FORUM NTIC BIG DATA, OPEN DATA
Big Data: les challenges, les défis

Joseph Salmon

Télécom ParisTech

Jeudi 6 Février

Agenda

- Contexte et opportunités
- Challenges scientifiques et collaborations industrielles
- Le défi de la formation
- Qelques actions en cours et à venir

"Big Data" - Le contexte

Une accumulation de données **massives** dans de nombreux domaines:

- Biologie/Médecine (génomique, essais cliniques, imagerie, *etc.*)
- Réseaux sociaux (Facebook, Tweeter)
- Moteurs de recherche internet (contenu multimédia)
- Administrations (santé publique, douanes)
- Sécurité (biométrie, vidéosurveillance, informatique)
- Banque/Finance (risque de marché/liquidité, accès au crédit)
- Grande distribution, marketing (CRM), e-commerce
- ...

"Big Data" - Le contexte

Un **déluge de données** qui rend inopérant:

- les outils basiques de
  - stockage de données
  - gestion de base de données (MySQL)
- le prétraitement reposant sur l'expertise humaine
  - indexation, analyse sémantique
  - modélisation
  - intelligence décisionnelle

"Big Data" - Le contexte

Une multitude de briques technologiques et de services disponibles pour:

- La parallélisation massive (Velocity)
- Le calcul distribué (Volume)
- La gestion de données sans schéma prédéfini (Variety)

parmi lesquels:

- Calculs parallélisés/distribués: MapReduce
- Framework Hadoop
- NoSQL: SGBD Cassandra, MongoDB, bases de données orientées graphe, moteur de recherche Elasticsearch, *etc.*
- Clouds: infrastructures, plate-formes

promus par Google, Amazon, Facebook, *etc.*

"Big Data" - Les opportunités

Des avancées spectaculaires

- la **collecte** et le **stockage** (distribué) des données
- la **recherche** automatique d'objets, de contenu
- le **partage** de données peu structurées

Le Big Data: un moteur pour la technologie, la science, l'économie

- Moteurs de recherche, moteurs de recommandation
- Maintenance prédictive
- Marketing viral à travers les réseaux sociaux
- Détection des fraudes
- Médecine individualisée
- Publicité en ligne

"Big Data" - Les opportunités

Ubiquité

De nombreux secteurs d'activité sont concernés:

- (e-) Commerce
- Gestion de la relation client
- Santé
- Défense, renseignement
- Banque/Finance
- Transports "intelligents"
- *etc.*

"Big Data" - Les besoins

Talents

Selon l'Institut McKinsey Global, d'ici à 2018 aux USA:

- 300 000 **architectes Big Data**
- 140 000-190 000 recrutements de **data scientists**
- 1,5 million de cadres/techniciens avec des connaissances générales dans ce domaine

"Big Data" - Recherche

Pour exploiter les données (prédiction, interpétation), il faut développer des technologies mathématiques permettant de résoudre les problèmes computationnels liés:

- aux contraintes du temps réel
  $\rightarrow$ apprentissage automatique séquentiel "on-line" ($\neq$ batch)
- au caractère distribué des données/ressources
  $\rightarrow$ apprentissage automatique distribué
- à la volumétrie des données
  $\rightarrow$ Impacte des techniques de sondages/échantillonnage sur la performance des algorithmes

"Big Data" - Recherche

Des techniques de visualisation, représentation de données complexes

- Graphes (évolutifs) - clustering, graph-mining
- Image, audio, vidéo - filtrage, compression
- Données textuelles (*e.g.* page web, tweet)

Domaines

- Probabilité, Statistique
- Machine-Learning
- Optimisation
- Traitement du signal et de l'image
- Analyse harmonique computationnelle
- Analyse sémantique
- Informatique, *etc.*

Formations "Data Science": un bref tour d'horizon

En France

- Masters **disciplinaires**: informatique ou mathématiques appliquées

A l'international

- Stanford University (Stanford Center for Professional Development): "Data Mining and Applications Graduate Certificate" en 3 ans (12 000 $ env.), avec Sony, Cisco
- Columbia Institute for Data Sciences and Engineering, Chicago Northwestern University (MS program in Predictive Analytics), North Carolina State University (MS in Analytics avec SAS), UC San Diego (certificate program in data mining), *etc.*
- Secteur privé: SAS, EMC (GreenPlum), IBM (Netezza), Cloudera, *etc.*

"Big Data": le défi de la formation

Compétences en maths/info

- Analyse sémantique
- Sécurité des données
- Systèmes répartis, parallélisation massive, clouds
- Analyse des réseaux sociaux, "graph-mining"
- Visualisation, interface homme-machine
- Ranking, moteurs de recommandation/recherche
- Machine-learning "distribué", en ligne, passage à l'échelle

Compétences en droit/économie

- Propriété des données personnelles
- Création de valeur, économie de l'internet

"Big Data"

Le phénomène "Big Data"

- appelle de nouvelles offres de formation
- pose de nouveaux problèmes scientifiques/technologiques/juridiques

Répondre à ces besoins requiert

- une collaboration étroite avec le monde de l'industrie et des services
- une mutualisation des moyens/forces

Actions dans le domaine de la formation

Mastère Spécialisé "Big Data" - ouvert en sept. 2013 30 stagiaires

- Multidisciplinarité: programmation, données, sécurité, visualisation, maths appliquées, droit, économie/business
- Recrutement: niveau M2
- Projets "Fil Rouge" tutorés par des professionnels (e.g. Capgemini, GdF Suez, EDF, Xebia, start-ups de l'incubateur de TPT), séminaires (Thalès, McKinsey, NextPerformance, *etc.*)
- Comité de perfectionnement: EADS, Thalès, IBM, Google, Critéo, Liligo, Safran, SAS, Capgemini
- Objectif: un savoir-faire opérationnel
- Stage et thèse professionnelle $\sim$ 6 mois

La formation - Projets "Fil Rouge" 7 mois

- travail en groupe (par ex. 4 à 5 stagiaires)
- vise à explorer un thème du Big Data à plusieurs facettes
  e.g. infrastructure, analyse, exploitation, mise en production, droit, business model
- recherches bibliographiques, étapes de modélisation, développement et expérimentation
- projets proposés et tutorés par des professionnels
- suivi régulier assuré par des enseignants-chercheurs
- Infrastructure Hadoop sur un cluster propriétaire, utilisation d'AWS

Actions dans le domaine de la formation

Certificat "Data Science" - prévu au printemps 2014

- Cible: professionnels en situation d'emploi
- 24 jours de présentiel
- En développement: plateforme d'e-learning

Formation Continue

- Stages ciblés
- Ex: machine-learning, valorisation des données personnelles, l'écosystème Big Data

Actions dans le domaine de la recherche

Partenariats avec l'industrie

- Chaires industrielles:
  - Réseaux sociaux - Télecom Ecole de Management
  - Valeurs et politiques des informations personnelles - Télecom ParisTech, Télecom Sud Paris & Télecom Ecole de Management
  - Machine-Learning & Big Data - Télecom ParisTech
  - Big Data for e-commerce - Télecom ParisTech
- Contrats bilatéraux (e.g. EADS, Orange, StreamWide)

Partenaires académiques

- IMT: Mines (Bio-info), TSP, Télecom Ecole de Management
- Au sein de ParisTech: X (CMAP), ENPC (Cermics), ENSAE (Crest)
- Saclay: Université Paris Orsay (LRI), ENS Cachan (CMLA), CEA, INRIA

Infrastructure - Partenariat IMT/GENES: TERALAB

Une plateforme IMT/GENES pour la recherche et l'enseignement