

# POC Coclico : risque(s) data, algorithme ?



<u>Auteur</u>	: LEPICAUT Didier
<u>Jury</u>	: MS Valdom - thèse professionnelle
<u>Objet</u>	: Salarié Groupe EDF - Marché d'Affaires - HUB2B - DataHUB
<u>Date</u>	: 03/10/2019
<u>Ecole</u>	: Philippe Besse
<u>Entreprise</u>	: Arnaud Bortolotti

Problème : La gestion des risques données et algorithmiques est-il un nouvel enjeu pour la gouvernance des données des entreprises ?

## CADRAGE

date	trim4.2019
version	version 3.1
source des données	prototype Coclico
nombre de variables	p = 2 : risque et robustesse
nombre d'enregistrements	n = 4 : expression de besoins
périmètre devis concerné	programme Marché Affaires Intelligence Artificielle
problème	piloter les risques d'un prototype de voice-computing
type d'analyse	modèle de risque management
<b>Note destiné au lecteur</b>	résultats clefs



## - SOMMAIRE -

(1) Contexte	p. 3
(2) Problème	p. 6
(3) Méthode	p. 8
(4) Données	p. 10
(5) Résultats	p. 11
A RETENIR	p. 15

\*\*\* AVERTISSEMENT : LE CONTEXTE ORGANISATIONNEL DE HUB2B / DATAHUB FACONNE LE PLAN DE TRAVAIL DE CE DOCUMENT \*\*\*

Note à l'attention des lecteurs : lors de la rédaction de ce document, j'ai pris le plus grand soin d'adopter un esprit scientifique rigoureux, de mobiliser les concepts et autres illustrations utiles à la compréhension du travail et de fournir le niveau de précision adapté à toutes les parties prenantes techniques et fonctionnelles du Marché d'Affaires concernées par ce document.

Lors de la séance de Grand Oral Valdom, afin de fluidifier la lecture de ce support et pointer les résultats clefs pour les membres du Jury, je concentrerais ma présentation sur les éléments repérés par la signalétique [ ] .

Groupe EDF ~ Direction Commerce ~ Marché d'Affaires: au 31/12/18, France.

. Volume CA = 13 Mds / €, en part de Marché = 65%,

Bilan concurrence sur le marché « B to B »:

- . La concurrence fait rage sur **l'innovation opérationnelle dans la conception des produits et des services,**
- . **La connaissance client est le facteur clef de succès,**
- . **La protection des données est le 1<sup>er</sup> besoin client.**



The New York Times

*San Francisco Bans Facial Recognition Technology*



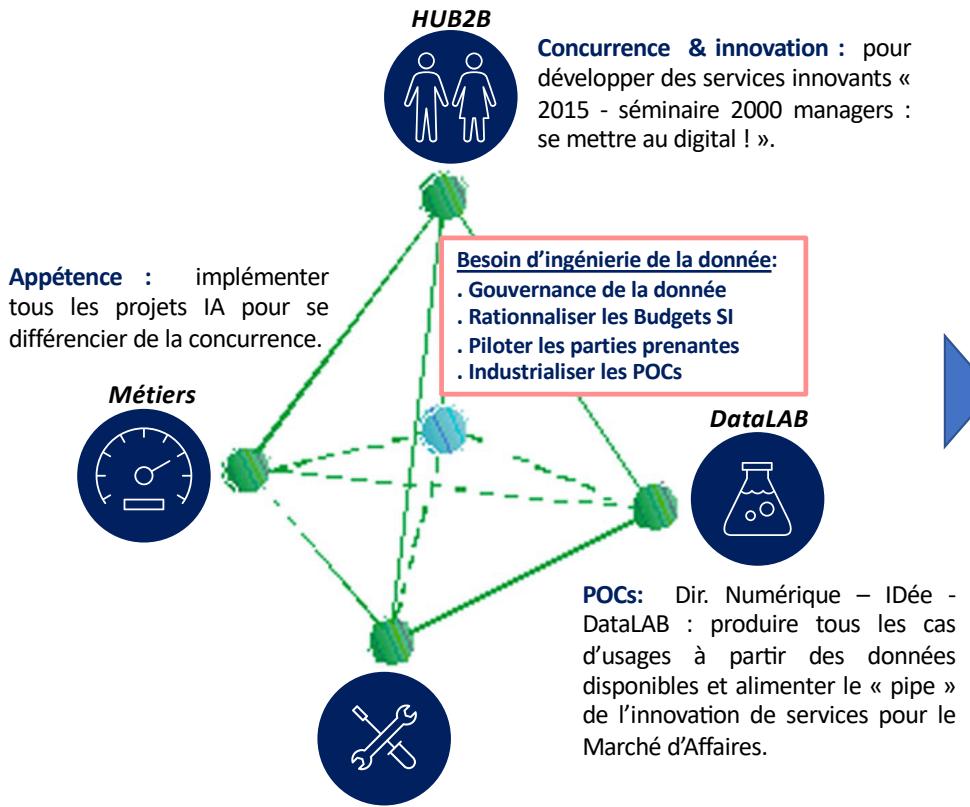
<https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html>



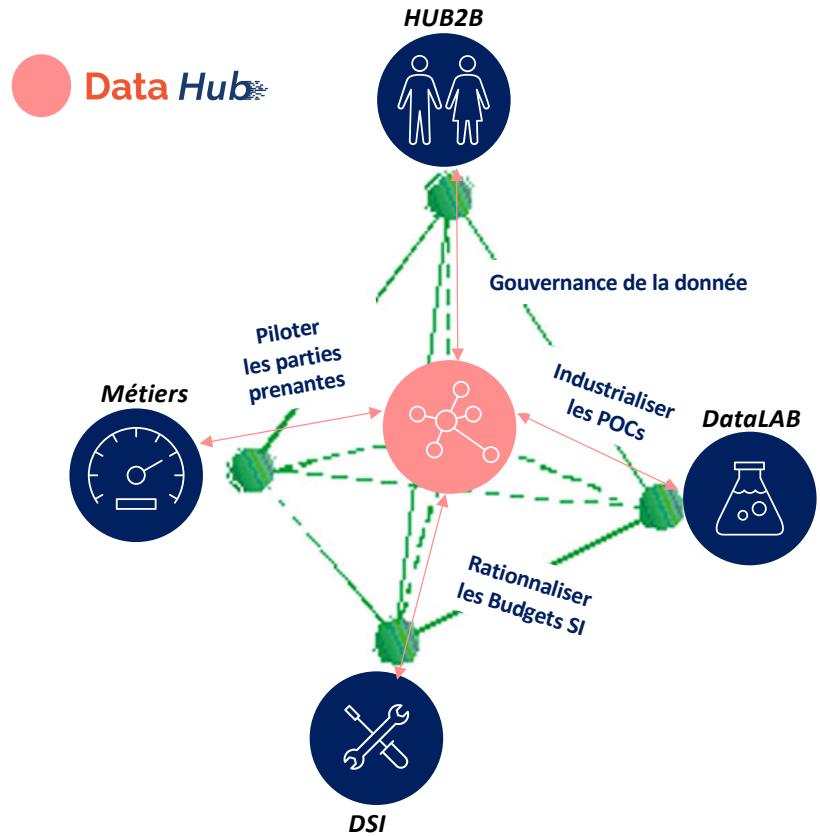
**« ... le pire ennemi d'EDF : c'est le commerçant final qui possède la connaissance client ... »**

## CONTEXTE INTERNE : POURQUOI LA CREATION DE DATAHUB ? (1/2)

2015 - 2017 : DATA & ALGO ~ ETAT DES LIEUX



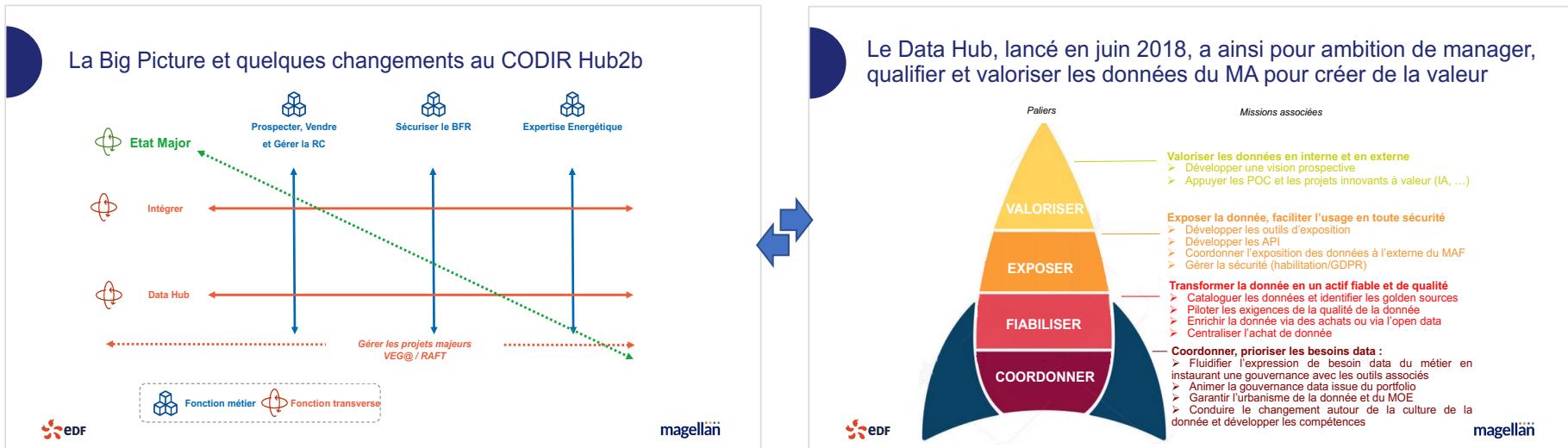
2018 - 2021 : STRATEGIE IA POUR MAGELLAN



**Investissement & exploitation :** 2015 -  
achat d'une solution stockage  
technique des données eDMA,  
construire une trajectoire IA.

**Marché d'Affaires -> HUB2B -> DataHUB -> Stratégie IA :**  
Identifier et démarrer une structure en charge de l'ingénierie des données du MA

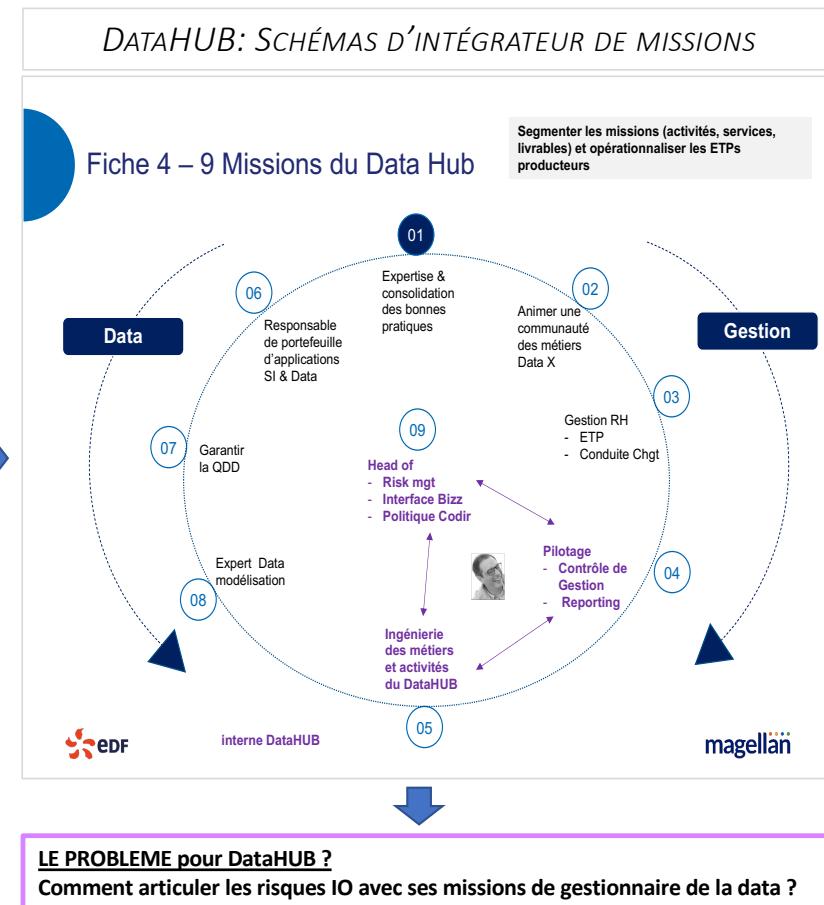
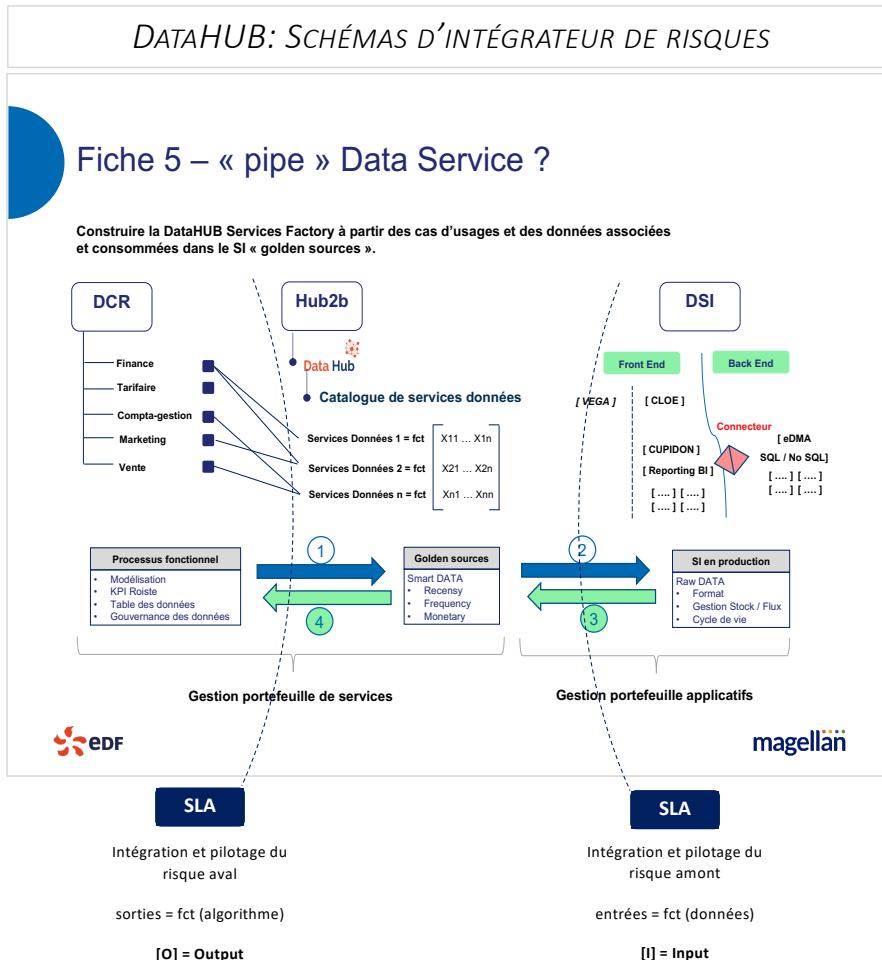
## CONTEXTE INTERNE : LES MISSIONS DE LA FUSEE DE DATAHUB (2/2)



**HUB2B:** direction en charge d'animer le plan stratégique Magellan 2017-21 pour le Marché d'Affaires, se transforme en créant un service qui va gérer « la gouvernance des données », DataHUB.

**DataHUB:** les 4 métiers (**Valoriser, Exposer, Fiabiliser et Coordonner**) doivent décliner la nouvelle gouvernance du modèle intégré de HUB2B.

## FORMULATION DU PROBLEME



## Risque donnée ?

→ Tableau synoptique de qualification du risque données :

FAMILLE	QUALIFICATION		CONTRE-MESURE	
	TYPE	DETAIL		
<b>(1) Défaut de qualité des données sur</b>				
	(notation matricielle) : $Y = XB + E$			
1.1. Y : erreur codage étiquette	. biais statistiques	. défaut d'audit et préparation des Y	. data : contrôle / audit (preprocessing)	
1.2. X : données manquantes (Tx NA)	. biais statistiques	. imputation « hotdeck » **	. appliquer l'imputation adaptée au problème (i.e. l'algorithme adhoc)	
1.3. X : erreur de mesure (acquisition)	. biais statistiques	. GIGO : Garbage in, Garbage out, données inexactes	. durcir le protocole de mesure [i.e. Plan d'Expérience]	
1.4. X : erreur de codage (preprocessing)	. biais statistiques	. défaut d'audit et préparation des Xi	. data : standardisée, normalisée, et nettoyée de valeur aberrante ...	
1.5. X : données corrompues - test	. biais statistiques	. adversarial attack	. algo. défenseur <sup>20</sup>	
1.6. X : données corrompues - train	. biais statistiques	. data poisoning attack	. algo. défenseur	
<b>(2) Défaut de quantité des données sur</b>				
2.1. Y et X : Nb d'enregistrements insuffisant pour la formation du modèle en apprentissage et en test	. biais statistiques	. défaut de généralisation	. durcir le protocole d'acquisition de données ou transfert learning	
2.2. Y : déséquilibre de classe – formation du modèle en apprentissage [sur représentation d'une classe Y vs les autres classes Y]	. biais statistiques	. la formation du modèle se fait en apprenant sur la classe dominante	. rééquilibrage de classe (équi-répartition) avec l'algorithme SMOTE : Synthetic Minority Over-sampling Technique.	
2.3. Y : biais de sélection classe – problème de représentativité de l'échantillon vs la population totale [Poids des classes Y dans l'échantillon non conforme au Poids des classes Y dans la population totale]	. biais statistiques	. biais de sélection : credit scoring  . biais de censure et de troncature : maintenance prédictive avec dérive temporelle (censure temps observation)	. échantillonnage stratifié pour former le dataset total (train + test)  . estimer la durée de vie totale de l'appareil [i.e. du phénomène observé]	

**ETRE PRATICO-PRATIQUE:** Le risque données et algorithme en entreprise ...

- . peut s'analyser dans une approche naïve comme un simple paramétrage de modèle,
- . est une activité récurrente, complexe et non automatique,
- . adresse un champ disciplinaire où la matérialité d'un risque se mesure a posteriori...

→ Tableau synoptique de qualification du risque algorithmique :

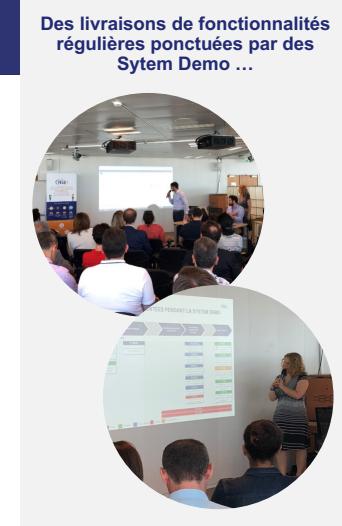
FAMILLE	QUALIFICATION		CONTRE-MESURE
	TYPE	DETAIL	
<b>(1) Défaillance de la conception</b>			
1.1. Défaut Ontologie de l'algorithme :	. biais cognitifs	. le biais de « bandwagon » ou « du mouton de Parnasse	. suivre des modélisations qui sont populaires sans s'assurer de leur exactitude. (i.e. faire du kmeans sur du supervisé)
<b>Risque algorithme ?</b>			
		. Les biais d'anticipation et de confirmation	= le programmeur à favoriser sa vision du monde même si des données disponibles peuvent remettre en question cette vision.
		. Le biais de « corrélations illusoires »	= conduire une personne à déceler des corrélations entre deux événements indépendants.
		. le biais de stéréotype (i.e. femme vs offre emploi)	= Ce dernier peut survenir lorsqu'un individu agit en référence au groupe social auquel il s'identifie plutôt que sur ses capacités individuelles
<b>(2) Défaillance dans l'utilisation</b>			
2.1. Défaut de spécification du Modèle :	. biais statistiques	. Le biais de variable omise	. non prise en compte d'une variable importante, i.e. var proxy à toutes les autres.
	. biais statistiques	. Le biais d'endogénéité	. utiliser les données du passé pour prédire le futur alors que la spécification du modèle a changé dans le temps
2.2. Défaut de réglage du modèle :	. biais statistiques	. la pratique d'Overfitting du modèle	. procédure validation croisée k-folds
<b>(3) De nouveaux enjeux à prendre en compte</b>			
3.1. L'équité algorithmique	. L'équité est concept éthique et pluriel . Algorithme = fonction (équité) en intégrant des contraintes . Les règles formelles d'équité sont incomptables et non universelles		
3.2. Interprétabilité ou l'explicabilité des algorithmes	. se réfère à la complexité des modèles, à la nécessité d'expliquer aux utilisateurs finaux comment et pourquoi un résultat a été obtenu.		
3.3. De la transparence à l'auditabilité des algorithmes	. L'idée est de rendre public, ou bien de mettre sous séquestre, des algorithmes en vue les auditer pour étudier des difficultés potentielles.		
3.4. La responsabilité algorithmique	. Principe du Pollueur / Payer : la responsabilité est le principe selon lequel une personne ou une organisation légalement responsable d'un préjudice doit fournir une explication ou une compensation au préjudice subi.		

Source : Algorithmes - Biais discrimination équité. p. 8. / HAL Id: hal-02077745 / Patrice Bertail, D. Bounie, Stephan Clemenccon, Patrick Waelbroeck

. l'analyse de risque doit porter sur un processus opérationnel « de tout petit scope »,  
avec une attention particulière sur l'explicabilité des données et des modèles pour éclairer la dimension éthique des prédictions (i.e. biais de sélection).

## METHODE DE RESOLUTION (1/2)

PHASE DU PROJET	OUTIL MOBILISE	DESCRIPTION	PREUVE
JALON – CODIR HUB2B [Amont Projet]	. Revue de « Portfolio » pour une prise de décision du Comité de Direction de la Direction HUB2B.	. Le Chef de Projet MAIA passe en revue toutes les fiches POC, pour prononcer le GO/NOGO « Business » de chaque POC, puis à chaque Rendez-vous de l'instance CODIR fait un point d'avancement projet sur tous les POCs engagés en SPRINT <sup>23</sup> .	. Etude d'opportunité d'un POC IA . Suivi de projet et coordination des parties prenantes aux projets POC IA du Marché d'affaires (DM, DSL, DN/IDée, HUB2B).
JALON – démarrage du POC « Prononce le GO de J0 »	. Comitologie : mise en place de l'instance décisionnelle Conseil Scientifique [CS] <sup>24</sup>	. Réunir toutes les compétences techniques et fonctionnelles en relation avec l'objet du POC, prononcer un GO de lancement du POC au sein du DN/IDée DataLAB.	. Constituer un comité d'experts techniques et fonctionnels qui va suivre « le déroulé » du Sprint projet du POC data sciences.  ➤ <i>To Do : équipe Sia Partners avec (Arnaud Bortolotti &amp; JM Gauthier) = autorité pour définir et retenir les personnes qualifiées pour chaque POC.</i>
	. Outil : <b>en amont projet</b> , utiliser la FICHE D'EVALUATION DU RISQUE MODELE.	. Grille de notation du risque modèle, qui croise les axes d'analyse - risque des impacts négatifs du modèle et la robustesse du modèle mobilisé.	. Le CS va passer en revue les conditions de réussite du POC, avant son démarrage au sein du DN / IDée / DataLAB.
JALON – suivi POC	. Reporting sur l'avancement du POC.	. Le responsable DN / Idée / DataLAB avec le Data-Steward (DataHUB) diffusera au fil de l'eau de la réalisation du POC, des éléments de point d'avancement.	. Identifier si point bloquant lors de la production du POC par le Data Scientist et mettre en place les méthodes correctives nécessaires.
	. Outil : à formaliser.	. Prise en charge de la réalisation d'un reporting partagé HUB2B / DataHUB et DN / IDée / DataLAB.	➤ <i>To Do : équipe Sia Partners</i>
JALON – recette du POC	. Conseil Scientifique : recette du POC.	. Une fois le POC finalisé au sein de la DN / IDée / DataLAB, réunir l'instance Conseil Scientifique pour prononcer un PV Recette pour « Mise en production du POC ».	. Réunir le CS qui a participé au Go de lancement J0, pour passer en revue les conditions de réalisation du POC data sciences et auditer les risques données et algorithme.
	. Outil : <b>en aval projet</b> , utiliser la FICHE CHECKLIST	. La CHECKLIST qui audite les risques données et algorithme du POC.	. Le CS va passer en revue tous les éléments de conception, de mise en œuvre, de validation des résultats des modèles d'apprentissage statistiques.



### CONTEXTUALISER:

L'ingénierie a été de transformer le problème de risque (data, algorithme) en modèle de gestion de risques adapté à l'organisation ...

... mise en œuvre avec un jalonnement projet

## METHODE RESOLUTION (2/2)

### • Modèles quantitatifs: REX du secteur financier

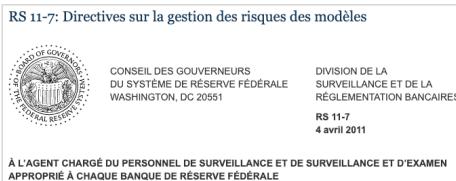


En août 2007, aux États-Unis, les défauts de paiement des emprunteurs conduisent à la **crise des « subprimes »**, c'est-à-dire des prêts hypothécaires à taux variable accordés à des familles américaines pauvres ...

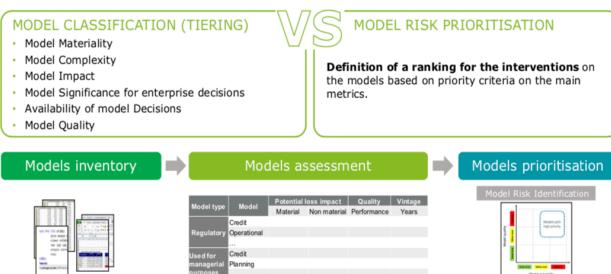
... de nombreux ménages américains sont dans l'incapacité de rembourser les échéances de ces emprunts ... **faillite de Lehman Brothers 15 sept. 2008**

... l'utilisation imprudente d'un modèle de risque unique qui a presque tué Wall Street. La **copule gaussienne**, utilisée indifféremment pour capturer la structure de dépendance sous-jacente de trillions de dollars de produits titrisés ... ( $10^{12}$ )

**Déf.**: famille de modèles de copule gaussienne, utilisés en finance pour estimer la distribution de probabilité de pertes sur un pool de prêts ou d'obligations, et qui ont été impliqués de manière centrale dans la crise du crédit



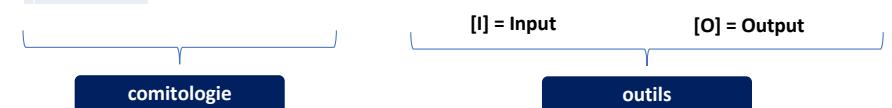
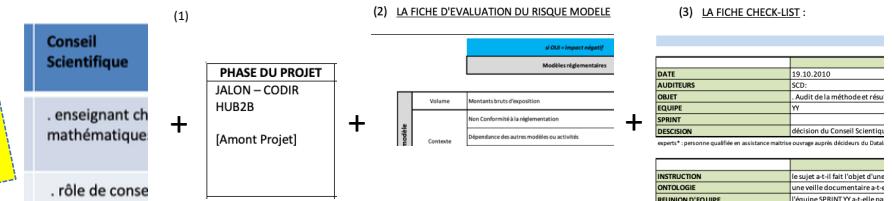
### Integrated Approach for Model Risk Assessment



### ITEMS DU CADRE GLOBAL DE RISQUE MANAGEMENT

CH 1. Gouvernance, politiques et contrôles:	CH 2. Développement, mise en œuvre et utilisation	CH 3. Processus de validation du modèle
<ul style="list-style-type: none"> <li>Politique</li> <li>Définition du modèle</li> <li>inventaire</li> <li>Les contrôles</li> <li>Rôles et responsabilités</li> <li>Documentation</li> <li><b>Évaluation du risque de modèle</b></li> <li>Agrégation des risques du modèle</li> <li>Processus de contrôle des changements</li> <li>Défi efficace</li> <li>Utilisation des vendeurs</li> <li>Identifiants des parties prenantes</li> <li>Processus du cycle de vie</li> <li>Interprétation réglementaire</li> </ul>	<ul style="list-style-type: none"> <li>Procédé de design</li> <li>évaluation des données</li> <li><b>Test de modèle</b></li> <li>Documentation</li> <li><b>Limites du modèle</b></li> <li><b>Évaluation du risque de modèle</b></li> <li>utilisation vs intention</li> <li>Processus de programmation</li> <li>Intégration au réseau</li> <li><b>Conception de contrôles</b></li> <li><b>Mise en œuvre des tests</b></li> <li>Documentation</li> <li><b>Processus d'erreur de modèle</b></li> </ul>	<ul style="list-style-type: none"> <li><b>Procédures de validation</b></li> <li>Documentation</li> <li>Résolution des constatations</li> <li>Nature de la surveillance</li> <li>étendue de la surveillance</li> <li>Fréquence de la surveillance</li> <li>Procédures de re-calcul</li> <li>Solidité conceptuelle</li> <li><b>Analyse des résultats</b></li> <li><b>Analyse de sensibilité</b></li> <li>Documentation</li> <li><b>Processus d'erreur de modèle</b></li> </ul>

Transfert de technologie



**METHODE DE L'INGENIEUR:**  
• Vérifier que les paramètres du modèle de gestion des risques sont nominaux

## DONNEES

Terrain d'étude:  
piloter les risques d'un prototype de voice-computing

### MATRICE D'ANALYSE QUI CARTOGRAPHIE LES DIFFERENTES EXPRESSIONS DE BESOINS DU POC MA'IA – périmètre (DN, DATAHUB)

Fonctionnement de l'algorithme Horizon backlog de production	<i>Temps réel</i>	<i>Temps différé</i>
Court terme	<b>Compte-Rendu Automatique : transcription automatique des conversations en compte-rendu d'appel (RPA)</b> <span style="border: 1px solid red; padding: 2px;">①</span>	<b>Cas d'entraînement pédagogique : collecter des données textuelles pour développer des cas pédagogiques de formation pour le perfectionnement des conseillers clients (RPA, textmining)</b> <span style="border: 1px solid red; padding: 2px;">②</span>
Moyen terme	<b>Sentiment Analysis : mettre en place une supervision des appels entrants des clients MA pour mieux qualifier le besoin client et adapter le type de réponse EDF expert niveau 1, expert niveau 2 ... à l'échelon centre d'appels (RPA)</b> <span style="border: 1px solid red; padding: 2px;">③</span>	
Long terme	<b>Sentiment Analysis : déclenchement d'alerte risque de départ client sur la base de détection de mots clefs (marquage client + mise en route du programme de fidélisation client &gt; RPA, textmining)</b> <span style="border: 1px solid red; padding: 2px;">④</span>	<b>Sentiment Analysis : mettre en place un programme marketing de fidélisation client pour lutter contre le CHURN client (scoring de risque départ client + plan commercial = action de fidélisation &gt; Modèle classique scoring = régression logistique)</b>

## RESULTATS: analyse intra, entre les 4 fiches d'évaluation du POC MA'IA (1/2)

Compte-Rendu Automatique & Cas d'entraînement pédagogique					
Evaluation du risque modèle		RISQUE (I)			
	% (Oui)	[0;25[	[25;50[	[50;75[	[75;100[
ROBUSTESSE (ii)	(% (Oui))	note	faible	moyen	fort
	[0;25[	très	X		
	[25;50[	robuste			
	[50;75[	moyen			
	[75;100[	faible			

Analyse a posteriori

Sentiment Analysis					
Evaluation du risque modèle		RISQUE (I)			
	% (Oui)	[0;25[	[25;50[	[50;75[	[75;100[
ROBUSTESSE (ii)	(% (Oui))	note	faible	moyen	fort
	[0;25[	très			
	[25;50[	robuste			
	[50;75[	moyen	X		
	[75;100[	faible			

Sentiment Analysis anti-CHURN					
Evaluation du risque modèle		RISQUE (I)			
	% (Oui)	[0;25[	[25;50[	[50;75[	[75;100[
ROBUSTESSE (ii)	(% (Oui))	note	faible	moyen	fort
	[0;25[	très			
	[25;50[	robuste			
	[50;75[	moyen			
	[75;100[	faible			X

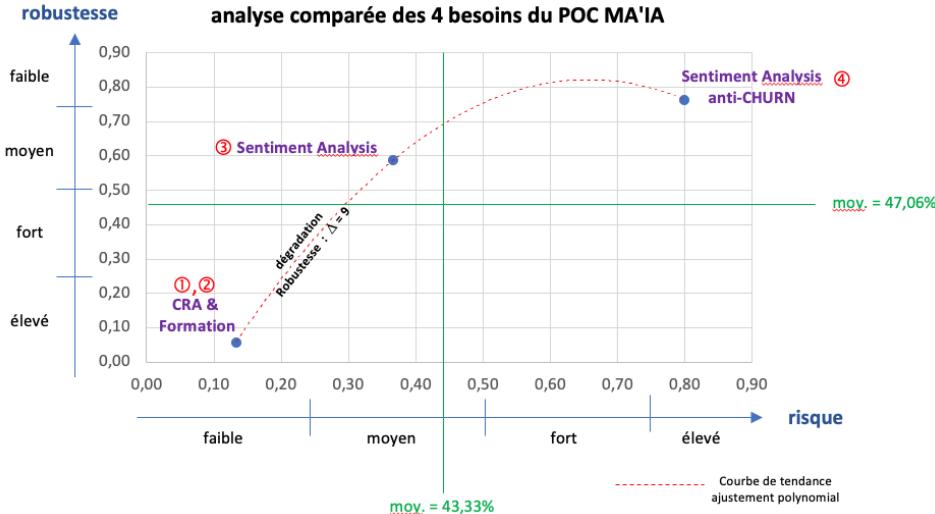
Compte-Rendu Automatique & Cas d'entraînement pédagogique					
Sentiment Analysis					
Sentiment Analysis anti-CHURN					

risque	delta	robustesse	delta
13,33%		5,88%	9,00
36,67%	1,75	58,82%	
80,00%	1,18	76,47%	0,30

moy.

43,33%

47,06%



Analyse:

. d'un point de vue qualitatif:

- l'outil fiche d'évaluation joue son rôle et permet de positionner les 4 besoins du POC MA'IA

. d'un point de vue quantitatif:

- sur la moyenne, le POC MA'IA est un modèle de risque moyennement robuste et avec un risque d'impacts négatifs moyen [vs. notre modèle de risque cible = risque faible + robustesse élevée].
- sur l'axe de risque: 3 sur 4 modèles sont en-dessous de la moyenne de risque,
- sur l'axe de robustesse: 2 sur 4 modèles sont au-dessus de la moyenne de robustesse,

On observe une forte accélération de la dégradation de la robustesse ( $\Delta = 9 \sim$  taux variation) entre les besoins ①, ② et le besoin ③.

i.e. l'outil « fiche d'évaluation du risque modèle » arrive bien à capturer l'accroissement de complexité introduit avec le besoin « Sentiment Analysis ».

## RESULTATS: analyse inter, entre l'analyse a posteriori et l'analyse a priori (2/2)

Compte-Rendu Automatique & Cas d'entraînement pédagogique						
Evaluation du risque modèle		RISQUE (I)				
	% (Oui)	note	faible	moyen	fort	élevé
ROBUSTESSE (II)	[0;25[	très	X			
	[25;50[	robuste				
	[50;75[	moyen				
	[75;100[	faible				

Analyse  
a posteriori

Sentiment Analysis						
Evaluation du risque modèle		RISQUE (I)				
	% (Oui)	note	faible	moyen	fort	élevé
ROBUSTESSE (II)	[0;25[	très				
	[25;50[	robuste				
	[50;75[	moyen	X			
	[75;100[	faible				

Analyse  
a priori

Sentiment Analysis anti-CHURN						
Evaluation du risque modèle		RISQUE (I)				
	% (Oui)	note	faible	moyen	fort	élevé
ROBUSTESSE (II)	[0;25[	très				
	[25;50[	robuste				
	[50;75[	moyen				
	[75;100[	faible				X

Fonctionnement de l'algorithme	Temps réel	Temps différé
Horizon backlog de production		
Court terme	Compte-Rendu Automatique :  Risque data : moyen   Risque algo. : faible	Cas d'entraînement pédagogique :  Risque data : faible   Risque algo. : faible
Moyen terme	Sentiment Analysis :  Risque data : moyen   Risque algo. : moyen	
Long terme	Sentiment Analysis anti-CHURN :  Risque data : fort   Risque algo. : fort	Sentiment Analysis anti-CHURN :  Risque data : fort   Risque algo. : fort

### Analyse:

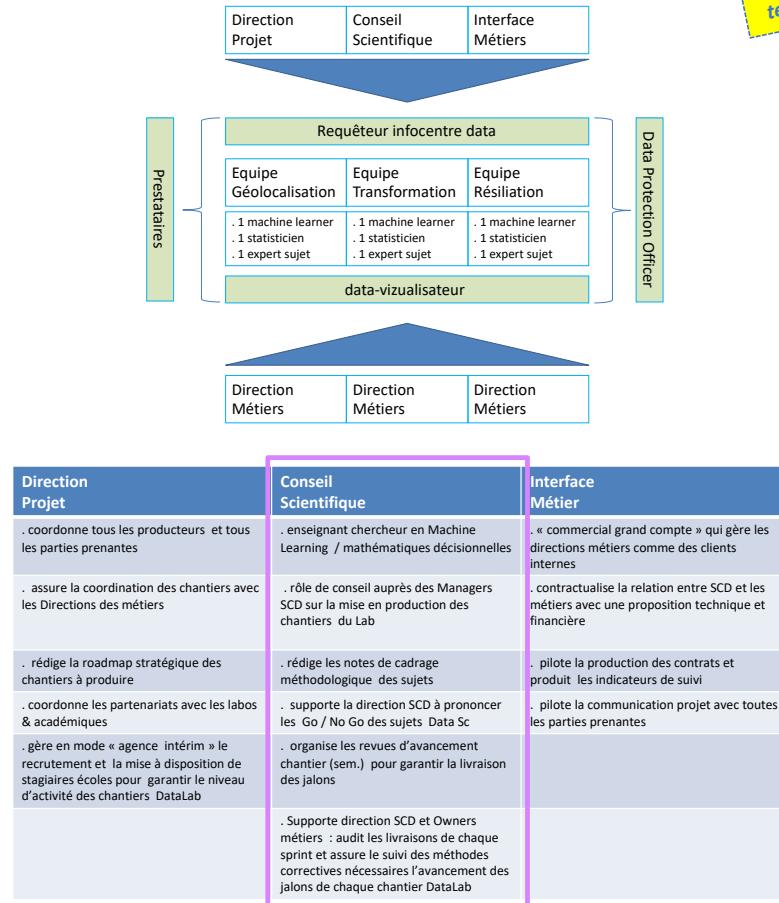
. lors de la comparaison des résultats des 2 analyses, l'analyse a posteriori (calculatoire avec l'incorporation des données via les fiches d'évaluation) confirme l'analyse a priori reposant sur « **un prior qualitatif** » apporté par la connaissance projet compilée avec la fiche d'instruction projet de l'instance Portfolio et le suivi de l'avancement projet Coclico.

Compte-Rendu Automatique & Cas d'entraînement pédagogique						
Sentiment Analysis						
Sentiment Analysis anti-CHURN						

moy.  
43,33%  
47,06%



> DataLAB



## RESULTATS: Conseil Scientifique



> Data Hub

Transfert  
de  
technologie

n°	POSITION	ENTITE	EXPERTISE
<b>PERSONNEL INTERNE EDF GROUPE</b>			
1	Chief Data Officer	COMMERCE	Stratégie de la gestion des données d'EDF Commerce
2	Chief Data Manager	MA / HUB2B / DataHUB	Définit, met en œuvre et fait appliquer les politiques et règles de gestion des données pour le MA
3	Customer Data Manager	MCP (Client Particulier)	Définit, met en œuvre et fait appliquer les politiques et règles de gestion des données pour le MCP
4	Chef du Pôle LID (Text mining)	COMMERCE/DN/IDée/Data LAB	Ingénierie des données Textuelles, intervient sur les marchés MCP et MA
5	Data Scientist	COMMERCE/DN/IDée/Data LAB	Expert en apprentissage statistique
6	Responsable Socié Données eDMA	MA / HUB2B / DataHUB	Expert en Data Gouvernance
7	BI Décisionnel + Data steward	MA / HUB2B / DataHUB	requête SQL + analyste solution fonctionnelle BI Décisionnel
8	Project Manager Office	Sia Partners / MA / HUB2B / DataHUB	Conduite de changement organisationnel
9	Architecte technique et fonctionnel	DSI	Product Owner solution de stockage des données (roadmap et industrialisation des POCs)
10	Charge de Mission	DSI	Chief de projet déploiement programme MAIA + en charge du CRM CLOE
11	Chef de projet	MA / HUB2B / animation commerce	Chef de projet animation multi-canaux [dont animation du métier conseiller en Centre d'Appel]
<b>PERSONNES QUALIFIÉES EXTERNES - EXPERTS TECHNIQUES - MOBILISABLES PAR LE CONSEIL SCIENTIFIQUE</b>			
1	doctorant [CIFRE - allomédia]	lium.univ-lemans.fr	Ingénieur R&D : analyse de données massives en temps réel pour l'extraction d'informations sémantiques et émotionnelles de la parole
2	ingénieur R&D en solution basée sur Théorie Sens-Texte (MTT)	<a href="https://www.inbenta.com/en/technology/the-meaning-text-theory/">https://www.inbenta.com/en/technology/the-meaning-text-theory/</a>	Expert solution de Speech-to-Text à base de règle de gestion
3	académique - chercheur	à définir	Expert en deep learning spécialiste du NLP
4	...	...	...

### Analyse:

- . La seule supervision du Conseil Scientifique garantit l'alignement modèle quantitatif et des IO,
- . **La qualité des personnes qualifiées externes est critique pour l'atteinte de cet objectif.**

RESULTATS DU PROTOTYPE COCLICO	EXPERIMENTATION
Reproductibilité	<input type="radio"/>
Généralisation	<input type="radio"/>
Robustesse	<input checked="" type="radio"/>
Explicabilité	<input checked="" type="radio"/>

30.09.19 : Quick Win  
Pas de lancement de projet de développement informatique  
sans que le critère d'acceptation des données soit validé

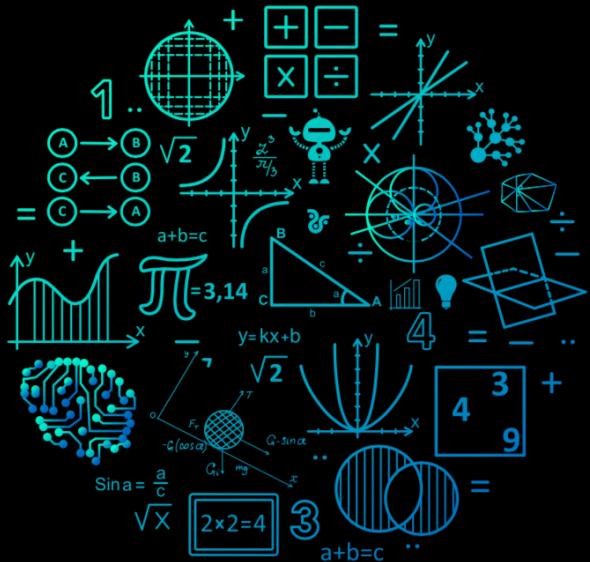


### Matérialiser des recommandations opérationnelles (A partir du contexte HUB2B / DataHUB en Trim4/2019)

- **Enabler 1** : mettre en œuvre la démarche de data lignage et de gouvernance des données,
- **Enabler 2** : produire le dossier Qualité des données,
- **Enabler 3** : orienter la roadmap d'architecture des infrastructures de stockage NoSQL,
- **Enabler 4** : en coordination avec le Chief Data Officer et sous le pilotage de Sia Partners, en 2020, organiser le premier « test à blanc » de mise en œuvre du Cadre Global de Risque Management avec un dossier à sensibilité risque données et risque algorithme avérée,
- **Enabler 5** : dans le cadre de la méthode agile « SAFE », en 2020, organiser la production des enablers 1 à 4 en intégrant les critères Technologique, Juridique, Financier, Organisationnel et de la GRH.

## - A RETENIR -

LIVRABLE	<p>L'objectif principal de ce travail, sur la base de l'étude de cas Coclito, <b>consiste à fournir les éléments de méthodes et les outils permettant à l'organisation HUB2B / DataHUB d'anticiper et de minimiser les risques données et algorithme qui sont contenus dans les dossiers « Proof of Concept » à base d'algorithme de Machine et Deep Learning du Programme d'Intelligence Artificiel du Marché d'Affaires du Groupe EDF.</b></p>
ENSEIGNEMENTS	<ul style="list-style-type: none"><li><b>Conduite du changement:</b> La transformation en modèle intégré de HUB2B... passe par la création de DataHUB en charge de l'ingénierie des données pour le Marché d'Affaires,</li><li><b>Organisation:</b> Le DataHUB pour fonctionner doit organiser son activité sous le format d'un DataHUB Services Factory,</li><li><b>Data lignage:</b> DataHUB doit être en capacité technique d'auditer les entrées, les sorties, le modèle d'apprentissage mobilisé ainsi que la cohérence de la méthodologie appliquée par les Data Scientists de la DN / IDée / DataLAB,</li><li><b>Risque algorithme:</b> La solution pour minimiser le risque algorithme passe par l'hybridation des performances des modèles de Machine / Deep learning avec l'outillage, les méthodes et les techniques de l'ingénieur en statistique -&gt; l'IA hybride,</li><li><b>Etude de cas:</b> L'outil fiche d'évaluation du risque modèle joue son rôle et permet de positionner les 4 expressions de besoins du POC MA'IA sur le « Mapping » d'analyse comparée.</li></ul>
POC SENTIMENT ANALYSIS	<ul style="list-style-type: none"><li><b>Etat de l'art:</b> sujet de recherche, peu de solution IT en production,</li><li><b>Mathématique:</b> un problème de classification multi-classes et un difficile apprentissage à former [peu de dataset normalisé],</li><li><b>Stack de calculs:</b> « end-to-end » Deep Speech 3 ~ <a href="http://research.baidu.com">http://research.baidu.com</a> [transcription auto. de la parole],</li><li><b>Architecture du réseau:</b> définir et exécuter un modèle unique capable de travailler sur des données mixtes (audio, textuelle) ?</li></ul>



## Managing algorithmic risks

Safeguarding the use of complex algorithms  
and machine learning

# Merci!

## - ANNEXES TECHNIQUES -

➔ Ce qu'il faut retenir de l'échange :

**1. Voice-computing en bidirectionnel :**

- c'est surtout un sujet de recherche,
- pas ou peu de solution réelle IT en production sur la question du « Sentiment Analysis »,

**2. Le socle d'apprentissage statistique support à la recherche :**

- sur le SA, c'est un problème difficile de classification multi-classes,
- difficile de former l'apprentissage, pas ou peu de dataset normalisé,

**3. Solution algorithmique pour le calcul numérique :**

- historiquement, tous les chercheurs du voice-computing ont démarré avec Kaldi,
- actuellement, on utilise le Stack de calculs « end-to-end » Deep Speech 2 et 3,  
(transcription automatique de la parole) de <http://research.baidu.com>

**4. Puissance de calcul pour former le classifieur :**

- besoin d'une puissance de calcul importante, avec un code d'exécution en parallèle,
- mobiliser un cluster de multi-GPUs,

**5. Retour d'Expérience :**

- il reste à définir et exécuter un modèle unique capable de travailler sur des données mixtes (audio, textuelle).

**Q1 = comment peut-on faire du SA sur la voix d'un humain ?**

- . dans les données de la voix, il y a 2 marqueurs utiles ( $X_i$ ), la fréquence de la voix et le rythme de la parole,
- . sur du signal audio, on a des courbes de signal, des spectrogrammes, on fait des fenêtres en « t » pour extraire des séries temporelles,
- . il y a 3 types d'étiquettes pour qualifier un locuteur : **frustration, satisfaction et neutre** (i.e. = problème multi-classes),
  - . pour mesurer l'émotion dans la voix, on utilise 2 métriques :
    - l'état émotionnel de la personne : positif vs. négatif → métrique = **valence**,
    - la force de l'activation de l'état : faible vs. fort → métrique = **arousal**,

**Q2 = concrètement comment fais-tu ta classification ?**

- . le plus gros problème est de trouver / construire un dataset pour entraîner le modèle car pour un fichier de 100 courbes audio, on a seulement 10 courbes audio avec des étiquettes exploitables → dans notre métier, on dit « qu'on a pas de Corpus »,
- . le 2<sup>e</sup> problème c'est la puissance de calcul pour former l'apprentissage : pour un dataset de 30 heures d'audio, avec un modèle « end-to-end » et un cluster de 4 GPU, l'entraînement se forme en 1.5 jours,

**Q3 = quelle architecture de classifieur utilises-tu dans ton modèle end-to-end baidu deep speech 2 ?**

- . c'est toujours la même, plusieurs layers définis comme il suit : CNN + RNN + softmax,
- . c'est l'architecture du réseau qui fonctionne le mieux,
- . ce modèle est surtout utilisé en recherche,

**Q4 = connais-tu d'autres solutions pour faire du SA ?**

- . oui, il existe la solution <https://www.audeering.com/opensmile/> → par contre, c'est très expérimental...

**Q5 = qui sont les meilleurs chercheurs sur ce domaine du voice-computing ?**

- . pour la France, le meilleur labo, c'est mon labo du Mans,
- . à l'international, « les chinois » sont « forts » dans le domaine.



L'effet Enceintes connectées : demain, tous assistés ?

30.09.19

## “Alexa, tell me what you heard.”

- Alexa peut contrôler votre Wi-Fi et vous informer des devoirs de votre enfant
- Voix de célébrités
- Soutien à domicile multilingue
- Sonnette Concierge
- La nouvelle détection d'Alexa Guard à domicile
- Commandes de confidentialité Alexa
- Alexa Communications pour les enfants
- Vidéo exclusive du réseau alimentaire et cours de cuisine en direct

<https://venturebeat.com/2019/09/25/alexa-can-soon-control-your-wi-fi-and-brief-you-on-your-kids-homework/>

le sujet de Machine Learning  
qui monte en entreprise ...



The screenshot shows the homepage of the Voicetech Paris website. At the top, there is a navigation bar with links: A propos, Programme, Speakers, Exposition, Networking, Content Factory, and Infos pratiques. Below the navigation, a yellow banner displays the dates "26-27 novembre 2019" and the location "Salons de l'Aveyron · Paris". The main title "VOICETECH corp PARIS" is prominently displayed, with "corp" in blue and "PARIS" in white. Below the title, the slogan "OUVRONS LA [VOIX] AUX USAGES DE DEMAIN" is written in large, bold, white letters. A blue arrow points to the right from the left side of the page. At the bottom, there is a purple button labeled "Inscription".

<https://www.voicetechparis.com/2019/>

## Définition(s)

### Copule gaussienne

Une **Copule gaussienne** est une mesure de dépendance entre deux variables, introduite en finance par David X. Li en 2000. La dépendance est décrite de la même façon que la distribution normale (d'où le nom de « gaussienne »), mais avec des distributions marginales arbitraires.

Les Copules gaussiennes ont été très largement utilisées **pour modéliser la corrélation** des défauts entre plusieurs obligations et donc les pertes potentielles d'un panier. Elles ont naturellement servies à l'évaluation de CDOs (Collateralized Debt Obligations).

<http://financedemarche.fr/definition/copule-gaussienne>

### Qu'est-ce le PI Planning en SAFe ?

Le PI Planning (PI pour **Program Increment**) est la cérémonie qui permet d'aligner l'ensemble des équipes de l'Agile Release Train (Art) sur la vision et les objectifs pour une durée de 8 semaines à 12 semaines (8 étant recommandé par le framework).

<https://blog.myagilepartner.fr/index.php/2018/06/14/le-pi-planning-en-safe/>