

war government united states military political party national president people world british country state minister union gene acid enzyme reaction chemical compounds municipality group class carbon family products republic synthesis reactions czech area formation fossils genus period back ground fossil died prince egyptian family charles earl century queen de lord king s empera eric institute des sciences lyon universite de lyon imperial empress yuan son jin cricket world cup championships medal men competi event games women olympic born gold time final competition nati russian soviet polish russia moscow union ukrainian ukraine party oblast poland communist city socialist center state german ps species plant plants found disease genus virus water human bacteria soil host leaves food common genetic cells infection animal county state states united served virginia american york district north war carolina school ohio john university march house sc building house built church historic buildings century site national located street style city construction tower stone hall ar roman bc greek empire city century emperor rome ancient king war greece byzantine period battle il time army ad music opera orchestra musical piano symphony composer jazz performed works concert performance major work theatre festival conc romanian hungarian hungary ti bucharest county slovakia village villages moldova part transylvania slovak composed hist school district pennsylvania education students state districts high year schools tax grade public received funding area studer ship navy shi fleet naval sea island british class june coast august service japanese battle port admiral crew ireland irish self year ga german german nazi communis football league team club season born played national career school won cup championship division teams world professional play family species genus moth sea marine found gastropod small mollusk beetle snail brown white distribution wingsp army division battle region israel jewish rabbis jewish judaism torah synagogue jerusalem bible yaj temple pen bible yeshiva orthodox son biblical jewish jews rabbi god hebrew israeli jewish jewish torah synagogue jerusalem bible yaj temple pen bible yeshiva orthodox son biblical poland village county voivodeship district east gmina administrative south north west br references lies approximately capital storm tropical hurricane winds mph tornado mph plant south typhoon food wine made meat rice beer potatoes additional data served white canada canadian ontario quebec toronto atign provincial montreal alberta election province north british nova manitoba ottawa s system data software systems computer user time based windows code network information users version support internet file acc protein cells cell gene proteins dna humans binding genes receptor rna function encoded domain shown expression membrane family energy water power high temperature gas nuclear light surface heat system low time process pressure current liquid material syst israel israeli palestinian arab al palestine jerusalem lebanon jordan tel nucl river lake water area park north island south creek national west region part state dan land valley forest east year card cards coins stamps calendar issued coin silver current events http://mediamining.univ-lyon2.fr/velcin/br sri lanka province located chess region municipality area clear lankan december population comune municipalities borders its station line railway trains service train rail route services opened lines built street company bridge class road north passeng armenian village azerbaijan district rayc formula function number dragon dragons dungeons king tolkien edition elves dragon norse icelandic city iceland lord world war saga battle setting earth middle engine car model cars models rear engines front production ford vehicle v24 and 25 mai 2016 – Lyon power design ethiopia ethiopian somali somalia ballet population region woreda zone reported damc eritria part total mandushu urban women league baseball season game games home runs team series run played major average hit pitcher era career pitched sox al pakistan islamic muslim muhammad all Khan ibn islam iraq iran muslims afghanistan shah arab abu iranian persian bin made called time design white mm black type long color small blue common large hand high similar century standard district province rural iran village population census county families romanized central ye khorasan noted reported kerman azer party election elected state democratic president elections vote house republican member votes candidate senate general politic game season football yards yard team nfl bowl field touchdown coach quarter pass state record goal games lead play language languages english word words form verb vowel spoken dialects forms speakers dialect vowels latin verbs written number

## Résumé de vastes corpus textuels : quelques cas d'application aux données ouvertes

Julien Velcin – Laboratoire ERIC  
<http://mediamining.univ-lyon2.fr/velcin/>

## Journées « Humanités numériques et données ouvertes » 24 et 25 mai 2016 – Lyon

# Plan de la présentation

- Résumer par des modèles thématiques
- Illustration sur plusieurs cas d'étude
- Aller plus loin : cas des données temporelles
- Conclusion et perspectives

# Plan de la présentation

- Résumer par des modèles thématiques
- Illustration sur plusieurs cas d'étude
- Aller plus loin : cas des données temporelles
- Conclusion et perspectives

The screenshot shows the Guardian news website. At the top right, it says 'the guardian'. Below that, there's a navigation menu with 'UK world sport football opinion culture business lifestyle fashion environment tech travel' and a 'browse all sections' button. A 'home' button is also present. On the left, there's a 'headlines' section with the date 'Thursday 28 January 2016' and a weather forecast for Lyon showing a high of 12°C and a low of 9°C. The main content area features several article teasers: 'Zika virus spreading 'explosively', says World Health Organisation', 'Denmark PM's tough stance criticised by international media but has popular support at home', 'Should I cancel my holiday?' (Latest advice for travellers), 'Apples, berries, peppers: Natural compound in fruit and veg could help prevent weight gain - study', and 'US: Marco Rubio, from 'Republican saviour' to prophet of gloom ... and back again'. At the bottom of the screenshot, there's a section for 'THE HUFFINGTON POST UNITED KINGDOM' with a search bar and social media icons for Facebook, Twitter, and Google+. The footer contains various category links like 'FRONT PAGE', 'NEWS', 'POLITICS', 'BUSINESS', 'TECH', 'YOUNG VOICES', 'COMEDY', 'ENTERTAINMENT', 'CELEBRITY', 'LIFESTYLE', 'PARENTS', 'BLOGS' and a list of specific articles like 'Copies', 'Building Modern Men', 'What's Working', 'Environment', 'Media', 'Women', 'Impact', 'Entrepreneurs', 'Young Talent', 'Christmas', and 'Smart Living'.

Google

Actualités

Édition France - Compacte

À la une

Christiane Taubira  
Roger Federer  
Laurent Gbagbo  
Hassan Rohani  
Areva  
Gaël Monfils  
Rihanna  
PSA Peugeot Citroën  
Angelique Kerber  
Transparency  
International

Recommandations

Bron, Rhône-Alpes  
International  
France  
Économie  
Science/High-Tech  
Culture  
Sport  
Santé

À la une

**Tapie: «Le départ de Taubira est mauvais pour tout le monde»**  
Libération - il y a 57 minutes  
Bernard Tapie et Christiane Taubira avaient fait campagne commune aux élections européennes de 1992, partageant entre autre «le même carac  
Valls sur le départ de Taubira: "Résister, c'est se confronter à la réalité" L'Express  
Manuel Valls: «Christiane Taubira va manquer au gouvernement, mais...» Le Figaro

Voir l'actu en direct

Visite de Rohani en France : premiers contrats pour PSA, la SNCF et Total  
Le Figaro - il y a 11 minutes

**INFOGRAPHIE. Zika, ce virus qui menace les bébés et débarque en Europe**  
L'Obs - il y a 1 heure

**À son procès, Laurent Gbagbo plaide non coupable**  
Les Échos - il y a 1 heure

**François Hollande va recevoir la famille de Jacqueline Sauvage à l'Élysée**  
Le Figaro - il y a 1 heure

**De nouvelles menaces visent six lycées parisiens**  
Le Figaro - il y a 51 minutes

Recommandations

**Laurent Ruquier: "Si le Front national passe, je me tire"**  
leJDD.fr - il y a 5 heures  
Fallait-il inviter Manuel Valls à On n'est pas couché? Deux semaines après, l'animateur de l'émission, Laurent Ruquier, répond dans une longue inter  
On n'est pas couché vous intéresse ? Oui | Non

**Touche Pas à Mon Poste : Nabilla invitée par Cyril Hanouna, 3 ...**  
mesty.fr - il y a 3 heures  
Touche pas à mon poste l vous intéresse ? Oui | Non

#journéedelalanguefrançaise

Top Direct Comptes Photos Vidéos Autres options

Suggestions · Actualiser · Tout afficher

4 nouveaux résultats

**NyTx** @NyTxSw · 1 min  
#JournéeDeLaLangueFrançaise on va donc éviter tous ces horribles anglicismes qui tuent lentement notre langue.

**servietsky** @servietsky74 · 1 min  
il va falloir fermer twitter #JournéeDeLaLangueFrançaise

**QUENTIN** @Nitneuc\_ · 1 min  
POUAHAHAHAH #JournéeDeLaLangueFrançaise

**ben&jerrys&ana** @cgdornan · 1 min  
Pourquoi faire une journée pour cette langue si c'est pour la massacrer avec une réforme par la suite? #JournéeDeLaLangueFrançaise

**Moins gentil ligné** @ParathorO · 2 min  
#JournéeDeLaLangueFrançaise zig

**ben&jerrys&ana** @cgdornan · 3 min  
Si vous voulez honorer la langue française alors s'il vous plaît pas de "ognon" #JournéeDeLaLangueFrançaise

Tendances · Modifier

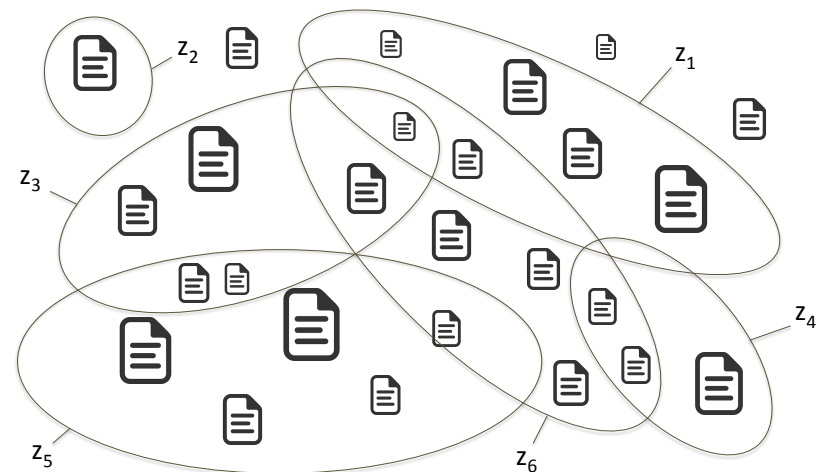
#JournéeDeLaLangueFrançaise  
#BourdinDirect  
#SRFCOL  
#SOSPascal  
#BrunoFunRadio  
Lacazette  
Troyes  
Olivier Bourdeaut  
Albert Einstein  
Dany Laferrière

## Apprendre les thématiques ?

- Les modèles de thématiques ont pour objectif d'organiser **automatiquement** de vastes collections de données textuelles
  - découvrir les thèmes qui « émergent »
  - indexer les documents à l'aide de ces thèmes
  - utiliser ces annotations pour résumer, naviguer, rechercher... donc mieux appréhender les données

## Clustering vs. thématiques

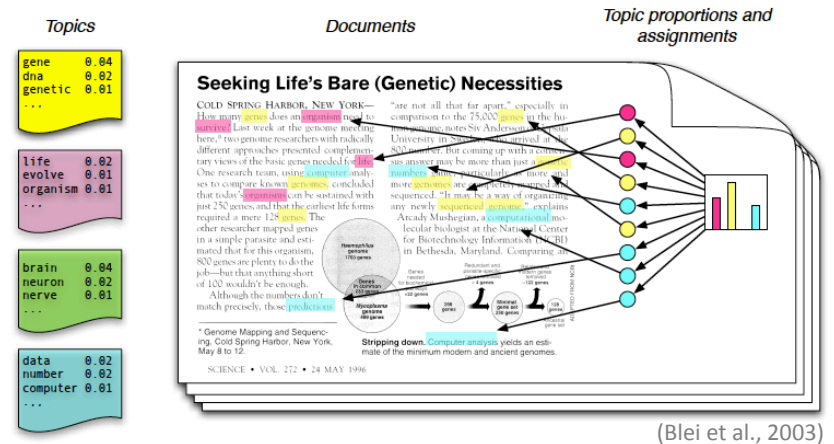
(Xie and Xing, 2013)



# Eléments d'état de l'art

- Approches algébriques
  - LSA (Deerwester et al., 1990)
  - NMF (Paatero et Tapper, 1994)
  - apprentissage de dictionnaires (Jenatton et al., 2010)
- Approches géométriques
  - issus du TDT (Allan et al., 1998) (Pons-Porrata et al., 2003)
  - AGAPE (Velcin and Ganascia, 2007)
- Approches probabilistes
  - pLSA, LDA... (voir la suite de l'exposé)

# Modèle de thématiques : cas de LDA



(Blei et al., 2003)

# Manipuler des données textuelles

- Hypothèse du sac de mots
- Etape de prétraitements :
  - suppression des « mots-outils »
  - suppression des marques de ponctuations
  - suppression des nombres
- En entrée :



| Terms        | Docs |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |
|--------------|------|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
|              | 1    | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| data         | 1    | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0  | 1  | 2  | 1  | 1  | 1  | 0  | 1  | 0  | 0  | 0  |
| examples     | 0    | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| introduction | 0    | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  |
| mining       | 0    | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  | 0  | 1  | 1  | 0  | 1  | 0  | 0  | 0  | 0  | 0  |
| network      | 0    | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 1  | 1  |
| package      | 0    | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

# Plan de la présentation

- Résumer par des modèles thématiques
- **Illustration sur plusieurs cas d'étude**
- Aller plus loin : cas des données temporelles
- Conclusion et perspectives

## Illustration avec quelques jeux de données (oui, qui ne sont pas tous libres...)

- Articles scientifiques
  - 20 Newsgroups
  - Série Harry Potter
- en collaboration avec P. Poncelet, M. Roche et J.A. Lossio (LIRMM)
- en collaboration avec C. Gravier (LHC)

13

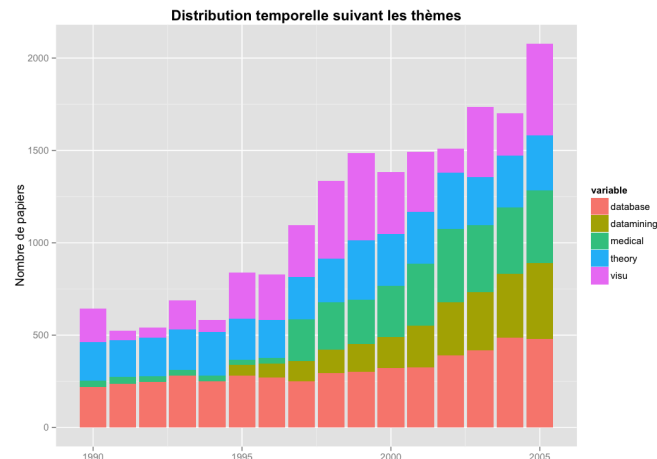
## Articles scientifiques

+ de 18000 titres et/ou résumés d'articles publiés entre 1990 et 2005 (Tang et al., 2012)

- base de données : ICDE, VLDB, SIGMOD...
- data mining (après 1994) : KDD, ICDM...
- visualisation : CVPR, InfoViz, ICCV...
- informatique théorique : FOCS, SODA...
- informatique médicale : JAMIA, AIME...

14

## Evolution des thématiques



15

## Thématiques extraites par LDA (vocabulaire de 5000 mots)

**Ida 0** : data - query - queries - database - performance - xml - system - processing - systems - relational - paper - efficient - databases - algorithms - memory - techniques - access - results - storage - time - optimization - present - index - distributed - show - structure - operations - approach - model - join...

**Ida 1** : algorithm - problem - time - algorithms - graph - show - number - problems - approximation - graphs - bound - lower - bounds - complexity - set - optimal - case - polynomial - random - log - constant - linear - results - size - result - network - general - present - tree - model...

**Ida 2** : image - images - method - surface - motion - model - object - algorithm - visualization - paper - volume - approach - data - rendering - objects - shape - points - present - models - flow - results - methods - technique - segmentation - reconstruction - surfaces - recognition - point - tracking - structure...

**Ida 3** : data - mining - algorithm - clustering - learning - paper - approach - classification - results - method - algorithms - methods - problem - patterns - large - set - model - sets - analysis - show - number - time - models - present - search - performance - detection - association - pattern - efficient...

**Ida 4** : data - information - system - systems - research - database - paper - web - visualization - user - model - application - management - knowledge - applications - design - users - databases - medical - analysis - integration - semantic - support - technology - development - network - environment - issues - language - process...

16

# Thématiques extraites par LDA

(vocabulaire de 5000 ngrams, n>1)

**lda 0** : volume rendering - case study - research paper - vector fields - decision support - information technology - visualization techniques - a case study - volume data - vector field...

**lda 1** : data mining - time series - experimental results - knowledge discovery - machine learning - nearest neighbor - support vector - feature selection - decision tree - association rule...

**lda 2** : lower bound - lower bounds - polynomial time - approximation algorithms - extended abstract - approximation algorithm - running time - upper bound - competitive ratio - high probability...

**lda 3** : database systems - query processing - database system - query optimization - xml data - data management - query language - database management - management systems - relational database...

**lda 4** : experimental results - object recognition - computer vision - image sequences - optical flow - extended abstract - image segmentation - pattern matching - real images - motion estimation (...) a new approach...

17

From: ahlenius@rtsg.mot.com (Mark Ahlenius)  
Subject: converting color gif to X pixmap

I have looked through the FAQ sections and have not seen a answer for this.

I have an X/Motif application that I have written. I have a couple of gif files (or pict) that I have scanned in with a color scanner. Now I would like to be able to convert the gif files into a format that could be read into my application and displayed on the background of its main window. Preferably with pixmaps, or perhaps as an XImage.

I have found functions in the pbmplus program suite to convert gif to xbm, but that is monochrome, and I really do need color.

I have looked at xv, which reads in gif, and writes out several formats, but have not found a way to write out a file which can be read in as a pixmap.

Is there an easy way to do this?

**catégorie :**  
**comp.windows.x**

19

# 20 NewsGroups

- 20 000 textes répartis dans 20 catégories  
<http://qwone.com/~jason/20Newsgroups/>

|   |  |   |
|---|--|---|
| comp.graphics<br>comp.os.ms-windows.misc<br>comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>comp.windows.x | rec.autos<br>rec.motorcycles<br>rec.sport.baseball<br>rec.sport.hockey | sci.crypt<br>sci.electronics<br>sci.med<br>sci.space        |
| misc.forsale  | talk.politics.misc<br>talk.politics.guns<br>talk.politics.mideast      | talk.religion.misc<br>alt.atheism<br>soc.religion.christian |

18

From: leech@cs.unc.edu (Jon Leech)  
Subject: Space FAQ 15/15 - Orbital and Planetary Launch Services

Archive-name: space/launchers  
Last-modified: \$Date: 93/04/01 14:39:11 \$

ORBITAL AND PLANETARY LAUNCH SERVICES

**catégorie :**  
**sci.space**

The following data comes from \_International Reference Guide to Space Launch Systems\_ by Steven J. Isakowitz, 1991 edition.

Notes:

- \* Unless otherwise specified, LEO and polar payloads are for a 100 nm orbit.
- \* Reliability data includes launches through Dec, 1990. Reliability for a family of vehicles includes launches by types no longer built when applicable
- \* Prices are in millions of 1990 \$US and are subject to change.
- \* Only operational vehicle families are included. Individual vehicles which have not yet flown are marked by an asterisk (\*) If a vehicle had first launch after publication of my data, it may still be marked with an asterisk.

20

| Vehicle (nation) | Payload kg (lbs) | Reliability | Price | Launch Site    |
|------------------|------------------|-------------|-------|----------------|
|                  | LEO Polar GTO    |             |       | (Lat. & Long.) |

|              |                 |                |                |                        |
|--------------|-----------------|----------------|----------------|------------------------|
| Ariane (ESA) | 35/40           | 87.5%          |                | Kourou (5.2 N, 52.8 W) |
| AR40         | 4,900 (10,800)  | 3,900 (8,580)  | 1,900 (4,190)  | 1/1 \$65m              |
| AR42P        | 6,100 (13,400)  | 4,800 (10,600) | 2,600 (5,730)  | 1/1 \$67m              |
| AR44P        | 6,900 (15,200)  | 5,500 (12,100) | 3,000 (6,610)  | 0/0 ? \$70m            |
| AR42L        | 7,400 (16,300)  | 5,900 (13,000) | 3,200 (7,050)  | 0/0 ? \$90m            |
| AR44LP       | 8,300 (18,300)  | 6,600 (14,500) | 3,700 (8,160)  | 6/6 \$95m              |
| AR44L        | 9,600 (21,100)  | 7,700 (16,900) | 4,200 (9,260)  | 3/4 \$115m             |
| * AR5        | 18,000 (39,600) | ???            | 6,800 (15,000) | 0/0 \$105m [300nm]     |

**catégorie :  
sci.space  
(suite)**

21

## Thématiques extraites par LDA (vocabulaire de 10000 mots)

Extrait des 20 thématiques demandées :

**Ida 5 :** window - file - program - server - set - motif - widget - application - problem - entry - display - code - sun - error - xterm - manager - running - work - subject - open - make - line - openwindows - number - size - x11r5 - function - run - version - client...

**Ida 6 :** image - file - jpeg - images - format - color - files - gif - program - display - version - bit - printer - convert - quality - programs - software - screen - formats - xv - good - colors - print - graphics - free - article - windows - postscript - tiff - fonts...

**Ida 18 :** god - jesus - church - bible - christ - christian - people - christians - sin - lord - faith - love - life - man - paul - word - law - time - article - good - heaven - hell - father - christianity - john - homosexuality - spirit - scripture - holy - things...

**Ida 19 :** space - nasa - launch - earth - article - orbit - shuttle - moon - mission - system - satellite - solar - time - spacecraft - data - years - lunar - station - flight - sky - cost - mars - project - venus - high - pat - surface - planet - program - henry...

22

## Et dans

extrait sur  
20 thématiques :

### Maisons de sorcellerie

house 0.04657586  
gryffindor 0.04424846  
points 0.03416309  
slytherin 0.03261149  
hundred 0.02252612  
hat 0.02175032  
will 0.02097452  
cup 0.01554393  
hufflepuff 0.01399234  
taken 0.01321654

### Quidditch

wood 0.04386071  
quidditch 0.03491612  
team 0.02060479  
quaffle 0.01702696  
snitch 0.01613250  
game 0.01523804  
catch 0.01344912  
play 0.01255466  
flint 0.01255466  
seeker 0.01166021

### Famille Weasley

weasley 0.04050274  
percy 0.03038466  
fred 0.02869831  
george 0.02448244  
twins 0.01942340  
year 0.018580??

### Professeurs de Poudlard

professor 0.141749006  
mcgonagall 0.074869763  
dumbledore 0.035060689  
quirrell 0.022321786  
flitwick 0.015952334  
turban 0.011175245  
reached 0.007990520  
teacher 0.007990520  
dumbledore's 0.007194338  
talking 0.007194338

### 4, Private Drive

uncle 0.065995844  
dudley 0.062179040  
vernon 0.057271720  
aunt 0.035461411  
petunia 0.031099349  
letter 0.018013164  
dudley's 0.012560586  
room 0.012015328  
cupboard 0.012015328

23

## Mais aussi...

looked 0.03222237  
like 0.02515218  
eyes 0.02122431  
long 0.02122431  
little 0.01886758  
black 0.01886758  
took 0.01611806  
hat 0.01572528  
face 0.01533249  
pulled 0.01533249  
said 0.08080992  
get 0.02566909  
know 0.02420258  
back 0.02317602  
see 0.01877648  
something 0.01862983  
think 0.01804323  
now 0.01730997  
going 0.01716332  
well 0.01511021

harry 0.07567148  
one 0.03453437  
first 0.02718846  
time 0.01739391  
much 0.01616959  
next 0.01543500  
never 0.01494527  
day 0.01298636

hagrid 0.06322251  
yeh 0.06136366  
ter 0.04835173  
yer 0.03719865  
said 0.02170826  
dragon 0.01861018  
fer 0.01799056  
gringotts 0.01737095  
got 0.01675133  
don 0.01613172

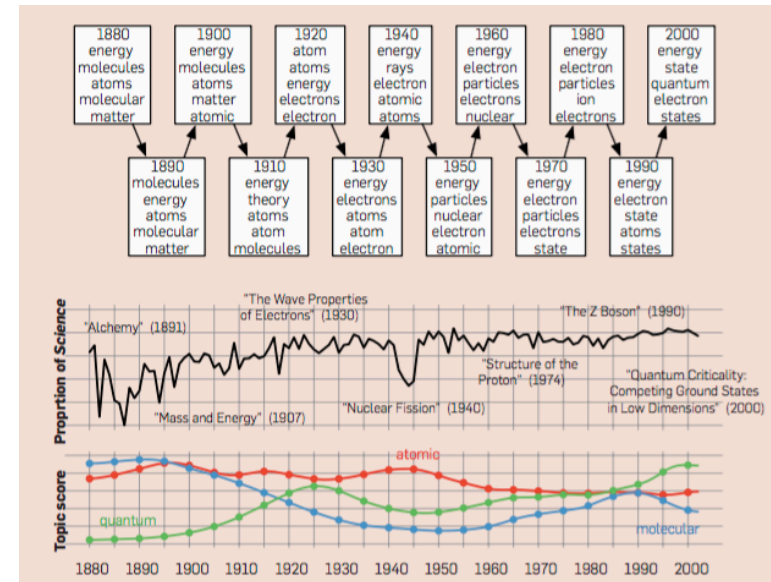
door 0.03846168  
open 0.02367537  
cloak 0.02121098  
looking 0.01874660  
two 0.01874660  
floor 0.01677509  
forward 0.01677509

24



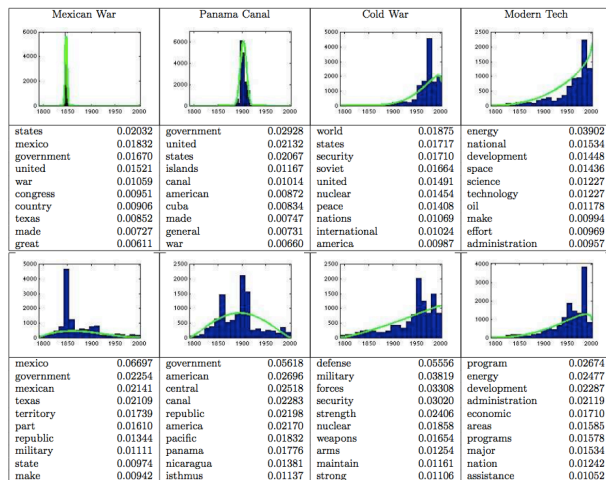
# Plan de la présentation

- Résumer par des modèles thématiques
- Illustration sur plusieurs cas d'étude
- **Aller plus loin : cas des données temporelles**
- Conclusion et perspectives

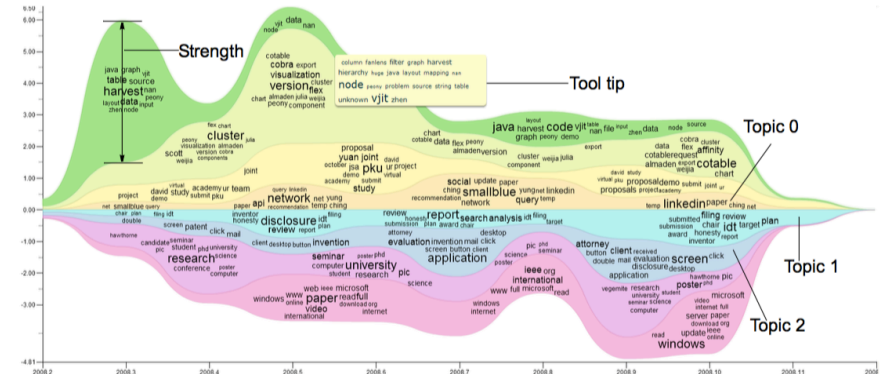


données tirées de *Science* entre 1880 et 2002 (Blei, 2012)

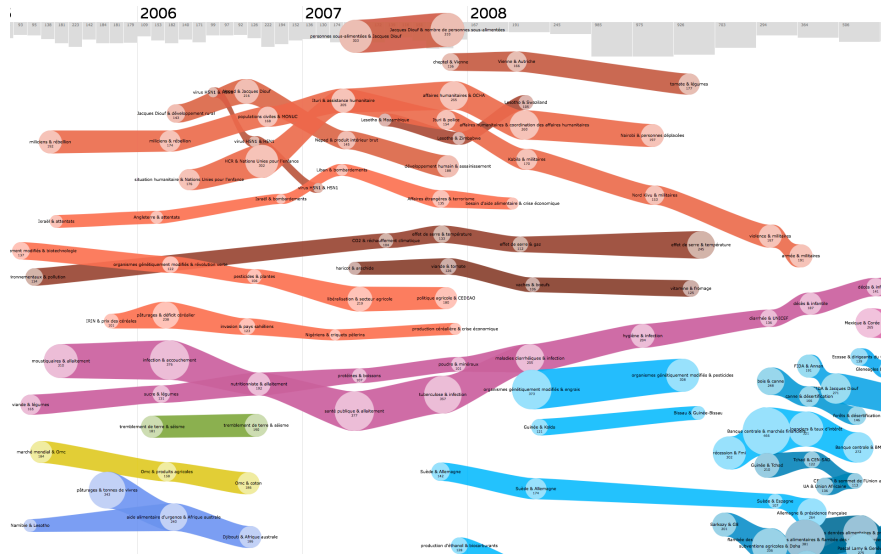
# Résultats sur le discours de l'union



# Dynamique des thématiques



TIARA (Liu et al., 2009) (Wei et al., 2010)



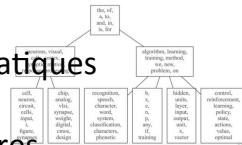
Projet Pulseweb : <http://pulseweb.cortex.net>

## Plan de la présentation

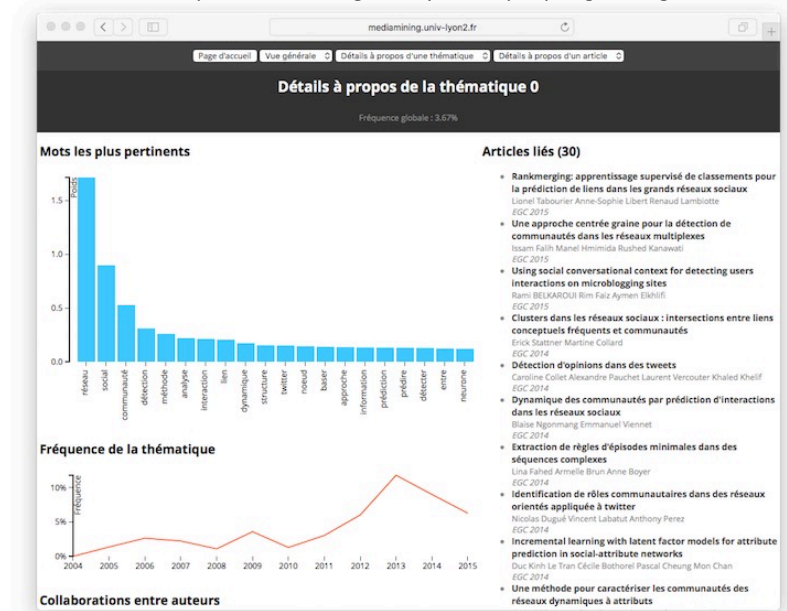
- Résumer par des modèles thématiques
- Illustration sur plusieurs cas d'étude
- Aller plus loin : cas des données temporelles
- **Conclusion et perspectives**

## Quelques questions intéressantes

- Trouver le « meilleur » nombre de thématiques  
HDP (Teh et al., 2004)
- Résumer l'information portée par une thématique  
Topic labeling (Mei et al., 2007)
- Trouver une structure entre les thématiques  
hLDA (Griffiths et al., 2004) CTM (Blei et al., 2006)
- Combiner les thématiques avec d'autres informations : auteur, opinion, structure, etc.  
Author-Topic (Rosen-Zvi et al., 2006)  
Topic-Opinion (Dermouche et al., 2014) (et tant d'autres...)
- Liens avec le « plongement de mots » (Das et al., 2015)



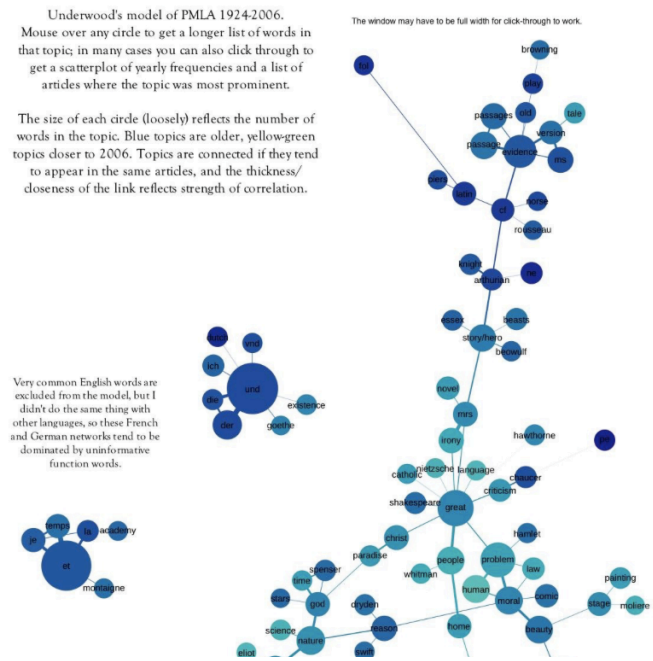
TOM @ERIC: <http://mediamining.univ-lyon2.fr/people/guille/egc2016/>





# Références orientées DH

- Blei, D.M., A.Y. Ng and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: pp. 993–1022.
- Burton M. (2013). The Joy of Topic Modeling: A bag of words. <http://mcburton.net/blog/joy-of-tm>
- Goldstone A. and T. Underwood (2012). What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship? <http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone>
- Graham S. et al. (2012). Getting Started with Topic Modeling and MALLET. Online lesson. <http://programminghistorian.org/lessons/topic-modeling-and-mallet>
- Weingart, S. (2012). Topic Modeling for Humanists: A Guided Tour. <http://www.scottbot.net/HIAL/?p=19113>



<http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone>