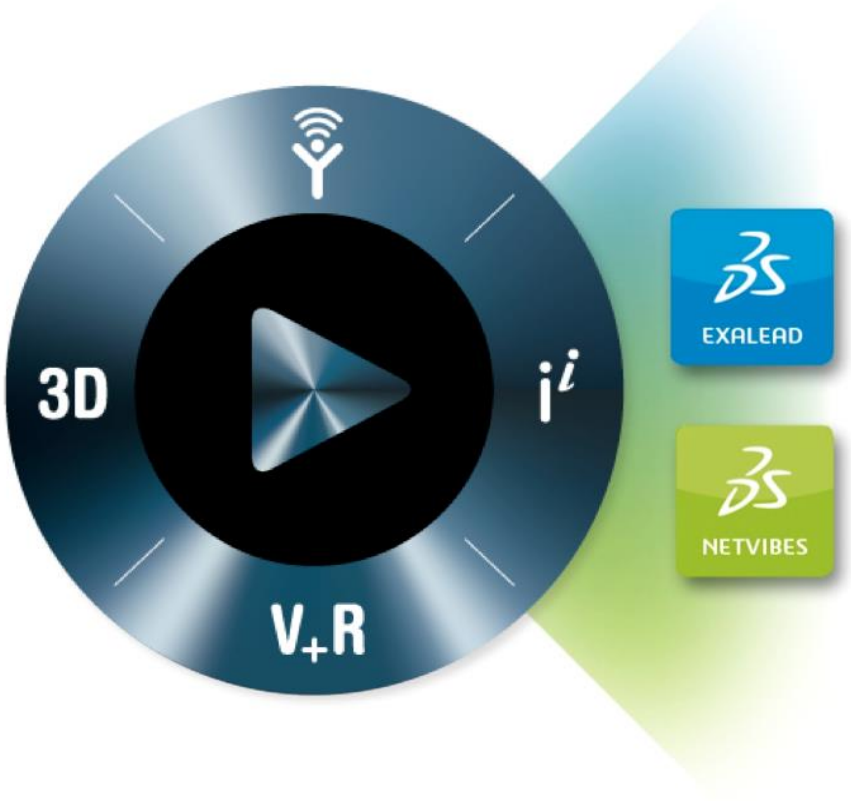
Les Entretiens de Toulouse
Rencontres Aérospatiales

Mo3: Big Data, Web & (Cyber)security

Laura WILBER
Director of Strategy, Dassault Systèmes EXALEAD

23/04/2013

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, « Securité » (« 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 !

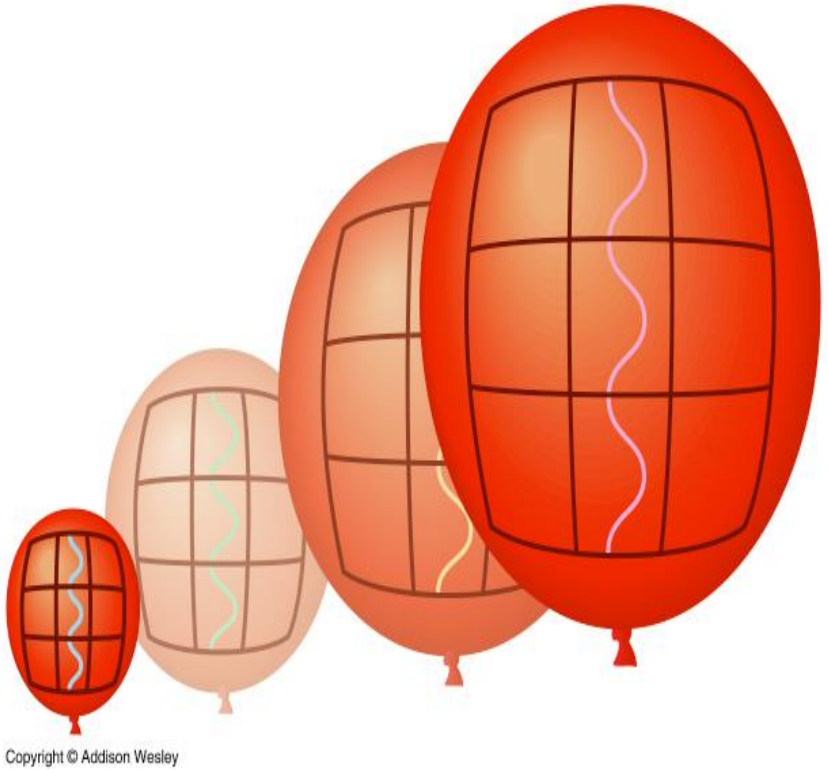
- Grid computing
- Super computers
- High-end datawarehouses

Quoi de nouveau ?

- Taux de croissance

1 Po/15 sec.

1 Po/sec

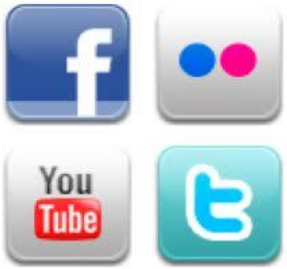


+40% an

2015: 8 Zo 8 *trillion* Go

Quoi de nouveau ?

- Types de données



Quoi de nouveau ?

- Vitesse (et parcours – P2P)

Quoi de nouveau ?

- Nouvelle technologies




Quoi de nouveau ?

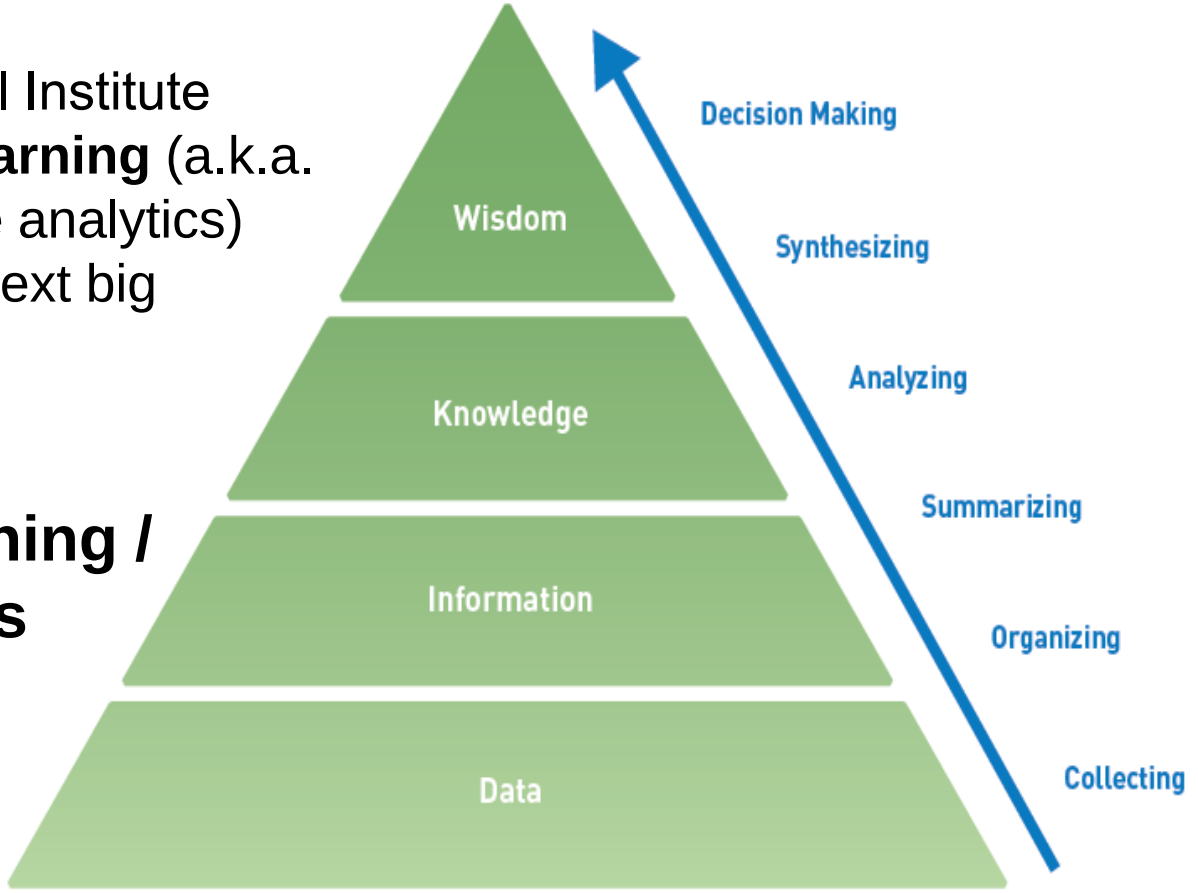
- « Nouvelle » techniques

Quoi de nouveau ?

- Trouver le trésor caché pour l'avantage concurrentielle

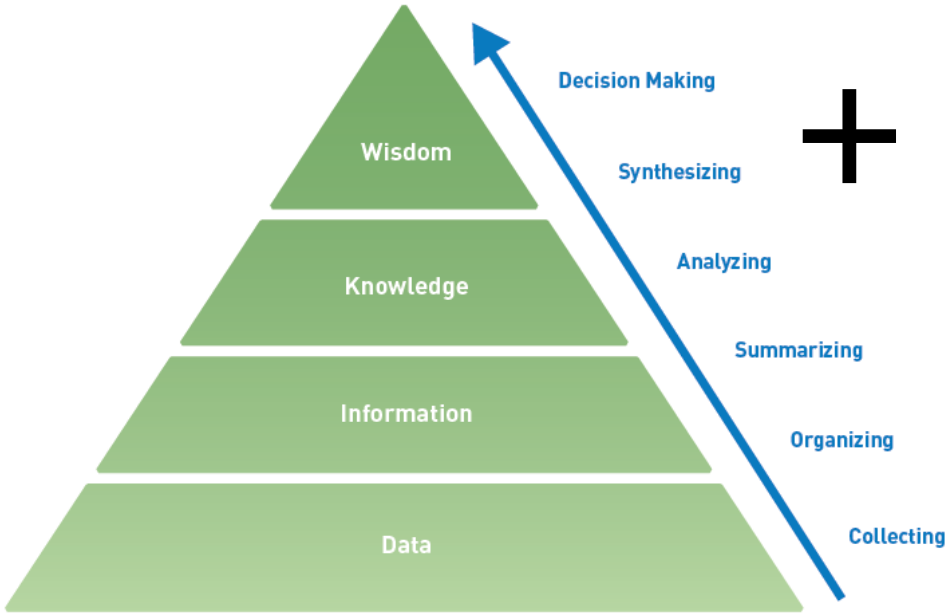
« …the McKinsey Global Institute asserts that **machine learning** (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of innovation »

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:

- Détecter
- 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

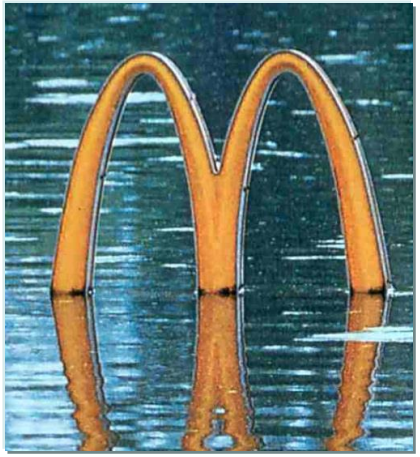
Menaces selon US Army

- 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 »

Traditionelle

- 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
2	Japan	4,395,083.00
3	Germany	2,081,202.00
4	France	
5	United Kingdom	
6	Italy	
7	China	
8	Brazil	
9	Canada	
10	Spain	
11	Mexico	
12	India	
13	Korea, Rep.	
14	Australia	
15	Netherlands	
16	Russian Federation	
17	Argentina	281,942.00
18	Switzerland	260,299.00
19	Belgium	245,706.00
20	Sweden	226,388.00
21	Austria	208,949.00
22	Turkey	188,374.00
23	*General Motors*	176,558.00
24	Denmark	174,363.00
25	*Wal-Mart*	166,809.00
39	*Toyota Motor*	115,670.90
40	*General Electric*	111,630.00
41	*Itochu*	109,068.90
42	Portugal	107,716.00
43	*Royal Dutch/Shell*	105,366.00
44	Venezuela	103,918.00
45	Iran, Islamic rep.	101,073.00
46	Israel	99,068.00
47	*Sumitomo*	95,701.60
48	*Nippon Tel & Tel*	93,591.70
49	Egypt, Arab Republic	92,413.00
50	*Marubeni*	91,807.40
51	Colombia	88,596.00
52	*AXA*	87,645.70
53	*IBM*	87,548.00
54	Singapore	84,945.00
		84,861.00
		83,556.00
		82,005.00
		80,072.70
	ce	78,515.10
		75,350.00
		75,337.00
		74,634.00
		74,178.20
		71,858.50
		71,092.00
	nd.	65,555.60
		65,393.20
68	*ING Group*	62,492.40
78	*Honda Motor*	54,773.50
79	*Assicurazioni Generali*	53,723.20
80	*Nissan Motor*	53,679.90
81	New Zealand	53,622.00
82	*E.On*	52,227.70
83	*Toshiba*	51,634.90
84	*Bank of America*	51,392.00
85	*Fiat*	51,331.70
86	*Nestle*	49,694.10
87	*SBC Communications*	49,489.00
88	*Credit Suisse*	49,362.00
89	Hungary	48,355.00
90	*Hewlett-Packard*	48,253.00
91	*Fujitsu*	47,195.90
92	Algeria	47,015.00
93	*Metro*	46,663.60
94	*Sumitomo Life Insur.*	46,445.10
95	Bangladesh	45,779.00
96	*Tokyo Electric Power*	45,727.70
		45,351.60
		44,990.30
		44,828.00
		44,637.20

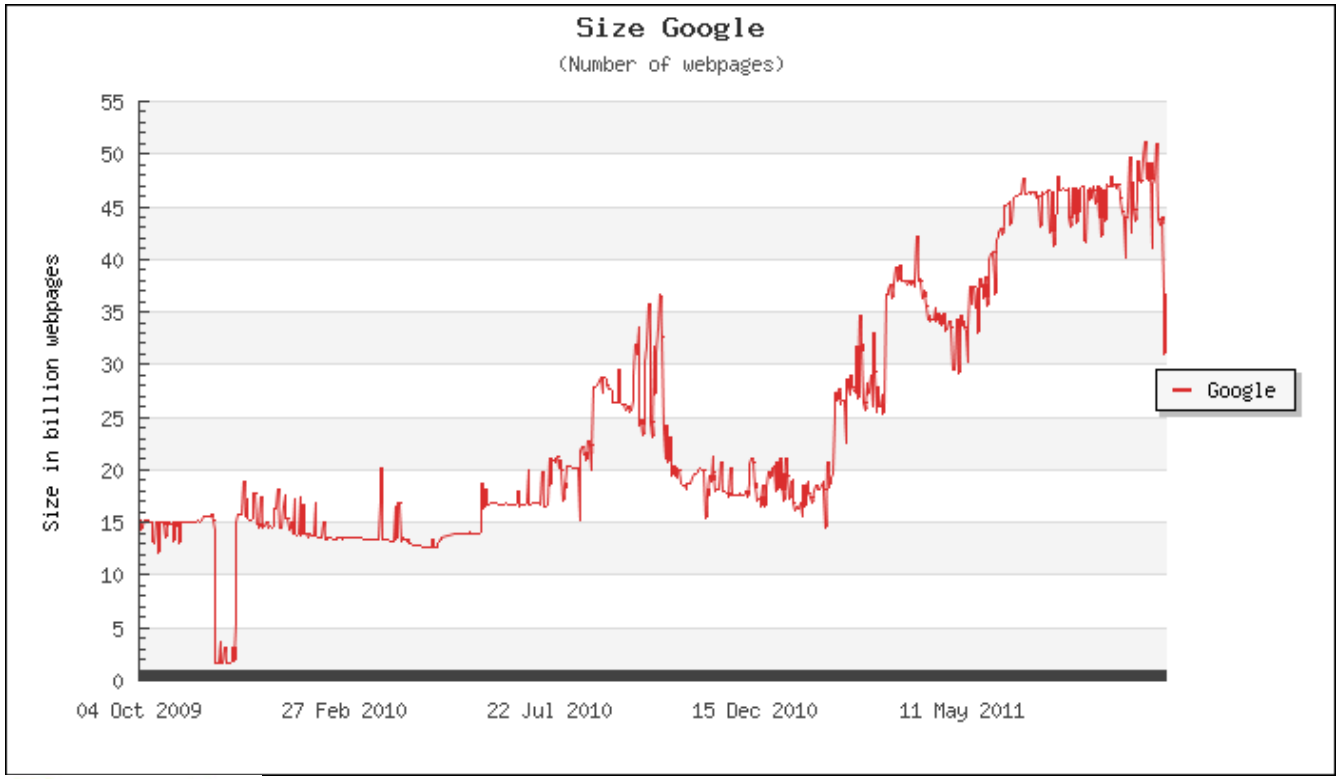
100 plus grandes entités économiques:
51 entreprises, 49 pays;
risques sans frontières

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'

> 1 billion users

> 300PB; +> 500TB/day

> 35% of world's photographs

Vraiment 'Big'

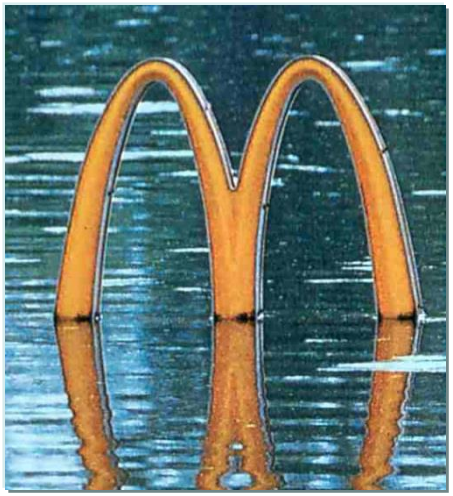
> 1000PB
+>72 hours/minute
>37 million hours/year
> 4 billion views/day

Vraiment ‘Big’

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

4. ON PEUT FAIRE QUOI AVEC TOUT CA?

Qu'est-ce qu'on peut faire?

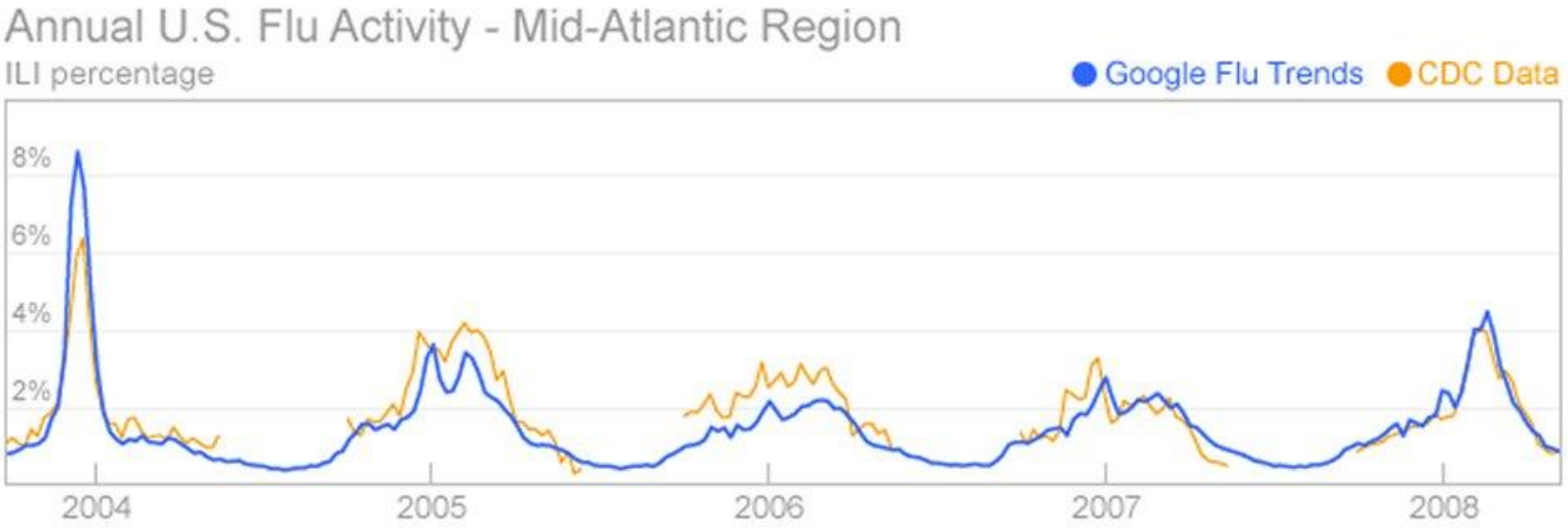


(OSINT – Biz et Govt)

Qu'est-ce qu'on peut faire?

EXALEAD

Nouveaux source & type de données, ‘nouvelles’ méthodes d’analyse



- 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	Analyse prédictive	Analyse prescriptive
Creuser des données à découvrir des faits, des tendances, des groupes, des patterns	Construire des modèles à prévoir qu’est-ce que se passera dans l’avenir	Construire des modèles pour faire des prescriptions/ recommendations
		
Actuelles et historique	Historique	Historique

EXALEAD

Analyse descriptive : Construire des profiles riches

BEFORE

AFTER

Restminer

Search

Advanced Search

Becco

355 W 46th St, New York, NY 10036, USA
Phone: (212) 397-7597
Subway: C, E at 50th St.; A, C, E at 42nd St.-Port Authority Bus Terminal
Website: http://www.becconyc.com - **Book Online !**
Kind of food: Italian

Comments | Menu

d her son joe's theater district mainstay. best known for the $21.95 pasta tasting menu, and the great 25 italian wine list.for 15 years becco has been serving consistently delicious regional italian cuisine. 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 pedastle 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)*

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

Details

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

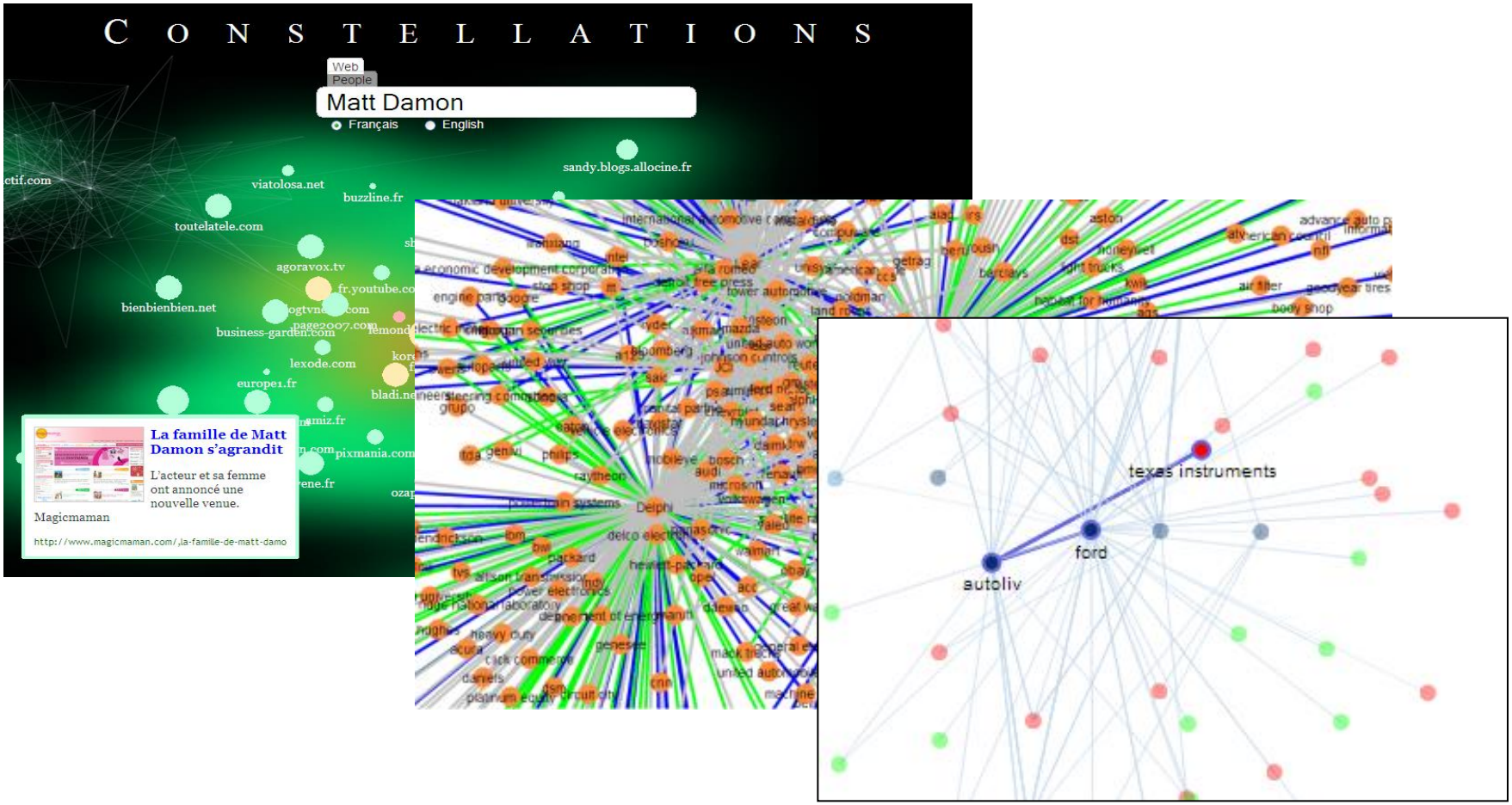
Chef: William Gallagher

Price range: Moderate, Expensive

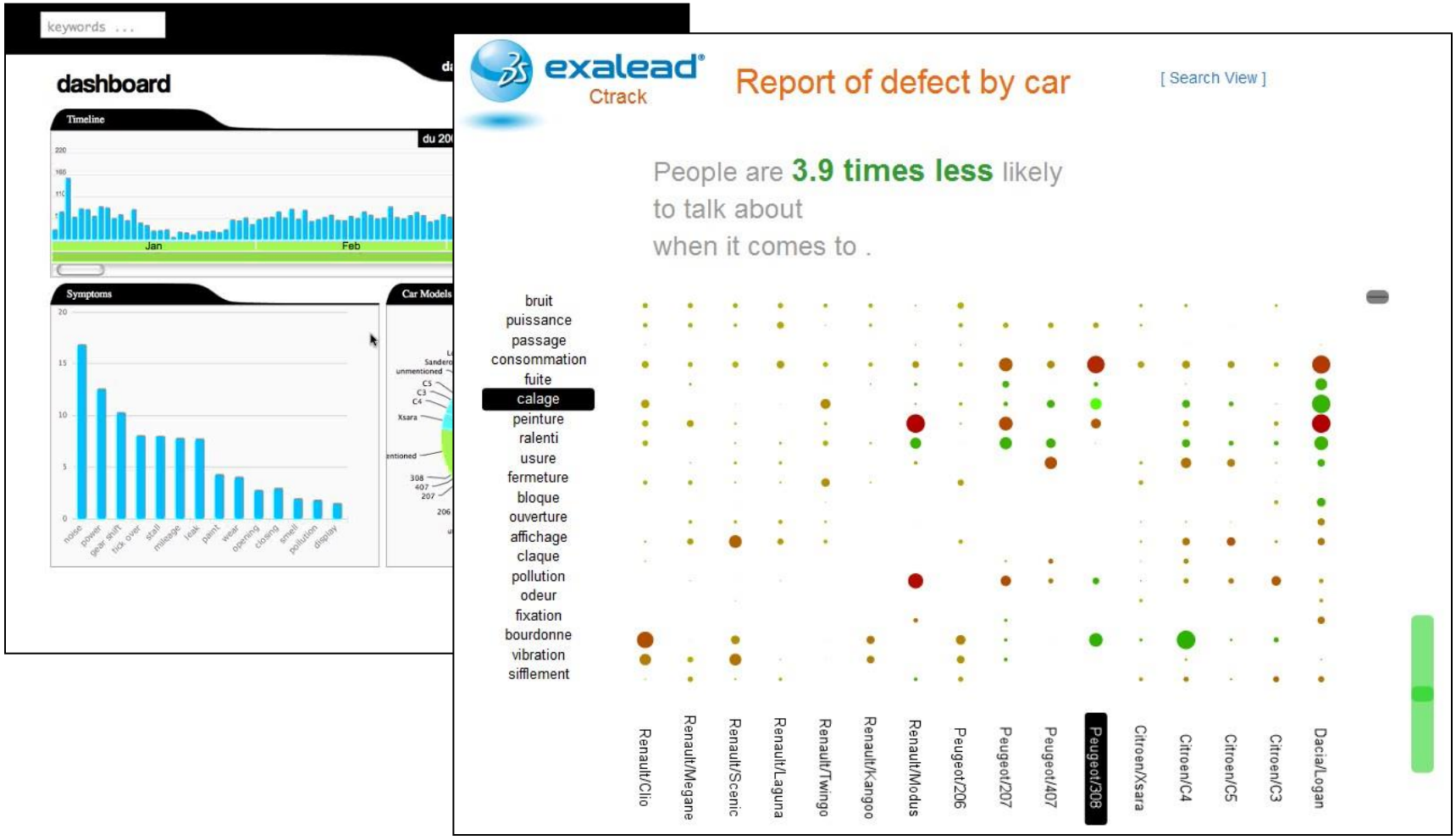
People sentiments

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

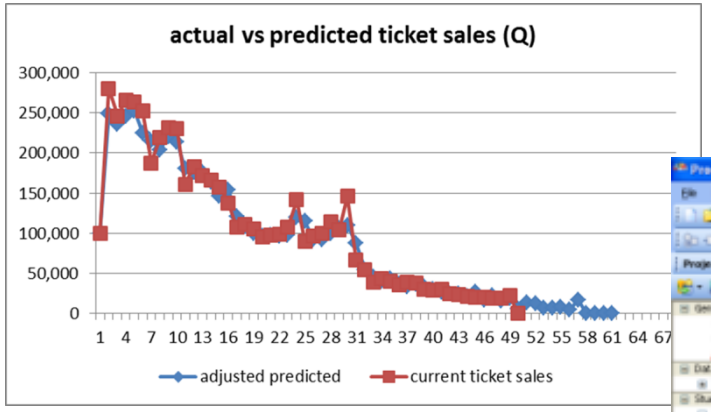
Analyse descriptive : Construire des graphes de rélations/réseaux

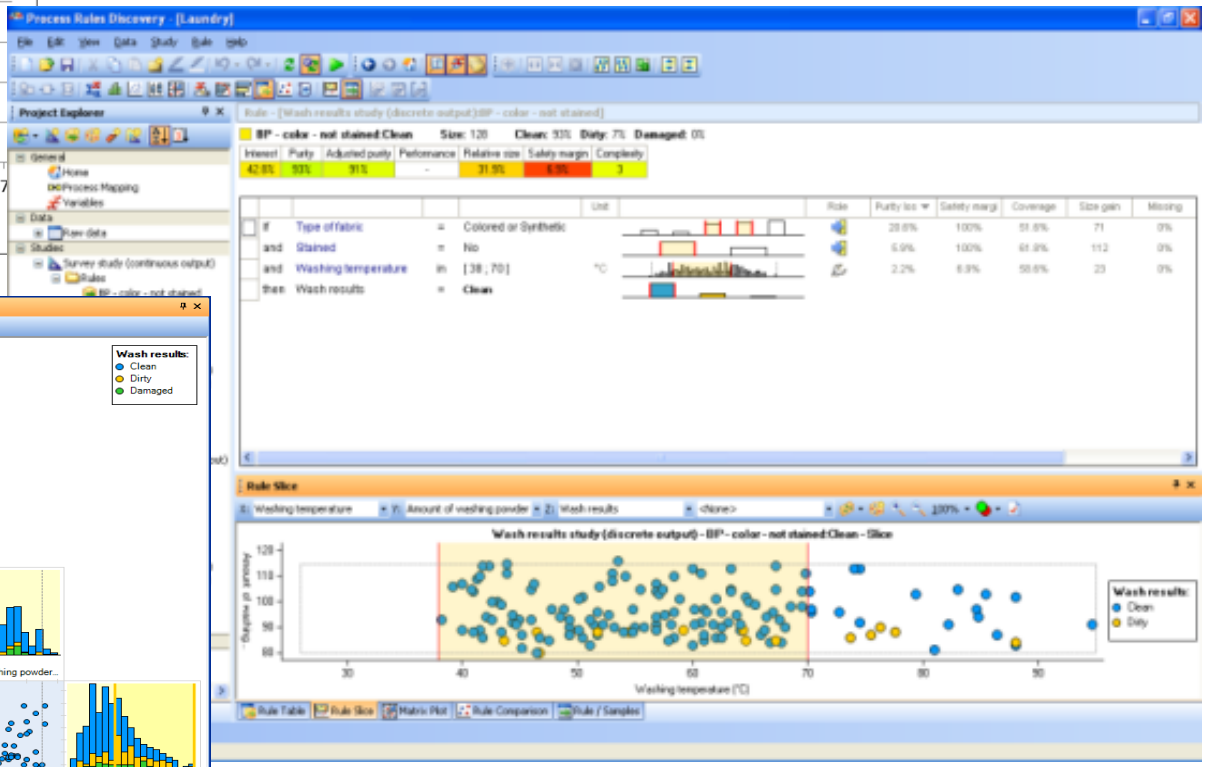


Analyse descriptive : Détecter des patterns

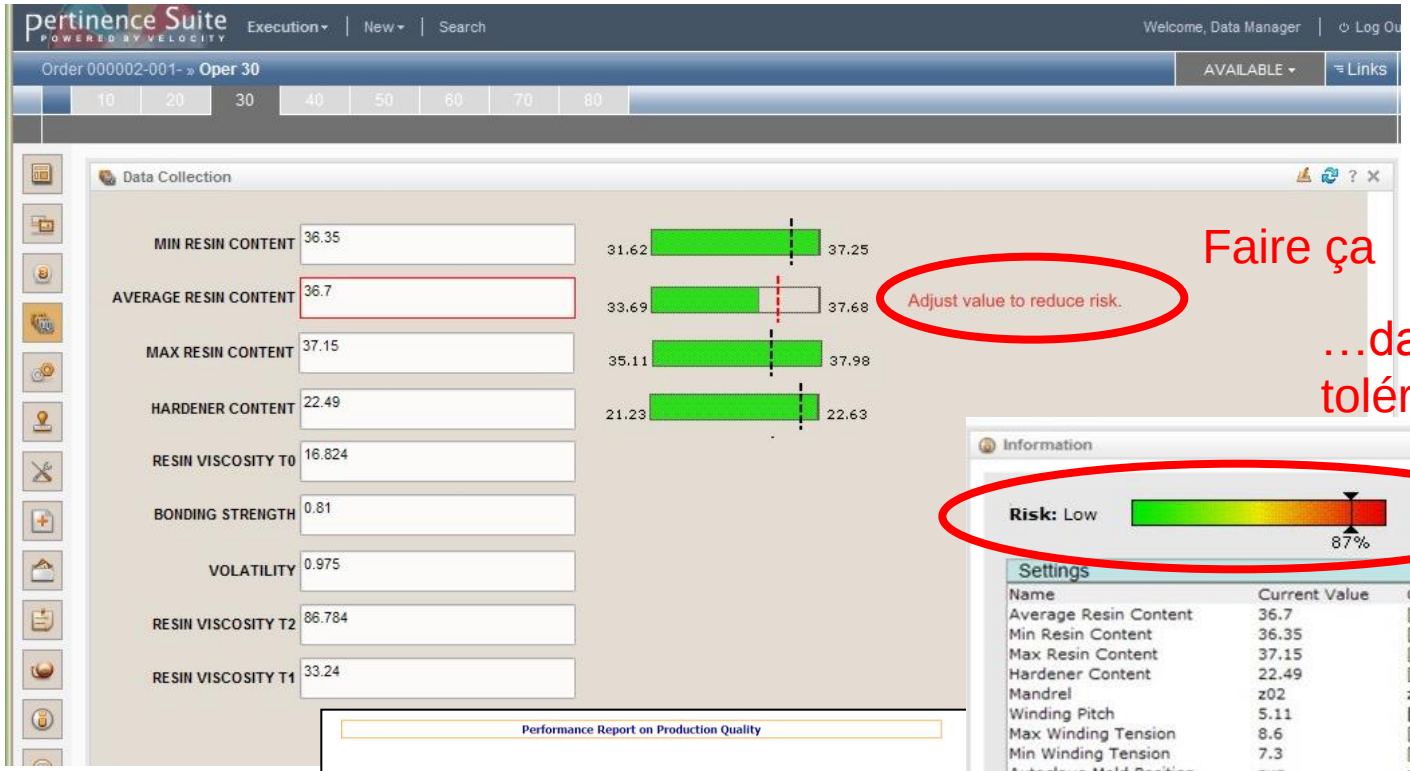


Analyse prédictive : Prévisions à partir des modèles



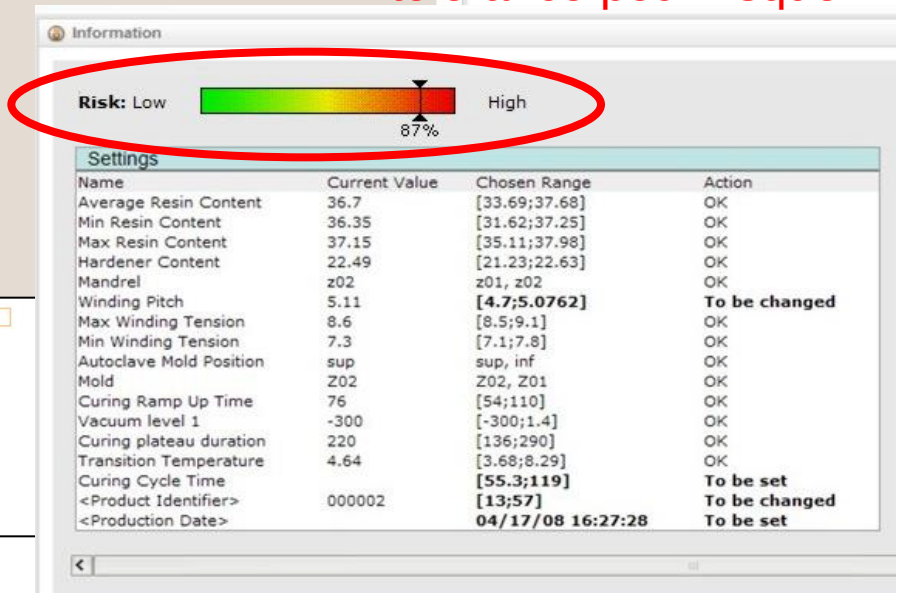


Analyse prescriptive : Recommendations à partir des modèles

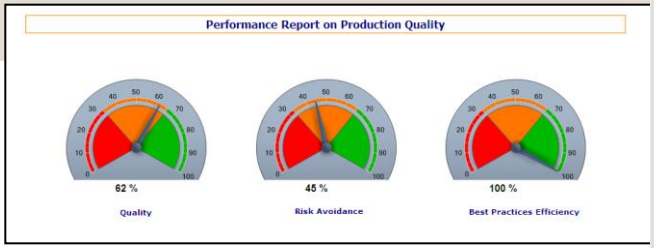


Faire ça

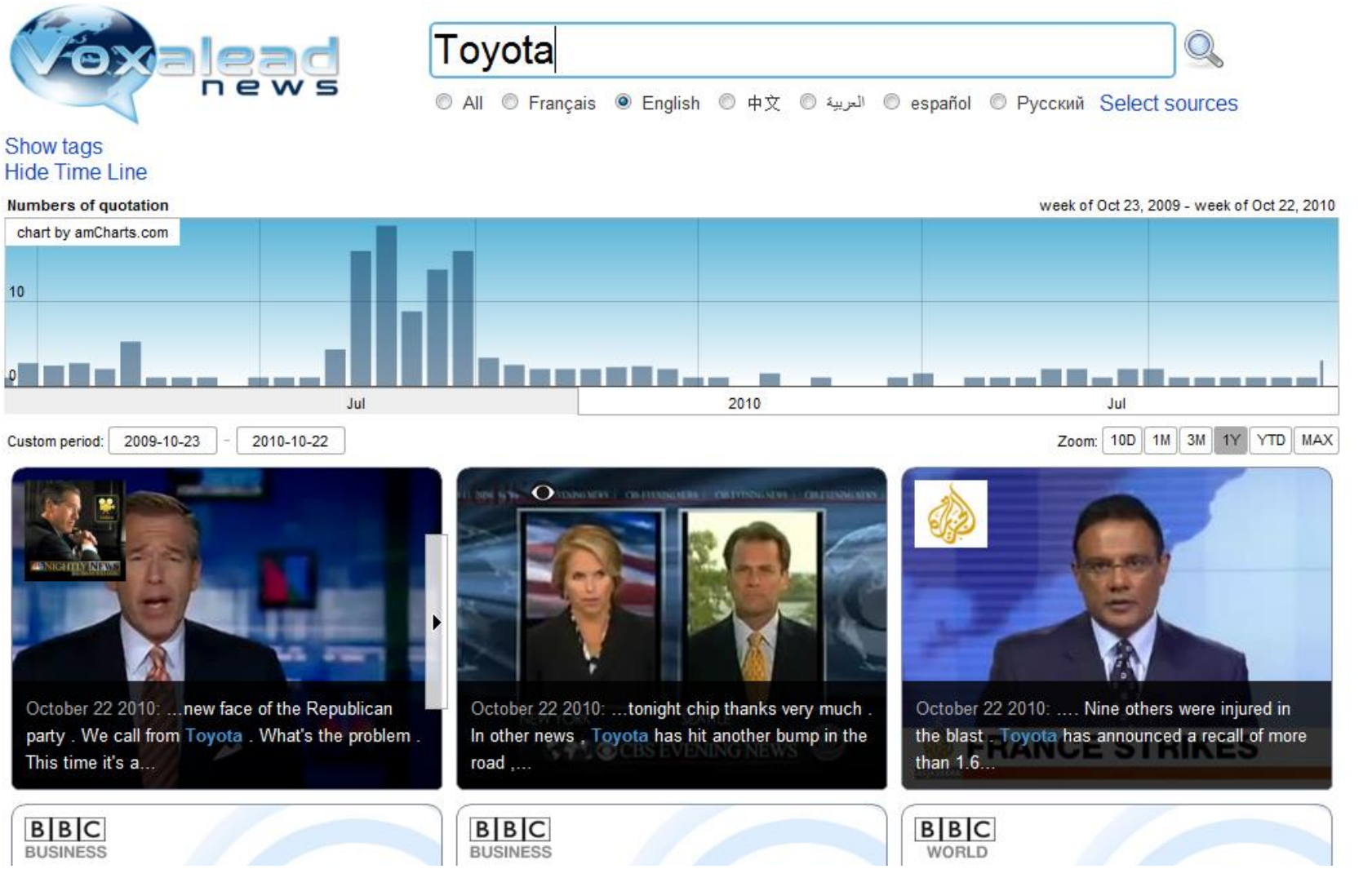
…dans le cadre de votre tolérance pour risque



Suivez vos résultats

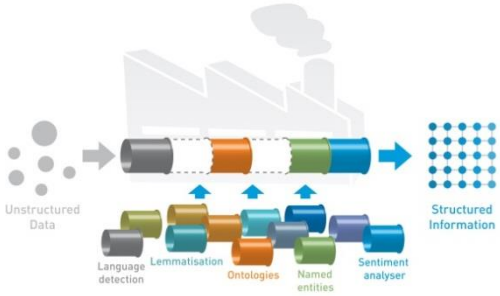


Surveillance/analyse en temps réel : Dashboards



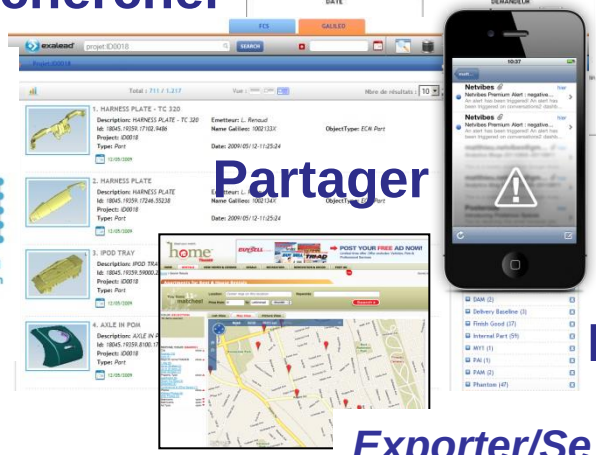
5. ON PEUT LE FAIRE COMMENT ?

Comment ?



Visualiser

Rechercher



Partager

Notifier

Explorer

Exporter/Se connecter

Capturer **Transformer** **Analyser/Action**

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

Infrastructure

WWW

App Internet

M2M

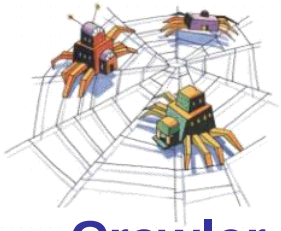
H2H

INTERNET

Capturer

Infrastructure

WWW



Crawler

API

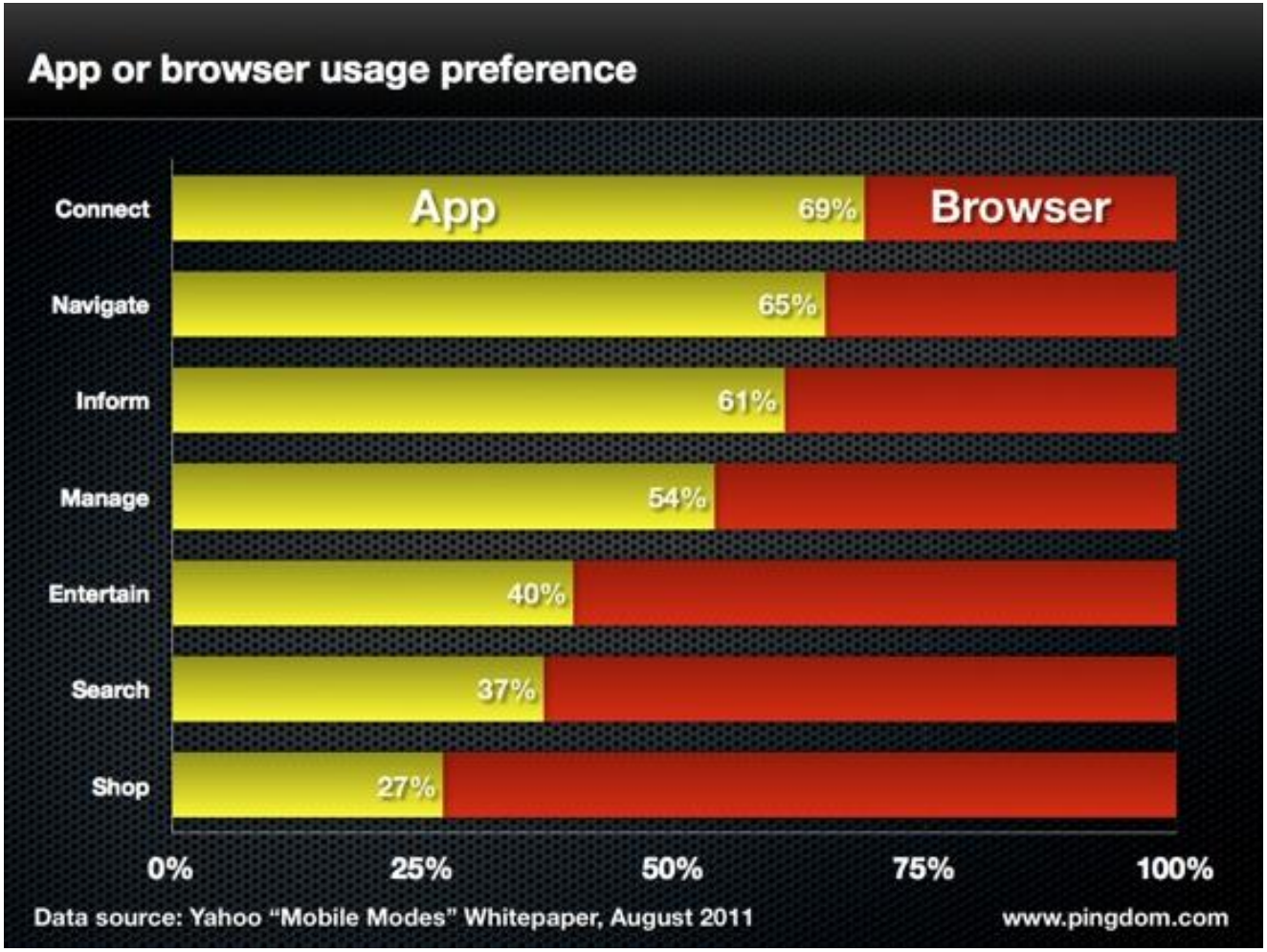
RSS

App Internet

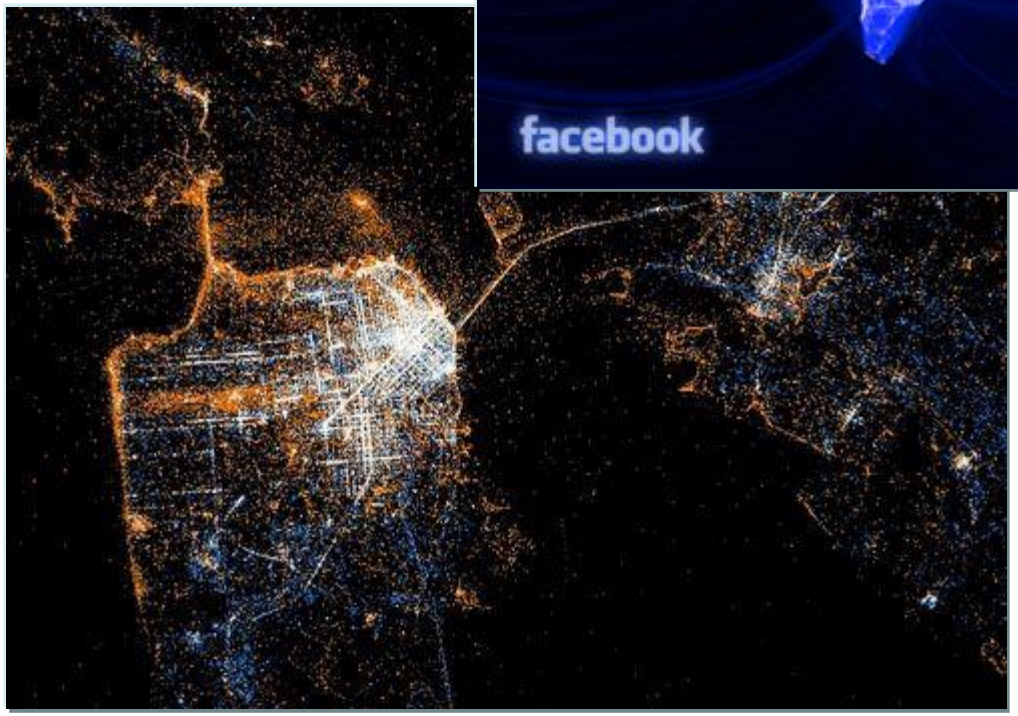
API

RSS

Apps: La fin du WWW?




facebook



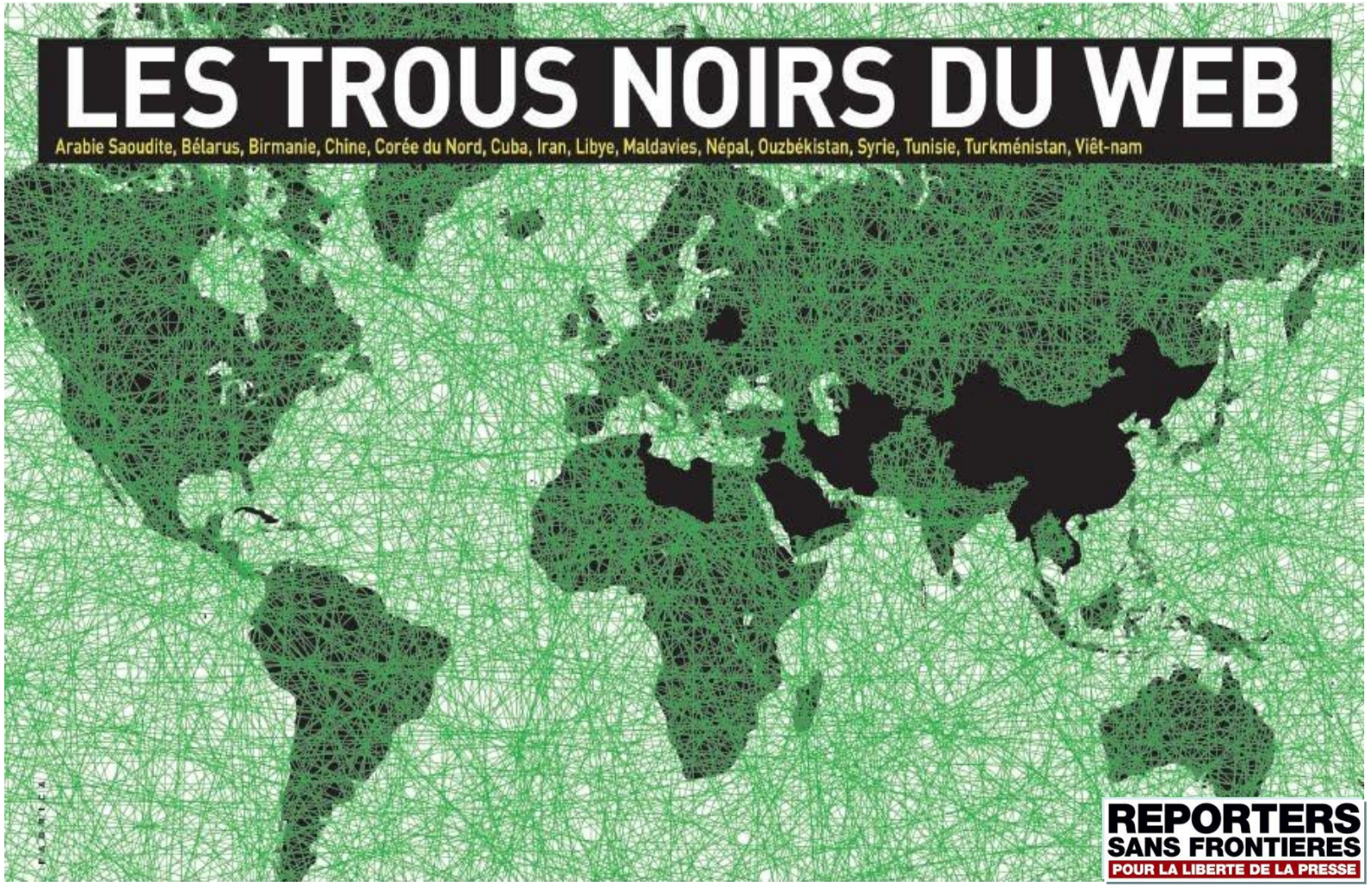
Crawler

Liste des URLs à crawler

Nouvelle liste des URLs à crawler

Nouveaux liens

Défis : Crawler



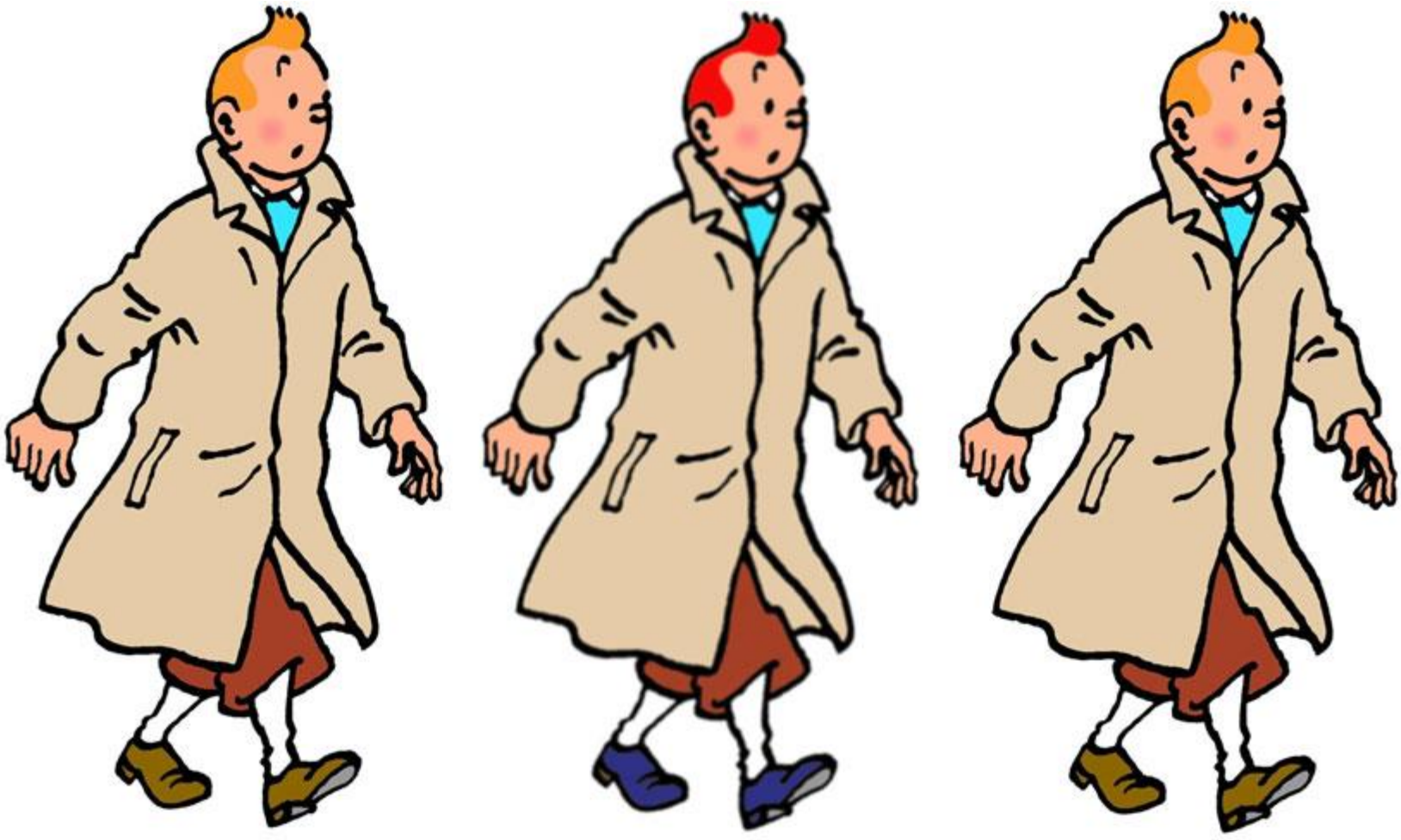
Défis : Crawler

Extraire du bon contenu

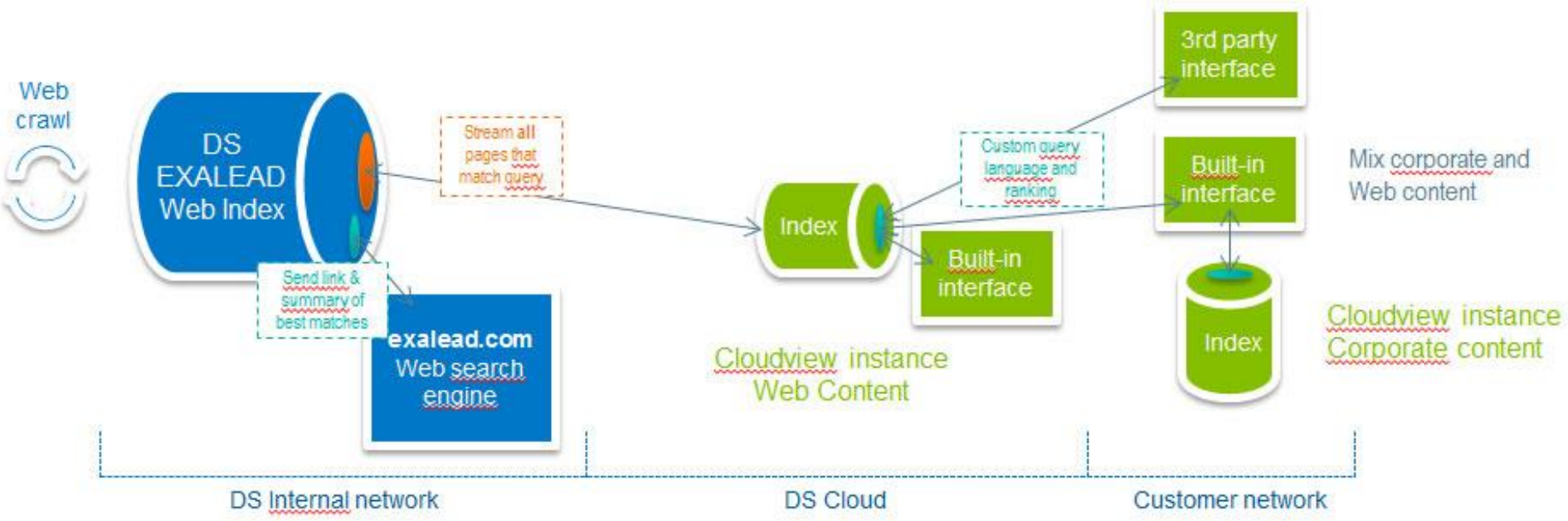
http://www.unixuser.org/~euske/python/webstemmer/howitworks.html

Crawler : Défis

Crawler : Défis Duplicata

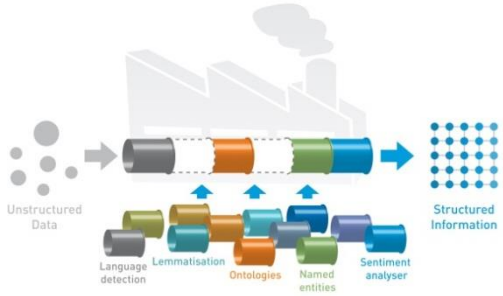


Crawler : Altérnatif



EXALEAD Web Mining Experience

Transformer : Traitement sémantique



Over 300+ file formats

Metadata extraction and normalization

125 languages supported

Split words and phrases

Detect parts of phrase (subject, verb, etc.)

Finds normalized form and radical

Document Conversion

Document Processors (Others & Custom)

Language Detection

Tokenizer

Part of Speech

Lemmatization / Stemming

Synonyms

Built-in Named Entities

Custom Named Entities (Ontology, Thesaurus)

Related Terms

Information Extraction

Indexing

Associate synonyms using configurable dictionaries

Identify people, places, organizations

Identify concepts from your application domain

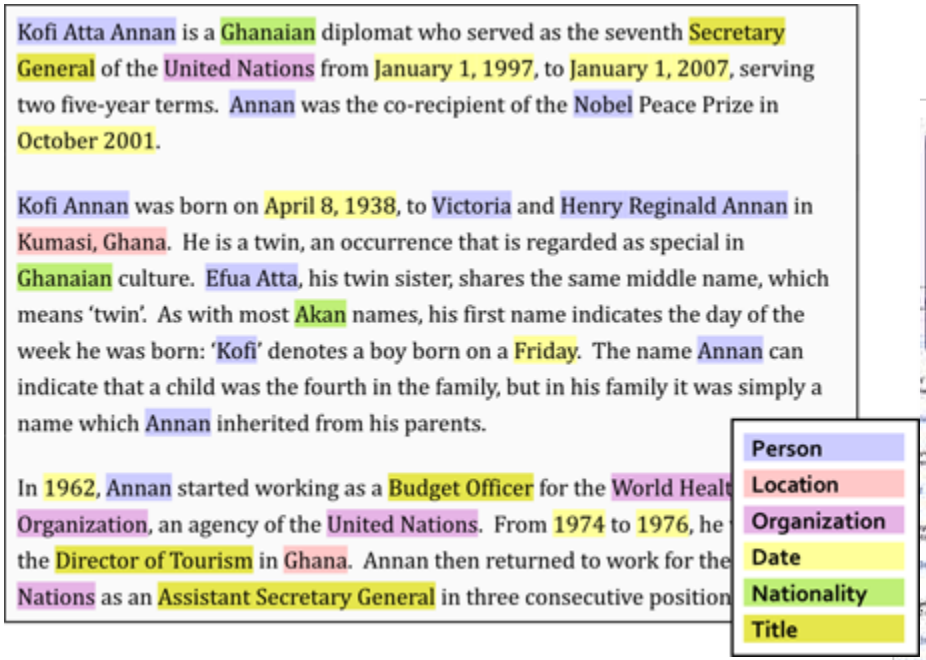
Identify frequent topics

Sentiment analysis Extraction rules

+ specialized or custom processors

EXALEAD

Transformer : Traitement sémantique



Entity extraction
Classification automatique
Analyse de sentiment…

EXALEAD

Transformer : Traitement sémantique

EXALEAD

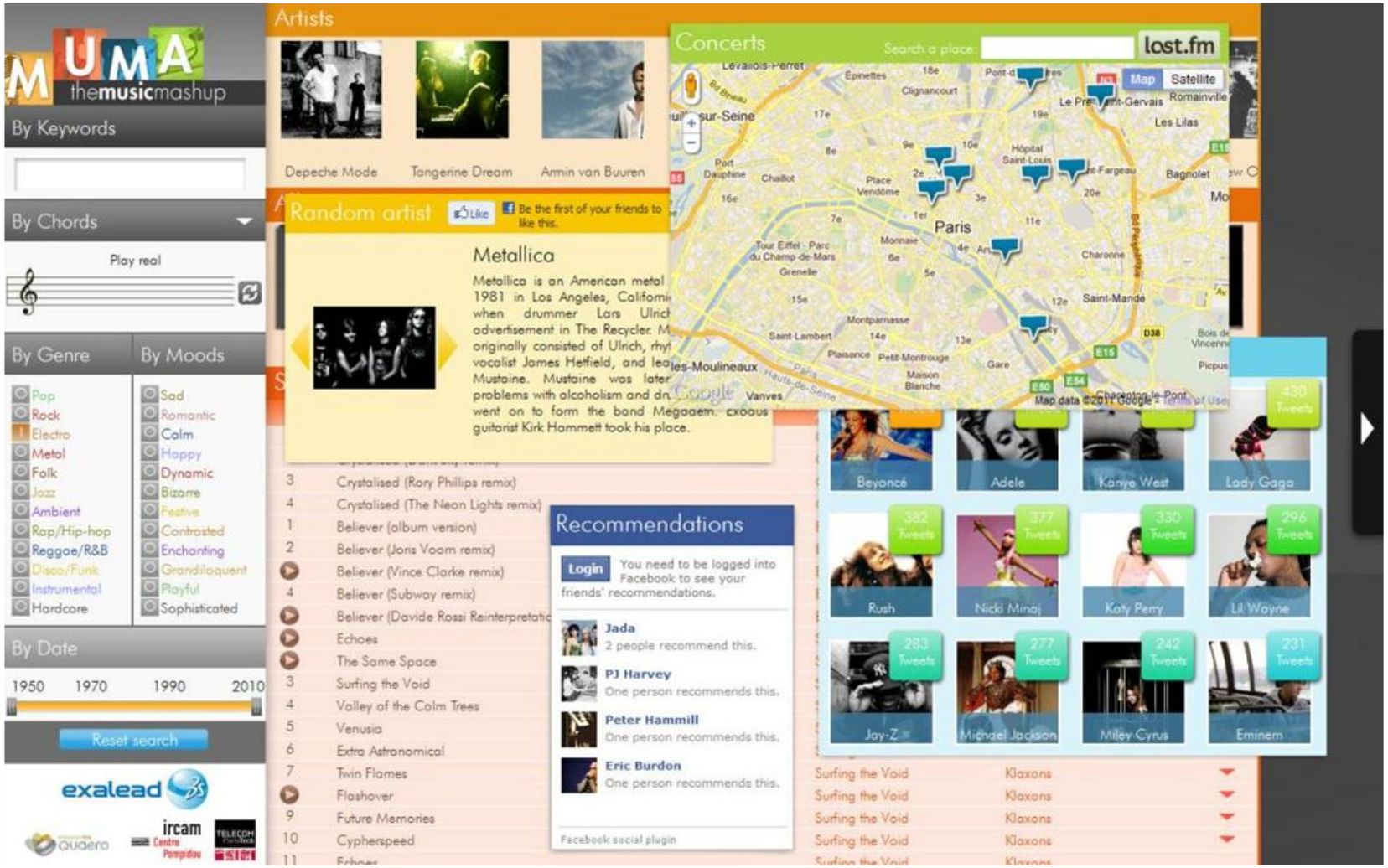
Transformer : Transcription ‘speech-to-text’; sémantique

EXALEAD

Transformer : Traduction

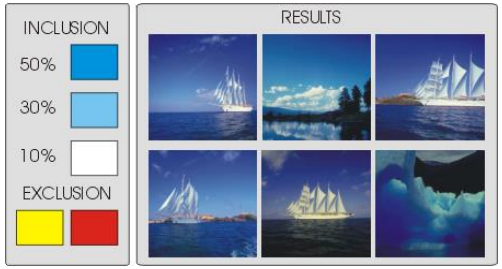
EXALEAD

Transformer : Analyse multimédia et sémantique



Images

COLEUR

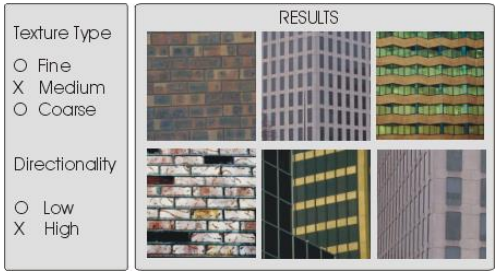


DESSIN

FORME

EXEMPLAIRE

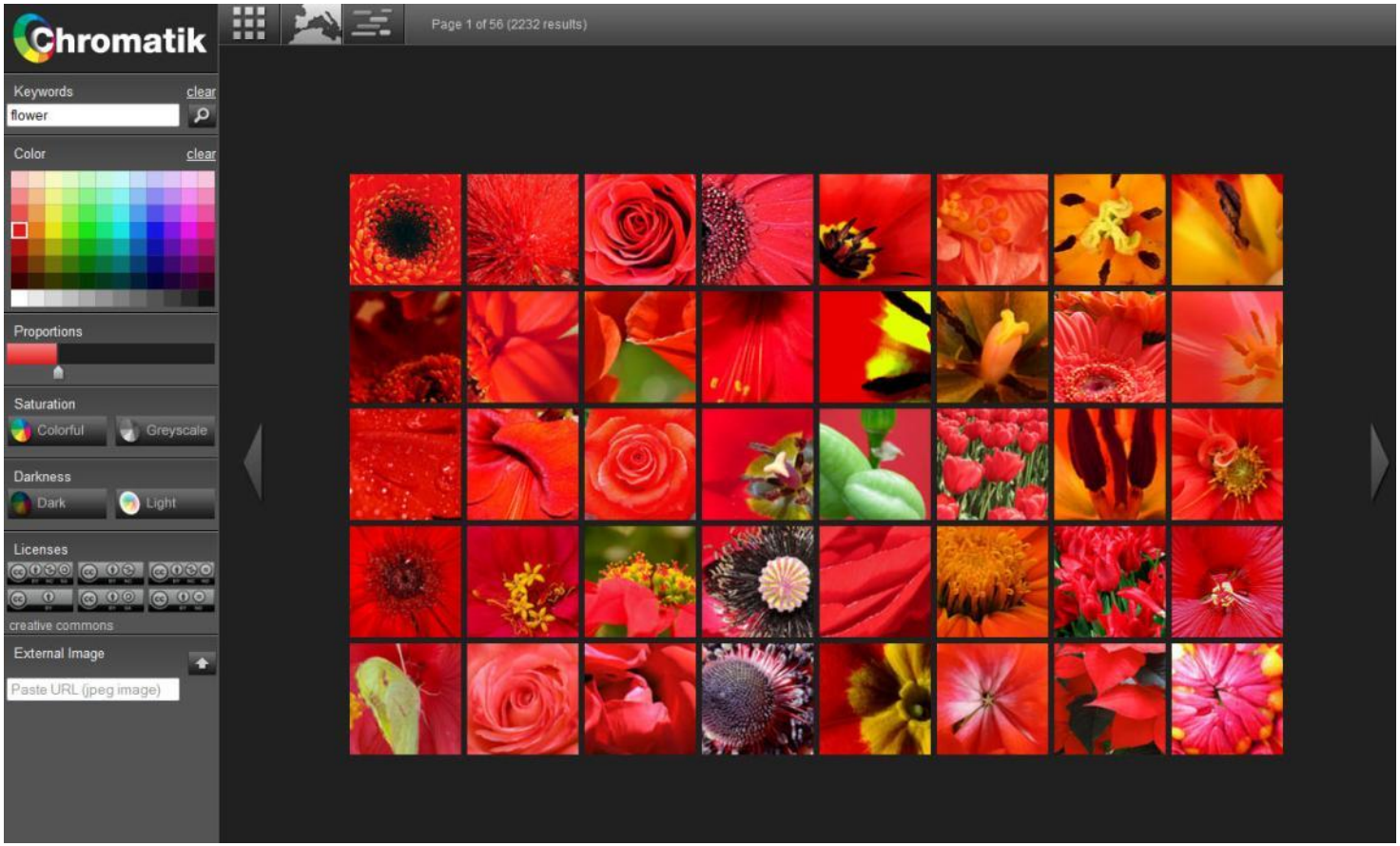
TEXTURE



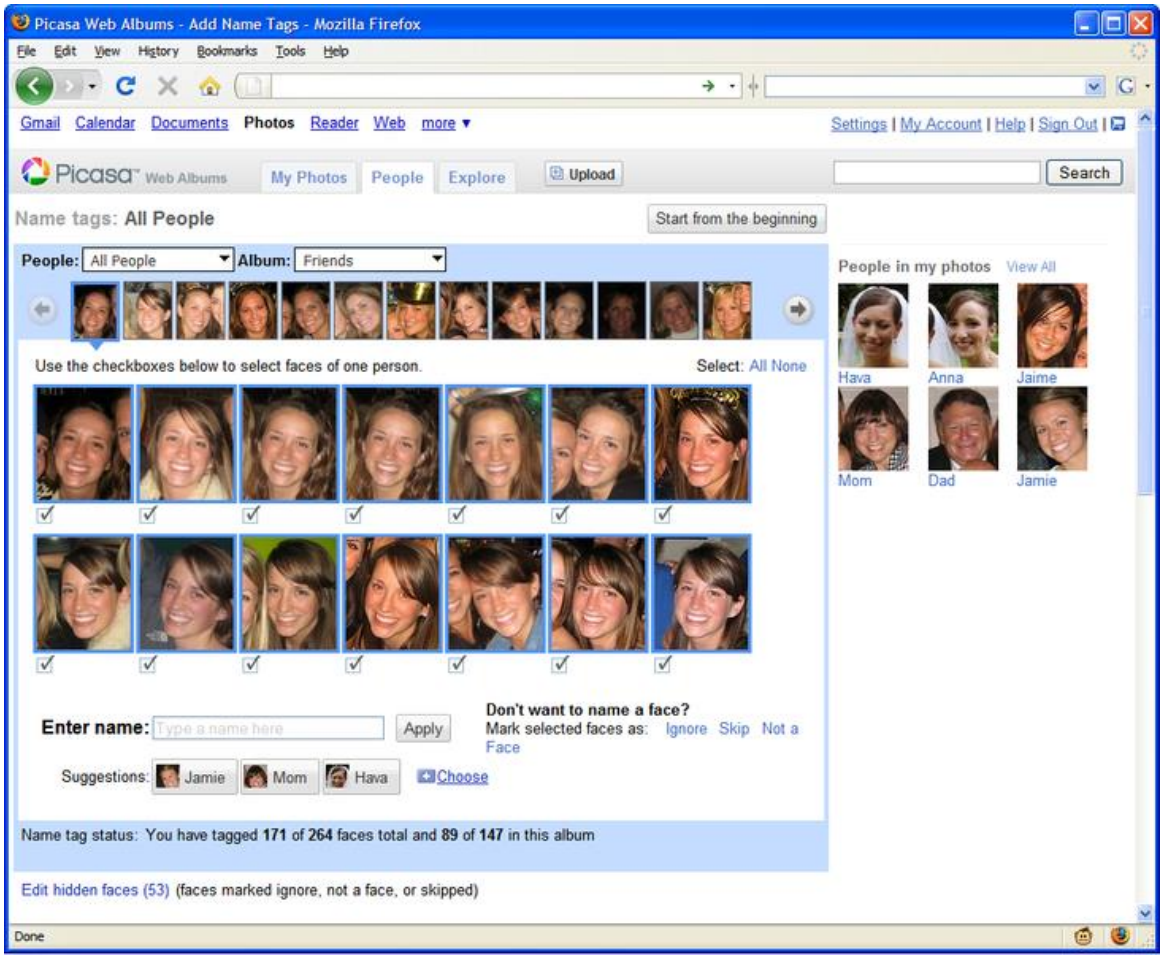
DES TYPES PLUS COMPLEXES EXISTENT MAIS ILS SONT DES PLUS FONDAMENTAUX & LES PLUS SOUVENT UTILSES

EXALEAD

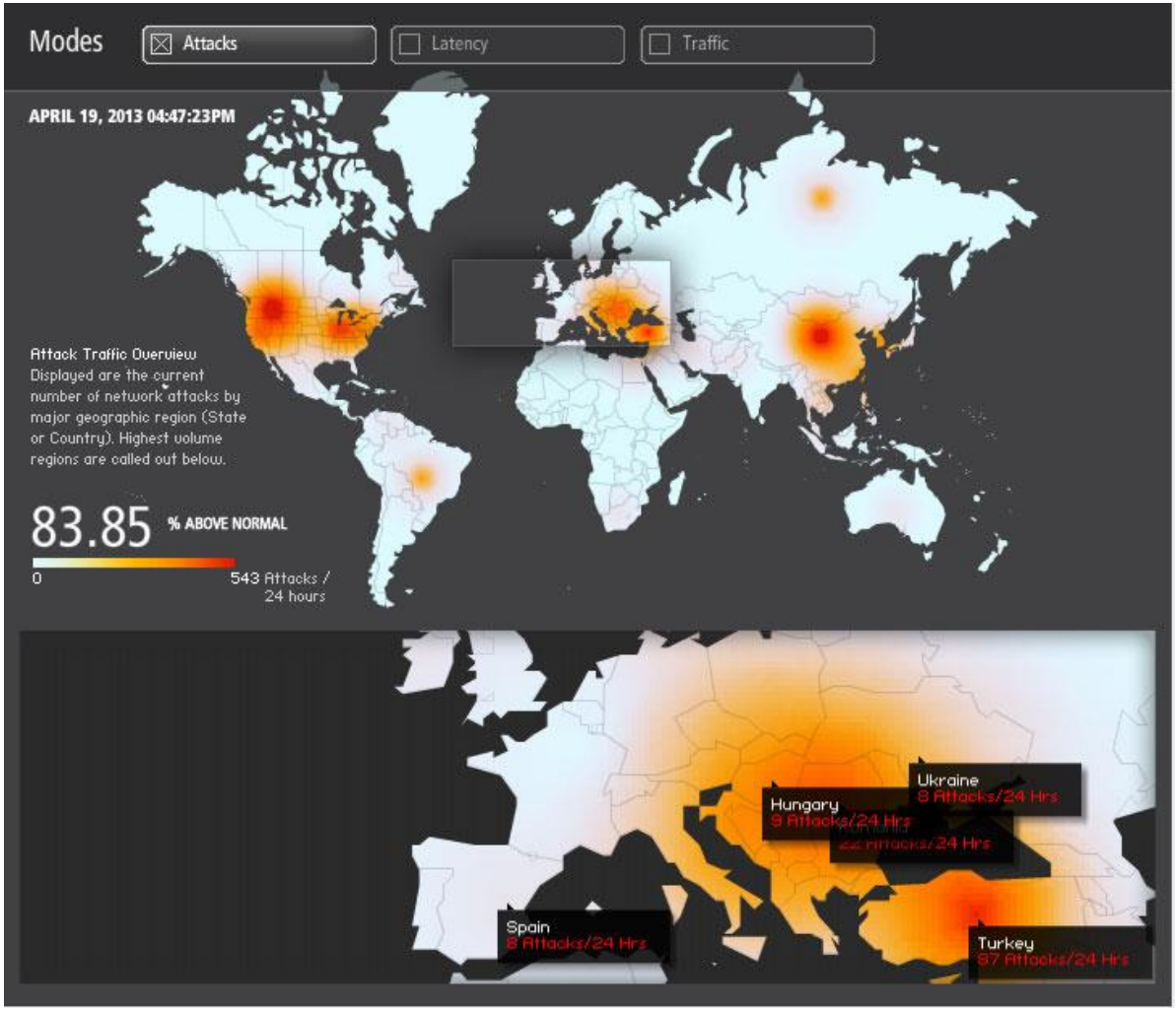
Transformer : Analyse multimédia et sémantique

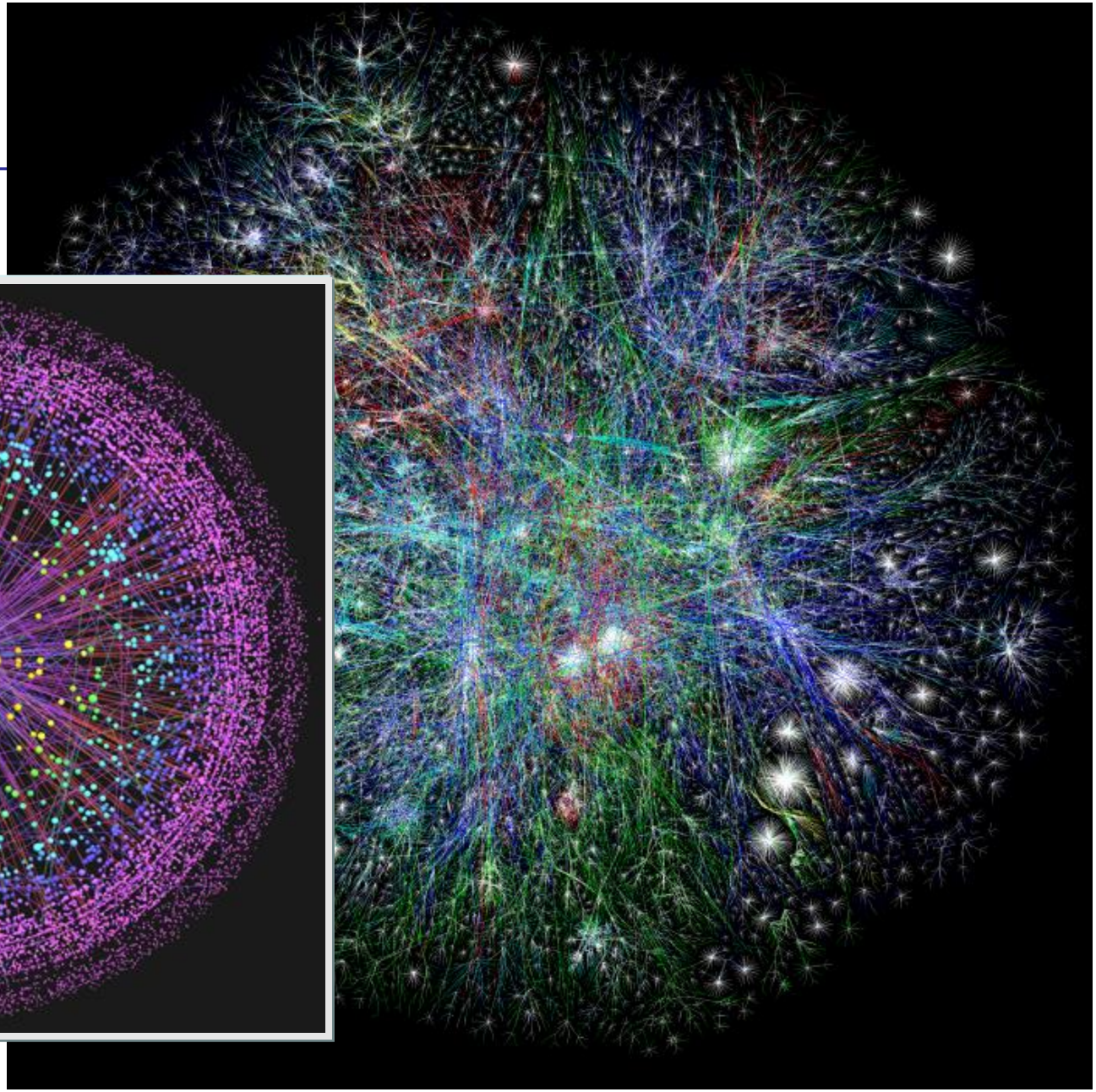


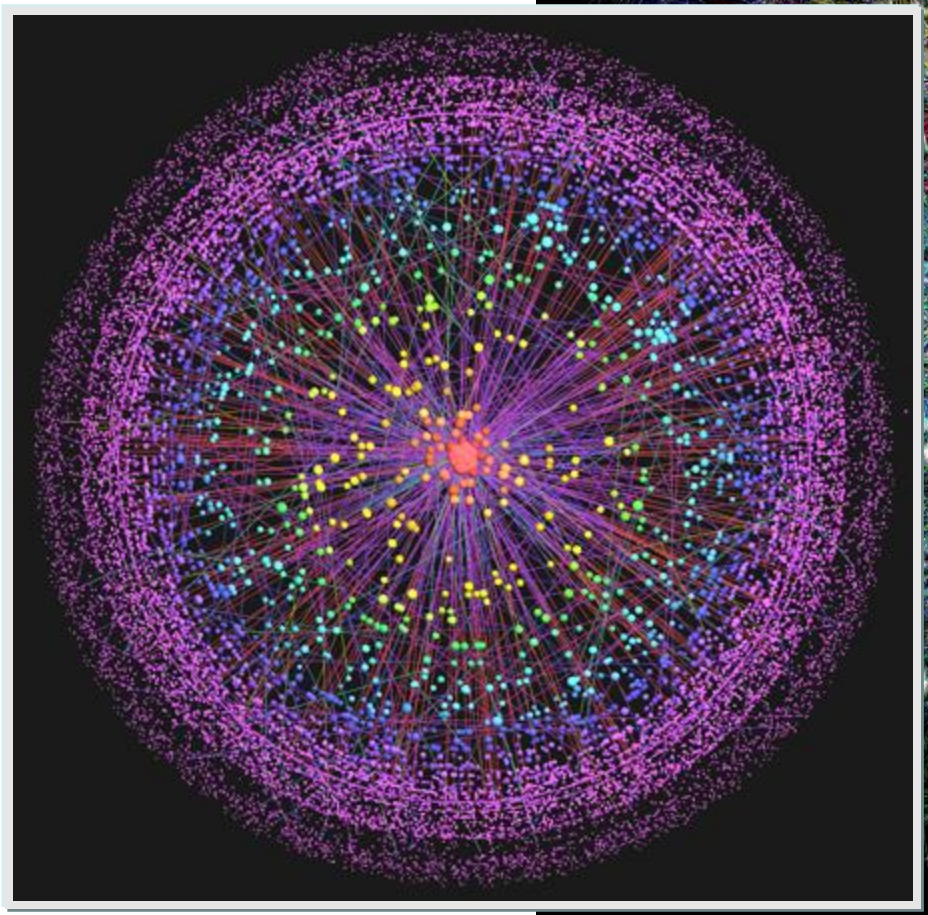
Transformer : Analyse multimédia et sémantique



Transformer : Analyse à soutenir la visualisation







Merci !

laura.wilber@3ds.com