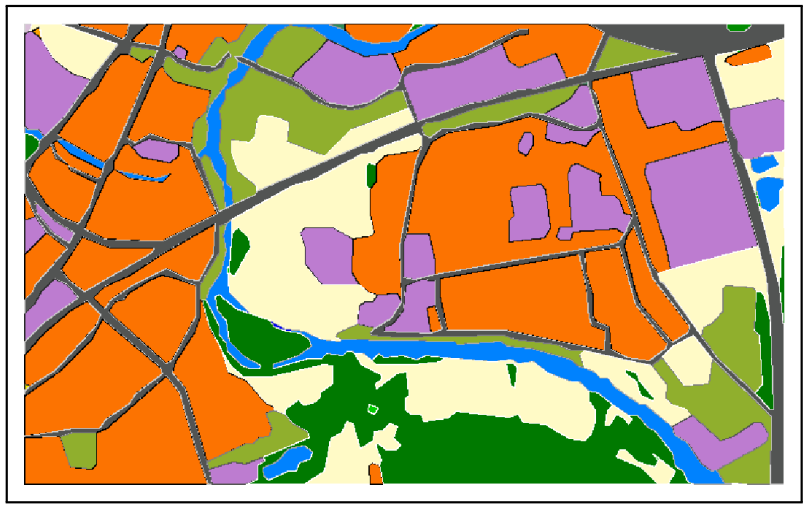
- 90% des données mondiales ont été créées ces 2 dernières années
- explosion de la quantité de données disponibles
- méthodes actuelles inadaptées (changement de paradigme)
- les 3V : volume, vélocité et variété

Travaux de recherche (thèse)

- analyse de données satellitaires
- sentinel, pléiades, CNES

Objectif: identification des types d'occupation du sol





G. Forestier, *Connaissances et clustering collaboratif d'objets complexes multisources*, Doctorat de l'Université de Strasbourg, 2011

Travaux de recherche (thèse)

- analyse de données satellitaires
- sentinel, pléiades, CNES

Enjeux sociétaux

- étude de l'évolution urbaine / végétation
- suivi de l'agriculture / environnement

Verrous scientifiques

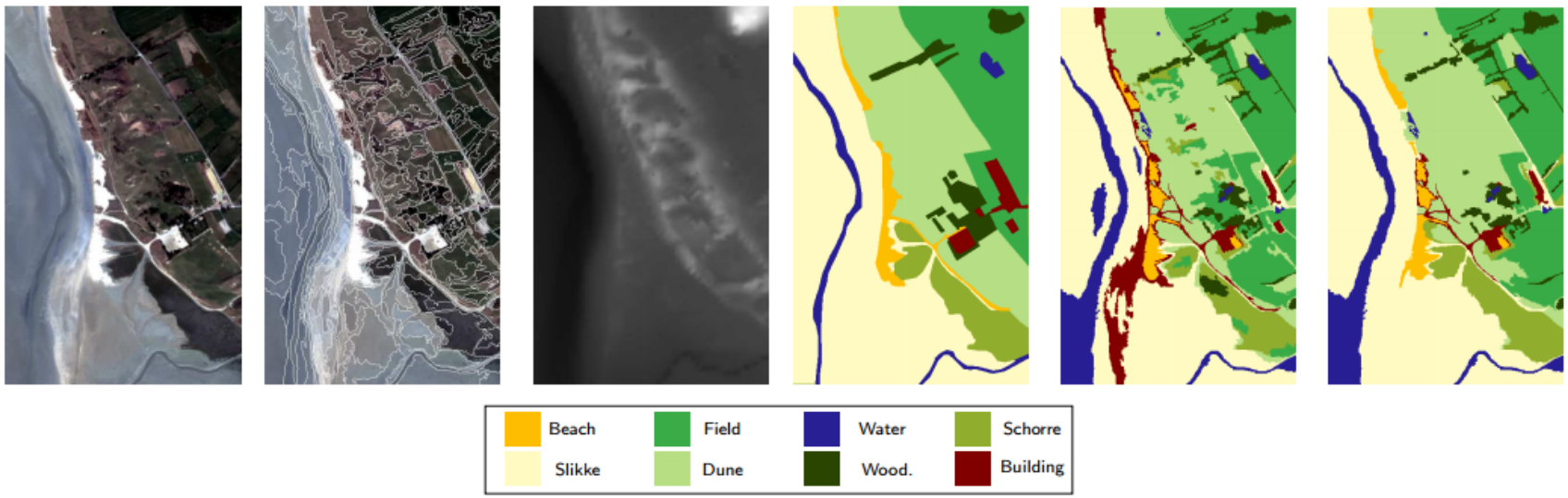
- masse de données importante (captures quotidiennes)
- hétérogénéité des données (saisons, capteurs)

G. Forestier et al. *Knowledge-based region labeling for remote sensing image interpretation*. Computers, Environment and Urban Systems, 36(5):470-480, 2012.

Travaux de recherche (thèse)

- analyse de données satellitaires
- sentinel, pléiades, CNES

Objectif: identification des types d'occupation du sol

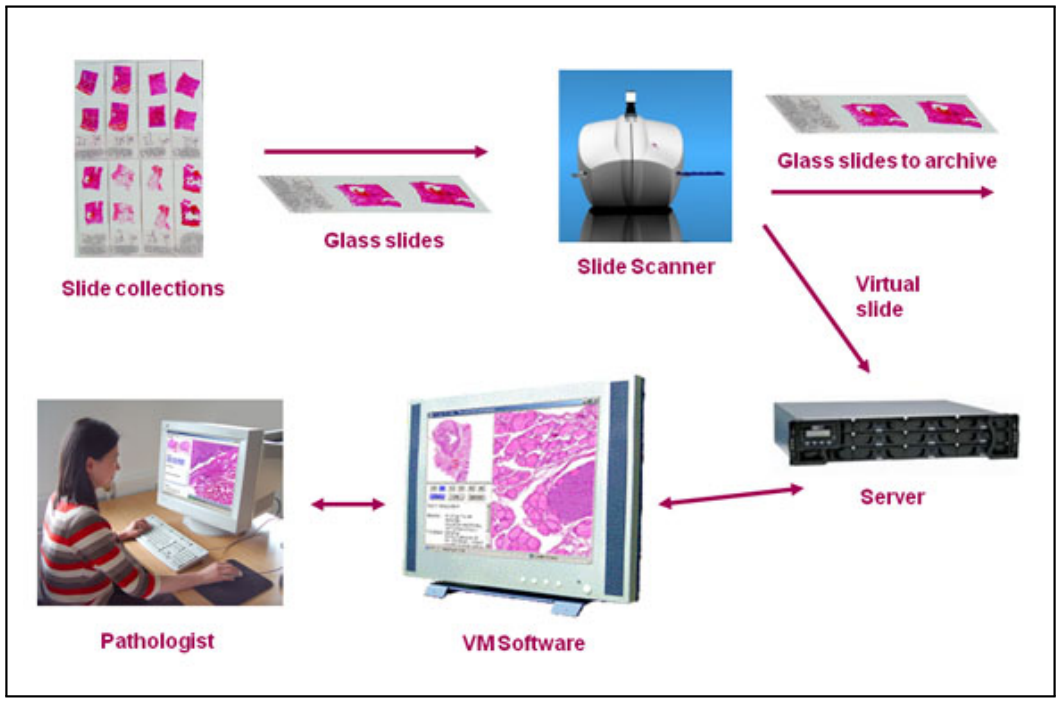


G. Forestier et al. *Coastal image interpretation using background knowledge and semantics.* Computers & Geosciences, 54(0):88-96, 2013.

Travaux de recherche (actuels)

- étude de données histopathologiques
- collaboration laboratoires Roche et Hôpital de Hanovre

Objectif: identification automatique de zones d'intérêt

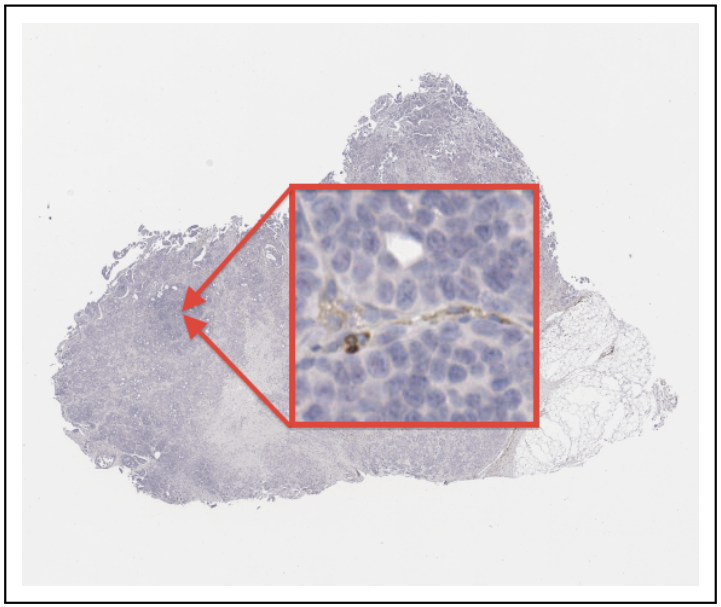


source : `http://vmscope.com/virtuelle-mikroskopie.html`

Travaux de recherche (actuels)

- étude de données histopathologiques
- collaboration laboratoires Roche et Hôpital de Hanovre

Objectif: identification automatique de zones d'intérêt

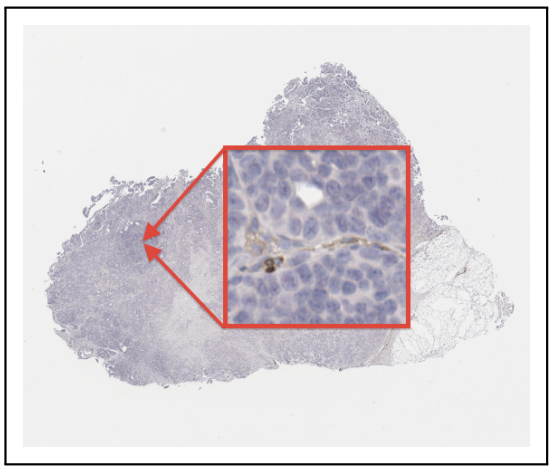
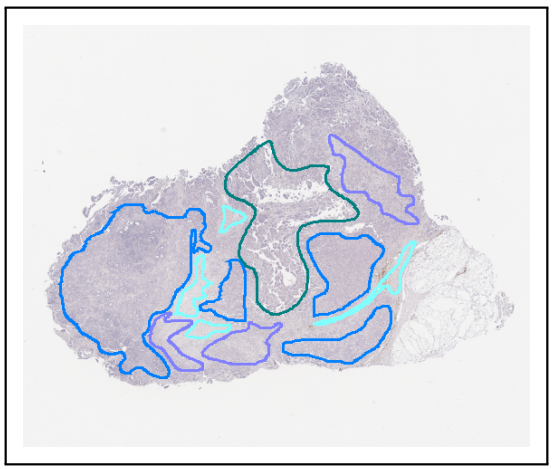
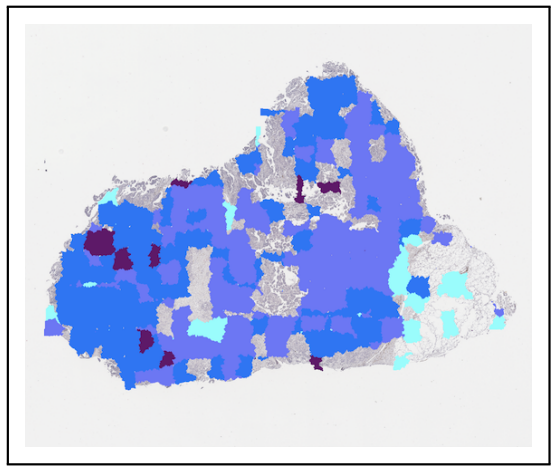


18000x15000 $\rightsquigarrow$ 270 Mo

Travaux de recherche (actuels)

- étude de données histopathologiques
- collaboration laboratoires Roche et Hôpital de Hanovre

Objectif: identification automatique de zones d'intérêt

Travaux de recherche (actuels)

- étude de données histopathologiques
- collaboration laboratoires Roche et Hôpital de Hanovre

Enjeux sociétaux

- aider la prise de décision (choix de traitement)
- télémédecine, diagnostic à distance

Verrous scientifiques

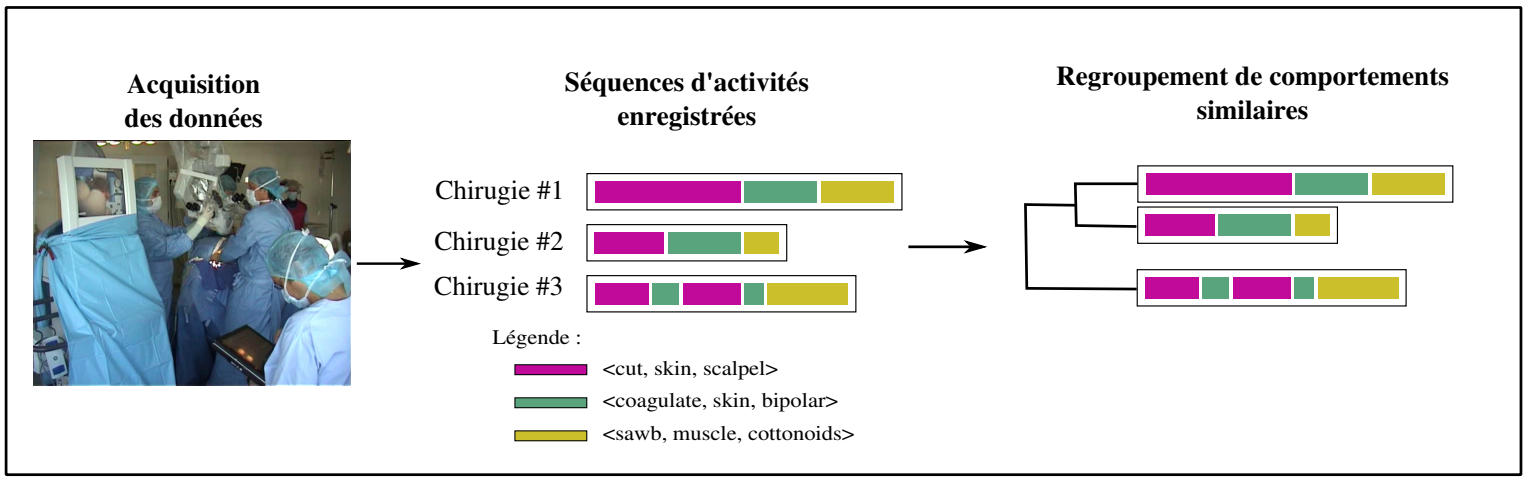
- données massives disponibles (WSI - *Whole Slide Image*)
- modélisation des connaissances médicales

G. Apou et al. *Fast segmentation for texture-based cartography of whole slide images.* In International Conference on Computer Vision Theory and Applications, pages 309-319, Lisbon, Portugal, 2014. (best paper award)

Travaux de recherche (actuels)

- étude du comportement des chirurgiens
- collaboration Université de Rennes

Objectif: comparer les gestes chirurgicaux



G. Forestier et al. *Multi-site study of surgical practice in neurosurgery based on surgical process models.* Journal of Biomedical Informatics, 46(5):822-829, 2013.

Travaux de recherche (actuels)

- étude du comportement des chirurgiens
- collaboration Université de Rennes

Enjeux sociétaux

- formation des chirurgiens
- amélioration des soins

Verrous scientifiques

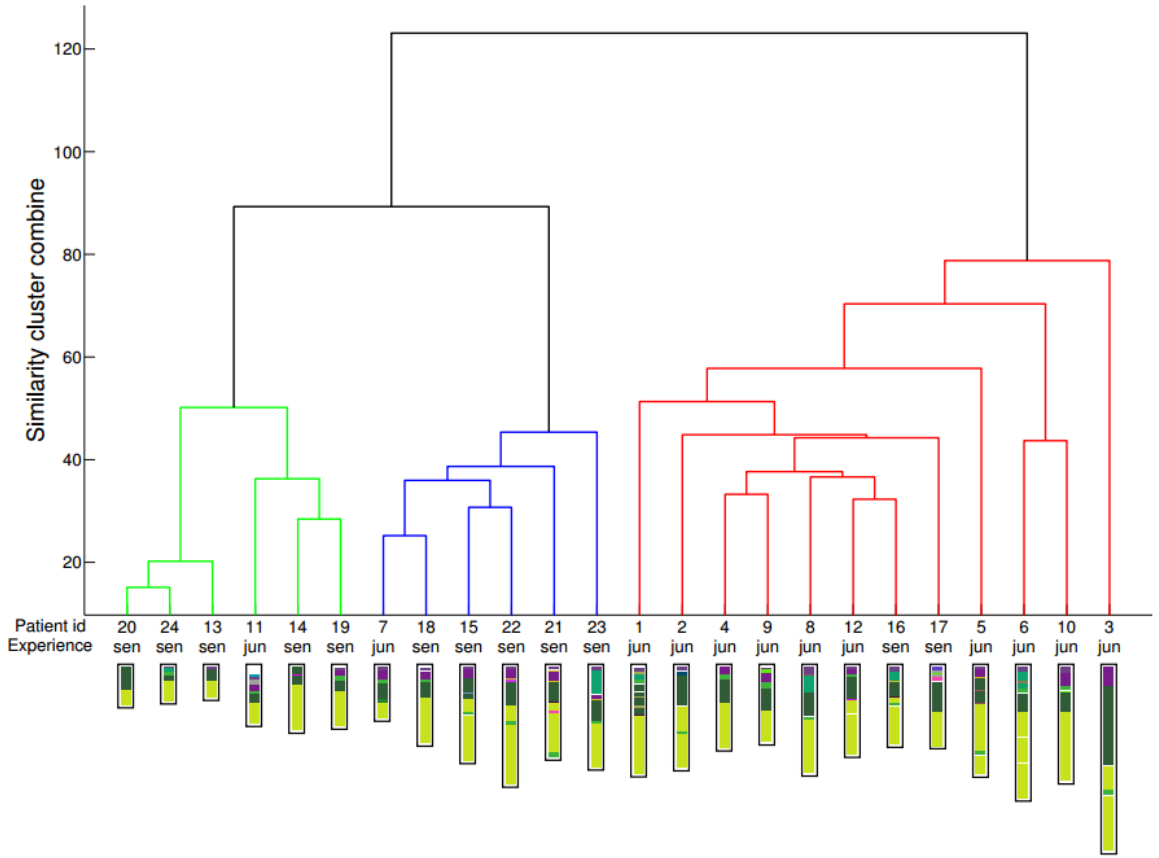
- données difficiles à acquérir
- données complexes (séries temporelles d'actions)

G. Forestier et al. *Multi-site study of surgical practice in neurosurgery based on surgical process models.* Journal of Biomedical Informatics, 46(5):822-829, 2013.

Travaux de recherche (actuels)

- étude du comportement des chirurgiens

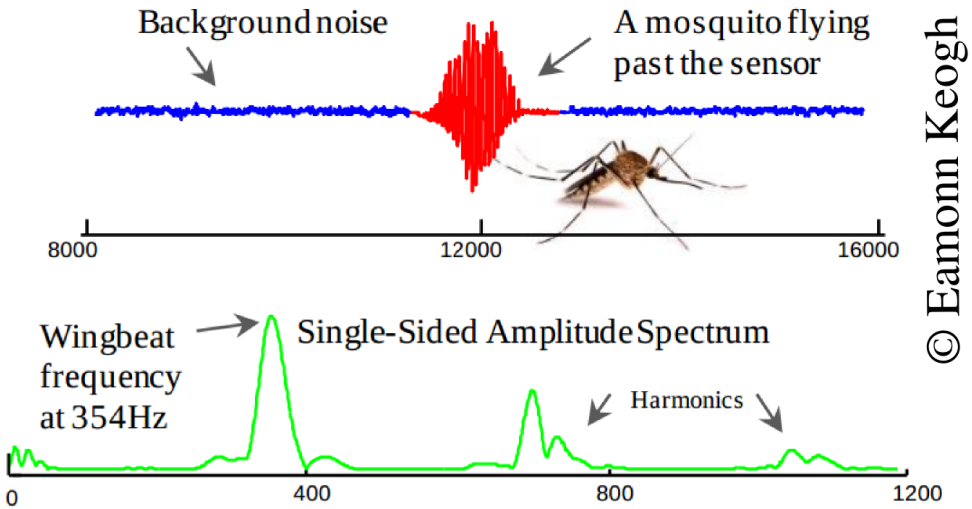
Objectif: comparer les gestes chirurgicaux



Travaux de recherche (actuels)

- analyse de séries temporelles
- collaboration internationale (Univ. of California, Monash Univ.)

Objectif: analyser des séquences numériques



UC RIVERSIDE UNIVERSITY OF CALIFORNIA MONASH University

F. Petitjean et al. *Dynamic time warping averaging of time series allows faster and more accurate classification.* IEEE International Conference on Data Mining, to appear, 2014.

Travaux de recherche (actuels)

- analyse de séries temporelles

Enjeux sociétaux

- détecter les types d'insecte (moustique) automatiquement
- enjeux pour la santé des plantes et des humains

Verrous scientifiques

- données très complexes (bruit)
- nécessité de classification rapide (prise de décision)

F. Petitjean et al. *Dynamic time warping averaging of time series allows faster and more accurate classification.* IEEE International Conference on Data Mining, to appear, 2014.

Campagne présidentielle de Barak Obama

"We could [predict] people who were going to give online. We could model people who were going to give through mail. We could model volunteers."

Source : *"Inside the Secret World of the Data Crunchers Who Helped Obama Win"*, Time Magazine

`http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/`

Big Data Is Changing the Way We Get Well

""Big data allows us to collect more data about patients than any clinical trial in history," said Vinnie Ramesh, CTO at Wellframe, via email. "We can start generating truly evidence-based insights and accelerate the pace of medical innovation in an unprecedented manner.""

Source : *"Big Data Is Changing the Way We Get Well"*, Mashable

`http://mashable.com/2014/09/17/big-data-health-care/`

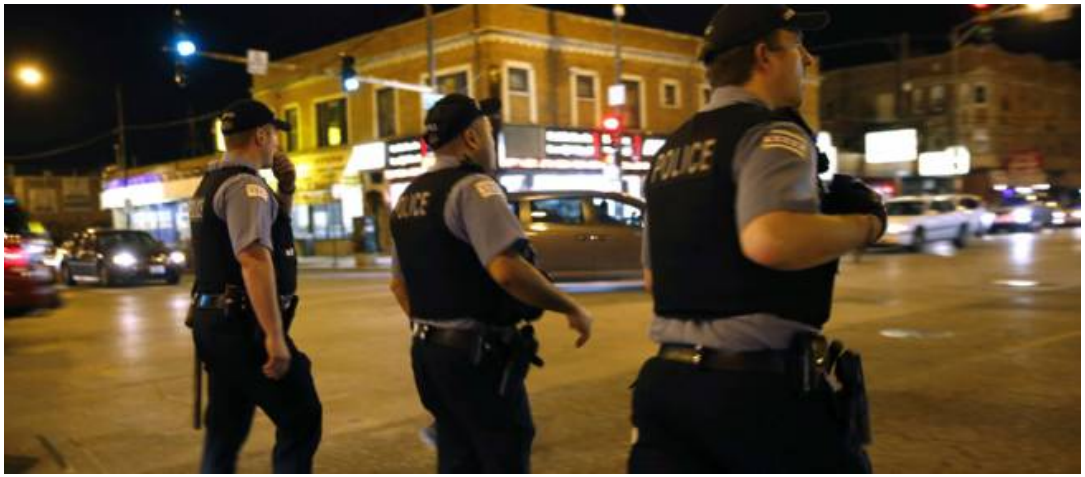
Tirer les cordons de la Bourse grâce à Twitter ?

"De plus en plus de spécialistes se tournent vers Twitter en espérant pouvoir y déceler les tendances de la Bourse avant les autres. Aux Etats-Unis, des start-up proposent des logiciels qui analysent le "sentiment" des prescripteurs influents qui sont sur le réseau."

Source : *"Tirer les cordons de la Bourse grâce à Twitter ?"*, Courrier International

`http://www.courrierinternational.com/article/2013/12/25/tirer-les-cordons-de-la-bourse-grace-a-twitter`

Et si Twitter... aidait à prévenir la criminalité ?



"L'analyse de tweets géolocalisés permet de prédire 19 à 25 formes de criminalité, en particulier le harcèlement, le vol et certains types d'agressions, selon ces travaux publiés dans le journal scientifique Decision Support Systems."

Source : *"Et si Twitter... aidait à prévenir la criminalité ?"*, Le Point

`http://www.lepoint.fr/high-tech-internet/et-si-twitter-aidait-a-prevenir-la-criminalite-20-04-2014-1814734_47.php`

Big data football club

"Le football moderne se convertit peu à peu au " big data ", la collecte massive de données. Au rythme des avancées technologiques, les performances des stars du ballon rond sont désormais scrutées à travers les statistiques."

Source : *"Big data football club"*, LeMonde

`http://www.lemonde.fr/coupe-du-monde/article/2014/06/12/big-data-football-club_4432507_1616627.html`

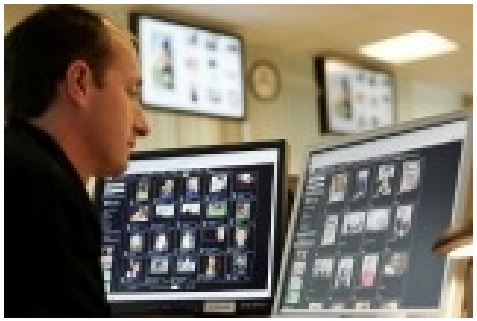
Pourquoi Netflix a (déjà) gagné

"Cette simplicité d'utilisation doit par ailleurs beaucoup au fameux algorithme créé par Netflix pour identifier les envies des spectateurs et leur proposer le contenu qu'ils veulent voir, avant même qu'ils n'aient conscience de le vouloir."

Source : *"Pourquoi Netflix a (déjà) gagné"*, Le Point

`http://www.lepoint.fr/high-tech-internet/pourquoi-netflix-a-deja-gagne-12-09-2014-1862528_47.php`

Le commerce en ligne français s'arrache les data miners



"Son site emploie 5 data miners sur 40 informaticiens, pour "draguer légèrement plutôt que lourdement" les clients, c'est-à-dire leur envoyer des publicités bien ciblées. "Tout le monde s'arrache les meilleurs", témoigne Romain Niccoli, cofondateur et patron de la R&D chez Criteo. Ce spécialiste français de la pub en ligne cherche à pourvoir 100 postes dans la recherche et développement."

Source : *"Le commerce en ligne français s'arrache les data miners"*, Les Echos

`http://www.lesechos.fr/15/07/2012/lesechos.fr/0202173368914_`

`le-commerce-en-ligne-francais-s-arrache-les---data-miners--.htm`

Comment notre ordinateur nous manipule

"L'objectif est de vous "profiler", c'est à dire de créer des fichiers personnalisés, stockés dans des bases de données. En d'autres termes, de mieux vous connaitre afin de vous présenter le bon message publicitaire au bon moment et dans le bon format."

Source : *"Comment notre ordinateur nous manipule"*, Le Monde

`http://www.lemonde.fr/technologies/article/2014/04/10/big-brother-ce-vendeur_4399335_651865.html`

Conclusion

- thème de recherche : la fouille de données
- plus d'information : `http://germain-forestier.info/`
- contact : `germain.forestier@uha.fr`
- merci !