

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



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

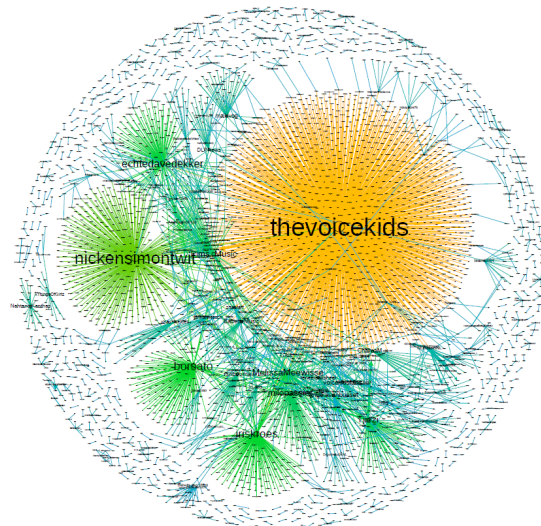
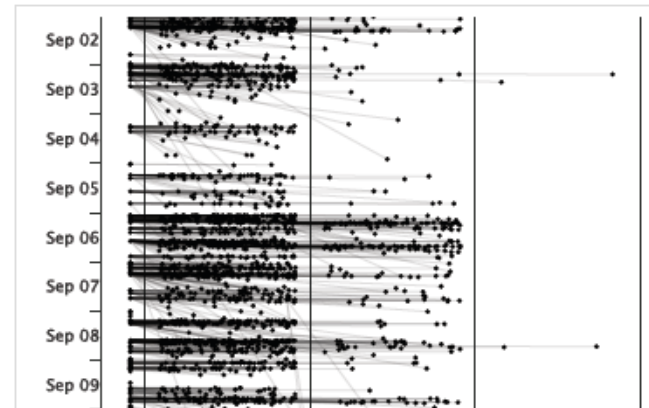
Données structurées

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

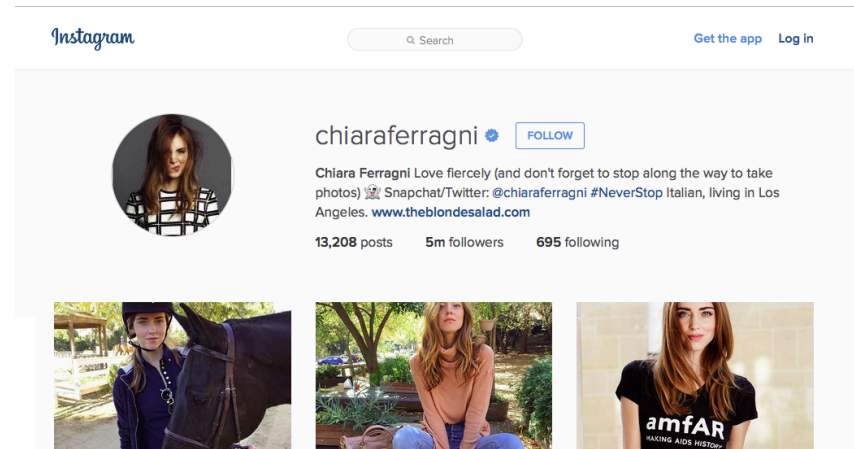
← Les enregistrements de la relation Personne



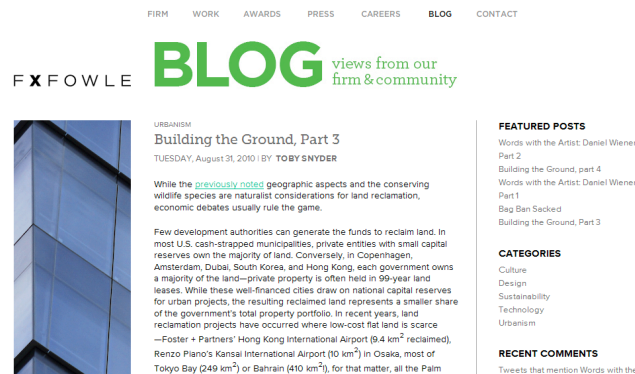
Données non structurées



A screenshot of a tweet from Barack Obama (@BarackObama) asking "How are you and your family celebrating this Independence Day?". The tweet has 1,340 retweets and 951 favorites. The interface includes a profile picture, a verified account badge, and interaction buttons for reply, retweet, favorite, and more.



A screenshot of the Instagram profile for Chiara Ferragni (@chiara ferragni). The profile shows a circular profile picture, a bio mentioning her location in Los Angeles and website, and statistics for 13,208 posts, 5m followers, and 695 following. Below the bio are three grid images: one of her with a horse, one of her sitting outdoors, and one of her wearing an amfAR t-shirt.



A screenshot of a blog post from Fx Fowle titled "Building the Ground, Part 3". The post is dated Tuesday, August 31, 2010, and is by Toby Snyder. The main text discusses urbanism, land reclamation, and the challenges of generating funds for land reclamation in cash-strapped municipalities. It mentions that in most U.S. municipalities, private entities own the majority of land, while in cities like Copenhagen, Amsterdam, Dubai, South Korea, and Hong Kong, the government owns a majority of the land. The post also notes that in recent years, land reclamation projects have occurred where low-cost flat land is scarce, citing Foster + Partners' Hong Kong International Airport (9.4 km² reclaimed), Renzo Piano's Kansai International Airport (10 km²) in Osaka, and Tokyo Bay (249 km²) or Bahrain (410 km²).

FEATURED POSTS

- Words with the Artist: Daniel Wiener, Part 2
- Building the Ground, part 4
- Words with the Artist: Daniel Wiener, Part 1
- Bag Ban Sacked
- Building the Ground, Part 3

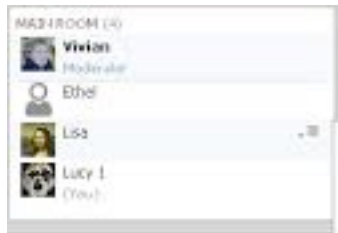
CATEGORIES

- Culture
- Design
- Sustainability
- Technology
- Urbanism

RECENT COMMENTS

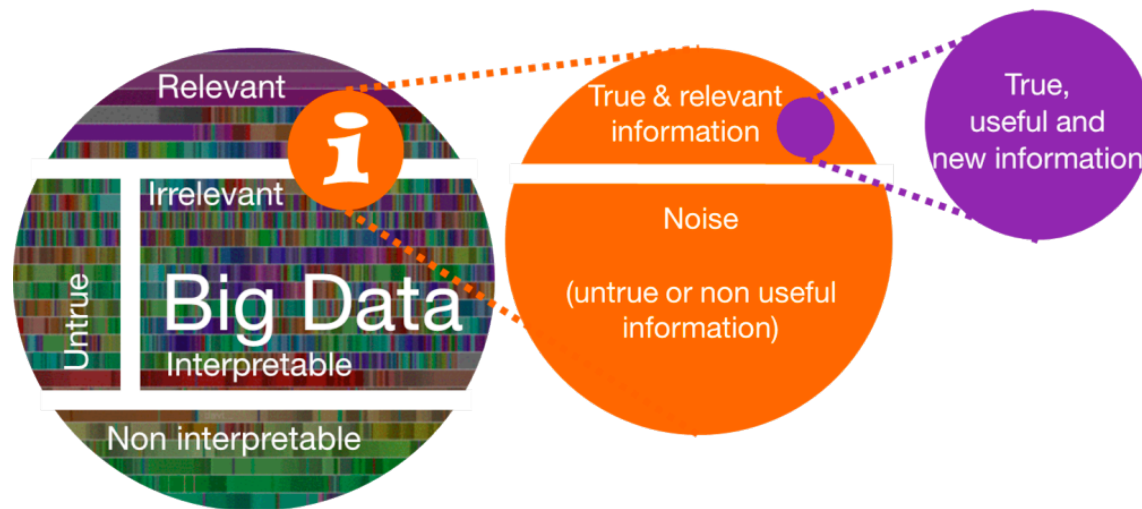
Tweets that mention Words with the

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

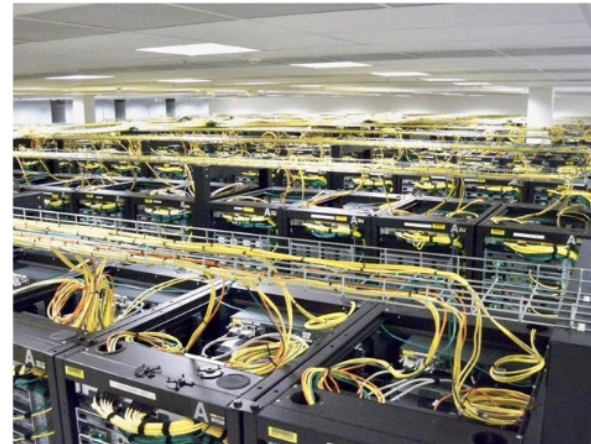
BD (classiques et propriétaires) (e.g. statistiques, archives
(numériques)

Capteurs – physiques, virtuels, sociaux

Données ouvertes - Open data

Crowdsourcing

BD / Serveurs / Cloud



Capteurs – physique

Intelligence ambiante – Internet of things



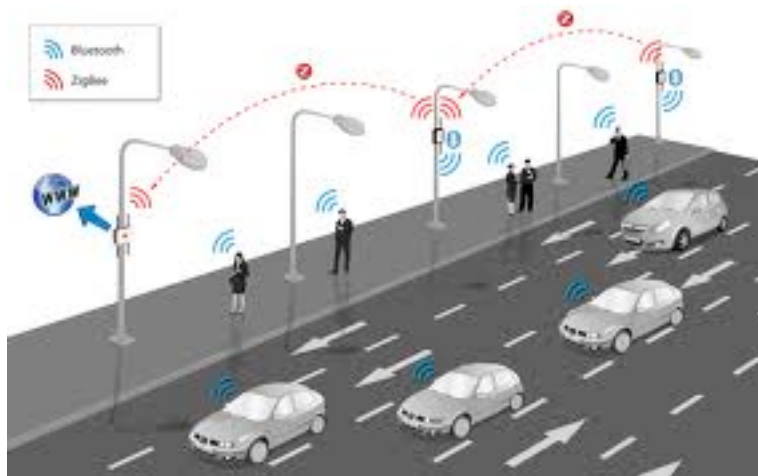
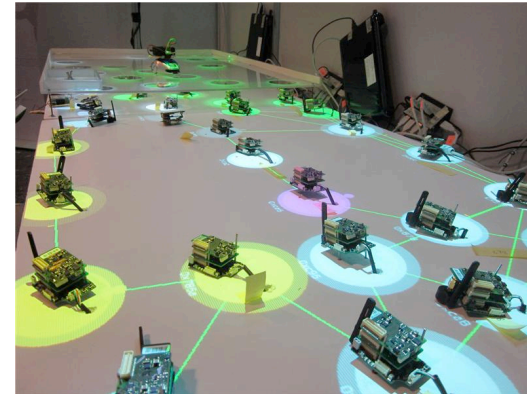
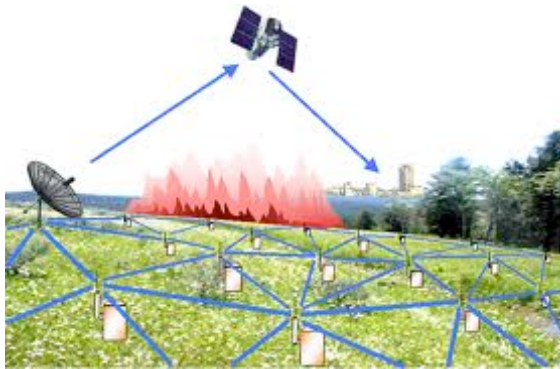
<http://vimeo.com/41363473>

Capteurs – physique

Internet of Things - Wearables



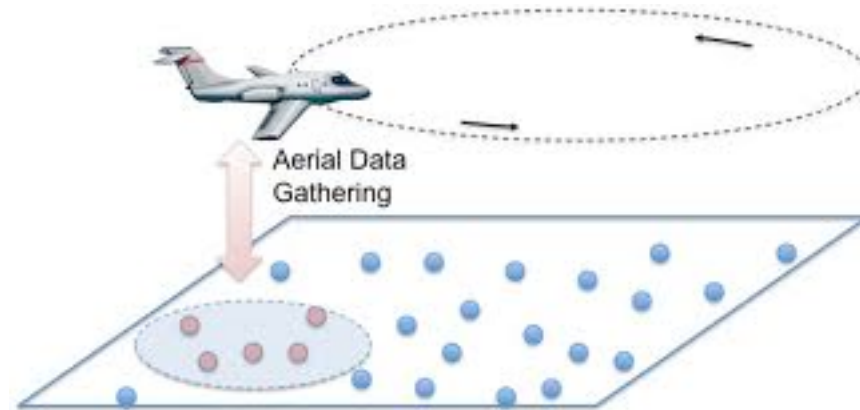
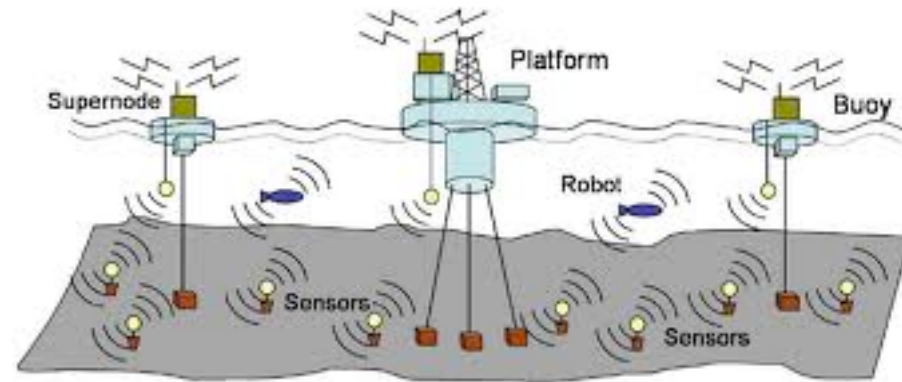
Capteurs physiques – environnement



Wireless ou non

Capteurs – physiques

Environnement



Terrestre
Sous-marin
Aérien

Captours – virtuels



- Jeremiah Owyang** @jowyang 21 Jan 10
Savvy Fishermen Fish where the Fish are. The distributed web strategy will involve widgets. Expect to see Widget advertising networks grow.
Expand
- Erica OGrady** @ericaogrady 21 Jan 10
@alsamko - I second that. In case you missed it she said: Social Media - Content isn't King, relationships are.
View conversation
- K.G. Schneider** @kgs 21 Jan 10
@bjjukeke I haven't had a job title my family has understood in about a decade!
View conversation
- Amy Gahrn** @agahrn 21 Jan 10
Just told my Poynter readers how to do news podcasts smart. Can't wait to hear their excuses: <http://tinyurl.com/2bpu25>
Expand
- Laura P Thomas** @lpt 22 Jan 10
me: "No, I can't stay home again today, have some face-to-face mrgs to make" gin: "can't you do those in Second Life and stay home again?"
Expand



Capteurs - sociaux



Médias sociaux et blogs



Open Data

Catégories plus

- Politique (1499)
- Espace et environnement (118)
- 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...

Choix Classe	Type	Nom	Len
<input type="checkbox"/>	IGISERVANTINTEACTOC	ORTHOPHOTOS 2008 (pmax 40 m)	
<input type="checkbox"/>	IGISERVANTINTEACTOC	ORTHOPHOTOS 2011 (pmax 40 m)	
<input type="checkbox"/>	IGISERVANTINTEACTOC	ND - ORTHOPHOTOS ANGLU 2012 (pmax 40 m)	

Crowdsourcing

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

The screenshot shows a FixMyStreet report interface. At the top, there is a navigation bar with the FixMyStreet logo, a 'Report a problem' button, and links for 'Your reports', 'All reports', 'Local alerts', and 'Help'. The report title is 'Green Traffic Light Out'. Below the title, it states 'Reported via iOS in the Traffic lights category by P Dixon at 18:44 today' and 'Sent to North Tyneside Borough Council 1 minute later'. A photograph of a street at night with a traffic light is included. The text describes the issue: 'Green traffic light out at pedestrian crossing on Park Lane near the junction with Lesbury Avenue. Traffic head faces traffic heading north'. There is a 'Provide an update' section with a note that updates are not sent to the council. On the right, a map shows the location with a yellow pin on Park Lane near the junction with Lesbury Avenue. The map includes street names like Park Lane, Lesbury Avenue, and Park Crescent, and landmarks like Shiremoor Centre and Emmons Court.

FixMyStreet Report a problem Your reports All reports Local alerts Help mySociety

Green Traffic Light Out

Reported via iOS in the Traffic lights category by P Dixon at 18:44 today
Sent to North Tyneside Borough Council 1 minute later

Green traffic light out at pedestrian crossing on Park Lane near the junction with Lesbury Avenue. Traffic head faces traffic heading north

Provide an update

Please note that updates are not sent to the council. Your information will only be used in accordance with our [privacy policy](#)

Shiremoor Centre
Emmons Court
Park Lane
Lesbury Avenue
Park Crescent
Park Grove
Park Avenue
Park Road
Lesdale Crescent
Glendale Road
Byewell Grove
Shiremoor

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

- Prédicatif
 - 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 connaître 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








- Ingénierie 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


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 Prevention**
 - CDC joins investigation as dengue fever cases in Hawaii tops 100**  CNN, 12 hours ago
 - Hawaii Officials Ask Public to Help Stop Dengue Fever Spread**  ABC News, 1 day ago
 - Health Officials Investigating Dengue Fever Outbreak in Hawaii**  Rapid News Network, 6 hours ago
- 5 **Food, Hepatitis A, Hepatitis**

Google Trends

Interest over time

● Search interest ● News articles



Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of

[Search lots of books](#) [Share](#) [Tweet](#) [Embed Chart](#)

Google Books Ngram Viewer



Google BigQuery Service

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



+ Briefcase

What's this?

1 Comment

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

- <http://www>

Visualisation

Réalité augmentée

Mashups

Représentations animées

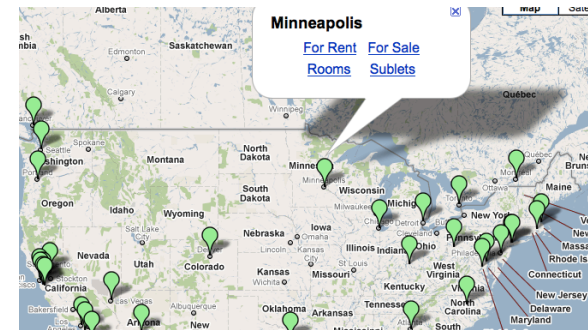
Réalité augmentée pour les journalistes



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



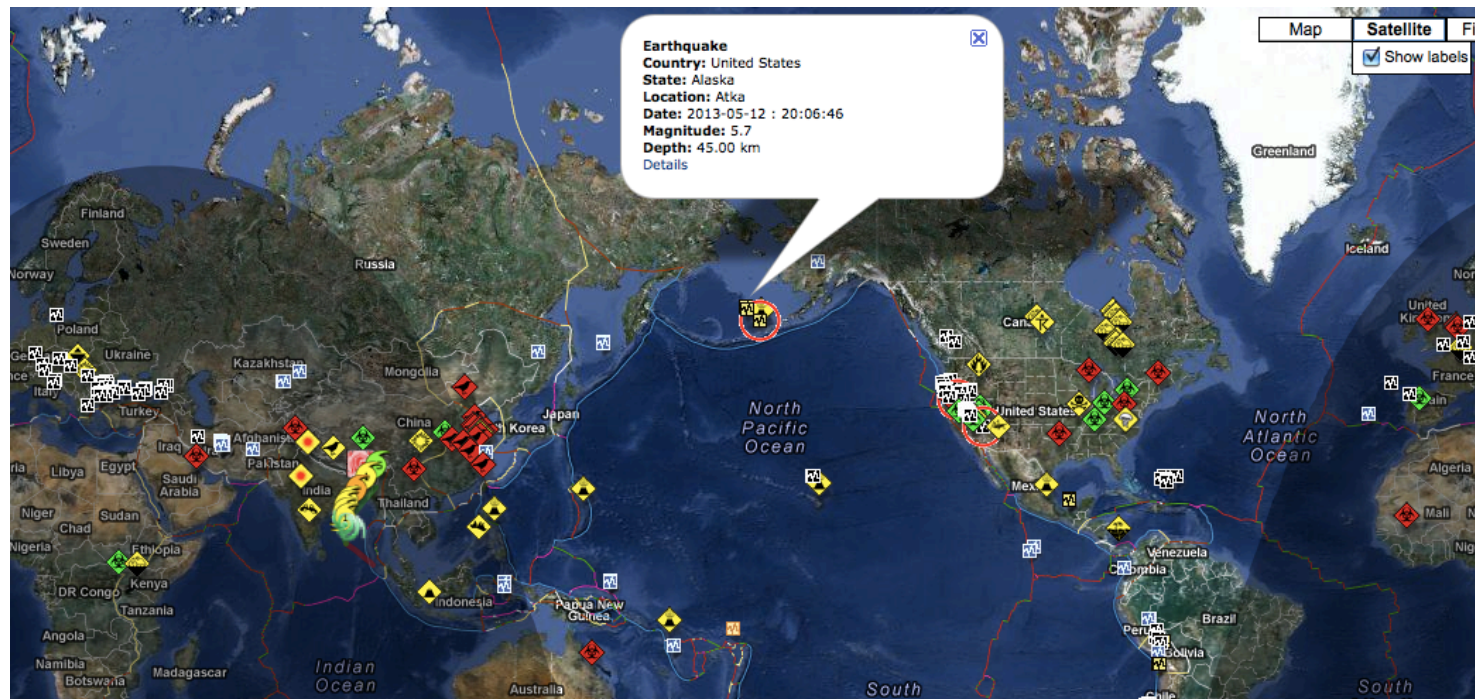
<http://bigbrowser.blog.lemonde.fr/2013/05/09/planisphere-le-monde-en-tweets/>

Real-time Tweeter Map



- <http://tweereal.com/>

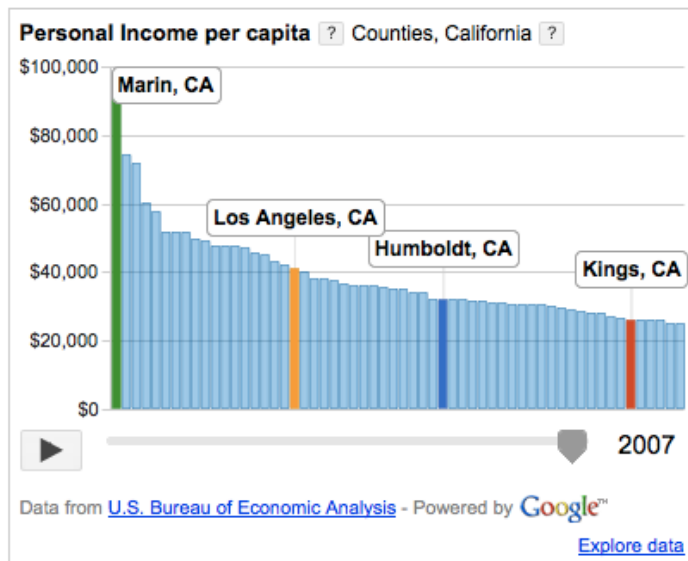
Senseurs virtuels – interrogeables sur une carte



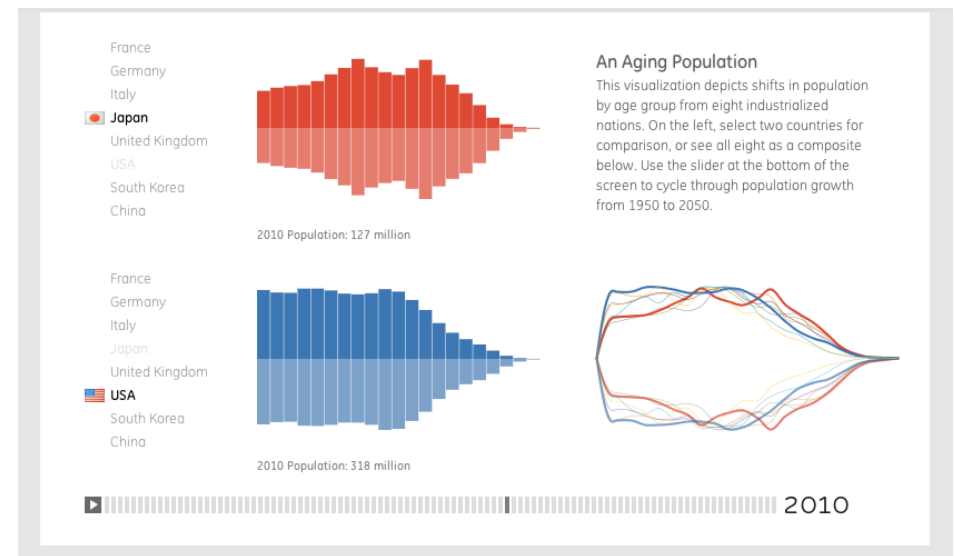
<http://hisz.rsoe.hu/alertmap/index2.php>

http://nathazmap.com/hazard_maps/interactive

DaaS - Examples



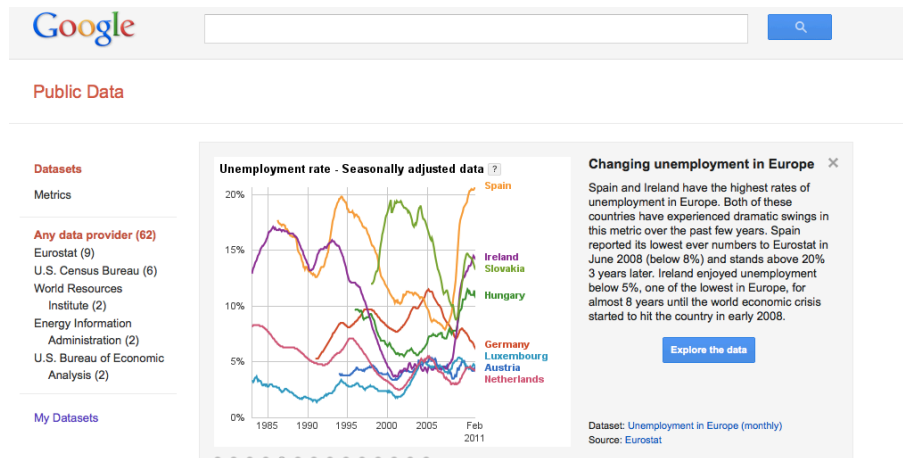
<http://radar.oreilly.com/2010/07/data-as-a-service.htm>



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

DaaS - Providers

Public Data Explorer (Google)



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

Exemples et applications

Jeux Sérieux

Science citoyenne, participative

Serious Games

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

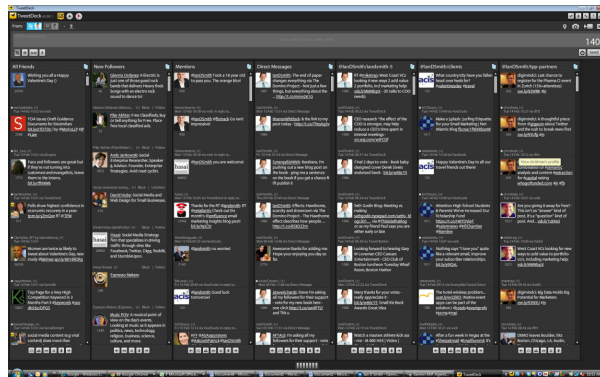


<http://www.rts.ch/info/suisse/6850695-la-rts-lance-tabula-rasa-le-jeu-web-ou-l-on-cree-sa-suisse-ideale.html>

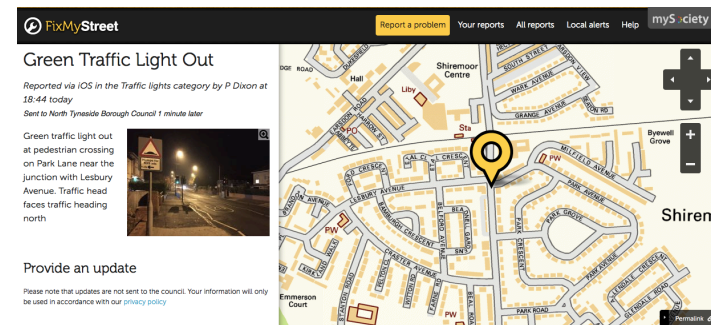
www.tabularasa.ch

Crowdsourcing / Science participative

Partager des informations



Citoyen est un “capteur”



Citoyen est un scientifique

Few have witnessed what you're about to see
Experience a privileged glimpse of the distant universe as observed by the SDSS, the Hubble Space Telescope, and UKIRT

Classify Galaxies
To understand how galaxies formed we need your help to classify them according to their shapes. If you're quick, you may even be the first person to see the galaxies you're asked to classify.

[Begin Classifying](#)

How Do Galaxies Form? History of Galaxy Zoo

Engagement démocratique

electionbuddy
1343114 votes cast in 8353 elections

Home Tour Customers Pricing & Signup Help Blog About

Easy, secure, anonymous online voting and elections

A quick, convenient, and cost-effective alternative to reduce volunteer hours and improve voter turnout ideal for associations, schools, unions or anyone

[TRY IT FOR FREE](#) [Feature Tour](#)

It's free to setup and test electionbuddy and it's **always free for less than 20 voters** — [test it for yourself](#)

Technology tools

Plateformes Web,
médias sociaux



Plateformes de crowdsourcing



Téléphones portables

Screenshot of the International Federation of Red Cross and Red Crescent Societies website showing the TERA (Trilogy Emergency Relief Application) and Beneficiary Communication page. The page includes a navigation menu, a search bar, and a main content area with a photo of a hand holding a mobile phone. The text on the page reads: "TERA (Trilogy Emergency Relief Application) and Beneficiary Communication". Below the photo, it says: "SMS messages are sent to communities across the country providing simple information on how to avoid cholera and other diseases." The author is listed as "Andrena Geffard".

Plateformes open source
Démocratie

Screenshot of the Loomio website homepage. The page features the Loomio logo, a navigation menu, and a main heading that reads: "The world needs a better way to make decisions together." Below the heading, it says: "Loomio is free and open source software for anyone, anywhere, to participate in decisions that affect them." The page also includes a globe icon and social media links.

Platforms for democracy and engagement

Votes / signatures électroniques Trouver parti (voting advice)



Vote préférentiel et opinions

Information, conversation, vote

LiquidFeedback News · Developers area · Donate · [Führ: ituu](#)

LiquidFeedback is an **open-source software**, powering internet platforms for **proposition development and decision making**

LiquidFeedback is an independent open source project published under MIT license by the Public Software Group of Berlin, Germany. The developers of LiquidFeedback have teamed up in the Association for Interactive Democracy to promote the use of electronic media for democratic processes.

LiquidFeedback is more than Liquid Democracy

Liquid Democracy Collective Moderation Fully Transparent Decision Process Preferential Voting

PARTIDO DE LA RED - DEMOCRACIA 05 PARTIDO DE LA RED INGRESAR

11 ABIERTOS 99+ CERRADOS

PRÓXIMOS A CERRAR

- Tarifa Social y Abonos para el Subte**
458 Participantes
- Protección de Nuestra Señora de Las Victorias
198 Participantes
- Basta de camiones con volquetes a cualquier hora
458 Participantes
- Extensión del horario del Subte
292 Participantes
- Una ley para los Museos Portefolios

Fecha de cierre desconocida

Tarifa Social y Abonos para el Subte

Despacho 3035 / 2013

Este proyecto apunta, por un lado, a la ampliación de la Tarifa Social del Subte y Premetro, destinada a usuarios en situación de "vulnerabilidad socioeconómica". Esta Tarifa equivale al 60% de la tarifa ordinaria.

Además, el proyecto establece la gratuidad del pasaje para alumnos de escuelas públicas, inicial, primaria y secundaria e incorpora 3 tipos de abonos del Subte: el "Abono por pago



Accueil

Consultation

Actualités

Évènements

À propos ▾



Inscription

Connexion

Les cookies assurent le bon fonctionnement de nos services. En utilisant ces derniers, vous acceptez l'utilisation des cookies.

[En savoir plus](#)

OK

La République numérique en actes

Écrivons ensemble la loi numérique

#Contribuer

© gouvernement.fr

Le projet de loi pour une République numérique

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



- ABOUT
- PROJECTS
- LABS
- BLOG**
- CHALLENGES
- PRIVACY
- PARTNERSHIPS
- CONTACT
- HOME

BLOG

COMBINING "BIG" AND "SMALL" DATA TO BUILD URBAN RESILIENCE IN JAKARTA

Giulio Quagotto Apr 9, 2014



Interview with Etienne Turpin and Tomas Holderness, directors of PetaJakarta.org

SUBSCRIBE TO OUR NEWSLETTER

GO

Crowd Sourcing

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

The screenshot displays the CitizenCyberScience website. At the top left is the logo for the CitizenCyberScience Centre, featuring a stylized figure holding a globe. The tagline "All for science, science for all" is centered above a navigation menu with links for Home, Cyberlab, Projects, Events, Blog, Follow, Contact, and About. Below this is a green banner for CrowdCrafting, with a search bar and a "Sign in" button. The main content area features a "crowdcrafting" section with a description: "Online assistance in performing tasks that require human cognition, knowledge or intelligence such as image classification, transcription, geocoding and more!". A list of bullet points follows: "Help advance research", "Everything is open and freely usable", and "Things computers can't do". To the right of the text is a grid of images, including a colorful 3D model, a close-up of a fingerprint, a green sign for "SPILL BIRD MICHIGAN 17-551-5030", a person's face, and a landscape. The bottom of the page has a dark footer with the text "CrowdCrafting.org" and "An open platform for volunteer thinking", flanked by navigation arrows.

<http://www.citizencyberscience.net/>

Aspects Transverses

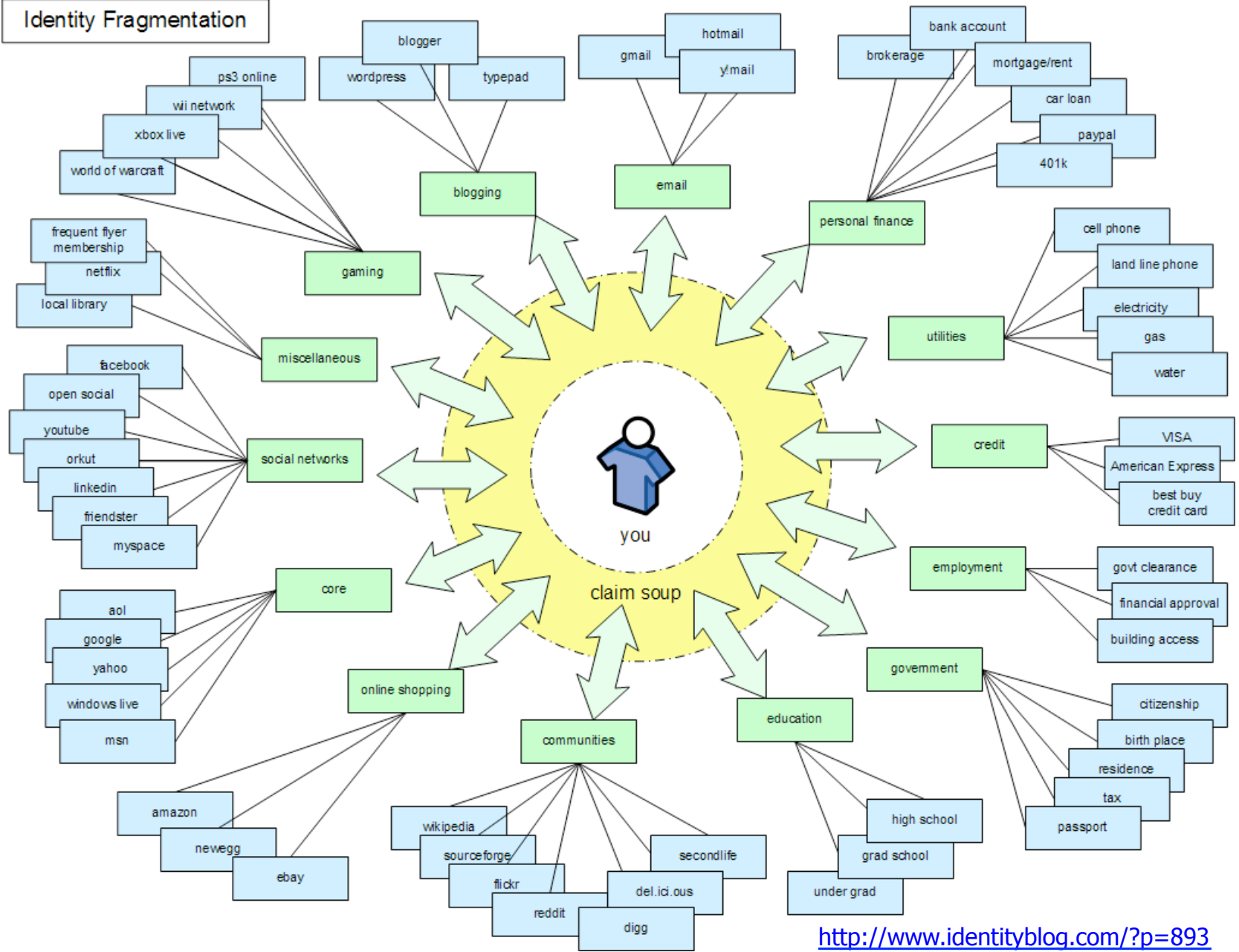
Identité numérique

Réputation en ligne

Protection de la sphère privée

Anonymisation vs Re-identification

Identity Fragmentation



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

Mobile Phones

lemonde.fr
Technologies

Abonnez-vous au journal **le Monde** : 17€/mois

Recherchez sur Le Monde.fr
Recevez les newsletters gratuites

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 Webdocs

A la Une > Technologies

Ces données privées que les applications mobiles transmettent à votre insu

LEMONDE.FR | 20.12.10 | 18h00 • Mis à jour le 20.12.10 | 20h09

ÉDITION ABONNÉS
6 € par mois

Partagez Facebook



Recommander 1054 recommandations. 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-privées-que-les-applications-mobiles-transmettent-a-votre-insu_1455982_651865.html

Les iPhone collectent l'historique des déplacements des usagers

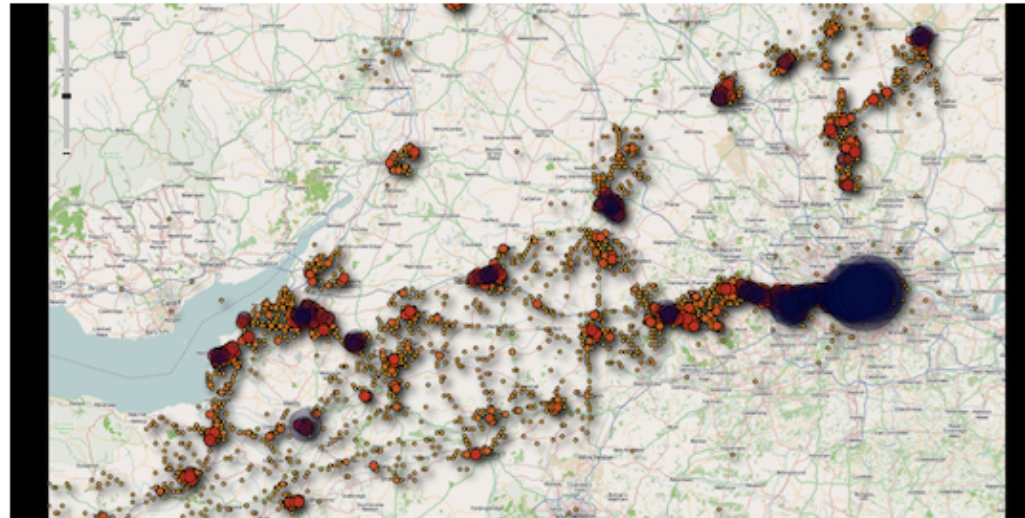
LEMONDE.FR | 21.04.11 | 10h51 • Mis à jour le 21.04.11 | 15h09

Abonnez-vous
15 € / mois

48



Partagez



Recommander

Envoyer

4 402 personnes recommandent ça.

- <http://www.lemc.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

[e-des-](#)

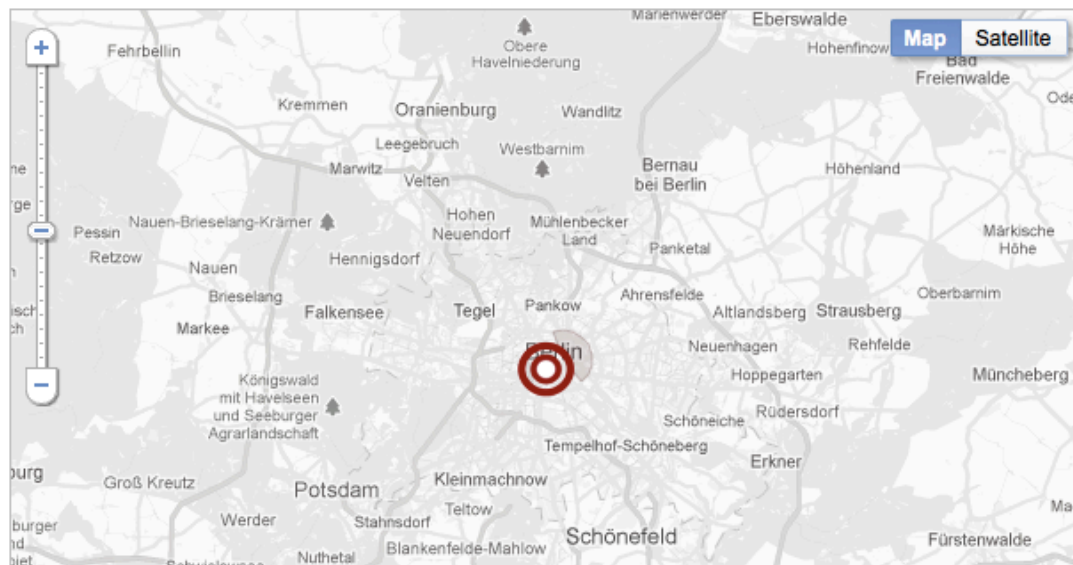
Correlating Data

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

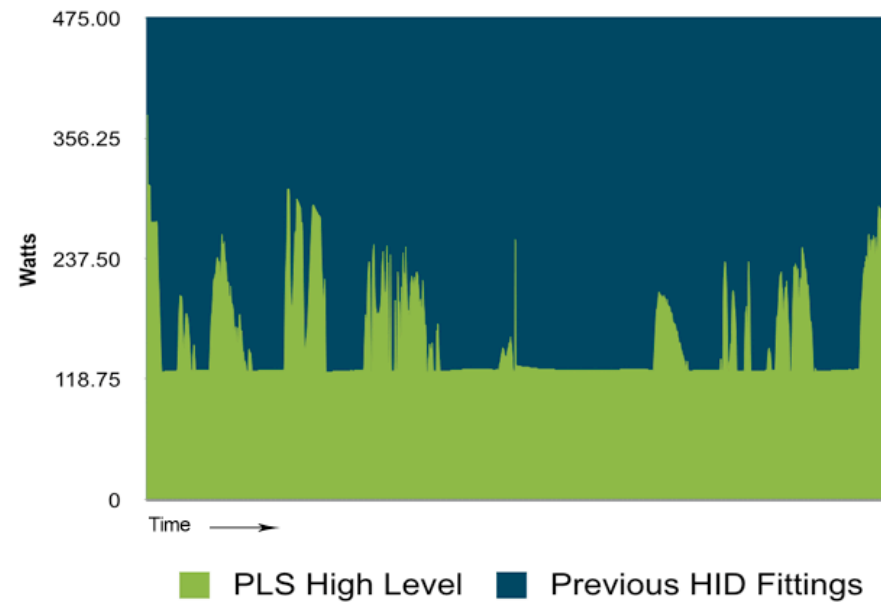
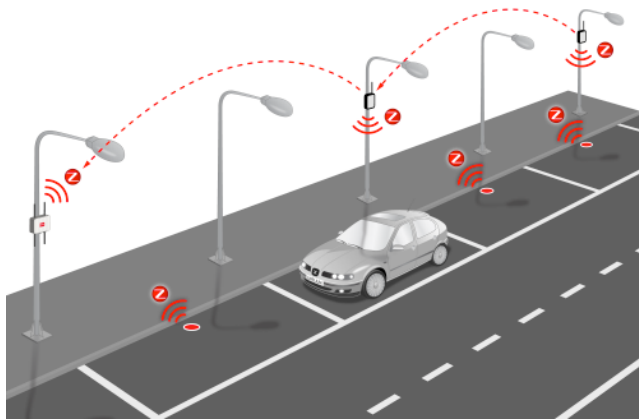
-  Final election campaign spurt: Spitz meets with regional Green party leaders and top candidates at the Jerusalem Church in Berlin's Kreuzberg district. (source: [Twitter](#))
-  12 incoming calls
13 outgoing calls
total time: 0h 35min 38s
-  36 incoming messages
35 outgoing messages
-  duration of 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