

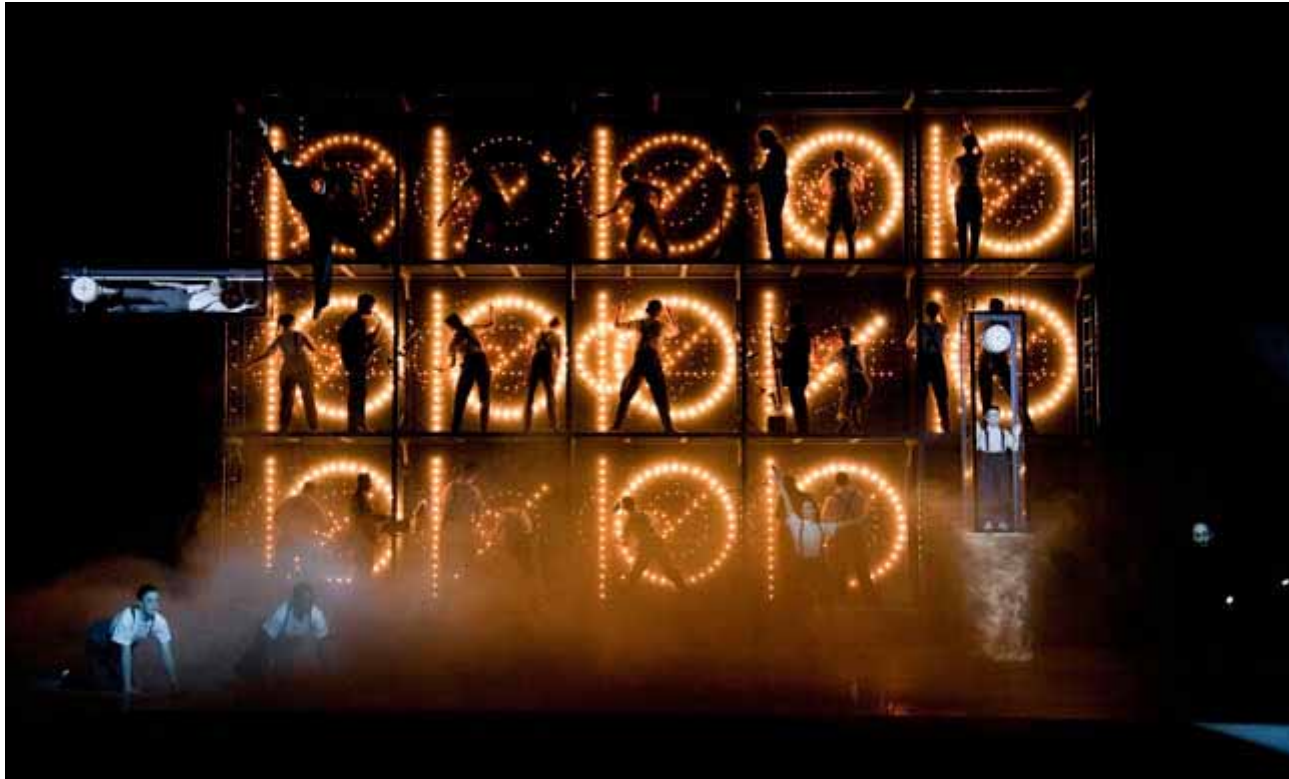
BIG DATA et sécurité

3 Paradoxes



Philippe WOLF

8es rencontres de l'ARCSI, Vendredi 13 juin 2014



« La photographie, c'est notre exorcisme. La société primitive avait ses masques, la société bourgeoise ses miroirs, nous avons nos images. »

Jean Baudrillard, Car l'illusion ne s'oppose pas à la réalité ..., 1998.

Avertissements



**Toute fonction de sécurité est à usage dual.
Elle servira aussi bien le criminel que l'honnête homme.**

Définition

- Le terme *big data* fait référence à des ensembles de données dont **la taille dépasse la capacité des logiciels usuels** pour collecter, gérer et traiter les données dans un temps raisonnable.
- Les traitements de masse impliquent **une nouvelle approche de la donnée** :
 - collecter et utiliser beaucoup de données plutôt que de se contenter d'échantillons comme l'ont fait des générations de statisticiens ;
 - accepter de traiter des **données imparfaites ou mal organisées**, une part d'inexactitude peut en général être tolérée car dans de nombreux cas, il est plus avantageux d'avoir bien plus de données qu'un plus petit nombre de données très exactes ou finement sélectionnées afin d'être représentatives ;
 - accepter de renoncer à rechercher des causalités au profit de la **recherche de corrélations, de motifs** qui peuvent aider à prédire le futur ;
 - le *big data* aide à répondre à la question du **quoi** et pas à celle du **comment**, ce qui est souvent suffisant.

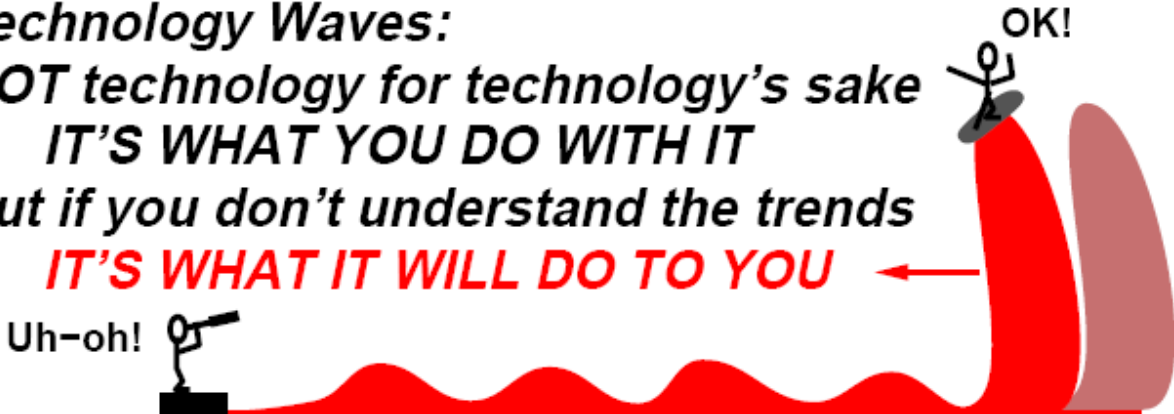
Archéologie du Big Data

sgi

Big Data ...
and the Next Wave of *InfraStress*

John R. Mashey
Chief Scientist, SGI

Technology Waves:
NOT technology for technology's sake
IT'S WHAT YOU DO WITH IT
But if you don't understand the trends
IT'S WHAT IT WILL DO TO YOU ←



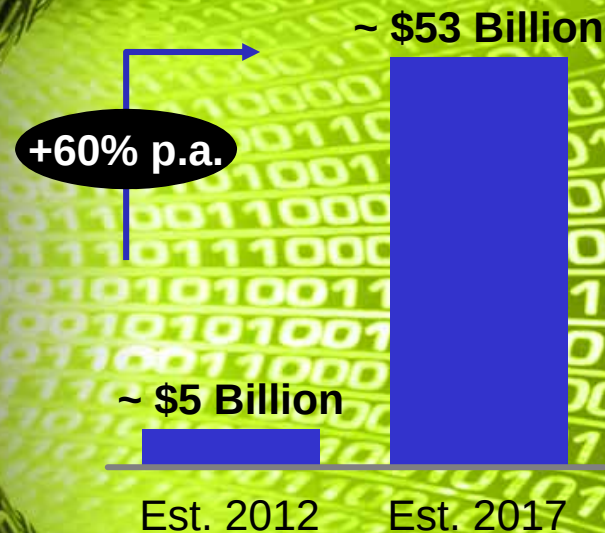
Big Data

Source : McKinsey

It is estimated that
4 exabytes
of unique information will
be generated this year

That is more than the
previous **5,000 years**

And this will grow
65 x by 2020



Quelques chiffres

Multiples de l'octet : préfixes SI et mésusages			
Nom	Symbole	Valeur	Mésusage
kilooctet	ko	10^3	2^{10}
mégaoctet	Mo	10^6	2^{20}
gigaoctet	Go	10^9	2^{30}
téraoctet	To	10^{12}	2^{40}
pétaoctet	Po	10^{15}	
exaoctet	Eo	10^{18}	
zettaoctet	Zo	10^{21}	
yottaoctet	Yo	10^{24}	

Multiples de l'octet : préfixes binaires		
Nom	Symbole	Valeur
kibioctet	Kio	2^{10}
mébioctet	Mio	2^{20}
gibioctet	Gio	2^{30}
tébioctet	Tio	2^{40}
pébioctet	Pio	2^{50}
exbioctet	Eio	2^{60}
zébioctet	Zio	2^{70}
yobioctet	Yio	2^{80}

1 099 511 627 776

Espérance vie	$2,5 \cdot 10^9$
Atomes dans l'univers	10^{79}
clés AES-256	$1,15 \cdot 10^{77}$
biclés RSA-2048	$2,5 \cdot 10^{305}$
Gogol	10^{100}

10^{120} octets	capacité mémoire de l'univers visible d'après Seth Lloyd
-------------------	--

La Jaune et la Rouge



REVUE MENSUELLE DE L'ASSOCIATION DES ANCIENS ÉLÈVES ET DIPLÔMÉS DE L'ÉCOLE

polytechnique

MARS 2014 - 8 € - N° 693



Dossier Jaune et Rouge

PAR FRANÇOIS BOURDONCLE (84)



co-chef de file
de la filière
Big Data française

Enjeux économiques et citoyens
d'une filière **Big Data** européenne

PAR PAUL HERMELIN (72)



P-DG de Capgemini,
cochef de file du plan
big data, dans le cadre
de l'initiative
La Nouvelle France
industrielle

Une opportunité pour la **France**
dans l'économie de l'**innovation**

PAR JEAN-PIERRE DARDAYROL (72)



ingénieur général
des Mines au Conseil
général de l'économie

Phénomène de **mode**
ou nouvelle **frontière** ?

PAR DIDIER KRAINIC



directeur général
d'IDC France

Un nouvel **eldorado** numérique

PAR DIDIER LOMBARD (62)



président du conseil
de surveillance de ST
Microelectronics

Une **révolution** majeure
pour la France et l'**Europe**

PAR CLAIRE LEVALLOIS-BARTH



maître de conférences
en droit
à Télécom ParisTech

L'**utilisation** raisonnée
des données **personnelles**

Les **trois** paradoxes
de la **sécurité**

PAR PHILIPPE WOLF (78)



ingénieur général
de l'armement, ANSSI

PAR STÉPHAN CLEMENÇON



professeur à Télécom
ParisTech, Institut
Mines-Télécom

Une **formation** spécifique
et **pluridisciplinaire**

LE MARCHÉ DES *BIG DATA*

Marché big data (M€)	2012	2013	2014	2015	2016	2017
Serveurs	10	13	16	19	22	25
Stockage	9	15	32	50	69	89
Équipements réseaux	3	4	6	8	11	13
Cloud	4	6	10	15	23	31
TOTAL INFRASTRUCTURES	26	38	64	93	125	158
Logiciels de gestion des données	15	20	26	34	44	58
Logiciels analytiques et d'exploration	20	24	29	37	48	63
Logiciels applicatifs	5	6	8	9	12	15
Total logiciels	40	50	63	80	104	136
Services	86	102	16	141	166	199
TOTAL	152	189	242	314	395	494

Source : IDC France.

Trois exemples d'effets escomptés

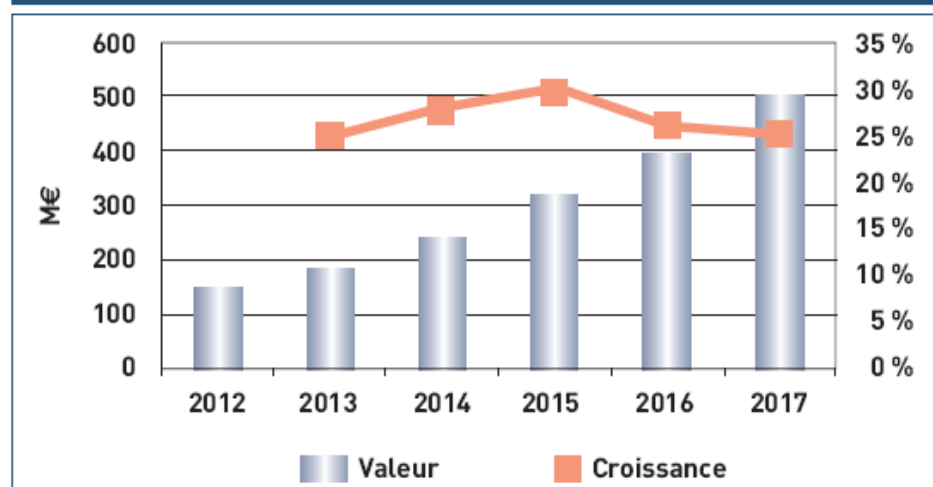
Finance : rationalisation du *post-trading* (*) permettant d'obtenir unicité des données, vitesse de *reporting* réglementaire; baisse du coût par événement.

Distribution : gestion des assortiments; réévaluation de la stratégie de *sourcing*; optimisation de la fidélité.

Réseaux utilitaires : multiplication de la fréquence de collecte; nouveaux critères de segmentation.

(*) Beaucoup de transactions se font rapidement de *trader à trader*, en temps réel, souvent par téléphone ou par échange de messages; une fois que la transaction est faite, il s'agit de la clarifier, d'en préciser rigoureusement les termes, d'en fixer le montant exact en fonction des cours, de vérifier sa faisabilité réglementaire, etc.

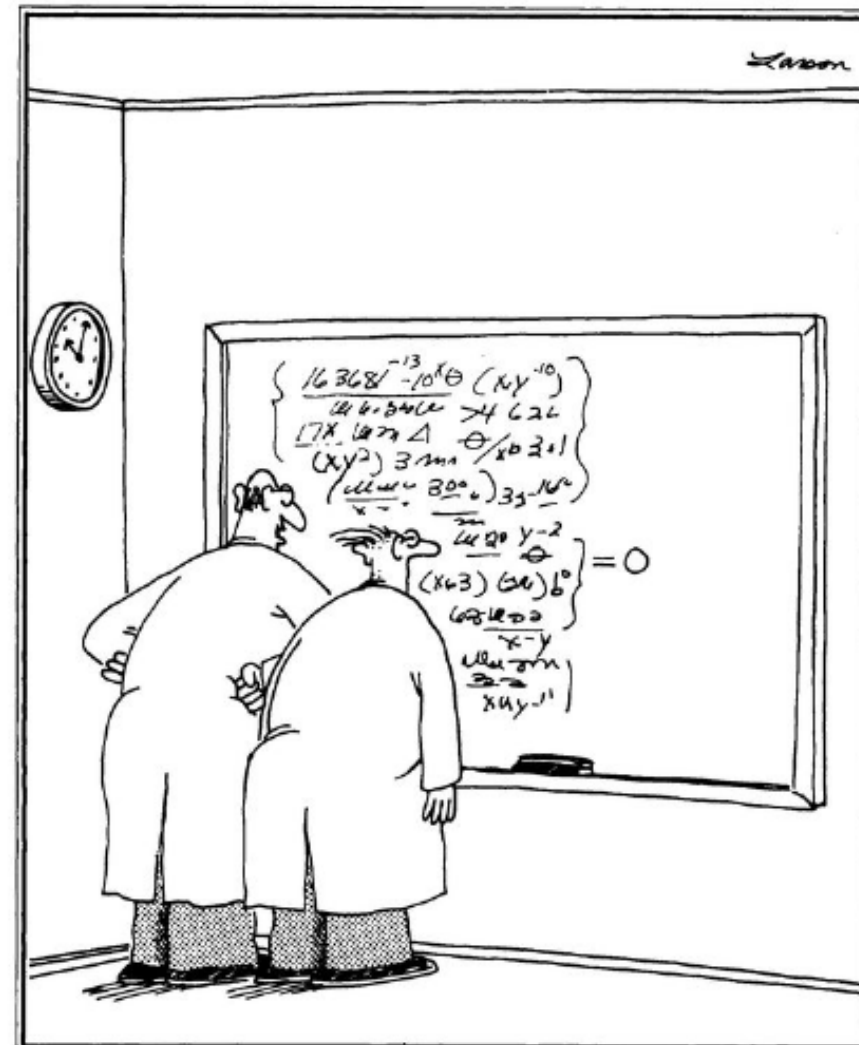
ÉVOLUTION DU MARCHÉ



Et l'homme ?

- Un humain absorbera au plus 40 pétaoctets (10^{15}) dans sa vie, à rapporter aux 200 yottaoctets (10^{24}) manipulés par Internet sur 60 ans.

10^{23} à 10^{29} inst/s	Puissance de tous les cerveaux humains
10^{30} inst/s	Puissance de calcul cumulée des systèmes génétiques de tous les êtres humains
10^{34} inst/s	Puissance de calcul cumulée des systèmes génétiques de tous les êtres vivant sur terre
10^{45} inst/s	Puissance de l'ordinateur de 1 kilogramme le plus puissant possible d'après un résultat de mécanique quantique de Margolus et Levitin
10^{115} instructions	En considérant que chaque mouvement dans l'Univers est équivalent à un calcul, S. Lloyd a évalué que l'Univers depuis sa naissance aurait produit un calcul d'environ 10^{115} instructions.



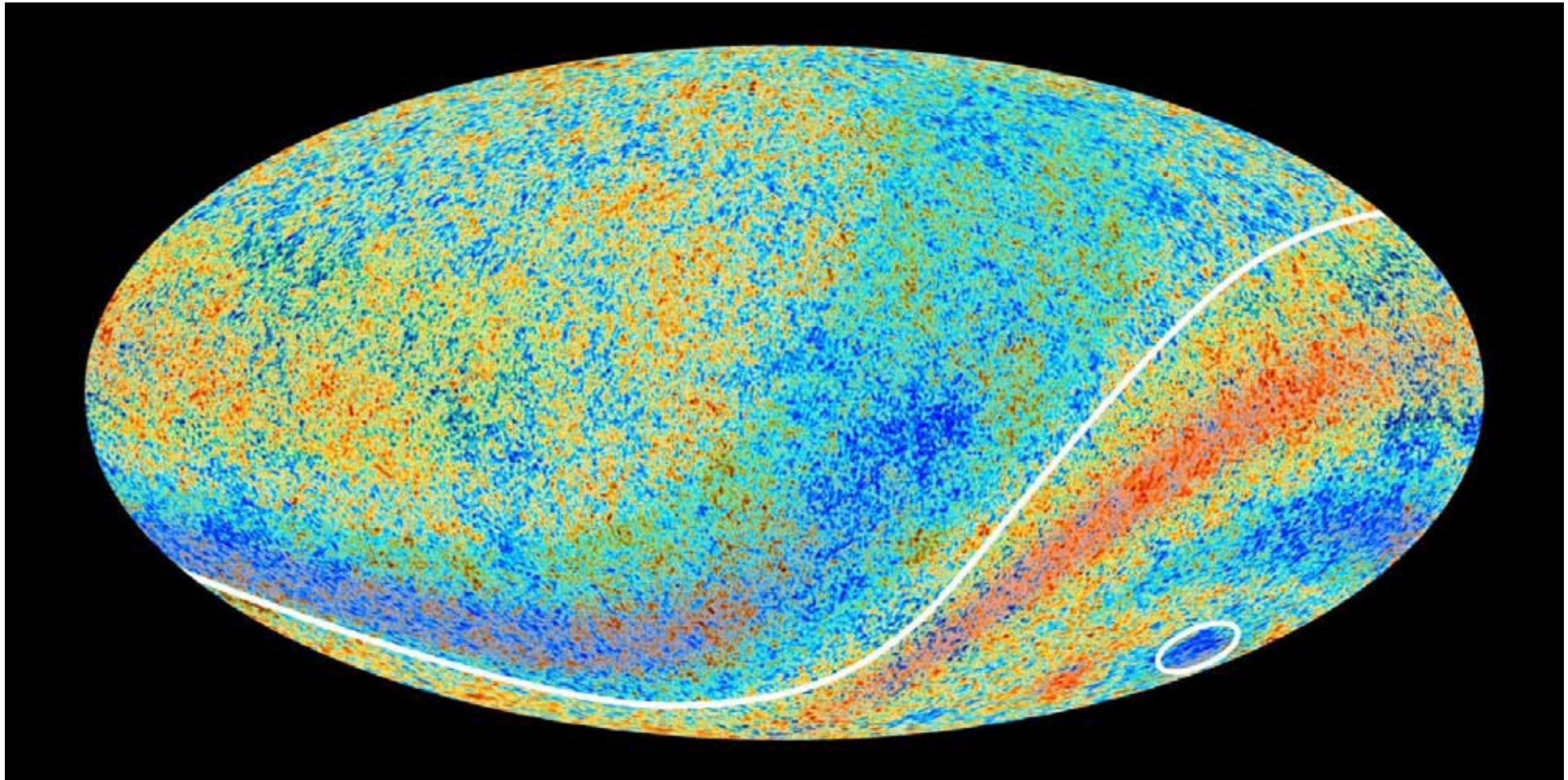
"No doubt about it, Ellington—we've mathematically expressed the purpose of the universe. Gad, how I love the thrill of scientific discovery!"

Demain : alimentation du Big Data par les objets connectés



Exemples : champ scientifique

- **Planck Microwave Background Radiation: Seeing the Big Bang**



Logarithme discret

- 33,5 Gb
- 46 ans/cpu

Subject: Discrete Logarithms in $GF(2^{9234})$
From: Jens Zumbrägel <[log in to unmask]>
Reply-To: Number Theory List <[log in to unmask]>
Date: Fri, 31 Jan 2014 07:59:39 -0600
Content-Type: text/plain
Parts/Attachments: text/plain (240 lines)

Dear Number Theorists,

We are pleased to announce a new record for the computation of discrete logarithms in finite fields. In particular, we were able to compute discrete logarithms in $GF(2^{9234})$ using about 400'000 core hours. To our knowledge the previous record was announced on 21 May 2013 in (a multiplicative subgroup of) the field $GF((2^{24})^{257})$ of 6168 bits [8].

[...]

The running time (in core hours) is as follows:

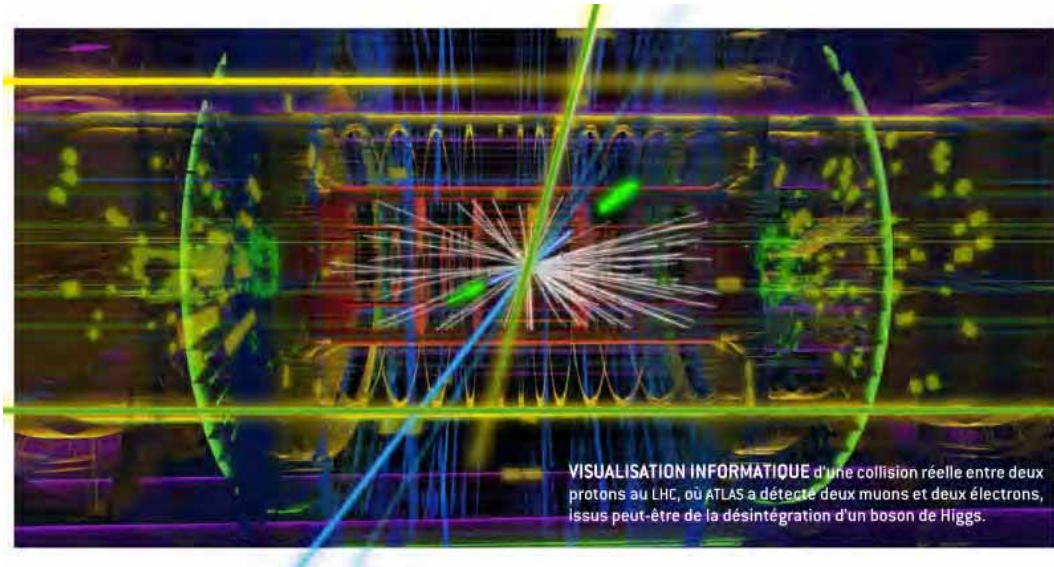
- relation generation 640 h (AMD: 6128 Opteron 2.0 GHz)
- linear algebra 258'048 h (Intel: Ivy Bridge 2.4 GHz)
- classical descent 134'889 h (Intel)
- Grobner basis descent 3'832 h (AMD)
- Pollard's rho 13 h (AMD)

totalling in 397'422 core hours.

Robert Granger*, Thorsten Kleinjung*, Jens Zumbrägel*

Antoine Joux, Prix Gödel 2013

Boson de Higgs



- **LHC de Genève (7 à 8 TeV)**
 - Collisions de 10^{11} protons à la seconde
 - pré-tri de 200 événements intéressants/s
 - 2 Go de données à la seconde -> Plusieurs dizaines de petaoctets par an
 - 1 boson observable toutes les 10^{13} collisions en théorie
 - quelques centaines observés réellement
 - Grid de 200 centres de calculs. En 2012 :
 - 500 000 ordinateurs + 200 petaoctets de stockage sur disques.

Génomes et Transcriptomes

- Séquence d'un génome humain
 - 3,4 milliards de paires de bases = 1 Goctet
- 3 centres publics de stockage
 - 18 petaoctets
 - Saturés!

The screenshot displays the Genamics Expression software interface. The main window shows a DNA sequence with various annotations. A circular map on the left indicates the location of a promoter and a misc feature. The right panel shows a sequence analysis window with a table of results.

Genamics Expression

File Edit Sequence Tools Internet Window Help

Calc Digest ORF Map Finder Primer RevTrans Arrange Back Forward

DNA Document

Protein5 Protein6 Protein7

cattgcgcattcaggctgcgaactgt
tgggaaggcgatcggtgcggcct
cttcgtattacgccagctggcgaaa
gggggatgtgctgcaaggcgattaag
ttgggtaacgccagggtttccagtc
acgacgttgtaaaacgacggccagt
gaattga<promoter "SP6 prom
oter">attaggtgacactata</pro
moter>gaagagctat<misc featu

Interactive Sequence

rep origin 1 promoter misc feature

S 2 DNA8 (4109) CC

Internet Tool Window

Enter Nucleotide

for [] Go Clear

☐ Limits Index History Clipboard

Display GenBank Show: 20

Add to Clipboard Save 1-1 items of 1

☐ 1 : [U12390](#) . Cloning vector pSp...[gi:531828]

LOCUS	PSPORT1	4109 bp	DNA
DEFINITION	Cloning vector pSport1, complete cds.		
ACCESSION	U12390		
NID	g531828		
VERSION	U12390.1 GI:531828		

Hint: You can annotate the sequence with the '<' key

Exemples : champ du renseignement



MYSTIC/SOMALGET



- **Bahamas: 307 451 habitants**
 - 100 millions d'appels par jour
 - 3 milliards de métadonnées par mois
 - Mystic est sous-traité à General Dynamics Advanced Information Systems

(S//SI//REL) SOMALGET collection systems ***forward full-take metadata in real time and buffer full-take audio for nominally 30 days.***³ It makes possible the selection of audio content against the buffered data after the fact, in near real-time, or up to 30 days later. This ability is dubbed "*retrospective retrieval*." The power of retrospective retrieval in facilitating target development or discovery lies in its ability to permit the analyst to selectively retrieve audio content and immediately validate his/her tentative analytic conclusions derived from metadata.

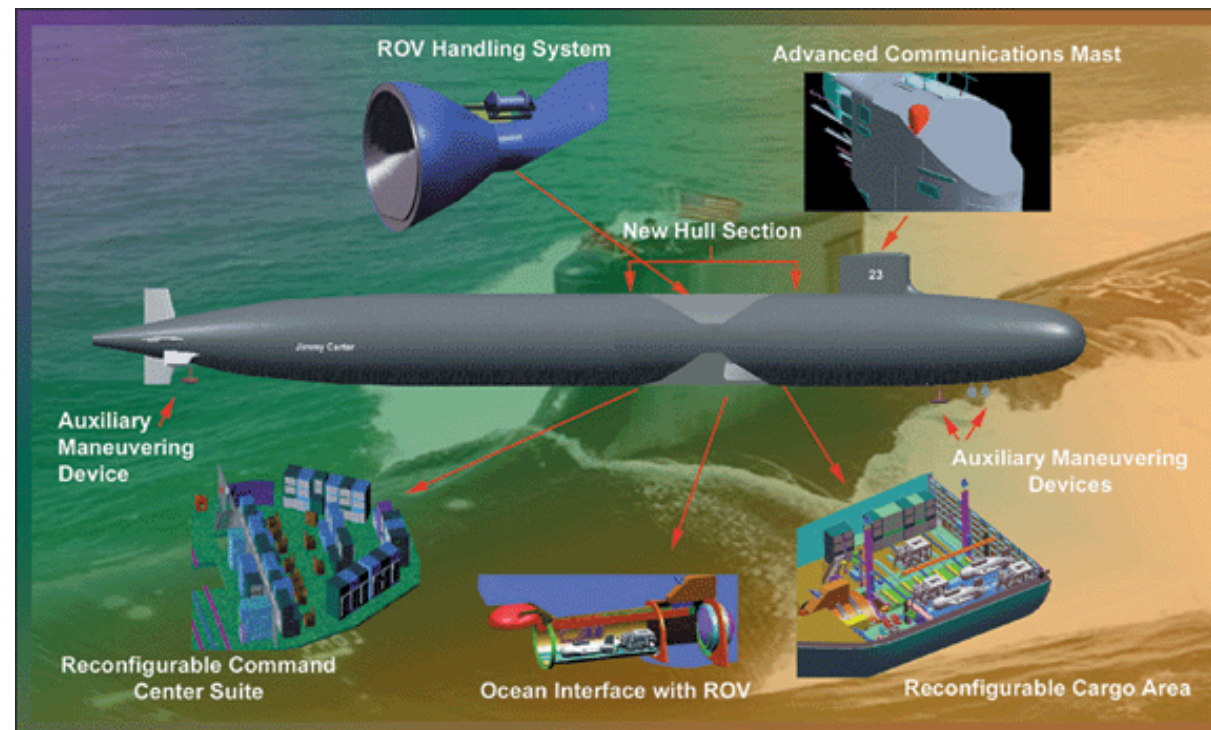
(S//SI//REL) SOMALGET collection systems ***forward full-take metadata in real time and buffer full-take audio for nominally 30 days.***³ It makes possible the selection of audio content against the buffered data after the fact, in near real-time, or up to 30 days later. This ability is dubbed "*retrospective retrieval*." The power of retrospective retrieval in facilitating target development or discovery lies in its ability to permit the analyst to selectively retrieve audio content and immediately validate his/her tentative analytic conclusions derived from metadata.

NSA Utah Data Center

- En juillet 2013, sur la base des plans de la structure, le magazine Forbes estimait la capacité de stockage entre 3 et 12 exabytes (milliards de gigabytes), à l'aide de 10 000 racks de serveurs.



DANCING OASIS (1)



DANCINGOASIS (2)

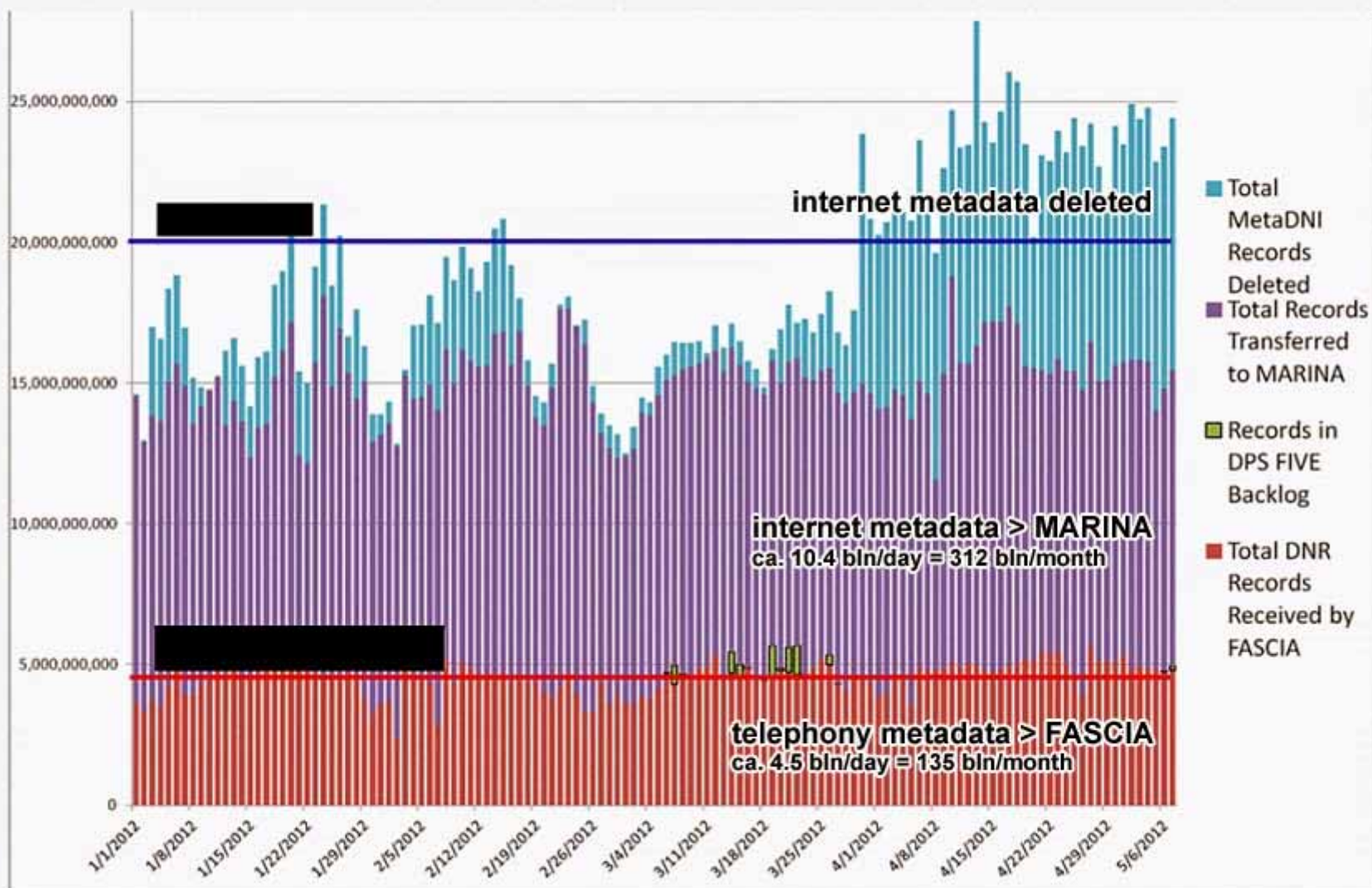
- 57 milliards d'enregistrements par mois, 684 milliards par an.
 - 27 milliards de données au Pakistan, 24 milliards en Afghanistan, 15 milliards en Iran et 13 milliards en Jordanie
- Tri par le programme SCISSOR
 - FO = 25 petaoctets/jour : entre 3 et 6 petaoctets/j traités par la NSA
 - Filtrés dans 2 bases de données: une de 43 gigaoctets une de 132 gigaoctets par jour (= 50 Teraoctets par an)

(S//SI//REL TO USA, FVEY) SHELLTRUMPET Processes it's One Trillionth Metadata Record

By NAME REDACTED on 2012-12-31 0738

(S//SI//REL TO USA, FVEY) On December 21, 2012 SHELLTRUMPET processed its One Trillionth metadata record. SHELLTRUMPET began as a near-real-time metadata analyzer on Dec 8, 2007 for a CLASSIC collection system. In its five year history, numerous other systems from across the Agency have come to use SHELLTRUMPET's processing capabilities for performance monitoring, direct E-Mail tip alerting, TRAFFICTHIEF tipping, and Real-Time Regional Gateway (RTRG) filtering and ingest. Though it took five years to get to the one trillion mark, almost half of this volume was processed in this calendar year, and half of that volume was from SSO's DANCINGOASIS. SHELLTRUMPET is currently processing Two Billion call events/day from select SSO (Ram-M, OAKSTAR, MYSTIC and NCSC enabled systems), MUSKETEER, and Second Party systems. We will be expanding its reach into other SSO systems over the course of 2013. The Trillion records processed have resulted in over 35 Million tips to TRAFFICTHIEF.

Example of Current Volumes and Limits

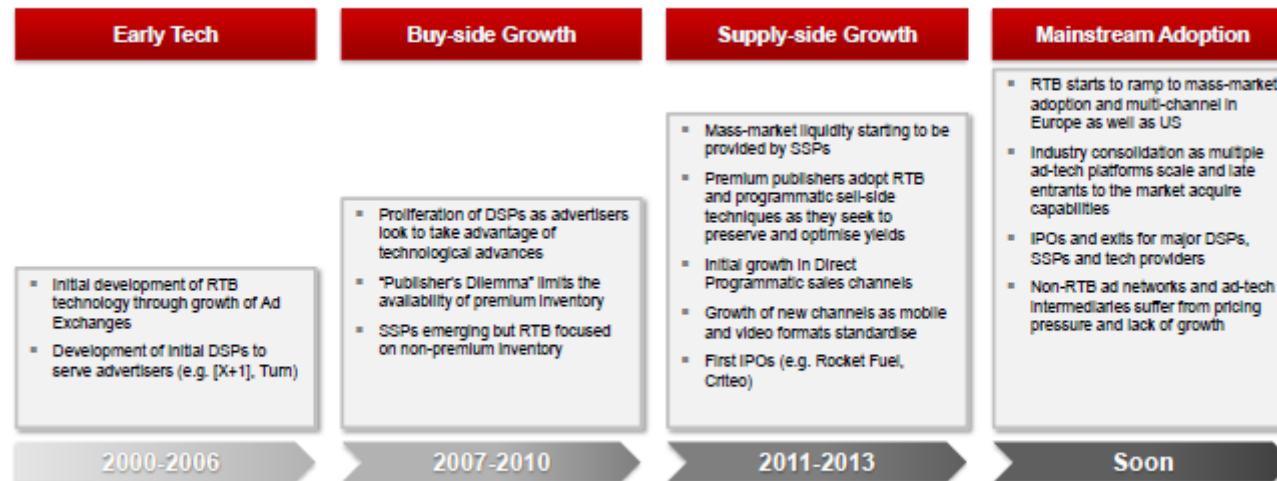


Exemples : champ mercatique

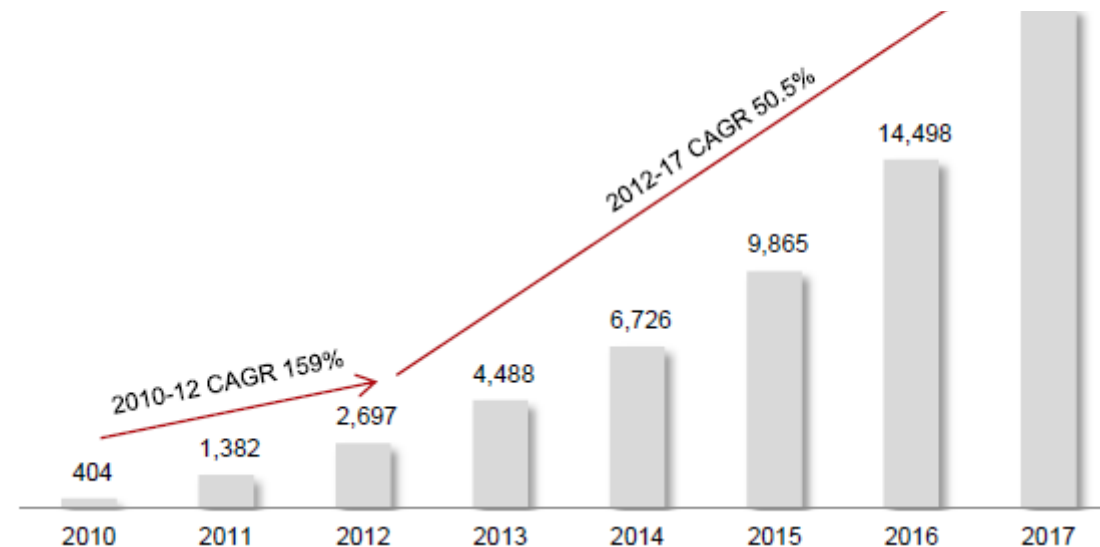


REAL-TIME BIDDING (1)

FIGURE 1 – EVOLUTION OF RTB

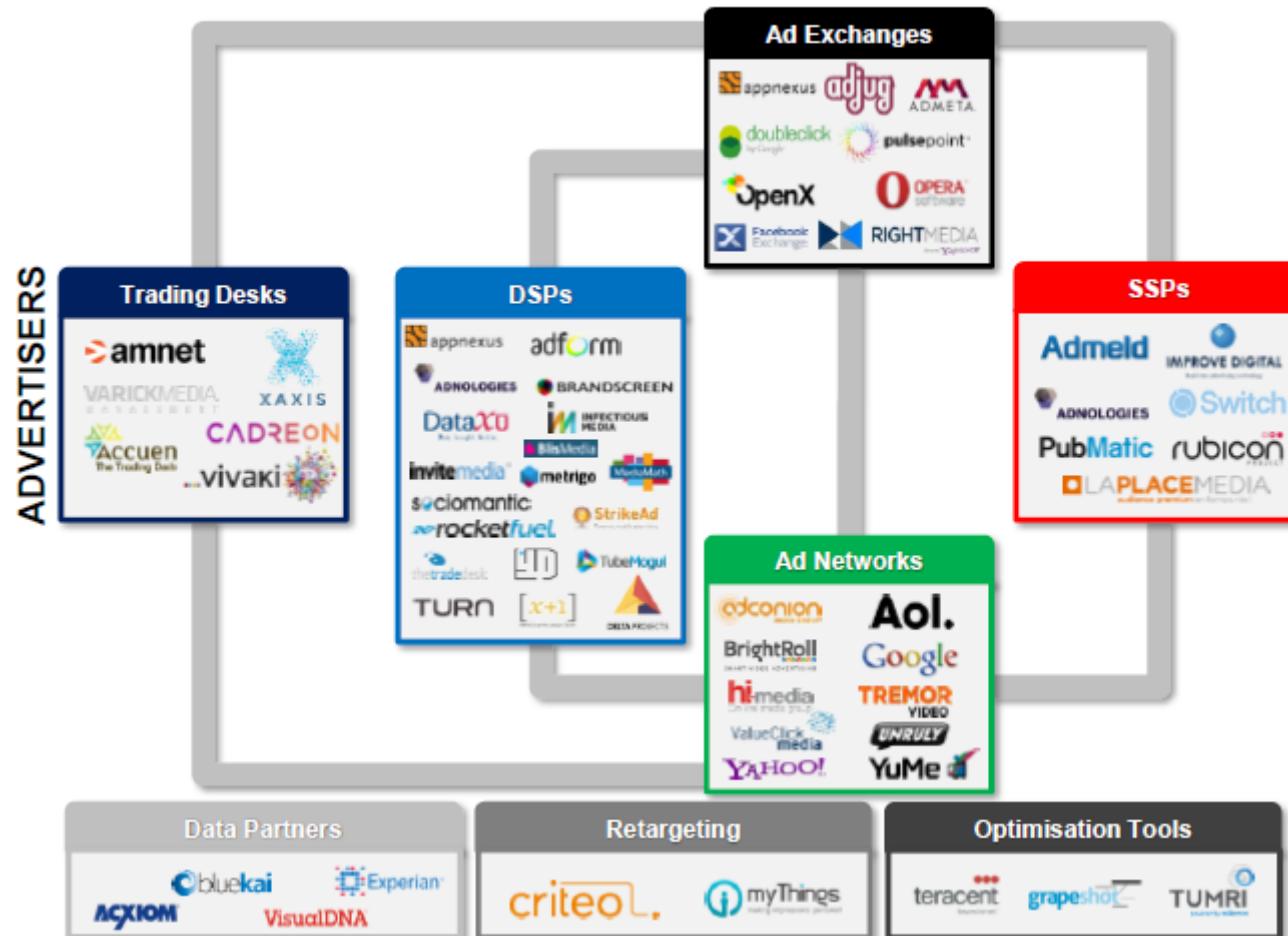


Source: GP Bullhound



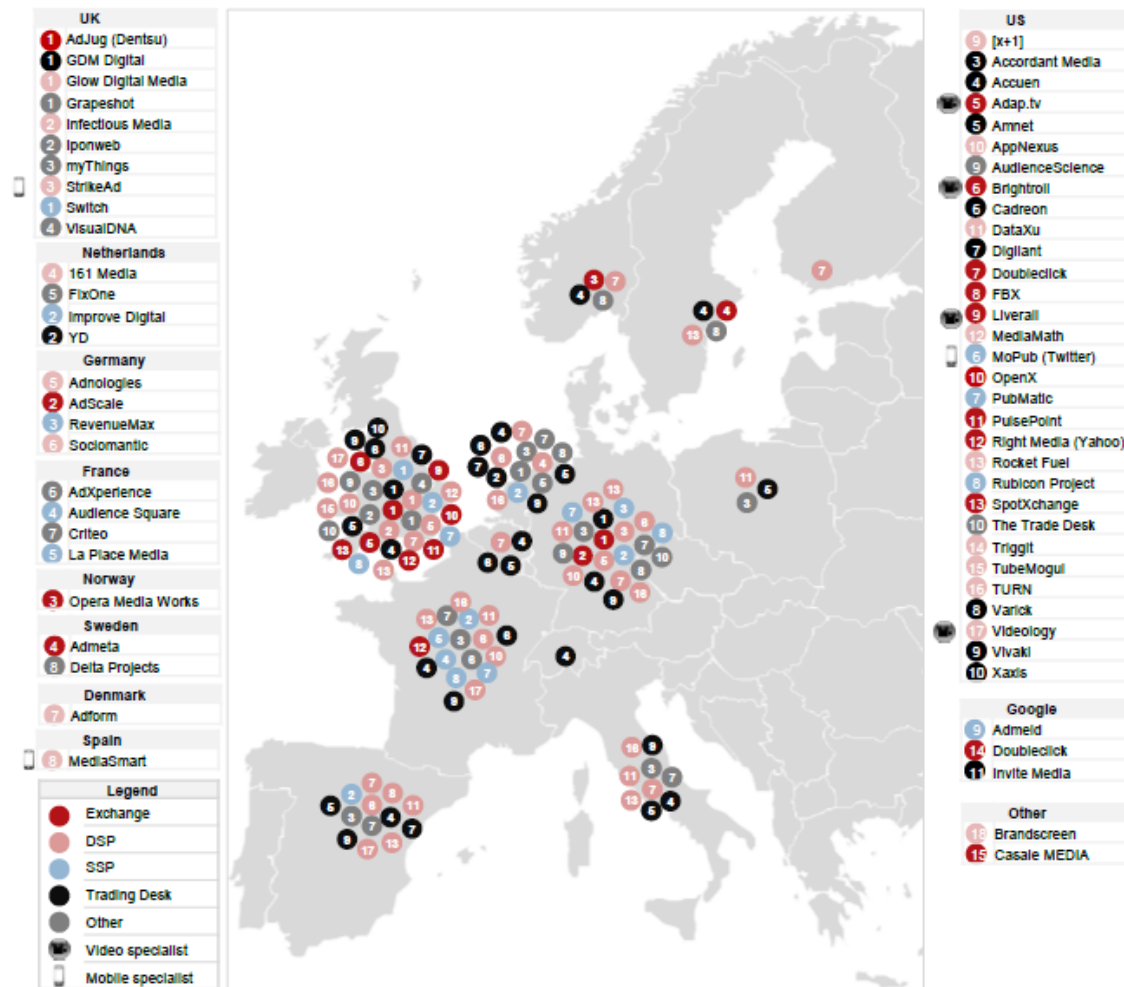
REAL-TIME BIDDING (2)

FIGURE 3 – THE RTB ECOSYSTEM, ILLUSTRATED WITH SELECTED PLAYERS



REAL-TIME BIDDING (3)

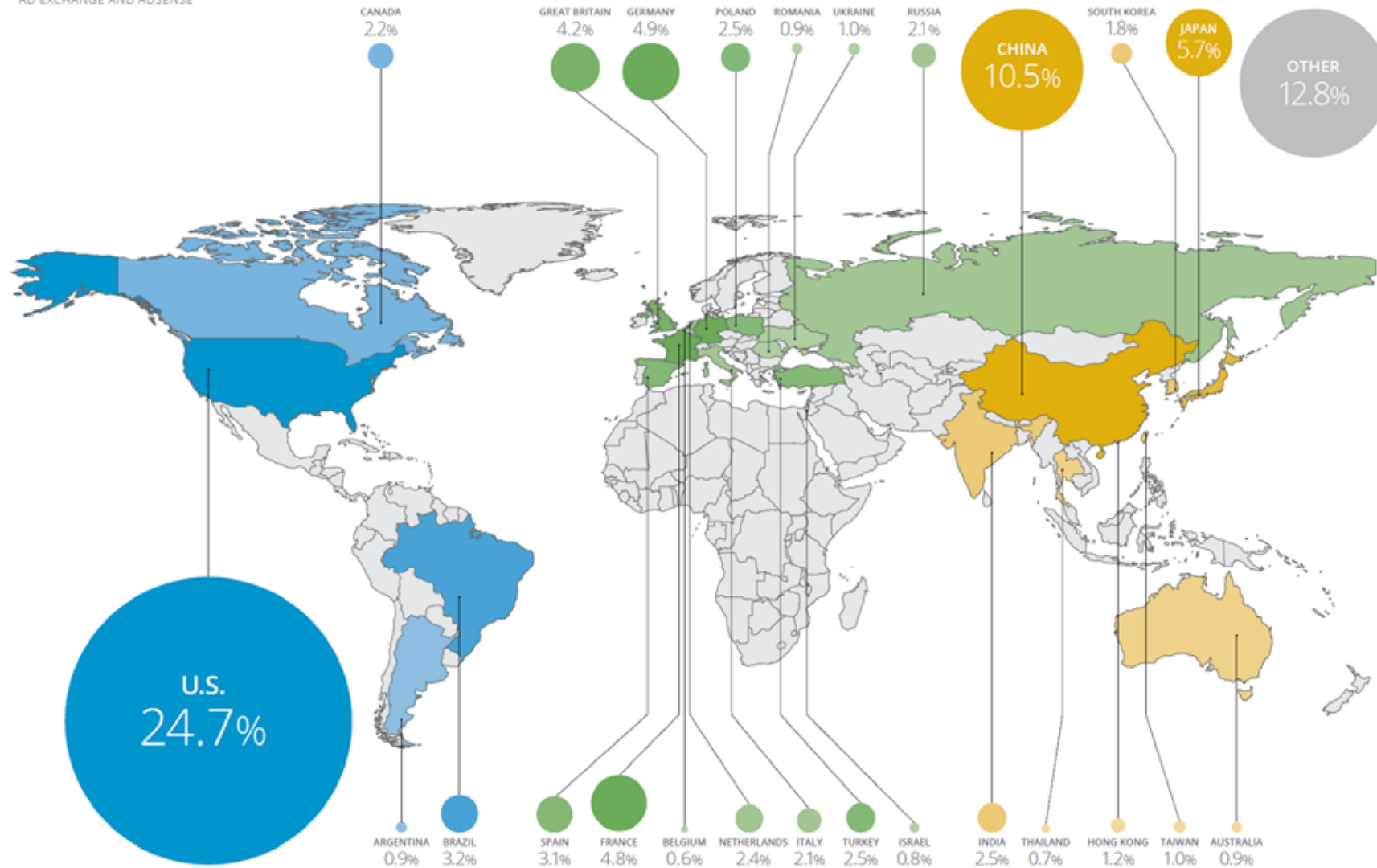
FIGURE 15 – EUROPEAN RTB MAP – SELECTED PLAYERS



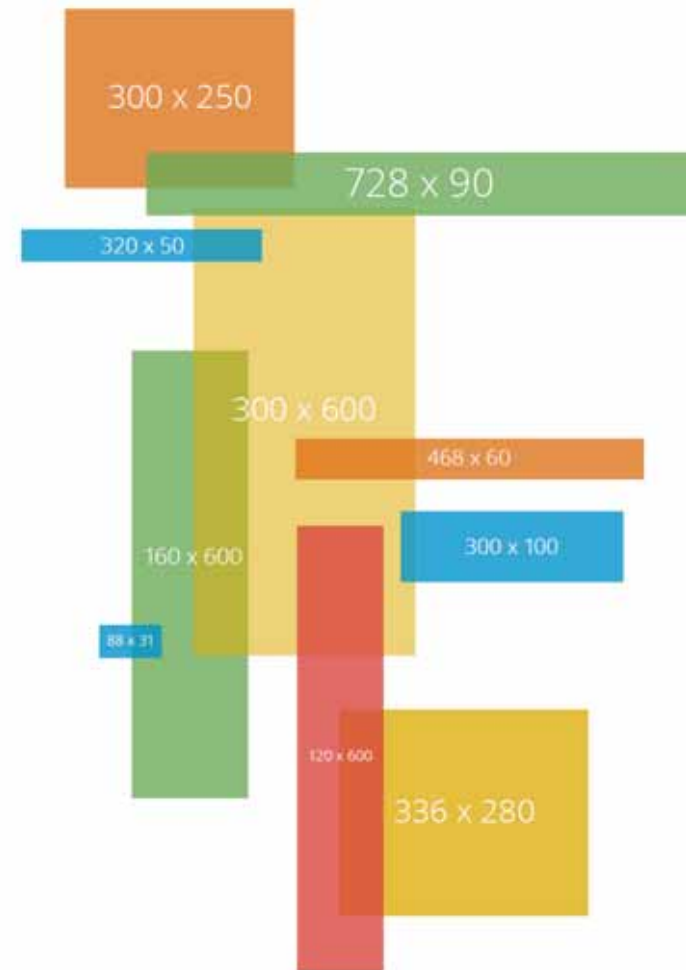
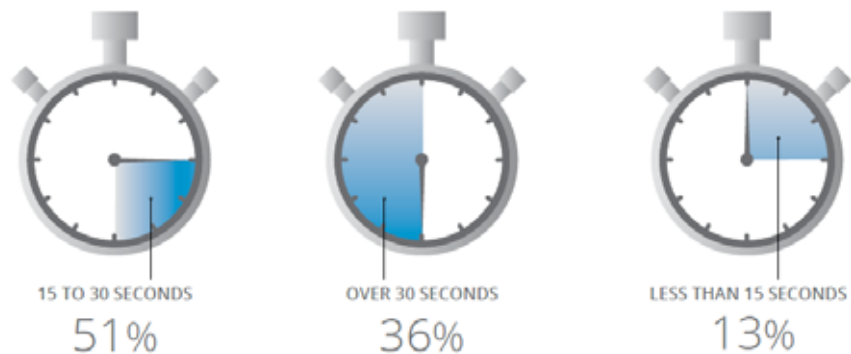
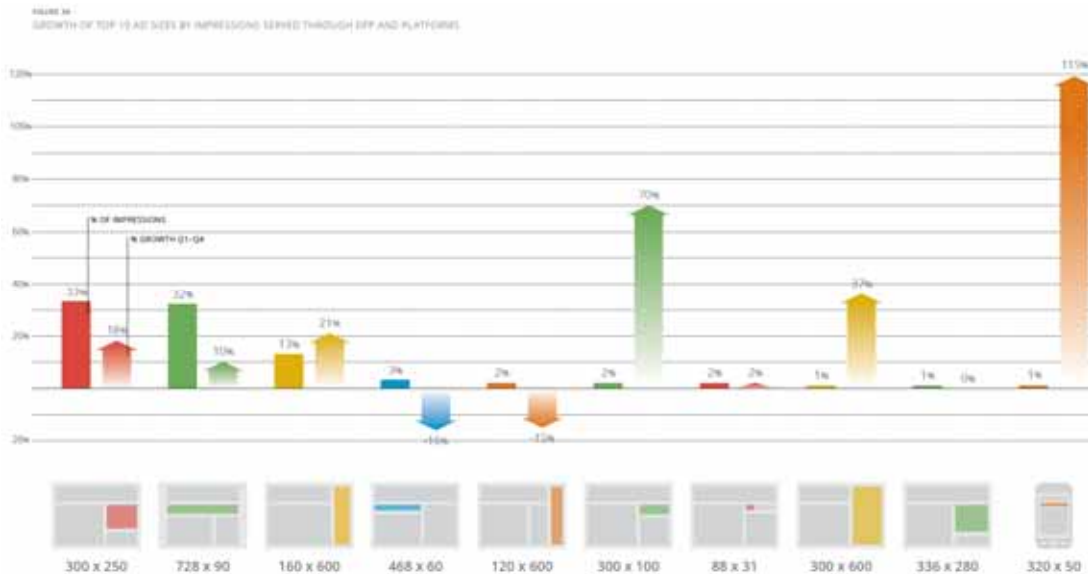
Source: GP Bullhound

REAL-TIME BIDDING (4)

FIGURE 2C
IMPRESSIONS BY COUNTRY ON THE
AD EXCHANGE AND ADSENSE

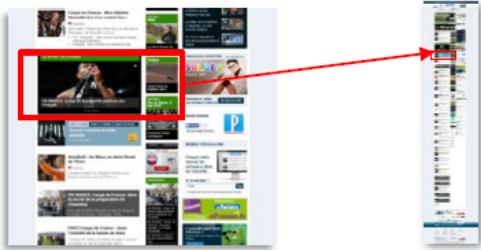


REAL-TIME BIDDING (5)



REAL-TIME BIDDING (6)

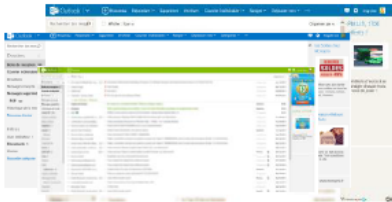
Native Ads, des formats toujours plus intégrés qui se fondent dans l'expérience client



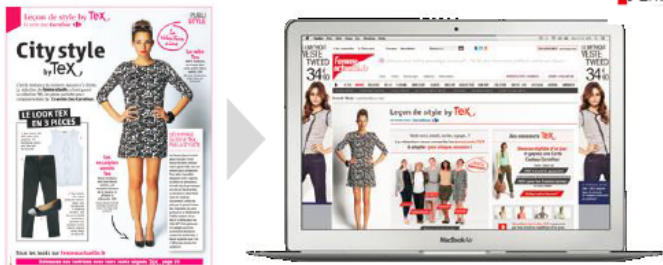
Des slots éditoriaux intégrés dans les flux d'actu qui renvoient vers une page pure « brand content » où un collaborateur de l'annonceur décrit son métier.
Amaury Media pour Renault



Des créations intégrées dans la feuille de route de Viamichelin : un format natif innovant (**Orange Advertising pour le Ministère de l'intérieur**)



Format totalement intégré à la structure : la colonne est utilisée pour présenter le profil associé aux emails quand un mail est affiché.
Format « native » 2 fois plus générateur de clics que la moyenne marché display.
Microsoft Advertising -Outlook



Leçon de style en cross-média de **Prisma Presse pour la marque TEX** : création d'un espace dédié aux couleurs de la marque incluant conseils de styliste et coups de cœur de la rédaction

REAL-TIME BIDDING (7)

DATA

BROKERS

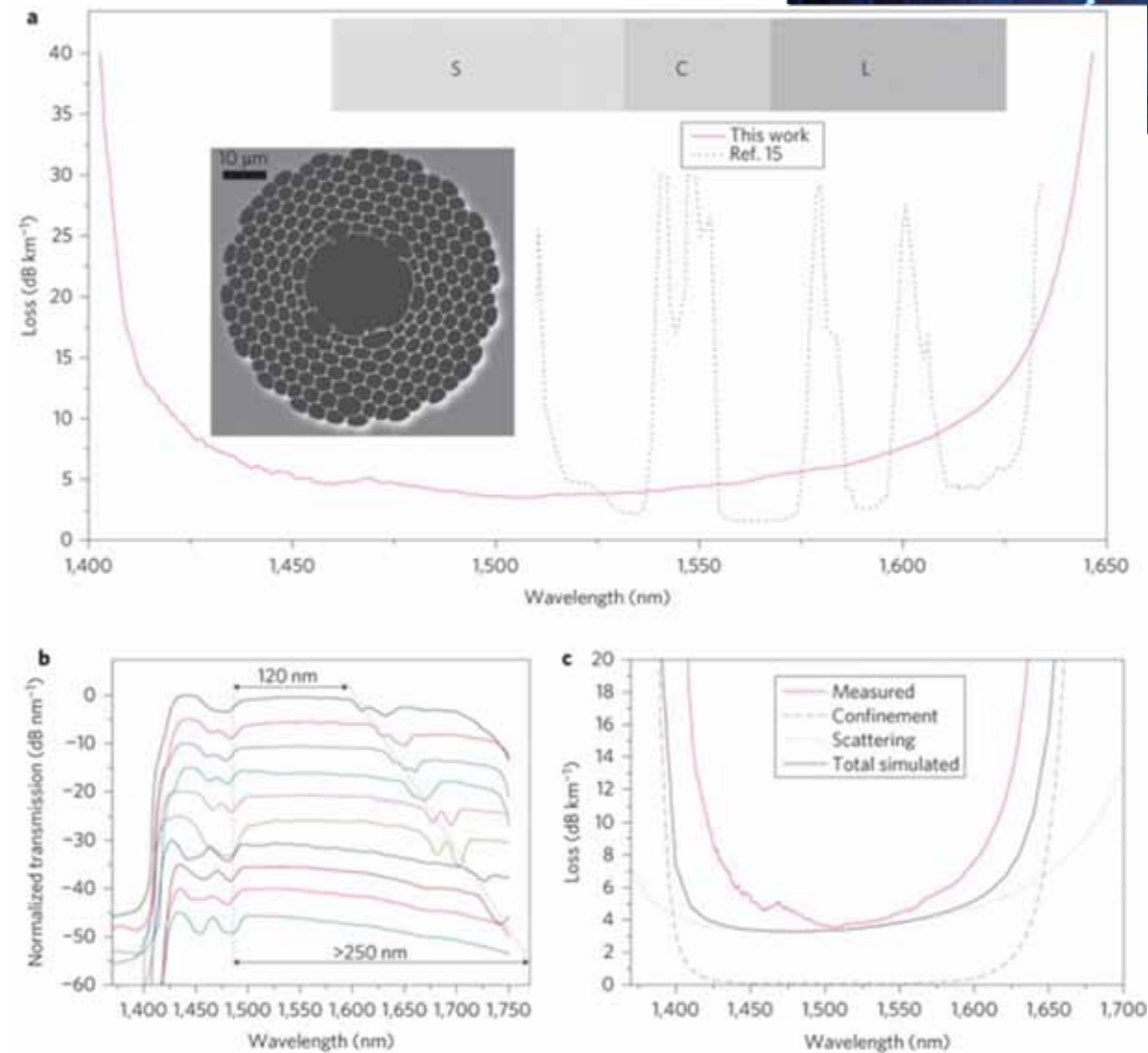
A Call for Transparency and Accountability

Federal Trade Commission
May 2014

HIGH FREQUENCY TRADING

- **Econophysique = proximity hosting**
 - Temps moyen d'exécution des ordres en bourse = 1s
- **2012 : 56% equities, 50% global futures, 35% foreign exchange**
 - 1 transaction : 5 μ s, record (FixNetix) : 740 ns
- **FO : vitesse $2/3*c$ alors que microwaves et nouvelle fibre proche de la vitesse de la lumière.**
 - 1 ms gagnée = 100 millions de dollars gagnés
- **Le délit d'initiés 2.0 se joue à la seconde pour les clients premium**

Nouvelle FIBRE ?



Econophysique

RELATIVISTIC STATISTICAL ARBITRAGE...

PHYSICAL REVIEW E 82, 056104 (2010)

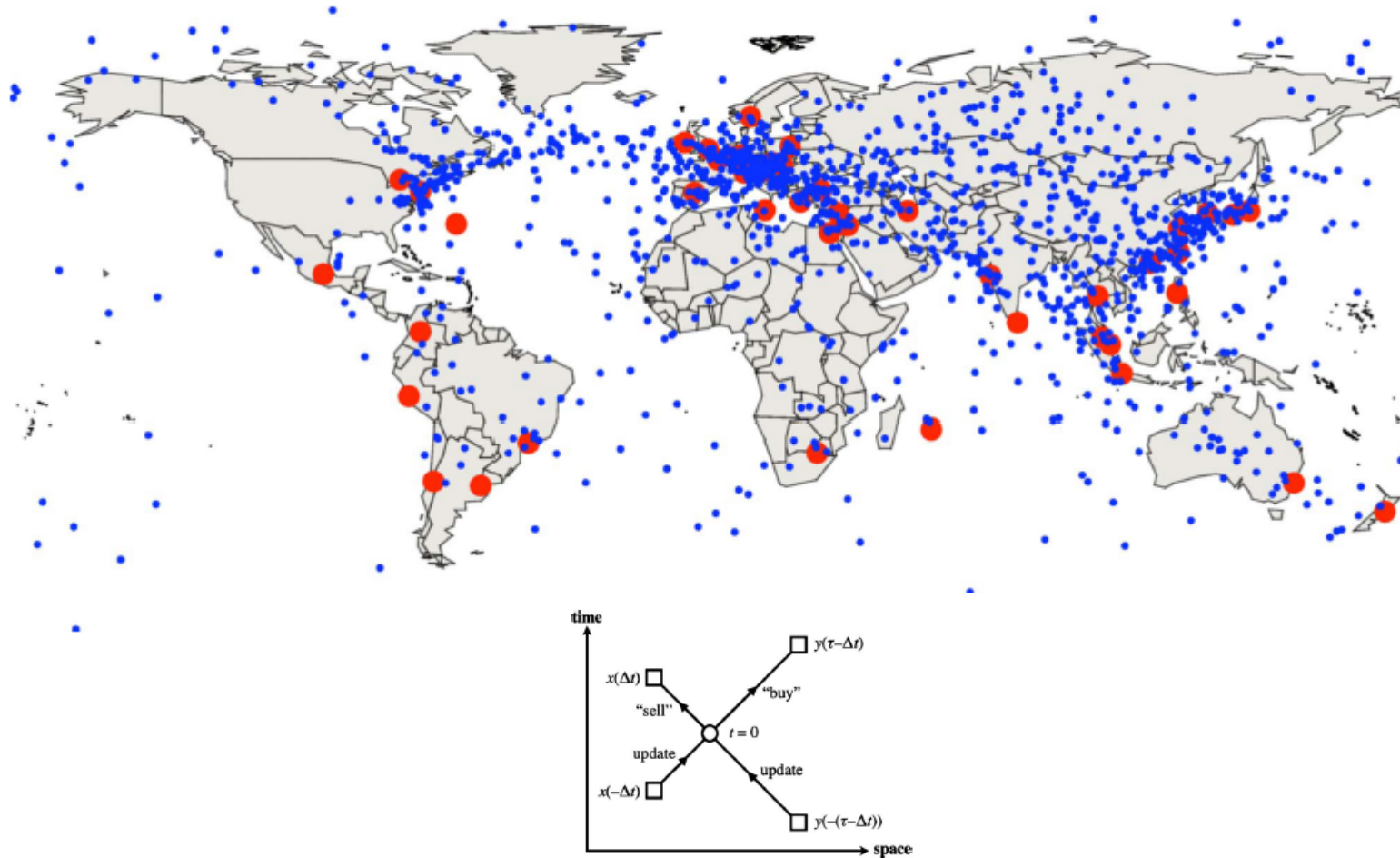


FIG. 1. Space-time diagram of relativistic arbitrage transaction. Security price updates from spacelike separated trading centers (squares) arrive with different light propagation delays at an intermediate node (circle) at time $t=0$, whereupon the node may issue a pair of "buy" and "sell" orders back to the centers.

Exemples : champ sociétal



NEW YORK: DU BIG DATA POUR LUTTER CONTRE LES INCENDIES

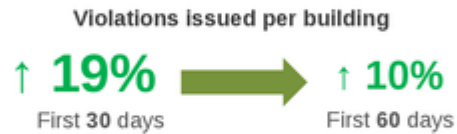


BIG DATA INNOVATION SUMMIT

January 22 & 23, 2014 | Las Vegas

Field Performance | Deployment Evaluation

Relative to FireCast 1.0, how did the new model perform?



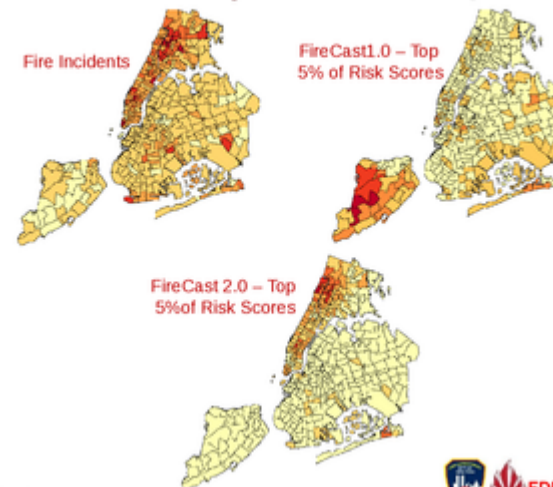
Other Benefits

Units visit high risk buildings that were not previously inspected

FireCast 2.0 | Results



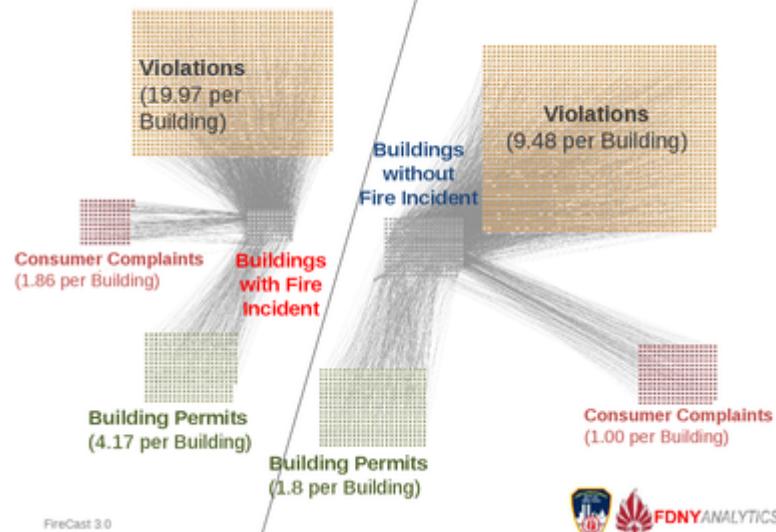
Model Performance | Geo Accuracy



FireCast 2.0 | Results



Exploratory Analysis: Buildings with Fire Incident = 2x more activity



FireCast 3.0



Evaluating the Risk Mitigation Strategy

Successes

- Culture shift from reporting to in-depth analytics
- **10% to 19%** increase in per-building violation issuance relative to former approach
- Higher hit rate of risky buildings

Areas for Improvement

- Citywide data share QAQC protocols
- Expanded data collection to feed FDNY's inspection findings into risk model

Obstacles

- Conventional wisdom
- Raising technical skills of entire workforce
- Data infrastructure not designed for this work

Opportunities

- Pioneer and test new approach for smart municipal government
- Approach can be applied to nearly any risk outcome to identify critical weaknesses in infrastructure to assessing the potential impact of natural disasters

FireCast 3.0



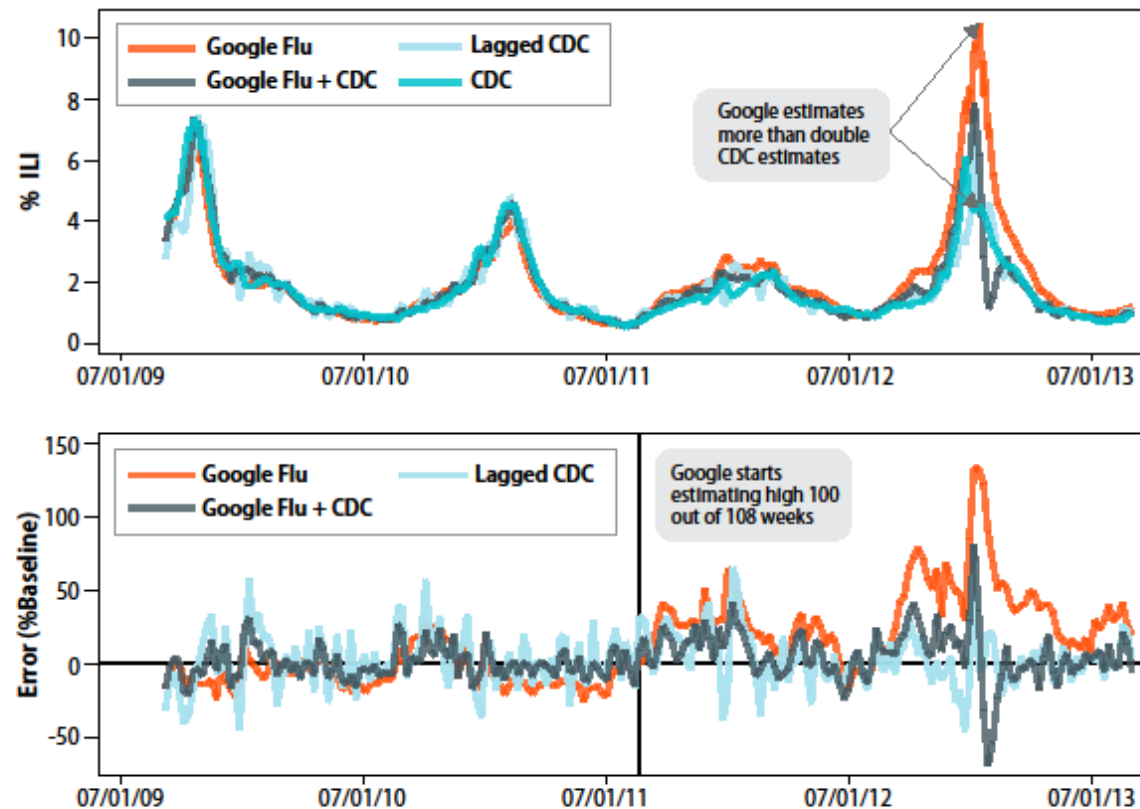


Fig. 1. GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. **(Top)** Estimates of doctor visits for ILI. “Lagged CDC” incorporates 52-week seasonality variables with lagged CDC data. “Google Flu + CDC” combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. **(Bottom)** Error [as a percentage of CDC baseline: (estimate by CDC)/CDC data]. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,2*} Ryan Kennedy,^{1,3,4} Gary King,² Alessandro Vespignani⁵