

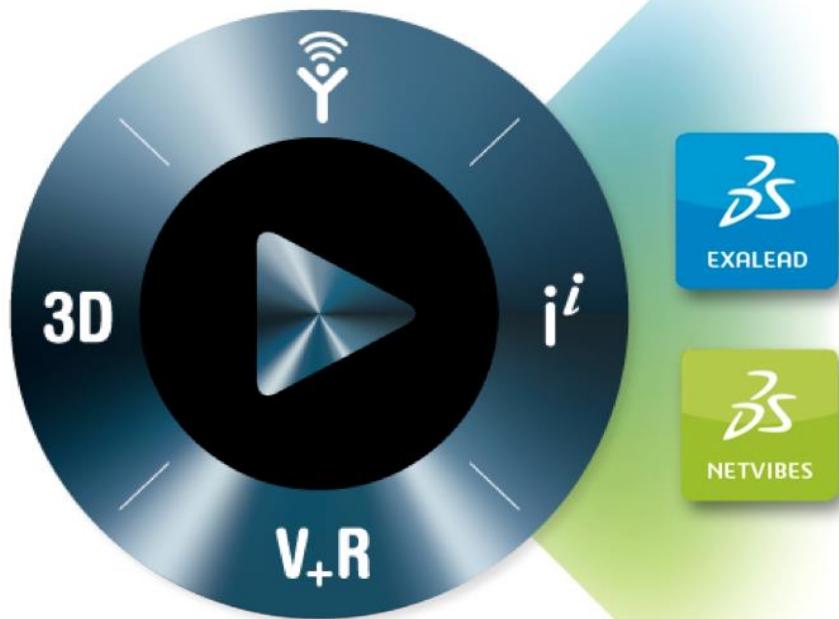


## Mo3: Big Data, Web & (Cyber)security

Laura WILBER

Director of Strategy, Dassault Systèmes EXALEAD

# Dassault Systèmes EXALEAD



## « Information Intelligence »

- Search & Discovery
  - Entreprise
  - Web
- « ii » du monde réel pour alimenter le monde virtuel de DS

# Agenda

---

1. Qu'est-ce que ça veut dire, « Big Data » ?
2. Qu'est-ce que ça veut dire, « Sécurité » (« Cybersécurité ») ?
3. Le Web comme source Big Data
4. On peut faire quoi avec le Web ?
5. On peut le faire comment ?

---

## 1. QU'EST-CE QUE ÇA VEUT DIRE, « BIG DATA » ?

# Big Data : de nouveau?

---

Gros volumes de données?

De nouveau ? Non !

- Grid computing
- Super computers
- High-end datawarehouses



# Quoi de nouveau ?

- Taux de croissance

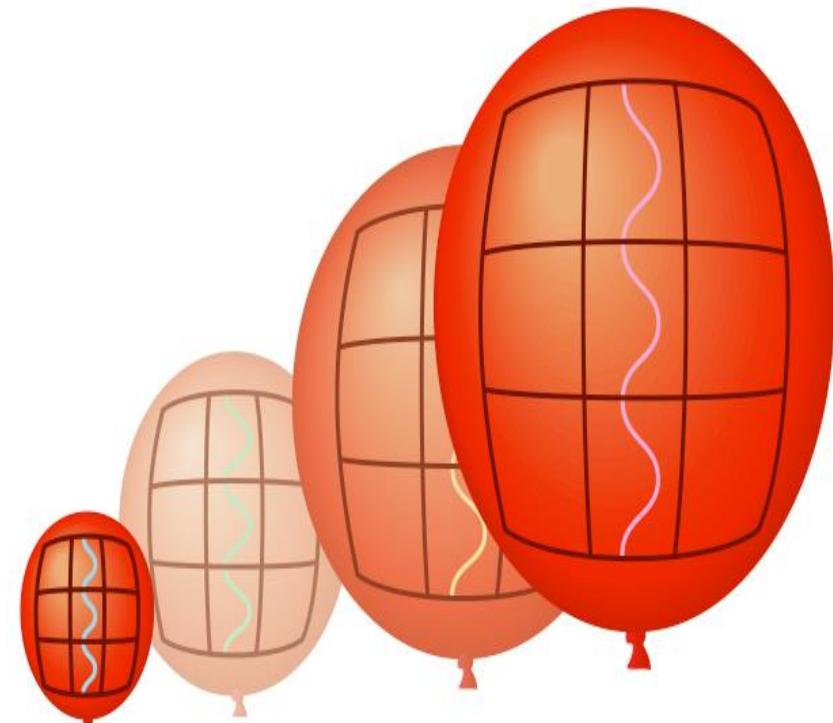


**1 Po/15 sec.**



LARGE HADRON COLLIDER

**1 Po/sec**

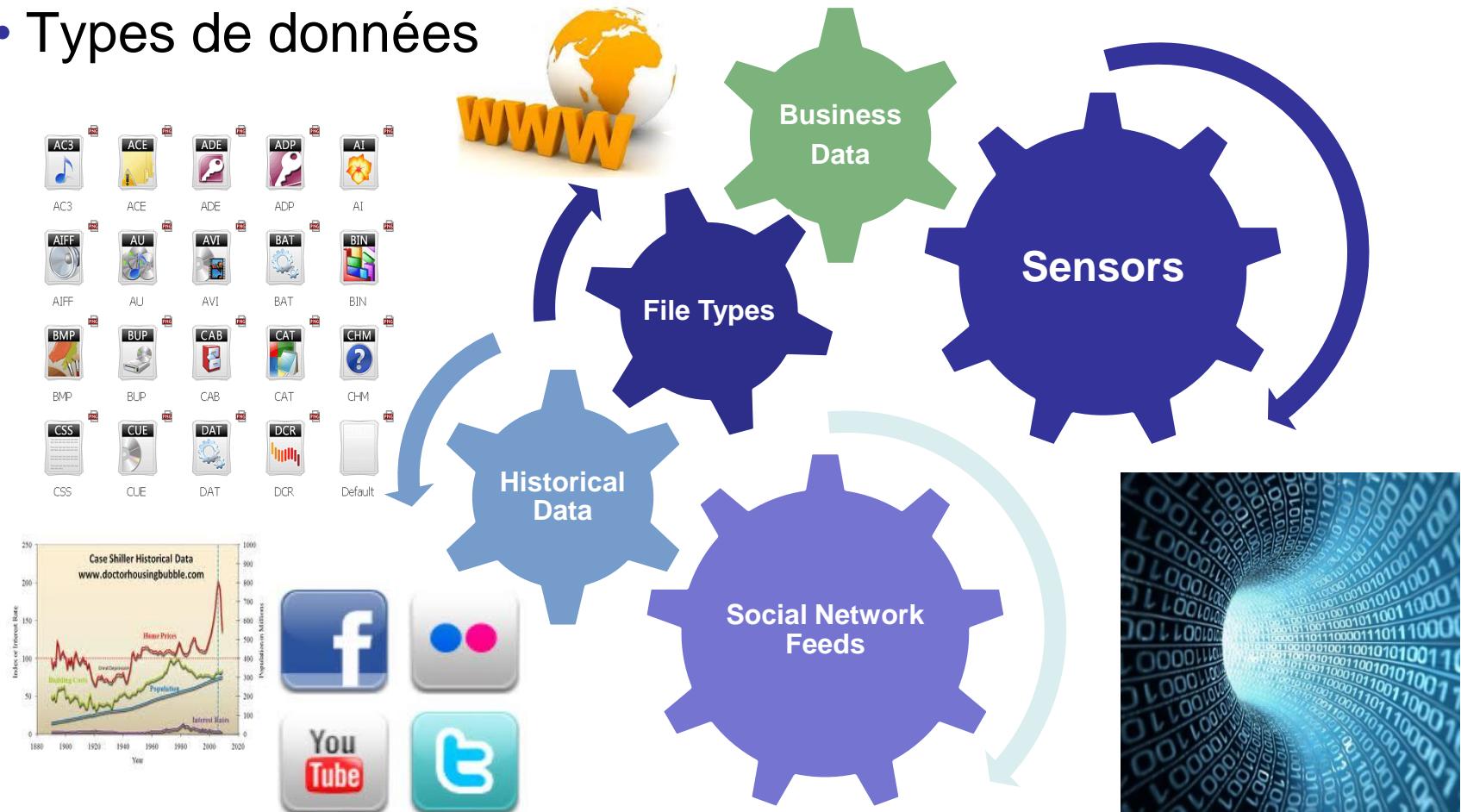


Copyright © Addison Wesley

**+40% an**  
**2015: 8 Zö 8 trillion Go**

# Quoi de nouveau ?

- Types de données



# Quoi de nouveau ?

- Vitesse (et parcours – P2P)



# Quoi de nouveau ?

- Nouvelle technologies



# Quoi de nouveau ?

---

- « Nouvelle » techniques

A word cloud diagram centered around the concept of data processing and machine learning. The words are arranged in a roughly circular pattern, with the most central and largest word being 'clustering'. Other prominent words include 'classification', 'data fusion', 'natural language processing', and 'machine learning'. Smaller words surrounding these include 'sentiment analysis', 'semantic processing', 'statistics', 'pattern recognition', 'association rule learning', 'indexing', 'distributed processing', 'data visualization', 'network analysis / relationship mapping', 'text mining', 'ensemble learning', and 'signal processing'.

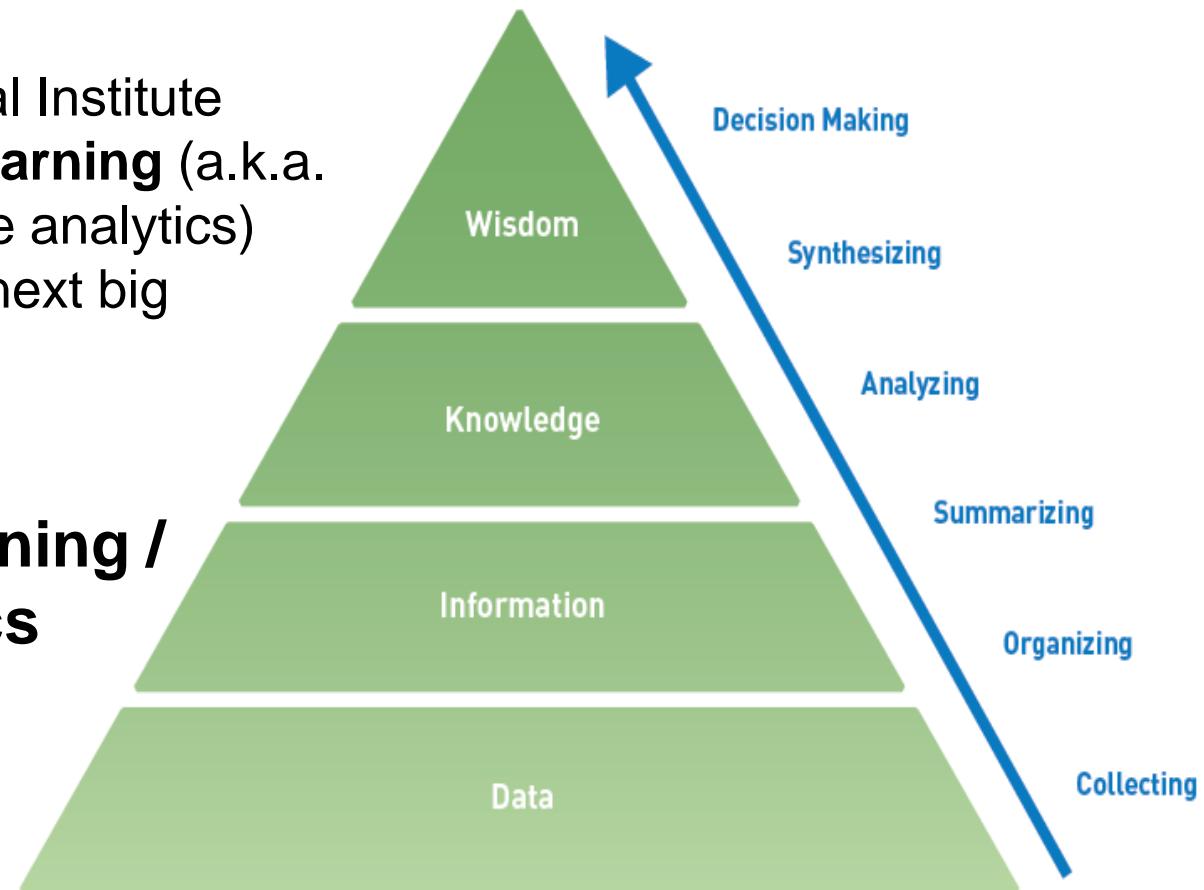
sentiment analysis  
spatial analysis semantic processing  
natural language processing statistics  
text mining clustering ensemble learning  
network analysis / relationship mapping pattern recognition  
classification machine learning  
distributed processing data fusion  
data visualization indexing association rule learning

# Quoi de nouveau ?

- Trouver le trésor caché pour l'avantage concurrentiel

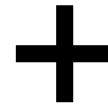
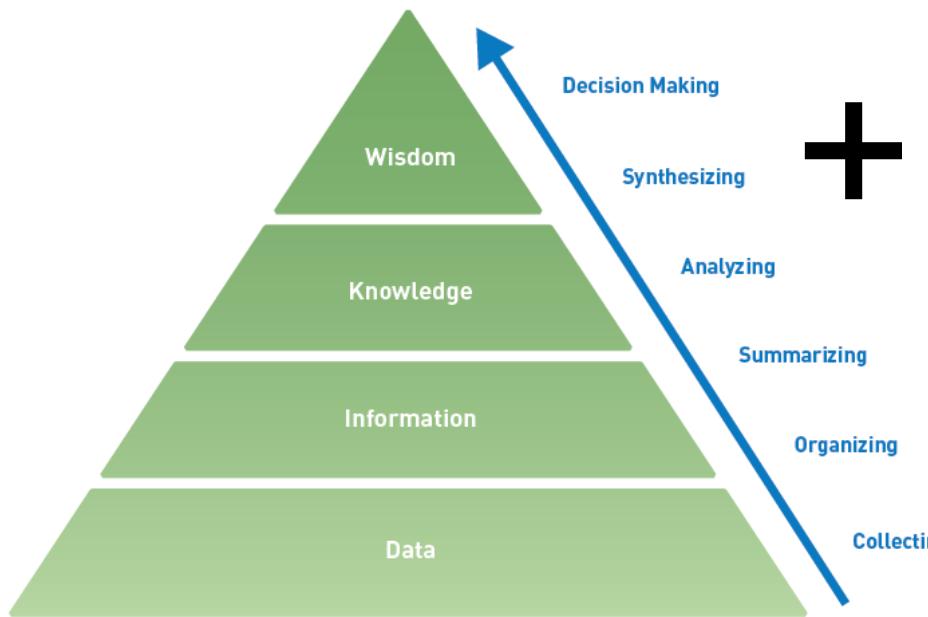
« ...the McKinsey Global Institute asserts that **machine learning** (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of innovation »

**Big Data = Data Mining / Predictive Analytics**



# Le « Saint Graal »

- Rendre accessible aux gens ordinaires, au quotidien





## 2. QU'EST-CE QUE ÇA VEUT DIRE, « SECURITÉ » ?

# La Sécurité

## Gestion des risques, les menaces:

- Déetecter
- Anticiper
- Analyser
- Agir
  - Éviter
  - Éliminer
  - Minimiser
- Réagir



# La Sécurité « Nationale »

---

- Classique



# La Sécurité « Nationale »

## Attentats aux civils

- Associé d'un état, ou non
- Domestique, ou non
- Organisé, ou non
- Individuel, ou non



# La Sécurité en Evolution

## Menaces selon US Army

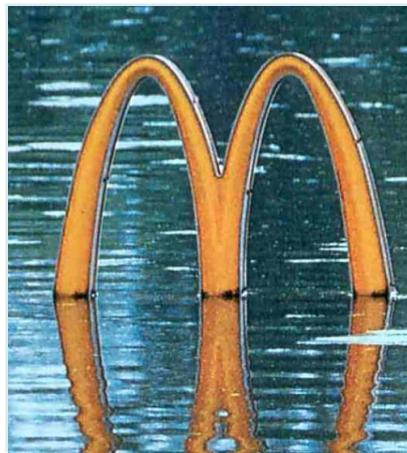
- Terrorisme
- Prolifération des armes, ADM
- Les organisations criminelles transnationales
- Cybersecurité
- Une croissance économique inégale
- Vulnérabilité du système financière mondiale
- Catastrophes naturelles
- Impacts de la démographie
- La rareté des ressources
- Les pressions environnementales croissantes  
(i.e., « changement climatique » - à huit clos)

# La Sécurité « Business »



## Traditionnelle

- Concurrence
- Défis financière
- L'offre de travail
- Réglementation
- Qualité, 'churn' clientale
- Changements dans la demande...



## Convergente

- Le changement climatique
  - Catastrophes, perturbations multivalents
  - Modifications de l'environnement durables
- La rareté des ressources
- Cybersecurité
- Terrorisme...

# Confluence sur « Sécurité »

## Rank Country / Corporation GDP / sales (\$mil)

1	United States	8,708,870.00	23	<i>General Motors</i>	176,558.00	52	<i>AXA</i>	87,645.70	80	<i>Nissan Motor</i>	53,679.90
2	Japan	4,395,083.00	24	Denmark	174,363.00	53	<i>IBM</i>	87,548.00	81	New Zealand	53,622.00
3	Germany	2,081,202.00	25	<i>Wal-Mart</i>	166,809.00	54	Singapore	84,945.00	82	<i>E.On</i>	52,227.70
4	France							84,861.00	83	<i>Toshiba</i>	51,634.90
5	United Kingdom							83,556.00	84	<i>Bank of America</i>	51,392.00
6	Italy							82,005.00	85	<i>Fiat</i>	51,331.70
7	China							80,072.70	86	<i>Nestle</i>	49,694.10
8	Brazil							78,515.10	87	<i>SBC Communications</i>	49,489.00
9	Canada							75,350.00	88	<i>Credit Suisse</i>	49,362.00
10	Spain							75,337.00	89	Hungary	48,355.00
11	Mexico							74,634.00	90	<i>Hewlett-Packard</i>	48,253.00
12	India							74,178.20	91	<i>Fujitsu</i>	47,195.90
13	Korea, Rep.							71,858.50	92	Algeria	47,015.00
14	Australia							71,092.00	93	<i>Metro</i>	46,663.60
15	Netherlands							65,555.60	94	<i>Sumitomo Life Insur.</i>	46,445.10
16	Russian Federation							65,393.20	95	Bangladesh	45,779.00
17	Argentina	281,942.00	39	<i>Toyota Motor</i>	115,670.90	67	<i>Nissan Motor</i>	62,492.40	96	<i>Tokyo Electric Power</i>	45,727.70
18	Switzerland	260,299.00	40	<i>General Electric</i>	111,630.00						45,351.60
19	Belgium	245,706.00	41	<i>Itochu</i>	109,068.90						44,990.30
20	Sweden	226,388.00	42	Portugal	107,716.00						44,828.00
21	Austria	208,949.00	43	<i>Royal Dutch/Shell</i>	105,366.00						44,637.20
22	Turkey	188,374.00	44	Venezuela	103,918.00						
				45 Iran, Islamic rep.	101,073.00						
				46 Israel	99,068.00						
				47 <i>Sumitomo</i>	95,701.60						
				48 <i>Nippon Tel &amp; Tel</i>	93,591.70						
				49 Egypt, Arab Republic	92,413.00	78	<i>Honda Motor</i>	54,773.30			
				50 <i>Marubeni</i>	91,807.40	79	<i>Assicurazioni Generali</i>	53,723.20			
				51 Colombia	88,596.00						

**100 plus grandes entités économiques:  
51 entreprises, 49 pays;  
risques sans frontières**

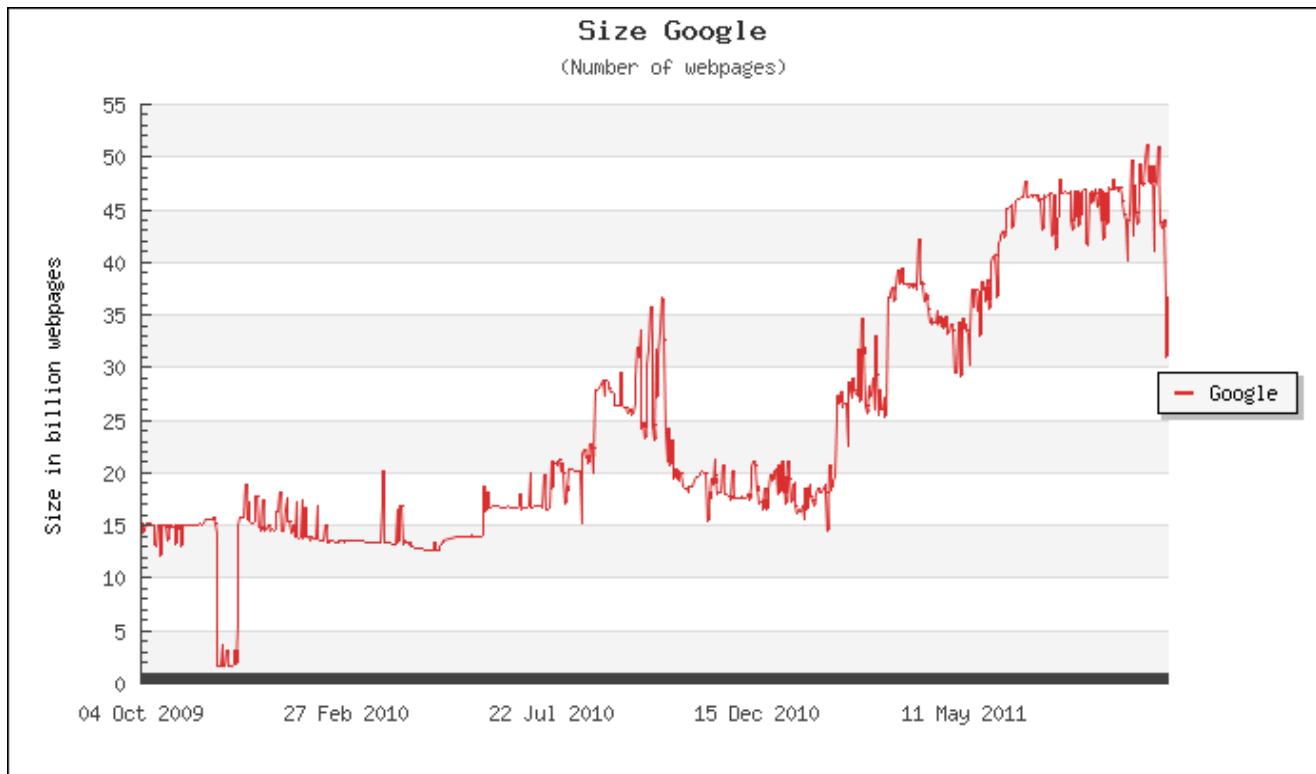
**Diff - “agir” et mission:  
Protéger le citoyen,  
Protéger l’entreprise (?)**

---

### **3. LE WEB COMME SOURCE BIG DATA**

## Vraiment 'Big'

### Moteurs de recherche



Google: 30MM (?)  
EXALEAD: 18MM  
Bing/Yahoo: 8MM (?)

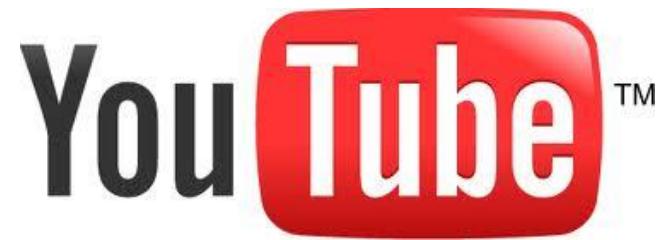
**Google™ Index 100Po, 1 trillion URLs indexés – 950 trillion non-indexés !**

## Vraiment 'Big'

The Facebook logo, featuring the word "facebook" in its signature white lowercase font, centered on a solid blue rectangular background.

- > 1 billion users
- > 300PB; +> 500TB/day
- > 35% of world's photographs

## Vraiment 'Big'



> 1000PB  
+>72 hours/minute  
>37 million hours/year  
> 4 billion views/day

## Vraiment 'Big'

---

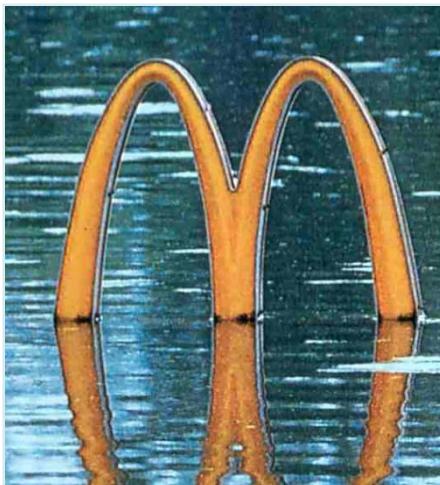


> 124B tweets/year  
> 390M/day  
~4500/sec

---

## 4. ON PEUT FAIRE QUOI AVEC TOUT CA?

# Qu'est-ce qu'on peut faire?



**(OSINT – Biz et Govt)**

# Qu'est-ce qu'on peut faire?



## 7 Universal Constructs for Analytics



People



Places



Organizations



Time



Events



Concepts



Things

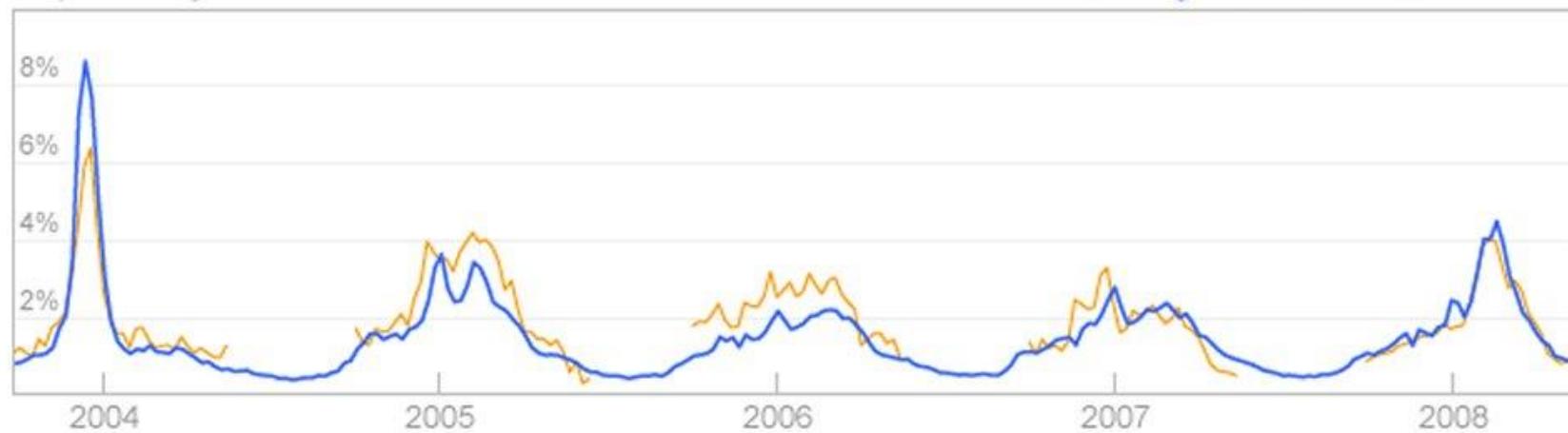
ClearStory DATA

# Nouveaux source & type de données, 'nouvelles' méthodes d'analyse

Annual U.S. Flu Activity - Mid-Atlantic Region

ILI percentage

● Google Flu Trends ● CDC Data



- Même précision que U.S. Centers for Disease Control and Prevention (CDC)
- Plus vite – 2 semaines

“Invariably, **simple models** and a **lot of data** trump more elaborate models based on less data.”

— Alon Halevy, Peter Norvig & Fernando Pereira

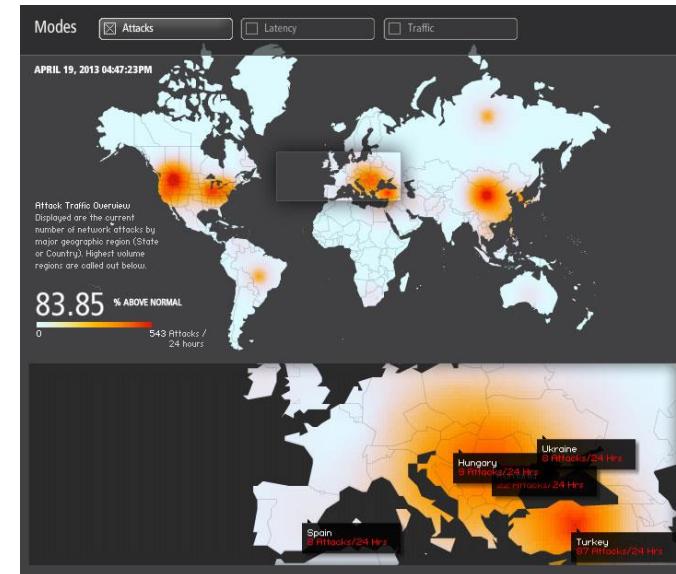
# On peut faire quoi ?

## Analyse



Plutôt historique

## Surveillance



Plutôt temps réel

# Analyse par 'data mining/machine learning'

## Analyse descriptive

Creuser des données à découvrir des faits, des tendances, des groupes, des patterns



## Analyse prédictive

Construire des modèles à prévoir qu'est-ce que se passera dans l'avenir



## Analyse prescriptive

Construire des modèles pour faire des prescriptions/ recommendations



Actuelles et historique

Historique

Historique



# Analyse descriptive : Construire des profiles riches

**BEFORE**

**AFTER**

**1 Becco**  
★★★★★ (124) [Read Reviews](#) | [Write a Review](#)  
355 West 46th Street, New York, NY zip code  
[Click here for Becco Website](#)

[► Phone](#) | [More Info](#) | [Map it](#) | [Features - Restaurant Reservations](#)

**Comments** | **Menu**

loved by locals and out of towners alike. enjoy signature dishes, like the osso bucco, as well as the bounty of fresh seasonal preparations.the reserve wine list has many of italy's best known producers, as well as some more eclectic wine makers. a knowledgeable staffer will be on hand to help you make a great choice.with 15 years on restaurant... [More >>](#)

From OpenTable

**Most viewed comment**

«We went here recently on a Friday night and were terribly disappointed. The restaurant was incredibly crowded, which we expected, but what made it unbearable was that the owner has used every square inch to squeeze a table into. We were in a main aisle table, so every time a waiter had to get to a table past us, we were bumped or reached over. Plus the tables are very small, and the pedestal is fat and in the middle, so you really don't have many choices as to where to put your legs/feet comfortably. We do eat in the city often, I understand the need for maximizing space, but this w... [More >>](#)

From New York Magazine

**Bloggers entries about Becco**

- Bastianich's Becco: Sinfonia di Pasta Dinner for \$21.95 ([New York City's Hottest Restaurant Deals](#))
- "Sinfonia di Pasta": \$21.95 prix fixe at Becco ([New York City's Hottest Restaurant Deals](#))
- Digging for nuggets of goodness at the Taste of Times Square ([Midtown Lunch](#))
- Who wants "A Taste of Times Square"? ([Midtown Lunch](#))

**Details**

Payment options: AMEX, Diners Club, Discover, MasterCard, Visa

Chef: William Gallagher

Price range: Moderate, Expensive

**Opening Hours**

monday	5pm - 12am / 12pm - 3pm
tuesday	5pm - 12am / 12pm - 3pm
wednesday	5pm - 12am / 12pm - 3pm
thursday	5pm - 12am / 12pm - 3pm
friday	5pm - 12am / 12pm - 3pm
saturday	5pm - 12am / 12pm - 3pm
sunday	5pm - 12am / 12pm - 3pm

**People sentiments**

reasonable okay spicy baked hot maximizing welcoming unbearable famous hard grilled impressed tasteful stated falling adjacent mundane relaxed

impressive surrounded **amazing** odd unimaginative negative extensive expensive superb unbeatable efficient snobby endless tiny available meager cute typical outrageous yummy adventurous

**good** tasteless rude super

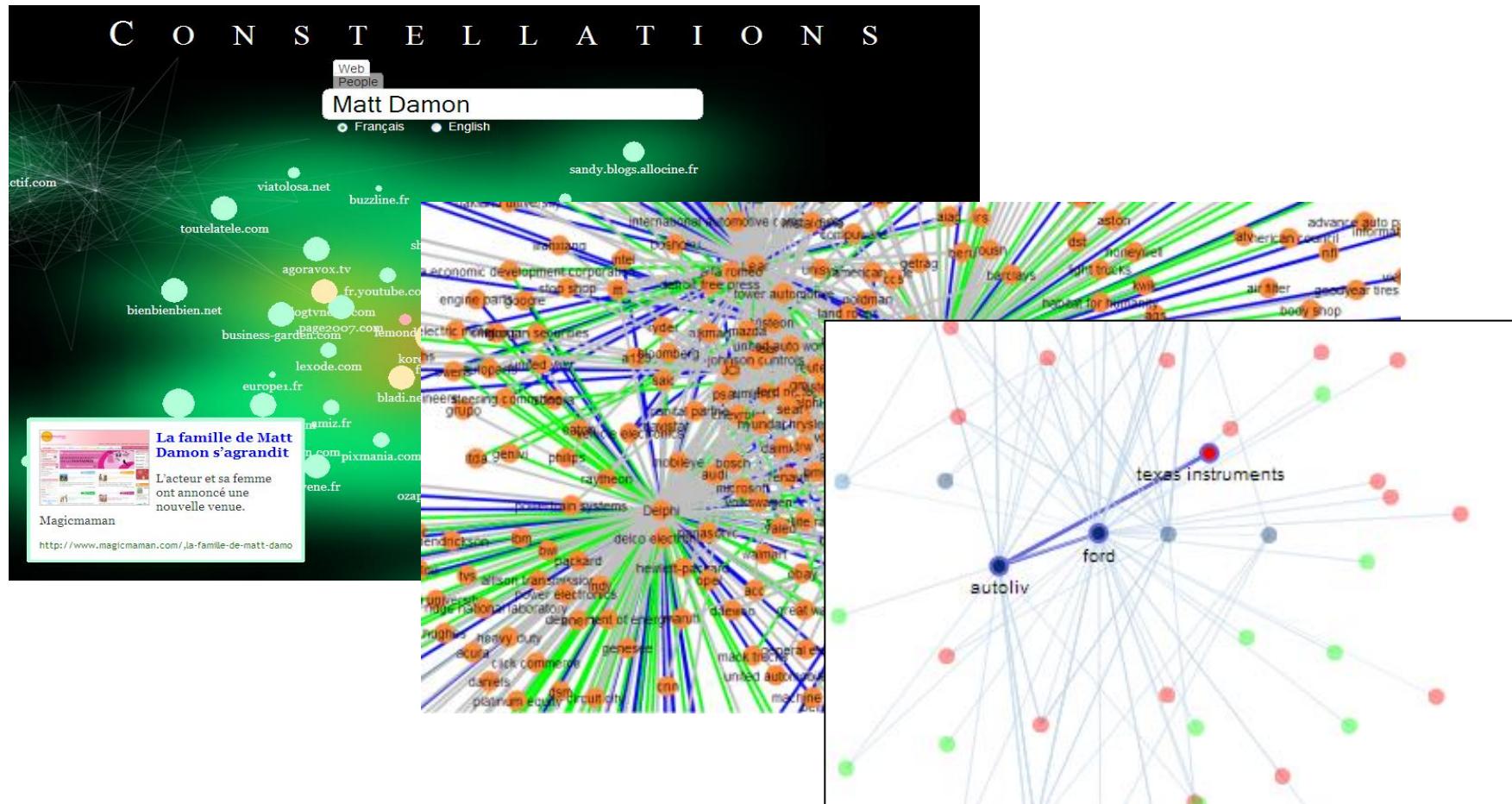
interesting overbearing unexpected pretty pleasant dished sour incredible extraordinary charmed low fixed comfortable boring suburban **OK** mixed surviving impeccable considered **quality** loving sorry unparalleled vinegary returning easy ended

**small** cleared included

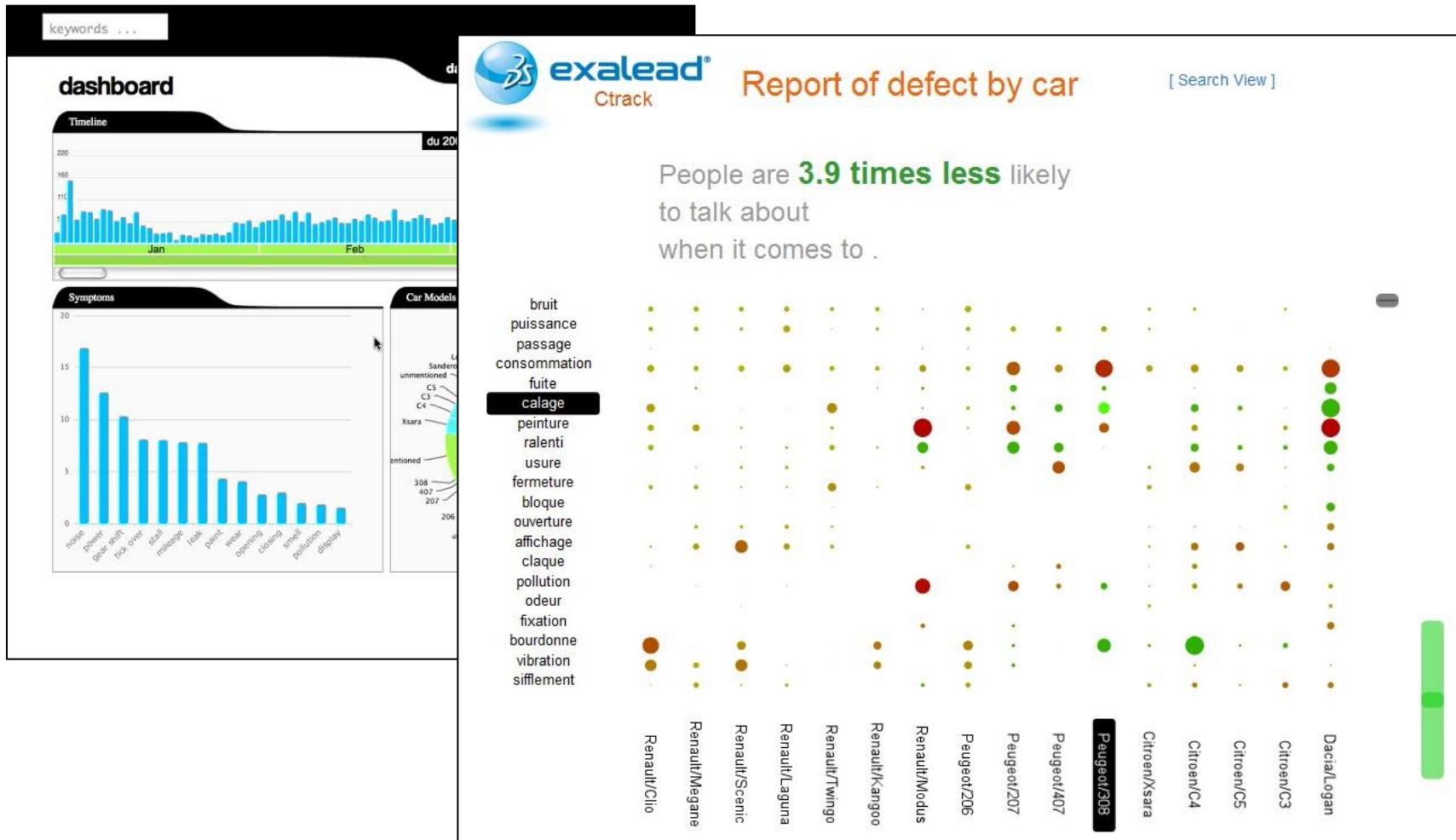
**delicious** sizable excited homey refreshing inconsistent over-the-top received velvety

Mo3

# Analyse descriptive : Construire des graphes de relations/réseaux



# Analyse descriptive : Déetecter des patterns



# Analyse prédictive : Prévisions à partir des modèles



# Analyse prescriptive : Recommendations à partir des modèles

Pertinence Suite POWERED BY VELOCITY Execution | New | Search Order 000002-001- » Oper 30 Welcome, Data Manager | Log Out AVAILABLE | Links

**Data Collection**

MIN RESIN CONTENT	36.35	31.62	37.25
AVERAGE RESIN CONTENT	36.7	33.69	37.68
MAX RESIN CONTENT	37.15	35.11	37.98
HARDENER CONTENT	22.49	21.23	22.63
RESIN VISCOSITY T0	16.824		
BONDING STRENGTH	0.81		
VOLATILITY	0.975		
RESIN VISCOSITY T2	86.784		
RESIN VISCOSITY T1	33.24		

**Information**

Risk:	Low	High
87%		

**Settings**

Name	Current Value	Chosen Range	Action
Average Resin Content	36.7	[33.69;37.68]	OK
Min Resin Content	36.35	[31.62;37.25]	OK
Max Resin Content	37.15	[35.11;37.98]	OK
Hardener Content	22.49	[21.23;22.63]	OK
Mandrel	z02		OK
Winding Pitch	5.11	[4.7;5.0762]	To be changed
Max Winding Tension	8.6	[8.5;9.1]	OK
Min Winding Tension	7.3	[7.1;7.8]	OK
Autoclave Mold Position	sup	sup, inf	OK
Mold	z02	z01, z02	OK
Curing Ramp Up Time	76	[54;110]	OK
Vacuum level 1	-300	[-300;1.4]	OK
Curing plateau duration	220	[136;290]	OK
Transition Temperature	4.64	[3.68;8.29]	OK
Curing Cycle Time	55.3;119	[55.3;119]	OK
<Product Identifier>	000002	[13;57]	To be changed
<Production Date>	04/17/08 16:27:28		To be set

**Performance Report on Production Quality**

**Faire ça**  
...dans le cadre de votre tolérance pour risque

Suivez vos résultats

# Surveillance/analyse en temps réel : Dashboards



Show tags  
Hide Time Line

Numbers of quotation



Custom period: 2009-10-23 - 2010-10-22

week of Oct 23, 2009 - week of Oct 22, 2010

Zoom: 10D 1M 3M 1Y YTD MAX

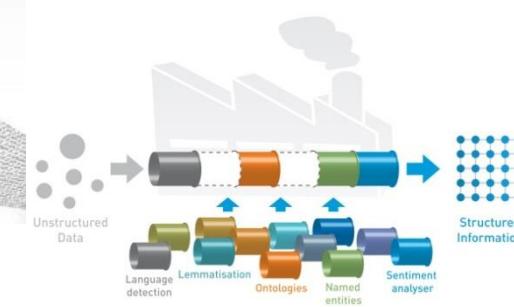




---

## 5. ON PEUT LE FAIRE COMMENT ?

# Comment ?



**Capturer**

**Transformer**

**Visualiser**



**Rechercher**



**Partager**

**Notifier**

**Explorer**

**Exporter/Se connecter**

**Analyser/Action**

# Qu'est-ce que c'est le Web ? Le Web ≠ L'Internet



Infrastructure

M2M



WWW

H2H



App Internet

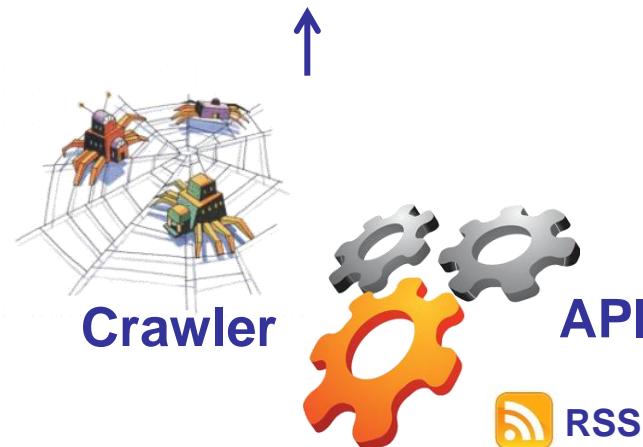
## INTERNET

# Capturer

## Infrastructure



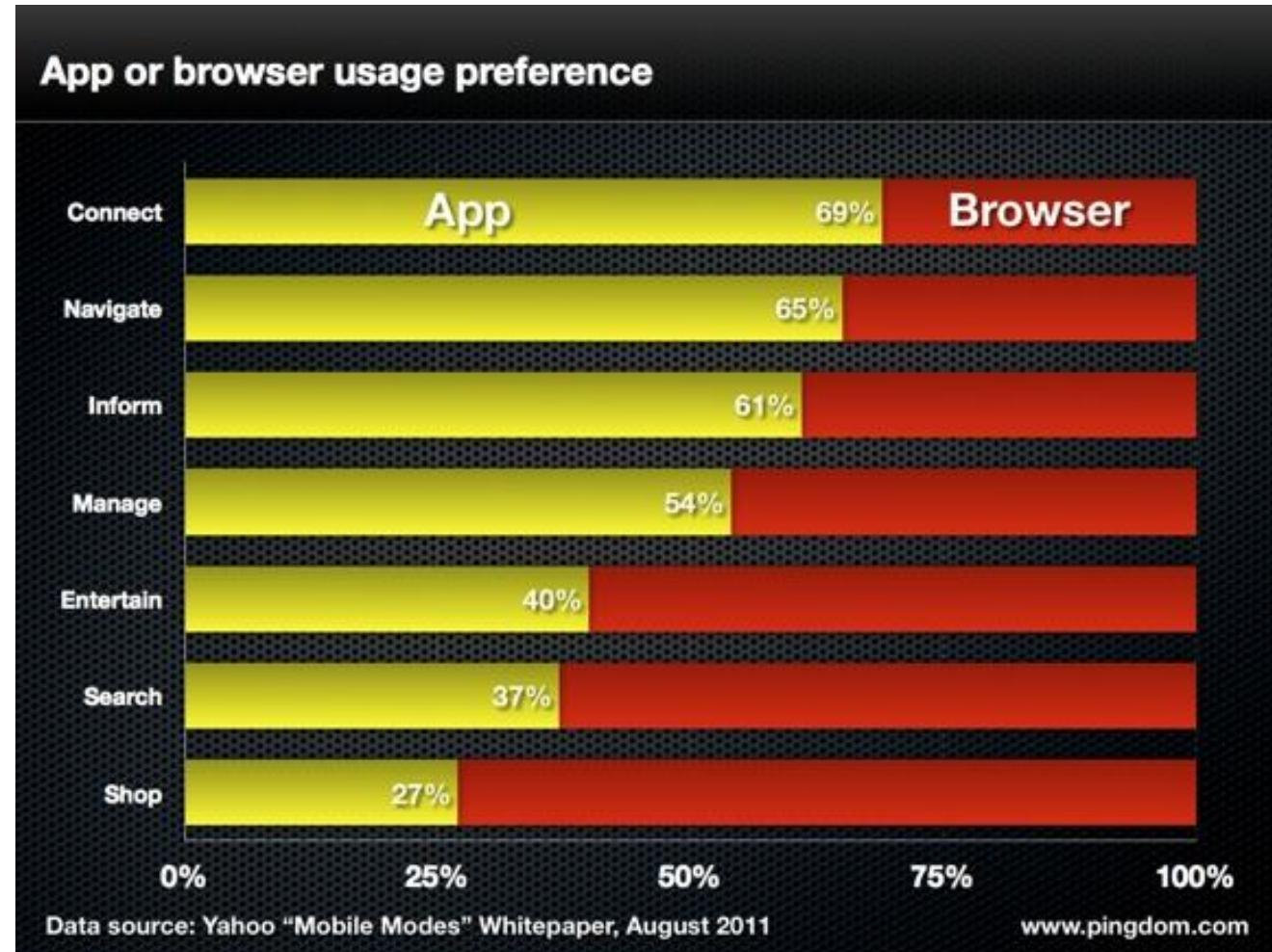
## WWW



## App Internet



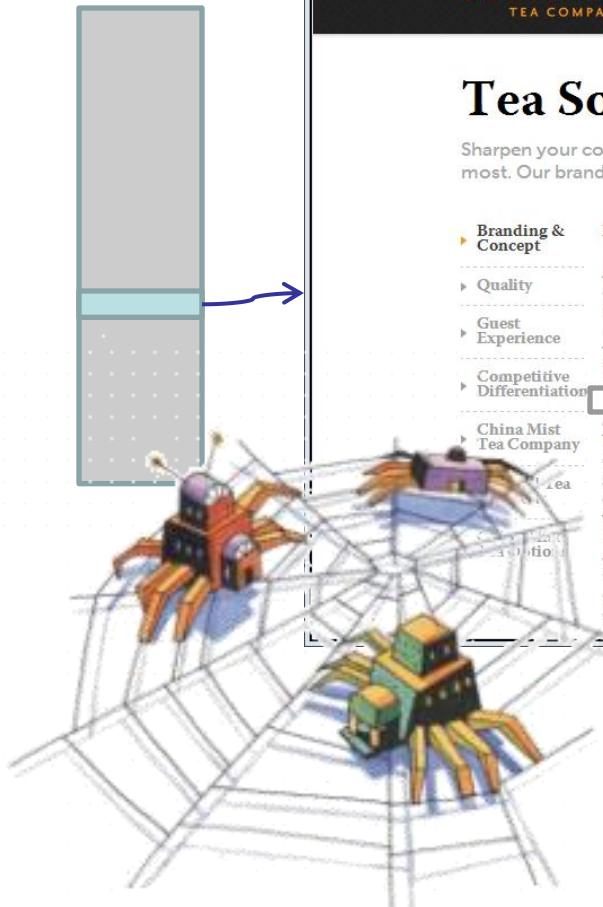
# Apps: La fin du WWW?





# Crawler

## Liste des URLs à crawler



Nouveaux liens



Nouvelle liste  
des URLs à  
crawler



## Défis : Crawler



# Défis : Crawler

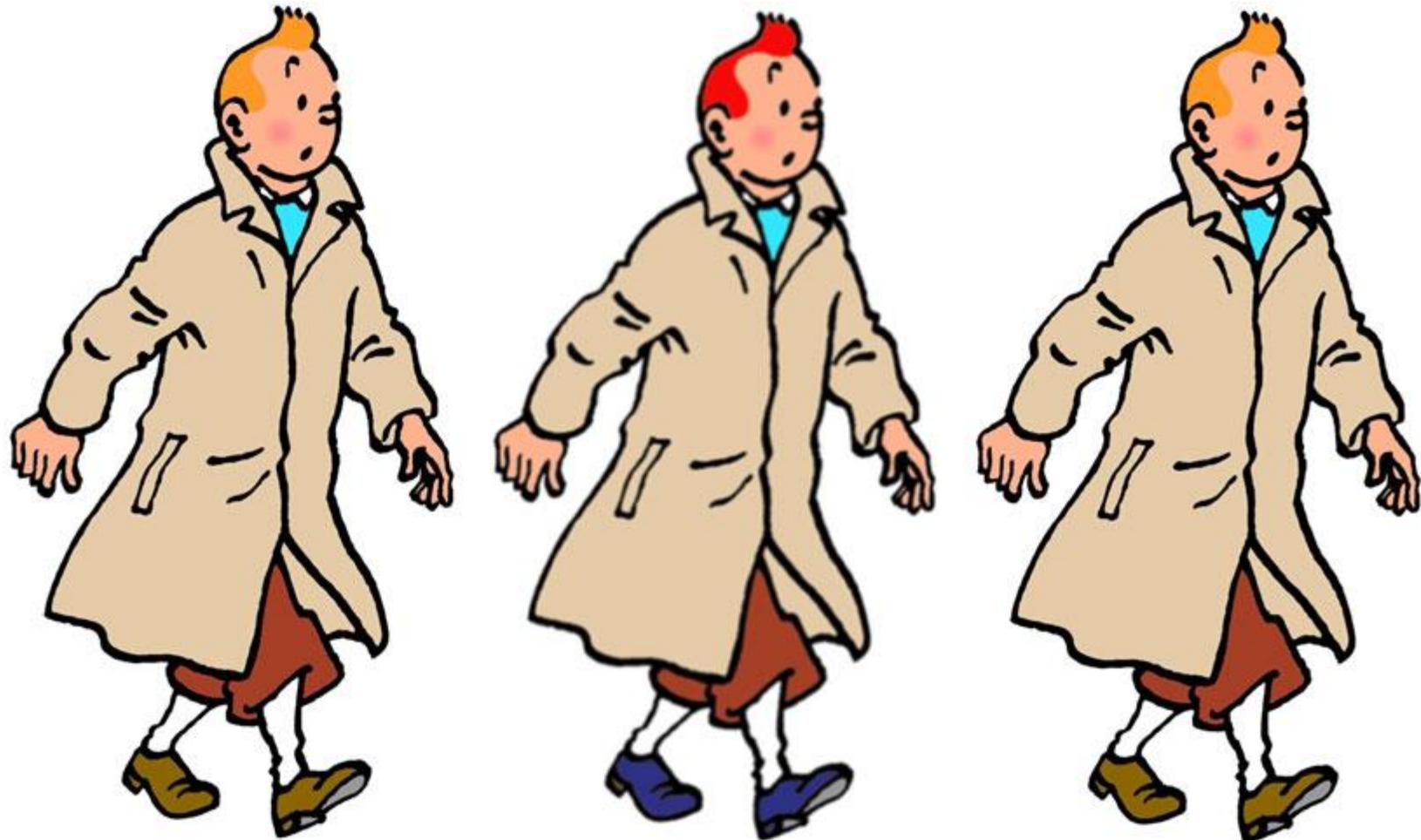


<http://www.unixuser.org/~euske/python/webstemmer/howitworks.html>

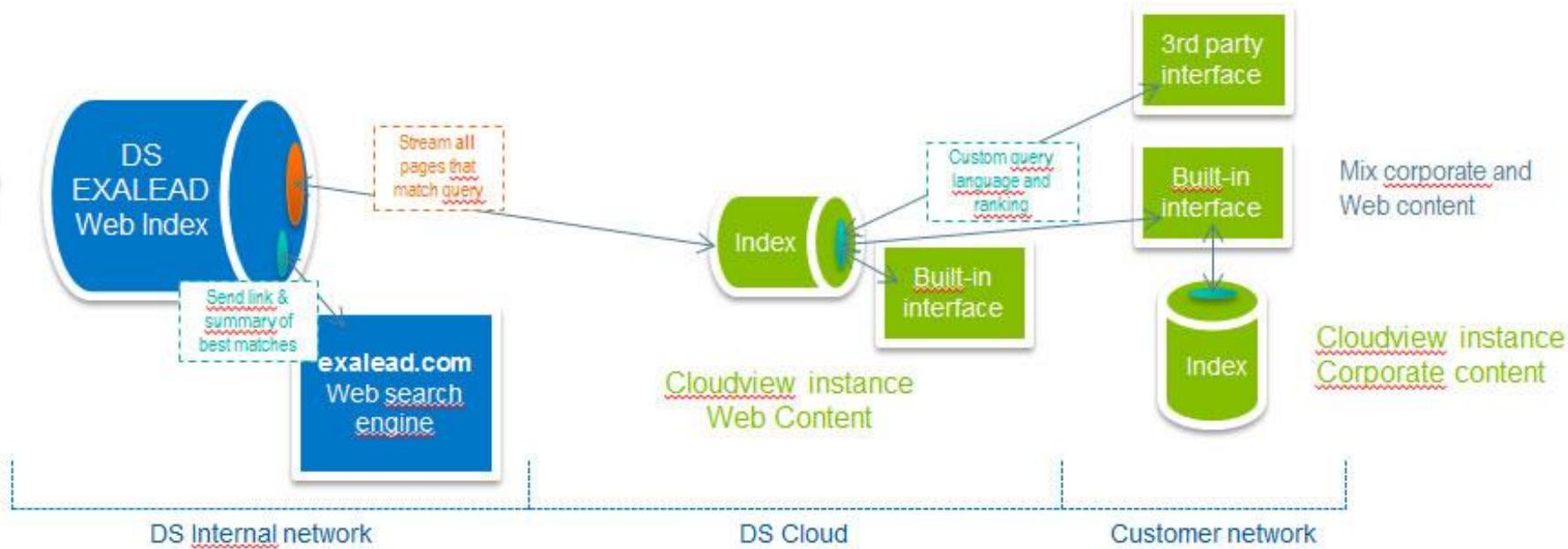
## Crawler : Défis



# Crawler : Défis Duplicata

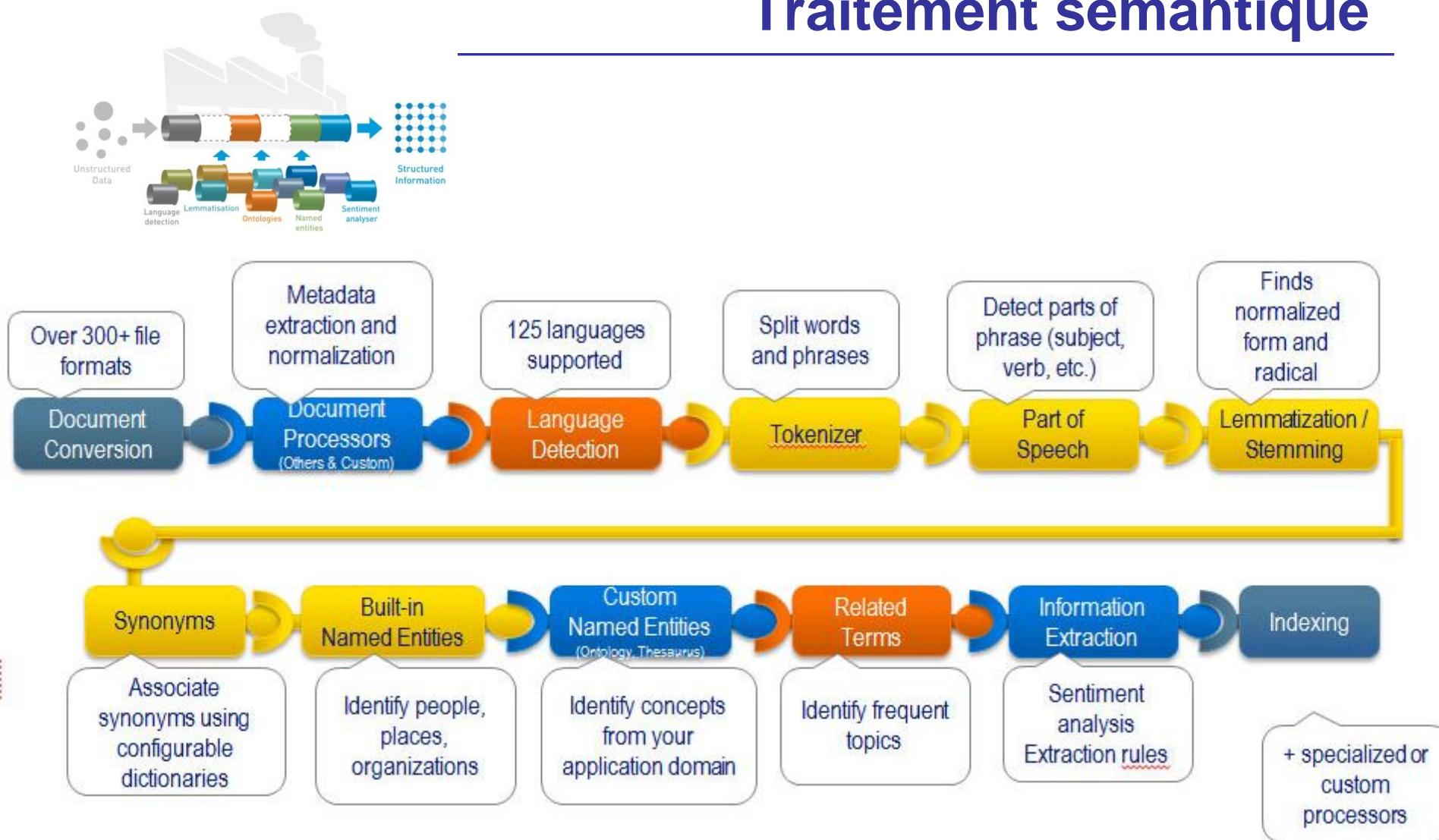


# Crawler : Altérnatif



## EXALEAD Web Mining Experience

# Transformer : Traitement sémantique



# Transformer : Traitement sémantique

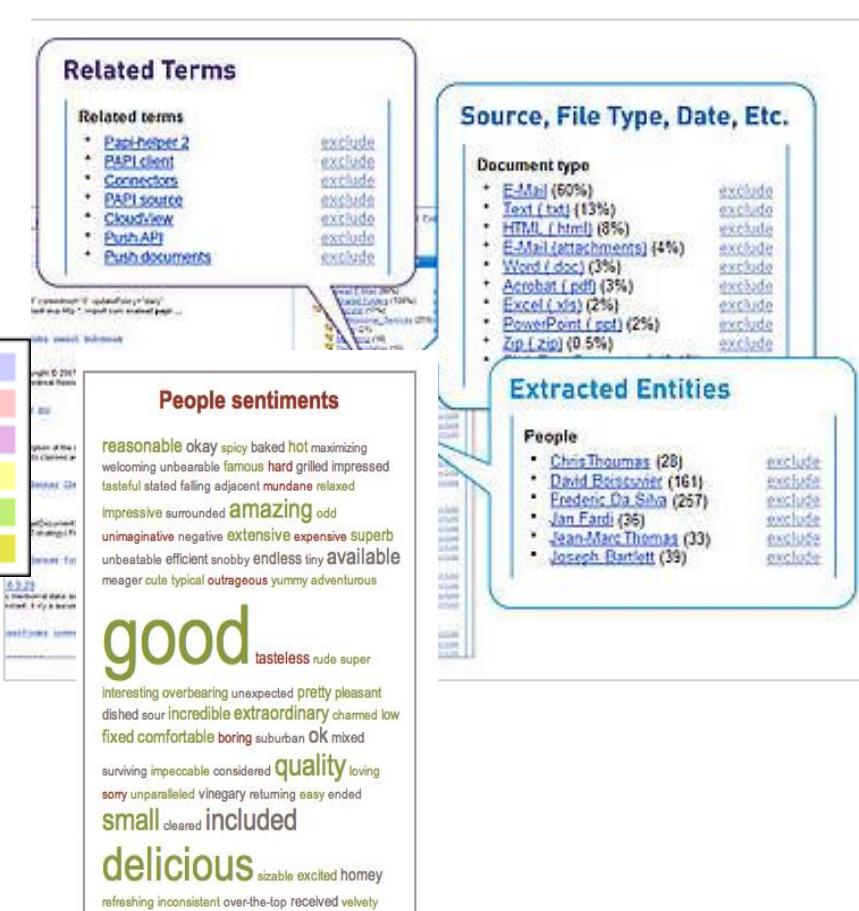
Kofi Atta Annan is a Ghanaian diplomat who served as the seventh Secretary General of the United Nations from January 1, 1997, to January 1, 2007, serving two five-year terms. Annan was the co-recipient of the Nobel Peace Prize in October 2001.

Kofi Annan was born on April 8, 1938, to Victoria and Henry Reginald Annan in Kumasi, Ghana. He is a twin, an occurrence that is regarded as special in Ghanaian culture. Efua Atta, his twin sister, shares the same middle name, which means 'twin'. As with most Akan names, his first name indicates the day of the week he was born: 'Kofi' denotes a boy born on a Friday. The name Annan can indicate that a child was the fourth in the family, but in his family it was simply a name which Annan inherited from his parents.

In 1962, Annan started working as a Budget Officer for the World Health Organization, an agency of the United Nations. From 1974 to 1976, he was the Director of Tourism in Ghana. Annan then returned to work for the United Nations as an Assistant Secretary General in three consecutive positions.



Entity extraction  
Classification automatique  
Analyse de sentiment...



# Transformer : Traitement sémantique

**MON GUIDE PERSONNALISÉ**

Je cherche un T2, T3 ou T4 sur 75006

1470 ANNONCES IMMOBILIÈRES ONT PERMIS DE GÉNÉRER LES STATISTIQUES CI-DESSOUS.

Quel loyer en moyenne par type de logement ?

Type	Moyenne
T2	12 €
T3	2594 €
T4	3061 €

Comment a évolué le nombre d'annonces ?

Combien a évolué le loyer moyen ?

CARTE VILLES VOISINES

LES VILLES VOISINES

Paris 75000

Quel est le nombre d'annonces par type

Location appartement 4 pièces 120 m² Paris

3.300 €

Paris 12e Face Gare de Lyon. Immeuble haussmannien ravalé. Luxueux 4 pièces 120 m² état neuf. Ascenseur. Grandes fenêtres. 3 balcons. Calme. Salon, salle à manger, 2 chambres, grande entrée (porte blindée), salle de... suite

3 photos



# Transformer : Transcription ‘speech-to-text’; sémantique

**Voxalead news**

« Return to results

6 days ago.



In this video :

People   Organization   Location

- AIG   Acura   Bahama   Barack Obama
- Bob Gucci Gianni   Brian Tom Costello
- Brian Williams   CNBC   California
- Chris Christie   Chris Jansing Nbc   FORD
- FOX   Fannie Mae   Fox News
- Freddie Mac   Gucci   Guinea
- Helen Thomas Jefferson   Honda   Hudson
- Hudson River   Hugh Hefner   Illinois
- Iowa   Jersey   Juan Williams
- Laura Schlessinger   Lexus   MSN
- Massachusetts   Mexico   Michael Isikoff
- Mike Huckabee   NASA   New Jersey
- New York   New York Times   News West
- Ohio   Pennsylvania   Playboy
- President Obama   Republican
- Rick Sanchez   Rome   San Diego
- Sarah Palin   Scotch Plains   Springfield
- Springfield , Massachusetts

Toyota

All   Français   English   中文   العربية   español   Русский   Select sources



On the broadcast tonight , politics and what's different this year . Hidden fortune being funneled into campaigns and is this the new face of the Republican party . We call from Toyota . What's the problem . This time it's a prominent media figure fired for what he said on the air . Is it fair punishment or political correctness gone and houses of worship . Americans finding a place to pray right in their own backyard . Also tonight something very exciting on the moon and lots of the nightly news begins now ..

More news in New York this is NBC nightly News with Brian Williams.



# Transformer : Traduction

[Exalabs](#) [Chromatik](#) [Constellations](#) [Exalead Light](#) [MiiGet](#) [Sourcier](#) [Tweepz](#) [Voxalead](#) [Wikifier](#) [Feedback](#)

The screenshot shows a search results page for 'afghanistan'. A red circle highlights the search term 'afghanistan' in the search bar. Another red circle highlights the date 'December 2 2009' in the third search result, which is also circled by a red arrow pointing from the search bar.

Voxalead news

afghanistan

All Français English 中文 العربية Select sources

Show tags Show Time Line

October 12 2009: ...参议院 军事 委员会 主席 莱 文 说 在 阿富汗 制定 更 好 的 战略 以 增加...

August 25 2009: ...一些 宇宙 阿富汗 总统 大选 后 的 情势 受到 各 方 高度 关切...

December 2 2009: ...美国总统 奥 巴 马 公布 了 对 阿富汗 新政 策 为 明年 夏天 之前 美国...

December 1 2009: ...好的 德 林 森 美国 总统 奥 巴 马 和 阿富汗 总统 卡 尔 扎 伊 进行 了...

December 1 2009: ...计划 于 日本 时间 明天 宣布 新的 阿富汗 战略 在此 之前 他 一直 是...

November 30 2009: ...如果 是 有 好 玩 得 更好 他 们 有 什 样 都 阿富汗 恐怖 活动 他们 只...

NHKG WORLD CH

RFA

VOA

NHKG WORLD CH

NHKG WORLD CH

VOA



# Transformer : Analyse multimédia et sémantique

MUMA themusicmashup

By Keywords

By Chords

Play real

By Genre

- Pop
- Rock
- Electro
- Metal
- Folk
- Jazz
- Ambient
- Rap/Hip-hop
- Reggae/R&B
- Disco/Funk
- Instrumental
- Hardcore

By Moods

- Sad
- Romantic
- Calm
- Happy
- Dynamic
- Bizarre
- Festive
- Contrasted
- Enchanting
- Grandiloquent
- Playful
- Sophisticated

By Date

1950 1970 1990 2010

Reset search

exalead

ircam

Centre Pompidou

TELECOM ParisTech

### Artists

Depeche Mode Tangerine Dream Armin van Buuren

Random artist

Be the first of your friends to like this.

**Metallica**

Metallica is an American metal band from Los Angeles, California. It was formed in 1981 when drummer Lars Ulrich advertisement in The Recycler. The band originally consisted of Ulrich, rhythm guitarist James Hetfield, and lead vocalist Robert Trujillo. Trujillo was later replaced by Robert Trujillo, who went on to form the band Megadeth; ex-guitarist Kirk Hammett took his place.

3 Crystallized (Rory Phillips remix)  
4 Crystallized (The Neon Lights remix)  
1 Believer (album version)  
2 Believer (Jonas Vroom remix)  
Believer (Vince Clarke remix)  
Believer (Subway remix)  
Believer (Davide Rossi Reinterpretation)  
Echoes  
The Same Space  
3 Surfing the Void  
4 Valley of the Calm Trees  
5 Venusia  
6 Extra Astronomical  
7 Twin Flames  
Flashover  
9 Future Memories  
10 Cypherspeed  
11 Fortune

### Concerts

Search a place

lost.fm

Map data ©2011 Google Terms of Use

430 Tweets

Beyoncé Adele Kanye West Lady Gaga

Rush Nicki Minaj Katy Perry Lil Wayne

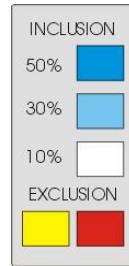
Jada PJ Harvey Peter Hammill Eric Burdon

Jay-Z Michael Jackson Miley Cyrus Eminem

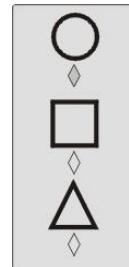
Surfing the Void Klaxons

# Images

COULEUR



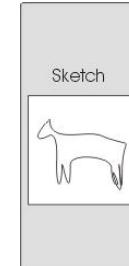
FORME



TEXTURE



DESSIN

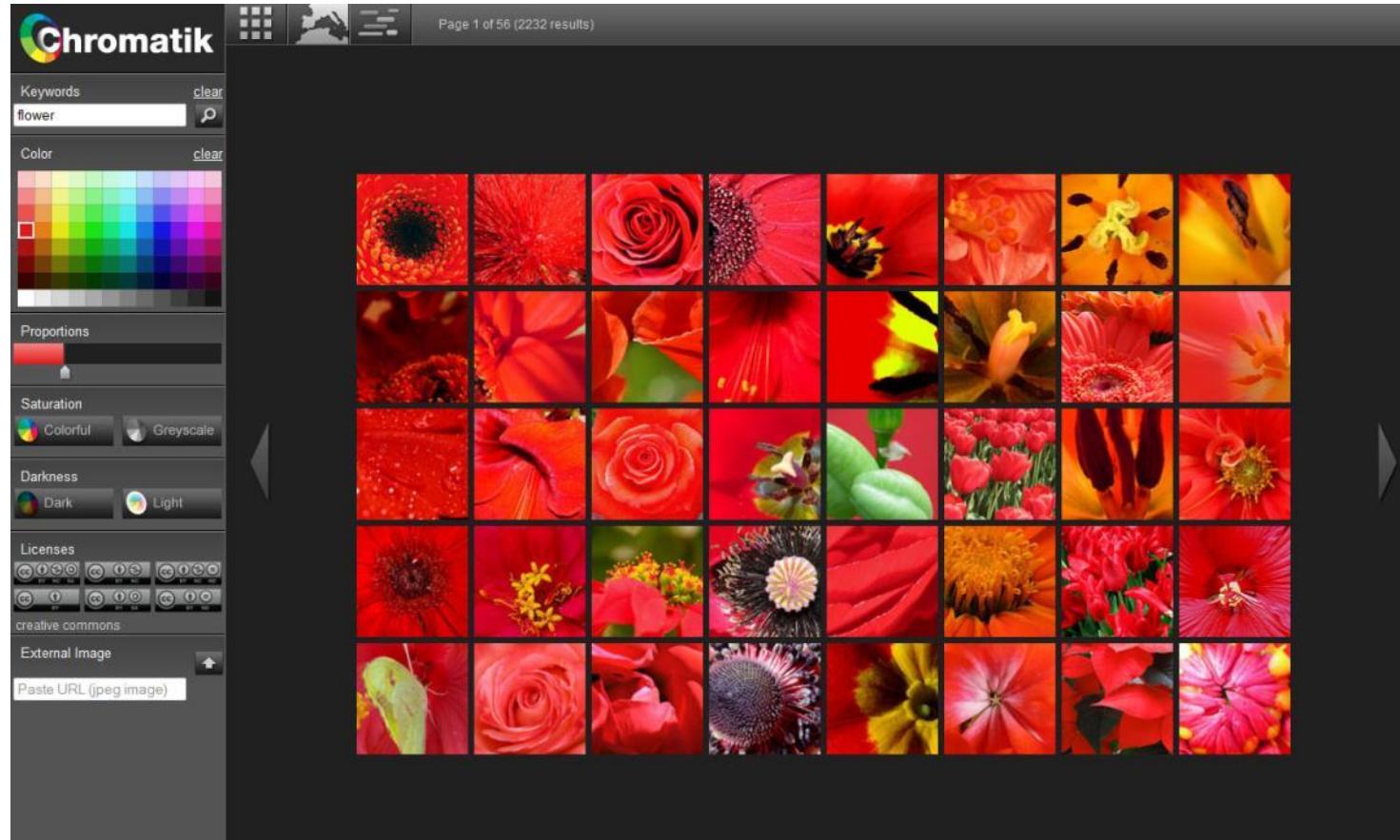


EXEMPLAIRE



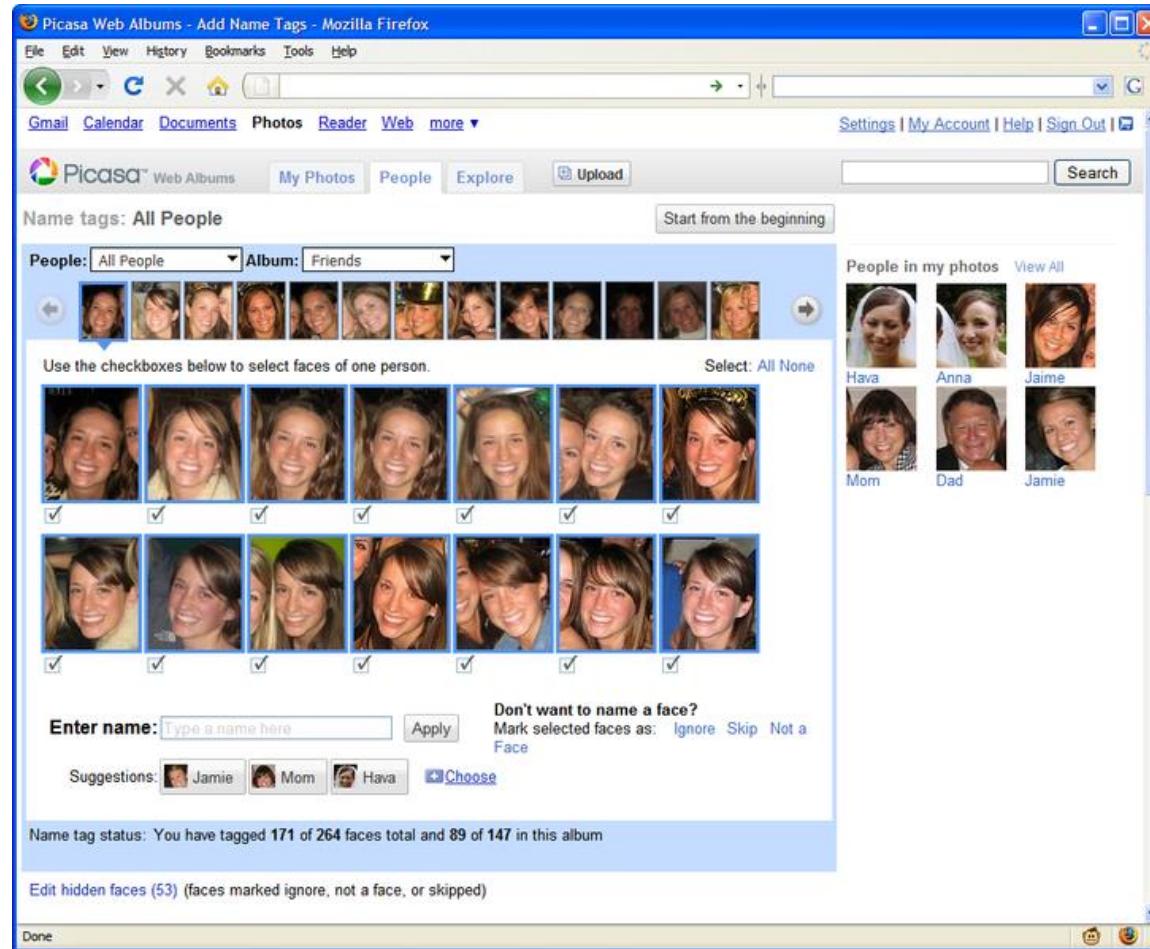
**DES TYPES PLUS COMPLEXES  
EXISTENT MAIS ILS SONT DES  
PLUS FONDAMENTAUX & LES  
PLUS SOUVENT UTILSES**

# Transformer : Analyse multimédia et sémantique



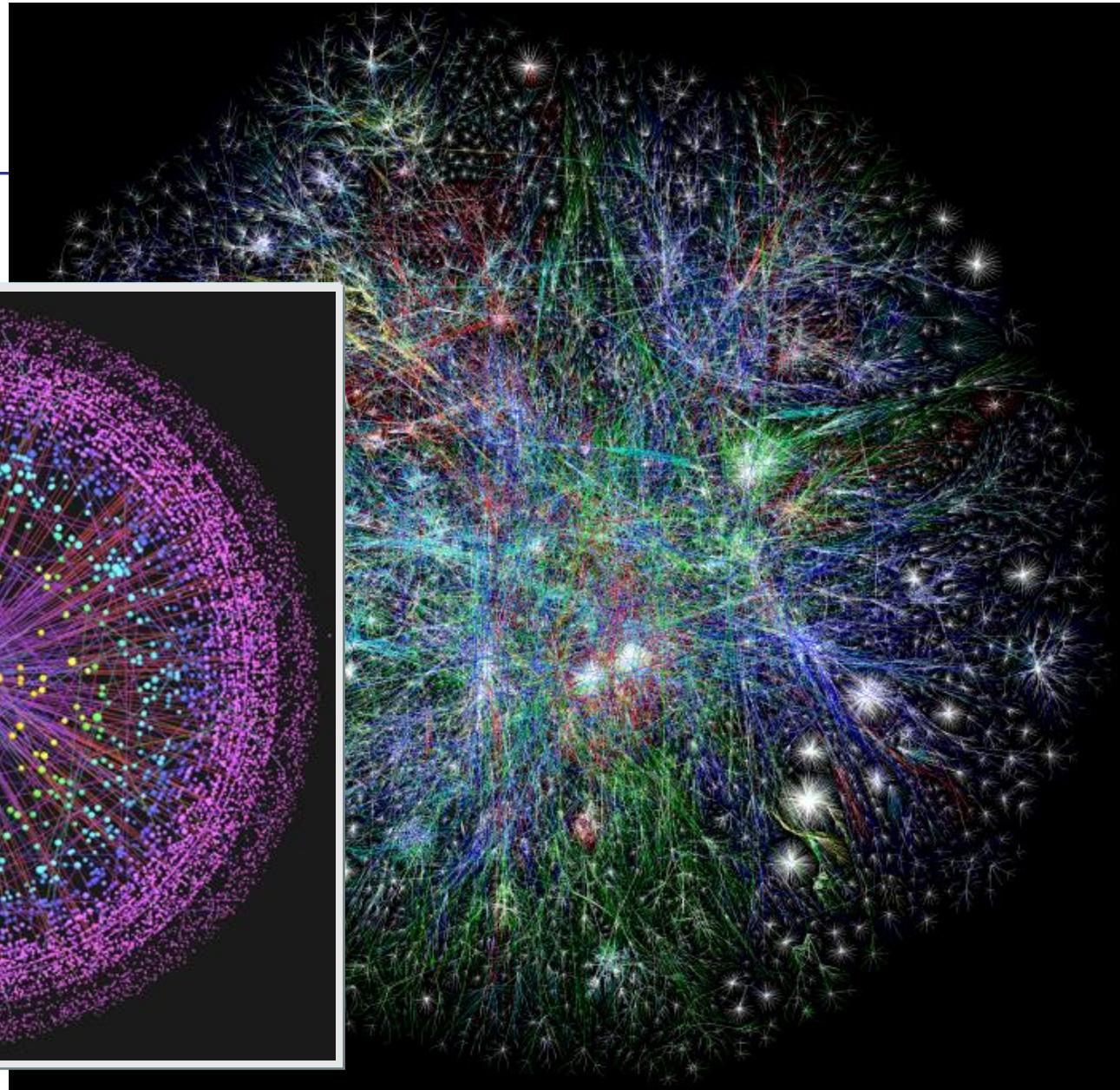


# Transformer : Analyse multimédia et sémantique



# Transformer : Analyse à soutenir la visualisation





Mo3

---

Merci !

[laura.wilber@3ds.com](mailto:laura.wilber@3ds.com)