
INSTITUT DES
ACTUAIRES

Formation DSA

Data Science pour l'Actuariat

Jérémie Jakubowicz

Direction des Etudes

Prix spécial

« Innovation formation digitale en Assurance »

de l'Université de l'Assurance 2016

OBJECTIF DU PROJET

- **Doter l'actuaire de la maîtrise du panel des nouveaux outils du monde de la data science :**
  - Extraction
  - Stockage
  - Filtrage
  - Analyse prédictive
  - Visualisation
- **Sensibiliser l'actuaire à communiquer avec l'ensemble des métiers de l'assurance à travers le vecteur des données.**
- **Identifier les enjeux économiques :**
  - Valeur client
  - Tarification personnalisée
  - Prévention ciblée des risques dans leur cadre juridique et déontologique

MODALITES DE LA FORMATION

- 12 mois de formation (à partir de mars) pour les 2 premières promotions, à compter de la 3ème promotion, début Janvier 2017
- 168 heures de formation, dispensées à temps partiel sous forme de sessions mensuelles de 2 jours consécutifs
- Les interventions sont assurées par des professionnels français et internationaux, universitaires et entrepreneurs

PROGRAMME DE LA FORMATION 1/2

- Eléments logiciels et programmation Python,
- Datamining et programmation R,
- Algorithmique du Machine Learning et participation à des concours de datascience (dont un obligatoire),
- Fondements théoriques de l'apprentissage statistique,
- Machine Learning distribué et applications,
- Extraction, utilisation, visualisation et valorisation des données.

PROGRAMME DE LA FORMATION 2/2

- Etudes de cas : application de méthodes statistiques sur données massives à des problématiques principalement actuarielles.

- Réalisation d'un projet sur un sujet actuariel, sous le tutorat d'un membre du corps professoral de la formation.

PROGRAMME

A) Éléments logiciels et programmation Python

Objectif : *Introduction au langage Python et sensibilisation aux grandeurs informatiques pertinentes.*

Éléments de programme :

- Initiation à la programmation Python
- Bibliothèque des méthodes statistiques usuelles
- Éléments logiciels pour grandes bases de donnée
- Hardware, performance machine et gestion de mémoire
- Efficacité d'un algorithme
- Complexité, accès mémoire, ordres de grandeur

PROGRAMME

B) Datamining et programmation R

Objectif : *Présenter les outils classiques d'exploration de données, sous un angle essentiellement descriptif. Ce module permettra une uniformisation du niveau en R.*

Éléments de programme :

- Manipuler des données sous R: données continues, facteurs (recodification), dates, heures
- Bases de la programmation avancée en R
- Méthodes non-supervisées
- Analyse factorielle et détection de clusters

PROGRAMME

C) Algorithmique en Learning & concours Kaggle

Objectif : *Approche par mise en situation via la participation à un concours type Kaggle.*

Éléments de programme :

- Exploration/ Sélection / Transformation / Nettoyage des données
- Principaux algorithmes de Machine Learning
- Procédures de validation / sélection de modèle
- Visualisation
- Retour d'expérience et analyse des résultats d'un Kaggle

PROGRAMME

D) Fondements théoriques du Machine learning

Objectif : *Présenter les fondements mathématiques des principaux algorithmes de Machine Learning*

Eléments de programme :

- Théorie de la décision, Perte, risque
- Machine Learning, Méthodes paramétriques, perceptron
- Classification, Convexification du risque, boosting et SVM
- Méthodes ascendantes et descendantes, Critères AIC et BIC
- kNN, Random Forest, Bagging, Arbres de décision

PROGRAMME

E) Machine Learning distribué et applications

Objectif : *Pour passer à l'échelle, les algorithmes de Machine Learning doivent être repensés. Un cadre efficace est celui des* **algorithmes distribués** *où on utilise plusieurs entités de calculs pour mener à bien l'objectif initial.*

Éléments de programme :

-Principes généraux de la distributions des algorithmes

-Le cas de Map-Reduce

-Applications en Machine Learning

PROGRAMME

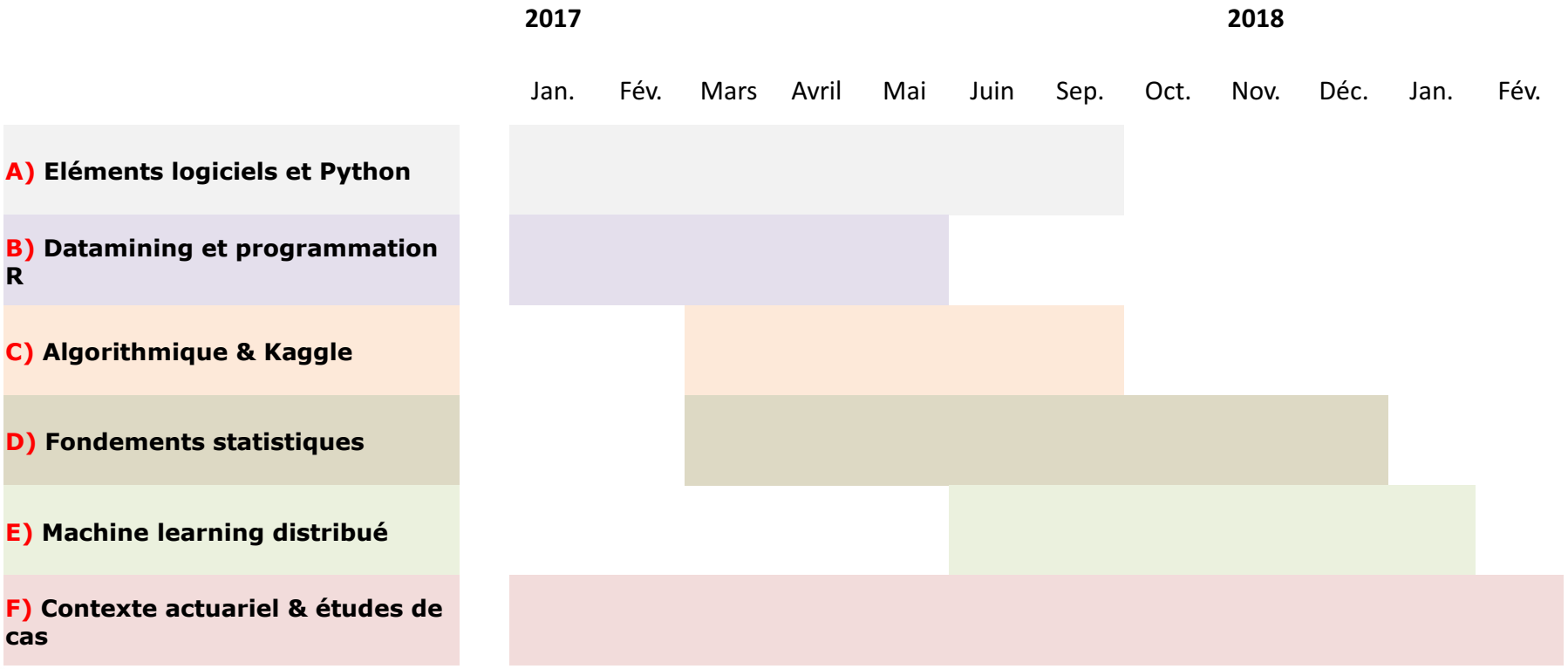
F) Contexte actuariel et études de cas

Objectif : *Présentations de sujets « métier » et mises en situation via des interventions courtes et techniques sur des thématiques précises.*

Exemple :

- Cartographie et GPS
- Géolocalisation et anonymisation
- Parallélisation massive pour la simulation Monte Carlo
- Health monitoring
- Visualisation de données et réseaux sociaux
- Applications en Génomique
- Investissement séquentiel en gestion de portefeuille

Organisation temporelle des modules



Corps professoral

Equilibre entre profils académiques et professionnels

Corps Académique:

- *Ecole Polytechnique, MIT, Sorbonne, Télécom Sud paris, ENSAE, CNRS…*

Corps professionnel:

- Du monde des données: *CNIL, Microsoft, Dataiku, Critea, Kamaleoon, Datarobot, Quantcube, Teralab …*

- Du monde actuariel et financier: *Milliman, COVEA, Deloitte, Ernst and Young, Advestis, Capital Fund Management, BNP, Cardiff …*

Réalisation d'un projet

- Les participants à la formation réaliseront un projet visant à exhiber **l'apport de ces nouvelles méthodologies statistiques et informatiques** pour la modélisation d'un **phénomène actuariel**
- Projet réalisé sous le **tutorat** d'un membre du corps enseignant de la formation
- Projet réalisé sur la deuxième moitié de la formation
- Rédaction d'un rapport et soutenance devant un **jury de membres du corps enseignant** ainsi que l'ensemble de la promotion *(sauf conflit trop important pour cause de confidentialité)*.

Parmi les projets de la 1ère promotion

- Détecter le projet d'achat d'un véhicule de nos clients en portefeuille – Applications au marketing direct,
- Utilisation de variables géographiques pour la modélisation de la survenance de sinistres en assurance automobile,
- La France découpée grâce à la télédétection dans le cadre de l'offre d'assurance climatique « prairie »,
- Prévision de sinistres sur les garanties ITT,
- Eclairer la connaissance client – étudier et comprendre la dynamique de multi-équipement

PUBLIC CIBLE

- Actuaires membres de l'Institut des actuaires

- non membres de l'Institut des actuaires, professionnels de l'assurance et de la finance ayant un niveau Bac + 5 (formation à dominante mathématiques et informatiques) avec une expérience professionnelle

FORMATION DSA

Unicity (ε)

Developed specifically for big/meta-data

Unicity evaluates the percentage of individuals in a dataset that would be uniquely identified by p points taken at random from their trace

Gives you:

1) The likelihood of re-identifying one person in a data knowing p points about him

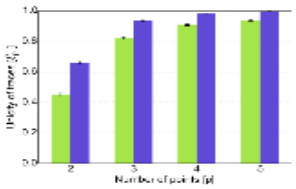
2) The approximate number of points needed to reconcile two datasets

Estimating Unicity (ε)

Algorithm S2 Unicity Estimation(p)

```
users ← select(D, p, 10000)
for u ∈ users do
    I_p ← draw(u, p)
    is_unique ← true
    for x ∈ D \ {u} do
        if I_p ⊂ x.trace then
            is_unique ← false
            break
    if is_unique then
        uniqueUsers ← uniqueUsers + {u}
return |uniqueUsers|/|users|
```

Unicity: credit card data



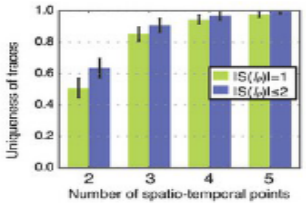
1 million people over 3 months

Points: shop and day

ε_4 = 90%

(blue bars)

Unicity: mobile phone data



1.5 million people over a year

Points: approximate place and time (~1km² with resolution of an hour)

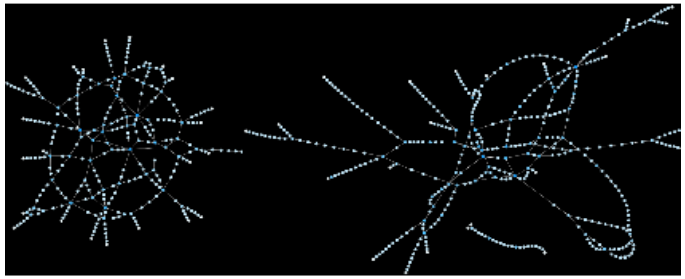
ε_4 = 95%

(green bars)

Anonymisation, Y.A. de Montjoye, MIT Media Lab

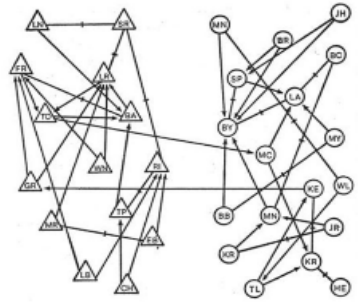
FORMATION DSA

VISUALISATION DE DONNÉES ET ANALYSE DE RÉSEAU - MARTIN GRANDJEAN
Penser en réseau : différentes conceptions



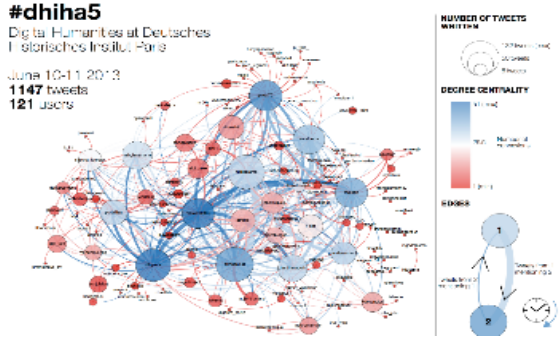
VISUALISATION DE DONNÉES ET ANALYSE DE RÉSEAU - MARTIN GRANDJEAN
Penser en réseau : différentes conceptions

Sociogramme d'une classe d'élèves de 11-12 ans
(critère : s'asseoir à côté des élèves choisis – 2 choix au maximum)

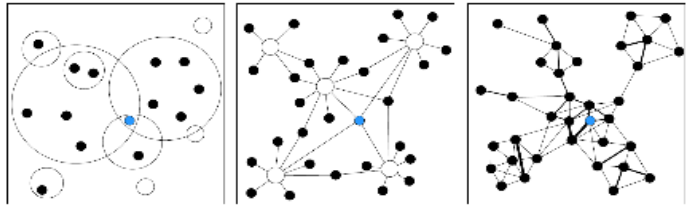


Moreno (1934)

VISUALISATION DE DONNÉES ET ANALYSE DE RÉSEAU - MARTIN GRANDJEAN
Penser en réseau : différentes conceptions



VISUALISATION DE DONNÉES ET ANALYSE DE RÉSEAU - MARTIN GRANDJEAN
Penser en réseau : différentes conceptions



Réseaux, M. Grandjean, Université de Lausanne

FORMATION DSA

INSTITUT DES ACTUAIRES

Où sont les données $(x_i, y_i)_{1 \leq i \leq n}$?

Nous avons vu lors du précédent cours, deux scénarios dans lesquels les données se retrouvaient stockées de façon distribuée :

- Le cas des systèmes de fichiers distribués (par ex : HDFS)
- Le cas des `Resilient Distributed Datasets` (RDD) de Spark

Optimisation distribuée : la fonction objectif

On a vu qu'un nombre important d'outils du ML s'écrivent comme la solution d'un **problème d'optimisation** :

$$\arg\min_{\theta} \sum_{i=1}^{n} f_i(\theta)$$

où f_i est une fonction qui dépend de la donnée (x_i, y_i). On peut **regrouper les données** qui dépendent d'un **même noeud de calcul** et le problème devient :

$$\arg\min_{\theta} \sum_{j=1}^{J} g_j(\theta)$$

où $g_j(\theta) = \sum_{i \in I_j} f_i(\theta)$. En d'autres termes, la *maille* considérée est celle de l'**entité de calcul** et pas celle de la **donnée unitaire**.

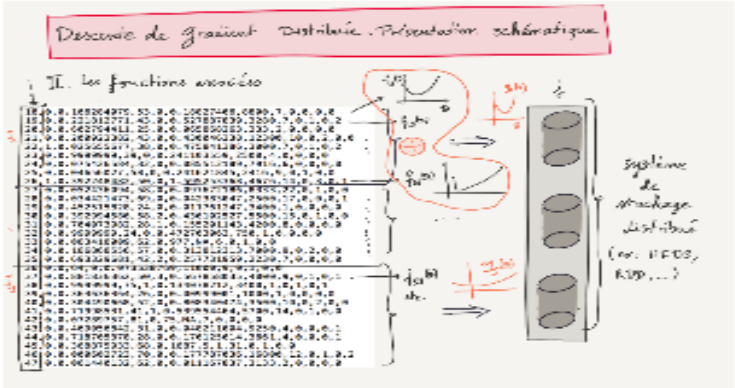
Optimisation distribuée : descente de gradient distribuée

Compte-tenu du contexte distribué (que ce soit via HDFS ou les RDD), l'algorithme de descente de gradient :

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \sum_{j=1}^{J} \nabla g_j(\theta_n)$$

s'écrit naturellement comme une itération de jobs Map Reduce.

1. Le client envoie θ_n au noeud maître
2. Le noeud maître diffuse la vecteur θ_n au noeuds travailleurs (étape de broadcast).
3. Le noeud maître lance les jobs :
   - 3.1 `MAP` : Calcul de $m_j = \nabla g_j(\theta_n) = \sum_{i \in I_j} \nabla f_i(\theta_n)$
   - 3.2 `REDUCE` : Calcul de $\sum_{j=1}^{J} m_j$



Parallélisation, JJ

FORMATION DSA

Le package ff (fast access files)

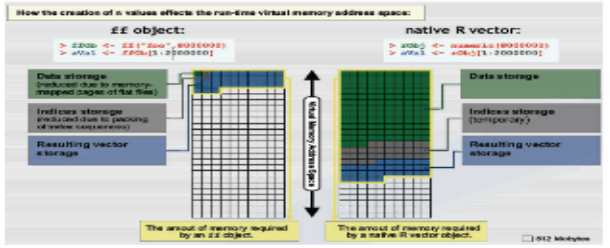
- Le package ff stocke les données sous forme de fichiers plats binaires sur le disque dur
- Sur un répertoire :
  - affiché : `getOption("fftempdir")`
  - modifié : `options(fftempdir = "D:\\Data\\ff")`
- Choisir un répertoire temporaire contenant suffisamment d'espace disque
- Elles sont traitées comme si elles étaient dans la RAM par un mécanisme « chunk-wise »
- ff charge en RAM les données au fur et à mesure de leur utilisation dans les calculs, en conservant les résultats intermédiaires et en restituant à la fin le résultat définitif, tel qu'il aurait été obtenu de façon classique par un calcul entièrement en RAM

Le package ff (suite)

- Fonction ff pour l'ouverture et la création de fichiers
- Les variables peuvent être de type numerical, integer, boolean ou factor
- Entrées/sorties par les opérateurs habituels `<-` et `[ ]`
- Conversion possible d'une matrice en dataframe par `as.ffdf`
- Import d'un jeu de données par `read.table.ffdf` ou d'un fichier csv par `read.csv.ffdf` ou `read.csv2.ffdf`
- Manipulation un peu difficile des objets ff et ffdf avec du code « chunk-wise » (du type `for (i in chunk(x)) { xxx }`)
- Différences avec bigmemory :
  - ff disponible sur Windows
  - données toujours stockées par ff sur disque dur
  - ff gère des objets de type numérique, entier, logique, facteur, matrice, et crée des objets ffdf de type data frame, alors que bigmemory ne gère que les objets matrices
  - ff non comparible avec BLAS et LAPACK
  - pas de code « chunk-wise » à écrire pour bigmemory

Principe du package ff

- Gère une couche basse en C++ et une couche haute en R



Utilisation conjointe de ff et ffbase

- Exemple :

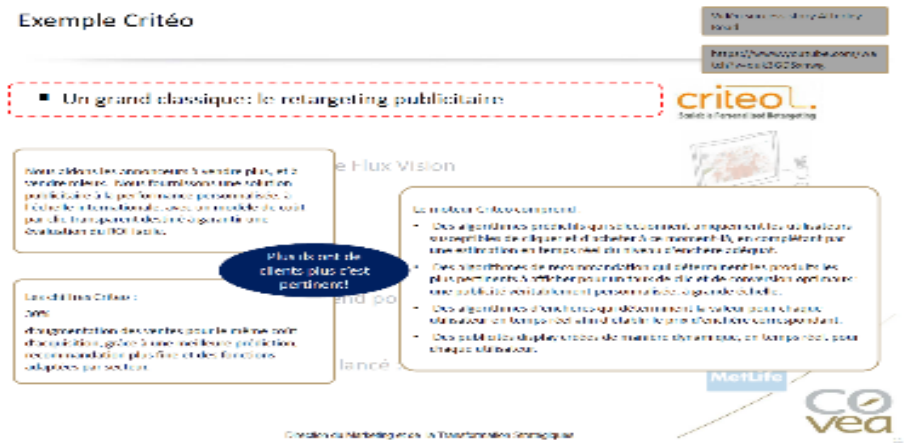
```
> x <- 1:1000000
> object.size(x)
4000040 bytes
> mean(x)
[1] 500000.5
> library(ff)
> x_ff <- ff(x)
> object.size(x_ff)
2984 bytes
> mean(x_ff)
[1] NA
Message d'avis :
In mean.default(x_ff) :
  l'argument n'est ni numérique, ni logique : renvoi de NA
```

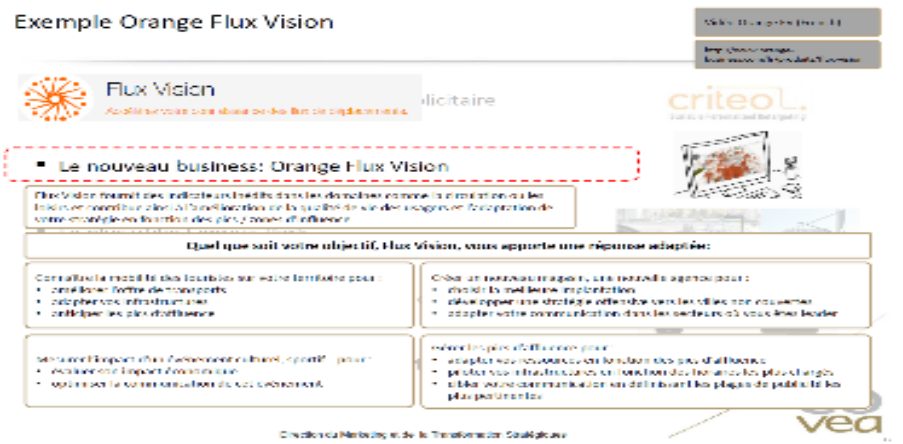
- Avec le package ffbase :

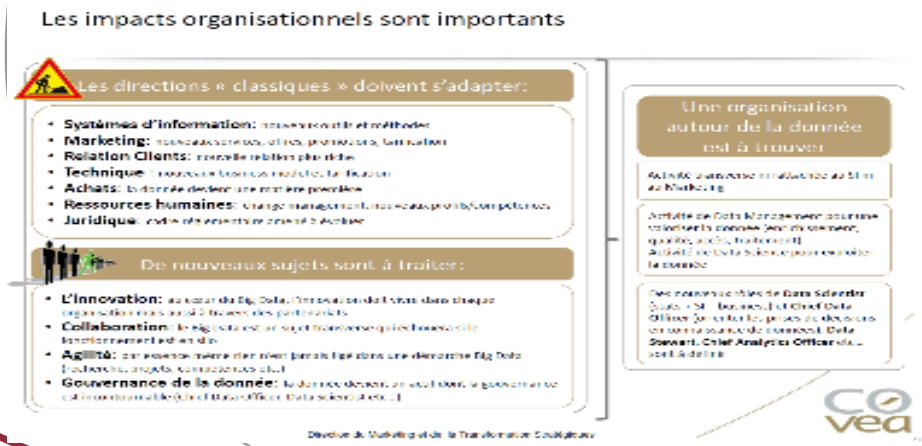
```
> library(ffbase)
> x_ff <- ff(x)
> mean(x_ff)
[1] 500000.5
```

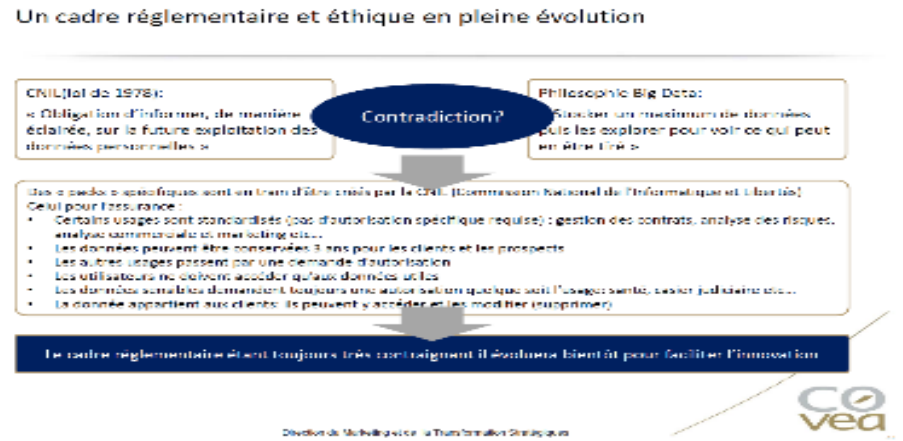
R & Data Mining, S. Tufféry, Crédit Mutuel & Université de Rennes 1

FORMATION DSA









Data et Assurance, B. Beaume, Covéa

Gouvernance de la formation

INSTITUT DES ACTUAIRES

Direction des Etudes:

Arthur Charpentier (Actuaire, Prof. Stat. Rennes et UQAM)
Romuald Elie (Actuaire, Prof. Maths, Université Paris-Est et ENSAE)
Jérémie Jakubowicz (Docteur Maths, Prof Télécom Paris Sud et École Polytechnique)

Comité Scientifique:

Michel Bois (Actuaire, DSI CNP, Membre du Comex CNP)
Renaud Dumora (Actuaire, DGA Paribas BNP Cardif)
Philippe Marie-Jeanne (Actuaire, Directeur Lab Big Data Innovation AXA Group)
Françoise Soulié-Fogelman (ENS, Consultante, ex KXEN)
Olivier Sorba (Actuaire, CRO Groupe Lagardère)
Marc Hoffmann (Professeur Stat Dauphine et X, Chaire Big Data Havas ILB)
Florence Picard (Actuaire, Commission Scientifique Institut des actuaires)

Sous le parrainage de la
Commission Scientifique de l'Institut des Actuaires

Points forts de la formation

INSTITUT DES ACTUAIRES

- Formation spécifiquement **dédiée aux actuaires**
- En prise directe avec les **besoins des entreprises**
- Mixant théorie et pratique pour optimiser l'**efficacité opérationnelle**
- Un contenu pédagogique validée par un **Comité Scientifique de haut niveau**
- Une direction des études à **double compétence: statistiques et informatique**
- Un **contrôle** de l'acquisition des connaissances
- La **réalisation d'un projet** sur un sujet actuariel encadré par un **tuteur**
- Recrutement de la 4ème promotion
- **Prix spécial « Innovation formation digitale en Assurance »** de l'Université de l'Assurance 2016

Promotion actuelle

- 24 stagiaires (dont 25% de femmes)
- Profil actuariel avec forte sensibilité informatique
- Entreprises représentées:

Axa, Groupama, MMA, Pacifica, Crédit Agricole, MAAF, CCR, GMF, Alptis, PricewaterhouseCoopers, Forsides, Actuaris, Optimind-Winter, ACPR, Mercer, CNP, BAO SCOR, Fixage, Banque Postale, Aviva, Milliman, PwC, Generali, BNP, Santiane, Allianz, AG2R…