Big data, les stéréotypes et l'opérationnalisation des recherches

Ecole Doctorale en Journalisme et Communication
3-4 décembre 2015

Giovanna Di Marzo Serugendo
Université de Genève
Giovanna.Dimarzo@unige.ch

ISS INSTITUTE OF INFORMATION SERVICE SCIENCE

UNIVERSITÉ DE GENÈVE
CENTRE UNIVERSITAIRE D'INFORMATIQUE

Big data – différents aspects

Données et leurs Sources

Traitement

Analyse, Algorithmes et Techniques

Outils disponibles sur le Web

Visualisation

Exemples

Aspects transverses

Données

Structurées, Non-structurées,
Persistentes, Transientes, Temps-réel
Bruit

Données structurées

Passager : Table

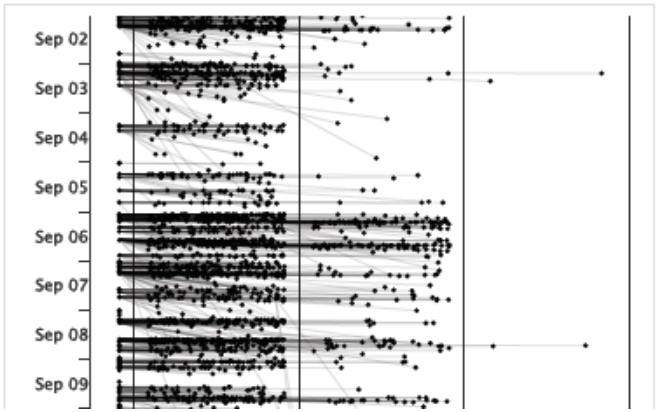
nom	prénom	n° vol	Nbr bagages	N° client
Einstein	Albert	45622	2	154565
Lavoisier	Antoine	45644	4	235002
Raimbault	Arthur	12896	2	544552
Poincaré	Henri	45644	3	781201
Lavoisier	Antoine	45644	1	785154
Einstein	Albert	75906	0	858547

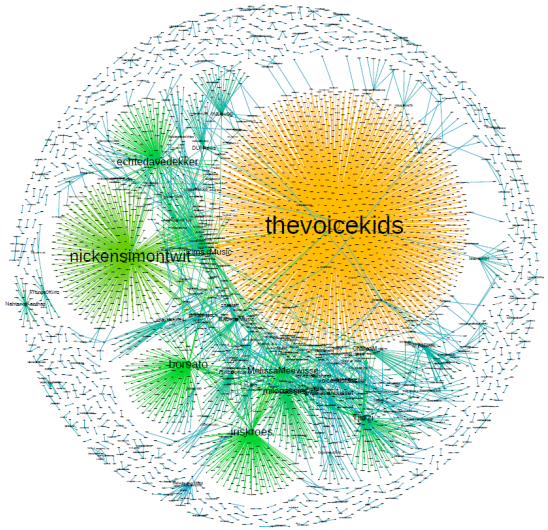
Les enregistrements de la relation **Passager**

Personne : Table

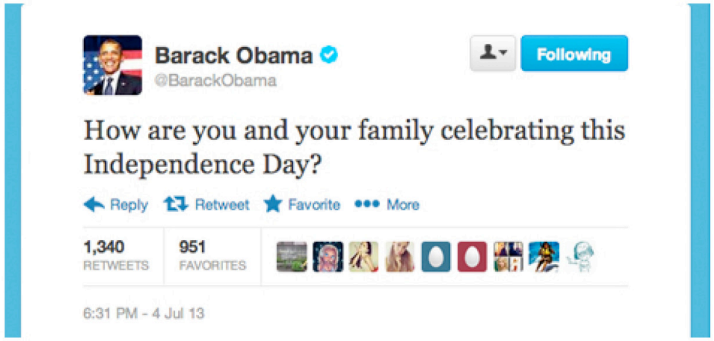
nom	prénom	age	civilité
Einstein	Albert	45	marié
Lavoisier	Antoine	41	marié
Planck	Max	52	veuf
Poincaré	Henri	45	marié
Raimbault	Arthur	25	célibataire

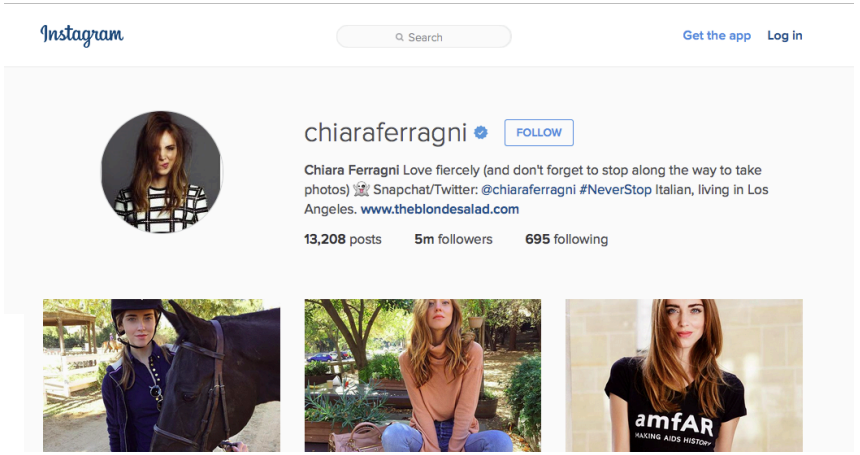
Les enregistrements de la relation **Personne**

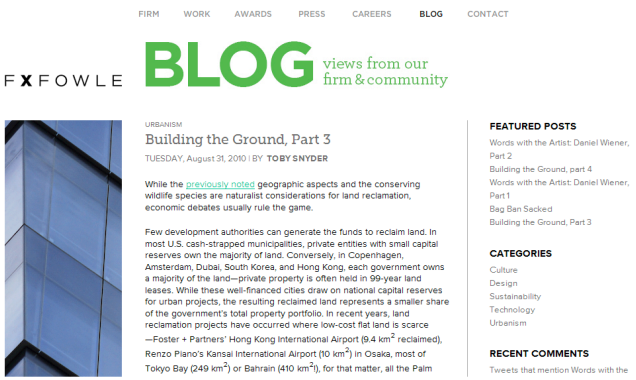




Données non structurées



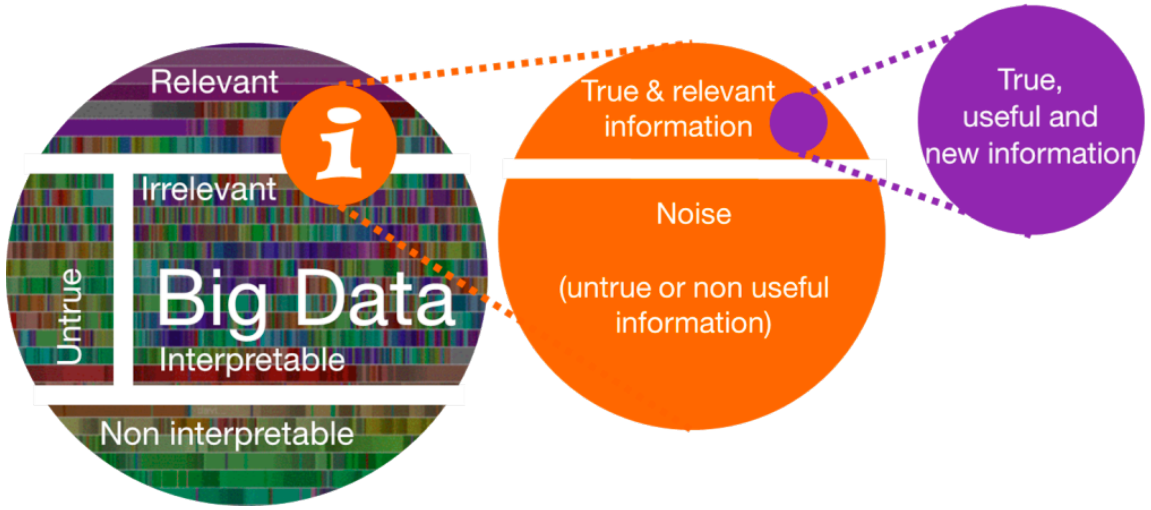




Persistent vs Transient vs Temps-réel

Bruit

- Bruit vs Signal intéressant
  - Sur 2’000’000 de tweets seuls 2000 sont pertinents (utiles) pour tirer des conclusions



Sources

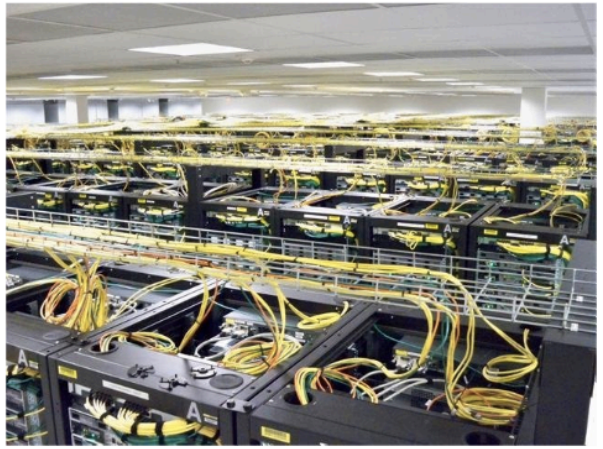
BD (classiques et propriétaires) (e.g. statistiques, archives (numériques)

Capteurs – physiques, virtuels, sociaux

Données ouvertes - Open data

Crowdsourcing

BD / Serveurs / Cloud

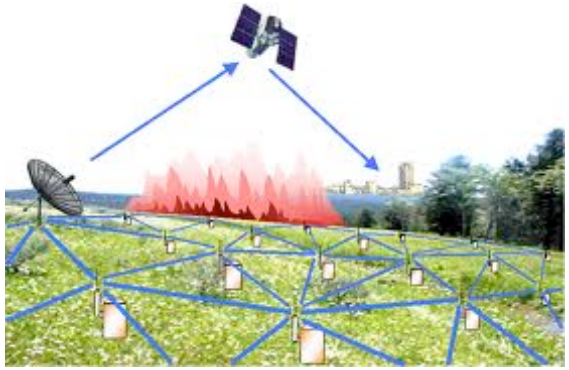


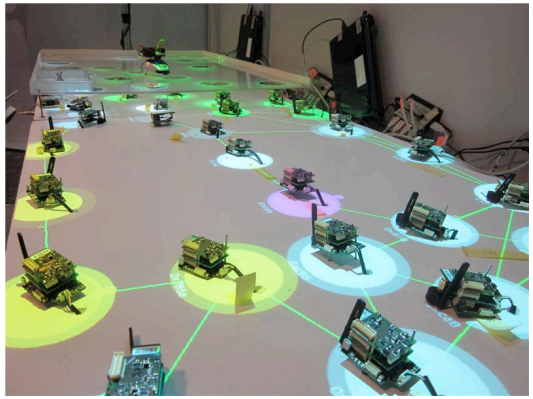
Capteurs – physique
Intelligence ambiante – Internet of things

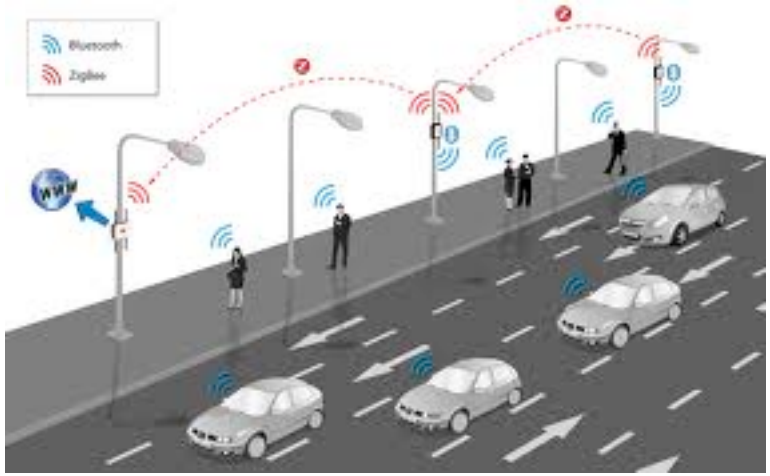
http://vimeo.com/41363473

Capteurs – physique
Internet of Things - Wearables

Capteurs physiques – environnement

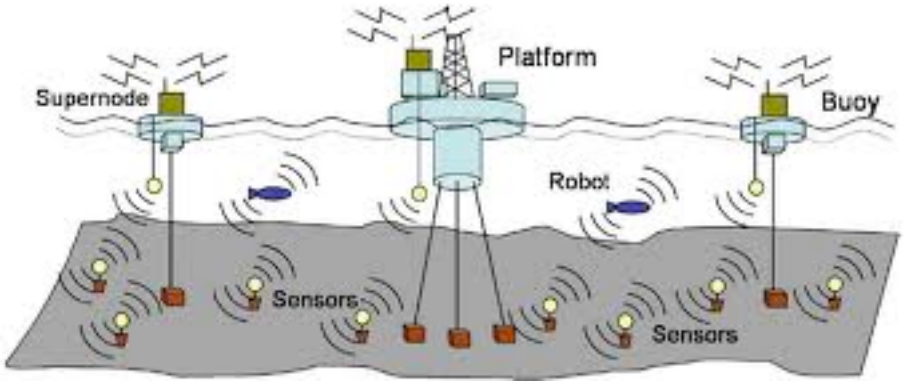




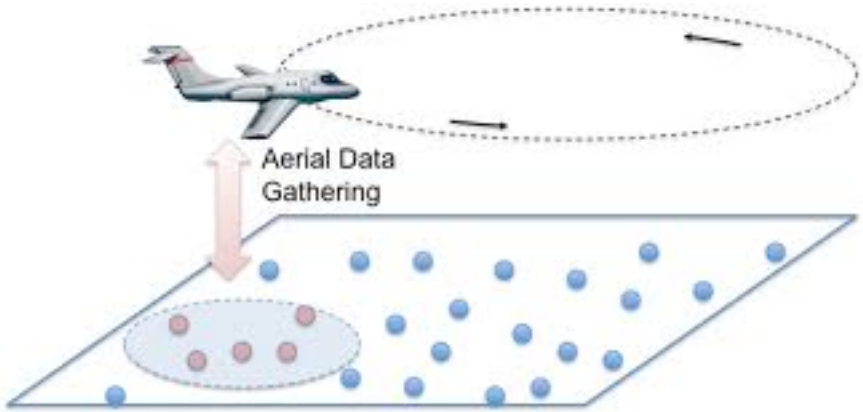


Wireless ou non

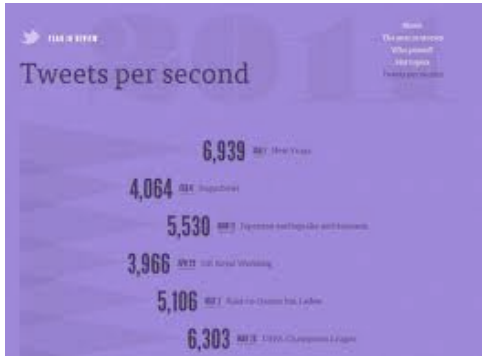
Capteurs – physiques
Environnement



Terrestre
Sous-marin
Aérien



Capteurs – virtuels



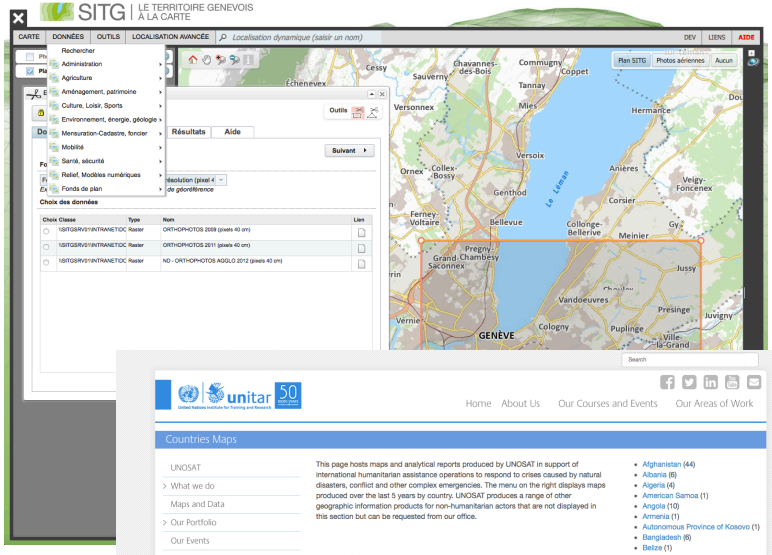
Capteurs - sociaux

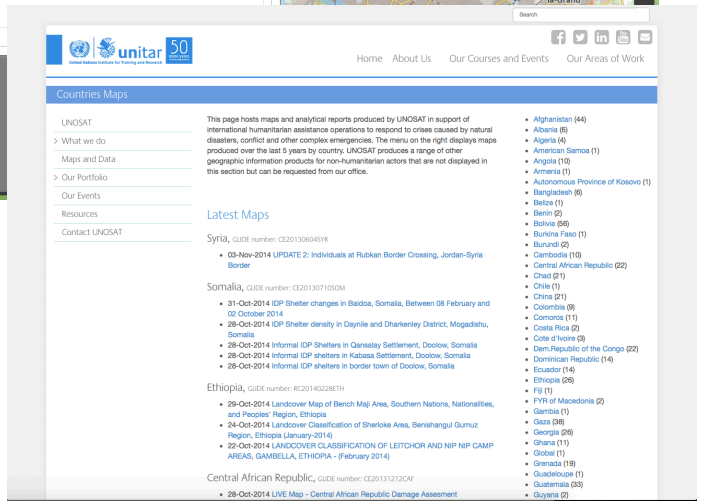
Médias sociaux et blogs

Open Data


tpg
</Open Data>
Accueil
Démarche
Données
Actualités
Concours
Forum
FAQ
INCUBATEUR D'IDÉES
Une API pour...
Créer...
Imaginer...
Partager...
Les tpg ouvrent leurs données...
En savoir plus
Problème technique
Le service est de nouveau opérationnel

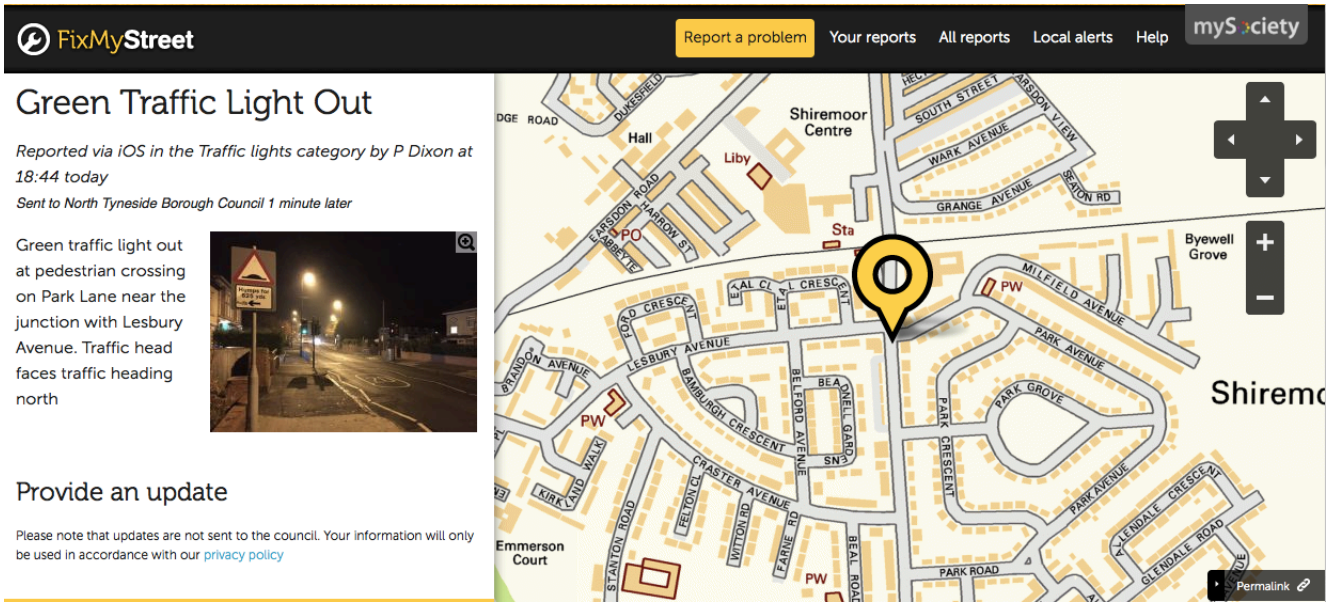

opendata.admin.ch
Administration fédérale admin.ch
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
opendata.admin.ch
Portail pilote des données ouvertes de l'administration publique suisse
Page d'accueil | Contact
Deutsch | Français | Italiano | English
Données
Applications
FAQ
Organisations
A propos du portail
Données ouvertes de l'administration publique suisse
Le portail pilote Open Government Data met à votre disposition 1841 jeux de données que vous pouvez analyser, combiner ou utiliser pour mettre au point des applications et des visualisations.
Rechercher p.ex. météo, initiatives, famille ...
Catégories
plus
Politique (1499)
Espace et environnement (116)
Santé (116)
Population (85)
Finances (8)
Education et science (4)
Administration (3)
Bases statistiques et généralités (2)
Législation (1)
Applications
plus
citymobile - die App für Schweizer Gemeinden
anthrazit
citymobile – Entdecken Sie di...


SITG
LE TERRITOIRE GENEVOIS À LA CARTE
CARTE
DONNÉES
OUTILS
LOCALISATION AVANCÉE
Résultats
Aide
Suivant
GENÈVE


unitar
Home
About Us
Our Courses and Events
Our Areas of Work
Countries Maps
UNOSAT
What we do
Maps and Data
Our Portfolio
Our Events
Resources
Contact UNOSAT
Latest Maps
Syria,
Somalia,
Ethiopia,
Central African Republic,

Crowdsourcing

Citoyens fournissent les données de manière participative – Citoyen est un capteur



Traitement

Collecte

Stockage

Exploitation – analyse, visualisation, services, distribution - DaaS

Analyse

- Descriptif
  - Trends, patterns, profil, sentiment analysis, opinion
    - On parle de la smart watch (twitter, google, …)
    - On parle en bien/mal/neutre de la smart watch
      - Plus dur: humour, sarcasme, …
    - Volume de discussion sur la smart watch et évolution du volume
    - Evolution des opinions
    - Qualité d'une discussion
      - Discourse Quality Index (politique)
    - Profil d'achat: famille avec petits enfants
    - Patient monitoré à la maison est tombé

Analyse

- Prédictif
  - Prédire le “futur”
    - La smartwatch ne va pas se vendre aussi bien que prévu
    - Google Flu – prédire deux semaines avant le début de l’épidémie de grippe
    - Prédire l’issue d’un vote
    - Prédire la densité d’une foule dans un événement pour éviter des désastres
    - Anticiper les maladies/crises cardiaques/AVC/ …

Analyse

- Prescriptif
  - Déterminer l'action à faire pour anticiper le "futur"
    - Identifier la politique publique qui va effectivement réduire le CO2
    - Quel prix pour assurer la vente d'un nouveau produit
    - Quelle décision à prendre face aux concurrents (effet Kodak)
    - Produire plus d'arômes de vanille

- Transformer, inférer, augmenter les données brutes

Algorithmes / Techniques

Machine Learning

- Supervisé
  - Construit un model à partir de données (input-output) d'exemple
- Non supervisé
  - Pas de données au départ
- Reinforcement Learning
  - Feedback de l'environnement qui permet de corriger les actions
- But: Faire des prédictions

Algorithmes / Techniques

Data Mining

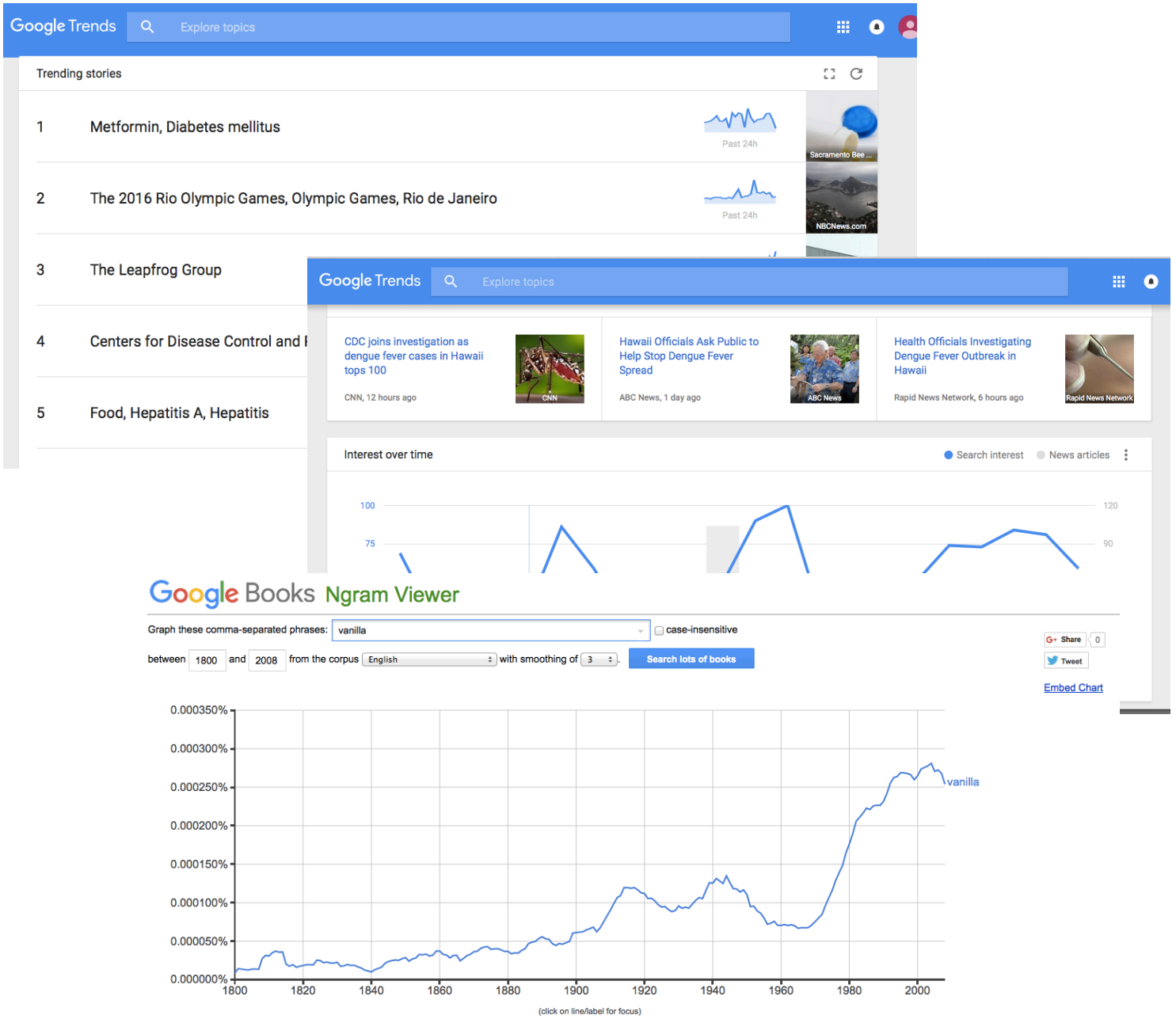
- Détection d'anomalies (déviations)
- Règles d'association entre variables (produits achetés simultanément)
- Clustering (grouper les données similaires sans les connaîtres à l'avance)
- Classification (mettre des labels, topic modelling)
- Régression (trouver une fonction qui modélise les données)
- Résumer (représentation compacte des données)
- But: découvrir des patterns

Algorithmes / Techniques

- Ingéniérie des connaissances – ontologies
- Analyse de réseaux (graphes)
- Analyse des opinions/sentiments (e.g. lexiques)
- Trends, n-grams, extraire des mécanismes/modèles, patterns,
- LDA – Latent Dirichlet Association

Outils sur le Web / Open Source

- Google Trends
  - https://www.google.com/trends/
- Google N-grams viewer
  - https://books.google.com/ngrams
- Weka machine learning – open source
  - http://www.cs.waikato.ac.nz/ml/weka/
- LDA / Twitter-LDA – Topic modelling / labelling
  - http://mallet.cs.umass.edu/


Google Trends
Explore topics
Trending stories
1 Metformin, Diabetes mellitus
Past 24h
2 The 2016 Rio Olympic Games, Olympic Games, Rio de Janeiro
Past 24h
3 The Leapfrog Group
4 Centers for Disease Control and
5 Food, Hepatitis A, Hepatitis
Sacramento Bee ...
NBCNews.com
Google Trends
Explore topics
CDC joins investigation as dengue fever cases in Hawaii tops 100
CNN, 12 hours ago
CNN
Hawaii Officials Ask Public to Help Stop Dengue Fever Spread
ABC News, 1 day ago
ABC News
Health Officials Investigating Dengue Fever Outbreak in Hawaii
Rapid News Network, 6 hours ago
Rapid News Network
Interest over time
Search interest
News articles
100
75
120
90
Google Books Ngram Viewer
Graph these comma-separated phrases: vanilla
case-insensitive
between 1800 and 2008 from the corpus English with smoothing of 3
Search lots of books
Share 0
Tweet
Embed Chart
0.000350%
0.000300%
0.000250%
0.000200%
0.000150%
0.000100%
0.000050%
0.000000%
vanilla
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
(click on line/label for focus)

Google BigQuery Service

- http://www

Google's BigQuery Offers Infrastructure to Crunch Big Data

Google today announced the general availability of its cloud-based BigQuery Service, an online analytical processing (OLAP) system designed for crunching terabyte-scale datasets using the search engine giant's infrastructure.

By Thor Olavsrud
Tue, May 01, 2012

1 Comment

+ Briefcase What's this?

30 Tweet

5 +1

CIO — Few companies in the world have access to datasets as large as Google does, and, unsurprisingly, Google is one of the companies at the forefront of Big Data analytics. Now Google plans to share the wealth by giving others access to its data crunching infrastructure with its new Google BigQuery Service.

The BigQuery service is an online analytical processing (OLAP) system designed for terabyte-scale datasets. It gives customers the capability to run SQL-like queries against massive datasets that potentially have billions of rows without requiring the hardware and software costs associated with an

Visualisation

Réalité augmentée

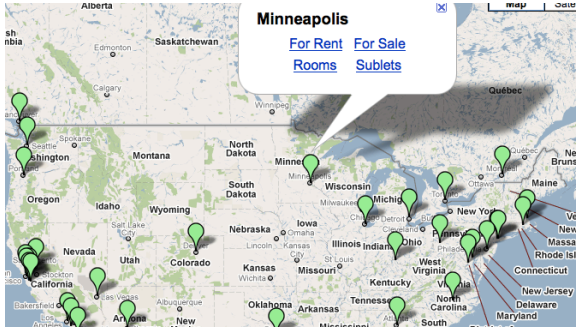
Mashups

Représentations animées

Réalité augmentée pour les journalistes

http://www.techrepublic.com/article/immersive-journalism-what-virtual-reality-means-for-the-future-of-storytelling-and-empathy-casting/?tag=nl.e101&s_cid=e101&ttag=e101&ftag=TRE684d531

Mashups



- **Combination with Google Maps**
- HousingMaps: combines rental listings from Craigslist with Google Maps for a visual representation of local apartments for rent.
- Wikipediavision: combines Google Map and a Wikipedia API
- List of such services: http://googlemapsmania.blogspot.com/

- **Combination with eBay**

- **RSS feeds** and Google Maps
- http://mapifiedrss.gmapify.fr/

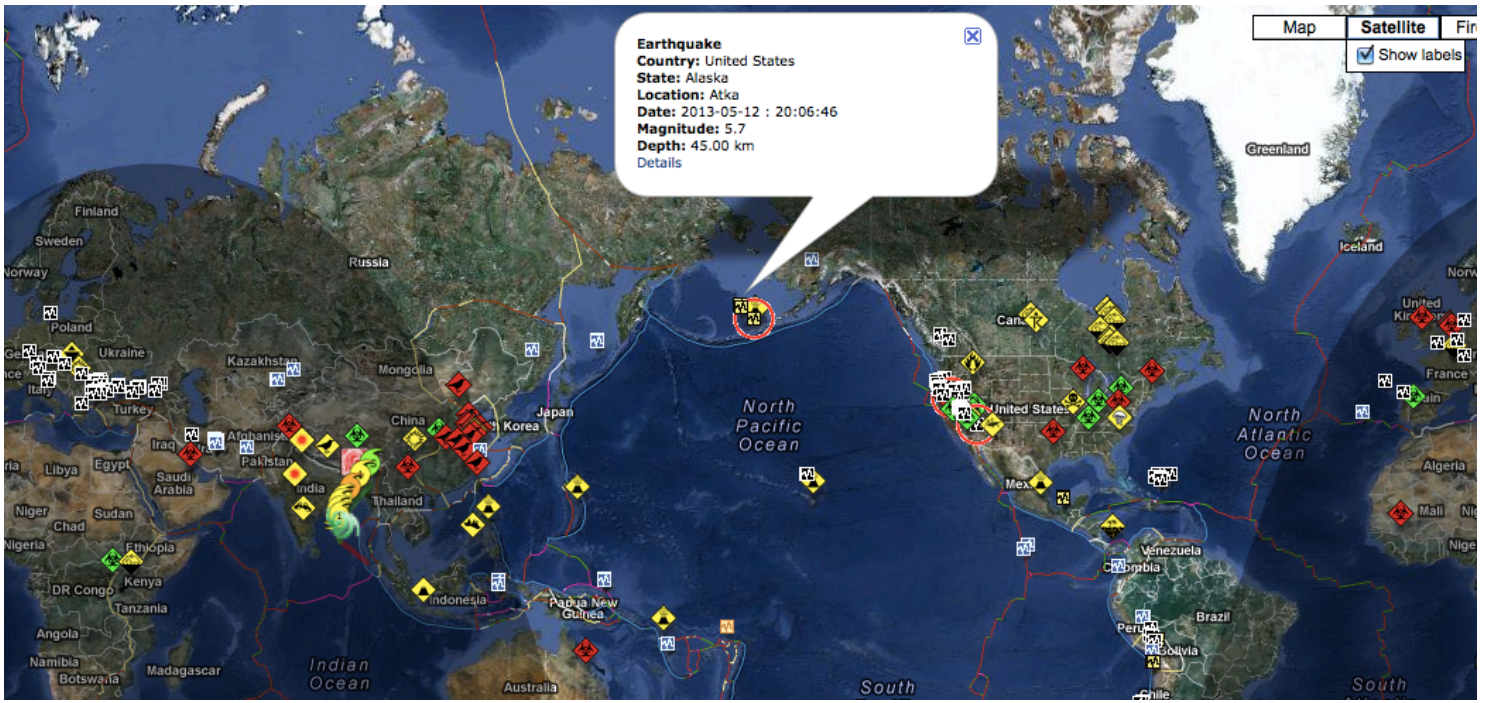
Planisphère des Tweets

http://bigbrowser.blog.lemonde.fr/2013/05/09/planisphere-le-monde-en-tweets/

Real-time Tweeter Map

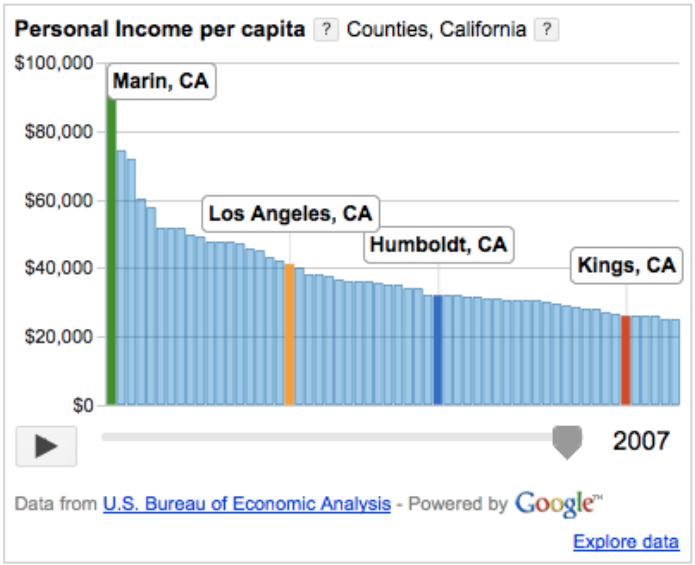
- http://tweereal.com/

Senseurs virtuels – interrogeables sur une carte

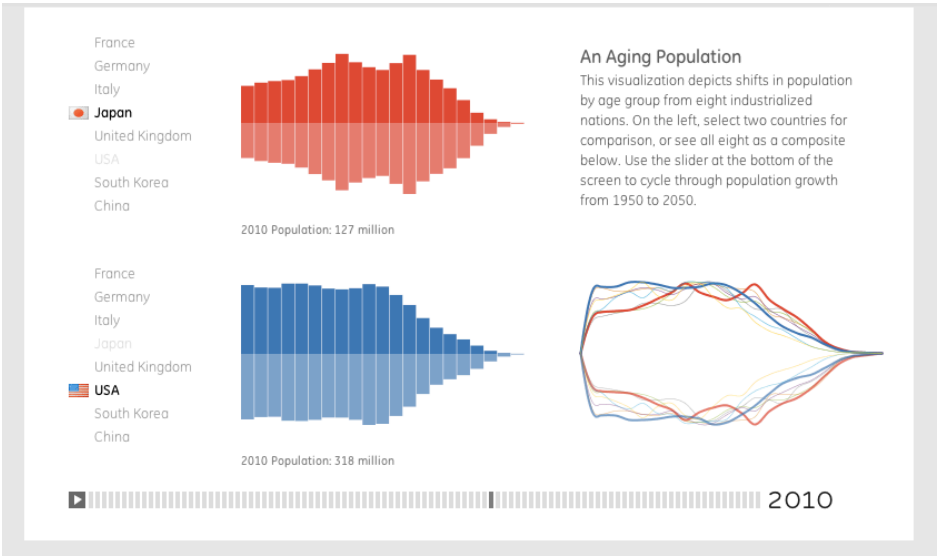


http://hisz.rsoe.hu/alertmap/index2.php
http://nathazmap.com/hazard_maps/interactive

DaaS - Examples



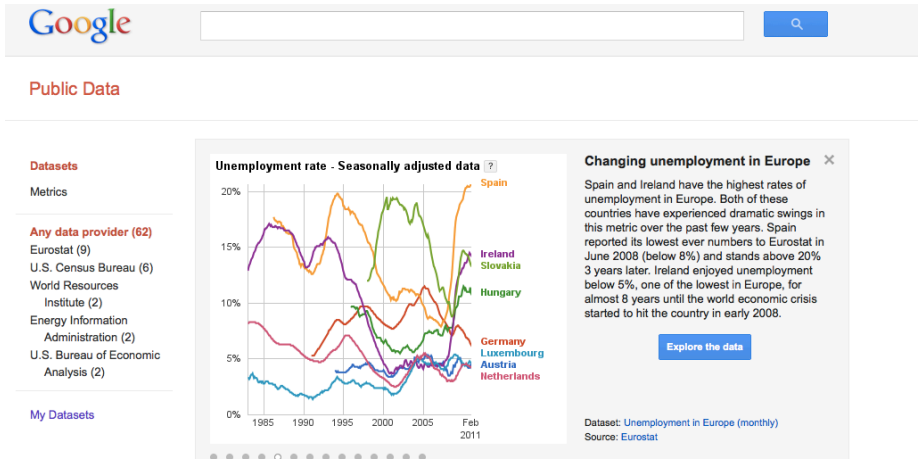
http://radar.oreilly.com/2010/07/data-as-a-service.htm
l



http://visualization.geblogs.com/visualization/aging/

DaaS - Providers

Public Data Explorer (Google)



http://www.google.com/publicdata/directory?hl=en&dl=en#!

Examples et applications

Jeux Sérieux

Science citoyenne, participative

Serious Games

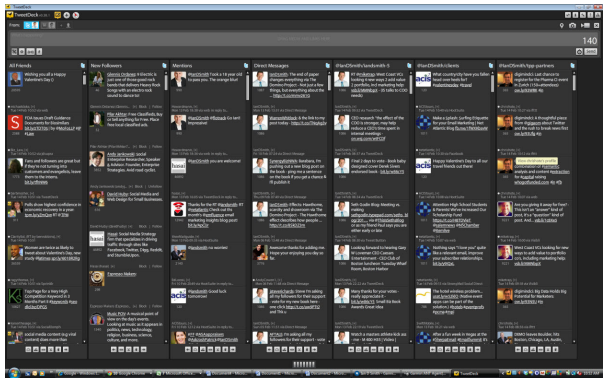
La RTS lance Tabula Rasa, le jeu web où l'on crée sa Suisse idéale

http://www.rts.ch/info/suisse/6850695-la-rts-lance-tabula-rasa-le-jeu-web-ou-l-on-cree-sa-suisse-ideale.html

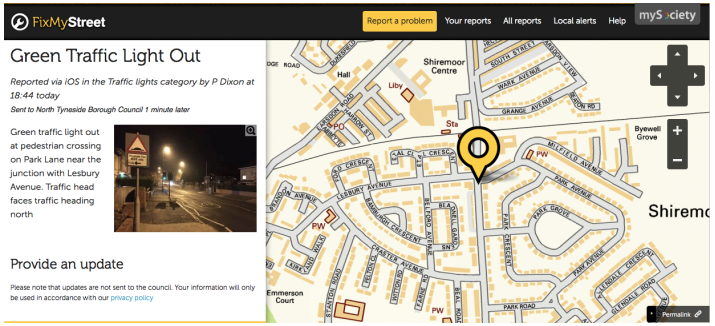
www.tabularasa.ch

Crowdsourcing / Science participative

Partager des informations



Citoyen est un “capteur”



Citoyen est un scientifique

Engagement démocratique

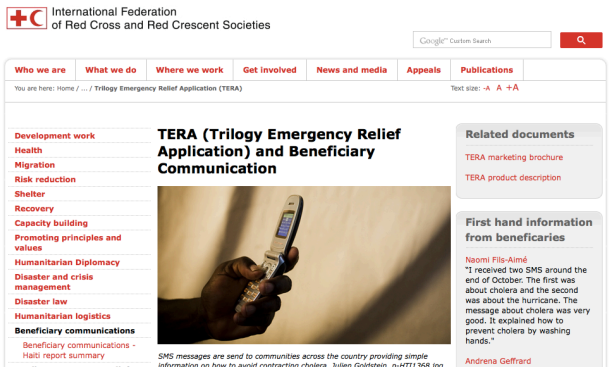
Technology tools

Plateformes Web, médias sociaux

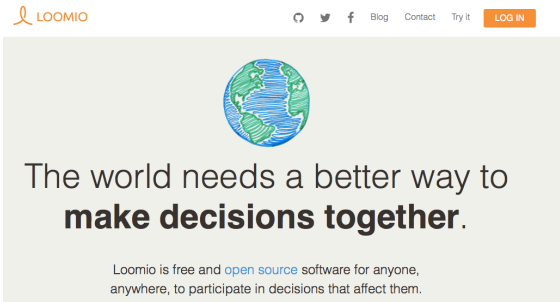
Plateformes de crowdsourcing

citizen cyberLab

Téléphones portables



Plateformes open source Démocratie

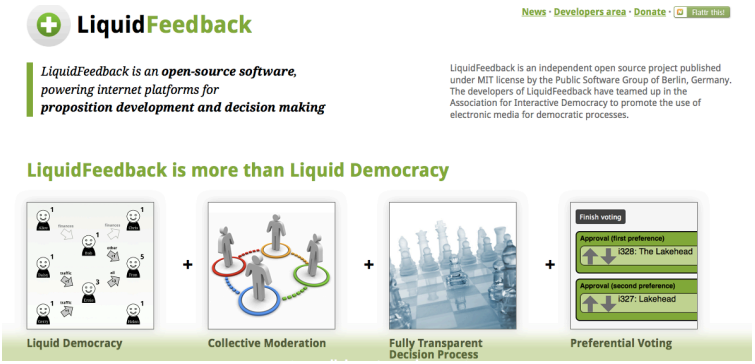


Platforms for democracy and engagement

Votes / signatures électroniques

Trouver parti (voting advice)

Vote préférentiel et opinions



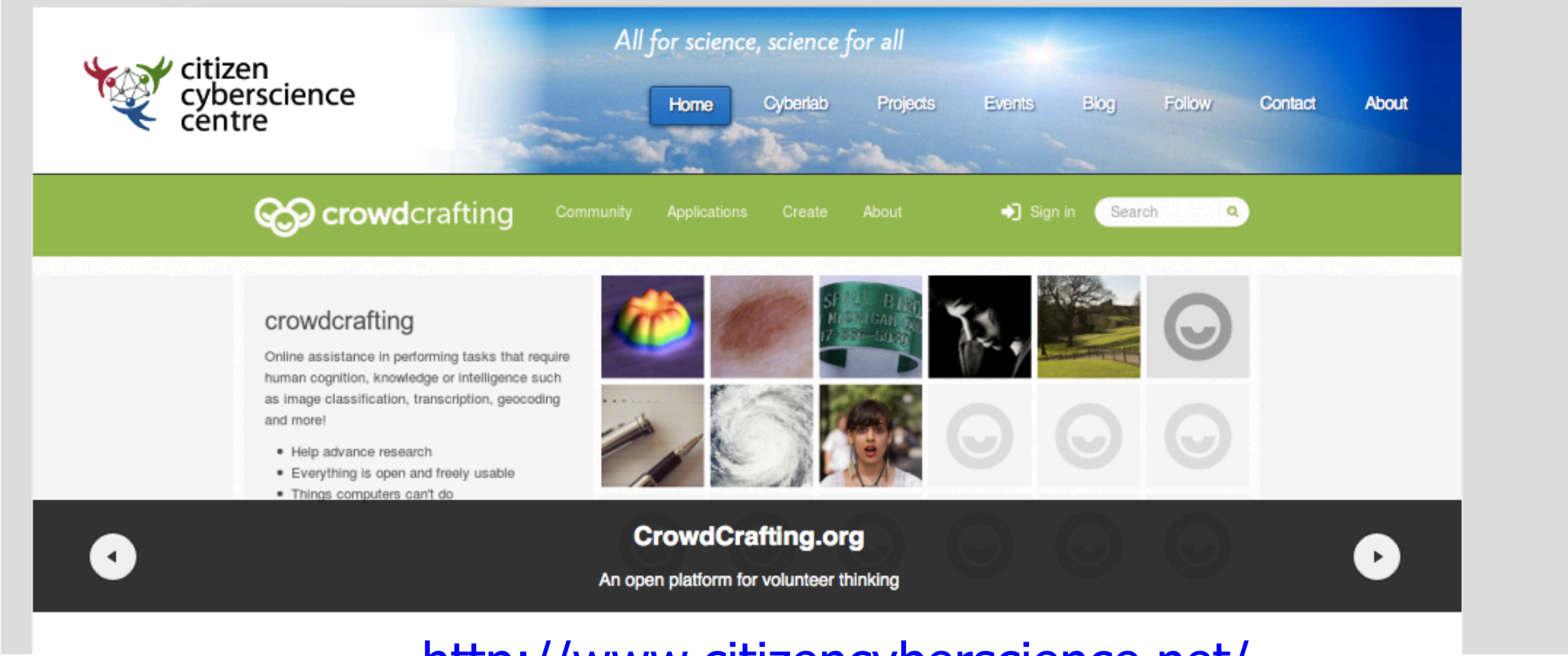
Information, conversation, vote

https://www.republique-numerique.fr/

http://www.unglobalpulse.org/urban-resilience-petajakarta (video)

Crowd Sourcing

- Systèmes socio-techniques
  - Connaissance/information détenue par les humains



http://www.citizencyberscience.net/

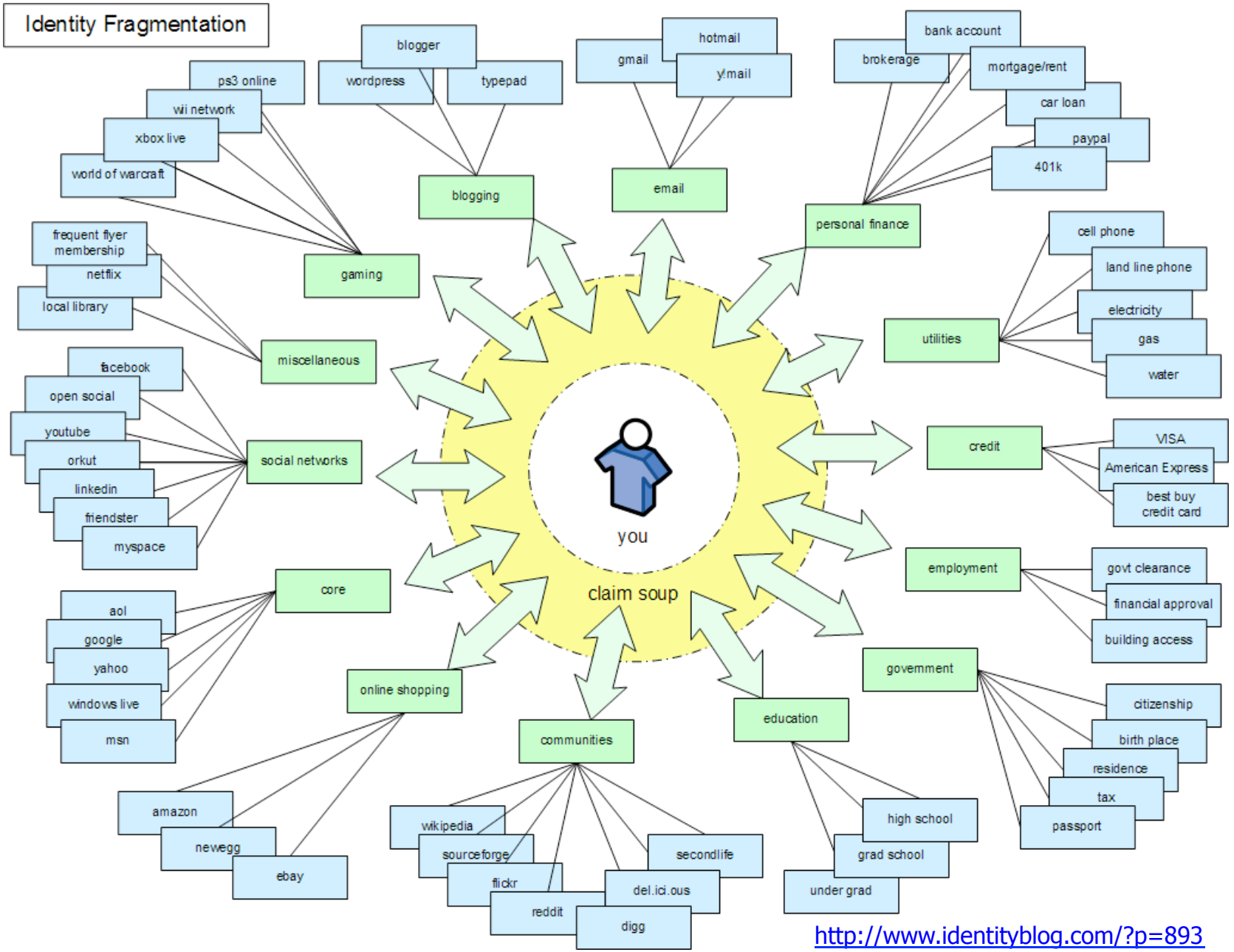
Aspects Transverses

Identité numérique

Réputation en ligne

Protection de la sphère privée

Anonymisation vs Re-identification



http://www.identityblog.com/?p=893

Mobile Phones

Le Monde.fr
Technologies

Abonnez-vous au journal Le Monde : 17€/mois

Recherchez sur Le Monde.fr

Recevez les newsletters gratuites

S'identifier

ACTUALITÉS | DÉBATS | sport | LOISIRS | PRATIQUE | VOUS | VOTRE INFO | LE MONDE | LES NEWSLETTERS | LES DOSSIERS

International | Planète | Europe | Politique | Société | Carnet | Economie | Médias | Sport | Technologies | Culture | Webdocus

A la Une > Technologies

Ces données privées que les applications mobiles transmettent à votre insu

LEMONDE.FR | 20.12.10 | 16h00 • Mis à jour le 20.12.10 | 20h09

ÉDITION ABONNÉS 4 € par mois

Partagez | Facebook

Recommander | 1054 recommendations. Inscription to see what your friends recommend.

Localisation, âge, sexe, identifiants : la plupart des applications sur téléphone mobile envoient des données privées à des régies publicitaires sans que l'utilisateur en soit informé, selon une enquête du *Wall Street Journal*. Sur 101 applications populaires étudiées par le journal américain, moitié sur iPhone, moitié sur Android, 56 transmettent l'identifiant unique du téléphone, 47 donnent la localisation de l'utilisateur, et 5 livrent l'âge et le sexe du mobinaute sans qu'il se doute de rien.

L'application la plus "généreuse" serait TextPlus 4, un service permettant d'envoyer gratuitement SMS et MMS. Le logiciel enverrait l'identifiant unique du téléphone à huit régies publicitaires différentes, et la localisation, l'âge et le sexe de l'utilisateur à deux autres régies. Même chose pour Pandora (musique en streaming), Paper Toss (jeu), ou Grindr (site de rencontre pour homosexuels et bisexuels).

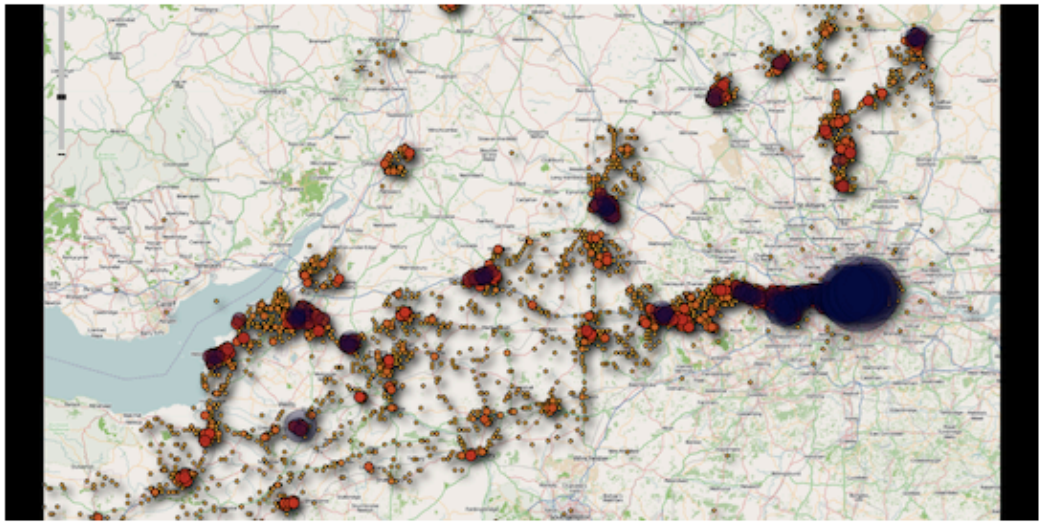
IMPOSSIBLE DE DÉSACTIVER LE TRAÇAGE

http://www.lemonde.fr/technologies/article/2010/12/20/ces-donnees-privees-que-les-applications-mobiles-transmettent-a-votre-insu_1455982_651865.html

Les iPhone collectent l'historique des déplacements des usagers

LEMONDE.FR | 21.04.11 | 10h51 • Mis à jour le 21.04.11 | 15h09

Abonnez-vous 15 € / mois 48 Partagez +1 0



Recommander Envoyer 4 402 personnes recommandent ça.

- http://www.lemc e-des-deplacements-de

Deux chercheurs ont révélé, mercredi 20 avril, que les terminaux mobiles d'Apple disposent d'un système permettant de suivre et de garder en mémoire les déplacements de leurs utilisateurs. Selon les experts en sécurité Alasdair Allan et Pete Warden, des indications comme la latitude et la longitude des usagers d'iPhone ou d'iPad 3G, associées à des informations horaires, sont consignées

Correlating Data

Tell-all telephone

deutsch | **english**

Green party politician Malte Spitz sued to have German telecoms giant Deutsche Telekom hand over six months of his phone data that he then made available to ZEIT ONLINE. We combined this geolocation data with information relating to his life as a politician, such as Twitter feeds, blog entries and websites, all of which is all freely available on the internet.

By pushing the play button, you will set off on a trip through Malte Spitz's life. The speed controller allows you to adjust how fast you travel, the pause button will let you stop at interesting points. In addition, a calendar at the bottom shows when he was in a particular location and can be used to jump to a specific time period. Each column corresponds to one day.

Friday, 4 September 2009

Final election campaign spurt: Spitz meets with regional Green party leaders and top candidates at the Jerusalem Church in Berlin's Kreuzberg district. (source: Twitter)

12 incoming calls
13 outgoing calls
total time: 0h 35min 38s

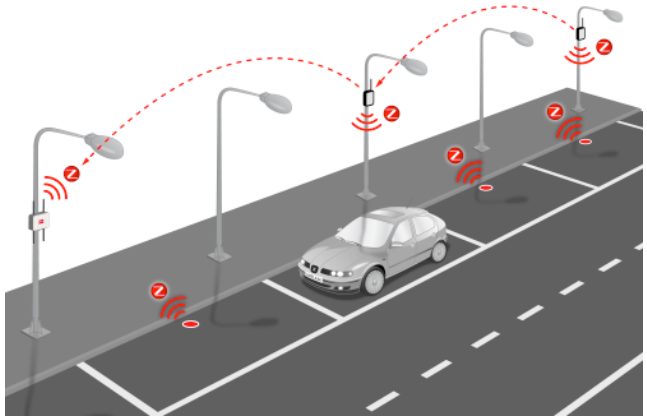
36 incoming messages
35 outgoing messages

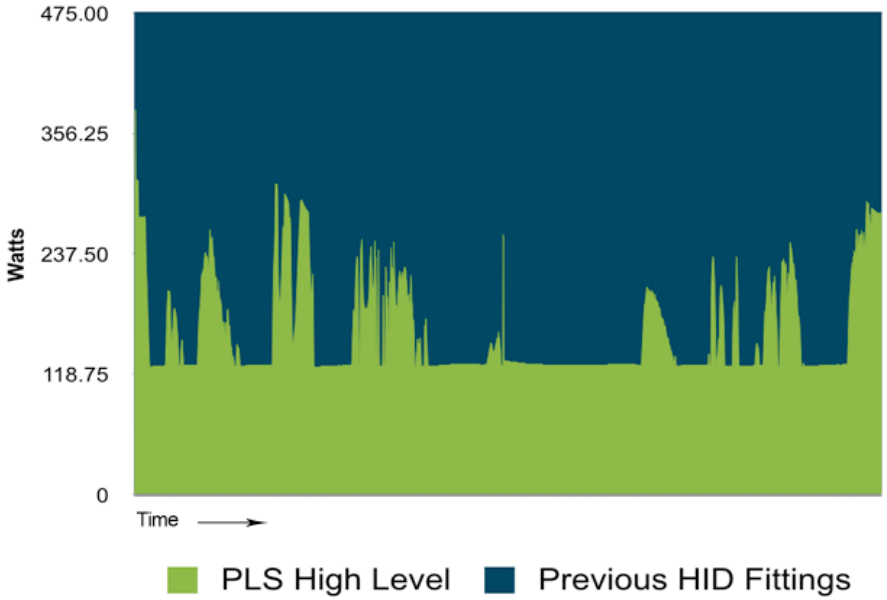
duration of internet connection:
16h 22min 57s

Correlating Data

- http://pleaserobme.com

Lampadaires intelligents mais indiscrets





Conclusion / Résumé

- Données: sources et stockage
- Traitement: analyse et enrichissement
- Analyse: diverses techniques et outils disponibles
- Exemples et services exploitant big data
- Aspects transverses