Mastère Spécialisé "Big Data"

Stéphan Clémençon

Télécom ParisTech

July 1, 2013

Agenda

- Contexte et Opportunité
- Les grandes lignes du programme de formation
- Recrutement - Partenariats

"Big Data":

Le contexte

- Accumulation de données **massives** dans de nombreux domaines:
  - Biologie/Médecine (génomique, métabolomique, essais cliniques, imagerie, *etc.*
  - Grande distribution, marketing (CRM), e-commerce
  - Moteurs de recherche internet (ex: contenu multimedia)
  - Réseaux sociaux (Facebook, Tweeter, ...)
  - Banque/Finance (risque de marché/liquidité, accès au crédit)
  - Sécurité (ex: biométrie, vidéosurveillance)
  - Administrations (Santé Publique, Douanes)
  - Risques opérationnels

- Un **déluge de données** qui rend inopérant:
  - les outils basiques de gestion de base de données
  - le prétraitement reposant sur l'expertise humaine (indexation, modélisation, intelligence décisionnelle)

"Big Data"

Les besoins

- Le phénomène "Big Data" requiert des techniques récentes/nouvelles pour:
  - La **collecte** et le **stockage** des données
  - La **recherche** automatique d'objets
  - Le **partage** des données ("cloud", fichiers distribués)
  - L'**analyse** (prédictive) et la **visualisation** de données massives
- Des avancées spectaculaires en informatique et en mathématiques appliquées ("machine-learning") en réponse à cette évolution
- "Big Data", un moteur pour la technologie, l'économie:
  - Moteurs de recherche, moteurs de recommandation
  - Marketing viral
  - Détection des fraudes
  - Médecine individualisée
  - Publicité en ligne
  - *etc.*

"Big Data"

Les cibles

- De nombreux secteurs d'activité sont concernés et susceptibles de recruter des ingénieurs/techniciens avec ce type de compétences, parmi lesquels:
  - (e-) Commerce (ex: Carrefour, Amazon)
  - Web (ex: Google)
  - Santé (ex: Merck, Glaxo, Novartis, SmithKline)
  - Sécurité (ex: Safran, Thalès)
  - Banque/Finance (ex: BNPP)

Perspectives:

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

- 140 000-190 000 recrutements de "data scientists"
- 15 millions de cadres avec des connaissances générales dans ce domaine

Les atouts de Télécom ParisTech

- Une **masse critique** d'enseignants-chercheurs compétents sur le plan académique
  - Administration de bases de données
  - Indexation, compression
  - Prédiction
  - Visualisation
- Une capacité reconnue à **intégrer les contraintes industrielles**
  - Projets collaboratifs avec l'industrie
  - Valorisation, brevets

Le Machine-Learning à Telecom ParisTech

Un thème omniprésent en recherche et dans les applications

- Applications en Traitement du Signal/Image, Réseaux, Communications Numériques, Sociologie
- **Nombreux spécialistes** du domaine: en particulier au sein des départements TSI (STA, TII, AAO et MM) et Infres (IC2)
- Une **compétence originale**: les méthodes d'apprentissage combinées à un savoir expert dans les domaines du traitement du signal, des réseaux, permet de traiter des **données structurées** (données dépendantes, séries temporelles, graphes par ex.)
- **Des équipes reconnues**: évaluées A+ par l'AERES, médaille d'argent du CNRS (E. Moulines), participation aux comités scientifiques des conférences majeures

Le Machine-Learning à Telecom ParisTech

Une couverture thématique très large

- **Graph-mining**: ANR Systèmes Complexes "Viroscopy", Futur & Rupture "Graphes et Réseaux"
- **Filtrage Collaboratif**: Projet Digiteo "Bemol"
- **Apprentissage par renforcement**: CRE avec Orange Lab (CRM, "Channel-sensing"),
- **Détection d'anomalies**: ANR-RNRT "Oscar"
- **Sélection de variables:** Cifre Renault Technocentre
- **Apprentissage non supervisé**: ANR "Tangerine"
- **Données structurées**: Cifre Exane, Cifre Natixis, Cifre Findworks Tech.
- **Ranking**: Cifre Renault Technocentre, Futur & Rupture "MetaRank"

Brevets

Logiciels TopRank, TreeRank, ...

Le Machine-Learning à Telecom ParisTech

Nos partenaires académiques

- Au sein de ParisTech:
  - Polytechnique (CMLA)
  - Mines (Centre de Bio-informatique)
  - ENPC (Cermics)
  - ENSAE (Crest)
- INRIA/ENS Ulm (SIERRA)
- CEA LIST (LIMA)
- UPMC (Lip6) - Groupe d'Intérêt Scientifique PARISTIC
- Université Paris Sud (LRI)

Co-animation du séminaire SMILE

Le Machine-Learning à Telecom ParisTech

Enseignement: une compétence reconnue

- Cours de Machine-Learning en Masters, co-habilités par Telecom ParisTech:
  - Modélisation Aléatoire (Université Paris 7),
  - Mathématiques Vision Apprentissage ENS Cachan
- Ensae ParisTech (Cours de 3ème année)
- ENPC ParisTech (Cours de 2ème année)
- **Un parcours dédié** à Telecom ParisTech: Apprentissage Fouille de Données et Applications
- Un cours de **Web-Mining** pour la formation des Ingénieurs du Corps Mines-Télécom
- Une **Formation Continue** destinée aux chefs de projet, ingénieurs techniciens

Compétences en Machine-learning: une demande croissante depuis 10 ans

- Nombreux débouchés professionnels (secteurs high-tech, biotech, marketing, finance, sécurité, *etc.*) dans l'Industrie (PME innovantes, grands groupes)
- Un domaine clef de l'innovation technologique (Carnegie Mellon, Waterloo, CalTech, MIT)
- Une offre aujourd'hui insuffisante en (Ile de) France

Les forces de Telecom ParisTech

- Une équipe de recherche performante
- Une **masse critique** de chercheurs (en comparaison des autres acteurs en France), permettant un rayonnement international
- Une expérience avérée dans le domaine des applications industrielles
- Une implication significative dans la formation (cours de spécialité)
- Des partenaires: X, Mines ParisTech, Ensae ParisTech, ENPC ParisTech, ENSC, Lip6 (UPMC)

Les objectifs de la chaire

- Mutualiser les expériences (industrielles et académiques), faire émerger des innovations
- Permettre un financement plus pérenne des activités de recherche en machine-learning, sur des sujets "en amont"
- Création d'un Mastère Spécialisé de référence "Big Data"
- Accroître la visibilité de la discipline auprès des étudiants (de ParisTech)

Les formations "concurrentes" : un bref tour d'horizon

En (Ile de) France

- Master informatique (Paris 6, Paris 7, ...)
- ESC Grenoble - MS "Business Intelligence"

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
- 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.*

Mastère Spécialisé "Big Data"
Les grandes lignes du programme de formation

Quatre volets

- Gestion/administration de bases de données massives (ex: web, cloud)
- Outils d'indexation, de recherche, de compression (ex: textmining, données multimedia), Visualisation
- "Business intelligence", apprentissage automatique
- Aspects juridiques et "business"

Animation

- Projets,"fil rouge", groupes de travail (en collaboration avec nos partenaires industriels)
- Séminaires avec des spécialistes du domaine

Recrutement - Effectif - Partenariats

Filières de recrutement

- Etudiants à la sortie d'un Master (informatique, modélisation statistique)
- Ingénieurs
- Salariés: cadres, chefs de projet, techniciens

Effectif: 20-25 stagiaires

Comité de Perfectionnement

- Google, BNP, Safran, Thalès, EADS, EDF, Criteo, Amazon, Liligo, IBM, Capgemini, SAS, PSA ...