

Sciences des Données et Données Massives

E. Le Pennec



20ème Journée nationale d'Intelligence Économique d'Entreprise
Intelligence Économique et Mégadonnées
12 décembre 2014

Plan

- 1 Big Data !
- 2 Data Science
- 3 Data Science à l'École polytechnique
- 4 Perspectives

- 1 Big Data !
- 2 Data Science
- 3 Data Science à l'École polytechnique
- 4 Perspectives

Le Big Data, nouvel eldorado des entreprises

Par Direct Matin, publié le 26 Septembre 2014 à 08:32



Les mégadonnées représentent un marché de plusieurs milliards d'euros[© infocux technologies]

Considéré comme le "nouveau pétrole du XXIe siècle", le Big data attise toutes les convoitises.

EN COMPLÉMENT



[INTERNATIONAL](#) [POLITIQUE](#) [SOCIÉTÉ](#) [ÉCONOMIE](#) [CULTURE](#) [IDÉES](#) [SPORT](#) [SCIENCES](#) [TECHNO](#) [STYLE](#) [VOUS](#) [ÉDITION ABONNÉS](#)

M Idées

[IDÉES](#) [Les débats](#) [Think tanks](#) [Points de vue](#) [Editoriaux](#) [Opinions du Monde](#) [Analyses](#) [Idées chroniques](#) [Chats](#) [Blogs](#) [Forums](#)

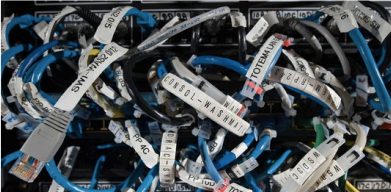
Les données, puissance du futur

LE MONDE | 07.01.2013 à 15h10 • Mis à jour le 07.01.2013 à 18h03

Par Stéphane Grumbach, Stéphane Frénôt

[Abonnez-vous à partir de 1 €](#) [Réagir](#) [Classer](#) [Imprimer](#) [Envoyer](#) [Partager](#) [f](#) [t](#) [g](#) [s](#) [in](#)

[Recommander](#) [Envoyer](#) [f](#) 467 personnes le recommandent.



Les plus partagés

- 1 Une équipe de scientifiques filme un calamar géant par 900 mètres de fond dans le Pacifique [2212](#)
- 2 Infirmiers et aides-soignants refusent d'être des "pigeons" [1664](#)
- 3 Messi remporte son 4e Ballon d'or consécutif [987](#)
- 4 Mariage homosexuel : Wauquiez veut "forcer" le débat sur un référendum [629](#)
- 5 La première Eglise athéiste ouvre à Londres [603](#)

Nous suivre

Retrouvez le meilleur de notre communauté

[f](#) [t](#) [g](#) [s](#) [in](#)

Bits

Search Bits

Go

OCTOBER 24, 2012, 9:00 AM | 4 Comments

Big Data in More Hands

By QUENTIN HARDY

FACEBOOK

TWITTER

GOOGLE+

SAVE

E-MAIL

SHARE

PRINT

Business people, Big Data is coming for you.

Software that captures lots of data and uses it to make predictions has mostly been the province of engineers skilled in arcane databases and statisticians capable of developing complex algorithms. As the business gets bigger, however, software makers are domesticating their products in the hope they will prove attractive to a broader population.

Cloudera, which offers a popular version of the open source database called Hadoop, released software on Wednesday that makes it possible to run queries from a more mainstream SQL programming language interface. SQL, thanks to its adoption by Oracle, Microsoft and others, is known to millions of business analysts.

"This enables us to talk to a whole other class of customer," said Mike Olson, the chief executive of Cloudera. "The knock against Hadoop was that it is too complex."

There is a reason for that. Hadoop is one of several so-called unstructured databases that were created at Yahoo and Google, after those two companies found they had previously unimaginable amounts of data about activities like people's Web-surfing habits. Put into databases designed to handle this unstructured behavior, then analyzed, this information was

PREVIOUS POST

Google Shifts Pitch for Its New Chromebooks

NEXT POST

In Contest for Rescue Robots, Darpa Offers \$2 Million Prize

AROUND THE WEB »

THE NEXT WEB

Google says Maps redirect on Windows Phone was a product decision, and will be removed



BLOOMBERG

HTC Posts Lowest Net Income in Eight Years After Revenue Drops



SCUTTLEBOT News from the Web, annotated by our staff

Google's Schmidt arrive in North Korea

REUTERS | From Mountain View to...errr, Pyongyang? - Somini Sengupta

AP provides sponsored tweets during electronics show

AP.ORG | The Associated Press is renting out its Twitter feed, with 1.5 million followers, to advertisers during C.E.S. - Joshua Brustein

A history of grieving

EDGE-ONLINE.COM | Meet the cult of gamers who want to ruin your day — just for kicks. - Jenna Wortham

A Million First Dates

THE ATLANTIC | Is online romance threatening monogamy? - Jenna Wortham

SEE MORE »

Stratégies.fr

RECHERCHER
Une agence, une marque, un produit...

[Accueil](#) [Créations](#) [Actualités](#) [Etudes/Tendances](#) [New Business](#) [Blogs/Opinions](#) [Guides](#)

[Grands Prix](#) [Marques](#) • [Agences](#) • [Médias](#)

[• Marques](#)

[• Agences](#)

[• Médias](#)

Rechercher dans Marques

Mots clés

Date de parution :
Entre JJ/MM/AAAA 
Et JJ/MM/AAAA 

Secteur d'activité :
-- Sélectionnez -- 

Rechercher

Plan de la rubrique [Marques](#)

Voir les dossiers [Marques](#)

 Votre Flux Rss Stratégies [Marques](#)

[Accueil](#) > [Actualités](#) > [Marques](#) > L'École polytechnique et Télécom Paris Tech ouvrent un master en big data

Liens sponsorisés de Stratégies

MAZARINE DIGITAL : Agence de communication dédiée aux marques de luxe et premium www.mazarine.com

PASSERELLES : Passerelles, agence de communication corporate et relations publiques www.passerelles.com

AGENCE MARKETING OPÉRATIONNEL BTOB : Agence BtoB, Adelanto est Spécialiste en Développement des Ventes, Lead Génération et Marketing Automation www.adelanto.fr

RETOUR    [Version imprimable](#)

L'École polytechnique et Télécom Paris Tech ouvrent un master en big data

03/09/2014

L'École polytechnique et Télécom Paris Tech ouvrent, dès cette rentrée, un nouveau master 2 baptisé « Mathématiques pour la science des masses de données ». Cette formation en big data a pour vocation de « devenir une référence dans le domaine des data sciences au niveau mondial », expliquent les deux établissements. Les enseignements théoriques se répartissent sur une session allant du 1^{er} octobre 2014 au 31 mars 2015. Ces cours sont

LE FIGARO-fr

NEWS - ORIENTATION - STAGE - 1ER EMPLOI - VIE ÉTUDIANTE - INTERNATIONAL

Les entreprises convoitent les futurs spécialistes du «big data»

Par Figaro Étudiant, AFP agence | Publié le 17/09/2014 à 12:35



Surnommées «le pétrole du XXI^e siècle», les données personnelles sont une mine d'or pour les entreprises. -17THOMAS SAMSON/AFP-

Votre adresse e-mail

RECEVOIR NOTRE NEWSLETTER

12 413 187 10

À LIRE AUSSI

- Sujets et corrigés de l'épreuve de français au brevet
- De plus en plus de sélection à l'université
- Sujets et corrigés de l'épreuve d'espagnol LV2

FIGARO ÉTUDIANT
SUR FACEBOOK



HEC, Essec, Polytechnique, Telecom ParisTech: depuis plusieurs mois, les grandes écoles multiplient les formations au «big data». Un phénomène suivi de près par les grandes entreprises.

Jonathan Bibas, 21 ans, sort de [Telecom ParisTech](#), l'une des premières écoles d'ingénieurs à avoir lancé en septembre 2013 une formation en big data («mégadonnées», en bon français). «Je n'ai pas cherché de travail, j'ai tout de suite été contacté par plusieurs agences de recrutement», explique-t-il. Depuis deux semaines, il travaille chez Air Liquide, leader mondial des gaz industriels, comme «data scientist». Son travail? Valoriser les données de l'entreprise en espèces sonnantes et trébuchantes, grâce à ses connaissances pointues en mathématiques et informatique. «En ce moment, je cherche à comprendre comment réduire la pollution en ville grâce au big data», confie-t-il.

Big Data !

Médias : l'X aussi



EducPros.fr

Au service des professionnels de l'enseignement supérieur

by l'Étudiant

Rechercher sur EducPros



ACTUALITÉS ENQUÊTES ENTRETIENS VEILLE BLOGS OPINIONS ÉVÉNEMENTS EDUCPROS

ON EN PARLE

Rentrée universitaire. Najat Vallaud-Belkacem défend son budget

Accueil > Actualité > Polytechnique et Télécom ParisTech plus que jamais dans le big data

Actu | Formations

Polytechnique et Télécom ParisTech plus que jamais dans le big data

Sophie Blitman | Publié le 28.08.2014 à 16H57, mis à jour le 28.08.2014 à 16H57

Recommander 34

Tweeter 68

+1 1

commentaire 0

A+ A- Imprimer Partager



L'École polytechnique et Télécom ParisTech ouvrent, à la rentrée 2014, un master dédié au big data. Un nouveau signe de l'intérêt des écoles pour ce secteur porteur d'emploi.

**“DATA
IS THE
NEW
OIL.”**

From the beginning of record time until 2003, we created
5 exabytes (5 billion gigabytes)
of data.

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to **10 minutes.**

Every hour, we
create enough
Internet traffic
to fill
7 billion

Side by side, that's
that's seven times
the height of

Coined in 2006 by Oliver Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

As of August 2001, there were just over **4 million**

There are
133 million
BLOGS
on the web

80%

66% of all Americans own mobile phones of one sort or another. Out of those with mobiles, 34% are flip phones. (In fact, 54% of all Americans own any type of mobile phone.)

English is the dominant language of the web.
But by 2014 it will be **Chinese**,
if its current rate of increase continues.

Top languages used on the web (May 2011):

Based on the data provided in the chart, the languages are ranked by the number of people who speak them. The languages and their approximate number of speakers are:

Language	Approximate Number of Speakers (in millions)
Korean	120
Russian	140
French	160
Arabic	220
German	240
Portuguese	260

247 billion EMAILS

are sold **every day.** (Up to 60% off.)

Language	Speakers (millions)
Spanish	450
Chinese	1,400
English	1,100

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders**, with the help of computer algorithms, use Big Data to follow trends and to act quickly on market swings.

These specialized algorithms
make split-second decisions to buy or sell a commodity.
New cable being laid under the Atlantic will share.

5 milliseconds

from the current 65 milliseconds
It takes for trading instructions
to travel between
New York City

With new fiber-optic cable

This 5-millisecond saving is worth many millions of dollars to the transatlantic finance firms who use the cable (and who pay millions to its cost).

How they save 5 milliseconds

The new cable will lie on areas of the ocean floor that are up to 1,800 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.

50%

of 5-year-old kids in the U.S. are given access to a smartphone.

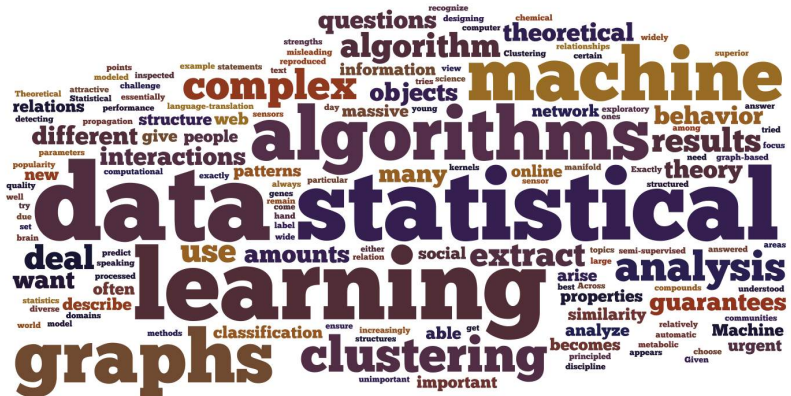
Information provided by Susan Lindberg





Big Data !

Beaucoup de mots !



Big data

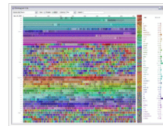
From Wikipedia, the free encyclopedia

This article is about large collections of data. For the band, see [Big Data \(band\)](#).

Big data^{[1][2]} is the term for a collection of **data sets** so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage,^[3] search, sharing, transfer, analysis^[4] and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."^{[5][6][7]}

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of **exabytes** of data.^[8] Scientists regularly encounter limitations due to large data sets in many areas, including [meteorology](#), [genomics](#),^[9] [connectomics](#), complex physics simulations,^[10] and biological and environmental research.^[11] The limitations also affect [Internet search](#), [finance](#) and [business informatics](#). Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies ([remote sensing](#)), software logs, cameras, microphones, [radio-frequency identification](#) readers, and [wireless sensor networks](#).^{[12][13]} The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s.^[14] as of 2012, every day 2.5 **exabytes** (2.5×10^{18}) of data were created.^[15] The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.^[16]

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers".^[17] What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."^[18]

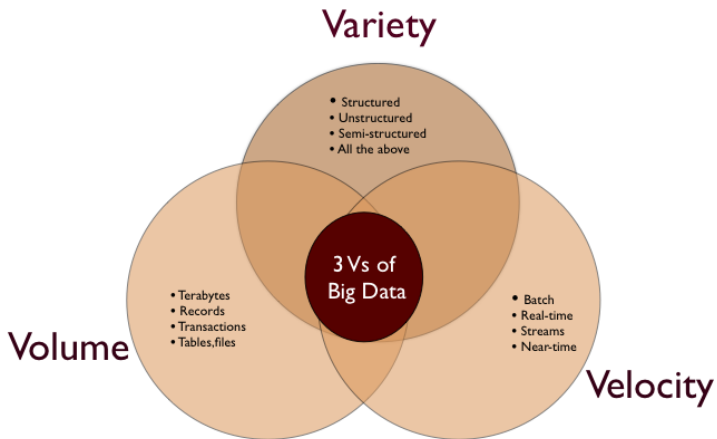


A visualization created by IBM of Wikipedia edits. At multiple **terabytes** in size, the text and images of Wikipedia are a classic example of big data.

- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.

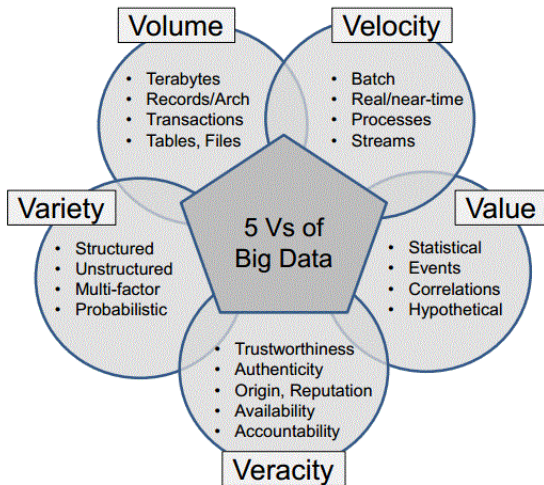
Big Data !

Les 3 Vs du Big Data



Big Data !

Les 5 Vs du Big Data



Big Data !

Doing Data Science

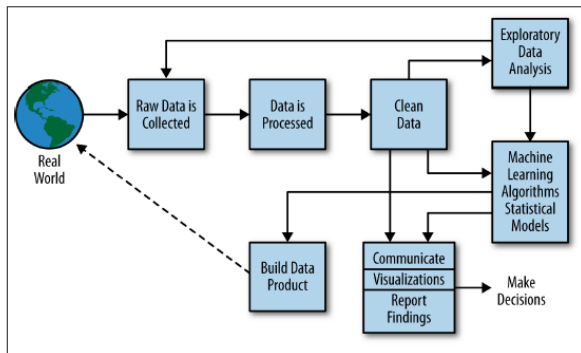


Figure 2-2. The data science process

- *Doing Data Science : Straight talk from the frontline* - Rachel Schutt, Cathy O'Neil - O'Reilly
- Art de la décision/évaluation à partir de données.

Des données aux produits

Moteur de recherches



 moteur de recherche 

[Web](#) [Actualités](#) [Images](#) [Vidéos](#) [Maps](#) [Plus ▾](#) [Outils de recherche](#)

Environ 10 100 000 résultats (0,24 secondes)

Moteur de recherche - Mozbot France - La recherche facile ...
www.mozbot.fr/ ▾
Moteur de recherche Mozbot en partenariat avec Brioude-Internet, Abondance et Google : résultats, synonymes, expressions connexes, statistiques mots clés, ...

Actualités correspondant à **moteur de recherche**



Le moteur de recherche DuckDuckGo bloqué en Chine
[Le Monde](#) - il y a 3 heures
Selon le site spécialisé TechnAsia, le **moteur de recherche** serait bloqué depuis le 4 septembre dans le pays. DuckDuckGo, qui se présente ...

[L'Allemagne souhaite que Google dévoile les algorithmes ...](#)
[Clubic.com](#) - il y a 5 jours

Plus d'actualités pour "**moteur de recherche**"

Moteur de recherche — Wikipédia
fr.wikipedia.org/wiki/Moteur_de_recherche ▾
Un **moteur de recherche** est une application web permettant de retrouver des ressources (pages web, articles de forums Usenet, images, vidéo, fichiers, etc.) ...

Moteur de Recherche SEEK.fr™
www.seek.fr/ ▾
Moteur de recherche alternatif français respectant la vie privée via un métamoteur utilisant les principaux **moteurs de recherche** ainsi qu'un annuaire ...
[Métamoteur Web SEEK.fr](#) - [A Propos de Seek](#) - [Horoscope](#) - [Seek annuaire](#)

More Ideas Based on Your Browsing History

You looked at



Thriving in the Knowledge Age: New... Paperback by John H. Falk
~~\$29.95~~

You might also consider



Museum Administration: An Introduction Paperback by Hugh H. Genoways
~~\$31.95~~ **\$28.75**



Exhibit Labels: An Interpretive Approach Paperback by Beverly Serrell
~~\$34.95~~ **\$27.85**

► [Find similar items](#)

Recommendations don't have to be
about showing you more of the same...

Des données aux produits

Publicité en ligne



Outlet
» Descubrielo

Innovatoren und Kleinunternehmer nutzen ihre Möglichkeiten bei Amazon
» Ihre Geschichten



Jetzt neu:
Schnell & einfach Ersatzteile finden
» Hier klicken



365 Tage im Jahr Licht bei 0€ Stromkosten
» Hier klicken



Libros universitarios y de estudios superiores a precios bajos
» Descubrielo



Neuheiten von Makita
» Hier klicken



fire + 12 MONTHS OF PRIME
PHONE
NOW ONLY \$0.99
with a two-year contract » Shop now



Fall Outlet Event
» Shop now



FALL COATS
» See more



New from iRobot:
Roomba 870
Vacuum Cleaning Robot
» Learn more



Save Big
on Outdoor Fire Pits from Strathwood
» Shop now



Rentrée des Conservatoires

-10% sur une sélection d'instruments*

*Voir conditions » Cliquez ici



Vos courses en livraisons gratuites et régulières

Economisez en vous Abonnant

» Cliquez ici



PROMOTIONS CHAUSSURES
-30% -40% -50%...
» J'en profite



PROMOTIONS SACS À MAIN
» J'en profite



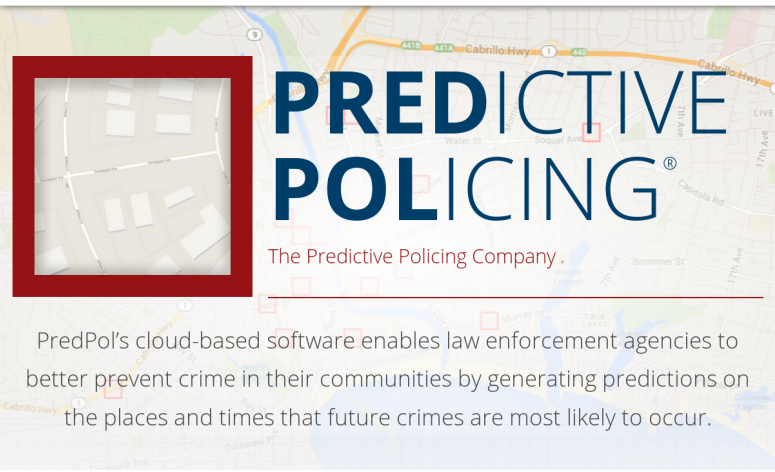
Des données aux produits

Détection d'intrusion





[ABOUT](#) [HOW PREDPOL WORKS](#) [PROVEN RESULTS](#) [TECHNOLOGY](#) [PRESS](#) [CONTACT US](#) [BLOG](#)



PREDICTIVE POLICING®

The Predictive Policing Company.

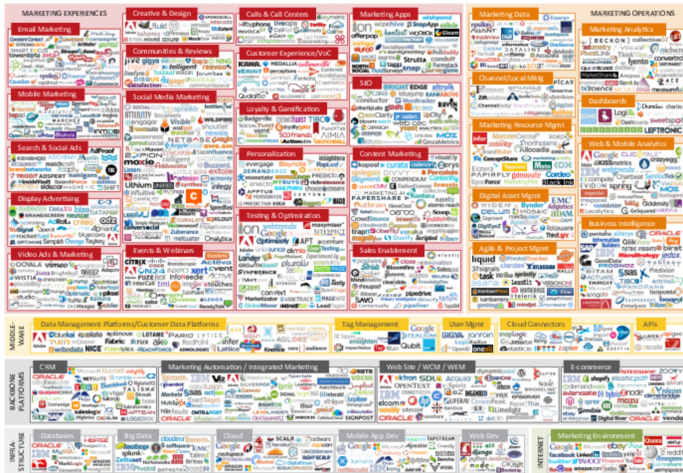
PredPol's cloud-based software enables law enforcement agencies to better prevent crime in their communities by generating predictions on the places and times that future crimes are most likely to occur.

Des données aux produits Marketing

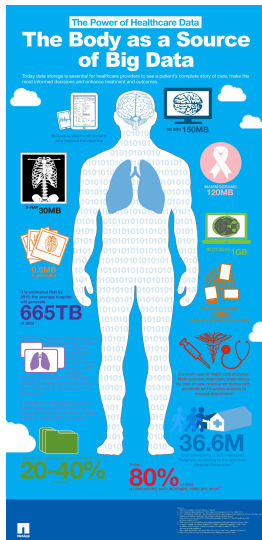


chiefmartec.com Marketing Technology Landscape

January 2014



by Scott Brinker @chiefmartec <http://chiefmartec.com>

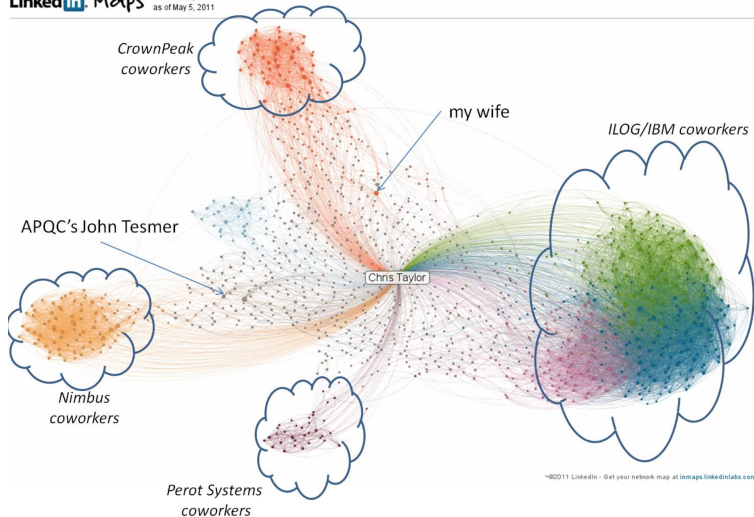


Des données aux produits

Réseaux sociaux



LinkedIn Maps Chris Taylor's Professional Network
as of May 5, 2011



©2011 LinkedIn - Get your network map at inmaps.linkedinlabs.com

Smarter Cities: Turning Big Data Into Insight

City Planning and Operations

\$1 Trillion

global annual savings could be attained by optimizing public infrastructure.
Source: McKinsey

\$57 Trillion

in infrastructure investments will be needed between 2013-2030.
Source: McKinsey

Transportation Analytics

50 Hours

of traffic delays per year are incurred, on average, by travelers.

30 Billion

people all over the world travel approximately 30 billion miles per year. By 2050, that figure will grow to over 150 billion miles.

Cloud is driving cities in their digital transformation.

Water Management

60%

of water allocated for domestic human use goes to urban cities.

\$14 Billion

in potable water is lost every year because of leaks, theft and unbilled usage.
Source: World Bank

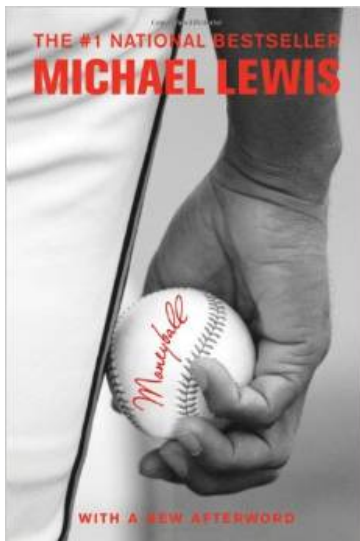
Open Cloud

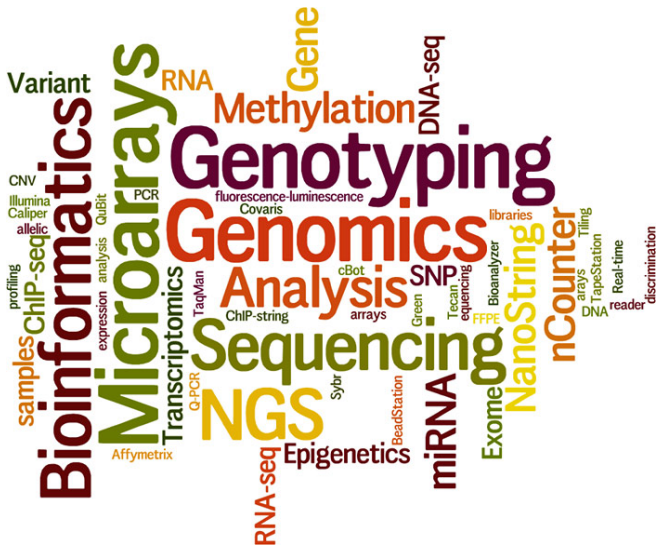
\$6 Billion

has been invested by IBM in more than a dozen acquisitions to accelerate its cloud initiatives.

IBM Intelligent Operations software is designed with cities, for cities, to provide the tools to monitor, visualize and analyze vital city services such as water and wastewater systems, transportation, infrastructure planning, permit management and emergency response.

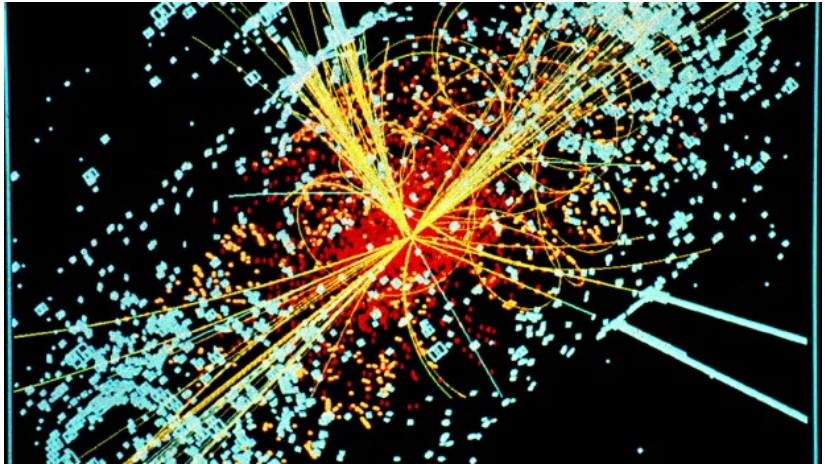






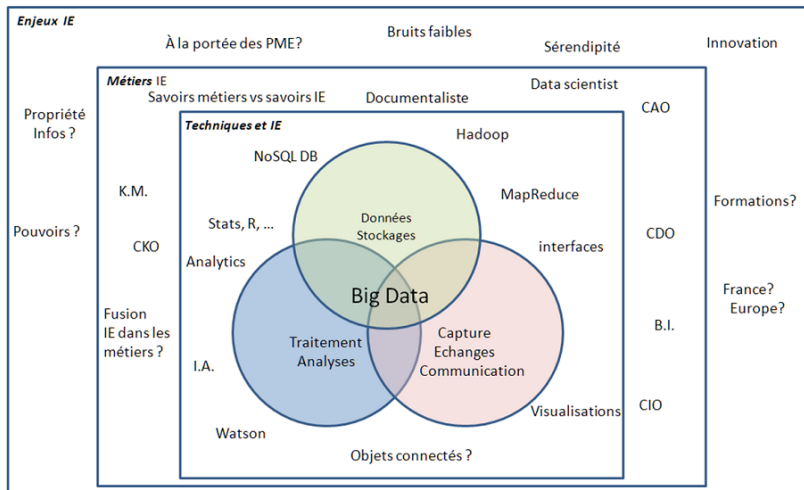
Des données aux produits

Physique



Des données aux produits

Intelligence économique



- 1 Big Data !
- 2 Data Science**
- 3 Data Science à l'École polytechnique
- 4 Perspectives

- Omni-présence des données :
 - Grande volumétrie,
 - Grande variété...
- Technologie de calcul abordable :
 - Cloud computing
 - Graphical Processor Units (GPU)...
- Intérêt grandissant industriel et académique

Un exemple de solution “clé en main”, code open-source :



```
def run(params: Params) {  
  val conf = new SparkConf()  
    .setAppName(s"BinaryClassification with $params")  
  val sc = new SparkContext(conf)  
  
  Logger.getRootLogger.setLevel(Level.WARN)  
  
  val examples = MLUtils.loadLibSVMFile(sc, params.input).cache()  
  
  val splits = examples.randomSplit(Array(0.8, 0.2))  
  val training = splits(0).cache()  
  val test = splits(1).cache()  
  val numTraining = training.count()  
  val numTest = test.count()  
  println(s"Training: $numTraining, test: $numTest.")  
  examples.unpersist(blocking = false)  
  
  val updater = params.regType match {  
    case L1 => new L1Updater()  
    case L2 => new SquaredL2Updater()  
  }  
  
  val algorithm = new LogisticRegressionWithSGD()  
    .optimizer  
    .setNumIterations(params.numIterations)  
    .setStepSize(params.stepSize)  
    .setUpdater(updater)  
    .setRegParam(params.regParam)  
  val model = algorithm.run(training).clearThreshold()  
  
  val prediction = model.predict(test.map(_.features))  
  val predictionAndLabel = prediction.zip(test.map(_.label))  
  
  val metrics = new BinaryClassificationMetrics(predictionAndLabel)  
  val myMetrics = new MyBinaryClassificationMetrics(predictionAndLabel)  
  
  println(s"Empirical CrossEntropy = ${myMetrics.crossEntropy().}")  
  println(s"Test areaUnderPR = ${metrics.areaUnderPR().}")  
  println(s"Test areaUnderROC = ${metrics.areaUnderROC().}")  
  
  sc.stop()  
}
```

Un exemple de solution “clé en main”, code open-source :



```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop distcp s3n://celluledecalcul/dataset/raw/train.csv /data/train.csv
scp -i cellule.pem cellule/challenge/target/scala-2.10/target/scala-2.10/challenges_2.10-0.0.jar

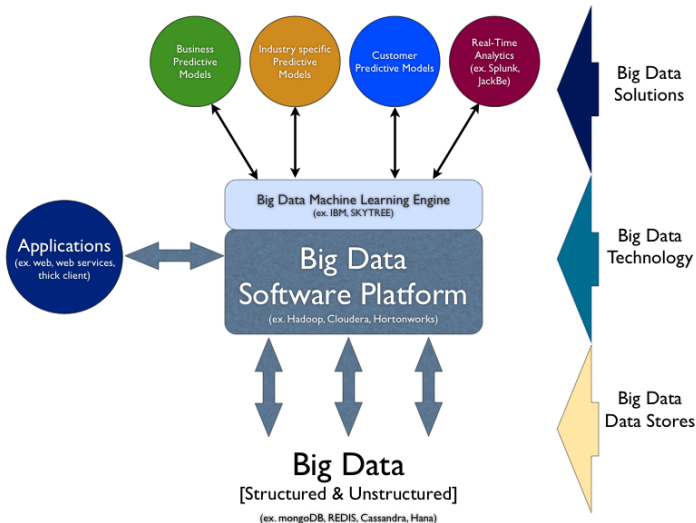
cellule/spark/bin/spark-submit \
    --class fr.cc.challenge.Preprocess \
    challenges_2.10-0.0.jar \
    /data/train.csv \
    /data/train2.csv

cellule/spark/bin/spark-submit \
    --class fr.cc.sparktest.LogisticRegression \
    challenges_2.10-0.0.jar \
    /data/train2.csv
```

⇒ Une régression logistique distribuée avec Spark

Data Science

Un écosystème complexe !



Infrastructure

- NoSQL Databases:** 10gen, COUCHBASE, boshjo, CLOUDANT, HYPERTABLE, Neo4j, MarkLogic, memsql, DRAVNSCALE, VoltDB, NUODB.
- NewSQL Databases:** MarkLogic, memsql, DRAVNSCALE, VoltDB, NUODB.
- MPP Databases:** VERTICA, kognitio, PARACCEL, GREENPLUM, N, NETAZZA, InfiniDB, SQL Server.
- Cross Infrastructure / Analytics:** SAP, Ssas, IBM, Google, Microsoft, vmware, amazon, Hadoop, META MARKETS, TERADATA, NetApp.

Analytics

- Analytics Solutions:** platforma, Palantir, PERVASIVE, Datameer, KARMASPHIRE, DataHive, Kitegroup, CENTRIFUGE, Ayasdi, ClearStory, metaLayer, tableau, ISS, Quantum4D, dataspot, PREDDIC, SKYTREE, sense, REVOLUTION, SAS, SPSS, SAS, SAS.
- Data Visualization:** Quid, ACTUAL.ly, VIZUAL, Kitegroup, CENTRIFUGE, Ayasdi, ClearStory, metaLayer, tableau, ISS, Quantum4D, dataspot, PREDDIC, SKYTREE, sense, REVOLUTION, SAS, SPSS, SAS, SAS.
- Statistical Computing:** SKYTREE, sense, REVOLUTION, SAS, SPSS, SAS, SAS.
- Sentiment Analysis:** GENERAL SENTIMENT, crimon hexagon, Owlin, Location / People / Events, RapLeaf, Flitop, Recorded Future, Place IQ, RADIUS.
- Real-Time Analytics:** ParStream, feedzy, Crowdsourced Analytics, DataKind, kaggle, SMB Analytics, sumail, RJMetrics, custora.

Applications

- Ad Optimization:** DataXu, m6d, ai Match, bluekai, rocketfuel, TURN, cross.
- Industry Applications:** KNEWTON, finance, Wongai, numberFire, MileSense, BILLIS, Climate Solutions, Bloomberg, GUARD.
- Application Service Providers:** collective[i].

Data Sources

- Data Marketplaces:** factual, DataMarket, Windows Azure Marketplace, BASIS, Personal Data, Withings, LAW BONE, RunKeeper, fitbit.
- Data Sources:** premise, DATA SIF, knoema, GNIP, infochimps, SPAC, COUL.

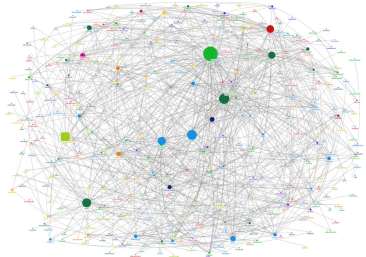
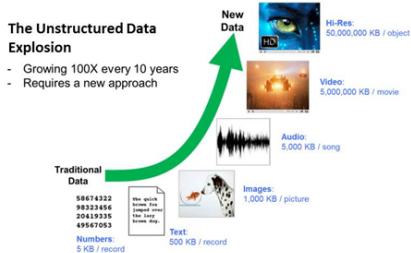
Open Source Projects

- Framework:** Hadoop, HDFS, Hive.
- Query / Data Flow:** riik, Cassandra, SciDB, HBASE, COUCHDB, Sqoop.
- Data Access:** mongoDB, SciDB, HBASE, COUCHDB, Sqoop.
- Coordination / Workflow:** ZooKeeper, talend, OGG.
- Real-Time:** Storm.
- Statistical Tools:** SciPy.
- Machine Learning:** Mahout.
- Cloud Deployment:** Amazon, Google, Microsoft, vmware, amazon, Hadoop, META MARKETS, TERADATA, NetApp.

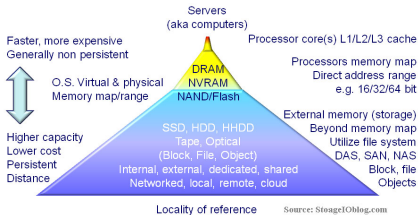
Matt Turck (@mattturck) and Shivon Zilis (@shivonz)

- Mathématiques appliquées ET informatique
- Recherche méthodologique nourrie par de nombreuses disciplines : marketing, traitement du signal, génétique, biologie, physique, médecine, ...
- Nouvelles expertises conjointes (maths/info)
- Nouveaux langages/frameworks MPP/GPU/Cloud : hadoop, spark, graphlab, cuda, ...
- Données massives : parallélisation, réduction du nombre de passages sur les données
- Gestion de la grande dimension : réduction de dimension
- Données très hétérogènes : représentation, bases de données noSQL

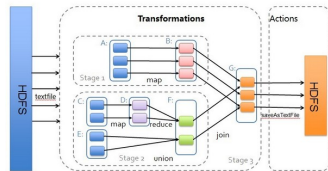
Besoin de **data-scientists**, besoin de formation



- Comment stocker efficacement les données ?
- Comment décrire les données pour les utiliser par la suite ?
- Comment mettre en relation des données de nature différente ?



Spark: Transformations & Actions

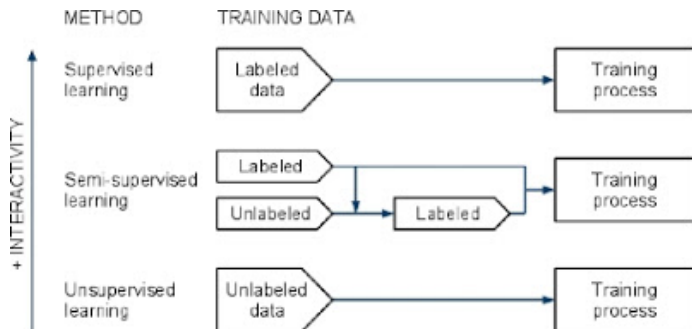


- Comment prendre en compte la localisation des données ?
- Comment construire des architecture de parallélisation ?
- Comment développer des algorithmes adaptés.

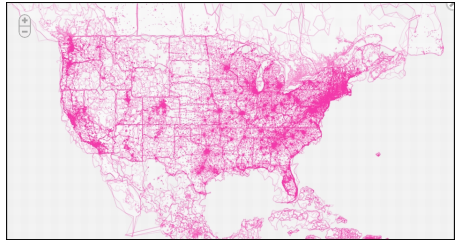
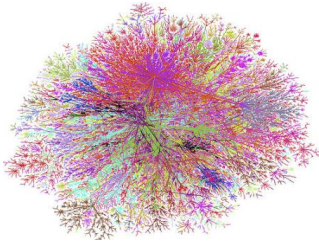
Main Paradigmatic Changes in Big Data Analytics Environment			
	Statistical Data Analysis <1985 (Pure Statistical Inference)	Business Intelligence 1985-2008 (Constrained Data Mining)	Big Analytics >2008 up to now (Unconstrained Data Mining)
Data types	Homogeneous Structured Data (proprietary)	Homogeneous Structured & Homogeneous Unstructured Data, separately	Mix of Heterogeneous Unstructured & Structured Data (proprietary + open data)
Data storing	Line & column dimensions fixed Flat Files, Hierarchical DBs, & First Relational DBs	Column dimensions fixed SQL DBs: MySQL, DB2, ORACLE & OLAP Cubes	No dimensions fixed NoSQL DBs: Column oriented DBs, object oriented DBs etc.
Volume Cost/volume	Exponential cost decrease		Exponential volume increase
Basic Analytical Principles	Hypotheses driven mode: Power use of sampling Techniques	Mix Hypotheses driven & Data driven: Dimensions Reduction & Populations Segmentations	Full Data driven mode: Power use of learning techniques, mainly unsupervised
Main Algorithmic approaches	Regression Analysis, Factorial Analysis, Statistical Inference the sampling, Linear general Models, Decision Trees Etc.	Clustering (K-means, K Neighbours), Classification & Support Vector Machines Multi layers Neural Nets, Scoring Techniques, Sequential Patterns, etc.	Deep adaptive learning techniques, Auto encoded neural Nets, Huge Graph Modularization, & Visual Analytics, Full unsupervised linear Clustering, etc.
New types of Business deliverables	Score Cards, Decisional Models based on sampling	etc. Populations Profiling: CRM, Churn & Attrition Analysis, Loyalty & Propensity Programs, Cross selling	New real time analysis for "microtargeting": adapted online marketing & sales, machine learning for various purposes, automated maintenance programs

THALES

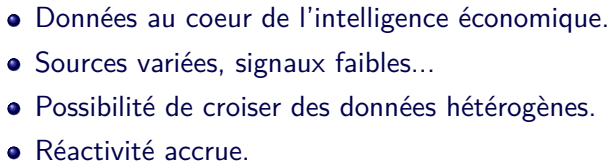
- Comment représenter les données ?
- Comment réduire la dimension des modèles utilisés ?
- Comment choisir les modèles ?

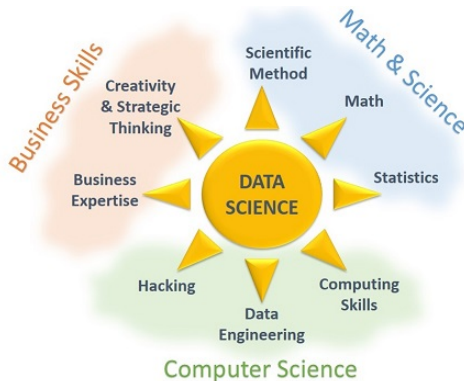


- Comment apprendre avec le moins d'interactions possibles ?



- Comment voir les données ?
- Comment présenter les résultats ?





- Le *Data Scientist* !
- Très difficile de maîtriser l'ensemble au niveau individuel.
- Nécessité de penser en terme d'équipe.
- Recrutements difficiles : peu de formation et forte compétition.

- 1 Big Data !
- 2 Data Science
- 3 Data Science à l'École polytechnique**
- 4 Perspectives

- 2ème année :
 - Cours fondamentaux dont statistique et informatique,
 - Projet de presque 1 an en équipe, de nombreux sujets en data-science.
- 3ème année :
 - **Programme d'approfondissement complet Data Science,**
 - Cours dédiés : Apprentissage, Big Data, Base de données...
 - Formation à l'entrepreneuriat
 - Stages de recherche de 3ème année en entreprise ou laboratoire académique
 - Projets de trois mois de recherche en machine learning
- 4ème année :
 - **Master Data Science**



<http://www.lix.polytechnique.fr/bigdata/mathbigdata/>

- Futur parcours du master Mathématiques et Applications de Paris Saclay dédié à la formation de Data Scientist.
- Ouvert dès cette année par l'École polytechnique et Telecom Paris Tech en collaboration avec l'ENSAE Paris Tech.
- Intervenants : École polytechnique, Telecom, Univ. Paris Sud, ENSAE, ENS Cachan...
- Statistique, Optimisation, Machine Learning, Visualisation, Graphes, Données textuelles, etc.
- 24 semaines + stage de 5 mois
- Briques technologiques permettant le "passage à l'échelle" : calcul et optimisation distribuée, Framework Hadoop, bases de données noSQL, stockage distribué.

Data Science Starter Program - Polytechnique DSSP

Un nouveau programme de formation continue en Data Science et Big Data, proposé par l'École Polytechnique

et ses départements de Mathématiques Appliquées et d'Informatique

<https://www.polytechnique.edu/bigdata/dssp>

- Programme de formation continue en Data Science et Big Data
- Proposé par les départements de mathématiques appliquées et d'informatique de l'X.
- 10 semaines (vendredi et samedi 9h–18h)
- Data Management, data analysis and machine learning, unsupervised learning, big data camp
- Intervenants industriels : AXA, Canal+, CDiscount, Criteo, DataKnowledge, Group HIL, Heva, Keyrus, MFGLabs, Microsoft, Orange, Predicis, Publicis, Thales, 1000mercis

- Chaire créée en octobre 2014 par l'École polytechnique, Keyrus, Orange et Thales et portée par la fondation de l'X
- But : former la prochaine génération de Data Scientist
- Actions :
 - Appuyer la formation initiale en Data Science proposée par l'École polytechnique
 - Soutenir les actions de formation continue en Data Science
 - Encourager la recherche en Data Science

Data Science à l'École polytechnique

Equipe SIMPAS au CMAP



Rapport d'activité

Informations Générales

Les équipes de recherche

Enseignement

Publications

Annuaire

Recrutement

Chaires

Webmail

Intranet

Planning des salles de réunion

Rechercher

sur ce site



[Accueil du site](#) > [Equipes](#) [Fr](#) [En](#) [It](#) > SIMPAS

SIMPAS - Signal Image Probabilités numériques Apprentissage Statistique

Equipe SIMPAS

(Signal Image Probabilités numériques Apprentissage Statistique)

Responsables scientifiques : E. Gobet et E. Le Pennec

Cette équipe regroupe des chercheurs du CMAP dans le domaine de l'aléatoire (au sens large), dont les travaux sont axés sur le traitement numérique des données ou des modèles aléatoires, allant des fondements théoriques des algorithmes et méthodes, aux développements informatiques efficaces.

Chercheurs confirmés :

- ▶ [Stéphanie Allasonnière](#) (Professeur Chargé de Cours Ecole Polytechnique)
- ▶ [Emmanuel Bacry](#) (Chargé de Recherches CNRS)
- ▶ [Antonin Chambolle](#) (Directeur de Recherches CNRS)
- ▶ [Stefano De Marco](#) (Professeur Chargé de Cours Ecole Polytechnique)
- ▶ [Stéphane Gaïffas](#) (Professeur Chargé de Cours Ecole Polytechnique)
- ▶ [Emmanuel Gobet](#) (Professeur Ecole Polytechnique)
- ▶ [Erwan Le Pennec](#) (Professeur Associé Ecole Polytechnique)
- ▶ [Guillaume Lecué](#) (Chargé de Recherches CNRS)

Post-doctorants :

- ▶ [Plamen Turkedjiev](#) (2013-2015) : simulation des processus non-linéaires
- ▶ [Jacopo Mastromatteo](#) (2014-) : statistique des carnets d'ordre

Doctorants encadrés au laboratoire :

- ▶ [Adrian Iuga](#) (2010-2013), encadré par E. Bacry et M. Hoffmann (Univ. Paris-Dauphine) : Modélisation et analyse statistique de la formation de prix à travers les échelles
- ▶ [Romain Bompis](#) (2010-2013), encadré par E. Gobet : Développements asymptotiques pour l'approximation de processus de diffusions
- ▶ [Hao Xu](#) (2011-2014), encadré par S. Allasonnière et B. Thirion (INRIA Parietal)
- ▶ [Thibault Jaisson](#) (2012-2015), encadré par E. Bacry et M. Rosenbaum (UPMC) : Etudes de problèmes statistiques venant de la microstructure des marchés
- ▶ [Gang Liu](#) (2013-2016), encadré par E. Gobet et P. Del Moral (INRIA Alea) : Simulation d'évènements rares pour les fonctionnelles browniennes
- ▶ [Etienne Corman](#), encadré par A. Chambolle et M. Ovsjanikov (LIX) : Matching de formes
- ▶ [Pauline Tan](#), encadrée par A. Chambolle et P. Monasse (CERTIS, ENPC) : vision stéréo.

Doctorants encadrés en dehors du laboratoire :

- ▶ [Islem Rekik](#) (2010-2013), encadré par S. Allasonnière et J. Wardlaw (Neuroradiologie, Univ. Edimbourg)

- **Apprentissage machine** : web mining, big data, grande dimension, apprentissage non supervisé ou faiblement supervisé
- **Approximations stochastiques** : méthodes asymptotiques, optimisation et algorithmes stochastiques, MCMC et inférence bayésienne
- **Méthodes de Monte-Carlo** : régression en grande dimension, processus stochastiques non-linéaires, méthodes particulières et évènements rares
- **Statistique mathématique** : estimation non-paramétrique, sélection de modèles, classification, réduction de dimension
- **Statistique des processus** : finance, turbulence, processus multifractaux
- **Traitement du signal et de l'image** : méthodes adaptatives, représentation adaptées, analyse multi-échelle, multi-fractales, compressed sensing, méthodes variationnelles. Applications à l'imagerie médicale.



[Home](#) [News](#) [Members](#) [Research](#) [Publications](#) [Teaching](#) [External services](#) [Contact](#)

This year our team will organize the [LIX colloquium 2014](#) under the theme: "Data Science" on November 28, 2014 in Inria Saclay Building.

Our team

The Data Science and Mining (DaSciM) team is part of the [Computer Science Laboratory \(LIX\)](#) of [École Polytechnique](#), and the successor and evolution of the Data and Web Mining team ([AUEB](#), Greece). In the previous years we have researched in the areas of databases and data mining. More specifically in unsupervised learning (clustering algorithms and validity measures), advanced data management and indexing (P2P systems, distributed indexing, distributed dimensionality reduction), text mining (word disambiguation for classification) and ranking algorithms (temporal extensions to PageRank). More recently, supported by the DIGITEO Chair Levetone grant, we are working in large scale graph mining (degeneracy based community detection and evaluation), text mining and retrieval for web advertising/marketing and recommendations. The leader of the team has supervised previously nine completed Ph.D. theses and supervises six more underway and has published chapters in books and encyclopedias, two international books and more than a hundred twenty (120) papers in international refereed journals and conferences. Also we have coauthored three patents and attracted significant R&D funding including national and international research & development projects. Members of our team have received the ERCIM, Marie Curie, and Google fellowships.

Latest News

LIX colloquium 2014

 [September 2014](#) This year our team will organize the [LIX colloquium 2014](#) under the theme: "Data Science" on November 28, 2014 in Inria Saclay Building.

New Master on "Mathematics for Data Science"

 [September 2014](#) The new Master degree on "Mathematics for Data Science" co-organized by Ecole Polytechnique and Telecom-Paristech is inaugurated in September 2014. See details [here](#).

 [IEEE](#) - Internet Computing special issue on "Web Graph and So-

- Mining large graphs – algorithms, foundations and applications
- Mining large evolving text collections
- Distributed algorithms for large scale data
- Applications (Web, Social Networks, Advertising, Bio-data, industrial data)

Data science & Big Data : domaine transverse

⇒ Fédérer les différents laboratoires à l'Ecole Polytechnique

- Centre de Mathématiques Appliquées
- DACSIM (Computer Science lab - LIX)
- Le Prince Ringuet Lab (CERN Data Science)
- Département d'Economie
- Département de Physique
- ...

Développer les partenariats avec l'industrie

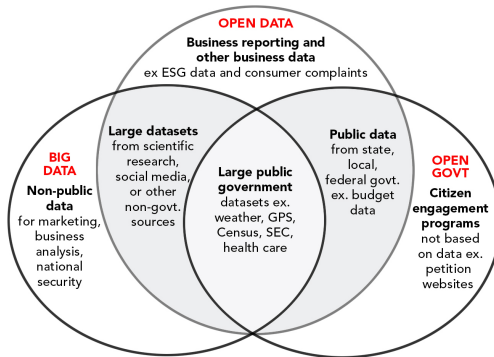
⇒ Chaires ou contrats de recherche

Exemples de partenariats :

- FBF (Fédération des banques Françaises)
Trading Haute Fréquence et régulation
- CNAM (Caisse Nationale d'Assurance Maladie)
Une des plus grandes bases de données de santé au monde ($\simeq$ 1 PetaOctet)
- Chaire Data Scientist
- Chaire Finance et Développement Durable (marchés de l'énergie, régulation)
- ...

- 1 Big Data !
- 2 Data Science
- 3 Data Science à l'École polytechnique
- 4 Perspectives**

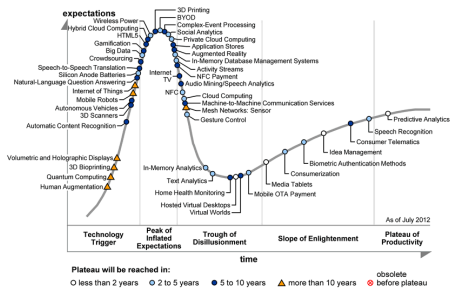
- Environnement français **très** favorable :
 - Formation scientifique de haut niveau,
 - Besoin mesuré d'investissement financier,
 - Évolution des mentalités.
- Interaction Startup/PME/Grandes entreprises.
- L'École polytechnique souhaite être en pointe sur ces domaines !



- Données : carburant indispensable du Big Data.
- Environnement législatif/administratif : vie privée, données désanonymisantes...

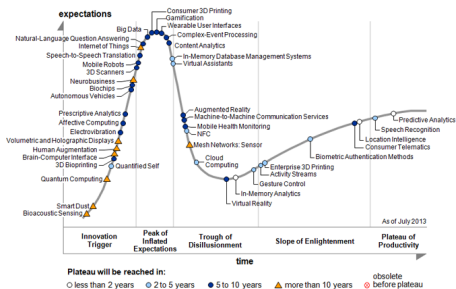
Perspectives

Don't believe the hype?



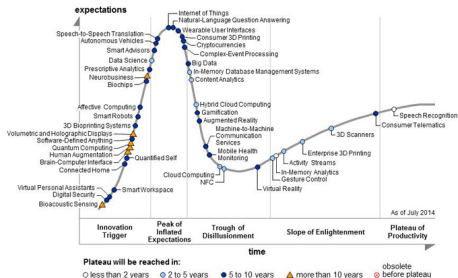
Perspectives

Don't believe the hype?



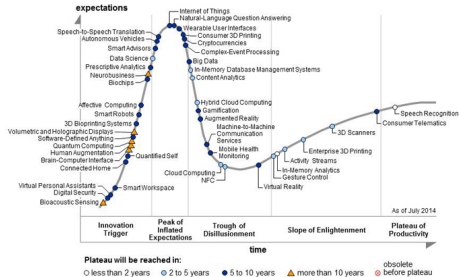
Perspectives

Don't believe the hype?



Perspectives

Don't believe the hype?



- Data Science et Big Data : beaucoup plus qu'un effet de mode !



- Énorme opportunité !
- Travaillons tous ensemble et ne pas la manquons pas...