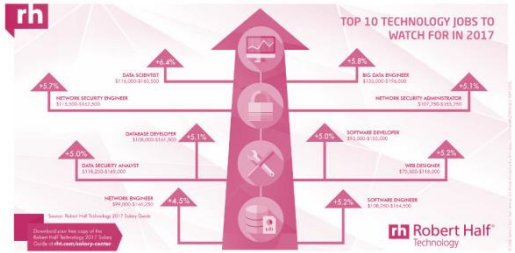
THALES

le métier de data scientist

INSTITUT DES ACTUAIRES

janvier 2017

catherine gouttas

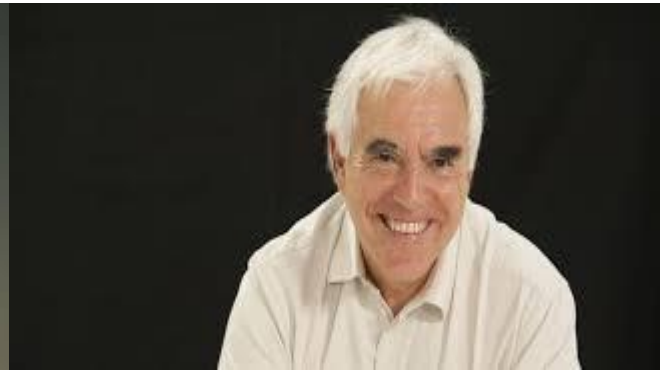


sommaire

- une petite introduction
- le laboratoire Big Data de Thales
- le métier de Data Scientist
- propositions d'accompagnement
- questions & réponses

pour introduire, la leçon inaugurale …





Prononcée solennellement par chaque nouveau professeur, la Leçon inaugurale est à la fois la description de l'état d'une discipline et la présentation d'un programme de recherche

à lire, à entendre absolument

Yann Le Cun Leçon inaugurale Collège de France sur l'Apprentissage profond

Josselin Garnier Leçon inaugurale Chaire Data Science X et AXA pour l'Assurance

Françoise Soulié-Fogelman Leçon inaugurale de l'Institut des Actuaires

pour introduire, la bio

- 1990 : après un DEA de Linguistique computationnelle, intègre le Centre Scientifique d'IBM France comme ingénieur de recherche en Linguistique computationnelle, détachée au Laboratoire de développement de Lidingo d'IBM Suède pour implémenter de grammaires d'analyse et de génération, pour l'interrogation en Langage Naturel des bases de données relationnelles

- 1993 : intègre le Centre Européen de Mathématiques Appliquées d'IBM France ; ses travaux et missions portent essentiellement sur l'analyse syntaxique automatique, le traitement des interactions clients/entreprise, l'analyse des opinions et la fusion de données structurées et non-structurées

- 2000 : se rapproche des besoins clients et intègre la Practice Customer Relationship Management Solutions d'IBM Business Consulting Services en tant que consultant certifié BP

- 2004 : rejoint le groupe Thales en tant que manager de projets R&D en 2004.

- 2006 : responsable du laboratoire Big Data de Thales et co-responsable du Laboratoire CLEAR – laboratoire joint Thales / LIP6-UPMC

pour introduire, les motivations

- **Issue d'une famille française traditionnelle dans laquelle les filles étaient vouées aux sciences molles et les garçons aux sciences dures**
- **khâgneuse dans les années 80 / vouée aux métiers féminins de l'Education Nationale, attirée par la diversité des métiers qui s'offraient aux matheux, attirée par les sciences et technologies**
- **ne pas se laisser enfermer dans la représentation des genres de la société française**
- **volonté d'aller sur le terrain des hommes et de l'intelligence collective, persuadée que les femmes « sont la deuxième moitié du ciel »**
  - Modèle des jeunes filles des sociétés asiatiques (Inde / Chine) qui vivent dans des sociétés patriarcales et qui malgré tout « prennent le pouvoir » via en outre la maîtrise des sciences et des technos
- **explorer le continuum ou la complémentarité sciences molles / sciences dures / faire un pont entre ces univers**

Le Big Data « prochaine révolution industrielle » : un excellent terrain de jeu (monde encore très masculin / pluridisciplinairité / enjeux scientifiques et techniques / enjeux métier / enjeux juridiques et éthiques / nouveaux métiers, etc …)

THALES

le LABO BD&BA (Big Data et Big Analytics …) / écosystème, mission

recherche, innovation, développement, transfert technologique, éthique

THALES

le LABO BD&BA (Big Data et Big Analytics) ….

Une équipe de Data Scientists

recherche appliquée dans les domaines du Big Data, du Big Analytics et du Visual Analytics

- développement d'une plate-forme Big Data fondée sur le pattern d'architecture Lambda et sur le framework Spark
- intégration au framework Spark d'algorithmes propriétaires et Open Source
  - approches statistiques et Machine learning, avec une spécialisation dans le domaine de la détection des anomalies et de l'analyse des grands graphes
- développement d'un portail de visualisation analytique fondé sur des technologies Web

innovation orientée utilisateur, en boucle courte, en intégrant les clients dans le développement

THALES



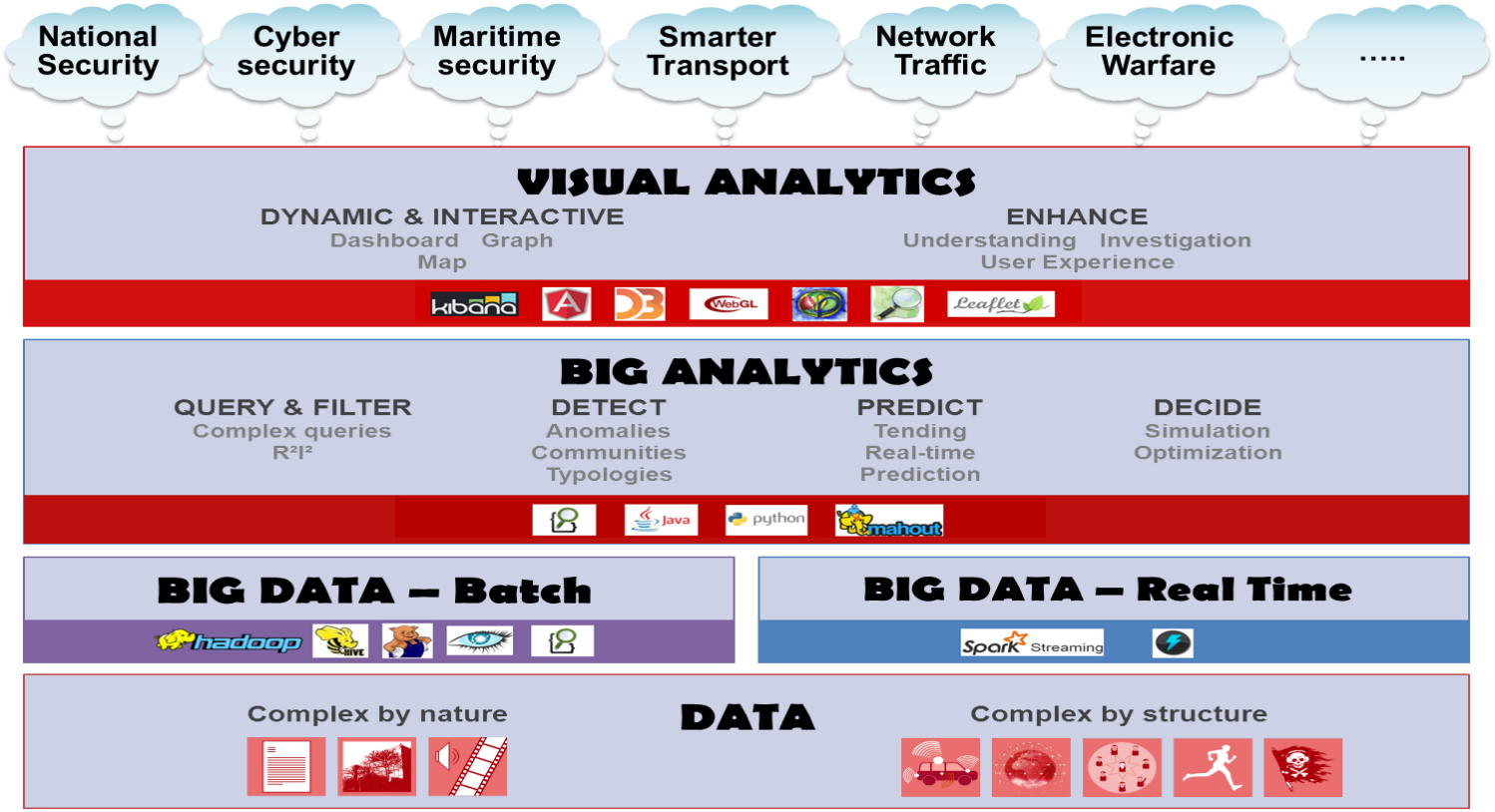
principaux domaines adressés

- *SECURITE : Cybersécurité, Cybercriminalité, Sécurité Maritime, Sécurité urbaine, Guerre électronique radio, CIP, Renseignement*
- *TRANSPORT terrestre et aérien : Mobilité, ATM, Maintenance prédictive, …*
- *SPATIAL : Maintenance*
- *des sujets transverses : par exemple, Anonymisation des données à caractère personnel, Détection et investigation des signaux faibles dans les réseaux sociaux, …*

développement et suivi de partenariats avec les laboratoires académiques

20 publications, 12 brevets en algorithmie, 4 talks aux Hadoop Summit et Spark Summit de 2015 et 2016
3 POCS Transformation digitale : Datalake Thales, Hunting, Reverse Mentoring en 2016
2 thales innovation awards (2014 & 2016)
1 thèse en Deep Learning, 1 thèse en cyber, 1 thèse en maintenance prédictive
etc …

la plate-forme



OPEN

THALES

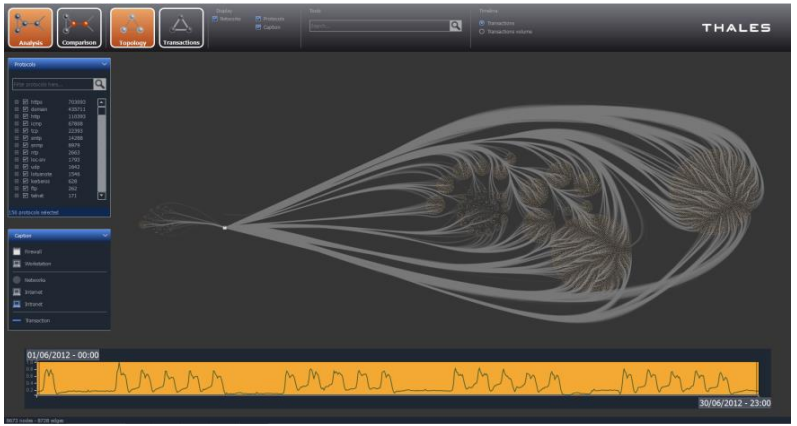
les algos

des algorithmes génériques (statistiques, machine learning et IA)

- analyse comportementale
  - segmentation/typologie
  - détection d'anomalies
- analyse relationnelle des grands graphes
- prédiction , levée d'alertes
- optimisation, aide à la décision

… adaptés aux besoins métier

la visualisation



des travaux sur les marchés fortement impactés par la transformation digitale

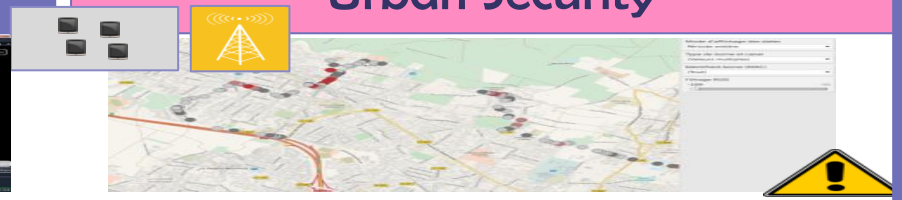
Social Network Analysis

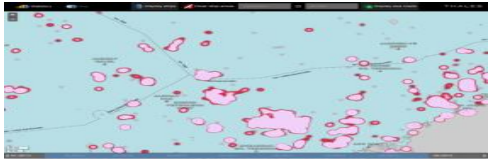
Ground Transportation

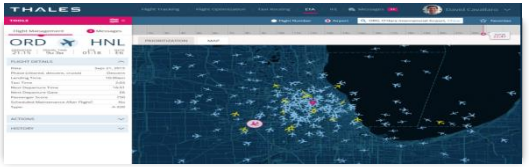
Air Transport

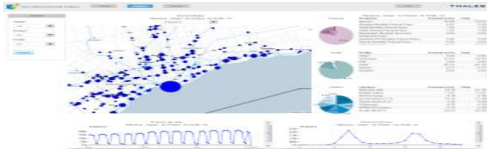
Cyber Security

Urban Security









eBorder

Electronic Warfare

Maritime Security

Predictive Maintenance

Airport Security

Communication Networks

Critical Infrastructure Protection

la Sécurité, fil d'Ariane de la transformation digitale / vers la confiance digitale / vecteur majeur pour le positionnement de Thales sur le Big Data

TROIS USE CASES POUR ILLUSTRER

1) PROTECTION DE LA VIE PRIVEE

- Anonymisation des données à caractère personnel & Lutte contre la violation des données

Accélérateur digital : Favorise la mise à disposition des data, grand frein aux chantiers Analytics en interne et en externe

2) PROTECTION DE L'ENTREPRISE

- Détection des attaques complexes

3) PROTECTION DU CITOYEN et DE LA VILLE

- Gestion et Anticipation des risques

1) PROTECTION DE LA VIE PRIVEE

anonymisation des données à caractère personnel

des besoins en matière de protection des données personnelles de plus en plus importants

BIG DATA

OPEN DATA

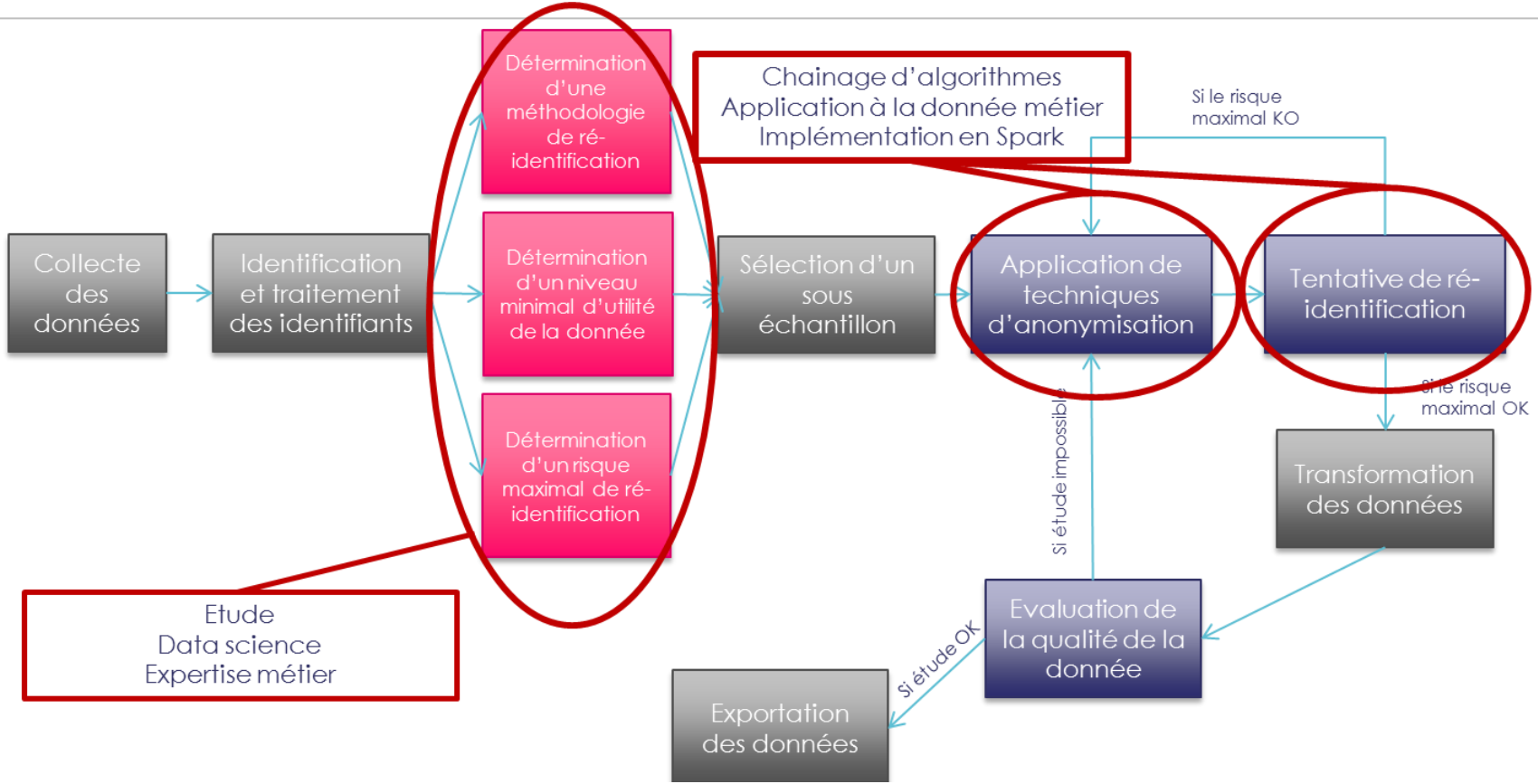
TRANSFORMATION DIGITALE

- Chantiers d'exploitation des données au cœur de la transformation digitale (data / catalyseur et produit)
- Principe de « désilotage » généralisé des données

dans un cadre juridique qui évolue : Nouveau règlement européen sur la protection des données personnelles, Loi Numérique / Axelle Lemaire, CNIL

THALES

proposition de valeur : une chaîne de traitement dédiée



appliquée aux données ENEDIS : 35M de courbes toutes les 30 minutes à la fin du déploiement de Linky

approche

ANONYMISATION

- chaînage d'Algorithmes
  - Transformation des données : Extraction d'information issues de la projection des courbes de charge en base d'ondelettes
  - Ré-utilisation de méthodes classiques (Agrégation, clustering et découpage temporel)
  - Passage à l'échelle (Spark/Scala) en particulier pour le clustering

RE-IDENTIFICATION

- algorithmes de Deep Learning (Problème vu comme de la classification d'images)
  - Modélisation de la courbe journalière sous forme d'image
  - Prédiction du client à partir d'un base de clients étiquetée

gains Client & Thales

ENEDIS

- autorisation de la CNIL / ouverture des données
- nouveaux usages, nouveaux services qui seront développés en interne et en externe
- typologie des risques de ré-indetification

THALES

- un nouveau volet du Secured by Thales avec un brevet en cours de dépôt
- des méthodes d'anonymisation formalisées et des composants algorithmiques Big Data réutilisables sur de nouvelles problématiques d'anonymisation
- passage à l'échelle et traitement du flux

2) PROTECTION DE L'ENTREPRISE

détection des attaques complexes (APT)

Jusqu'à aujourd'hui, des approches à base de règles pour détecter les comportements anormaux

détection d'anomalies

- avec des approches algorithmiques complémentaires permettant de traiter le flux
  - Copule, Clustering, Scoring et Deep Learning

détection statique de fichiers infectés

- par apprentissage génétique

détection distribuée de balayages de ports

- par détection de composantes connexes dans un cache distribué

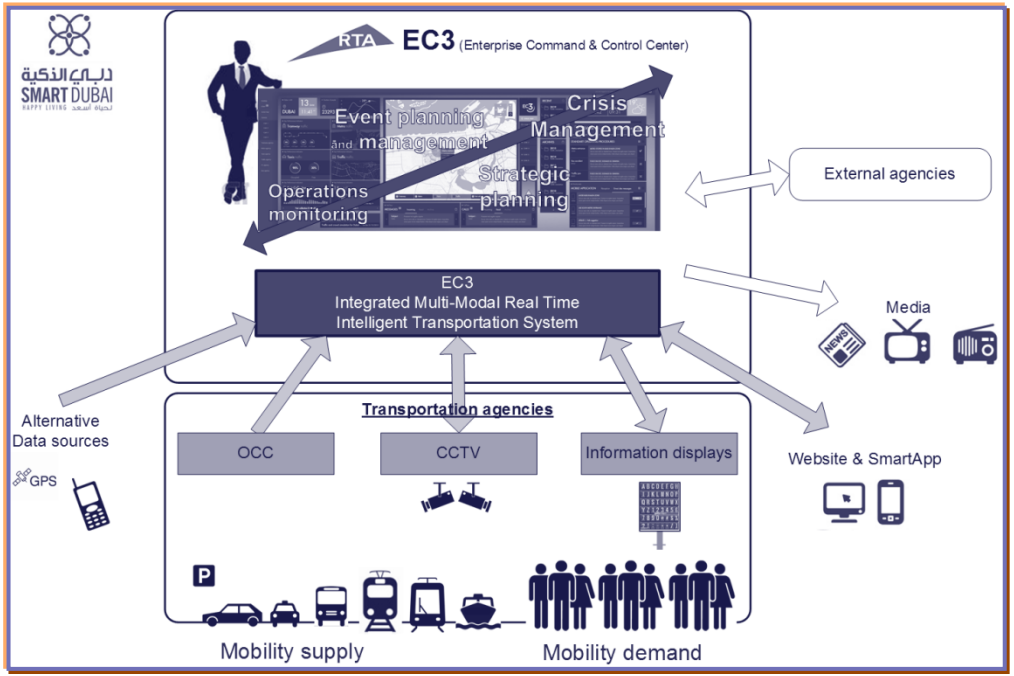
➔ intégration SOC & Sonde souveraine Thales

3) PROTECTION DU CITOYEN ET DE LA VILLE

DUBAI : SUPERVISION, PILOTAGE de la MOBILITE & SECURITE DU CITOYEN

- collecte des données de mobilité multimodales (route, tp, taxi, parking) sur l'offre et la demande

- anticipation des risques
  - événements climatiques: tempêtes de sable, inondations
  - afflux de personnes en zones sensibles
  - incendies (Ex : INCENDIE TOUR 2016)
  - etc …



هيئة الطرق والمواصلات
ROADS & TRANSPORT AUTHORITY
RTA

THALES

le labo dans le projet : Optimisation du pilotage de la mobilité et de la sécurité dans la ville

2014 & 2015 : transfert de la plate-forme Big Data du labo vers le domaine opérationnel et apport d'expertise

2016 : apport d'expertise métier et Développement

- développement d'Indicateurs multimodaux pour la levée d'alertes
- détection et Investigation des Transport Patterns pour optimiser l'anticipation des risques et la mise en place des plans de réponse
  - recherche de situations permettant de prédire les problèmes survenant dans le réseau de transport
- visualisation sur l'évolution de la densité de population dans la ville et sur l'orignine/destination des parcours des citoyens

DOMAINE BANQUE & ASSURANCE

E-Fraud Box – Fraude à la carte bancaire sur Internet

la fraude sur Internet très différente de la fraude de proximité

- fraude de proximité
  - fraude localisée dans l'espace et le temps : carte perdue ou volée avec code confidentiel,
  - mécanismes bien compris : par ex. piratage des distributeurs et duplication de la piste sur un faux support et utilisation à l'étranger…
- fraude sur Internet
  - les comportements de fraude sont diffus, vagues, mouvants et changent fréquemment
  - les origines des compromissions de données sensibles sont très diverses, beaucoup plus largement distribuées géographiquement

E-Fraud Box – Fraude à la carte bancaire sur Internet

une boîte à outils de techniques

- fouille de données, Analyse des réseaux sociaux & Informatique décisionnelle

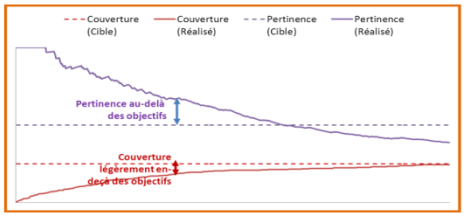
pour la détection de la fraude à la carte bancaire sur Internet

- identifier plus rapidement les cartes utilisées frauduleusement sur Internet et ainsi prévenir les porteurs de carte plus tôt

& pour l'investigation de la fraude

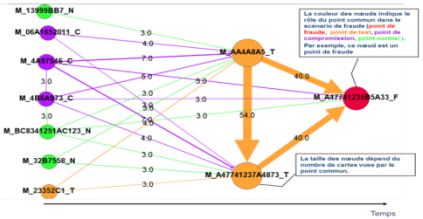
- identifier plus automatiquement des points de compromission
- détecter plus rapidement les nouveaux modes opératoires
- identifier plus rapidement les affaires pour les transmettre aux forces de l'ordre

MOTEUR DE DETECTION



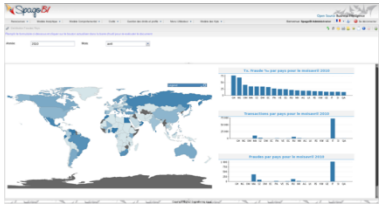
Identifier les cartes utilisées frauduleusement en garantissant de bonnes performances

MOTEUR D'INVESTIGATION



Identifier les compromissions des coordonnées bancaires

PORTAIL DECISIONNEL



Suivre et visualiser les statistiques de fraude

assurance – Etude actuarielle du cyber risque (stage avec un étudiant de l'Institut des Actuaires et la SCOR)

contexte : la cyberassurance

- un risque qui a émergé avec l'arrivée des premiers réseaux, et qui s'est développé avec Internet
- les réseaux mobiles et les objets connectés constituent l'évolution actuelle

données : Open data

- base de vulnérabilités
- base de violations de SI
- nombre d'entreprises aux USA
- Institut Ponemon

modélisation : Scoring

- évaluation de modèles

banques – Crédit agricole

web tracking : faciliter la transformation d'une visite sur le site web en demande de devis, de souscription de produits

- analyse des trajectoires des clients sur le site web du site pour identifier ce qui a amené le client à faire une demande de devis, une demande de souscription…
- algorithmes en temps réel

réseaux sociaux : analyse des réseaux sociaux pour identifier les ambassadeurs, les détracteurs de la banque

- text mining et Social mining (détection de communautés…)
- segmentations comportementales

exploitation DW : Analyse des mouvements de comptes pour identifier des motifs de contacts

- détection d'anomalies, d'atypis
- segmentations comportementales

autres pistes Assurance

modélisation du risque Santé avec enrichissement des modèles de risque existants

- utilisation d'Open data
  - météo
  - données socioéconomiques au niveau code postal : revenu, taux de chômage…
  - données professionnels de santé (via OpenStreetMap, ameli.fr…)
  - données statistiques santé (exemple SENTINELLE pour les épidémies de grippe…)
- utilisation de modèles plus performants
  - modèles d'apprentissage : Arbres de décision, réseaux de neurones…

en conclusion

- **un labo qui a introduit et lancé le Big Data chez Thales (2006)**
- **un effort soutenu et continu sur les technologies Big Data, les algorithmes, les technologies de visualisation, les nouveaux usages métier et l'éthique de la data science**
- **une volonté de se positionner sur de nouveaux domaines par exemple : la banque et l'assurance**
- **les grands enjeux 2017 : les nouvelles technos big data (BD hte perf. BD Graphes, la variété des données, le deep learning) …**
- **les difficultés majeures**
  - les jeux de données (Big Data, Small data, Fast data)
    - pas de données, pas d'algos, pas de monétisation
  - le rebouclage court entre les experts de la data science et les experts métier
  - **le recrutement, la formation et la fidélisation de data scientists de haut niveau**

LE METIER DE DATA SCIENTIST

les origines : KDD, data science, data scientist

- **la petite histoire : un terme « inventé » par deux ingénieurs de chez Facebook et LinkedIn, en 2008. Il a depuis fait école et a été élu « métier le plus sexy du XXIe siècle » par la Harvard Business Review. / pas du tout**

- **en fait : proposition de C. F. Jeff Wu en 1997 dans sa leçon inaugurale / Chaire de statistiques H. C. Carver de l'université du Michigan / renommer la statistique DATA SCIENCE et les statisticiens DATA SCIENTISTS** (Réf : A very short history of data science, www.forbes.com)

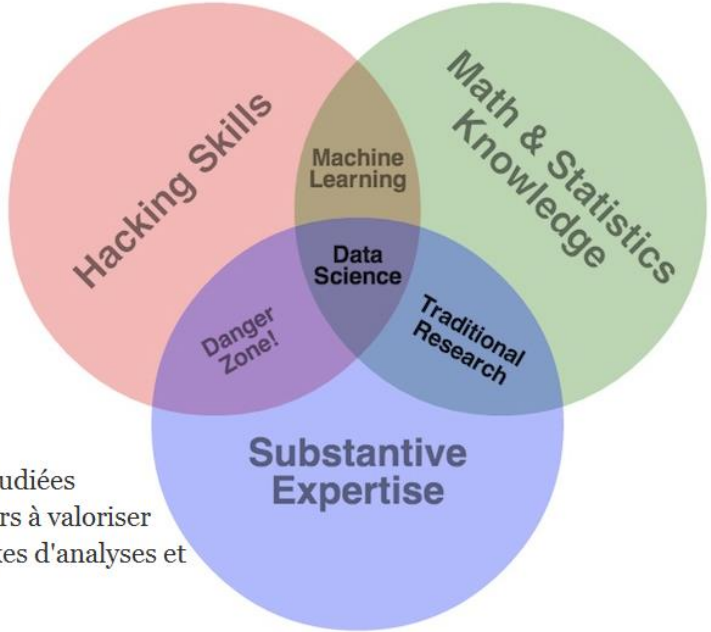
- **lancement des premières filières de formation en France en 2014** (*Serge Abiteboul, Francois Bancilhon, Francois Bourdoncle, Stephan Clemencon, Colin De La Higuera, et al. L'émergence d'une nouvelle filière de formation : " data scientists ". [Interne] INRIA Saclay. 2014. <hal-01092062>)*

THALES

le diagramme de Venn

les data, les technologies informatiques, les algorithmes, le métier

- Accéder à des bases de données
- Prototyper des applications
- Coder des visualisations
- Manipuler les technologies Big Data



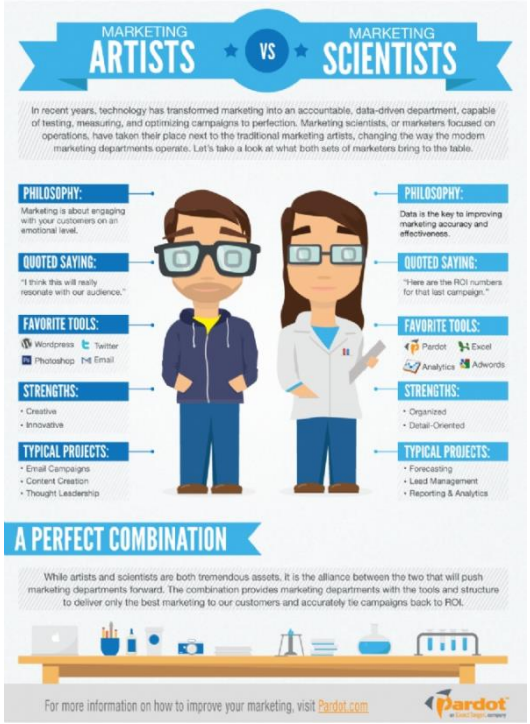
- Comprendre les modèles, leurs différences et leurs conditions d'application
- Avoir une intuition géométrique quant à leur fonctionnement
- Être capable de customiser certaines parties calculatoires des algorithmes

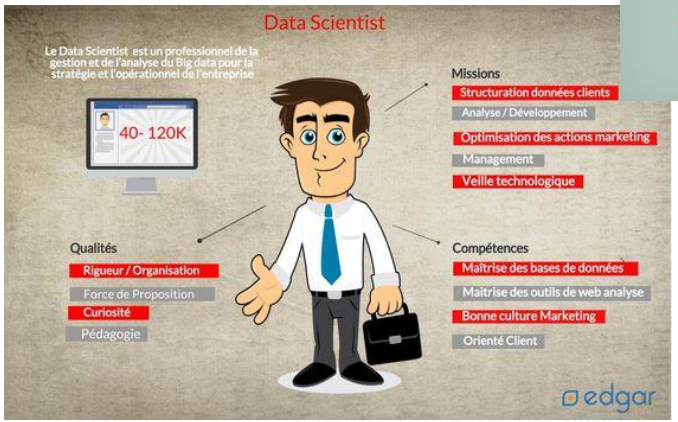
- Connaitre les données étudiées
- Maitriser les cibles métiers à valoriser
- Imaginer de nouveaux axes d'analyses et variables explicatives

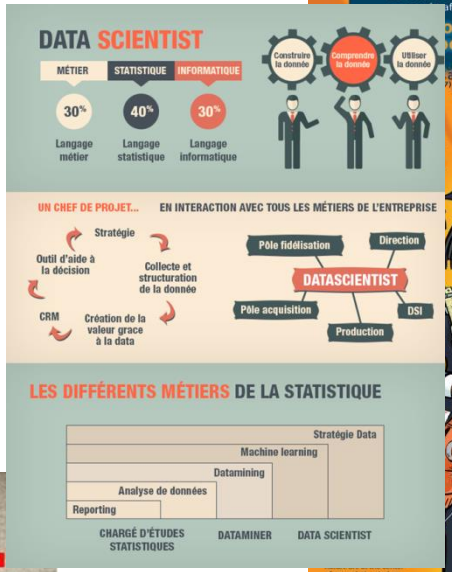
et les enjeux étiques : respect de la vie privée, loyauté des algorithmes

THALES

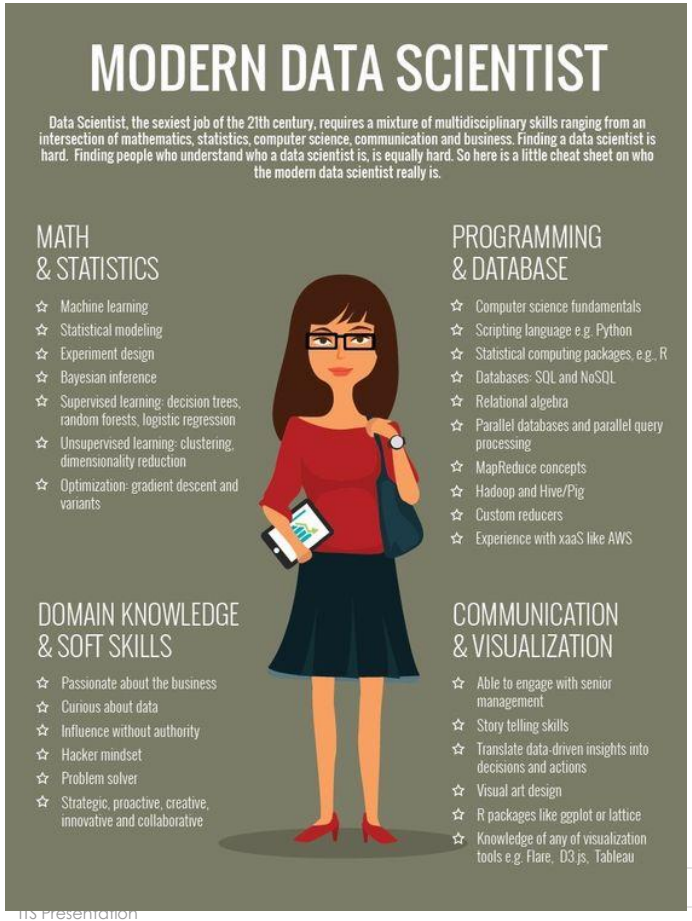
des infographies qui cherchent à caractériser le métier







ma préférée (les femmes et une tentative d'élargissement du champ de travail / toujours rien sur l'éthique!)



http://www.decideo.fr/La-premiere-femme-Data-Scientist-de-France-Irene-Balmes-se-trouve-chez-Kynapse-Consulting-Innovation-Big-Data-d-Open_a8584.html

paroles de data scientist

Benoit=Data scientist
Age=28
S=Masculin
F=ENSAI
Depuis 5 ans au labo
En partance pour Thales US

*« Un data scientist est **un matheux et un développeur**. Il cherche à répondre à un problème métier donné en maximisant un **compromis entre performances algorithmiques et performances d'implémentation** (l'implémentation doit être capable de traiter X gb de données en X secondes ou le système doit supporter x données par secondes…).*
*Les compétences techniques ? Je dirais qu'il faut avoir un **bon background mathématique (statistique machine, learning, etc…)** et les capacités de les faire sortir des logiciels d'analyse de données. Il faut aussi être capable de **comprendre le problème métier** pour savoir quels critères permet de **gagner de l'argent**.*
*Le data-scientist à la différence d'un statisticien ne produit pas uniquement un rapport mais **des algorithmes qui visent à être intégrés dans un contexte de production**.*
*Sur l'aspect psychologique je dirais qu'il faut une **très bonne capacité d'écoute et de compréhension**. Il faut aussi avoir une bonne **confiance en soi** parce qu'il faut présenter à des gens qui n'y croient pas forcément la façon dont la **data science va changer les process** qu'ils ont depuis des années. Il faut apprendre **la pédagogie ou la vulgarisation** pour pouvoir convaincre les décideurs qui n'ont pas forcément les connaissances techniques.*
*Il faut aussi une **bonne capacité de maîtrise de soi et de résistance**. »*

les data scientists du labo BD&BA

organisation

- Le labo : data scientists, architectes, développeurs logiciel
- Les domaines opérationnels
- Les académiques et les SMEs

formation

- ENSAI, Polytechnique, Normale, Université

recrutement

- Immersion stages et thèses puis CDI

profil « psychologique »

- curieux, agile, innovant, pédagogue, communicant, multi-tâches, résistant

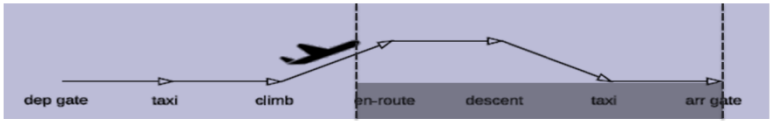
évolution

- Le labo : un lieu de passage et de formation
- Des experts qui partent dans les domaines opérationnels

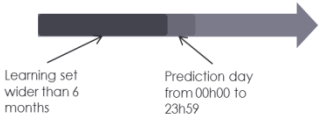
au quotidien

données, besoins
chaines de traitement
évaluation, tests
accompagnement vers l'industrialisation
veille, benchmarking, publications, brevets

Prédiction du temps de vol restant

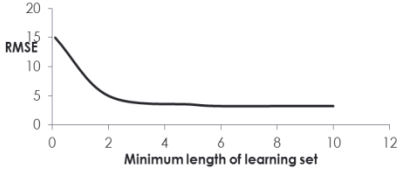


Phase d'apprentissage



Les meilleurs résultats sont obtenus avec 6 mois d'historiques

- L'erreur de prévision est stabilisée après 6 mois d'historique
- Réduction de l'erreur de l'ordre de 10% en passant d'un historique de 2 mois à 6 mois



RMSE	Deep Learning	Kernel Regression	Random Forest	Stochastic calculus
Juste après le départ de la gate	3.25 min	3.41 min	4.30 min	3.32 min
60 min avant l'arrivée à la gate	1.63 min	1.70 min	3.20 min	1.81 min
20 min avant l'arrivée à la gate	1.53 min	1.42 min	3.12 min	1.02 min
10 min avant l'arrivée à la gate	1.12 min	1.20 min	3.11 min	46 sec

OPEN

THALES

data science et actuariat
digitalisation de l'Assurance / assurance connectée / e-assurance …

transformation du monde de l'assurance induite par le potentiel des data sciences et des nouvelles technologies (Big Data, IoT, assurance en ligne, etc …)

data science au cœur du métier de l'assurance et de l'actuariat

- une vraie culture de la data et de la statistique
- des pistes pour
  - optimiser la gestion des risques en croisant les données, en allant plus vite sur le développement des modèles
  - réduire les coûts pour l'assureur tout en garantissant une amélioration des services proposés
  - s'ajuster de façon individuelle aux usages et besoins des bénéficiaires
  - améliorer l'expérience client et capter de nouveaux marchés
- tout en prenant en compte le cadre réglementaire et éthique

INSTITUT DES ACTUAIRES

le programme de formation en data science

- développer les compétences techniques des actuaires en informatique permettant d'aborder de façon opérationnelle les problématiques liées aux nouveaux usages du numérique et à la nouvelle typologie des données : diversité et volumétrie
- acquérir les éléments de la chaîne de traitement de données massives : stockage, filtrage, analyse prédictive, validation, visualisation, valorisation
- identifier des enjeux économiques (valeur client, tarification personnalisée, prévention ciblée des risques) dans un cadre juridique et déontologique en évolution

THALES

en conclusion
PROPOSITIONS D'ACCOMPAGNEMENT

il ne s'agit pas d'une proposition commerciale mais d'une proposition de collaboration Institut des Actuaires & Labo

THALES

paroles d'un ancien étudiant de l'Institut des Actuaires

ANGELO CARIA
CREDIT AGRCIOLE NORD DE FRANCE
ARRAS, France
Poste actuel RESPONSABLE CRM ANALYTIQUE
Formation ACTUAIRE DATA SCIENTIST

le labo fort de son expérience nous a permis

- de passer en mode projet pour bien aborder notre problématique du webtracking :

en nous formant à l'utilisation de hive un outil simple d'utilisation permettant de manipuler de gros volumes de données et de les structurer

en nous proposant des solutions pour identifier des parcours client qui ont donné lieu à des souscriptions d'un produit lors de la réalisation d'un devis (identification de patterns en utilisant des packages présents dans des versions récentes de spark mais pas dans celle que nous avions à disposition)

- de mieux appréhender l'utilisation d'une plateforme big data en l'occurrence ici Hortonworks dans le cadre d'un POC BIG DATA

en nous sensibilisant aux problèmes d'architecture et de version des composants de la plateforme (allers retours avec l'IT pour résoudre des problèmes de dysfonctionnements de la plateforme impactant la réalisation du POC)

en nous accompagnant lors de la mise en place de code scala pour la problématique par exemple de sélection de variables dans un modèle de score

en nous aidant à mettre en place des ponts entre les données du datawarehouse et les données

en nous aidant à mettre en place des solutions d'industrialisation pour rendre opérationnelle notre démarche

- de réfléchir à la mise en œuvre de l'anonymisation des données

Pour conclure : le labo nous a permis de conforter notre expertise en data science, de mieux appréhender les problématiques liées à un environnement Big Data et de rendre opérationnelle notre démarche Big Data

MERCI A TOUS!
Q&A

OPEN

THALES